The first class on estimation within the course Random Signals and Noise

Mark Shtaif

School of Electrical Engineering, Dept. of Physical Electron. Tel-Aviv University, Tel-Aviv, Israel shtaif@eng.tau.ac.il

Introduction

Our life is full of situations in which we need to estimate things that we do not know for sure. We do it based on related information that is available to us. Essentially what we mean by 'estimating' is the same as what we mean by 'guessing', and therefore I will use these terms interchangeably in what follows (although I must admit that the latter term does sound less scientific than the former). As an example, assume that we want to estimate someone's annual income. If the only thing that we know about him is the country where he lives, we will probably just guess the average income in that country (in fact, that is indeed what many students suggested when I ask them this question). If we have additional information, like the value of that person's house or car, or the neighborhood where he lives, we usually expect to arrive at a better guess. Intuitively, we would often estimate the man's income to be the average income of all people that we know that have a similar value house and car, and who live in a similar style of a neighborhood. We will see in what follows that this intuition is not a bad approach for estimating things.

We now wish to formulate these notions better. Assume that the thing that we want to know is X (e.g. someone's annual income), but the information that is available to us is Y (the value of his car). We want to use Y in order to make the best possible estimate of X. Of course, in general there may be several things that we want to guess (e.g. the annual income and the total worth), so that X may be a vector, and there can be several

parameters that we want to base our estimate on (e.g. the value of his car and the value of his house) so that Y may be a vector as well. But we need to start somewhere and we will assume for now that both X and Y are scalars. We will denote our guess of X by $\hat{X}$ and call the difference

$$e = X - \hat{X} \tag{1}$$

the estimation error. The intuitive approach that we have discussed earlier and to which we have referred as "natural" comes down to $\hat{X} = \mathbf{E}[X|Y]$. But in what sense is this solution "natural"? What we need in order to answer this question is some reasonably objective criterion of optimality, which is what we are going to present in what follows.

First of all, we should decide what is the quantity that we want to optimize. Since 'optimize' means 'maximize' or 'minimize' something, this quantity needs to be scalar. The error e would be a good candidate if it weren't random (namely e is not known to us, so we can't really minimize it directly in a meaningful manner). Assuming that we know something about the distribution of X and Y (which implies some knowledge of the distribution of e) we could minimize $\mathbf{E}[e]$, which is a deterministic quantity. But that wouldn't make much sense (think for yourselves why this is so). Alternatively, a much more meaningful choices would be to minimize $\mathbf{E}[|e|]$ or $\mathbf{E}[e^2]$, but even these will make sense in some cases, and no sense in others,[1] and we really need a better, or a more general framework.

So the idea is the following. We define a quantity $d(e)$, whose formal name is 'distortion measure', and which 'sort-of' represents the price that we will have to pay for being wrong when guessing the value of X. Then we look for an estimator which will minimize its expectation value $D = \mathbf{E}[d(e)]$.

[1] One example when these choices are not very good is when the consequences of having a positive error or a negative error are drastically different, like in the case of a pilot trying to estimate the height of his airplane above ground when planning an aerobatic maneuver. Clearly, in this case choosing a symmetric criterion like $\mathbf{E}[|e|]$ or $\mathbf{E}[e^2]$ doesn't seem like a very good idea.

Minimizing D is equivalent to minimizing the overall price of our guessing errors if we had to perform these guesses a very large number of times. Clearly, there are going to be some restrictions on $d(e)$. First we are going to require that $d(0) = 0$, namely making the correct guess should not infer any cost. Secondly, since making the correct guess is always better than having an error, we must require that $d(e) \geq 0$. These very natural restrictions will allow us to come up with a theory for the estimation process. Examples for possible choices for $d(e)$ are $d(e) = |e|$, $d(e) = e^2$, $d(e) = e^2u(e) + 20e^2u(-e)$ (where $u(e)$ is the Heaviside step function), and we will see more example in what follows. Notice that the last of the above examples is asymmetric, such that the cost of a negative error is 20 times larger than when the error is positive. Another example would be to choose $d(e) = 1 - \delta_{e,0}$, where $\delta_{e,0}$ is the Kronecker delta function, which is equal to 1, whenever $e = 0$. In this case $d(e)$ is zero only when the guess is exactly correct, and it is equal to 1 in all other cases. As you may guess, this last choice can be appropriate only in the case of discrete variables, or else the event $e = 0$ will almost never take place.

Let us now formulate the above discussion and see how it goes

$$\begin{aligned} D = \mathbf{E}[X - \hat{X}(Y)] &= \int \mathrm{d}x \int \mathrm{d}y f_{X,Y}(x,y) d(x - \hat{X}(y)) \\ &= \int \mathrm{d}y f_Y(y) \int \mathrm{d}x f_{X|Y}(x|y) d(x - \hat{X}(y)). \\ &= \int \mathrm{d}y f_Y(y) \mathbf{E}\Big[d(X - \hat{X}(y))|Y = y\Big]. \end{aligned} \tag{2}$$

Notice that $f_Y(y)$ is not something that we have control of and $\mathbf{E}\Big[d(x - \hat{X}(y))|Y = y\Big]$ is a non-negative number. Hence the estimator that will minimize D will be the one for which $\mathbf{E}\Big[d(x - \hat{X}(y))|Y = y\Big]$ is the smallest possible. The formal way of expressing it is

$$\hat{X}(y) = \underset{\alpha}{\operatorname{argmin}} \left\{ \mathbf{E}\Big[d(X-\alpha) | Y = y \Big] \right\}. \tag{3}$$

You may not be used to such expressions, so lets go over what they mean. The expression in the curly brackets on the right-hand-side is a function of y and of α (it is not a function of X because we are averaging over X). So pick y and then draw this expression in the curly brackets as a function of α. The estimator $\hat{X}(y)$ for that particular value of y corresponds to the value of α for which the curve that you drew assumes its minimum value. It would be best to add a drawing here, but I will do it on the whiteboard in class instead. This way one can find the best estimator for each value of y.

The maximum a posteriori probability (MAP) estimator

There are many cases in which a discrete random variable X that assumes one of N possible values $x_1, x_2, \ldots, x_N$ represents an entity whose numerical representation is irrelevant. For example, the marital status of people applying for a certain job can be encoded as follows $X = 1$ represents single, $X = 2$ represents married, and $X = 3$ represents divorced. Clearly, the numbers 1,2 and 3 have no physical meaning, as they have been arbitrarily selected in order to represent the different groups. If a person whose status we wish to estimate is single and we guessed married (so that $e = 1$) the situation may be just as bad as it would be if we guessed that he was divorced, in spite of the fact that the error in the latter case is larger ($e = 2$). Since there is no importance to the magnitude of the error in this case, a sensible choice would be

$$d(e) = 1 - \delta_{e,0} = \begin{cases} 1 & e \neq 0 \\ 0 & e = 0 \end{cases}. \tag{4}$$

This implies that

$$\begin{aligned} \mathbf{E}\Big[d(e) | Y = y \Big] &= P(e \neq 0 | Y = y) \\ &= 1 - P(e = 0 | Y = y), \end{aligned} \tag{5}$$

which is just the probability of an error given the measured value of the quantity Y. Namely, the minimization of D is equivalent to minimizing the probability of an error, or alternatively maximizing the probability of being correct — which justifies the title given to this subsection. Simply "guess the value of X for which the probability that you are wrong, given the knowledge of Y is the smallest."

Example: Assume a binary random variable

$$X = \begin{cases} 1, & \text{with probability } p \\ -1, & \text{with probability } 1-p \end{cases}.$$

The measured quantity is $Y = X + N$ where $N \sim N[0, \sigma^2]$ is statistically independent of X. We need to guess the value of X when given the value of Y. Before we start to do this formally, lets think of what would be intuitive for us. If $p = 1/2$, X is symmetrically distributed and then it seems natural to guess $\hat{X} = 1$ if $Y > 0$ and $\hat{X} = -1$ if $Y < 0$. If $p > 1/2$ we may prefer to guess $\hat{X} = 1$ even if Y is slightly negative (after all if we take it to the extreme and assume $p = 1$, we would be very unwise to guess $X = -1$ regardless of the value of Y). Similarly, we will likely guess $\hat{X} = -1$ even for slightly positive Y when $p < 1/2$. The variance of N also matters. If it is small, we will have more confidence in our choices than if it is large. Lest see now how it turns out to work in practice. Formally

$$\hat{X}(y) = \underset{\alpha}{\operatorname{argmax}} \left\{ P(X = \alpha | Y = y) \right\}.$$

Since X is binary, there are only two possibilities for α. So what we need to see is which one is bigger $P(X = 1|Y = y)$ or $P(X = -1|Y = y)$. Lets massage this a little bit

$$P(X = \alpha | Y = y) = \frac{P(Y = y|X = \alpha)P(X = \alpha)}{P(Y = y)},$$

but since Y is a continuous variable we have to interpret $P(Y = y|X = \alpha) \to f_{Y|X}(y|\alpha)\mathrm{d}y$ and $P(Y = y) \to f_Y(y)\mathrm{d}y$, which produces

$$P(X=1|Y=y)=\frac{f_{Y|X}(y|1)p}{f_Y(y)}=\frac{p\exp\left(-\frac{(y-1)^2}{2\sigma^2}\right)}{p\exp\left(-\frac{(y-1)^2}{2\sigma^2}\right)+(1-p)\exp\left(-\frac{(y+1)^2}{2\sigma^2}\right)}$$

$$P(X=-1|Y=y)=\frac{f_{Y|X}(y|-1)p}{f_Y(y)}=\frac{(1-p)\exp\left(-\frac{(y+1)^2}{2\sigma^2}\right)}{p\exp\left(-\frac{(y-1)^2}{2\sigma^2}\right)+(1-p)\exp\left(-\frac{(y+1)^2}{2\sigma^2}\right)},$$

where we have used the fact that $Y|X\sim N[X,\sigma^2]$. Now we compare the two expressions. Since the denominators are identical, we only need to compare the numerators, which comes down to testing whether the probability ratio

$$\left(\frac{p}{1-p}\right)\exp\left(\frac{(y+1)^2-(y-1)^2)}{2\sigma^2}\right)=\left(\frac{p}{1-p}\right)\exp\left(2y/\sigma^2\right),$$

is larger or smaller than 1. Notice that if $p=1/2$ the above is greater than 1 whenever $y>0$ and vice versa, just as we have predicted intuitively, and in general the threshold value of y is

$$y_{th}=\frac{\sigma^2}{2}\ln\left(\frac{1-p}{p}\right). \tag{6}$$

Hence, our guess for the value of X will be

$$\hat{X}(y)=\begin{cases}1, & y>y_{th}\\ -1, & y<y_{th}\end{cases},$$

and of course, when $y=y_{th}$ (an event that may happen, although it has zero probability), we may just flip a coin.

The minimum mean-square error (MMSE) estimator

Perhaps the most popular choice when estimating continuous variables is $d(e)=e^2$. It corresponds well with our affinity for second order statistics (variances etc.), and unlike the choice of $d(e)=|e|$, it is analytic and hence conveniently lends itself to manipulation. Of course, it is not by any means

an exclusive, or even the most appropriate, choice in all cases, as we have discussed earlier.

Formally, what we are looking at is

$$\hat{X}(y) = \underset{\alpha}{\operatorname{argmin}} \left\{ \mathbf{E}\big[(X-\alpha)^2 | Y = y\big] \right\}. \tag{7}$$

Namely, what we need to do is minimize the function

$$\mathbf{E}\big[(X-\alpha)^2 | Y = y\big] = \alpha^2 - 2\alpha \mathbf{E}[X|Y=y] + \mathbf{E}[X^2|Y=y] \tag{8}$$

with respect to α when y is a parameter. This is a trivial procedure and you may readily check that the result is

$$\hat{X}(y) = \mathbf{E}\big[X|Y=y\big]. \tag{9}$$

Namely, the optimal guess for X using the MMSE criterion is just the average of X when conditioned on the measured value of Y — the exact result that we have suggested intuitively at the beginning of this lesson.

Example: Assume $X \sim N[0,1]$ and $Y = X + N$ with $N \sim N[0,\sigma^2]$, where X and N are statistically independent. The independence of X and N implies that they are jointly Gaussian and hence X and Y are jointly Gaussian as well (make sure that you understand why this is so). Therefore, using the relations that we have seen previously for jointly Gaussian variables, we have

$$\hat{X}(y) = \mathbf{E}[X|Y=y] = \frac{\mathrm{Cov}(X,Y)}{\mathrm{Var(Y)}} y = \frac{1}{1+\sigma^2} y.$$

In order to obtain some insight, note that when $\sigma^2 \ll 1$, the noise is not noticeable and the best thing to do is just guess $X = y$. On the other hand, when $\sigma^2 \gg 1$ the noise dominates the measurement and the knowledge of Y tells us very little about what X is. In that case the best guess will be close to the mean of X, namely it would be close to 0.

Example: Lets solve the same problem but with the binary variable $X = \pm 1$,

as we have had in the case of MAP. Only this time we will use the MMSE estimator. The result, as we shall see, will be less meaningful than it was in the case of MAP estimation. Here is the question again for your convenience,

$$X = \begin{cases} 1, & \text{with probability } p \\ -1, & \text{with probability } 1-p \end{cases},$$

and the measured quantity is $Y = X + N$ where $N \sim N[0, \sigma^2]$ is statistically independent of X.

$$\hat{X}(y) = \mathbf{E}[X|Y = y] = P(1|Y = y) - P(-1|Y = y),$$

where the conditional probabilities $P(\pm 1|Y = y)$ have been found earlier, so that

$$\begin{aligned} \hat{X}(y) &= \frac{p\exp\left(-\frac{(y-1)^2}{2\sigma^2}\right) - (1-p)\exp\left(-\frac{(y+1)^2}{2\sigma^2}\right)}{p\exp\left(-\frac{(y-1)^2}{2\sigma^2}\right) + (1-p)\exp\left(-\frac{(y+1)^2}{2\sigma^2}\right)} \\ &= \frac{p\exp\left(\frac{y}{\sigma^2}\right) - (1-p)\exp\left(-\frac{y}{\sigma^2}\right)}{p\exp\left(\frac{y}{\sigma^2}\right) + (1-p)\exp\left(-\frac{y}{\sigma^2}\right)}. \end{aligned}$$

For some insight, lets consider the special case where $p = 1/2$ and

$$\hat{X}(y) = \tanh(y/\sigma^2). \tag{10}$$

This is a smoothened version of a step-function and it is very clear that the values of $\hat{X}(y)$ are always between -1 and 1. For example $\hat{X}(0) = 0$, which is not a very meaningful estimate for X when the only possible values are $X = \pm 1$. Indeed, this example demonstrates the fact that the MMSE approach may not be very meaningful in some cases.

The linear MMSE estimator (LMMSE)

The MMSE estimator is a very powerful tool, but it is not always easy to extract it. Moreover, in some cases, when estimation needs to be done in

practice and in real time, there are complexity constraints that prevent the implementation of complicated functions. In those cases we may want to look at a simplified estimator; for example one in which the relation between $\hat{X}$ and y is linear

$$\hat{X}(Y) = aY + b. \tag{11}$$

The only thing that needs to be done in this case is determine the best values of a and b, i.e. the ones for which $\mathbf{E}[e^2]$ is minimal. Lets find these values directly in a brute-force procedure.

$$\begin{aligned}\mathbf{E}[e^2] &= \mathbf{E}[(X - aY - b)^2] \\ &= E[X^2] - 2a\mathbf{E}[XY] - 2b\mathbf{E}[X] + a^2\mathbf{E}[Y^2] + 2ab\mathbf{E}[Y] + b^2. \end{aligned} \tag{12}$$

By taking a derivative with respect to b and equating it to 0 we find that

$$b = \mathbf{E}[X] - a\mathbf{E}[Y], \tag{13}$$

and when equating to zero the derivative with respect to a, we find that $a\mathbf{E}[Y^2] + b\mathbf{E}[Y] = \mathbf{E}[XY]$. Substituting Eq. (13) for b, one obtains

$$a = \frac{\mathrm{Cov}(X, Y)}{\mathrm{Var}(Y)}. \tag{14}$$

The final result for the linear MMSE estimator is therefore

$$\hat{X} = \mathbf{E}[X] + \frac{\mathrm{Cov}(X, Y)}{\mathrm{Var}(Y)} \left(Y - \mathbf{E}[Y]\right). \tag{15}$$

Substituting Eq. (15) for the estimation error we obtain

$$\mathbf{E}[(X - \hat{X})^2] = \mathrm{Var}(X) - \frac{\mathrm{Cov}^2(X, Y)}{\mathrm{Var}(Y^2)} = \mathrm{Var}(X)\left(1 - \rho_{X,Y}^2\right), \tag{16}$$

where $\rho_{X,Y}$ is Pierson's covariance coefficient. Notice that the MSE is 0 only if $|\rho_{X,Y}| = 1$, which (as we have shown in the past) happens only when the ratio between X and Y is a deterministic constant.

Some properties of MMSE estimators

Bias: It is easy to see that a MMSE estimator (whether linear or general) is unbiased. By unbiased we mean that $\mathbf{E}[e] = 0$. Assume that it where different and $\mathbf{E}[e] = m \neq 0$. Then we would have that $\mathbf{E}[(e-m)^2] < \mathbf{E}[e^2]$ (make sure that you know how to show it), which implies that our estimator was not optimal in the first place.[2]

Orthogonality: The optimal MMSE estimator is always orthogonal to the estimation error, namely $\mathbf{E}[e\hat{X}] = 0$. In fact, we can make a stronger statement than this by saying that *the error of the optimal MMSE estimator is orthogonal to any function of the measurement* Y and that *the error of the Linear MMSE estimator is orthogonal to any linear function of* Y. The straightforward way of seeing this would be to verify that the optimal MMSE estimator of Eq. (9) and the linear MMSE estimator of Eq. (15) both satisfy these respective properties (please do it at home). But let me show you now (first for the case of the optimal MMSE estimator) that any function of y that satisfies the orthogonality condition is an optimal MMSE estimator. Assume that $\hat{X}(y)$ is an optimal estimator whereas $\hat{X}'(y)$ is a function of y that satisfies $\mathbf{E}[(X - \hat{X}'(y))g(y)] = 0$ for any function $g(y)$. Lets express the mean-square estimation error as follows

$$\begin{aligned}\mathbf{E}[(X-\hat{X})^2] &= \mathbf{E}\Big[[(X-\hat{X}') + (\hat{X}'-\hat{X})]^2\Big] \\ &= \mathbf{E}[(X-\hat{X}')^2] + \underbrace{\mathbf{E}[(\hat{X}'-\hat{X})^2]}_{\geq 0} + \underbrace{\mathbf{E}[(X-\hat{X}')(\hat{X}'-\hat{X})]}_{=0}. \quad (17)\end{aligned}$$

Since by assumption $(X - \hat{X}')$ is orthogonal to any function of y, it is orthogonal to $(\hat{X}' - \hat{X})$, implying that the last term on the right-hand-side of (17) is 0. The first term on the right-hand-side of (17) is the MSE of $\hat{X}'(y)$ and the second term is never negative. Hence, the smallest error is obtained when

[2] After all, we could just subtract m from it and it would become better.

$\hat{X} = \hat{X}'$, which concludes the proof of our statement. Notice now that very minor adaptations (find out which) the same proof can be used to demonstrate that a linear estimator $\hat{X}(y)$ whose error is orthogonal to any linear function of y is the best possible linear MMSE estimator. In the linear case, one can come up with a geometric interpretation of the orthogonality principle. One that gives insight into the idea of linear estimation in general. Can you try to picture what this geometric representation might be? In any case, we will discuss this in class.

The Pythagorean theorem: One of the consequences of the orthogonality relation is that e, X and $\hat{X}$ satisfy the Pythagorean theorem. Namely

$$\mathbf{E}[X^2] = \mathbf{E}[\hat{X}^2] + \mathbf{E}[e^2], \tag{18}$$

as readily follows from orthogonality together with the relation $X = \hat{X} + e$.

The Gaussian case: When X and Y are jointly Gaussian, we have seen in class that

$$\mathbf{E}[X|Y = y] = \mathbf{E}[X] + \frac{\mathrm{Cov}(X,Y)}{\mathrm{Var}(Y)}\left(Y - \mathbf{E}[Y]\right), \tag{19}$$

and therefore in the case of jointly Gaussian variables, the optimal estimator is also a linear one.

This is perhaps a good point to prove Eq. (19), which has been stated without proof in the past. Of course, the brute-force proof was outlined to you before,[3] which is why here we will give an alternative and much more elegant proof, that goes as follows. If X, Y are jointly Gaussian, the linear MMSE estimator $\hat{X}_{\mathrm{lin}}$ of X on the basis of Y and the linear estimation error $e_{\mathrm{lin}} = X - \hat{X}_{\mathrm{lin}}$ are also jointly Gaussian with X and Y (make sure that you understand why). The linear orthogonality condition states that e_{lin} and Y are uncorrelated and since they are jointly Gaussian, they must also be

[3] Express $f_{X|Y}(x|y)$ using the usual procedure, and demonstrate that it is Gaussian with the mean being given by Eq. (19), and with the variance given by Eq. (16).

statistically independent, in which case $\mathbf{E}[e_{\text{lin}}|Y=y]=0$. Using the relation $X=\hat{X}_{\text{lin}}+e_{\text{lin}}$, we can express the optimal MMSE estimator as

$$\mathbf{E}[X|Y=y]=\mathbf{E}[\hat{X}_{\text{lin}}(y)|Y=y]+\mathbf{E}[e_{\text{lin}}|Y=y]=\hat{X}_{\text{lin}}(y). \tag{20}$$

This concludes the proof of Eq. (19).