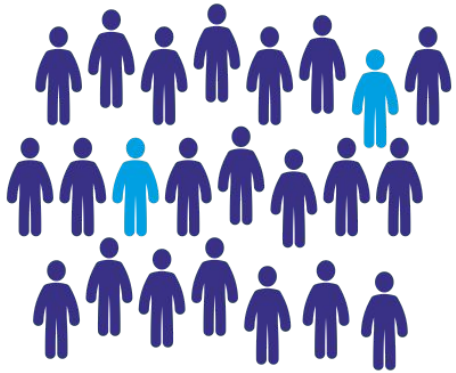
Theoretical Limitations of Encoder-Decoder GAN architectures

Sanjeev Arora, Andrej Risteski, Yi Zhang

Presented by Calvin Xu (cx23)

Birthday Paradox Test

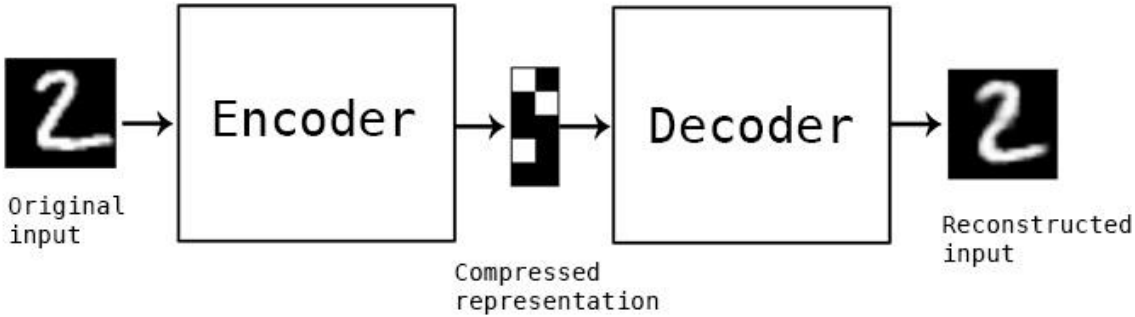
The Birthday Paradox How many people do you need before the probability that two of them share a birthday exceeds 50%?



Birthday Paradox Test (1) Pick a sample of size *s* from the distribution
(2) Measure similarity and flag *k* most similar pairs
(3) Visually inspect pairs for duplicates
(4) Repeat

Question If you find a duplicate, what does that mean?

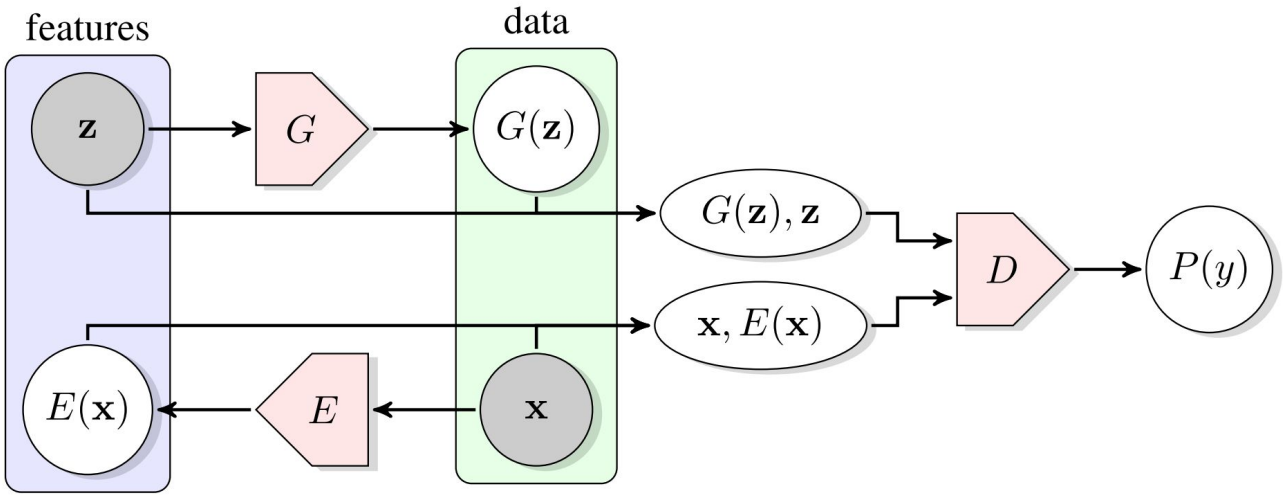
Background - Encoder Decoder Pair



Manifold Assumption High dimensional data (i.e. images) lies on a low dimensional manifold

Semantic Meaning The low dimensional representation is ‘meaningful’, the encoding is a ‘meaningful’ description of the original

Background - Encoder-Decoder GANs



Real Image - *x*

Fake Image - *G(z)*

GAN

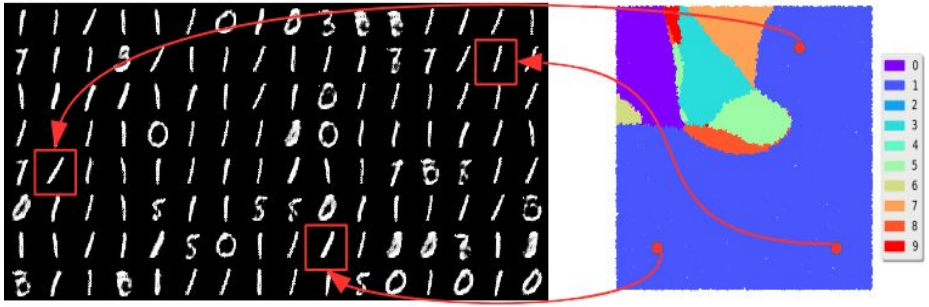
Encoder Output - (*x*, *E(x)*)

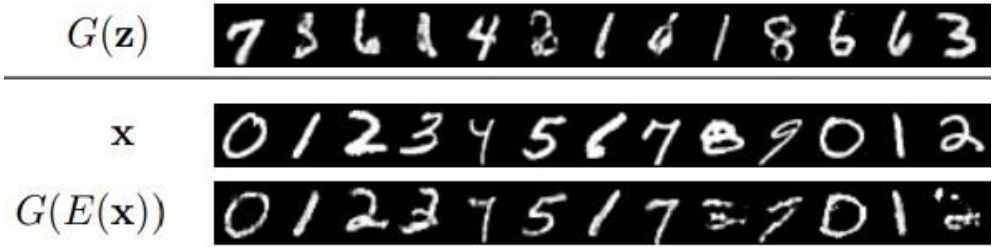
Generator Output - (*G(z)*, *z*)

Encoder-Decoder GAN

Background - Encoder-Decoder GANs

Mode Collapse Generative models which learn to only generate one class





Generator G(z) distributed p(x|z)
Encoder E(x) distributed p(z|x)
(z, G(z)) and (E(x), x) equal to p(z, x)

BiGAN Objective $$\min_{G,E} \max_{D} \left| \mathop{\mathbb{E}}_{x \sim \hat{\mu}} \phi(D(x, E(x))) - \mathop{\mathbb{E}}_{z \sim \hat{\nu}} \phi(D(G(z), z)) \right|$$

Main Theorem

Theorem There exists a generator with small support (far from true data distribution) and an encoder with small complexity s.t. the BiGAN objective can be made arbitrarily small for all discriminators

Question What does this theorem imply? Does this mean BiGAN is unusable?

Theorem 1 (Main). *There exists a generator G of support $\frac{p\Delta^2 \log^2(p\Delta L L_\phi/\epsilon)}{\epsilon^2}$ and an encoder E with at most $\tilde{d}$ non-zero weights, s.t. for all discriminators D that are L-Lipschitz and have capacity less than p, it holds that*

$$|\mathop{\mathbb{E}}_{x\sim\mu} \phi(D(x, E(x))) - \mathop{\mathbb{E}}_{z\sim\nu} \phi(D(G(z), z))| \leq \epsilon$$

Noise Assumption

Noise Assumption Assume images come noised, imagine replacing every 100th pixel with Gaussian noise

Question Should this change the image content?

$$x_i = \begin{cases} z_{\frac{i}{\lfloor \frac{d}{\tilde{d}} \rfloor}}, & \text{if } i \equiv 0(\bmod \lfloor \frac{d}{\tilde{d}} \rfloor) \\ \tilde{x}_i, & \text{otherwise} \end{cases}$$

Natural Images We can extend the proof non-noised images by assuming natural images have innate stochasticity

Original Image

Noised Image

Proof Sketch- Building the Encoder

Encoder Extract noise from noised image, i.e. $E(\tilde{x} \circledast z) = z$

Interpretation The code is just noise, so meaningless
A trivial network of ReLU's can emulate this

Proof Sketch- Building the Generator

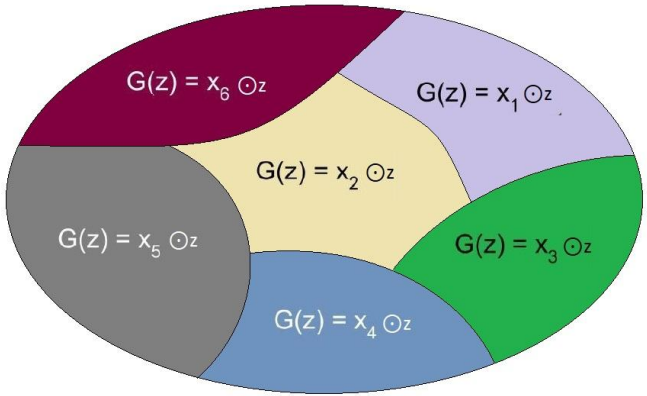
Generator Memorize a hash function that partitions all codes, *z*, into *m* equal sized blocks. Also memorize *m* unnoised images. Then create a mapping, i.e. $G(z) = \tilde{x}_i \circledast z$ where *i* is *z*'s partition

$$m := \frac{p\Delta^2 \log^2(p\Delta L L_\phi/\epsilon)}{\epsilon^2}$$

Interpretation Distribution of generators

Will prove with high probability that one of these satisfies the theorem

ReLU network with *O(md)*



Proof Sketch- Putting it all together

We can observe that the expected encoder matches the expectation of *D(x, E(x))*

$$\mathop{\mathbb{E}}_{x\sim\mu}\phi(D(x,E(x))) = \mathop{\mathbb{E}}_{x\sim\tilde{\mu},z\sim\nu}\phi(D(x\circledast z,z)) = \mathop{\mathbb{E}}_{G}\mathop{\mathbb{E}}_{z\sim\nu}\phi(G(z),z)$$

Want to show a particular *G* works for all discriminators, detailed proof in paper

- *D* is *L*-Lipschitz, bounding the parameters, allowing use of ε-net + union bound
- Reformulate expectation with non-colliding sets, T, that are independently drawn, thus can use McDiarmid's inequality to show concentration around expectation
- This allows us to use Markov's inequality to show that all but exponentially small fraction of encoders make the below equation arbitrarily small

$$\left|\mathop{\mathbb{E}}_{T\sim\mathcal{T}_{nc}}\mathop{\mathbb{E}}_{z\sim T}\phi(D(G(z),z)) - \mathop{\mathbb{E}}_{G}\mathop{\mathbb{E}}_{T\sim\mathcal{T}_{nc}}\mathop{\mathbb{E}}_{z\sim T}\phi(D(G(z),z))\right|$$

Conclusion

Summary Encoder-Decoder training objectives cannot avoid mode collapse and cannot enforce meaningful manifold spaces as we have shown a relatively small finite support generator and white noise encoder that satisfy the training objective

Questions Why do these architectures work in practice?
Do other GAN variations suffer from the same fate?

References & Further Reading

- Theoretical Limitations of Encoder-Decoder GAN architectures
  - Professor Arora and Risteski's Blog Post Summary
- Adversarial Feature Learning (BiGAN)
  - Medium Post Review for BiGAN
- Adversarially Learned Inference (ALI)
- Do gans actually learn the distribution? An empirical study
- Generalization and Equilibrium in Generative Adversarial Nets (GANs)

Certifying Some Distributional Robustness with Principled Adversarial Training

Aman Sinha, Hongseok Namkoong, John Duchi

Presented by Calvin Xu (cx23)

Empirical vs Formal Robust Training

Metric	Empirical	Formal
Standard Accuracy	Loss in Accuracy	Greater Loss in Accuracy
Adversarial Accuracy	Adversarially Robust	Adversarially Robust
Formal Guarantees	Harder to show	Easier to show
Scalability	More Scalable	Less Scalable
Examples	PGD, TRADES, MART, AWP	DiffAI, COLT, CROWN-IBP, L_∞-nets

Question How do you balance efficiency and formal guarantees?

Distributionally Robust Optimization

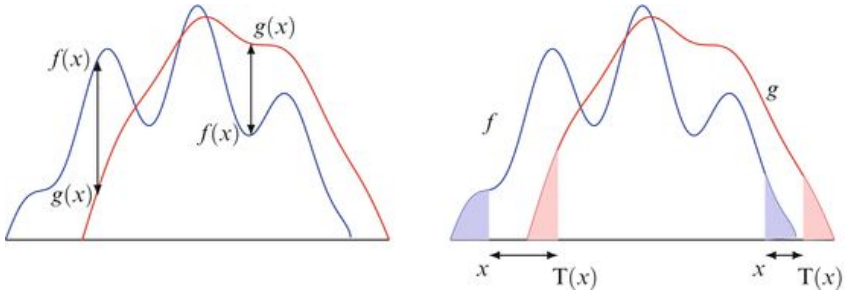
Stochastic $\min \mathbb{E}_{P_0}[l(\theta; Z)]$

DRO $\min_{\theta \in \Theta} \sup_{P \in \mathcal{P}} \mathbb{E}_P[l(\theta; Z)]$

Assumptions

1. $\mathcal{P} = \{P : W_c(P, P_0) \leq \rho\}$
2. ρ defines neighborhood
3. $c(z, z_o) = || z - z_0||_p^2$

Review - Wasserstein Distance



Solving DRO

Question Can we solve this min-max problem?

DRO $\min_{\theta \in \Theta} \sup_{P \in \mathcal{P}} \mathbb{E}_P[l(\theta; Z)]$

Solution

Use Lagrangian relaxation to replace loss with robust surrogate

$$\min_{\theta \in \Theta} \left\{ F(\theta) := \sup_{P \in \mathcal{P}} \{ \mathbb{E}_p[l(\theta; Z)] - \gamma W_c(P, P_0) \} = \mathbb{E}_p[\phi_\gamma(\theta; Z)] \right\}$$

$$\phi_\gamma(\theta; z_0) := \sup_{z \in Z} l(\theta; z) - \gamma c(z, z_0)$$

Key Insight

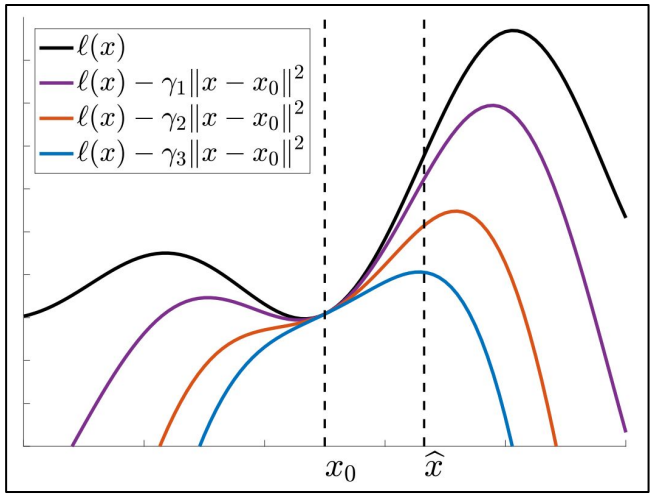
If we choose a large γ and a smooth loss, ϕ is concave and easy to optimize!

Why?

$$\phi_\gamma(\theta; z_0) := \sup_{z \in Z} l(\theta; z) - \gamma c(z, z_0)$$

1. c is convex, so -c is concave
2. Loss is smooth so gradient is L-Lipschitz
3. For large γ, the second term dominates

Thus, we have a surrogate strongly concave optimization problem



Stochastic Gradient Descent for DRO

Repeat:

1. Draw $Z_k \overset{\text{iid}}{\sim} P$
2. Compute (approximate) maximizer

$$\widehat{Z}_k \approx \underset{z}{\operatorname{argmax}} \left\{ \ell(\theta; z) - \frac{\lambda}{2} \|z - Z_k\|_2^2 \right\}$$

3. Update

$$\theta_{k+1} := \theta_k - \alpha_k \nabla_\theta \ell(\theta_k; \widehat{Z}_k)$$

where α_k is a stepsize

Theorem This process converges 'quickly'

Note

Only works if cost function is continuous and strongly convex and if the loss is Lipschitz smooth

$$\Delta_\theta \phi_\gamma(\theta; z_0) = \Delta_\theta l(\theta; z^\star(z_0, \theta))$$

$$z^\star(z_0, \theta) = \arg\max_{z \in Z} \{ l(\theta; z) - \gamma c(z, z_0) \}$$

Robustness Certification

Theorem (Robustness Certificate)

With high probability, for all $\theta \in \Theta$

$$\frac{1}{n}\sum_{i\equiv 1}^{n} \sup_{\Delta}\left\{ l(\theta; Z_i + \Delta) - \frac{\lambda}{2}||\Delta||_2^2 \right\} + \lambda \hat{W}(\theta) \geq \sup_{P: W(P,P_0)\leq \hat{W}(\theta)} \{\mathbb{E}_P[l(\theta; Z)]\} - \frac{O(1)}{\sqrt{(n)}}$$

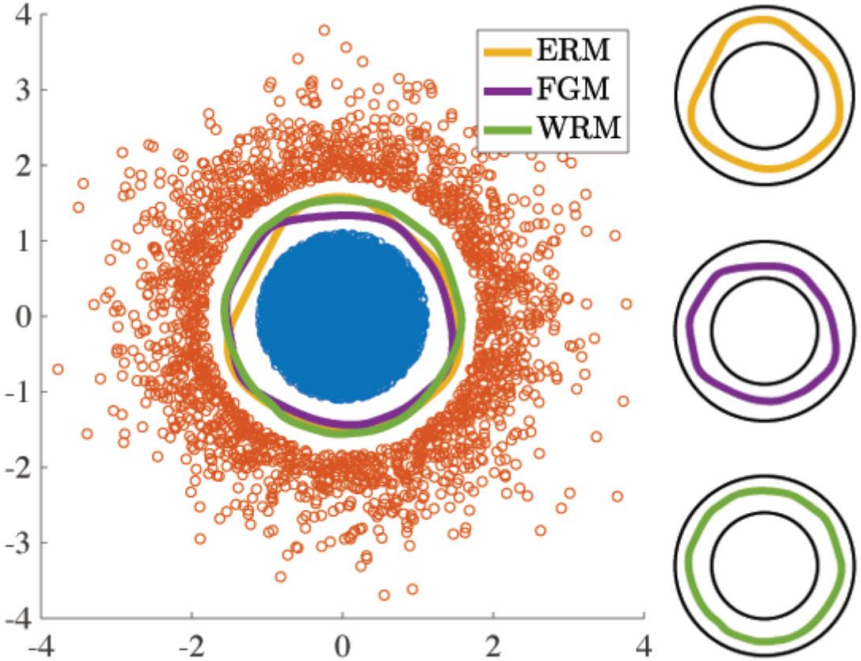
Theorem (Robustness Certificate - Empirical)

Can approximate the divergence by

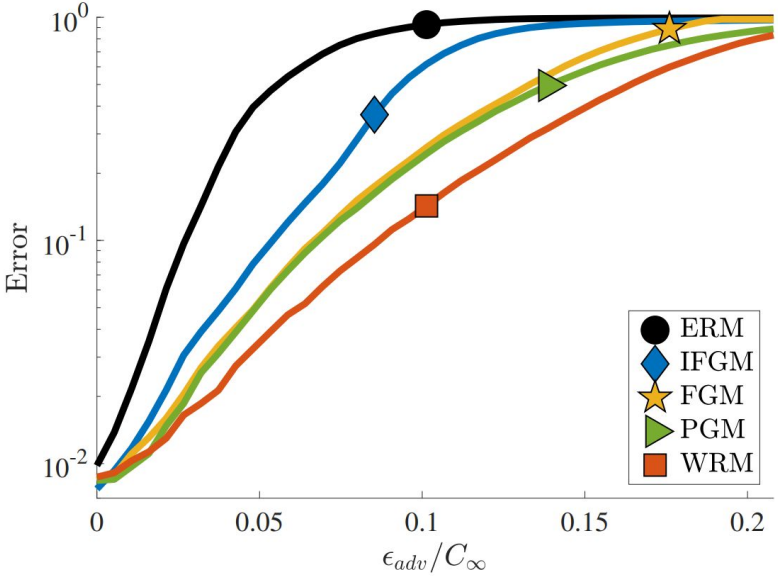
$$\hat{W}(\theta) := \frac{1}{2n}\sum_{i=1}^{n} \left|\left| \hat{Z}_i(\theta) - Z_i(\theta) \right|\right|_2^2 \text{ where } \hat{Z}_i = \arg\max_{z}\left\{ l(\theta; z) - \frac{\lambda}{2}||z - Z_i(\theta)||_2^2 \right\}$$

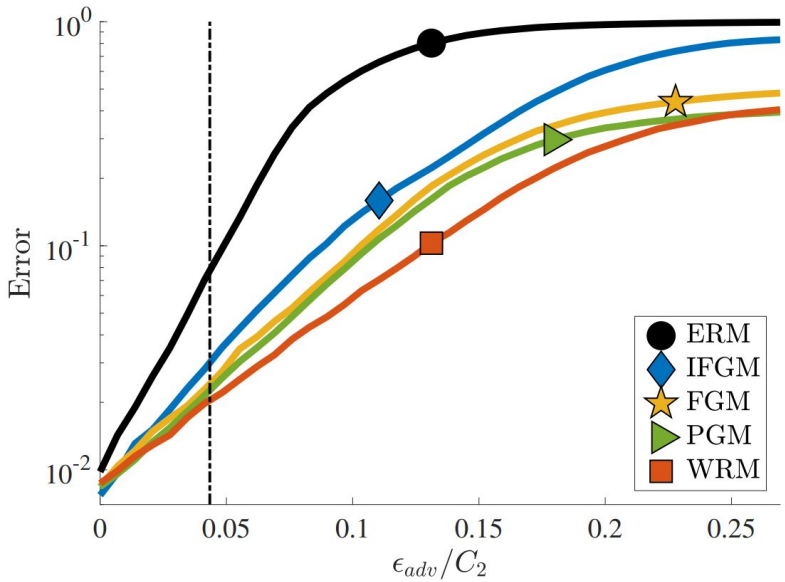
Visualizing Benefits

$$y = \text{sign}(\|x\|_2 - \sqrt{2})$$

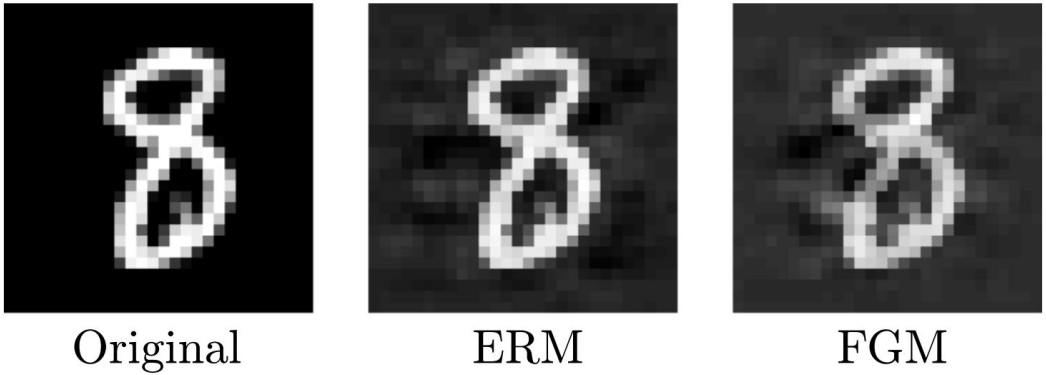


Visualizing Benefits

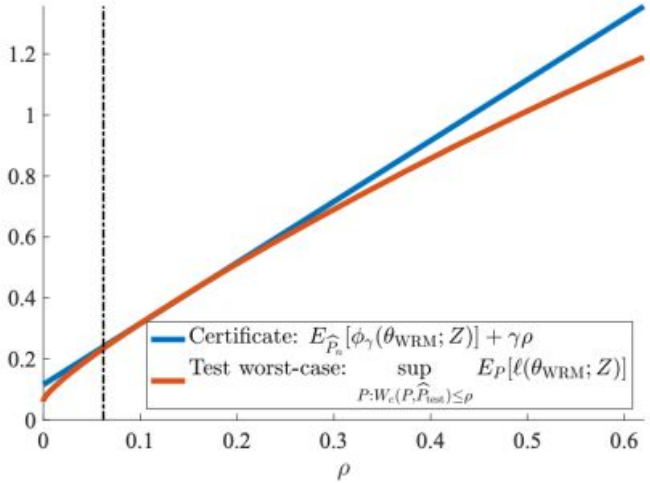




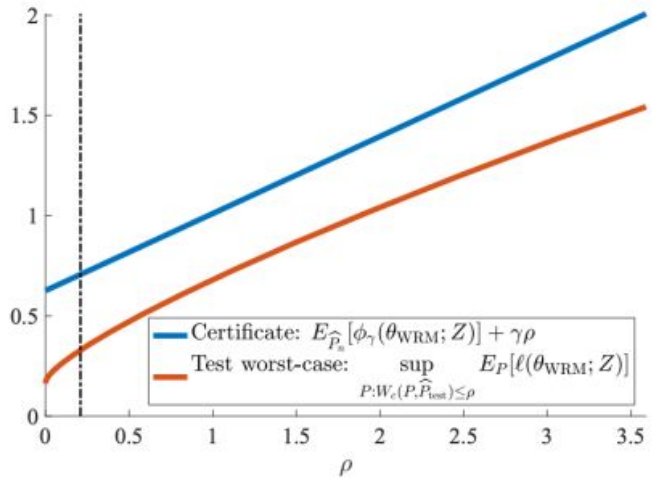
Visualizing Benefits



Worst Case Performance



(a) Synthetic data



(b) MNIST

Conclusion

Summary We can create a surrogate objective function that is no longer NP-hard by using a Lagrangian relaxation .This allows us to efficiently compute the training objective and provides us a way to obtain statistically significant certificates of robustness

Limits
1. This only applies for smooth losses (i.e. no ReLU)
2. Convergence depends on small values of robustness bounds
3. $\hat{\gamma}$ is bounded by a Lipschitz constant so can be expensive for large networks

References & Further Reading

- Certifying Some Distributional Robustness with Principled Adversarial Training
  - Aman Sinha's Slides
  - A blog post summarizing the paper
  - Professor Duchi's Talk and Slides
- Stochastic Gradient Methods for Distributionally Robust Optimization with *f*-Divergences
- Variance-based Regularization with Convex Objectives