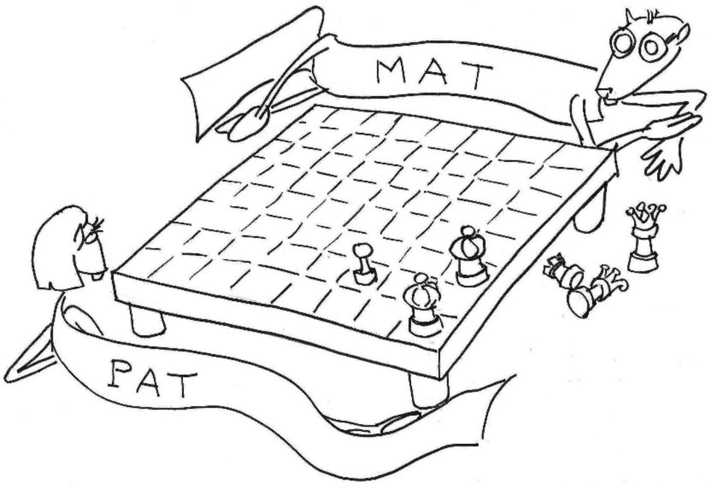
Part I

ANALYSIS

CHAPTER 1

Probabilistic Models



In this book we study pattern matching problems in a probabilistic framework. We first introduce some general probabilistic models of generating sequences. The reader is also referred to Alon and Spencer (1992), Reinert, Schbath, and Waterman (2000), Szpankowski (2001) and Waterman (1995a) for a brief introduction to probabilistic models. For the convenience of the reader, we recall here some definitions.

We also briefly discuss some analytic tools such as generating functions, the residue theorem, and the Cauchy coefficient formula. For in-depth discussions the reader is referred to Flajolet and Sedgewick (2009) and Szpankowski (2001).

1.1. Probabilistic models on words

Throughout we shall deal with sequences of discrete random variables. We write $(X_k)_{k=1}^{\infty}$ for a one-sided infinite sequence of random variables; however, we will often abbreviate it as X provided that it is clear from the context that we are talking about a sequence, not a single variable. We assume that the sequence $(X_k)_{k=1}^{\infty}$ is defined over a finite alphabet $\mathcal{A} = \{a_1, \ldots, a_V\}$ of size V. A partial sequence is denoted as $X_m^n = (X_m, \ldots, X_n)$ for $m < n$. Finally, we shall always assume that a probability measure exists, and we will write $P(x_1^n) = P(X_k = x_k,\ 1 \le k \le n,\ x_k \in \mathcal{A})$ for the probability mass, where we use lower-case letters for a realization of a stochastic process.

Sequences are generated by information sources, usually satisfying some constraints. We also refer to such sources as *probabilistic models*. Throughout, we assume the existence of a stationary probability distribution; that is, for any string w we assume that the probability that the text X contains an occurrence of w at position k is equal to $P(w)$ independently of the position k. For $P(w) > 0$, we denote by $P(u|w)$ the conditional probability, which equals $P(wu)/P(w)$.

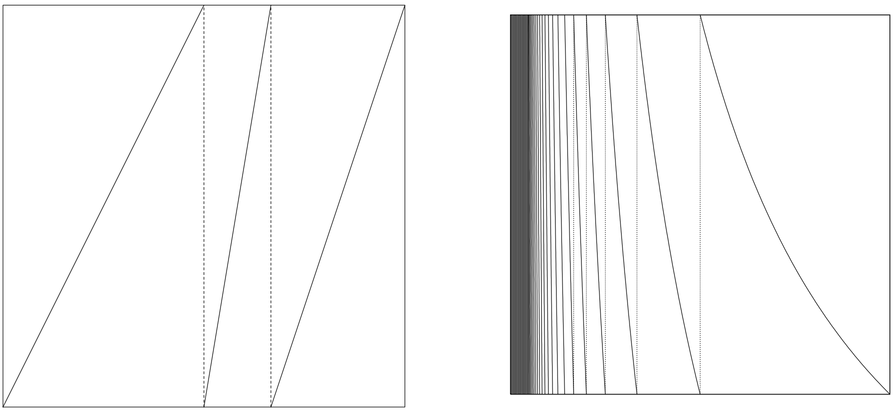
The most elementary information source is a *memoryless source*; also known as a *Bernoulli source*:

(B) Memoryless or Bernoulli Source. The symbols from the alphabet $\mathcal{A} = \{a_1, \ldots, a_V\}$ occur independently of one another; thus the string $X = X_1X_2X_3\cdots$ can be described as the outcome of an infinite sequence of Bernoulli trials in which $P(X_j = a_i) = p_i$ and $\sum_{i=1}^{V} p_i = 1$. Throughout, we assume that at least for one i we have $0 < p_i < 1$.

In many cases, assumption (B) is not very realistic. When this is the case, assumption (B) may be replaced by:

(M) Markov source. There is a Markov dependency between the consecutive symbols in a string; that is, the probability $p_{ij} = P(X_{k+1} = a_j | X_k = a_i)$ describes the conditional probability of sampling symbol a_j immediately after symbol a_i. We denote by $\mathrm{P} = \{p_{ij}\}_{i,j=1}^{V}$ transition matrix and by $\boldsymbol{\pi} = (\pi_1, \ldots, \pi_V)$ the stationary row vector satisfying $\boldsymbol{\pi}\mathrm{P} = \boldsymbol{\pi}$. (Throughout, we assume that the Markov chain is irreducible and aperiodic.) A general Markov source of order r is characterized by the $V^r \times V$ the transition matrix with coefficients $P(j \in \mathcal{A} \mid u)$ for $u \in \mathcal{A}^r$.

In some situations more general sources must be considered (for which one still can obtain a reasonably precise analysis). Recently, Vallée (2001) introduced *dynamic sources*, which we briefly describe here and will use in the analysis of the generalized subsequence problem in Section 5.6. To introduce such sources we start with a description of a *dynamic system* defined by:



(a)

(b)

Figure 1.1. Plots of $\mathcal{I}_m$ versus x for the dynamic sources discussed in Example 1.1.1: (a) a memoryless source with shift mapping $T_m(x) = (x - q_m)/p_{m+1}$ for $p_1 = 1/2$, $p_2 = 1/6$, and $p_3 = 1/3$; (b) a continued fraction source with $T_m(x) = 1/x - m = \langle 1/x \rangle$.

- a topological partition of the unit interval $\mathcal{I} = (0, 1)$ into a disjoint set of open intervals $\mathcal{I}_a, a \in \mathcal{A}$;
- an encoding mapping e which is constant and equal to $a \in \mathcal{A}$ on each $\mathcal{I}_a$;
- a shift mapping $T : \mathcal{I} \to \mathcal{I}$ whose restriction to $\mathcal{I}_a$ is a bijection of class $\mathcal{C}^2$ from $\mathcal{I}_a$ to $\mathcal{I}$; the local inverse of T restricted to $\mathcal{I}_a$ is denoted by h_a.

Observe that such a dynamic system produces infinite words of $\mathcal{A}^\infty$ through the encoding e. For an initial $x \in \mathcal{I}$ the source outputs a word, say $w(x) = (e(x), (e(T(x)), \ldots)$.

(DS) PROBABILISTIC DYNAMIC SOURCE. A source is called a *probabilistic dynamic source* if the unit interval of a dynamic system is endowed with a probability density f.

Example 1.1.1. A memoryless source associated with the probability distribution $\{p_i\}_{i=1}^{V}$, where V can be finite or infinite, is modeled by a dynamic source in which the components $w_k(x)$ are independent and the corresponding topological partition of $\mathcal{I}$ is defined as follows:

$$\mathcal{I}_m := (q_m, q_{m+1}], \qquad q_m = \sum_{j<m} p_j.$$

In this case the shift mapping restricted to $\mathcal{I}_m$, $m \in \mathcal{A}$, is defined as

$$T_m = \frac{x - q_m}{p_{m+1}}, \qquad q_m < x \leq q_{m+1}.$$

In particular, a symmetric V-ary memoryless source can be described by

$$T(x) = \langle Vx \rangle, \qquad e(x) = \lfloor Vx \rfloor,$$

where $\lfloor x \rfloor$ is the integer part of x and $\langle x \rangle = x - \lfloor x \rfloor$ is the fractional part of x. In Figure 1.1(a) we assume $\mathcal{A} = \{1, 2, 3\}$ and $p_1 = 1/2$, $p_2 = 1/6$, and $p_3 = 1/3$, so that $q_1 = 0$, $q_1 = 1/2$, $q_2 = 2/3$, and $q_4 = 1$. Then

$$T_1 = 2x, \qquad T_2 = 6(x - 1/2), \qquad T_3 = 3(x - 2/3).$$

For example, if $x = 5/6$ then

$$w(x) = (e(x), e(T_3(x)), e(T_2(T_3(x))), \ldots) = (3, 2, 1, \ldots).$$

Here is another example of a source with a memory related to continued fractions. The alphabet $\mathcal{A}$ is the set of all natural numbers and the partition of $\mathcal{I}$ is defined as $\mathcal{I}_m = (1/(m+1), 1/m)$. The restriction of T to $\mathcal{I}_m$ is the decreasing linear fractional transformation $T(x) = 1/x - m$, that is,

$$T(x) = \langle 1/x \rangle, \qquad e(x) = \lfloor 1/x \rfloor.$$

Observe that the inverse branches h_m of the mapping $T(x)$ are defined as $h_m(x) = 1/(x + m)$ (see Figure 1.1(b)). ■

Let us observe that a word of length k, say $w = w_1 w_2 \ldots w_k$, is associated with the mapping $h_w := h_{w_1} \circ h_{w_2} \circ \cdots \circ h_{w_k}$, which is an inverse branch of T^k. In fact all words that begin with the same prefix w belong to the same *fundamental interval*, defined as $\mathcal{I}_w = (h_w(0), h_w(1))$. Furthermore, for probabilistic dynamic sources with density f one easily computes the probability of w as the measure of the interval $\mathcal{I}_w$.

The probability $P(w)$ of a word w can be explicitly computed through a special *generating operator* $\mathbf{G}_w$, defined as follows

$$\mathbf{G}_w[f](t) := |h'_w(t)| f \circ h_w(t). \qquad (1.1)$$

One recognizes in $\mathbf{G}_w[f](t)$ a density mapping, that is, $\mathbf{G}_w[f](t)$ is the density of f mapped over $h_w(t)$. The probability of w can then be computed as

$$P(w) = \left| \int_{h_w(0)}^{h_w(1)} f(t)dt \right| = \int_0^1 |h'_w(t)| f \circ h_w(t)dt = \int_0^1 \mathbf{G}_w[f](t)dt. \qquad (1.2)$$

Let us now consider a concatenation of two words w and u. For memoryless sources $P(wu) = P(w)P(u)$. For Markov sources one still obtains the product of the *conditional* probabilities. For dynamic sources the product of probabilities is replaced by the product (composition) of the generating operators. To see this, we observe that

$$\mathbf{G}_{wu} = \mathbf{G}_u \circ \mathbf{G}_w, \qquad (1.3)$$

where we write $\mathbf{G}_w := \mathbf{G}_w[f](t)$. Indeed, $h_{wu} = h_w \circ h_u$ and

$$\mathbf{G}_{wu} = h'_w \circ h_u h'_u f \circ h_w \circ h_u$$

while $\mathbf{G}_w = h'_w f \circ h_w$ and so

$$\mathbf{G}_u \circ \mathbf{G}_w = h'_u h'_w \circ h_u f \circ h_w \circ h_u,$$

as desired.

Another generalization of Markov sources, namely the *mixing source*, is very useful in practice, especially for dealing with problems of data compression or molecular biology when one expects long(er) dependency between the symbols of a string.

(MX) (STRONGLY) ψ-MIXING SOURCE. Let $\mathcal{F}_m^n$ be a σ-field generated by the sequence $(X_k)_{k=m}^n$ for $m \le n$. The source is called *mixing* if there exists a bounded function $\psi(g)$ such that, for all $m, g \ge 1$ and any two elements of the σ-field (events) $A \in \mathcal{F}_1^m$ and $B \in \mathcal{F}_{m+g}^\infty$, the following holds:

$$(1 - \psi(g))P(A)P(B) \le P(AB) \le (1 + \psi(g))P(A)P(B). \qquad (1.4)$$

If, in addition, $\lim_{g\to\infty} \psi(g) = 0$ then the source is called *strongly* mixing.

In words, the model (MX) postulates that the dependency between $(X_k)_{k=1}^m$ and $(X_k)_{k=m+g}^\infty$ gets weaker and weaker as g becomes larger (note that when the sequence (X_k) is independent and identically distributed (i.i.d.) we have $P(AB) = P(A)P(B)$). The degree of dependency is characterized by $\psi(g)$. A weaker mixing condition, namely ϕ-*mixing*, is defined as follows:

$$-\phi(g) \le P(B|A) - P(B) \le \phi(g), \qquad P(A) > 0, \qquad (1.5)$$

provided that $\phi(g) \to 0$ as $g \to \infty$. In general, strong ψ-mixing implies the ϕ-mixing condition but not vice versa.

1.2. Probabilistic tools

In this section, we briefly review some probabilistic tools used throughout the book. We will concentrate on the different types of *stochastic convergence.*

The first type of convergence of a sequence of random variables is known as *convergence in probability.* The sequence X_n converges to a random variable X in probability, denoted $X_n \to X$ (pr.) or $X_n \overset{pr}{\to} X$, if, for any $\varepsilon > 0$,

$$\lim_{n\to\infty} P\left(|X_n - X| < \varepsilon\right) = 1.$$

It is known that *if* $X_n \overset{pr}{\to} X$ *then* $f(X_n) \overset{pr}{\to} f(X)$ *provided that* f *is a continuous function* (see Billingsley (1968)).

Note that convergence in probability does not say that the difference between X_n and X becomes very small. What converges here is the *probability* that the difference between X_n and X becomes very small. It is, therefore, possible, although unlikely, for X_n and X to differ by a significant amount and for such differences to occur infinitely often. A stronger kind of convergence that does not allow such behavior is *almost sure convergence* or *strong convergence.* This convergence ensures that *the set of sample points for which* X_n *does not converge to* X *has probability zero.* In other words, a sequence of random variables X_n converges to a random variable X almost surely, denoted $X_n \to X$ a.s. or $X_n \overset{(a.s.)}{\to} X$, if, for any $\varepsilon > 0$,

$$\lim_{N\to\infty} P(\sup_{n\geq N} |X_n - X| < \varepsilon) = 1.$$

From this formulation of almost sure convergence, it is clear that if $X_n \to X$ (a.s.), the probability of infinitely many large differences between X_n and X is zero. As the term "strong" implies, almost sure convergence implies convergence in probability.

A simple sufficient condition for almost sure convergence can be inferred from the *Borel–Cantelli lemma*, presented below.

Lemma 1.2.1 (Borel–Cantelli)**.** *If* $\sum_{n=0}^{\infty} P(|X_n - X| > \varepsilon) < \infty$ *for every* $\varepsilon > 0$ *then* $X_n \overset{a.s.}{\to} X$*.*

Proof. This follows directly from the following chain of relationships:

$$P\left(\sup_{n\geq N} |X_n - X| \geq \varepsilon\right) = P\left(\bigcup_{n\geq N} |X_n - X| \geq \varepsilon\right) \leq \sum_{n\geq N} P(|X_n - X| \geq \varepsilon) \to 0.$$

The inequality above is a consequence of the fact that the probability of a union of events is smaller than the sum of the probability of the events (see Boole's or the union inequality below). The last convergence is a consequence of our assumption that $\sum_{n=0}^{\infty} P(|X_n - X| > \varepsilon) < \infty$. ■

A third type of convergence is defined on distribution functions $F_n(x)$. The sequence of random variables X_n *converges in distribution* or *converges in law* to the random variable X, this convergence being denoted as $X_n \overset{d}{\to} X$, if

$$\lim_{n\to\infty} F_n(x) = F(x) \tag{1.6}$$

for each point of continuity of $F(x)$. In Billingsley (1968) it was proved that the above definition is equivalent to the following: $X_n \overset{d}{\to} X$ *if*

$$\lim_{n\to\infty} \mathbf{E}[f(X_n)] = \mathbf{E}[f(X)] \tag{1.7}$$

for all bounded continuous functions f.

The next type of convergence is *convergence in mean of order p* or *convergence in L^p*, which postulates that $\mathbf{E}[|X_n - X|^p] \to 0$ as $n \to \infty$. We write this type of convergence as $X_n \overset{L^p}{\to} X$. Finally, we introduce *convergence in moments* for which $\lim_{n\to\infty} \mathbf{E}[X_n^p] = \mathbf{E}[X^p]$ for any $p \geq 1$.

We now describe the relationships (implications) between the various types of convergence. The reader is referred to Billingsley (1968) for a proof.

Theorem 1.2.2. *We have the following implications:*

$$X_n \overset{a.s.}{\to} X \quad \Rightarrow \quad X_n \overset{pr}{\to} X, \tag{1.8}$$

$$X_n \overset{L^p}{\to} X \quad \Rightarrow \quad X_n \overset{pr}{\to} X, \tag{1.9}$$

$$X_n \overset{pr}{\to} X \quad \Rightarrow \quad X_n \overset{d}{\to} X, \tag{1.10}$$

$$X_n \overset{L^p}{\to} X \quad \Rightarrow \quad \mathbf{E}[X_n^p] \to \mathbf{E}[X^p]. \tag{1.11}$$

No other implications hold in general.

It is easy to devise an example showing that convergence in probability does not imply convergence in mean (e.g., take $X_n = n$ with probability $1/n$ and $X_n = 0$ with probability $1 - 1/n$). To obtain convergence in mean from convergence in probability one needs somewhat stronger conditions. For example, if $|X_n| \leq Y$ and $\mathbf{E}[Y] < \infty$ then, by the *dominated convergence theorem*, we know that convergence in probability implies convergence in mean. To generalize, one introduces so-called uniform integrability. *It is said that a sequence* $\{X_n,\ n \geq 1\}$ *is uniformly integrable if*

$$\sup_{n\geq 1} \mathbf{E}\left[|X_n| I(|X_n| > a)\right] \to 0 \tag{1.12}$$

when $a \to \infty$; here $I(B)$ is the indicator function, equal to 1 if A holds and 0 otherwise. The above is equivalent to

$$\lim_{a\to\infty} \sup_{n\geq 1} \int_{|x|>a} x dF_n(x) = 0.$$

Then the following is true, as shown in Billingsley (1968): *if* X_n *is uniformly integrable then* $X_n \overset{pr}{\to} X$ *implies that* $X_n \overset{L^1}{\to} X$.

In the probabilistic analysis of algorithms, inequalities are very useful for establishing these stochastic convergences. We review now some inequalities used in the book.

Boole's or the union inequality. For any set of events $A_1, \ldots, A_n$ the following is true:

$$P(A_1 \cup A_2 \cup \cdots \cup A_n) \leq P(A_1) + P(A_2) + \cdots + P(A_n). \tag{1.13}$$

The proof follows from iterative applications of $P(A_n \cup A_2) \leq P(A_1) + P(A_2)$, which is obvious.

Markov's inequality. For a nonnegative function $g(\cdot)$ and a random variable X,

$$P(g(X) \geq t) \leq \frac{\mathbf{E}[g(X)]}{t} \tag{1.14}$$

holds for any $t > 0$. Indeed, we have the following chain of obvious inequalities:

$$\mathbf{E}[g(X)] \geq \mathbf{E}\left[g(X) I(g(X) \geq t)\right] \geq t\mathbf{E}[I(g(X) \geq t)] = tP(g(X) \geq t),$$

where we recall that $I(A)$ is the indicator function of event A.

Chebyshev's inequality. If one replaces $g(X)$ by $|X - \mathbf{E}[X]|^2$ and t by t^2 in Markov's inequality then we have

$$P(|X - \mathbf{E}[X]| \geq t) \leq \frac{\mathrm{Var}[X]}{t^2}, \tag{1.15}$$

which is known as Chebyshev's inequality.

Schwarz's inequality (also called the **Cauchy–Schwarz** inequality). Let X and Y be such that $\mathbf{E}[X^2] < \infty$ and $\mathbf{E}[Y^2] < \infty$. Then

$$\mathbf{E}[|XY|]^2 \leq \mathbf{E}[X^2]\mathbf{E}[Y^2]\ , \tag{1.16}$$