AMSC 607 / CMSC 764 Advanced Numerical Optimization
Fall 2008
UNIT 3: Constrained Optimization
PART 1: Characterizing a solution
Dianne P. O'Leary

Fundamentals for constrained optimization

- Characterizing a solution
- Duality

Our approach: Always try to reduce the problem to one with a known solution.

Reference: N&S, Chapter 14

Our problem

$$\min_{\mathbf{x}} f(\mathbf{x})$$

$$\begin{aligned} c_i(\mathbf{x}) &= 0\,, \quad i \in \mathcal{E} \\ c_i(\mathbf{x}) &\geq 0\,, \quad i \in \mathcal{I} \end{aligned}$$

where f and c_i are $\mathcal{C}^2$ functions from $\mathcal{R}^n$ into $\mathcal{R}^1$.

Definition of a solution

We say that $\mathbf{x}^*$ is a solution to our problem if

- $\mathbf{x}^*$ satisfies all of the constraints.
- For some $\epsilon > 0$, if $\|\mathbf{y} - \mathbf{x}^*\| \leq \epsilon$, and if $\mathbf{y}$ satisfies the constraints, then $f(\mathbf{y}) \geq f(\mathbf{x}^*)$.

In other words, $\mathbf{x}^*$ is feasible and locally optimal.

The plan

We will develop necessary and sufficient optimality conditions so that we can recognize solutions and develop algorithms to find solutions.

We do this in several stages.

- Case 1: Linear equality constraints only.
- Case 2: Linear inequality constraints.
- Case 3: General constraints.

Then we will discuss duality.

Case 1: Optimality Conditions for Linear equality constraints only

---

Our problem

Reference: Some of this material can be found in N&S Chapter 3.

Our problem:

$$\min_{\mathbf{x}} f(\mathbf{x})$$

$$\mathbf{A}\mathbf{x} = \mathbf{b}$$

where $\mathbf{A}$ is a matrix of dimension $m \times n$.

We also assume a constraint qualification or regularity condition: assume that $\mathbf{A}$ has rank m.

Unquiz:

- What happens if $\mathbf{A}$ has rank n?
- What happens if $\mathbf{A}$ has rank less than m?

[]

An example

Let

$$\begin{aligned} f(\mathbf{x}) &= x_1^2 - 2x_1x_2 + x_2^2 \\ c_1(\mathbf{x}) &= x_1 + x_2 - 1 = \begin{bmatrix} 1 & 1 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} - 1 \end{aligned}$$

We'll consider two approaches to the problem.

Approach 1: Variable Reduction

If $x_1 + x_2 = 1$, then all feasible points have the form

$$\begin{bmatrix} x_1 \\ 1 - x_1 \end{bmatrix}.$$

Therefore, the possible function values are

$$\begin{aligned} f(\mathbf{x}) &= x_1^2 - 2x_1 x_2 + x_2^2 \\ &= x_1^2 - 2x_1(1 - x_1) + (1 - x_1)^2 \end{aligned}$$

We now have an unconstrained minimization problem involving a function of a single variable, and we know how to solve this!

picture

This is called the reduced variable method.

Approach 2: The feasible direction formulation

If $x_1 + x_2 = 1$, then all feasible points have the form

$$\mathbf{x} = \begin{bmatrix} 0 \\ 1 \end{bmatrix} + \alpha \begin{bmatrix} 1 \\ -1 \end{bmatrix}.$$

This formulation works because

$$\mathbf{A}\mathbf{x} = \begin{bmatrix} 1 & 1 \end{bmatrix} x = \begin{bmatrix} 1 & 1 \end{bmatrix} \begin{bmatrix} 0 \\ 1 \end{bmatrix} + \alpha \begin{bmatrix} 1 & 1 \end{bmatrix} \begin{bmatrix} 1 \\ -1 \end{bmatrix} = 1$$

and all vectors $\mathbf{x}$ that satisfy the constraints have this form.

We obtain this formulation for feasible $\mathbf{x}$ by taking a particular solution

$$\begin{bmatrix} 0 \\ 1 \end{bmatrix}$$

and adding on a linear combination of vectors that span the null space of the matrix

$$\begin{bmatrix} 1 & 1 \end{bmatrix}.$$

The null space defines the set of feasible directions, the directions in which we can step without immediately stepping outside the feasible space.

End example []

What we have accomplished

In general, if our constraints are $\mathbf{Ax} = \mathbf{b}$, to get feasible directions, we express $\mathbf{x}$ as

$$\mathbf{x} = \bar{\mathbf{x}} + \mathbf{Zv}$$

where

- $\bar{\mathbf{x}}$ is a particular solution to the equations $\mathbf{Ax} = \mathbf{b}$ (any one will do),
- the columns of $\mathbf{Z}$ form a basis for the nullspace of $\mathbf{A}$ (any basis will do),
- $\mathbf{v}$ is an arbitrary vector of dimension $(n - m) \times 1$.

Then we have succeeded in reformulating our constrained problem as an unconstrained one:

$$\min_{\mathbf{v}} f(\bar{\mathbf{x}} + \mathbf{Zv})$$

Where does Z come from?

N&S, Section 3.3.4

Suppose we have a QR factorization of the matrix $\mathbf{A}^T$:

$$\mathbf{A}^T = \mathbf{Q}\hat{\mathbf{R}} \equiv \begin{bmatrix} \mathbf{Q}_1 & \mathbf{Q}_2 \end{bmatrix} \begin{bmatrix} \mathbf{R} \\ \mathbf{0} \end{bmatrix} = \mathbf{Q}_1\mathbf{R} + \mathbf{Q}_2\mathbf{0}$$

where

- $\mathbf{Q}_1 \in \mathcal{R}^{n \times m}$,
- $\mathbf{Q}_2 \in \mathcal{R}^{n \times (n-m)}$,
- $\mathbf{R} \in \mathcal{R}^{m \times m}$ is upper triangular,
- $\mathbf{0} \in \mathcal{R}^{(n-m) \times m}$,
- $\mathbf{Q}^T\mathbf{Q} = \mathbf{I}$.

Then

$$\mathbf{Ax} = (\mathbf{R}^T\mathbf{Q}_1^T + \mathbf{0}\mathbf{Q}_2^T)\mathbf{x} = \mathbf{R}^T\mathbf{Q}_1^T\mathbf{x}$$

and the columns of $\mathbf{Q}_2$ form a basis for the nullspace of $\mathbf{A}$.

Therefore, to determine $\mathbf{Z}$, we do a QR factorization of $\mathbf{A}^T$ and set $\mathbf{Z} = \mathbf{Q}_2$.

Algorithms for QR factorization: Gram-Schmidt, Givens, Householder, ...

What are the optimality conditions for our reformulated problem?

$$\min_{\mathbf{v}} f(\bar{\mathbf{x}} + \mathbf{Z}\mathbf{v})$$

Let

$$F(\mathbf{v}) = f(\bar{\mathbf{x}} + \mathbf{Z}\mathbf{v}) .$$

Then

$$\begin{aligned} \nabla_{\mathbf{v}} F(\mathbf{v}) &= \mathbf{Z}^T \nabla_{\mathbf{x}} f(\bar{\mathbf{x}} + \mathbf{Z}\mathbf{v}) = \mathbf{Z}^T \mathbf{g}(\mathbf{x}) \\ \nabla_{\mathbf{v}}^2 F(\mathbf{v}) &= \mathbf{Z}^T \nabla_{\mathbf{x}^2} f(\bar{\mathbf{x}} + \mathbf{Z}\mathbf{v})\mathbf{Z} = \mathbf{Z}^T \mathbf{H}(\mathbf{x})\mathbf{Z} \end{aligned}$$

since $\bar{\mathbf{x}} + \mathbf{Z}\mathbf{v} = \mathbf{x}$.

Our theory for unconstrained optimization now gives us necessary conditions for optimality:

- Reduced gradient is zero: $\mathbf{Z}^T \nabla f(\mathbf{x}) = \mathbf{0}$.
- Reduced Hessian $\mathbf{Z}^T \nabla^2 f(\mathbf{x})\mathbf{Z}$ is positive semidefinite.

We also have sufficient conditions for optimality:

- Reduced gradient is zero: $\mathbf{Z}^T \nabla f(\mathbf{x}) = \mathbf{0}$.
- Reduced Hessian $\mathbf{Z}^T \nabla^2 f(\mathbf{x})\mathbf{Z}$ is positive definite.

An alternate approach

Recall what you know, from advanced calculus, about Lagrange multipliers: to minimize a function subject to equality constraints, we set up the Lagrange function, with one Lagrange multiplier per constraint, and find a point where its partial derivatives are all zero.

Note: We'll sketch the proof of why this works when we consider nonlinear constraints later in this set of notes.

The Lagrange function for our problem

$$\min_{\mathbf{x}} f(\mathbf{x})$$

$$\mathbf{A}\mathbf{x} = \mathbf{b}$$

is

$$L(\mathbf{x}, \boldsymbol{\lambda}) = f(\mathbf{x}) - \boldsymbol{\lambda}^T(\mathbf{A}\mathbf{x} - \mathbf{b}),$$

and setting the partials to zero yields

$$\begin{aligned} \nabla_{\mathbf{x}} L = & \quad \nabla f(\mathbf{x}) - \mathbf{A}^T \boldsymbol{\lambda} & = \mathbf{0}, \\ -\nabla_{\boldsymbol{\lambda}} L = & \quad \mathbf{A}\mathbf{x} - \mathbf{b} & = \mathbf{0}. \end{aligned}$$

These are the first order necessary conditions for optimality.

What does this mean geometrically? The solution is characterized by this:

- It satisfies the constraints.
- The gradient of f at $\mathbf{x}^*$ is a linear combination of the rows of $\mathbf{A}$, which are the gradients of the constraints.

We can also express this in terms of our QR factorization: $\mathbf{A}^T \boldsymbol{\lambda} = \mathbf{g}(\mathbf{x})$, means

$$\mathbf{Q}_1 \mathbf{R} \boldsymbol{\lambda} = \mathbf{g}(\mathbf{x})$$

so $\mathbf{g}(\mathbf{x})$ is in the range of the columns of $\mathbf{Q}_1$ and this is equivalent to

$$\mathbf{Q}_2^T \mathbf{g}(\mathbf{x}) = \mathbf{0}$$

or, in our earlier notation,

$$\mathbf{Z}^T \mathbf{g}(\mathbf{x}) = \mathbf{0}.$$

So we have an alternate formulation of our first order necessary conditions for optimality:

$$\begin{aligned} \mathbf{Z}^T \mathbf{g}(\mathbf{x}) &= \mathbf{0}, \\ \mathbf{A}\mathbf{x} &= \mathbf{b}. \end{aligned}$$

Three digressions

Digression 1: There are cheaper but less stable alternatives to QR.

The QR factorization gives a very nice basis for the nullspace: its columns are mutually orthogonal and therefore computing with them is stable.

There are alternative approaches.

Option 1: Partitioning

Let

$$\mathbf{A} = \left[\begin{array}{cc} \mathbf{B} & \mathbf{N} \end{array} \right]$$

where $\mathbf{B} \in \mathcal{R}^{m \times m}$ and $\mathbf{N} \in \mathcal{R}^{m \times (n-m)}$.

Partition $\mathbf{x}$ similarly, with $\mathbf{x}_1 \in \mathcal{R}^m$ and $\mathbf{x}_2 \in \mathcal{R}^{n-m}$.

Assume that $\mathbf{B}$ is nonsingular. (If not, rearrange the columns of $\mathbf{A}$ until it is.)

Then $\mathbf{A}\mathbf{x} = \mathbf{0}$ if and only if

$$\mathbf{B}\mathbf{x}_1 + \mathbf{N}\mathbf{x}_2 = \mathbf{0}\,,$$

and this means

$$\mathbf{x}_1 + \mathbf{B}^{-1}\mathbf{N}\mathbf{x}_2 = \mathbf{0}\,,$$

so

$$\mathbf{x}_1 = -\mathbf{B}^{-1}\mathbf{N}\mathbf{x}_2$$

and

$$\mathbf{x} = \left[\begin{array}{c} -\mathbf{B}^{-1}\mathbf{N} \\ \mathbf{I} \end{array} \right] \mathbf{v}\,.$$

Therefore, the columns of

$$\left[\begin{array}{c} -\mathbf{B}^{-1}\mathbf{N} \\ \mathbf{I} \end{array} \right]$$

must be a basis for the nullspace of $\mathbf{A}$!

Caution: This basis is sometimes very ill-conditioned, and working with it can lead to unnecessary round-off error.

Option 2: Orthogonal projection

Let

$$\mathbf{x} = \mathbf{p} + \mathbf{q}$$

where $\mathbf{p}$ is in the nullspace of $\mathbf{A}$ and $\mathbf{q}$ is in the range of $\mathbf{A}^T$.

Then

$$\mathbf{A}\mathbf{p} = \mathbf{0}$$

and $\mathbf{q}$ can be expressed as

$$\mathbf{q} = \mathbf{A}^T \boldsymbol{\lambda}$$

for some vector $\boldsymbol{\lambda}$.

Now

$$\mathbf{A}\mathbf{x} = \mathbf{A}(\mathbf{A}^T \boldsymbol{\lambda})$$

so

$$\boldsymbol{\lambda} = (\mathbf{A}\mathbf{A}^T)^{-1}\mathbf{A}\mathbf{x}.$$

Let's look at

$$\begin{aligned}\mathbf{p} &= \mathbf{x} - \mathbf{q} \\ &= \mathbf{x} - \mathbf{A}^T(\mathbf{A}\mathbf{A}^T)^{-1}\mathbf{A}\mathbf{x} \\ &= (\mathbf{I} - \mathbf{A}^T(\mathbf{A}\mathbf{A}^T)^{-1}\mathbf{A})\mathbf{x} \\ &\equiv \mathbf{P}\mathbf{x}\,.\end{aligned}$$

The matrix $\mathbf{P}$ is an orthogonal projection that takes $\mathbf{x}$ into the null space of $\mathbf{A}$.

Thus we have reduced our problem to an unconstrained one, where $\mathbf{x} = \mathbf{x}_b + (\mathbf{I} - \mathbf{A}^T(\mathbf{A}\mathbf{A}^T)^{-1}\mathbf{A})\mathbf{y}$ where $\mathbf{x}_b$ is a particular solution to $\mathbf{A}\mathbf{x} = \mathbf{b}$ and $\mathbf{y}$ is any n-vector.

Unquiz: Prove that

1. $\mathbf{P}^2 = \mathbf{P}$.
2. $\mathbf{P}^T = \mathbf{P}$.

but note that in general $\mathbf{P}^T\mathbf{P} \neq \mathbf{I}$, so $\mathbf{P}$ itself is not an orthogonal matrix. []

The projector $\mathbf{P}$ is usually applied using a Cholesky factorization.

Digression 2: the meaning of the Lagrange multipliers

Our optimality conditions:

$$\begin{aligned}\mathbf{g}(\mathbf{x}^*) - \mathbf{A}^T\boldsymbol{\lambda}^* &= \mathbf{0} \\ \mathbf{A}\mathbf{x}^* - \mathbf{b} &= \mathbf{0}\end{aligned}$$

Sensitivity analysis: Suppose we have a point $\hat{\mathbf{x}}$ satisfying

$$\|\mathbf{x}^* - \hat{\mathbf{x}}\| \leq \epsilon$$

and

$$\mathbf{A}\hat{\mathbf{x}} = \mathbf{b} + \boldsymbol{\delta}$$

where ϵ and $\|\boldsymbol{\delta}\|$ are small.

Then Taylor series expansion tells us

$$\begin{aligned}f(\hat{\mathbf{x}}) &= f(\mathbf{x}^*) + (\hat{\mathbf{x}} - \mathbf{x}^*)^T\mathbf{g}(\mathbf{x}^*) + O(\epsilon^2) \\ &= f(\mathbf{x}^*) + (\hat{\mathbf{x}} - \mathbf{x}^*)^T\mathbf{A}^T\boldsymbol{\lambda}^* + O(\epsilon^2) \\ &= f(\mathbf{x}^*) + \boldsymbol{\delta}^T\boldsymbol{\lambda}^* + O(\epsilon^2)\,.\end{aligned}$$

What this tells us: If we wiggle b_i by δ_i, then we wiggle f by $\delta_i \lambda_i^*$.

Therefore, λ_i^* is the change in f per unit change in b_i. It tells us the sensitivity of f to b_i.

Jargon: λ_i is called a dual variable or a shadow price.

Digression 3

It is important to realize that we do not minimize the Lagrangian function

$$L(\mathbf{x}, \boldsymbol{\lambda}) = f(\mathbf{x}) - \boldsymbol{\lambda}^T(\mathbf{A}\mathbf{x} - \mathbf{b}) .$$

We find a saddlepoint of this function.

So far...

- We have optimality conditions for unconstrained problems.
- We have optimality conditions for linear equality constraints.

Case 2: Optimality conditions for linear inequality constraints

---

A big "if"

IF we knew

$$\mathcal{W} = \{i \in \mathcal{I} : c_i(\mathbf{x}^*) = 0\} ,$$

where $\mathbf{c}(\mathbf{x}^*) = \mathbf{A}\mathbf{x}^* - \mathbf{b}$, then we could set up the Lagrange multiplier problem and have optimality conditions for our problem.

Let $\bar{W}$ denote the subscripts not in $\mathcal{W}$.

But we don't know the set $\mathcal{W}$ of constraints that are active at the solution.

Let's guess!

Suppose we take a guess at the active set. This gives us a set of equations to solve:

$$\begin{aligned} \mathbf{g}(\mathbf{x}) - \mathbf{A}_w^T \boldsymbol{\lambda}_w &= \mathbf{0}, \\ \mathbf{A}_w \mathbf{x} &= \mathbf{b}_w . \end{aligned}$$

Assume that $\mathbf{A}_w$ has full row rank. This implies that $\mathcal{W}$ has at most n elements.

Suppose this system has a solution $\hat{\mathbf{x}}$, $\hat{\boldsymbol{\lambda}}$. Also suppose that $\mathbf{A}_{\bar{w}}\hat{\mathbf{x}} > \mathbf{b}_{\bar{w}}$, so that $\hat{\mathbf{x}}$ is feasible. Do we have a solution to our minimization problem?

Suppose we find that $\hat{\lambda}_j < 0$.

Let $\mathbf{p}$ solve $\mathbf{A}_w\mathbf{p} = \mathbf{e}_j$.

(This has a solution since $\mathbf{A}_w$ is full rank.)

Then

$$\mathbf{A}_w(\hat{\mathbf{x}} + \alpha\mathbf{p}) = \mathbf{b}_w + \alpha\mathbf{e}_j \geq \mathbf{b}_w,$$

so $\hat{\mathbf{x}} + \alpha\mathbf{p}$ satisfies the $\mathcal{W}$ inequality constraints as long as $\alpha > 0$, and it satisfies the other inequalities as long as α is small enough. Thus, $\mathbf{p}$ is a feasible direction.

Also, by Digression 2, we know that

$$f(\hat{\mathbf{x}} + \alpha\mathbf{p}) \approx f(\hat{\mathbf{x}}) + \alpha\mathbf{e}_j^T\hat{\boldsymbol{\lambda}} = f(\hat{\mathbf{x}}) + \alpha\hat{\boldsymbol{\lambda}}_j < f(\hat{\mathbf{x}})$$

(for small enough α) so we have found a better point!

We'll come back to the algorithmic use of this idea later. For now, we seek insight on recognizing an optimal point.

We have just shown that if $\mathbf{x}$ is a minimizer, then the multipliers $\boldsymbol{\lambda}_w$ that satisfy $\mathbf{A}_w^T\boldsymbol{\lambda}_w = \mathbf{g}(\mathbf{x})$ must be nonnegative.

(The multipliers for the $\bar{w}$ indices must be zero, since these constraints do not appear in the Lagrangian.)

A fancy way of writing this

Current formulation of (first order) necessary conditions for optimality:

$$\begin{aligned}
\mathbf{A}_w^T\boldsymbol{\lambda}_w &= \mathbf{g}(\mathbf{x}) \\
\boldsymbol{\lambda}_w \geq \mathbf{0} \quad &, \quad \boldsymbol{\lambda}_{\bar{w}} = \mathbf{0} \\
\mathbf{A}_w\mathbf{x} &= \mathbf{b}_w \\
\mathbf{A}_{\bar{w}}\mathbf{x} &> \mathbf{b}_{\bar{w}}
\end{aligned}$$

where $\bar{w}$ denotes the subscripts not in $\mathcal{W}$.

Equivalently,

$$\begin{aligned}
\mathbf{A}^T\boldsymbol{\lambda} &= \mathbf{g}(\mathbf{x}) \\
\boldsymbol{\lambda} &\geq \mathbf{0} \\
\mathbf{A}\mathbf{x} &\geq \mathbf{b} \\
\boldsymbol{\lambda}^T(\mathbf{A}\mathbf{x} - \mathbf{b}) &= 0
\end{aligned}$$

This last condition is called complementarity.

The second order necessary condition: (from the reduced variable derivation above) The reduced variable Hessian matrix

$$\mathbf{Z}_w^T \mathbf{H}(\mathbf{x}) \mathbf{Z}_w$$

must be positive semidefinite.

Sufficient conditions for optimality: All of this, plus $\mathbf{Z}_w^T \mathbf{H}(\mathbf{x}) \mathbf{Z}_w$ positive definite, where W indicates active constraints for which the multipliers are positive).

Case 3: Optimality conditions for general constraints

$$\min_{\mathbf{x}} f(\mathbf{x})$$

$$\mathbf{c}(\mathbf{x}) \geq \mathbf{0}$$

A constraint qualification

Let the $m \times n$ matrix $\mathbf{A}(\mathbf{x})$ be defined by

$$a_{ij}(\mathbf{x}) = \frac{\partial c_i(\mathbf{x})}{\partial x_j}.$$

Assume that $\mathbf{A}(\mathbf{x})$ has linearly independent rows.

Again, this is a constraint qualification, saying that the gradients of the active constraints are linearly independent.

picture.

Optimality conditions

$$L(\mathbf{x}, \boldsymbol{\lambda}) = f(\mathbf{x}) - \boldsymbol{\lambda}^T \mathbf{c}(\mathbf{x})$$

Theorem: Necessary conditions for a feasible point $\mathbf{x}$ to be a minimizer:

- $\mathbf{g}(\mathbf{x}) - \mathbf{A}^T(\mathbf{x})\boldsymbol{\lambda} = \mathbf{0}$
- $\lambda_j \geq 0$ if j is an inequality constraint.
- λ_j unrestricted in sign for equality constraints.
- $\boldsymbol{\lambda}^T \mathbf{c}(\mathbf{x}) = 0$ (complementarity)

- $\mathbf{Z}^T \bigtriangledown_{xx} L(\mathbf{x}, \boldsymbol{\lambda})\mathbf{Z}$ is positive semidefinite, where the columns of $\mathbf{Z}$ are a basis for the null space of $\mathbf{A}_w$, the gradients of the active constraints for which the multipliers are positive.

Theorem: Sufficient conditions: Add positive definiteness of $\mathbf{Z}^T \bigtriangledown_{xx} L(\mathbf{x}, \boldsymbol{\lambda})\mathbf{Z}$.

We won't prove these theorems, but we will sketch the proof of a piece of a special case: that for equality constraints, if $\mathbf{x}^*$ is a local minimizer of f, then there is a vector of multipliers satisfying

$$\mathbf{A}^T(\mathbf{x}^*)\boldsymbol{\lambda} = \mathbf{g}(\mathbf{x}^*)\,.$$

Goal:

To prove: If all constraints are equalities, then

$$\mathbf{A}^T(\mathbf{x}^*)\boldsymbol{\lambda} = \mathbf{g}(\mathbf{x}^*)\,.$$

Note: We are proving the correctness of the Lagrange multiplier formulation for solving equality constrained problems as promised earlier in this set of notes.

Proof ingredient 1: a pitfall

With nonlinear constraints, there may be no feasible directions!

picture

So we need to work with feasible curves $\mathbf{x}(t)$, $0 \le t \le t_1$, with $\mathbf{x}(0)$ being our current point. A curve is feasible if it stays tangent to our (active) constraints.

Example 1: The curve

$$\mathbf{x}(t) = \left[\begin{array}{c} \cos t \\ \sin t \end{array} \right]$$

stays tangent to the unit circle $x_1^2 + x_2^2 = 1$.

This is true since

$$\mathbf{x}(0) = \left[\begin{array}{c} 1 \\ 0 \end{array} \right]$$

and

$$\mathbf{x}'(t) = \left[\begin{array}{c} -\sin t \\ \cos t \end{array} \right]$$

which is tangent to the circle. []

Example 2: The curve

$$\mathbf{x}(t) = \begin{bmatrix} t \\ 2t \end{bmatrix} + \begin{bmatrix} 0 \\ 4 \end{bmatrix}$$

stays tangent to the line

$$x_2 - 2x_1 = 4\,.$$

[]

Proof ingredient 2: Some unstated machinery that N&S use:

- For $\mathbf{x}(t)$ to be a feasible curve, it must be defined for $t \in [t_0, t_1]$, where $t_0 < 0 < t_1$.
- Every feasible point in a neighborhood of the current point is on some feasible curve.

Proof ingredient 3: the tangent cone

Define the tangent cone

$$T(\mathbf{x}^*) = \{\mathbf{p} : \mathbf{p} = \mathbf{x}'(0) \text{ for some feasible curve at } \mathbf{x}^*\}\,.$$

This is a cone because

- $\mathbf{0} \in T$ (because we could define the curve $\mathbf{x}(t) = \mathbf{x}^*$ for all t).
- If $\mathbf{p} \in T$, then $\alpha\mathbf{p} \in T$ for positive scalars α.

picture

Now the constraints are equalities, so

$$c_i(\mathbf{x}(t)) = 0, \;\; t \in [t_0, t_1]\,,$$

so

$$\frac{d\, c_i(\mathbf{x}(t))}{dt} = \mathbf{x}'(t)^T \bigtriangledown c_i(\mathbf{x}(t)) = 0. \;\; t \in [t_0, t_1]\,.$$

Therefore, at $t = 0$, for all feasible curves,

$$\mathbf{x}'(0)^T \bigtriangledown c_i(\mathbf{x}^*) = 0\,.$$

Thus, for all $\mathbf{p}$ in the tangent cone T of $\mathbf{x}^*$,

$$\mathbf{p}^T \bigtriangledown c_i(\mathbf{x}^*) = 0, \;\; i = 1, \ldots, m,$$

so

$$\mathbf{A}(\mathbf{x}^*)\mathbf{p} = \mathbf{0}\,.$$

Therefore, if $\mathbf{p}$ is in the tangent cone, then $\mathbf{p}$ is in the null space of the matrix of constraint gradients!

If the rows of $\mathbf{A}$ are linearly independent, then we can reverse the argument and show that if $\mathbf{p}$ is in the null space of $\mathbf{A}$, then $\mathbf{p}$ is in the tangent cone.

Therefore, when $\mathbf{A}(\mathbf{x}^*)$ is full rank, the tangent cone $T(\mathbf{x}^*)$ equals the nullspace of $\mathbf{A}(\mathbf{x}^*)$.

Finally, the sketch of proof for equality constraints

Suppose $\mathbf{x}^*$ is a local minimizer of $f(\mathbf{x})$ over $\{\mathbf{x} : \mathbf{c}(\mathbf{x}){=}\mathbf{0}\}$.

Then, for all feasible curves $\mathbf{x}(t)$ with $\mathbf{x}(0) = \mathbf{x}^*$, it must be true that

$$f(\mathbf{x}(t)) {\geq} f(\mathbf{x}^*)$$

for $t > 0$ sufficiently small.

The chain rule tells us

$$\frac{d}{dt} f(\mathbf{x}(t)) = \mathbf{x}'(t)^T \bigtriangledown_{\mathbf{x}} f(\mathbf{x}(t))\,,$$

and optimality implies that

$$\frac{d}{dt} f(\mathbf{x}(t))|_{t=0} = \mathbf{x}'(0)^T \bigtriangledown_{\mathbf{x}} f(\mathbf{x}^*) = 0\,.$$

Therefore $\mathbf{p}^T\mathbf{g}(\mathbf{x}^*) = 0$ for all $\mathbf{p}$ in the nullspace of $\mathbf{A}(\mathbf{x}^*)$.

Therefore, a necessary condition for optimality is that the reduced gradient is zero:

$$\mathbf{Z}(\mathbf{x}^*)^T\mathbf{g}(\mathbf{x}^*) = \mathbf{0}\,.$$

Equivalently, there must be a vector $\boldsymbol{\lambda}$ so that

$$\mathbf{A}(\mathbf{x}^*)^T\boldsymbol{\lambda} = \mathbf{g}(\mathbf{x}^*)$$

so that $\mathbf{g}(\mathbf{x}^*)$ is in the span of the constraint gradients.

[]

picture

Notes on the proof for inequality constraints

- To prove the sign conditions on $\boldsymbol{\lambda}$, the argument is the same as for linear constraints.
- To prove the second derivative conditions, see N&S p. 461.

Duality

Duality

Idea: Problems come in pairs, linked through the Lagrangian.

We need two theorems about this linkage, or duality:

- weak duality
- strong duality

and then two theorems about dual problems:

- weak dual
- convex duality

and finally an alternate dual problem, the Wolfe dual, that depends on differentiability.

Weak duality

Theorem: (Weak Duality) (N&S p466)
Let $F(\mathbf{x}, \boldsymbol{\lambda})$ be a function from $\mathcal{R}^{n+m} \rightarrow R^1$ with $\mathbf{x} \in \mathcal{R}^n$ and $\boldsymbol{\lambda} \in \mathcal{R}^m$. Then

$$\max_{\boldsymbol{\lambda}} \min_{\mathbf{x}} F(\mathbf{x}, \boldsymbol{\lambda}) \leq \min_{\mathbf{x}} \max_{\boldsymbol{\lambda}} F(\mathbf{x}, \boldsymbol{\lambda}) .$$

Notes:

- Really, the max should be sup and the min should be inf, so substitute this terminology if you are comfortable with it.
- The function F does not need to be defined everywhere; we could restate the theorem with $\mathbf{x}$ and $\boldsymbol{\lambda}$ restricted to smaller domains.

Proof: Given any $\hat{\mathbf{x}}$ and $\hat{\boldsymbol{\lambda}}$,

$$\min_{\mathbf{x}} F(\mathbf{x}, \hat{\boldsymbol{\lambda}}) \leq F(\hat{\mathbf{x}}, \hat{\boldsymbol{\lambda}}) \leq \max_{\boldsymbol{\lambda}} F(\hat{\mathbf{x}}, \boldsymbol{\lambda}) .$$

Now let's make a specific choice:

- Let $\hat{\boldsymbol{\lambda}}$ be the $\boldsymbol{\lambda}$ that maximizes the left-hand side.
- Let $\hat{\mathbf{x}}$ be the $\mathbf{x}$ that minimizes the right-hand side.

Then

$$\max_{\boldsymbol{\lambda}} \min_{\mathbf{x}} F \leq \min_{\mathbf{x}} \max_{\boldsymbol{\lambda}} F\,.$$

[]

Strong duality

Theorem: (Strong Duality) (N&S p.468)
Let $F(\mathbf{x}, \boldsymbol{\lambda})$ be a function from $\mathcal{R}^{n+m} \to \mathcal{R}^1$. Then the condition

$$\max_{\boldsymbol{\lambda}} \min_{\mathbf{x}} F(\mathbf{x}, \boldsymbol{\lambda}) = \min_{\mathbf{x}} \max_{\boldsymbol{\lambda}} F(\mathbf{x}, \boldsymbol{\lambda})$$

holds if and only if there exists a point $(\mathbf{x}^*, \boldsymbol{\lambda}^*)$ such that

$$F(\mathbf{x}^*, \boldsymbol{\lambda}) \leq F(\mathbf{x}^*, \boldsymbol{\lambda}^*) \leq F(\mathbf{x}, \boldsymbol{\lambda}^*)$$

for all points $\mathbf{x}$ and $\boldsymbol{\lambda}$ in the domain of F.

In words: We can reverse the order of the max and the min if and only if there exists a saddle point for F.

Proof: $(\leftarrow)$ Suppose $(\mathbf{x}^*, \boldsymbol{\lambda}^*)$ is a saddle point. Then

$$\begin{aligned}
\min_{\mathbf{x}} \max_{\boldsymbol{\lambda}} F(\mathbf{x}, \boldsymbol{\lambda}) &\leq \max_{\boldsymbol{\lambda}} F(\mathbf{x}^*, \boldsymbol{\lambda}) \\
&\leq F(\mathbf{x}^*, \boldsymbol{\lambda}^*) \\
&\leq \min_{\mathbf{x}} F(\mathbf{x}, \boldsymbol{\lambda}^*) \\
&\leq \max_{\boldsymbol{\lambda}} \min_{\mathbf{x}} F(\mathbf{x}, \boldsymbol{\lambda})
\end{aligned}$$

Now, considering the result of the weak duality theorem, we can conclude that the first term must equal the last.

$(\rightarrow)$ Suppose

$$\max_{\boldsymbol{\lambda}} \min_{\mathbf{x}} F(\mathbf{x}, \boldsymbol{\lambda}) = \min_{\mathbf{x}} \max_{\boldsymbol{\lambda}} F(x, \boldsymbol{\lambda})$$

and that this is equal to the value $F(\mathbf{x}^*, \boldsymbol{\lambda}^*)$. Then, for any $\hat{\mathbf{x}}$ and $\hat{\boldsymbol{\lambda}}$,

$$\begin{aligned}
F(\mathbf{x}^*, \hat{\boldsymbol{\lambda}}) &\leq \max_{\boldsymbol{\lambda}} F(\mathbf{x}^*, \boldsymbol{\lambda}) \\
&= \max_{\boldsymbol{\lambda}} \min_{\mathbf{x}} F(\mathbf{x}, \boldsymbol{\lambda}) \\
&= F(\mathbf{x}^*, \boldsymbol{\lambda}^*) \\
&= \min_{\mathbf{x}} \max_{\boldsymbol{\lambda}} F(\mathbf{x}, \boldsymbol{\lambda}) \\
&= \min_{\mathbf{x}} F(\mathbf{x}, \boldsymbol{\lambda}^*) \\
&\leq F(\hat{\mathbf{x}}, \boldsymbol{\lambda}^*)
\end{aligned}$$

[]

So what?

Consider our original problem:

$$\min_{\mathbf{x}} f(\mathbf{x})$$

$$\mathbf{c}(\mathbf{x}) \geq \mathbf{0}$$

The Lagrangian for this problem is

$$L(\mathbf{x}, \boldsymbol{\lambda}) = f(\mathbf{x}) - \boldsymbol{\lambda}^T \mathbf{c}(\mathbf{x}) .$$

A new problem to play with: Lagrange duality

Define

$$L^*(\mathbf{x}) = \max_{\boldsymbol{\lambda} \geq \mathbf{0}} L(\mathbf{x}, \boldsymbol{\lambda}) = \max_{\boldsymbol{\lambda} \geq \mathbf{0}} f(\mathbf{x}) - \boldsymbol{\lambda}^T \mathbf{c}(\mathbf{x}) .$$

Case 1: If $\mathbf{x}$ is feasible, then $\mathbf{c}(\mathbf{x}) \geq \mathbf{0}$, so the max occurs when $\boldsymbol{\lambda} = \mathbf{0}$.

Case 2: If $\mathbf{x}$ is not feasible, then some $c_i(\mathbf{x})$ is negative, so the max is infinite.

Therefore,

$$L^*(\mathbf{x}) = \begin{cases} f(\mathbf{x}) & \text{if } \mathbf{c}(\mathbf{x}) \geq \mathbf{0} , \\ \infty & \text{otherwise} . \end{cases}$$

Therefore, the solution to the original problem is the same as the solution to the primal problem

$$\min_{\mathbf{x}} L^*(\mathbf{x}) = \min_{\mathbf{x}} \max_{\boldsymbol{\lambda} \geq \mathbf{0}} L(\mathbf{x}, \boldsymbol{\lambda}) .$$

A dual problem

Suppose $\boldsymbol{\lambda} \geq \mathbf{0}$. Define

$$L_*(\boldsymbol{\lambda}) = \min_{\mathbf{x}} L(\mathbf{x}, \boldsymbol{\lambda}) = \min_{\mathbf{x}} f(\mathbf{x}) - \boldsymbol{\lambda}^T \mathbf{c}(\mathbf{x}) ,$$

and consider the problem

$$\max_{\lambda \geq 0} L_*(\boldsymbol{\lambda}) = \max_{\lambda \geq 0} \min_{\mathbf{x}} L(\mathbf{x}, \boldsymbol{\lambda}).$$

Weak Lagrange duality

Theorem: (Weak Lagrange duality) (N&S p. 471)
Let $\tilde{\mathbf{x}}$ be primal feasible, so that $\mathbf{c}(\tilde{\mathbf{x}}) \geq \mathbf{0}$.
Let $\bar{\mathbf{x}}, \bar{\boldsymbol{\lambda}}$ be dual feasible, so that $\bar{\boldsymbol{\lambda}} \geq \mathbf{0}$, and $\bar{\mathbf{x}}$ minimizes $L(\mathbf{x}, \bar{\boldsymbol{\lambda}})$.

Then

$$f(\bar{\mathbf{x}}) - \bar{\boldsymbol{\lambda}}^T \mathbf{c}(\bar{\mathbf{x}}) \leq f(\tilde{\mathbf{x}}) .$$

Note:

- For dual feasibility, it is not necessary that $\mathbf{c}(\mathbf{x}) \geq \mathbf{0}$.
- Sometimes we require that our solution, in addition to satisfying $\mathbf{c}(\mathbf{x}) \geq \mathbf{0}$, satisfies $\mathbf{x} \in S \subset \mathcal{R}^n$. If the problem is formulated this way, then a dual feasible point must have $\mathbf{x} \in S$, but it is not necessary that $\mathbf{c}(\mathbf{x}) \geq \mathbf{0}$.

Proof: Let's recall what we know. The Lagrangian is

$$L(\mathbf{x}, \boldsymbol{\lambda}) = f(\mathbf{x}) - \boldsymbol{\lambda}^T \mathbf{c}(\mathbf{x}) .$$

The Weak Duality Theorem, and the fact that $\tilde{\mathbf{x}}$ is feasible, tells us

$$\begin{aligned} f(\bar{\mathbf{x}}) - \bar{\boldsymbol{\lambda}}^T \mathbf{c}(\bar{\mathbf{x}}) &= L(\bar{\mathbf{x}}, \bar{\boldsymbol{\lambda}}) \\ &\leq \max_{\boldsymbol{\lambda} \geq \mathbf{0}} \min_{\mathbf{x}} L(\mathbf{x}, \boldsymbol{\lambda}) \\ &\leq \min_{x} \max_{\boldsymbol{\lambda} \geq \mathbf{0}} L(\mathbf{x}, \boldsymbol{\lambda}) \\ &\leq \max_{\boldsymbol{\lambda} \geq \mathbf{0}} L(\tilde{\mathbf{x}}, \boldsymbol{\lambda}) \\ &= f(\tilde{\mathbf{x}}) \end{aligned}$$

[]

Corollary: If the primal is unbounded, then the dual is infeasible.
If the dual is unbounded, then the primal is infeasible.

Example: Consider the primal problem

$$\min_{x} -x$$

(with $x \in \mathcal{R}^1$) subject to $x \geq 0$. The Lagrangian is

$$L(x, \lambda) = -x - \lambda x .$$

Then $\bar{x}, \lambda$ is dual feasible if $\bar{x}$ satisfies

$$\min_{x} -(\lambda + 1)x$$

where λ is a fixed nonnegative number. There are no dual feasible points, and the primal has no minimum. []

An important example: Linear programming duality

Example: Duality for linear programming

Consider the linear programming problem

$$\min_{\mathbf{x}} \mathbf{c}^T \mathbf{x}$$

$$\mathbf{A}\mathbf{x} - \mathbf{b} \geq \mathbf{0}$$

The Lagrangian is

$$L(\mathbf{x}, \boldsymbol{\lambda}) = \mathbf{c}^T\mathbf{x} - \boldsymbol{\lambda}^T(\mathbf{A}\mathbf{x} - \mathbf{b})\,.$$

The primal problem is

$$\min_{\mathbf{x}} \max_{\boldsymbol{\lambda} \geq \mathbf{0}} \mathbf{c}^T\mathbf{x} - \boldsymbol{\lambda}^T(\mathbf{A}\mathbf{x} - \mathbf{b})$$

which is equivalent to our original problem.

The dual problem is

$$\max_{\boldsymbol{\lambda} \geq \mathbf{0}} \min_{\mathbf{x}} \mathbf{c}^T\mathbf{x} - \boldsymbol{\lambda}^T(\mathbf{A}\mathbf{x} - \mathbf{b})\,.$$

Fix $\boldsymbol{\lambda} \geq \mathbf{0}$. Then we need to minimize

$$(\mathbf{c} - \mathbf{A}^T\boldsymbol{\lambda})^T\mathbf{x} + \boldsymbol{\lambda}^T\mathbf{b}$$

and this value is $L_*(\boldsymbol{\lambda})$.

But

$$L_*(\boldsymbol{\lambda}) = \begin{cases} -\infty & \text{if } \mathbf{c} - \mathbf{A}^T\boldsymbol{\lambda} \neq \mathbf{0}\,, \\ \boldsymbol{\lambda}^T\mathbf{b} & \text{if } \mathbf{c} - \mathbf{A}^T\boldsymbol{\lambda} = \mathbf{0}\,. \end{cases}$$

Therefore, if $\boldsymbol{\lambda}^* \geq \mathbf{0}$ and $\mathbf{c} - \mathbf{A}^T\boldsymbol{\lambda}^* = \mathbf{0}$, then the dual problem solution value is $\boldsymbol{\lambda}^{*T}\mathbf{b}$.

Thus, the dual problem is equivalent to

$$\max_{\boldsymbol{\lambda} \geq \mathbf{0}} \boldsymbol{\lambda}^T\mathbf{b}$$

$$\mathbf{A}^T\boldsymbol{\lambda} - \mathbf{c} = \mathbf{0}$$

Check strong duality for LP:

Suppose $\mathbf{x}^*$ solves the primal and $\boldsymbol{\lambda}^*$ solves the dual.

Then

$$\mathbf{c}^T\mathbf{x}^* = \boldsymbol{\lambda}^{*T}\mathbf{b}$$

so we can solve either one and know the solution to the other!

For example, if we know $\boldsymbol{\lambda}^*$, then the components that are positive determine the active set of constraints and enable us to determine $\mathbf{x}^*$.

Remember that the dual variables also give us sensitivity information, so they are important to know.

Caution: Usually the variables $\mathbf{x}$ and $\boldsymbol{\lambda}$ cannot be uncoupled in the dual. Linear programming is an exception to this.

End of linear programming example. []

Final words

Final words

- We have derived optimality conditions so that we can recognize a solution when we find one.
- We have derived a partner to our original (primal) problem, called the dual problem.
- We have hinted at some algorithmic approaches:
  - Idea 1: Eliminate constraints by reducing the number of variables.
  - Idea 2: Walk in feasible descent directions.
  - Idea 3: Eliminate constraints through Lagrangians.

Next we will discuss these algorithmic approaches.