On calibration and out-of-domain generalization

(paper at NeurIPS 2021)

Uri Shalit
Technion – Israel Institute of Technology

Bellairs Workshop

March 2022

Team

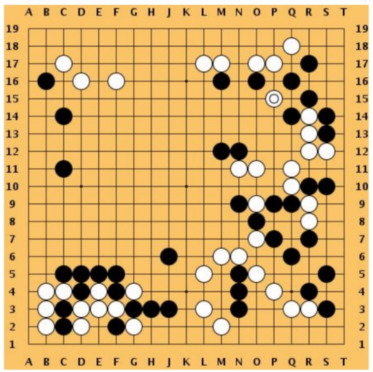
Yoav Wald
(Hebrew University →
Johns Hopkins University)

Amir Feder
(Technion)

Daniel Greenfeld
(Jether Energy Research)

Machine learning:
some remarkable successes

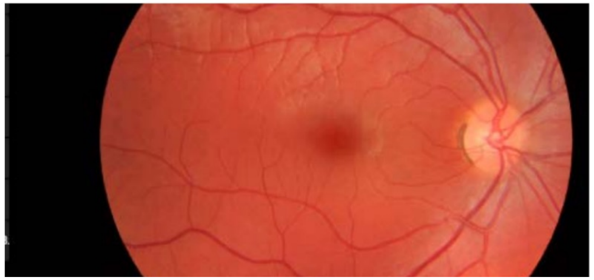
- Learning to classify
- Learning to act
  (when a perfect simulator is available)

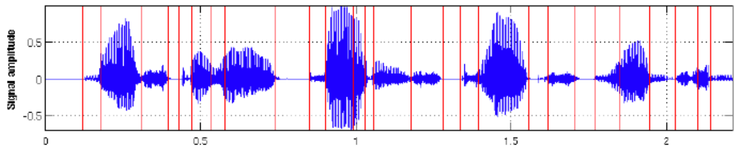


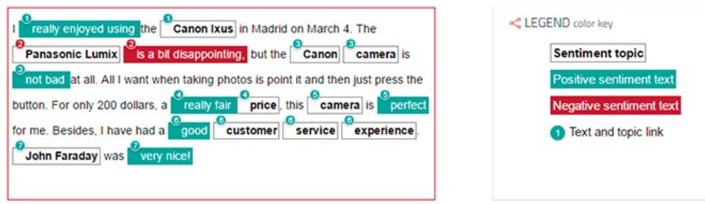
JAMA | Original Investigation | INNOVATIONS IN HEALTH CARE DELIVERY

Development and Validation of a Deep Learning Algorithm for Detection of Diabetic Retinopathy in Retinal Fundus Photographs

Varun Gulshan, PhD; Lily Peng, MD, PhD; Marc Coram, PhD; Martin C. Stumpe, PhD; Derek Wu, BS; Arunachalam Narayanaswamy, PhD; Subhashini Venugopalan, MS; Kasumi Widner, MS; Tom Madams, MEng; Jorge Cuadros, OD, PhD; Ramasamy Kim, OD, DNB; Rajiv Raman, MS, DNB; Philip C. Nelson, BS; Jessica L. Mega, MD, MPH; Dale R. Webster, PhD







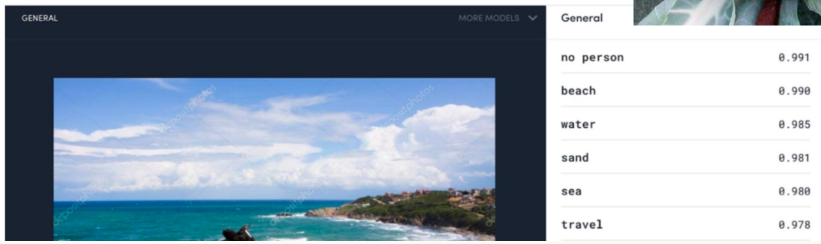
The next step:
Some things we can't do yet

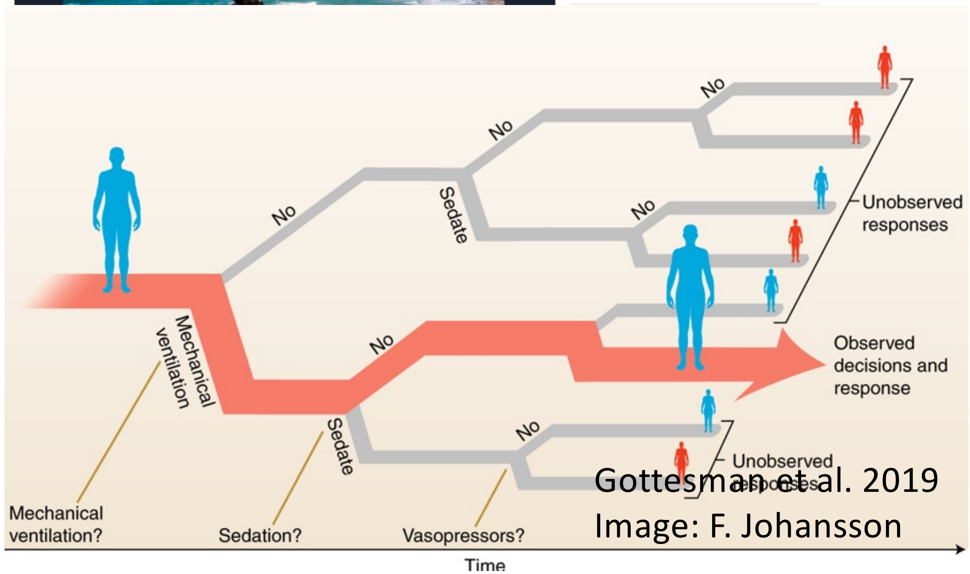
- Learn how to act optimally without access to a simulator
    - Based on observational data

- Unsupervised domain adapatation
    - Eg: classify images in a-priori unknown contexts

- My research is often motivated by problems in healthcare, where both subjects come up

Difficult cows





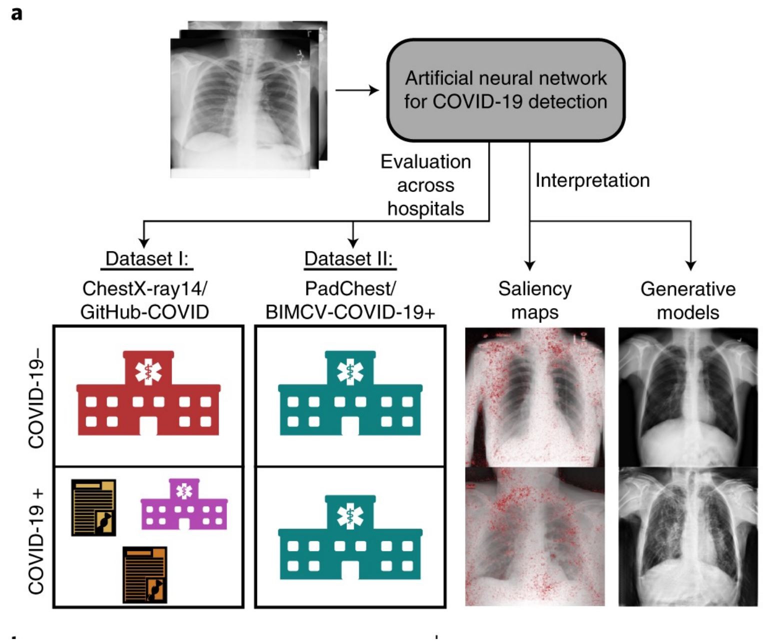
Gottesman et al. 2019
Image: F. Johansson

Article | Published: 31 May 2021

AI for radiographic COVID-19 detection selects shortcuts over signal

Alex J. DeGrave, Joseph D. Janizek & Su-In Lee

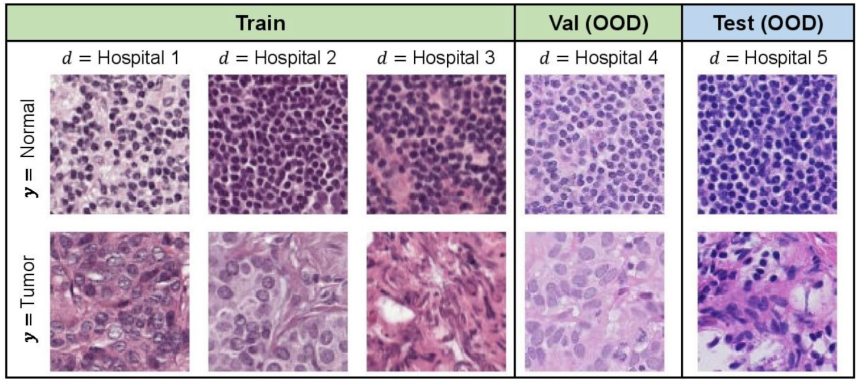
Nature Machine Intelligence **3**, 610–619 (2021) | Cite this article

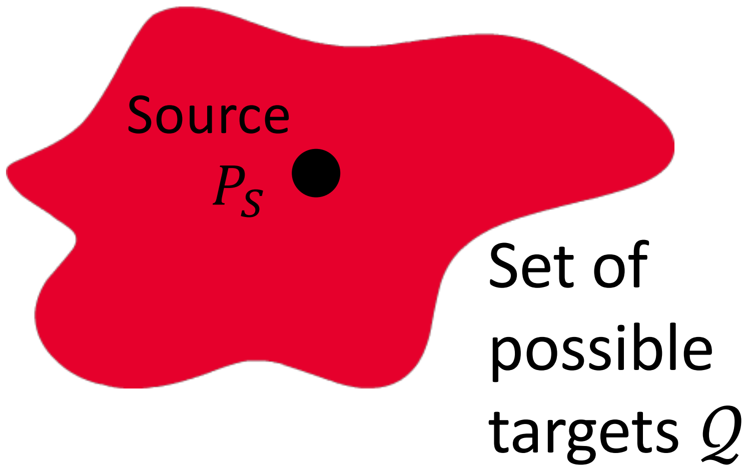


- Testing state-of-the-art deep learning models for COVID-19 detection
- Sharp drop in performance across hospitals and datasets
- Turns out the models often rely on spurious features outside the lungs
  - E.g.: Laterality markers, presence of shoulder region, known to be clinically irrelevant for COVID-19

Out-of-Domain (OOD) Generalization

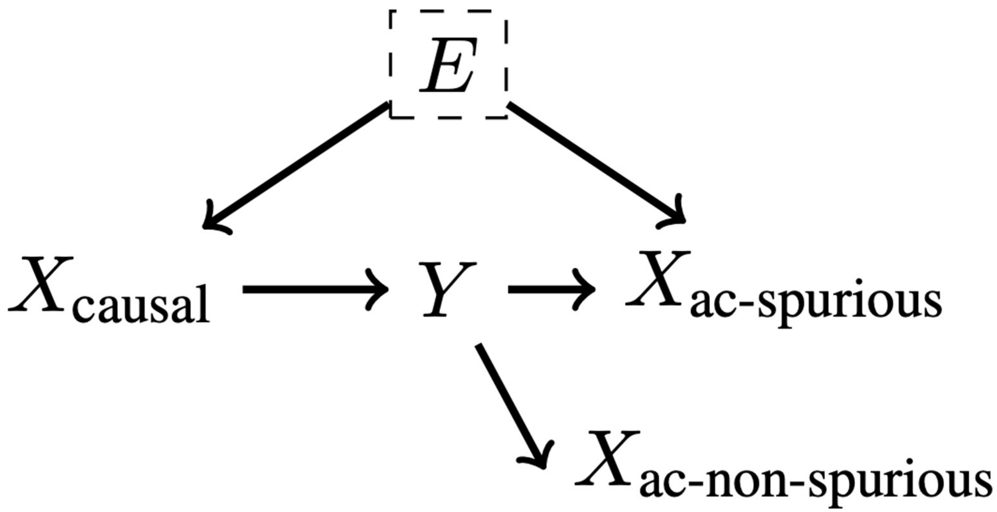
- X: features, Y: label (usually discrete)
- *Source* distributions $P_{S_k}(X,Y)$
- Learn model that works well on unknown *Target* distributions $P'(X,Y) \in Q$
- We allow $P'(X,Y)$ to change in certain ways relative to $P_{S_k}(X,Y)$ (defined via causal graphs)
  - Including changes to $P'(Y|X)$
- Our approach relies on *multi-environment calibration*





Formalizing OOD and spurious correlations

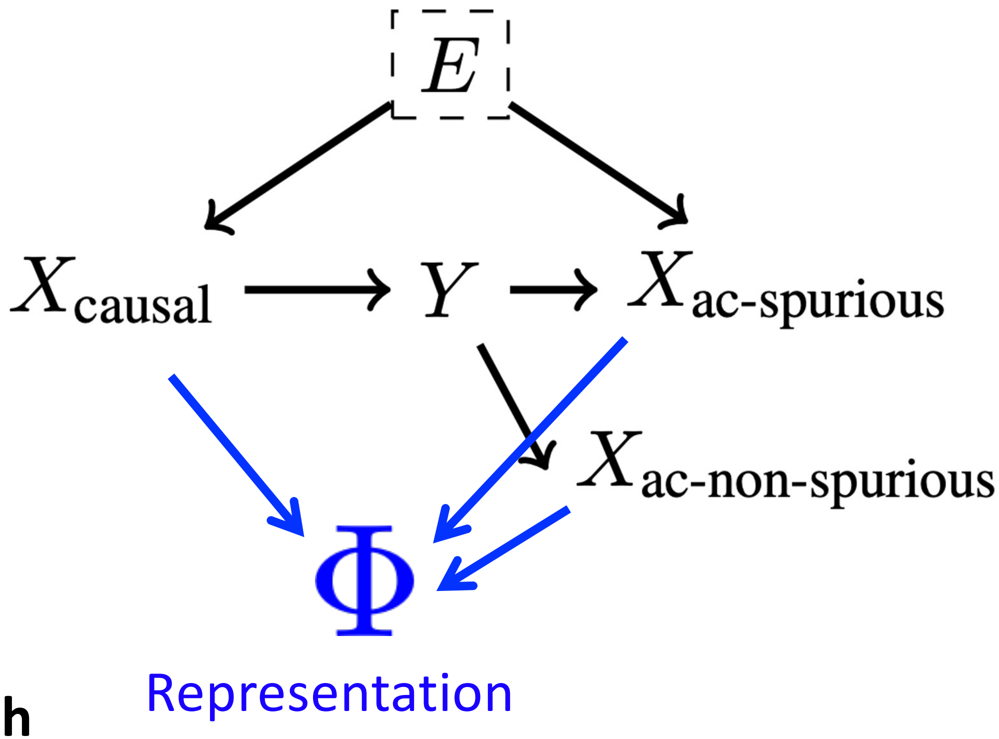
- Causal graph encoding assumptions about how target domain can differ from source (train) domains
- Example:
  E: hospital
  Y: disease
  X_{causal}: patient demographics
  $X_{\text{ac-non-spurious}}$: "disease pixels"
  $X_{\text{ac-spurious}}$: pixels caused by hospital specific imaging setup



- **We don't know a-priori which is which**
- Note no arrow from E to Y!
- At test time we observe a **new environment** $E = e, e \in \mathcal{E}$
  ($do(E = e)$ for previously unseen value e)

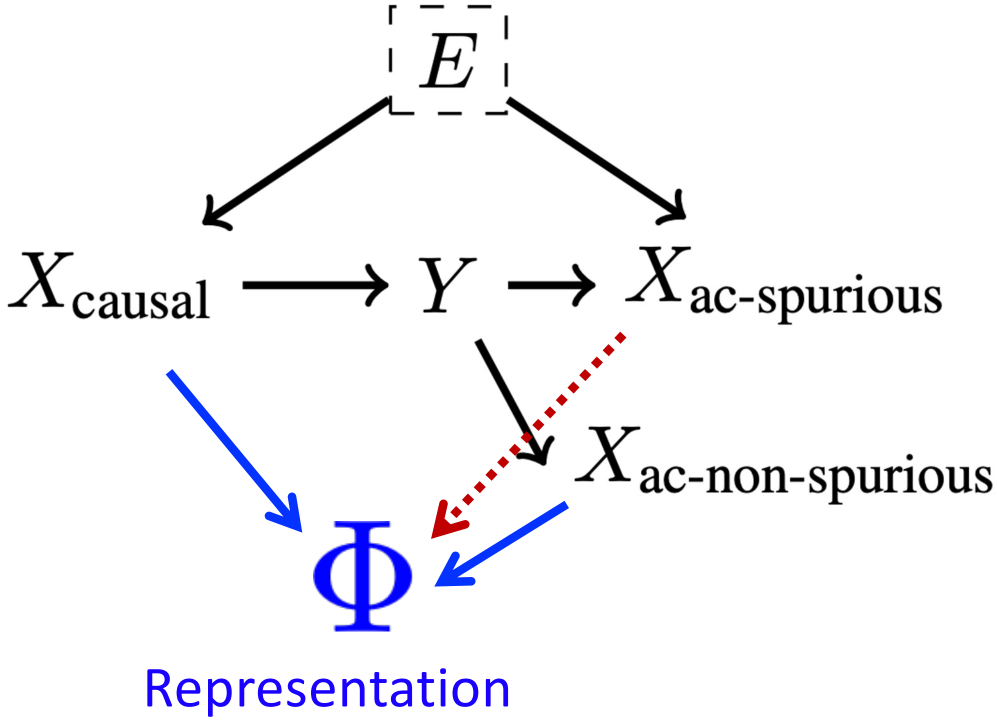
Spurious-free representations

- Causal graph encoding assumptions about how target domain can differ from source (train) domains
- Example:
  E: hospital
  Y: disease
  X_{causal}: patient demographics
  $X_{\text{ac-non-spurious}}$: "disease pixels"
  $X_{\text{ac-spurious}}$: pixels caused by hospital specific imaging setup
- **We don't know a-priori which is which**
- Note no arrow from E to Y!
- At test time we observe a **new environment** $E = e, e \in \mathcal{E}$ ($do(E = e)$ for previously unseen value e)



Formalizing spuriousness

- Models using spurious features can incur arbitrarily high risk when test is previously unseen environment $E = e$
- The problem occures when using $X_{\text{ac}-\text{spurious}}$ (Collider)
- and when not using X_{causal}

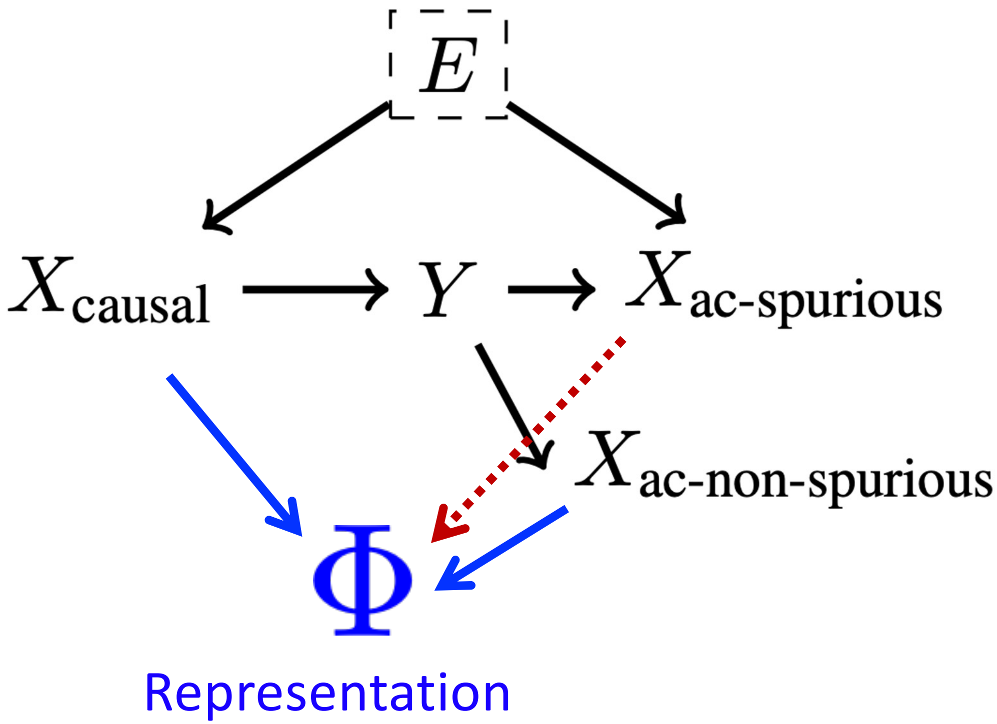


A representation $\Phi(x)$ has spurious correlatons w.r.t. to Y and E if $Y \not\perp E | \Phi(x)$

Formalizing spuriousness

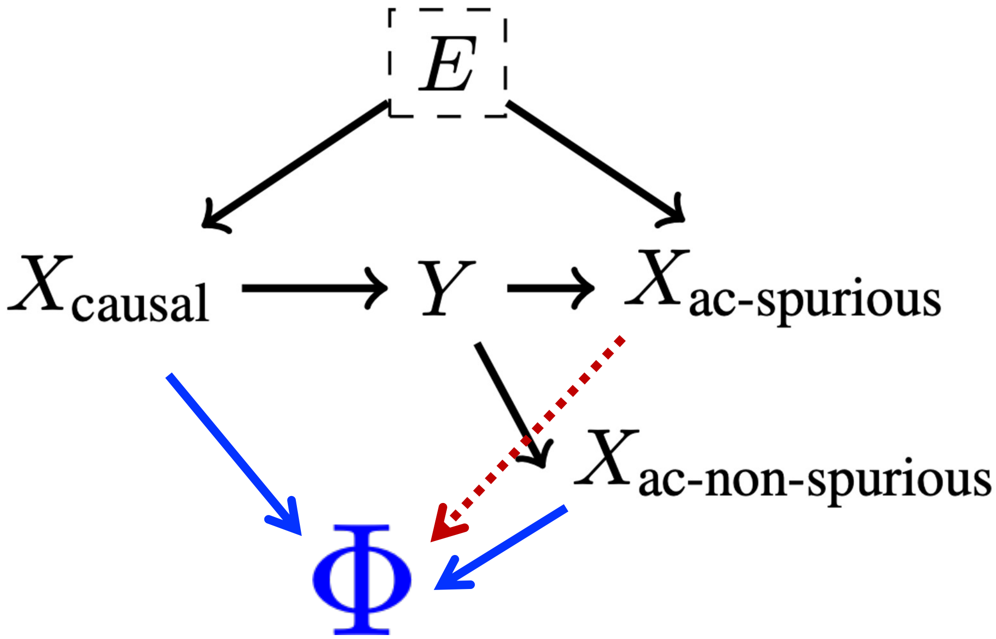
- The problem occures when using $X_{\text{ac}-\text{spurious}}$
  - Collider!
- Shares the spirit of Invariant Causal Prediction (ICP) (*Peters et al. 16*) and Invariant Risk Minimization (IRM) (*Arjovsky et al. 19*)

A representation $\Phi(x)$ has spurious correlatons w.r.t. to Y and E if $Y \not\perp E \mid \Phi(x)$



Optimizing for stability

- Assume we have access to samples from multiple e ∈ E
- How can we learn an **informative** representation $\Phi(x)$ such that $Y \perp\!\!\!\perp E \mid \Phi(x)$?
  (no spurious correlations)
- Seems like a difficult optimization problem
- We show this is equivalent to a more approachable problem: **Multi-environment Calibration**
- Allows us to adapt a huge set of pre-existing tools from the calibration literature



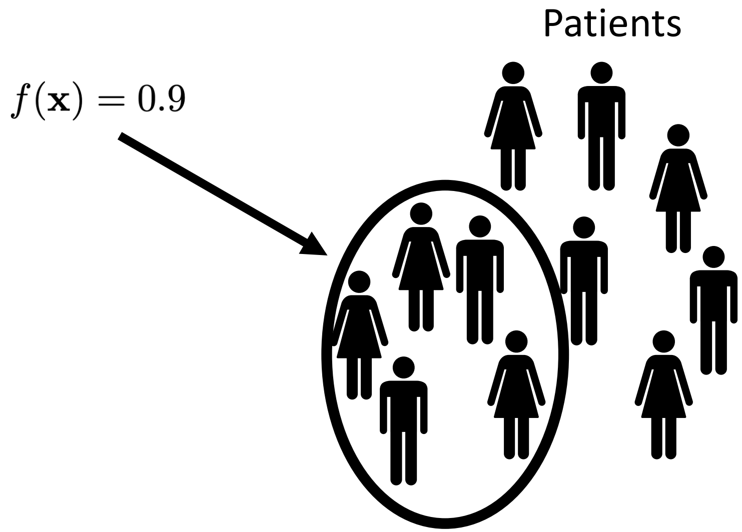
A representation $\Phi(x)$ has spurious correlatons w.r.t. to Y and E if $Y \not\perp\!\!\!\perp E \mid \Phi(x)$

Calibration

- A classifier $f: \mathcal{X} \to [0,1]$ is **calibrated** if $\mathbb{E}[Y|f(X)] = f(x)$
- Calibration: probabilities of events match predictions

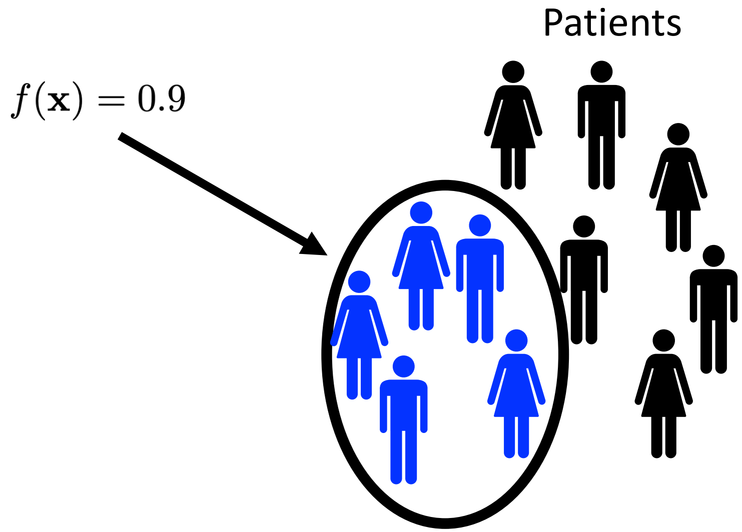
Calibration

- A classifier $f: \mathcal{X} \to [0,1]$ is **calibrated** if $\mathbb{E}[Y|f(X)] = f(x)$
- Calibration: probabilities of events match predictions
- $f(x)$: probability of tumor



Calibration

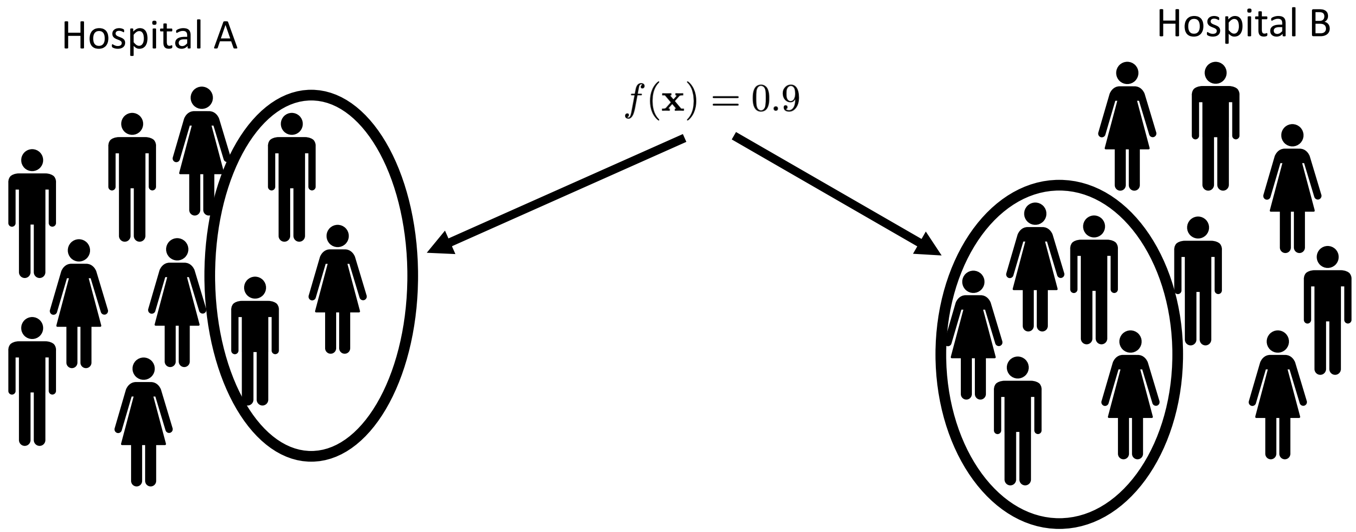
- A classifier $f: \mathcal{X} \to [0,1]$ is **calibrated** if $\mathbb{E}[Y|f(X)] = f(x)$
- Calibration: probabilities of events match predictions
- $f(x)$: probability of tumor



- Calibration: 90% of patients with prediction 0.9 indeed have tumor

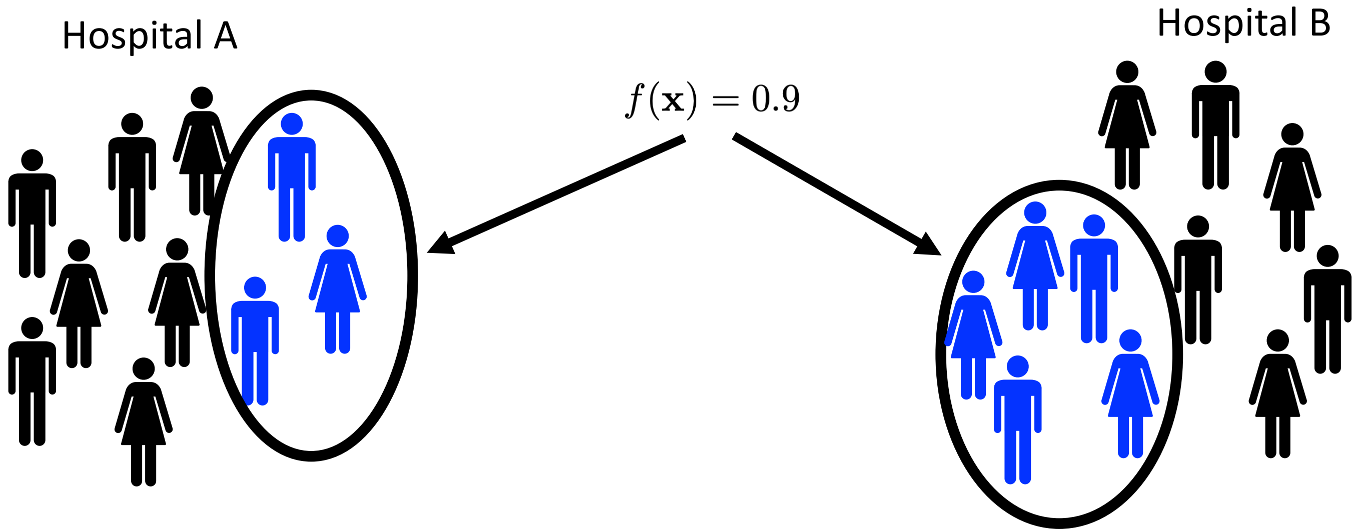
Multi Environment Calibration

- A classifier $f: \mathcal{X} \to [0,1]$ is **calibrated** if $\mathbb{E}[Y|f(X)] = f(x)$
- Calibration: probabilities of events match predictions
- $f(x)$: probability of tumor



Calibration

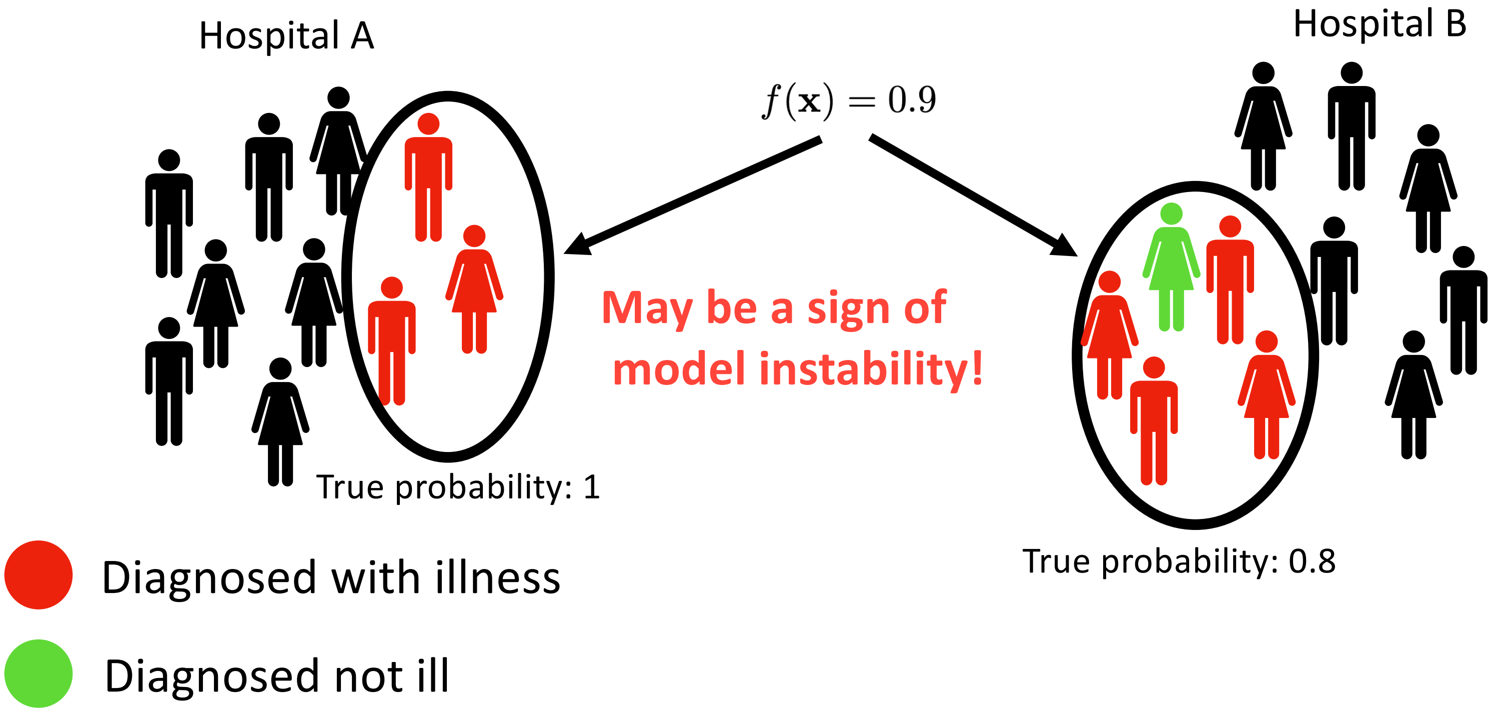
- A classifier $f: \mathcal{X} \to [0,1]$ is **calibrated** if $\mathbb{E}[Y|f(X)] = f(x)$
- Calibration: probabilities of events match predictions
- $f(x)$: probability of tumor



- Calibration: 90% of patients with prediction 0.9 indeed have tumor

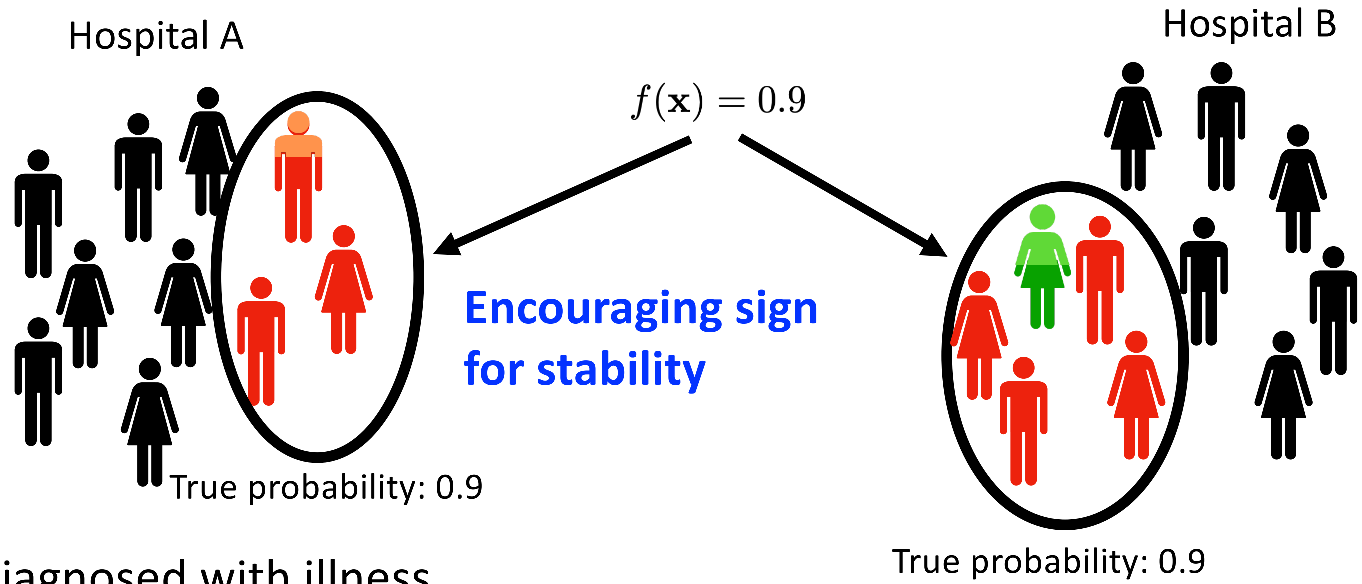
Calibration

- A classifier $f: \mathcal{X} \rightarrow [0,1]$ is **calibrated** if $\mathbb{E}[Y|f(X)] = f(x)$
- Calibration: probabilities of events match predictions
- $f(x)$: probability of tumor



Calibration

- A classifier $f : \mathcal{X} \rightarrow [0,1]$ is **calibrated** if $\mathbb{E}[Y|f(X)] = f(x)$
- Calibration: probabilities of events match predictions
- $f(x)$: probability of tumor



Diagnosed with illness

Diagnosed not ill

Invariance on Training Environments E_{train}

- Consider the representation $\Phi(x) = f(x)$, where $f(x)$ is a binary classifier
- Avoid spurious correlations by enforcing $Y \perp\!\!\!\perp E | f(x)$ on training environments E_{train}
- Let us call such classifiers *invariant classifiers.*

Definition. *Let* $f : \mathcal{X} \to [0, 1]$, *it is an invariant classifier w.r.t* E_{train} *if for all* $\alpha \in [0, 1]$ *and environments* $e_i, e_j \in E_{train}$ *where* α *is in the range of* f *restricted to each of them:*

$$\mathbb{E}[Y \mid f(X) = \alpha, E = e_i] = \mathbb{E}[Y \mid f(X) = \alpha, E = e_j].$$

Calibration on Training Environments E_{train}

- Seemingly unrelated to spurious correlations
- We are interested in calibration on *all training environments simultaneously*

Definition. *Let* $f : \mathcal{X} \rightarrow [0,1]$ *and* $P[X,Y]$ *be a joint distribution over the features and label. Then* $f(\mathbf{x})$ *is calibrated w.r.t to* P *if for all* $\alpha \in [0,1]$ *in the range of* f*:*

$$\mathbb{E}_P[Y \mid f(X) = \alpha] = \alpha.$$

In the multiple environments setting, $f(\mathbf{x})$ *is calibrated on* E_{train} *if for all* $e_i \in E_{train}$ *and* α *in the range of* f *restricted to* e_i*:*

$$\mathbb{E}[Y \mid f(X) = \alpha, E = e_i] = \alpha.$$

Invariance and Calibration on E_{train}

$\boxed{\text{Invariance: } Y \perp\!\!\!\perp E \mid f(x)}$

$\boxed{\text{Calibration: } \mathbb{E}[Y \mid f(x)] = f(x)]}$

Invariance and Calibration on E_{train}

Invariance:
$Y \perp\!\!\!\perp E \mid f(x)$

Calibration:
$\mathbb{E}[Y \mid f(x)] = f(x)]$

Lemma

If a binary classifier f *is invariant w.r.t* E_{train} *then there exists a function* $g: [0,1] \to [0,1]$ *such that:*

(i) $g \circ f$ *is calibrated on all training environments, and*

(ii) *the MSE of* $g \circ f$ *on each environment does not exceed that of* f

Invariance and Calibration on E_{train}

Invariance:
$Y \perp\!\!\!\perp E \mid f(x)$

$\widetilde{\Longleftrightarrow}$

Calibration:
$\mathbb{E}[Y \mid f(x)] = f(x)$

Lemma

If a binary classifier f *is invariant w.r.t* E_{train} *then there exists a function* $g: [0,1] \rightarrow [0,1]$ *such that:*

(i) $g \circ f$ *is calibrated on all training environments, and*

(ii) *the MSE of* $g \circ f$ *on each environment does not exceed that of* f

Conversely, if a classifier is calibrated on all training environments, it is invariant w.r.t. E_{train}

Invariance and Calibration on E_{train}

Lemma
If a binary classifier f is invariant w.r.t E_{train} then there exists a function $g: [0,1] \to [0,1]$ such that:
(i) $g \circ f$ *is calibrated on all training environments, and*
(ii) *the MSE of $g \circ f$ on each environment does not exceed that of f*

Conversely, if a classifier is calibrated on all training environments, it is invariant w.r.t. E_{train}

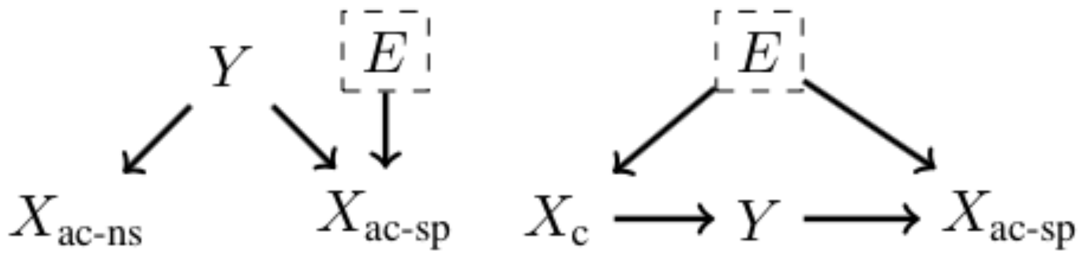
- Similar to invariant representations of Invariant Risk Minimization (IRM) [*Arjovsky 19*], yet with several differences:
  - Calibration does not involve optimality with respect to a specific loss function: IRM results in invariant classifier only when applied with logistic or squared loss
  - IRM cannot be effectively optimized, while more tractable IRMv1 does not guarantee invariances (Kamath et al. 2021)
  - We show multi-domain calibration correctly identfies Kamath et al. 2021 invariances

Questions of Interest

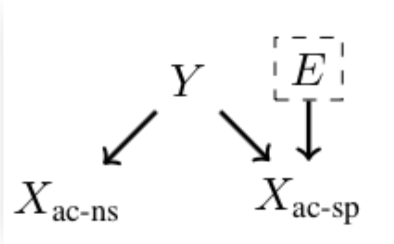
- **Generalization:** assume $f(x)$ is calibrated on all E_{train} , when does it imply calibration on $\mathcal{E}$?
- **Spurious correlations:** what can we formally claim about a calibrated classifier's use of $X_{\mathrm{ac-spurious}}$?

Simplified Settings: Linear-Gaussian Models

- We consider settings with features generated from multivariate Gaussians
  - Each environment parameterized by mean vectors and covariance matrices, $\mathcal{E} = \{(\mu, \Sigma) | \mu \in \mathbb{R}^d, \Sigma \in PSD_{d \times d}\}$
  - Two scenarios:



Classification with Invariant and Spurious Features



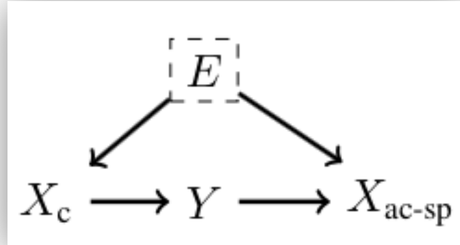
- $\mathbf{E}_{train}$ consists of k training environments
- Dimension of spurious features is d_{sp}
- For each environment (μ_i, Σ_i), data is generated by:

$$y = \begin{cases} 1 & \text{w.p } \eta \\ -1 & \text{o.w} \end{cases} \qquad \begin{array}{l} X_{\text{ac-ns}} \mid Y = y \sim \mathcal{N}(y\mu_{\text{ns}}, \Sigma_{\text{ns}}), \\ X_{\text{ac-sp}} \mid Y = y \sim \mathcal{N}(y\mu_i, \Sigma_i) \end{array}$$

- Learn linear classifier $f(\mathbf{x}) = \sigma(\mathbf{w}^\top \mathbf{x} + b)$, where $\sigma : \mathbb{R}^{d_{\text{sp}} + d_{\text{ns}}} \to [0, 1]$ invertible
- **Theorem:** given $k > 2d_{\text{sp}}$ training environments, under mild non-degeneracy conditions* any classifier that is calibrated on $\mathbf{E}_{train}$ has weights zero on $X_{\text{ac-sp}}$

*a general position assumption (μ_i, Σ_i)

Regression with Covariate Shift and Spurious Features



- Similar setting, but for regression with causal features
- Dimensions of features are d_c, d_{sp}
- For each environment $(\mu_i^c, \Sigma_i^c, \mu_i, \Sigma_i)$ data is generated by:

$$X_c \sim \mathcal{N}(\mu_i^c, \Sigma_i^c) \quad Y = \mathbf{w}_c^{*\top} \mathbf{x}_c + \xi,\ \xi \sim \mathcal{N}(0, \sigma_y^2) \quad X_{\text{ac-sp}} = y\mu_i + \eta,\ \eta \sim \mathcal{N}(\mathbf{0}, \Sigma_i)$$

- **Theorem:** given $k > \max\{d_c + 2, d_{sp}\}$ training environments, under mild non-degeneracy conditions the **only** multi-environment calibrated predictor is

$$f^*(\mathbf{x}) = \mathbf{w}_c^{*\top} \mathbf{x}_c$$

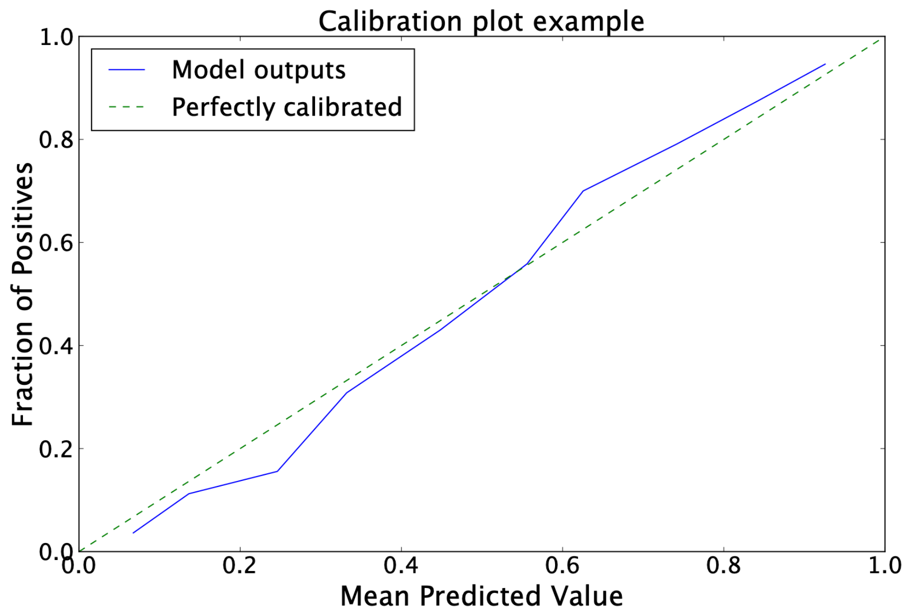
Conclusions from Motivating Examples

- In simple cases, calibration across training domains:
  - Discards $X_{\text{ac-spurious}}$
  - Achieves OOD calibration if number of environments is linear in number of features
- Here calibration = discarding $X_{\text{ac-spurious}}$ = bounded worst-case risk
  - This is not trivial for non-linear models, calls for further analysis
- ***Theory → Practice?***

Tools for Calibration

- **Calibration Plots** visual representation of calibration in binary problems
  - $[0,1]$ interval divided to B bins, $f(x)$ placed into appropriate bin
  - Average confidence in each bin plotted against accuracy.
- **ECE Score**
  - Scalar summary of the curve, averaging deviations from diagonal

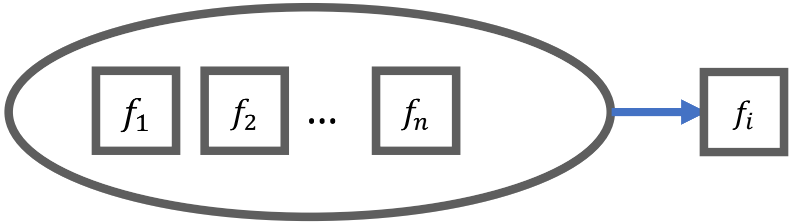
$$ECE = \sum_{b=1}^{B} \frac{n_b}{N} |acc(b) - co$$



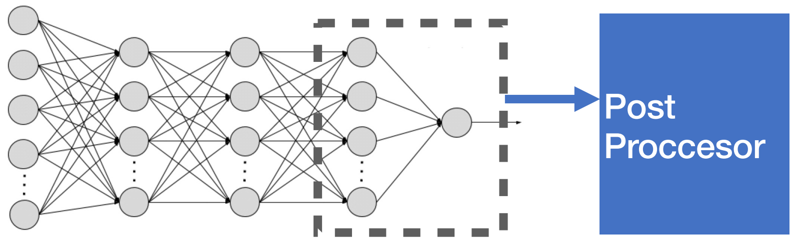
$acc(b)$ is mean number of errors in bin b
$conf(b)$ is mean of $f(x)$ in bin b

Achieving OOD generalization via improved multi-domain calibration: in practice

- **Model Selection**



- **Post-Processing**



- **Training full model:** learn classifier $f_\theta(x)$

$$\min_\theta \sum_{e \in E_{\text{train}}} l^e(f_\theta) + \lambda \cdot r(f_\theta)$$

Tools for Calibration - Post Processing

- **Isotonic Regression:** classic tools that learns *monotone* transformation $z : \mathbb{R} \to \mathbb{R}$ on model outputs f_i to minimize squared error from label:

$$\arg\min_{z} \frac{1}{N} \sum_{i=1}^{N} \left(z(f_i) - y_i\right)^2$$

- **Robust Isotonic Regression (new):** we suggest a variation to bound worst-domain calibration error:

$$\arg\min_{z} \max_{e \in \mathbf{E}_{\text{train}}} \frac{1}{N} \sum_{i=1}^{N} \left(z(f_i) - y_i\right)^2$$

A Multi-Domain Calibration Regularizer

- Our regularizer builds on the kernel based regularizer of [Kumar et al. 18]

  Maximum Mean Calibration Error (*MMCE*): for dataset $D = \{\mathbf{x}_i, y_i\}_{i=1}^{m}$

$$r_{\mathrm{MMCE}}^{D}(f_\theta) = \frac{1}{m^2} \sum_{i,j \in D} (c_i - f_{\theta;i})(c_j - f_{\theta;j}) k(f_{\theta;i}, f_{\theta;j})$$

c_i : correctness of $f_\theta(x)$ on example i

$f_{\theta;i}$: confidence of $f_\theta(x)$ on example i

$k(\cdot,\cdot)$: universal kernel

A Multi-Domain Calibration Regularizer

- Our regularizer builds on the kernel based regularizer of [Kumar et al. 18]

  Maximum Mean Calibration Error (*MMCE*): for dataset $D = \{\mathbf{x}_i, y_i\}_{i=1}^{m}$

$$r_{\mathrm{MMCE}}^{D}(f_\theta) = \frac{1}{m^2} \sum_{i,j \in D} (c_i - f_{\theta;i})(c_j - f_{\theta;j}) k(f_{\theta;i}, f_{\theta;j})$$

- Calibration Loss Over Environments (*CLOvE*): datasets D_e **for each** $e \in E_{\mathrm{train}}$

$$r_{\mathrm{CLOvE}}(f_\theta) = \sum_{e \in E_{\mathrm{train}}} r_{\mathrm{MMCE}}^{D_e}(f_\theta)$$

Key property:
$r_{\mathrm{CLOvE}}(f_\theta) = 0$ **if and only if** $f_\theta(x)$ **is calibrated on** E_{train}

Colored MNIST

- Example from [Kim et al. 18, Arjovsky et al. 19], introduce spurious correlations with color to MNIST digits
- Further simplified by [Kamath et al. 21] to “Two-Bit” environments”

$$Y \leftarrow \mathrm{Rad}(0.5), X_1 \leftarrow Y \cdot \mathrm{Rad}(\alpha), X_2 \leftarrow Y \cdot \mathrm{Rad}(\beta)$$

0-4 5-9

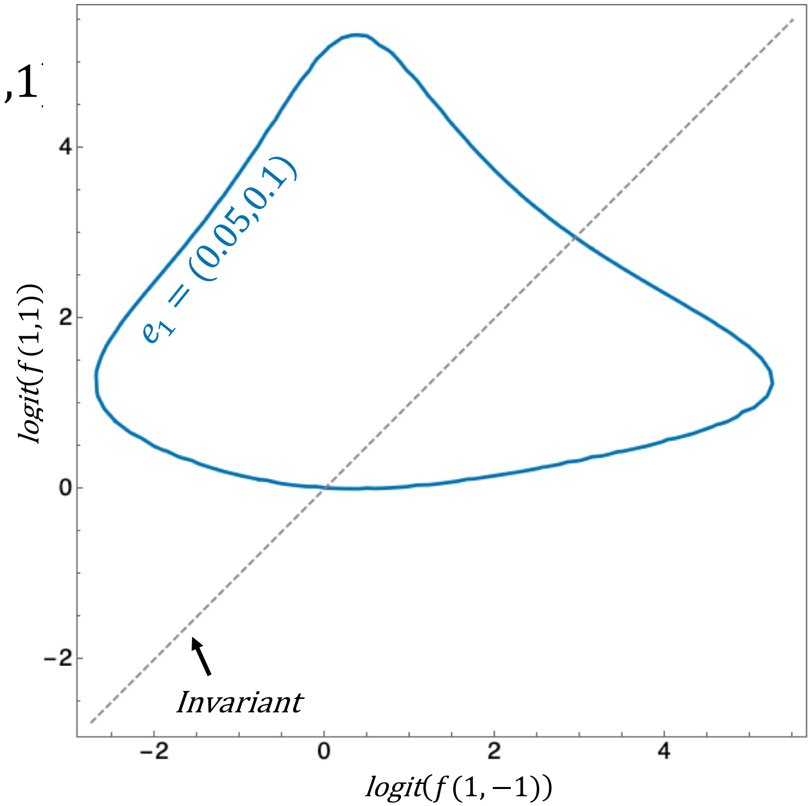
Correlation w/ Digit

Correlation w/ Color

- Train on $e_1 = (\alpha, \beta_1), e_2 = (\alpha, \beta_2)$, test on $e_{test} = (\alpha, \beta_{test})$
- Motivation for IRM [Arjovsky et al.], however turns out IRM is not a solution!

Colored MNIST

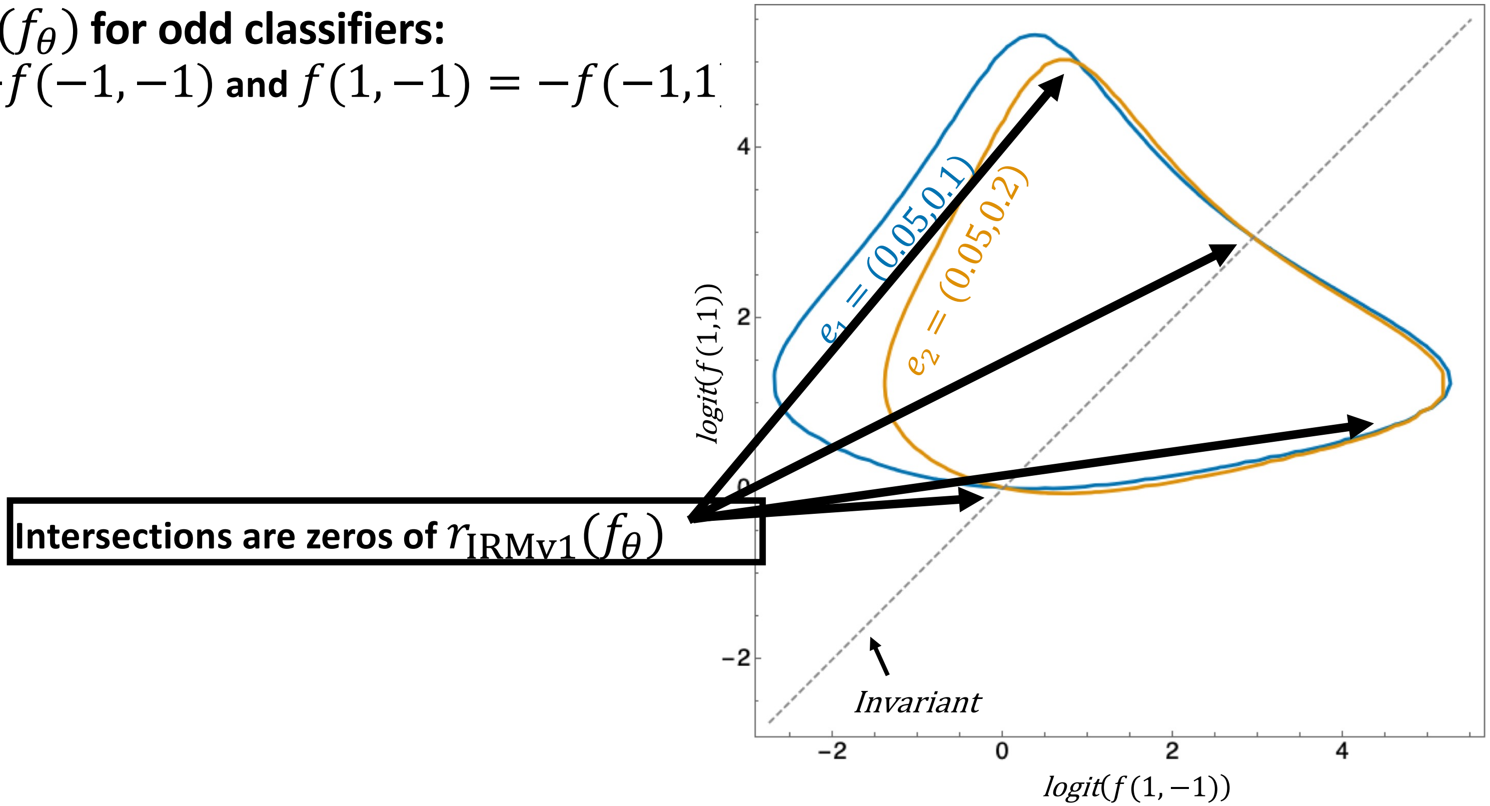
Zeros of $r^e_{\text{IRMv1}}(f_\theta)$ for odd classifiers:
i.e. $f(1,1) = -f(-1,-1)$ **and** $f(1,-1) = -f(-1,1)$



P. Kamath, A. Tangella, D. J. Sutherland, and N. Srebro. Does invariant risk minimization capture invariance? In AISTATS, 2021

Colored MNIST

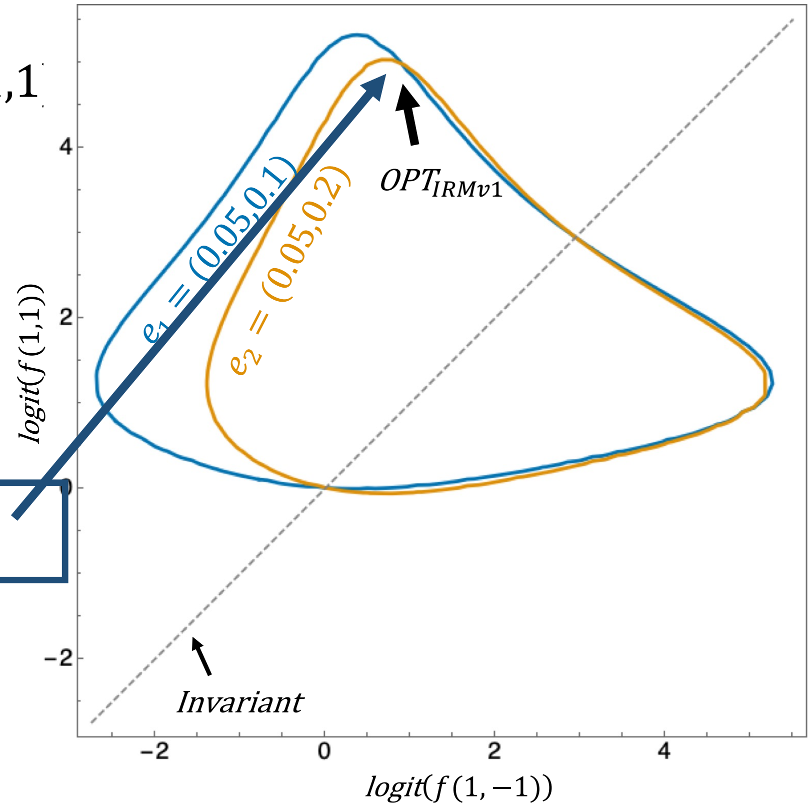
Zeros of $r^e_{\text{IRMv1}}(f_\theta)$ for odd classifiers:
i.e. $f(1,1) = -f(-1,-1)$ **and** $f(1,-1) = -f(-1,1)$



P. Kamath, A. Tangella, D. J. Sutherland, and N. Srebro. Does invariant risk minimization capture invariance? In AISTATS, 2021

Colored MNIST

Zeros of $r^{e}_{\mathrm{IRMv1}}(f_\theta)$ for odd classifiers:
i.e. $f(1,1) = -f(-1,-1)$ **and** $f(1,-1) = -f(-1,1$



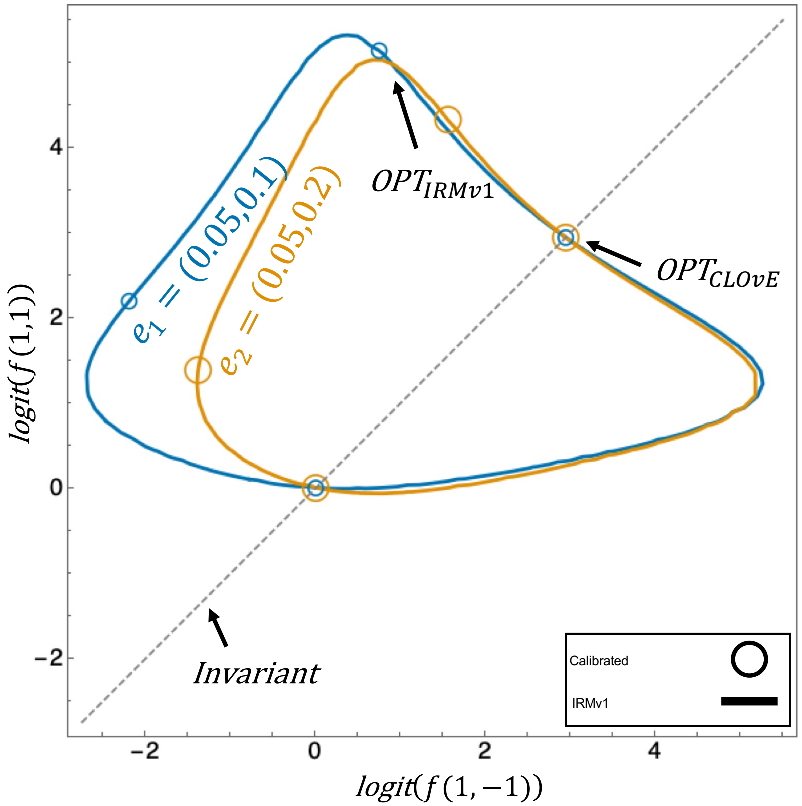
Intersection with minimal empirical loss is not invariant!

P. Kamath, A. Tangella, D. J. Sutherland, and N. Srebro. Does invariant risk minimization capture invariance? In AISTATS, 2021

CLOvE Achieves Invariance in Colored MNIST

Zeros of $r^e_{\mathrm{IRMv1}}(f_\theta)$ and $r^e_{\mathrm{CLOvE}}(f_\theta)$ for odd classifiers:

CLOvE discards spurious feature when IRMv1 doesn't!



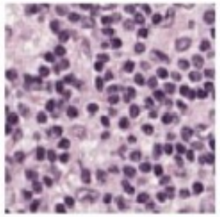
*On actual dataset: comparable performance in most settings, failure case can be reproduced

P. Kamath, A. Tangella, D. J. Sutherland, and N. Srebro. Does invariant risk minimization capture invariance? In AISTATS, 2021

Experiments on Large Scale Datasets

WILDS
[Koh et al. 20]

Camelyon17



d = Hospital 1

FMoW

2002 / Americas

+ other tasks that involve regression or sub-population shift

Fine-tune representation from last layer with 3 fully-connected layers

Our variation on Isotonic Regression

Post Processing Isotonic Regression

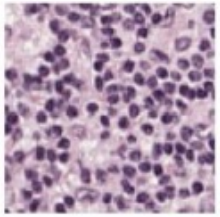
Algorithm	*FMoW* Orig.	Naive Cal.	Rob. Cal.	CLOvE	*Camelyon17* Orig.	Naive Cal.	Rob. Cal.	CLOvE
ERM	32.63 (<u>0.016</u>)	33.09 (0.021)	37.19 (0.035)	**44.16** (0.018)	66.66 (0.144)	71.23 (0.089)	71.22 (0.086)	**75.75** (<u>0.049</u>)

Experiments on Large Scale Datasets

WILDS

[Koh et al. 20]

Camelyon17



d = Hospital 1

FMoW

2002 / Americas

Fine-tune representation from last layer with 3 fully-connected layers

Our variation on Isotonic Regression

Post Processing Isotonic Regression

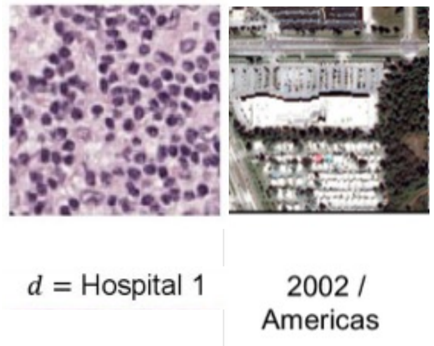
Algorithm	*FMoW* Orig.	Naive Cal.	Rob. Cal.	CLOvE	*Camelyon17* Orig.	Naive Cal.	Rob. Cal.	CLOvE
ERM	32.63 (0.016)	33.09 (0.021)	37.19 (0.035)	**44.16** (0.018)	66.66 (0.144)	71.23 (0.089)	71.22 (0.086)	**75.75** (0.049)
DeepCORAL	31.73 (0.01)	31.75 (0.01)	33.86 (0.016)	**40.05** (0.009)	72.44 (0.044)	75.97 (0.054)	76.8 (0.065)	**79.96** (0.039)
IRM	31.33 (0.012)	31.81 (0.016)	34.41 (0.015)	**42.24** (0.014)	70.87 (0.068)	73.25 (0.066)	73.4 (0.069)	**73.95** (0.061)

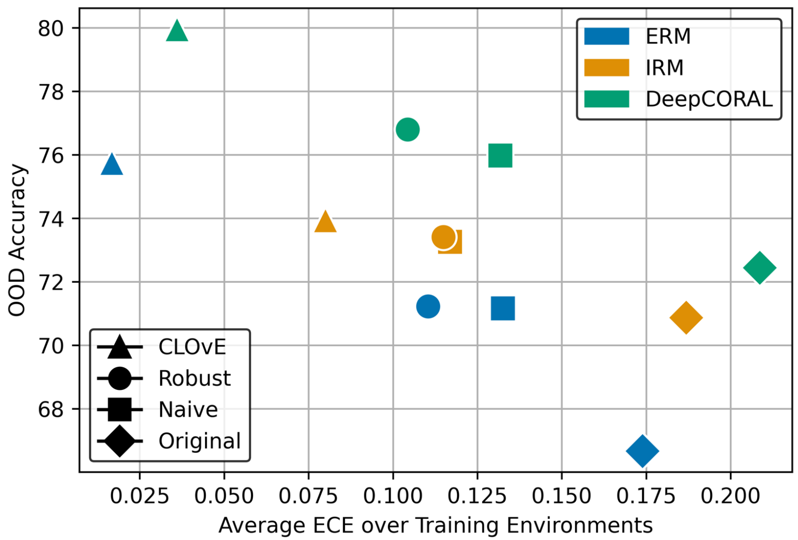
Experiments on Large Scale Datasets

WILDS

[Koh et al. 20]

Camelyon17 FMoW

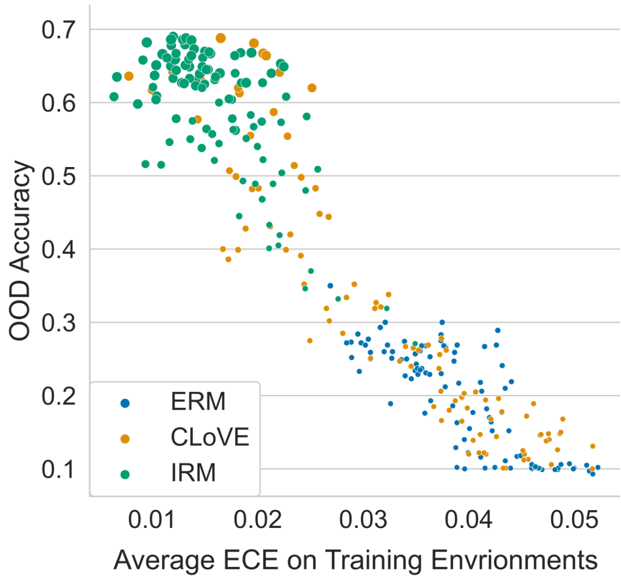




Assessing Stability with Calibration Error

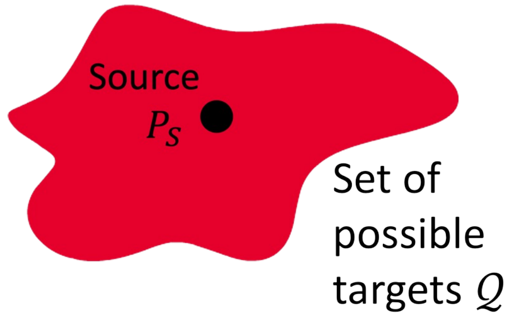
- Suggestion: balance in-domain accuracy and Expected Calibration Error (ECE)
- Colored MNIST: trained 100 models w/ IRM, CLOvE, ERM and random hyperparams

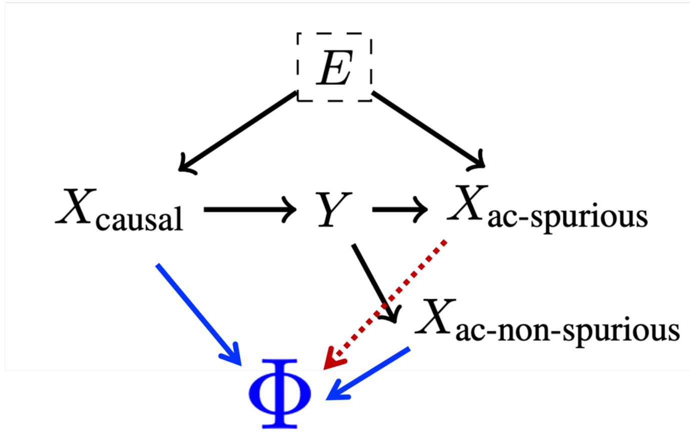
Pearson corr: -0.92



Summary

- Definition of spurious correlations of a representation w.r.t. an "environment" variable E : $Y \not\perp\!\!\!\perp E|\Phi(x)$
- Calibration ≈ invariance w.r.t. E
- With diverse environments: multi-environment calibration $\Longrightarrow$ no spurious correlations (in linear-Gaussian and some other simplified settings)
- Multi-environment calibration improves results on existing (flawed) OOD benchamrks





Open questions

- Is calibration a red herring here?
- Non-linear models
- Number of training environments
- The role of unobserved confounders
- High-dimensional representations
- Problems with overparameterized models
- What if I know about some interventions in the dataset?
  - Generally many ways to add side-information, e.g. other labels