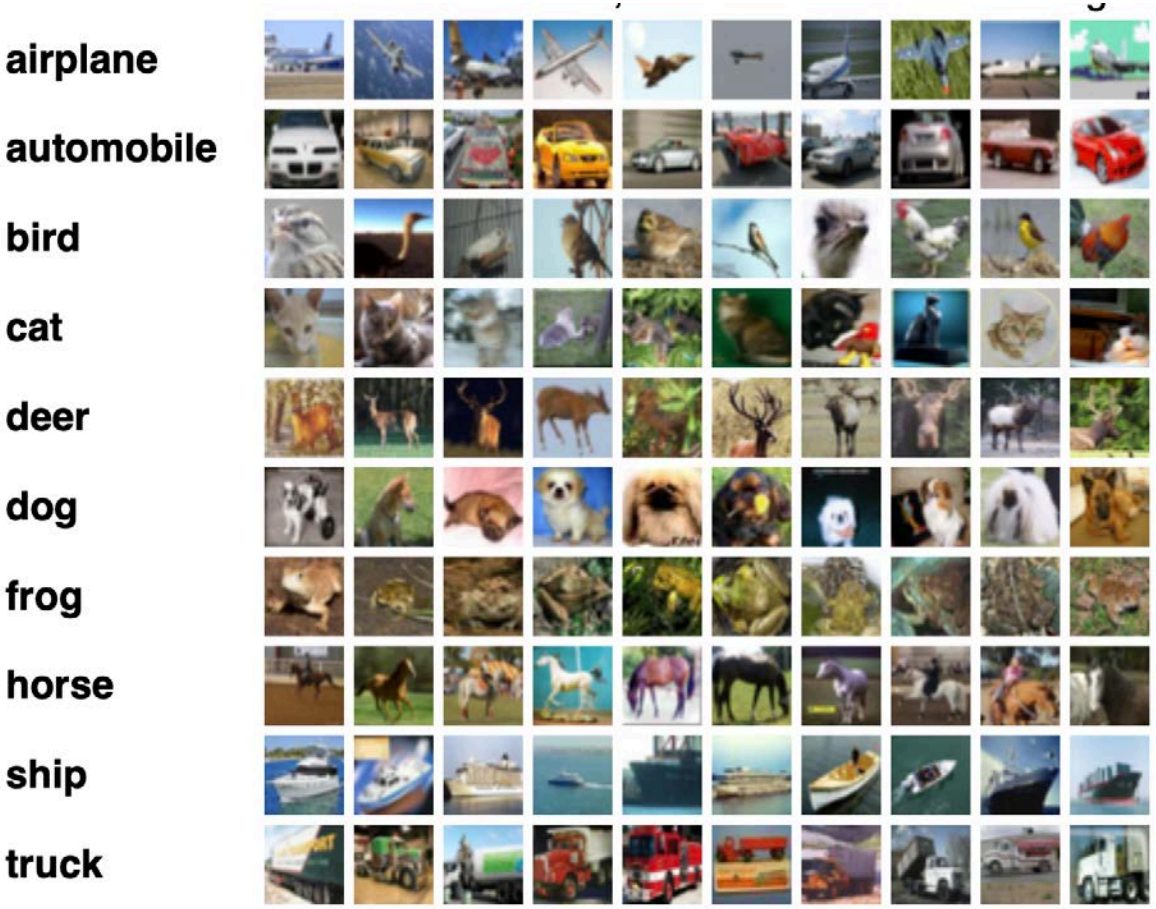
Fixing Bias in Reconstruction-Based Anomaly Detection with Lipschitz Discriminators

Alexander Tong, Guy Wolf, Smita Krishnaswamy

High Level Problem:

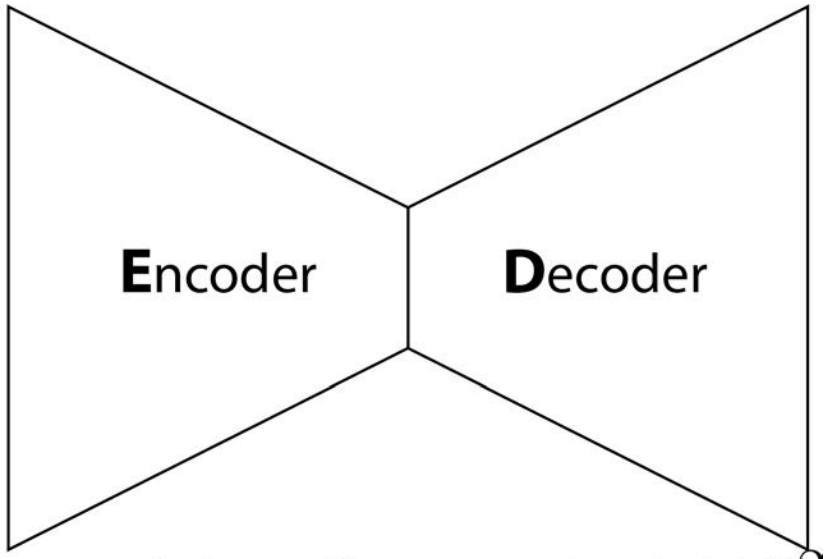
Given samples from the nominal distribution produce an anomaly scoring function that is high on anomalous points and low on nominal points.

Deep anomaly detection (on image classification data)



Deep anomaly detection (on image classification data)

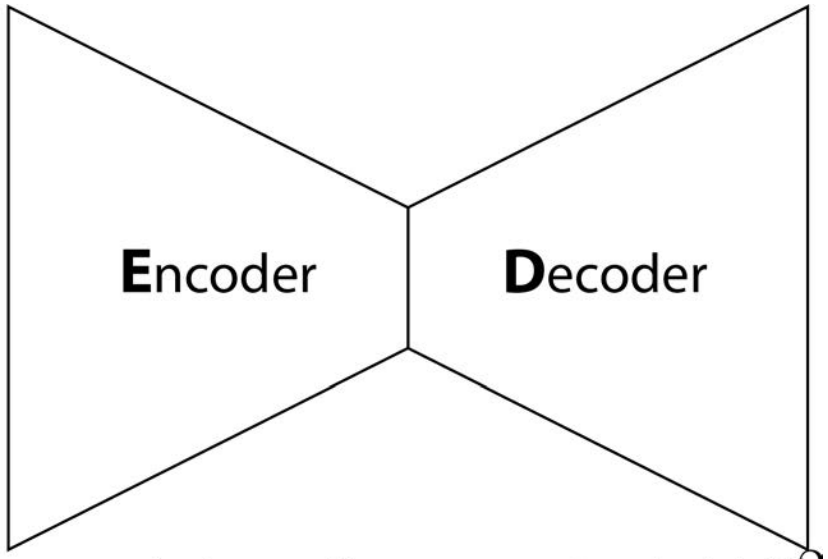
Reconstruction-based anomaly detection



$$\text{Loss}(x) = \|x - \mathbf{D}(\mathbf{E}(x))\|_2^2$$

$$\text{Score}(x) = \|x - \mathbf{D}(\mathbf{E}(x))\|_2^2$$

Reconstruction-based anomaly detection



$$\text{Loss}(x) = \|x - \mathbf{D}(\mathbf{E}(x))\|_2^2$$

$$\text{Score}(x) = \|x - \mathbf{D}(\mathbf{E}(x))\|_2^2$$

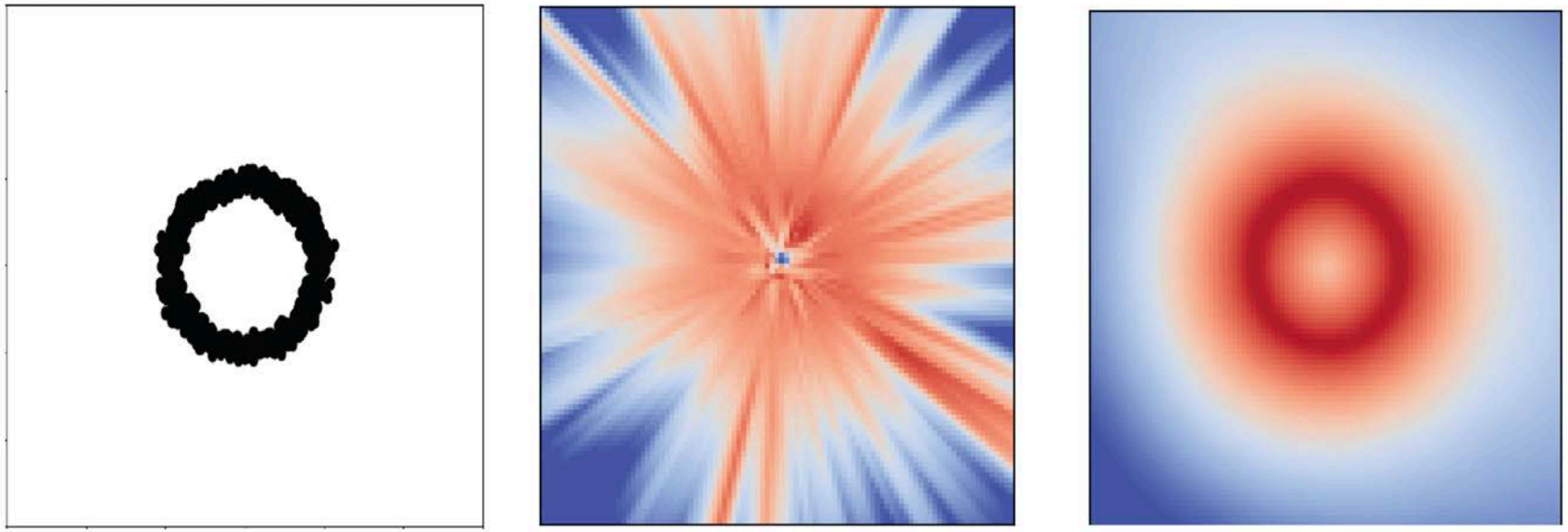
Anomalies are *implicitly* defined as inputs that are difficult to reconstruct

Desirable properties from a scoring function

- Similar inputs have similar scores
- Robust to a small number of anomalous samples in the training data
- Very distant points from the training set are classified as anomalous

Desirable properties from a scoring function

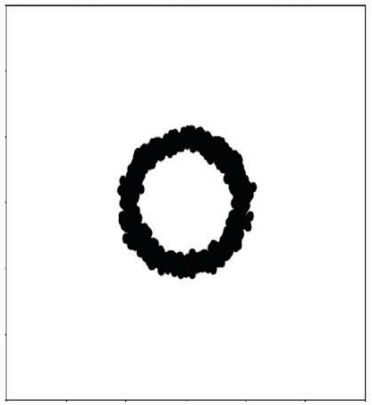
- Similar inputs have similar scores
- Robust to a small number of anomalous samples in the training data
- Very distant points from the training set are classified as anomalous



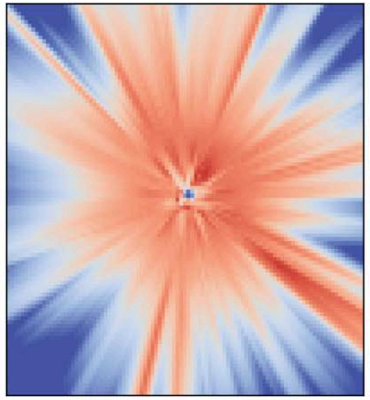
(a) Input Data (b) AE (c) LAD

A transport view of anomaly detection

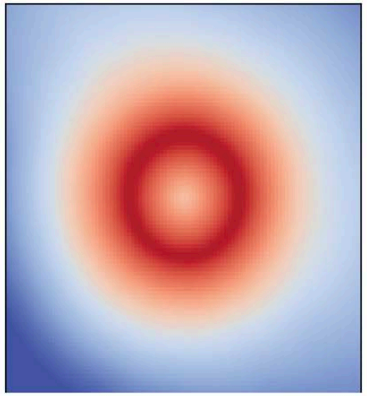
The optimal 1-Lipschitz scoring function is (up to a constant) the Kantorovich-Rubenstein witness function between the nominal and (unknown) anomalous distribution.



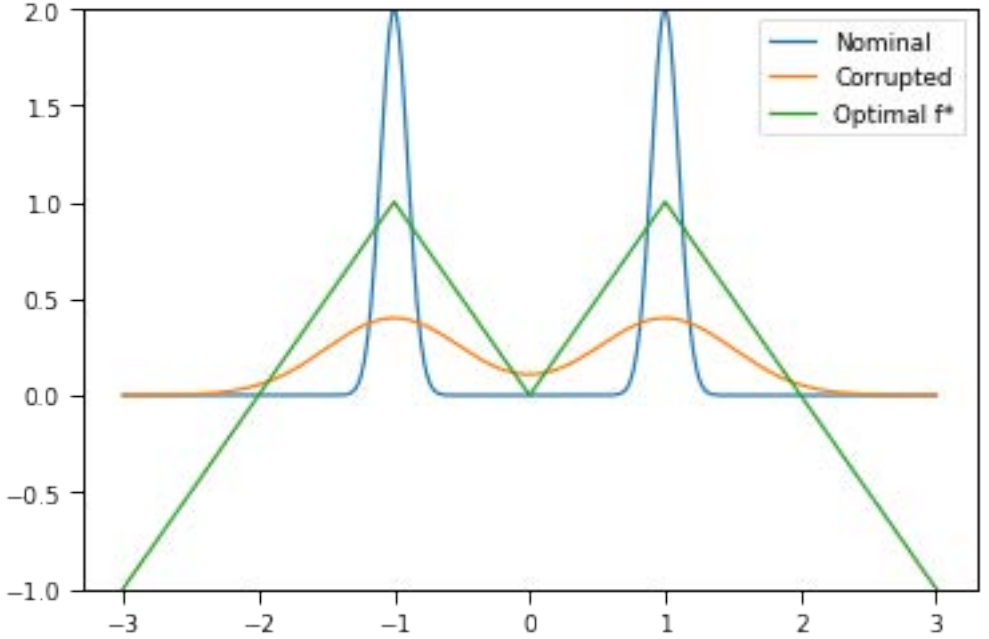
(a) Input Data



(b) AE



(c) LAD



Transport and the Kantorovich-Rubenstein Dual

- 1-Wasserstein distance between P and Q is defined as:

$$W(P,Q) = \inf_{\pi \in \Pi(P,Q)} \mathbb{E}_{(x,y)\sim\pi}\left[d(x,y)\right]$$

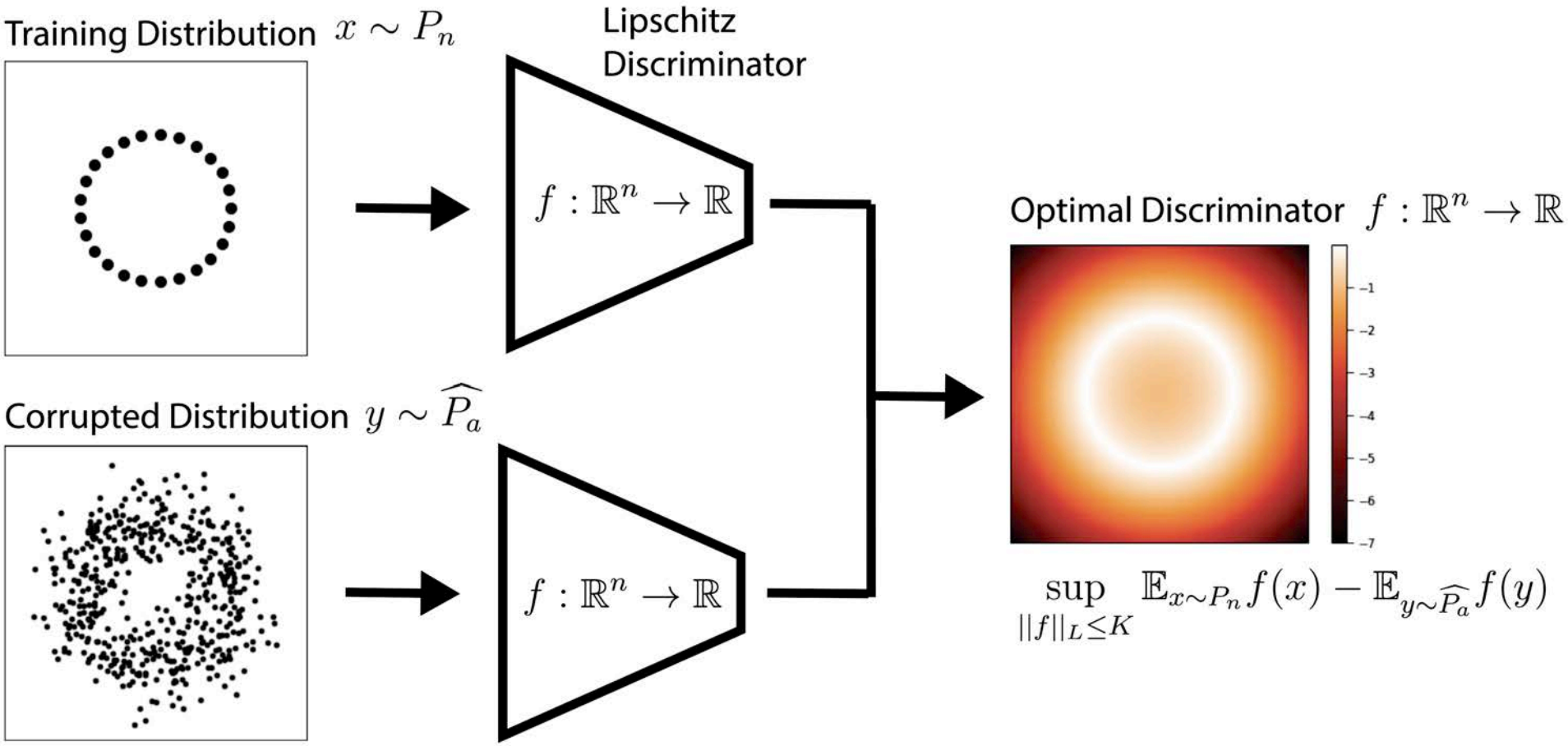
Where $\mathbf{\Pi}$ is the family of joint distribution between P and Q, and d(x,y) is a ground distance between points

- For d(x,y) = |x – y|, Kantorovich-Rubenstein duality says

$$W(P,Q) = \sup_{\|f\|_L \leq 1} \mathbb{E}_{x\sim P} f(x) - \mathbb{E}_{x\sim Q} f(x)$$

f is known as the “witness function”

The Lipschitz anomaly discriminator



How to train a Lipschitz neural network?

- Gradient Clipping [Arjovsky et al. 2017]

$$w \leftarrow \mathrm{clip}(w, -c, c)$$

- Gradient norm penalization [Gulrajani et al. 2017]

$$\lambda \mathbb{E}_{x \sim P_x} \left[(\|\nabla_x f(x)\|_2 - 1)^2 \right]$$

- Spectral normalization [Miyato et al. 2018]

$$w \leftarrow w / \sigma(w)$$

How to train a Lipschitz neural network?

- Gradient Clipping [Arjovsky et al. 2017]

$$w \leftarrow \mathrm{clip}(w, -c, c)$$

- **Gradient norm penalization [Gulrajani et al. 2017]**

$$\lambda \mathbb{E}_{x \sim P_x} \left[(\|\nabla_x f(x)\|_2 - 1)^2 \right]$$

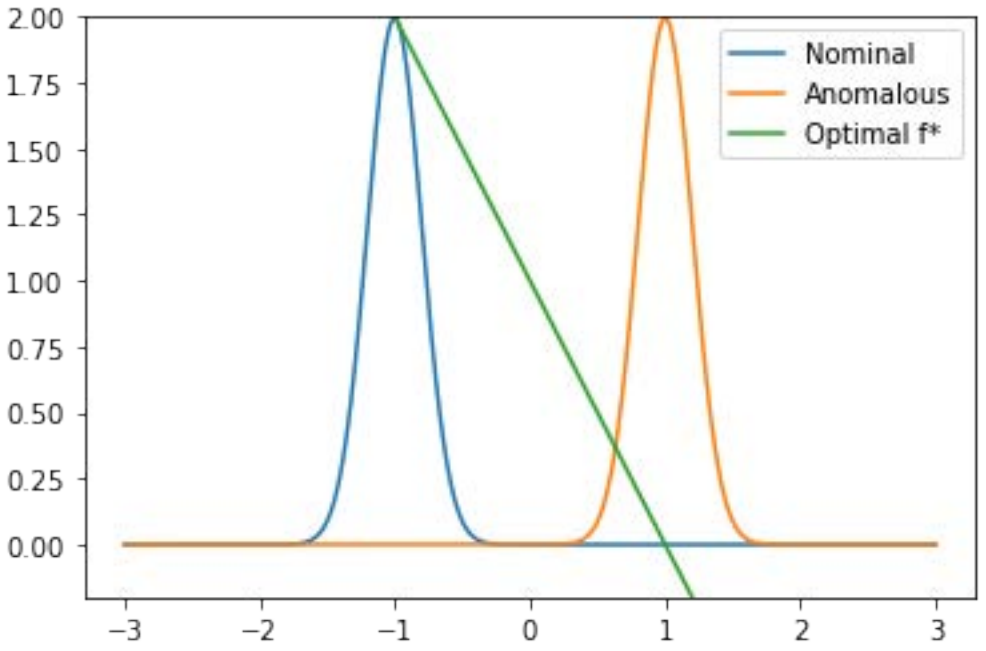
- Spectral normalization [Miyato et al. 2018]

$$w \leftarrow w / \sigma(w)$$

Robustness of the optimal score to training set corruption

Prop 1: Adding anomalies to the training set does not affect the scores very much. Let f* be optimal scoring function, and f** be optimal under corrupted training set $(1-\gamma)P_n + \gamma P_a$

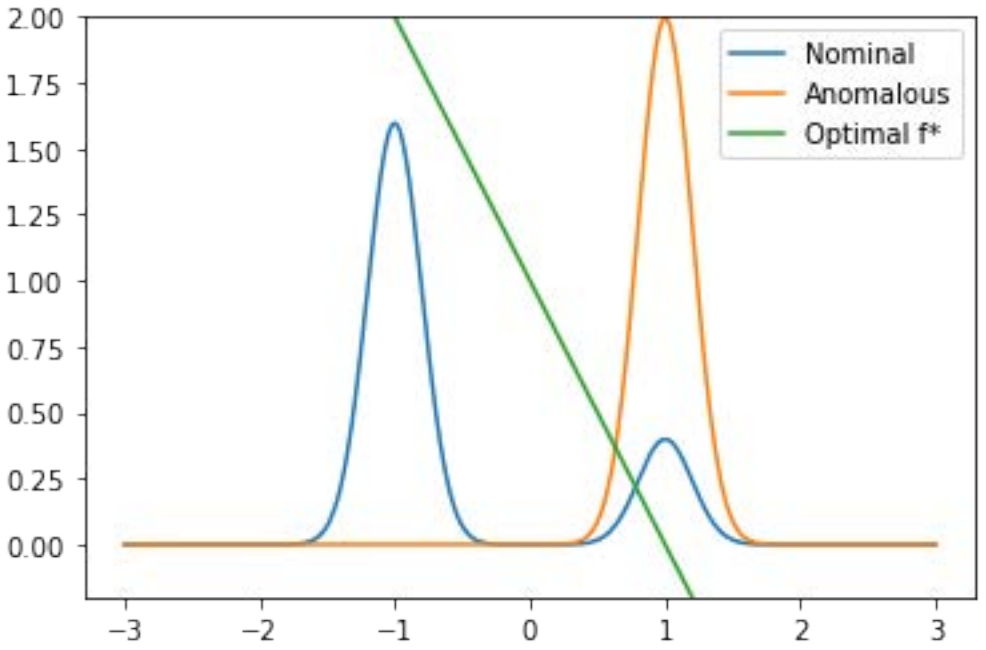
$$\left|\mathbb{E}_{x\sim P_n}[f^*(x) - f^{**}(x)] + \mathbb{E}_{x\sim P_a}[f^{**}(x) - f^*(x)]\right| \leq \frac{1}{1-\gamma} W(P_n, (1-\gamma)P_n + \gamma P_a)$$



Robustness of the optimal score to training set corruption

Prop 1: Adding anomalies to the training set does not affect the scores very much. Let f* be optimal scoring function, and f** be optimal under corrupted training set $(1-\gamma)P_n + \gamma P_a$

$$\left|\mathbb{E}_{x\sim P_n}[f^*(x) - f^{**}(x)] + \mathbb{E}_{x\sim P_a}[f^{**}(x) - f^*(x)]\right| \leq \frac{1}{1-\gamma} W(P_n, (1-\gamma)P_n + \gamma P_a)$$

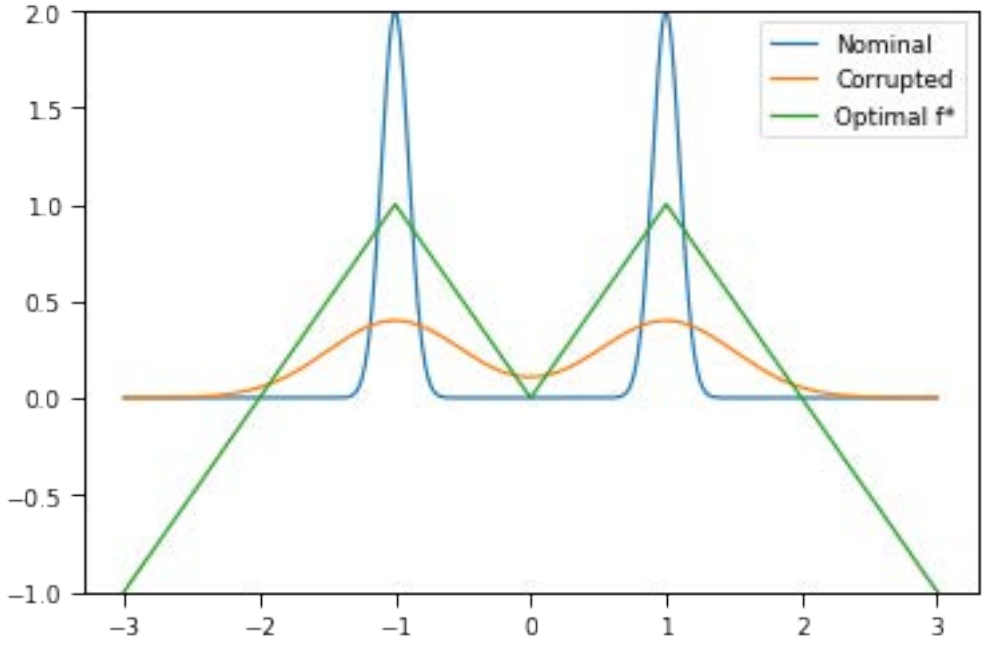


Outside a radius C, all points are scored anomalous

Prop 2: There exists a constant C such that for the optimal f*, nominal distribution P_n with support S_n, and anomalous distribution P_a

$$f^*(y) \leq C - \inf_{x \in S_n} \{\|x - y\|\} \text{ for } P_a\text{-almost every } y$$

Corollary: Outside a radius R, every point is scored anomalous



Revisiting properties

- Similar inputs have similar scores

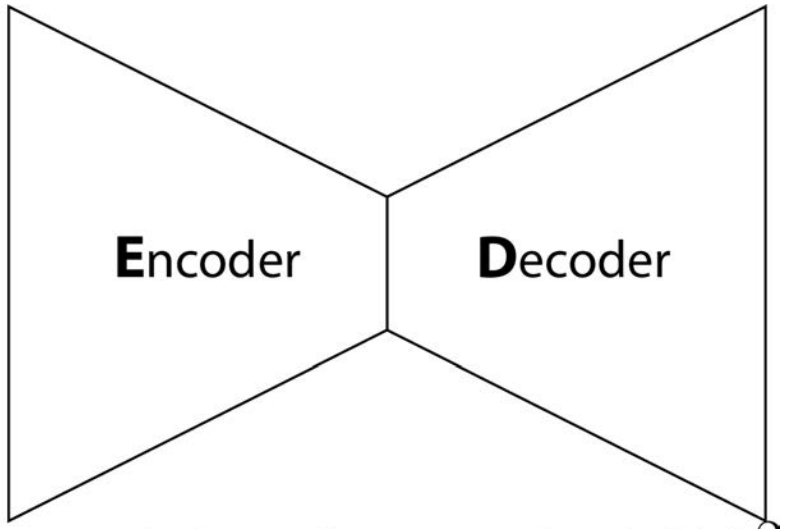
  Enforced by gradient penalty
- Robust to a small number of anomalous samples in the training data

  Prop 1.
- Very distant points from the training set are classified as anomalous

  Prop 2.

Training set contamination

- It is unrealistic to assume a large training set only contains points from the nominal distribution
- MSE training approximately equally low anomaly scores across data



$$\text{Loss}(x) = \|x - \mathbf{D}(\mathbf{E}(x))\|_2^2$$

$$\text{Score}(x) = \|x - \mathbf{D}(\mathbf{E}(x))\|_2^2$$

Training set contamination

MNIST training set corruption

Use discriminator loss

Use reconstruction score

Traditional baselines

Deep SVM models

Train Corrupt.	0.00	0.01	0.05	0.10
ALOCC [6]	0.694	0.511	0.539	0.509
AND [5]	0.975	-	-	-
AnoGAN [7]	0.913	-	-	-
CAE [11]	0.965	0.925	0.868	0.832
DCAE [12]	0.967	0.925	0.865	0.829
DSVDD [14]	0.748	0.788	0.718	0.696
IF [15]	0.853	0.853	0.837	0.822
LOF [16]	0.973	0.958	0.789	0.709
OCSVM [13]	0.954	0.895	0.828	0.794
RCAE [4]	0.957	0.934	0.870	0.832
LAD (ours)	0.940	0.937	0.923	0.901
LAD + CAE (ours)	**0.981**	**0.965**	**0.936**	**0.912**

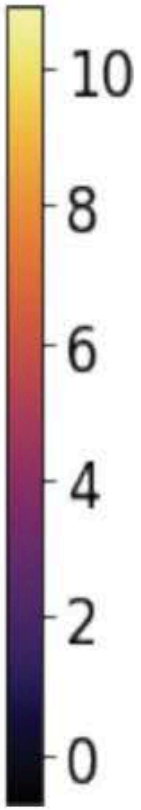
Mean AUROC over digits from 3 seeds for 0%-10% training set corruption on MNIST

MNIST – Relative score of the all black image

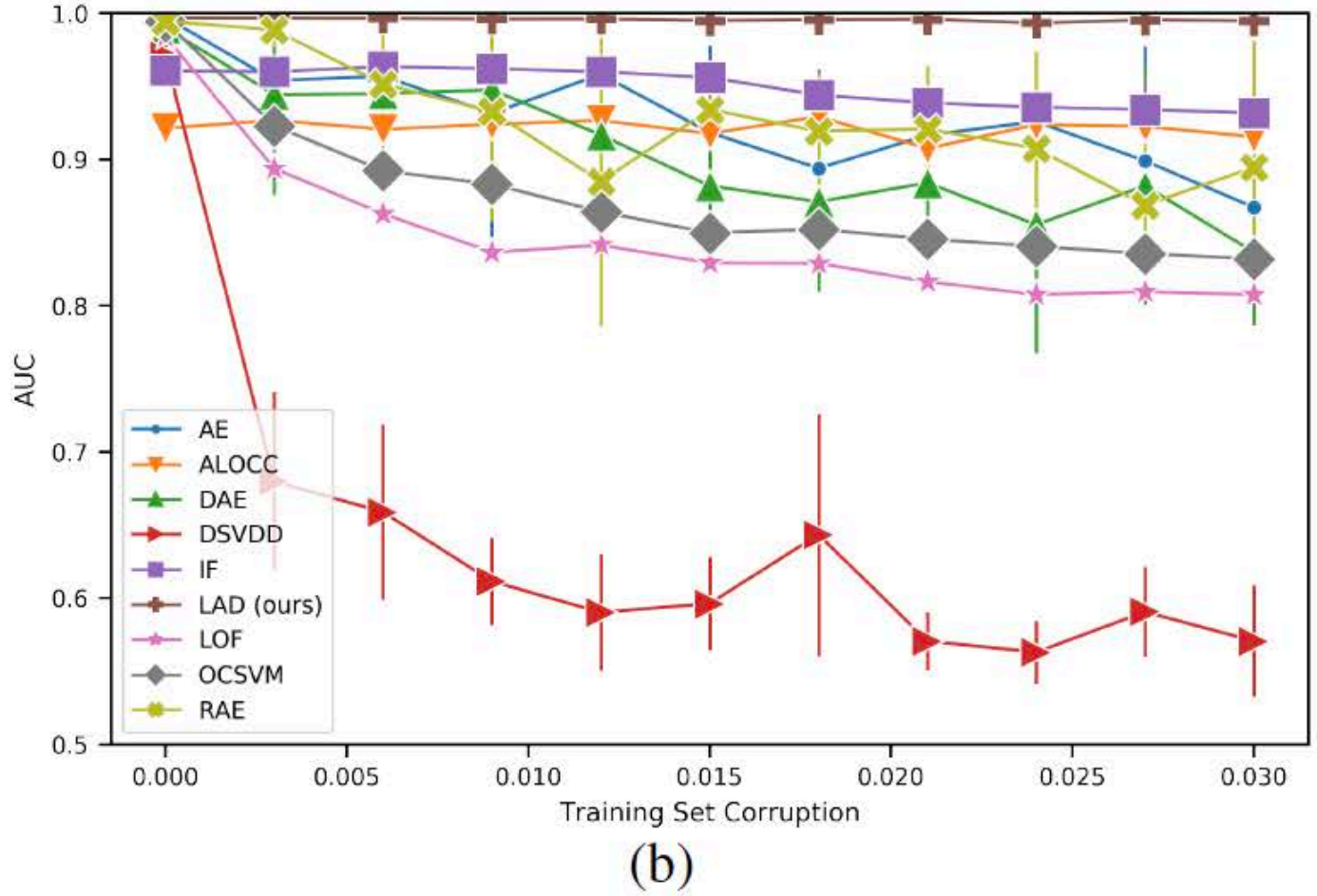
Train Corrupt.	0.00	0.01	0.05	0.10	Black
ALOCC [6]	0.694	0.511	0.539	0.509	0.168
AND [5]	0.975	-	-	-	-
AnoGAN [7]	0.913	-	-	-	-
CAE [11]	0.965	0.925	0.868	0.832	0.067
DCAE [12]	0.967	0.925	0.865	0.829	0.059
DSVDD [14]	0.748	0.788	0.718	0.696	0.571
IF [15]	0.853	0.853	0.837	0.822	0.312
LOF [16]	0.973	0.958	0.789	0.709	0.695
OCSVM [13]	0.954	0.895	0.828	0.794	0.677
RCAE [4]	0.957	0.934	0.870	0.832	0.049
LAD (ours)	0.940	0.937	0.923	0.901	**1.000**
LAD + CAE (ours)	**0.981**	**0.965**	**0.936**	**0.912**	**1.000**

Mean AUROC over digits from 3 seeds for 0%-10% training set corruption on MNIST

VACS data Training set Corruption



(a)



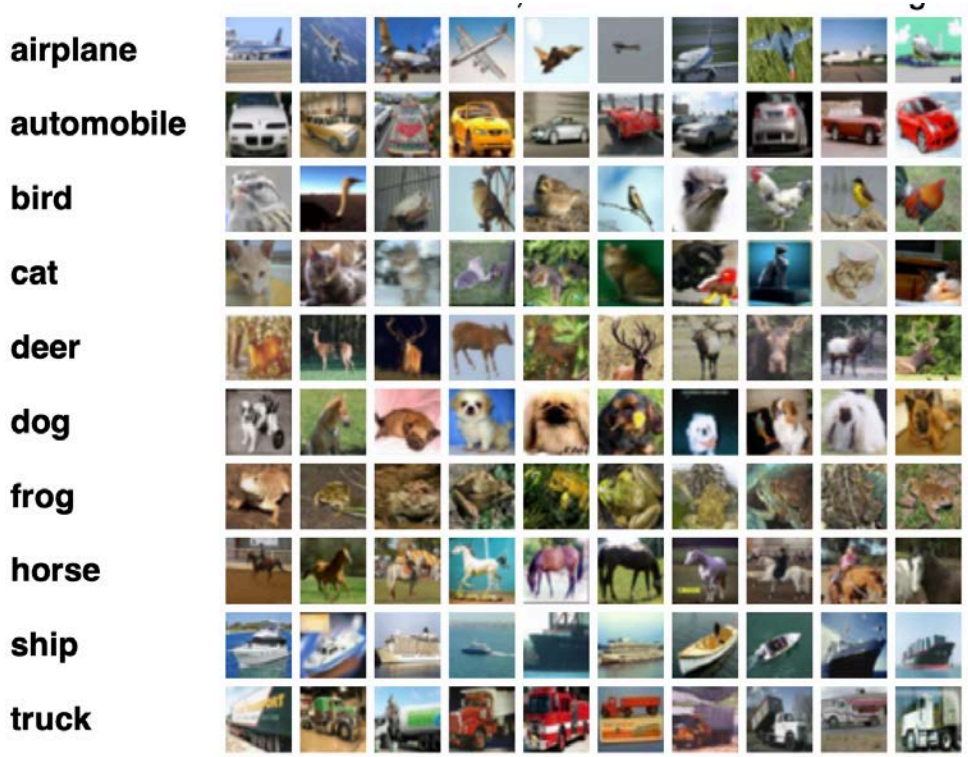
(b)

(a) Creatinine levels on electronic health record dataset. Creatinine > 2 was taken as anomalous
(b) AUROC for various models with some a small percentage of the training data containing high creatinine patients

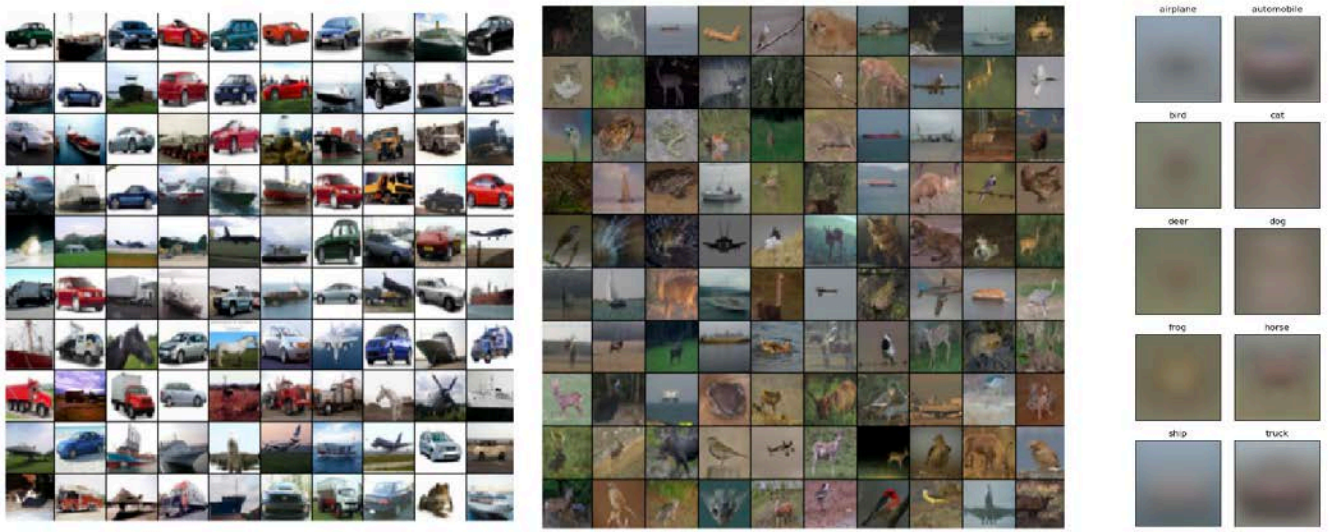
CIFAR10 – performance depending on class

Class	plane	car	bird	dog	mean
ALOCC [6]	0.421	0.439	0.530	0.473	0.463
AND [5]	0.717	0.494	0.662	0.504	0.617
AnoGAN [7]	0.671	0.547	0.529	**0.603**	0.618
CAE [11]	0.683	0.454	0.677	0.525	0.604
DCAE [12]	0.689	0.447	**0.679**	0.526	0.605
DSVDD [14]	0.518	0.656	0.528	0.568	0.571
IF [15]	0.670	0.442	0.645	0.516	0.599
LOF [16]	0.661	0.440	0.649	0.511	0.575
OCSVM [17]	0.684	0.456	0.674	0.502	0.590
RCAE [4]	0.675	0.429	0.669	0.531	0.592
LAD (ours)	0.597	**0.663**	0.411	0.561	0.565
LAD + CAE (ours)	**0.723**	0.497	0.652	0.544	**0.635**

Table 2: AUC on CIFAR10 for representative classes over 3 seeds with no corruption.



Reconstruction-based methods prefer images like the mean



(a) (b) (c)

Fig. 4: Top 100 nominal images in test set of LAD (a) and DCAE (b) trained on the automobile class. (c) Mean over examples of pixel values for each class. Many car images have white background and/or bright colors that are far from the mean image. LAD does better at modeling such images.

Conclusion

- Standard autoencoder based anomaly detection implicitly defines anomalies as hard for the model to reconstruct, leading to a few issues.
- These issues can be fixed with the addition of a neural network based on optimal transport theory which we call LAD.
- Combining LAD with existing models gives state of the art results on standard MNIST and CIFAR10 benchmarks.

Thanks!

- Lab Website: https://www.krishnaswamylab.org
- Email: alexander.tong@yale.edu
- Code: https://github.com/krishnaswamylab/LAD