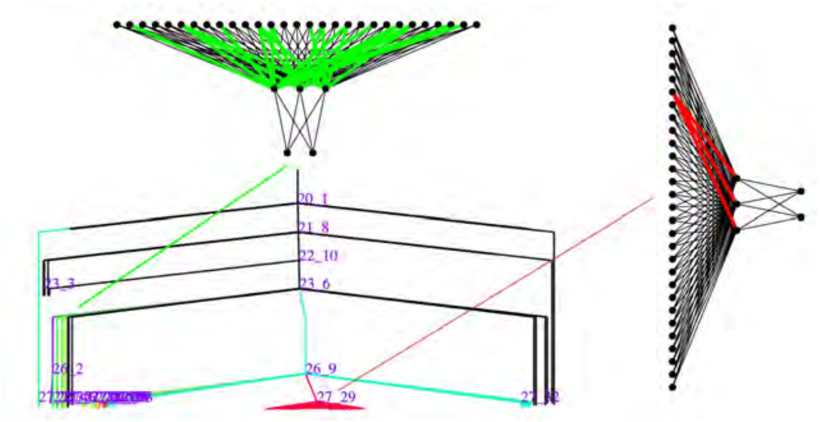
Some Applications of Energy Landscapes in Machine Learning

Maximilian Niroomand

Downing College
University of Cambridge

June 6, 2023


20_1
21_8
22_10
23_3
23_6
26_2
26_9
27_29

Research interest and motivation

Problem setting

- We can use the energy landscapes approach to study ML just like molecules
  - Energy function = loss function
- ML minimises loss function to find best parameters to explain relationship between data (X) and outcome (y)

Why use energy landscapes? 2 central questions in ML:

- Can we explain ML better (move away from 'black-box')?
- Can we improve learning (faster and/or more accurate)?

Loss landscapes for ML

Current state of research in ML-LLs

- Not too many people look into LFLs for ML in general
- ML practitioners use standard python packages (PyTorch)
  - Single minimisation, find one minimum, done
  - Surprisingly works well, and is much faster than computing whole LFL
  - Fast, easy to use, yet hard to explain **why** it works

(Selected) related work

- Loss landscapes have been used to:
  - Study optimisation methods [1, 2]
  - Explain why single minimum is often sufficient [3–5]
  - Improve accuracy [1, 6, 7]

Machine Learning methods

Neural networks

- Standard ML method
- Learn parameters for complex non-linear function
- Network architecture: design choice
  - Number and depth of layers

Gaussian Processes

- Non-parametric Bayesian ML
- Learn hyperparameters for covariance kernel
- Allows construction of confidence interval around prediction
- Scales $\mathcal{O}(n^3)$ for n datapoints

GP details

Bayes theorem allows

$$\begin{aligned}\mu(\mathbf{x}_*) &= \boldsymbol{\Sigma}_{\mathbf{x}_*,\mathbf{x}}(\boldsymbol{\Sigma}_{\mathbf{x},\mathbf{x}} + \sigma^2 \mathrm{I})^{-\mathbf{1}}\mathbf{y} \\ \sigma^2(\mathbf{x}_*) &= \boldsymbol{\Sigma}_{\mathbf{x}_*,\mathbf{x}_*} - \boldsymbol{\Sigma}_{\mathbf{x}_*,\mathbf{x}}(\boldsymbol{\Sigma}_{\mathbf{x},\mathbf{x}} + \sigma^2 \mathrm{I})^{-\mathbf{1}}\boldsymbol{\Sigma}_{\mathbf{x},\mathbf{x}_*}\end{aligned} \tag{1}$$

for new (test) datapoint $\mathbf{x}_*$.

Loss function

$$\log p(\mathbf{y}|\mathcal{X},\theta) = -\frac{1}{2}\mathbf{y}^\intercal \boldsymbol{\Sigma}^{-1}\mathbf{y} - \frac{1}{2}\log|\boldsymbol{\Sigma}| - \frac{n}{2}\log 2\pi, \tag{2}$$

Matern kernel

$$k_{ij} = \sigma^2 \frac{1}{\Gamma(\nu)2^{\nu-1}}\left(\frac{\sqrt{2\nu}}{\ell}d\left(x_i,x_j\right)\right)^{\nu} K_\nu\left(\frac{\sqrt{2\nu}}{\ell}d\left(x_i,x_j\right)\right) \tag{3}$$

Energy Landscapes vs Loss Landscapes

Landscape characteristics:

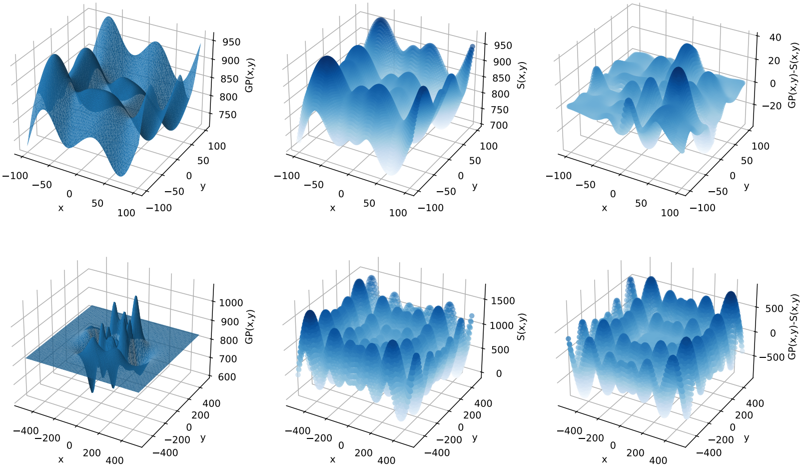
Feature	molecular PES	ML LL
energy	potential energy	loss value
temperature	physical temperature	fictitious parameter
coordinates	atomistic coordinates	weights/hyperparameters
local minimum	locally-stable molecular isomer	locally optimal weights
global minimum	energetically most favourable molecular isomer	best weights for given loss function

Landscape metrics:

Feature	molecular PES	ML LL
basin volume	entropic contribution to occupation probability	connection to robustness
heat capacity	change in occupied minima as a function of temperature	identification of minima with complementary properties

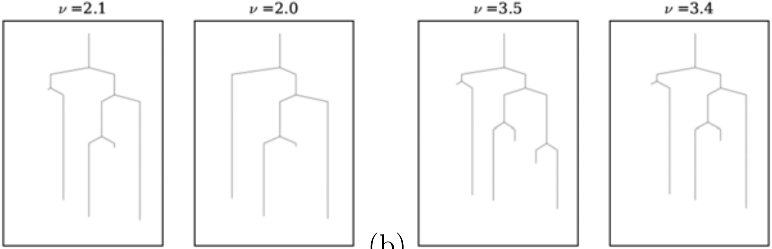
Schwefel function and GP fits

$$f(x) = 418.9829d - \sum_{i=1}^{d} x_i \sin\sqrt{|x_i|}$$



GP loss landscapes exhibit interesting structures

Catastrophe theory:

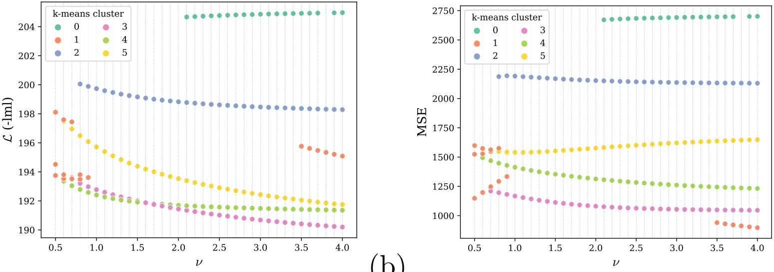


(a) (b)

Figure 2: Loss landscape fold catastrophes illustrated in disconnectivity graphs. Each leaf node of the graphs is a minimum in hyperparameter space Θ. The disappearance of minima from ν=2.1 $\rightarrow$ ν=2.0 (a) and ν=3.5 $\rightarrow$ ν=3.4 (b) corresponds to a reduction in the number of leaf nodes.

GP loss landscapes exhibit interesting structures II

ν-continuity



(a) (b)

Figure 3: Loss value (-lml) (a) and mean squared error (b) for all identified minima at all values of ν. k-means clustering is performed on the minima, where k is chosen to be the maximum number of minima at any ν. The data shown here were obtained from the $3d$ Schwefel function.

GP ensembles

Ensembles

- 2 heads are better than one
- Commonly used approach in ML field
- Combine multiple, orthogonal predictors for improved accuracy

Landscape ensembles

- Don't rerun same model from different initialisations
- Exploit multi-funneled landscapes and use unique minima
- Avoid single point estimate for solution

GP ensembles substantially improve accuracy

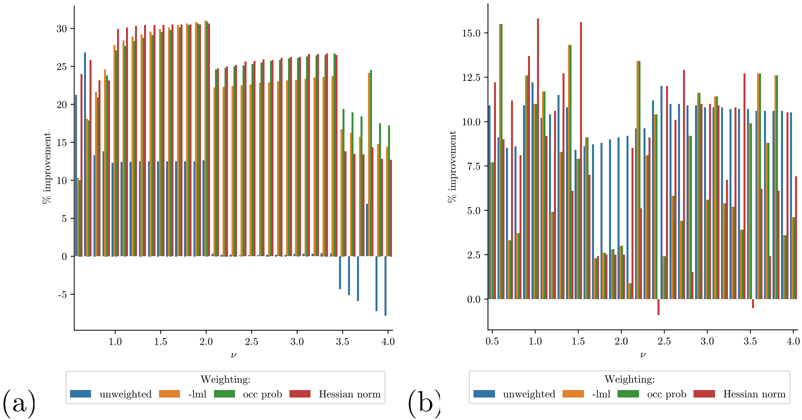


Figure 4: Percentage improvement over single best minimum of GP ensembles for fitting the (a) $3d$ and (b) $4d$ Schwefel function for various parameterisations of ν.

How can landscapes make GP learning more Bayesian

Hyperparameter sampling

- Single-point estimates are not Bayesian
- SPE is standard practice for SGD optimisation
- Bayesian treatment would be sampling hyperparameters from distribution
- Current methods use HMC: extremely show and expensive

Landscape approach

- Look at landscape to decide whether to run HMC?
- Or sample from reconstructed landscape?
- Perhaps only useful if multi-funneled?

Ensemble effectiveness increases when landscape has more minima

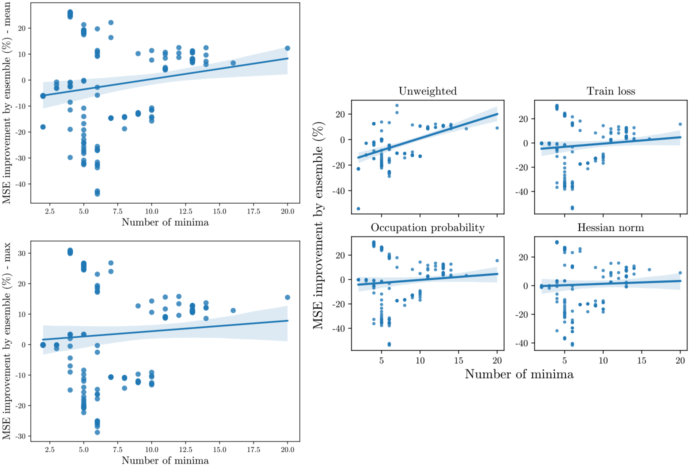
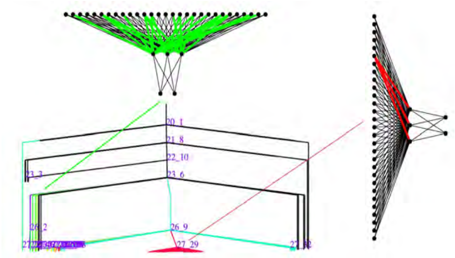


Figure 5: Correlation between MSE improvement achieved by ensemble methods and the number of minima in a loss landscape.

Conserved weights can help to interpret ML



- Identify conserved weights between groups of minima
- Are conserved weights those that are relevant to classification?
- Changing conserved weights seems to suggest that (accuracy decreases)

ML loss landscapes exhibit interesting properties and are not well explored

Landscapes and ML

- Various analogies between ELs and ML-LLs
- Methods can be used to improve interpretability and accuracy
- ML-LLs exhibit various interesting properties
- Understanding loss landscape crucial to understand system
- → Learn why machine learning works so well

Future work

- Reconstruct landscapes for fully Bayesian treatment
- Identify further analogies between ELs and ML-LLs

Acknowledgements

Thanks to great collaborators and supervisors:

- Prof David Wales
- Dr Edward Pyzer-Knapp
- Dr Luke Dicks
- Dr John Morgan
- All other Wales group members

Relevant readings I

[1] Pratik Chaudhari et al. "Entropy-sgd: Biasing gradient descent into wide valleys". In: *Journal of Statistical Mechanics: Theory and Experiment* 2019.12 (2019), p. 124018.

[2] Sepp Hochreiter and Jürgen Schmidhuber. "Flat minima". In: *Neural computation* 9.1 (1997), pp. 1–42.

[3] Felix Draxler et al. "Essentially no barriers in neural network energy landscape". In: *International conference on machine learning*. PMLR. 2018, pp. 1309–1318.

[4] Ian J Goodfellow, Oriol Vinyals, and Andrew M Saxe. "Qualitatively characterizing neural network optimization problems". In: *arXiv preprint arXiv:1412.6544* (2014).

Relevant readings II

[5] James Lucas et al. "Analyzing monotonic linear interpolation in neural network loss landscapes". In: *arXiv preprint arXiv:2104.11044* (2021).

[6] Maximilian P Niroomand et al. "On the capacity and superposition of minima in neural network loss function landscapes". In: *Machine Learning: Science and Technology* 3.2 (2022), p. 025004.

[7] Maximilian P Niroomand et al. "Characterising the area under the curve loss function landscape". In: *Machine Learning: Science and Technology* 3.1 (2022), p. 015019.