Gradient Methods (§8.1)

- Given $\boldsymbol{x}^{(k)}$.
- The vector $-\nabla f(\boldsymbol{x}^{(k)})$ points in the direction of maximum rate of decrease.
- Makes sense to choose $\boldsymbol{d}^{(k)} = -\nabla f(\boldsymbol{x}^{(k)})$.
- Gradient algorithm:
$$\boldsymbol{x}^{(k+1)} = \boldsymbol{x}^{(k)} - \alpha_k \nabla f(\boldsymbol{x}^{(k)}).$$
- Step size α_k can be chosen in many different ways.
- For sufficiently small step size, the gradient algorithm has descent property.
- Prop.: Suppose $\nabla f(\boldsymbol{x}^{(k)}) \neq \mathbf{0}$. There exists $\bar{\alpha} > 0$ such that for all $\alpha_k \in (0, \bar{\alpha})$, we have
$$f(\boldsymbol{x}^{(k+1)}) < f(\boldsymbol{x}^{(k)}).$$
- Remark: if $\nabla f(\boldsymbol{x}^{(k)}) = \mathbf{0}$, the FONC holds. Can use as basis for stopping.

Proof of prop.:

- Proof: Consider $\phi(\alpha) = f(\boldsymbol{x}^{(k)} - \alpha \nabla f(\boldsymbol{x}^{(k)}))$.
- By chain rule, we have
$$\phi'(0) = -\|\nabla f(\boldsymbol{x}^{(k)})\|^2 < 0.$$
- Hence, there exists $\bar{\alpha} > 0$ such that for all $\alpha_k \in (0, \bar{\alpha})$, we have
$$\phi(\alpha_k) < \phi(0).$$
- Rewriting, we obtain
$$f(\boldsymbol{x}^{(k+1)}) < f(\boldsymbol{x}^{(k)}).$$
- Several possible choices for α_k.
- If α_k too small, we need to iterate many times to get to the solution.
- If α_k too big, algorithm may zig-zag around the solution (overshoot).
- We can either fix $\alpha_k = \alpha$ for all k, or let α_k vary from iteration to iteration.
- Greedy scheme:
$$\alpha_k = \arg\min_{\alpha \geq 0} f(\boldsymbol{x}^{(k)} - \alpha \nabla f(\boldsymbol{x}^{(k)})).$$
Name: *Steepest descent algorithm*

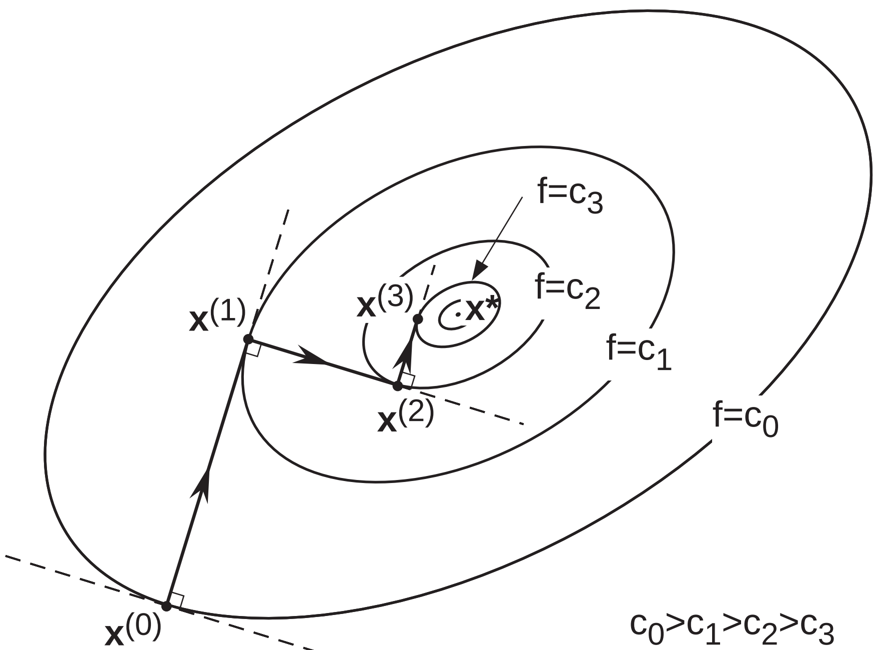
Steepest descent algorithm (§8.2)

- See Example 8.1.
- The steepest descent algorithm has the descent property. Why?
- Prop. (8.1): In the steepest descent algorithm,

$$\boldsymbol{x}^{(k+1)} - \boldsymbol{x}^{(k)}$$

  is orthogonal to

$$\boldsymbol{x}^{(k+2)} - \boldsymbol{x}^{(k+1)}.$$



Proof of prop.:

- We have

$$\begin{aligned}&\left\langle \boldsymbol{x}^{(k+1)} - \boldsymbol{x}^{(k)}, \boldsymbol{x}^{(k+2)} - \boldsymbol{x}^{(k+1)} \right\rangle \\ &\qquad = \alpha_k \alpha_{k+1} \left\langle \nabla f(\boldsymbol{x}^{(k)}), \nabla f(\boldsymbol{x}^{(k+1)}) \right\rangle .\end{aligned}$$

- To complete the proof it is enough to show that

$$\left\langle \nabla f(\boldsymbol{x}^{(k)}), \nabla f(\boldsymbol{x}^{(k+1)}) \right\rangle = 0.$$

- Let $\phi(\alpha) = f(\boldsymbol{x}^{(k)} - \alpha \nabla f(\boldsymbol{x}^{(k)}))$.
- By FONC, $\phi'(\alpha_k) = 0$.

- By chain rule, $\nabla f(\boldsymbol{x}^{(k)})^T \nabla f(\boldsymbol{x}^{(k+1)}) = 0$.

- Typical stopping criteria:
$$\|\nabla f(\boldsymbol{x}^{(k)})\| \leq \varepsilon,$$
or
$$\|\boldsymbol{x}^{(k)} - \boldsymbol{x}^{(k-1)}\| \leq \varepsilon,$$
or
$$\frac{\|\nabla f(\boldsymbol{x}^{(k)})\|}{\|\nabla f(\boldsymbol{x}^{(0)})\|} \leq \varepsilon,$$
or
$$\frac{\|\boldsymbol{x}^{(k)} - \boldsymbol{x}^{(k-1)}\|}{\|\boldsymbol{x}^{(k)}\|} \leq \varepsilon,$$
where ε is prespecified.

- The latter two are usually preferable, because they are "scale-free."

Analysis of optimization algorithms

- Rely heavily on mathematical tools.
- "Do we really have to go through this?"
- Analysis provides insight into:
  - Range of applicability of an algorithm.
  - Appropriate choice of algorithm for a given problem.
  - Qualitative behavior of an algorithm.
- We must be able to answer:
  - Does the method work?
  - When does it work?
  - How well does it work?
- Not good enough to superficially use commercial optimization software package.

Several characterizations of performance:

- *Globally convergent*: starting from any initial point, the algorithm converges to a "solution."
- Usually, by "solution" we mean a point satisfying the FONC.
- *Locally convergent*: starting from an initial point that is close enough to a solution, the algorithm converges to the solution.
- *Rate of convergence*: how fast an algorithm converges.

Analysis of gradient methods (§8.3)

- We analyze gradient algorithms applied to quadratics only:

$$f(\boldsymbol{x}) = \frac{1}{2}\boldsymbol{x}^T\boldsymbol{Q}\boldsymbol{x} - \boldsymbol{b}^T\boldsymbol{x}$$

  where $\boldsymbol{Q} > 0$.

- We restrict our attention to quadratics because:
  - Simplifies analysis.
  - Local behavior near solution. (Global convergence for quadratics tells us something about local convergence in more general functions. How?)

Steepest descent method applied to quadratics

- Consider

$$f(\boldsymbol{x}) = \frac{1}{2}\boldsymbol{x}^T\boldsymbol{Q}\boldsymbol{x} - \boldsymbol{b}^T\boldsymbol{x},$$

  where $\boldsymbol{Q} > 0$.

- We have $\nabla f(\boldsymbol{x}) = \boldsymbol{Q}\boldsymbol{x} - \boldsymbol{b}$ and $\boldsymbol{F}(\boldsymbol{x}) = \boldsymbol{Q}$.
- For simplicity, write $\boldsymbol{g}^{(k)} = \nabla f(\boldsymbol{x}^{(k)})$.
- We can find an explicit formula for α_k.

- Let $\phi(\alpha) = f(\boldsymbol{x}^{(k)} - \alpha\boldsymbol{g}^{(k)})$.
- $\phi(\alpha)$ is a quadratic:

$$\phi(\alpha) = \left(\frac{1}{2}\boldsymbol{g}^{(k)T}\boldsymbol{Q}\boldsymbol{g}^{(k)}\right)\alpha^2 - \left(\boldsymbol{g}^{(k)T}\boldsymbol{g}^{(k)}\right)\alpha + \text{constant}.$$

- Hence, we get

$$\alpha_k = \frac{\boldsymbol{g}^{(k)T}\boldsymbol{g}^{(k)}}{\boldsymbol{g}^{(k)T}\boldsymbol{Q}\boldsymbol{g}^{(k)}}.$$

- Algorithm applied to quadratic:

$$\boldsymbol{x}^{(k+1)} = \boldsymbol{x}^{(k)} - \left(\frac{\boldsymbol{g}^{(k)T}\boldsymbol{g}^{(k)}}{\boldsymbol{g}^{(k)T}\boldsymbol{Q}\boldsymbol{g}^{(k)}}\right)\boldsymbol{g}^{(k)}.$$

- For convenience, instead of working with f, we work with

$$\begin{aligned} V(\boldsymbol{x}) &= f(\boldsymbol{x}) + \frac{1}{2}\boldsymbol{x}^{*T}\boldsymbol{Q}\boldsymbol{x}^* \\ &= \frac{1}{2}(\boldsymbol{x}-\boldsymbol{x}^*)^T\boldsymbol{Q}(\boldsymbol{x}-\boldsymbol{x}^*), \end{aligned}$$

  where $\boldsymbol{x}^* = \boldsymbol{Q}^{-1}\boldsymbol{b}$.

- The constant we add to f does not change solution. Why?

- Lemma (8.1): We have

$$V(\boldsymbol{x}^{(k+1)}) = (1-\gamma_k)\, V(\boldsymbol{x}^{(k)}),$$

  where γ_k is defined as

$$\gamma_k = \alpha_k \frac{\boldsymbol{g}^{(k)T}\boldsymbol{Q}\boldsymbol{g}^{(k)}}{\boldsymbol{g}^{(k)T}\boldsymbol{Q}^{-1}\boldsymbol{g}^{(k)}}\left(2\frac{\boldsymbol{g}^{(k)T}\boldsymbol{g}^{(k)}}{\boldsymbol{g}^{(k)T}\boldsymbol{Q}\boldsymbol{g}^{(k)}} - \alpha_k\right)$$

  if $\boldsymbol{g}^{(k)} \neq \boldsymbol{0}$, and $\gamma_k = 1$ if $\boldsymbol{g}^{(k)} = \boldsymbol{0}$.

- Proof: By substitution and algebraic manipulations.

Remarks:

- γ_k is simply a (complicated) function of α_k.
- Note that $\gamma_k \leq 1$ always.
- $\gamma_k = 1$ implies that $V(\boldsymbol{x}^{k+1}) = 0$, which means $\boldsymbol{x}^{k+1} = \boldsymbol{x}^*$.
- The previous lemma has the following strong consequence.
- Theorem (8.1): Suppose $\gamma_k > 0$ for all k. Then, $\boldsymbol{x}^{(k)} \to \boldsymbol{x}^*$ for any initial condition $\boldsymbol{x}^{(0)}$ if and only if

$$\sum_{k=0}^{\infty} \gamma_k = \infty.$$

- To apply the theorem, we just check if our step size sequence $\{\alpha_k\}$ satisfies the above.

Proof of theorem:

- Note that $\boldsymbol{x}^{(k)} \to \boldsymbol{x}^*$ if and only if $V(\boldsymbol{x}^{(k)}) \to 0$.
- By previous lemma,

$$V(\boldsymbol{x}^{(k)}) = \left(\Pi_{i=0}^{k-1}(1-\gamma_i)\right) V(\boldsymbol{x}^{(0)}).$$

- Assume $\gamma_k < 1$ (otherwise, the result holds trivially).
- Hence,

$$\begin{aligned} \boldsymbol{x}^{(k)} \to \boldsymbol{x}^* \text{ for all } \boldsymbol{x}^{(0)} &\Leftrightarrow \Pi_{i=0}^{\infty}(1-\gamma_i) = 0 \\ &\Leftrightarrow \sum_{i=0}^{\infty} \gamma_i = \infty, \end{aligned}$$

  where the last line is obtained by taking logs.

Application of convergence theorem

- We can apply the previous theorem to answer the following questions (for quadratics):
  - Is the steepest descent algorithm globally convergent?
  - In a fixed step size gradient algorithm (i.e., $\alpha_k = \alpha$ for all k), for what values of the step size α is the algorithm globally convergent?

Convergence of steepest descent algorithm

- We now apply the previous theorem to show convergence of the steepest descent algorithm.
- Recall that in this case α_k is given by

$$\alpha_k = \frac{\boldsymbol{g}^{(k)T}\boldsymbol{g}^{(k)}}{\boldsymbol{g}^{(k)T}\boldsymbol{Q}\boldsymbol{g}^{(k)}}.$$

- Substituting into the formula for γ_k yields

$$\gamma_k = \frac{(\boldsymbol{g}^{(k)T}\boldsymbol{g}^{(k)})^2}{(\boldsymbol{g}^{(k)T}\boldsymbol{Q}\boldsymbol{g}^{(k)})(\boldsymbol{g}^{(k)T}\boldsymbol{Q}^{-1}\boldsymbol{g}^{(k)})}.$$

Rayleigh's inequality

- Given $\boldsymbol{Q} > 0$.
- Rayleigh's inequality:

$$\lambda_{min}(\boldsymbol{Q})\|\boldsymbol{x}\|^2 \leq \boldsymbol{x}^T\boldsymbol{Q}\boldsymbol{x} \leq \lambda_{max}(\boldsymbol{Q})\|\boldsymbol{x}\|^2,$$

  where

  $\lambda_{min}(\boldsymbol{Q})$ is the smallest eigenvalue of $\boldsymbol{Q}$

  $\lambda_{max}(\boldsymbol{Q})$ is the largest eigenvalue of $\boldsymbol{Q}$

- See p. 34.

- Note that $\lambda_{max}(\boldsymbol{Q}^{-1}) = 1/\lambda_{min}(\boldsymbol{Q})$.
- Applying Rayleigh's inequality to $\boldsymbol{Q}$ and $\boldsymbol{Q}^{-1}$, we obtain

$$\gamma_k \geq \frac{\lambda_{min}(\boldsymbol{Q})}{\lambda_{max}(\boldsymbol{Q})} > 0.$$

- Hence, $\gamma_k > 0$ for all k, and also

$$\sum_{k=0}^{\infty} \gamma_k = \infty.$$

- By previous theorem, $\boldsymbol{x}^{(k)} \to \boldsymbol{x}^*$ for all $\boldsymbol{x}^{(0)}$.
  [Theorem 8.2].

Convergence of fixed step size gradient algorithm

- Consider the case where we fix $\alpha_k = \alpha$ for all k.
- Theorem (8.3): $\boldsymbol{x}^{(k)} \to \boldsymbol{x}^*$ for any $\boldsymbol{x}^{(0)}$ if and only if

$$0 < \alpha < \frac{2}{\lambda_{max}(\boldsymbol{Q})}.$$

- The theorem gives some idea of how large the step size is allowed to be.

Proof of theorem:

- We have

$$\gamma_k = \alpha \frac{\boldsymbol{g}^{(k)T}\boldsymbol{Q}\boldsymbol{g}^{(k)}}{\boldsymbol{g}^{(k)T}\boldsymbol{Q}^{-1}\boldsymbol{g}^{(k)}} \left(2\frac{\boldsymbol{g}^{(k)T}\boldsymbol{g}^{(k)}}{\boldsymbol{g}^{(k)T}\boldsymbol{Q}\boldsymbol{g}^{(k)}} - \alpha \right).$$

- To prove $\Leftrightarrow$, we have two directions to prove:
  $\Rightarrow$ and $\Leftarrow$.

To prove $\Leftarrow$:

- Apply Rayleigh's inequality to get

$$\gamma_k \geq \alpha \left(\lambda_{min}(\boldsymbol{Q})\right)^2 \left(\frac{2}{\lambda_{max}(\boldsymbol{Q})} - \alpha \right) > 0.$$

- Hence, $\gamma_k > 0$ for all k, and also
$$\sum_{k=0}^{\infty} \gamma_k = \infty.$$

To prove $\Rightarrow$:

- Use contraposition.
- Suppose either $\alpha \leq 0$ or $\alpha \geq 2/\lambda_{max}(\boldsymbol{Q})$.
- It suffices to find a single $\boldsymbol{x}^{(0)}$ that makes the algorithm not converge.
- So, choose $\boldsymbol{x}^{(0)}$ such that $\boldsymbol{x}^{(0)} - \boldsymbol{x}^*$ is an eigenvector of $\boldsymbol{Q}$ corresponding to the eigenvalue $\lambda_{max} = \lambda_{max}(\boldsymbol{Q})$.
- Note that for all k, $\boldsymbol{x}^{(k)} - \boldsymbol{x}^*$ is an eigenvector of $\boldsymbol{Q}$ corresponding to λ_{max}.
- Moreover, for all k, $\boldsymbol{g}^{(k)}$ is an eigenvector of $\boldsymbol{Q}$ corresponding to λ_{max}.
- Hence, using the formula for γ_k, we get
$$\gamma_k = \alpha\lambda_{max}^2 \left(\frac{2}{\lambda_{max}} - \alpha\right) \leq 0.$$
- By previous lemma, $V(\boldsymbol{x}^{(k+1)}) \geq V(\boldsymbol{x}^{(k)})$, which implies that $\boldsymbol{x}^{(k)} \not\to \boldsymbol{x}^*$.

Example (8.3):

- Consider
$$f(\boldsymbol{x}) = \frac{1}{2}\boldsymbol{x}^T \begin{bmatrix} 8 & 2\sqrt{2} \\ 2\sqrt{2} & 10 \end{bmatrix} \boldsymbol{x} + \boldsymbol{x}^T \begin{bmatrix} 3 \\ 6 \end{bmatrix} + 24.$$
- Eigenvalues of $\boldsymbol{Q}$: 6 and 12.
- Fixed step size gradient algorithm converges if and only if $0 < \alpha < 2/12 = 1/6$.

Other insights into convergence

- Theorem (8.4): In the steepest descent algorithm,
$$V(\boldsymbol{x}^{(k+1)}) \leq \left(1 - \frac{1}{r}\right) V(\boldsymbol{x}^{(k)}),$$
where
$$r = \frac{\lambda_{max}(\boldsymbol{Q})}{\lambda_{min}(\boldsymbol{Q})}$$
(called the *condition number* of $\boldsymbol{Q}$).

- If r is small (close to 1), convergence is fast.
- If r is large, convergence is slow.
- See book for proof.

Order of convergence

- Given: a sequence $\{\boldsymbol{x}^{(k)}\}$ converging to $\boldsymbol{x}^*$.
- The *order of convergence* is p (where $1 \leq p < \infty$) if

$$0 < \lim_{k\to\infty} \frac{\|\boldsymbol{x}^{(k+1)} - \boldsymbol{x}^*\|}{\|\boldsymbol{x}^{(k)} - \boldsymbol{x}^*\|^p} < \infty.$$

  We say that the order of convergence is ∞ if for all $p \geq 1$,

$$\lim_{k\to\infty} \frac{\|\boldsymbol{x}^{(k+1)} - \boldsymbol{x}^*\|}{\|\boldsymbol{x}^{(k)} - \boldsymbol{x}^*\|^p} = 0.$$

- Order of convergence is one measure of "speed" of convergence.

Example:

- Given: $x^{(k)} = 1/k$.
- Thus, $x^{(k)} \to 0$. Then,

$$\frac{|x^{(k+1)}|}{|x^{(k)}|^p} = \frac{1/(k+1)}{1/k^p} = \frac{k^p}{k+1}.$$

- If $p > 1$, it grows to ∞.
- If $p = 1$, the sequence converges to 1.
- Hence, the order of convergence is 1.

Example:

- Given: $x^{(k)} = \gamma^k$, where $0 < \gamma < 1$.
- Thus, $x^{(k)} \to 0$. Then,

$$\frac{|x^{(k+1)}|}{|x^{(k)}|^p} = \frac{\gamma^{k+1}}{(\gamma^k)^p} = \gamma^{k+1-kp} = \gamma^{k(1-p)+1}.$$

- If $p > 1$, it grows to ∞.

- If $p = 1$, the sequence converges to γ (actually, it remains constant).
- Hence, the order of convergence is 1.

Example:

- Given: $x^{(k)} = \gamma^{(q^k)}$, where $q > 1$ and $0 < \gamma < 1$.
- Thus, $x^{(k)} \to 0$. Then,

$$\frac{|x^{(k+1)}|}{|x^{(k)}|^p} = \frac{\gamma^{(q^{k+1})}}{(\gamma^{(q^k)})^p} = \gamma^{(q^{k+1}-pq^k)} = \gamma^{(q-p)q^k}.$$

- If $p < q$, the above sequence converges to 0.
- If $p > q$, it grows to ∞.
- If $p = q$, the sequence converges to 1 (actually, it remains constant).
- Hence, the order of convergence is q.

Bounding the order of convergence

- $g(h) = O(h)$ means that there exists a constant c such that $|g(h)| \leq c|h|$ for sufficiently small h.
- Given $\boldsymbol{x}^{(k)} \to \boldsymbol{x}^*$. If
$$\|\boldsymbol{x}^{(k+1)} - \boldsymbol{x}^*\| = O(\|\boldsymbol{x}^{(k)} - \boldsymbol{x}^*\|^p),$$
then the order of convergence is at least p.
- Example: Order of convergence is at least 2 if
$$\|\boldsymbol{x}^{(k+1)} - \boldsymbol{x}^*\| = O(\|\boldsymbol{x}^{(k)} - \boldsymbol{x}^*\|^2).$$
- $g(h) = \Omega(h)$ means there exists a constant $c > 0$ such that $|g(h)| \geq c|h|$ for sufficiently small h.
- Given $\boldsymbol{x}^{(k)} \to \boldsymbol{x}^*$. If
$$\|\boldsymbol{x}^{(k+1)} - \boldsymbol{x}^*\| = \Omega(\|\boldsymbol{x}^{(k)} - \boldsymbol{x}^*\|^p),$$
then the order of convergence is at most p.

Order of convergence of steepest descent

Theorem: The steepest descent algorithm has order of convergence of 1 in the worst case.

Proof:

- Consider only quadratic case. Assume $\lambda_{max}(\boldsymbol{Q}) > \lambda_{min}(\boldsymbol{Q})$.
- Suffices to show that there exists $\boldsymbol{x}^{(0)}$ such that $\|\boldsymbol{x}^{(k+1)} - \boldsymbol{x}^*\| = \Omega(\|\boldsymbol{x}^{(k)} - \boldsymbol{x}^*\|)$; i.e.,

$$\|\boldsymbol{x}^{(k+1)} - \boldsymbol{x}^*\| \geq c\|\boldsymbol{x}^{(k)} - \boldsymbol{x}^*\|$$

  for some $c > 0$.
- By Rayleigh's inequality,

$$\begin{aligned} V(\boldsymbol{x}^{(k+1)}) &= \frac{1}{2}(\boldsymbol{x}^{(k+1)} - \boldsymbol{x}^*)^T\boldsymbol{Q}(\boldsymbol{x}^{(k+1)} - \boldsymbol{x}^*) \\ &\leq \frac{\lambda_{max}(\boldsymbol{Q})}{2}\|\boldsymbol{x}^{(k+1)} - \boldsymbol{x}^*\|^2. \end{aligned}$$

- Similarly,

$$V(\boldsymbol{x}^{(k)}) \geq \frac{\lambda_{min}(\boldsymbol{Q})}{2}\|\boldsymbol{x}^{(k)} - \boldsymbol{x}^*\|^2.$$

- Therefore, by Lemma 8.1, it suffices to show that $\gamma_k \leq d$ for some $d < 1$.
- Recall that for the steepest descent algorithm, γ_k depends on $\boldsymbol{g}^{(k)}$:

$$\gamma_k = \frac{(\boldsymbol{g}^{(k)T}\boldsymbol{g}^{(k)})^2}{(\boldsymbol{g}^{(k)T}\boldsymbol{Q}\boldsymbol{g}^{(k)})(\boldsymbol{g}^{(k)T}\boldsymbol{Q}^{-1}\boldsymbol{g}^{(k)})}.$$

- First consider the case where $n = 2$.
- Suppose $\boldsymbol{x}^{(0)}$ is chosen such that $\boldsymbol{g}^{(0)}$ is not an eigenvector of $\boldsymbol{Q}$. By Prop. 8.1, $\boldsymbol{g}^{(k)}$ is also not an eigenvector of $\boldsymbol{Q}$ for all k (because any two eigenvectors corresponding to $\lambda_{max}(\boldsymbol{Q})$ and $\lambda_{min}(\boldsymbol{Q})$ are mutually orthogonal).
- Also, $\boldsymbol{g}^{(k)}$ lies in one of 2 mutually orthogonal directions. Therefore, the value of γ_k is one of 2 numbers, both of which are < 1. This proves the $n = 2$ case.
- For the general n case, let $\boldsymbol{v}_1$ and $\boldsymbol{v}_2$ be mutually orthogonal eigenvectors corresponding to $\lambda_{max}(\boldsymbol{Q})$ and $\lambda_{min}(\boldsymbol{Q})$. Choose $\boldsymbol{x}^{(0)}$ such that $\boldsymbol{g}^{(0)}$ lies in the span of $\boldsymbol{v}_1$ and $\boldsymbol{v}_2$ but is not equal to either. Then, $\boldsymbol{g}^{(k)}$ lies in the span of $\boldsymbol{v}_1$ and $\boldsymbol{v}_2$ for all k. To see this, note that $\boldsymbol{g}^{(k+1)} = (\boldsymbol{I} - \alpha_k\boldsymbol{Q})\boldsymbol{g}^{(k)}$, and the span of $\boldsymbol{v}_1$ and $\boldsymbol{v}_2$ is invariant under operations of $\boldsymbol{I} - \alpha_k\boldsymbol{Q}$. We can now proceed as in the $n = 2$ case.