Probability and computation *

Luděk Kučera

April 10, 2001

1 Tools

1.1 Basic definitions

Definition 1.1.1 *A* probability space *is a nonempty set X together with a function $\pi : X \to [0,1]$ such that the set $\{x \in X \mid \pi(x) \neq 0\}$ is finite, and*

$$\sum_{x \in X} \pi(x) = 1.$$

A uniform *probability space over a nonempty set X is a space (X, π), where $\pi(x) = |X|^{-1}$ for each $x \in X$.*

The definition is much weaker than the usual definition of a probability space because of the condition of finiteness of carrier of the probability function π, but quite sufficient for our purposes. The advantage of the finiteness assumption is that we can avoid most difficult (though rather technical) parts of the theory of probability. What remains is essentially the art of counting, based on a very limited number of probability theory theorems.

The class of real (natural, resp.) numbers will be denoted by $\mathcal{R}$ (N, resp.).

Definition 1.1.2 *Given a probability space $S = (X, \pi)$, a* random variable *on S is a real-valued function $f : X \to \mathcal{R}$.*

---

*Lecture held at Universität des Saarlandes, Saarbrücken in winter semester 1992/93

The k-th moment *of* f *is the sum*

$$\sum_{x \in X} f^k(x)\pi(x).$$

The first moment of f *is usually denoted by* $E(f)$ *and called the* expectation *of* f*.*

Note that the k-th moment of f is the expectation $E(f^k)$ of the k-th power of f. It is also clear that

Lemma 1.1.3 *If* c *is a real constant, and* f, C *are random variables such that* $C(x) = c$ *for each* x*, then*

$$E(cf) = cE(f), \qquad E(C) = c.$$

Definition 1.1.4 *A* predicate *on a probability space* $S = (X, \pi)$ *is a random variable with values in the set* $\{0, 1\}$*.*

A predicate P on (X, π) is usually viewed as a certain property, that is verified for $x \in X$ such that $P(x) = 1$. E.g. if X is a collection of all graphs with a given n element vertex set, P might be equal to 1 for those graphs that can be colored by 4 colors. Values $0, 1$ are often identified with boolean values *false, true*, and therefore we can say that $P(x)$ holds if $P(x) = 1$, and $P(x)$ is not verified if $P(x) = 0$.

Instead of speaking about the expectation of predicate P, we use words *probability of* P, i.e. if P is a predicate on (X, π), then

$$\mathbf{Prob}(P) = E(P) = \sum_{x \in X} P(x)\pi(x) = \sum_{P(x)=1} \pi(x).$$

Definition 1.1.5 *We say that random variables* $f_1, \ldots, f_n$ *defined on a probability space* $S = (X, \pi)$ *are* independent*, if*

$$E(f_1 f_2 \ldots f_n) = E(f_1)E(f_2) \ldots E(f_n).$$

Definition 1.1.6 *We say that a probability space* (Y, ρ) *is a subspace of a probability space* (X, π) *if* $Y \subset X$*,* $\sum_{y \in Y} \pi(y) > 0$*, and*

$$\rho(y) = \pi(y) / \sum_{y \in Y} \pi(y) \qquad \textit{for each } y \in Y.$$

Note that a subspace (Y, ρ) of (X, π) is uniquely determined by the set Y.

Essentially a subspace of (X, π) is obtained by taking a subset Y of X, and a restriction $\overline{\rho}$ of π to Y. However, since the sum of probabilities of all elements of a probability space should be 1, the function $\overline{\rho}$ must be normalized by dividing by $\sum_{y \in Y} \pi(y)$ (which also implies the condition in the definition).

Definition 1.1.7 *Let P, Q be two predicates on a probability space (X, π) such that $\mathbf{Prob}(Q) > 0$. The (conditional) probability of P, conditioned on Q, denoted as $\mathbf{Prob}(P|Q)$, is defined as*

$$\mathbf{Prob}(P|Q) = \frac{\mathbf{Prob}(PQ)}{\mathbf{Prob}(Q)}.$$

Note that the conditional probability $\mathbf{Prob}(P|Q)$ is a probability of the restriction of P to the probability space that is a subspace of (X, π), determined by those elements of X that $Q(x)$ is verified.

Definition 1.1.8 *Let $S_1 = (X_1, \pi_1), \ldots, S_n = (X_n, \pi_n)$ be probability spaces. Their product $S_1 \ldots S_n$ is a probability space (X, π), where $X = X_1 \times \ldots \times X_n$, and $\pi((x_1, \ldots, x_n)) = \pi_1(x_1) \ldots \pi_n(x_n)$ for each element $(x_1, \ldots, x_n) \in X_1 \times \ldots \times X_n$.*

Note that random variables σ_i, $i = 1, \ldots, n$, defined on the product of $(X_1, \pi_1), \ldots, (X_n, \pi_n)$ by $\sigma_i((x_1, \ldots, x_n)) = \pi_i(x_i)$ are independent random variables.

1.2 Inequalities of Markov and Chebychev

In this chapter we prove three simple but very important theorems.

Theorem 1.2.1 *Let $f_1, \ldots, f_n$ be random variables on a probability space $S = (X, \pi)$. Then*

$$E(f_1 + \cdots + f_n) = E(f_1) + \cdots + E(f_n).$$

Remark: It is very important that the theorem requires no conditions on random variables $f_1, \ldots, f_n$, e.g. their independence.

Proof:

$$E(f_1(x)+\cdots+f_n(x)) = \sum_{x\in X}(f_1+\cdots+f_n)\pi(x) = \sum_{x\in X}(f_1\pi(x)+\cdots+f_n\pi(x)) =$$

$$= \sum_{x\in X} f_1\pi(x) + \cdots + \sum_{x\in X} f_n\pi(x) = E(f_1) + \cdots + E(f_n).$$

♣

Theorem 1.2.2 [Markov] *Let f be a non-negative integer valued random variable on a probability space $S = (X, \pi)$. Then*

$$\mathbf{Prob}(f \neq 0) \leq E(f).$$

Proof:

$$\mathbf{Prob}(f \neq 0) = \sum_{f(x)\neq 0} \pi(x) = \sum_{f(x)\geq 1} \pi(x) \leq$$

$$\leq \sum_{f(x)\geq 1} f(x)\pi(x) = \sum_{x\in X} f(x)\pi(x) = E(f).$$

♣

Theorem 1.2.3 [Chebychev] *Let f be a non-negative integer valued random variable on a probability space $S = (X, \pi)$, $t > 0$ be a real number. Then*

$$\mathbf{Prob}(|f(x) - E(f)| \geq t) \leq \frac{E(f^2) - E^2(f)}{t^2}.$$

Proof: Given $x \in X$, define $\varphi(x) = 1$ if $|f(x) - E(f)| \geq t$, $\varphi(x) = 0$ otherwise, $\psi(x) = t^{-2}(f(x) - E(f))^2$. Clearly $\varphi \leq \psi$.

$$\mathbf{Prob}(|f(x) - E(f)| \geq t) = E(\varphi) \leq E(\psi) = \frac{E(f^2) - 2E(fE(f)) + E(E^2(f))}{t^2} =$$

$$= \frac{E(f^2) - 2E(f)E(f) + E^2(f)}{t^2} = \frac{E(f^2) - E^2(f)}{t^2}.$$

♣

Corollary 1.2.4

$$\mathbf{Prob}(f=0) \le \frac{E(f^2)-E^2(f)}{E^2(f)}$$

Proof: It is sufficient to realize that $f(x)=0$ implies $|f(x)-E(f)| \ge E(f)$, and put $t=E(f)$. ♣

We see that knowledge of the first two moments of $f(x)$ makes it possible to give an upper and a lower bound to the probability that $f(x)$ is not equal to 0. This is especially important if $f(x)$ is the number of some objects, because in this case $f(x) \ne 0$ means that such an object exists.

We give an example that gives a standard way of computing moments in many practical situations.

Example

Let G_n be the collection of all graphs with the n-element vertex set $\{1,\ldots,n\}$, $\mathcal{G}_n$ be the uniform probability space over G_n. We want to know how large cliques are likely to appear in graphs in $\mathcal{G}_n$.

Given a graph $G \in G_n$ and a non-negative integer k, denote by $Y_k(G)$ the number of different cliques with k elements in the graph G. Therefore Y_k is a random variable on $\mathcal{G}_n$, and $Y_k(G) \ne 0$ means that the graph G posses a clique of size k.

Let us first compute the first moment of Y_k: Denote $\{1,\ldots,n\}$ by X, the set of k-element subsets of X by $\binom{X}{k}$, if $A \in \binom{X}{k}$ then $Y_A(G)$ denotes the number of cliques of G that are equal to A. The last definition is tricky, the only clique that is equal to A might be A only, i.e. $Y_A=1$ if A is a clique, and $Y_A=0$ if A is not a clique.

Note that

$$Y_k = \sum_{A \in \binom{X}{k}} Y_A,$$

and therefore the expectation of Y_k is the sum of expectations of predicates Y_A. What is the value of $\mathbf{Prob}(Y_A)$? Since our probability space is uniform, it is sufficient to determine the ratio of the number of graphs G such that A is a clique of G (i.e. $Y_A(G)=1$) and the number of all graphs with the vertex set X.

The number of all graphs is $2^{\binom{n}{2}}$, because there are $\binom{n}{2}$ pairs of vertices that might be edges. However, if we suppose that A is a clique, then $\binom{k}{2}$

pairs that are contained in A must be edges, and we have free choice for $\binom{n}{2} - \binom{k}{2}$ pairs only, therefore $Y_A(G)$ is 1 for $2^{\binom{n}{2}-\binom{k}{2}}$ graphs, which implies that

$$\mathbf{Prob}(Y_A = 1) = 2^{-\binom{k}{2}},$$

and therefore

$$E(Y_k) = \binom{n}{k} 2^{-\binom{k}{2}}.$$

We will give more detailed analysis of the last equation later, but now we mention that if $k = 1 + 2(1+\varepsilon)\log_2 n$, where $\varepsilon > 0$ is a constant, then it follows from Markov inequality that

$$\mathbf{Prob}(\mathcal{G}_n \text{ has a clique of size } k) = \mathbf{Prob}(Y_k \neq 0) \leq E(Y_k) = \binom{n}{k} 2^{-\binom{k}{2}} \leq$$

$$\leq \left[n 2^{-(k-1)/2}\right]^k = n^{-\varepsilon k} \to 0$$

as n tends to infinity, i.e. cliques larger than $2(1+\varepsilon)\log_2 n$ are unlikely.

Now to the second moment: It is important to realize that $Y_k^2(G)$ is the number of ordered pairs $< A, B >$, where A, B are cliques of B. Similarly as above, put $Y_{A,B}(G) = 1$ for $A, B \in \binom{X}{k}$ if both A and B are cliques of G, $Y_{A,B}(G) = 0$ otherwise. Now

$$\mathbf{Prob}(Y_{A,B} = 1) = 2^{-2\binom{k}{2}+\binom{j}{2}},$$

where j is the size of the intersection of A and B, and therefore

$$E(Y_k^2) = \sum_{A,B \in \binom{X}{2}} \mathbf{Prob}(Y_{A,B}) = \sum_{j=0}^{k} \binom{n}{k}\binom{n-k}{k-j}\binom{k}{j} 2^{-2\binom{k}{2}+\binom{j}{2}},$$

which implies that

$$\mathbf{Prob}(\mathcal{G}_n \text{ has no clique of size } k) = \mathbf{Prob}(Y_k = 0) \leq \frac{E(Y_k^2)}{E^2(Y_k)} - 1 =$$

$$= \sum_{j=1}^{k} \frac{\binom{n-k}{k-j}\binom{k}{j}}{\binom{n}{k}} 2^{\binom{j}{2}}.$$

We will prove later that if $k \leq 2(1-\varepsilon)\log_2 n$, then the probability that $\mathcal{G}_n$ has a clique with k elements tends to 1 as n tends to infinity. Thus, most graphs with n vertices have the largest clique of size $2(1+o(1))\log_2 n$.

1.3 Bounds of Chernoff and Azuma

Let $X_1, \ldots, X_n$ be a sequence of real-valued random variables. We will suppose that values of any X_i are always close to its expectation, e.g. $|X_i - E(X_i)| \leq 1$, and we ask about concentration of their sum $S = X_1 + \cdots + X_n$ around its expectation. In the simplest case, when the variables are independent, the Chernoff bound says that $|S - E(S)|$ is likely to be of order $O(\sqrt{n})$. A generalization of a sequence of independent variables is a martingale, and Azuma's inequality says exactly the same for martingales as the Chernoff bound in the independent case. However, we will start with what could be called an upper semi-martingale, prove a generalization of Azuma bound to this case, and give the inequalities of Chernoff and Azuma as corollaries.

Once again we are going to show that the applications of the (discrete) probability theory are more art than a science, because the tools are quite simple (though powerful), but the way they are used is often quite nontrivial.

Theorem 1.3.1 *Let $0 = X_0, \ldots, X_n$ be random variables such that $|X_i| \leq 1$ for $i = 0, \ldots, n$. Denote $Y_i = X_0 + \cdots + X_i$. If*

$$E(X_i \mid Y_{i-1} = \xi) \leq 0$$

for $i = 1, \ldots, n$ and for all ξ such that $\mathbf{Prob}(Y_{i-1} = \xi) > 0$, then

$$\mathbf{Prob}(Y_0, \ldots, Y_n > \varepsilon\sqrt{n}) < e^{\varepsilon^2/2}$$

for arbitrary number $\varepsilon > 0$.

Proof: Denote $\varepsilon/\sqrt{n}$ by α. Put

$$f(x) = \frac{e^{\alpha} + e^{-\alpha}}{2} + \frac{x}{\alpha}\frac{e^{\alpha} - e^{-\alpha}}{2}.$$

Since $e^{-\alpha} = f(-\alpha)$, $e^{\alpha} = f(\alpha)$, and the function e^x is convex, the inequality $e^x \leq f(x)$ is true for all $-\alpha \leq x \leq \alpha$.

Suppose that ξ is such that $\mathbf{Prob}(Y_{i-1} = \xi) > 0$.

$$E(e^{\alpha X_i} \mid Y_{i-1} = \xi) \leq E(f(\alpha X_i) \mid Y_{i-1} = \xi) = f(E(\alpha X_i \mid Y_{i-1} = \xi)) =$$

$$= f(\alpha E(X_i \mid Y_{i-1} = \xi)) \leq f(0) = \frac{e^{\alpha} + e^{-\alpha}}{2} \leq e^{\alpha^2/2},$$

where the last inequality follows from comparing the corresponding terms of the Taylor expansions of both expressions.

Let $0 < i \leq n$.

$$E(e^{\alpha Y_i}) = E(e^{\alpha(Y_{i-1}+X_i)}) = E(e^{\alpha Y_{i-1}} e^{\alpha X_i}) =$$

$$= \sum_\xi E(e^{\alpha Y_{i-1}} e^{\alpha X_i} \mid Y_{i-1} = \xi)\mathbf{Prob}(Y_{i-1} = \xi) =$$

$$= \sum_\xi E(e^{\alpha \xi} e^{\alpha X_i} \mid Y_{i-1} = \xi)\mathbf{Prob}(Y_{i-1} = \xi) \leq$$

$$\leq \sum_\xi e^{\alpha \xi} E(e^{\alpha X_i} \mid Y_{i-1} = \xi)\mathbf{Prob}(Y_{i-1} = \xi) \leq$$

$$\leq \sum_\xi e^{\alpha \xi} e^{\alpha^2/2}\mathbf{Prob}(Y_{i-1} = \xi) =$$

$$= e^{\alpha^2/2} \sum_\xi e^{\alpha \xi}\mathbf{Prob}(Y_{i-1} = \xi) = e^{\alpha^2/2} E(e^{\alpha Y_{i-1}}),$$

which gives by induction

$$E(e^{\alpha Y_n}) \leq e^{\alpha^2 n/2} E(e^{\alpha Y_0}) = e^{\alpha^2 n/2}.$$

Finally, the Markov inequality implies that

$$\mathbf{Prob}(Y_i > \varepsilon\sqrt{n}) = \mathbf{Prob}(e^{\alpha S} > e^{\alpha\varepsilon\sqrt{n}}) \leq E(e^{\alpha S})e^{-\alpha\varepsilon\sqrt{n}} \leq$$

$$\leq e^{\alpha^2 n/2} e^{-\alpha^2 n} = e^{-\alpha^2 n/2} = e^{-\varepsilon^2/2}.$$

♣

Now let us go to the special cases.

Definition 1.3.2 *A sequence $Y_0, \ldots, Y_n$ of real-valued random variables is called a* martingale *if*

$$E(Y_i \mid Y_{i-1} = \xi) = \xi$$

for $i = 1, \ldots, n$ and for all ξ such that $\mathbf{Prob}(Y_{i-1} = \xi) > 0$.

It is clear that if $Y_0, \ldots, Y_n$ is a martingale such that $Y_0 = 0$ and $|Y_i - Y_{i-1}| \leq 1$ for $i = 1, \ldots, n$, then the sequence $Y_0, Y_1 - Y_0, Y_2 - Y_1, \ldots, Y_n - Y_{n-1}$ verifies the assumptions of Theorem /refsumbound, and therefore $\mathbf{Prob}(Y_n > \varepsilon\sqrt{n}) \leq e^{\varepsilon^2/2}$. If the theorem is applied to the sequence $-Y_0, -Y_1, \ldots, -Y_n$, which is obviously a martingale as well, we obtain a lower bound $\mathbf{Prob}(Y_n < -\varepsilon\sqrt{n}) \leq e^{\varepsilon^2/2}$, which implies that the probability that $|Y_n|$ is greater than $\varepsilon\sqrt{n}$ is at most $2e^{\varepsilon^2/2}$. Well known *Azuma inequality* is nothing else than Theorem 1.3.1 applied to martingales.

If $0 = X_0, \ldots, X_n$ are independent random variables verifying $|X_i| \leq 1$, $E(X_i) = 0$, then both this sequence and its negation verify the assumption of Theorem 1.3.1, and we obtain a variant of the *Chernoff bound*.

We will use the Chernoff bound so often that we will give no examples of its use here. However, we are going to show a nontrivial example of a martingale, that has been used with a great success in the random graph theory.

Theorem 1.3.3 *Let (X, p) be a probability space, f be a real-valued function on X, and $E_0, \ldots, E_n$ be equivalences on X such that E_i is finer than E_{i-1} for $i = 1, \ldots, n$, E_k is the identity relation and xE_0y for any $x, y \in X$. Define*

$$Y_i(x) = E(f(y) \mid yE_kx).$$

Then the sequence $Y_0, \ldots, Y_n$ forms a martingale.

Proof: Suppose without loss of generality that $p(x) > 0$ for each $x \in X$. We will denote $\sum_{x \in C} p(x)$ by $\| C \|$. Given an equivalence class C of E_j, the function Y_j is constant on C, and therefore

$$E(Y_j|C) = Y_j(x) = E(f(y) \mid yE_jx) = E(f|C) \quad \text{for any } x \in C.$$

Let ξ is such that $Y_{i-1}(x) = \xi$ for some $x \in X$. Denote be A the set of all $x \in X$ such that $Y_{i-1} = \xi$. It is clear that Y_j is constant on any equivalence class of Y_j for arbitrary j. Therefore A is a disjoint union of some classes of E_{i-1}; let us denote them by $A_1, \ldots, A_m$. Each A_ℓ is in turn a disjoint union of classes $B_{\ell,k_1} \ldots, B_{\ell,k_\ell}$ of E_i.

Now,

$$\| A \| \, E(Y_i|A) = \sum_\ell \| A_\ell \| \, E(Y_i|A_\ell) = \sum_\ell \sum_k \| B_{\ell,k} \| \, E(Y_i|B_{\ell,k}) =$$

$$= \sum_{\ell} \sum_{k} \parallel B_{\ell,k} \parallel E(f|B_{\ell,k}) = \sum_{\ell} \parallel A_\ell \parallel E(f|A_\ell) = \sum_{\ell} \parallel A_\ell \parallel E(Y_{i-1}|A_\ell) =$$

$$= \sum_{\ell} \parallel A_\ell \parallel \xi = \parallel A \parallel \xi,$$

where the sums are over $\ell = 1, \ldots, m$ and $k = 1, \ldots, k_\ell$. Thus, we have proved that $E(Y_i|Y_{i-1} = \xi) = \xi$. ♣

A nice property of the martingale from the last theorem is that $Y_0 = E(f)$, and $Y_n = f$, and the Azuma inequality can be used to bound the probability that $f - E(f)$ is large.

Two important special classes of martingales on the graphs are defined as follows: Let X be the set of all graphs with the vertex set $\{1, \ldots, n\}$, $\mathcal{G}_n$ be the uniform probability space over G_n, and f be a real-valued function on G_n. Order two-element subsets of $\{1, \ldots, n\}$ into a sequence $e_1, e_2, \ldots, e_N$, where $N = \binom{n}{2}$.

Given $i = 0, \ldots, n$, we write GE_iH if $\{x, y\}$ is an edge of G if and only if $\{x, y\}$ is an edge of H for any two $1 \leq x, y \leq i$.

Given $i = 0, \ldots, N$, we write $G\overline{E}_iH$ if e_j is an edge of G if and only if e_j is an edge of H for each $1 \leq j \leq i$.

The sequence E_i gives rise to a *vertex exposure* martingale of f, while the sequence $\overline{E}_i$ defines an *edge exposure* martingale.

Azuma inequality assumes a certain bound to $|Y_i - Y_{i-1}|$, which can be proved under certain conditions on the function f. Suppose that if two graphs $G, H \in G_n$ differ in only one edge, then $|f(G) - f(H)| \leq 1$. In this case the edge martingale $\mathcal{M}(f, n)$ satisfies the inequality $|Y_i - Y_{i-1}| \leq 1$ for all $i = 1, \ldots, N$, and therefore the value of $|f - E(f)|$ can be bounded by Theorem 1.3.1. Similar statement is true for the vertex martingale.

Finally we will show how Azuma inequality can be used to prove much sharper bounds to the the size of the largest clique of a random graph.

Let $\overline{Y}_k(G)$ be the largest size of a collection of edge-disjoint cliques of size k. Note that $\overline{Y}_k(G) > 0$ if and only if G has a clique of size k. Moreover, if one edge is added to a graph G, $\overline{Y}_k(G)$ can not change by more than 1, since we consider edge-disjoint families of cliques. Therefore Azuma inequality implies that

Lemma 1.3.4

$$\mathbf{Prob}(\overline{Y}_k = 0) \leq e^{E^2(\overline{Y}_k)/(n(n-1))}.$$

Proof: Substitute $\varepsilon = E(\overline{Y}_k)/\sqrt{n(n-1)/2}$ into

$$\mathbf{Prob}(\overline{Y}_k = 0) \leq \mathbf{Prob}(E(\overline{Y}_k) - \overline{Y}_k > \varepsilon\sqrt{n(n-1)/2}) \leq e^{\varepsilon^2/2}.$$

♣

Let k_0 be a solution of the equation $\binom{n}{k}2^{-\binom{k}{2}} = 1$. We have shown in the previous section that if $k \leq k_0 - \delta$ for some constant $\delta > 0$, the expected number of k cliques in a random graph grows to infinity as $n \to \infty$. Now we prove that $E(\overline{Y}) = \Omega(n^2/\log^8 n)$. As a consequence we get a very sharp bound to the probability of the existence of cliques of sufficiently large size.

Theorem 1.3.5 *If k_0 is a solution of $\binom{n}{k}2^{-\binom{k}{2}} = 1$ and $k \leq k_0 - 4$ then the probability that a random graph with n vertices has no clique of size k is at most $exp(-\Omega(n^2/\log^8 n))$.*

Proof: Denote

$$M = \binom{n}{k}2^{-\binom{k}{2}},$$

$$a_j = \frac{\binom{n-k}{k-j}\binom{k}{j}}{\binom{n}{k}}2^{\binom{j}{2}}.$$

We have already shown in section on Chebychev inequality that M is the expected number of k-cliques in a random graph with n vertices, $M^2 a_j$ is the expected number of pairs (A, B) of k-cliques such that $|A \cap B| = j$. Since

$$b_j = \frac{a_{j+1}}{a_j} = \frac{(k-j)^2}{(j+1)(n-2k+j+1)}2^j,$$

$$\frac{b_{j+1}}{b_j} = \frac{\left(1 - \frac{1}{k-j}\right)^2}{\left(1 + \frac{1}{j+1}\right)\left(1 + \frac{1}{n-2k+j+1}\right)} \cdot 2,$$

$b_{j+1}/b_j > 1$ for $j \geq 1$ and sufficiently large n, which means that b_j is an increasing sequence, and therefore a_j is convex, i.e. the sequence a_j is first (possibly) decreasing and then (possibly) increasing, which imply that $a_j \leq \max(a_3, a_k)$ for $j = 4, 5, \ldots, k-1$,

$$a_3 + \cdots + a_k \leq k \max(a_3, a_k) = O(\log(n)) \max(a_3, a_k).$$

Direct evaluation of a_j, $j = 0, 1, 2, 3$ gives

$$a_j = \frac{n^{k-j}k!}{n^k(k-j)!}\left(1 - O\left(\frac{k^2}{n}\right)\right)\binom{k}{j}2^{\binom{j}{2}} =$$

$$= \frac{k^{2j}}{n^j}\frac{2^{\binom{j}{2}}}{j!}\left(1 - O\left(\frac{k^2}{n}\right)\right)\left(1 - O\left(\frac{j}{k}\right)\right),$$

$$a_0 = 1 + o(1), \quad a_1 = \frac{k^2}{n}(1 + o(1)), \quad a_2 = \frac{k^4}{n^2}(1 + o(1)), \quad a_3 = \frac{4k^6}{3n^3}(1 + o(1)),$$

moreover

$$a_k = \frac{1}{M}.$$

Let $\mathcal{C}$ be a set of k-cliques generated at random so that $\mathbf{Prob}(C \in \mathcal{C}) = q$ independently for different cliques C, where q is a number that will be determined later, $\mathcal{D}$ be a set of unordered pairs $\{A, B\}$, $A, B \in \mathcal{C}$, $|A \cap B| \geq 2$, $\mathcal{E}$ be an arbitrary subset of $\mathcal{C}$ such that either $A \in \mathcal{E}$ and/or $B \in \mathcal{E}$ for each $(A, B) \in \mathcal{D}$.

$\mathcal{E}$ is a set of edge disjoint k-cliques, and therefore $|\mathcal{E}| \leq \overline{Y}_k$,

$$E(\overline{Y}_k) \geq E(|\mathcal{E}|) \geq E(|\mathcal{C}| - |\mathcal{D}|) = qM - q^2M^2\left(\frac{a_2}{2} + \cdots + \frac{a_{k-1}}{2} + a_k\right) =$$

$$= qM - \frac{q^2M^2a_2}{2}(1 + o(1)) = \frac{1}{a_2} - \frac{1}{2a_2} = \frac{1}{2a_2} = \frac{n^2}{k^4}(1 + o(1))$$

if we choose $q = (Ma_2)^{-1}$.

We finish the proof by substituting the last bound to $E(\overline{Y}_k)$ to the inequality of the preceding lemma. ♣

An easy consequence of the last theorem we obtain a precise estimation of the expected chromatic number of a random graph, thus giving an answer to a problem that was open for many years

Theorem 1.3.6 *With probability* $1 - o(1)$*, the chromatic number of a random graph is*

$$\frac{n}{2\log_2 n}(1 + o(1)).$$

Proof: Let $\mathbf{G}$ be a random graph with a vertex set $\{1, \ldots, n\}$, $m = n/\log_2^2 n$, k_0 be a solution of $\binom{m}{x}2^{-\binom{x}{2}}$, k be the largest integer less or equal to $k_0 - 4$. Suppose that X is an arbitrary m-element subset of the vertex set of $\mathbf{G}$. The subgraph induced by $\mathbf{G}$ on X is a uniform random graph on X, and therefore it follows from the preceding theorem that the probability that the subgraph has no k clique is

$$\exp(-\Omega(m^2/\log^8 m)) = \exp(-\Omega(n^{2+o(1)})).$$

Therefore the probability that there exists an m-element set of vertices which contains no k-clique is at most

$$\binom{n}{m}\exp(-\Omega(n^{2+o(1)})) \leq \exp(O(m \cdot \log n))\exp(-\Omega(n^{2+o(1)})) = o(1).$$

We can therefore suppose that each m element subset of vertices contains a clique of size k. Since the distribution of the complement of $\mathbf{G}$ is the same as the distribution of $\mathbf{G}$, the probability that each m-element set contains an independent set of size k is $1 - o(1)$ as well. However, in this case we can color $\mathbf{G}$ by $n/k + m$ colors using the next algorithm

$Z :=$ the set of vertices; $i := 0$;
while $|Z| \geq m$ **do begin**
  $S :=$ an independent set of k vertices of Z;
  $i := i + 1$; color vertices of S by i;
  end; color vertices of Z by defferent colors;

♣

1.4 Rapidly mixing Markov chains

Several interesting probabilistic approximation algorithms were obtained using random walks in a certain space. Suppose that V is a set of known size and $A \subset V$ such that the membership problem for A is not too difficult. We can approximate the size of A as follows: define a random walk process in V in such a way that if t is sufficiently large, probabilities $p_t(x)$ that after t steps of the walk we are in $x \in V$ are approximately the same. Repeat a random walk m times, and denote by k the number of walks that arrive into an element of A after t steps. If m is sufficiently large, then k/m is a good approximation of $|A|/|V|$, and therefore $|A|$ can be approximated by $|V|k/m$.

We will show later that a permanent of a given matrix can be approximated in polynomial time using an algorithm of this type.

The efficiency of random walk algorithm depends strongly on necessary values of t and m. While bounds to m can be obtained using the Chernoff bound, it is much more complicated to obtain good bounds to t. The tool which we will use is called a rapidly mixing Markov chain.

Definition 1.4.1 *A matrix* $P = (p_{ij})_{i,j=1,\dots}$ *is called* stochastic *if*

$$\sum_{j=1}^{n} p_{ij} = 1 \qquad \textit{for all } i = 1, \dots, n.$$

A stochastic matrix P with non-negative elements can be viewed as a matrix of transition probabilities of a random walk on an n-element set $X = \{1, \dots, n\}$: elements of X denote places that can be visited during the walk, and the probability that we move from x_i to x_j is p_{ij}. The probability of transition $x_i \to x_j$ does not depend on the way that lead us to x_i.

A *distribution* of a random walk described by P is a function $e : X \to [0, 1]$ such that $\sum_{i=1}^{n} e(i) = 1$ (i.e. (X, e) forms a probability space). $e(x_i)$ is interpreted as a probability that we are in a state i, and will be denoted simply by e_i. Initial position of the walk is usually such that $e_i = 1$ for some i. A random walk is described by a sequence $e^{(0)}, e^{(1)}, \dots$ of positions such that $e^{(k)} = e^{(k-1)}P$, i.e.

$$e_i^{(k)} = \sum_{j=1}^{n} e_j^{(k-1)} p_{ji}.$$

We say that a Markov chain $P = (p_{ij})$ is strongly aperiodic if $p_{ii} > 0$ for each $i = 1, \dots, n$, and is irreducible if the directed graph with the vertex set X and edges represented by all pairs (i, j) such that $p_{ij} > 0$ is strongly connected. In other words, P is irreducible if for each $i, j = 1, \dots, n$ there is a sequence $i = i_0, \dots, i_m = j$ such that $p_{i_{k-1}i_k} > 0$ for $k = 1, \dots, m$. The *stationary distribution* of a Markov chain P is a vector π such that $\pi = \pi P$ (i.e. $\pi_i = \sum_{i=j}^{n} \pi_j p_{ji}$ for $i = 1, \dots, n$), $\sum_i \pi_i = 1$ and $\pi_i \geq 0$ for $i = 1, \dots, n$.

The classical theory of Markov chains shows that a strongly aperiodic irreducible Markov chain has always a unique stationary distribution π such that all π_i are greater than 0, and if e is an arbitrary distribution and we define $e(0) = e$, $e^{(t+1)} = e^{(t)}P$, then the sequence $e^{(0)}, e^{(1)}, \ldots$ converges to the stationary distribution. The aim of this section is to prove that, under some conditions, the convergence is sufficiently fast.

The requirement of irreducibility will be satisfied in all our applications. If P is not strongly aperiodic, then we usually replace it by the chain with matrix $(I + P)/2$, where I is the unity matrix. Note that π_i is a stationary distribution of P if and only if it is a stationary distribution of $(I + P)/2$. However, it might happen that the sequence of distributions defined by $e^{(t+1)} = e^{(t)}(I + P)/2$ converges to π, while the sequence $e^{(t+1)} = e^{(t)}P$ does not, see e.g. the matrix

$$P = \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}.$$

Therefore we will study properties of a sequence defined by $e^{(t+1)} = e^{(t)}(I + P)/2$, where P is a given irreducible chain.

From now on, π will denote the stationary distribution. We will also denote $w_{ij} = \pi_i p_{ij}$.

Given a vector e, we will denote by $\hat{e}$ the vector with coordinates $\hat{e}_i = (e_i - \pi_i)/\pi_i$. The *norm* of a vector e is defined by

$$\|e\|^2 = \sum_{i=1}^{n} \pi_i \hat{e}_i^2 = \sum_{i=1}^{n} \frac{(e_i - \pi_i)^2}{\pi_i}.$$

Note that

$$\sum_{j=1}^{n} w_{ij} = \sum_{j=1}^{n} w_{ji} = \pi_i \quad \text{for each } i = 1, \ldots, n,$$

because

$$\sum_j w_{ij} = \sum_j \pi_i p_{ij} = \pi_i \sum_j p_{ij} = \pi_i = \sum_j \pi_j p_{ji} = \sum_j w_{ji}.$$

Given $A \subset X$, denote

$$C(A) = \sum_{i \in A} \pi_i, \qquad F(A) = \sum_{i \in A} \sum_{j \notin A} w_{ij}.$$

Note that $C(A) = 1 - C(X - A)$,

$$F(A) + \sum_{i,j \in A} w_{ij} = \sum_{i \in A} \sum_j w_{ij} = \sum_{i \in A} \sum_j w_{ji} =$$

$$= \sum_{i,j\in A} w_{ji} + \sum_{i\in A}\sum_{j\notin A} w_{ji} = \sum_{i,j\in A} w_{ij} + F(X-A),$$

which implies $F(A) = F(X-A)$ for each $A \subset X$.

From now on we will suppose that $F(A) \geq \alpha C(A)$ for each $A \subset X$ such that $C(A) \leq 1/2$. Let $e = (e_1, \ldots, e_n)$ be a vector, S_k be the set of all integers i such that $1 \leq i \leq k$.

We will prove first

Lemma 1.4.2 *If there is r such that $0 \leq r < n$, $C(S_r) \leq 1/2$, and $e_1 \geq e_2 \geq \ldots \geq e_r \geq e_{r+1} = \ldots = e_n = 0$, then*

$$\sum_{i,j=1}^{n} w_{ij}|e_i^2 - e_j^2| \geq 2\alpha \sum_{i=1}^{n} \pi_i e_i^2.$$

Proof:

$$\sum_{ij} w_{ij}|e_i^2 - e_j^2| = \sum_{i<j} w_{ij}(e_i^2 - e_j^2) + \sum_{i<j} w_{ji}(e_i^2 - e_j^2) =$$

$$= \sum_{i<j} w_{ij} \sum_{i\leq k<j} (e_k^2 - e_{k+1}^2) + \sum_{i<j} w_{ji} \sum_{i\leq k<j} (e_k^2 - e_{k+1}^2) =$$

$$= \sum_{k=1}^{n-1} (e_k^2 - e_{k+1}^2) \sum_{\substack{i\leq k\\ k<j}} w_{ij} + \sum_{k=1}^{n-1} (e_k^2 - e_{k+1}^2) \sum_{\substack{i\leq k\\ k<j}} w_{ji} =$$

$$= \sum_{k=1}^{r} (e_k^2 - e_{k+1}^2)\left[F(S_k) + F(X-S_k)\right] =$$

$$= 2\sum_{k=1}^{r} (e_k^2 - e_{k+1}^2) F(S_k) \geq$$

$$\geq 2\alpha \sum_{k=1}^{r} (e_k^2 - e_{k+1}^2) C(S_k) = 2\alpha \sum_{i=1}^{r} e_i^2 \pi_i = 2\alpha \sum_{i=1}^{n} \pi_i e_i^2.$$

♣

As a consequence we obtain

Lemma 1.4.3 *If there is r such that $0 \leq r < n$, $C(S_r) \leq 1/2$, and $e_1 \geq e_2 \geq \ldots \geq e_r \geq e_{r+1} = \ldots = e_n = 0$, then*

$$\sum_{i,j=1}^{n} w_{ij}(e_i - e_j)^2 \geq \alpha^2 \sum_{i} \pi_i e_i^2.$$

Proof: Since $e_i^2 + e_j^2 = \frac{1}{2}\left[(e_i+e_j)^2 + (e_i-e_j)^2\right] \geq (e_i+e_j)^2/2$,

$$\sum_{ij} w_{ij}(e_i+e_j)^2 \leq 2\sum_{ij} w_{ij}(e_i^2+e_j^2) = 2\sum_i e_i^2 \sum_j w_{ij} + 2\sum_j e_j^2 \sum_i w_{ij} =$$

$$= 2\sum_i \pi_i e_i^2 + 2\sum_j \pi_j e_j^2 = 4\sum_i \pi_i e_i^2,$$

and using the preceding lemma and the Cauchy-Schwartz inequality, we obtain

$$\sum_{ij} w_{ij}(e_i-e_i)^2 \geq \frac{\sum_{ij} w_{ij}(e_i-e_j)^2 \sum_{ij} w_{ij}(e_i+e_j)^2}{4\sum_i \pi_i e_i^2} \geq$$

$$\geq \frac{\left[\sum_{ij} w_{ij}|e_i-e_j|(e_i+e_j)\right]^2}{4\sum_i \pi_i e_i^2} =$$

$$= \frac{\left[\sum_{ij} w_{ij}|e_i^2-e_j^2|\right]^2}{4\sum_i \pi_i e_i^2} \geq$$

$$\geq \frac{4\alpha^2\left[\sum_i \pi_i e_i^2\right]^2}{4\sum_i \pi_i e_i^2} = \alpha^2 \sum_i \pi_i e_i^2.$$

♣

Lemma 1.4.4 *Denote* $S^+ = \{i \mid e_i > 0\}$, $S^- = \{i \mid e_i < 0\}$. *If both* $C(S^+) \leq 1/2$ *and* $C(S^-) \leq 1/2$, *then*

$$\sum_{i,j=1}^{n} w_{ij}(e_i-e_j)^2 \geq \alpha^2 \sum_i \pi_i e_i^2.$$

Proof: Put $e_i^+ = \max(e_i, 0)$, $e_i^- = \max(-e_i, 0)$. It follows from the preceding lemma that the inequality of the present lemma holds for both e^+ and e^-. Therefore

$$\sum_{ij} w_{ij}(e_i-e_j)^2 \geq \sum_{ij} w_{ij}(e_i^+-e_j^+)^2 + \sum_{ij} w_{ij}(e_i^- - e_j^-)^2 \geq$$

$$\geq \alpha^2 \sum_i \pi_i (e_i^+)^2 + \alpha^2 \sum_i \pi_i (e_i^-)^2 = \alpha^2 \sum_i \pi_i e_i^2.$$

♣

Finally, we obtain

Theorem 1.4.5 *If $e = (e_1, \ldots, e_n)$ is a distribution, then*

$$\sum_{i,j=1}^{n} w_{ij}(\hat{e}_i - \hat{e}_j)^2 \geq \alpha^2 \sum_i \pi_i \hat{e}_i^2.$$

Proof: Note that

$$\sum_i \pi_i \hat{e}_i = \sum_i (e_i - \pi_i) = \sum_i e_i - \sum_i \pi_i = 1 - 1 = 0.$$

Choose $1 \leq m \leq n$ such that $C(S_{m-1}) \leq \frac{1}{2} < C(S_m)$, and put $f_i = \hat{e}_i - \hat{e}_m$. In view of the preceding lemma applied to f instead of e,

$$\sum_{i,j=1}^{n} w_{ij}(\hat{e}_i - \hat{e}_j)^2 = \sum_{i,j=1}^{n} w_{ij}(f_i - f_j)^2 \geq \alpha^2 \sum_i \pi_i(\hat{e}_i - \hat{e}_m)^2 \geq$$

$$\geq \alpha^2 \sum_i \pi_i \hat{e}_i^2 - 2\alpha^2 \hat{e}_m \sum_i \pi_i \hat{e}_i = \alpha^2 \sum_i \pi_i \hat{e}_i^2.$$

♣

Theorem 1.4.6 *Let $e = (e_1, \ldots, e_n)$ be a distribution, $Q = (I+P)/2$. Then*

$$\|e\|^2 - \|eQ\|^2 \geq \frac{1}{4} \sum_{i,j=1}^{n} w_{ij}(\hat{e}_i - \hat{e}_j)^2.$$

Proof:

$$\|e\|^2 = \sum_i \pi_i \hat{e}_i^2 = \frac{1}{2}\left(\sum_i \hat{e}_i^2 \sum_j w_{ij} + \sum_i \hat{e}_i^2 \sum_j w_{ji}\right) =$$

$$= \frac{1}{2}\left(\sum_i \hat{e}_i^2 \sum_j w_{ij} + \sum_j \hat{e}_j^2 \sum_i w_{ij}\right) = \sum_{ij} w_{ij} \frac{\hat{e}_i^2 + \hat{e}_j^2}{2},$$

if $h = eP$, $k = e(I+P)/2$, then

$$\hat{e}_j = \frac{1}{\pi_j} \hat{e}_j \sum_i w_{ij} = \frac{1}{\pi_j} \sum_i w_{ij} \hat{e}_j,$$

$$\hat{h}_j = \frac{1}{\pi_j}(h_j - \pi_j) = \frac{1}{\pi_j}(\sum_i e_i p_{ij} - \sum \pi_i p_{ij}) =$$

$$= \frac{1}{\pi_j}\sum_i \frac{w_{ij}}{\pi_i}(e_i - \pi_i) = \frac{1}{\pi_j}\sum_i w_{ij}\hat{e}_i,$$

$$\hat{k}_j = \frac{k_j - \pi_j}{\pi_j} = \frac{1}{2\pi_j}[(e_j + h_j) - 2\pi_j] = \frac{1}{2\pi_j}[(e_j - \pi_j) + (h_j - \pi_j)] =$$

$$= \frac{\hat{e}_j + \hat{h}_j}{2} = \frac{1}{\pi_j}\sum_i w_{ij}\frac{\hat{e}_i + \hat{e}_j}{2},$$

the convexity of $f(x) = x^2$ implies that

$$\|\hat{k}\|^2 = \sum_j \pi_j \hat{k}_j^2 = \sum_j \pi_j \left(\sum_i \frac{w_{ij}}{\pi_j}\frac{\hat{e}_i + \hat{e}_j}{2}\right)^2 \le \sum_j \pi_j \sum_i \frac{w_{ij}}{\pi_j}\left(\frac{\hat{e}_i + \hat{e}_j}{2}\right)^2 =$$

$$= \sum_{ij} w_{ij}\left(\frac{\hat{e}_i + \hat{e}_j}{2}\right)^2,$$

and therefore

$$\|e\|^2 - \|k\|^2 \ge \sum_{ij} w_{ij}\left[\frac{\hat{e}_i^2 + \hat{e}_j^2}{2} - \left(\frac{\hat{e}_i + \hat{e}_j}{2}\right)^2\right] = \sum_{ij} w_{ij}\left(\frac{\hat{e}_i^2}{4} - \frac{\hat{e}_i\hat{e}_j}{2} + \frac{\hat{e}_j^2}{4}\right) =$$

$$= \frac{1}{4}\sum_{ij} w_{ij}(\hat{e}_i - \hat{e}_j)^2.$$

♣

Theorem 1.4.7 *Let $e = (e_1, \ldots, e_n)$ be a distribution, define distributions $e^{(0)}, e^{(1)}, \ldots$ by*

$$e^{(0)} = e, \qquad e^{(t+1)} = e^{(t)}\frac{I + P}{2}.$$

Then

$$\|e^{(t)}\|^2 \le \|e^{(0)}\|^2 \left(1 - \left(\frac{\alpha}{2}\right)^2\right)^t.$$

Proof: It follows from the preceding two theorems that

$$\|e^{(t)}\|^2 - \|e^{(t+1)}\|^2 \ge \frac{1}{4}\sum_{ij} w_{ij}(\hat{e}_i^{(t)} - \hat{e}_j^{(t)})^2 \ge \frac{1}{4}\alpha^2\|e^{(t)}\|^2,$$

$$\|e^{(t+1)}\|^2 \leq \left(1 - \frac{\alpha^2}{4}\right) \|e^{(t)}\|^2.$$

♣

The largest constant α such that $F(A) \geq \alpha C(A)$ for each $A \subset X$ such that $C(A) \leq 1/2$ is called a *conductivity* of the chain P. Theorem 1.4.7 shows that the convergence of a random walk determined by $(I + P)/2$ is fast, provided the chain P has a sufficiently large conductivity.

A rapid mixing property of a chain with large conductivity similar to Theorem 1.4.7 can be proved in another way, using spectral theory, if the chain is *time reversible*, which means that $w_{ij} = w_{ji}$ (or equivalently $\pi_i p_{ij} = \pi_j p_{ji}$) for each $i, j = 1, \ldots, n$.

Let $D^{1/2}$ be a diagonal matrix with diagonal $(\sqrt{\pi_1}, \ldots, \sqrt{\pi_n})$. The time reversibility is equivalent to the requirement that $A = D^{1/2} P D^{-1/2}$ is symmetric, where $D^{-1/2}$ is the inverse of $D^{1/2}$, because

$$A_{ij} = p_{ij}\sqrt{\frac{\pi_i}{\pi_j}}.$$

It is well known that eigenvalues of a symmetric matrix are all real, and if $\lambda_1 \geq \ldots \geq \lambda_n$ are eigenvalues of $D^{1/2} P D^{-1/2}$ (each appearing with its multiplicity), we can choose left eigenvectors $\mathbf{e}^{(i)}$ of $D^{1/2} P D^{-1/2}$ corresponding to λ_i so that they form an orthonormal basis. It is easy to check that

$$D^{1/2} P D^{-1/2} = \sum_{i=1}^{n} \lambda_i (\mathbf{e}^{(i)})^T \mathbf{e}^{(i)}$$

(T used for the transpose), because multiplying both sides by any $\mathbf{e}^{(j)}$ from left gives the same vectors equal to $\lambda_j \mathbf{e}^{(j)}$. Denote $E^{(i)} = (\mathbf{e}^{(i)})^T \mathbf{e}^{(i)}$. Since obviously $E^{(i)} E^{(j)} = 0$ for $i \neq j$ and $E^{(i)} E^{(i)} = E^{(i)}$, this representation of A is extremely well suited for computing powers:

$$A^k = \sum_{i=1}^{n} \lambda_i^k (\mathbf{e}^{(i)})^T \mathbf{e}^{(i)}.$$

and hence

$$P^k = D^{-1/2} A^k D^{1/2} = \sum_{i=1}^{n} \lambda_i (D^{-1/2} (\mathbf{e}^{(i)})^T \mathbf{e}^{(i)} D^{1/2}).$$

Now let us look more closely to the eigenvalues of P and A (which are the same). We know that the stationary distribution is a left eigenvector of P with eigenvalue $\lambda_1 = 1$. The left eigenvector of A corresponding to the eigenvalue $\lambda_1 = 1$ is $\pi D^{-1/2}$, i.e. the vector $(\sqrt{\pi_1}, \dots, \sqrt{\pi_n})$. By standard theory of non-negative matrices, the remaining eigenvalues of A verify $|\lambda_i| < 1$ for $i = 2, \dots, n$. Therefore if we denote $P^k = (p_{ij}^{(k)})$, then

$$p_{uv}^{(k)} = \sum_{i=1}^{n} \sqrt{\frac{\pi_v}{\pi_u}} \lambda_i^k \mathbf{e}_u^{(i)} \mathbf{e}_v^{(i)} = \pi_v + \sqrt{\frac{\pi_v}{\pi_u}} \sum_{i=2}^{n} \lambda_i^k \mathbf{e}_u^{(i)} \mathbf{e}_v^{(i)},$$

and therefore if $e^{(0)}$ is a distribution and $e^{(t+1)} = e^{(t)}P$, then

$$\max_u \left| \frac{e_u^{(t)} - \pi_u}{\pi_u} \right| \leq \max_{u,v} \left| \frac{p - \pi_u}{\pi_u} \right| = \max_{u,v} \frac{|\sum_{i=2}^{n} \lambda_i^t \mathbf{e}_u^{(i)} \mathbf{e}_v^{(i)}|}{\sqrt{\pi_u \pi_v}} \leq$$

$$\leq |\lambda_2| \frac{\max_{u,v} \sum_{i=2}^{n} |\mathbf{e}_u^{(i)}||\mathbf{e}_v^{(i)}|}{\min_u \pi_u} \leq \frac{|\lambda_2|}{\min_u \pi_u}.$$

2 Average case completeness

2.1 Basic definition

The aim of this chapter is to present the theory of average case completeness, that was started by one-page paper of L.Levin [9, 10] as an analog of the (worst-case) NP-completeness theory. The chapter is divided into four parts. The first section deals with main definitions and simple properties of introduced complexity classes. The second one brings Levin's proof of the existence of a complete problems in the class of problems with polynomially computable distribution. The third section shows that problems with samplable distribution can be reduced to problems with polynomially computable distribution. The last part of the chapter shows some problems hard on average, that are known to date.

Whenever we are studying the average case behavior of algorithm, we have to make some assumptions on the probabilistic distribution of inputs. The next definitions follow [4, 6].

Definition 2.1.1 *An* alphabet *is a finite ordered set A. A* word *over an alphabet is a sequence $u = u_1 \dots u_k$ of elements of A. k is the* length *of u, and is denoted by $|u|$. Given words $u = u_1 \dots u_k$, $v = v_1 \dots v_\ell$, we write $u < v$ if either $k < \ell$ or $k = \ell$ and there is j, $1 \leq j \leq k$ such that $u_i = v_i$ for $i < j$, and $u_j < v_j$. A successor of w, i.e. the smallest among words larger than w is denoted by w^+.*

Definition 2.1.2 *A* probability function *over an alphabet A is a nonnegative real valued function μ on A^* such that $\sum_{w \in A^*} \mu(w) = 1$. The function $\mu^*(w) = \sum_{v<w} \mu(v)$ is the probability distribution corresponding to μ.*

We suppose that a sum over the empty set is 0, i.e. $\mu^*(\Lambda) = 0$. Note that $\mu(w) = \mu^*(w^+) - \mu^*(w)$. Definition of probability function is nothing else than a redefinition of a probability space. Therefore if f is any real-valued function defined on A^*, we can determine its moments, e.g. its expectation. Especially, f might represent the running time of an algorithm; in this case the expectation of f is called the *average* running time of the algorithm. Similarly as in the theory of **NP**-completeness, we are interested in algorithms with polynomially bounded running time.

Definition 2.1.3 *We say that a nonnegative real valued function defined on A^* is polynomial on average with respect to a probability function μ, if*

$$\sum_{w \in A^*} \frac{f^{1/k}(w)}{|w|} \mu(w) < \infty$$

for some k.

It is obvious that

Lemma 2.1.4 *If a function f is bounded by a polynomial, then f is polynomial on average with respect to any probability function.*

This definition makes it possible to define problems that can be solved in polynomially bounded average time as follows

Definition 2.1.5 *A* distributed problem *over alphabet A is a couple (L, μ), where $L \subset A^*$, and μ is a probabilistic function over A.*

We say that a distributed problem (L, μ) over A can be solved (deterministically) in polynomial average time, if there is a deterministic algorithm accepting the language L such that the running time of this algorithm is polynomial on average with respect to the probability function μ.

The class of distributed problems that are solvable deterministically in polynomially bounded average time will be denoted by $\mathbf{P}_{\text{ave}}$.

Similarly as in the classical theory of **NP**-completeness, we need a notion of a reduction of one problem to another such that if a problem P_1 can be reduced to P_2, and $P_2 \in \mathbf{P}_{\text{ave}}$ then $P_1 \in \mathbf{P}_{\text{ave}}$ as well.

Definition 2.1.6 *Given probability functions μ_1, μ_2 over an alphabet A, we write $\mu_1 \leq \mu_2$ if there exists a polynomial p such that $\mu_1(w) \leq \mu_2(w)p(|w|)$ for each $w \in A^*$.*

If μ is a probability function over A_1, and $f : A_1^* \to A_2^*$ is a function, then we denote by $f(\mu)$ a probability function on A_2 defined by $f(\mu)(w) = \sum_{f(v)=w} \mu(v)$.

Recall again that the sum over the empty set of terms is 0. An important observation is that

Lemma 2.1.7 *If a function f is polynomial on average with respect to μ_2 and $\mu_1 \leq \mu_2$, then f is polynomial on average with respect to μ_1 as well.*

Proof: There is k such that $\sum_w f^{1/k}(w)|w|^{-1}\mu_2(w) < \infty$, and there is a polynomial p such that $\mu_1(w) \leq \mu_2(w)p(|w|)$. Put $\ell = k(1 + \deg(p))$, and define W as the set of all $w \in A^*$ such that

$$f^{1/\ell}(w)p(|w|) \leq f^{1/k}(w).$$

Then

$$\sum_{w \in W} \frac{f^{1/\ell}(w)}{|w|}\mu_1(w) \leq \sum_{w \in W} \frac{f^{1/\ell}(w)}{|w|}p(|w|)\mu_2(w) \leq \sum_{w \in W} \frac{f^{1/k}(w)}{|w|}\mu_2(w) < \infty,$$

while $w \notin W$ implies

$$f^{1/k}(w) < f^{1/\ell}(w)p(|w|), \quad f^{1/k-1/\ell}(w) \leq p(|w|), \quad f^{\frac{\ell-k}{\ell k}}(w) \leq p(|w|),$$

$$f^{1/\ell}(w) \leq p^{\frac{k}{\ell-k}}(w) = p^{\frac{1}{\deg(p)}}(w) \leq C$$

for some constant C, and therefore

$$\sum_{w \notin W} \frac{f^{1/\ell}(w)}{|w|} \mu_1(w) \leq \sum_{w \notin W} C\mu_1(w) \leq C.$$

♣

Definition 2.1.8 *Given distributed problems $P_1 = (L_1, \mu_1)$ over an alphabet A_1, and $P_2 = (L_2, \mu_2)$ over A_2, we say that P_1 reduces to P_2, if there is a function $\Phi : A_1^* \to A_2^*$ such that*

- *Φ can be computed deterministically in polynomially bounded worst case time,*
- *$w \in L_1$ if and only if $\Phi(w) \in L_2$, and*
- *$\Phi(\mu_1) \leq \mu_2$.*

If $\mathcal{C}$ is a class of distributed problems, we say that a distributed problem P is hard *for $\mathcal{C}$, if any problem in $\mathcal{C}$ reduces to P, and is* complete *for $\mathcal{C}$, if it is hard for $\mathcal{C}$, and is a member of $\mathcal{C}$.*

As a corollary of Lemma 2.1.7, we obtain

Lemma 2.1.9 *If $P_1 = (L_1, \mu_1)$ reduces to $P_2 = (L_2, \mu_2)$ and $P_2 \in \mathbf{P}_{\text{ave}}$, then $P_1 \in \mathbf{P}_{\text{ave}}$ as well.*

Proof: Let Φ be a reduction of P_1 to P_2 in the sense of Lemma 2.1.8. There is an algorithm $\mathcal{A}$ solving the problem L_2 in time t such that t is polynomial on average with respect to μ_2. It follows from Lemma 2.1.7 that t is polynomial on average with respect to $\Phi(\mu_1)$, too. This means that if $T(w)$ is the time that $\mathcal{A}$ needs when applied to $\Phi(w)$, then T is polynomial on average with respect to μ_1. Since Φ can be computed in polynomial worst case time, the composition of Φ and $\mathcal{A}$ is an algorithm that solves L_1 in time polynomial on average with respect to μ_1. ♣

Finally we need a distributed analog to the classical class **NP** of problems solvable nondeterministically in polynomial time. We will define the class **NP** in a slightly unusual way, which is however equivalent to the classical one.

Definition 2.1.10 *The class* **NP** *is a collection of all* $L \subset \{0,1\}^*$ *such that there is a polynomial* p*, and a function* $\phi : \{0,1\}^* \to \{0,1\}^*$ *computable in polynomially bounded worst case deterministic time such that* $u \in L$ *if and only if* $u = \phi(w)$ *for some* $w \in \{0,1\}^*$ *such that* $|w| \leq p(|u|)$*. The string* w *is called a* witness *of membership of* u *in* L*.*

The class **DNP** *is a collection of all distributed problems* (L, μ) *such that* $L \in$ **NP***.*

The definition is equivalent to the classical one, because we can test the membership of u in the range of ϕ by generating nondeterministically all strings w of the length at most $p(|u|)$ and checking whether $u = \phi(w)$; on the other hand, a nondeterministic machine can be viewed as a deterministic algorithm with two inputs: the input itself, and a sequence telling how to continue in nondeterministic forks, and if L is accepted by a nondeterministic algorithm $\mathcal{A}$, then ϕ be a function that applied to (suitable encoded) pair (u, w) simulates computation of $\mathcal{A}$ with inputs u and w, and outputs u iff the corresponding computational path of $\mathcal{A}$ accepts, otherwise outputs a fixed word $u_0 \in L$.

The definition of **DNP** is too weak, because we assume no condition on the probability function μ. We will now define two important classes of probability functions.

Definition 2.1.11 *The* uniform *probability function on* A^* *is defined by*

$$\mu(w) = \frac{|A|^{-|w|}}{(|w|+1)(|w|+2)}.$$

We say that a probability function μ *on* A^* *is* P-distributed*, if it is a rational valued function such that* μ^* *can be computed deterministically in polynomially bounded worst case time.*

A function f *is called* polynomially approximable *if there is a function* $g = g(x, \varepsilon)$ *such that* $g(x, \varepsilon)$ *can be computed in time polynomial in* $|x|$ *and* ε^{-1}*, and* $|f(x) - g(x, \varepsilon)| < \varepsilon$ *for each* $x \in \{0,1\}^*$ *and* $\varepsilon > 0$*.*

We say that μ *is samplable, if* $\mu = f(\nu)$*, where* ν *is a* P*-distributed probability function on* A^**, and* $f : A^* \to A^*$ *is a polynomially approximable function.*

We say that a problem $(L, \mu) \in$ **DNP** *is* P*-distributed, samplable, resp. if* μ *is* P*-distributed, samplable, resp.*

Lemma 2.1.12 *The uniform probability function is well defined and P-distributed.*

Proof: Let μ be the uniform probability function on A^*.

$$s_n = \sum_{|w|=n} \mu(w) = \frac{1}{(n+1)(n+2)} = \frac{1}{n+1} - \frac{1}{n+2},$$

$$s_0 + \cdots + s_n = 1 - \frac{1}{n+2},$$

$$\sum_{i=0}^{\infty} s_i = 1.$$

This also shows that if $w = w_1 \ldots w_n$, then

$$\mu^*(w) = 1 - (n+1)^{-1} + (|A|^{n-1}w_1 + \cdots + |A|w_{n-1} + w_n)|A|^{-n},$$

which is an easily computable function. ♣

Note that any P-distributed probability function is samplable (use the identity function f in the definition of a samplable function).

The notion of P-distributed problem is very restrictive, because even the requirement of polynomial time computability of μ would be too strong. A classical example of a P-distributed probability function is the uniform distribution (see below), but many other important distributions are not P-distributed. However, it is quite easy to show that there is a problem complete in the class of P-distributed **DNP** problems.

On the other hand, samplable functions are exactly those that are interesting from practical point of view. Roughly speaking, they represent the probability distribution of outputs of a randomized polynomial worst case time algorithm, using a P-distributed (e.g. uniform) random bits. Again, it is not very difficult to prove that there is a complete problem of the class of samplable **DNP** problems. However, Impagliazzo and Levin have shown [7] much stronger result, namely that if we generalize the notion of reduction of distributed problem, then there is a P-distributed problem, which is complete for the class of all samplable **DNP** problems.

We conclude this section by a short discussion of relation of our definitions to a usual way of analyzing the expected case behavior. Quite often we define a probability functions separately for strings of a given length, i.e.

we have a collection of functions $\nu_i : \{0,1\}^i \to [0,1]$. We can transform this collection to a probability function by Definition 2.1.2 by choosing a function $h : \{0, 1, \ldots, \} \to [0,1]$ such that $\sum_{i=0}^{\infty} h(i) = 1$, and put $\mu(w) = \nu_{|w|}(w)h(|w|)$.

Conversely, if μ is a probability function, we can put $h(i) = \sum_{|w|=i} \mu(w)$, $\nu_i(w) = h^{-1}(i)\mu(w)$ for each w such that $|w| = i$ (the only problem is the case when $h(i) = 0$, but in this case ν_i can be chosen arbitrarily, because it has no influence on μ).

If p_1 is a polynomial, p_2 is a function bounded from below by a polynomial, and $\sum_{i=0}^{\infty} h_1(i) = \sum_{i=0}^{\infty} h_2(i) = 1$, where $h_i = p_i^{-1}$, then if μ_i are obtained from a sequence $f_0, f_1, \ldots$ of probability functions by h_i, then $\mu_2 \leq \mu_1$. Especially if p_2 is a polynomial, too, then there are no differences between classes of functions polynomial on average with respect to both μ_1 and μ_2. Especially, it was not important which polynomial was used in the definition of the uniform distribution (the advantage of $n(n+1)$ are low degree and the fact that $\sum[n(n+1)]^{-1} = 1$.

There are several possible ways of defining the notion of "polynomial on average", which are not equivalent. Frequently used definitions are to require that $\sum_w f(w)|w|^{-k}\mu(w) < \infty$, or $\sum_{|w|=n} f(w)\nu_n(w) = O(n^k)$ for some k. They are weaker then Definition 2.1.2 (see example below), but their disadvantage is that they do not give a robust notion of polynomiality on average. Suppose that t is running time of an algorithm, assuming RAM as a model of computation. The same algorithm on a Turing machine might require t^2 time, and therefore it would be good, if t is polynomial on average if and only if t^2 is. However, if for each n $t(w) = 2^n$ for one fixed string of the length n (e.g. for the first one of length n), and 0 otherwise, then t is polynomial on average with respect to the uniform distribution, but t^2 is not.

2.2 Levin's theorem

In this paragraph we will prove that the problem defined below is complete for the class of P-distributed **DNP** problems.

In order to formulate the problem, we have to choose a way of encoding deterministic Turing machines by strings. Without loss of generality we suppose a one-tape Turing machine with input alphabet $\{0, 1, \#\}$, set of states $\{0, \ldots, k-1\}$ for some k, 0 being the terminal state, 1 the initial state. An instruction of such a machine is a 5-tuple (s, a, t, b, c), indicating that if the machine is in the state s, and its head reads a, then the machine changes its state to t, writes b on the tape, and the head moves left, right, or stays, resp. if $c = 0, 1, \#$, resp. Such an instruction can be encoded by a string of $(2\lceil \log_2 k \rceil + 3)$ elements of the alphabet $\{0, 1, \#\}$ by replacing states s, t by numbers s, t written in binary. An encoding of the machine is then a string $I_1|I_2|\ldots|I_\ell$, where I_i are encodings of its single instructions. It is clear that any machine is encoded by at least one string, and it is easy to determine whether a string over $\{0, 1, \#, |\}$ encodes a Turing machine.

We prove completeness of a restricted version of the halting problem. The language $\mathcal{L}$ is the class of words of the form $(x\#M\#0^T)$, where $x \in \{0, 1\}^*$, M is an encoding of a deterministic Turing machine with input alphabet $\{0, 1, \#\}$, T is a natural number, and there is a string $w \in \{0, 1\}^*$ (the witness), such that the machine M *accepts* the string $x\#w$ in at most T steps.

These language are clearly in **NP**, because it is easy to find out whether the string is in the form $(x\#M\#0^T)$, to extract x, M, and T, guess w, and simulate computation of M with the input $x\#w$ for T steps. Since the time bound T is given in unary notation, this algorithm runs clearly in worst case time bounded by a polynomial in the length of the input.

We put

$$\nu(x) = \frac{2 \cdot 2^{-|u|} 4^{-|v|}}{(|x| + 2)(|x| + 1)|x|(|x| - 1)}$$

if $x = u\#v\#0^T$, $u \in \{0, 1\}^*$, $v \in \{0, 1, \#, |\}$, $\nu(x) = 0$ otherwise. Note that the function is very "wild" on any class of strings of the same length, because the halting problem is algorithmically unsolvable, and therefore there is no computable bound to T in terms of $|u|$, which means that u-part of long strings from $\mathcal{L}_0$ can be short or long, i.e. likely or unlikely.

We have to show that ν is well defined and P-distributed. Let $s_{n,i,j}$ be the sum of $\nu(x)$ over all $x = u\#v\#O^T$, such that $u \in \{0, 1\}^*$, $v \in \{0, 1, \#, |\}$, $|u| = i$, $|v| = j$, $|x| = n$, $0 \le i, j$, $i + j \le n - 2$. Then obviously $s_{n,i,j} = 2((n + 2)(n + 1)n(n - 1))^{-1}$, and since there are $n(n - 1)/2$ possible combinations of nonnegative i, j such that $i + j \le n - 2$,

$$\sum_{|x|=n} \nu(x) = \frac{1}{(n + 1)(n + 2)} = \frac{1}{n + 1} - \frac{1}{n + 2},$$

$$\nu^*(O^{|x|}) = \sum_{|x|\le n} \nu(x) = 1 - \frac{1}{n+2}, \qquad \sum_x \nu(x) = 1,$$

and it is not difficult to evaluate $\nu^*(x)$ for general x.

The proof of completeness of this problems follows easily from the next technical lemma:

Lemma 2.2.1 *Let μ be a P-distributed probability function on $\{0,1\}^*$. Then there exist two functions $f, g : \{0,1\}^* \to \{0,1\}^*$ and a polynomial h such that*

- *$g(f(x)) = x$ for each $x \in \{0,1\}^*$,*
- *$f(x)$ can be computed in time $h(x)$,*
- *$g(x)$ is computable in time $O(h(|x| + |g(x)|))$, and*
- *$\mu(x) \le 4 \cdot 2^{-|f(x)|}$ for each $x \in \{0,1\}^*$.*

We will prove the lemma later, first we will show how it implies the following

Theorem 2.2.2 *The above defined halting problem $(\mathcal{L}, \nu)$ is complete for the class of P-distributed* **DNP** *problems.*

Proof: We already know that $\mathcal{L} \in$ **NP**, and ν^* can be computed in polynomial time.

Suppose that (L, μ) is a P-distributed **DNP** problem. There exists a polynomial time bounded Turing machine $\mathcal{M}$, and a polynomial p such that $x \in L$ iff $\mathcal{M}$ accepts $x\#w$ for some $w \in \{0,1\}^*$, $|w| \le p(|x|$. Let f, g be functions from Lemma 2.2.1, and M be an encoding of an arbitrary Turing machine that works in the following way: given a word $x\#w$, $x, w \in \{0,1\}$, compute $g(w)$, and then apply $\mathcal{M}$ to $g(x)\#w$. Let k be a polynomial such that if the machine with description M is applied to $f(w)$, it finishes in at most $k(|w|)$ steps. Such k exists, because the running time of the machine is $O(h(|f(w)| + |w|)) + O(p(|w|)$. It is clear that $w \in \mathcal{L}$ iff $\mathcal{M}$ accepts w in time $p(|w|)$, which is if and only if the machine encoded by M accepts $f(w)$ in time $k(|w|)$.

Put $\Phi(w) = (f(w)\#M\#0^{k(|w|)})$. The function Φ clearly verifies the conditions of Definition 2.1.8. ♣

Now we give the proof of Lemma 2.2.1:

Given a string $x = x_1 \dots x_k \in \{0,1\}^*$, denote by $\xi(x)$ the rational number with binary notation $0.x_1 \dots x_k$, i.e. the number $\sum_{i=1}^{k} 2^{-i} x_i$, by I_x the half-open interval $[\mu^*(x), \mu^*(x^+))$. Note that $x_1 < x_2$, $|x_1| = |x_2|$ implies $\xi(x_1) < \xi(x_2)$. If $\mu(x) \le 2^{-|x|+1}$, then put $f(x) = 0x$, otherwise $f(x) = 1y$, where y is the shortest string over $\{0,1\}$ such that $\xi(y) \in I_x$. $g(0y) = y$, $g(1y)$ is the unique string z such that $\xi(y) \in I_z$.

First we have to show that f is well defined. Note that if y is the shortest string such that $\xi(y) \in I_x$, and y is nonempty, that y ends by 1 (otherwise the rightmost 0 could be deleted to get shorter string with the same value of $\xi(y)$). Suppose that there are two shortest strings $y_1 < y_2$ such that $\xi(y_1), \xi(y_2) \in I_x$. Then they are both nonempty, hence ending by 1, and y_3 obtained from y_2 by deleting the trailing 1 verifies $y_1 < y_3 < y_2$, $\xi(y_1) < \xi(y_3) < \xi(y_2)$, $y_3 \in I_x$, a contradiction, because y_3 is shorter than y_1, y_2.

It is clear that $g(f(x)) = x$ for all x. The inequality of the lemma follows from the observation that the width of I_x is equal to $\mu(x)$, and if it is large, then I_x contains a number with short binary expansion. More precisely, there are four possibilities

- if $f(x) = 0x$, then $\mu(x) \le 2^{-|x|+1}$, $\mu(x) \le 4 \cdot 2^{-(1+|x|)} = 4 \cdot 2^{-|f(x)|}$,
- if $f(x) = 1$, then $\mu(x) \le 1 = 2 \cdot 2^{-|f(x)|}$,
- if $f(x) = 1^k$, $k > 1$, then $\xi(1^{k-2}) \notin I_x$, because it is too short, and hence $\xi(1^{k-2}) \le \mu^*(x)$, since $1^{k-2} < 1^{k-1}$, and $\xi(1^{k-1}) \in I_x$, and therefore

$$\mu(x) = \mu^*(x^+) - \mu^*(x) \le 1 - \mu^*(x) \le 1 - \xi(1^{k-2}) =$$

$$= 2^{-(k-2)} = 4 \cdot 2^{-k} = 4 \cdot 2^{-|f(x)|},$$

- if $f(x) = 1y$, y nonempty string containing 0, then $y = z1$, $|z| = |z^+| < |y|$, and therefore $\xi(z), \xi(z^+) \notin I_x$,

$$\xi(z) < \mu^*(x) \le \xi(y) < \mu^*(x) + \mu(x) \le \xi(z^+),$$

$$\mu(x) \le \xi(z^+) - \xi(z) = 2^{-|z|} = 4 \cdot 2^{-|f(x)|},$$

Finally we have to show that both f and g are easily computable, provided μ^* is. f can be computed as follows

begin $y :=$ the empty string;
while $\xi(y) < \mu^*(x)$ **do if** $\xi(y1) < \mu^*(x^+)$ **then** $y := y1$;
end;

It is clear that we have to put $g(0y)$. In order to compute $g(1y)$, let us note that $1^{|z|-1} < z \leq 1^{|z|}$ for each z, and therefore if $\xi(y) \in I_z$, then $\mu^*(1^{|z|-1}) < \xi(y) \leq \mu^*(1^{|z|})$, and therefore $g(1y)$ can be obtained as follows

begin $k := 0$;
while $\mu^*(1^k) < \xi(y)$ **do** $k := k + 1$;
$z :=$ the empty string;
for $i := 1$ **to** k **do**
  if $\mu^*(z) < \xi(y)$ **then** $z := z1$ **else** $z := z0$;
end;

Now it is clear that it is sufficient to be able to approximate $\mu^*(x)$ for arbitrary x with precision 2^{-k} in time $O(|x| + k)$.

2.3 Samplable problems

We first show that there is a problem complete for the class of samplable problems. The proof is based on the existence of the largest samplable probability function.

Lemma 2.3.1 *Let μ be a samplable probability function. Then there exists a polynomially computable function f such that $\mu = f(\mathcal{U})$, where $\mathcal{U}$ is the uniform probability function.*

Proof:
♣

Theorem 2.3.2 *There is a samplable distribution μ_0, such that $\mu \leq \mu_0$ for each samplable function μ.*

Proof: Given a Turing machine M and a polynomial p, let $M|p$ be a machine that simulates M, but as a result of the computation we consider the contents of the output tape after $p(|w|)$ steps of the simulated machine.

It is easy to find an encoding of pairs (M, p) by strings from $\{0, 1\}^*$ such that

- each pair (M, p) has at least one encoding,
- it can be decided in polynomial time whether a given string encodes (M, p) for some Turing machine M and polynomial p,
- if w encodes (M, p), then M and p can be found in polynomial time.

Now, given a natural number $i > 0$, let w_i be the string such that $1w_i$ is the binary representation of i. If w_i is an encoding of no (M, p), then μ_i is the uniform probability function, otherwise μ_i is the samplable distribution, generated by the machine $M|p$ and the uniform probability function.

Put

$$\mu_0(x) = \sum_{i=1}^{\infty} \frac{\mu_i(x)}{2^i}.$$

It is clear that $\mu_0(x)$ can be evaluated with precision $\varepsilon > 0$ in time polynomial in $|x|$ and $\log 1/\varepsilon$: it is sufficient to evaluate $\mu_i(x)$ for $1 \le i$, $2^i \le 1\varepsilon$ and add the terms together.

Moreover, $\mu_i(x)2^{-i} \le \mu_0(x)$ for every fixed i, which implies $\mu_i \le \mu_0$. If μ is a samplable probability function, then $\mu = f(\nu)$, where f is a function computable in polynomial time, and ν is a P-distributed probability function. In view of the preceding lemma, $\nu = g(\mathcal{U})$, where g is computable in polynomial time, and $\mathcal{U}$ is the uniform distribution. The composition of g and f can be computed by some Turing machine M in time bounded by a polynomial p. Let w be the above described encoding of the pair (M, p), i be the number with binary representation $1w$. Then $\mu = \mu_i$, and therefore $\mu \le \mu_0$. ♣

Theorem 2.3.3 *There is a samplable distribution μ_0, such that (L, μ_0) is complete in the class of samplable problems for any* **NP***-complete problem L.*

Proof: Let μ_0 be the distribution from the previous theorem, L be an **NP**-complete problem. Suppose that (K, μ) is a samplable problem, $\mu = f(\nu)$, where f is polynomially computable, ν is P-distributed, and let g be a polynomially computable reduction of K to L. Then g is also a reduction of (K, μ) to the samplable problem $(L, h(\nu))$, where $h(x) = g(f(x))$, and the identity mapping reduces $(L, h(\nu))$ to (L, μ_0). ♣

Samplable probability functions represent much larger class than P-distributed ones. Let us suppose that a function g is defined as follows: if

$x = y\#w$, where y, w, resp. represent (under some encoding) a graph G, and a Hamiltonian cycle in G, then $g(x) = y$, otherwise $g(x)$ is the empty string. In this case the task to determine $\mu(y)$ for $\mu = g(\mathcal{U})$, where $\mathcal{U}$ is the uniform probability function, implies determining the number of Hamiltonian cycles in the graph represented by y. Therefore the problem to determine the value of a samplable probability function is $\#\mathbf{P}$-hard (and even the problem if $\mu(x) \neq 0$ is **NP**-hard), as opposed to polynomial time computability of both μ and μ^* for P-distributed functions. It might therefore seem that using samplable distributions one can construct distributed problems, that are more difficult that any P-distributed one. The aim of the paragraph is to show that this is not true: there are P-distributed problems complete for the class of samplable problems. In other words, samplable problems are not more difficult than problems with P-distributed probability functions. However, we will need to generalize the notion of reduction of problems, as defined in Definition 2.1.8.

The reduction of a source problem (L_1, μ_1) to a target problem (L_2, μ_2) that will be studied in this paragraph is a randomized algorithm, and the notion is more general than the reduction defined in Lemma 2.1.8. The reduction consists of two polynomial time computable functions ϕ, ψ that are used as follows: given a string x of the length k, concatenate it with a random string r chosen uniformly among 0,1-strings of the length $p(k)$, where p is a polynomial, compute $\phi(xr)$, try to find a witness w of membership $\phi(xr) \in L_2$, and if w is found, then check if $\psi(w)$ is a witness of membership $x \in L_1$. It is obvious that if $x \notin L_1$, then either $\phi(xr) \notin L_2$ or $\psi(w)$ is not a witness for x. Our aim is to construct ϕ, ψ so that the witness of $x \in L_1$ is found with reasonably large probability.

We will require that for each string x there is a set $G_x \subset \{0,1\}^{p(k)}$ such that

- $|G_x| \geq \frac{2^{p(k)}}{q(k)}$ for some polynomial q (i.e. it is sufficiently likely that r hits G_x),
- if r is "good", i.e. belongs to G_x, then $x \in L_1$ implies that $\phi(xr) \in L_2$ and its witness maps to a witness of x by ψ
- the distribution of strings $\phi(xr)$, x good, is consistent with μ_2.

Note that we do not need to require that $r \in G_x$ is easy to test; it is not necessary to know when a random r is good, provided there is a guarantee that this happens often.

Now, let us give the formal definition of a generalized reduction

Definition 2.3.4 *Let $P_1 = (L_1, \mu_1)$ and $P_2 = (L_2, \mu_2)$ be distributed problems. We say that P_1* **R***-reduces to P_2, if there are polynomially computable functions ϕ, ψ, polynomials p, q, h and sets $G_x \subset \{0,1\}^{p(|x|)}$, $x \in \{0,1\}^*$, such that*

$$|G_x| \geq q^{-1}(|x|),$$

if $x \in L_1$, $\mu_1(x) > 0$, and $r \in G_x$, then $\phi(xr) \in L_2$, and if w is a witness of $\phi(xr)$, then $\psi(w)$ is a witness of x, and

$$\sum_{(z,r)\in S(x)} \mu_1(z) 2^{-p(|z|)} \leq \mu_2(x) h(|x|),$$

where $S(x)$ is the set of all pairs (z, r) such that $r \in G_z$ and $\phi(zr) = x$.

We say that a problem A is **R***-hard for a class $\mathcal{C}$, if any problem in $\mathcal{C}$* **R***-reduces to A, and we say that A is* **R***-complete for $\mathcal{C}$, if $A \in \mathcal{C}$ and A is hard for $\mathcal{C}$.*

Lemma 2.3.5 *If μ is a samplable distribution, then there is a polynomially computable length preserving function $h : \{0,1\}^* \to \{0,1\}^*$ such that $\mu = h(\mathcal{U})$, where $\mathcal{U}$ is the uniform distribution.*

Proof:

♣

Theorem 2.3.6 *Any problem complete for the class of P-distributed problems is* **R***-complete for the class of samplable problems.*

Proof: It is clear that a composition of an **R**-reduction and reduction is an **R**-reduction. Therefore it is sufficient to find an **R**-reduction of any samplable problem to some P-distributed problem.

Let $P = (L, \mu_0)$ be a samplable problem. In view of Lemma 2.3.5 there is a polynomially computable function h such that $|h(x)| = |x|$ for each x, and $\mu_0 = h(\mathcal{U})$, where $\mathcal{U}$ is the uniform distribution. Let $f : \{0,1\}^* \to \{0,1\}^*$ be a polynomially computable function such that $x \in L$ iff $x = f(w)$ for some w.

Suppose that $H_{i,j}$ is a family of functions from $\{0,1\}^i$ into $\{0,1\}^j$. $H_{i,j}$ will be chosen so that they are universal hashing families in the sense of Carter and Wegman [3]. Especially we will suppose that there are constants $0 < c_1, c_2$ such that

- if $S \subset \{0,1\}^j$, $|S| \geq 2^{j-i}$, then the probability that S intersects with $\{h(x) | x \in \{0,1\}^i\}$, where $h \in H_{i,j}$ is chosen randomly with uniform probability, is at least c_1,
- given $x \in \{0,1\}^i$, if $h \in H_{i,j}$ chosen randomly with uniform probability, then the probability that $h(x) \neq h(y)$ for a arbitrary $y \neq x$ is at least c_2,

Example of such a class are functions $h(x) = ((ax + b) \bmod P) \bmod 2^j$, $a = 1, \ldots, P-1$, $b = 0, \ldots, P-1$, where P is a prime such that $2^i \leq P \leq 2^{i+1}$.

Our target problem is defined as follows: its instances are represented by tuples (n, k, y, h_1, h_2), where $k \leq n$ are integers, y is a string of length k, $h_1 \in H_{k,n}$, $h_2 \in H_{n,k}$ such that for some x, w and s, $|s| = k$, $f(w) = h(h_1(r))$, and $h_2(f(w)) = y$, see picture.

$$
\begin{tikzcd}
w \arrow[ddr, "f"'] & & s \arrow[d, "h_1"] \\
 & & z \arrow[dl, "h"] \\
 & x \arrow[d, "h_2"] & \\
 & y &
\end{tikzcd}
$$

An $\mathbf{R}$-reduction will be constructed as follows: given a string x, put $n = |x|$, choose a random $1 \leq k \leq n$, a random $h_1 \in H_{k,n}$, a random $h_2 \in H_{n,k}$, and output (n, k, y, h_1, h_2), where $y = h_2(x)$.

A witness for $(n, k, y = h_2(x), h_1, h_2) \in L_2$, are strings x, w, and s described above. We will take w as a temptative witness of $x \in L_1$.

Let $\mu_0(x) > 0$ (i.e. $h^{-1}(x) \neq \emptyset$). We will say that a randomization given by k, h_1, h_2 is *good*, if

1. $2^{n-k-1} \le |h^{-1}(x)| \le 2^{n-k}$,

2. $h^{-1}(x)$ intersects with the range of h_1, or, equivalently, $x = h(h_1(s))$ for some s,

3. $h_2(x) = h_2(z)$ for no $z \ne x$,

If $x \in L_1$ and k, h_1, h_2 are good, then $x = f(w)$ for some w, and $x = h(h_2(s))$ for some s, which implies that x, w, s represent a witness for $(n, k, h_2(x), h_1, h_2)$. If another witness x', w', s' for $(n, k, h_2(x), h_1, h_2)$ is found, then $x = x'$, and therefore w' is another witness for $x = x' = f(w') \in L_1$. Therefore in this case any witness for the image of x gives a witness for x itself.

Given a string x, if k, h_1, h_2 are good, then $(n, k, h_2(x), h_1, h_2)$ can not be an image of another string, see condition (3). Therefore the probability that $y = h_2(x)$, where x is generated by $\mu = h(\mathcal{U})$, is equal to

$$|h^{-1}(x)|2^{-n} \le 2^{n-k}2^{-n} = 2^{-k},$$

but 2^{-k} is the probability that a uniformly generated element of $\{0,1\}^k$ is equal to x. Therefore the distribution of the original and the target problems are consistent with our randomized reduction, restricted to good random strings.

Finally, we will prove that k, h_1, h_2 are good with probability $\Omega(1/n)$:

- k verifies (1) with probability $1/n$.

- Since probabilities that conditions (2), (3) hold are independent, it follows from properties of classes $H_{i,j}$ that they are fulfilled with probability at least $c_1 c_2$.

- With probability at least c_2, h_2 is such that if $(n, k, h_2(x), h_1, h_2)$ is a "yes" instance of the target problem, and x', w', s' is another witness for that, then $x = x'$, i.e. w' is a witness of $x \in L_1$, too.

♣

References

[1] Feller, Introduction into Probability Theory

[2] Ben-David, S., Chor, B., Goldreich, O., and Luby, M., On the theory of average case complexity, 21st *Symposium on Theory of Computing*, 1989, 204-216.

[3] Carter, J., and Wegman, M., Universal classes of hash functions, *J. of Computer and System Sci.*, 18 (1979), 143-154.

[4] Gurevich, Y., Complete and incomplete randomized NP problems, 28th *Symposium on Foundations of Computer Science*, 1987, 111-117.

[5] Gurevich, Y., Matrix decomposition problem is complete for average case, 31st *Symposium on Foundations of Computer Science*, 1990, 802-811.

[6] Gurevich, Y., Average case completeness, *J. of Computer and System Sci.*, 42 (1991), 346-398.

[7] Impagliazzo, R., and Levin, L.A., No better ways to generate hard NP instances than picking uniformly at random, 31st *Symposium on Foundations of Computer Science*, 812-821.

[8] Johndon, D.S., The NP-Complete Column - an ongoing guide, *J. of Algorithms*, 5(1984), 284-299.

[9] Levin, L.A., Average case complete problems, 16th *Symposium on Theory of Computing*, 1984, 465.

[10] Levin, L.A., Average case complete problems, *SIAM J. on Computing*, 15(1986), 285-286.

[11] A theory of learning simple concepts under simple distributions and average case complexity for the universal distribution, 30th *Symposium on Foundations of Computer Science*, 1989.

[12] Venkatesan, R., Levin, L.A., Random instances of a graph coloring problem are hard, 20th *Symposium on Theory of Computing*, 1988, 217-222.

[13] Venkatesan, R., Rajagopalan, S., Average case intractability of matrix and Diofantine problems, 24th *Symposium on Theory of Computing*, 1992, 632-642.