Winter Term 21/22

Adversarial Self-Supervised Learning with Digital Twins

Lecture-4: Underspecification

Prof. Dr. Holger Giese (holger.giese@hpi.uni-potsdam.de)

Christian Medeiros Adriano (christian.adriano@hpi.de) - **"Chris"**

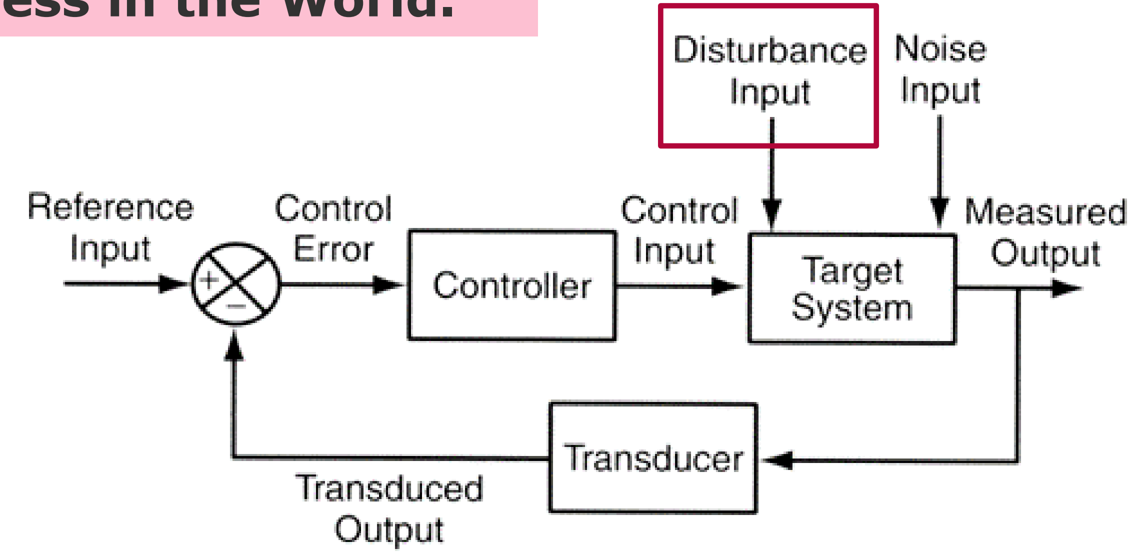
He Xu (He.Xu@hpi.de)

Infrastructure to run experiments

"Success in the Lab is not guarantee of success in the World."

Feedback loop models

"When solving a problem of interest, do not solve a more general problem as an intermediate step. Try to get the answer that you really need but not a more general one." **Vladimir Vapnik**

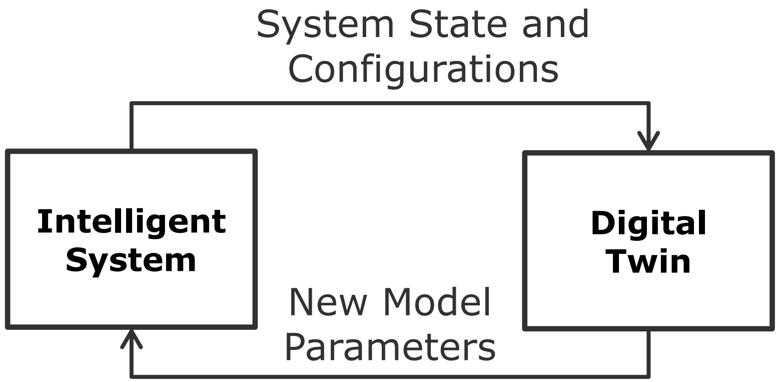


Simulation models

"Thinking is acting in an imagined space" **Konrad Lorenz**

"Perception is a generative act" – [Gross et al. 1999]

" Consciousness is a controlled hallucination " - [Seth et al. 2000]



Failure Modes [D’Amour 2020]

1- Structural Conflicts
Evidence: Solving the wrong problem
Root-causes: Measurement error, Selection bias, etc.

BUG

Solution: follow ML best practices, e.g., more and better data, methods more robust to overfitting, i.i.d. evaluations (train-test-validation splits)

2- Under-Specification
Evidence: Solving a fragment of the problem
Root-causes: lack of robustness, fairness, causal grounding

Untested code

Solution: stress test, sensitivity analysis, prior-knowledge, proper visualization

D'Amour, A., et al., 2020, Underspecification Presents Challenges for Credibility in Modern Machine Learning

Not a new problem but overlooked!

- Underspecification is well-documented in the ML literature, and is a core idea in deep ensembles, double descent, Bayesian deep learning, and loss landscape analysis
  - □ Lakshminarayanan, B., 2017, Simple and scalable predictive uncertainty estimation using deep ensembles.
  - □ Fort, S., 2019, Deep ensembles: A loss landscape perspective.
  - □ Belkin, M., 2018, Reconciling modern machine learning and the bias-variance trade-off.
  - □ Nakkiran, P., 2020, Deep double descent: Where bigger models and more data hurt.

However, its implications for the gap between i.i.d. and application-specific generalization are **neglected**.

Under-specification has downstream effects on robustness, fairness, and causal grounding.

D'Amour, A., et al., 2020, Underspecification Presents Challenges for Credibility in Modern Machine Learning

Under-specification [D’Amour 2020]

Definition: *An ML pipeline is underspecified when it generates predictors with equivalently held-out performance in the training domain, but that behave very differently in deployment.*

Impact: poor model behavior, instability at deployment, lack of system reliability

Keywords: distribution shift, spurious correlation, fairness, identifiability,

Domains: computer vision, natural language processing, medical imaging, electronic health records, genomics

Example: *The solution to an underdetermined system of linear equations (i.e.,) is under-specified, because*

- □ *it has more unknowns than linearly independent equations*
- □ *it represents an **equivalence class** of solutions given by a linear subspace of the variables.*

D'Amour, A., et al., 2020, Underspecification Presents Challenges for Credibility in Modern Machine Learning

Reichenbach's Common Cause Principle

Assume that $X \not\!\perp\!\!\!\perp Y$ (X and Y are dependent)

- X causes Y
- Y causes X
- There is a third hidden common cause
- Combination all the above

In other words, "there is not correlation without causation"

Hans
Reichenbach
1891-1953

Markov factorization

$$P(x_1, \dots x_n) = \prod_i P\underbrace{(x_i \,|Parents(x_i)}_{\text{Causal Markov kernels}})$$

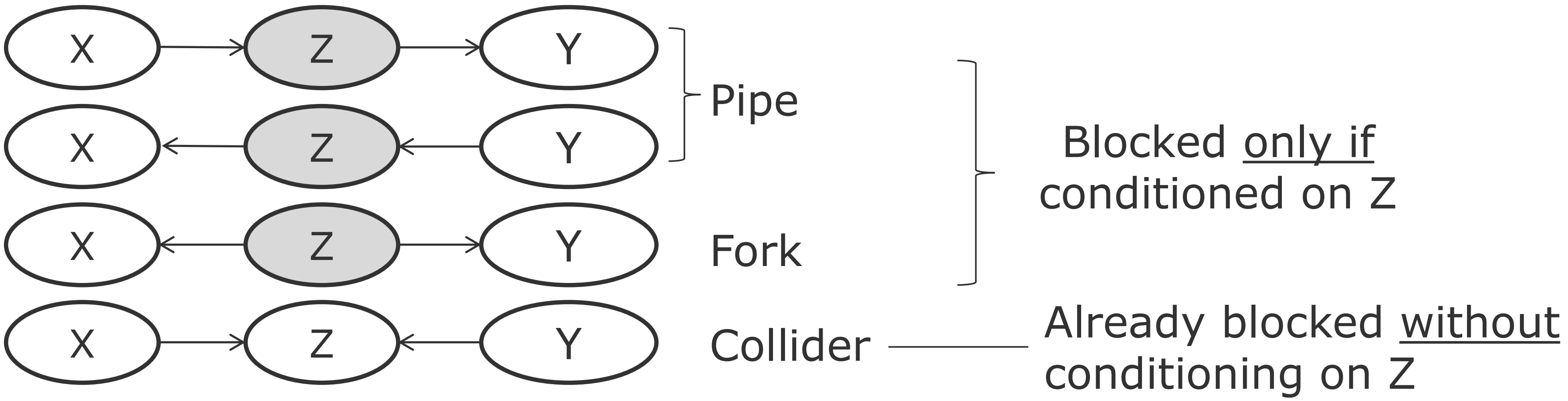
If P admits the factorization relative to a DAG G, we can say that the DAG represents the probability function P, i.e., G and P are compatible or that P is Markov relative to G.

We can also say that the Graph G **induces** the probability P.

d-separation

X and Y are d-separated if all paths between X and Y are blocked by a set Z of nodes, i.e., $X \perp Y \mid Z$

The blocking* situations are:



Pipe

Fork

Collider

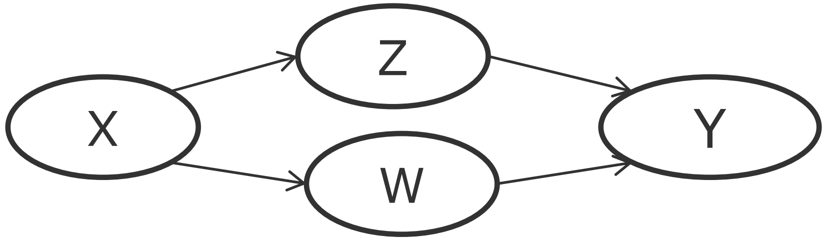
Blocked only if conditioned on Z

Already blocked without conditioning on Z

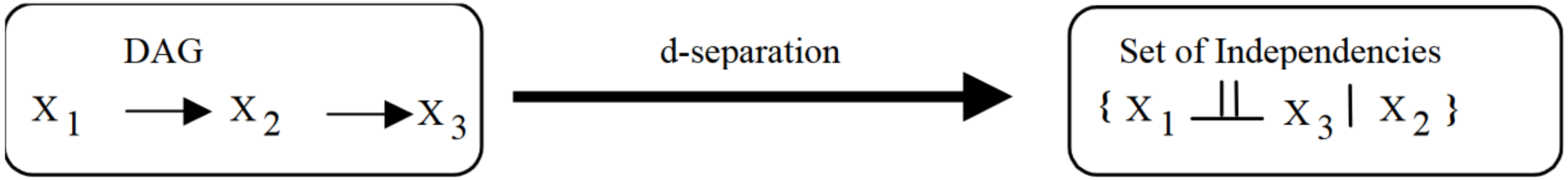
*grey means blocked (usually by conditioning on it)

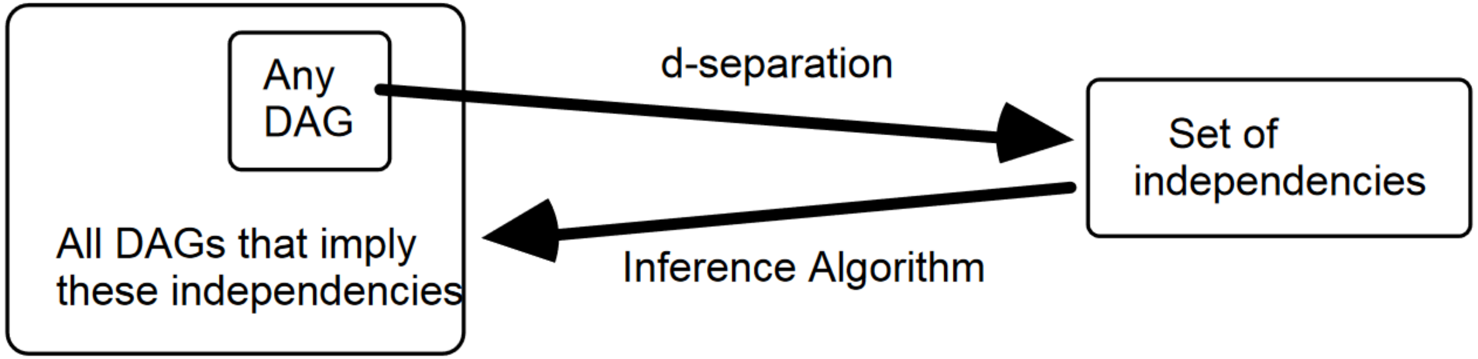
Faithfulness

If P(X_1,X_2,X_3) is faithful, then whatever the independencies that occur in P, they arise not from incredible coincidence, but rather from structure (a causal graph) [Scheines 1997].



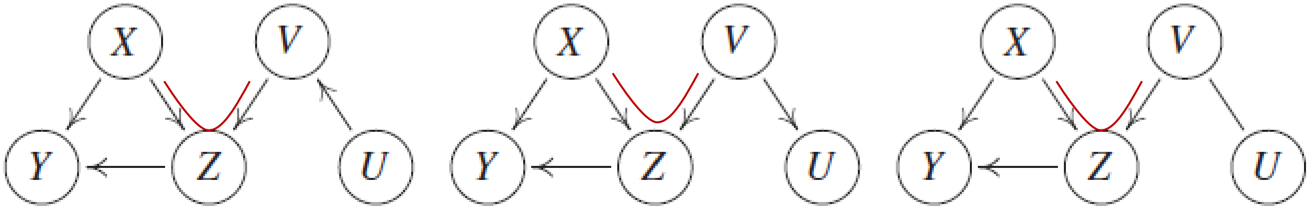
Exception, cancelling causal paths:
$Y \perp X$ but Y is not d-separated from X


DAG
$X_1 \rightarrow X_2 \rightarrow X_3$
d-separation
Set of Independencies
$\{ X_1 \perp\!\!\!\perp X_3 \mid X_2 \}$



Markov equivalence classes

- Graphs are equivalent if they have the same set of nodes and present the same colliders.



source: [Peters, Janzing & Schölkopf 2017]

Example of Under-Specification [D’Amour 2020]

- Epidemiological Model SIR $$\frac{dS}{dt} = -\beta \left(\frac{I}{N}\right) S, \qquad \frac{dI}{dt} = -\frac{I}{D} + \beta \left(\frac{I}{N}\right) S, \qquad \frac{dR}{dt} = \frac{I}{D}.$$

Which correspond to transmission rate (β), number of susceptible (S), infected (I), and recovered (R) individuals in a population of size N, and average duration of infection (D).

- Simulate with different initializations for D (D_0) to generate a full trajectory from the SIR model for a full time-course *T*, but learn the parameters (β;D) from data of observed infections up to some time $T_{obs} < T$

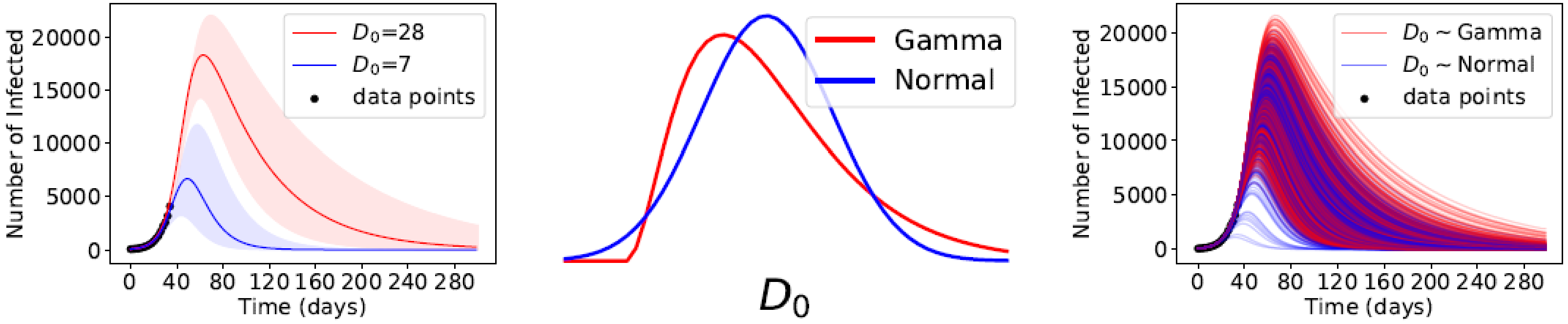


Figure 1: **Underspecification in a simple epidemiological model.** A training pipeline that only minimizes predictive risk on early stages of the epidemic leaves key parameters underspecified, making key behaviors of the model sensitive to arbitrary training choices. Because many parameter values are equivalently compatible with fitting data from early in the epidemic, the trajectory returned by a given training run depends on where it was initialized, and different initialization distributions result in different distributions of predicted trajectories.

Gendered Correlations in Downstream Tasks [D’Amour 2020]

- Gender-based shortcuts on two previously proposed benchmarks: a semantic textual similarity (STS) task and a pronoun resolution task **“The engineer alerted the scientist that her design specifications were not compliant.”**

- similarity delta = sim(“a woman is walking”; “a doctor is walking”) - sim(“a man is walking”; “a doctor is walking”)
- pronoun resolution task = a sentence with a pronoun that could refer to one of two possible antecedents (one is a profession), and the predictor must determine which of the antecedents is the correct one.
- goal = how the sensitivy are predictions to gender signals? How similarity delta for each profession correlates with the percentage of women actually employed in that profession (U.S. Bureau of Labor Statistics [BLS; Rudinger et al., 2018).

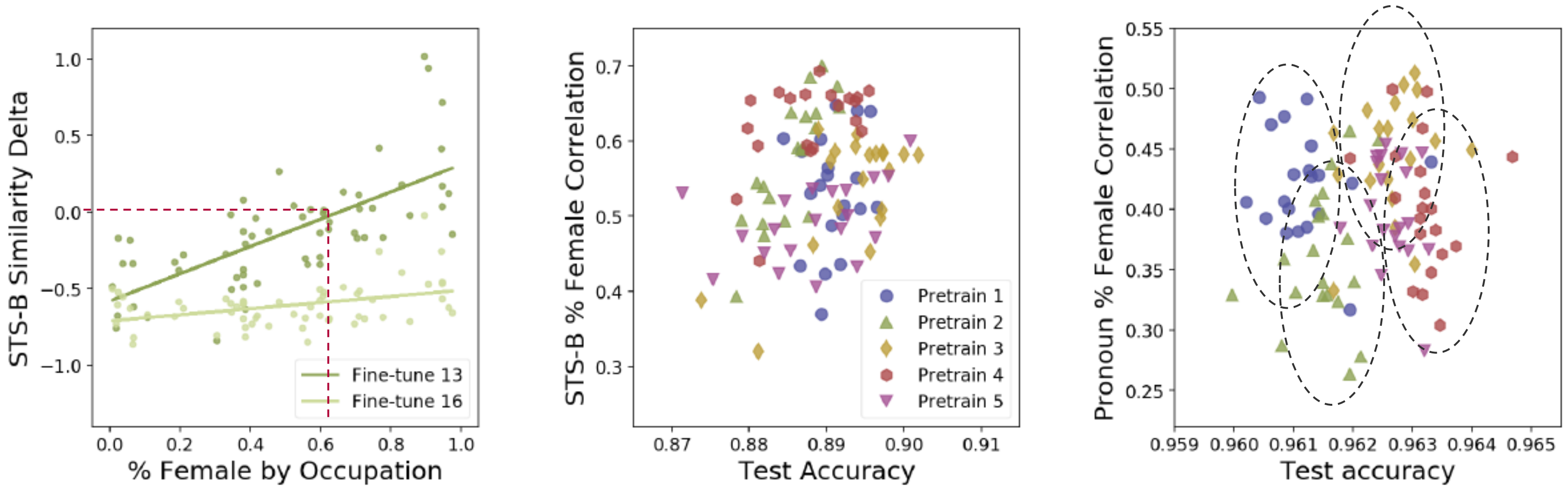


Figure 8: **Reliance on gendered correlations is affected by random initialization. (Left)** The gap in similarity for female and male template sentences is correlated with the gender statistics of the occupation, shown in two randomly-initialized fine-tunes. **(Right)** Pretraining initialization significantly affects the distribution of gender biases encoded at the fine-tuning stage.

Image classification [D'Amour 2020]

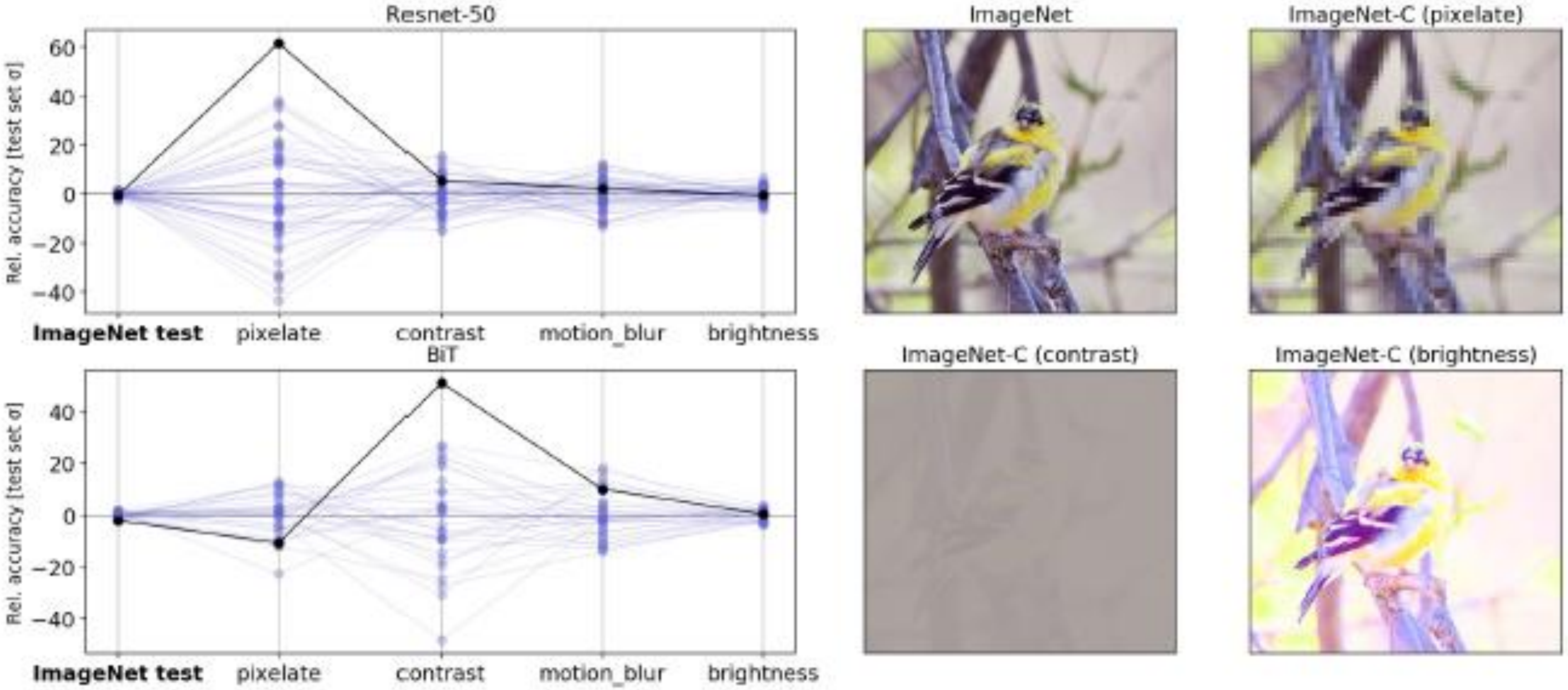


Figure 4: **Image classification model performance on stress tests is sensitive to random initialization in ways that are not apparent in iid evaluation. (Top Left)** Parallel axis plot showing variation in accuracy between identical, randomly initialized ResNet 50 models on several ImageNet-C tasks at corruption strength 5. Each line corresponds to a particular model in the ensemble; each each parallel axis shows deviation from the ensemble mean in accuracy, scaled by the standard deviation of accuracies on the "clean" ImageNet test set. On some tasks, variation in performance is orders of magnitude larger than on the standard test set. **(Right)** Example image from the standard ImageNet test set, with corrupted versions from the ImageNet-C benchmark.

Fundamental difficulty -1

- Misalignment between the predictor learned by empirical risk minimization and the causal structure of the desired predictor(Schölkopf, 2019; Arjovsky et al., 2019).
  - source of misalignment spurious correlations (called shortcuts in Deep Learning) that are result of selection bias and hidden confounders
- Usual approaches: use data collected in multiple environments to identify causal invariances:
  - Peters, J., et al., 2016, Causal inference by using invariant prediction: identification and confidence intervals. Journal of the Royal Statistical Society, 78(5):947–1012.
  - Arjovsky, M., et al., 2019, Invariant risk minimization.
  - Magliacane, S., et al., 2018, Domain adaptation by using causal inference to predict invariant conditional distributions, NeurIPS2018, pages 10869–10879.
- However, these approaches do not always work!

Fundamental difficulty-2

- Why might domain shift approaches not work?
  - **Reason-1**: they do not cover all cases where predictors trained to minimize predictive risk encode poor inductive biases. In many settings where ML excels, the structural issues identified above are not present.
  - **Reason-2**: usually, there is not enough information in the training distribution to distinguish between the inductive biases and spurious relationships, which are necessary to make the connection to causal reasoning,
  - Hence, this underspecified failure mode corresponds to a lack of positivity, not a structural defect in the learning problem.
    - The shortcut learning in deep neural networks stems from this difficulty. (Geirhos et al 2020)
    - Geirhos, R., et al., 2020, Shortcut learning in deep neural networks. arXiv preprint arXiv:2004.07780

Approach – Stress Test (Sensitivity Analysis)

- Probe the model's inductive biases on practically relevant dimensions—is sensitive to arbitrary, i.i.d.-performance-preserving choices, e.g., random seed.
- A key point is that the stress tests induce variation between predictors' behavior, not simply a uniform degradation of performance. This variation distinguishes under-specification-induced failure from the more familiar case of structural-change induced failure.

Procedure

1. Generate a suite of stress tests
   - Stratified
   - Shifted
   - Contrasted
2. Construct an ensemble of models
3. Confirm that i.i.d. performance if equivalent
4. Measure variation in the stress test performance

Solution – Stress Tests

Stratified Performance Evaluations

- We choose a particular feature A and stratify a standard test dataset D' into strata $D'_a = \{(x_i; y_i) : A_i = a\}$. A performance metric can then be calculated and compared across different values of a.

Shifted Performance Evaluations

- tests whether the average performance of a predictor f generalizes when the test distribution differs in a specific way from the training distribution. Specifically, these tests define a new data distribution $P' <> P$ from which to draw the test dataset D', then evaluate a performance metric with respect to this shifted dataset There are several strategies for generating P' to test different properties of the predictor f.
- For example, to test whether f exhibits invariance to a particular transformation $T(x)$ of the input, one can define P' to be the distribution of the variables $(T(x); y)$ when $(x; y)$ are drawn from the training distribution P_D

Contrastive Evaluation [D'Amour et al. 2020]

Contrastive evaluation relies on a dataset of matched sets C

$$\mathcal{C} = \{z_i\}_{i=1}^{|\mathcal{C}|}$$

where each matched set *z_i* consists of a base input x_i that is modified by a set of transformations *T*

$$\mathcal{T},\ z_i = (T_j(x_i))_{T_j \in \mathcal{T}}.$$

The evaluation metrics are computed with respect to matched sets and can include measures of similarity or ordering among the examples in the matched set.

For example, if one assumes that each transformation in *T* should be label-preserving, then a measurement of disagreement within the matched sets can reveal a poor inductive bias.

Contrastive Evaluation [D'Amour et al. 2020]

Common uses:

- ML Counterfactual notions of **Fairness** (Garg et al., 2019; Kusner et al., 2017).
  - Garg, S., et al., 2019, Counterfactual fairness in text classification through robustness. In Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society, pages 219–226.
  - Kusner, M. J., 2017, Counterfactual fairness, in Advances in neural information processing systems, pages 4066–4076.

- NLP **Robustness** and Testing (Ribeiro et al., 2020; Kaushik et al., 2020).
  - Ribeiro, T., et al., 2020, Beyond Accuracy: Behavioral Testing of NLP Models with CheckList
  - Kaushik, D., et al., 2020, Learning the difference that makes a difference with counterfactually-augmented data. In International Conference on Learning Representations.

Project Goals - Research Problems

1- Under-specification Problem **Simulation**

Goal: Show that different prediction models solve the task well for testing data, however, perform very differently in two distinct situations:

1.1 distinct hyper-parameters (prior knowledge)

1.2 out-of-distribution data (distribution shifts)

2- Value-at-Risk Problem **Sim2Real**

Goal: Show different rates of synchronization between Production and Simulation can lead to:

2.1 excessive cost of training and redeployment

2.2 increase in the risk of under-performance

3- Learning to Synchronize Problem **Feedback Loop**

Goal: Show that different strategies to learn when to train and redeploy require:

3.1 more data to achieve an average value-at-risk

3.2 longer time to converge

Next tasks

Think about answers to the for the following questions:

1. How to generate test-data?
2. How to generate distribution shifts? How much shift (how to measure it)?
3. How to train distinct models that perform equally well on test data?
4. How to do you know that two models are distinct?
5. How to do you know that they perform equally well on the test data?

Suggested Deadline = Nov 16 (present during lecture)

END