Wouter Koolen
Machine Learning Group
Centrum Wiskunde & Informatica, Amsterdam
wmkoolen@cwi.nl

Sub-Gaussians in game-theoretic probability

In their elegant *Game-Theoretic Probability* framework, Shafer and Vovk interpret probabilistic assumptions as the availability of certain elementary bets, and derive probabilistic consequences by strategically combining these bets. In this article Veni laureate Wouter Koolen of the Machine Learning Group at CWI takes this framework for a spin by deriving (game-theoretic) deviation inequalities for sub-Gaussian random variables.

The purpose of this article is threefold. First, I will derive deviation inequalities for sub-Gaussian random variables. Such statements find application in statistics and machine learning, for example in hypothesis tests, confidence intervals and optional stopping. So if you have not seen sub-Gaussian variables or deviation inequalities (or both) before, this will be useful. Second, these results will illustrate how one can systematically exploit the (weak) assumption that the distribution belongs to a given set. And finally, the unified way in which the results are derived illustrates the power and intuitiveness of the game-theoretic probability framework.

The article is structured as follows. We will first review the sub-Gaussianity assumption, and investigate its game-theoretic interpretation as a collection of available bets. Subsequently, assuming that S is sub-Gaussian, we will construct betting strategies that lead to upper bounds on $\mathbb{P}\{S \geq c\}$, $\mathbb{P}\{S^2 \geq c\}$, $\mathbb{E}[e^{\lambda S}]$, $\mathbb{E}[e^{\lambda S^2}]$ as well as on $\mathbb{P}\{\sum_{i=1}^K S_i^2 \geq c\}$ for independent sub-Gaussian S_i. Each of these bounds expresses the intuition that S cannot be extreme with high probability.

Setup

The goal is to showcase the game-theoretic probability framework and techniques, and the associated way of thinking in terms of intuitive bets. Moreover, as we will see the constructions will come with natural certificates of tightness.

A minimal assumption

In this article we will not commit to a single distribution, but instead work with a (non-parametric) class of distributions. Here, the assumption that we are willing to make is that of sub-Gaussianity. Let's review the definition.

Definition. A variable S is *t-sub-Gaussian* if for each $\eta \in \mathbb{R}$,

$$\mathbb{E}[e^{\eta S}] \leq e^{t\eta^2/2}. \tag{1}$$

Sub-Gaussian random variables are ubiquitous. They are often used as models for noise (sub-Gaussianity implies mean zero). The centred Gaussian distribution with variance t satisfies (1) with equality, hence the name. Moreover, by Hoeffding's Inequality, any zero-mean random variable with bounded support $[a,b]$ is $(b-a)^2/4$-sub-Gaussian. For independent t_i-sub-Gaussian S_i, the sum $\sum_i S_i$ is $\sum_i t_i$-sub-Gaussian. For any $\alpha \in \mathbb{R}$ the scaling αS is $\alpha^2 t$-sub-Gaussian. The canonical t-sub-Gaussian random variable in applications is $S = \sum_{i=1}^t (X_i - \mu)$ where the $X_i - \mu$ are independent 1-sub-Gaussian.

Next, we will review the game-theoretic-probability interpretation of our assumption (1).

Game-theoretic probability

Following [3], we interpret the sub-Gaussian condition (1) as a collection of bets that are offered regarding S. Namely, for each $\eta \in \mathbb{R}$ we can buy any positive number of η-tickets that cost $e^{t\eta^2/2}$ and pay out $e^{\eta S}$. In other words, each unit capital invested in η-tickets yields $e^{\eta S - t\eta^2/2}$. A *strategy* for the learner is a *portfolio*, specified by a positive measure $p(\eta) \geq 0$, indicating how much capital should be invested in η-tickets for each $\eta \in \mathbb{R}$. (Our strategies are relatively simple as we are considering just a single round. See [3] for general multi-round protocols. We will use the notation for densities throughout for simplicity, although we will find we need both continuous and discrete measures.) The cost of the portfolio $p(\eta)$ is hence

$$\int p(\eta)\, d\eta,$$

and for each possible outcome $S \in \mathbb{R}$, its payoff is

$$\int p(\eta)\, e^{\eta S - t\eta^2/2} d\eta.$$

We will use portfolios to price arbitrary variables.

Upper price

Our goal is to show that S cannot be extreme with high probability. Our approach will be to fix a function $Y(S)$, expressing how extreme we deem S to be. The mechanism is then to construct a portfolio of tickets such that we end up with payoff at least $Y(S)$ no matter the outcome S. Given that all bets are fair at best, it will be highly unlikely that the strategy pays off significantly more than its cost. Formally, we define the *upper price* of Y to be the minimum cost portfolio

$$\bar{\mathbb{E}}[Y] := \min_{p(\eta)\geq 0} \int p(\eta)\, d\eta \qquad (2a)$$

subject to the 'super-replication' constraint

$$\forall S \in \mathbb{R}: \ \int p(\eta)\, e^{\eta S - t\eta^2/2} d\eta \geq Y(S). \qquad (2b)$$

As the name suggests, the upper price bounds the expectation from above. To see why, fix any optimiser p^* of (2). Taking expectation of (2b) under any t-sub-Gaussian distribution on S, we find

$$\begin{aligned}\mathbb{E}[Y] &\leq \int p^*(\eta)\, \mathbb{E}\big[e^{\eta S - t\eta^2/2}\big] d\eta \\ &\leq \int p^*(\eta)\, d\eta = \bar{\mathbb{E}}[Y].\end{aligned}$$

Now how to approach the optimisation problem above?

Duality

As the objective and constraint in (2) are *linear* in the portfolio $p(\eta)$, this is an (infinite) linear program. Like finite linear programs, this problem has an associated *dual problem* where the role of variables and constrains are swapped. (Duality is a rich concept in mathematical optimisation, see for example [1]. A useful analogy is perhaps the simplest duality relation, namely that the *maximum* over a set is also the *minimum* number that is larger than each member.) In our case the dual problem asks for a positive measure $b(S)$ on outcomes S that maximises

$$\max_{b(S)\geq 0} \int b(S)\, Y(S)\, dS \qquad (3a)$$

subject to the 'fair ticket pricing' constraint

$$\forall \eta \in \mathbb{R}: \ \int e^{\eta S - \frac{t\eta^2}{2}} b(S)\, dS \leq 1. \qquad (3b)$$

We will be using duality to certify optimality. A pair $p^*(\eta) \geq 0$ and $b^*(S) \geq 0$ satisfying the constraints (2b) and (3b) for which the values (2a) and (3a) coincide simultaneously solves (2) and (3) to optimality. We will call such a pair a *saddle point*.

Applications in the univariate case

We now use the above duality relationship to compute the sub-Gaussian upper price $\bar{\mathbb{E}}[Y]$ of the five variables Y of interest from the introduction. Throughout we assume that S is t-sub-Gaussian.

One-sided tail

Fix a threshold $c \geq 0$, and let

$$Y := \mathbb{1}\{S \geq c\}.$$

We claim that the upper price is $\bar{\mathbb{E}}[Y] = e^{-\frac{c^2}{2t}}$, as witnessed by the following saddle point:

$$p^* = e^{-\frac{c^2}{2t}} \delta_{\eta = \frac{c}{t}} \quad \text{and} \quad b^* = e^{-\frac{c^2}{2t}} \delta_{S=c},$$

where we write $\delta_{S=c}$ for the Dirac point-mass at c. Primal feasibility (2b) follows from

$$\min_{S\geq c} e^{-\frac{c^2}{2t}} e^{\frac{c}{t}S - \frac{c^2}{2t}} = 1$$

and dual feasibility (3b) from

$$\max_{\eta} e^{-\frac{c^2}{2t}} e^{\eta c - \frac{\eta^2}{2} t} = 1.$$

(Exercise: what are upper price $\bar{\mathbb{E}}[Y]$ and saddle point for $c < 0$?)

Primal optimality tells us that $\mathbb{P}\{S \geq c\} \leq e^{-\frac{c^2}{2t}}$, while dual optimality tells us that we cannot prove a tighter bound without changing the assumptions or technique. For example, if we know that $S \sim \mathcal{N}(0,t)$ is Gaussian, we find $\mathbb{P}\{S \geq c\} = \Psi(-c/\sqrt{t})$.

Two-sided tail

Fix $c \geq 0$. Let's look at the two-sided threshold

$$Y := \mathbb{1}\{S^2 \geq c\}. \qquad (4)$$

It would be natural to conjecture that the upper price $\bar{\mathbb{E}}[Y]$ is just twice that of a single tail. But in actuality it is less, especially so for small c. To say what it is exactly, let v and z be the value and optimiser of

$$v = \max_{z} \cosh(z\sqrt{c})\, e^{-\frac{z^2}{2}t}.$$

We claim that the value is $\bar{\mathbb{E}}[Y] = \frac{1}{v}$, as witnessed by the saddle point

$$p^* = \frac{\delta_{\eta=z} + \delta_{\eta=-z}}{2v} \ \text{and}$$
$$b^* = \frac{\delta_{S=\sqrt{c}} + \delta_{S=-\sqrt{c}}}{2v}. \qquad (5)$$

Let's first check dual feasibility,

$$\begin{aligned}&\max_{\eta} \frac{e^{\eta\sqrt{c} - \frac{\eta^2}{2}t} + e^{-\eta\sqrt{c} - \frac{\eta^2}{2}t}}{2v} \\ &= \frac{\max_\eta \cosh(\eta\sqrt{c})\, e^{-\frac{\eta^2}{2}t}}{v} = 1.\end{aligned}$$

And let's check primal feasibility,

$$\begin{aligned}&\min_{S^2\geq c} \frac{e^{zS - \frac{z^2}{2}t} + e^{-zS - \frac{z^2}{2}t}}{2v} \\ &= \frac{1}{v} \min_{S^2 \geq c} \cosh(zS)\, e^{-\frac{z^2}{2}t} \\ &= \frac{1}{v} \cosh(z\sqrt{c})\, e^{-\frac{z^2}{2}t} = 1.\end{aligned}$$

For exact Gaussian $S \sim \mathcal{N}(0,t)$, we have $\mathbb{P}\{S^2 \geq c\} = 2\Psi(-\sqrt{c/t})$.

Moment Generating Function

In the previous two sections we quantified that S cannot be extreme by giving upper bounds on probabilities of its tail events. Another way of expressing that S cannot be extreme is to bound its moment generating function. (Tail bounds would then follow by Chernoff's method). Fix $\lambda \in \mathbb{R}$. Let's consider

$$Y := e^{\lambda S}.$$

For a centred Gaussian $S \sim \mathcal{N}(0,t)$ with variance t, we would find $\mathbb{E}[e^{\lambda S}] = e^{t\lambda^2/2}$. Here we show that $\bar{\mathbb{E}}[Y] = e^{t\lambda^2/2}$ for t-sub-Gaussian S as well. The following is a witnessing saddle point

$$p^* = e^{\frac{\lambda^2}{2}t} \delta_{\eta=\lambda} \quad \text{and} \quad b^* = e^{\frac{\lambda^2}{2}t} \delta_{S=t\lambda}.$$

Dual feasibility follows by

$$\max_{\eta} e^{\frac{\lambda^2}{2}t} e^{\eta t\lambda - \frac{\eta^2}{2}t} = 1,$$

while primal feasibility is established by

$$\forall S \in \mathbb{R}: \ e^{\frac{\lambda^2}{2}t} e^{\lambda S - \frac{\lambda^2}{2}t} = e^{\lambda S}.$$

We find that the upper moment-generating function $\bar{\mathbb{E}}[e^{\lambda S}]$ is exactly that of a Gaussian. Hence the generalisation to sub-Gaussian comes for free.

Moment generating function of square

Now it gets interesting. Fix $\lambda \in [0,1]$. Let's consider

$$Y := e^{\lambda \frac{S^2}{2t}}. \qquad (6)$$

In contrast to what happened before, here the supports of the components of the saddle point are *continuous measures*. We claim the value is

$$\bar{\mathbb{E}}[Y] = \frac{1}{\sqrt{1-\lambda}}$$

as witnessed by the saddle point

$$p^* = \frac{1}{\sqrt{1-\lambda}}\mathcal{N}\Big(0, \frac{\lambda}{t(1-\lambda)}\Big) \text{ and}$$
$$b^* = \mathcal{N}(0,t).$$

Let's check primal feasibility. For all S,

$$\int e^{\eta S - \frac{\eta^2}{2}t} p^*(\eta)\, d\eta = e^{\lambda \frac{S^2}{2t}}.$$

Now let's check dual feasibility. For all η,

$$\int e^{\eta S - \frac{\eta^2}{2}t} b^*(S)\, dS = 1.$$

Finally, the values indeed agree and are equal to

$$\begin{aligned}\bar{\mathbb{E}}[Y] &= \int p^*(\eta)\, d\eta \\ &= \int e^{\lambda \frac{S^2}{2t}} b^*(S)\, dS = \frac{1}{\sqrt{1-\lambda}}.\end{aligned}$$

Interestingly, if S is Gaussian then S^2/t has a χ^2 distribution, and hence the moment-generating function of $\frac{S^2}{2t}$ is equal to $(1-\lambda)^{-1/2}$. So we are not losing anything by generalising to sub-Gaussian.

Application in the multivariate case

We conclude the exposition by looking at the simplest multi-variate case. For here something very interesting happens. The setup will be as follows. We consider independent $S_1, \ldots, S_K$ where S_i is t_i-sub-Gaussian. The joint outcome $\boldsymbol{S} = (S_1, \ldots, S_K)$ will be revealed at once, so there is no sequentiality to the problem. Before it is revealed, we can engage in a collection of bets on the outcome. For every $\boldsymbol{\eta} \in \mathbb{R}^K$, we will be able to buy any number of $\boldsymbol{\eta}$-ticket, which each pay off $\prod_{i=1}^K e^{\eta_i S_i}$ and cost $\prod_{i=1}^K e^{\frac{\eta_i^2}{2} t_i}$. So now a strategy for the learner is a positive measure $p(\boldsymbol{\eta})$ on $\mathbb{R}^K$. We will be interested in the statistic

$$Z := \sum_{i=1}^K \frac{S_i^2}{2t_i}.$$

This statistic arises for example as the maximum log-likelihood value when comparing arbitrary mean models with mean zero models.

Products

The univariate price for $e^{\lambda \frac{S^2}{2t}}$ developed below (6) immediately gives us a price for the product $\prod_{i=1}^K e^{\lambda \frac{S_i^2}{2t_i}} = e^{\lambda Z}$, namely

$$\bar{\mathbb{E}}[e^{\lambda Z}] = (1-\lambda)^{-K/2}.$$

Self-normalised sums of squares

We finally consider thresholding Z in the form of

$$Y := \mathbb{1}\{Z \geq c\}.$$

The upper price $\bar{\mathbb{E}}[Y]$ and witnessing strategies will need to generalise those below (4), which cover the case $K = 1$. This indeed happens, but in a curious way. Namely, the general pattern is to have $p(\boldsymbol{\eta})$ and $b(\boldsymbol{S})$ mix over certain ellipses. In the special case $K = 1$ we indeed recover the mixtures over 2 symmetrically placed points that we found in (5). To express the result, let v and d be the value and optimiser of

$$v = \max_{d \geq 0} e^{-d} \mathbb{E}_q\big[e^{2\sqrt{cd}q}\big]$$

where $q \in [-1, 1]$ has density

$$\frac{\Gamma\left(\frac{K}{2}\right)}{\sqrt{\pi}\,\Gamma\left(\frac{K-1}{2}\right)}(1-q^2)^{\frac{K-3}{2}}, \tag{7}$$

which we may recognise as the marginal density of a point drawn uniformly from the unit sphere $\mathcal{S}^K$ in $\mathbb{R}^K$. The final claim is that the upper price is $\bar{\mathbb{E}}[Y] = \frac{1}{v}$, as witnessed by the saddle point

$$p^* = \frac{1}{v}\mathcal{L}\Big(\boldsymbol{\eta}\colon \eta_i = \sqrt{2dt_i^{-1}}\, Y_i \text{ where } \boldsymbol{Y} \overset{\text{unif}}{\sim} \mathcal{S}^K\Big),$$
$$b^* = \frac{1}{v}\mathcal{L}\Big(\boldsymbol{S}\colon S_i = \sqrt{2ct_i}\, X_i \text{ where } \boldsymbol{X} \overset{\text{unif}}{\sim} \mathcal{S}^K\Big),$$

where $\boldsymbol{X}$ and $\boldsymbol{Y}$ are uniformly distributed on the unit sphere. Here $\mathcal{L}(\cdot)$ denotes the law of the sampling procedure specified in the argument.

First, let's check dual feasibility. For all $\boldsymbol{\eta}$, abbreviating $z = \frac{1}{2}\sum_i \eta_i^2 t_i$ and $q = X_1$ (which has the density given in (7) above),

$$\begin{aligned}\frac{1}{v}\mathbb{E}_{\boldsymbol{S}}\Big[e^{\sum_i\left(\eta_i S_i - \frac{\eta_i^2}{2}t_i\right)}\Big] &= \frac{1}{v}\mathbb{E}_{\boldsymbol{X}}\big[e^{\sqrt{2c}\sum_i \eta_i \sqrt{t_i} X_i - z}\big] \\ &= \frac{1}{v}\mathbb{E}_{\boldsymbol{X}}\big[e^{2\sqrt{cz}X_1 - z}\big] \\ &= \frac{1}{v}e^{-z}\mathbb{E}_q\big[e^{2\sqrt{cz}q}\big] \leq 1.\end{aligned}$$

Okay, good. Now primal feasibility. We have

$$\begin{aligned}\frac{1}{v}\mathbb{E}_{\boldsymbol{\eta}} e^{\sum_i\left(\eta_i S_i - \frac{\eta_i^2}{2}t_i\right)} &= \frac{1}{v}\mathbb{E}_{\boldsymbol{Y}}\Big[e^{2\sqrt{d}\sum_i \frac{S_i}{\sqrt{2t_i}} Y_i - d}\Big] \\ &= \frac{1}{v}\mathbb{E}_{\boldsymbol{Y}}\Big[e^{2\sqrt{d}\sqrt{\sum_i \frac{S_i^2}{2t_i}}\, Y_1 - d}\Big] \\ &\geq \frac{1}{v}\mathbb{E}_q\big[e^{2\sqrt{cd}q - d}\big] = 1.\end{aligned}$$

In both cases the crucial step is to use rotational symmetry: for $\boldsymbol{c} \in \mathbb{R}^K$, the inner product $\sum_i c_i X_i$ has the same distribution as $\|c\| X_1$. Finally, note that

$$\mathbb{E}_q\big[e^{2\sqrt{s}q}\big] = {}_0F_1\Big(\frac{K}{2}, s\Big).$$

is the confluent hypergeometric limit function, for which computer support is readily available. For example, Mathematica calls it `Hypergeometric0F1`, Matlab calls it `hypergeom` and Octave has `gsl_sf_hyperg_0F1` in package `gsl`.

We found $\bar{\mathbb{E}}[Y] = \frac{1}{v}$, and hence $\mathbb{P}\{Z \geq c\} \leq \frac{1}{v}$. We can also reason backwards and find the threshold c corresponding to a given confidence $\frac{1}{v} = \delta$. We obtain

$$\begin{aligned}c^* &= \min\Big\{c \,\Big|\, \max_{d \geq 0} e^{-d}\mathbb{E}_q\big[e^{2\sqrt{cd}q}\big] \geq \frac{1}{\delta}\Big\} \\ &= \inf_{s \geq 0} \frac{s}{\ln_+\big(\delta \mathbb{E}_q\big[e^{2\sqrt{s}q}\big]\big)},\end{aligned}$$

which can be implemented numerically using a binary search for the zero of the derivative.

Conclusion

We illustrated the power of the game-theoretic probability framework by deriving in a uniform fashion a series of deviation inequalities for sub-Gaussian random variables. We covered just the tip of a giant (and partially unexplored) iceberg. In more advanced sequential settings, taking bets and observing outcomes are interleaved, and more elaborate strategies beyond mixtures are possible and necessary, naturally leading to martingales. Analogues of the methods showcased here can be used to prove more advanced deviation inequalities that e.g. hold for arbitrary exponential families, and hold uniformly over time [2].

Acknowledgements

This article benefited from discussions with Emilie Kaufmann (Inria Lille), Aurélien Garivier (IMT Toulouse) and Peter Grünwald (CWI).

References

1 S. Boyd and L. Vandenberghe, *Convex Optimization*, Cambridge University Press, 2004.

2 E. Kaufmann, W.M. Koolen and A. Garivier, Sequential test for the lowest mean: From Thompson to Murphy sampling, arXiv: 1806.00973, 2018.

3 G. Shafer and V. Vovk, *Probability and Finance — It's Only a Game!*, Wiley, 2001.