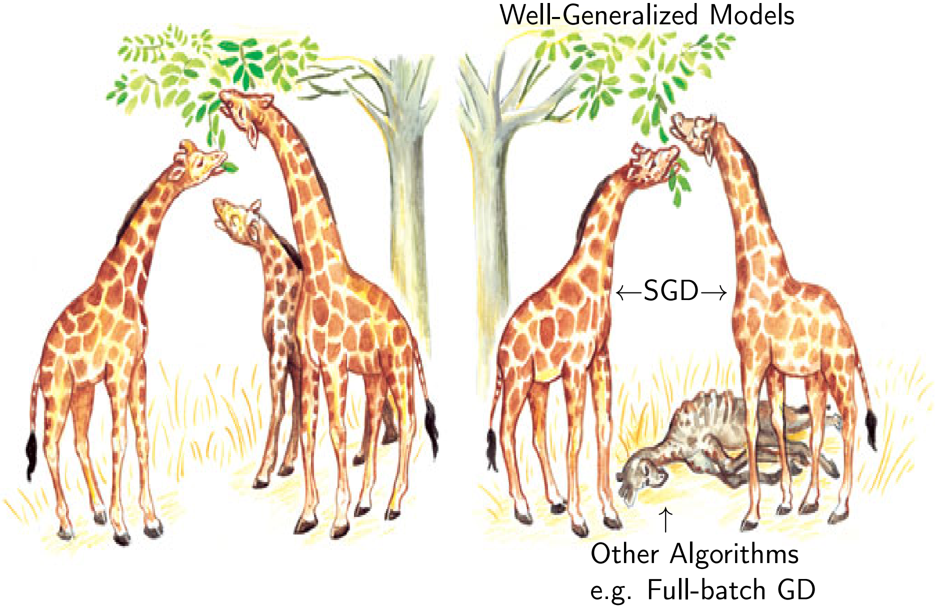
Implicit Jacobian Regularization

Weighted with Impurity of Probability Output

Sungyoon Lee[1] Jinseong Park[2] Jaewook Lee[2]

[1]Dept. of Computer Science, Hanyang University
[2]Dept. of Industrial Engineering, Seoul National University


Well-Generalized Models
←SGD→
↑
Other Algorithms
e.g. Full-batch GD

- Natural Selection = Selection by Practitioners
- Leaves = Well-Generalized Models (goal)
- Giraffes = Learning Algorithms
- Giraffes w/ long necks = SGD (and its variants)
- Long neck = ?

Q. What are the advantageous features of SGD to find well-generalized models?

•	Natural Selection	= Selection by Practitioners
•	Leaves	= Well-Generalized Models (goal)
•	Giraffes	= Learning Algorithms
•	Giraffes w/ long necks	= SGD (and its variants)
•	Long neck	= Implicit Regularization in SGD!

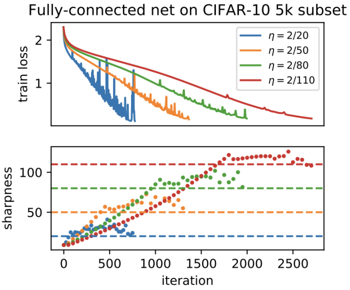
Q. What are the advantageous features of SGD to find well-generalized models?
A. Implicit Regularization in SGD!

Implicit Jacobian Regularization

Weighted with Impurity of Probability Output

At the Edge of Stability [CKL+21],

$$\frac{2}{\eta} \approx \|H\|$$



At the Edge of Stability [CKL+21],

$$\frac{2}{\eta} \approx \|H\| \approx \|G\|$$

$$\uparrow$$

Gauss-Newton Approximation

$$H := \langle \nabla_\theta^2 \ell \rangle$$

$$G := \langle \nabla_\theta z \nabla_z^2 \ell \nabla_\theta z^\top \rangle = \langle JMJ^\top \rangle$$

$$J := \nabla_\theta z$$

$$M := \nabla_z^2 \ell$$

where $\langle \cdot \rangle = \mathbb{E}_{\mathcal{D}}[\cdot]$

At the Edge of Stability [CKL+21],

$$J := \nabla_\theta z$$

$$\downarrow$$

$$\frac{2}{\eta} \approx \|H\| \approx \|G\| \, {\color{red}=} \, \langle \lambda^* \|J\|^2 \rangle$$

$$\uparrow$$

Our Main Theorem

At the Edge of Stability [CKL+21],

$$\frac{2}{\eta} \approx \|H\| \approx \|G\| = \langle \lambda^* \|J\|^2 \rangle$$

- λ^* controls the effectiveness of IJR (high $\lambda^* \Rightarrow$ low $\|J\|^2$).

$\uparrow$

Implicit **J**acobian **R**egularization

At the Edge of Stability [CKL+21],

$$\frac{2}{\eta} \approx \|H\| \approx \|G\| = \langle \lambda^* \|J\|^2 \rangle$$

- λ^* controls the effectiveness of IJR (high $\lambda^* \Rightarrow$ low $\|J\|^2$).
- λ^* is bounded above by the norm $\|M\|$.

$$\uparrow$$
$$M := \nabla_z^2 \ell$$

At the Edge of Stability [CKL+21],

$$\frac{2}{\eta} \approx \|H\| \approx \|G\| = \langle \lambda^* \|J\|^2 \rangle$$

- λ^* controls the effectiveness of IJR (high $\lambda^* \Rightarrow$ low $\|J\|^2$).
- λ^* is bounded above by the norm $\|M\|$.
- The lower the norm $\|M\| \downarrow$, the weaker the regularization effect $\downarrow$.
- The norm $\|M\|$ acts as an adaptive regularization weight.

At the Edge of Stability [CKL+21],

$$\frac{2}{\eta} \approx \|H\| \approx \|G\| = \langle \lambda^* \|J\|^2 \rangle$$

- λ^* controls the effectiveness of IJR (high $\lambda^* \Rightarrow$ low $\|J\|^2$).
- λ^* is bounded above by the norm $\|M\|$.
- The lower the norm $\|M\| \downarrow$, the weaker the regularization effect $\downarrow$.
- The norm $\|M\|$ acts as an adaptive regularization weight.
- How does $\|M\|$ evolve during training?

Implicit Jacobian Regularization

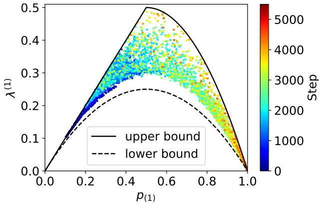
Weighted with Impurity of Probability Output

$$M := \nabla_z^2 \ell = \mathrm{diag}(p) - pp^\top$$

$p_{(1)}$ is the proability of the most probable class.

Theorem ($\|M\|$ as Impurity of Probability Output)

$$\underbrace{\frac{1}{2} Gini(p_{(1)})}_{\textit{lower bound}} \le \|M\| \le \overbrace{\min(p_{(1)}, Gini(p_{(1)}))}^{\textit{upper bound}}$$



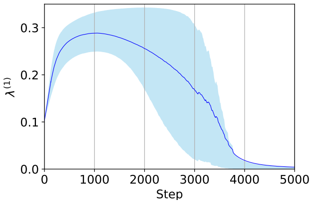


Figure: Inverted U-shaped curve of evolution of $\|M\|$

$$M := \nabla_z^2 \ell = \mathrm{diag}(p) - pp^\top$$

$p_{(1)}$ is the proability of the most probable class.

Theorem ($\|M\|$ as Impurity of Probability Output)

$$\underbrace{\frac{1}{2} Gini(p_{(1)})}_{\textit{lower bound}} \leq \|M\| \leq \overbrace{\min(p_{(1)}, Gini(p_{(1)}))}^{\textit{upper bound}}$$

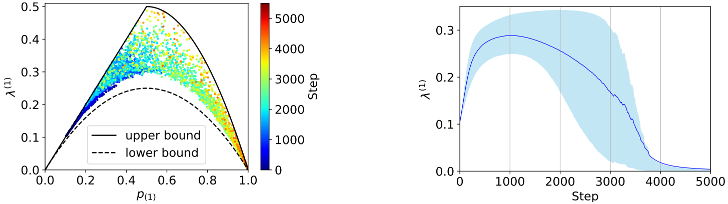


Figure: Inverted U-shaped curve of evolution of $\|M\|$

Active Regularization Period (ARP)

I beginning → not at the Edge of Stability yet.

II **early (ARP)** → **high** impurity → strong regularization (IJR)

III later → low impurity → weak regularization (IJR)

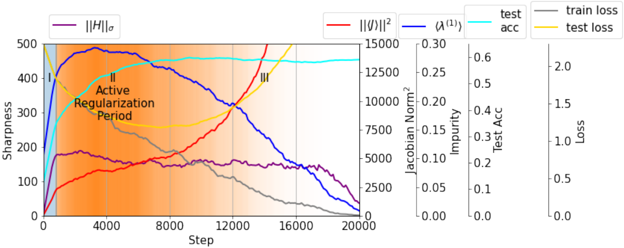


Figure: Dark orange color indicates a high impurity $\langle \lambda^{(1)} \rangle$.

More in the Paper ...

- Explicit Jacobian Regularization (two-step update like SAM [FKM+21])

The End

sungyoonlee@hanyang.ac.kr

References

[CKL+21] Jeremy Cohen et al. "Gradient Descent on Neural Networks Typically Occurs at the Edge of Stability". In: *International Conference on Learning Representations*. 2021. URL: https://openreview.net/forum?id=jh-rTtvkGeM.

[FKM+21] Pierre Foret et al. "Sharpness-aware Minimization for Efficiently Improving Generalization". In: *International Conference on Learning Representations*. 2021. URL: https://openreview.net/forum?id=6Tm1mposlrM.