Random Smoothing Regularization in Kernel Gradient Descent Learning

Wenjia Wang

The Hong Kong University of Science and Technology (Guangzhou)

April 7th, 2025

Ding, L., Hu, T., Jiang, J., Li, D., Wang, W., & Yao, Y. (2023) (Alphabetical order). Random Smoothing Regularization in Kernel Gradient Descent Learning. JMLR accepted

Data augmentation

- An effective regularization technique, contributing to the empirical success of deep learning models across various applications.
- Making the model more robust to small perturbations.

Data augmentation

- An effective regularization technique, contributing to the empirical success of deep learning models across various applications.
- Making the model more robust to small perturbations.

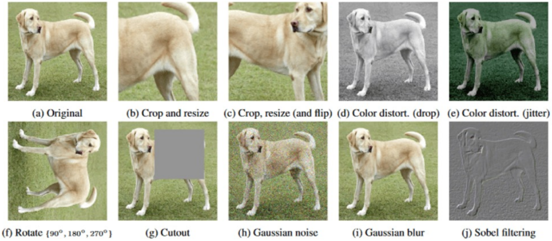

Figure: Data Augmentation.

Random smoothing data augmentation

- Random smoothing data augmentation involves adding random noise, such as Gaussian or Laplace noise, to the input data during the training process.
  - Address the adversarial vulnerability;
  - Applied in self-supervised contrastive learning methods;
- A simple example: adding Gaussian noise $\varepsilon_i \sim \mathcal{N}(0, \lambda \mathrm{I})$ to $\mathbf{x}_i$ to linear regression:

$$\min_{\mathbf{W}} \mathbb{E}_{\varepsilon} \frac{1}{n} \sum_{i=1}^{n} \left| \mathbf{W}^T(\mathbf{x}_i + \varepsilon_i) - y_i \right|^2 = \min_{\mathbf{W}} \frac{1}{n} \sum_{i=1}^{n} \left| \mathbf{W}^T \mathbf{x}_i - y_i \right|^2 + \lambda \mathbf{W}^T \mathbf{W}$$

  LSE $\rightarrow$ ridge regression, more robust

- In spite of the empirical success of random smoothing in various applications, there is a lack of systematic research on the regularization effect of random smoothing in the literature.

Our contributions

- We examine the classic *nonparametric regression* problem from the perspective of random smoothing regularization.
- We present a unified framework that can learn a wide range of D-dimensional ground truth functions belonging to the classical Sobolev spaces $(\mathcal{W}^{m_f})$ in an effective and adaptive manner.
- Optimal convergence rates can be achieved by utilizing random smoothing regularization and appropriate early stopping and/or weight decay techniques.

Our contributions

We investigate two possible function spaces that may contain the target function.

Sobolev space of low intrinsic dimensionality $d \leq D$

- Gaussian random smoothing: $n^{-m_f/(2m_f+d)}(\log n)^{D+1}$;
- Polynomial random smoothing with data size adaptive smoothing degree: $n^{-m_f/(2m_f+d)}(\log n)^{2m_f+1}$.

Mixed smooth Sobolev spaces

- Polynomial random smoothing of degree m_ε: $n^{-2m_f/(2m_f+1)}(\log n)^{\frac{2m_f}{2m_f+1}\left(D-1+\frac{1}{2(m_0+m_\varepsilon)}\right)}$

Problem Formulation

Suppose we have observed data $(\boldsymbol{x}_j, y_j)$ for $j = 1, \ldots, n$, which follows the relationship given by

$$y_j = f^*(\boldsymbol{x}_j) + \epsilon_j. \tag{1}$$

Here, $\boldsymbol{x}_j$'s are independent and identically distributed (i.i.d.) following a marginal distribution $P_{\mathbf{X}}$ with support $\mathrm{supp}(P_{\mathbf{X}}) = \Omega \subset \mathbb{R}^D$.

We employ reproducing kernel Hilbert spaces (can be viewed as two-layer infinite wide neural network).

(Example: $K = \mathbb{E}_{\phi \sim S}[\phi \otimes \phi]$ where ϕ is feature map and S spectral density.)

Reproducing kernel Hilbert space (RKHS)

Let $K : \Omega \times \Omega \to \mathbb{R}$ be a symmetric positive definite kernel function. Define the linear space

$$F_K(\Omega) = \left\{ \sum_{k=1}^{n} \beta_k K(\cdot, \boldsymbol{x}_k) : \beta_k \in \mathbb{R}, \boldsymbol{x}_k \in \Omega, n \in \mathbb{N} \right\},$$

and equip this space with the bilinear form

$$\left\langle \sum_{k=1}^{n} \beta_k K(\cdot, \boldsymbol{x}_k), \sum_{j=1}^{m} \gamma_j K(\cdot, \boldsymbol{x}_j') \right\rangle_K := \sum_{k=1}^{n} \sum_{j=1}^{m} \beta_k \gamma_j K(\boldsymbol{x}_k, \boldsymbol{x}_j').$$

- RKHS $\mathcal{H}_K(\Omega)$ generated by K: the closure of $F_K(\Omega)$ under the inner product $\langle \cdot, \cdot \rangle_K$;
- Inner product: $\langle \cdot, \cdot \rangle_{\mathcal{H}_K(\Omega)}$ is induced by $\langle \cdot, \cdot \rangle_K$;
- Norm: $\|f\|_{\mathcal{H}_K(\Omega)} = \sqrt{\langle f, f \rangle_{\mathcal{H}_K(\Omega)}}$;

- Consider loss function: $C[f] = \frac{1}{n}\sum_{j=1}^{n}(y_j - f)^2$.
- Representation theorem $f_t = K(\cdot, \boldsymbol{X})\boldsymbol{W}$:

$$\Theta_{t+1} = (1-\alpha_t)\Theta_t + \beta_t\left(\sqrt{\boldsymbol{K}}\boldsymbol{y} - \boldsymbol{K}\Theta_t\right)$$

where $\Theta_t = \sqrt{\boldsymbol{K}}\boldsymbol{W} \in \mathbb{R}^n$, $\boldsymbol{K} = [K(\boldsymbol{x}_i, \boldsymbol{x}_j)]_{i,j}$, and $K(\cdot, \boldsymbol{X}) = [K(\cdot, \boldsymbol{x}_1), \cdots, K(\cdot, \boldsymbol{x}_n)]$

- Asymptotically equivalent to Kernel Ridge regression

$$f_t = \operatorname*{arginf}_{h\in\mathcal{H}_K} \frac{1}{n}\sum_{j=1}^{n}(h(\boldsymbol{x}_j) - y_j)^2 + \lambda\|h\|_{\mathcal{H}_K}^2$$

where $\lambda = \lambda(t, \alpha_t, \beta_t)$ with $\lim_{t\to\infty}\lambda = 0$

- Early stop $t << n$

Random smoothing

- We construct N augmentations for each observed input point $\boldsymbol{x}_j$ by adding i.i.d. noise ε_{jk} with a continuous probability density function p_ε.
- Then take the average, i.e., the estimator is constructed as

$$f(\boldsymbol{x}) = \frac{1}{N}\sum_{k=1}^{N} h(\boldsymbol{x} + \varepsilon_k) \tag{2}$$

for $h \in \mathcal{H}_K(\Omega)$.

- Augmentations smooth the estimator. In the case of NTK, we have

$$K_s = \mathbb{E}_{f\sim S,\varepsilon,\varepsilon'\sim\hat{p}_{\varepsilon,N}}[f(\cdot+\varepsilon)\otimes f(\cdot+\varepsilon')]$$

Random Smoothing KGD with Early Stopping

- The kernel becomes

$$K_S(\boldsymbol{x}_l - \boldsymbol{x}_j) := \frac{1}{N^2}\sum_{k_1=1}^{N}\sum_{k_2=1}^{N} K(\boldsymbol{x}_l + \varepsilon_{k_1} - (\boldsymbol{x}_j + \varepsilon_{k_2})). \quad (3)$$

- The same update rule with a smooth path:

$$f_{t+1} = (1-\alpha_t)f_t + \beta_t \Phi_{K_S}(\langle y - f, \cdot\rangle_n)$$
$$\Theta_{t+1} = (1-\alpha_t)\Theta_t + \beta_t\left(\sqrt{\boldsymbol{K_S}}\boldsymbol{y} - \boldsymbol{K_S}\Theta_t\right)$$

- We are interested in the prediction error with early stop

$$\|f^* - f_t\|_{L_2(P_{\mathbf{X}})}, \quad (4)$$

Random smoothing noise

The elements of ε_k are i.i.d. mean zero sub-Gaussian random variables.

(C1) (Polynomial noise) There exists $m_\varepsilon > D/2$ such that the characteristic function of ε_k satisfies

$$c_1(1+\sigma_n^2\|\boldsymbol{\omega}\|_2^2)^{-m_\varepsilon} \le \mathbb{E}(e^{i\boldsymbol{\omega}^T\varepsilon_k}) \le c_2(1+\sigma_n^2\|\boldsymbol{\omega}\|_2^2)^{-m_\varepsilon}, \forall \boldsymbol{\omega} \in \mathbb{R}^D.$$

(C2) (Tensor Polynomial noise) There exists $m_\varepsilon > 1/2$ such that the characteristic function of ε_k satisfies

$$c_1 \prod_{j=1}^D (1+\sigma_n^2\omega_j^2)^{-m_\varepsilon} \le \mathbb{E}(e^{i\boldsymbol{\omega}^T\varepsilon_k}) \le c_2 \prod_{j=1}^D (1+\sigma_n^2\omega_j^2)^{-m_\varepsilon}, \forall \boldsymbol{\omega} \in \mathbb{R}^D.$$

(C3) (Gaussian noise) The elements of ε_k are normally distributed with variance σ_n^2.

Here the constants c_1 and c_2 do not depend on σ_n and m_ε. We call σ_n the smoothing scale in this work.

Sobolev space of low intrinsic dimensionality

Low intrinsic dimension

There exist positive constants c_1 and $d \leq D$ such that for all $\delta \in (0, 1)$, we have

$$\mathcal{N}_{\ell_\infty^D}(\delta, \Omega) \leq c_1 \delta^{-d},$$

where ℓ_∞^D is the $\mathbb{R}^D$ space equipped with ℓ_∞ norm.

- $\Omega \subset \mathbb{R}^D$ is bounded and has a positive Lebesgue measure, then $d = D$;
- Ω is a bounded D'-dimensional differentiable manifold, then $d = D'$.

Polynomial smoothing

- We have

$$\|f_t - f^*\|^2_{L_2(P_{\mathbf{x}})} = O_{\mathbb{P}}\left(n^{-\frac{2m_f}{2m_f+d}} (\log n)^{2m_f+1}\right).$$

  for $N > N_0$, where N is the number of augmentations.
- The above statements hold for both cases where early stopping is
  - without weight decay (iteration number $t \asymp n^{\frac{2(m_0+m_\varepsilon)}{2m_f+d}} \sigma_n^{2m_\varepsilon}$);
  - with weight decay (weight decay rate $\alpha \asymp n^{-1-\frac{2(m_0+m_\varepsilon)}{2m_f+d}} \sigma_n^{-2m_\varepsilon}$, and iteration number $t \geq C_2(\frac{m_f}{2m_f+d} + 1/2) \log n/(\log(1-\alpha))$).

Gaussian smoothing

- We have

$$\|f_t - f^*\|_{L_2(P_{\mathbf{x}})}^2 = O_{\mathbb{P}}\left(n^{-\frac{2m_f}{2m_f+d}}(\log n)^{D+1}\right).$$

  for $N > N_0$, where N is the number of augmentations.
  - Stepsize: $\beta = n^{-1}C_1$ with $C_1 \le (2\sup_{\boldsymbol{x}\in\mathbb{R}^D} K_S(\boldsymbol{x}))^{-1}$;
  - Smoothing scale: $\sigma_n \asymp n^{-\frac{1}{2m_f+d}}$.
- The above statements hold for both cases where early stopping is
  - without weight decay (iteration number $t \asymp n^{\frac{2m_0+2m_f}{2m_f+d}}$);
  - with weight decay (weight decay rate $\alpha \asymp n^{-1-\frac{2(m_0+m_\varepsilon)}{2m_f+d}}$, and iteration number $t \ge C_2(\frac{m_f}{2m_f+d} + 1/2)\log n/(\log(1-\alpha)).$).

Mixed smooth Sobolev Space

For a function f defined on $\mathbb{R}^D$, the mixed smooth Sobolev norm is defined as

$$\|f\|_{\mathcal{MW}^m(\mathbb{R}^D)} = \left(\int_{\mathbb{R}^D} |\mathcal{F}(f)(\omega)|^2 \prod_{j=1}^{D} (1 + |\omega_j|^2)^m \mathrm{d}\omega \right)^{1/2}. \quad (5)$$

- Under appropriate choice of the smoothing scale, stepsize, and iteration number, we have

$$\|f_t - f^*\|_{L_2(P_{\mathbf{x}})}^2 = O_{\mathbb{P}} \left(n^{-\frac{2m_f}{2m_f+1}} (\log n)^{\frac{2m_f}{2m_f+1}\left(D-1+\frac{1}{2(m_0+m_\varepsilon)}\right)} \right). \quad (6)$$

  for $N > N_0$, where N is the number of augmentations.
- The above statements hold for both cases where early stopping is with weight decay and is without weight decay.

Numerical experiments

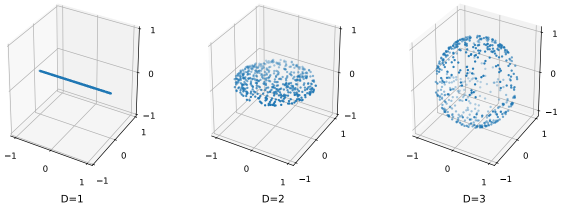

Figure: Simulated data spaces in the forms of: line ($D = 1$), circle ($D = 2$) and sphere ($D = 3$).

Numerical experiments

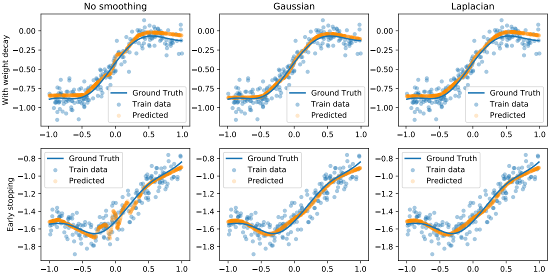

Figure: Underlying truth (blue curve), training data (blue dots), and neural network predictions (orange dots) when training size is 200.

Conclusions

- This work studies random smoothing kernel and random smoothing regularization, which have a natural relationship with data augmentations.
- We show that by applying random smoothing, with appropriate early stopping and/or weight decay techniques, the resulting estimator can achieve fast convergence rates, regardless of the kernel function used in the construction of the random smoothing kernel estimator.