A Linear-Time Kernel Goodness-of-Fit Test

Wittawat Jitkrittum\textsuperscript{1}  \quad Wenkai Xu\textsuperscript{1}  \quad Zoltán Szabó\textsuperscript{2}
Kenji Fukumizu\textsuperscript{3}  \quad Arthur Gretton\textsuperscript{1}

wittawat@gatsby.ucl.ac.uk

\textsuperscript{1}Gatsby Unit, University College London
\textsuperscript{2}CMAP, École Polytechnique
\textsuperscript{3}The Institute of Statistical Mathematics, Tokyo

NIPS 2017, Long Beach
5 December 2017
Model Criticism

Lake Michigan
Model Criticism

Data = robbery events in Chicago in 2016.
Model Criticism

Is this a good model?
Model Criticism

Goals:

1. Test if a (complicated) model fits the data.
2. If it does not, show a location where it fails.
Model Criticism

Goals:
1. Test if a (complicated) model fits the data.
2. If it does not, show a location where it fails.
Problem Setting: Goodness-of-Fit Test

Test goal: Are data from the model $p$?

1. Nonparametric.
2. Linear-time. Runtime is $O(n)$. Fast.
3. Interpretable. Model criticism by finding $F$. 

\[ q \quad \text{(unknown)} \quad \downarrow \quad ? \quad \sim \quad p \quad \text{(model)} \]

\[ x_1, x_2, \ldots, x_n \]
Problem Setting: Goodness-of-Fit Test

\[
\begin{align*}
q & \quad \text{(unknown)} \\
\downarrow & \\
\sim & \\
\{x_1, x_2, \ldots, x_n\} & \quad \text{(model)}
\end{align*}
\]

Test goal: Are data from the model \( p \)?

1. Nonparametric.
2. Linear-time. Runtime is \( \mathcal{O}(n) \). Fast.
3. Interpretable. Model criticism by finding \( \star \).
Test goal: Are data from the model \( p \)?

1. **Nonparametric.**
2. **Linear-time.** Runtime is \( \mathcal{O}(n) \). Fast.
3. **Interpretable.** Model criticism by finding \( \star \).
Model Criticism by Maximum Mean Discrepancy [Gretton et al., 2012]

- Find a location \( \mathbf{v} \) at which \( q \) and \( p \) differ most [Jitkrittum et al., 2016].
Model Criticism by Maximum Mean Discrepancy [Gretton et al., 2012]

- Find a location $v$ at which $q$ and $p$ differ most [Jitkrittum et al., 2016].

\[
\text{witness}(v) = \mathbb{E}_{x \sim q}[k_v(x)] - \mathbb{E}_{y \sim p}[k_v(y)]
\]
Model Criticism by Maximum Mean Discrepancy [Gretton et al., 2012]

- Find a location $v$ at which $q$ and $p$ differ most [Jitkrittum et al., 2016].

\[
witness(v) = \mathbb{E}_{x \sim q}[v] - \mathbb{E}_{y \sim p}[v]
\]
Model Criticism by Maximum Mean Discrepancy [Gretton et al., 2012]

- Find a location $v$ at which $q$ and $p$ differ most [Jitkrittum et al., 2016].

\[
\text{witness}(v) = \mathbb{E}_{x \sim q}[v] - \mathbb{E}_{y \sim p}[v]
\]

\[
\text{score}(v) = \frac{|\text{witness}(v)|}{\text{standard deviation}(v)}.
\]
Model Criticism by Maximum Mean Discrepancy \cite{Gretton2012}

- Find a location $v$ at which $q$ and $p$ differ most \cite{Jitkrittum2016}.

\[
\text{witness}(v) = E_{x \sim q}[v] - E_{y \sim p}[v] \\
\text{score}(v) = \frac{|\text{witness}(v)|}{\text{standard deviation}(v)}.
\]
Model Criticism by Maximum Mean Discrepancy [Gretton et al., 2012]

- Find a location $v$ at which $q$ and $p$ differ most [Jitkrittum et al., 2016].

**score: 0.008**

\[
k_v(x) = v
\]

\[
\text{witness}(v) = \mathbb{E}_{x \sim q}[-v] - \mathbb{E}_{y \sim p}[v]
\]

\[
\text{score}(v) = \frac{|\text{witness}(v)|}{\text{standard deviation}(v)}.
\]
Model Criticism by Maximum Mean Discrepancy [Gretton et al., 2012]

- Find a location $v$ at which $q$ and $p$ differ most [Jitkrittum et al., 2016].

**score: 1.6**

\[
\text{witness}(v) = \mathbb{E}_{x \sim q}[v] - \mathbb{E}_{y \sim p}[v]
\]

\[
\text{score}(v) = \frac{|\text{witness}(v)|}{\text{standard deviation}(v)}.
\]
Model Criticism by Maximum Mean Discrepancy [Gretton et al., 2012]

- Find a location $v$ at which $q$ and $p$ differ most [Jitkrittum et al., 2016].

\[
\text{score: 13}
\]

\[
\begin{align*}
\text{witness}(v) &= \mathbb{E}_{x \sim q}[v] - \mathbb{E}_{y \sim p}[v] \\
\text{score}(v) &= \frac{|\text{witness}(v)|}{\text{standard deviation}(v)}.
\end{align*}
\]
Model Criticism by Maximum Mean Discrepancy [Gretton et al., 2012]

- Find a location $v$ at which $q$ and $p$ differ most [Jitkrittum et al., 2016].

\[
\begin{align*}
\text{score}(v) &= \frac{|\text{witness}(v)|}{\text{standard deviation}(v)}.
\end{align*}
\]
Model Criticism by **Maximum Mean Discrepancy** [Gretton et al., 2012]

- Find a location \( v \) at which \( q \) and \( p \) differ most [Jitkrittum et al., 2016].

\[
\text{score}(v) = \frac{|\text{witness}(v)|}{\text{standard deviation}(v)}.
\]

\[
\text{witness}(v) = \mathbb{E}_{x \sim q}[v] - \mathbb{E}_{y \sim p}[v].
\]

No sample from \( p \). Difficult to generate.
The Stein Witness Function [Liu et al., 2016, Chwialkowski et al., 2016]

Problem: No sample from $p$. Cannot estimate $\mathbb{E}_{y \sim p}[k_{v}(y)]$. 
The Stein Witness Function [Liu et al., 2016, Chwialkowski et al., 2016]

Problem: No sample from \( p \). Cannot estimate \( \mathbb{E}_{y \sim p}[k_v(y)] \).

\[
(\text{Stein) witness}(v) = \mathbb{E}_{x \sim q}[T_p k_v(x)] - \mathbb{E}_{y \sim p}[T_p k_v(y)]
\]
The Stein Witness Function [Liu et al., 2016, Chwialkowski et al., 2016]

Problem: No sample from $p$. Cannot estimate $\mathbb{E}_{y \sim p}[k_v(y)]$.

$$(\text{Stein}) \text{ witness}(v) = \mathbb{E}_{x \sim q}[T_p v] - \mathbb{E}_{y \sim p}[T_p v]$$
Problem: No sample from $p$. Cannot estimate $\mathbb{E}_{y \sim p}[k_v(y)]$.

(Stein) witness$(v) = \mathbb{E}_{x \sim q}[\mathbb{E}_{y \sim p}[k_v(y)]] - \mathbb{E}_{y \sim p}[k_v(y)]$
The Stein Witness Function [Liu et al., 2016, Chwialkowski et al., 2016]

Problem: No sample from $p$. Cannot estimate $\mathbb{E}_{y \sim p}[k_v(y)]$.

(Stein) witness($v$) = $\mathbb{E}_{x \sim q}[v] - \mathbb{E}_{v \sim p}[v]

Idea: Define $T_p$ such that $\mathbb{E}_{y \sim p}(T_p k_v)(y) = 0$, for any $v$. 
The Stein Witness Function [Liu et al., 2016, Chwialkowski et al., 2016]

Problem: No sample from \( p \). Cannot estimate \( \mathbb{E}_{y \sim p}[k_v(y)] \).

(Stein) witness\( (v) = \mathbb{E}_{x \sim q}[v] \)

Idea: Define \( T_p \) such that \( \mathbb{E}_{y \sim p}(T_p k_v)(y) = 0 \), for any \( v \).
The Stein Witness Function [Liu et al., 2016, Chwialkowski et al., 2016]

Problem: No sample from $p$. Cannot estimate $\mathbb{E}_{y \sim p}[k_v(y)]$.

Idea: Define $T_p$ such that $\mathbb{E}_{x \sim q}[T_p k_v(x)] = 0$, for any $v$.

$$(\text{Stein})\ \text{witness}(v) = \mathbb{E}_{x \sim q}[T_p k_v(x)]$$
The Stein Witness Function \cite{Liuetal2016,Chwialkowskietal2016}

Problem: No sample from $p$. Cannot estimate $\mathbb{E}_{y \sim p}[k_v(y)]$.

(Stein) witness($v$) = $\mathbb{E}_{x \sim q}[T_p k_v(x)]$

Idea: Define $T_p$ such that $\mathbb{E}_{y \sim p}(T_p k_v)(y) = 0$, for any $v$.

Proposal: Good $v$ should have high

$$\text{score}(v) = \frac{|\text{witness}(v)|}{\text{standard deviation}(v)}.$$
The Stein Witness Function [Liu et al., 2016, Chwialkowski et al., 2016]

Problem: No sample from $p$. Cannot estimate $\mathbb{E}_{y \sim p}[k_v(y)]$.

Idea: Define $T_p$ such that $\mathbb{E}_{y \sim p}(T_p k_v)(y) = 0$, for any $v$.

$$\text{(Stein) witness}(v) = \mathbb{E}_{x \sim q}[T_p k_v(x)]$$

Proposal: Good $v$ should have high

$$\text{score}(v) = \frac{|\text{witness}(v)|}{\text{standard deviation}(v)}.$$
The Stein Witness Function [Liu et al., 2016, Chwialkowski et al., 2016]

Problem: No sample from \( p \). Cannot estimate \( \mathbb{E}_{y \sim p}[k_v(y)] \).

\[
(\text{Stein})\ \text{witness}(v) = \mathbb{E}_{x \sim q}[\ T_p k_v(x) \ ]
\]

Idea: Define \( T_p \) such that \( \mathbb{E}_{y \sim p}(T_p k_v)(y) = 0 \), for any \( v \).

Proposal: Good \( v \) should have high

\[
\text{score}(v) = \frac{|\text{witness}(v)|}{\text{standard deviation}(v)}.
\]

\[ \text{witness}(v) \text{ and standard deviation}(v) \text{ can be estimated in linear-time.} \]
Proposal: Model Criticism with the Stein Witness

\[ \text{score}(\mathbf{v}) = \frac{|\text{witness}(\mathbf{v})|}{\text{standard deviation}(\mathbf{v})}. \]
Proposal: Model Criticism with the Stein Witness

\[ \text{score}(v) = \frac{|\text{witness}(v)|}{\text{standard deviation}(v)}. \]
Proposal: Model Criticism with the Stein Witness

\[(T_p k_v)(x) = v\]

\[\text{score}(v) = \frac{|\text{witness}(v)|}{\text{standard deviation}(v)}.\]
score: 0.089

\[ \text{score}(\mathbf{v}) = \frac{|\text{witness}(\mathbf{v})|}{\text{standard deviation}(\mathbf{v})}. \]
Proposal: Model Criticism with the Stein Witness

score: 0.17

\[
score(v) = \frac{|\text{witness}(v)|}{\text{standard deviation}(v)}.\]
Proposal: Model Criticism with the Stein Witness

\[ \text{score: } 0.26 \]

\[ \text{score}(\mathbf{v}) = \frac{|\text{witness}(\mathbf{v})|}{\text{standard deviation}(\mathbf{v})}. \]
Proposal: Model Criticism with the Stein Witness

**score:** 0.33

\[
\text{score}(v) = \frac{|\text{witness}(v)|}{\text{standard deviation}(v)}.
\]
Proposal: Model Criticism with the Stein Witness

score: 0.37

\[
score(v) = \frac{|\text{witness}(v)|}{\text{standard deviation}(v)}.
\]
Proposal: Model Criticism with the Stein Witness

score: 0.37

\[ \text{score}(\mathbf{v}) = \frac{|\text{witness}(\mathbf{v})|}{\text{standard deviation}(\mathbf{v})}. \]
score: 0.45

\[ \text{score}(\mathbf{v}) = \frac{|\text{witness}(\mathbf{v})|}{\text{standard deviation}(\mathbf{v})}. \]
Proposal: Model Criticism with the Stein Witness

score: 0.44

\[
\text{score}(v) = \frac{|\text{witness}(v)|}{\text{standard deviation}(v)}.
\]
score: 0.39

$$\text{score}(v) = \frac{|\text{witness}(v)|}{\text{standard deviation}(v)}.$$
score: 0.31

\[
\text{score}(v) = \frac{|\text{witness}(v)|}{\text{standard deviation}(v)}.
\]
**Proposal: Model Criticism with the Stein Witness**

score: 0.32

\[
\text{score}(\mathbf{v}) = \frac{|\text{witness}(\mathbf{v})|}{\text{standard deviation}(\mathbf{v})}.
\]
Proposal: Model Criticism with the Stein Witness

\[
\text{score}(v) = \frac{|\text{witness}(v)|}{\text{standard deviation}(v)}.
\]
score: 0.37

$$\text{score}(v) = \frac{|\text{witness}(v)|}{\text{standard deviation}(v)}.$$
score: 0.48

\[
score(v) = \frac{|\text{witness}(v)|}{\text{standard deviation}(v)}.
\]
Proposal: Model Criticism with the Stein Witness

\[
\text{score}(v) = \frac{\text{witness}(v)}{\text{standard deviation}(v)}.
\]
score: 0.47

\[ \text{score}(v) = \frac{|\text{witness}(v)|}{\text{standard deviation}(v)}. \]
Proposal: Model Criticism with the Stein Witness

score: 0.44

\[
score(v) = \frac{|\text{witness}(v)|}{\text{standard deviation}(v)}.
\]
Proposal: Model Criticism with the Stein Witness

\[ \text{score}(v) = \frac{|\text{witness}(v)|}{\text{standard deviation}(v)}. \]
Proposal: Model Criticism with the Stein Witness

\[ \text{score}(v) = \frac{|\text{witness}(v)|}{\text{standard deviation}(v)}. \]
Proposal: Model Criticism with the Stein Witness

score: 0.16

$$\text{score}(\mathbf{v}) = \frac{|\text{witness}(\mathbf{v})|}{\text{standard deviation}(\mathbf{v})}.$$
score: 0.44

\[
\text{score}(v) = \frac{|\text{witness}(v)|}{\text{standard deviation}(v)}.
\]
What is $T_p k_v$?

Recall \( \text{witness}(v) = \mathbb{E}_{x \sim q}(T_p k_v)(x) - \mathbb{E}_{y \sim p}(T_p k_v)(y) \)
What is $T_p k_v$?

Recall $\text{witness}(v) = \mathbb{E}_{x \sim q(T_p k_v)(x)} - \mathbb{E}_{y \sim p(T_p k_v)(y)}$

$$(T_p k_v)(y) = \frac{1}{p(y)} \frac{d}{d y} [k_v(y) p(y)].$$

Then, $\mathbb{E}_{y \sim p(T_p k_v)(y)} = 0$.

[Liu et al., 2016, Chwialkowski et al., 2016]
What is $T_p k_v$?

Recall $\text{witness}(v) = \mathbb{E}_{x \sim q}(T_p k_v)(x) - \mathbb{E}_{y \sim p}(T_p k_v)(y)$

$$ (T_p k_v)(y) = \frac{1}{p(y)} \frac{d}{dy} [k_v(y)p(y)]. $$

Then, $\mathbb{E}_{y \sim p}(T_p k_v)(y) = 0$.

[Liu et al., 2016, Chwialkowski et al., 2016]
Technical Details

Theorem: Maximizing

\[ \text{score}(\mathbf{v}) = \frac{|\text{witness}(\mathbf{v})|}{\text{uncertainty}(\mathbf{v})} \]

- Increases true positive rate
  \[ = \Pr(\text{detect difference when } p \neq q), \]
- Does not affect false positive rate.

- General form of score(...) can consider more than one location \( \mathbf{v} \).
Technical Details

**Theorem:** Maximizing

\[
\text{score}(v) = \frac{|\text{witness}(v)|}{\text{uncertainty}(v)}
\]

- increases true positive rate
  \[= \mathbb{P}(\text{detect difference when } p \neq q),\]
- does not affect false positive rate.

- General form of score(...) can consider more than one location \(v\).
Technical Details

**Theorem:** Maximizing

\[
\text{score}(v) = \frac{\text{witness}(v)}{\text{uncertainty}(v)}
\]

- increases true positive rate
  \[= \mathbb{P}(\text{detect difference when } p \neq q),\]
- does not affect false positive rate.

- General form of \text{score}(\ldots) can consider more than one location \(v\).
Experiment: Restricted Boltzmann Machine (RBM)

Model $p =$

40 hidden units
50 visible units
Experiment: Restricted Boltzmann Machine (RBM)

- Model $p = \cdot \cdot \cdot$
- 40 hidden units
- 50 visible units

Perturb one weight

Sample from

9/11
Experiment: Restricted Boltzmann Machine (RBM)

40 hidden units
50 visible units

Model $p =$

Perturb one weight

Sample from

Sample size $n$

$P(\text{detect difference})$

MMD test (quadratic-time)

[Gretton et al., 2012]
Experiment: Restricted Boltzmann Machine (RBM)

Model $p = \cdots$

40 hidden units

50 visible units

Perturb one weight

Sample from

Sample size $n$

0.00

0.25

0.50

0.75

$P(\text{detect difference})$

Better

$\frac{9}{11}$

MMD test (quadratic-time)

[Gretton et al., 2012]

Proposed (linear-time)
Interpretable Features: Chicago Crime

Learned test locations are interpretable.
Interpretable Features: Chicago Crime

Learned test locations are interpretable.
Interpretable Features: Chicago Crime

- \( n = 11957 \) robbery events in Chicago in 2016.
  - lat/long coordinates = sample from \( q \).
- Model spatial density with Gaussian mixtures.
Interpretable Features: Chicago Crime

Model $p = 2$-component Gaussian mixture.
Interpretable Features: Chicago Crime

Score surface
Interpretable Features: Chicago Crime

$$F = \text{optimized } v.$$
Interpretable Features: Chicago Crime

\[ F = \text{optimized } v. \]

No robbery in Lake Michigan.
Interpretable Features: Chicago Crime

Model $p = 10$-component Gaussian mixture.
Interpretable Features: Chicago Crime

Capture the right tail better.
Still, does not capture the left tail.
Interpretable Features: Chicago Crime

Still, does not capture the left tail.

Learned test locations are interpretable.
Conclusions

Proposed a new goodness-of-fit test.

2. Linear-time
3. Interpretable

Poster #57 at Pacific Ballroom tonight.
Python code: https://github.com/wittawatj/kernel-gof
Questions?

Thank you
FSSD and KSD in 1D Gaussian Case

Consider $p = \mathcal{N}(0, 1)$ and $q = \mathcal{N}(\mu_q, \sigma_q^2)$.

- Assume $J = 1$ feature for $nFSSD^2$. Gaussian kernel (bandwidth $= \sigma_k^2$).

\[ FSSD^2 = \frac{\sigma_k^2 e^{-\frac{(v-\mu_q)^2}{\sigma_k^2+\sigma_q^2}}}{(\sigma_k^2+\sigma_q^2)^3} \left( (\sigma_k^2 + 1) \mu_q + v (\sigma_q^2 - 1) \right)^2. \]

- If $\mu_q \neq 0, \sigma_q^2 \neq 1$, and $v = -\frac{(\sigma_k^2+1)\mu_q}{(\sigma_q^2-1)}$, then $FSSD^2 = 0$!
  - This is why $v$ should be drawn from a distribution with a density.

- For KSD, Gaussian kernel (bandwidth $= \kappa^2$).

\[ S^2 = \frac{\mu_q^2 (\kappa^2 + 2\sigma_q^2) + (\sigma_q^2 - 1)^2}{(\kappa^2 + 2\sigma_q^2) \sqrt{\frac{2\sigma_q^2}{\kappa^2} + 1}}. \]
FSSD and KSD in 1D Gaussian Case

Consider \( p = \mathcal{N}(0, 1) \) and \( q = \mathcal{N}(\mu_q, \sigma_q^2) \).

- Assume \( J = 1 \) feature for \( nFSSD^2 \). Gaussian kernel (bandwidth = \( \sigma_k^2 \)).

\[
FSSD^2 = \frac{\sigma_k^2 e^{-\frac{(v-\mu_q)^2}{\sigma_k^2 + \sigma_q^2}} \left( (\sigma_k^2 + 1) \mu_q + v (\sigma_q^2 - 1) \right)^2}{(\sigma_k^2 + \sigma_q^2)^3}.
\]

- If \( \mu_q \neq 0, \sigma_q^2 \neq 1 \), and \( v = -\frac{(\sigma_k^2 + 1)\mu_q}{(\sigma_q^2 - 1)} \), then \( FSSD^2 = 0 \)!
  - This is why \( v \) should be drawn from a distribution with a density.

- For KSD, Gaussian kernel (bandwidth = \( \kappa^2 \)).

\[
S^2 = \frac{\mu_q^2 (\kappa^2 + 2\sigma_q^2) + (\sigma_q^2 - 1)^2}{(\kappa^2 + 2\sigma_q^2) \sqrt{\frac{2\sigma_q^2}{\kappa^2} + 1}}.
\]
What is $T_p k_v$?

Recall $\text{witness}(v) = \mathbb{E}_{x \sim q}(T_p k_v)(x) - \mathbb{E}_{y \sim p}(T_p k_v)(y)$
What is $T_p k_v$?

Recall $\text{witness}(v) = \mathbb{E}_{x \sim q} (T_p k_v)(x) - \mathbb{E}_{y \sim p} (T_p k_v)(y)$

$$(T_p k_v)(y) = \frac{1}{p(y)} \frac{d}{dy} [k(y, v)p(y)].$$

Then, $\mathbb{E}_{y \sim p} (T_p k_v)(y) = 0.$

[Liu et al., 2016, Chwialkowski et al., 2016]
What is $T_p k_v$?

Recall $\text{witness}(v) = \mathbb{E}_{x \sim q}(T_p k_v)(x) - \mathbb{E}_{y \sim p}(T_p k_v)(y)$

$$(T_p k_v)(y) = \frac{1}{p(y)} \frac{d}{dy}[k(y, v)p(y)].$$

Then, $\mathbb{E}_{y \sim p}(T_p k_v)(y) = 0$.

[Liu et al., 2016, Chwialkowski et al., 2016]
What is $T_p k_v$?

Recall $\text{witness}(v) = \mathbb{E}_{x \sim q}(T_p k_v)(x) - \mathbb{E}_{y \sim p}(T_p k_v)(y)$

$$(T_p k_v)(y) = \frac{1}{p(y)} \frac{d}{dy} [k(y, v)p(y)].$$

Then, $\mathbb{E}_{y \sim p}(T_p k_v)(y) = 0.$

[Chwialkowski et al., 2016, Liu et al., 2016]

Proof:

$$\mathbb{E}_{y \sim p} [(T_p k_v)(y)]$$
What is $T_p k_v$?

Recall \( \text{witness}(v) = \mathbb{E}_{x \sim q} (T_p k_v)(x) - \mathbb{E}_{y \sim p} (T_p k_v)(y) \)

\[
(T_p k_v)(y) = \frac{1}{p(y)} \frac{d}{dy} [k(y, v) p(y)].
\]

Then, \( \mathbb{E}_{y \sim p} (T_p k_v)(y) = 0. \)

[Liu et al., 2016, Chwialkowski et al., 2016]

Proof:

\[
\mathbb{E}_{y \sim p} [(T_p k_v)(y)] = \int_{-\infty}^{\infty} [(T_p k_v)(y)] p(y) dy
\]
What is $T_p k_v$?

Recall $\text{witness}(v) = \mathbb{E}_{x \sim q} (T_p k_v)(x) - \mathbb{E}_{y \sim p} (T_p k_v)(y)$

$$
(T_p k_v)(y) = \frac{1}{p(y)} \frac{d}{dy} [k(y, v) p(y)].
$$

Then, $\mathbb{E}_{y \sim p} (T_p k_v)(y) = 0$.

[Liu et al., 2016, Chwialkowski et al., 2016]

Proof:

$$
\mathbb{E}_{y \sim p} [(T_p k_v)(y)] = \int_{-\infty}^{\infty} \left[ \frac{1}{p(y)} \frac{d}{dy} [k_v(y) p(y)] \right] p(y) \, dy
$$
What is $T_p k_v$?

Recall $\text{witness}(v) = \mathbb{E}_{x \sim q}(T_p k_v)(x) - \mathbb{E}_{y \sim p}(T_p k_v)(y)$

$$(T_p k_v)(y) = \frac{1}{p(y)} \frac{d}{dy} [k(y, v)p(y)].$$

Then, $\mathbb{E}_{y \sim p}(T_p k_v)(y) = 0.$

[Liu et al., 2016, Chwialkowski et al., 2016]

Proof:

$$\mathbb{E}_{y \sim p}
[(T_p k_v)(y)] = \int_{-\infty}^{\infty} \left[ \frac{1}{p(y)} \frac{d}{dy} [k_v(y)p(y)] \right] p(y) dy$$
What is $T_p k_v$?

Recall \[ \text{witness}(v) = \mathbb{E}_{x \sim q}(T_p k_v)(x) - \mathbb{E}_{y \sim p}(T_p k_v)(y) \]

\[ (T_p k_v)(y) = \frac{1}{p(y)} \frac{d}{dy} [k(y, v)p(y)]. \]

Then, \[ \mathbb{E}_{y \sim p}(T_p k_v)(y) = 0. \]

[Liu et al., 2016, Chwialkowski et al., 2016]

Proof:

\[ \mathbb{E}_{y \sim p} [(T_p k_v)(y)] = \int_{-\infty}^{\infty} \left[ \frac{1}{p(y)} \frac{d}{dy} [k_v(y)p(y)] \right] p(y) dy \]

\[ = \int_{-\infty}^{\infty} \frac{d}{dy} [k_v(y)p(y)] dy \]
What is $T_p k_v$?

Recall $\text{witness}(v) = \mathbb{E}_{x \sim q}(T_p k_v)(x) - \mathbb{E}_{y \sim p}(T_p k_v)(y)$

$$(T_p k_v)(y) = \frac{1}{p(y)} \frac{d}{dy} [k(y, v) p(y)].$$

Normalizer cancels

Then, $\mathbb{E}_{y \sim p}(T_p k_v)(y) = 0$.

[Liu et al., 2016, Chwialkowski et al., 2016]

Proof:

$$\mathbb{E}_{y \sim p}[(T_p k_v)(y)] = \int_{-\infty}^{\infty} \left[ \frac{1}{p(y)} \frac{d}{dy} [k_v(y) p(y)] \right] p(y) \, dy$$

$$= \int_{-\infty}^{\infty} \frac{d}{dy} [k_v(y) p(y)] \, dy$$

$$= [k_v(y) p(y)]_{y=-\infty}^{y=\infty}$$
What is $T_p k_v$?

Recall $\text{witness}(v) = \mathbb{E}_{x \sim q}(T_p k_v)(x) - \mathbb{E}_{y \sim p}(T_p k_v)(y)$

$$(T_p k_v)(y) = \frac{1}{p(y)} \frac{d}{dy} [k(y, v) p(y)].$$

Then, $\mathbb{E}_{y \sim p}(T_p k_v)(y) = 0$.

[Liu et al., 2016, Chwialkowski et al., 2016]

Proof:

$$\mathbb{E}_{y \sim p} [(T_p k_v)(y)] = \int_{-\infty}^{\infty} \left[ \frac{1}{p(y)} \frac{d}{dy} [k_v(y)p(y)] \right] p(y) \, dy$$

$$= \int_{-\infty}^{\infty} \frac{d}{dy} [k_v(y)p(y)] \, dy$$

$$= [k_v(y)p(y)]_{y=-\infty}^{y=\infty}$$

$$= 0$$

(assume $\lim_{|y| \to \infty} k(y, v)p(y)$)
FSSD is a Discrepancy Measure

Theorem 1.

Let $V = \{v_1, \ldots, v_J\} \subset \mathbb{R}^d$ be drawn i.i.d. from a distribution $\eta$ which has a density. Let $\mathcal{X}$ be a connected open set in $\mathbb{R}^d$. Assume

1. (Nice RKHS) Kernel $k: \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ is $C_0$-universal, and real analytic.
2. (Stein witness not too rough) $\|g\|_{k}^2 < \infty$.
3. (Finite Fisher divergence) $\mathbb{E}_{x \sim q} \|\nabla_x \log \frac{p(x)}{q(x)}\|^2 < \infty$.
4. (Vanishing boundary) $\lim_{\|x\| \to \infty} p(x)g(x) = 0$.

Then, for any $J \geq 1$, $\eta$-almost surely

$$FSSD^2 = 0 \text{ if and only if } p = q.$$
What Are “Blind Spots”?

\[
g(v) : = \mathbb{E}_{x \sim q} \left[ \frac{1}{p(x)} \frac{d}{dx} [k_v(x)p(x)] \right] \\
= \mathbb{E}_{x \sim q} \left[ \left( \frac{d}{dx} \log p(x) \right) k_v(x) + \partial_x k_v(x) \right] \in \mathbb{R}^d.
\]

Consider \( p = \mathcal{N}(0, 1) \) and \( q = \mathcal{N}(0, \sigma_q^2) \). Use unit-width Gaussian kernel.

\[
g(v) = v \exp \left( -\frac{v^2}{2 + 2\sigma_q^2} \right) \left( \sigma_q^2 - 1 \right) \\
\quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \frac{1}{(1 + \sigma_q^2)^{3/2}}
\]

- If \( v = 0 \), then \( \text{FSSD}^2 = g^2(v) = 0 \) regardless of \( \sigma_q^2 \).
- If \( g \neq 0 \), and \( k \) is real analytic, \( R = \{ v \mid g(v) = 0 \} \) (blind spots) has 0 Lebesgue measure.
- So, if \( v \sim \) a distribution with a density, then \( v \notin R \).
What Are “blind spots”?

\[
g(v) : = \mathbb{E}_{x \sim q} \left[ \frac{1}{p(x)} \frac{d}{dx} [k_v(x)p(x)] \right] \\
= \mathbb{E}_{x \sim q} \left[ \left( \frac{d}{dx} \log p(x) \right) k_v(x) + \partial_x k_v(x) \right] \in \mathbb{R}^d.
\]

Consider \( p = \mathcal{N}(0, 1) \) and \( q = \mathcal{N}(0, \sigma_q^2) \). Use unit-width Gaussian kernel.

\[
g(v) = \frac{v \exp \left( -\frac{v^2}{2 + 2\sigma_q^2} \right) (\sigma_q^2 - 1)}{(1 + \sigma_q^2)^{3/2}}
\]

- If \( v = 0 \), then \( \text{FSSD}^2 = g^2(v) = 0 \) regardless of \( \sigma_q^2 \).
- If \( g \neq 0 \), and \( k \) is real analytic, \( R = \{ v \mid g(v) = 0 \} \) (blind spots) has 0 Lebesgue measure.
- So, if \( v \sim \) a distribution with a density, then \( v \notin R \).
What Are “Blind Spots”?  

\[
g(v) : = \mathbb{E}_{x \sim q} \left[ \frac{1}{p(x)} \frac{d}{dx} [k_v(x)p(x)] \right] \\
= \mathbb{E}_{x \sim q} \left[ \left( \frac{d}{dx} \log p(x) \right) k_v(x) + \partial_x k_v(x) \right] \in \mathbb{R}^d.
\]

Consider \( p = \mathcal{N}(0, 1) \) and \( q = \mathcal{N}(0, \sigma^2_q) \). Use unit-width Gaussian kernel.

\[
g(v) = v \exp \left( -\frac{v^2}{2 + 2\sigma^2_q} \right) \left( \sigma^2_q - 1 \right) \frac{1 + \sigma^2_q}{(1 + \sigma^2)^{3/2}}
\]

- If \( v = 0 \), then \( \text{FSSD}^2 = g^2(v) = 0 \) regardless of \( \sigma^2_q \).
- If \( g \neq 0 \), and \( k \) is real analytic, \( R = \{ v \mid g(v) = 0 \} \) (blind spots) has 0 Lebesgue measure.
- So, if \( v \sim \) a distribution with a density, then \( v \not\in R \).
What Are “Blind Spots”? 

\[ g(v) : = \mathbb{E}_{x \sim q} \left[ \frac{1}{p(x)} \frac{d}{dx} [k_v(x)p(x)] \right] \]

\[ = \mathbb{E}_{x \sim q} \left[ \left( \frac{d}{dx} \log p(x) \right) k_v(x) + \partial_x k_v(x) \right] \in \mathbb{R}^d. \]

Consider \( p = \mathcal{N}(0, 1) \) and \( q = \mathcal{N}(0, \sigma_q^2) \). Use unit-width Gaussian kernel.

\[ g(v) = \frac{v \exp \left( -\frac{v^2}{2 + 2\sigma_q^2} \right) (\sigma_q^2 - 1)}{(1 + \sigma_q^2)^{3/2}} \]

- If \( v = 0 \), then \( \text{FSSD}^2 = g^2(v) = 0 \) regardless of \( \sigma_q^2 \).
- If \( g \neq 0 \), and \( k \) is real analytic, \( R = \{ v \mid g(v) = 0 \} \) (blind spots) has 0 Lebesgue measure.
- So, if \( v \sim \) a distribution with a density, then \( v \notin R \).
What Are “Blind Spots”? 

\[
g(v) := \mathbb{E}_{x \sim q} \left[ \frac{1}{p(x)} \frac{d}{dx} [k_v(x)p(x)] \right] \\
= \mathbb{E}_{x \sim q} \left[ \left( \frac{d}{dx} \log p(x) \right) k_v(x) + \partial_x k_v(x) \right] \in \mathbb{R}^d.
\]

Consider \( p = \mathcal{N}(0, 1) \) and \( q = \mathcal{N}(0, \sigma_q^2) \). Use unit-width Gaussian kernel.

\[
g(v) = \frac{v \exp \left( -\frac{v^2}{2+2\sigma_q^2} \right) (\sigma_q^2 - 1)}{(1 + \sigma_q^2)^{3/2}}
\]

- If \( v = 0 \), then \( \text{FSSD}^2 = g^2(v) = 0 \) regardless of \( \sigma_q^2 \).
- If \( g \neq 0 \), and \( k \) is real analytic, \( R = \{ v \mid g(v) = 0 \} \) (blind spots) has 0 Lebesgue measure.
- So, if \( v \sim \) a distribution with a density, then \( v \notin R \).
Asymptotic Distributions of \( \text{FSSD}^2 \)

- Recall \( \xi(x, v) := \frac{1}{p(x)} \frac{d}{dx} [k(x, v)p(x)] \in \mathbb{R}^d \).
- \( \tau(x) := \) vertically stack \( \xi(x, v_1), \ldots \xi(x, v_J) \in \mathbb{R}^{dJ} \). Features of \( x \).
- Mean feature: \( \mu := \mathbb{E}_{x \sim q}[\tau(x)] \).
- \( \Sigma_r := \text{cov}_{x \sim r}[\tau(x)] \in \mathbb{R}^{dJ \times dJ} \) for \( r \in \{p, q\} \).

Proposition 1 (Asymptotic distributions).

Let \( Z_1, \ldots, Z_{dJ} \overset{i.i.d.}{\sim} \mathcal{N}(0, 1) \), and \( \{\omega_i\}_{i=1}^{dJ} \) be the eigenvalues of \( \Sigma_p \).

1. Under \( H_0 : p = q \), asymptotically \( n \text{FSSD}^2 \overset{d}{\to} \sum_{i=1}^{dJ} (Z_i^2 - 1)\omega_i \).
   - Simulation cost independent of \( n \).

2. Under \( H_1 : p \neq q \), we have \( \sqrt{n} (\text{FSSD}^2 - \text{FSSD}^2) \overset{d}{\to} \mathcal{N}(0, \sigma^2_{H_1}) \)
   where \( \sigma^2_{H_1} := 4\mu^\top \Sigma_q \mu \). Implies \( \mathbb{P}(\text{reject } H_0) \to 1 \) as \( n \to \infty \).

But, how to estimate \( \Sigma_p \)? No sample from \( p \)!

- Theorem: Using \( \hat{\Sigma}_q \) (computed with \( \{x_i\}_{i=1}^n \sim q \)) still leads to a consistent test.
Asymptotic Distributions of $FSSD^2$

- Recall $\xi(x, v) := \frac{1}{p(x)} \frac{d}{dx}[k(x, v)p(x)] \in \mathbb{R}^d$.
- $\tau(x) := \text{vertically stack } \xi(x, v_1), \ldots \xi(x, v_J) \in \mathbb{R}^{dJ}$. Features of $x$.
- Mean feature: $\mu := \mathbb{E}_{x \sim q}[\tau(x)]$.
- $\Sigma_r := \text{cov}_{x \sim r}[\tau(x)] \in \mathbb{R}^{dJ \times dJ}$ for $r \in \{p, q\}$

Proposition 1 (Asymptotic distributions).

Let $Z_1, \ldots, Z_{dJ} \overset{i.i.d.}{\sim} \mathcal{N}(0, 1)$, and $\{\omega_i\}_{i=1}^{dJ}$ be the eigenvalues of $\Sigma_p$.

1. Under $H_0 : p = q$, asymptotically $nFSSD^2 \xrightarrow{d} \sum_{i=1}^{dJ}(Z_i^2 - 1)\omega_i$.
   - Simulation cost independent of $n$.

2. Under $H_1 : p \neq q$, we have $\sqrt{n}(\text{FSSD}^2 - \text{FSSD}^2) \xrightarrow{d} \mathcal{N}(0, \sigma_{H_1}^2)$ where $\sigma_{H_1}^2 := 4\mu^\top \Sigma_q \mu$. Implies $\mathbb{P}(\text{reject } H_0) \rightarrow 1$ as $n \rightarrow \infty$.

But, how to estimate $\Sigma_p$? No sample from $p$!

- Theorem: Using $\widehat{\Sigma}_q$ (computed with $\{x_i\}_{i=1}^n \sim q$) still leads to a consistent test.
Asymptotic Distributions of $\text{FSSD}^2$

Recall $\xi(x, v) := \frac{1}{p(x)} \frac{d}{dx}[k(x, v)p(x)] \in \mathbb{R}^d$.

$\tau(x) := \text{vertically stack } \xi(x, v_1), \ldots \xi(x, v_J) \in \mathbb{R}^{dJ}$. Features of $x$.

Mean feature: $\mu := \mathbb{E}_{x \sim q}[\tau(x)]$.

$\Sigma_r := \text{cov}_{x \sim r}[\tau(x)] \in \mathbb{R}^{dJ \times dJ}$ for $r \in \{p, q\}$

Proposition 1 (Asymptotic distributions).

Let $Z_1, \ldots, Z_{dJ} \overset{i.i.d.}{\sim} \mathcal{N}(0, 1)$, and $\{\omega_i\}_{i=1}^{dJ}$ be the eigenvalues of $\Sigma_p$.

1. Under $H_0 : p = q$, asymptotically $n\text{FSSD}^2 \xrightarrow{d} \sum_{i=1}^{dJ}(Z_i^2 - 1)\omega_i$.
   
   - Simulation cost independent of $n$.

2. Under $H_1 : p \neq q$, we have $\sqrt{n}(\text{FSSD}^2 - \text{FSSD}^2) \xrightarrow{d} \mathcal{N}(0, \sigma_{H_1}^2)$ where $\sigma_{H_1}^2 := 4\mu^\top\Sigma_q\mu$. Implies $\mathbb{P}(\text{reject } H_0) \to 1$ as $n \to \infty$.

But, how to estimate $\Sigma_p$? No sample from $p$!

Theorem: Using $\hat{\Sigma}_q$ (computed with $\{x_i\}_{i=1}^n \sim q$) still leads to a consistent test.
Asymptotic Distributions of $\text{FSSD}^2$

- Recall $\xi(x, v) := \frac{1}{p(x)} \frac{d}{dx} [k(x, v)p(x)] \in \mathbb{R}^d$.
- $\tau(x) := \text{vertically stack } \xi(x, v_1), \ldots, \xi(x, v_J) \in \mathbb{R}^{dJ}$. Features of $x$.
- Mean feature: $\mu := \mathbb{E}_{x \sim q} [\tau(x)]$.
- $\Sigma_r := \text{cov}_{x \sim r} [\tau(x)] \in \mathbb{R}^{dJ \times dJ}$ for $r \in \{p, q\}$

**Proposition 1 (Asymptotic distributions).**

Let $Z_1, \ldots, Z_{dJ} \overset{i.i.d.}{\sim} \mathcal{N}(0, 1)$, and $\{\omega_i\}_{i=1}^{dJ}$ be the eigenvalues of $\Sigma_p$.

1. Under $H_0: p = q$, asymptotically $n \text{FSSD}^2 \xrightarrow{d} \sum_{i=1}^{dJ} (Z_i^2 - 1) \omega_i$.
   - Simulation cost independent of $n$.

2. Under $H_1: p \neq q$, we have $\sqrt{n}(\text{FSSD}^2 - \text{FSSD}^2) \xrightarrow{d} \mathcal{N}(0, \sigma_{H_1}^2)$
   where $\sigma_{H_1}^2 := 4\mu^\top \Sigma_q \mu$. Implies $\mathbb{P}(\text{reject } H_0) \rightarrow 1$ as $n \rightarrow \infty$.

But, how to estimate $\Sigma_p$? No sample from $p$!

- **Theorem:** Using $\hat{\Sigma}_q$ (computed with $\{x_i\}_{i=1}^n \sim q$) still leads to a consistent test.
Bahadur Slope and Bahadur Efficiency

- Bahadur slope $\approx$ rate of p-value $\to 0$ under $H_1$ as $n \to \infty$.
- Measure a test’s sensitivity to the departure from $H_0$.

\[ H_0: \theta = 0, \]
\[ H_1: \theta \neq 0. \]

- Typically $pval_n \approx \exp \left( -\frac{1}{2} c(\theta)n \right)$ where $c(\theta) > 0$ under $H_1$, and $c(0) = 0$ [Bahadur, 1960].
- $c(\theta)$ higher $\implies$ more sensitive. Good.

Bahadur slope

\[ c(\theta) := -2 \lim_{n \to \infty} \frac{\log (1 - F(T_n))}{n}, \]

where $F(t) = \text{CDF of } T_n$ under $H_0$.

- Bahadur efficiency $= \text{ratio of slopes of two tests.}$
Bahadur Slope and Bahadur Efficiency

- Bahadur slope $\approx$ rate of p-value $\to 0$ under $H_1$ as $n \to \infty$.
- Measure a test’s sensitivity to the departure from $H_0$.

\[ H_0: \theta = 0, \]
\[ H_1: \theta \neq 0. \]

- Typically $pval_n \approx \exp \left(-\frac{1}{2}c(\theta)n\right)$ where $c(\theta) > 0$ under $H_1$, and $c(0) = 0$ [Bahadur, 1960].
- $c(\theta)$ higher $\implies$ more sensitive. Good.

Bahadur slope

\[ c(\theta) := -2 \lim_{n \to \infty} \frac{\log (1 - F(T_n))}{n}, \]

where $F(t) = \text{CDF of } T_n$ under $H_0$.

- Bahadur efficiency = ratio of slopes of two tests.
Bahadur Slope and Bahadur Efficiency

- Bahadur slope $\approx$ rate of p-value $\to 0$ under $H_1$ as $n \to \infty$.
- Measure a test’s sensitivity to the departure from $H_0$.
  
  $H_0: \theta = 0,$
  
  $H_1: \theta \neq 0.$

- Typically $pval_n \approx \exp\left(-\frac{1}{2} c(\theta)n\right)$ where $c(\theta) > 0$ under $H_1$, and $c(0) = 0$ [Bahadur, 1960].
- $c(\theta)$ higher $\implies$ more sensitive. Good.

Bahadur slope

\[ c(\theta) := -2 \lim_{n \to \infty} \frac{\log(1 - F(T_n))}{n}, \]

where $F'(t) =$ CDF of $T_n$ under $H_0$.

- Bahadur efficiency = ratio of slopes of two tests.
Bahadur Slope and Bahadur Efficiency

- Bahadur slope \( \approx \) rate of p-value \( \to 0 \) under \( H_1 \) as \( n \to \infty \).
- Measure a test’s sensitivity to the departure from \( H_0 \).

\[
H_0 : \theta = 0, \\
H_1 : \theta \neq 0.
\]

- Typically \( pval_n \approx \exp\left(-\frac{1}{2}c(\theta)n\right) \) where \( c(\theta) > 0 \) under \( H_1 \), and \( c(0) = 0 \) [Bahadur, 1960].
- \( c(\theta) \) higher \( \implies \) more sensitive. Good.

![Bahadur slope graph](image)

Bahadur slope

\[
c(\theta) := -2 \lim_{n \to \infty} \frac{\log (1 - F(T_n))}{n},
\]

where \( F'(t) = \text{CDF of } T_n \text{ under } H_0 \).

- Bahadur efficiency = ratio of slopes of two tests.
Gaussian Mean Shift Problem

Consider $p = \mathcal{N}(0, 1)$ and $q = \mathcal{N}(\mu_q, 1)$.

- Assume $J = 1$ location for $n\text{FSSD}^2$. Gaussian kernel (bandwidth $= \sigma_k^2$)

  \[
  c^{(\text{FSSD})}(\mu_q, v, \sigma_k^2) = \frac{\sigma_k^2 \left( \sigma_k^2 + 2 \right)^3 \mu_q^2 e^{\frac{v^2}{\sigma_k^2 + v^2}} - \frac{(v - \mu_q)^2}{\sigma_k^2 + 1}}{\sqrt{\frac{2}{\sigma_k^2} + 1 \left( \sigma_k^2 + 1 \right) \left( \sigma_k^6 + 4\sigma_k^4 + (v^2 + 5)\sigma_k^2 + 2 \right)}}.
  \]

- For LKS, Gaussian kernel (bandwidth $= \kappa^2$).

  \[
  c^{(\text{LKS})}(\mu_q, \kappa^2) = \frac{(\kappa^2)^{5/2} (\kappa^2 + 4)^{5/2} \mu_q^4}{2 (\kappa^2 + 2) (\kappa^8 + 8\kappa^6 + 21\kappa^4 + 20\kappa^2 + 12)}.
  \]

**Theorem 2 (FSSD is at least two times more efficient).**

*Fix $\sigma_k^2 = 1$ for $n\text{FSSD}^2$. Then, $\forall \mu_q \neq 0, \exists v \in \mathbb{R}, \forall \kappa^2 > 0$, we have Bahadur efficiency*

\[
\frac{c^{(\text{FSSD})}(\mu_q, v, \sigma_k^2)}{c^{(\text{LKS})}(\mu_q, \kappa^2)} > 2.
\]
Gaussian Mean Shift Problem

Consider \( p = \mathcal{N}(0, 1) \) and \( q = \mathcal{N}(\mu_q, 1) \).

- Assume \( J = 1 \) location for \( nFSSD^2 \). Gaussian kernel (bandwidth = \( \sigma_k^2 \))

\[
c^{(FSSD)}(\mu_q, \nu, \sigma_k^2) = \frac{\sigma_k^2 (\sigma_k^2 + 2)^3 \mu_q^2 e^{\frac{\nu^2}{\sigma_k^2 + 2}} - \frac{(\nu - \mu_q)^2}{\sigma_k^2 + 1}}{\sqrt{\frac{2}{\sigma_k^2 + 1} (\sigma_k^2 + 1) (\sigma_k^2 + 4\sigma_k^4 + (\nu^2 + 5)\sigma_k^2 + 2)}}.
\]

- For LKS, Gaussian kernel (bandwidth = \( \kappa^2 \)).

\[
c^{(LKS)}(\mu_q, \kappa^2) = \frac{(\kappa^2)^{5/2} (\kappa^2 + 4)^{5/2} \mu_q^4}{2 (\kappa^2 + 2) (\kappa^8 + 8\kappa^6 + 21\kappa^4 + 20\kappa^2 + 12)}.
\]

**Theorem 2 (FSSD is at least two times more efficient).**

Fix \( \sigma_k^2 = 1 \) for \( nFSSD^2 \). Then, \( \forall \mu_q \neq 0, \exists \nu \in \mathbb{R}, \forall \kappa^2 > 0 \), we have Bahadur efficiency

\[
\frac{c^{(FSSD)}(\mu_q, \nu, \sigma_k^2)}{c^{(LKS)}(\mu_q, \kappa^2)} > 2.
\]
Gaussian Mean Shift Problem

Consider $p = \mathcal{N}(0, 1)$ and $q = \mathcal{N}(\mu_q, 1)$.

- Assume $J = 1$ location for $n\text{FSSD}^2$. Gaussian kernel (bandwidth $= \sigma_k^2$)
  
  $$c^{(\text{FSSD})}(\mu_q, \nu, \sigma_k^2) = \frac{\sigma_k^2 (\sigma_k^2 + 2)^3 \mu_q^2 e^{\frac{\nu^2}{\sigma_k^2+2}} - \frac{(\nu-\mu_q)^2}{\sigma_k^2+1}}{\sqrt{\frac{2}{\sigma_k^2} + 1 (\sigma_k^2 + 1) (\sigma_k^6 + 4\sigma_k^4 + (\nu^2 + 5) \sigma_k^2 + 2)}}.$$  

- For LKS, Gaussian kernel (bandwidth $= \kappa^2$).
  
  $$c^{(\text{LKS})}(\mu_q, \kappa^2) = \frac{(\kappa^2)^{5/2} (\kappa^2 + 4)^{5/2} \mu_q^4}{2 (\kappa^2 + 2) (\kappa^8 + 8\kappa^6 + 21\kappa^4 + 20\kappa^2 + 12)}.$$  

Theorem 2 (FSSD is at least two times more efficient).

Fix $\sigma_k^2 = 1$ for $n\text{FSSD}^2$. Then, $\forall \mu_q \neq 0$, $\exists \nu \in \mathbb{R}$, $\forall \kappa^2 > 0$, we have Bahadur efficiency

$$\frac{c^{(\text{FSSD})}(\mu_q, \nu, \sigma_k^2)}{c^{(\text{LKS})}(\mu_q, \kappa^2)} > 2.$$
[Liu et al., 2016] also proposed a linear version of KSD. For \( \{x_i\}_{i=1}^{n} \sim q \), KSD test statistic is

\[
\frac{2}{n(n-1)} \sum_{i<j} h_p(x_i, x_j).
\]

LKS test statistic is a “running average”

\[
\frac{2}{n} \sum_{i=1}^{n/2} h_p(x_{2i-1}, x_{2i}).
\]

Both unbiased. LKS has \( \mathcal{O}(d^2n) \) runtime.

\( \times \) LKS has high variance. Poor test power.
Bahadur Slopes of FSSD and LKS

Theorem 3.

The Bahadur slope of $n^{\text{FSSD}^2}$ is

$$c^{(\text{FSSD})} := \frac{\text{FSSD}^2}{\omega_1},$$

where $\omega_1$ is the maximum eigenvalue of $\Sigma_p := \text{cov}_{x \sim p}[\tau(x)]$.

The Bahadur slope of the linear-time kernel Stein (LKS) statistic $\sqrt{nS_i^2}$ is

$$c^{(\text{LKS})} = \frac{1}{2} \frac{[\mathbb{E}_q h_p(x, x')]^2}{\mathbb{E}_p \left[h_p^2(x, x')\right]}',$$

where $h_p$ is the U-statistic kernel of the KSD statistic.
Consider $J = 1$ location.

Training objective $\frac{\hat{\text{FSSD}}_2(v)}{\hat{\sigma}_{H_1}(v)}$ (gray), $p$ in wireframe, $\{x_i\}_{i=1}^n \sim q$ in purple, $\star = \text{best } v$.

$$p = \mathcal{N} \left( 0, \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix} \right) \quad \text{vs.} \quad q = \mathcal{N} \left( 0, \begin{pmatrix} 2 & 0 \\ 0 & 1 \end{pmatrix} \right).$$
Illustration: Optimization Objective

- Consider $J = 1$ location.
- Training objective $\frac{\text{FSSD}^2(v)}{\sigma_{H_1}(v)}$ (gray), $p$ in wireframe, $\{x_i\}_{i=1}^n \sim q$ in purple, $\star$ = best $v$.

$$p = \mathcal{N}(0, I) \text{ vs. } q = \text{Laplace with same mean & variance}.$$
## Simulation Settings

- Gaussian kernel \( k(x, v) = \exp \left( -\frac{||x-v||^2}{2\sigma_k^2} \right) \)

<table>
<thead>
<tr>
<th>Method</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>1 FSSD-opt</td>
<td>Proposed. With optimization. ( J = 5 ).</td>
</tr>
<tr>
<td>2 FSSD-rand</td>
<td>Proposed. Random test locations.</td>
</tr>
<tr>
<td>3 KSD</td>
<td>Quadratic-time kernel Stein discrepancy [Liu et al., 2016, Chwialkowski et al., 2016]</td>
</tr>
<tr>
<td>4 LKS</td>
<td>Linear-time running average version of KSD.</td>
</tr>
<tr>
<td>5 MMD-opt</td>
<td>MMD two-sample test [Gretton et al., 2012]. With optimization.</td>
</tr>
<tr>
<td>6 ME-test</td>
<td>Mean Embeddings two-sample test [Jitkrittum et al., 2016]. With optimization.</td>
</tr>
</tbody>
</table>

- Two-sample tests need to draw sample from \( p \).
- Tests with optimization use 20% of the data.
- \( \alpha = 0.05 \). 200 trials.
## Simulation Settings

- Gaussian kernel $k(x, v) = \exp \left( -\frac{||x-v||^2}{2\sigma_k^2} \right)$

<table>
<thead>
<tr>
<th>Method</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>2</td>
<td><strong>FSSD-rand</strong> Proposed. Random test locations.</td>
</tr>
<tr>
<td>3</td>
<td><strong>KSD</strong> Quadratic-time kernel Stein discrepancy</td>
</tr>
<tr>
<td></td>
<td>[Liu et al., 2016, Chwialkowski et al., 2016]</td>
</tr>
<tr>
<td>4</td>
<td><strong>LKS</strong> Linear-time running average version of KSD.</td>
</tr>
<tr>
<td>5</td>
<td><strong>MMD-opt</strong> MMD two-sample test [Gretton et al., 2012]. With optimization.</td>
</tr>
<tr>
<td>6</td>
<td><strong>ME-test</strong> Mean Embeddings two-sample test</td>
</tr>
<tr>
<td></td>
<td>[Jitkrittum et al., 2016]. With optimization.</td>
</tr>
</tbody>
</table>

- Two-sample tests need to draw sample from $p$.
- Tests with optimization use 20% of the data.
- $\alpha = 0.05$. 200 trials.
## Simulation Settings

- Gaussian kernel \( k(x, v) = \exp \left( -\frac{\|x - v\|^2}{2\sigma_k^2} \right) \)

<table>
<thead>
<tr>
<th>Method</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>1 FSSD-opt</td>
<td>Proposed. With optimization. ( J = 5 ).</td>
</tr>
<tr>
<td>2 FSSD-rand</td>
<td>Proposed. Random test locations.</td>
</tr>
<tr>
<td>3 KSD</td>
<td>Quadratic-time kernel Stein discrepancy</td>
</tr>
<tr>
<td></td>
<td>[Liu et al., 2016, Chwialkowski et al., 2016]</td>
</tr>
<tr>
<td>4 LKS</td>
<td>Linear-time running average version of KSD.</td>
</tr>
<tr>
<td>5 MMD-opt</td>
<td>MMD two-sample test [Gretton et al., 2012]. With</td>
</tr>
<tr>
<td></td>
<td>optimization.</td>
</tr>
<tr>
<td>6 ME-test</td>
<td>Mean Embeddings two-sample test</td>
</tr>
<tr>
<td></td>
<td>[Jitkrittum et al., 2016]. With optimization.</td>
</tr>
</tbody>
</table>

- Two-sample tests need to draw sample from \( p \).
- Tests with optimization use 20% of the data.
- \( \alpha = 0.05 \). 200 trials.
**Gaussian Vs. Laplace**

- \( p = \text{Gaussian} \). \( q = \text{Laplace} \). Same mean and variance. High-order moments differ.
- Sample size \( n = 1000 \).

![Graph showing rejection rate vs. dimension](diagram.png)

- Optimization increases the power.
- Two-sample tests can perform well in this case (\( p, q \) clearly differ).
**Harder RBM Problem**

- Perturb only one entry of $\mathbf{B} \in \mathbb{R}^{50 \times 40}$ (in the RBM).
- $B_{1,1} \leftarrow B_{1,1} + \mathcal{N}(0, \sigma_{\text{per}}^2 = 0.1^2)$.

Two-sample tests fail. Samples from $p, q$ look roughly the same.

FSSD-opt is comparable to KSD at low $n$. One order of magnitude faster.
Harder RBM Problem

- Perturb only one entry of $B \in \mathbb{R}^{50 \times 40}$ (in the RBM).
- $B_{1,1} \leftarrow B_{1,1} + \mathcal{N}(0, \sigma_{\text{per}}^2 = 0.1^2)$.

Two-sample tests fail. Samples from $p, q$ look roughly the same.
- FSSD-opt is comparable to KSD at low $n$. One order of magnitude faster.

