7-2019

# Why LASSO, EN, and CLOT: Invariance-Based Explanation

Hamza Alkhatib
*Leibniz University Hannover*, alkhatib@gih.uni-hannover.de

Ingo Neumann
*Leibniz University Hannover*, neumann@gih.uni-hannover.de

Vladik Kreinovich
*The University of Texas at El Paso*, vladik@utep.edu

Chon Van Le
*International University VNU HCMC*, lvchon@hcmiu.edu.vn



Comments:
Technical Report: UTEP-CS-19-72

# Why LASSO, EN, and CLOT: Invariance-Based Explanation

Hamza Alkhatib[1], Ingo Neumann[1],
Vladik Kreinovich[2], and Chon Van Le[3]
[1]Geodetic Institute
Leibniz University of Hannover
Nienburger Str. 1, 30167 Hannover, Germany
alkhatib@gih.uni-hannover.de, neumann@gih.uni-hannover.de
[2]Department of Computer Science
University of Texas at El Paso
500 W. University, El Paso, TX 79968, USA
vladik@utep.edu
[3]International University – VNU HCMC
Ho Chi Minh City, Vietnam
lvchon@hcmiu.edu.vn

**Abstract**

In many practical situations, observations and measurement results are consistent with many different models – i.e., the corresponding problem is ill-posed. In such situations, a reasonable idea is to take into account that the values of the corresponding parameters should not be too large; this idea is known as regularization. Several different regularization techniques have been proposed; empirically the most successful are LASSO method, when we bound the sum of absolute values of the parameters, and EN and CLOT methods in which this sum is combined with the sum of the squares. In this paper, we explain the empirical success of these methods by showing that they are the only ones which are invariant with respect to natural transformations – like scaling which corresponds to selecting a different measuring unit.

## 1 Formulation of the Problem

**Need for solve the inverse problem.** Once we have a model of a system, we can use this model to predict the system's behavior, in particular, to predict the results of future measurements and observations of this system. The problem of estimating future measurement results based on the model is known as the *forward problem*.

In many practical situations, we do not know the exact model. To be more precise, we know the general form of a dependence between physical quantities, but the parameters of this dependence need to be determined from the observations and from the results of the experiment. For example, often, we have a linear model $y = a_0 + \sum\limits_{i=1}^{n} a_i \cdot x_i$, in which the parameters $a_i$ need to be experimentally determined. The problem of determining the parameters of the model based on the measurement results is known as the *inverse problem*.

To actually find the parameters, we can use, e.g., the Maximum Likelihood method. For example, when the errors are normally distributed, then the Maximum Likelihood procedure results in the usual Least Squares estimates; see, e.g., [7]. For example, for a general linear model with parameters $a_i$, once we know several tuples of corresponding values $(x_1^{(k)}, \ldots, x_n^{(k)}, y^{(k)})$, $1 \le k \le K$, then we can find the parameters from the condition that

$$\sum_{k=1}^{K} \left( y^{(k)} - \left( a_0 + \sum_{i=1}^{n} a_i \cdot x_i^{(k)} \right) \right)^2 \to \min_{a_0, \ldots, a_n}. \tag{1}$$

**Need for regularization.** In some practical situations, based on the measurement results, we can determine all the model's parameters with reasonably accuracy. However, in many other situations, the inverse problem is *ill-defined* in the sense that several very different combinations of parameters are consistent with all the measurement results.

This happens, e.g., in dynamical systems, when the observations provide a smoothed picture of the system's dynamics. For example, if we are tracing the motion of a mechanical system caused by an external force, then a strong but short-time force in one direction followed by a similar strong and short-time force in the opposite direction will (almost) cancel each other, so the same almost-unchanging behavior is consistent both with the absence of forces and with the above wildly-oscillating force. A similar phenomenon occurs when, based on the observed economic behavior, we try to reconstruct the external forces affecting the economic system.

In such situations, the only way to narrow down the set of possible solution is to take into some general a priori information. For example, for forces, we may know – e.g., from experts – the upper bound on each individual force, or the upper bound on the overall force. The use of such a priori information is known as *regularization*; see, e.g., [10].

**Which regularizations are currently used.** There are many possible regularizations. Many of them have been tried, and, based on the results of these tries, a few techniques turned out to be empirically successful.

The most widely used technique of this type is known as LASSO technique (short of Least Absolute Shrinkage and Selection Operator), where we look for solutions for which the sum of the absolute values $\|a\|_1 \stackrel{\text{def}}{=} \sum\limits_{i=0}^{n} |a_i|$ is bounded

by a given value; see, e.g., [8]. Another widely used method is a *ridge regression* method, in which we limit the sum of the squares $S \stackrel{\text{def}}{=} \sum_{i=0}^{n} a_i^2$ or, equivalently, its square root $\|a\|_2 \stackrel{\text{def}}{=} \sqrt{S}$; see, e.g., [4, 9]. Very promising are also:

- the *Elastic Net* (EN) method, in which we limit a linear combination $\|a\|_1 + c \cdot S$ (see, e.g., [6, 11]), and
- the *Combined L-One and Two* (CLOT) method in which we limit a linear combination $\|a\|_1 + c \cdot \|a\|_2$; see, e.g., [2].

**Why: remaining question and what we do in this paper.** The above empirical facts prompt a natural question: why the above regularization techniques work the best? In this paper, we show that the efficiency of these methods can be explained by the natural invariance requirements.

## 2 General and Probabilistic Regularizations

**General idea of regularization and its possible probabilistic background.** In general, regularization means that we dismiss values $a_i$ which are too large or too small. In some cases, this dismissal is based on subjective estimations of what is large and what is small. In other cases, the conclusion about what is large and what is not large is based on past experience of solving similar problem – i.e., on our estimate of the frequencies (= probabilities) with which different values have been observed in the past. In this paper, we consider both types of regularization.

**Probabilistic regularization: towards a precise definition.** There is no a priori reason to believe that different parameters have different distributions. So, in the first approximation, it makes sense to assume that they have the same probability distribution. Let us denote the probability density function of this common distribution by $\rho(a)$.

In more precise terms, the original information is invariant with respect to all possible permutations of the parameters; thus, it makes sense to conclude that the resulting joint distribution is also invariant with respect to all these permutations – which implies, in particular, that all the marginal distributions are the same.

Similarly, in general, we do not have a priori reasons to prefer positive or negative values of each the coefficients, i.e., the a priori information is invariant with respect to changing the sign of each of the variables: $a_i \to -a_i$. It is therefore reasonable to conclude that the marginal distribution should also be invariant, i.e., that we should have $\rho(-a) = \rho(a)$, and thus, $\rho(a) = \rho(|a|)$.

Also, there is no reason to believe that different parameters are positively or negatively correlated, so it makes sense to assume that their distributions are statistically independent. This is in line with the general Maximum Entropy (= Laplace Indeterminacy Principle) ideas [5], according to which we should

pretend to be certain – to be more precise, if several different probability distributions are consistent with our knowledge:

- we should *not* select distributions with small entropy (measure of uncertainty),
- we *should* select the one for which the entropy is the largest.

If all we know are marginal distributions, then this principle leads to the conclusion that the corresponding variables are independent; see, e.g., [5].

Due to the independence assumption, the joint distribution of $n$ variables $a_i$ take the form $\rho(a_0, a_1, \ldots, a_n) = \prod_{i=0}^{n} \rho(|a_i|)$. In applications of probability and statistics, it is usually assumed, crudely speaking, that events with very small probability are not expected to happen. This is the basis for all statistical tests – e.g., if we assume that the distribution is normal with given mean and standard deviation, and the probability that this distribution will lead to the observed data is very small (e.g., if we observe a 5-sigma deviation from the mean), then we can conclude, with high confidence that experiments disprove our assumption. In other words, we take some threshold $t_0$, and we consider only the tuples $a = (a_0, a_1, \ldots, a_n)$ for which $\rho(a_0, a_1, \ldots, a_n) = \prod_{i=0}^{n} \rho(|a_i|) \geq t_0$. By taking logarithms of both sides and changing signs, we get an equivalent inequality

$$\sum_{i=0}^{n} \psi(|a_i|) \leq p_0, \tag{2}$$

where we denoted $\psi(z) \stackrel{\text{def}}{=} -\ln(\rho(z))$ and $p_0 \stackrel{\text{def}}{=} -\ln(t_0)$. (The sign is changed for convenience, since for small $t_0 \ll 1$, logarithm is negative, and it is more convenient to deal with positive numbers.)

Our goal is to avoid coefficients $a_i$ whose absolute values are too large. Thus, if the absolute values $(|a_0|, |a_1| \ldots, |a_n|)$ satisfy the inequality (2), and we decrease one of the absolute values, the result should also satisfy the same inequality. Thus the function $\psi(z)$ must be increasing.

We want to find the minimum of the usual least squares (or similar) criterion under the constraint (2). The minimum is attained:

- either when in (2), we have strict inequality
- or when we have equality.

If we have a strict inequality, then we get a local minimum, and for convex criteria like least squares (where there is only one local minimum which is also global), this means that we have the solution of the original constraint-free problem – and we started this whole discussion by considering situations in which this straightforward approach does not work. Thus, we conclude that the minimum under constraint (2) is attained when we have the equality, i.e., when

we have

$$\sum_{i=0}^{n} \psi(|a_i|) = p_0 \tag{3}$$

for some function $\psi(z)$ and for some value $p_0$.

In practice, most probability distributions are continuous – step-wise and point-wise distributions are more typically found in textbooks than in practice. Thus, it is reasonable to assume that the probability density $\rho(x)$ is continuous. Then, its logarithm $\psi(z) = \ln(\rho(z))$ is continuous as well. Thus, we arrive at the following definition.

**Definition 1.** *By a* probabilistic constraint*, we mean a constraint of the type* (3) *corresponding to some continuous increasing function $\psi(z)$ and to some number $p_0$.*

**General regularization.** In the general case, we do not get any probabilistic justification of our approach, we just deal with the values $|a_i|$ themselves, without assigning probability to different possible values. In general, similarly to the probabilistic case, there is no reason to conclude that large positive values of $a_i$ are better or worse than negative values with similar absolute value. Thus, we can say that a very large value $a$ and its opposite $-a$ are equally impossible. The absolute value of each coefficient can be thus used as its "degree of impossibility": the larger the number, the less possible it is that this number will appear as the absolute value of a coefficient $a_i$.

Based on the degrees of impossibility of $a_0$ and $a_1$, we need to estimate the degree of impossibility of the pair $(a_0, a_1)$. Let us denote the corresponding estimate by $|a_0| * |a_1|$. If the second coefficient $a_1$ is 0, it is reasonable to say that the degree of impossibility of the pair $(a_0, 1_1)$ is the same as the degree of impossibility of $a_0$, i.e., equal to $|a_0|$: $|a_0| * 0 = |a_0|$. f the second coefficient is not 0, the situation becomes slightly worse that when it was 0, so: if $a_1 \neq 0$, then $|a_0| * |a_1| > |a_0| * 0 = |a_0|$. In general, if the absolute value of one of the coefficients increases, the overall degree of impossibility should increase.

Once we know the degree of impossibility $|a_0| * |a_1|$ of a pair, we can combine it with the degree of impossibility $|a_2|$ of the third coefficient $a_2$, and get the estimated degree of impossibility $(|a_0| * |a_1|) * |a_2|$ of a triple $(a_0, a_1, a_2)$, etc., until we get the degree of impossibility of the whole tuple.

The result of applying this procedure should not depend on the order in which we consider the coefficients, i.e., we should be $a * b = b * a$ (commutativity) and $(a * b) * c = a * (b * c)$ (associativity).

We should consider only the tuples for which the degree of impossibility does not exceed a certain threshold $t_0$:

$$|a_0| * |a_1| * \ldots * |a_n| \leq t_0 \tag{4}$$

for some $t_0$. Thus, we arrive at the following definitions.

**Definition 2.** *By a* combination operation*, we mean a function $*$ that maps two non-negative numbers into a new non-negative number which is:*

- *commutative,*
- *associative,*
- *has the property* $a * 0 = a$ *and*
- *is* monotonic *in the sense that if* $a < a'$*, then* $a * b < a' * b$*.*

**Definition 3.** *By a* general constraint*, we means a constraint of the type* (4) *for some combination operation* $*$ *and for some number* $t_0 > 0$*.*

# 3 Natural Invariances

**Scale-invariance: general idea.** The numerical values of physical quantities depend on the selection of a measuring unit. For example, if we previously used meters and now start using centimeters, all the physical quantities will remain the same, but the numerical values will change – they will all get multiplied by 100.

In general, if we replace the original measuring unit with a new measuring unit which is $\lambda$ times smaller, then all the numerical values get multiplied by $\lambda$:

$$x \to x' = \lambda \cdot x.$$

Similarly, if we change the original measuring units for the quantity $y$ to a new unit which is $\lambda$ times smaller, then all the coefficients $a_i$ in the corresponding dependence $y = a_0 + \ldots + a_i \cdot x_i + \ldots$ will also be multiplied by the same factor: $a_i \to \lambda \cdot a_i$.

**Scale-invariance: case of probabilistic constraints.** It is reasonable to require that the corresponding constraints should not depend on the choice of a measuring unit. Of course, if we change $a_i$ to $\lambda \cdot a_i$, then the value $p_0$ may also need to be accordingly changed, but overall, the constraint should remain the same. Thus, we arrive at the following definition.

**Definition 4.** *We say that probability constraints corresponding to the function* $\psi(z)$ *are* scale-invariant *if for every* $p_0$ *and for every* $\lambda > 0$*, there exists a value* $p_0'$ *such that*

$$\sum_{i=0}^{n} \psi(|a_i|) = p_0 \Leftrightarrow \sum_{i=0}^{n} \psi(\lambda \cdot |a_i|) = p_0'. \tag{5}$$

**Scale-invariance: case of general constraints.** In general, the degree of impossibility is described in the same units as the coefficients themselves. Thus, invariance would mean that if replace $a$ and $b$ with $\lambda \cdot a$ and $\lambda \cdot b$, then the combined value $a * b$ will be replaced by a similarly re-scaled value $\lambda \cdot (a * b)$. Thus, we arrive at the following definition:

**Definition 5.** *We say that a general constraint corresponding to a combination operation $*$ is* scale-invariance *if for every $a$, $b$, and $\lambda$, we have*

$$(\lambda \cdot a) * (\lambda \cdot b) = \lambda \cdot (a * b). \tag{6}$$

In this case, the corresponding constraint is naturally scale-invariant: if $*$ is scale-invariant operation, then, for all $a_i$ and for all $\lambda$, we have

$$|\lambda \cdot a_0| * |\lambda \cdot a_1| * \ldots * |\lambda \cdot a_n| = \lambda * (|a_0| * |a_1| * \ldots * |a_n|)$$

and thus,

$$|a_0| * |a_1| * \ldots * |a_n| = t_0 \Leftrightarrow |\lambda \cdot a_0| * |\lambda \cdot a_1| * \ldots * |\lambda \cdot a_n| = t_0' \stackrel{\text{def}}{=} \lambda \cdot t_0.$$

**Shift-invariance: general idea.** Our goal is to minimize the deviations of the coefficients $a_i$ from 0. In the ideal case, when the model is exact and when measurement errors are negligible, in situations when there is no signal at all (i.e., when $a_i = 0$ for all $i$), we will measure exactly 0s and reconstruct exactly 0 values of $a_i$. In this case, even if we do not measure some of the quantities, we should also return all 0s. In this ideal case, any deviation of the coefficients from 0 is an indication that something is not right.

In practice, however, all the models are approximate. Because of the model's imperfection and measurement noise, even if we start with a case when $a_i = 0$ for all $i$, we will still get some non-zero values of $y$ and thus, some non-zero values of $a_i$ (hopefully, small, but still non-zero). In such situations, small deviations from 0 are OK, they do not necessarily indicate that something is wrong.

We can deal with this phenomenon in two different ways:

- we can simply have this phenomenon in mind when dealing with the original values of the coefficients $a_i$ – and do not change any formulas,
- or we can explicitly subtract an appropriate small tolerance level $\varepsilon > 0$ from the absolute values of all the coefficients, i.e., replace the original values $|a_i|$ with the new values $|a_i| - \varepsilon$ thus explicitly taking into account that deviations smaller that this tolerance level are OK, and only values above this level are problematic.

It is reasonable to require that the corresponding constraints do not change under this shift $|a| \to |a| - \varepsilon$.

**Shift-invariance: case of probabilistic constraints.** If we change $|a_i|$ to $|a_i| - \varepsilon$, then the coefficient $p_0$ may also need to be accordingly changed, but overall, the constraint should remain the same. Thus, we arrive at the following definition.

**Definition 6.** *We say that probability constraints corresponding to the function $\psi(z)$ are* shift-invariant *if for every $p_0$ and for every sufficiently small $\varepsilon > 0$, there exists a value $p_0'$ such that*

$$\sum_{i=0}^{n} \psi(|a_i|) = p_0 \Leftrightarrow \sum_{i=0}^{n} \psi(|a_i| - \varepsilon) = p_0'. \tag{7}$$

**Shift-invariance: case of general constraints.** In general, the degree of impossibility is described in the same units as the coefficients themselves. Thus, invariance would mean that if replace $a$ and $b$ with $a - \varepsilon$ and $b - \varepsilon$, then the combined value $a * b$ will be replaced by a similarly re-scaled value $(a * b) - \varepsilon'$. Here, $\varepsilon'$ may be different from $\varepsilon$, since it represents deleting two small values, not just one. A similar value should exist for all $n$. Thus, we arrive at the following definition:

**Definition 7.** *We say that a general constraint corresponding to a combination operation $*$ is* shift-invariance *if for every $n$ and for all sufficiently small $\varepsilon > 0$ there exists a value $\varepsilon' > 0$ such that for every $a_0, a_1, \ldots, a_n > 0$, we have*

$$(a_0 - \varepsilon) * (a_1 - \varepsilon) * \ldots * (a_n - \varepsilon) = (a_0 * a_1 * \ldots * a_n) - \varepsilon'. \tag{8}$$

In this case, the corresponding constraint is naturally shift-invariant: if $*$ is a shift-invariant operation, then, for all $a_i$ and for all sufficiently small $\varepsilon > 0$, we have

$$|a_0| * |a_1| * \ldots * |a_n| =$$

$$t_0 \Leftrightarrow (|a_0| - \varepsilon) * (|a_1| - \varepsilon) * \ldots * |(|a_n| - \varepsilon) = t_0' \stackrel{\text{def}}{=} t_0 - \varepsilon.$$

# 4 Why LASSO: First Result

Let us show that for both types of constraints, natural invariance requirements lead to LASSO formulas.

**Proposition 1.** *Probabilistic constraints corresponding to a function $\psi(x)$ are shift- and scale-invariant if and only $\psi(z)$ is a linear function $\psi(z) = k \cdot a + \ell$.*

**Discussion.** For a linear function, the corresponding constraint $\sum\limits_{i=0}^{n} \psi(|a_i|) = p_0$ is equivalent to the LASSO constraints $\sum\limits_{i=1}^{n} |a_i| = t_0'$, with $t_0' \stackrel{\text{def}}{=} (t_0 - \ell)/k$. Thus, Proposition 1 explains why probabilistic constraints should be LASSO constraints: LASSO constraints are the only probabilistic constraints that satisfy natural invariance requirements.

**Proposition 2.** *General constraints corresponding to a combination function $*$ are shift- and scale-invariant if and only if the operation $*$ is addition $a*b = a+b$.*

**Discussion.** For addition, the corresponding constraint $\sum_{i=0}^{n} |a_i| = t_0$ is exactly the LASSO constraints. Thus, Proposition 2 explains why general constraints should be LASSO constraints: LASSO constraints are the only general constraints that satisfy natural invariance requirements.

**Proof of Proposition 1.** Scale-invariance implies that if $\psi(a)+\psi(b) = \psi(c)+\psi(0)$, then, for every $\lambda > 0$, we should have $\psi(\lambda\cdot a)+\psi(\lambda\cdot b) = \psi(\lambda\cdot c)+\psi(0)$. If we subtract $2\psi(0)$ from both sides of each of these equalities, then we can conclude that for the auxiliary function $\Psi(z) \stackrel{\text{def}}{=} \psi(a)-\psi(0)$, if $\Psi(a)+\Psi(b) = \Psi(c)$, then $\Psi(\lambda\cdot a)+\Psi(\lambda\cdot b) = \Psi(\lambda\cdot c)$. So, for the mapping $f(z)$ that transforms $z = \Psi(a)$ into $f(z) = \Psi(\lambda\cdot a)$ – i.e., for $f(z) \stackrel{\text{def}}{=} \Psi(\lambda\cdot\Psi^{-1}(z))$, where $\Psi^{-1}(z)$ denotes the inverse function – we conclude that if $z+z' = z''$ then $f(z)+f(z') = f(z'')$. In other words, $f(z+z') = f(z)+f(z')$. It is known that the only monotonic functions with this property are linear functions $f(z) = c\cdot z$; see, e.g., [1].

Since $z = \Psi(a)$ and $f(z) = \Psi(\lambda\cdot a)$, we thus conclude that for every $\lambda$, there exists a value $c$ (which, in general, depends on $\lambda$) for which $\Psi(\lambda\cdot a) = c(\lambda)\cdot\Psi(a)$. Every monotonic solution to this functional equation has the form $\Psi(a) = A\cdot a^\alpha$ for some $A$ and $\alpha$, so $\psi(a) = \Psi(a)+\psi(0) = A\cdot a^\alpha + B$, where $B \stackrel{\text{def}}{=} \psi(0)$.

Similarly, shift-invariance implies that if $\psi(a)+\psi(b) = \psi(c)+\psi(d)$, then, for each sufficiently small $\varepsilon > 0$, we should have

$$\psi(a-\varepsilon)+\psi(b-\varepsilon) = \psi(c-\varepsilon)+\psi(d-\varepsilon).$$

The inverse is also true, so the same property holds for $\varepsilon = -\delta$, i.e., if $\psi(a)+\psi(b) = \psi(c)+\psi(d)$, then, for each sufficiently small $\delta > 0$, we should have

$$\psi(a+\delta)+\psi(b+\delta) = \psi(c+\delta)+\psi(d+\delta).$$

Substituting the expression $\psi(a) = A\cdot a^\alpha + B$, subtracting $2B$ from both sides of each equality and dividing both equalities by $A$, we conclude that if $a^\alpha+b^\alpha = c^\alpha+d^\alpha$, then $(a+\delta)^\alpha+(b+\delta)^\alpha = (c+\delta)^\alpha+(d+\delta)^\alpha$. In particular, the first equality is satisfied if we have $a = b = 1$, $c = 2^{1/\alpha}$, and $d = 0$. Thus, for all sufficiently small $\delta$, we have $2\cdot(1+\delta)^\alpha = (2^{1/\alpha}+\delta)^\alpha+\delta^\alpha$.

On both sides, we have analytical expressions. When $\alpha < 1$, then for small $\delta$, the left-hand side term and the first term in the right0hand side start with linear term $\delta$, and the terms $\delta^\alpha \gg \delta$ is not compensated. If $\alpha > 1$, then by equating terms linear in $\delta$ in the corresponding expansions, we get $2\alpha\cdot\delta$ in the left-hand side and $\alpha\cdot(2^{1/\alpha})^{\alpha-1}\cdot\delta = 2^{1-1/\alpha}\cdot\alpha\cdot\delta$ in the right-hand side – the coefficients are different, since the corresponding powers of two are different: $1 \neq 1-1/\alpha$. Thus, the only possibility is $\alpha = 1$. The proposition is proven.

**Proof of Proposition 2.** It is known (see, e.g., [3]) that every scale-invariant combination operation has the form $a*b = (a^\alpha+b^\alpha)^{1/\alpha}$ or $a*b = \max(a,b)$. The second case contradicts the requirement that $a*b$ be strictly increasing in both variables. For the first case, similarly to the proof of Proposition 1, we conclude that $\alpha = 1$. The proposition is proven.

# 5 Why EN and CLOT

**Need to go beyond LASSO.** In the previous section, we showed that if we need to select a single method for all the problems, then natural invariance requirements lead to LASSO, i.e., to bounds on the sum of the absolute values of the parameters. In some practical situations, this works, while in others, it does not lead to good results. To deal with such situations, instead of fixing a *single* method for all the problems, a natural idea is to select a *family* of methods, so that in each practical situation, we should select an appropriate method from this family. Let us analyze how we can do it both for probabilistic and for general constraints.

**Probabilistic case.** Constraints in the probabilistic case are described by the corresponding function $\psi(z)$. The LASSO case corresponds to a 2-parametric family $\psi(z) = c_0 + c_1 \cdot z$. In terms of the corresponding constraints, all these functions from this family are equivalent to $\psi(z) = z$.

To get a more general method, a natural idea is to consider a 3-parametric family, i.e., a family of the type $\psi(z) = c_0 + c_1 \cdot z + c_2 \cdot f(z)$ for some function $f(z)$. Constraints related to this family are equivalent to using the functions $\psi(z) = z + c \cdot f(z)$ for some function $f(z)$. Which family – i.e., which function $f(z)$ – should we choose? A natural idea is to again use scale-invariance and shift-invariance.

**Definition 8.** *We say that functions $\psi_1(z)$ and $\psi_2(z)$ are* constraint-equivalent *if:*

- *for each $n$ and for each $c_1$, there exists a value $c_2$ such that the condition $\sum_{i=0}^{n} \psi_1(a_i) = c_1$ is equivalent to $\sum_{i=0}^{n} \psi_2(a_i) = c_2$, and*
- *for each $n$ and for each $c_2$, there exists a value $c_1$ such that the condition $\sum_{i=0}^{n} \psi_2(a_i) = c_2$ is equivalent to $\sum_{i=0}^{n} \psi_1(a_i) = c_1$.*

**Definition 9.**

- *We say that a family $\{z + c \cdot f(z)\}_c$ is* scale-invariant *if for each $c$ and $\lambda$, there exists a value $c'$ for which the re-scaled function $\lambda \cdot z + c \cdot f(\lambda \cdot z)$ is constraint-equivalent to $z + c' \cdot f(z)$.*
- *We say that a family $\{z + c \cdot f(z)\}_c$ is* shift-invariant *if for each $c$ and for each sufficiently small number $\varepsilon$, there exists a value $c'$ for which the shifted function $z - \varepsilon + c \cdot f(z - \varepsilon)$ is constraint-equivalent to $z + c' \cdot f(z)$.*

**Proposition 3.** *A family $\{z + c \cdot f(z)\}_c$ corresponding to a smooth function $f(z)$ is scale- and shift-invariant if and only if the function $f(z)$ is quadratic.*

**Discussion.** Thus, it is sufficient to consider functions $\psi(z) = z + c \cdot z^2$. This is exactly the EN approach – which is thus justified by the invariance requirements.

**Proof.** Similarly to the proof of Proposition 1, from the shift-invariance, for $c = 1$, we conclude that $z - \varepsilon + f(z - \varepsilon) = A + B \cdot (z + c' \cdot f(z))$ for some values $A$, $B$, and $c'$ which are, in general, depending on $\varepsilon$. Thus,

$$f(z - \varepsilon) = A_0(\varepsilon) + A_1(\varepsilon) \cdot z + A_2(\varepsilon) \cdot f(z), \tag{9}$$

where $A_0(\varepsilon) \stackrel{\text{def}}{=} A + \varepsilon$, $A_1(\varepsilon) \stackrel{\text{def}}{=} B - 1$, and $A_2(\varepsilon) \stackrel{\text{def}}{=} B \cdot c'$.

By considering three different values $x_k$ ($k = 1, 2, 3$), we get a system of three linear equations for three unknowns $A_i(\varepsilon)$. Thus, by using Cramer's rule, we get an explicit formula for each $A_i$ in terms of the values $x_k$, $f(x_k)$, and $f(x_k - \varepsilon)$. Since the function $f(z)$ is smooth (differentiable), these expressions are differentiable too. Thus, we can differentiate both sides of the formula (9) with respect to $\varepsilon$. After taking $\varepsilon = 0$, we get the following differential equation $f'(z) = B_0 + B_1 \cdot z + B_2 \cdot f(z)$, where we denoted $B_i \stackrel{\text{def}}{=} A_i'(0)$. For $B_2 = 0$, we get $f'(z) = B_0 + B_1(z)$, so $f(z)$ is a quadratic function.

Let us show that the case $B_2 \neq 0$ is not possible. Indeed, in this case, by moving all the terms containing $f$ to the left-hand side, we get $f'(z) - B_2 \cdot f(z) = B_0 + B_1 \cdot z$. Thus, for the auxiliary function $F(z) \stackrel{\text{def}}{=} \exp(-B_2 \cdot z) \cdot f(z)$, we get

$$F'(z) = \exp(-B_2 \cdot z) \cdot f'(z) - B_2 \cdot \exp(-B_2 \cdot z) \cdot f(z) =$$

$$\exp(-B_2 \cdot z) \cdot (f'(z) - B_2 \cdot f(z)) = \exp(-B_2 \cdot z) \cdot (B_0 + B_1 \cdot z).$$

Integrating both sides, we conclude that

$$F(z) = f(z) \cdot \exp(-B_2 \cdot z) = (c_0 + c_1 \cdot z) \cdot \exp(-B_2 \cdot z) + c_2$$

for some constants $c_i$, thus

$$f(z) = c_0 + c_1 \cdot z + c_2 \cdot \exp(B_2 \cdot z). \tag{10}$$

From scale-invariance for $c = 1$, we similarly get

$$\lambda \cdot z + f(\lambda \cdot z) = D + E \cdot (z + c' \cdot f(z))$$

for some values $D$, $E$, and $c'$ which are, in general, depending on $\lambda$. Thus,

$$f(\lambda \cdot z) = D_0(\lambda) + D_1(\lambda) \cdot z + D_2(\lambda) \cdot f(z) \tag{11}$$

for appropriate $D_i(\lambda)$. Similarly to the case of shift-invariance, we can conclude that the functions $D_i$ are differentiable. Thus, we can differentiate both sides of the formula (11) with respect to $\lambda$. After taking $\lambda = 1$, we get the following differential equation $x \cdot f'(z) = D_0 + D_1 \cdot z + D_2 \cdot f(z)$ for appropriate values $D_i$. Substituting the expression (10) with $B_2 \neq 0$ into this formula, we can see that this equation is not satisfied. Thus, the case $B_2 \neq 0$ is indeed not possible, so the only possible case is $B_2 = 0$ which leads to a quadratic function $f(z)$. The proposition is proven.

*Comment.* The general expression $\psi(z) = g_0 + g_1 \cdot z + g_2 \cdot z^2$ is very natural for a different reason as well: it can be viewed as keeping the first terms in the Taylor expansion of a general function $\psi(z)$.

**Case of general constraints.** For the case of probabilistic constraints, we used a linear combination of different functions $\psi(z)$. For the case of general constraints, it is natural to use a linear combination of combination operations. As we have mentioned in the proof of Proposition 2, scale-invariant combination operations have the form $\|a\|_p \stackrel{\text{def}}{=} \left(\sum\limits_{i=0}^{n} |a_i|^p\right)^{1/p}$. According to Proposition 3, it makes sense to use quadratic terms, i.e., $\|a\|_2$. Thus, it makes sense to consider the combination $\|a\|_1 + c \cdot \|a\|_2$ – which is exactly CLOT.

Another interpretation of CLOT is that we combine $\|a\|_1$ and $c \cdot \|a\|_2$ by using shift- and scaling-invariant combination rule – which is, according to Proposition 2, simply addition.

*Comments.*

- An interesting feature of CLOT – as opposed to EN – is that it is scale-invariant.
- Not only we got a justification of EN and CLOT, we also got an understanding of when we should use EN and when CLOT: for probabilistic constraints, it is more appropriate to use EN, while for general constraints, it is more appropriate to use CLOT.

# 6 Beyond EN and CLOT?

**Discussion.** What if 1-parametric families like EN and CLOT are not sufficient? In this case, we need to consider families

$$F = \{z + c_1 \cdot f_1(z) + \ldots + c_n \cdot f_m(z)\}_{c_1,\ldots,c_m}$$

with more parameters.

**Definition 10.**

- *We say that a family* $\{z + c_1 \cdot f_1(z) + \ldots + c_m \cdot f_m(z)\}_{c_1,\ldots,c_m}$ *is* scale-invariant *if for each* $c = (c_1, \ldots, c_m)$ *and* $\lambda$*, there exists a tuple* $c' = (c'_1, \ldots, c'_m)$ *for which the re-scaled function*

$$\lambda \cdot z + c_1 \cdot f_1(\lambda \cdot z) + \ldots + c_m \cdot f_m(\lambda \cdot z)$$

*is constraint-equivalent to* $z + c'_1 \cdot f_1(z) + \ldots + c'_m \cdot f_m(z)$.

- *We say that a family* $\{z + c_1 \cdot f_1(z) + \ldots + c_m \cdot f_m(z)\}_{c_1,\ldots,c_m}$ *is* shift-invariant *if for each tuple* $c$ *and for each sufficiently small number* $\varepsilon$*, there exists a tuple* $c'$ *for which the shifted function*

$$z - \varepsilon + c_1 \cdot f_1(z - \varepsilon) + \ldots + c_m \cdot f_m(z - \varepsilon)$$

*is constraint-equivalent to* $z + c'_1 \cdot f_1(z) + \ldots + c'_m \cdot f_m(z)$.

**Proposition 4.** *A family* $\{z + c_1 \cdot f_1(z) + \ldots + c_m \cdot f_m(z)\}_{c_1,\ldots,c_m}$ *corresponding to a smooth functions* $f_i(z)$ *is scale- and shift-invariant if and only if all the functions* $f_i(z)$ *are polynomials of order* $\le m + 1$.

**Discussion.** So, if EN and CLOT are not sufficient, our recommendation is to use a constraint $\sum_{i=0}^{n} \psi(|a_i|)$ for some higher order polynomial $\psi(z)$.

**Proof of Proposition 4** is similar to the proof of Proposition 3, the only difference is that instead of a single differential equation, we will have a system of linear differential equations.

*Comment.* Similarly to the quadratic case, the resulting general expression $\psi(z) = g_0 + g_1 \cdot z + \ldots + a_{m+1} \cdot z^{m+1}$ can be viewed as keeping the first few terms in the Taylor expansion of a general function $\psi(z)$.

# Acknowledgments

This work was supported by the Institute of Geodesy, Leibniz University of Hannover. It was also supported in part by the US National Science Foundation grants 1623190 (A Model of Change for Preparing a New Generation for Professional Practice in Computer Science) and HRD-1242122 (Cyber-ShARE Center of Excellence). This paper was written when V. Kreinovich was visiting Leibniz University of Hannover.

# References

[1] J. Aczel and J. Dhombres, *Functional Equations in Several Variables*, Cambridge University Press, Cambridge, UK, 1989.

[2] M. E. Ahsen, N. Challapalli, and M. Vidyasagar, "Two new approaches to compressed sensing exhibiting both robust sparse recovery and the grouping effect", *Journal of Machine Learning Research*, 2017, Vol. 18, pp. 1–24.

[3] K. Autchariyapanitkul, O. Kosheleva, V. Kreinovich, and S. Sriboonchitta, "Quantum econometrics: how to explain its quantitative successes and how the resulting formulas are related to scale invariance, entropy, and fuzziness", In: V.-N. Huynh, M. Inuiguchi, D.-H. Tran, and Th. Denoeux (eds.), *Proceedings of the International Symposium on Integrated Uncertainty in Knowledge Modelling and Decision Making IUKM'2018*, Hanoi, Vietnam, March 13–15, 2018.

[4] A. E. Hoerl and R. W. Kennard, "Ridge regression: Biased estimation for nonorthogonal problems", *Technometrics*, 1970, Vol. 12, No. 1, pp. 55–67.

[5] E. T. Jaynes and G. L. Bretthorst, *Probability Theory: The Logic of Science*, Cambridge University Press, Cambridge, UK, 2003.

[6] B. Kargoll, M. Omidalizarandi, I. Loth, J.-A. Paffenholz, and H. Alkhatib, "An iteratively reweighted least-squares approach to adaptive robust adjustment of parameters in linear regression models with autoregressive and t-distributed deviations", *Journal of Geodesy*, 2018, Vol. 92, No. 3, pp. 271–297.

[7] D. J. Sheskin, *Handbook of Parametric and Nonparametric Statistical Procedures*, Chapman and Hall/CRC, Boca Raton, Florida, 2011.

[8] R. Tibshirani, "Regression shrinkage and selection via the lasso", *Journal of the Royal Statistical Society*, 1996, Vol. 58, No. 1, pp. 267–288.

[9] A. N. Tikhonov, "On the stability of inverse problems", *Doklady Akademii Nauk SSSR*, 1943, Vol. 39, No. 5, pp. 195–198.

[10] A. N. Tikhonov and V. Y. Arsenin, *Solutions of Ill-Posed Problems*, Winston and Sons, Washington, DC, 1977.

[11] H. Zou and T. Hastie, "Regularization and variable selection via the elastic net", *Journal of the Royal Statistical Society B*, 2005, Vol. 67, pp. 301–320.