Applied Optimization: Formulation and Algorithms for Engineering Systems Slides

Ross Baldick
Department of Electrical and Computer Engineering
The University of Texas at Austin
Austin, TX 78712

Part IV

Equality-constrained optimization

12

Case studies of equality-constrained optimization

(i) Production, at least-cost, of a commodity from machines, while meeting a total demand (Section 12.1), and
(ii) State estimation in an electric power system where the power injections at some of the buses are known to high accuracy (Section 12.2).

12.1 Least-cost production

12.1.1 Motivation

- Consider a machine that makes a certain product, requiring some costly input to produce.
- If **stock-piling** of the product is costly or inconvenient or if demand for the product varies rapidly, then to avoid over-supplies and shortages we must vary production to follow variations in demand.
- An extreme example of this problem is in the production of electricity.
- Typically the fuel cost is non-zero and it is not practical to stock-pile electrical energy over even very short periods.
- Electric generators also have efficiencies that vary markedly with output.
- In electric power, the problem of least-cost production is called **economic dispatch.**

12.1.2 Formulation

12.1.2.1 Variables

- Suppose that we own n machines or plants that are producing a commodity or product.
- We consider the production over a particular period of time.
- The length T of this period of time should be chosen to be short enough so that the production *per unit time* for the commodity or product by each machine can be well approximated by a constant over the time period T.
- That is, we are assuming that the plant is in **quasi-steady state**.
- Define $x_k \in \mathbb{R}$ to be the total amount of the commodity produced by machine k over the time period.
- We collect the production decisions of machines $k = 1, \ldots, n$, into a vector $x \in \mathbb{R}^n$, so that $x = \begin{bmatrix} x_1 \\ \vdots \\ x_n \end{bmatrix}$.

12.1.2.2 Production costs

- We suppose that for $k = 1, \ldots, n$ there are functions $f_k : \mathbb{R} \to \mathbb{R}$ such that $f_k(x_k)$ is the cost for machine k to produce x_k over the time period T.

12.1.2.3 Objective

- We want to minimize the objective $f : \mathbb{R}^n \to \mathbb{R}$ defined by:

$$\forall x \in \mathbb{R}^n, f(x) = \sum_{k=1}^{n} f_k(x_k). \tag{12.1}$$

Machine

- We assume that machine k has:

  a maximum production capacity, say $\overline{x}_k$, and
  a minimum production capacity, $\underline{x}_k \geq 0$.

$$\underline{x}_k \leq x_k \leq \overline{x}_k. \tag{12.2}$$

- The feasible operating set for machine k is therefore:

$$\mathbb{S}_k = \{0\} \cup [\underline{x}_k, \overline{x}_k].$$

- The set $\mathbb{S}_k$ is not convex if $\underline{x}_k > 0$.
- In specifying (12.1) we assumed that each function f_k was defined on the whole of $\mathbb{R}$; however, only the values of f_k on $\mathbb{S}_k$ are relevant to the solution of the problem.
- In defining f, we have implicitly extrapolated the cost function of each machine from its operating range, as specified by $\mathbb{S}_k$, to the whole of $\mathbb{R}$.

Production

- Let us assume that during the time period T we face a total demand for the commodity of quantity D.
- To meet demand, we must satisfy the constraint:

$$D = \sum_{k=1}^{n} x_k. \tag{12.3}$$

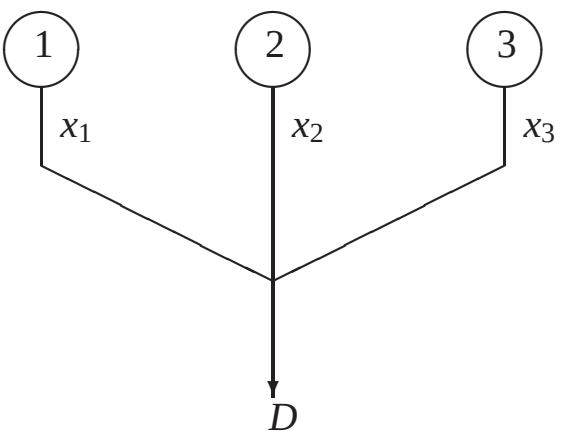

Fig. 12.1. Production from three machines.

Production, continued

- We can write the constraint in the form $Ax = b$ with either of the following two choices for $A \in \mathbb{R}^{1\times n}$ and $b \in \mathbb{R}$:

  $A = \mathbf{1}^\dagger, b = [D]$, or
  $A = -\mathbf{1}^\dagger, b = [-D]$.
- For reasons that will be made clear in Section 13.5 when we discuss an economic interpretation of the problem, we prefer to use the second choice for A and b.

Machine and production combined

- The feasible operating set for all the machines is: $(\prod_{k=1}^{n} \mathbb{S}_k) \subset \mathbb{R}^n$, where the symbol $\prod$ means the **Cartesian product**, so that the feasible set for the problem is:

$$\underline{\mathbb{S}} = \left(\prod_{k=1}^{n} \mathbb{S}_k \right) \cap \{x \in \mathbb{R}^n | Ax = b\}.$$

Relaxation

- For the discussion in this chapter, however, we are going to:
  - assume that each machine is in-service and operating, and
  - ignore minimum and maximum production capacity constraints.
- That is, we are going to relax the set of feasible operating points for machine k from the set $\mathbb{S}_k$ to the whole of $\mathbb{R}$ and correspondingly relax the feasible set for the problem from $\underline{\mathbb{S}}$ to:

$$\mathbb{S} = \{x \in \mathbb{R}^n | Ax = b\}.$$

Relaxation, continued

- Part of the feasible set $\mathbb{S}$ lying in the non-negative orthant is illustrated in Figure 12.2 for $n = 3$ and $D = 10$.

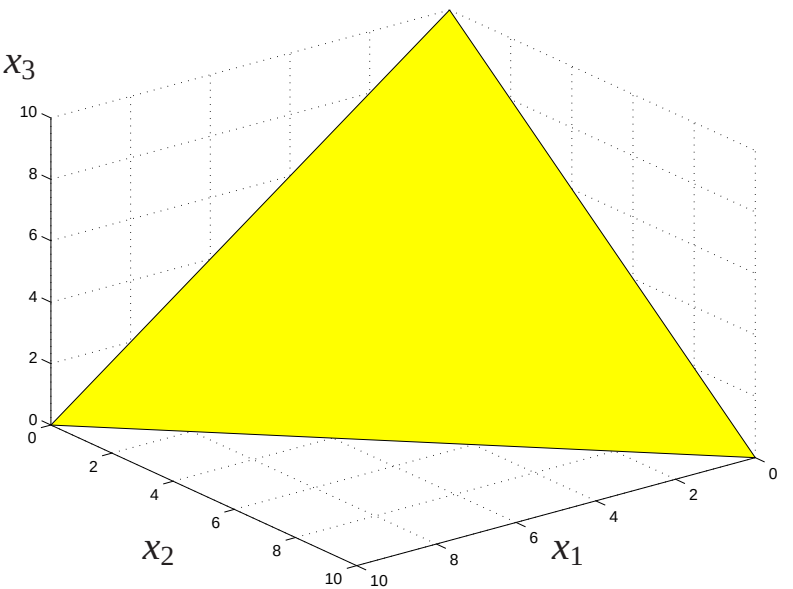

Fig. 12.2. Part of feasible set $\mathbb{S}$ for least-cost production case study.

12.1.2.5 Problem

- Our relaxed optimization problem is:

$$\min_{x \in \mathbb{R}^n} \{ f(x) | Ax = b \} . \tag{12.4}$$

- We have implicitly assumed that each function f_k has been **extrapolated** to being a function defined on the whole of $\mathbb{R}$.

12.1.2.6 Alternative formulation

- If the cost function for each machine increases monotonically with production, we could also consider solving the inequality-constrained problem:

$$\min_{x \in \mathbb{R}^n} \left\{ f(x) \,\middle|\, D \leq \sum_{k=1}^{n} x_k \right\}, \tag{12.5}$$

- which is a further relaxation of our constraints, but which has the same minimum and minimizer as Problem (12.4) if costs are strictly monotonically increasing.
- The flexibility in the choice of formulation can sometimes be useful in adapting a problem formulation to an algorithm or in proving results about the problem.
- However, in this chapter we will only consider the equality-constrained version, Problem (12.4).

12.1.2.7 Discussion

- Suppose that the solution $x^\star$ of the relaxed Problem (12.4) happens to satisfy the omitted minimum and maximum capacity constraints (12.2).
- That is, $x_k^\star \in \mathbb{S}_k$.
- Then the solution of the relaxed Problem (12.4) is optimal for the complete problem including the machine constraints:

$$\min_{x \in \underline{\mathbb{S}}} f(x).$$

- If the omitted constraints are not satisfied, then we must consider them explicitly.
- We will explicitly consider inequality constraints such as the minimum and maximum production capacity constraints (12.2) in Part V, but the feasible set $\mathbb{S}_k$ for machine k is non-convex since it includes the points 0 and $\underline{x}_k$ but not any points between 0 and $\underline{x}_k$.

12.1.3 Change in demand

- We can expect that demand will change over time.
- Consequently, it is important to be able to estimate the change in the costs due to a change in demand from D to $D + \Delta D$, say.

12.1.4 Problem characteristics

12.1.4.1 Objective

Separability

- It is expressed as the sum of functions, f_k, each of which depends only on a single entry, x_k, of x.
- That is, the objective is **additively separable**.

Average production costs

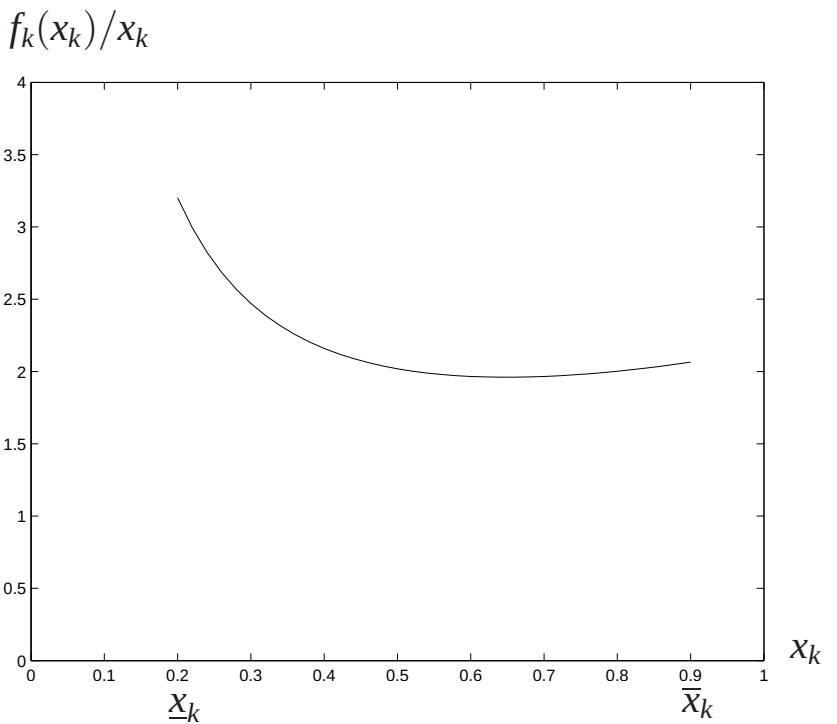

Fig. 12.3. The average production cost $f_k(x_k)/x_k$ versus production x_k for a typical machine for $\underline{x}_k \leq x_k \leq \overline{x}_k$.

Average production costs, continued

- Consider the **average cost per unit of production** $f_k(x_k)/x_k$ for machine k producing x_k.
- At low levels of production, we would expect the average production cost to be relatively high.
- This is because there are usually costs that must be incurred whenever the plant is in-service and producing non-zero levels of output.
- As x_k increases from low levels, the average production costs typically decrease because the costs of operating the auxiliary equipment are averaged over a greater amount of production.
- For some x_k, the average costs $f_k(x_k)/x_k$ reach a minimum and then begin to increase again for larger values of x_k.
- The point where $f_k(x_k)/x_k$ is at a minimum is the point of maximum efficiency of the machine.

Production costs

- If we multiply the values of $f_k(x_k)/x_k$ in Figure 12.3 by x_k, we obtain the production costs $f_k(x_k)$ as illustrated in Figure 12.4.

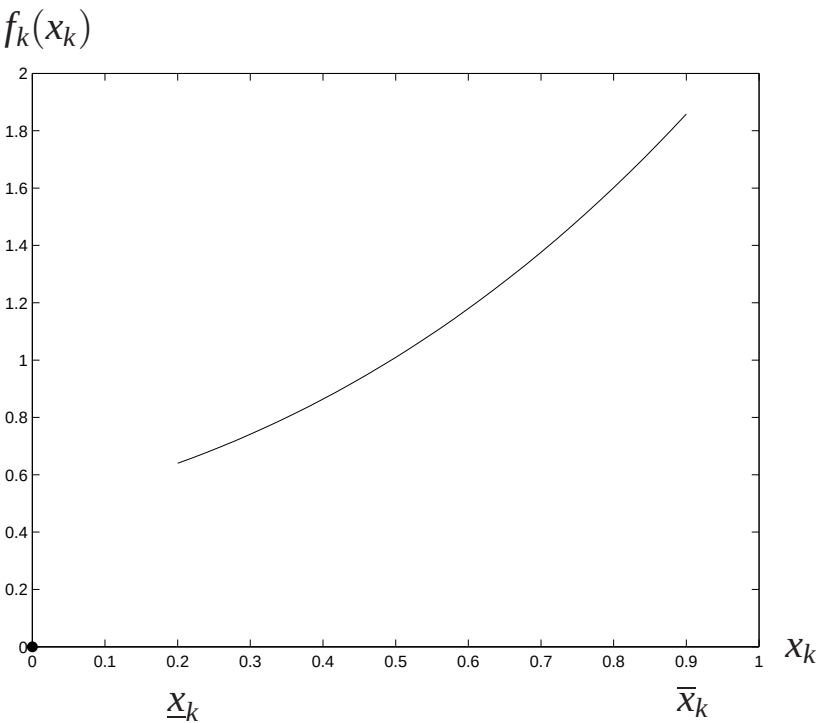

Fig. 12.4. Production cost $f_k(x_k)$ versus production x_k for a typical machine.

Production costs, continued

- Extrapolating the shape of f_k from $\underline{x}_k$ to values $x_k < \underline{x}_k$ we find that at $x_k = 0$ the extrapolated value of the production cost function would be greater than zero due to the auxiliary operating costs.

Convexity

- If $\underline{x}_k > 0$ then $\mathbb{S}_k = \{0\} \cup [\underline{x}_k, \overline{x}_k]$ is not convex.
- If $\underline{x}_k = 0$ then $\mathbb{S}_k = \{0\} \cup [\underline{x}_k, \overline{x}_k] = [\underline{x}_k, \overline{x}_k] = [0, \overline{x}_k]$ *is* convex.
- Even in this case, however, if there are non-zero auxiliary operating costs then f_k is not a convex function on $[0, \overline{x}_k]$ because of the discontinuity in f_k.

Convexity, continued

- To identify a test set on which the objective might be convex, suppose that:

  $\underline{x}_k = 0$ and consider the set $\underline{\mathbb{S}}_k = \{x_k \in \mathbb{R} | 0 < x_k \leq \overline{x}_k\} \subset \mathbb{S}_k$, or
  $\underline{x}_k > 0$ and consider the set $\underline{\mathbb{S}}_k = \{x_k \in \mathbb{R} | \underline{x}_k \leq x_k \leq \overline{x}_k\} \subset \mathbb{S}_k$.
- In both cases, $\underline{\mathbb{S}}_k$ is a convex set.
- Moreover, for both these cases, Figure 12.4 suggests that f_k is convex on $\underline{\mathbb{S}}_k$.
- We will assume that the cost function of each machine has been extrapolated to a function that is convex on the *whole* of $\mathbb{R}$.
- We have effectively redefined $f_k(0)$.
- It is often reasonable to assume that $f_k : \underline{\mathbb{S}}_k \to \mathbb{R}$ is quadratic:

$$\forall x_k \in \underline{\mathbb{S}}_k, f_k(x_k) = \frac{1}{2} Q_{kk} (x_k)^2 + c_k x_k + d_k. \tag{12.6}$$

- For convex costs, $Q_{kk} \geq 0$.
- With non-zero auxiliary costs, $d_k > 0$.
- We also usually expect that $c_k > 0$.

Convexity, continued

- Adding together the cost functions for all machines, we obtain:

$$\forall x \in \mathbb{R}^n, f(x) = \frac{1}{2} x^\dagger Q x + c^\dagger x + d,$$

- where $Q \in \mathbb{R}^{n \times n}$ is a diagonal matrix with k-th diagonal entry equal to Q_{kk},
- $c \in \mathbb{R}^n$ has k-th entry equal to c_k, and
- $d = \sum_{k=1}^{n} d_k \in \mathbb{R}$.

Eliminating a variable

- By Corollary 3.7, we can use the equality constraint $Ax = b$ to eliminate one of the variables, say x_1, by writing:

$$x_1 = D - x_2 - \cdots - x_n.$$

- Expressing the objective in terms of $x_2, \ldots, x_n$ yields an unconstrained problem with objective $f(\tilde{x})$ where:

$$\begin{aligned} \tilde{x} &= \begin{bmatrix} D - x_2 - \cdots - x_n \\ x_2 \\ \vdots \\ x_n \end{bmatrix} = \begin{bmatrix} D \\ \mathbf{0} \end{bmatrix} + \begin{bmatrix} -\mathbf{1}^\dagger \\ \mathbf{I} \end{bmatrix} \begin{bmatrix} x_2 \\ \vdots \\ x_n \end{bmatrix}, \\ &= \hat{x} + Z\xi, \\ &= \tau(\xi), \end{aligned}$$

Eliminating a variable, continued

- where:

$$\begin{aligned} \hat{x} &= \begin{bmatrix} D \\ \mathbf{0} \end{bmatrix} \in \mathbb{S}, \\ Z &= \begin{bmatrix} -\mathbf{1}^{\dagger} \\ \mathbf{I} \end{bmatrix} \in \mathbb{R}^{n \times (n-1)}, \\ \xi &= \begin{bmatrix} x_2 \\ \vdots \\ x_n \end{bmatrix} \in \mathbb{R}^{n-1}, \end{aligned}$$

- and where $\tau : \mathbb{R}^{n-1} \to \mathbb{S}$ is defined by:

$$\forall \xi \in \mathbb{R}^{n-1}, \tau(\xi) = \hat{x} + Z\xi,$$

- and we note that τ is onto $\mathbb{S}$.

Eliminating a variable, continued

- The point $\hat{x}$ is a particular solution of the equations $Ax = b$.
- The matrix Z has columns that form a basis for the null space of A.
- The objective $f(\tilde{x})$ depends only on $\xi \in \mathbb{R}^{n-1}$.
- We have transformed the equality-constrained problem into an unconstrained problem with objective $\phi : \mathbb{R}^{n-1} \to \mathbb{R}$ defined by:

$$\begin{aligned} \forall \xi \in \mathbb{R}^{n-1}, \phi(\xi) &= f(\tilde{x}), \\ &= f\begin{pmatrix} D - \mathbf{1}^{\dagger}\xi \\ \xi \end{pmatrix}, \\ &= f(\tau(\xi)). \end{aligned}$$

- The unconstrained problem:

$$\min_{\xi \in \mathbb{R}^{n-1}} f(\tilde{x}) = \min_{\xi \in \mathbb{R}^{n-1}} \phi(\xi),$$

- could then be solved using the techniques developed in Chapter 10.

Eliminating a variable, continued

- Elimination of variables is often an effective way to solve a problem with linear constraints.
- If there were, say, m equality constraints eliminated, then there would be $(n-m)$ variables in the resulting transformed problem, assuming that the corresponding rows of A were linearly independent.

Treating the constraint directly

- We will also explore approaches that treat the equality constraints directly.

12.1.4.3 Solvability

- Since:
  - (i) we have defined the objective function f on the whole of $\mathbb{R}^n$,
  - (ii) the objective increases with increasing values of $x_k \geq 0$, for each k, and
  - (iii) the constraint has a particularly simple form,
- there will always be a solution to Problem (12.4).
- However, the solution might not satisfy the minimum and maximum machine constraints (12.2).

12.2 Power system state estimation with zero injection buses

12.2.1 Motivation

12.2.1.1 Zero injection buses

- Recall the power system state estimation problem introduced in Section 9.2.
- Consider the situation in Figure 12.5.
- Bus 2 does not have any load nor generation nor any measurement devices.
- Such buses are common at intermediate points in electric power systems between generators and load.
- We called this bus a **zero injection bus**.

12.2.1.2 Ignoring zero injection buses

- Suppose we use only the measurements shown explicitly in Figure 12.5 in the objective of Problem (9.8).
- We do not have enough information to uniquely determine the voltage magnitudes and angles at buses 2 and 3.

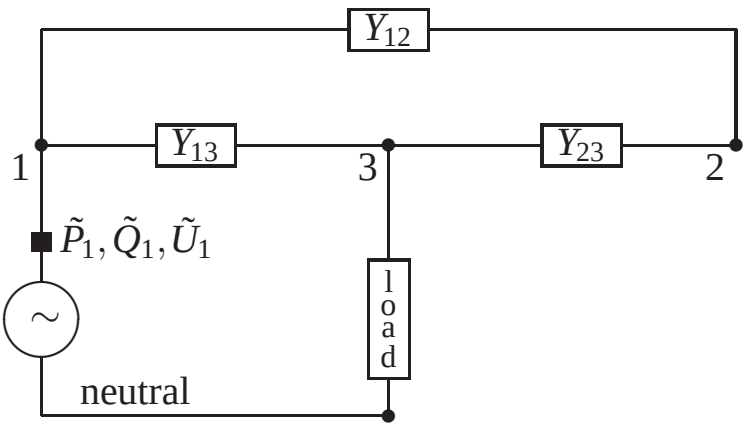

Fig. 12.5. Three-bus electric power system with a bus, bus 2, having neither load nor generation.

12.2.1.3 Treating zero injection buses as accurate measurements

- Alternatively, we could think of the zero injection at bus 2 as a pair of very accurate real and reactive power measurements having zero value and zero measurement error.

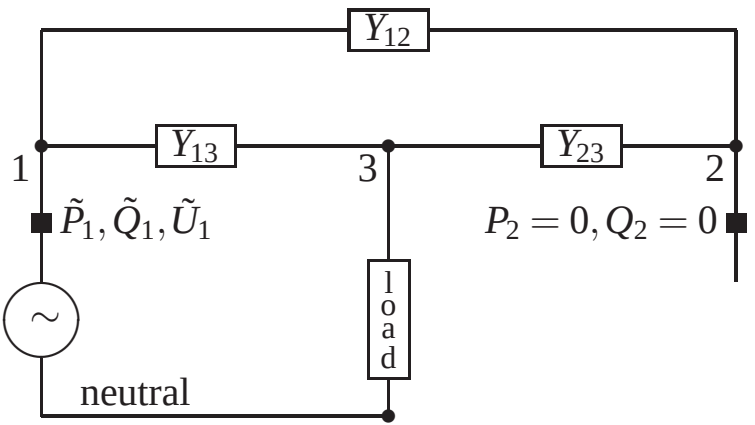

Fig. 12.6. Zero injection bus re-interpreted as an exact measurement.

Treating zero injection buses as accurate measurements, continued

- We pick a small but non-zero value of measurement error σ_ℓ for each zero injection bus measurement.
- We must then compromise between:
  - (i) making σ_ℓ small enough to approximately represent our certainty that the measurement is zero, and
  - (ii) making σ_ℓ large enough so that the entries in $[\Sigma]^{-1}$ are not too large.
- The entry in $[\Sigma]^{-1}$ corresponding to the zero injection bus measurement $\tilde{G}_\ell = 0$ is $(\sigma_\ell)^{-1}$, which must be "approximately" infinity to enforce satisfaction of the constraint $\tilde{g}_\ell(x) = 0$.
- We are effectively using a penalty function approach, as discussed in Section 3.1.2.1.
- The optimality conditions and algorithms developed in Section 11.2.3 involved factorizing either $\tilde{J}(x)^\dagger [\Sigma]^{-2} \tilde{J}(x)$ or $[\Sigma]^{-1} \tilde{J}(x)$, where $\tilde{J}$ is the Jacobian of $\tilde{g}$.
- The presence of widely differing values in Σ will lead to an ill-conditioned coefficient matrix as discussed in Section 3.1.2.1.

12.2.1.4 Treating zero injection buses as equality constraints

- The approach we will follow is to explicitly represent the zero injection buses as pairs of equality constraints each of the form $g_\ell(x) = 0$.

12.2.2 Formulation

12.2.2.1 Objective

- Let $\mathbb{M}$ be the set of measurements in the system, not including the injection measurements at the zero injection buses.
- The maximum likelihood objective can again be transformed into:

$$\forall x \in \mathbb{R}^n, f(x) = \sum_{\ell \in \mathbb{M}} \frac{(\tilde{g}_\ell(x) - \tilde{G}_\ell)^2}{2\sigma_\ell^2}. \tag{12.7}$$

12.2.2.2 Constraints

- Let $\mathbb{M}^0$ be the set of real and reactive injections at the zero injection buses.
- For each $\ell \in \mathbb{M}^0$, let $g_\ell : \mathbb{R}^n \to \mathbb{R}$ be the function representing an injection at a zero injection bus.
- The power flow equations require that $\forall \ell \in \mathbb{M}^0, g_\ell(x) = 0$, so that our estimate of the state x should be consistent with these constraints.
- We can collect the functions associated with the zero injection buses together into a vector function $g : \mathbb{R}^n \to \mathbb{R}^m$, where m is the number of zero injection bus measurements, which is the number of elements in $\mathbb{M}^0$.
- That is, $g : \mathbb{R}^n \to \mathbb{R}^m$ is defined by:

$$\forall x \in \mathbb{R}^n, g(x) = (g_\ell(x))_{\ell \in \mathbb{M}^0}. \tag{12.8}$$

12.2.2.3 Problem

- Our problem is therefore:

$$\min_{x \in \mathbb{R}^n} \{ f(x) | g(x) = \mathbf{0} \}. \tag{12.9}$$

12.2.3 Change in measurement data

- Over time, the state of the power system changes as demand and supply situations change.
- Consequently, the measured data will change.
- We will consider how a change in measurement data affects the result.

12.2.4 Problem characteristics

12.2.4.1 Objective

- The objective of Problem (12.9) defined in (12.7) is approximately quadratic.

12.2.4.2 Constraints

- The constraints $g(x) = \mathbf{0}$ are approximately linear.
- However, since they are not exactly linear we cannot eliminate them and re-write the problem as an unconstrained optimization in fewer variables.

12.2.4.3 Solvability

- The constraints in the problem are consistent with Kirchhoff's laws and we know from physical principles that there are solutions to Kirchhoff's laws.

13

Algorithms for linear equality-constrained minimization

- In this chapter we will develop algorithms for constrained optimization problems of the form:

$$\min_{x \in \mathbb{R}^n} \{ f(x) | Ax = b \}, \tag{13.1}$$

- where $f : \mathbb{R}^n \to \mathbb{R}$, $A \in \mathbb{R}^{m \times n}$, and $b \in \mathbb{R}^m$.

Key issues

- Consideration of **descent directions** for the objective that also maintain feasibility for the constraints,
- consideration of the **null space** of the coefficient matrix A to transform the constrained problem into an unconstrained problem,
- optimality conditions and the definition and interpretation of the **dual variables** and the **Lagrange multipliers**,
- optimality conditions for **convex problems**, and
- **duality** and **sensitivity analysis**.

Analysis

- Consider a feasible point $\hat{x}$ that is a candidate solution to Problem (13.1).
- By the discussion in Section 5.8.1.2, every feasible point is of the form $\hat{x}+\Delta x$ where:

$$\begin{aligned} \Delta x \in \mathcal{N}(A) &= \{\Delta x \in \mathbb{R}^n | A\Delta x = \mathbf{0}\}, \\ &= \{Z\Delta\xi | \Delta\xi \in \mathbb{R}^{n'}\}, \end{aligned}$$

- where $Z \in \mathbb{R}^{n\times n'}$, with $n' \geq n-m$, is a matrix with columns that form a basis for the null space of A.

Analysis, continued

- Suppose that a vector $\Delta x \in \mathcal{N}(A)$ happened to also satisfy $\nabla f(\hat{x})^{\dagger}\Delta x < 0$.
- By Lemma 10.1, such a direction is a descent direction for f at $\hat{x}$.
- That is:

$$\exists \overline{\alpha} \in \mathbb{R}_{++} \text{ such that } (0 < \alpha \leq \overline{\alpha}) \Rightarrow (f(\hat{x} + \alpha \Delta x) < f(\hat{x})). \qquad (13.2)$$

- We also have that:

$$\begin{aligned} \forall \alpha \in \mathbb{R}, A(\hat{x} + \alpha \Delta x) &= b + \alpha A \Delta x, \\ &= b. \end{aligned}$$

- If $\Delta x \in \mathcal{N}(A)$ and $\nabla f(\hat{x})^{\dagger}\Delta x < 0$ then $\hat{x}$ cannot be a minimizer.

Example

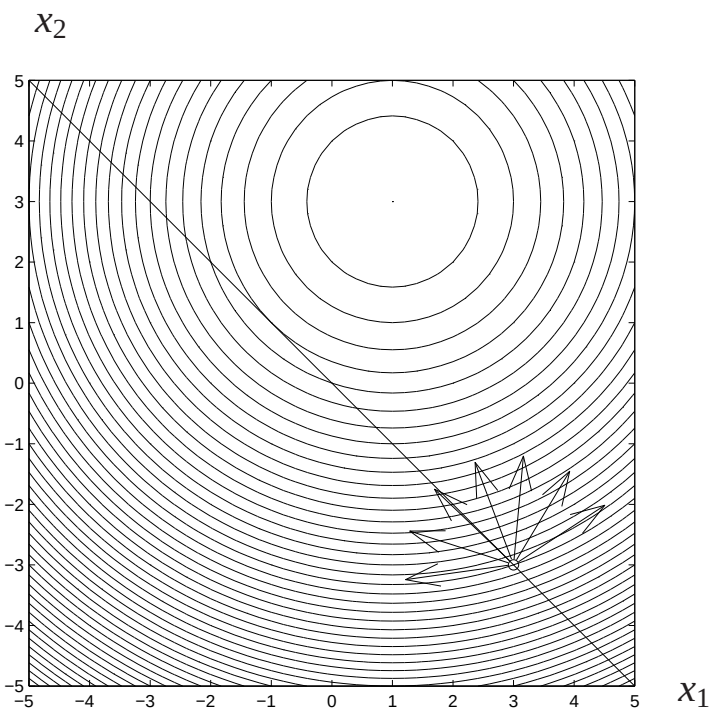

Fig. 13.1. Descent directions for a function at a point $\hat{x} = \begin{bmatrix} 3 \\ -3 \end{bmatrix}$, indicated by the $\circ$, and one descent direction that maintains feasibility for the equality constraint corresponding to the feasible set illustrated by the line.

Analysis

- Suppose $x^\star$ is a minimizer of the linear equality-constrained problem.
- Then for any direction $\Delta x \in \mathcal{N}(A)$, that is, such that $A\Delta x = \mathbf{0}$, we must have that:

$$\nabla f(x^\star)^\dagger \Delta x \not< 0.$$

- Applying the same argument to the vector $(-\Delta x) \in \mathcal{N}(A)$, we must have that:

$$\nabla f(x^\star)^\dagger (-\Delta x) \not< 0.$$

- Combining these two observations, we have that:

$$\nabla f(x^\star)^\dagger \Delta x = 0.$$

- If $x^\star$ is a minimizer of the linear equality-constrained problem then for each $\Delta x \in \mathcal{N}(A)$ we must have that $\nabla f(x^\star)^\dagger \Delta x = 0$.
- That is, $\mathcal{N}(A) \subseteq \{\Delta x \in \mathbb{R}^n | \nabla f(x^\star)^\dagger \Delta x = 0\}$.

Example

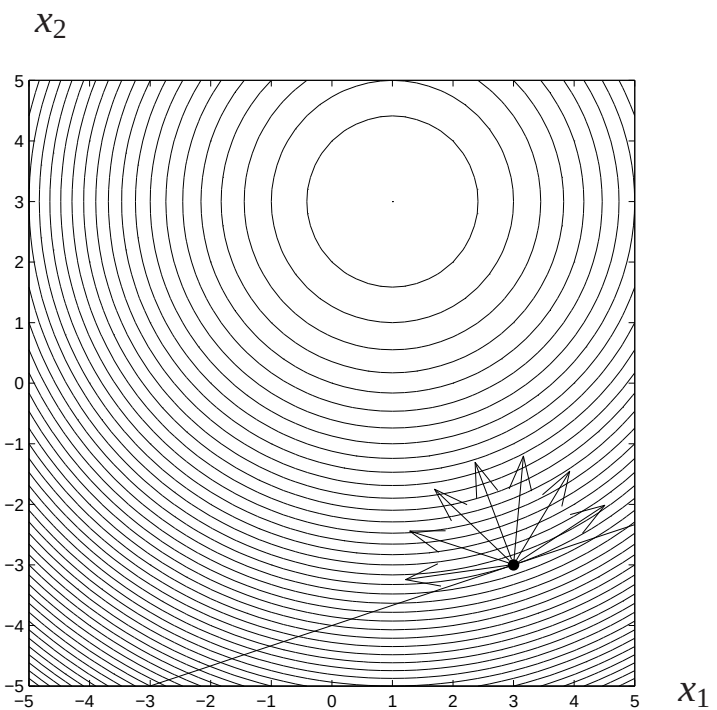

Fig. 13.2. Descent directions for a function at a point $x^\star = \begin{bmatrix} 3 \\ -3 \end{bmatrix}$, indicated by the •, none of which maintains feasibility for the equality constraint corresponding to the feasible set $\mathbb{S}$ illustrated by the line.

Tangent plane

Definition 13.1 Let $f : \mathbb{R}^n \to \mathbb{R}$ be partially differentiable, $x^\star \in \mathbb{R}^n$, and suppose that $\nabla f(x^\star) \neq \mathbf{0}$. Let $f^\star = f(x^\star)$. Then the **tangent plane to the contour set** $\mathbb{C}_f(f^\star) = \{x \in \mathbb{R}^n | f(x) = f^\star\}$ of f at the point $x^\star$ is the set:

$$\mathbb{P} = \{x \in \mathbb{R}^n | \nabla f(x^\star)^\dagger (x - x^\star) = 0\}.$$

For brevity, we will often refer to $\mathbb{P}$ as "the tangent plane to the contour set of f at $x^\star$." If a set $\mathbb{S} \subseteq \mathbb{R}^n$ is contained in $\mathbb{P}$ then we say that "the contour set of f is **tangential** to $\mathbb{S}$ at $x^\star$." □

Example

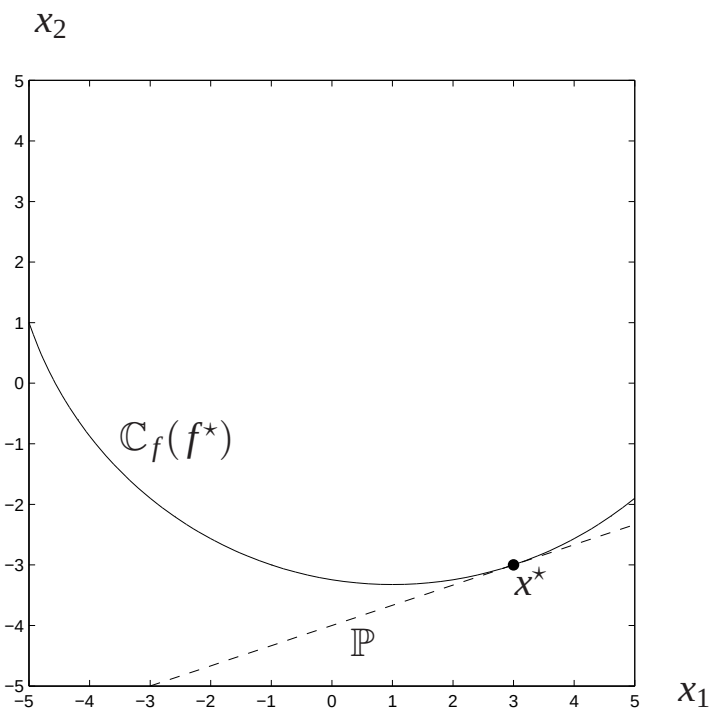

Fig. 13.3. Tangent plane $\mathbb{P}$ (shown dashed) to contour set $\mathbb{C}_f(f^\star)$ of f (shown solid) at a point $x^\star = \begin{bmatrix} 3 \\ -3 \end{bmatrix}$, indicated by the $\bullet$.

Example in higher dimension

- Consider $x^\star = \begin{bmatrix} 1 \\ 1 \\ 0 \end{bmatrix}$ and objective function $f : \mathbb{R}^3 \to \mathbb{R}$ defined by:

$$
\begin{aligned}
\forall x \in \mathbb{R}^3, f(x) &= (x_1)^2 + (x_2)^2 + (x_3)^2, \qquad (13.3) \\
\forall x \in \mathbb{R}^3, \nabla f(x) &= \begin{bmatrix} 2x_1 \\ 2x_2 \\ 2x_3 \end{bmatrix}, \\
\nabla f(x^\star) &= \begin{bmatrix} 2 \\ 2 \\ 0 \end{bmatrix}, \\
\mathbb{P} &= \{x \in \mathbb{R}^3 | \nabla f(x^\star)^\dagger (x - x^\star) = 0\}, \\
&= \left\{ x \in \mathbb{R}^3 \left| \begin{bmatrix} 2 \\ 2 \\ 0 \end{bmatrix}^\dagger \begin{bmatrix} x_1 - 1 \\ x_2 - 1 \\ x_3 - 0 \end{bmatrix} = 0 \right. \right\}, \\
&= \{x \in \mathbb{R}^3 | x_1 + x_2 = 2\}.
\end{aligned}
$$

Example in higher dimension, continued

- In $\mathbb{R}^n$, the tangent plane is a **hyperplane**, that is, a space of dimension $n-1$ defined by a single equality constraint.
- Descent directions for f at $x^\star$ point from $x^\star$ into the sphere.

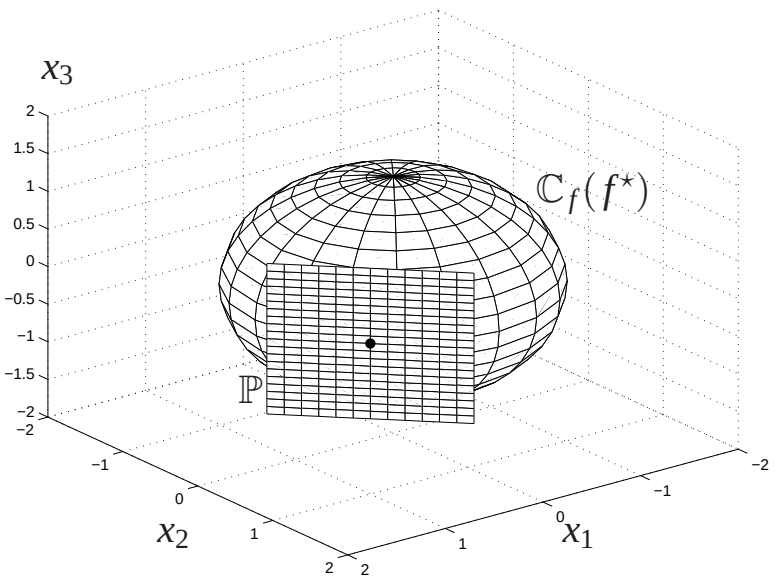

Fig. 13.4. Tangent plane $\mathbb{P}$ to contour set $\mathbb{C}_f(f^\star)$ of f at a point $x^\star = \begin{bmatrix} 1 \\ 1 \\ 0 \end{bmatrix}$, indicated by the •. The contour set is the sphere and the tangent plane is the plane.

13.1.1.4 Geometric interpretation

Analysis

- In Section 13.1.1.2, we showed that if $x^\star$ is a constrained minimizer of f then:

$$\mathcal{N}(A) \subseteq \{\Delta x \in \mathbb{R}^n | \nabla f(x^\star)^\dagger \Delta x = 0\}.$$

- Translating both of these sets by adding $x^\star$ to every element in both sets and noting that $Ax^\star = b$, we have that:

$$\begin{aligned}
\mathbb{S} &= \{x \in \mathbb{R}^n | Ax = b\}, \\
&= \{x \in \mathbb{R}^n | x = x^\star + \Delta x, \Delta x \in \mathcal{N}(A)\}, \text{ since } Ax^\star = b, \\
&\subseteq \{x \in \mathbb{R}^n | x = x^\star + \Delta x, \nabla f(x^\star)^\dagger \Delta x = 0\}, \\
&\qquad \text{since } x^\star \text{ is a constrained minimizer of } f, \\
&= \{x \in \mathbb{R}^n | \nabla f(x^\star)^\dagger (x - x^\star) = 0\}, \\
&= \mathbb{P},
\end{aligned}$$

- which is the tangent plane to the contour set of f at $x^\star$.

Analysis, continued

- Geometrically, we can say that the feasible set, $\mathbb{S} = \{x \in \mathbb{R}^n | Ax = b\}$, is contained in the set $\mathbb{P}$, which is the tangent plane to the contour set of f at $x^\star$.
- We can also say that the contour set of f is tangential to the feasible set at $x^\star$.
- This observation is consistent with Figures 13.2 and 13.3.

Example

- Recall the example equality-constrained Problem (2.13):

$$\min_{x\in\mathbb{R}^2}\{f(x)|Ax=b\},$$

$$\begin{aligned}\text{where: } \forall x\in\mathbb{R}^2, f(x) &= (x_1-1)^2+(x_2-3)^2,\\ A &= [1\ \ -1],\\ b &= [0].\end{aligned}$$

- The (unique) local minimizer is at $x^\star=\begin{bmatrix}2\\2\end{bmatrix}$ with minimum $f^\star=2$.
- The tangent plane to the contour set of f at $x^\star$ is:

$$\begin{aligned}\mathbb{P} &= \{x\in\mathbb{R}^2|\nabla f(x^\star)^\dagger(x-x^\star)=\mathbf{0}\},\\ &= \left\{x\in\mathbb{R}^2\left|\begin{bmatrix}2\\-2\end{bmatrix}^\dagger\left(x-\begin{bmatrix}2\\2\end{bmatrix}\right)=0\right.\right\},\\ &= \{x\in\mathbb{R}^2|x_1-x_2=0\},\end{aligned}$$

- which is the same set as the feasible set.

Example, continued

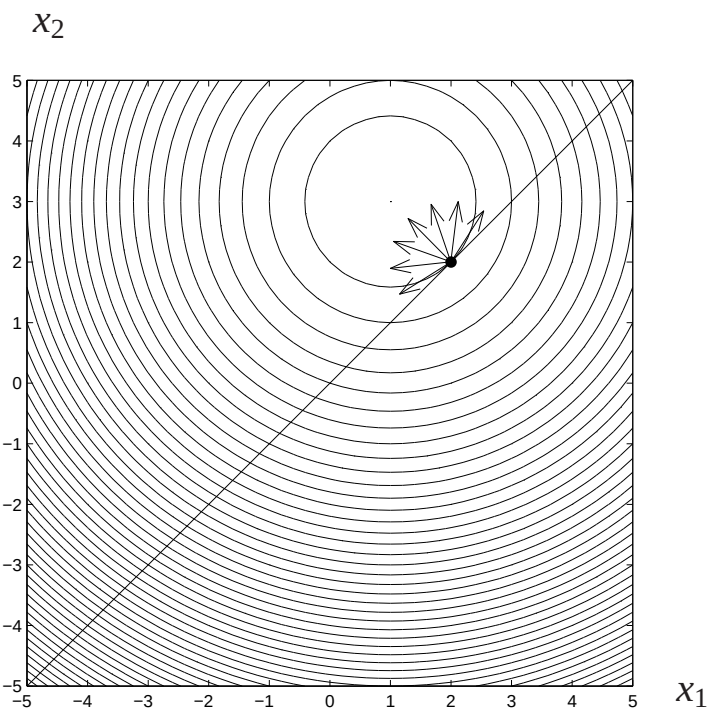

Fig. 13.5. Descent directions for a function at a point $x^\star = \begin{bmatrix} 2 \\ 2 \end{bmatrix}$, (indicated by the •), none of which maintains feasibility for the equality constraint illustrated by the line.

Example of strict containment

- In higher dimensions, it can typically be the case that the feasible set $\mathbb{S} = \{x \in \mathbb{R}^n | Ax = b\}$ is *strictly* contained in $\mathbb{P} = \{x \in \mathbb{R}^n | \nabla f(x^\star)^\dagger (x - x^\star) = \mathbf{0}\}$.
- For example, consider again the objective function $f : \mathbb{R}^3 \to \mathbb{R}$ defined in (13.3):

$$\forall x \in \mathbb{R}^3, f(x) = (x_1)^2 + (x_2)^2 + (x_3)^2.$$

- Moreover, suppose that the equality constraints $Ax = b$ are defined by:

$$\begin{aligned} A &= \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \end{bmatrix}, \\ b &= \mathbf{1}. \end{aligned}$$

- The constraints specify that $x_1 = x_2 = 1$, so that the feasible set is the line in $\mathbb{R}^3$ that is parallel to the x_3-axis and that passes through $x_1 = x_2 = 1$.
- By inspection, the minimizer of $\min_{x \in \mathbb{R}^3} \{f(x) | Ax = b\}$ is $x^\star = \begin{bmatrix} 1 \\ 1 \\ 0 \end{bmatrix}$.

Example of strict containment, continued

- In this case:

$$\begin{aligned}
\mathbb{S} &= \{x \in \mathbb{R}^3 | Ax = b\}, \\
&= \{x \in \mathbb{R}^3 | x_1 = x_2 = 1\}, \\
\mathbb{P} &= \{x \in \mathbb{R}^3 | \nabla f(x^\star)^\dagger (x - x^\star) = \mathbf{0}\}, \\
&= \left\{ x \in \mathbb{R}^3 \left| \begin{bmatrix} 2 & 2 & 0 \end{bmatrix} \left(x - \begin{bmatrix} 1 \\ 1 \\ 0 \end{bmatrix} \right) = \mathbf{0} \right. \right\}, \\
&= \{x \in \mathbb{R}^3 | x_1 + x_2 = 2\}.
\end{aligned}$$

- That is, the tangent plane to the contour set of f at $x^\star$ is a plane, $\mathbb{P}$, in $\mathbb{R}^3$, which strictly contains the feasible set $\mathbb{S}$, which is a line.
- The situation is illustrated in Figure 13.6, which repeats Figure 13.4 but adds a line that represents $\mathbb{S}$.
- Descent directions for f at $x^\star$ point into the sphere.
- No descent directions point along the feasible set $\mathbb{S}$.

Example of strict containment, continued

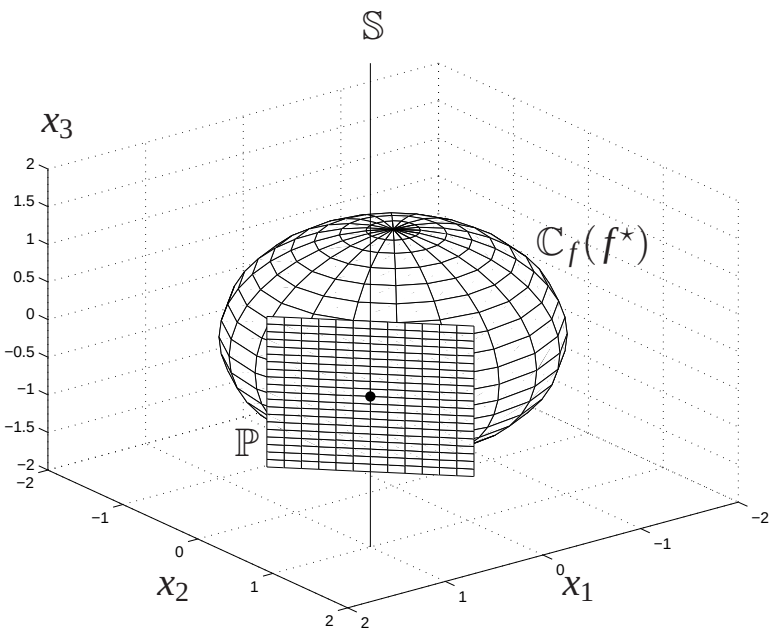

Fig. 13.6. Feasible set strictly $\mathbb{S}$ contained in tangent plane $\mathbb{P}$ to contour set $\mathbb{C}_f(f^\star)$ of f at a point $x^\star = \begin{bmatrix} 1 \\ 1 \\ 0 \end{bmatrix}$, indicated by the $\bullet$. The contour set is the sphere; the tangent plane is the plane; and, the feasible set is the vertical line.

13.1.1.5 Summary

- At a minimizer $x^\star$ of Problem (13.1), every descent direction for f at $x^\star$ must lie outside the null space of A.
- At a minimizer, the contour set of f is tangential to the feasible set.

13.1.2 First-order necessary conditions

13.1.2.1 Transformation of problem

- Let $Z \in \mathbb{R}^{n \times n'}$, with $n' \geq n - m$, be a matrix with columns that form a basis for the null space of A. Then:

$$\begin{aligned} \mathcal{N}(A) &= \{\Delta x \in \mathbb{R}^n | A\Delta x = \mathbf{0}\}, \\ &= \{Z\Delta\xi | \Delta\xi \in \mathbb{R}^{n'}\}. \end{aligned}$$

- Suppose that $\hat{x} \in \mathbb{R}^n$ is a particular solution to $Ax = b$.

$$\begin{aligned} \mathbb{S} &= \{x \in \mathbb{R}^n | Ax = b\}, \\ &= \{x \in \mathbb{R}^n | x = \hat{x} + \Delta x, A\Delta x = \mathbf{0}, \Delta x \in \mathbb{R}^n\}, \\ &= \{\hat{x} + Z\Delta\xi | \Delta\xi \in \mathbb{R}^{n'}\}. \end{aligned}$$

- We can define an **onto function** $\tau : \mathbb{R}^{n'} \to \mathbb{S}$ by:

$$\forall \xi \in \mathbb{R}^{n'}, \tau(\xi) = \hat{x} + Z\xi.$$

- Varying ξ over $\mathbb{R}^{n'}$ allows $\tau(\xi)$ to explore over the feasible set $\mathbb{S}$.

Transformation of problem, continued

- We use Theorem 3.5 to transform the equality-constrained Problem (13.1) into an unconstrained problem.
- In the hypothesis of Theorem 3.5, let $\mathbb{P} = \mathbb{R}^{n'}$ and define $\phi : \mathbb{R}^{n'} \to \mathbb{R}$ by:

$$\forall \xi \in \mathbb{R}^{n'}, \phi(\xi) = f(\tau(\xi)). \tag{13.4}$$

- The function ϕ is called the **reduced function**.

- By Theorem 3.5:
  - (i) $\min_{x\in\mathbb{R}^n}\{f(x)|Ax=b\}$ has a minimum if and only if $\min_{\xi\in\mathbb{R}^{n'}}\phi(\xi)$ has a minimum.
  - (ii) If either one of the problems in Item (i) possesses a minimum (and consequently, by Item (i), each one possesses a minimum), then:

$$\min_{\xi\in\mathbb{R}^{n'}}\phi(\xi) = \min_{x\in\mathbb{S}} f(x),$$

$$\arg\min_{x\in\mathbb{R}^n}\{f(x)|Ax=b\} = \left\{\tau(\xi)\,\middle|\,\xi\in\arg\min_{\xi\in\mathbb{R}^{n'}}\phi(\xi)\right\}.$$

- $\min_{\xi\in\mathbb{R}^{n'}}\phi(\xi)$ is an *unconstrained* problem.

Transformation of problem, continued

- The gradient of ϕ, $\nabla\phi(\bullet) = Z^\dagger \nabla f(\tau(\bullet))$, is called the **reduced gradient** or the **projected gradient**.
- Consider the direction corresponding to the reduced gradient in the original decision variables $x \in \mathbb{R}^n$.
- Referred to the original decision variables x, the reduced gradient $\nabla\phi$ corresponding to a point $\hat{x} \in \mathbb{R}^n$ lies in the direction $ZZ^\dagger \nabla f(\hat{x}) \in \mathbb{R}^n$.
- The vector $\Delta x = -ZZ^\dagger \nabla f(\hat{x})$, which is opposite to the direction corresponding to the reduced gradient, is a descent direction for f at $\hat{x}$ unless the reduced gradient $Z^\dagger \nabla f(\hat{x}) = \mathbf{0}$.
- Moreover, if $A\hat{x} = b$ then, for any α, $\hat{x} + \alpha\Delta x$ also satisfies the equality constraints.

13.1.2.2 Necessary conditions in terms of original problem

Analysis

Theorem 13.1 *Suppose that $f : \mathbb{R}^n \to \mathbb{R}$ is partially differentiable with continuous partial derivatives, $A \in \mathbb{R}^{m \times n}$, and $b \in \mathbb{R}^m$. Let $Z \in \mathbb{R}^{n \times n'}$ be a matrix with columns that form a basis for the null space of A. If $x^\star \in \mathbb{R}^n$ is a local minimizer of the problem:*

$$\min_{x \in \mathbb{R}^n} \{ f(x) | Ax = b \},$$

then:

$$\begin{aligned} Z^\dagger \nabla f(x^\star) &= \mathbf{0}, \\ Ax^\star &= b. \end{aligned} \tag{13.5}$$

□

Example

- We continue with the previous equality-constrained Problem (2.13).
- By inspection, $Z = \begin{bmatrix} 1 \\ 1 \end{bmatrix} \in \mathbb{R}^{2\times 1}$ is a matrix with columns that form a basis for the null space:

$$\mathcal{N}(A) = \{\Delta x \in \mathbb{R}^n | A\Delta x = \mathbf{0}\},$$

- since:

  $A\Delta x = \mathbf{0}$ if and only if $\Delta x_1 = \Delta x_2$, and

  for $\xi \in \mathbb{R}, Z\xi = \begin{bmatrix} \xi \\ \xi \end{bmatrix}$.

Example

- Also:

$$\begin{aligned} \forall x \in \mathbb{R}^2, \nabla f(x) &= \begin{bmatrix} 2(x_1 - 1) \\ 2(x_2 - 3) \end{bmatrix}, \\ \nabla f(x^\star) &= \begin{bmatrix} 2 \\ -2 \end{bmatrix}, \end{aligned}$$

- so that $x^\star = \begin{bmatrix} 2 \\ 2 \end{bmatrix}$ is not an *unconstrained* minimizer of f.
- Using these calculations, we obtain:

$$\begin{aligned} Z^\dagger \nabla f(x^\star) &= \begin{bmatrix} 1 & 1 \end{bmatrix} \begin{bmatrix} 2 \\ -2 \end{bmatrix}, \\ &= [0]. \end{aligned}$$

- consistent with the conclusion of Theorem 13.1.

Analysis

Theorem 13.2 *Suppose that $f : \mathbb{R}^n \to \mathbb{R}$ is partially differentiable with continuous partial derivatives, $A \in \mathbb{R}^{m\times n}$, and $b \in \mathbb{R}^m$. If $x^\star \in \mathbb{R}^n$ is a local minimizer of the problem:*

$$\min_{x\in\mathbb{R}^n}\{f(x)|Ax = b\},$$

then:

$$\exists\lambda^\star \in \mathbb{R}^m \text{ such that } \nabla f(x^\star) + A^\dagger\lambda^\star = \mathbf{0}, \tag{13.6}$$

$$Ax^\star = b. \tag{13.7}$$

Proof By Theorem 13.1:

$$Z^\dagger\nabla f(x^\star) = \mathbf{0}, \tag{13.8}$$

$$Ax^\star = b,$$

where $Z \in \mathbb{R}^{n\times n'}$ is a matrix with columns that form a basis for the null space of A. Any vector in $\mathbb{R}^n$ can be written in the form $Zu - A^\dagger\lambda$ for

some $u \in \mathbb{R}^{n'}$ and $\lambda \in \mathbb{R}^m$. In particular, since $\nabla f(x^\star) \in \mathbb{R}^n$, we have:

$$\exists u^\star \in \mathbb{R}^{n'}, \exists \lambda^\star \in \mathbb{R}^m \text{ such that } \nabla f(x^\star) = Zu^\star - A^\dagger \lambda^\star.$$

Multiplying this expression through by $Z^\dagger$ we obtain:

$$Z^\dagger \nabla f(x^\star) = Z^\dagger Z u^\star - Z^\dagger A^\dagger \lambda^\star.$$

But $Z^\dagger \nabla f(x^\star) = \mathbf{0}$ by (13.8), so:

$$Z^\dagger Z u^\star - Z^\dagger A^\dagger \lambda^\star = \mathbf{0}.$$

Also $AZ = \mathbf{0}$, so $Z^\dagger A^\dagger \lambda^\star = \mathbf{0}$ and $Z^\dagger Z u^\star = \mathbf{0}$. But this means that $u^\star = \mathbf{0}$ since Z has linearly independent columns. That is,

$$\exists \lambda^\star \in \mathbb{R}^m \text{ such that } \nabla f(x^\star) + A^\dagger \lambda^\star = \mathbf{0},$$

which is (13.6). We already have that $Ax^\star = b$, which is (13.7). □

- A vector $\lambda^\star$ satisfying (13.6), given an $x^\star$ that also satisfies (13.7), is called a vector of **Lagrange multipliers** for the problem.
- The conditions (13.6)–(13.7) are called the **first-order necessary conditions** (or **FONC**) for Problem (13.1).

Example

- Continuing with the previous equality-constrained Problem (2.13), we obtain:

$$\begin{aligned} \nabla f(x^\star) + A^\dagger[-2] &= \begin{bmatrix} 2 \\ -2 \end{bmatrix} + \begin{bmatrix} 1 \\ -1 \end{bmatrix}[-2], \\ &= \mathbf{0}, \end{aligned}$$

- which is consistent with Theorem 13.2 for $\lambda^\star = [-2]$.

13.1.2.4 Analytic interpretation

The Lagrangian

- Recall Definition 3.2 of the **Lagrangian**.
- For a problem with objective $f : \mathbb{R}^n \to \mathbb{R}$ and equality constraints $Ax = b$, with $A \in \mathbb{R}^{m\times n}$ and $b \in \mathbb{R}^m$ the Lagrangian $\mathcal{L} : \mathbb{R}^n \times \mathbb{R}^m \to \mathbb{R}$ is defined by:

$$\forall x \in \mathbb{R}^n, \forall \lambda \in \mathbb{R}^m, \mathcal{L}(x,\lambda) = f(x) + \lambda^\dagger (Ax - b), \tag{13.9}$$

  where λ is called the vector of **dual variables** for the problem.
- We also define the gradients of $\mathcal{L}$ with respect to x and λ by, respectively, $\nabla_x \mathcal{L} = \left[\dfrac{\partial \mathcal{L}}{\partial x}\right]^\dagger$ and $\nabla_\lambda \mathcal{L} = \left[\dfrac{\partial \mathcal{L}}{\partial \lambda}\right]^\dagger$.
- That is:

$$\begin{aligned} \nabla_x \mathcal{L}(x,\lambda) &= \nabla f(x) + A^\dagger \lambda, \\ \nabla_\lambda \mathcal{L}(x,\lambda) &= Ax - b. \end{aligned}$$

- We can interpret the first-order necessary conditions (13.6)–(13.7) in two ways using the Lagrangian $\mathcal{L}$.

Minimization of Lagrangian over primal variables

- The first-order necessary conditions imply that $x^\star$ is a critical point of the function $\mathcal{L}(\bullet, \lambda^\star)$ that also satisfies the constraints $Ax = b$.
- We seek a point $x^\star$ that minimizes $\mathcal{L}(\bullet, \lambda^\star)$.
- The vector of Lagrange multipliers $\lambda^\star$ "adjusts" the unconstrained optimality conditions by $A^\dagger \lambda^\star$ to "balance" the minimization of the objective against satisfaction of the constraints.

Critical point of the Lagrangian

- The first-order necessary conditions also imply that $\begin{bmatrix} x^\star \\ \lambda^\star \end{bmatrix}$ is a solution of the simultaneous equations:

$$\nabla_x \mathcal{L}(x, \lambda) = \mathbf{0}, \tag{13.10}$$
$$\nabla_\lambda \mathcal{L}(x, \lambda) = \mathbf{0}. \tag{13.11}$$

- The second set of equations requires that $x^\star$ be feasible and are linear equations.
- We seek $\begin{bmatrix} x^\star \\ \lambda^\star \end{bmatrix}$ satisfying $\nabla \mathcal{L}(x^\star, \lambda^\star) = \mathbf{0}$, where $\nabla \mathcal{L} = \begin{bmatrix} \nabla_x \mathcal{L} \\ \nabla_\lambda \mathcal{L} \end{bmatrix}$.
- That is, $\begin{bmatrix} x^\star \\ \lambda^\star \end{bmatrix}$ is a critical point of $\mathcal{L}$.
- However, $\begin{bmatrix} x^\star \\ \lambda^\star \end{bmatrix}$ is *not* a minimizer of $\mathcal{L}(\bullet, \bullet)$ over values of $\begin{bmatrix} x \\ \lambda \end{bmatrix}$.

Algorithms

- As in the unconstrained case, these two interpretations lead us to two (of several) classes of algorithms for solving Problem (13.1):
  - (i) minimize the Lagrangian over x for a fixed λ and then adjust λ until feasibility is obtained, (Sections 13.3.1.4 and 13.3.2.4), and
  - (ii) solve the necessary conditions (13.10)–(13.11) for x and λ, (Sections 13.3.1.3 and 13.3.2.3).

Example

- Continuing with the previous equality-constrained Problem (2.13), the Lagrangian $\mathcal{L} : \mathbb{R}^2 \times \mathbb{R} \to \mathbb{R}$ is defined by:

$$\forall x \in \mathbb{R}^2, \forall \lambda \in \mathbb{R}, \mathcal{L}(x, \lambda) = (x_1 - 1)^2 + (x_2 - 3)^2 + \lambda(x_1 - x_2). \quad (13.12)$$

- Setting the value of the dual variable in the Lagrangian equal to the Lagrange multiplier, $\lambda^\star = [-2]$, we have:

$$\forall x \in \mathbb{R}^2, \mathcal{L}(x, \lambda^\star) = (x_1 - 1)^2 + (x_2 - 3)^2 + (-2)(x_1 - x_2).$$

- The first-order necessary conditions for minimizing $\mathcal{L}(x, \lambda^\star)$ with respect to x is that:

$$\begin{aligned} \nabla_x \mathcal{L}(x, \lambda^\star) &= \begin{bmatrix} 2(x_1 - 1) - 2 \\ 2(x_2 - 3) + 2 \end{bmatrix}, \\ &= \mathbf{0}, \end{aligned}$$

- which yields a solution of $x^\star = \begin{bmatrix} 2 \\ 2 \end{bmatrix}$.

Example, continued

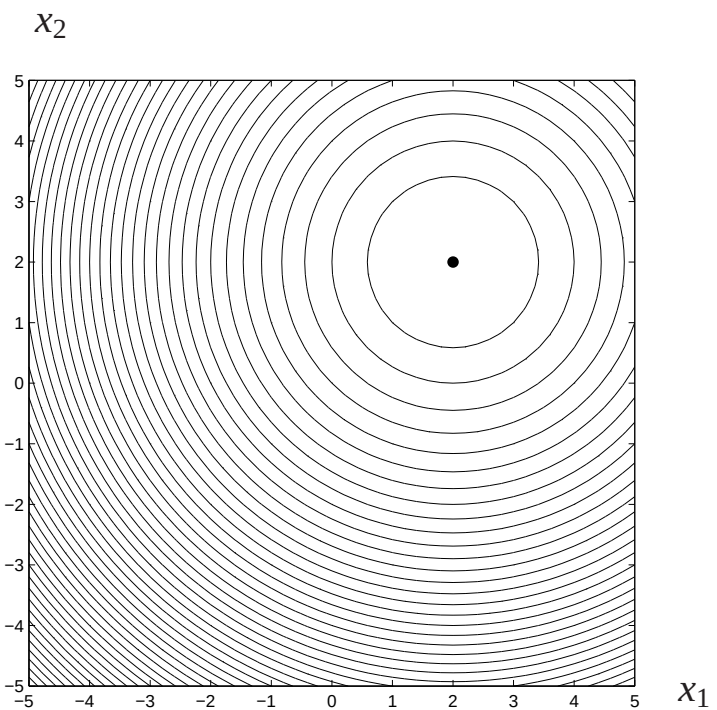

Fig. 13.7. Contour sets for Lagrangian $\mathcal{L}(\bullet, \lambda^\star)$ evaluated at the Lagrange multipliers $\lambda^\star = [-2]$.

Example, continued

- For other values of the dual variables λ not equal to the Lagrange multipliers $\lambda^\star$, the corresponding minimizer of $\mathcal{L}(\bullet, \lambda)$ will differ from the minimizer of Problem (2.13).
- For $\tilde{\lambda} = [-5]$, the contour sets of $\mathcal{L}(\bullet, \tilde{\lambda})$ are illustrated in Figure 13.8.
- The unconstrained minimizer of this function is at $\tilde{x} = \begin{bmatrix} 3.5 \\ 0.5 \end{bmatrix}$, illustrated with a $\circ$ in Figure 13.8, which differs from $x^\star$.

Example, continued

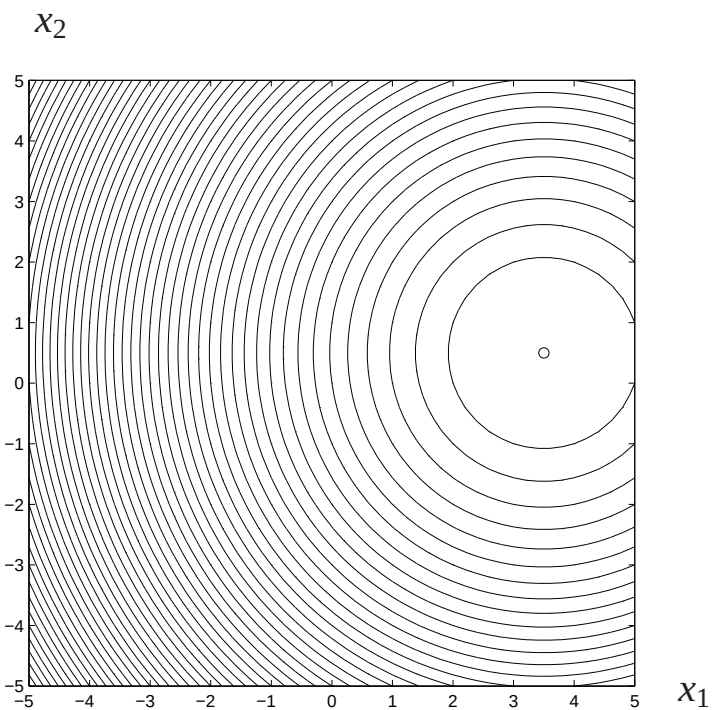

Fig. 13.8. Contour sets for Lagrangian $\mathcal{L}(\bullet, \tilde{\lambda})$ evaluated at value of dual variables $\tilde{\lambda} = [-5]$ not equal to Lagrange multiplers.

13.1.2.5 Relation to geometric interpretation

- To see that the first-order necessary conditions imply the geometric observation made in Section 13.1.1.4, suppose that $\hat{x} \in \mathbb{R}^n$ satisfies:

$$\hat{x} \in \mathbb{S} = \{x \in \mathbb{R}^n | Ax = b\}.$$

- Then $A(\hat{x} - x^\star) = \mathbf{0}$ and so $[\lambda^\star]^\dagger A(\hat{x} - x^\star) = 0$.
- The necessary conditions require that $\nabla f(x^\star)^\dagger + [\lambda^\star]^\dagger A = \mathbf{0}$.
- Multiplying by $(\hat{x} - x^\star)$ on the right we obtain:

$$\begin{aligned} 0 &= \left(\nabla f(x^\star)^\dagger + [\lambda^\star]^\dagger A\right)(\hat{x} - x^\star), \\ &= \nabla f(x^\star)^\dagger (\hat{x} - x^\star). \end{aligned}$$

- Therefore:

$$\hat{x} \in \mathbb{P} = \{x \in \mathbb{R}^n | \nabla f(x^\star)^\dagger (x - x^\star) = \mathbf{0}\}.$$

- The contour set of f is tangential to the feasible set $\mathbb{S}$ at $x^\star$.

13.1.2.6 First-order necessary conditions are not sufficient

Discussion

- As with unconstrained problems, it is possible for a point $\hat{x}$ to satisfy the first-order necessary conditions (13.6)–(13.7) and yet not be a local minimizer of Problem (13.1).

Example

- Consider the case of Problem (13.1) with $n = 2$ and $m = 1$ and:

$$\begin{aligned} \forall x \in \mathbb{R}^2, f(x) &= -\frac{1}{2}(x_1)^2 - \frac{1}{2}(x_2)^2, \\ A &= [1 \quad -1], \\ b &= [0]. \end{aligned}$$

- $\hat{x} = \mathbf{0}$ is not the minimizer of the problem.

Example, continued

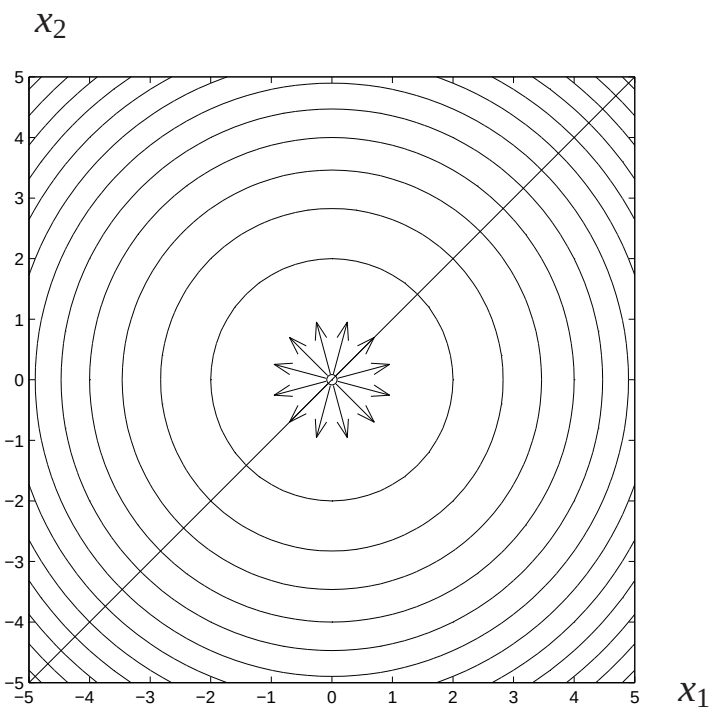

Fig. 13.9. Contour sets for non-convex objective. The objective decreases away from $\hat{x} = \mathbf{0}$.

Example, continued

- The point $\hat{x} = \mathbf{0}$ satisfies (13.6)–(13.7) with $\hat{\lambda} = 0$ since:

$$\begin{aligned}
\nabla f(\hat{x}) + A^T\hat{\lambda} &= -\hat{x} + \begin{bmatrix} 1 \\ -1 \end{bmatrix}\hat{\lambda}, \\
&= \mathbf{0} + \begin{bmatrix} 1 \\ -1 \end{bmatrix} 0, \\
&= \mathbf{0}, \\
A\hat{x} &= A\mathbf{0}, \\
&= \mathbf{0}, \\
&= b.
\end{aligned}$$

- That is, $\hat{x} = \mathbf{0}$ and $\hat{\lambda} = 0$ satisfy the first-order necessary conditions for Problem (13.1), but $\hat{x} = \mathbf{0}$ is not a minimizer of this problem.
- In fact, it is a *maximizer* of f over the feasible set.

13.1.3 Second-order sufficient conditions

13.1.3.1 Null space basis

Analysis

Theorem 13.3 *Suppose that $f : \mathbb{R}^n \to \mathbb{R}$ is twice partially differentiable with continuous second partial derivatives, $A \in \mathbb{R}^{m \times n}$, and $b \in \mathbb{R}^m$. Let $Z \in \mathbb{R}^{n \times n'}$ be a matrix with columns that form a basis for the null space of A. Let $x^\star \in \mathbb{R}^n$ and suppose that:*

$$\begin{aligned} Z^\dagger \nabla f(x^\star) &= \mathbf{0}, \\ Ax^\star &= b, \\ Z^\dagger \nabla^2 f(x^\star) Z & \quad \text{is positive definite.} \end{aligned}$$

Then $x^\star \in \mathbb{R}^n$ is a strict local minimizer of the problem:

$$\min_{x \in \mathbb{R}^n} \{ f(x) | Ax = b \}.$$

Proof The conditions follow from the second-order sufficient conditions presented in Theorem 10.5 for unconstrained minimization applied to the problem of minimizing the reduced function $\phi : \mathbb{R}^{n'} \to \mathbb{R}$ defined in (13.4):

$$\forall \xi \in \mathbb{R}^{n'}, \phi(\xi) = f(\tau(\xi)).$$

□

Example

- Continuing with the previous equality-constrained Problem (2.13),

$$\begin{aligned} \forall x \in \mathbb{R}^2, \nabla^2 f(x) &= \begin{bmatrix} 2 & 0 \\ 0 & 2 \end{bmatrix}, \\ Z^\dagger \nabla^2 f(x^\star) Z &= \begin{bmatrix} 1 & 1 \end{bmatrix} \begin{bmatrix} 2 & 0 \\ 0 & 2 \end{bmatrix} \begin{bmatrix} 1 \\ 1 \end{bmatrix}, \\ &= [4], \end{aligned}$$

- which is positive definite.
- Applying Theorem 13.3, we conclude that $x^\star$ is a local minimizer of Problem (2.13).

13.1.3.2 Lagrange multipliers

Analysis As previously, we can also develop second-order sufficient conditions in terms of Lagrange multipliers:

Corollary 13.4 *Suppose that $f : \mathbb{R}^n \to \mathbb{R}$ is twice partially differentiable with continuous second partial derivatives, $A \in \mathbb{R}^{m\times n}$, and $b \in \mathbb{R}^m$. Let $x^\star \in \mathbb{R}^n$ and $\lambda^\star \in \mathbb{R}^m$ satisfy:*

$$\begin{aligned} \nabla f(x^\star) + A^\dagger \lambda^\star &= \mathbf{0}, \\ Ax^\star &= b, \\ (A\Delta x = \mathbf{0} \text{ and } \Delta x \neq \mathbf{0}) &\Rightarrow (\Delta x^\dagger \nabla^2 f(x^\star)\Delta x > 0). \end{aligned}$$

Then $x^\star \in \mathbb{R}^n$ is a strict local minimizer of the problem:

$$\min_{x\in\mathbb{R}^n} \{ f(x) | Ax = b \}.$$

Proof The hypotheses of this corollary imply the hypotheses of Theorem 13.3. □

- We refer to the conditions in Corollary 13.4 as the **second-order sufficient conditions** (or **SOSC**).

Example

- Continuing with the previous equality-constrained Problem (2.13),

$$\nabla^2 f(x^\star) = \begin{bmatrix} 2 & 0 \\ 0 & 2 \end{bmatrix},$$

- which is positive definite on $\mathbb{R}^2$ and, in particular, on the null space $\mathcal{N}(A) = \{\Delta x \in \mathbb{R}^n | A\Delta x = \mathbf{0}\}$.
- Applying Corollary 13.4, we conclude that $x^\star$ is a local minimizer of Problem (2.13).

Theorem 13.5 *Suppose that $f : \mathbb{R}^n \to \mathbb{R}$ is partially differentiable with continuous partial derivatives, $A \in \mathbb{R}^{m\times n}$, and $b \in \mathbb{R}^m$. Consider points $x^\star \in \mathbb{R}^n$ and $\lambda^\star \in \mathbb{R}^m$. Suppose that:*

(i) *f is convex on $\{x \in \mathbb{R}^n | Ax = b\}$,*
(ii) *$\nabla f(x^\star) + A^\dagger \lambda^\star = \mathbf{0}$, and*
(iii) *$Ax^\star = b$.*

Then $x^\star$ is a global minimizer of the problem:

$$\min_{x\in\mathbb{R}^n} \{ f(x) | Ax = b \}.$$

Proof Consider any feasible point $x \in \{x \in \mathbb{R}^n | Ax = b\}$. We have:

$$\begin{aligned} f(x) &\geq f(x^\star) + \nabla f(x^\star)^\dagger (x - x^\star), \text{ by Theorem 2.6, noting that:} \\ &\quad f \text{ is partially differentiable with continuous partial derivatives,} \\ &\quad f \text{ is convex on the convex set } \{x \in \mathbb{R}^n | Ax = b\} \text{ by Item (i); and} \\ &\quad x, x^\star \in \{x \in \mathbb{R}^n | Ax = b\} \text{ by Item (iii) of the hypothesis,} \\ &= f(x^\star) - [\lambda^\star]^\dagger A(x - x^\star), \text{ by Item (ii) of the hypothesis,} \\ &= f(x^\star), \text{ since } Ax = Ax^\star = b \text{ by Item (iii) and construction.} \end{aligned}$$

Therefore $x^\star$ is a global minimizer of f on $\{x \in \mathbb{R}^n | Ax = b\}$. $\square$

Corollary 13.6 *Suppose that* $f : \mathbb{R}^n \to \mathbb{R}$ *is partially differentiable with continuous partial derivatives,* $A \in \mathbb{R}^{m \times n}$, *and* $b \in \mathbb{R}^m$. *Let* $Z \in \mathbb{R}^{n \times n'}$ *have columns that form a basis for the null space* $\{\Delta x \in \mathbb{R}^n | A\Delta x = \mathbf{0}\}$. *Consider a point* $x^\star \in \mathbb{R}^n$. *Suppose that:*

(i) f *is convex on* $\{x \in \mathbb{R}^n | Ax = b\}$,
(ii) $Z^\dagger \nabla f(x^\star) = \mathbf{0}$, *and*
(iii) $Ax^\star = b$.

Then $x^\star$ *is a global minimizer of the problem:*

$$\min_{x \in \mathbb{R}^n} \{ f(x) | Ax = b \}.$$

Proof Items (i) and (iii) of the hypothesis of this corollary are the same as the corresponding Items (i) and (iii) of the hypothesis of Theorem 13.5.
Item (ii) of the hypothesis of this corollary says that $Z^\dagger \nabla f(x^\star) = \mathbf{0}$. In the proof of Theorem 13.2, it was proven that:

$$(Z^\dagger \nabla f(x^\star) = \mathbf{0}) \Rightarrow (\exists \lambda^\star \in \mathbb{R}^m \text{ such that } \nabla f(x^\star) + A^\dagger \lambda^\star = \mathbf{0}.)$$

That is, Item (ii) of the hypothesis of Theorem 13.5 holds. Therefore, the result then follows from Theorem 13.5. □

13.2.1.2 Example

- Continuing with the previous equality-constrained Problem (2.13), we have already verified that $x^\star = \begin{bmatrix} 2 \\ 2 \end{bmatrix}$ and $\lambda^\star = [-2]$ satisfy the first-order necessary conditions.
- We also have that f is convex.
- By Theorem 13.5, $x^\star$ is a global minimizer of Problem (2.13).

13.2.2 Duality

- The discussion in Section 13.1.2.4 suggests that if we knew the vector of Lagrange multipliers $\lambda^\star$ we could avoid explicit consideration of the equality constraints if f was convex.
- Here we discuss one method to find the Lagrange multipliers and indicate some of the issues that arise.
- In particular, we will see that we generally require strict convexity of f to yield useful results.

13.2.2.1 Dual function

Analysis

- As we discussed in Section 3.4, we can define a dual problem where the role of variables and constraints is partly or fully swapped.
- Recall Definition 3.3 of the **dual function** and **effective domain**.
- For Problem (13.1), the dual function $\mathcal{D} : \mathbb{R}^m \to \mathbb{R} \cup \{-\infty\}$ is defined by:

$$\forall \lambda \in \mathbb{R}^m, \mathcal{D}(\lambda) = \inf_{x \in \mathbb{R}^n} \mathcal{L}(x, \lambda), \tag{13.13}$$

- while the effective domain is:

$$\mathbb{E} = \{\lambda \in \mathbb{R}^m | \mathcal{D}(\lambda) > -\infty\},$$

- so that the restriction of $\mathcal{D}$ to $\mathbb{E}$ is a function $\mathcal{D} : \mathbb{E} \to \mathbb{R}$.

Example

$$\begin{aligned}
\forall x \in \mathbb{R}^2, \forall \lambda \in \mathbb{R}, \mathcal{L}(x,\lambda) &= (x_1-1)^2+(x_2-3)^2+\lambda(x_1-x_2), \\
\forall \lambda \in \mathbb{R}, \mathcal{D}(\lambda) &= \inf_{x\in\mathbb{R}^2} \mathcal{L}(x,\lambda), \\
&= \inf_{x\in\mathbb{R}^2} \{(x_1-1)^2+(x_2-3)^2+\lambda(x_1-x_2)\}.
\end{aligned}$$

- $\mathcal{L}(\bullet,\lambda)$ is partially differentiable with continuous partial derivatives and is strictly convex.
- By Corollary 10.6 the first-order necessary conditions are sufficient for global optimality:

$$\begin{aligned}
\nabla_x \mathcal{L}(x,\lambda) &= \begin{bmatrix} 2(x_1-1)+\lambda \\ 2(x_2-3)-\lambda \end{bmatrix}, \\
&= \mathbf{0}.
\end{aligned}$$

Example, continued

- For any given $\lambda \in \mathbb{R}$, the unique solution is $x^{(\lambda)} = \begin{bmatrix} 1-\lambda/2 \\ 3+\lambda/2 \end{bmatrix}$.

$$\begin{aligned} \forall \lambda \in \mathbb{R}, \mathcal{D}(\lambda) &= \left(1-\frac{\lambda}{2}-1\right)^2 + \left(3+\frac{\lambda}{2}-3\right)^2 + \lambda\left(1-\frac{\lambda}{2}-3-\frac{\lambda}{2}\right), \\ &= -\frac{(\lambda)^2}{2} - 2\lambda. \end{aligned} \tag{13.14}$$

13.2.2.2 Dual problem

Analysis

- Under certain conditions, Lagrange multipliers can be found as the maximizer, over the dual variables λ, of the following problem:

$$\max_{\lambda \in \mathbb{E}} \mathcal{D}(\lambda). \tag{13.15}$$

- Problem (13.15) is called the **dual problem** to Problem (13.1).
- Problem (13.1) is called the **primal problem**.

Theorem 13.7 *Suppose that $f : \mathbb{R}^n \to \mathbb{R}$ is convex and partially differentiable with continuous partial derivatives, $A \in \mathbb{R}^{m\times n}$, and $b \in \mathbb{R}^m$. Consider primal problem, Problem (13.1):*

$$\min_{x\in\mathbb{R}^n}\{f(x)|Ax = b\}.$$

Also, consider the dual problem, Problem (13.15). We have that:

(i) *If the primal problem possesses a minimum then the dual problem possesses a maximum and the optima are equal. That is:*

$$\min_{x\in\mathbb{R}^n}\{f(x)|Ax = b\} = \max_{\lambda\in\mathbb{E}} \mathcal{D}(\lambda). \tag{13.16}$$

(ii) *If:*

- $\lambda \in \mathbb{E}$,
- $\min_{x\in\mathbb{R}^n} \mathcal{L}(x,\lambda)$ *exists, and*
- f *is twice partially differentiable with continuous second partial derivatives and* $\nabla^2 f$ *is positive definite,*

then $\mathcal{D}$ is partially differentiable at λ with continuous partial derivatives and

$$\nabla\mathcal{D}(\lambda) = Ax^{(\lambda)} - b, \tag{13.17}$$

where $x^{(\lambda)}$ *is the unique minimizer of* $\min_{x \in \mathbb{R}^n} \mathcal{L}(x, \lambda)$.

Proof This is a special case of Theorem 17.4 be presented in Chapter 17. □

- For some $\lambda \in \mathbb{R}^m$ it is possible for $\inf_{x \in \mathbb{R}^n} \mathcal{L}(x, \lambda)$ to be a real number, so that $\lambda \in \mathbb{E}$, yet for there to be no minimum of $\min_{x \in \mathbb{R}^n} \mathcal{L}(x, \lambda)$ or for $\nabla^2 f$ to fail to be positive definite so that there are multiple minimizers of $\min_{x \in \mathbb{R}^n} \mathcal{L}(x, \lambda)$.
- In either case, the dual function $\mathcal{D}$ may be non-differentiable at $\lambda \in \mathbb{E}$.
- Recall from Theorem 3.12 that the effective domain $\mathbb{E}$ of the dual function is a convex set and that the dual function is concave on $\mathbb{E}$.

Corollary 13.8 *Let $f : \mathbb{R}^n \to \mathbb{R}$ be twice partially differentiable with continuous second partial derivatives and with $\nabla^2 f$ positive definite, $A \in \mathbb{R}^{m \times n}$, and $b \in \mathbb{R}^m$. Let $\mathbb{E}$ be the effective domain of the dual function.*
If:

- $\mathbb{E} = \mathbb{R}^m$, *and*
- $\forall \lambda \in \mathbb{R}^m$, $\min_{x \in \mathbb{R}^n} \mathcal{L}(x, \lambda)$ *exists,*

then necessary and sufficient conditions for $\lambda^\star \in \mathbb{R}^m$ to be the maximizer of the dual function are that:

$$Ax^{(\lambda^\star)} - b = \mathbf{0},$$

where $\{x^{(\lambda^\star)}\} = \operatorname{argmin}_{x \in \mathbb{R}^n} \mathcal{L}(x, \lambda^\star)$. Moreover, if $\lambda^\star$ maximizes the dual then $x^{(\lambda^\star)}$ and $\lambda^\star$ satisfy the first-order necessary conditions for Problem (13.1).

Proof Note that the hypothesis implies that the dual function is finite for all λ so that Problem (13.15) is an unconstrained maximization of a real-valued function and, moreover, by Theorem 3.12, $-\mathcal{D}$ is convex and partially differentiable with continuous partial derivatives. By Theorem 10.3 and Corollary 10.6, $\nabla\mathcal{D}(\lambda) = \mathbf{0}$ is necessary and sufficient for λ to be a global maximizer of $\mathcal{D}$. By Theorem 13.7, $\nabla\mathcal{D}(\lambda) = Ax^{(\lambda)} - b$, so the necessary and sufficient conditions for maximizing the dual are that $Ax^{(\lambda)} - b = \mathbf{0}$. Direct substitution shows that $x^{(\lambda^\star)}$ and $\lambda^\star$ satisfy the first-order necessary conditions for Problem (13.1). □

- Theorem 13.7 shows that an alternative approach to finding the minimum of Problem (13.1) involves finding the *maximum* of the dual function over $\lambda \in \mathbb{R}^m$.
- Theorem 3.12 shows that the dual function has at most one local maximum, with necessary and sufficient conditions for the maximizer specified in Corollary 13.8.

Example

- Continuing with the previous equality-constrained Problem (2.13), we note that $\nabla^2 f$ is positive definite and, for each λ, $\mathcal{L}(\bullet, \lambda)$ has a unique minimizer, specified by the solution of $\nabla_x \mathcal{L}(x, \lambda) = \mathbf{0}$, so that, by Theorem 13.7, $\mathbb{E} = \mathbb{R}$ and the dual function is partially differentiable with continuous partial derivatives on the whole of $\mathbb{R}$.
- Moreover, since the dual function is concave, the first-order necessary conditions to maximize $\mathcal{D}$ are also sufficient.
- Partially differentiating $\mathcal{D}$ we obtain:

$$\nabla \mathcal{D}(\lambda) = [-\lambda - 2].$$

- This is consistent with Theorem 13.7, since:

$$\begin{aligned} Ax^{(\lambda)} - b &= [1 \quad -1] \begin{bmatrix} 1 - \lambda/2 \\ 3 + \lambda/2 \end{bmatrix} - [0], \\ &= [-\lambda - 2]. \end{aligned}$$

Example, continued

- Moreover, $\nabla \mathcal{D}(\lambda) = [0]$ for $\lambda^\star = [-2]$.
- Also, $\mathcal{D}(\lambda^\star) = 2$, which is equal to the minimum of Problem (2.13) and $x^{(\lambda^\star)} = \begin{bmatrix} 2 \\ 2 \end{bmatrix}$, which is the minimizer of Problem (2.13).

Wolfe dual

- In some cases we can write down conditions characterizing the value of the dual function more explicitly than in (13.13).
- Suppose that f is partially differentiable with continuous partial derivatives and that it is convex on $\mathbb{R}^n$.
- Then by Corollary 10.6, the first-order necessary conditions $\nabla_x \mathcal{L}(x, \lambda) = 0$ are sufficient for minimizing $\mathcal{L}(\bullet, \lambda)$.
- Given $\lambda \in \mathbb{R}^m$, if there is a solution to $\nabla_x \mathcal{L}(x, \lambda) = \mathbf{0}$ then we can evaluate the dual function by:

$$\mathcal{D}(\lambda) = \{\mathcal{L}(x, \lambda) | \nabla_x \mathcal{L}(x, \lambda) = \mathbf{0}\},$$

- where by the notation on the right-hand side we mean the *value* of $\mathcal{L}(x, \lambda)$ evaluated for a value of x that satisfies $\nabla_x \mathcal{L}(x, \lambda) = \mathbf{0}$, assuming a solution for x exists.

Wolfe dual, continued

- Using Theorem 13.7, this observation means that under the same assumptions, we can solve for the minimum of Problem (13.1) by using the **Wolfe dual**:

$$\min_{x\in\mathbb{R}^n}\{f(x)|Ax=b\} = \max_{\lambda\in\mathbb{R}^m}\{\mathcal{L}(x,\lambda)|\nabla_x\mathcal{L}(x,\lambda)=\mathbf{0}\}, \tag{13.18}$$

- where we again use the equation $\nabla_x\mathcal{L}(x,\lambda)=\mathbf{0}$ to evaluate x and have tacitly assumed that $\nabla_x\mathcal{L}(x,\lambda)=\mathbf{0}$ has a solution for each λ.

Discussion

- It is essential in Theorem 13.7 for f to be convex on the *whole* of $\mathbb{R}^n$, not just on the feasible set.
- The reason is that the inner minimization of $\mathcal{L}(\bullet, \lambda)$ is taken over the whole of $\mathbb{R}^n$.
- Unfortunately, if f is not *strictly* convex then $\mathcal{L}(\bullet, \lambda)$ may have multiple minimizers over x for fixed λ.
- In this case, it may turn out that some of the minimizers of $\mathcal{L}(\bullet, \lambda^\star)$ do not actually minimize (13.1).
- Even when the objective is not strictly convex we can still try to solve the dual problem to obtain $\lambda^\star$ and extract a corresponding value of $x^{(\lambda^\star)}$.
- This approach forms the basis of **Lagrangian relaxation**, the **sub-gradient method**, and other methods to solve non-differentiable problems that result from "dualizing" a problem that has an objective that is not convex or which has a feasible set that is not convex.

13.2.2.3 Separable objective

Analysis

- Suppose that $f : \mathbb{R}^n \to \mathbb{R}$ is additively separable, so that:

$$\forall x \in \mathbb{R}^n, f(x) = \sum_{k=1}^{n} f_k(x_k),$$

- where $f_k : \mathbb{R} \to \mathbb{R}, k = 1, \ldots, n$.
- We consider the dual.

Analysis, continued

$$
\begin{aligned}
\forall \lambda \in \mathbb{E}, \mathcal{D}(\lambda) &= \inf_{x \in \mathbb{R}^n} \mathcal{L}(x, \lambda), \\
&= \min_{x \in \mathbb{R}^n} \mathcal{L}(x, \lambda), \text{ assuming that the minimum exists,} \\
&= \min_{x \in \mathbb{R}^n} f(x) + \lambda^\dagger (Ax - b), \text{ by definition of } \mathcal{L}, \\
&= \min_{x \in \mathbb{R}^n} \left\{ \sum_{k=1}^{n} f_k(x_k) + \lambda^\dagger \left(\sum_{k=1}^{n} A_k x_k - b \right) \right\}, \\
&\quad \text{where } A_k \text{ is the } k\text{-th column of } A, \\
&= \min_{x \in \mathbb{R}^n} \left\{ \sum_{k=1}^{n} \left(f_k(x_k) + \lambda^\dagger A_k x_k \right) \right\} - \lambda^\dagger b, \\
&= \sum_{k=1}^{n} \min_{x_k \in \mathbb{R}} \{ f_k(x_k) + \lambda^\dagger A_k x_k \} - \lambda^\dagger b. \qquad (13.19)
\end{aligned}
$$

Analysis, continued

- For each fixed $\lambda \in \mathbb{R}^m$, the dual function $\mathcal{D}(\lambda)$ is the sum of:

  a constant $(-\lambda^\dagger b)$, and
  n one-dimensional optimization "sub-problems" that can each be evaluated independently.
- We have **decomposed** the problem by exploiting the separability of the objective.
- If there are relatively few constraints but many variables and the objective is separable then maximizing the dual problem involves optimization in a smaller dimension than minimizing the primal problem.

Example

- Continuing with the previous equality-constrained Problem (2.13), note that the objective is separable.
- The dual function is:

$$\begin{aligned} \forall \lambda \in \mathbb{R}, \mathcal{D}(\lambda) &= \min_{x \in \mathbb{R}^2} \mathcal{L}(x, \lambda), \\ &= \min_{x_1 \in \mathbb{R}} \{(x_1 - 1)^2 + \lambda x_1\} + \min_{x_2 \in \mathbb{R}} \{(x_2 - 3)^2 - \lambda x_2\}. \end{aligned} \tag{13.20}$$

- Each of the two convex sub-problems can be solved separately and the result is the same as obtained previously.

13.2.2.4 Penalty functions and augmented Lagrangians

Discussion

- In Section 3.4.5 in discussing duality, we interpreted terms in the Lagrangian as functioning as a penalty.
- In Section 3.1.2.1, we discussed an approach to approximately solving constrained problems by defining an unconstrained problem with a **penalized objective**.
- In Section 3.1.2.1 we also observed that we could consider the penalized objective $f + \Pi \|g\|^2$ for some suitable value of the penalty coefficient $\Pi \in \mathbb{R}_{++}$ and retain the constraints.
- Here we will consider the combined use of penalty functions and duality.

Example

$$\forall x \in \mathbb{R}^2, f(x) = -2(x_1 - x_2)^2 + (x_1 + x_2)^2.$$

- The objective is not convex and is not bounded below.

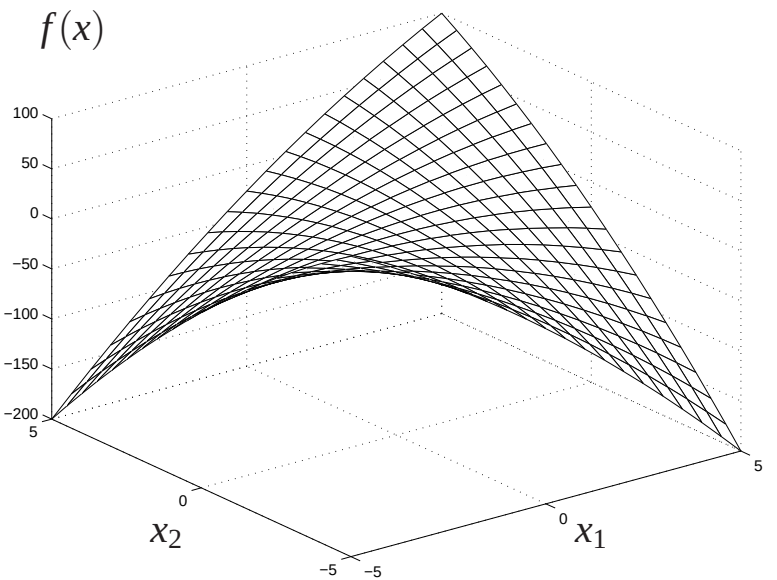

Fig. 13.10. The non-convex objective function defined in section 13.2.2.4.

Example, continued

$$\begin{aligned} A &= [1 \quad -1], \\ b &= [0], \\ \forall x \in \mathbb{R}^2, \forall \lambda \in \mathbb{R}, \mathcal{L}(x,\lambda) &= f(x) + \lambda^\dagger (Ax - b), \\ &= -2(x_1 - x_2)^2 + (x_1 + x_2)^2 + \lambda(x_1 - x_2). \end{aligned}$$

- For any given $\lambda \in \mathbb{R}$, $\mathcal{L}(\bullet, \lambda)$ is not bounded below.
- Therefore:

$$\forall \lambda \in \mathbb{R}, \inf_{x \in \mathbb{R}^n} \mathcal{L}(x,\lambda) = -\infty,$$

- and $\mathbb{E} = \emptyset$.
- We cannot usefully apply Theorem 13.7.

Example, continued

- However, we know that the solution to the equality-constrained optimization problem:

$$\min_{x \in \mathbb{R}^n} \{ f(x) | Ax = b \}$$

- is $x^\star = \mathbf{0}$.
- Substitution into the necessary conditions shows that corresponding value of the Lagrange multiplier is $\lambda^\star = [0]$, so that $\mathcal{L}(\bullet, \lambda^\star) = f(\bullet)$.
- The primal problem is well-defined, the first-order necessary conditions hold at the minimizer, and $x^\star$ and $\lambda^\star$ satisfy the second-order sufficient conditions.
- The difficulties in applying Theorem 13.7 arise here because the objective is not convex on $\mathbb{R}^n$.

Example, continued

- Suppose that instead we consider a penalized objective.
- That is, we modify the objective to be $f + \Pi f_{\mathrm{p}}$, where $\Pi \in \mathbb{R}_{++}$ and $f_{\mathrm{p}} : \mathbb{R}^n \to \mathbb{R}_+$ is defined by:

$$\begin{aligned} \forall x, f_{\mathrm{p}}(x) &= \|Ax - b\|_2^2, \\ &= (x_1 - x_2)^2. \end{aligned}$$

- For example, suppose that we choose $\Pi = 3$.

Example, continued

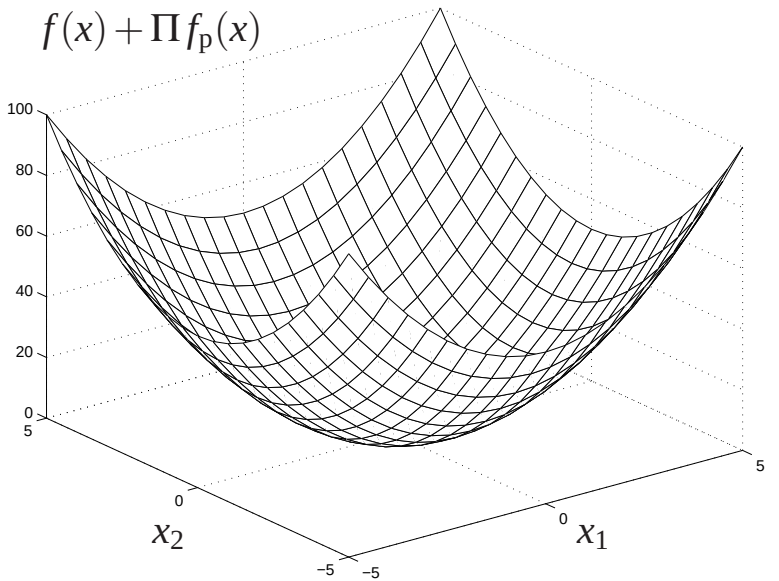

Fig. 13.11. Convex penalized objective function $f+\Pi f_p$ for $\Pi=3$.

Example, continued

- The Lagrangian of the corresponding problem is called the **augmented Lagrangian**, $\mathcal{L}_\mathrm{p} : \mathbb{R}^2 \times \mathbb{R} \to \mathbb{R}$, defined by:

$$\begin{aligned} \forall x \in \mathbb{R}^2, \forall \lambda \in \mathbb{R}, \mathcal{L}_\mathrm{p}(x,\lambda) &= \mathcal{L}(x,\lambda) + \Pi f_\mathrm{p}(x), \\ &= (x_1+x_2)^2 + \lambda(x_1-x_2) + (x_1-x_2)^2, \end{aligned}$$

- which is strictly convex as a function of x for fixed λ.
- Moreover, for each $\lambda \in \mathbb{R}$, the minimizer of $\mathcal{L}_\mathrm{p}(\bullet,\lambda)$ exists, so that $\mathbb{E} = \mathbb{R}$, and the minimizer is unique, so that the dual function is partially differentiable.
- In particular, for the example shown, minimizing $\mathcal{L}_\mathrm{p}(\bullet,\lambda^\star)$ over x now yields the optimal $x^\star$.
- We must pick Π large enough so that:
  - the augmented Lagrangian $\mathcal{L}_\mathrm{p}(\bullet,\lambda)$ is strictly convex for each given λ (so that there is at most one minimizer of $\mathcal{L}_\mathrm{p}(\bullet,\lambda)$ for each given λ), and
  - there is a minimizer of the augmented Lagrangian $\mathcal{L}_\mathrm{p}(\bullet,\lambda)$ for each λ.

Analysis

- Consider a quadratic $f : \mathbb{R}^n \to \mathbb{R}$ with quadratic coefficient matrix $Q \in \mathbb{R}^{n \times n}$ and the use of a penalty function $\Pi \|Ax - b\|_2^2$.
- The Hessian of the augmented Lagrangian $\mathcal{L}_p(\bullet, \lambda)$ for fixed λ is $Q + 2\Pi A^\dagger A$.

Theorem 13.9 *Suppose that $Q \in \mathbb{R}^{n \times n}$ is positive definite on the null-space of $A \in \mathbb{R}^{m \times n}$. Then there exists $\Pi > 0$ such that $Q + 2\Pi A^\dagger A$ is positive definite.* $\square$

- Theorem 13.9 shows that we can find Π such that the augmented Lagrangian is strictly convex as a function of x for fixed λ.
- We typically must apply an adjustment procedure to find a suitable value of Π.

Separability and the augmented Lagrangian

- Augmented Lagrangians have a drawback for separable objectives since the penalty function adds “cross-terms” between variables, which prevent decomposition into sub-problems.
- One approach to preserving separability while maintaining the advantages of augmented Lagrangians involves linearizing the cross-terms.

13.3 Approaches to finding minimizers

- For each case considered, we will tacitly assume that a minimum and minimizer exists.
- The algorithms will either be:
  - direct, typically involving solution of a linear system of equations, or
  - iterative, typically requiring at each iteration the solution of a linear equation representing a Newton–Raphson update for solving non-linear equations or an approximation to the Newton–Raphson update.
- We will proceed as though these Hessians are available and that the resulting linear systems can be conveniently factorized using the basic *LU* factorization.
- However, in practice, it may be necessary or desirable to:
  - use a variation on the basic Newton–Raphson update along the lines described in Section 10.2.3.3 to avoid the computational effort of evaluation and factorization of the Hessian at each iteration, or
  - use a different factorization method such as *QR* if the equations are ill-conditioned.

13.3.1 Convex quadratic objective

13.3.1.1 Problem

$$\forall x \in \mathbb{R}^n, f(x) = \frac{1}{2} x^\dagger Q x + c^\dagger x,$$

- with $c \in \mathbb{R}^n$ and $Q \in \mathbb{R}^{n \times n}$ and symmetric.
- We assume that Q is positive semi-definite, or at least positive semi-definite on the null space $\mathcal{N}(A) = \{\Delta x \in \mathbb{R}^n | A\Delta x = \mathbf{0}\}$.

13.3.1.2 Null space basis

Optimality conditions

- Let $Z \in \mathbb{R}^{n \times n'}$ be a matrix with columns that form a basis for the null space $\mathcal{N}(A) = \{\Delta x \in \mathbb{R}^n | A\Delta x = \mathbf{0}\}$.

$$Z^\dagger Q x^\star = -Z^\dagger c, \tag{13.21}$$

$$Ax^\star = b. \tag{13.22}$$

Algorithm

- Equations (13.21) and (13.22) are linear and involve $n' + m$ equations in n variables.

Example

$$\min_{x\in\mathbb{R}^2}\{f(x)|Ax=b\},$$

$$\forall x\in\mathbb{R}^2, f(x)=\frac{1}{2}x^\dagger Qx+c^\dagger x,$$

$$Q=\begin{bmatrix}2 & 0\\ 0 & 2\end{bmatrix}, c=\begin{bmatrix}-2\\ -6\end{bmatrix}.$$

$$A=\begin{bmatrix}1 & -1\end{bmatrix}, b=[0].$$

$$\begin{aligned} Z &= \begin{bmatrix}1\\ 1\end{bmatrix}\in\mathbb{R}^{2\times 1},\\ Z^\dagger Q &= \begin{bmatrix}1 & 1\end{bmatrix}\begin{bmatrix}2 & 0\\ 0 & 2\end{bmatrix},\\ &= \begin{bmatrix}2 & 2\end{bmatrix},\end{aligned}$$

Example, continued

$$
\begin{aligned}
-Z^\dagger c &= -\begin{bmatrix} 1 & 1 \end{bmatrix} \begin{bmatrix} -2 \\ -6 \end{bmatrix}, \\
&= [8] \\
A &= [1 \quad -1], \\
b &= [0], \\
\begin{bmatrix} 2 & 2 \\ 1 & -1 \end{bmatrix} x^\star &= \begin{bmatrix} 8 \\ 0 \end{bmatrix}, \\
x^\star &= \begin{bmatrix} 2 \\ 2 \end{bmatrix}.
\end{aligned}
$$

Discussion

- The main drawback of this approach is the need to construct the matrix Z and then form and factorize the coefficient matrix $\begin{bmatrix} Z^\dagger Q \\ A \end{bmatrix}$.

13.3.1.3 Lagrange multipliers

Optimality conditions

$$Qx^\star + A^\dagger \lambda^\star = -c, \tag{13.23}$$
$$Ax^\star = b. \tag{13.24}$$

Algorithm

- Equations (13.23) and (13.24) are linear and involve $n+m$ equations in $n+m$ variables.
- The coefficient matrix of this system:

$$\mathcal{A} = \begin{bmatrix} Q & A^\dagger \\ A & \mathbf{0} \end{bmatrix}, \tag{13.25}$$

- is indefinite, so that a special purpose algorithm for factorization of indefinite matrices should be used, as mentioned in Section 5.4.7.
- Performing a single forwards and backwards substitution then solves:

$$\begin{bmatrix} Q & A^\dagger \\ A & \mathbf{0} \end{bmatrix} \begin{bmatrix} x \\ \lambda \end{bmatrix} = \begin{bmatrix} -c \\ b \end{bmatrix}. \tag{13.26}$$

Example

$$\forall x \in \mathbb{R}^2, f(x) = \frac{1}{2} x^\dagger Q x + c^\dagger x,$$

$$Q = \begin{bmatrix} 2 & 0 \\ 0 & 2 \end{bmatrix}, c = \begin{bmatrix} -2 \\ -6 \end{bmatrix}.$$

$$A = \begin{bmatrix} 1 & -1 \end{bmatrix}, b = [0].$$

$$\begin{bmatrix} 2 & 0 & 1 \\ 0 & 2 & -1 \\ 1 & -1 & 0 \end{bmatrix} \begin{bmatrix} x \\ \lambda \end{bmatrix} = \begin{bmatrix} 2 \\ 6 \\ 0 \end{bmatrix}.$$

$$x^\star = \begin{bmatrix} 2 \\ 2 \end{bmatrix}, \lambda^\star = [-2].$$

Discussion

- The coefficient matrix $\mathcal{A}$ in (13.25) is sparse if Q and A are sparse.
- Although (13.23)–(13.24) has more equations than (13.21)–(13.22), if (13.23)–(13.24) is sparse then it can be much easier to solve than (13.21)–(13.22).
- If Q is positive semi-definite but not positive definite, then it may be the case that the minimizer of Problem (13.1) is non-unique. A QR factorization of $\mathcal{A}$ specialized to indefinite matrices can be used. (See Section 5.4.7.)
- If $Q = \mathbf{0}$ so that the problem is actually linear, then it is usually the case that no minimum exists.
- If some of the rows of A are linearly dependent, then the Lagrange multipliers are not unique.

13.3.1.4 Dual maximization

Optimality conditions

- The dual function $\mathcal{D} : \mathbb{R}^m \to \mathbb{R} \cup \{-\infty\}$ is defined by:

$$\forall \lambda \in \mathbb{R}^m, \mathcal{D}(\lambda) = \inf_{x \in \mathbb{R}^n} \left\{ \frac{1}{2} x^\dagger Q x + c^\dagger x + \lambda^\dagger (Ax - b) \right\}. \tag{13.27}$$

- The dual problem is:

$$\max_{\lambda \in \mathbb{E}} \mathcal{D}(\lambda).$$

- The first-order necessary conditions for the unconstrained minimization problem on the right-hand side of (13.27) are:

$$\nabla_x \mathcal{L}(x, \lambda) = Qx + c + A^\dagger \lambda = \mathbf{0}. \tag{13.28}$$

Optimality conditions, continued

- For the rest of the analysis of dual maximization, we will assume that Q is positive definite so that the unconstrained problem on the right-hand side of (13.27) is strictly convex and (13.28) has a unique solution:

$$x^{(\lambda)} = -Q^{-1}(c + A^{\mathsf{T}}\lambda).$$

- The necessary conditions for maximizing the dual are that:

$$\nabla \mathcal{D}(\lambda) = Ax^{(\lambda)} - b = \mathbf{0}.$$

- Each entry in λ can be increased or decreased depending on whether the corresponding entry of $Ax^{(\lambda)} - b$ is greater than or less than zero.

Algorithm

- A steepest ascent algorithm:

$$
\begin{aligned}
x^{(\nu)} &= -Q^{-1}(c + A^{\dagger}\lambda^{(\nu)}), && (13.29)\\
\Delta\lambda^{(\nu)} &= Ax^{(\nu)} - b, && (13.30)\\
\lambda^{(\nu+1)} &= \lambda^{(\nu)} + \alpha^{(\nu)}\Delta\lambda^{(\nu)},
\end{aligned}
$$

- where $\alpha^{(\nu)}$ should be chosen to ensure a *sufficient increase* in $\mathcal{D}(\lambda^{(\nu+1)})$ compared to $\mathcal{D}(\lambda^{(\nu)})$ using, for example, the Armijo criterion described in Section 10.2.4.2.

Stopping criterion

- By Theorem 3.13, $\mathcal{D}(\lambda^{(\nu+1)})$ provides a lower bound on the value of the minimum.
- This lower bound can be incorporated into a stopping criterion.

Example

$$\forall x \in \mathbb{R}^2, f(x) = \frac{1}{2} x^\dagger Q x + c^\dagger x,$$

$$Q = \begin{bmatrix} 2 & 0 \\ 0 & 2 \end{bmatrix}, c = \begin{bmatrix} -2 \\ -6 \end{bmatrix}.$$

$$A = \begin{bmatrix} 1 & -1 \end{bmatrix}, b = [0].$$

- Let $\lambda^{(0)} = [0]$.

Example, continued

$$
\begin{aligned}
x^{(0)} &= -Q^{-1}(c + A^{\dagger}\lambda^{(0)}), \\
&= -\begin{bmatrix} \frac{1}{2} & 0 \\ 0 & \frac{1}{2} \end{bmatrix} \begin{bmatrix} -2 \\ -6 \end{bmatrix}, \\
&= \begin{bmatrix} 1 \\ 3 \end{bmatrix}, \\
\Delta\lambda^{(0)} &= Ax^{(0)} - b, \\
&= \begin{bmatrix} 1 & -1 \end{bmatrix} \begin{bmatrix} 1 \\ 3 \end{bmatrix} - 0, \\
&= [-2].
\end{aligned}
$$

Example, continued

$$
\begin{aligned}
\lambda^{(1)} &= \lambda^{(0)} + \alpha^{(0)} \Delta\lambda^{(0)}, \\
&= [0] + 1 \times [-2], \text{ picking } \alpha^{(0)} = 1, \\
&= [-2], \\
x^{(1)} &= -Q^{-1}(c + A^{\dagger}\lambda^{(1)}), \\
&= -\begin{bmatrix} \frac{1}{2} & 0 \\ 0 & \frac{1}{2} \end{bmatrix} \left(\begin{bmatrix} -2 \\ -6 \end{bmatrix} + \begin{bmatrix} 1 \\ -1 \end{bmatrix} - 2 \right), \\
&= \begin{bmatrix} 2 \\ 2 \end{bmatrix}, \\
\Delta\lambda^{(1)} &= Ax^{(1)} - b, \\
&= [0],
\end{aligned}
$$

- and the dual algorithm has converged in one iteration.
- Usually, the dual iteration using steepest ascent requires more than one iteration to converge, even if an optimal step-size is chosen, because the level sets of the dual function are elliptical and not spherical.

Discussion

- The algorithm adjusts λ until the optimality conditions for the dual are satisfied.
- Maximizing the dual involves:
  - choosing x to satisfy $\nabla f(x) + A^T \lambda^{(\nu)} = \mathbf{0}$ at each iteration, given the current estimate of the Lagrange multiplier, $\lambda^{(\nu)}$, and
  - updating the Lagrange multiplier estimate at each iteration so as to more nearly satisfy the constraint (13.7), that is, $Ax = b$.

13.3.2 Non-quadratic objective

13.3.2.1 Problem

- Suppose that the objective $f : \mathbb{R}^n \to \mathbb{R}$ is partially differentiable with continuous partial derivatives.
- We will consider several approaches to this problem.

Optimality conditions

- Let $Z \in \mathbb{R}^{n \times n'}$ be a matrix with columns that form a basis for the null space $\mathcal{N}(A) = \{\Delta x \in \mathbb{R}^n | A\Delta x = \mathbf{0}\}$.

$$Z^\dagger \nabla f(x^\star) = \mathbf{0}, \tag{13.31}$$

$$Ax^\star = b. \tag{13.32}$$

Algorithm

- Suppose we construct an initial guess $x^{(0)}$ that satisfies the equality constraints.
- The set of all solutions to the linear equations is given by:

$$\{x^{(0)} + Z\xi | \xi \in \mathbb{R}^{n'}\}.$$

- We can now proceed to minimize the reduced function $\phi : \mathbb{R}^{n'} \to \mathbb{R}$ defined by:

$$\forall \xi \in \mathbb{R}^{n'}, \phi(\xi) = f(x^{(0)} + Z\xi).$$

- Any of the unconstrained minimization methods developed in Section 10.2 can be used to minimize this function.
- A natural initial guess for ξ is $\xi^{(0)} = \mathbf{0}$, corresponding to an initial guess of $x^{(0)}$.

Algorithm, continued

- A steepest descent algorithm using the reduced gradient $\nabla\phi$ would involve the following recursion to define the iterates:

$$\xi^{(\nu+1)} = \xi^{(\nu)} - \alpha^{(\nu)} \nabla\phi(\xi^{(\nu)}),$$

- or equivalently:

$$\begin{aligned} \xi^{(\nu+1)} &= \xi^{(\nu)} - \alpha^{(\nu)} Z^{\dagger} \nabla f(x^{(0)} + Z\xi^{(\nu)}), \\ &= \xi^{(\nu)} - \alpha^{(\nu)} Z^{\dagger} \nabla f(x^{(\nu)}), \end{aligned}$$

- where $x^{(\nu)} = x^{(0)} + Z\xi^{(\nu)}$ and the step-size $\alpha^{(\nu)}$ should be chosen to achieve sufficient decrease in the reduced function $\phi(\xi^{(\nu+1)})$ according to, for example, the Armijo criterion.

Algorithm, continued

- A Newton–Raphson algorithm would involve:

$$\begin{aligned} \nabla^2\phi(\xi^{(\nu)})\Delta\xi^{(\nu)} &= -\nabla\phi(\xi^{(\nu)}), \\ \xi^{(\nu+1)} &= \xi^{(\nu)} + \alpha^{(\nu)}\Delta\xi^{(\nu)}, \end{aligned}$$

- or equivalently:

$$\begin{aligned} Z^{\dagger}\nabla^2 f(x^{(\nu)})Z\Delta\xi^{(\nu)} &= -Z^{\dagger}\nabla f(x^{(\nu)}), \\ \xi^{(\nu+1)} &= \xi^{(\nu)} + \alpha^{(\nu)}\Delta\xi^{(\nu)}. \end{aligned}$$

Example

- Consider $f : \mathbb{R}^2 \to \mathbb{R}$ defined by:

$$\forall x \in \mathbb{R}^2, f(x) = 0.01 \times (x_1 - 1)^4 + 0.01 \times (x_2 - 3)^4 + (x_1 - 1)^2 + (x_2 - 3)^2 - 1.8(x_1 - 1)(x_2 - 3).$$

- Consider the problem $\min_{x \in \mathbb{R}^2} \{ f(x) | Ax = b \}$, where $A \in \mathbb{R}^{1 \times 2}$ and $b \in \mathbb{R}^1$ are defined by:

$$\begin{aligned} A &= [1 \quad -1], \\ b &= [8]. \end{aligned}$$

- By inspection, $Z = \begin{bmatrix} 1 \\ 1 \end{bmatrix}$ is a matrix with columns that form a basis for the null space of A.

Example, continued

- Consider the initial guess $x^{(0)} = \begin{bmatrix} 3 \\ -5 \end{bmatrix}$, which is feasible for the equality constraint.
- We perform one iteration of a steepest descent algorithm to minimize the reduced function with initial guess $\xi^{(0)} = [0]$.

$$
\begin{aligned}
f(x^{(0)}) &= 137.77, \\
\nabla f(x) &= \begin{bmatrix} 0.04 \times (x_1 - 1)^3 + 2 \times (x_1 - 1) - 1.8 \times (x_2 - 3) \\ 0.04 \times (x_2 - 3)^3 - 1.8 \times (x_1 - 1) + 2 \times (x_2 - 3) \end{bmatrix}, \\
\nabla f(x^{(0)}) &= \begin{bmatrix} 18.72 \\ -40.08 \end{bmatrix}, \\
Z^{\dagger} \nabla f(x^{(0)}) &= [-21.36].
\end{aligned}
$$

Example, continued

- Using a step-size of 1, we obtain a tentative update of:

$$\begin{aligned} \xi^{(1)} &= \xi^{(0)} - Z^{\dagger}\nabla f(x^{(0)}), \\ &= [21.36], \\ x^{(0)} + Z\xi^{(1)} &= \begin{bmatrix} 24.36 \\ 16.36 \end{bmatrix}, \\ f(x^{(0)} + Z\xi^{(1)}) &= 3458.8. \end{aligned}$$

- This is larger than $f(x^{(0)})$, so we must consider a step-size rule.
- We use the Armijo rule, with the step-size halved until the Armijo condition (10.14) is satisfied.
- For $\alpha^{(0)} = 0.25$, the Armijo condition is satisfied and we obtain:

$$\begin{aligned} \xi^{(1)} &= [5.34], \\ x^{(1)} &= \begin{bmatrix} 8.34 \\ 0.34 \end{bmatrix}, \\ f(x^{(1)}) &= 125.6. \end{aligned}$$

Stopping criterion

- The algorithm involved unconstrained minimization of the reduced function ϕ.
- Stopping criteria for unconstrained problems as discussed in Section 10.2.5 can be used for this algorithm.

Discussion

- Whatever algorithm is used for minimizing ϕ, at each iteration the iterate $x^{(\nu)} = x^{(0)} + Z\xi^{(\nu)}$ is feasible for the equality constraints.
- In summary, we generate iterates that are:
  - feasible at each iteration, satisfying (13.32), and
  - in principle, become closer to satisfying the condition (13.31).

13.3.2.3 Lagrange multipliers

Optimality conditions

$$\nabla f(x^\star) + A^\dagger \lambda^\star = \mathbf{0}, \tag{13.33}$$

$$Ax^\star - b = \mathbf{0}. \tag{13.34}$$

Algorithm

- Equations (13.33)–(13.34) are non-linear in $\begin{bmatrix} x \\ \lambda \end{bmatrix}$, involve $n+m$ equations in $n+m$ variables, and can be solved iteratively using the Newton–Raphson method.
- Ideally, convergence is quadratic.
- Since the constraints $Ax = b$ are linear, we can construct an initial point $x^{(0)}$ that satisfies $Ax^{(0)} = b$ using the techniques discussed in Section 5.8.1.

Algorithm, continued

- At each subsequent iteration, we try to solve for the Newton–Raphson step direction using:

$$\begin{aligned} \mathcal{A}\begin{bmatrix} \Delta x^{(\nu)} \\ \Delta\lambda^{(\nu)} \end{bmatrix} &= -\begin{bmatrix} \nabla f(x^{(\nu)}) + A^{\dagger}\lambda^{(\nu)} \\ Ax^{(\nu)} - b \end{bmatrix}, \\ &= -\begin{bmatrix} \nabla f(x^{(\nu)}) + A^{\dagger}\lambda^{(\nu)} \\ \mathbf{0} \end{bmatrix}, \end{aligned} \tag{13.35}$$

- where $\mathcal{A} \in \mathbb{R}^{(n+m)\times(n+m)}$ is defined by:

$$\mathcal{A} = \begin{bmatrix} \nabla^2 f(x^{(\nu)}) & A^{\dagger} \\ A & \mathbf{0} \end{bmatrix}$$

- and where we have assumed that $Ax^{(\nu)} - b = \mathbf{0}$.

Example

- Continuing with the non-quadratic objective $f : \mathbb{R}^2 \to \mathbb{R}$ we consider the initial guess $x^{(0)} = \begin{bmatrix} 3 \\ -5 \end{bmatrix}$ and $\lambda^{(0)} = [0]$.
- We perform one Newton–Raphson update.
- The coefficient matrix $\mathcal{A}$ and right-hand side in (13.35) is given by:

$$\begin{aligned} \mathcal{A} &= \begin{bmatrix} \nabla^2 f(x^{(0)}) & A^\dagger \\ A & \mathbf{0} \end{bmatrix}, \\ &= \begin{bmatrix} 2.48 & -1.8 & 1 \\ -1.8 & 9.68 & -1 \\ 1 & -1 & 0 \end{bmatrix}, \\ -\begin{bmatrix} \nabla f(x^{(\nu)}) + A^\dagger \lambda^{(\nu)} \\ \mathbf{0} \end{bmatrix} &= \begin{bmatrix} -18.72 \\ 40.08 \\ 0 \end{bmatrix}. \end{aligned}$$

Example, continued

- Solving (13.35) for these values yields:

$$\begin{bmatrix} \Delta x^{(\nu)} \\ \Delta \lambda^{(\nu)} \end{bmatrix} = \begin{bmatrix} 2.4953 \\ 2.4953 \\ -20.4168 \end{bmatrix}.$$

- Using a step-size of one, we obtain:

$$\begin{aligned} x^{(1)} &= \begin{bmatrix} 5.4953 \\ -2.5047 \end{bmatrix}, \\ \lambda^{(1)} &= [-20.4168], \end{aligned}$$

- with objective value $f(x^{(1)}) = 108.3163$.

Stopping criterion

- Suppose that:

  f is convex,

  there is a known bound on where the minimizer can lie of the form $\left\| x^\star - x^{(\nu)} \right\| \leq \overline{\rho}$, and

  we want to ensure that $f(x^{(\nu)})$ is within ε_f of the minimum.

- Then we should iterate until $\left\| \nabla f(x^{(\nu)}) + A^\dagger \lambda^{(\nu)} \right\| \leq \varepsilon_f / \overline{\rho}$.

Discussion

- As in the case of the quadratic objective, even if $\nabla^2 f(x^{(\nu)})$ is positive definite, the coefficient matrix $\mathcal{A}$ in (13.35) is indefinite.
- To factorize it, we should use a special purpose algorithm as mentioned in Section 5.4.7.
- The iterates are:
  - feasible at each iteration, satisfying (13.7), and
  - in principle, become closer to satisfying the condition (13.6).

13.3.2.4 Dual maximization

Optimality conditions

- The dual function in this case is $\mathcal{D} : \mathbb{R}^m \to \mathbb{R} \cup \{-\infty\}$ defined by:

$$\forall \lambda \in \mathbb{R}^m, \mathcal{D}(\lambda) = \inf_{x \in \mathbb{R}^n} \{ f(x) + \lambda^\dagger (Ax - b) \}.$$

- The dual problem is:

$$\max_{\lambda \in \mathbb{E}} \mathcal{D}(\lambda).$$

- If we assume that there is a minimum and minimizer of the primal problem and that there is no duality gap, then maximizing the dual function yields the minimum of the primal problem.
- If the conditions of Corollary 13.8 hold then the optimality conditions for the dual problem are that:

$$\begin{aligned} \nabla \mathcal{D}(\lambda) &= A x^{(\lambda)} - b, \\ &= \mathbf{0}, \end{aligned}$$

- where $x^{(\lambda)}$ is the unique minimizer of $\min_{x \in \mathbb{R}^n} \{ f(x) + \lambda^\dagger (Ax - b) \}$.

Algorithm

$$
\begin{aligned}
x^{(\nu)} &\in \arg\min_{x\in\mathbb{R}^n}\{f(x)+[\lambda^{(\nu)}]^{\dagger}(Ax-b)\}, \qquad (13.36)\\
\Delta\lambda^{(\nu)} &= Ax^{(\nu)}-b,\\
\lambda^{(\nu+1)} &= \lambda^{(\nu)}+\alpha^{(\nu)}\Delta\lambda^{(\nu)}.
\end{aligned}
$$

Example

- Continuing with the objective defined in (10.9),

$$\forall x \in \mathbb{R}^2, f(x) = 0.01 \times (x_1 - 1)^4 + 0.01 \times (x_2 - 3)^4 + (x_1 - 1)^2 + (x_2 - 3)^2 - 1.8(x_1 - 1)(x_2 - 3),$$

- and constraints defined by:

$$A = [1 \quad -1],$$
$$b = [8],$$

- we let $\lambda^{(0)} = [0]$, and perform one (outer) iteration of a dual maximization algorithm.

Example, continued

- Since $\lambda^{(0)} = [0]$, Problem (13.36) is equivalent to unconstrained minimization of f.
- The minimizer is $x^{(0)} = \begin{bmatrix} 1 \\ 3 \end{bmatrix}$, and we have:

$$\begin{aligned} \Delta\lambda^{(0)} &= Ax^{(0)} - b, \\ &= \begin{bmatrix} 1 & -1 \end{bmatrix} \begin{bmatrix} 1 \\ 3 \end{bmatrix} - [8], \\ &= [-10]. \end{aligned}$$

- Using a step-size of $\alpha^{(0)} = 1$, this yields:

$$\begin{aligned} \lambda^{(1)} &= \lambda^{(0)} + \alpha^{(0)} \Delta\lambda^{(0)}, \\ &= [0] + 1[-10], \\ &= [-10]. \end{aligned}$$

Stopping criterion

- Again Theorem 3.13 can be used to show that $\mathcal{D}(\lambda^{(\nu+1)})$ provides a lower bound on the value of the minimum.

Discussion

- Maximizing the dual involves:
  - satisfying $\nabla f(x^{(\nu)}) + A^{\dagger}\lambda^{(\nu)} = \mathbf{0}$ at each outer iteration, given the current estimate of the Lagrange multiplier, $\lambda^{(\nu)}$, and
  - updating the Lagrange multiplier estimate at each outer iteration so as to more nearly satisfy the constraint (13.7).
- For each update of λ there are a number of inner iterations to solve Problem (13.36) to sufficient accuracy.
- Once a minimizer of Problem (13.36) is obtained then, to update λ, $\alpha^{(\nu)}$ should be chosen to yield a *sufficient increase* in the dual function using, for example, the Armijo condition as described in Section 10.2.4.2.

13.4 Sensitivity

- We imagine that we have solved the problem:

$$\min_{x \in \mathbb{R}^n} \{ f(x;\chi) | A(\chi)x = b(\chi) \},$$

- for the base-case value of the parameters $\chi = \mathbf{0}$.
- We now consider the sensitivity of the minimizer and minimum to variation of the parameters around $\chi = \mathbf{0}$.

13.4.1 General case

Corollary 13.10 *Let $f : \mathbb{R}^n \times \mathbb{R}^s \to \mathbb{R}$ be twice partially differentiable with continuous second partial derivatives and let $A : \mathbb{R}^s \to \mathbb{R}^{m\times n}$ and $b : \mathbb{R}^s \to \mathbb{R}^m$ be partially differentiable with continuous partial derivatives. Consider the minimization problem:*

$$\min_{x\in\mathbb{R}^n}\{ f(x;\chi)|A(\chi)x = b(\chi)\}, \tag{13.37}$$

where χ is a parameter. Suppose that $x^\star \in \mathbb{R}^n$ is a local minimizer of Problem (13.37) for the base-case value of the parameters $\chi = \mathbf{0}$ with corresponding Lagrange multipliers $\lambda^\star \in \mathbb{R}^m$. We call $x = x^\star$ a base-case minimizer and call $\lambda = \lambda^\star$ the base-case Lagrange multipliers. Define the (parameterized) Hessian $\nabla^2_{xx} f : \mathbb{R}^n \times \mathbb{R}^s \to \mathbb{R}^{n\times n}$ by:

$$\forall x \in \mathbb{R}^n, \forall \chi \in \mathbb{R}^s, \nabla^2_{xx} f(x;\chi) = \frac{\partial^2 f}{\partial x^2}(x;\chi).$$

Suppose that:

- $\nabla^2_{xx} f(x^\star; \mathbf{0})$ *is positive definite on the null space of* $A(\mathbf{0})$*, so that* $x^\star$ *and* $\lambda^\star$ *satisfy the second-order sufficient conditions for the base-case problem, and*
- $A(\mathbf{0})$ *has linearly independent rows.*

Then, for values of χ *in a neighborhood of the base-case value of the parameters* $\chi = \mathbf{0}$*, there is a local minimum and corresponding local minimizer and Lagrange multipliers for Problem (13.37). Moreover, the local minimum, local minimizer, and Lagrange multipliers are partially differentiable with respect to* χ *and have continuous partial derivatives in this neighborhood.*

We consider the sensitivity with respect to χ_j, the j-th entry of χ. The sensitivity of the local minimizer $x^\star$ and Lagrange multipliers $\lambda^\star$ to χ_j, evaluated at the base-case $\chi = \mathbf{0}$, is given by the solution of:

$$\mathcal{A}\begin{bmatrix} \dfrac{\partial x^\star}{\partial \chi_j}(\mathbf{0}) \\ \dfrac{\partial \lambda^\star}{\partial \chi_j}(\mathbf{0}) \end{bmatrix} = \begin{bmatrix} -K_j(x^\star;\mathbf{0}) - \left[\dfrac{\partial A}{\partial \chi_j}(\mathbf{0})\right]^\dagger \lambda^\star \\ -\dfrac{\partial A}{\partial \chi_j}(\mathbf{0})x^\star + \dfrac{\partial b}{\partial \chi_j}(\mathbf{0}) \end{bmatrix}, \tag{13.38}$$

where:

$$\mathcal{A} = \begin{bmatrix} \nabla^2_{xx} f(x^\star;\mathbf{0}) & [A(\mathbf{0})]^\dagger \\ A(\mathbf{0}) & \mathbf{0} \end{bmatrix},$$

and $K_j : \mathbb{R}^n \times \mathbb{R}^s \to \mathbb{R}^n$ is defined by:

$$\forall x \in \mathbb{R}^n, \forall \chi \in \mathbb{R}^s, K_j(x;\chi) = \frac{\partial^2 f}{\partial x \partial \chi_j}(x;\chi).$$

The sensitivity of the local minimum $f^\star$ to χ, evaluated at the base-case $\chi = \mathbf{0}$, is given by:

$$\frac{\partial f^\star}{\partial \chi}(\mathbf{0}) = \frac{\partial \mathcal{L}}{\partial \chi}(x^\star, \lambda^\star; \mathbf{0}),$$

where $\mathcal{L} : \mathbb{R}^n \times \mathbb{R}^m \times \mathbb{R}^s \to \mathbb{R}$ is the parameterized Lagrangian defined by:

$$\forall x \in \mathbb{R}^n, \forall \lambda \in \mathbb{R}^m, \forall \chi \in \mathbb{R}^s, \mathcal{L}(x, \lambda; \chi) = f(x; \chi) + \lambda^\dagger (A(\chi)x - b(\chi)).$$

If $f(\bullet; \chi)$ is convex for χ in a neighborhood of $\mathbf{0}$ then the minimizers and minima are global in this neighborhood.

Proof The sensitivity of the local minimizer follows from Corollary 7.5, noting that:

- the Hessian $\nabla^2_{xx} f$ is positive definite on the null space of $A(\chi)$ for x in a neighborhood of the base-case minimizer $x^\star$ and χ in a neighborhood of $\chi = \mathbf{0}$, and
- the coefficient matrix $\mathcal{A}$ is non-singular in a neighborhood of the base-case minimizer and parameters,

so that the first-order necessary conditions (13.6)–(13.7) for Problem (13.37) are well-defined and satisfied in a neighborhood of $\chi = \mathbf{0}$ and the sensitivity of the first-order necessary conditions at $\chi = \mathbf{0}$ is given by the solution of (13.38). Moreover, the second-order sufficient conditions for Problem (13.37) given in Corollary 13.4 are satisfied in this neighborhood.

The sensitivity of the local minimum follows by totally differentiating the value of the local minimum $f^\star(\chi) = f(x^\star(\chi);\chi)$ with respect to χ and noting that the first-order necessary conditions for the local minimizer mean that $\dfrac{\partial f}{\partial x}(x^\star;\mathbf{0}) = -[\lambda^\star]^\dagger A(\mathbf{0})$. But:

$$A(\mathbf{0})\frac{\partial x^\star}{\partial \chi}(\mathbf{0}) = -\frac{\partial A}{\partial \chi}(\mathbf{0})x^\star + \frac{\partial b}{\partial \chi}(\mathbf{0}), \tag{13.39}$$

by the second block row of (13.38) evaluated for $j = 1,\ldots,s$ and where, abusing notation, we interpret $\dfrac{\partial A}{\partial \chi}(\mathbf{0})x^\star \in \mathbb{R}^{m\times s}$ as having ℓj-th entry equal to $\sum_{k=1}^n \dfrac{\partial A_{\ell k}}{\partial \chi_j}(\mathbf{0})x_k^\star$.

Therefore,

$$\begin{aligned}
\frac{\partial f^\star}{\partial \chi}(\mathbf{0}) &= \frac{\partial f}{\partial x}(x^\star;\mathbf{0})\frac{\partial x^\star}{\partial \chi}(\mathbf{0}) + \frac{\partial f}{\partial \chi}(x^\star;\mathbf{0}), \text{ since } f^\star(\chi) = f(x^\star(\chi);\chi), \\
&= -[\lambda^\star]^\dagger A(\mathbf{0})\frac{\partial x^\star}{\partial \chi}(\mathbf{0}) + \frac{\partial f}{\partial \chi}(x^\star;\mathbf{0}), \\
&= \frac{\partial f}{\partial \chi}(x^\star;\mathbf{0}) - [\lambda^\star]^\dagger \left(-\frac{\partial A}{\partial \chi}(\mathbf{0})x^\star + \frac{\partial b}{\partial \chi}(\mathbf{0}) \right), \text{ by (13.39)}, \\
&= \frac{\partial \mathcal{L}}{\partial \chi}(x^\star, \lambda^\star;\mathbf{0}).
\end{aligned}$$

□

- The sensitivity of the local minimum is again called the **envelope theorem**.

13.4.2 Special case

Corollary 13.11 *Consider Problem (13.1), a perturbation vector $\gamma \in \mathbb{R}^m$, and a perturbed version of Problem (13.1) defined by:*

$$\min_{x \in \mathbb{R}^n} \{ f(x) | Ax = b - \gamma \}. \tag{13.40}$$

Suppose that $f : \mathbb{R}^n \to \mathbb{R}$ is twice partially differentiable with continuous second partial derivatives, $A \in \mathbb{R}^{m \times n}$, and $b \in \mathbb{R}^m$, with the rows of A linearly independent. Let $x^\star \in \mathbb{R}^n$ and $\lambda^\star \in \mathbb{R}^m$ satisfy the second-order sufficient conditions in Corollary 13.4 for Problem (13.1):

$$\begin{aligned} \nabla f(x^\star) + A^\dagger \lambda^\star &= \mathbf{0}, \\ Ax^\star &= b, \\ ((A\Delta x = \mathbf{0}) \text{ and } (\Delta x \neq \mathbf{0})) &\Rightarrow (\Delta x^\dagger \nabla^2 f(x^\star) \Delta x > 0). \end{aligned}$$

Consider Problem (13.40). For values of γ in a neighborhood of the base-case value of the parameters $\gamma = \mathbf{0}$, there is a local minimum and corresponding local minimizer and Lagrange multipliers for Problem (13.40). Moreover, the local minimum, local minimizer, and

Lagrange multipliers are partially differentiable with respect to γ and have continuous partial derivatives in this neighborhood. The sensitivity of the local minimum to γ, evaluated at the base-case $\gamma = \mathbf{0}$, is equal to $[\lambda^\star]^\dagger$. If f is convex then the minimizers and minima are global. $\square$

13.4.3 Discussion

- A significant part of the effort in proving Corollary 13.10 and Corollary 13.11 is using the implicit function theorem to show that the sensitivity of the minimizer is well-defined.
- If we assume that the minimizer and minimum are partially differentiable with respect to χ, then the following argument explains why the sensitivity is given by the value of the Lagrange multipliers.
- Consider Problem (13.40), a perturbation γ, and the corresponding change $\Delta x^\star$ in the minimizer of the perturbed problem.
- The change in the minimum is:

$$\begin{aligned} f(x^\star + \Delta x^\star) - f(x^\star) &\approx \nabla f(x^\star)^\dagger \Delta x^\star, \text{ with equality as } \Delta x^\star \to \mathbf{0}, \\ &= -[\lambda^\star]^\dagger A \Delta x^\star, \text{ by the first-order} \\ &\quad \text{necessary condition } \nabla f(x^\star) + A^\dagger \lambda^\star = \mathbf{0}, \\ &= [\lambda^\star]^\dagger \gamma, \end{aligned}$$

- since $A(x^\star + \Delta x^\star) = b - \gamma$, so that $-A\Delta x^\star = \gamma$.
- But this is true for any such perturbation γ. In the limit as $\gamma \to \mathbf{0}$, the change in the minimum approaches $[\lambda^\star]^\dagger \gamma$.

Discussion, continued

- We can interpret the Lagrange multipliers as the sensitivity of the minimum to changes in γ.
- In many problems, the specification of constraints represents some judgment about the availability of resources.
- Then we can use the Lagrange multipliers to help in trading off the change in the optimal objective against the cost of the purchase of additional resources.

13.4.4 Example

- Consider the equality-constrained Problem (2.13) from Section 2.3.2.2:

$$\min_{x \in \mathbb{R}^2} \{ f(x) | Ax = b \},$$

$$\begin{aligned} \forall x \in \mathbb{R}^2, f(x) &= (x_1 - 1)^2 + (x_2 - 3)^2, \\ A &= [1 \quad -1], \\ b &= [0]. \end{aligned}$$

- Suppose that the equality constraints changed from $Ax = b$ to $Ax = b - \gamma$.
- Then, if γ is small enough, the minimum of the perturbed problem differs from the minimum of the original problem by approximately $[\lambda^\star]^\dagger \gamma = (-2)\gamma$.

13.5 Solution of the least-cost production case study

13.5.1 Problem

$$\min_{x \in \mathbb{R}^n} \{ f(x) | Ax = b \}.$$

- Suppose that $n = 3$.
- Then the coefficient matrix and right-hand side can be specified as:

$$\begin{aligned} A &= [-1 \quad -1 \quad -1], \\ b &= [-D]. \end{aligned}$$

- In summary, this problem has a convex separable objective and only one equality constraint.
- Furthermore, the equality constraint is linear.
- That is, the problem is convex.

13.5.2 Algorithms

13.5.2.1 Null space basis

- We first construct an initial guess $x^{(0)}$ that is feasible for the equality constraint:

$$x^{(0)} = \begin{bmatrix} D \\ 0 \\ 0 \end{bmatrix}.$$

- A matrix Z with columns that form a basis for the null space is $Z = \begin{bmatrix} -1 & -1 \\ 1 & 0 \\ 0 & 1 \end{bmatrix}$.
- We can form the reduced gradient and update ξ to decrease the reduced objective.
- This is equivalent to expressing x_1 in terms of x_2 and x_3 as discussed in Section 12.1.4.2.

13.5.2.2 Lagrange multipliers

$$\begin{aligned} \forall k = 1, \ldots, n, \frac{d f_k}{d x_k}(x_k^\star) - \lambda^\star &= 0, \\ D - \sum_{k=1}^{n} x_k^\star &= 0. \end{aligned}$$

13.5.2.3 Dual maximization

$$
\begin{aligned}
\forall k = 1, \ldots, n, x_k^{(\nu)} &\in \underset{x_k \in \mathbb{R}}{\arg\min} \{ f_k(x_k) - \lambda^{(\nu)} x_k \}, \qquad (13.41) \\
\Delta\lambda^{(\nu)} &= A x^{(\nu)} - b, \\
&= D - \sum_{k=1}^{n} x_k^{(\nu)}, \\
\lambda^{(\nu+1)} &= \lambda^{(\nu)} + \alpha^{(\nu)} \Delta\lambda^{(\nu)}.
\end{aligned}
$$

- If f_k is quadratic then, at each iteration ν, the k-th sub-problem on the right-hand side of (13.41) can be solved directly in one step by solving the linear necessary conditions.
- If f_k is not quadratic then (13.41) can be solved by applying the Newton–Raphson update until a value of $x_k^{(\nu)}$ is obtained that satisfies the necessary conditions to within a tolerance.
- That is, if f_k is non-quadratic, then at each outer iteration ν and for each k we must perform several inner iterations to solve the necessary conditions of (13.41).

13.5.3 Discussion

- Maximizing the dual has a suggestive economic interpretation if we think of λ as the price paid for producing the commodity.
- The values $\lambda^{(\nu)}$ are tentative prices that are suggested at each iteration by a central purchaser.
- The goal of the central purchaser is to pick prices such that supply matches demand.
- The Lagrange multiplier $\lambda^\star$ is the final price that matches supply to demand.
- Each cost function f_k is associated with a decision-making agent that makes decisions based on:

  its own cost function, and
  the tentative prices.

Discussion, continued

- Each decision-making agent sells a quantity of product x_k to maximize its profits, which is equivalent to minimizing the difference between:

  the *cost* of production $f_k(x_k)$ for the quantity x_k, minus
  the *revenues* $x_k\lambda^{(\nu)}$, based on the current value of the dual variable, $\lambda^{(\nu)}$.

- The solution of (13.41) maximizes the agent's *profit*, that is, revenues minus costs, for the given value of the dual variable.
- At each iteration, the central agent adjusts the tentative prices based on comparing the sum of offered productions by the agents to the target value D:
  - price is raised or lowered to encourage or discourage production.
- At the optimum, the "marginal cost of production" for each agent, that is, the derivative of its cost function, is the same for all agents.
- The value of the Lagrange multiplier is sometimes called the **shadow price**.

13.6 Summary

- We have discussed descent directions for linear equality-constrained optimization problems.
- Analysis of descent directions yielded optimality conditions, which in turn led to algorithms.
- We also discussed sensitivity analysis.
- Finally, we discussed solution of the least-cost production case study.

14

Algorithms for non-linear equality-constrained minimization

- In this chapter we will develop algorithms for constrained optimization problems of the form:

$$\min_{x \in \mathbb{R}^n} \{ f(x) | g(x) = \mathbf{0} \}, \tag{14.1}$$

- where $f : \mathbb{R}^n \to \mathbb{R}$ and $g : \mathbb{R}^n \to \mathbb{R}^m$.

Key issues

- The notion of a **regular point of constraints** as a characterization of suitable formulations of non-linear equality constraint functions,
- linearization of non-linear constraint functions and consideration of the **null space of the coefficient matrix** of the linearized constraints and the associated **tangent plane**,
- optimality conditions and the definition and interpretation of the **Lagrange multipliers**,
- algorithms that seek points that satisfy the optimality conditions,
- use of a **merit function** in the trade-off between satisfaction of constraints and improvement of the objective, and
- **duality** and **sensitivity analysis**.

14.1 Geometry and analysis of constraints

- In the case of *linear* equality constraints, the convexity of the feasible set allowed us to consider step directions such that successive iterates were always feasible.
- With non-linear constraints, movement from a feasible point along a line segment will usually take us outside the feasible set.
- Nevertheless, our approach to non-linear equality constraints will be to linearize the equality constraint function g about a current iterate.
- We must explore conditions under which this linearization yields a useful approximation to the original feasible set.

14.1.1 Regular point of constraints

14.1.1.1 Definition

Definition 14.1 Let $g : \mathbb{R}^n \to \mathbb{R}^m$. Then we say that $x^\star$ is a **regular point** of the equality constraints $g(x) = \mathbf{0}$ if:

(i) $g(x^\star) = \mathbf{0}$,

(ii) g is partially differentiable with continuous partial derivatives at $x^\star$, and

(iii) the m rows of the Jacobian $J(x^\star)$ of g evaluated at $x^\star$ are linearly independent.

□

14.1.1.2 Example

- Consider the function $g : \mathbb{R}^3 \to \mathbb{R}$ defined by:

$$\forall x \in \mathbb{R}^3, g(x) = (x_1)^2 + (x_2+1)^2 - x_3 - 4,$$

- and the point $x^\star = \begin{bmatrix} 1 \\ 3 \\ 13 \end{bmatrix}$.
- We observe that $x^\star = \begin{bmatrix} 1 \\ 3 \\ 13 \end{bmatrix}$ is a regular point of the equality constraints $g(x) = 0$ because:
  - (i) $g(x^\star) = (1)^2 + (3+1)^2 - 13 - 4 = 0$,
  - (ii) g is partially differentiable with Jacobian $J : \mathbb{R}^3 \to \mathbb{R}^{1\times 3}$ defined by $\forall x \in \mathbb{R}^3, J(x) = \begin{bmatrix} 2x_1 & 2(x_2+1) & -1 \end{bmatrix}$, which is continuous at $x^\star$, and
  - (iii) the one row of the Jacobian $J(x^\star)$ of g evaluated at $x^\star$ is $J(x^\star) = \begin{bmatrix} 2 & 8 & -1 \end{bmatrix}$, which is a linearly independent row.

14.1.2 Tangent plane

14.1.2.1 Definition

- We make the following generalization of Definition 13.1.

Definition 14.2 Let $g : \mathbb{R}^n \to \mathbb{R}^m$ be partially differentiable and $x^\star \in \mathbb{R}^n$. Let $J : \mathbb{R}^n \to \mathbb{R}^{m \times n}$ be the Jacobian of g. Suppose that $x^\star$ is a regular point of the constraints $g(x) = \mathbf{0}$. Then the **tangent plane** to the set $\mathbb{S} = \{x \in \mathbb{R}^n | g(x) = \mathbf{0}\}$ at the point $x^\star$ is the set $\mathbb{T} = \{x \in \mathbb{R}^n | J(x^\star)(x - x^\star) = \mathbf{0}\}$. □

- The tangent plane at $x^\star$ is the set of points such that the first-order Taylor approximation to g about $x^\star$ has value $\mathbf{0}$.

14.1.2.2 Example

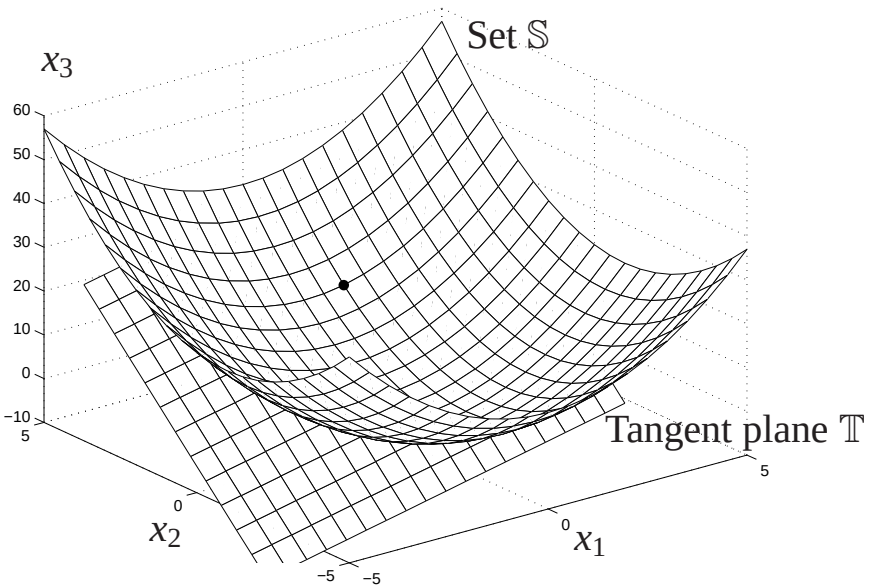

Fig. 14.1. Tangent plane $\mathbb{T}$ to a set $\mathbb{S}$ in $\mathbb{R}^3$ at the point $x^\star = \begin{bmatrix} 1 \\ 3 \\ 13 \end{bmatrix} \in \mathbb{S}$, shown as a ●.

14.1.2.3 Affine case

- In the case that g is affine of the form:

$$\forall x \in \mathbb{R}^n, g(x) = Ax - b,$$

then $J(x) = A$ and the tangent plane $\mathbb{T}$ at a point $x^\star \in \mathbb{S} = \{x \in \mathbb{R}^n | Ax = b\}$ is given by:

$$\begin{aligned} \mathbb{T} &= \{x \in \mathbb{R}^n | A(x - x^\star) = \mathbf{0}\}, \\ &= \{x \in \mathbb{R}^n | Ax = b\}, \\ &= \mathbb{S}, \end{aligned}$$

- since $Ax^\star = b$ at a feasible point $x^\star$.
- That is, in the case that g is affine, the tangent plane $\mathbb{T}$ is the same as the feasible set $\mathbb{S} = \{x \in \mathbb{R}^n | g(x) = \mathbf{0}\}$.
- In contrast, for non-linear g such as shown in Figure 14.1, the tangent plane $\mathbb{T}$ to $\{x \in \mathbb{R}^n | g(x) = \mathbf{0}\}$ at $x^\star$ is usually different to $\mathbb{S} = \{x \in \mathbb{R}^n | g(x) = \mathbf{0}\}$.

14.1.2.4 Discussion

- The concept of a regular point will help us to characterize when the tangent plane $\mathbb{T}$ is a good approximation to the feasible set $\mathbb{S}$.

14.1.3 Relationship of regular points to seeking minimizers

14.1.3.1 Movement from a feasible point

- If $x^\star$ is a regular point of $g(x) = \mathbf{0}$ then we will be close to satisfying the constraints so long as we stay near to $x^\star$ and in the tangent plane $\mathbb{T} = \{x \in \mathbb{R}^n | J(x^\star)(x - x^\star) = \mathbf{0}\}$.
- Consider $n = 2$ and $f : \mathbb{R}^2 \to \mathbb{R}$ and $g : \mathbb{R}^2 \to \mathbb{R}$ defined by:

$$\forall x \in \mathbb{R}^2, f(x) = -x_1, \tag{14.2}$$

$$\forall x \in \mathbb{R}^2, g(x) = x_2 - \sin(x_1). \tag{14.3}$$

- Figure 14.2 shows part of the set of points $\mathbb{S}$ satisfying the equality constraint $g(x) = 0$ as a solid curve.
- Also shown is the feasible point $x^\star = \begin{bmatrix} 5 \\ \sin(5) \end{bmatrix}$, shown as a $\circ$, and the tangent plane $\mathbb{T}$ to the feasible set $\mathbb{S}$ at $x^\star$, shown dashed.
- For this problem, the tangent plane $\mathbb{T}$ is only a good approximation to the feasible for points that are close to $x^\star$.

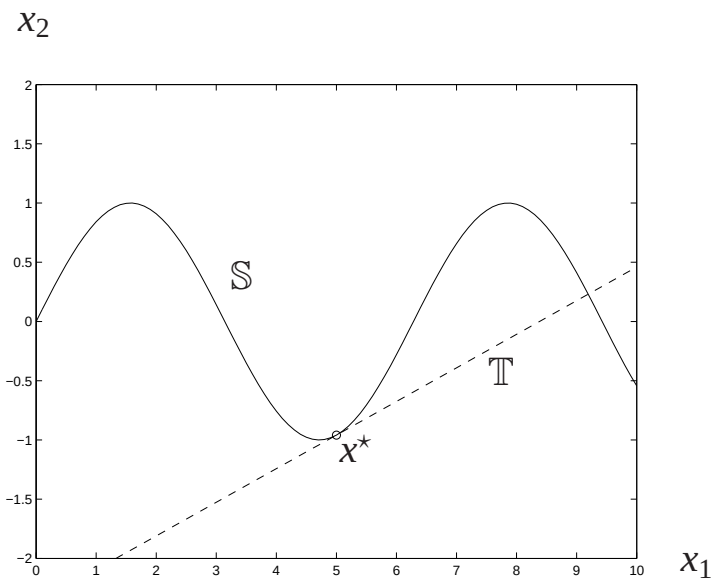

Fig. 14.2. Feasible point $x^\star \in \mathbb{S}$ and tangent plane $\mathbb{T}$ (shown dashed) to $\mathbb{S}$ at $x^\star$ (shown solid).

14.1.3.2 Descent

- Figure 14.3 shows the same feasible set as illustrated in Figure 14.2.
- The arrows emanating from the feasible points illustrate directions along the tangent plane at these points.
- Moving along these directions takes us outside the feasible set but reduces the objective f.
- Paths that stay on the feasible set must follow the curve $g(x) = \mathbf{0}$ and therefore depart from straight line segments.
- Nevertheless, we will consider paths that, at least initially, follow the tangent plane $\mathbb{T}$.

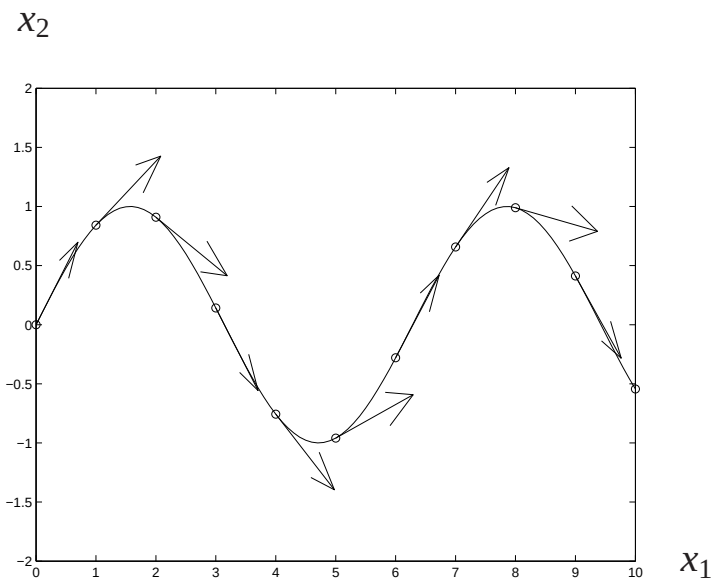

Fig. 14.3. Feasible points and directions along the corresponding tangent planes.

14.1.3.3 Movement from an infeasible point

- We will again approximate the feasible points by a set defined in terms of linear equalities:

$$\mathbb{T} = \{x \in \mathbb{R}^n | J(\hat{x})(x - \hat{x}) = -g(\hat{x})\}. \tag{14.4}$$

- Linear independence of the rows of J and proximity of $\hat{x}$ to the feasible set will guarantee that this set closely approximates the feasible set in the vicinity of $\hat{x}$.
- Figure 14.4 again shows the part of the set of points satisfying the equality constraint $g(x) = 0$.
- Also shown is an infeasible point $\hat{x} = \begin{bmatrix} 5 \\ -1.5 \end{bmatrix}$ and the set $\mathbb{T}$ defined according to (14.4).
- In this particular case, the set $\mathbb{T}$ is tangential to the feasible set $\mathbb{S}$; however, in general this is not the case.

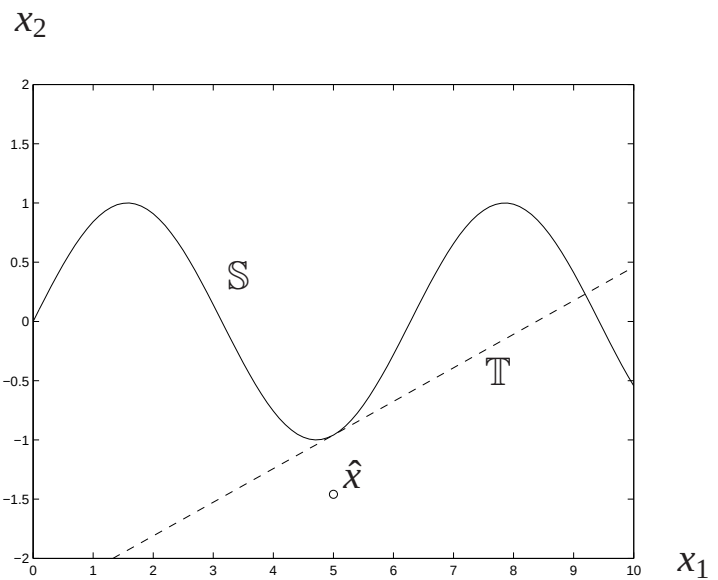

Fig. 14.4. Movement from an infeasible point $\hat{x} \in \mathbb{S}$ and approximation $\mathbb{T}$ (shown dashed) to feasible set $\mathbb{S}$ (shown solid).

14.1.3.4 Linear constraints

- If g is affine and $g(x^\star) = \mathbf{0}$ then $\mathbb{T} = \{x \in \mathbb{R}^n | A(x - x^\star) = \mathbf{0}\}$ is the same as the feasible set, whether or not A has linearly independent rows.
- However, if the linear coefficient matrix does not have linearly independent rows, then a slight perturbation of the coefficient matrix will make the linear approximation to the feasible set empty.

14.1.3.5 Formulation of problems

- Whether or not g is affine, we should try to formulate the problem to avoid linear dependence of the rows of J since, analogously to the case of simultaneous equations, redundant linearized constraints make the linearized problem ill-conditioned.

14.2 Optimality conditions

14.2.1 First-order necessary conditions

14.2.1.1 Analysis

Theorem 14.1 *Consider Problem (14.1) and a point $x^\star \in \mathbb{R}^n$. Suppose that:*

(i) *f is partially differentiable with continuous partial derivatives,*

(ii) *$x^\star$ is a regular point of the equality constraints $g(x) = \mathbf{0}$. That is:*

(a) *$g(x^\star) = \mathbf{0}$,*

(b) *g is partially differentiable with continuous partial derivatives, and*

(c) *the m rows of the Jacobian $J(x^\star)$ of g evaluated at $x^\star$ are linearly independent.*

Then if $x^\star$ is a local minimizer of Problem (14.1) then:

$$\exists \lambda^\star \in \mathbb{R}^m \text{ such that } \nabla f(x^\star) + J(x^\star)^\dagger \lambda^\star = \mathbf{0}. \tag{14.5}$$

□

Analysis, continued

- The vector $\lambda^\star$ is again called the vector of **Lagrange multipliers** for the constraints $g(x) = \mathbf{0}$.
- We will refer to:

$$\nabla f(x^\star) + J(x^\star)^\dagger \lambda^\star = \mathbf{0}, \tag{14.6}$$

$$g(x^\star) = \mathbf{0}, \tag{14.7}$$

- as the **first-order necessary conditions** or **FONC**.

14.2.1.2 Lagrangian

- Recall Definition 3.2 of the **Lagrangian**.
- For Problem (14.1) the Lagrangian $\mathcal{L} : \mathbb{R}^n \times \mathbb{R}^m \to \mathbb{R}$ is defined by:

$$\forall x \in \mathbb{R}^n, \forall \lambda \in \mathbb{R}^m, \mathcal{L}(x, \lambda) = f(x) + \lambda^\dagger g(x).$$

- As in the linear case, we can reproduce the first-order necessary conditions (14.6)–(14.7) by setting the gradients of $\mathcal{L}$ with respect to x and λ, respectively, equal to zero.

14.2.1.3 Relationship to linearly constrained problems

- The condition (14.5) is the same as the corresponding first-order condition for the *linearly* constrained problem:

$$\min_{x \in \mathbb{R}^n} \{ f(x) | J(x^\star)(x - x^\star) = \mathbf{0} \}. \tag{14.8}$$

- Regularity of $x^\star$, in addition to the hypotheses for the linear case, ensures that (14.5) characterizes the necessary conditions in the non-linear equality-constrained case.
- Unlike in the linear case, the assumption of regularity is important to ensure that there are Lagrange multipliers satisfying (14.5).

14.2.1.4 Geometric interpretation

- In the linear equality-constrained case, we interpreted the first-order necessary conditions as requiring that the feasible set be a subset of the tangent plane to the contour set of the objective.
- We said that the contour set of f was tangential to the feasible set at $x^\star$.
- In the non-linear equality-constrained case, we can similarly interpret (14.5) as requiring that the feasible set and the contour set be tangential at $x^\star$.

14.2.1.5 Example

- If the objective is non-convex then a maximizer can satisfy the necessary conditions.
- In the case of non-linear equality constraints, however, we may have an objective $f : \mathbb{R}^n \to \mathbb{R}$ that is convex on $\mathbb{R}^n$, but have a non-convex feasible set.
- For example:

$$\forall x \in \mathbb{R}^2, f(x) = \frac{1}{2}(x_1)^2 + \frac{1}{2}(x_2)^2, \tag{14.9}$$

$$\forall x \in \mathbb{R}^2, g(x) = \frac{1}{4}(x_1)^2 + (x_2)^2 - 1. \tag{14.10}$$

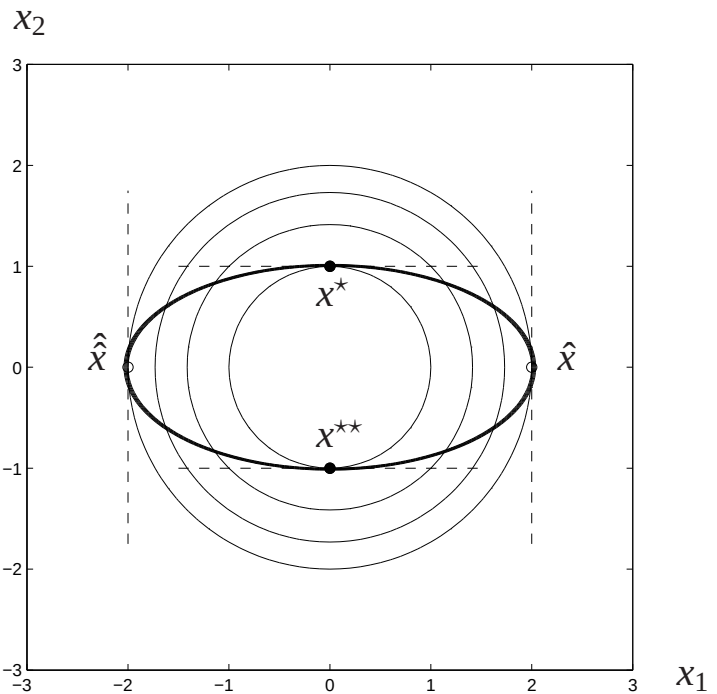

Fig. 14.5. Points $x^\star$, $x^{\star\star}$, $\hat{x}$, and $\hat{\hat{x}}$ that satisfy the first-order necessary conditions but which may or may not be minimizers.

Example, continued

- There are four points that satisfy the first-order necessary conditions.
  - Two of the points are $x^\star = \begin{bmatrix} 0 \\ 1 \end{bmatrix}$ and $x^{\star\star} = \begin{bmatrix} 0 \\ -1 \end{bmatrix}$, both with Lagrange multiplier $\lambda^\star = \lambda^{\star\star} = [-0.5]$, which corresponds to a *minimum* $f^\star = 0.5$ of the objective over the feasible set. The points $x^\star$ and $x^{\star\star}$ are illustrated with $\bullet$ in Figure 14.5.
  - The other two points are $\hat{x} = \begin{bmatrix} 2 \\ 0 \end{bmatrix}$ and $\hat{\hat{x}} = \begin{bmatrix} -2 \\ 0 \end{bmatrix}$, both with dual variables $\hat{\lambda} = \hat{\hat{\lambda}} = [-2]$, which corresponds to a *maximum* $\hat{f} = 2$ of the objective over the feasible set. The points $\hat{x}$ and $\hat{\hat{x}}$ are illustrated with $\circ$ in Figure 14.5.

14.2.2 Second-order sufficient conditions

14.2.2.1 Analysis

Theorem 14.2 *Suppose that $f : \mathbb{R}^n \to \mathbb{R}$ and $g : \mathbb{R}^n \to \mathbb{R}^m$ are twice partially differentiable with continuous second partial derivatives. Let $J : \mathbb{R}^n \to \mathbb{R}^{m\times n}$ be the Jacobian of g. Consider Problem (14.1) and points $x^\star \in \mathbb{R}^n$ and $\lambda^\star \in \mathbb{R}^m$. Suppose that:*

$$\begin{aligned} \nabla f(x^\star) + J(x^\star)^\dagger \lambda^\star &= \mathbf{0}, \\ g(x^\star) &= \mathbf{0}, \end{aligned}$$

$\nabla^2 f(x^\star) + \sum_{\ell=1}^{m} \lambda_\ell^\star \nabla^2 g_\ell(x^\star)$ is positive definite on the null space:

$$\mathcal{N} = \{\Delta x \in \mathbb{R}^n | J(x^\star)\Delta x = \mathbf{0}\}. \quad (14.11)$$

Then $x^\star$ is a strict local minimizer of Problem (14.1). □

- Compared to the first-order necessary conditions, the second-order sufficient conditions in addition require that:
  - the objective and constraint functions are twice partially differentiable with continuous second partial derivatives, and
  - $x^\star$ and $\lambda^\star$ satisfy (14.11).
- In (14.11), the function $\nabla^2_{xx}\mathcal{L} : \mathbb{R}^n \times \mathbb{R}^m \to \mathbb{R}^{n\times n}$ defined by:

$$\forall x \in \mathbb{R}^n, \forall \lambda \in \mathbb{R}^m, \nabla^2_{xx}\mathcal{L}(x,\lambda) = \nabla^2 f(x) + \sum_{\ell=1}^{m} \lambda_\ell \nabla^2 g_\ell(x),$$

- is called the **Hessian of the Lagrangian**.
- The condition (14.11) is analogous to the corresponding condition in Corollary 13.4 for linear constraints.
- It requires that the Hessian of the Lagrangian evaluated at the minimizer and corresponding Lagrange multipliers, $\nabla^2_{xx}\mathcal{L}(x^\star, \lambda^\star)$, be positive definite on the null space $\mathcal{N}$ defined in the theorem.

14.2.2.2 Example

- Continuing with the example from Section 14.2.1.5, the two minimizers $x^\star$ and $x^{\star\star}$ satisfy the second-order sufficient conditions.
- However, both of the other points, $\hat{x}$ and $\hat{\hat{x}}$, that satisfy the first-order necessary conditions do not satisfy the second-order sufficient conditions.

14.3 Approaches to finding minimizers

- If the constraints are non-linear, we cannot expect to exactly satisfy them.
- We can consider algorithms that attempt to satisfy the first-order necessary conditions or use step directions based on the Newton–Raphson update for solving the first-order necessary conditions.

14.3.1 Solution of first-order necessary conditions

$$\begin{aligned} \nabla f(x) + J(x)^{\dagger}\lambda &= \mathbf{0}, \\ g(x) &= \mathbf{0}. \end{aligned}$$

14.3.1.1 Newton–Raphson step direction

$$\begin{bmatrix} \nabla^2_{xx}\mathcal{L}(x^{(\nu)},\lambda^{(\nu)}) & J(x^{(\nu)})^\dagger \\ J(x^{(\nu)}) & \mathbf{0} \end{bmatrix} \begin{bmatrix} \Delta x^{(\nu)} \\ \Delta\lambda^{(\nu)} \end{bmatrix} = - \begin{bmatrix} \nabla f(x^{(\nu)}) + J(x^{(\nu)})^\dagger \lambda^{(\nu)} \\ g(x^{(\nu)}) \end{bmatrix}. \tag{14.12}$$

- An indefinite factorization algorithm should be used.
- As discussed for the unconstrained case in Section 10.2.3.2, zero and negative pivots in the top left-hand block should be modified to be positive to ensure that $\Delta x^{(\nu)}$ is a descent direction for $f + [\lambda^{(\nu)}]^\dagger g$ at $x^{(\nu)}$.
- We can approximate the solution of (14.12).
- The approximations will, in principle, inherit the corresponding convergence rates described for the solution of non-linear equations.
- The update is then:

$$\begin{bmatrix} x^{(\nu+1)} \\ \lambda^{(\nu+1)} \end{bmatrix} = \begin{bmatrix} x^{(\nu)} \\ \lambda^{(\nu)} \end{bmatrix} + \alpha^{(\nu)} \begin{bmatrix} \Delta x^{(\nu)} \\ \Delta\lambda^{(\nu)} \end{bmatrix}.$$

14.3.1.2 Selection of step-size

- In choosing a step-size, we cannot just seek reduction in f because if we are far from satisfying the constraints then we may have to accept an increase in f to obtain a feasible point.
- We must trade-off the tension between satisfaction of the constraints and improvement in the objective.
- A standard approach to this trade-off is to define a **merit function** $\phi : \mathbb{R}^n \to \mathbb{R}$ of the form, for example:

$$\forall x \in \mathbb{R}^n, \phi(x) = f(x) + \Pi \|g(x)\|^2, \tag{14.13}$$

- for some norm $\|\bullet\|$ and some $\Pi \in \mathbb{R}_{++}$ and use a rule analogous to the Armijo rule or variants to seek a step that leads to sufficient reduction in the merit function ϕ at each iteration.
- We will discuss the choice of Π in the context of the power system state estimation with zero injection buses case study in Section 14.5.

Selection of step-size, continued

- A variant on the merit function approach is to replace the objective in Problem (14.1) with the merit function $\phi : \mathbb{R}^n \to \mathbb{R}$ defined in (14.13).
- In this case, the Newton–Raphson update will explicitly seek a direction that reduces the merit function.
- There are other approaches including:
  - a **filter**, where the step-size is selected to improve satisfaction of the constraints or the value of the objective or both at each iteration, and
  - a **watchdog**, where the merit function is allowed to increase for a limited number of iterations.

14.3.1.3 Feasibility

- In some applications, we might want to be able to terminate at any iteration with an iterate that is close to being feasible.
- In this case, at each iteration we can first update x to reduce the objective or reduce a merit function and then do a subsidiary search using an iterative technique to return to the feasible set.
- This approach is used in the **generalized reduced gradient** algorithm.

14.3.1.4 Stopping criteria

- We iterate until the first-order necessary conditions are satisfied to sufficient accuracy.
- Unless the second-order sufficient conditions hold or approximately hold, we cannot be certain that we are at or close to a local optimum.

14.3.2 Dual maximization

- Recall Definition 3.3 of the **dual function**.
- For Problem (14.1), the dual function $\mathcal{D} : \mathbb{R}^m \to \mathbb{R} \cup \{-\infty\}$ is defined by:

$$\forall \lambda \in \mathbb{R}^m, \mathcal{D}(\lambda) = \inf_{x \in \mathbb{R}^n} \mathcal{L}(x, \lambda),$$

- where $\mathcal{L} : \mathbb{R}^n \times \mathbb{R}^m \to \mathbb{R}$ is the Lagrangian.
- Although the problem is not convex, we can try to maximize the dual function.
- The following recursion can be used to define the iterates:

$$\begin{aligned} x^{(\nu)} &\in \operatorname{argmin}_{x \in \mathbb{R}^n} \{ f(x) + [\lambda^{(\nu)}]^\dagger g(x) \}, \\ \Delta\lambda^{(\nu)} &= g(x^{(\nu)}), \\ \lambda^{(\nu+1)} &= \lambda^{(\nu)} + \alpha^{(\nu)} \Delta\lambda^{(\nu)}. \end{aligned}$$

- If f or g is non-quadratic then we will have to perform several inner iterations to approximately minimize the Lagrangian for each outer iteration to update λ.

Dual maximization, continued

- There can be a duality gap.
- Nevertheless, by Theorem 3.13, the maximum of the dual is a lower bound for the minimum of the primal problem and the solution of the dual can be a useful guide to the solution of the primal.

14.4 Sensitivity

14.4.1 Analysis

- Suppose that the objective f and equality constraint function g are parameterized by a parameter $\chi \in \mathbb{R}^s$.
- We imagine that we have solved the non-linear equality-constrained minimization problem:

$$\min_{x \in \mathbb{R}^n} \{ f(x;\chi) | g(x;\chi) = \mathbf{0} \}, \tag{14.14}$$

- for a base-case value of the parameters, say $\chi = \mathbf{0}$, to find the base-case local minimizer $x^\star$ and the base-case Lagrange multipliers $\lambda^\star$.
- We now consider the sensitivity of the local minimum of Problem (14.14) to variation of the parameters about $\chi = \mathbf{0}$.
- That is, we also consider perturbations $\gamma \in \mathbb{R}^m$ and the problem:

$$\min_{x \in \mathbb{R}^n} \{ f(x) | g(x) = -\gamma \}. \tag{14.15}$$

- For the parameter values $\gamma = \mathbf{0}$, Problem (14.15) is the same as Problem (14.1).

Corollary 14.3 *Consider Problem (14.14) and suppose that the functions $f : \mathbb{R}^n \times \mathbb{R}^s \to \mathbb{R}$ and $g : \mathbb{R}^n \times \mathbb{R}^s \to \mathbb{R}^m$ are twice partially differentiable with continuous second partial derivatives. Also consider Problem (14.15) and suppose that the functions $f : \mathbb{R}^n \to \mathbb{R}$ and $g : \mathbb{R}^n \to \mathbb{R}^m$ are twice partially differentiable with continuous second partial derivatives. Suppose that $x^\star \in \mathbb{R}^n$ and $\lambda^\star \in \mathbb{R}^m$ satisfy:*

- *the second-order sufficient conditions for Problem (14.14) for the base-case value of parameters $\chi = \mathbf{0}$, and*
- *the second-order sufficient conditions for Problem (14.15) for the base-case value of parameters $\gamma = \mathbf{0}$.*

In particular:

- *$x^\star$ is a local minimizer of Problem (14.14) for $\chi = \mathbf{0}$, and*
- *$x^\star$ is a local minimizer of Problem (14.15) for $\gamma = \mathbf{0}$,*

in both cases with associated Lagrange multipliers $\lambda^\star$. Moreover, suppose that the rows of the Jacobians $J(x^\star; \mathbf{0})$ and $J(x^\star)$, respectively, are linearly independent so that $x^\star$ is a regular point of the constraints for the base-case problems.

Then, for values of χ in a neighborhood of the base-case value of the parameters $\chi = \mathbf{0}$, there is a local minimum and corresponding local minimizer and Lagrange multipliers for Problem (14.14). Moreover, the local minimum, local minimizer, and Lagrange multipliers are partially differentiable with respect to χ and have continuous partial derivatives in this neighborhood. The sensitivity of the local minimum $f^\star$ to χ, evaluated at the base-case $\chi = \mathbf{0}$, is given by:

$$\frac{\partial f^\star}{\partial \chi}(\mathbf{0}) = \frac{\partial \mathcal{L}}{\partial \chi}(x^\star, \lambda^\star; \mathbf{0}),$$

where $\mathcal{L} : \mathbb{R}^n \times \mathbb{R}^m \times \mathbb{R}^s \to \mathbb{R}$ is the **parameterized Lagrangian** *defined by:*

$$\forall x \in \mathbb{R}^n, \forall \lambda \in \mathbb{R}^m, \forall \chi \in \mathbb{R}^s, \mathcal{L}(x, \lambda; \chi) = f(x; \chi) + \lambda^\dagger g(x; \chi).$$

Furthermore, for values of γ in a neighborhood of the base-case value of the parameters $\gamma = \mathbf{0}$, there is a local minimum and corresponding local minimizer and Lagrange multipliers for Problem (14.15). Moreover, the local minimum, local minimizer, and Lagrange multipliers are partially differentiable with respect to γ and have continuous partial derivatives. The sensitivity of the local minimum to γ, evaluated at the base-case $\gamma = \mathbf{0}$, is equal to $[\lambda^\star]^\dagger$.

□

14.4.2 Discussion

- As in the case of linear equality constraints, we can interpret the Lagrange multipliers as the sensitivity of the minimum to changes in γ.
- Again, this allows us to trade-off the change in the optimal objective against the cost of changing the constraint.

14.4.3 Example

- Consider the example equality-constrained Problem (2.14) first mentioned in Section 2.3.2:

$$\min_{x \in \mathbb{R}^2} \{ f(x) | g(x) = 0 \},$$

- where $f : \mathbb{R}^2 \to \mathbb{R}$ and $g : \mathbb{R}^2 \to \mathbb{R}$ were defined by:

$$\begin{aligned} \forall x \in \mathbb{R}^2, f(x) &= (x_1 - 1)^2 + (x_2 - 3)^2, \\ \forall x \in \mathbb{R}^2, g(x) &= (x_1)^2 + (x_2)^2 + 2x_2 - 3. \end{aligned}$$

- The minimizers and Lagrange multipliers of Problem (2.14) satisfy the second-order sufficient conditions and the minimizers are regular points of the constraints.
- If the equality constraint changes to $g(x) = -\gamma$, where $\gamma = 0.1$, then we can use Corollary 14.3 to approximate the change in the minimum by $0.1\lambda^\star$.

14.5 Solution of power system state estimation with zero injection buses case study

14.5.1 Problem

- Recall Problem (12.9):

$$\min_{x\in\mathbb{R}^n}\{f(x)|g(x)=\mathbf{0}\},$$

- where $f:\mathbb{R}^n\to\mathbb{R}$ and $g:\mathbb{R}^n\to\mathbb{R}^m$ were defined in (12.7) and (12.8), respectively:

$$\begin{aligned}\forall x\in\mathbb{R}^n, f(x) &= \sum_{\ell\in\mathbb{M}}\frac{(\tilde{g}_\ell(x)-\tilde{G}_\ell)^2}{2\sigma_\ell^2},\\ \forall x\in\mathbb{R}^n, g(x) &= (g_\ell(x))_{\ell\in\mathbb{M}^0}.\end{aligned}$$

14.5.2 Algorithms

14.5.2.1 Newton–Raphson step direction

- The most straightforward way to solve this problem is to seek a solution of the necessary conditions (14.6)–(14.7) using the Newton–Raphson step direction given by the solution of (14.12) or some approximation to it that ensures that a descent direction is found for $f + [\lambda^{(\nu)}]^{\dagger} g$.
- Possible approximations to the coefficient matrix for the Newton–Raphson step direction include:
  - using the fast-decoupled or other approximations to the Jacobian of the power flow equations, as in the discussion of the solution of the power flow equations in Section 8.2.4.2, and
  - using the Gauss–Newton or Levenberg–Marquardt approximation to the Hessian of the objective, as in the discussion of the state estimation problem in Section 11.2.3.2.

14.5.2.2 Merit function and step-size

- f consists of (half of) the sum of squares of terms each of which represent a measurement error for measurement ℓ divided by the standard deviation σ_ℓ of the measurement error.
- Consequently, each term has expected value of 1 if evaluated at the true value of the voltage angles and magnitudes in the system.
- The terms in g represent real and reactive power values that are exactly equal to zero when evaluated at the true value of the voltage angles and magnitudes in the system.
- We can use a merit function (14.13) with the L_2 norm $\|\bullet\|_2$ and a value of penalty coefficient Π that is somewhat larger than the inverse of the square of a typical real and reactive power measurement error standard deviation.
- We can interpret the merit function as being a penalized objective, as discussed in Section 12.2.1.3, that uses modest values of the penalty coefficient.
- The step-size should be selected to ensure sufficient reduction in (14.13) using the Armijo rule.

14.5.2.3 Observability

- To ensure that there is a unique maximum likelihood estimator there must be enough measurements and zero bus injections spread around the system to make it observable.

14.6 Summary

- In this chapter we considered the notion of a regular point of constraints as a bridge between equality-constrained problems with linear constraints and equality-constrained problems with non-linear constraints.
- We developed optimality conditions, algorithms, and sensitivity analysis.
- We then applied the algorithms to the power system state estimation with zero injection buses case study.