Generative Models and Model Criticism via Optimized Maximum Mean Discrepancy

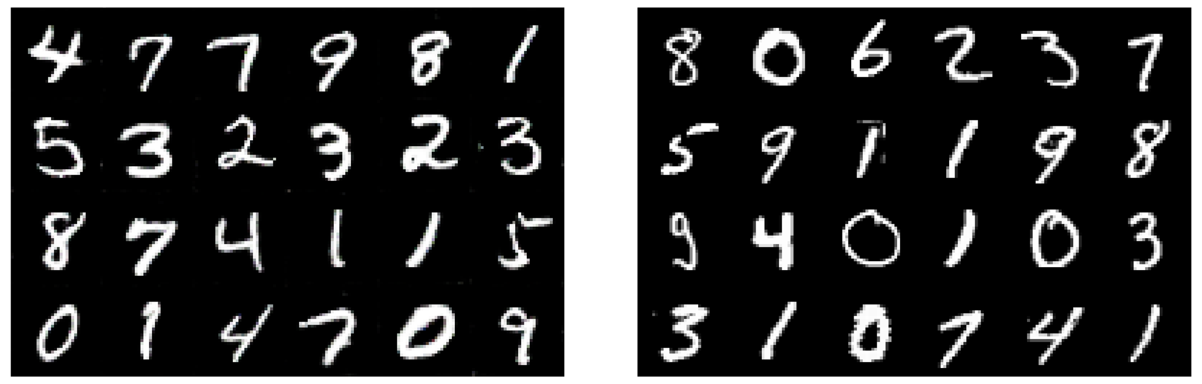
Dougal J. Sutherland, Hsiao-Yu Tung, Heiko Strathmann, Soumyajit De, Aaditya Ramdas, Alex Smola, Arthur Gretton

dougal@gmail.com; github.com/dougalsutherland/opt-mmd

GATSBY · Berkeley University of California · Carnegie Mellon University

Two-sample testing

Say we observe two different datasets:



$X \sim \mathbb{P}$ (model of [6]) $\quad Y \sim \mathbb{Q}$ (MNIST samples)

Our question: is $\mathbb{P} = \mathbb{Q}$?

- Did my generative model actually learn the distribution I wanted it to?
- Do smokers and non-smokers have different distributions of cancers?
- Do these neurons fire differently when the subject is looking at image A instead of B?
- Are these different data sources the same?

We want

- to be able to detect any possible difference,
- without making parametric assumptions,
- on high-dimensional data.

Maximum mean discrepancy

Distance between distributions [2] based on a kernel on sample points $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$:

$$\mathrm{MMD}^2(\mathbb{P}, \mathbb{Q}) = -2\,\mathbb{E}_{X\sim\mathbb{P}, Y\sim\mathbb{Q}}[k(X,Y)] + \mathbb{E}_{X,X'\sim\mathbb{P}}[k(X,X')] + \mathbb{E}_{Y,Y'\sim\mathbb{Q}}[k(Y,Y')].$$

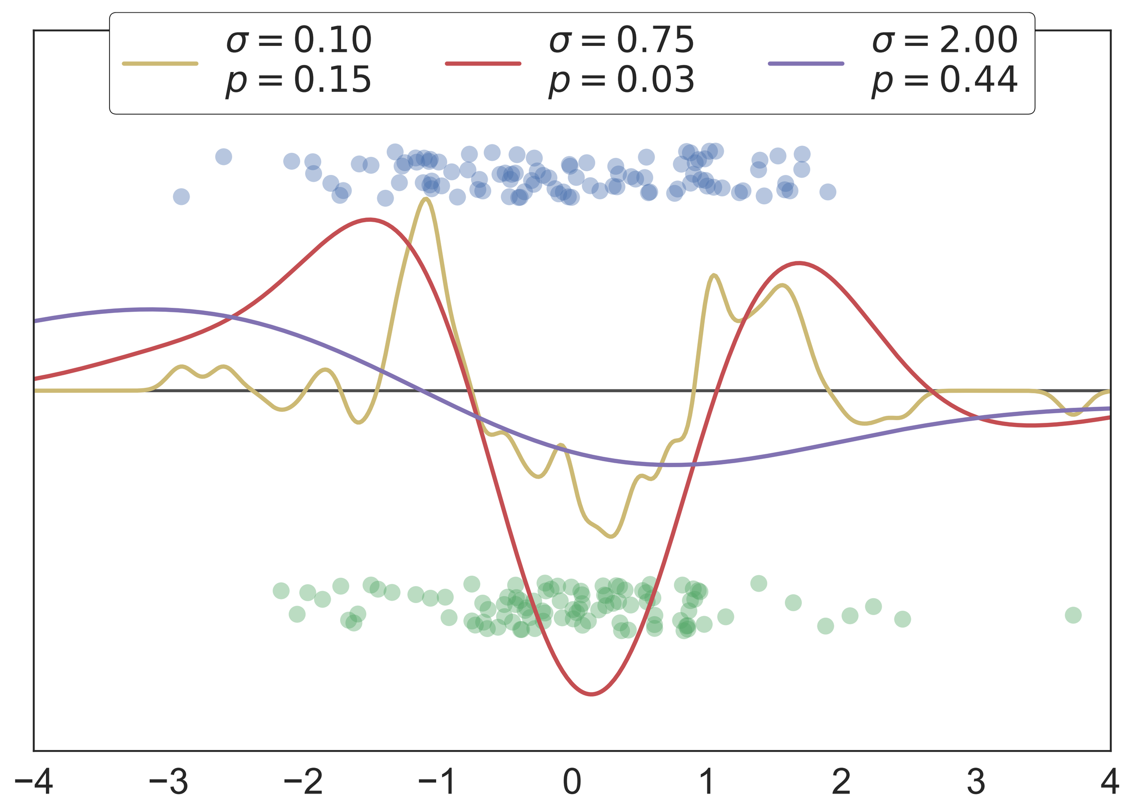
Estimate the MMD by taking sample means.

MMD tests

- Estimate $\mathrm{MMD}(\mathbb{P},\mathbb{Q})$ with $\widehat{\mathrm{MMD}}(X,Y)$.
- Estimate a threshold $\hat{c}_\alpha$:
  - shuffle up $X \cup Y$ into random halves many times;
  - take $\hat{c}_\alpha$ as the $1-\alpha$th quantile of the $\widehat{\mathrm{MMD}}^2$s.
- Say $\mathbb{P} \neq \mathbb{Q}$ if $m\,\widehat{\mathrm{MMD}}^2(X,Y) > \hat{c}_\alpha$.

Kernel choice matters!

- Test blue $\mathbb{P}$ versus green $\mathbb{Q}$ with Gaussian kernels.
- *Witness function* $\mathbb{E}_{X\sim\mathbb{P}}[k(X,\cdot)] - \mathbb{E}_{Y\sim\mathbb{Q}}[k(Y,\cdot)]$ shows which locations more indicative of $\mathbb{P}$ (top) or of $\mathbb{Q}$ (bottom).



- Too small bandwidth: overfits to minor variation.
- Too wide a bandwidth: not confident enough.

Optimizing MMD test power

- When $\mathbb{P} \neq \mathbb{Q}$, $\widehat{\mathrm{MMD}}^2$ is asymptotically normal:

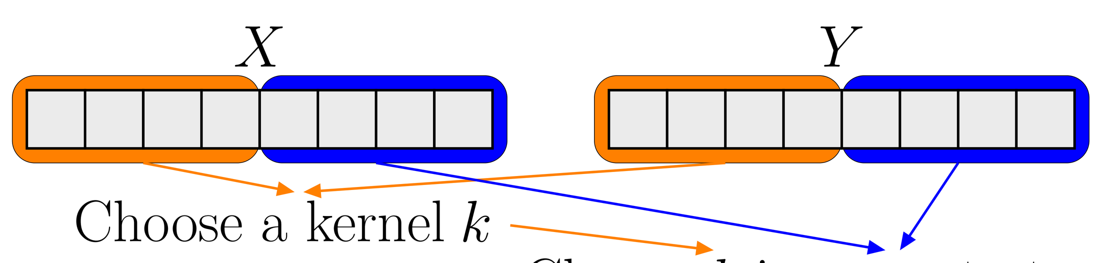
$$\frac{\widehat{\mathrm{MMD}}^2(X,Y) - \mathrm{MMD}^2(\mathbb{P},\mathbb{Q})}{\sqrt{V_m(\mathbb{P},\mathbb{Q})}} \xrightarrow{D} \mathcal{N}(0,1).$$

- Then test power $\Pr\left(m\,\widehat{\mathrm{MMD}}^2(X,Y) > \hat{c}_\alpha\right)$ goes to

$$\Phi\left(\frac{\mathrm{MMD}^2(\mathbb{P},\mathbb{Q})}{\sqrt{V_m(\mathbb{P},\mathbb{Q})}} - \frac{c_\alpha}{m\sqrt{V_m(\mathbb{P},\mathbb{Q})}}\right).$$

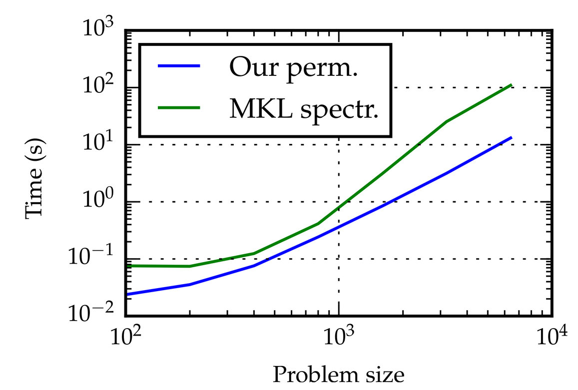
- $V_m = O(m^{-1})$; MMD, c_α are constant in m.
- So, maximize $\hat{t} = \widehat{\mathrm{MMD}}^2(X,Y) / \sqrt{\hat{V}_m(X,Y)}$.
- $\hat{V}_m$: quadratic-time, unbiased estimator of V_m.
- Maximize kernel parameters with backprop.

Train-test splits



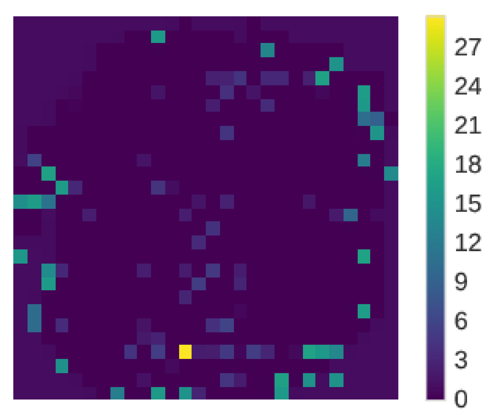
Efficient permutation tests

- Current ways to compute permutations very slow.
- Inefficient memory access pattern.
- Wrote a cache-aware implementation in Shogun.
- 15-30x the speed of existing implementations.
- Faster, more scalable than spectral approximations.

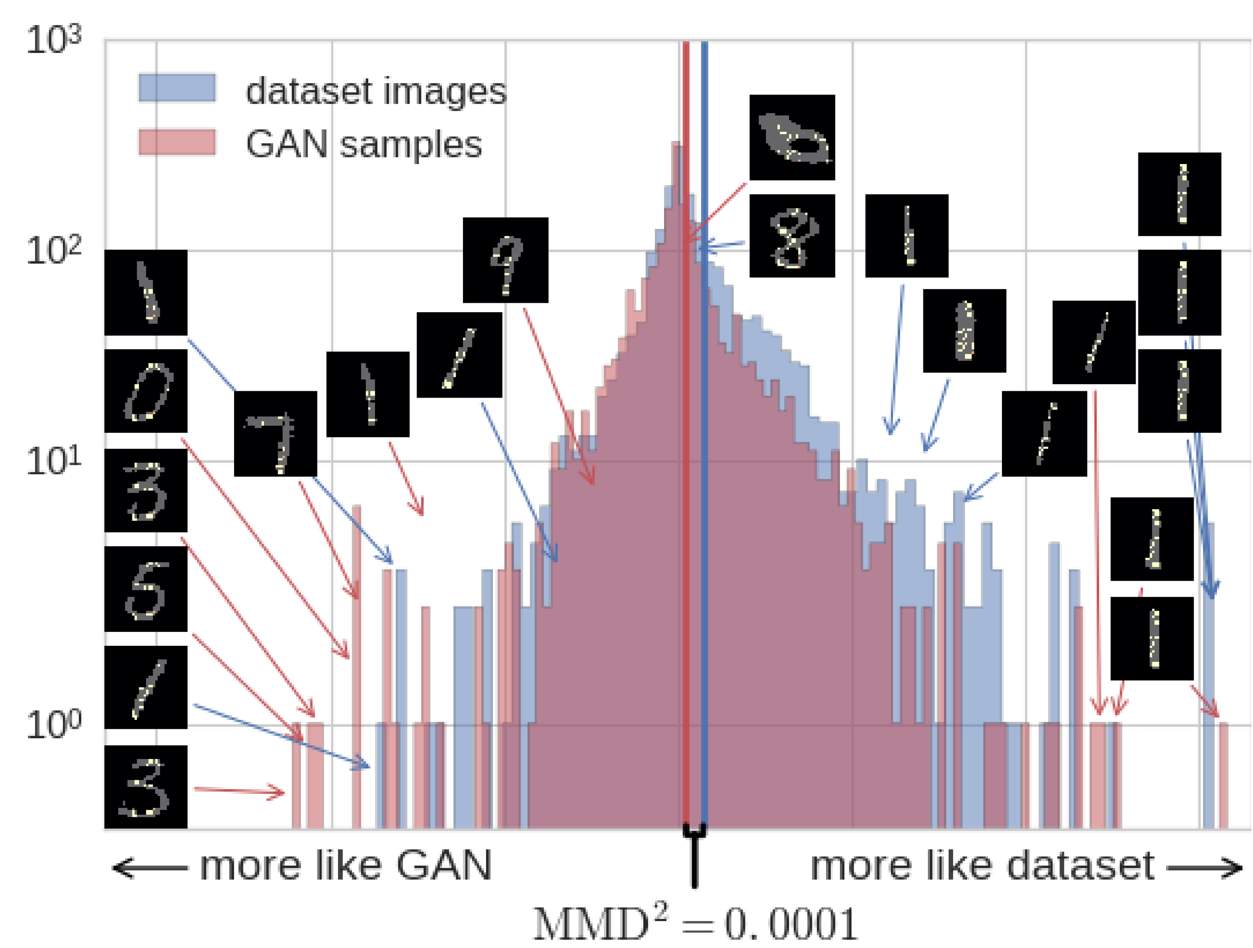


Model criticism

- [6]'s MNIST GAN is really good (top-left).
- Can we tell the distributions apart? Yes!
- Gaussian-ARD kernel: p-values almost exactly 0.
- Pixel weights (right) show where the model's distribution differs.

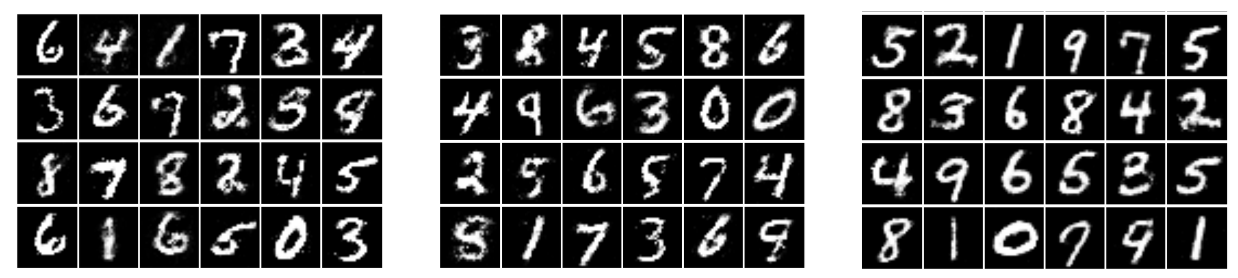


- Just optimizing bandwidth: 57% power at $\alpha = .01$.
- Median heuristic: 42% power.
- Looking at points with high/low witness function values from ARD kernel (like [5]) gives more insight:
  - Model underproduces vertical 1s,
  - Overproduces right-slanted digits.
- MMD value very small, but very consistent.



As a GAN objective

- Discriminators in standard GANs [3] look at one sample at a time.
- Problem: generator incentivized to produce just one sample that the discriminator likes, then gets stuck.
- Generator distribution should match true one.
- Use a two-sample test as the discriminator!
- [1, 4]: doing this by maximizing the MMD.
  - Generative Moment Matching Network (GMMN)
- Instead, optimize $\hat{t}$ criterion (t-GMMN).
- Or, do distributional feature matching (like [6]):
  - Train discriminator normally.
  - Generator uses $\hat{t}$ with kernel from discriminator.



(a) GMMN (b) t-GMMN (c) Feature match

- Used sum of Gaussian kernels:
  - Not a great kernel on MNIST pixels.
  - Nearly useless on natural image pixels.
  - Gradients decay too fast.
- We're trying out better kernels.

References

[1] Dziugaite, Roy, and Ghahramani. Training generative neural networks via Maximum Mean Discrepancy optimization. UAI 2015.

[2] Gretton, Borgwardt, Rasch, Schölkopf, and Smola. A kernel two-sample test. JMLR 2012.

[3] Goodfellow, Pouget-Abadie, Mirza, Xu, Warde-Farley, Ozair, Bengio, Generative Adversarial Nets. NIPS 2014.

[4] Li, Swersky, and Zemel. Generative moment matching networks. UAI 2105.

[5] Lloyd and Ghahramani. Statistical model criticism using kernel two sample tests. NIPS 2015.

[6] Salimans, Goodfellow, Zaremba, Cheung, Radford, and Chen. Improved techniques for training GANs. NIPS 2016.