Adaptive linear-time nonparametric t-test

Zoltán Szabó (École Polytechnique)

Wittawat Jitkrittum
Kacper Chwialkowski
Arthur Gretton

Facebook AI Research, Paris
November 21, 2016
Motivating examples: NLP, computer vision.
Two-sample test: t-test → 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 → 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.
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:
  - $X = \{x_i\}_{i=1}^n \overset{i.i.d.}{\sim} P$, $Y = \{y_j\}_{j=1}^n \overset{i.i.d.}{\sim} Q$.
  - Example: $x_i = i^{th}$ happy face, $y_j = j^{th}$ sad face.
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

\[ H_0 : P = Q, \quad \text{vs} \]
\[ H_1 : P \neq Q. \]
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 $X, Y \subset \mathbb{R}^d$. 

Zoltán Szabó Adaptive linear-time nonparametric t-test
Ingredients of two-sample test

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

![Graph showing distribution of test statistic $\hat{\lambda}_n$ under $H_0$ and $H_1$.]

- Correctly accepting $H_0$.
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 $H_1$: $P_{H_1}(T_\alpha < \hat{\lambda}_n) = P(\text{correctly rejecting } H_0) =: \text{power}$. 

[Graph and diagram showing the distribution of $\hat{\lambda}_n$, $P_{H_0}(\hat{\lambda}_n)$, $P_{H_1}(\hat{\lambda}_n)$, $T_\alpha$, and $\hat{\lambda}_n$.]

Zoltán Szabó  Adaptive linear-time nonparametric t-test
Given: 2 Gaussians with (possibly) different means.
Solution: \( t \)-test.
Towards representations of distributions: $E X^2$

- 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 $X^2$
Setup: a Gaussian and a Laplacian distribution.
Challenge: their means and variances are the same.
Idea: look at higher-order features.
Given: \( x \) and \( x' \) objects (images or texts).
Kernel: similarity between features

- Given: \( x \) and \( x' \) 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:

\( \varphi_x : \) features of \( x \),

\( \varphi_{x'} : \) 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

$k(dog_i, dog_j)$  $k(dog_i, fish_j)$

$k(fish_j, dog_i)$  $k(fish_i, fish_j)$
Towards distribution features

Zoltán Szabó

Adaptive linear-time nonparametric t-test
Towards distribution features

\[ \hat{\text{MMD}}^2(P, Q) = \hat{K}_{P,P} + \hat{K}_{Q,Q} - 2\hat{K}_{P,Q} \]  
(without diagonals in \( \hat{K}_{P,P}, \hat{K}_{Q,Q} \))

\[ \hat{MMD} \] illustration credit: Arthur Gretton

\[ \text{Adaptive linear-time nonparametric t-test} \]
Kernel recall: $k(x, x') = \langle \phi_x, \phi_{x'} \rangle$. 
Kernel → distribution feature

- Kernel recall: \( k(\mathbf{x}, \mathbf{x}') = \langle \varphi_\mathbf{x}, \varphi_\mathbf{x}' \rangle \).
- Feature of \( \mathbb{P} \) (mean embedding):

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

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

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

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

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

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

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

where $\hat{\mu}_P(v) = \frac{1}{n} \sum_{i=1}^{n} k(x_i, v)$. Example using $k(x, v) = e^{-\frac{||x-v||^2}{2\sigma^2}}$.
Estimation – continued

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

Good news: estimation is linear in \( n \)
Bad news: intractable null distr. = (?)
Estimation – continued

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

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

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

\[ = \frac{1}{J} \sum_{j=1}^{J} \left[ \frac{1}{n} \sum_{i=1}^{n} k(\mathbf{x}_i, \mathbf{v}_j) - \frac{1}{n} \sum_{i=1}^{n} k(\mathbf{y}_i, \mathbf{v}_j) \right]^2 = \frac{1}{J} \sum_{j=1}^{J} (\tilde{z}_n)_j^2 = \frac{1}{J} \tilde{z}_n^T \tilde{z}_n, \]

where \( \tilde{z}_n = \frac{1}{n} \sum_{i=1}^{n} \left[ k(\mathbf{x}_i, \mathbf{v}_j) - k(\mathbf{y}_i, \mathbf{v}_j) \right]_j^J \in \mathbb{R}^J. \)
Estimation – continued

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

\[
    = \frac{1}{J} \sum_{j=1}^{J} \left[ \frac{1}{n} \sum_{i=1}^{n} k(\mathbf{x}_i, \mathbf{v}_j) - \frac{1}{n} \sum_{i=1}^{n} k(\mathbf{y}_i, \mathbf{v}_j) \right]^2
    = \frac{1}{J} \sum_{j=1}^{J} (\mathbf{z}_n)_j^2
    = \frac{1}{J} \mathbf{z}_n^T \mathbf{z}_n,
\]

where \( \mathbf{z}_n = \frac{1}{n} \sum_{i=1}^{n} \left[ k(\mathbf{x}_i, \mathbf{v}_j) - k(\mathbf{y}_i, \mathbf{v}_j) \right]_{j=1}^{J} \in \mathbb{R}^J. \)

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

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

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

Under \( H_0 \):

- \( \hat{\lambda}_n \xrightarrow{w} \chi^2(J) \). \( \Rightarrow \) Easy to get the \((1 - \alpha)\)-quantile!
Our idea
Until this point: test locations ($\mathcal{V}$) are fixed.
Instead: choose $\theta = \{\mathcal{V}, \sigma\}$ to maximize lower bound on the test power.
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 $(X_{te}, Y_{te})$.

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

2. Test statistic: $\hat{\lambda}^{te}_n(\hat{\theta})$. 

Examples:

1. $K = \|x - y\|^2$:
   
2. $K = \|A^{\top}x - A^{\top}y\|^2$:

Zoltán Szabó

Adaptive linear-time nonparametric t-test
Convergence of the $\lambda_n$ estimator

But $\lambda_n$ is unknown. Split $pX, Yq$ into $pX_{tr}, Y_{tr}q$ and $pX_{te}, Y_{te}q$.

1 Locations, kernel parameter: $\hat{\theta}$

$\hat{\lambda}_{tr}n^2 p_{\theta}q$

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

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

But $\lambda_n$ is unknown. Split $pX, Yq$ into $pX_{tr}, Y_{tr}q$ and $pX_{te}, Y_{te}q$.

1. Locations, kernel parameter: $\hat{\theta}$

$\arg \max \theta \hat{\lambda}_{tr} n^2 p \theta q \arg \max \theta \hat{\lambda}_{te} n^2 p \theta q \approx \hat{\theta}$.

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

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

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

Examples:

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

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

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

- 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.
NLP: discrimination of document categories

- 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:
  
  \[ \text{spike, markov, cortex, dropout, recurr, iii, gibb}. \]

  - 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 ($\approx \text{MMD-quad}$),

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

- \( \mathcal{V} = \{v_1, v_2\} \). Fix \( v_1 \) to \( \blacktriangle \).

- \( v_2 \mapsto \hat{\lambda}_{n/2}(\{v_1, v_2\}) \): contour plot.
Non-convexity, informative features

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

- **Small $J$:**
  - often enough to detect the difference of $\mathbb{P}$ & $\mathbb{Q}$.
  - few distinguishing regions to reject $H_0$.
  - faster test.
Number of locations \( (J) \)

- **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

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

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

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

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

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

(* in details:

\[ \langle \mu_\mathbb{P}, f \rangle_{\mathcal{H}(k)} = \langle \int k(\cdot, x)d\mathbb{P}(x), f \rangle_{\mathcal{H}(k)} \]
MMD: IPM representation

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

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

\((*)\) in details:

\[ \langle \mu_\mathbb{P}, f \rangle_{\mathcal{H}(k)} = \left\langle \int k(\cdot, x) d\mathbb{P}(x), f \right\rangle_{\mathcal{H}(k)} = \int \left\langle k(\cdot, x), f \right\rangle_{\mathcal{H}(k)} d\mathbb{P}(x) = f(x) \]
MMD: IPM representation

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

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

(*) in details:

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

\[
MMD^2(P, Q) = \| \mu_P - \mu_Q \|^2_H = \langle \mu_P - \mu_Q, \mu_P - \mu_Q \rangle_H
\]
\[
= \langle \mu_P, \mu_P \rangle_H + \langle \mu_Q, \mu_Q \rangle_H - 2 \langle \mu_P, \mu_Q \rangle_H
\]
\[
= \mathbb{E}_{P,P} k(x, x') + \mathbb{E}_{Q,Q} k(y, y') - 2\mathbb{E}_{P,Q} k(x, y).
\]
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 \(\{x_i\}_{i=1}^n \sim P, \{y_j\}_{j=1}^n \sim Q:\)

\[
\widehat{MMD^2}(P, Q) = \widehat{K}_{P,P} + \widehat{K}_{Q,Q} - 2\widehat{K}_{P,Q}.
\]
Proof idea

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

- Optimization & testing: linear in $n$.
- Testing: $O(ndJ + nJ^2 + J^3)$.
- Optimization: $O(ndJ^2 + J^3)$ per gradient ascent.


