

Instability is all you need: *the surprising dynamics of learning in deep models*

Stephen Roberts
University of Oxford



With thanks to co-authors

Lawrence Wang

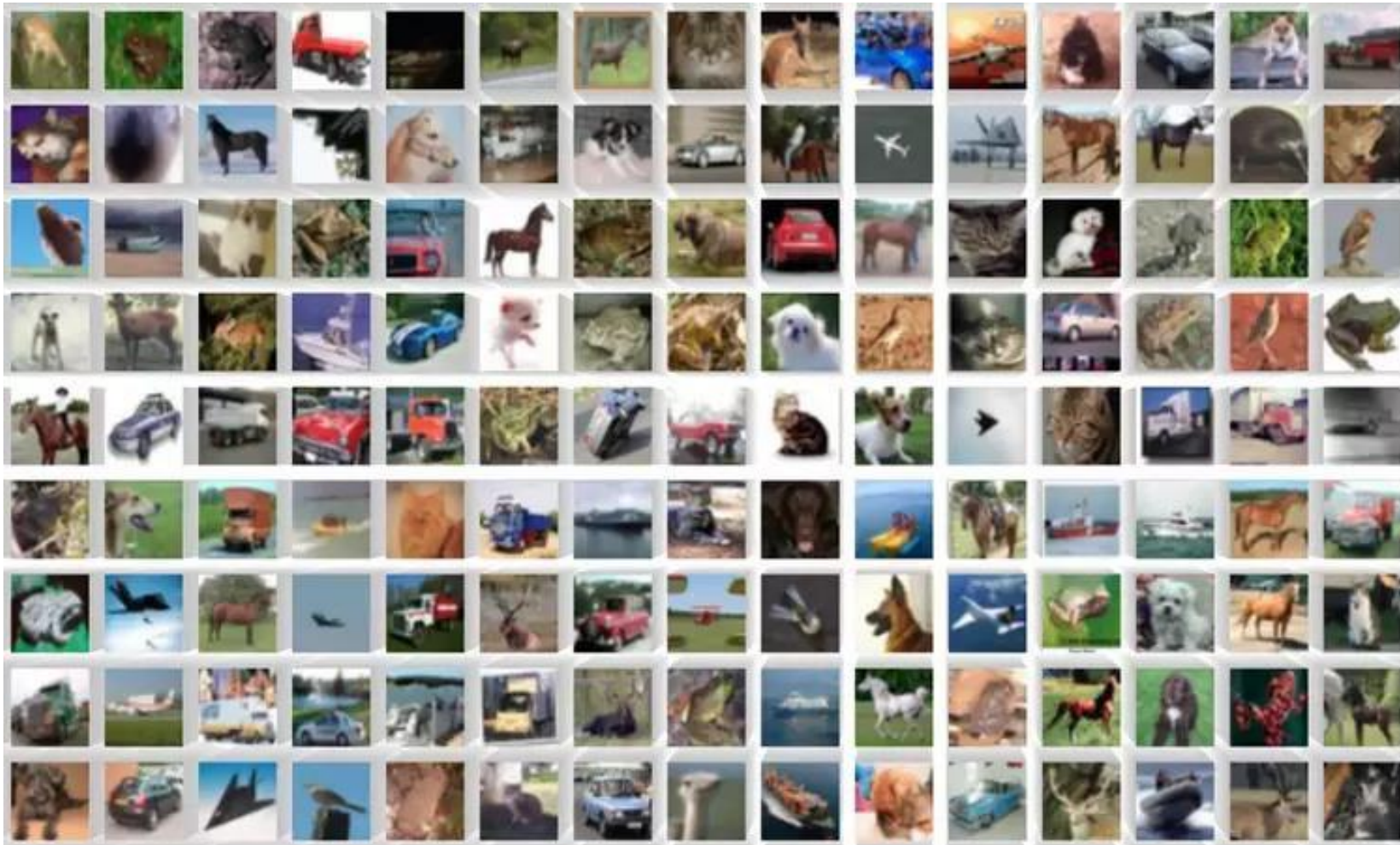
Diego Granziol

John Williams

Preamble: Generalization

Why worry?

We want models that can perform well across different data sets



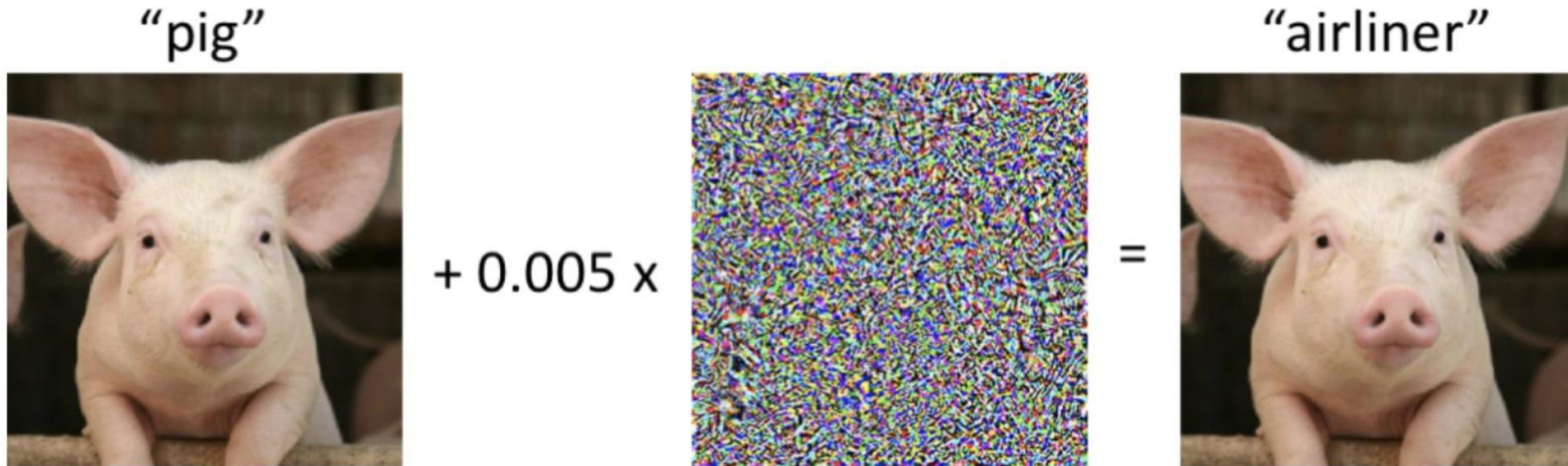
Why worry?

(Re)-training models is *costly*



Why worry?

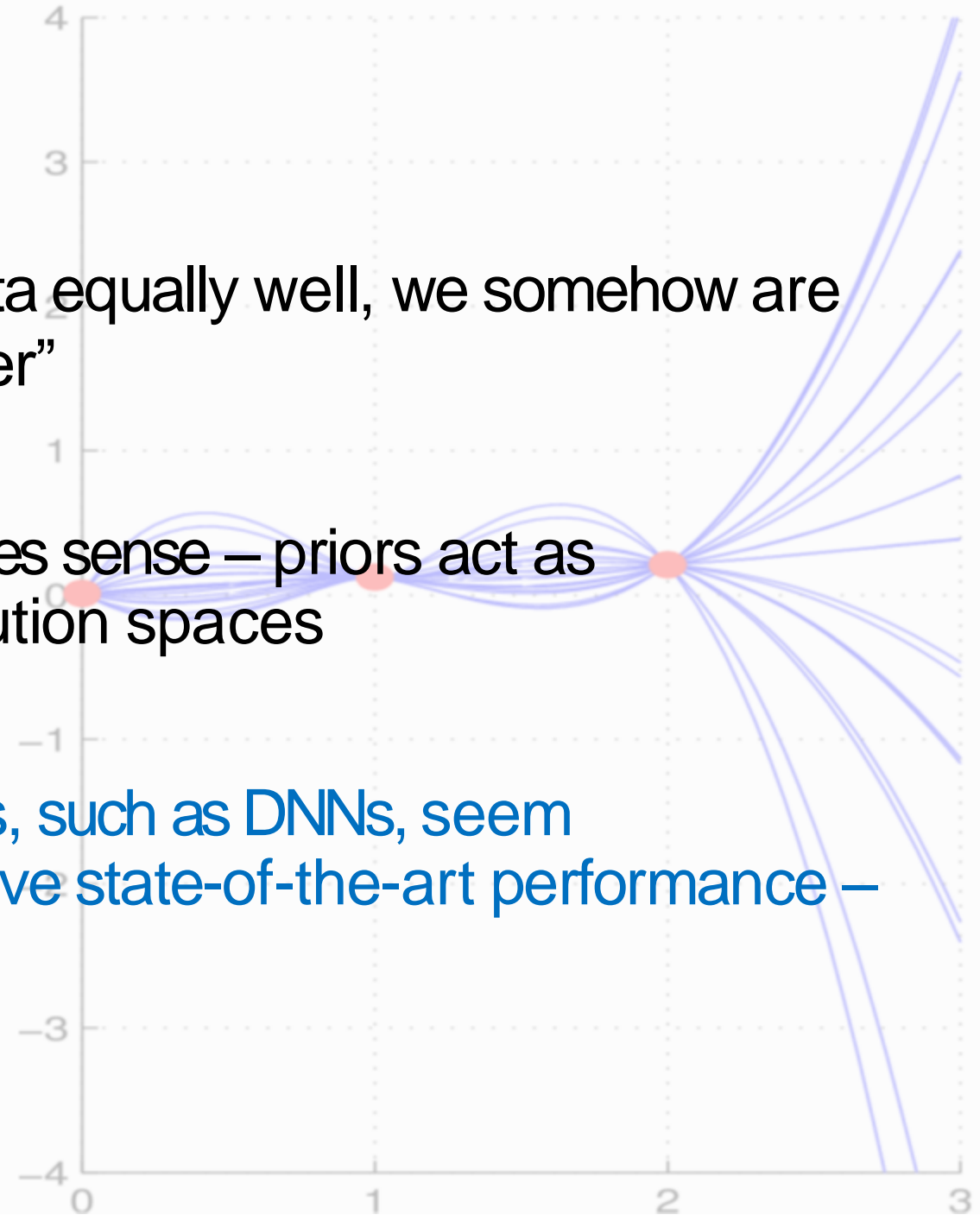
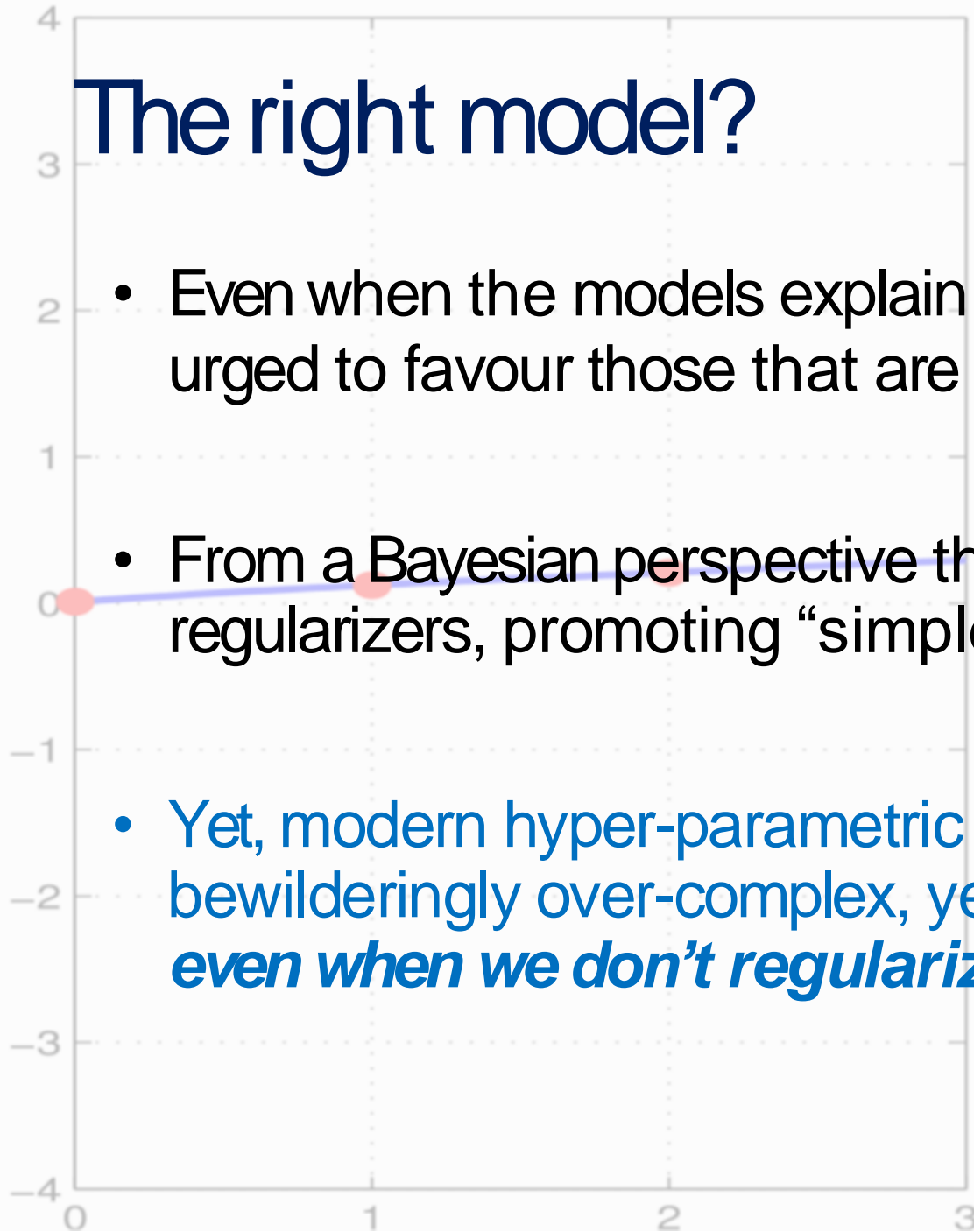
We want models that are hard to spoof



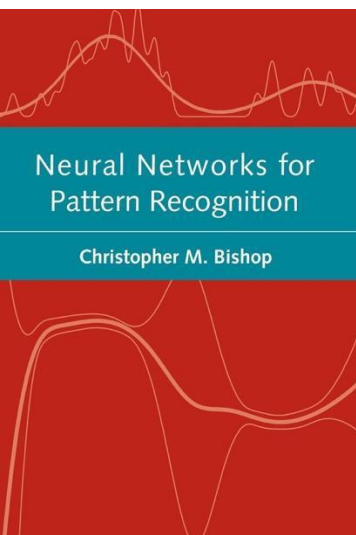
Adversarial example for Inception v3 network on ImageNet example of class pig (Source: Madry Lab at MIT)

The right model?

- Even when the models explain the data equally well, we somehow are urged to favour those that are “simpler”
- From a Bayesian perspective this makes sense – priors act as regularizers, promoting “simpler” solution spaces
- Yet, modern hyper-parametric models, such as DNNs, seem bewilderingly over-complex, yet achieve state-of-the-art performance – *even when we don't regularize!*



The Bayesian view



Bishop, 1995

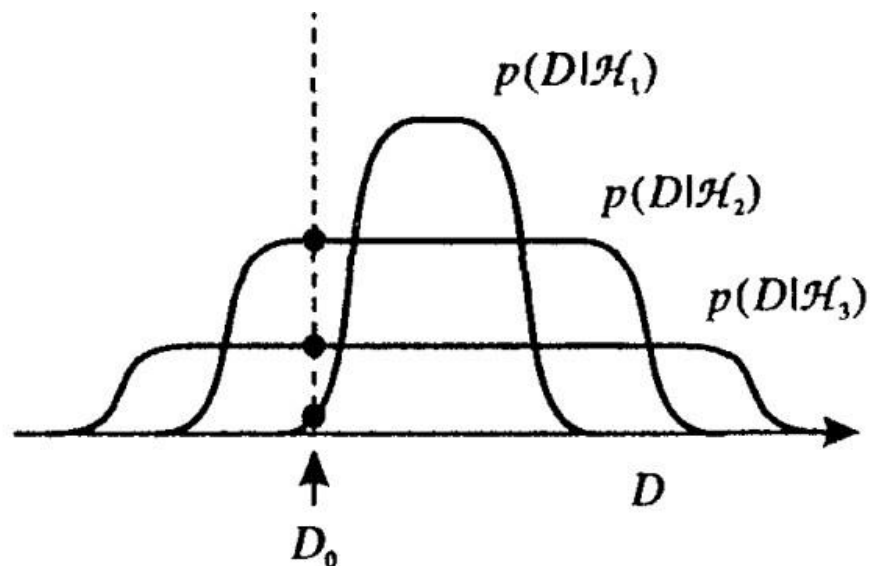
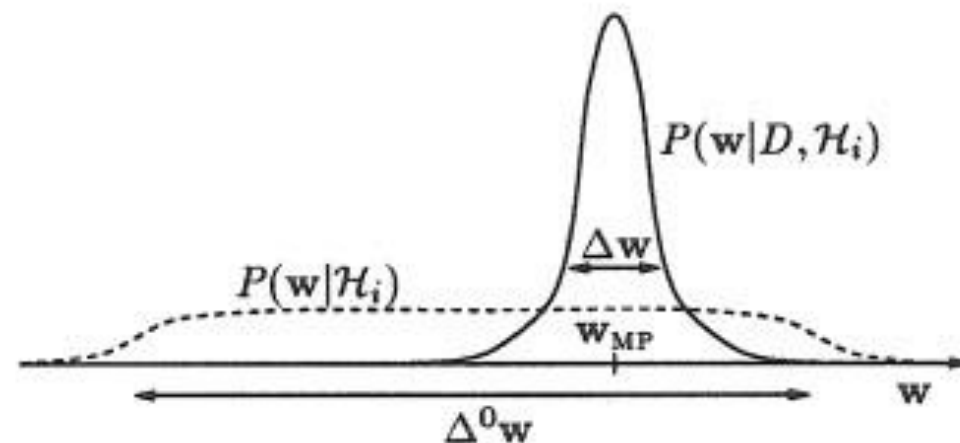


Figure 10.1. Schematic example of three models, \mathcal{H}_1 , \mathcal{H}_2 and \mathcal{H}_3 , which have successively greater complexity, showing the probability (known as the *evidence*) of different data sets D given each model \mathcal{H}_i . We see that more complex models can describe a greater range of data sets. Note, however, that the distributions are normalized. Thus, when a particular data set D_0 is observed, the model \mathcal{H}_2 has a greater evidence than either the simpler model \mathcal{H}_1 or the more complex model \mathcal{H}_3 .



Bayesian Methods for Adaptive Models

Thesis by David John Cameron MacKay

In Partial Fulfillment of the Requirements
for the Degree of Doctor of Philosophy

California Institute of Technology
Pasadena, California

1992

(Submitted December 10, 1991)

Advisor: Prof. J.J. Hopfield

The Bayesian view

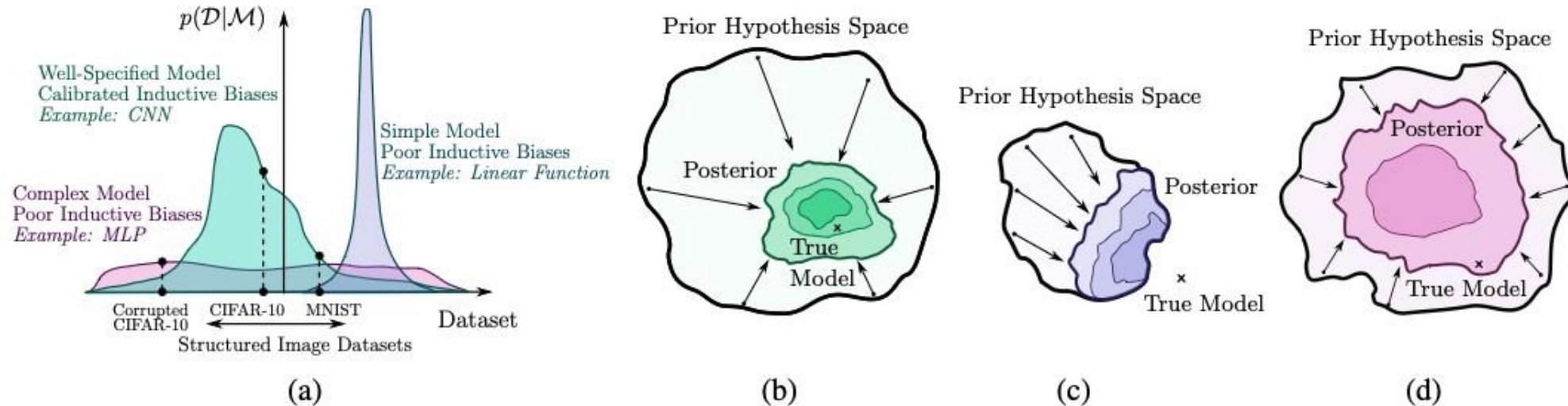
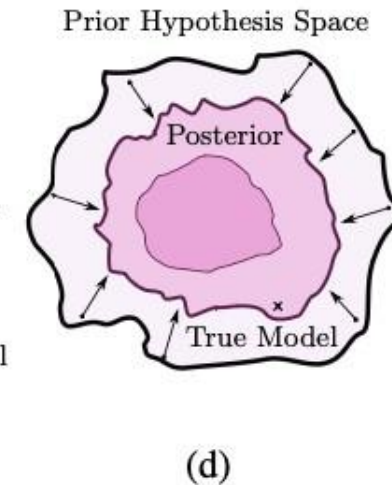
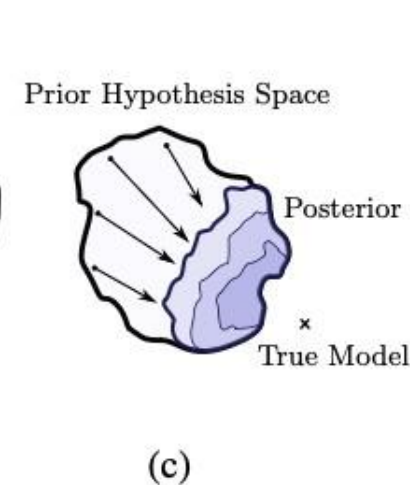
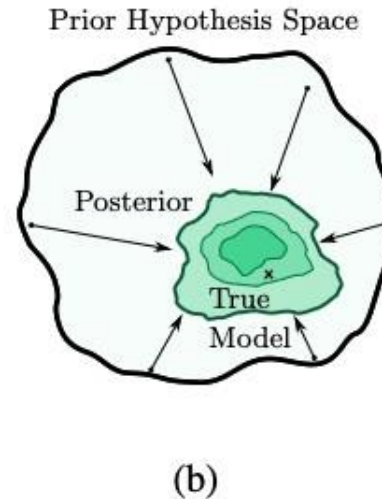
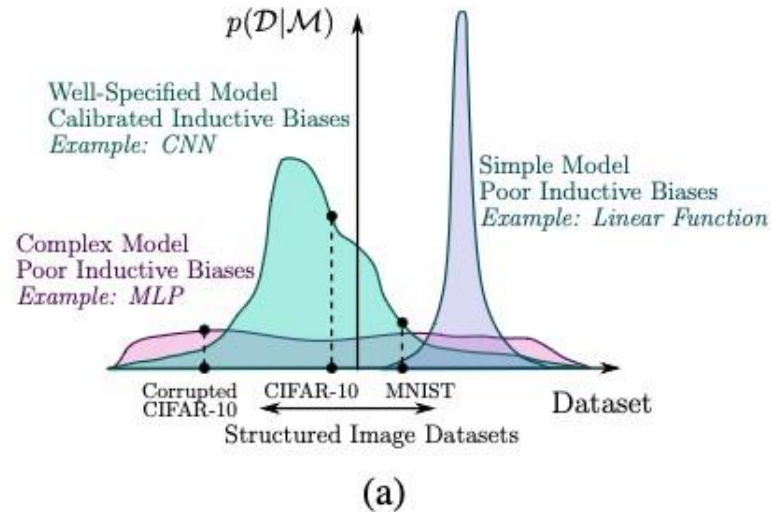


Figure 2. A probabilistic perspective of generalization. (a) Ideally, a model supports a wide range of datasets, but with inductive biases that provide high prior probability to a particular class of problems being considered. Here, the CNN is preferred over the linear model and the fully-connected MLP for CIFAR-10 (while we do not consider MLP models to in general have poor inductive biases, here we are considering a hypothetical example involving images and a very large MLP). (b) By representing a large hypothesis space, a model can contract around a true solution, which in the real-world is often very sophisticated. (c) With truncated support, a model will converge to an erroneous solution. (d) Even if the hypothesis space contains the truth, a model will not efficiently contract unless it also has reasonable inductive biases.

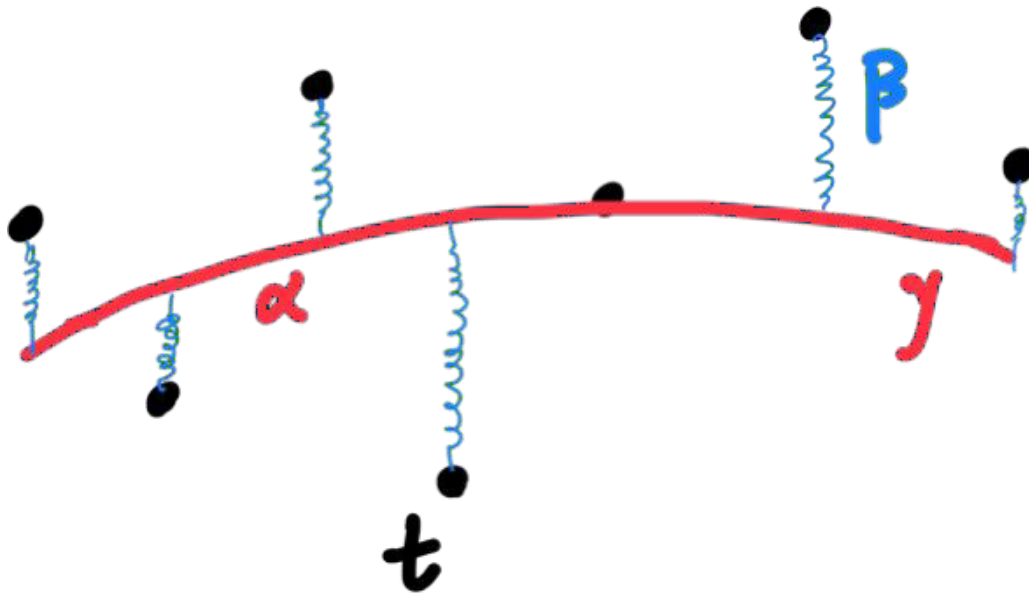
The Bayesian view



(b) By representing a large hypothesis space, a model can **contract** around a true solution, which in the real-world is often very sophisticated.

(d) Even if the hypothesis space contains the truth, a model will not efficiently contract unless it also has reasonable **inductive biases**.

Regularization – a classic inductive bias



Bending Energy is function of $\left| \frac{d^2 y}{dx^2} \right|^2$

$$y = w\phi(x, \sigma)$$

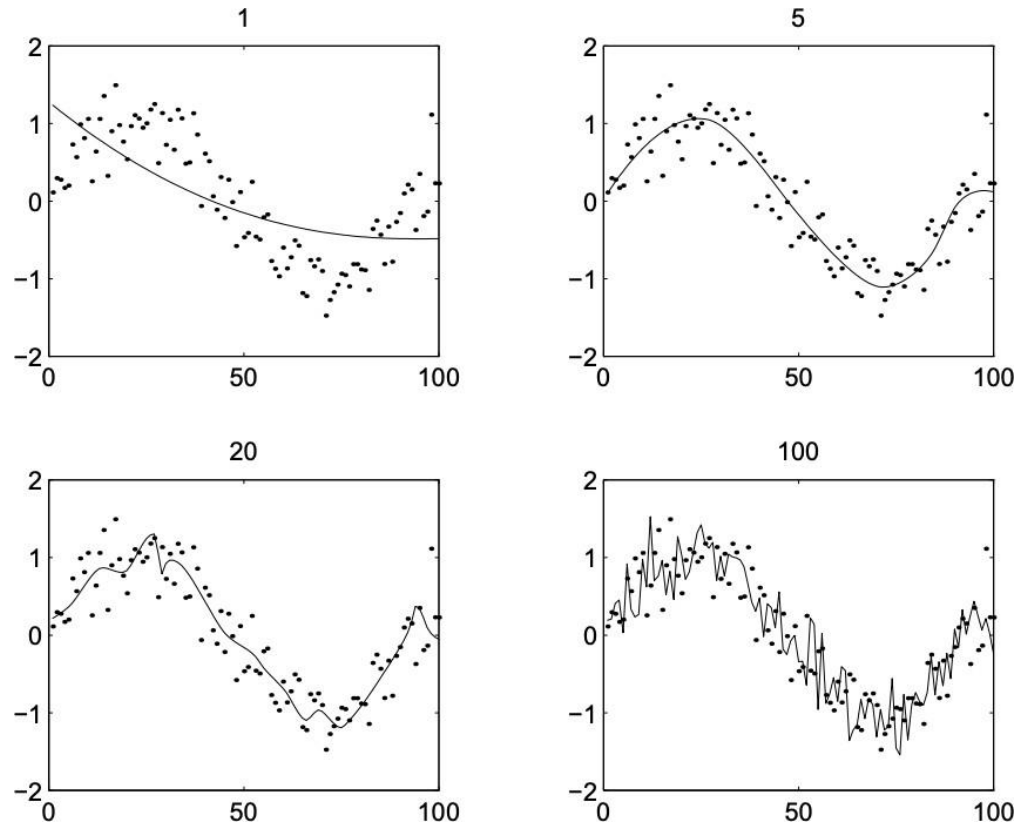
$$\left| \frac{d^2 y}{dx^2} \right| \propto |w| \text{ and } \sigma^{-1}$$

$$\beta(y - t)^2 + \alpha \left| \frac{d^2 y}{dx^2} \right|^2$$

Norm on weights -> ridge regression, Lasso etc
Width of basis -> kernel lengthscales etc

(Tikhonov and Arsenin, 1977)

Regularization – priors -> inductive bias



$$D = \{x, t\}$$

$$y = f(D, w)$$

$$p(D, w) \propto p(D|w)p(w)$$

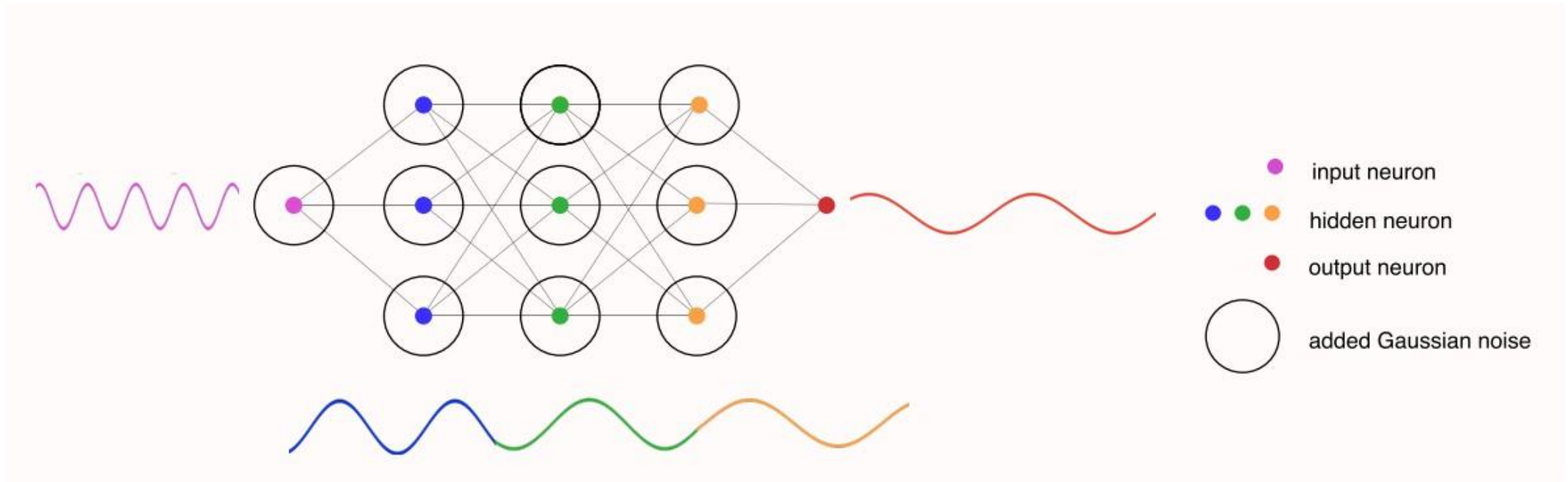
$$E(D, w) = \underbrace{-\log p(D|w)}_{\text{Data error term}} - \underbrace{\log p(w)}_{\text{Penalty term}}$$

Data error term

Penalty term

In effect, we induce a bias such that “simpler” solutions are preferred (more on that later)

Regularization – noise injection



Alexander Camuto, Matthew Willetts, Umut AzimaYekli, Stephen Roberts, Chris Holmes (2020). Explicit Regularisation in Gaussian Noise Injections. *Proceedings of NeurIPS 2020*.

Regularization

- Early stopping
- Momentum terms
- (Stochastic) weight averaging
- (Bayesian) model averaging

- Sharpness (much more later...)

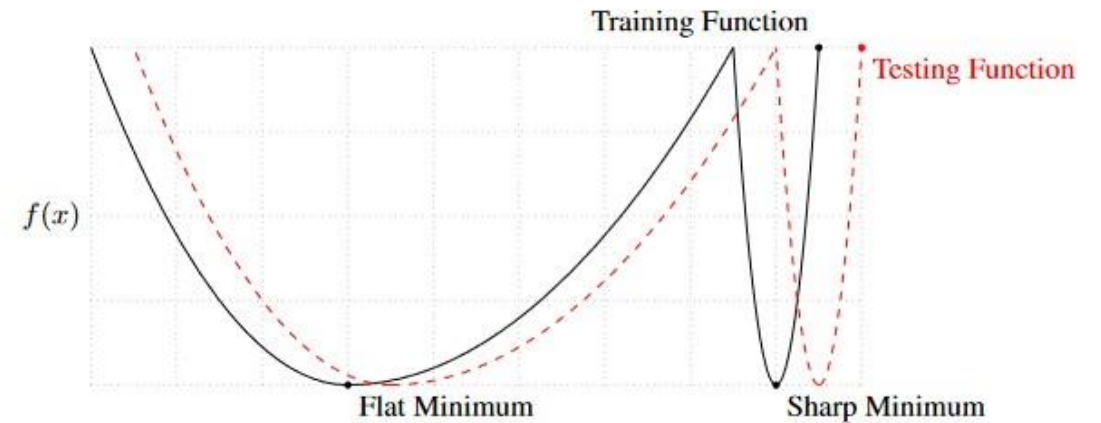
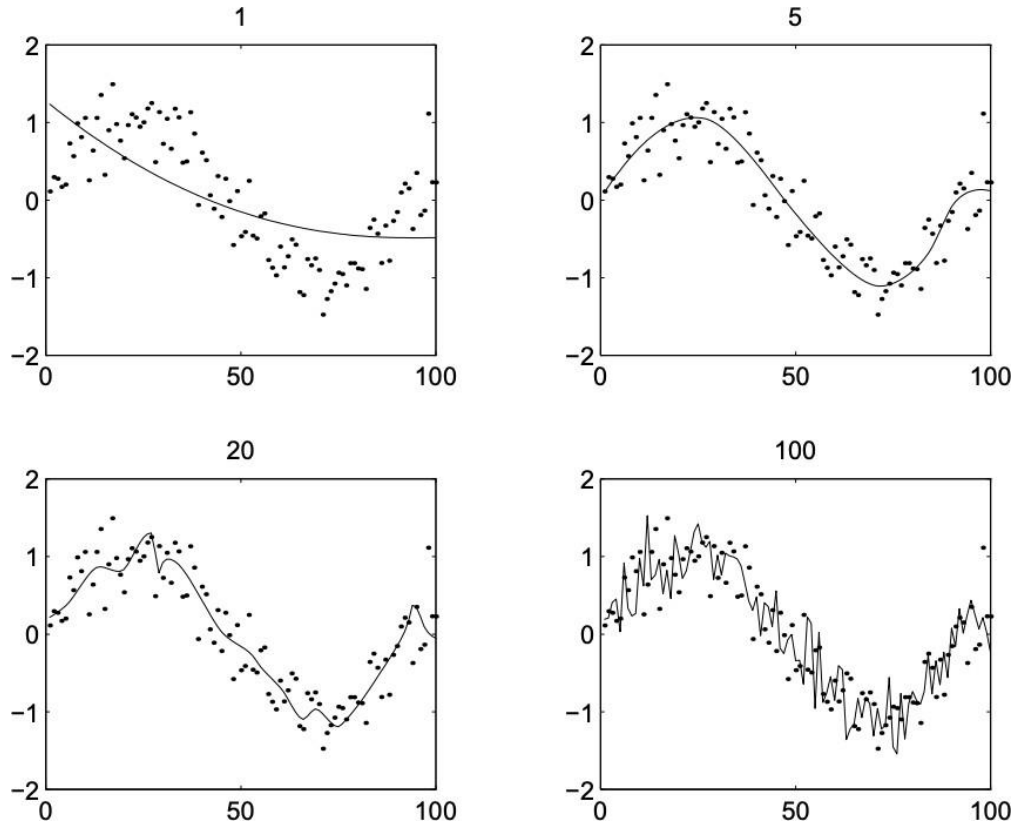


Figure 1: A Conceptual Sketch of Flat and Sharp Minima. The Y-axis indicates value of the loss function and the X-axis the variables (parameters)

The Dynamics of Learning

“Yet, modern hyper-parametric models, such as DNNs, seem bewilderingly over-complex, yet achieve state-of-the-art performance – ***even when we don’t regularize!***”

Sharpness

- Generalization gap = |In-sample error – ‘Unseen’ set error|
- Large training temperatures¹ (λ/B) seem to lead to flatter basins -> “simpler” solutions, potentially with a *lower generalization gap*²

Figure imported from Figure 1 of: NS Keskar et al. *On Large-Batch Training for Deep Learning: Generalization Gap and Sharp Minima*. (2017)

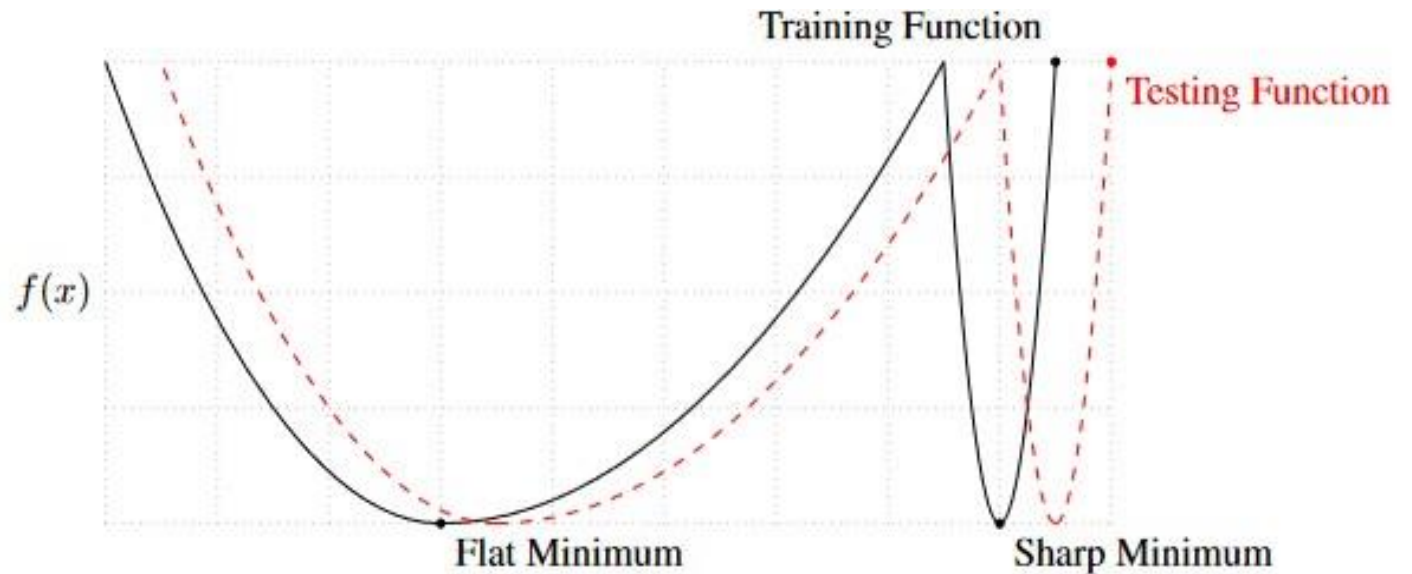
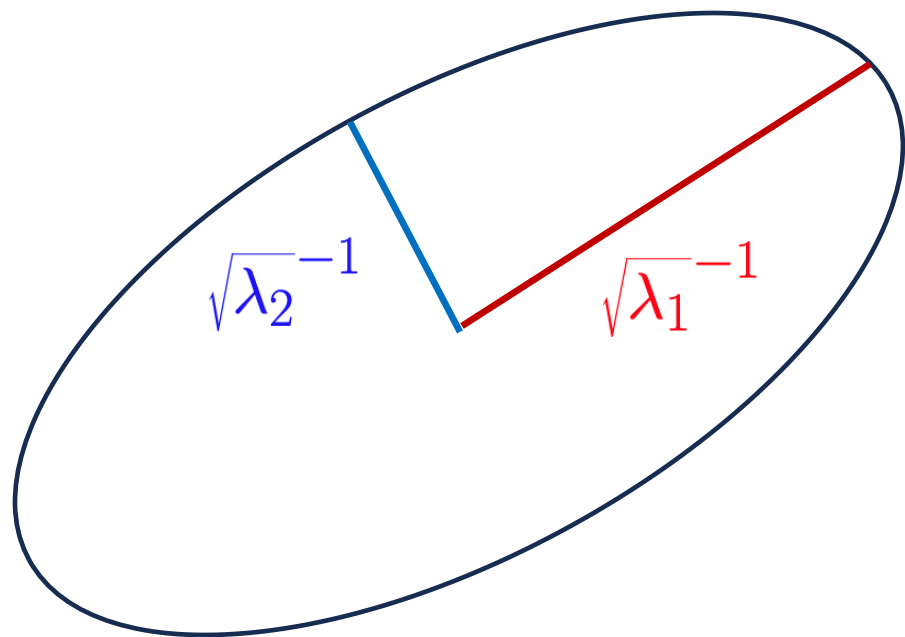


Figure 1: A Conceptual Sketch of Flat and Sharp Minima. The Y-axis indicates value of the loss function and the X-axis the variables (parameters)

1. D Granzio et al. *Learning Rates as a Function of Batch Size: A Random Matrix Theory Approach to Neural Network Training*. (2020)

2. S Hochreiter, J Schmidhuber. *Flat minima*. (1997)

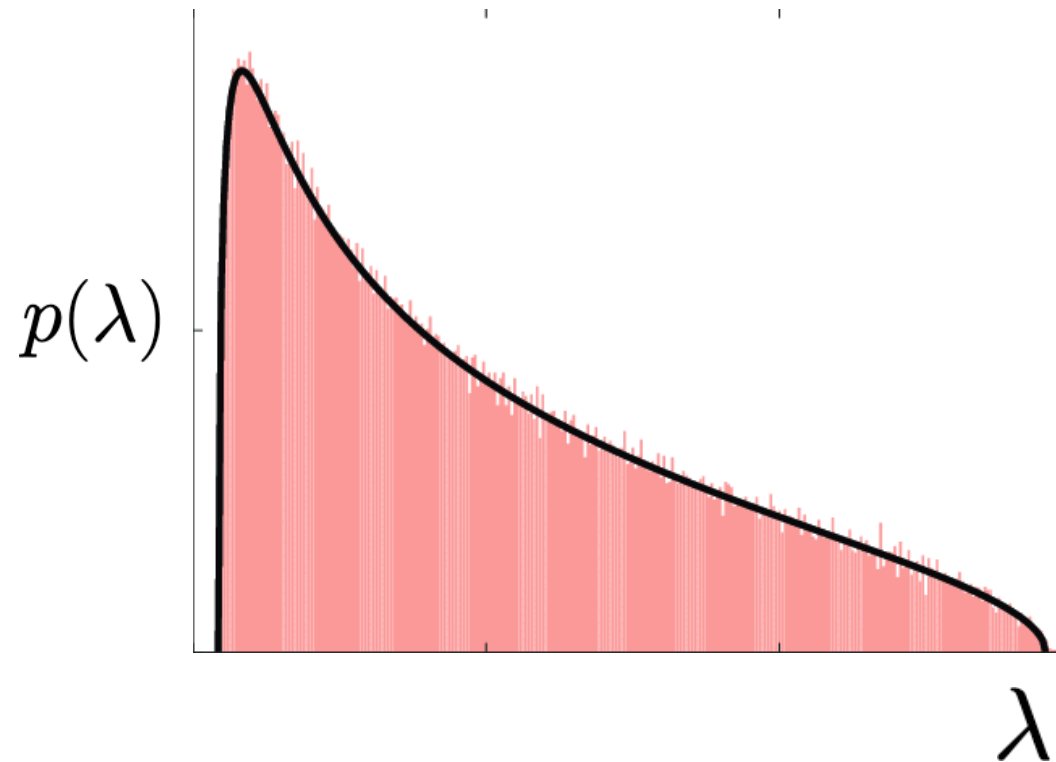
Sharpness



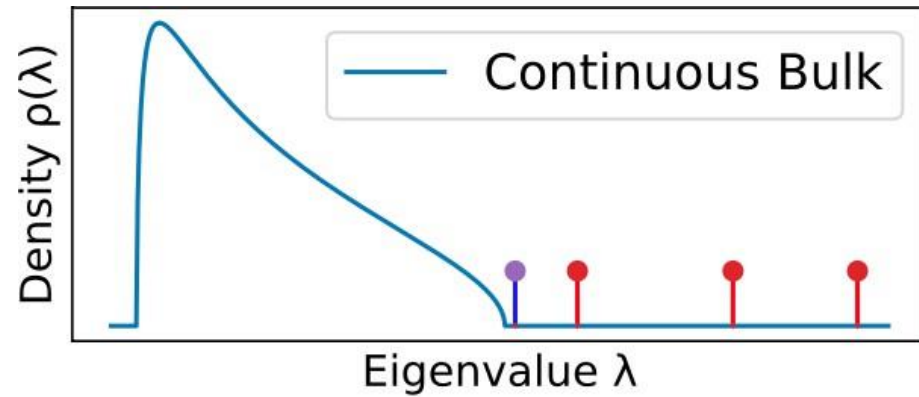
Hessian is inverse covariance, so larger eigenvalues \rightarrow sharper solutions

$$H_{ij} = \left. \frac{\partial^2 E}{\partial w_i \partial w_j} \right|_{\mathbf{w}^*}$$

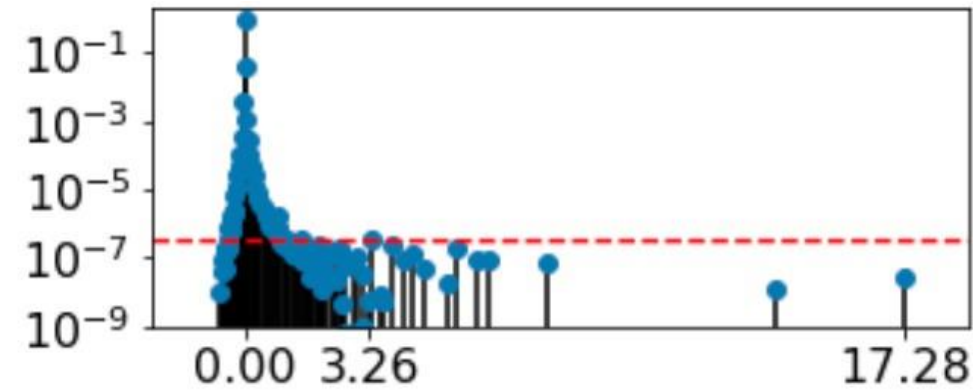
Hessian is a general Wishart matrix, so eigen density given by Marchenko-Pastur distribution



Sharpness



(a) Hypothetical $\rho(\lambda)$



(b) VGG-16 C-100 Hessian

Figure from: *Diego Granziol, Stefan Zohren, Stephen Roberts (2022). Learning Rates as a Function of Batch Size: A Random Matrix Theory Approach to Neural Network Training. Journal of Machine Learning Research 23(173): 1-65, 2022.*

SHARPNESS-AWARE MINIMIZATION FOR EFFICIENTLY IMPROVING GENERALIZATION

Pierre Foret *
Google Research
pierre.pforet@gmail.com

Ariel Kleiner
Google Research
akleiner@gmail.com

Hossein Mobahi
Google Research
hmobahi@google.com

Behnam Neyshabur
Blueshift, Alphabet
neyshabur@google.com

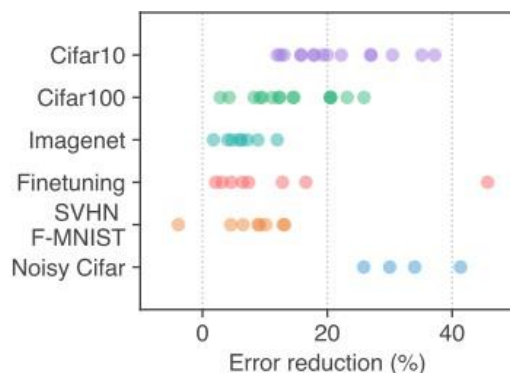
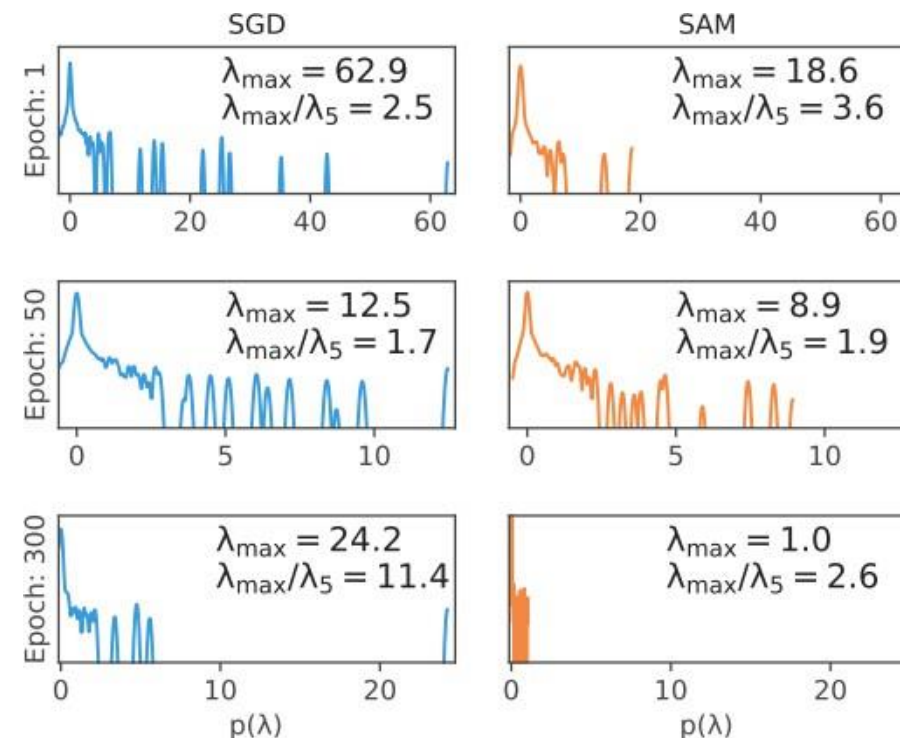


Figure 1: (left) Error rate reduction obtained by switching to SAM. Each point is a different dataset / model / data augmentation. (middle) A sharp minimum to which a ResNet trained with SGD converged. (right) A wide minimum to which the same ResNet trained with SAM converged.



Goal – tame the maximum eigenvalue

SAM – shows impressive performance boosts (though there are mixed reports of performance gains)

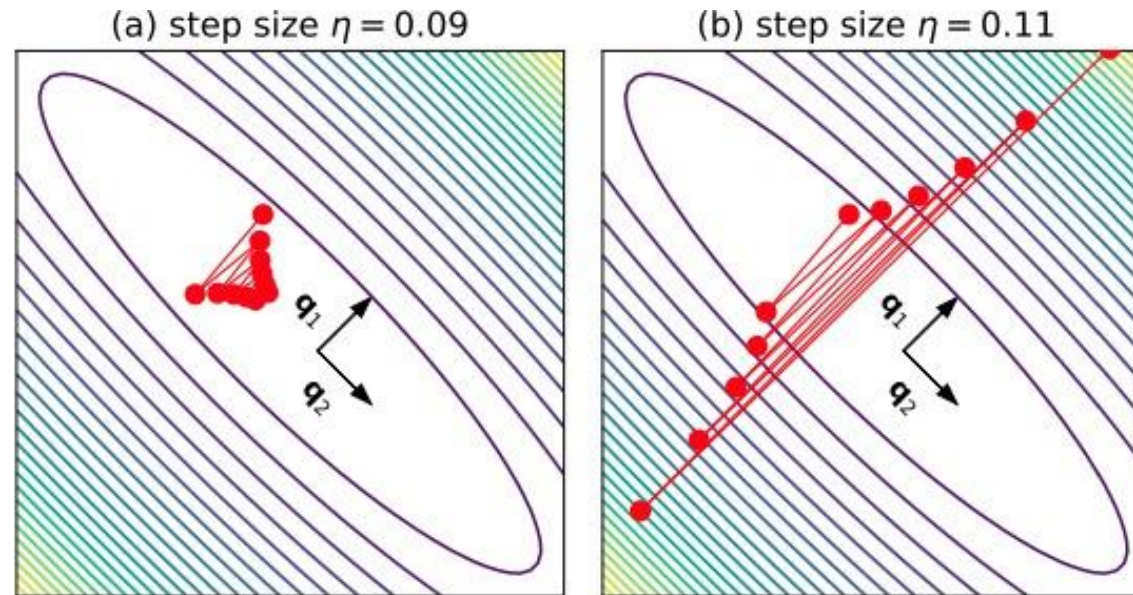
On the Maximum Hessian Eigenvalue and Generalization

Simran Kaur[†], Jeremy Cohen[†], Zachary C. Lipton[†]

[†]Carnegie Mellon University
{skaur, jeremycohen, zlipton}@cmu.edu

The edge of stability

$$\eta < \frac{2}{\lambda_{max}}$$



$$\eta_{\text{eos}} = 0.1$$

Figure 2: Gradient descent on a quadratic with eigenvalues $a_1 = 20$ and $a_2 = 1$.

The phases of learning – phase 1

We see *progressive sharpening* whilst

$$\lambda_{\max} < \frac{2}{\eta}$$

Then there is a distinct transition to another regime...

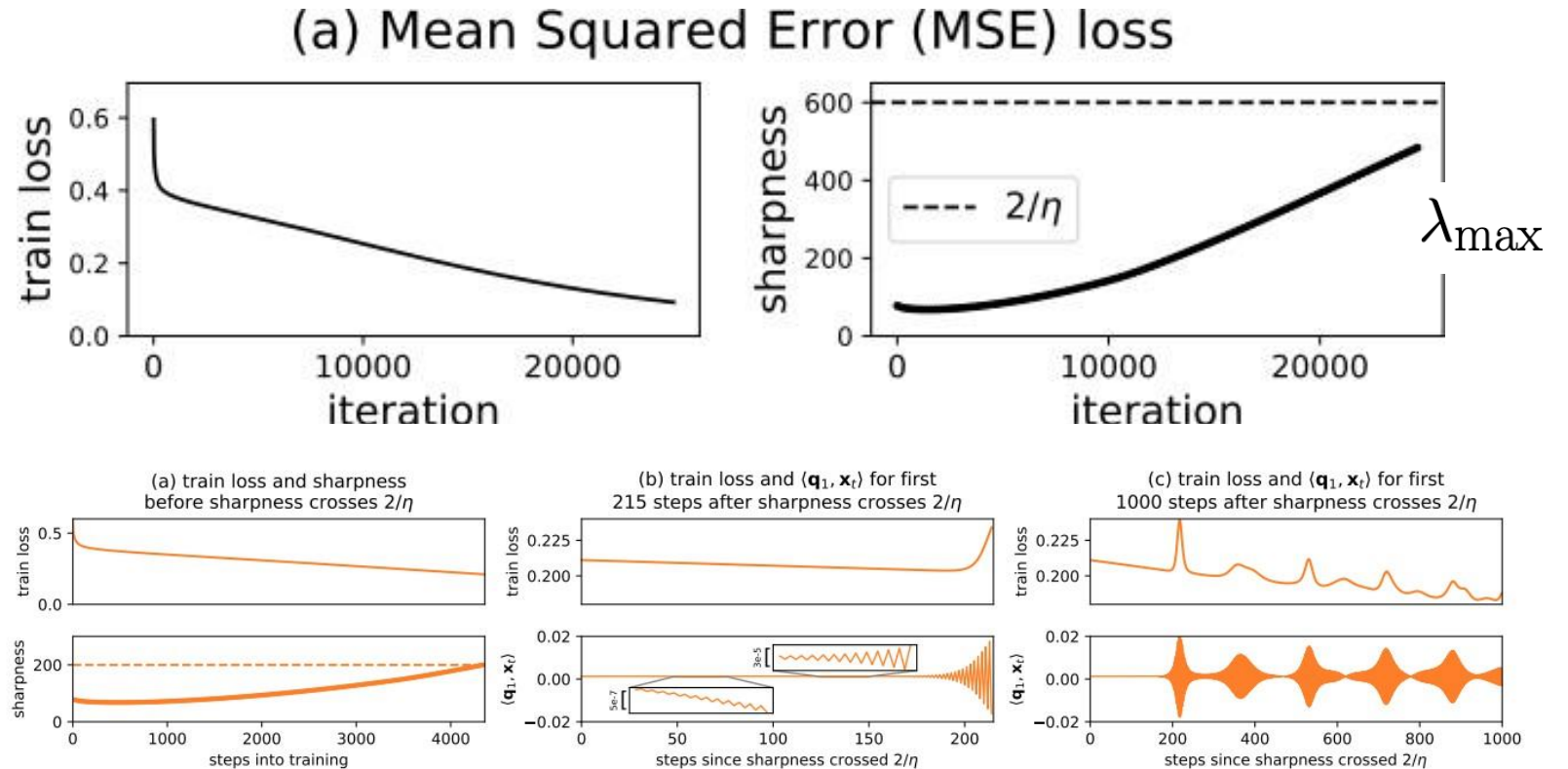
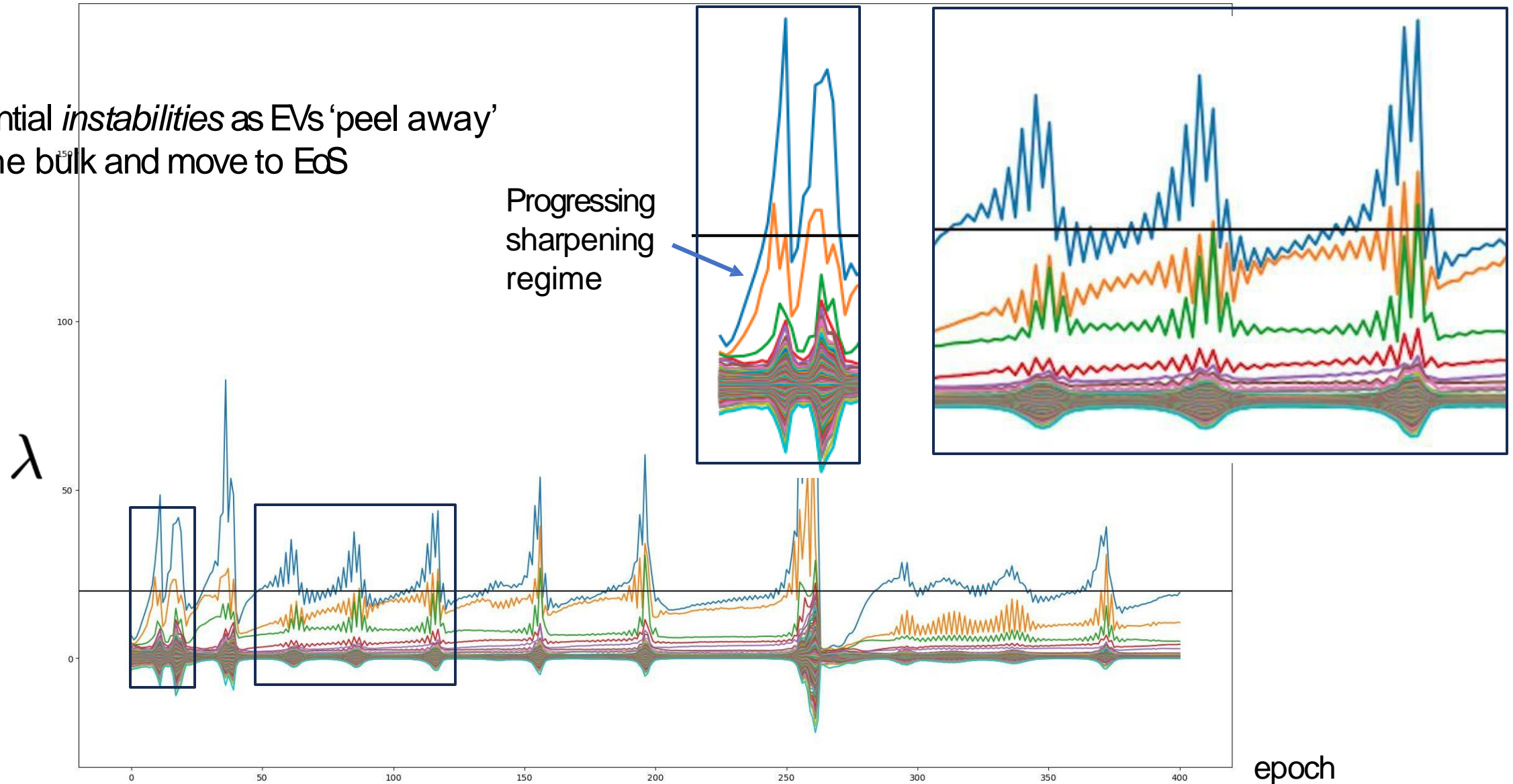


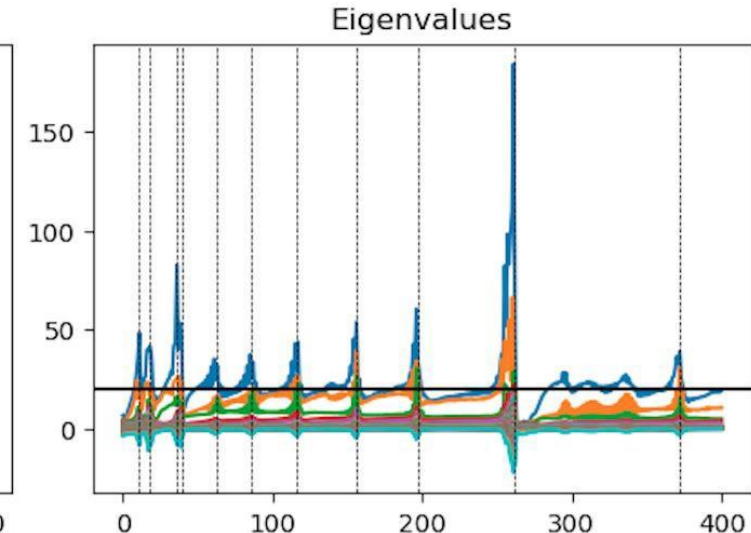
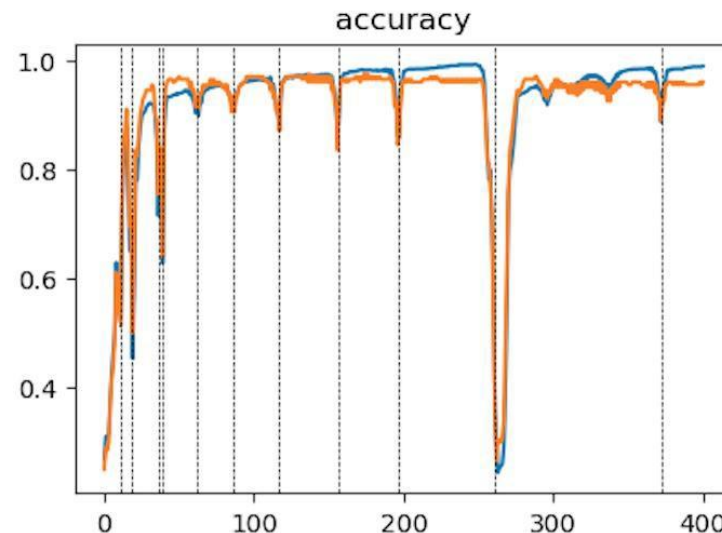
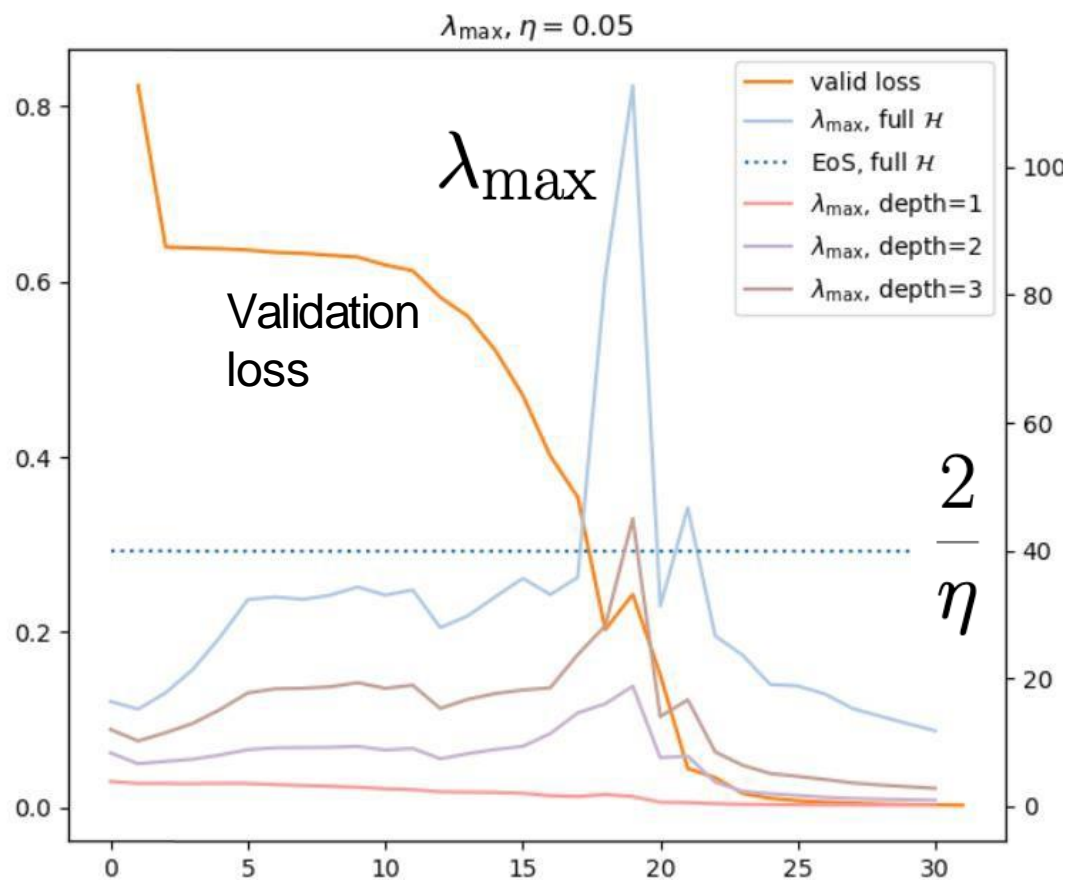
Figure 4: **Once the sharpness crosses $2/\eta$, gradient descent becomes destabilized.** We run gradient descent at $\eta = 0.01$. (a) The sharpness eventually reaches $2/\eta$. (b) Once the sharpness crosses $2/\eta$, the iterates start to oscillate along \mathbf{q}_1 with ever-increasing magnitude. (c) Somehow, GD does not diverge entirely; instead, the train loss continues to decrease, albeit non-monotonically.

The phases of learning – phase 2

Sequential *instabilities* as EVs ‘peel away’ from the bulk and move to EoS



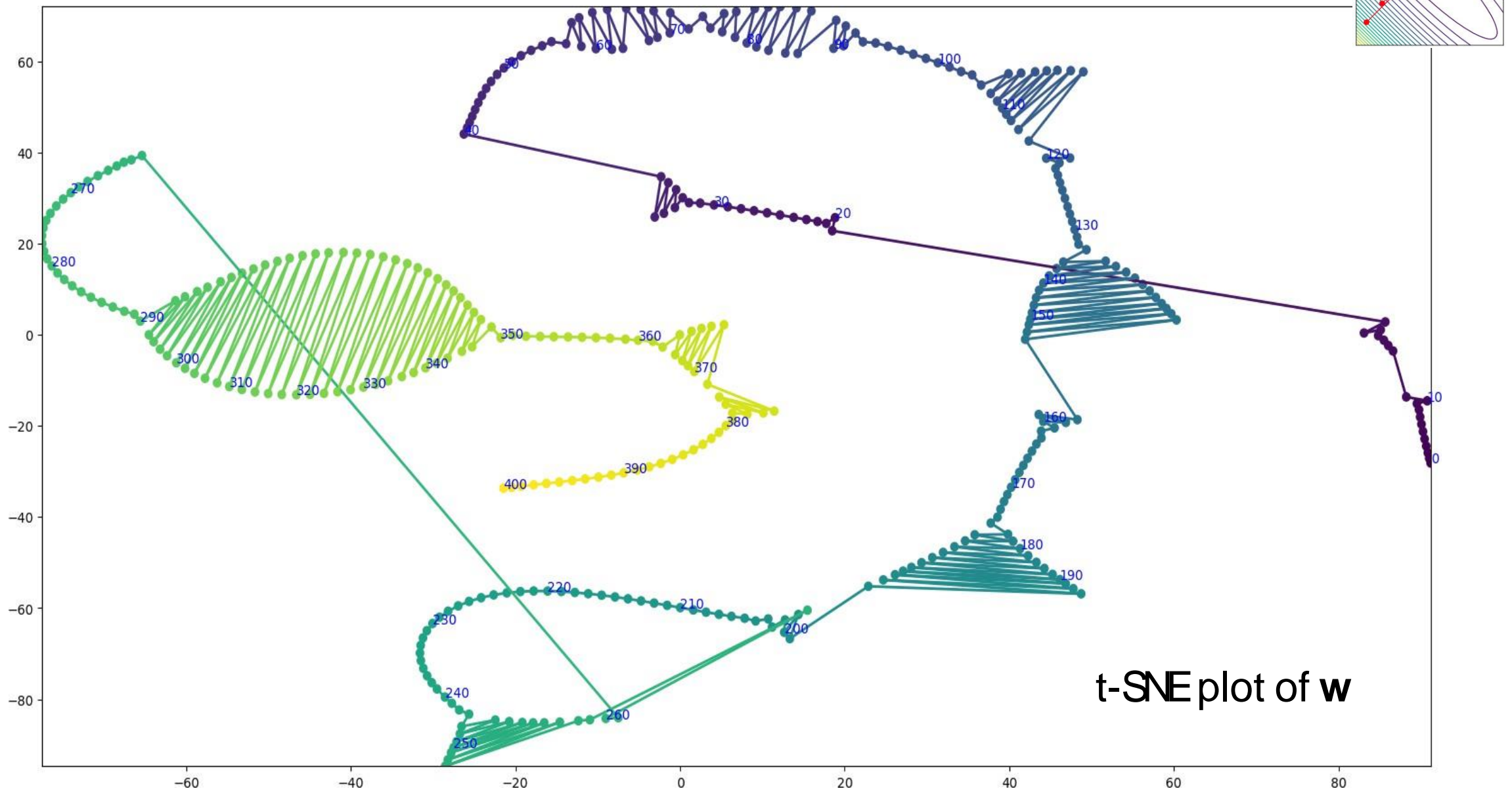
Instabilities



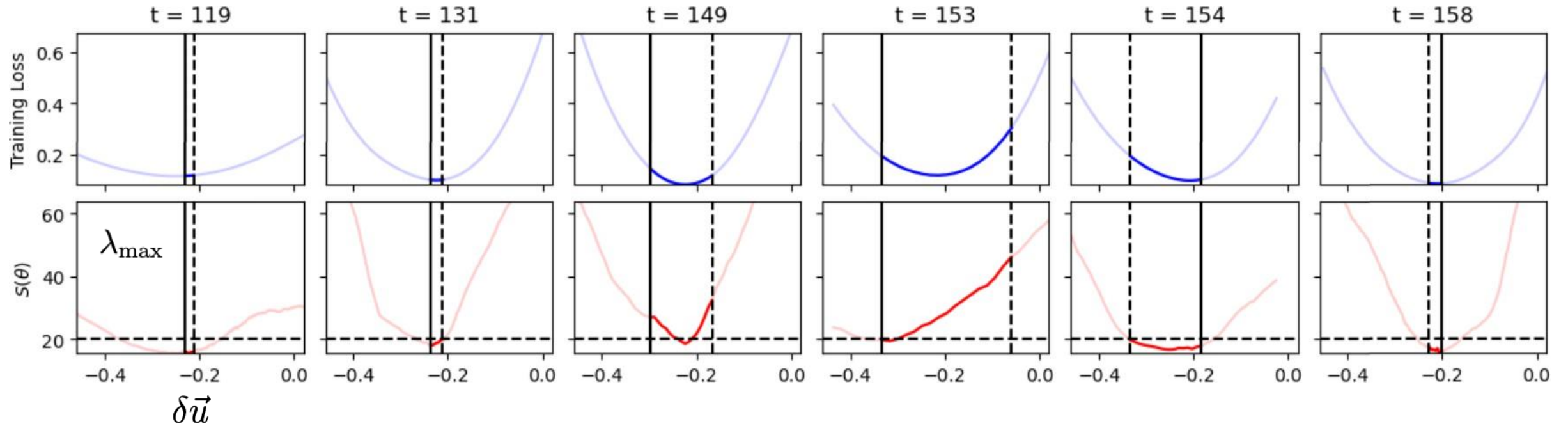
Phase 2 leads to progressive instabilities in learning

Oscillations in w (mainly in directions of maximum sharpness) induce phase changes leading to *less-sharp* solutions with improved performance

Instabilities



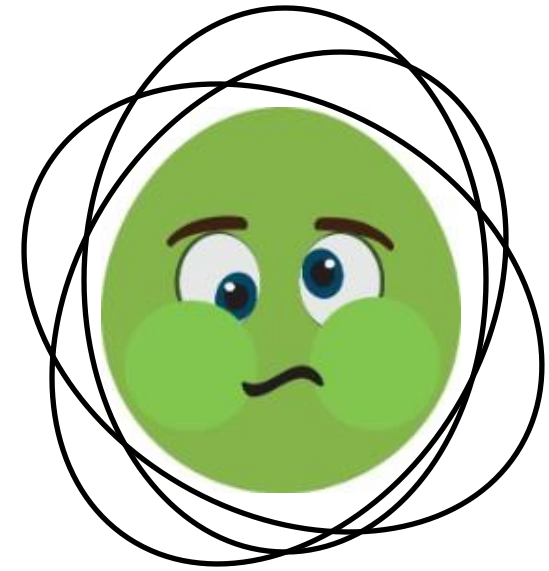
Passing through instabilities



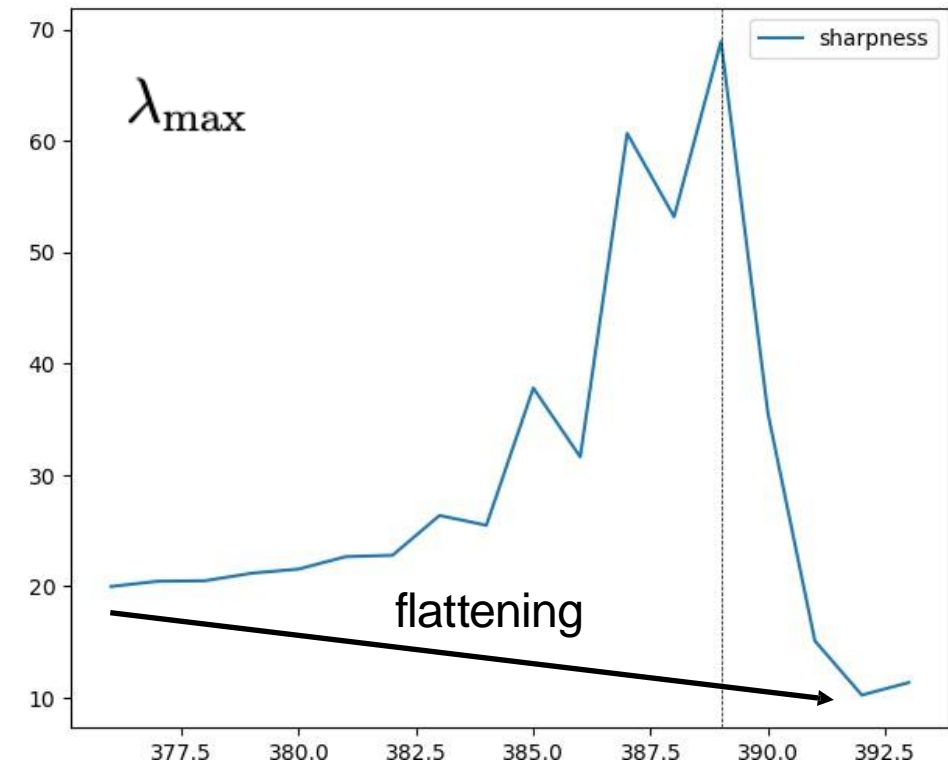
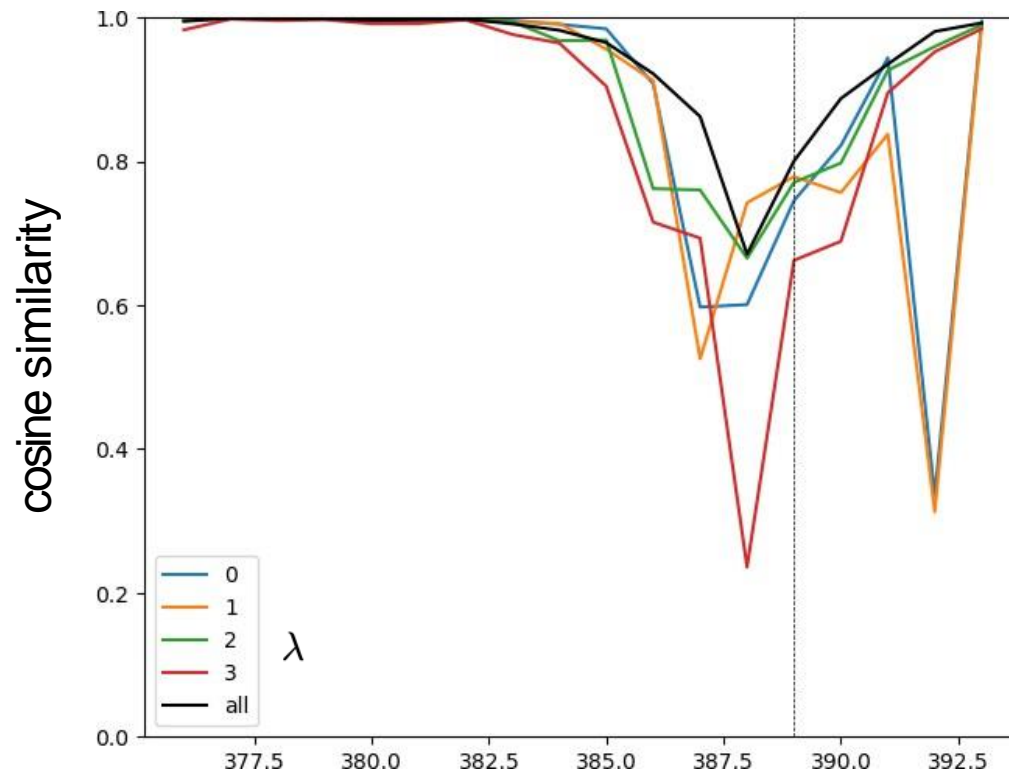
Dashed line = starting point in direction of parameter update
Solid line = end point after update using standard SGD

Instabilities - wobbly Hessians

The Hessian matrix becomes highly variable through an instability – in the directions of large λ



Noise-injection regularization along directions of maximum sharpness



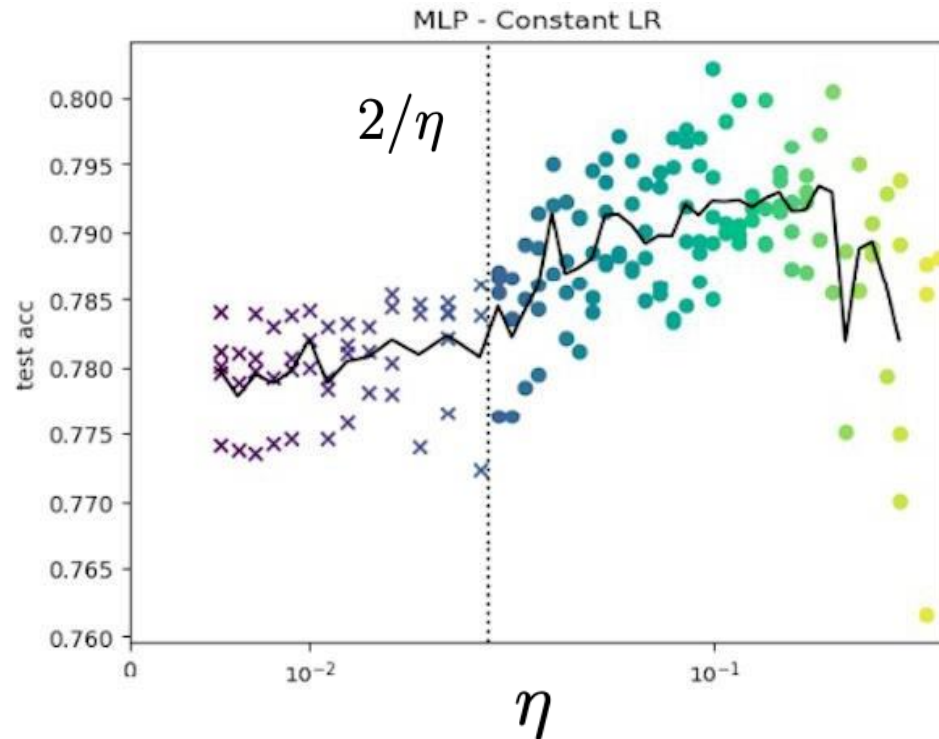
Aha!

Yet, modern hyper-parametric models, such as DNNs, seem bewilderingly over-complex, yet achieve state-of-the-art performance – *even when we don't regularize*

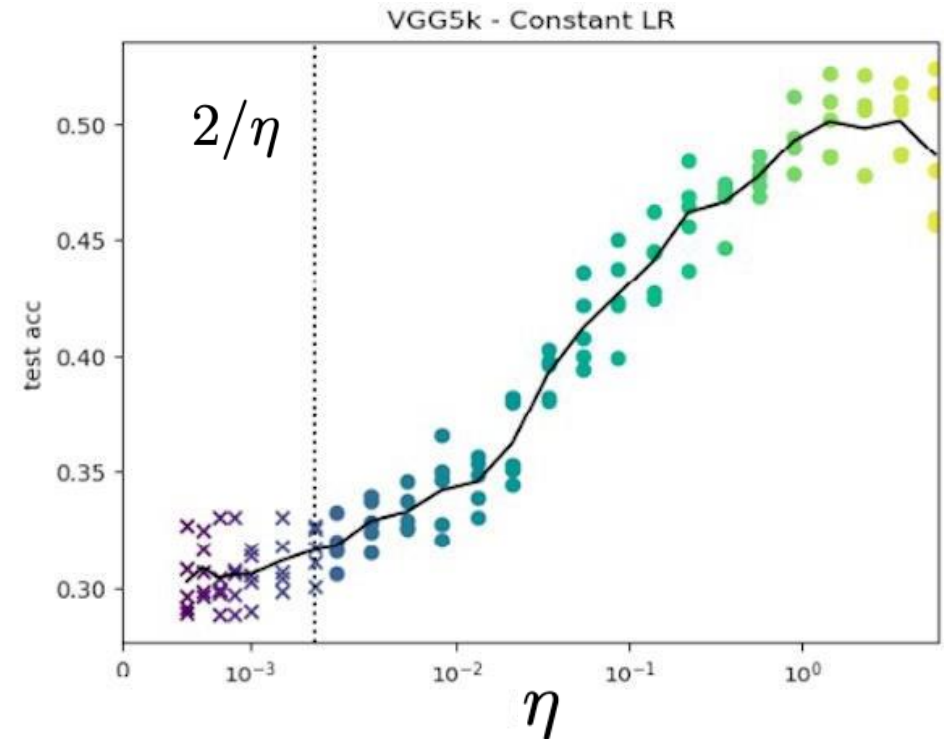
(Deep) Neural Networks – so long as we have instabilities – are self-regularizing

So long as we have instabilities...

We can promote this by having *large learning rates*, as the edge of stability is given by $2/\eta$



(a) MLP-fMNIST



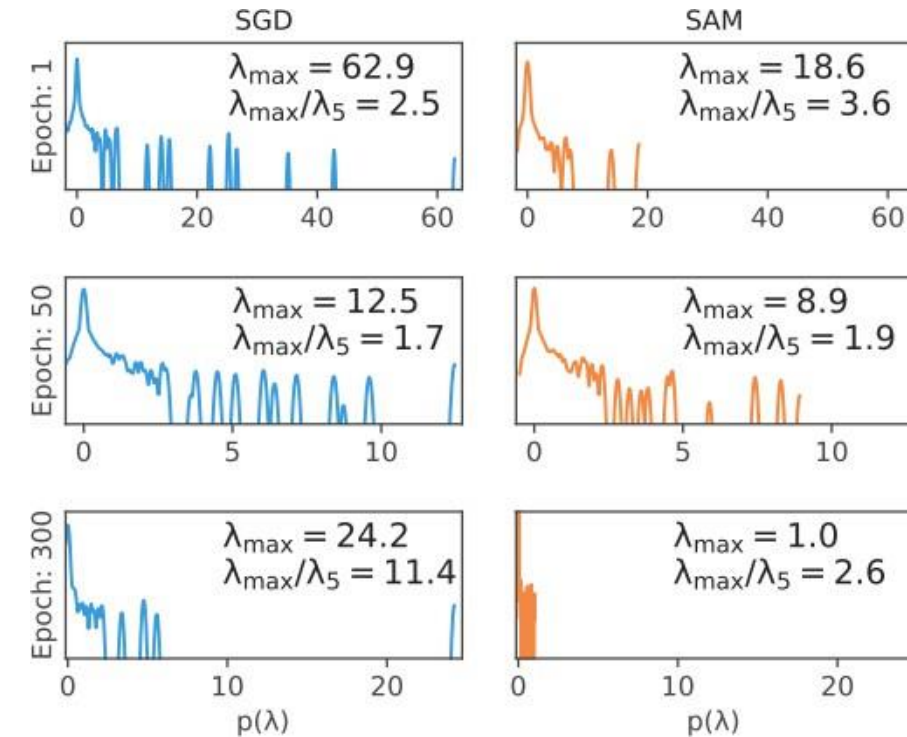
(b) VGG-CIFAR5k

Sharpness - revisited

Do we really want to tame λ_{\max} ?

No! As we want to induce instabilities

But, we want to tame the *number* of large λ



The notion of flatness has been challenged by Dinh et al. (2017), who argued that the different flatness measures proposed are not invariant under reparametrization of the parameter space and questioned the assumption that flatness directly causes generalization.

(Alison Pouplin, Hritik Roy, Sidak Pal Singh, Georgios Arvanitidis, On the curvature of the loss landscape, 2023)

On the Maximum Hessian Eigenvalue and Generalization

Simran Kaur[†], Jeremy Cohen[†], Zachary C. Lipton[†]

[†]Carnegie Mellon University
{skaur, jeremycohen, zlipton}@cmu.edu

May 25, 2023

SHARPNESS-AWARE MINIMIZATION FOR EFFICIENTLY IMPROVING GENERALIZATION

Pierre Foret *
Google Research
pierre.pforet@gmail.com

Ariel Kleiner
Google Research
akleiner@gmail.com

Hossein Mobahi
Google Research
hmobahi@google.com

Behnam Neyshabur
Blueshift, Alphabet
neyshabur@google.com



Back in 1991, at NeurlPS

Neural Information Processing Systems: *Natural & Synthetic*

Monday - Thursday, December 2 - 5, 1991; Denver, Colorado • Friday - Saturday, December 6 - 7, 1991; Vail, Colorado


The *Effective* Number of Parameters: An Analysis of Generalization and Regularization in Nonlinear Learning Systems

John E. Moody

Department of Computer Science, Yale University
P.O. Box 2158 Yale Station, New Haven, CT 06520-2158
Internet: moody@cs.yale.edu, Phone: (203)432-1200

The relationship between expected training set and expected test set errors for *linear models* trained using the *SSE* error function with no regularizer is well known in statistics (Akaike 1970, Barron 1984, Eubank 1988). The exact relation for test and training sets with density (9):

$$\langle \mathcal{E}_{test} \rangle_{\xi\xi'} = \langle \mathcal{E}_{train} \rangle_{\xi} + 2\sigma^2 \frac{p}{n} .$$

$$\sum_{\alpha=1}^p \frac{\kappa^{\alpha}}{\kappa^{\alpha} + \lambda}$$


$$\langle \mathcal{E}_{test}(\lambda) \rangle_{\xi\xi'} \approx \langle \mathcal{E}_{train}(\lambda) \rangle_{\xi} + 2\sigma_{eff}^2 \frac{p_{eff}(\lambda)}{n}$$

Moody referred to this as the **Generalized Prediction Error (GPE)**

Effective number of parameters

$$\sum_{\alpha=1}^p \frac{\kappa^\alpha}{\kappa^\alpha + \lambda}$$

MacKay also derived this
as part of his thesis

$$\gamma = k - \alpha \text{Trace} \mathbf{A}^{-1} = k - \sum_{a=1}^k \frac{\alpha}{\lambda_a + \alpha} = \sum_{a=1}^k \frac{\lambda_a}{\lambda_a + \alpha}.$$

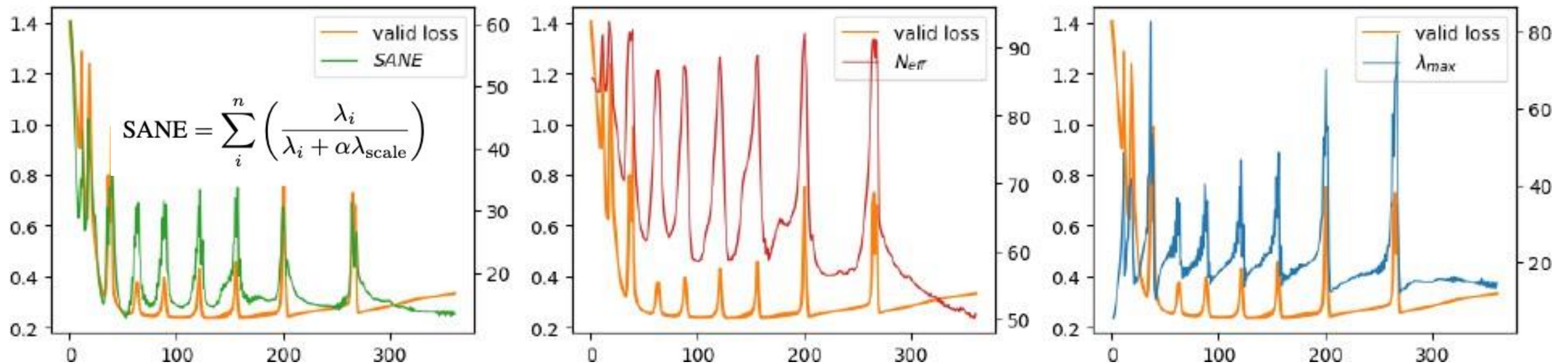


Figure from Lawrence Wang, Stephen J. Roberts (2023). SANE: The phases of gradient descent through Sharpness Adjusted Number of Effective parameters. <https://arxiv.org/abs/2305.18490>

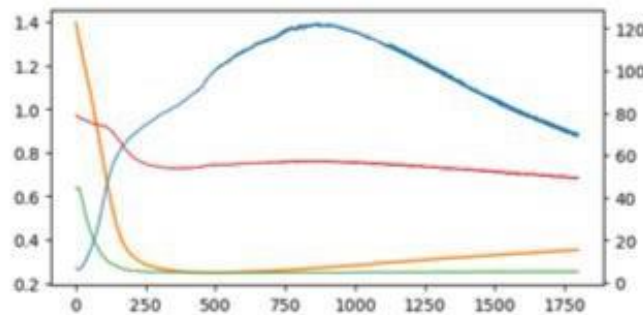
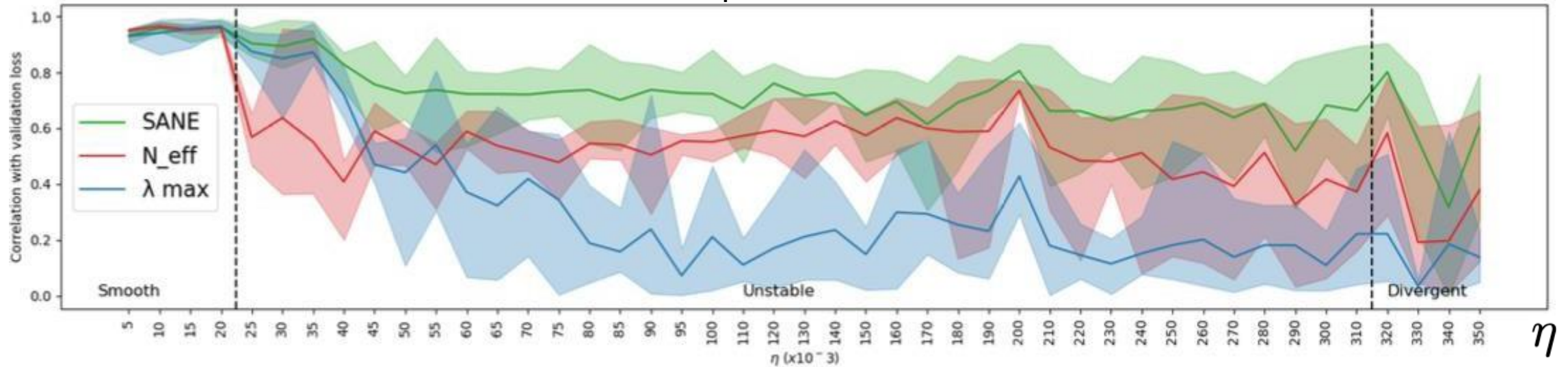
MacKay, David J.C. (1991) Bayesian methods for adaptive models. Dissertation (Ph.D.), California Institute of Technology.

John Moody, NIPS 1991. The Effective Number of Parameters: An Analysis of Generalization and Regularization in Nonlinear Learning Systems

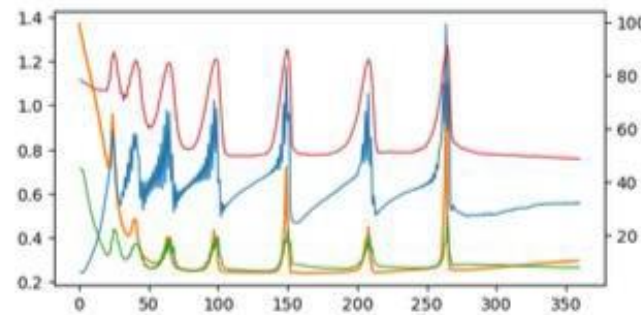
Effective number of parameters

$$\text{SANE} = \sum_i^n \left(\frac{\lambda_i}{\lambda_i + \alpha \lambda_{\text{scale}}} \right)$$

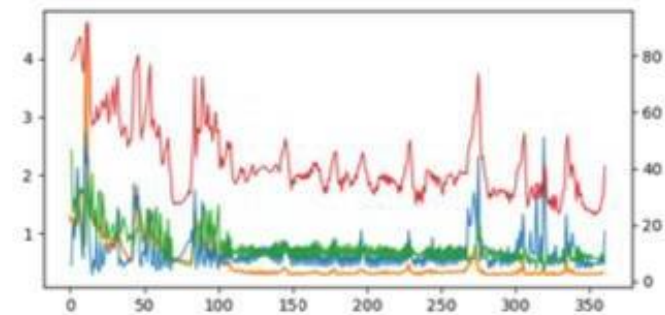
Correlation with out of sample loss



(a) $\eta = 0.01$, *smooth*



(b) $\eta = 0.05$, *low instability*



(c) $\eta = 0.35$, *high instability*

Figure from Lawrence Wang, Stephen J. Roberts (2023). SANE: The phases of gradient descent through Sharpness Adjusted Number of Effective parameters. <https://arxiv.org/abs/2305.18490>

What about the GPE?

$$\langle \mathcal{E}_{test}(\lambda) \rangle_{\xi\xi'} \approx \langle \mathcal{E}_{train}(\lambda) \rangle_{\xi} + 2\sigma_{eff}^2 \frac{p_{eff}(\lambda)}{n}$$

$$N_{\text{eff}} = \sum_i \frac{\lambda_i}{\lambda_i + \alpha} = \sum_i \frac{\lambda_i}{\nu_i}$$

$$\text{Tr}(G^{-1}H)$$

- Measure of sharpness (flatness) focused on the density of outliers
- Invariant under (affine) reparametrization
- SAM can be recovered as special case

The notion of flatness has been challenged by Dinh et al. (2017), who argued that the different flatness measures proposed are not invariant under reparametrization of the parameter space and questioned the assumption that flatness directly causes generalization.

(Alison Pouplin, Hritik Roy, Sidak Pal Singh, Georgios Arvanitidis, On the curvature of the loss landscape, 2023)

Could we improve on SAM?

Use a smoothed (diagonal) FIM to precondition the Hessian

→ G-ADAM & G-TRACER

$$L^{\mathcal{G}}(\mathbf{w}) = L(\mathbf{w}) + \rho \text{Tr}(\mathbf{G}^{-1} \mathbf{H}(\mathbf{w}))$$

(from G-TRACER, John Williams)

$$\begin{aligned} \mathbf{w} &\leftarrow \mathbf{w} - \alpha \nabla_{\mathbf{w}} [L(\mathbf{w}) + \rho \text{Tr}(\mathbf{G}^{-1} \mathbf{H}(\mathbf{w}))] \\ \mathbf{G} &\leftarrow (1 - \beta) \mathbf{G} + \beta \mathbf{F} \end{aligned}$$



Hessian regularization of deep neural networks: A novel approach based on stochastic estimators of Hessian trace

Yucong Liu^a, Shixing Yu^b, Tong Lin^c ✉

Of course, this is a popular topic!

* Diego Granziol, Stefan Zohren, Stephen Roberts (2022). Learning Rates as a Function of Batch Size: A Random Matrix Theory Approach to Neural Network Training. *Journal of Machine Learning Research* 23(173):1-65

* Diego Granziol, Nicholas Baskerville, Xingchen Wan, Samuel Albanie, Stephen Roberts (2024). Iterative Averaging in the Quest for Best Test Error. *Journal of Machine Learning Research (JMLR)*, 25(20):1-55

* John Williams, Stephen Roberts (2023). G-TRACER Expected Sharpness Optimization. <https://arxiv.org/abs/2306.13914>

Yes!

Table 3: CIFAR-100: ResNet20, accuracy (standard error)

	no aug	with aug	50% noise & no aug
SGD	51.43 % (0.41)	70.02% (0.36)	21.96% (0.36)
SAM	58.98 % (0.52)	70.33% (0.22)	49.89% (0.32)
SGD-TRACER	63.47% (0.32)	70.71% (0.36)	51.62% (0.18)

Table 4: CIFAR-100: ViT, accuracy (standard error)

	with aug
SGD	37.7 % (0.71)
SAM	38.2 % (0.52)
SAM batch-split	38.7 % (0.44)
SGD-TRACER	39.1 % (0.32)
SGD-TRACER batch-split	41.6 % (0.28)

Table 5: NLP tasks BERT base-uncased results, accuracy (standard error)

	BOOLQ	WIC	RTE
Adam	73.84% (0.14)	69.36% (0.08)	69.18% (0.33)
SAM	73.95% (0.13)	69.06% (0.07)	69.54% (0.28)
Adam-TRACER	75.09% (0.04)	70.01% (0.06)	70.13% (0.18)

In conclusion

- Learning to **generalize** from complex data sets requires thought!
- **Instabilities in learning**, far from being a problem, are **beneficial to generalization**
- If we harness **unstable dynamics** in learning, we **avoid costly hyper-parameter tuning**
- (D)NNs **self-regularize** if allowed (consider **very large learning rates!**)

“Civilization advances by extending the number of important operations which we can perform without thinking of them.”

- Alfred North Whitehead

Thank you!

