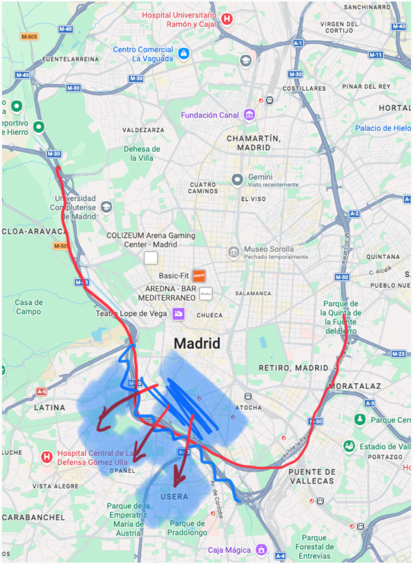
PML 7. Beyond the i.i.d. assumption

Probabilistic Machine Learning Reading Group

Carlos García Meixide

February 4, 2026

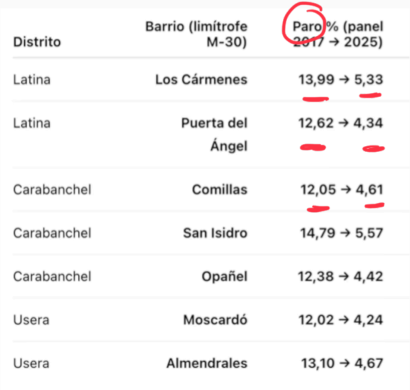
ICMAT



SPILLOVER EFFECT

Unemployment rate decrease

unemployment



Distrito	Barrio (limítrofe M-30)	Paro % (panel 2017 → 2025)
Latina	Los Cármenes	13,99 → 5,33
Latina	Puerta del Ángel	12,62 → 4,34
Carabanchel	Comillas	12,05 → 4,61
Carabanchel	San Isidro	14,79 → 5,57
Carabanchel	Opañel	12,38 → 4,42
Usera	Moscardó	12,02 → 4,24
Usera	Almendrales	13,10 → 4,67

N-550
DP-7803
Museo do Pobo Galego
Catedral de Santiago de Compostela
Mercado de Abastos
Miradoiro da Catedral (Santiago de Compostela)
Hostal Santa Cruz
Santiago de Compostela
Postgrado en Odontología - UCAM - PgO UCAM (IDSQ)
Campus Vida (USC)
Goiko
Lux Apartamentos Santiago
Synergym Santiago Ensanche
Cidade
Galuresa Pontepedriña (Santiago de Compostela)
Hospital Clínico Universitario de Santiago
SAR
SC-20
7B
7A
Hotel Santiago Plaza Affiliated by Meliá
El Corte Inglés

topography

demographics

salaries

Y : "unemployment rate"

A : "building ringroad"

W :

$$\boxed{p(Y,A,W)} = \underset{\text{market}}{\boxed{p(Y|A,W)}}\,\overset{\text{policy}}{\boxed{p(A|W)}}\,\boxed{p(W)}$$

Vitamin C

SCURVY

INTERNAL VALIDITY

Y: "collagen"

A: "oC y/n"

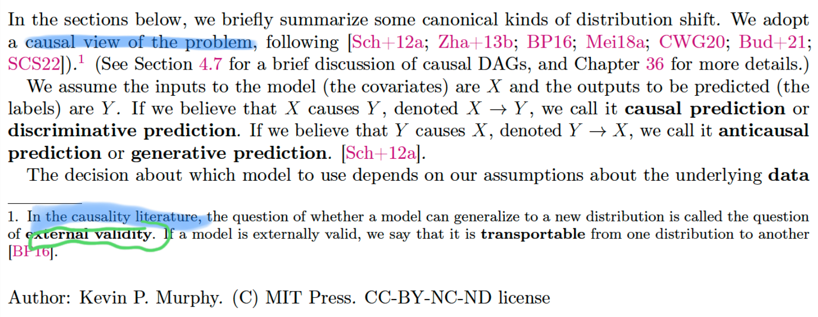
W: "health"

rand controlled trial

$\mathbb{E}Y^1 - \mathbb{E}Y^0$

$$\begin{tikzcd}
Y & A \arrow[l] \arrow[l, bend right, dashed, no head] \\
& W \arrow[ul] \arrow[u]
\end{tikzcd}$$

$$\mathbb{E}\, Y^{A=1} = \mathbb{E}_{W} \boxed{\mathbb{E}[Y \mid A=1, W]}$$

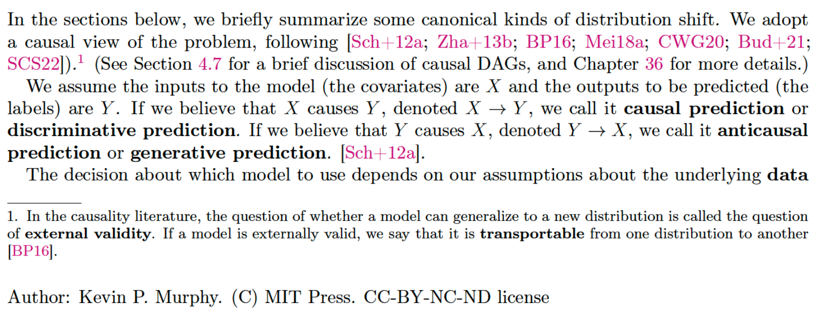


In the sections below, we briefly summarize some canonical kinds of distribution shift. We adopt a causal view of the problem, following [Sch+12a; Zha+13b; BP16; Mei18a; CWG20; Bud+21; SCS22]).[1] (See Section 4.7 for a brief discussion of causal DAGs, and Chapter 36 for more details.)

We assume the inputs to the model (the covariates) are X and the outputs to be predicted (the labels) are Y. If we believe that X causes Y, denoted $X \rightarrow Y$, we call it **causal prediction** or **discriminative prediction**. If we believe that Y causes X, denoted $Y \rightarrow X$, we call it **anticausal prediction** or **generative prediction**. [Sch+12a].

The decision about which model to use depends on our assumptions about the underlying **data**

[1] In the causality literature, the question of whether a model can generalize to a new distribution is called the question of **external validity**. If a model is externally valid, we say that it is **transportable** from one distribution to another [BP16].



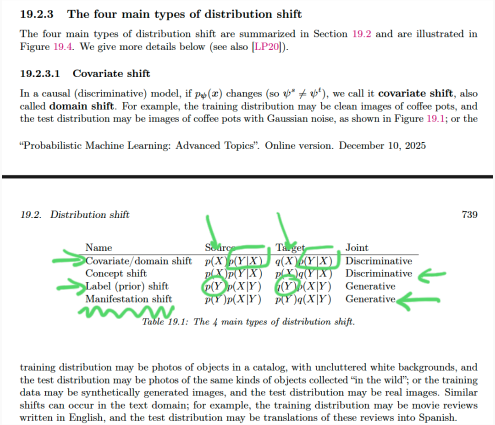
In the sections below, we briefly summarize some canonical kinds of distribution shift. We adopt a causal view of the problem, following [Sch+12a; Zha+13b; BP16; Mei18a; CWG20; Bud+21; SCS22]).[1] (See Section 4.7 for a brief discussion of causal DAGs, and Chapter 36 for more details.)

We assume the inputs to the model (the covariates) are X and the outputs to be predicted (the labels) are Y. If we believe that X causes Y, denoted $X \rightarrow Y$, we call it **causal prediction** or **discriminative prediction**. If we believe that Y causes X, denoted $Y \rightarrow X$, we call it **anticausal prediction** or **generative prediction**. [Sch+12a].

The decision about which model to use depends on our assumptions about the underlying **data**

[1] In the causality literature, the question of whether a model can generalize to a new distribution is called the question of **external validity**. If a model is externally valid, we say that it is **transportable** from one distribution to another [BP16].

> *"A method, algorithm or theory should be judged from its ability to predict in new contexts."*
>
> — Sir David R. Cox



19.2.3 The four main types of distribution shift

The four main types of distribution shift are summarized in Section 19.2 and are illustrated in Figure 19.4. We give more details below (see also [LP20]).

19.2.3.1 Covariate shift

In a causal (discriminative) model, if $p_\psi(\boldsymbol{x})$ changes (so $\psi^s \neq \psi^t$), we call it **covariate shift**, also called **domain shift**. For example, the training distribution may be clean images of coffee pots, and the test distribution may be images of coffee pots with Gaussian noise, as shown in Figure 19.1; or the

“Probabilistic Machine Learning: Advanced Topics”. Online version. December 10, 2025

19.2. Distribution shift 739

Name	Source	Target	Joint
Covariate/domain shift	$p(X)p(Y\|X)$	$q(X)p(Y\|X)$	Discriminative
Concept shift	$p(X)p(Y\|X)$	$p(X)q(Y\|X)$	Discriminative
Label (prior) shift	$p(Y)p(X\|Y)$	$q(Y)p(X\|Y)$	Generative
Manifestation shift	$p(Y)p(X\|Y)$	$p(Y)q(X\|Y)$	Generative

Table 19.1: The 4 main types of distribution shift.

training distribution may be photos of objects in a catalog, with uncluttered white backgrounds, and the test distribution may be photos of the same kinds of objects collected “in the wild”; or the training data may be synthetically generated images, and the test distribution may be real images. Similar shifts can occur in the text domain; for example, the training distribution may be movie reviews written in English, and the test distribution may be translations of these reviews into Spanish.

The war for generalizability

Heterogeneity

- Average effects are a convenience. Bayesian
- Clinical trials use convenience samples through weird inclusion criteria.
- Marginalization is useless if there is distribution shift.

Randomization

freq.

- Modelling heterogeneity turns noise into findings
- Slicing into subgroups breaks randomization
- Biological variation is too small in comparison to noise.

The war for generalizability

Heterogeneity

- Average effects are a convenience.
- Clinical trials use convenience samples through weird inclusion criteria.
- Marginalization is useless if there is distribution shift.

$$\boxed{P_{\mathbf{X},\theta}(\mathcal{I}(\mathbf{X}) \ni \theta_0) \geq 1 - \alpha}$$

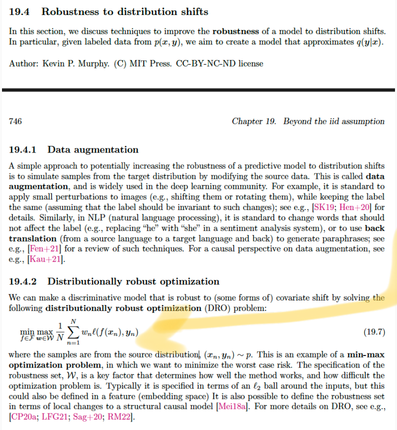
Randomization

- Modelling heterogeneity turns noise into findings
- Slicing into subgroups breaks randomization
- Biological variation is too small in comparison to noise.

$$P_{\mathbf{X}|\theta}(\mathcal{I}(\mathbf{X}) \ni \theta_0) \geq 1 - \alpha \; {+ o_p(1)}$$

ado
iid

Adversarial regimes



19.4 Robustness to distribution shifts

In this section, we discuss techniques to improve the **robustness** of a model to distribution shifts. In particular, given labeled data from $p(\boldsymbol{x}, y)$, we aim to create a model that approximates $q(y|\boldsymbol{x})$.

Author: Kevin P. Murphy. (C) MIT Press. CC-BY-NC-ND license

746 *Chapter 19. Beyond the iid assumption*

19.4.1 Data augmentation

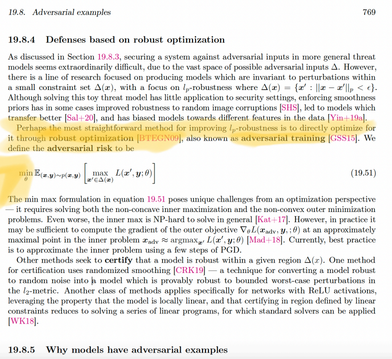
A simple approach to potentially increasing the robustness of a predictive model to distribution shifts is to simulate samples from the target distribution by modifying the source data. This is called **data augmentation**, and is widely used in the deep learning community. For example, it is standard to apply small perturbations to images (e.g., shifting them or rotating them), while keeping the label the same (assuming that the label should be invariant to such changes); see e.g., [SK19; Hen+20] for details. Similarly, in NLP (natural language processing), it is standard to change words that should not affect the label (e.g., replacing "he" with "she" in a sentiment analysis system), or to use **back translation** (from a source language to a target language and back) to generate paraphrases; see e.g., [Fen+21] for a review of such techniques. For a causal perspective on data augmentation, see e.g., [Kau+21].

19.4.2 Distributionally robust optimization

We can make a discriminative model that is robust to (some forms of) covariate shift by solving the following **distributionally robust optimization** (DRO) problem:

$$\min_{f \in \mathcal{F}} \max_{\boldsymbol{w} \in \mathcal{W}} \frac{1}{N} \sum_{n=1}^{N} w_n \ell(f(\boldsymbol{x}_n), \boldsymbol{y}_n) \tag{19.7}$$

where the samples are from the source distribution, $(\boldsymbol{x}_n, \boldsymbol{y}_n) \sim p$. This is an example of a **min-max optimization problem**, in which we want to minimize the worst case risk. The specification of the robustness set, $\mathcal{W}$, is a key factor that determines how well the method works, and how difficult the optimization problem is. Typically it is specified in terms of an ℓ_2 ball around the inputs, but this could also be defined in a feature (embedding space) It is also possible to define the robustness set in terms of local changes to a structural causal model [Mei18a]. For more details on DRO, see e.g., [CP20a; LFG21; Sag+20; RM22].



19.8. Adversarial examples 769

19.8.4 Defenses based on robust optimization

As discussed in Section 19.8.3, securing a system against adversarial inputs in more general threat models seems extraordinarily difficult, due to the vast space of possible adversarial inputs Δ. However, there is a line of research focused on producing models which are invariant to perturbations within a small constraint set $\Delta(\boldsymbol{x})$, with a focus on l_p-robustness where $\Delta(\boldsymbol{x}) = \{\boldsymbol{x}' : ||\boldsymbol{x} - \boldsymbol{x}'||_p < \epsilon\}$. Although solving this toy threat model has little application to security settings, enforcing smoothness priors has in some cases improved robustness to random image corruptions [SHS], led to models which transfer better [Sal+20], and has biased models towards different features in the data [Yin+19a].

Perhaps the most straightforward method for improving l_p-robustness is to directly optimize for it through **robust optimization** [BTEGN09], also known as **adversarial training** [GSS15]. We define the **adversarial risk** to be

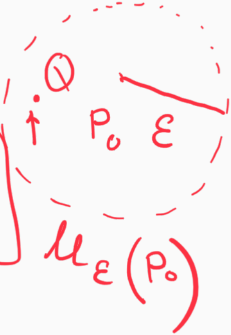
$$\min_{\theta} \mathbb{E}_{(\boldsymbol{x},\boldsymbol{y})\sim p(\boldsymbol{x},\boldsymbol{y})} \left[\max_{\boldsymbol{x}' \in \Delta(\boldsymbol{x})} L(\boldsymbol{x}', \boldsymbol{y}; \theta) \right] \tag{19.51}$$

The min max formulation in equation 19.51 poses unique challenges from an optimization perspective — it requires solving both the non-concave inner maximization and the non-convex outer minimization problems. Even worse, the inner max is NP-hard to solve in general [Kat+17]. However, in practice it may be sufficient to compute the gradient of the outer objective $\nabla_{\theta} L(\boldsymbol{x}_{\text{adv}}, \boldsymbol{y}, ; \theta)$ at an approximately maximal point in the inner problem $\boldsymbol{x}_{\text{adv}} \approx \text{argmax}_{\boldsymbol{x}'} L(\boldsymbol{x}', \boldsymbol{y}; \theta)$ [Mad+18]. Currently, best practice is to approximate the inner problem using a few steps of PGD.

Other methods seek to **certify** that a model is robust within a given region $\Delta(x)$. One method for certification uses randomized smoothing [CRK19] — a technique for converting a model robust to random noise into a model which is provably robust to bounded worst-case perturbations in the l_2-metric. Another class of methods applies specifically for networks with ReLU activations, leveraging the property that the model is locally linear, and that certifying in region defined by linear constraints reduces to solving a series of linear programs, for which standard solvers can be applied [WK18].

19.8.5 Why models have adversarial examples

$$\inf_{\theta} \sup_{Q \in \mathcal{U}_\varepsilon(P_0)} \boxed{\mathbb{E}_{Z\sim Q}\, \ell(Z,\theta)}$$



$Z = (X,Y)$

coffee

$$X = H + \gamma$$

h.att

$$Y = \boxed{\underbrace{\beta^*}_{\text{causal eff}} X + H + \varepsilon}$$

(H, X, Y)

(X, Y) identif.

$\mathcal{U} = \{Q : \tilde{X} = X + \nu\}$

$\mathbb{E}_Q[(y - x^\top\beta)^2] = \nu + \varepsilon$

$$\boxed{E[(y - x^\top\beta_{LS})^2] \leq E[(y - x^\top\beta^*)^2]}$$

$$\underbrace{(\beta^* - \beta)^\top \mathbb{E}[(X+\nu) + (X+\nu)^\top](\beta^* - \beta)}_{\text{Shift penalty}} + \underbrace{\mathbb{E}[H^2] + \sigma_\epsilon^2}_{\text{Minimum risk}}$$

$$+ \underbrace{2(\beta^* - \beta)^\top \mathbb{E}[(X+\nu)H]}_{\text{Confounding bias term}}$$

$\|\beta^* - \beta^*$

$\nu = \kappa(\beta - \beta^*) \sim \kappa^2\|\beta^* - \beta^*\|^4 \longrightarrow$

white box ∞

ROBUSTNESS ← INVARIANT → CAUSALITY

Questions?

β^*

β^* $\longrightarrow$ β

$EY^1 - EY^0$

identi

$\overset{\downarrow}{=} E_P E_P Y | A=1, W$

Thank you

$Y = f_Y(1, W, \varepsilon)$ β^*_1

Questions?

Thank you

Questions?

Next Session

Gaussian Processes & Structured Prediction (Ch.18)

Feb 18, 2026

Simón Rodríguez Santana