7 Projection methods for linear equality constrained problems

7.1 Optimization over linear equality constraints

Suppose we want to solve

\[(P) \quad \min f(x) \quad \text{s.t.} \quad Ax = b, \quad x \in X,\]

where \(X\) is an open set. Assume that the problem is feasible. Then, without loss of generality matrix \(A\) has full row rank.

The KKT conditions are necessary for this problem and are as follows (we will drop the explicit requirement \(x \in X\) for brevity):

\[
\begin{align*}
A\bar{x} &= b \\
\nabla f(\bar{x}) + A^T \bar{\pi} &= 0.
\end{align*}
\]

We therefore wish to find such a KKT point \((\bar{x}, \bar{\pi})\).

Suppose we are at an iterate \(x\) where \(Ax = b\), i.e., \(x\) is a feasible point. Just like in the steepest descent algorithm, we wish to find a direction \(d\) which is a direction of steepest descent of the objective function, but in order to stay in the feasible region, we also need to have \(Ad = 0\). Therefore, the direction-finding problem takes form

\[
\min \nabla f(x)^T d \quad \text{s.t.} \quad d^T Id \leq 1, \quad Ad = 0.
\]

The first constraint of the problem requires that the search direction \(d\) has length 1 in the Euclidean norm. We can, however, adapt a more generalized approach and replace the Euclidean norm \(\|d\| = \sqrt{d^T Id} = \sqrt{d^T d}\) with a more general norm \(\|d\|_Q = \sqrt{d^T Qd}\), where \(Q\) is an arbitrary symmetric p.d. matrix. Using this general norm in the direction-finding problem, we can state the projected steepest descent algorithm:

**Step 0** Given \(x^0\), set \(k \leftarrow 0\)

**Step 1** Solve the direction-finding problem defined at point \(x^k\):

\[
\text{DFP}_{x^k} \quad d^k = \arg\min \nabla f(x^k)^T d \quad \text{s.t.} \quad d^T Id \leq 1, \quad Ad = 0.
\]

If \(\nabla f(x^k)^T d^k = 0\), stop. \(x^k\) is a KKT point.

**Step 2** Choose stepsize \(\alpha^k\) by performing an exact (or inexact) line search.

**Step 3** Set \(x^{k+1} \leftarrow x^k + \alpha^k d^k\), \(k \leftarrow k + 1\). Go to **Step 1**.

Notice that if \(Q = I\) and the equality constraints are absent, the above is just the steepest descent algorithm. The choice of name “projected” steepest descent may not be apparent at this point, but will be clarified later.
7.2 Analysis of (DFP)

Notice that the search direction at each iteration is the solution of a nonlinearly constrained optimization problem (DFP,\(x_k\)) above. Note that (DFP,\(x_k\)) is a convex program, and \(d = 0\) is a Slater point. Therefore, the KKT conditions are necessary and sufficient for optimality. These conditions are (we omit the superscript \(k\) for simplicity):

\[
\begin{align*}
Ad &= 0 \\
d^T Qd &\leq 1 \\
\nabla f(x) + A^T \pi + 2\beta Qd &= 0 \\
\beta &\geq 0 \\
\beta (1 - d^T Qd) &= 0.
\end{align*}
\]

(4)

Let \(d\) solve these equations with multipliers \(\beta\) and \(\pi\).

**Proposition 7.1** \(x\) is a KKT point of \((P)\) if and only if \(\nabla f(x)^T d = 0\), where \(d\) is the KKT point of \((DFP_x)\).

**Proof:** First, suppose \(x\) is a KKT point of \((P)\). Then there exists \(v\) such that

\[Ax = b, \quad \nabla f(x) + A^T v = 0.\]

Let \(d\) be any KKT point of \((DFP_x)\) together with multipliers \(\pi\) and \(\beta\). Then, in particular, \(Ad = 0\). Therefore,

\[\nabla f(x)^T d = -(A^T v)^T d = v^T Ad = 0.\]

To prove the converse, suppose that \(d\) (together with multipliers \(\pi\) and \(\beta\)) is a KKT point of \((DFP_x)\), and \(\nabla f(x)^T d = 0\). Then

\[0 = (\nabla f(x) + A^T \pi + 2\beta Qd)^T d = \nabla f(x)^T d + \pi^T Ad + 2\beta d^T Qd = 2\beta,\]

and so \(\beta = 0\). Therefore,

\[Ax = b \text{ and } \nabla f(x) + A^T \pi = 0,\]

i.e., the point \(x\) (together with multiplier vector \(\pi\)) is a KKT point of \((P)\).  

**Proposition 7.2** \(x\) is a KKT point of \((P)\) if and only if \(\beta = 0\), where \(\beta\) is the appropriate KKT multiplier of \((DFP_x)\).

**Proposition 7.3** If \(x\) is not a KKT point of \((P)\), then \(d\) is a descent direction, where \(d\) is the KKT point of \((DFP_x)\).

**Proposition 7.4** The projected steepest descent algorithm has the same convergence properties and the same linear convergence as the steepest descent algorithm. Under the same conditions as in the steepest descent algorithm, the iterates converge to a KKT point of \((P)\), and the convergence rate is linear, with a convergence constant that is bounded in terms of eigenvalues identically as in the steepest descent algorithm.

7.3 Solving (DFP,\(x\))

Approach 1 to solving DFP: solving linear equations
Create the system of linear equations:

\[ Q \tilde{d} + A^T \tilde{\pi} = -\nabla f(x) \]
\[ A \tilde{d} = 0 \]

(5)

and solve this linear system for \((\tilde{d}, \tilde{\pi})\) by any method at your disposal.

If \(Q\tilde{d} = 0\), then \(\nabla f(x) + A^T \tilde{\pi} = 0\) and so \(x\) is a KKT point of \((P)\).

If \(Q\tilde{d} \neq 0\), then rescale the solution as follows:

\[ d = \frac{\tilde{d}}{\sqrt{\tilde{d}^T Q \tilde{d}}}, \]
\[ \pi = \tilde{\pi}, \]
\[ \beta = \frac{1}{2 \sqrt{\tilde{d}^T Q \tilde{d}}}. \]

Proposition 7.5 \((d, \pi, \beta)\) defined above satisfy \((4)\).

Note that the rescaling step is not necessary in practice, since we use a line-search.

**Approach 2 to solving DFP: Formulas**

Let

\[ P_Q = Q^{-1} - Q^{-1} A^T (AQ^{-1} A^T)^{-1} A Q^{-1} \]
\[ \beta = \frac{\sqrt{(\nabla f(x))^T P_Q (\nabla f(x))}}{2} \]
\[ \pi = -(AQ^{-1} A^T)^{-1} A Q^{-1} (\nabla f(x)) \]

If \(\beta > 0\), let

\[ d = \frac{-P_Q \nabla f(x)}{\sqrt{\nabla f(x)^T P_Q \nabla f(x)}}. \]

If \(\beta = 0\), let \(\tilde{d} = 0\).

Proposition 7.6 \(P_Q\) is symmetric and positive semi-definite. Hence \(\beta \geq 0\).

Proposition 7.7 \((d, \pi, \beta)\) defined above satisfy \((4)\).

With many efficient implementations of procedures for solving systems on linear equations, Approach 1 is the one typically implemented in practice, while formulas of Approach 2 can be utilized in analysis.

7.4 The Variable Metric Method

In the projected steepest descent algorithm, the direction \(d\) must be contained in the ellipsoid \(E_Q = \{d \in \mathbb{R}^n : d^T Q d \leq 1\}\), where \(Q\) is fixed for all iterations. In a variable metric method, \(Q\) can vary at every iteration. The variable metric algorithm is:

**Step 0** Given \(x^0\), set \(k \leftarrow 0\)
Step 1  Choose a p.d. symmetric matrix $Q$. (Perhaps $Q = Q(\bar{x})$, i.e., the choice of $Q$ may depend on the current point.) Solve the direction-finding problem defined at point $x^k$:

$$(DFP_{x^k}) \quad d^k = \arg\min_d \nabla f(x^k)^T d \quad \text{s.t.} \quad d^T Qd \leq 1$$

$$Ad = 0.$$ 

If $\nabla f(x^k)^T d^k = 0$, stop. $x^k$ is a KKT point.

Step 2  Choose stepsize $\alpha^k$ by performing an exact (or inexact) line search.

Step 3  Set $x^{k+1} \leftarrow x^k + \alpha^k d^k$, $k \leftarrow k + 1$. Go to Step 1.

All properties of the projected steepest descent algorithm still hold here.

Some strategies for choosing $Q$ at each iteration are:

- $Q = I$
- $Q$ is a given matrix held constant over all iterations
- $Q = \nabla^2 f(x^k)$ where $\nabla^2 f(x)$ is the Hessian of $f(x)$. It is easy to show that in this case, the variable metric algorithm is equivalent to Newton’s method with a line-search, see the proposition below.
- $Q = \nabla^2 f(x^k) + \delta I$, where $\delta$ is chosen to be large for early iterations, but $\delta$ is chosen to be small for later iterations. One can think of this strategy as approximating the projected steepest descent algorithm in early iterations, followed by approximating Newton’s method in later iterations.

**Proposition 7.8** Suppose that $Q = \nabla^2 f(x^k)$ in the variable metric algorithm. Then the direction $\bar{d}$ in the variable metric method is a positive scalar times the Newton direction.

**Proof:** If $Q = \nabla^2 f(x^k)$, then the vector $\bar{d}$ of the variable metric method is the optimal solution of $(DFP_{x^k})$:

$$\bar{d} = \arg\min_d \nabla f(x^k)^T d \quad \text{s.t.} \quad Ad = 0$$

$$d^T \nabla^2 f(x^k)d \leq 1.$$ 

The Newton direction $\hat{d}$ for (P) at the point $\bar{x}$ is the solution of the following problem:

$$(NDP_{x^k}) : \quad \hat{d} = \arg\min_d \nabla f(x^k)^T d + \frac{1}{2}d^T \nabla^2 f(x^k)d \quad \text{s.t.} \quad Ad = 0.$$ 

(6)

If you write down the KKT conditions for each of these two problems, you then can easily verify that $d = \gamma \hat{d}$ for some scalar $\gamma > 0$. ■