Regularized f-Divergence Kernel Tests

Google Research DeepMind

UNIVERSITY OF CAMBRIDGE

Mónica Ribero*, Antonin Schrab*, Arthur Gretton (mribero@google.com)

*Authors contributed equally to this work.

A New Unified Framework

The problem: two sample testing

- Given samples X1, X2,..., Xn~ P , Y1, Y2,..., Yn~ Q

  Is there evidence that P≠Q ?

- In practice... Pick your favorite divergence or distance measure D,

  Is there evidence that D(P, Q) > 0 ?

- Traditional MMD two-sample tests
  - often fail to detect specific localized differences and,
  - are not suitable for privacy or unlearning applications, that allow a small difference parameterized by a budget.

Variational Representation

- We constructs practical kernel-based two-sample tests from the family of f-divergences using a regularized variational representation.
- Given distributions P and Q, the f-divergence is:

$$D_f(P \parallel Q) = \mathbb{E}_P[g^*(X)] - \mathbb{E}_Q[f^*(g^*(Y))]$$
$$= \mathbb{E}_P\left[f'\left(\frac{dP}{dQ}(X)\right)\right] - \mathbb{E}_Q\left[f^*\left(f'\left(\frac{dP}{dQ}(Y)\right)\right)\right].$$

- The witness function g^* is estimated using kernel methods and regularization.
- **Special cases:**
  1. The Hockey stick divergence opens practical applications such Privacy and Unlearning.

$$g^*_{HS}(x) = \mathbb{1}\left(\frac{dP}{dQ}(x) \geq \gamma\right)$$

  2. The Chi-squared (DrMMD) divergence is more sensitive for localized differences.

Methods & Guarantees

Theoretical Guarantees

- **Significance:** Permutation tests naturally control Type I error at any sample size.
- **Power:** Non-asymptotic power guarantees: tests are consistent and asymptotically powerful.
- **Convergence rate:** The estimator converges to the true ratio as O($N^{-1/6}$).

Algorithm 1: Single f-divergence two-sample test

Input: Samples $\mathbf{X} = (x_1, \ldots, x_n) \overset{\text{iid}}{\sim} P$, $\mathbf{Y} = (y_1, \ldots, y_m) \overset{\text{iid}}{\sim} Q$, significance level $\alpha \in (0,1)$, list of hyperparameter configurations $\mathcal{C}$, number of permutations B, f-divergence estimator $\tau : \mathcal{Y}^{m+n} \times \mathcal{C}$ that receives a sample as the first argument and a hyperparameter configuration as the second argument, Fuse function.

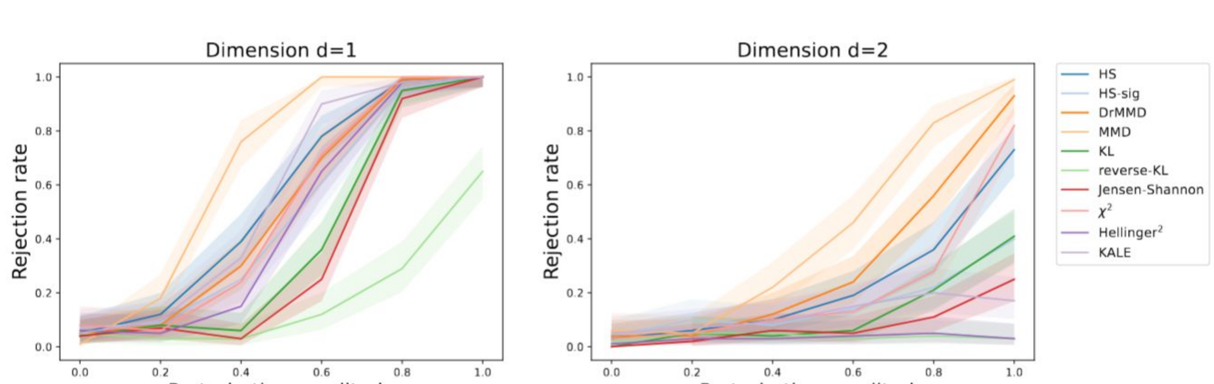
Output: Boolean decision: 1 (rejecting H_0), 0 otherwise.

```
Z ← (X, Y)                                   // Combine samples.
T ← {τ(Z, c) | c ∈ C}                        // Compute true statistics.
T^fuse ← Fuse(T)                             // Compute the fused observed statistic.
G ← {σ_1, ..., σ_B}                          // Generate permutations on n + m.
for σ ∈ G do
    for c ∈ C do
        T_{σ,c} ← τ(σZ, c)                   // Compute statistic on all configurations.
    end
    T_σ ← {T_{σ,c} | c ∈ C}                  // Collect statistics for permutation σ.
    T_σ^fuse ← Fuse(T_σ)                     // Compute fused statistic for permutation σ.
end
p_val ← (1 + |{σ_i ∈ G : T_σ^fuse ≥ T^fuse}|)/(B + 1)   // Compute p-value.
return 1(p_val ≤ α)
```

Experiments & Results

Smooth vs. localized differences

- Performance on perturbed d-dimensional uniform alternatives [1,2] with varying perturbation amplitudes. As the amplitude increases, the deviation from uniformity grows, leading to higher statistical power.
- Both MMD and DrMMD perform well in this setting, and the performance decreases for all methods in two dimensions. The f-Agg test maintains almost the same performance as the best method.



- Detection rate comparison on the Expo1D task, a New Physics model [4], varying the multiplier k of the perturbation. The DrMMD test achieves the highest power across all values.

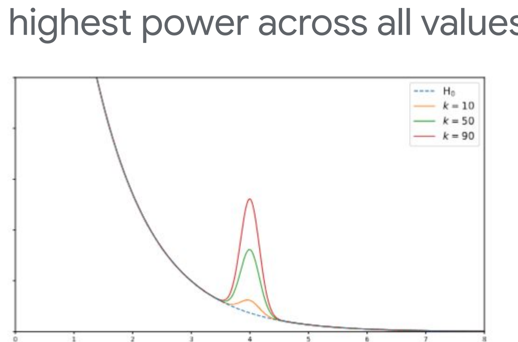
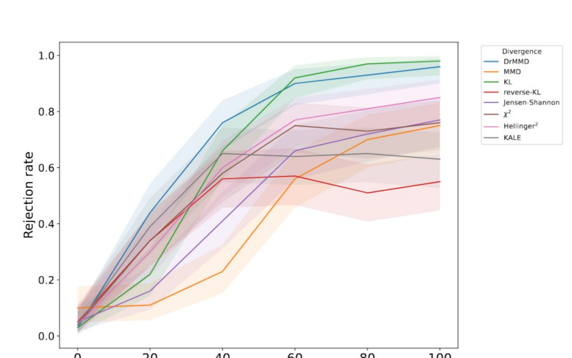


Figure 2: Expo-1D null (H_0) and alternative hypothesis density functions varying the multiplier k.



Auditing Privacy

- A randomized mechanism M is (ε,δ)-***differentially private*** if for any pair of datasets ***D*** and ***D'*** differing in *only one record* and any S

$$\mathrm{HS}_{e^\varepsilon}(M(D)||M(D')) = \mathrm{HS}_{e^\varepsilon}(M(D')||M(D)) = 0.$$

- Privacy violations detection rate for several non-private mechanisms. HS improves detection rate over DP-auditorium [1] which requires hundreds of thousand of samples to detect some of the mechanisms below.

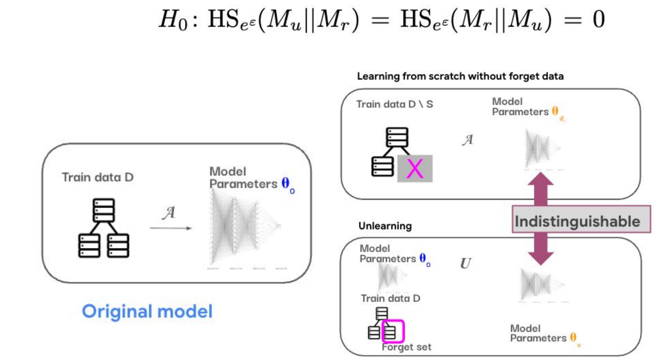
	Test	SVT3	SVT4	SVT5	SVT6	mean1	mean2
$n = 500$	HS	0.095	0.070	0.785	0.085	0.970	1.0
	HS-sig	0.075	0.050	0.505	0.100	0.705	1.0
	f-Agg	0.060	0.050	0.765	0.080	0.960	1.0
	DP-Aud	0.0	0.0	0.0	0.0	0.768	0.071
$n = 5000$	HS	~~0.175~~	~~0.090~~	~~0.995~~	~~0.165~~	~~1.0~~	~~1.0~~
	HS-sig	0.125	0.050	0.485	0.110	1.0	1.0
	f-Agg	0.170	0.045	0.995	0.125	1.0	1.0
	DP-Aud	0.0	0.0	0.010	0.0	1.0	0.717

- Previous methods required 100x more data points to achieve positive detection rates. In particular, SVT3 required millions of samples in previous work.

Auditing Unlearning

Two sample test

$$H_0 : \mathrm{HS}_{e^\varepsilon}(M_u||M_r) = \mathrm{HS}_{e^\varepsilon}(M_r||M_u) = 0$$

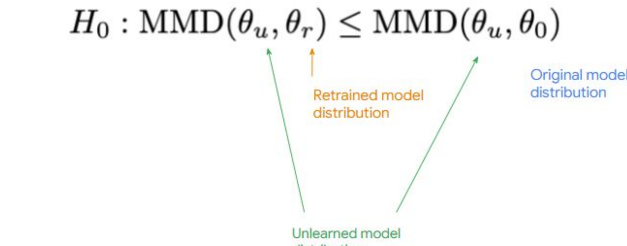


Unlearning mechanism	DrMMD	KALE	MMD	HS
Finetune	1.0	0.2	1.0	1.0
Pruning	1.0	0.0	1.0	1.0
SSD	1.0	0.0	1.0	1.0
Random label	1.0	0.0	1.0	1.0
Retrained-less	1.0	0.5	1.0	1.0
Retrained-batch	1.0	0.0	1.0	1.0

- Two-sample test power with ε = 0 and ε = 0.1. A rejection rate of 1.0 means the test detects a distributional difference.

Three sample test

$$H_0 : \mathrm{MMD}(\theta_u, \theta_r) \leq \mathrm{MMD}(\theta_u, \theta_0)$$



Method	Power of relative test
Finetune	1.0
Pruning	1.0
SSD	1.0
Random-label	0.0
Retrained-less	0.0
Retrained-batch	0.0

- Rejection rate for the relative similarity test. A rejection rate of 1.0 means the test detects that the unlearned model is closer to the original model than to the retrained model, a proxy for evidence that the unlearning method is ineffective. The test correctly identifies retraining from scratch on retain data as a safe procedure (rejection rate = 0.0).

[1] Kong, W., Medina, A. M., Ribero, M., & Syed, U. DP-auditorium: ***A large-scale library for auditing differential privacy.*** IEEE S&P (2024).
[2] Schrab, A., Kim, I., Albert, M., Laurent, B., Guedj, B., & Gretton, A. (2023). ***MMD aggregated two-sample test.*** Journal of Machine Learning Research
[3] Biggs, F., Schrab, A., & Gretton, A. ***MMD-FUSE: Learning and combining kernels for two-sample testing without data splitting***. NeurIPS (2023).
[4] Grosso, Gaia, and Marco Letizia. ***Multiple testing for signal-agnostic searches for new physics with machine learning.*** The European Physical Journal (2025)