MDPI

Article

Limit Theorems as Blessing of Dimensionality: Neural-Oriented Overview

Vladik Kreinovich and Olga Kosheleva

University of Texas at El Paso, El Paso TX 79968, USA; vladik@utep.edu, olgak@utep.edu
* Correspondence: vladik@utep.edu (V.K.)

Abstract: As a system becomes more complex, at first, its description and analysis becomes more complicated. However, a further increase in the system's complexity often makes this analysis simpler. A classical example is Central Limit Theorem: when we have a few independent sources of uncertainty, the resulting uncertainty is very difficult to describe, but as the number of such sources increases, the resulting distribution get close to an easy-to-analyze normal one – and indeed, normal distributions are ubiquitous. We show that such limit theorems make analysis of complex systems easier – i.e., lead to blessing of dimensionality phenomenon – for all the aspects of these systems: the corresponding transformation, the system's uncertainty, and the desired result of the system's analysis.

Keywords: limit theorems; curse and blessing of dimensionality; neural networks

1. Introduction: From Curse of Dimensionality to Blessing of Dimensionality

First, a curse. As a system becomes more complex, at first, its description and analysis becomes more complicated.

Then, a blessing. However, a further increase in the system's complexity often makes this analysis simpler.

Example. A classical example of this first-curse-then-blessing phenomenon is the joint effect of many random phenomena. When we know the probability distribution of each phenomenon, in principle, we can compute their joint effect – but, as the number of these phenomena becomes larger and larger, the corresponding computations become more and more complicated. At first glance, this is a classical example of the curse of dimensionality.

However, as the number of these phenomena increases further, we start seeing the effect of the Central Limit Theorem (see, e.g., [26]), according to which, under reasonable conditions, the joint effect of many small independent random phenomena is close to Gaussian. The resulting distribution becomes very close to the easy-to-analyze Gaussian distribution – and this is one of the main reasons why normal distributions are ubiquitous.

Is this a lucky example or a general trend? At first glance, it may appear that the Central Limit Theorem is a lucky break in the dark world of curse-of-dimensionality phenomena.

It is a general trend: what we show in this paper. In this paper, we show that the above pessimistic viewpoint is – well – unnecessarily pessimistic. Actually, as we will show, similar limit theorems are ubiquitous – and their use can (and do) help in data processing in general and, in particular, in neural data processing. These limit theorems are not only helpful – they explain a surprising empirical success of many techniques, from traditional neural networks to convex techniques and clustering.

Citation: Kreinovich, V.; Kosheleva, O. Limit Theorems as Blessing of Dimensionality: Neural-Oriented Overview. *Entropy* **2021**, *1*, 0. https://doi.org/

Received:
Accepted:
Published:

2. Two Main Sources of Dimensionality: Spatial and Temporal

To provide an adequate analysis of the situation, let us first observe that in general, there are two main sources of dimensionality:

- First, at each moment of time, there is usually a large number of phenomena – located, in general, at different points in space – that need to be taken into account. Even if we use a few parameters to describe each of these phenomena, overall, we will need a very large number of parameters to describe all these phenomena – and thus, the dimensionality of the problems grows. We will call this dimensionality of *spatial origin*, or simply spatial dimensionality, for short. The above-mentioned Central Limit Theorem is a good example of spatial dimensionality.
- Also, there may be parameters describing the history of this phenomenon – which also affect its current state. For example, in a sensor, the original signal may be transformed many times, and what we see is the result of all these past transformations. We will call this dimensionality of *temporal origin*, or simply *temporal* dimensionality.

In this paper, we will study the limit theorems related to both sources of dimensionality.

Comment. Limit theorems are often somewhat complicated to understand and prove. In our experience, a better understanding of a complex multi-dimensional phenomena is usually achieved if we consider easier-to-analyze few-dimensional particular cases or analogues. For limit theorems, a natural few-dimensional analogues are iterative methods in numerical mathematics, such as:

- the Newton's iterative method

$$x^{(k+1)} = x^{(k)} - \frac{f\left(x^{(k)}\right)}{f'\left(x^{(k)}\right)}$$

  for finding the solution to the equation $f(x) = 0$ or
- since we are interested in neural network applications – the gradient method

$$x_i^{(k+1)} = x_i^{(k)} - \alpha \cdot \frac{\partial f}{\partial x_i}_{|x=x^{(k)}}$$

  for finding the minimum of a function $f(x)$.

In both examples, convergence is not guaranteed, and the results explaining when there is convergence are often difficult to prove. However, what is much easier to prove is that *if* there is a convergence, then the limit satisfies the desired property – or at least some part of this desired property. Indeed:

- For the Newton's method, if $x^{(k)} \to x$, then, in the limit, we get $x = x - \frac{f(x)}{f'(x)}$, which implies that $f(x) = 0$.
- For the gradient descent, if $x^{(k)} \to x$, then, in the limit, we get $x_i = x_i - \alpha \cdot \frac{\partial f}{\partial x_i}$, which implies that $\frac{\partial f}{\partial x_i} = 0$. Thus, the limit point is always a stationary point, which is a necessary (but, as is well known, not sufficient) condition for it being the location of the minimum.

Similarly to these cases, in this paper, we will concentrate not so much on the conditions under which the processes converge, but rather on the description of the limit cases *when* there is convergence.

3. Dimensionality of Spatial Origin

As we have mentioned, the standard Central Limit Theorem is an example of what we called dimensionality of spatial origin. While many consequences of this theorem are well known, as we will show, there are many aspects of this theorem which still need

exploring. So, the first thing we will consider – in the first subsection of this section – is what are the less known consequences of the Central Limit Theorem.

Of course, the limit distribution does not have to be normal: as we have mentioned, the convergence to the normal distribution happens only under certain conditions. For situations when these conditions are not satisfied, there are more general limit theorems. Applications of these more general theorems – mostly to uncertainty quantification – is what we will overview in the second subsection of this section.

All this assumes that we know the probability distributions that we are trying to combine. But what if we do not know the probabilities, what if we only know the corresponding range of possible values – and we do not know the probabilities of different points from this range? This situation is discussed in the third subsection of this section.

This section ends with related open questions.

3.1. Not-Well-Known Consequences of the Central Limit Theorem

Why are many things in the words discrete? Outside quantum physics, most physical processes are continuous, most probability distributions are continuous – so what we should observe should be continuous as well. However, in reality, many things in the real world are discrete. We do not have weather continuously changing from sunny to rain: most of the time, we either have a sunny day or a rainy day. Yes, it is possible to have hybrid animals like mules, but most of the time, animals we see fall into one of the precise categories.

In general, when we use a neural network (or any similar tool) for classification, what this network actually produces are continuous numbers – that can be converted, e.g., to degrees to which the object belongs to different categories. However, usually, we do not return these degrees to the user. What usually do at the end is selecting one of these categories (e.g., the most probable one) – and in most cases, this is exactly the desired classification, cat or dog, car or not-a-car, disease or healthy.

This discreteness definitely helps when making decisions – instead of a continuum of possible values, we need to deal with only a few discrete ones. So, this discreteness can be viewed as an example of a blessing of dimensionality.

But why are we mentioning this discreteness? At first glance, it may seem to be unrelated to the Central Limit Theorem – which is all about the normal distribution, which is, of course, absolutely continuous. Interestingly, there is a relation. Let us describe it.

This puzzling discreteness has been observed before. Of course, we are not the first ones who noticed that, in spite of the the fact that many processes are continuous, what we observe is often discrete. For example, B. S. Tsirelson noticed in [28] that in many cases, when we reconstruct the signal from the noisy data, and we assume that the resulting signal belongs to a certain class, the reconstructed signal is often an *extreme* point from this class – i.e., is one of the discrete extreme points. For example:

- when we assume that the reconstructed signal is monotonic, the reconstructed function is often (piece-wise) constant;
- if we additional assume that the signal is one time differentiable, the result is usually one time differentiable but rarely twice differentiable, etc.

Tsirelson's explanation. Out of many papers that mention the puzzling discreteness, we cited the paper [28] – because this paper not only *mentions* the fact of discreteness, it also provides an *explanation* for this discreteness, and this explanation is closely related to the Central Limit Theorem (see also [20]).

Indeed, when we reconstruct a signal from a mixture of the signal and a Gaussian noise, then the *maximum likelihood* estimation (a traditional statistical technique; see, e.g., [26]) means that our of all possible signals from the given class of signals, we look for

the signal which is the closest (in the least squares — i.e., in effect, Euclidean – metric) to the observed "signal + noise" combination.

In particular, if the signal is determined by finitely many (say, d) parameters, we must look for a signal $\vec{s} = (s_1, \ldots, s_d)$ from the a priori set $A \subseteq R^d$ that is the closest (in the usual Euclidean sense) to the observed values

$$\vec{o} = (o_1, \ldots, o_d) = (s_1 + n_1, \ldots, s_d + n_d),$$

where n_i denotes the (unknown) values of the noise.

Since the noise is Gaussian, we can conclude that the average value of $(n_i)^2$ is close to σ^2, where σ is the standard deviation of the noise. In other words, we can conclude that

$$(n_1)^2 + \ldots + (n_d)^2 \approx d \cdot \sigma^2.$$

In geometric terms, this means that the distance

$$\sqrt{\sum_{i=1}^{d} (o_i - s_i)^2} = \sqrt{\sum_{i=1}^{d} n_i^2}$$

between $\vec{s}$ and $\vec{o}$ is $\approx \sigma \cdot \sqrt{d}$. Let us denote this distance $\sigma \cdot \sqrt{d}$ by ε.

For simplicity of explanation, let us consider the case when $d = 2$, and when A is a convex polygon. Then, we can divide all points p from the exterior of A that are ε-close to A into several zones depending on what part of A is the closest to p: one of the *sides*, or one of the *edges*.

Geometrically, the set of all points for which the closest point $a \in A$ belongs to the *side* e is bounded by the straight lines orthogonal (perpendicular) to e. The total length of this set is is therefore equal to the length of this particular side; hence, the total length of all the points that are the closest to all the sides is equal to the *perimeter* of the polygon. This total length thus does not depend on ε at all.

On the other hand, the set of all the points at the distance ε from A grows with the increase in ε; its length grows approximately as the length of a circle, i.e., as const·ε.

When ε increases, the (constant) perimeter is a vanishing part of the total length. Hence, for large ε:

- the fraction of the points that are the closest to one of the sides tends to 0, while
- the fraction of the points p for which the *closest* is one of the *edges* tends to 1.

Thus, with high probability, the reconstructed signal corresponds to one of the edges (extreme points) of the set A.

Similar arguments can be repeated for any dimension. For the same noise level σ, when d increases, the distance $\varepsilon = \sigma \cdot \sqrt{d}$ also increases, and therefore, for large d, for "almost all" observed points $\vec{o}$, the reconstructed signal is one of the extreme points of the a priori set A.

Much less probable is that the reconstructed signal belongs to the 1-dimensional face of the set A, even much less probable that s belongs to a 2-D face, etc.

Methodological consequence. So, when the dimension increases, we have a clear example of blessing of dimensionality: instead of having to consider a continuum of possible states, we only have to deal with a much smaller discrete set of extreme points – edges of the corresponding polyhedron.

So, all observed phenomena falls into a few clusters – exactly as we observe in many cases.

Comment. This idea helps even in the quantum case. Namely, in quantum physics, there is a known paradox formulated by Schroedinger himself (the author of the main equation of quantum physics): while in quantum physics, we can have a superposition of any two states, how come we never see a superposition of two macro-states, e.g., of

the state in which a cat is live and the state in which the same cat is dead? This is indeed a serious problem, it was one of the reasons why Einstein did not believe that quantum physics is an adequate description of reality.

The above idea explains this paradox – with very high probability, we will observe one of the two original states and not their convex combination (i.e., in this case, not their superposition).

3.2. Uncertainty Quantification and Probabilistic Limit Theorems – Including Theorems Beyond Normal Distributions

Need for uncertainty quantification. Whether we use neural networks of whether we use other algorithms for data processing, the inputs to all these algorithms are real numbers. These real numbers mostly come from measurements, and measurements are never absolutely accurate; see, e.g., [23]. There is always noise. As a result, the measurement results $\widetilde{x}_i$ are, in general, somewhat different from the actual (unknown) values x_i of the corresponding quantities, and the difference $\Delta x_i \stackrel{\text{def}}{=} \widetilde{x}_i - x_i$ – known as *measurement error* – is, in general, different from 0. So, when we apply the data processing algorithm f to the measurement result, the algorithm's output $\widetilde{y} = f(\widetilde{x}_1, \ldots, \widetilde{x}_n)$ is, in general, different from the value $y = f(x_1, \ldots, x_n)$ that we would have obtained if we knew the actual values x_i.

In practice, it is important to know how close is our estimate $\widetilde{y}$ to the desired value y, i.e., in other words, how big can the difference $\Delta y \stackrel{\text{def}}{=} \widetilde{y} - y$ be. For example, suppose that we are prospecting for oil, and our estimate $\widetilde{y}$ for the amount of oil y in the given region is 150 million ton. Then, if the accuracy is 10 million tons, this estimate is good news, and we can start exploiting this region. On the other hand, if it is 150 ± 200, then maybe there is no oil at all, so before we invest a lot of money into digging deep wells, we better perform more measurements to make sure that this money will not be wasted.

Estimating Δy is one the most important aspects of *uncertainty quantification*.

Possibility of linearlization. We are interested in estimating the quantity

$$\Delta y = f(\widetilde{x}_1, \ldots, \widetilde{x}_n) - f(x_1, \ldots, x_n) = f(\widetilde{x}_1, \ldots, \widetilde{x}_n) - f(\widetilde{x}_1 - \Delta x_1, \ldots, \widetilde{x}_n - \Delta x_n).$$

Measurements are usually reasonable accurate, so the measurement errors Δx_i are relatively small. For small values, their squares are much smaller than the values themselves – and can therefore be usually safely ignored. For example, if $\Delta x_i \approx 10\%$, then $(\Delta x_i)^2 \approx 1\% \ll \Delta x_i$. Thus, a reasonable idea is to expand the above expression for Δy in Taylor series and ignore terms which are quadratic (or of higher order) in terms of the measurement errors Δx_i. As a result, we get a linear dependence:

$$\Delta y = \sum_{i=1}^{n} c_i \cdot \Delta x_i, \text{ where } c_i \stackrel{\text{def}}{=} \frac{\partial f}{\partial x_i}.$$

Comment. This linearization – replacing the generic dependence with a linear one – is a usual idea in applications. Actually, it one of the main ideas in many applications; see, e.g., [6].

Here, the Central Limit Theorem can help. Let us first consider an important case when we know the probability distribution of each measurement error Δx_i. Usually, each measuring instrument is *calibrated* – if it has a *bias*, i.e., if the mean value $E[\Delta x_i]$ of the measurement error is not 0, we simply subtract this mean value from all the measurement results and thus, reduce it to 0.

In many practical applications, the number n of inputs is large, and the role of each of these inputs is relatively small. For example, one of the important data when prospecting for oil is seismograms – several-times-a-second recordings of the seismic signal. There are thousands of the corresponding values, and the effect of each indi-

vidual value of the result of data processing is indeed small. The measurement errors corresponding to different measurements are usually reasonably independent. Thus, we are under the condition of the Central Limit Theorem – so we can conclude that the desired estimation error Δy is normally distributed.

A normal distribution is uniquely determined by its mean μ and its standard deviation σ. When each measurement error Δx_i has mean value 0, the mean value of their linear combination Δy is also 0, and the variance σ of this linear combination can be determined from the known fact that the variance of the sum of independent random variables is equal to the sum of variances:

$$\sigma^2 = \sum_{i=1}^{n} c_i^2 \cdot \sigma_i^2.$$

How can actually estimate σ? In principle, we can directly use the above formula to estimate the standard deviation σ of the approximation error Δy. The main computational difficulty is that the data processing algorithm f is usually very complicated (especially in case of neural networks), so it is not possible to compute the partial derivatives analytically. We can, however, use the fact that a partial derivative is defined as the limit of the ratios

$$\frac{\partial f}{\partial x_i} = \lim_{h \to 0} \frac{f(\widetilde{x}_1, \ldots, \widetilde{x}_{i-1}, \widetilde{x}_i + h, \widetilde{x}_{i+1}, \ldots, \widetilde{x}_n) - \widetilde{y}}{h},$$

and thus, for a sufficiently small h, the value of the ratio is very close to the desired partial derivative. Thus, we can estimate c_i as

$$c_i \approx \frac{f(\widetilde{x}_1, \ldots, \widetilde{x}_{i-1}, \widetilde{x}_i + h, \widetilde{x}_{i+1}, \ldots, \widetilde{x}_n) - \widetilde{y}}{h}.$$

The problem with this idea is that it takes too long. Indeed, if we have several thousand inputs, then, to compute all the corresponding values c_i, we need to call the data processing algorithm f (which often takes hours to compute) $n + 1$ times: one time to compute $\widetilde{y}$ and n time to compute the corresponding n ratios c_i. For several thousand inputs, this is not realistic.

Good news is that we can instead use Monte-Carlo techniques: instead of computing n partial derivatives, we can simply emulate, certain number of times K, measurement errors $\delta x_i^{(k)}$ which are normally distributed with standard deviation σ_i, and compute the differences

$$\delta y^{(k)} = \widetilde{y} - f(\widetilde{x}_1 - \delta x_1^{(k)}, \ldots, \widetilde{x}_n - \delta x_n^{(k)}).$$

By the same logic as before, the differences $\delta y^{(k)}$ are normally distributed with the desired standard deviation σ. Thus, from a sample of K values, we can estimate σ with accuracy $\approx 1/\sqrt{K}$ [26]. So, if we want to estimate σ with relative accuracy $1/\sqrt{K} \approx 20\%$, it is sufficient to call the algorithm f $K = 25$ times – which is much much smaller than thousands needed for exact estimation.

So what? Why are we spending so much time on the ideas that are well known to many readers? Because this will prepare readers to something that – unfortunately – not too many readers now: that we can use limit theorems beyond normal distributions to cover other realistic cases of uncertainty quantification.

Interval uncertainty. In the previous text, we assumed that for each measurement, we know the probability distribution of the corresponding measurement error. Sometimes we do know this distribution, but often, this distribution is not known. Indeed, nowadays, many sensors are cheap, but determining the probability distribution for each sensor would cost too much and is thus often not done. Instead, we have to rely on the information provided by the manufacturers of the corresponding measuring instruments, and this information often consists of simply providing an upper bound Δ_i for

the absolute value of the corresponding measurement error; see, e.g., [23]. (At least such an upper bound needs to be provided – otherwise, it is not a measuring instrument, it is a wild guess.)

Once we know the upper bound Δ_i on the absolute value $|\Delta x_i| = |\widetilde{x}_i - x_i|$ of the measurement error, then, based on the measurement result $\widetilde{x}_i$, the only information we gain about the actual (unknown) value x_i of the corresponding quantity is that this value belongs to the interval $[\underline{x}_i, \overline{x}_i] \stackrel{\text{def}}{=} [\widetilde{x}_i - \Delta_i, \widetilde{x}_i + \Delta_i]$. Because of this fact, such a situation is known as *interval uncertainty*.

Under such interval uncertainty, the only thing we can conclude about the value $y = f(x_1, \ldots, x_n)$ is that it belongs to the *range* $[\underline{y}, \overline{y}]$ of possible values of the function f when x_i are in the corresponding intervals:

$$[\underline{y}, \overline{y}] = \{f(x_1, \ldots, x_n) : x_i \in [\underline{x}_i, \overline{x}_i] \text{ for all } i\}.$$

The problem of computing this interval is known as the problem of *interval computation*; see, e.g., [17,18].

In general, this problem is NP-hard [14] – which means that, unless P = NP (which most computer scientists do not believe to be possible), no feasible algorithm is possible for solving all particular cases of this problems. However, in the linearized case, a feasible algorithm *is* possible. Indeed, since the expression $\sum\limits_i c_i \cdot \Delta x_i$ is linear (thus monotonic) in the variables Δx_i, its largest value is attained:

- for $c_i > 0$, when the value Δx_i is the largest, i.e., when $\Delta x_i = \Delta_i$, and
- for $c_i < 0$, when the value Δx_i is the smallest, i.e., when $\Delta x_i = -\Delta_i$.

Thus, the largest possible value Δ of Δy is equal to

$$\Delta = \sum_{i=1}^{n} |c_i| \cdot \Delta_i.$$

Similarly, one can easily show that the smallest possible value of Δy is equal to $-\Delta$.

How to estimate uncertainty in the interval case. How can we compute this sum Δ? We can directly use this formula – i.e., use numerical differentiation to compute all the partial derivatives c_i and then compute the sum. However, as we have mentioned earlier, in many practical situations, this approach is not realistic. What can we do?

Another limit distribution comes to the rescue. As we have mentioned, the convergence to a normal distribution only happens under certain conditions. In other cases, we may have convergence to other so-called *infinitely divisible* distributions [26]. One of such distributions is the *Cauchy distribution*, in which the probability density $\rho(x)$ has the following form:

$$\rho(x) = \text{const} \cdot \frac{1}{1 + \left(\frac{x}{\Delta}\right)^2},$$

for some parameter Δ.

An important feature of the Cauchy distribution is that if we have several independent Cauchy distributed random variables r_i with parameters Δ_i, then their linear combination $\sum\limits_i c_i \cdot r_i$ is also Cauchy distributed, with parameter $\Delta = \sum\limits_i |c_i| \cdot \Delta_i$ – which is exactly the value that we want to compute. This feature leads to the following Monte-Carlo method for computing Δ: we emulate, certain number of times K, measurement errors $\delta x_i^{(k)}$ which are Cauchy distributed with paremeters Δ_i, and compute the differences

$$\delta y^{(k)} = \widetilde{y} - f(\widetilde{x}_1 - \delta x_1^{(k)}, \ldots, \widetilde{x}_n - \delta x_n^{(k)}).$$

Then, due to the above feature, the differences $\delta y^{(k)}$ are Cauchy distributed with the desired parameter Δ. Thus, to a sample of K values, we can apply, e.g., the maximum

likelihood method [26], and thus estimate Δ with accuracy $\approx 1/\sqrt{K}$. Similarly to the case of normal distributions, this drastically speeds up computations: if we want to estimate Δ with relative accuracy 20%, it is sufficient to call the algorithm f 25 times – which is much much smaller than thousands needed for exact estimation.

This method has been successfully used in many applications; see, e.g., [13].

Comment. Note that, in contrast to many simulation techniques, the use of Cauchy distribution in interval-related uncertainty quantification is *not* a realistic simulation: the actual measurement error is always located inside the interval $[-\Delta, \Delta]$, while the Cauchy-distributed random variable has a non-zero probability to be anywhere, in particular, outside the interval. In other words, this is one of the cases when the blessing of dimensionality comes not from the direct use of the corresponding limit theorem, but from its indirect use.

3.3. What If We Have No Information about Probabilities

Formulation of the problem. What if we know that the disturbance $x = (x_1, \ldots, x_n)$ is a joint effect of several very independent small ones: $x = x^{(1)} + \ldots + x^{(N)}$, where about each component $x^{(i)}$, we only know the set $X^{(i)}$ of its possible values – and we do not have any information about probabilities of different points within each set. The only constraint is that all the points from each set $X^{(i)}$ are small, i.e., that for some small values $\varepsilon > 0$, the length $\left\|x^{(i)}\right\|$ of each vector $x^{(i))} \in X^{(i)}$ does not exceed ε. We will call such sets *ε-small*.

In this case, the set X of all possible values of the sum x is the set of all possible sums $x^{(1)} + \ldots + x^{(N)}$, where $x^{(i)} \in X^{(i)}$ for all i. In mathematics, the set of all such sums is known as the *Minkowski sum* of the sets $X^{(i)}$. The Minkowski sum is usually denoted by $X^{(1)} + \ldots + X^{(N)}$.

What can we say about such set X?

1-D case. The 1-D case $n = 1$ was studied in [11]. This paper showed that if a set X is the Minkowski sum of several ε-small closed sets, then it is ε-close to some interval $I = [a, b]$, i.e.:

- every point from the set X is ε-close to some point from the interval I, and
- every point from the interval I is ε-close to some point from the set X.

In the limit $\varepsilon \to 0$, we conclude that the Minkowski sum tends to the interval.

Comment. This limit theorem is similar, in formulation, to the Central Limit Theorem and its generalizations: it shows that if a quantity can be represented as the sum of many small components, then the set of all possible values of this quantity is close to an interval – and the smaller the components, the closer is the resulting set to an interval.

Similarly to the fact that the original Central Limit Theorem explains the real-life ubiquity of normal distributions, this limit theorem explains the ubiquity of interval uncertainty; see, e.g., [17,18,23].

General case. It is well known that every convex set X containing 0 can be represented, for every $\varepsilon > 0$, as a Minkowski sum of ε-small sets: indeed, it is sufficient to take $X^{(i)} = N^{-1} \cdot X$ for a sufficiently large N, then:

- the inclusion $X \subseteq X^{(1)} + \ldots + X^{(N)}$ follows from the fact that each element x can be represented as the sum $x = N^{-1} \cdot x + \ldots + N^{-1} \cdot x$; and
- the opposite inclusion $X^{(1)} + \ldots + X^{(N)} \subseteq X$ follows from the fact that the set X is convex and thus, once the elements $x^{(1)}, \ldots, x^{(N)}$ belong to this set, their convex combination $N^{-1} \cdot x^{(1)} + \ldots + N^{-1} \cdot x^{(N)}$ also belongs to X.

Whether the opposite is true – i.e., whether only convex sets can be represented as sums of small sets – remained an open problem. This problem – first formulated in [11] – was resolved in [25], where it was shown that indeed, if a set X can be represented, for each ε, as the Minkowski sum of ε-small closed sets, then this set X is convex.

To be more precise, this paper showed that for every $\gamma > 0$, if a set $X \subset \mathbb{R}^d$ of diameter < 1 is δ-close to the Minkowski sum of sets of diameter $\leq \varepsilon$, then X is γ-close to a convex set, for $\delta = \gamma/3$ and $\varepsilon = \gamma^2/(20d)$.

Comment. This limit theorem explains the ubiquity of convex set in real-life problems. This is very good news, since it is known that convexity makes many computational problems easier to solve; see, e.g., [21].

3.4. Important Open Questions

What is we only have partial information about probabilities? In the first two subsections of the current section, we considered the case when we know the probability distributions of the aggregated factors. In the third subsection, we considered the case when we only know the ranges, and we have no information about the probability of different values from these ranges.

These are two extreme situations – either we know everything about the probabilities, or we have no information about these probabilities at all. In practice, we often have intermediate situations, when we have *partial* information about the probabilities. It is therefore desirable to extend the limit results from both extreme cases to the such intermediate situations as well.

Possible approach and natural generalizations of the Central Limit Theorem. As we have mentioned earlier, when we know all the probabilities, then for uncertainty quantification, we can use Monte-Carlo approach with normal distributions. When we only know the upper bounds, we can use Cauchy distributions. What if for some components, we know the probabilities, and for others, we only know bounds? The resulting random variable is the sum of two partial sums, for which the first partial sum can be handled by the normal distribution, while the second partial sum can be handled by the Cauchy distribution. In this case, it seems reasonable to use the distributions corresponding to the sum of normally and Cauchy distributed random variables.

The family of such distribution is also a natural limit – the limit of sums in which the first partial sum tends to normal distribution and the second partial sum tends to the Cauchy one. Such mixed distributions are not covered by the usual limit theorems, since these limit theorems consider 2-parametric limit families of probability distributions: e.g., a normal distribution is determined by two parameters – mean and standard deviation. Sums would require more parameters: we need mean and standard deviation of the normal part and the parameter Δ of the Cauchy part.

Possible generalizations of the traditional limit theorems to such multi-parametric families have been analyzed in [29]. It turns out that, in general, in this case, the resulting distribution is equivalent to the distribution of the sum of several different infinitely divisible distributions: e.g., to the sum of normally and Cauchy distributed variables. So maybe other distributions of this time can be used for uncertainty quantification in other cases when we only have partial information about probabilities?

What if we are interested in the extreme case? Very often, we are interested in the extreme case: e.g., when we design a bridge, we want it to withstand the strongest possible winds that can happen in this area. In such situations, we are interested not in the summary effect of several random variables, but rather in the largest value of several random variables – e.g., variables describing the wind on different days. When all these variables are identically distributed, then, similarly to the Central Limit Theorem, we have a finite-parametric family of distributions that represents the distribution of such extreme events; see, e.g.. [1,3–5,9,10,22,24].

But what happens in a realistic setting when the distributions may be different? The Central Limit Theorem still works in such cases, but the Extreme Value Theorem stops working – and moreover, there are fundamental limitations to how far we can go in this direction: we cannot go as far as the Central Limit Theorem and consider all possible cases when distributions are different; see [15].

An important question is: how far *can* we go?

4. Dimensionality of Temporal Origin

Idea: reminder. In general, when a signal goes through a multi-layer sensor, it undergoes a sequence of transformations, and these transformations are, in general, nonlinear. In mathematical terms, this means that the resulting transformation $f(x)$ of input to output is a composition of several different nonlinear functions

$$f(x) = f_n(f_{n-1}(\ldots f_2(f_1(x))\ldots)).$$

If n is large, what can we say about the resulting transformation?

Let us formulate this idea in precise terms. As we have mentioned earlier, in this paper, we do not focus on *conditions* when there is a convergence, we only focus on the resulting limit. In line with this approach, let us assume that we have a finite-parametric family F of limit functions.

If we have two sequences of transformations:

- a sequence f_i whose compositions tend to some function $f \in F$ and
- a sequence g_i whose composition tends to some function $g \in F$,

then in the case when we first apply all f_i-transformations and then all g_i-transformations, then the resulting limit function $g(f(x))$ should also belong to the family F. Thus, the desired family F of all possible limit functions should be closed under composition.

Most transformations in sensors are reversible. So, if we limit ourselves to such transformations, and instead of first applying f_1, then f_2, etc., we change the direction of signal processing and first apply f_n^{-1}, then f_{n-1}^{-1}, etc., then, in the limit, instead of the original limit function f we will get the inverse function $f^{-1}(x)$. So, the class F of all possible limit function should contain, with each function f, its inverse function as well. So, the class F must be closed under composition and inverse. Such classes are known as *transformation groups*.

Also, linear transformations are ubiquitous. Thus, it make sense to consider finite-parametric groups that contain all linear transformations. What are these groups?

Enter Norbert Wiener. Interestingly, the answer to this question is related to Norbert Wiener, the father of cybernetics. As he describes in his pioneering monograph [30] on cybernetics, when he started working on engineering problems, at first, he trusted exact mathematical models much more than vague biological analogies. And then, when he came up with a draft design of a system for automatic vision, a neurophysiologist colleague – who saw the corresponding picture – asked him with surprise since when Wiener has become interested in human vision: because it turned out that what Wiener came up with after many thoughts and tries was exactly the scheme implemented in human vision. This experience lead to Wiener's idea of *cybernetics*, a science studying both engineering and biological systems, in which one of the main ideas is that since we the humans are the product of billion years of improving evolution, our biology should be close to optimal – and thus simulating this biology can be very helpful in engineering.

In some cases, this optimality was indeed confirmed. In some other cases, Wiener became so confident in the related optimality that he made several mathematical hypotheses based on this confidence. For example, he learned, from the psychologists, that when we get closer and closer to an object, there are several clearly distinct phrases in our visual perception (which, by the way, again fits with the above explanation of discreteness):

- When the object is very far, all we see is a formless blurb – in other words, objects obtained from one another by arbitrary smooth transformations cannot be distinguished.

- When the object gets closer, we can detect whether it is smooth or has sharp angles. We may see a circle as an ellipse, a square as a rhombus (diamond). At this stage, images obtained by a projective transformation are indistinguishable.
- When the object gets even closer, we can detect which lines are parallel but we may not yet detect the angles. For example, we are not sure whether what we see is a rectangle or a parallelogram. This stage corresponds to affine transformation.
- Then, we have a stage of similarity transformations – when we detect the shape but cannot yet detect its size.
- Finally, when the object is close enough, we can detect both its shape and its size.

Each stage can be thus described by an appropriate transformation group. So, Wiener conjectured that if there was a group intermediate between, e.g., all projective and all continuous transformations, our vision mechanism – the result of millions of years of improving evolution – would have used it. Thus, he formulated a hypothesis that such intermediate transformation groups are not possible [30].

Many mathematicians did not take this hypothesis too seriously – while they appreciated Wiener's engineering ideas, they thought that he was going too far in his analogies. But other mathematicians took it seriously – and, two decades after the first edition of Wiener's book, they came up with a formal proof that, indeed, under reasonable conditions, there is only one transformation group that contains all linear (= affine) transformations and some non-linear ones: namely, the group of all projective transformations [8,27].

The general proof is very complicated – e.g., the paper [27] consists of more than 100 pages of dense mathematics. But good news is that at present, we are only interested in the transformations of 1D signals. In this case, projective transformations are nothing else but fractional-linear ones

$$f(x) = \frac{a \cdot x + b}{c \cdot x + d},$$

and the corresponding proof can be shortened to a few pages; see, e.g., [19,31].

So, we arrive at the following conclusion.

So what are the limit transformations? We have shown that limit transformations form a finite-parametric transformation group that contains all linear transformations, and that all transformations from such a group are fractional linear – with linear ones being a particular case.

Thus, we conclude that all limit transformations are fractional-linear.

A similar conclusion can be made about all possible reasonable transformations. Instead of looking for *limit* transformations, we can consider a different problem: to describe a class of all transformations which are, in some sense, *reasonable*. Linear transformations are reasonable: shift corresponds to the changing the starting point and a multiplication by a number corresponds to changing a measuring unit. A good example of both transformations are transformation between Celsius and Fahrenheit temperature scales.

It is also natural to conclude that a composition of two reasonable transformations is reasonable, and that a transformation which is inverse to a reasonable transformation is also reasonable. If we want to use computers to deal with reasonable transformations, it also makes sense to require that the reasonable transformations form a finite-parametric family – since in a computer, we can only stored finitely many parameter values.

Thus, the class of all reasonable transformations forms a finite-parametric transformation group containing all linear transformations. So, we conclude that every reasonable transformation is fractional linear.

What are the implications for neural networks. Artificial neural networks – a perfect example of Wiener's belief that emulating biological systems can be beneficial – are formed of *neurons*. In a neuron, first, we form a linear combination x of the inputs x_i,

and then we apply some non-linear transformation $y = s(x)$ to this linear combination. In neural networks, this linear transformation is known as an *activation function*.

Which activation function should we use? The first nonlinear neurons use *sigmoid* activation function

$$s(x) = \frac{1}{1 + \exp(-x)},$$

because, in the first approximation, this is how signals are processed in biological neurons; see, e.g., [2]. This activation function worked very well – much better than other activation functions that have been tried. This activation function is still often used in some layers of deep neural networks [7], where they are also very successful. How can we explain this success?

A natural explanation comes from the fact that, as we have mentioned earlier, all inputs come with noise. The simplest case is when, for each measurement, we just have a constant noise $n_i = \text{const}$, when instead of the actual values x_i, the measurement results are shifted by this value n_i, to $x_i + n_i$. As a result, the linear combination x is also shifted by some constant n (which is the similar linear combination of noises n_i):

$$x \to x + n.$$

We do not know the exact value of this noise – if we knew, we could simply subtract it from all the measured values. It is therefore reasonable to require that the result of applying the activation functions should be insensitive to this noise as much as possible.

Of course, we cannot simply require that $s(x + n) = s(x)$ for all x and n – this would imply that the function $s(x)$ is a constant that does not depend on the input at all. This makes sense: for example, the formula $d = v \cdot t$ showing that the distance can be obtained by multiplying velocity and time does not change when we change the unit of time, e.g., from hours to seconds. However, this invariance does not mean that the formula remains exactly the same when we change the unit of time: to keep the formula the same, we also need to apply an appropriate transformation to velocity as well: namely, replace the values in km/h with a value in km/sec. Similarly here, a natural idea is to require that if we apply a shift $x \to x' = x + n$ to the input, the formula remains the same if we applying an appropriate transformation to y as well, i.e., that $y' = s(x')$, where $y' = T(y)$ for some reasonable transformation y.

In other words, we conclude that for every value n, there exists some reasonable transformation T_n for which $s(x') = T_n(y)$. Here, $x' = x + n$, and $y = s(x)$, so $s(x + n) = T_n(s(x))$. We have already concluded that reasonable transformations are fractional linear, thus we have

$$s(x + n) = \frac{a(n) \cdot x + b(n)}{c(n) \cdot x + d(n)}$$

for some values $a(n)$ through $d(n)$. To describe all the functions $s(x)$ that have this property, we can differentiate both side of this equation by n and take $n = 0$. The resulting differential equation can then be explicitly solved; see, e.g., [12,16,19]. The generic monotonic solution to this equation indeed differs from the sigmoid activation functions only by linear transformations of x and y.

This explains why the sigmoid activation function indeed works well.

Comment. Similar invariance ideas can explain the rectified linear activation functions used in deep learning – as well many other empirically successful features of deep learning algorithms; see, e.g., [12].

5. Conclusions

In this paper, we showed that limit theorems – similar to the Central Limit Theorem from statistics – make analysis of complex systems easier – i.e., lead to the blessing-of-dimensionality phenomenon. We showed that this simplification happens for all the aspects of these systems:

- for the corresponding transformations – as shown, e.g., by the description of all possible limit and/or reasonable transformations, and by the resulting theoretical explanation of the efficiency of sigmoid activation functions;
- for the system's uncertainty – as shown, e.g., by the use of limit distributions such as normal and Cauchy to make uncertainty quantification more efficient, and by the use of limit theorems to explain the ubiquity of interval uncertainty, and
- the desired result of the system's analysis – as shown, e.g., by a limit-theorem-based explanation of why it is usually possible to meaningfully classify objects into a small finite number of classes.

Author Contributions: Both authors contributed equally to this paper. Both authors have read and agreed to the published version of the manuscript.

Funding: This work was supported in part by the National Science Foundation grants 1623190 (A Model of Change for Preparing a New Generation for Professional Practice in Computer Science), and HRD-1834620 and HRD-2034030 (CAHSI Includes). It was also supported by the program of the development of the Scientific-Educational Mathematical Center of Volga Federal District No. 075-02-2020-1478.

Acknowledgments: The authors are greatly thankful to Alexander Gorban for his encouragement and valuable discussions.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Beirlant, J.; Goegebeur, Y.; Segers, J.; Teugels, J. *Statistics of Extremes: Theory and Applications*, Wiley: New York, 2004.
2. Bishop, C.M. *Pattern Recognition and Machine Learning*, Springer: New York, 2006.
3. Coles, S. *An Introduction to Statistical Modeling of Extreme Values*, Springer Verlag: London, 2001.
4. de Haan, L.; Ferreira, A. *Extreme Value Theory: An Introduction*, Springer Verlag: New York, 2006.
5. Embrechts, P.; Klüppelberg, C.; Mikosch, T. *Modelling Extremal Events for Insurance and Finance*, Springer Verlag: Berlin, 1997.
6. Feynman, R.; Leighton, R.; Sands, M. *The Feynman Lectures on Physics*, Addison Wesley: Boston, Massachusetts, 2005.
7. Goodfellow, I.; Bengio, Y.; Courville, A. *Deep Leaning*, MIT Press: Cambridge, Massachusetts, 2016.
8. Guillemin, V.M.; Sternberg, S. An algebraic model of transitive differential geometry, *Bulletin of American Mathematical Society* **1964**, *70*(1), 16–47.
9. Gumbel, E.J. *Statistics of Extremes*, Dover: New York, 2013.
10. Kotz, S.; Nadarajah, S. *Extreme Value Distributions: Theory and Applications*, Imperial College Press: London, UK, 2000.
11. Kreinovich, V. Why intervals? A simple limit theorem that is similar to limit theorems from statistics", *Reliable Computing* **1995**, *1*(1), 33–40.
12. Kreinovich, V.; Kosheleva, O. Optimization under uncertainty explains empirical success of deep learning heuristics", In: Pardalos, P.; Rasskazova, V.; Vrahatis, M.N. (eds.), *Black Box Optimization, Machine Learning and No-Free Lunch Theorems*, Springer: Cham, Switzerland, 2021, pp. 195–220.
13. Kreinovich, V.; Ferson, S. A new Cauchy-based black-box technique for uncertainty in risk analysis, *Reliability Engineering and Systems Safety* **2004**, *85*(1–3), 267–279.
14. Kreinovich, V.; Lakeyev, A.; Rohn, J.; Kahl, P. *Computational Complexity and Feasibility of Data Processing and Interval Computations*, Kluwer: Dordrecht, 1998.
15. Kreinovich, V.; Nguyen, H.T.; Sriboonchitta, S.; Kosheleva, O. Modeling extremal events is not easy: why the extreme value theorem cannot be as general as the central limit theorem. In: Kreinovich, V. (ed.), *Uncertainty Modeling*, Springer Verlag: Cham, Switzerland, 2017, 123–134.
16. Kreinovich, V.; Quintana, C. Neural networks: what non-linearity to choose?, *Proceedings of the 4th University of New Brunswick Artificial Intelligence Workshop*, Fredericton, N.B., Canada, 1991, 627–637.
17. Mayer, G. *Interval Analysis and Automatic Result Verification*, de Gruyter: Berlin, 2017.
18. Moore, R.E.; Kearfott, R.B.; Cloud, M.J. *Introduction to Interval Analysis*, SIAM: Philadelphia, 2009.
19. Nguyen, H.T.; Kreinovich, V. *Applications of Continuous Mathematics to Computer Science*, Kluwer: Dordrecht, 1997.
20. Nguyen, H.T.; Wu, B.: Kreinovich, V. Our reasoning is clearly fuzzy, so why is crisp logic so often adequate?, *International Journal of Intelligent Technologies and Applied Statistics (IJITAS)* **2015**, *8*(2), 133–137.
21. Nocedal, G.; Wright, S.J. *Numerical Optimization*, Springer: New York, 2006.
22. Novak, S.Y. *Extreme Value Methods with Applications to Finance*, Chapman & Hall/CRC Press: London, 2011.
23. Rabinovich, S.G. *Measurement Errors and Uncertainties: Theory and Practice*, Springer: New York, 2005.
24. Resnick, S.I. *Extreme Values, Regular Variation and Point Processes*, Springer Verlag: New York, 2008.

25. Roginskaya, M.M.; Shulman, V.S. On Minkoswki sums of many small sets, *Functional Analysis and Its Applications* **2018**, *52*(3), 233–235.
26. Sheskin, D.J. *Handbook of Parametric and Non-Parametric Statistical Procedures*, Chapman & Hall/CRC: London, UK, 2011.
27. Singer, I.M.; Sternberg, S. Infinite groups of Lie and Cartan, Part 1, *Journal d'Analyse Mathematique* **1965**, *XV*, 1–113.
28. Tsirel'son, B.S. A geometrical approach to maximum likelihood estimation for infinite-dimensional Gaussian location. I, *Theory of Probability and its Applications* **1982**, *27*, 411–418.
29. Urenda, J.C.; Kosheleva, O.; Kreinovich, V. How to describe measurement errors: a natural generalization of the Central Limit Theorem beyond normal (and other infinitely divisible) distributions, In: Pavese, F.; Forbes, A.B.; Zhang, N.F.; Chunovkina, A.G. (eds.), *Advanced Mathematical and Computational Tools in Metrology and Testing XII*, World Scientific: Singapore, to appear.
30. Wiener, N. *Cybernetics, or Control and Communication in the Animal and the Machine*, 3rd edition, MIT Press: Cambridge, Massachusetts, 1962.
31. Zapata, F.; Kosheleva, O.; Kreinovich, V. Wiener's conjecture about transformation groups helps predict which fuzzy techniques work better, *Proceedings of the 2014 Annual Conference of the North American Fuzzy Information Processing Society NAFIPS'2014*, Boston, Massachusetts, June 24–26, 2014.