Parameter Estimation I: Classical

Dr. Edmund Lam

Department of Electrical and Electronic Engineering
The University of Hong Kong

ELEC2844: Probabilistic Systems Analysis
(Second Semester, 2020–21)

https://www.eee.hku.hk/~elec2844

Motivation

Suppose we have a possibly biased coin with unknown probability p that Head will show up. Our best strategy to estimate p is perhaps to experiment:

- What if we flip 1000 times and you get 500 heads?
- What if we flip 999 times and you get 500 heads?
- What if we flip once and you get head?
- What if we flip 10 times and you get 4 heads, but the coin looks "normal" and you would normally think that it is a fair coin?

In this lecture, we will explore how to make estimate on probabilistic quantities based on observations. We begin with the basic *language* and *context* of this class of problems.

1 Estimator and estimation error
2 Maximum likelihood (ML)
3 Examples of computing the ML estimates
4 Linear regression with Gaussian noise using ML
5 Linear regression with Gaussian noise in higher dimensions
6 Linear regression with Laplacian noise in higher dimensions
7 (Advanced Topic) Information theory and ML

Definition

The "classical" method of (statistical) parameter estimation goes as follows.

We want to estimate an unknown constant θ,
based on observations $X = (X_1, \ldots, X_n)$ that depend on this parameter.

$$\boxed{\text{Estimator: } \hat{\Theta} = g(X)} \tag{1}$$

- X depends on θ, so $\hat{\Theta}$ depends on θ
- X is a random variable, so $\hat{\Theta}$ is also a random variable
  - We can ask for $\mathbb{E}\left(\hat{\Theta}\right)$ and $\mathrm{var}\left(\hat{\Theta}\right)$, for example
- The actual realized value of $\hat{\Theta}$ is the estimate

Definition

Often, the number of individual observations n has a significant influence on the "goodness" of the estimator, so we emphasize that in the notation:

$$\boxed{\text{Estimator: } \hat{\Theta}_n = g(X)} \tag{2}$$

For different values of n:

- We can view $\hat{\Theta}_n$ as *a sequence of estimators*
- We can also have a sequence of $\mathbb{E}\left(\hat{\Theta}_n\right)$ and $\text{var}\left(\hat{\Theta}_n\right)$

Definition

Estimates are not 100% accurate:

$$\boxed{\text{Estimation error: } \tilde{\Theta}_n = \hat{\Theta}_n - \theta} \tag{3}$$

The estimation error depends on θ and observations X.

Bias is the expected value of the estimation error:

$$\boxed{b(\tilde{\Theta}_n) = \mathbb{E}\left(\hat{\Theta}_n\right) - \theta} \tag{4}$$

Expected value, variance, and bias depend on θ.

We have an unbiased estimator if $b(\tilde{\Theta}_n) = 0$ for all possible values of θ.

Definition

The general question:

How do we design the function $g(\cdot)$, such that the estimator $\hat{\Theta}_n = g(X)$ is "good" in some sense? (e.g. small estimation error, unbiased)

The classical approach is something called maximum likelihood.

1. Estimator and estimation error
2. Maximum likelihood (ML)
3. Examples of computing the ML estimates
4. Linear regression with Gaussian noise using ML
5. Linear regression with Gaussian noise in higher dimensions
6. Linear regression with Laplacian noise in higher dimensions
7. (Advanced Topic) Information theory and ML

Motivating example

Suppose there is a lucky draw. The participant reaches into a box and picks up a ball; if the ball is gold in color, he or she wins a prize. The ball is then returned into the box. You somehow know that altogether there are 7 identical balls in the box, but you do not know how many of them are gold.

You observed that five people participated in the lucky draw, and the first and fourth won a prize, while the second, third, and fifth didn't. What would be your "best estimate" of the number of gold balls in the box?

Ans: Let k be the number of gold balls. It seems $\frac{2}{5}$ wins a prize. So, $k = 7 \times \frac{2}{5} \approx 3$? Is there a way to formalize this argument?

Motivating example

We know the following:

- Every lucky draw is independent, with identical distribution.
- Probability of getting a gold ball is $p = \frac{k}{7}$.
- Each draw is a Bernoulli,

$$X \sim \text{Ber}\left(\frac{k}{7}\right)$$

  and we label $X = 1$ for prize, and $X = 0$ for no prize.
- The probability of observing $\{1, 0, 0, 1, 0\}$ is

$$p \cdot (1-p) \cdot (1-p) \cdot p \cdot (1-p) = \frac{k^2(7-k)^3}{7^5}$$

Motivating example

It makes sense to ask: What value of k would make the observation $\{1, 0, 0, 1, 0\}$ most likely to happen? Let's try to evaluate $\frac{k^2(7-k)^3}{7^5}$ for different values of k.

k	0	1	2	3	4	5	6	7
$\frac{k^2(7-k)^3}{7^5}$	0	0.012	0.030	0.034	0.026	0.012	0.002	0

So, $k = 3$ is indeed the most sensible answer.

Definition

Depending on an unknown parameter θ, the observations $X = (X_1, \ldots, X_n)$ can be described by a joint PMF / PDF:

$$\begin{aligned} \text{discrete:} \quad & p_X(x_1, \ldots, x_n; \theta) \\ \text{continuous:} \quad & f_X(x_1, \ldots, x_n; \theta) \end{aligned}$$

We call these the likelihood function.

Interpretation: $p_X(x; \theta)$ *or* $f_X(x; \theta)$ *is the probability of the observed value* x *when the parameter is* θ*, but* ***not*** *the probability that the unknown parameter is* θ*.*

Main idea

What is the value of θ such that the observations $x = x_1, \ldots, x_n$ are ***most likely*** *to appear?*

ANS: We find the θ that maximizes the likelihood function, hence the method is known as maximum likelihood (ML).

$$\text{discrete:} \quad \hat{\theta}_n = \arg\max_{\theta} p_X(x_1, \ldots, x_n; \theta) \tag{5}$$

$$\text{continuous:} \quad \hat{\theta}_n = \arg\max_{\theta} f_X(x_1, \ldots, x_n; \theta) \tag{6}$$

Note: The notation arg max *means we want to maximize the quantity, but we don't quite care the value of that expression, but are interested in the argument that leads to such maximum.*

Main idea

Often, the observations X_i are assumed to be independent, and perhaps even identically distributed. The likelihood function is then

$$\text{discrete:} \quad p_X(x_1, \ldots, x_n; \theta) = \prod_{i=1}^{n} p_{X_i}(x_i; \theta)$$

$$\text{continuous:} \quad f_X(x_1, \ldots, x_n; \theta) = \prod_{i=1}^{n} f_{X_i}(x_i; \theta)$$

Computationally, for ML, we often maximizes the log-likelihood:

$$\text{discrete:} \quad \log p_X(x; \theta) = \log \prod_{i=1}^{n} p_{X_i}(x_i; \theta) = \sum_{i=1}^{n} \log p_{X_i}(x_i; \theta)$$

$$\text{continuous:} \quad \log f_X(x; \theta) = \log \prod_{i=1}^{n} f_{X_i}(x_i; \theta) = \sum_{i=1}^{n} \log f_{X_i}(x_i; \theta)$$

1. Estimator and estimation error
2. Maximum likelihood (ML)
3. Examples of computing the ML estimates
4. Linear regression with Gaussian noise using ML
5. Linear regression with Gaussian noise in higher dimensions
6. Linear regression with Laplacian noise in higher dimensions
7. (Advanced Topic) Information theory and ML

Example 1: Poisson

Determine the ML estimator of the parameter of a Poisson random variable based on independent identically distributed observations $X_1, \ldots, X_n$. Is the estimator unbiased?

ANS: Let us proceed step-by-step.

1. Let the unknown Poisson parameter be θ.
2. The PMF of X_i is $p_{X_i}(x_i) = \dfrac{e^{-\theta}\theta^{x_i}}{x_i!} \qquad x_i = 0, 1, \ldots$
3. The log-likelihood function is

$$\begin{aligned} \log p_X(x_1, \ldots, x_n; \theta) &= \sum_{i=1}^{n} \log p_{X_i}(x_i; \theta) \\ &= -n\theta + \sum_{i=1}^{n} x_i \log \theta - \sum_{i=1}^{n} \log (x_i!) \end{aligned}$$

Example: Poisson

4. We maximize it by setting the derivative (w.r.t. θ) to 0:

$$0 = -n + \frac{1}{\theta}\sum_{i=1}^{n} x_i$$

so the estimator is $\hat{\Theta}_n = \frac{1}{n}\sum_{i=1}^{n} X_i$

5. Taking expectation,

$$\mathbb{E}\left(\hat{\Theta}_n\right) = \frac{1}{n}\sum_{i=1}^{n}\mathbb{E}(X_i) = \theta$$

since $\mathbb{E}(X_i) = \theta$ for all i. Therefore, the estimator is unbiased.

Example 2: Uniform

Observations $X_1, \ldots, X_n$ are independent identically distributed, being uniform over $[0, \theta]$, where θ is unknown. Find the ML estimator of θ.

ANS:

1. The PDF of X_i is $f_{X_i}(x_i) = \begin{cases} 1/\theta & \text{if } 0 \leqslant x_i \leqslant \theta \\ 0 & \text{otherwise} \end{cases}$
2. The likelihood function is

$$\begin{aligned} f_X(x_1, \ldots, x_n; \theta) &= f_{X_1}(x_1; \theta) \cdots f_{X_n}(x_n; \theta) \\ &= \begin{cases} 1/\theta^n & \text{if } 0 \leqslant \max_{i=1,\ldots,n} x_i \leqslant \theta \\ 0 & \text{otherwise} \end{cases} \end{aligned}$$

3. We maximize the likelihood function by noting that we need θ to be as small as possible, hence $\hat{\Theta}_n = \max_{i=1,\ldots,n} X_i$

Example 3: Laplacian

Observations $X_1, \ldots, X_n$ are independent identically distributed, following the (mean-shifted) Laplacian PDF $f_{X_i}(x_i) = \frac{1}{2}e^{-|x_i - \theta|}$, where θ is unknown. Find the ML estimator of θ.

ANS: From the PDF, we can find the log-likelihood function

$$\begin{aligned} \log f_X(x; \theta) &= \sum_{i=1}^{n} \log \tfrac{1}{2}e^{-|x_i - \theta|} \\ &= n \log \tfrac{1}{2} - \sum_{i=1}^{n} |x_i - \theta| \end{aligned}$$

This is maximized when the second term is minimized, which occurs when

$$\hat{\Theta}_n = \text{median}(X_1, \ldots, X_n)$$

Example 4: Binomial

A coin has an unknown probability θ to result in head.

(a) We toss the coin n times, where random variable K denotes the resulting number of heads. Find the ML estimator of θ based on K.

(b) We fix k, and let N be a random variable denoting the number of tosses until the kth head occurs. Find the ML estimator of θ based on N.

Example 4: Binomial

ANS: (a) The PMF of K is binomial:

$$p_K(k;\theta) = \binom{n}{k}\theta^k(1-\theta)^{n-k} \qquad k = 0, 1, 2, \ldots, n$$

This is also the likelihood function (to get our observation $K = k$).

Taking the derivative w.r.t. θ and setting it to zero,

$$0 = \binom{n}{k}k(1-\theta)^{n-k}\theta^{k-1} - \binom{n}{k}(n-k)(1-\theta)^{n-k-1}\theta^k$$

which yields the ML estimator $\hat{\Theta} = \dfrac{K}{n}$.

This is the fraction of heads in the observed trials.

Example 4: Binomial

ANS: (b) We need to first find the PMF of N. The kth head occurs at $N = n$ toss means that exactly $(k-1)$ heads have occurred among the first $(n-1)$ tosses, and that the nth toss is a head. Therefore,

$$p_N(n;\theta) = \binom{n-1}{k-1}\theta^{k-1}(1-\theta)^{n-k}\cdot\theta = \binom{n-1}{k-1}\theta^k(1-\theta)^{n-k}$$

for $n = k, k+1, \ldots$.

Taking the derivative w.r.t. θ and setting it to zero,

$$0 = \binom{n-1}{k-1}k(1-\theta)^{n-k}\theta^{k-1} - \binom{n-1}{k-1}(n-k)(1-\theta)^{n-k-1}\theta^k$$

which yields the ML estimator $\hat{\Theta} = \dfrac{k}{N}$.

This is again the fraction of heads in the observed trials.

Example 5: Hypergeometric

Suppose we want to estimate the population (n) of a certain endangered species in a habitat. One strategy to do that is to first catch a number of them (m), tag each of them, and then release them back to the wild. Some time later (assuming that the population stays the same, with negligible new birth and death during the intervening period of time), we go there and catch these animals again (j of them), among which k of them are found to be tagged. What is the ML estimate of N?

Example 5: Hypergeometric

Ans: Let K be the random variable of the number of tagged animals found. The original n animals are divided into two sets: m being tagged, and $(n-m)$ being untagged. So, there are $\binom{m}{k}$ possible ways of having k tagged animals, and $\binom{n-m}{j-k}$ untagged, out of a total of $\binom{n}{j}$ possibilities. So, the PMF of K is

$$p_K(k;n) = \frac{\binom{m}{k}\binom{n-m}{j-k}}{\binom{n}{j}}$$

this is known as a *hypergeometric random variable*. We also denote the above as $L(n)$ because it is the likelihood function.

Example 5: Hypergeometric

We need to find n that maximizes $L(n)$. Note that n is integer-valued. A trick here is to note that

$$\frac{L(n)}{L(n-1)} = \frac{(n-m)(n-j)}{n(n-m-j+k)}$$

This ratio is greater than one if and only if

$$(n-m)(n-j) \geqslant n(n-m-j+k)$$

which reduces to $mj \geqslant nk$, or,

$$n \leqslant \frac{mj}{k}$$

So, $L(n)$ is increasing at first, and later decreasing, where the maximum is at the largest integer value not exceeding mj/k. This is the ML estimate!

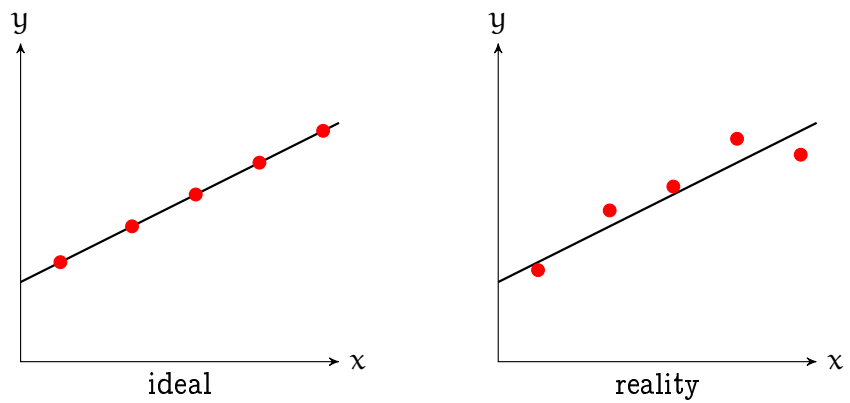
1. Estimator and estimation error
2. Maximum likelihood (ML)
3. Examples of computing the ML estimates
4. Linear regression with Gaussian noise using ML
5. Linear regression with Gaussian noise in higher dimensions
6. Linear regression with Laplacian noise in higher dimensions
7. (Advanced Topic) Information theory and ML

Motivating example

Suppose that in a semester, a pandemic causes all final exams to be canceled. The students, however, still need grades. In a particular course, fortunately the students have already obtained marks for their homework. In addition, the teacher has past statistics of student performance in previous years.

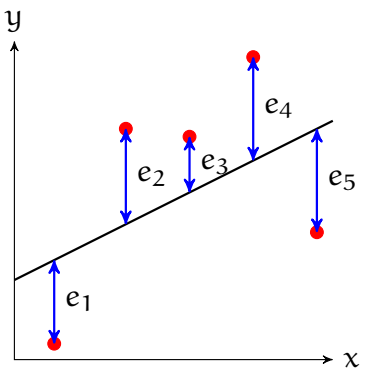
So, the teacher proceeds as follows: Let (x_i, y_i) be the actual homework and final marks of student i who took the course last year, where $i = 1, 2, \ldots, N$. He *postulates* that there is a linear relationship between the homework and final marks. Based on these data points, he tries to find the line. Then, he can use this to estimate the final exam of every student this semester, based on the student's homework performance.

Motivating example



How do we find a line that somehow "fits" the data points?

Motivating example



We want to minimize the "sum" of these errors. Practically, we will see soon that one good approach is to minimize the sum of the *square* of these errors, i.e., $\sum_{i=1}^{5} e_i^2$. The reason has to do with *why* there are errors: they are due to randomness in the data points, which follow a certain kind of distributions!

Setting

We now consider a general scenario, and invoke probability.

- We have two variables of interest, x and y, with a collection of data points, (x_i, y_i) for $i = 1, \ldots, n$.
- We *assume* a linear relationship between them,

$$y \approx \theta_0 + \theta_1 x$$

  where θ_0 and θ_1 are *unknown parameters to be estimated.*
- We approach the problem by assuming that each y_i is a realization of a random variable Y_i, generated according to

$$y_i = \theta_0 + \theta_1 x_i + \epsilon_i \qquad i = 1, \ldots, n$$

  where ϵ_i's are independent, identically distributed Gaussian random variables with mean zero and variance σ^2.
- We compute the ML estimates of θ_0 and θ_1.

Computing the ML estimates

1. Given θ_0 and θ_1, we have $Y_i = \mathcal{N}\left(\theta_0 + \theta_1 x_i, \sigma^2\right)$
2. Likelihood function is therefore

$$f_Y(y; \theta_0, \theta_1) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi}\sigma} \exp\left\{-\frac{(y_i - \theta_0 - \theta_1 x_i)^2}{2\sigma^2}\right\}$$

3. Log-likelihood function is

$$\log f_Y(y; \theta_0, \theta_1) = \sum_{i=1}^{n} \left(\log \frac{1}{\sqrt{2\pi}\sigma} - \frac{(y_i - \theta_0 - \theta_1 x_i)^2}{2\sigma^2}\right)$$

4. Maximizing the log-likelihood is equivalent to

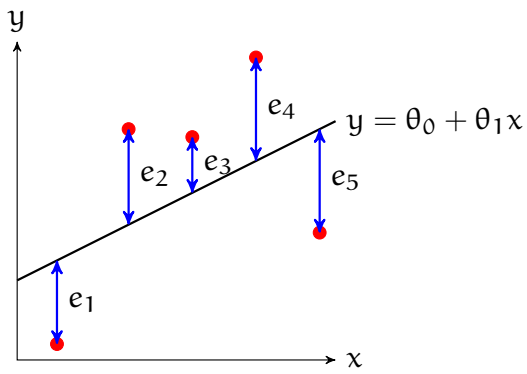
$$\boxed{\text{minimize} \quad \sum_{i=1}^{n} (y_i - \theta_0 - \theta_1 x_i)^2} \tag{7}$$

Computing the ML estimates

The computation

$$\text{minimize} \quad \sum_{i=1}^{n} (y_i - \theta_0 - \theta_1 x_i)^2$$

is essentially minimizing the square of the errors:



Computing the ML estimates

5. Let $S = \sum_{i=1}^{n} (y_i - \theta_0 - \theta_1 x_i)^2$.

$$\frac{\partial S}{\partial \theta_0} = \sum_{i=1}^{n} 2\left(y_i - \theta_0 - \theta_1 x_i\right)(-1) = 0$$

$$\sum_{i=1}^{n} y_i - n\theta_0 - \theta_1 \sum_{i=1}^{n} x_i = 0$$

$$\theta_0 = \frac{1}{n}\sum_{i=1}^{n} y_i - \theta_1 \cdot \frac{1}{n}\sum_{i=1}^{n} x_i$$

By defining $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$ and $\bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i$,

$$\boxed{\hat{\Theta}_0 = \bar{y} - \hat{\Theta}_1 \bar{x}} \tag{8}$$

Computing the ML estimates

$$\frac{\partial S}{\partial \theta_1} = \sum_{i=1}^{n} 2\left(y_i - \theta_0 - \theta_1 x_i\right)(-x_i) = 0$$

$$\sum_{i=1}^{n} x_i y_i - \theta_0 \sum_{i=1}^{n} x_i - \theta_1 \sum_{i=1}^{n} x_i^2 = 0$$

$$\sum_{i=1}^{n} x_i y_i - (\bar{y} - \theta_1 \bar{x}) \sum_{i=1}^{n} x_i - \theta_1 \sum_{i=1}^{n} x_i^2 = 0$$

$$\theta_1 = \frac{\sum_{i=1}^{n} x_i y_i - n\bar{x}\,\bar{y}}{\sum_{i=1}^{n} x_i^2 - n\bar{x}^2}$$

A more "aesthetic" formula is

$$\boxed{\hat{\Theta}_1 = \frac{\sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n} (x_i - \bar{x})^2}} \tag{9}$$

1. Estimator and estimation error
2. Maximum likelihood (ML)
3. Examples of computing the ML estimates
4. Linear regression with Gaussian noise using ML
5. Linear regression with Gaussian noise in higher dimensions
6. Linear regression with Laplacian noise in higher dimensions
7. (Advanced Topic) Information theory and ML

Higher dimensions

We can generalize the above to a linear regression of the form

$$y \approx w_0 x_0 + w_1 x_1 + \ldots + w_D x_D$$

where we can assume $x_0 = 1$ to take care of the constant, and $x_1, \ldots, x_D$ are D "features". Sometimes, we simply ignore the $w_0 x_0$ term in the derivations.

This is extremely useful when we want to estimate the weights w_i for each feature to match our observation y, so that in the future, when we have a new set of features, we can generate a new y easily.

Linear regression in higher dimensions

For linear regression, we approach the problem by assuming that each y_i is a realization of a random variable Y_i, generated according to

$$y_i = \theta^T x_i + \epsilon_i \qquad i = 1, \ldots, n$$

where ϵ_i's are independent, identically distributed Gaussian random variables with mean zero and variance σ^2. Here, θ is a length-D vector of parameters to estimate; x_i is a length-D vector of features.

We therefore aim to compute the ML estimates of θ.

Linear regression in higher dimensions

1. Given the vector θ, we have $Y_i = \mathcal{N}\left(\theta^T x_i, \sigma^2\right)$
2. Likelihood function is therefore

$$f_Y(y;\theta) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi}\sigma} \exp\left\{-\frac{(y_i - \theta^T x_i)^2}{2\sigma^2}\right\}$$

3. Log-likelihood function is

$$\log f_Y(y;\theta) = \sum_{i=1}^{n} \left(\log \frac{1}{\sqrt{2\pi}\sigma} - \frac{(y_i - \theta^T x_i)^2}{2\sigma^2}\right)$$
$$= n \log \frac{1}{\sqrt{2\pi}\sigma} - \frac{1}{2\sigma^2}(X\theta - y)^T(X\theta - y)$$

where y is a length-n vector containing all y_i, and X is a $n \times D$ matrix containing all x_i.

Linear regression in higher dimensions

Maximizing the log-likelihood is equivalent to

$$\boxed{\text{minimize} \quad \sum_{i=1}^{n} \left(y_i - \theta^T x_i\right)^2} \tag{10}$$

which can also be written as

$$\text{minimize} \quad (X\theta - y)^T(X\theta - y)$$

This has an analytic solution

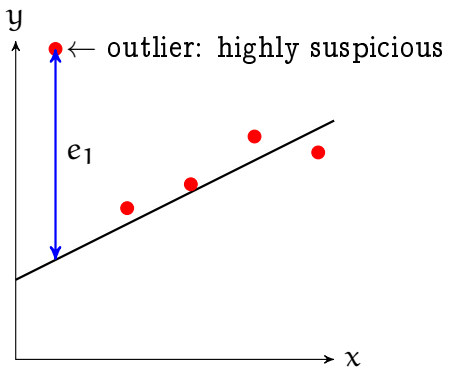
$$\boxed{\hat{\Theta} = \left(X^T X\right)^{-1} X^T y} \tag{11}$$

It is also known as the minimum mean square error (MMSE) or ℓ_2-norm solution.

1. Estimator and estimation error
2. Maximum likelihood (ML)
3. Examples of computing the ML estimates
4. Linear regression with Gaussian noise using ML
5. Linear regression with Gaussian noise in higher dimensions
6. Linear regression with Laplacian noise in higher dimensions
7. (Advanced Topic) Information theory and ML

Motivating example

Back to our pandemic-need-to-infer-final-from-homework example. Suppose we get the following data:



The line will be significantly influenced by the outlier, because we are minimizing the square of e_1.

Motivating example

One technique that is more robust against outliers is to minimize the absolute error of the noise, i.e.,

$$\sum_{i=1}^{n} |e_i|$$

It turns out we can again relate that to an assumption of the noise distributions.

Linear regression in higher dimensions

We again approach the problem by assuming that each y_i is a realization of a random variable Y_i, generated according to

$$y_i = \theta^T x_i + \epsilon_i \qquad i = 1, \ldots, n$$

but this time, ϵ_i's are independent, identically distributed Laplacian random variables with mean zero and parameter λ. We will see that a Laplacian distribution indeed is robust against outliers.

Once again, we aim to compute the ML estimates of θ.

Linear regression in higher dimensions

1. Given the vector θ, we have $Y_i = \text{Lap}\left(\theta^T x_i, \lambda\right)$
2. Likelihood function is therefore

$$f_Y(y;\theta) = \prod_{i=1}^{n} \frac{\lambda}{2} \exp\left\{-\lambda \left|y_i - \theta^T x_i\right|\right\}$$

3. Log-likelihood function is

$$\begin{aligned} \log f_Y(y;\theta) &= \sum_{i=1}^{n} \left(\log \frac{\lambda}{2} - \lambda \left|y_i - \theta^T x_i\right|\right) \\ &= n \log \frac{\lambda}{2} - \lambda \sum_{i=1}^{n} \left|y_i - \theta^T x_i\right| \end{aligned}$$

Linear regression in higher dimensions

Maximizing the log-likelihood is equivalent to

$$\boxed{\text{minimize} \quad \sum_{i=1}^{n} |y_i - \theta^T x_i|} \tag{12}$$

This is also known as the minimum absolute difference (MAD) or minimum absolute error (MAE), or ℓ_1-norm solution.

ℓ_p-norm and generalized Gaussian distribution

We can generalize to an ℓ_p-norm solution:

$$\boxed{\text{minimize} \quad \sum_{i=1}^{n} |y_i - \theta^T x_i|^p} \tag{13}$$

We have already seen that $p = 2$ corresponds to Gaussian, and $p = 1$ corresponds to Laplacian. Any others we can relate?

ANS: Generalized Gaussian distribution, where β controls the shape.

1. Estimator and estimation error
2. Maximum likelihood (ML)
3. Examples of computing the ML estimates
4. Linear regression with Gaussian noise using ML
5. Linear regression with Gaussian noise in higher dimensions
6. Linear regression with Laplacian noise in higher dimensions
7. (Advanced Topic) Information theory and ML

Information theory

Our goal in this part is to introduce a few key concepts in information theory, and relate in particular to the ML estimation for further insight.

- Entropy
- Joint entropy
- Conditional entropy
- Relative entropy, also known as the Kullback-Leibler divergence
- Mutual information

Entropy

Entropy is a measure of *uncertainty of a random variable*. Let X be a random variable with PMF $p_X(x)$. Then the entropy of X is defined as

$$\boxed{H(X) = -\sum_x p_X(x) \log_2 p_X(x) = \mathbb{E}\left(\log \frac{1}{p_X(x)}\right)} \tag{14}$$

The entropy is a measure of uncertainty of a random variable. The logarithm is base 2, and entropy is measured in bits.

- By convention, $0 \log_2 0 = 0$, which is justified because $\lim_{x \to 0} x \log x = 0$.
- Interpretation: Adding terms with zero probability should not change the entropy.
- $H(X) \geqslant 0$ because $-\log_2 p_X(x) \geqslant 0$, since $0 \leqslant p_X(x) \leqslant 1$.

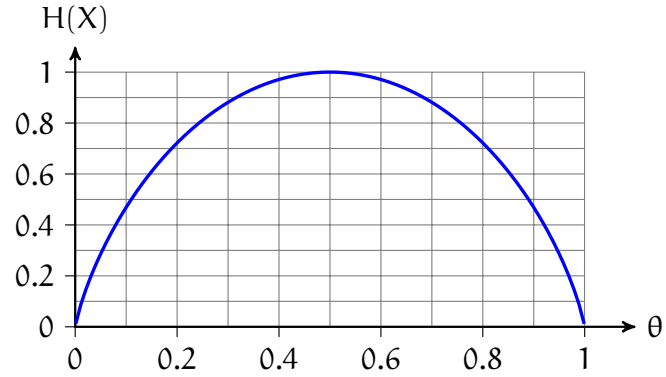
Entropy

For a Bernoulli random variable with parameter θ,

$$p_X(x) = \theta^x (1-\theta)^{1-x} \qquad x = 0, 1$$

Therefore, the entropy is

$$H(X) = -\theta \log_2(\theta) - (1-\theta) \log_2(1-\theta)$$



Example

Entropy of flipping coins:

- Unbiased coin: $H(X) = -0.5 \log_2 0.5 - 0.5 \log_2 0.5 = 1$ bit.
- Biased coin: $H(X) = -0.9 \log_2 0.9 - 0.1 \log_2 0.1 = 0.469$ bit.

Another example: $X = \begin{cases} a & \text{with probability } \frac{1}{2} \\ b & \text{with probability } \frac{1}{4} \\ c & \text{with probability } \frac{1}{8} \\ d & \text{with probability } \frac{1}{8}. \end{cases}$ Then,

$$H(X) = -\tfrac{1}{2} \log_2 \tfrac{1}{2} - \tfrac{1}{4} \log_2 \tfrac{1}{4} - \tfrac{1}{8} \log_2 \tfrac{1}{8} - \tfrac{1}{8} \log_2 \tfrac{1}{8} = \tfrac{7}{4} \text{ bit.}$$

$H(X)$ measures the number of bits needed to represent the information. *The more uneven the distribution, the better!* (smaller $H(X)$)

Joint entropy

The joint entropy is a measure of *uncertainty of a pair of random variables*. Let (X, Y) be a pair of random variables with PMF $p_{X,Y}(x, y)$. Then the joint entropy of (X, Y) is defined as

$$\boxed{H(X, Y) = -\sum_{x,y} p_{X,Y}(x, y) \log_2 p_{X,Y}(x, y) = \mathbb{E}\left(\log \frac{1}{p_{X,Y}(x, y)}\right)} \tag{15}$$

Conditional entropy

The conditional entropy is defined as

$$\begin{aligned} H(Y|X) &= \sum_x p_X(x)\, H(Y|X=x) \\ &= -\sum_x p_X(x) \left(\sum_y p_{Y|X}(y|x) \log p_{Y|X}(y|x) \right) \\ &= -\sum_x \sum_y p_{X,Y}(x,y) \log p_{Y|X}(y|x) \\ &= -\mathbb{E}\big(\log p_{Y|X}(y|x)\big) \end{aligned}$$

which leads to the chain rule

$$\boxed{H(X,Y) = H(X) + H(Y|X)} \tag{16}$$

since $\log p_{X,Y}(x,y) = \log p_X(x) + \log p_{Y|X}(y|x)$.

Relative entropy

To simplify notation, we use

$p(x)$ for one PMF over different values of $\{X = x\}$

$q(x)$ for another PMF over the same set of $\{X = x\}$

The relative entropy, a.k.a. the Kullback-Leibler (or KL) divergence, between two distributions $p(x)$ and $q(x)$ is defined as

$$\boxed{D(p\|q) = \sum_x p(x) \log_2 \left(\frac{p(x)}{q(x)} \right) = \mathbb{E}_p \left[\log_2 \left(\frac{p(x)}{q(x)} \right) \right]} \tag{17}$$

Note: $D(p\|q) \geqslant 0$ and equals to zero if and only if $p(x) = q(x)$.

Relative entropy

The KL divergence can be thought of as measuring the "distance" between two distributions, but mathematically does not qualify as a distance metric.

Example: Suppose we have two Bernoulli distributions $p \sim \mathrm{Ber}(r)$ and $q \sim \mathrm{Ber}(s)$.

$$\begin{aligned} D(p\|q) &= (1-r) \log \frac{1-r}{1-s} + r \log \frac{r}{s} \\ D(q\|p) &= (1-s) \log \frac{1-s}{1-r} + s \log \frac{s}{r} \end{aligned}$$

If $r = s$, we can see $D(p\|q) = D(q\|p) = 0$.

If $r = 0.5$ and $s = 0.25$, we have $D(p\|q) = 0.2075$ and $D(q\|p) = 0.1887$. Note that $D(p\|q) \neq D(q\|p)$ in general.

Relative entropy and mutual information

The relative entropy leads us to the definition of mutual information, which is the relative entropy between the joint distribution and the product distribution:

$$\boxed{I(X;Y) = \sum_{x,y} p_{X,Y}(x,y) \log \frac{p_{X,Y}(x,y)}{p_X(x)\, p_Y(y)} = D(p_{X,Y}(x,y) \,\|\, p_X(x)\, p_Y(y))} \tag{18}$$

We can further derive that

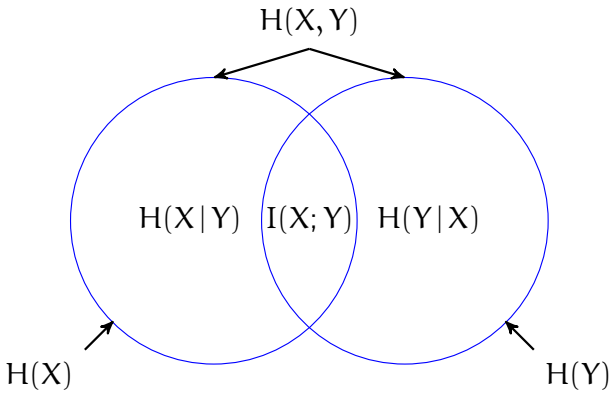
$$\boxed{I(X;Y) = H(X) - H(X|Y) = H(Y) - H(Y|X)} \tag{19}$$

Mutual information $I(X;Y)$ *is the reduction in uncertainty of* X *due to knowledge of* Y.

Relative entropy and mutual information

We can also substitute $H(X,Y) = H(X) + H(Y|X)$ to get

$$\boxed{I(X;Y) = H(X) + H(Y) - H(X,Y)} \tag{20}$$



Relative entropy and ML estimation

Relationship between relative entropy and ML estimation:

Problem setting:

- We get data $x_1, \ldots, x_N$ from some unknown distribution $p(x)$ (not necessarily of any particular parametric form).
- We wish to fit a distribution $q(x;\theta)$ that has some parametric form (e.g. Gaussian) that best explains the data.
- We derive the ML estimate for the parameters of distributions (θ) of a certain parametric form.

Relative entropy and ML estimation

The ML estimate is given by:

$$\begin{aligned}
\hat{\Theta} &= \arg\max_{\theta} \prod_{i=1}^{N} q(x_i \,|\, \theta) \\
&= \arg\max_{\theta} \sum_{i=1}^{N} \log q(x_i \,|\, \theta) \\
&= \arg\max_{\theta} \frac{1}{N} \sum_{i=1}^{N} \log q(x_i \,|\, \theta) - \frac{1}{N} \sum_{i=1}^{N} \log p(x_i) \\
&= \arg\min_{\theta} \frac{1}{N} \sum_{i=1}^{N} \log \left(\frac{p(x_i)}{q(x_i \,|\, \theta)} \right)
\end{aligned}$$

Note that the highlighted term does not affect the arg max.

Relative entropy and ML estimation

In the limit as $N \to \infty$, we replace the *average* by the *expectation*,

$$\begin{aligned}
\hat{\Theta} &\approx \arg\min_{\theta} \mathbb{E}\left[\log \left(\frac{p(x)}{q(x \,|\, \theta)} \right) \right] \\
&= \arg\min_{\theta} D(p(x) \| q(x \,|\, \theta))
\end{aligned}$$

Interpretation: The ML estimate can be viewed as finding parameters (from some family of distributions) that minimizes the relative entropy between the true distribution generating the data and the modeled distribution from this family.