1 Suppose $Q \succ 0$ is of $n \times n$ size. Given a set of linearly independent vectors $\{\mathbf{p}^0, \dots, \mathbf{p}^{n-1}\}$ in $\mathbb{R}^n$, perform the Gram-Schmidt procedure as follows:

$$\mathbf{d}^0 = \mathbf{p}^0,$$
$$\mathbf{d}^{k+1} = \mathbf{p}^{k+1} - \sum_{i=0}^{k} \frac{(\mathbf{p}^{k+1})^\top Q \mathbf{d}^i}{(\mathbf{d}^i)^\top Q \mathbf{d}^i} \mathbf{d}^i, \quad \forall k = 0, \dots, n-2$$

Show that $\mathbf{d}^0, \dots, \mathbf{d}^{n-1}$ are Q-conjugate.

Solution We prove this by induction.

Base step:

Let $\mathbf{d}^0$ and $\mathbf{d}^1$ be vectors generated from the Gram-Schmidt process. Then

$$\begin{aligned}
(\mathbf{d}^0)^\top Q \mathbf{d}^1 &= (\mathbf{p}^0)^\top Q \left(\mathbf{p}^1 - \frac{(\mathbf{p}^1)^\top Q \mathbf{d}^0}{(\mathbf{d}^0)^\top Q \mathbf{d}^0} \mathbf{d}^0 \right) \\
&= (\mathbf{p}^0)^\top Q \left(\mathbf{p}^1 - \frac{(\mathbf{p}^1)^\top Q \mathbf{p}^0}{(\mathbf{p}^0)^\top Q \mathbf{p}^0} \mathbf{p}^0 \right) \\
&= (\mathbf{p}^0)^\top Q \mathbf{p}^1 - \frac{(\mathbf{p}^1)^\top Q \mathbf{p}^0}{(\mathbf{p}^0)^\top Q \mathbf{p}^0} (\mathbf{p}^0)^\top Q \mathbf{p}^0 \\
&= (\mathbf{p}^0)^\top Q \mathbf{p}^1 - (\mathbf{p}^1)^\top Q \mathbf{p}^0 \\
&= 0
\end{aligned}$$

The last step holds since Q is symmetric and each term is a scalar, so taking the transpose of a scalar gives us the same number.

Inductive step:

Assume that we have $\mathbf{d}^0, \dots, \mathbf{d}^k$ as generated from the Gram-Schmidt procedure and are Q-conjugate. We wish to show that $\mathbf{d}^1, \dots, \mathbf{d}^k, \mathbf{d}^{k+1}$ are also Q-conjugate.

Fix $\mathbf{d}^j$, where $j < k+1$. Then

$$\begin{aligned}
(\mathbf{d}^j)^\top Q \mathbf{d}^{k+1} &= (\mathbf{d}^j)^\top Q \left(\mathbf{p}^{k+1} - \sum_{i=0}^{k} \frac{(\mathbf{p}^{k+1})^\top Q \mathbf{d}^i}{(\mathbf{d}^i)^\top Q \mathbf{d}^i} \mathbf{d}^i \right) \\
&= (\mathbf{d}^j)^\top Q \mathbf{p}^{k+1} - \sum_{i=0}^{k} \frac{(\mathbf{p}^{k+1})^\top Q \mathbf{d}^i}{(\mathbf{d}^i)^\top Q \mathbf{d}^i} (\mathbf{d}^j)^\top Q \mathbf{d}^i \\
&= (\mathbf{d}^j)^\top Q \mathbf{p}^{k+1} - \frac{(\mathbf{p}^{k+1})^\top Q \mathbf{d}^j}{(\mathbf{d}^j)^\top Q \mathbf{d}^j} (\mathbf{d}^j)^\top Q \mathbf{d}^j \\
&= (\mathbf{d}^j)^\top Q \mathbf{p}^{k+1} - (\mathbf{p}^{k+1})^\top Q \mathbf{d}^j \\
&= 0
\end{aligned}$$

Thus, $\mathbf{d}^1, \dots, \mathbf{d}^k, \mathbf{d}^{k+1}$ are Q-conjugate.

By induction, the vectors generated by the Gram-Schmidt process are Q-conjugate.

---

2 Given $f \colon \mathbb{R}^n \to \mathbb{R}$, $f \in C^1$, consider the algorithm

$$\mathbf{x}^{k+1} = \mathbf{x}^k + \alpha_k \mathbf{d}^k,$$

where $\mathbf{d}^1, \mathbf{d}^2, \dots$ are vectors in $\mathbb{R}^n$, and $\alpha_k \geq 0$ is chosen to minimize $f(\mathbf{x}^k + \alpha \mathbf{d}^k)$, that is,

$$\alpha_k = \arg\min_{\alpha \geq 0} f(\mathbf{x}^k + \alpha \mathbf{d}^k).$$

Note that the general algorithm above encompasses almost all algorithms that we discussed in this part, including the steepest descent, Newton, conjugate gradient, and quasi-Newton algorithms.

Let $\mathbf{g}^k = \nabla f(\mathbf{x^k})$, and assume that $(\mathbf{d}^k)^\top \mathbf{g}^k < 0$.

a. Show that $\mathbf{d}^k$ is a descent direction for f in the sense that there exists $\bar{\alpha} > 0$ such that for all $\alpha \in (0, \bar{\alpha}]$, $f(\mathbf{x}^k + \alpha \mathbf{d}^k) < f(\mathbf{x}^k)$.

b. Show that $\alpha_k > 0$.

c. Show that $(\mathbf{d}^k)^\top \mathbf{g}^{k+1} = 0$.

d. Show that the following algorithms all satisfy the condition $(\mathbf{d}^k)^\top \mathbf{g}^k < 0$, if $\mathbf{g}^k \neq 0$:
   1. Steepest descent algorithm.
   2. Newton's method, assuming that the Hessian is positive definite.
   3. Conjugate gradient algorithm.
   4. Quasi-Newton algorithm, assuming that $H_k \succ 0$.

e. For the case where $f(x) = \frac{1}{2}\mathbf{x}^\top Q\mathbf{x} - \mathbf{x}^\top \mathbf{b}$, with $Q = Q^T \succ 0$, derive an expression for α_k in terms of Q, $\mathbf{d}^k$, and $\mathbf{g}^k$.

Solution a. Consider the Taylor expansion of f about $\mathbf{x}^k$:

$$f(\mathbf{x}^k + \alpha \mathbf{d}^k) = f(\mathbf{x}^k) + (\mathbf{g}^k)^\top (\alpha \mathbf{d}^k) + O(\alpha^2)$$

For small α, the first order term $\alpha (\mathbf{g}^k)^\top \mathbf{d}^k < 0$ will have a larger magnitude than higher order terms. Hence, there exists $\bar{\alpha} > 0$ such that if $\alpha \in (0, \bar{\alpha}]$,

$$f(\mathbf{x}^k + \alpha \mathbf{d}^k) < f(\mathbf{x}^k)$$

as desired.

b. If $\alpha_k = 0$, the $\mathbf{x}^k$ will be our minimizer, meaning $\mathbf{g}^k = 0$. But that would imply that $(\mathbf{d}^k)^\top \mathbf{g}^k = 0$, which is a contradiction. Hence, $\alpha_k > 0$.

c. By definition, α_k minimizes $\varphi(\alpha) := f(\mathbf{x}^k + \alpha \mathbf{d}^k)$. Hence,

$$\varphi'(\alpha_k) = (\mathbf{d}^k)^\top \nabla f(\mathbf{x}^k + \alpha_k \mathbf{d}^k) = 0 \implies (\mathbf{d}^k)^\top \mathbf{g}^{k+1} = 0$$

as desired.

d. 1. This is obviously true since the algorithm is the same as the above.

2. For Newton's method, $\mathbf{d}^k = -H_k^{-1}\mathbf{g}^k$. Then

$$(\mathbf{d}^k)^\top \mathbf{g}^k = -(\mathbf{g}^k)^\top H_k^{-1} \mathbf{g}^k < 0$$

since H_k is positive definite $\implies H_k^{-1}$ is positive definite.

3. For the conjugate gradient algorithm, $\mathbf{d}^k = -\mathbf{g}^k + \beta_{k-1}\mathbf{d}^{k-1}$, $\mathbf{d}^0 = -\mathbf{g}^0$. Then

$$\begin{aligned}
(\mathbf{d}^{k+1})^\top \mathbf{g}^{k+1} &= \left(-\mathbf{g}^{k+1} + \frac{(\mathbf{g}^{k+1})^\top Q \mathbf{d}^k}{(\mathbf{d}^k)^\top Q \mathbf{d}^k} \mathbf{d}^k\right)^\top \mathbf{g}^{k+1} \\
&= -\|\mathbf{g}^{k+1}\|^2 + \frac{(\mathbf{g}^{k+1})^\top Q \mathbf{d}^k}{(\mathbf{d}^k)^\top Q \mathbf{d}^k} (\mathbf{d}^k)^\top \mathbf{g}^{k+1} \\
&= -\|\mathbf{g}^{k+1}\|^2 < 0 \quad \text{(by part (c))}
\end{aligned}$$

4. In a Quasi-Newton algorithm, we have $\mathbf{d}^k = -H_k \mathbf{g}^k$. Then

$$(\mathbf{d}^k)^\top \mathbf{g}^k = -(\mathbf{g}^k)^\top H_k \mathbf{g}^k < 0$$

as desired.

e. We wish to find $\alpha_k = \arg\min_{\alpha \geq 0} f(\mathbf{x}^k + \alpha \mathbf{d}^k)$. Define $\varphi(\alpha) = f(\mathbf{x}^k + \alpha \mathbf{d}^k)$. Then

$$\begin{aligned}
\varphi'(\alpha) &= (\mathbf{d}^k)^\top \nabla f(\mathbf{x}^k + \alpha \mathbf{d}^k) \\
&= (\mathbf{d}^k)^\top (Q(\mathbf{x}^k + \alpha \mathbf{d}^k) - \mathbf{b}) \\
&= (\mathbf{d}^k)^\top (\mathbf{g}^k + Q\alpha \mathbf{d}^k) = 0 \\
\implies \alpha_k &= -\frac{(\mathbf{d}^k)^\top \mathbf{g}^k}{(\mathbf{d}^k)^\top Q \mathbf{d}^k}
\end{aligned}$$

This is the only FONC point, so it must be our minimizer.

---

3 Let $Q \succ 0$. Suppose the nonzero vectors $\mathbf{d}^0, \ldots, \mathbf{d}^{n-1}$ are Q-conjugate. Show that they are linearly independent.

Solution Consider

$$\sum_{i=0}^{n-1} c_i \mathbf{d}^i = 0.$$

We wish to show that $c_i = 0$ for all i. Left-multiply by $(\mathbf{d}^k)^\top Q$ to get

$$\sum_{i=0}^{n-1} c_i (\mathbf{d}^k)^\top Q \mathbf{d}^i = c_k (\mathbf{d}^k)^\top Q \mathbf{d}^k = 0$$

Q is positive definite, so we must have that $c_k = 0$ for the equation to hold. Hence, $c_k = 0$ for all $0 \leq k \leq n-1$.

---

4 Consider the algorithm

$$\mathbf{x}^{k+1} = \mathbf{x}^k - \alpha_k P_k \nabla f(\mathbf{x}^k),$$

where $f\colon \mathbb{R}^2 \to \mathbb{R}$, $P_k = \begin{pmatrix} 1 & 0 \\ 0 & a \end{pmatrix}$ with $a \in \mathbb{R}$, and

$$\alpha_k = \arg\min_{\alpha \geq 0} f(\mathbf{x}^k - \alpha P_k \nabla f(\mathbf{x}^k)).$$

Suppose that at some iteration k, we have $\nabla f(\mathbf{x}^k) = (1, 2)^\top$. Find the largest range of values of a that guarantees that $\alpha_k > 0$ for any f.

Solution We argue by contrapositive.

Consider $\varphi(\alpha) = f(\mathbf{x}^k - \alpha P_k \nabla f(\mathbf{x}^k))$. Then

$$\begin{aligned}
\varphi'(0) = -(P_k \mathbf{g}^k)^\top \mathbf{g}^k &= 0 \\
\implies -(1 + 4a) &= 0 \\
\implies a &= -\frac{1}{4}
\end{aligned}$$

If $a > 0$, then $P_k \succ 0$, which guarantees descent. Hence, we must have $a > -\frac{1}{4}$ in order to have $\alpha_k > 0$.