Adaptive linear-time nonparametric two-sample testing

Zoltán Szabó (École Polytechnique)

Wittawat Jitkrittum
Kacper Chwialkowski
Arthur Gretton

CMAP seminar, Palaiseau
École Polytechnique November 22, 2016
Motivating examples: NLP, computer vision.
Two-sample test: t-test $\rightarrow$ distribution features.
Linear-time, interpretable, high-power, nonparametric t-test.
Numerical illustrations.
Motivating examples
Motivating example-1: NLP

Given: two categories of documents (Bayesian inference, neuroscience).

Task:

- test their distinguishability,
- most discriminative words $\rightarrow$ interpretability.
Motivating example-2: computer vision

Given: two sets of faces (happy, angry).

Task:
- check if they are different,
- determine the most discriminative features/regions.
One-page summary

Contribution:

- We propose a nonparametric t-test.
- It gives a reason why $H_0$ is rejected.
- It has high test power.
- It runs in linear time.
Contribution:

- We propose a nonparametric t-test.
- It gives a reason why $H_0$ is rejected.
- It has high test power.
- It runs in linear time.

Dissemination, code:

- NIPS-2016 [Jitkrittum et al., 2016]: full oral = top 1.84%.
Two-sample test, distribution features
What is a two-sample test?

Given:

- $X = \{x_i\}_{i=1}^n \overset{i.i.d.}{\sim} \mathbb{P}$, $Y = \{y_j\}_{j=1}^n \overset{i.i.d.}{\sim} \mathbb{Q}$.
- Example: $x_i$ = $i^{th}$ happy face, $y_j$ = $j^{th}$ sad face.

Problem: using $X$, $Y$ test $H_0$: $\mathbb{P} = \mathbb{Q}$, vs $H_1$: $\mathbb{P} \neq \mathbb{Q}$.
What is a two-sample test?

- **Given:**
  - $X = \{x_i\}_{i=1}^n \overset{i.i.d.}{\sim} P$, $Y = \{y_j\}_{j=1}^n \overset{i.i.d.}{\sim} Q$.
  - Example: $x_i = i^{th}$ happy face, $y_j = j^{th}$ sad face.

- **Problem:** using $X$, $Y$ test

  
  $H_0 : P = Q$, vs  
  $H_1 : P \neq Q$. 
What is a two-sample test?

Given:
- \( X = \{x_i\}_{i=1}^{n} \overset{i.i.d.}{\sim} P, \ Y = \{y_j\}_{j=1}^{n} \overset{i.i.d.}{\sim} Q. \)
- Example: \( x_i = i^{th} \) happy face, \( y_j = j^{th} \) sad face.

Problem: using \( X, \ Y \) test

\[ H_0 : P = Q, \ \text{vs} \ \ H_1 : P \neq Q. \]

Assume \( X, \ Y \subset \mathbb{R}^d. \)
Ingredients of two-sample test

- Test statistic: \( \hat{\lambda}_n = \hat{\lambda}_n(X, Y) \), random.
- Significance level: \( \alpha = 0.01 \).
- Under \( H_0 \): \( P_{H_0}(\hat{\lambda}_n \leq T_\alpha) = 1 - \alpha \).

\[ \text{correctly accepting } H_0 \]

\[ \text{power} = P_{H_1}(\hat{\lambda}_n) \]

\[ \text{Type I error rate} = P_{H_0}(\hat{\lambda}_n) \]

\[ \text{Type II error rate} = 1 - P_{H_1}(\hat{\lambda}_n) \]

\[ \text{Power} = 1 - \text{Type II error rate} \]

\[ \text{Type I error rate} = \alpha \]

\[ \text{Power} = \text{Type II error rate} \]

\[ \text{Power} = 1 - \text{Type II error rate} \]

\[ \text{Type I error rate} = \alpha = 0.01 \]

\[ \text{Power} = 1 - \alpha = 0.99 \]
Ingredients of two-sample test

- Test statistic: $\hat{\lambda}_n = \hat{\lambda}_n(X, Y)$, random.
- Significance level: $\alpha = 0.01$.
- **Under $H_0$:** $P_{H_0}(\hat{\lambda}_n \leq T_\alpha) = 1 - \alpha$.
  
  (Correctly accepting $H_0$)

- **Under $H_1$:** $P_{H_1}(T_\alpha < \hat{\lambda}_n) = P($correctly rejecting $H_0) =: $ power.
Given: 2 Gaussians with (possibly) different means.
Solution: $t$-test.
Setup: 2 Gaussians; same means, different variances.
Idea: look at 2nd-order features of RVs.
Towards representations of distributions: $\mathbb{E}X^2$

- Setup: 2 Gaussians; same means, different variances.
- Idea: look at 2nd-order features of RVs.
- $\varphi_x = x^2 \Rightarrow$ difference in $\mathbb{E}X^2$. 

**Two Gaussians with different variances**

**Densities of feature $X^2$**
Towards representations of distributions: further moments

- Setup: a Gaussian and a Laplacian distribution.
- Challenge: their means and variances are the same.
- Idea: look at higher-order features.

Let us consider feature/distribution representations!
Kernel: similarity between features

- Given: $x$ and $x'$ objects (images or texts).
Kernel: similarity between features

- Given: $\mathbf{x}$ and $\mathbf{x}'$ objects (images or texts).
- Question: how similar they are?
Given: \( x \) and \( x' \) objects (images or texts).

Question: how similar they are?

Define features of the objects:

\[
\varphi_x : \text{features of } x,
\]
\[
\varphi_{x'} : \text{features of } x'.
\]

Kernel: inner product of these features

\[
k(x, x') := \langle \varphi_x, \varphi_{x'} \rangle.
\]
Kernel examples on $\mathbb{R}^d$ ($\gamma > 0$, $p \in \mathbb{Z}^+$)

- Polynomial kernel:

$$k(x, y) = (\langle x, y \rangle + \gamma)^p.$$ 

- Gaussian kernel:

$$k(x, y) = e^{-\gamma \|x - y\|_2^2}.$$
Towards distribution features

\[ \sim \sim P \sim Q \]
Towards distribution features

\begin{align*}
k(\text{dog}_i, \text{dog}_j) & \quad k(\text{dog}_i, \text{fish}_j) \\
n(\text{fish}_j, \text{dog}_i) & \quad k(\text{fish}_i, \text{fish}_j)
\end{align*}
\[ \hat{\mathsf{MMD}}^2(P, Q) = \hat{K}_{P,P} + \hat{K}_{Q,Q} - 2\hat{K}_{P,Q} \] (without diagonals in $\hat{K}_{P,P}, \hat{K}_{Q,Q}$)

\[\uparrow \hat{\mathsf{MMD}} \text{ illustration credit: Arthur Gretton}\]
Kernel recall: $k(x, x') = \langle \varphi_x, \varphi_{x'} \rangle$. 
Kernel → distribution feature

- Kernel recall: \( k(x, x') = \langle \phi_x, \phi_{x'} \rangle \).
- Feature of \( \mathbb{P} \) (mean embedding):
  \[
  \mu_\mathbb{P} := \mathbb{E}_{x \sim \mathbb{P}}[\phi_x].
  \]
Kernel recall: \( k(x, x') = \langle \varphi_x, \varphi_{x'} \rangle \).

Feature of \( \mathbb{P} \) (mean embedding):

\[
\mu_\mathbb{P} := \mathbb{E}_{x \sim \mathbb{P}}[\varphi_x].
\]

Previous quantity: unbiased estimate of

\[
MMD^2(\mathbb{P}, \mathbb{Q}) = \| \mu_\mathbb{P} - \mu_\mathbb{Q} \|^2.
\]
Kernel → distribution feature

- Kernel recall: \( k(\mathbf{x}, \mathbf{x}') = \langle \varphi_{\mathbf{x}}, \varphi_{\mathbf{x}'} \rangle \).
- Feature of \( \mathbb{P} \) (mean embedding):
  \[
  \mu_{\mathbb{P}} := \mathbb{E}_{\mathbf{x} \sim \mathbb{P}}[\varphi_{\mathbf{x}}].
  \]
- Previous quantity: unbiased estimate of
  \[
  \text{MMD}^2(\mathbb{P}, \mathbb{Q}) = \| \mu_{\mathbb{P}} - \mu_{\mathbb{Q}} \|^2 .
  \]

Valid test [Gretton et al., 2012]. Challenges:

1. Threshold choice: 'ugly' asymptotics of \( n \text{MMD}^2(\mathbb{P}, \mathbb{P}) \).
2. Test statistic: quadratic time complexity.
3. Witness \( \in \mathcal{H}(k) \): can be hard to interpret.
Linear-time tests
Linear-time 2-sample test

- Recall:

\[
MMD^2(\mathbb{P}, \mathbb{Q}) = \| \mu_\mathbb{P} - \mu_\mathbb{Q} \|^2_{\mathcal{H}(k)}.
\]

- Changing [Chwialkowski et al., 2015] this to

\[
\rho^2(\mathbb{P}, \mathbb{Q}) := \frac{1}{J} \sum_{j=1}^{J} [\mu_\mathbb{P}(v_j) - \mu_\mathbb{Q}(v_j)]^2
\]

with random \( \{v_j\}_{j=1}^{J} \) test locations.

\( \rho \) is a metric (a.s.). How do we estimate it? Distribution under \( H_0 \)?
Estimation

Compute

\[ \rho^2(\mathcal{P}, \mathcal{Q}) = \frac{1}{J} \sum_{j=1}^{J} [\hat{\mu}_\mathcal{P}(\mathbf{v}_j) - \hat{\mu}_\mathcal{Q}(\mathbf{v}_j)]^2, \]

where \( \hat{\mu}_\mathcal{P}(\mathbf{v}) = \frac{1}{n} \sum_{i=1}^{n} k(\mathbf{x}_i, \mathbf{v}) \). Example using \( k(\mathbf{x}, \mathbf{v}) = e^{-\frac{||\mathbf{x} - \mathbf{v}||^2}{2\sigma^2}} \):
Estimation – continued

\[ \rho^2(\mathbb{P}, \mathbb{Q}) = \frac{1}{J} \sum_{j=1}^{J} [\hat{\mu}_\mathbb{P}(v_j) - \hat{\mu}_\mathbb{Q}(v_j)]^2 \]
\[ \rho^2(\mathbb{P}, Q) = \frac{1}{J} \sum_{j=1}^{J} \left[ \hat{\mu}_\mathbb{P}(v_j) - \hat{\mu}_Q(v_j) \right]^2 \]

\[ = \frac{1}{J} \sum_{j=1}^{J} \left[ \frac{1}{n} \sum_{i=1}^{n} k(x_i, v_j) - \frac{1}{n} \sum_{i=1}^{n} k(y_i, v_j) \right]^2 \]
Estimation – continued

\[
\rho^2(P, Q) = \frac{1}{j} \sum_{j=1}^{J} \left[ \hat{\mu}_P(v_j) - \hat{\mu}_Q(v_j) \right]^2
\]

\[
= \frac{1}{j} \sum_{j=1}^{J} \left[ \frac{1}{n} \sum_{i=1}^{n} k(x_i, v_j) - \frac{1}{n} \sum_{i=1}^{n} k(y_i, v_j) \right]^2
= \frac{1}{j} \sum_{j=1}^{J} (\bar{z}_n)_j^2 = \frac{1}{j} \bar{z}_n^T \bar{z}_n,
\]

where \( \bar{z}_n = \frac{1}{n} \sum_{i=1}^{n} \left[ k(x_i, v_j) - k(y_i, v_j) \right]_{j=1}^{J} \in \mathbb{R}^J. \)
\[ \rho^2(\bar{P}, \bar{Q}) = \frac{1}{J} \sum_{j=1}^{J} [\hat{\mu}_p(v_j) - \hat{\mu}_q(v_j)]^2 \]

\[ = \frac{1}{J} \sum_{j=1}^{J} \left[ \frac{1}{n} \sum_{i=1}^{n} k(x_i, v_j) - \frac{1}{n} \sum_{i=1}^{n} k(y_i, v_j) \right]^2 = \frac{1}{J} \sum_{j=1}^{J} (\bar{z}_n)_j^2 = \frac{1}{J} \bar{z}_n^T \bar{z}_n, \]

where \( \bar{z}_n = \frac{1}{n} \sum_{i=1}^{n} [k(x_i, v_j) - k(y_i, v_j)]_j \in \mathbb{R}^J. \)

- Good news: estimation is linear in \( n! \)
- Bad news: intractable null distr. = \( \sqrt{n} \rho^2(\bar{P}, \bar{P}) \overset{w}{\longrightarrow} \) sum of \( J \) correlated \( \chi^2 \).
Modified test statistic:

\[ \hat{\lambda}_n = n \bar{z}_n^T \Sigma_n^{-1} \bar{z}_n, \]

where \( \Sigma_n = \text{cov} \left( \{z_i\}_{i=1}^n \right) \).

Under \( H_0 \):

\[ \hat{\lambda}_n \xrightarrow{w} \chi^2(J). \Rightarrow \text{Easy to get the} \ (1 - \alpha)\text{-quantile!} \]
Our idea
Until this point: test locations ($V$) are \textbf{fixed}.

Instead: choose $\theta = \{V, \sigma\}$ to

\begin{center}
maximize lower bound on the test power.
\end{center}
Until this point: test locations ($\mathcal{V}$) are fixed.

Instead: choose $\theta = \{\mathcal{V}, \sigma\}$ to maximize lower bound on the test power.

**Theorem (Lower bound on power, for large $n$)**

**Test power** $\geq L(\lambda_n)$; $L$: explicit function, increasing.

Here,

- $\lambda_n = n\mu^T\Sigma^{-1}\mu$: population version of $\hat{\lambda}_n$.
- $\mu = \mathbb{E}_{xy}[z_1]$, $\Sigma = \mathbb{E}_{xy}[(z_1 - \mu)(z_1 - \mu)^T]$. 
Convergence of the $\lambda_n$ estimator

But $\lambda_n$ is unknown. Split $(X, Y)$ into $(X_{tr}, Y_{tr})$ and $(X_{te}, Y_{te})$.

Locations, kernel parameter: $\hat{\theta} = \arg\max_{\theta} \hat{\lambda}_{n/2}^{tr}(\theta)$. 

Examples:

- $K = \kappa_{\sigma}^p x, y^q e^{x \cdot y}^2$ : $\sigma \equiv 0$
- $K = \kappa_A^p x, y^q e^{p x \cdot y^T A p x \cdot y}$ : $A \equiv 0$
Convergence of the $\lambda_n$ estimator

But $\lambda_n$ is unknown. Split $(X, Y)$ into $(X_{tr}, Y_{tr})$ and $(X_{te}, Y_{te})$.

1. Locations, kernel parameter: $\hat{\theta} = \arg\max_{\theta} \hat{\lambda}_{tr}^{\frac{n}{2}}(\theta)$.

2. Test statistic: $\hat{\lambda}_{te}^{\frac{n}{2}}(\hat{\theta})$. 
Theorem (Guarantee on objective approximation, $\gamma_n \to 0$)

$$\sup_{\nu,\mathcal{K}} |\mathbf{z}_n^T(\Sigma_n + \gamma_n)^{-1}\mathbf{z}_n - \mu^T \Sigma^{-1} \mu| = O\left(n^{-\frac{1}{4}}\right).$$
Convergence of the $\lambda_n$ estimator

But $\lambda_n$ is unknown. Split $p \mathbf{X}, Y q$ into $p \mathbf{X}_{\text{tr}}, Y_{\text{tr}} q$ and $p \mathbf{X}_{\text{te}}, Y_{\text{te}} q$.

Locations, kernel parameter: $\hat{\theta} = \arg \max_{\theta} \lambda_{\text{tr}} n^2 p \theta q$.

Test statistic: $\hat{\lambda}_{\text{te}} n^2 \hat{\theta}$.

Theorem (Guarantee on objective approximation, $\gamma_n \to 0$)

$$\sup_{\nu, \mathcal{K}} \left| \bar{\mathbf{z}}_n^T (\Sigma_n + \gamma_n)^{-1} \bar{\mathbf{z}}_n - \mu^T \Sigma^{-1} \mu \right| = \mathcal{O}(n^{-\frac{1}{4}}).$$

Examples:

$$\mathcal{K} = \left\{ k_{\sigma}(\mathbf{x}, \mathbf{y}) = e^{-\frac{\| \mathbf{x} - \mathbf{y} \|^2}{2\sigma^2}} : \sigma > 0 \right\},$$

$$\mathcal{K} = \left\{ k_{\mathbf{A}}(\mathbf{x}, \mathbf{y}) = e^{-(\mathbf{x} - \mathbf{y})^T \mathbf{A}(\mathbf{x} - \mathbf{y})} : \mathbf{A} > 0 \right\}.$$
Numerical demos
Parameter settings

- Gaussian kernel ($\sigma$). $\alpha = 0.01$. $J = 1$. Repeat 500 trials.
- Report
  \[ P(\text{reject } H_0) \approx \frac{\#\text{times } \hat{\lambda}_n > T_\alpha}{\#\text{trials}}. \]

- Compare 4 methods
  - **ME-full**: Optimize $\mathcal{V}$ and Gaussian bandwidth $\sigma$.
  - **ME-grid**: Optimize $\sigma$. Random $\mathcal{V}$ [Chwialkowski et al., 2015].
  - **MMD-quad**: Test with quadratic-time MMD [Gretton et al., 2012].
  - **MMD-lin**: Test with linear-time MMD [Gretton et al., 2012].

- Optimize kernels to power in MMD-lin, MMD-quad.

Keyword-based category assignment into 4 groups:

- Bayesian inference, Deep learning, Learning theory, Neuroscience

$d = 2000$ nouns. TF-IDF representation.

<table>
<thead>
<tr>
<th>Problem</th>
<th>$n^{te}$</th>
<th>ME-full</th>
<th>ME-grid</th>
<th>MMD-quad</th>
<th>MMD-lin</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. Bayes-Bayes</td>
<td>215</td>
<td>.012</td>
<td>.018</td>
<td>.022</td>
<td>.008</td>
</tr>
<tr>
<td>2. Bayes-Deep</td>
<td>216</td>
<td>.954</td>
<td>.034</td>
<td>.906</td>
<td>.262</td>
</tr>
<tr>
<td>3. Bayes-Learn</td>
<td>138</td>
<td>.990</td>
<td>.774</td>
<td>1.00</td>
<td>.238</td>
</tr>
<tr>
<td>4. Bayes-Neuro</td>
<td>394</td>
<td>1.00</td>
<td>.300</td>
<td>.952</td>
<td>.972</td>
</tr>
<tr>
<td>5. Learn-Deep</td>
<td>149</td>
<td>.956</td>
<td>.052</td>
<td>.876</td>
<td>.500</td>
</tr>
<tr>
<td>6. Learn-Neuro</td>
<td>146</td>
<td>.960</td>
<td>.572</td>
<td>1.00</td>
<td>.538</td>
</tr>
</tbody>
</table>

Performance of ME-full [$O(n)$] is comparable to MMD-quad [$O(n^2)$].
NLP: most/least discriminative words

- Aggregating over trials; example: 'Bayes-Neuro'.
- Most discriminative words:
  
  - spike, markov, cortex, dropout, recurr, iii, gibb.
  - learned test locations: highly interpretable,
  - 'markov', 'gibb' (⇔ Gibbs): Bayesian inference,
  - 'spike', 'cortex': key terms in neuroscience.
NLP: most/least discriminative words

- Aggregating over trials; example: 'Bayes-Neuro'.

- Least discriminative ones:
  
  circumfer, bra, dominiqu, rhino, mitra, kid, impostor.
Distinguish positive/negative emotions

- Karolinska Directed Emotional Faces (KDEF) [Lundqvist et al., 1998].
- 70 actors = 35 females and 35 males.
- $d = 48 \times 34 = 1632$. Grayscale. Pixel features.

```
+ : happy neutral surprised
- : afraid angry disgusted
```

<table>
<thead>
<tr>
<th>Problem</th>
<th>$n^{te}$</th>
<th>ME-full</th>
<th>ME-grid</th>
<th>MMD-quad</th>
<th>MMD-lin</th>
</tr>
</thead>
<tbody>
<tr>
<td>± vs. ±</td>
<td>201</td>
<td>.010</td>
<td>.012</td>
<td>.018</td>
<td>.008</td>
</tr>
<tr>
<td>+ vs. −</td>
<td>201</td>
<td>.998</td>
<td>.656</td>
<td>1.00</td>
<td>.578</td>
</tr>
</tbody>
</table>

Learned test location (averaged) =
We proposed a nonparametric t-test:
- linear time,
- high-power (≈ 'MMD-quad'),
2 demos: discriminating
- documents of different categories,
- positive/negative emotions.
Thank you for the attention!

Acknowledgements: This work was supported by the Gatsby Charitable Foundation.
Non-convexity, informative features.
Number of locations \((J)\).
MMD: IPM representation.
Estimation of \(\text{MMD}^2\).
Proof idea.
Computational complexity: \((J, n, d)\)-dependence.
Non-convexity, informative features

2D problem:

\[ \mathbb{P} := \mathcal{N}(0, I), \quad \mathbb{Q} := \mathcal{N}(e_1, I). \]

\[ \mathcal{V} = \{\mathbf{v}_1, \mathbf{v}_2\}. \text{ Fix } \mathbf{v}_1 \text{ to } \blacktriangle. \]

\[ \mathbf{v}_2 \mapsto \hat{\lambda}_{n/2}(\mathbf{v}_1, \mathbf{v}_2): \text{ contour plot.} \]
Non-convexity, informative features

- Nearby locations: do not increase discriminability.
- Non-convexity: reveals multiple ways to capture the difference.
Small $J$:

- often enough to detect the difference of $\mathbb{P}$ & $\mathbb{Q}$.
- few distinguishing regions to reject $H_0$.
- faster test.
Very large $J$:

- Test power need not increase monotonically in $J$ (more locations $\Rightarrow$ statistic can gain in variance).
- Defeats the purpose of a linear-time test.
MMD: IPM representation

\[ MMD^2(\mathbb{P}, \mathbb{Q}) = \| \mu_\mathbb{P} - \mu_\mathbb{Q} \|^2_{\mathcal{H}(k)} \]
MMD: IPM representation

\[ MMD^2(\mathbb{P}, \mathbb{Q}) = \|\mu_P - \mu_Q\|^2_{\mathcal{H}(k)} = \left[ \sup_{\|f\|_{\mathcal{H}(k)} \leq 1} \langle \mu_P - \mu_Q, f \rangle_{\mathcal{H}(k)} \right]^2 \]
\[ \text{MMD}^2(\mathbb{P}, \mathbb{Q}) = \| \mu_\mathbb{P} - \mu_\mathbb{Q} \|^2_{\mathcal{H}(k)} = \left[ \sup_{\| f \|_{\mathcal{H}(k)} \leq 1} \langle \mu_\mathbb{P} - \mu_\mathbb{Q}, f \rangle_{\mathcal{H}(k)} \right]^2 \]

\[ \overset{(*)}{=} \left[ \sup_{\| f \|_{\mathcal{H}(k)} \leq 1} \mathbb{E}_{x \sim \mathbb{P}} f(x) - \mathbb{E}_{y \sim \mathbb{Q}} f(y) \right]^2. \]
MMD: IPM representation

\[
MMD^2(\mathbb{P}, \mathbb{Q}) = \|\mu_\mathbb{P} - \mu_\mathbb{Q}\|_{\mathcal{H}(k)}^2 = \left[ \sup_{\|f\|_{\mathcal{H}(k)} \leq 1} \langle \mu_\mathbb{P} - \mu_\mathbb{Q}, f \rangle_{\mathcal{H}(k)} \right]^2
\]

\[
\overset{(\ast)}{=} \left[ \sup_{\|f\|_{\mathcal{H}(k)} \leq 1} \mathbb{E}_{x \sim \mathbb{P}} f(x) - \mathbb{E}_{y \sim \mathbb{Q}} f(y) \right]^2.
\]

\((\ast)\) in details:

\[
\langle \mu_\mathbb{P}, f \rangle_{\mathcal{H}(k)} = \left\langle \int k(\cdot, x) d\mathbb{P}(x), f \right\rangle_{\mathcal{H}(k)}
\]
$$MMD^2(\mathbb{P}, \mathbb{Q}) = \|\mu_\mathbb{P} - \mu_\mathbb{Q}\|_{\mathcal{H}(k)}^2 = \left[ \sup_{\|f\|_{\mathcal{H}(k)} \leq 1} \langle \mu_\mathbb{P} - \mu_\mathbb{Q}, f \rangle_{\mathcal{H}(k)} \right]^2$$

\[\leq \sup_{\|f\|_{\mathcal{H}(k)} \leq 1} \mathbb{E}_{x \sim \mathbb{P}} f(x) - \mathbb{E}_{y \sim \mathbb{Q}} f(y) \]

\((\ast)\) in details:

$$\langle \mu_\mathbb{P}, f \rangle_{\mathcal{H}(k)} = \left\langle \int k(\cdot, x) d\mathbb{P}(x), f \right\rangle_{\mathcal{H}(k)} = \int \left\langle k(\cdot, x), f \right\rangle_{\mathcal{H}(k)} d\mathbb{P}(x) = f(x)$$

Zoltán Szabó
Adaptive linear-time nonparametric two-sample testing
\[ MMD^2(\mathbb{P}, \mathbb{Q}) = \|\mu_\mathbb{P} - \mu_\mathbb{Q}\|_{\mathcal{H}(k)}^2 = \left[ \sup_{\|f\|_{\mathcal{H}(k)} \leq 1} \langle \mu_\mathbb{P} - \mu_\mathbb{Q}, f \rangle_{\mathcal{H}(k)} \right]^2 \]

\[ \overset{(*)}{=} \left[ \sup_{\|f\|_{\mathcal{H}(k)} \leq 1} \mathbb{E}_{x \sim \mathbb{P}} f(x) - \mathbb{E}_{y \sim \mathbb{Q}} f(y) \right]^2. \]

(*) in details:

\[ \langle \mu_\mathbb{P}, f \rangle_{\mathcal{H}(k)} = \left\langle \int k(\cdot, x) d\mathbb{P}(x), f \right\rangle_{\mathcal{H}(k)} = \int \underbrace{\langle k(\cdot, x), f \rangle_{\mathcal{H}(k)} d\mathbb{P}(x)}_{=f(x)} = \mathbb{E}_{x \sim \mathbb{P}} f(x). \]
Squared difference between feature means:

\[
MMD^2(P, Q) = \| \mu_P - \mu_Q \|^2_H = \langle \mu_P - \mu_Q, \mu_P - \mu_Q \rangle_H \\
= \langle \mu_P, \mu_P \rangle_H + \langle \mu_Q, \mu_Q \rangle_H - 2 \langle \mu_P, \mu_Q \rangle_H \\
= \mathbb{E}_{P,P} k(x, x') + \mathbb{E}_{Q,Q} k(y, y') - 2\mathbb{E}_{P,Q} k(x, y).
\]
Estimation of $MMD^2$

Squared difference between feature means:

$$MMD^2(P, Q) = \| \mu_P - \mu_Q \|_H^2 = \langle \mu_P - \mu_Q, \mu_P - \mu_Q \rangle_H$$

$$= \langle \mu_P, \mu_P \rangle_H + \langle \mu_Q, \mu_Q \rangle_H - 2 \langle \mu_P, \mu_Q \rangle_H$$

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

Unbiased empirical estimate for $\{x_i\}_{i=1}^n \sim P$, $\{y_j\}_{j=1}^n \sim Q$: 

$$\hat{MMD}^2(P, Q) = \bar{K}_{P,P} + \bar{K}_{Q,Q} - 2\bar{K}_{P,Q}. $$
Proof idea

1. Lower bound on the test power:
   1. \(|\hat{\lambda}_n - \lambda_n| \preceq \|\bar{z}_n - \mu\|_2 + \|\Sigma_n - \Sigma\|_F\).
   2. Bound the r.h.s. by Hoeffding inequality $\Rightarrow P(|\hat{\lambda}_n - \lambda_n| \geq t)$.
   3. By reparameterization: $P(\hat{\lambda}_n \geq T_\alpha)$ bound.
Proof idea

1. Lower bound on the test power:
   1. \(|\hat{\lambda}_n - \lambda_n| \lesssim \|\bar{z}_n - \mu\|_2 + \|\Sigma_n - \Sigma\|_F.\)
   2. Bound the r.h.s. by Hoeffding inequality \(\Rightarrow P(|\hat{\lambda}_n - \lambda_n| \geq t).\)
   3. By reparameterization: \(P(\hat{\lambda}_n \geq T_{\alpha})\) bound.

2. Uniformly \(\hat{\lambda}_n \approx \lambda_n:\)
   - Reduction to bounding \(\sup_{\mathcal{V},\mathcal{K}} \|\bar{z}_n - \mu\|_2, \sup_{\mathcal{V},\mathcal{K}} \|\Sigma_n - \Sigma\|_F.\)
   - Empirical processes, Dudley entropy bound.
Optimization & testing: linear in \( n \).

Testing: \( \mathcal{O} (ndJ + nJ^2 + J^3) \).

Optimization: \( \mathcal{O} (ndJ^2 + J^3) \) per gradient ascent.


