Generative Models and Model Criticism via Optimized Maximum Mean Discrepancy

Dougal J. Sutherland[1,2] Hsiao-Yu Tung[2] Heiko Strathmann[1] Soumyajit De[1] Aaditya Ramdas[3] Alex Smola[2] Arthur Gretton[1]

[1]Gatsby Computational Neuroscience Unit, University College London

[2]School of Computer Science, Carnegie Mellon University

[3]Departments of EECS and Statistics, University of California at Berkeley

ICLR, 2017/ Presenter: Anant Kharkar

Outline

Outline

Motivation

- Generative models produce data - attempt to match ground truth distribution
    - How can we distinguish real from generated data?
    - How to train generator network?
- Proposal: use Maximum Mean Discrepancy (MMD) to distinguish distributions

Outline

Divergences

- Comparison of two probability distributions
- Kullback-Leibler (KL)

$$D_{KL}(P||Q) = -\sum_i P(i) log \frac{Q(i)}{P(i)}$$

- Jensen-Shannon

$$JSD(P||Q) = \frac{1}{2}KL(P||M) + \frac{1}{2}KL(Q||M)$$

$$M = \frac{1}{2}(P + Q)$$

- Integral probability metrics
  - Witness function distinguishes P from Q
  - Includes MMD

Outline

- Definition

$$MMD_k^2(P,Q) = \mathbb{E}_{x,x'}[k(x,x')] + \mathbb{E}_{y,y'}[k(y,y')] - 2\mathbb{E}_{x,y}[k(x,y)]$$

- Estimate

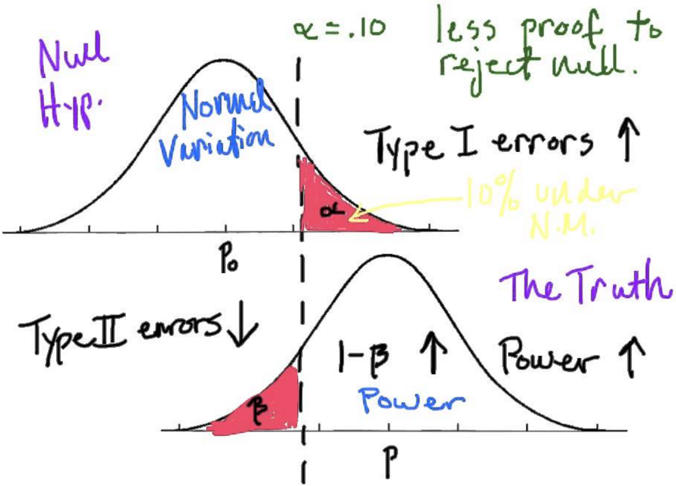
$$\widehat{MMD_U}^2(P,Q) = \frac{1}{\binom{m}{2}} \sum_{j \neq j'} k(X_i, X_i') + \frac{1}{\binom{m}{2}} \sum_{i \neq i'} k(Y_j, Y_j')$$

$$-\frac{2}{\binom{m}{2}} \sum_{i \neq j} k(X_i, Y_j)$$

Outline

Test Power

- Power: probability of rejecting H_0 given that H_A is true
- Measure of effectiveness of hypothesis



Hypothesis Test

- $H_0 : P = Q$
- $H_1 : P \neq Q$

$$\frac{\widehat{\mathrm{MMD}}_{\mathrm{U}}^2(X,Y) - \mathrm{MMD}^2(P,Q)}{\sqrt{V_m(P,Q)}} \xrightarrow{D} \mathcal{N}(0,1)$$

$$\mathrm{Pr}_1\left(m\,\widehat{\mathrm{MMD}}_{\mathrm{U}}^2(X,Y) > \hat{c}_\alpha\right) = \mathrm{Pr}_1\left(\frac{\widehat{\mathrm{MMD}}_{\mathrm{U}}^2(X,Y) - \mathrm{MMD}^2(P,Q)}{\sqrt{V_m(P,Q)}} > \frac{\hat{c}_\alpha/m - \mathrm{MMD}^2(P,Q)}{\sqrt{V_m(P,Q)}}\right)$$
$$\to \Phi\left(\frac{\mathrm{MMD}^2(P,Q)}{\sqrt{V_m(P,Q)}} - \frac{c_\alpha}{m\sqrt{V_m(P,Q)}}\right) \quad (4)$$

t-statistic

- $H_0 : P = Q, H_1 : P \neq Q$

$$t_k(P,Q) := \mathrm{MMD}_k^2(P,Q) / \sqrt{V_m^{(k)}(P,Q)}$$

t-statistic

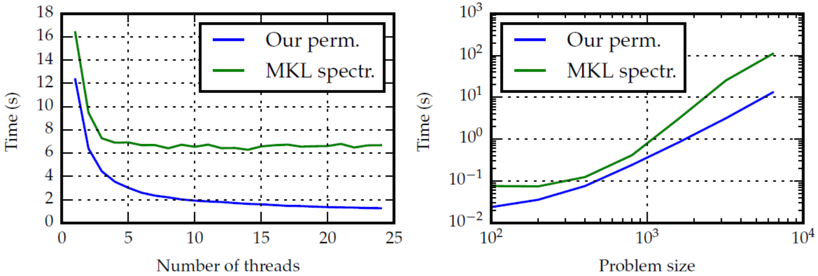
- $H_0 : P = Q, H_1 : P \neq Q$

$$t_k(P,Q) \; := \; \mathrm{MMD}_k^2(P,Q)/\sqrt{V_m^{(k)}(P,Q)}$$

$$\begin{aligned}
\widehat{V}_m :=& \frac{2}{m^2(m-1)^2}\left(2\|\tilde{K}_{XX}e\|^2 - \|\tilde{K}_{XX}\|_F^2 + 2\|\tilde{K}_{YY}e\|^2 - \|\tilde{K}_{YY}\|_F^2\right) \\
&- \frac{4m-6}{m^3(m-1)^3}\left[\left(e^\mathsf{T}\tilde{K}_{XX}e\right)^2 + \left(e^\mathsf{T}\tilde{K}_{YY}e\right)^2\right] + \frac{4(m-2)}{m^3(m-1)^2}\left(\|K_{XY}e\|^2 + \|K_{XY}^\mathsf{T}e\|^2\right) \\
&- \frac{4(m-3)}{m^3(m-1)^2}\|K_{XY}\|_F^2 - \frac{8m-12}{m^5(m-1)}\left(e^\mathsf{T}K_{XY}e\right)^2 \qquad (5) \\
&+ \frac{8}{m^3(m-1)}\left(\frac{1}{m}\left(e^\mathsf{T}\tilde{K}_{XX}e + e^\mathsf{T}\tilde{K}_{YY}e\right)\left(e^\mathsf{T}K_{XY}e\right) - e^\mathsf{T}\tilde{K}_{XX}K_{XY}e - e^\mathsf{T}\tilde{K}_{YY}K_{XY}^\mathsf{T}e\right).
\end{aligned}$$

Optimization

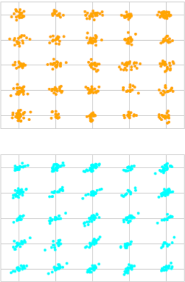
- CPU cache optimization, multithreading
- Improved performance over Intel MKL spectral solver



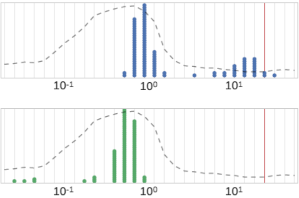
Outline

Synthetic Data

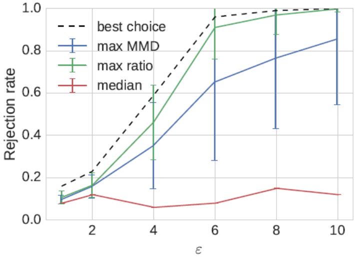
- Bandwidth selection for Gaussian RBF kernels - Blobs dataset



(a) Samples of size 500 with $\varepsilon = 6$ from P (top) and Q (bottom).



(b) Bandwidths chosen by maximizing $\widehat{\mathrm{MMD}}_U^2$ (top, blue) and $\hat{t}$ (bottom, green), as well as the median heuristic (red), for $\varepsilon = 6$. Gray lines show the power of each bandwidth: $\sigma = 0.67$ had power 96%, $\sigma = 10$ had 10%.



(c) Mean and standard deviation of rejection rate as ε increases. "Best choice" shows the mean power of the bandwidth with highest rejection rate for each problem.

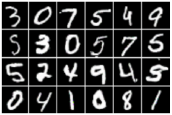
- Optimizing $\hat{t}$ better than optimizing MMD

Outline

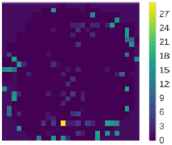
Model Criticism

- Automatic relevance determination (ARD) kernel - MNIST

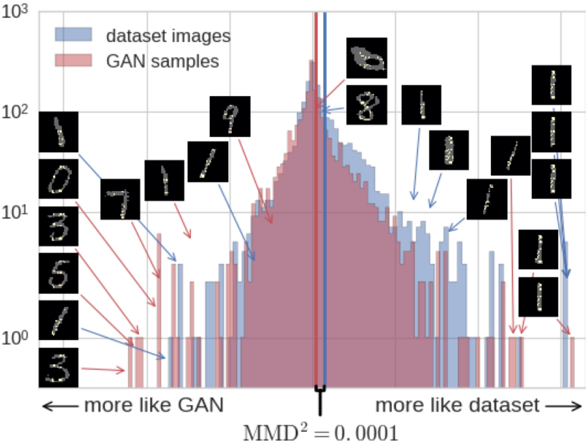
(a) Dataset samples.



(b) GAN samples.



(c) ARD weights.

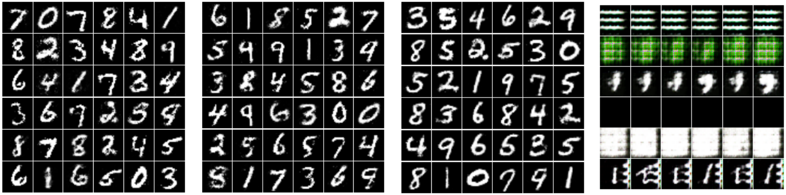


(d) The witness function $\mathbb{E}_{x \sim P}\, k(x, \cdot) - \mathbb{E}_{y \sim Q}\, k(y, \cdot)$ evaluated on a test set. Images are shown with ARD weights highlighted. Vertical lines show distribution means; the distance between them is small but highly consistent.

Outline

GAN

- Generative moment matching network (GMMN) uses MMD
- *t*-GMMN - minimizes *t*-statistic



(a) GMMN (b) *t*-GMMN (c) Feature matching (d) Improved GAN

Figure 4: MNIST digits from various models. Part d shows six runs of the minibatch discrimination model of Salimans et al. (2016), trained without labels — the same model that, with labels, generated Figure 3b. (The third row is the closest we got the model to generating digits without any labels.)

Summary

- MMD is a divergence metric
- Constructed MMD optimization t-test
- Propose MMD t-test as tool for GANs
  - Model criticism
  - GAN optimizer