Hierarchical Prior Models and Krylov-Bayes Iterative Methods: Part 2

Daniela Calvetti

Case Western Reserve University

- D Calvetti, E Somersalo: Priorconditioners for linear systems. Inverse Problems 21 (2005) 13971418
- D Calvetti, F Pitolli, E Somersalo, B Vantaggi: Bayes meets Krylov: preconditioning CGLS for underdetermined systems http://arxiv.org/abs/1503.06844

In the mood for Bayes

In the Bayesian framework for linear discrete inverse problems we need to solve a linear system of equations

- The linear solver apparatus (least-squares solvers, iterative methods, etc) can be used to update z in the MAP calculation
- Conversely, the Bayesian framework can be very helpful to solve linear systems of equations
- More generally, the Bayesian approach can be used in the solution of nonlinear systems
- The Bayesian approach is particularly well-suited for under-determined linear systems

General setting

Consider the problem of estimating $x \in \mathbb{R}^n$ from

$$b = \mathbf{F}(x) + \epsilon, \qquad \mathbf{F} : \mathbb{R}^n \longrightarrow \mathbb{R}^m,$$

Here we focus the attention on the special case where

$$\mathbf{F}(x) = \mathrm{A}x$$

with A is an $m \times n$ matrix of rank m, typically badly conditioned and of ill-determined rank.

We solve the linear system with a Krylov subspace iterative method.

Bayesian solution of inverse problems

In the Bayesian framework for the solution of inverse problems,

- All unknown parameters are modeled as random variables and described in terms of their probability density functions;
- Here the unknowns are ϵ and x
- $\pi_{\mathrm{noise}}(\epsilon)$ describes what we know about the statistics of the noise and defines the *likelihood*;
- $\pi_{\mathrm{prior}}(x)$ expresses what we know about x before taking into consideration the data and is called the *prior*
- The solution of the inverse problem is $\pi(x \mid b)$ and is called the *posterior*.

It follows from Bayes' formula that

$$\pi(x \mid b) \propto \pi_{\mathrm{prior}}(x)\pi_{\mathrm{noise}}(b - Ax).$$

The noise in a Bayesian way

The linear discrete inverse problem that we consider is $b = Ax + \epsilon$.

- The noise term ϵ accounts for inaccuracies in the measurements as well as model uncertainties, i.e. discrepancy between reality and the model. (more on this tomorrow)
- We assume that $\epsilon \sim \mathrm{N}(0, I_m)$.
- If $\epsilon = \epsilon_c \sim \mathrm{N}(\mu_\epsilon, \Gamma)$, we can proceed as follows:
  - Compute a symmetric factorization of the precision matrix of ϵ $\Gamma^{-1} = G^T G$
  - Make the change of variables

$$\epsilon = G(\epsilon_c - \mu_c) = G(b - Ax) - G\mu_c$$

- In the linear system

$$Gb = GAx + \epsilon$$

  the noise is zero-mean white Gaussian.

The unknown in a Bayesian way

Assume that $x \sim \mathrm{N}(0, C)$, where C is symmetric positive definite. It follows from Bayes formula that the posterior density is of the form

$$\pi(x \mid b) \propto \exp\left(-\frac{1}{2}\|\mathrm{A}x - b\|^2 - \frac{1}{2}x^T C^{-1} x\right).$$

Give a symmetric factorization of the precision matrix of x

$$C^{-1} = B^{\mathsf{T}} B$$

we can write the negative logarithm of the posterior, or *Gibbs energy* in the form

$$G(x) = \|Ax - b\|^2 + \|Bx\|^2 = \left\| \begin{bmatrix} A \\ B \end{bmatrix} x - \begin{bmatrix} b \\ 0 \end{bmatrix} \right\|^2.$$

MAP estimate

The Maximum a Posteriori (MAP) estimate of x, x_{MAP} is the value of highest posterior probability, or equivalently, the minimizer of G(x) .

The value of x_{MAP} is the *least squares* solution of the linear system

$$\begin{bmatrix} A \\ B \end{bmatrix} x = \begin{bmatrix} b \\ 0 \end{bmatrix},$$

or, equivalently, the solution of the square linear system

$$(A^{\mathsf{T}}A + B^{\mathsf{T}}B)x = A^{\mathsf{T}}b.$$

… and Tikhonov regularization

The latter are the normal equations associated with the problem

$$x_{\mathrm{MAP}} = \mathrm{argmin}\left\{\|Ax - b\|^2 + \lambda\|Bx\|^2\right\} \qquad (1)$$

which is Tikhonov regularized solution with regularization parameter $\lambda = 1$ and linear regularization operator B.

- The computation of Tikhonov regularized solution with B different from I requires attention: moreover, a suitable value of the regularization parameter λ must be chosen.
- This question has been studied extensively in the literature.

Regularization operator beyond Tikhonov

- The operator B brings into the solution additional information about x
- When the matrix A is underdetermined, the operator B boosts the rank of the matrix of the linear system actually solved.
- In Tikhonov regularization when n is large the introduction of B may lead to a very large linear system

Question: How can we retain the benefits of B while containing the computational costs?

The alternative: Krylov subspace methods

As an alternative to Tikhonov regularization consider solving the linear system

$$b = \mathsf{A}x + \epsilon$$

with an iterative solver using the matrix B as a right preconditioned. More specifically

- Consider the Conjugate Gradient for Least Squares method (CGLS)
- WLOG assume the initial approximate solution is $x_0 = 0$
- Define a termination rule based on the discrepancy

Standard CGLS method

At the kth iteration step the approximate solution x_k computed by the CGLS method satisfies

$$x_k = \mathrm{argmin}\big\{\|b - \mathsf{A}x\| \mid x \in \mathscr{K}_k\big\},$$

where the kth Krylov subspace is

$$\mathscr{K}_k = \mathrm{span}\big\{\mathsf{A}^\mathsf{T} b, (\mathsf{A}^\mathsf{T}\mathsf{A})\mathsf{A}^\mathsf{T} b, \ldots, (\mathsf{A}^\mathsf{T}\mathsf{A})^k \mathsf{A}^\mathsf{T} b\big\}.$$

The noise is additive, zero-mean white Gaussian, thus

$$E\left\{\|\epsilon\|^2\right\} = m;$$

we stop iterating as soon as

$$\|\mathsf{A}x - b\|^2 < \tau m,$$

where $\tau = 1.2$. Typically, $k_{\mathrm{last}} \ll m$.

The question of the null space

- It follows from the canonical orthogonal decomposition in terms of fundamental subspaces that

$$\mathbb{R}^n = \mathscr{N}(\mathsf{A}) \oplus \mathscr{R}(\mathsf{A}^\mathsf{T})$$

- In standard CGLS any contribution to the solution from the null space must be added separately
- The right CGLS priorconditioner implicitly selects null space components based on the information contained in the data with the belief about x.

CGLS with a whitened unknown

Assume that a prior we believe that $x \sim N(0, \mathsf{C})$. If

$$\mathsf{C}^{-1} = \mathsf{B}^{\mathsf{T}}\mathsf{B}$$

then

$$w = Bx, \qquad w \sim \mathrm{N}(0, I_n).$$

Make the change of variable from x to w in the linear system

$$AB^{-1}w = b \qquad x = B^{-1}w, \tag{2}$$

let $\widetilde{\mathsf{A}} = \mathsf{A}\mathsf{B}^{-1}$ and solve by CGLS for w. The jth iterate of the whitened problem satisfies

$$w_j = \operatorname{argmin}\big\{\|\widetilde{\mathsf{A}}w - b\| \mid w \in \mathcal{K}_j(\widetilde{\mathsf{A}}^{\mathsf{T}}b, \widetilde{\mathsf{A}}^{\mathsf{T}}\widetilde{\mathsf{A}})\big\}.$$

Priorconditioning and the null space

The corresponding jth priorconditioned CGLS solution $\widetilde{x}_j = \mathsf{B}^{-1} w_j$ satisfies

$$\widetilde{x}_j \in \operatorname{span}\left\{\mathsf{B}^{-1}(\widetilde{A}^\mathsf{T}\widetilde{A})^\ell \widetilde{A}^\mathsf{T} b \mid 0 \le \ell \le j-1\right\}.$$

It follows from

$$\mathsf{B}^{-1}\widetilde{A}^\mathsf{T} = \mathsf{B}^{-1}\mathsf{B}^{-\mathsf{T}}\mathsf{A}^\mathsf{T} = \mathsf{C}\mathsf{A}^\mathsf{T},$$

that

$$\mathsf{B}^{-1}\left(\widetilde{A}^\mathsf{T}\widetilde{\mathsf{A}}\right)^\ell \widetilde{\mathsf{A}}^\mathsf{T} = \left(\mathsf{C}\mathsf{A}^\mathsf{T}\mathsf{A}\right)^\ell \mathsf{C}\mathsf{A}^\mathsf{T}, \quad 0 \le \ell \le j-1.$$

Therefore

$$\widetilde{x}_j \in \mathsf{C}\left(\mathscr{N}(\mathsf{A})^\perp\right),$$

hence $\widetilde{x}_j$ is not necessarily orthogonal to the null space of A.

Analizing the Krylov subspaces with the GSVD

Theorem
Given (A,B) with $\mathsf{A} \in \mathbb{R}^{m\times n}$, $\mathsf{B} \in \mathbb{R}^{n\times n}$, $m < n$, there is a factorization of the form

$$\mathsf{A} = \mathsf{U}\begin{bmatrix} 0_{m,n-m} & \Sigma_\mathsf{A} \end{bmatrix}\mathsf{X}^{-1}, \quad \mathsf{B} = \mathsf{V}\begin{bmatrix} \mathsf{I}_{n-m} & \\ & \Sigma_\mathsf{B} \end{bmatrix}\mathsf{X}^{-1},$$

called the *generalized singular value decomposition*, where $\mathsf{U} \in \mathbb{R}^{m\times m}$ and $\mathsf{V} \in \mathbb{R}^{n\times n}$ are orthogonal matrices, $\mathsf{X} \in \mathbb{R}^{n\times n}$ is invertible, and $\Sigma_\mathsf{A} \in \mathbb{R}^{m\times m}$ and $\Sigma_\mathsf{B} \in \mathbb{R}^{m\times m}$ are diagonal matrices.

The diagonal entries $s_1^{(\mathrm{A})}, \ldots, s_m^{(\mathrm{A})}$ and $s_1^{(\mathrm{B})}, \ldots, s_m^{(\mathrm{B})}$ of the matrices Σ_{A} and Σ_{B} are real, nonnegative and satisfy

$$\begin{aligned} s_1^{(\mathrm{A})} &\leq s_2^{(\mathrm{A})} \leq \ldots \leq s_m^{(\mathrm{A})} \\ s_1^{(\mathrm{B})} &\geq s_2^{(\mathrm{B})} \geq \ldots \geq s_m^{(\mathrm{B})} \\ (s_j^{(\mathrm{A})})^2 &+ (s_j^{(\mathrm{B})})^2 = 1, \qquad 1 \leq j \leq m. \end{aligned} \tag{3}$$

thus $0 < s_j^{(\mathrm{A})} \leq 1$ and $0 < s_j^{(\mathrm{B})} \leq 1$. The ratios $s_j^{(\mathrm{A})}/s_j^{(\mathrm{B})}$ for $1 \leq j \leq m$ are the generalized singular values of (A, B).
If A has full rank, the diagonal entries of Σ_{A} are positive.

C-orthogonality

Theorem

If we partition the matrix $\mathsf{X} \in \mathbb{R}^{n\times n}$ in GSVD above as

$$\mathsf{X} = \left[\begin{array}{cc} \mathsf{X}' & \mathsf{X}'' \end{array}\right], \quad \mathsf{X}' \in \mathbb{R}^{n\times(n-m)},\ \mathsf{X}'' \in \mathbb{R}^{n\times m},$$

it follows that

$$\mathrm{span}\{\mathsf{X}'\} = \mathscr{N}(\mathsf{A}),$$

and we can express $\mathbb{R}^n$ as a C-orthogonal direct sum,

$$\mathbb{R}^n = \mathrm{span}\{\mathsf{X}'\} \oplus_{\mathsf{C}} \mathrm{span}\{\mathsf{X}''\} = \mathscr{N}(\mathsf{A}) \oplus_{\mathsf{C}} \mathrm{span}\{\mathsf{X}''\}.$$

Orthogonality and not

Corollary 1

$$\mathscr{N}(\mathsf{A})^{\perp} = \mathscr{R}(\mathsf{A}^{\mathsf{T}}), \quad \mathscr{N}(\mathsf{A})^{\perp \mathsf{c}} = \mathrm{span}\{\mathsf{X}''\}.$$

Corollary 2
If $\mathscr{R}(\mathsf{A}^{\mathsf{T}})$ is an invariant subspace of the covariance matrix C, then the iterates $\widetilde{x}_j$ are orthogonal to the null space of A.

Corollary 3
When $\mathsf{C}\big(\mathscr{R}(\mathsf{A}^{\mathsf{T}})\big)$ is not C-orthogonal to $\mathscr{N}(\mathsf{A})$, $\widetilde{x}_j$ may have a component in the null space of A. This component is invisible to the data.

Priorconditionting and the Lanczos process

The first k residual vectors computed by CGLS normalized to have unit length $v_0, v_1, \ldots, v_{k-1}$ form an orthonormal basis for the Krylov subspace $\mathcal{K}_k(\mathsf{A}^\mathsf{T} b, \mathsf{A}^\mathsf{T}\mathsf{A})$.

It can be shown that

$$\mathsf{A}^\mathsf{T}\mathsf{A} V_k = V_k \mathsf{T}_k - \frac{\sqrt{\beta_{k-1}}}{\alpha_{k-1}} v_k e_k^\mathsf{T}., \qquad \mathsf{V}_k = [v_0, v_1, \ldots, v_{k-1}].$$

It follows from the orthogonality of the v_j that the tridiagonal matrix T_k is the projection of $\mathsf{A}^\mathsf{T}\mathsf{A}$ onto the Krylov subspace $\mathcal{K}_k(\mathsf{A}^\mathsf{T} b, \mathsf{A}^\mathsf{T}\mathsf{A})$.

$$\mathsf{V}_k^\mathsf{T}(\mathsf{A}^\mathsf{T}\mathsf{A}) V_k = \mathsf{T}_k.$$

The Lanczos tridiagonal matrix

The kth CGLS iterate can be expressed as

$$x_k = \mathsf{V}_k y_k,$$

where y_k solves the $k \times k$ linear system

$$\mathsf{T}_k y = \|r_0\| e_1.$$

Thus the kth CGLS iterate x_k is the lifting of y_k via V_k.

The eigenvalues of T_k are the Ritz values of $\mathsf{A}^\mathsf{T}\mathsf{A}$ and approximate of the eigenvalues of $\mathsf{A}^\mathsf{T}\mathsf{A}$.

Ritz values and convergence rate

Theorem
For all k, $1 \leq k \leq r$, where r is the rank of A, there exists $\xi_k, \lambda_1 \leq \xi_k \leq \lambda_r$ such that the norm of the residual vector satisfies

$$\|r_k\|^2 = \frac{1}{\xi_k^{2k+1}} \sum_{i=1}^{n} \left[\prod_{j=1}^{k} \left(\lambda_i - \theta_j^{(k)} \right)^2 \right] \left(r_0^{\mathsf{T}} q_i \right)^2 ,$$

where q_i is the eigenvector of $\mathrm{A}^{\mathsf{T}}\mathrm{A}$ corresponding to the eigenvalue λ_i, and $\theta_j^{(k)}$ is the jth eigenvalue of the tridiagonal matrix T_k.

The quality of the eigenvalues approximations in the projected problem affects the number of iterations needed to meet the stopping rule.

A simple deconvolution problem

Forward model: Deconvolution problem with few data,

$$g(t) = \int_0^1 a(t-s)f(s)ds, \quad a(t) = \left(\frac{J_1(\kappa t)}{\kappa t}\right)^2,$$

Discretize:

$$g(t) \approx \frac{1}{n}\sum_{k=1}^{n} a(t-s_k)f(s_k), \quad 1 \leq j \leq n,$$

Discrete noisy observations at $t_1, \ldots, t_m$, $m \ll n$.

$$b_\ell = g(t_\ell) + \varepsilon_\ell, \quad 1 \leq \ell \leq m,$$

or, in matrix notation, $\mathrm{A} \in \mathbb{R}^{m \times n}$,

$$b = \mathrm{A}x + \varepsilon, \quad x_k = f(s_k).$$

Computed examples: Deconvolution

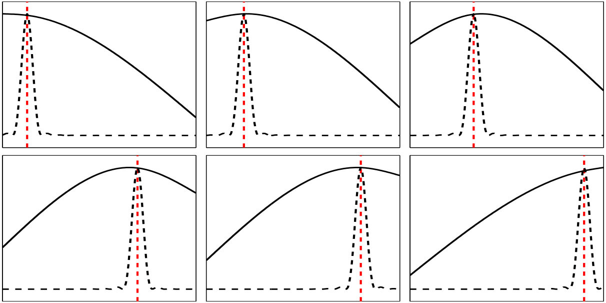
Prior: Define the precision matrix C^{-1} as

$$\mathsf{C}^{-1} = \mathsf{B}^{\mathsf{T}}\mathsf{B}, \quad \mathsf{B} = \beta \begin{bmatrix} \alpha & & & & \\ -1 & 2 & -1 & & \\ & & \ddots & & \\ & & -1 & 2 & -1 \\ & & & & \alpha \end{bmatrix},$$

where $\alpha > 0$ is chosen so that prior variance is as uniform as possible over the interval.

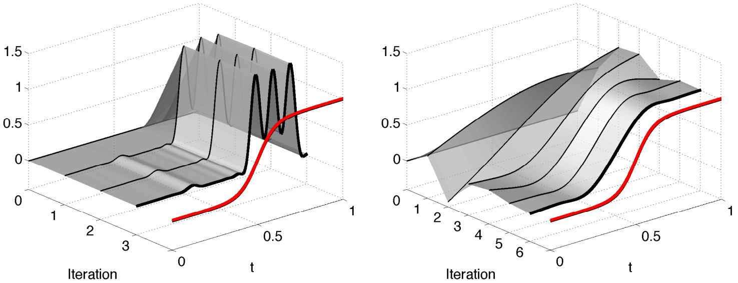
Parameters: Set $n = 150$, $m = 6$.

Basis vectors



The six basis vectors that span $\mathscr{R}(\mathsf{A}^\mathsf{T})$ (dashed line), and the vectors that span $\mathsf{C}\big(\mathscr{R}(\mathsf{A}^\mathsf{T})\big)$ (solid line).

Approximate solutions

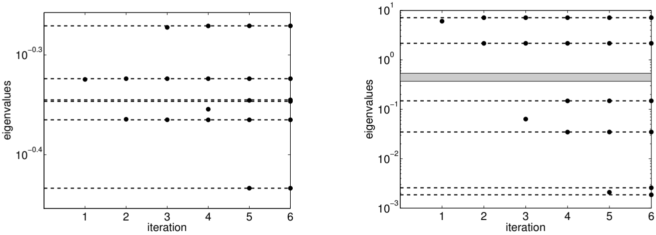


Iterations with low additive noise ($\sigma = 5 \times 10^{-5}$) without prior conditioner (left) and with preconditioner.

Observations

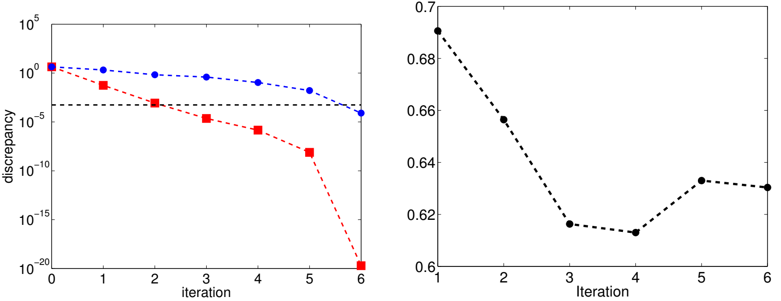
- Every vector whose support consists of points where all six basis functions of $\mathscr{R}(\mathsf{A}^\mathsf{T})$ vanish is in the null space $\mathscr{N}(\mathsf{A})$
- Consequently, plain CGLS produces approximate solutions that are zero at those points
- The basis functions of $\mathsf{C}\big(\mathscr{R}(\mathsf{A}^\mathsf{T})\big)$ are non-zero everywhere
- Consequently, priorconditioned CGLS has no blind spots
- The price to pay is that priorconditioned CGLS requires more iterations

Spectral approximation



Spectral approximation: Plain CGLS (left) and preconditioned CGLS (right). The grey band on the right is the spectral interval of the non-preconditioned matrix $A^T A$.

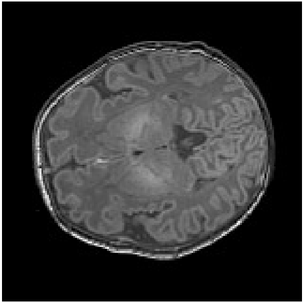
Convergence history and null space contributions



Left: Convergence rates of the two algorithms. The dashed line marks the stopping criterion. Right: Component of the computed solution in the null space measured as

$$\nu_k = \frac{\|\mathsf{P}\widetilde{x}_k\|}{\|\widetilde{x}_k\|}, \quad \mathsf{P} : \mathbb{R}^n \longrightarrow^{\perp} \mathscr{N}(\mathsf{A}).$$

Computed examples: X-ray tomography



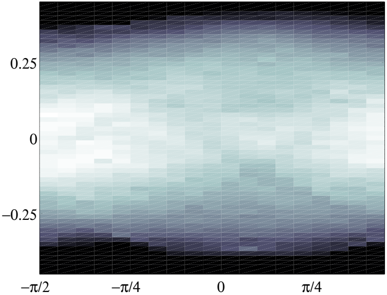


Image size: $N = 160 \times 160$ pixels. 20 illumination angles, 60 parallel beams per illumination angle.

Correlation priors

Matèrn-Whittle correlation priors: Define the precision matrix as

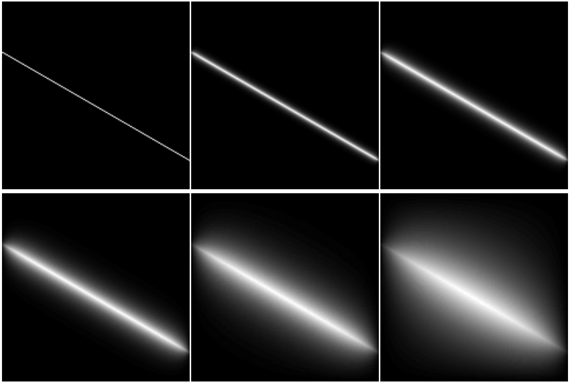
$$\mathsf{C}^{-1} = -\mathsf{I}_n \otimes \mathsf{D} - \mathsf{D} \otimes \mathsf{I}_n + \frac{1}{\lambda^2}\mathsf{I}_N,$$

where $\mathsf{D} \in \mathbb{R}^{n \times n}$ is the three-point finite difference approximation of the one-dimensional Laplacian with Dirichlet boundary conditions,

$$\mathsf{D} = \frac{1}{n^2}\begin{bmatrix} -2 & 1 & & \\ 1 & -2 & \ddots & \\ & \ddots & & 1 \\ & & 1 & -2 \end{bmatrix},$$

and $\lambda > 0$ is the correlation length.

Basis functions

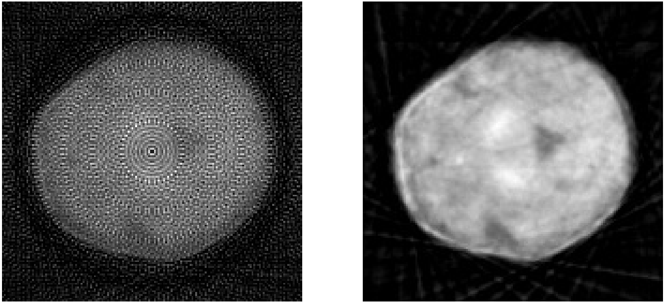


Basis vector with no priorconditioning (upper left) and with priorconditioning. Correlation length 2, 4, 8, 16 and 32 pixels.

Observations

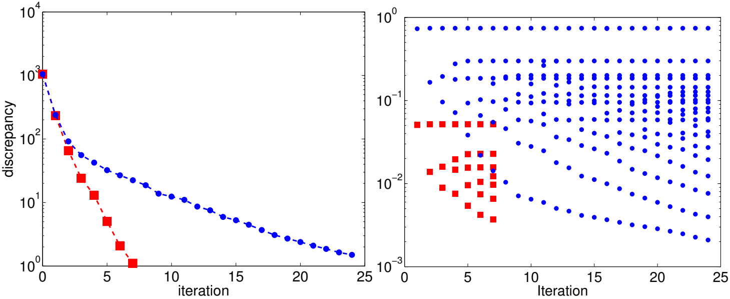
- Every image whose support is on pixels not touched by a ray is in the null space of A
- $\Rightarrow$ plain CGLS iterates will be zero at those pixels
- Preconditioning makes the rays fuzzy, illuminating the dark pixels
- Reconstruction will be slightly blurred, but has fewer geometric artifacts
- Number of iterations needed will increase.

Computed solutions



Reconstructions with plain CGLS (left) and priorconditioned CGLS (right).

Converge history and spectral approximation



Convergence and spectral approximation.