CMO: Conjugate Descent

Eklavya Sharma

Objective: Minimize $f(x) = \frac{1}{2}x^TQx - b^Tx$, where Q is symmetric and positive definite.

Contents

1 Q-conjugate vectors **1**
2 Descent algorithm using Q-conjugate vectors **2**
3 Proof of convergence **2**
4 Rate of convergence **4**
5 Choosing Q-conjugate pairs **4**
6 Faster convergence for structured eigenvalues **5**
6.1 Q has r distinct eigenvalues . . . 7
6.2 Theorem for a polynomial . . . 7
6.3 Q has some clustered eigenvalues . . . 9

1 Q-conjugate vectors

Definition 1. *A set of d-dimensional non-0 vectors $U = \{u_0, u_1, \ldots, u_{k-1}\}$ is Q-conjugate iff $\forall i \neq j, u_i^TQu_j = 0$.*

Theorem 1. *If $U = \{u_0, \ldots, u_{d-1}\}$ is Q-conjugate, then U is a basis of $\mathbb{R}^d$.*

Proof. Assume U is linearly dependent. Then one of the vectors in U can be represented as a linear combination of the other (proof). Without loss of generality, assume $u_{d-1} = \sum_{i=0}^{d-2} \alpha_i u_i$.

$\forall i \neq d-1$,

$$0 = u_i^TQu_{d-1} = u_i^TQ\left(\sum_{j=0}^{d-2} \alpha_j u_j\right) = \sum_{j=0}^{d-2} \alpha_j u_i^TQu_j = \alpha_i u_i^TQu_i \implies \alpha_i = 0$$

Hence, $u_{d-1} = 0 \Rightarrow \perp$.

On assuming U to be linearly dependent, we got a contradiction. Therefore, U is linearly independent.

Since $|U| = d = \dim(\mathbb{R}^d)$, U is a basis of $\mathbb{R}^d$ (proof). □

Since Q is positive definite, $u_i^TQu_i > 0$ for all i.

2 Descent algorithm using Q-conjugate vectors

We'll develop a descent algorithm which uses u_k in the k^{th} iteration with exact line search. The name of this algorithm will be 'Conjugate Gradient Algorithm'.

Let $g(\alpha) = f(x_k + \alpha u_k)$ and $g_k = \nabla_f(x_k)^T$ (sorry for overloading variables; the subscript will help distinguish them though). Therefore, $g'(0) = \nabla_f(x_k) = g_k$ and $g''(0) = u_k^T Q u_k$.

By univariate Taylor series, we get

$$g(\alpha) = g(0) + \alpha g'(0) + \frac{\alpha^2}{2} g''(0)$$

Let $\alpha_k^* = \operatorname{argmin}_\alpha f(x_k + \alpha u_k)$. Therefore,

$$\alpha_k^* = -\frac{g'(0)}{g''(0)} = -\frac{g_k^T u_k}{u_k^T Q u_k}$$

We'll choose $x_{k+1} = x_k + \alpha_k^* u_k$. Therefore, $x_k = x_0 + \sum_{i=0}^{k-1} \alpha_i^* u_i$.

3 Proof of convergence

Theorem 2.

$$u_j^T g_k = \begin{cases} 0 & \textit{if } j < k \\ u_j^T g_0 & \textit{if } j \geq k \end{cases}$$

Proof.

$$\begin{aligned} g_k &= \nabla_f(x_k) = Q x_k - b \\ &= Q\left(x_0 + \sum_{i=0}^{k-1} \alpha_i^* u_i\right) - b \\ &= (Q x_0 - b) + \sum_{i=0}^{k-1} \alpha_i^* Q u_i \\ &= g_0 + \sum_{i=0}^{k-1} \alpha_i^* Q u_i \end{aligned}$$

$$
\begin{aligned}
u_j^T g_k &= u_j^T \left(g_0 + \sum_{i=0}^{k-1} \alpha_i^* Q u_i \right) \\
&= u_j^T g_0 + \sum_{i=0}^{k-1} \alpha_i^* u_j^T Q u_i \\
&= u_j^T g_0 + \sum_{i=0}^{k-1} \alpha_i^* \begin{Bmatrix} u_j^T Q u_j & i = j \\ 0 & i \neq j \end{Bmatrix} \\
&= u_j^T g_0 + \begin{Bmatrix} \alpha_j^* u_j^T Q u_j & j < k \\ 0 & j \geq k \end{Bmatrix} \\
&= u_j^T g_0 - \begin{Bmatrix} u_j^T g_j & j < k \\ 0 & j \geq k \end{Bmatrix}
\end{aligned}
$$

When $j = k$, we get $u_k^T g_k = u_k^T g_0$. Therefore,

$$
\begin{aligned}
u_j^T g_k &= u_j^T g_0 - \begin{Bmatrix} u_j^T g_j & j < k \\ 0 & j \geq k \end{Bmatrix} \\
&= u_j^T g_0 - \begin{Bmatrix} u_j^T g_0 & j < k \\ 0 & j \geq k \end{Bmatrix} \\
&= \begin{Bmatrix} 0 & j < k \\ u_j^T g_0 & j \geq k \end{Bmatrix}
\end{aligned}
$$

□

Corollary 2.1. $g_d = 0$. *This means that the conjugate descent algorithm converges in d iterations.*

Proof. By the previous theorem (2), $\forall 0 \leq j \leq d-1, u_j^T g_d = 0$. Since $U = \{u_0, u_1, \ldots, u_{d-1}\}$ forms a basis of $\mathbb{R}^d$, we get that $\forall x \in \mathbb{R}^d, x^T g_d = 0$. Therefore, $g_d^T g_d = 0 \implies g_d = 0$. □

We'll now look at an alternative way of proving convergence which will give us more insight.

Let $B_k = \{x_0 + \sum_{i=0}^{k-1} \beta_i u_i : \beta_i \in \mathbb{R}\}$. Since U is a basis of $\mathbb{R}^d$, $B_d = \mathbb{R}^d$. Therefore, to prove convergence of this algorithm, we'll prove the following theorem.

Theorem 3 (Expanding subspace theorem)**.** $\forall k, x_k = \operatorname{argmin}_{x \in B_k} f(x)$.

$x_k = x_0 + \sum_{i=0}^{k-1} \alpha_i^* u_i$. Let $\alpha^* = [\alpha_0^*, \ldots, \alpha_{k-1}^*]$. Let $h(\beta) = f(x_0 + \sum_{i=0}^{k-1} \beta_i u_i)$. Then $\min_{x \in B_k} f(x) = \min_{\beta \in \mathbb{R}^k} h(\beta)$. Since $h(\alpha^*) = f(x_k)$, if we prove that $\alpha^* = \operatorname{argmin}_{\beta \in \mathbb{R}^k} h(\beta)$, then $x_k = \operatorname{argmin}_{x \in B_k} f(x)$.

Lemma 4. *$h(\beta)$ is a convex function.*

Proof. Let $U = [u_0, u_1, \ldots, u_{k-1}]$ be a d by k matrix. Then

$$
(U\beta)_j = \sum_{i=0}^{k-1} U[j,i]\beta_i = \sum_{i=0}^{k-1} (u_i)_j \beta_i = \left(\sum_{i=0}^{k-1} u_i \beta_i \right)_j
$$

$$\implies h(\beta) = f\left(x_0 + \sum_{i=0}^{k-1} \beta_i u_i\right) = f(x_0 + U\beta)$$

$$\begin{aligned}
h(\beta) &= f(x_0 + U\beta) \\
&= f(x_0) + \nabla_f(x_0)^T(U\beta) + \frac{1}{2}(U\beta)^T Q(U\beta) && \text{(by Taylor series)} \\
&= f(x_0) + (\nabla_f(x_0)^T U)\beta + \frac{1}{2}\beta^T(U^T QU)\beta
\end{aligned}$$

This is a quadratic function in β. It is convex iff $U^T QU$ is positive definite.

By the rules for multiplying stacked matrices, we get that $(U^T QU)_{i,j} = u_i^T Q u_j$. Since vectors in U are Q-conjugate, $u_i^T Q u_j = 0$ when $i \neq j$. Therefore, $U^T QU$ is a diagonal matrix. Also, $\forall i, u_i^T Q u_i > 0$ because Q is positive definite. Therefore, all diagonal entries of $U^T QU$ are positive. Therefore, $U^T QU$ is positive definite. □

Since $h(\beta)$ is convex, $\nabla_h(\beta) = 0$ is a necessary and sufficient condition for minimum.

For all $j \in [0, k-1]$

$$h(\beta)_j = \frac{\partial f(x_0 + \sum_{i=0}^{k-1} \beta_i u_i)}{\partial \beta_j} = u_j^T \nabla_f \left(x_0 + \sum_{i=0}^{k-1} \beta_i u_i\right)$$

$$h(\alpha^*)_j = u_j^T \nabla_f \left(x_0 + \sum_{i=0}^{k-1} \alpha_i^* u_i\right) = u_j^T \nabla_f(x_k) = u_j^T g_k = 0 \qquad \text{(by theorem 2)}$$

Therefore, α^* minimizes h, so x_d minimizes f.

4 Rate of convergence

Unlike the previous algorithms, this algorithm:

- Converges exactly (instead of only 'approaching' the solution).
- Converges very fast – in exactly d steps.

5 Choosing Q-conjugate pairs

We will find U as follows: $u_0 = -g_0$ and $u_{k+1} = -g_{k+1} + \beta_k u_k$. We'll choose β_k such that $u_k^T Q u_{k+1} = 0$.

$$0 = u_k^T Q u_{k+1} = -u_k^T Q g_{k+1} + \beta_k u_k^T Q u_k \implies \beta_k = \frac{u_k^T Q g_{k+1}}{u_k^T Q u_k}$$

Algorithm 1 CGA(x_0): Conjugate Gradient Algorithm for $f(x) = \frac{1}{2}x^TQx - b^Tx$. Takes starting point as input.

$g_0 = Qx_0 - b$
if g_0 == 0 **then**
  return x_0
end if
$u_0 = -g_0$
for $i \in [0, \infty)$ **do**
  $\alpha_i = \dfrac{-g_i^T u_i}{u_i^T Q u_i}$
  $x_{i+1} = x_i + \alpha_i u_i$
  $g_{i+1} = Qx_{i+1} - b$
  if g_{i+1} == 0 **then**
    return x_{i+1}
  end if
  $\beta_i = \dfrac{u_i^T Q g_{i+1}}{u_i^T Q u_i}$
  $u_{i+1} = -g_{i+1} + \beta_i u_i$
end for

Theorem 5. *U is Q-conjugate.*

Proof. Proof can be found in the lecture notes for the course 'Optimization II - Numerical Methods for Nonlinear Continuous Optimization' by A. Nemirovski, in Theorem 5.4.1, page 95. □

Proof sketch. First induct on k to prove that for all k,

$$\text{span}(\{g_0, g_1, \dots, g_k\}) = \text{span}(\{g_0, Qg_0, \dots, Q^k g_0\}) = \text{span}(\{u_0, u_1, \dots, u_k\})$$

This can be done using the facts that $g_{k+1} - g_k = Q(x_{k+1} - x_k) = \alpha_k Q u_k$ and that $v_{k+1} = -g_{k+1} + \beta_k v_k$.

Then induct on k to prove that

$$\forall k, \forall i < k, u_k^T Q u_i = 0$$

To do this, express v_{k+1} as $-g_{k+1} + \beta_k v_k$, write Qv_i as a linear combination of $\{v_0, v_1, \dots, v_{i+1}\}$ and carefully invoke theorem 2. □

6 Faster convergence for structured eigenvalues

When the eigenvalues of Q have certain properties, we can guarantee faster convergence.

$B_{k+1} = x_0 + \text{span}(u_0, \dots, u_k)$. Therefore, any vector $x \in B_{k+1}$ can be expressed as $x_0 + \sum_{i=0}^k \gamma_i u_i$. Since $\text{span}(u_0, \dots, u_k) = \text{span}(g_0, \dots, Q^k g_0)$, $x = x_0 + \left(\sum_{i=0}^k \delta_i Q^i\right) g_0$.

Let Poly^k be the set of univariate polynomials of degree at most k where the coefficients are from $\mathbb{R}$ and the variable is an n by n matrix over $\mathbb{R}$. Therefore,

$$x \in B_{k+1} \implies \left(\exists P_k \in \text{Poly}^k, x = x_0 + P_k(Q) g_0\right)$$

$$\begin{aligned} x - x^* &= (x_0 - x^*) + P_k(Q)g_0 = (x_0 - x^*) + P_k(Q)Q(x_0 - x^*) \\ &= (I + QP_k(Q))(x_0 - x^*) \end{aligned}$$

Define $E(x) = f(x) - f(x^*)$. By Taylor series,

$$\begin{aligned} E(x) &= \frac{1}{2}(x - x^*)^T Q(x - x^*) \\ &= \frac{1}{2}(x_0 - x^*)^T (I + QP_k(Q))^T Q(I + QP_k(Q))(x_0 - x^*) \\ &= \frac{1}{2}(x_0 - x^*)^T Q(I + QP_k(Q))^2(x_0 - x^*) \end{aligned}$$

Let $R = \{e_1, e_2, \ldots, e_d\}$ be the set of orthonormal eigenvectors of Q. Let $\lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_d$ be the corresponding eigenvalues. Since R forms a basis of $\mathbb{R}^d$, $x_0 - x^*$ can be represented as a linear combination of R. Let $x_0 - x^* = \sum_{i=1}^d \zeta_i e_i = \zeta_i$.

Lemma 6. $E(x_0) = \frac{1}{2}\sum_{i=1}^d \zeta_i^2 \lambda_i$

Proof. Let R be a matrix whose i^{th} column is e_i. Since the eigenvectors are orthonormal, $RR^T = R^T R = I$. Let $\zeta = [\zeta_1, \ldots, \zeta_d]^T$. Then

$$R\zeta = \sum_{i=1}^d \zeta_i e_i = x_0 - x^*$$

Since Q is symmetric, $Q = RDR^T$, Where D is a diagonal matrix whose i^{th} entry is λ_i. Therefore,

$$\begin{aligned} 2E(x_0) &= (x_0 - x^*)^T Q(x_0 - x^*) = (R\zeta)^T (RDR^T)(R\zeta) \\ &= \zeta^T (R^T R) D (R^T R)\zeta = \zeta^T D\zeta = \sum_{i=1}^d \zeta_i^2 \lambda_i \end{aligned}$$

□

Lemma 7 (Homework)**.** *Let T be a polynomial where $T(X) = X(I + XP_k(X))^2$. Then $E(x) = \frac{1}{2}\sum_{i=1}^d \zeta_i^2 T(\lambda_i)$.*

Hint. Use the fact that for all $j \in \mathbb{N}$, R is also the set of eigenvectors of Q^j and the corresponding eigenvalues are $\lambda_1^j, \ldots, \lambda_d^j$. □

Lemma 8. *For any polynomial $P_k \in Poly^k$,*

$$\frac{E(x_{k+1})}{E(x_0)} \leq \max_{i=0}^{d}(1 + \lambda_i P_k(\lambda_i))^2$$

Proof.

$$\begin{aligned}
E(x_{k+1}) &= \min_{x \in B_{k+1}} E(x) && \text{(Expanding subspace theorem)} \\
&= \min_{P_k \in \text{Poly}^k} \frac{1}{2} \sum_{i=1}^{d} \zeta_i^2 \lambda_i (1 + \lambda_i P_k(\lambda_i))^2 \\
&\leq \min_{P_k \in \text{Poly}^k} \frac{1}{2} \sum_{i=1}^{d} \left(\zeta_i^2 \lambda_i \left(\max_{i=0}^{d} (1 + \lambda_i P_k(\lambda_i))^2 \right) \right) \\
&= \min_{P_k \in \text{Poly}^k} \left(\frac{1}{2} \sum_{i=1}^{d} \zeta_i^2 \lambda_i \right) \left(\max_{i=0}^{d} (1 + \lambda_i P_k(\lambda_i))^2 \right) \\
&= E(x_0) \min_{P_k \in \text{Poly}^k} \max_{i=0}^{d} (1 + \lambda_i P_k(\lambda_i))^2
\end{aligned}$$

□

Therefore, by cleverly choosing a polynomial, we can prove useful bounds on convergence.

6.1 Q has r distinct eigenvalues

Suppose Q has r distinct eigenvalues $\mu_1 > \mu_2 > \ldots > \mu_r$. Let $\overline{P}_r(x) = 1 + xP_{r-1}(x)$.

We'll construct P_{r-1} such that $\overline{P}_r(x) = 0$ for all $1 \leq i \leq r$. This would mean that $\frac{E(x_r)}{E(x_0)} = 0$, so the conjugate gradient algorithm will converge in r iterations.

Define $\overline{P}_r$ and P_{r-1} as follows:

$$\overline{P}_r(x) = \prod_{j=1}^{r} \left(1 - \frac{x}{\mu_j} \right) \qquad\qquad P_{r-1}(x) = \frac{\overline{P}_r(x) - 1}{x}$$

Lemma 9. *P_{r-1} is a polynomial of degree $r - 1$ such that $\forall 0 \leq i \leq r, \overline{P}_r(\mu_i) = 0$.*

Proof. Clearly, $\overline{P}_r(\mu_i) = 0$ for all i. Also, the degree of $\overline{P}$ is r.

Next, we must prove that P_{r-1} is a polynomial. Note that $\overline{P}_r(0) = 1$, so 0 is a root of $\overline{P}_r(x) - 1$. Therefore, x is a factor of $\overline{P}_r(x) - 1$ and hence P_{r-1} is a polynomial.

Since the degree of $\overline{P}_r$ is r, the degree of P_{r-1} is $r - 1$. □

6.2 Theorem for a polynomial

In this section, we'll prove a theorem for a certain polynomial which we'll use in the next section.

Theorem 10. *Let $n \geq 2$. Let $0 < a_1 < a_2 < \ldots < a_n$. Let $p_1, p_2, \ldots, p_n$ be positive integers and let $p_1 = 1$.*

$$f(x) = \prod_{i=1}^{n} \left(1 - \frac{x}{a_i} \right)^{p_i} \qquad\qquad g(x) = f(x) - 1 + \frac{x}{a_1}$$

Then

1. *f is positive in $(-\infty, a_1)$, negative in (a_1, a_2) and 0 at a_1 and a_2.*

2. *$g(x) \leq 0$ for $x \in [0, a_1]$ and $g(x) \geq 0$ for $x \in [a_1, a_2]$.*

Proof. Since a_1 and a_2 are zeros of f, $f(a_1) = f(a_2) = 0$. Since a_1 is the leftmost zero of f, f has the same sign in $(-\infty, a_1)$ (by intermediate value theorem). Since $f(0) = 1$, f is positive in $(-\infty, a_1)$.

$$\frac{f'(x)}{f(x)} = \sum_{i=1}^{n} \frac{p_i}{x - a_i}$$

Let

$$h_1(x) = \prod_{i=1}^{n} (x - a_i)^{p_i - 1}$$

Then $h_1(x)$ divides $f'(x)$.

By Rolle's theorem, there must be points $b_1 < b_2 < \ldots < b_{n-1}$ such that for all i, $f'(b_i) = 0$ and $b_i \in (a_i, a_{i+1})$. Let

$$h_2(x) = \prod_{i=1}^{n-1} (x - b_i)$$

So $h_2(x)$ divides $f'(x)$.

Let $N = \sum_{i=1}^{n} p_i$. Then $\deg(f) = N$. Also

$$\deg(h_1 h_2) = \deg(h_1) + \deg(h_2) = (N - n) + (n - 1) = N - 1 = \deg(f')$$

Therefore, $f'(x) = \gamma h_1(x) h_2(x)$ for some $\gamma \in \mathbb{R}$.

Since $p_1 = 1$, b_1 is the leftmost zero of f' and it is the only zero in $(-\infty, a_2)$. Therefore, $f'(x)$ has the same sign for $x \in (-\infty, b_1)$. Since $f(0) = 1$, $f'(0) = -\sum_{i=1}^{n} \frac{1}{a_i} < 0$. Therefore, $f'(x) < 0$ for $x \in (-\infty, b_1)$.

Since $f(a_1) = 0$ and $f'(a_1) < 0$, $f(a_1 + \epsilon) < 0$ for all very small ϵ. Also, f has the same sign in (a_1, a_2), otherwise it would have a root in (a_1, a_2), which we know is false. Therefore, $f(x) < 0$ for $x \in (a_1, a_2)$. This completes the proof of part 1 of this theorem.

Applying Rolle's theorem to $f'(x)$ and by a similar argument (todo: expand this), we get that $f''(x)$ must have its leftmost root in (b_1, a_2). Therefore, $f''(x)$ has the same sign in $(-\infty, b_1]$.

$$\frac{f''(x)}{f(x)} = \left(\sum_{i=1}^{n} \frac{p_i}{a_i - x} \right)^2 - \sum_{i=1}^{n} \frac{p_i}{(a_i - x)^2}$$

$$\implies f''(0) = \left(\sum_{i=1}^{n} \frac{p_i}{a_i} \right)^2 - \sum_{i=1}^{n} \frac{p_i}{a_i^2} > 0$$

Therefore, $f''(x) > 0$ for $x \in (-\infty, b_1]$.

$f'(b_1) = 0$ and $f''(b_1) > 0$. Therefore, $f'(b_1 + \epsilon) > 0$ for all very small ϵ. $f'(x)$ has the same sign in (b_1, a_2) because b_1 is the only root of $f'(x)$ in $[b_1, a_2)$. Therefore, $f'(x) > 0$ for $x \in (b_1, a_2)$.

Since f is convex in $(-\infty, b_1]$, for $\alpha \in [0, 1]$,

$$f(\alpha a_1) = f((1-\alpha)0 + \alpha a_1) \leq (1-\alpha)f(0) + \alpha f(a_1) = (1-\alpha)$$

Setting α to x/a_1, we get that for $x \in [0, a_1]$, $f(x) \leq 1 - \frac{x}{a_1} \Rightarrow g(x) \leq 0$.

$g(0) = g(a_1) = 0$. By Rolle's theorem, $\exists x_0 \in (0, a_1), g'(x_0) = 0$. Since $g''(x) = f''(x) > 0$ for $x \in (-\infty, b_1]$, $g'(x) > 0$ for $x \in (x_0, b_1]$.

$g'(x) = f'(x) + \frac{1}{a_1}$. For $x \in (b_1, a_2)$, $f'(x) > 0 \Rightarrow g'(x) > 0$. Therefore, $g'(x) > 0$ for $x \in [a_1, b_1)$.

Since $g(a_1) = 0$ and $g'(x) > 0$ for $x \in [a_1, b_1)$, $g(x) > 0$ for $x \in (a_1, b_1)$. □

6.3 Q has some clustered eigenvalues

Suppose Q has eigenvalues $\lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_d$, where for some constants a and b,

$$0 < a \leq \lambda_d \leq \ldots \leq \lambda_{r+1} < b < \lambda_r \leq \ldots \leq \lambda_1$$

Let $\mu_i = \lambda_i$ for i from 1 to r. Let $\mu_{r+1} = \frac{a+b}{2}$.

$$\overline{P}_{r+1}(x) = \prod_{j=1}^{r+1} \left(1 - \frac{x}{\mu_j}\right) \qquad P_r(x) = \frac{P_{r+1}(x) - 1}{x} \qquad h(x) = 1 - \frac{x}{\mu_{r+1}}$$

It's easy to prove (similar to lemma 9) that P_r is a polynomial and has degree r.

Since $\overline{P}_{r+1}$ is of the right form, we can apply theorem 10.

By part 1 of theorem 10, we get that for $x \in [a, \frac{a+b}{2}]$, $\overline{P}_{r+1}(x) \geq 0$. By part 2 of theorem 10, we get that for $x \in [a, \frac{a+b}{2}]$,

$$\overline{P}_{r+1}(x) \leq h(x) \leq h(a) = \frac{b-a}{b+a}$$

By part 1 of theorem 10, we get that for $x \in [\frac{a+b}{2}, b]$, $\overline{P}_{r+1}(x) \leq 0$. By part 2 of theorem 10, we get that for $x \in [\frac{a+b}{2}, b]$,

$$\overline{P}_{r+1}(x) \geq h(x) \geq h(b) = -\frac{b-a}{b+a}$$

Therefore, for $x \in [a, b]$, $\left|\overline{P}_{r+1}(x)\right| \leq \frac{b-a}{b+a}$. Therefore,

$$\frac{E(x_{r+1})}{E(x_0)} \leq \left(\frac{b-a}{b+a}\right)^2$$

We can use the above fact to design an algorithm called the 'partial conjugate gradient' algorithm. In this algorithm, we'll start at the point z_0 and run the conjugate gradient algorithm for $r+1$ steps to reach the point z_1. Then we'll rerun the conjugate gradient algorithm for $r+1$ steps from z_1 to reach a point z_2, then we'll rerun the conjugate gradient algorithm for $r+1$ steps from z_2 to reach a point z_3, and so on. We'll do this l times. After l iterations $\frac{E(z_l)}{E(z_0)} = \left(\frac{b-a}{b+a}\right)^{2l}$. This will give us linear convergence.