Spikes in the training loss, catapults and feature learning

Libin Zhu, Chaoyue Liu, Adityanarayanan Radhakrishnan, Misha Belkin

Department of Computer Science and Halicioğlu Data Science Institute

University of California, San Diego

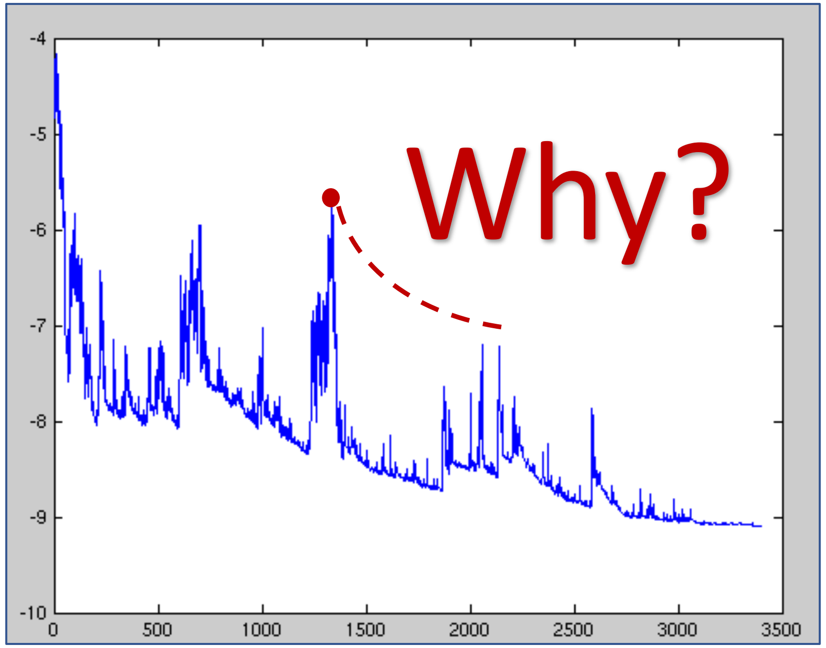
UC San Diego

HALICIOĞLU DATA SCIENCE INSTITUTE

Pretty common when training a neural network:

Training loss of SGD(Source: Wikipedia)

Y axis: **Training Loss**



X axis: **Iteration**

"If the learning rate is too large, the learning curve will show violent oscillations, with the cost function often increasing significantly."

-- *Goodfellow et al. 2016 Deep learning.*

Small mini-batch size leads to better generalization

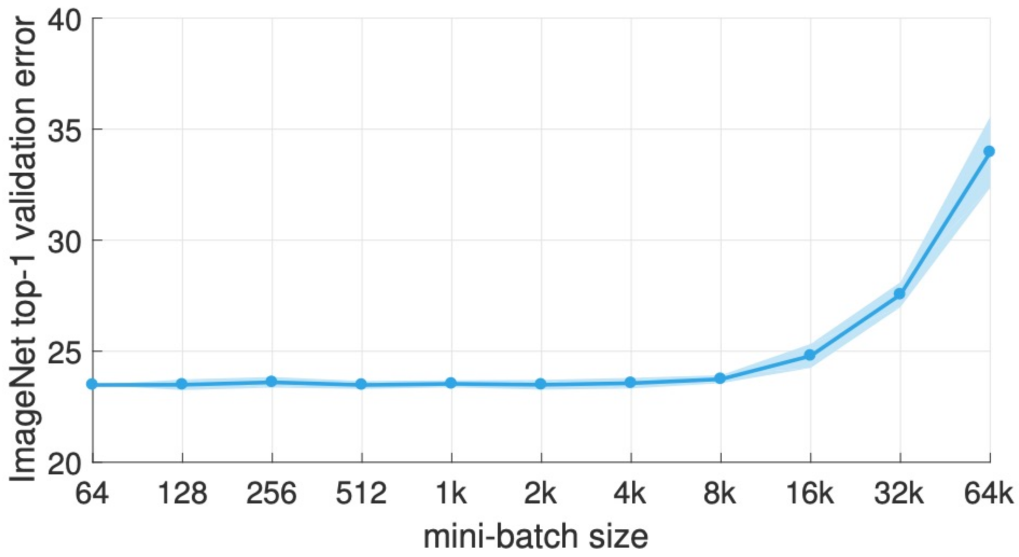


Figure 1. **ImageNet top-1 validation error *vs*. minibatch size.**

Figure from Goyal et al. 2017

"The lack of generalization ability is due to the fact that large-batch methods tend to converge to sharp minimizers of the training function…. In contrast, small-batch methods converge to flat minimizers…
-- *Keskar et al. ICLR 2017*

AGOP: A feature learning measurement

AGOP(Average Gradient Outer Product) [Triveti et al. 2014],[Xia et al. 2002]…

For a parameterized model $f(w;\cdot)$: $\mathbb{R}^{p\times d} \to \mathbb{R}$, the AGOP of it w.r.t. n input data $\mathrm{x}_1, \mathrm{x}_2, \dots, x_n \in \mathbb{R}^d$ is

$$G(\mathbf{w}) = \frac{1}{n}\sum_{i=1}^{n} \frac{\partial f(\mathbf{w};\boldsymbol{x}_i)}{\partial \boldsymbol{x}_i} \frac{\partial f(\mathbf{w};\boldsymbol{x}_i)}{\partial \boldsymbol{x}_i}^T \in \mathbb{R}^{d\times d}$$

Example: $f^*(x) = x_1 x_2$ with $x \in \mathbb{R}^{100}$

$$\frac{\partial f^*(x)}{\partial x} = (x_2, x_1, 0, \dots, 0) \qquad \frac{\partial f^*(x)}{\partial x}\left(\frac{\partial f^*(x)}{\partial x}\right)^T = \begin{pmatrix} x_2^2 & x_1 x_2 & \cdots & 0 \\ x_1 x_2 & x_1^2 & \cdots & 0 \\ \vdots & & \ddots & \vdots \\ 0 & 0 & \cdots & 0 \end{pmatrix}$$

For n samples i.i.d. from $N(0, I_{100})$, the AGOP of $f^*(x)$ will converge to $\begin{pmatrix} I_2 & 0 \\ 0 & 0 \end{pmatrix}$ as $n \to \infty$.

$$G_{f^*} = \frac{1}{n}\begin{pmatrix} \sum_{i=1}^{n}(x_i)_2^2 & \sum_{i=1}^{n}(x_i)_1(x_i)_2 & \cdots & 0 \\ \sum_{i=1}^{n}(x_i)_1(x_i)_2 & \sum_{i=1}^{n}(x_i)_1^2 & & 0 \\ \vdots & & \ddots & \vdots \\ 0 & 0 & \cdots & 0 \end{pmatrix}$$

Connecting loss spikes, generalization and feature learning

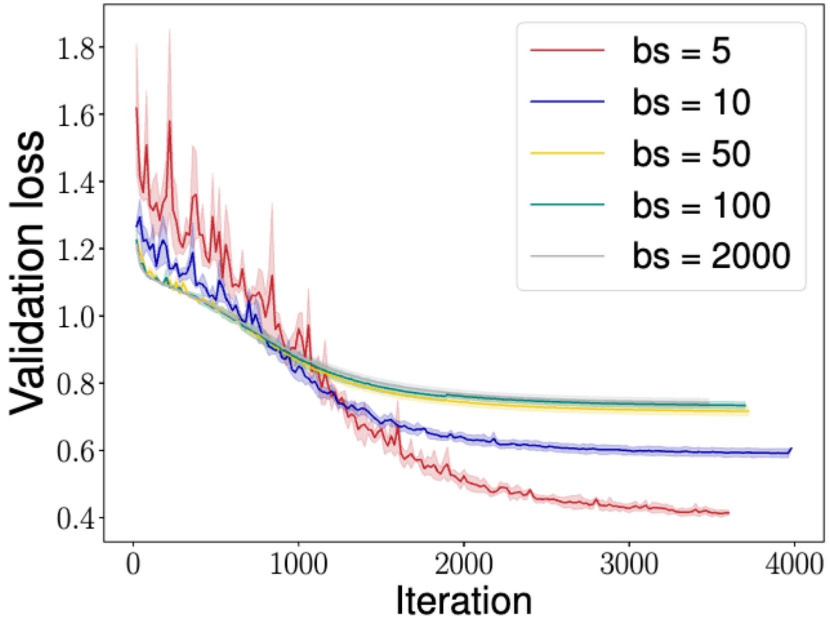


Figure: Smaller SGD batch size has more loss spikes and better generalization. Rank-2 dataset is used ($f^*(x) = x_1 x_2$).

Feature learning

Batch size	AGOP alignment	Test loss
2000 (GD)	0.81	0.74
50	0.84	0.71
10	0.89	0.59
5	0.95	0.42

Table: Smaller SGD batch size leads to a higher (better) AGOP alignment and smaller (better) test loss.

AGOP alignment: $\cos(G, G^*) = \frac{\langle G, G^* \rangle}{\|G\|_F, \|G^*\|_F}$,
where G, G^* are the empirical and true AGOP.

Outline:

1. **Catapult dynamics**
2. Catapults in SGD: spikes in the training loss
3. Feature learning of catapults: alignment of Average Gradient Outer Product(AGOP)

[1] Zhu, Liu, Radhakrishnan, Belkin, Quadratic models for understanding neural network dynamics
[2] Zhu, Liu, Radhakrishnan, Belkin, Catapults in SGD: spikes in the training loss and their impact on generalization through feature learning

Linear training dynamics

Linear training dynamics (GD with constant step size): if $f(w;\cdot)$ is a **linear model**, minimizing squared loss $\frac{1}{2}\|F(w)-y\|^2$ using GD with constant learning rate η**:**

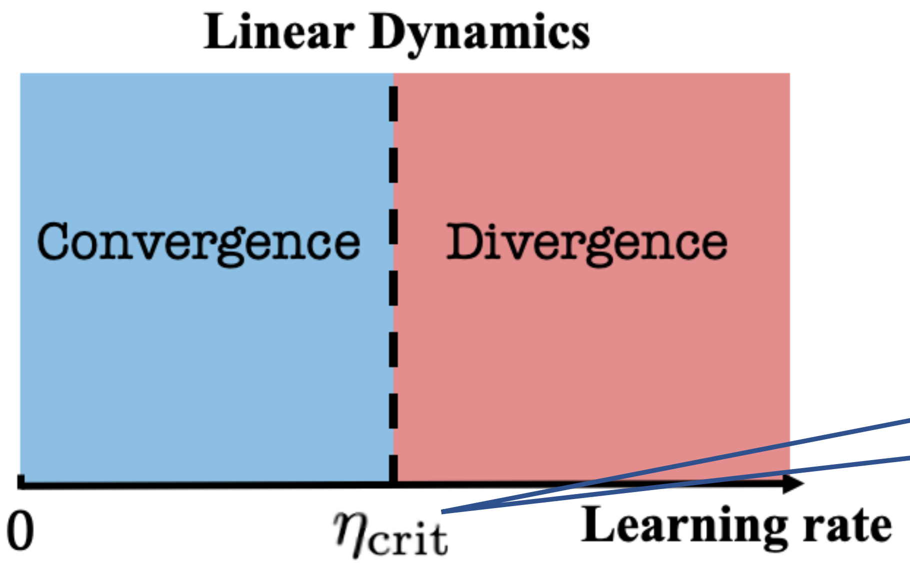
$$F(\mathbf{w}(t+1)) - \mathbf{y} = (I - \eta K(\mathbf{w}_0))(F(\mathbf{w}(t)) - \mathbf{y}), \quad \forall t \geq 0,$$

$$K(\mathbf{w}_0) = \left\langle \frac{dF(\mathbf{w}_0)}{d\mathbf{w}}, \frac{dF(\mathbf{w}_0)}{d\mathbf{w}} \right\rangle$$

$$F(\mathbf{w}(t)) = (f(\mathbf{w}(t); x_1), f(\mathbf{w}(t); x_2), \cdots, f(\mathbf{w}(t); x_n))^T$$

Convergence: $\|I - \eta K(\mathbf{w}_0)\| < 1$

Divergence: $\|I - \eta K(\mathbf{w}_0)\| > 1$



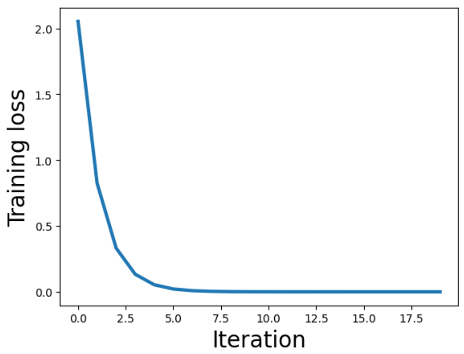
$$\frac{2}{\lambda_{\max}(K(\mathbf{w}_0))}$$

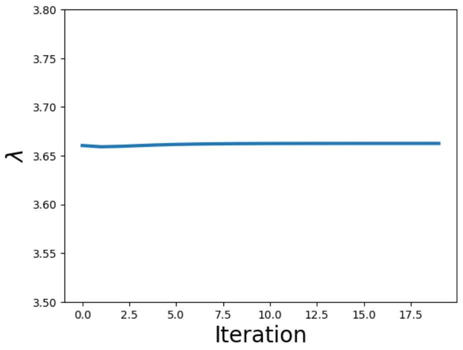
Optimization of Neural Networks

Neural tangent kernel(NTK) [Jacot et al. 2018] $K(\mathbf{w}) := \left\langle \frac{dF(\mathbf{w})}{d\mathbf{w}}, \frac{dF(\mathbf{w})}{d\mathbf{w}} \right\rangle \in \mathbb{R}^{n \times n}$

Linear dynamics: $\eta < \eta_{crit} \approx \frac{2}{\|K(w_0)\|_2}$:

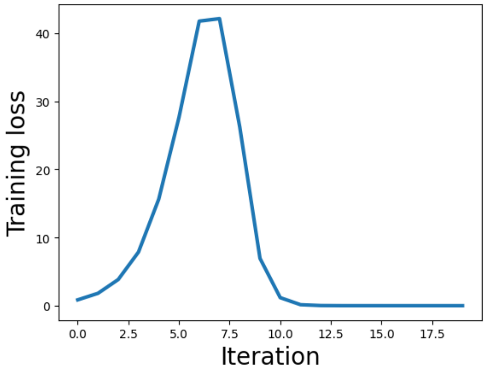
Loss $L(t)$ decreases to zero monotonically and the NTK $\lambda(t)$ almost does not change [Lee et al. 2019],[Liu, **Zhu**, Belkin 2022]…

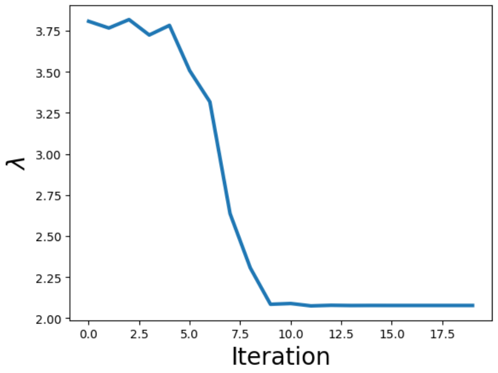




Catapult dynamics: $\eta_{crit} < \eta < \eta_{max}$:

Loss $L(t)$ increases then decreases, the largest eigenvalue of NTK $\lambda(t)$ decreases [Lewkowycz et al. 2020],[**Zhu**, Liu,Radhakrishnan,Belkin 2022].





Non-linear training dynamics of networks

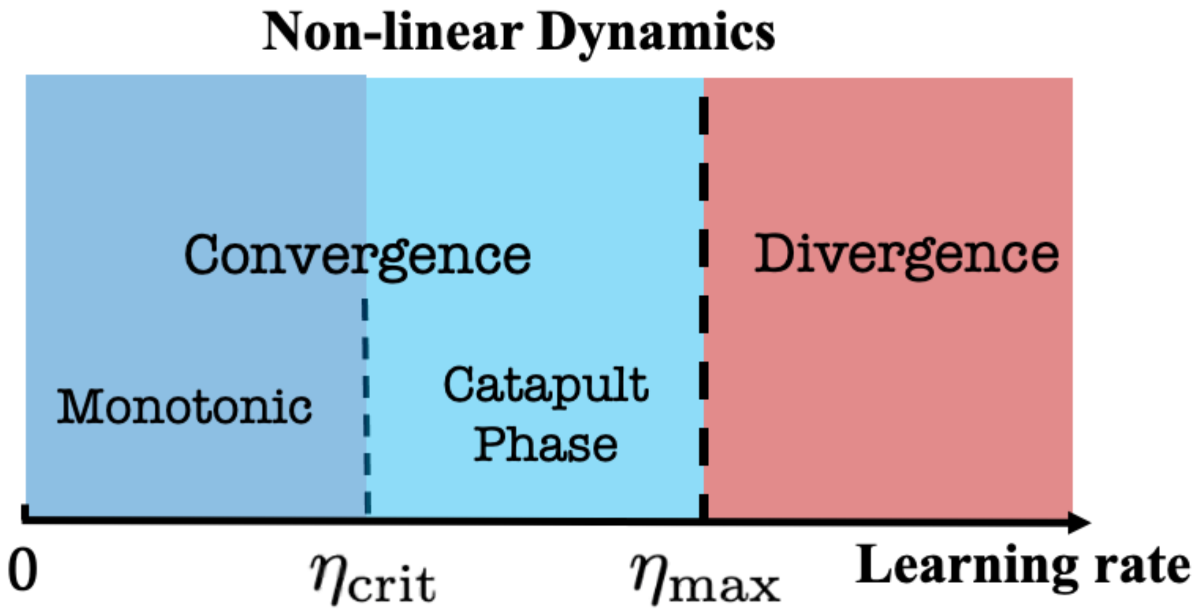
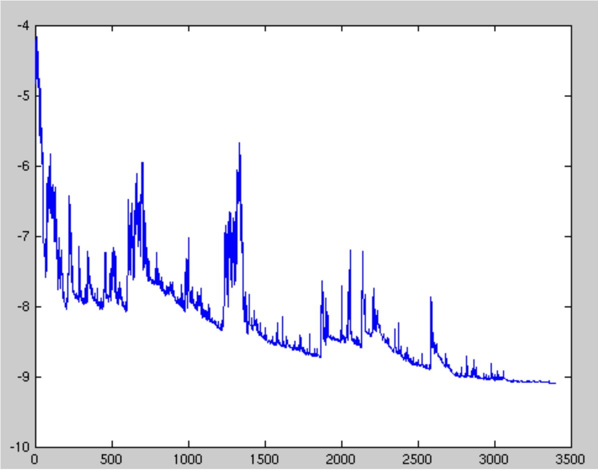


Figure: Neural networks exhibit non-linear dynamics
[Zhu, Liu,Radhakrishnan,Belkin 2022].

Outline:

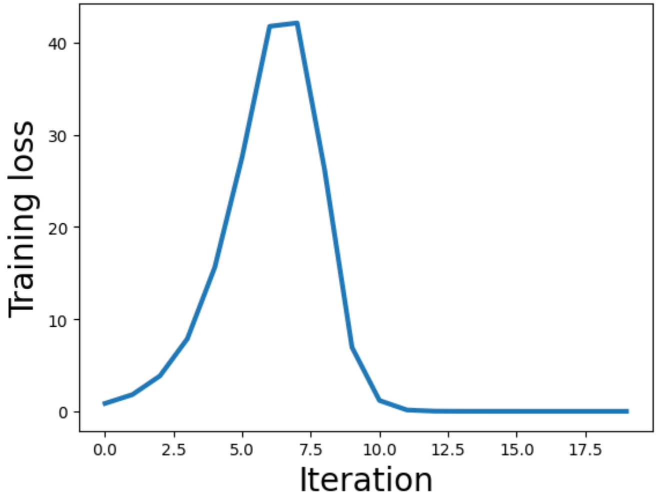
1. Catapult dynamics
2. **Catapults in SGD: spikes in the training loss**
3. Feature learning of catapults: alignment of Average Gradient Outer Product(AGOP)

Spikes in the training loss of SGD:



Loss spikes in SGD

Similar?



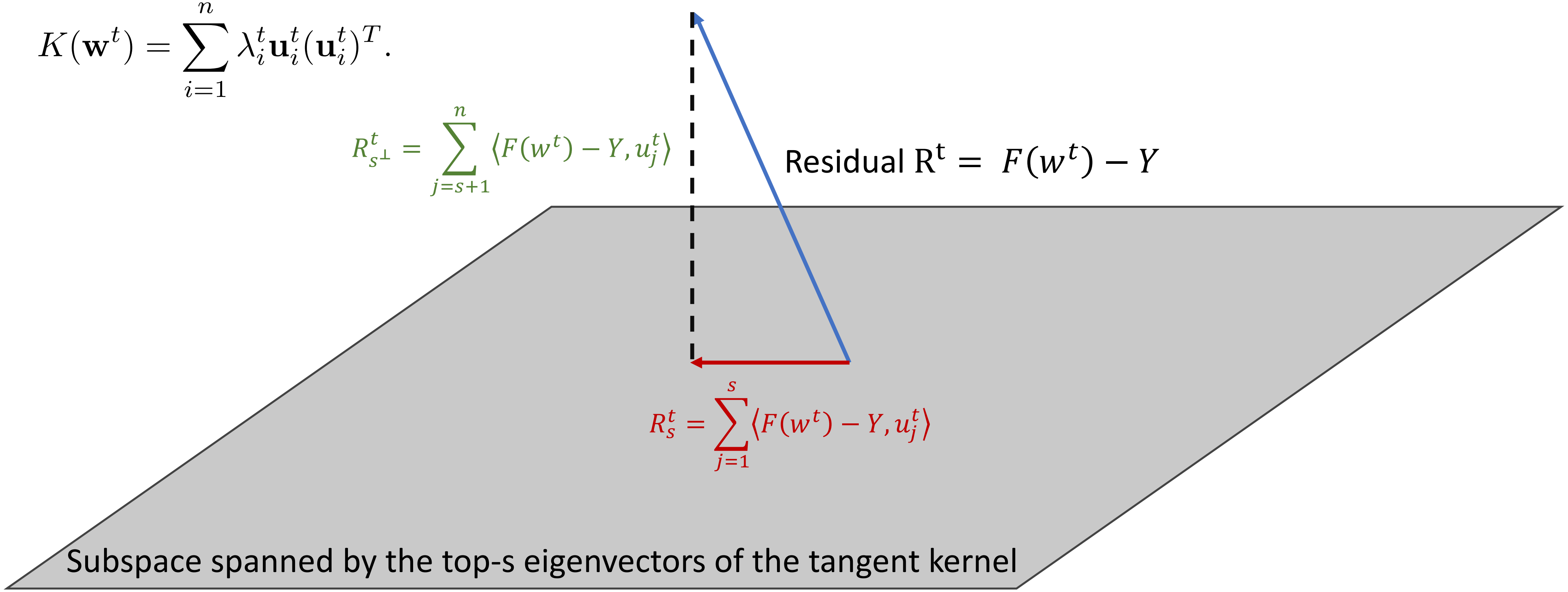
Catapult phenomenon in GD

-- Why does the loss drop so quickly at the peak of the spike?
-- Catapults occur in a low-dimensional subspace.

Loss decomposition based on the tangent kernel

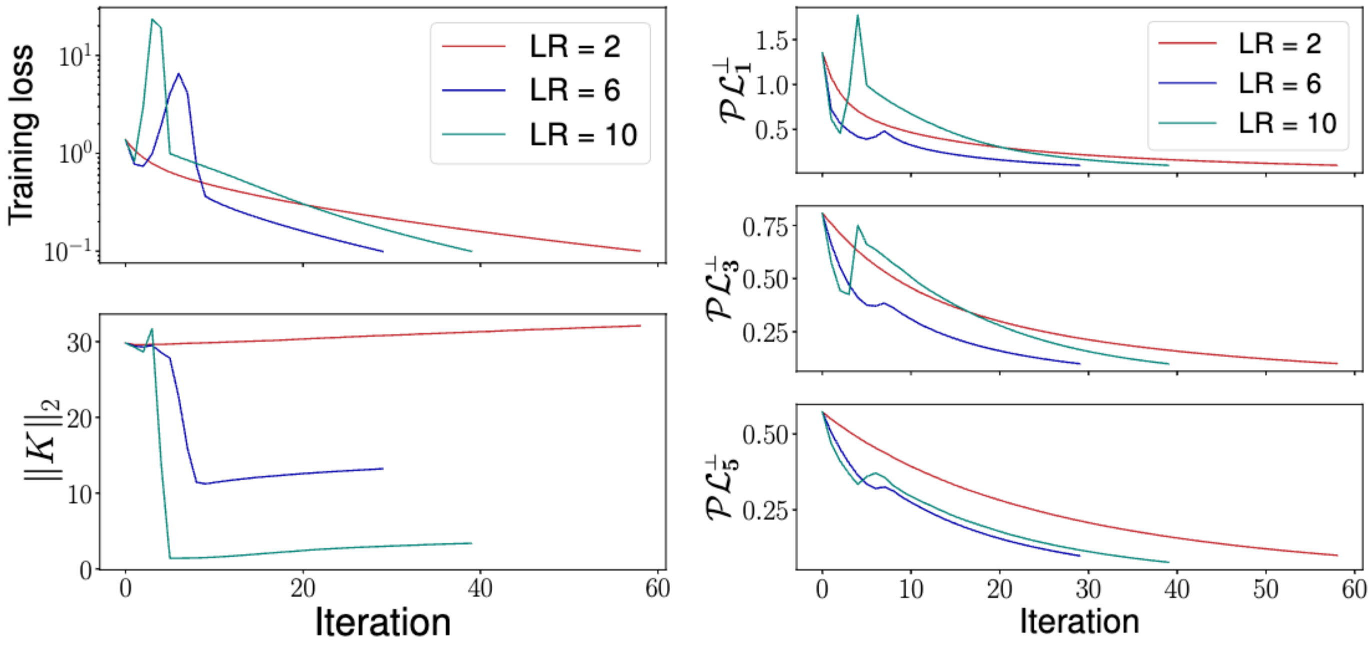
Eigen decomposition of the tangent kernel K^t

$$K(\mathbf{w}^t) = \sum_{i=1}^{n} \lambda_i^t \mathbf{u}_i^t (\mathbf{u}_i^t)^T.$$



Definition: $PL_s = \frac{1}{n}\|R_s^t\|^2$, $PL_s^\perp = \frac{1}{n}\left\|R_{s^\perp}^t\right\|^2$, hence Loss L $= PL_s + PL_s^\perp$.

Spikes occur in the top eigendirection:



(a) Training loss(FCN) (b) Loss decomposition

Figure: Training a 5-layer Fully Connected Neural network (FCN) on 128 data points from CIFAR-10. The networks are trained by GD with a constant learning rate.

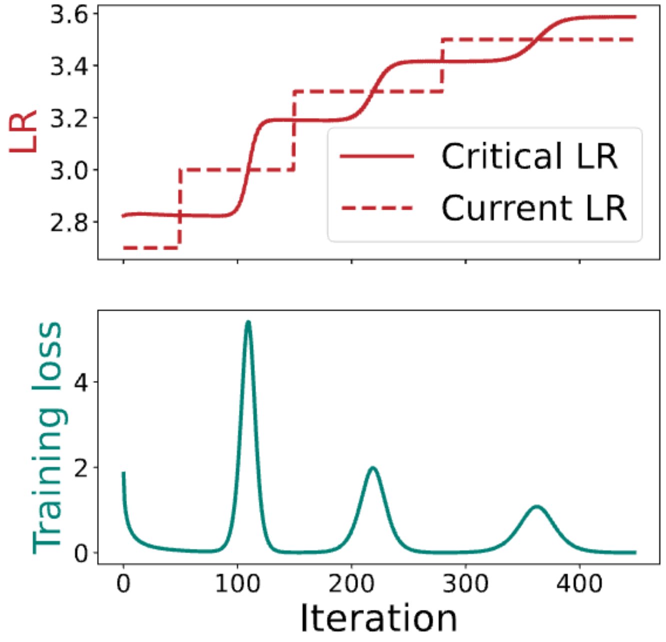
Multiple catapults with increasing learning rates

Recall that $\eta_{crit} \approx \frac{2}{\lambda_{max}(K(w))}$. Therefore,

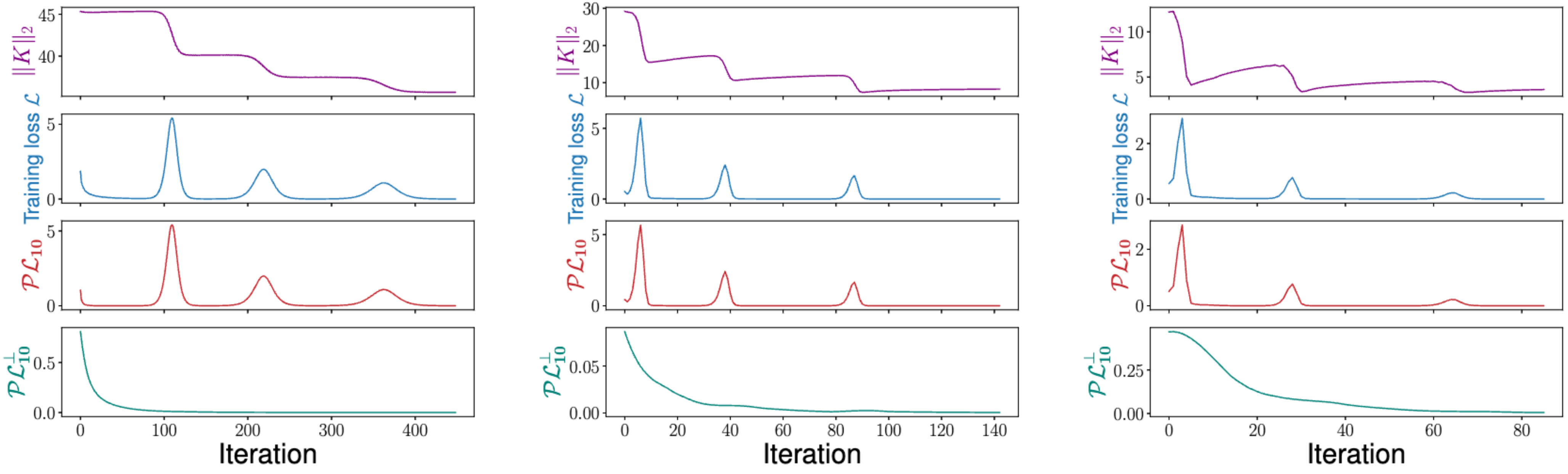
Multiple catapults:

$\eta > \eta_{crit}$ → Catapult occurs → η_{crit} increases

Increase η

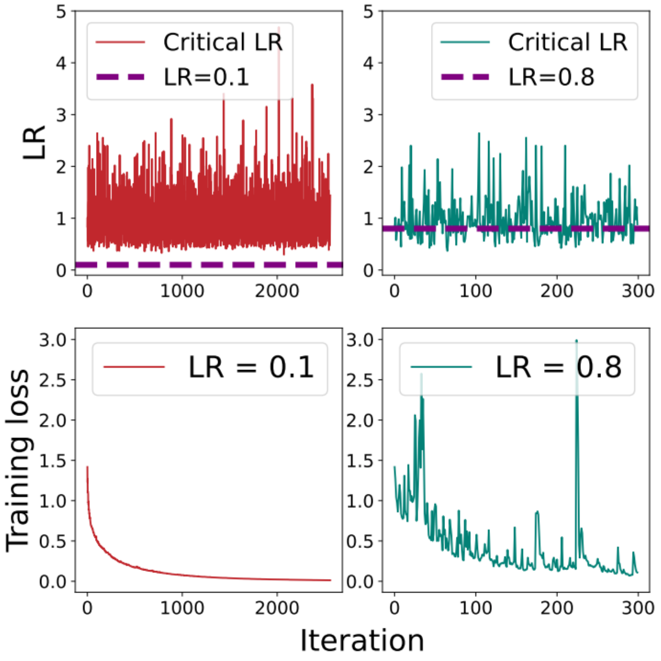


Multiple catapults with increasing learning rates

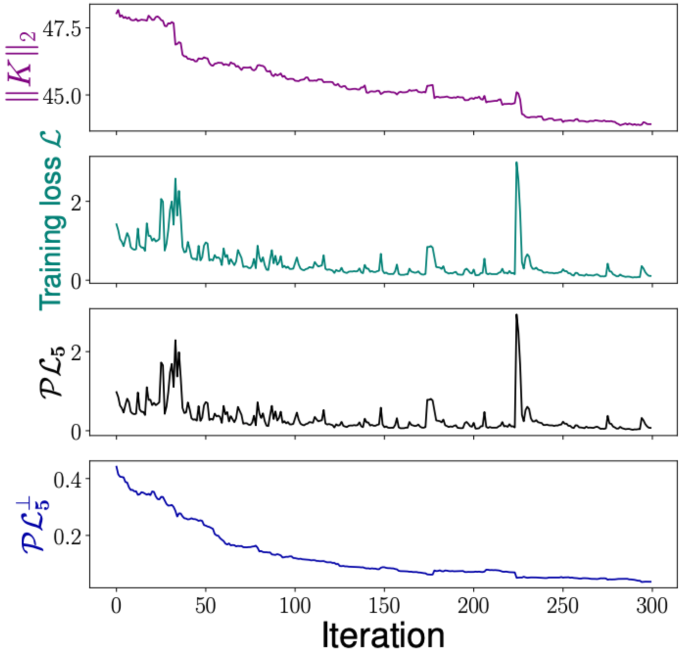


(a) Shallow network (b) 5-layer FCN (c) 5-layer CNN

Mechanism of catapults in SGD



(b) η_{crit} and training loss with $\eta = 0.1, 0.8$

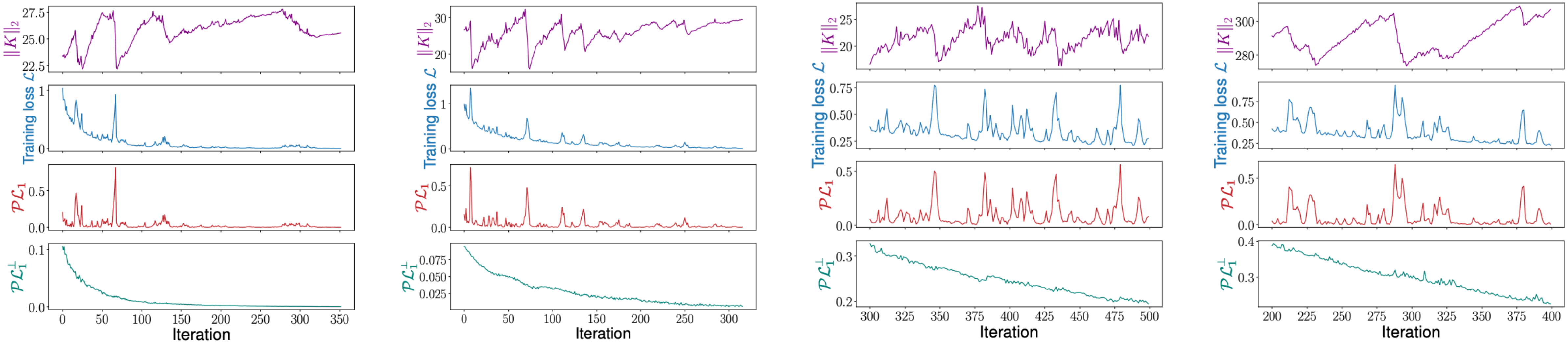


(c) Loss decomposed into $\mathcal{PL}_5$ and $\mathcal{PL}_5^{\perp}$ with $\eta = 0.8$

Observation:

1. Loss spikes occur when the LR is larger than the critical LR for each batch.
2. Loss spikes exist in the top eigenspace of the tangent kernel.

Mechanism of catapults in SGD



(a) 5-layer FCN (b) 5-layer CNN (c) WRN (zoomed-in) (d) ViT-4(zoomed-in)

Observation:

1. Each loss spike corresponds to a decrease in the ||K||.
2. Loss spikes exist in the top eigenspace of the tangent kernel.

Outline:

1. Catapult dynamics

2. Catapults in SGD: spikes in the training loss

3. **Feature learning of catapults: alignment of Average Gradient Outer Product(AGOP)**

Catapults improve generalization

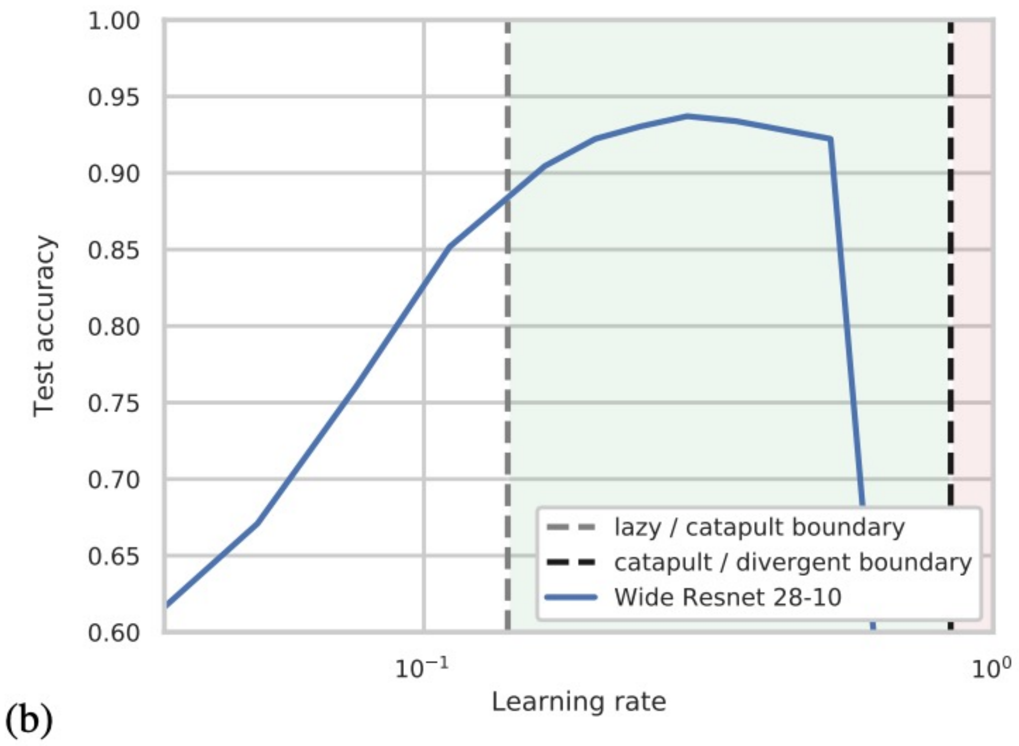


Figure from [Lewkowycz et al. 2020]

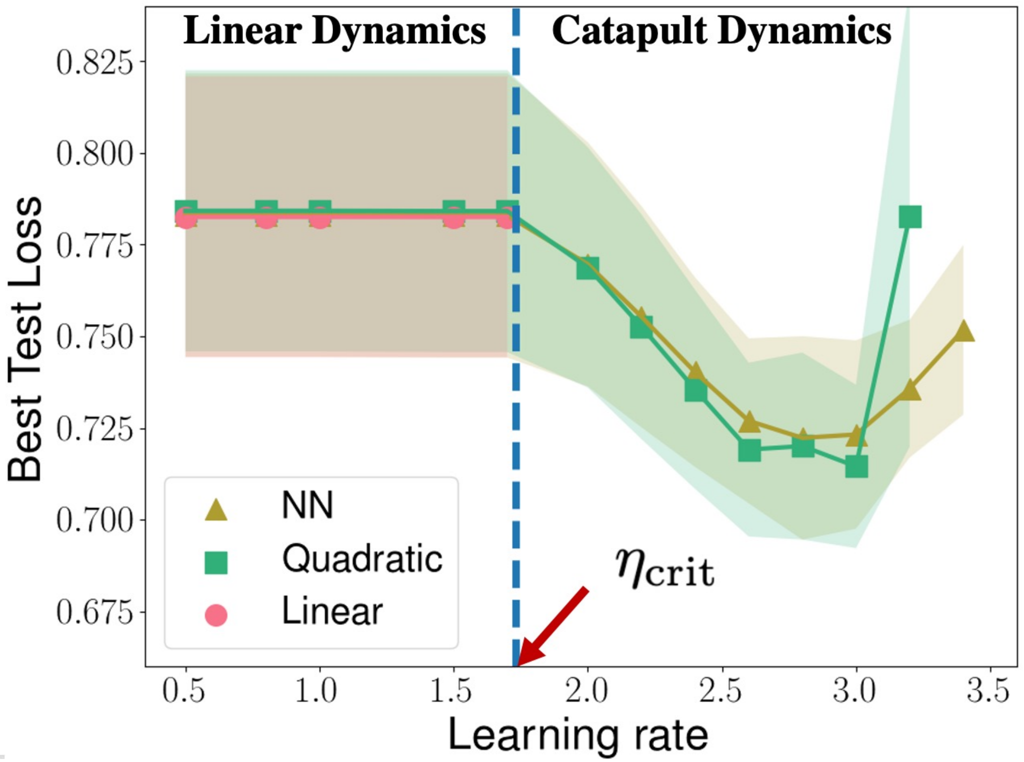


Figure from [Zhu et al. 2022]

Feature learning through catapults

AGOP(Average Gradient Outer Product) [Triveti et al. 2014],[Xia et al. 2002]…

For a parameterized model $f(w;\cdot)\colon \mathbb{R}^{p\times d} \to \mathbb{R}$, the AGOP of it w.r.t. input data $X \in \mathbb{R}^{n\times d}$ is

$$G(\mathbf{w}) = \frac{1}{n}\sum_{i=1}^{n} \frac{\partial f(\mathbf{w};\boldsymbol{x}_i)}{\partial \boldsymbol{x}_i}\frac{\partial f(\mathbf{w};\boldsymbol{x}_i)}{\partial \boldsymbol{x}_i}^T.$$

Example:

$$f(x) = x_1 x_2 \text{ with } x \in \mathbb{R}^{100}$$

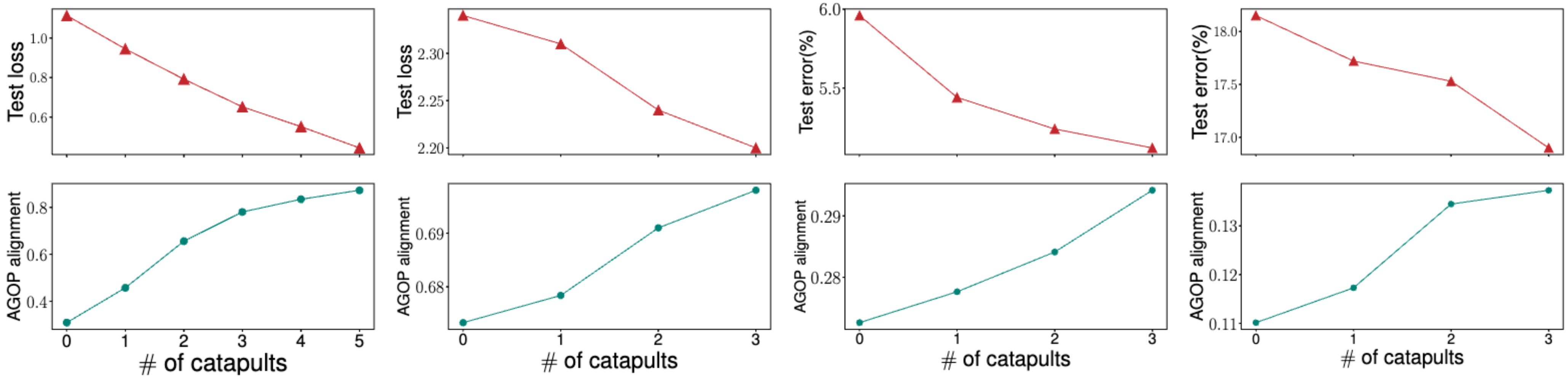
For n samples i.i.d. from $N(0, I_{100})$, the AGOP of $f(x)$ will converge to $\begin{pmatrix} I_2 & 0 \\ 0 & 0 \end{pmatrix}$ as $n \to \infty$.

Feature learning through catapults in GD

Tasks:

Rank-2 regression: $f^*(x) = x_1x_2$ with $x \in \mathbb{R}^{100}$
Rank-4 regression: $f^*(x) = x_1 + x_1x_2 + x_1x_2x_3 + x_1x_2x_3x_4$ with $x \in \mathbb{R}^{100}$

If the neural network $f(w;\cdot)$ learns the feature, its AGOP G_f should be close to the true AGOP G_{f^*}.



(a) Rank-2 regression (b) Rank-4 regression (c) SVHN-2 (d) CelebA

Feature learning through catapults in GD

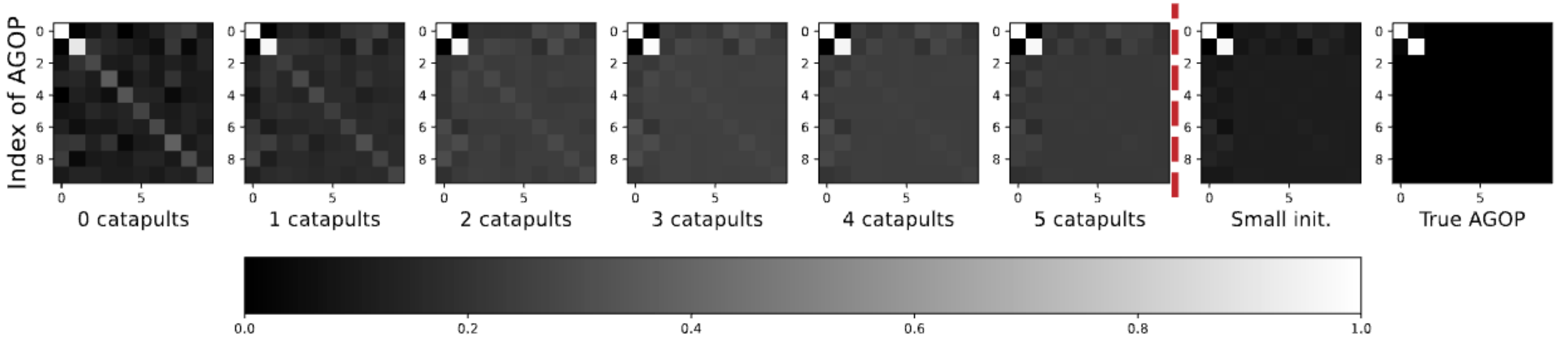
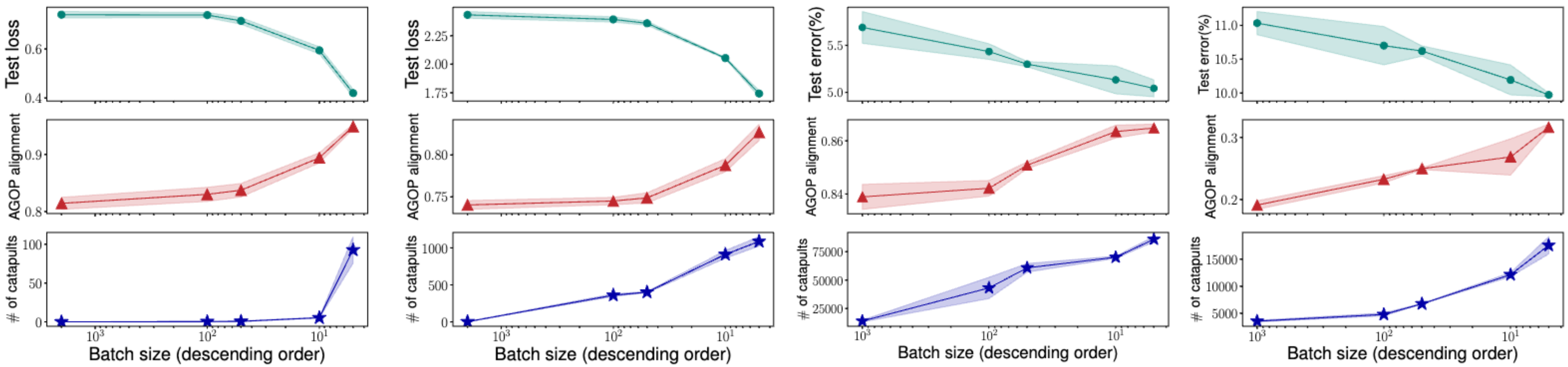


Figure 17: **Visualization of AGOP for rank-2 regression task.** All pixels are normalized to the range $[0, 1]$ and the top 10 rows and columns of the AGOP are plotted.

Small batch size leads to more catapults, hence better generalization

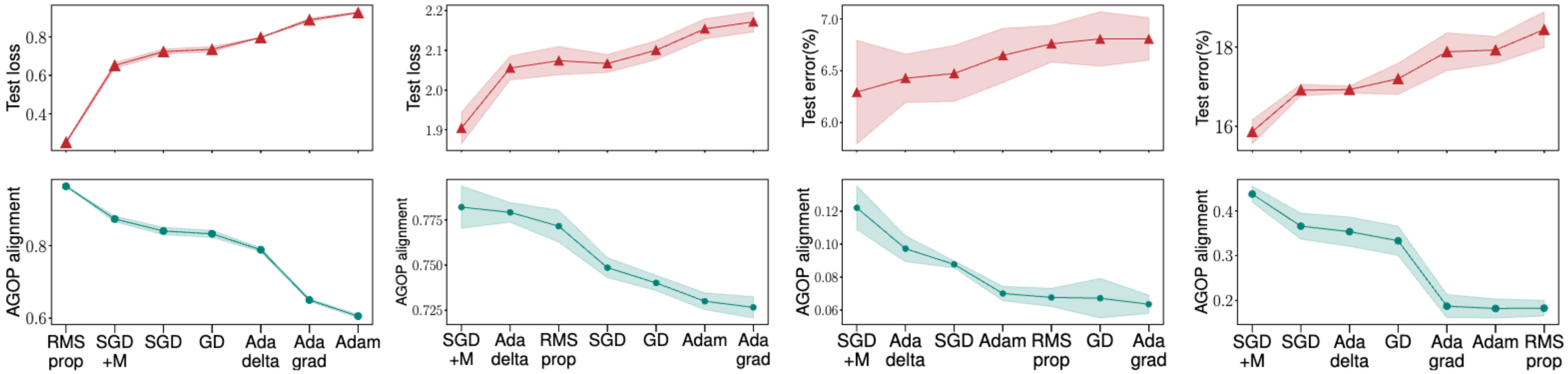


(a) Rank-2 regression (b) Rank-4 regression (c) SVHN (d) CelebA

Observation:

1. Test loss/error correlates with AGOP alignment.
2. A smaller batch size corresponds to a greater number of catapults, hence leading to better generalization.

Generalization with different optimizers correlates with AGOP alignment



(a) Rank-2 regression (b) Rank-4 regression (c) SVHN-2 (d) CelebA

Thanks