Analyzing Generalization of Neural Networks through Loss Path Kernels

Yilan Chen[1], Wei Huang[2], Hao Wang[3], Charlotte Loh[3], Akash Srivastava[3], Lam M. Nguyen[4], and Tsui-Wei Weng[1]

[1]UCSD, [2]RIEKN AIP, [3]MIT-IBM Watson AI Lab, [4]IBM Research

paper twitter

Kernel machine and generalization theory of neural networks (NNs)

Kernel: $K(x,x') = \langle \Phi(x), \Phi(x') \rangle$, $\Phi: \mathcal{X} \to \mathcal{H}$ maps the data to a feature space.

Kernel machine (KM): linear function in the feature space:

$$g(x) = \langle \beta, \Phi(x) \rangle + b = \sum_{i=1}^{n} a_i K(x, x_i) + b, \quad \text{where } \beta = \sum_{i=1}^{n} a_i \Phi(x_i)$$

Neural Tangent Kernel (NTK) (Jacot et al., 2018):

$$\hat{\Theta}(w;\, x, x') = \langle \nabla_w f(w,x), \nabla_w f(w,x') \rangle$$

Infinite-width NN trained by gradient descent with square loss ⇔ kernel regression with NTK [Jacot et al., 2018; Arora et al., 2019]

Infinite-width NN trained with ℓ_2 regularized loss ⇔ ℓ_2 regularized KMs with NTK, e.g. SVM [Chen et al., 2021]

Generalization gap:

$$GAP = \mathbb{E}_{z\sim\mu}[\ell(w,z)] - \frac{1}{n}\sum_{i=1}^{n} \ell(w, z_i) \le ?$$

- VC dimension
- Norm-based bounds
- NTK-based bounds for ultra-wide NNs

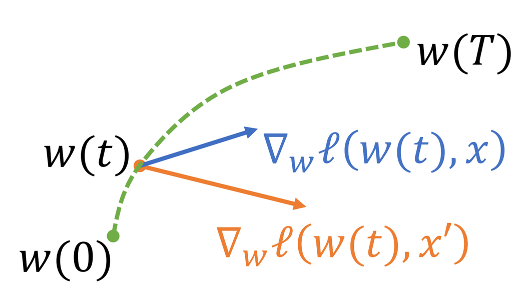
Motivations:

1. Can we establish a connection or equivalence between general NNs (vs ultra-wide NNs) and Kernel machines (KMs)?
2. Can we establish tight (vs vacuous) generalization bounds for general NNs (vs ultra-wide NNs)?

Contribution 1: Equivalence between NN and KM

Loss Path Kernel (LPK):

$$\mathrm{K}_T(z, z'; S) = \int_0^T \langle \nabla_w \ell(w,x), \nabla_w \ell(w,x') \rangle \, dt$$



With gradient flow (gradient descent with infinitesimal step size):

$$\frac{w(t+1) - w(t)}{\eta} = -\nabla_w L_S(w(t)) \xrightarrow{\eta \to 0} \frac{dw(t)}{dt} = -\nabla_w L_S(w(t))$$

$$\ell(w_T, z) = \sum_{i=1}^{n} -\frac{1}{n} \mathrm{K}_T(z, z_i; S) + \ell(w_0, z)$$

Loss function at time T — Kernel machine with **LPK** — Loss function at initialization

Stochastic gradient flow:

$$\ell(w_T, z) = \sum_{t=1}^{T-1} \sum_{i \in S_t} -\frac{1}{m} \mathrm{K}_T(z, z_i; S) + \ell(w_0, z)$$

Sum of KMs with **LPK**

Contribution 2: Generalization bound for NN trained by gradient flow

Different training set induces distinct LPK. Set of LPKs with constrained RKHS norm: $\mathcal{K}_T = \{\mathrm{K}_T(\cdot,\cdot; S'): S' \in \mathrm{supp}(\mu^{\otimes n}), \frac{1}{n^2}\sum_{i,j} \mathrm{K}_T(z_i', z_j'; S') \le B^2\}$

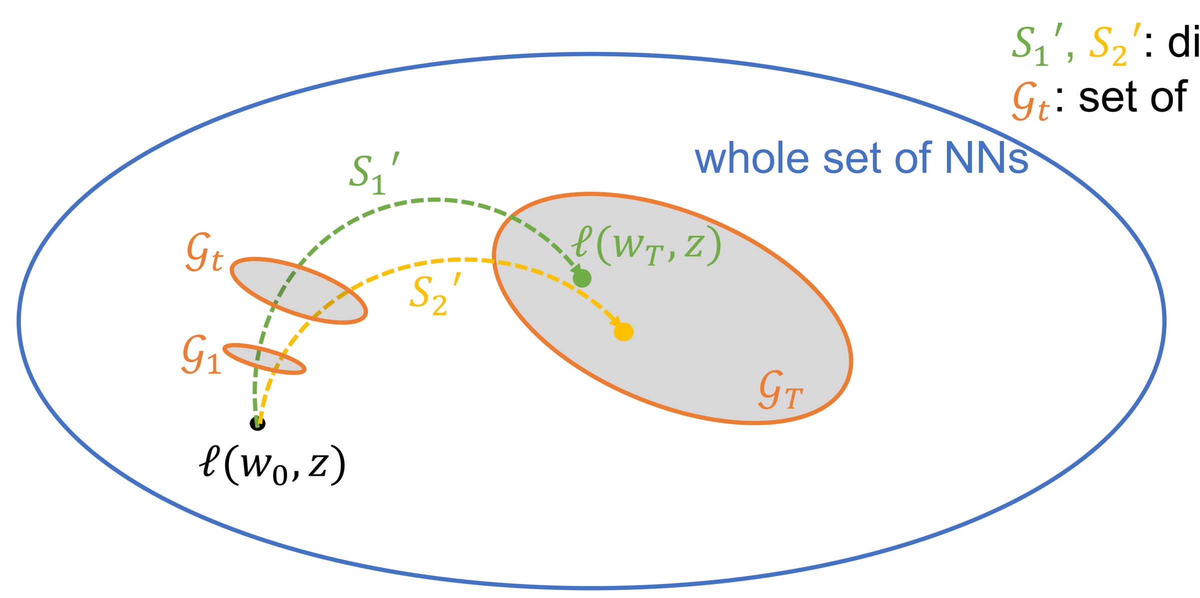
Set of NNs trained to time T:

$$\mathcal{G}_T = \left\{ g(z) = \sum_{i=1}^{n} -\frac{1}{n} \mathrm{K}(z, z_i'; S') + \ell(w_0, z) : \mathrm{K}(\cdot,\cdot; S') \in \mathcal{K}_T \right\}$$

$\ell(w_T, z)$ trained from S'

S_1', S_2': different set of samples

$\mathcal{G}_t$: set of NNs at time t



$$GAP \le 2\, min(U_1, U_2)$$

$$\to U_1 = \frac{B}{n} \sqrt{\sup_{\mathrm{K}\in\mathcal{K}_T} \sum_{i=1}^{n} \mathrm{K}(z_i, z_i; S') + \sum_{i \ne j} \Delta(z_i, z_j)} \qquad \Delta(z_i, z_j) = \frac{1}{2}[\sup_{\mathrm{K}\in\mathcal{K}_T} \mathrm{K}(z_i, z_j; S') - \inf_{\mathrm{K}\in\mathcal{K}_T} \mathrm{K}(z_i, z_j; S')]$$

maximum magnitude of the loss gradient in $\mathcal{K}_T$ evaluated with S throughout the training trajectory.

range of variation of LPK in $\mathcal{K}_T$

Compare with the bound of KM with a fixed kernel K: $GAP \le \frac{B}{n}\sqrt{\sum_{i=1}^{n} K(x_i, x_i)}$.

When $|\mathcal{K}_T| = 1$, our bound recovers KM's bound.

$$\to U_2 = \inf_{\epsilon > 0} \left(\frac{\epsilon}{n} + \sqrt{\frac{2 \ln \mathcal{N}(\mathcal{G}_T^S, \epsilon, \|\,\|_1)}{n}} \right)$$

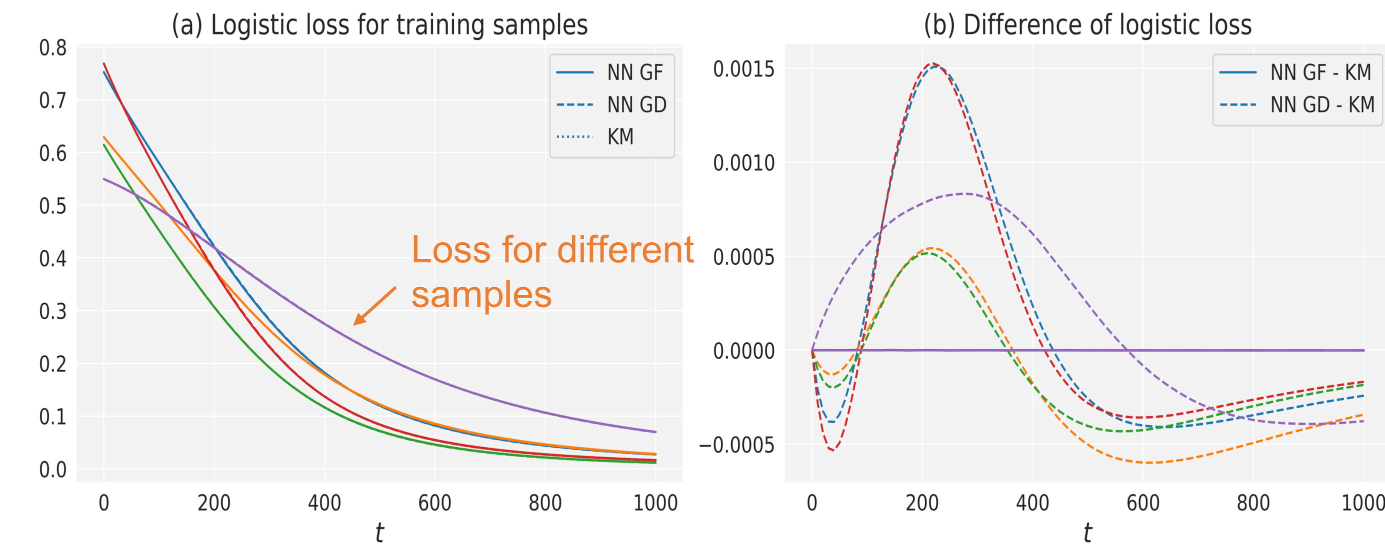
If the variation of the loss dynamics of gradient flow with different training data is small, U_2 will be small.

Compare with previous NTK-based bounds: Much more general results!

	Arora et al.	Cao & Gu	**Ours**
Bound	$\sqrt{\frac{2\mathbf{Y}^\top(\mathbf{H}^\infty)^{-1}\mathbf{Y}}{n}}$	$\tilde{O}(L \cdot \sqrt{\frac{\mathbf{Y}^\top(\mathbf{\Theta})^{-1}\mathbf{Y}}{n}})$	Theorem 3, Theorem 5
Model	Ultra-wide two-layer FCNN	Ultra-wide FCNN	**General continuously differentiable NN**
Data	i.i.d. data with $\|\boldsymbol{x}\| = 1$	i.i.d. data with $\|\boldsymbol{x}\| = 1$	i.i.d. data
Loss	Square loss	Logistic loss	**Continuously differentiable & bounded loss**
During training	No	No	**Yes**
Multi-outputs	No	No	**Yes**
Training algorithm	GD	SGD	(Stochastic) gradient flow

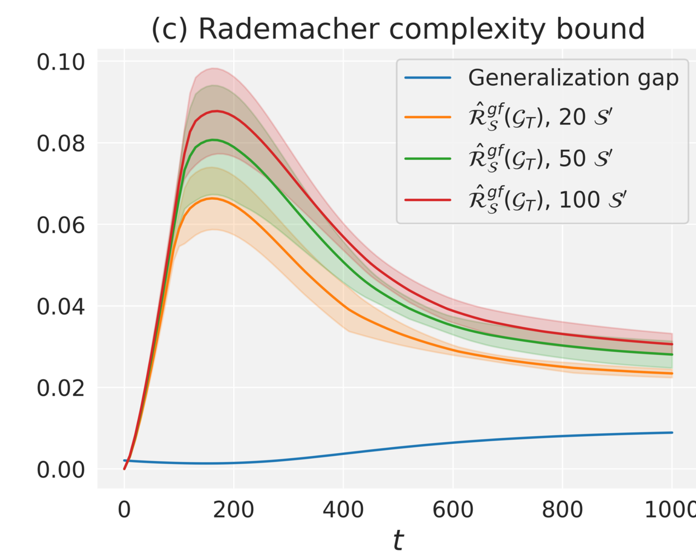
Experiments

a. Verify the equivalence:



- NN trained by gradient flow (GF) overlaps with the KM
- NN trained by gradient descent (GD) is also close with the KM

b. Generalization bound:



Our bound: ~0.03

VC dimension bound: 55957.3

Norm-based bound: 140.7

NTK-based bound (ultra-wide NN): 1.44

Tight bound!

Contribution 3: Neural architecture search

Use the bound to estimate the test loss and design minimum-training NAS algorithms: $\mathrm{Gene}(w, S) = L_S(w) + 2U_{sgd}$

U_{sgd}: simplified from the bound of stochastic gradient flow

Algorithm	CIFAR-10 Accuracy	Best	CIFAR-100 Accuracy	Best
Baselines				
TENAS [13]	93.08±0.15	93.25	70.37±2.40	**73.16**
RS + LGA$_3$ [39]	93.64		69.77	
Ours				
RS + Gene$(\boldsymbol{w}, \mathcal{S})_1$	93.68±0.12	93.84	72.02±1.43	73.15
RS + Gene$(\boldsymbol{w}, \mathcal{S})_2$	**93.79**±0.18	**94.02**	**72.76**±0.33	73.15
Optimal	94.37		73.51	

References

1. Jacot et al. Neural tangent kernel: Convergence and generalization in neural networks. NeurIPS 2018.
2. Arora et al. On exact computation with an infinitely wide neural net. NeurIPS 2019.
3. Chen et al. On the equivalence between neural network and support vector machine. NeurIPS 2021.
4. Arora et al. Fine-grained analysis of optimization and generalization for overparameterized two-layer neural networks. ICML 2019.
5. Cao, Y. and Gu, Q. Generalization bounds of stochastic gradient descent for wide and deep neural networks. NeurIPS 2019.