OVERVIEW

- MMD GANs are related to WGANs, but with part of critic function optimization done in closed form.
- Outperform WGAN-GP, especially with smaller critic network.
- Clarify gradient bias situation: “outer loop” generator gradients are biased, but each step is unbiased.
- New GAN performance metric, KID, with better estimator than FID; use it to adapt the learning rate during training.

RELATION TO WASSERSTEIN AND CRAMÉR GANS

Integral Probability Metrics (IPMs) are distances between distributions defined by a class of critic functions \( F \):

\[
D(P, Q) = \sup_{f \in F} \mathbb{E} \left[ f(X) - f(Y) \right],
\]

- Wasserstein distance has \( F \) the set of 1-Lipschitz functions

\[
F = \left\{ f : \sup_{x, y} \frac{|f(x) - f(y)|}{\|x - y\|} \leq 1 \right\}.
\]

WGANs approximate \( f \) with a critic network, made approximately Lipschitz with weight clipping [1] or gradient penalty [4].

- Maximum Mean Discrepancy (MMD) has \( F \) a unit ball in a Reproducing Kernel Hilbert Space (RKHS) \( H \) with kernel \( k \):

\[
f(f) = \mathbb{E} \left[ k(X, t) - k(Y, t) \right],
\]

\[
\text{MMD}^2_{\delta} = \mathbb{E} \left[ \mathbb{E} \left[ f(X) - f(Y) \right] \right].
\]

- MMD GANs [6] optimize representation in kernel

\[
k_{\theta}(x, y) = k_{\text{base}}(h_\theta(x), h_\theta(y)),
\]

corresponding to distance

\[
D(P, Q) = \sup_{\theta} D_\theta(P, Q) = \sup_{\theta} \text{MMD}^2_{\delta}(P, Q).
\]

- Cramér GAN [2] almost same, with Energy Distance \( k_{\text{base}} \).

MMD GAN WITH GRADIENT PENALTY

Like WGAN-GPs [4], we penalize gradient of the critic function:

\[
\text{Loss}^{\text{critic}}(\theta) = \text{MMD}^2_{\delta}(P, \mathbb{E} X \sim P) + \lambda \mathbb{E} \left[ \|\nabla_X f(X)\| - 1\right]^2.
\]

With \( k_{\text{base}} \), almost the same as a WGAN-GP.