University of Michigan - NAME 568/EECS 568/ROB 530 **Winter 2022**

Note: Estimation and Information Theory

Lecturer and Author: Maani Ghaffari

1 Estimation

A static estimation problem involve estimation of the system parameters. State estimation is a dynamic estimation problem, that is the state variables of the system evolve over time [1].

Definition 1 (Parameter)**.** *A quantity that is assumed to be time invariant or its time variation is slow compared to the state variables of a system.*

The PDF of the measurements, z, conditioned on the parameter, x, is called the *likelihood function* of the parameter

$$l(x) \triangleq p(z|x). \tag{1}$$

The likelihood measures that how likely a parameter value is given the obtained observations (evidence from the data).

1.1 Maximum Likelihood (ML) Estimator

The maximum likelihood method maximizes the likelihood function, leading to the ML estimator

$$x^\star(z) = \arg\max\ l(x) = \arg\max_x\ p(z|x), \tag{2}$$

where $x^\star(z)$ is a random variable as it is a function of random observations, z. The ML estimate is the solution of the *likelihood equation* (necessary condition)

$$\frac{\mathrm{d}l(x)}{\mathrm{d}x} = \frac{\mathrm{d}p(z|x)}{\mathrm{d}x} = 0. \tag{3}$$

Problem 1 (ML estimate of n Gaussian samples)**.** *Find the ML estimator of a sample of n i.i.d. normal random variables (univariate).*

1.2 Maximum A Posteriori (MAP) Estimator

In the Bayesian framework, one can place a prior distribution over the parameter. Then, the MAP estimator maximizes the posterior PDF

$$x^\star(z) = \arg\max_x\ p(x|z) = \arg\max_x\ p(z|x)p(x), \tag{4}$$

were the last equality is true because the normalization constant in the Bayes' formula is independent of x.

Remark 1. *Since* log *is a monotonic function, it is often the case that we use the logarithm of the likelihood or posterior for maximization.*

Problem 2 (MAP estimate of n Gaussian samples)**.** *Find the MAP estimator of a sample of n i.i.d. normal random variables (univariate). Assume the prior is also Gaussian with mean μ_0 and variance σ_0.*

1.3 Least Squares (LS) Estimator

A common estimation procedure for nonrandom parameters is the LS method. Let $h_k(x)$ be a nonlinear measurement model and $z_k = h_k(x) + v$ the scalar measurements. The LS estimator of x is the solution of the following problem known as the nonlinear LS.

$$x_k^\star = \arg\min_x \sum_{k=1}^{t} [z_k - h_k(x)]^2. \tag{5}$$

If $h_k(x)$ is linear in x, (5) becomes the linear LS problem. Furthermore, there is no assumption on the distribution of the noise term, v, in the LS problem. If we assume $v \sim \mathcal{N}(0, \sigma)$, then the LS estimator coincides with the ML estimator.

Problem 3 (Polynomial regression)**.** *Formulate the polynomial regression problem of order n using the LS estimator. Write a function that takes data and the order of the polynomial, and outputs the coefficients. Compare your results with MATLAB's built-in polyfit. What can you say about the goodness of fit? What happens when the order of the polynomial is exactly equal to the number of data points?*

1.4 Minimum Mean Square Error (MMSE) Estimator

For random parameters, the MMSE estimator is

$$x^\star(z) = \arg\min_{\hat{x}} \mathbb{E}[(\hat{x} - x)^2 | z], \tag{6}$$

where $\hat{x}$ is an estimator of x. The MMSE estimator's solution corresponds to the conditional mean of x, that is

$$x^\star(z) = \mathbb{E}[x|z] = \int_{-\infty}^{\infty} x p(x|z) dx, \tag{7}$$

and can be obtained by

$$\frac{\partial \mathbb{E}[(\hat{x} - x)^2 | z]}{\partial \hat{x}} = \mathbb{E}[2(\hat{x} - x)|z] = 2(\hat{x} - \mathbb{E}[x|z]) = 0.$$

Remark 2. *If $p(x|z)$ is Gaussian, then the MMSE estimator (the conditional mean) coincides with the MAP estimator since the mode and mean of the Gaussian distribution are the same.*

Example 1. *The MMSE estimate - the conditional mean - of a Gaussian random vector in terms of another Gaussian random vector (the measurement).*

1.5 Central Limit Theorem

If the random variables X_i are independent, under general conditions the distribution of

$$Y_n = \frac{1}{n}\sum_{i=1}^{n} X_i \tag{8}$$

is approximately Gaussian, of mean $\frac{1}{n}\sum_{i=1}^{n}\mu_i$ and variance $\frac{1}{n^2}\sum_{i=1}^{n}\sigma_i^2$, and μ_i and σ_i^2 are the mean and variance of X_i. As $n \to \infty$, the approximation becomes more accurate [2].

The Central Limit Theorem (CLT) has a very important role in characterizing many real-world sources of uncertainty: It is used as the justification/excuse to make the omnipresent Gaussian assumption. For example, the thermal noise in electronic devices, as the sum of many "small contributions," is indeed close to Gaussian [1].

1.6 Filtering, Smoothing, and Prediction

This part is a summary from Anderson and Moore [2, Chapter 2]. In general and a broad sense, *filtering* may refer to extraction of information about the internal state of a system from noisy measurements. Technically, filtering refer to a specific type of information processing (inference). In particular, in the latter definition, filtering means recovery of the state at time t, using measurements up to time t.

Example 2. *An example of the application of filtering in everyday life is in radio reception. Here the signal of interest is the voice signal. This signal is used to modulate a high frequency carrier that is transmitted to a radio receiver. The received signal is inevitably corrupted by noise, and so, when demodulated, it is filtered to recover as well as possible the original signal.*

Smoothing is another form of information processing that differs from filtering. In smoothing, information about the state need not become available at time t, and measurements derived later than time t can be used in obtaining information about the state. The trade-off here is that there must be a delay in producing the information about the state, as compared with the filtering case, but we can use more measurement data than in the filtering case in producing the information about the state. Since in smoothing we can also use measurements after time t, one should expect the smoothing process to be more accurate in some sense than the filtering process.

Example 3. *An example of smoothing is provided by the way the human brain tackles the problem of reading hastily written handwriting. Each word is tackled sequentially, and when word is reached that is particularly difficult to interpret, several words after the difficult word, as well as those before it, may be used to attempt to deduce the word.*

Prediction is the forecasting side of information processing. Here we intend to obtain information about the state at some time after t, while we may use measurements up to time t.

Example 4. *When attempting to catch a ball, we have to predict the future trajectory of the ball in order to position a catching hand correctly. This task becomes more difficult the more the ball is subject to random disturbances such as wind gusts. Generally, any prediction task becomes more difficult as the environment becomes noisier.*

2 Solving System of Linear Equations

For an accessible background on linear algebra and solving systems of linear equations, see the University of Michigan ROB 101 course materials [3].

2.1 Cholesky Decomposition

I took this part from the appendix of the well-known book on Gaussian processes [4]. The Cholesky decomposition of a symmetric, positive definite matrix A decomposes A into a product of a lower triangular matrix L and its transpose

$$LL^{\mathsf{T}} = A \tag{9}$$

where L is called the Cholesky factor. The Cholesky decomposition is useful for solving linear systems with symmetric, positive definite coefficient matrix A. To solve $Ax = b$ for x, first solve the triangular system $Ly = b$ by forward substitution and then the triangular system $L^{\mathsf{T}}x = y$ by back substitution. Using the backslash operator, we write the solution as $x = L^{\mathsf{T}}\backslash(L\backslash b)$, where the notation $A\backslash b$ is the vector x which solves $Ax = b$. Both the forward and backward substitution steps require $n^2/2$ operations, when A is of size $n \times n$. The computation of the Cholesky factor L takes time $n^3/6$, so it is the method of choice when it can be applied.

2.2 QR Decomposition

This part is taken from Lecture Notes for EE263, Stephen Boyd, Stanford 2008 [5]. Consider the case of overdetermined set of linear equations in which there are more equations than unknowns. We wish to approximately solve $Ax = b$ where now $A \in \mathbb{R}^{m\times n}$ is skinny, i.e., $m > n$. The QR decomposition or factorization finds $A = QR$ where $Q \in \mathbb{R}^{m\times n}$ is an orthogonal matrix, i.e., $Q^{\mathsf{T}}Q = I_n$, and $R \in \mathbb{R}^{n\times n}$ is an upper triangular and invertible matrix.

To approximately solve $Ax = b$, find $A = QR$ and substitute it into the pseudo-inverse of A, i.e., $(A^{\mathsf{T}}A)^{-1}A^{\mathsf{T}}$, resulting in $\hat{x} = R\backslash Q^{\mathsf{T}}b$.

3 Information Theory

Entropy is a measure of the uncertainty of a random variable [6]. The entropy $H(X)$ of a discrete random variable X is defined as

$$H(X) = \mathbb{E}_{p(x)}[\log \frac{1}{p(x)}] = -\sum_{\mathcal{X}} p(x) \log p(x). \tag{10}$$

The joint entropy $H(X, Y)$ of discrete random variables X and Y with a joint distribution $p(x, y)$ is defined as

$$H(X, Y) = -\sum_{x\in\mathcal{X}} \sum_{y\in\mathcal{Y}} p(x, y) \log p(x, y). \tag{11}$$

The chain rule implies that

$$H(X,Y) = H(X) + H(Y|X) = H(Y) + H(X|Y), \tag{12}$$

where $H(X|Y)$ is the conditional entropy and is defined as

$$H(Y|X) = -\sum_{x\in\mathcal{X}}\sum_{y\in\mathcal{Y}} p(x,y)\log p(y|x) \tag{13}$$

Theorem 1 (Chain rule for entropy)**.** *Let $X_1, X_2, ..., X_n$ be drawn according to the joint probability distribution $p(x_1, x_2, ..., x_n)$. Then*

$$H(X_1, X_2, ..., X_n) = \sum_{i=1}^{n} H(X_i|X_{i-1}, ..., X_1) \tag{14}$$

Proof. Please refer to [6] for the proof. □

The relative entropy or Kullback–Leibler Divergence (KLD) is a measure of distance between two distributions $p(x)$ and $q(x)$. It is defined as

$$D(p||q) = \mathbb{E}_{p(x)}[\log\frac{p(x)}{q(x)}]. \tag{15}$$

Theorem 2 (Information inequality)**.** *Let X be a discrete random variable. Let $p(x)$ and $q(x)$ be two probability mass functions. Then*

$$D(p||q) \geq 0, \tag{16}$$

with equality if and only if $p(x) = q(x)$ for all x.

Proof. Please refer to [6] for the proof. □

The mutual information $I(X;Y)$ is the reduction in the uncertainty of one random variable due to the knowledge of the other. The mutual information is non-negative and can be written as

$$I(X;Y) = D(p(x,y)||p(x)p(y)) = \mathbb{E}_{p(x,y)}[\log\frac{p(x,y)}{p(x)p(y)}], \tag{17}$$

$$I(X;Y) = H(X) - H(X|Y). \tag{18}$$

Corollary 3 (Nonnegativity of mutual information)**.** *For any two random variables X and Y,*

$$I(X;Y) \geq 0, \tag{19}$$

with equality if and only if X and Y are independent.

Proof. $I(X;Y) = D(p(x,y)||p(x)p(y)) \geq 0$, with equality if and only if $p(x,y) = p(x)p(y)$. □

Some immediate consequences of the provided definitions are as follows.

Lemma 4. *For any discrete random variable X, we have $H(X) \geq 0$.*

Proof. $0 \leq p(x) \leq 1$ implies that $\log \frac{1}{p(x)} \geq 0$. □

Theorem 5 (Conditioning reduces entropy)**.** *For any two random variables X and Y,*

$$H(X|Y) \leq H(X), \tag{20}$$

with equality if and only if X and Y are independent.

Proof. $0 \leq I(X;Y) = H(X) - H(X|Y)$. □

We now define the equivalent of the functions mentioned above for probability density functions.

Definition 2 (Differential entropy)**.** *Let X be a continuous random variable whose support set is $\mathcal{S}$. Let $p(x)$ be the probability density function for X. The differential entropy $h(X)$ of X is defined as*

$$h(X) = -\int_{\mathcal{S}} p(x) \log p(x) dx. \tag{21}$$

Remark 3. *The differential entropy of a set of continuous random variables that have a joint distribution is also defined using* (21)*.*

Definition 3 (Conditional differential entropy)**.** *Let X and Y be continuous random variables that have a joint probability density function $p(x, y)$. The conditional differential entropy $h(X|Y)$ is defined as*

$$h(X|Y) = -\int p(x, y) \log p(x|y) dx dy. \tag{22}$$

Definition 4 (Relative entropy (KLD))**.** *The relative entropy (KLD) between two probability density functions p and q is defined as*

$$D(p||q) = \int p \log \frac{p}{q}. \tag{23}$$

Definition 5 (Mutual information)**.** *The mutual information $I(X;Y)$ between two continuous random variables X and Y with joint probability density function $p(x, y)$ is defined as*

$$I(X;Y) = \int p(x, y) \log \frac{p(x, y)}{p(x)p(y)} dx dy. \tag{24}$$

References

[1] Y. Bar-Shalom, X. R. Li, and T. Kirubarajan, *Estimation with applications to tracking and navigation: theory algorithms and software*. John Wiley & Sons, 2001.

[2] B. D. Anderson and J. B. Moore, *Optimal filtering*. Englewood Cliffs, 1979.

[3] "ROB 101: Computational linear algebra," https://robotics.umich.edu/academic-program/course-offerings/rob101-fall-2021/, accessed: 2022-01-01.

[4] C. Rasmussen and C. Williams, *Gaussian processes for machine learning*. MIT press, 2006, vol. 1.

[5] S. Boyd, "Lecture notes for EE263," *Introduction to Linear Dynamical Systems*, 2008.

[6] T. M. Cover and J. A. Thomas, *Elements of information theory*. John Wiley & Sons, 1991.