Observing the Population Dynamics in GE by means of the Intrinsic Dimension

E. Medvet, **A. Bartoli**, F. Tarlao, A. Ansuini

University of Trieste, Italy

Context: Diversity in EC (I)

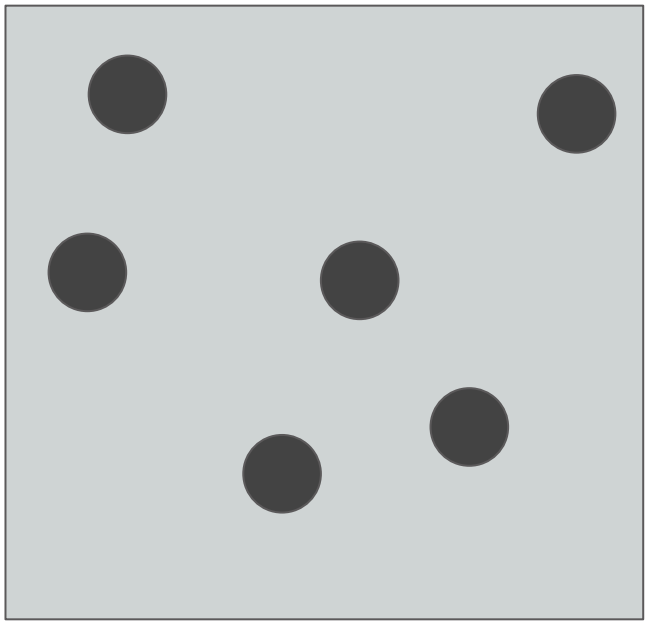
- We consider **population-based** Evolutionary Computation (EC)
- Population with many "**similar**" individuals is **ineffective**
  - Premature convergence
- Population shall have "**high diversity**"

Context: Diversity in EC (II)

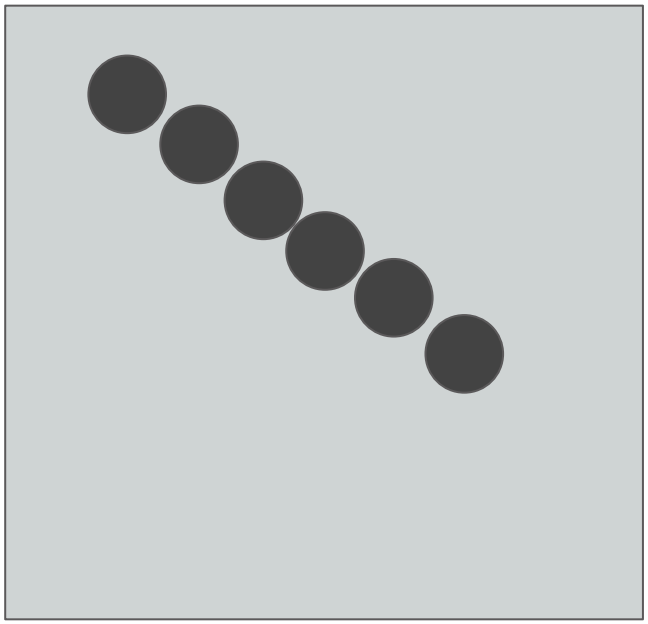
- Many different ways for quantifying diversity
  - Genotype
  - Phenotype
  - Behavior
- Frequent idea
  - Fraction of “**unique**” individuals w.r.t. Population size

Our (exploratory) idea (I)

- Is there any "**meaningful difference**" between their diversities?



Population 1



Population 2

Our (exploratory) idea (II)

- Is a “highly diverse” population effectively exploring the search space…
- ...or is it “compressed” in a subspace of **much smaller dimension**?

Intrinsic Dimension (ID)

- N-dimensional dataset D
- Its ID is the **minimum number of variables** required for **describing** D
- Intuition
  - All dataset points on a hyperplane: ID=2
  - All dataset points on a line: ID=1
  - **Irrespective** of the dimension N of the space
- Increasingly used in several machine learning-related problems
- Algorithm for estimating ID

Our (exploratory) work

- Grammatical Evolution
  - Genotype: bit string
  - Phenotype: program in language with Context Free Grammar
- Why?
  - Interesting applications
  - Simple distance in genotype space
  - Diversity is important
    - Suffer when lack of diversity
    - Several mechanisms for promoting diversity
  - ...and because we have much experience...

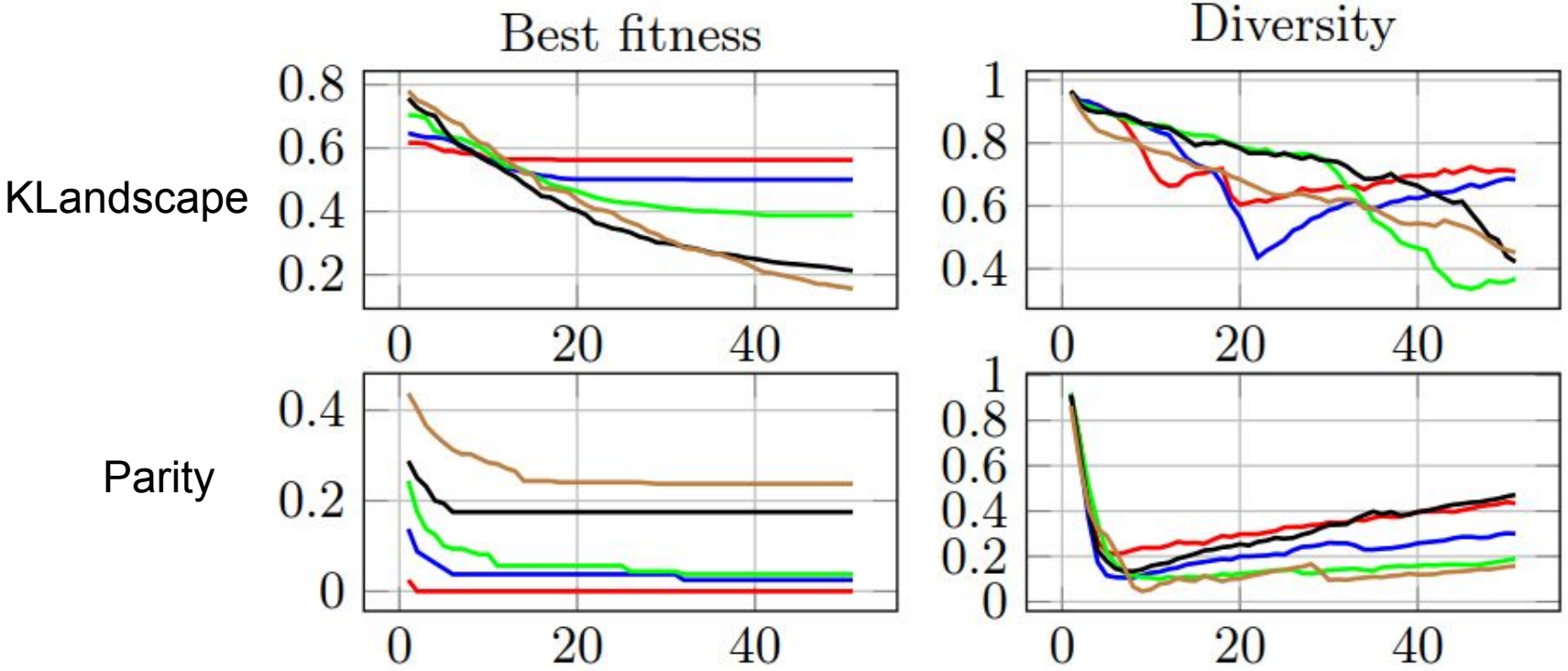
Methodology

- Two benchmark problems
  - Tunable hardness
- We measured at each generation:
  - Best **fitness**
  - **Diversity** of population (based on uniqueness)
  - **Intrinsic Dimension** of population

Methodology: some details

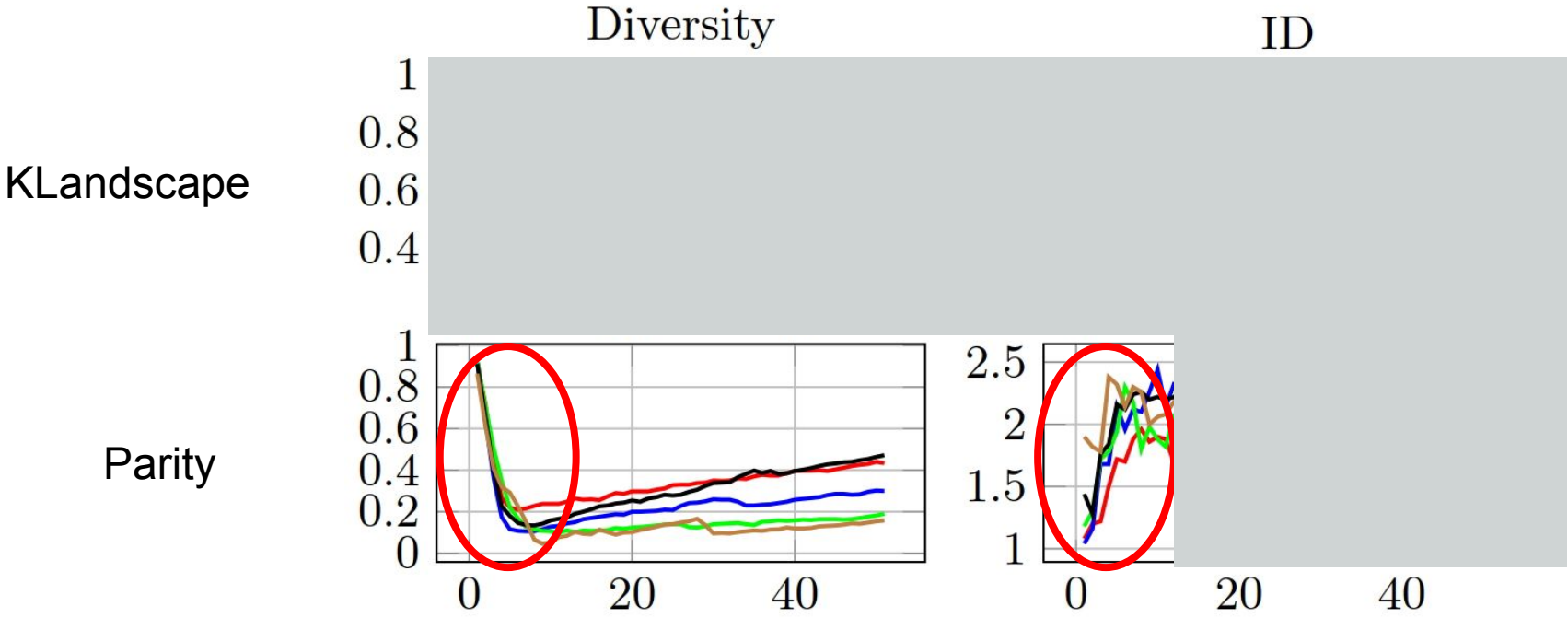
- Weighted Hierarchical GE
  - Recent variant
  - Different representation
  - Better effectiveness
- Parity and KLandscape
  - Hardness 3-7
- Average indexes on 5 independent runs
- 500 individuals, 50 generations

Results: Fitness vs Diversity



- Diversity initially high
- Reaches a minimum (more or less quickly)
- Grow again when fitness no longer increases

Results: Diversity vs ID (I)

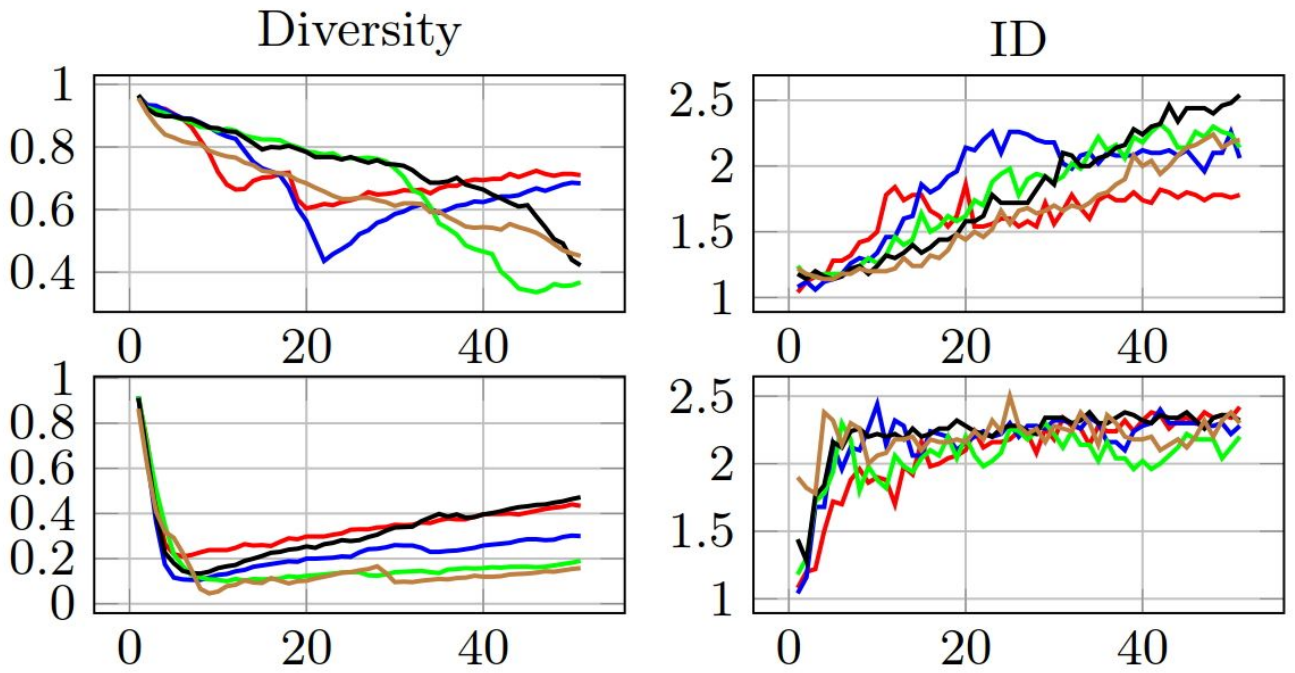


- ID **grows quickly** when diversity **drops quickly**
- Suggests that population distributes itself on a "**flat**" region

Results: Diversity vs ID (II)

KLandscape

Parity

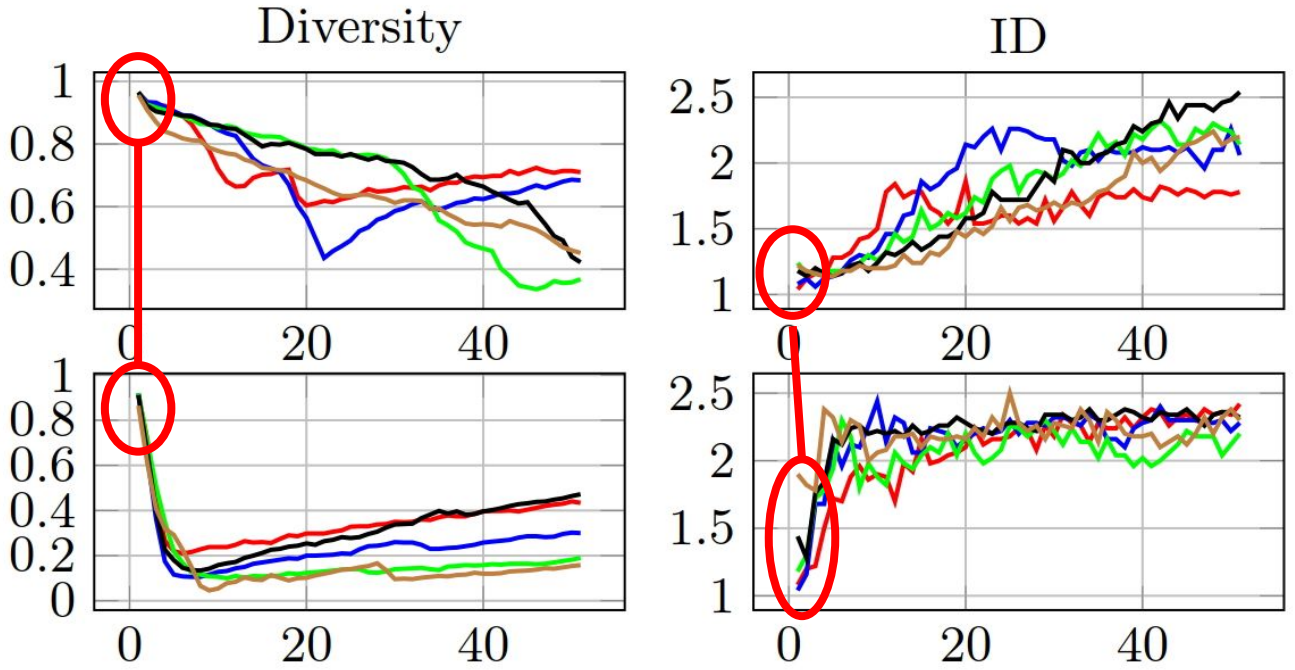


- ID tends to always grow (more or less)
- Never exceeds 2-2.5
  - Flat region???

Results: Diversity vs ID (III)

KLandscape

Parity



- Diversity initially high, but ID=1
  - Distributed along a line
- Maybe it is “**not diverse enough**”???

Summary

- ID **seem** to provide information
  - Useful
  - Complementary to Diversity
- **Many more** experiments needed
- **Big warning**
  - Is Hamming distance in genotype space meaningful for the ID estimation algorithm?

Thanks for your attention

http://machinelearning.inginf.units.it