EUROPEAN CONFERENCE ON COMPUTER VISION

MILANO 2024

Model Stock: All we need is just a few fine-tuned models

Dong-Hwan Jang[1,2], Sangdoo Yun[1†], Dongyoon Han[1†]

[1]NAVER AI Lab, [2]Samsung Advanced Institute of Technology (SAIT)
[†] corresponding authors

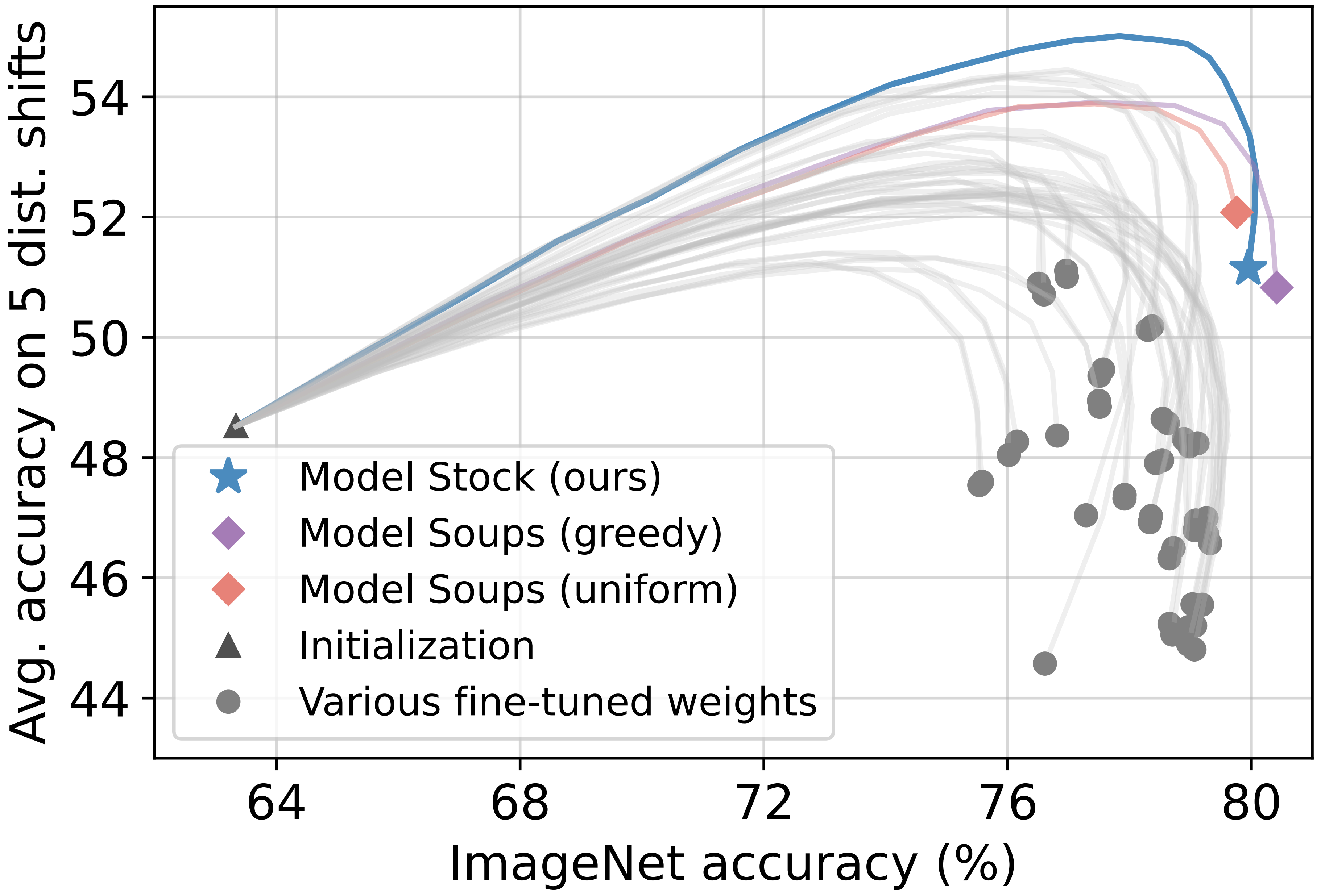
* Work done during an internship at NAVER AI Lab

Introduction

Robust Fine-tuning

- Traditional **robust fine-tuning methods** like Model Soup require dozens of fine-tuned weights

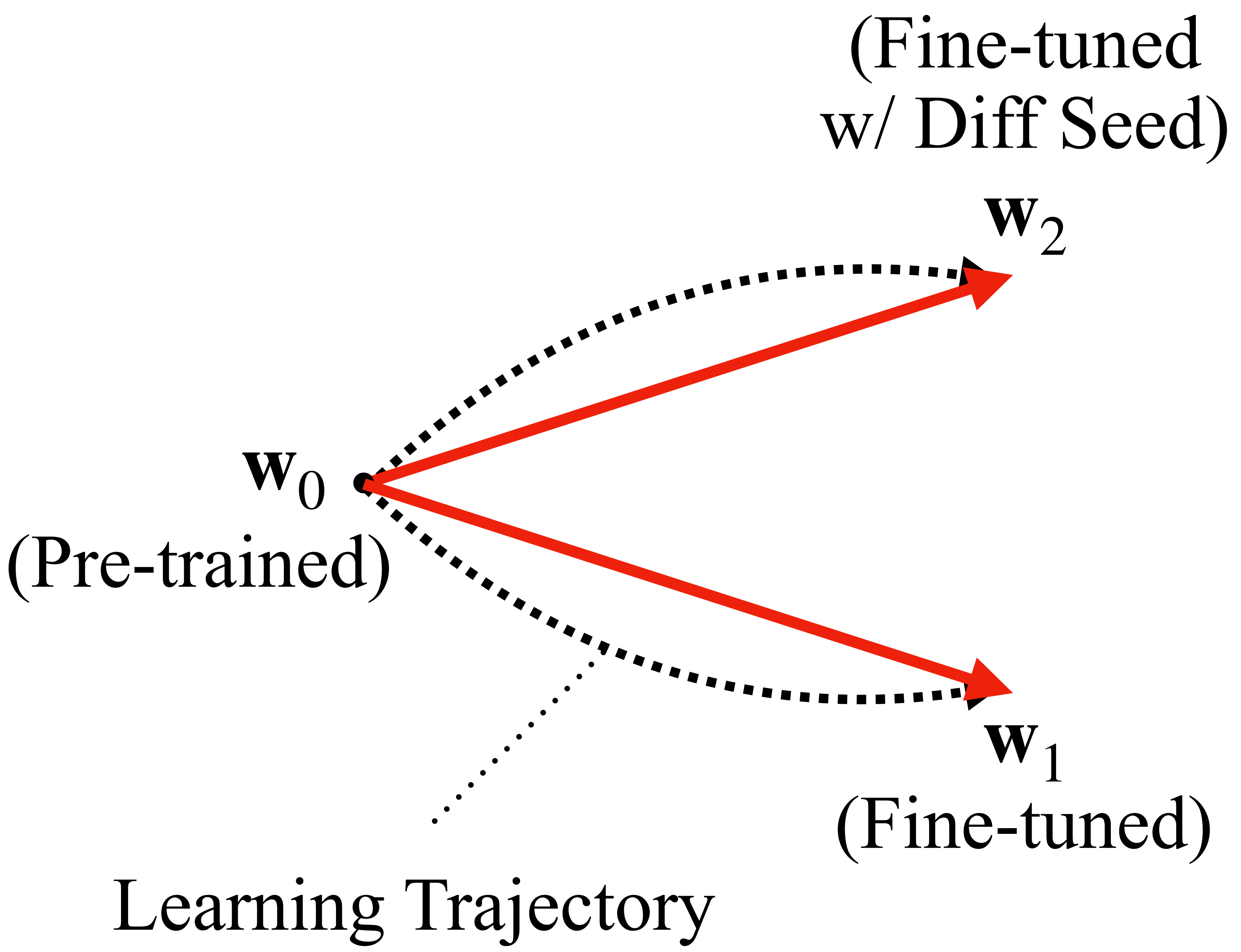
- We introduce **Model Stock**, a novel fine-tuning method that enhances both in-distribution and out-of-distribution performance while drastically reducing computational costs.



Observation 1: ***Angle and Norm Consistency among Fine-tuned Weights***

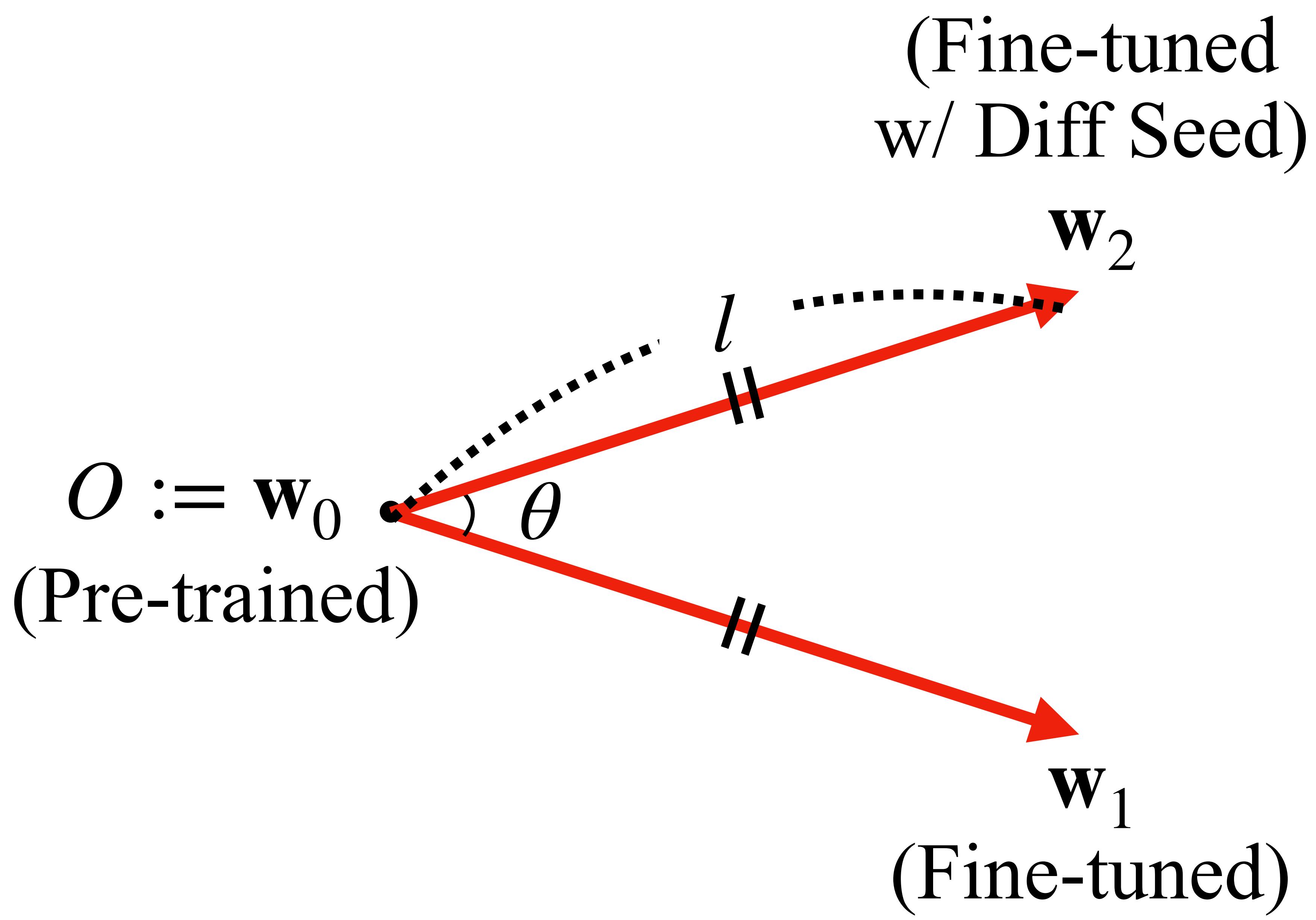
Observation 1: **Angle and Norm Consistency**

Pretrain-Finetune Paradigm in Weight Space



Observation 1: **Angle and Norm Consistency**

Geometric Relations between Fine-tuned Weights

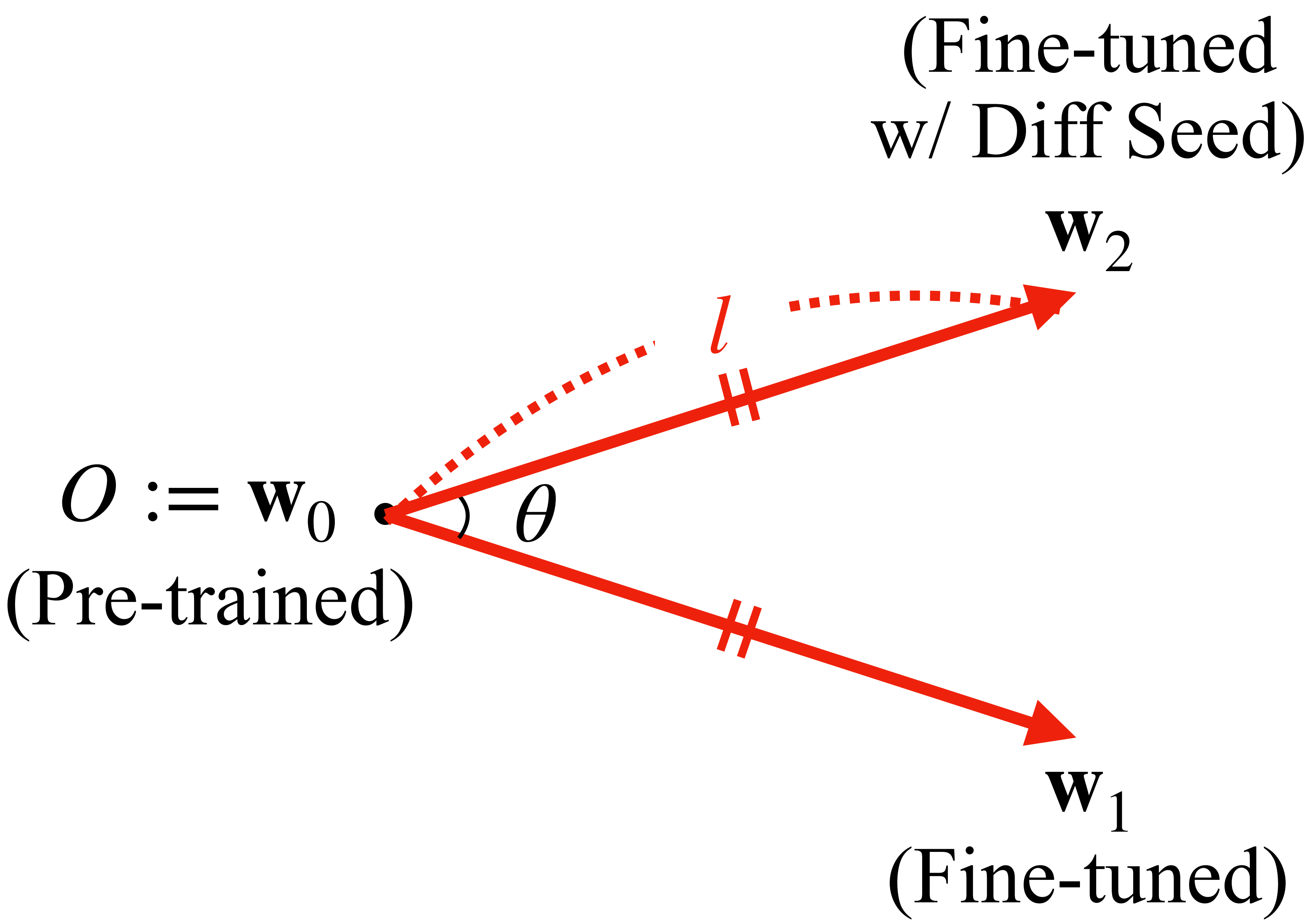


$\forall$ random seeds i and j,

$$\mathbf{w}_i \cdot \mathbf{w}_j = \begin{cases} l^2 & \text{if } i = j, \\ l^2 \cos\theta & \text{otherwise,} \end{cases}$$

Observation 1: **Angle and Norm Consistency**

Geometric Relations between Fine-tuned Weights



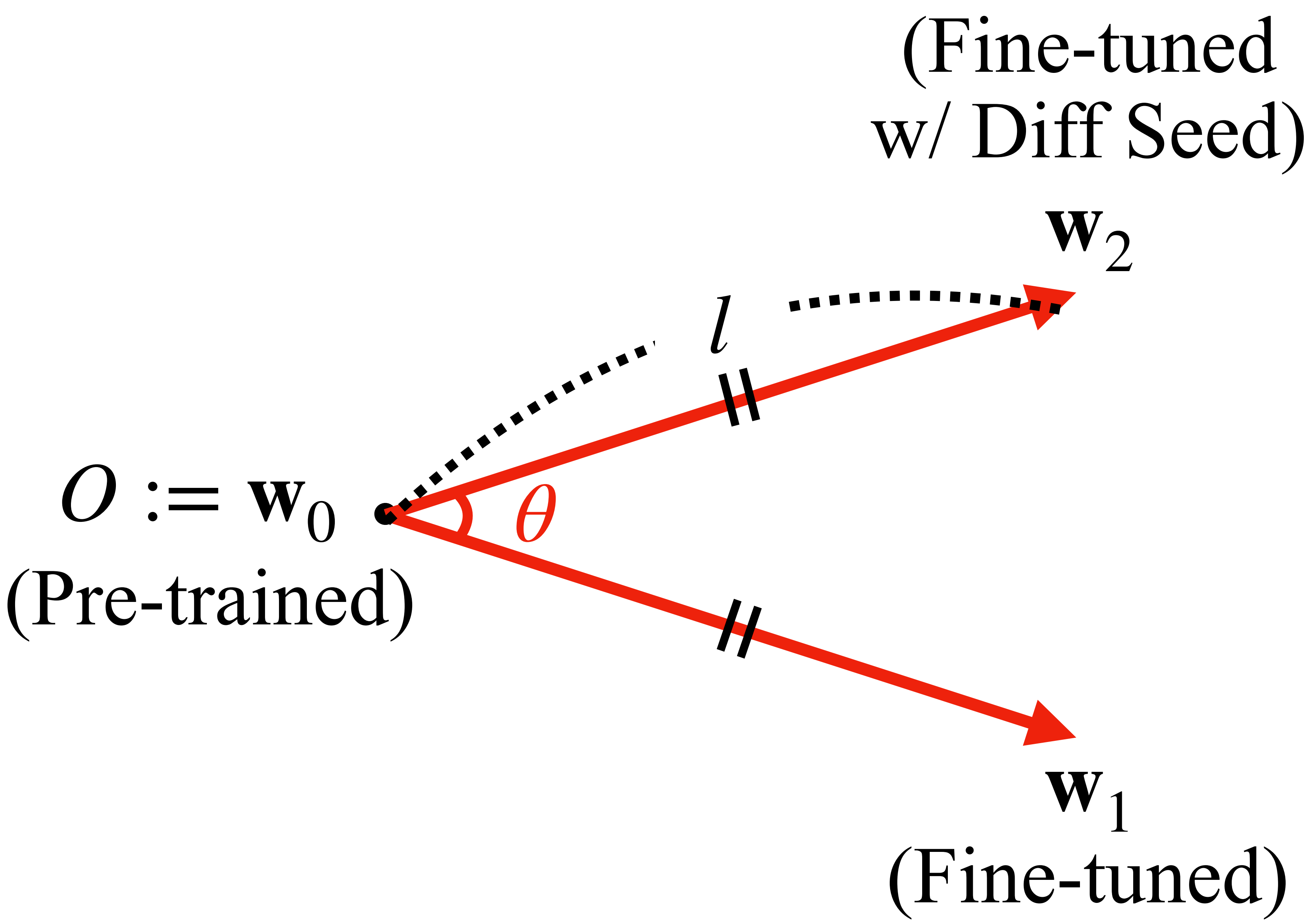
$\forall$ random seeds i and j,

Norm Consistency

$$\mathbf{w}_i \cdot \mathbf{w}_j = \begin{cases} l^2 & \text{if } i = j, \\ l^2 \cos\theta & \text{otherwise,} \end{cases}$$

Observation 1: **Angle and Norm Consistency**

Geometric Relations between Fine-tuned Weights



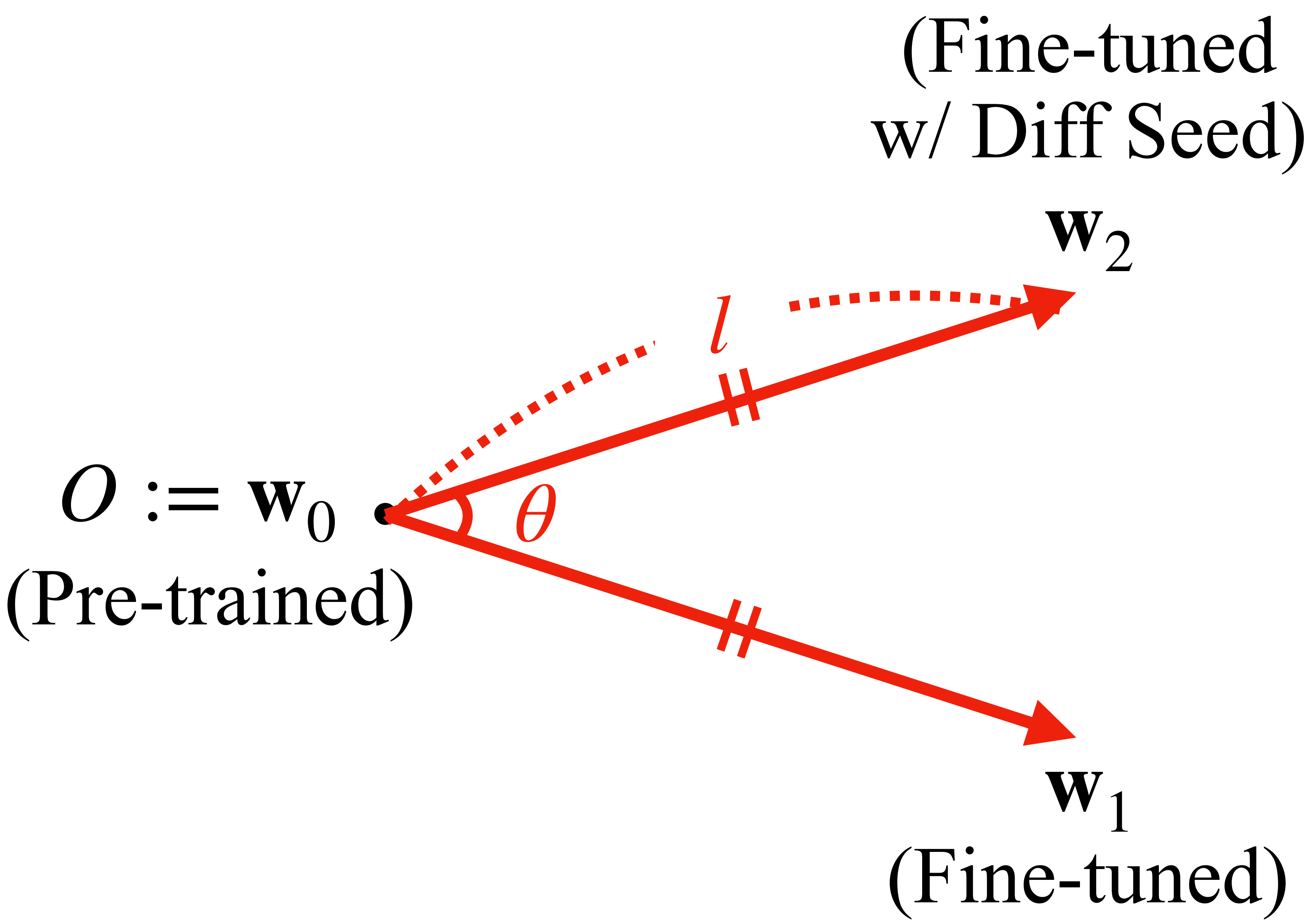
$\forall$ random seeds i and j,

$$\mathbf{w}_i \cdot \mathbf{w}_j = \begin{cases} l^2 & \text{if } i = j, \\ l^2 \cos\theta & \text{otherwise,} \end{cases}$$

Angle Consistency

Observation 1: **Angle and Norm Consistency**

Geometric Relations between Fine-tuned Weights



$\forall$ random seeds i and j,

$$\mathbf{w}_i \cdot \mathbf{w}_j = \begin{cases} l^2 & \text{if } i = j, \\ l^2 \cos\theta & \text{otherwise,} \end{cases}$$

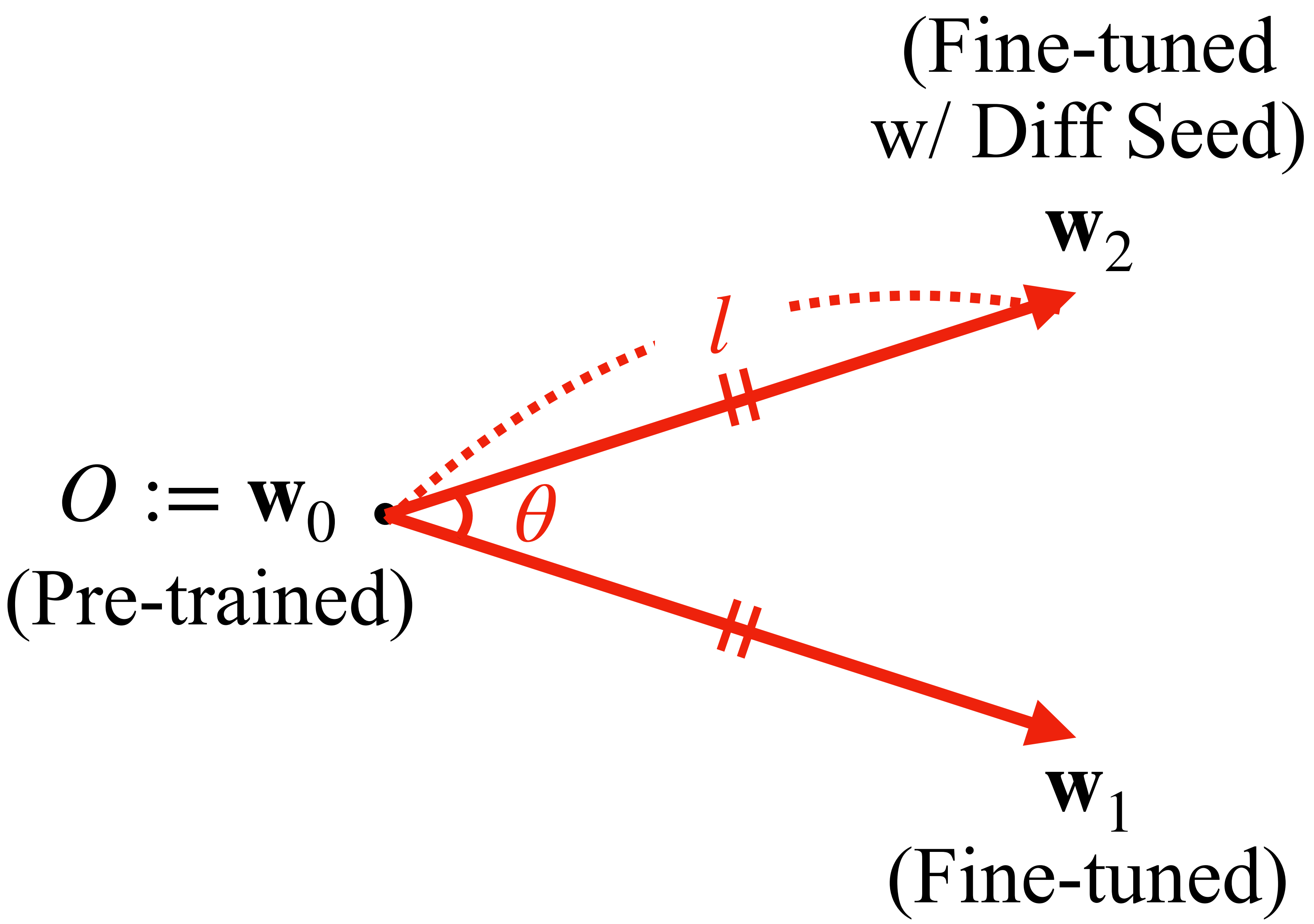
(i) **Various Setups** (arch., optim., h-params)

(ii) **Layer-wise**

(iii) **During Training**

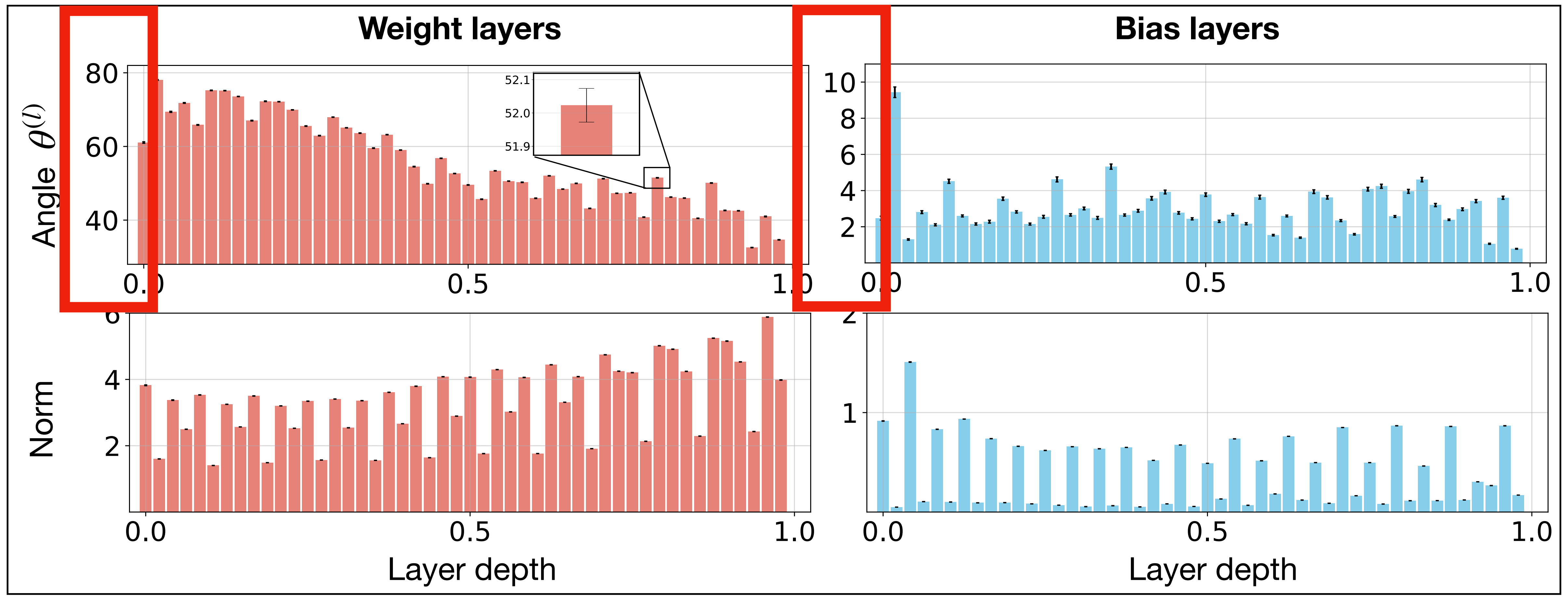
Observation 1: **Angle and Norm Consistency**

Geometric Relations between Fine-tuned Weights



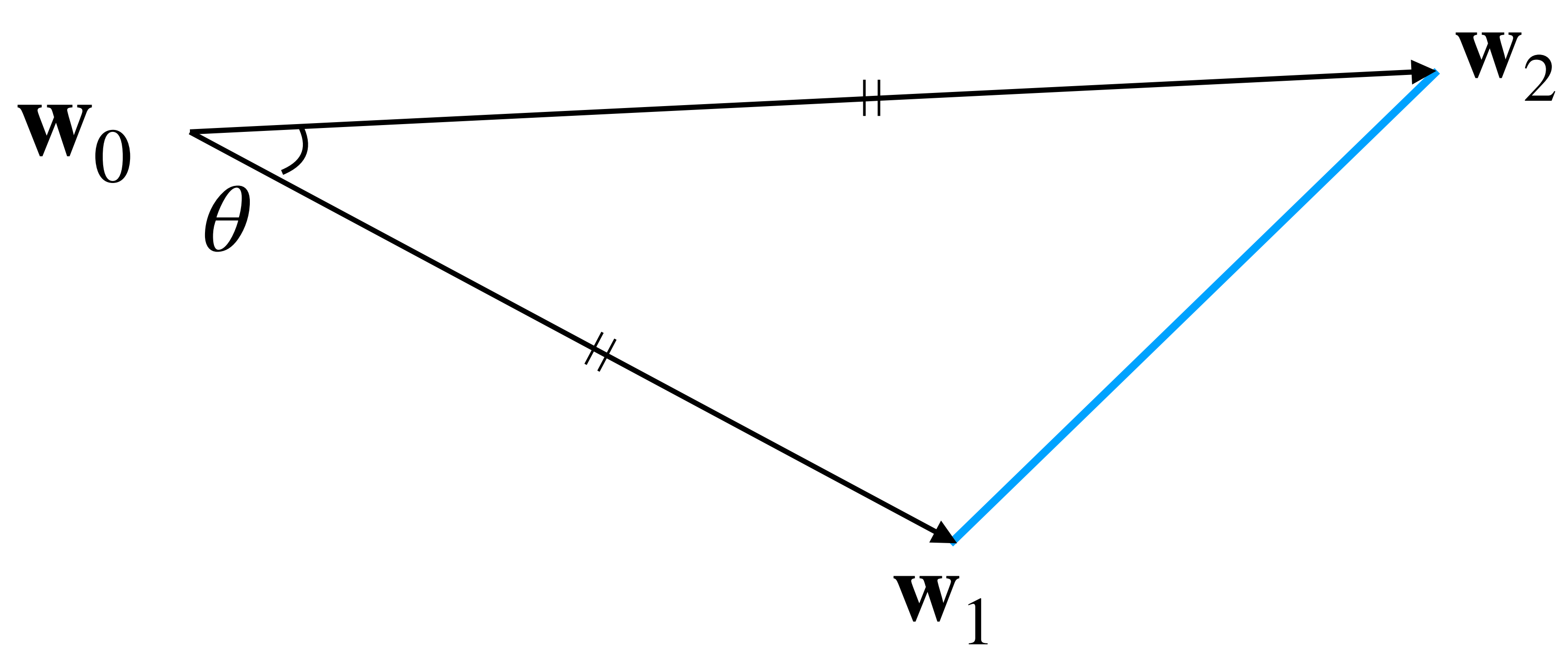
$\forall$ random seeds i and j,

$$\mathbf{w}_i \cdot \mathbf{w}_j = \begin{cases} l^2 & \text{if } i = j, \\ l^2 \cos\theta & \text{otherwise,} \end{cases}$$



Observation 1: **Angle and Norm Consistency**

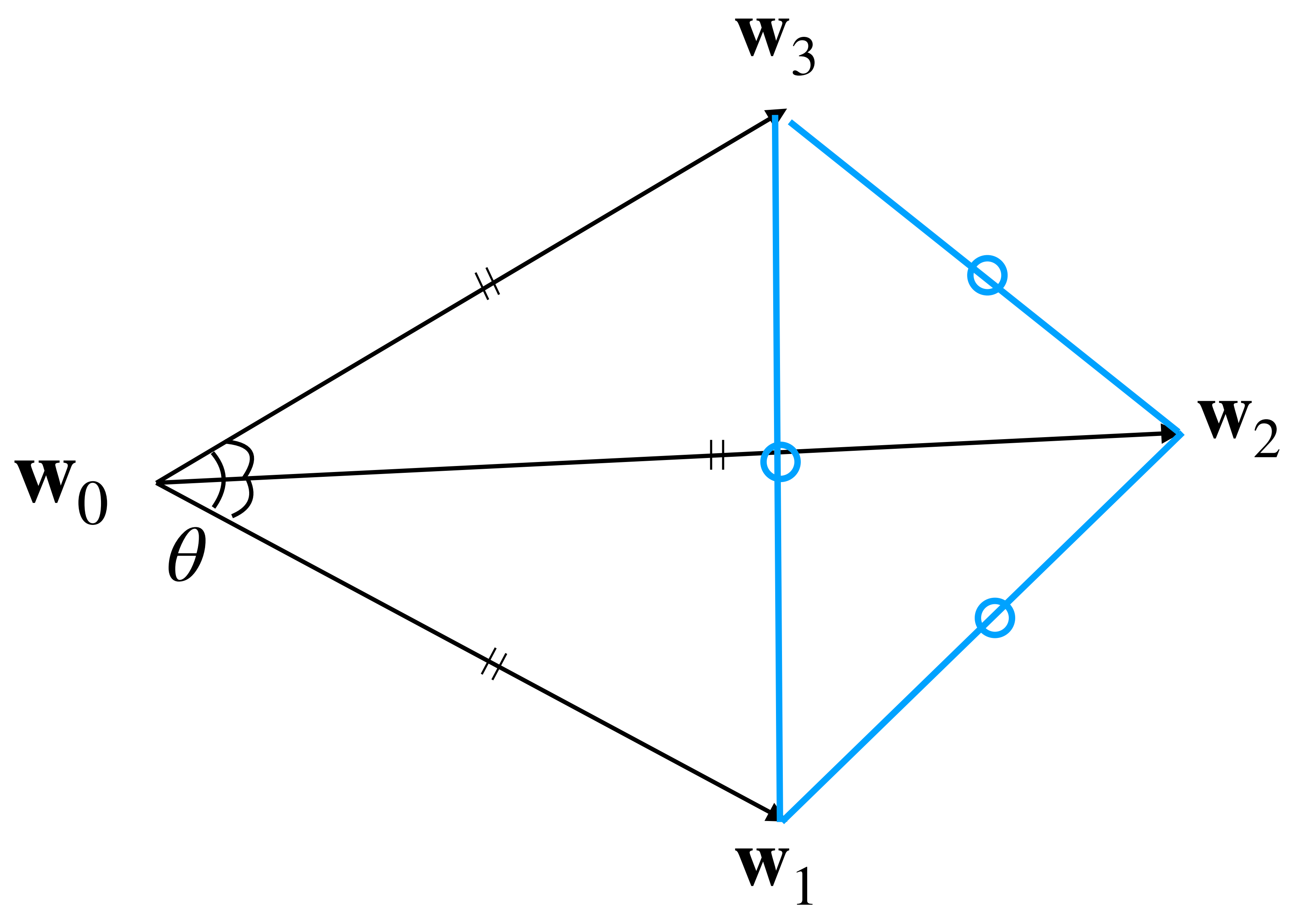
Geometric Relations between Fine-tuned Weights



N=2

Observation 1: **Angle and Norm Consistency**

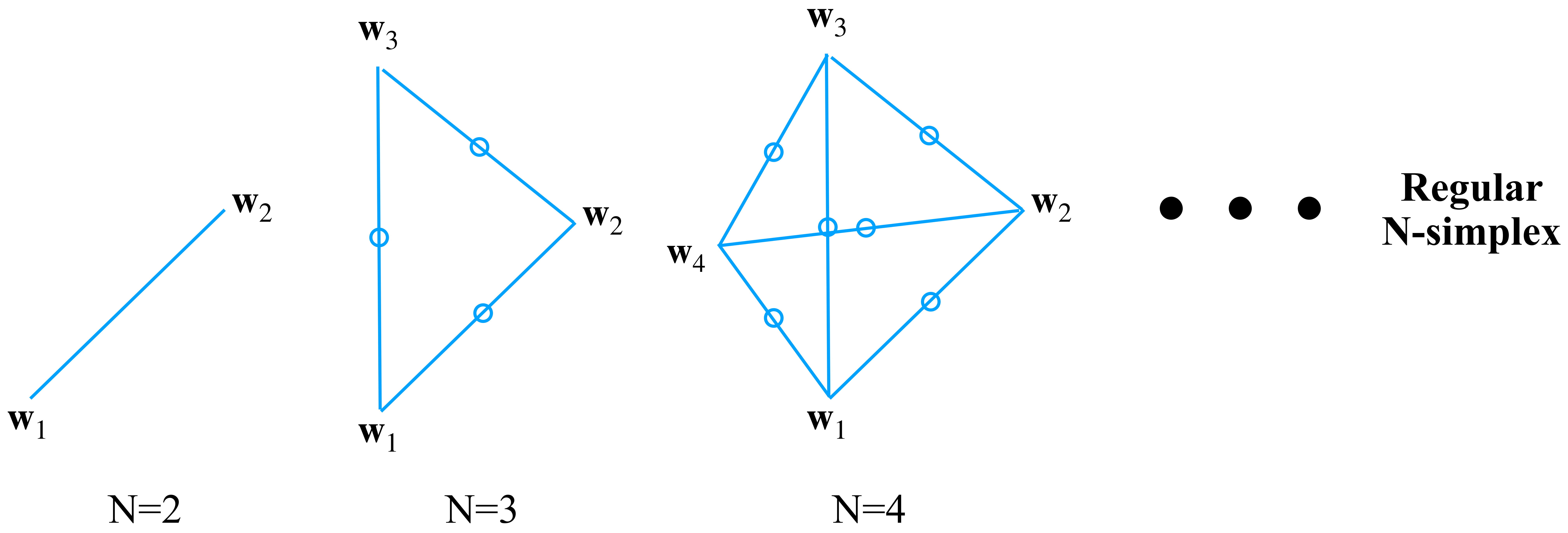
Geometric Relations between Fine-tuned Weights



N=3

Observation 1: **Angle and Norm Consistency**

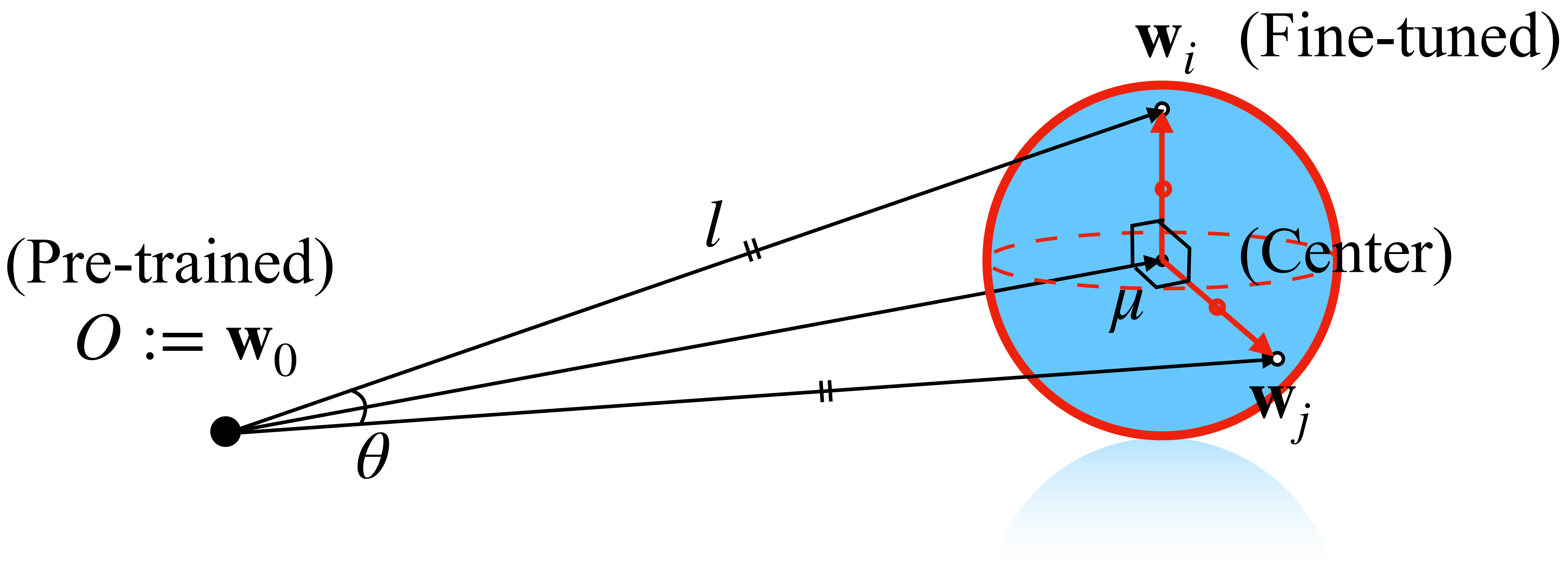
Geometric Relations between Fine-tuned Weights



Observation 1: **Angle and Norm Consistency**

Geometric Relations between Fine-tuned Weights

Let us define **the center** of fine-tuned weights as $\boldsymbol{\mu} = \lim_{N\to\infty} \frac{1}{N} \sum_{i=1}^{N} \mathbf{w}_i$,



(i) $\|\mathbf{w}_i - \boldsymbol{\mu}\| =$ constant *(thin shell)*

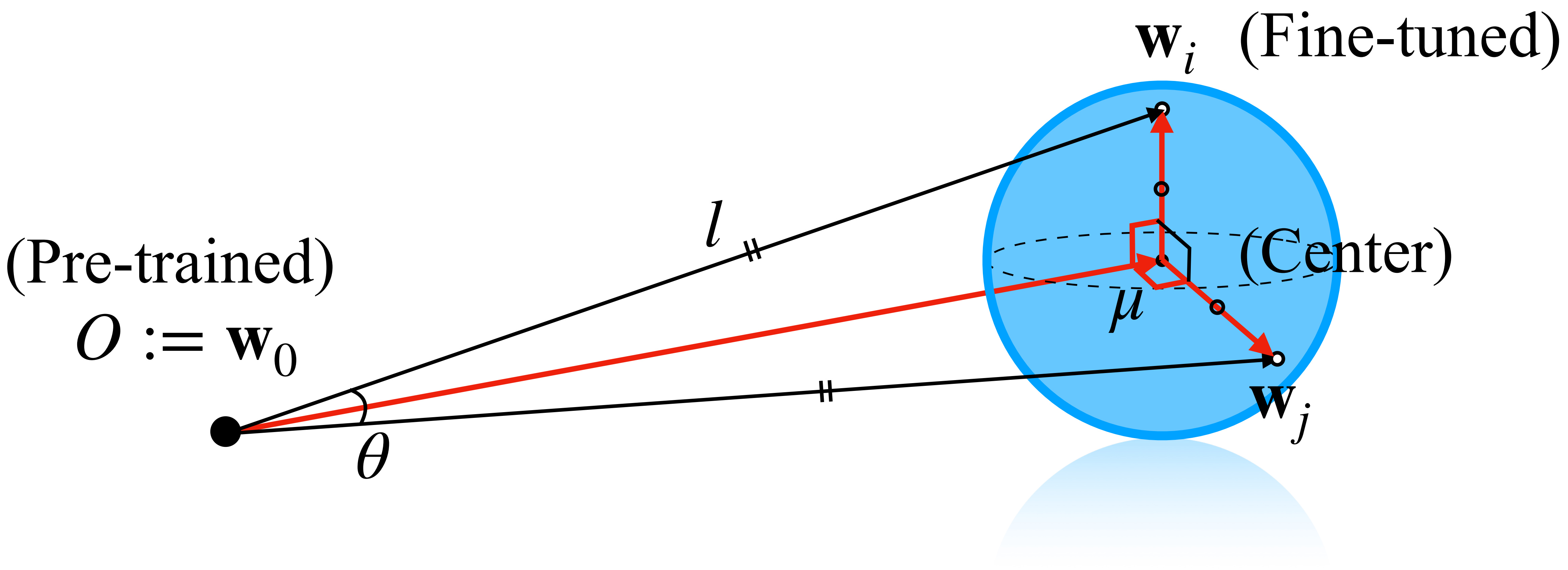
(ii) $(\mathbf{w}_0 - \boldsymbol{\mu}) \perp (\mathbf{w}_i - \boldsymbol{\mu})$

(iii) $(\mathbf{w}_i - \boldsymbol{\mu}) \perp (\mathbf{w}_j - \boldsymbol{\mu})$

Observation 1: **Angle and Norm Consistency**

Geometric Relations between Fine-tuned Weights

Let us define **the center** of fine-tuned weights as $\boldsymbol{\mu} = \lim_{N\to\infty} \frac{1}{N} \sum_{i=1}^{N} \mathbf{w}_i,$



(i) $\|\mathbf{w}_i - \boldsymbol{\mu}\| = \text{constant}$
(thin shell)

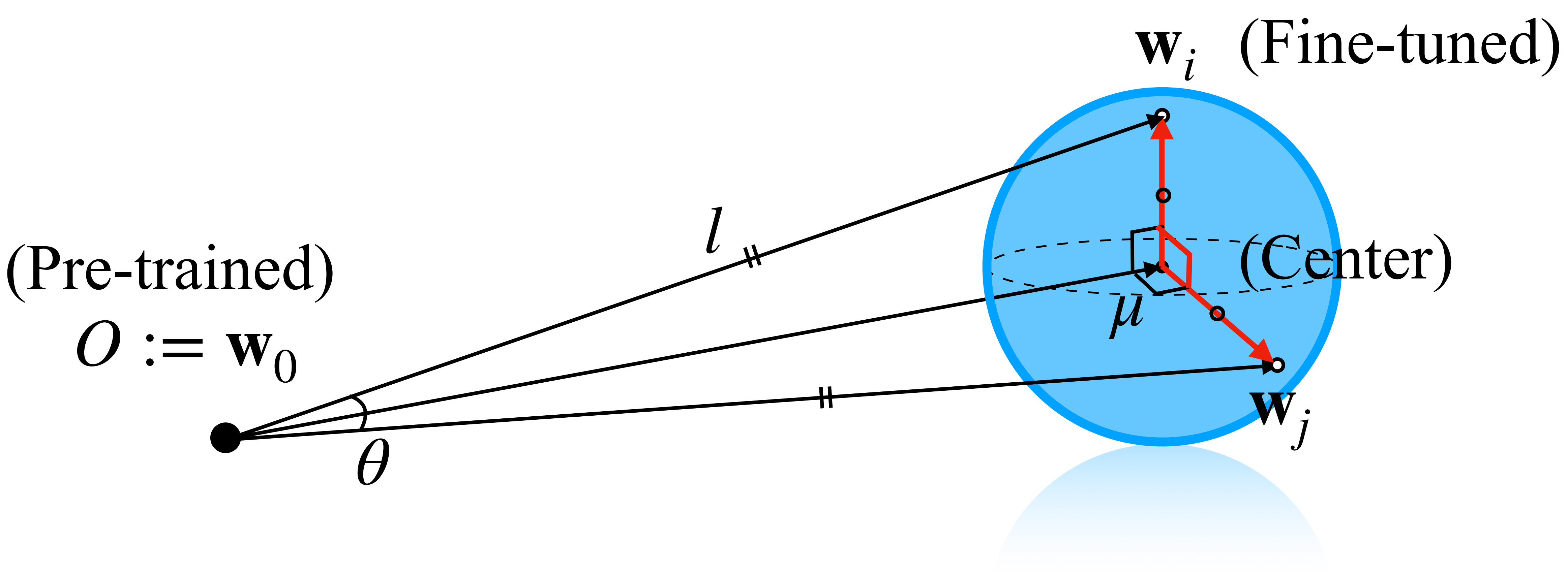
(ii) $(\mathbf{w}_0 - \boldsymbol{\mu}) \perp (\mathbf{w}_i - \boldsymbol{\mu})$

(iii) $(\mathbf{w}_i - \boldsymbol{\mu}) \perp (\mathbf{w}_j - \boldsymbol{\mu})$

Observation 1: **Angle and Norm Consistency**

Geometric Relations between Fine-tuned Weights

Let us define **the center** of fine-tuned weights as $\boldsymbol{\mu} = \lim_{N\to\infty} \frac{1}{N} \sum_{i=1}^{N} \mathbf{w}_i,$



(i) $\|\mathbf{w}_i - \boldsymbol{\mu}\| = \text{constant}$
(thin shell)

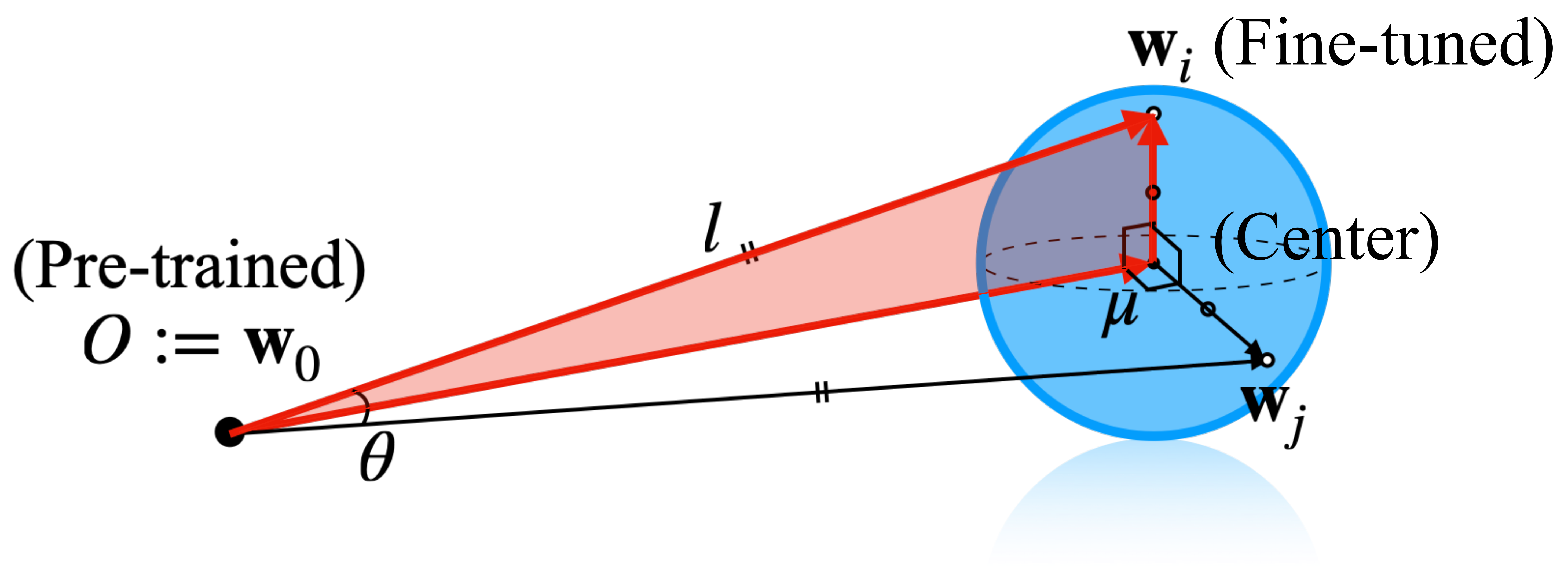
(ii) $(\mathbf{w}_0 - \boldsymbol{\mu}) \perp (\mathbf{w}_i - \boldsymbol{\mu})$

(iii) $(\mathbf{w}_i - \boldsymbol{\mu}) \perp (\mathbf{w}_j - \boldsymbol{\mu})$

Observation 2: ***Distance from the Center of Weights and Performance***

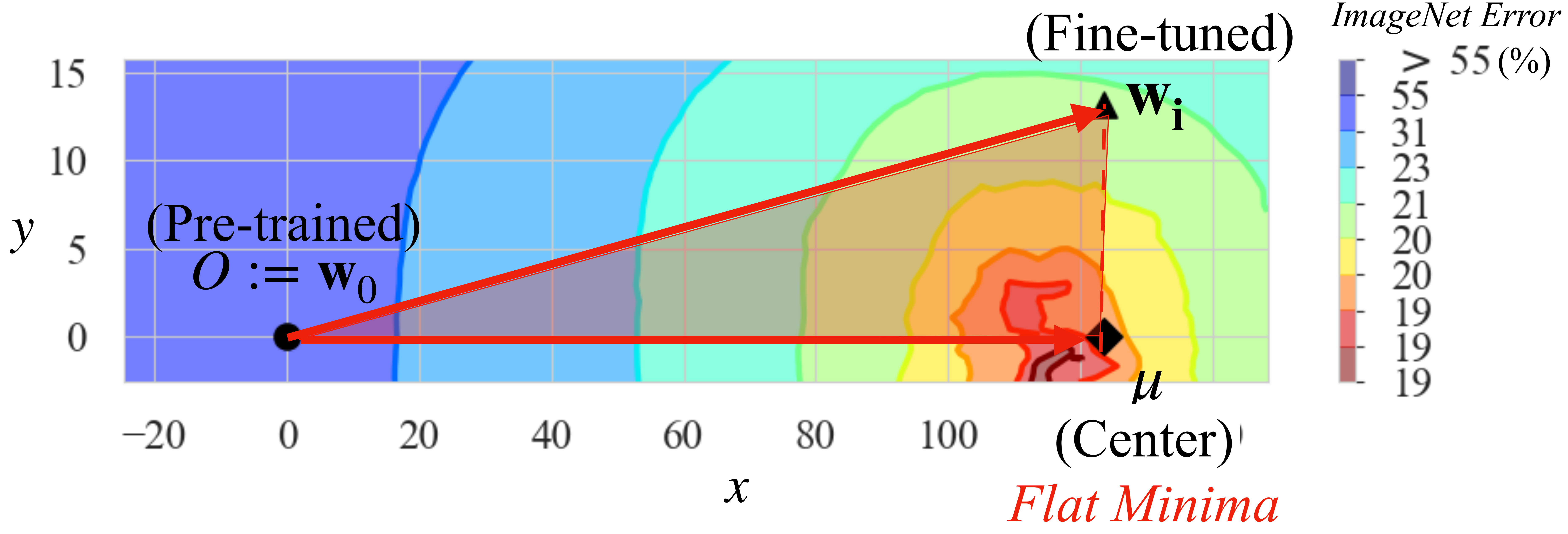
Observation 2: **Distance and Performance**

Test Error Landscape (ImageNet)



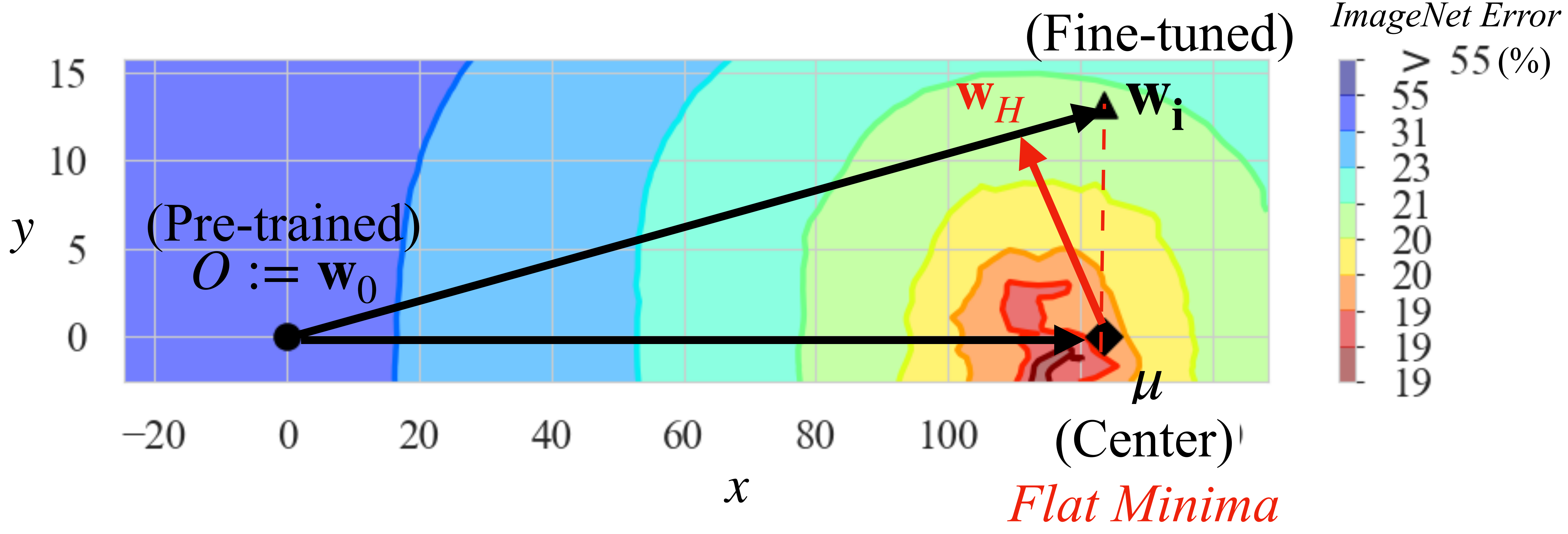
Observation 2: **Distance and Performance**

Test Error Landscape (ImageNet)



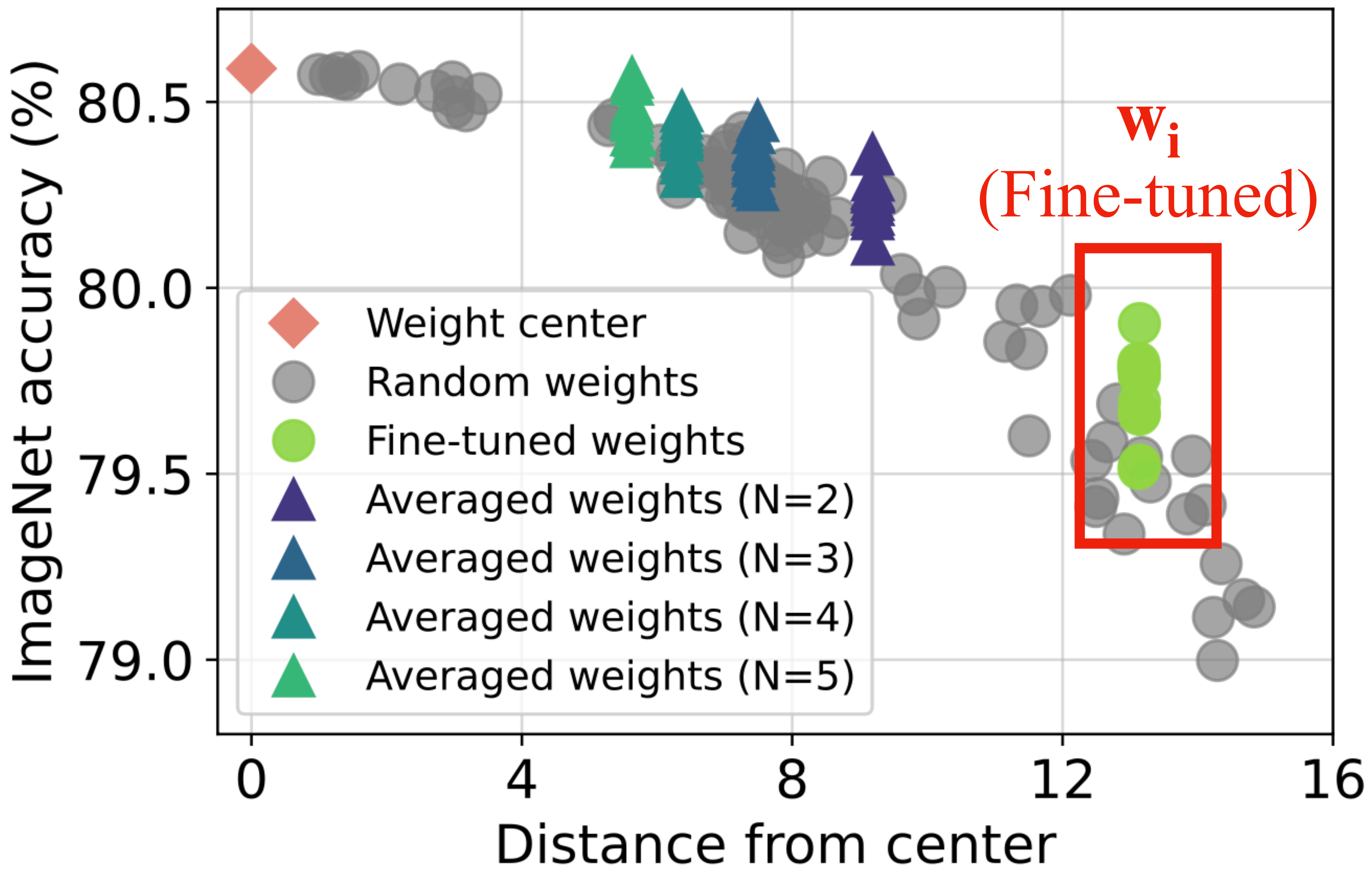
Observation 2: **Distance and Performance**

Test Error Landscape (ImageNet)



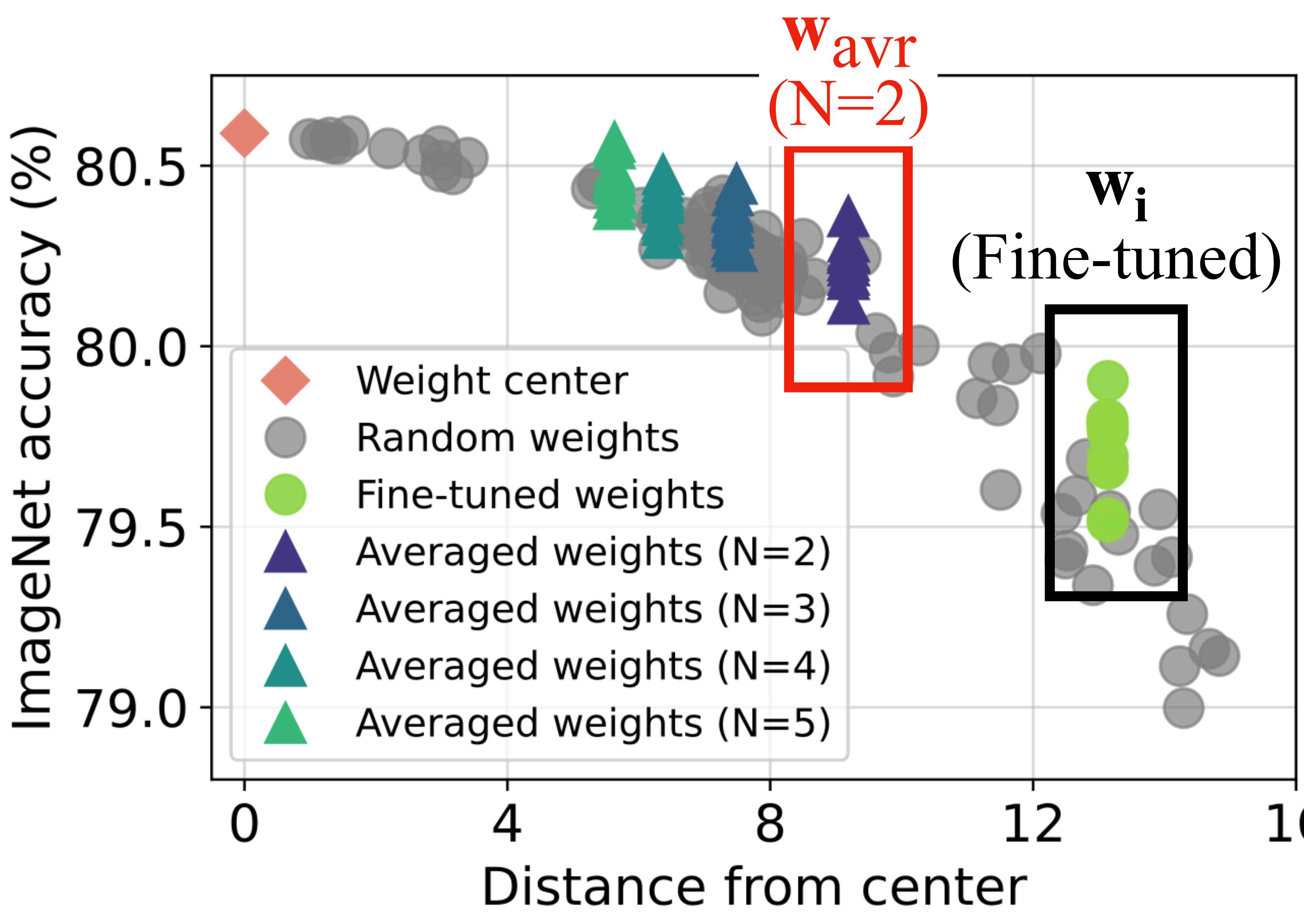
Observation 2: **Distance and Performance**

Distance vs. Random Weights' Performance



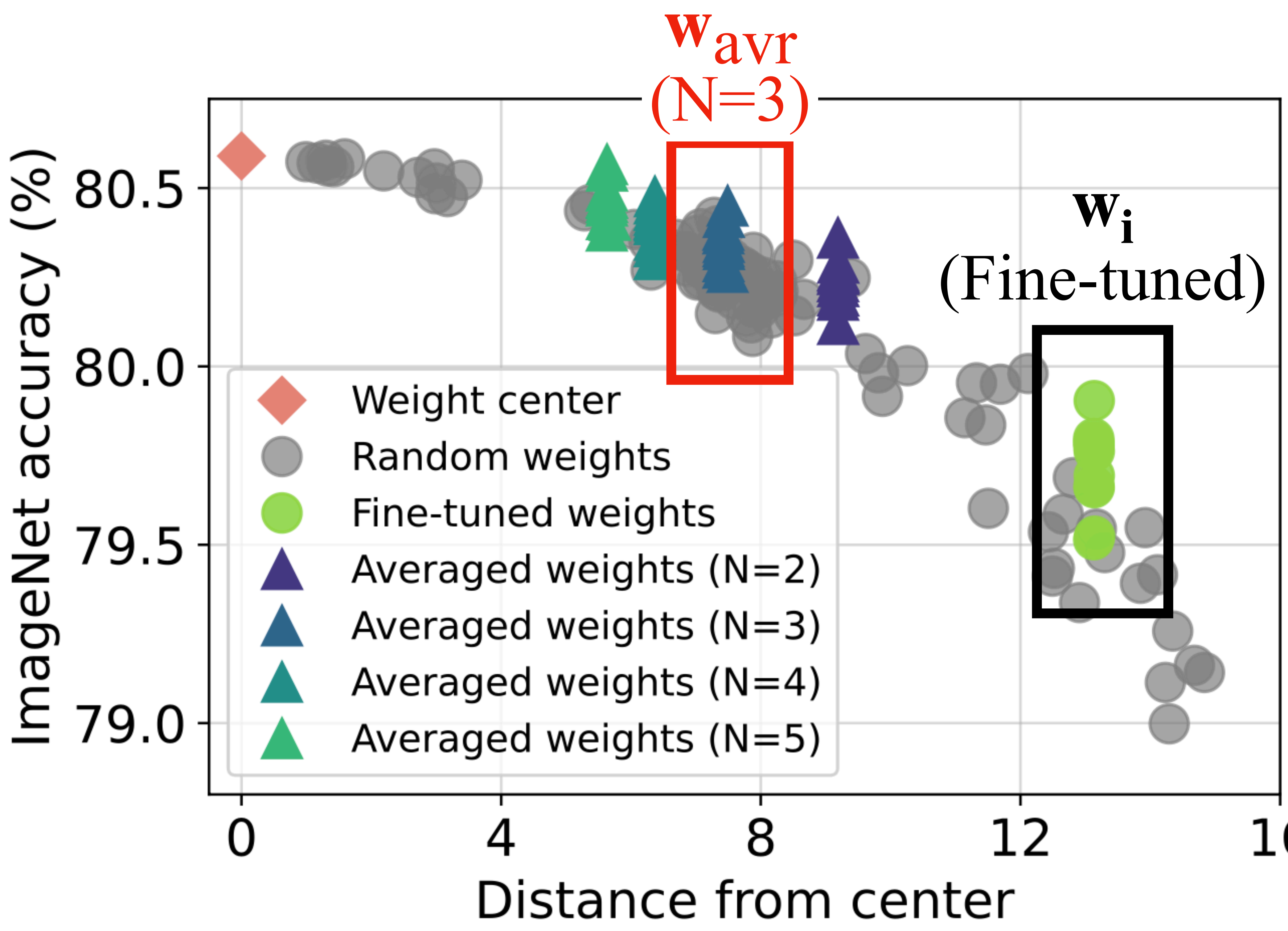
Observation 2: **Distance and Performance**

Distance vs. Random Weights' Performance



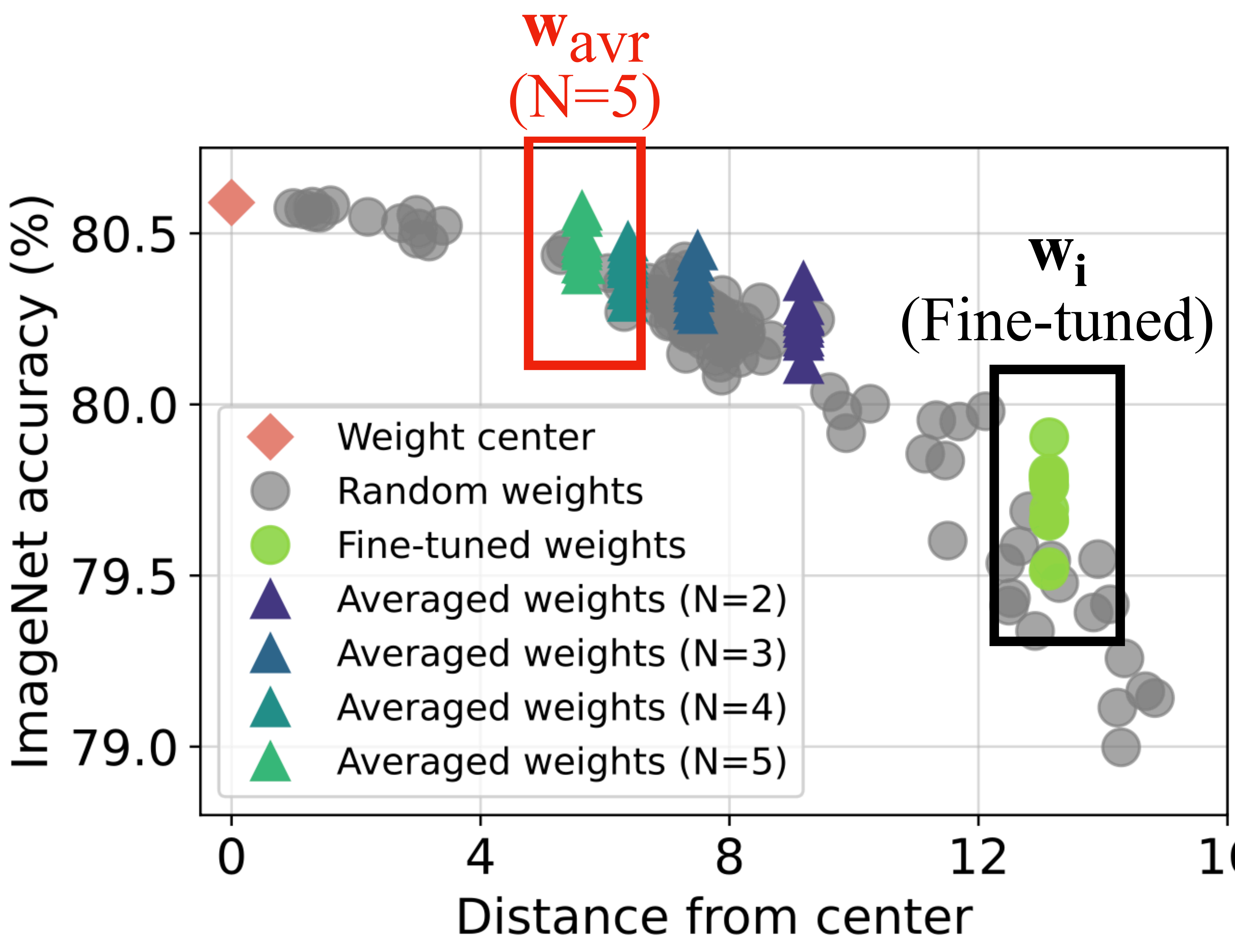
Observation 2: **Distance and Performance**

Distance vs. Random Weights' Performance



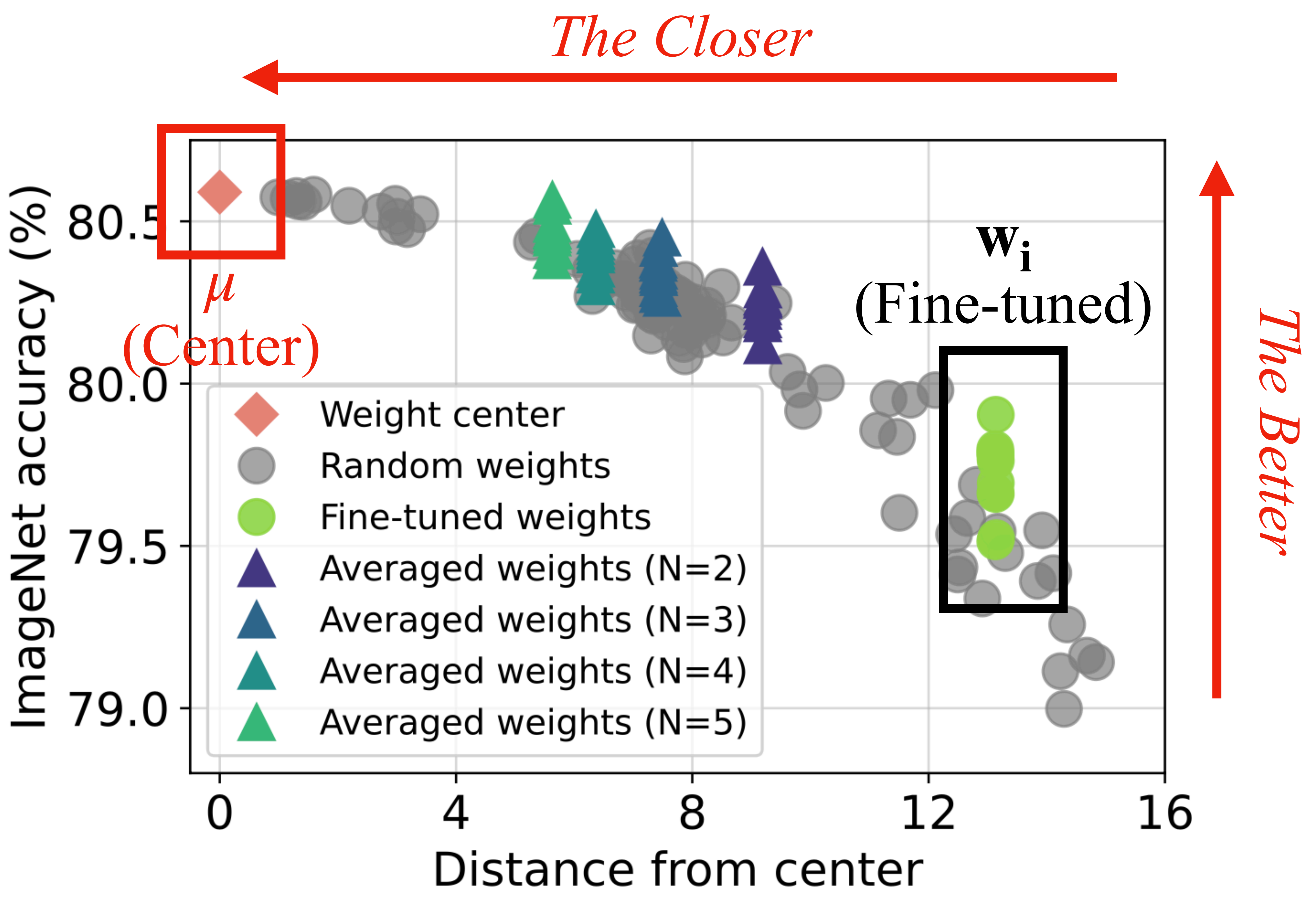
Observation 2: **Distance and Performance**

Distance vs. Random Weights' Performance



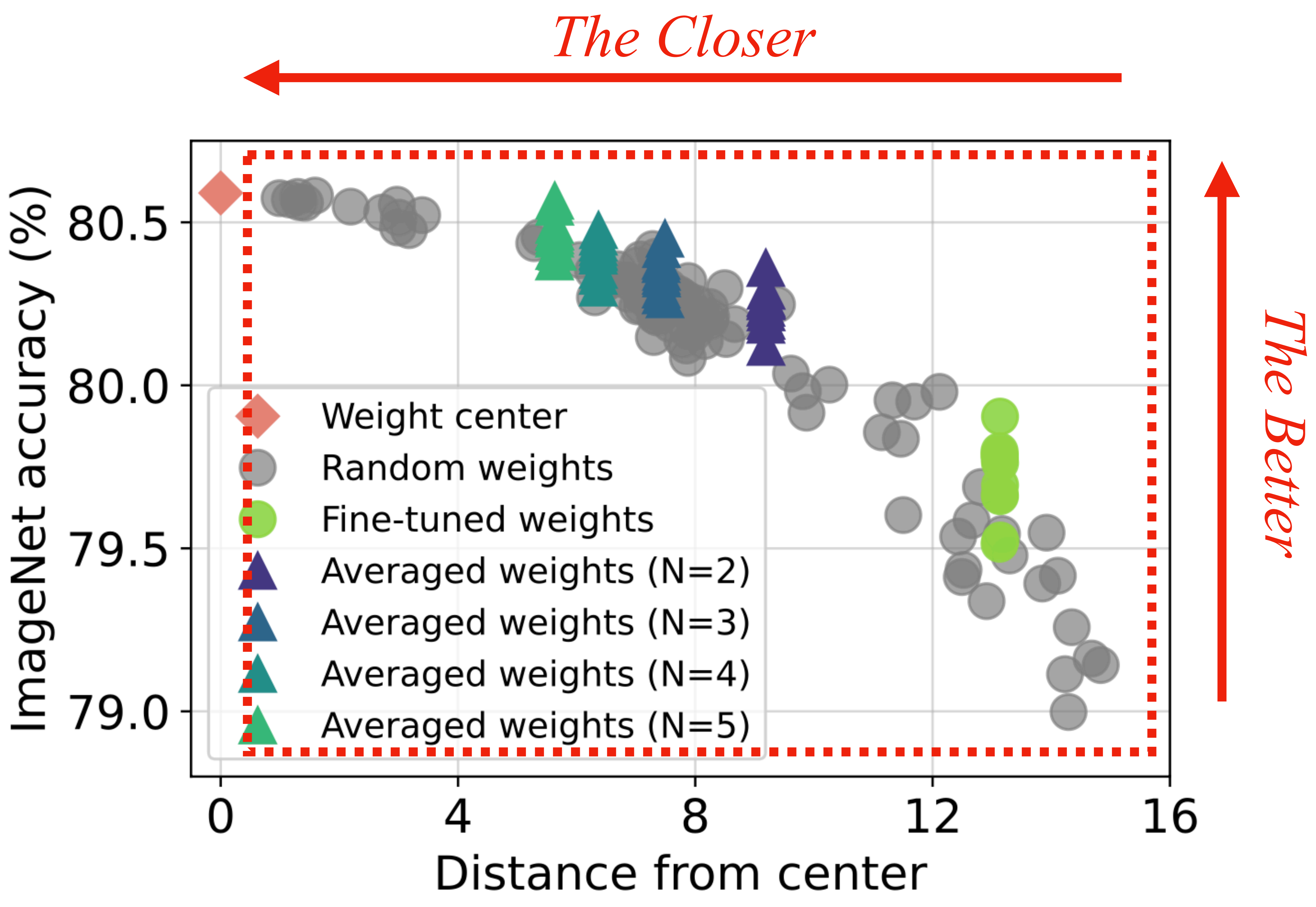
Observation 2: **Distance and Performance**

Distance vs. Random Weights' Performance



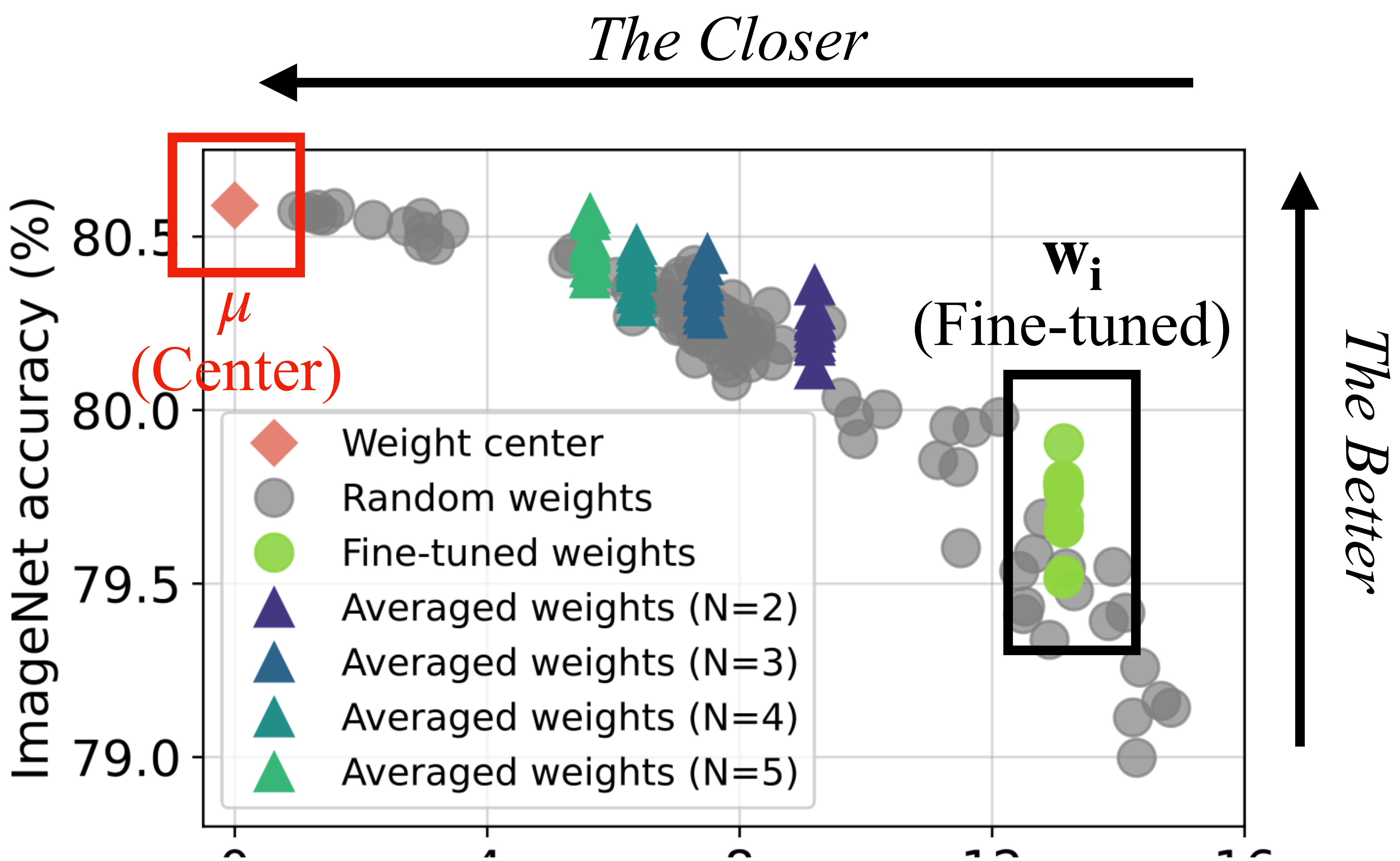
Observation 2: **Distance and Performance**

Distance vs. Random Weights' Performance



Observation 2: **Distance and Performance**

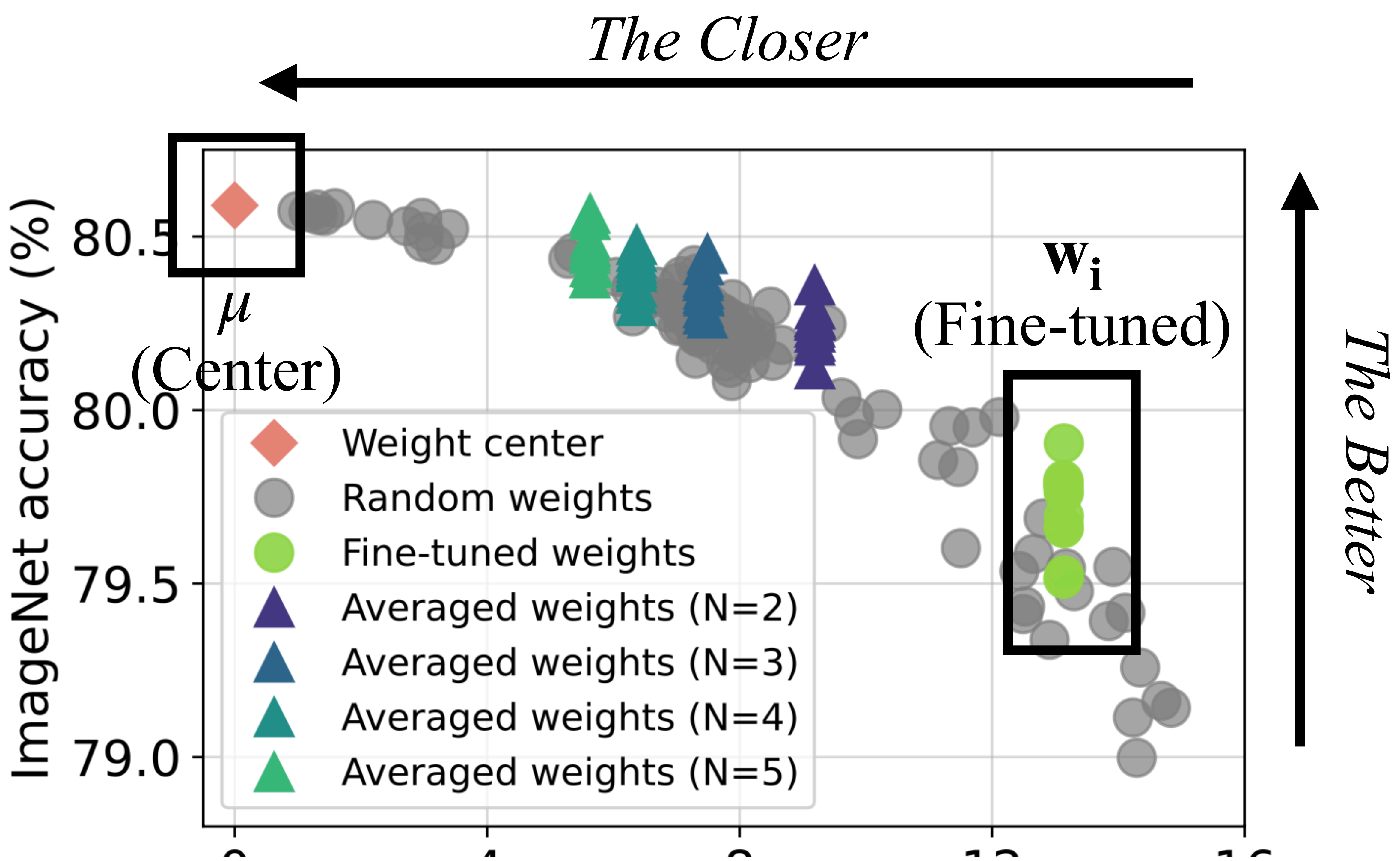
Distance vs. Random Weights' Performance



- Naive averaging is NOT scalable
    - Distance $\propto 1/\sqrt{N}$
- Gradient-based method is NOT reachable

Observation 2: **Distance and Performance**

Distance vs. Random Weights' Performance



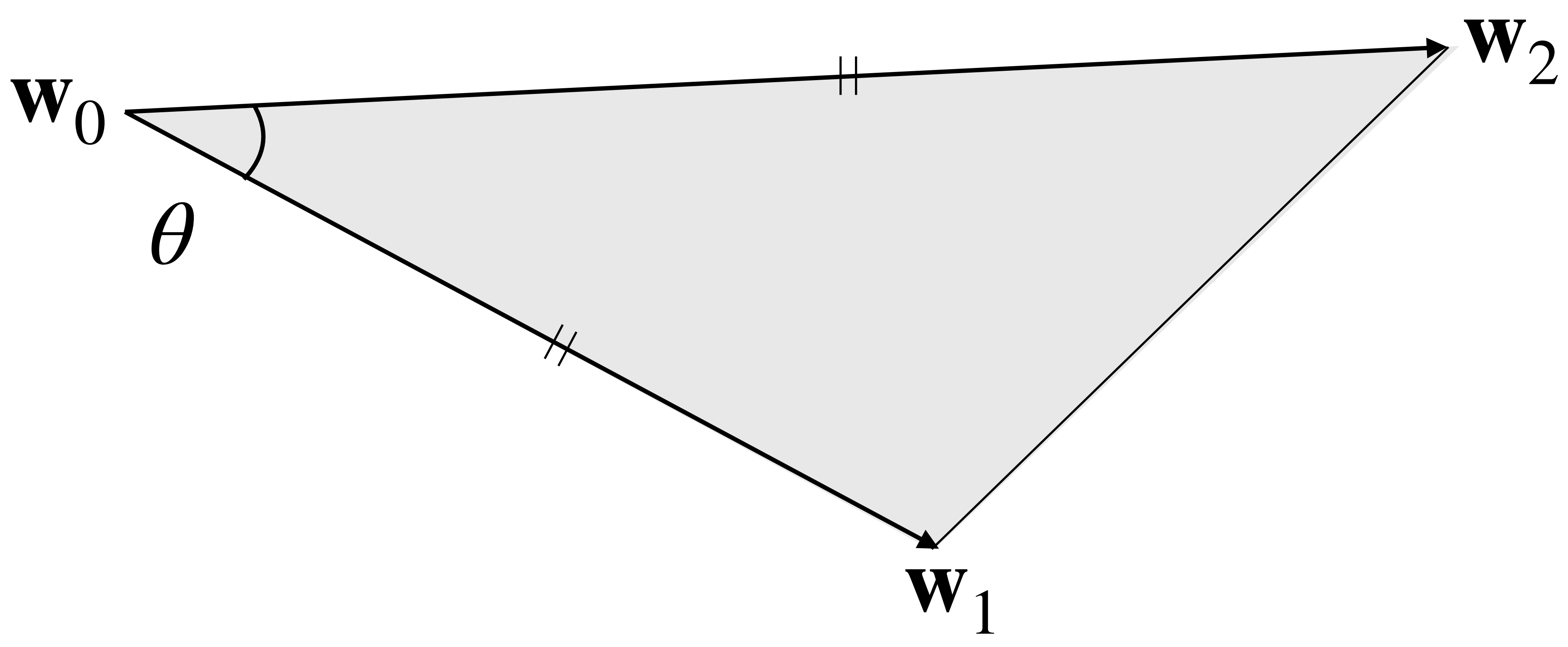
- Naive averaging is NOT scalable
  - Distance $\propto 1/\sqrt{N}$
- Gradient-based method is NOT reachable

Any Better Idea?

Method: ***Find the Closest Weight to the Center using Pre-trained Weight***

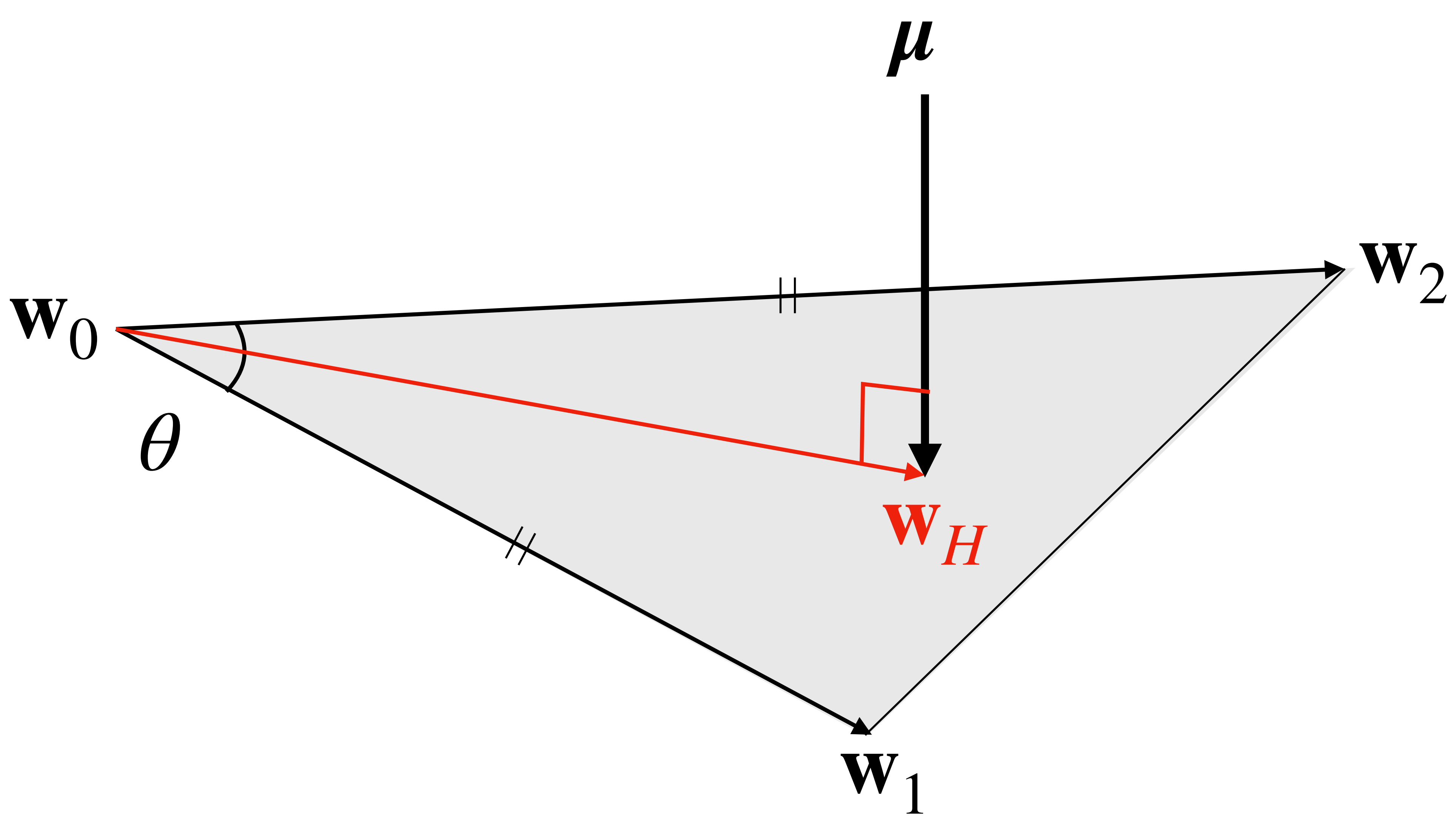
Method: **Model Stock**

N=2 fine-tuned weights



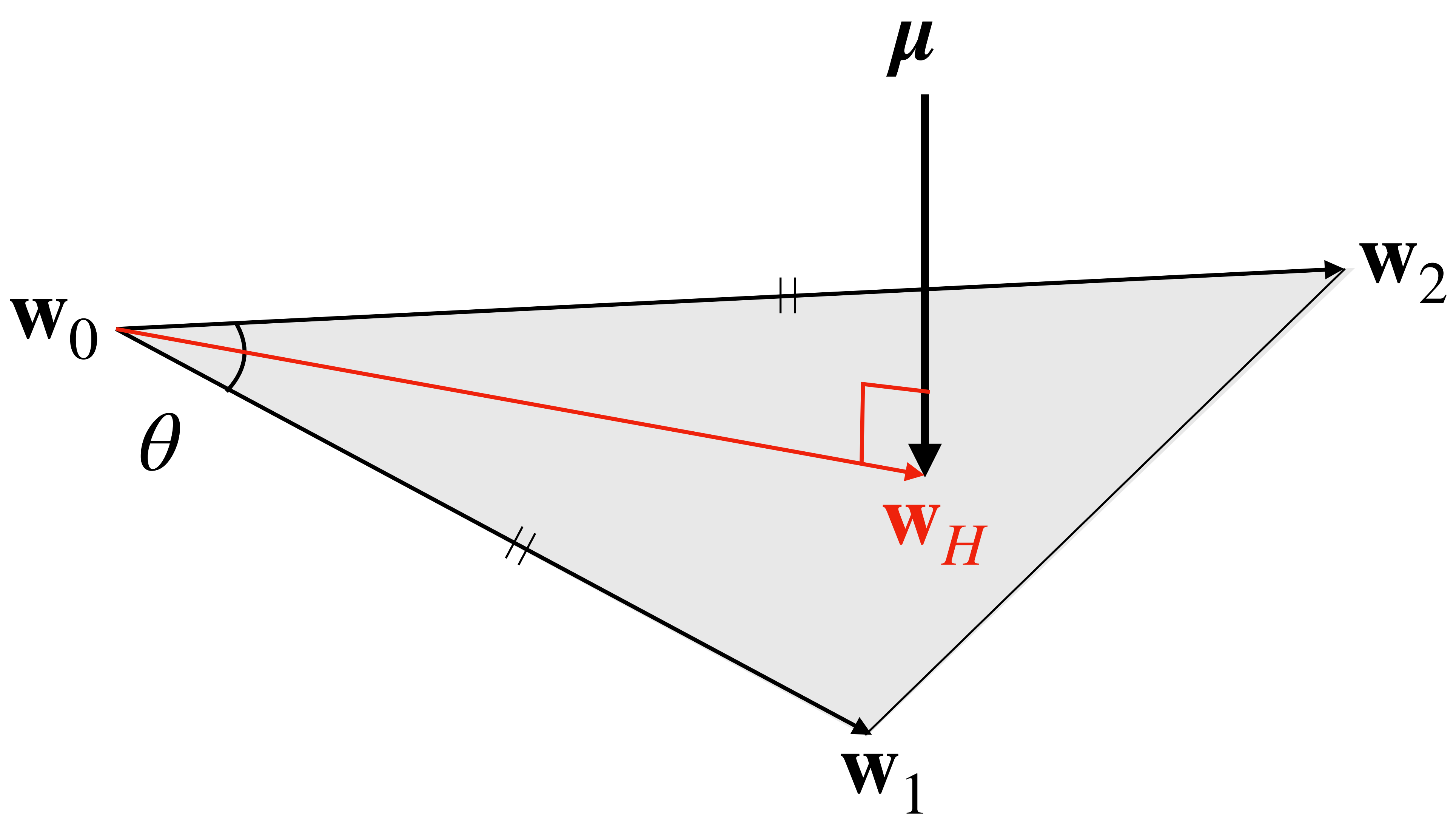
Method: **Model Stock**

N=2 fine-tuned weights



Method: **Model Stock**

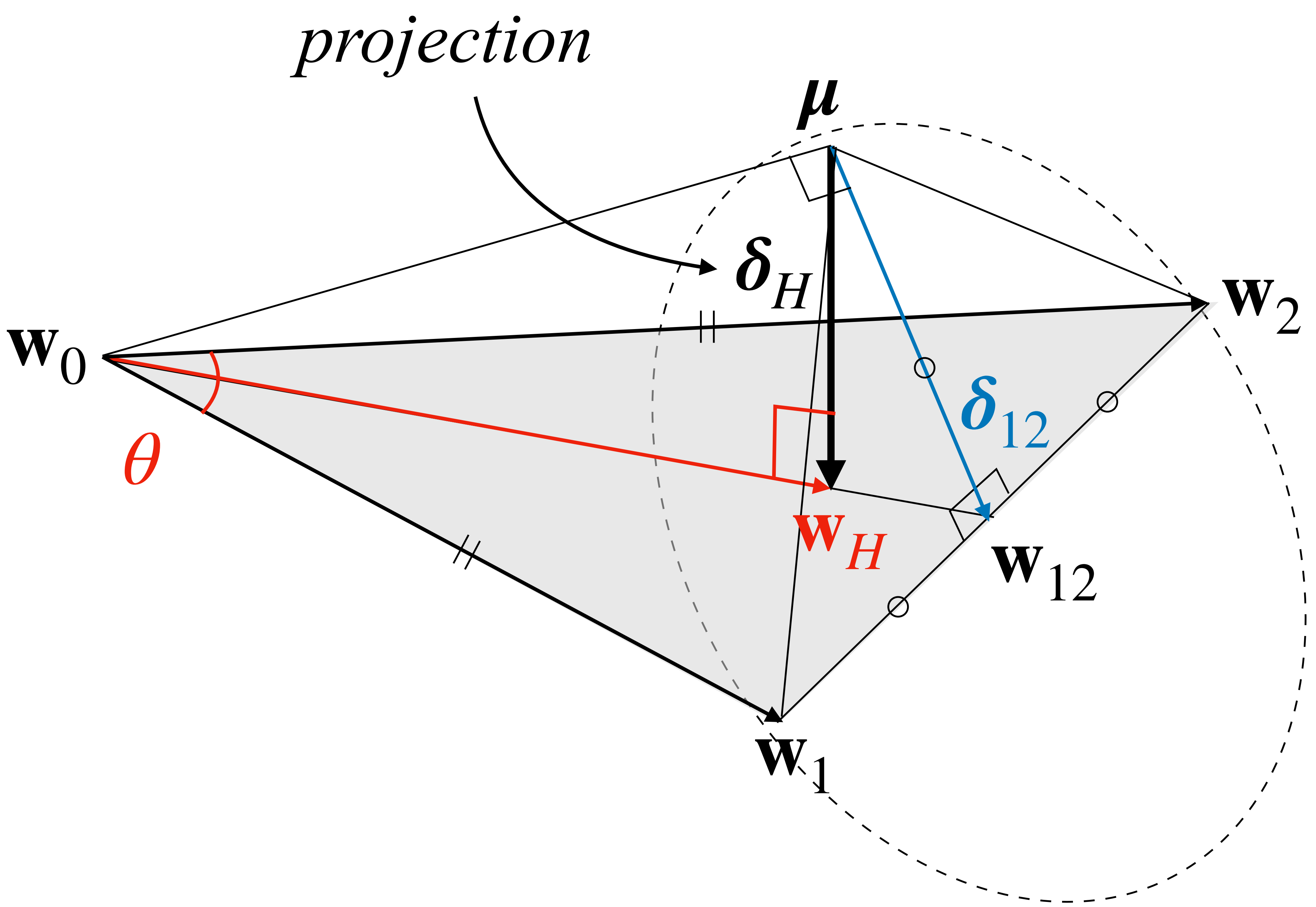
N=2 fine-tuned weights



- Do NOT need exact position of μ
- $\mathbf{w}_H$ is deriven from:
  - (i) $\|\mathbf{w}_i - \boldsymbol{\mu}\| =$ constant *(thin shell)*
  - (iii) $(\mathbf{w}_i - \boldsymbol{\mu}) \perp (\mathbf{w}_j - \boldsymbol{\mu})$
  - (ii) $(\mathbf{w}_0 - \boldsymbol{\mu}) \perp (\mathbf{w}_i - \boldsymbol{\mu})$

Method: **Model Stock**

N=2 fine-tuned weights



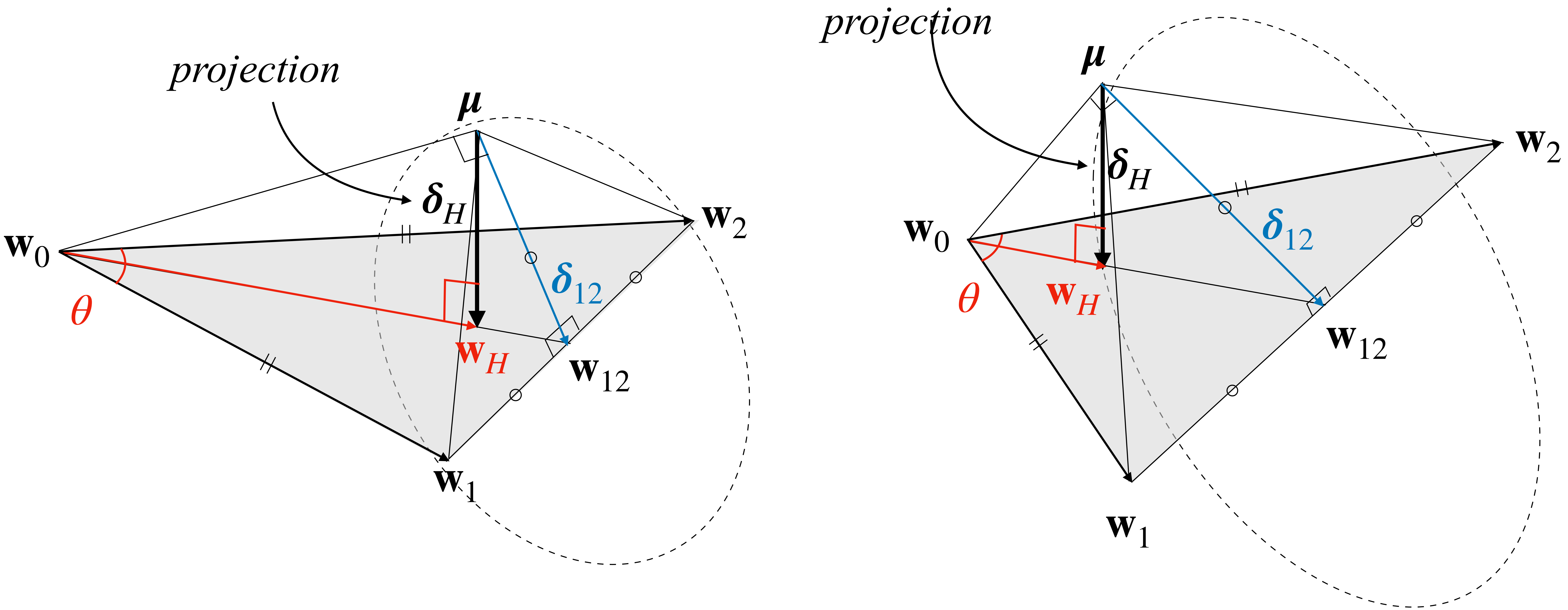
- Do NOT need exact position of μ
- $\mathbf{w}_H$ is deriven from:
  - (i) $\|\mathbf{w}_i - \boldsymbol{\mu}\| = \text{constant}$ *(thin shell)*
  - (iii) $(\mathbf{w}_i - \boldsymbol{\mu}) \perp (\mathbf{w}_j - \boldsymbol{\mu})$
  - (ii) $(\mathbf{w}_0 - \boldsymbol{\mu}) \perp (\mathbf{w}_i - \boldsymbol{\mu})$

$$\mathbf{w}_H = \frac{2\cos\theta}{1+\cos\theta} \cdot \mathbf{w}_{12} + \left(1 - \frac{2\cos\theta}{1+\cos\theta}\right) \cdot \mathbf{w}_0 \quad \text{[layer-wise]}$$

Only depends on θ

Method: **Model Stock**

N=2 fine-tuned weights



Small $\theta \rightarrow \mathbf{w}_1, \mathbf{w}_2 \uparrow$
(e.g., bias layers)

Large $\theta \rightarrow \mathbf{w}_1, \mathbf{w}_2 \downarrow$
(e.g., attention layers)

Method: **Model Stock**

N fine-tuned weights

$$\mathbf{w}_H = \frac{2\cos\theta}{1+\cos\theta} \cdot \mathbf{w}_{12} + \left(1 - \frac{2\cos\theta}{1+\cos\theta}\right) \cdot \mathbf{w}_0$$

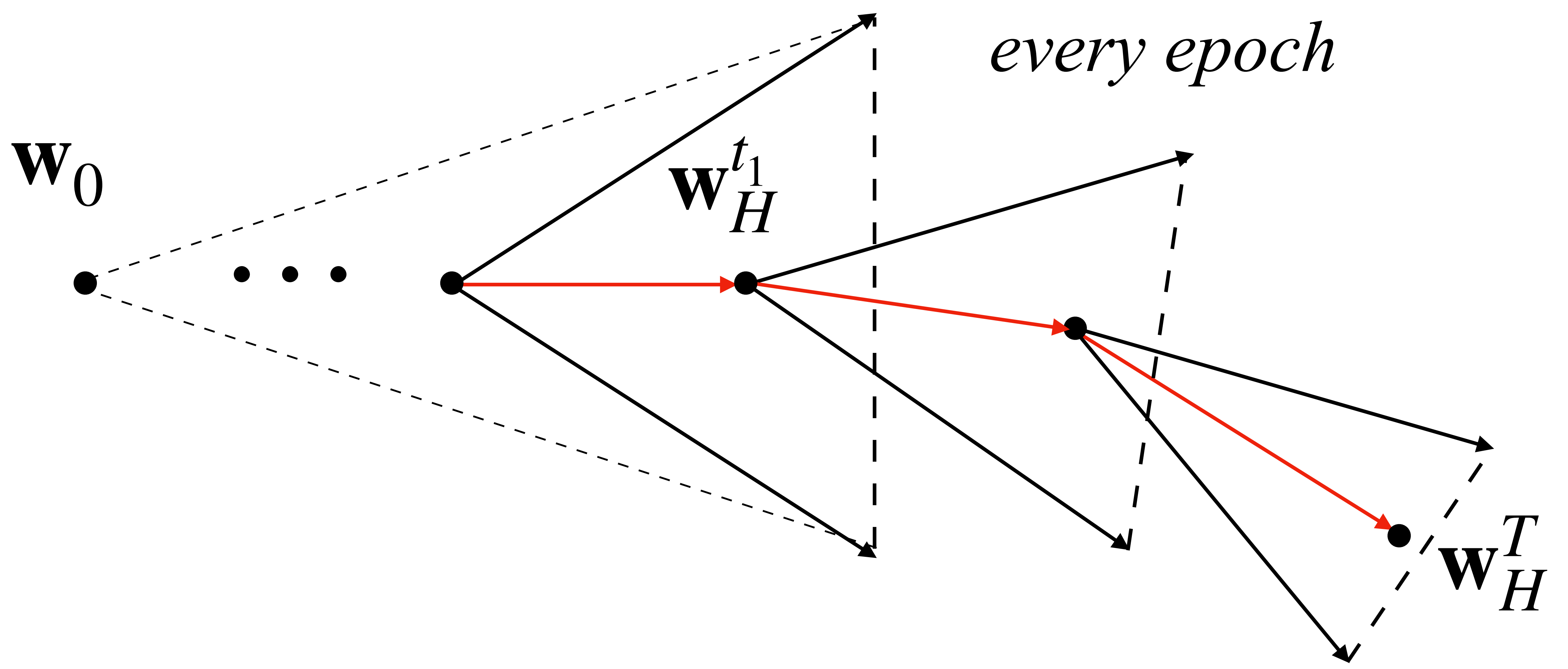
Generalize (merging N models)

$$\mathbf{w}_H^{(N)} = t \cdot \mathbf{w}_{\text{avr}}^{(N)} + (1-t) \cdot \mathbf{w}_0, \qquad \text{s.t.} \quad t = \frac{N\cos\theta}{1+(N-1)\cos\theta}.$$

Method: **Model Stock**

Periodic Merging

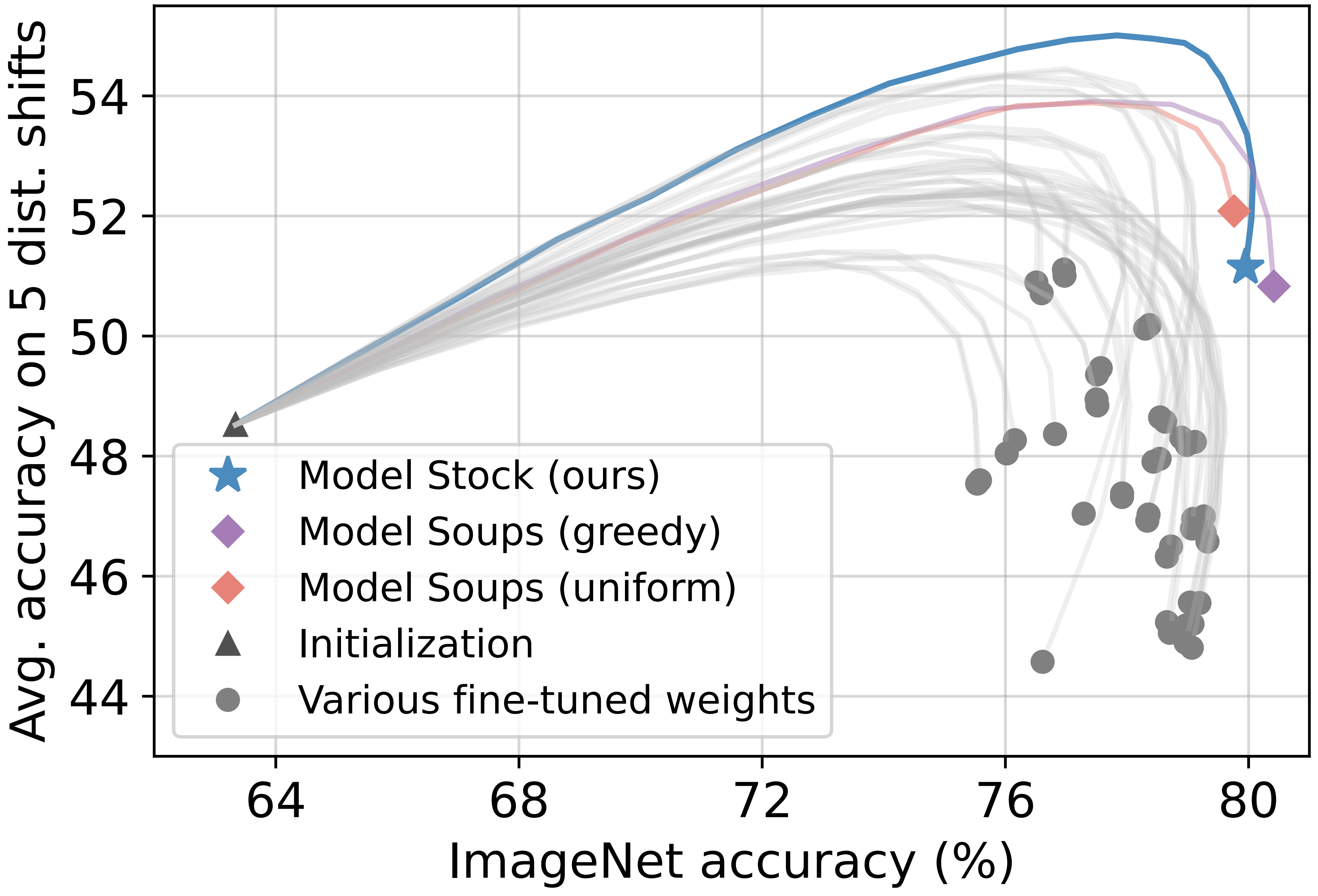
- Leveraging the fact that norm and angle **consistencies hold even during training**, we adopt **periodic merging** to gradually approach the weight center at each epoch.



Experimental Results

Experiments

CLIP ViT-B/32 fine-tuned on ImageNet



Method	ImageNet	Avg. shifts	Cost
Comparing with Model Soups from zero-shot init.			
CLIP zero-shot Initialization	63.34	48.51	0
Vanilla FT	78.35	47.03	1
Uniform Model Soup (from zero-shot)	79.76	**52.08**	48
Greedy Model Soup (from zero-shot)	**80.42**	50.83	48
Model Stock	<u>79.89</u>	<u>50.99</u>	2 ↓
Comparing with Model Soups from LP init.			
CLIP LP initialization	75.57	47.21	α
Vanilla FT*	79.72	46.37	1
Uniform Model Soup (from LP init)	79.97	**51.45**	$71{+}\alpha$
Greedy Model Soup (from LP init)	<u>81.03</u>	<u>50.75</u>	$71{+}\alpha$
Model Stock*	**81.19**	48.69	2 ↓

Experiments

CLIP ViT-B/16 and ViT-L/14 Results

CLIP ViT-B/16

Method	ImageNet	Distribution shifts				
		Avg. shifts	IN-V2	IN-R	IN-A	IN-Sketch
Zero-shot	68.3	59.5	62.0	**77.7**	<u>49.9</u>	48.3
Vanilla FT	82.8	57.7	72.9	66.4	43.7	48.0
Vanilla FT⋆	83.7	57.4	73.5	67.6	40.0	48.6
LP [18]	79.7	48.1	71.5	52.4	27.8	40.5
LP-FT [18]	81.7	<u>60.5</u>	71.6	<u>72.9</u>	49.1	48.4
CAR-FT [27]	83.2	59.4	73.0	71.3	43.7	49.5
FTP [37]	<u>84.2</u>	49.7	74.6	47.2	26.5	50.2
FLYP [7]	82.6	<u>60.5</u>	73.0	71.4	48.1	49.6
Model Stock	84.1	**62.4**	<u>74.8</u>	71.8	**51.2**	**51.8**
Model Stock⋆	**85.2**	60.1	**75.3**	68.7	45.0	<u>51.3</u>

CLIP ViT-L/14

	IN	Avg. shifts
Zero-shot	75.0	63.0
Vanilla FT	85.8	66.8
Vanilla FT⋆	87.1	68.0
TPGM [36]	87.0	69.4
CAR-FT [27]	87.1	67.8
Model Stock	87.0	71.6
Model Stock⋆	**87.7**	**73.5**

Experiments

Post-training Merging

	Uniform averaging ($\mathbf{w}_{\text{avg}}^N$)			Model Stock (post-training)		
	ImageNet	Avg. Shifts	$\|\mathbf{w} - \boldsymbol{\mu}\|$	ImageNet	Avg. Shifts	$\|\mathbf{w} - \boldsymbol{\mu}\|$
N=2	80.2	47.8	9.19	80.3(+0.1)	**50.4(+2.6)**	**7.62(-1.57)**
N=3	80.4	48.2	7.44	80.4(+0.0)	**50.2(+2.0)**	**6.49(-0.95)**
N=4	80.5	48.5	5.63	80.5(+0.0)	**49.8(+1.4)**	**5.16(-0.47)**

Oral 7C / Poster #110

Fri October 4, 08:30 - 10:30 / 10:30 - 12:30 (respectively)

See you at the poster
donghwanjang.github.io

Poster

Project Page