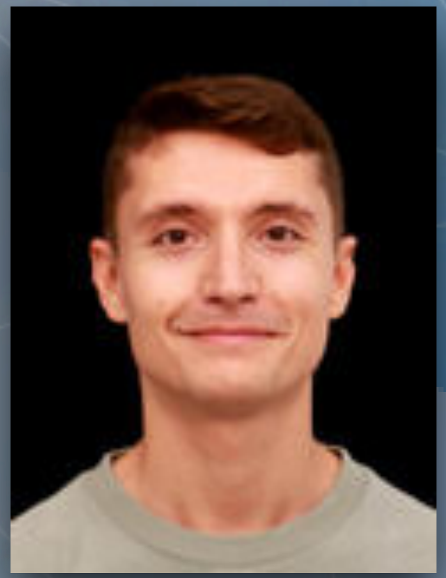
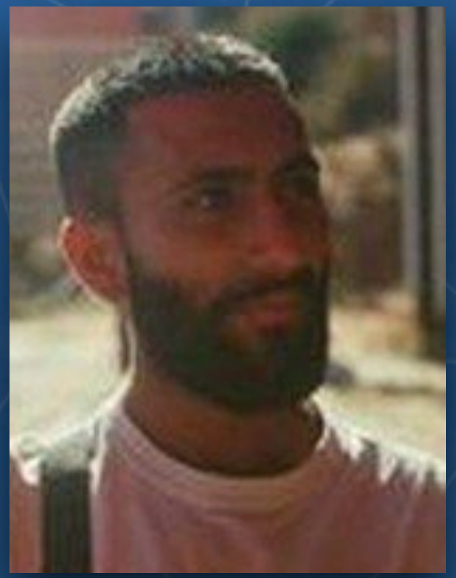
Only Strict Saddles in the Energy Landscape of Predictive Coding Networks?


Francesco Innocenti


El Mehdi Achour


Ryan Singh


Christopher L. Buckley

NEURAL INFORMATION PROCESSING SYSTEMS

NeurIPS 2024 (Main Track)

UNIVERSITY OF SUSSEX

sussex NEUROSCIENCE

RWTH AACHEN UNIVERSITY

TL;DR

Predictive coding inference seems to make the loss landscape of feedforward neural networks more benign and robust to vanishing gradients.

Overview

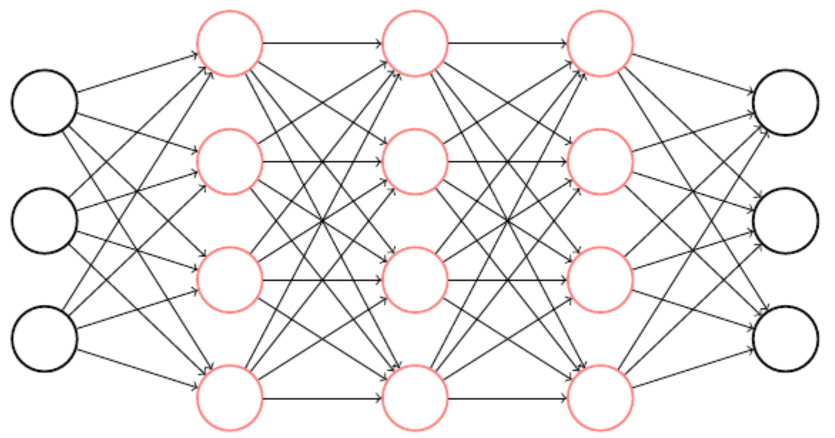
1. Introduction
2. Preliminaries
3. Theoretical results
4. Experiments
5. Conclusion

Overview

1. **Introduction**
2. Preliminaries
3. Theoretical results
4. Experiments
5. Conclusion

Introduction: **predictive coding**

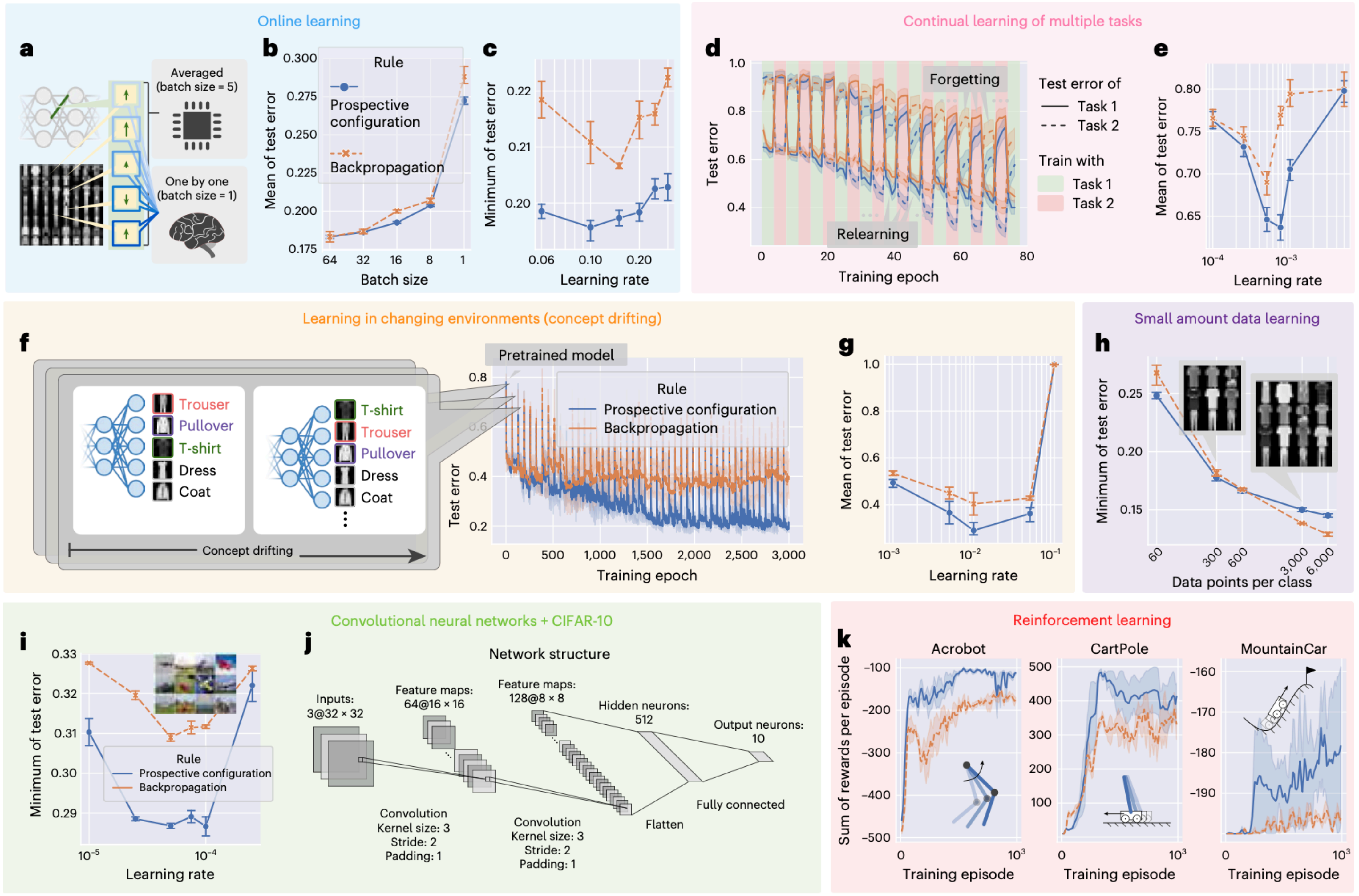
- **Predictive coding** (PC) is a brain-inspired learning algorithm that can train deep neural networks (DNNs) as an alternative to backpropagation (BP)



- In contrast to BP, PC **iteratively infers network activities** before updating weights

- This incurs an extra compute cost, but it has been argued to provide many benefits such as **faster learning convergence** [Song et al. '22]

- However, these speed-ups are not always observed, and the impact of PC inference on learning is not theoretically well understood



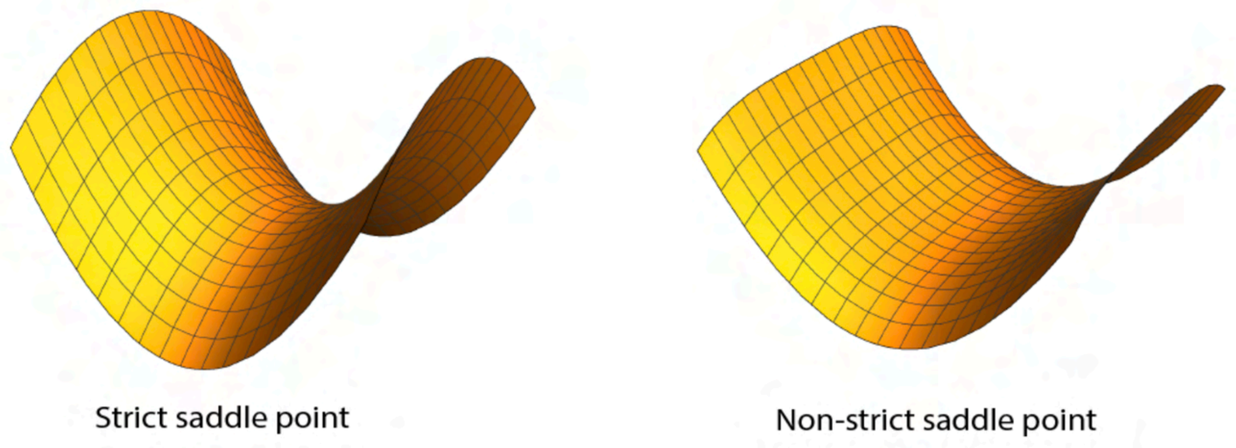
Introduction: **approach**

- To address this gap, we study the geometry of the effective landscape on which PC learns: ***the weight landscape at the equilibrium of the network activities***

- We focus on **saddle points** of the equilibrated energy

Introduction: **saddles & neural networks**

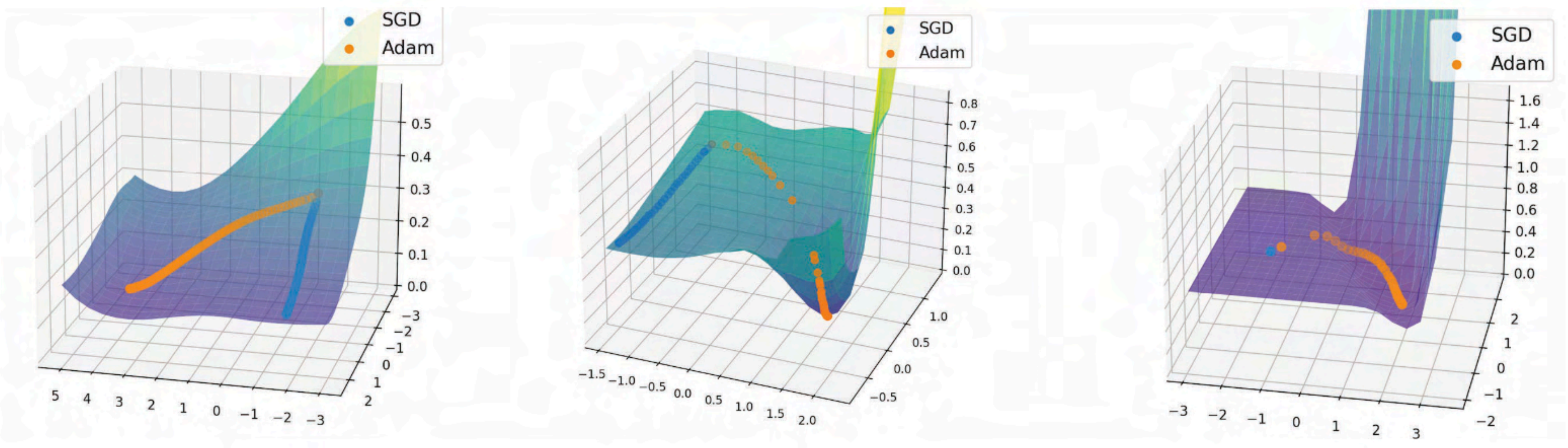
- Saddles are ubiquitous in the loss landscape of DNNs [Dauphin et al. '14]
- They have been characterised as [e.g. Get et al. '15]:

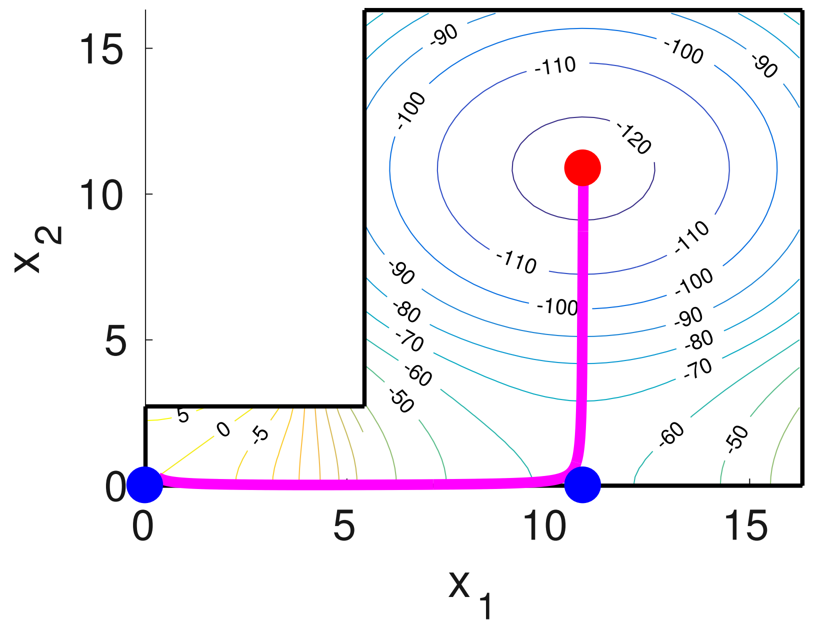


  i) "**Strict**", with negative curvature (indefinite Hessian), or

  ii) "**Non-strict**", where an escape (negative) direction is found in higher-order (n>2) derivatives

- Stochastic gradient descent (SGD) can be exponentially slowed by strict saddles [Du et al. '17] and effectively get stuck in non-strict ones [e.g. Böttcher & Wheeler '24]
  - (This is vanishing gradients from a landscape perspective [Orvieto et al. '22].)





Introduction: **contributions**

- For DLNs, we first show that, at the equilibrium of the network activities, the PC energy is equal to a rescaled mean squared error (MSE) loss with a weight-dependent rescaling
- We then prove that many highly degenerate (non-strict) saddles of the loss become much easier to escape (strict) in the equilibrated energy
- We empirically verify that our linear theory holds for non-linear networks
- We provide evidence that other non-strict saddles of the loss that we do not address theoretically also become strict in the equilibrated energy

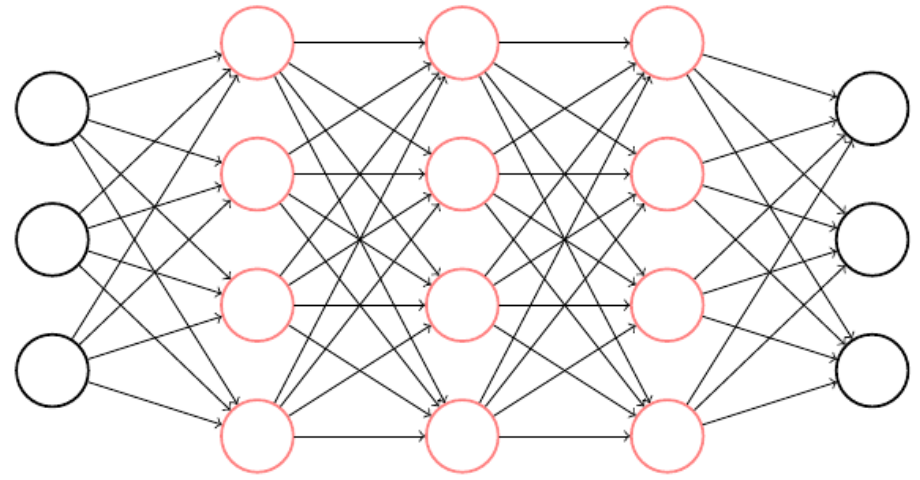
Overview

1. Introduction
2. **Preliminaries**
3. Theoretical results
4. Experiments
5. Conclusion

Preliminaries

- MSE loss for DLNs: $$\mathcal{L} = \frac{1}{2N}\sum_{i=1}^{N} ||\mathbf{y}_i - \mathbf{W}_{L:1}\mathbf{x}_i||^2$$
- PC energy for DLNs: $$\mathcal{F} = \frac{1}{2N}\sum_{i=1}^{N}\sum_{\ell=1}^{L} ||\mathbf{z}_{\ell,i} - \mathbf{W}_{\ell}\mathbf{z}_{\ell-1,i}||^2$$
- Minimised in 2 phases:

Inference: $\Delta \mathbf{z}_\ell \propto -\frac{\partial \mathcal{F}}{\partial \mathbf{z}_\ell}$ *Learning:* $\Delta \mathbf{W}_\ell \propto -\frac{\partial \mathcal{F}}{\partial \mathbf{W}_\ell}$



- In practice, inference is run to convergence until $\Delta \mathbf{z}_\ell \approx 0$ before updating the weights
- Importantly, the **effective landscape** on which PC learns is the energy at the inference equilibrium $\mathcal{F}|_{\partial\mathcal{F}/\partial\mathbf{z}=0}(\boldsymbol{\theta})$ which we will abbreviate as $\mathcal{F}^*(\boldsymbol{\theta})$

Overview

1. Introduction
2. Preliminaries
3. **Theoretical results**
4. Experiments
5. Conclusion

Theoretical results: **equilibrated energy as rescaled MSE**

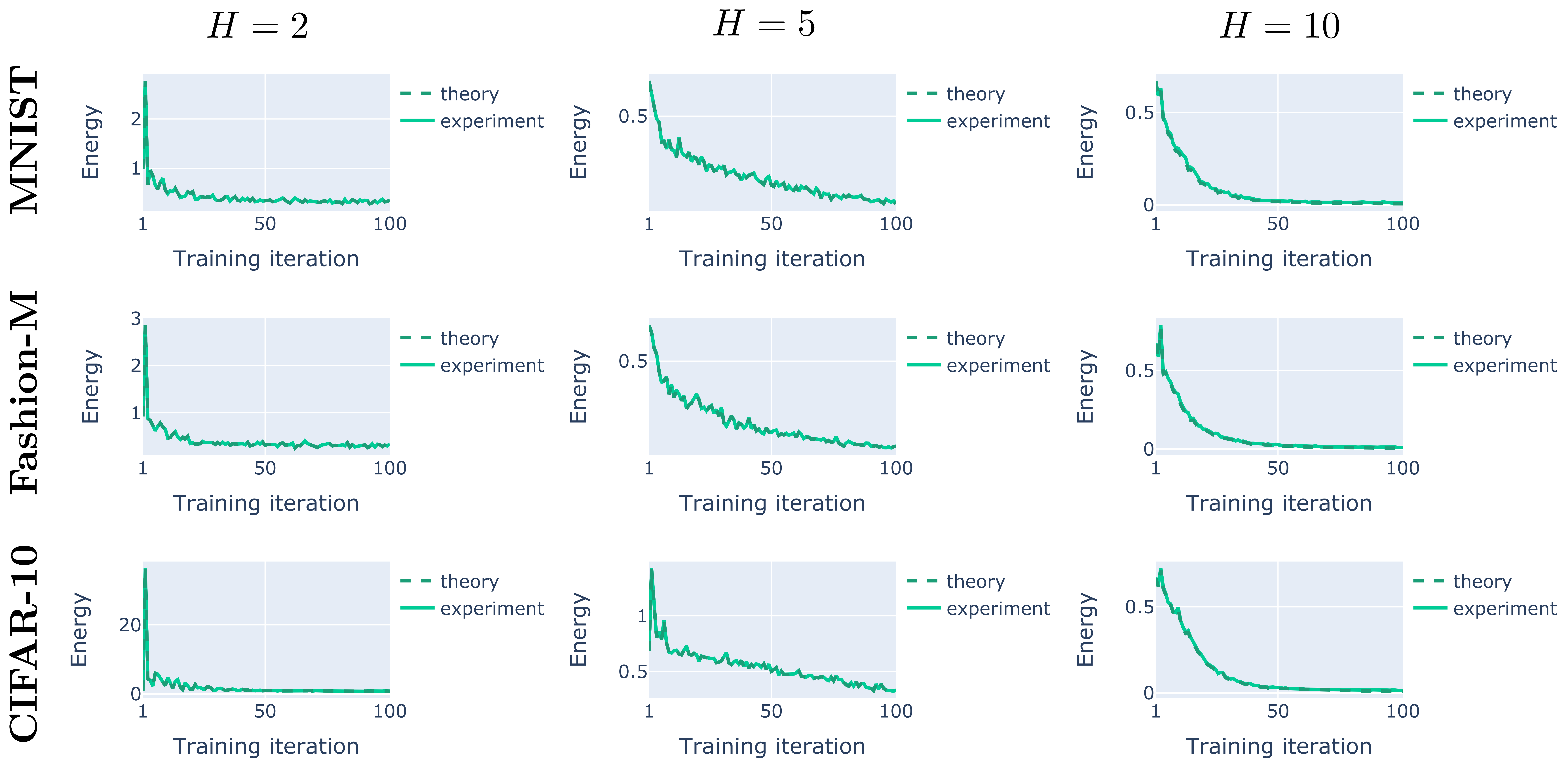
- At the inference equilibrium, the PC energy turns out to be equal to a rescaled MSE loss

Theorem 1 (Equilibrated energy for DLNs)**.** *For any DLN parameterised by* $\boldsymbol{\theta} := (\mathbf{W}_1, \dots, \mathbf{W}_L)$ *with input and output* $(\mathbf{x}_i, \mathbf{y}_i)$*, the PC energy (Eq. 2) at the exact inference equilibrium* $\partial \mathcal{F} / \partial \mathbf{z} = \mathbf{0}$ *is the following rescaled MSE loss (see §A.3.2 for derivation)*

$$\mathcal{F}^* = \frac{1}{2N} \sum_{i=1}^{N} (\mathbf{y}_i - \mathbf{W}_{L:1}\mathbf{x}_i)^T \mathbf{S}^{-1} (\mathbf{y}_i - \mathbf{W}_{L:1}\mathbf{x}_i) \tag{5}$$

where the rescaling is $\mathbf{S} = \mathbf{I}_{d_y} + \sum_{\ell=2}^{L} (\mathbf{W}_{L:\ell})(\mathbf{W}_{L:\ell})^T$.

Theoretical results: **equilibrated energy as rescaled MSE**



Theoretical results: **saddle analysis**

- Many highly degenerate (non-strict) saddles of the MSE loss become much easier to escape (strict) in the equilibrated energy

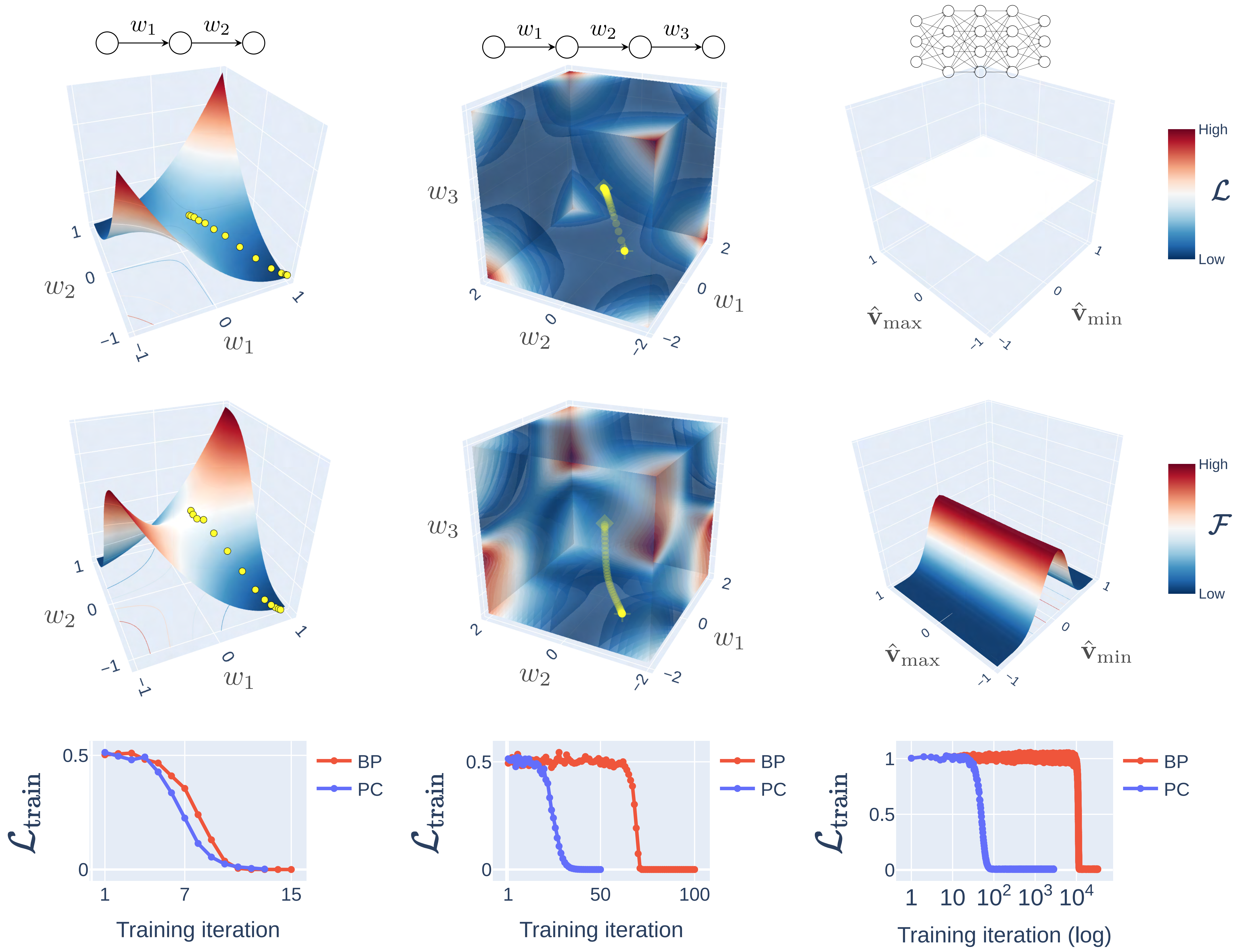
Theorem 3 (Strictness of zero-rank saddles of the equilibrated energy)**.** *Consider the set of critical points of the equilibrated energy (Eq. 5)* $\boldsymbol{\theta}^*(\mathbf{W}_L = \mathbf{0}, \mathbf{W}_{L-1:1} = \mathbf{0})$ *where* $\mathbf{g}_{\mathcal{F}^*}(\boldsymbol{\theta}^*) = \mathbf{0}$*. The Hessian at these points has at least one negative eigenvalue (see §A.3.6 for proof)*

$$\exists \lambda(\mathbf{H}_{\mathcal{F}^*}(\boldsymbol{\theta}^*)) < 0 \quad \textit{[strict saddles, Def. 1]} \tag{10}$$

- These saddles include the origin, effectively making PC more robust to vanishing gradients

Theoretical results: **saddle analysis**

- Toy examples illustrating the result for the origin saddle



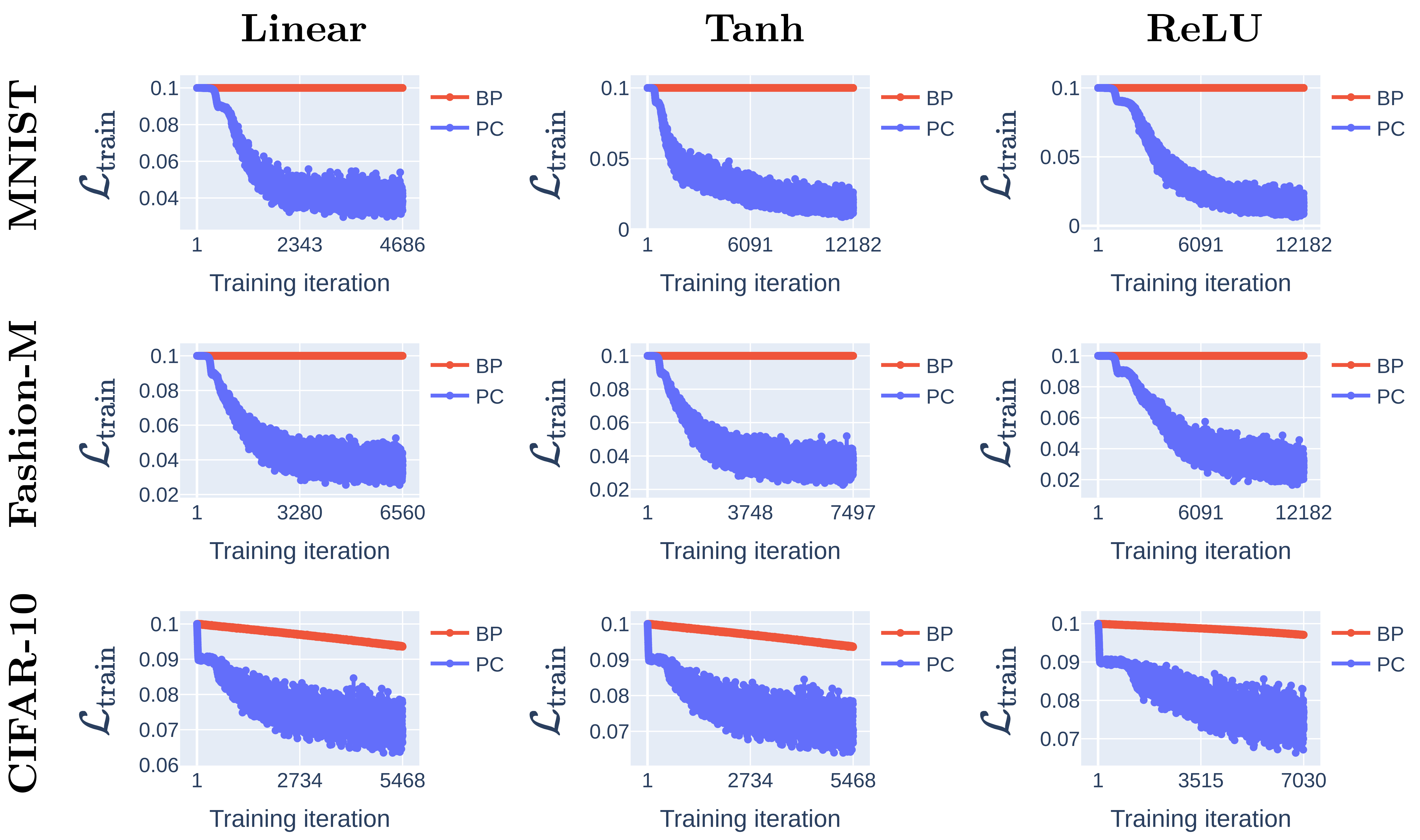
Overview

1. Introduction
2. Preliminaries
3. Theoretical results
4. **Experiments**
5. Conclusion

Experiments: **what about non-linear networks?**

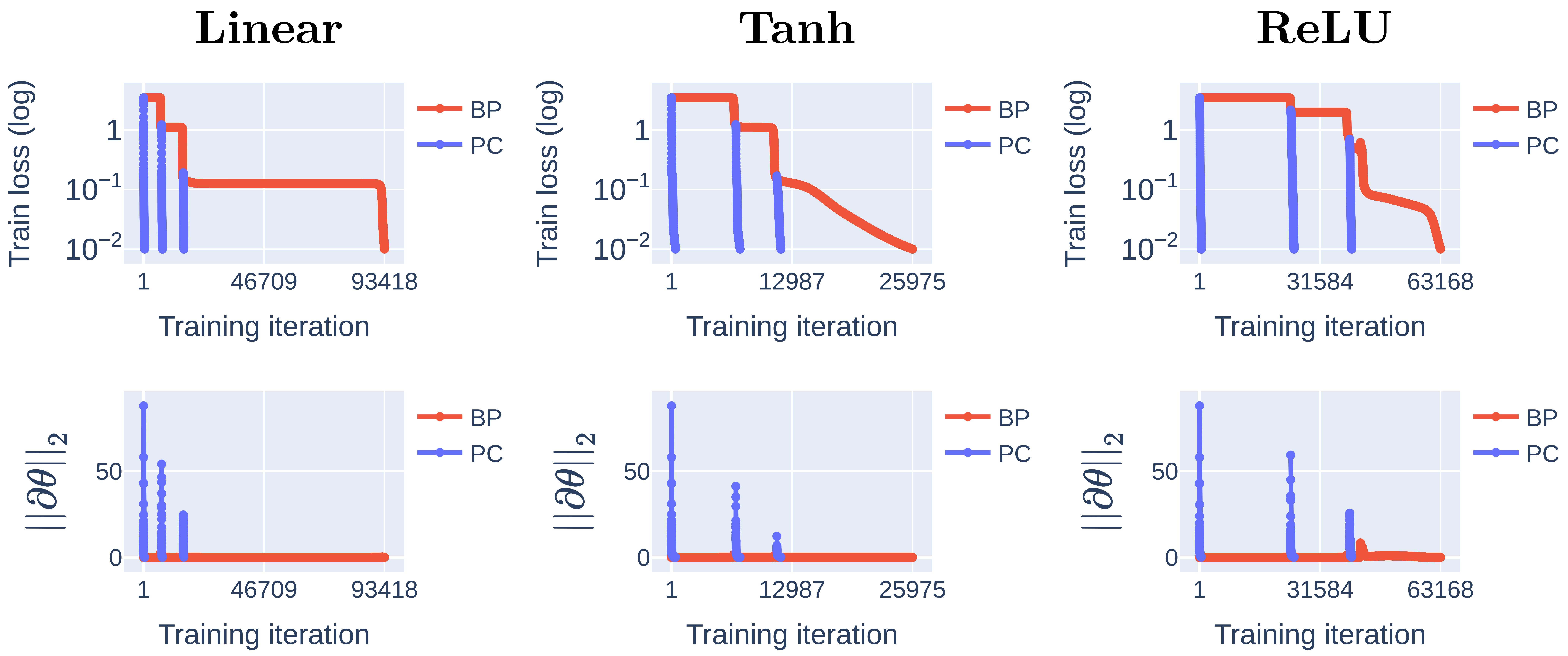
- To test the theory, we train various networks on standard datasets by initialising close to the considered saddles (e.g. origin)

- We find that, for the same learning rate, SGD on the equilibrated energy (PC) escapes much faster than on the loss (BP)



Experiments: **what about other saddles?**

- To test other non-strict saddles of the loss that we do not address theoretically, we train networks on a matrix completion task, where we know that starting near origin GD goes through these other saddles



Overview

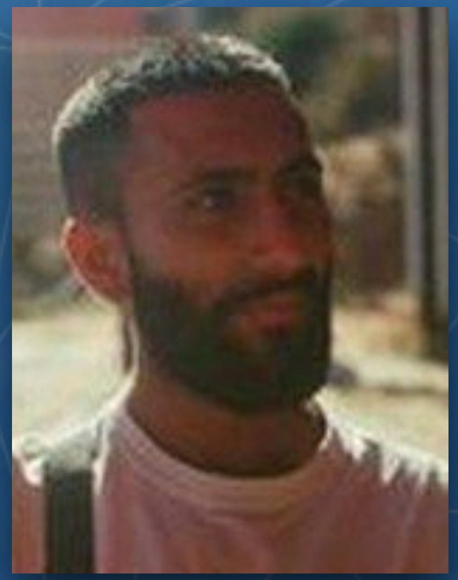
1. Introduction
2. Preliminaries
3. Theoretical results
4. Experiments
5. **Conclusion**

Conclusion

- **Summary:** we provided theoretical and empirical evidence that the effective landscape on which PC learns has only strict saddles and is more robust to vanishing gradients
- **Conjecture**: all the saddles of the equilibrated energy are strict
- **Conclusion**: our work suggests that PC inference makes the loss landscape of feedforward neural networks more benign or easier to navigate
- **Limitation**: inference convergence significantly slows down with network depth and remains a key challenge for scaling PC to large tasks

Thank you for your attention!

El Mehdi Achour



Ryan Singh

Christopher L. Buckley

NEURAL INFORMATION PROCESSING SYSTEMS

UNIVERSITY OF SUSSEX

sussex NEUROSCIENCE

RWTH AACHEN UNIVERSITY