Looking for the Missing Signal

LÉON BOTTOU
FACEBOOK AI RESEARCH

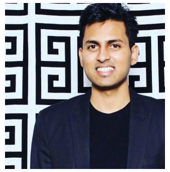
In Search for Lost Signal

LÉON BOTTOU
FACEBOOK AI RESEARCH

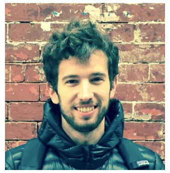
People

Martin Arjovsky
NYU

Léon Bottou
Facebook AI Research



Soumith Chintala
Facebook AI Research



David Lopez-Paz
Facebook AI Research

Robert Nishihara
Berkeley

Maxime Oquab
Inria

Alex Peysakhovich
Facebook AI Research

Bernhard Schölkopf
MPI Tübingen

Motivation

Machine learning success stories

- Recognizing objects in images

  after training on more images than a human can see.

- Translating natural languages (somehow)

  after training on more text than a human can read.

- Playing Atari games

  after playing more games than any teenager can endure.

- Playing Go (famously)

  after playing more grandmaster level games than mankind.

What are we doing wrong?

Are our learning algorithms so inefficient?

- Hard to say for the most complex learning systems.
- For simpler systems, *in the absence of a strong prior*, the Cramer-Rao bound suggests that this is not the case.

Transfer learning?

Does transfer learning give strong enough priors?

- Transfer learning works well across similar tasks.
- Transfer learning across all human experiences is hypothetical.
- Could there be something else?

Another viewpoint

Is there more signal in data than we think?

- Where to find it?

- Beyond correlations…
  - → Causation
  - → Complex moments

Typical supervised machine learning systems use

- $E[\phi(x)]$ $E[y]$ $E[y\,\phi(x)]$

What about

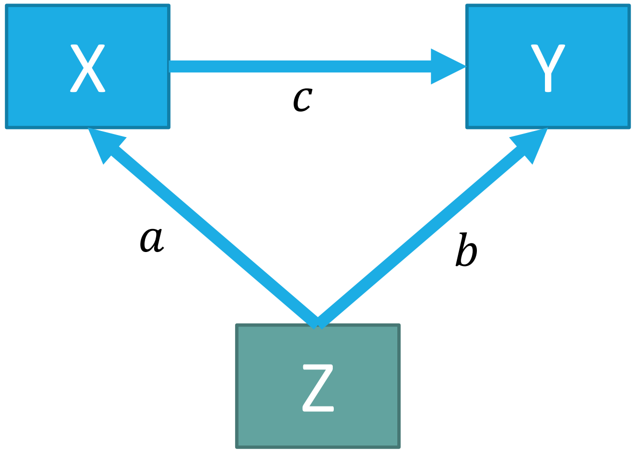
- $E[\phi(x,y)]$

Causation and Moments

Causal confounding

$X = aZ + \mathcal{U}(-s_1, s_1)$

$Y = bZ + cX + \mathcal{U}(-s_2, s_2)$

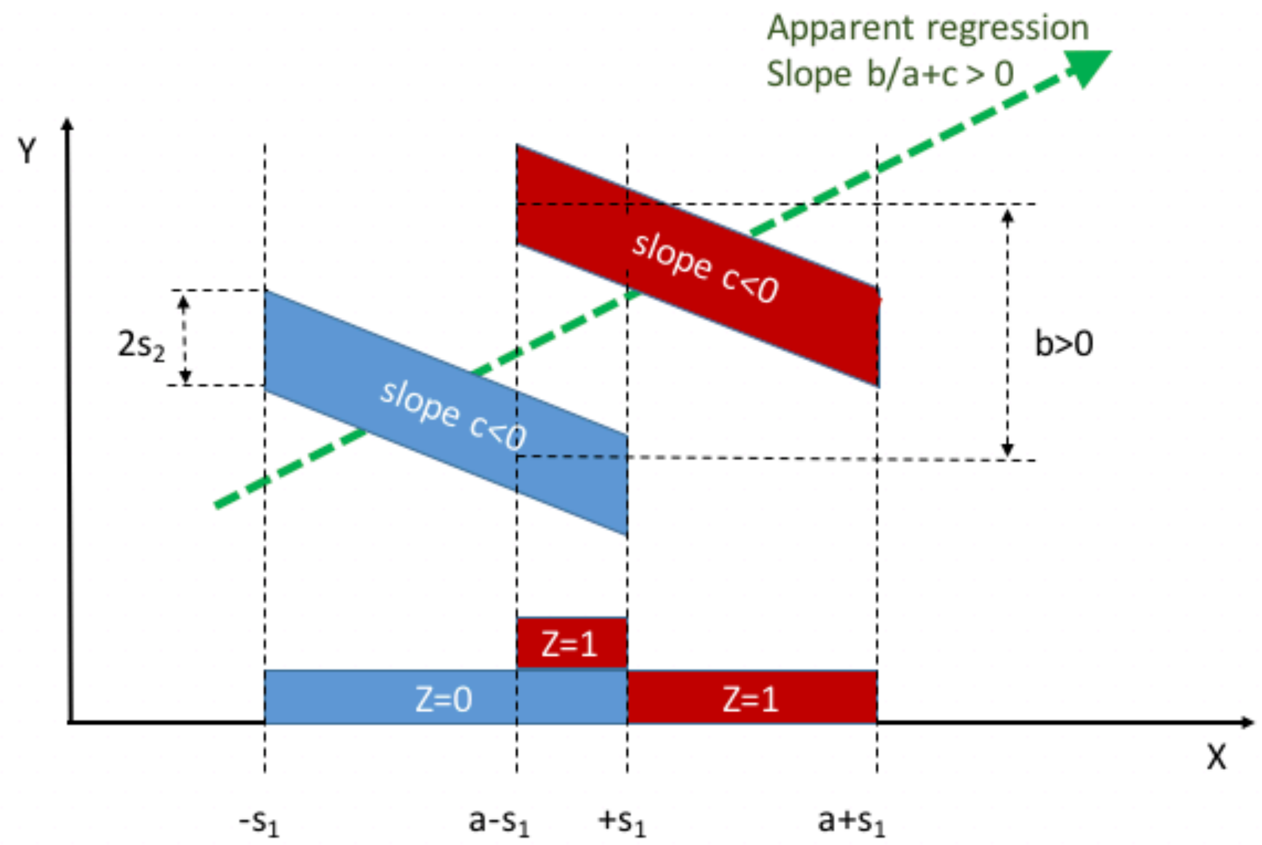


$Z \sim$ Bernoulli, $p = \frac{1}{2}$

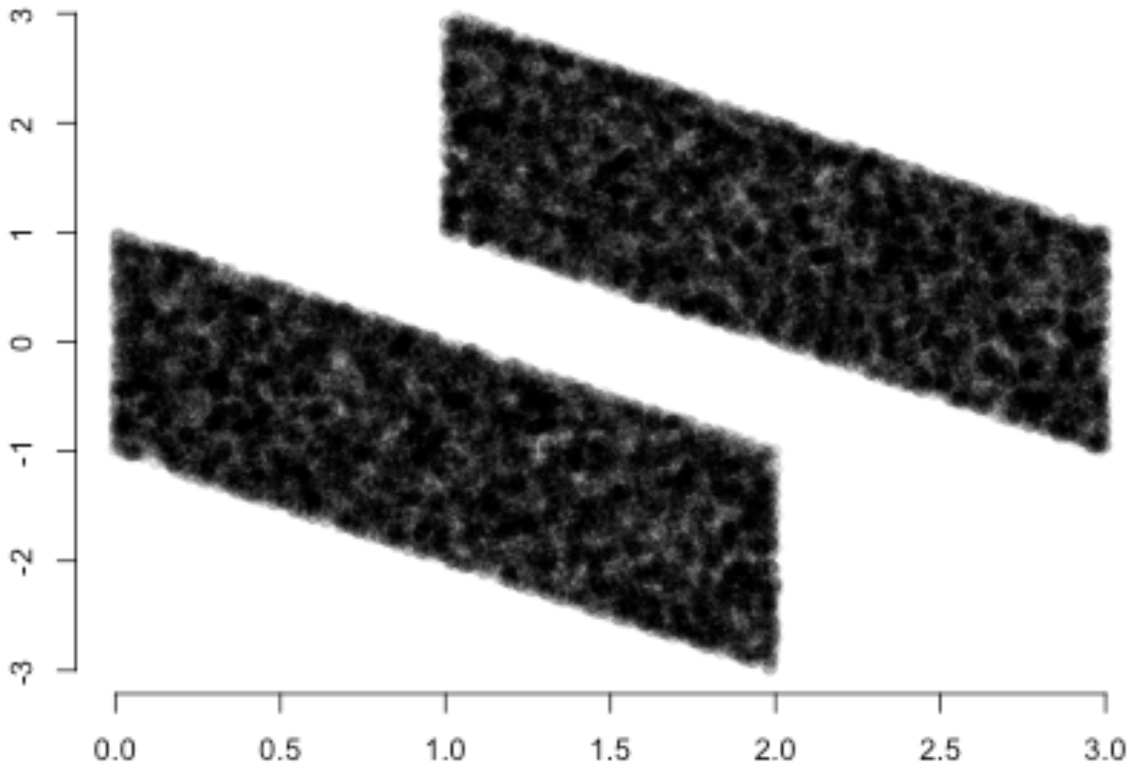
Simpson effect

- Suppose we only observe X, Y.
- If $c < 0$ and $b + ca > 0$, then $Cor(X, Y) > 0$

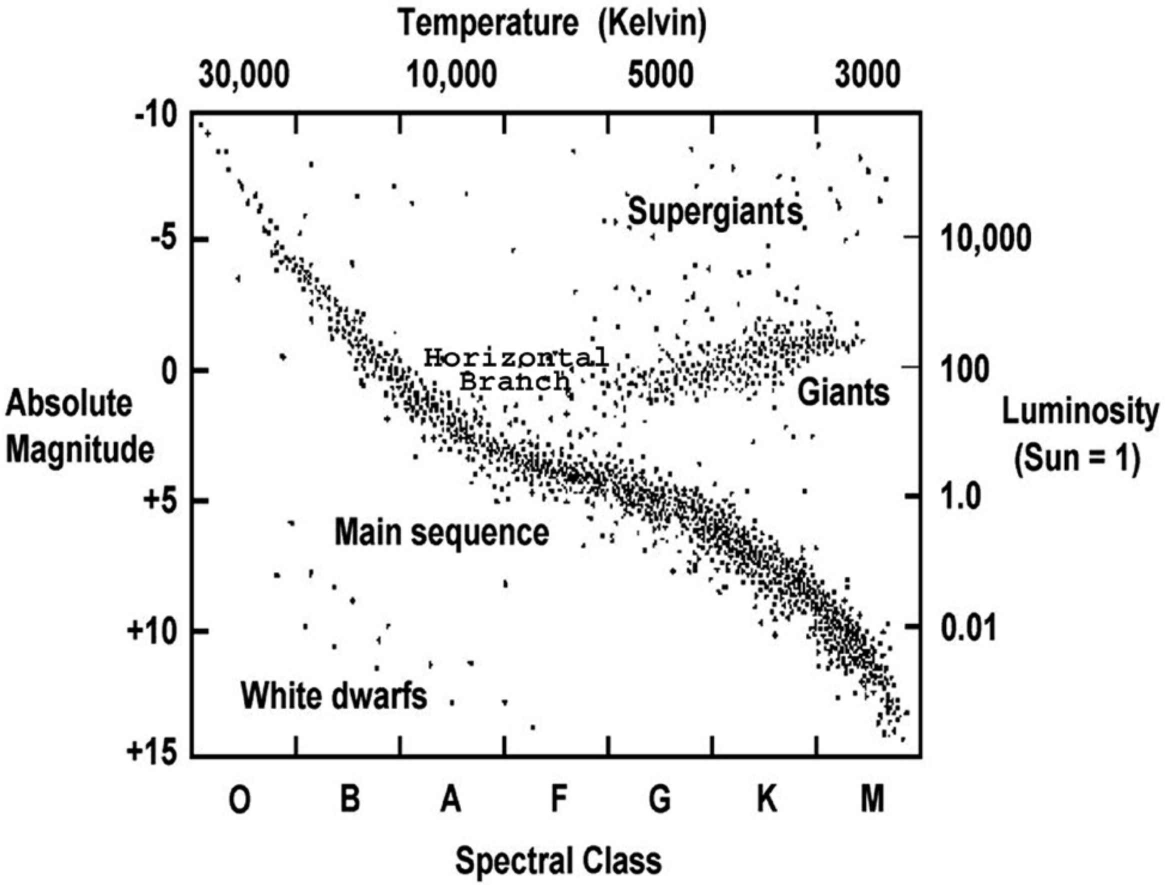
Causal confounding



Look at the scatterplot!



More scatterplots



The Hertzsprung–Russell diagram shows the relationship between the stars' absolute magnitudes or luminosities versus their stellar classifications or effective temperatures.

Scientists clearly draw causal conclusions from a scatterplot, even when interventions are impossible.

Causal problems with two variables

Given two observed variables X, Y

I. Either X causes Y,

II. or Y causes X,

III. or X and Y have unobserved common causes,

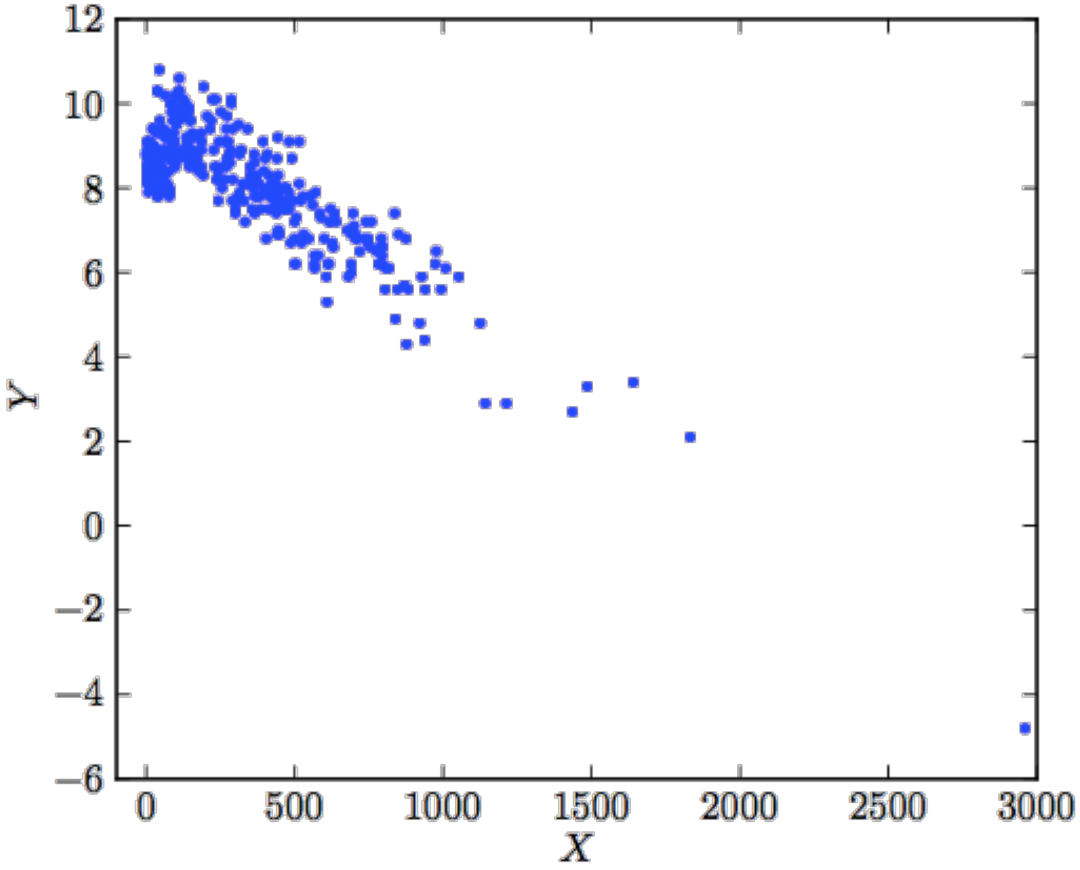
IV. or X and Y are independent.

Reichenbach

potentially confounding

Let’s focus on causal direction detection (I and II)

How does causal direction look like?



In this scatter plot

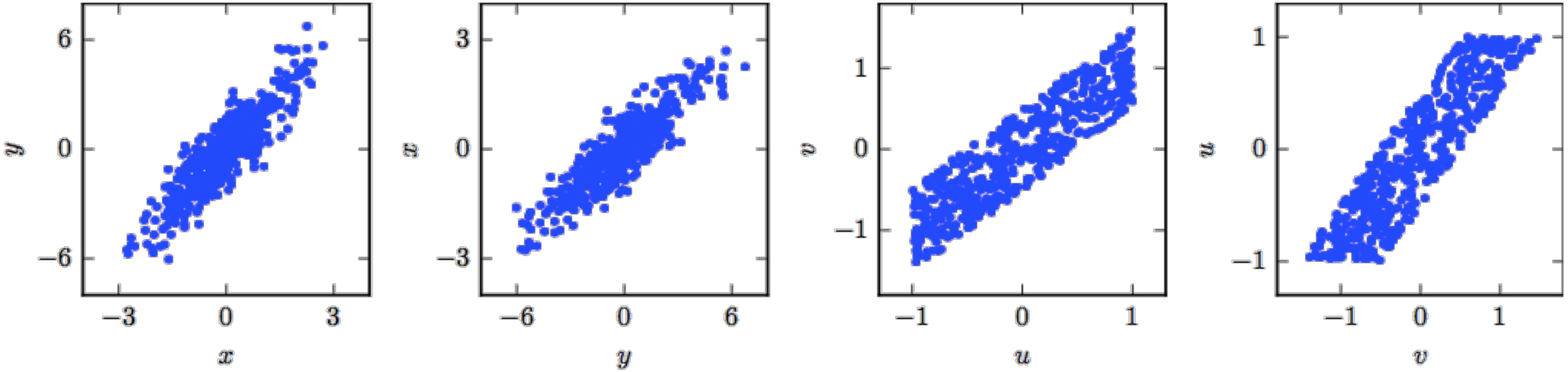
- X is altitude.
- Y is average temperature.

Does the scatter plot reveal whether

- X causes Y
- or Y causes X ?

Footprint example 1 – additive noise

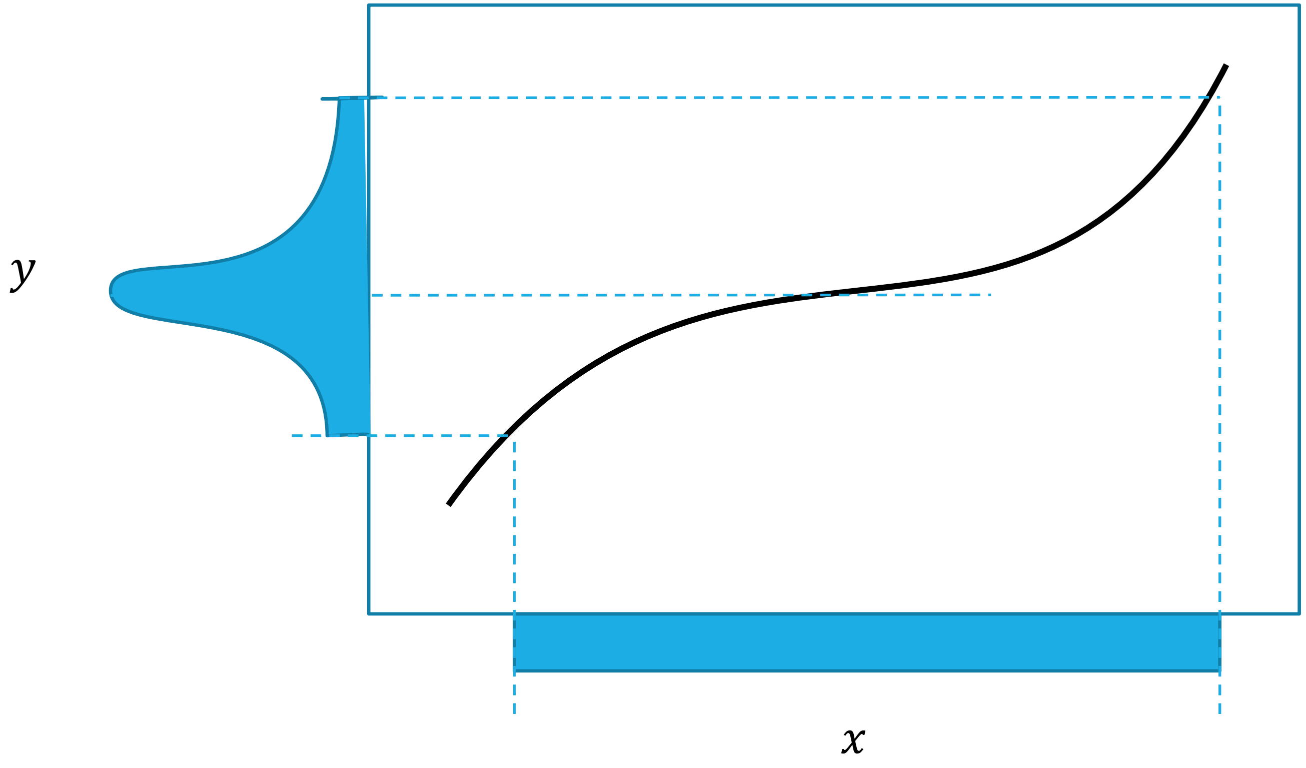
$Y = \alpha\, X + \beta + Noise$



Sometimes the high moments (the corners) reveal something.

(PETERS ET AL., 14)

Footprint example 2 -- coincidences



(Janzing et al., 2011)

From scatterplot to causation direction

Detecting causation direction at scale

- We could build a long list of causal footprint examples, then decide which example is most appropriate for a given scatterplot, etc.
- Or we can construct a classifier...

(Lopez-Paz, et al., 2015)

Featurizing a scatterplot

High moments?

- $F_{rs} = \frac{1}{m}\sum_{j=1}^{m} x_j^r y_j^s$ for well chosen r and s.

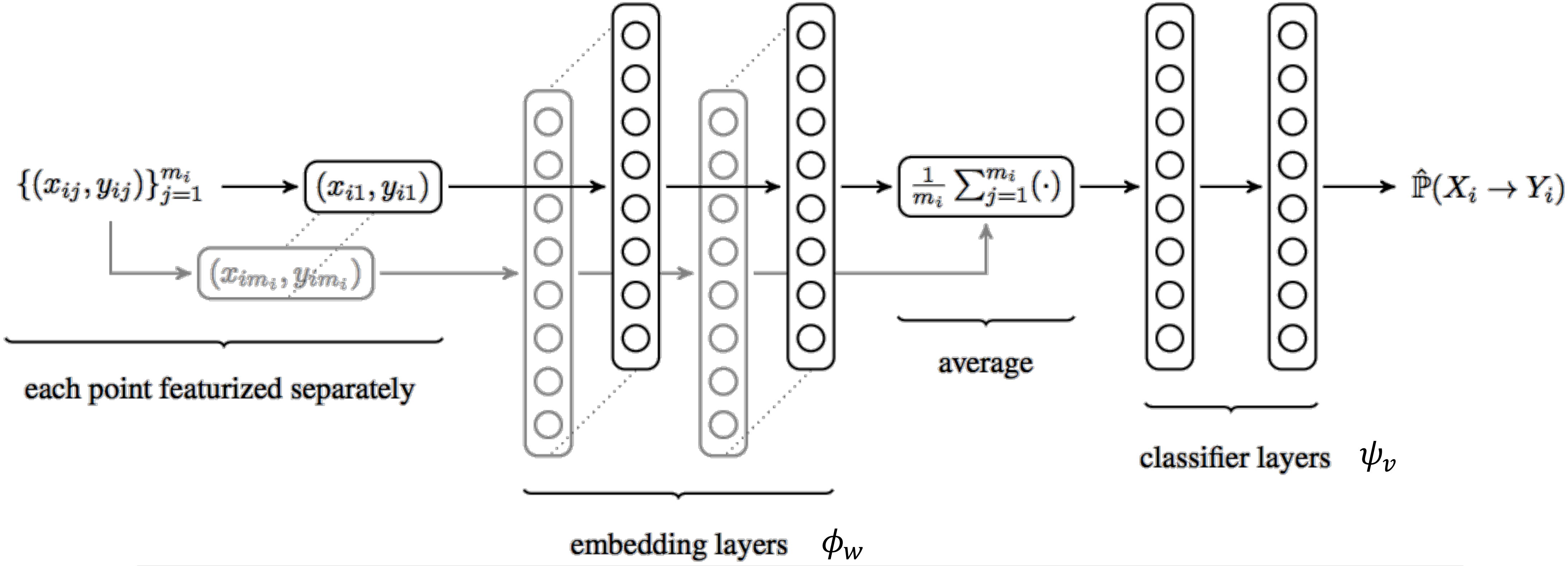
Reproducing Kernel Hilbert space?

- $F = \frac{1}{m}\sum_{j=1}^{m} \phi(x_j,\ y_j) \in \mathcal{H}_K$ with $\langle \phi(.), \phi(.) \rangle_K = K(.,.)$

Learning the features and the classifier

- $F_w = \frac{1}{m}\sum_{j=1}^{m} \phi_w(x_j,\ y_j)$

Neural Causation Classifier


$\{(x_{ij}, y_{ij})\}_{j=1}^{m_i}$
(x_{i1}, y_{i1})
(x_{im_i}, y_{im_i})
$\frac{1}{m_i}\sum_{j=1}^{m_i}(\cdot)$
$\hat{\mathbb{P}}(X_i \to Y_i)$
each point featurized separately
embedding layers ϕ_w
average
classifier layers ψ_v

Training NCC

We do not have access to large causal direction datasets

But we can generate artificial scatterplots.

$$Y = f(X) + v(X)\varepsilon$$

Step 1 - draw distribution on X

- Draw $k \sim \mathcal{U}\{1,2,3,4,5\}$ $\mathrm{r}, s \sim \mathcal{U}[0,5]$
- Take a mixture of k Gaussians with $\mu \sim \mathcal{N}(0, r)$ and $\sigma \sim \mathcal{N}(0, s)$

Training NCC

Step 2 - draw mechanism f

- Cubic spline with random number of random knots…

Step 3 - draw noise

- Noise ε is Gaussian with random variance $\sim\mathcal{U}[0,5]$
- Function $v(X)$ is another cubic spline with random knots.

Step 4 – generate causal scatter plot $X \rightarrow Y$

- Draw x_j, ε_j then compute $y_j = f(x_j) + v(x_j)\varepsilon_j$
- Rescale x_j, y_j to enforce marginal mean 0 and sdev 1

Training NCC

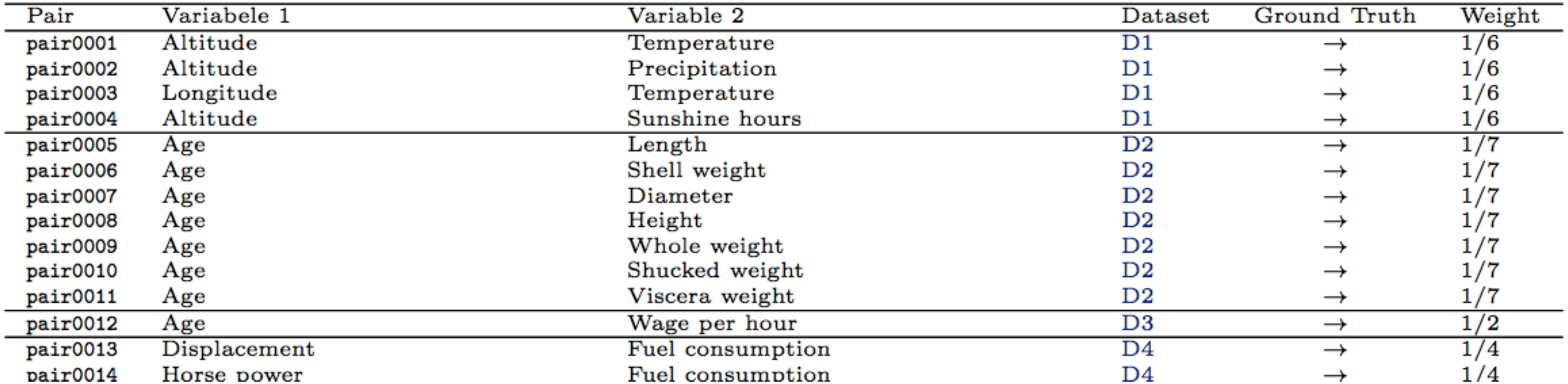
Step 6 – Generate training examples

- Scatterplot $\{(x_j, y_j)\}$ is associated with target label 1
- Scatterplot $\{(y_j, x_j)\}$ is associated with target label 0

Repeat 100000 to generate a training set.
Train the neural network classifier with the usual bag of tricks.
(dropout regularization, rmsprop, cross-validation, etc.)

Sanity check

- After training on artificial data, NCC achieves state-of-the-art [79%] performance on the *Tübingen cause-effect dataset”,* which contains 100 cause-effect pairs (https://webdav.tuebingen.mpg.de/cause-effect)



Pair	Variabele 1	Variable 2	Dataset	Ground Truth	Weight
pair0001	Altitude	Temperature	D1	→	1/6
pair0002	Altitude	Precipitation	D1	→	1/6
pair0003	Longitude	Temperature	D1	→	1/6
pair0004	Altitude	Sunshine hours	D1	→	1/6
pair0005	Age	Length	D2	→	1/7
pair0006	Age	Shell weight	D2	→	1/7
pair0007	Age	Diameter	D2	→	1/7
pair0008	Age	Height	D2	→	1/7
pair0009	Age	Whole weight	D2	→	1/7
pair0010	Age	Shucked weight	D2	→	1/7
pair0011	Age	Viscera weight	D2	→	1/7
pair0012	Age	Wage per hour	D3	→	1/2
pair0013	Displacement	Fuel consumption	D4	→	1/4
pair0014	Horse power	Fuel consumption	D4	→	1/4

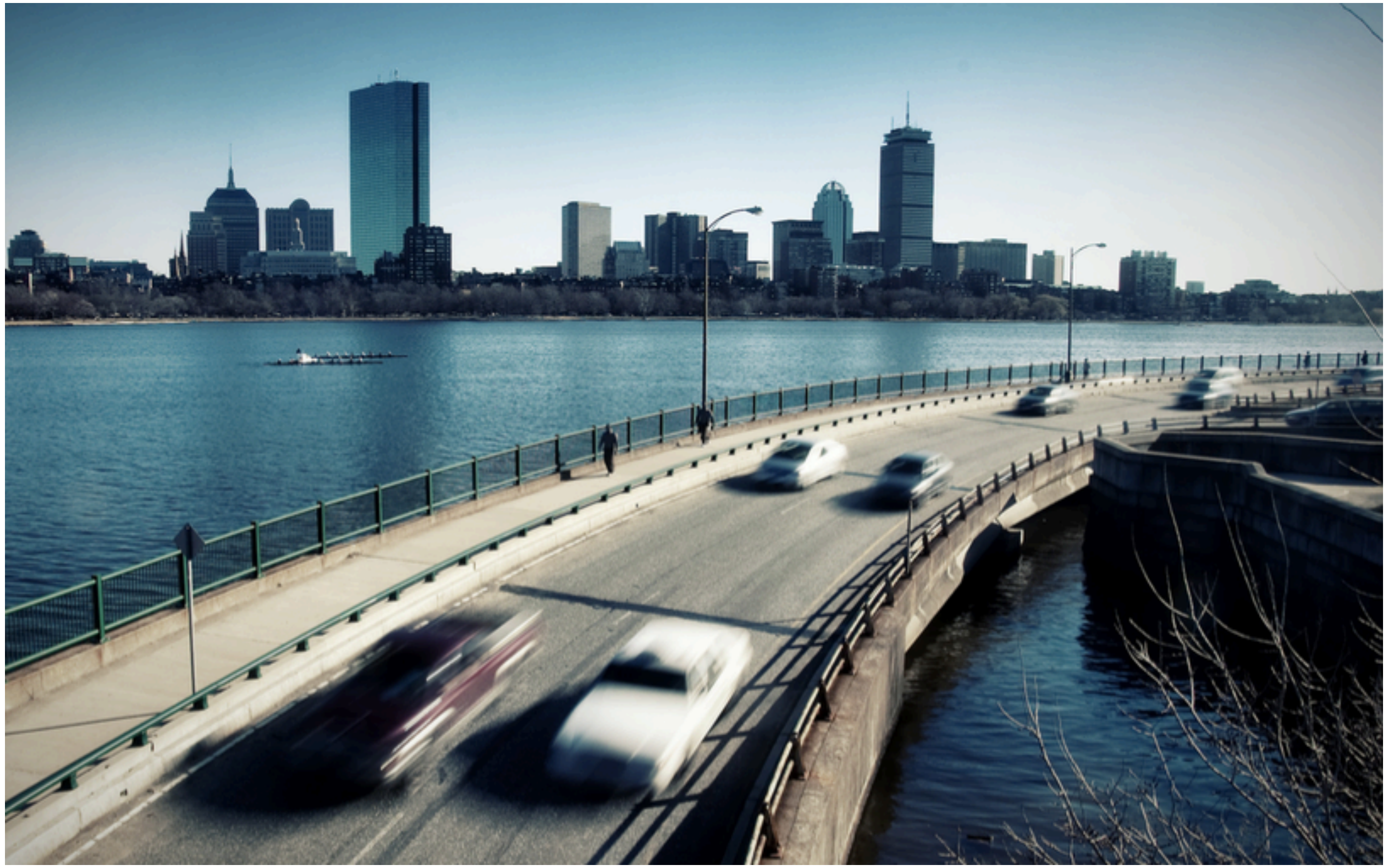
Remarks

- This works also for detecting confounding variables. How to validate that?
- Two-dimensional scatterplots are limited…

Finding
a causal signal
in static images

(LOPEZ-PAZ, NISHIHARA, CHINTALA, SCHOELKOPF, BOTTOU, TO APPEAT IN CVPR17)

Counterfactual on images



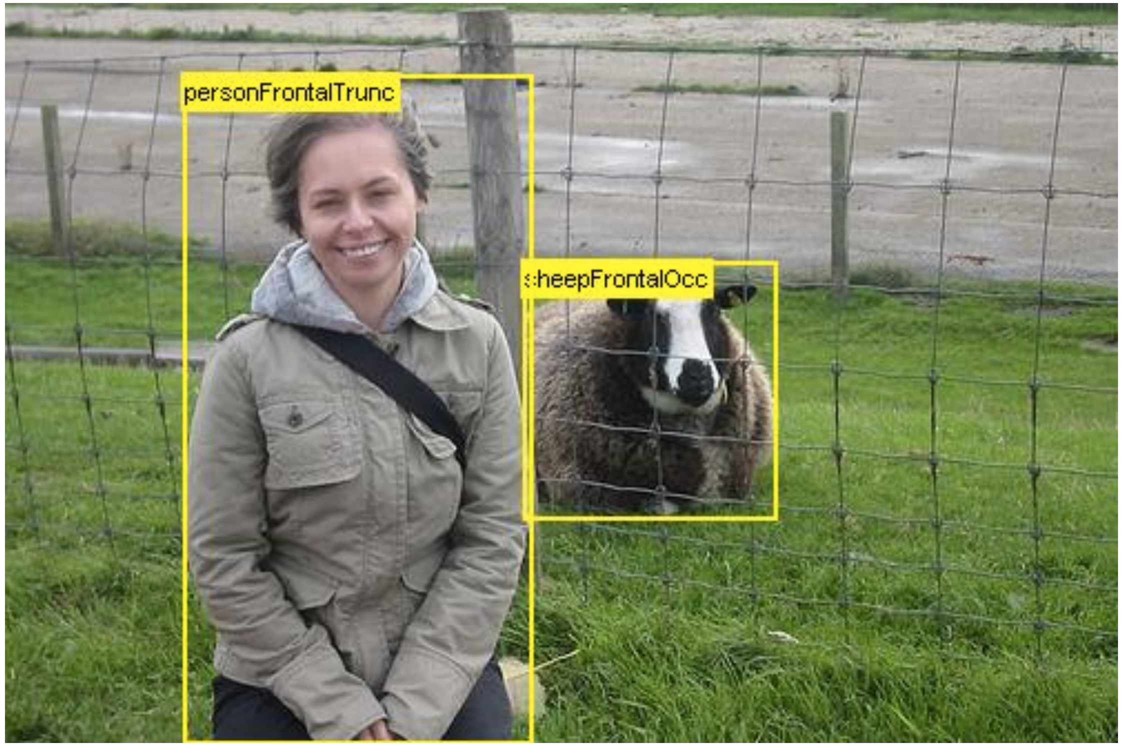
Asymmetric relation

- How would this image would have looked like if one had removed the cars?
- How would this image would have looked like if one had removed the bridge?

Can we use image datasets to identify the *causal dispositions* of object categories?

How to validate a result?

Image datasets



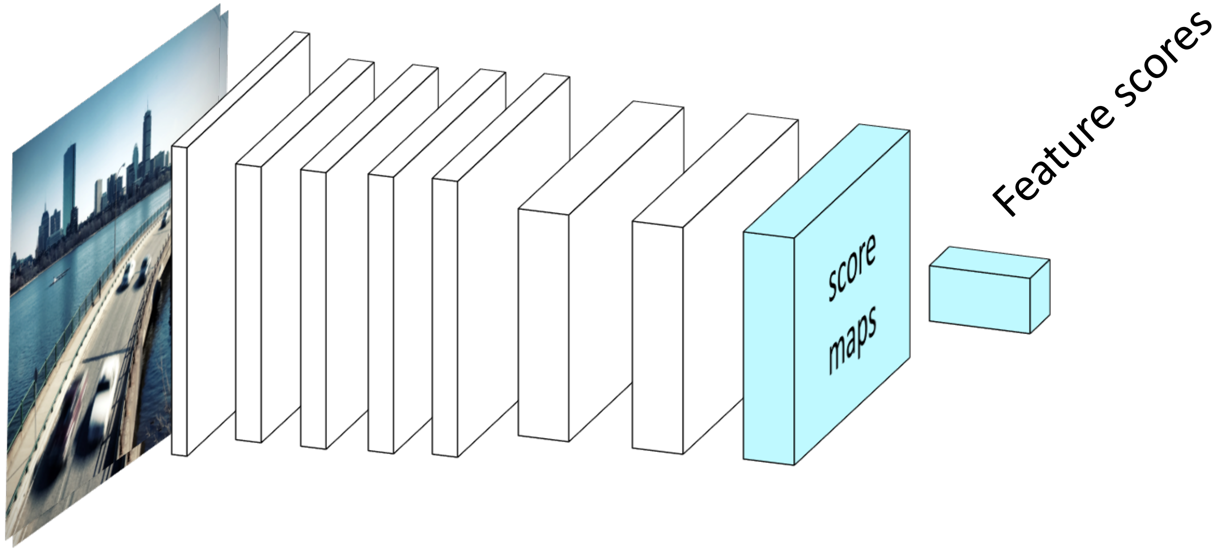
Images labeled with

- Object of interests (cat, dog, …)
- Bounding boxes.

The PASCAL VOC dataset contains 20 categories, 11541 images

The COCO dataset is much larger. After restricting to the same categories than PASCAL VOC, we have 99309 images.

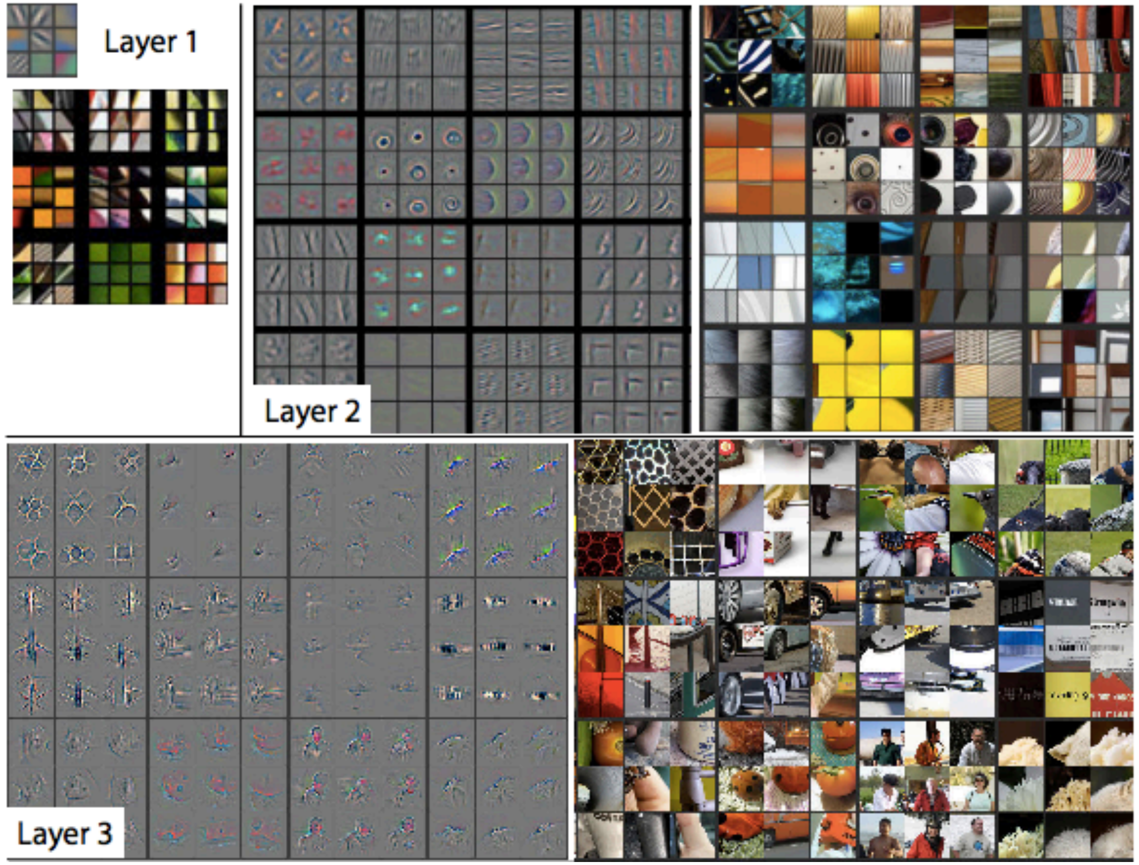
Featurizing the images



All images are preprocessed using a state-of-the-art pretrained CNN.

Each image is then represented by a vector of 512 features.

Features scores are often interpretable



Features scores
are often interpretable
as features of the scene.

(Zeiler & Fergus, 2013)

Causal and anti-causal features

For each object category, we can also define two sets of features

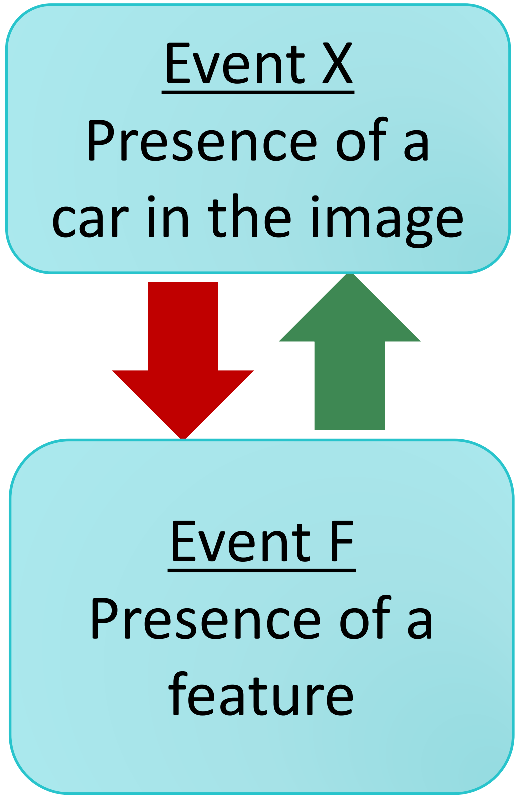
- The causal features are those that cause the presence of the object of interest. *If the object of interest had not been present in the image, these feature would still have appeared*.
- The anticausal features are those that are caused by the presence of the object of interest. *If the object of interest had not been present in the image, these feature would not have appeared*.

Causal and anti-causal features

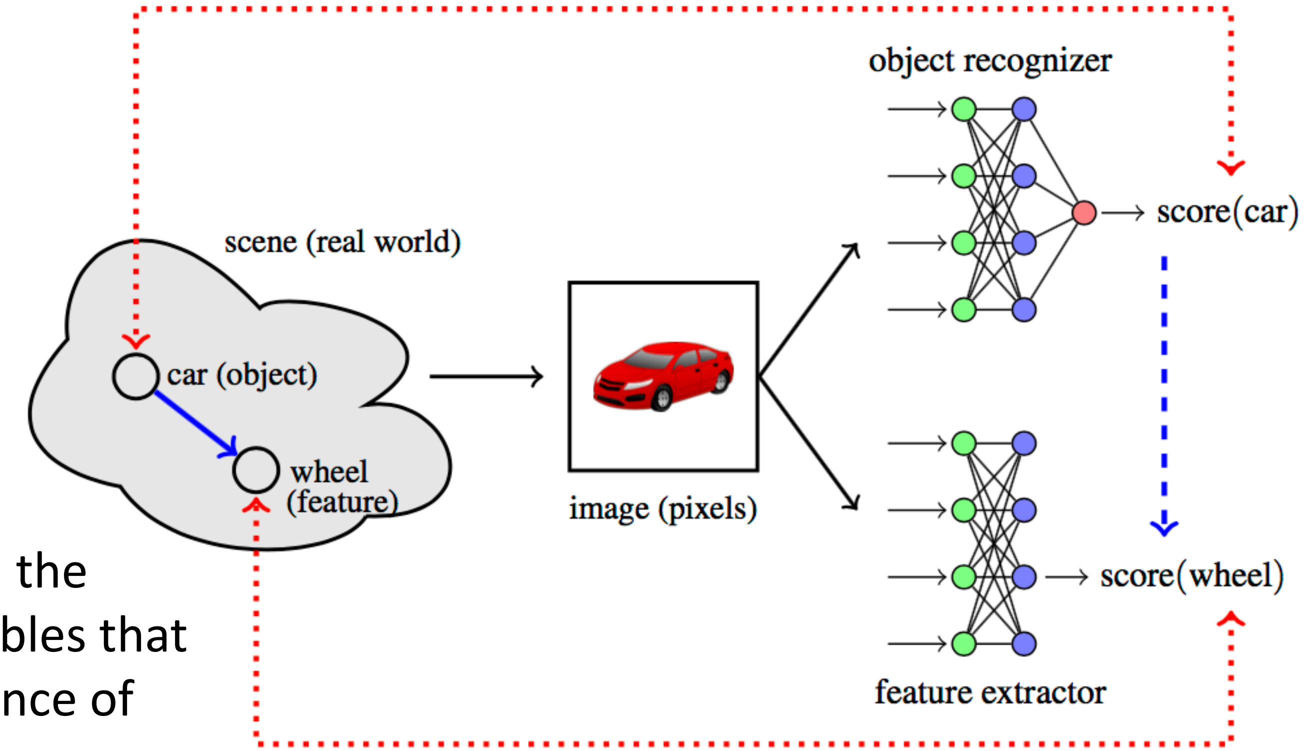
If X and F are positively correlated,
a trained classifier may rely on SCORE(F).

This correlation may occur because

- X → F (anticausal feature)
  Example : F = presence of wheels.
- F → X (causal feature)
  Example: F = presence of road.
- F ← C → X (something else)
  Example: F = bike, C=street



Proxy variables



We hope to see a similar footprint between the scores computed by a well tuned classifier.

Assume there is
a causal footprint in the distribution of variables that represent the presence of an object or a feature

Empirically identifying causal and anti-causal features

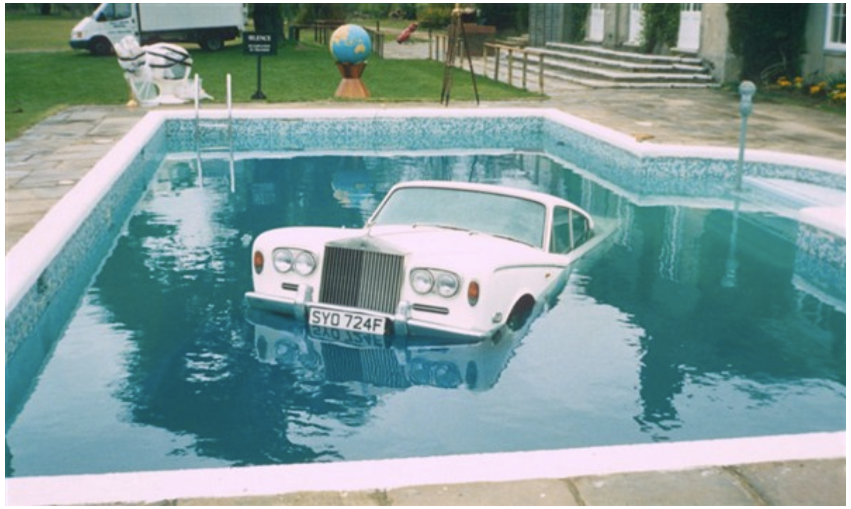
- We apply NCC to the feature scores and object scores to identify the top 1% causal and anticausal features for each of the twenty categories.

- *NCC was trained using artificial data only (not image specific)*
- *The same NCC classifier is used for all categories.*

Computer vision ≠ Statistics

Context features vs Object features

Car examples in ImageNet



Is this less of a car
because the context is wrong?

Object features and context features

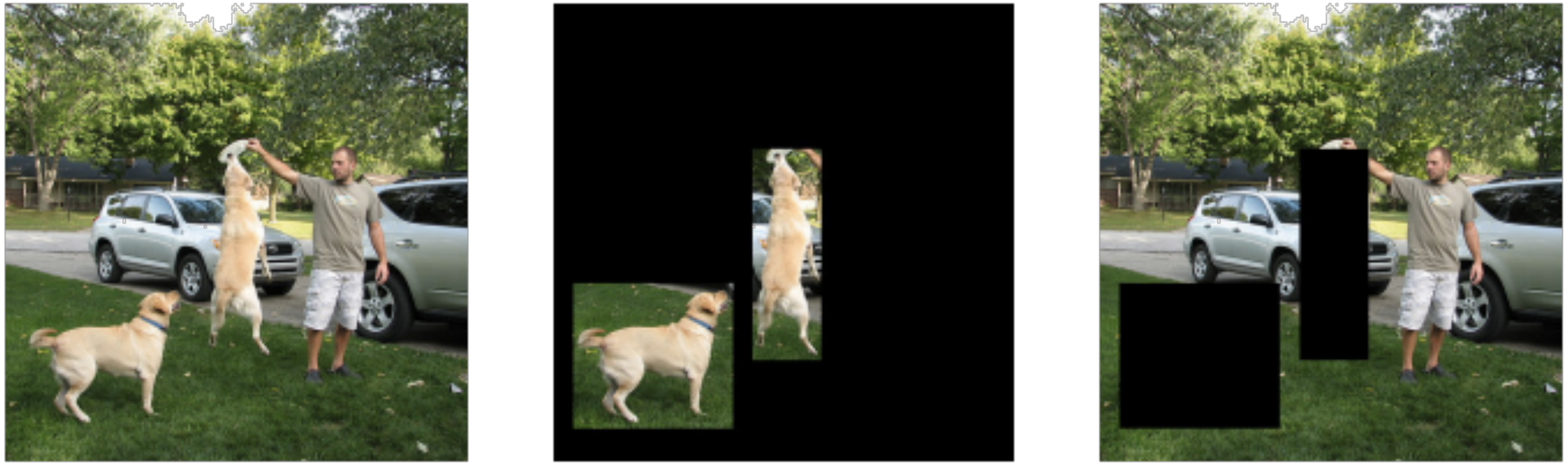
In computer vision, one is often interested in another distinction

- The object features "belong" to the object and are most often activated inside the object bounding box.
  Example: car wheels, person eyes, etc.
- The context features are those most often activated outside the bounding box.
  Example: road under a car, car shadow

Background story
"bags of visual words"

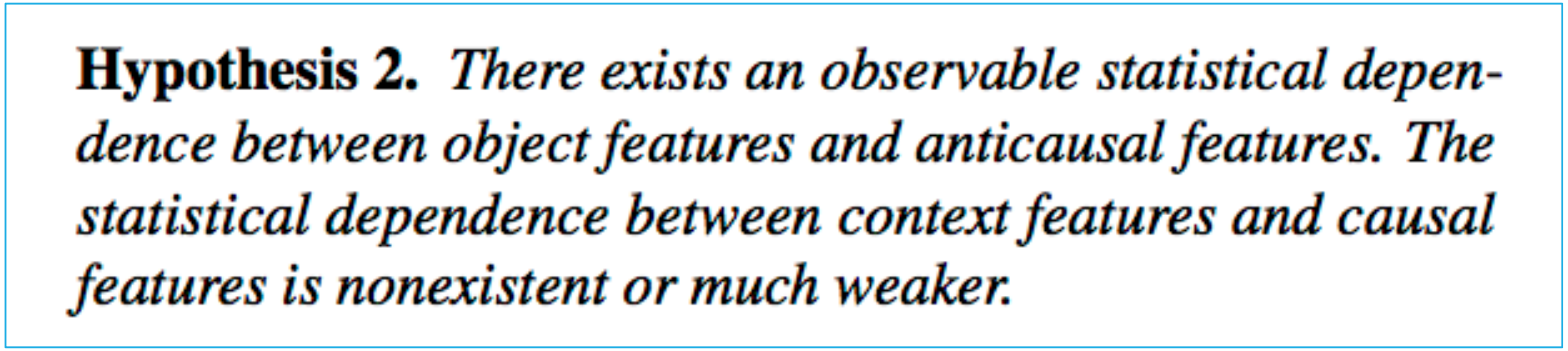
Empirically identifying object and context-features

Since we know the bounding boxes, we can observe how the feature values change when we black out image parts. Averaging and normalizing these variations gives us the object-feature ratio and context-feature ratio.



(a) Original image x_j (b) Object image x_j^o (c) Context image x_j^c

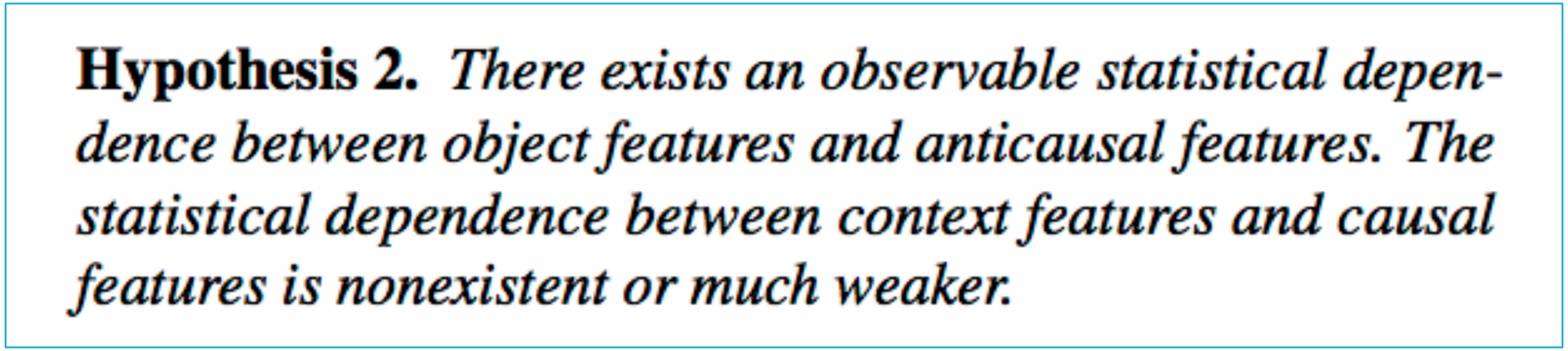
Hypotheses



Hypothesis 2. *There exists an observable statistical dependence between object features and anticausal features. The statistical dependence between context features and causal features is nonexistent or much weaker.*

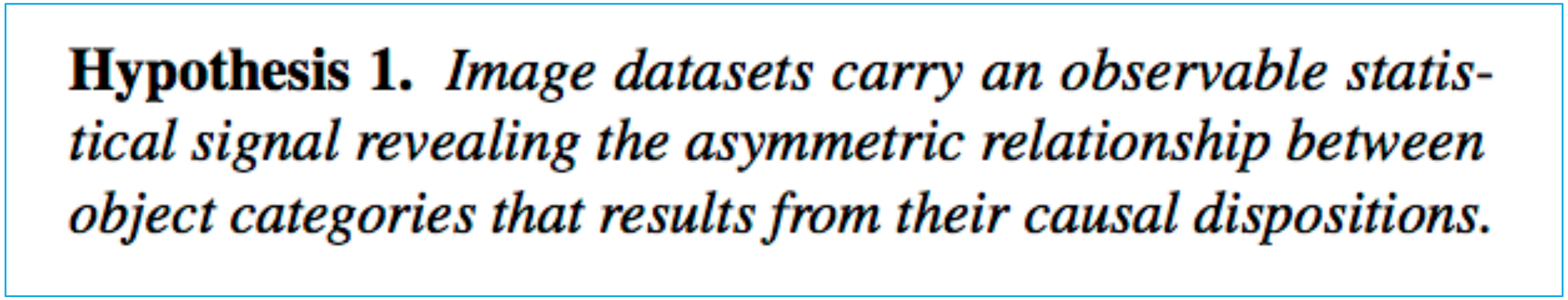
We expect this because anticausal features should often be features of subparts of the object, likely to be contained in the bounding box. Context features may cause or be caused by the presence of an object (e.g., the shadow of a car).

Hypotheses



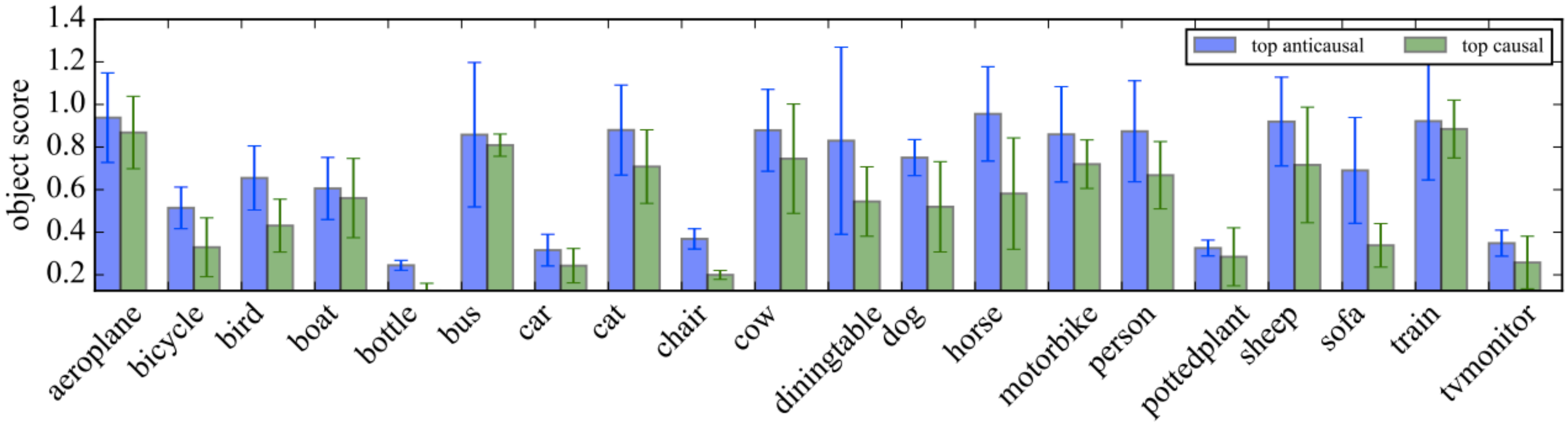
Hypothesis 2. *There exists an observable statistical dependence between object features and anticausal features. The statistical dependence between context features and causal features is nonexistent or much weaker.*

$\Rightarrow$



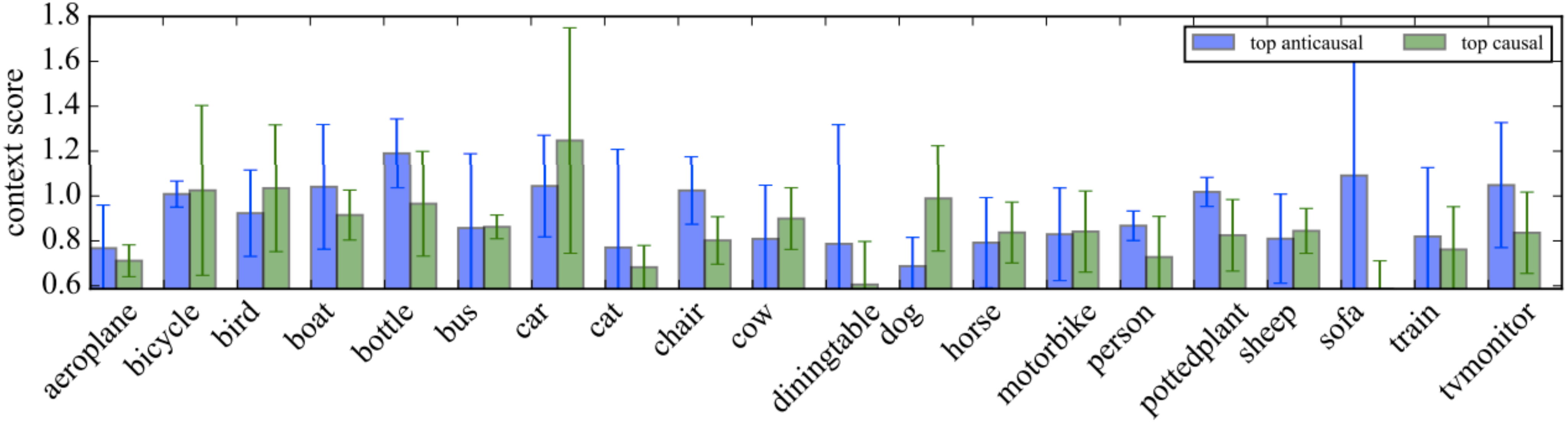
Hypothesis 1. *Image datasets carry an observable statistical signal revealing the asymmetric relationship between object categories that results from their causal dispositions.*

Results



- Top anticausal features have higher object scores for all twenty categories.
- The probability that this happens out of chance is $2^{-20} \approx 10^{-6}$.

Results



No clear relation between top causal features and context scores.

More information

- The effect disappears completely if we replace NCC by the correlation coefficient (or its absolute value) between the feature and the category.
- The effect appears to be robust to many details of the experiment such as the precise composition of the NCC data, the precise computation of object/context scores, the methods we use to determine a continuous proxy for the categories, etc.

Causal signal in images

- We have indirectly shown that high order statistics in image datasets can inform us about causation in the scenes. To our knowledge, no prior work has established or even considered the existence of such a signal.
- We don’t know how to use it.
- Our detection method is cumbersome.

But there is signal.

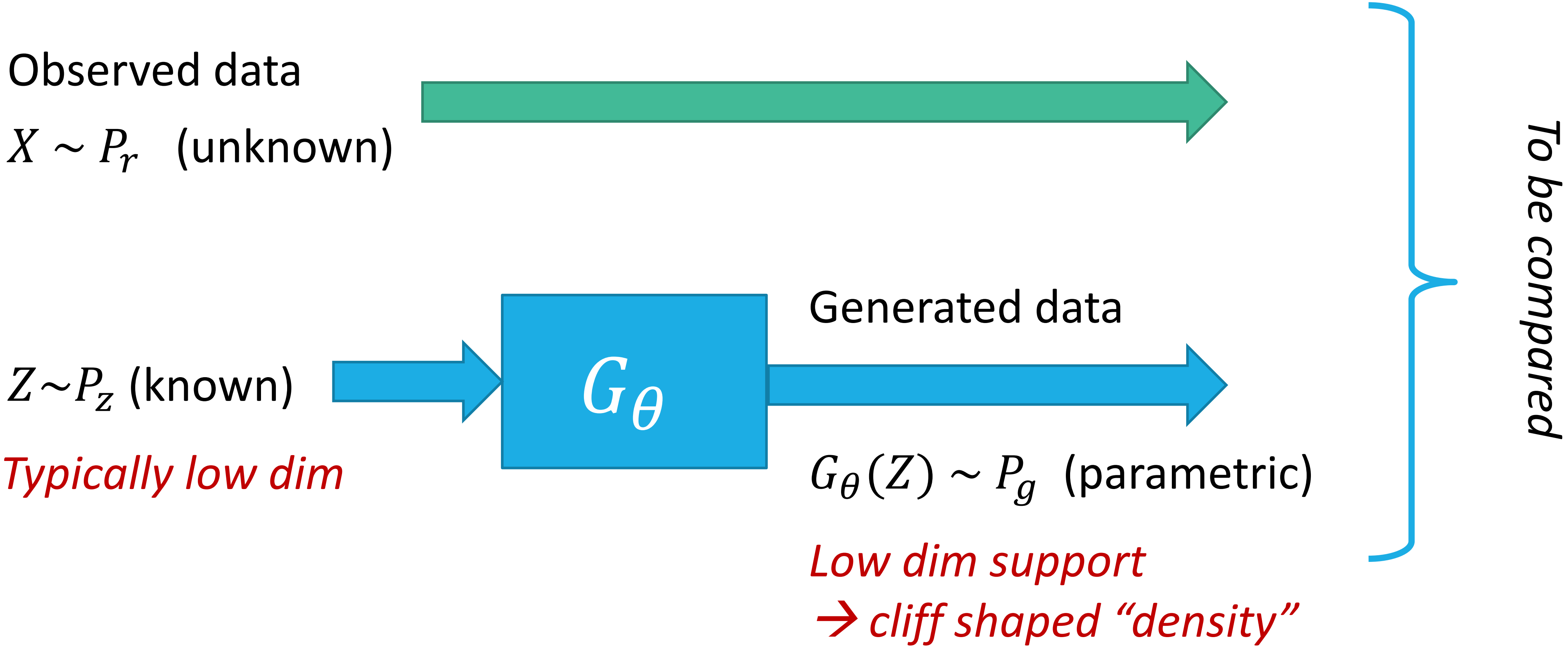
On the uses of a Wasserstein(ish) distance

(ARJOVSKY, BOTTOU, ICLR 2017)
(ARJOVSKY, CHINTALA, BOTTOU, SUBMITTED).

The “mythical” unsupervised learning

- This is not about using unlabeled data to discover probability ratios.
- This is about using unlabeled data to discover the (causal) generating mechanism.
- Causal footprints
  - → corners, cliffs, shocks, …
  - → low dimensional causal models

The generator approach (VAE, GAN, …)



Observed data

$X \sim P_r$ (unknown)

$Z \sim P_z$ (known)

Typically low dim

G_θ

Generated data

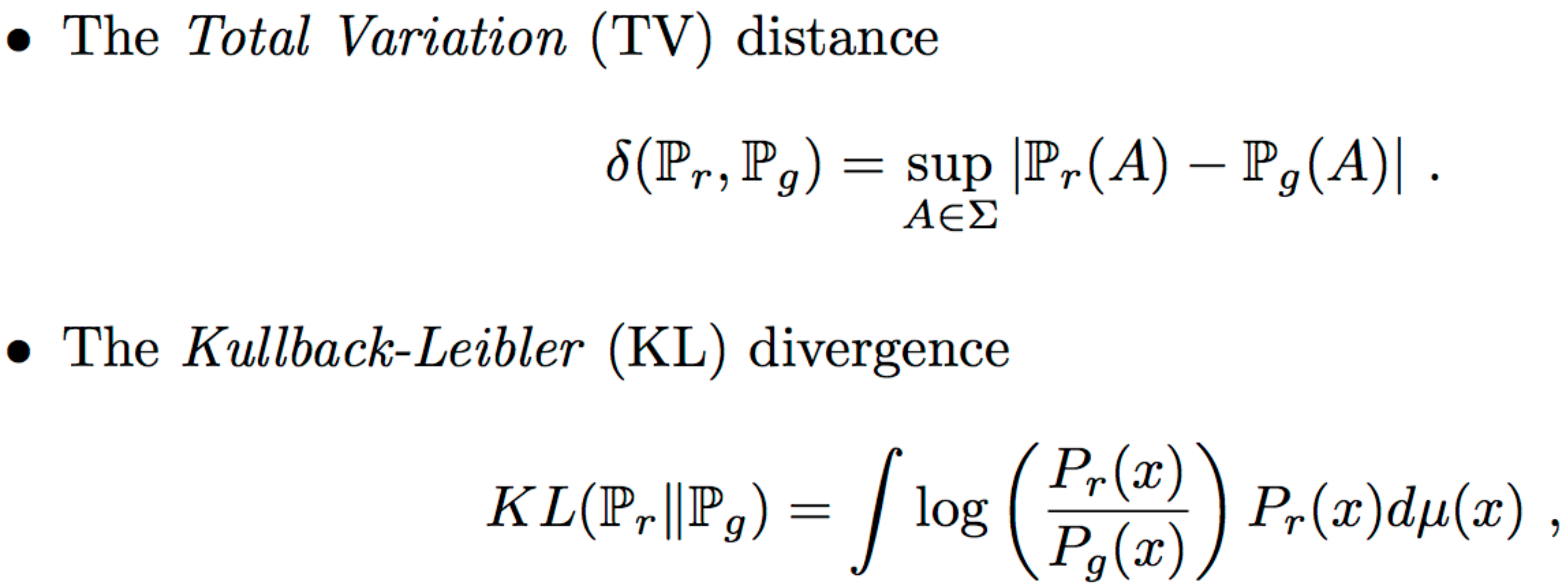
$G_\theta(Z) \sim P_g$ (parametric)

Low dim support

→ cliff shaped "density"

To be compared

Comparing distributions



- The *Total Variation* (TV) distance

$$\delta(\mathbb{P}_r, \mathbb{P}_g) = \sup_{A \in \Sigma} |\mathbb{P}_r(A) - \mathbb{P}_g(A)| \ .$$

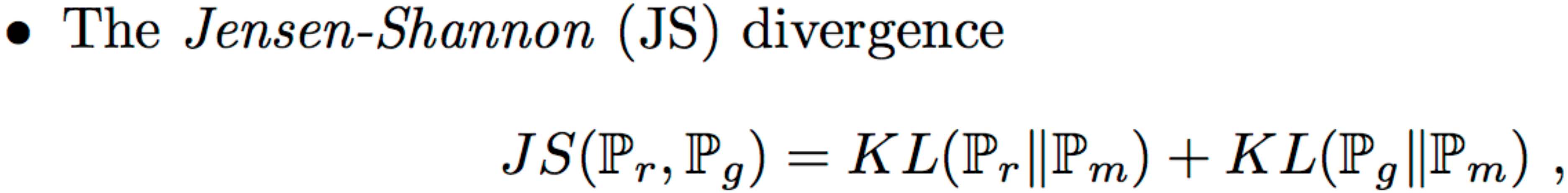
- The *Kullback-Leibler* (KL) divergence

$$KL(\mathbb{P}_r \| \mathbb{P}_g) = \int \log\left(\frac{P_r(x)}{P_g(x)}\right) P_r(x) d\mu(x) \ ,$$

VAE

requires densities, asymmetric, possibly infinite

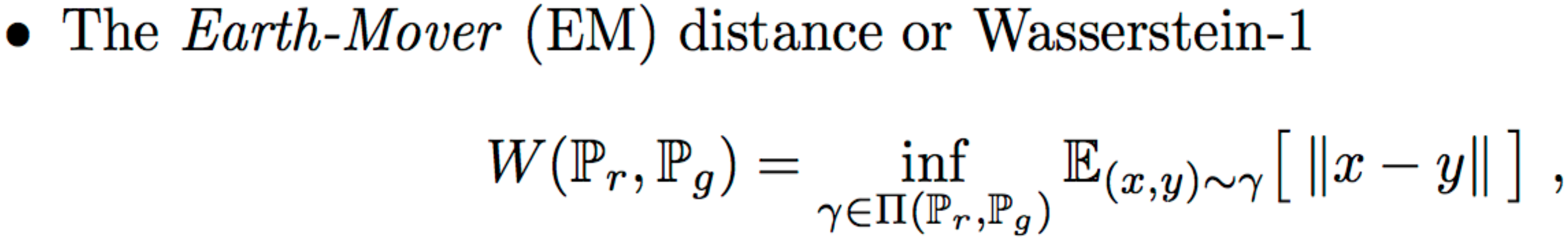
Comparing distributions



- The *Jensen-Shannon* (JS) divergence

$$JS(\mathbb{P}_r, \mathbb{P}_g) = KL(\mathbb{P}_r \| \mathbb{P}_m) + KL(\mathbb{P}_g \| \mathbb{P}_m) \ ,$$

symmetric, does not require densities, $0 \leq JS \leq \log(2)$



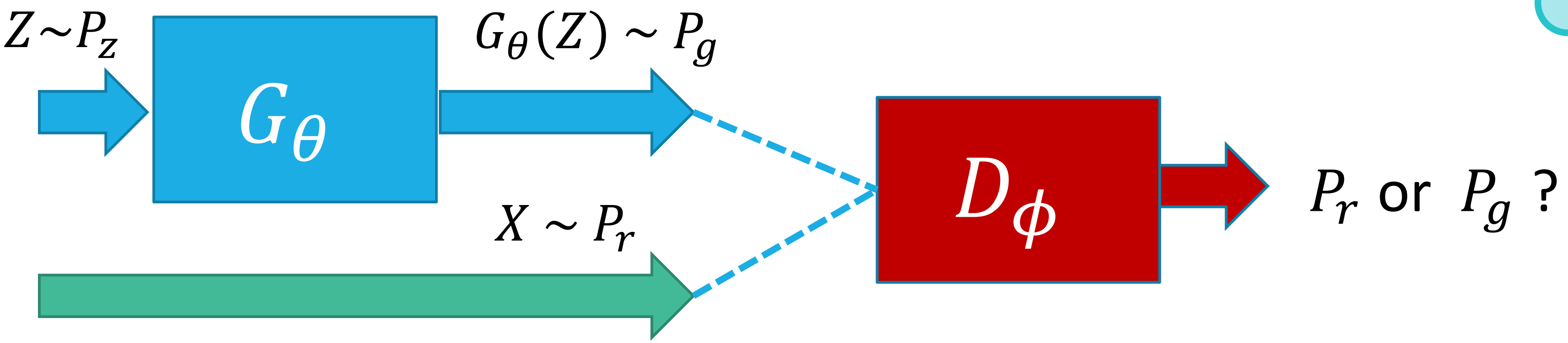
- The *Earth-Mover* (EM) distance or Wasserstein-1

$$W(\mathbb{P}_r, \mathbb{P}_g) = \inf_{\gamma \in \Pi(\mathbb{P}_r, \mathbb{P}_g)} \mathbb{E}_{(x,y)\sim\gamma} \big[\|x - y\| \big] \ ,$$

always defined, involves metric on underlying space.

Generative adversarial network

GAN



Discriminator maximizes and generator minimizes

$$L(\phi, \theta) = \mathbb{E}_{x \sim \mathbb{P}_r}[\log D_\phi(x)] + \mathbb{E}_{z \sim p_Z}[\log(1 - D_\phi(g_\theta(z)))]$$

Generative adversarial network

Discriminator maximizes and generator minimizes

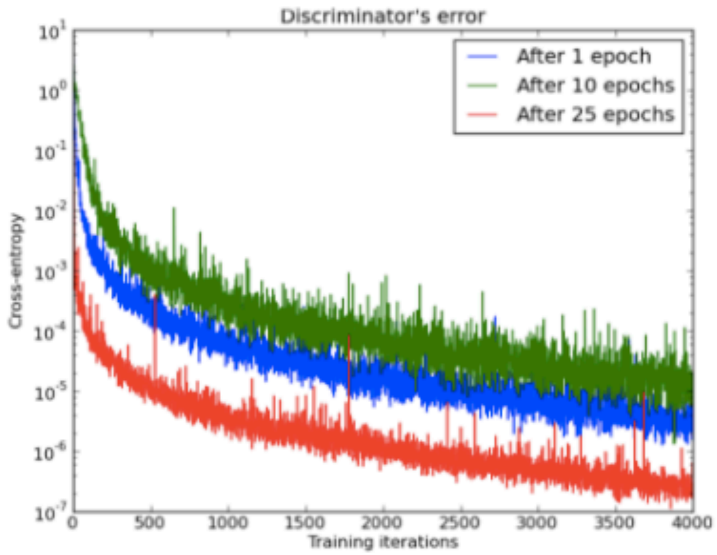
$$L(\phi, \theta) = \mathbb{E}_{x\sim\mathbb{P}_r}[\log D_\phi(x)] + \mathbb{E}_{z\sim p_Z}[\log(1 - D_\phi(g_\theta(z)))]$$

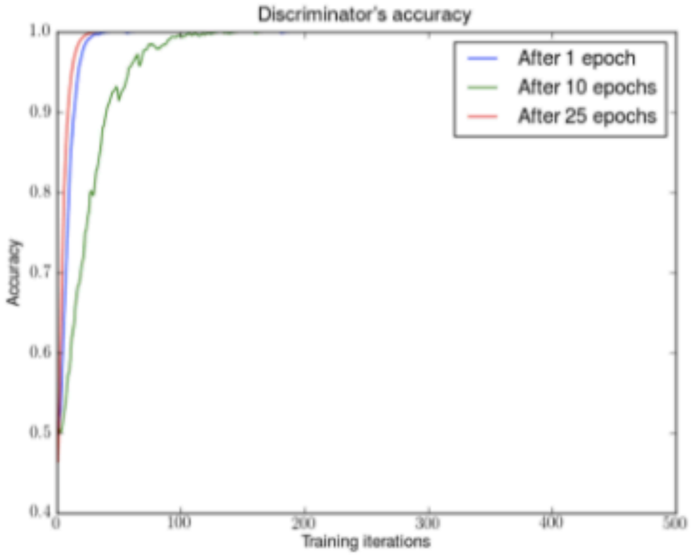
Nasty saddle point problem

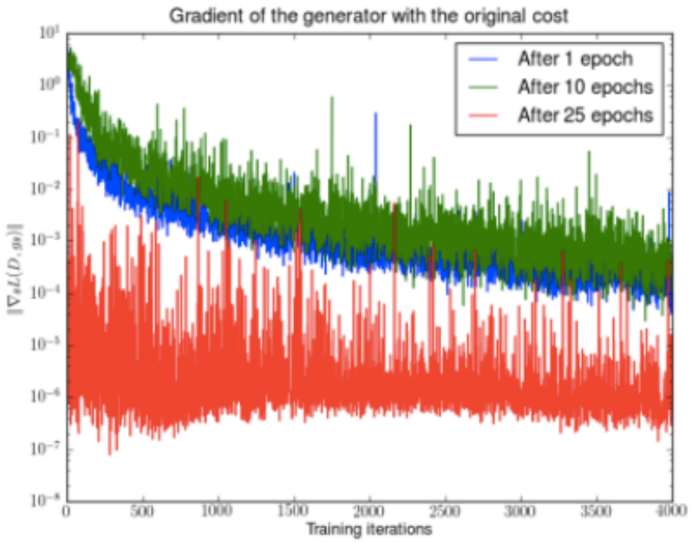
- Keeping the discriminator optimal : $\min_\theta L(\phi^*(\theta), \theta)$ minimizes $JS(P_r, P_g)$
- Keeping the generator optimal $\max_\phi L(\phi, \theta^*(\phi))$ yields garbage

Problem with GAN training

If one trains the discriminator thoroughly, the generator receives no gradient…







Alternate GAN training

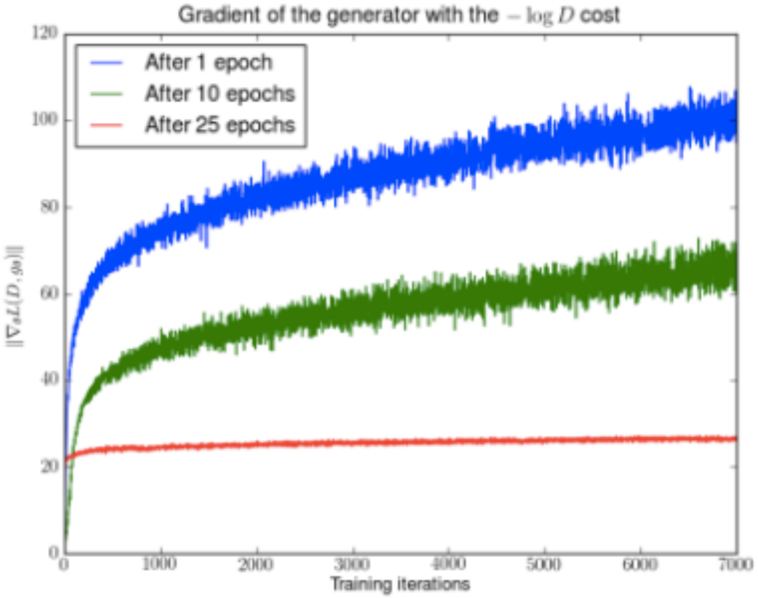
Alternate update that has less vanishing gradients

$$\Delta\theta \propto \mathbb{E}_{z\sim p_Z}[\nabla_\theta \log(D_\phi(g_\theta(z)))]$$

Under optimality optimizes

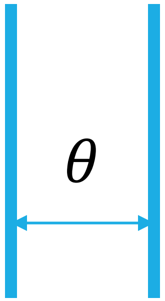
$$KL(\mathbb{P}_\theta||\mathbb{P}_r) - 2JSD(\mathbb{P}_r||\mathbb{P}_\theta)$$

Problems: JSD with the wrong sign, reverse KL has high mode dropping. Still unstable when D is good.



Distributions with low dimensional support

Let P_0 and P_θ be two uniform distributions supported by parallel line segments separated by distance θ.



Continuous in θ

- $W(\mathbb{P}_0, \mathbb{P}_\theta) = |\theta|,$
- $JS(\mathbb{P}_0, \mathbb{P}_\theta) = \begin{cases} \log 2 & \text{if } \theta \neq 0 \ , \\ 0 & \text{if } \theta = 0 \ , \end{cases}$
- $KL(\mathbb{P}_\theta \| \mathbb{P}_0) = KL(\mathbb{P}_0 \| \mathbb{P}_\theta) = \begin{cases} +\infty & \text{if } \theta \neq 0 \ , \\ 0 & \text{if } \theta = 0 \ , \end{cases}$
- and $\delta(\mathbb{P}_0, \mathbb{P}_\theta) = \begin{cases} 1 & \text{if } \theta \neq 0 \ , \\ 0 & \text{if } \theta = 0 \ . \end{cases}$

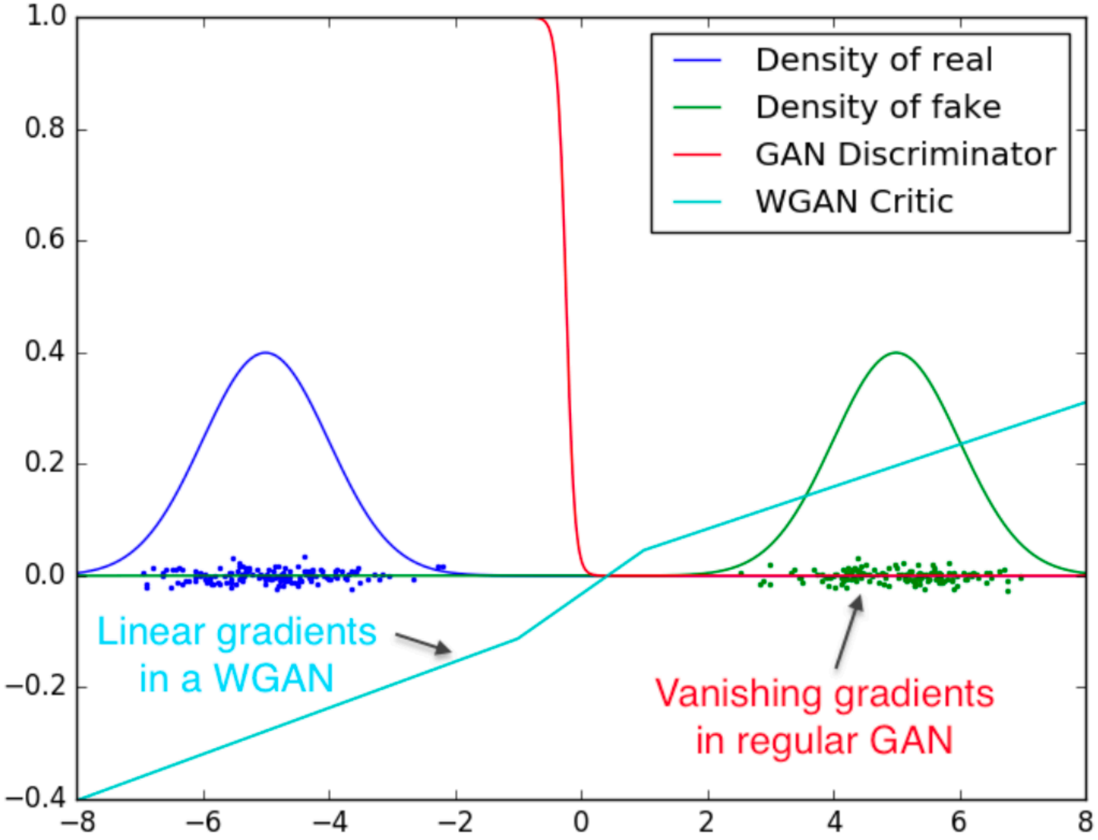
Optimizing a Wasserstein(ish) distance

Wasserstein-1 has a simple dual formulation (Kantorovich)

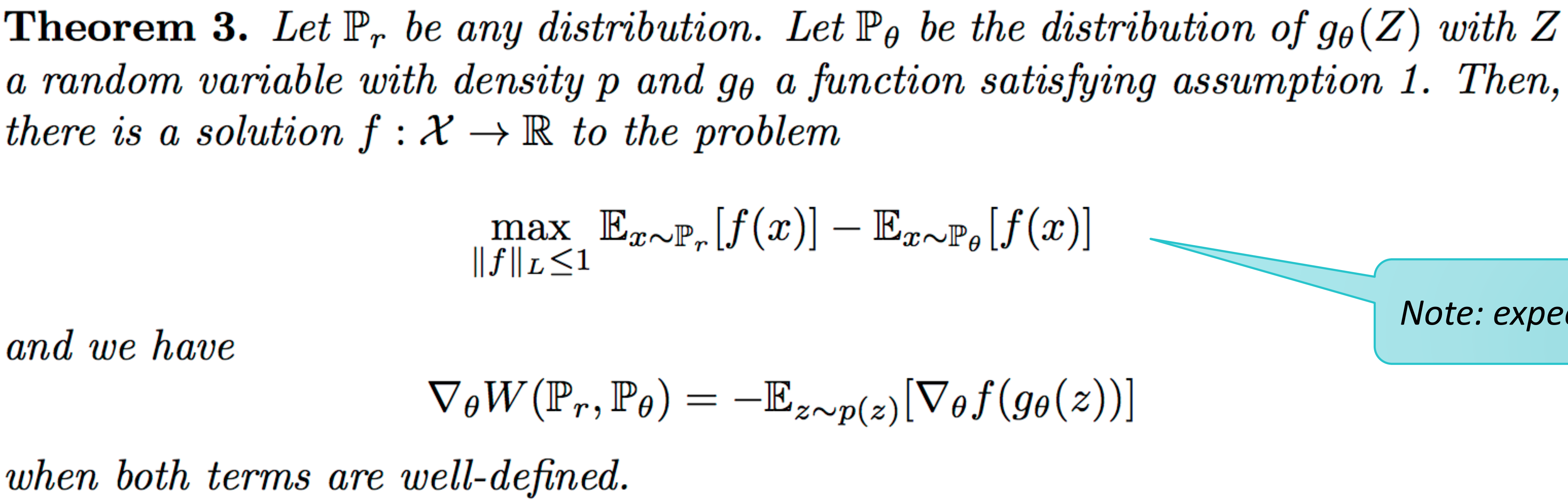
$$W(\mathbb{P}_r, \mathbb{P}_\theta) = \max_{\|f\|_L \leq 1} \mathbb{E}_{x \sim \mathbb{P}_r}[f(x)] - \mathbb{E}_{x \sim \mathbb{P}_\theta}[f(x)]$$

- Parametrize $f(x)$, for instance with a neural network.
- Enforce Lipschitz constraint, for instance by aggressively clipping the weights.
- Maintain $f(x)$ well trained, and train $G_\theta(z)$ by back-prop through $f(x)$.
- No vanishing gradients!

No vanishing gradients



Theorem



Theorem 3. *Let $\mathbb{P}_r$ be any distribution. Let $\mathbb{P}_\theta$ be the distribution of $g_\theta(Z)$ with Z a random variable with density p and g_θ a function satisfying assumption 1. Then, there is a solution $f : \mathcal{X} \to \mathbb{R}$ to the problem*

$$\max_{\|f\|_L \leq 1} \mathbb{E}_{x \sim \mathbb{P}_r}[f(x)] - \mathbb{E}_{x \sim \mathbb{P}_\theta}[f(x)]$$

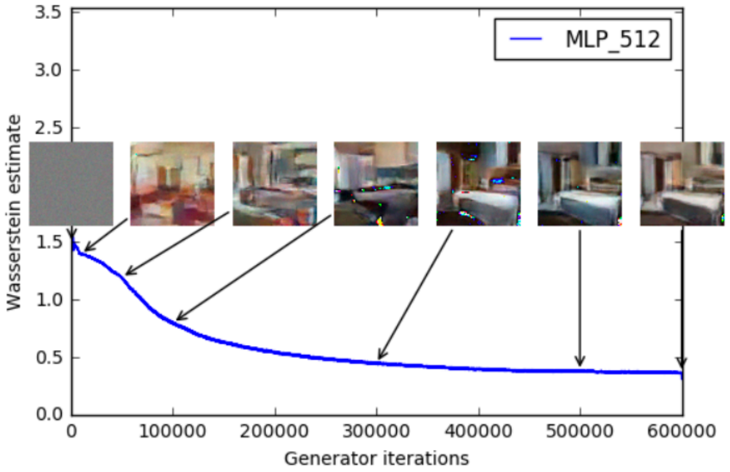
and we have

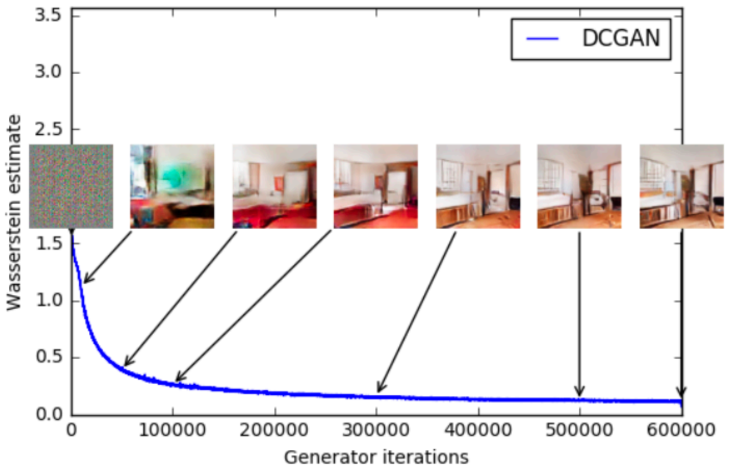
$$\nabla_\theta W(\mathbb{P}_r, \mathbb{P}_\theta) = -\mathbb{E}_{z \sim p(z)}[\nabla_\theta f(g_\theta(z))]$$

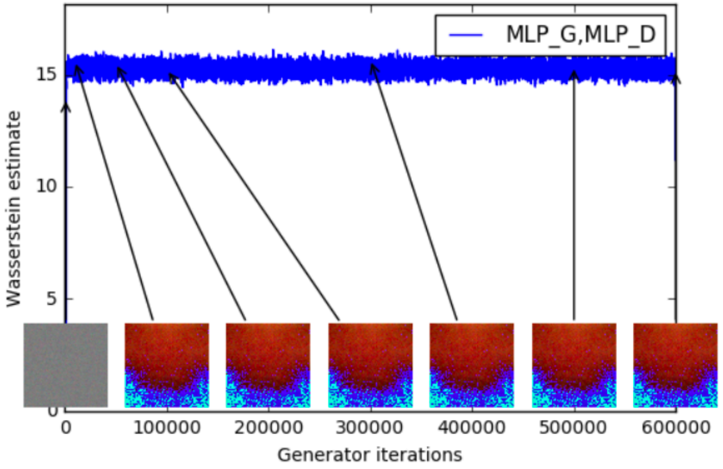
when both terms are well-defined.

Note: expectations

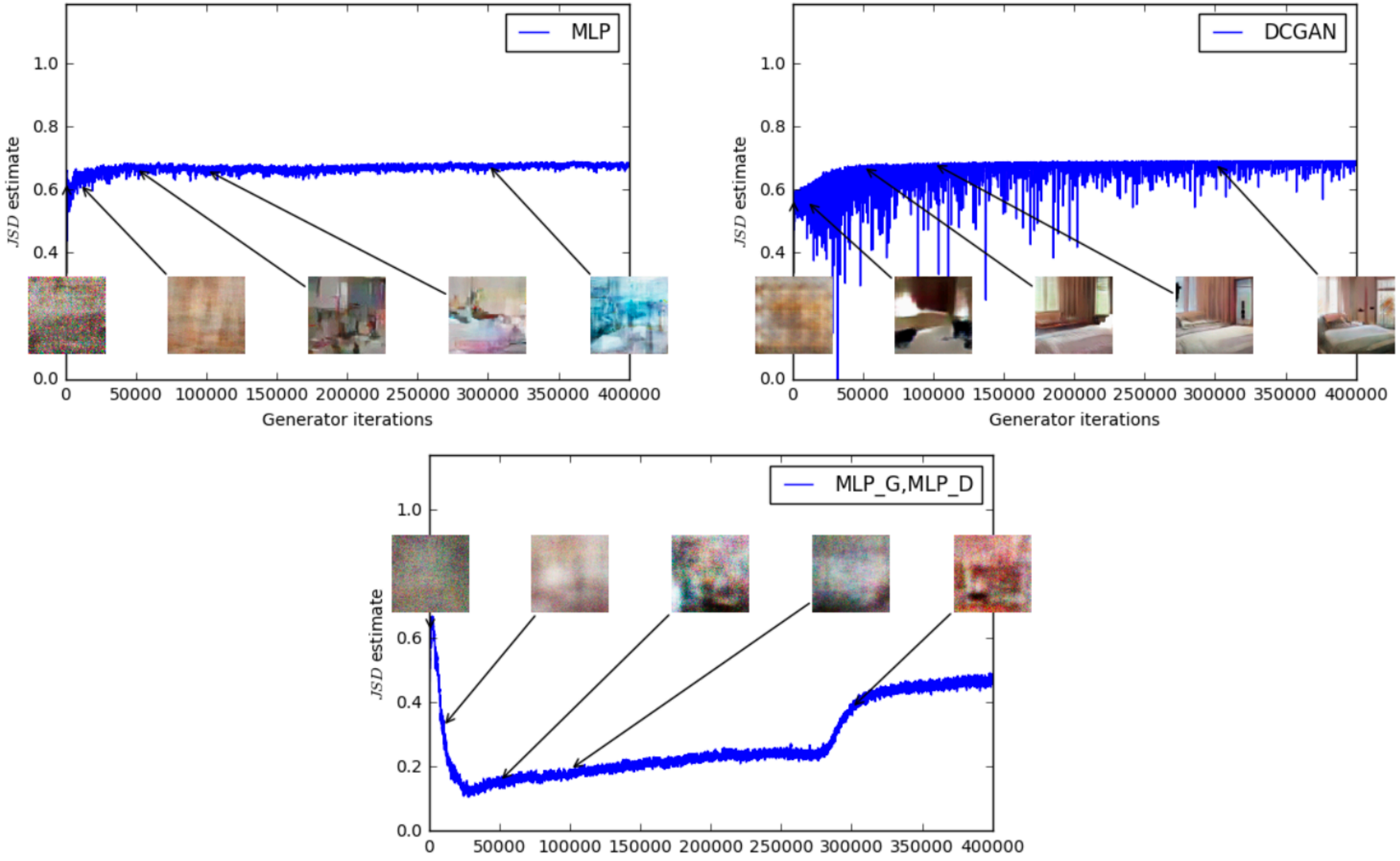
WGAN loss correlates with sample quality







GAN loss does not correlate with sample quality



WGAN is less sensitive to modeling choices

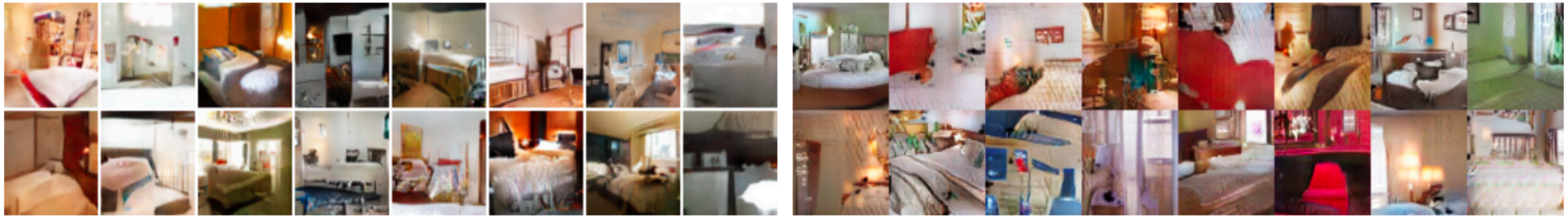


Figure 5: Algorithms trained with a DCGAN generator. Left: WGAN algorithm. Right: standard GAN formulation. Both algorithms produce high quality samples.

WGAN is less sensitive to modeling choices

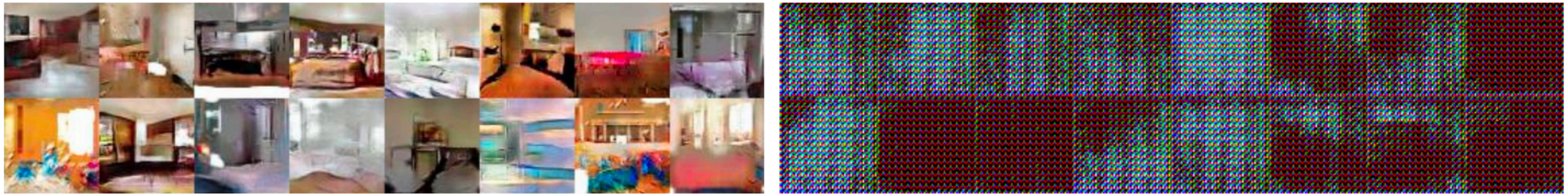


Figure 6: Algorithms trained with a generator without batch normalization and constant number of filters at every layer (as opposed to duplicating them every time as in [18]). Aside from taking out batch normalization, the number of parameters is therefore reduced by a bit more than an order of magnitude. Left: WGAN algorithm. Right: standard GAN formulation. As we can see the standard GAN failed to learn while the WGAN still was able to produce samples.

WGAN is less sensitive to modeling choices

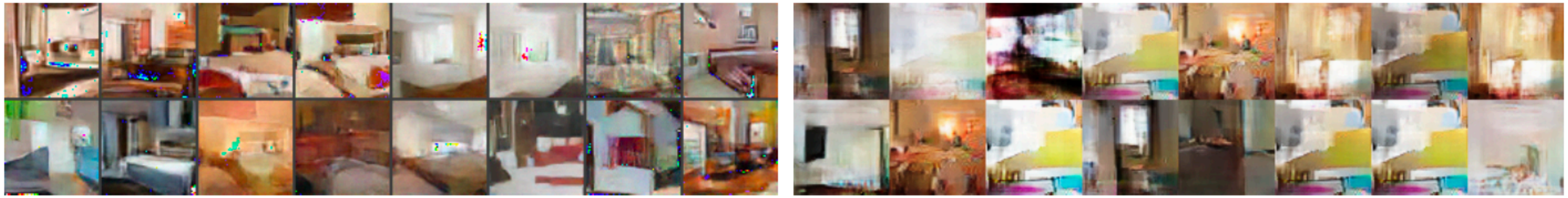


Figure 7: Algorithms trained with an MLP generator with 4 layers and 512 units with ReLU nonlinearities. The number of parameters is similar to that of a DCGAN, but it lacks a strong inductive bias for image generation. Left: WGAN algorithm. Right: standard GAN formulation. The WGAN method still was able to produce samples, lower quality than the DCGAN, and of higher quality than the MLP of the standard GAN. Note the significant degree of mode collapse in the GAN MLP.

WGAN

- Many authors have advocated using W distance to estimate densities.
  (Rozasco et al, 2012, Cuturi et al, 2015, …)
- Maximum Mean Discrepancy
  (Gretton et al, 2012)
- Our originality is a focus on continuous distributions with low dim support, and the idea to parametrize f in order to obtain a fast algorithm.

Conclusion

In Search for Lost Signal

- There is a causal signal in the high moments.
- It takes the form of cliffs, corners, shocks, etc.
- This has everything to do with the mythical unsupervised learning
- Weak distribution distances such as Wasserstein seem more able to catch it.
- This is just a beginning.