Regularized Feature Selection Landscapes: An Empirical Study of Multimodality

Xavier F. C. Sánchez-Díaz, **Corentin Masson**, **Ole Jakob Mengshoel**

Department of Computer Science
Norwegian University of Science and Technology

xavier.sanchezdz@ntnu.no, camasson@stud.ntnu.no, ole.j.mengshoel@ntnu.no

Abstract

The processing of **features** in data is among the key topics in machine learning. While a broad range of heuristics for feature processing and selection have been developed and experimented with, less research has been concerned with the underlying fitness landscape. We perform a fitness **landscape analysis of feature selection**, using local optima networks and other methods. We focus on the impact of **regularization**, a central machine learning topic. Our study, using decision trees, confirms and adds to previous findings that **feature selection landscapes are highly multimodal**. In the ten UCI datasets studied, we find a high degree of multimodality when there is no regularization. With increasing regularization, the degree of multimodality generally drops off but remains substantial.

Problem Definition

Consider a bitstring $\boldsymbol{b} = b_1, \dots b_n$ indicating which features are included ($b_i = 1$) or not ($b_i = 0$). We model the feature selection problem as an *energy* function to **minimize**:

$$h(\boldsymbol{b}) = h_E(T(\boldsymbol{b})) + \epsilon \cdot h_P(\boldsymbol{b}),$$

where $h_E(T(\boldsymbol{b}))$ is the **classification error** over a given dataset using a decision tree, $h_P(\boldsymbol{b})$ is a penalty depending on the number of features used for training with the feature subset $\boldsymbol{b}$, and ϵ controls the degree of regularization.

Method and Datasets

- **Datasets**: 10 classification datasets from UCI
- **Model**: A decision tree trained on all 2^n combinations of features for different values of $\epsilon \in \{0, 1/32, 1/16, 1/8\}$.
- Accuracy tables can be downloaded using the QR code

Table 1: Two of the ten UCI datasets used in this study, sorted by number of features (n). We present the number of examples m, local optima L, and global optima G for various values of the regularization term ϵ. We focus only on **4-glass** in this poster.

Name	n	m	Number of optima $\epsilon = 0$ L	G	$\epsilon = 1/32$ L	G	$\epsilon = 1/16$ L	G	$\epsilon = 1/8$ L	G
4-glass	9	214	65	1	51	2	22	2	7	2
5-heart-c	13	303	700	1	407	1	117	1	13	1

Results and Findings

Finding 1: The landscape changes under regularization

A steep reduction in the number of optima suggests that the landscapes undergo several changes due to increasing regularization. See for example Table 2.

Table 2: The tree lowest-energy optima in 4-glass, for regularization values $\epsilon = 0$ (top three rows) and $\epsilon = 1/8$ (bottom three rows). Redundant or unimportant features are highlighted in red when there is a tie, i.e., two different feature subsets $\boldsymbol{b}_i^*$ and $\boldsymbol{b}_j^*$ have the same energy $h(\boldsymbol{b}_i^*) = h(\boldsymbol{b}_j^*)$.

	ID i	x-axis	y-axis	Bitstring original	x-axis	y-axis	Energy $\epsilon = 0$	$\epsilon = 1/8$	Accuracy
$\boldsymbol{b}_{273}^*$	273	17	1	100010001	10001	0001	**0.2188**	0.5938	0.7813
$\boldsymbol{b}_{63}^*$	63	3	15	000111111	00011	1111	**0.2344**	0.9844	0.7656
$\boldsymbol{b}_{235}^*$	235	14	11	011101011	01110	1011	**0.2344**	0.9844	0.7656
$\boldsymbol{b}_{1}^*$	1	0	1	000000001	00000	0001	0.2656	**0.3906**	0.7344
$\boldsymbol{b}_{2}^*$	2	0	2	000000010	00000	0010	0.2656	**0.3906**	0.7344
$\boldsymbol{b}_{16}^*$	16	1	0	000010000	00001	0000	0.3125	**0.4375**	0.6875

Figure 1 shows an **overview of the fitness landscape** in the 4-glass dataset. In this 2D bitmap, we *slice* the bitstring in two: the first half mapped to the x-axis and the second half mapped to the y-axis (rounding up in favor of the x-axis when n is odd). We call this visualization a *hinged bitstring map* or HBM.

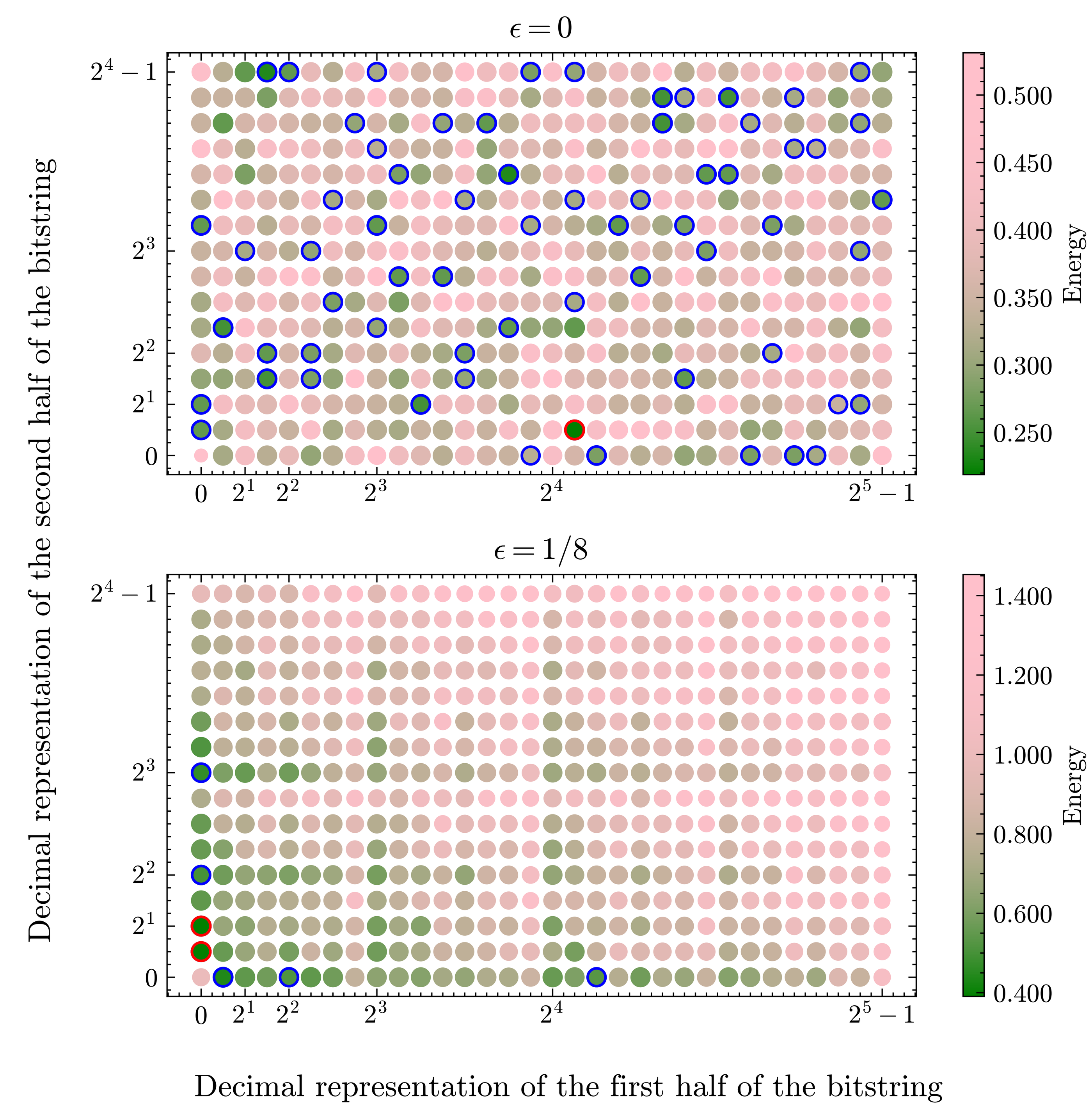

Figure 1: Hinged bitstring map of the 4-glass dataset with $n = 9$ features. The local and global optima are highlighted with blue and red outlines, respectively.

Finding 2: The distribution and concentration of local optima changes too

Basins of attraction undergo some changes when regularization varies. Figure 2 shows how the concentration of optima around certain 'basins' varies in the 4-glass dataset when the regularization parameter ϵ is modified. Additional plots of partial LONs on an HBM can be found in the paper.

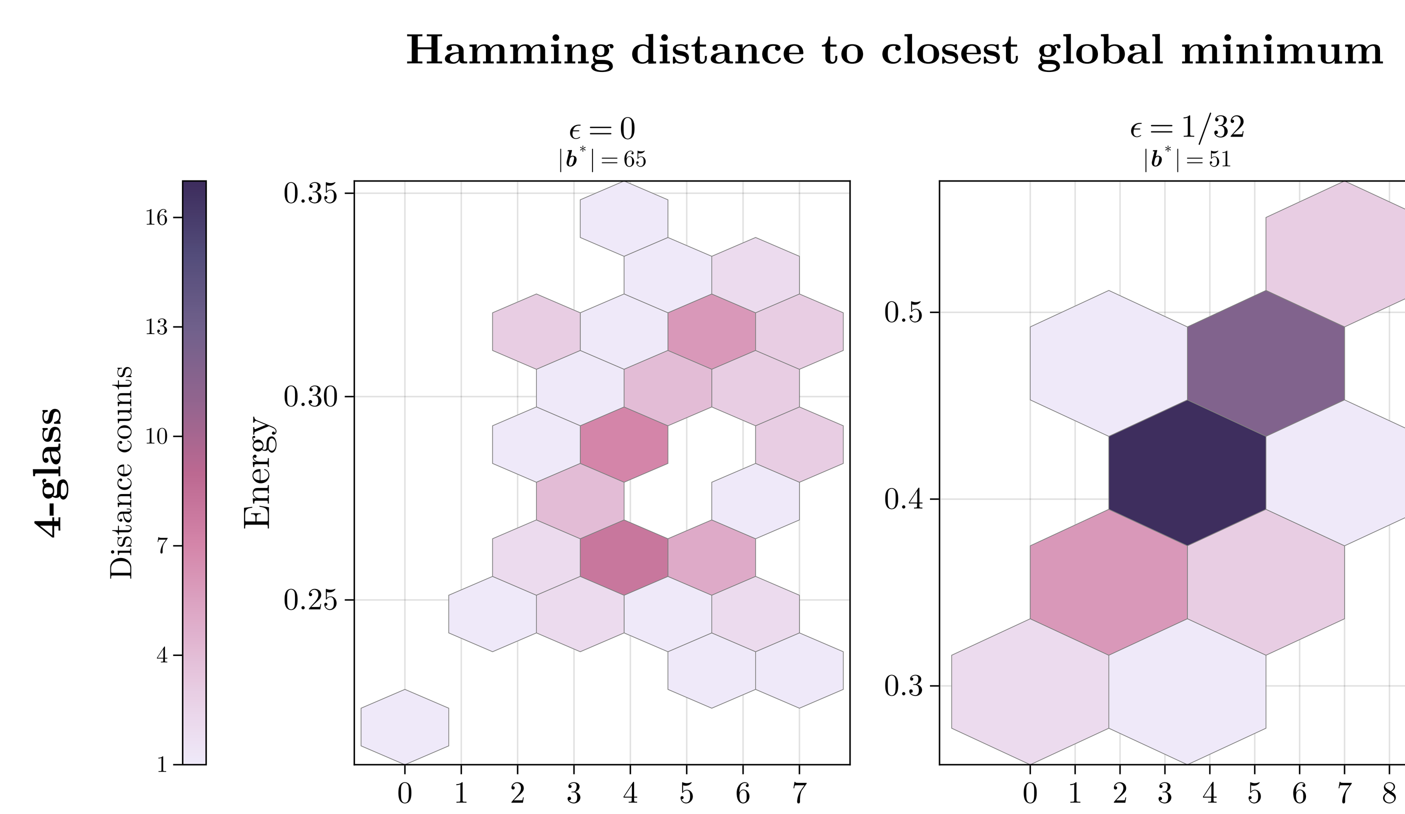

Figure 2: Hexagonal binned plot of the Hamming distance from all local optima to their closest global optimum of the 4-glass dataset. Each bin aggregates distance counts, where a darker shade means more local optima are at that given distance to the global optimum, hinting at a structure containing 'big valleys'.

Conclusion and Future Work

This work improves the understanding of the **multimodal** nature of the **feature selection** problem by addressing how the **landscape changes under regularization**.

Some avenues for future work include carrying out similar analyses on the remainder of the datasets, as well as studying different machine learning methods (other than decision trees). Combining the analysis with other landscape features (including ruggedness and deception) is also a possibility.

Download the datasets

Parallel Problem Solving From Nature 2024
September 14–18, 2024, Hagenberg, Austria

Coded with julia