Probability and Computing

Randomization and Probabilistic Techniques in Algorithms and Data Analysis

Second Edition

Michael Mitzenmacher   Eli Upfal

CAMBRIDGE
UNIVERSITY PRESS

CAMBRIDGE
UNIVERSITY PRESS

University Printing House, Cambridge CB2 8BS, United Kingdom

One Liberty Plaza, 20th Floor, New York, NY 10006, USA

477 Williamstown Road, Port Melbourne, VIC 3207, Australia

4843/24, 2nd Floor, Ansari Road, Daryaganj, Delhi - 110002, India

79 Anson Road, #06-04/06, Singapore 079906

Cambridge University Press is part of the University of Cambridge.

It furthers the University's mission by disseminating knowledge in the pursuit of education, learning, and research at the highest international levels of excellence.

www.cambridge.org
Information on this title: www.cambridge.org/9781107154889
10.1017/9781316651124

First published 2017

Printed in the United Kingdom by Clays, St Ives plc

A catalogue record for this publication is available from the British Library.

Library of Congress Cataloging in Publication Data
Names: Mitzenmacher, Michael, 1969– author. | Upfal, Eli, 1954– author.
Title: Probability and computing / Michael Mitzenmacher Eli Upfal.
Description: Second edition. | Cambridge, United Kingdom ;
New York, NY, USA : Cambridge University Press, [2017] |
Includes bibliographical references and index.
Identifiers: LCCN 2016041654 | ISBN 9781107154889
Subjects: LCSH: Algorithms. | Probabilities. | Stochastic analysis.
Classification: LCC QA274.M574 2017 | DDC 518/.1 – dc23
LC record available at https://lccn.loc.gov/2016041654

ISBN 978-1-107-15488-9 Hardback

Additional resources for this publication at www.cambridge.org/Mitzenmacher.

CHAPTER THREE

Moments and Deviations

In this and the next chapter we examine techniques for bounding the *tail distribution,* the probability that a random variable assumes values that are far from its expectation. In the context of analysis of algorithms, these bounds are the major tool for estimating the failure probability of algorithms and for establishing high probability bounds on their run-time. In this chapter we study Markov's and Chebyshev's inequalities and demonstrate their application in an analysis of a randomized median algorithm. The next chapter is devoted to the Chernoff bound and its applications.

3.1. Markov's Inequality

Markov's inequality, formulated in the next theorem, is often too weak to yield useful results, but it is a fundamental tool in developing more sophisticated bounds.

Theorem 3.1 [Markov's Inequality]: *Let X be a random variable that assumes only nonnegative values. Then, for all $a > 0$,*

$$\Pr(X \geq a) \leq \frac{\mathbf{E}[X]}{a}.$$

Proof: For $a > 0$, let

$$I = \begin{cases} 1 & \text{if } X \geq a, \\ 0 & \text{otherwise,} \end{cases}$$

and note that, since $X \geq 0$,

$$I \leq \frac{X}{a}. \tag{3.1}$$

Because I is a 0–1 random variable, $\mathbf{E}[I] = \Pr(I = 1) = \Pr(X \geq a)$.

Taking expectations in (3.1) thus yields

$$\Pr(X \geq a) = \mathbf{E}[I] \leq \mathbf{E}\left[\frac{X}{a}\right] = \frac{\mathbf{E}[X]}{a}.$$

■

For example, suppose we use Markov's inequality to bound the probability of obtaining more than $3n/4$ heads in a sequence of n fair coin flips. Let

$$X_i = \begin{cases} 1 & \text{if the } i\text{th coin flip is heads,} \\ 0 & \text{otherwise,} \end{cases}$$

and let $X = \sum_{i=1}^{n} X_i$ denote the number of heads in the n coin flips. Since $\mathbf{E}[X_i] = \Pr(X_i = 1) = 1/2$, it follows that $\mathbf{E}[X] = \sum_{i=1}^{n} \mathbf{E}[X_i] = n/2$. Applying Markov's inequality, we obtain

$$\Pr(X \geq 3n/4) \leq \frac{\mathbf{E}[X]}{3n/4} = \frac{n/2}{3n/4} = \frac{2}{3}.$$

3.2. Variance and Moments of a Random Variable

Markov's inequality gives the best tail bound possible when all we know is the expectation of the random variable and that the variable is nonnegative (see Exercise 3.16). It can be improved upon if more information about the distribution of the random variable is available.

Additional information about a random variable is often expressed in terms of its *moments*. The expectation is also called the *first moment* of a random variable. More generally, we define the moments of a random variable as follows.

Definition 3.1: *The* kth *moment of a random variable* X *is* $\mathbf{E}[X^k]$.

A significantly stronger tail bound is obtained when the second moment ($\mathbf{E}[X^2]$) is also available. Given the first and second moments, one can compute the *variance* and *standard deviation* of the random variable. Intuitively, the variance and standard deviation offer a measure of how far the random variable is likely to be from its expectation.

Definition 3.2: *The* variance *of a random variable* X *is defined as*

$$\mathbf{Var}[X] = \mathbf{E}[(X - \mathbf{E}[X])^2] = \mathbf{E}[X^2] - (\mathbf{E}[X])^2.$$

The standard deviation *of a random variable* X *is*

$$\sigma[X] = \sqrt{\mathbf{Var}[X]}.$$

The two forms of the variance in the definition are equivalent, as is easily seen by using the linearity of expectations. Keeping in mind that $\mathbf{E}[X]$ is a constant, we have

$$\begin{aligned} \mathbf{E}[(X - \mathbf{E}[X])^2] &= \mathbf{E}[X^2 - 2X\mathbf{E}[X] + \mathbf{E}[X]^2] \\ &= \mathbf{E}[X^2] - 2\mathbf{E}[X\mathbf{E}[X]] + \mathbf{E}[X]^2 \\ &= \mathbf{E}[X^2] - 2\mathbf{E}[X]\mathbf{E}[X] + \mathbf{E}[X]^2 \\ &= \mathbf{E}[X^2] - (\mathbf{E}[X])^2. \end{aligned}$$

If a random variable X is constant – so that it always assumes the same value – then its variance and standard deviation are both zero. More generally, if a random variable X takes on the value $k\mathbf{E}[X]$ with probability $1/k$ and the value 0 with probability

$1 - 1/k$, then its variance is $(k-1)(\mathbf{E}[X])^2$ and its standard deviation is $\sqrt{k-1}\mathbf{E}[X]$. These cases help demonstrate the intuition that the variance (and standard deviation) of a random variable are small when the random variable assumes values close to its expectation and are large when it assumes values far from its expectation.

We have previously seen that the expectation of the sum of two random variables is equal to the sum of their individual expectations. It is natural to ask whether the same is true for the variance. We find that the variance of the sum of two random variable has an extra term, called the covariance.

Definition 3.3: *The* covariance *of two random variables X and Y is*

$$\mathbf{Cov}(X,Y) = \mathbf{E}[(X - \mathbf{E}[X])(Y - \mathbf{E}[Y])].$$

Theorem 3.2: *For any two random variables X and Y,*

$$\mathbf{Var}[X+Y] = \mathbf{Var}[X] + \mathbf{Var}[Y] + 2\,\mathbf{Cov}(X,Y).$$

Proof:

$$\begin{aligned}
\mathbf{Var}[X+Y] &= \mathbf{E}[(X+Y-\mathbf{E}[X+Y])^2] \\
&= \mathbf{E}[(X+Y-\mathbf{E}[X]-\mathbf{E}[Y])^2] \\
&= \mathbf{E}[(X-\mathbf{E}[X])^2 + (Y-\mathbf{E}[Y])^2 + 2(X-\mathbf{E}[X])(Y-\mathbf{E}[Y])] \\
&= \mathbf{E}[(X-\mathbf{E}[X])^2] + \mathbf{E}[(Y-\mathbf{E}[Y])^2] + 2\mathbf{E}[(X-\mathbf{E}[X])(Y-\mathbf{E}[Y])] \\
&= \mathbf{Var}[X] + \mathbf{Var}[Y] + 2\,\mathbf{Cov}(X,Y). \qquad \blacksquare
\end{aligned}$$

The extension of this theorem to a sum of any finite number of random variables is proven in Exercise 3.14.

The variance of the sum of two (or any finite number of) random variables does equal the sum of the variances when the random variables are independent. Equivalently, if X and Y are independent random variables, then their covariance is equal to zero. To prove this result, we first need a result about the expectation of the product of independent random variables.

Theorem 3.3: *If X and Y are two independent random variables, then*

$$\mathbf{E}[X \cdot Y] = \mathbf{E}[X] \cdot \mathbf{E}[Y].$$

Proof: In the summations that follow, let i take on all values in the range of X, and let j take on all values in the range of Y:

$$\begin{aligned}
\mathbf{E}[X \cdot Y] &= \sum_i \sum_j (i \cdot j) \cdot \Pr((X=i) \cap (Y=j)) \\
&= \sum_i \sum_j (i \cdot j) \cdot \Pr(X=i) \cdot \Pr(Y=j) \\
&= \left(\sum_i i \cdot \Pr(X=i)\right)\left(\sum_j j \cdot \Pr(Y=j)\right) \\
&= \mathbf{E}[X] \cdot \mathbf{E}[Y],
\end{aligned}$$

where the independence of X and Y is used in the second line. ■

Unlike the linearity of expectations, which holds for the sum of random variables whether they are independent or not, the result that the expectation of the product of two (or more) random variables is equal to the product of their expectations does not necessarily hold if the random variables are dependent. To see this, let Y and Z each correspond to fair coin flips, with Y and Z taking on the value 0 if the flip is heads and 1 if the flip is tails. Then $\mathbf{E}[Y] = \mathbf{E}[Z] = 1/2$. If the two flips are independent, then $Y \cdot Z$ is 1 with probability $1/4$ and 0 otherwise, so indeed $\mathbf{E}[Y \cdot Z] = \mathbf{E}[Y] \cdot \mathbf{E}[Z]$. Suppose instead that the coin flips are dependent in the following way: the coins are tied together, so Y and Z either both come up heads or both come up tails together. Each coin considered individually is still a fair coin flip, but now $Y \cdot Z$ is 1 with probability $1/2$ and so $\mathbf{E}[Y \cdot Z] \neq \mathbf{E}[Y] \cdot \mathbf{E}[Z]$.

Corollary 3.4: *If X and Y are independent random variables, then*

$$\mathbf{Cov}(X, Y) = 0$$

and

$$\mathbf{Var}[X + Y] = \mathbf{Var}[X] + \mathbf{Var}[Y].$$

Proof:

$$\begin{aligned}\mathbf{Cov}(X, Y) &= \mathbf{E}[(X - \mathbf{E}[X])(Y - \mathbf{E}[Y])] \\ &= \mathbf{E}[X - \mathbf{E}[X]] \cdot \mathbf{E}[Y - \mathbf{E}[Y]] \\ &= 0.\end{aligned}$$

In the second equation we have used the fact that, since X and Y are independent, so are $X - \mathbf{E}[X]$ and $Y - \mathbf{E}[Y]$ and hence Theorem 3.3 applies. For the last equation we use the fact that, for any random variable Z,

$$\mathbf{E}[(Z - \mathbf{E}[Z])] = \mathbf{E}[Z] - \mathbf{E}[\mathbf{E}[Z]] = 0.$$

Since $\mathbf{Cov}(X, Y) = 0$, we have $\mathbf{Var}[X + Y] = \mathbf{Var}[X] + \mathbf{Var}[Y]$. ■

By induction we can extend the result of Corollary 3.4 to show that the variance of the sum of any finite number of independent random variables equals the sum of their variances.

Theorem 3.5: *Let $X_1, X_2, \dots, X_n$ be mutually independent random variables. Then*

$$\mathbf{Var}\left[\sum_{i=1}^{n} X_i\right] = \sum_{i=1}^{n} \mathbf{Var}[X_i].$$

3.2.1. *Example: Variance of a Binomial Random Variable*

The variance of a binomial random variable X with parameters n and p can be determined directly by computing $\mathbf{E}[X^2]$:

$$\begin{aligned}
\mathbf{E}[X^2] &= \sum_{j=0}^{n} \binom{n}{j} p^j (1-p)^{n-j} j^2 \\
&= \sum_{j=0}^{n} \frac{n!}{(n-j)!\,j!} p^j (1-p)^{n-j} ((j^2 - j) + j) \\
&= \sum_{j=0}^{n} \frac{n!\,(j^2 - j)}{(n-j)!\,j!} p^j (1-p)^{n-j} + \sum_{j=0}^{n} \frac{n!\,j}{(n-j)!\,j!} p^j (1-p)^{n-j} \\
&= n(n-1)p^2 \sum_{j=2}^{n} \frac{(n-2)!}{(n-j)!\,(j-2)!} p^{j-2} (1-p)^{n-j} \\
&\quad + np \sum_{j=1}^{n} \frac{(n-1)!}{(n-j)!\,(j-1)!} p^{j-1} (1-p)^{n-j} \\
&= n(n-1)p^2 + np.
\end{aligned}$$

Here we have simplified the summations by using the binomial theorem. We conclude that

$$\begin{aligned}
\mathbf{Var}[X] &= \mathbf{E}[X^2] - (\mathbf{E}[X])^2 \\
&= n(n-1)p^2 + np - n^2 p^2 \\
&= np - np^2 \\
&= np(1-p).
\end{aligned}$$

An alternative derivation makes use of independence. Recall from Section 2.2 that a binomial random variable X can be represented as the sum of n independent Bernoulli trials, each with success probability p. Such a Bernoulli trial Y has variance

$$\mathbf{E}[(Y - \mathbf{E}[Y])^2] = p(1-p)^2 + (1-p)(-p)^2 = p - p^2 = p(1-p).$$

By Theorem 3.5, the variance of X is then $np(1-p)$.

3.3. Chebyshev's Inequality

Using the expectation and the variance of the random variable, one can derive a significantly stronger tail bound known as Chebyshev's inequality.

Theorem 3.6 [Chebyshev's Inequality]: *For any* $a > 0$,

$$\Pr(|X - \mathbf{E}[X]| \geq a) \leq \frac{\mathbf{Var}[X]}{a^2}.$$

Proof: We first observe that

$$\Pr(|X - \mathbf{E}[X]| \geq a) = \Pr((X - \mathbf{E}[X])^2 \geq a^2).$$

Since $(X - \mathbf{E}[X])^2$ is a nonnegative random variable, we can apply Markov's inequality to prove:

$$\Pr((X - \mathbf{E}[X])^2 \geq a^2) \leq \frac{\mathbf{E}[(X - \mathbf{E}[X])^2]}{a^2} = \frac{\mathbf{Var}[X]}{a^2}. \qquad \blacksquare$$

The following useful variants of Chebyshev's inequality bound the deviation of the random variable from its expectation in terms of a constant factor of its standard deviation or expectation.

Corollary 3.7: *For any $t > 1$,*

$$\Pr(|X - \mathbf{E}[X]| \geq t \cdot \sigma[X]) \leq \frac{1}{t^2} \quad \textit{and}$$

$$\Pr(|X - \mathbf{E}[X]| \geq t \cdot \mathbf{E}[X]) \leq \frac{\mathbf{Var}[X]}{t^2(\mathbf{E}[X])^2}.$$

Let us again consider our coin-flipping example, and this time use Chebyshev's inequality to bound the probability of obtaining more than $3n/4$ heads in a sequence of n fair coin flips. Recall that $X_i = 1$ if the ith coin flip is heads and 0 otherwise, and that $X = \sum_{i=1}^{n} X_i$ denotes the number of heads in the n coin flips. To use Chebyshev's inequality we need to compute the variance of X. Observe first that, since X_i is a 0–1 random variable,

$$\mathbf{E}[(X_i)^2] = \mathbf{E}[X_i] = \frac{1}{2}.$$

Thus,

$$\mathbf{Var}[X_i] = \mathbf{E}[(X_i)^2] - (\mathbf{E}[X_i])^2 = \frac{1}{2} - \frac{1}{4} = \frac{1}{4}.$$

Now, since $X = \sum_{i=1}^{n} X_i$ and the X_i are independent, we can use Theorem 3.5 to compute

$$\mathbf{Var}[X] = \mathbf{Var}\left[\sum_{i=1}^{n} X_i\right] = \sum_{i=1}^{n} \mathbf{Var}[X_i] = \frac{n}{4}.$$

Applying Chebyshev's inequality then yields

$$\begin{aligned}\Pr(X \geq 3n/4) &\leq \Pr(|X - \mathbf{E}[X]| \geq n/4)\\ &\leq \frac{\mathbf{Var}[X]}{(n/4)^2}\\ &= \frac{n/4}{(n/4)^2}\\ &= \frac{4}{n}.\end{aligned}$$

In fact, we can do slightly better. Chebyshev's inequality yields that $4/n$ is actually a bound on the probability that X is either smaller than $n/4$ or larger than $3n/4$, so by symmetry the probability that X is greater than $3n/4$ is actually $2/n$. Chebyshev's inequality gives a significantly better bound than Markov's inequality for large n.

3.3.1. *Example: Coupon Collector's Problem*

We apply Markov's and Chebyshev's inequalities to the coupon collector's problem. Recall that the time X to collect n coupons has expectation nH_n, where $H_n = \sum_{i=1}^{n} 1/n = \ln n + O(1)$. Hence Markov's inequality yields

$$\Pr(X \geq 2nH_n) \leq \frac{1}{2}.$$

To use Chebyshev's inequality, we need to find the variance of X. Recall again from Section 2.4.1 that $X = \sum_{i=1}^{n} X_i$, where the X_i are geometric random variables with parameter $(n - i + 1)/n$. In this case, the X_i are independent because the time to collect the ith coupon does not depend on how long it took to collect the previous $i - 1$ coupons. Hence

$$\mathbf{Var}[X] = \mathbf{Var}\left[\sum_{i=1}^{n} X_i\right] = \sum_{i=1}^{n} \mathbf{Var}[X_i],$$

so we need to find the variance of a geometric random variable.

Let Y be a geometric random variable with parameter p. As we saw in Section 2.4, $\mathbf{E}[X] = 1/p$. We calculate $\mathbf{E}[Y^2]$. The following trick proves useful. We know that, for $0 < x < 1$,

$$\frac{1}{1-x} = \sum_{i=0}^{\infty} x^i.$$

Taking derivatives, we find:

$$\begin{aligned}
\frac{1}{(1-x)^2} &= \sum_{i=0}^{\infty} ix^{i-1} \\
&= \sum_{i=0}^{\infty} (i+1)x^i; \\
\frac{2}{(1-x)^3} &= \sum_{i=0}^{\infty} i(i-1)x^{i-2} \\
&= \sum_{i=0}^{\infty} (i+1)(i+2)x^i.
\end{aligned}$$

We can conclude that

$$\begin{aligned}\sum_{i=1}^{\infty} i^2x^i &= \sum_{i=0}^{\infty} i^2x^i\\ &= \sum_{i=0}^{\infty}(i+1)(i+2)x^i - 3\sum_{i=0}^{\infty}(i+1)x^i + \sum_{i=0}^{\infty}x^i\\ &= \frac{2}{(1-x)^3} - 3\frac{1}{(1-x)^2} + \frac{1}{(1-x)}\\ &= \frac{x^2+x}{(1-x)^3}.\end{aligned}$$

We now use this to find

$$\begin{aligned}\mathbf{E}[Y^2] &= \sum_{i=1}^{\infty} p(1-p)^{i-1}i^2\\ &= \frac{p}{1-p}\sum_{i=1}^{\infty}(1-p)^i i^2\\ &= \frac{p}{1-p}\frac{(1-p)^2+(1-p)}{p^3}\\ &= \frac{2-p}{p^2}.\end{aligned}$$

Finally, we reach

$$\begin{aligned}\mathbf{Var}[Y] &= \mathbf{E}[Y^2] - \mathbf{E}[Y]^2\\ &= \frac{2-p}{p^2} - \frac{1}{p^2}\\ &= \frac{1-p}{p^2}.\end{aligned}$$

We have just proven the following useful lemma.

Lemma 3.8: *The variance of a geometric random variable with parameter p is $(1-p)/p^2$.*

For a geometric random variable Y, $\mathbf{E}[Y^2]$ can also be derived using conditional expectations. We use that Y corresponds to the number of flips until the first heads, where each flip is heads with probability p. Let $X = 0$ if the first flip is tails and $X = 1$ if the first flip is heads. By Lemma 2.5,

$$\begin{aligned}\mathbf{E}[Y^2] &= \Pr(X=0)\mathbf{E}[Y^2 \mid X=0] + \Pr(X=1)\mathbf{E}[Y^2 \mid X=1]\\ &= (1-p)\mathbf{E}[Y^2 \mid X=0] + p\mathbf{E}[Y^2 \mid X=1].\end{aligned}$$

If $X = 1$, then $Y = 1$ and so $\mathbf{E}[Y^2 \mid X = 1] = 1$. If $X = 0$, then $Y > 1$. In this case, let the number of remaining flips after the first flip until the first head be Z. Then

$$\begin{aligned}\mathbf{E}[Y^2] &= (1-p)\mathbf{E}[(Z+1)^2] + p \cdot 1 \\ &= (1-p)\mathbf{E}[Z^2] + 2(1-p)\mathbf{E}[Z] + 1 \qquad (3.2)\end{aligned}$$

by the linearity of expectations. By the memoryless property of geometric random variables, Z is also a geometric random variable with parameter p. Hence $\mathbf{E}[Z] = 1/p$ and $\mathbf{E}[Z^2] = \mathbf{E}[Y^2]$. Plugging these values into Eqn. (3.2), we have

$$\mathbf{E}[Y^2] = (1-p)\mathbf{E}[Y^2] + \frac{2(1-p)}{p} + 1 = (1-p)\mathbf{E}[Y^2] + \frac{2-p}{p},$$

which yields $\mathbf{E}[Y^2] = (2-p)/p^2$, matching our other derivation.

We return now to the question of the variance in the coupon collector's problem. We simplify the argument by using the upper bound $\mathbf{Var}[Y] \le 1/p^2$ for a geometric random variable, instead of the exact result of Lemma 3.8. Then

$$\mathbf{Var}[X] = \sum_{i=1}^{n} \mathbf{Var}[X_i] \le \sum_{i=1}^{n} \left(\frac{n}{n-i+1}\right)^2 = n^2 \sum_{i=1}^{n} \left(\frac{1}{i}\right)^2 \le \frac{\pi^2 n^2}{6}.$$

Here we have used the identity

$$\sum_{i=1}^{\infty} \left(\frac{1}{i}\right)^2 = \frac{\pi^2}{6}.$$

Now, by Chebyshev's inequality,

$$\Pr(|X - nH_n| \ge nH_n) \le \frac{n^2\pi^2/6}{(nH_n)^2} = \frac{\pi^2}{6(H_n)^2} = O\left(\frac{1}{\ln^2 n}\right).$$

In this case, Chebyshev's inequality again gives a much better bound than Markov's inequality. But it is still a fairly weak bound, as we can see by considering instead a fairly simple union bound argument.

Consider the probability of not obtaining the ith coupon after $n \ln n + cn$ steps. This probability is

$$\left(1 - \frac{1}{n}\right)^{n(\ln n + c)} < \mathrm{e}^{-(\ln n + c)} = \frac{1}{\mathrm{e}^c n}.$$

By a union bound, the probability that some coupon has not been collected after $n \ln n + cn$ steps is only e^{-c}. In particular, the probability that all coupons are not collected after $2n \ln n$ steps is at most $1/n$, a bound that is significantly better than what can be achieved even with Chebyshev's inequality.

3.4. Median and Mean

Let X be a random variable. The *median* of X is defined to be any value m such that

$$\Pr(X \le m) \ge 1/2 \quad \text{and} \quad \Pr(X \ge m) \ge 1/2.$$

For example, for a discrete random variable that is uniformly distributed over an odd number of distinct, sorted values $x_1, x_2, \ldots, x_{2k+1}$, the median is the middle value x_{k+1}. For a discrete random variable that is uniformly distributed over an even number of values $x_1, x_2, \ldots, x_{2k}$, *any* value in the range (x_k, x_{k+1}) would be a median.

The expectation $\mathbf{E}[X]$ and the median are usually different numbers. For distributions with a unique median that are symmetric around either the mean or median, the median is equal to the mean. For some distributions, the median can be easier to work with than the mean, and in some settings it is a more natural quantity to work with.

The following theorem gives an alternate characterization of the mean and median:

Theorem 3.9: *For any random variable X with finite expectation* $\mathbf{E}[X]$ *and finite median m,*

1. *the expectation* $\mathbf{E}[X]$ *is the value of c that minimizes the expression*

$$\mathbf{E}[(X-c)^2],$$

and

2. *the median m is a value of c that minimizes the expression*

$$\mathbf{E}[|X-c|].$$

Proof: The first result follows from linearity of expectations.

$$\mathbf{E}[(X-c)^2] = \mathbf{E}[X^2] - 2c\mathbf{E}[X] + c^2,$$

and taking the derivative with respect to c shows that $c = \mathbf{E}[X]$ yields the minimum.

For the second result, we want to show that that for any value c that is not a median and for any median m, we have $\mathbf{E}[|X-c|] > \mathbf{E}[|X-m|]$, or equivalently that $\mathbf{E}[|X-c| - |X-m|] > 0$. In that case the value of c that minimizes $\mathbf{E}[|X-c|]$ will be a median. (In fact, as a by-product, we show that for any two medians m and m', $\mathbf{E}[|X-m|] = \mathbf{E}[|X-m'|]$.)

Let us take the case where $c > m$ for a median m, and c is not a median, so $\Pr(X \geq c) < 1/2$. A similar argument holds for any value of c such that $\Pr(X \leq c) < 1/2$.

For $x \geq c$, $|x-c| - |x-m| = m-c$. For $m < x < c$, $|x-c| - |x-m| = c+m-2x > m-c$. Finally, for $x \leq m$, $|x-c| - |x-m| = c-m$. Combining the three cases, we have

$$\begin{aligned}
&\mathbf{E}[|X-c| - |X-m|] \\
&\quad = \Pr(X \geq c)(m-c) + \sum_{x:m<x<c} \Pr(X=x)(c+m-2x) + \Pr(X \leq m)(c-m).
\end{aligned}$$

We now consider two cases. If $\Pr(m < X < c) = 0$, then

$$\begin{aligned}
\mathbf{E}[|X-c| - |X-m|] &= \Pr(X \geq c)(m-c) + \Pr(X \leq m)(c-m) \\
&> \frac{1}{2}(m-c) + \frac{1}{2}(c-m) \\
&= 0,
\end{aligned}$$

where the inequality comes from $\Pr(X \geq c) < 1/2$ and $m < c$. (Note here that if c were another median, so $\Pr(X \geq c) = 1/2$, we would obtain $\mathbf{E}[|X - c| - |X - m|] = 0$, as stated earlier.)

If $\Pr(m < X < c) \neq 0$, then

$$\begin{aligned}
&\mathbf{E}[|X - c| - |X - m|] \\
&\quad = \Pr(X \geq c)(m - c) + \sum_{x:m<x<c} \Pr(X = x)(c + m - 2x) + \Pr(X \leq m)(c - m) \\
&\quad > \Pr(X > m)(m - c) + \Pr(X \leq m)(c - m) \\
&\quad > \frac{1}{2}(m - c) + \frac{1}{2}(c - m) \\
&\quad = 0,
\end{aligned}$$

where here the first inequality comes from $c + m - 2x > m - c$ for any value of x with non-zero probability in the range $m < x < c$. (This case cannot hold if c and m are both medians, as in this case we cannot have $\Pr(X \geq m) = 1/2$ and $\Pr(X \geq c) = 1/2$.) ■

Interestingly, for well-behaved random variables, the median and the mean cannot deviate from each other too much.

Theorem 3.10: *If X is a random variable with finite standard deviation σ, expectation μ, and median m, then*

$$|\mu - m| \leq \sigma.$$

Proof: The proof follows from the following sequence:

$$\begin{aligned}
|\mu - m| &= |\mathbf{E}[X] - m| \\
&= |\mathbf{E}[X - m]| \\
&\leq \mathbf{E}[|X - m|] \\
&\leq \mathbf{E}[|X - \mu|] \\
&\leq \sqrt{\mathbf{E}[(X - \mu)^2]} \\
&= \sigma.
\end{aligned}$$

Here the first inequality follows from Jensen's inequality, the second inequality follows from the result that the median minimizes $\mathbf{E}[|X - c|]$, and the third inequality is again Jensen's inequality. ■

In Exercise 3.19, we suggest another way of proving this result.

3.5. Application: A Randomized Algorithm for Computing the Median

Given a set S of n elements drawn from a totally ordered universe, the *median* of S is an element m of S such that at least $\lfloor n/2 \rfloor$ elements in S are less than or equal to m and at least $\lfloor n/2 \rfloor + 1$ elements in S are greater than or equal to m. If the elements in S are distinct, then m is the $(\lceil n/2 \rceil)$th element in the sorted order of S. Note that the median of a set is similar to but slightly different from the the median of a random variable defined in Section 3.4.

The median can be easily found deterministically in $O(n \log n)$ steps by sorting, and there is a relatively complex deterministic algorithm that computes the median in $O(n)$ time. Here we analyze a randomized linear time algorithm that is significantly simpler than the deterministic one and yields a smaller constant factor in the linear running time. To simplify the presentation, we assume that n is odd and that the elements in the input set S are distinct. The algorithm and analysis can be easily modified to include the case of a multi-set S (see Exercise 3.24) and a set with an even number of elements.

3.5.1. *The Algorithm*

The main idea of the algorithm involves *sampling,* which we first discussed in Section 1.2. The goal is to find two elements that are close together in the sorted order of S and that have the median lie between them. Specifically, we seek two elements $d, u \in S$ such that:

1. $d \leq m \leq u$ (the median m is between d and u); and
2. for $C = \{s \in S : d \leq s \leq u\}$, $|C| = o(n/\log n)$ (the total number of elements between d and u is small).

Sampling gives us a simple and efficient method for finding two such elements.

We claim that, once these two elements are identified, the median can easily be found in linear time with the following steps. The algorithm counts (in linear time) the number ℓ_d of elements of S that are smaller than d and then sorts (in sublinear, or $o(n)$, time) the set C. Notice that, since $|C| = o(n/\log n)$, the set C can be sorted in time $o(n)$ using any standard sorting algorithm that requires $O(m \log m)$ time to sort m elements. The $(\lfloor n/2 \rfloor - \ell_d + 1)$th element in the sorted order of C is m, since there are exactly $\lfloor n/2 \rfloor$ elements in S that are smaller than that value $(\lfloor n/2 - \ell_d \rfloor$ in the set C and ℓ_d in $S - C)$.

To find the elements d and u, we sample with replacement a multi-set R of $\lceil n^{3/4} \rceil$ elements from S. Recall that sampling with replacement means each element in R is chosen uniformly at random from the set S, independent of previous choices. Thus, the same element of S might appear more than once in the multi-set R. Sampling without replacement might give marginally better bounds, but both implementing and analyzing it are significantly harder. It is worth noting that we assume that an element can be sampled from S in constant time.

Since R is a random sample of S we expect m, the median element of S, to be close to the median element of R. We therefore choose d and u to be elements of R surrounding the median of R.

We require all the steps of our algorithm to work *with high probability,* by which we mean with probability at least $1 - O(1/n^c)$ for some constant $c > 0$. To guarantee that with high probability the set C includes the median m, we fix d and u to be respectively the $\lfloor n^{3/4}/2 - \sqrt{n} \rfloor$th and the $\lceil n^{3/4}/2 + \sqrt{n} \rceil$th elements in the sorted order of R. With

```
Randomized Median Algorithm:

Input: A set $S$ of $n$ elements over a totally ordered universe.
Output: The median element of $S$, denoted by $m$.

1. Pick a (multi-)set $R$ of $\lceil n^{3/4} \rceil$ elements in $S$, chosen independently and uniformly
   at random with replacement.
2. Sort the set $R$.
3. Let $d$ be the $\left(\lfloor \frac{1}{2}n^{3/4} - \sqrt{n} \rfloor\right)$th smallest element in the sorted set $R$.
4. Let $u$ be the $\left(\lceil \frac{1}{2}n^{3/4} + \sqrt{n} \rceil\right)$th smallest element in the sorted set $R$.
5. By comparing every element in $S$ to $d$ and $u$, compute the set
   $C = \{x \in S : d \le x \le u\}$ and the numbers $\ell_d = |\{x \in S : x < d\}|$ and
   $\ell_u = |\{x \in S : x > u\}|$.
6. If $\ell_d > n/2$ or $\ell_u > n/2$ then FAIL.
7. If $|C| \le 4n^{3/4}$ then sort the set $C$, otherwise FAIL.
8. Output the $(\lfloor n/2 \rfloor - \ell_d + 1)$th element in the sorted order of $C$.
```

Algorithm 3.1: Randomized median algorithm.

this choice, the set C includes all the elements of S that are between the $2\sqrt{n}$ sample points surrounding the median of R. The analysis will clarify that the choice of the size of R and the choices for d and u are tailored to guarantee both that (a) the set C is large enough to include m with high probability and (b) the set C is sufficiently small so that it can be sorted in sublinear time with high probability.

A formal description of the procedure is presented as Algorithm 3.1. In what follows, for convenience we treat $\sqrt{n}$ and $n^{3/4}$ as integers.

3.5.2. *Analysis of the Algorithm*

Based on our previous discussion, we first prove that – regardless of the random choices made throughout the procedure – the algorithm (a) always terminates in linear time and (b) outputs either the correct result or FAIL.

Theorem 3.11: *The randomized median algorithm terminates in linear time, and if it does not output FAIL then it outputs the correct median element of the input set S.*

Proof: Correctness follows because the algorithm could only give an incorrect answer if the median were not in the set C. But then either $\ell_d > n/2$ or $\ell_u > n/2$ and thus step 6 of the algorithm guarantees that, in these cases, the algorithm outputs FAIL. Similarly, as long as C is sufficiently small, the total work is only linear in the size of S. Step 7 of the algorithm therefore guarantees that the algorithm does not take more than linear time; if the sorting might take too long, the algorithm outputs FAIL without sorting. ■

The interesting part of the analysis that remains after Theorem 3.11 is bounding the probability that the algorithm outputs FAIL. We bound this probability by identifying three "bad" events such that, if none of these bad events occurs, the algorithm does not fail. In a series of lemmas, we then bound the probability of each of these events and show that the sum of these probabilities is only $O(n^{-1/4})$.

Consider the following three events:

$\mathcal{E}_1$: $Y_1 = |\{r \in R \mid r \leq m\}| < \frac{1}{2}n^{3/4} - \sqrt{n}$;
$\mathcal{E}_2$: $Y_2 = |\{r \in R \mid r \geq m\}| < \frac{1}{2}n^{3/4} - \sqrt{n}$;
$\mathcal{E}_3$: $|C| > 4n^{3/4}$.

Lemma 3.12: *The randomized median algorithm fails if and only if at least one of $\mathcal{E}_1$, $\mathcal{E}_2$, or $\mathcal{E}_3$ occurs.*

Proof: Failure in step 7 of the algorithm is equivalent to the event $\mathcal{E}_3$. Failure in step 6 of the algorithm occurs if and only if $\ell_d > n/2$ or $\ell_u > n/2$. But for $\ell_d > n/2$, the $\left(\frac{1}{2}n^{3/4} - \sqrt{n}\right)$th smallest element of R must be larger than m; this is equivalent to the event $\mathcal{E}_1$. Similarly, $\ell_u > n/2$ is equivalent to the event $\mathcal{E}_2$. ■

Lemma 3.13:

$$\Pr(\mathcal{E}_1) \leq \frac{1}{4}n^{-1/4}.$$

Proof: Define a random variable X_i by

$$X_i = \begin{cases} 1 & \text{if the } i\text{th sample is less than or equal to the median,} \\ 0 & \text{otherwise.} \end{cases}$$

The X_i are independent, since the sampling is done with replacement. Because there are $(n-1)/2+1$ elements in S that are less than or equal to the median, the probability that a randomly chosen element of S is less than or equal to the median can be written as

$$\Pr(X_i = 1) = \frac{(n-1)/2+1}{n} = \frac{1}{2} + \frac{1}{2n}.$$

The event $\mathcal{E}_1$ is equivalent to

$$Y_1 = \sum_{i=1}^{n^{3/4}} X_i < \frac{1}{2}n^{3/4} - \sqrt{n}.$$

Since Y_1 is the sum of Bernoulli trials, it is a binomial random variable with parameters $n^{3/4}$ and $1/2 + 1/2n$. Hence, using the result of Section 3.2.1 yields

$$\begin{aligned}\mathbf{Var}[Y_1] &= n^{3/4}\left(\frac{1}{2} + \frac{1}{2n}\right)\left(\frac{1}{2} - \frac{1}{2n}\right) \\ &= \frac{1}{4}n^{3/4} - \frac{1}{4n^{5/4}} \\ &< \frac{1}{4}n^{3/4}.\end{aligned}$$

Applying Chebyshev's inequality then yields

$$\begin{aligned}\Pr(\mathcal{E}_1) &= \Pr\left(Y_1 < \frac{1}{2}n^{3/4} - \sqrt{n}\right)\\ &\le \Pr\left(|Y_1 - \mathbf{E}[Y_1]| > \sqrt{n}\right)\\ &\le \frac{\mathbf{Var}[Y_1]}{n}\\ &< \frac{\frac{1}{4}n^{3/4}}{n} = \frac{1}{4}n^{-1/4}. \qquad \blacksquare\end{aligned}$$

We similarly obtain the same bound for the probability of the event $\mathcal{E}_2$. We now bound the probability of the third bad event, $\mathcal{E}_3$.

Lemma 3.14:

$$\Pr(\mathcal{E}_3) \le \frac{1}{2}n^{-1/4}.$$

Proof: If $\mathcal{E}_3$ occurs, so $|C| > 4n^{3/4}$, then at least one of the following two events occurs:

$\mathcal{E}_{3,1}$: at least $2n^{3/4}$ elements of C are greater than the median;
$\mathcal{E}_{3,2}$: at least $2n^{3/4}$ elements of C are smaller than the median.

Let us bound the probability that the first event occurs; the second will have the same bound by symmetry. If there are at least $2n^{3/4}$ elements of C above the median, then the order of u in the sorted order of S was at least $\frac{1}{2}n + 2n^{3/4}$ and thus the set R has at least $\frac{1}{2}n^{3/4} - \sqrt{n}$ samples among the $\frac{1}{2}n - 2n^{3/4}$ largest elements in S.

Let X be the number of samples among the $\frac{1}{2}n - 2n^{3/4}$ largest elements in S. Let $X = \sum_{i=1}^{n^{3/4}} X_i$, where

$$X_i = \begin{cases} 1 & \text{if the } i\text{th sample is among the } \frac{1}{2}n - 2n^{3/4} \text{ largest elements in } S,\\ 0 & \text{otherwise.}\end{cases}$$

Again, X is a binomial random variable, and we find

$$\mathbf{E}[X] = \frac{1}{2}n^{3/4} - 2\sqrt{n}$$

and

$$\mathbf{Var}[X] = n^{3/4}\left(\frac{1}{2} - 2n^{-1/4}\right)\left(\frac{1}{2} + 2n^{-1/4}\right) = \frac{1}{4}n^{3/4} - 4n^{1/4} < \frac{1}{4}n^{3/4}.$$

Applying Chebyshev's inequality yields

$$\Pr(\mathcal{E}_{3,1}) = \Pr\left(X \ge \tfrac{1}{2}n^{3/4} - \sqrt{n}\right) \tag{3.3}$$

$$\le \Pr\left(|X - \mathbf{E}[X]| \ge \sqrt{n}\right) \le \frac{\mathbf{Var}[X]}{n} < \frac{\frac{1}{4}n^{3/4}}{n} = \frac{1}{4}n^{-1/4}. \tag{3.4}$$

Similarly,

$$\Pr(\mathcal{E}_{3,2}) \le \frac{1}{4}n^{-1/4}$$

and

$$\Pr(\mathcal{E}_3) \le \Pr(\mathcal{E}_{3,1}) + \Pr(\mathcal{E}_{3,2}) \le \frac{1}{2}n^{-1/4}.$$ ■

Combining the bounds just derived, we conclude that the probability that the algorithm outputs FAIL is bounded by

$$\Pr(\mathcal{E}_1) + \Pr(\mathcal{E}_2) + \Pr(\mathcal{E}_3) \le n^{-1/4}.$$

This yields the following theorem.

Theorem 3.15: *The probability that the randomized median algorithm fails is bounded by* $n^{-1/4}$.

By repeating Algorithm 3.1 until it succeeds in finding the median, we can obtain an iterative algorithm that never fails but has a random running time. The samples taken in successive runs of the algorithm are independent, so the success of each run is independent of other runs, and hence the number of runs until success is achieved is a geometric random variable. As an exercise, you may wish to show that this variation of the algorithm (that runs until it finds a solution) still has linear expected running time.

Randomized algorithms that may fail or return an incorrect answer are called *Monte Carlo* algorithms. The running time of a Monte Carlo algorithm often does not depend on the random choices made. For example, we showed in Theorem 3.11 that the randomized median algorithm always terminates in linear time, regardless of its random choices.

A randomized algorithm that always returns the right answer is called a *Las Vegas* algorithm. We have seen that the Monte Carlo randomized algorithm for the median can be turned into a Las Vegas algorithm by running it repeatedly until it succeeds. Again, turning it into a Las Vegas algorithm means the running time is variable, although the expected running time is still linear.

3.6. Exercises

Exercise 3.1: Let X be a number chosen uniformly at random from $[1, n]$. Find $\mathbf{Var}[X]$.

Exercise 3.2: Let X be a number chosen uniformly at random from $[-k, k]$. Find $\mathbf{Var}[X]$.

Exercise 3.3: Suppose that we roll a standard fair die 100 times. Let X be the sum of the numbers that appear over the 100 rolls. Use Chebyshev's inequality to bound $\Pr(|X - 350| \ge 50)$.

Exercise 3.4: Prove that, for any real number c and any discrete random variable X, $\mathbf{Var}[cX] = c^2\,\mathbf{Var}[X]$.

Exercise 3.5: Given any two random variables X and Y, by the linearity of expectations we have $\mathbf{E}[X - Y] = \mathbf{E}[X] - \mathbf{E}[Y]$. Prove that, when X and Y are independent, $\mathbf{Var}[X - Y] = \mathbf{Var}[X] + \mathbf{Var}[Y]$.

Exercise 3.6: For a coin that comes up heads independently with probability p on each flip, what is the variance in the number of flips until the kth head appears?

Exercise 3.7: A simple model of the stock market suggests that, each day, a stock with price q will increase by a factor $r > 1$ to qr with probability p and will fall to q/r with probability $1 - p$. Assuming we start with a stock with price 1, find a formula for the expected value and the variance of the price of the stock after d days.

Exercise 3.8: Suppose that we have an algorithm that takes as input a string of n bits. We are told that the expected running time is $O(n^2)$ if the input bits are chosen independently and uniformly at random. What can Markov's inequality tell us about the worst-case running time of this algorithm on inputs of size n?

Exercise 3.9: (a) Let X be the sum of Bernoulli random variables, $X = \sum_{i=1}^{n} X_i$. The X_i do not need to be independent. Show that

$$\mathbf{E}[X^2] = \sum_{i=1}^{n} \Pr(X_i = 1)\mathbf{E}[X \mid X_i = 1]. \tag{3.5}$$

Hint: Start by showing that

$$\mathbf{E}[X^2] = \sum_{i=1}^{n} \mathbf{E}[X_i X],$$

and then apply conditional expectations.

(b) Use Eqn. (3.5) to provide another derivation for the variance of a binomial random variable with parameters n and p.

Exercise 3.10: For a geometric random variable X, find $\mathbf{E}[X^3]$ and $\mathbf{E}[X^4]$. (*Hint:* Use Lemma 2.5.)

Exercise 3.11: Recall the Bubblesort algorithm of Exercise 2.22. Determine the variance of the number of inversions that need to be corrected by Bubblesort.

Exercise 3.12: Find an example of a random variable with finite expectation and unbounded variance. Give a clear argument showing that your choice has these properties.

Exercise 3.13: Find an example of a random variable with finite jth moments for $1 \le j \le k$ but an unbounded $(k + 1)$th moment. Give a clear argument showing that your choice has these properties.

Exercise 3.14: Prove that, for any finite collection of random variables $X_1, X_2, \ldots, X_n$,

$$\mathbf{Var}\left[\sum_{i=1}^{n} X_i\right] = \sum_{i=1}^{n} \mathbf{Var}[X_i] + 2\sum_{i=1}^{n}\sum_{j>i} \mathbf{Cov}(X_i, X_j).$$

Exercise 3.15: Let the random variable X be representable as a sum of random variables $X = \sum_{i=1}^{n} X_i$. Show that, if $\mathbf{E}[X_iX_j] = \mathbf{E}[X_i]\mathbf{E}[X_j]$ for every pair of i and j with $1 \leq i < j \leq n$, then $\mathbf{Var}[X] = \sum_{i=1}^{n} \mathbf{Var}[X_i]$.

Exercise 3.16: This problem shows that Markov's inequality is as tight as it could possibly be. Given a positive integer k, describe a random variable X that assumes only nonnegative values such that

$$\Pr(X \geq k\mathbf{E}[X]) = \frac{1}{k}.$$

Exercise 3.17: Can you give an example (similar to that for Markov's inequality in Exercise 3.16) that shows that Chebyshev's inequality is tight? If not, explain why not.

Exercise 3.18: Show that, for a random variable X with standard deviation $\sigma[X]$ and any positive real number t:

(a) $\Pr(X - \mathbf{E}[X] \geq t\sigma[X]) \leq \dfrac{1}{1+t^2}$;

(b) $\Pr(|X - \mathbf{E}[X]| \geq t\sigma[X]) \leq \dfrac{2}{1+t^2}$.

Exercise 3.19: Using Exercise 3.18, show that $|\mu - m| \leq \sigma$ for a random variable with finite standard deviation σ, expectation μ, and median m.

Exercise 3.20: Let Y be a nonnegative integer-valued random variable with positive expectation. Prove

$$\frac{\mathbf{E}[Y]^2}{\mathbf{E}[Y^2]} \leq \Pr[Y \neq 0] \leq \mathbf{E}[Y].$$

Exercise 3.21: **(a)** Chebyshev's inequality uses the variance of a random variable to bound its deviation from its expectation. We can also use higher moments. Suppose that we have a random variable X and an even integer k for which $\mathbf{E}[(X - \mathbf{E}[X])^k]$ is finite. Show that

$$\Pr\left(|X - \mathbf{E}[X]| > t\sqrt[k]{\mathbf{E}[(X - \mathbf{E}[X])^k]}\right) \leq \frac{1}{t^k}.$$

(b) Why is it difficult to derive a similar inequality when k is odd?

Exercise 3.22: A *fixed point* of a permutation $\pi : [1, n] \rightarrow [1, n]$ is a value for which $\pi(x) = x$. Find the variance in the number of fixed points of a permutation chosen uniformly at random from all permutations. (*Hint:* Let X_i be 1 if $\pi(i) = i$, so that $\sum_{i=1}^{n} X_i$

is the number of fixed points. You cannot use linearity to find $\mathbf{Var}\left[\sum_{i=1}^{n} X_i\right]$, but you can calculate it directly.)

Exercise 3.23: Suppose that we flip a fair coin n times to obtain n random bits. Consider all $m = \binom{n}{2}$ pairs of these bits in some order. Let Y_i be the exclusive-or of the ith pair of bits, and let $Y = \sum_{i=1}^{m} Y_i$ be the number of Y_i that equal 1.

(a) Show that each Y_i is 0 with probability $1/2$ and 1 with probability $1/2$.
(b) Show that the Y_i are not mutually independent.
(c) Show that the Y_i satisfy the property that $\mathbf{E}[Y_iY_j] = \mathbf{E}[Y_i]\mathbf{E}[Y_j]$.
(d) Using Exercise 3.15, find $\mathbf{Var}[Y]$.
(e) Using Chebyshev's inequality, prove a bound on $\Pr(|Y - \mathbf{E}[Y]| \geq n)$.

Exercise 3.24: Generalize the median-finding algorithm for the case where the input S is a multi-set. Bound the error probability and the running time of the resulting algorithm.

Exercise 3.25: Generalize the median-finding algorithm to find the kth largest item in a set of n items for any given value of k. Prove that your resulting algorithm is correct, and bound its running time.

Exercise 3.26: The weak law of large numbers states that, if $X_1, X_2, X_3, \ldots$ are independent and identically distributed random variables with mean μ and standard deviation σ, then for any constant $\varepsilon > 0$ we have

$$\lim_{n\to\infty} \Pr\left(\left|\frac{X_1 + X_2 + \cdots + X_n}{n} - \mu\right| > \varepsilon\right) = 0.$$

Use Chebyshev's inequality to prove the weak law of large numbers.