

# Don't Compare Averages

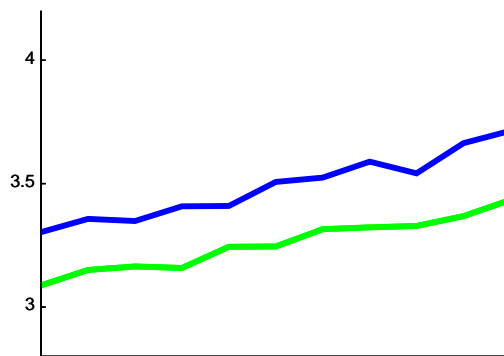
Holger Bast and Ingmar Weber

Max-Planck-Institut für Informatik  
Saarbrücken, Germany  
bast@mpi-sb.mpg.de  
iweber@mpi-sb.mpg.de

**Abstract.** We point out that for two sets of measurements, it can happen that the average of one set is larger than the average of the other set on one scale, but becomes smaller after a non-linear monotone transformation of the individual measurements. We show that the inclusion of error bars is no safeguard against this phenomenon. We give a theorem, however, that limits the amount of “reversal” that can occur; as a by-product we get two non-standard one-sided tail estimates for arbitrary random variables which may be of independent interest. Our findings suggest that in the not infrequent situation where more than one cost measure makes sense, there is no alternative other than to explicitly compare averages for each of them, much unlike what is common practice. The presentation at the workshop will have a guaranteed surprise effect!

## 1 Introduction

Fig. 1 shows us a typical performance statistic as we find it in many papers.

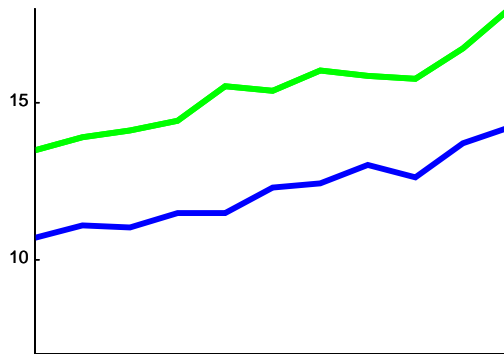


**Fig. 1.** The green algorithm is clearly better than the blue one ...

For the sake of concreteness, let us assume that the two graphs pertain to two different numerical algorithms that compute with large integers, and that it was

measured how large these numbers get in the internal computations. More precisely, the number of bits needed to represent the largest integer were measured, and each point in the graph is actually an average taken over a number of problem instances. The fewer bits, the better of course. Along the  $x$ -axis the input size is varied. The message conveyed by the figure is clear: the “green (light gray)” algorithm performs consistently, that is for all considered problem sizes, about 10% better than the “blue (dark gray)” algorithm.

Now the cost measure is somewhat arbitrary in the sense that we might as well have chosen to record the largest integer used and not the number of bits used to represent it, that is, to consider costs  $2^c$  instead of costs  $c$ . What graph do we expect then? Well, if on some instance the one algorithm needs 3 bits and the other 4 bits, the modified costs would be  $2^3 = 8$  versus  $2^4 = 16$ , that is, not surprisingly the gap between the two becomes larger. Now let us take a look at the graph for the same data but with the modified cost measure.



**Fig. 2.** ... or isn't it?

Indeed, the gap has increased (from 10% to about 30%), but moreover, *the order of the two graphs has changed!* How is that possible?

There is, of course, nothing wrong with the figures, which are from authentic data; details are given in Appendix A. The reason for the reversal is that for two random variables  $X$  and  $Y$ ,  $\mathbf{E}X \leq \mathbf{E}Y$  does *not*, in general, imply that for an (even strictly) increasing function  $f$ ,  $\mathbf{E}f(X) \leq \mathbf{E}f(Y)$ . For a simple counterexample, consider two runs of our two algorithms above, where the first algorithm required once 1 and once 5 bits, and the second algorithm required 4 bits twice. Then clearly, on average the first algorithm required *one bit less*. Considering the second cost measure, the first algorithm on average required numbers up to  $(2^1 + 2^5)/2 = 17$ , which is *one more* than the  $(2^4 + 2^4)/2 = 16$  required by the second algorithm.

Alternative cost measures are actually quite frequent: to assess the quality of a language model, for example, both *cross-entropy* ( $c$ ) and *perplexity* ( $2^c$ ) are equally meaningful and both used frequently [1]. Example publications with comparison graphs (or tables) of the very same kind as in Fig. 1 and 2 are [2] [3] [4] [1]. To give a concrete numerical example also, one of these papers in one of their graphs states average perplexities of  $\approx 3200$  and  $\approx 2900$  for two competing methods. This appears to indicate a solid 10%-improvement of the one method over the other, but first, note that the difference in cross-entropy is merely 0.14, and second, these perplexities would also result if, for example, the cross-entropies were normally distributed with a mean and standard deviation of 11.4 and 0.6, respectively, for the apparently superior method, and 11.3 and 1.0 for the apparently inferior method; detailed calculations can be found in Appendix A.

But language modeling is just one prominent example. Another frequent scenario is that one (basic) algorithm is used as a subroutine in another (more complex) algorithm in such a way that the complexity of the latter depends on the complexity of the former via a non-linear, for example quadratic, function  $f$ . Then, of course, an average complexity of  $c$  of the basic algorithm does not simply translate to an average complexity of  $f(c)$  of the more complex one. But isn't it very tempting to assume that a subroutine with an improved average complexity will at least improve the program that uses it? Well, but that is just not necessarily true.

Now it is (or at least should be) common practice when plotting averages to also provide so-called *error bars*, indicating some average deviation from the average. The following theorem, which is the main result of this paper, says that the "bands" formed by such error bars at least cannot be reversed completely, that is, without intersection, by any monotone transformation  $f$ . As is also stated in the theorem, however, the obvious strengthenings of this statement do *not* hold: for example, it can very well happen that the bands do not intersect in one measure, yet the means reverse in another measure. The theorem is stated in terms of expected *absolute deviations*  $\delta X = \mathbf{E}|X - \mathbf{E}X|$  and  $\delta Y = \mathbf{E}|Y - \mathbf{E}Y|$ , which are never more than the standard deviation; see Fact 1 further down.

**Theorem 1.** *For any two random variables  $X$  and  $Y$ , and for any function  $f$  that is strictly increasing, we have*

$$\mathbf{E}X + \delta X \leq \mathbf{E}Y - \delta Y \implies \mathbf{E}f(X) - \delta f(X) \leq \mathbf{E}f(Y) + \delta f(Y) .$$

*This result cannot be strengthened in the sense that if we drop any one of  $\delta X$ ,  $\delta Y$ ,  $\delta f(X)$ , or  $\delta f(Y)$  to obtain weaker conditions, we can find a counter-example to the statement.*

The proof for Theorem 1, which we give in the following Sect. 2, is elementary but not obvious. Indeed, on their way the authors switched several times between striving for a proof, and being close to finding a counterexample. In Sect. 3, we

give an alternative, more elegant proof in terms of the median. The first proof is more direct, however, while the second proof owes its elegance and brevity to the insight gained from the first; this is why we give the two proofs in that order.

To establish Theorem 1, we will derive two non-standard one-sided tail estimates for general random variables, namely for  $a > 0$ ,

$$\begin{aligned}\Pr(X \geq \mathbf{E}X + a) &\leq \delta X / (2a); \\ \Pr(X \leq \mathbf{E}X - a) &\leq \delta X / (2a) .\end{aligned}$$

These bounds, which are reminiscent of but incomparable to the one-sided version of the Chebyshev inequality (cf. Appendix B), seem to be little known and may be of independent interest.

## 2 Proof of the Main Theorem

All the proofs we give in this paper are for continuous random variables. In all cases it will be obvious how to modify the proofs to work for the discrete case by replacing integrals by sums. For a random variable  $X$ , we write  $\mathbf{E}X$  for its expected value (mean),  $\sigma X$  for its standard deviation, that is  $\sqrt{\mathbf{E}(|X - \mathbf{E}X|^2)}$ , and  $\delta X$  for the mean absolute deviation  $\mathbf{E}|X - \mathbf{E}X|$ . We will throughout assume that these entities exist. The following simple fact relates the two deviation measures.

**Fact 1** *For every random variable  $X$ , it holds that  $\delta X \leq \sigma X$ .*

*Proof.* By Jensen's inequality,  $(\delta X)^2 = (\mathbf{E}|X - \mathbf{E}X|)^2 \leq \mathbf{E}(|X - \mathbf{E}X|^2) = (\sigma X)^2$ .  $\square$

Generally, this inequality will be strict. To get a feeling for the difference, check that for a normal distribution  $N(\mu, \sigma)$  we have  $\delta = \sqrt{2/\pi} \sigma \approx 0.8 \sigma$  and for an exponential distribution  $\text{Exp}(\lambda)$  we have  $\delta = 2/e \sigma = 2/(e \lambda) \approx 0.7 \sigma$ .

As a consequence of Fact 1 all our results still hold if we replace  $\delta$  by  $\sigma$ , that is, we will be proving the stronger form of all results.

We first prove the following non-standard tail estimates, which might be of independent interest. There is a one-sided version of Chebyshev's inequality [5] which looks similar to Lemma 1 below, but the two are incomparable: Lemma 1 is stronger for deviation up to at least  $\sigma X$ , while the Chebyshev tail bounds are stronger for large deviations; see Appendix B.

**Lemma 1.** *For any random variable  $X$  and for every  $a > 0$ , it holds that*

$$\begin{aligned}(a) \quad &\Pr(X \geq \mathbf{E}X + a) \leq \delta X / (2a); \\ (b) \quad &\Pr(X \leq \mathbf{E}X - a) \leq \delta X / (2a) .\end{aligned}$$

*Proof.* Since  $\delta X$  is invariant under shifting  $X$  by a constant, we may assume without loss of generality that  $\mathbf{E}X = 0$ .

Then, with  $\varphi$  denoting the density function pertaining to  $X$ ,

$$\begin{aligned} 0 = \mathbf{E}X &= \int_{-\infty}^0 t \cdot \varphi(t) dt + \int_0^{\infty} t \cdot \varphi(t) dt \\ \delta X &= \int_{-\infty}^0 (-t) \cdot \varphi(t) dt + \int_0^{\infty} t \cdot \varphi(t) dt. \end{aligned}$$

Adding the two equations gives us

$$\begin{aligned} \delta X &= 2 \cdot \int_0^{\infty} t \cdot \varphi(t) dt \\ &\geq 2 \cdot \int_a^{\infty} t \cdot \varphi(t) dt \\ &\geq 2a \cdot \int_a^{\infty} \varphi(t) dt \\ &= 2a \cdot \mathbf{Pr}(X \geq a), \end{aligned}$$

and hence  $\mathbf{Pr}(X \geq a) \leq \delta X / (2a)$ , which establishes (a). The proof for (b) is analogous.  $\square$

Armed with Lemma 1 we can now establish a relation between  $f(\mathbf{E}X)$  and  $\mathbf{E}f(X)$  for a monotone function  $f$ .

**Lemma 2.** *For any random variable  $X$ , and for any function  $f$  that is strictly increasing, it holds that*

$$\begin{aligned} (a) \quad \mathbf{E}f(X) - \delta f(X) &\leq f(\mathbf{E}X + \delta X); \\ (b) \quad \mathbf{E}f(X) + \delta f(X) &\geq f(\mathbf{E}X - \delta X). \end{aligned}$$

*Proof.* Let  $a = \mathbf{E}f(X) - f(\mathbf{E}X + \delta X)$ . If  $a \leq 0$ , there is nothing to show for (a), otherwise two applications of the previous Lemma 1 give us

$$\begin{aligned} 1/2 &\leq \mathbf{Pr}(X \leq \mathbf{E}X + \delta X) \\ &= \mathbf{Pr}(f(X) \leq f(\mathbf{E}X + \delta X)) \\ &= \mathbf{Pr}(f(X) \leq \mathbf{E}f(X) - a) \\ &\leq \delta f(X) / (2a), \end{aligned}$$

and hence  $\mathbf{E}f(X) - f(\mathbf{E}X + \delta X) = a \leq \delta f(X)$ , which is exactly part (a) of the lemma. More generally, we could in fact get that for any  $t$ ,

$$f(t) - \frac{\delta f(X)}{2\mathbf{Pr}(X \geq t)} \leq \mathbf{E}f(X) \leq f(t) + \frac{\delta f(X)}{2\mathbf{Pr}(X \leq t)}.$$

The proof of part (b) is analogous.  $\square$

Theorem 1 is now only two application of Lemma 2 away. Let  $\mathbf{E}X + \delta X \leq \mathbf{E}Y - \delta Y$ , like in the theorem, that is, the “bands” formed by the error bars do not intersect. Then

$$\begin{aligned} \mathbf{E}f(X) - \delta f(X) &\leq f(\mathbf{E}X + \delta X) \\ &\leq f(\mathbf{E}Y - \delta Y) \\ &\leq \mathbf{E}f(Y) + \delta f(Y), \end{aligned}$$

where the first inequality is by part (a) of Lemma 2, the second inequality follows from the monotonicity of  $f$ , and the third inequality is by part (b) of Lemma 2. This finishes the proof of our main theorem.

### 3 The Median

There is an elegant alternative proof of Lemma 2 in terms of the median.

**Fact 2** *For any random variable  $X$  and any strictly monotone function  $f$  we have  $\mathbf{m}f(X) = f(\mathbf{m}X)$ . In the discrete case the medians can be chosen to have this property.*

*Proof.* Simply observe that for any  $a$  we have  $\Pr(X \leq a) = \Pr(f(X) \leq f(a))$ . Here we do require the strict monotonicity.  $\square$

**Fact 3** *For any random variable  $X$ , the median  $\mathbf{m}X$  deviates from the mean  $\mathbf{E}X$  by at most  $\delta X$ , i.e.  $\mathbf{m}X \in [\mathbf{E}X - \delta X, \mathbf{E}X + \delta X]$ .*

*Remark.* This also establishes the (weaker) fact that for any random variable  $X$ , the median  $\mathbf{m}X$  always lies in the interval  $[\mathbf{E}X - \sigma X, \mathbf{E}X + \sigma X]$ , which is mentioned in the literature [6], but, according to a small survey of ours, seems to be little known among theoretical computer scientists. When the distribution of  $X$  is unimodal, the difference between the mean and the median can even be bounded by  $\sqrt{3/5} \cdot \sigma$  [7]. By what is shown below, we may in that case replace  $\delta$  by  $\sqrt{3/5} \cdot \sigma$  in Theorem 1.

*Proof.* Fact 3 is an immediate consequence of Lemma 1 by noting that (for continuous random variables)  $\Pr(X \leq \mathbf{m}X) = \Pr(X \geq \mathbf{m}X) = 1/2$  and taking  $a = \delta X$ . Alternatively, we could mimic the proof of that lemma.  $\square$

These two simple facts are the heart and soul underlying Theorem 1 in the sense that the two inequalities of Lemma 2 now have the following very short and elegant alternative proofs:

$$\begin{aligned} \mathbf{E}f(X) - \delta f(X) &\leq \mathbf{m}f(X) = f(\mathbf{m}X) \leq f(\mathbf{E}X + \delta X) \\ \mathbf{E}f(X) + \delta f(X) &\geq \mathbf{m}f(X) = f(\mathbf{m}X) \geq f(\mathbf{E}X - \delta X) \end{aligned}$$

where the inequalities follow from Fact 3 and the monotonicity of  $f$ , and the equalities are just restatements of Fact 2.

Given Theorem 1 and Fact 2, the question arises whether not the median should generally be preferred over the mean when looking for an “average” value?

One strong argument that speaks against the median is the following. By the (weak) law of large numbers, the average over a large number of independent trials will be close to the mean, not to the median. In fact, by exactly the kind of considerations given in our introduction, the order of the medians could be the opposite of the order of the averages, which would be deceptive when in practice there were indeed a large number of independent runs of the algorithm.

A pragmatic argument is that the mean can be computed much easier: the values to be averaged over can simply be summed up without a need to keep them in memory. For the median, it is known that such a memoryless computation does not exist [8]; even approximations have to use a non-constant number of intermediate variables, and the respective algorithms are far from being simple [9].

## 4 Relaxations of the Main Theorem

In this section, we show that the result from the previous section cannot be relaxed in any obvious way, as stated in Theorem 1.

We try to find examples which are realistic in the sense that the  $f$  is well-behaved and the distributions are simple. We do so to emphasize that all conditions are also of practical relevance. First, observe that if the function  $f$  is strictly increasing it also has a strictly increasing inverse function  $f^{-1}$ . Replacing  $f(X) \rightarrow U$  and  $f(Y) \rightarrow V$ , where  $U$  and  $V$  are also random variables, we immediately see that we have halved the number of cases to consider. If we can find a counterexample for the case when  $\delta X$  is dropped we have also found one for the case when  $\delta f(Y)$  is dropped. The same symmetry relates  $\delta Y$  to  $\delta f(X)$ .

To prove that the  $\delta X$  (and hence the  $\delta f(y)$ ) cannot be removed from the statement of the theorem we consider an example with  $Y$  constant. Then we find an example of a distribution for  $X$  and a strictly increasing function  $f$  such that

$$\mathbf{E}X < Y \quad \text{and} \\ \mathbf{E}f(X) - \delta f(X) > f(Y).$$

The obvious thing works: We let  $X$  have a two-point distribution with points  $x_1 < Y$  and  $x_2 > Y$  and consider a function which is convex, e.g.  $f(x) = e^x$ . For this setting we try to solve the system

$$\begin{aligned} p_1 x_1 + p_2 x_2 &< Y \\ p_1 f(x_1) + p_2 f(x_2) - 2 p_1 p_2 (f(x_2) - f(x_1)) &< f(Y) \end{aligned} \quad (1)$$

It becomes slightly easier to spot solutions to this if we write  $p_1 = \frac{1}{2} - \delta$  and  $p_2 = \frac{1}{2} + \delta$ . Then (1) becomes

$$2 p_1 f(x_1) (1 + \delta) + 2 p_2 f(x_2) \delta < f(Y) \quad (2)$$

Thus as long as  $\delta > 0$  and  $f$  increases ‘fast enough’ in the region between  $Y$  and  $x_2$  we can always construct a simple counter-example as  $f(x_2) \gg f(Y)$ . E.g. take  $Y = 2$ ,  $p_1 = \frac{1}{4}$ ,  $p_2 = \frac{3}{4}$ ,  $x_1 = -2$ ,  $x_2 = 3$ . Similarly, we can find a two point counter-example for the case without the  $\delta Y$  by considering a logarithmic function. One such example consists of a constant  $X = 1$ ,  $p_1 = \frac{3}{4}$ ,  $p_2 = \frac{1}{4}$ ,  $y_1 = .5$ ,  $y_2 = 3$  and  $f(x) = \log(x)$ .

If we restrict ourselves, as we have done, to the case where only one of  $X$  and  $Y$  is random we see from Jensen’s inequality that we indeed must consider examples with the curvatures as chosen above. Otherwise, it would be impossible to find a counter-example.

The same examples still work if we allow  $Y$  to have a small degree of variation.

## 5 Conclusions

Theorem 1 ensures that when conclusions are drawn *only* when the error bands do not intersect, there will at least never be contradictions from the angle of different measurement scales. The bad news is that, even when the error bands do not intersect in one scale, in general nothing can be inferred about the order of the averages after a monotone transformation.

Obviously, when two sets of measurements are completely separated in the sense that the largest measurement of one set is smaller than the smallest measurement of the other set, then no monotone transformation can reverse the order of the averages. Beyond that, however, there does not seem to be any less restrictive natural precondition, which most datasets would fulfill and under which average reversal provably cannot occur.

What *can* be proven is that for two random variables  $X$  and  $Y$ , if  $0 \leq \mathbf{E}(X - \mathbf{E}X)^k \leq \mathbf{E}(Y - \mathbf{E}Y)^k$  for all  $k \in \mathbb{N}$ , then for a monotonously increasing function  $f$ , with all derivatives also monotone (as is the case for any monomial  $x \mapsto x^k$  with  $k \in \mathbb{N}$ , and for any exponential  $x \mapsto b^x$  with  $b > 1$ ), indeed  $\mathbf{E}(X) \leq \mathbf{E}(Y) \Rightarrow \mathbf{E}f(X) \leq \mathbf{E}f(Y)$ . Unfortunately, this precondition is neither practical to check nor typically fulfilled: even when restricting to classes of distributions with only two parameters (mean and standard deviation), it is not hard to come up with two (realistic) such distributions where one has smaller mean *and* variance than the other, but still the relative order of the means changes by a monotone transformation.

The bottom line of our findings could therefore be put as follows: if the cost measure is perfectly unique, e.g., we measure running time and what is of interest

in the application is nothing but this very running time, then average comparison is fine (as long, of course, as the standard precautions of considering error bars are taken; but see, for example, <http://www.graphpad.com/articles/errorbars.htm>). In all other cases there is no alternative but to explicitly provide the comparison in every cost measure that is of interest.

## References

1. Manning, C.D., Schütze, H.: Foundations of statistical natural language processing. MIT Press (1999)
2. Teh, Y.W., Jordan, M.I., Beal, M.J., Blei, D.M.: Sharing clusters among related groups: Hierarchical dirichlet processes. In: Proceedings of the Advances in Neural Information Processings Systems Conference (NIPS'04), MIT Press (2004)
3. Lavrenko, V., Croft, W.B.: Relevance based language models. In: Proceedings of the 24th annual international ACM SIGIR conference on Research and development in information retrieval (SIGIR'01), ACM Press (2001) 120–127
4. Mori, S., Nagao, M.: A stochastic language model using dependency and its improvement by word clustering. In: Proceedings of the 17th international conference on Computational linguistics (COLING'98), Association for Computational Linguistics (1998) 898–904
5. Grimmett, G., Stirzaker, D.: Probability and Random Processes. Oxford University Press (1992)
6. Siegel, A.: Median bounds and their application. *Journal of Algorithms* **38** (2001) 184–236
7. Basu, S., Dasgupta, A.: The mean, median and mode of unimodal distributions: A characterization. *Theory of Probability and its Applications* **41** (1997) 210–223
8. Munro, J.I., Paterson, M.S.: Selection and sorting with limited storage. *Theoretical Computer Science* **12** (1980) 315–323
9. Manku, G.S., Rajagopalan, S., Lindsay, B.G.: Approximate medians and other quantiles in one pass and with limited memory. In: Proceedings of the ACM SIGMOD International Conference on Management of Data (SIGMOD'98). (1998) 426–435

## A The Example from the Introduction

Each point in the figures in the introduction was computed as the average over points distributed as  $Z = z_0 + \text{Exp}(\lambda)$ , where  $\text{Exp}(\lambda)$  denotes the exponential distribution with mean  $1/\lambda$  (density is  $\varphi(t) = \lambda e^{-\lambda t}$ ; variance is  $1/\lambda^2$ ).

For the mean of  $2^Z$ , or more generally,  $e^{\kappa Z}$ , we have that

$$\begin{aligned} \mathbf{E}e^{\kappa Z} &= \int_0^\infty e^{\kappa(z_0+t)} \lambda e^{-\lambda t} dt \\ &= e^{\kappa z_0} \cdot \lambda/(\lambda - \kappa) \\ &\approx e^{\kappa(z_0 + 1/(\lambda - \kappa))} \\ &= e^{\kappa(z_0 + 1/\lambda + \kappa/(\lambda(\lambda - \kappa)))}. \end{aligned}$$

For the figures in the introduction, we chose  $z_0$  so that the means for each curve would lie on a slightly perturbed line. For the green (light gray) curve, we chose  $\lambda = 1$ , for the blue (dark gray) curve we chose  $\lambda = 2$ . For example for  $X = 3 + \text{Exp}(1)$  and  $Y = 5 + \text{Exp}(2)$ , we then have

$$\begin{aligned} \mathbf{E}X &= 3 + 1/1 = 4 \\ \mathbf{E}Y &= 5 + 1/2 = 5.5, \end{aligned}$$

and for  $\kappa = 3/4$  (then  $e^\kappa \approx 2$ ),

$$\begin{aligned} \mathbf{E}e^{\kappa X} &\approx e^{\kappa(3 + 1.0 + 3.0)} \approx 27.0 \\ \mathbf{E}e^{\kappa Y} &\approx e^{\kappa(4 + 0.5 + 0.3)} \approx 25.8. \end{aligned}$$

Observe that in this setting we need  $\kappa < \lambda_1$  and  $\kappa < \lambda_2$  to ensure that both  $\mathbf{E}e^{\kappa X}$  and  $\mathbf{E}e^{\kappa Y}$  exist.

One objection against the exponential distribution might be that its exponentiation is too heavy-tailed in the sense that not all its moments exist. However, the same calculations as above can also be carried out for two, say, normal distributions, which are free from this taint. Let  $Z = N(z_0, \sigma)$ , that is,  $Z$  has a normal distribution with mean  $z_0$  and standard deviation  $\sigma$ . A straightforward calculation shows that the mean of  $e^{\kappa Z}$ , which obeys a lognormal distribution, is given by

$$\begin{aligned} \mathbf{E}e^{\kappa Z} &= \int_{-\infty}^\infty e^{\kappa(z_0+t)} \frac{1}{\sqrt{2\pi\sigma^2}} e^{-(t-z_0)/(2\sigma^2)} dt \\ &= e^{\kappa z_0 + \kappa^2 \sigma^2 / 2}. \end{aligned}$$

For example, taking  $X = N(4, 1.5)$  and  $Y = N(4.5, 1.0)$  and  $\kappa = 1$ , the order of the means then changes:

$$\begin{aligned} \mathbf{E}e^{\kappa X} &= e^{\kappa 4 + \kappa^2 1.5^2 / 2} = e^{5.125} \\ \mathbf{E}e^{\kappa Y} &= e^{\kappa 4.5 + \kappa^2 / 2} = e^5. \end{aligned}$$

## B One-sided Chebyshev Bounds

For the sake of completeness, we state the one-sided version of Chebyshev's inequality, which looks similar to Lemma 1 in Sect. 2. As mentioned in that section, Lemma 1 is stronger for deviation up to at least  $\sigma X$ , while the lemma below is stronger for large deviations.

**Lemma 3.** *For any random variable  $X$  and for every  $a > 0$ , it holds that*

$$(a) \Pr(X \geq \mathbf{E}X + a) \leq \frac{(\sigma X)^2}{a^2 + (\sigma X)^2};$$

$$(b) \Pr(X \leq \mathbf{E}X - a) \leq \frac{(\sigma X)^2}{a^2 + (\sigma X)^2}.$$

*Proof.* See e.g. [5]. The main idea is to write  $\Pr(X \geq \mathbf{E}X + a) = \Pr((X - \mathbf{E}X + c)^2 \geq (a + c)^2)$ , then apply Markov's inequality and determine that  $c$  which gives the best bound; similarly for (b).  $\square$