Informative Features for Model Comparison

Wittawat Jitkrittum[1] Heishiro Kanagawa[2] Patsorn Sangkloy[3] James Hays[3] Bernhard Schölkopf[1] Arthur Gretton[2]

[1]Max Planck Institute for Intelligent Systems [2]Gatsby Unit, University College London [3]Georgia Institute of Technology

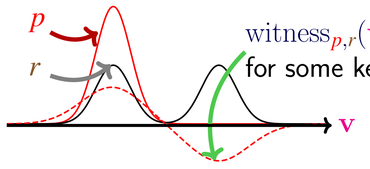
When comparing complex generative models in high dimensions, the question to ask is not "which model is correct" (neither), or "which model is better," but rather **"where does each model do better than the other?"**

- **Given**: Two candidate models p, q, a sample $\{\mathbf{z}_i\}_{i=1}^n$ from an unknown distribution r.
- **Do**: Test H_0: p fits sample better **vs** H_1: q fits sample better.
- **Propose**: Two new model comparison tests:
  1. Rel-UME: Represent p, q by i.i.d. samples $\{\mathbf{x}_i\}_{i=1}^n$, $\{\mathbf{y}_i\}_{i=1}^n$.
  2. Rel-FSSD: p, q are unnormalized probability densities.
- **Advantages**:
  1. Nonparametric: mild assumptions on p, q. Domain $\mathcal{X} \subseteq \mathbb{R}^d$.
  2. Linear-time: $\mathcal{O}(n)$ runtime complexity. Fast. ☺
  3. **Informative**: show where q fits better than p (or vice versa) with a set of points (**features**).
- Test power matches that of the state-of-the-art quadratic-time relative MMD test [Bounliphone et al., 2015].

Test Statistics

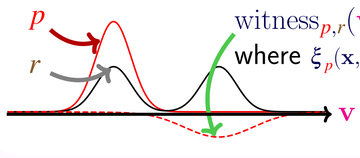
- Let $D_V(p, r)$:= distance between p, r measured at $V = \{\mathbf{v}_1, \dots, \mathbf{v}_J\}$ (**features**). D_V can be UME or FSSD.
- Statistic $S := D_V(p, r) - D_V(q, r) \implies H_0: S \leq 0$ vs $H_1: S > 0$.
- $D_V(p, r)$ = average evaluation of the squared **witness function** $= \frac{1}{J}\sum_{j=1}^J \text{witness}_{p,r}^2(\mathbf{v}_j)$.

UME (Unnormalized Mean Embeddings) [Jitkrittum et al., 2016]



$\text{witness}_{p,r}(\mathbf{v}) = \mathbb{E}_{\mathbf{x}\sim p} k(\mathbf{x}, \mathbf{v}) - \mathbb{E}_{\mathbf{z}\sim r} k(\mathbf{z}, \mathbf{v})$ for some kernel $k(\mathbf{x}, \mathbf{v})$.

FSSD (Finite-Set Stein Discrepancy) [Jitkrittum et al., 2017]



$\text{witness}_{p,r}(\mathbf{v}) = \mathbb{E}_{\mathbf{x}\sim p}\xi_p(\mathbf{x}, \mathbf{v}) - \mathbb{E}_{\mathbf{z}\sim r}\xi_p(\mathbf{z}, \mathbf{v})$ where $\xi_p(\mathbf{x}, \mathbf{v}) := k(\mathbf{x}, \mathbf{v})\nabla_{\mathbf{x}} \log p(\mathbf{x}) + \nabla_{\mathbf{x}} k(\mathbf{x}, \mathbf{v})$

(Normalizer of p not needed.)

Relative Goodness-of-Fit Testing

Proposition. *Empirical statistic $S = S(V)$ for both* Rel-UME *and* Rel-FSSD *follows a normal distribution as $n \to \infty$.*

- Estimate S. Reject H_0 if $S > \text{threshold} = (1-\alpha)$-quantile of the normal. False rejection rate $< \alpha$ (asymptotically).
- Reject $H_0 \implies q$ is closer to r as measured at V.

Informative Features $= V$ which maximizes

Test Power $= \mathbb{P}(\text{detect better fit of } q \mid q \text{ is better})$.

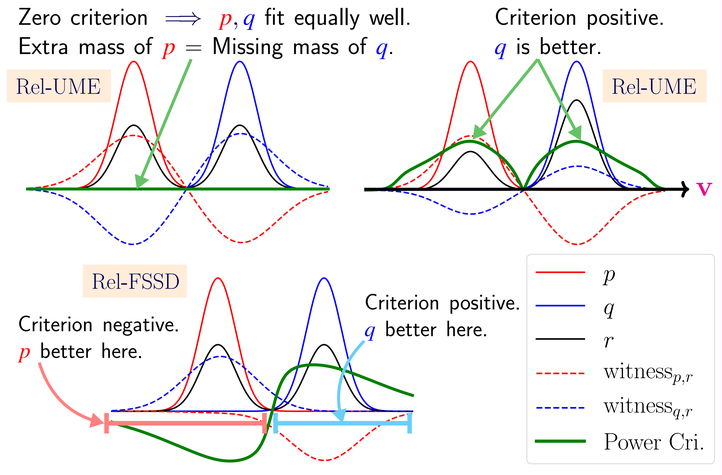
Equivalently, find V which maximizes the **power criterion**:

$$\text{Power Criterion}(V) := \frac{S(V)}{\text{uncertainty}(V)} = \left(\frac{\text{signal}}{\text{noise}}\right),$$

where $\text{uncertainty}(V)$ = variance of $S(V)$ under H_1.

- $\mathcal{O}(n)$ complexity to evaluate power criterion. Fast.

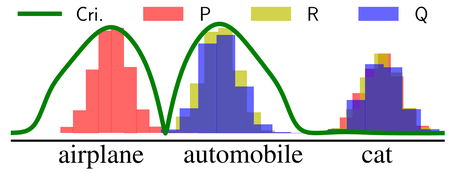
Rel-UME and Rel-FSSD Power Criteria

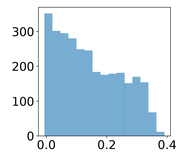


- Rel-UME: better model produces mass closer to the test sample from r.
- Rel-FSSD: better model has shape (given by $\nabla_{\mathbf{x}} \log p(\mathbf{x})$ and $\nabla_{\mathbf{y}} \log q(\mathbf{y})$) closer to r.

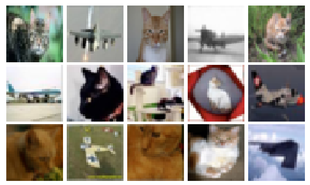
Informative Features (CIFAR-10)

$\{\mathbf{x}_i\}_{i=1}^n = \{\text{airplane, cat}\}$,
$\{\mathbf{y}_i\}_{i=1}^n = \{\text{automobile, cat}\}$,
$\{\mathbf{z}_i\}_{i=1}^n = \{\text{automobile, cat}\}$.

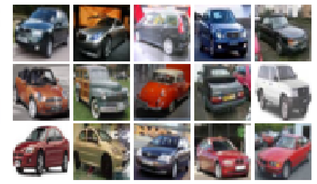




Power Criterion



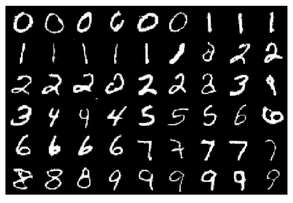
Close to 0



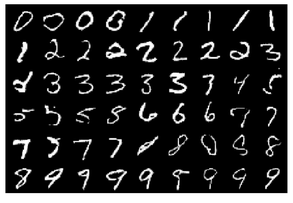
Large positive value

- Evaluate Power Criterion$(\{\mathbf{v}\})$ where $\mathbf{v}$ = image in each category.
- Power Criterion$(\{\mathbf{v}\}) > 0$ everywhere $\implies q$ is better.
- q is better at generating automobile images.
- Criterion$(\{\mathbf{v}\}) \approx 0 \implies p, q$ equally good at generating cat images.

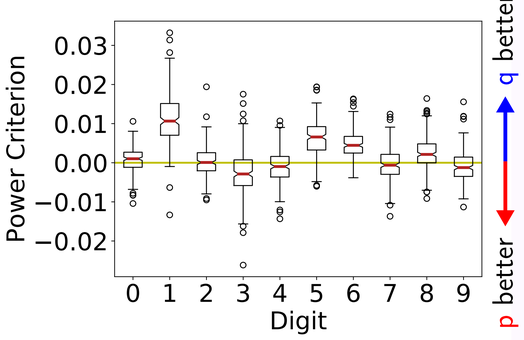
Where Does Each GAN Do Better?



q = LSGAN



p = GAN



- q = LSGAN [Mao et al., 2017]
- p = GAN [Goodfellow et al., 2014]
- Trained for 40 epochs. Evaluate power criterion with $n = 2000$.
- Set $V = 40$ (real) images of digit $i = 0, \dots, 9$.
- q is better at "1" and "5". p is slightly better at "3". **Interpretable**.

Contact: wittawat@tuebingen.mpg.de **Code:** github.com/wittawatj/kernel-mod