Statistics and Machine Learning 1

Lecture 1A: Randomness and Its Uses

Mark Muldoon
Department of Mathematics, Alan Turing Building
University of Manchester

Week 1

Motivation

There are three main ways that randomness comes into data science:

1. The *data themselves* are often best understood as random. For example, surveys are often considered a random sample; measurements affected by noise or error can also be viewed as random.
2. When we want to reason under *subjective uncertainty* (for example in Bayesian approaches) then unknown quantities can be represented as random. Often when we make predictions they will be *probabilitic*.
3. Many of the most effective / efficient / commonly-used algorithms in data science—typically called *Monte Carlo* algorithms—exploit randomness.

The mathematical foundations of the subject are relatively recent: modern axioms for probability were introduced by Kolmogorov in 1933.

True Randomness

- As far as we know, the only truly random phenomena are quantum mechanical.
- It is possible to buy hardware that generates ‘physical’ / ‘true’ / ‘quantum’ random numbers.
- We will not be working with such equipment, so sometimes we’ll have to *simulate* randomness.

What is a probability?

- There is a lot of academic / philosophical discussion about this question!
- We will be somewhat pragmatic, and will consider two versions of the prototypical example, a *flip of a coin*:

Unpredictability

- Flipping a standard coin many times (like before a sporting event) does not give a predictable outcome.
- This is because there are too many variables out of our control—the number of flips is very sensitive to the initial force, the wind, the placement of the coin on the thumb …

- Unless the coin is oddly balanced, for all practical purposes, each toss is equally likely (has a probability of half) to come up heads or tails, and this does not not depend on any previous results or influence later ones.

Subjective uncertainty

- Flipping a very large coin over once gives us an outcome we can easily predict *given knowledge of its current state*—heads and tails will switch each time.
- We can make this prediction because we control and measure all of the variables involved in flipping a large coin over once. The probability that the next face up after a flip is tails *given that* the current face up is heads is one.

- But if someone phones you and says 'I have a giant coin, what will the next face up be?' then in the absence of additional information it is rational to give each of heads and tails probability one half.

Unpredictable dynamics: the logistic map

The idea of *deterministic chaos* was much studied a few years ago, but has somewhat given way to random models. Here we're going to consider a simple model of population dynamics—the *logistic model*—that, although purely deterministic, can behave in a way that appears random.

The model describes a population that reproduces once per year and says that if the population in year n is x_n, then the population the next year is given by

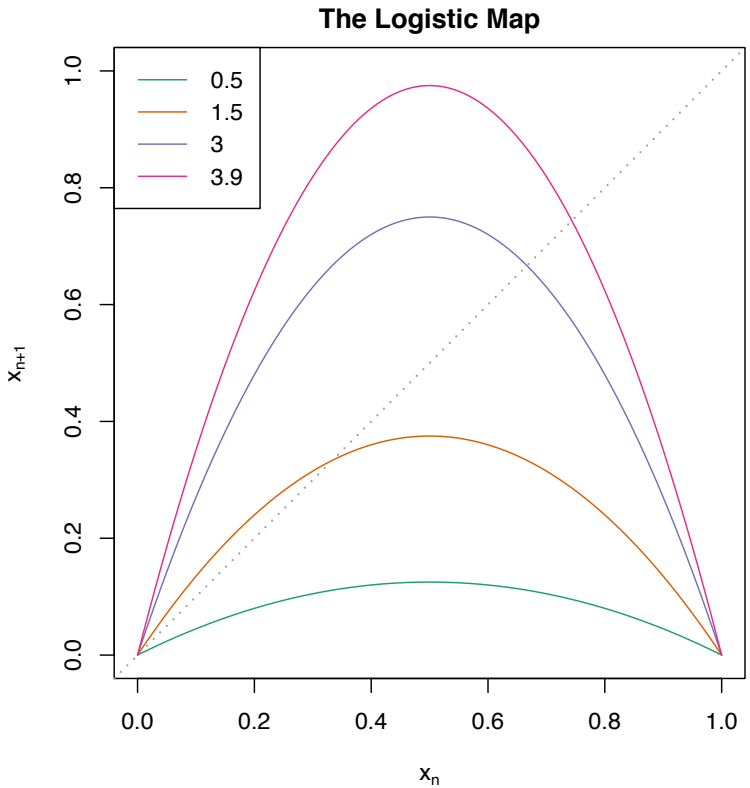
$$x_{n+1} = r x_n (1 - x_n),$$

where:

- we'll imagine that we're the measuring population as a fraction of some maximal one, so $0 \leq x_n \leq 1$;
- to keep x_{n+1} in the same range, we need $0 \leq r \leq 4$.

As $r \to 4$, we'll see apparently *unpredictable* behaviour, even though the sequence arises from a *simple rule*.

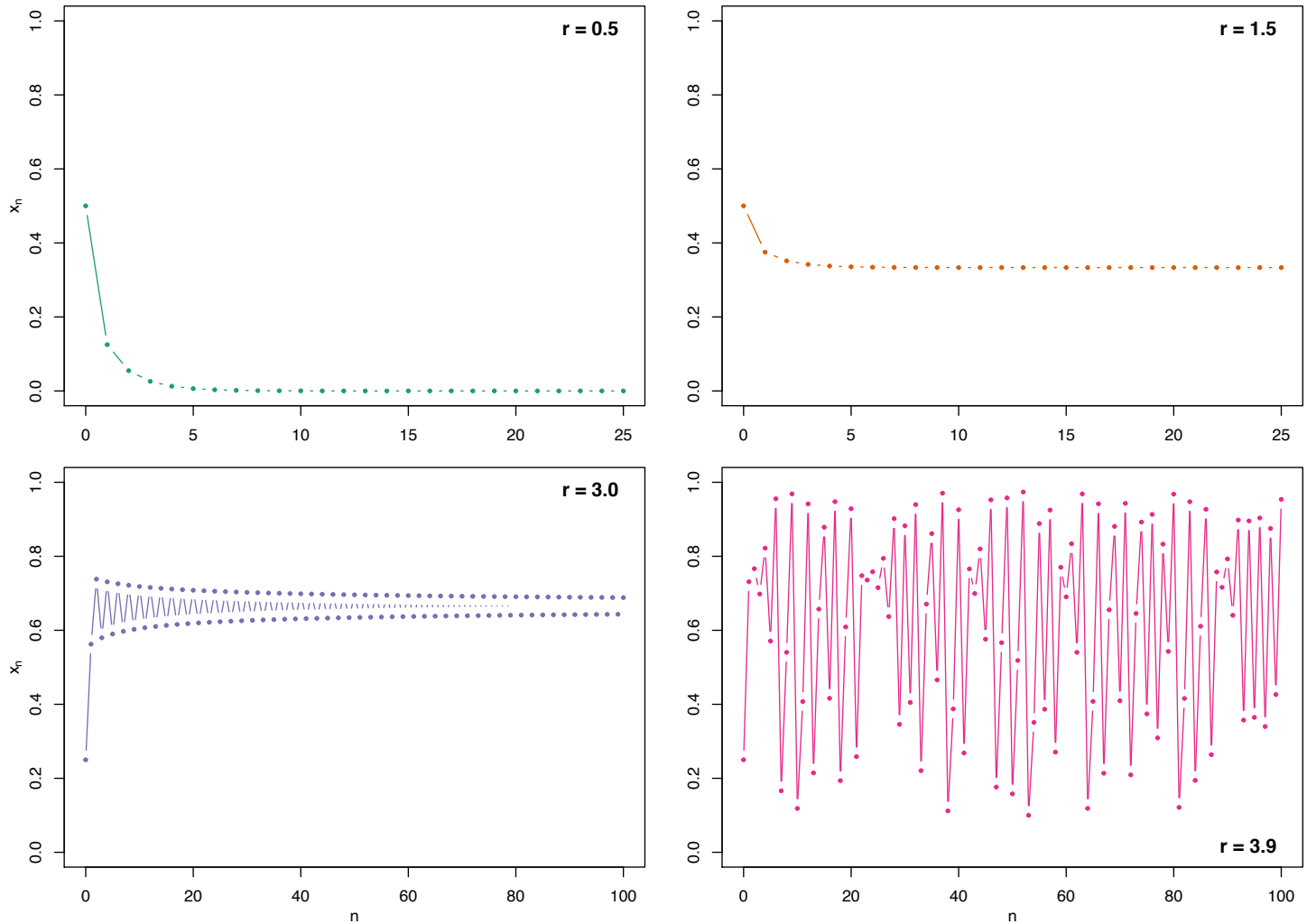
Logistic map, qualitatively



The curve giving x_{n+1} as a function of x_n for various values of r.

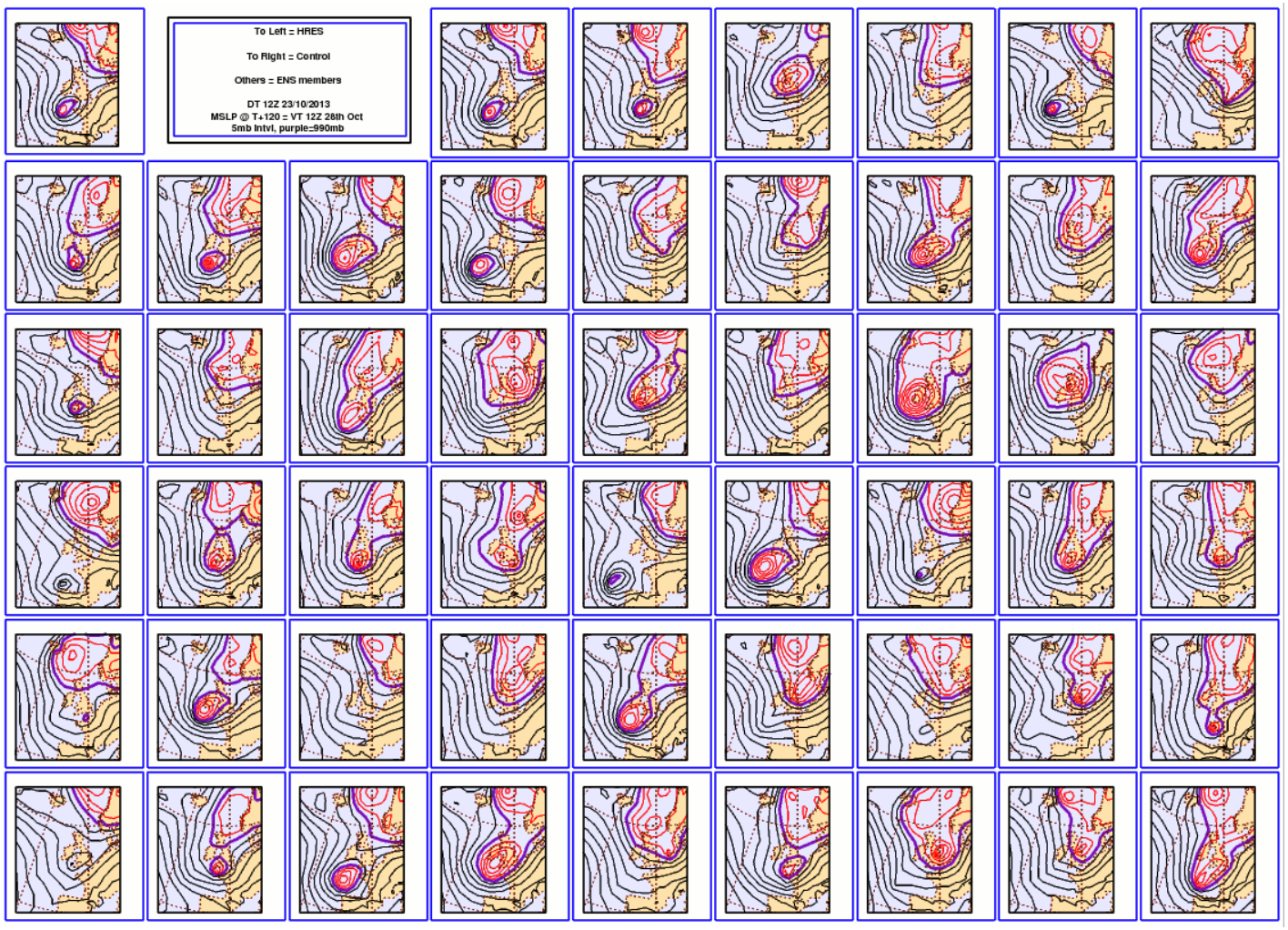
- We can generate a sequence of x's—which corresponds to a predicted history for the population—by:
  - Choosing some starting population x_0.
  - Applying $x_{n+1} = rx_n(1 - x_n)$ with $n = 0, 1, \ldots$ to get $x_1, x_2, \ldots$
- If $x_n = 0$, then $x_{n+1} = 0$ too: the population has gone extinct.
- If $x_n = 1$, then $x_{n+1} = 0$: overcrowding can cause extinction.
- When r is small ($r < 1$), x_{n+1} is always less than x_n, so the population dies away.

Logistic map: deterministic chaos



When $r > 3.7$ the sequence $x_0, x_1, x_2, \ldots$ hops around the whole interval, $0 < x_n < 1$, apparently randomly.

Weather forecasting: chaos, simulation & probabilistic prediction



ECMWF

Electoral politics: probabilistic prediction

Elections simulated using polling data from the period up to 9 October 2020.