Watching CG's convergence

Marc Aurele Gilles

May 3, 2017

1 Introduction

The convergence of Krylov subspace methods is notoriously hard to understand and reason about. For some algorithms, bounds on convergence exist but are typically not informative, as the actual convergence is much faster than the one predicted by the bound. The convergence of Krylov subspace method is so unpredictable that the advice given to practitioneer when they would like to know if a particular method or preconditioner would be effective for a specific problem is to try it out and see if it works. In this report we state a theorem which relates the convergence of a particular Krylov subspace method: the Conjugate-Gradient (CG) method to the approximation of an integral through Gauss-Christoffel quadrature. Neither of these concepts is particularily intuitive, but the Gauss-Christoffel quadrature has a nice geometric interpretation, which allows us to visualize it. The hope is that the we can get some intuition about the convergence of CG for a specific problem by watching the associated Gauss-Christoffel quadrature problem. The rest of the report is organized as follows: in section 2 we briefly decribe CG and some of its properties, in section 3 we define the Gauss-Christoffel quadrature and in section 4 we link the two notions using the theorem and provide a visualization.

2 The Conjugate-Gradient Method

Krylov subspace methods are iterative algorithms for solving linear systems $Ax = b$. They operate by approximating the exact solution of the linear system by the best solution over low dimensional subspaces, called Krylov subspaces. The k-th Krylov subspace of A and r_0 is:

$$\mathcal{K}_k(A, r_0) = \text{span}\left\{r_0, Ar_0, \ldots, A^{k-1}r_0\right\} . \tag{1}$$

The vector r_0 is the initial residual $r_0 = b - Ax_0$, associated with some initial guess x_0.

The conjugate gradient method is a Krylov subspace method which applies when the matrix A is symmetric positive definite. At the k-th iteration, the CG method computes the approximate solution x_k that has the smallest error

$$e_k = x - x_k \tag{2}$$

as measured in the A-norm defined by $\langle x, y \rangle_A = \langle x, Ay \rangle$ where $\langle x, y \rangle$ denotes the standard Euclidean inner product over $\mathbb{R}^n$. That is, x_k is uniquely defined as:

$$x_k = \min_{x_k \in \mathcal{K}_k(A, r_0)} \|x - x_k\|_A \tag{3}$$

Some algebra on this expression leads to the 3-term recurrence which us allows to compute x_{k+1} using only information from the previous two iterations . In this report we ignore many details that come into the efficient implementation of CG such as preconditioning or stopping criteria.

2.1 Convergence Bounds

A well-known convergence bound of the Conjugate Gradient method is:

$$\|e_k\|_A \leq 2\|e_0\|_A \left(\frac{\sqrt{\kappa(A)} - 1}{\sqrt{\kappa(A)} + 1} \right)^k \tag{4}$$

where $\kappa(A) = \frac{\lambda_{\max}}{\lambda_{\min}}$ denotes the condition number of A. Even though the CG method has the best convergence theory out of all the Krylov subspace method, it is still a rather poor one for several reasons:

1. It ignores the discrete nature of the spectrum of the matrix A. It is known that the discrete distribution of the eigenvalues of A is a very important to the convergence of CG. For example, if $A \in \mathbb{R}^{n \times n}$ has only two distinct eigenvalues λ_1, λ_2, then its condition number can be made arbitrarily large by making the ratio λ_2/λ_1 large leading to a very weak bound, but CG will always converge in two steps.

2. The bound ignores the contribution of the right hand side. We know that the convergence of CG depends strongly on the right hand side. In particular if b can be written down as a linear sum of K eigenvectors of A, then CG provides the exact solution in at most K steps.

3. It is usually a very bad bound.

Unfortunately the convergence of Krylov subspace method is hard to think about. Better bounds than the one in 4 can be derived for specific problems, but each case requires a lot of work [2, 1]. Even if we can compute the exact solution and observe the true convergence curve, it is often difficult to understand the behavior of CG on a specific example. This motivates the theorem described in section 4, which provides a visual way to understand the convergence of CG.

3 Approximation of Riemann-Stieltjes Integrals

In this section we introduce the notion of Riemann-Stieltjes Integrals, and their approximation by quadrature. We do not give the most general definitions of the concepts, but merely the ones needed to make sense of the theorem described in section 4.

3.1 Riemann-Stieltjes integrals

Let $\omega(\lambda)$ be a non-decreasing function on $[a, b] \subset \mathbb{R}$ with $\omega(a) = 0$, $\omega(b) = 1$, and n points of increase at $\{\lambda_i\}_{i=1}^n$. That is $\omega(\lambda)$ is of the form:

$$\omega(\lambda) = \begin{cases} 0 & \text{if } a \leq \lambda \leq \lambda_1 \\ \sum_{j=1}^{i} \omega_j & \text{if } \lambda_i \leq \lambda \leq \lambda_{i+1}, \quad i = 1, \ldots, n-1 \\ \sum_{j=1}^{n} \omega_j = 1 & \text{if } \lambda_n \leq \lambda \leq b \end{cases} \tag{5}$$

where $\omega_j \geq 0$. The function $\omega(\lambda)$ is often refered to as a distribution function. Figure 1 shows an example of a distribution function.

Then for some function f, we can define the Riemann-Stieljes integral of f with respect to ω as:

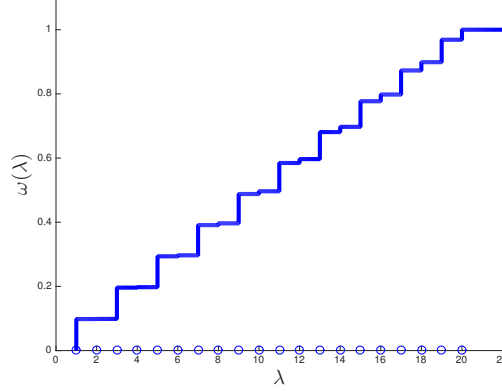$$\int_a^b f(\lambda)d\omega(\lambda) = \sum_{i=1}^n f(\lambda_i)\omega_i \tag{6}$$

Figure 1: An example distribution function $\omega(\lambda)$

If $f > 0$, the Riemann-Stieljes integral can be visualized as the area under the curve defined by $(x, y) = (\omega(\lambda), f(\lambda))$ for $\lambda \in [a, b]$. Figure 2 shows a visualization of the Riemann-Stieljes integral of $f(\lambda) = \frac{1}{\lambda}$ with respect to the distribution function shown in figure 1.
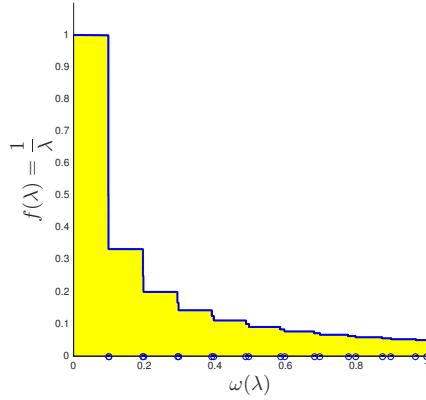

Figure 2: Riemann–Stieltjes integral of $f(\lambda) = \frac{1}{\lambda}$ with respect to the distribution function in figure 1. The blue curve is the curve $(x, y) = (\omega(\lambda), f(\lambda))$, and the area of the yellow shaded region is the Riemann-Stieltjes integral.

3.2 Gauss-Christoffel Quadrature

Suppose we would like to approximate the Riemann-Stieltjes integral by a weighted sum of k samples of the functions. That is, for some $\theta_j^{(k)}$ (called nodes), and $w_j^{(k)}$ (called weights) , $j \in \{1, \ldots, k\}$, we would like to approximate the Riemann Stieljes integral by:

$$\int_a^b f(\lambda) d\omega(\lambda) \approx \sum_{i=1}^k w_i^{(k)} f(\theta_i^{(k)}) \tag{7}$$

3

We define the signed error of the approximation in the following way:

$$E_k(f) = \int_a^b f(\lambda)d\omega(\lambda) - \sum_{i=1}^k w_i^{(k)} f(\theta_i^k) \tag{8}$$

$$\iff \int_a^b f(\lambda)d\omega(\lambda) = \sum_{i=1}^k w_i^{(k)} f(\theta_i^{(k)}) + E_k(f) \tag{9}$$

This naturally raises the question: where should we put the nodes (i.e. where should we sample the function)? What should the weights be?

Note that the answer is trivial if k, the number of nodes, is equal to n, the number of points of increase of ω, since in that case we can set $\theta_j^{(k)} = \lambda_j$ and $\omega_j^{(k)} = w_j$ and we have:

$$\int_a^b f(\lambda)d\omega(\lambda) \coloneqq \sum_{i=1}^n \omega_i f(\lambda_i) = \sum_{i=1}^k w_i^{(k)} f(\theta_i^k) \tag{10}$$

for all functions f. For other cases, one way to define a "best" quadrature rule is to impose that the integral be exact for low degree polynomials. If this is the choice we make for optimality, then Gauss-Christoffel quadrature is a rule which provides the optimal choice. The k-node Gauss-Christoffel quadrature is exact for polynomials of degree $2k - 1$, which can be shown to be optimal. That is, if we pick $\theta_j^{(k)}$, $\lambda_j^{(k)}$ according to the Gauss-Christoffel quadrature rule then for all polynomials p of degree at most $2k - 1$:

$$\int_a^b p(\lambda)d\omega(\lambda) = \sum_{i=1}^k w_i^{(k)} f(\theta_i^k) \tag{11}$$

$$\iff E_k(p) = 0 \tag{12}$$

It can be shown that the weights w_j are non-negative and sum up to one, and as a consequence the nodes and weights nodes of the k-node Gauss-Christoffel quadrature define a new distribution function $w^k(\lambda)$:

$$w^k(\lambda) = \begin{cases} 0 & \text{if } a \le \lambda \le \theta_1^k \\ \sum_{j=1}^i w_i^{(k)} & \text{if } \theta_i^k \le \lambda \le \theta_{i+1}^k, \quad i = 1, \dots, k-1 \\ \sum_{j=1}^n w_j^k = 1 & \text{if } \theta_k^k \le \lambda \le b \end{cases} \tag{13}$$

The error in the Gauss-Christoffel quadrature of the function f, $E_k(f)$, can be visualized using this new distribution function. It signed area between the curves defined $(x,y) = (\omega(\lambda), f(\lambda))$ and $(x,y) = (\Theta^k(\lambda), f(\lambda))$. Figure 3 shows a visualization of 1-node Gauss-Christoffel quadrature.

4 Linking CG and Gauss-Christoffel Quadrature

4.1 A special distribution function

In this section we define a special distribution function in terms of a matrix A and a vector r_0. Let $A = \sum_{i=1}^n \lambda_i v_i v_i^T$ be an eigendecomposition of $A \in \mathbb{R}^{n \times n}$. We define the distribution function $\omega(\lambda)$ by setting the points of increase to λ_i, the eigenvalues of A, and setting $\omega_i = \left(v_i^T \frac{r_0}{\|r_0\|_2} \right)^2$.
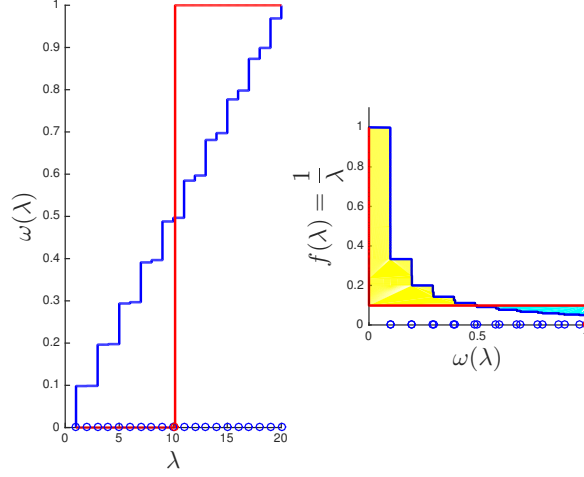
Figure 3: Example of 1-node Gauss-Christoffel quadrature. Left figure: the blue plot is $\omega(\lambda)$ and the red plot is $w^1(\lambda)$. Right figure: the blue plot is the curve $(x, y) = \left(\omega(\lambda), \frac{1}{\lambda} \right)$, the red plot is $(x, y) = \left(w^1(\lambda), \frac{1}{\lambda} \right)$. The yellow and cyan regions are the regions of positive error and negative error respectively between the true integral and Gauss-Christoffel quadrature. The net error $E_1(f)$ is the area of the yellow region minus the area of the cyan region

For example, we take A to be the finite-difference discretization of the differential operator

$$L[u] = -\frac{\partial^2 u}{\partial x^2} - \frac{\partial^2 u}{\partial y^2} + \frac{1}{2}u \quad \text{in } [-1, 1] \times [-1, 1] \tag{14}$$

with homogenenous dirichlet boundary conditions, and b to be the vector of all ones. The resulting distribution function $\omega(\lambda)$ is plotted is the one plotted in figure 1.

4.2   Theorem by G. Meurant, 1997

We are now ready to state the theorem linking the conjugate gradient method to Gauss-Christoffel quadrature:

Theorem 1 (G. Meurant, 1997) *The squared A-norm of the nth error in the CG method, scaled by the squared Euclidean norm of the initial residual, is equal to the approximation error of the n-node Gauss-Christoffel quadrature for $f(\lambda) = \lambda^{-1}$, with the distribution function defined by the A matrix and the normalized initial residual. i.e.:*

$$\int \lambda^{-1} d\omega(\lambda) = \sum_{i=1}^{n} \omega_i^{(k)} \{\theta_i^{(k)}\}^{-1} + \frac{\|x - x_k\|_A^2}{\|r_0\|^2} \tag{15}$$

The theorem was proved by Gerard Meurant in 1997, but a completely equivalent statement which relates the Jacobi matrices which arise by doing Lanczos algorithm to CG was apparently already known by Stiefel, one of the inventor of the Conjugate-Gradient method [3].

In this report we view this theorem as a theoretical tool that provides a visualization, not as a useful bound for computation. Indeed, the computation of Gauss-Christoffel quadrature is

much more expensive computationally than to run CG since it involves an eigenvalue computation. Nevertheless, this theorem has been used to derive estimates on the error which are computable during the iteration, and can provide good stopping criteria to the iteration. The resulting algorithm is called CGQL (Conjugate Gradient with Quadrature and Lanczos) [5], and was propsed by Gerard Meurant.

Note that this theorem does not have any of the issues of the bound in equation 4 that were pointed out in section 2: it takes into account the discrete distribution of eigenvalues, depends heavily on the right hand side, and by definition is tight since the theorem is an equality.

4.3 Visualization

We consider the matrix A defined as the finite difference approximation of (14), the vector b of all ones, and an initial guess of $x_0 = 0$. Figure 4 shows the visualization for this choice of A and b.

In each subfigure:

1. The left plot shows the distribution function $\omega(\lambda)$ in blue, the distribution function obtained $w^{(k)}(\lambda)$ corresponding to the Gauss-Christoffel quadrature in red. The blue circles on the x-axis indicate the eigenvalues λ_i of the matrix A, and the red circles on the x-axis indicate the nodes $\theta_i^{(k)}$.

2. The middle plot is depiction of the Riemann-Stieltjes integrals of the function $f(\lambda) = \frac{1}{\lambda}$ with respect to the distribution function $\omega(\lambda)$ (area under blue curve) and $w^{(k)}(\lambda)$ (area under red curve). The positive part of the error of the Gauss-Christoffel quadrature is the area of the yellow region, and the negative part of the error is the area of the cyan region. The total error of the Gauss-Christoffel quadrature is the area of the yellow region minus the area of the cyan region. The blue circles on the x-axis indicate the eigenvalues $\omega(\lambda_i)$, the red circles on the x-axis indicate the points of increase of the $w^{(k)}(\theta_i^{(k)})$.

3. The right plot is the square of the squared A-norm of the k-th CG iterate, scaled by the initial residual.

The theorem states that the net error depicted in the middle plot, is equal to the value of the iterate on the right plot (depicted as the filled red dot).

Although the visualization illustrates the theorem, it is arguably not easy to get intuition about the convergence of CG from it. One fact that we can note for this specific problem is that the distribution function associated with the problem, has only around 15 "visible" jumps, which is much less than 144, the total number of points of increase of the function, so we can expect that we only need a small number of nodes to correctly approximate the distribution function, and therefore the integral. This is confirmed by seeing that the squared error of CG at the 19-th step is very small: around 10^{-20}.

5 Conclusion

In this report we stated a theorem which relates the error in CG to a Riemann-Stieljes integral, which can be visualized as the net area under a curve. The code for the visualization is available at https://github.com/mtg79/Meurant. It contains a few more test examples and can be run for matrix A and starting residual r_0. For a more thorough explanation of all the concepts presented in this report, see [4].
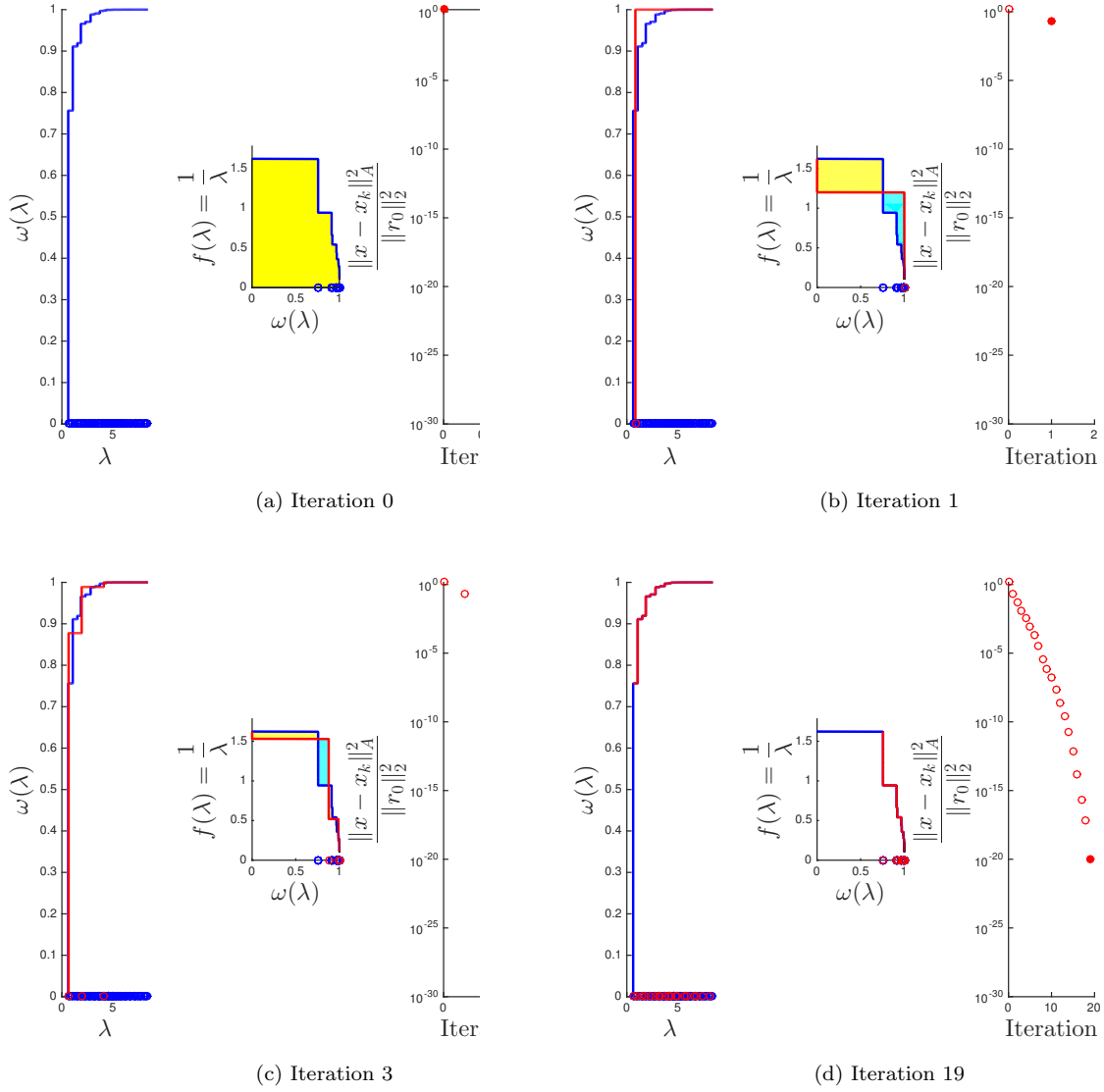
(a) Iteration 0

(b) Iteration 1

(c) Iteration 3

(d) Iteration 19

Figure 4: Visualization of Meurant's theorem.

References

[1] Bernhard Beckermann and Arno BJ Kuijlaars. "Superlinear CG convergence for special right-hand sides". In: *Electron. Trans. Numer. Anal* 14 (2002), pp. 1–19.

[2] Bernhard Beckermann and Arno BJ Kuijlaars. "Superlinear convergence of conjugate gradients". In: *SIAM Journal on Numerical Analysis* 39.1 (2001), pp. 300–329.

[3] Gene H Golub and Gérard Meurant. *Matrices, moments and quadrature with applications.* Princeton University Press, 2009.

[4] Jorg Liesen and Strakos Zdenek. *Krylov subspace methods: principles and analysis.* Oxford University Press, 2013.

[5] Gérard Meurant. "The computation of bounds for the norm of the error in the conjugate gradient algorithm". In: *Numerical Algorithms* 16.1 (1997), pp. 77–87.