Chapter 10

Equality constrained minimization

10.1 Equality constrained minimization problems

In this chapter we describe methods for solving a convex optimization problem with equality constraints,

$$\begin{array}{ll} \text{minimize} & f(x) \\ \text{subject to} & Ax = b, \end{array} \tag{10.1}$$

where $f : \mathbf{R}^n \to \mathbf{R}$ is convex and twice continuously differentiable, and $A \in \mathbf{R}^{p\times n}$ with $\mathbf{rank}\, A = p < n$. The assumptions on A mean that there are fewer equality constraints than variables, and that the equality constraints are independent. We will assume that an optimal solution $x^\star$ exists, and use $p^\star$ to denote the optimal value, $p^\star = \inf\{f(x) \mid Ax = b\} = f(x^\star)$.

Recall (from §4.2.3 or §5.5.3) that a point $x^\star \in \mathbf{dom}\, f$ is optimal for (10.1) if and only if there is a $\nu^\star \in \mathbf{R}^p$ such that

$$Ax^\star = b, \qquad \nabla f(x^\star) + A^T\nu^\star = 0. \tag{10.2}$$

Solving the equality constrained optimization problem (10.1) is therefore equivalent to finding a solution of the KKT equations (10.2), which is a set of $n+p$ equations in the $n+p$ variables $x^\star$, $\nu^\star$. The first set of equations, $Ax^\star = b$, are called the *primal feasibility equations*, which are linear. The second set of equations, $\nabla f(x^\star) + A^T\nu^\star = 0$, are called the *dual feasibility equations*, and are in general nonlinear. As with unconstrained optimization, there are a few problems for which we can solve these optimality conditions analytically. The most important special case is when f is quadratic, which we examine in §10.1.1.

Any equality constrained minimization problem can be reduced to an equivalent unconstrained problem by eliminating the equality constraints, after which the methods of chapter 9 can be used to solve the problem. Another approach is to solve the dual problem (assuming the dual function is twice differentiable) using an unconstrained minimization method, and then recover the solution of the

equality constrained problem (10.1) from the dual solution. The elimination and dual methods are briefly discussed in §10.1.2 and §10.1.3, respectively.

The bulk of this chapter is devoted to extensions of Newton's method that directly handle equality constraints. In many cases these methods are preferable to methods that reduce an equality constrained problem to an unconstrained one. One reason is that problem structure, such as sparsity, is often destroyed by elimination (or forming the dual); in contrast, a method that directly handles equality constraints can exploit the problem structure. Another reason is conceptual: methods that directly handle equality constraints can be thought of as methods for directly solving the optimality conditions (10.2).

10.1.1 Equality constrained convex quadratic minimization

Consider the equality constrained convex quadratic minimization problem

$$\begin{array}{ll} \text{minimize} & (1/2)x^T Px + q^T x + r \\ \text{subject to} & Ax = b, \end{array} \tag{10.3}$$

where $P \in \mathbf{S}_+^n$ and $A \in \mathbf{R}^{p\times n}$. This problem is important on its own, and also because it forms the basis for an extension of Newton's method to equality constrained problems.

Here the optimality conditions (10.2) are

$$Ax^\star = b, \qquad Px^\star + q + A^T\nu^\star = 0,$$

which we can write as

$$\left[\begin{array}{cc} P & A^T \\ A & 0 \end{array}\right]\left[\begin{array}{c} x^\star \\ \nu^\star \end{array}\right] = \left[\begin{array}{c} -q \\ b \end{array}\right]. \tag{10.4}$$

This set of $n+p$ linear equations in the $n+p$ variables $x^\star$, $\nu^\star$ is called the *KKT system* for the equality constrained quadratic optimization problem (10.3). The coefficient matrix is called the *KKT matrix*.

When the KKT matrix is nonsingular, there is a unique optimal primal-dual pair $(x^\star, \nu^\star)$. If the KKT matrix is singular, but the KKT system is solvable, any solution yields an optimal pair $(x^\star, \nu^\star)$. If the KKT system is not solvable, the quadratic optimization problem is unbounded below or infeasible. Indeed, in this case there exist $v \in \mathbf{R}^n$ and $w \in \mathbf{R}^p$ such that

$$Pv + A^T w = 0, \qquad Av = 0, \qquad -q^T v + b^T w > 0.$$

Let $\hat{x}$ be any feasible point. The point $x = \hat{x} + tv$ is feasible for all t and

$$\begin{aligned} f(\hat{x}+tv) &= f(\hat{x}) + t(v^T P\hat{x} + q^T v) + (1/2)t^2 v^T Pv \\ &= f(\hat{x}) + t(-\hat{x}^T A^T w + q^T v) - (1/2)t^2 w^T Av \\ &= f(\hat{x}) + t(-b^T w + q^T v), \end{aligned}$$

which decreases without bound as $t \to \infty$.

Nonsingularity of the KKT matrix

Recall our assumption that $P \in \mathbf{S}_+^n$ and $\mathbf{rank}\, A = p < n$. There are several conditions equivalent to nonsingularity of the KKT matrix:

- $\mathcal{N}(P) \cap \mathcal{N}(A) = \{0\}$, *i.e.*, P and A have no nontrivial common nullspace.
- $Ax = 0$, $x \neq 0 \implies x^T P x > 0$, *i.e.*, P is positive definite on the nullspace of A.
- $F^T P F \succ 0$, where $F \in \mathbf{R}^{n \times (n-p)}$ is a matrix for which $\mathcal{R}(F) = \mathcal{N}(A)$.

(See exercise 10.1.) As an important special case, we note that if $P \succ 0$, the KKT matrix must be nonsingular.

10.1.2 Eliminating equality constraints

One general approach to solving the equality constrained problem (10.1) is to eliminate the equality constraints, as described in §4.2.4, and then solve the resulting unconstrained problem using methods for unconstrained minimization. We first find a matrix $F \in \mathbf{R}^{n \times (n-p)}$ and vector $\hat{x} \in \mathbf{R}^n$ that parametrize the (affine) feasible set:

$$\{x \mid Ax = b\} = \{Fz + \hat{x} \mid z \in \mathbf{R}^{n-p}\}.$$

Here $\hat{x}$ can be chosen as any particular solution of $Ax = b$, and $F \in \mathbf{R}^{n \times (n-p)}$ is any matrix whose range is the nullspace of A. We then form the reduced or eliminated optimization problem

$$\text{minimize} \quad \tilde{f}(z) = f(Fz + \hat{x}), \tag{10.5}$$

which is an unconstrained problem with variable $z \in \mathbf{R}^{n-p}$. From its solution $z^\star$, we can find the solution of the equality constrained problem as $x^\star = Fz^\star + \hat{x}$.

We can also construct an optimal dual variable $\nu^\star$ for the equality constrained problem, as

$$\nu^\star = -(AA^T)^{-1} A \nabla f(x^\star).$$

To show that this expression is correct, we must verify that the dual feasibility condition

$$\nabla f(x^\star) + A^T(-(AA^T)^{-1} A \nabla f(x^\star)) = 0 \tag{10.6}$$

holds. To show this, we note that

$$\begin{bmatrix} F^T \\ A \end{bmatrix} \left(\nabla f(x^\star) - A^T (AA^T)^{-1} A \nabla f(x^\star)\right) = 0,$$

where in the top block we use $F^T \nabla f(x^\star) = \nabla \tilde{f}(z^\star) = 0$ and $AF = 0$. Since the matrix on the left is nonsingular, this implies (10.6).

Example 10.1 *Optimal allocation with resource constraint.* We consider the problem

$$\begin{array}{ll} \text{minimize} & \sum_{i=1}^n f_i(x_i) \\ \text{subject to} & \sum_{i=1}^n x_i = b, \end{array}$$

where the functions $f_i : \mathbf{R} \to \mathbf{R}$ are convex and twice differentiable, and $b \in \mathbf{R}$ is a problem parameter. We interpret this as the problem of optimally allocating a single resource, with a fixed total amount b (the *budget*) to n otherwise independent activities.

We can eliminate x_n (for example) using the parametrization

$$x_n = b - x_1 - \cdots - x_{n-1},$$

which corresponds to the choices

$$\hat{x} = be_n, \qquad F = \begin{bmatrix} I \\ -\mathbf{1}^T \end{bmatrix} \in \mathbf{R}^{n\times(n-1)}.$$

The reduced problem is then

$$\text{minimize} \quad f_n(b - x_1 - \cdots - x_{n-1}) + \textstyle\sum_{i=1}^{n-1} f_i(x_i),$$

with variables $x_1, \ldots, x_{n-1}$.

Choice of elimination matrix

There are, of course, many possible choices for the elimination matrix F, which can be chosen as any matrix in $\mathbf{R}^{n\times(n-p)}$ with $\mathcal{R}(F) = \mathcal{N}(A)$. If F is one such matrix, and $T \in \mathbf{R}^{(n-p)\times(n-p)}$ is nonsingular, then $\tilde{F} = FT$ is also a suitable elimination matrix, since

$$\mathcal{R}(\tilde{F}) = \mathcal{R}(F) = \mathcal{N}(A).$$

Conversely, if F and $\tilde{F}$ are any two suitable elimination matrices, then there is some nonsingular T such that $\tilde{F} = FT$.

If we eliminate the equality constraints using F, we solve the unconstrained problem

$$\text{minimize} \quad f(Fz + \hat{x}),$$

while if $\tilde{F}$ is used, we solve the unconstrained problem

$$\text{minimize} \quad f(\tilde{F}\tilde{z} + \hat{x}) = f(F(T\tilde{z}) + \hat{x}).$$

This problem is equivalent to the one above, and is simply obtained by the change of coordinates $z = T\tilde{z}$. In other words, changing the elimination matrix can be thought of as changing variables in the reduced problem.

10.1.3 Solving equality constrained problems via the dual

Another approach to solving (10.1) is to solve the dual, and then recover the optimal primal variable $x^\star$, as described in §5.5.5. The dual function of (10.1) is

$$\begin{aligned} g(\nu) &= -b^T\nu + \inf_x (f(x) + \nu^T Ax) \\ &= -b^T\nu - \sup_x \left((-A^T\nu)^T x - f(x)\right) \\ &= -b^T\nu - f^*(-A^T\nu), \end{aligned}$$

where f^* is the conjugate of f, so the dual problem is

$$\text{maximize} \quad -b^T\nu - f^*(-A^T\nu).$$

Since by assumption there is an optimal point, the problem is strictly feasible, so Slater's condition holds. Therefore strong duality holds, and the dual optimum is attained, *i.e.*, there exists a $\nu^\star$ with $g(\nu^\star) = p^\star$.

If the dual function g is twice differentiable, then the methods for unconstrained minimization described in chapter 9 can be used to maximize g. (In general, the dual function g need not be twice differentiable, even if f is.) Once we find an optimal dual variable $\nu^\star$, we reconstruct an optimal primal solution $x^\star$ from it. (This is not always straightforward; see §5.5.5.)

Example 10.2 *Equality constrained analytic center.* We consider the problem

$$\begin{array}{ll} \text{minimize} & f(x) = -\sum_{i=1}^n \log x_i \\ \text{subject to} & Ax = b, \end{array} \tag{10.7}$$

where $A \in \mathbf{R}^{p\times n}$, with implicit constraint $x \succ 0$. Using

$$f^*(y) = \sum_{i=1}^n (-1 - \log(-y_i)) = -n - \sum_{i=1}^n \log(-y_i)$$

(with $\mathbf{dom}\, f^* = -\mathbf{R}^n_{++}$), the dual problem is

$$\text{maximize} \quad g(\nu) = -b^T\nu + n + \textstyle\sum_{i=1}^n \log(A^T\nu)_i, \tag{10.8}$$

with implicit constraint $A^T\nu \succ 0$. Here we can easily solve the dual feasibility equation, *i.e.*, find the x that minimizes $L(x,\nu)$:

$$\nabla f(x) + A^T\nu = -(1/x_1, \ldots, 1/x_n) + A^T\nu = 0,$$

and so

$$x_i(\nu) = 1/(A^T\nu)_i. \tag{10.9}$$

To solve the equality constrained analytic centering problem (10.7), we solve the (unconstrained) dual problem (10.8), and then recover the optimal solution of (10.7) via (10.9).

10.2 Newton's method with equality constraints

In this section we describe an extension of Newton's method to include equality constraints. The method is almost the same as Newton's method without constraints, except for two differences: The initial point must be feasible (*i.e.*, satisfy $x \in \mathbf{dom}\, f$ and $Ax = b$), and the definition of Newton step is modified to take the equality constraints into account. In particular, we make sure that the Newton step $\Delta x_{\rm nt}$ is a feasible direction, *i.e.*, $A\Delta x_{\rm nt} = 0$.

10.2.1 The Newton step

Definition via second-order approximation

To derive the Newton step $\Delta x_{\rm nt}$ for the equality constrained problem

$$\begin{array}{ll} \text{minimize} & f(x) \\ \text{subject to} & Ax = b, \end{array}$$

at the feasible point x, we replace the objective with its second-order Taylor approximation near x, to form the problem

$$\begin{array}{ll} \text{minimize} & \widehat{f}(x+v) = f(x) + \nabla f(x)^T v + (1/2) v^T \nabla^2 f(x) v \\ \text{subject to} & A(x+v) = b, \end{array} \tag{10.10}$$

with variable v. This is a (convex) quadratic minimization problem with equality constraints, and can be solved analytically. We define $\Delta x_{\rm nt}$, the Newton step at x, as the solution of the convex quadratic problem (10.10), assuming the associated KKT matrix is nonsingular. In other words, the Newton step $\Delta x_{\rm nt}$ is what must be added to x to solve the problem when the quadratic approximation is used in place of f.

From our analysis in §10.1.1 of the equality constrained quadratic problem, the Newton step $\Delta x_{\rm nt}$ is characterized by

$$\left[\begin{array}{cc} \nabla^2 f(x) & A^T \\ A & 0 \end{array} \right] \left[\begin{array}{c} \Delta x_{\rm nt} \\ w \end{array} \right] = \left[\begin{array}{c} -\nabla f(x) \\ 0 \end{array} \right], \tag{10.11}$$

where w is the associated optimal dual variable for the quadratic problem. The Newton step is defined only at points for which the KKT matrix is nonsingular.

As in Newton's method for unconstrained problems, we observe that when the objective f is exactly quadratic, the Newton update $x + \Delta x_{\rm nt}$ exactly solves the equality constrained minimization problem, and in this case the vector w is the optimal dual variable for the original problem. This suggests, as in the unconstrained case, that when f is nearly quadratic, $x + \Delta x_{\rm nt}$ should be a very good estimate of the solution $x^\star$, and w should be a good estimate of the optimal dual variable $\nu^\star$.

Solution of linearized optimality conditions

We can interpret the Newton step $\Delta x_{\rm nt}$, and the associated vector w, as the solutions of a linearized approximation of the optimality conditions

$$Ax^\star = b, \qquad \nabla f(x^\star) + A^T \nu^\star = 0.$$

We substitute $x + \Delta x_{\rm nt}$ for $x^\star$ and w for $\nu^\star$, and replace the gradient term in the second equation by its linearized approximation near x, to obtain the equations

$$A(x + \Delta x_{\rm nt}) = b, \qquad \nabla f(x + \Delta x_{\rm nt}) + A^T w \approx \nabla f(x) + \nabla^2 f(x) \Delta x_{\rm nt} + A^T w = 0.$$

Using $Ax = b$, these become

$$A \Delta x_{\rm nt} = 0, \qquad \nabla^2 f(x) \Delta x_{\rm nt} + A^T w = -\nabla f(x),$$

which are precisely the equations (10.11) that define the Newton step.

The Newton decrement

We define the Newton decrement for the equality constrained problem as

$$\lambda(x) = (\Delta x_{\rm nt}^T \nabla^2 f(x) \Delta x_{\rm nt})^{1/2}. \tag{10.12}$$

This is exactly the same expression as (9.29), used in the unconstrained case, and the same interpretations hold. For example, $\lambda(x)$ is the norm of the Newton step, in the norm determined by the Hessian.

Let

$$\widehat{f}(x+v) = f(x) + \nabla f(x)^T v + (1/2) v^T \nabla^2 f(x) v$$

be the second-order Taylor approximation of f at x. The difference between $f(x)$ and the minimum of the second-order model satisfies

$$f(x) - \inf\{\widehat{f}(x+v) \mid A(x+v) = b\} = \lambda(x)^2/2, \tag{10.13}$$

exactly as in the unconstrained case (see exercise 10.6). This means that, as in the unconstrained case, $\lambda(x)^2/2$ gives an estimate of $f(x) - p^\star$, based on the quadratic model at x, and also that $\lambda(x)$ (or a multiple of $\lambda(x)^2$) serves as the basis of a good stopping criterion.

The Newton decrement comes up in the line search as well, since the directional derivative of f in the direction $\Delta x_{\rm nt}$ is

$$\frac{d}{dt} f(x + t\Delta x_{\rm nt})\bigg|_{t=0} = \nabla f(x)^T \Delta x_{\rm nt} = -\lambda(x)^2, \tag{10.14}$$

as in the unconstrained case.

Feasible descent direction

Suppose that $Ax = b$. We say that $v \in \mathbf{R}^n$ is a *feasible direction* if $Av = 0$. In this case, every point of the form $x + tv$ is also feasible, *i.e.*, $A(x+tv) = b$. We say that v is a *descent direction* for f at x, if for small $t > 0$, $f(x+tv) < f(x)$.

The Newton step is always a feasible descent direction (except when x is optimal, in which case $\Delta x_{\rm nt} = 0$). Indeed, the second set of equations that define $\Delta x_{\rm nt}$ are $A\Delta x_{\rm nt} = 0$, which shows it is a feasible direction; that it is a descent direction follows from (10.14).

Affine invariance

Like the Newton step and decrement for unconstrained optimization, the Newton step and decrement for equality constrained optimization are affine invariant. Suppose $T \in \mathbf{R}^{n\times n}$ is nonsingular, and define $\bar{f}(y) = f(Ty)$. We have

$$\nabla \bar{f}(y) = T^T \nabla f(Ty), \qquad \nabla^2 \bar{f}(y) = T^T \nabla^2 f(Ty) T,$$

and the equality constraint $Ax = b$ becomes $ATy = b$.

Now consider the problem of minimizing $\bar{f}(y)$, subject to $ATy = b$. The Newton step $\Delta y_{\rm nt}$ at y is given by the solution of

$$\begin{bmatrix} T^T \nabla^2 f(Ty) T & T^T A^T \\ AT & 0 \end{bmatrix} \begin{bmatrix} \Delta y_{\rm nt} \\ \bar{w} \end{bmatrix} = \begin{bmatrix} -T^T \nabla f(Ty) \\ 0 \end{bmatrix}.$$

Comparing with the Newton step $\Delta x_{\rm nt}$ for f at $x = Ty$, given in (10.11), we see that

$$T\Delta y_{\rm nt} = \Delta x_{\rm nt}$$

(and $w = \bar{w}$), *i.e.*, the Newton steps at y and x are related by the same change of coordinates as $Ty = x$.

10.2.2 Newton's method with equality constraints

The outline of Newton's method with equality constraints is exactly the same as for unconstrained problems.

Algorithm 10.1 *Newton's method for equality constrained minimization.*

given starting point $x \in \mathbf{dom}\, f$ with $Ax = b$, tolerance $\epsilon > 0$.

repeat

1. Compute the Newton step and decrement $\Delta x_{\rm nt}$, $\lambda(x)$.
2. *Stopping criterion.* **quit** if $\lambda^2/2 \leq \epsilon$.
3. *Line search.* Choose step size t by backtracking line search.
4. *Update.* $x := x + t\Delta x_{\rm nt}$.

The method is called a *feasible descent method*, since all the iterates are feasible, with $f(x^{(k+1)}) < f(x^{(k)})$ (unless $x^{(k)}$ is optimal). Newton's method requires that the KKT matrix be invertible at each x; we will be more precise about the assumptions required for convergence in §10.2.4.

10.2.3 Newton's method and elimination

We now show that the iterates in Newton's method for the equality constrained problem (10.1) coincide with the iterates in Newton's method applied to the reduced problem (10.5). Suppose F satisfies $\mathcal{R}(F) = \mathcal{N}(A)$ and $\mathbf{rank}\, F = n - p$, and $\hat{x}$ satisfies $A\hat{x} = b$. The gradient and Hessian of the reduced objective function $\tilde{f}(z) = f(Fz + \hat{x})$ are

$$\nabla \tilde{f}(z) = F^T \nabla f(Fz + \hat{x}), \qquad \nabla^2 \tilde{f}(z) = F^T \nabla^2 f(Fz + \hat{x})F.$$

From the Hessian expression, we see that the Newton step for the equality constrained problem is defined, *i.e.*, the KKT matrix

$$\begin{bmatrix} \nabla^2 f(x) & A^T \\ A & 0 \end{bmatrix}$$

is invertible, if and only if the Newton step for the reduced problem is defined, *i.e.*, $\nabla^2 \tilde{f}(z)$ is invertible.

The Newton step for the reduced problem is

$$\Delta z_{\rm nt} = -\nabla^2 \tilde{f}(z)^{-1} \nabla \tilde{f}(z) = -(F^T \nabla^2 f(x) F)^{-1} F^T \nabla f(x), \tag{10.15}$$

where $x = Fz + \hat{x}$. This search direction for the reduced problem corresponds to the direction

$$F\Delta z_{\rm nt} = -F(F^T\nabla^2 f(x)F)^{-1}F^T\nabla f(x)$$

for the original, equality constrained problem. We claim this is precisely the same as the Newton direction $\Delta x_{\rm nt}$ for the original problem, defined in (10.11).

To show this, we take $\Delta x_{\rm nt} = F\Delta z_{\rm nt}$, choose

$$w = -(AA^T)^{-1}A(\nabla f(x) + \nabla^2 f(x)\Delta x_{\rm nt}),$$

and verify that the equations defining the Newton step,

$$\nabla^2 f(x)\Delta x_{\rm nt} + A^T w + \nabla f(x) = 0, \qquad A\Delta x_{\rm nt} = 0, \tag{10.16}$$

hold. The second equation, $A\Delta x_{\rm nt} = 0$, is satisfied because $AF = 0$. To verify the first equation, we observe that

$$\begin{aligned}
&\begin{bmatrix} F^T \\ A \end{bmatrix} \left(\nabla^2 f(x)\Delta x_{\rm nt} + A^T w + \nabla f(x)\right) \\
&\quad = \begin{bmatrix} F^T\nabla^2 f(x)\Delta x_{\rm nt} + F^T A^T w + F^T\nabla f(x) \\ A\nabla^2 f(x)\Delta x_{\rm nt} + AA^T w + A\nabla f(x) \end{bmatrix} \\
&\quad = 0.
\end{aligned}$$

Since the matrix on the left of the first line is nonsingular, we conclude that (10.16) holds.

In a similar way, the Newton decrement $\tilde{\lambda}(z)$ of $\tilde{f}$ at z and the Newton decrement of f at x turn out to be equal:

$$\begin{aligned}
\tilde{\lambda}(z)^2 &= \Delta z_{\rm nt}^T\nabla^2\tilde{f}(z)\Delta z_{\rm nt} \\
&= \Delta z_{\rm nt}^T F^T\nabla^2 f(x)F\Delta z_{\rm nt} \\
&= \Delta x_{\rm nt}^T\nabla^2 f(x)\Delta x_{\rm nt} \\
&= \lambda(x)^2.
\end{aligned}$$

10.2.4 Convergence analysis

We saw above that applying Newton's method with equality constraints is exactly the same as applying Newton's method to the reduced problem obtained by eliminating the equality constraints. Everything we know about the convergence of Newton's method for unconstrained problems therefore transfers to Newton's method for equality constrained constrained problems. In particular, the practical performance of Newton's method with equality constraints is exactly like the performance of Newton's method for unconstrained problems. Once $x^{(k)}$ is near $x^\star$, convergence is extremely rapid, with a very high accuracy obtained in only a few iterations.

Assumptions

We make the following assumptions.

- The sublevel set $S = \{x \mid x \in \mathbf{dom}\, f,\ f(x) \leq f(x^{(0)}),\ Ax = b\}$ is closed, where $x^{(0)} \in \mathbf{dom}\, f$ satisfies $Ax^{(0)} = b$. This is the case if f is closed (see §A.3.3).

- On the set S, we have $\nabla^2 f(x) \preceq MI$, and

$$\left\| \begin{bmatrix} \nabla^2 f(x) & A^T \\ A & 0 \end{bmatrix}^{-1} \right\|_2 \leq K, \tag{10.17}$$

  i.e., the inverse of the KKT matrix is bounded on S. (Of course the inverse must exist in order for the Newton step to be defined at each point in S.)

- For $x,\ \tilde{x} \in S$, $\nabla^2 f$ satisfies the Lipschitz condition $\|\nabla^2 f(x) - \nabla^2 f(\tilde{x})\|_2 \leq L\|x - \tilde{x}\|_2$.

Bounded inverse KKT matrix assumption

The condition (10.17) plays the role of the strong convexity assumption in the standard Newton method (§9.5.3, page 488). When there are no equality constraints, (10.17) reduces to the condition $\|\nabla^2 f(x)^{-1}\|_2 \leq K$ on S, so we can take $K = 1/m$, if $\nabla f(x)^2 \succeq mI$ on S, where $m > 0$. With equality constraints, the condition is not as simple as a positive lower bound on the minimum eigenvalue. Since the KKT matrix is symmetric, the condition (10.17) is that its eigenvalues, n of which are positive, and p of which are negative, are bounded away from zero.

Analysis via the eliminated problem

The assumptions above imply that the eliminated objective function $\tilde{f}$, together with the associated initial point $z^{(0)} = \hat{x} + Fx^{(0)}$, satisfy the assumptions required in the convergence analysis of Newton's method for unconstrained problems, given in §9.5.3 (with different constants $\tilde{m}$, $\tilde{M}$, and $\tilde{L}$). It follows that Newton's method with equality constraints converges to $x^\star$ (and $\nu^\star$ as well).

To show that the assumptions above imply that the eliminated problem satisfies the assumptions for the unconstrained Newton method is mostly straightforward (see exercise 10.4). Here we show the one implication that is tricky: that the bounded inverse KKT condition, together with the upper bound $\nabla^2 f(x) \preceq MI$, implies that $\nabla^2 \tilde{f}(z) \succeq mI$ for some positive constant m. More specifically we will show that this inequality holds for

$$m = \frac{\sigma_{\min}(F)^2}{K^2 M}, \tag{10.18}$$

which is positive, since F is full rank.

We show this by contradiction. Suppose that $F^T HF \not\succeq mI$, where $H = \nabla^2 f(x)$. Then we can find u, with $\|u\|_2 = 1$, such that $u^T F^T HFu < m$, *i.e.*, $\|H^{1/2}Fu\|_2 < m^{1/2}$. Using $AF = 0$, we have

$$\begin{bmatrix} H & A^T \\ A & 0 \end{bmatrix} \begin{bmatrix} Fu \\ 0 \end{bmatrix} = \begin{bmatrix} HFu \\ 0 \end{bmatrix},$$

and so

$$\left\| \begin{bmatrix} H & A^T \\ A & 0 \end{bmatrix}^{-1} \right\|_2 \geq \frac{\left\| \begin{bmatrix} Fu \\ 0 \end{bmatrix} \right\|_2}{\left\| \begin{bmatrix} HFu \\ 0 \end{bmatrix} \right\|_2} = \frac{\|Fu\|_2}{\|HFu\|_2}.$$

Using $\|Fu\|_2 \geq \sigma_{\min}(F)$ and

$$\|HFu\|_2 \leq \|H^{1/2}\|_2 \|H^{1/2}Fu\|_2 < M^{1/2}m^{1/2},$$

we conclude

$$\left\| \begin{bmatrix} H & A^T \\ A & 0 \end{bmatrix}^{-1} \right\|_2 \geq \frac{\|Fu\|_2}{\|HFu\|_2} > \frac{\sigma_{\min}(F)}{M^{1/2}m^{1/2}} = K,$$

using our expression for m given in (10.18).

Convergence analysis for self-concordant functions

If f is self-concordant, then so is $\tilde{f}(z) = f(Fx + \hat{x})$. It follows that if f is self-concordant, we have the exact same complexity estimate as for unconstrained problems: the number of iterations required to produce a solution within an accuracy ϵ is no more than

$$\frac{20 - 8\alpha}{\alpha\beta(1-2\alpha)^2}(f(x^{(0)}) - p^\star) + \log_2 \log_2(1/\epsilon),$$

where α and β are the backtracking parameters (see (9.56)).

10.3 Infeasible start Newton method

Newton's method, as described in §10.2, is a feasible descent method. In this section we describe a generalization of Newton's method that works with initial points, and iterates, that are not feasible.

10.3.1 Newton step at infeasible points

As in Newton's method, we start with the optimality conditions for the equality constrained minimization problem:

$$Ax^\star = b, \qquad \nabla f(x^\star) + A^T\nu^\star = 0.$$

Let x denote the current point, which we do *not* assume to be feasible, but we do assume satisfies $x \in \mathbf{dom}\, f$. Our goal is to find a step Δx so that $x + \Delta x$ satisfies (at least approximately) the optimality conditions, *i.e.*, $x + \Delta x \approx x^\star$. To do this

we substitute $x + \Delta x$ for $x^\star$ and w for $\nu^\star$ in the optimality conditions, and use the first-order approximation

$$\nabla f(x + \Delta x) \approx \nabla f(x) + \nabla^2 f(x)\Delta x$$

for the gradient to obtain

$$A(x + \Delta x) = b, \qquad \nabla f(x) + \nabla^2 f(x)\Delta x + A^T w = 0.$$

This is a set of linear equations for Δx and w,

$$\left[\begin{array}{cc} \nabla^2 f(x) & A^T \\ A & 0 \end{array} \right] \left[\begin{array}{c} \Delta x \\ w \end{array} \right] = - \left[\begin{array}{c} \nabla f(x) \\ Ax - b \end{array} \right]. \tag{10.19}$$

The equations are the same as the equations (10.11) that define the Newton step at a feasible point x, with one difference: the second block component of the righthand side contains $Ax - b$, which is the residual vector for the linear equality constraints. When x is feasible, the residual vanishes, and the equations (10.19) reduce to the equations (10.11) that define the standard Newton step at a feasible point x. Thus, if x is feasible, the step Δx defined by (10.19) coincides with the Newton step described above (but defined only when x is feasible). For this reason we use the notation $\Delta x_{\rm nt}$ for the step Δx defined by (10.19), and refer to it as the Newton step at x, with no confusion.

Interpretation as primal-dual Newton step

We can give an interpretation of the equations (10.19) in terms of a *primal-dual method* for the equality constrained problem. By a primal-dual method, we mean one in which we update both the primal variable x, and the dual variable ν, in order to (approximately) satisfy the optimality conditions.

We express the optimality conditions as $r(x^\star, \nu^\star) = 0$, where $r : \mathbf{R}^n \times \mathbf{R}^p \to \mathbf{R}^n \times \mathbf{R}^p$ is defined as

$$r(x, \nu) = (r_{\rm dual}(x, \nu), r_{\rm pri}(x, \nu)).$$

Here

$$r_{\rm dual}(x, \nu) = \nabla f(x) + A^T \nu, \qquad r_{\rm pri}(x, \nu) = Ax - b$$

are the *dual residual* and *primal residual*, respectively. The first-order Taylor approximation of r, near our current estimate y, is

$$r(y + z) \approx \hat{r}(y + z) = r(y) + Dr(y)z,$$

where $Dr(y) \in \mathbf{R}^{(n+p)\times(n+p)}$ is the derivative of r, evaluated at y (see §A.4.1). We define the primal-dual Newton step $\Delta y_{\rm pd}$ as the step z for which the Taylor approximation $\hat{r}(y + z)$ vanishes, *i.e.*,

$$Dr(y)\Delta y_{\rm pd} = -r(y). \tag{10.20}$$

Note that here we consider both x and ν as variables; $\Delta y_{\rm pd} = (\Delta x_{\rm pd}, \Delta \nu_{\rm pd})$ gives both a primal and a dual step.

Evaluating the derivative of r, we can express (10.20) as

$$\begin{bmatrix} \nabla^2 f(x) & A^T \\ A & 0 \end{bmatrix} \begin{bmatrix} \Delta x_{\rm pd} \\ \Delta \nu_{\rm pd} \end{bmatrix} = - \begin{bmatrix} r_{\rm dual} \\ r_{\rm pri} \end{bmatrix} = - \begin{bmatrix} \nabla f(x) + A^T \nu \\ Ax - b \end{bmatrix}. \tag{10.21}$$

Writing $\nu + \Delta\nu_{\rm pd}$ as ν^+, we can express this as

$$\begin{bmatrix} \nabla^2 f(x) & A^T \\ A & 0 \end{bmatrix} \begin{bmatrix} \Delta x_{\rm pd} \\ \nu^+ \end{bmatrix} = - \begin{bmatrix} \nabla f(x) \\ Ax - b \end{bmatrix}, \tag{10.22}$$

which is exactly the same set of equations as (10.19). The solutions of (10.19), (10.21), and (10.22) are therefore related as

$$\Delta x_{\rm nt} = \Delta x_{\rm pd}, \qquad w = \nu^+ = \nu + \Delta\nu_{\rm pd}.$$

This shows that the (infeasible) Newton step is the same as the primal part of the primal-dual step, and the associated dual vector w is the updated primal-dual variable $\nu^+ = \nu + \Delta\nu_{\rm pd}$.

The two expressions for the Newton step and dual variable (or dual step), given by (10.21) and (10.22), are of course equivalent, but each reveals a different feature of the Newton step. The equation (10.21) shows that the Newton step and the associated dual step are obtained by solving a set of equations, with the primal and dual residuals as the righthand side. The equation (10.22), which is how we originally defined the Newton step, gives the Newton step and the updated dual variable, and shows that the current value of the dual variable is not needed to compute the primal step, or the updated value of the dual variable.

Residual norm reduction property

The Newton direction, at an infeasible point, is not necessarily a descent direction for f. From (10.21), we note that

$$\begin{aligned} \frac{d}{dt} f(x + t\Delta x)\Big|_{t=0} &= \nabla f(x)^T \Delta x \\ &= -\Delta x^T \left(\nabla^2 f(x) \Delta x + A^T w \right) \\ &= -\Delta x^T \nabla^2 f(x) \Delta x + (Ax - b)^T w, \end{aligned}$$

which is not necessarily negative (unless, of course, x is feasible, *i.e.*, $Ax = b$). The primal-dual interpretation, however, shows that the norm of the residual decreases in the Newton direction, *i.e.*,

$$\frac{d}{dt} \left\| r(y + t\Delta y_{\rm pd}) \right\|_2^2 \Big|_{t=0} = 2r(y)^T Dr(y) \Delta y_{\rm pd} = -2r(y)^T r(y).$$

Taking the derivative of the square, we obtain

$$\frac{d}{dt} \left\| r(y + t\Delta y_{\rm pd}) \right\|_2 \Big|_{t=0} = -\|r(y)\|_2. \tag{10.23}$$

This allows us to use $\|r\|_2$ to measure the progress of the infeasible start Newton method, for example, in the line search. (For the standard Newton method, we use the function value f to measure progress of the algorithm, at least until quadratic convergence is attained.)

Full step feasibility property

The Newton step $\Delta x_{\rm nt}$ defined by (10.19) has the property (by construction) that

$$A(x + \Delta x_{\rm nt}) = b. \tag{10.24}$$

It follows that, if a step length of one is taken using the Newton step $\Delta x_{\rm nt}$, the following iterate will be feasible. Once x is feasible, the Newton step becomes a feasible direction, so all future iterates will be feasible, regardless of the step sizes taken.

More generally, we can analyze the effect of a damped step on the equality constraint residual $r_{\rm pri}$. With a step length $t \in [0, 1]$, the next iterate is $x^+ = x + t\Delta x_{\rm nt}$, so the equality constraint residual at the next iterate is

$$r_{\rm pri}^+ = A(x + \Delta x_{\rm nt} t) - b = (1 - t)(Ax - b) = (1 - t)r_{\rm pri},$$

using (10.24). Thus, a damped step, with length t, causes the residual to be scaled down by a factor $1 - t$. Now suppose that we have $x^{(i+1)} = x^{(i)} + t^{(i)}\Delta x_{\rm nt}^{(i)}$, for $i = 0, \ldots, k - 1$, where $\Delta x_{\rm nt}^{(i)}$ is the Newton step at the point $x^{(i)} \in \mathbf{dom}\, f$, and $t^{(i)} \in [0, 1]$. Then we have

$$r^{(k)} = \left(\prod_{i=0}^{k-1} (1 - t^{(i)}) \right) r^{(0)},$$

where $r^{(i)} = Ax^{(i)} - b$ is the residual of $x^{(i)}$. This formula shows that the primal residual at each step is in the direction of the initial primal residual, and is scaled down at each step. It also shows that once a full step is taken, all future iterates are primal feasible.

10.3.2 Infeasible start Newton method

We can develop an extension of Newton's method, using the Newton step $\Delta x_{\rm nt}$ defined by (10.19), with $x^{(0)} \in \mathbf{dom}\, f$, but not necessarily satisfying $Ax^{(0)} = b$. We also use the dual part of the Newton step: $\Delta\nu_{\rm nt} = w - \nu$ in the notation of (10.19), or equivalently, $\Delta\nu_{\rm nt} = \Delta\nu_{\rm pd}$ in the notation of (10.21).

Algorithm 10.2 *Infeasible start Newton method.*

given starting point $x \in \mathbf{dom}\, f$, ν, tolerance $\epsilon > 0$, $\alpha \in (0, 1/2)$, $\beta \in (0, 1)$.

repeat

1. Compute primal and dual Newton steps $\Delta x_{\rm nt}$, $\Delta\nu_{\rm nt}$.
2. *Backtracking line search on* $\|r\|_2$.
   $t := 1$.
   while $\|r(x + t\Delta x_{\rm nt}, \nu + t\Delta\nu_{\rm nt})\|_2 > (1 - \alpha t)\|r(x, \nu)\|_2$, $t := \beta t$.
3. *Update.* $x := x + t\Delta x_{\rm nt}$, $\nu := \nu + t\Delta\nu_{\rm nt}$.

until $Ax = b$ and $\|r(x, \nu)\|_2 \leq \epsilon$.

This algorithm is very similar to the standard Newton method with feasible starting point, with a few exceptions. First, the search directions include the extra correction terms that depend on the primal residual. Second, the line search is carried out using the norm of the residual, instead of the function value f. Finally, the algorithm terminates when primal feasibility has been achieved, and the norm of the (dual) residual is small.

The line search in step 2 deserves some comment. Using the norm of the residual in the line search can increase the cost, compared to a line search based on the function value, but the increase is usually negligible. Also, we note that the line search must terminate in a finite number of steps, since (10.23) shows that the line search exit condition is satisfied for small t.

The equation (10.24) shows that if at some iteration the step length is chosen to be one, the next iterate will be feasible. Thereafter, all iterates will be feasible, and therefore the search direction for the infeasible start Newton method coincides, once a feasible iterate is obtained, with the search direction for the (feasible) Newton method described in §10.2.

There are many variations on the infeasible start Newton method. For example, we can switch to the (feasible) Newton method described in §10.2 once feasibility is achieved. (In other words, we change the line search to one based on f, and terminate when $\lambda(x)^2/2 \leq \epsilon$.) Once feasibility is achieved, the infeasible start and the standard (feasible) Newton method differ only in the backtracking and exit conditions, and have very similar performance.

Using infeasible start Newton method to simplify initialization

The main advantage of the infeasible start Newton method is in the initialization required. If $\mathbf{dom}\, f = \mathbf{R}^n$, then initializing the (feasible) Newton method simply requires computing a solution to $Ax = b$, and there is no particular advantage, other than convenience, in using the infeasible start Newton method.

When $\mathbf{dom}\, f$ is not all of $\mathbf{R}^n$, finding a point in $\mathbf{dom}\, f$ that satisfies $Ax = b$ can itself be a challenge. One general approach, probably the best when $\mathbf{dom}\, f$ is complex and not known to intersect $\{z \mid Az = b\}$, is to use a phase I method (described in §11.4) to compute such a point (or verify that $\mathbf{dom}\, f$ does not intersect $\{z \mid Az = b\}$). But when $\mathbf{dom}\, f$ is relatively simple, and known to contain a point satisfying $Ax = b$, the infeasible start Newton method gives a simple alternative.

One common example occurs when $\mathbf{dom}\, f = \mathbf{R}^n_{++}$, as in the equality constrained analytic centering problem described in example 10.2. To initialize Newton's method for the problem

$$\begin{array}{ll} \text{minimize} & -\sum_{i=1}^n \log x_i \\ \text{subject to} & Ax = b, \end{array} \tag{10.25}$$

requires finding a point $x^{(0)} \succ 0$ with $Ax = b$, which is equivalent to solving a standard form LP feasibility problem. This can be carried out using a phase I method, or alternatively, using the infeasible start Newton method, with any positive initial point, *e.g.*, $x^{(0)} = \mathbf{1}$.

The same trick can be used to initialize unconstrained problems where a starting point in $\mathbf{dom}\, f$ is not known. As an example, we consider the dual of the equality

constrained analytic centering problem (10.25),

$$\text{maximize} \quad g(\nu) = -b^T\nu + n + \sum_{i=1}^n \log(A^T\nu)_i.$$

To initialize this problem for the (feasible start) Newton method, we must find a point $\nu^{(0)}$ that satisfies $A^T\nu^{(0)} \succ 0$, *i.e.*, we must solve a set of linear inequalities. This can be done using a phase I method, or using an infeasible start Newton method, after reformulating the problem. We first express it as an equality constrained problem,

$$\begin{array}{ll} \text{maximize} & -b^T\nu + n + \sum_{i=1}^n \log y_i \\ \text{subject to} & y = A^T\nu, \end{array}$$

with new variable $y \in \mathbf{R}^n$. We can now use the infeasible start Newton method, starting with any positive $y^{(0)}$ (and any $\nu^{(0)}$).

The disadvantage of using the infeasible start Newton method to initialize problems for which a strictly feasible starting point is not known is that there is no clear way to detect that there exists no strictly feasible point; the norm of the residual will simply converge, slowly, to some positive value. (Phase I methods, in contrast, can determine this fact unambiguously.) In addition, the convergence of the infeasible start Newton method, before feasibility is achieved, can be slow; see §11.4.2.

10.3.3 Convergence analysis

In this section we show that the infeasible start Newton method converges to the optimal point, provided certain assumptions hold. The convergence proof is very similar to those for the standard Newton method, or the standard Newton method with equality constraints. We show that once the norm of the residual is small enough, the algorithm takes full steps (which implies that feasibility is achieved), and convergence is subsequently quadratic. We also show that the norm of the residual is reduced by at least a fixed amount in each iteration before the region of quadratic convergence is reached. Since the norm of the residual cannot be negative, this shows that within a finite number of steps, the residual will be small enough to guarantee full steps, and quadratic convergence.

Assumptions

We make the following assumptions.

- The sublevel set

$$S = \{(x,\nu) \mid x \in \mathbf{dom}\, f,\ \|r(x,\nu)\|_2 \leq \|r(x^{(0)},\nu^{(0)})\|_2\} \tag{10.26}$$

  is closed. If f is closed, then $\|r\|_2$ is a closed function, and therefore this condition is satisfied for any $x^{(0)} \in \mathbf{dom}\, f$ and any $\nu^{(0)} \in \mathbf{R}^p$ (see exercise 10.7).

- On the set S, we have

$$\|Dr(x,\nu)^{-1}\|_2 = \left\| \begin{bmatrix} \nabla^2 f(x) & A^T \\ A & 0 \end{bmatrix}^{-1} \right\|_2 \leq K, \tag{10.27}$$

for some K.

- For (x, ν), $(\tilde{x}, \tilde{\nu}) \in S$, Dr satisfies the Lipschitz condition

$$\|Dr(x, \nu) - Dr(\tilde{x}, \tilde{\nu})\|_2 \leq L\|(x, \nu) - (\tilde{x}, \tilde{\nu})\|_2.$$

  (This is equivalent to $\nabla^2 f(x)$ satisfying a Lipschitz condition; see exercise 10.7.)

As we will see below, these assumptions imply that $\mathbf{dom}\, f$ and $\{z \mid Az = b\}$ intersect, and that there is an optimal point $(x^\star, \nu^\star)$.

Comparison with standard Newton method

The assumptions above are very similar to the ones made in §10.2.4 (page 529) for the analysis of the standard Newton method. The second and third assumptions, the bounded inverse KKT matrix and Lipschitz condition, are essentially the same. The sublevel set condition (10.26) for the infeasible start Newton method is, however, more general than the sublevel set condition made in §10.2.4.

As an example, consider the equality constrained maximum entropy problem

$$\begin{array}{ll} \text{minimize} & f(x) = \sum_{i=1}^n x_i \log x_i \\ \text{subject to} & Ax = b, \end{array}$$

with $\mathbf{dom}\, f = \mathbf{R}^n_{++}$. The objective f is *not* closed; it has sublevel sets that are not closed, so the assumptions made in the standard Newton method may not hold, at least for some initial points. The problem here is that the negative entropy function does not converge to ∞ as $x_i \to 0$. On the other hand the sublevel set condition (10.26) for the infeasible start Newton method *does* hold for this problem, since the norm of the gradient of the negative entropy function does converge to ∞ as $x_i \to 0$. Thus, the infeasible start Newton method is guaranteed to solve the equality constrained maximum entropy problem. (We do not know whether the standard Newton method can fail for this problem; we are only observing here that our convergence analysis does not hold.) Note that if the initial point satisfies the equality constraints, the only difference between the standard and infeasible start Newton methods is in the line searches, which differ only during the damped stage.

A basic inequality

We start by deriving a basic inequality. Let $y = (x, \nu) \in S$ with $\|r(y)\|_2 \neq 0$, and let $\Delta y_{\rm nt} = (\Delta x_{\rm nt}, \Delta \nu_{\rm nt})$ be the Newton step at y. Define

$$t_{\max} = \inf\{t > 0 \mid y + t\Delta y_{\rm nt} \notin S\}.$$

If $y + t\Delta y_{\rm nt} \in S$ for all $t \geq 0$, we follow the usual convention and define $t_{\max} = \infty$. Otherwise, $t_{\max}$ is the smallest positive value of t such that $\|r(y + t\Delta y_{\rm nt})\|_2 = \|r(y^{(0)})\|_2$. In particular, it follows that $y + t\Delta y_{\rm nt} \in S$ for $0 \leq t \leq t_{\max}$.

We will show that

$$\|r(y + t\Delta y_{\rm nt})\|_2 \leq (1 - t)\|r(y)\|_2 + (K^2 L/2)t^2\|r(y)\|_2^2 \tag{10.28}$$

for $0 \leq t \leq \min\{1, t_{\max}\}$.

We have

$$\begin{aligned} r(y + t\Delta y_{\rm nt}) &= r(y) + \int_0^1 Dr(y + \tau t \Delta y_{\rm nt}) t \Delta y_{\rm nt}\, d\tau \\ &= r(y) + tDr(y)\Delta y_{\rm nt} + \int_0^1 (Dr(y + \tau t \Delta y_{\rm nt}) - Dr(y)) t \Delta y_{\rm nt}\, d\tau \\ &= r(y) + tDr(y)\Delta y_{\rm nt} + e \\ &= (1-t)r(y) + e, \end{aligned}$$

using $Dr(y)\Delta y_{\rm nt} = -r(y)$, and defining

$$e = \int_0^1 (Dr(y + \tau t \Delta y_{\rm nt}) - Dr(y)) t \Delta y_{\rm nt}\, d\tau.$$

Now suppose $0 \leq t \leq t_{\max}$, so $y + \tau t \Delta y_{\rm nt} \in S$ for $0 \leq \tau \leq 1$. We can bound $\|e\|_2$ as follows:

$$\begin{aligned} \|e\|_2 &\leq \|t\Delta y_{\rm nt}\|_2 \int_0^1 \|Dr(y + \tau t \Delta y_{\rm nt}) - Dr(y)\|_2\, d\tau \\ &\leq \|t\Delta y_{\rm nt}\|_2 \int_0^1 L\|\tau t \Delta y_{\rm nt}\|_2\, d\tau \\ &= (L/2)t^2 \|\Delta y_{\rm nt}\|_2^2 \\ &= (L/2)t^2 \|Dr(y)^{-1} r(y)\|_2^2 \\ &\leq (K^2 L/2) t^2 \|r(y)\|_2^2, \end{aligned}$$

using the Lipschitz condition on the second line, and the bound $\|Dr(y)^{-1}\|_2 \leq K$ on the last. Now we can derive the bound (10.28): For $0 \leq t \leq \min\{1, t_{\max}\}$,

$$\begin{aligned} \|r(y + t\Delta y_{\rm nt})\|_2 &= \|(1-t)r(y) + e\|_2 \\ &\leq (1-t)\|r(y)\|_2 + \|e\|_2 \\ &\leq (1-t)\|r(y)\|_2 + (K^2 L/2) t^2 \|r(y)\|_2^2. \end{aligned}$$

Damped Newton phase

We first show that if $\|r(y)\|_2 > 1/(K^2 L)$, one iteration of the infeasible start Newton method reduces $\|r\|_2$ by at least a certain minimum amount.

The righthand side of the basic inequality (10.28) is quadratic in t, and monotonically decreasing between $t = 0$ and its minimizer

$$\bar{t} = \frac{1}{K^2 L \|r(y)\|_2} < 1.$$

We must have $t_{\max} > \bar{t}$, because the opposite would imply $\|r(y + t_{\max}\Delta y_{\rm nt})\|_2 < \|r(y)\|_2$, which is false. The basic inequality is therefore valid at $t = \bar{t}$, and therefore

$$\begin{aligned} \|r(y + \bar{t}\Delta y_{\rm nt})\|_2 &\leq \|r(y)\|_2 - 1/(2K^2 L) \\ &\leq \|r(y)\|_2 - \alpha/(K^2 L) \\ &= (1 - \alpha\bar{t})\|r(y)\|_2, \end{aligned}$$

which shows that the step length $\bar{t}$ satisfies the line search exit condition. Therefore we have $t \geq \beta\bar{t}$, where t is the step length chosen by the backtracking algorithm. From $t \geq \beta\bar{t}$ we have (from the exit condition in the backtracking line search)

$$\begin{aligned}
\|r(y + t\Delta y_{\rm nt})\|_2 &\leq (1 - \alpha t)\|r(y)\|_2 \\
&\leq (1 - \alpha\beta\bar{t})\|r(y)\|_2 \\
&= \left(1 - \frac{\alpha\beta}{K^2 L\|r(y)\|_2}\right)\|r(y)\|_2 \\
&= \|r(y)\|_2 - \frac{\alpha\beta}{K^2 L}.
\end{aligned}$$

Thus, as long as we have $\|r(y)\|_2 > 1/(K^2L)$, we obtain a minimum decrease in $\|r\|_2$, per iteration, of $\alpha\beta/(K^2L)$. It follows that a maximum of

$$\frac{\|r(y^{(0)})\|_2 K^2 L}{\alpha\beta}$$

iterations can be taken before we have $\|r(y^{(k)})\|_2 \leq 1/(K^2L)$.

Quadratically convergent phase

Now suppose $\|r(y)\|_2 \leq 1/(K^2L)$. The basic inequality gives

$$\|r(y + t\Delta y_{\rm nt})\|_2 \leq (1 - t + (1/2)t^2)\|r(y)\|_2 \tag{10.29}$$

for $0 \leq t \leq \min\{1, t_{\max}\}$. We must have $t_{\max} > 1$, because otherwise it would follow from (10.29) that $\|r(y + t_{\max}\Delta y_{\rm nt})\|_2 < \|r(y)\|_2$, which contradicts the definition of $t_{\max}$. The inequality (10.29) therefore holds with $t = 1$, *i.e.*, we have

$$\|r(y + \Delta y_{\rm nt})\|_2 \leq (1/2)\|r(y)\|_2 \leq (1 - \alpha)\|r(y)\|_2.$$

This shows that the backtracking line search exit criterion is satisfied for $t = 1$, so a full step will be taken. Moreover, for all future iterations we have $\|r(y)\|_2 \leq 1/(K^2L)$, so a full step will be taken for all following iterations.

We can write the inequality (10.28) (for $t = 1$) as

$$\frac{K^2 L\|r(y^+)\|_2}{2} \leq \left(\frac{K^2 L\|r(y)\|_2}{2}\right)^2,$$

where $y^+ = y + \Delta y_{\rm nt}$. Therefore, if $r(y^{+k})$ denotes the residual k steps after an iteration in which $\|r(y)\|_2 \leq 1/K^2L$, we have

$$\frac{K^2 L\|r(y^{+k})\|_2}{2} \leq \left(\frac{K^2 L\|r(y)\|_2}{2}\right)^{2^k} \leq \left(\frac{1}{2}\right)^{2^k},$$

i.e., we have quadratic convergence of $\|r(y)\|_2$ to zero.

To show that the sequence of iterates converges, we will show that it is a Cauchy sequence. Suppose y is an iterate satisfying $\|r(y)\|_2 \leq 1/(K^2L)$, and y^{+k} denotes

the kth iterate after y. Since these iterates are in the region of quadratic convergence, the step size is one, so we have

$$\begin{aligned}
\|y^{+k} - y\|_2 &\leq \|y^{+k} - y^{+(k-1)}\|_2 + \cdots + \|y^{+} - y\|_2 \\
&= \|Dr(y^{+(k-1)})^{-1} r(y^{+(k-1)})\|_2 + \cdots + \|Dr(y)^{-1} r(y)\|_2 \\
&\leq K\left(\|r(y^{+(k-1)})\|_2 + \cdots + \|r(y)\|_2\right) \\
&\leq K\|r(y)\|_2 \sum_{i=0}^{k-1} \left(\frac{K^2 L \|r(y)\|_2}{2}\right)^{2^i - 1} \\
&\leq K\|r(y)\|_2 \sum_{i=0}^{k-1} \left(\frac{1}{2}\right)^{2^i - 1} \\
&\leq 2K\|r(y)\|_2
\end{aligned}$$

where in the third line we use the assumption that $\|Dr^{-1}\|_2 \leq K$ for all iterates. Since $\|r(y^{(k)})\|_2$ converges to zero, we conclude $y^{(k)}$ is a Cauchy sequence, and therefore converges. By continuity of r, the limit point $y^\star$ satisfies $r(y^\star) = 0$. This establishes our earlier claim that the assumptions at the beginning of this section imply that there is an optimal point $(x^\star, \nu^\star)$.

10.3.4 Convex-concave games

The proof of convergence for the infeasible start Newton method reveals that the method can be used for a larger class of problems than equality constrained convex optimization problems. Suppose $r : \mathbf{R}^n \to \mathbf{R}^n$ is differentiable, its derivative satisfies a Lipschitz condition on S, and $\|Dr(x)^{-1}\|_2$ is bounded on S, where

$$S = \{x \in \mathbf{dom}\, r \mid \|r(x)\|_2 \leq \|r(x^{(0)})\|_2\}$$

is a closed set. Then the infeasible start Newton method, started at $x^{(0)}$, converges to a solution of $r(x) = 0$ in S. In the infeasible start Newton method, we apply this to the specific case in which r is the residual for the equality constrained convex optimization problem. But it applies in several other interesting cases. One interesting example is solving a *convex-concave game*. (See §5.4.3 and exercise 5.25 for discussion of other, related games).

An unconstrained (zero-sum, two-player) game on $\mathbf{R}^p \times \mathbf{R}^q$ is defined by its *payoff function* $f : \mathbf{R}^{p+q} \to \mathbf{R}$. The meaning is that player 1 chooses a value (or move) $u \in \mathbf{R}^p$, and player 2 chooses a value (or move) $v \in \mathbf{R}^q$; based on these choices, player 1 makes a payment to player 2, in the amount $f(u, v)$. The goal of player 1 is to minimize this payment, while the goal of player 2 is to maximize it.

If player 1 makes his choice u first, and player 2 knows the choice, then player 2 will choose v to maximize $f(u, v)$, which results in a payoff of $\sup_v f(u, v)$ (assuming the supremum is achieved). If player 1 assumes that player 2 will make this choice, he should choose u to minimize $\sup_v f(u, v)$. The resulting payoff, from player 1 to player 2, will then be

$$\inf_u \sup_v f(u, v) \tag{10.30}$$

(assuming that the supremum is achieved). On the other hand if player 2 makes the first choice, the strategies are reversed, and the resulting payoff from player 1 to player 2 is

$$\sup_v \inf_u f(u, v). \tag{10.31}$$

The payoff (10.30) is always greater than or equal to the payoff (10.31); the difference between the two payoffs can be interpreted as the advantage afforded the player who makes the second move, with knowledge of the other player's move. We say that $(u^\star, v^\star)$ is a *solution* of the game, or a *saddle-point* for the game, if for all $u,\ v$,

$$f(u^\star, v) \leq f(u^\star, v^\star) \leq f(u, v^\star).$$

When a solution exists, there is no advantage to making the second move; $f(u^\star, v^\star)$ is the common value of both payoffs (10.30) and (10.31). (See exercise 3.14.)

The game is called *convex-concave* if for each v, $f(u, v)$ is a convex function of u, and for each u, $f(u, v)$ is a concave function of v. When f is differentiable (and convex-concave), a saddle-point for the game is characterized by $\nabla f(u^\star, v^\star) = 0$.

Solution via infeasible start Newton method

We can use the infeasible start Newton method to compute a solution of a convex-concave game with twice differentiable payoff function. We define the residual as

$$r(u, v) = \nabla f(u, v) = \begin{bmatrix} \nabla_u f(u, v) \\ \nabla_v f(u, v) \end{bmatrix},$$

and apply the infeasible start Newton method. In the context of games, the infeasible start Newton method is simply called Newton's method (for convex-concave games).

We can guarantee convergence of the (infeasible start) Newton method provided $Dr = \nabla^2 f$ has bounded inverse, and satisfies a Lipschitz condition on the sublevel set

$$S = \{(u, v) \in \mathbf{dom}\, f \mid \|r(u, v)\|_2 \leq \|r(u^{(0)}, v^{(0)})\|_2\},$$

where $u^{(0)},\ v^{(0)}$ are the starting players' choices.

There is a simple analog of the strong convexity condition in an unconstrained minimization problem. We say the game with payoff function f is strongly convex-concave if for some $m > 0$, we have $\nabla^2_{uu} f(u, v) \succeq mI$ and $\nabla^2_{vv} f(u, v) \preceq -mI$, for all $(u, v) \in S$. Not surprisingly, this strong convex-concave assumption implies the bounded inverse condition (exercise 10.10).

10.3.5 Examples

A simple example

We illustrate the infeasible start Newton method on the equality constrained analytic center problem (10.25). Our first example is an instance with dimensions $n = 100$ and $m = 50$, generated randomly, for which the problem is feasible and bounded below. The infeasible start Newton method is used, with initial primal

and dual points $x^{(0)} = \mathbf{1}$, $\nu^{(0)} = 0$, and backtracking parameters $\alpha = 0.01$ and $\beta = 0.5$. The plot in figure 10.1 shows the norms of the primal and dual residuals separately, versus iteration number, and the plot in figure 10.2 shows the step lengths. A full Newton step is taken in iteration 8, so the primal residual becomes (almost) zero, and remains (almost) zero. After around iteration 9 or so, the (dual) residual converges quadratically to zero.

An infeasible example

We also consider a problem instance, of the same dimensions as the example above, for which $\mathbf{dom}\, f$ does not intersect $\{z \mid Az = b\}$, *i.e.*, the problem is infeasible. (This violates the basic assumption in the chapter that problem (10.1) is solvable, as well as the assumptions made in §10.2.4; the example is meant only to show what happens to the infeasible start Newton method when $\mathbf{dom}\, f$ does not intersect $\{z \mid Az = b\}$.) The norm of the residual for this example is shown in figure 10.3, and the step length in figure 10.4. Here, of course, the step lengths are never one, and the residual does not converge to zero.

A convex-concave game

Our final example involves a convex-concave game on $\mathbf{R}^{100} \times \mathbf{R}^{100}$, with payoff function

$$f(u, v) = u^T Av + b^T u + c^T v - \log(1 - u^T u) + \log(1 - v^T v), \tag{10.32}$$

defined on

$$\mathbf{dom}\, f = \{(u, v) \mid u^T u < 1,\ v^T v < 1\}.$$

The problem data A, b, and c were randomly generated. The progress of the (infeasible start) Newton method, started at $u^{(0)} = v^{(0)} = 0$, with backtracking parameters $\alpha = 0.01$ and $\beta = 0.5$, is shown in figure 10.5.

10.4 Implementation

10.4.1 Elimination

To implement the elimination method, we have to calculate a full rank matrix F and an $\hat{x}$ such that

$$\{x \mid Ax = b\} = \{Fz + \hat{x} \mid z \in \mathbf{R}^{n-p}\}.$$

Several methods for this are described in §C.5.

10.4.2 Solving KKT systems

In this section we describe methods that can be used to compute the Newton step or infeasible Newton step, both of which involve solving a set of linear equations

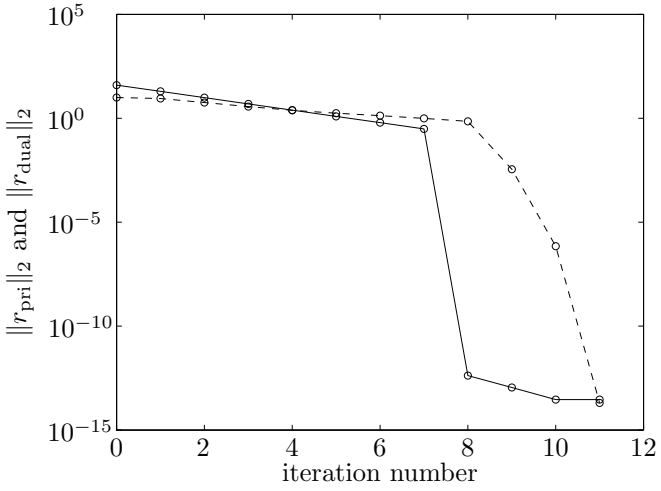

Figure 10.1 Progress of infeasible start Newton method on an equality constrained analytic centering problem with 100 variables and 50 constraints. The figure shows $\|r_{\mathrm{pri}}\|_2$ (solid line), and $\|r_{\mathrm{dual}}\|_2$ (dashed line). Note that feasibility is achieved (and maintained) after 8 iterations, and convergence is quadratic, starting from iteration 9 or so.

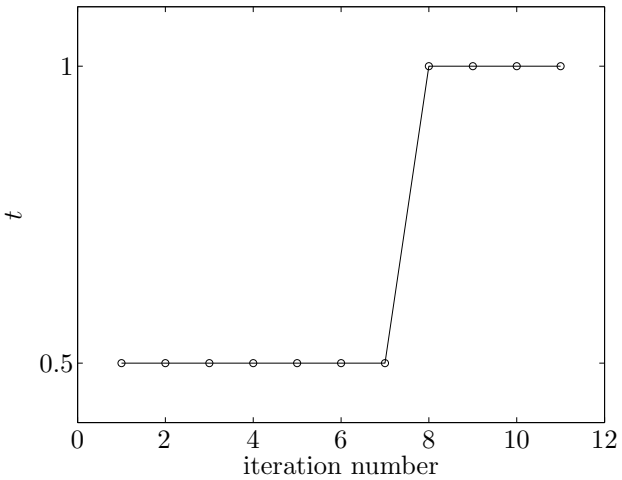

Figure 10.2 Step length versus iteration number for the same example problem. A full step is taken in iteration 8, which results in feasibility from iteration 8 on.

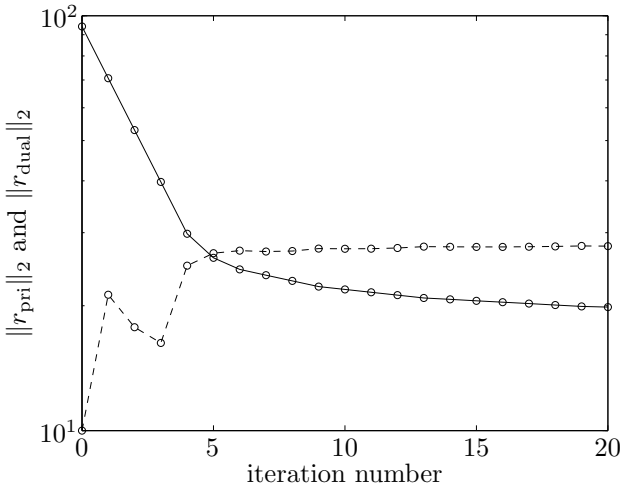

Figure 10.3 Progress of infeasible start Newton method on an equality constrained analytic centering problem with 100 variables and 50 constraints, for which $\mathbf{dom}\, f = \mathbf{R}^{100}_{++}$ does not intersect $\{z \mid Az = b\}$. The figure shows $\|r_{\rm pri}\|_2$ (solid line), and $\|r_{\rm dual}\|_2$ (dashed line). In this case, the residuals do not converge to zero.

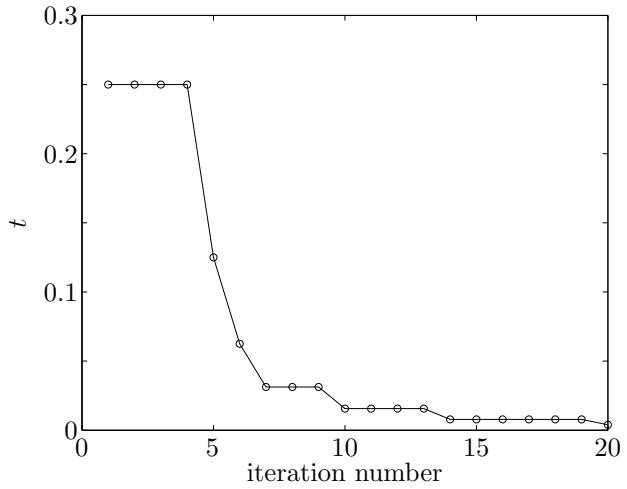

Figure 10.4 Step length versus iteration number for the infeasible example problem. No full steps are taken, and the step lengths converge to zero.

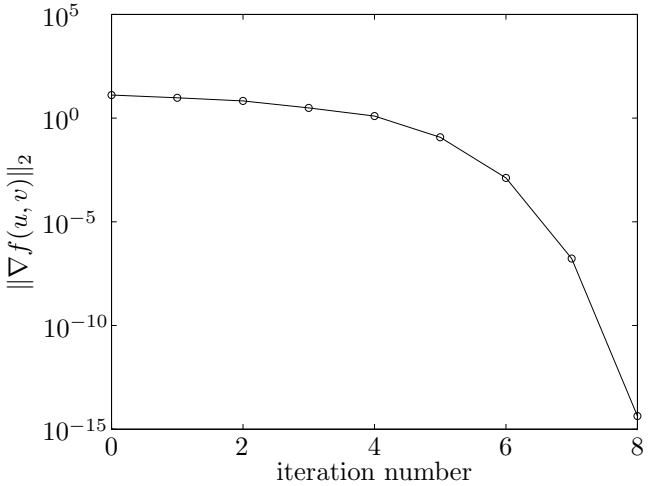

Figure 10.5 Progress of (infeasible start) Newton method on a convex-concave game. Quadratic convergence becomes apparent after about 5 iterations.

with KKT form

$$\begin{bmatrix} H & A^T \\ A & 0 \end{bmatrix} \begin{bmatrix} v \\ w \end{bmatrix} = - \begin{bmatrix} g \\ h \end{bmatrix}. \tag{10.33}$$

Here we assume $H \in \mathbf{S}^n_+$, and $A \in \mathbf{R}^{p\times n}$ with $\mathbf{rank}\, A = p < n$. Similar methods can be used to compute the Newton step for a convex-concave game, in which the bottom right entry of the coefficient matrix is negative semidefinite (see exercise 10.13).

Solving full KKT system

One straightforward approach is to simply solve the KKT system (10.33), which is a set of $n + p$ linear equations in $n + p$ variables. The KKT matrix is symmetric, but not positive definite, so a good way to do this is to use an $\mathrm{LDL}^{\mathrm{T}}$ factorization (see §C.3.3). If no structure of the matrix is exploited, the cost is $(1/3)(n + p)^3$ flops. This can be a reasonable approach when the problem is small (*i.e.*, n and p are not too large), or when A and H are sparse.

Solving KKT system via elimination

A method that is often better than directly solving the full KKT system is based on eliminating the variable v (see §C.4). We start by describing the simplest case, in which $H \succ 0$. Starting from the first of the KKT equations

$$Hv + A^T w = -g, \qquad Av = -h,$$

we solve for v to obtain

$$v = -H^{-1}(g + A^T w).$$

Substituting this into the second KKT equation yields $AH^{-1}(g+A^Tw) = h$, so we have

$$w = (AH^{-1}A^T)^{-1}(h - AH^{-1}g).$$

These formulas give us a method for computing v and w.

The matrix appearing in the formula for w is the Schur complement S of H in the KKT matrix:

$$S = -AH^{-1}A^T.$$

Because of the special structure of the KKT matrix, and our assumption that A has rank p, the matrix S is negative definite.

Algorithm 10.3 *Solving KKT system by block elimination.*

given KKT system with $H \succ 0$.

1. Form $H^{-1}A^T$ and $H^{-1}g$.
2. Form Schur complement $S = -AH^{-1}A^T$.
3. Determine w by solving $Sw = AH^{-1}g - h$.
4. Determine v by solving $Hv = -A^Tw - g$.

Step 1 can be done by a Cholesky factorization of H, followed by $p+1$ solves, which costs $f + (p+1)s$, where f is the cost of factoring H and s is the cost of an associated solve. Step 2 requires a $p \times n$ by $n \times p$ matrix multiplication. If we exploit no structure in this calculation, the cost is p^2n flops. (Since the result is symmetric, we only need to compute the upper triangular part of S.) In some cases special structure in A and H can be exploited to carry out step 2 more efficiently. Step 3 can be carried out by Cholesky factorization of $-S$, which costs $(1/3)p^3$ flops if no further structure of S is exploited. Step 4 can be carried out using the factorization of H already calculated in step 1, so the cost is $2np + s$ flops. The total flop count, assuming that no structure is exploited in forming or factoring the Schur complement, is

$$f + ps + p^2n + (1/3)p^3$$

flops (keeping only dominant terms). If we exploit structure in forming or factoring S, the last two terms are even smaller.

If H can be factored efficiently, then block elimination gives us a flop count advantage over directly solving the KKT system using an $\mathrm{LDL^T}$ factorization. For example, if H is diagonal (which corresponds to a separable objective function), we have $f = 0$ and $s = n$, so the total cost is $p^2n + (1/3)p^3$ flops, which grows only linearly with n. If H is banded with bandwidth $k \ll n$, then $f = nk^2$, $s = 4nk$, so the total cost is around $nk^2 + 4nkp + p^2n + (1/3)p^3$ which still grows only linearly with n. Other structures of H that can be exploited are block diagonal (which corresponds to block separable objective function), sparse, or diagonal plus low rank; see appendix C and §9.7 for more details and examples.

Example 10.3 *Equality constrained analytic center.* We consider the problem

$$\begin{array}{ll} \text{minimize} & -\sum_{i=1}^n \log x_i \\ \text{subject to} & Ax = b. \end{array}$$

Here the objective is separable, so the Hessian at x is diagonal:

$$H = \mathbf{diag}(x_1^{-2}, \ldots, x_n^{-2}).$$

If we compute the Newton direction using a generic method such as an $\mathrm{LDL}^{\mathrm{T}}$ factorization of the KKT matrix, the cost is $(1/3)(n+p)^3$ flops.

If we compute the Newton step using block elimination, the cost is $np^2 + (1/3)p^3$ flops. This is much smaller than the cost of the generic method.

In fact this cost is the same as that of computing the Newton step for the dual problem, described in example 10.2 on page 525. For the (unconstrained) dual problem, the Hessian is

$$H_{\mathrm{dual}} = -ADA^T,$$

where D is diagonal, with $D_{ii} = (A^T\nu)_i^{-2}$. Forming this matrix costs np^2 flops, and solving for the Newton step by a Cholesky factorization of $-H_{\mathrm{dual}}$ costs $(1/3)p^3$ flops.

Example 10.4 *Minimum length piecewise-linear curve subject to equality constraints.* We consider a piecewise-linear curve in $\mathbf{R}^2$ with knot points $(0,0)$, $(1,x_1)$, $\ldots$, (n,x_n). To find the minimum length curve that satisfies the equality constraints $Ax = b$, we form the problem

$$\begin{array}{ll} \text{minimize} & \left(1+x_1^2\right)^{1/2} + \sum_{i=1}^{n-1}\left(1+(x_{i+1}-x_i)^2\right)^{1/2} \\ \text{subject to} & Ax = b, \end{array}$$

with variable $x \in \mathbf{R}^n$, and $A \in \mathbf{R}^{p\times n}$. In this problem, the objective is a sum of functions of pairs of adjacent variables, so the Hessian H is tridiagonal. Using block elimination, we can compute the Newton step in around $p^2n + (1/3)p^3$ flops.

Elimination with singular H

The block elimination method described above obviously does not work when H is singular, but a simple variation on the method can be used in this more general case. The more general method is based on the following result: The KKT matrix is nonsingular if and only $H+A^TQA \succ 0$ for some $Q \succeq 0$, in which case, $H+A^TQA \succ 0$ for all $Q \succ 0$. (See exercise 10.1.) We conclude, for example, that if the KKT matrix is nonsingular, then $H + A^TA \succ 0$.

Let $Q \succeq 0$ be a matrix for which $H+A^TQA \succ 0$. Then the KKT system (10.33) is equivalent to

$$\left[\begin{array}{cc} H+A^TQA & A^T \\ A & 0 \end{array}\right]\left[\begin{array}{c} v \\ w \end{array}\right] = -\left[\begin{array}{c} g+A^TQh \\ h \end{array}\right],$$

which can be solved using elimination since $H + A^TQA \succ 0$.

10.4.3 Examples

In this section we describe some longer examples, showing how structure can be exploited to efficiently compute the Newton step. We also include some numerical results.

Equality constrained analytic centering

We consider the equality constrained analytic centering problem

$$\begin{array}{ll} \text{minimize} & f(x) = -\sum_{i=1}^n \log x_i \\ \text{subject to} & Ax = b. \end{array}$$

(See examples 10.2 and 10.3.) We compare three methods, for a problem of size $p = 100$, $n = 500$.

The first method is Newton's method with equality constraints (§10.2). The Newton step $\Delta x_{\rm nt}$ is defined by the KKT system (10.11):

$$\left[\begin{array}{cc} H & A^T \\ A & 0 \end{array}\right] \left[\begin{array}{c} \Delta x_{\rm nt} \\ w \end{array}\right] = \left[\begin{array}{c} -g \\ 0 \end{array}\right],$$

where $H = \mathbf{diag}(1/x_1^2, \ldots, 1/x_n^2)$, and $g = -(1/x_1, \ldots, 1/x_n)$. As explained in example 10.3, page 546, the KKT system can be efficiently solved by elimination, *i.e.*, by solving

$$AH^{-1}A^T w = -AH^{-1}g,$$

and setting $\Delta x_{\rm nt} = -H^{-1}(A^T w + g)$. In other words,

$$\Delta x_{\rm nt} = -\mathbf{diag}(x)^2 A^T w + x,$$

where w is the solution of

$$A\,\mathbf{diag}(x)^2 A^T w = b. \tag{10.34}$$

Figure 10.6 shows the error versus iteration. The different curves correspond to four different starting points. We use a backtracking line search with $\alpha = 0.1$, $\beta = 0.5$.

The second method is Newton's method applied to the dual

$$\text{maximize} \quad g(\nu) = -b^T\nu + \textstyle\sum_{i=1}^n \log(A^T\nu)_i + n$$

(see example 10.2, page 525). Here the Newton step is obtained from solving

$$A\,\mathbf{diag}(y)^2 A^T \Delta\nu_{\rm nt} = -b + Ay \tag{10.35}$$

where $y = (1/(A^T\nu)_1, \ldots, 1/(A^T\nu)_n)$. Comparing (10.35) and (10.34) we see that both methods have the same complexity. In figure 10.7 we show the error for four different starting points. We use a backtracking line search with $\alpha = 0.1$, $\beta = 0.5$.

The third method is the infeasible start Newton method of §10.3, applied to the optimality conditions

$$\nabla f(x^\star) + A^T\nu^\star = 0, \qquad Ax^\star = b.$$

The Newton step is obtained by solving

$$\left[\begin{array}{cc} H & A^T \\ A & 0 \end{array}\right] \left[\begin{array}{c} \Delta x_{\rm nt} \\ \Delta\nu_{\rm nt} \end{array}\right] = -\left[\begin{array}{c} g + A^T\nu \\ Ax - b \end{array}\right],$$

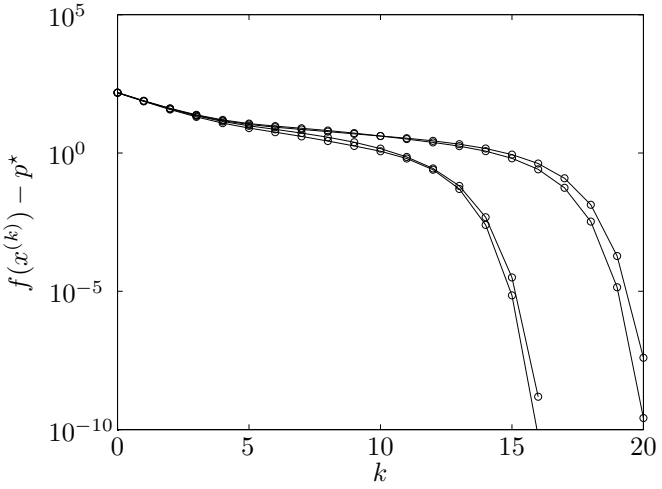

Figure 10.6 Error $f(x^{(k)}) - p^\star$ in Newton's method, applied to an equality constrained analytic centering problem of size $p = 100$, $n = 500$. The different curves correspond to four different starting points. Final quadratic convergence is clearly evident.

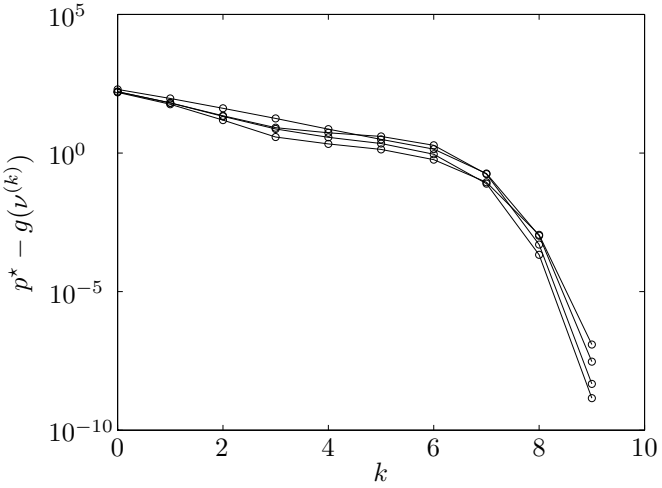

Figure 10.7 Error $|g(\nu^{(k)}) - p^\star|$ in Newton's method, applied to the dual of the equality constrained analytic centering problem.

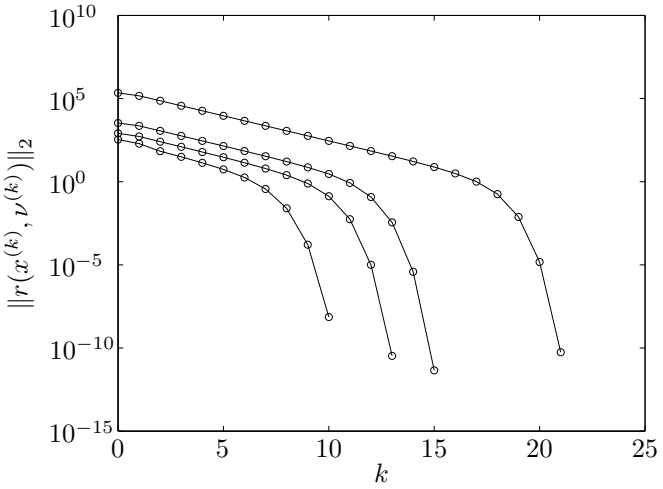

Figure 10.8 Residual $\|r(x^{(k)}, \nu^{(k)})\|_2$ in the infeasible start Newton method, applied to the equality constrained analytic centering problem.

where $H = \mathbf{diag}(1/x_1^2, \ldots, 1/x_n^2)$, and $g = -(1/x_1, \ldots, 1/x_n)$. This KKT system can be efficiently solved by elimination, at the same cost as (10.34) or (10.35). For example, if we first solve

$$A\,\mathbf{diag}(x)^2 A^T w = 2Ax - b,$$

then $\Delta\nu_{\rm nt}$ and $\Delta x_{\rm nt}$ follow from

$$\Delta\nu_{\rm nt} = w - \nu, \qquad \Delta x_{\rm nt} = x - \mathbf{diag}(x)^2 A^T w.$$

Figure 10.8 shows the norm of the residual

$$r(x, \nu) = (\nabla f(x) + A^T\nu, Ax - b)$$

versus iteration, for four different starting points. We use a backtracking line search with $\alpha = 0.1$, $\beta = 0.5$.

The figures show that for this problem, the dual method appears to be faster, but only by a factor of two or three. It takes about six iterations to reach the region of quadratic convergence, as opposed to 12–15 in the primal method and 10–20 in the infeasible start Newton method.

The methods also differ in the initialization they require. The primal method requires knowledge of a primal feasible point, *i.e.*, satisfying $Ax^{(0)} = b$, $x^{(0)} \succ 0$. The dual method requires a dual feasible point, *i.e.*, $A^T\nu^{(0)} \succ 0$. Depending on the problem, one or the other might be more readily available. The infeasible start Newton method requires no initialization; the only requirement is that $x^{(0)} \succ 0$.

Optimal network flow

We consider a connected directed graph or network with n edges and $p + 1$ nodes. We let x_j denote the flow or traffic on arc j, with $x_j > 0$ meaning flow in the

direction of the arc, and $x_j < 0$ meaning flow in the direction opposite the arc. There is also a given external source (or sink) flow s_i that enters (if $s_i > 0$) or leaves (if $s_i < 0$) node i. The flow must satisfy a conservation equation, which states that at each node, the total flow entering the node, including the external sources and sinks, is zero. This conservation equation can be expressed as $\tilde{A}x = s$ where $\tilde{A} \in \mathbf{R}^{(p+1)\times n}$ is the *node incidence matrix* of the graph,

$$\tilde{A}_{ij} = \begin{cases} 1 & \text{arc } j \text{ leaves node } i \\ -1 & \text{arc } j \text{ enters node } i \\ 0 & \text{otherwise.} \end{cases}$$

The flow conservation equation $\tilde{A}x = s$ is inconsistent unless $\mathbf{1}^T s = 0$, which we assume is the case. (In other words, the total of the source flows must equal the total of the sink flows.) The flow conservation equations $\tilde{A}x = s$ are also redundant, since $\mathbf{1}^T \tilde{A} = 0$. To obtain an independent set of equations we can delete any one equation, to obtain $Ax = b$, where $A \in \mathbf{R}^{p\times n}$ is the *reduced node incidence matrix* of the graph (*i.e.*, the node incidence matrix with one row removed) and $b \in \mathbf{R}^p$ is reduced source vector (*i.e.*, s with the associated entry removed).

In summary, flow conservation is given by $Ax = b$, where A is the reduced node incidence matrix of the graph and b is the reduced source vector. The matrix A is very sparse, since each column has at most two nonzero entries (which can only be $+1$ or -1).

We will take traffic flows x as the variables, and the sources as given. We introduce the objective function

$$f(x) = \sum_{i=1}^{n} \phi_i(x_i),$$

where $\phi_i : \mathbf{R} \to \mathbf{R}$ is the flow cost function for arc i. We assume that the flow cost functions are strictly convex and twice differentiable.

The problem of choosing the best flow, that satisfies the flow conservation requirement, is

$$\begin{array}{ll} \text{minimize} & \sum_{i=1}^{n} \phi_i(x_i) \\ \text{subject to} & Ax = b. \end{array} \tag{10.36}$$

Here the Hessian H is diagonal, since the objective is separable.

We have several choices for computing the Newton step for the optimal network flow problem (10.36). The most straightforward is to solve the full KKT system, using a sparse $\mathrm{LDL^T}$ factorization.

For this problem it is probably better to compute the Newton step using block elimination. We can characterize the sparsity pattern of the Schur complement $S = -AH^{-1}A^T$ in terms of the graph: We have $S_{ij} \neq 0$ if and only if node i and node j are connected by an arc. It follows that if the network is sparse, *i.e.*, if each node is connected by an arc to only a few other nodes, then the Schur complement S is sparse. In this case, we can exploit sparsity in forming S, and in the associated factorization and solve steps, as well. We can expect the computational complexity of computing the Newton step to grow approximately linearly with the number of arcs (which is the number of variables).

Optimal control

We consider the problem

$$\begin{array}{ll}\text{minimize} & \sum_{t=1}^{N}\phi_t(z(t)) + \sum_{t=0}^{N-1}\psi_t(u(t)) \\ \text{subject to} & z(t+1) = A_t z(t) + B_t u(t), \quad t = 0, \ldots, N-1.\end{array}$$

Here

- $z(t) \in \mathbf{R}^k$ is the system state at time t
- $u(t) \in \mathbf{R}^l$ is the input or control action at time t
- $\phi_t : \mathbf{R}^k \to \mathbf{R}$ is the state cost function
- $\psi_t : \mathbf{R}^l \to \mathbf{R}$ is the input cost function
- N is called the *time horizon* for the problem.

We assume that the input and state cost functions are strictly convex and twice differentiable. The variables in the problem are $u(0), \ldots, u(N-1)$, and $z(1), \ldots, z(N)$. The initial state $z(0)$ is given. The linear equality constraints are called the *state equations* or *dynamic evolution equations*. We define the overall optimization variable x as

$$x = (u(0), z(1), u(1), \ldots, u(N-1), z(N)) \in \mathbf{R}^{N(k+l)}.$$

Since the objective is block separable (*i.e.*, a sum of functions of $z(t)$ and $u(t)$), the Hessian is block diagonal:

$$H = \mathbf{diag}(R_0, Q_1, \ldots, R_{N-1}, Q_N),$$

where

$$R_t = \nabla^2\psi_t(u(t)), \quad t = 0, \ldots, N-1, \qquad Q_t = \nabla^2\phi_t(z(t)), \quad t = 1, \ldots, N.$$

We can collect all the equality constraints (*i.e.*, the state equations) and express them as $Ax = b$ where

$$A = \begin{bmatrix} -B_0 & I & 0 & 0 & 0 & \cdots & 0 & 0 & 0 \\ 0 & -A_1 & -B_1 & I & 0 & \cdots & 0 & 0 & 0 \\ 0 & 0 & 0 & -A_2 & -B_2 & \cdots & 0 & 0 & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots & & \vdots & \vdots & \vdots \\ 0 & 0 & 0 & 0 & 0 & \cdots & I & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & \cdots & -A_{N-1} & -B_{N-1} & I \end{bmatrix}$$

$$b = \begin{bmatrix} A_0 z(0) \\ 0 \\ 0 \\ \vdots \\ 0 \\ 0 \end{bmatrix}.$$

The number of rows of A (*i.e.*, equality constraints) is Nk.

Directly solving the KKT system for the Newton step, using a dense $\mathrm{LDL}^{\mathrm{T}}$ factorization, would cost

$$(1/3)(2Nk + Nl)^3 = (1/3)N^3(2k + l)^3$$

flops. Using a sparse $\mathrm{LDL}^{\mathrm{T}}$ factorization would give a large improvement, since the method would exploit the many zero entries in A and H.

In fact we can do better by exploiting the special block structure of H and A, using block elimination to compute the Newton step. The Schur complement $S = -AH^{-1}A^T$ turns out to be block tridiagonal, with $k \times k$ blocks:

$$\begin{aligned} S &= -AH^{-1}A^T \\ &= \begin{bmatrix} S_{11} & Q_1^{-1}A_1^T & 0 & \cdots & 0 & 0 \\ A_1Q_1^{-1} & S_{22} & Q_2^{-1}A_2^T & \cdots & 0 & 0 \\ 0 & A_2Q_2^{-1} & S_{33} & \cdots & 0 & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & 0 & \cdots & S_{N-1,N-1} & Q_{N-1}^{-1}A_{N-1}^T \\ 0 & 0 & 0 & \cdots & A_{N-1}Q_{N-1}^{-1} & S_{NN} \end{bmatrix} \end{aligned}$$

where

$$\begin{aligned} S_{11} &= -B_0R_0^{-1}B_0^T - Q_1^{-1}, \\ S_{ii} &= -A_{i-1}Q_{i-1}^{-1}A_{i-1}^T - B_{i-1}R_{i-1}^{-1}B_{i-1}^T - Q_i^{-1}, \quad i = 2, \ldots, N. \end{aligned}$$

In particular, S is banded, with bandwidth $2k - 1$, so we can factor it in order k^3N flops. Therefore we can compute the Newton step in order k^3N flops, assuming $k \ll N$. Note that this grows linearly with the time horizon N, whereas for a generic method, the flop count grows like N^3.

For this problem we could go one step further and exploit the block tridiagonal structure of S. Applying a standard block tridiagonal factorization method would result in the classic Riccati recursion for solving a quadratic optimal control problem. Still, using only the banded nature of S yields an algorithm that is the same order.

Analytic center of a linear matrix inequality

We consider the problem

$$\begin{array}{ll} \text{minimize} & f(X) = -\log\det X \\ \text{subject to} & \mathbf{tr}(A_iX) = b_i, \quad i = 1, \ldots, p, \end{array} \tag{10.37}$$

where $X \in \mathbf{S}^n$ is the variable, $A_i \in \mathbf{S}^n$, $b_i \in \mathbf{R}$, and $\mathbf{dom}\, f = \mathbf{S}^n_{++}$. The KKT conditions for this problem are

$$-X^{\star -1} + \sum_{i=1}^m \nu_i^\star A_i = 0, \qquad \mathbf{tr}(A_iX^\star) = b_i, \quad i = 1, \ldots, p. \tag{10.38}$$

The dimension of the variable X is $n(n + 1)/2$. We could simply ignore the special matrix structure of X, and consider it as (vector) variable $x \in \mathbf{R}^{n(n+1)/2}$,

and solve the problem (10.37) using a generic method for a problem with $n(n+1)/2$ variables and p equality constraints. The cost for computing a Newton step would then be at least

$$(1/3)(n(n+1)/2+p)^3$$

flops, which is order n^6 in n. We will see that there are a number of far more attractive alternatives.

A first option is to solve the dual problem. The conjugate of f is

$$f^*(Y) = \log\det(-Y)^{-1} - n$$

with $\mathbf{dom}\, f^* = -\mathbf{S}^n_{++}$ (see example 3.23, page 92), so the dual problem is

$$\text{maximize} \quad -b^T\nu + \log\det(\textstyle\sum_{i=1}^p \nu_i A_i) + n, \tag{10.39}$$

with domain $\{\nu \mid \sum_{i=1}^p \nu_i A_i \succ 0\}$. This is an unconstrained problem with variable $\nu \in \mathbf{R}^p$. The optimal $X^\star$ can be recovered from the optimal $\nu^\star$ by solving the first (dual feasibility) equation in (10.38), *i.e.*, $X^\star = (\sum_{i=1}^p \nu_i^\star A_i)^{-1}$.

Let us work out the cost of computing the Newton step for the dual problem (10.39). We have to form the gradient and Hessian of g, and then solve for the Newton step. The gradient and Hessian are given by

$$\begin{aligned} \nabla^2 g(\nu)_{ij} &= -\mathbf{tr}(A^{-1}A_iA^{-1}A_j), \quad i,\ j = 1,\ldots,p, \\ \nabla g(\nu)_i &= \mathbf{tr}(A^{-1}A_i) - b_i, \quad i = 1\ldots,p, \end{aligned}$$

where $A = \sum_{i=1}^p \nu_i A_i$. To form $\nabla^2 g(\nu)$ and $\nabla g(\nu)$ we proceed as follows. We first form A (pn^2 flops), and $A^{-1}A_j$ for each j ($2pn^3$ flops). Then we form the matrix $\nabla^2 g(\nu)$. Each of the $p(p+1)/2$ entries of $\nabla^2 g(\nu)$ is the inner product of two matrices in $\mathbf{S}^n$, each of which costs $n(n+1)$ flops, so the total is (dropping dominated terms) $(1/2)p^2n^2$ flops. Forming $\nabla g(\nu)$ is cheap since we already have the matrices $A^{-1}A_i$. Finally, we solve for the Newton step $-\nabla^2 g(\nu)^{-1}\nabla g(\nu)$, which costs $(1/3)p^3$ flops. All together, and keeping only the leading terms, the total cost of computing the Newton step is $2pn^3 + (1/2)p^2n^2 + (1/3)p^3$. Note that this is order n^3 in n, which is far better than the simple primal method described above, which is order n^6.

We can also solve the primal problem more efficiently, by exploiting its special matrix structure. To derive the KKT system for the Newton step $\Delta X_{\rm nt}$ at a feasible X, we replace $X^\star$ in the KKT conditions by $X + \Delta X_{\rm nt}$ and $\nu^\star$ by w, and linearize the first equation using the first-order approximation

$$(X + \Delta X_{\rm nt})^{-1} \approx X^{-1} - X^{-1}\Delta X_{\rm nt} X^{-1}.$$

This gives the KKT system

$$-X^{-1} + X^{-1}\Delta X_{\rm nt}X^{-1} + \sum_{i=1}^p w_i A_i = 0, \qquad \mathbf{tr}(A_i\Delta X_{\rm nt}) = 0, \quad i = 1,\ldots,p. \tag{10.40}$$

This is a set of $n(n+1)/2 + p$ linear equations in the variables $\Delta X_{\rm nt} \in \mathbf{S}^n$ and $w \in \mathbf{R}^p$. If we solved these equations using a generic method, the cost would be order n^6.

We can use block elimination to solve the KKT system (10.40) far more efficiently. We eliminate the variable $\Delta X_{\rm nt}$, by solving the first equation to get

$$\Delta X_{\rm nt} = X - X\left(\sum_{i=1}^{p} w_i A_i\right) X = X - \sum_{i=1}^{p} w_i X A_i X. \tag{10.41}$$

Substituting this expression for $\Delta X_{\rm nt}$ into the other equation gives

$$\mathbf{tr}(A_j \Delta X_{\rm nt}) = \mathbf{tr}(A_j X) - \sum_{i=1}^{p} w_i \,\mathbf{tr}(A_j X A_i X) = 0, \quad j = 1, \ldots, p.$$

This is a set of p linear equations in w:

$$Cw = d$$

where $C_{ij} = \mathbf{tr}(A_i X A_j X)$, $d_i = \mathbf{tr}(A_i X)$. The coefficient matrix C is symmetric and positive definite, so a Cholesky factorization can be used to find w. Once we have w, we can compute $\Delta X_{\rm nt}$ from (10.41).

The cost of this method is as follows. We form the products $A_i X$ ($2pn^3$ flops), and then form the matrix C. Each of the $p(p+1)/2$ entries of C is the inner product of two matrices in $\mathbf{R}^{n\times n}$, so forming C costs p^2n^2 flops. Then we solve for $w = C^{-1}d$, which costs $(1/3)p^3$. Finally we compute $\Delta X_{\rm nt}$. If we use the first expression in (10.41), *i.e.*, first compute the sum and then pre- and post-multiply with X, the cost is approximately $pn^2 + 3n^3$. All together, the total cost is $2pn^3 + p^2n^2 + (1/3)p^3$ flops to form the Newton step for the primal problem, using block elimination. This is far better than the simple method, which is order n^6. Note also that the cost is the same as that of computing the Newton step for the dual problem.

Bibliography

The two key assumptions in our analysis of the infeasible start Newton method (the derivative Dr has a bounded inverse and satisfies a Lipschitz condition) are central to most convergence proofs of Newton's method; see Ortega and Rheinboldt [OR00] and Dennis and Schnabel [DS96].

The relative merits of solving KKT systems via direct factorization of the full system, or via elimination, have been extensively studied in the context of interior-point methods for linear and quadratic programming; see, for example, Wright [Wri97, chapter 11] and Nocedal and Wright [NW99, §16.1-2]. The Riccati recursion from optimal control can be interpreted as a method for exploiting the block tridiagonal structure in the Schur complement S of the example on page 552. This observation was made by Rao, Wright, and Rawlings [RWR98, §3.3].

Exercises

Equality constrained minimization

10.1 *Nonsingularity of the KKT matrix.* Consider the KKT matrix

$$\begin{bmatrix} P & A^T \\ A & 0 \end{bmatrix},$$

where $P \in \mathbf{S}^n_+$, $A \in \mathbf{R}^{p\times n}$, and $\mathbf{rank}\, A = p < n$.

(a) Show that each of the following statements is equivalent to nonsingularity of the KKT matrix.

- $\mathcal{N}(P) \cap \mathcal{N}(A) = \{0\}$.
- $Ax = 0$, $x \neq 0 \Longrightarrow x^T P x > 0$.
- $F^T P F \succ 0$, where $F \in \mathbf{R}^{n\times(n-p)}$ is a matrix for which $\mathcal{R}(F) = \mathcal{N}(A)$.
- $P + A^T Q A \succ 0$ for some $Q \succeq 0$.

(b) Show that if the KKT matrix is nonsingular, then it has exactly n positive and p negative eigenvalues.

10.2 *Projected gradient method.* In this problem we explore an extension of the gradient method to equality constrained minimization problems. Suppose f is convex and differentiable, and $x \in \mathbf{dom}\, f$ satisfies $Ax = b$, where $A \in \mathbf{R}^{p\times n}$ with $\mathbf{rank}\, A = p < n$. The Euclidean projection of the negative gradient $-\nabla f(x)$ on $\mathcal{N}(A)$ is given by

$$\Delta x_{\rm pg} = \operatorname*{argmin}_{Au=0} \|-\nabla f(x) - u\|_2.$$

(a) Let (v, w) be the unique solution of

$$\begin{bmatrix} I & A^T \\ A & 0 \end{bmatrix} \begin{bmatrix} v \\ w \end{bmatrix} = \begin{bmatrix} -\nabla f(x) \\ 0 \end{bmatrix}.$$

Show that $v = \Delta x_{\rm pg}$ and $w = \operatorname{argmin}_y \|\nabla f(x) + A^T y\|_2$.

(b) What is the relation between the projected negative gradient $\Delta x_{\rm pg}$ and the negative gradient of the reduced problem (10.5), assuming $F^T F = I$?

(c) The *projected gradient method* for solving an equality constrained minimization problem uses the step $\Delta x_{\rm pg}$, and a backtracking line search on f. Use the results of part (b) to give some conditions under which the projected gradient method converges to the optimal solution, when started from a point $x^{(0)} \in \mathbf{dom}\, f$ with $Ax^{(0)} = b$.

Newton's method with equality constraints

10.3 *Dual Newton method.* In this problem we explore Newton's method for solving the dual of the equality constrained minimization problem (10.1). We assume that f is twice differentiable, $\nabla^2 f(x) \succ 0$ for all $x \in \mathbf{dom}\, f$, and that for each $\nu \in \mathbf{R}^p$, the Lagrangian $L(x, \nu) = f(x) + \nu^T(Ax - b)$ has a unique minimizer, which we denote $x(\nu)$.

(a) Show that the dual function g is twice differentiable. Find an expression for the Newton step for the dual function g, evaluated at ν, in terms of f, ∇f, and $\nabla^2 f$, evaluated at $x = x(\nu)$. You can use the results of exercise 3.40.

(b) Suppose there exists a K such that

$$\left\| \begin{bmatrix} \nabla^2 f(x) & A^T \\ A & 0 \end{bmatrix}^{-1} \right\|_2 \leq K$$

for all $x \in \mathbf{dom}\, f$. Show that g is strongly concave, with $\nabla^2 g(\nu) \preceq -(1/K)I$.

10.4 *Strong convexity and Lipschitz constant of the reduced problem.* Suppose f satisfies the assumptions given on page 529. Show that the reduced objective function $\tilde{f}(z) = f(Fz+\hat{x})$ is strongly convex, and that its Hessian is Lipschitz continuous (on the associated sublevel set $\tilde{S}$). Express the strong convexity and Lipschitz constants of $\tilde{f}$ in terms of K, M, L, and the maximum and minimum singular values of F.

10.5 *Adding a quadratic term to the objective.* Suppose $Q \succeq 0$. The problem

$$\begin{array}{ll} \text{minimize} & f(x) + (Ax-b)^T Q(Ax-b) \\ \text{subject to} & Ax = b \end{array}$$

is equivalent to the original equality constrained optimization problem (10.1). Is the Newton step for this problem the same as the Newton step for the original problem?

10.6 *The Newton decrement.* Show that (10.13) holds, *i.e.*,

$$f(x) - \inf\{\widehat{f}(x+v) \mid A(x+v) = b\} = \lambda(x)^2/2.$$

Infeasible start Newton method

10.7 *Assumptions for infeasible start Newton method.* Consider the set of assumptions given on page 536.

(a) Suppose that the function f is closed. Show that this implies that the norm of the residual, $\|r(x,\nu)\|_2$, is closed.

(b) Show that Dr satisfies a Lipschitz condition if and only if $\nabla^2 f$ does.

10.8 *Infeasible start Newton method and initially satisfied equality constraints.* Suppose we use the infeasible start Newton method to minimize $f(x)$ subject to $a_i^T x = b_i$, $i = 1, \ldots, p$.

(a) Suppose the initial point $x^{(0)}$ satisfies the linear equality $a_i^T x = b_i$. Show that the linear equality will remain satisfied for future iterates, *i.e.*, if $a_i^T x^{(k)} = b_i$ for all k.

(b) Suppose that one of the equality constraints becomes satisfied at iteration k, *i.e.*, we have $a_i^T x^{(k-1)} \neq b_i$, $a_i^T x^{(k)} = b_i$. Show that at iteration k, *all* the equality constraints are satisfied.

10.9 *Equality constrained entropy maximization.* Consider the equality constrained entropy maximization problem

$$\begin{array}{ll} \text{minimize} & f(x) = \sum_{i=1}^n x_i \log x_i \\ \text{subject to} & Ax = b, \end{array} \tag{10.42}$$

with $\mathbf{dom}\, f = \mathbf{R}^n_{++}$ and $A \in \mathbf{R}^{p\times n}$. We assume the problem is feasible and that $\mathbf{rank}\, A = p < n$.

(a) Show that the problem has a unique optimal solution $x^\star$.

(b) Find A, b, and feasible $x^{(0)}$ for which the sublevel set

$$\{x \in \mathbf{R}^n_{++} \mid Ax = b,\ f(x) \leq f(x^{(0)})\}$$

is *not* closed. Thus, the assumptions listed in §10.2.4, page 529, are not satisfied for some feasible initial points.

(c) Show that the problem (10.42) satisfies the assumptions for the infeasible start Newton method listed in §10.3.3, page 536, for any feasible starting point.

(d) Derive the Lagrange dual of (10.42), and explain how to find the optimal solution of (10.42) from the optimal solution of the dual problem. Show that the dual problem satisfies the assumptions listed in §10.2.4, page 529, for *any* starting point.

The results of part (b), (c), and (d) do not mean the standard Newton method will fail, or that the infeasible start Newton method or dual method will work better in practice. It only means our convergence analysis for the standard Newton method does not apply, while our convergence analysis does apply to the infeasible start and dual methods. (See exercise 10.15.)

10.10 *Bounded inverse derivative condition for strongly convex-concave game.* Consider a convex-concave game with payoff function f (see page 541). Suppose $\nabla^2_{uu} f(u,v) \succeq mI$ and $\nabla^2_{vv} f(u,v) \preceq -mI$, for all $(u,v) \in \mathbf{dom}\, f$. Show that

$$\|Dr(u,v)^{-1}\|_2 = \|\nabla^2 f(u,v)^{-1}\|_2 \leq 1/m.$$

Implementation

10.11 Consider the resource allocation problem described in example 10.1. You can assume the f_i are strongly convex, *i.e.*, $f_i''(z) \geq m > 0$ for all z.

(a) Find the computational effort required to compute a Newton step for the reduced problem. Be sure to exploit the special structure of the Newton equations.

(b) Explain how to solve the problem via the dual. You can assume that the conjugate functions f_i^*, and their derivatives, are readily computable, and that the equation $f_i'(x) = \nu$ is readily solved for x, given ν. What is the computational complexity of finding a Newton step for the dual problem?

(c) What is the computational complexity of computing a Newton step for the resource allocation problem? Be sure to exploit the special structure of the KKT equations.

10.12 Describe an efficient way to compute the Newton step for the problem

$$\begin{array}{ll} \text{minimize} & \mathbf{tr}(X^{-1}) \\ \text{subject to} & \mathbf{tr}(A_i X) = b_i, \quad i = 1, \ldots, p \end{array}$$

with domain $\mathbf{S}^n_{++}$, assuming p and n have the same order of magnitude. Also derive the Lagrange dual problem and give the complexity of finding the Newton step for the dual problem.

10.13 *Elimination method for computing Newton step for convex-concave game.* Consider a convex-concave game with payoff function $f : \mathbf{R}^p \times \mathbf{R}^q \to \mathbf{R}$ (see page 541). We assume that f is *strongly convex-concave*, *i.e.*, for all $(u,v) \in \mathbf{dom}\, f$ and some $m > 0$, we have $\nabla^2_{uu} f(u,v) \succeq mI$ and $\nabla^2_{vv} f(u,v) \preceq -mI$.

(a) Show how to compute the Newton step using Cholesky factorizations of $\nabla^2_{uu} f(u,v)$ and $-\nabla^2 f_{vv}(u,v)$. Compare the cost of this method with the cost of using an $\mathrm{LDL}^{\mathrm{T}}$ factorization of $\nabla f(u,v)$, assuming $\nabla^2 f(u,v)$ is dense.

(b) Show how you can exploit diagonal or block diagonal structure in $\nabla^2_{uu} f(u,v)$ and/or $\nabla^2_{vv} f(u,v)$. How much do you save, if you assume $\nabla^2_{uv} f(u,v)$ is dense?

Numerical experiments

10.14 *Log-optimal investment.* Consider the log-optimal investment problem described in exercise 4.60. Use Newton's method to compute the solution, with the following problem

data: there are $n = 3$ assets, and $m = 4$ scenarios, with returns

$$p_1 = \begin{bmatrix} 2 \\ 1.3 \\ 1 \end{bmatrix}, \quad p_2 = \begin{bmatrix} 2 \\ 0.5 \\ 1 \end{bmatrix}, \quad p_3 = \begin{bmatrix} 0.5 \\ 1.3 \\ 1 \end{bmatrix}, \quad p_4 = \begin{bmatrix} 0.5 \\ 0.5 \\ 1 \end{bmatrix}.$$

The probabilities of the four scenarios are given by $\pi = (1/3, 1/6, 1/3, 1/6)$.

10.15 *Equality constrained entropy maximization.* Consider the equality constrained entropy maximization problem

$$\begin{array}{ll} \text{minimize} & f(x) = \sum_{i=1}^n x_i \log x_i \\ \text{subject to} & Ax = b, \end{array}$$

with $\mathbf{dom}\, f = \mathbf{R}^n_{++}$ and $A \in \mathbf{R}^{p\times n}$, with $p < n$. (See exercise 10.9 for some relevant analysis.)

Generate a problem instance with $n = 100$ and $p = 30$ by choosing A randomly (checking that it has full rank), choosing $\hat{x}$ as a random positive vector (*e.g.*, with entries uniformly distributed on $[0, 1]$) and then setting $b = A\hat{x}$. (Thus, $\hat{x}$ is feasible.)

Compute the solution of the problem using the following methods.

(a) *Standard Newton method.* You can use initial point $x^{(0)} = \hat{x}$.

(b) *Infeasible start Newton method.* You can use initial point $x^{(0)} = \hat{x}$ (to compare with the standard Newton method), and also the initial point $x^{(0)} = \mathbf{1}$.

(c) *Dual Newton method*, *i.e.*, the standard Newton method applied to the dual problem.

Verify that the three methods compute the same optimal point (and Lagrange multiplier). Compare the computational effort per step for the three methods, assuming relevant structure is exploited. (Your implementation, however, does not need to exploit structure to compute the Newton step.)

10.16 *Convex-concave game.* Use the infeasible start Newton method to solve convex-concave games of the form (10.32), with randomly generated data. Plot the norm of the residual and step length versus iteration. Experiment with the line search parameters and initial point (which must satisfy $\|u\|_2 < 1$, $\|v\|_2 < 1$, however).