How Does Loss Function Affect Generalization Performance of Deep Learning? Application to Human Age Estimation

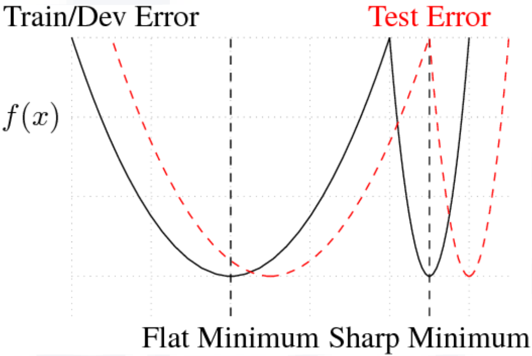
Ali Akbari, Muhammad Awais, Manijeh Bashar and Josef Kittler

Research Fellow
Centre for Vision, Speech and Signal Processing (CVSSP)
University of Surrey
Guildford, UK

ICML: International Conference on Machine Learning
July 2021

Outline

A Conceptual Sketch of Flat and Sharp Minima.



Outline

1. **Generalisation Analysis**
2. Numerical Results
3. Conclusion

Generalisation Error

Given a training set $\mathcal{S}$, the generalisation error of the output model $f_{\mathcal{S}}^{\theta}$, trained using the learning algorithm $\mathcal{A}$ on $\mathcal{S}$, is the difference between the empirical and true risk:

$$E = R_{\text{true}}(f_{\mathcal{S}}^{\theta}) - R_{\text{emp}}(f_{\mathcal{S}}^{\theta})$$

Generalisation Error and Uniform Stability

We use the notion of uniform stability to uncover the link between the generalisation error of SGD and loss function.

Generalisation Error Bound

Consider a loss function ℓ such that $0 \leq \ell(f(\cdot;\mathbf{z}) \leq L$ for any point $\mathbf{z}$. Suppose that SGD update rule is executed for T iterations with an annealing learning rate λ_t. Then, we have the following generalisation error bound with probability at least $1-\delta$:

$$E(f_{\mathcal{S}}) = R_{\text{true}}(f_{\mathcal{S}}) - R_{\text{emp}}(f_{\mathcal{S}}) \leq$$
$$2\gamma^2 \sum_{t=1}^{T} \lambda_t \left(2\sqrt{\frac{\log(2/\delta)}{T}} + \sqrt{\frac{2\log(2/\delta)}{N}} + \frac{1}{N} \right) + L\sqrt{\frac{\log(2/\delta)}{2N}}$$

What factors make generalisation error bound tighter?

- Number of training samples N
- Number of SGD iteration T
- Lipschitz constant γ

Lipschitz Loss Function

A loss function $\ell(\hat{\mathbf{y}}, \mathbf{y})$ is γ-Lipschitz with respect to the output vector $\hat{\mathbf{y}}$, if for $\gamma \geq 0$ and $\forall \mathbf{u}, \mathbf{v} \in \mathbb{R}^K$ we have

$$|\ell(\mathbf{u}, \mathbf{y}) - \ell(\mathbf{v}, \mathbf{y})| \leq \gamma \|\mathbf{u} - \mathbf{v}\|.$$

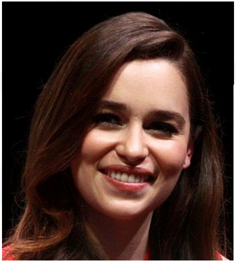
We use $\|\cdot\|$ to denote the ℓ_2-norm of vectors.

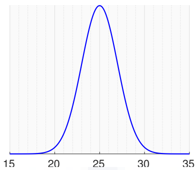
Intuitively, γ is related to how fast ℓ is allowed to change.

Learning Problem: Age Estimation

Semantic Similarity

Characterising the semantic similarity among classes.





- Due to similarity between neighbouring classes, the label is a Gaussian distribution for a facial image at the age of 25.

Loss Functions

Existing Loss Function

Kullback-Leibler divergence (KL)

$$L(\mathbf{p}, \mathbf{q}) = \sum_{k=1}^{L} q_k \log(\frac{q_k}{p_k})$$

Jensen-Shannon divergence (JS)

$$L = \frac{1}{2} \sum_{k=1}^{L} q_k \log\left(\frac{q_k}{\frac{p_k+q_k}{2}}\right) + p_k \log\left(\frac{p_k}{\frac{p_k+q_k}{2}}\right)$$

Distribution Cognisant Loss (GJM)

$$L = \sum_{k=1}^{L} |q_k^{\alpha} - p_k^{\alpha}|^{\frac{1}{\alpha}} = \sum_{k=1}^{L} q^k \left|1 - (\frac{p_k}{q_k})^{\alpha}\right|^{\frac{1}{\alpha}} \quad 0 \leq \alpha \leq 1$$

Theoretical Results

Our Main Result

Given that the GJM, JS and KL loss functions are γ_{GJM}-Lipschitz, γ_{JS}-Lipschitz and γ_{KL}-Lipschitz, respectively, the following inequality holds:

$$\gamma_{GJM} \leq \gamma_{JS} \leq \gamma_{KL}$$

and then we have:

$$E(f_{\mathcal{S}})_{GJM} \leq E(f_{\mathcal{S}})_{JS} \leq E(f_{\mathcal{S}})_{KL}.$$

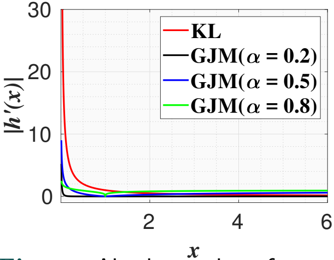


Figure: Absolute value of derivative of loss functions at different points x.

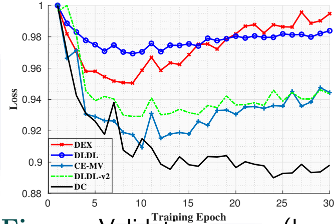


Figure: Validation curves (loss scores vs. epoch).

Outline

Cross-database Evaluation (MAE & CS) on the Target Databases

	FG-NET		MORPH		FACES		SC-ROT		SC-SUR	
Method	**MAE**	**CS (%)**	**MAE**	**CS (%)**	**MAE**	**CS (%)**	**MAE**	**CS (%)**	**MAE**	**CS (%)**
Human Workers	4.70	69.5	6.30	51.0	NA	NA	NA	NA	NA	NA
Microsoft API	6.20	53.80	6.59	46.00	-	-	-	-	-	-
CE	3.20	82.14	5.50	60.34	5.33	61.60	6.07	53.59	5.44	66.76
Ranking	3.12	83.80	5.28	62.55	4.83	65.74	5.29	63.92	5.41	64.90
KL	3.08	83.83	5.27	62.43	4.72	66.76	5.25	63.93	5.46	65.71
JS	2.99	83.53	4.81	65.83	4.68	66.52	4.54	69.23	4.98	67.59
GJM	**2.93**	**84.43**	**4.63**	**66.03**	**4.47**	**69.88**	4.72	**71.19**	**4.78**	**71.75**

Outline

1. Generalisation Analysis
2. Numerical Results
3. **Conclusion**

Our main statement in this paper is:

1. Choose a Lipschitz loss function, get model with higher generalisation.

Thank You!