Wasserstein adversarial examples via projected Sinkhorn iterations

Eric Wong[1] Frank R. Schmidt[2] J. Zico Kolter[13]

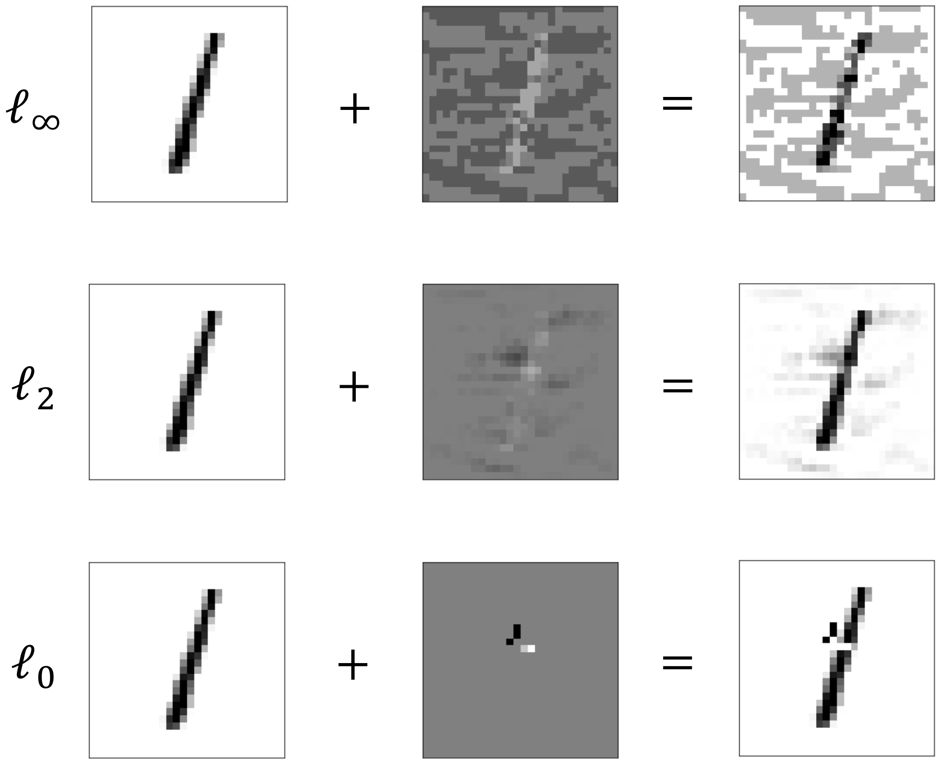
Code: https://github.com/locuslab/projected_sinkhorn/

[1]School of Computer Science
Carnegie Mellon University
Pittsburgh, PA, 15213, USA

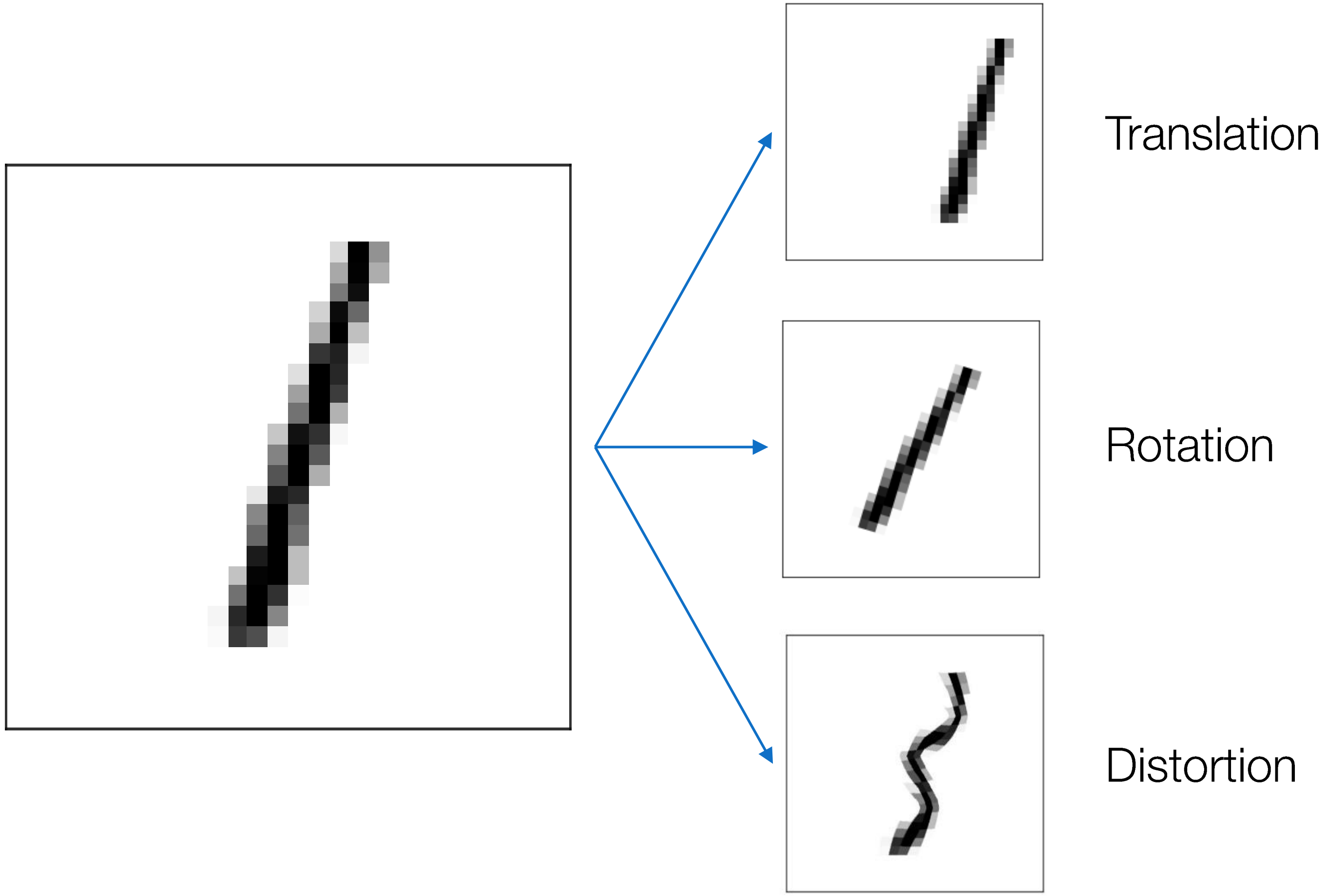
[2]Bosch Center for Artificial Intelligence
Pittsburgh, PA 15222, USA

[3]Bosch Center for Artificial Intelligence
Renningen, Germany

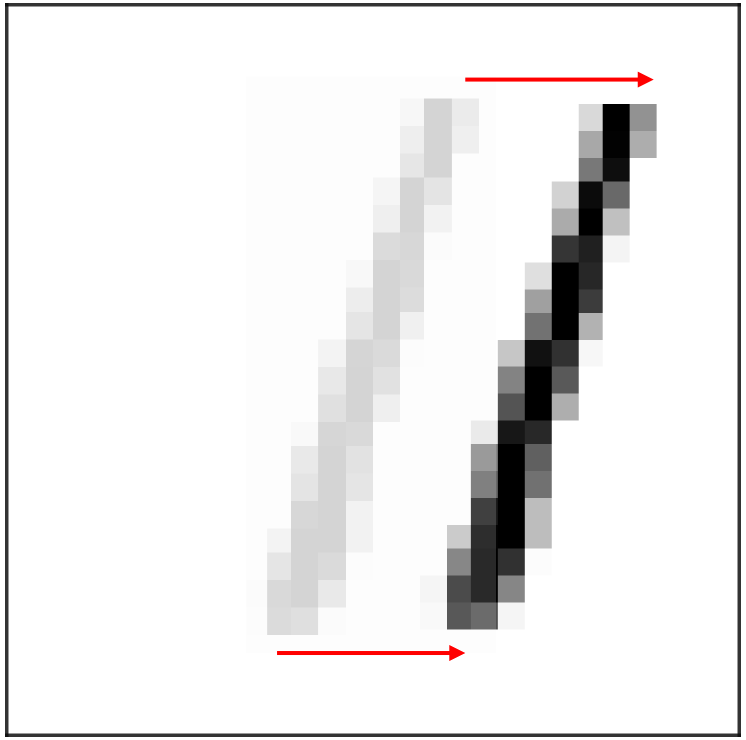
Typical threat model: norm-bounded perturbation



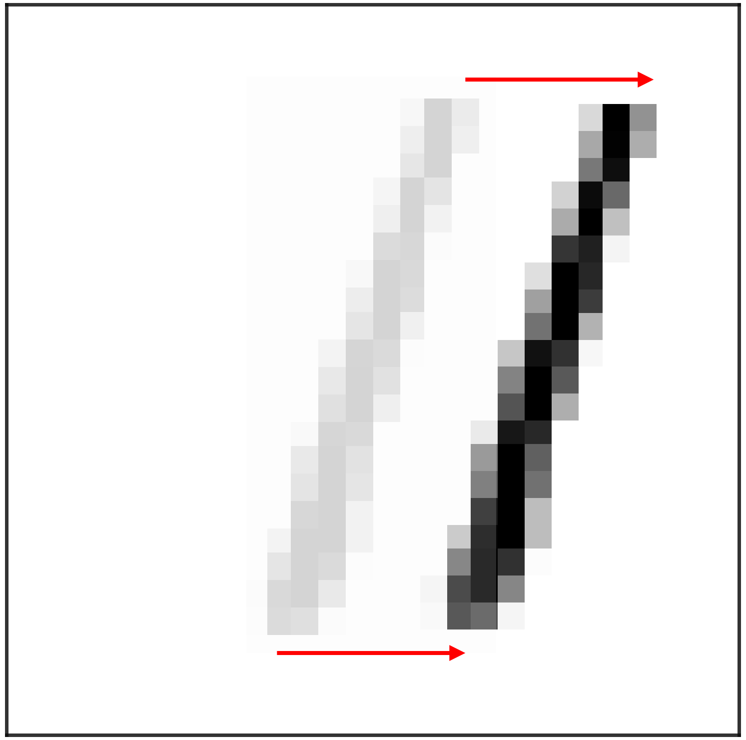
ℓ_p norms don't capture typical image transforms



These transforms move pixel mass short distances…

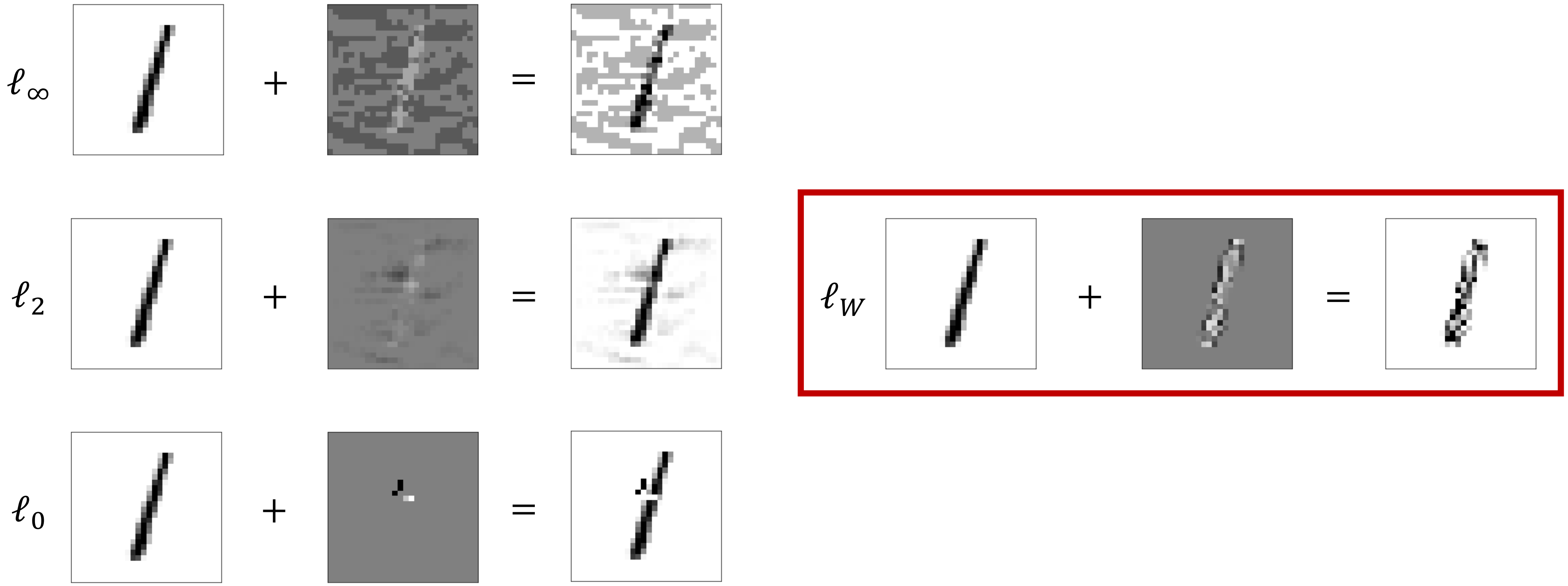


These transforms move pixel mass short distances…

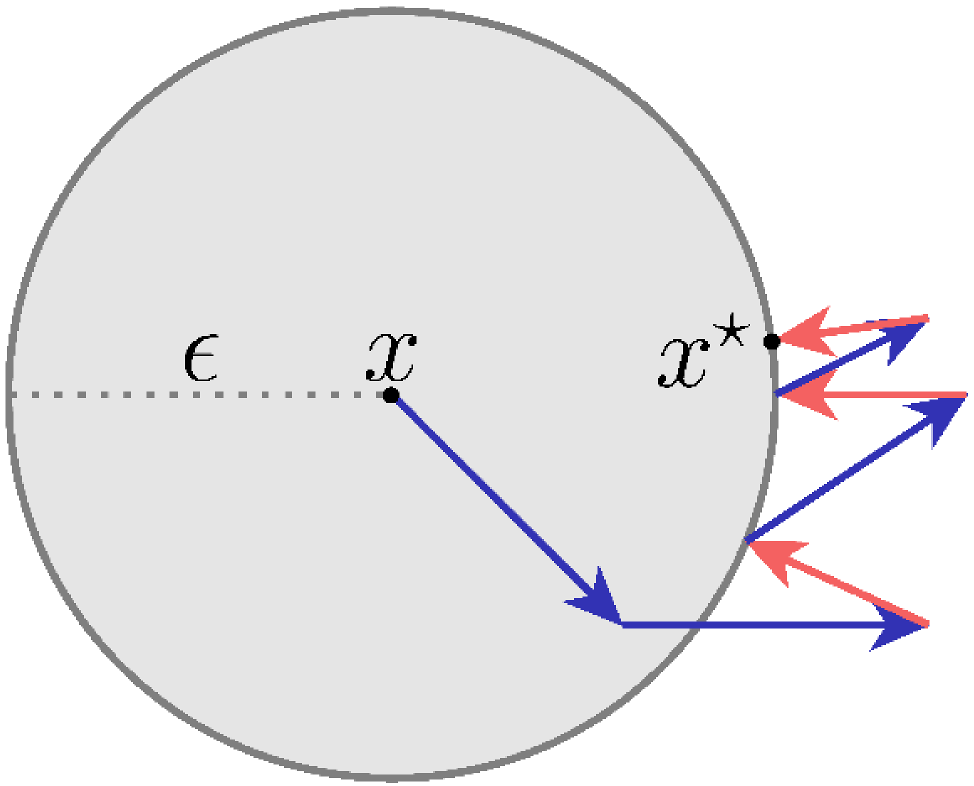


and the Wasserstein metric measures “moving mass”

We propose Wasserstein balls as a threat model



The strongest known method for generating adversarial examples is projected gradient descent



How to project onto the Wasserstein ball?

How to project onto the Wasserstein ball?

$$\underset{z \in \mathbb{R}^n,\, \Pi \in \mathbb{R}^{n \times n}}{\text{minimize}} \; ||w \; - z||_2^2$$

$$\text{subject to} \quad \begin{aligned} \Pi 1 &= x \\ \Pi^T 1 &= z \\ \langle \Pi, C \rangle &\leq \epsilon \end{aligned}$$

How to project onto the Wasserstein ball?

$$\underset{z \in \mathbb{R}^n,\, \Pi \in \mathbb{R}^{n \times n}}{\text{minimize}} \quad \boxed{||w - z||_2^2}$$

Closest point

$$\text{subject to} \quad \boxed{\begin{aligned} \Pi 1 &= x \\ \Pi^T 1 &= z \\ \langle \Pi, C \rangle &\leq \epsilon \end{aligned}}$$

Wasserstein ball constraints

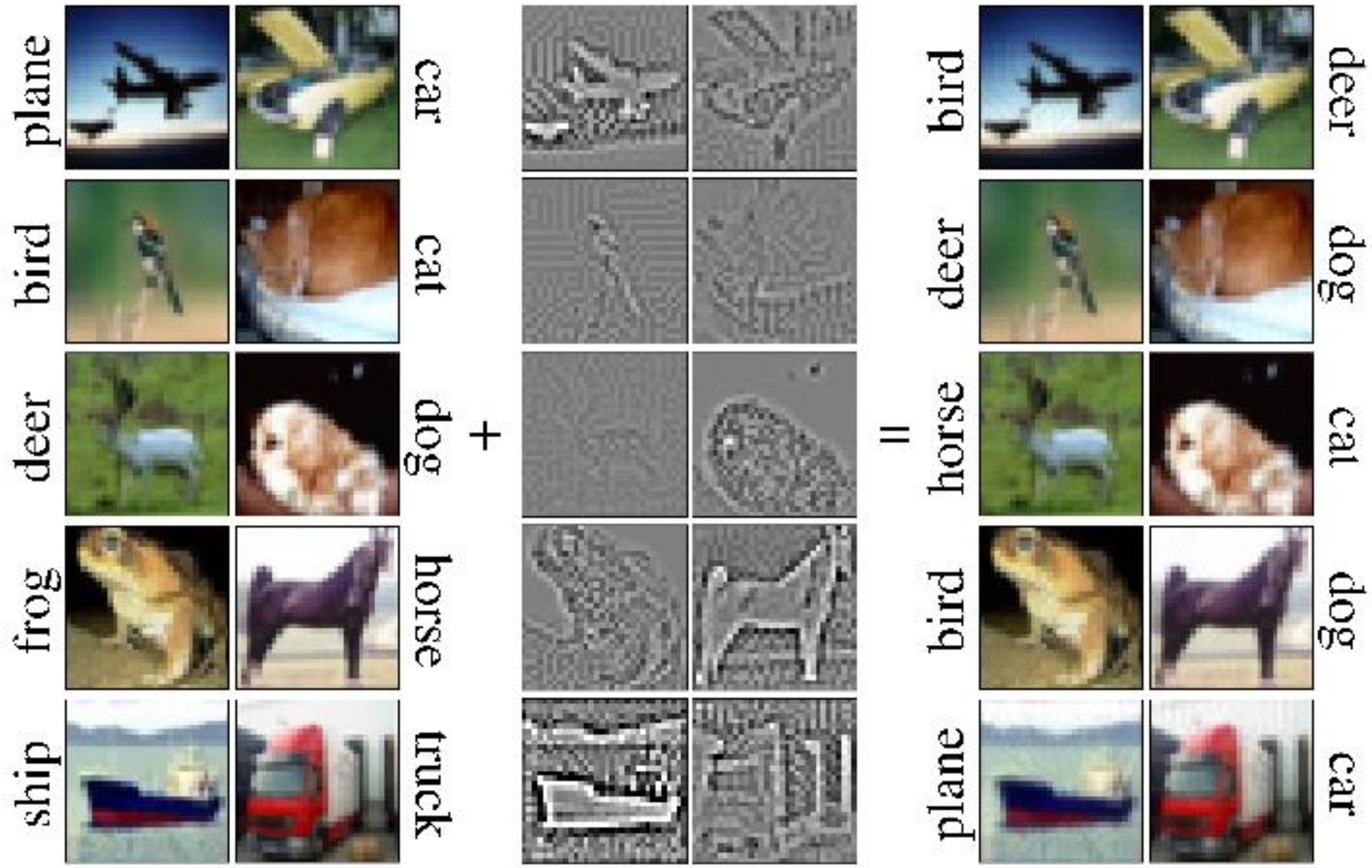
Quadratic program, quadratic number of variables → Costly!

Projected Sinkhorn Iteration: a fast (approximate) projection algorithm onto the Wasserstein ball

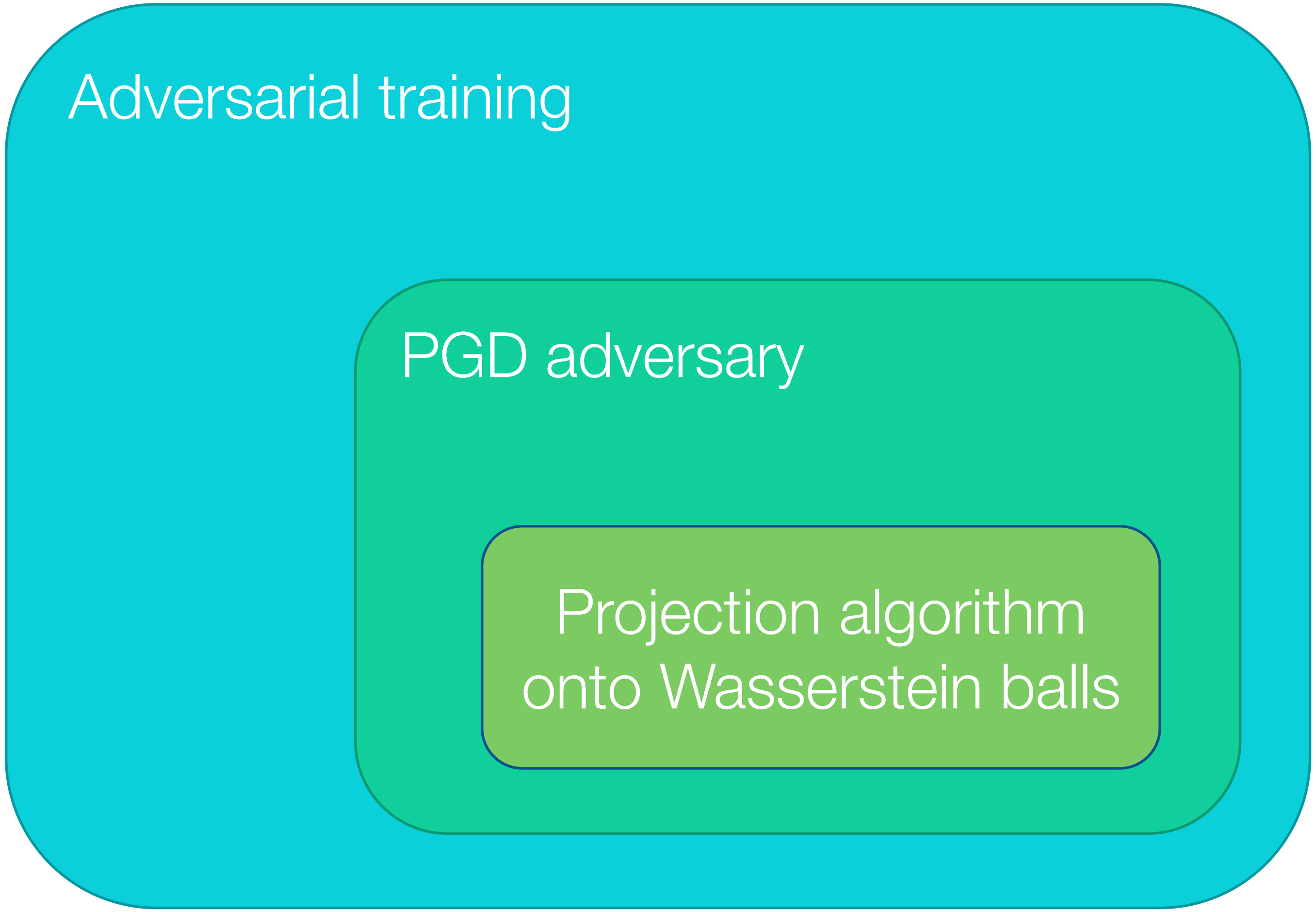
$$\begin{aligned}
\underset{z\in\mathbb{R}^n,\ \Pi\in\mathbb{R}^{n\times n}}{\text{minimize}} \quad & ||w - z||_2^2 + \boxed{\frac{1}{\lambda}\sum_{ij} \Pi_{ij}\log(\Pi_{ij})} \\
\text{subject to} \quad & \Pi 1 = x \\
& \Pi^T 1 = z \\
& \langle \Pi, C \rangle \leq \epsilon
\end{aligned}$$

- Entropy regularization term
- Local transport plans
- Block coordinate descent on the dual problem

CIFAR10 Wasserstein adversarial examples

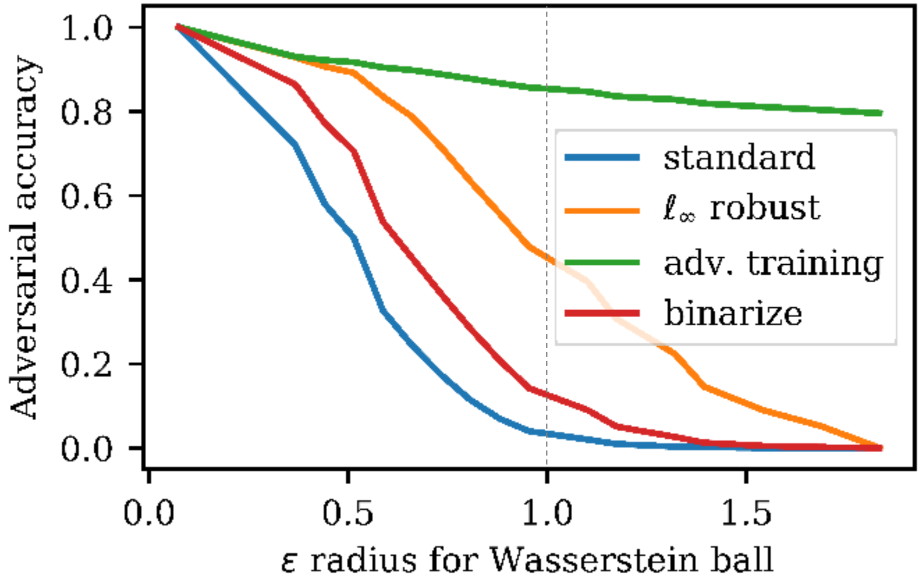


We can also adversarially train robust networks

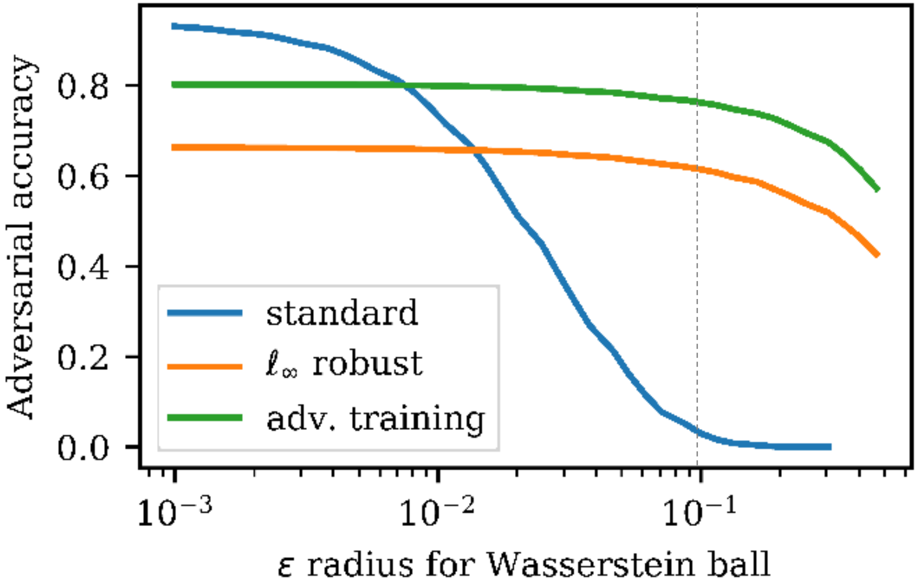


MNIST and CIFAR10 robustness curves

MNIST



CIFAR10



Wasserstein adversarial examples via projected Sinkhorn iterations

Poster #67 in the Pacific Ballroom

Code: https://github.com/locuslab/projected_sinkhorn