The law of large numbers, Markov's inequality and Chebyshev's inequality

Ten problems on the law of large numbers and the inequalities behind it: Markov's bound on a latency, Chebyshev's bound against the normal's exact tail, Chebyshev proved from Markov with the distribution that makes it tight, the bound against a uniform's exact tail, the weak law for die rolls, the sample size for estimating a coin's bias, a Monte Carlo estimate of π, the batch size for a minibatch gradient, the variance of an average of correlated samples and why proportions settle while counts spread, with worked solutions and the mistakes of using Markov on a variable that can be negative, putting σ where σ² belongs, dividing by n instead of √n, averaging correlated samples as if they were independent and expecting a coin to even out.

BEFORE YOU START

The law of large numbers is the reason an average of many noisy measurements can be trusted: as the sample grows, the sample mean settles on the true mean. Every Monte Carlo estimate, every minibatch gradient and every test-set accuracy relies on it. Its simplest proof is two lines long and rests on two inequalities that hold for any distribution with a mean and, for the second, a variance: Markov's, which bounds a tail using only the mean, and Chebyshev's, which bounds the distance from the mean using the variance. These ten problems use both bounds, compare them with exact answers to see how much they give away, prove the weak law and use it to size a sample, a Monte Carlo run and a minibatch, and then look at what changes when samples are correlated and why a fair coin's proportion of heads settles while its count of extra heads does not. The five mistakes at the end are the ones that give a believable number: Markov applied to a variable that can be negative, σ written where σ^2 belongs, an error divided by n instead of $\sqrt{n}$, correlated samples averaged as if independent, and a coin expected to make up for its past.

- **Markov's inequality.** If $X \geq 0$ and $a > 0$, then $P(X \geq a) \leq \dfrac{E[X]}{a}$.
- **Chebyshev's inequality.** If X has mean μ and variance σ^2, then for every $t > 0$, $P(|X - \mu| \geq t) \leq \dfrac{\sigma^2}{t^2}$. With $t = k\sigma$ this reads $P(|X - \mu| \geq k\sigma) \leq \dfrac{1}{k^2}$.
- **The sample mean.** For independent, identically distributed (i.i.d.) $X_1, \ldots, X_n$ with mean μ and variance σ^2, $\bar{X}_n = \frac{1}{n}\sum_i X_i$ has $E[\bar{X}_n] = \mu$ and $\mathrm{Var}(\bar{X}_n) = \sigma^2/n$, so its standard deviation is $\sigma/\sqrt{n}$.
- **The weak law of large numbers.** For every $\varepsilon > 0$, $P(|\bar{X}_n - \mu| \geq \varepsilon) \leq \dfrac{\sigma^2}{n\varepsilon^2}$, which tends to 0 as $n \to \infty$.

- **Proportions.** The fraction $\hat{p}$ of n independent trials in which an event of probability p happens is the sample mean of indicators, each with mean p and variance $p(1-p) \le \frac{1}{4}$.
- **Correlation.** For a sum, $\operatorname{Var}\left(\sum_i X_i\right) = \sum_i \operatorname{Var}(X_i) + \sum_{i \neq j} \operatorname{Cov}(X_i, X_j)$. Φ is the standard normal CDF, as on the normal-distribution page.

Builds on: Discrete random variables: expectation, variance and the binomial, geometric and Poisson, Variance, covariance and correlation

Problems

1. A service's response time T is never negative and has mean 200 ms. What does Markov's inequality say about $P(T \ge 1000)$ and about $P(T \ge 100)$? If T is in fact exponential with mean 200 ms, what is $P(T \ge 1000)$? difficulty ·

2. A measurement has mean 50 and variance 25. Bound $P(|X - 50| \ge 15)$ with Chebyshev's inequality, and give the guarantee it implies for $P(35 < X < 65)$. Compare with the exact value if X is normal. difficulty ·

3. Prove Chebyshev's inequality from Markov's. Then let $k > 1$ and let X equal $\mu - k\sigma$ and $\mu + k\sigma$ with probability $\dfrac{1}{2k^2}$ each, and μ otherwise. Show that X has mean μ and variance σ^2, and that Chebyshev's bound is an equality for it. difficulty ··

4. Let X be uniform on $(0, 1)$. Compute $P(|X - \frac{1}{2}| \ge 0.4)$ and $P(|X - \frac{1}{2}| \ge 0.3)$ exactly, and compare each with Chebyshev's bound. difficulty ··

5. A fair die is rolled n times. Use the weak law to bound the probability that the average roll is 0.1 or more away from 3.5 when $n = 1000$, and find the n that makes the bound 0.01. What does the normal approximation give for $n = 1000$? difficulty ··

6. A coin of unknown bias p is flipped n times and $\hat{p}$ is the fraction of heads. Using only $p(1-p) \le \frac{1}{4}$, how large must n be for Chebyshev to guarantee $P(|\hat{p} - p| \ge 0.05) \le 0.05$? What n does the normal approximation suggest? difficulty ··

7. To estimate π, draw $n = 10\,000$ points uniformly in the unit square and let $\hat{p}$ be the fraction with $x^2 + y^2 \le 1$; the estimate is $4\hat{p}$. Find its standard deviation, and bound $P(|4\hat{p} - \pi| \ge 0.05)$. difficulty ··

8. Each example's gradient component has mean g and variance 4, independently across examples. Find the standard deviation of the minibatch average over B examples, the B that makes it 0.1, and Chebyshev's bound on $P(|\bar{g} - g| \ge 0.25)$ at that B. difficulty ··

9. Let $X_1, \dots, X_n$ each have variance σ^2, with $\operatorname{Cov}(X_i, X_j) = \rho\sigma^2$ for every $i \ne j$. Find $\operatorname{Var}(\bar{X}_n)$ and its limit as $n \to \infty$. Evaluate for $\sigma^2 = 1$, $\rho = 0.1$ and $n = 100$, and find the number of independent samples that would give the same variance. difficulty ···

10. A fair coin is flipped n times, with H heads. Find the standard deviation of the proportion H/n and of the excess $H - n/2$ for $n = 100$ and $n = 10\,000$. Then find $P(H = n/2)$ for both, given that it is about $\sqrt{2/(\pi n)}$ for large n. difficulty ···

Answers

1. Markov: $P(T \geq 1000) \leq 0.2$, and nothing useful about $P(T \geq 100)$; for an exponential the true value is $e^{-5} \approx 0.0067$
2. $P(|X - 50| \geq 15) \leq \frac{1}{9} \approx 0.111$, so $P(35 < X < 65) \geq \frac{8}{9} \approx 0.889$; for a normal the tail is only 0.0027
3. Chebyshev is Markov applied to $(X - \mu)^2$, and the three-point distribution attains $1/k^2$ exactly, so no better bound holds for every distribution
4. Exact 0.2 against a bound of 0.521, and exact 0.4 against 0.926
5. The bound is 0.292 at $n = 1000$, and $n \geq 29\,167$ makes it 0.01; the normal approximation gives about 0.064
6. Chebyshev needs $n \geq 2000$; the normal approximation suggests 385
7. Standard deviation ≈ 0.0164; $P(|4\hat{p} - \pi| \geq 0.05) \leq 0.108$, so the estimate is within 0.05 of π at least 89% of the time
8. $2/\sqrt{B}$; $B = 400$; the bound is 0.16
9. $\operatorname{Var}(\bar{X}_n) = \dfrac{\sigma^2}{n} + \dfrac{n-1}{n}\rho\sigma^2 \to \rho\sigma^2$; here 0.109, the same as about 9 independent samples
10. Proportion: 0.05 then 0.005; excess heads: 5 then 50; exactly half heads: 0.080 then 0.008

Worked solutions

Problem 1

> A service's response time T is never negative and has mean 200 ms. What does Markov's inequality say about $P(T \geq 1000)$ and about $P(T \geq 100)$? If T is in fact exponential with mean 200 ms, what is $P(T \geq 1000)$?

$P(T \geq 1000) \leq \dfrac{200}{1000} = 0.2$	Markov with $a = 1000$; $T \geq 0$, so it applies.
$P(T \geq 100) \leq \dfrac{200}{100} = 2$	Markov with $a = 100$.
A bound of 2 says nothing	Every probability is at most 1 already. Markov is only useful for a above the mean.
Exponential: $P(T \geq t) = e^{-t/200}$, so $P(T \geq 1000) = e^{-5}$	The exponential's tail.
Markov: $P(T \geq 1000) \leq 0.2$, and nothing useful about $P(T \geq 100)$; for an exponential the true value is $e^{-5} \approx 0.0067$	The bound is thirty times the truth here because it uses only the mean. It holds for every nonnegative distribution with that mean, including the worst one: $T = 1000$ with probability 0.2 and $T = 0$ otherwise, where it is exact.

Problem 2

> A measurement has mean 50 and variance 25. Bound $P(|X - 50| \geq 15)$ with Chebyshev's inequality, and give the guarantee it implies for $P(35 < X < 65)$. Compare with the exact value if X is normal.

$P(\lvert X - 50\rvert \geq 15) \leq \dfrac{25}{15^2} = \dfrac{1}{9}$	Chebyshev with $\sigma^2 = 25$ and $t = 15$.
$15 = 3\sigma$, so this is $1/k^2$ with $k = 3$	$\sigma = 5$.

$P(35 < X < 65) = 1 - P(\lvert X - 50 \rvert \geq 15) \geq \dfrac{8}{9}$	The complement.
For a normal: $P(\lvert X - 50 \rvert \geq 15) = 2(1 - \Phi(3)) = 0.0027$	Three standard deviations, both tails.
$P(\lvert X - 50 \rvert \geq 15) \leq \frac{1}{9} \approx 0.111$, **so** $P(35 < X < 65) \geq \frac{8}{9} \approx 0.889$; **for a normal the tail is only** 0.0027	Chebyshev knows nothing about the shape, so it must allow for heavy tails. It is the guarantee to quote when the distribution is unknown, and the normal's 99.7% is the one to quote when it is known.

Problem 3

> Prove Chebyshev's inequality from Markov's. Then let $k > 1$ and let X equal $\mu - k\sigma$ and $\mu + k\sigma$ with probability $\dfrac{1}{2k^2}$ each, and μ otherwise. Show that X has mean μ and variance σ^2, and that Chebyshev's bound is an equality for it.

Let $Y = (X - \mu)^2$, so $Y \geq 0$ and $E[Y] = \sigma^2$	The definition of the variance.
$\lvert X - \mu \rvert \geq t$ exactly when $Y \geq t^2$	Both sides are nonnegative, so squaring keeps the inequality.
$P(\lvert X - \mu \rvert \geq t) = P(Y \geq t^2) \leq \dfrac{E[Y]}{t^2} = \dfrac{\sigma^2}{t^2}$	Markov on Y with $a = t^2$.
For the three-point X: $E[X] = \mu$	The two outer values are symmetric about μ and equally likely.
$\operatorname{Var}(X) = 2 \cdot \dfrac{1}{2k^2}(k\sigma)^2 = \sigma^2$	Only the outer values are away from μ, each by $k\sigma$.
$P(\lvert X - \mu \rvert \geq k\sigma) = 2 \cdot \dfrac{1}{2k^2} = \dfrac{1}{k^2}$	Exactly the bound.

Chebyshev is Markov applied to $(X-\mu)^2$, and the three-point distribution attains $1/k^2$ exactly, so no better bound holds for every distribution	For $k=2$ it puts $\frac{1}{8}$ at each of $\mu \pm 2\sigma$ and $\frac{3}{4}$ at μ. A smaller bound, like the one in Mistake 2, fails on a distribution of this kind.

Problem 4

> Let X be uniform on $(0,1)$. Compute $P(|X-\frac{1}{2}| \geq 0.4)$ and $P(|X-\frac{1}{2}| \geq 0.3)$ exactly, and compare each with Chebyshev's bound.

$\mu = \frac{1}{2}$ and $\sigma^2 = \frac{1}{12}$	The uniform's mean and variance on an interval of length 1.
$\lvert X-\frac{1}{2}\rvert \geq 0.4$ means $X \leq 0.1$ or $X \geq 0.9$: probability 0.2	Two end pieces of length 0.1.
Chebyshev: $\dfrac{1/12}{0.16} \approx 0.521$	$t = 0.4$, $t^2 = 0.16$.
$\lvert X-\frac{1}{2}\rvert \geq 0.3$ means $X \leq 0.2$ or $X \geq 0.8$: probability 0.4	Two end pieces of length 0.2.
Chebyshev: $\dfrac{1/12}{0.09} \approx 0.926$	$t = 0.3$.
Exact 0.2 against a bound of 0.521, and exact 0.4 against 0.926	The bounds hold, with room to spare: a uniform has no tails at all. In units of $\sigma \approx 0.289$ these are $k \approx 1.39$ and $k \approx 1.04$, where $1/k^2$ is barely below 1.

Problem 5

A fair die is rolled n times. Use the weak law to bound the probability that the average roll is 0.1 or more away from 3.5 when $n = 1000$, and find the n that makes the bound 0.01. What does the normal approximation give for $n = 1000$?

$\mu = 3.5$ and $\sigma^2 = \frac{1}{6}\sum_{k=1}^{6} k^2 - 3.5^2 = \frac{91}{6} - 12.25 = \frac{35}{12}$	$E[X^2] - (E[X])^2$.
$P(\lvert\bar{X} - 3.5\rvert \ge 0.1) \le \dfrac{35/12}{1000 \cdot 0.01} \approx 0.292$	The weak law with $\varepsilon = 0.1$.
$\dfrac{35/12}{n \cdot 0.01} \le 0.01$ needs $n \ge 29\,166.7$	Solve for n.
Normal: $\bar{X}$ has standard deviation $\sqrt{(35/12)/1000} \approx 0.0540$	$\sigma/\sqrt{n}$.
$2\big(1 - \Phi(0.1/0.0540)\big) = 2\big(1 - \Phi(1.85)\big)$	Both tails, at $z \approx 1.85$.
The bound is 0.292 at $n = 1000$, and $n \ge 29\,167$ makes it 0.01; the normal approximation gives about 0.064	Chebyshev's guarantee is about four and a half times the likely value. The law of large numbers says the average converges; the central limit theorem says how fast in a form sharp enough to plan with.

Problem 6

A coin of unknown bias p is flipped n times and $\hat{p}$ is the fraction of heads. Using only $p(1-p) \le \frac{1}{4}$, how large must n be for Chebyshev to guarantee $P(\lvert\hat{p} - p\rvert \ge 0.05) \le 0.05$? What n does the normal approximation suggest?

$\operatorname{Var}(\hat{p}) = \dfrac{p(1-p)}{n} \le \dfrac{1}{4n}$	The sample mean of indicators.
$P(\lvert\hat{p} - p\rvert \ge 0.05) \le \dfrac{1}{4n \cdot 0.0025} = \dfrac{100}{n}$	Chebyshev with $t = 0.05$.
$\dfrac{100}{n} \le 0.05$ needs $n \ge 2000$	Solve for n.

Normal: $1.96\sqrt{\frac{1}{4n}} \le 0.05$ needs $n \ge \left(\frac{1.96}{0.05}\right)^2 \cdot \frac{1}{4} = 384.16$	A two-sided 95% margin of 0.05.
Chebyshev needs $n \ge 2000$; the normal approximation suggests 385	The worst case $p = \frac{1}{2}$ is built into both. A poll of about 400 for a ± 5 point margin is the normal figure; 2000 is what can be promised with no assumption about the shape of the sampling distribution.

Problem 7

> To estimate π, draw $n = 10\,000$ points uniformly in the unit square and let $\hat{p}$ be the fraction with $x^2 + y^2 \le 1$; the estimate is $4\hat{p}$. Find its standard deviation, and bound $P(|4\hat{p} - \pi| \ge 0.05)$.

Each point lands in the quarter disk with probability $p = \frac{\pi}{4} \approx 0.7854$	The quarter disk has area $\pi/4$ and the square area 1.
$E[4\hat{p}] = \pi$ and $\operatorname{Var}(4\hat{p}) = \dfrac{16\,p(1-p)}{n}$	Scaling by 4 multiplies the variance by 16.
$16 \cdot 0.7854 \cdot 0.2146 / 10\,000 \approx 2.70 \times 10^{-4}$	$p(1-p) \approx 0.1685$.
Standard deviation ≈ 0.0164	The square root.
$P(\lvert 4\hat{p} - \pi \rvert \ge 0.05) \le \dfrac{2.70 \times 10^{-4}}{0.0025} \approx 0.108$	Chebyshev with $t = 0.05$.
Standard deviation ≈ 0.0164; $P(\lvert 4\hat{p} - \pi \rvert \ge 0.05) \le 0.108$, so the estimate is within 0.05 of π at least 89% of the time	The error shrinks like $1/\sqrt{n}$: one more correct digit costs a hundred times as many points. That rate, and not the dimension of the problem, is why Monte Carlo is used for high-dimensional integrals.

Problem 8

> Each example's gradient component has mean g and variance 4, independently across examples. Find the standard deviation of the minibatch average over B examples, the B that makes it 0.1, and Chebyshev's bound on $P(|\bar{g} - g| \geq 0.25)$ at that B.

$\mathrm{Var}(\bar{g}) = 4/B$, so its standard deviation is $2/\sqrt{B}$	The sample mean of B i.i.d. values.
$2/\sqrt{B} = 0.1$ gives $\sqrt{B} = 20$	Solve.
$B = 400$	Square.
$P(\lvert\bar{g} - g\rvert \geq 0.25) \leq \dfrac{4/400}{0.25^2} = \dfrac{0.01}{0.0625}$	Chebyshev with $\sigma^2 = 0.01$.
$2/\sqrt{B}$; $B = 400$; **the bound is** 0.16	Halving the noise of a gradient estimate takes four times the batch, which is why very large batches give diminishing returns per example. Examples drawn without replacement from a finite dataset, or correlated as in Problem 9, change the $4/B$.

Problem 9

> Let $X_1, \ldots, X_n$ each have variance σ^2, with $\mathrm{Cov}(X_i, X_j) = \rho\sigma^2$ for every $i \neq j$. Find $\mathrm{Var}(\bar{X}_n)$ and its limit as $n \to \infty$. Evaluate for $\sigma^2 = 1$, $\rho = 0.1$ and $n = 100$, and find the number of independent samples that would give the same variance.

$\mathrm{Var}\left(\sum_i X_i\right) = n\sigma^2 + n(n-1)\rho\sigma^2$	n variances and $n(n-1)$ ordered pairs $i \neq j$.
$\mathrm{Var}(\bar{X}_n) = \dfrac{\sigma^2}{n} + \dfrac{n-1}{n}\rho\sigma^2$	Divide by n^2.
As $n \to \infty$ this tends to $\rho\sigma^2$, not 0	The first term vanishes and the second tends to $\rho\sigma^2$.

For $\sigma^2 = 1$, $\rho = 0.1$, $n = 100$: $0.01 + 0.99 \cdot 0.1 = 0.109$	Step 2.
Independent samples: $\sigma^2/m = 0.109$ gives $m \approx 9.2$	The effective sample size.
$\mathrm{Var}(\bar{X}_n) = \dfrac{\sigma^2}{n} + \dfrac{n-1}{n}\rho\sigma^2 \to \rho\sigma^2$; **here** 0.109, **the same as about** 9 **independent samples**	With a shared positive correlation the average never settles: the part of the noise that all samples share does not average out. Neighboring frames of a video, or a minibatch drawn from one document, behave like this.

Problem 10

> A fair coin is flipped n times, with H heads. Find the standard deviation of the proportion H/n and of the excess $H - n/2$ for $n = 100$ and $n = 10\,000$. Then find $P(H = n/2)$ for both, given that it is about $\sqrt{2/(\pi n)}$ for large n.

$\mathrm{Var}(H) = n \cdot \frac{1}{2} \cdot \frac{1}{2} = \frac{n}{4}$, so H has standard deviation $\frac{1}{2}\sqrt{n}$	A binomial count.
$H - n/2$ has the same standard deviation, $\frac{1}{2}\sqrt{n}$: 5 and 50	Subtracting a constant changes nothing but the mean.
H/n has standard deviation $\dfrac{\sqrt{n}}{2n} = \dfrac{1}{2\sqrt{n}}$: 0.05 and 0.005	Dividing by n divides the standard deviation by n.
$P(H = 50 \mid n = 100) = \binom{100}{50} 2^{-100} \approx 0.0796$	Exact; $\sqrt{2/(100\pi)} \approx 0.0798$.
$P(H = 5000 \mid n = 10\,000) \approx \sqrt{2/(10\,000\pi)} \approx 0.0080$	The approximation; the exact value is 0.00798.

Proportion: 0.05 **then** 0.005; **excess heads:** 5 **then** 50; **exactly half heads:** 0.080 **then** 0.008	The law of large numbers is about the proportion. The excess of heads over half grows like $\sqrt{n}$, and an exact tie becomes ten times less likely. The proportion settles because the excess grows more slowly than n, not because the coin corrects itself (Mistake 5).

Where this goes wrong

1. Using Markov on a variable that can be negative

Markov's inequality needs only the mean, and a change in temperature has a mean.

A daily change X has $E[X] = 1$	Right so far.
"By Markov, $P(X \geq 4) \leq \frac{1}{4}$."	The condition that is missing: Markov needs $X \geq 0$. Large negative values can pull the mean down while the upper tail stays heavy.
$P(X \geq 4) \leq 0.25$ ✗	Let $X = -10$ or 12 with probability $\frac{1}{2}$ each. Its mean is 1, and $P(X \geq 4) = 0.5$, twice the claimed bound.

2. Putting σ where σ² belongs

Problem 2's numbers are given as a mean and a variance, and a standard deviation of 5 is the number that comes to mind.

$\mu = 50$, $\sigma = 5$, $t = 15$	Right so far.
"Chebyshev: $P(\lvert X - 50 \rvert \geq 15) \leq \sigma / t^2$."	The slip that causes the mistake: the numerator is the variance $\sigma^2 = 25$, which is what makes the two sides of the inequality the same units.

$P(\lvert X - 50\rvert \geq 15) \leq \dfrac{5}{225} \approx 0.0222$ ✗	The distribution of Problem 3 with $k = 3$, putting $\frac{1}{18}$ at each of 35 and 65 and the rest at 50, has variance 25 and $P(\lvert X - 50\rvert \geq 15) = \frac{1}{9} \approx 0.111$, five times the claimed bound.

3. Dividing by n instead of √n

The average divides the sum by n, and the same n seems to divide the noise.

$\mathrm{Var}(\bar{g}) = 4/B$	Right so far: Problem 8.
"So the standard deviation of $\bar{g}$ is $2/B$."	The slip that causes the mistake: the variance is divided by B, so the standard deviation is divided by $\sqrt{B}$.
a batch of $B = 20$ gives standard deviation 0.1 ✗	At $B = 20$ it is $2/\sqrt{20} \approx 0.447$, more than four times the target. The batch that gives 0.1 is 400.

4. Averaging correlated samples as if they were independent

The formula σ^2/n is the one remembered, and the samples in Problem 9 each have variance σ^2.

$n = 100$ samples, each with variance 1	Right so far.
"$\mathrm{Var}(\bar{X}) = \sigma^2/n = 0.01$, so the average has standard deviation 0.1."	The condition that is missing: σ^2/n needs the covariances to be 0. Here every pair has covariance 0.1.
standard deviation of $\bar{X} = 0.1$ ✗	It is $\sqrt{0.109} \approx 0.330$, more than three times larger. Error bars computed this way look tight on data that is far less informative than its size suggests.

5. Expecting the coin to even out

The law of large numbers says the proportion of heads tends to $\frac{1}{2}$, and after a run of heads it seems that tails must now be due.

After 100 flips there have been 60 heads	Right so far.

"To bring the proportion back to $\frac{1}{2}$, the coming flips must favor tails."	The habit that causes the mistake: reading a statement about the long-run proportion as a force acting on the next flips. Each flip is still fair.
the next 100 flips are expected to give fewer than 50 heads ✗	They are expected to give exactly 50. After 10 000 more flips the expected total is 5060 heads out of 10 100, a proportion of 0.501: the excess of 10 is still there in expectation, but it is swamped by the new flips, not canceled.

More at eulerpad.app. This is one of 82 free practice sets with worked solutions at eulerpad.app/practice, alongside guides to the topics behind them.

EulerPad, the app, reads your handwritten math on iPad, iPhone and Android and checks it line by line. It's free to start.

This set is also Chapter 14 of *Probability & Statistics for Machine Learning*, a 250-problem workbook with worked solutions: eulerpad.app/workbooks

eulerpad.app/practice/law-of-large-numbers-and-chebyshevs-inequality