1 Foundations

This sections focuses on important mathematical entities that are important for formulating, analyzing the Conjugate Gradient and the Lanczos Algorithm. Major parts of this sections cited from...

1.1 Projectors

There are 2 types of projector, an oblique Projector and Orthgonal Projector. A matrix P is called a projector if:

Definition 1.

$$P^2 = P \tag{1.1.1}$$

This property is sometimes referred as idempotent. As a consequence, $\operatorname{ran}(I-P) = \operatorname{null}(P)$ and here is the proof:

Proof.

$$\forall x \in \mathbb{C}^n : P(I-P)x = \mathbf{0} \implies \operatorname{ran}(I-P) \subseteq \operatorname{null}(P) \tag{1.1.2}$$

$$\forall x \in \operatorname{null}(P) : Px = \mathbf{0} \implies (I-P)x = x \implies x \in \operatorname{ran}(I-P) \tag{1.1.3}$$

$$\implies \operatorname{ran}(I-P) = \operatorname{null}(P) \tag{1.1.4}$$

□

This consequence states the fact that any vector x can be represented in the form of: $x = Px + (I-P)x$, and every projector will be defined via the range of $I-P$ and P.

1.1.1 Orthogonal Projector

An orthogonal projector is a projector such that:

Definition 2.

$$\operatorname{null}(P) \perp \operatorname{ran}(P) \tag{1.1.5}$$

This property is in fact, very special. A good example of an orthogonal projector would be the Householder Reflector Matrix. Or just any $\hat{u}\hat{u}^H$ where $\hat{u}$ is being an unitary vector. For convenience of proving, assume subspace $M = \operatorname{ran}(P)$. Consider the following lemma:

Lemma 1.

$$\operatorname{null}(P^H) = \operatorname{ran}(P)^\perp \tag{1.1.6}$$

$$\operatorname{null}(P) = \operatorname{ran}(P^H)^\perp \tag{1.1.7}$$

Using (1.1.4) and consider the proof:

Proof.

$$\langle P^H x, y\rangle = \langle x, Py\rangle \tag{1.1.8}$$
$$\forall x \in \text{null}(P^H), y \in \mathbb{C}^n \tag{1.1.9}$$
$$\implies \langle P^H x, y\rangle = 0 = \langle x, Py\rangle \tag{1.1.10}$$
$$\implies \text{null}(P^H) \perp \text{ran}(P) \tag{1.1.11}$$
$$\forall y \in \text{null}(P), x \in \mathbb{C}^n : \tag{1.1.12}$$
$$\langle x, Py\rangle = 0 = \langle P^H x, y\rangle \tag{1.1.13}$$
$$\implies \text{ran}(P^H) \perp \text{null}(P) \tag{1.1.14}$$

□

Proposition 1. A projector is orthogonal iff it's Hermitian.

Proof. $\impliedby$ Assuming the matrix is Hermitian and it's a projector, then we wish to prove that it's an orthogonal projector. Let's recall:

$$\text{null}(P^H) = \text{ran}(P)^\perp \tag{1.1.15}$$
$$\text{null}(P) = \text{ran}(P^H)^\perp \tag{1.1.16}$$

Substituting $P^H = P$, we have $\text{null}(P) = \text{ran}(P)^\perp$, Which is the definition of Orthogonal Projector. Therefore, P is an orthogonal projector by the definition of the projector.

For the $\implies$ direction, we assume that P is an Orthogonal Projector, then we wish to show that it's also Hermitian. Observe that P^H is also a projector because $(P^H)^2 = (P^2)^H$. Then, using the definition of orthogonal projector:

$$\text{null}(P) \perp \text{ran}(P) \tag{1.1.17}$$
$$\text{null}(P^H) \perp \text{ran}(P^H) \tag{1.1.18}$$

Notice that using above statement together with Lemma 1 means $\text{null}(P) = \text{ran}(P)^\perp = \text{null}(P^H)$, and then $\text{ran}(P) = \text{null}(P)^\perp = \text{ran}(P^H)$. Therefore, P^H is an projector such that: $\text{ran}(P) = \text{ran}(P^H) \wedge \text{null}(P) = \text{null}(P^H)$. The range and null space of P^H and P is the same therefore P has to be Hermitian. □

1.1.2 Oblique Projector

An oblique projector is not orthogonal, and vice versa. It's a projector that satisfies the following conditions:

Definition 3.

$$Px \in M \quad (I - P)x \perp L \quad \text{where: } M \neq L \tag{1.1.19}$$

An orthgonal projector is the case when the subspace $M = L$.

A famous example of an orthogonal projector is QQ^H where Q is an Unitary Matrix. This is a Hermitian Matrix and it's idempotent, making it an orthogonal projector.

1.1.3 Projector Geometric Intuitions

A projector describes a given vector using some elements from another basis. The oblique projector creates a light sources in the form of the subspace L and it shoots parrell light ray orthogonal to L, crossing vectors and projecting their shadow onto subspace M.

1.2 Projectors and Norm Minimizations

An orthogonal projector always reduce the 2 norm of a vector. Given any subspace M, we can create a basis of vectors packing into the some matrix, say A, then P_M as a projector onto the basis M one example can be: $A(AA^T)^{-1}A^T$. Let's consider the claim:

$$\|P_M x\|^2 \le \|x\|^2 \tag{1.2.1}$$

Proof:

$$x = Px + (I-P)x \tag{1.2.2}$$
$$\|x\|^2 = \|Px\|^2 + \|(I-P)x\|^2 \tag{1.2.3}$$
$$\|x\|^2 \ge \|Px\|^2 \tag{1.2.4}$$

Using this property of the Orthogonal Projector, we consider the following minimizations problem:

$$\min_{x\in M} \|y-x\|_2^2 = \|y - P_M(y)\|_2^2 \tag{1.2.5}$$

Proof:

$$\|y-x\|_2^2 = \|y - P_My + P_My - x\|_2^2 \tag{1.2.6}$$
$$\|y-x\|_2^2 = \|y - P_My\|_2^2 + \|P_My - x\|_2^2 \tag{1.2.7}$$
$$\implies \|y - P_My\|_2^2 \le \|y-x\|_2^2 \tag{1.2.8}$$

That concludes the proof. Observe that, $y - P_My \perp M$ and $P_My - x \in M$ because $P_My, x \in M$, which allows us to split the norm of $y - x$ into 2 components. In addition using the fact that the projector is orthogonal. That concludes the proof.

1.3 Subspace Orthogonality Framework

Let $\mathcal{K}, \mathcal{L}$ be subspaces where candidates soltuions are chosen and residuals are orthogonalized against. Under the idea case the 2 subspaces spans all dimensions, and it's able to approximate all solutions and forcing the residual vector $(b - Ax)$ to be zero. This is a description of this framework:

$$\tilde{x} \in x_0 + \mathcal{K} \text{ s.t: } b - A\tilde{x} \perp \mathcal{L} \tag{1.3.1}$$

it looks for an x in the affine linear subspace $\mathcal{K}$ such that it's perpendicular to the subspace $\mathcal{L}$, or, equivalently, minimizing the projection onto the subspace $\mathcal{L}$. One interpretation of it is an projection of residual onto the basis that is orthogonal to $\mathcal{L}$.

Sometimes, for convenience and the exposition and exposing hidden connections between ideas, the above conditions can be expressed using matrix.

$$\text{Let } V \in \mathbb{C}^{n\times m} \text{ be a aasis for: } \mathcal{K} \tag{1.3.2}$$
$$\text{Let } W \in \mathbb{C}^{n\times m} \text{ be a basis for: } \mathcal{L} \tag{1.3.3}$$

We can then make us of (1.3.1) and express it in the form of:

$$\tilde{x} = x^{(0)} + Vy \tag{1.3.4}$$
$$b - A\tilde{x} \perp (\text{span} \leftarrow \text{col})(W) \tag{1.3.5}$$
$$W^T(b - A\tilde{x} - AVy) = \mathbf{0} \tag{1.3.6}$$
$$W^T r^{(0)} - W^T AVy = \mathbf{0} \tag{1.3.7}$$
$$W^T AVy = W^T r^{(0)} \tag{1.3.8}$$

1.3.1 Prototype Algorithm

And from here, we can define a simple prototype algorithm using this framworks.

$$\begin{aligned}
&\text{While not converging :}\\
&\quad \text{Increase Span for: } \mathcal{K}, \mathcal{L}\\
&\quad \text{Choose: } V, W \text{ for } \mathcal{K}, \mathcal{L}\\
&\quad r := b - Ax\\
&\quad y := (W^T AV)^{-1} W^T r\\
&\quad x := x + Vy
\end{aligned} \tag{6}$$

Each time, we increase the span of the subspace $\mathcal{K}, \mathcal{L}$, which gives us more space to choose the solution x, and more space to reduce the residual vector r. This idea is incredibly flexible, and we will see in later part where it reduces to a more concrete algorithm. Finally, when $\mathcal{K} = \mathcal{L}$, this is referred to as a Petrov Galerkin's Conditions.

1.4 Subspace Minimization Framework

Other times, iterative method will choose to build up a subspace for each step with a subspace generator, and build up the solution on this expanding subspace, but with the additional objective of minimizing the residual under certain norm. Assuming that the vector $x \in x_0 + \mathcal{K}$, and we want to minimize the residual under a norm induced by positive definite operator B. Let it be the case that the columns of matrix K span subspace $\mathcal{K}$ with $\dim(\mathcal{K}) = k$.

$$\min_{x\in x_0+\mathcal{K}} \|b - Ax\|_B^2 \tag{1.4.1}$$
$$= \min_{w\in\mathbb{R}^k} \|b - A(x_0 + Kw)\|_B^2 \tag{1.4.2}$$
$$= \min_{w\in\mathbb{R}^k} \|r_0 - AKw\|_B^2 \tag{1.4.3}$$

We take the derivative of it and set the derivative to zero, this translate the problem to a projection problem under the A norm.

$$\nabla_w \left[\|r_0 - AKx\|_B^2\right] = \mathbf{0} \tag{1.4.4}$$

$$(AK)^T B(r_0 - AKx) = \mathbf{0} \tag{1.4.5}$$

$$(AK)^T Br_0 - (AK)^T BAKx = \mathbf{0} \tag{1.4.6}$$

$$(AK)^T Br_0 = (AK)^T BAKx \tag{1.4.7}$$

The above formulation is tremendously powerful. I used gradient instead of projector for the simplicity of the argument. One can derive the same using projector but the math is bit more hedious.

1.5 Krylov Subspace

A Krylov Subspace is a sequence basis paramaterized by A, an linear operator, v an initial vector, and k, which basis in the sequece of basis that we are looking at.

Definition 4 (Krylov subspace)**.**

$$\mathcal{K}_k(A|b) = \operatorname{span}(b, Ab, A^2b, \cdots A^{k-1}b) \tag{1.5.1}$$

Please immediately observe that from the definition we have:

$$\forall v : \mathcal{K}_1(A|v) \subseteq \mathcal{K}_2(A|v) \subseteq \mathcal{K}_3(A|v) \cdots \tag{1.5.2}$$

Please also observe that, every element inside of krylov subspace generated by matrix A, and an initial veoctr v can be represented as a polynomial of matrix A multiplied by the vector v and vice versa.

$$\forall x \in \mathcal{K}_j(A|v)\ \exists\ w : p_k(A|w)v = x \tag{1.5.3}$$

We use $p_k(A|w)$ to denotes a matrix polynomial with coefficients w, where w is a vector. No proof this is trivial. Take note that, we can change the field of where the scalar w is coming from, but for discussion below, $\mathbb{R}, \mathbb{C}$ doesn't matter and won't change the results.

$$p_k(A|w)v = \sum_{j=0}^{k-1} w_j A^j v \tag{1.5.4}$$

1.5.1 The Grade of a Krylov Subspace

The most important porperty of the subspace is the idea of grade denoted as: $\operatorname{grade}(A|v)$, indicating when the Krylov Subspace of A wrt to v stops expanding after a certain size. To show this idea, we consider the following 3 statements about Krylov Subspace which we will proceed to prove.

Proposition 2.

$$\exists 1 \leq k \leq m+1 : \mathcal{K}_k(A|v) = \mathcal{K}_{k+1}(A|v) \tag{1.5.5}$$

There exists an natural number between 1 and $m+1$ such that, the successive krylov subspace span the same space asthe previous one.

Proposition 3.

$$\exists \min k \text{ s.t: } \mathcal{K}_k(A|v) = \mathcal{K}_{k+1}(A|v) \implies \mathcal{K}_k(A|v) \text{ is Lin Ind} \wedge \mathcal{K}_{k+1}(A|v) \text{ is Lin Dep.} \tag{1.5.6}$$

There eixsts a minimum such k where the immediate next krylov subspace is linear dependent.

Proposition 4.

$$\mathcal{K}_k(A|v) \text{ Lin Dep} \implies \mathcal{K}_{k+1}(A|v) = \mathcal{K}_k(A|v) \tag{1.5.7}$$

if the k krylov subspace is linear dependent, then it stops expanding and the successive krylov subspace spans the same space.

Theorem 1 (Existence of Grade for a Krylov Subspace)**.** Let k be the minumum number when the krylov subspace stops expanding, then all successive krylov subspace spand the same space. $\mathcal{K}_k(A|v) = \mathcal{K}_{k+j}(A|v)\ \forall j \geq 0$. The number k is regarded as the grade of krylov subspace wrt to v denoted using grade$(A|v)$.

Proposition 2, 3 ensures that there exists a term in the sequence of krylov subspace stops expanding, and when that happens all subsequent Krylov Subspace will span the same subspace, this is by Proposition 4.

Next, let's consider the proof of theorem 1 by proving all proposition 2, 3, and 4.

Proposition 2. For notational simplicity, $\mathcal{K}_k$ now denotes $\mathcal{K}_k(A|v)$. Let's start the considerations from the definition of the Krylov Subspace:

$$\forall\ k : \mathcal{K}_k \subseteq \mathcal{K}_{k+1} \implies \dim(\mathcal{K}_k) \leq \dim(\mathcal{K}_{k+1}) \tag{1.5.8}$$

$$\mathcal{K}_{k+1} \setminus \mathcal{K}_k = \operatorname{span}(A^k v) \tag{1.5.9}$$

$$\implies \dim(\mathcal{K}_{k+1}) - \dim(\mathcal{K}_k) \leq 1 \tag{1.5.10}$$

Therefore, the dimension of the successive krylov subspace forms a sequence of positive integer that is monotonically increasing. By the Cayley's Hamilton's theorem (will be stated later), the sequence is bounded by m, since there are $m + 1$ terms, it must be the case that at least 2 of the krylov subspace has the same dimension (And the earliest such occurance will exist), implying the the fact that the new added vector from k to $k + 1$ is in the span of the previous subspace. □

Proposition 4.

$$\mathcal{K}_k \text{Lin Dep} \tag{1.5.11}$$

$$\implies \exists w_k \neq \mathbf{0} : p_k(A|w_k^+)v = \mathbf{0} \tag{1.5.12}$$

$$\implies Ap_k(A|w_k^+)v = \mathbf{0} \tag{1.5.13}$$

$$p_{k+1}(A|[0\ (w_k^+)^T]) = \mathbf{0} \tag{1.5.14}$$

$$\mathcal{K}_{k+1} \text{ is Lin Dep} \tag{1.5.15}$$

The recurrence of multplying by A allows the krylov subspace to grow and the new bigger subspace will contain the previous one. Therefore inheriting the linear dependence, we use the idea of matrix polynomial for the proof. □

Proposition 3. Assuming that prop 2, 4 are true. prop 3 implies the existence of the smallest such k. For contradiction, we only have one case to assume, that is $\mathcal{K}_k$ and $\mathcal{K}_{k+1}$ are linear dependence. Then $\mathcal{K}_k$ is either Linear Dependence, or Independence.

If $\mathcal{K}_{k-1}$ is linear dependence, then by (3) $\mathcal{K}_{k-1} = \mathcal{K}_k$, hence k is not the minimum. Else assume $\mathcal{K}_{k-1}$ is linear independence, however $\mathcal{K}_k$ is linear dependence, and $\mathcal{K}_k \setminus \mathcal{K}_{k-1} = \text{span}(A^{k-1}v)$; therefore, $A^{k-1}v$ is in the span of $\mathcal{K}_{k-1}$, hence $\mathcal{K}_{k-1} = \mathcal{K}_k$, contradicing again that k is the minimum such k. □

1.6 Minimal Polynomial of a Matrix

Definition 5. A minimal polynomial is a polynomial $p_k(x)$ with degree k that is monic such that $p_k(A) = \mathbf{0}$.

1.7 Useful Theorems

1.7.1 Cauchy Interlace Theorem

1.7.2 Caley Hamilton's Theorem

Definition 6. A matrix satisfies it's own characteristic equation, let $p(x)$ be the characteristic polynomial for the matrix A, then $p(A) = \mathbf{0}$.

1.8 Deriving Conjugate Gradient from First Principles

1.8.1 CG Objective and Framework

We introduce the algorithm as an attempt to minimize the energy norm of the error for a linear equation $Ax = b$, here we make the assumptions:

1) The matrix A is symmetric semi-positive definite.

2) Further assume another matrix $P_k = [p_0 \ p_1 \ \cdots p_{k-1}]$ as a matrix whose columns is a basis.

$$\min_{w \in \mathbb{R}^k} \|A^{-1}b - (x_0 + P_k w)\|_A^2 \iff P_k^T r_0 = P_k^T A P_k w \tag{1.8.1}$$

Refer back to (1.4) for how to deal with the above minimization objective. Using the matrix form for the Petrov Galerkin Conditions where W, V are both P_k, we have this orthgonality formulations:

$$\text{choose: } x \in x_0 + \text{ran}(P_k) \text{ s.t: } b - Ax \perp \text{ran}(P_k) \tag{1.8.2}$$

Take note that this link between a norm minimzation and an equivalent Orthogonality condition doesn't garantee to happen, for example the FOM and Bi-Lanczos Method are orthogonalizations method that doesn't directly link to a norm minimization objective.

To solve for w, we wish to make $P_k^T A P_k$ to be an easy to solve matrix. Let the easy to solve matrix to be a diagonal matrx and hence we let P_k to be a *matrix whose columns are A-Orthogonal vectors.*

$$P_k^T A P_k = D_k \text{ where: } (D_k)_{i,i} = \langle p_{i-1}, Ap_{i-1}\rangle \tag{1.8.3}$$

$$P_k r_0 = P_k^T A P_k w = D_k w \tag{1.8.4}$$

$$w = D_k^{-1} P_k^T r_0 \tag{1.8.5}$$

The idea here is: Accumulating vectors p_j into the matrix P_k and then iterative improve the solution x_k, by reducing the error denote as e_k which is defined as $A^{-1}b - x_k$. Then, we can derive the following expression for the solution at step k x_k and the residual at step $r_k = b - Ax_k$ for the algorithm:

$$\begin{cases} x_k = x_0 + P_k D_k^{-1} P_k^T r_0 \\ r_k = r_0 - AP_k D_k^{-1} P_k^T r_0 \\ P_k^T A P_k = D_k \end{cases} \tag{1.8.6}$$

1.8.2 Using the Projector

Here, we consider the simple algorithm from (1.8.6). Please observe that $AP_k D_k^{-1} P_k$ is a projector, and so is $P_k D_k^{-1} P_k^T A$.

Proof.

$$AP_k D_k^{-1} P_k^T (AP_k D_k^{-1} P_k^T) = AP_k D_k^{-1} P_k^T A P_k D_k^{-1} P_k^T \tag{1.8.7}$$

$$= AP_k D_k^{-1} D_k D_k^{-1} P_k^T \tag{1.8.8}$$

$$= AP_k D_k^{-1} P_k^T \tag{1.8.9}$$

$$P_k D_k^{-1} P_k^T A (P_k D_k^{-1} P_k^T A) = P_k D_k^{-1} D_k D_k^{-1} P_k^T A \tag{1.8.10}$$

$$= P_k D_k^{-1} P_k^T A \tag{1.8.11}$$

□

Both matrices are indeed projectors. Please take note that they are not Hermitian, which would mean that they are not orthogonal projector, hence, oblique projectors. For notational convenience, we denote $\overline{P}_k = P_k D_k^{-1} P_k^T$; then these 2 projectros are:

$$AP_k D_k^{-1} P_k^T = A\overline{P}_k \tag{1.8.12}$$

$$P_k D_k^{-1} P_k^T A = \overline{P}_k A \tag{1.8.13}$$

One immediate consequence is:

$$\text{ran}(I - A\overline{P}_k) \perp \text{ran}(P_k) \tag{1.8.14}$$

$$\text{ran}(I - \overline{P}_k A) \perp \text{ran}(AP_k) \tag{1.8.15}$$

Proof.

$$
\begin{aligned}
P_k^T(I - A\overline{P}_k) &= P_k^T - P_k^T A\overline{P}_k && (1.8.16)\\
&= P_k^T - D_k D_k^{-1} P_k^T && (1.8.17)\\
&= \mathbf{0} && (1.8.18)\\
(AP_k)^T(I - \overline{P}_k A) &= P_k^T A - P_k^T A\overline{P}_k A && (1.8.19)\\
&= P_k^T A - P_k^T A P_k D_k^{-1} P_k^T A && (1.8.20)\\
&= P_k^T A - P_k^T A && (1.8.21)\\
&= \mathbf{0} && (1.8.22)
\end{aligned}
$$

□

Using the properties of the oblique projector, we can proof 2 facts about this simple norm minimization method we developed:

Proposition 5 (Residuals are Orthogonal to P_k)**.**

$$
\begin{aligned}
r_k &= r_0 - A\overline{P}_k r_0 = (I - A\overline{P}_k)r_0 && (1.8.23)\\
\implies r_k &\perp \operatorname{ran}(P_k) && (1.8.24)
\end{aligned}
$$

Proposition 6 (Generating A Orthogonal Vectors)**.** Given any set of basis vector, for example $\{u_k\}_{i=0}^{n-1}$, one can generate a set of A-Orthogonal vectors from it. More specifically:

$$
\begin{aligned}
p_k &= (I - \overline{P}_k A)u_k && (1.8.25)\\
\operatorname{span}(p_k) &\perp \operatorname{ran}(AP_k) && (1.8.26)
\end{aligned}
$$

1.8.3 Assisted Conjugate Gradient

So far, we have this particular scheme of solving the optimization problem, coupled with the way to computing the solution x_k at each step, and the residual at each step, while also getting the residual vector at each step too. However, it would be great if we can accumulate on the same subspace P_k and look for a chance to reuse the computational results from the previous iterations of the algorithm:

$$
\begin{cases}
x_k = x_0 + \overline{P}_k r_0\\
r_k = (I - A\overline{P}_k)r_0\\
P_k^T A P_k = D_k\\
\overline{P}_k = P_k D_k^{-1} P_k^T\\
p_k = (I - \overline{P}_k A)u_k \quad \{u_i\}_{i=0}^{n-1} \text{ is a Basis}
\end{cases} \tag{1.8.27}
$$

With the assitance of a set of basis vector that span the whole space, this algorithm is possible to achieve the objective. Take note that we can accumulate the solution for x_k accumulatively, instead of computing the whole projector process, we have the choice to update it recursively as the newest p_k vector is introduced at that step. Let's Call this formulation of the algorithm: *Assisted Conjugate Gradient.*

1.8.4 Properties of Assisted Conjugate Gradient

Proposition 7.

$$p_{k+j}^T r_k = p_{k+j}^T r_0 \quad \forall\, 0 \le j \le n-k \tag{1.8.28}$$

$$\begin{aligned} p_{k+j}^T r_k &= p_k^T (I - A\overline{P}_k) r_0 && (1.8.29)\\ &= (p_{k+j}^T - p_{k+j}^T A\overline{P}_k) r_0 && (1.8.30)\\ &= p_{k+j}^T r_0 && (1.8.31) \end{aligned}$$

This made the recurrence between successive residual from the ACG possible.

Next, we wish to use this property to find out a recurrences for the residuals of ACG, and here is how we do it:

$$\begin{aligned} r_k - r_{k-1} &= r_0 - A\overline{P}_k r_0 - (r_0 - A\overline{P}_{k-1} r_0) && (1.8.32)\\ &= A\overline{P}_k r_0 - A\overline{P}_{k-1} r_0 && (1.8.33)\\ &= -Ap_{k-1}\frac{\langle p_{k-1}, r_0\rangle}{\langle p_{k-1}, Ap_{k-1}\rangle} && (1.8.34)\\ \implies x_k - x_{k-1} &= p_{k-1}\frac{\langle p_{k-1}, r_0\rangle}{\langle p_{k-1}, Ap_{k-1}\rangle} && (1.8.35)\\ \text{def: } a_{k-1} &:= \frac{\langle p_{k-1}, r_0\rangle}{\langle p_{k-1}, Ap_{k-1}\rangle} = \frac{\langle p_{k-1}, r_{k-1}\rangle}{\langle p_{k-1}, Ap_{k-1}\rangle} && (1.8.36) \end{aligned}$$

We define the value of a_{k-1}, and in above, we have 2 equivalent representation. Please take note that, Proposition still remains true for the ACG algorithm we just developed here.

1.8.5 Residual Assisted Conjugate Gradient

Now, consider the case where, the set of basis vector: $\{u\}_{i=0}^{n-1}$ to be the residual vector generated from the ACG itself. Then there are a series of new added lemmas that are true. However, this is where things started to get exciting, because a short recurrence for p_k during each iteration arised and residuals are all orthgonal. We wish to proceed to prove that part.

Lemma 2.

$$\langle p_{k+j}, Ap_k\rangle = \langle r_k, Ap_{k+j}\rangle = \langle p_{k+j}, Ar_k\rangle \quad \forall\, 0 \le j \le n-k \tag{1.8.37}$$

Proof.

$$\begin{aligned} p_{k+j}Ap_k &= p_{k+j}^T A r_k - p_{k+j}^T A\overline{P}_k A r_k \quad \forall\, 0 \le j \le n-k && (1.8.38)\\ &= p_{k+j}^T A r_k && (1.8.39)\\ \langle p_{k+j}, Ap_k\rangle &= \langle r_k, Ap_{k+j}\rangle = \langle p_{k+j}, Ar_k\rangle && (1.8.40) \end{aligned}$$

□

Lemma 3.

$$\langle r_k, p_k\rangle = \langle r_k, r_k\rangle \tag{1.8.41}$$

Proof.

$$\langle r_k, p_k \rangle = \langle r_k, p_k \rangle \tag{1.8.42}$$
$$= \langle r_k, r_k \rangle - \langle r_k, \overline{P}_k A r_k \rangle \tag{1.8.43}$$
$$= \langle r_k, r_k \rangle \tag{1.8.44}$$

From the first line to the second line, we make use of the definition proposed. □

Theorem 2 (Residual Assisted CG Generates Orthogonal Residuals)**.**

$$\langle r_k, r_j \rangle = 0 \quad \forall\, 0 \le j \le k-1 \tag{1.8.45}$$

Let this above claim be inductively true then consider the following proof:

Proof.

$$r_{k+1} = r_k - a_k A p_k \tag{1.8.46}$$
$$\implies \langle r_{k+1}, r_k \rangle = \langle r_k, r_k \rangle - a_k \langle r_k, A p_k \rangle \tag{1.8.47}$$
$$= \langle r_k, r_k \rangle - \frac{\langle r_k, r_k \rangle}{\langle p_k, A p_k \rangle} \langle r_k, A p_k \rangle \tag{1.8.48}$$
$$= 0 \tag{1.8.49}$$

The first line is from the reccurrence of ACG residuals, and then next we make use of the updated definition for a_k. Next we consider:

$$p_j = (I - \overline{P}_j A) r_j \quad \forall\, 0 \le j \le k-1 \tag{1.8.50}$$
$$r_j = p_j + \overline{P}_j A r_j \tag{1.8.51}$$
$$r_k = (I - A\overline{P}_k) P_0 \tag{1.8.52}$$
$$r_k \perp \operatorname{ran}(P_k) \implies \langle r_k, r_j \rangle = \langle r_k, p_j + \overline{P}_j A r_j \rangle = 0 \tag{1.8.53}$$

Here we again make use of the projector $I - A\overline{P}_k$. The base case of the argument is simple, because $p_0 = r_0$, and by the property of the projector, $\langle r_1, r_0 \rangle = 0$. The theorem is now proven. □

Proposition 8 (RACG Recurrences)**.**

$$p_k = r_k + b_{k-1} p_{k-1} \quad b_{k-1} = \frac{\|r_k\|_2^2}{\|r_{k-1}\|_2^2} \tag{1.8.54}$$

The proof is direct and we start with the definition of ACG, which is given as:

Proof.

$$p_k = (I - \overline{P}_k A) r_k \tag{1.8.55}$$
$$r_k - \overline{P}_k A r_k = r_k - P_k D_k^{-1} P_k^T A r_k \tag{1.8.56}$$
$$= r_k - P_k D_k^{-1} (A P_k)^T r_k \tag{1.8.57}$$

Observe that the term $(AP_k)^T$ can be expanded and we can make use of the Symmetric Property of the operator A_k.

$$(AP_k)^T r_k = \begin{bmatrix} \langle p_0, Ar_k\rangle \\ \langle p_1, Ar_k\rangle \\ \vdots \\ \langle p_{k-1}, Ar_k\rangle \end{bmatrix} \tag{1.8.58}$$

Next, we can make use of Lemma 2 to get rid of Ar_k. Please consider:

$$(AP_k)^T r_k = \begin{bmatrix} \langle p_0, Ar_k\rangle \\ \langle p_1, Ar_k\rangle \\ \vdots \\ \langle p_{k-1}, Ar_k\rangle \end{bmatrix} \tag{1.8.59}$$

The second line is using the property that the matrix A is symmetric, the third line is using the recurrence of the residual of ACG, and the last line is true for all $0 \le j \le k-2$ by the orhogonality of the residual proved in Claim 1. Therefore we have:

$$(AP_k)^T r_k = \begin{bmatrix} \langle p_0, Ar_k\rangle \\ \langle p_1, Ar_k\rangle \\ \vdots \\ \langle p_{k-1}, Ar_k\rangle \end{bmatrix} = a_{k-1}^{-1}\langle r_k, (r_{k-1} - r_k)\rangle \xi_k \tag{1.8.60}$$

Take note that the vector ξ_k is the k th standard basis vector in $\mathbb{R}^k$. And using this we can simplify the expression for p_k into:

$$\begin{aligned}
p_k &= r_k - P_k D_k^{-1}(AP_k)^T r_k && (1.8.61)\\
&= r_k - P_k D_k^{-1} a_{k-1}^{-1}(\langle r_k, (r_{k-1} - r_k)\rangle)\xi_k && (1.8.62)\\
&= r_k - \frac{a_{k-1}^{-1}\langle -r_k, r_k\rangle}{\langle p_{k-1}, Ap_{k-1}\rangle} p_k && (1.8.63)\\
&= r_k + \frac{a_{k-1}^{-1}\langle r_k, r_k\rangle}{\langle p_{k-1}, Ap_{k-1}\rangle} p_k && (1.8.64)\\
&= r_k + \left(\frac{\langle r_{k-1}, r_{k-1}\rangle}{\langle p_{k-1}, Ap_{k-1}\rangle}\right)^{-1} \frac{\langle r_k, r_k\rangle}{\langle p_{k-1}, Ap_{k-1}\rangle} p_k && (1.8.65)\\
&= r_k + \frac{\langle r_k, r_k\rangle}{\langle r_{k-1}, r_{k-1}\rangle} p_k && (1.8.66)
\end{aligned}$$

We make use of the definition for a_{k-1} in figured out for the *ACG* algorithm. At this point, we have proven the RACG recurrences. □

Up until this point we have proven the usual version of conjugate gradient, we started with the minimizations objective and the properties of P_k, then we define a recurrences for the residual (Simultaneously the solution x_k), and the A-Orthogonal vectors using a basis as assistance for the generations process. Next, we make the key changes of the assistance basis, making it equalto the set of residuals vector generated from the algorithm itself; after some proof, we uncovered the exact same parameters found in most of the definitions of the CG algorithm, which we refers to as Residual Assisted Conjugate Gradient. Here we proposed the RACG:

$$p^{(0)} = b - Ax^{(0)} \tag{1.8.67}$$

$$\text{For } i = 0, 1, \cdots \tag{1.8.68}$$

$$\begin{aligned}
a_i &= \frac{\|r^{(i)}\|^2}{\|p^{(i)}\|_A^2} \\
x^{(i+1)} &= x^{(i)} + a_i p^{(i)} \\
r^{(i+1)} &= r^{(i)} - a_i A p^{(i)} \\
b_i &= \frac{\|r^{(j+1)}\|_2^2}{\|r^{(i)}\|_2^2} \\
p^{(i+1)} &= r^{(i+1)} + b_i p^{(i)}
\end{aligned} \tag{1.8.69}$$

That is the algorithm, stated with all the iteration number listed as a super script inside of a parenthesis. Which is equivalent to what we have proven for the Residual Assisted Conjugate Gradient.

1.8.6 RACG and Krylov

The conjugate Gradient Algorithm is actually a residual assisted conjugate gradient, a special case of the algorithm we derived at the start of the excerp. The full algorithm can be seem by the short recurrence for the residual and the conjugation vector. This part is trivial. Next, we want to show the relations to the Krylov Subspace, which only occurs for the Residual Assisted Conjugate Gradient algorithm.

Proposition 9.

$$p_k \in \mathcal{K}_{k+1}(A|r_0) \tag{1.8.70}$$

$$r_k \in \mathcal{K}_{k+1}(A|r_0) \tag{1.8.71}$$

Proof. The base case is tivial and it's directly true from the definition of Residual Assisted Conjugate Gradient: $r_0 \in \mathcal{K}_1(A|r_0), p_0 = r_0 \in \mathcal{K}_1(A|r_0)$. Next, we inductively assume that $r_k \in \mathcal{K}_{k+1}(A|r_0), p_k \in \mathcal{K}_{k+1}(A|r_0)$, then we consider:

$$r_{k+1} = r_k - a_k A p_k \tag{1.8.72}$$

$$\in r_k + A\mathcal{K}_{k+1}(A|r_0) \tag{1.8.73}$$

$$\in r_k + \mathcal{K}_{k+2}(A|r_0) \tag{1.8.74}$$

$$r_k \in \mathcal{K}_{k+1}(A|r_0) \subseteq \mathcal{K}_{k+2}(A|r_0) \tag{1.8.75}$$

$$\implies r_{k+1} \in \mathcal{K}_{k+2}(A|r_0) \tag{1.8.76}$$

At the same time the update of p_k would asserts the property that:

$$p_{k+1} = r_{k+1} + b_k p_k \tag{1.8.77}$$

$$\in r_{k+1} + \mathcal{K}_{k+1}(A|r_0) \tag{1.8.78}$$

$$\in \mathcal{K}_{k+2}(A|r_0) \tag{1.8.79}$$

This is true because r_{k+1} is already a member of the expanded subspace $\mathcal{K}_{k+2}(A|r_0)$. And from this formulation of the algorithm, we can update the Petrov Galerkin's Conditions to be:

$$\text{choose: } x_k \in x_0 + \mathcal{K}_k(A|r_0) \text{ s.t: } r_k \perp \mathcal{K}_k(A|r_0) \tag{1.8.80}$$

Take note that, $\text{ran}(P_k) = \mathcal{K}_k(A|r_0)$ because the index starts with zero. The above formulations gives theoretical importance for the Conjugate Gradient Algorithm. $\square$