Problem Setup:

Consider the quadratic minimization problem

$$\min_{\vec{w} \in \mathbb{R}^d} \frac{1}{2} \| X\vec{w} - \vec{y} \|_2^2 ,$$

where $X \in \mathbb{R}^{n \times d}$ is a data matrix, and $\vec{y} \in \mathbb{R}^n$ is a vector of observations. This problem is equivalent the following.

$$\min_{\vec{w} \in \mathbb{R}^d} \frac{1}{2} \vec{w}^T X^T X \vec{w} - \vec{w}^T X^T \vec{y} + \frac{1}{2} \vec{y}^T \vec{y} .$$

Let us denote $A := X^T X$ and $\vec{b} := X^T \vec{y}$. Then, an equivalent optimization objective can be written as

$$\min_{\vec{w} \in \mathbb{R}^d} \mathcal{L}(\vec{w}) := \frac{1}{2} \vec{w}^T A \vec{w} - \vec{b}^T \vec{w} .$$

Gradient Descent with Fixed Step Size:

Algorithm:

Initialized at $\vec{w}_0 \in \mathbb{R}^d$, gradient descent (GD) with step size $\eta > 0$ iterates

$$\vec{w}_{k+1} = \vec{w}_k - \eta (A\vec{w}_k - \vec{b})$$

$$= \vec{w}_k - \eta A (\vec{w}_k - \vec{w}^*) ,$$

where $\vec{w}^*$ is any minimizer of $\mathcal{L}$, i.e. one satisfying $A\vec{w}^* = \vec{b}$.

To analyze GD,

$$\begin{aligned}\vec{w}_{k+1} - \vec{w}^* &= \vec{w}_k - \vec{w}^* - \eta A(\vec{w}_k - \vec{w}^*) \\ &= (I - \eta A)(\vec{w}_k - w^*).\end{aligned}$$

Denote $\vec{\varepsilon}_k = \vec{w}_k - \vec{w}^*$. Then,

$$\begin{aligned}\vec{\varepsilon}_{k+1} &= (I - \eta A)\,\vec{\varepsilon}_k \\ \Rightarrow \quad \vec{\varepsilon}_{k+1} &= (I - \eta A)^k \vec{\varepsilon}_k .\end{aligned}$$

Now, observe that

$$\begin{aligned}\mathcal{L}(\vec{w}_k) - \mathcal{L}^* &= \frac{1}{2}\vec{w}_k^T A \vec{w}_k - \vec{b}^T \vec{w}_k - \frac{1}{2}\vec{w}^{*T} A \vec{w} + \vec{b}^T \vec{w}^* \\ &= \frac{1}{2}\vec{w}_k^T A \vec{w}_k - \vec{b}^T \vec{w}_k - \frac{1}{2}\vec{w}^{*T} A \vec{w} + \vec{w}^{*T} A \vec{w}^* \\ &\overset{\varepsilon}{=} \frac{1}{2}\vec{w}_k^T A \vec{w}_k - \vec{b}^T \vec{w}_k + \frac{1}{2}\vec{w}^{*T} A \vec{w} \\ &= \frac{1}{2}\vec{w}_k^T A \vec{w}_k - \vec{w}^{*T} A \vec{w}_k + \frac{1}{2}\vec{w}^{*T} A \vec{w} \\ &= \frac{1}{2}(\vec{w}_k - \vec{w}^*)^T A (\vec{w}_k - \vec{w}^*) \\ &= \frac{1}{2}\vec{\varepsilon}_k^T A \varepsilon_k \\ &= \frac{1}{2}\|\varepsilon_k\|_A^2 ,\end{aligned}$$

where $\mathcal{L}^* := \mathcal{L}(\vec{w}^*)$, and $\|\vec{v}\|_A = (\vec{v}^T A \vec{v})^{1/2}$.

We denote the eigenvalues of A by

$$0 < \alpha = \lambda_1 \leq \lambda_2 \leq \dots \leq \lambda_d = \beta$$

Visualization in $\mathbb{R}^2$:

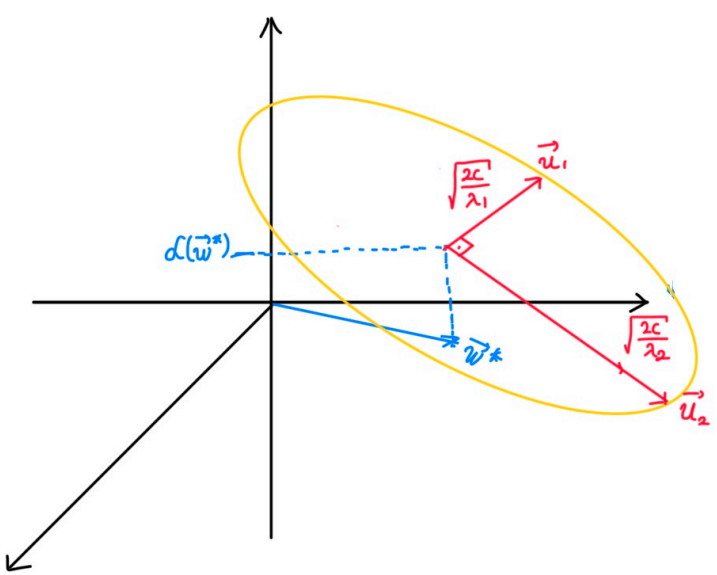
Suppose $d = 2$. To visualize level sets, let $\mathcal{L}(\vec{w}) - \mathcal{L}^* = c$.

Then, $\frac{1}{2}\|\vec{\varepsilon}\|_A^2 = c \Rightarrow \vec{\varepsilon}^T A \vec{\varepsilon} = 2c$. This implies that

$$[\langle \vec{\varepsilon}, \vec{u}_1 \rangle, \langle \vec{\varepsilon}, \vec{u}_2 \rangle] \begin{bmatrix} \lambda_1 & 0 \\ 0 & \lambda_2 \end{bmatrix} \begin{bmatrix} \langle \vec{\varepsilon}, \vec{u}_1 \rangle \\ \langle \vec{\varepsilon}, \vec{u}_2 \rangle \end{bmatrix} = 2c.$$

where $A = U \Lambda U^T \in \mathbb{R}^{2 \times 2}$.

$$\Leftrightarrow \sum_{i=1}^{2} \lambda_i \langle \vec{\varepsilon}, \vec{u}_i \rangle^2 = 2c$$



Convergence Rate for GD with Fixed Step-Size:

Let $\vec{v}_1, \vec{v}_2, \ldots, \vec{v}_d$ be an eigenbasis of A with $A\vec{v}_i = \lambda \vec{v}_i$.

Let $\vec{\varepsilon}_0 = \sum_{i=1}^{d} c_i \vec{v}_i$. Then,

$$\vec{\varepsilon}_k = (I - \eta A)^k \vec{\varepsilon}_0 = (I - \eta A)^k \sum_{i=1}^{d} c_i \vec{v}_i$$
$$= \sum_{i=1}^{d} (1 - \eta \lambda_i)^k c_i v_i .$$

Therefore,

$$\|\vec{\varepsilon}_k\|_A = \sum_{i=1}^{d} \lambda_i (1 - \eta \lambda_i)^{2k} c_i^2$$
$$\leq \max_{1 \leq i \leq d} (1 - \eta \lambda_i)^{2k} \sum_{i=1}^{d} \lambda_i c_i^2$$
$$= \rho(\eta)^{2k} \|\vec{\varepsilon}_0\|_A^2 ,$$

where

$$\rho(\eta) = \max_{1 \leq i \leq d} |1 - \eta \lambda_i| = \max\{|1 - \eta\alpha|, |1 - \eta\beta|\} .$$

Hence if $\alpha \neq 0$, we observe the following.

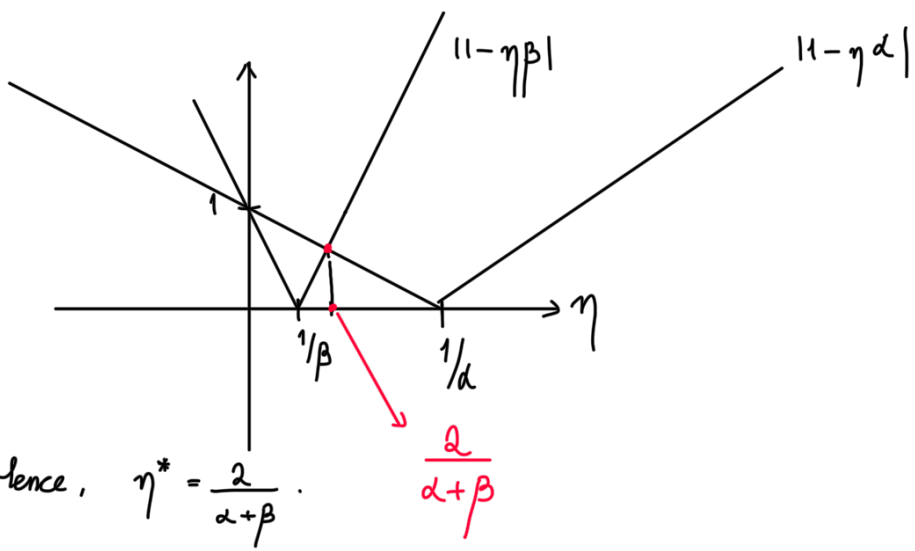
Theorem: For any $\eta \in (0, 2/\beta)$, $\rho(\eta) \in (0,1)$. Thus, GD iterates enjoy the linear rate of convergence:

$$\mathcal{L}(\vec{w}_k) - \mathcal{L}^* \leq \rho(\eta)^{2k} (\mathcal{L}(\vec{w}_0) - \mathcal{L}^*) .$$

Optimal Step Size:

When $\alpha \neq 0$, the optimal step size is determined by the following objective.

$$\min_{\eta \in (0, 2/\beta)} \max \left\{ |1-\eta\alpha| , |1-\eta\beta| \right\}$$



Hence, $\eta^* = \frac{2}{\alpha+\beta}$.

Corollary: Suppose $\alpha > 0$ and set $\eta = \eta^* = \frac{2}{\beta+\alpha}$. Then GD satisfies

$$\mathcal{L}(\vec{w}_k) - \mathcal{L}^* \leq \left(\frac{\kappa - 1}{\kappa + 1} \right)^{2k} (\mathcal{L}(\vec{w}_0) - \mathcal{L}^*) .$$

Proof:

$$\rho(\eta^*) = \left| 1 - \frac{2\alpha}{\alpha+\beta} \right| = \left| \frac{\beta-\alpha}{\beta+\alpha} \right| = \left| \frac{\kappa-1}{\kappa+1} \right|$$

Practical Step Size:

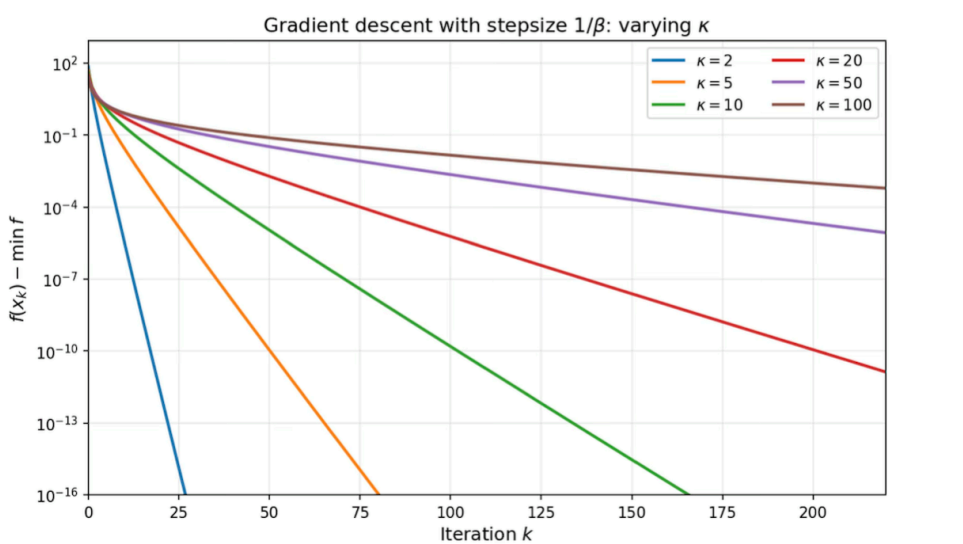
In practice, the smallest eigenvalue α is often unknown or expensive to estimate. Therefore, a natural choice is $\eta = \frac{1}{\beta}$, which requires only an upper bound on the spectrum.

Corollary: Suppose $\alpha > 0$ and $\eta = \frac{1}{\beta}$. Then GD iterates enjoy the following convergence rate:

$$\mathcal{L}(\vec{w}_k) - \mathcal{L}^* \leq \left(1 - \frac{1}{\kappa}\right)^{2k} \left(\mathcal{L}(\vec{w}_0) - \mathcal{L}^*\right).$$

Proof: $\eta = 1/\beta$ yields

$$\rho(\eta = 1/\beta) = \max\left\{\left|1 - \frac{1}{\kappa}\right|, 0\right\}$$



Acceleration by Chebyshev Step Sizes:

We now show that by considering the performance of GD on entire time horizon, it is possible to choose a time-varying step size that yields a faster convergence. To see this, consider time-varying step sizes $\eta_0, \eta_1, \dots, \eta_{k-1}$. Therefore

$$\begin{aligned}\vec{\varepsilon}_k &= (I - \eta_{k-1}A)(I - \eta_{k-2}A)\dots(I - \eta_0 A)\vec{\varepsilon}_0 \\ &= p_k(A)\vec{\varepsilon}_0 ,\end{aligned}$$

where p_k is the degree-k polynomial such that

$$p_k(\lambda) = \prod_{i=1}^{d} (1 - \eta_i \lambda) .$$

Note that we have $p_k(0) = 1$ regardless of choices of step sizes.

Expanding $\vec{\varepsilon}_k$ in the eigenbasis of A as before yields

$$\mathcal{L}(\vec{w}_k) - \mathcal{L}^* = \frac{1}{2}\|\vec{\varepsilon}_k\|_A^2 = \frac{1}{2}\vec{\varepsilon}_0^T p_k(A) A p_k(A)\vec{\varepsilon}_0 .$$ Let $\vec{\varepsilon}_0 = \sum_{i=1}^{d} c_i \vec{v}_i$.

Then, $p_k(A)\vec{\varepsilon}_0 = \sum_{i=1}^{d} c_i p_k(A)\vec{v}_i = \sum_{i=1}^{d} c_i p_k(\lambda_i)\vec{v}_i$. This implies that

$$\mathcal{L}(\vec{w}_k) - \mathcal{L}^* = \frac{1}{2}\|\vec{\varepsilon}_k\|_A^2 = \frac{1}{2}\sum_{i=1}^{d} \lambda_i \rho_k(\lambda_i)^2 c_i^2$$

$$\leq \max_{\lambda \in [\alpha, \beta]} \rho_k(\lambda)^2 \frac{1}{2}\|\vec{\varepsilon}_0\|_A^2 .$$

We can rearrange the inequality as follows:

$$\frac{\mathcal{L}(\vec{w}_k) - \mathcal{L}^*}{\mathcal{L}(\vec{w}_0) - \mathcal{L}^*} \leq \max_{\lambda \in [\alpha, \beta]} \rho_k(\lambda)^2 .$$

Notice that as we vary the step sizes $\eta_0, \eta_1, \cdots, \eta_{k-1}$, any polynomial $p(\lambda)$ of degree at most k, having all real roots, and satisfying $p_k(0) = 1$ can be realized as $p_k(\lambda)$. Thus, choosing time-varying step-sizes is equivalent to choosing such a polynomial. The best possible convergence after k steps is, therefore, determined by the minimax polynomial problem.

$$\min_{\substack{p \in P_k \\ p(0)=1}} \max_{\lambda \in [\alpha, \beta]} p(\lambda)^2 ,$$

where $\mathbb{P}_k$ denotes the set of all polynomials of degree at most k with all real roots. The solution to this classical variational problem is described through so-called Chebyshev polynomials of the first kind.

Chebyshev Polynomials :

The Chebyshev polynomials of first kind of degree k, denoted T_k, is defined recursively : set $T_0(x) = 1$ and $T_1(x) = x$ and define

$$T_{k+1} = 2x\, T_k(x) - T_{k-1}(x) \quad .$$

An equivalent characterization is

$$T_k(\cos(\theta)) = \cos(k\theta) \qquad \forall \theta \in [0, \pi] \, .$$

Chebyshev polynomials play a special role since they solve the following extremal problem :

Remark : Any degree-k polynomial $p(x)$ with the same leading coefficient as T_k satisfies

$$\max_{x \in [-1,1]} |p(x)| \geq \max_{x \in [-1,1]} |T_k(x)| = 1$$

In other words, among all degree-k polynomials with the same leading coefficient as T_k, the Chebyshev polynomial has the smallest absolute value on $[-1,1]$.

Chebyshev polynomials satisfy the following key properties:

① **Boundedness**: The inequality $|T_k(t)| \leq 1$ for all $t \in [-1,1]$.

② **Roots**: T_k has k roots in $(-1,1)$ at $t_j = \cos\left(\frac{(2j-1)\pi}{2k}\right)$ for $j = 1, 2, \dots, k$.

③ **Explosion**: For $t > 1$, we have $T_k(t) = \cosh(k \operatorname{arccosh}(t))$.

The Optimal Polynomial:

Let us see how Chebyshev polynomials yield a solution to the minimax problem. We rescale the interval $[\alpha, \beta]$ to $[-1,1]$ with the affine change of coordinates $\varphi(\lambda) = \frac{\beta + \alpha - 2\lambda}{\beta - \alpha}$. Note that $\varphi(0) = \delta := \frac{\beta+\alpha}{\beta-\alpha} = \frac{\kappa+1}{\kappa-1}$. Thus, under this substitution, any degree-k polynomial $p(\lambda)$ with $p(0) = 1$ corresponds to a degree-k polynomial $q = p \circ \varphi^{-1}$ with $q(\delta) = 1$, and

$$\max_{\lambda \in [\alpha,\beta]} |p(\lambda)| = \max_{t \in [-1,1]} |q(t)|$$

We must, therefore, find the degree-k polynomial q with $q(\delta) = 1$ that has the smallest maximum on $[-1, 1]$. By properties ① and ③ above, T_k is bounded by 1 on $[-1, 1]$ while $T_k(\delta) \gg 1$ for large k.

This makes the rescaled polynomial

$$q_k^*(t) := \frac{T_k(t)}{T_k(\delta)}$$

an excellent candidate: it satisfies $q_k^*(\delta) = 1$ and $\max_{t \in [-1,1]} |q_k^*(t)| = \frac{1}{T_k(\delta)}$, which is small because $T_k(\delta)$ grows exponentially in k. Transforming back to the λ-variable, the optimal polynomial is

$$p_k^*(\lambda) := (q_k^* \circ \varphi)(\lambda) = \frac{T_k\left(\frac{\beta + \alpha - 2\lambda}{\beta - \alpha}\right)}{T_k\left(\frac{\kappa + 1}{\kappa - 1}\right)}$$

Summarizing, we have the following lemma.

Lemma (Chebyshev minimax): Suppose $\alpha > 0$. Then with $\delta = \frac{\kappa+1}{\kappa-1}$, the minimax value satisfies

$$\min_{\substack{p \in \mathbb{P}_k \\ p(0)=1}} \max_{\lambda \in [\alpha,\beta]} |p^*(\lambda)| \leq \frac{1}{T_k(\delta)} \leq 2\left(\frac{\sqrt{\kappa}-1}{\sqrt{\kappa}+1}\right)^k.$$

Proof: We already proved the first inequality. For the second inequality, we use the identity $T_k(x) = \cosh(k \operatorname{arccosh}(x))$ valid for every real number $x > 1$. Applying this identity with the quantity δ gives the relation

$$\operatorname{arccosh}(\delta) = \ln\left(\delta + \sqrt{\delta^2-1}\right) = \ln\left(\frac{\sqrt{\kappa}+1}{\sqrt{\kappa}-1}\right).$$

Consequently,

$$T_k(\delta) = \cosh\left(k \ln\left(\frac{\sqrt{\kappa}+1}{\sqrt{\kappa}-1}\right)\right) = \frac{1}{2}\left[\left(\frac{\sqrt{\kappa}+1}{\sqrt{\kappa}-1}\right)^k + \left(\frac{\sqrt{\kappa}-1}{\sqrt{\kappa}+1}\right)^k\right]$$

$$\geqslant \frac{1}{2}\left(\frac{\sqrt{\kappa}+1}{\sqrt{\kappa}-1}\right)^k.$$

The roots of $p_k^*(\lambda)$ on $[\alpha, \beta]$ are the images of the Chebyshev roots t_j under the inverse map $\varphi(t)^{-1} = \frac{\beta+\alpha}{2} - \frac{\beta-\alpha}{2} t$, yielding the step sizes $\eta_j = \frac{1}{\lambda_j}$. Thus, we have the following

Theorem (Chebyshev Step Sizes)

Define the step sizes

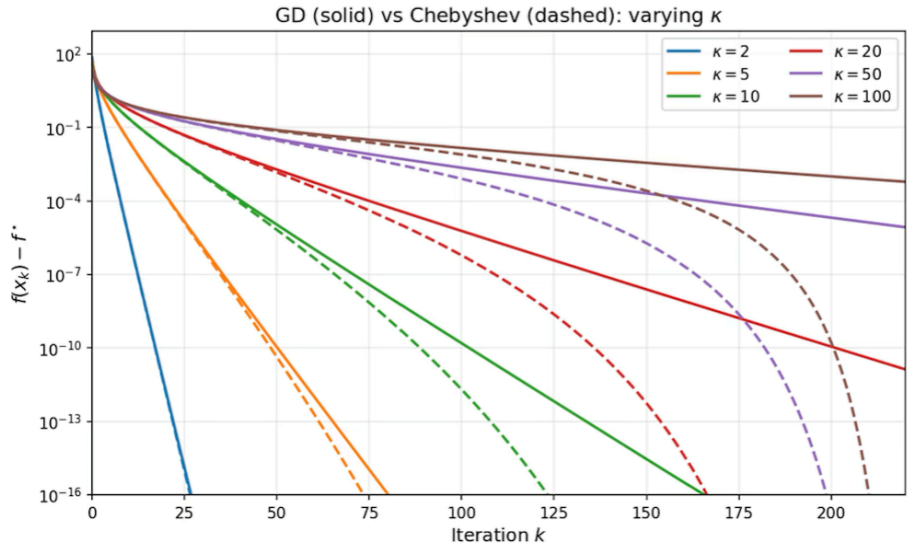
$$\eta_j = \frac{1}{\lambda_j} \quad \text{where} \quad \lambda_j = \frac{\beta+\alpha}{2} - \frac{\beta-\alpha}{2} \cos\left(\frac{(2j-1)\pi}{2k}\right) \quad \text{for } j \in [k].$$

Then, as long as $\alpha > 0$ the GD iterates satisfy

$$\mathcal{L}(\vec{w}_k) - \mathcal{L}^* \leq 4\left(\frac{\sqrt{\kappa}-1}{\sqrt{\kappa}+1}\right)^{2k} (\mathcal{L}(\vec{w}_0) - \mathcal{L}^*)$$

Remark: Thus, the iteration complexity of Chebyshev-accelerated GD is $O\left(\frac{\sqrt{\kappa}\log(\mathcal{L}(\vec{w}_0)-\mathcal{L}^*)}{\varepsilon}\right)$ - a square root improvement over the $O\left(\frac{\kappa\log(\mathcal{L}(\vec{w}_0)-\mathcal{L}^*)}{\varepsilon}\right)$ complexity of fixed-step-size GD.

As a final illustration, the plot below overlays GD with stepsize $1/\beta$ (solid) and Chebyshev-accelerated GD with $k = 220$ (dashed) for varying condition numbers. The Chebyshev curves stay nearly flat during the cycle and then drop sharply near the final iteration, reaching machine precision much sooner than GD for every value of κ.



Krylov Subspace Method and Conjugate Gradient (CG):

From polynomials to Krylov Subspaces:

The Chebyshev method requires advance knowledge of the extreme eigenvalues α and β. Moreover, the total number of iterations must be set in advance in order to define the step sizes. Therefore, a natural question arises

"Can we design an adaptive algorithm that matches this rate adaptively, without knowing the spectrum nor setting the time horizon?"

The key observation is that GD with any sequence of step sizes produces iterates that lie in a specific linear subspace. Due to the recursion $x_{j+1} = x_j - \eta_j (A\vec{x}_j - \vec{b})$, one readily verifies the inclusion

$$\vec{x}_k \in \vec{x}_0 + K_k(A, \vec{r}_0),$$

where $r_0 := \vec{b} - A\vec{x}_0$ is the initial residual, and

$$K_k(A, \vec{r}_0) := \text{span}\{\vec{r}_0, A\vec{r}_0, A^2\vec{r}_0, \dots, A^{k-1}\vec{r}_0\}$$

is the Krylov subspace of order k. Both fixed-step-size GD and the Chebyshev method search within this subspace but do not fully exploit it. The natural idea is to search optimally within the Krylov subspace at each step.

Krylov Subspace Method:

$$\vec{x}_k = \underset{\vec{x} \in \vec{x}_0 + K_k(A, \vec{r}_0)}{\text{argmin}} f(\vec{x})$$

Theorem: Assuming $\alpha > 0$, the Krylov subspace method satisfies

$$f(\vec{x}_k) - f^* \leq 4\left(\frac{\sqrt{\kappa} - 1}{\sqrt{\kappa} + 1}\right)^{2k} (f(\vec{x}_0) - f^*) \quad \forall k \geq 0.$$

Moreover, the method converges in at most m iterations, where m is the number of distinct eigenvalues of A.

Proof: The k-th iterate produced by the Chebyshev step sizes lies in $\vec{x}_0 + K_k(A, \vec{r}_0)$, whereas the Krylov method minimizes f over that entire affine space, and so cannot do worse.

To prove finite termination, we need to show that $\vec{x}^*$ lies in $\vec{x}_0 + K_k(A, \vec{r}_0)$. Define $\vec{\varepsilon}_0 := \vec{x}_0 - \vec{x}^*$. Observe that a point lies in $\vec{x}_0 + K_k(A, \vec{r}_0)$ iff it can be written as $\vec{x}_0 + q(A)\vec{r}_0$ for some polynomial q of degree at most $k-1$. Note that $-\vec{r}_0 = A\vec{x}_0 - \vec{b} - A\vec{x}_0 - A\vec{x}^* = A(\vec{x}_0 - \vec{x}^*) = A\vec{\varepsilon}_0$. Then,

$$\vec{x}_0 - x^* + q(A)\vec{r}_0 = \vec{\varepsilon}_0 - q(A)A\vec{\varepsilon}_0 = p(A)\vec{\varepsilon}_0,$$

where $p(\lambda) = (1 - \lambda q(\lambda))$ has degree at most k and satisfies $p(0) = 1$. Observe that the polynomials p that have this form are exactly the polynomials of degree at most k having $p(0) = 1$.

With this in mind, define

$$p(\lambda) := \prod_{i=1}^{m} \left(1 - \frac{\lambda}{\lambda_i}\right),$$

where $\lambda_1, \lambda_2, \ldots, \lambda_m$ are the distinct eigenvalues of A. Since this polynomial p vanishes at every eigenvalue of A, we deduce that $p(A)\vec{\varepsilon}_0 = 0$. This implies that

$$\vec{x}_0 + q(A)\vec{r}_0 = \vec{x}^*$$

The Conjugate Gradient Algorithm

The conjugate gradient algorithm is an implementation of the Krylov method that uses only one matrix-vector product per iteration.

The key idea is to iteratively build a basis of the Krylov subspaces that is orthogonal with respect to the inner product $\langle \vec{x}, \vec{y} \rangle_A = \vec{x}^T A \vec{y}$ so that each succesive minimization reduces to a single line search.

Suppose that we have constructed an A-orthogonal basis $\{p_i\}_{i=0}^{k-1}$ for K_{k-1} and we have an available minimizer $\vec{x}_k$ of f on $\vec{x}_0 + K_k$. Let us see how we can efficiently extend the A orthogonal basis to K_k and construct the minimizer $\vec{x}_{k+1}$ of f on $\vec{x}_0 + K_{k+1}$. To this end, define the residuals

$$r_i = -\nabla f(\vec{x}_i) = b - A\vec{x}_i$$

Observe that we may write $\vec{r}_k = \vec{b} - A\vec{x}_k = \vec{r}_0 - A(\vec{x}_k - \vec{x}_0)$ and therefore $\vec{r}_k$ lies in K_{k+1}.

Why does $\vec{r}_k$ lie in K_{k+1}?

Note that $\vec{x}_k = \underset{\vec{x} \in \vec{x}_0 + K_k(A, \vec{r}_0)}{\operatorname{argmin}} f(\vec{x})$. This implies that $\vec{x}_k \in \vec{x}_0 + K_k(A, \vec{r}_0)$. This implies that $\vec{x}_k - \vec{x}_0 \in K_k$. Note that for any $\vec{x} \in K_k$, $\gamma \vec{r}_0 + cA\vec{x} \in K_{k+1}$ for any $\gamma, c \in \mathbb{R}$.

Note that $\{\vec{p}_i\}_{i=0}^{k-1}$ forms an A-orthogonal basis for K_k.

To find the $\vec{p}_k$, we can apply Gram-Schmidt procedure.

$$\vec{p}_k = \vec{r}_k - \sum_{i=0}^{k-1} \frac{\langle \vec{p}_i, \vec{r}_k \rangle_A}{\langle \vec{p}_i, \vec{p}_i \rangle_A} \vec{p}_i$$

Since $\vec{x}_k$ minimizes $f(\vec{x})$ over the affine subspace $\vec{x}_0 + K_k$, we need every directional derivative along every direction in K_k to vanish. Since for any $i \in [k-2]$: $Ap_i \in K_k$,

$$\vec{p}_i^T A \vec{r}_k = \vec{r}_k^T \underbrace{A p_i}_{\in K_k} = 0 \quad \forall i \in [k-2].$$

Hence, suppose

$$\vec{p}_k = \vec{r}_k + \beta_{k-1} \vec{p}_{k-1},$$

for a constant β_{k-1}. Since $\{p_i\}_{i=0}^{k}$ are A-orthogonal,

$$\beta_{k-1} = \frac{-\vec{r}_k^T A \vec{p}_{k-1}}{\vec{p}_{k-1}^T A \vec{p}_{k-1}}$$

It remains to declare

$$\eta_k = \arg\min_{\eta} f(\vec{x}_k + \eta \vec{p}_k).$$

Note that $f(\vec{x}_k + \eta\vec{p}_k) = \frac{1}{2}\|D(\vec{x}_k + \eta\vec{p}_k) - \vec{y}\|_2^2$.

$$\frac{\partial f(\vec{x}_k + \eta\vec{p}_k)}{\partial \eta} = \langle \nabla f(\vec{x}_k + \eta\vec{p}_k), \vec{p}_k \rangle = \langle A\vec{x}_k + \eta A\vec{p}_k - \vec{b}, \vec{p}_k \rangle$$

$$= \vec{p}_k^{\,T} A \vec{x}_k + \eta \|\vec{p}_k\|_A^2 - \vec{p}_k^{\,T}\vec{b}$$

$$\Rightarrow \frac{\partial^2 f(\vec{x}_k + \eta\vec{p}_k)}{\partial \eta^2} > 0 \Rightarrow g(\eta) := f(\vec{x}_k + \eta\vec{p}_k) \text{ is convex.}$$

$$\Rightarrow \eta_k = \frac{\vec{p}_k^{\,T}\vec{r}_k}{\vec{p}_k^{\,T} A \vec{p}_k}$$

Algorithm 1 (Conjugate Gradient Method)

Input: $x_0 \in \mathbb{R}^d$

1. Set $r_0 = b - Ax_0$, $p_0 = r_0$
2. **For** $k = 0, 1, 2, \ldots$ do:
3. $\eta_k = \dfrac{r_k^\top r_k}{p_k^\top A p_k} = \dfrac{\vec{r}_k^{\,T}\vec{r}_k}{\vec{p}_k^{\,T} A \vec{p}_k}$
4. $x_{k+1} = x_k + \eta_k p_k$
5. $r_{k+1} = r_k - \eta_k A p_k$
6. $\beta_k = -\dfrac{r_{k+1}^\top A p_k}{p_k^\top A p_k}$
7. $p_{k+1} = r_{k+1} + \beta_k p_k$