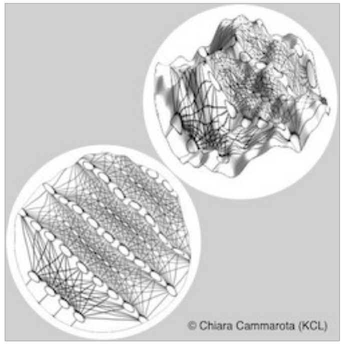
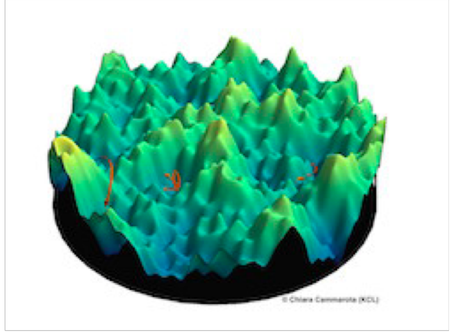
Measuring the Spectrum of Deepnet Hessians

Vardan Papyan
Postdoc advisor: David Donoho

The Rough High-Dimensional Landscape Problem

UC SANTA BARBARA
Kavli Institute for Theoretical Physics



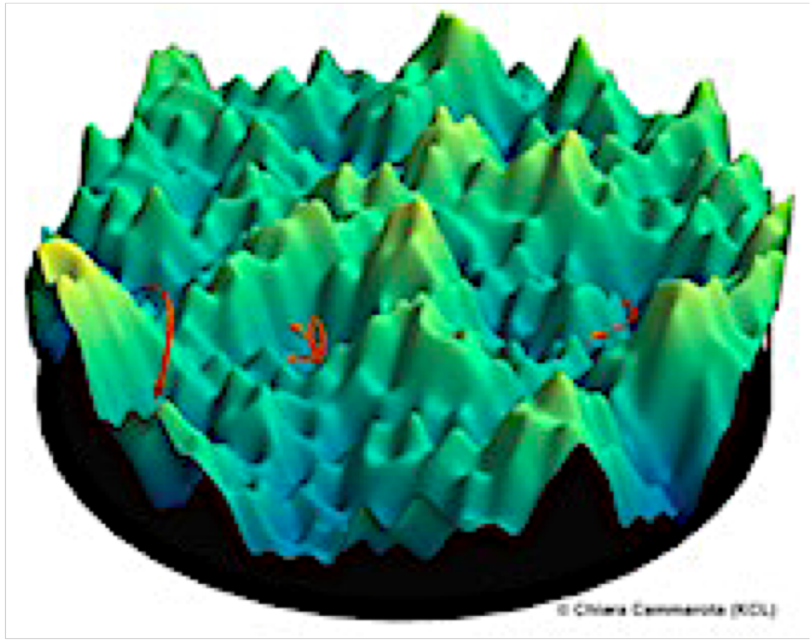


Outline

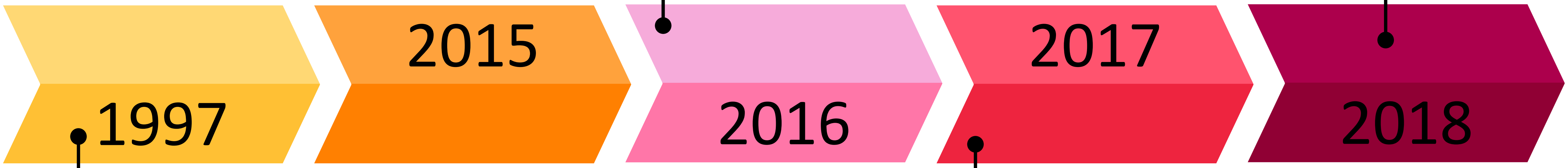
- Rough landscapes in deep learning
- Hessians in deep learning
- Measurements of Hessians at large scale
- Structure in the outliers

Deepnet Loss surfaces have rough landscapes, however…

- **Traditional notion of landscape assumes:**
  - One is exploring the *whole* landscape
- **Deep learning:**
  - Run SGD and converge to some solution
  - Observe a range of behaviors *along the path*
  - No exploration of the *whole* landscape
- **This talk:**
  - The global topology of the landscape will not be at issue
  - The path will not be at issue
  - We focus on the *converged solution*



Landscapes & generalization performance



Flat Minima
Hochreiter & Schmidhuber

flat minima lead to better generalization

On Large-Batch Training for Deep Learning: Generalization Gap and Sharp Minima
Keskar et. al

large batch SGD leads to sharp minima

Sharp Minima Can Generalize For Deep Nets
Dinh et. al

most notions of flatness are problematic

Hessian-based Analysis of Large Batch Training and Robustness to Adversaries
Yao et. al

large batch SGD converges to higher Hessian spectrum

Landscapes & speed of training

1997

2015

2016

Entropy-SGD: Biasing Gradient Descent Into Wide Valleys
Chaudhari et. al

modification to SGD that favors flat minima

2017

Three Factors Influencing Minima in SGD
Jastrzębski et. al

$$\text{generalization} \approx \text{flatness} \approx \frac{\text{learning rate}}{\text{batch size}}$$

2018

Gradient Descent Happens in a Tiny Subspace
Gur-Ari et. al

gradients of SGD spanned by top eigenvectors of the Hessian

An Empirical Model of Large-Batch Training
McCandlish et. al

$\frac{tr(H\Sigma)}{G^T HG}$ *predicts the largest "useful" batch size*

Landscapes & optimization guarantees

The Loss Surfaces of Multilayer Networks
Choromanska, Henaff, Mathieu, Ben Arous & LeCun

lowest critical values are located in a band near the global minimum

1997 — 2015 — 2016 — 2017 — 2018

Geometry of Neural Network Loss Surfaces via Random Matrix Theory
Pennington & Bahri

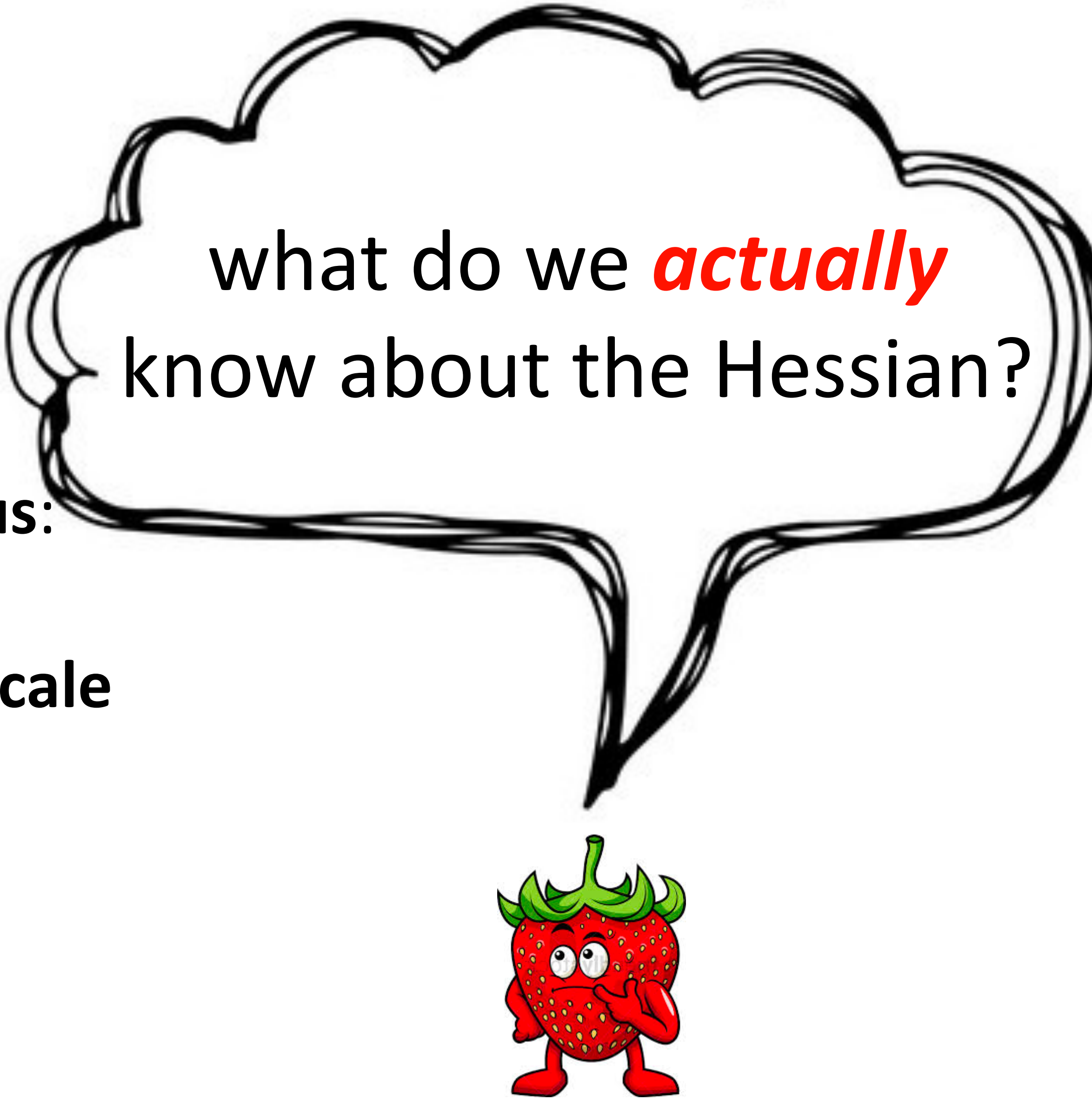
number of negative eigenvalues at critical points of small index scales like the 3/2 power of the energy

Outline

- Rough landscapes in deep learning
- Hessians in deep learning
- Measurements of Hessians at large scale
- Structure in the outliers

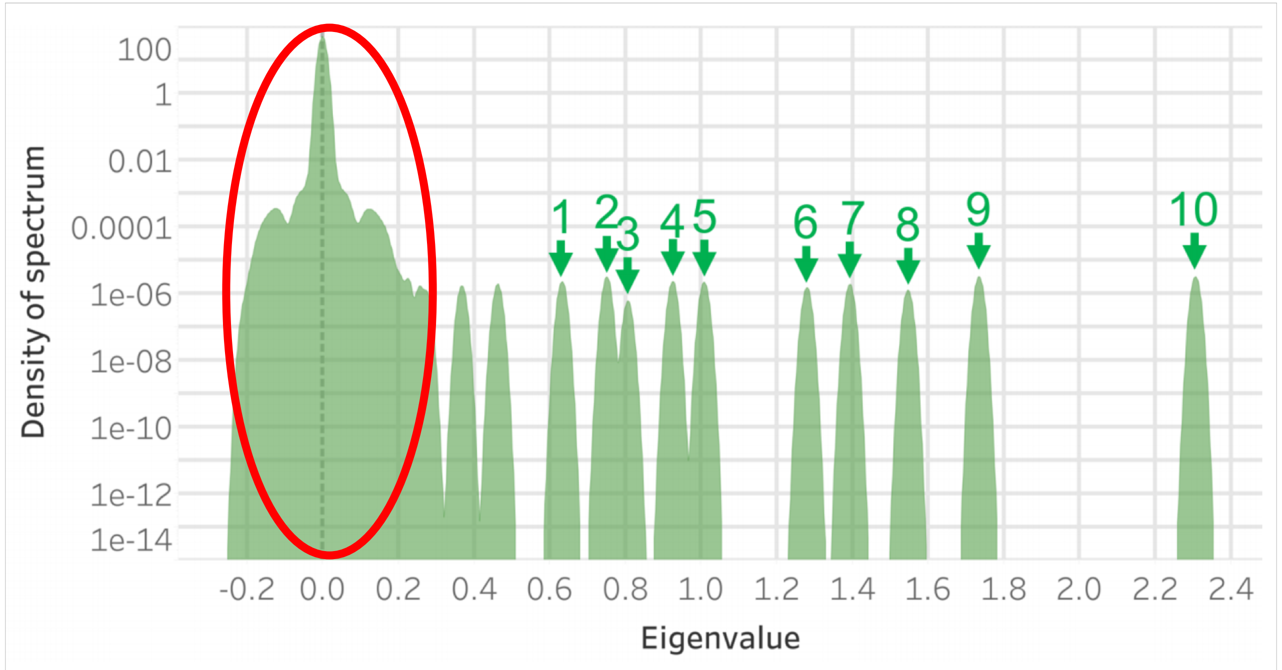
Today’s question

- Properties of Hessian crucial to:
  - Generalization performance
  - Training speed
  - Optimization guarantees
- Hessians of today's deepens **enormous**: e.g., 30 million x 30 million!
- Not previously widely studied **at full scale**



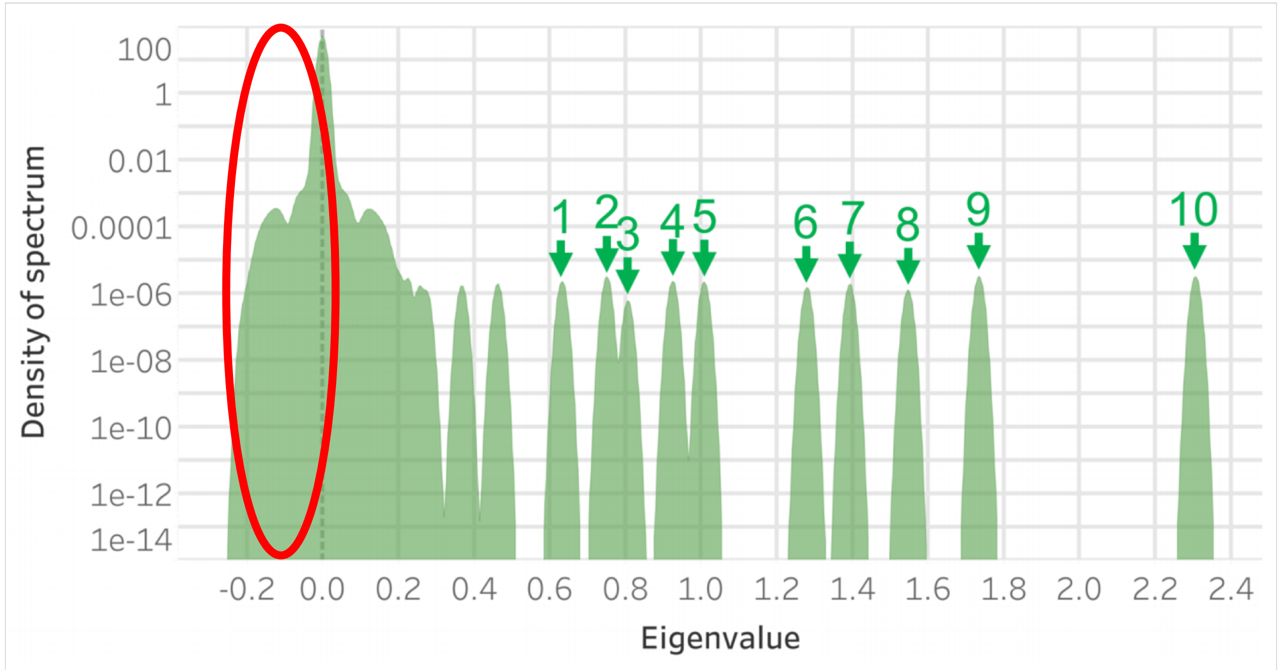
Slogans about eigenvalue distributions

- Bulk distribution



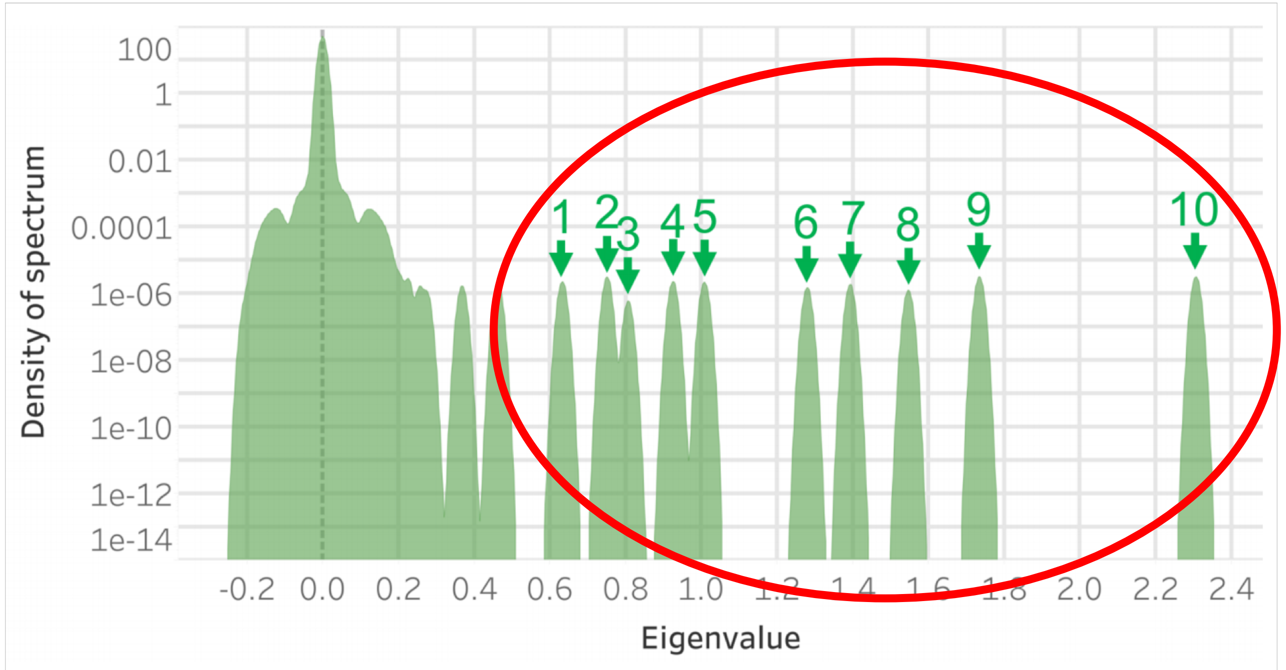
Slogans about eigenvalue distributions

- Bulk distribution
- Many negative eigenvalues



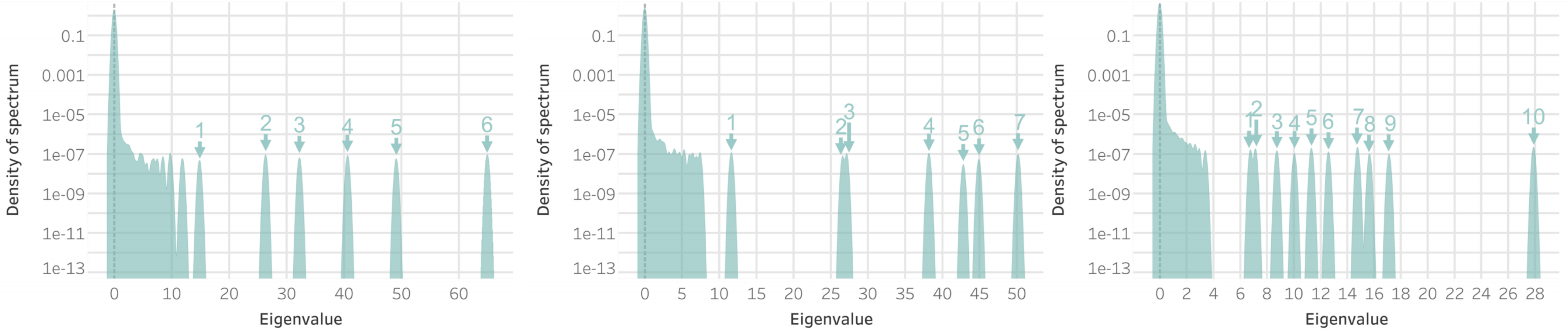
Slogans about eigenvalue distributions

- Bulk distribution
- Many negative eigenvalues
- Number of outliers = number of classes



Slogans about eigenvalue distributions

- Bulk distribution
- Many negative eigenvalues
- Number of outliers = number of classes
- Scaling of outliers with training/sample size



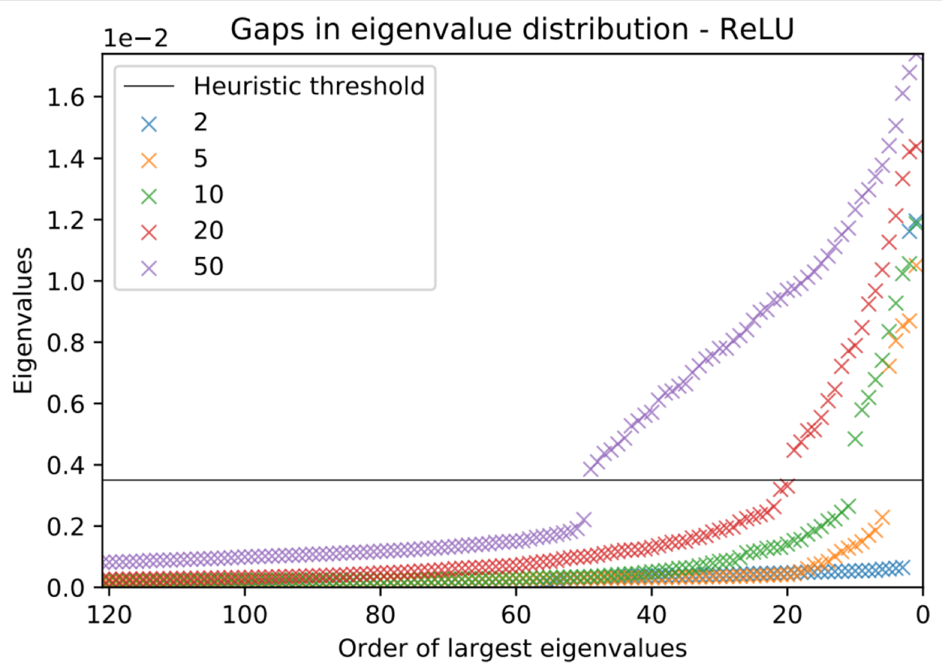
(a) 10 examples per class. (b) 51 examples per class. (c) 506 examples per class.

Outline

- Rough landscapes in deep learning
- Hessians in deep learning
- Measurements of Hessians at large scale
- Structure in the outliers

Number of outliers = number of classes

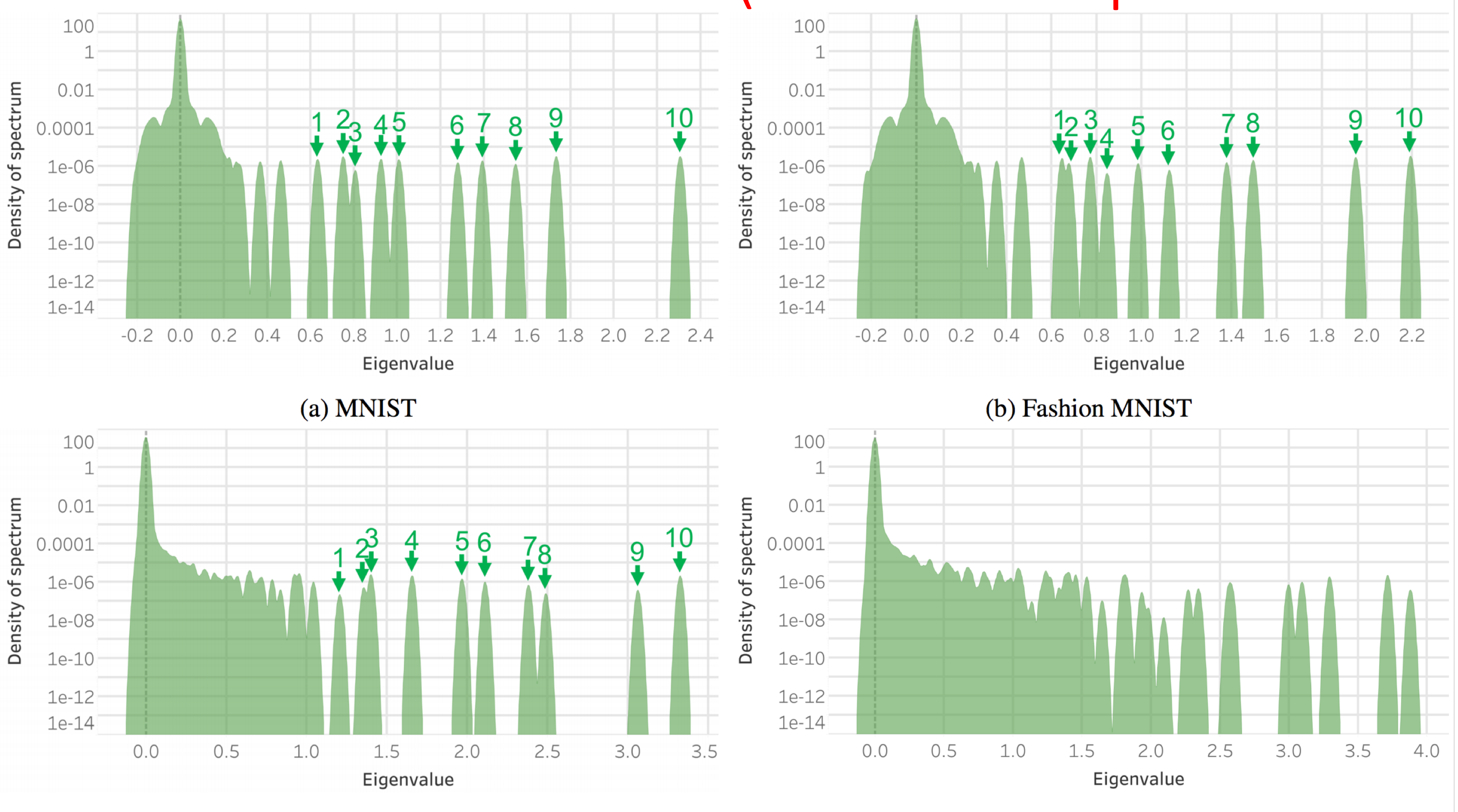
- Empirical Analysis of the Hessian of Over-Parametrized Neural Networks [Sagun et. al ’17]
- 100 dimensional Gaussian mixture model with $C \in \{2,5,10,20,50\}$ classes
- Two hidden layers with 30 neurons each



Measurements at scale

- Recent paper
- The Full Spectrum of Deep Net Hessians At Scale: Dynamics With Sample Size [Papyan ‘18]
- https://arxiv.org/abs/1811.07062

Train Hessian of VGG11 (28 million parameters)



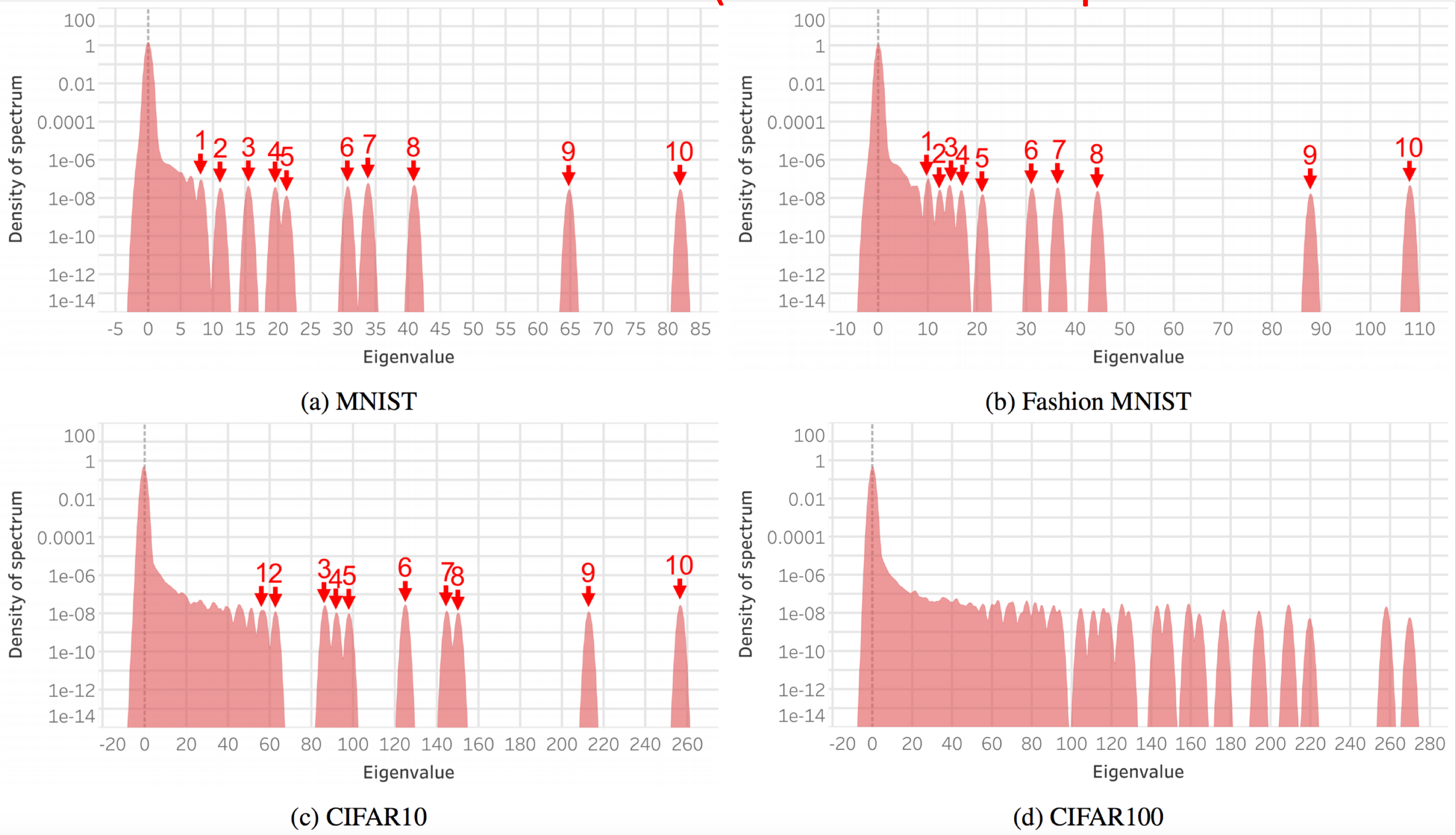
(a) MNIST

(b) Fashion MNIST

(c) CIFAR10

(d) CIFAR100

Test Hessian of VGG11 (28 million parameters)



(a) MNIST

(b) Fashion MNIST

(c) CIFAR10

(d) CIFAR100

Decomposing the Hessian into two components

$$\text{Hessian} = \underbrace{\text{Ave}_{i,c}\left\{\frac{\partial \ell(z;\theta)}{\partial z}\bigg|_{z=f(x_{i,c};\theta)} \frac{\partial^2 f(x_{i,c};\theta)}{\partial^2 \theta}\right\}}$$

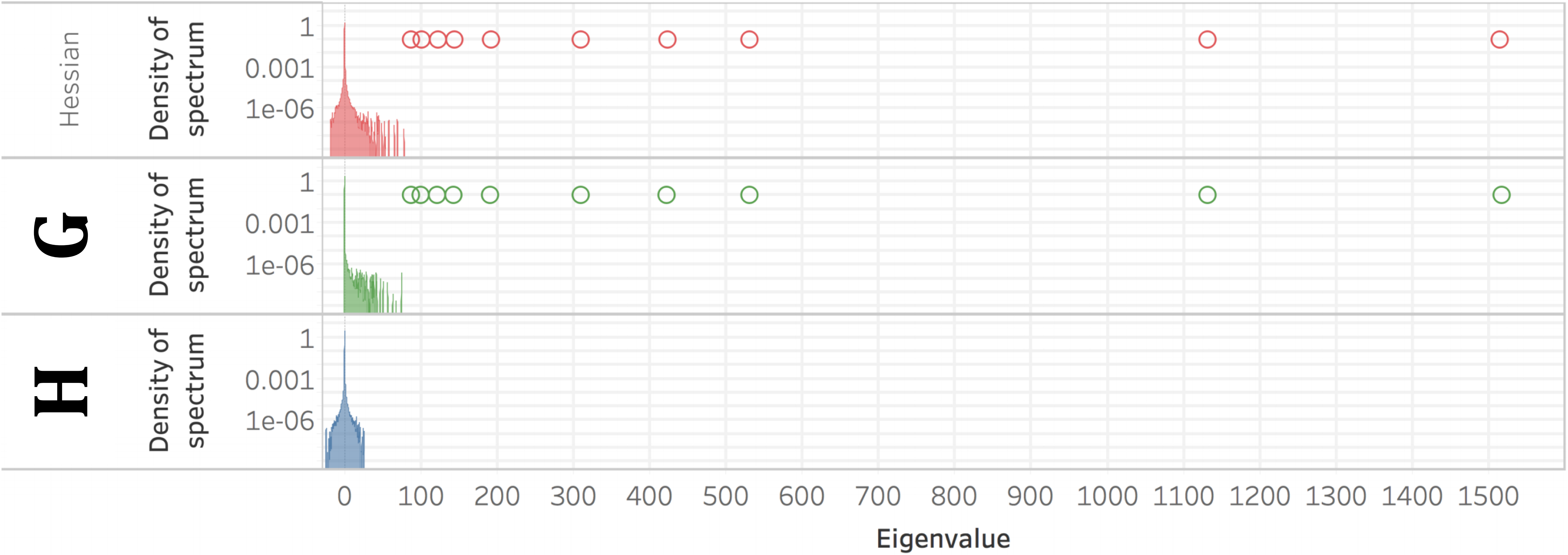
H – Hessian of predictions

$$+\underbrace{\text{Ave}_{i,c}\left\{\frac{\partial f(x_{i,c};\theta)}{\partial \theta}^T \frac{\partial^2 \ell(z;\theta)}{\partial z^2}\bigg|_{z=f(x_{i,c};\theta)} \frac{\partial f(x_{i,c};\theta)}{\partial \theta}\right\}}$$

G – covariance of gradients

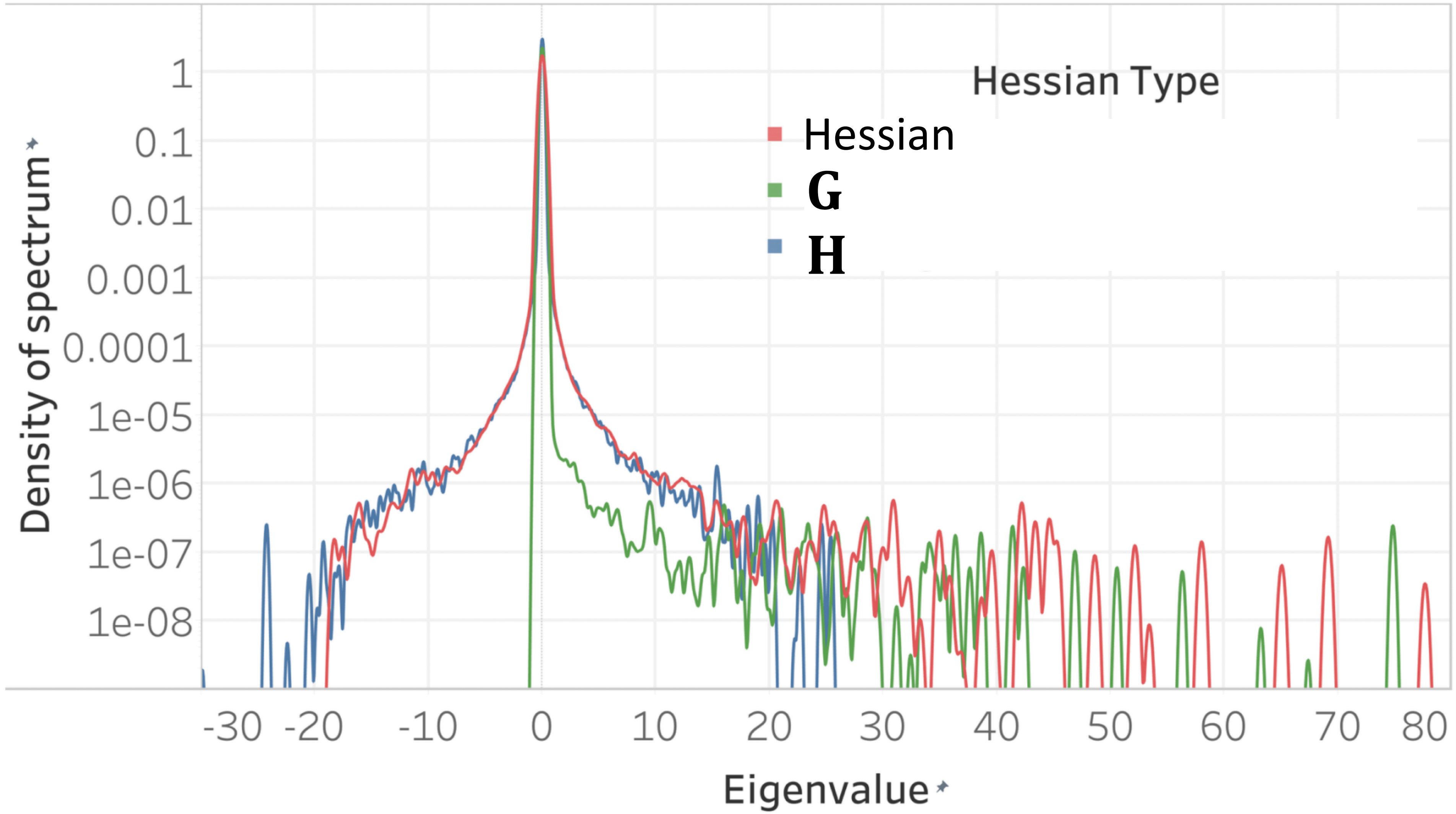
Attribution of outliers

VGG11 trained on MNIST sub-sampled to 2599 examples per class



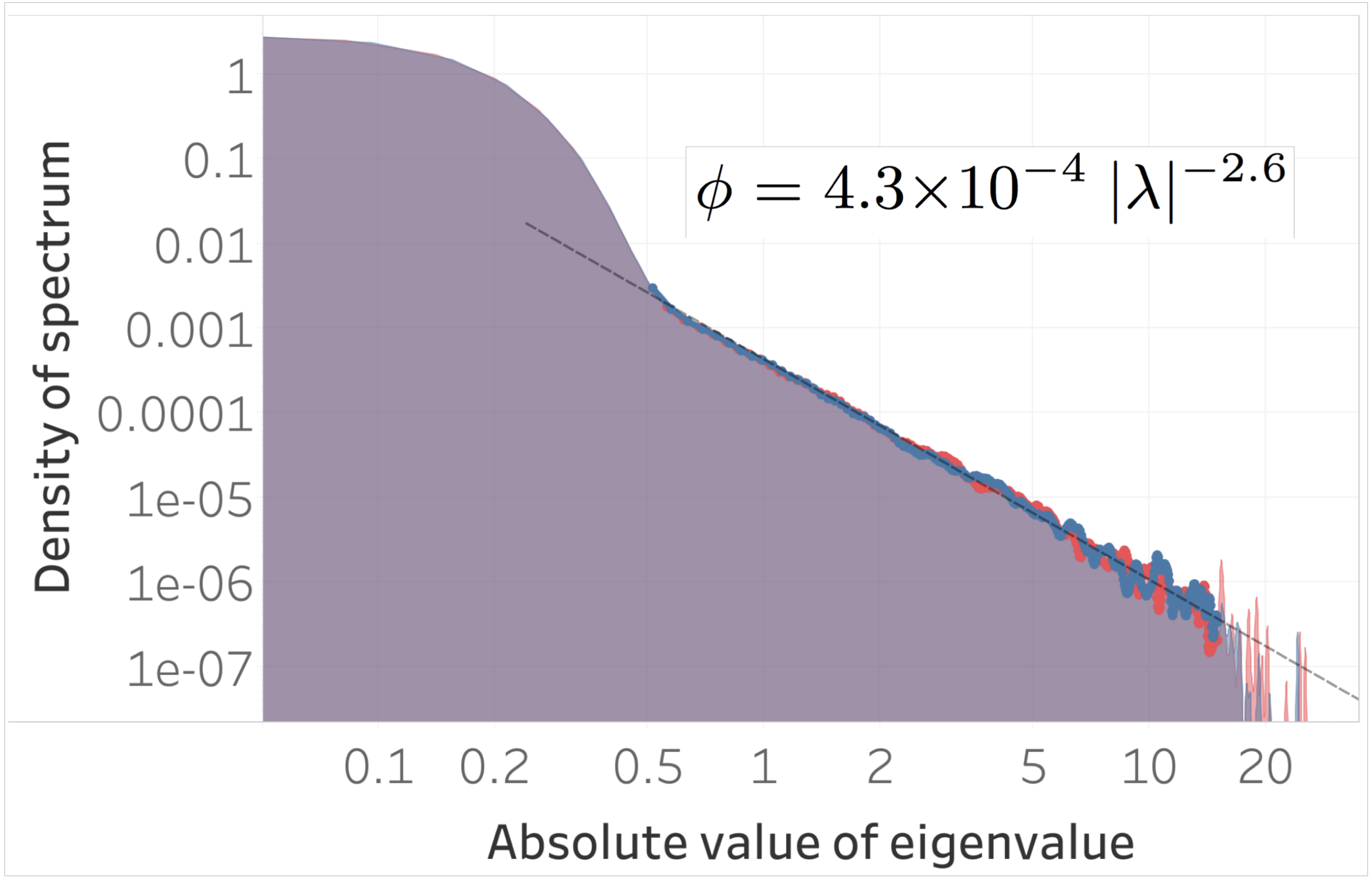
Attribution of bulk

VGG11 trained on MNIST sub-sampled to 2599 examples per class



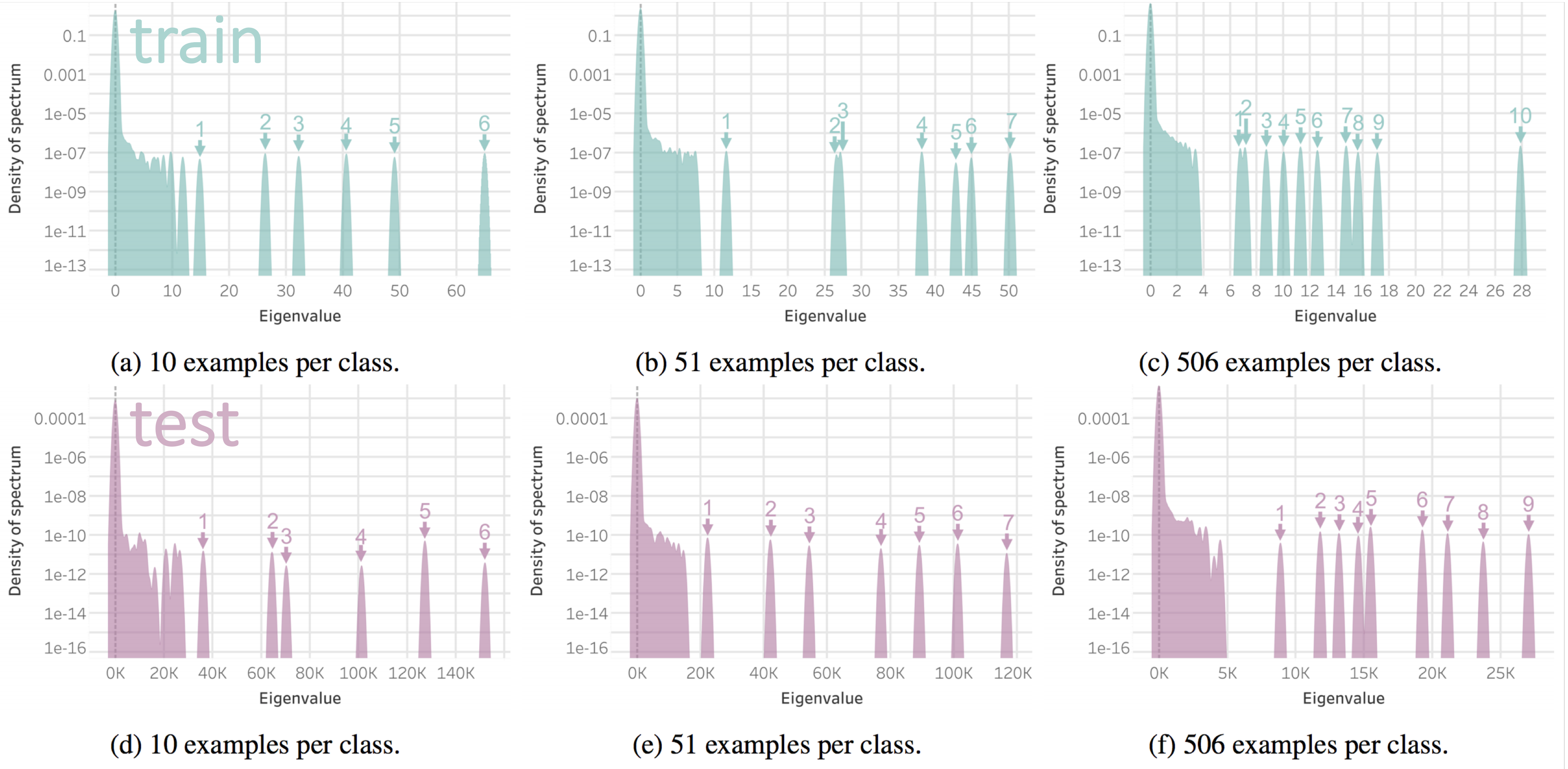
Tail properties

VGG11 trained on MNIST sub-sampled to 2599 examples per class



Scaling of outliers with training/sample size

VGG11 trained on CIFAR10



(a) 10 examples per class. (b) 51 examples per class. (c) 506 examples per class.

(d) 10 examples per class. (e) 51 examples per class. (f) 506 examples per class.

We show here

- Outliers are induced by G, covariance of gradients
- Bulk is induced by H, hessian of predictions
- Tail of bulk follows power law

How did we make measurements at such massive scale?

- Algorithms that **do not** work:
  - Power method – will get you 1/30,000,000 eigenvalues
  - Subspace iteration – will get you 10/30,000,000 eigenvalues
  - SVD – will get you spectra of **small** Hessians (thousands of eigenvalues)
- Comparison:
  - Previous work: thousands of parameters
  - Our work: 30 million parameters
- How???

How did we make measurements at such massive scale?

- 1970's:

  Quantum mechanics and physicists study the energy levels of Hamiltonians

- 2018:

  We leverage these ideas to ***approximate*** the spectrum of deepnet Hessians

- Survey of algorithms used [Approximating Spectral Densities of Large Matrices, '14]

Lanczos

- We implemented Lanczos in PYTORCH
- Many non-trivial engineering tricks
- We plan to release a package so anyone can compute spectra of deepnet Hessians
- Complexity similar to training a model

Algorithm 2: FastLanczos(H, M)

Input: Linear operator $H \in \mathbb{R}^{p \times p}$ with spectrum in the range $[-1, 1]$.
Number of iterations M.

Result: Eigenvalues and eigenvectors of the tridiagonal matrix T_m.

for $m = 1, \ldots, M$ **do**
  if $m == 1$ **then**
    sample $v \sim \mathcal{N}(0, I)$;
    $v = \frac{v}{\|v\|_2}$;
    $v_{\text{next}} = Hv$;
  else
    $v_{\text{next}} = Hv - \beta_{m-1} v_{\text{prev}}$;
  end
  $\alpha_m = v_{\text{next}}^T v$;
  $v_{\text{next}} = v_{\text{next}} - \alpha_m v$;
  $\beta_m = \|v_{\text{next}}\|_2$;
  $v_{\text{next}} = \frac{v_{\text{next}}}{\beta_m}$;
  $v_{\text{prev}} = v$;
  $v = v_{\text{next}}$;
end

$$T_M = \begin{bmatrix} \alpha_1, & \beta_1, & & & \\ \beta_1, & \alpha_2, & \beta_2, & & \\ & \beta_2, & \alpha_3, & & \\ & & & \ddots & \beta_{M-1} \\ & & & \beta_{M-1} & \alpha_M \end{bmatrix};$$

$\{\theta_m\}_{m=1}^M, \{y_m\}_{m=1}^M = \text{eig}(T_M)$;

return $\{\theta_m\}_{m=1}^M, \{y_m\}_{m=1}^M$;

Outline

- Rough landscapes in deep learning
- Hessians in deep learning
- Measurements of Hessians at large scale
- Structure in the outliers

Importance of spectral outliers

- Outliers due to $\mathbf{G}$
- The only generalizable eigenspaces of $\mathbf{G}$ are the outlying ones
- Gur-Ari et. al show that SGD trapped in tiny subspace
- This subspace is outlier subspace!

Insights

- Gradients have structured means
- Mean structure induces outliers
- Outliers cause low-dimensionality
- Low-dimensionality causes slow SGD
- Possibilities to exploit means in SGD?

What is causing the outliers to appear?

- Recent paper
- A Three-Level Hierarchical Model for the Outliers in the Spectrum of Deepnet Hessians
- arXiv posting soon

Observation 1: gradient vectors have a structure on indices

- δ_k = gradient induced by k-th element
- Coordinate index k has this structure: $c, c', i = [c(k), c'(k), i(k)]$
  - i = observation (i.e. image)
  - c = true class of observation
  - c' = classifier coordinate (i.e. potential class)

  there are $I \times C \times C$ elements
- Define:

$$\delta_{c,c'} = \text{Ave}\{\delta_k : c(k) = c, c'(k) = c'\}$$
$$\Sigma_{c,c'} = \text{Covar}\{\delta_k : c(k) = c, c'(k) = c'\}$$

- Gradient induced by observation k is sampled from population with mean $\delta_{c,c'}$ and covariance $\Sigma_{c,c'}$

Observation 2:
$\mathbf{G}$ is a second moment matrix

$$\mathbf{G} = c \sum_{k} \delta_k \delta_k^T$$

$$= \sum_{c} \sum_{c'} \sum_{i} \delta_{i,c,c'} \delta_{i,c,c'}^T$$

$$= c \sum_{c} \sum_{c'} \delta_{c,c'} \delta_{c,c'}^T + c \sum_{c} \sum_{c'} \Sigma_{\mathrm{c,c'}}$$

Observation 3:
The means $\delta_{c,c'}$ themselves have structure

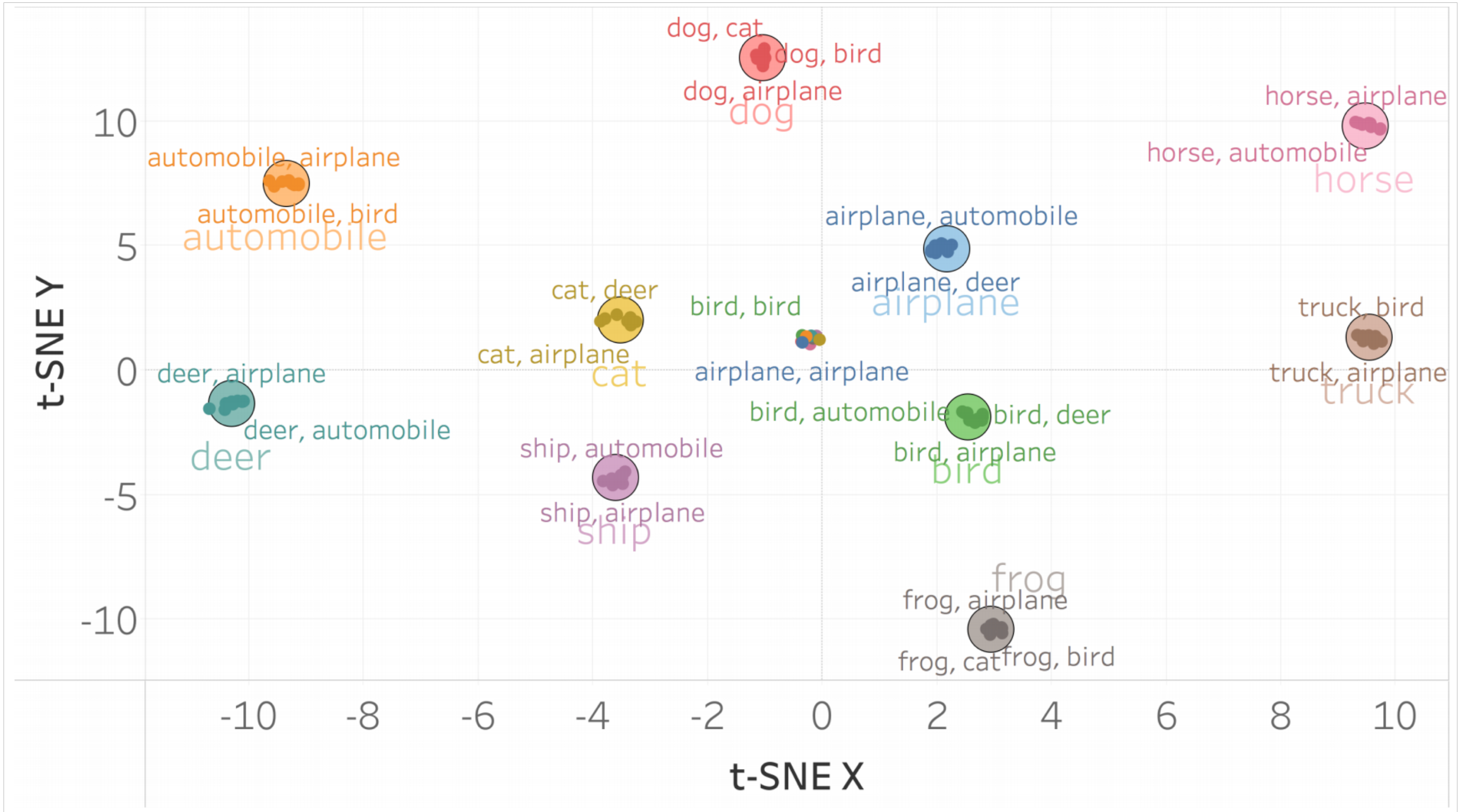
- Define:

$$\delta_c = \text{Ave}\{\delta_{c,c'} \,:\, c' = 1, \ldots, C\}$$

- The means $\delta_{c,c'}$ can be viewed as sampled from a population with mean δ_c

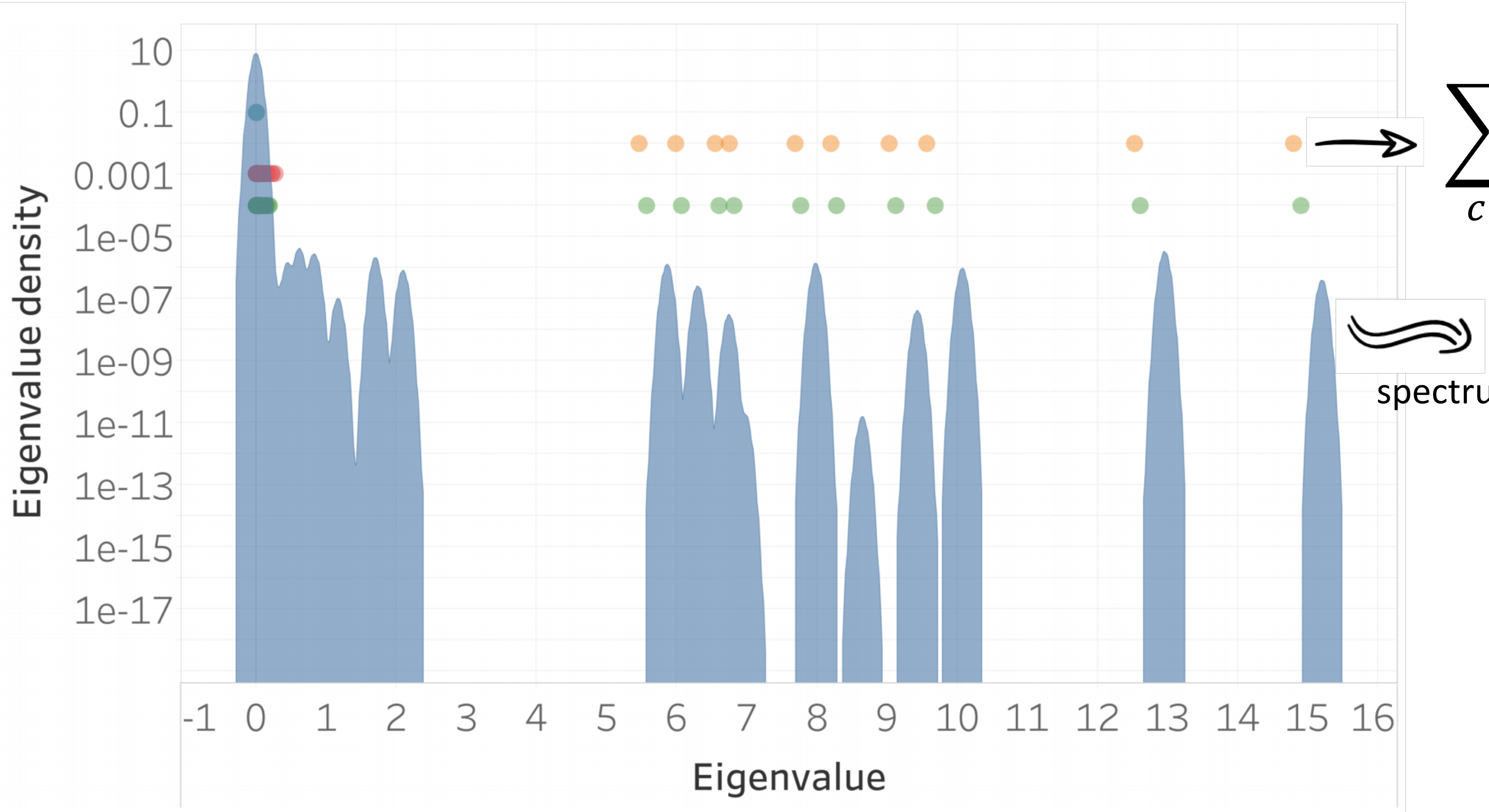
t-SNE visualization of $\{\delta_c\}_c$ and $\{\delta_{c,c'}\}_{c,c'}$

ResNet18 trained on CIFAR10 with 365 examples per class



Decomposing $\mathbf{G}$

ResNet18 trained on MNIST with 702 examples per class



$$\sum_c \delta_c \delta_c^T$$

spectrum approximation of $\mathbf{G}$

Proof that mean structure induces outliers

- Spiked second moment model [Benaych-Georges and Raj Rao Nadakuditi, '09]
- $P + ZZ^T$
- Z or P orthogonally invariant
- P low rank

Deliverables

- Measurements of spectral distributions of Hessians of **modern deepnets** at full scale on **real data**
- Confirmation of characteristics observed in toy models:
  - Bulk
  - Negative eigenvalues
  - C outliers
- Attribution of characteristics to substructure of gradient and hessian:
  - Bulk and negative eigenvalues due to H - hessian of predictions
  - Outliers due to second moment of gradients G
  - Outliers due to mean structure of embedding
- Exciting opportunities for optimization
  - We are told Google researchers have made related observations (e.g. Behrooz Ghorbani and the team he works in)