Bounds on the Expectation of the Maximum of Samples from a Gaussian

Gautam Kamath

In this document, we will provide bounds on the expected maximum of n samples from a Gaussian distribution.

Theorem 1. *Let* $Y = \max_{1 \leq i \leq n} X_i$, *where* $X_i \sim \mathcal{N}(0, \sigma^2)$ *are i.i.d. random variables. Then*

$$\frac{1}{\sqrt{\pi \log 2}} \sigma \sqrt{\log n} \leq \mathbf{E}[Y] \leq \sqrt{2} \sigma \sqrt{\log n}.$$

We comment that both constants which multiply $\sigma \sqrt{\log n}$ are tight. Indeed, as $n \to \infty$, $\mathbf{E}[Y]/\sqrt{\log n}$ converges to $\sqrt{2}\sigma$. On the other hand, by explicit calculations, one can verify the cases $n = 1$ and 2, for which $\mathbf{E}[Y]$ is 0 and $\sigma/\sqrt{\pi}$, respectively. In the former case, the inequality trivially holds for any multiplying constant, and in the latter, our inequality is tight.

First, we show $\mathbf{E}[Y] \leq \sigma\sqrt{2}\sqrt{\log n}$. This result and method are folklore, but we include them here for completeness.

$$
\begin{aligned}
\exp\left(t\mathbf{E}[Y]\right) &\leq \mathbf{E}[\exp\left(tY\right)] \\
&= \mathbf{E}[\max \exp\left(tX_i\right)] \\
&\leq \sum_{i=1}^{n} \mathbf{E}[\exp\left(tX_i\right)] \\
&= n \exp\left(t^2\sigma^2/2\right)
\end{aligned}
$$

The first inequality is Jensen's inequality, the second is the union bound, and the final equality follows from the definition of the moment generating function.

Taking the logarithm of both sides of this inequality, we get

$$\mathbf{E}[Y] \leq \frac{\log n}{t} + \frac{t\sigma^2}{2}.$$

This can be minimized by setting $t = \frac{\sqrt{2 \log n}}{\sigma}$, which gives us the desired result

$$\mathbf{E}[Y] \leq \sigma\sqrt{2}\sqrt{\log n}.$$

Next, we show the more difficult direction, the lower bound. We have already established that it holds for $n = 1$ and 2. It can be verified for $n = 3$ to 2834 using the Python 3 code provided in Section A. Thus, for the remainder of the proof, we assume $n \geq 2835$.

Note that we have the following crude bound, which uses the Chernoff bound and the lower bound on n:

$$
\mathbf{E}[Y] \geq \Pr(|\{i \ : \ X_i \geq 0\}| \geq \lceil n/3 \rceil) \cdot \mathbf{E}\left[\max_{1 \leq i \leq \lceil n/3 \rceil} |X_i|\right] + \Pr(|\{i \ : \ X_i \geq 0\}| < \lceil n/3 \rceil) \cdot \mathbf{E}[-|X_i|]
$$

$$
\geq 0.999 \cdot \mathbf{E}\left[\max_{1 \leq i \leq \lceil n/3 \rceil} |X_i|\right] - 0.001\sigma \cdot \sqrt{\frac{2}{\pi}}.
$$

The second inequality uses the expected value of the half-normal distribution.

It remains to lower bound $\mathbf{E}\left[\max_{1\leq i\leq k}|X_i|\right]$. We will show that

$$\Pr(|X_i|\geq \sigma\sqrt{\log n})\geq \frac{9}{n}.$$

This will imply the following lower bound:

$$\mathbf{E}\left[\max_{1\leq i\leq \lceil n/3\rceil}|X_i|\right]\geq \sigma\sqrt{\log n}\cdot \Pr(\exists i:|X_i|\geq \sigma\sqrt{\log n})$$

$$\geq \sigma\sqrt{\log n}\cdot\left(1-\left(1-\frac{9}{n}\right)^{\lceil n/3\rceil}\right)$$

$$\geq \left(1-\frac{1}{e^2}\right)\sigma\sqrt{\log n}$$

We compute the CDF of $|X_i|$ at the point $\sigma\sqrt{\log n}$.

$$\Pr(|X_i|\geq \sigma\sqrt{\log n})=1-\mathrm{erf}\left(\frac{\sqrt{\log n}}{\sqrt{2}}\right)$$

$$\geq 1-\sqrt{1-\exp\left(-\frac{2}{\pi}\log n\right)}$$

$$=1-\sqrt{1-n^{-\frac{2}{\pi}}},$$

where the first equality is based on the CDF of the half-normal distribution and the inequality is from the bound on the error function, $\mathrm{erf}(x)\leq \sqrt{1-\exp\left(-\frac{4}{\pi}x^2\right)}$ [Bul]. We require this value to be at least $\frac{9}{n}$:

$$1-\sqrt{1-n^{-\frac{2}{\pi}}}\geq \frac{9}{n} \qquad\qquad \Leftrightarrow$$

$$1-\frac{9}{n}\geq \sqrt{1-n^{-\frac{2}{\pi}}} \qquad\qquad \Leftrightarrow$$

$$1-\frac{18}{n}+\frac{81}{n^2}\geq 1-\frac{1}{n^{\frac{2}{\pi}}} \qquad\qquad \Leftrightarrow$$

$$n^{2-\frac{2}{\pi}}\geq 18n-81 \qquad\qquad \Leftrightarrow$$

$$\left(2-\frac{2}{\pi}\right)\log n\geq \log(18n-81) \qquad\qquad \Leftrightarrow$$

$$\left(2-\frac{2}{\pi}\right)\frac{\log n}{\log(18n-81)}\geq 1$$

This inequality holds for all $n\geq 2835$, as desired.

Putting these inequalities together, we have

$$\mathbf{E}[Y]\geq 0.999\left(1-\frac{1}{e^2}\right)\sigma\sqrt{\log n}-0.001\sigma\cdot\sqrt{\frac{2}{\pi}}\geq \frac{1}{\sqrt{\pi\log 2}}\sigma\sqrt{\log n},$$

where the second inequality holds for any integer $n>1$.

Other proofs. Proofs of qualitatively similar lower bounds also appear in [vH14] and [OP15].

Acknowledgments

References

[Bul]   Yaroslav Bulatov.   Proving $1 - \exp(-4x^2/\pi) \geq \mathrm{erf}(x)^2$.   Mathematics Stack Exchange. URL:https://math.stackexchange.com/q/6908 (version: 2018-12-27).

[OP15] Francesco Orabona and Dávid Pál. Optimal non-asymptotic lower bound on the minimax regret of learning with expert advice. *arXiv preprint arXiv:1511.02176*, 2015.

[vH14] Ramon van Handel. Probability in high dimension. Technical report, Princeton University, 2014.

A   Code for verifying small n

We derive a simple method for computing the expected maximum of a set of random variables. This will not, in general, give closed form expressions, but will be amenable to numerical evaluation.

Let $F_X(x)$ be the CDF of a random variable X, i.e., $F_X(x) = \Pr[X \leq x]$. If $Y = \max_{1 \leq i \leq n} X_i$ where the X_i are all i.i.d. according to X, then this simplifies to $F_Y(x) = F_X(x)^n$. Taking the derivative of the CDF gives us the PDF f, and thus by chain rule, $f_Y(x) = n f_X(x) F_X(x)^{n-1}$. Integrating over the entire domain gives us the expected value, and thus

$$\mathbf{E}[Y] = \int_{-\infty}^{\infty} x n f_X(x) F_X(x)^{n-1} dx.$$

We substitute in the following expressions for $N(0,1)$:

$$f_X(x) = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{x^2}{2}\right);$$

$$F_X(x) = \frac{1}{2}\left[1 + \mathrm{erf}\left(\frac{x}{\sqrt{2}}\right)\right],$$

giving the following expression:

$$\mathbf{E}[Y] = \int_{-\infty}^{\infty} x \cdot n \cdot \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{x^2}{2}\right) \cdot \left(\frac{1}{2}\left[1 + \mathrm{erf}\left(\frac{x}{\sqrt{2}}\right)\right]\right)^{n-1} dx.$$

As mentioned before, it is not generally possible to evaluate this integral in closed form. Instead, we use numerical integration from Python 3's SciPy library. In particular, we verify the theorem for $n = 3$ to 2834.

```python
import scipy.integrate as integrate
import scipy.special as special
import numpy as np
import math

for n in range(3,2835):
  result = integrate.quad(lambda x: x*n*(1/np.sqrt(2*math.pi))*math.exp(-math.pow(x,2)/2)\
  *math.pow(0.5*(1 + special.erf(x/np.sqrt(2))),n-1), -np.inf, np.inf)
  if ((result[0]-result[1])/np.sqrt(np.log(n))) <= 1/np.sqrt(math.pi*np.log(2)):
    print("The inequality may not be true for n = " + str(n))
```