Paper Seminar

SMART: Robust and Efficient Fine-Tuning for Pre-trained Natural Language Models through Principled Regularized Optimization

Jiang et al., 2020, ACL

Myeongsup Kim

Integrated M.S./Ph.D. Student
Data Science & Business Analytics Lab.
School of Industrial Management Engineering
Korea University

Myeongsup_kim@korea.ac.kr

Introduction

- **Complexity of Deep Learning Model**
- **Complexity of Language Model**

<What This Seminar Does Not Cover>

- **Details of BERT**

  Devlin et al., BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding, NAACL, 2019

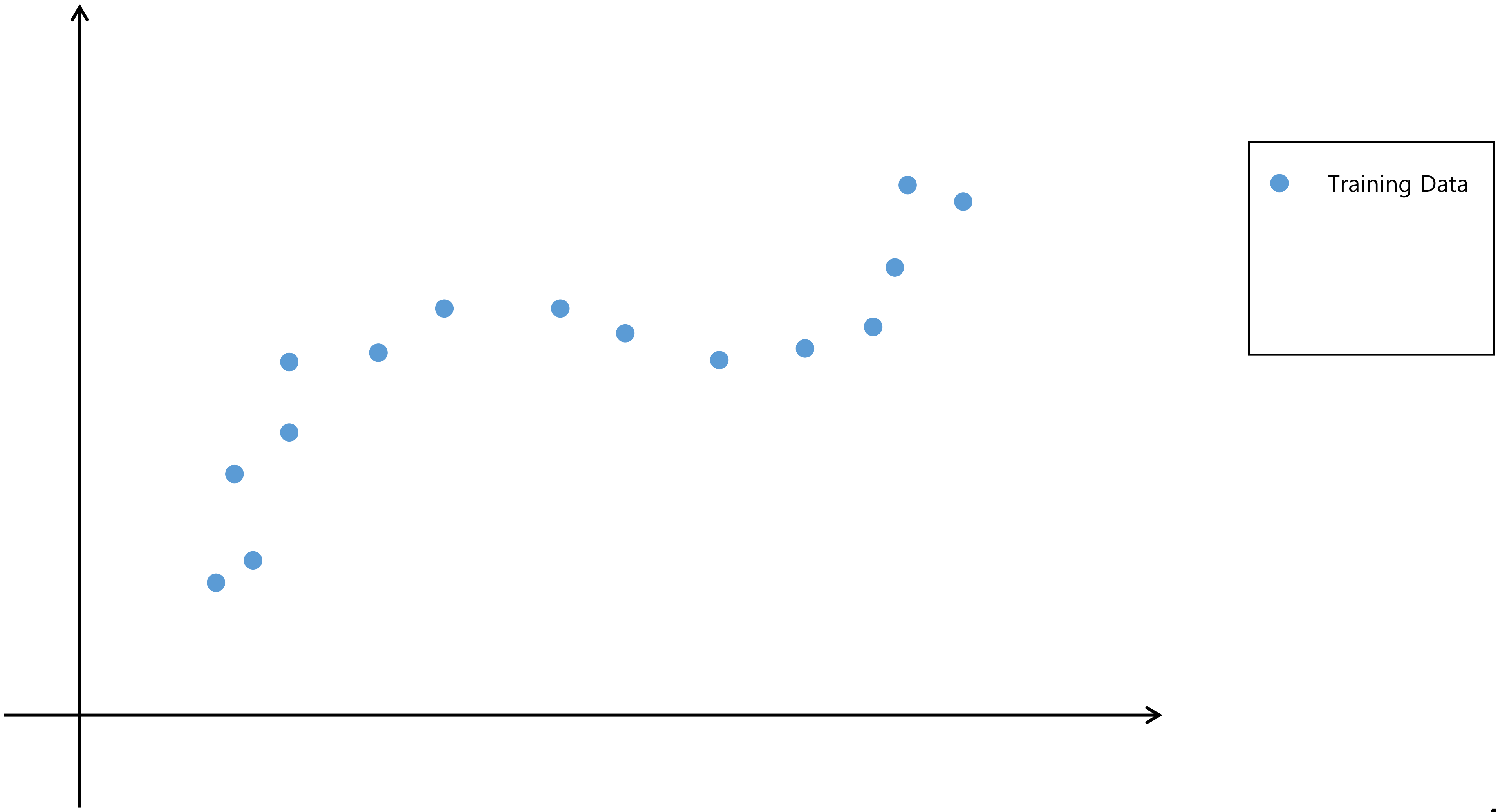
- **Details of RoBERTa**

  Liu et al., RoBERTa: A Robustly Optimized BERT Pretraining Approach, arXiv, 2019

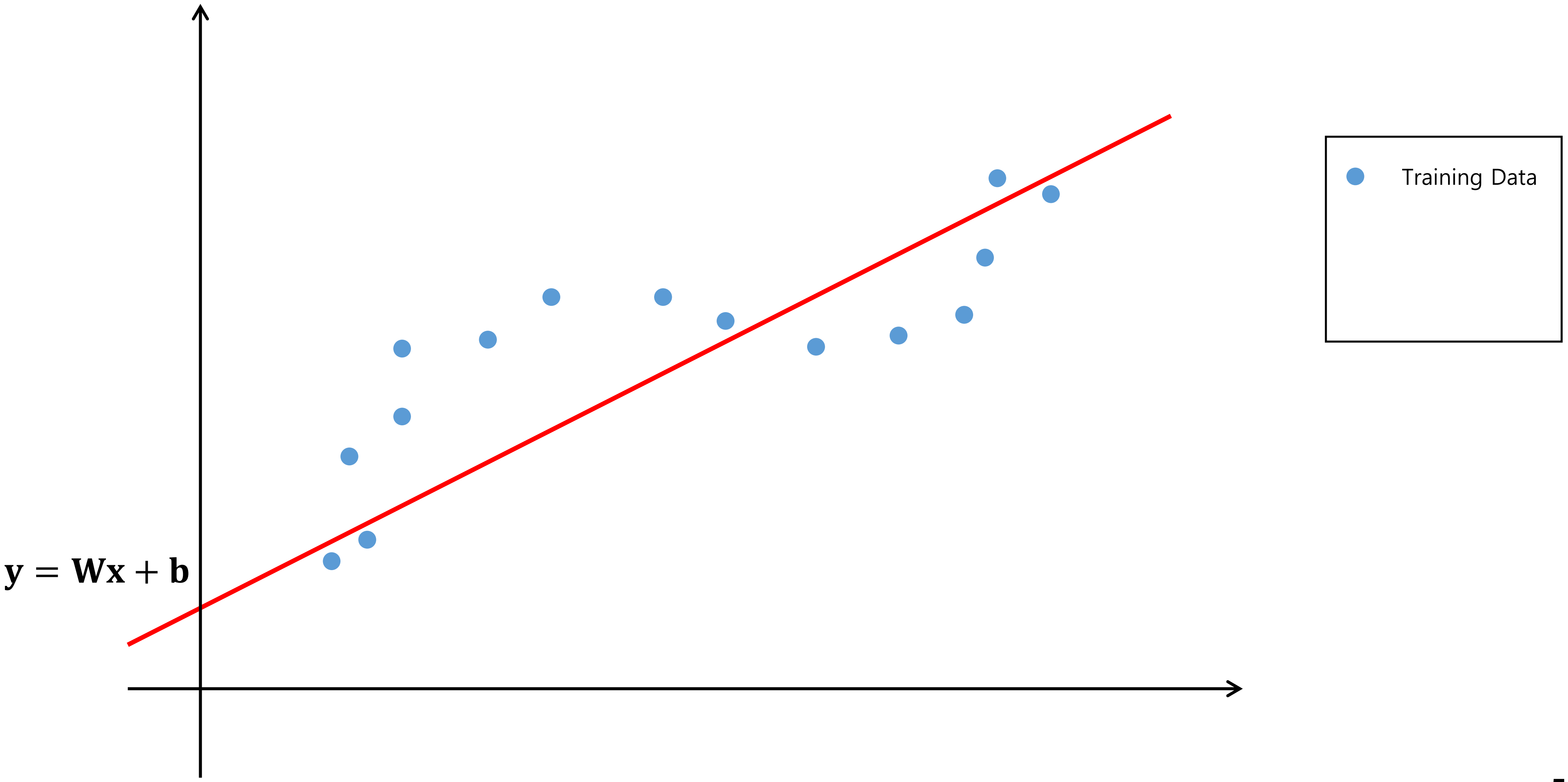
- **Details of FreeLB**

  Zhu et al., FreeLB: Enhanced Adversarial Training for Natural Language Understanding, ICLR, 2020

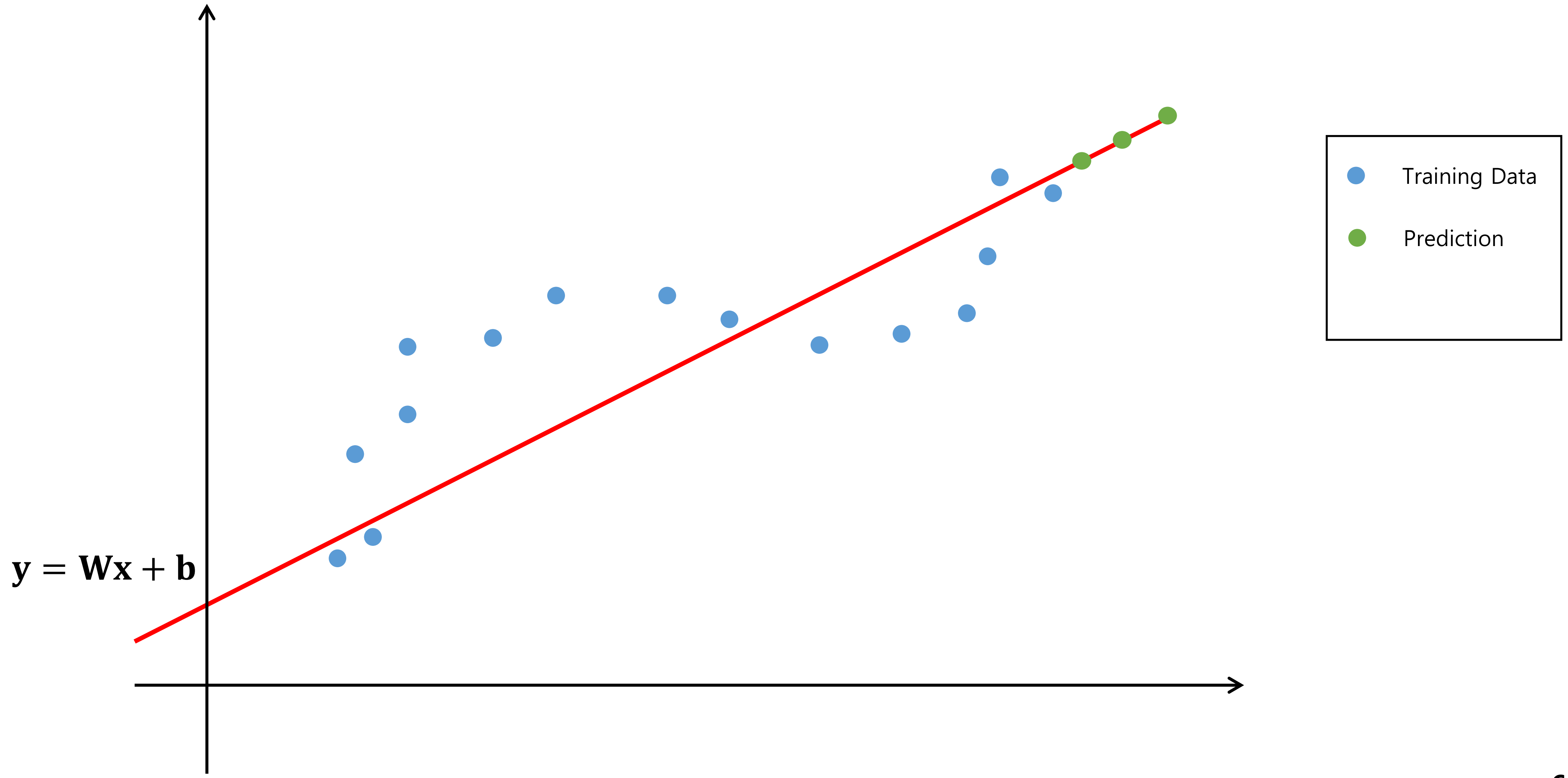
<Machine Learning Model>



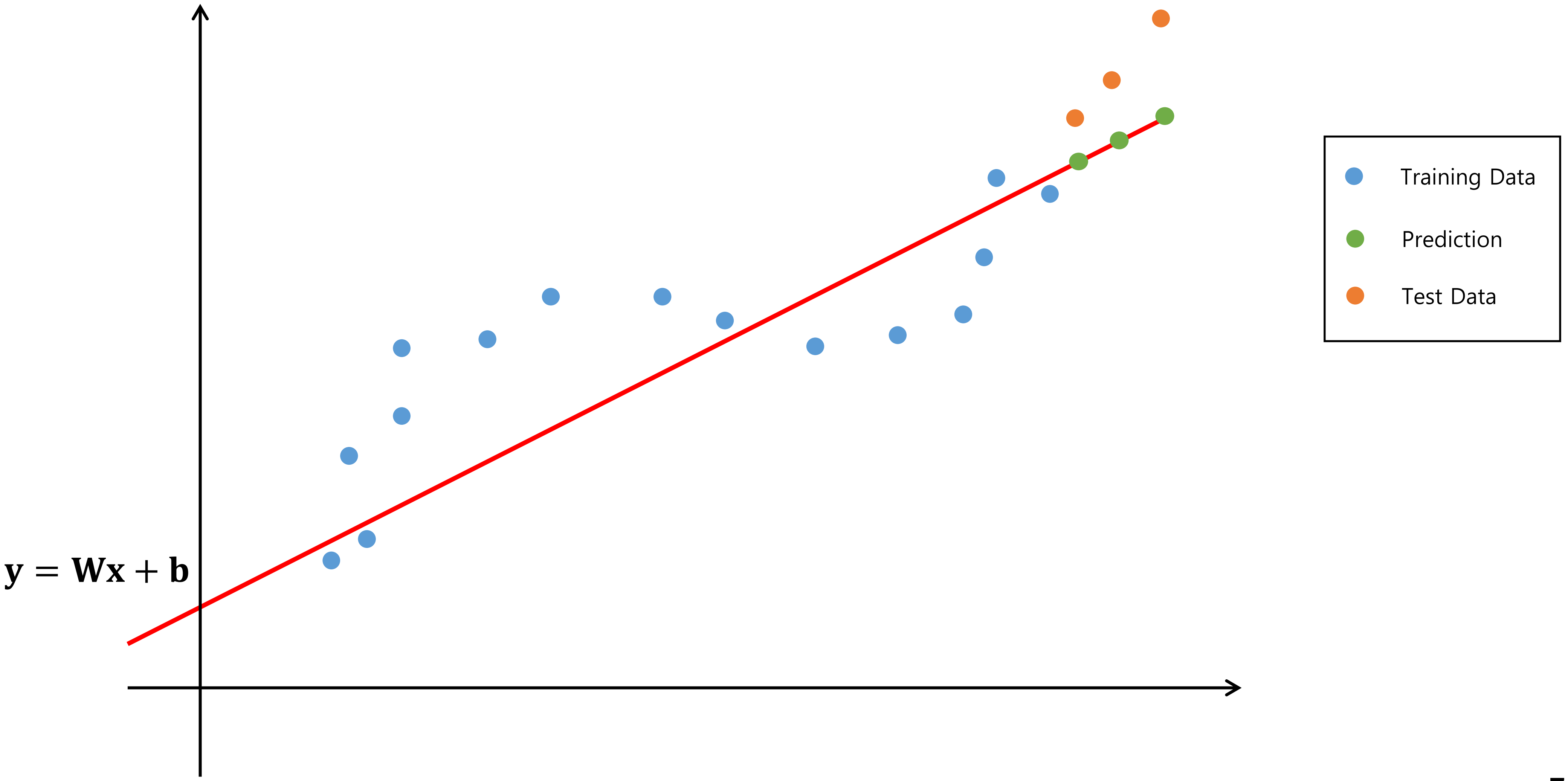
<Machine Learning Model>



<Machine Learning Model>



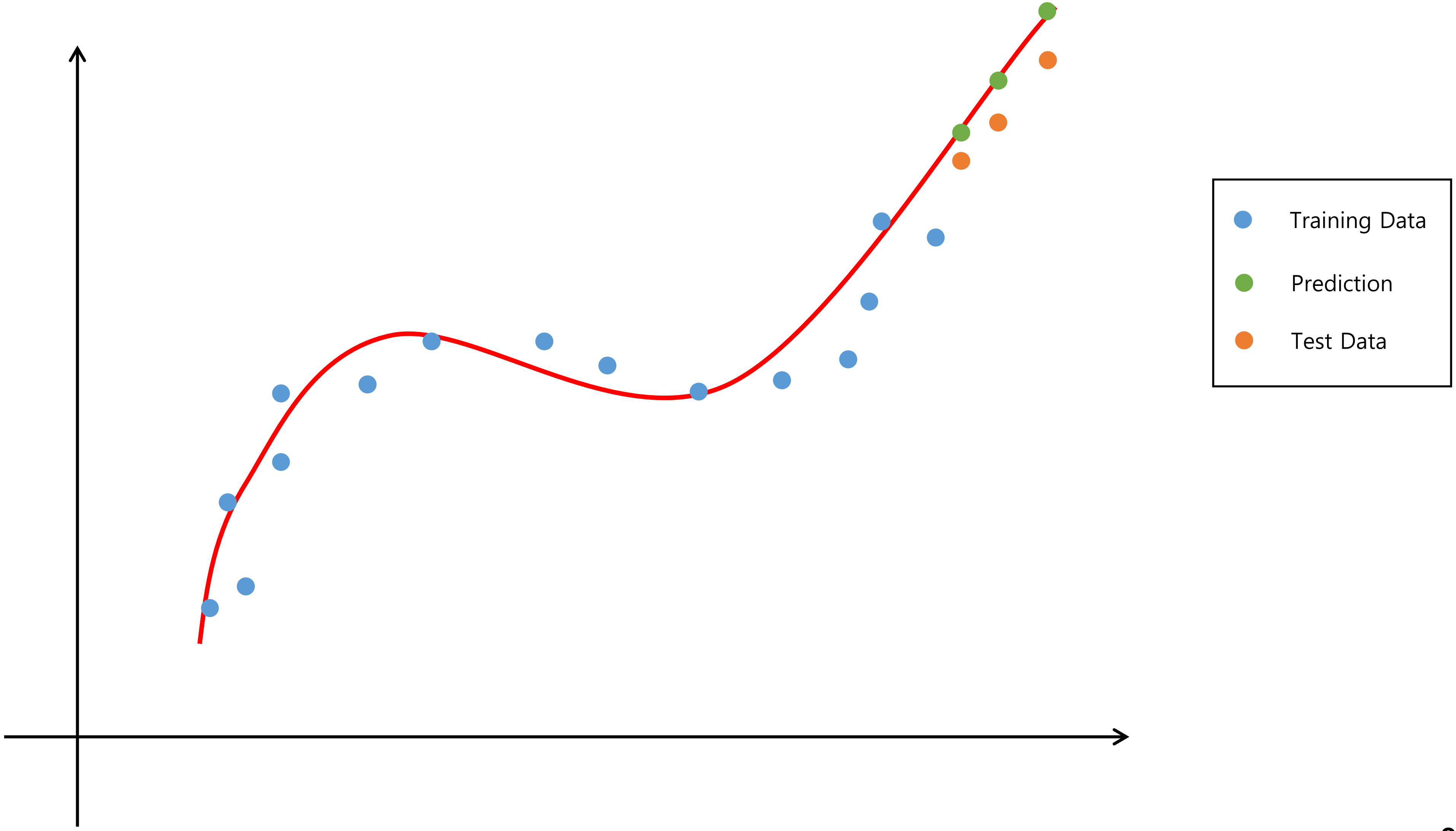
<Machine Learning Model>



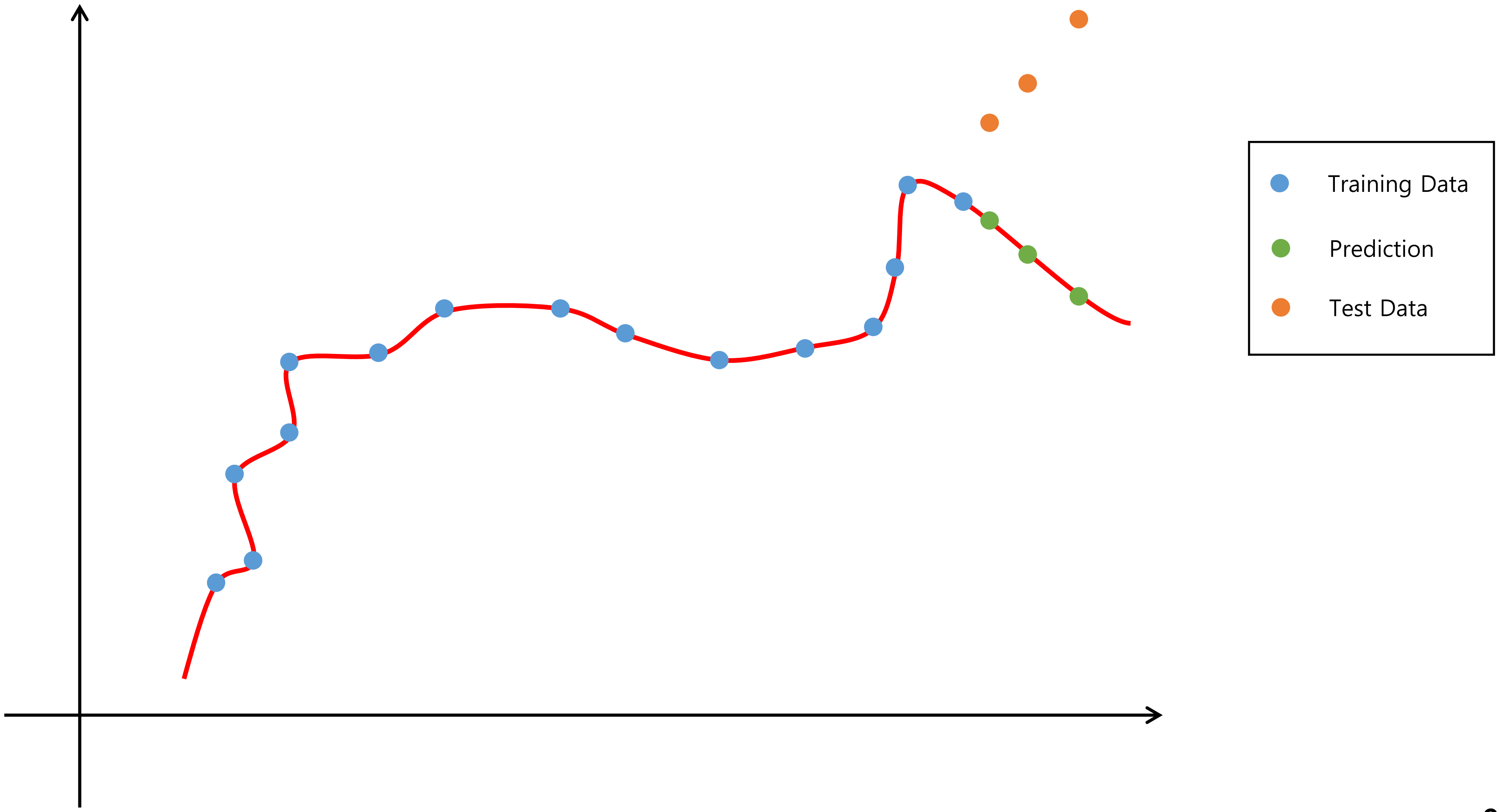
Introduction

-Complexity of Deep Learning Model

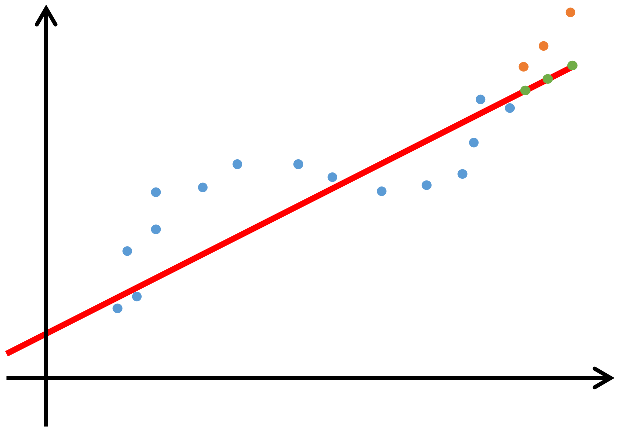
<Machine Learning Model>



<Machine Learning Model>

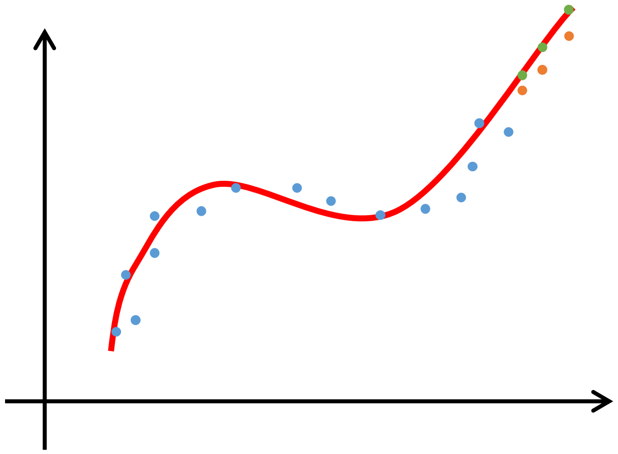


<Complexity>



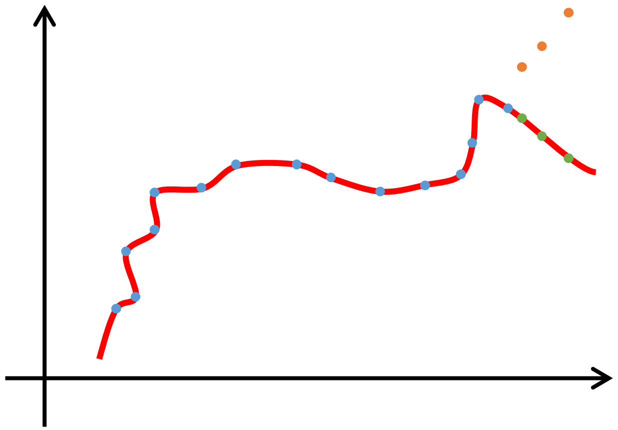
Too Low Complexity

Underfitting



Appropriate Complexity

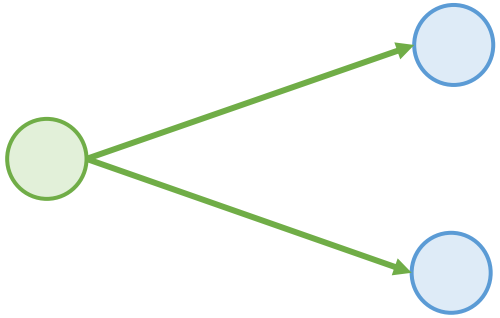
Fitted Well



Too High Complexity

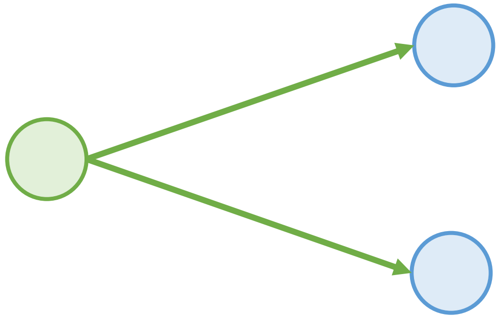
Overfitting

<Deep Learning Model>



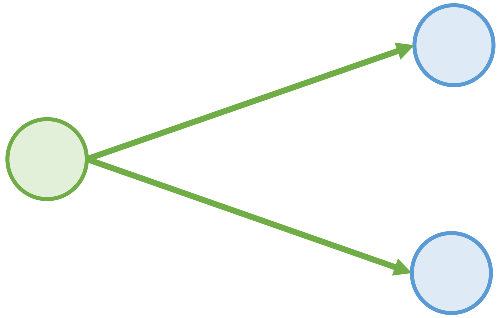
$$\mathbf{y} = \mathbf{W}\mathbf{x} + \mathbf{b}$$

<Deep Learning Model>

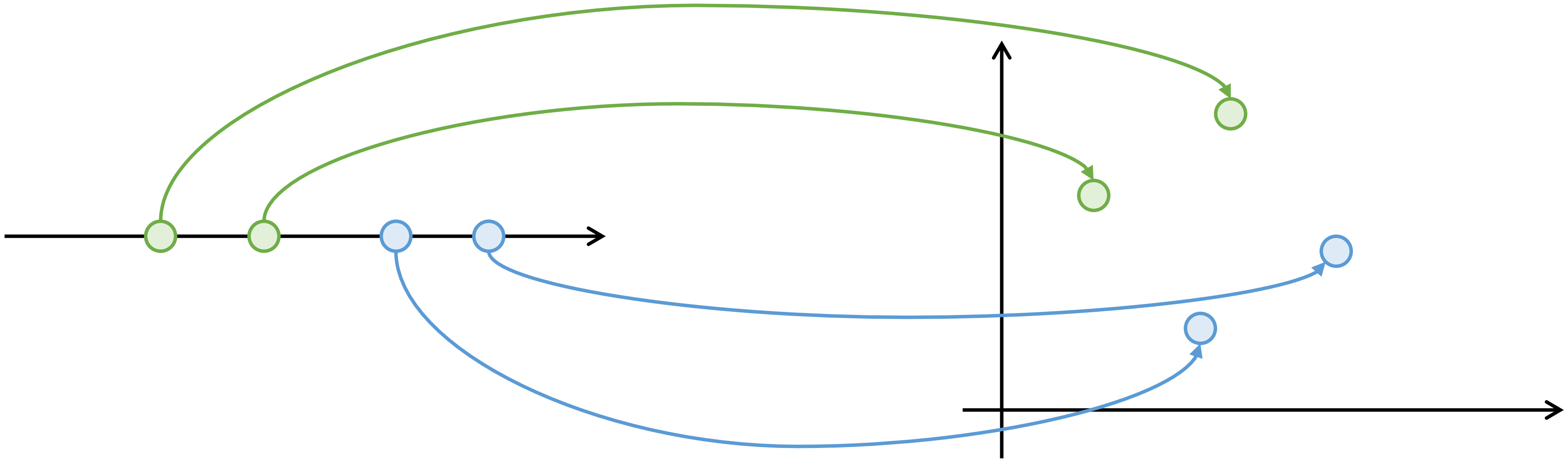


$$\mathbf{y} = \mathbf{W}\mathbf{x} + \mathbf{b}$$

<Deep Learning Model>



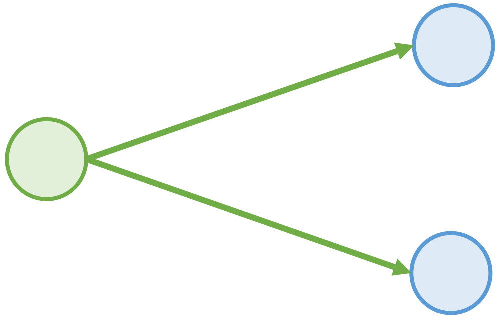
$$\mathbf{y} = \mathbf{W}\mathbf{x} + \mathbf{b}$$



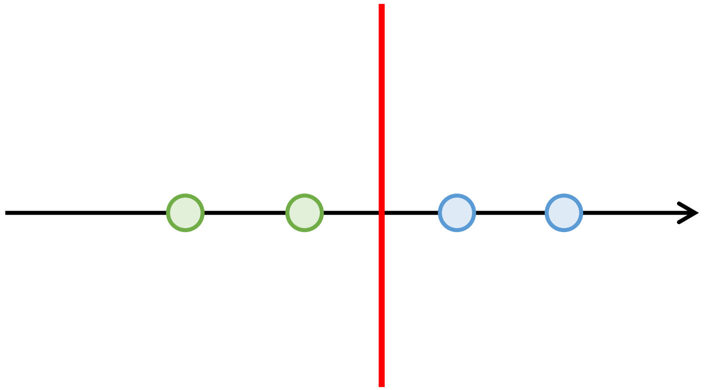
Introduction

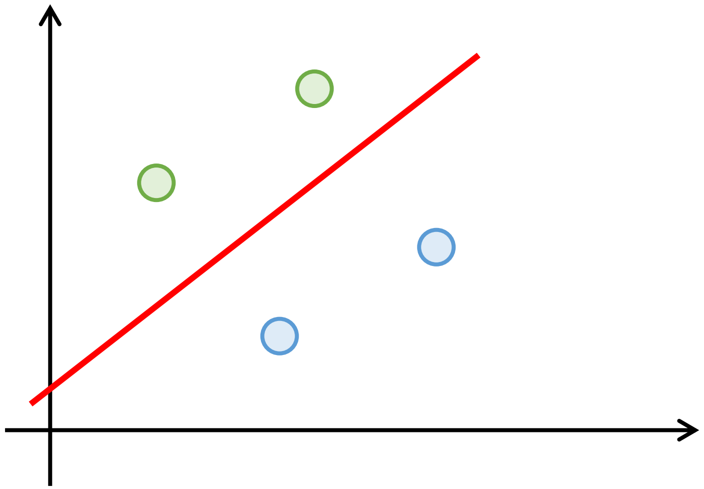
-Complexity of Deep Learning Model

<Deep Learning Model>

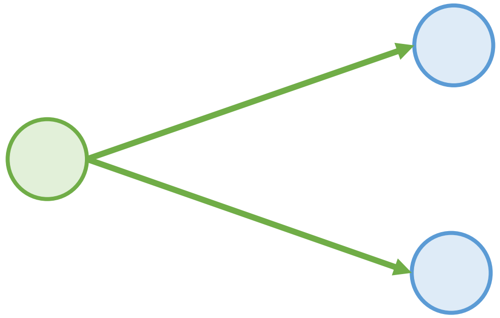


$$\mathbf{y} = \mathbf{W}\mathbf{x} + \mathbf{b}$$

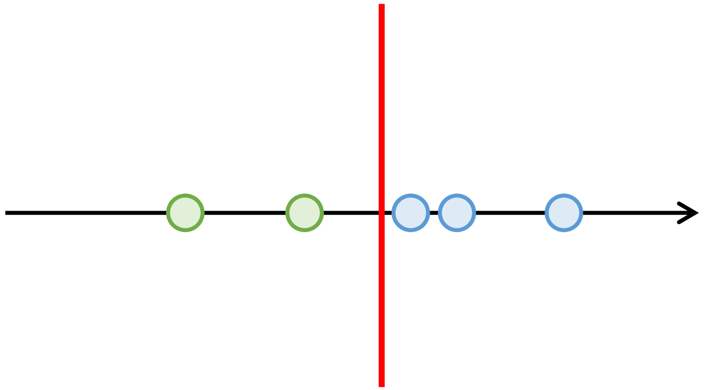


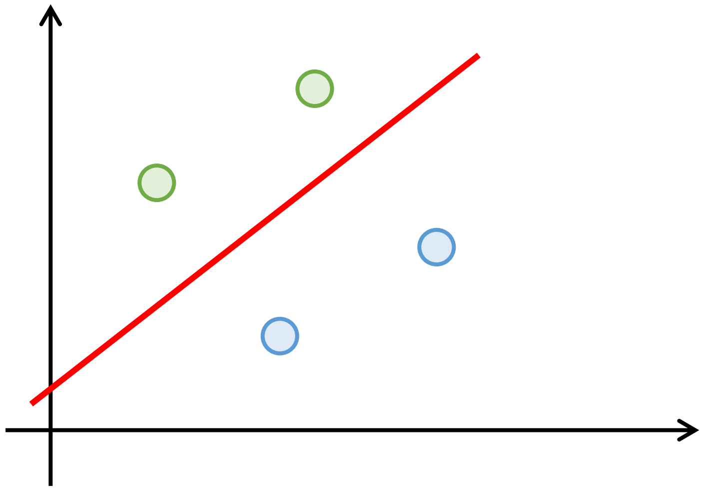


<Deep Learning Model>

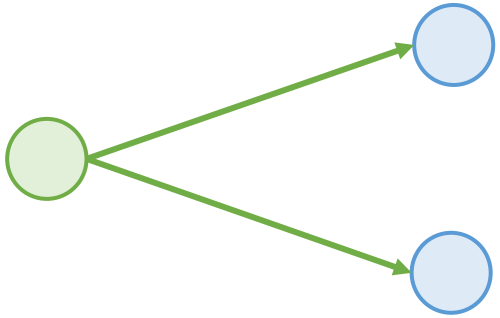


$$\mathbf{y} = \mathbf{W}\mathbf{x} + \mathbf{b}$$

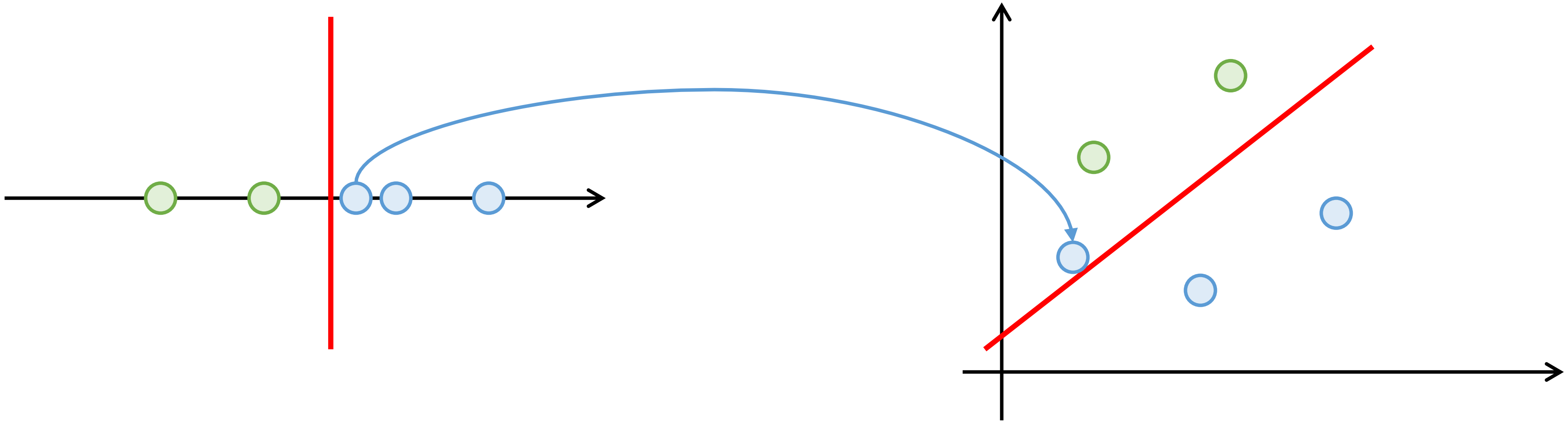




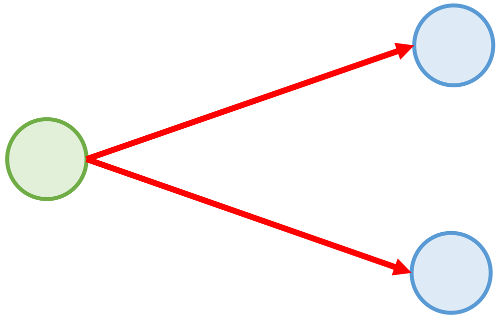
<Deep Learning Model>



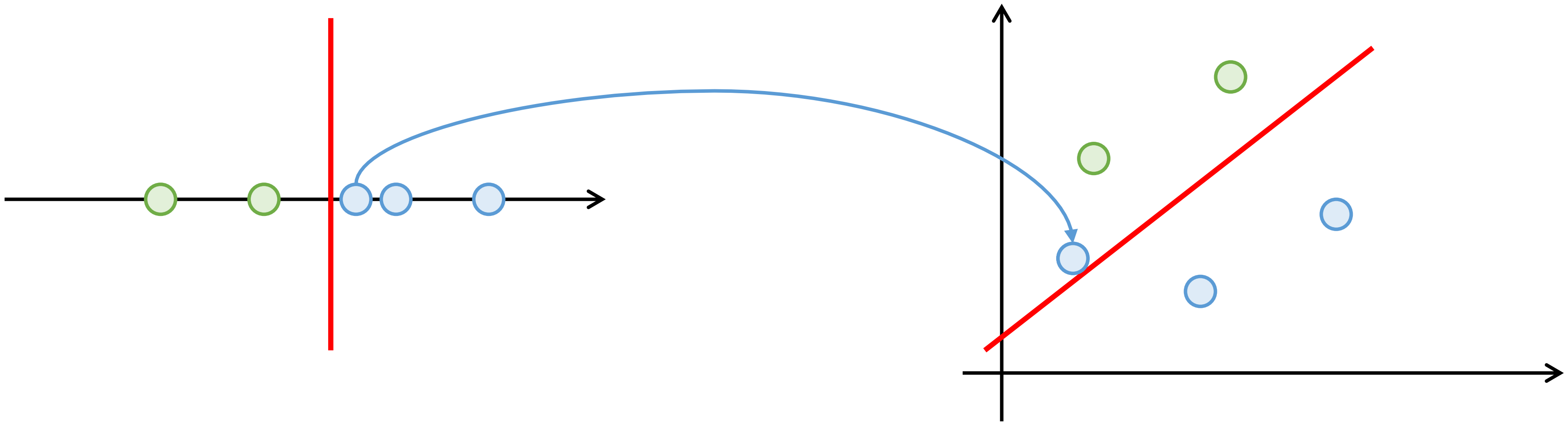
$$\mathbf{y} = \mathbf{W}\mathbf{x} + \mathbf{b}$$



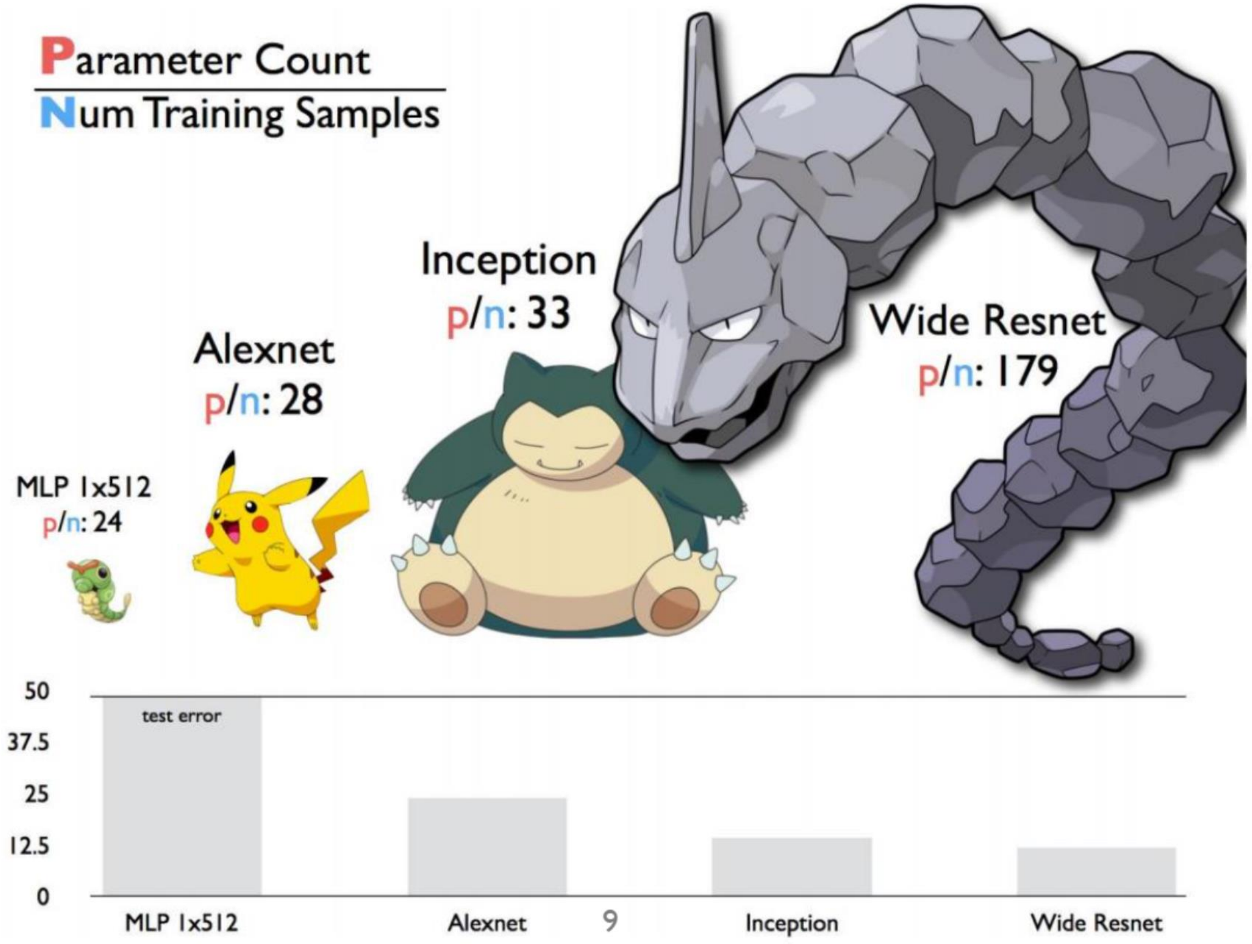
<Complexity of Deep Learning Model>



$$y = \mathbf{W}x + b$$



<Complexity of Deep Learning Model>

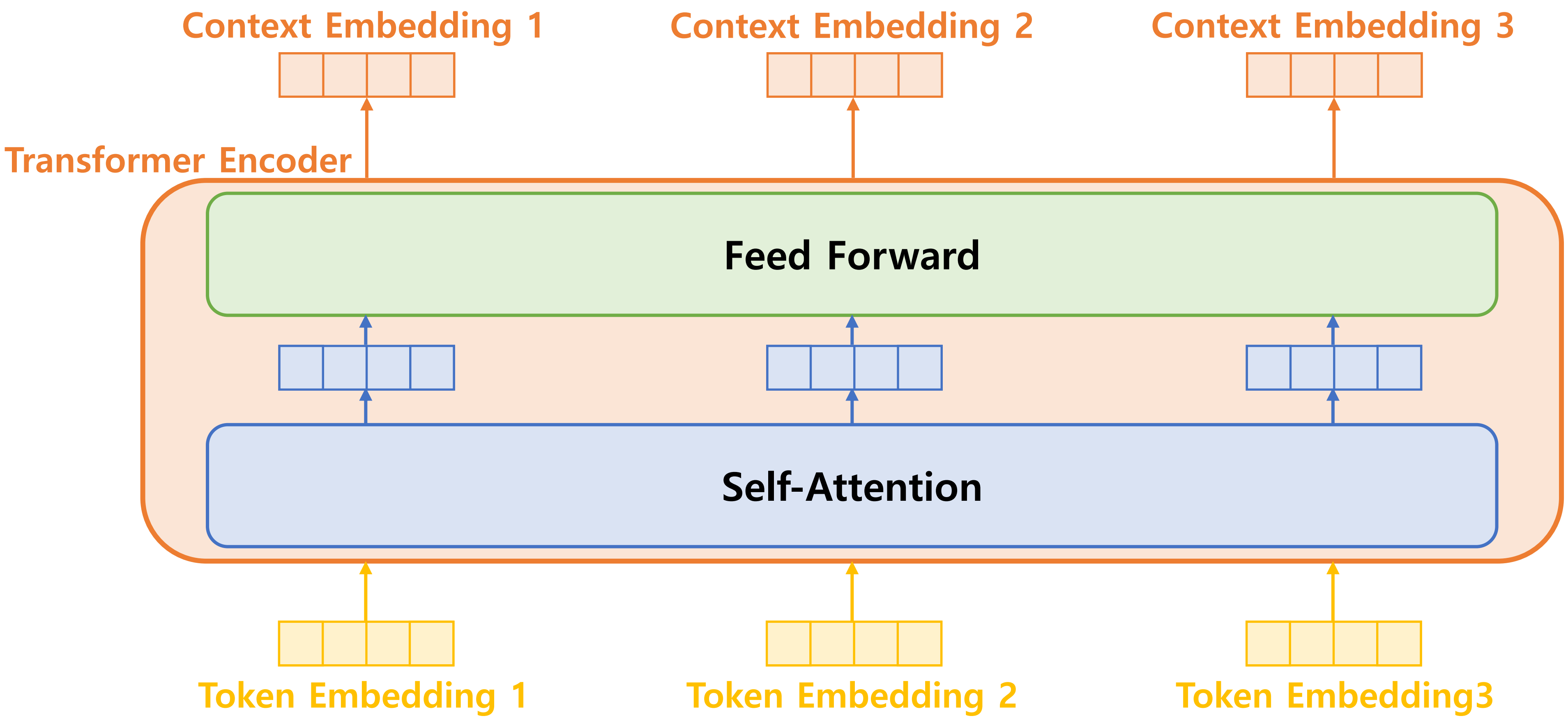


Introduction

-Complexity of Language Model

From [Paper Review] FreeLB: Enhanced Adversarial Training for Natural Language Understanding (Myeongsup Kim, 2021)

<Language Model>

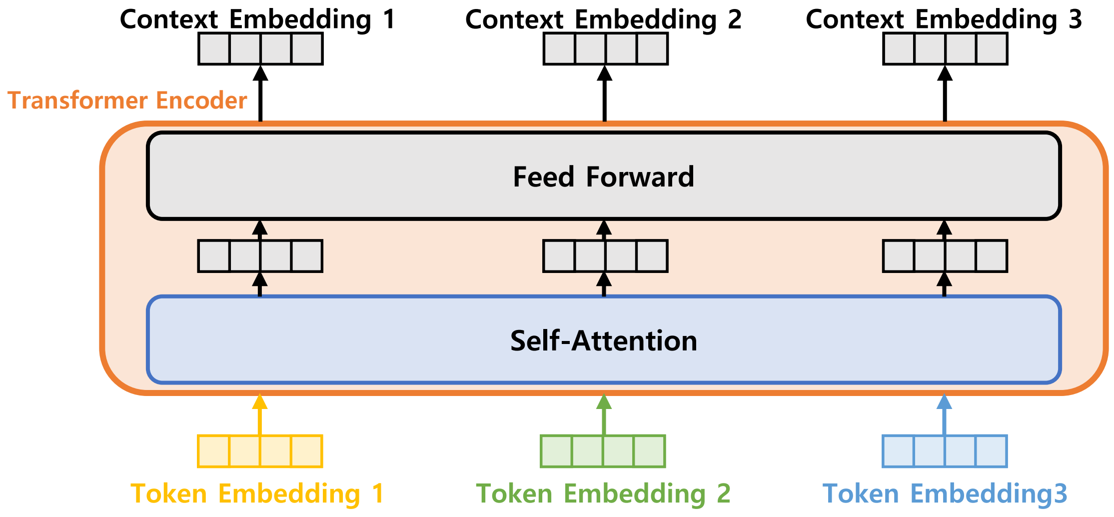


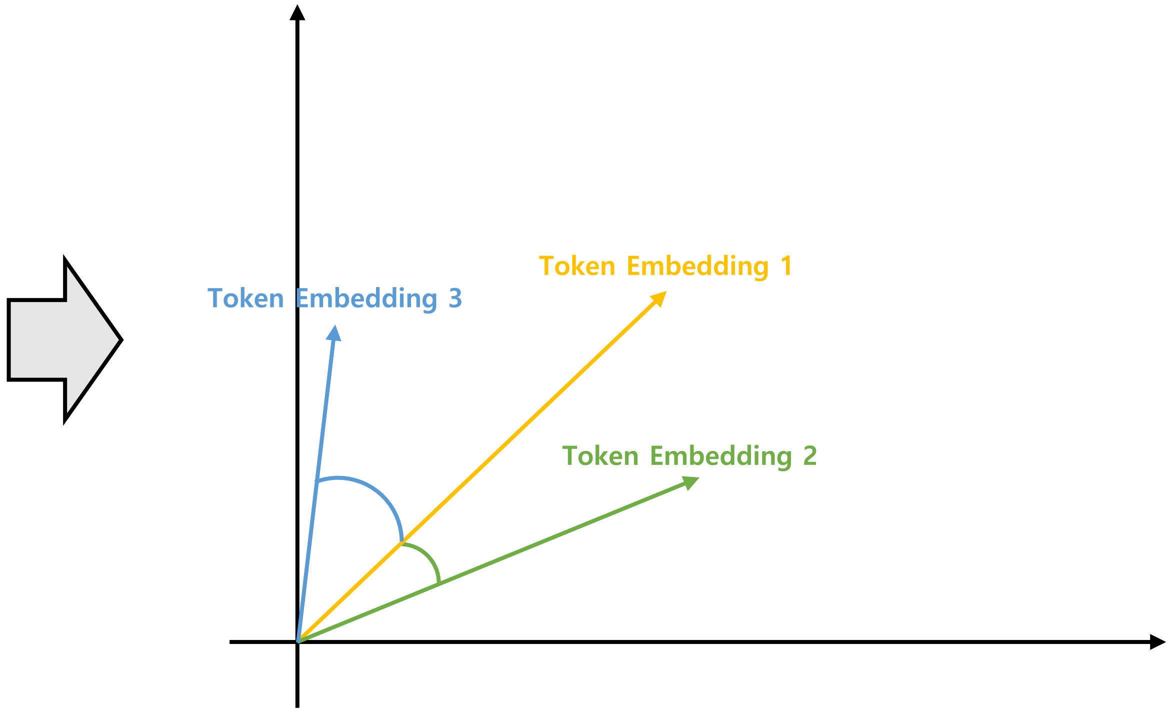
From [Paper Review] FreeLB: Enhanced Adversarial Training for Natural Language Understanding (Myeongsup Kim, 2021)

Introduction

-Complexity of Language Model

<Self Attention>



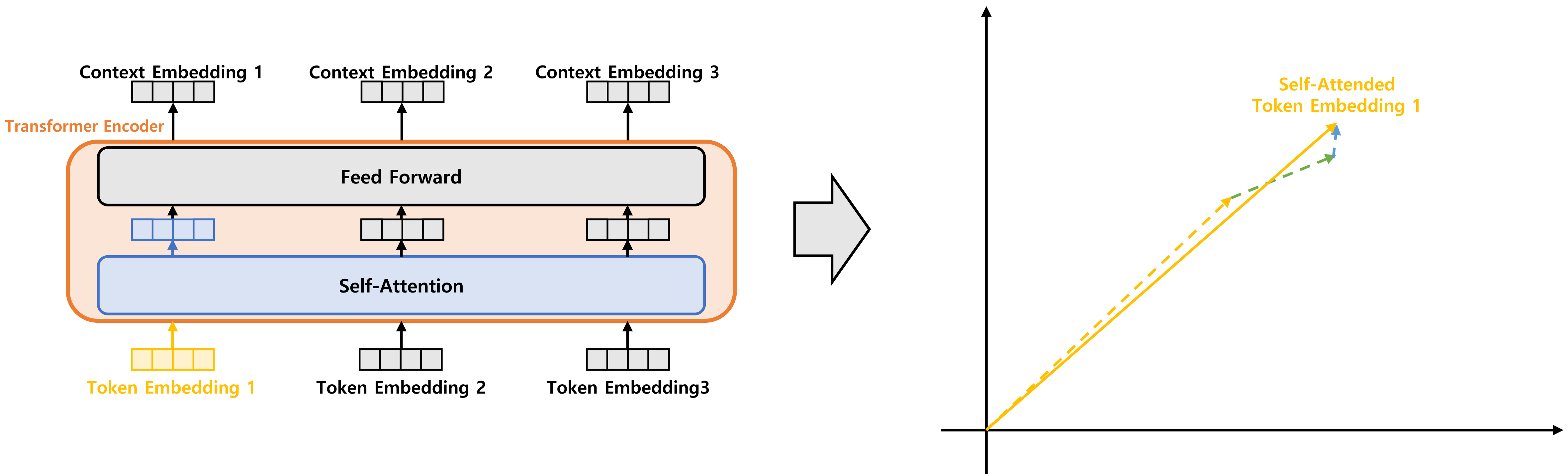


From [Paper Review] FreeLB: Enhanced Adversarial Training for Natural Language Understanding (Myeongsup Kim, 2021)

Introduction

-Complexity of Language Model

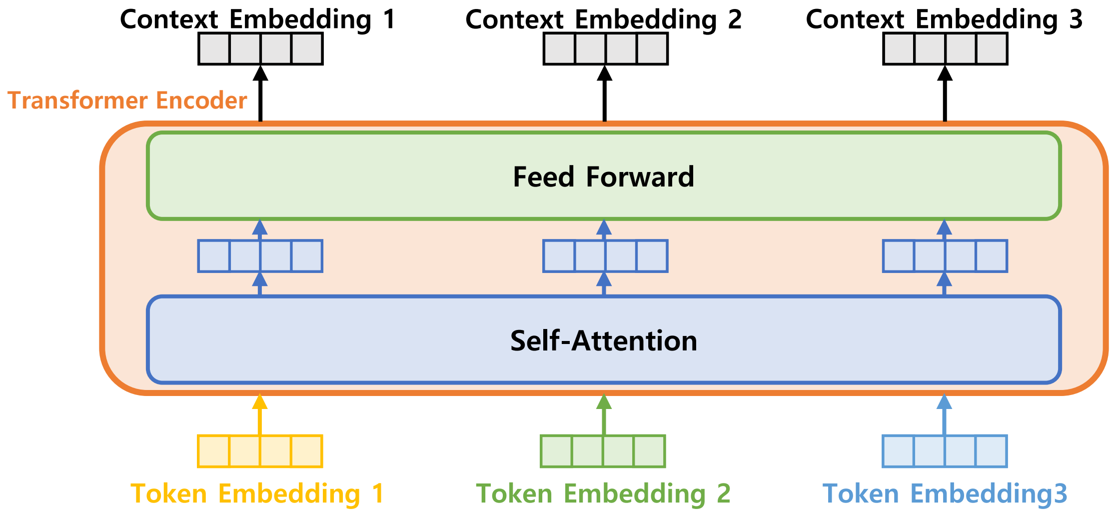
<Self Attention>

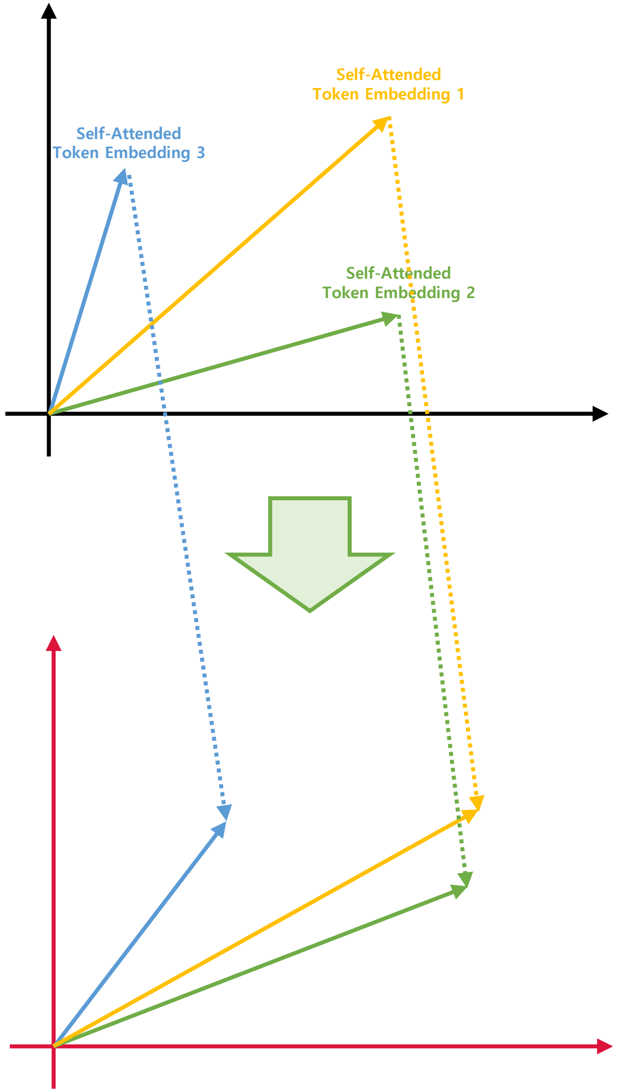


Introduction

-Complexity of Language Model

<Feed Forward>

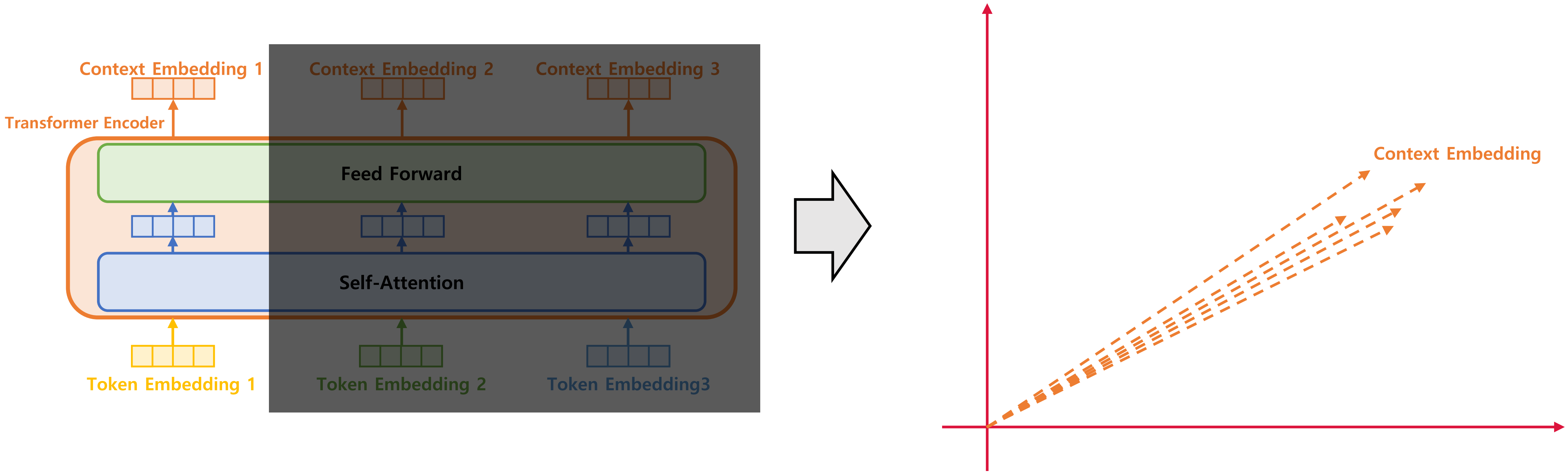




Introduction

-Complexity of Language Model

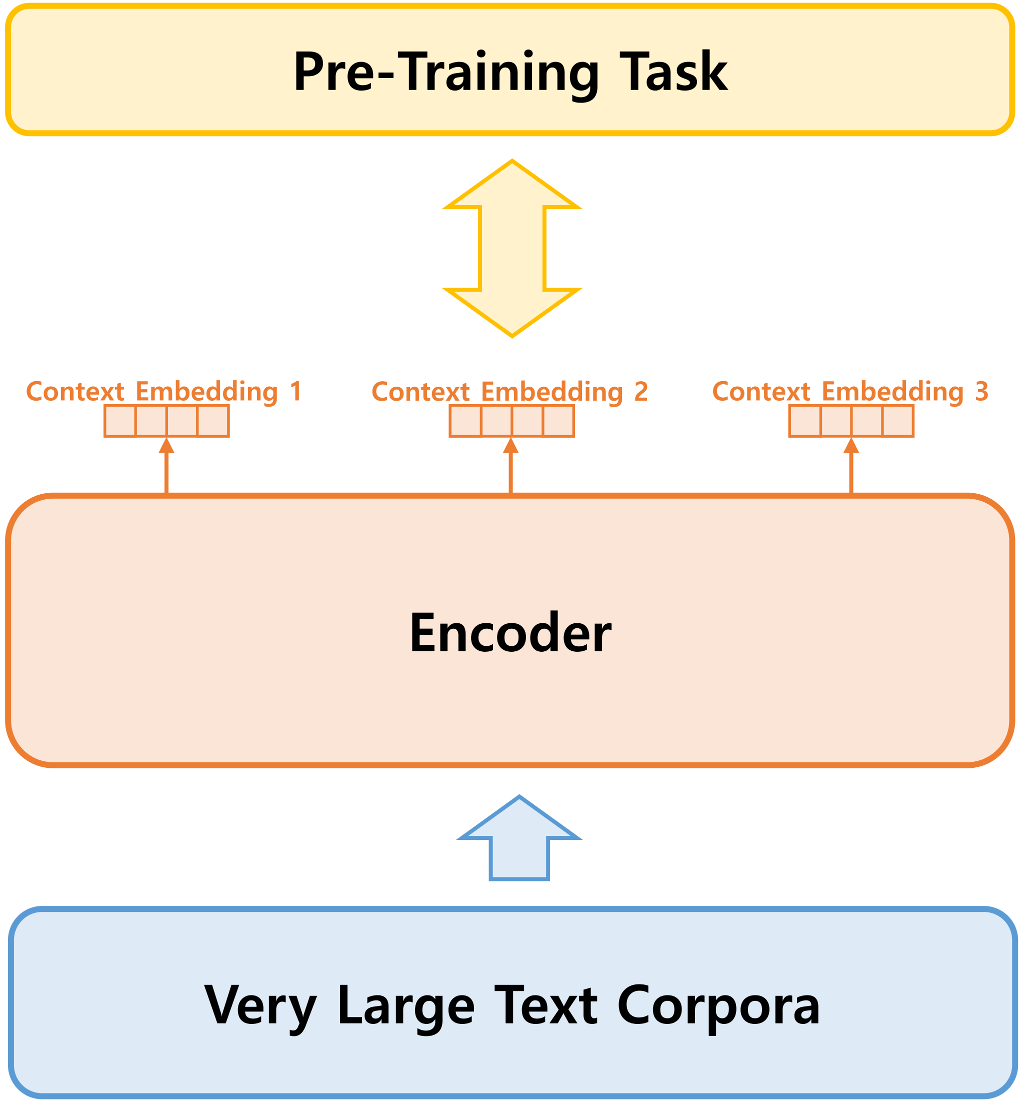
<Contextualized Representation>



Introduction

-Transformer-Based Language Model

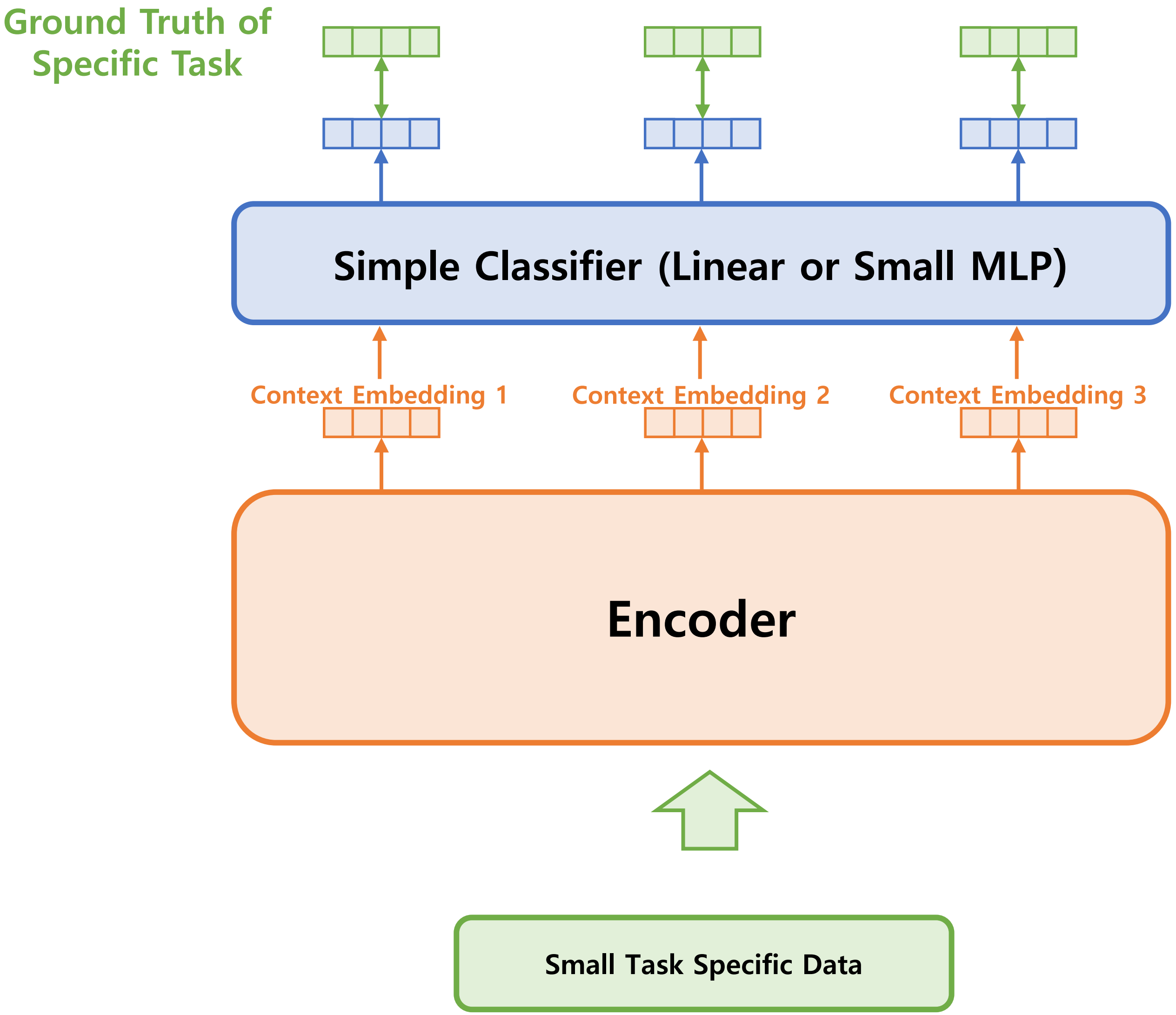
<Pre-Training>



Introduction

-Transformer-Based Language Model

<Fine-Tuning>



Introduction

-Transformer-Based Language Model

<Complexity of Language Model>

Pre-Training

Fine-Tuning

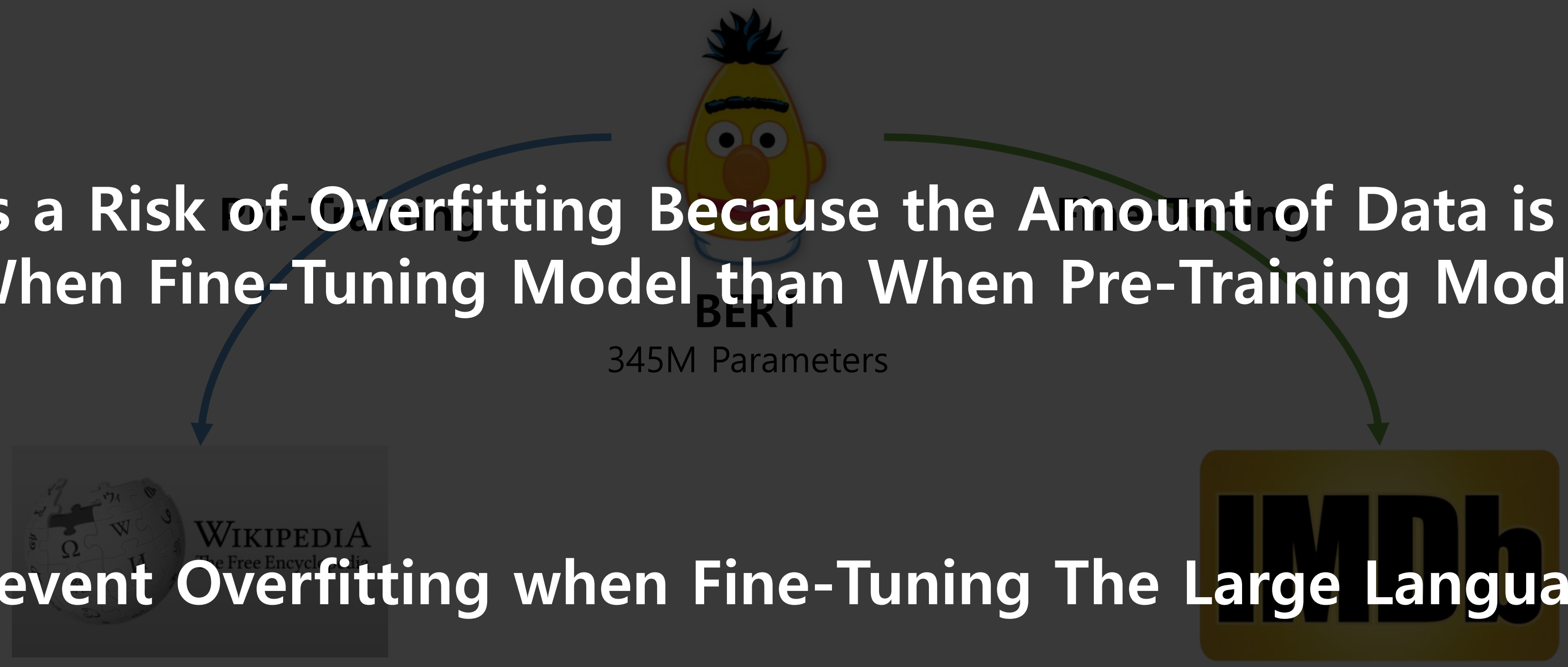
BERT

345M Parameters

Wikipedia + Book Corpus
Data Size: 20GB

GLUE Benchmark (WNLI)
Data Size: 98KB

<Complexity of Language Model>



There is a Risk of Overfitting Because the Amount of Data is Smaller When Fine-Tuning Model than When Pre-Training Model

How to Prevent Overfitting when Fine-Tuning The Large Language Model?

Wikipedia: 2,500M words
Book Corpus: 800M words

IMDB: 50,000 Text Data

SMART: Robust and Efficient Fine-Tuning for Pre-Trained Natural Language Models through Principled Regularized Optimization

Jiang et al., 2020, ACL

SMART: Robust and Efficient Fine-Tuning for Pre-Trained Natural Language Models through Principled Regularized Optimization

Jiang et al., 2020, ACL

Method

- Smoothness-Inducing Adversarial Regularization
- Bregman Proximal Point Optimization

<Overall Purpose of SMART>

SMART

Smoothness-Inducing Adversarial Regularization

"Control Model Capacity"

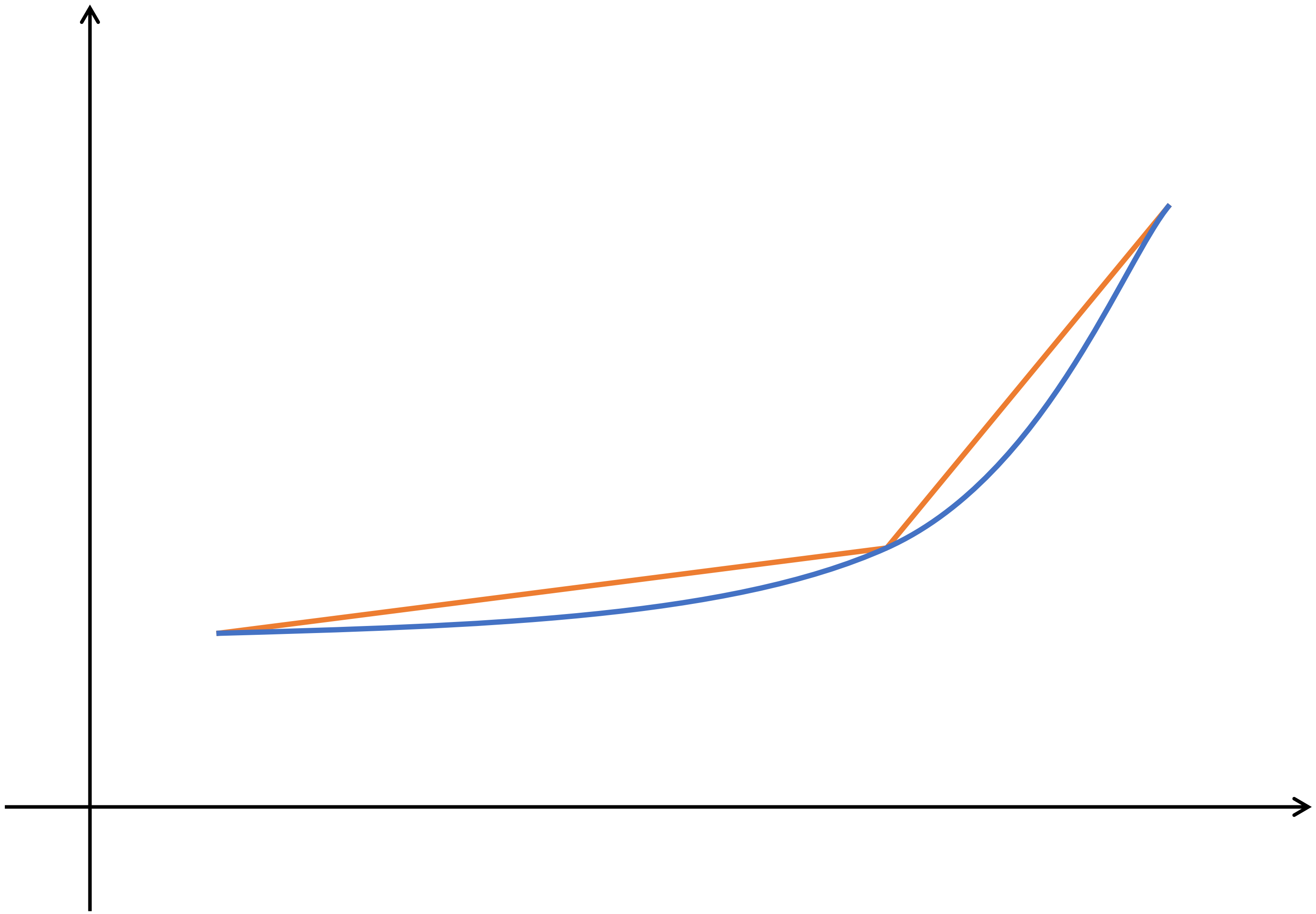
Bregman Proximal Point Optimization

"Prevent Aggressive Update"

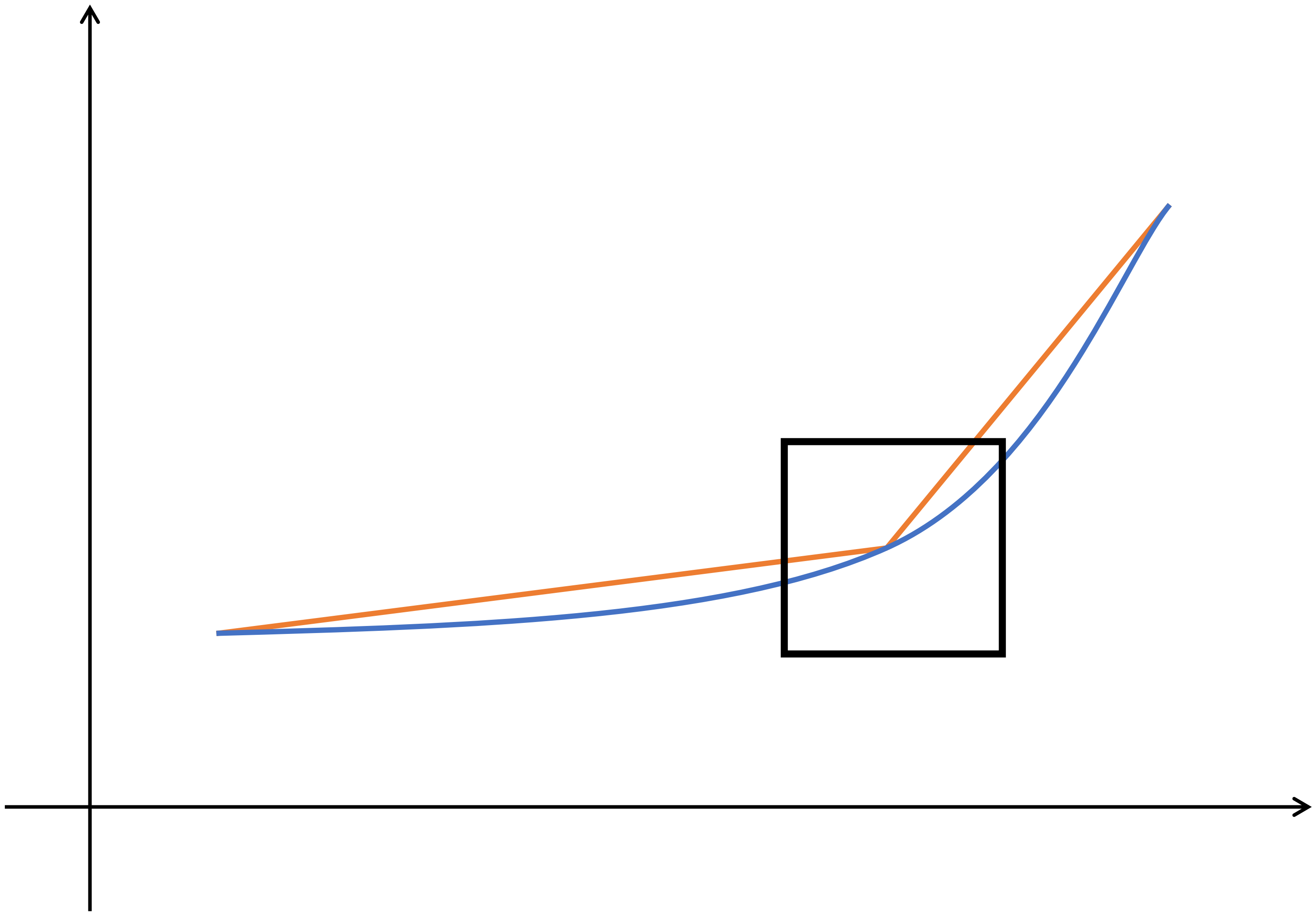
Method

-Smoothness-Inducing Adversarial Regularization

<Smoothness>



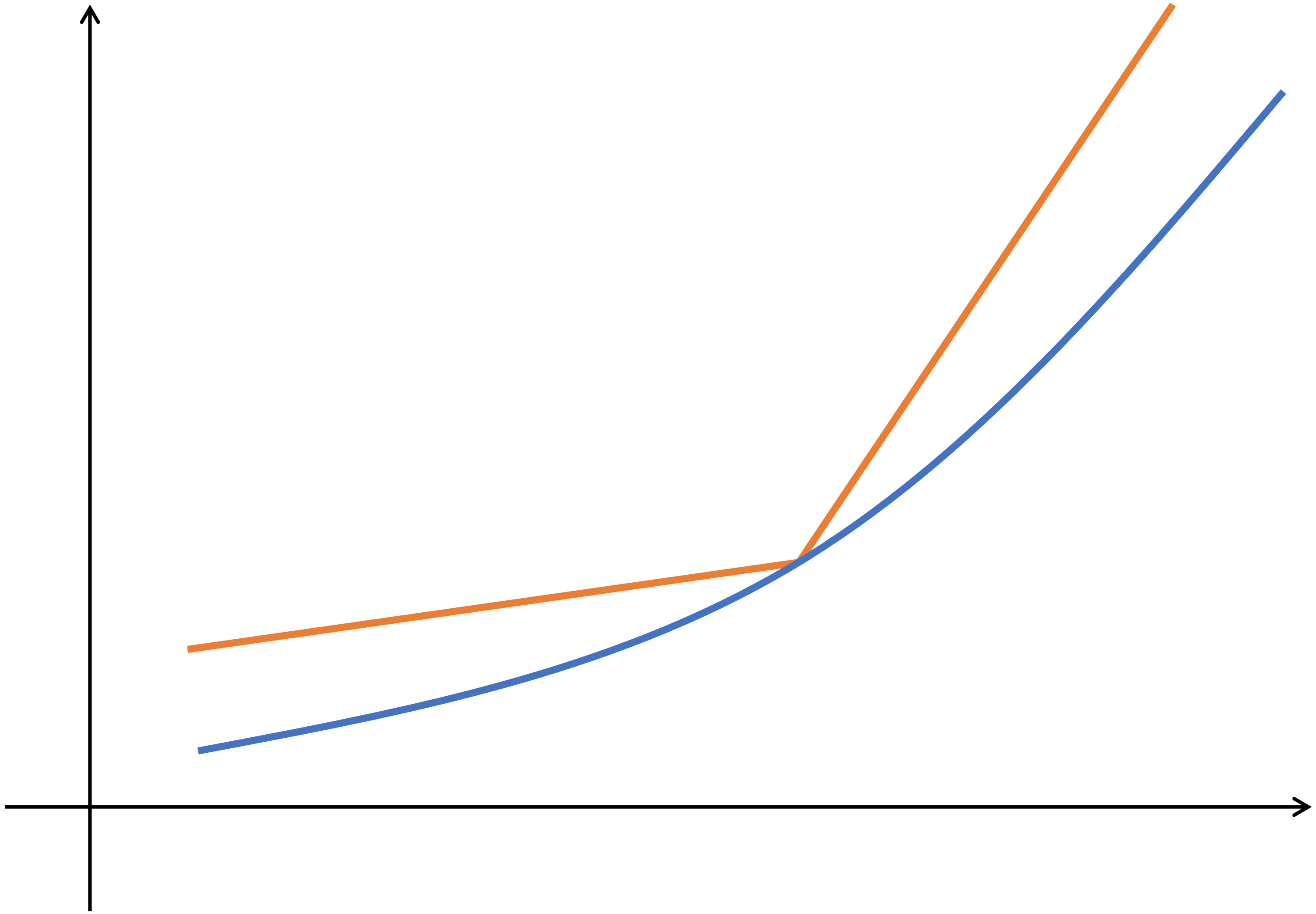
<Smoothness>



Method

-Smoothness-Inducing Adversarial Regularization

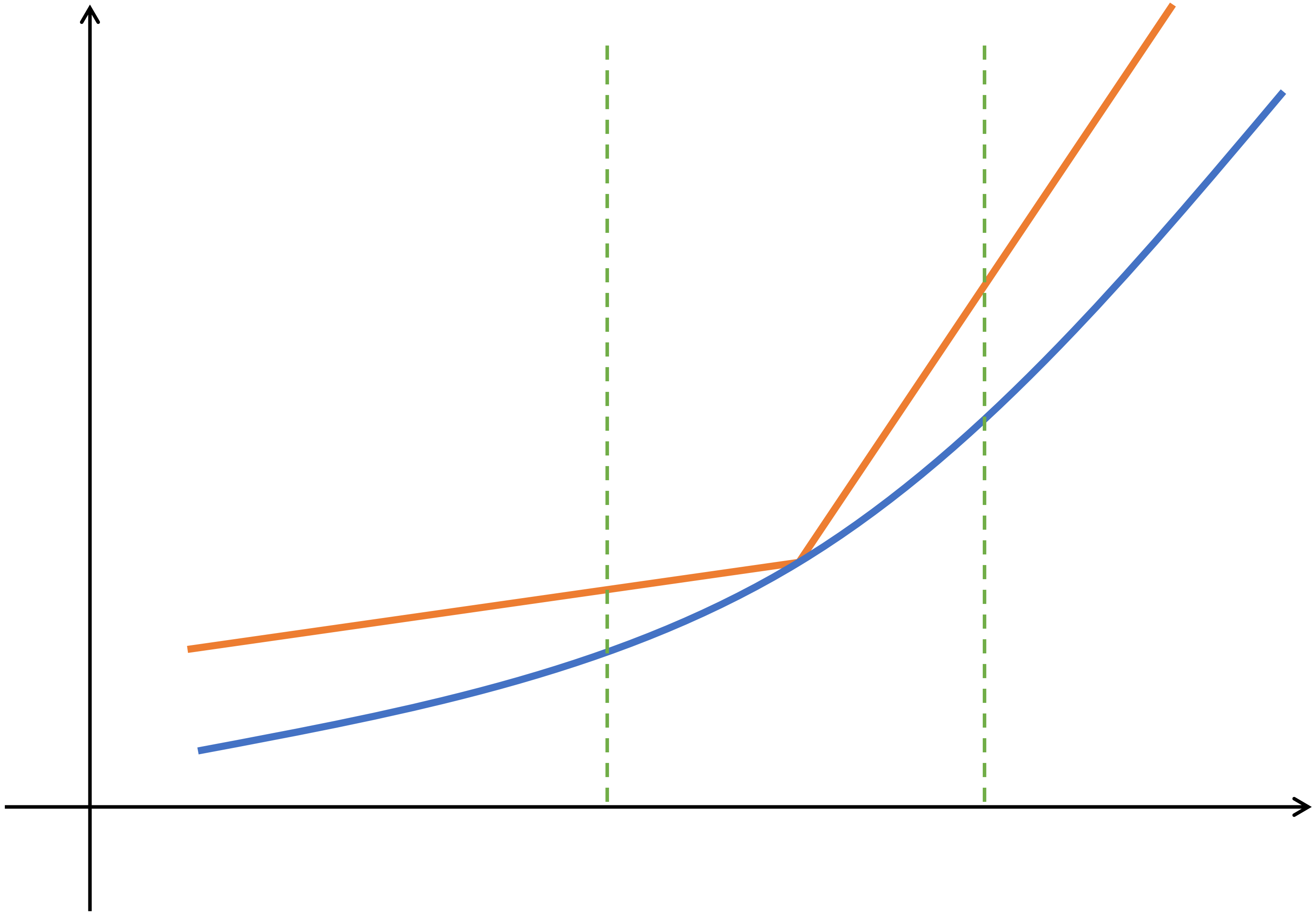
<Smoothness>



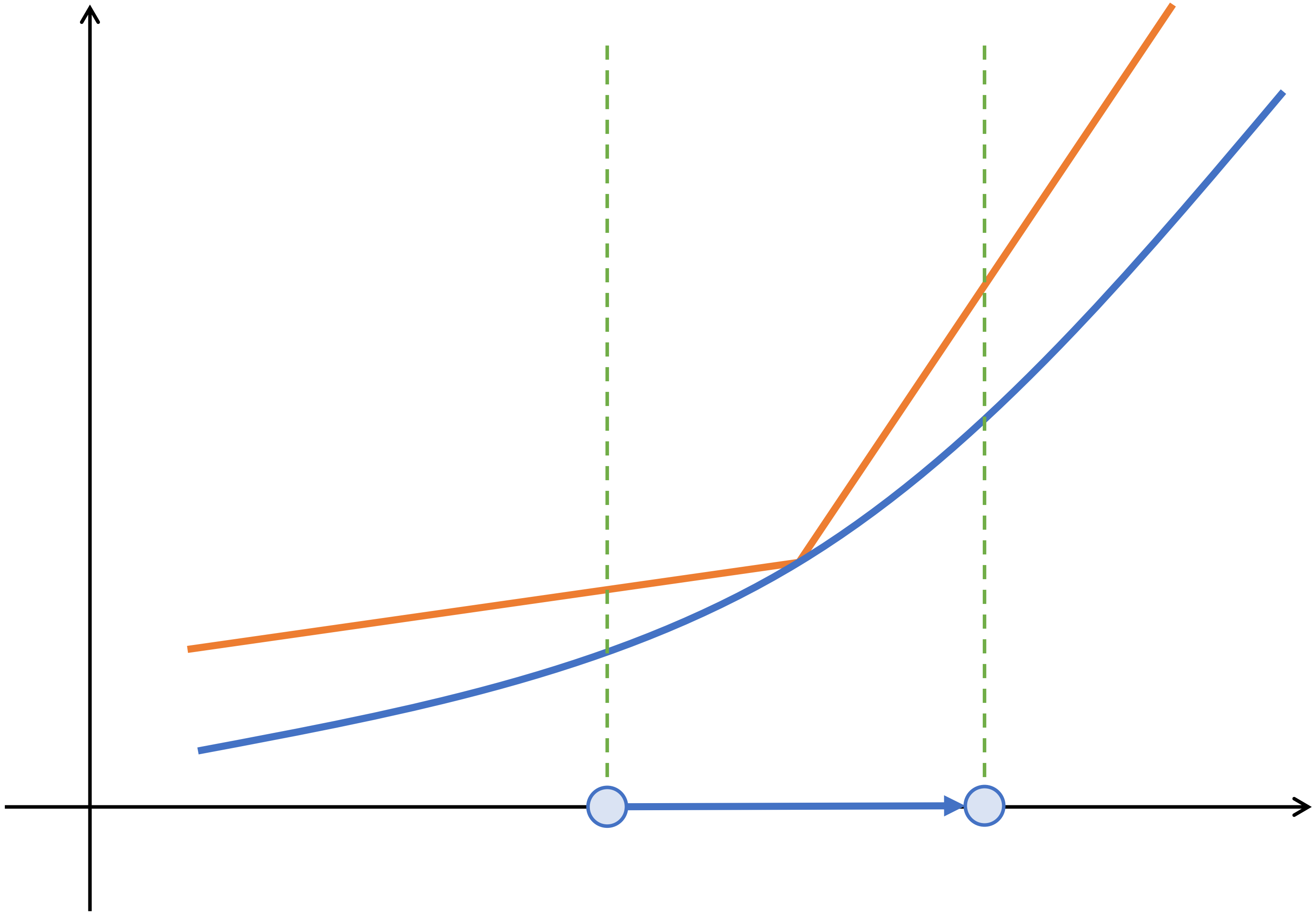
Method

-Smoothness-Inducing Adversarial Regularization

<Smoothness>



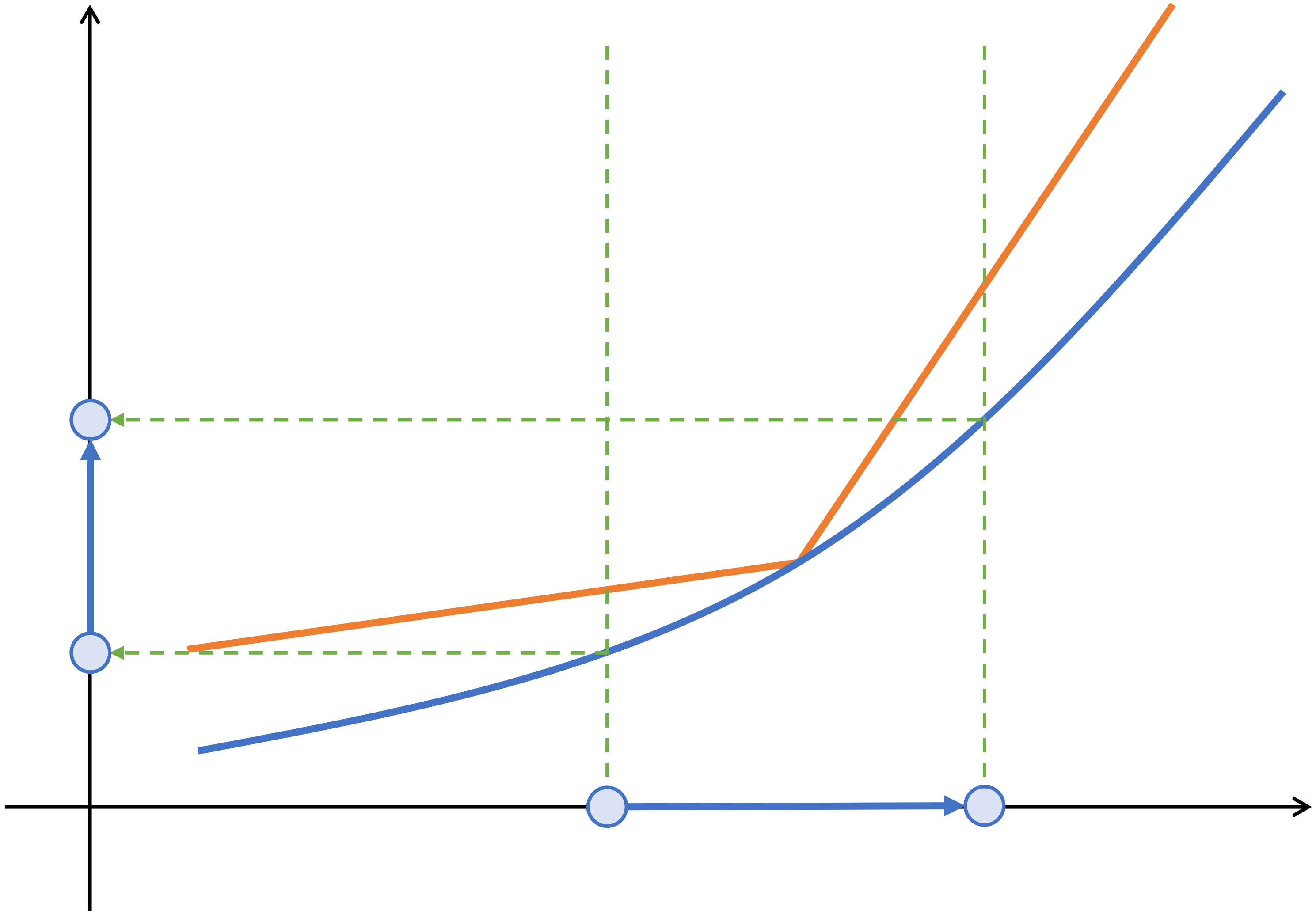
<Smoothness>



Method

-Smoothness-Inducing Adversarial Regularization

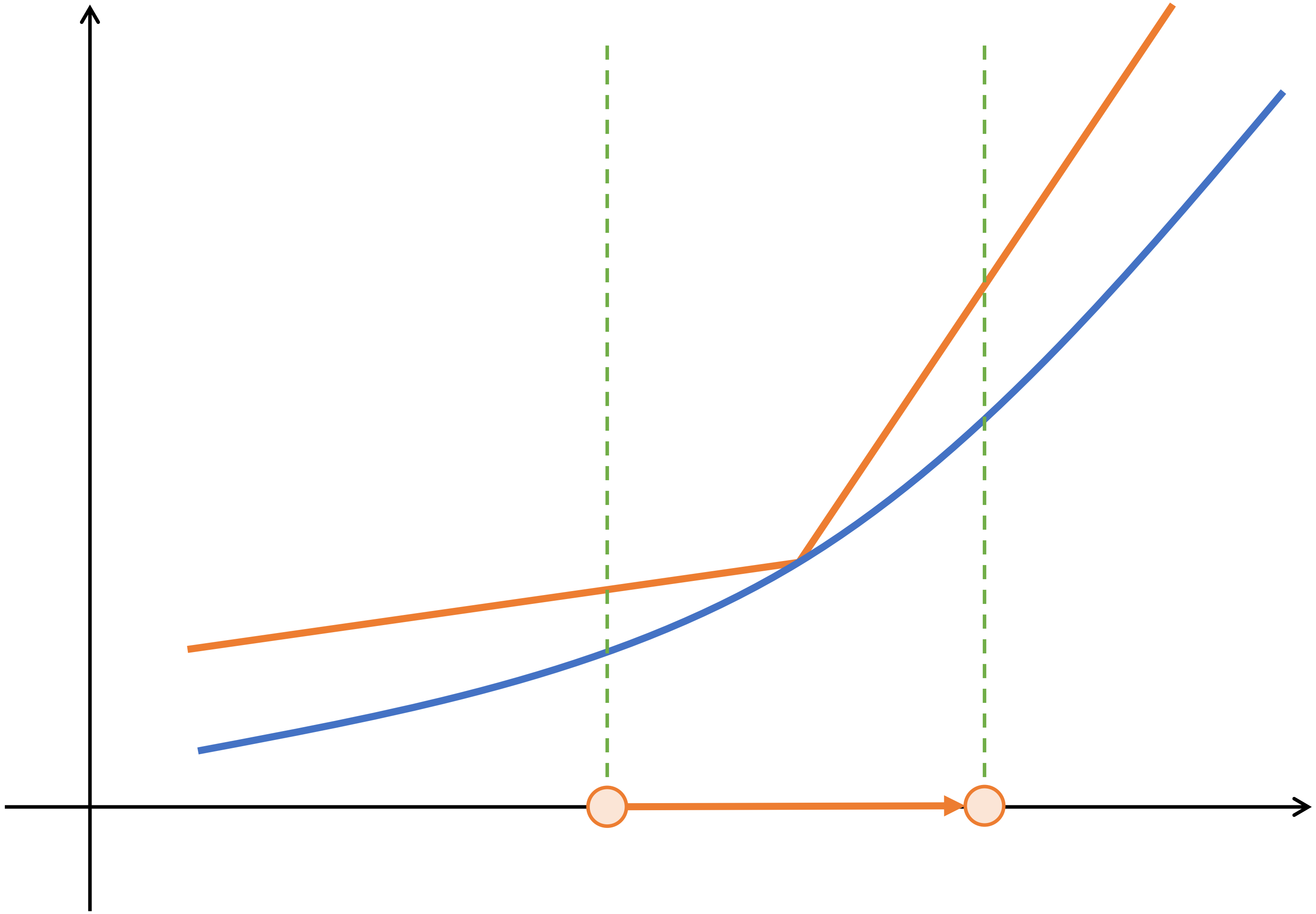
<Smoothness>



Method

-Smoothness-Inducing Adversarial Regularization

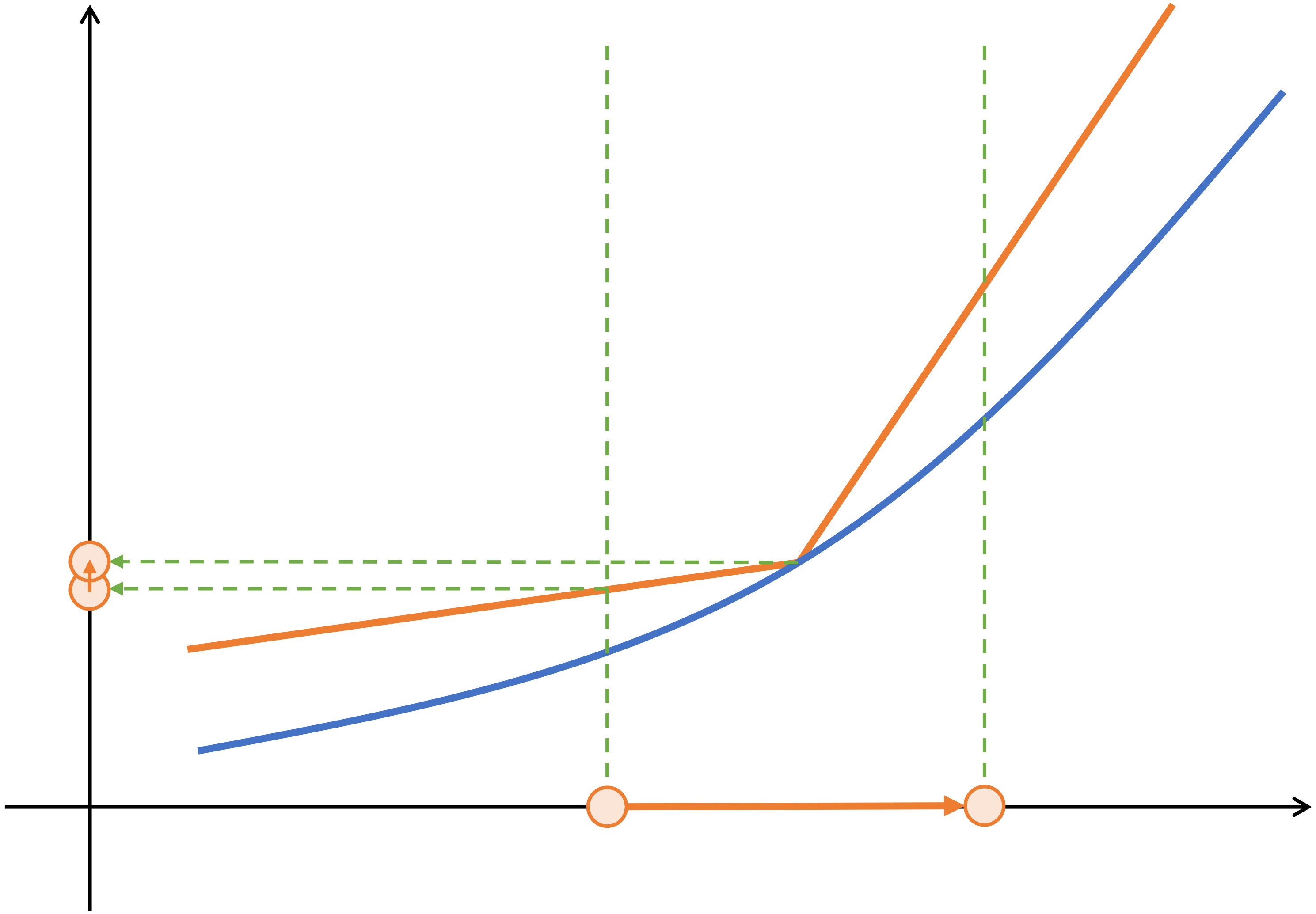
<Smoothness>



Method

-Smoothness-Inducing Adversarial Regularization

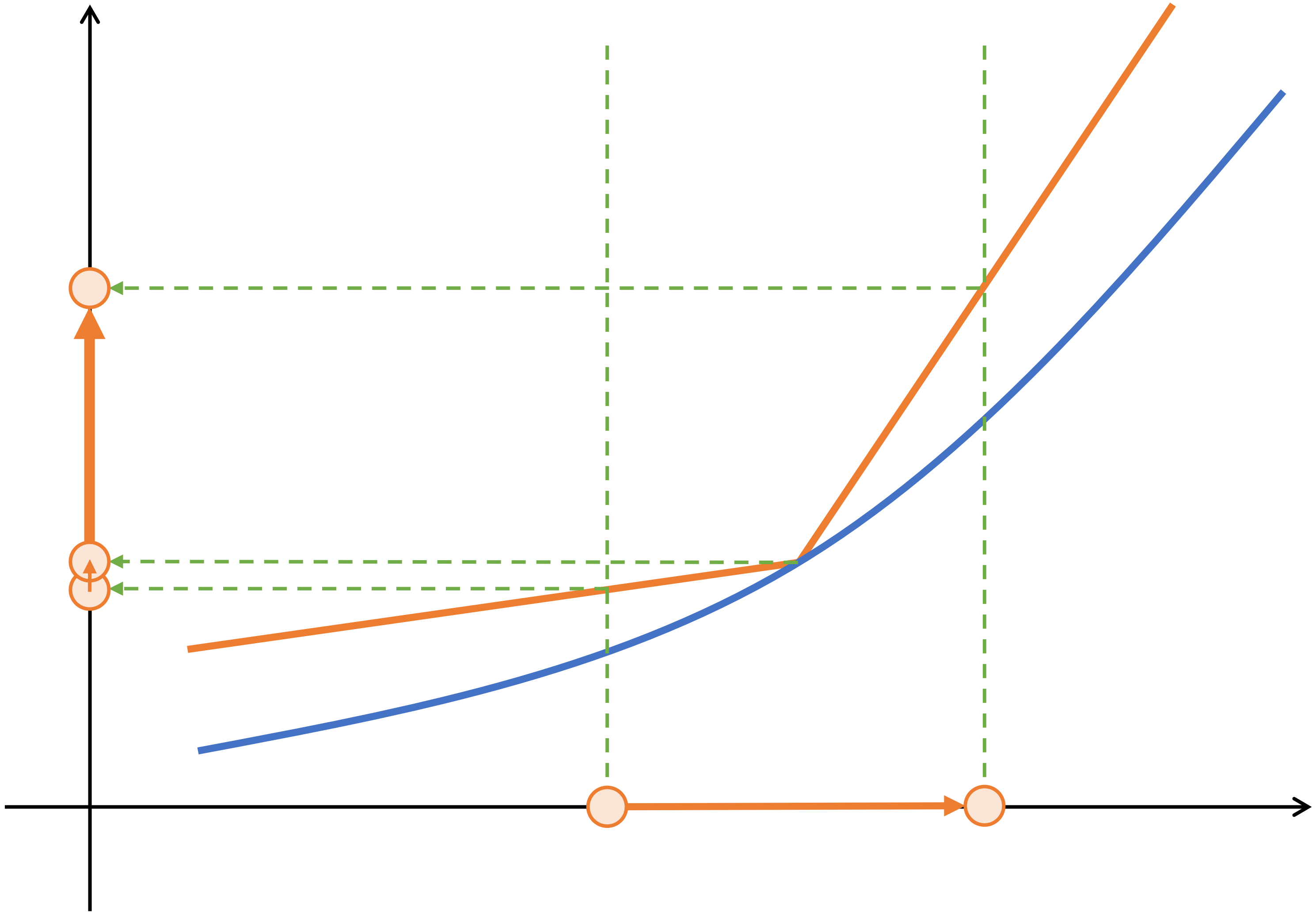
<Smoothness>



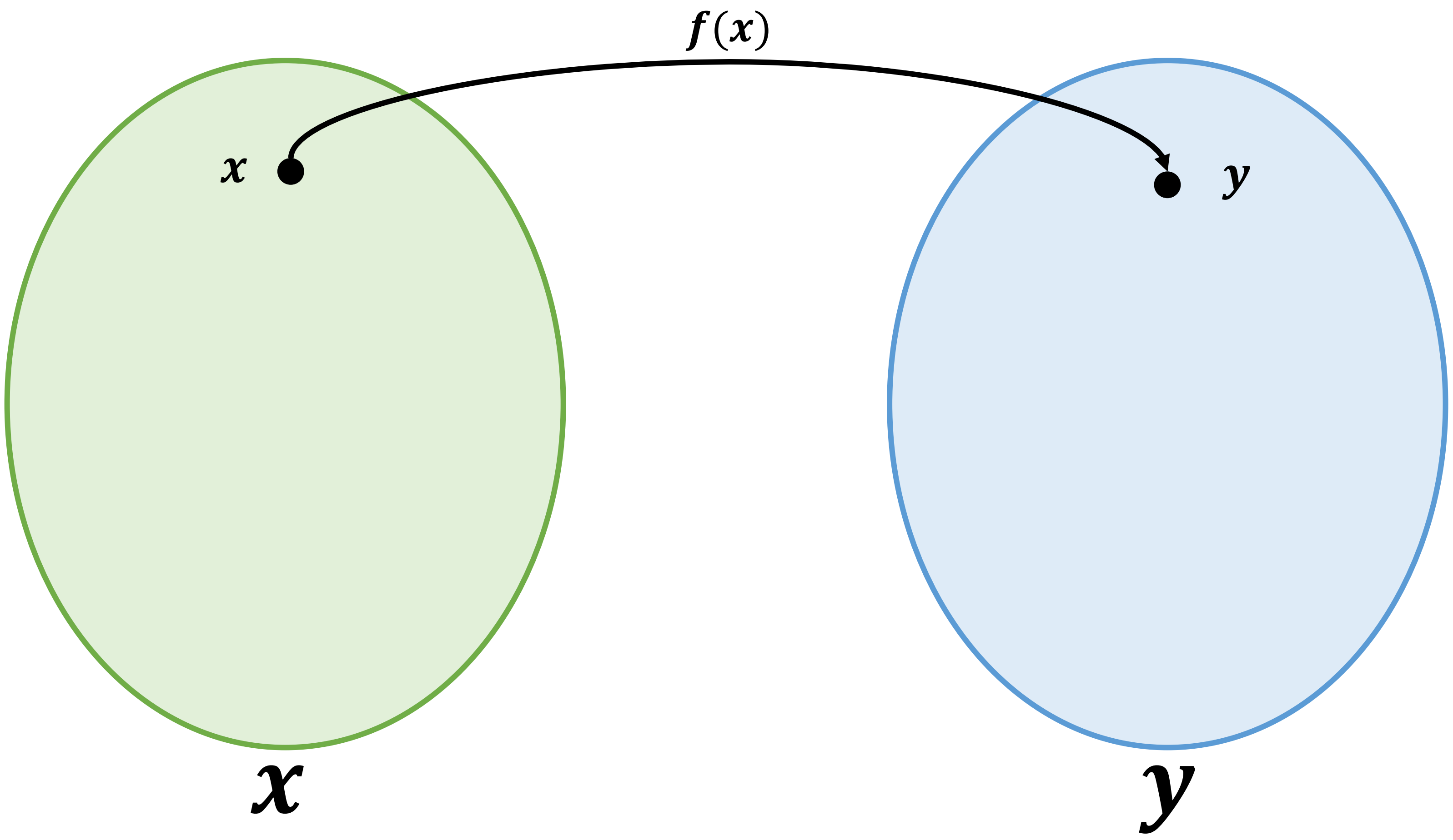
Method

-Smoothness-Inducing Adversarial Regularization

<Smoothness>



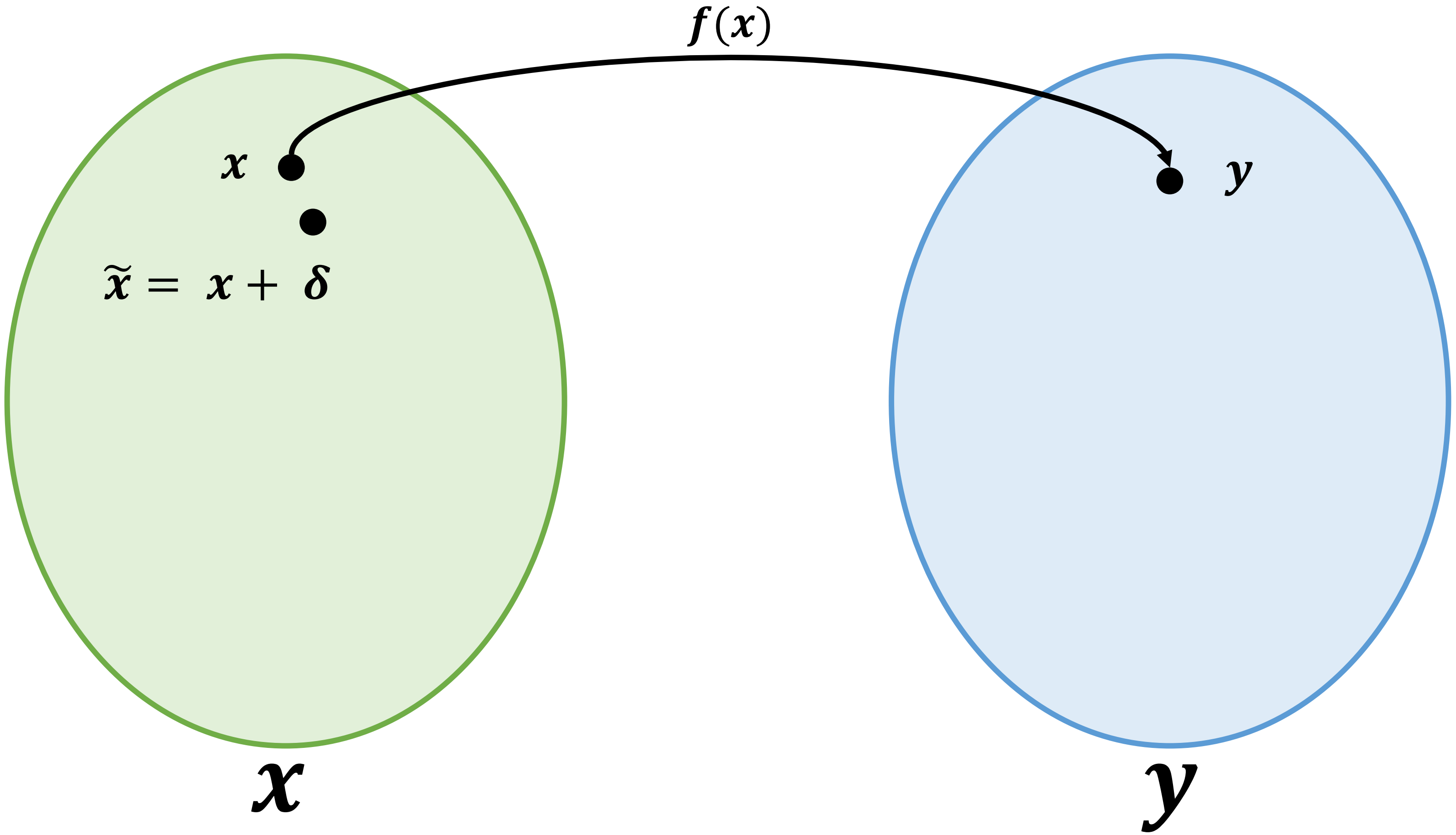
<Adversarial Regularization>



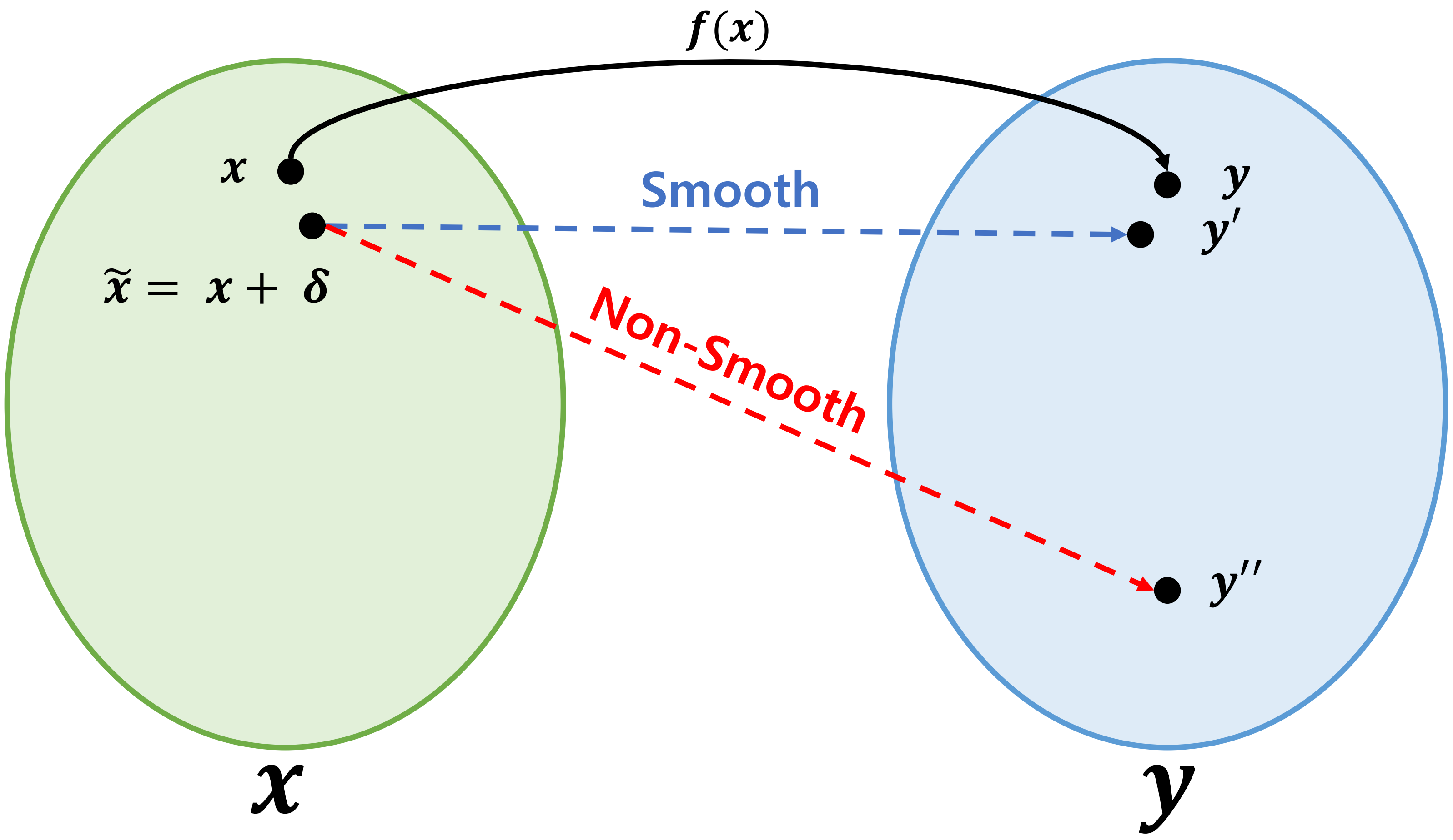
Method

-Smoothness-Inducing Adversarial Regularization

<Adversarial Regularization>



<Adversarial Regularization>



Method
-Smoothness-Inducing Adversarial Regularization

<Notations>

x_i : $Embedding$

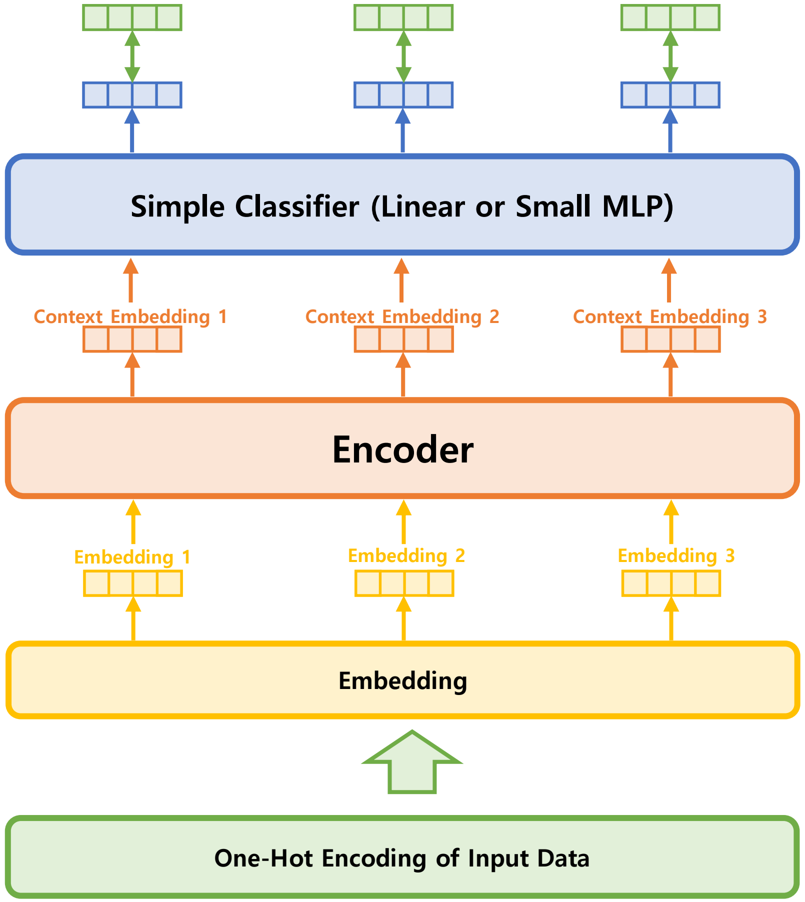
$f(x_i;\theta)$: $Language\ Model\ (Encoder)\ as\ Function$

θ: $All\ Learnable\ Parameter\ in\ Language\ Model$

y_i: $Label$

δ: $Perturbation$

$\tilde{x}_i = x_i + \delta$: $Adversarial\ Example$



<Adversarial Regularization>

$$\min_{\theta} \mathcal{F}(\theta) = \mathcal{L}(\theta) + \lambda_s \mathcal{R}_s(\theta)$$

$$\mathcal{L}(\theta) = \frac{1}{n}\sum_{i=1}^{n} \ell(f(x_i;\theta), y_i)$$

$$\mathcal{R}_s(\theta) = \frac{1}{n}\sum_{i=1}^{n} \max_{\|\tilde{x}_i - x_i\|_p \leq \epsilon} \ell_s\big(f(\tilde{x}_i;\theta), f(x_i;\theta)\big)$$

$$\ell_s(P,Q) = \mathcal{D}_{KL}(P||Q) + \mathcal{D}_{KL}(Q||P)$$

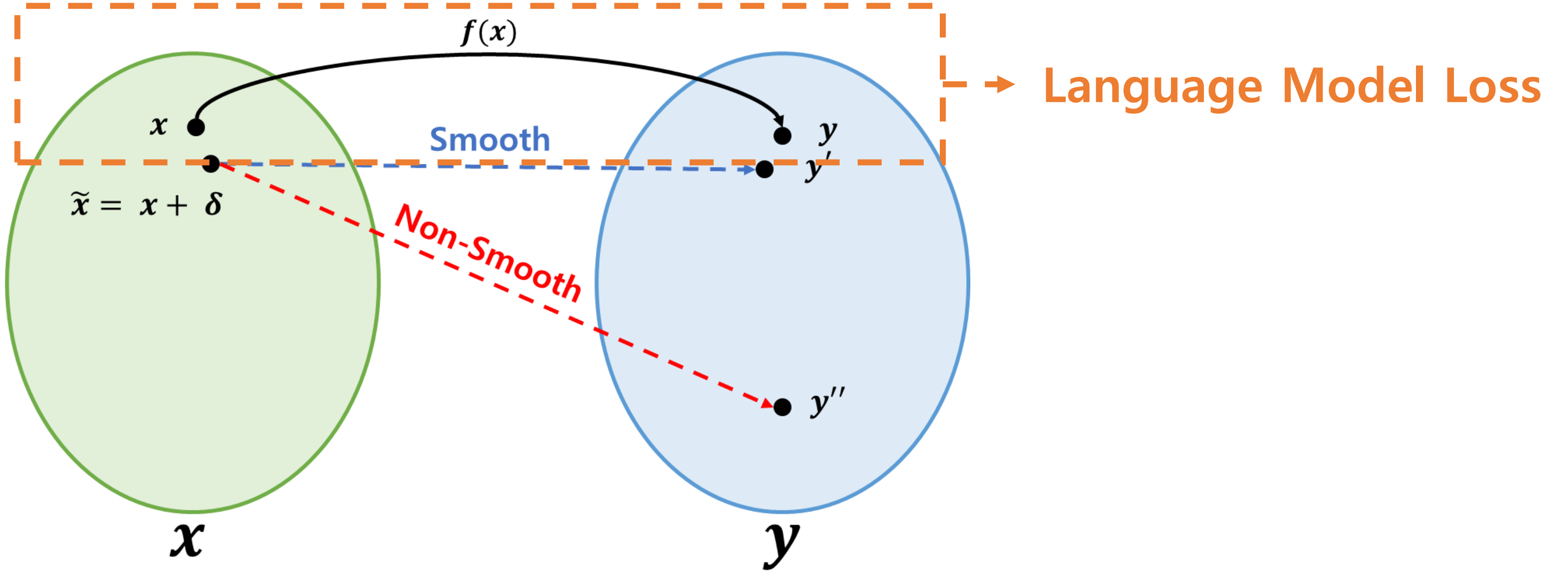
<Adversarial Regularization>

$$\min_{\Theta} \mathcal{F}(\theta) = \mathcal{L}(\theta) + \lambda_s \mathcal{R}_s(\theta)$$

$$\boldsymbol{\mathcal{L}(\theta) = \frac{1}{n}\sum_{i=1}^{n} \ell(f(x_i;\theta), y_i)}$$

$$\mathcal{R}_s(\theta) = \frac{1}{n}\sum_{i=1}^{n} \max_{\|\tilde{x}_i - x_i\|_p \le \epsilon} \ell_s\big(f(\tilde{x}_i;\theta), f(x_i;\theta)\big)$$

$$\ell_s(P,Q) = \mathcal{D}_{KL}(P||Q) + \mathcal{D}_{KL}(Q||P)$$



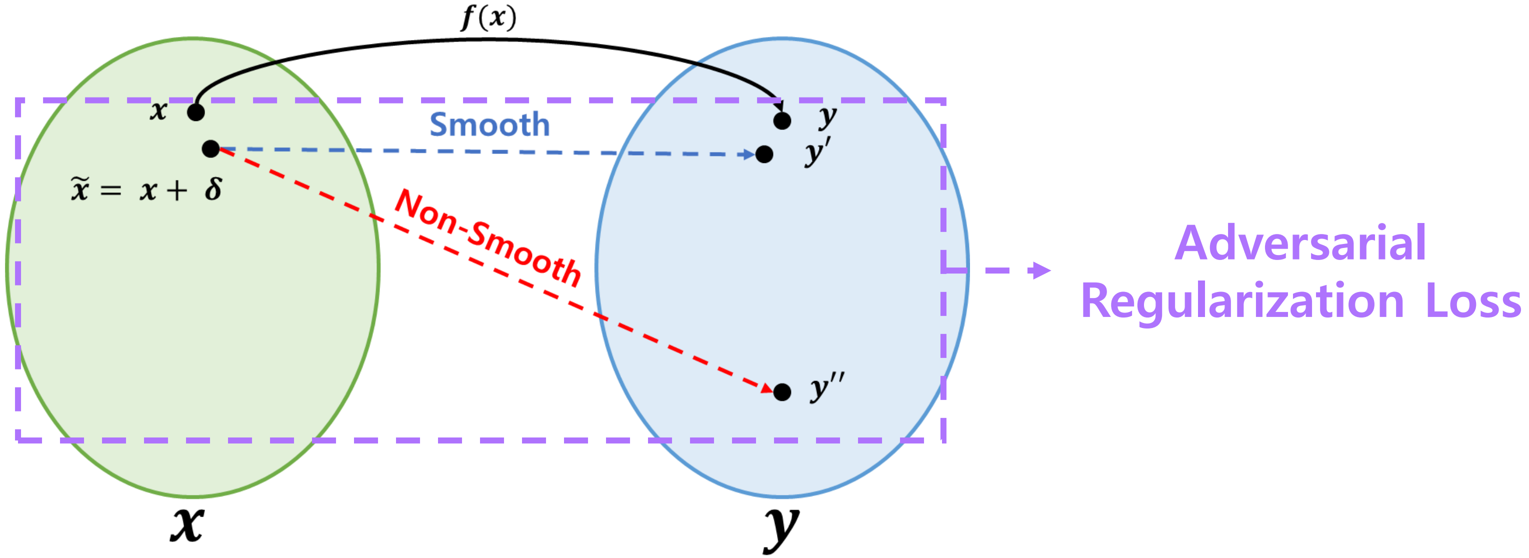
<Adversarial Regularization>

$$\min_{\Theta} \mathcal{F}(\theta) = \mathcal{L}(\theta) + \lambda_s \mathcal{R}_s(\theta)$$

$$\mathcal{L}(\theta) = \frac{1}{n}\sum_{i=1}^{n} \ell(f(x_i;\theta), y_i)$$

$$\boldsymbol{\mathcal{R}_s(\theta) = \frac{1}{n}\sum_{i=1}^{n} \max_{\|\tilde{x}_i - x_i\|_p \le \epsilon} \ell_s\big(f(\tilde{x}_i;\theta), f(x_i;\theta)\big)}$$

$$\ell_s(P,Q) = \mathcal{D}_{KL}(P||Q) + \mathcal{D}_{KL}(Q||P)$$



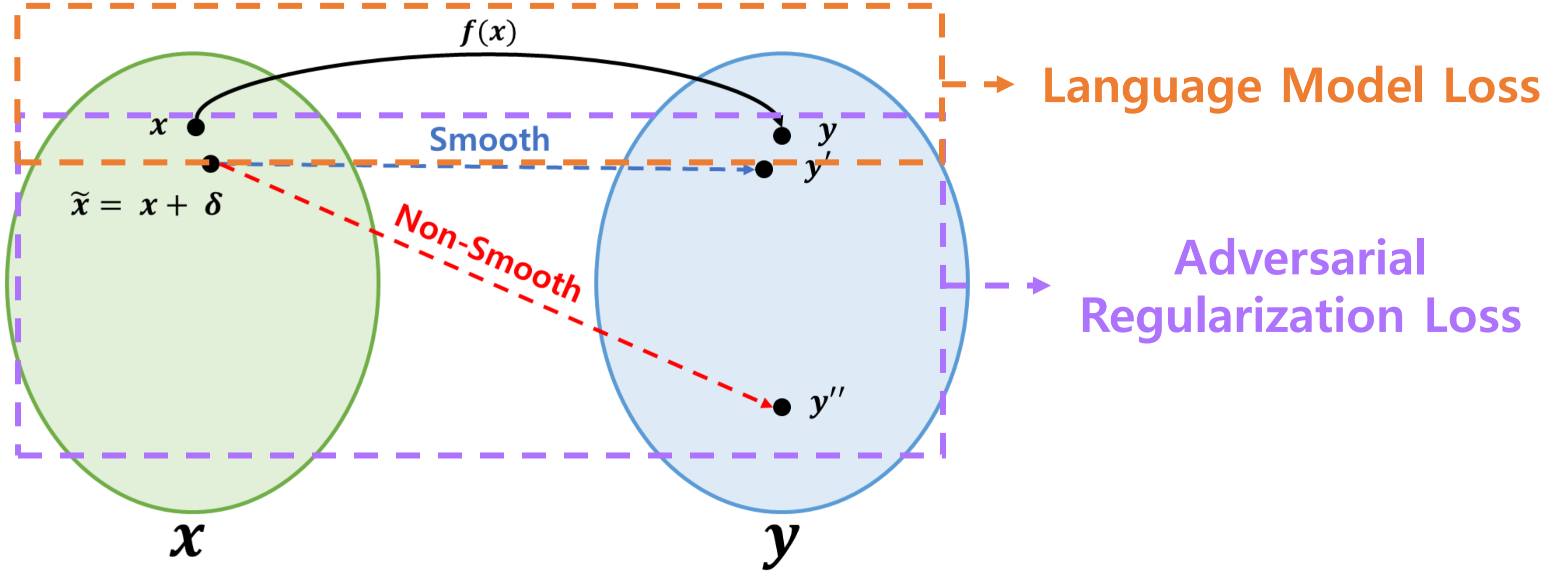
<Adversarial Regularization>

$$\min_{\theta} \mathcal{F}(\theta) = \mathcal{L}(\theta) + \lambda_s \mathcal{R}_s(\theta)$$

$$\mathcal{L}(\theta) = \frac{1}{n}\sum_{i=1}^{n} \ell(f(x_i;\theta), y_i)$$

$$\mathcal{R}_s(\theta) = \frac{1}{n}\sum_{i=1}^{n} \max_{\|\tilde{x}_i - x_i\|_p \le \epsilon} \ell_s\big(f(\tilde{x}_i;\theta), f(x_i;\theta)\big)$$

$$\ell_s(P,Q) = \mathcal{D}_{KL}(P||Q) + \mathcal{D}_{KL}(Q||P)$$



Method

-Smoothness-Inducing Adversarial Regularization

<SMART VS FreeLB>

$$\min_{\theta} \mathcal{F}(\theta) = \mathcal{L}(\theta) + \lambda_s \mathcal{R}_s(\theta)$$

$$\mathcal{L}(\theta) = \frac{1}{n}\sum_{i=1}^{n} \ell(f(x_i;\theta), y_i)$$

$$\mathcal{R}_s(\theta) = \frac{1}{n}\sum_{i=1}^{n} \max_{\|\tilde{x}_i - x_i\|_p \le \epsilon} \boldsymbol{\ell_s\left(f(\tilde{x}_i;\theta), f(x_i;\theta)\right)}$$

$$\boldsymbol{\ell_s(P, Q) = \mathcal{D}_{KL}(P||Q) + \mathcal{D}_{KL}(Q||P)}$$

<SMART>

$$\min_{\theta} \mathbb{E}_{(Z,y)\sim\mathcal{D}}\left[\max_{||\delta|| \le \varepsilon} \boldsymbol{L(f_{\theta}(X+\delta), y)}\right]$$

$$\delta_{t+1} = \Pi_{||\delta||_F \le \varepsilon}(\delta_t + \alpha g(\delta_t) \,/\, ||g(\delta_t)||_F)$$

<FreeLB>

"Adversarial Training to **Probability**"

Clean	0.1	0.7	0.1	0.1
Adversarial	0.2	0.6	0.1	0.1

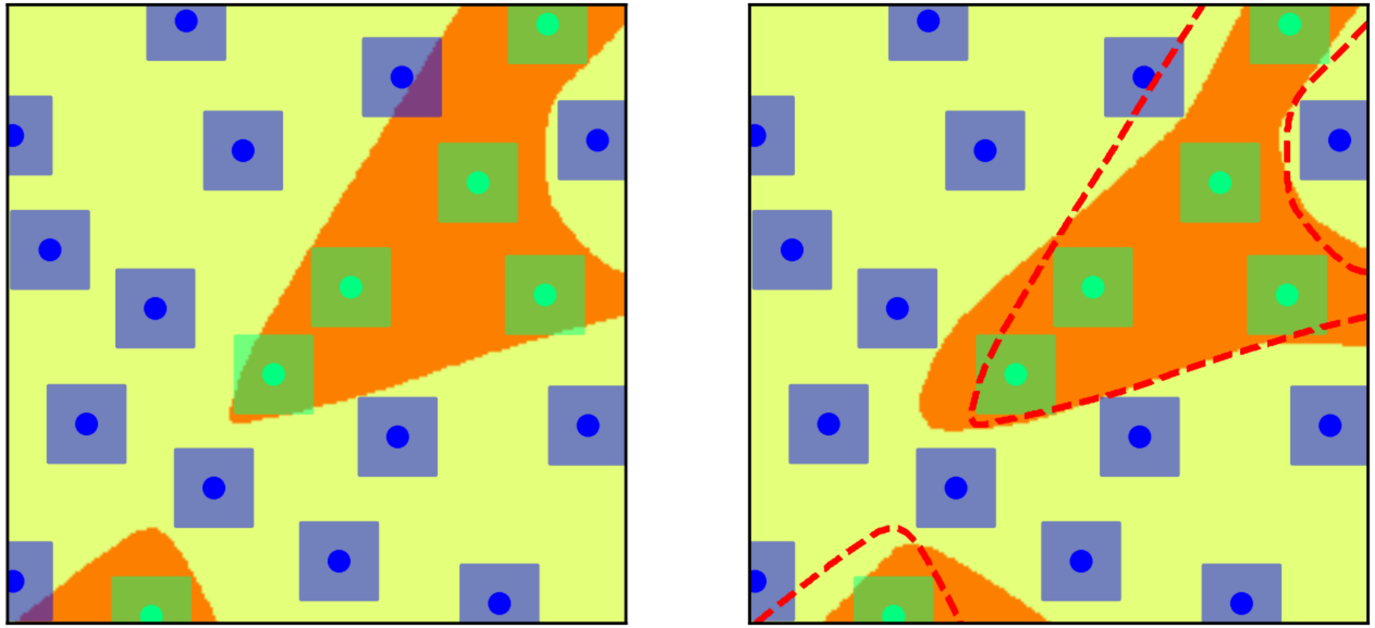
"Adversarial Training to **Label**"

Label	0	1	0	0
Adversarial	0.2	0.6	0.1	0.1

Method

-Smoothness-Inducing Adversarial Regularization

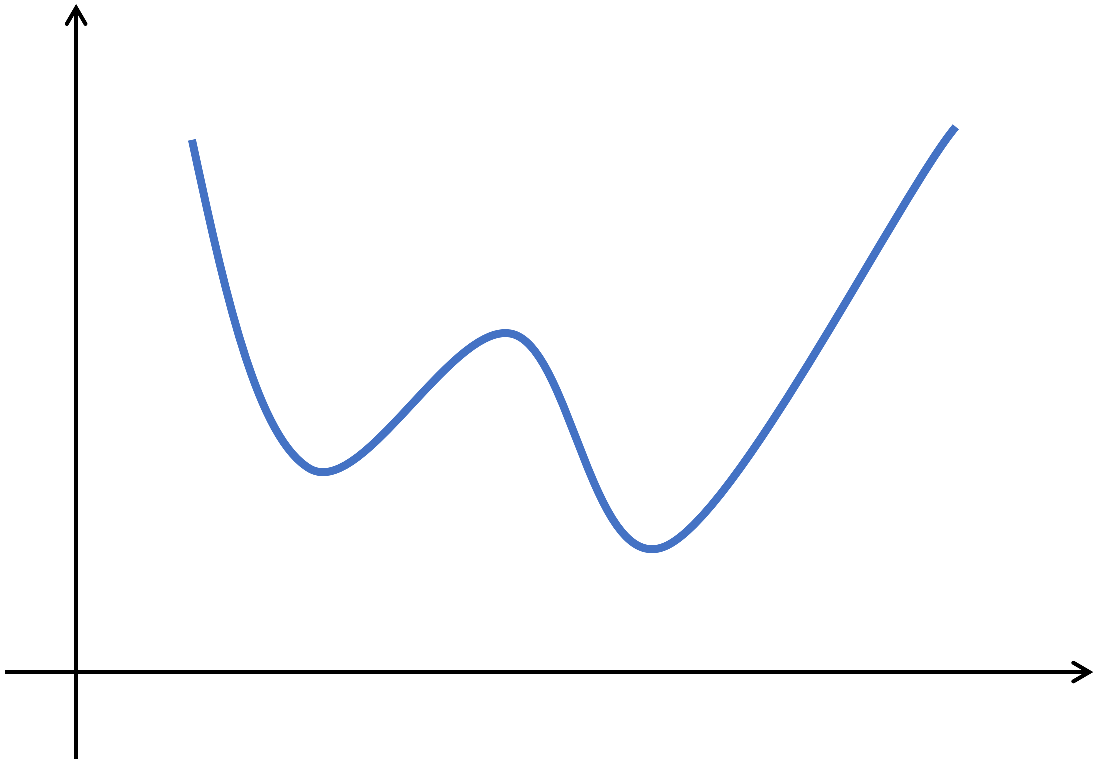
<Adversarial Regularization>



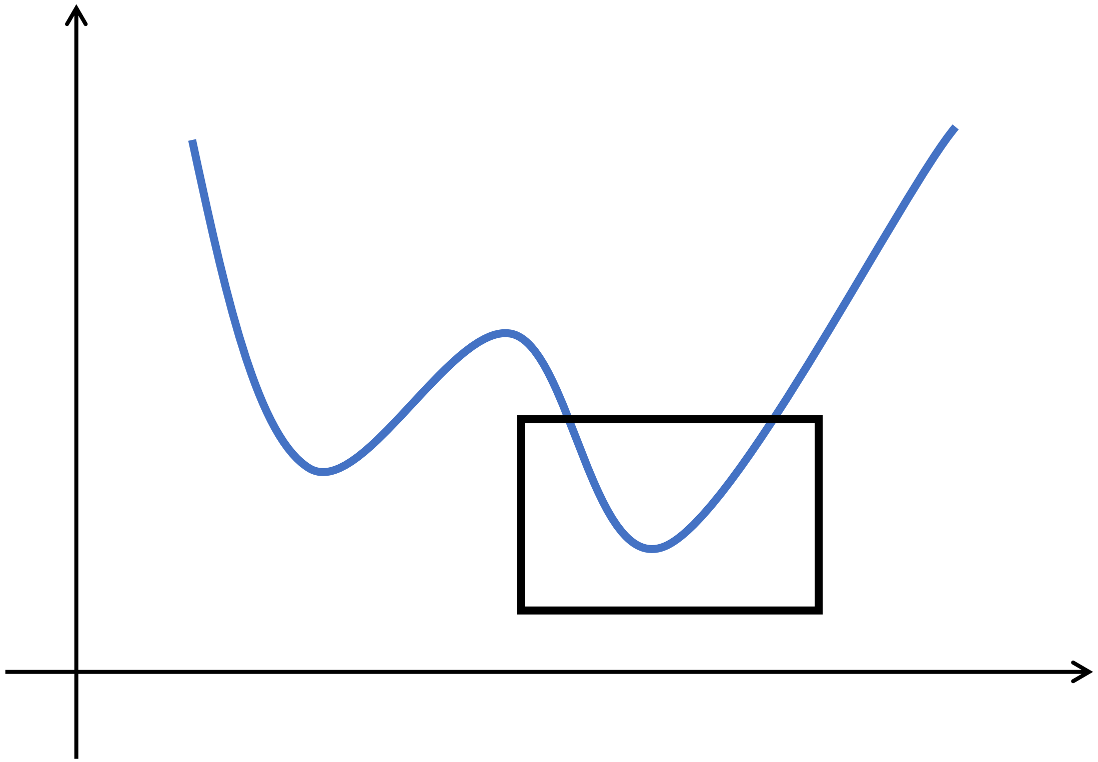
(a) (b)

Decision Boundaries Learned without (a) and with (b) Smoothness-Inducing Adversarial Regularization

<Gradient Descent>



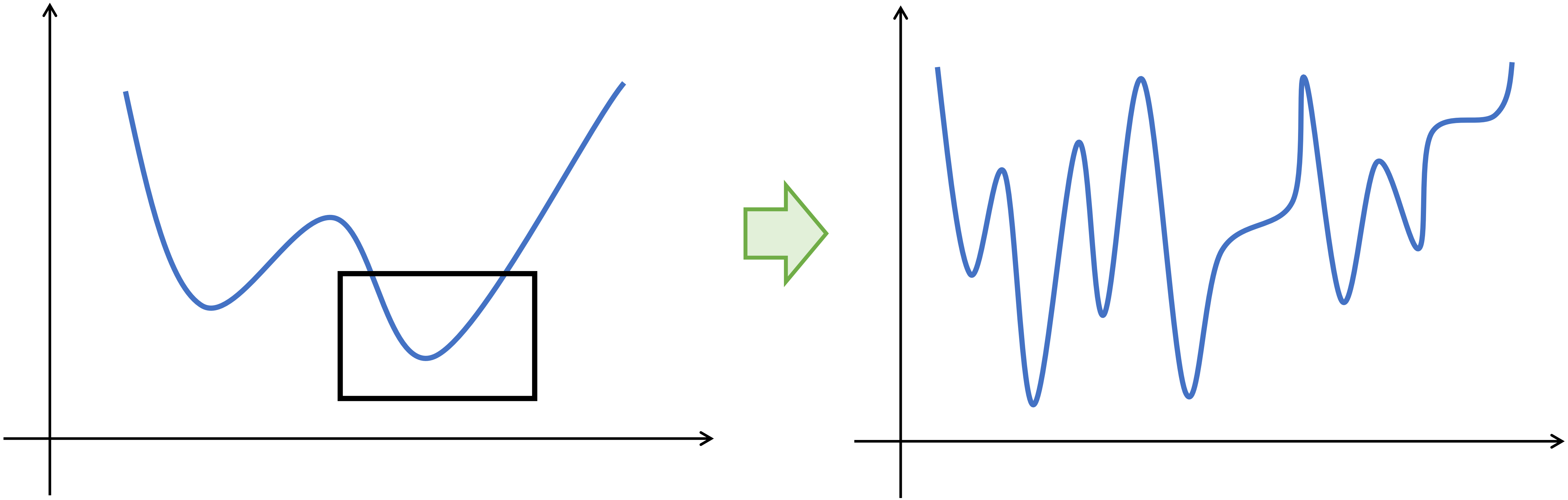
<Gradient Descent>



Method

-Bregman Proximal Point Optimization

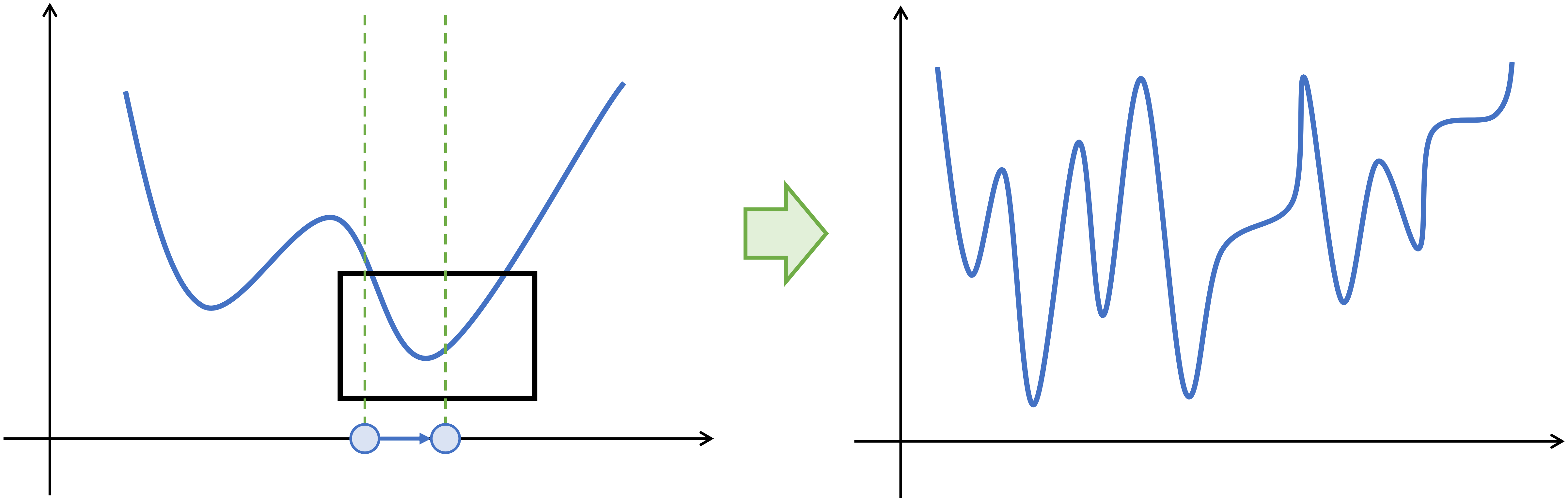
<Gradient Descent>



Method

-Bregman Proximal Point Optimization

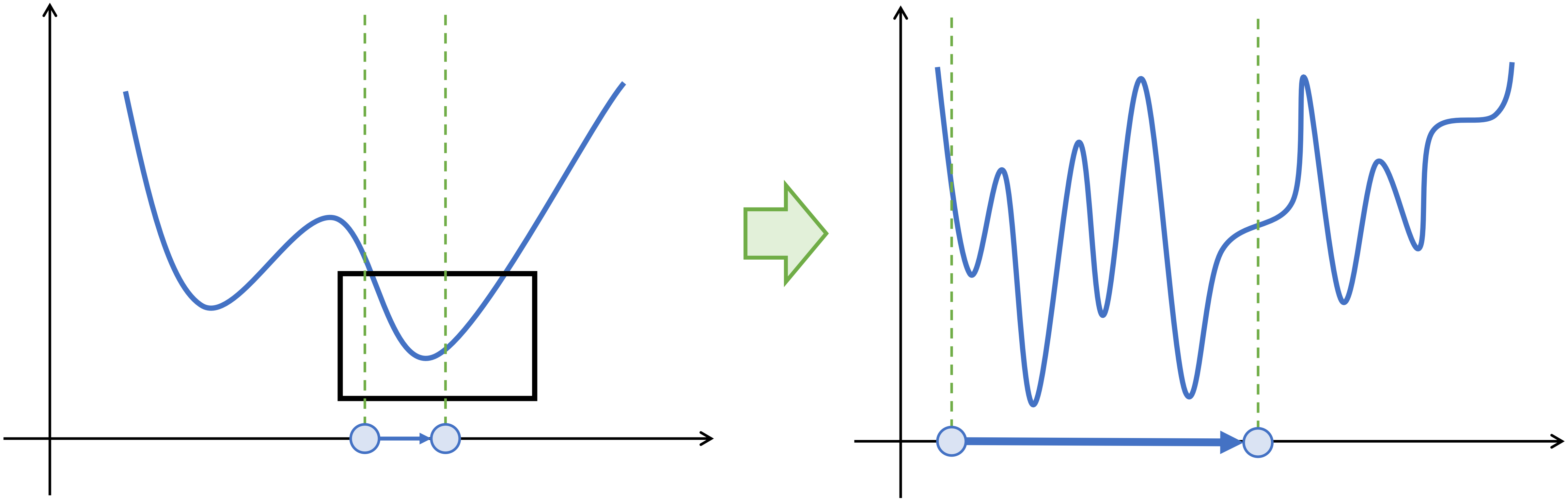
<Gradient Descent>



Method

-Bregman Proximal Point Optimization

<Gradient Descent>



<Bregman Proximal Point Optimization>

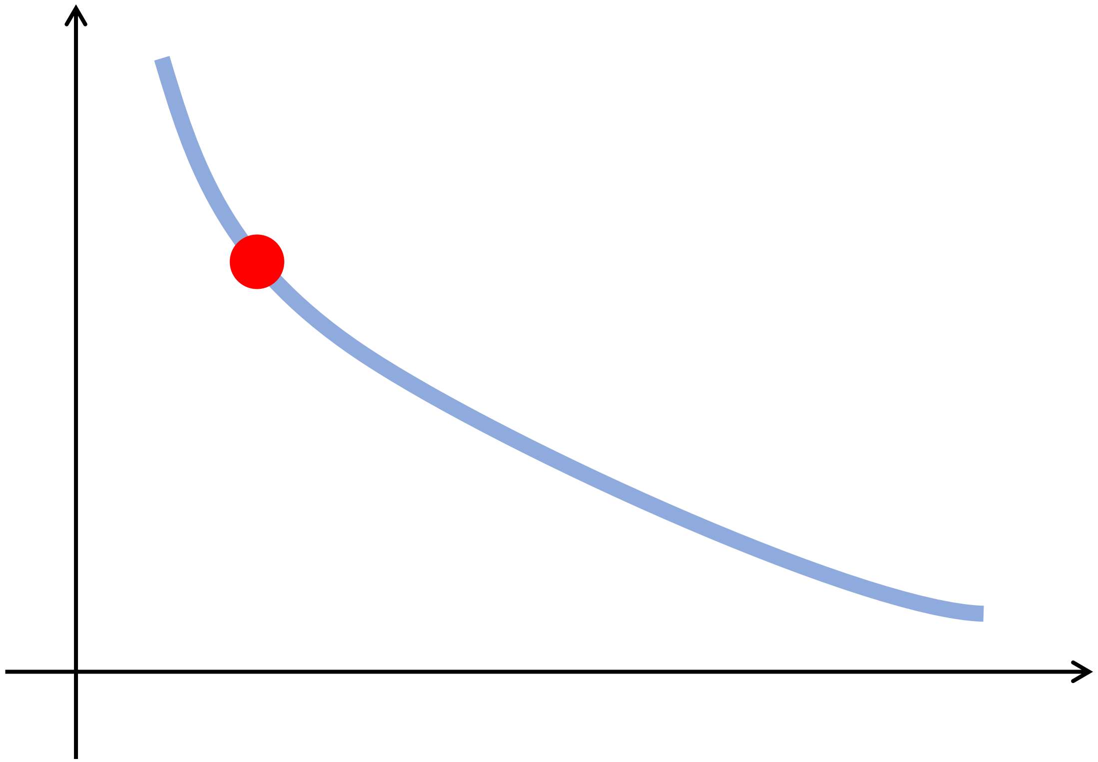
$$f(\cdot;\theta_0): Pre\text{-}Trained\ Model, Initialization$$
$$\theta_{t+1} = arg\min_{\theta} \mathcal{F}(\theta) + \mu\mathcal{D}_{\text{Breg}}(\theta,\theta_t)$$
$$\mathcal{D}_{\text{Breg}}(\theta,\theta_t) = \frac{1}{n}\sum_{i=1}^{n} \ell_s(f(x_i;\theta), f(x_i;\theta_t))$$

<Bregman Proximal Point Optimization>

$$f(\cdot;\theta_0): Pre\text{-}Trained\ Model, Initialization$$

$$\theta_{t+1} = arg\min_{\theta} \mathcal{F}(\theta) + \mu\mathcal{D}_{\text{Breg}}(\theta,\theta_t)$$

$$\mathcal{D}_{\text{Breg}}(\theta,\theta_t) = \frac{1}{n}\sum_{i=1}^{n} \ell_s(\boldsymbol{f(x_i;\theta)}, f(x_i;\theta_t))$$

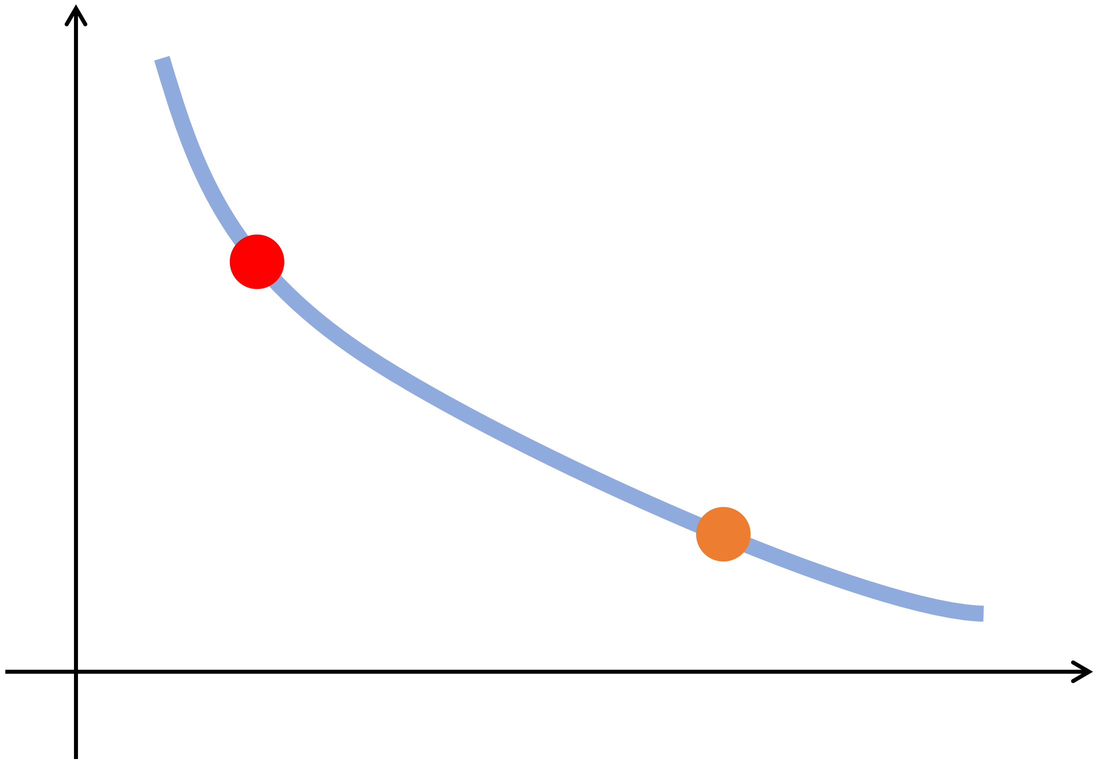


<Bregman Proximal Point Optimization>

$$f(\cdot;\theta_0): Pre\text{-}Trained\ Model, Initialization$$

$$\theta_{t+1} = arg\min_{\theta} \mathcal{F}(\theta) + \mu\mathcal{D}_{\text{Breg}}(\theta, \theta_t)$$

$$\mathcal{D}_{\text{Breg}}(\theta, \theta_t) = \frac{1}{n}\sum_{i=1}^{n} \ell_s(\boldsymbol{f(x_i;\theta)}, \boldsymbol{f(x_i;\theta_t)})$$

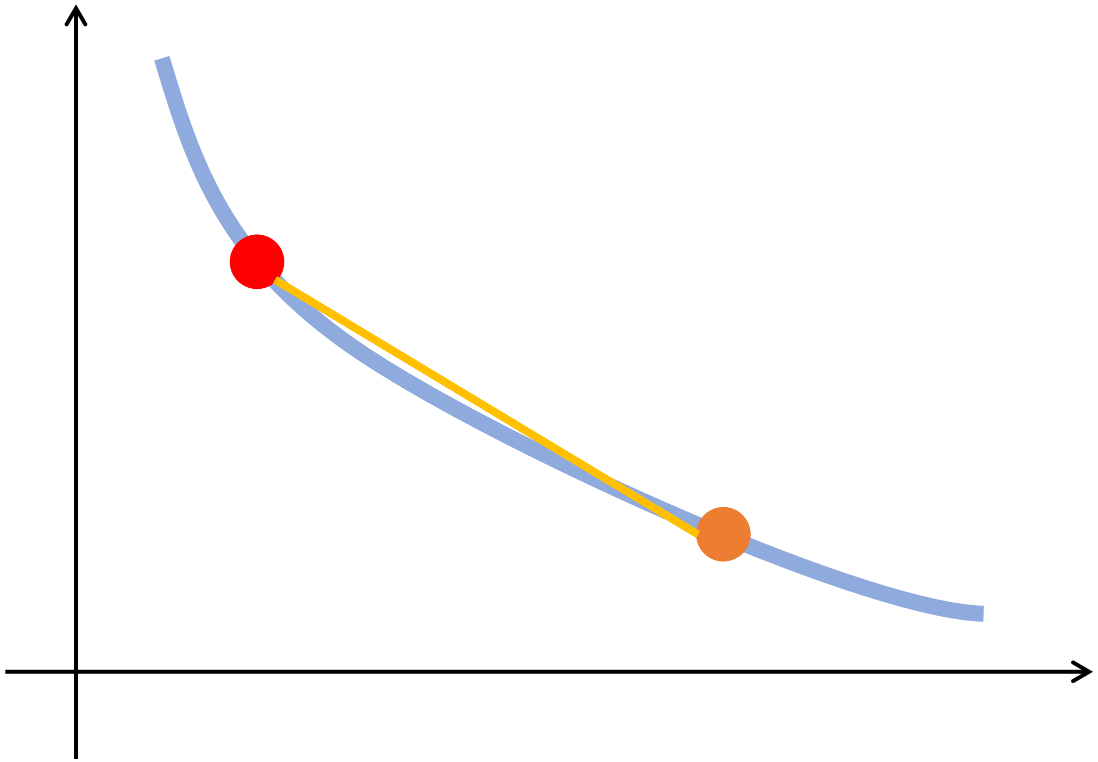


<Bregman Proximal Point Optimization>

$$f(\cdot;\theta_0): \textit{Pre-Trained Model, Initialization}$$

$$\theta_{t+1} = arg\min_{\theta} \mathcal{F}(\theta) + \mu\mathcal{D}_{\mathrm{Breg}}(\theta, \theta_t)$$

$$\boldsymbol{\mathcal{D}_{Breg}(\theta, \theta_t)} = \frac{1}{n}\sum_{i=1}^{n} \ell_s(\boldsymbol{f(x_i;\theta)}, \boldsymbol{f(x_i;\theta_t)})$$

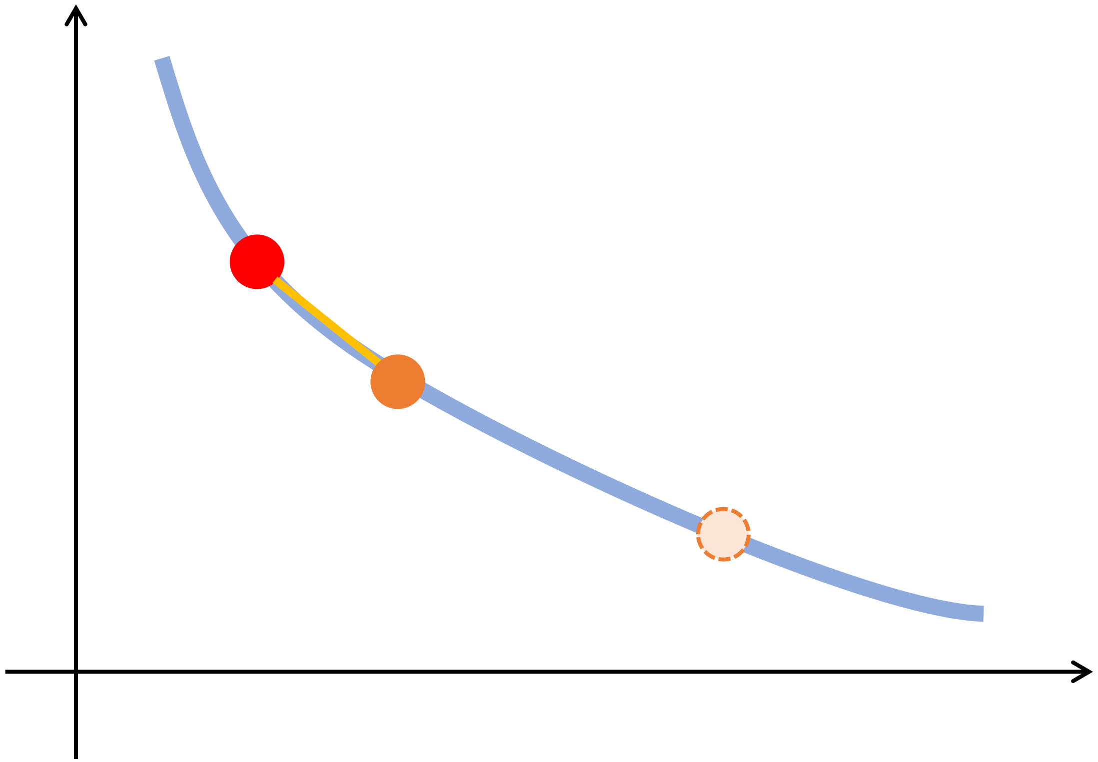


<Bregman Proximal Point Optimization>

$$f(\cdot;\theta_0): \textit{Pre-Trained Model, Initialization}$$

$$\theta_{t+1} = \arg\min_{\theta} \mathcal{F}(\theta) + \mu \mathcal{D}_{\mathbf{Breg}}(\boldsymbol{\theta}, \boldsymbol{\theta}_t)$$

$$\mathcal{D}_{\mathbf{Breg}}(\boldsymbol{\theta}, \boldsymbol{\theta}_t) = \frac{1}{n}\sum_{i=1}^{n} \ell_s(\boldsymbol{f}(\boldsymbol{x}_i;\boldsymbol{\theta}), \boldsymbol{f}(\boldsymbol{x}_i;\boldsymbol{\theta}_t))$$



<Momentum Bregman Proximal Point Optimization>

$$f(\cdot;\theta_0): \textit{Pre-Trained Model, Initialization}$$

$$\theta_{t+1} = arg\min_{\theta} \mathcal{F}(\theta) + \mu\mathcal{D}_{\mathrm{Breg}}(\theta, \boldsymbol{\widetilde{\theta}_t})$$

$$\mathcal{D}_{\mathrm{Breg}}(\theta, \theta_t) = \frac{1}{n}\sum_{i=1}^{n} \ell_s(f(x_i;\theta), f(x_i;\theta_t))$$

$$\boldsymbol{\widetilde{\theta}_t = (1-\beta)\theta_t + \beta\widetilde{\theta}_{t-1}}$$

Experiments

- GLUE Benchmark
- Ablation Study

<GLUE Benchmark>

Model	MNLI-m/mm Acc	QQP ACC/F1	RTE Acc	QNLI Acc	MRPC Acc/F1	CoLA Mcc	SST Mcc	STS-B P/S Corr
				$\text{BERT}_{\text{BASE}}$				
BERT(Devlin et al., 2019)	84.4/-	-	-	88.4	-/86.7	-	92.7	-
$\text{BERT}_{\text{ReImp}}$	84.5/84.4	90.9/88.3	63.5	91.1	84.1/89.0	54.7	92.9	89.2/88.8
$\text{SMART}_{\text{BERT}}$	**85.6/86.0**	**91.5/88.5**	**71.2**	**91.7**	**87.7/91.3**	**59.1**	**93**	**90.0/89.4**
				$\text{RoBERTa}_{\text{LARGE}}$				
RoBERTa(Liu et al., 2019)	90.2/-	92.2/-	86.6	94.7	-/90.9	68	96.4	92.4/-
PGD(Zhu et al., 2020)	90.5/-	92.5/-	87.4	94.9	-/90.9	69.7	96.4	92.4/-
FreeAT(Zhu et al., 2020)	90.0/-	92.5/-	86.7	94.7	-/90.7	68.8	96.1	92.4/-
FreeLB(Zhu et al., 2020)	90.6/-	**92.6/-**	88.1	95	-/91.4	**71.1**	96.7	92.7/-
$\text{SMART}_{\text{RoBERTa}}$	**91.1/91.3**	92.4/89.8	**92**	**95.6**	**89.2/92.1**	70.6	**96.9**	**92.8/92.6**

<Main Result on GLUE Development Set>

<GLUE Benchmark>

Model	MNLI-m/mm Acc	QQP ACC/F1	RTE Acc	QNLI Acc	MRPC Acc/F1	CoLA Mcc	SST Mcc	STS-B P/S Corr
$\mathbf{BERT_{BASE}}$								
BERT(Devlin et al., 2019)	84.4/-	-	-	88.4	-/86.7	-	92.7	-
$\mathbf{BERT_{ReImp}}$	84.5/84.4	90.9/88.3	63.5	91.1	84.1/89.0	54.7	92.9	89.2/88.8
$\mathbf{SMART_{BERT}}$	**85.6/86.0**	**91.5/88.5**	**71.2**	**91.7**	**87.7/91.3**	**59.1**	**93**	**90.0/89.4**
$\mathbf{RoBERTa_{LARGE}}$								
RoBERTa(Liu et al., 2019)	90.2/-	92.2/-	86.6	94.7	-/90.9	68	96.4	92.4/-
PGD(Zhu et al., 2020)	90.5/-	92.5/-	87.4	94.9	-/90.9	69.7	96.4	92.4/-
FreeAT(Zhu et al., 2020)	90.0/-	92.5/-	86.7	94.7	-/90.7	68.8	96.1	92.4/-
FreeLB(Zhu et al., 2020)	90.6/-	**92.6/-**	88.1	95	-/91.4	**71.1**	96.7	92.7/-
$\mathbf{SMART_{RoBERTa}}$	**91.1/91.3**	92.4/89.8	**92**	**95.6**	**89.2/92.1**	70.6	**96.9**	**92.8/92.6**

<Main Result on GLUE Development Set>

Adversarial Training for NLU

- GLUE Benchmark

<GLUE Benchmark>

Model	MNLI-m/mm Acc	QQP ACC/F1	RTE Acc	QNLI Acc	MRPC Acc/F1	CoLA Mcc	SST Mcc	STS-B P/S Corr
				$\mathbf{BERT_{BASE}}$				
BERT(Devlin et al., 2019)	84.4/-	-	-	88.4	-/86.7	-	92.7	-
$\mathbf{BERT_{ReImp}}$	84.5/84.4	90.9/88.3	63.5	91.1	84.1/89.0	54.7	92.9	89.2/88.8
$\mathbf{SMART_{BERT}}$	**85.6/86.0**	**91.5/88.5**	**71.2**	**91.7**	**87.7/91.3**	**59.1**	**93**	**90.0/89.4**
				$\mathbf{RoBERTa_{LARGE}}$				
RoBERTa(Liu et al., 2019)	90.2/-	92.2/-	86.6	94.7	-/90.9	68	96.4	92.4/-
PGD(Zhu et al., 2020)	90.5/-	92.5/-	87.4	94.9	-/90.9	69.7	96.4	92.4/-
FreeAT(Zhu et al., 2020)	90.0/-	92.5/-	86.7	94.7	-/90.7	68.8	96.1	92.4/-
FreeLB(Zhu et al., 2020)	90.6/-	**92.6/-**	88.1	95	-/91.4	**71.1**	96.7	92.7/-
$\mathbf{SMART_{RoBERTa}}$	**91.1/91.3**	92.4/89.8	**92**	**95.6**	**89.2/92.1**	70.6	**96.9**	**92.8/92.6**

<Main Result on GLUE Development Set>

<GLUE Benchmark>

Model	MNLI-m/mm Acc	QQP ACC/F1	RTE Acc	QNLI Acc	MRPC Acc/F1	CoLA Mcc	SST Mcc	STS-B P/S Corr
				$BERT_{BASE}$				
BERT(Devlin et al., 2019)	84.4/-	-	-	88.4	-/86.7	-	92.7	-
$BERT_{ReImp}$	84.5/84.4	90.9/88.3	63.5	91.1	84.1/89.0	54.7	92.9	89.2/88.8
$SMART_{BERT}$	**85.6/86.0**	**91.5/88.5**	**71.2**	**91.7**	**87.7/91.3**	**59.1**	**93**	**90.0/89.4**
				$RoBERTa_{LARGE}$				
RoBERTa(Liu et al., 2019)	90.2/-	92.2/-	86.6	94.7	-/90.9	68	96.4	92.4/-
PGD(Zhu et al., 2020)	90.5/-	92.5/-	87.4	94.9	-/90.9	69.7	96.4	92.4/-
FreeAT(Zhu et al., 2020)	90.0/-	92.5/-	86.7	94.7	-/90.7	68.8	96.1	92.4/-
FreeLB(Zhu et al., 2020)	90.6/-	**92.6/-**	88.1	95	-/91.4	**71.1**	96.7	92.7/-
$SMART_{RoBERTa}$	**91.1/91.3**	92.4/89.8	**92**	**95.6**	**89.2/92.1**	70.6	**96.9**	**92.8/92.6**

<Main Result on GLUE Development Set>

<GLUE Benchmark>

Model /#Train	CoLA	SST	MRPC	STS-B	QQP	MNLI-m/mm	QNLI	RTE	WNLI	AX	Score	#params
	8.5k	**67k**	**3.7k**	**7k**	**634k**	**393k**	**108k**	**2.5k**	**634**			
Human Performance	66.4	97.8	86.3/80.8	92.7/92.6	59.5/80.4	92.0/92.8	91.2	93.6	95.9	-	87.1	
Ensemble Models												
RoBERTa	67.8	96.7	92.3/89.8	92.2/91.9	74.3/90.2	90.8/90.2	98.9	88.2	89	48.7	88.5	356M
FreeLB	68	96.8	93.1/90.8	92.4/92.2	**74.8**/90.3	91.1/90.7	98.8	88.7	89	50.1	88.8	356M
ALICE	69.2	97.1	93.6/91.5	92.7/92.3	74.4/**90.7**	90.7/90.2	**99.2**	87.3	89.7	47.8	89	340M
ALBERT	69.1	97.1	93.4/91.2	92.5/92.0	74.2/90.5	91.3/91.0	**99.2**	89.2	91.8	50.2	89.4	235M
MT-DNN-SMART	69.5	**97.5**	**93.7/91.6**	**92.9/92.5**	73.9/90.2	91.0/90.8	**99.2**	89.7	94.5	50.2	**89.9**	356M
Single Model												
BERT$_{\text{LARGE}}$	60.5	94.9	89.3/85.4	87.6/86.5	72.1/89.3	86.7/85.9	92.7	70.1	65.1	39.6	80.5	335M
MT-DNN	62.5	95.6	90.0/86.7	88.3/87.7	72.4/89.6	86.7/86.0	93.1	75.5	65.1	40.3	82.7	335M
T5	**70.8**	97.1	91.9/89.2	92.5/92.1	74.6/90.4	**92.0/91.7**	96.7	**92.5**	**93.2**	**53.1**	89.7	11,000M
SMART$_{\text{RoBERTa}}$	65.1	**97.5**	**93.7/91.6**	**92.9/92.5**	74.0/90.1	91.0/90.8	95.4	87.9	87.9	50.2	88.4	356M

<GLUE Test Set Results Scored Using the GLUE Evaluation Server>
-December 5, 2019-

<GLUE Benchmark>

Model /#Train	CoLA	SST	MRPC	STS-B	QQP	MNLI-m/mm	QNLI	RTE	WNLI	AX	Score	#params
	8.5k	67k	3.7k	7k	634k	393k	108k	2.5k	634			
Human Performance	66.4	97.8	86.3/80.8	92.7/92.6	59.5/80.4	92.0/92.8	91.2	93.6	95.9	-	87.1	
Ensemble Models												
RoBERTa	67.8	96.7	92.3/89.8	92.2/91.9	74.3/90.2	90.8/90.2	98.9	88.2	89	48.7	88.5	356M
FreeLB	68	96.8	93.1/90.8	92.4/92.2	**74.8**/90.3	91.1/90.7	98.8	88.7	89	50.1	88.8	356M
ALICE	69.2	97.1	93.6/91.5	92.7/92.3	74.4/**90.7**	90.7/90.2	**99.2**	87.3	89.7	47.8	89	340M
ALBERT	69.1	97.1	93.4/91.2	92.5/92.0	74.2/90.5	91.3/91.0	**99.2**	89.2	91.8	50.2	89.4	235M
MT-DNN-SMART	69.5	**97.5**	**93.7/91.6**	**92.9/92.5**	73.9/90.2	91.0/90.8	**99.2**	89.7	94.5	50.2	**89.9**	356M
Single Model												
$\text{BERT}_{\text{LARGE}}$	60.5	94.9	89.3/85.4	87.6/86.5	72.1/89.3	86.7/85.9	92.7	70.1	65.1	39.6	80.5	335M
MT-DNN	62.5	95.6	90.0/86.7	88.3/87.7	72.4/89.6	86.7/86.0	93.1	75.5	65.1	40.3	82.7	335M
T5	**70.8**	97.1	91.9/89.2	92.5/92.1	74.6/90.4	**92.0/91.7**	96.7	**92.5**	**93.2**	**53.1**	89.7	11,000M
$\text{SMART}_{\text{RoBERTa}}$	65.1	**97.5**	**93.7/91.6**	**92.9/92.5**	74.0/90.1	91.0/90.8	95.4	87.9	87.9	50.2	88.4	356M

<GLUE Test Set Results Scored Using the GLUE Evaluation Server>
-December 5, 2019-

<GLUE Benchmark>

Model /#Train	CoLA 8.5k	SST 67k	MRPC 3.7k	STS-B 7k	QQP 634k	MNLI-m/mm 393k	QNLI 108k	RTE 2.5k	WNLI 634	AX	Score	#params
Human Performance	66.4	97.8	86.3/80.8	92.7/92.6	59.5/80.4	92.0/92.8	91.2	93.6	95.9	-	87.1	
Ensemble Models												
RoBERTa	67.8	96.7	92.3/89.8	92.2/91.9	74.3/90.2	90.8/90.2	98.9	88.2	89	48.7	88.5	356M
FreeLB	68	96.8	93.1/90.8	92.4/92.2	**74.8**/90.3	91.1/90.7	98.8	88.7	89	50.1	88.8	356M
ALICE	69.2	97.1	93.6/91.5	92.7/92.3	74.4/**90.7**	90.7/90.2	**99.2**	87.3	89.7	47.8	89	340M
ALBERT	69.1	97.1	93.4/91.2	92.5/92.0	74.2/90.5	91.3/91.0	**99.2**	89.2	91.8	50.2	89.4	235M
MT-DNN-SMART	69.5	**97.5**	**93.7/91.6**	**92.9/92.5**	73.9/90.2	91.0/90.8	**99.2**	89.7	94.5	50.2	**89.9**	356M
Single Model												
$BERT_{LARGE}$	60.5	94.9	89.3/85.4	87.6/86.5	72.1/89.3	86.7/85.9	92.7	70.1	65.1	39.6	80.5	335M
MT-DNN	62.5	95.6	90.0/86.7	88.3/87.7	72.4/89.6	86.7/86.0	93.1	75.5	65.1	40.3	82.7	335M
T5	**70.8**	97.1	91.9/89.2	92.5/92.1	74.6/90.4	**92.0/91.7**	96.7	**92.5**	**93.2**	**53.1**	89.7	11,000M
$SMART_{RoBERTa}$	65.1	**97.5**	**93.7/91.6**	**92.9/92.5**	74.0/90.1	91.0/90.8	95.4	87.9	87.9	50.2	88.4	356M

<GLUE Test Set Results Scored Using the GLUE Evaluation Server>
-December 5, 2019-

<GLUE Benchmark>

Model /#Train	CoLA 8.5k	SST 67k	MRPC 3.7k	STS-B 7k	QQP 634k	MNLI-m/mm 393k	QNLI 108k	RTE 2.5k	WNLI 634	AX	Score	#params
Human Performance	66.4	97.8	86.3/80.8	92.7/92.6	59.5/80.4	92.0/92.8	91.2	93.6	95.9	-	87.1	
Ensemble Models												
RoBERTa	67.8	96.7	92.3/89.8	92.2/91.9	74.3/90.2	90.8/90.2	98.9	88.2	89	48.7	88.5	356M
FreeLB	68	96.8	93.1/90.8	92.4/92.2	**74.8**/90.3	91.1/90.7	98.8	88.7	89	50.1	88.8	356M
ALICE	69.2	97.1	93.6/91.5	92.7/92.3	74.4/**90.7**	90.7/90.2	**99.2**	87.3	89.7	47.8	89	340M
ALBERT	69.1	97.1	93.4/91.2	92.5/92.0	74.2/90.5	91.3/91.0	**99.2**	89.2	91.8	50.2	89.4	235M
MT-DNN-SMART	69.5	**97.5**	**93.7/91.6**	**92.9/92.5**	73.9/90.2	91.0/90.8	**99.2**	89.7	94.5	50.2	**89.9**	356M
Single Model												
$BERT_{LARGE}$	60.5	94.9	89.3/85.4	87.6/86.5	72.1/89.3	86.7/85.9	92.7	70.1	65.1	39.6	80.5	335M
MT-DNN	62.5	95.6	90.0/86.7	88.3/87.7	72.4/89.6	86.7/86.0	93.1	75.5	65.1	40.3	82.7	335M
T5	**70.8**	97.1	91.9/89.2	92.5/92.1	74.6/90.4	**92.0/91.7**	96.7	**92.5**	**93.2**	**53.1**	89.7	11,000M
$SMART_{RoBERTa}$	65.1	**97.5**	**93.7/91.6**	**92.9/92.5**	74.0/90.1	91.0/90.8	95.4	87.9	87.9	50.2	88.4	356M

<GLUE Test Set Results Scored Using the GLUE Evaluation Server>
-December 5, 2019-

<GLUE Benchmark>

Model /#Train	CoLA 8.5k	SST 67k	MRPC 3.7k	STS-B 7k	QQP 634k	MNLI-m/mm 393k	QNLI 108k	RTE 2.5k	WNLI 634	AX	Score	#params
Human Performance	66.4	97.8	86.3/80.8	92.7/92.6	59.5/80.4	92.0/92.8	91.2	93.6	95.9	-	87.1	
Ensemble Models												
RoBERTa	67.8	96.7	92.3/89.8	92.2/91.9	74.3/90.2	90.8/90.2	98.9	88.2	89	48.7	88.5	356M
FreeLB	68	96.8	93.1/90.8	92.4/92.2	**74.8**/90.3	91.1/90.7	98.8	88.7	89	50.1	88.8	356M
ALICE	69.2	97.1	93.6/91.5	92.7/92.3	74.4/**90.7**	90.7/90.2	**99.2**	87.3	89.7	47.8	89	340M
ALBERT	69.1	97.1	93.4/91.2	92.5/92.0	74.2/90.5	91.3/91.0	**99.2**	89.2	91.8	50.2	89.4	235M
MT-DNN-SMART	69.5	**97.5**	**93.7/91.6**	**92.9/92.5**	73.9/90.2	91.0/90.8	**99.2**	89.7	94.5	50.2	**89.9**	356M
Single Model												
$\text{BERT}_{\text{LARGE}}$	60.5	94.9	89.3/85.4	87.6/86.5	72.1/89.3	86.7/85.9	92.7	70.1	65.1	39.6	80.5	335M
MT-DNN	62.5	95.6	90.0/86.7	88.3/87.7	72.4/89.6	86.7/86.0	93.1	75.5	65.1	40.3	82.7	335M
T5	**70.8**	97.1	91.9/89.2	92.5/92.1	74.6/90.4	**92.0/91.7**	96.7	**92.5**	**93.2**	**53.1**	89.7	11,000M
$\text{SMART}_{\text{RoBERTa}}$	65.1	**97.5**	**93.7/91.6**	**92.9/92.5**	74.0/90.1	91.0/90.8	95.4	87.9	87.9	50.2	88.4	356M

<GLUE Test Set Results Scored Using the GLUE Evaluation Server>
-December 5, 2019-

<Ablation Study>

Model	MNLI Acc	RTE Acc	QNLI Acc	SST Acc	MRPC Acc
BERT	84.5	63.5	91.1	92.9	89
SMART	**95.6**	**71.2**	**91.7**	**93**	**91.3**
-$\mathcal{R}_s$	84.8	70.8	91.3	92.8	90.8
-$\mathcal{D}_{\mathbf{Breg}}$	85.4	**71.2**	91.6	92.9	91.2

<Ablation Study of SMART on 5 GLUE Task>
Backbone: BERT

<Ablation Study>

Model	MNLI Acc	RTE Acc	QNLI Acc	SST Acc	MRPC Acc
BERT	84.5	63.5	91.1	92.9	89
SMART	**95.6**	**71.2**	**91.7**	**93**	**91.3**
-$\mathcal{R}_s$	84.8	70.8	91.3	92.8	90.8
-$\mathcal{D}_{\mathbf{Breg}}$	85.4	**71.2**	91.6	92.9	91.2

<Ablation Study of SMART on 5 GLUE Task>
Backbone: BERT

<Ablation Study>

Model	MNLI Acc	RTE Acc	QNLI Acc	SST Acc	MRPC Acc
BERT	84.5	63.5	91.1	92.9	89
SMART	**95.6**	**71.2**	**91.7**	**93**	**91.3**
-$\mathcal{R}_s$	84.8	70.8	91.3	92.8	90.8
-$\mathcal{D}_{\mathbf{Breg}}$	85.4	**71.2**	91.6	92.9	91.2

<Ablation Study of SMART on 5 GLUE Task>
Backbone: BERT

Conclusion

<Conclusion>

- Proposed a *Smoothness-Inducing Adversarial Regularization* Technique to Effectively Control the **Extremely High Complexity** of the Model
- Proposed a Class of *Bregman Proximal Point Optimization* Method to Prevent **Aggressive Updating**
- Achieved State-of-the-art Results on Several Popular NLP Benchmarks (e.g. GLUE, …)

Any Questions?

Thank You