JOHANNES KEPLER
UNIVERSITÄT LINZ

Loss Landscapes under Distribution Shift

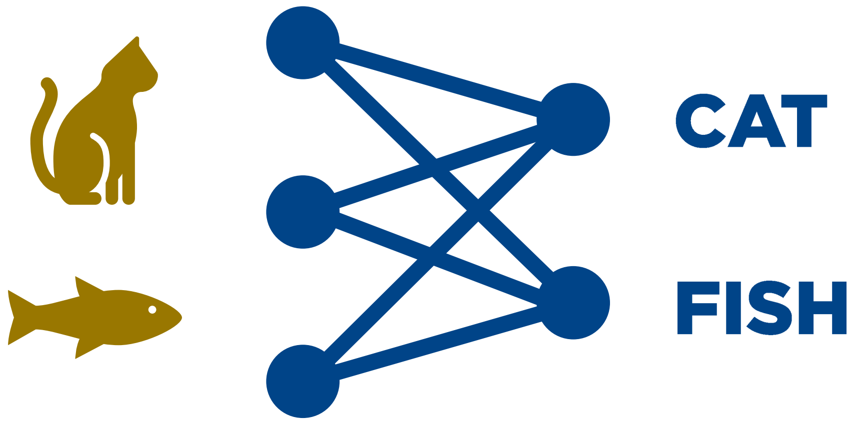
PhD Seminar Talk

Maximilian Beck, beck@ml.jku.at, maxmbeck

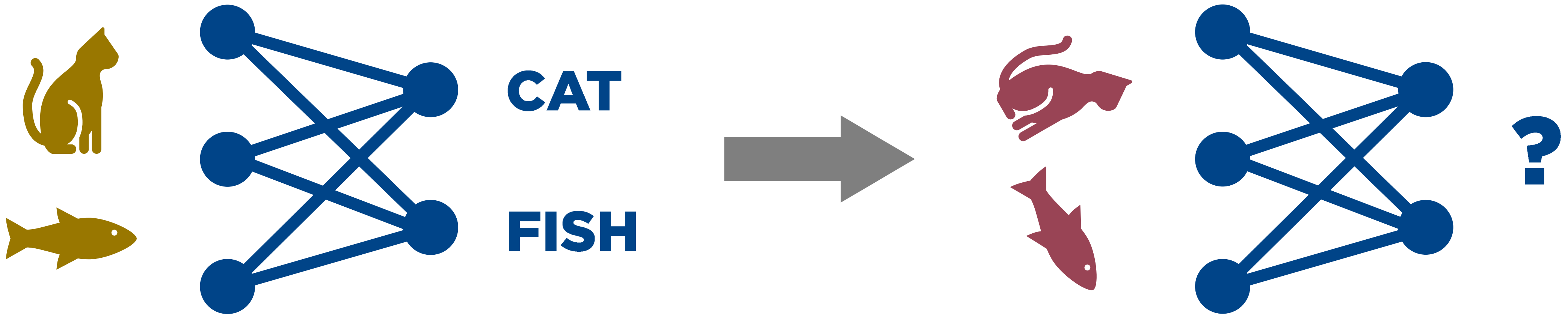
Joint work with Sebastian Lehner and Sepp

Institute for Machine Learning, November 2022

Challenge with Neural Networks

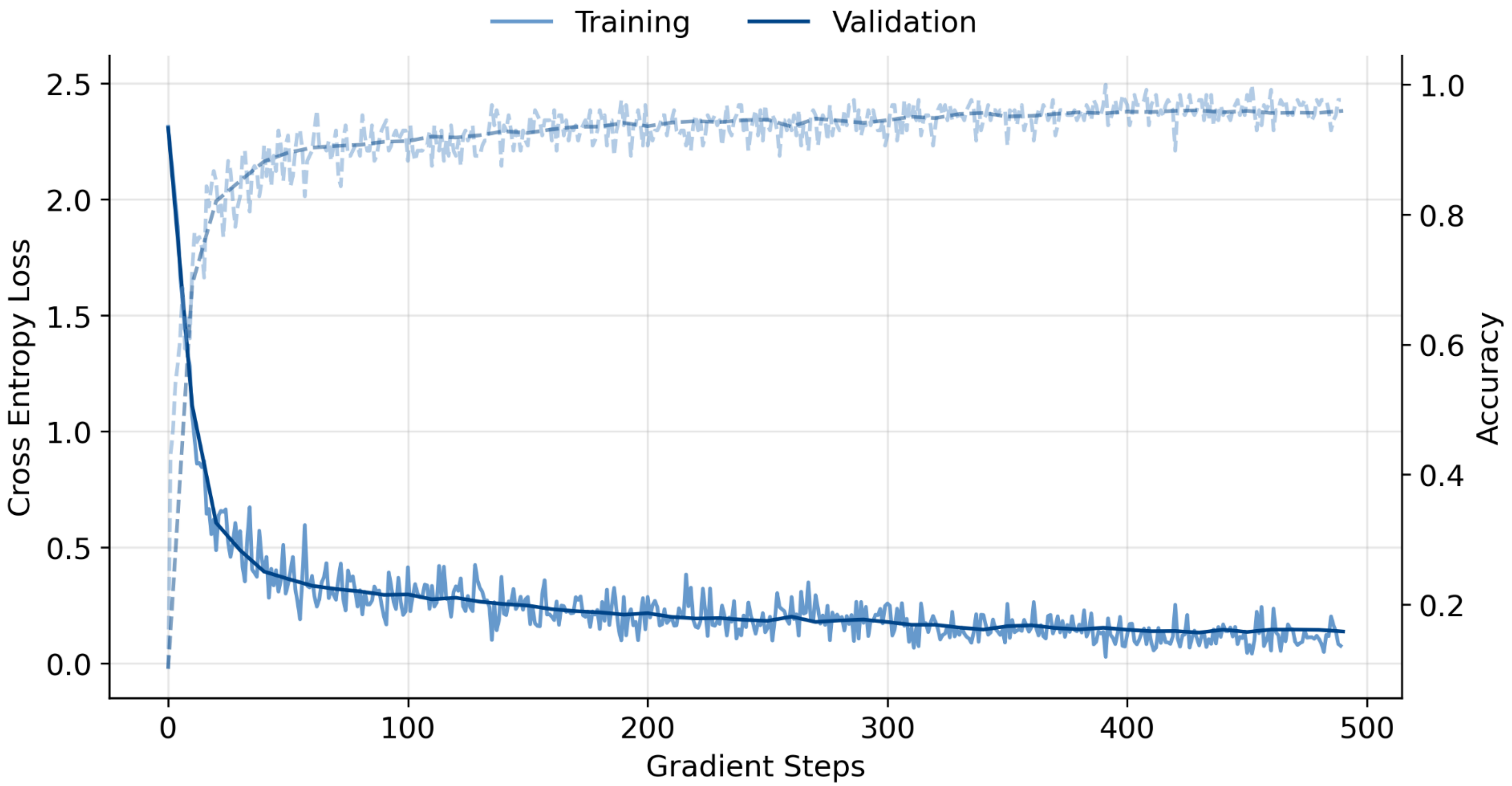


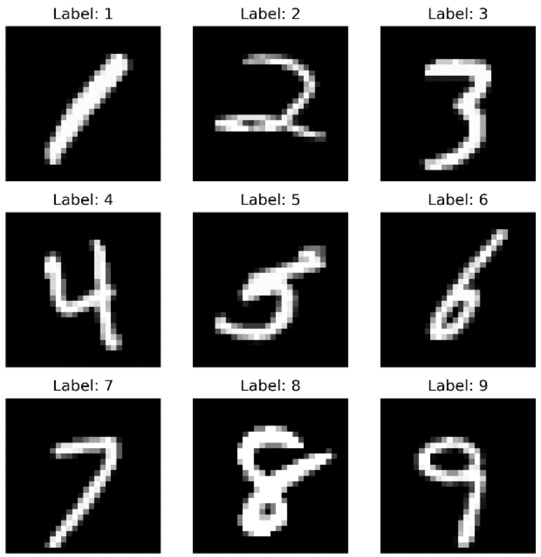
Challenge with Neural Networks



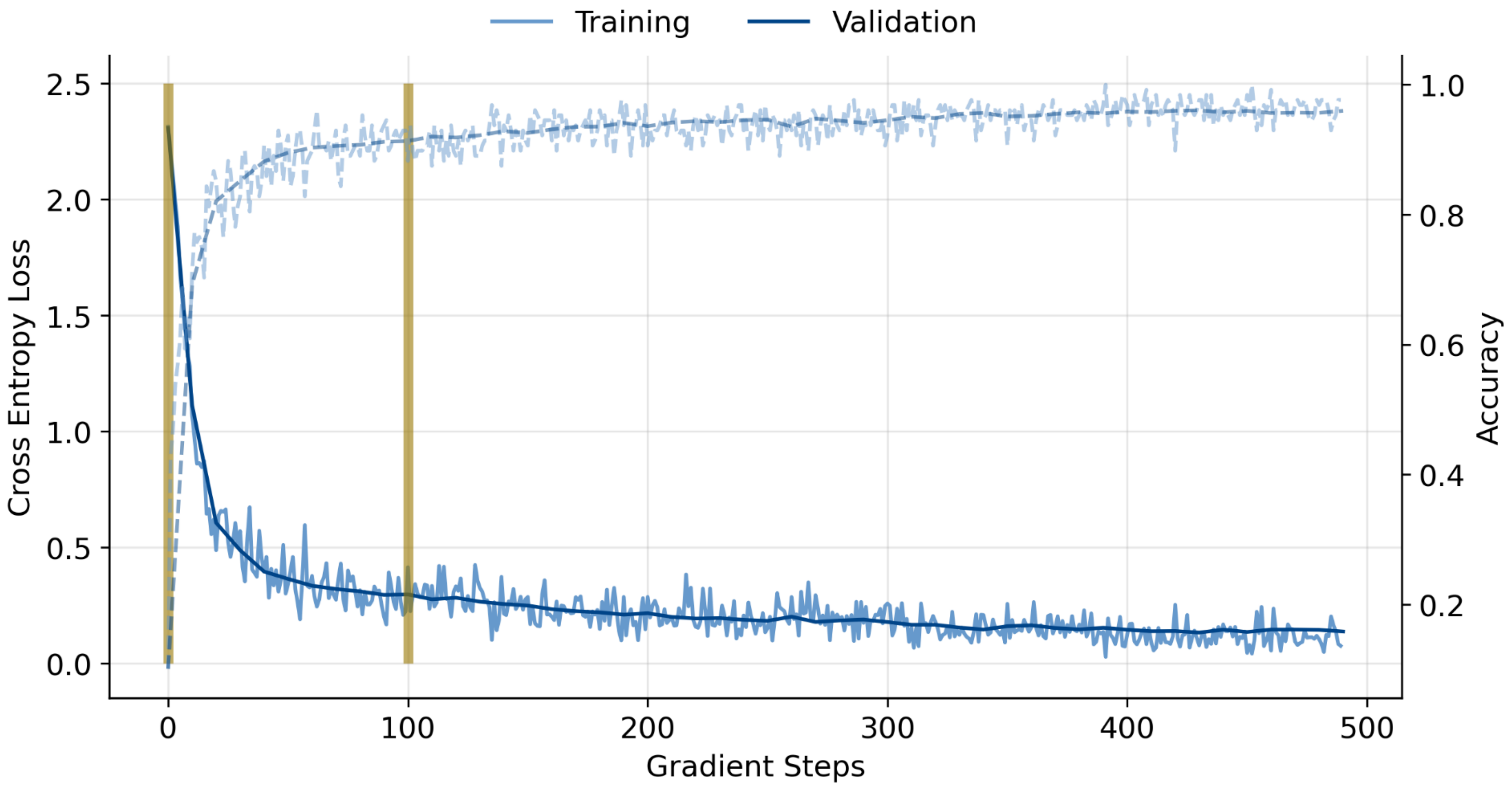
Distribution Shift

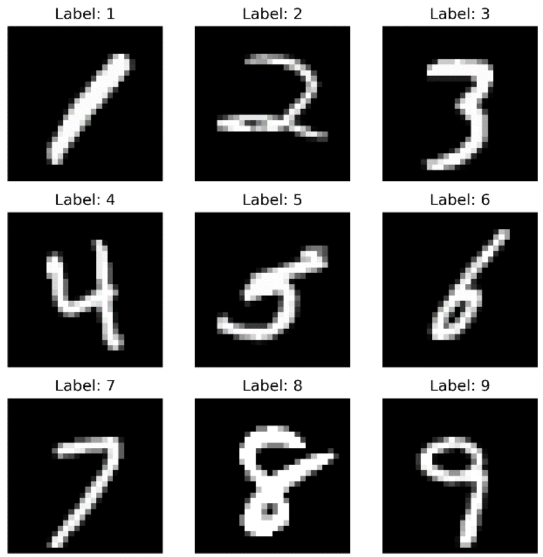
Toy Example: Rotated MNIST – Pretraining



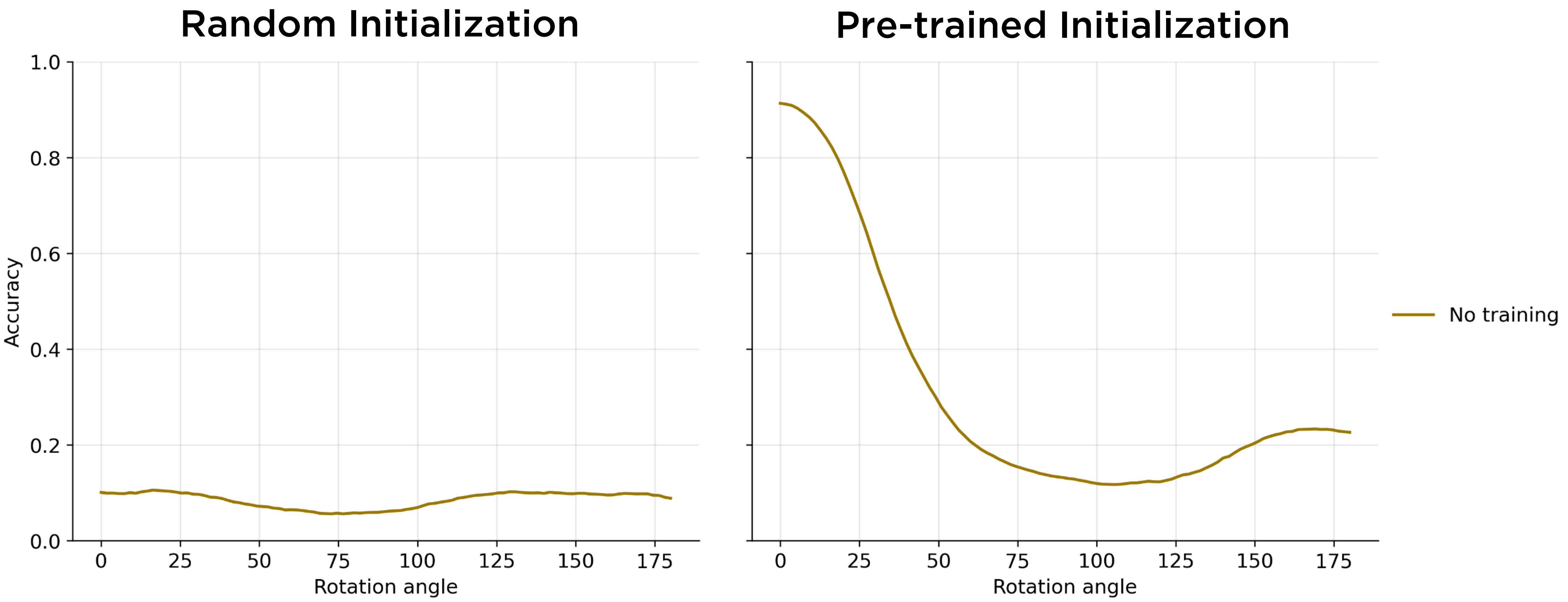


Toy Example: Rotated MNIST – Pretraining





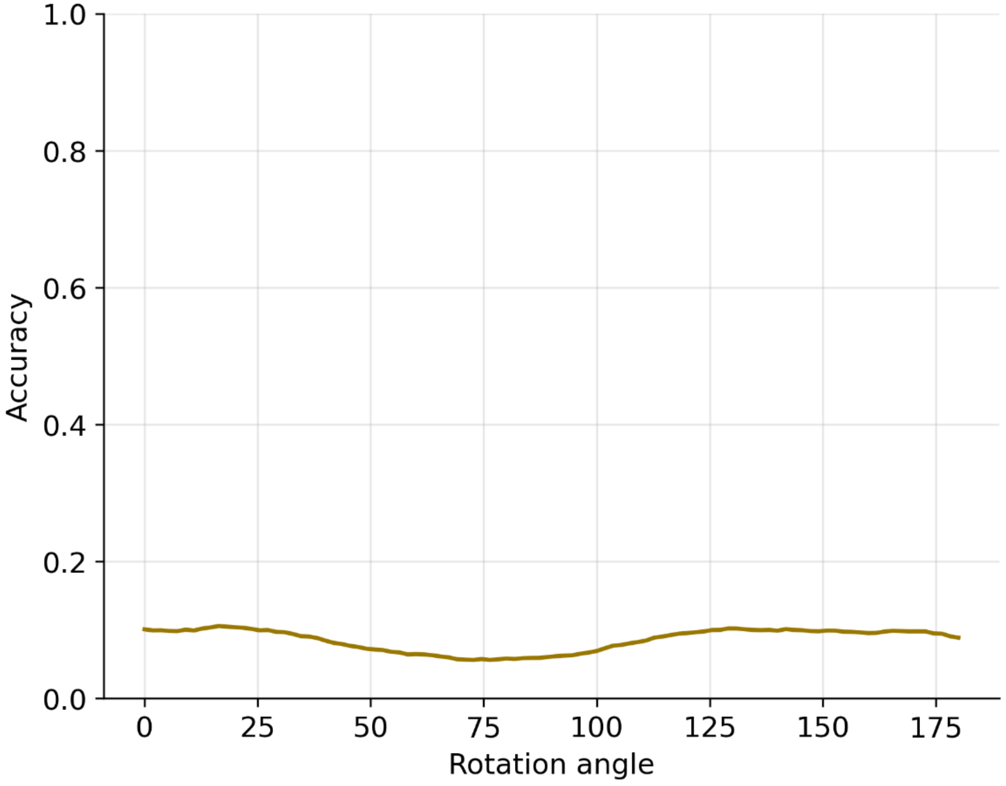
Toy Example: Rotated MNIST – Finetuning



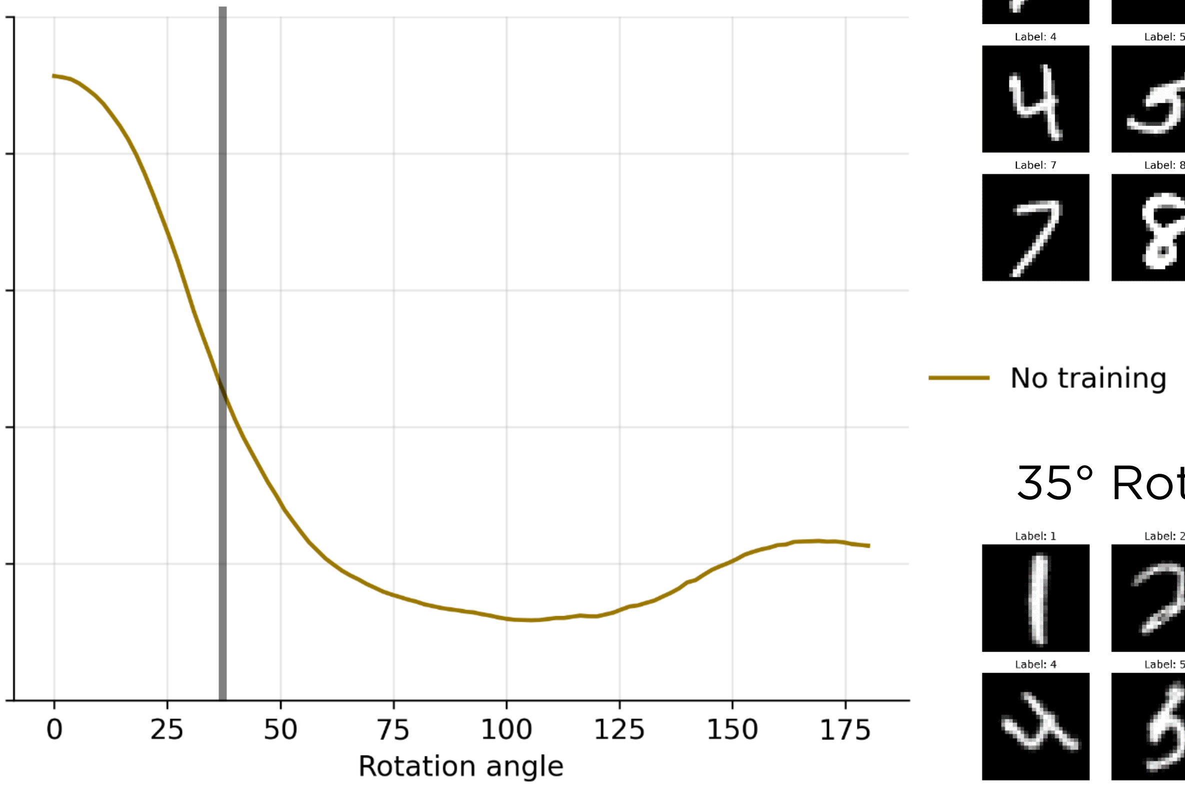
100 Gradient Steps on 0° Rotation

Toy Example: Rotated MNIST – Finetuning

Random Initialization

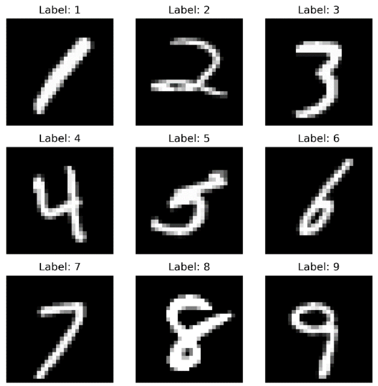


Pre-trained Initialization

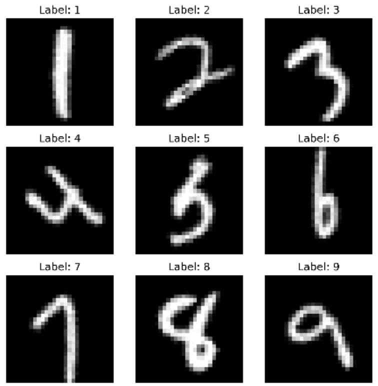


100 Gradient Steps on 0° Rotation

0° Rotation

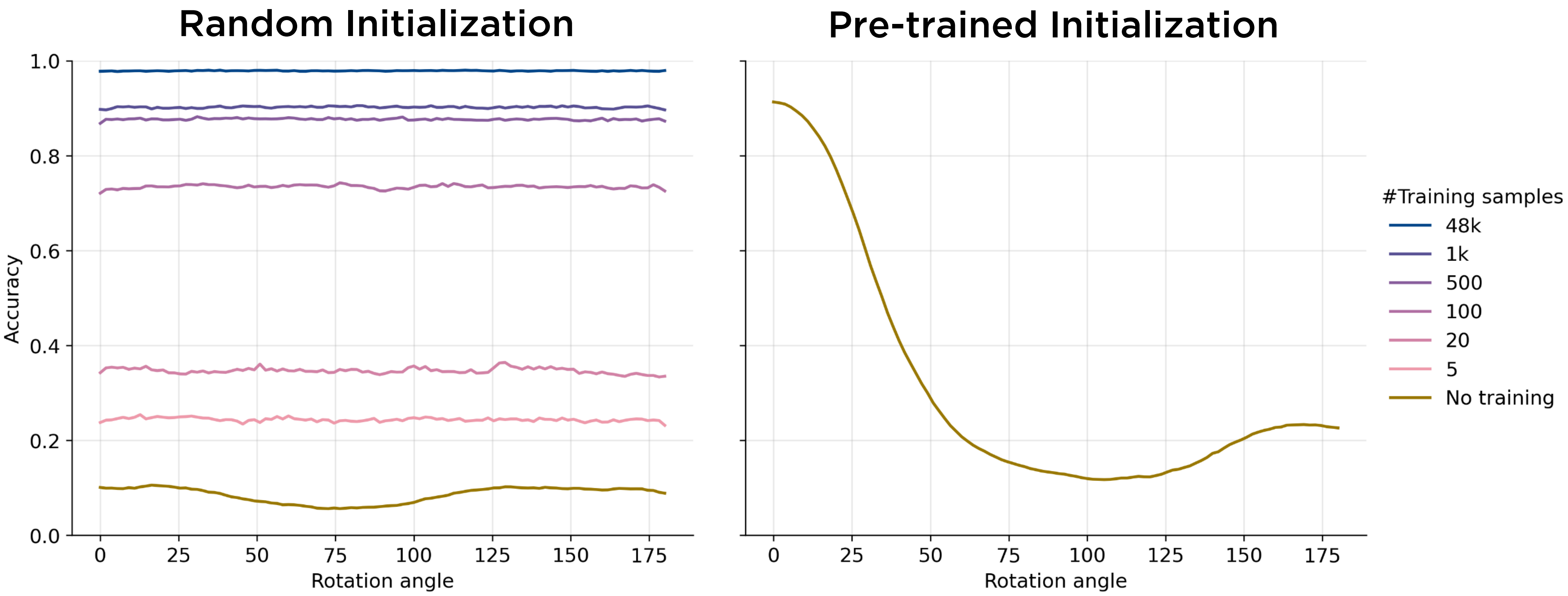


35° Rotation



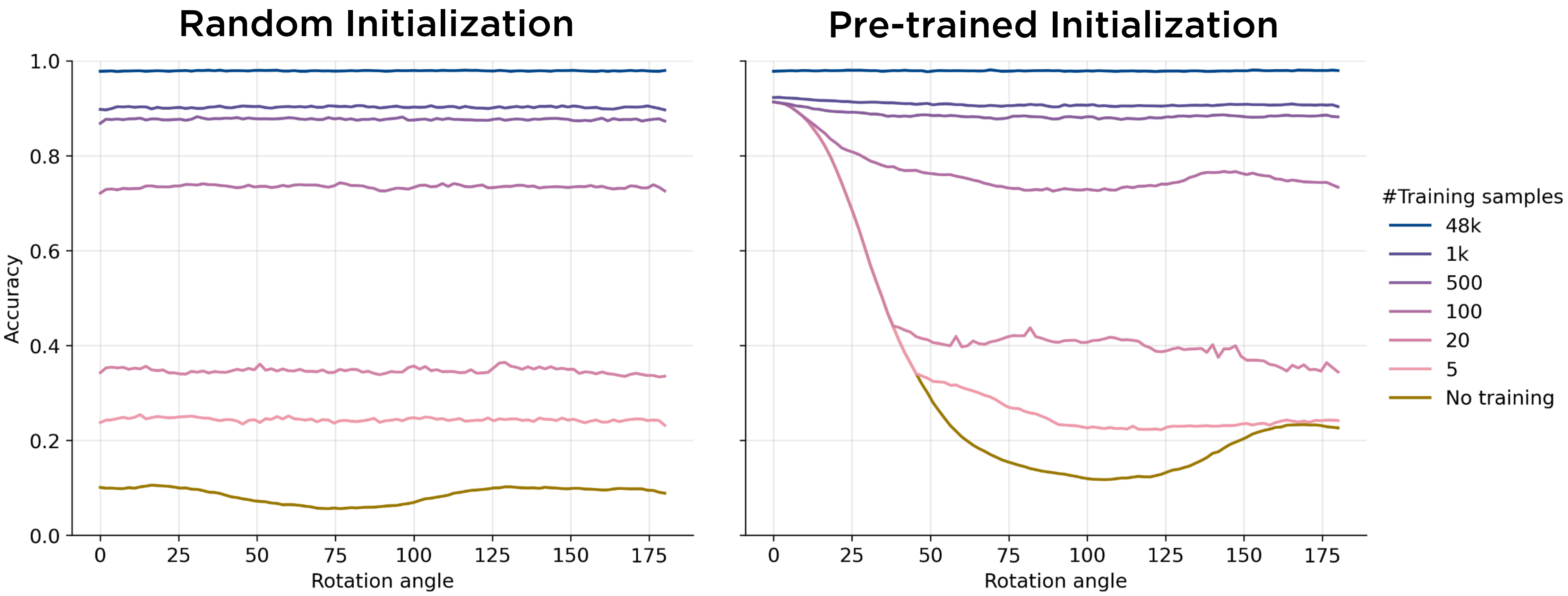
JKU

Toy Example: Rotated MNIST – Finetuning



100 Gradient Steps on 0° Rotation

Toy Example: Rotated MNIST – Finetuning



100 Gradient Steps on 0° Rotation

JKU

What I am excited about

What I am excited about

Learning from Multiple Distributions

- Few-shot Learning
- Domain Adaptation
- Transfer Learning

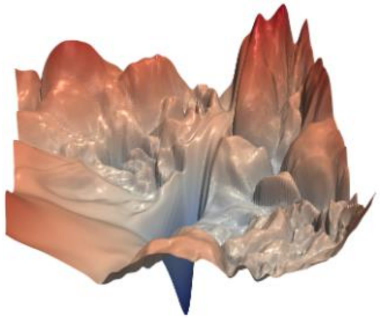
JKU

What I am excited about

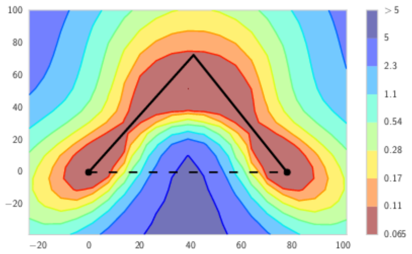
Learning from Multiple Distributions

- Few-shot Learning
- Domain Adaptation
- Transfer Learning

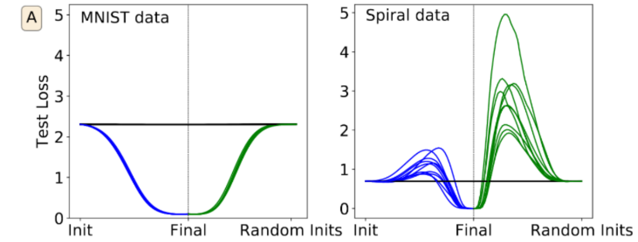
Empirical Phenomena of Deep Learning



Li et al., 2018



Garipov et al., 2018



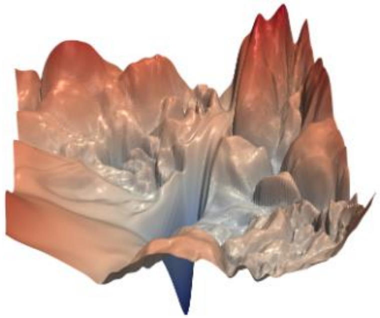
Vlaar and Frankle, 2022

What I am excited about

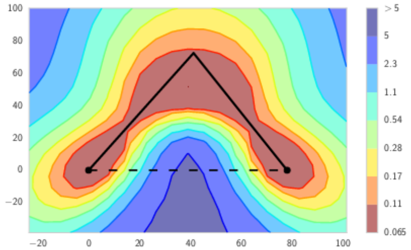
Learning from Multiple Distributions

- Few-shot Learning
- Domain Adaptation
- Transfer Learning

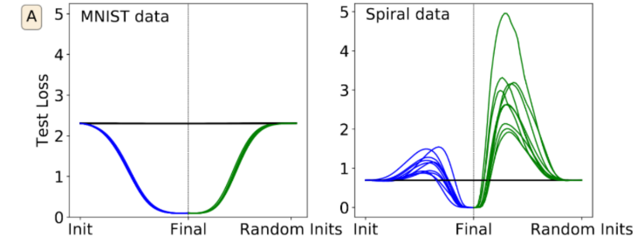
Empirical Phenomena of Deep Learning



Li et al., 2018



Garipov et al., 2018



Vlaar and Frankle, 2022

Outline

- Transfer Learning: Setting and Challenges
- Empirical findings about loss landscapes
- Our approach:
  Use loss landscape information for improved fine-tuning

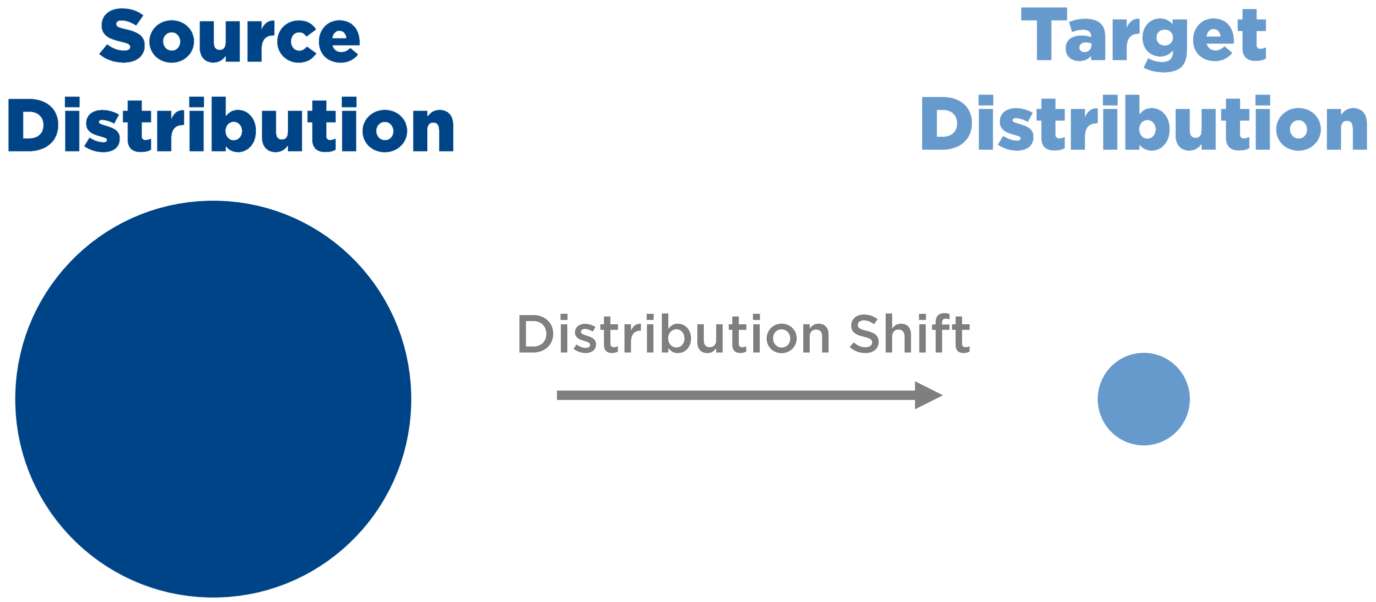
Transfer Learning – Setting and Challenges

Transfer Learning – Setting

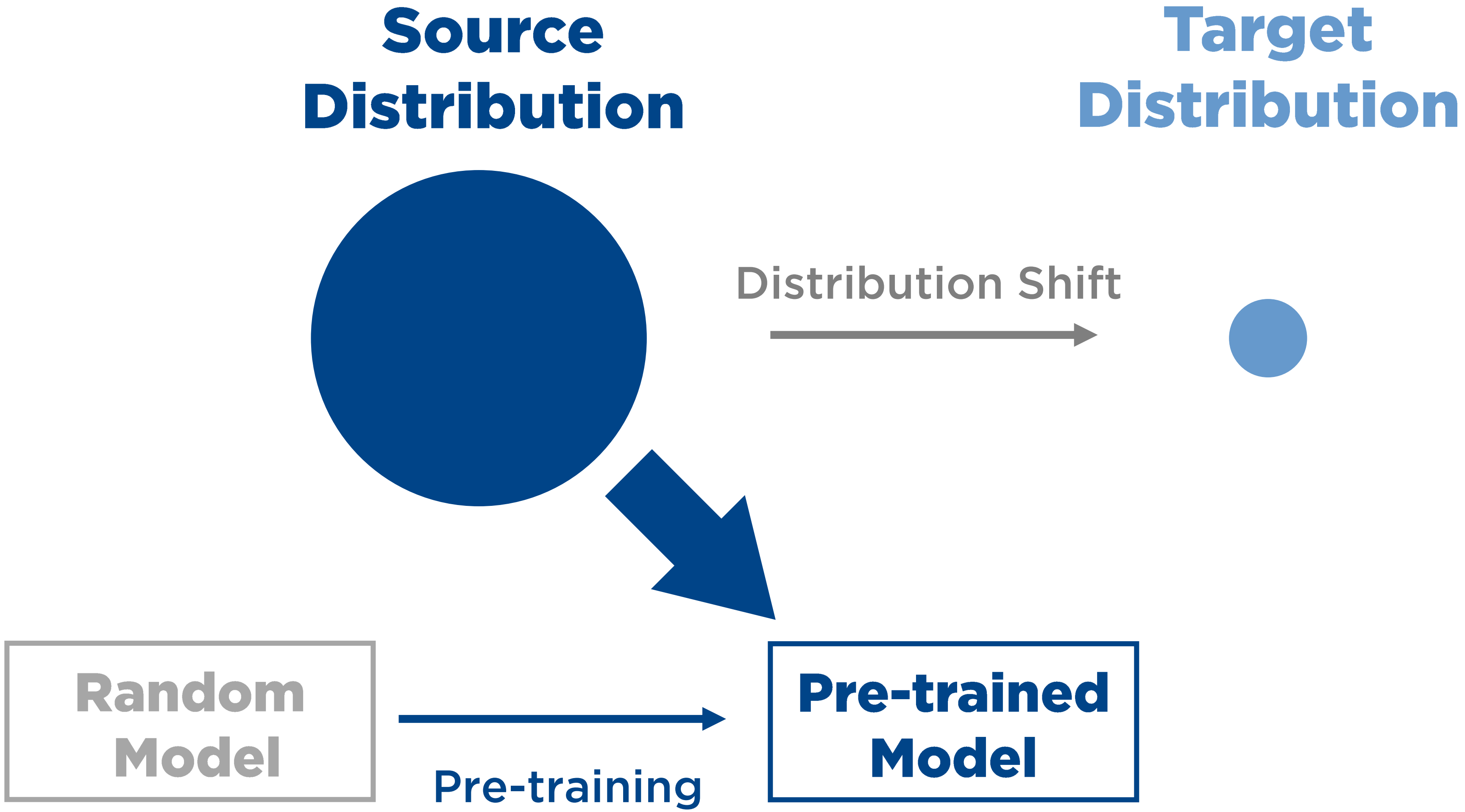
Source Distribution

Target Distribution

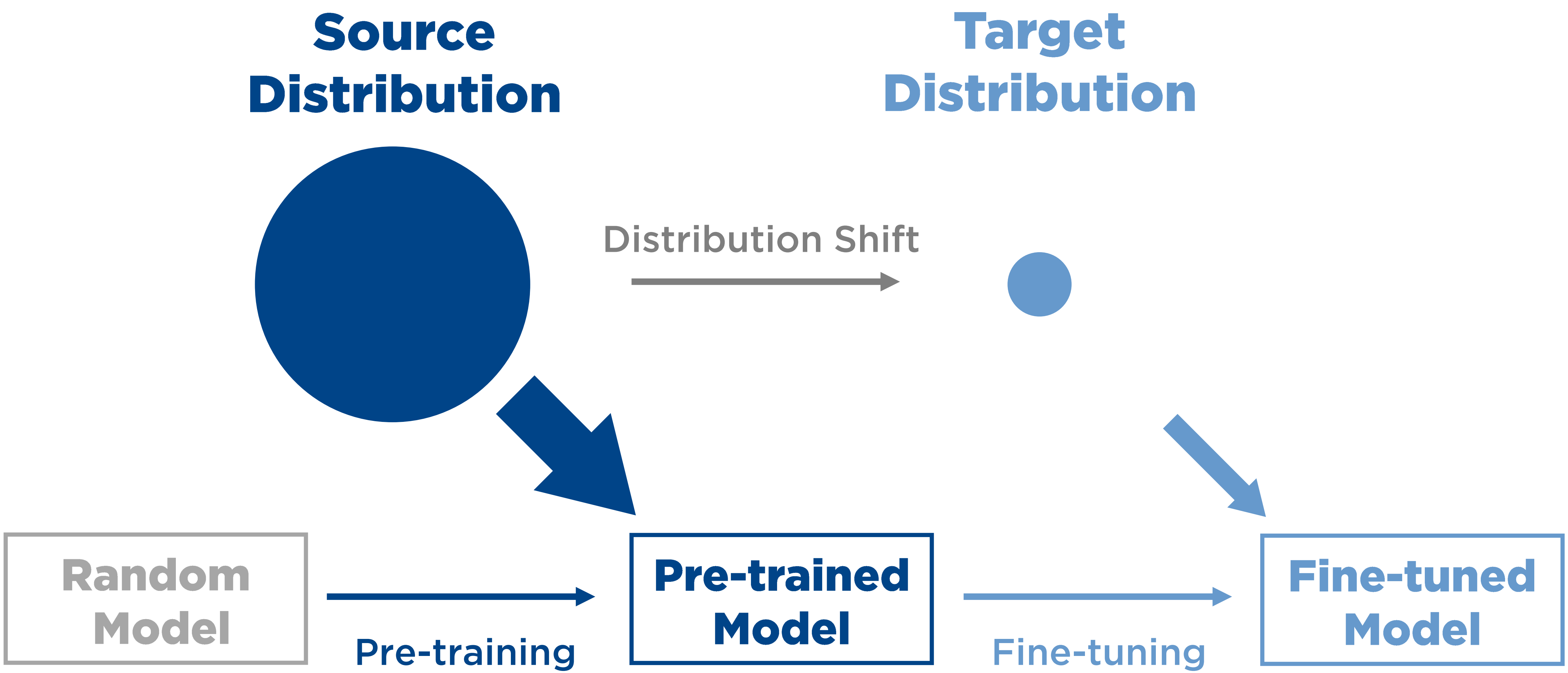
Transfer Learning – Setting



Transfer Learning – Setting



Transfer Learning – Setting



Transfer Learning – Approaches and Challenges

- **Standard Approaches:**
  - Fine-tuning: Gradient descent on all network parameters
  - Linear Probing: Tuning the head but freezing lower layers

Transfer Learning – Approaches and Challenges

- **Standard Approaches:**
  - Fine-tuning: Gradient descent on all network parameters
  - Linear Probing: Tuning the head but freezing lower layers

- **Challenges:**
  - Distribution shift between source and target distributions [Koh et al., 2022]
  - Spurious correlations in training datasets [Kirichenko et al. 2022]
  - Fine-tuning can distort pre-trained features [Kumar et al. 2022]

Transfer Learning – (Some) Recent Works

- **Recently proposed approaches:**
  - *LP-FT:* First Linear Probing then full Fine-tuning [Kumar et al., 2022]

Transfer Learning – (Some) Recent Works

- **Recently proposed approaches:**
  - *LP-FT:* First Linear Probing then full Fine-tuning [Kumar et al., 2022]
  - *Surgical fine-tuning:* Fine-tuning only a small contiguous subset of all layers [Lee et al., 2022]

Transfer Learning – (Some) Recent Works

- **Recently proposed approaches:**
  - *LP-FT:* First Linear Probing then full Fine-tuning [Kumar et al., 2022]
  - *Surgical fine-tuning:* Fine-tuning only a small contiguous subset of all layers [Lee et al., 2022]
  - *Deep feature reweighting:* Last-layer retraining on a small dataset without any spurious correlations [Kirichenko et al., 2022]

Transfer Learning – (Some) Recent Works

- **Recently proposed approaches:**
  - *LP-FT:* First Linear Probing then full Fine-tuning [Kumar et al., 2022]
  - *Surgical fine-tuning:* Fine-tuning only a small contiguous subset of all layers [Lee et al., 2022]
  - *Deep feature reweighting:* Last-layer retraining on a small dataset without any spurious correlations [Kirichenko et al., 2022]

Bottom line:

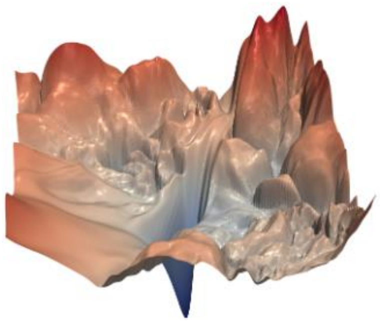
These methods add to a growing evidence in the literature that lightweight fine-tuning, where only a small part of a pre-trained model are updated, can perform better under distribution shifts.

Empirical findings about loss landscapes

Analyzing the loss landscape of Neural Networks

Analyzing the loss landscape of Neural Networks

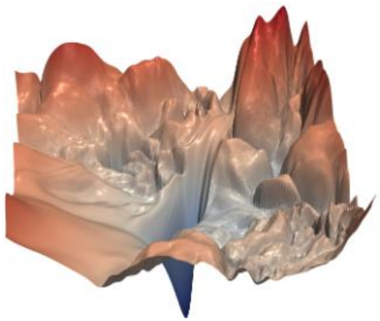
Visualization of the loss landscape



Li et al., 2018

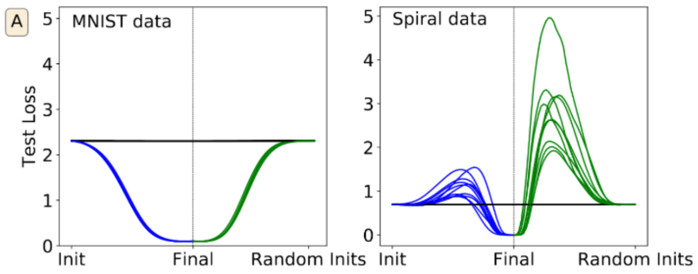
Analyzing the loss landscape of Neural Networks

Visualization of the loss landscape



Li et al., 2018

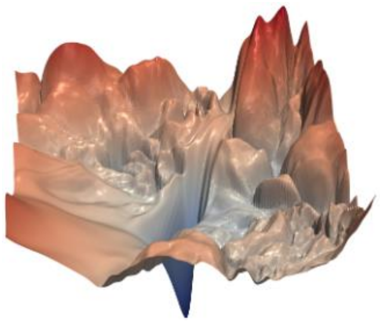
Interpolation between initial and final model states



Vlaar and Frankle, 2022

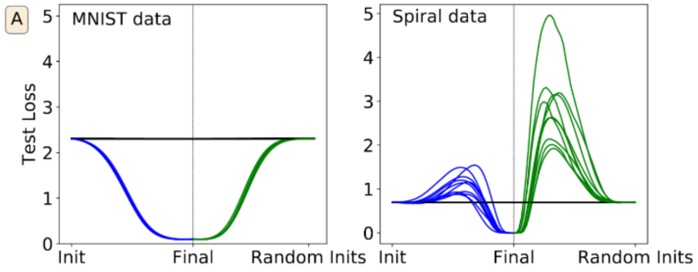
Analyzing the loss landscape of Neural Networks

Visualization of the loss landscape



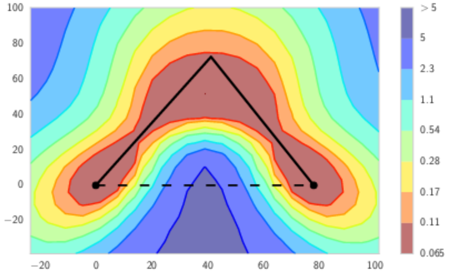
Li et al., 2018

Interpolation between initial and final model states



Vlaar and Frankle, 2022

Interpolation between different optima



Garipov et al., 2018

Focus in this talk!

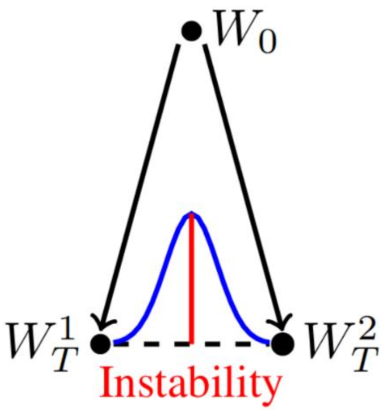
Instability Analysis of SGD a la Frankle et al., 2020

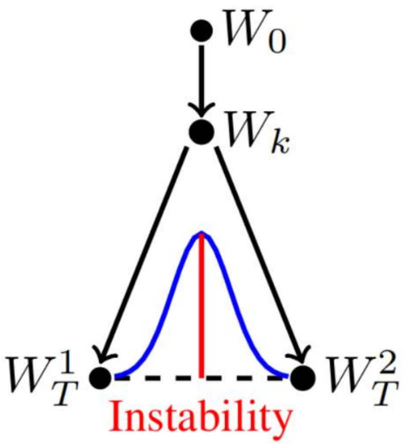
- **Goal:**
  Determine whether the outcome of optimizing a particular network N is stable to SGD noise

Instability Analysis of SGD a la Frankle et al., 2020

- **Goal:**
  Determine whether the outcome of optimizing a particular network N is stable to SGD noise

- **Procedure:**
  - Make **two copies of N** and **train** them with **different random samples** of SGD noise
  - **Compare these weights** with a **function** to produce a value called **instability of N**





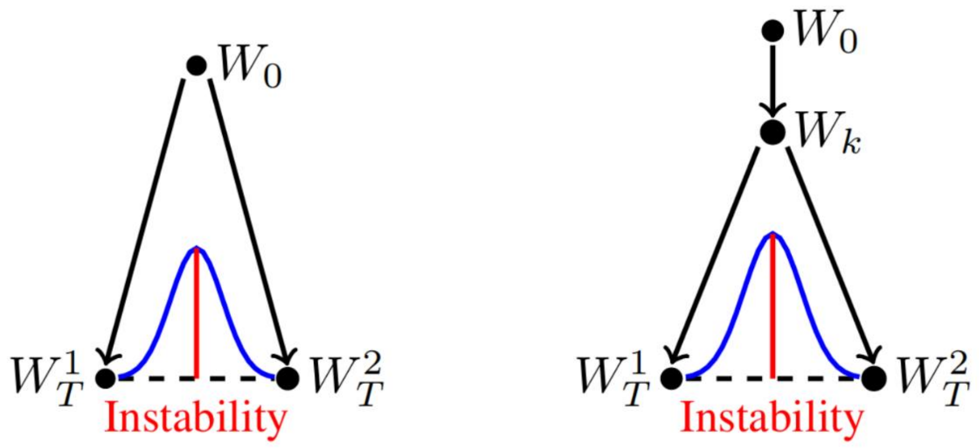
Frankle et al., 2020

Instability Analysis of SGD a la Frankle et al., 2020

- **Goal:**
  Determine whether the outcome of optimizing a particular network N is stable to SGD noise

- **Procedure:**
  - Make **two copies of N** and **train** them with **different random samples** of SGD noise
  - **Compare these weights** with a **function** to produce a value called **instability of N**

- **Linear interpolation instability:**
  Maximum increase in error along linear interpolation path between w_T^1 and w_T^2



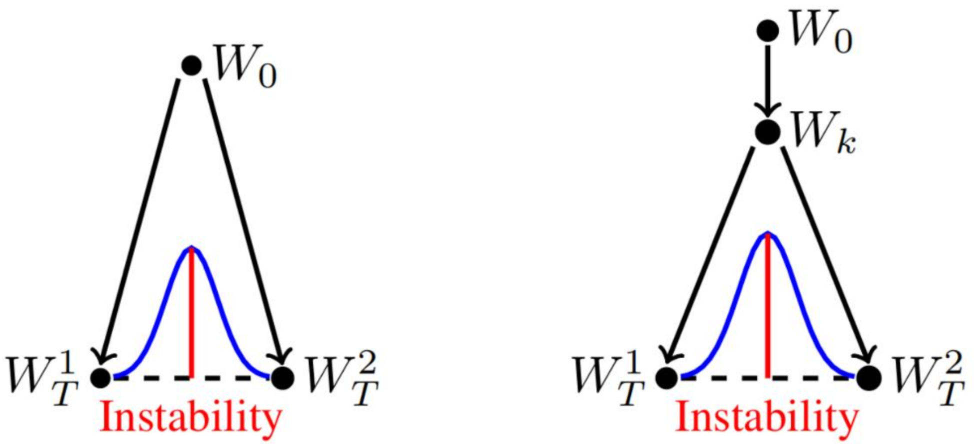
Frankle et al., 2020

Instability Analysis of SGD a la Frankle et al., 2020

- **Goal:**
  Determine whether the outcome of optimizing a particular network N is stable to SGD noise

- **Procedure:**
  - Make **two copies of N** and **train** them with **different random samples** of SGD noise
  - **Compare these weights** with a **function** to produce a value called **instability of N**

- **Linear interpolation instability:**
  Maximum increase in error along linear interpolation path between w_T^1 and w_T^2



Frankle et al., 2020

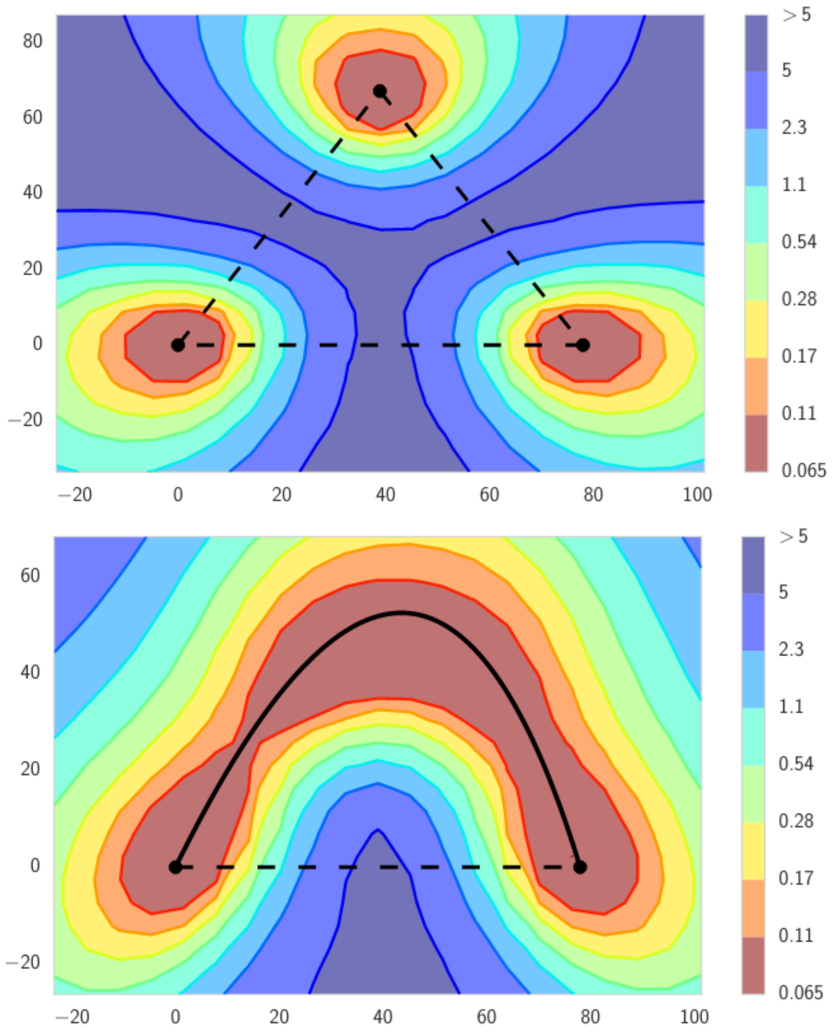
Outcome for MNIST, CIFAR10, ImageNet:
All but the smallest MNIST networks are unstable at initialization.
By a point early in training all networks become stable to SGD noise.

Loss Basin

Loose Definition:

“Area in the parameter space where the loss function has relatively low values.”

Neyshabur et al., 2020



Garipov et al., 2018

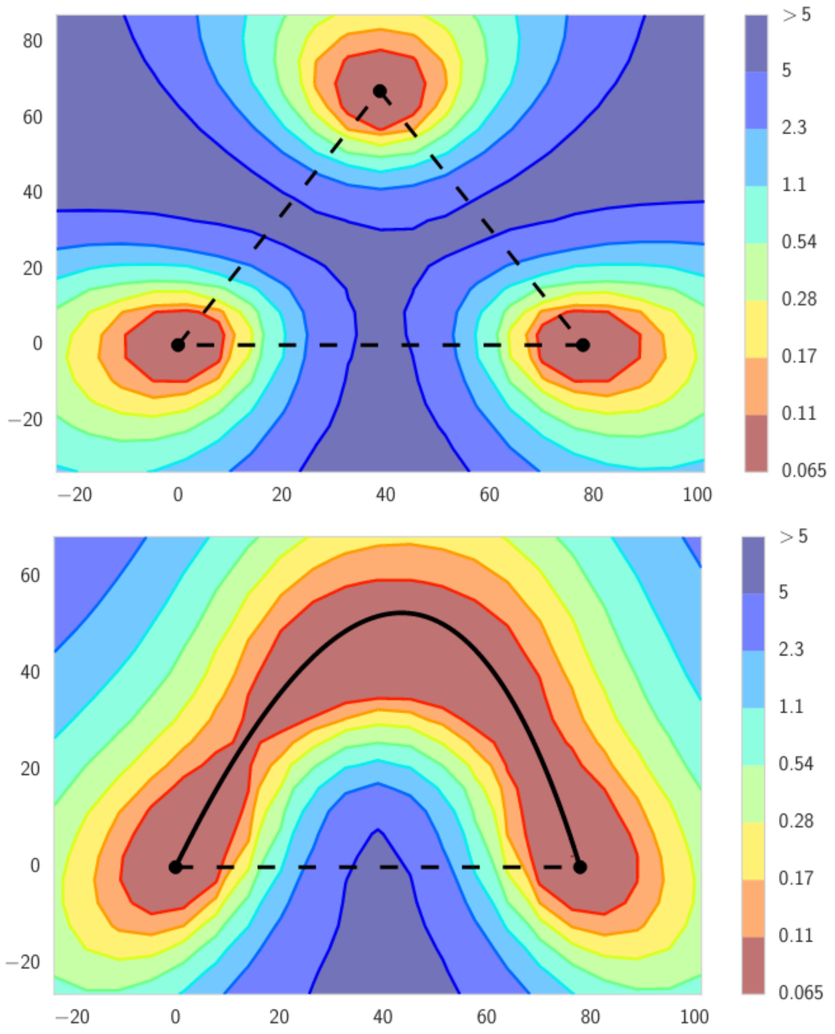
Loss Basin

Loose Definition:

"Area in the parameter space where the loss function has relatively low values."

Neyshabur et al., 2020

SGD solutions that are **linearly connected** with no barrier are in the **same basin** of the loss landscape. Entezari et al., 2022; Frankle et al., 2020



Garipov et al., 2018

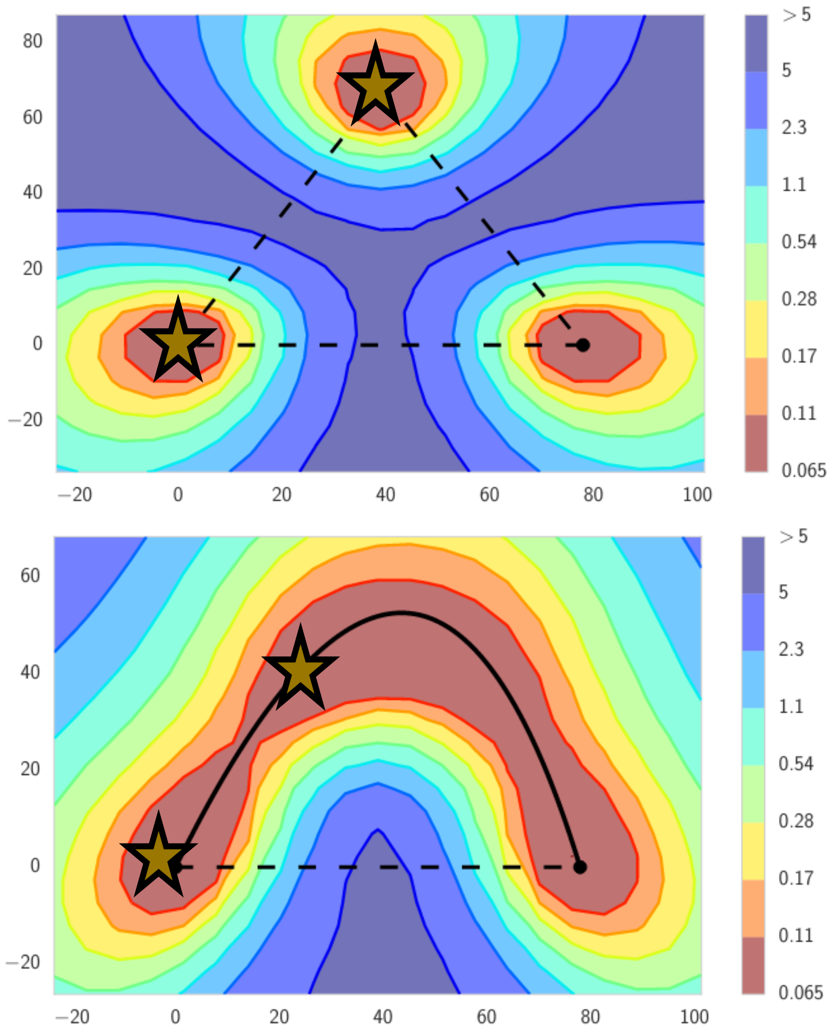
Loss Basin

Loose Definition:

“Area in the parameter space where the loss function has relatively low values.”

Neyshabur et al., 2020

SGD solutions that are **linearly connected** with no barrier are in the **same basin** of the loss landscape. Entezari et al., 2022; Frankle et al., 2020



Garipov et al., 2018

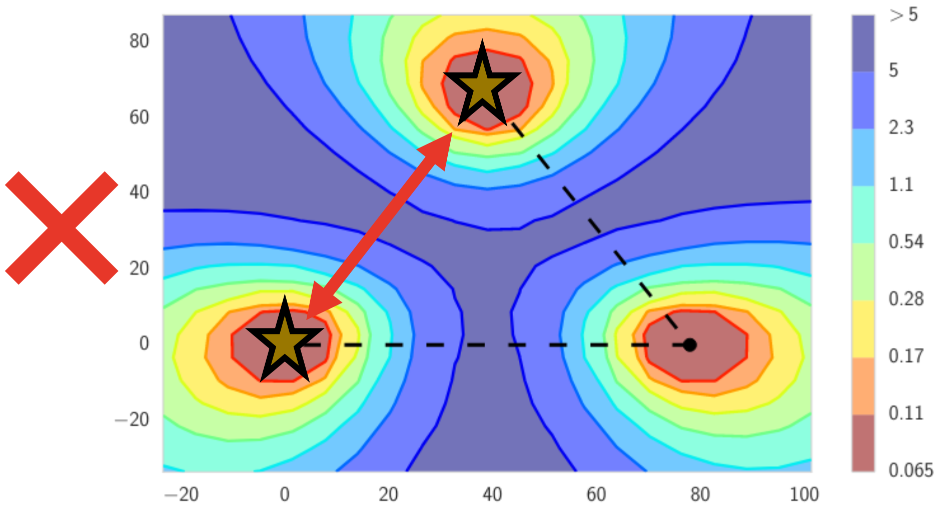
Loss Basin

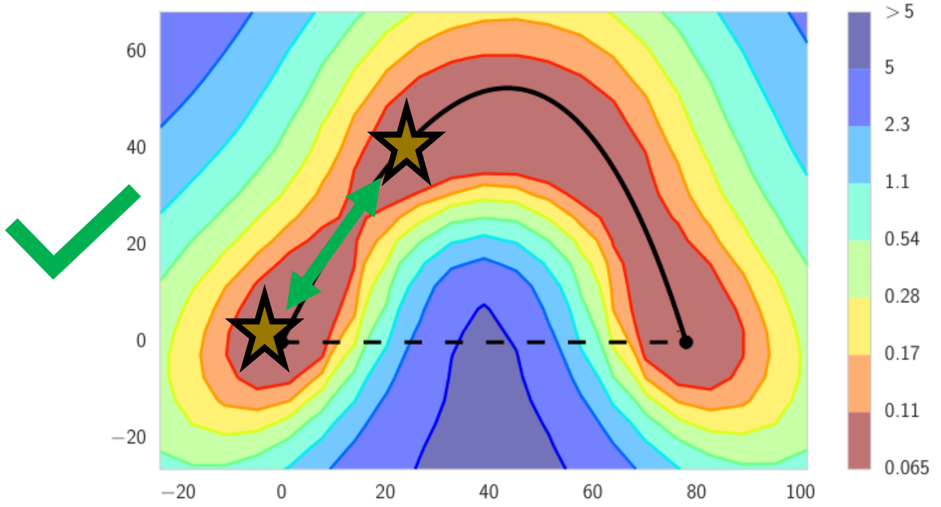
Loose Definition:

"Area in the parameter space where the loss function has relatively low values."

Neyshabur et al., 2020

SGD solutions that are **linearly connected** with no barrier are in the **same basin** of the loss landscape. Entezari et al., 2022; Frankle et al., 2020





Garipov et al., 2018

Training dynamics of SGD

Implications from Instability Analysis

Training can be divided in **two phases:**

- **Unstable phase:** Network finds linearly unconnected minima due to SGD noise
- **Stable phase:** Linearly connected minimum is determined

Frankle et al., 2020

Training dynamics of SGD

Implications from Instability Analysis

Training can be divided in **two phases:**

- **Unstable phase:** Network finds linearly unconnected minima due to SGD noise
- **Stable phase:** Linearly connected minimum is determined

Frankle et al., 2020

- **Similar findings & connections to other papers:**
  - Gradient descent happens in a subspace. [Gur-Ari et al., 2018]

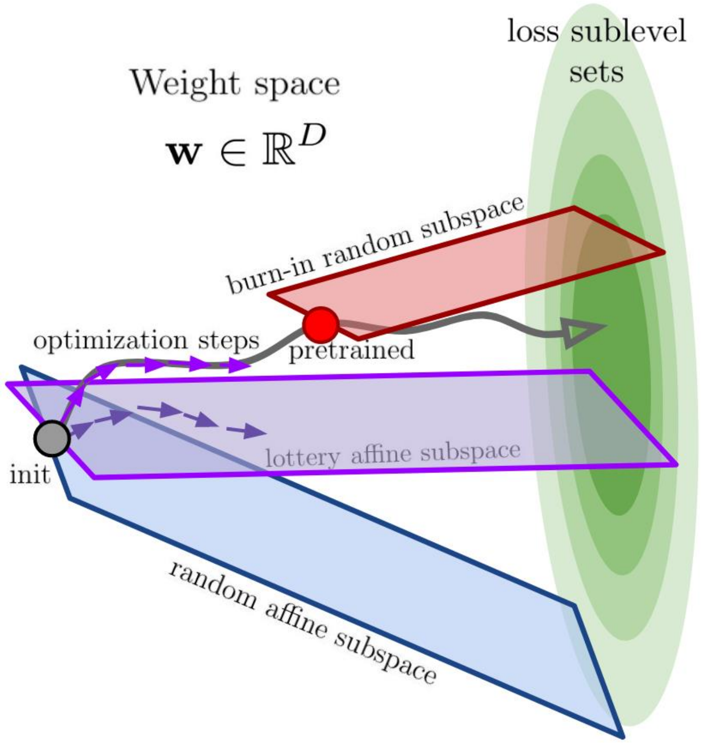
Training dynamics of SGD

Implications from Instability Analysis

Training can be divided in **two phases:**

- **Unstable phase:** Network finds linearly unconnected minima due to SGD noise
- **Stable phase:** Linearly connected minimum is determined

Frankle et al., 2020



Larsen et al., 2022

- **Similar findings & connections to other papers:**
  - Gradient descent happens in a subspace. [Gur-Ari et al., 2018]
  - Longer burn-in lowers the number of degrees of freedom required to train to a given accuracy. [Larsen et al., 2022]

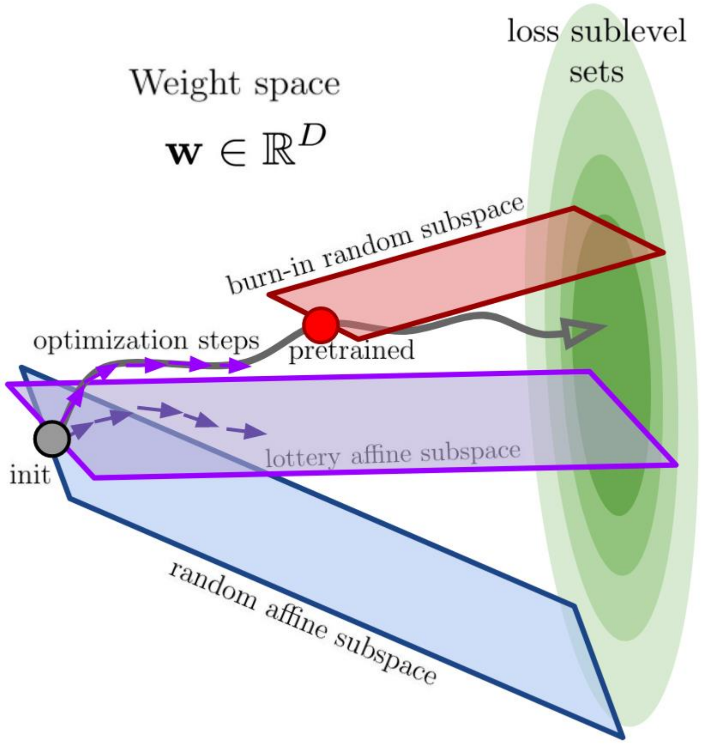
Training dynamics of SGD

Implications from Instability Analysis

Training can be divided in **two phases:**

- **Unstable phase:** Network finds linearly unconnected minima due to SGD noise
- **Stable phase:** Linearly connected minimum is determined

Frankle et al., 2020



Larsen et al., 2022

- **Similar findings & connections to other papers:**
  - Gradient descent happens in a subspace. [Gur-Ari et al., 2018]
  - Longer burn-in lowers the number of degrees of freedom required to train to a given accuracy. [Larsen et al., 2022]
  - There exists a “break-even point” on the training trajectory. Hyperparameters in the early phase control the mini-batch noise and the local curvature of loss surface after this “break-even point”. [Jastrzebski et al., 2020]

Connecting Loss Landscapes and Transfer Learning
by Neyshabur et al., 2020

When **training from pre-trained weights**, the **model stays in the same basin** in the loss landscape and different instances of such model are similar in feature space and close in parameter space.

Connecting Loss Landscapes and Transfer Learning
by Neyshabur et al., 2020

When **training from pre-trained weights**, the **model stays in the same basin** in the loss landscape and different instances of such model are similar in feature space and close in parameter space.

- **Other observations:**
  - Benefits of transfer learning come not only from feature reuse, but also from low-level data statistics.

Connecting Loss Landscapes and Transfer Learning

by Neyshabur et al., 2020

When **training from pre-trained weights**, the **model stays in the same basin** in the loss landscape and different instances of such model are similar in feature space and close in parameter space.

- **Other observations:**
  - Benefits of transfer learning come not only from feature reuse, but also from low-level data statistics.
  - Two instances of models trained from the same pre-trained weights make more common mistakes.

Connecting Loss Landscapes and Transfer Learning
by Neyshabur et al., 2020

> When **training from pre-trained weights**, the **model stays in the same basin** in the loss landscape and different instances of such model are similar in feature space and close in parameter space.

- **Other observations:**
  - Benefits of transfer learning come not only from feature reuse, but also from low-level data statistics.
  - Two instances of models trained from the same pre-trained weights make more common mistakes.
  - One can start fine-tuning from earlier pre-training checkpoints without loosing accuracy in the target domain.

Our Approach:

Use loss landscape information for improved fine-tuning

Summary so far

Summary so far

Problem setting:

Transfer Learning

Lightweight fine-tuning can perform better under distribution shift

Summary so far

Problem setting:

Transfer Learning

Lightweight fine-tuning can perform better under distribution shift

Conceptual model:

Loss basin view on SGD

Training consists of a stable and unstable phase; fine-tuning stays in same basin

Summary so far

Problem setting:

Transfer Learning

Lightweight fine-tuning
can perform better
under distribution shift

Conceptual model:

Loss basin view on SGD

Training consists of a
stable and unstable phase;
fine-tuning stays in same basin

Possible contribution(s):

New Methods

Improved fine-tuning

New Insights

Loss landscape
under distribution shift

Research Questions

New Methods

New Insights

JKU

Research Questions

New Methods

New Insights

- Can we find subnetworks based on local loss surface information for better fine-tuning?

Research Questions

New Methods

New Insights

- Can we find subnetworks based on local loss surface information for better fine-tuning?

- Can we adapt a model inside a basin, i.e. use basin information as a type of regularization?

Research Questions

New Methods

- Can we find subnetworks based on local loss surface information for better fine-tuning?

- Can we adapt a model inside a basin, i.e. use basin information as a type of regularization?

New Insights

- Do pre-trained weights fit to the target distribution?

JKU

Research Questions

New Methods

- Can we find subnetworks based on local loss surface information for better fine-tuning?
- Can we adapt a model inside a basin, i.e. use basin information as a type of regularization?

New Insights

- Do pre-trained weights fit to the target distribution?
- When does "staying in the basin" break?
  e.g. in Meta-Learning setting

Conclusion

- Fine-tuning only a small part of the model can perform better under distribution shift
- Fine-tuning stays within the same loss basin
- We want to use insights on the loss landscape for transfer learning
- Discuss and send papers! ☺

Researchers & Workshops@NeurIPS22

Researchers

- Mitchell Wortsman, PhD University of Washington
- Benahm Neyshabur, Research Scientist,Google Research
- Jonathan Frankle, Ass. Prof Harvard
- Stanislav Fort, Research Scientist, Anthropic
- Stanisław Jastrzębski, CTO Molecule.one
- Chelsea Finn, Ass. Prof Stanford
- Andrew Gordon Wilson, Prof NYU
- Michael I. Jordan, Prof UC Berkeley
- Hugo Larochelle, Prof Mila & Google Brain
- Samuel Ainsworth, Research Scientist Cruise AI
- Surya Ganguli, Prof Stanford
- Vincent Dumoulin, Research Scientist, Google Brain
- Pavel Izmailov, PhD NYU
- Timur Garipov, PhD MIT
- Guy Gur-Ari, Research Scientist Google Research
- Ali Farhadi, Prof University of Washington
- Mohammad Rastegari, Apple
- Martin Jaggi, Prof EPFL Lausanne
- Felix Draxler, PhD Heidelberg University
- Brett W. Larsen, PhD Stanford
- Gabriel Ilharco, PhD University of Washington
- Hanie Sedghi, Research Scientist Google Brain
- Roger Grosse, Prof University of Toronto
- James Lucas, Research Scientist NVIDIA
- Gintare Karolina Dziugaite, Research Scientist Google Brain
- Ludwig Schmidt, Ass. Prof University of Washington
- Pang Wei Koh, Ass. Prof University of Washington
- Percy Liang, Prof Stanford
- Shiori Sagawa, PhD Stanford
- Rahim Entezari, PhD TU Graz

Workshops@NeurIPS22

- Workshop on Distribution Shifts: Connecting Methods and Applications
- Workshop on Meta-Learning
- INTERPOLATE – First Workshop on Interpolation Regularizers and Beyond
- Transfer Learning for Natural Language Processing
- Federated Learning: Recent Advances and New Challenges
- OPT2022: Optimization for Machine Learning
- Order up! The Benefits of Higher-Order Optimization in Machine Learning
- Has it Trained Yet? A Workshop for Algorithmic Efficiency in Practical Neural Network Training

Thank You

PhD Seminar Talk

Maximilian Beck, beck@ml.jku.at, maxmbeck

Joint work with Sebastian Lehner and Sepp

Institute for Machine Learning, November 2022

Backup Slides

Loss Basin – Formal Definition

Definition 3.1. *Given a loss function $\ell : \mathbb{R}^n \to \mathbb{R}^+$ and a closed convex set $S \subset \mathbb{R}^n$, we say that S is a (ϵ, δ)-basin for ℓ if and only if S has all following properties:*

1. *Let U_S be the uniform distribution over set S and $\mu_{S,\ell}$ be the expected value of the loss ℓ on samples generated from U_S. Then,*

$$\mathbb{E}_{\mathbf{w} \sim U_S}[|\ell(\mathbf{w}) - \mu_{S,\ell}|] \leq \epsilon \tag{1}$$

2. *For any two points $w_1, w_2 \in S$, let $f(w_1, w_2) = w_1 + \tilde{\alpha}(w_2 - w_1)$, where $\tilde{\alpha} = \max\{\alpha | w_1 + \alpha(w_2 - w_1) \in S\}$. Then,*

$$\mathbb{E}_{\mathbf{w_1}, \mathbf{w_2} \sim U_S, \nu \sim \mathcal{N}(0, (\delta^2/n) I_n)}[\ell(f(\mathbf{w_1}, \mathbf{w_2}) + \nu) - \mu_{S,\ell}] \geq 2\epsilon \tag{2}$$

3. *Let $\kappa(\mathbf{w_1}, \mathbf{w_2}, \nu) = f(\mathbf{w_1}, \mathbf{w_2}) + \frac{\nu}{\|f(\mathbf{w_1}, \mathbf{w_2}) - \mathbf{w_1}\|_2}(f(\mathbf{w_1}, \mathbf{w_2}) - \mathbf{w_1})$. Then,*

$$\mathbb{E}_{\mathbf{w_1}, \mathbf{w_2} \sim U_S, \nu \sim \mathcal{N}(0, \delta^2)}[\ell(\kappa(\mathbf{w_1}, \mathbf{w_2}, |\nu|)) - \mu_{S,\ell}] \geq 2\epsilon \tag{3}$$

3 requirements for a convex set to be a basin

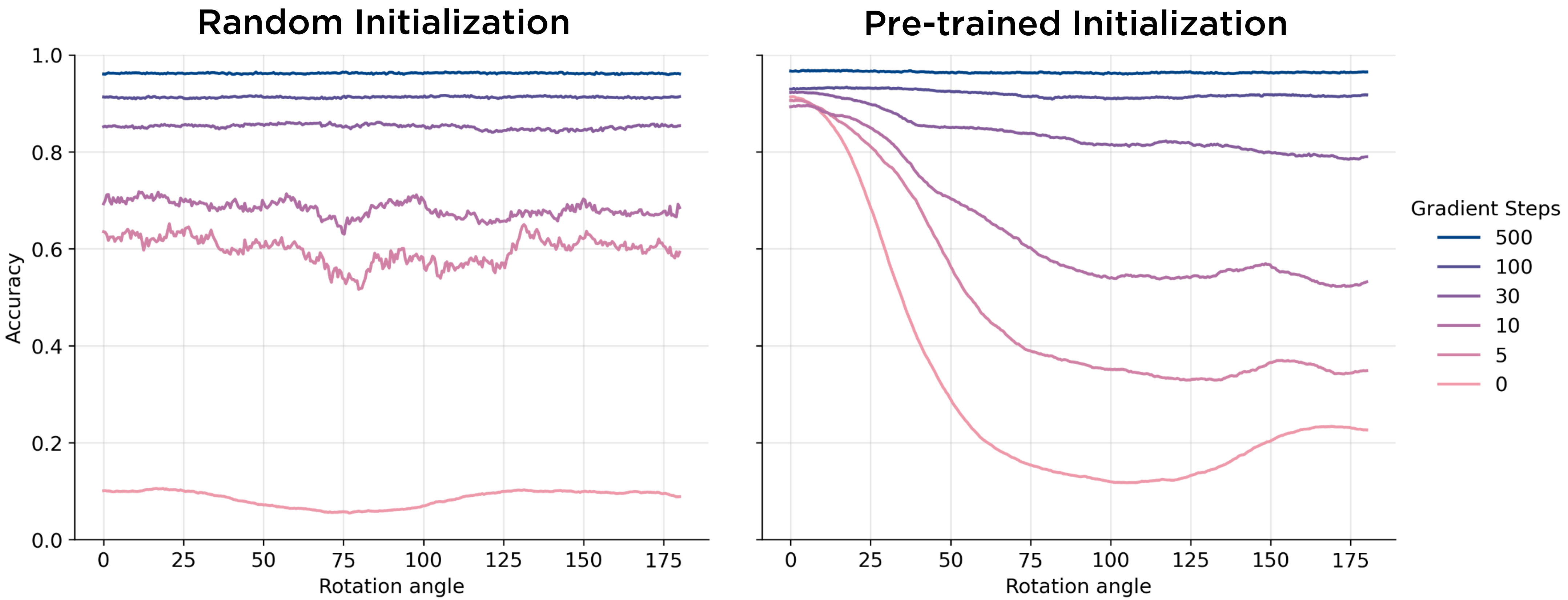
Neyshabur et al., 2020

Explanation:

1. Most points in the basin have a loss close to expected value of the loss in the basin.

2.-3. Loss of points in the vicinity of the basin is higher than the expected loss in the basin.

Toy Example: Rotated MNIST – Finetuning



100 Gradient Steps on 0° Rotation