Don’t let your Discriminator be fooled
Brady Zhou, Philipp Krähenbühl
brady.zhou@utexas.edu, philkr@cs.utexas.edu

Experiments

Robust discriminator

Typical GAN objective

\[ L(G, D) = \mathbb{E}_{x \sim P_X}[f(-D(x))] + \mathbb{E}_{z \sim P_Z}[f(D(G(z)))] \]

Generator minimizes

\[ \ell(G) = \max_D L(G, D) \]

A more robust discriminator objective

\[ \ell_r(G) = \max_D L(D, G) - \lambda \rho(D, G) \]

Adversarial Regularization (AR)

\[ \rho(D, G) = \mathbb{E}_{z \sim P_Z} \left[ \max_{\|v\|_p < \delta} \left( D(D(z)) - D(G(z) + v) \right)^2 \right] \]

Robust Feature Matching (RFM)

\[ \rho_r(D, G) = \mathbb{E}_{z \sim P_Z} \left[ \max_{\|v\|_p < \delta} \| \phi(G(z)) - \phi(G(z) + v) \|_2^2 \right] \]

If D is robust,

\[ E_{z \sim P_Z}[D(G(z)) - D(G(z) + u(z))] < \varepsilon \]

the GAN objective is robust

\[ |\ell(G) - \ell(G + u)| < \varepsilon C \]

A toy example (Dirac-GAN)

\[ P_R = \delta_0 \quad D_R(x) = \psi \cdot x \]

\[ P_G = \delta_\theta \quad L(\theta, \psi) = f(\psi(0)) + f(0) \]

Consistency across hyperparameters

Randomly vary:
- loss
- normalization
- nonlinearity
- channel width
- batch size
- learning rate

and measure best performing method