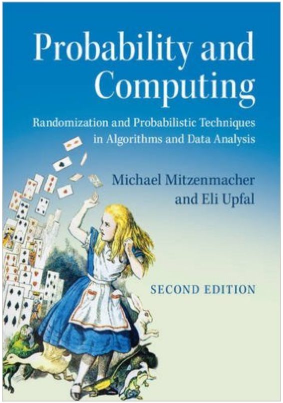
CS155/254: Probabilistic Methods in Computer Science

Chapter 6: The Probabilistic Method



The Probabilistic Method

Paul Erdös 1913 - 1996

Let X be a random variable defined on a discrete sample space $(\Omega, Pr(\cdot))$.

- Assume $X : \Omega \to \{0,1\}$.
  If $Pr(X=1) > 0$, then there is $\omega \in \Omega$ such that $X(\omega) = 1$.
  Example: Ω is a collection of graphs, $X(\omega) = 1$ if graph ω is connected.
- If $E[X] = c$, then there are $\omega_1, \omega_2 \in \Omega$ such that $X(\omega_1) \leq c$ and $X(\omega_2) \geq c$.
  Example: Assume that X is a gain in a sequence of games. There are sequences that yield $\leq c$ and $\geq c$

Probability Argument Example: Edge Coloring

$K_k =$ the complete graph on k vertices (a clique of k nodes) - K_k has all the $\binom{n}{2}$ edges between its k vertices.

Can we color the edges of K_{1000} with two colors so that no K_{20} is edge monochromatic?

Theorem

If $\binom{n}{k}2^{-\binom{k}{2}+1} < 1$, then it is possible to color the edges of K_n so that it has no monochromatic K_k subgraph.

Can we color the edges of K_{1000} with two colors so that no K_{20} is edge monochromatic?

Color the edges of K_{1000} randomly with two colors. The probability that at least one K_{20} is edge monochromatic is bounded by

$$\begin{aligned}\binom{1000}{20}2^{-\binom{20}{2}+1} &\leq \frac{1000^{20}}{20!}2^{-(20(20-1)/2)+1} \\ &\leq \frac{2^{10\cdot 20}}{20!}2^{-10(20-1)+1} \leq \frac{2^{10+1}}{20!} < 1.\end{aligned}$$

The probability that no K_{20} that is edge monochromatic is > 0.

Therefore, the space of all $2^{\binom{1000}{2}}$ coloring of the edges in K_{1000} has at least one assignment such that no K_{20} is edge monochromatic

Proof

Define a sample space:

- $\Omega =$ all $2^{\binom{n}{2}}$ coloring with two colors of all the edges in K_n.
- The probability of each coloring in Ω is $2^{-\binom{n}{2}}$.

This model is equivalent to coloring each edge independently with equal probabilities to the two colors.

For $i = 1, \ldots, \binom{n}{k}$, let A_i be the event that clique i is monochromatic. $\Pr(A_i) = 2^{-\binom{k}{2}+1}$. The probability that at least one K_k is monochromatic

$$\leq \Pr\left(\bigcup_{i=1}^{\binom{n}{k}} A_i\right) \leq \sum_{i=1}^{\binom{n}{k}} \Pr(A_i) = \binom{n}{k} 2^{-\binom{k}{2}+1} < 1,$$

$$\Pr\left(\bigcap_{i=1}^{\binom{n}{k}} \overline{A_i}\right) = 1 - \Pr\left(\bigcup_{i=1}^{\binom{n}{k}} A_i\right) > 0.$$

For $i = 1, \ldots, \binom{n}{k}$, let A_i be the event that clique i is monochromatic. $\Pr(A_i) = 2^{-\binom{k}{2}+1}$.

$$\Pr\left(\bigcap_{i=1}^{\binom{n}{k}} \overline{A_i}\right) = 1 - \Pr\left(\bigcup_{i=1}^{\binom{n}{k}} A_i\right) > 0.$$

Thus, there is a coloring $\omega \in \Omega$ of the $\binom{n}{2}$ edges with the required property.

Theorem

If $\binom{n}{k} 2^{-\binom{k}{2}+1} < 1$, then it is possible to color the edges of K_n so that it has no monochromatic K_k subgraph.

The Expectation Argument: Large Cut-Set in a Graph.

Theorem

Given any graph $G = (V, E)$ with n vertices and m edges, there is a partition of V into two disjoint sets A and B such that at least $m/2$ edges connect a vertex in A to a vertex in B.

Proof.

Construct sets A and B by randomly assign each vertex to one of the two sets.
The probability that a given edge connect A to B is $1/2$, thus the expected number of such edges is a random partition is $m/2$.
Thus, there exists such a partition. $\square$

How do we find such a partition?

Derandomization using Conditional Expectations

$C(A, B) =$ number of edges connecting A to B.
If A, B is a random partition $E[C(A, B)] = \frac{m}{2}$.
Algorithm:

1. Let $v_1, v_2, \ldots, v_n$ be an arbitrary enumeration of the vertices.
2. Let x_i be the set where v_i is placed ($x_i \in \{A, B\}$).
3. For $i = 1$ to n do:
    1. Place v_i such that

$$E[C(A, B) \mid x_1, x_2, \ldots, x_i]$$
$$\geq E[C(A, B) \mid x_1, x_2, \ldots, x_{i-1}] \geq m/2.$$

Conditional Expectation

Definition

$$E[Y \mid Z = z] = \sum_y y \Pr(Y = y \mid Z = z),$$

where the summation is over all y in the range of Y.

Lemma

For any random variables X and Y,

$$E[X] = \sum_y \Pr(Y = y) E[X \mid Y = y],$$

where the sum is over all values in the range of Y.

Lemma

For all $i = 1, \ldots, n$ *there is an assignment of* v_i *such that*

$$E[C(A, B) \mid x_1, x_2, \ldots, x_i] \\ \geq E[C(A, B) \mid x_1, x_2, \ldots, x_{i-1}] \geq m/2.$$

Proof.

By induction on i.

For $i = 1$, $E[E[C(A,B) \mid X_1]] = E[C(A,B)] = m/2$

For $i > 1$, if we place v_i randomly in one of the two sets,

$$\begin{aligned} & E[C(A,B) \mid x_1, x_2, \ldots, x_{i-1}] \\ = \; & \frac{1}{2} E[C(A,B) \mid x_1, x_2, \ldots, x_i = A] + \frac{1}{2} E[C(A,B) \mid x_1, x_2, \ldots, x_i = B] \\ = \; & m/2. \end{aligned}$$

$$\begin{aligned} & \max\left(E[C(A,B) \mid x_1, x_2, \ldots, x_i = A], E[C(A,B) \mid x_1, x_2, \ldots, x_i = B]\right) \\ \geq \; & E[C(A,B) \mid x_1, x_2, \ldots, x_{i-1}] \\ \geq \; & m/2 \end{aligned}$$

$\square$

How do we compute

$$\max(E[C(A,B) \mid x_1, x_2, \ldots, x_i = A], E[C(A,B) \mid x_1, x_2, \ldots, x_i = B])$$
$$\geq E[C(A,B) \mid x_1, x_2, \ldots, x_{i-1}] \geq m/2$$

We just need to consider edges between v_i and $v_1, \ldots, v_{i-1}$.
Simple Algorithm:

1. Place v_1 arbitrarily.
2. For $i = 2$ to n do
    1. Place v_i in the set with smaller number of neighbors.

Sample and Modify

An *independent set* in a graph G is a set of vertices with no edges between them.
Finding the largest independent set in a graph is an NP-hard problem.

Theorem

Let $G = (V, E)$ be a graph on n vertices with $dn/2$ edges. Then G has an independent set with at least $n/2d$ vertices.

Algorithm:

1. Delete each vertex of G (together with its incident edges) independently with probability $1 - 1/d$.
2. For each remaining edge, remove it and one of its adjacent vertices.

$X =$ number of vertices that survive the first step of the algorithm.

$$E[X] = \frac{n}{d}.$$

$Y =$ number of edges that survive the first step.
An edge survives if and only if its two adjacent vertices survive.

$$E[Y] = \frac{nd}{2}\left(\frac{1}{d}\right)^2 = \frac{n}{2d}.$$

The second step of the algorithm removes all the remaining edges, and at most Y vertices.
Size of output independent set:

$$E[X - Y] = \frac{n}{d} - \frac{n}{2d} = \frac{n}{2d}.$$

Sets with Distinct Sums

A set $S = \{x_1, \ldots x_k\} \subset \{1, \ldots, n\}$ has the *distinct sums property if for any* $S_1, S_2 \subset S$, $S_1 \neq S_2$

$$\sum_{x_i \in S_1} x_i \neq \sum_{x_j \in S_2} x_j.$$

Theorem

Let $f(n)$ be the maximum size of a distinct sums set that is a subset of $\{1, \ldots, n\}$.

$$f(n) \leq \log_2 n + \frac{1}{2} \log \log n + O(1).$$

Simple Argument

Assume that $S = \{x_1, \ldots x_k\} \subset \{1, \ldots, n\}$ has the *distinct sums property.*

For $i = 1, \ldots, k$*, let* $Y_i \in \{0, 1\}$*.*

There are 2^n *assignments for* $Y_1, \ldots, Y_k$*, and for each assignment* $X = \sum_{i=1}^{k} x_i Y_i$ *must give a different value.*

There are no more than kn *possible different values.*

$$2^k \leq nk$$

$$k = \log_2 n + \log_2 k.$$

$$k \leq \log_2 n + \log \log n$$

Adding the Variance

Assume that $S = \{x_1, \ldots x_k\} \subset \{1, \ldots, n\}$ has the *distinct sums property.*

Define k *random random variable* $Pr(Y_i = 1) = Pr(Y_i = 0) = \frac{1}{2}$.

Let $X = \sum_{i=1}^{k} x_i Y_i$. *Then*

$$\mu = E[X] = \frac{1}{2}\sum_{i=1}^{k} x_i, \quad \text{and} \quad Var[X] = \frac{1}{4}\sum_{i=1}^{k} x_i^2 \leq \frac{n^2 K}{4}.$$

Applying Chebyschev's Inequality, for any $\lambda > 0$

$$Pr(|X - \mu| \geq \lambda \frac{n\sqrt{k}}{2}) \leq \frac{1}{\lambda^2}$$

$$Pr(|X-\mu| \le \lambda \frac{n\sqrt{k}}{2}) \ge 1 - \frac{1}{\lambda^2}$$

Since S has the distinct sums property, for any x, $Pr(X = x)$ is ether 2^{-k} or 0. Thus,

$$Pr(|X-\mu| \le \lambda \frac{n\sqrt{k}}{2}) \le 2^{-k}(\lambda n\sqrt{k} + 1),$$

$$1 - \frac{1}{\lambda^2} \le Pr(|X-\mu| \le \lambda \frac{n\sqrt{k}}{2}) \le 2^{-k}(\lambda n\sqrt{k} + 1),$$

$$n \ge \frac{2^k(1-\lambda^{-2}) - 1}{\lambda\sqrt{k}}$$

.

For $\lambda = \sqrt{3}$,

$$k \le \log_2 n + \frac{1}{2}\log\log n + O(1).$$

The First and Second Moment Methods

Theorem

For an integer random variable $X \geq 0$,

1. $Pr(X > 0) = Pr(X \geq 1) \leq E[X]$
2. $Pr(X = 0) \leq Pr(|X - E[X]| \geq E[X]) \leq \frac{Var[X]}{(E[X])^2} \leq \frac{E[X^2]}{(E[X])^2}$
3. $Pr(X \geq 1) \geq \frac{(E[X])^2}{E[X^2]}$

Proof: For $X \geq 0$ and integer:

1. $Pr(X \geq 1) \leq \sum_{i \geq 1} Pr(X \geq i) = E[X]$.
2. Chebyshev bound.
3. Using Cauchy–Schwarz inequality:

$\sum_{i=1}^{n} a_i b_i \leq \sqrt{\sum_{i=1}^{n} a_i^2} \sqrt{\sum_{i=1}^{n} b_i^2}$

$$E[X] = E[X \cdot 1_{X \geq 1}] \leq \sqrt{E[X^2]} \sqrt{Pr(X \geq 1)}$$

Application: Number of Isolated Nodes

Let $G_{n,p} = (V, E)$ be a random graph generated as follows:

- The graph has n nodes.
- Each of the $\binom{n}{2}$ pairs of vertices are connected by an edge with probability p independently of any other edge in the graph.

A node is isolated if it is adjacent to no edges.

If $p = 0$ all vertices are isolated (have no edges). If $p = 1$ no vertex is isolated. What can we say for $0 < p < 1$?

Application: Number of Isolated Nodes

Let $G_{n,p} = (V, E)$ be a random graph generated as follows:

- The graph has n nodes.
- Each of the $\binom{n}{2}$ pairs of vertices are connected by an edge with probability p independently of any other edge in the graph.

A node is isolated if it has no edges.

Theorem

For any function $w(n) \to \infty$

- *If $p = \frac{\log n - w(n)}{n}$, then with high probability the graph has isolated nodes.*
- *If $p = \frac{\log n + w(n)}{n}$, then with high probability the graph has no isolated nodes.*

High Probability = probability converging to 1 as $n \to \infty$

Proof

For $i = 1, \ldots, n$, let $X_i = 1$ if node i is isolated, otherwise $X_i = 0$.
Let $X = \sum_{i=1}^{n} X_i$.

$$E[X] = n(1-p)^{n-1}$$

For $p = \frac{\log n + w(n)}{n}$

$$E[X] = n(1-p)^{n-1} \leq e^{\log n - (n-1)p} \leq e^{-w(n)} \to 0$$

Thus, for $p = \frac{\log n + w(n)}{n}$,

$$Pr(X > 0) \leq E[X] \to 0$$

To use the second moment method we need to bound $Var[X]$.

$$Var[X_i] \leq E[X_i^2] = E[X_i] = (1-p)^{n-1}$$

$$Cov(X_i, X_j) = E[X_i X_j] - E[X_i]E[X_j] = (1-p)^{2n-3} - (1-p)^{2n-2}$$

$$\begin{aligned} Var[X] &\leq \sum_{i=1}^{n} Var[X_i] + \sum_{i \neq j} Cov(X_i, X_i) \\ &= n(1-p)^{n-1} + n(n-1)(1-p)^{2n-3} - n(n-1)(1-p)^{2n-2} \\ &= n(1-p)^{n-1} + n(n-1)p(1-p)^{2n-3} \end{aligned}$$

$$
\begin{aligned}
Var[X] &= \sum_{i=1}^{n} Var[X_i] + \sum_{i \neq j} Cov(X_i, X_i) \\
&= n(1-p)^{n-1} + n(n-1)p(1-p)^{2n-3}
\end{aligned}
$$

$$
\begin{aligned}
Pr(X=0) &\leq Pr(|X - E[X]| \geq E[X]) \leq \frac{Var[X]}{(E[X])^2} \\
&= \frac{n(1-p)^{n-1} + n(n-1)p(1-p)^{2n-3}}{n^2(1-p)^{2n-2}} \\
&= \left(1 - \frac{1}{n}\right) \frac{p}{1-p} + \frac{1}{n(1-p)^{n-1}}
\end{aligned}
$$

For $p = \frac{\log n - w(n)}{n}$,

$$\begin{aligned} Pr(X=0) &\leq \frac{Var[X]}{(E[X])^2} \\ &= \left(1 - \frac{1}{n}\right) \frac{p}{1-p} + \frac{1}{n(1-p)^{n-1}} \rightarrow 0 \end{aligned}$$

Since

$$n(1-p)^{n-1} \geq ne^{-p(n-1)}(1 - \frac{p^2}{n}) \geq \frac{1}{2}e^{w(n)}$$

We use: for $|X| \leq 1$

$$e^x \left(1 - \frac{x^2}{n}\right) \leq \left(1 + \frac{x}{n}\right)^n \leq e^x$$