A Derivation of Scale-Invariant Properties

Lemma 1. *For a scale-invariant loss function $\mathcal{L}(\boldsymbol{x})$ and $\alpha > 0$, we have*

$$1. \quad \nabla\mathcal{L}(\alpha\boldsymbol{x}) = \frac{1}{\alpha}\nabla\mathcal{L}(\boldsymbol{x}) \tag{28}$$

$$2. \quad \langle \boldsymbol{x}, \nabla\mathcal{L}(\boldsymbol{x})\rangle = 0 \tag{29}$$

$$3. \quad \|\boldsymbol{\Sigma}(\boldsymbol{x})^{1/2}\boldsymbol{x}\| = \sqrt{\boldsymbol{x}^\top\boldsymbol{\Sigma}(\boldsymbol{x})\boldsymbol{x}} = 0. \tag{30}$$

Proof. Given $\boldsymbol{x} \in \mathbb{R}^d$, we have $\mathcal{L}(\alpha\boldsymbol{x}) = \mathcal{L}(\boldsymbol{x})$, $\forall\alpha > 0$. The first property can be proved by taking gradients with respect to $\boldsymbol{x}$ on both sides of $\mathcal{L}(\alpha\boldsymbol{x}) = \mathcal{L}(\boldsymbol{x})$. Taking derivatives with respect to α for the both sides of $\mathcal{L}(\alpha\boldsymbol{x}) = \mathcal{L}(\boldsymbol{x})$, we have $\nabla_\alpha\mathcal{L}(\alpha\boldsymbol{x}) = 0$. Left hand side equals $\nabla_{\alpha\boldsymbol{x}}\mathcal{L}(\alpha\boldsymbol{x})^\top\boldsymbol{x}$, so the second property follows by taking $\alpha = 1$. Since $\langle \boldsymbol{x}, \nabla\mathcal{L}(\boldsymbol{x})\rangle = 0$ and $\langle \boldsymbol{x}, \nabla\mathcal{L}(\boldsymbol{x}; \mathcal{B})\rangle = 0$ we have $\langle \boldsymbol{x}, \nabla\mathcal{L}(\boldsymbol{x}; \mathcal{B}) - \nabla\mathcal{L}(\boldsymbol{x})\rangle = 0$. Thus we get the third property $\boldsymbol{x}^\top\boldsymbol{\Sigma}(\boldsymbol{x})\boldsymbol{x} = \mathbb{E}\left[\langle \boldsymbol{x}, \nabla\mathcal{L}(\boldsymbol{x}; \mathcal{B}) - \nabla\mathcal{L}(\boldsymbol{x})\rangle^2\right] = 0$.

B Additional Experimental Results

Experimental setup. We presents results for a ResNet-18 [5] (or scale-invariant ResNet-18) trained on CIFAR-10 [13], a scale-invariant DenseNet-100 [11] trained on CIFAR-100 [13], and scale-invariant ResNet-18 trained on Tiny ImageNet [14]. For ResNet-18 on CIFAR-10 and DenseNet-100 on CIFAR-100, we train for 300 epochs regardless of the number of training samples and the learning rate is divided by 10 at epochs 150 and 225. For ResNet-18 on Tiny ImageNet, we train for 120 epochs and the learning rate is divided by 10 at epochs 60 and 90. For SGDM, we set base value as LR 0.1, WD 0.0001 and search them by multiplying the factor of 2 or $\sqrt{2}$ and we set momentum coefficient 0.9. For SGD, we set base value as LR 1, which makes the value $\frac{N\eta\lambda}{B(1-\mu)}$ the same as LR 0.1 for the SGDM setting. We apply data augmentation including padding, random crops and left-right flips. When using a scale-invariant network, all parameters are used to calculate the angular update per epoch. On the other hand, when using an unmodified network, only the parameters of the convolutional layer are used to calculate the angular update per epoch.

B.1 Scale-Invariant Network

Experiment on fixed B. Table 8 and Table 9 show the results on fixed batch sizes of DenseNet-100 trained on CIFAR-100 and ResNet-18 trained on Tiny ImageNet respectively. In Table 8, we colored yellow to cells which satisfy the tuning factor $\frac{N\eta\lambda}{B(1-\mu)} = \frac{5}{64}$ and in Table 9, we colored yellow to cells which satisfy the tuning factor $\frac{N\eta\lambda}{B(1-\mu)} = \frac{5}{32}$. Bold values represent the highest accuracy for each column.

Experiment on fixed N. Table 10 shows the results on fixed number of training samples of DenseNet-100 trained on CIFAR-100 and we colored yellow to cells which satisfy the tuning factor $\frac{N\eta\lambda}{B(1-\mu)} = \frac{5}{64}$.

B.2 Unmodified Network

Table 11 shows the results on fixed number of training samples of unmodified ResNet-18 trained on CIFAR-10 and we colored yellow to cells which satisfy the tuning factor $\frac{N\eta\lambda}{B(1-\mu)} = \frac{5}{64}$. Figure 9 shows the angular update per epoch of the yellow cells of Table 11.

Table 8: Scale-invariant DenseNet-100 on CIFAR-100 with SGDM and B=128.

LR	WD	3125	6250	12500	25000	50000
6.4	0.0001	22.73				
3.2	0.0001	32.29	41.29	45.99	40.74	
1.6	0.0001	**32.31**	45.93	52.90	20.42	61.50
0.8	0.0001	29.73	**46.16**	**57.50**	65.53	71.95
0.4	0.0001	27.32	44.36	57.17	**67.78**	73.62
0.2	0.0001	26.81	43.10	55.67	66.56	**74.03**
0.1	0.0001	27.84	41.62	55.51	65.56	73.73

(a) Tuning LR

LR	WD	3125	6250	12500	25000	50000
0.1	0.0064	29.44	28.18	22.52	15.17	
0.1	0.0032	**36.28**	43.08	49.96	48.63	42.68
0.1	0.0016	35.48	**47.43**	55.94	62.50	64.59
0.1	0.0008	32.95	47.14	**60.21**	67.03	72.24
0.1	0.0004	30.88	45.96	58.77	**68.27**	74.05
0.1	0.0002	28.87	44.47	57.01	67.52	**74.35**
0.1	0.0001	27.84	41.62	55.51	65.56	73.73

(b) Tuning WD

LR	WD	3125	6250	12500	25000	50000
0.8	0.0008	25.39	20.44			
0.5657	0.0005657	31.75	41.69	46.66	42.67	40.84
0.4	0.0004	**33.74**	46.53	54.86	59.94	71.98
0.2828	0.0002828	30.91	**46.89**	57.96	67.04	71.98
0.2	0.0002	29.32	45.92	**58.82**	67.47	73.58
0.1414	0.0001414	27.66	43.77	56.30	**67.60**	**74.13**
0.1	0.0001	27.84	41.62	55.51	65.56	73.73

(c) Tuning LR,WD

Table 9: Scale-invariant ResNet-18 on Tiny ImageNet with SGDM and B=256.

LR	WD	12500	25000	50000	100000
6.4	0.0001	30.61	36.90	40.50	38.42
3.2	0.0001	**36.25**	44.52	51.99	58.19
1.6	0.0001	35.17	**47.70**	55.77	62.20
0.8	0.0001	34.54	46.01	**56.67**	**64.38**
0.4	0.0001	34.68	44.36	55.47	63.67
0.2	0.0001	33.04	43.54	53.58	62.08

(a) Tuning LR

LR	WD	12500	25000	50000	100000
0.1	0.0064	32.31	39.98	41.51	39.05
0.1	0.0032	**37.04**	45.99	53.04	58.67
0.1	0.0016	35.56	**47.75**	56.07	62.20
0.1	0.0008	34.34	45.43	**56.89**	**64.38**
0.1	0.0004	32.69	43.78	54.59	63.67
0.1	0.0002	31.60	43.09	52.57	62.08

(b) Tuning WD

LR	WD	12500	25000	50000	100000
0.8	0.0008	32.96	38.80	41.58	39.90
0.5657	0.0005657	**37.02**	45.60	52.58	58.74
0.4	0.0004	35.88	**47.77**	56.06	62.75
0.2828	0.0002828	34.95	45.97	**57.06**	**64.19**
0.2	0.0002	34.15	44.50	55.06	64.18
0.1414	0.0001414	32.74	42.91	52.65	61.85

(c) Tuning LR,WD

Table 10: Scale-invariant DenseNet-100 on CIFAR-100 with SGDM and 50k data samples.

LR	WD	2048	1024	512	256	128
6.4	0.0001	71.27	66.43			
3.2	0.0001	72.65	72.78	69.51		
1.6	0.0001	**72.88**	**73.76**	73.17	70.55	61.50
0.8	0.0001	72.35	73.35	73.20	73.57	71.95
0.4	0.0001	71.37	72.66	**73.90**	**73.78**	73.62
0.2	0.0001	70.45	71.93	72.42	73.57	**74.03**
0.1	0.0001	69.01	70.31	71.71	72.55	73.73

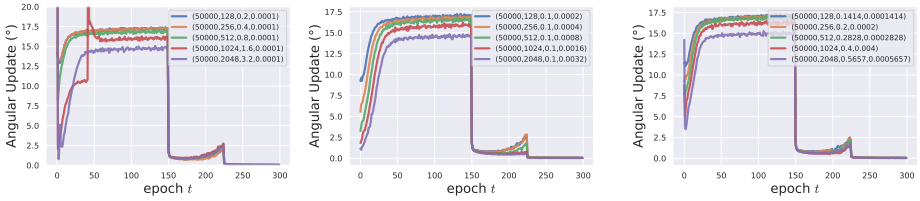
(a) Tuning LR

LR	WD	2048	1024	512	256	128
0.1	0.0064	73.84	70.91	61.19	41.16	
0.1	0.0032	**73.89**	73.54	72.33	62.46	42.68
0.1	0.0016	73.29	**74.11**	73.65	72.50	64.59
0.1	0.0008	72.12	73.73	**74.31**	74.12	72.24
0.1	0.0004	70.67	72.41	73.63	**74.16**	74.05
0.1	0.0002	69.50	70.80	72.69	73.40	**74.35**
0.1	0.0001	69.01	70.31	71.71	72.55	73.73

(b) Tuning WD

LR	WD	2048	1024	512	256	128
0.8	0.0008	72.95	69.36	57.74		
0.5657	0.0005657	**73.69**	73.49	71.27	59.22	40.84
0.4	0.0004	73.26	**74.37**	73.56	71.67	71.98
0.2828	0.0002828	72.00	73.68	**73.93**	73.37	71.98
0.2	0.0002	70.87	72.21	73.67	**74.22**	73.58
0.1414	0.0001414	70.00	71.33	72.81	73.67	**74.13**
0.1	0.0001	69.01	70.31	71.71	72.55	73.73

(c) Tuning LR,WD



(a) (b) (c)

Fig. 9: Angular update per epoch of unmodified ResNet-18 on CIFAR-10 with the number of data 50k (a) tuning LR, (b) tuning WD, (c) tuning both the LR and WD.

Table 11: Unmodified ResNet-18 on CIFAR-10 with SGDM and 50k data samples.

LR	WD	2048	1024	512	256	128
6.4	0.0001	93.98	92.94	89.97		
3.2	0.0001	94.41	**94.51**	93.86	91.47	63.49
1.6	0.0001	**94.42**	94.17	**94.80**	94.05	92.23
0.8	0.0001	93.71	94.50	94.72	94.88	94.62
0.4	0.0001	93.55	94.18	94.75	**94.89**	95.16
0.2	0.0001	92.57	93.56	94.32	94.80	**95.25**
0.1	0.0001	71.15	81.52	87.52	92.16	94.86

(a) Tuning LR

LR	WD	2048	1024	512	256	128
0.1	0.0064	**94.82**	94.09	90.76	86.30	78.39
0.1	0.0032	94.76	**95.13**	94.51	92.42	89.11
0.1	0.0016	94.53	95.06	95.10	94.73	93.45
0.1	0.0008	93.91	94.60	**95.14**	**95.27**	95.08
0.1	0.0004	92.93	94.00	94.63	95.22	95.16
0.1	0.0002	91.80	93.34	94.11	94.73	**95.25**
0.1	0.0001	71.15	81.52	87.52	92.16	94.86

(b) Tuning WD

LR	WD	2048	1024	512	256	128
0.8	0.0008	94.44	93.40	91.33	81.50	74.19
0.5657	0.0005657	**94.71**	94.80	94.11	92.37	89.28
0.4	0.0004	94.57	**94.91**	94.91	94.69	93.66
0.2828	0.0002828	93.89	94.71	**94.97**	95.10	94.87
0.2	0.0002	93.19	94.20	94.71	**95.28**	**95.17**
0.1414	0.0001414	92.15	93.34	94.24	94.76	94.99
0.1	0.0001	71.15	81.52	87.52	92.16	94.86

(c) Tuning LR,WD