Contents

Preface

This book presents the key ideas and algorithms that are used to estimate an unknown real vector x from a set of indirect, inexact measurements $b_1, \ldots, b_m$. The book illustrates and motivates these ideas and algorithms by showing how they are applied to the estimation of geometric locations and paths. The estimation techniques that we discuss make two important assumptions. First, we assume that the quantities that we measure are generated from the unknown vector in a known way. For example, a GPS receiver at a location x measures the arrival times $b_1, \ldots, b_m$ of signals that were transmitted from known locations (in space) at known times. The exact arrival times are functions of the unknown location, $b_i = M_i(x)$, where M_i is a known function, at least approximately. Second, we assume that the statistical behavior of measurement errors is known. For example, we might assume that while our GPS receiver cannot determine $M_i(x)$ exactly, it can observe or estimate $b_i = M_i(x) + \epsilon_i$, where ϵ is a random vector with a known distribution (again, perhaps only approximately known). The techniques and algorithms that this book teaches use the vector b, our knowledge of M, and knowledge of the distribution of ϵ to estimate x and to assess the accuracy of this estimate.

Estimation problems of this type are important and ubiquitous in science and engineering because many quantities of interest cannot be measured directly and because measurements, both direct and indirect, are inexact. The length of a box can be measured directly with a ruler. The measurement is direct in the sense that we measure the length of the box by comparing it to many fixed lengths marked on the ruler. In contrast, we measure temperature indirectly. To measure the temperature of a liquid, we can place the bulb of a mercury-in-glass thermometer (or a less hazardous modern equivalent) in the liquid and read the temperature from a scale that relates the volume of mercury to its temperature. In this case, the function M transforms a temperature x to a volume $M(x)$. In most cases, both direct and indirect measurements of continuous quantities are inexact. When measuring lengths with a ruler, the inexactness is due to the limited resolution of markings on the ruler's edge and from possible changes in the length of the ruler itself (it may have expanded or shrunk).

Estimation of locations, in particular, is a special case that is important and ubiquitous on its own. It is an old problem with a fascinating history. In some settings, it still poses significant challenges today. To appreciate the historical context, think of marine navigators who must determine their location on an open featureless sea. Mariners were able to estimate latitude from measurements of the inclined position of stars since the 15th century. They only gained the ability to accurately estimate longitude in the 18th century, after chronometers (accurate clocks) were invented. Today, every smartphone includes a GPS receiver that can estimate the position of the phone from measurements of the arrival times of radio signals transmitted by satellites. However, GPS receivers do not work well inside buildings, so indoor location estimation remains an active area of research.

By focusing on location estimation, this book not only covers an important application of estimation but also teaches estimation without requiring any background in physics, chemistry, or signal processing. In location estimation, the vector that we need to estimate and the measurements that we use to estimate it are related in simple geometric ways. Location estimation provides a wealth of interesting and easy-to-model estimation problems.

The background that the book does require includes some linear algebra, calculus, and a little bit of continuous probability. These topics are normally taught in the first year of most bachelor degree programs in math, computer science, physics, and engineering. Therefore, the material that the book covers should be accessible to students in these fields starting from their second year of study, as well as to more advanced undergraduates, to graduate students, and to professionals. An appendix enumerates all the mathematical concepts and results that the book assumes that the reader knows.

The scope of the book is defined roughly by the range of algorithms and estimation techniques that are used in GPS receivers, including high-end ones. The book also covers some location-estimation techniques that are not used in GPS receivers, mainly when these techniques provide standalone motivation for key building blocks (e.g., leveling is used to motivate linear least squares). After reading this book you will understand how GPS receivers work at a deep algorithmic and mathematical level, you will understand how several other localization systems work, and you will be able to understand and develop models and algorithms for new location-estimation systems.

The book focuses on modeling and on algorithms. Most of the theory that explains the general statistical properties of estimators is omitted.

The book emphasizes recurring themes in estimation and in location estimation, especially least-squares minimization and the use of orthogonal factorizations, especially the QR and SVD factorizations. Least-squares minimization is used to solve linear problems (leveling), nonlinear problems (e.g., GPS), arrival-time estimation, dynamical systems (Kalman filtering and smoothing), and problems with integer parameters. At the same time, the book also highlights the differences between these problems: linear problems are convex, nonlinear problems can be (and often are) nonconvex, arrival-time estimation problems are highly nonconvex and require a grid search, and so on. The QR factorization is another recurring theme. It is used to solve linear least-squares problems, to compute the correction step in solvers for nonlinear problems, to perform Kalman filtering and smoothing, and to enable efficient search for integer solutions.

Several topics are presented in the book in a unique way. The QR factorization and the way that it is used to solve linear least-squares problems is presented in a particularly succinct way. Estimation of the covariance matrix of an estimator is presented using both the Jacobian of the model function and using the implicit function approach, which is more accurate. Arrival-time estimation of known pseudorandom signals is presented in a way that exposes the relationship between the maximum-likelihood criterion and cross correlation; this makes it clear how to produce subsample estimates without resorting to the sampling theorem (which is indeed not required). The significance of the discrete Fourier transform is explained by showing that it diagonalizes circulant matrices. Kalman filtering and smoothing is presented as a linear least-squares problem that can be efficiently solved using a structured sparse QR factorization.

Chapters end with notes that direct the reader to original sources and to books that cover the material in more depth or more breadth, and with problems that the reader is invited to solve. Many of the problems constitute small programming projects. The text of the problem provides detailed guidance and background on how to use MATLAB to solve the problem. Some of these problems focus on modeling, some on algorithms, and some on visualization of the behavior and results of algorithms. These problems complement the text in the sense that reading the text and solving the problems prepares the reader to developing new location estimators, all the way from the model, through the code, to the evaluation of the method. Other problems focus on the theory and are designed to help readers deepen their understanding of the material. Some problems rely on data files or MATLAB source-code files that are provided on the book's website.[1]

[1] `https://bookstore.siam.org/fa17/bonus`

I wrote the book intending that it be used as the main textbook in an elective single-semester course on location estimation in undergraduate or graduate programs in computer science and applied math. For such programs, the scope, the required background, and the level of rigor are just right. I have taught three such courses using drafts of the book. I believe that it is also suitable for electives in other majors, especially if augmented with some additional discipline-specific material. For example, students of physics or geodesy will probably benefit from some additional material on geodetic systems and on orbit representation.

Specific chapters, especially those that present material in a unique way, can be used as supplementary material in other courses. For example, the chapter on leveling can motivate linear least squares in courses on numerical linear algebra. The material on nonlinear location-estimation problems can motivate the study of nonlinear optimization algorithms. The material on cross correlation together with some of the material on arrival-time estimation can be used to motivate the fast Fourier transform. The chapter on Kalman filtering and smoothing can be used in courses on robotics and on signal processing.

I am grateful to Nicolàs de Hilster, a collector and scholar of navigation and surveying instruments, for allowing me to use images of instruments from his collection.[2] The function of an instrument is often more evident from an image of an early model than from an image of a modern model. (The old instruments whose images are shown in the book are also beautiful and display remarkable craftsmanship.) The book also shows a few old maps from the Historical Map & Chart Collection[3] of the Office of Coast Survey, part of the U.S. National Oceanic and Atmospheric Administration.

[2]http://www.dehilster.info

[3]https://historicalcharts.noaa.gov

Notation

Tables 1 and 2 show the usual meaning of letters that are used in mathematical expressions in this book. As much as possible, each capital or small letter represents at most a single concept. There is usually no connection between what a small letter represents and what the corresponding capital represents. For example, M always represents a model function, which maps unknowns to observable quantities, and m always represents the number of observations or constraints. Some letters represent more than one concept but using very different typefaces. For example, I represents the identity matrix, and $\mathcal{I}$ represents a Fisher information matrix. In a few cases keeping the notation reasonably conventional required us to use the same letter for different concepts in different chapters. For example, we use x to denote a generic (usually unknown) vector, but in location-estimation problems we use x to denote the first coordinate (x coordinate) of a location, and we use $\ell = \begin{bmatrix} x & y \end{bmatrix}^T$ or $\ell = \begin{bmatrix} x & y & z \end{bmatrix}^T$ to denote the entire vector of coordinates.

We use conventional notation from linear algebra, calculus, and probability, such as A^T and A^* (matrix transpose and conjugate transpose), $\partial/\partial x$ (partial derivatives), and $\mathrm{E}(x)$ (expectation). We also use conventional asymptotic notation for the complexity of algorithms, such as $O(n)$ and $\Theta(n^2)$; this notation is defined mathematically in virtually every textbook on algorithms, such as [19].

Subscripts usually denote elements of vectors and matrices, so b_5 is the fifth element of the vector b and $A_{i,j}$ or A_{ij} is the element of the matrix A in row i and column j. Vectors are always column vectors. We use so-called MATLAB notation for subvectors and submatrices: $x_{j:k}$ represents elements j through k of the vector x, and $A_{:,j:k}$ represents all the rows in columns j through k of the matrix A. Row and column indexes usually start from 1, except in Chapter 14, where they start from 0 because this is more convenient in Fourier transforms. Two chapters, 5 and 15, use block matrix notation, in which we use B_{ij} to denote a submatrix in block row i and block column j, as in

$$B = \begin{bmatrix} B_{11} & B_{12} \\ B_{21} & B_{22} \end{bmatrix} .$$

Here B_{11} refers to the first k rows and columns of B for some k, and so on. The same block B_{11} can also be denoted $B_{1:k,1:k}$ but the former notation is more compact. In a few places we use a single subscript to denote the index in a sequence of matrices, such as $Q_1, Q_2, \ldots, Q_k$. Chapter 14 uses subscripts to denote the dimension of matrices, so $F_{m\times m}$ is m-by-m.

We use $\mathrm{diag}(A)$ to denote the vector containing the diagonal elements of a matrix A and $\mathrm{diag}(x)$ to denote the diagonal matrix with $A_{i,i} = x_i$, where x is a vector. For clarity, we sometimes drop zero entries from a matrix, so

$$\begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} = \begin{bmatrix} 1 & \\ & 1 \end{bmatrix} .$$

Table 1. *Greek letters used in mathematical expressions.*

		α	complex amplitude
		β	phase
		γ	phase
Δ	perturbation or displacement; sampling period	δ	a small perturbation
		ϵ	vector of error or noise terms
		ζ	shrinkage bound in LLL algorithm
		η	satellite clock error
Θ	asymptotic growth rate	θ	an angle
Λ	$\mathrm{diag}(\begin{bmatrix}\lambda_1 & \lambda_2 & \cdots\end{bmatrix})$	λ	eigenvalue, wavelength in Chapter 16
		μ	micro, 10^{-6}
		ν	vector of integer parameters
		ξ	tropospheric delay
		π	pi, $\pi = 3.14\ldots$
		ρ	location of reference points
Σ	diagonal matrix of singular values	σ	standard deviation or singular value
		τ	propagation delay, tropospheric delay
		ϕ	objective function
		φ	phase
		ψ	ionospheric delay
Ω	diagonal matrix of FFT multipliers	ω	root of unity, $\omega = \omega_m = e^{2\pi\mathbf{i}/m}$

Constants like 0 and 1 denote either a scalar, a vector, or a matrix; the intention is usually clear from the context. In a few cases, we use bold $\mathbf{0}$ and $\mathbf{1}$ to denote vectors of zeros and ones, to distinguish them from scalar zeros and ones that appear nearby.

Unless marked otherwise, norms refer to the Euclidean norm,

$$\|r\| = \|r\|_2 = \sqrt{r_1^2 + r_2^2 + \cdots + r_n}.$$

We use $\lfloor a \rfloor$ to the denote the floor of a real number a (the largest integer smaller or equal to a), $\lceil a \rceil$ to denote the ceiling of a, and $\lfloor a \rceil$ the denote the rounding of a, the closest integer to a.

One aspect of the notation is specialized to estimation problems. We decorate the true value of the quantity or vector that we are trying to estimate with a ring above it (the ring reminds us that this is the target), for example, $\mathring{x}$. Other values that the quantity of vector might take (hypotheses) are denoted by the same letter without decoration, x in our example. The value of an estimator for $\mathring{x}$ is denoted by $\hat{x}$.

Table 2. *Latin letters used in mathematical expressions.*

A	matrix, linear part of a model	a	real amplitude
B	matrix	b	vector of observations
C	covariance matrix	c	propagation speed
$\mathbb{C}$	the set of complex numbers		
D	diagonal matrix	d	vector of differences or differential, as in $\int f(x)dx$
E	matrix, usually for elimination		
E	expectation (note the upright typeface)	e	base of the natural logarithm, $e = 2.71\ldots$; e_i is a unit vector
F	matrix	f	a function, often an estimator $\hat{x} = f(b)$; also frequency in Chapters 13 and 16
G	matrix	g	a function
H	circulant matrix	h	height in Chapter 3; vector in Chapter 14; a function elsewhere
I	identity matrix	i	row index, index of a constraint
$\mathcal{I}$	Fisher information matrix	$\mathbf{i}$	$\sqrt{-1}$
J	Jacobian (note the upright typeface)	j	column index, index of an unknown
K	matrix	k	index; number of time stamps in Chapter 15
L	lower-triangular matrix	ℓ	location vector
$\mathcal{L}$	log-likelihood function		
M	model function	m	number of constraints
N	nonlinear part of model function	n	number of unknowns
$\mathbb{N}$	the set of natural numbers		
O	asymptotic upper bound	o	clock offset
P	projection matrix	p	probability density function
Q	orthonormal matrix, usually from QR factorization	q	vector of nuisance parameters
R	upper-triangular matrix, usually from QR factorization	r	residual or rank
$\mathbb{R}$	the set of real numbers		
S	matrix, often representing weighting and projection $S = (I - QQ^T)W$	s	continuous signal (a real/complex function of time)
$\mathcal{S}$	score function (gradient of $\mathcal{L}$)		
T	transposition, as in A^T	t	time
U	orthonormal matrix, usually left singular vectors	u	vector
V	orthonormal matrix, usually right singular vectors	v	vector
W	weight matrix, $W^TW = C^{-1}$	w	vector
		x	unknown vector or x coordinate
		y	y coordinate
$\mathbb{Z}$	the set of integers	z	z coordinate

Index