Understanding the Generalization Ability of Deep Learning Algorithms: A Kernelized Rényi's Entropy Perspective

Yuxin Dong, Tieliang Gong, Hong Chen, Chen Li

Xi'an Jiaotong University

MLA 2023

Generalization Analysis

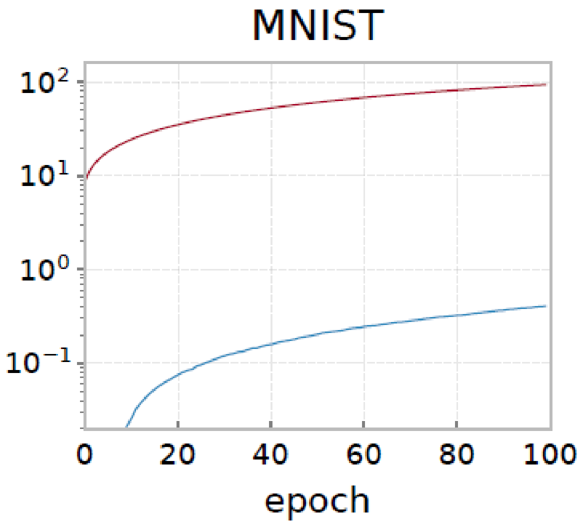
Method	Uniform Stability	Information Theory
Representative Work	[2016] Train faster, generalize better: Stability of stochastic gradient descent	[2017] Information-theoretic analysis of generalization capability of learning algorithms
Assumptions	Strong Assumptions Lipschitz Condition Smoothness (Strong) Convexity	Weak Assumptions Sub-gaussian
Computability	Not computable	Computable

Existing Problems

Information-theoretic generalization bound for SGLD [1]:

$$
\begin{aligned}
|\mathrm{gen}(W,S)| &\triangleq \left|\mathbb{E}_{W,S}[L(W)-L_S(W)]\right| && \text{Generalization Error} \\
&\le \sqrt{\frac{2R^2 I(W;S)}{n}} && \text{Not computable} \\
&\le \sqrt{\frac{2R^2 \sum_{t=1}^{T} I(W_t;B_t|W_{t-1})}{n}} && \text{Not computable} \\
&\le \sqrt{\frac{R^2 \sum_{t=1}^{T} \frac{\eta_t^2 V_t}{\sigma_t^2}}{n}} && \text{Computable Upper Bound}
\end{aligned}
$$

10^2x ~ 10^3x Gap



[1] Hao Wang, et al. Analyzing the generalization capability of SGLD using properties of gaussian channels. NeurIPS, 2021.

Kernelized Rényi's Entropy

Shannon's entropy is not computable for high-dimensional random variables:

Shannon's Entropy

$$H(X) = -\int_{\mathcal{X}} p(x) \log p(x)\, dx$$

Recently, matrix-based Rényi's entropy enables direct estimation of entropy from data [2].

Take the infinite-sample case, we now have kernelized Rényi's entropy of order $\alpha \to 1$:

Kernelized Rényi's Entropy

$$S_1(X) = -C_\phi \iint_{\mathcal{X}^2} p(x) \log p(x')\, \phi^2(x, x')\, dx\, dx'$$

[2] Giraldo et al. Measures of entropy from data using infinitely divisible kernels. IEEE Transactions on Information Theory. 2014.

Kernelized Rényi's Entropy

Essential properties of Shannon's entropy are inherited by our definition:

- $I_1(X;Y) = D_1\left(P_{X,Y} \parallel P_X \otimes P_Y\right) \geq 0.$
- $I_1(X;Y) = S_1(X) + S_1(Y) - S_1(X,Y).$
- $I_1(X;Y|Z) = I_1(X;Y,Z) - I_1(X;Z) = I_1(X,Z;Y) - I_1(Z;Y).$
- Given Markov chain relationship $X \rightarrow Y \rightarrow Z$, we have $\min\left(I_1(X;Y), I_1(Y;Z)\right) \geq I_1(X;Z).$

Theorem 1

Let $\{x_i\}_{i=1}^m$ be i.i.d data points sampled from X, and let $K \in \mathbb{R}^{m\times m}$ be the kernel matrix constructed by $K_{ij} = \frac{1}{m}\phi(x_i, x_j)$. Then with confidence $1-\delta$,

$$\left|S_1(X) - \hat{S}_1(X)\right| \leq \frac{9C_\phi\sqrt{2\log\frac{2}{\delta}}}{\sqrt[3]{m}},$$

Where $\hat{S}_1(X) = -C_\phi\, \mathrm{tr}(K\log K)$.

Our Contributions

Introduce kernelized Rényi's mutual information: $I \leftrightarrow I_1$

$$|\mathrm{gen}(W,S)| \triangleq \left|\mathbb{E}_{W,S}[L(W) - L_S(W)]\right|$$ Generalization Error

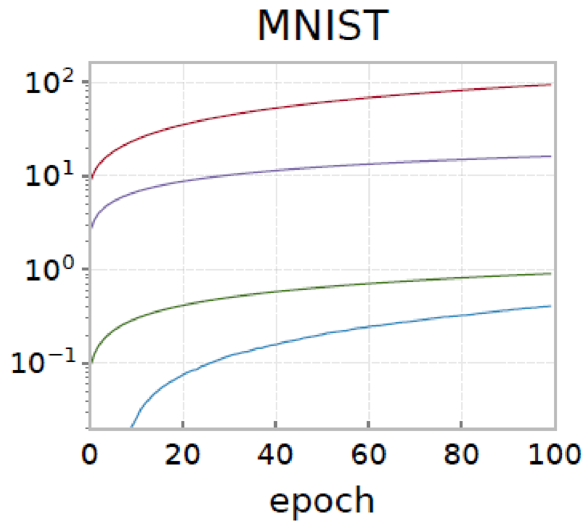
$$\leq \sqrt{\frac{2R^2 I_1(W;S)}{n}}$$ Computable

$$\leq \sqrt{\frac{2R^2 \sum_{t=1}^{T} I_1(W_t; B_t|W_{t-1})}{n}}$$ Computable

$$\leq \sqrt{\frac{R^2 \sum_{t=1}^{T} \log\left|\frac{\eta_t^2}{\sigma_t^2}\mathbb{V}_t + I\right|}{n}}$$ Ours Upper Bound

5x ~ 10x Improvement

$$\leq \sqrt{\frac{R^2 \sum_{t=1}^{T} \frac{\eta_t^2 V_t}{\sigma_t^2}}{n}}$$ Previous Upper Bound



Thank You !

For questions, please write to dongyuxin@stu.xjtu.edu.cn.

MLA 2023