2 Recap of Basic Concepts of Statistics

In this document we review some preliminary concepts of statistics used in what follows of the course and how statistical tools can be used in MATLAB . After that, we briefly mention some of the most known optimization techniques. This is not supposed to be an exhaustive document on the two topics, for more information you may refer to:

- Bishop, C.M., "Pattern recognition and machine learning", 2006, Springer;
- Montgomery, D.C., Runger G.C., "Applied statistics and probability for engineers", 2010, John Wiley & Sons.

2.1 Discrete Random Variables

A discrete random variable X is a variable with values in a discrete set E whose value is determined by a stochastic phenomenon, i.e., we are not able to predict its value even if we are given precise information about the phenomenon. For instance, consider a 20-faced dice: the event of trowing it can be modeled as a random variable X taking values in the finite set of events $E = \{1, \dots, 20\}$, since we are not able to predict precisely which value might occur, even if we are given all the characteristics of the dice (e.g., dimensions, initial position, speed).

To properly model this phenomenons, we define a *probability function* $\mathbb{P} : E \to [0, 1]$ which tells you how often the event i belonging to a discrete set of events E occurs (e.g., probability that by trowing the dice you get 3) as:

$$\mathbb{P}(X = i) := \frac{|i|}{|E|},$$

where $|i|$ is the measure of the favorable events set and $|E|$ is the measure of set of the possible events. For instance, for a 20-faced dice we have:

$$\mathbb{P}(X = i) = \frac{1}{20},$$

since all the faces have the same probability to occur. In this case we have some properties that the probability function should have:

- $0 \leq \mathbb{P}(X = i) \leq 1$: an event can occur at least with zero probability and at most with probability one (the favorable cases are at most equal to the possible ones);
- $\sum_{i \in E} \mathbb{P}(X = i) = 1$: if we consider probability of the set of all the possible events E we should get one.

While this is the probability of a single events we might be interested in the probability of multiple events. For instance one might be interested in the probability that the dice roll is too small, say less than a specific value. This probability is captured by the *cumulative function* $F : E \to [0, 1]$, which specifies the probability that the random variable is lower than i:

$$F(i) := \mathbb{P}(X \leq i) = \sum_{h=1}^{i} \mathbb{P}(X = h) = \sum_{h \in E, h \leq i} \frac{|h|}{|E|}.$$

This newly defined function satisfies:

- $0 \leq F(i) \leq 1$: the sum of the probabilities should be between zero and one:
- $F(i) = 0, \ \forall i < \min_{h \in E} h$: if we consider a value small enough (smaller than the smaller element in the event space) the cumulative function should have value zero;
- $F(i) = 1, \ \forall i \geq \max_{h \in E} h$: if we consider a value large enough (larger or equal than the element in the event space) the cumulative function should have value one.

In the 20-faced dice case we have:

$$F(i) = \sum_{h=1}^{i} \frac{1}{20} = \frac{i}{20}.$$

There are two quantities which might be of major interest in the study of a random variable: the *expected value* and the *variance*. In the case of the dice, the former tells you what is value on average one could get from trowing the dice repeatedly, the latter gives us information about the spread in the single results we would have. Formally, for a generic random variable we have:

$$\mathbb{E}[X] := \sum_{i \in E} i \ P(X = i);$$
$$Var(X) := \sum_{i \in E} (\mathbb{E}[X] - i)^2 P(X = i).$$

In the case of the 20-faced dice we have:

$$\mathbb{E}[X] := \sum_{i=1}^{20} \frac{i}{20} = \frac{1}{20}\frac{20(20+1)}{2} = \frac{21}{2};$$
$$Var(X) := \sum_{i=1}^{20} \frac{\left(\frac{21}{2} - i\right)^2}{20} = \frac{57}{4}.$$

Sometimes the "spread" of a random variable is evaluated with the *standard deviation*, which is the square root of the variance $std(X) = \sqrt{Var(X)}$. This is due to the fact that if we are considering random variables expressed in some unit of measure the variance is not compatible with the measurement itself, but its squared root does.

Remark 1. *Notice that these are the values obtained by knowing the random variable. If we only have some samples coming from the random variable, we are not able to compute the expected value and the variance, but we would be able to estimate their real values. We will see how in what follows.*

2.2 Continuous Random Variables

All the concept presented for a random variable X taking discrete values in a set E can be extended to the ones taking values in a continuous 1D set $\Omega \subseteq \mathbb{R}$. For instance, when we perform a length measurement, its values belongs to the interval $[0, +\infty)$. Similarly to what we did for discrete random variables with the probability function, we define the a *probability density function* (pdf) as follows:

$$f(x) := \lim_{\delta x \to 0} \mathbb{P}(x \leq X \leq x + \delta x),$$

since, in this case, the probability of the event $X = x$ (having zero measure in a 1D space) is zero. In this case, the properties of the pdf are:

$$f(x) \geq 0 \ \ \forall x \in \Omega,$$
$$\int_{x \in \Omega} f(x)dx = 1.$$

A definition similar to what has been provided with the cumulative function for discrete variables can be provided also in the countinuous case. When we want to evaluate the probabilities of intervals we might resort to the *Cumulative Distribution Function* (CDF), defined as:

$$F(x) := \int_{s \in \Omega, s \leq x} f(s)ds$$

having the following properties:

$$0 \leq F(x) \leq 1 \; \forall x \in \Omega,$$
$$F\left(\min_{x\in\Omega} x - \varepsilon\right) = 0,$$
$$F\left(\max_{x\in\Omega} x\right) = 1,$$

where $\varepsilon > 0$.

Similarly to the discrete case, the expected value and the variance are defined as:

$$\mu = \mathbb{E}[X] := \int_{x\in\Omega} x f(x) dx; \qquad \sigma^2 = Var(X) := \int_{x\in\Omega} (\mathbb{E}[X] - x)^2 f(x) dx.$$

Among the most used continuous distributions we have the Gaussian one $X \sim \mathcal{N}(\mu, \sigma^2)$ defined over $\Omega = \mathbb{R}$ and having:

$$f(x; \mu, \sigma) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$
$$F(x; \mu, \sigma) = \int_{-\infty}^{x} f(t; \mu, \sigma) dt,$$

and the uniform random variable $X \sim \mathcal{U}([0, 1])$ defined over $\Omega = [0; 1]$ and having:

$$f(x) = 1$$
$$F(x) = x.$$

2.3 Univariate Distributions in MATLAB

In `MATLAB` it is possible to define objects to handle different distributions with the instruction `makedist(distname)`, among the others:

- `'Normal'` Gaussian distribution (`'mu'` and `'sigma'` parameters);
- `'Uniform'` Uniform distribution (`'lower'` and `'upper'` parameters);
- `'Binomial'` Binomial distribution (`'N'` and `'p'` parameters);
- `'Beta'` Beta distribution (`'a'` and `'b'` parameters);

each of which will have specific parameters.

Remark 2. *Here the parameter `'sigma'` is the standard deviation. In some other `MATLAB` functions the second parameter of the Gaussian distribution is the variance. Check the documentation to know which one of the two is needed.*

For some of the aforementioned distributions there exist MATLAB functions which does not require to instantiate a random variable handle.

If we want to instantiate a random variable with Gaussian distribution with mean 3 and standard deviation 4 we write:

```
X = makedist('Normal','mu',3,'sigma',4);
```

We get the pdf of a given point of the random variable, find the point where the CDF function takes a specific value or sample from the distribution in the following way:

```
norm_pd.pdf(5); %pdf at x = 5
norm_pd.cdf(3); %CDF for x = 3
norm_pd.icdf(0.05); %inverse CDF for alpha = 0.05
```

alternatively:

```
normpdf(5,3,4); %pdf at x = 5
normcdf(3,3,4); %CDF for x = 3
norminv(0.05,3,4); %inverse CDF for alpha = 0.05
```

Another useful functionality available in MATLAB is the possibility to draw samples from a specific distribution. For instance, if we want to sample 100 realizations of the Gaussian variable X we previously defined we could do:

```
X.random(100,1)
```

or

```
normrnd(3,4,100,1)
```

2.4 Multivariate Distributions in MATLAB

There exists also some distributions which takes values in $\Omega \subseteq \mathbb{R}^n$ with $n \in \mathbb{N}$, $n > 1$. they are usually addressed as *multivariate distributions*. If we want to resort to such random variable in MATLAB we should use different functions. For instance for the multivariate Gaussian we have:

```
mvnpdf(X,MU,SIGMA);          %pdf
mvncdf(X,MU,SIGMA);          %CDF
mvnrnd(MU,SIGMA,n_samples)  %random sampling
```

where `X` are the points considered, `MU` is the mean vector, `SIGMA` is the covariance matrix, and `n_samples` is the number of instances to be sampled.

For instance, if we want to sample 100 points from a 5-variate normal distribution with mean $[1\ 1\ 1\ 1\ 1]^T$ with identity covariance matrix and plot these points, we can use:

```
rand_samples = mvnrnd(ones(1,5), eye(5), 100);
gplotmatrix(rand_samples);
```

2.5 Central Limit Theorem

In the previous sections we assumed that the distribution was known, i.e., that the parameters characterizing the distribution were known. Otherwise we built some *estimators* to infer them from data coming from the specified distribution. For instance, consider a set of N samples $\{x_1, \ldots, x_n\}$ coming from the same distribution, the consistent estimators for the expected value and for the (sample) variance are:

$$\bar{X} := \frac{\sum_{i=1}^{n} x_i}{n}$$
$$s^2 := \frac{\sum_{i=1}^{n} (\bar{X} - x_i)^2}{n-1},$$

respectively.[1]

Let us consider the empirical expected value $\bar{X}$. We recall the *Central Limit Theorem* (CLT):

Theorem 1 (Central Limit Theorem)**.** *Assume $\{X_1, \ldots X_n\}$ is a sequence of independent and identically distributed (i.i.d.) random variables with $\mathbb{E}[X_i] = \mu$ and $Var[X_i] = \sigma^2 < \infty$, then:*

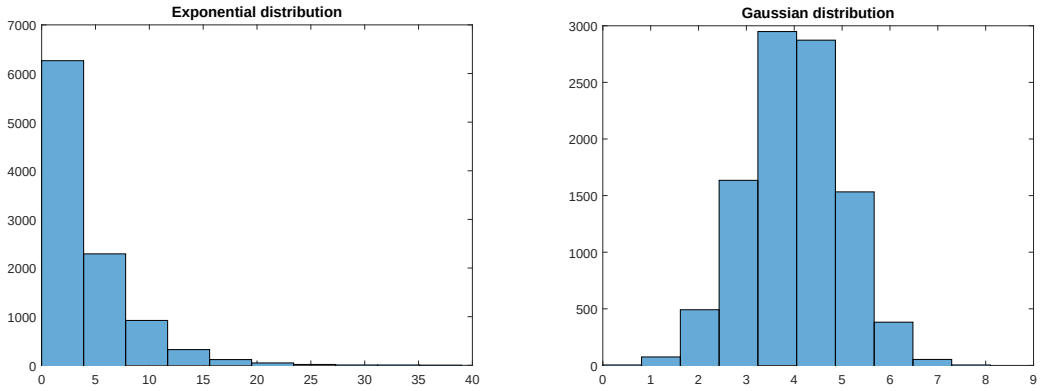
$$\sqrt{n}\left(\frac{\sum_{i=1}^{n} X_i}{n} - \mu\right) \to \mathcal{N}(0; \sigma^2),$$

where the convergence holds in distribution.

To better understand this concept let us sample from an *exponential* distribution with $\mu = 4$ and a *Gaussian* distribution with $\mu = 4$ and $\sigma = 1$. These are the histograms of the sampled distributions (we considered $n = 10000$ samples). The results are shown in Figure 2.1. Since the distributions have the same expected value, the empiric mean $\bar{X} = \frac{\sum_{i=1}^{n} X_i}{n}$ will concentrate around this value. If we repeatedly sample from these distributions and consider the empiric mean obtained at each repetition we will have

[1] The term *consistent* means that if we have infinite number of samples we would converge (in probability) to the true value of the parameter.

the results shown in Figure 2.2, which confirm that the distribution of the empirical mean is approximated by a Gaussian, as stated in the CLT. Moreover, we can see how the first one is more spread since the variance of the exponential distribution is 4 times the one of the Gaussian we considered.



(a) Exponential (b) Gaussian

Figure 2.1: Histograms of the samples coming from two different distributions.

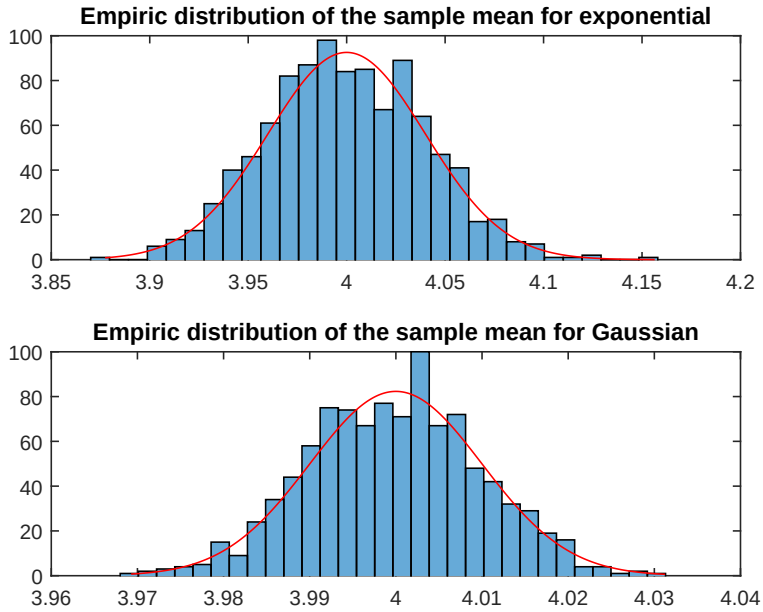


Figure 2.2: Histograms of the estimated means $\bar{X}$ coming from two different distributions.

2.6 Confidence Intervals

Once we have some estimates of the distribution parameters we would like to understand how much we should rely on them. For instance, if we used few data to estimate the true expected value $\mathbb{E}[X]$ it is likely that the true value might be far from the estimated one, while if we used N large enough we are more certain about the true value. Another factor determining how much we can rely on the estimator is the variance of the phenomenon itself: with high variance we will have samples which are more likely to be far from each other and, thus, we are less certain about the value of the expected value (as it was shown in the previous section). Since we are in a stochastic environment, we need to set a level identifying that our estimator is "good enough". The probability that $\mathbb{E}[X]$ is exactly $\bar{X}$ is zero since the expected realization is a continuous random variable itself. Thus, we need to built some intervals, where we have high confidence that the true mean $\mathbb{E}[X]$ is in.

Since we have the characterization of the distribution of the empirical mean, we can built intervals in which the expected value $\mathbb{E}[X]$ is with a specific confidence α. For instance, if we want to consider a confidence interval for the empirical mean with confidence at least $1-\alpha$:

$$\bar{X} - \frac{z_{\alpha/2}\sigma}{\sqrt{n}} \leq \mu \leq \bar{X} + \frac{z_{\alpha/2}\sigma}{\sqrt{n}} \tag{2.1}$$

where, $z_{\alpha/2}$ is the $1-\alpha/2$ quantile (i.e., the point where the CDF takes value $1-\alpha/2$ or $F^{-1}(1-\alpha/2)$) for a Normal distribution $\mathcal{N}(0,1)$, or equivalently:

$$\mathbb{P}\left(|\bar{X} - \mu| \geq \frac{z_{\alpha/2}\sigma}{\sqrt{n}}\right) \leq \alpha.$$

Remark 3. *The previous inequality is true only under Gaussian assumption. Since we are provided a finite number of samples, the confidence intervals are only an approximations of the real ones.*

Remark 4. *In the derivation of the confidence interval we assumed that the standard deviation σ was known. In the case we resort to its estimates s^2, there exist different confidence intervals for the mean of the Gaussian distribution.*

If we are considering unbounded domains (e.g., $\Omega = \mathbb{R}$ or $\Omega = \mathbb{R}^+$) we also might resort to the following inequality to design confidence bounds:

Theorem 2 (Chebichev Inequality)**.** *Suppose X is random variables with $\mathbb{E}[X] = \mu < \infty$ and $Var[X] = \sigma^2 < \infty$, then:*

$$\mathbb{P}\left(|\mu - X| \geq \frac{\sigma}{\sqrt{\alpha}}\right) \leq \alpha.$$

which, if we consider the empirical mean $\bar{X}$ and a symmetric bound, leads to:

$$\bar{X} - \frac{\sigma}{\sqrt{n}\sqrt{\alpha}} \leq \mu \leq \bar{X} + \frac{\sigma}{\sqrt{n}\sqrt{\alpha}} \tag{2.2}$$

Remark 5. *The Chebichev inequality is more general w.r.t. the one derived from CLT, since it can be applied to generic random variables. In fact, CLT holds asymptotically for every random variable and exactly for Gaussian ones.*

Finally, if we also assume that the random variables have have finite support (e.g., $\Omega = [a, b]$ or $\Omega = \{0, 1\}$), we might rely on:

Theorem 3 (Chernoff-Hoeffding bound)**.** *Assume to have a sequence $\{X_1, X_2 \dots X_n\}$ of n i.i.d. random variables with support in $[a, b]$ and $\mathbb{E}[X_i] = \mu, \ \forall i$, then for each $\varepsilon > 0$ we have:*

$$\mathbb{P}\left(|\bar{X} - \mu| \geq \varepsilon\right) \leq 2 \exp\left\{-\frac{2n\varepsilon^2}{(b-a)^2}\right\} = \alpha.$$

This statistical bound leads to the following confidence intervals with confidence at least $1 - \alpha$:

$$\bar{X} - (b-a)\sqrt{\frac{-\log(\alpha/2)}{2n}} \leq \mu \leq \bar{X} + (b-a)\sqrt{\frac{-\log(\alpha/2)}{2n}} \tag{2.3}$$

2.6.1 Hypothesis Testing

Sometimes we would like to make statement like:

The estimated parameter $\bar{X}$ is equal to μ.
The estimated parameter $\bar{X}'$ is different from $\bar{X}''$.

In the case stochastic quantities are involved, the answer to such questions is the *test of hypotheses*. More specifically, we need to specify a null hypothesis H_0 and an alternative hypothesis H_1, e.g.,:

$$H_0 \ : \ \mu = \mu_0 \quad vs. \quad H_1 \ : \ \mu \neq \mu_0.$$

Given the data $\{x_1, \dots, x_n\}$, we need that some of them support either the null or the alternative hypothesis. If we are able to compute an estimates for μ and we know its distribution, we are also able to say how likely is that the estimates has been drawn by the distribution. For instance, by considering the CLT we might say that:

$$\bar{X} \sim \mathcal{N}\left(\mu, \frac{\sigma^2}{n}\right) \rightarrow t = \frac{\bar{X} - \mu}{\frac{\sigma}{\sqrt{n}}} \sim \mathcal{N}(0, 1)$$

and thus that there is α probability that the test statistic t is lower than the quantile of order $\alpha/2$ of the standard Gaussian distribution or that it is larger than the quantile of order $1 - \alpha/2$, more formally

$$\mathbb{P}(t < z_{\alpha/2} \vee t > z_{1-\alpha/2}) = \alpha.$$

		Decision	
		Fail to reject H_0	reject H_0
True	H_0	Correct	Type I error
	H_1	Type II error	Correct

Table 2.1: Possible situations for a hypothesis test.

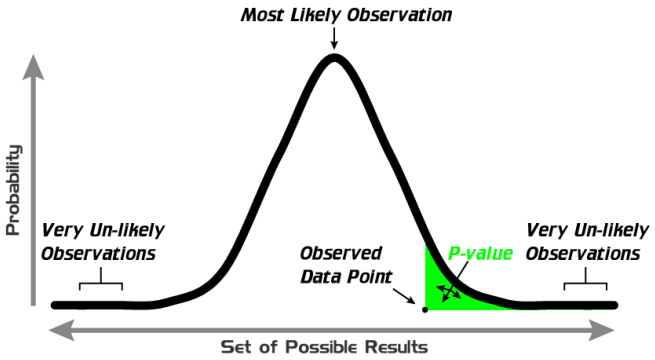


Figure 2.3: P-value in a right tailed test.

If the test statistic has absolute value greater than $z_{1-\alpha/2}$ or smaller than $z_{\alpha/2}$ there is small (α) probability that the data are coming from a distribution where H_0 holds. Nonetheless, if we sample data from distribution in the case H_0 holds repeatedly we have that α of the times the test will say that the data are not coming from H_0. One might think to use values of $\alpha \approx 0$ to reduce this probability and solve the problem. This would decrease the so called "type I error", but increase the "type II error", that even if the data are coming from a distribution for which H_1 holds we are not able to state that. Table 2.1 summarize all the possible situations.

The same reasoning used for the (two-tail) test could be used to develop test for hypothesis of the kind:

$$H_0 : \mu \leq \mu_0 \quad vs. \quad H_1 : \mu > \mu_0,$$
$$H_0 : \mu \geq \mu_0 \quad vs. \quad H_1 : \mu < \mu_0,$$

by considering only quantiles of order $1 - \alpha$ and α, respectively.

If we want to avoid to define a specific confidence α and let the data tell us how much we might be confident about their correspondence to a specific hypothesis, we could compute the p-value. The p-value is defined as the smallest confidence $\bar{\alpha}$ s.t. we are still able to reject the null hypothesis H_0. If we want to visualize the p-value in a right tailed hypothesis test (see Figure 2.3), it is the area under the distribution pdf and above the test statistics we computed

2.7 Frequentist vs. Bayesian Approach

The bounds we derived allow one to consider the cases in which we have only information about the bound of the variable which is considered. They do not allow to incorporate in a straightforward way information about the data distribution. In the case we have further information we might resort to a Bayesian approach for the parameter estimation. Indeed, by adopting the Bayesian framework, the value of the expected value of the random variable μ is a random variable itself. This is particularly interesting if we have information coming from the domain or from previously observed data.

For instance, let us say that we are considering a Bernoulli variable and we have some information coming from the past that tells us that a previously analysed phenomenon, **similar** to the one in analysis, had 3 successes over 10 trials. It would be wrong to consider these samples as drawn from the considered variable (i.e., using them to compute the empirical mean).

Consider the Bayes formula:

$$\begin{aligned}\mathbb{P}(\mu|x_1,\dots,x_t) &= \frac{\mathbb{P}(x_1,\dots,x_{t-1},x_t|\mu)\mathbb{P}(\mu)}{\mathbb{P}(x_1,\dots,x_t)}\\ &\propto \mathbb{P}(x_t|\mu)\mathbb{P}(x_1,\dots,x_{t-1}|\mu) \propto \mathbb{P}(x_t|\mu)\mathbb{P}(\mu|x_1,\dots,x_{t-1})\\ &\propto \mathbb{P}(\mu)\prod_{h=1}^{t}\mathbb{P}(x_h|\mu),\end{aligned}$$

where we assumed conditional independence of x_t from all the other data.[2] This way we are able to incorporate information starting from a prior distribution $\mathbb{P}(\mu)$ incrementally.

In the case we consider Bernoulli realizations, if we consider a Beta distribution as prior for the expected value μ ($\mu \sim \text{Beta}(3,7)$ in the example), we have that the posterior is still a Beta (i.e., Bernoulli and Beta are conjugate prior-posterior), thus allowing us to have an update rule for the next datum x_t.

Remark 6. *In the case we incorporate meaningful information the Bayesian learning process could be faster than the frequentist one. Though, if the information provided by the prior are misleading, the process could slow down and in some case prevent the estimation process to converge to the real value (e.g., if the prior assigns zero probability to the real value of the parameter).*

[2] We do not report the denominators since they do not change the idea of the Bayes update.

2.8 Optimization

One of the main element of *Machine Learning* techniques consists in the method used to optimize a given function $J(\cdot)$ over the available data. If we want to optimize a known loss function we should at first analyse type of objective function we want to solve and the constraints we want to satisfy. There are some situations where there exists a closed form solution for the optimal solution, thus one can use the analytical solution. For instance, if we are given a problem of the form:

$$\min_{\theta} J(\theta, X, y) = \frac{1}{2}||\theta^T X - y||_2^2,$$

where $\theta = (\theta_1, \dots, \theta_c)^T$, $\theta \in \mathbb{R}^c$, $X \in \mathbb{R}^{c \times n}$ and $y \in \mathbb{R}^n$ and $||\mathbf{a}||_2 := \sqrt{\sum_{i=1}^n a_i^2}$, we might resort to the *Least Square Method*. In fact the closed form solution of the problem is:

$$\theta := (X^T X)^{-1} X^T y,$$

where we are required to invert a matrix of order $c \times c$.

In the case a closed form solution is not available, it is possible to use other methods which exploit the properties of the function we are using. For instance for convex functions we might resort to the *Gradient Descent Method*. the idea behind this method is that we want to update an initial guess of the parameter $\theta^{(0)}$ by following the direction where the function $J(\cdot, X, y)$ decreases the most. By starting from an initial parameter vector $\theta^{(0)}$ we update it, by looking at the incoming data, more formally:

$$\theta^{(k+1)} = \theta^{(k)} - \alpha \left. \frac{\partial J(\theta, X, y)}{\partial \theta} \right|_{\theta = \theta^{(k)}}.$$

Finally, if you need to solve more complex optimization problems, we might resort to some advanced techniques which have been already implemented in the *Optimization toolbox* (Apps tab $\rightarrow$ Optimization) in `MATLAB` . This toolbox provides you different functions able to deal with different types of problem:

- Scalar bounded minimization (`fminbnd`):

$$\min_{\theta} J(\theta, X, y)$$
$$\text{s.t. } lb < \theta < ub$$

- Unconstrained minimization (`fminsearch`):

$$\min_{\theta} J(\theta, X, y)$$

- Linear programming (`linprog`):

$$\begin{aligned}&\min_{\theta} J(X,y)^T\theta\\ &\text{s.t. } A\,\theta \leq b\\ &Aeq\,\theta = beq\\ &lb \leq \theta \leq ub\end{aligned}$$

- Quadratic programming (`quadprog`):

$$\begin{aligned}&\min_{x} \frac{1}{2}\theta^T H(X,y)\theta + J(X,y)^T \cdot \theta\\ &\text{s.t. } A\,\theta \leq b\\ &Aeq\,\theta = beq\\ &lb \leq \theta \leq ub\end{aligned}$$

- Constrained minimization (`fmincon`):

$$\begin{aligned}&\min_{\theta} J(\theta,X,y)\\ &\text{s.t. } A\,\theta \leq b\\ &Aeq\,\theta = beq\\ &lb \leq \theta \leq ub\end{aligned}$$

They usually require to provide a feasible solution $\theta^{(0)}$ as starting point of the minimization process.

Remark 7. *The use of the previously shown minimization tools requires that the considered functions (both objective and constraints) have some sort of regularity. In the case we are considering highly variable functions or even discontinuous ones, we should resort to different techniques.*

Exercise 2.1

Model a fair 6-faced dice. Then, find the best strategy for playing dice (you bet on the sum of a couple of dice and win only if the predicted number is equal to the result).

What happens if the dice has 20 faces.

Exercise 2.2

Consider a simplified version of the game of the roulette in which you can only bet on specific numbers and you win 35 times what you bet if you guessed right and 0 if you guessed wrong.

What is the expected value of playing 1 euro at the roulette? What about the variability of this phenomenon?

Write a script emulating a player, with and initial budget of 100 euros that always bets 1 euro per turn on the number 17, and record the amount of money she/he has over time. Plot the amount of money in her/his wallet with a red line. Are you sure that the experiment ends at eventually?

Exercise 2.3

Given a set of $n = 1000$ realization from a Bernoulli distribution with $\mu = 0.5$ compute the following bounds for the expected value with confidence $\alpha = 0.01$:

- confidence interval from CLT;
- Hoeffding bounds;
- Chebichev bounds.

There is any theoretical difference among them?