Compress Then Test: Powerful Kernel Testing in Near-linear Time




Carles Domingo-Enrich[1] Raaz Dwivedi[2,3] Lester Mackey[4]

1 NYU 2 HARVARD UNIVERSITY 3 MIT 4 Microsoft Research

github.com/microsoft/goodpoints

Kernel two-sample testing

- $\mathbb{X}_n = (X_i)_{1\le i\le n}$ i.i.d. sample of $\mathbb{P}$, $\mathbb{Y}_n = (Y_i)_{1\le i\le n}$ i.i.d. sample of $\mathbb{Q}$.
- Null hypothesis: $\mathscr{H}_0 : \mathbb{P} = \mathbb{Q}$
- Non-parametric form via kernel maximum mean discrepancy (MMD)

$$\mathscr{H}_0 : \mathrm{MMD}^2_{\mathbf{k}}(\mathbb{P},\mathbb{Q}) \triangleq \mathbb{E}_{X,X'\sim\mathbb{P}}\mathbf{k}(X,X') + \mathbb{E}_{Y,Y'\sim\mathbb{P}}\mathbf{k}(Y,Y') - 2\mathbb{E}_{X\sim\mathbb{P},Y\sim\mathbb{Q}}\mathbf{k}(X,Y) = 0$$

with Test statistic: $$\Delta(\mathbb{X}_n,\mathbb{Y}_n) = \begin{cases} 0 & \text{if } \mathrm{MMD}^2_{\mathbf{k}}(\mathbb{X}_n,\mathbb{Y}_n) < t_\alpha \text{ (accept } \mathscr{H}_0) \\ 1 & \text{if } \mathrm{MMD}^2_{\mathbf{k}}(\mathbb{X}_n,\mathbb{Y}_n) \ge t_\alpha \text{ (reject } \mathscr{H}_0) \end{cases}$$

Prior MMD strategies are slow

- Computing MMD takes $O(n^2)$ time

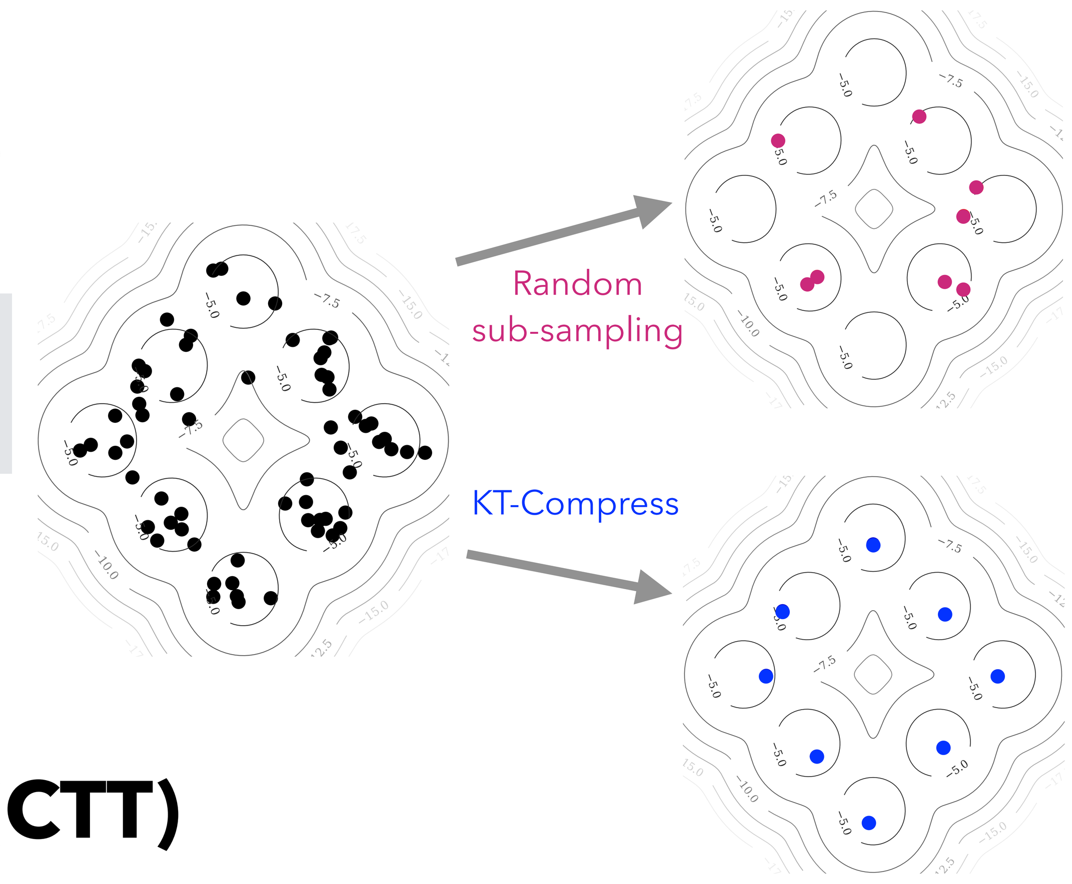
$$\mathrm{MMD}^2_{\mathbf{k}}(\mathbb{X}_n,\mathbb{Y}_n) = \frac{\sum_{i,i'=1}^n \mathbf{k}(X_i,X_{i'})}{n^2} + \frac{\sum_{j,j'=1}^n \mathbf{k}(Y_j,Y_{j'})}{n^2} - \frac{2\sum_{i=1}^n\sum_{j=1}^n \mathbf{k}(X_i,Y_j)}{n^2}$$

- Can we speed up the test while
  - A. respecting Type I error: reject null rarely when $\mathbb{P} = \mathbb{Q}$, and
  - B. keeping our test powerful: reject null often when $\mathbb{P} \neq \mathbb{Q}$?
- Prior speedup strategies sacrifice power!

Speed up testing by compressing

- Compress $\mathbb{X}_n$ to $\hat{\mathbb{X}}_m$ of size $m = \sqrt{n}$ with small $\mathrm{MMD}^2_{\mathbf{k}}(\mathbb{X}_n,\hat{\mathbb{X}}_m)$

	Random subsampling	KT-Compress (Shetty et al., 2023)
$\mathrm{MMD}_{\mathbf{k}}(\mathbb{X}_n,\hat{\mathbb{X}}_m)$	$O(n^{-1/4})$	$\tilde{O}(n^{-1/2})$
Time	$O(n^{1/2})$	$\tilde{O}(n)$



Compress Then Test (CTT)

1. Run KT-Compress to get $\mathbb{X}_n \to \hat{\mathbb{X}}_m$ & $\mathbb{Y}_n \to \hat{\mathbb{Y}}_m$ with $m = 2^{\mathfrak{g}}\sqrt{n}$
2. Use $\mathrm{MMD}^2_{\mathbf{k}}(\hat{\mathbb{X}}_m,\hat{\mathbb{Y}}_m)$ instead of $\mathrm{MMD}^2_{\mathbf{k}}(\mathbb{X}_n,\mathbb{Y}_n)$
3. Compute threshold t_α via *cheap permutations*: Group data into $s \ll \sqrt{n}$ bins, sample $\mathscr{B}$ permutations of $[s]$, and permute the s bins

CTT runtime: $\tilde{O}(n) + O(s^2B)$ if $\mathfrak{g} = \log\log n$

Original test runtime: $O(n^2B)$

Compress (the Data and) Then Test

1. Turns quadratic time tests to near-linear time tests
2. Provides up to 200x speed-up (1 hour → 20 sec)
3. Maintains level *and* power provably
4. Works with kernel approximations (Low-rank CTT)
5. Applies for kernel selection (Aggregated CTT)

CTT guarantees

A. Exact Type 1 error:

If $\mathrm{MMD}_{\mathbf{k}}(\mathbb{P},\mathbb{Q}) = 0$, then $\Pr[\Delta(\mathbb{X}_n,\mathbb{Y}_n) = 1] = \alpha$

B. Power:

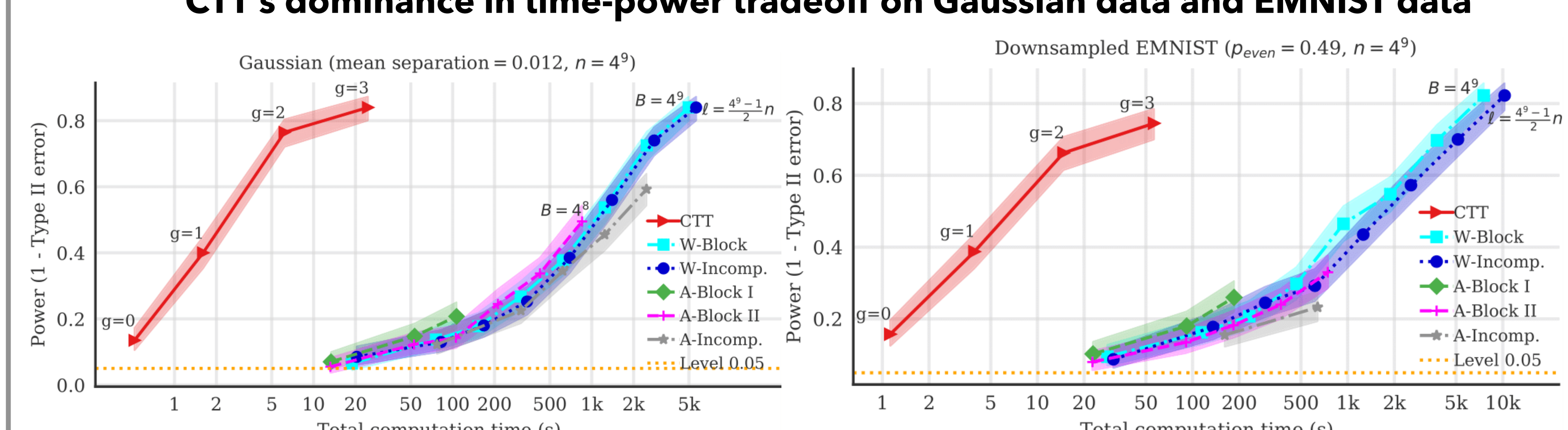
If $\mathrm{MMD}_{\mathbf{k}}(\mathbb{P},\mathbb{Q}) \ge$ Separation(β), then $\Pr[\Delta(\mathbb{X}_n,\mathbb{Y}_n) = 1] \ge 1-\beta$

Test name	MMD separation	Runtime
CTT (ours, Thm. 1)	$\frac{\mathrm{R}_{\mathbf{k},\mathbf{k}'}(\mathbb{P},n,\delta,\mathfrak{g})}{2^{\mathfrak{g}}\sqrt{n}} + n^{-\frac{1}{2}}$	$4^{\mathfrak{g}} n \log n$
Complete MMD (Gretton et al., 2012a)	$n^{-\frac{1}{2}}$	n^2
Block MMD (Zaremba et al., 2013)	$(Bn)^{-\frac{1}{4}}$	Bn
Incomplete MMD (Yamada et al., 2019)	$\ell^{-\frac{1}{4}}$	ℓ

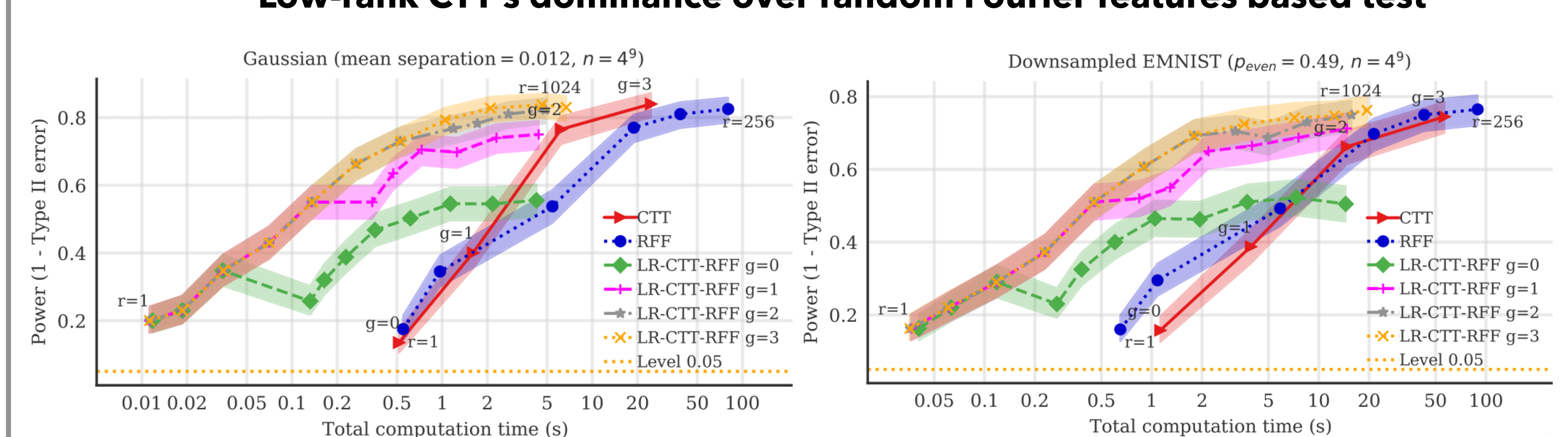
Tails of $\mathbb{P}$ on $\mathbb{R}^d$	Choice of $\mathbf{k}'$	$\mathrm{R}_{\mathbf{k},\mathbf{k}'}(\mathbb{P},n,\delta,\mathfrak{g})$
Compact	Compact $\mathbf{k}_{\mathrm{rt}}$	$(\log\frac{n}{\delta})^2$
Subexp	Analytic $\mathbf{k}$	$(\log\frac{n}{\delta})^{\frac{3d+5}{2}}$
Subexp	Subexp $\mathbf{k}_{\mathrm{rt}}$	$c_{m,\delta}(\log\frac{n}{\delta})^{\frac{d+5}{2}}$
ρ-Heavy-tail	ρ-Heavy-tail $\mathbf{k}_{\mathrm{rt}}$	$(\frac{n}{\delta})^{\frac{d}{2\rho}}(\log\frac{n}{\delta})^{\frac{d+5}{2}}$

B = Number of blocks used in Block MMD test
ℓ = Number of ordered index pairs in Incomplete MMD test
δ = Failure probability for CTT guarantees
$\mathbf{k}'$ = auxiliary kernel used by KT-Compress & $\mathbf{k}(x,y) = \int \mathbf{k}_{\mathrm{rt}}(x,z)\mathbf{k}_{\mathrm{rt}}(z,y)dz$

CTT's dominance in time-power tradeoff on Gaussian data and EMNIST data



Low-rank CTT's dominance over random Fourier features based test



Aggregated CTT's dominance for kernel selection on Blobs and Higgs data