Decision Making Under Uncertainty with Special Emphasis on Geosciences and Education

Laxman Bokati
Computational Science Program
University of Texas at El Paso
El Paso, TX 79968, USA
lbokati@miners.utep.edu

1. One of the main objective of science

- One of the main objectives of science is to help people make good decisions.
- Because of the ubiquity and importance of decision making, it has been the subject of intensive research.
- This research can be roughly divided into two categories:
  - analysis of how a rational person *should* make a decision, and
  - analysis of people *actually* make decisions.
- The main objective of this dissertation is:
  - to expand on both research categories,
  - with the ultimate objective of providing the corresponding practical recommendations.

2. Decision making under uncertainty

- One of the major difficulties in decision making is that usually, we do not have full information:
  - about the situation and
  - about possible consequences of our decisions.
- The larger this uncertainty, the more difficult it is to make decisions.

3. Selecting case studies

- The larger the uncertainty, the more difficult it is to make a right decision.
- So, a natural case study for new decision making techniques are situations:
  - where decisions are the most difficult – i.e.,
  - where there is the largest amount of uncertainty.
- In most practical problems:
  - even if we do not have the full information about the situation, i.e., even if we do not know the values of some quantities,
  - we can, in principle, measure these values and get a better understanding of the situation.
- For example, we often do not have enough information about the weather – i.e., about the current values of temperature, etc.

4. Selecting case studies (cont-d)

- However, we can, if needed, measure these values are thus, decrease the uncertainty.
- There is, however, an application area where such measurements are not possible: namely, geosciences.
- For example, oil companies would like to know whether it makes sense to start digging an oil well at a prospective location.
- When we make this decision, we do not have full information on what is happening at the corresponding depths.
- In principle, it is possible to perform direct measurements that can determine this information; however:
  - this measurement would require, in effect, digging a deep well and placing instruments down below, while
  - the whole purpose of this analysis is to decide whether it is worth investing significant resources in this possible well.

5. Selecting case studies (cont-d)

- Because of this, geosciences are among the most challenging areas for decision making.
- So, we have selected geosciences as the main case study for our results.
- Another area when measurements are difficult is education.
- In education, we can gauge the observable results of teaching but not the internal process that lead to more or less successful teaching.
- This is similar to geosciences, where:
  - we can measure the seismic waves reaching the surface, but
  - we cannot directly measure the processes leading to these waves.

6. Structure of the dissertation

- In line with all this, first we explain the general structure of the dissertation.
- Then, we will provide a more detailed description of several results.

7. Part I: Introduction

- First, we provide a brief reminder of decision theory – that explains how rational people should make decisions.
- The main ideas related to (rational) individual decision making are described in Chapter 2.
- Chapter 3 covers the general ideas behind (rational) group decision making.
- In general, the corresponding formulas are known.
- However, this chapter already contains some new material – namely, we provide a new simplified derivation of these formulas.
- Finally, Chapter 4 explains how we can control group decision making by modifying the proposed options.
- This chapter contains both an empirical dependence – and our explanation of this dependence.
- This is the first of the chapters that contains completely new results.

8. Part II: How people actually make decisions

- In this part:
  - we analyze how people actually make decisions – in general and, in particular, in economy-related situations, and
  - we explain why people's actual decisions differ from recommendations of decision theory.
- This part covers all possible deviations of actual decisions from the ideal ones.
- In the ideal case:
  - first, we find the exact value of each item in each alternative,
  - then, we combine these values into exact values of each alternative,
  - third, we find future consequences of different actions, and preferences of other people, and
  - fourth, based on all this information, we select the optimal alternative.

9. Part II: How people actually make decisions (cont-d)

- In real life, human decision making deviates from the ideal on all there four stages.
- On the first stage, we have to base our decisions on incomplete, approximate knowledge:
  - either because information leading more accurate estimates are not available,
  - or because, while this information is available, there is not enough time to process this information.
- In such case:
  - instead of coming up with the exact values of each item,
  - people come up with approximate estimates,
  - i.e., in effect, bounds on possible values.
- As we show in Chapter 5, this explains the empirical fact that people's selling prices are usually higher than their buying prices.

10. Part II: How people actually make decisions (cont-d)

- This fact seems to contradict the basic economic ideas.
- Also, since the information is usually incomplete, different people come up with different prices for the same item.
- This explains the constant buying and selling, something that also seems to contradict the basic economic ideas; see Chapter 6.
- Instead (or in lieu of) eliciting the accurate values, people make decisions based on clusters containing the actual values.
- For example, they use the so-called 7 ± 2 approach; see Chapters 7 and 8.
- On the second stage, when people combine utility values, they use approximate processing techniques; see, e.g., Chapter 9.
- On the third stage, people use biased perceptions of the future time; see Chapter 10, resulting in non-optimal solutions (Chapter 11).

11. Part II: How people actually make decisions (cont-d)

- They also have a biased perception of other people's utility, which also leads to non-optimal solutions; see Chapter 12.
- Finally, on the fourth stage, instead of going for an optimal solution:
  - people use approximately optimal solutions (Chapter 13)
  - or even use heuristics instead of looking for optimal or approximately optimal solutions (Chapter 14 and 15).
- In most of these cases, there are known empirical formulas describing actual human behavior.
- In our analysis, we provide possible theoretical explanations for these formulas.

12. Part 3: applications to geosciences

- After this general description of human decision making, we focus on our main application area: geosciences.
- In geosciences, like in many other application areas, we encounter two types of situations.
- In some cases, we have a relatively small number of observations – only sufficiently many to estimate the values of a few parameters of the model.
- In such cases, it is desirable to come up with the most adequate few-parametric model.
- We analyze the corresponding problem of select an optimal model on two examples:
  - of spatial dependence (Chapter 16) and
  - of temporal dependence (Chapter 17).

13. Part 3: applications to geosciences (cont-d)

- As an example of a temporal dependence problem, we consider the problem of earthquake prediction.
- This is one of the most challenging and the most important geophysical problems.
- Specifically, we analyze the problem of selecting the most adequate probabilistic distribution of between-earthquakes time intervals.
- In other cases, we already have many observations covering many locations and many moments of time.
- In such cases, we can look for the best ways to extend this knowledge:
  - to other spatial locations (Chapter 18) and
  - to future moments of time (Chapter 19).
- As an example of extending knowledge to future moments of time – i.e., prediction – we deal with earthquakes triggering earthquakes.
- This is one of the least studied seismic phenomena.

14. Part IV: applications to teaching

- Our analysis cover all three related major questions:
  - what to teach (Chapters 20 and 21),
  - how to teach (Chapter 22), and
  - how to grade, i.e., how to gauge the results of teaching (Chapter 23).

15. Part V: applications to computing

- Most of these and other applications involve intensive computing.
- In the final Part V, we show that the above-analyzed ideas can be used in all aspects of computing:
  - in analyzing the simplest (linear) models (Chapter 24),
  - in analyzing more realistic non-linear models (Chapter 25), and even
  - in exploring perspective approaches to computing (Chapter 26).

16. Appendix

- In all these parts, several of our applications are based on common (or at least similar) mathematical results.
- These results are summarized in a special mathematical Appendix.

First Detailed Example: Economics of Reciprocity

17. What Is Reciprocity

- Usually, people have reasonably fixed attitude to others.
- They feel empathy towards members of their family, members of their tribe, usually citizens of their country.
- They may also be consistently negative towards their country's competitors.
- However, they also have widely fluctuating attitudes towards people with whom they work.
- It is difficult to predict how these attitudes will evolve – even in what direction they will evolve.
- Usually, people are nice to those who treat them nicely and negative to those who treat them badly.

18. Utility in the Traditional Economic Models

- In the traditional economic models, it is usually assumed that a decision maker maximizes his/her gain.
- This gain is numerically expressed as utility u.
- This utility value describe the effect of this decision on this person at this particular moment of time.

19. Dependence on Others' Utilities

- Let $u_i^{(0)}$ be approximate utilities that come only from this person's consumption.
- How can we take into account other people's feelings?
- A natural way is to add, to $u_i^{(0)}$, terms proportional to other people's utilities:
$$u_i = u_i^{(0)} + \sum_{j \ne i} \alpha_{ij} \cdot u_j.$$
- Here each coefficient α_{ij} describes how the utility of the i-th person depends on the utility of the j-th person.
- This phenomenon is known by a polite term *empathy*:
  - for positive α_{ij}, this describes how people feel better if others around them are happier;
  - it is also possible to have $\alpha_{ij} < 0$, when someone's happiness makes the other person unhappy.

20. What Is Reciprocity (cont-d)

- In terms of the coefficients α_{ij} it means that:
  - if α_{ji} is positive, then we expect α_{ij} to be positive;
  - if α_{ji} is negative, then we expect α_{ij} to be negative.
- This *reciprocity* phenomenon is intuitively clear – this is, after all, a natural human behavior.
- But how can we explain it in economic terms?

21. Let Us Formulate the Problem in Precise Terms

- Let us consider the simplest case, when we have only two people. Then:
$$u_1 = u_1^{(0)} + \alpha_{12} \cdot u_2; \quad u_2 = u_2^{(0)} + \alpha_{21} \cdot u_1.$$
- Since each person tries to maximize his/her utility, a natural question is as follows:
  - suppose that Person 1 knows the attitude α_{21} of Person 2 towards him/her;
  - what value α_{12} describing his/her attitude should Person 1 select to maximize his/her utility u_1?

22. Analysis of the Problem

- The above system of equations is easy to solve, we get

$$u_1 = \frac{u_1^{(0)} + \alpha_{12} \cdot u_2^{(0)}}{1 - \alpha_{12} \cdot \alpha_{21}}.$$

- This expression can take infinite value – i.e., as large a value as possible – if we take $\alpha_{12} = \dfrac{1}{\alpha_{21}}$.
- We can make it positive – and as large as possible – if we take α_{12} close to the inverse $1/\alpha_{21}$.
- Then, the difference $1 - \alpha_{12} \cdot \alpha_{21}$ will not be exactly 0, but be close to 0, with the same sign as the expression

$$u_1^{(0)} + \alpha_{12} \cdot u_2^{(0)}.$$

23. This Explains Reciprocity

- Indeed, according to the formula $\alpha_{12} = \dfrac{1}{\alpha_{21}}$:
  - if α_{21} is positive, then the selected value α_{12} is also positive, and
  - if α_{21} is negative, then the selected value α_{12} is also negative.

Second Detailed Example: Scale-Invariance Explains the Empirical Success of Inverse Distance Weighting in Geosciences

24. Need for Interpolation of Spatial Data

- Often, we are interested in the value of a certain physical quantity at different spatial locations.
- In geosciences, we may be interested in how depths of different geological layers depend of the spatial location.
- In environmental sciences, we may be interested in the concentration of substances in the atmosphere, etc.
- In principle, at each location, we can measure – directly or indirectly – the value of the corresponding quantity.
- However, we can only perform the measurement at a finite number of locations.
- But we are interested in the values of the quantity at all possible locations.

25. Need for Interpolation (cont-d)

- So, we need to estimate these values based on the measurement results – *interpolate* and *extrapolate*.
- In precise terms:
  - We know the values $q_i = q(x_i)$ of the quantity of interest q at several locations x_i, $i = 1, 2, \ldots, n$.
  - We would like to estimate the value $q(x)$ of this quantity at a given location x.

26. Inverse Distance Weighting

- A reasonable estimate q for $q(x)$ is a weighted average of the known values $q(x_i)$: $q = \sum_{i=1}^{n} w_i \cdot q_i$, with $\sum_{i=1}^{n} w_i = 1$.
- Naturally, the closer is the point x to the point x_i, the larger should be the weight w_i.
- So, the weight w_i with which we take the value q_i should decrease with the distance.
- Empirically, the best interpolation is attained when $w_i \sim (d(x, x_i))^{-p}$ for some $p > 0$.
- Since the weights have to add up to 1, we thus get

$$w_i = \frac{(d(x, x_i))^{-p}}{\sum_{j=1}^{n} (d(x, x_j))^{-p}}.$$

- This method is known as *inverse distance weighting*.

27. Challenge: Why Inverse Distance Weighting?

- In general, the fact that some algorithm is empirically the best means that:
  - we tried many other algorithms, and
  - this particular algorithm worked better than everything else we tried.
- In practice, we cannot try all possible algorithms, we can only try finitely many different algorithms.
- So, in principle, there could be an algorithm:
  - that we did not try and
  - that will work better than the one which is currently empirically the best.

28. Challenge (cont-d)

- Because of this:
  - every time we have some empirically best alternative,
  - it is desirable to come up with a theoretical explanation of why this alternative is indeed the best.
- And if such an explanation cannot be found, maybe it this alternative is actually not the best? Thus:
  - the empirical success of inverse distance weighting prompts a natural question:
  - is this indeed the best method?
- This is the challenge that we will deal with in this part of the talk.

29. What Is Scale Invariance

- When we process the values of physical quantities, we process real numbers.
- The numerical value of each quantity depends on the measuring unit.
- For example, suppose that we measure the distance in kilometers and get a numerical value d such as 2 km.
- Alternatively, we could use meters instead of kilometers.
- In this case, the exact same distance will be described by a different number: 2000 m.

30. What Is Scale Invariance (cont-d)

- In general:
  - if we replace the original measuring unit with a new one which is λ times smaller,
  - all numerical values will be multiplied by λ:

$$x \to \lambda \cdot x.$$

- Scale-invariance means that the result of interpolation should not change if we change the measuring unit.
- Let us analyze how this natural requirement affects interpolation.

31. General Case of Distance-Dependent Interpolation

- Let us consider the general case, when the further the point, the smaller the weight.
- In precise terms, the weight w_i is proportional to $f(d(x, x_i))$ for some decreasing $f(z)$: $w_i \sim f(d(x, x_i))$.
- Since the weights should add up to 1, we conclude that:

$$w_i = \frac{f(d(x, x_i))}{\sum\limits_j f(d(x, x_j))}, \text{ so } q = \sum_{i=1}^{n} \frac{f(d(x, x_i))}{\sum\limits_j f(d(x, x_j))} \cdot q_i.$$

- In this case, scale-invariance means that:

$$\sum_{i=1}^{n} \frac{f(\lambda \cdot d(x, x_i))}{\sum\limits_j f(\lambda \cdot d(x, x_j))} \cdot q_i = \sum_{i=1}^{n} \frac{f(d(x, x_i))}{\sum\limits_j f(d(x, x_j))} \cdot q_i.$$

32. Let Us Show That Scale-Invariance Leads to Inverse Distance Weighting

- Indeed, let us consider the case when we have only two measurement results:
  - at the point x_1, we got the value $q_1 = 1$, and
  - at point x_2, we got the value $q_2 = 0$.
- Then, for any point x, if we use the original distance values $d_1 \stackrel{\text{def}}{=} d(x, x_1)$ and $d_2 \stackrel{\text{def}}{=} d(x, x_2)$, we get: $q = \dfrac{f(d_1)}{f(d_1) + f(d_2)}$.
- So, scale invariance implies $\dfrac{f(\lambda \cdot d_1)}{f(\lambda \cdot d_1) + f(\lambda \cdot d_2)} = \dfrac{f(d_1)}{f(d_1) + f(d_2)}$.
- If we take the inverse of both sides, we get:

$$\frac{f(\lambda \cdot d_1) + f(\lambda \cdot d_2)}{f(\lambda \cdot d_1)} = \frac{f(d_1) + f(d_2)}{f(d_1)}.$$

33. Scale-Invariance Proof (cont-d)

- Subtracting number 1 from both sides, we get:

$$\frac{f(\lambda\cdot d_2)}{f(\lambda\cdot d_1)}=\frac{f(d_2)}{f(d_1)}.$$

- If we divide both sides by $f(d_2)$ and multiply by $f(\lambda\cdot d_1)$, we separate d_1 and d_2:

$$\frac{f(\lambda\cdot d_2)}{f(d_2)}=\frac{f(\lambda\cdot d_1)}{f(d_1)}.$$

- The left-hand side does not depend on d_1; thus, the right-hand side does not depend on d_1 either.
- It must thus depend only on λ; let us denote it by $c(\lambda)$.
- Then, from $\dfrac{f(\lambda\cdot d_1)}{f(d_1)}=c(\lambda)$, we conclude that

$$f(\lambda\cdot d_1)=c(\lambda)\cdot f(d_1).$$

34. Scale-Invariance Proof (cont-d)

- It is known that for decreasing functions $f(z)$, the only solutions to this functional equation are:

$$f(z) = c \cdot z^{-p} \text{ for some } p > 0.$$

- For this function $f(z)$, the extrapolated value has the form $\sum w'_i \cdot q_i$, with

$$w'_i = \frac{c \cdot (d(x, x_i))^{-p}}{\sum_{j=1}^{n} c \cdot (d(x, x_j))^{-p}}.$$

- If we divide both numerator and denominator by c, we get exactly the inverse distance weighting formula.

35. Comment

- The equation $f(\lambda \cdot d_1) = c(\lambda) \cdot f(d_1)$ is easy to solve for smooth function $f(x)$.
- Indeed, differentiating both sides by λ and taking $\lambda = 1$, we get $f'(d_1) \cdot d_1 = \alpha \cdot f(d_1)$, where $\alpha \stackrel{\text{def}}{=} c'(1)$.
- So, $\dfrac{df}{dd_1} = \alpha \cdot f$.
- If we divide both sides by f and multiply by dd_1, we separate d_1 and f: $\dfrac{df}{f} = \alpha \cdot \dfrac{dd_1}{d_1}$.
- Integrating both sides, we get $\ln(f) = \alpha \cdot \ln(d_1) + C$, where C is the integration constant.
- Applying $\exp(z)$ to both sides, we get $f(d_1) = c \cdot d_1^{\alpha}$, where $c \stackrel{\text{def}}{=} \exp(C)$.
- Since the function $f(z)$ is decreasing, we should have $\alpha < 0$, i.e., $\alpha = -p$ for some $p > 0$. Q.E.D.

Third Detailed Example: Why Geometric Progression in Selecting the LASSO Parameter – A Theoretical Explanation

36. Need for Regression

- In many real-life situations:
  - we know that the quantity y is uniquely determined by the quantities $x_1, \ldots, x_n$, but
  - we do not know the exact formula for this dependence.
- For example, in physics:
  - we know that the aerodynamic resistance increases with the body's velocity, but
  - we often do not know how exactly.
- In economics:
  - we know that a change in tax rate influences the economic growth, but
  - we often do not know how exactly.

37. Need for Regression (cont-d)

- In all such cases, we need to find the dependence $y = f(x_1, \ldots, x_n)$ between several quantities.
- This dependence must be determined based on the available data.
- We need to use previous observations $(x_{k1}, \ldots, x_{kn}, y_k)$ in each of which we know both:
  - the values x_{ki} of the input quantities x_i and
  - the value y_k of the output quantity y.
- In statistics, determining the dependence from the data is known as *regression*.

38. Need for Linear Regression

- In most cases, the desired dependence is smooth – and usually, it can even be expanded in Taylor series.
- In many practical situations, the range of the input variables is small, i.e., we have $x_i \approx x_i^{(0)}$ for some $x_i^{(0)}$.
- In such situations, after we expand the desired dependence in Taylor series, we can:
  - safely ignore terms which are quadratic or of higher order with respect to the differences $x_i - x_i^{(0)}$ and
  - only keep terms which are linear in terms of these differences:

$$y = f(x_1, \ldots, x_n) = c_0 + \sum_{i=1}^{n} a_i \cdot \left(x_i - x_i^{(0)}\right).$$

- Here $c_0 \stackrel{\text{def}}{=} f\left(x_1^{(0)}, \ldots, x_n^{(0)}\right)$ and $a_i \stackrel{\text{def}}{=} \dfrac{\partial f}{\partial x_i}_{|x_i = x_i^{(0)}}$.

39. Need for Linear Regression (cont-d)

- This expression can be simplified into:

$$y = a_0 + \sum_{i=1}^{n} a_i \cdot x_i, \text{ where } a_0 \stackrel{\text{def}}{=} c_0 - \sum_{i=1}^{n} a_i \cdot x_i^{(0)}.$$

- In practice, measurements are never absolutely precise.
- So, when we plug in the actually measured values x_{ki} and y_k, we will only get an approximate equality:

$$y_k \approx a_0 + \sum_{i=1}^{m} a_i \cdot x_{ki}.$$

- Thus, the problem of finding the desired dependence can be reformulated as follows:
  - given the values y_k and x_{ki},
  - find the coefficients a_i for which the approximate equality holds for all k.

40. The Usual Least Squares Approach

- We want each left-and side y_k of the approximate equality to be close to the corresponding right-hand side.
- In other words, we want the left-hand-side tuple $(y_1, \ldots, y_K)$ to be close to the right-hand-sides tuple

$$\left(\sum_{i=1}^{m} a_i \cdot x_{1i}, \ldots, \sum_{i=1}^{m} a_i \cdot x_{Ki}\right).$$

- It is reasonable to select a_i for which the distance between these two tuples is the smallest possible.
- Minimizing the distance is equivalent to minimizing the square of this distance, i.e., the expression

$$\sum_{k=1}^{K} \left(y_k - \left(a_0 + \sum_{i=1}^{m} a_i \cdot x_{ki}\right)\right)^2.$$

- This minimization is know as the *Least Squares method*.

41. The Least Squares Approach (cont-d)

- This is the most widely used method for processing data.
- The corresponding values a_i can be easily found if:
  - we differentiate the quadratic expression with respect to each of the unknowns a_i and then
  - equate the corresponding linear expressions to 0.
- Then, we get an easy-to-solve systems of linear equations.

42. Discussion

- The above heuristic idea becomes well-justified:
  - when we consider the case when the measurement errors are normally distributed
  - with 0 mean and the same standard deviation σ.
- This indeed happens:
  - when the measuring instrument's bias has been carefully eliminated, and
  - most major sources of measurement errors have been removed.
- In such situations, the resulting measurement error is a joint effect of many similarly small error components.
- For such joint effects, the Central Limit Theorem states that the resulting distribution is close to Gaussian.

43. Discussion (cont-d)

- Once we know the probability distributions, a natural idea is to select the most probable values a_i.
- In other words, we select the values for which the probability to observe the values y_k is the largest.
- For normal distributions, this idea leads exactly to the least squares method.

44. Need to Go Beyond Least Squares

- Sometimes, we know that all the inputs x_i are essential to predict the value y of the desired quantity.
- In such cases, the least squares method works reasonably well.
- The problem is that in practice, we often do not know which inputs x_i are relevant and which are not.
- As a result, to be on the safe side, we include as many inputs as possible.
- Many of them will turn out to be irrelevant.
- If all the measurements were exact, this would not be a problem:
  - for irrelevant inputs x_i, we would get $a_i = 0$, and
  - the resulting formula would be the desired one.

45. Need to Go Beyond Least Squares (cont-d)

- However, because of the measurement errors, we do not get exactly 0s.
- Moreover, the more such irrelevant variables we add:
  - the more non-zero "noise" terms $a_i \cdot x_i$ we will have, and
  - the larger will be their sum.
- This will negatively affecting the accuracy of the formula,
- Thus, it will negative affect the accuracy of the resulting desired (non-zero) coefficients a_i.

46. LASSO Method

- We know that many coefficients will be 0; so, a natural idea is:
  - instead of considering all possible tuples

$$a \stackrel{\text{def}}{=} (a_0, a_1, \ldots, a_n),$$

  - to only consider tuples for which a bounded number of coefficients is non-0: $\|a\|_0 \le B$ for some constant B.
- Here, $\|a\|_0$ (known as the ℓ_0-norm) denotes the number of non-zero coefficients in a tuple.
- The problem with this natural idea is that the resulting optimization problem becomes NP-hard.
- This means, crudely speaking, that:
  - no feasible algorithm is possible
  - that would always solve all the instances of this problem.

47. LASSO Method (cont-d)

- A usual way to solve such problem is:
  - by replacing the ℓ_0-norm with an ℓ_1-norm $\sum\limits_{i=0}^{n} |a_i|$;
  - this norm is convex, therefore, the optimization problem is easier to solve.
- So:
  - instead of solving the problem of unconditionally minimizing the quadratic expression,
  - we minimize this expression under the constraint $\sum\limits_{i=0}^{n} |a_i| \le B$ for some constant B.
- This minimum can be attained when we have strict inequality or when the constraint becomes an equality.
- If the constraint is a strict inequality, then we have a local minimum.

48. LASSO Method (cont-d)

- For quadratic functions, a local minimum is exactly the global minimum that we try to avoid.
- Thus, we must consider the case when the constraint becomes an equality $\sum\limits_{i=0}^{n} |a_i| = B$.
- The Lagrange multiplier method leads to minimizing the expression:

$$\sum_{k=1}^{K} \left(y_k - \left(a_0 + \sum_{i=1}^{m} a_i \cdot x_{ki} \right) \right)^2 + \lambda \cdot \sum_{i=0}^{n} |a_i|.$$

- This minimization is known as the *Least Absolute Shrinkage and Selection Operator* method – *LASSO*, for short.

49. How λ Is Selected: Main Idea

- The success of the LASSO method depends on what value λ we select.
- When λ is close to 0, we retain all the problems of the usual least squares method.
- When λ is too large, the λ-term dominates.
- So we select all the values $a_i = 0$, which do not provide any good description of the desired dependence.
- In different situations, different values λ will work best.
- The more irrelevant inputs we have:
  - the more important it is to deviate form the least squares, and
  - thus, the larger the parameter λ – that describes this deviation – should be.

50. How λ Is Selected: Main Idea (cont-d)

- We rarely know beforehand which inputs are relevant – this is the whole problem.
- So we do now know beforehand what value λ we should use.
- The best value λ needs to be decided based on the data.
- A usual way of testing any dependence is by randomly dividing the data into:
  - a (larger) training set and
  - a (smaller) testing set.
- We use the training set to find the value of the desired parameters (in our case, the parameters a_i).
- Then we use the testing set to gauge how good is the model.

51. How λ Is Selected: Main Idea (cont-d)

- To get more reliable results, we can repeat this procedure several times.
- In precise terms, we select several training subsets

$$S_1, \ldots, S_m \subseteq \{1, \ldots, K\}.$$

- For each of these subsets S_j, we find the values $a_{ij}(\lambda)$ that minimize the functional

$$\sum_{k \in S_j} \left(y_k - \left(a_0 + \sum_{i=1}^{m} a_i \cdot x_{ki} \right) \right)^2 + \lambda \cdot \sum_{i=0}^{n} |a_i|.$$

- We can then compute the overall inaccuracy, as

$$\Delta(\lambda) \stackrel{\text{def}}{=} \sum_{j=1}^{m} \left(\sum_{k \notin S_j} \left(y_k - \left(a_{j0}(\lambda) + \sum_{i=1}^{m} a_{ji}(\lambda) \cdot x_{ki} \right) \right)^2 \right).$$

- We then select λ for which $\Delta(\lambda)$ is the smallest.

52. How λ Is Selected: Details

- In the ideal world, we should be able to try all possible real values λ.
- However, there are infinitely many real numbers, and in practice, we can only test finitely many of them.
- Which set of values λ should we choose?
- Empirically, the best results are obtained if we use the values λ from a geometric progression $\lambda_n = c_0 \cdot q^n$.
- Of course, a geometric progression also has infinitely many values, but we do not need to test all of them.
- Usually, as λ increases from 0, the value $\Delta(\lambda)$ first decreases then increases again.
- So, it is enough to catch a moment when this value starts increasing.

53. How λ Is Selected: Details (cont-d)

- A natural question is: why geometric progression works best?
- In this part of the talk, we provide a theoretical explanation for this empirical fact.

54. What Do We Want?

- At first glance, the answer is straightforward: we want to select a discrete set of values, i.e., a set

$$S = \{\ldots < \lambda_n < \lambda_{n+1} < \ldots\}.$$

- However, a deeper analysis shows that the answer is not so simple.
- Indeed, what we are interested in is the dependence between the quantities y and x_i.
- However, what we have to deal with is not the quantities themselves, but their numerical values.
- And the numerical values depend on what unit we choose for measuring these quantities; for example:
  - a person who is 1.7 m high is also 170 cm high,
  - an April 2020 price of 2 US dollars is the same as the price of $2 \cdot 23500 = 47000$ Vietnam Dong, etc.

55. What Do We Want (cont-d)

- In most cases, the choice of the units is rather arbitrary.
- It is therefore reasonable to require that the results of data processing should not depend on the unit.
- And hereby lies a problem.
- Suppose that we keep the same units for x_i but change a measuring unit for y to a one which is α times smaller.
- In this case, the new numerical values of y become α times larger: $y \to y' = \alpha \cdot y$.
- To properly capture these new values, we need to increase the original values a_i by the same factor:

$$a_i \to a_i' = \alpha \cdot a_i.$$

56. What Do We Want (cont-d)

- In terms of these new values, the minimized expression takes the form

$$\sum_{k=1}^{K}\left(y'_k-\left(a'_0+\sum_{i=1}^{m}a'_i\cdot x_{ki}\right)\right)^2+\lambda\cdot\sum_{i=0}^{n}|a'_i|.$$

- Taking into account that $y'_k=\alpha\cdot y_k$ and $a'_i=\alpha\cdot a_i$, we get:

$$\alpha^2\cdot\sum_{k=1}^{K}\left(y_k-\left(a_0+\sum_{i=1}^{m}a_i\cdot x_{ki}\right)\right)^2+\alpha\cdot\lambda\cdot\sum_{i=0}^{n}|a_i|.$$

- Minimizing an expression is the same as minimizing α^{-2} times this expression, i.e., the modified expression

$$\sum_{k=1}^{K}\left(y_k-\left(a_0+\sum_{i=1}^{m}a_i\cdot x_{ki}\right)\right)^2+\alpha^{-1}\cdot\lambda\cdot\sum_{i=0}^{n}|a_i|.$$

57. What Do We Want (cont-d)

- This new expression is similar to the original one, but with a new value of the LASSO parameter $\lambda' = \alpha^{-1} \cdot \lambda$.
- So, when we change the measuring units, the values of λ are also re-scaled – i.e., multiplied by a constant.
- What was the set $\{\lambda_n\}$ in the old units becomes the re-scaled set $\{\alpha^{-1} \cdot \lambda_n\}$ in the new units.
- This is, in effect, the same set but corresponding to different measuring units.
- So, we cannot say that one of these sets is better than the other, they clearly have the same quality.
- Thus, we cannot choose a single set S, we must choose a family of sets $\{c \cdot S\}_c$, where

$$c \cdot S \stackrel{\text{def}}{=} \{c \cdot \lambda : \lambda \in S\}.$$

58. Natural Uniqueness Requirement

- Eventually, we need to select some set S.
- We cannot select one set a priori, since with every set S, a set $c \cdot S$ also has the same quality.
- To fix a unique set, we can, e.g., fix one of the values

$$\lambda \in S.$$

- Let us require that with this fixture, we will be end up with a unique optimal set S.
- This means, in particular, that:
  - if we select a real number $\lambda \in S$,
  - then the only set $c \cdot S$ that contains this number will be the same set S.
- Let us describe this requirement in precise terms.

59. Definitions and the Main Result

- *A set* $S \subseteq \mathbb{R}^+$ *is called* discrete *if:*
  - *for every* $\lambda \in S$,
  - *there exists a* $\varepsilon > 0$ *such that* $|\lambda - \lambda'| > \varepsilon$ *for all other* $\lambda' \in S$.
- For such sets, for each element λ:
  - if there are larger elements,
  - then there is the "next" element – i.e., the smallest element which is larger than λ.
- Similarly:
  - if there are smaller elements,
  - then there exists the "previous" element – i.e., the largest element which is smaller than λ.
- Thus, such sets have the form

$$\{\ldots < \lambda_{n-1} < \lambda_n < \lambda_{n+1} < \ldots\}.$$

60. Definitions and the Main Result (cont-d)

- *A discrete set S is called* uniquely determined *if for every $\lambda \in S$ and $c > 0$, if $\lambda \in c \cdot S$, then $c \cdot S = S$.*

- **Proposition.** *A set S is uniquely determined if and only if it is a geometric progression, i.e.:*

$$S = \{c_0 \cdot q^n : n = \ldots, -2, -1, 0, 1, 2, \ldots\} \text{ for some } c_0 \text{ and } q.$$

- This results explains why geometric progression is used to select the LASSO parameter λ.

61. Proof

- It is easy to prove that every geometric progression is uniquely determined.
- Indeed, if for $\lambda = c_0 \cdot q^n$, we have $\lambda \in c \cdot S$, this means that $\lambda = c \cdot c_0 \cdot q^m$ for some m, i.e., $c_0 \cdot q^n = c \cdot c_0 \cdot q^m$.
- Dividing both sides by $c_0 \cdot q^m$, we conclude that $c = q^{n-m}$ for some integer $n - m$.
- Let us show that in this case, $c \cdot S = S$.
- Indeed, each element x of the set $c \cdot S$ has the form $x = c \cdot c_0 \cdot q^k$ for some integer k.
- Substituting $c = q^{n-m}$ into this formula, we conclude that $x = c_0 \cdot q^{k+(n-m)}$, i.e., that $x \in S$.
- Similarly, we can prove that if $x \in S$, then $x \in c \cdot S$.

62. Proof (cont-d)

- Vice versa, let us assume that the set S is uniquely determined.
- Let us pick any element $\lambda \in S$ and denote it by λ_0.
- The next element we will denote by λ_1, the next to next by λ_2, etc.
- Similarly, the element previous to λ_0 will be denoted by λ_{-1}, previous to previous by λ_{-2}, etc.
- Thus, $S = \{\ldots, \lambda_{-2}, \lambda_{-1}, \lambda_0, \lambda_1, \lambda_2, \ldots\}$.
- Clearly, $\lambda_1 \in S$, and for $q \stackrel{\text{def}}{=} \lambda_1/\lambda_0$, we have $\lambda_1 \in q \cdot S$ – since $\lambda_1 = (\lambda_1/\lambda_0) \cdot \lambda_0 = q \cdot \lambda_0$ for $\lambda_0 \in S$.
- Since the set S is uniquely determined, this implies that $q \cdot S = S$.
- Since $S = \{\ldots, \lambda_{-2}, \lambda_{-1}, \lambda_0, \lambda_1, \lambda_2, \ldots\}$, we have

$$q \cdot S = \{\ldots, q \cdot \lambda_{-2}, q \cdot \lambda_{-1}, q \cdot \lambda_0, q \cdot \lambda_1, q \cdot \lambda_2, \ldots\}.$$

63. Proof (cont-d)

- The sets S and $q \cdot S$ coincide.
- We know that $q \cdot \lambda_0 = \lambda_1$; thus:
  - the element next to $q \cdot \lambda_0$ in the set $q \cdot S$ – i.e., the element $q \cdot \lambda_1$,
  - must be equal to the element which is next to λ_1 in the set S, i.e., to the element λ_2:
  $$\lambda_2 = q \cdot \lambda_1.$$
- For next to next elements, we get $\lambda_3 = q \cdot \lambda_2$ and, in general, we get $\lambda_{n+1} = q \cdot \lambda_n$ for all n.
- This is exactly the definition of a geometric progression.
- The proposition is proven.

64. Discussion

- Machine learning (e.g., deep learning) uses the gradient method $x_{i+1} = x_i - \lambda_i \cdot \dfrac{\partial J}{\partial x_i}$ to minimize J.
- Empirically the best strategy for selecting λ_i also follows approximately a geometric progression.
- For example, some algorithms use:
  - $\lambda_i = 0.1$ for the first ten iterations,
  - $\lambda_i = 0.01$ for the next ten iterations,
  - $\lambda_i = 0.001$ for the next ten iterations, etc.
- In this case, similarly, re-scaling of J is equivalent to re-scaling of λ.
- Thus, we need to have a family of sequences $\{c \cdot \lambda_i\}$ corresponding to different $c > 0$.
- A natural uniqueness requirement – as we have shown – leads to the geometric progression.

65. Acknowledgments

I want to thank:

- my advisor Dr. Vladik Kreinovich;
- my committee members Drs. Martine Ceberio, Eric D. Smith, and Aaron Velasco;
- my wife Pooja Bhatt for supporting me throughout the process of this dissertation;
- my friends Pawan Koirala, Kamal Nyaupane, Neelam Dumre, Bibek Aryal, and Subharaj Ranabhat; and
- my father Damber Dutta Bokati, my mother Ishwori Devi Bokati, and my brother Prakash Bokati.

Without your inspiration and support, I would not have been able to pursue my academic goals.