Benign Overfitting in Token Selection of Attention Mechanism

Keitaro Sakamoto, Issei Sato

The University of Tokyo

Paper Link

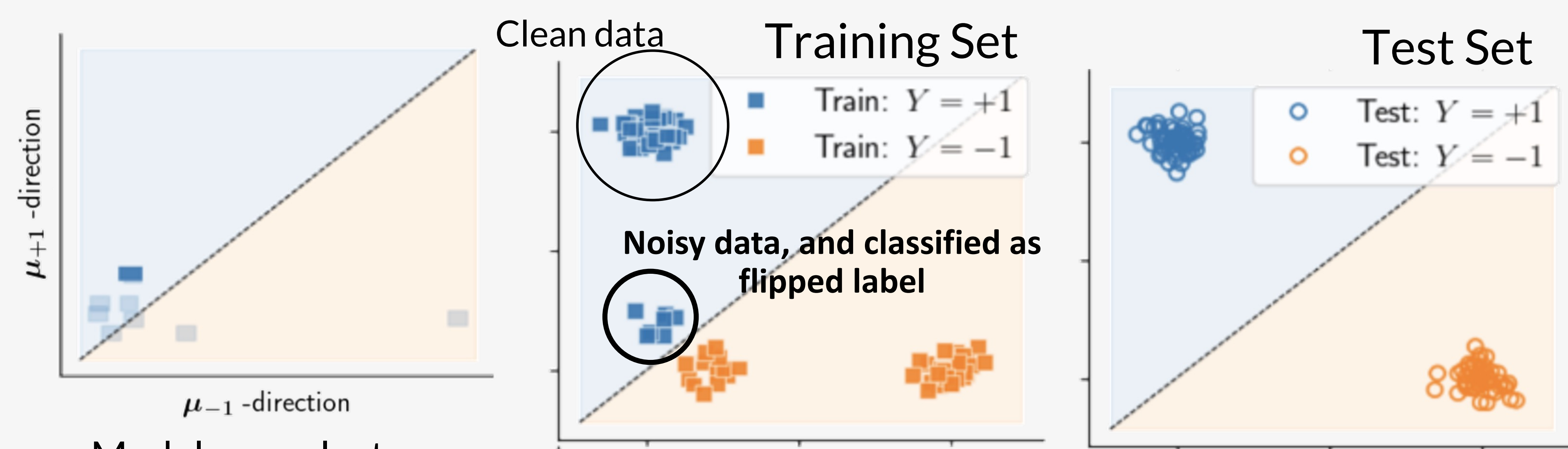
Summary

Analysis of "*benign overfitting*" in the token selection of attention mechanism under label noise setting.

Benign overfitting: Achieve high generalization while perfectly fitting training data in an over-parameterized model.

→ Overfits training data, but surprisingly, without hurting generalization.

1. *How do the training dynamics of token selection in attention evolve under label noise?*
2. *Does the obtained solution generalize well?*



Model can select one token for each input $\mathbf{X}^{(i)} = (\mathbf{x}_1^{(i)}, \dots, \mathbf{x}_T^{(i)})^\top \in \mathbb{R}^{T \times d}$

Overfits to label noise (Memorizes training labels)

Still generalizes well

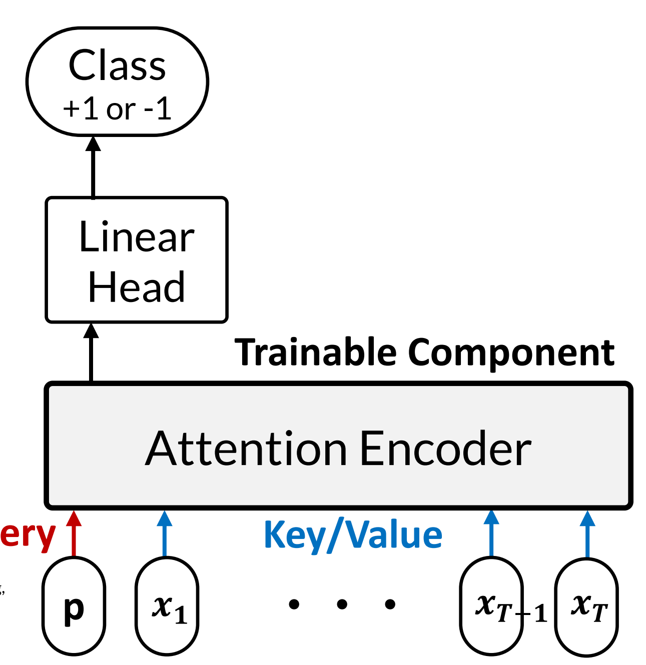
Problem Setting

Model

$$f(\mathbf{X}) = \boldsymbol{\nu}^\top \mathbf{X}^\top \mathbb{S}(\mathbf{X}\mathbf{W}^\top \mathbf{p})$$

The output corresponding to the [CLS] token.

$\mathbb{S}(\cdot)$	Softmax function.
$\mathbf{X} \in \mathbb{R}^{T \times d}$	Sequence of input tokens $(\mathbf{x}_1, \dots, \mathbf{x}_T)^\top$
$\mathbf{W} \in \mathbb{R}^{d \times d}$	Key-query weight matrix $\mathbf{W}_Q \mathbf{W}_K^\top$
$\mathbf{p} \in \mathbb{R}^d$	Tunable token [CLS] token [Devlin+, 2018; Dosovitskiy+, 2021] or prompt tuning [Li & Liang, 2021; Lester+, 2021].



Training

$$\widehat{\mathcal{L}}(\mathbf{W}, \mathbf{p}) = \frac{1}{n}\sum_{i=1}^{n} \ell\left(Y^{(i)} \cdot f(\mathbf{X}^{(i)})\right), \quad \ell(z) = \log(1+\exp(-z)) \quad \text{Binary cross-entropy}$$

Gradient descent with a step size $\alpha > 0$.

$$\mathbf{W}(\tau+1) = \mathbf{W}(\tau) - \alpha \nabla_{\mathbf{W}} \widehat{\mathcal{L}}(\mathbf{W}(\tau), \mathbf{p}(\tau)), \quad \mathbf{p}(\tau+1) = \mathbf{p}(\tau) - \alpha \nabla_{\mathbf{p}} \widehat{\mathcal{L}}(\mathbf{W}(\tau), \mathbf{p}(\tau))$$

Data

1. True label $Y^* \sim \mathrm{Unif}(\{\pm 1\})$, $Y = \begin{cases} Y^* & \text{with probability } 1-\eta \\ -Y^* & \text{with probability } \eta \end{cases}$, $\mathcal{C}$: Clean examples, $\mathcal{N}$: Noisy examples
2. Class signals $\boldsymbol{\mu}_{+1}$ and $\boldsymbol{\mu}_{-1}$, such that $\langle \boldsymbol{\mu}_{+1}, \boldsymbol{\mu}_{-1} \rangle = 0$ and $\|\boldsymbol{\mu}\|_2 = \|\boldsymbol{\mu}_{+1}\|_2 = \|\boldsymbol{\mu}_{-1}\|_2$
3. Input $\mathbf{X} = (\mathbf{x}_1, \mathbf{x}_2, \mathbf{x}_3, \dots, \mathbf{x}_T)^\top$

Relevant token	Confusing token	Weakly Relevant / Irrelevant token	
$\boldsymbol{\mu}_{Y^*} + \boldsymbol{\epsilon}_1$	$\rho\boldsymbol{\mu}_{-Y^*} + \boldsymbol{\epsilon}_2$	$\rho\boldsymbol{\mu}_{Y^*} + \boldsymbol{\epsilon}_t$	$\boldsymbol{\epsilon}_t$

Noise vectors $\boldsymbol{\epsilon}_t \sim N(0, \sigma_\epsilon^2 I)$

Signal-to-noise ratio $\mathrm{SNR} = \|\boldsymbol{\mu}\|_2 / (\sigma_\epsilon \sqrt{d})$

*$\rho \ll 1$: Small scale parameter representing weak class information

Main Result

Theorem (Informal)

Suppose that the norm of the linear head scales as $\|\boldsymbol{\nu}\|_2 = O(1/\|\boldsymbol{\mu}\|_2)$. Under some parameter assumptions (*, see our paper for details), we have

1. (Not overfitting) If $\mathrm{SNR}^2 = \omega(n^{-1})$, then with probability at least $1-\delta$, there exists a time

$$\tau = \Theta\left(\frac{1}{\alpha \|\boldsymbol{\nu}\|_2 \|\boldsymbol{\mu}\|_2^3 d \max\{\sigma_w^2, \sigma_p^2\}}\right) \text{ such that:}$$

$$\forall i \in \mathcal{C},\ f_\tau(\mathbf{X}^{(i)}) = Y^{(i)},\ \forall j \in \mathcal{N},\ f_\tau(\mathbf{X}^{(j)}) \neq Y^{(j)},\ \Pr_{(\mathbf{X}, Y^*) \sim P^*}[\mathrm{sign}(f_\tau(\mathbf{X})) \neq Y^*] < \delta.$$

2. (Benign overfitting) If $\mathrm{SNR}^2 = o(n^{-1})$, then with probability at least $1-\delta$, there exists a time

$$\tau = \Theta\left(\frac{\exp(n^{-1}\mathrm{SNR}^{-2})}{\alpha n^{-1} \sigma_\epsilon^2 \|\boldsymbol{\nu}\|_2 \|\boldsymbol{\mu}\|_2 d^2 \max\{\sigma_w^2, \sigma_p^2\}}\right) \text{ such that:}$$

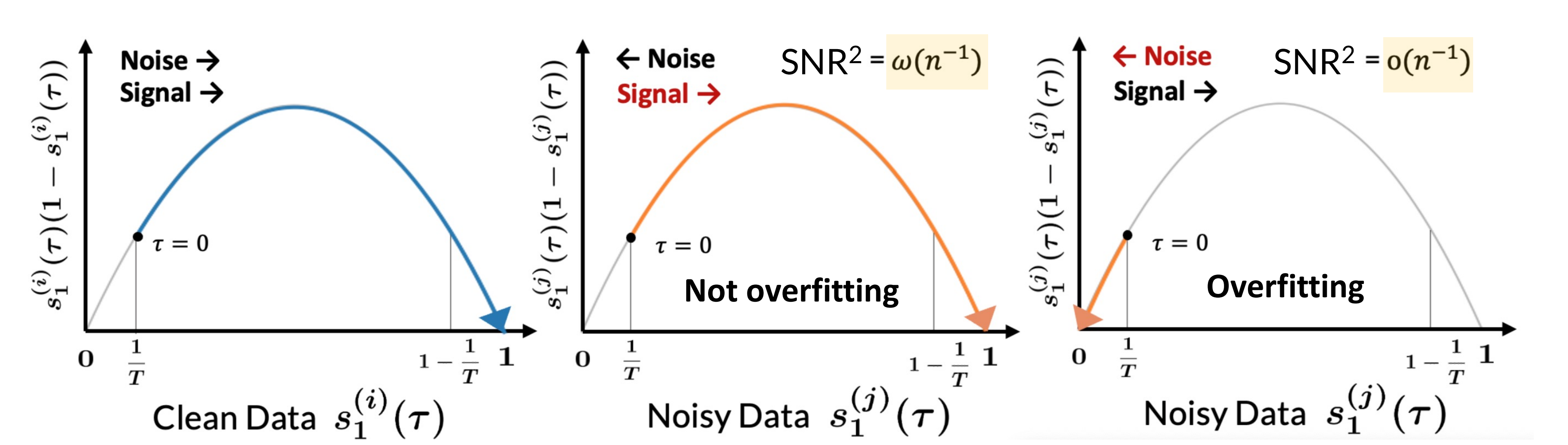
Generalization after overfitting requires **exponentially long training** (see Grokking).

$$\forall i \in [n],\ f_\tau(\mathbf{X}^{(i)}) = Y^{(i)},\ \Pr_{(\mathbf{X}, Y^*) \sim P^*}[\mathrm{sign}(f_\tau(\mathbf{X})) \neq Y^*] < \delta$$

* For example, we have $\mathrm{SNR}^2 = \Omega(d^{-1/4})$

For noisy data $j \in \mathcal{N}$, the class relevant token $\mathbf{x}_1^{(j)}$ should **NOT** be picked to decrease the training loss.

→ Noise memorization suppresses the probability of selecting $\mathbf{x}_1^{(j)}$ to zero (Figure, right). Furthermore, benign overfitting claims that such memorization does not adversely affect generalization.



* Y-axis essentially represents the magnitude of gradient updates.

Difficulties Specific to Attention

We must handle **two competing directions** in the same training run.

1. Clean samples $\mathcal{C}$ vs Noisy samples $\mathcal{N}$ to learn signals
2. Signal learning $\boldsymbol{\mu}$ vs Memorization $\{\boldsymbol{\epsilon}_t\}_{t \in [T]}$ in token selection

Two-layer NN $f(\mathbf{x}) = \boldsymbol{\nu}^\top \sigma(\mathbf{W}\mathbf{x}) = \sum_{j=1}^{m} \nu_j \sigma(\mathbf{w}_j^\top \mathbf{x})$, where $\mathbf{W} = \begin{pmatrix} \mathbf{w}_1^\top \\ \vdots \\ \mathbf{w}_m^\top \end{pmatrix}$

$$-\frac{\partial \widehat{\mathcal{L}}}{\partial \mathbf{w}_j} = \frac{1}{n}\sum_{i=1}^{n} \underbrace{(-\ell_i'(Y^{(i)} f(\mathbf{x}^{(i)})))}_{\approx \text{Loss at } (\mathbf{x}^{(i)}, Y^{(i)})} \cdot Y^{(i)} \nu_j \cdot \underbrace{\sigma'(\mathbf{w}_j^\top \mathbf{x}^{(i)})}_{= 1 \text{ or } 0 \text{ if ReLU}} \mathbf{x}^{(i)}$$

Once neuron $j \in [m]$ is activated (i.e., $\sigma'(\cdot) = 1$), the weights are updated until the loss decreases.

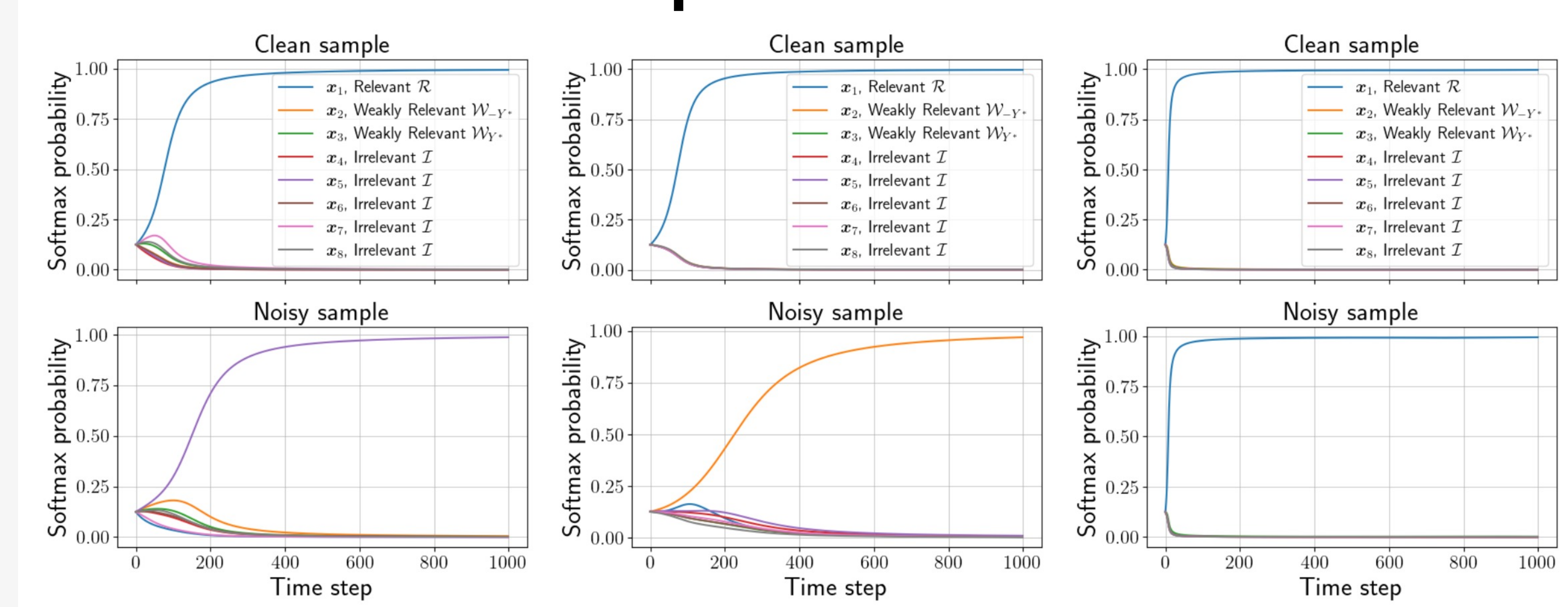
Attention $f(\mathbf{X}) = \boldsymbol{\nu}^\top \mathbf{X}^\top \mathbb{S}(\mathbf{X}\mathbf{W}^\top \mathbf{p})$, where $\mathbf{X} = \begin{pmatrix} \mathbf{x}_1^\top \\ \vdots \\ \mathbf{x}_T^\top \end{pmatrix}$

$$-\frac{\partial \widehat{\mathcal{L}}}{\partial \mathbf{p}} = \frac{1}{n}\sum_{i=1}^{n} \underbrace{(-\ell_i'(Y^{(i)} f(\mathbf{X}^{(i)})))}_{\approx \text{Loss at } (\mathbf{X}^{(i)}, Y^{(i)})} \left(\sum_{t=1}^{T} \mathbb{S}(\mathbf{X}^{(i)}\mathbf{W}^\top \mathbf{p})_t \left(\left(Y^{(i)} \boldsymbol{\nu}^\top \mathbf{x}_t^{(i)}\right) - \sum_{u=1}^{T} \mathbb{S}(\mathbf{X}^{(i)}\mathbf{W}^\top \mathbf{p})_u \left(Y^{(i)} \boldsymbol{\nu}^\top \mathbf{x}_u^{(i)}\right) \right) \mathbf{W}\mathbf{x}_t^{(i)} \right)$$

This term approaches **zero** both $\mathbb{S}(\mathbf{X}^{(i)}\mathbf{W}^\top \mathbf{p})_t \to 1$ (selected), $\mathbb{S}(\mathbf{X}^{(i)}\mathbf{W}^\top \mathbf{p})_t \to 0$ (not-selected)

- Learning direction depends intricately on softmax values.
- **Contribution to learning decreases as more desirable token are selected.**

Experiments



Harmful Overfitting	Benign Overfitting	Not Overfitting
train acc 1.0 / test acc 0.87	train acc 1.0 / test acc 1.0	train acc 0.8 / test acc 1.0

This result validates our theorem.

Additional experiments

- Heat-map experiments when changing SNR (Right figure)
- Real-world experiments when finetuning noisy data (MNIST, CIFAR10, MedMNIST, AG-news, TREC)