

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 $x_{i,c}$
$p(x_{i,c}; \theta) \in \mathbb{R}^C$	Softmax($f(x_{i,c}; \theta)$)
$p_{c'}(x_{i,c}; \theta) \in \mathbb{R}$	c' -th entry of Softmax($f(x_{i,c}; \theta)$)
$\delta_{i,c,c'} \in \mathbb{R}^p$	$\sqrt{p_{c'}(x_{i,c}; \theta)}(y_{c'} - p(x_{i,c}; \theta))^T \frac{\partial f(x_{i,c}; \theta)}{\partial \theta}$
$\delta_{c,c'} \in \mathbb{R}^p$	Ave $_i \{ \delta_{i,c,c'} \}$
$\Sigma_{c,c'} \in \mathbb{R}^{p \times p}$	Ave $_i \{ (\delta_{i,c,c'} - \delta_{c,c'}) (\delta_{i,c,c'} - \delta_{c,c'})^T \}$
$\delta_c \in \mathbb{R}^p$	Ave $_{c' \neq c} \{ \delta_{c,c'} \}$
$\Sigma_c \in \mathbb{R}^{p \times p}$	Ave $_{c' \neq c} \{ (\delta_{c,c'} - \delta_c) (\delta_{c,c'} - \delta_c)^T \}$
$G_0 \in \mathbb{R}^{p \times p}$	Ave $_c \{ \delta_{c,c} \delta_{c,c}^T \}$
$G_1 \in \mathbb{R}^{p \times p}$	$(C - 1)$ Ave $_c \{ \delta_c \delta_c^T \}$
$G_2 \in \mathbb{R}^{p \times p}$	$(C - 1)$ Ave $_c \{ \Sigma_c \}$
$G_3 \in \mathbb{R}^{p \times p}$	$\frac{1}{C} \sum_{c,c'} \Sigma_{c,c'}$

Table 1: Summary of notations.

