Toward Understanding Self-Supervised Learning: The Roles of Losses and Optimizers

Tengyu Ma (Stanford)

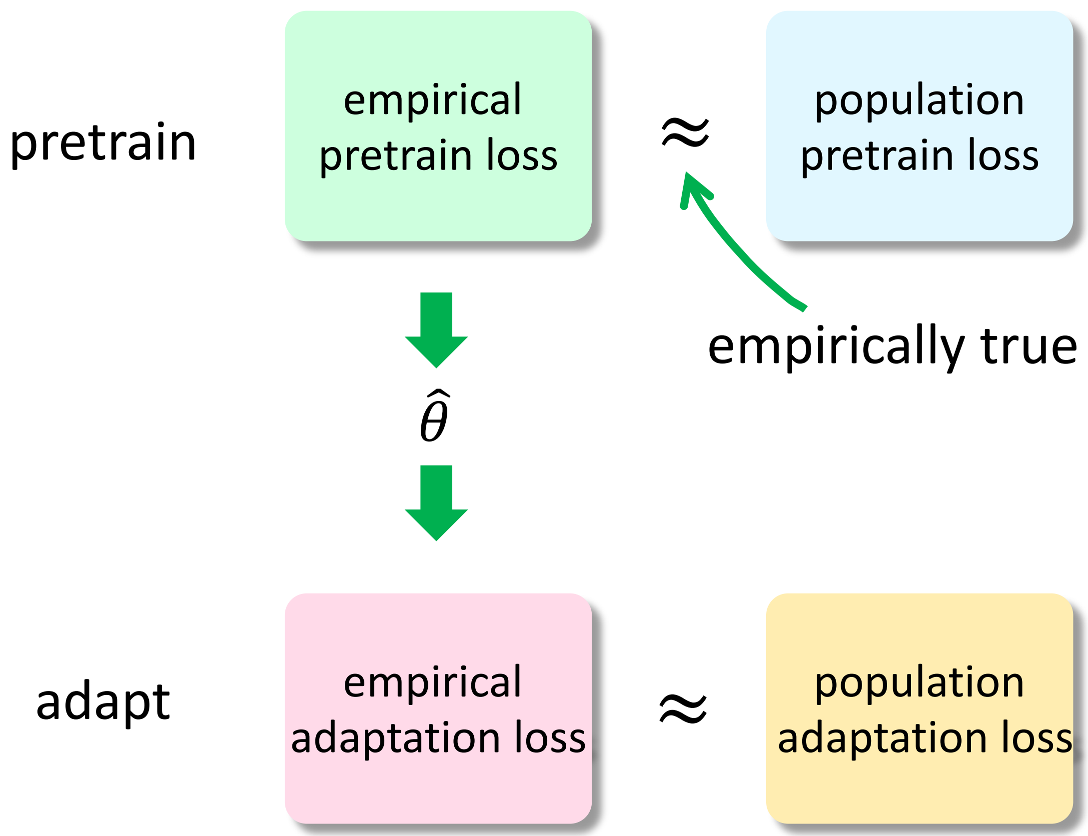
Why/When/How Does Pretraining Work?

1. The Role of Self-supervised Losses: What Structures of Data Do They Learn?

2. The Roles of Implicit Bias of Optimizers

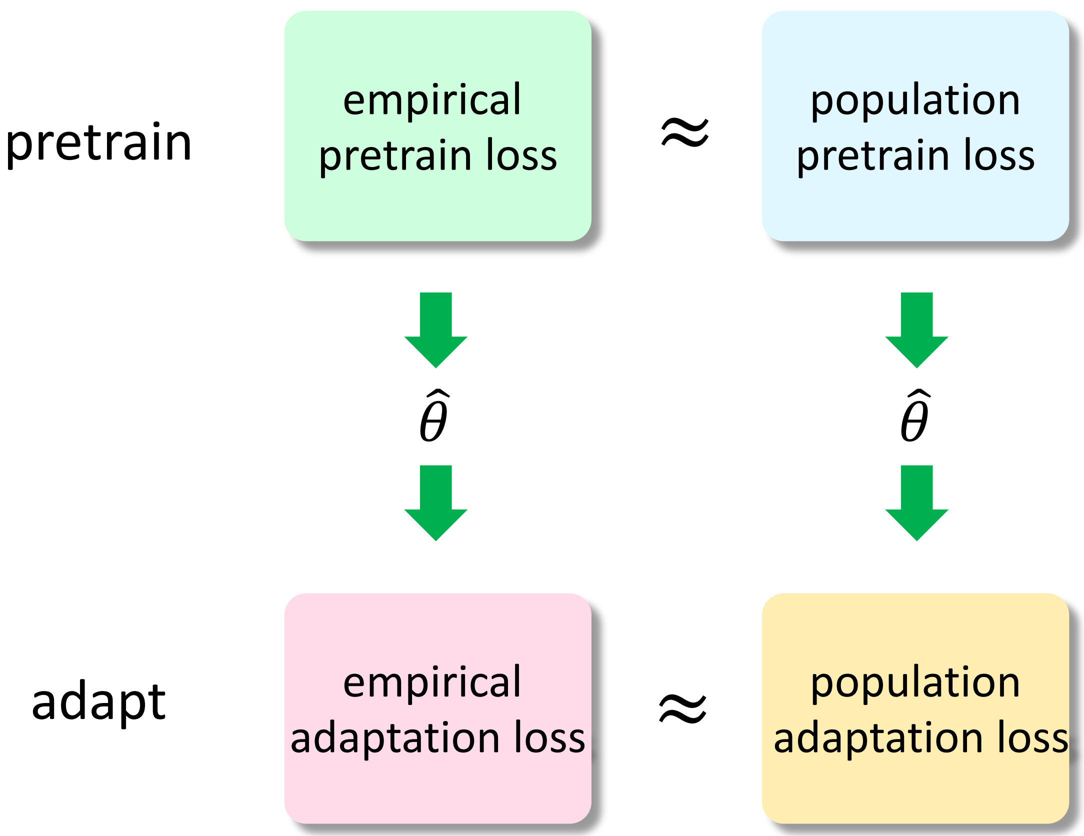
Isolating the Role of Losses, with Sufficient (Polynomial) Data

- Assume sufficient pretraining and downstream data (≫ the complexity of the model class)



Isolating the Role of Losses, with ~~Sufficient (Polynomial)~~ Infinite Data

- Might just as well assume infinite data

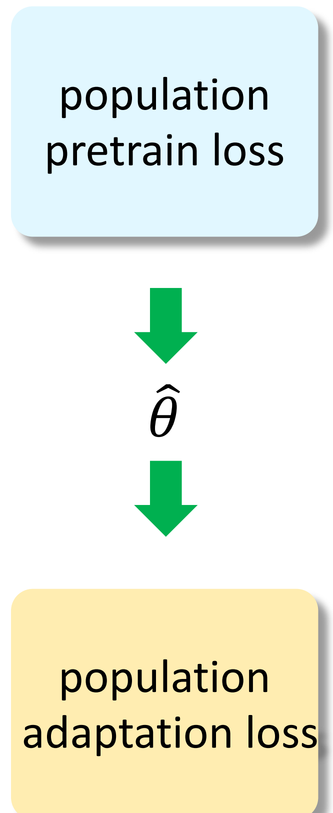


Isolating the Role of Losses, with ~~Sufficient (Polynomial)~~ Infinite Data

- Might just as well assume infinite data

Question:

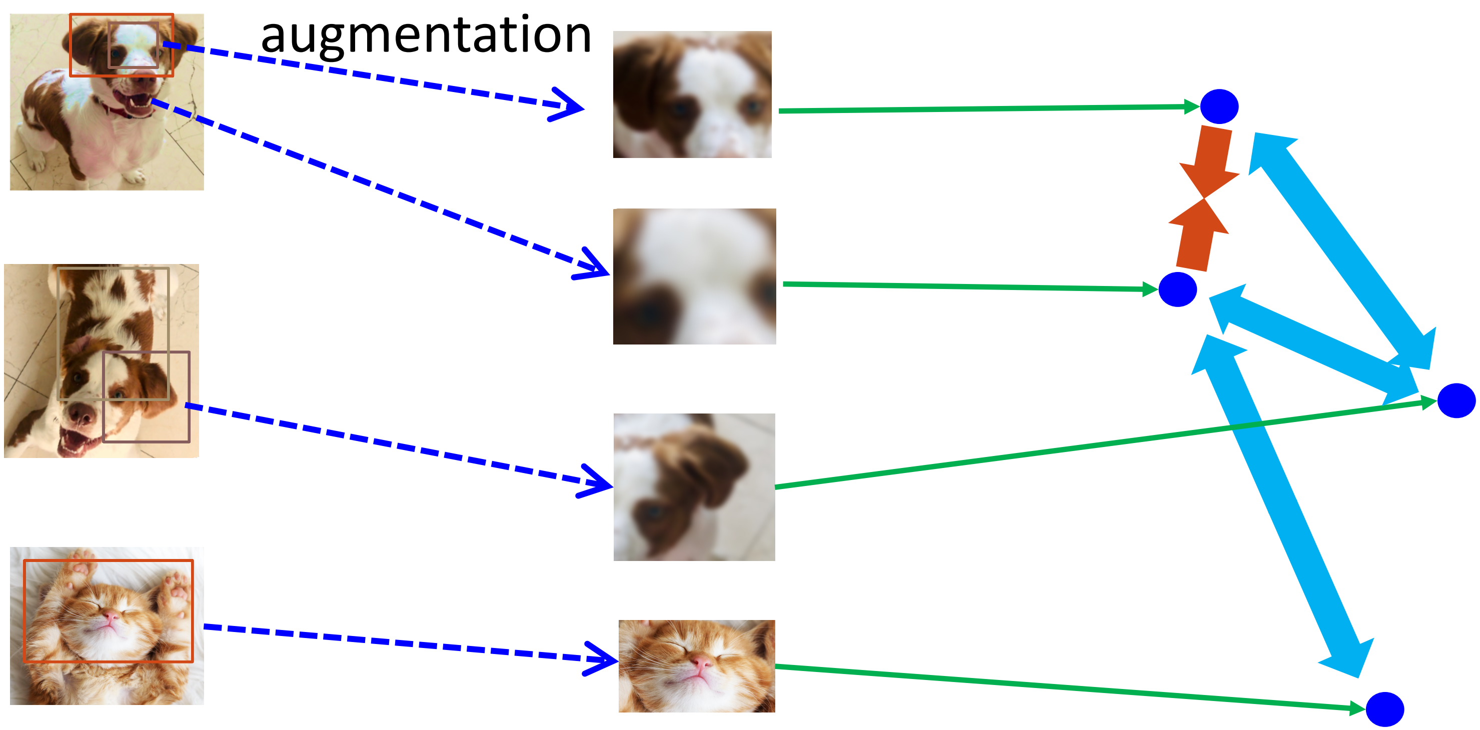
- Why does $\hat{\theta}$ give representations that are linearly separable on downstream tasks?



The Role of Contrastive Loss

Principles of contrastive loss:

- Pull representations of augmentations of the same image closer
- Push representations of augmentations of diff images further

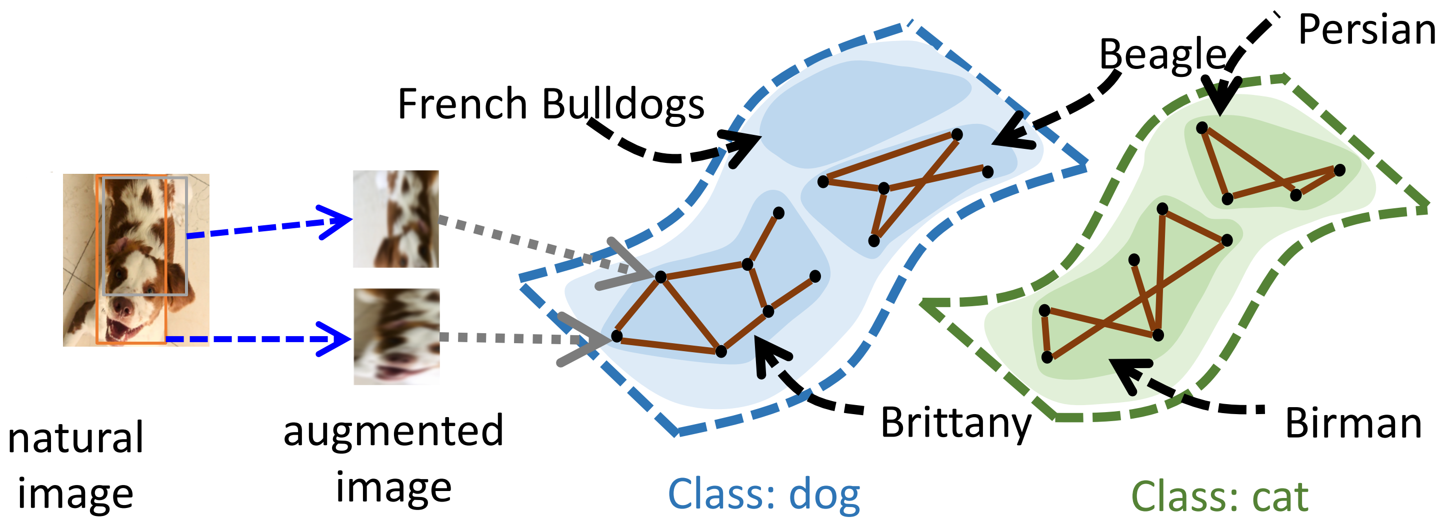


Various implementations: SimCLR [Chen et al'19], MoCo [He et al.'19], BYOL [Grill et al.'20], SimSiam [Chen et al.'20], SwAV [Caron et al.'20]

Contrastive Learning $=$ Spectral Clustering on an Infinite Graph

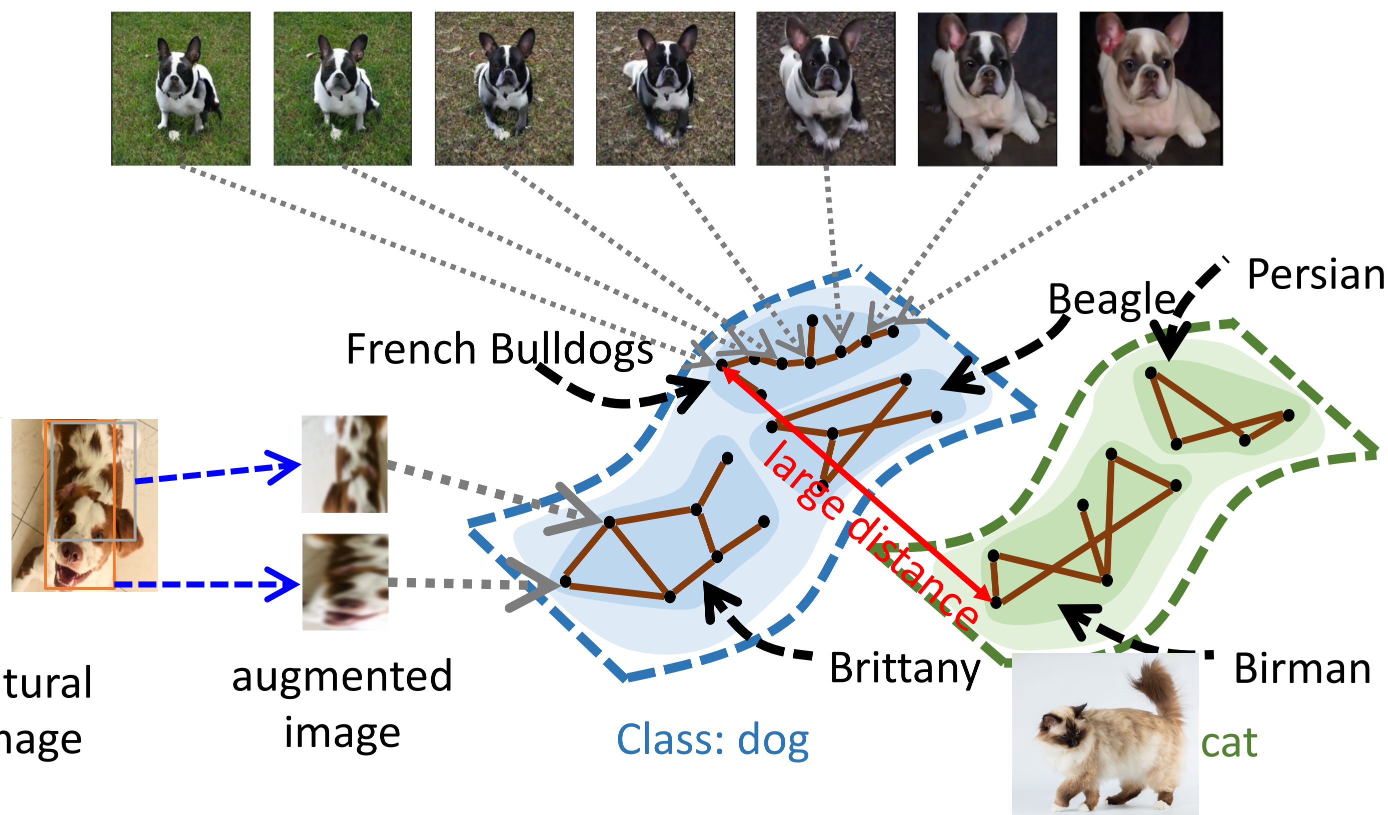
Population Positive-Pair Graph

- Vertex set: all images patches
- Edges: connect two patches if they can share an original image (i.e. they are positive pairs)
- Positive-pair graph is very sparse



Clustering Structures: Sub-clusters with Good Intra-connectivity

- Very few edges between different underlying classes
- Connectivity/expansions within the same classes or sub-classes
  - Two bulldogs can be connected via a sequence of bulldogs
- Graph distance is semantically meaningful



Main Results:
Contrastive Learning $\approx$ Spectral Clustering on Positive-Pair Graph

Theorem (informally stated):

With infinite data, minimizing the spectral contrastive loss is equivalent to spectral clustering on the positive-pair graph (up to rotations).

- We analyze the spectral contrastive loss (that also works empirically)

$$\min_{f} L(f) = -2\mathbb{E}_{x,x^+} f(x)^\top f(x^+) + \mathbb{E}_{x,x'} \left(f(x)^\top f(x')\right)^2$$

positive pair
(aug. of same image)

random pair
(aug. of random pairs of images)

What Downstream Tasks Can Be Solved Linearly?

Theorem (informally stated):

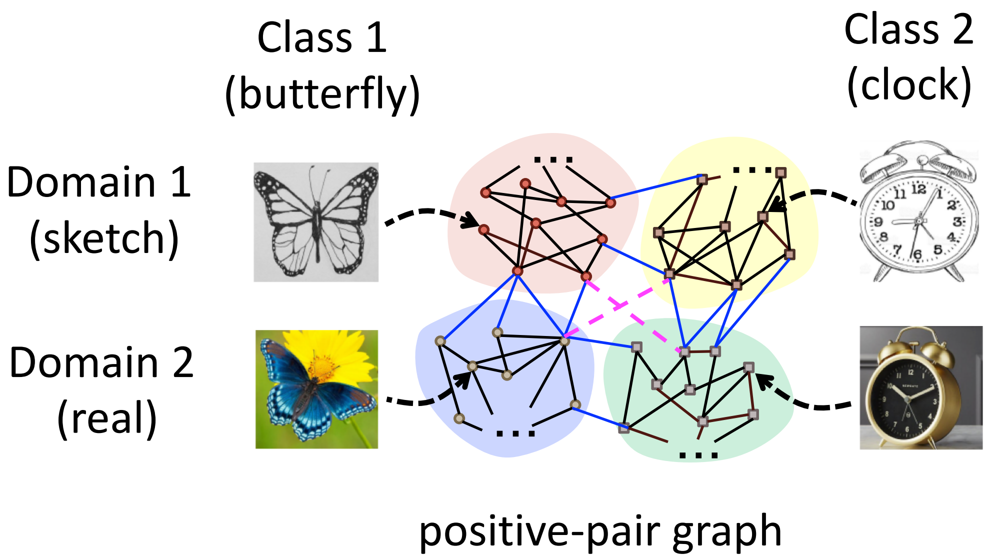
Suppose the positive-pair graph contains r major clusters, and representation dimension $k \geq 2r$.

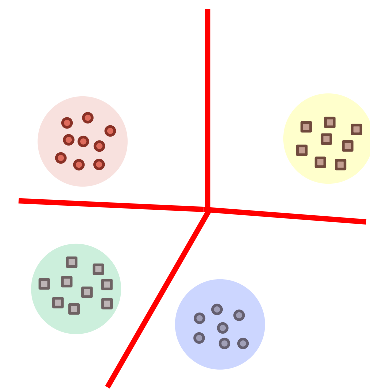
Then, linear classification on representations can solve any downstream task s.t. each cluster has the same label.

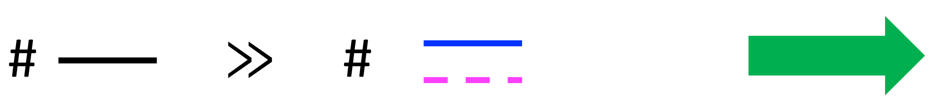
- relationship between task labels and structures of pretraining data

- A new but simple proof, using spectral graph theory tools
- Past works on spectral clustering don't analyze linear separability of the embeddings

Follow-up Work: Direction in Embedding Space Also Capture Relationship

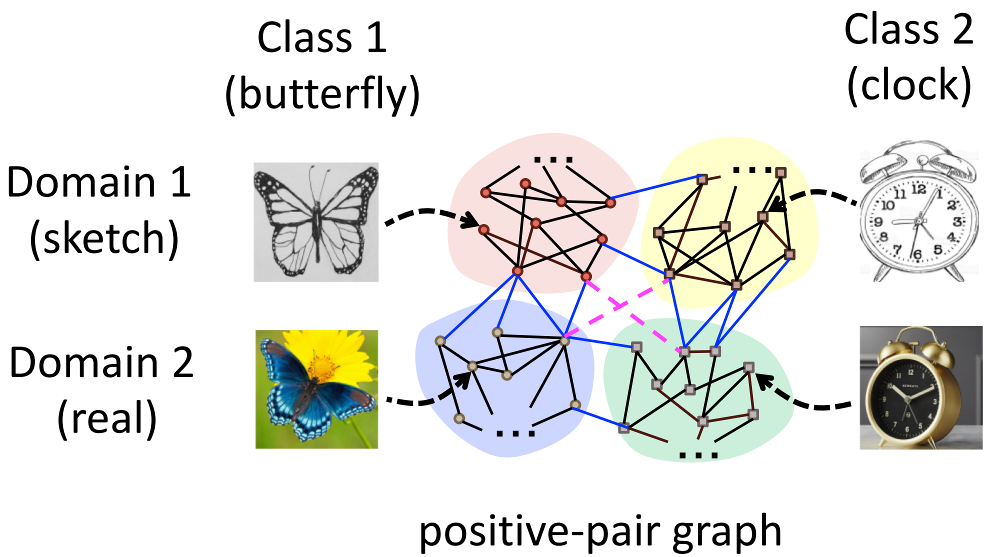


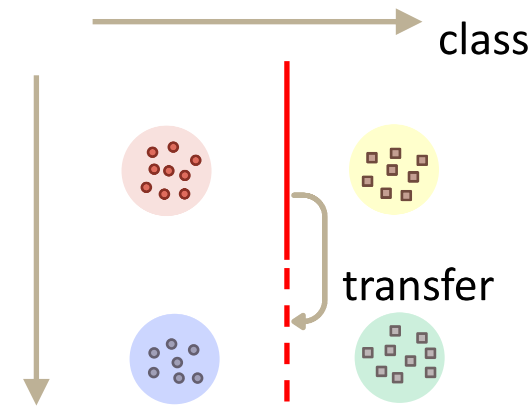




linear separability

Follow-up Work: Direction in Embedding Space Also Capture Relationship





—— ≫ # - - - ➡ linear transferability

➢ Pretraining features + finetuning on source gives SOTA performance for unsupervised domain adaptation

[HaoChen-Wei-Kumar-M.'2022, Shen-Jones-Kumar-Xie-HaoChen-M.-Liang.'22]

Why/When/How Does Pretraining Work?

1. The Role of Self-supervised Losses: What Structures of Data Do They Learn?

2. The Roles of Implicit Bias of Optimizers

- Previous slides and prior works: good pre-training loss => good downstream performance [Saunshi et al.'20, Wei et al.'21, Xie et al.'21, Haochen et al.'21]
- Common practice: use validation pre-training loss as an indicator for downstream performance

Is Pre-training Loss Always Correlated with Downstream Perf?

- Some counter-exapmles: ∃ models with different architectures, the same pre-training loss, and different downstream performances [Tay et al.'21, Zhang et al.'22, Saunshi et al.'22]
  - A deep and narrow transformer > a wide and shallow one [Tay et al.'21]
  - AlBert > Bert, on a synthetic reasoning task [Zhang et al.'22]

Q.: can two models with the same architecture and the same pre-training loss still have different downstream perf?

- Spoiler: yes!
  - There is an implicit bias from the optimizers/algorithms
- Understanding SSL require studying the roles of
  - self-supervised losses [Arora et al.'19, Lee et al.'20, Tosh et al.20,21, Haochen et al.'21, 22 …]
  - inductive bias of the architectures [Haochen et al.'22]
  - implicit bias of the optimizers: the rest of this talk [Liu et al.'22]

Is Pre-training Loss Always Correlated with Downstream Perf (Cont'd)?

- A priori, many models of the same architecture can have the same pre-training loss.
  - Why should they have the same downstream performance?
- Our findings: indeed, ∃ models with the same pre-training loss and architecture but different downstream perf.
  - especially when the pre-training loss is near optimal

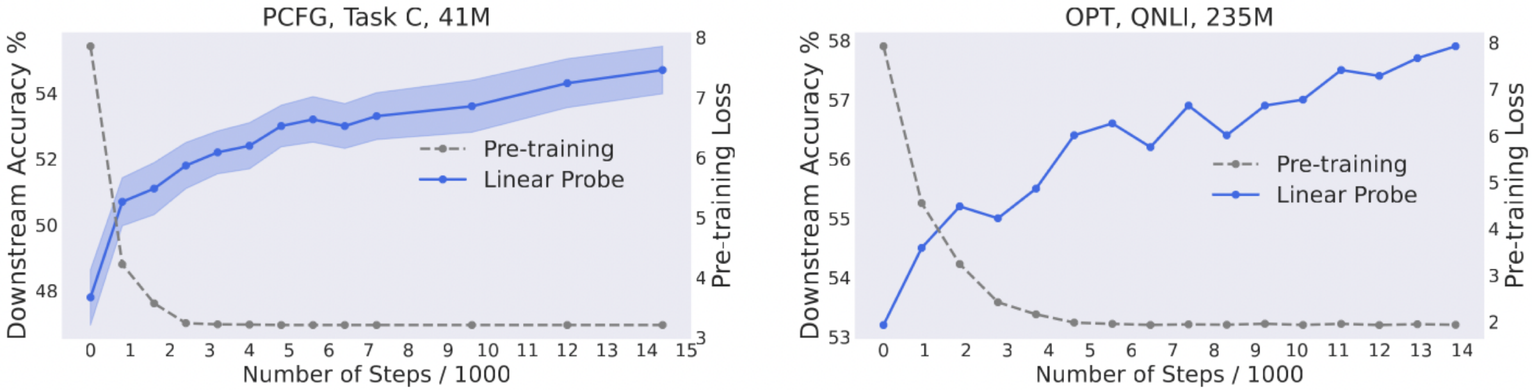
Experimental Setup

- Pre-training with MLM
- Simplified datasets from generative models
  - benefit: can compute the true MLM conditional prob.
- Downstream evaluation: fine-tuning and linear probe
- Saturation regime: ensure the pre-training loss is almost the same.
  - prediction = true conditional prob.
  - pre-training loss = entropy of true conditional prob.

Experimental Setup (Cont'd)

- Different factors in pre-training
  - # of steps after the pre-training loss converges
  - Training algorithms
    - "Natural" algorithms: AdamW, and SGD
    - Adversarial algorithms: add an objective to mess up downstream performance without changing the pre-training loss
    - Look-up table: a hypothetical model encoded in large transformers
      - memorizes all the inputs sequences
      - outputs the ground truth conditional prob. as features
  - Model sizes: transformers with sizes from 4M to 950M
- Note (again): in all experiments, the pre-training loss are the same

Varying # of Pre-training Steps:
Pre-training Loss Plateaus, But Downstream Perf Improves



(a) PCFG→Task C (b) OPT→QNLI

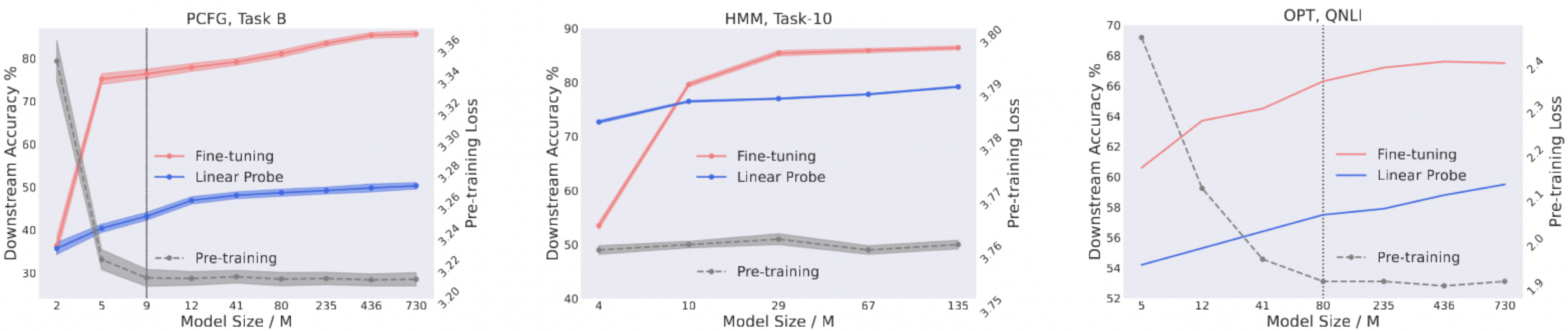
Changing the Algorithms:
Good pre-training loss ≠> Good downstream performance

Different pre-training algorithms on PCFG

Algorithm	Pre-training Loss	Task A Acc %	Task B Acc %
AdamW	3.204	89.9	49.2
Adversarial	3.206	83.1	42.3
Lookup table	3.196	71.2	39.7

- Adversarial algorithms indeed mess up the downstream perf. while keeping pre-training loss the same
- Lookup table has perfect pre-training loss but worst downstream perf.
  - Representations of transformers are better than the true conditional probability.

Varying the Model-size:
Large Models Is Better Than Smaller Ones



(a) PCFG→Task B　　(b) HMM→Task-10　　(c) OPT→QNLI

- Caveat: should transformers with different sizes be considered as the same arch.?

The Existence of Implicit Bias in Language Modeling

- ✓ There exists implicit bias in MLM.
  - Only training algorithms are different => they break ties among global minimizers differently

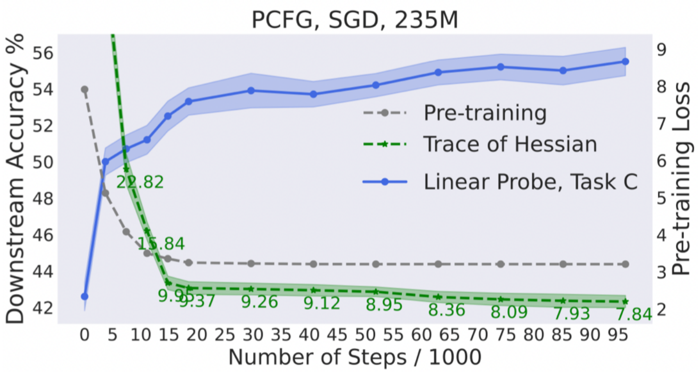
- **?** What's the role of implicit bias in MLM?
  - Theorem: In the saturation regime, SGD finds the flattest minimizer

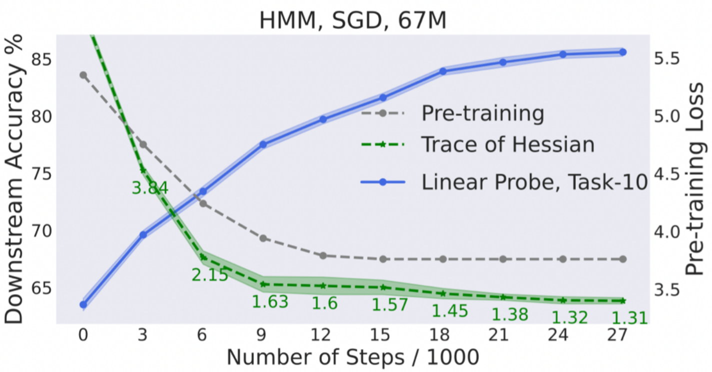
The Relationship between Downstream Perf and Flatness

- ✓ There exists implicit bias in language model pre-training.
- ✓ Implicit bias leads to flatter models.
- **?** Is downstream perf correlated with flatness?
  - We will evaluate the flatness of the models in the previous settings

Flatter Models Have Better Downstream Performance

Models at different pre-training time steps





Different pre-training algorithms on PCFG

Algorithm	Pre-training Loss	Task A Acc %	Task B Acc %	Task C Acc %	Trace of Hessian
AdamW	3.204	89.9	49.2	55.7	8.01
Adversarial	3.206	83.1	42.3	50.2	19.34

The Relationship between Downstream Perf and Flatness

- ✓ There exists implicit bias in language model pre-training.
- ✓ Implicit bias leads to flatter models.
- ✓ Downstream perf is correlated with flatness?
  - The paper also has some theory that explains this on toy language

Summary

Role of contrastive loss: spectral clustering on the positive-pair graph

Role of the optimizers:

- prefer flatter local minima
- flatness correlates with downstream perf (when the pretraining losses are the same)

Open questions

- The theory for implicit bias only works for SGD; how about AdamW?
- Theoretical results for “flatter models ⇒ better transferability”