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
Hannover, Germany
alkhatib@gih.uni-hannover.de
neumann@gih.uni-hannover.de
[2]University of Texas at El Paso
El Paso, Texas 79968, USA, vladik@utep.edu
[3]International University – VNU HCMC
Ho Chi Minh City, Vietnam, lvchon@hcmiu.edu.vn

Need for Regularization
Currently Used . . .
Why: Remaining . . .
Probabilistic . . .
General Regularization
Scale-Invariance: . . .
Shift-Invariance: . . .
Why LASSO
Beyond EN and CLOT?

1. Need for Solving 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.

2. Need for Inverse Problem (cont-d)

- For example, often, we have a linear model

$$y = a_0 + \sum_{i=1}^{n} a_i \cdot x_i.$$

- The parameters a_i need to be experimentally determined.
- In general, we need to determine the parameters of the model based on the measurement results.
- This problem is known as the *inverse problem*.
- To actually find the parameters, we can use, e.g., the Maximum Likelihood method.

Need for Regularization
Currently Used . . .
Why: Remaining . . .
Probabilistic . . .
General Regularization
Scale-Invariance: . . .
Shift-Invariance: . . .
Why LASSO
Beyond EN and CLOT?

3. Need for Inverse Problem (cont-d)

- For example:
  - when the errors are normally distributed,
  - the Maximum Likelihood procedure results in the usual Least Squares estimates.
- 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

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

Need for Regularization
Currently Used . . .
Why: Remaining . . .
Probabilistic . . .
General Regularization
Scale-Invariance: . . .
Shift-Invariance: . . .
Why LASSO
Beyond EN and CLOT?

4. Need for Regularization

- In some practical situations:
  - based on the measurement results,
  - we can determine all the model's parameters with reasonably accuracy.
- Often, several different combinations of parameters are consistent with all the measurement results.
- Such inverse problems are called *ill-defined*.
- E.g., in dynamical systems, the observations provide a smoothed picture of the system's dynamics.
- For example, we can be tracing the motion of a mechanical system caused by an external force.

Need for Regularization
Currently Used . . .
Why: Remaining . . .
Probabilistic . . .
General Regularization
Scale-Invariance: . . .
Shift-Invariance: . . .
Why LASSO
Beyond EN and CLOT?

5. Need for Regularization (cont-d)

- 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.

Need for Regularization
Currently Used . . .
Why: Remaining . . .
Probabilistic . . .
General Regularization
Scale-Invariance: . . .
Shift-Invariance: . . .
Why LASSO
Beyond EN and CLOT?

6. Need for Regularization (cont-d)

- In such situations:
  - the only way to narrow down the set of possible solution
  - is to take into account some general a priori information.
- For example, for forces, we may know – e.g., from experts – the upper bound.
- The use of such a priori information is known as *regularization*.

Need for Regularization
Currently Used . . .
Why: Remaining . . .
Probabilistic . . .
General Regularization
Scale-Invariance: . . .
Shift-Invariance: . . .
Why LASSO
Beyond EN and CLOT?

7. Which Regularizations Are Currently Used

- There are many possible regularizations.
- Many of them have been tried.
- 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.
- LASSO is short of Least Absolute Shrinkage and Selection Operator.
- We require that the sum of the absolute values $\|a\|_1 \stackrel{\text{def}}{=} \sum_{i=0}^{n} |a_i|$ be bounded by some number.

Need for Regularization
Currently Used . . .
Why: Remaining . . .
Probabilistic . . .
General Regularization
Scale-Invariance: . . .
Shift-Invariance: . . .
Why LASSO
Beyond EN and CLOT?

8. Currently Used Regularizations (cont-d)

- Another widely used method is a *ridge regression* method, in which we limit the sum of the squares $S \stackrel{\text{def}}{=} \sum_{i=0}^{n} a_i^2$.
- This is equivalent to bounding its square root

$$\|a\|_2 \stackrel{\text{def}}{=} \sqrt{S}.$$

- Very promising are also:
  - the *Elastic Net* (EN) method, in which we limit a linear combination $\|a\|_1 + c \cdot S$, and
  - the *Combined L-One and Two* (CLOT) method in which we limit a linear combination $\|a\|_1 + c \cdot \|a\|_2$.

Need for Regularization
Currently Used . . .
Why: Remaining . . .
Probabilistic . . .
General Regularization
Scale-Invariance: . . .
Shift-Invariance: . . .
Why LASSO
Beyond EN and CLOT?

9. Why: Remaining Question and What We Do in This Talk

- The above empirical facts prompt a natural question: why the above regularization techniques work the best?
- We show that the efficiency of these methods can be explained by the natural invariance requirements.

Need for Regularization
Currently Used . . .
Why: Remaining . . .
Probabilistic . . .
General Regularization
Scale-Invariance: . . .
Shift-Invariance: . . .
Why LASSO
Beyond EN and CLOT?

10. 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.
- So, it is based on the frequencies (= probabilities) with which different values have been observed in the past.
- In this talk, we consider both types of regularization.

Need for Regularization
Currently Used . . .
Why: Remaining . . .
Probabilistic . . .
General Regularization
Scale-Invariance: . . .
Shift-Invariance: . . .
Why LASSO
Beyond EN and CLOT?

11. 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 other words, the original information is invariant w.r.t. all possible permutations of parameters.
- Then, the resulting joint distribution should also be invariant with respect to all the permutations.
- This implies, in particular, that all the marginal distributions are the same.

Need for Regularization
Currently Used . . .
Why: Remaining . . .
Probabilistic . . .
General Regularization
Scale-Invariance: . . .
Shift-Invariance: . . .
Why LASSO
Beyond EN and CLOT?

12. Probabilistic Regularization (cont-d)

- Similarly, in general, we do not have a priori reasons to prefer positive or negative values of each a_i.
- So, 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.
- So, we should have $\rho(-a) = \rho(a)$, i.e., $\rho(a) = \rho(|a|)$.
- 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.

Need for Regularization
Currently Used . . .
Why: Remaining . . .
Probabilistic . . .
General Regularization
Scale-Invariance: . . .
Shift-Invariance: . . .
Why LASSO
Beyond EN and CLOT?

13. Probabilistic Regularization (cont-d)

- According to these ideas, we should not 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 independence.
- Due to independence, the joint distribution of n variables a_i is $\rho(a_0, a_1, \ldots, a_n) = \prod_{i=0}^{n} \rho(|a_i|)$.
- In applications, it is usually assumed that events with very small probability cannot happen.

Need for Regularization
Currently Used . . .
Why: Remaining . . .
Probabilistic . . .
General Regularization
Scale-Invariance: . . .
Shift-Invariance: . . .
Why LASSO
Beyond EN and CLOT?

14. Probabilistic Regularization (cont-d)

- This is the basis for all statistical tests.
- *Example:* assume that the distribution is normal with given mean and standard deviation.
- Assume also that the probability that this distribution will lead to the observed data is very small.
- E.g., 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

$$\rho(a_0, a_1, \ldots, a_n) = \prod_{i=0}^{n} \rho(|a_i|) \geq t_0.$$

Need for Regularization
Currently Used . . .
Why: Remaining . . .
Probabilistic . . .
General Regularization
Scale-Invariance: . . .
Shift-Invariance: . . .
Why LASSO
Beyond EN and CLOT?

15. Probabilistic Regularization (cont-d)

- By taking logarithms of both sides and changing signs, we get an equivalent inequality

$$\sum_{i=0}^{n} \psi(|a_i|) \le p_0, \text{ where } \psi(z) \stackrel{\text{def}}{=} -\ln(\rho(z)) \text{ and } p_0 \stackrel{\text{def}}{=} -\ln(t_0).$$

- The sign is changed for convenience:
  - 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,
  - and we decrease one of the absolute values,
  - the result should also satisfy the same inequality.

Need for Regularization
Currently Used . . .
Why: Remaining . . .
Probabilistic . . .
General Regularization
Scale-Invariance: . . .
Shift-Invariance: . . .
Why LASSO
Beyond EN and CLOT?

16. Probabilistic Regularization (cont-d)

- So, the function $\psi(z)$ must be increasing.
- We want to find the minimum of the usual least squares (or similar) criterion under this constraint.
- The minimum is attained:
  - either when in the constraint, we have strict inequality
  - or when we have equality.
- If we have a strict inequality, then we get a local minimum.
- For convex criteria like least squares, there is only one local minimum which is also global.
- So, this means that we have the solution of the original constraint-free problem.

Need for Regularization
Currently Used . . .
Why: Remaining . . .
Probabilistic . . .
General Regularization
Scale-Invariance: . . .
Shift-Invariance: . . .
Why LASSO
Beyond EN and CLOT?

17. Probabilistic Regularization (cont-d)

- However, we consider situations in which this straightforward approach does not work.
- Thus, we conclude that the minimum under constraint is attained when we have the equality:

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

- 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.

Need for Regularization
Currently Used . . .
Why: Remaining . . .
Probabilistic . . .
General Regularization
Scale-Invariance: . . .
Shift-Invariance: . . .
Why LASSO
Beyond EN and CLOT?

18. Probabilistic Regularization (cont-d)

- Thus, we arrive at the following definition.
- *By a* probabilistic constraint, *we mean the following constraint:* $\sum_{i=0}^{n} \psi(|a_i|) = p_0$.
- Here, $\psi(z)$ is a continuous increasing function, and p_0 is a number.

Need for Regularization
Currently Used . . .
Why: Remaining . . .
Probabilistic . . .
General Regularization
Scale-Invariance: . . .
Shift-Invariance: . . .
Why LASSO
Beyond EN and CLOT?

19. 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.
- 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.

Need for Regularization
Currently Used . . .
Why: Remaining . . .
Probabilistic . . .
General Regularization
Scale-Invariance: . . .
Shift-Invariance: . . .
Why LASSO
Beyond EN and CLOT?

20. General Regularization (cont-d)

- 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 $a_1 = 0$, it is reasonable to say that:
  - the degree of impossibility of the pair $(a_0, 0)$ is the same as
  - the degree of impossibility of a_0, i.e., equal to $|a_0|$:

$$|a_0| * 0 = |a_0|.$$

Need for Regularization
Currently Used . . .
Why: Remaining . . .
Probabilistic . . .
General Regularization
Scale-Invariance: . . .
Shift-Invariance: . . .
Why LASSO
Beyond EN and CLOT?

21. General Regularization (cont-d)

- If 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).

Need for Regularization
Currently Used . . .
Why: Remaining . . .
Probabilistic . . .
General Regularization
Scale-Invariance: . . .
Shift-Invariance: . . .
Why LASSO
Beyond EN and CLOT?

22. General Regularization (cont-d)

- We can combine again and again, 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.
- So, we should have $a * b = b * a$ (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:

$$|a_0| * |a_1| * \ldots * |a_n| \le t_0.$$

- Thus, we arrive at the following definitions.

Need for Regularization
Currently Used . . .
Why: Remaining . . .
Probabilistic . . .
General Regularization
Scale-Invariance: . . .
Shift-Invariance: . . .
Why LASSO
Beyond EN and CLOT?

23. General Regularization (cont-d)

- *By a* combination operation, *we mean a function* $* : \mathbb{R} \times \mathbb{R} \to \mathbb{R}$ *which is:*
  - *commutative,*
  - *associative,*
  - *has the property* $a * 0 = a$ *and*
  - monotonic *in the sense that if* $a < a'$, *then*

$$a * b < a' * b.$$

- *By a* general constraint, *we means a constraint of the type* $|a_0| * |a_1| * \ldots * |a_n| \le t_0$.
- Here, $*$ is a combination operation, and $t_0 > 0$ is a number.

Need for Regularization
Currently Used . . .
Why: Remaining . . .
Probabilistic . . .
General Regularization
Scale-Invariance: . . .
Shift-Invariance: . . .
Why LASSO
Beyond EN and CLOT?

24. 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 λ times smaller,
  - then all the numerical values get multiplied by λ:

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

Need for Regularization
Currently Used . . .
Why: Remaining . . .
Probabilistic . . .
General Regularization
Scale-Invariance: . . .
Shift-Invariance: . . .
Why LASSO
Beyond EN and CLOT?

25. Scale-Invariance (cont-d)

- Similarly, we can change the original measuring units for y to a new unit which is λ times smaller.
- Then, all the coefficients a_i in the dependence $y = a_0 + \ldots + a_i \cdot x_i + \ldots$ will also change: $a_i \to \lambda \cdot a_i$.

Need for Regularization
Currently Used . . .
Why: Remaining . . .
Probabilistic . . .
General Regularization
Scale-Invariance: . . .
Shift-Invariance: . . .
Why LASSO
Beyond EN and CLOT?

26. Scale-Invariance: Case of Probabilistic Constraints

- It is reasonable to require that the 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.
- However, overall, the constraint should remain the same.
- Thus, we arrive at the following definition.
- *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*

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

Need for Regularization
Currently Used . . .
Why: Remaining . . .
Probabilistic . . .
General Regularization
Scale-Invariance: . . .
Shift-Invariance: . . .
Why LASSO
Beyond EN and CLOT?

27. 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.
- *We say that a constraint corr. to $*$ is* scale-invariance *if for every a, b, and λ, we have*

$$(\lambda \cdot a) * (\lambda \cdot b) = \lambda \cdot (a * b).$$

Need for Regularization
Currently Used . . .
Why: Remaining . . .
Probabilistic . . .
General Regularization
Scale-Invariance: . . .
Shift-Invariance: . . .
Why LASSO
Beyond EN and CLOT?

28. Scale-Invariance (cont-d)

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

$$|\lambda\cdot a_0|*|\lambda\cdot a_1|*\ldots*|\lambda\cdot a_n| = \lambda*(|a_0|*|a_1|*\ldots*|a_n|);\ \text{so}$$

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

Need for Regularization
Currently Used . . .
Why: Remaining . . .
Probabilistic . . .
General Regularization
Scale-Invariance: . . .
Shift-Invariance: . . .
Why LASSO
Beyond EN and CLOT?

29. 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.

Need for Regularization
Currently Used . . .
Why: Remaining . . .
Probabilistic . . .
General Regularization
Scale-Invariance: . . .
Shift-Invariance: . . .
Why LASSO
Beyond EN and CLOT?

30. Shift-Invariance (cont-d)

- 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.
- These values are small, but still non-zero.
- In such situations, small deviations from 0 are OK, they do not necessarily indicate that something is wrong.
- To deal with this phenomenon, we can:
  - explicitly subtract an appropriate small tolerance level $\varepsilon > 0$
  - from the absolute values of all the coefficients.
- In other words, we can replace the original values $|a_i|$ with the new values $|a_i| - \varepsilon$.

Need for Regularization
Currently Used . . .
Why: Remaining . . .
Probabilistic . . .
General Regularization
Scale-Invariance: . . .
Shift-Invariance: . . .
Why LASSO
Beyond EN and CLOT?

31. Shift-Invariance (cont-d)

- This will explicitly take into account that:
  - deviations smaller that this tolerance level are OK, and
  - only deviations above this level are problematic.
- It is reasonable to require that the corresponding constraints do not change under this shift $|a| \to |a| - \varepsilon$.

Need for Regularization
Currently Used . . .
Why: Remaining . . .
Probabilistic . . .
General Regularization
Scale-Invariance: . . .
Shift-Invariance: . . .
Why LASSO
Beyond EN and CLOT?

32. 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.
- However, overall, the constraint should remain the same.
- Thus, we arrive at the following definition.
- *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*

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

Need for Regularization
Currently Used . . .
Why: Remaining . . .
Probabilistic . . .
General Regularization
Scale-Invariance: . . .
Shift-Invariance: . . .
Why LASSO
Beyond EN and CLOT?

33. 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, ε' may be different from ε, 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.

Need for Regularization
Currently Used . . .
Why: Remaining . . .
Probabilistic . . .
General Regularization
Scale-Invariance: . . .
Shift-Invariance: . . .
Why LASSO
Beyond EN and CLOT?

34. Shift-Invariance (cont-d)

- *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 $a_0, \ldots, a_n > 0$, we have*

$$(a_0 - \varepsilon) * \ldots * (a_n - \varepsilon) = (a_0 * \ldots * a_n) - \varepsilon'.$$

- 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 $\varepsilon > 0$:

$$|a_0| * |a_1| * \ldots * |a_n| = t_0 \Leftrightarrow$$

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

Need for Regularization
Currently Used . . .
Why: Remaining . . .
Probabilistic . . .
General Regularization
Scale-Invariance: . . .
Shift-Invariance: . . .
Why LASSO
Beyond EN and CLOT?

35. Why LASSO

- Let us show that for both types of constraints, natural invariance requirements lead to LASSO formulas.
- **Proposition 1.** *Probabilistic constraints corr. to $\psi(x)$ are shift- and scale-invariant if and only $\psi(x) = k \cdot x + \ell$.*
- For a linear function, the constraint $\sum_{i=0}^{n} \psi(|a_i|) = p_0$ is equivalent to the LASSO constraint

$$\sum_{i=1}^{n} |a_i| = t_0', \text{ with } t_0' \stackrel{\text{def}}{=} (t_0 - \ell)/k.$$

- Thus, we explained why probabilistic constraints should be LASSO constraints.

Need for Regularization
Currently Used . . .
Why: Remaining . . .
Probabilistic . . .
General Regularization
Scale-Invariance: . . .
Shift-Invariance: . . .
Why LASSO
Beyond EN and CLOT?

36. Why LASSO (cont-d)

- **Proposition 2.** *General constraints corr. to $*$ are shift- and scale-invariant if and only if $a * b = a + b$.*
- For addition, the corresponding constraint $\sum_{i=0}^{n} |a_i| = t_0$ is exactly the LASSO constraint.
- Thus, we also explained why general constraints should be LASSO constraints.

Need for Regularization
Currently Used . . .
Why: Remaining . . .
Probabilistic . . .
General Regularization
Scale-Invariance: . . .
Shift-Invariance: . . .
Why LASSO
Beyond EN and CLOT?

37. Need to Go Beyond LASSO

- We showed that:
  - if we need to select a single method,
  - 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,
  - a natural idea is to select a *family* of methods.
- So, 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.

Need for Regularization
Currently Used . . .
Why: Remaining . . .
Probabilistic . . .
General Regularization
Scale-Invariance: . . .
Shift-Invariance: . . .
Why LASSO
Beyond EN and CLOT?

38. 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

$$\psi(z) = c_0 + c_1 \cdot z + c_2 \cdot f(z).$$

- Constraints related to this family are equivalent to using the functions $\psi(z) = z + c \cdot f(z)$ for some $f(z)$.
- Which family – i.e., which function $f(z)$ – should we choose?

Need for Regularization
Currently Used . . .
Why: Remaining . . .
Probabilistic . . .
General Regularization
Scale-Invariance: . . .
Shift-Invariance: . . .
Why LASSO
Beyond EN and CLOT?

39. Probabilistic Case (cont-d)

- A natural idea is to again use scale-invariance and shift-invariance.
- *We say that functions* $\psi_1(z)$ *and* $\psi_2(z)$ *are* constraint-equivalent *($\psi_1 \sim \psi_2$) if:*
  - *for each* n *and for each* c_1*, there exists a value* c_2 *such that* $\sum\limits_{i=0}^{n} \psi_1(a_i) = c_1 \Leftrightarrow \sum\limits_{i=0}^{n} \psi_2(a_i) = c_2$,
  - *and for each* n *and for each* c_2*, there exists a value* c_1 *such that* $\sum\limits_{i=0}^{n} \psi_2(a_i) = c_2 \Leftrightarrow \sum\limits_{i=0}^{n} \psi_1(a_i) = c_1$.
- *We say that a family* $\{z + c \cdot f(z)\}_c$ *is* scale-invariant *if for each* c *and* λ*, there exists a value* c' *for which the*

$$\lambda \cdot z + c \cdot f(\lambda \cdot z) \sim z + c' \cdot f(z).$$

Need for Regularization
Currently Used . . .
Why: Remaining . . .
Probabilistic . . .
General Regularization
Scale-Invariance: . . .
Shift-Invariance: . . .
Why LASSO
Beyond EN and CLOT?

40. Probabilistic Case (cont-d)

- *We say that* $\{z + c \cdot f(z)\}_c$ *is* shift-invariant *if:*
  - *for each* c *and for each sufficiently small number* ε,
  - *there exists a value* c' *for which*

$$z - \varepsilon + c \cdot f(z - \varepsilon) \sim z + c' \cdot f(z).$$

- **Proposition 3.** *For smooth* $f(z)$, $\{z + c \cdot f(z)\}_c$ *is scale- and shift-invariant* $\Leftrightarrow$ $f(z)$ *is quadratic.*
- 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.

Need for Regularization
Currently Used . . .
Why: Remaining . . .
Probabilistic . . .
General Regularization
Scale-Invariance: . . .
Shift-Invariance: . . .
Why LASSO
Beyond EN and CLOT?

41. Probabilistic Case (cont-d)

- 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.
- Namely, it can be viewed as keeping the first terms in the Taylor expansion of a general function $\psi(z)$.

Need for Regularization
Currently Used . . .
Why: Remaining . . .
Probabilistic . . .
General Regularization
Scale-Invariance: . . .
Shift-Invariance: . . .
Why LASSO
Beyond EN and CLOT?

42. 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 mention in the proof of Proposition 2, scale-invariant combination operations have the form

$$\|a\|_p \stackrel{\text{def}}{=} \left(\sum_{i=0}^{n} |a_i|^p\right)^{1/p}.$$

- 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.

Need for Regularization
Currently Used . . .
Why: Remaining . . .
Probabilistic . . .
General Regularization
Scale-Invariance: . . .
Shift-Invariance: . . .
Why LASSO
Beyond EN and CLOT?

43. Case of General Constraints (cont-d)

- Another interpretation of CLOT is that:
  - we combine $\|a\|_1$ and $c \cdot \|a\|_2$ by using shift- and scaling-invariant combination rule,
  - since, according to Proposition 2, such a rule is simply addition.
- An interesting feature of CLOT – as opposed to EN – is that it is scale-invariant.
- 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.

Need for Regularization
Currently Used . . .
Why: Remaining . . .
Probabilistic . . .
General Regularization
Scale-Invariance: . . .
Shift-Invariance: . . .
Why LASSO
Beyond EN and CLOT?

44. Beyond EN and CLOT?

- What if 1-parametric families like EN and CLOT are not sufficient?
- In this case, we need to consider families with more parameters $\{z + c_1 \cdot f_1(z) + \ldots + c_n \cdot f_m(z)\}_{c_1,\ldots,c_m}$.
- *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* λ,
  - *there exists a tuple* $c' = (c'_1, \ldots, c'_m)$ *for which*

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

$$z + c'_1 \cdot f_1(z) + \ldots + c'_m \cdot f_m(z).$$

Need for Regularization
Currently Used . . .
Why: Remaining . . .
Probabilistic . . .
General Regularization
Scale-Invariance: . . .
Shift-Invariance: . . .
Why LASSO
Beyond EN and CLOT?

45. Beyond EN and CLOT (cont-d)

- *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* ε*,*
  - *there exists a tuple* c' *for which*

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

$$z+c'_1\cdot f_1(z)+\ldots+c'_m\cdot f_m(z).$$

- Let us consider smooth $f_i(z)$.
- **Proposition 4.** $\{z+c_1\cdot f_1(z)+\ldots+c_m\cdot f_m(z)\}_{c_1,\ldots,c_m}$ *is scale- and shift-invariant* $\Leftrightarrow$ *all* $f_i(z)$ *are polynomials.*
- These polynomials must be of order $\le m+1$.

Need for Regularization
Currently Used . . .
Why: Remaining . . .
Probabilistic . . .
General Regularization
Scale-Invariance: . . .
Shift-Invariance: . . .
Why LASSO
Beyond EN and CLOT?

46. Beyond EN and CLOT (cont-d)

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

Need for Regularization
Currently Used . . .
Why: Remaining . . .
Probabilistic . . .
General Regularization
Scale-Invariance: . . .
Shift-Invariance: . . .
Why LASSO
Beyond EN and CLOT?

47. 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.

Need for Regularization
Currently Used . . .
Why: Remaining . . .
Probabilistic . . .
General Regularization
Scale-Invariance: . . .
Shift-Invariance: . . .
Why LASSO
Beyond EN and CLOT?

48. 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

$$\psi(\lambda \cdot a) + \psi(\lambda \cdot b) = \psi(\lambda \cdot c) + \psi(0).$$

- Let's 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 $\Psi(\lambda \cdot a) + \Psi(\lambda \cdot b) = \Psi(\lambda \cdot c)$.
- Let $\Psi^{-1}(z)$ denote the inverse function.

Need for Regularization
Currently Used . . .
Why: Remaining . . .
Probabilistic . . .
General Regularization
Scale-Invariance: . . .
Shift-Invariance: . . .
Why LASSO
Beyond EN and CLOT?

49. Proof of Proposition 1 (cont-d)

- Then, for the mapping $f(z) \stackrel{\text{def}}{=} \Psi(\lambda \cdot \Psi^{-1}(z))$ that transforms $z = \Psi(a)$ into $f(z) = \Psi(\lambda \cdot a)$:
  - 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 $f(z) = c \cdot z$.
- Since $z = \Psi(a)$ and $f(z) = \Psi(\lambda \cdot a)$:
  - for every λ,
  - there exists a value c (which, in general, depends on λ) 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 α.

Need for Regularization
Currently Used . . .
Why: Remaining . . .
Probabilistic . . .
General Regularization
Scale-Invariance: . . .
Shift-Invariance: . . .
Why LASSO
Beyond EN and CLOT?

50. Proof of Proposition 1 (cont-d)

- 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 $\delta > 0$:

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

Need for Regularization
Currently Used . . .
Why: Remaining . . .
Probabilistic . . .
General Regularization
Scale-Invariance: . . .
Shift-Invariance: . . .
Why LASSO
Beyond EN and CLOT?

51. Proof of Proposition 1 (cont-d)

- Let's substitute $\psi(a) = A \cdot a^\alpha + B$, subtract $2B$ from both sides, and dividing both equalities by A, then:
  - 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 δ, 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 δ:
  - the left-hand side term and the first term in the right-hand side start with linear term δ, and
  - the terms $\delta^\alpha \gg \delta$ is not compensated.

Need for Regularization
Currently Used . . .
Why: Remaining . . .
Probabilistic . . .
General Regularization
Scale-Invariance: . . .
Shift-Invariance: . . .
Why LASSO
Beyond EN and CLOT?

52. Proof of Proposition 1 (cont-d)

- If $\alpha > 1$, then by equating terms linear in δ in the corresponding expansions:
  - we get $2\alpha \cdot \delta$ in the left-hand side and
  - we get $\alpha \cdot (2^{1/\alpha})^{\alpha-1} \cdot \delta = 2^{1-1/\alpha} \cdot \alpha \cdot \delta$ 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.

Need for Regularization
Currently Used . . .
Why: Remaining . . .
Probabilistic . . .
General Regularization
Scale-Invariance: . . .
Shift-Invariance: . . .
Why LASSO
Beyond EN and CLOT?

53. Proof of Proposition 2

- It is known that every scale-invariant combination operation has the form

$$a * b = (a^\alpha + b^\alpha)^{1/\alpha} \text{ or } a * b = \max(a, b).$$

- 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.

Need for Regularization
Currently Used . . .
Why: Remaining . . .
Probabilistic . . .
General Regularization
Scale-Invariance: . . .
Shift-Invariance: . . .
Why LASSO
Beyond EN and CLOT?

54. Proof of Proposition 3

- Similarly to the proof of Proposition 1, from the shift-invariance, for $c = 1$, we conclude that

$$z - \varepsilon + f(z - \varepsilon) = A + B \cdot (z + c' \cdot f(z).$$

- Here, the values A, B, and c', in general, depend on ε, so:

$$f(z - \varepsilon) = A_0(\varepsilon) + A_1(\varepsilon) \cdot z + A_2(\varepsilon) \cdot f(z).$$

- Here, $A_0(\varepsilon) \stackrel{\text{def}}{=} A + \varepsilon$, $A_1(\varepsilon) \stackrel{\text{def}}{=} B - 1$, and $A_2(\varepsilon) \stackrel{\text{def}}{=} B \cdot c'$.
- Let us consider three different values x_k $(k = 1, 2, 3)$.
- Then, 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

$$x_k, f(x_k), \text{ and } f(x_k - \varepsilon).$$

Need for Regularization
Currently Used . . .
Why: Remaining . . .
Probabilistic . . .
General Regularization
Scale-Invariance: . . .
Shift-Invariance: . . .
Why LASSO
Beyond EN and CLOT?

55. Proof of Proposition 3 (cont-d)

- Since the function $f(z)$ is smooth (differentiable), these expressions are differentiable too.
- Thus, we can differentiate both sides of the above formula with respect to ε.
- After taking $\varepsilon = 0$, we get $f'(z) = B_0 + B_1 \cdot z + B_2 \cdot f(z)$, where $B_i \stackrel{\text{def}}{=} A_i'(0)$.
- For $B_2 = 0$, we get $f'(z) = B_0 + B_1 \cdot z$, 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

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

Need for Regularization
Currently Used . . .
Why: Remaining . . .
Probabilistic . . .
General Regularization
Scale-Invariance: . . .
Shift-Invariance: . . .
Why LASSO
Beyond EN and CLOT?

56. Proof of Proposition 3 (cont-d)

- Thus, for the auxiliary function

$$F(z) \stackrel{\text{def}}{=} \exp(-B_2 \cdot z) \cdot f(z), \text{ we get}$$

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

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

- Integrating both sides, we conclude that

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

- Thus, $f(z) = c_0 + c_1 \cdot z + c_2 \cdot \exp(B_2 \cdot z)$.
- From scale-invariance for $c = 1$, we similarly get

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

- Here, the values D, E, and c' which are, in general, depending on λ; thus:

$$f(\lambda \cdot z) = D_0(\lambda) + D_1(\lambda) \cdot z + D_2(\lambda) \cdot f(z) \text{ for some } D_i(\lambda).$$

Need for Regularization
Currently Used . . .
Why: Remaining . . .
Probabilistic . . .
General Regularization
Scale-Invariance: . . .
Shift-Invariance: . . .
Why LASSO
Beyond EN and CLOT?

57. Proof of Proposition 3 (cont-d)

- 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 above formula with respect to λ.
- After taking $\lambda = 1$, we get:

$$z \cdot f'(z) = D_0 + D_1 \cdot z + D_2 \cdot f(z) \text{ for some } D_i.$$

- Substituting the expression 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.

Need for Regularization
Currently Used . . .
Why: Remaining . . .
Probabilistic . . .
General Regularization
Scale-Invariance: . . .
Shift-Invariance: . . .
Why LASSO
Beyond EN and CLOT?

58. Proof of Proposition 4

- This proof 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.

Need for Regularization
Currently Used . . .
Why: Remaining . . .
Probabilistic . . .
General Regularization
Scale-Invariance: . . .
Shift-Invariance: . . .
Why LASSO
Beyond EN and CLOT?