

# Derivation: How much data do I need?

A rederivation for @nntaleb's Technical Incerto Chapter 8.

[Updated 2021-02-04] Please help debug my derivation! I get an extra term which alters the required sample size.

## Equation Definition

- Let  $MAD(n)$  be mean absolute deviation of  $n$  summed variables.
- Let  $X_{g1}, X_{g2}, X_{g3}, \dots$  be gaussian variables with mean 0 and scale 1 (mean absolute deviation)
- Let  $X_{v1}, X_{v2}, X_{v3}, \dots$  be **non-gaussian** variables with mean 0 and scale 1
- The convergence speed is  $\kappa_{1,n}$ . It is defined as the value which solves the following:

$$MAD(n) = MAD(1) n^{\left(\frac{1}{2-\kappa_{1,n}}\right)} \quad (1)$$

**Equation 2 tells you what sample size  $n_v$  you need.** To be precise,  $n_v$  is the number of samples you need so that your sample average has the same mean absolute deviation as  $n_g$  gaussian variables.

$$n_v = n_g^{-\frac{1}{\kappa_{1,n_g}-1}} \quad (2)$$

## Derivation Attempt

Both the gaussian and non-gaussian distributions are normalized to  $MAD(1)=1$ .

We know that  $\kappa_{1,n} = 0$  for the Gaussian. Inserting this to Equation (1) gives:

$$MAD(n_g) = \sqrt{n_g} \quad (3)$$

We are looking for the value  $n_v$  which makes the MAD of the sample averages match, i.e.:

$$\frac{MAD(n_v)}{n_v} = \frac{MAD(n_g)}{n_g} \quad (4)$$

Inserting Equations (1) and (3) into Equation (4) gives:

$$\frac{n_v^{\left(\frac{1}{2-\kappa_{1,n_v}}\right)}}{n_v} = \frac{\sqrt{n_g}}{n_g} \quad (5)$$

This Simplifies to:

$$n_v^{\left(\frac{1}{2-\kappa_{1,n_v}} - \frac{2-\kappa_{1,n_v}}{2-\kappa_{1,n_v}}\right)} = n_g^{-\frac{1}{2}} \quad (6)$$

Then to:

$$n_v^{\left(\frac{\kappa_{1,n_v}-1}{2-\kappa_{1,n_v}}\right)} = n_g^{-\frac{1}{2}} \quad (7)$$

And then to:

$$n_v = n_g^{-\frac{1}{2}\left(\frac{2-\kappa_{1,n_v}}{\kappa_{1,n_v}-1}\right)} \quad (8)$$

Finally, to:

$$n_v = n_g^{-\left(\frac{1-0.5\kappa_{1,n_v}}{\kappa_{1,n_v}-1}\right)} \quad (9)$$

## Conclusion

1. There is an extra  $-0.5\kappa_{1,n_v}$  term in my derivation compared to Equation (2).
  - This reduces the required sample size  $n_v$  (In a significant way)
2. The subscript in my derivation is different. I have  $\kappa_{1,n_v}$  instead of  $\kappa_{1,n_g}$ .
  - The subscript in my example implies smaller sample size  $n_v$

Let me know if you understand why my derivation doesn't match the book!