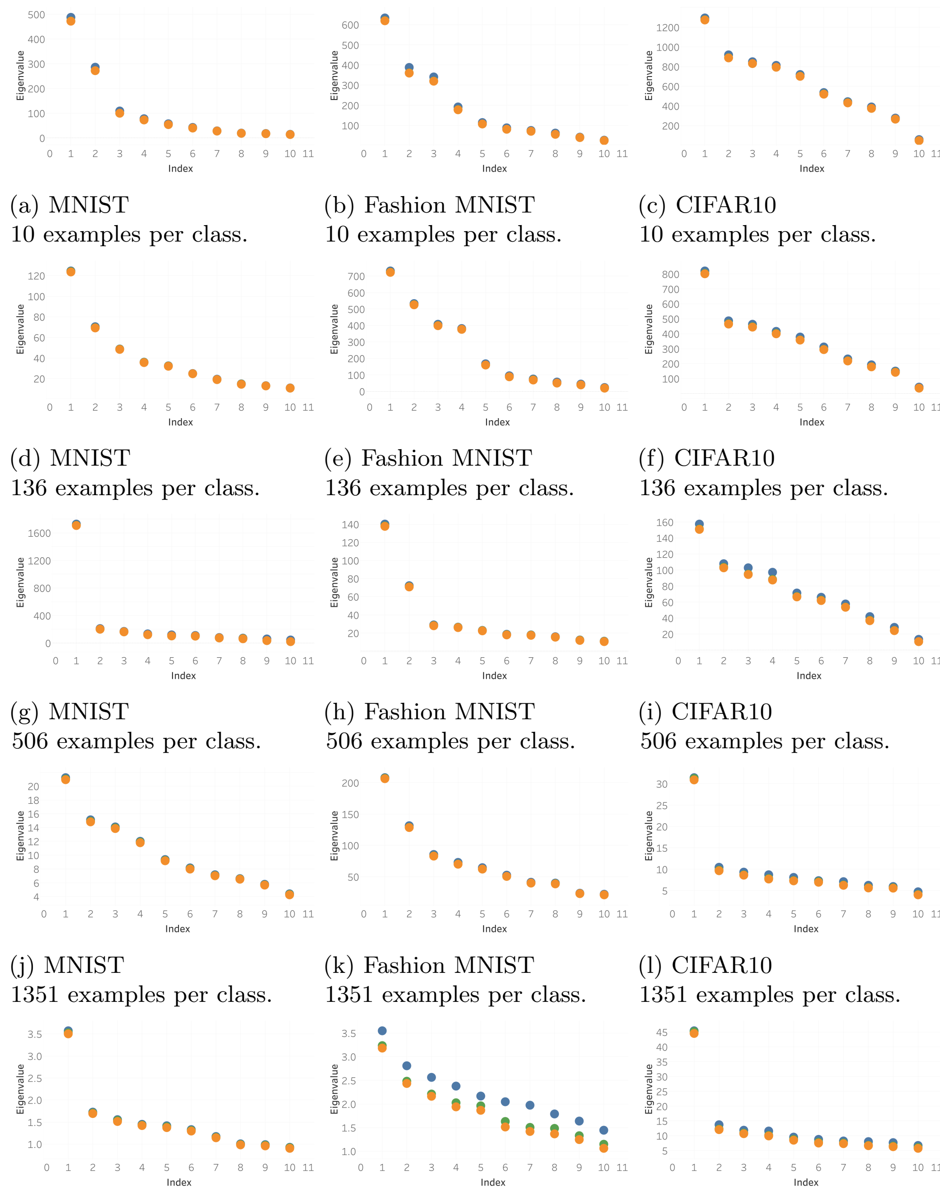
Measurements of Three-Level Hierarchical Structure in the Outliers in the Spectrum of Deepnet Hessians Appendix

p	number of parameters
C	number of classes
n	training examples
n_c	training examples per class
$x_{i,c}$	i-th example in c-th class
y_c	one hot vector corresponding to the c-th class
$\theta \in \mathbb{R}^p$	concatenation of all the parameters
$f(x_{i,c};\theta) \in \mathbb{R}^C$	logits (predictions prior to softmax) of $x_{i,c}$
$f_{c'}(x_{i,c};\theta) \in \mathbb{R}$	c'-th logit of $x_{i,c}$
$\frac{\partial f(x_{i,c};\theta)}{\partial \theta} \in \mathbb{R}^{C \times p}$	logit derivatives of $x_{i,c}$
$\frac{\partial f_{c'}(x_{i,c};\theta)}{\partial \theta} \in \mathbb{R}^p$	c'-th logit derivative of $x_{i,c}$
$p(x_{i,c};\theta) \in \mathbb{R}^C$	$\text{Softmax}(f(x_{i,c};\theta))$
$p_{c'}(x_{i,c};\theta) \in \mathbb{R}$	c'-th entry of $\text{Softmax}(f(x_{i,c};\theta))$
$\delta_{i,c,c'} \in \mathbb{R}^p$	$\sqrt{p_{c'}(x_{i,c};\theta)}(y_{c'} - p(x_{i,c};\theta))^T \frac{\partial f(x_{i,c};\theta)}{\partial \theta}$
$\delta_{c,c'} \in \mathbb{R}^p$	$\text{Ave}_i \left\{\delta_{i,c,c'}\right\}$
$\Sigma_{c,c'} \in \mathbb{R}^{p \times p}$	$\text{Ave}_i \left\{(\delta_{i,c,c'} - \delta_{c,c'})(\delta_{i,c,c'} - \delta_{c,c'})^T\right\}$
$\delta_c \in \mathbb{R}^p$	$\text{Ave}_{c' \neq c}\{\delta_{c,c'}\}$
$\Sigma_c \in \mathbb{R}^{p \times p}$	$\text{Ave}_{c' \neq c}\{(\delta_{c,c'} - \delta_c)(\delta_{c,c'} - \delta_c)^T\}$
$G_0 \in \mathbb{R}^{p \times p}$	$\text{Ave}_c\{\delta_{c,c}\delta_{c,c}^T\}$
$G_1 \in \mathbb{R}^{p \times p}$	$(C-1)\,\text{Ave}_c\left\{\delta_c\delta_c^T\right\}$
$G_2 \in \mathbb{R}^{p \times p}$	$(C-1)\,\text{Ave}_c\left\{\Sigma_c\right\}$
$G_3 \in \mathbb{R}^{p \times p}$	$\frac{1}{C}\sum_{c,c'} \Sigma_{c,c'}$

Table 1: Summary of notations.



(a) MNIST
10 examples per class.

(b) Fashion MNIST
10 examples per class.

(c) CIFAR10
10 examples per class.

(d) MNIST
136 examples per class.

(e) Fashion MNIST
136 examples per class.

(f) CIFAR10
136 examples per class.

(g) MNIST
506 examples per class.

(h) Fashion MNIST
506 examples per class.

(i) CIFAR10
506 examples per class.

(j) MNIST
1351 examples per class.

(k) Fashion MNIST
1351 examples per class.

(l) CIFAR10
1351 examples per class.

(m) MNIST
3605 examples per class.

(n) Fashion MNIST
3605 examples per class.

(o) CIFAR10
3605 examples per class.

Figure 1: *Scree plots of G_1, G_{1+2} and G for the VGG11 architecture.* Each column of panels corresponds to a different dataset, and each row to a different sample size. Each panel plots the top-C eigenvalues of G_1 in orange, G_{1+2} in **green** and G in **blue**. The top eigenvalues in G – which correspond to the outliers in the approximated spectrum of G – were computed using the LowRankDeflation procedure. For every $1 \leq c \leq C$, we have $\lambda_c(G) \geq \lambda_c(G_{1+2}) \geq \lambda_c(G_1)$. Moreover, $\lambda_c(G_{1+2})$ and $\lambda_c(G_1)$ are usually very close.