Contents

A Linear algebra basics **1**
A.1 Basic operations 1
A.2 Fundamental theorem of linear algebra 4
A.2.1 Part one 4
A.2.2 Part two 5
A.3 Nature of the solution 7
A.3.1 Exactly-identified 7
A.3.2 Under-identified 8
A.3.3 Over-identified 10
A.4 Matrix decomposition and inverse operations 12
A.4.1 LU factorization 12
A.4.2 Cholesky decomposition 16
A.4.3 Eigenvalues and eigenvectors 17
A.4.4 Singular value decomposition 23
A.4.5 Spectral decomposition 28
A.4.6 quadratic forms, eigenvalues, and positive definiteness 29
A.4.7 similar matrices, Jordan form, and generalized eigenvectors 29
A.5 Gram-Schmidt construction of an orthogonal matrix 32
A.5.1 QR decomposition 34
A.5.2 Gram-Schmidt QR algorithm 34
A.5.3 Accounting example 35
A.5.4 The Householder QR algorithm 36
A.5.5 Accounting example 36

A.6 Computing eigenvalues 39
A.6.1 Schur's lemma 39
A.6.2 Power algorithm 40
A.6.3 QR algorithm 40
A.6.4 Schur decomposition 42
A.7 Some determinant identities 46
A.7.1 Determinant of a square matrix 46
A.7.2 Identities 47
A.8 Matrix exponentials and logarithms 49

B Iterated expectations 51
B.1 Decomposition of variance 53
B.2 Jensen's inequality 54

C Multivariate normal theory 55
C.1 Conditional distribution 57
C.2 Special case of precision 61
C.3 Truncated normal distribution 63

D Projections and conditional expectations 71
D.1 Gauss-Markov theorem 71
D.2 Generalized least squares (GLS) 74
D.3 Recursive least squares 76

E Two stage least squares IV (2SLS-IV) 79
E.1 General case 79
E.2 Special case 81

F Seemingly unrelated regression (SUR) 83
F.1 Classical 84
F.2 Bayesian 84
F.3 Bayesian treatment effect application 85

G Maximum likelihood estimation of discrete choice models 87

H Optimization 89
H.1 Linear programming 89
H.1.1 basic solutions or extreme points 90
H.1.2 fundamental theorem of linear programming 91
H.1.3 duality theorems 91
H.1.4 example 92
H.1.5 complementary slackness 92
H.2 Nonlinear programming 93
H.2.1 unconstrained 93
H.2.2 convexity and global minima 93
H.2.3 example 94

H.2.4 constrained — the Lagrangian 94
H.2.5 Karash-Kuhn-Tucker conditions 95
H.2.6 example 96
H.3 Theorem of the separating hyperplane 97

I Quantum information 99
I.1 Quantum information axioms 99
I.1.1 The superposition axiom 99
I.1.2 The transformation axiom 100
I.1.3 The measurement axiom 100
I.1.4 The combination axiom 101
I.2 Summary of quantum "rules" 103
I.3 Observables and expected payoffs 104
I.4 Density operators and quantum entropy 105

J Common distributions 109

Appendix H
Optimization

In this appendix, we briefly review optimization. First, we'll take up linear programming then we'll review nonlinear programming.[1]

H.1 Linear programming

A linear program (LP) is any optimization frame which can be described by a linear objective function and linear constraints. Linear refers to choice variables, say x, of no greater than first degree (affine transformations which allow for parallel lines are included). Prototypical examples are

$$\begin{array}{cc} \max\limits_{x \geq 0} & \pi^T x \\ s.t. & Ax \leq r \end{array}$$

or

$$\begin{array}{cc} \min\limits_{x \geq 0} & \pi^T x \\ s.t. & Ax \geq r \end{array}$$

[1] For additional details, consult, for instance, Luenberger and Ye, 2010, *Linear and Nonlinear Programming*, Springer, or Luenberger, 1997 *Optimization by Vector Space Methods*, Wiley.

H.1.1 basic solutions or extreme points

Basic solutions are typically determined from the standard form for an *LP*. Standard form involves equality constraints except non-negative choice variables, $x \geq 0$. That is, $Ax \leq r$ is rewritten in terms of slack variables, s, such that $Ax + s = r$. The solution to this program is the same as the solution to the inequality program.

A basic solution or extreme point is determined from an $m \times m$ submatrix of A composed of m linearly independent columns of A. The set of basic feasible solutions then is the collection of all basic solutions involving $x \geq 0$.

Consider an example. Suppose $A = \begin{bmatrix} 2 & 1 \\ 1 & 2 \end{bmatrix}$, $r = \begin{bmatrix} r_1 \\ r_2 \end{bmatrix}$, $x = \begin{bmatrix} x_1 \\ x_2 \end{bmatrix}$ and $s = \begin{bmatrix} s_1 \\ s_2 \end{bmatrix}$. Then, $Ax + s = r$ can be written $Bx_s = r$ where $B = \begin{bmatrix} A & I_2 \end{bmatrix}$, I_2 is a 2×2 identity matrix, and $x_s = \begin{bmatrix} x \\ s \end{bmatrix}$. The matrix B has two linearly independent columns so each basic solution works with two columns of B, say B_{ij}, and the elements other than i and j of x_s are set to zero. For instance, B_{12} leads to basic solution $x_1 = \frac{2r_1 - r_2}{3}$ and $x_2 = \frac{2r_2 - r_1}{3}$. The basic solutions are tabulated below.

B_{ij}	x_1	x_2	s_1	s_2
B_{12}	$\frac{2r_1 - r_2}{3}$	$\frac{2r_2 - r_1}{3}$	0	0
B_{13}	r_2	0	$r_1 - 2r_2$	0
B_{14}	$\frac{r_1}{2}$	0	0	$\frac{2r_2 - r_1}{2}$
B_{23}	0	$\frac{r_2}{2}$	$\frac{2r_1 - r_2}{2}$	0
B_{24}	0	r_1	0	$r_2 - 2r_1$

To test feasibility consider specific values for r. Suppose $r_1 = r_2 = 10$. The table with a feasibility indicator $(1(x_s \geq 0))$ becomes

B_{ij}	x_1	x_2	s_1	s_2	feasible
B_{12}	$\frac{10}{3}$	$\frac{10}{3}$	0	0	yes
B_{13}	10	0	-10	0	no
B_{14}	5	0	0	5	yes
B_{23}	0	5	5	0	yes
B_{24}	0	10	0	-10	no

Notice, when $x_2 = 0$ there is slack in the second constraint ($s_2 > 0$) and similarly when $x_1 = 0$ there is slack in the first constraint ($s_1 > 0$). Basic feasible solutions, an algebraic concept, are also referred to by their geometric counterpart, extreme points.

Identification of basic feasible solutions or extreme points combined with the fundamental theorem of linear programming substantially reduce the search for an optimal solution.

H.1.2 fundamental theorem of linear programming

For a linear program in standard form where A is an $m \times n$ matrix of rank m,

i) if there is a feasible solution, there is a basic feasible solution;

ii) if there is an optimal solution, there is a basic feasible optimal solution.

Further, if more than one basic feasible solution is optimal, the edge between the basic feasible optimal solutions is also optimal. The theorem means the search for the optimal solution can be restricted to basic feasible solutions — a finite number of points.

H.1.3 duality theorems

Optimality programs come in pairs. That is, there is a complementary or dual program to the primary (primal) program. For instance, the dual to the maximization program is a minimization program, and vice versa.

$$\begin{array}{cccc} \multicolumn{2}{c}{\text{primal program}} & \multicolumn{2}{c}{\text{dual program}} \\ \max\limits_{x\geq 0} & \pi^T x & \min\limits_{\lambda\geq 0} & r^T\lambda \\ s.t. & Ax \leq r & s.t. & A^T\lambda \geq \pi \end{array}$$

or

$$\begin{array}{cccc} \multicolumn{2}{c}{\text{primal program}} & \multicolumn{2}{c}{\text{dual program}} \\ \min\limits_{x\geq 0} & \pi^T x & \max\limits_{\lambda\geq 0} & r^T\lambda \\ s.t. & Ax \geq r & s.t. & A^T\lambda \leq \pi \end{array}$$

where λ is a vector of shadow prices or dual variable values. The dual of the dual program is the primal program.

strong duality theorem

If either the primal or dual has an optimal solution so does the other and their optimal objective function values are equal. If one of the programs is unbounded the other has no feasible solution.

weak duality theorem

For feasible solutions, the objective function value of the minimization program (say, dual) is greater than or equal to the maximization program (say, primal).

The intuition for the duality theorems is straightforward. Begin with the constraints

$$Ax \leq r \quad A^T\lambda \geq \pi$$

Transposing both sides of the first constraint leaves the inequality unchanged.

$$x^T A^T \leq r^T \quad A^T\lambda \geq \pi$$

Now, post-multiply both sides of the first constraint by λ and pre-multiply both sides of the second constraint by x^T, since both λ and x are nonnegative the inequality is preserved.

$$x^T A^T \lambda \leq r^T \lambda \quad x^T A^T \lambda \geq x^T \pi$$

Since $x^T \pi$ is a scalar, $x^T \pi = \pi^T x$. Now, combine the results and we have the relation we were after.

$$\pi^T x \leq x^T A^T \lambda \leq r^T \lambda$$

The solution to the dual lies above that for the primal except when they both reside at the optimum solution, in which case their objective function values are equal.

H.1.4 example

Suppose we wish to solve

$$\begin{array}{cc} \max\limits_{x \geq 0} & 10x + 12y \\ s.t & \begin{bmatrix} 2 & 1 \\ 1 & 2 \end{bmatrix} \begin{bmatrix} x \\ y \end{bmatrix} \leq \begin{bmatrix} 10 \\ 10 \end{bmatrix} \end{array}$$

We only need evaluate the objective function at each of the basic feasible solutions we earlier identified: $10\left(\frac{10}{3}\right)+12\left(\frac{10}{3}\right) = \frac{220}{3}$, $10\,(5)+12\,(0) = 50 = \frac{150}{3}$, and $10\,(0) + 12\,(5) = 60 = \frac{180}{3}$. The optimal solution is $x = y = \frac{10}{3}$.

H.1.5 complementary slackness

Suppose $x \geq 0$ is an n element vector containing a feasible primal solution, $\lambda \geq 0$ is an m element vector containing a feasible dual solution, $s \geq 0$ is an m element vector containing primal slack variables, and $t \geq 0$ is an n element vector containing dual slack variables. Then, x and λ are optimal if and only if (element-by-element)

$$xt = 0$$

and

$$\lambda s = 0$$

These conditions are economically sensible as either the scarce resource is exhausted ($s = 0$) or if the resource is plentiful it has no value ($\lambda = 0$).

H.2 Nonlinear programming

H.2.1 unconstrained

Nonlinear programs involve nonlinear objective functions. For instance,

$$\max_{x} \quad f\left(x\right)$$

If the function is continuously differentiable, then a local optimum can be found by the first order approach. That is, equate the gradient (a vector of partial derivatives composed of terms, $\frac{\partial f(x)}{\partial x_i}$, $i = 1, \ldots, n$ where there are n choice variables, x).

$$\begin{aligned} \nabla f\left(x^*\right) &= 0 \\ \begin{bmatrix} \frac{\partial f(x)}{\partial x_1} \\ \vdots \\ \frac{\partial f(x)}{\partial x_n} \end{bmatrix} &= \begin{bmatrix} 0 \\ \vdots \\ 0 \end{bmatrix} \end{aligned}$$

Second order (sufficient) conditions involve the Hessian, a matrix of second partial derivatives.

$$H\left(x^*\right) = \begin{bmatrix} \frac{\partial^2 f(x)}{\partial x_1 \partial x_1} & \cdots & \frac{\partial^2 f(x)}{\partial x_1 \partial x_n} \\ \vdots & \ddots & \vdots \\ \frac{\partial^2 f(x)}{\partial x_n \partial x_1} & \cdots & \frac{\partial^2 f(x)}{\partial x_n \partial x_n} \end{bmatrix}$$

For a local minimum, the Hessian is positive definite (the eigenvalues of H are positive). While for a local maximum, the Hessian is negative definite (the eigenvalues of H are negative).

H.2.2 convexity and global minima

If f is a convex function (defined below), then the set where f achieves its local minimum is convex and any local minimum is a global minimum. A function f is convex if for every x_1, x_2, and α , $0 \leq \alpha \leq 1$,

$$f\left(\alpha x_1 + \left(1 - \alpha\right) x_2\right) \leq \alpha f\left(x_1\right) + \left(1 - \alpha\right) f\left(x_2\right)$$

If $x_1 \neq x_2$, and $0 < \alpha < 1$,

$$f\left(\alpha x_1 + \left(1 - \alpha\right) x_2\right) < \alpha f\left(x_1\right) + \left(1 - \alpha\right) f\left(x_2\right)$$

then f is strictly convex. If $g = -f$ and f is (strictly) convex, then g is (strictly) concave.

H.2.3 example

Suppose we face the problem

$$\min_{x,y} \quad f(x,y) = x^2 - 10x + y^2 - 10y + xy$$

The first order conditions are

$$\nabla f(x,y) = 0$$

or

$$\begin{aligned} \frac{\partial f}{\partial x} &= 2x - 10 + y = 0 \\ \frac{\partial f}{\partial y} &= 2y - 10 + x = 0 \end{aligned}$$

Since the problem is quadratic and the gradient is composed of linearly independent equations, a unique solution is immediately identifiable.

$$\begin{bmatrix} 2 & 1 \\ 1 & 2 \end{bmatrix} \begin{bmatrix} x \\ y \end{bmatrix} = \begin{bmatrix} 10 \\ 10 \end{bmatrix}$$

or $x = y = \frac{10}{3}$ with objective function value $-\frac{100}{3}$. As the Hessian is positive definite, this solution is a minimum.

$$H = \begin{bmatrix} 2 & 1 \\ 1 & 2 \end{bmatrix}$$

Positive definiteness of the Hessian follows as the eigenvalues of H are positive. To see this, recall the sum of the eigenvalues equals the trace of the matrix and the product of the eigenvalues equals the determinant of the matrix. The eigenvalues of H are 1 and 3, both positive.

H.2.4 constrained — the Lagrangian

Nonlinear programs involve either nonlinear objective functions, constraints, or both. For instance,

$$\begin{array}{cc} \max\limits_{x \geq 0} & f(x) \\ s.t. & G(x) \leq r \end{array}$$

Suppose the objective function and constraints are continuously differentiable concave and an optimal solution exists, then the optimal solution can be found via the Lagrangian. The Lagrangian writes the objective function less a Lagrange multiplier times each of the constraints. As either the multiplier is zero or the constraint is binding, each constraint term equals zero.

$$\mathcal{L} = f(x) - \lambda_1 [g_1(x) - r_1] - \cdots - \lambda_n [g_n(x) - r_n]$$

where $G(x)$ involves n functions, $g_i(x)$, $i = 1, \ldots, n$. Suppose x involves m choice variables. Then, there are m Lagrange equations plus the n constraints that determine the optimal solution.

$$\begin{aligned}
\frac{\partial \mathcal{L}}{\partial x_1} &= 0 \\
&\vdots \\
\frac{\partial \mathcal{L}}{\partial x_m} &= 0 \\
\lambda_1 \left[g_1(x) - r_1\right] &= 0 \\
&\vdots \\
\lambda_n \left[g_n(x) - r_n\right] &= 0
\end{aligned}$$

The Lagrange multipliers (shadow prices or dual variable values) represent the rate of change in the optimal objective function for each of the constraints.

$$\lambda_i = \frac{\partial f(r^*)}{\partial r_i}$$

where r^* refers to rewriting the optimal solution x^* in terms of the constraint values, r. If a constraint is not binding, it's multiplier is zero as it has no impact on the optimal objective function value.[2]

H.2.5 Karash-Kuhn-Tucker conditions

Originally, the Lagrangian only allowed for equality constraints. This was generalized to include inequality constraints by Karash and separately Kuhn and Tucker. Of course, some regularity conditions are needed to ensure optimality. Various necessary and sufficient conditions have been proposed to deal with the most general settings. The Karash-Kuhn-Tucker theorem supplies first order necessary conditions for a local optimum (gradient of the Lagrangian and the Lagrange multiplier times the inequality constraint equal zero when the Lagrange multipliers on the inequality constraints are non-negative evaluated at x^*). Second order necessary (positive semi-definite Hessian for the Lagrangian at x^*) and sufficient (positive definite Hessian for the Lagrangian at x^*) conditions are roughly akin to those for unconstrained local minima.

[2] Of course, these ideas regarding the multipliers apply to the shadow prices of linear programs as well.

H.2.6 example

Continue with the unconstrained example above with an added constraint.

$$\begin{array}{ll} \min\limits_{x,y} & f(x,y) = x^2 - 10x + y^2 - 10y + xy \\ s.t. & \qquad\qquad xy \geq 10 \end{array}$$

Since the unconstrained solution satisfies this constraint, $xy = \frac{100}{9} > 10$, the solution remains $x = y = \frac{10}{3}$.

However, suppose the problem is

$$\begin{array}{ll} \min\limits_{x,y} & f(x,y) = x^2 - 10x + y^2 - 10y + xy \\ s.t. & \qquad\qquad xy \geq 20 \end{array}$$

The constraint is now active. The Lagrangian is

$$\mathcal{L} = x^2 - 10x + y^2 - 10y + xy - \lambda(xy - 20)$$

and the first order conditions are

$$\nabla \mathcal{L} = 0$$

or

$$\begin{aligned} \frac{\partial \mathcal{L}}{\partial x} &= 2x - 10 + y - \lambda y = 0 \\ \frac{\partial \mathcal{L}}{\partial y} &= 2y - 10 + x - \lambda x = 0 \end{aligned}$$

and constraint equation

$$\lambda(xy - 20) = 0$$

A solution to these nonlinear equations is

$$\begin{aligned} x &= y = 2\sqrt{5} \\ \lambda &= 3 - \sqrt{5} \end{aligned}$$

The objective function value is -29.4427 which, of course, is greater than the objective function value for the unconstrained problem, -33.3333. The Hessian is

$$H = \begin{bmatrix} 2 & 1-\lambda \\ 1-\lambda & 2 \end{bmatrix}$$

with eigenvalues evaluated at the solution, $3 - \lambda = \sqrt{5}$ and $1 + \lambda = 4 - \sqrt{5}$, both positive. Hence, the solution is a minimum.

H.3 Theorem of the separating hyperplane

The theorem of the separating hyperplane states either there exists a nonnegative y such that $Ay = x$ or there exists λ such that $A^T\lambda \geq 0$ and $\lambda^T x < 0$. The theorem is about mutual exclusivity — one or the other is true not both. This is similar to the way in which orthogonal complements are mutually exclusive. If one subspace contains all positive vectors the orthogonal complement cannot contain positive vectors. Otherwise, their inner products would be positive and inner products of orthogonal subspaces are zero.

The intuition follows from the idea that vector inner products are proportional to the cosine of the angle between them; if the angle is less (greater) than 90 degrees the cosine is positive (negative). $A^T\lambda \geq 0$ means the columns of A are less than or equal to 90 degrees relative a fixed vector λ while $\lambda^T x < 0$ implies the angle between x and λ exceeds 90 degrees. The separating hyperplane (hyper simply refers to high dimension) is composed of all vectors orthogonal to a fixed vector λ.

Consider a simple example.

Example 16 (simple example) *Suppose* $A = I$ *and* $x = \begin{bmatrix} 2 \\ 3 \end{bmatrix}$*, then* $y = 2\begin{bmatrix} 1 \\ 0 \end{bmatrix} + 3\begin{bmatrix} 0 \\ 1 \end{bmatrix}$ *and there exists no* λ *from which to form a plane separating the positive quadrant from* x*. On the other hand, suppose* $x = \begin{bmatrix} -2 \\ 3 \end{bmatrix}$ *—* x *lies outside the positive quadrant and* $\lambda = \begin{bmatrix} 1 \\ 0 \end{bmatrix}$ *satisfies the theorem's alternative,* $A^T\lambda = \begin{bmatrix} 1 \\ 0 \end{bmatrix} \geq 0$ *and* $\lambda^T x = -2 < 0$*.*

Next, consider a simple accounting (incidence matrix) example. That is, a case in which A has a nullspace and a left nullspace.

Example 17 (simple accounting example) *If* $A = \begin{bmatrix} -1 & 0 & 1 \\ 1 & -1 & 0 \\ 0 & 1 & -1 \end{bmatrix}$ *and* $x = \begin{bmatrix} 1 \\ 2 \\ -3 \end{bmatrix}$*, then* $y = \begin{bmatrix} 2 \\ 0 \\ 3 \end{bmatrix} + k\begin{bmatrix} 1 \\ 1 \\ 1 \end{bmatrix}$ *and any* $k \geq 0$ *satisfies* $Ay = x$ *and* $y \geq 0$*. Hence, there exists no separating plane based on* λ*. On the other hand, suppose* $A = \begin{bmatrix} -1 & 0 & -1 \\ 1 & -1 & 0 \\ 0 & 1 & 1 \end{bmatrix}$*. Now,* $y = \begin{bmatrix} 2 \\ 0 \\ -3 \end{bmatrix} + k\begin{bmatrix} 1 \\ 1 \\ -1 \end{bmatrix}$ *and no* k *satisfies* $Ay = x$ *and* $y \geq 0$*. Any number of* λ*s exist.*

For example, $\lambda = \begin{bmatrix} 0 \\ 0 \\ 1 \end{bmatrix}$ *produces* $A^T\lambda = \begin{bmatrix} 0 \\ 1 \\ 1 \end{bmatrix} \geq 0$ *and* $\lambda^T x = -3 < 0$.

Hence, any $\lambda = \begin{bmatrix} 0 \\ 0 \\ 1 \end{bmatrix} + k \begin{bmatrix} 1 \\ 1 \\ 1 \end{bmatrix}$ *(drawing on the left nullspace of* A*) where* $k > -1$ *satisfies the theorem's alternative.*