Неравенства концентрации

Часто в приложениях нам необходимо показать, что некоторая случайная величина (количество очков в эксперименте с бросанием монеты или кости, продолжительность телефонного разговора, верхний уровень воды в реке и т.д.) лежит в каком-то небольшом интервале с большой вероятностью. Результаты такого типа называются неравенствами концентрации.

Теорема 1 (Неравенство Маркова)**.** *Для любого* $t > 0$

$$\mathbb{P}(|X| > t) \leq \frac{\mathbb{E}|X|}{t}.$$

Доказательство. Если $\mathbb{E}|X| = \infty$, неравенство становится тривиальным. Поэтому будем считать, что $\mathbb{E}X$ существует.

Нам потребуется следующее понятие. Назовём индикатором события A случайную величину $I(A)$, равную единице, если событие A произошло, и нулю, если A не произошло.

По определению, величина $I(A)$ имеет распределение Бернулли с параметром $p = \mathbb{P}(A)$, и её математическое ожидание равно вероятности успеха $p = \mathbb{P}(A)$. Индикаторы прямого и противоположного событий связаны равенством $I(A) + I(\bar{A}) = 1$. Поэтому для $A = \{|X| > t\}$ имеем:

$$|X| = |X| \cdot \big(I(A) + I(\bar{A})\big) \geq |X| \cdot I(A) \geq t \cdot I(A).$$

Тогда по свойствам математического ожидания

$$\mathbb{E}|X| \geq \mathbb{E}(t \cdot I(A)) = t \cdot \mathbb{P}(A) = t \cdot \mathbb{P}(|X| > t).$$

Осталось разделить обе части неравенства на положительное t. □

Следствие (Неравенство Чебышёва)**.** *Пусть* $\mathrm{Var}X$ *существует. Тогда для любого* $t > 0$

$$\mathbb{P}(|X - \mathbb{E}X| > t) \leq \frac{\mathrm{Var}X}{t^2}.$$

Доказательство. Для $t > 0$ неравенство $|X - \mathbb{E}X| > t$ равносильно неравенству $(X - \mathbb{E}X)^2 > t^2$, поэтому по неравенству Маркова

$$\mathbb{P}(|X - \mathbb{E}X| > t) = \mathbb{P}((X - \mathbb{E}X)^2 > t^2) \leq \frac{\mathbb{E}(X - \mathbb{E}X)^2}{t^2} = \frac{\mathrm{Var}X}{t^2}.$$

□

Упражнение 1. *Небольшая типография печатает случайное число* X *книг в месяц. Мы знаем, что среднее число книг в месяц —* 500.

1. *Что можно сказать о вероятности напечатать более* 1000 *книг в месяц?*
2. *Дисперсия случайной величины* X *равна* 100*. Что можно сказать о вероятности напечатать от* 400 *до* 600 *книг в месяц?*

Пример 1. Правильная монета подбрасывается 10 000 раз. Оценим вероятность того, что число гербов отличается от 5000 не более, чем на 100.

Если переформулировать задачу на вероятностный язык, то нам необходимо просто записать неравенство концентрации для случайной величины X, имеющей биномиальное распределение с $n =$

10 000 и $p = 1/2$. Напомним, что $\mathbb{E}X = np = 5\,000$ и $\mathrm{Var}X = np(1-p) = 2\,500$. По неравенству Чебышёва:

$$\mathbb{P}(|X - 500| \le 100) = 1 - \mathbb{P}(|X - 500| > 100) \ge 1 - \frac{2\,500}{10\,000} = \frac{3}{4}.$$

Итак, неравенство Чебышёва позволяет заключить, что в среднем не более чем в четверти случаев при 10 000 подбрасываниях монеты частота выпадения герба будет отличаться от 1/2 на одну сотую или больше. На самом деле искомую вероятность можно вычислить явно, и она равна примерно 0,9545. Разница этого ответа с оценкой 3/4 из неравенства Чебышёва показывает, насколько это неравенство является грубым. Существуют более тонкие оценки, мы их рассмотрим далее.

Производящая функция моментов

Перед тем, как перейти к более сложным (но и более эффективным) неравенствам концентрации, введем новый объект — производящую функцию моментов.

Определение. *Производящей функцией моментов* случайной величины X называется функция $M_X(\lambda)$ действительного аргумента λ, равная

$$M_X(\lambda) = \mathbb{E}e^{\lambda X}.$$

Эта функция определена, если указанное математическое ожидание существует для всех λ из некоторого интервала, включающего $\lambda = 0$.

Рассмотрим случайную величину X, у которой конечны моменты всех порядков. Тогда экспоненту мы можем представить в виде суммы ряда:

$$\mathbb{E}e^{\lambda X} = \mathbb{E}\Big(1 + \lambda X + \frac{\lambda^2 X^2}{2!} + \frac{\lambda^3 X^3}{3!} + \dots\Big) = 1 + \lambda \mathbb{E}X + \frac{\lambda^2}{2!}\mathbb{E}X^2 + \frac{\lambda^3}{3!}\mathbb{E}X^3 + \dots$$

Мы не будем подробно останавливаться на законности изменения порядка суммирования и взятия математического ожидания (так делать можно не всегда, так как сумма бесконечна). Во всех интересующих нас случаях так делать можно.

Смысл введения этой дополнительной функции кроется в её великолепных свойствах, главное из которых: разным распределениям отвечают разные производящие функции моментов. Название производящей функции моментов указывает ещё на одно полезное свойство: разлагая её в ряд Тейлора по степеням λ, получаем коэффициенты разложения, равные моментам случайной величины. Соответственно, k-я производная этой функции в нуле равна моменту k-го порядка:

$$\mathbb{E}X = M_X'(0), \quad \mathbb{E}X^2 = M_X''(0), \quad \mathbb{E}X^k = M_X^{(k)}(0).$$

Еще одно полезное свойство производящей функции моментов — производящая функция моментов суммы независимых случайных величин X и Y равна произведению их производящих функций:

$$M_{X+Y}(\lambda) = M_X(\lambda) \cdot M_Y(\lambda).$$

Упражнение 2. *Докажите последние два свойства производящей функции моментов (используйте тот факт, что порядок взятия производной и математического ожидания можно менять местами)*

Упражнение 3. *Пусть X имеет экспоненциальное распределение с параметром α, то есть плотность имеет вид*

$$f(u) = \begin{cases} \alpha e^{-\alpha u}, & u \ge 0, \\ 0, & u < 0. \end{cases}$$

Покажите, что производящая функция моментов $M(\lambda)$ имеет вид $M(\lambda) = \left(1 - \frac{\lambda}{\alpha}\right)^{-1}$. Используя этот результат, найдите третий момент $\mathbb{E}X^3$ разложив $M(\lambda)$ в ряд Тейлора.

Упражнение 4. *Покажите, что производящая функция моментов стандартной нормальной случайной величины $X \sim \mathcal{N}(0, 1)$ равна $M(\lambda) = e^{\lambda^2/2}$. Найдите третий момент $\mathbb{E}X^3$.*

Упражнение 5. *Покажите, что производящая функция моментов гауссовской случайной величины $X \sim \mathcal{N}(a, \sigma^2)$ равна $M(\lambda) = e^{\lambda^2\sigma^2/2+\lambda a}$.*

Закон больших чисел

Законами больших чисел (ЗБЧ) принято называть утверждения о том, при каких условиях среднее арифметическое случайных величин «стабилизируется» с ростом числа слагаемых. Всюду, где будут встречаться одинаково распределённые случайные величины X_i, $i = 1, 2, \ldots$, число $\mu = \mathbb{E}X_i$ будет обозначать их (у всех одинаковое) математическое ожидание, а $\sigma^2 = \mathrm{Var}X_i$ — их дисперсию.

Теорема 2 (ЗБЧ Чебышёва)**.** *Пусть $X_1, X_2, \ldots$ — последовательность независимых и одинаково распределённых случайных величин с конечной дисперсией, $\sigma^2 < \infty$. Тогда для любого фиксированного $\varepsilon > 0$*

$$\mathbb{P}\left(\left|\frac{X_1 + X_2 + \ldots + X_n}{n} - \mu\right| > \varepsilon\right) \to 0 \quad \textit{при } n \to \infty.$$

ЗБЧ утверждает, что среднее арифметическое большого числа одинаково устроенных случайных слагаемых сближается со средним значением одного слагаемого. Как бы сильно каждая случайная величина не отклонялась от своего математического ожидания, при суммировании эти отклонения «взаимно гасятся», так что среднее арифметическое приближается к постоянной величине.

Требование конечности дисперсии связано исключительно со способом доказательства: утверждение останется верным, если требовать существования только первого момента.

Доказательство. Обозначим через $S_n = X_1 + \ldots + X_n$ сумму первых n случайных величин. Из линейности математического ожидания получим:

$$\mathbb{E}\left(\frac{S_n}{n}\right) = \frac{\mathbb{E}X_1 + \ldots + \mathbb{E}X_n}{n} = \frac{n\mu}{n} = \mu.$$

Теперь посчитаем дисперсию S_n, пользуясь тем, что дисперсия суммы независимых слагаемых равна сумме их дисперсий:

$$\mathrm{Var}\left(\frac{S_n}{n}\right) = \frac{1}{n^2}\Big(\mathrm{Var}X_1 + \ldots + \mathrm{Var}X_n\Big) = \frac{n\sigma^2}{n^2} = \frac{\sigma^2}{n}.$$

Пусть $\varepsilon > 0$. Воспользуемся неравенством Чебышёва:

$$\mathbb{P}\left(\left|\frac{S_n}{n} - \mu\right| > \varepsilon\right) \le \frac{\mathrm{Var}(S_n/n)}{\varepsilon^2} = \frac{\sigma^2}{n\varepsilon^2} \to 0 \quad \text{при } n \to \infty.$$

□

Замечание. Мы не только доказали сходимость, но и получили оценку для вероятности среднему арифметическому любого числа независимых и одинаково распределённых величин отличаться от $\mu = \mathbb{E}X_1$ более чем на заданное $\varepsilon > 0$:

$$\mathbb{P}\left(\left|\frac{S_n}{n} - \mu\right| > \varepsilon\right) \le \frac{\sigma^2}{n\varepsilon^2}.$$

Замечание. Сходимость, которую мы получили в законе больших чисел, называется «сходимостью по вероятности». Это утверждение мы дальше будем писать так:

$$\frac{X_1 + X_2 + \ldots + X_n}{n} \xrightarrow{p} \mu,$$

имея ввиду то, что для любого фиксированного $\varepsilon > 0$

$$\mathbb{P}\left(\left|\frac{X_1 + X_2 + \ldots + X_n}{n} - \mu\right| > \varepsilon\right) \to 0 \quad \text{при } n \to \infty.$$

Центральная предельная теорема

Если в условиях ЗБЧ среднее арифметическое $\frac{S_n}{n} = \frac{X_1 + \ldots + X_n}{n}$ сходится к μ, то их разность $\frac{S_n}{n} - \mu$ сходится к нулю. Оказывается, что эта разность ведёт себя подобно величине $\frac{\sigma}{\sqrt{n}} N$, где N — случайная величина с нормальным распределением $\mathcal{N}(0,1)$. Иначе говоря, с ростом n распределение величины

$$\frac{\sqrt{n}}{\sigma}\left(\frac{S_n}{n} - \mu\right) = \frac{S_n - n\mu}{\sigma\sqrt{n}}$$

становится всё более похожим на стандартное нормальное распределение. Это утверждение и называется «центральной предельной теоремой» (ЦПТ).

Теорема 3 (ЦПТ). *Пусть $X_1, X_2, \ldots$ — независимые и одинаково распределённые случайные величины с конечной и ненулевой дисперсией σ^2 и математическим ожиданием μ. Тогда для любых $x < y$ имеет место сходимость при $n \to \infty$*

$$\mathbb{P}\left(x < \frac{S_n - n\mu}{\sigma\sqrt{n}} < y\right) \to \Phi(y) - \Phi(x) = \frac{1}{\sqrt{2\pi}} \int_x^y e^{-t^2/2} dt.$$

Заметим, что дробь $\frac{S_n - n\mu}{\sigma\sqrt{n}}$ получилась стандартизацией величины S_n : мы просто вычли из S_n её математическое ожидание $\mathbb{E}S_n = n\mu$ и поделили эту разность на корень из дисперсии $\sqrt{\mathrm{Var} S_n} = \sqrt{n\sigma^2}$.

Доказательство. Сначала сделаем удобную замену

$$Y_i = \frac{X_i - \mu}{\sigma}, \; i = 1, 2 \ldots.$$

Заметим, что случайные величины Y_i тоже будут независимы и одинаково распределены, $\mathbb{E}Y_i = 0$, $\mathrm{Var} Y_i = 1$, а S_n можно записать в виде

$$\frac{S_n - n\mu}{\sigma\sqrt{n}} = \frac{(X_1 - \mu) + \ldots (X_n - \mu)}{\sigma\sqrt{n}} = \frac{Y_1 + \ldots + Y_n}{\sqrt{n}}.$$

Предположим, что у суммируемых случайных величин существует производящая функция моментов. Тогда

$$\mathbb{E}e^{\lambda \frac{Y_1 + \ldots + Y_n}{\sqrt{n}}} = \mathbb{E}\left(e^{\lambda \frac{Y_1}{\sqrt{n}}}\right) \cdot \ldots \cdot \mathbb{E}\left(e^{\lambda \frac{Y_n}{\sqrt{n}}}\right) = \left(\mathbb{E}\left(e^{\lambda \frac{Y_1}{\sqrt{n}}}\right)\right)^n,$$

где мы воспользовались независимостью и одинаковой распределенностью случайный величин $Y_1, \ldots, Y_n$.

Теперь все, что осталось нам сделать, разложить полученное выражение в ряд Тейлора при $n \to \infty$

$$
\begin{aligned}
e^{\lambda \frac{Y_1}{\sqrt{n}}} &= 1 + \lambda \frac{Y_1}{\sqrt{n}} + \lambda^2 \frac{Y_1^2}{2n} + o\left(\frac{1}{n}\right), \\
\mathbb{E} e^{\lambda \frac{Y_1}{\sqrt{n}}} &= 1 + \frac{\lambda^2}{2n} + o\left(\frac{1}{n}\right).
\end{aligned}
$$

Теперь заметим, что при $n \to \infty$

$$
\left(\mathbb{E}\left(e^{\lambda \frac{Y_1}{\sqrt{n}}}\right)\right)^n = \left(1 + \frac{\lambda^2}{2n} + o\left(\frac{1}{n}\right)\right)^n \to e^{\lambda^2/2},
$$

а это и есть производящая функция моментов стандартной нормальной случайной величины $N \sim \mathcal{N}(0,1)$. Так как производящая функция моментов определяет распределение, утверждение теоремы можно считать доказанным. $\square$

С помощью центральной предельной теоремы можно записать приближенное неравенство концентрации:

$$
\mathbb{P}(x < S_n < y) = \mathbb{P}\left(\frac{x - n\mu}{\sigma\sqrt{n}} < \frac{Y - n\mu}{\sigma\sqrt{n}} < \frac{y - n\mu}{\sigma\sqrt{n}}\right) \approx \Phi\left(\frac{y - n\mu}{\sigma\sqrt{n}}\right) - \Phi\left(\frac{x - n\mu}{\sigma\sqrt{n}}\right).
$$

Замечание. Ещё раз напомним, что функция распределения стандартного нормального закона ищется либо по таблице, либо с помощью какого-либо программного обеспечения, но никак не путем нахождения первообразной от плотности нормального стандартного распределения.

Упражнение 6. *Правильная монета подбрасывается* 10 000 *раз. Оцените с помощью центральной предельной теоремы вероятность того, что число гербов отличается от* 5000 *не более, чем на* 100*. Сравните полученный результат с результатом из Примера 1.*

Неравенства концентрации (продолжение)

Как уже отмечалось, неравенства концентрации, полученные с помощью неравенства Чебышёва, не всегда дают удовлетворительный ответ (см. Пример 1). Центральная предельная теорема позволяет выписать неравенство концентрации, но приближенно. Оказывается, что для случайных величин, у которых существует производящая функция моментов, можно записать более эффективные оценки.

Лемма (Неравенство Чернова)**.** *Пусть у случайной величины X существует производящая функция моментов $M_X(\lambda)$. Обозначим $\mu = \mathbb{E}X$. Тогда для любых $t > 0$ и $\lambda > 0$*

$$
\mathbb{P}(X - \mu > t) \le \frac{\mathbb{E}e^{\lambda(X-\mu)}}{\lambda t}.
$$

Доказательство. Фиксируем произвольное $\lambda > 0$. Заметим, что следующие события эквивалентны:

$$
\{X - \mu > t\} = \{\lambda(X - \mu) > \lambda t\} = \{e^{\lambda(X-\mu)} > e^{\lambda t}\}.
$$

Первое равенство в этой цепочке следует из факта, что умножение обоих частей неравенства на положительное число λ не меняет его истинности, а второе — из монотонности экспоненты. Теперь найдем вероятность этих с помощью неравенства Маркова:

$$
\mathbb{P}(X - \mu > t) = \mathbb{P}(e^{\lambda(X-\mu)} > e^{\lambda t}) = \frac{\mathbb{E}e^{\lambda(X-\mu)}}{e^{\lambda t}}.
$$

Лемма доказана. $\square$

Замечание. Заметим, что доказательство этой леммы практически дословно повторяет доказательство неравенства Чебышёва. Единственное отличие заключается в том, что вместо того, чтобы возвести обе части неравенства $X - \mu > t$ в квадрат, мы берем функцию $f_\lambda(x) = e^{\lambda x}$ от обоих его частей. Как вы думаете, какие другие функции допустимо брать?

Оценка Чернова зависит от производящей функции моментов, так как в правой части мы имеем

$$\mathbb{E}e^{\lambda(X-\mu)} = \mathbb{E}e^{\lambda X}e^{-\lambda\mu} = e^{-\lambda\mu}M_X(\lambda).$$

Поэтому естественно классифицировать случайные величины по росту производящей функции (чтобы записывать неравенства концентрации сразу для целого класса случайных величин). Вспомним, что производящая функция $M_X(\lambda)$ нормальной случайной величины $X \sim \mathcal{N}(\mu, \sigma^2)$ (их еще иногда называют гауссовскими случайными величинами) равна

$$M_X(\lambda) = e^{\frac{\lambda^2\sigma^2}{2}+\lambda\mu}.$$

Введем следующее определение.

Определение. Случайную величину X со средним $\mu = \mathbb{E}X$ мы назовем суб-гауссовской, если существует такое число $\sigma > 0$, что

$$\mathbb{E}e^{\lambda(X-\mu)} \le e^{\frac{\lambda^2\sigma^2}{2}} \quad \text{для всех } \lambda \in \mathbb{R}.$$

Число $\sigma > 0$ мы будем называть параметром суб-гауссовской случайной величины. Очевидно, что каждая гауссовская случайная величина является суб-гауссовской с параметром σ равным дисперсии. Кроме того, любая ограниченная случайная величина тоже будет суб-гауссовской.

Утверждение. *Пусть X — случайная величина со средним $\mu = \mathbb{E}X$, принимающая значения в отрезке $[a, b]$, то есть $X \in [a, b]$. Тогда*

$$\mathbb{E}e^{\lambda(X-\mu)} \le e^{\frac{\lambda^2\sigma^2}{2}} \quad \textit{для всех } \lambda \in \mathbb{R},$$

где $\sigma = (b - a)/2$.

Этот факт мы доказывать не будем. Теперь перейдем к доказательству неравенства концентрации для всего класса суб-гауссовских величин.

Теорема 4. *Пусть X — суб-гауссовская случайная величина с параметром $\sigma > 0$. Обозначим $\mu = \mathbb{E}X$. Тогда для любого $t > 0$*

$$\mathbb{P}(X - \mu > t) \le e^{-\frac{\lambda^2}{2\sigma^2}} \quad \textit{и} \quad \mathbb{P}(|X - \mu| > t) \le 2e^{-\frac{\lambda^2}{2\sigma^2}}.$$

Доказательство. Запишем оценку Чернова и воспользуемся суб-гауссовостью X:

$$\mathbb{P}(X - \mu > t) \le \frac{\mathbb{E}e^{\lambda(X-\mu)}}{e^{\lambda t}} \le \frac{e^{\frac{\lambda^2\sigma^2}{2}}}{e^{\lambda t}} = e^{\frac{\lambda^2\sigma^2}{2}-\lambda t}.$$

Так как $\lambda > 0$ — произвольное в этой оценке, мы можем выбрать его так, чтобы наша оценка была максимально точна (то есть вероятность была минимальна). Так как в экспоненте стоит квадратный многочлен относительно λ, то минимум достигается на $\lambda_{\min} = t/\sigma^2$. Подставим,

$$\mathbb{P}(X - \mu > t) \le e^{\frac{\lambda_{\min}^2\sigma^2}{2}-\lambda_{\min}t} = e^{-\frac{\lambda^2}{2\sigma^2}}.$$

Теперь перейдем к доказательству второго неравенства. Начнем с оценки

$$\mathbb{P}(|X - \mu| > t) = \mathbb{P}(X - \mu > t, -X + \mu > t) \le \mathbb{P}(X - \mu > t) + \mathbb{P}(-X + \mu > t).$$

Первую вероятность мы уже оценили. Оказывается, вторая вероятность равна первой. Чтобы это доказать, заметим, что случайная величина $-X$ тоже является суб-гауссовской с параметром $\sigma > 0$

$$\mathbb{E}e^{\lambda(-X+\mu)} = \mathbb{E}e^{-\lambda(X-\mu)} \leq \frac{(-\lambda)^2\sigma^2}{2} = \frac{\lambda^2\sigma^2}{2}.$$

Следовательно, для нее тоже верна оценка $\mathbb{P}(-X + \mu > t) \leq e^{-\frac{\lambda^2}{2\sigma^2}}$. Поэтому

$$\mathbb{P}(|X - \mu| > t) \leq 2e^{-\frac{\lambda^2}{2\sigma^2}}.$$

Теорема доказана. □

Одна из причин, почему оценки с помощью производящей функции моментов популярны, — удобство работы с суммами случайных величин. Следующий факт даже не нужно доказывать: он следует сразу из определения.

Пусть дана последовательность независимых суб-гауссовских случайных величин X_i, $i = 1, \ldots, n$, со средними $\mu_i = \mathbb{E}X_i$ и параметрами $\sigma_i > 0$. Тогда сумма $\sum_{i=1}^n (X_i - \mu_i)$ является суб-гауссовской с параметром $\sigma^2 = \sum_{i=1}^n \sigma_i^2$ и средним 0.

Следствие (неравенство Хёфдинга)**.** *Пусть X_i — субгауссовские случайные величины с параметрами $\sigma_i > 0$ и средними $\mu_i \in \mathbb{R}$, $i = 1, \ldots, n$. Тогда для любого $t > 0$*

$$\mathbb{P}\left(\sum_{i=1}^n (X_i - \mu_i) > t\right) \leq \exp\left\{-\frac{t^2}{2\sum_{i=1}^n \sigma_i^2}\right\}.$$

Упражнение 7. *Правильная монета подбрасывается* 10 000 *раз. Оценим вероятность того, что число гербов отличается от* 5000 *не более, чем на* 100*. Сравните полученный результат с результатом из Примера 1 и Упражнения 6 .*

Лемма Джонсона-Линденштрауса

Мы сейчас рассмотрим один из методов снижения размерности, который основан на взятии случайных проекций, — вложение Джонсона-Линденштрауса.

Пусть нам данны точки $u_1, \ldots, u_n \in \mathbb{R}^D$. Если размерность пространства D велика, хранить координаты всех n точек может быть «дорого». Поэтому мы хотим придумать такое отображение $F : \mathbb{R}^D \to \mathbb{R}^d$ для $d \ll D$, которое лучше всего сохранит «структуру» данных. (Знак $\ll$ читается как «значительно меньше».)

Тут, конечно, возникает естественный вопрос: что понимать под «структурой»? Один из вариантов ответа на этот вопрос — попарные расстояния между точками (так как многие алгоритмы используют именно их). Поэтому нам интересно найти такое отображение F, которое слабо портит эту характеристику данных, то есть

$$(1 - \delta)\|u_i - u_j\|^2 \leq \|F(u_i) - F(u_j)\|^2 \leq (1 + \delta)\|u_i - u_j\|^2, \tag{1}$$

для любых $i \neq j$ и некоторого $\delta \in (0, 1)$. Здесь под $\|\cdot\|$ понимается обычная евклидова норма:

$$\|x\|^2 = x_1^2 + \ldots + x_d^2, \quad \text{где } x = (x_1, \ldots, x_d).$$

Конечно, можно взять $d = D$ и $F(u) = u$, и неравенство (5) будет выполнено для $\delta = 0$. Но нашей задачей является поиск небольшого значения d для относительно малой ошибки δ. В следующей теореме мы явно построим такое случайное отображение F, для которого неравенство (5) будет выполняться с большой вероятностью.

Теорема 5. *Пусть X — матрица размера $d \times D$, элементами которой являются независимые нормальные случайные величины, $X_{ij} \sim \mathcal{N}(0,1)$. Определим $F(u) = d^{-1/2}Xu$ и зафиксируем произвольное $\varepsilon \in (0,1)$. Покажите, что для любого $d > \frac{16}{\delta^2}\log(n/\varepsilon)$ неравенство (5) выполняется с вероятностью не меньше $1-\varepsilon$.*

Замечание. Обратите внимание, что полученная оценка на d (размерность пространства куда мы проецировали точки), совсем не зависит от исходной размерности D. А от количества точек n зависит только логарифмически. Это восхитительно!

Перед тем, как мы перейдем к доказательству этой теоремы, введем новый класс случайных величин, более широкий, чем суб-гауссовский.

Определение. Случайную величину X со средним $\mu = \mathbb{E}X$ мы назовем суб-экспоненциальной, если существуют такие числа $\sigma > 0$ и $b > 0$, что

$$\mathbb{E}e^{\lambda(X-\mu)} \le e^{\frac{\lambda^2\sigma^2}{2}} \quad \text{для всех } |\lambda| < \frac{1}{b}.$$

Числа $\sigma > 0$ и $b > 0$ мы будем называть параметрами суб-экспоненциальной случайной величины. Из определения видно, что каждая суб-гауссовская случайная величина является суб-экспоненциальной. Но несмотря на то, что производящая функция моментов ограничена только в окрестности нуля, мы все равно можем воспользоваться оценкой Чернова и записать неравенство концентрации.

Упражнение 8. *Пусть X — суб-экспоненциальная случайная величина с параметрами $\sigma > 0$, $b > 0$ и средним $\mu = \mathbb{E}X$. Тогда*

$$\mathbb{P}(|X-\mu| > t) \le \begin{cases} 2e^{-\frac{t^2}{2\sigma^2}}, & \text{для } 0 \le t < \frac{\sigma^2}{b}, \\ 2e^{-\frac{t}{2b}}, & \text{для } t \ge \frac{\sigma^2}{b}. \end{cases}$$

Доказательство этого утверждения практически ничем не отличается от доказательства Теоремы 4 выше. Поменяется только выбор оптимального λ: в данном случае $0 < \lambda < 1/b$, куда может не входить оптимальное значение $\lambda_{\min} = t/\sigma^2$. Это объясняет наличие двух «режимов»: если $\lambda_{\min} \in (0; 1/b)$, тогда мы получаем оценку из Теоремы 4, в противном случае — что-то иное.

Чтобы доказать Теоруму 5 нам понадобится концентрация распределения хи-квадрат.

Упражнение 9 (Концентрация для распределения хи-квадрат). *Пусть $X \sim \chi^2_1$ — хи-квадрат с одной степенью свободы, то есть $X = Y^2$ для $Y \sim \mathcal{N}(0,1)$. Покажите, что*

1. *Случайная величина X не является суб-гауссовской.*
2. *Случайная величина X является суб-экспоненциальной с параметрами $\sigma = 2$, $b = 4$. Для этого используйте следующую оценку*

$$\frac{e^{-\lambda}}{\sqrt{1-2\lambda}} \le e^{2\lambda^2} \quad \text{для всех } |\lambda| < \frac{1}{4}.$$

3. *Пусть $X_1, \ldots, X_d$ — независимые случайные χ^2_1 величины. Тогда сумма $\sum_{i=1}^d X_i$ является суб-экспоненциальной величиной со средним d и параметрами $\sigma = 2\sqrt{d}$, $b = 4$. Напомним, что это распределение называется распределением хи-квадрат с d степенями свободы и обозначается как χ^2_d.*

4. Используя этот факт, выпишите концентрационное неравенство

$$\mathbb{P}\left(\left|\sum_{i=1}^{d} X_i - d\right| > t\right) \le \begin{cases} 2e^{-\frac{t^2}{8d}}, & \text{для } 0 \le t < d, \\ 2e^{-\frac{t}{8}}, & \text{для } t \ge d, \end{cases}$$

и следствие из него

$$\mathbb{P}\left(\left|\frac{1}{d}\sum_{i=1}^{d} X_i - 1\right| > \delta\right) \le 2e^{-\frac{d\delta^2}{8}}, \quad \text{для } \delta = \frac{t}{d} \in (0,1).$$

Доказательство Теоремы 5. Найдем сначала вероятность того, что неравенство (5) не выполняется для фиксированных i и j. Обозначим $v = u_i - u_j$, $v = (v_1, \ldots, v_D)$. Заметим, что если $u_i = u_j$ то неравенство (5) тривиально, поэтому есть смысл рассматривать случай $v \neq 0$.

Неравенство (5) для $F(v) = d^{-1/2} X v$ можно записать следующим образом

$$1 - \delta \le \frac{1}{d}\frac{\|Xv\|^2}{\|v\|^2} \le 1 + \delta.$$

(Мы разделили на $\|v\| > 0$ обе части неравенства.) Положим $Y = \frac{Xv}{\|v\|} \in \mathbb{R}^d$, $Y = (Y_1, \ldots, Y_d)$, и перепишем неравенство в виде

$$1 - \delta \le \frac{1}{d}\|Y\|^2 \le 1 + \delta.$$

Нам необходимо найти вероятность, с которой данное соотношение не выполняется. Тут нам помогут неравенства концентрации. Заметим, что компоненты вектора Y равны

$$Y_i = \frac{1}{\|v\|}\sum_{j=1}^{D} X_{ij} v_j,$$

где X_{ij} — элементы матрицы X. Заметим, что Y_i является гауссовкой как сумма гауссовских случайныйх величин. Поэтому, чтобы определеить распределение Y_i, достаточно посчитать математическое ожидание и дисперсию (так как гауссовское распределение определяется только этими двумя параметрами).

$$\mathbb{E}Y_i = \frac{1}{\|u\|}\sum_{j=1}^{D} u_j \, \mathbb{E}X_{ij} = 0,$$

$$\mathrm{Var}Y_i = \frac{1}{\|u\|^2}\sum_{j=1}^{D} u_j^2 \, \mathrm{Var}X_{ij} = \frac{1}{\|u\|^2}\sum_{j=1}^{D} u_j^2 = 1.$$

То есть каждая компонента Y_i вектора Y является стандартной нормальной величиной. Кроме того, все компоненты Y_i независимы, так как X_{ij} независимы. Поэтому

$$\|Y\|^2 = Y_1^2 + \ldots Y_d^2 \sim \chi_d^2.$$

Воспользуемся полученной концентрацией для χ_d^2:

$$\mathbb{P}\left(\frac{1}{d}\|Y\| \notin [1-\delta, 1+\delta]\right) = \mathbb{P}\left(\left|\frac{1}{d}\|Y\|^2 - 1\right| > \delta\right) \le 2e^{-\frac{d\delta^2}{8}}, \quad \text{для } \delta \in (0,1).$$

Мы получили, что неравенство (5) для фиксированных u_i и u_j не выполняется с вероятностью $2e^{-\frac{d\delta^2}{8}}$. Но сколько всевозможных таких пар? Ответ: $C_n^2 = \frac{n(n-1)}{2}$. Так как вероятность объединения событий не превосходит сумму вероятностей каждого события, то

$$\mathbb{P}\left(\frac{1}{d}\left\|\frac{X(u_i-u_j)}{\|u_i-u_j\|}\right\|^2 \notin [1-\delta, 1+\delta] \text{ для какой-то пары } u_i \text{ и } u_j\right) \le n(n-1)e^{-\frac{d\delta^2}{8}}.$$

Нам осталось только разрешить неравенство

$$n(n-1)e^{-\frac{d\delta^2}{8}} \le \varepsilon$$

относительно d. Это даст нам искомый ответ. $\square$

Сумма независимых случайных величин

Остановимся на еще одном применении «аппарата» производящих функций моментов, в этот раз уже к задаче нахождения распределения суммы независимых случайных величин. Начнем с подхода «в лоб».

Допустим, что дискретные случайные величины X и Y независимы и принимают значения от 0 до n. Наша задача найти распределение суммы $X+Y$, то есть вероятности $\mathbb{P}(X+Y=k)$, где k — любое число от 0 до $2n$ (сумма не может превышать $2n$). Заметим, что при выполнении равенства $X=j$, $j \le k$, событие $X+Y=k$ происходит тогда и только тогда, когда $Y=k-j$. Тогда

$$\mathbb{P}(X+Y=k) = \sum_{j=0}^{k}\mathbb{P}(X=j, Y=k-j) = \sum_{j=0}^{k}\mathbb{P}(X=j)\cdot\mathbb{P}(Y=k-j),$$

где в последнем равенстве мы воспользовались независимостью X и Y. Эта формула называется «формулой свертки», ее естественным образом можно обобщить на произвольные дискретные случайные величины. В случае, когда у случайных величин существует плотность, формула выглядит аналогично, только сумму нужно заменить на интеграл:

$$f_{X+Y}(z) = \int_{-\infty}^{+\infty} f_X(u)f_Y(z-u)du = \int_{-\infty}^{+\infty} f_X(z-u)f_Y(u)du,$$

где f_X, f_Y и f_{X+Y} — плотности распределений случайных величин X, Y и $X+Y$ соответственно.

Обратите внимание, чтобы найти распределение суммы независимых случайных величин в дискретном случае необходимо посчитать сумму ряда (если случайные величины принимают счетное количество значений), а в случае, когда существует плотность, — интеграл.

Оказывается, что производящая функция моментов значительно может упростить вычисление распределения суммы. Напомним, что производящая функция моментов суммы независимых случайных величин X и Y равна произведению их производящих функций:

$$M_{X+Y}(\lambda) = M_X(\lambda)\cdot M_Y(\lambda).$$

Кроме того, разным распределениям отвечают разные производящие функции моментов. Эти два свойства можно применить следующим образом.

Пример 2. Пусть даны независимые случайные величины $X \sim \mathcal{N}(a_1, \sigma_1^2)$ и $Y \sim \mathcal{N}(a_2, \sigma_2^2)$. Каким будет распределение у случайной величины $X+Y$?

В Упражнении 5 были посчитаны производящие функции моментов для нормального распреде-

ления
$$M_X(\lambda) = e^{\lambda^2\sigma_1^2/2+\lambda a_1} \quad \text{и} \quad M_Y(\lambda) = e^{\lambda^2\sigma_2^2/2+\lambda a_2}.$$
Тогда по выше упомянутому свойству
$$M_{X+Y}(\lambda) = M_X(\lambda) \cdot M_Y(\lambda) = e^{\lambda^2(\sigma_1^2+\sigma_2^2)/2+\lambda(a_1+a_2)}.$$
Что совпадает с производящей функцией моментов для распределения $\mathcal{N}(a_1+a_2, \sigma_1^2+\sigma_2^2)$. Далее все следует из единственности производящей функции моментов: если бы у суммы $X+Y$ распределение было бы отлично от $\mathcal{N}(a_1+a_2, \sigma_1^2+\sigma_2^2)$, то у нее была бы другая производящая функция моментов. Следовательно, $X+Y \sim \mathcal{N}(a_1+a_2, \sigma_1^2+\sigma_2^2)$, то есть у суммы независимых гауссовских случайных величин распределение тоже будет гауссовским и, более того, параметры a и σ^2 у суммы будут в точности равны сумме параметров a_i и σ_i^2 соответственно.

Список литературы

[1] К. Л. Чжун, Ф. АитСахлиа. Элементарный курс теории вероятностей. Стохастические процессы и финансовая математика. Бином, 2011.

[2] Н. И. Чернова. Теория вероятностей. Учебное пособие. Новосибирск, 2009.

[3] М. Я. Кельберт, Ю. М. Сухов. Вероятность и статистика в примерах и задачах. Том 1. МЦНМО, 2007.