A Linear-Time Kernel Goodness-of-Fit Test

Wittawat Jitkrittum[1,*] Wenkai Xu[1] Zoltán Szabó[2]
Kenji Fukumizu[3] Arthur Gretton[1]

NIPS 2017
Best paper!



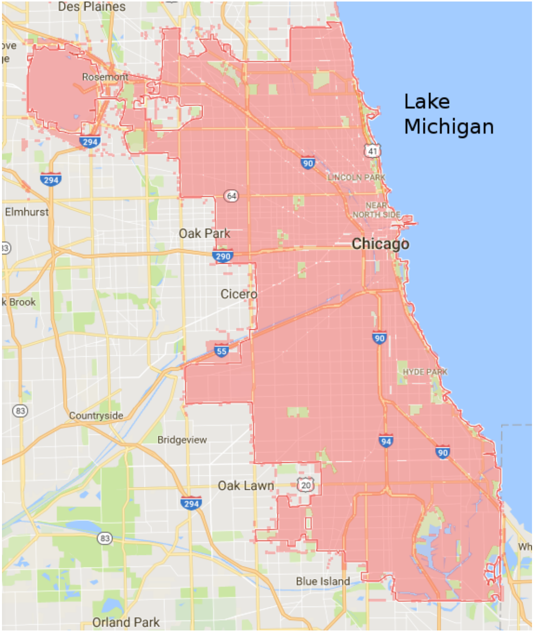


wittawatj@gmail.com

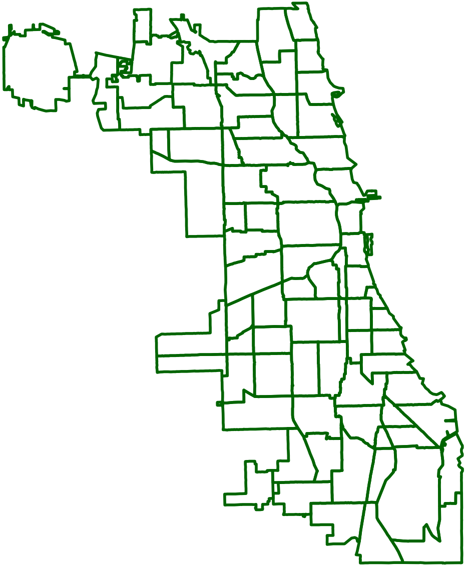
[1]Gatsby Unit, University College London
*(Now at Max Planck Institute for Intelligent Systems)
[2]CMAP, École Polytechnique
[3]The Institute of Statistical Mathematics, Tokyo

Workshop on Functional Inference and Machine Intelligence, Tokyo
20 February 2018

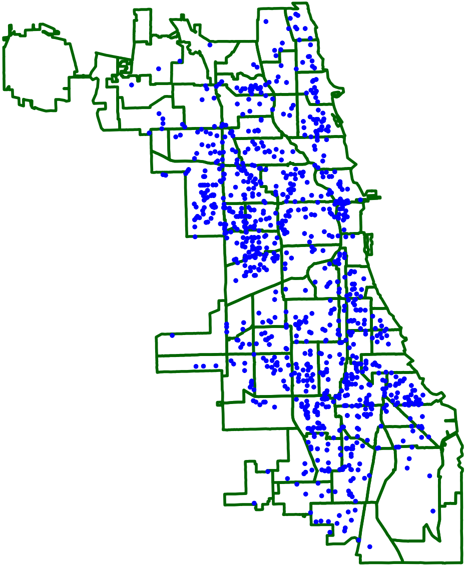
Model Criticism



Model Criticism

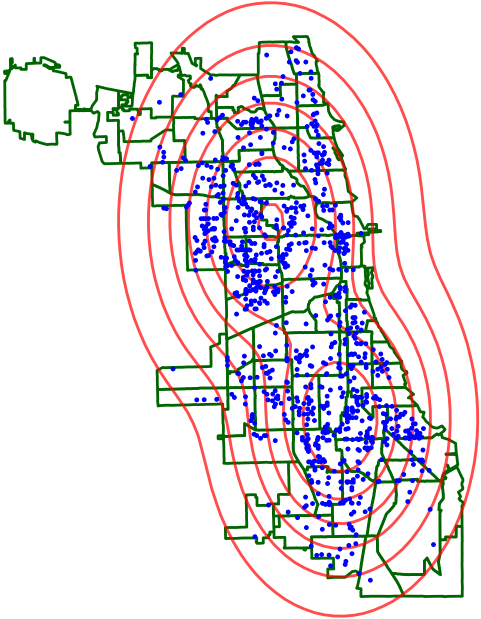


Model Criticism



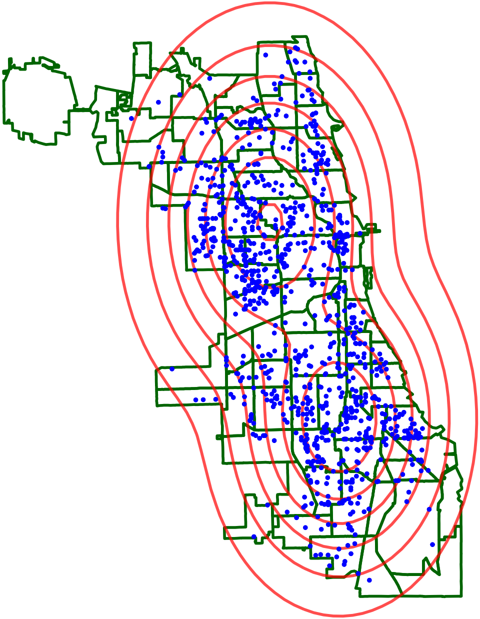
Data = robbery events in Chicago in 2016.

Model Criticism



Is this a good model?

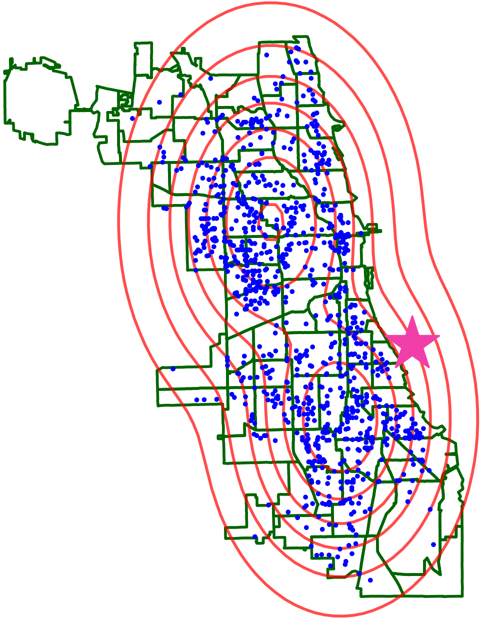
Model Criticism



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.

Goodness-of-fit Testing

Given:

1. Sample $\{\mathbf{x}_i\}_{i=1}^{n} \overset{i.i.d.}{\sim} q$ (unknown) on $\mathbb{R}^d$,
2. Unnormalized density p (known model).

$$H_0 : p = q$$
$$H_1 : p \neq q$$

Want a test . . .

1. Nonparametric.
2. Linear-time. Runtime is $\mathcal{O}(n)$. Fast.
3. Interpretable. Model criticism by finding ★.

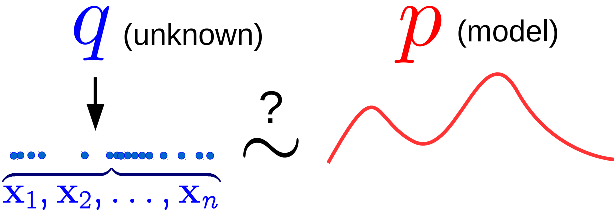
Goodness-of-fit Testing

Given:

1. Sample $\{\mathbf{x}_i\}_{i=1}^n \overset{i.i.d.}{\sim} q$ (unknown) on $\mathbb{R}^d$,
2. Unnormalized density p (known model).

$H_0 : p = q$

$H_1 : p \neq q$



Want a test . . .

1. Nonparametric.
2. Linear-time. Runtime is $\mathcal{O}(n)$. Fast.
3. Interpretable. Model criticism by finding ★.

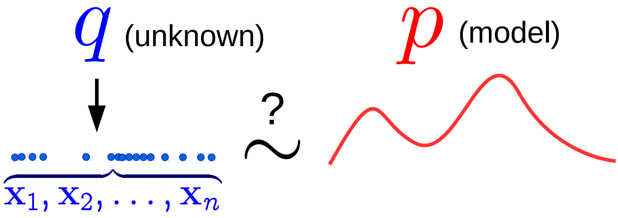
Goodness-of-fit Testing

Given:

1. Sample $\{\mathbf{x}_i\}_{i=1}^n \overset{i.i.d.}{\sim} q$ (unknown) on $\mathbb{R}^d$,
2. Unnormalized density p (known model).

$H_0: p = q$

$H_1: p \neq q$



Want a test ...

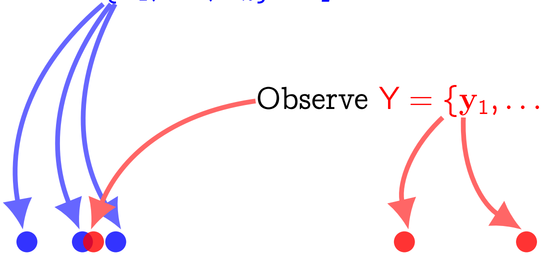
1. Nonparametric.
2. Linear-time. Runtime is $\mathcal{O}(n)$. Fast.
3. Interpretable. Model criticism by finding ★.

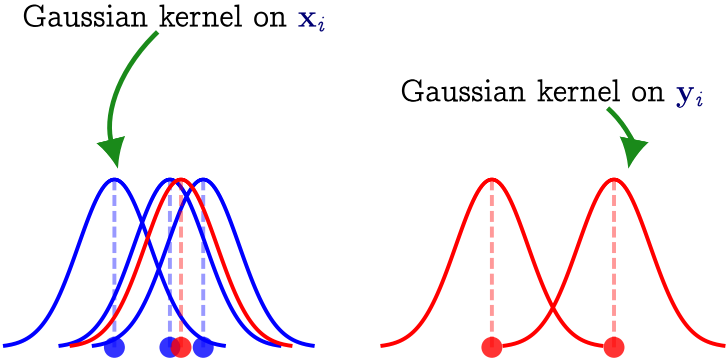
Maximum Mean Discrepancy (MMD) Witness Function (Gretton et al., 2012)

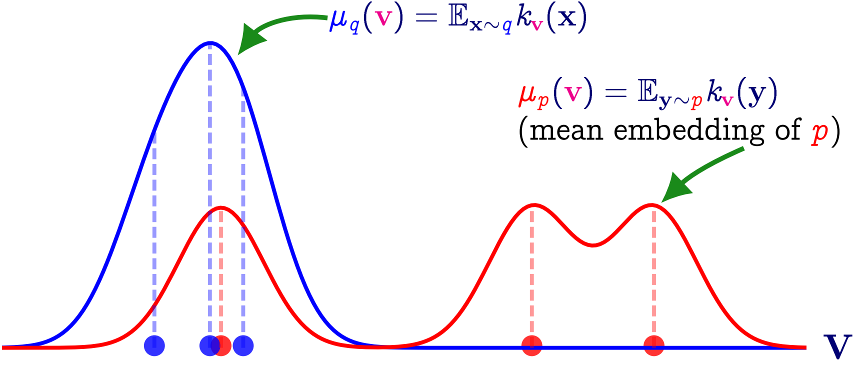
Maximum Mean Discrepancy (MMD) Witness Function (Gretton et al., 2012)

Observe $\mathsf{X} = \{\mathbf{x}_1, \ldots, \mathbf{x}_n\} \sim q$

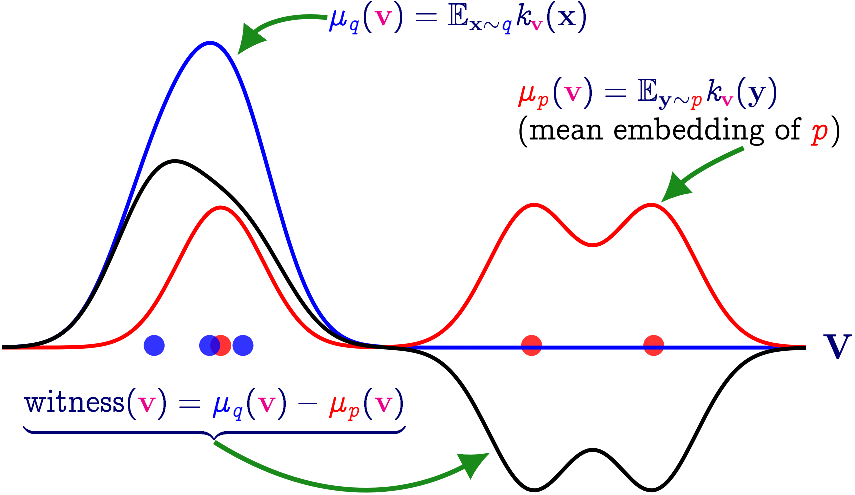
Observe $\mathsf{Y} = \{\mathbf{y}_1, \ldots, \mathbf{y}_n\} \sim p$



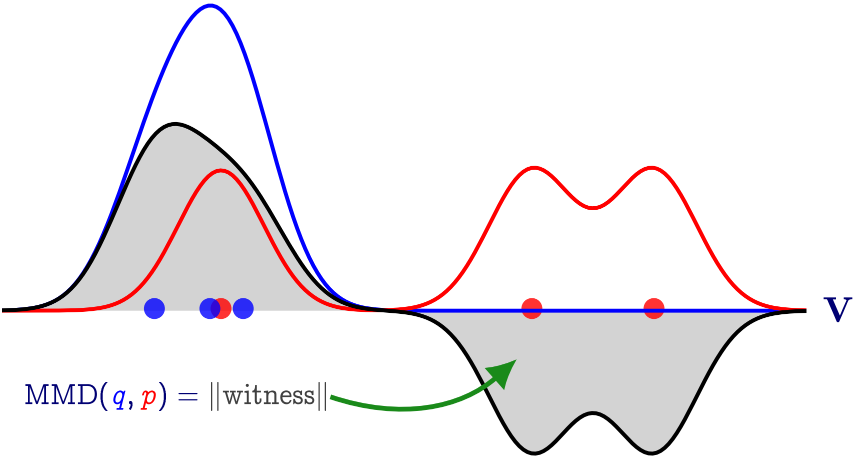

Gaussian kernel on $\mathbf{x}_i$
Gaussian kernel on $\mathbf{y}_i$


$\mu_q(\mathbf{v}) = \mathbb{E}_{\mathbf{x}\sim q} k_{\mathbf{v}}(\mathbf{x})$
$\mu_p(\mathbf{v}) = \mathbb{E}_{\mathbf{y}\sim p} k_{\mathbf{v}}(\mathbf{y})$
(mean embedding of p)
$\mathbf{V}$

Maximum Mean Discrepancy (MMD) Witness Function (Gretton et al., 2012)

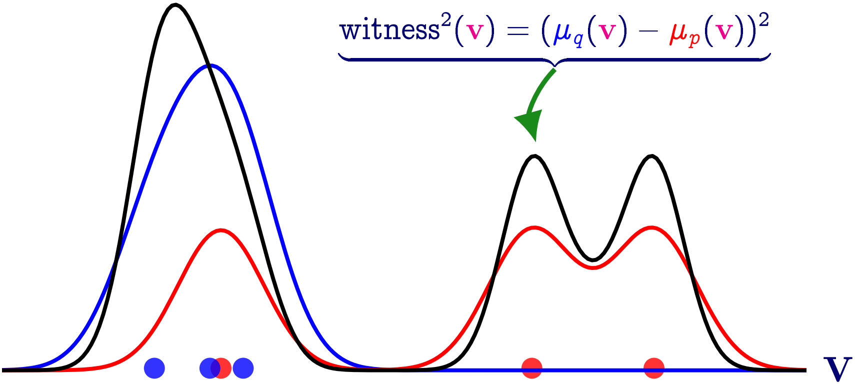


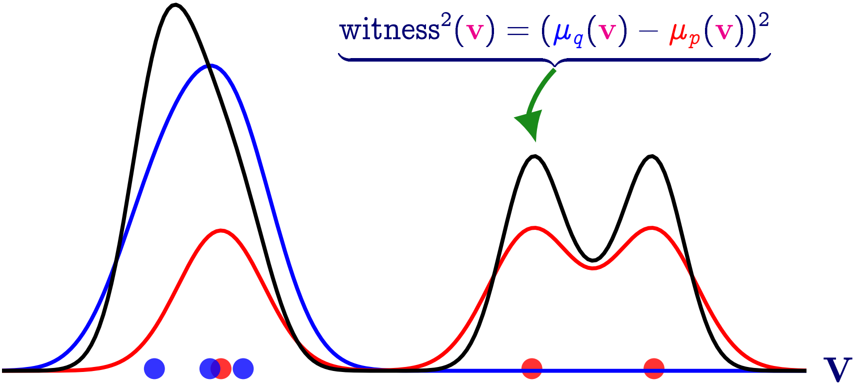
Maximum Mean Discrepancy (MMD) Witness Function (Gretton et al., 2012)



$\text{MMD}(q, p) = ||\text{witness}||$

Maximum Mean Discrepancy (MMD) Witness Function (Gretton et al., 2012)





- $\text{witness}^2(\mathbf{v})$ can be used to find a good test location $\mathbf{v}^* =$ ★ .

Model Criticism by the MMD Witness

- Find a location $\mathbf{v}$ at which q and p differ most (ME test) [Jitkrittum et al., 2016].

Model Criticism by the MMD Witness

- Find a location $\mathbf{v}$ at which q and p differ most (ME test) [Jitkrittum et al., 2016].

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

Model Criticism by the MMD Witness

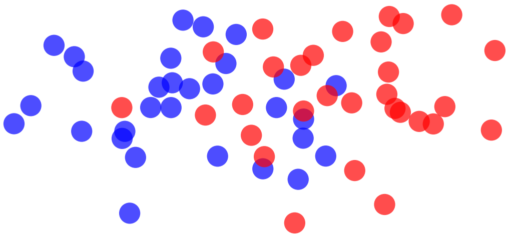
- Find a location $\mathbf{v}$ at which q and p differ most (ME test) [Jitkrittum et al., 2016].

$$\mathrm{witness}(\mathbf{v}) = \mathbb{E}_{\mathbf{x}\sim q}[\quad k_{\mathbf{v}}(\mathbf{x}) \quad] - \mathbb{E}_{\mathbf{y}\sim p}[\quad k_{\mathbf{v}}(\mathbf{y}) \quad]$$

$$\mathrm{score}(\mathbf{v}) = \frac{\mathrm{witness}^2(\mathbf{v})}{\mathrm{noise}(\mathbf{v})} = \frac{\mathrm{witness}^2(\mathbf{v})}{\sqrt{\mathbb{V}_{\mathbf{x}\sim q}[k_{\mathbf{v}}(\mathbf{x})] + \mathbb{V}_{\mathbf{y}\sim p}[k_{\mathbf{v}}(\mathbf{y})]}}.$$

Model Criticism by the MMD Witness

- Find a location $\mathbf{v}$ at which q and p differ most (ME test) [Jitkrittum et al., 2016].

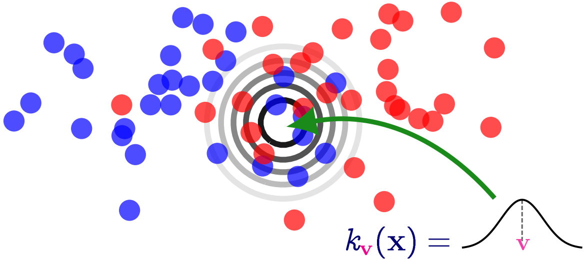


$$\text{witness}(\mathbf{v}) = \mathbb{E}_{\mathbf{x}\sim q}[\quad k_{\mathbf{v}}(\mathbf{x}) \quad] - \mathbb{E}_{\mathbf{y}\sim p}[\quad k_{\mathbf{v}}(\mathbf{y}) \quad]$$

$$\text{score}(\mathbf{v}) = \frac{\text{witness}^2(\mathbf{v})}{\text{noise}(\mathbf{v})} = \frac{\text{witness}^2(\mathbf{v})}{\sqrt{\mathbb{V}_{\mathbf{x}\sim q}[k_{\mathbf{v}}(\mathbf{x})] + \mathbb{V}_{\mathbf{y}\sim p}[k_{\mathbf{v}}(\mathbf{y})]}}.$$

- Find a location $\mathbf{v}$ at which q and p differ most (ME test) [Jitkrittum et al., 2016].

score: 0.008



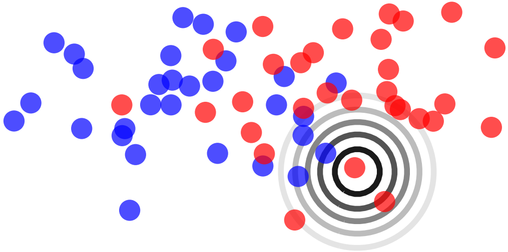
$$\mathrm{witness}(\mathbf{v}) = \mathbb{E}_{\mathbf{x}\sim q}[\quad k_{\mathbf{v}}(\mathbf{x}) \quad] - \mathbb{E}_{\mathbf{y}\sim p}[\quad k_{\mathbf{v}}(\mathbf{y}) \quad]$$

$$\mathrm{score}(\mathbf{v}) = \frac{\mathrm{witness}^2(\mathbf{v})}{\mathrm{noise}(\mathbf{v})} = \frac{\mathrm{witness}^2(\mathbf{v})}{\sqrt{\mathbb{V}_{\mathbf{x}\sim q}[k_{\mathbf{v}}(\mathbf{x})] + \mathbb{V}_{\mathbf{y}\sim p}[k_{\mathbf{v}}(\mathbf{y})]}}.$$

Model Criticism by the MMD Witness

- Find a location $\mathbf{v}$ at which q and p differ most (ME test) [Jitkrittum et al., 2016].

score: 1.6



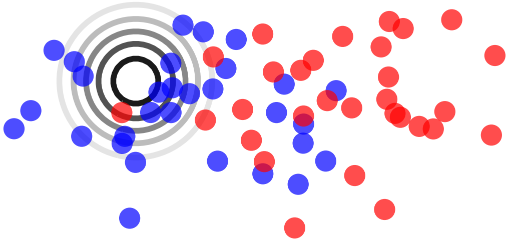
$$\text{witness}(\mathbf{v}) = \mathbb{E}_{\mathbf{x}\sim q}[\quad k_{\mathbf{v}}(\mathbf{x}) \quad] - \mathbb{E}_{\mathbf{y}\sim p}[\quad k_{\mathbf{v}}(\mathbf{y}) \quad]$$

$$\text{score}(\mathbf{v}) = \frac{\text{witness}^2(\mathbf{v})}{\text{noise}(\mathbf{v})} = \frac{\text{witness}^2(\mathbf{v})}{\sqrt{\mathbb{V}_{\mathbf{x}\sim q}[k_{\mathbf{v}}(\mathbf{x})] + \mathbb{V}_{\mathbf{y}\sim p}[k_{\mathbf{v}}(\mathbf{y})]}}.$$

Model Criticism by the MMD Witness

- Find a location $\mathbf{v}$ at which q and p differ most (ME test) [Jitkrittum et al., 2016].

score: 13



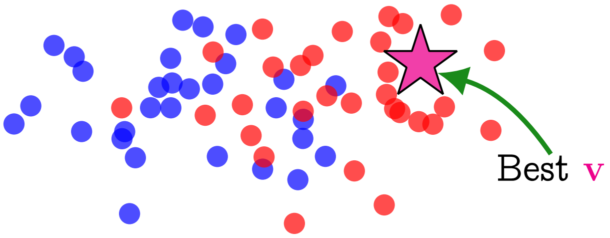
$$\text{witness}(\mathbf{v}) = \mathbb{E}_{\mathbf{x}\sim q}[\quad k_{\mathbf{v}}(\mathbf{x}) \quad] - \mathbb{E}_{\mathbf{y}\sim p}[\quad k_{\mathbf{v}}(\mathbf{y}) \quad]$$

$$\text{score}(\mathbf{v}) = \frac{\text{witness}^2(\mathbf{v})}{\text{noise}(\mathbf{v})} = \frac{\text{witness}^2(\mathbf{v})}{\sqrt{\mathbb{V}_{\mathbf{x}\sim q}[k_{\mathbf{v}}(\mathbf{x})] + \mathbb{V}_{\mathbf{y}\sim p}[k_{\mathbf{v}}(\mathbf{y})]}}.$$

Model Criticism by the MMD Witness

- Find a location $\mathbf{v}$ at which q and p differ most (ME test) [Jitkrittum et al., 2016].

score: 25



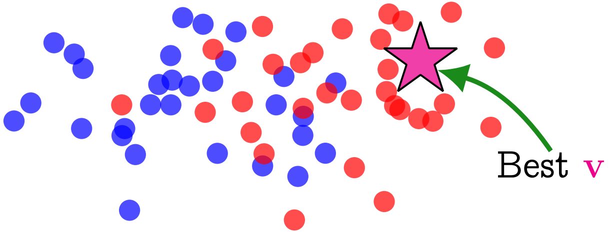
$$\text{witness}(\mathbf{v}) = \mathbb{E}_{\mathbf{x}\sim q}[\quad k_{\mathbf{v}}(\mathbf{x}) \quad] - \mathbb{E}_{\mathbf{y}\sim p}[\quad k_{\mathbf{v}}(\mathbf{y}) \quad]$$

$$\text{score}(\mathbf{v}) = \frac{\text{witness}^2(\mathbf{v})}{\text{noise}(\mathbf{v})} = \frac{\text{witness}^2(\mathbf{v})}{\sqrt{\mathbb{V}_{\mathbf{x}\sim q}[k_{\mathbf{v}}(\mathbf{x})] + \mathbb{V}_{\mathbf{y}\sim p}[k_{\mathbf{v}}(\mathbf{y})]}}.$$

Model Criticism by the MMD Witness

- Find a location $\mathbf{v}$ at which q and p differ most (ME test) [Jitkrittum et al., 2016].

score: 25



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

$$\text{score}(\mathbf{v}) = \frac{\text{witness}^2(\mathbf{v})}{\text{noise}(\mathbf{v})} =$$

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 $\mathbb{E}_{\mathbf{y}\sim p}[k_{\mathbf{v}}(\mathbf{y})]$.

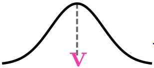
The Stein Witness Function [Liu et al., 2016, Chwialkowski et al., 2016]

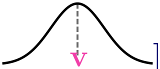
Problem: No sample from p. Cannot estimate $\mathbb{E}_{\mathbf{y}\sim p}[k_{\mathbf{v}}(\mathbf{y})]$.

$$\text{(Stein) witness}(\mathbf{v}) = \mathbb{E}_{\mathbf{x}\sim q}[\quad T_p k_{\mathbf{v}}(\mathbf{x}) \quad] - \mathbb{E}_{\mathbf{y}\sim p}[\quad T_p k_{\mathbf{v}}(\mathbf{y}) \quad]$$

The Stein Witness Function [Liu et al., 2016, Chwialkowski et al., 2016]

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

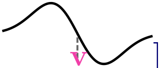
(Stein) witness($\mathbf{v}$) $= \mathbb{E}_{\mathbf{x}\sim q}[\, T_p$ 

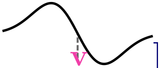
$] - \mathbb{E}_{\mathbf{y}\sim p}[\, T_p$ 

$]$

The Stein Witness Function [Liu et al., 2016, Chwialkowski et al., 2016]

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

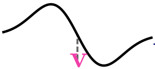
(Stein) witness($\mathbf{v}$) $= \mathbb{E}_{\mathbf{x}\sim q}[$ 

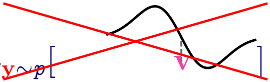
$] - \mathbb{E}_{\mathbf{y}\sim p}[$ 

$]$

The Stein Witness Function [Liu et al., 2016, Chwialkowski et al., 2016]

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

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

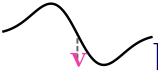
$] - \mathbb{E}_{\mathbf{y}\sim p}[$ 

$]$

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

The Stein Witness Function [Liu et al., 2016, Chwialkowski et al., 2016]

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

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

$]$

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

The Stein Witness Function [Liu et al., 2016, Chwialkowski et al., 2016]

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

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

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

The Stein Witness Function [Liu et al., 2016, Chwialkowski et al., 2016]

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

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

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

Proposal: Good $\mathbf{v}$ should have high

$$\text{score}(\mathbf{v}) = \frac{\text{witness}^2(\mathbf{v})}{\text{noise}(\mathbf{v})}.$$

The Stein Witness Function [Liu et al., 2016, Chwialkowski et al., 2016]

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

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

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

Proposal: Good $\mathbf{v}$ should have high

$$\text{score}(\mathbf{v}) = \frac{\text{witness}^2(\mathbf{v})}{\text{noise}(\mathbf{v})}.$$

signal-to-noise ratio

The Stein Witness Function [Liu et al., 2016, Chwialkowski et al., 2016]

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

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

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

Proposal: Good $\mathbf{v}$ should have high

$$\text{score}(\mathbf{v}) = \frac{\text{witness}^2(\mathbf{v})}{\text{noise}(\mathbf{v})}.$$

signal-to-noise ratio

- score($\mathbf{v}$) can be estimated in linear-time.

The Stein Witness Function [Liu et al., 2016, Chwialkowski et al., 2016]

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

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

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

Proposal: Good $\mathbf{v}$ should have high

$$\text{score}(\mathbf{v}) = \frac{\text{witness}^2(\mathbf{v})}{\text{noise}(\mathbf{v})}.$$

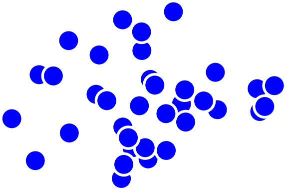
signal-to-noise ratio

- score($\mathbf{v}$) can be estimated in linear-time.

Goodness-of-fit test:

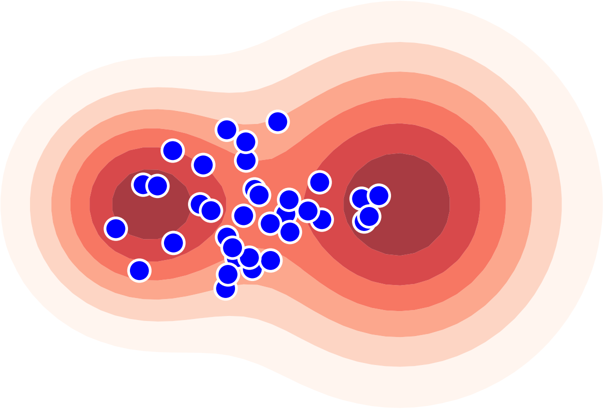
1. Find $\mathbf{v}^* = \arg\max_{\mathbf{v}} \text{score}(\mathbf{v})$.
2. Reject H_0 if $\text{witness}^2(\mathbf{v}^*) > \text{threshold}$.

Proposal: Model Criticism with the Stein Witness



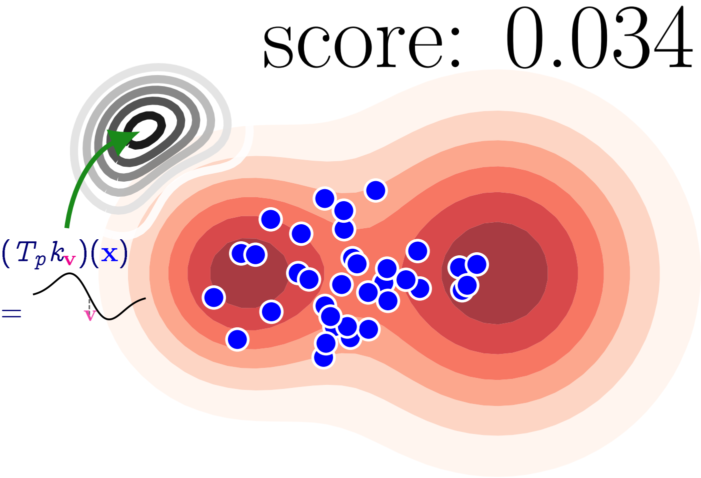
$$\text{score}(\mathbf{v}) = \frac{\text{witness}^2(\mathbf{v})}{\text{noise}(\mathbf{v})}.$$

Proposal: Model Criticism with the Stein Witness



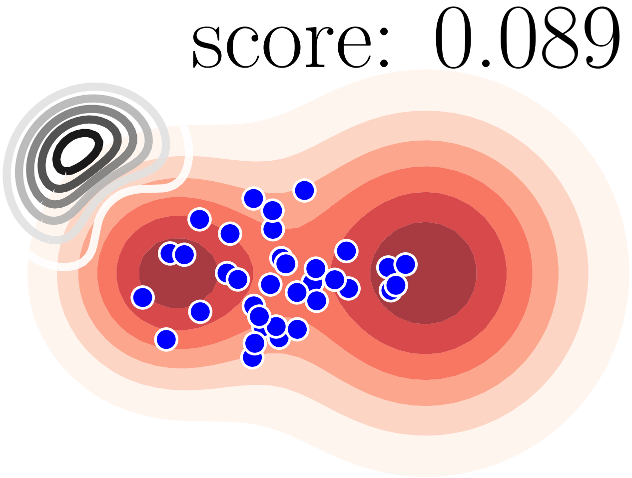
$$\text{score}(\mathbf{v}) = \frac{\text{witness}^2(\mathbf{v})}{\text{noise}(\mathbf{v})}.$$

Proposal: Model Criticism with the Stein Witness



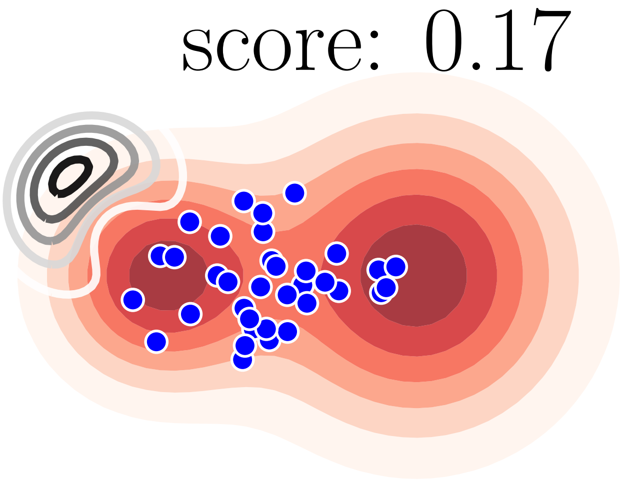
$$\text{score}(\mathbf{v}) = \frac{\text{witness}^2(\mathbf{v})}{\text{noise}(\mathbf{v})}.$$

Proposal: Model Criticism with the Stein Witness



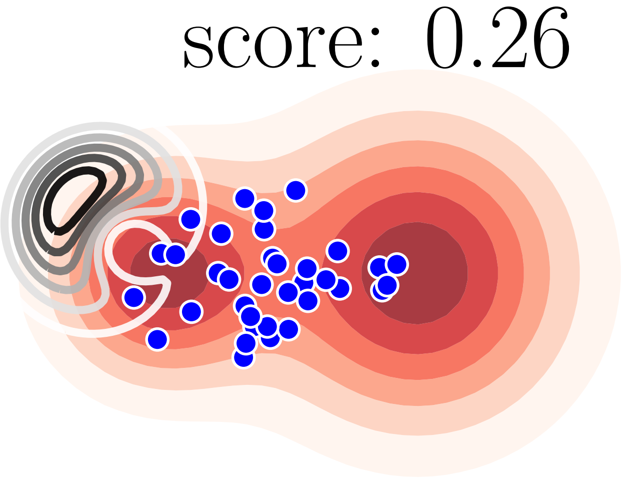
$$\text{score}(\mathbf{v}) = \frac{\text{witness}^2(\mathbf{v})}{\text{noise}(\mathbf{v})}.$$

Proposal: Model Criticism with the Stein Witness



$$\text{score}(\mathbf{v}) = \frac{\text{witness}^2(\mathbf{v})}{\text{noise}(\mathbf{v})}.$$

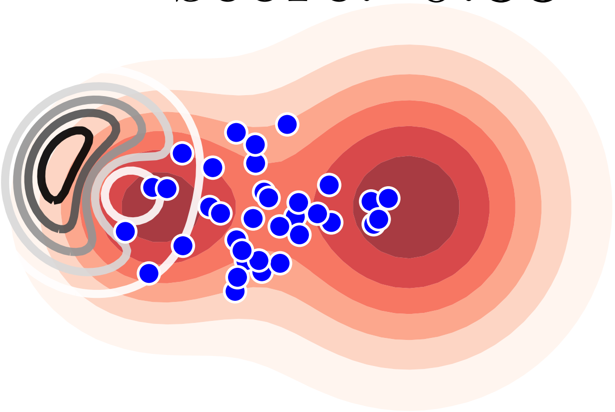
Proposal: Model Criticism with the Stein Witness



$$\text{score}(\mathbf{v}) = \frac{\text{witness}^2(\mathbf{v})}{\text{noise}(\mathbf{v})}.$$

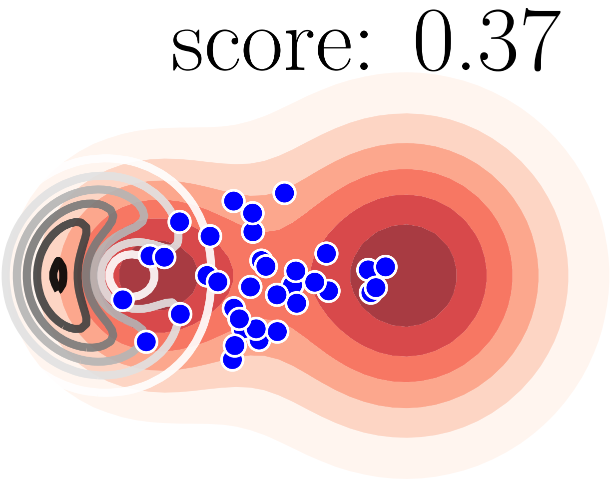
Proposal: Model Criticism with the Stein Witness

score: 0.33



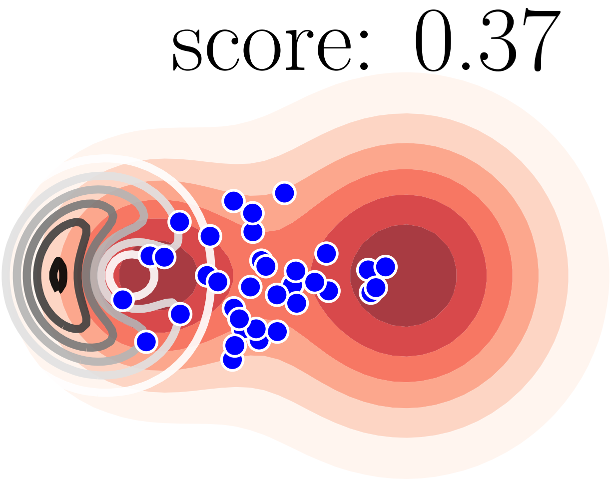
$$\text{score}(\mathbf{v}) = \frac{\text{witness}^2(\mathbf{v})}{\text{noise}(\mathbf{v})}.$$

Proposal: Model Criticism with the Stein Witness



$$\text{score}(\mathbf{v}) = \frac{\text{witness}^2(\mathbf{v})}{\text{noise}(\mathbf{v})}.$$

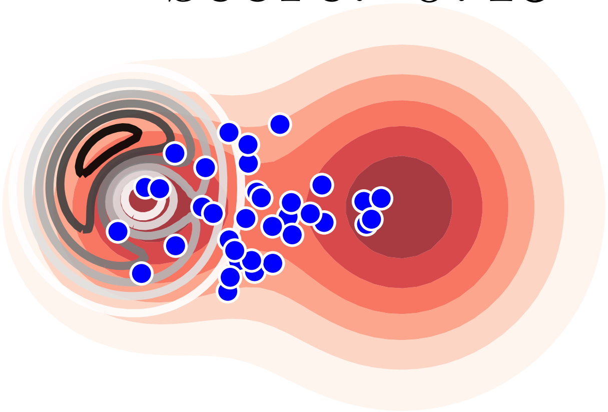
Proposal: Model Criticism with the Stein Witness



$$\text{score}(\mathbf{v}) = \frac{\text{witness}^2(\mathbf{v})}{\text{noise}(\mathbf{v})}.$$

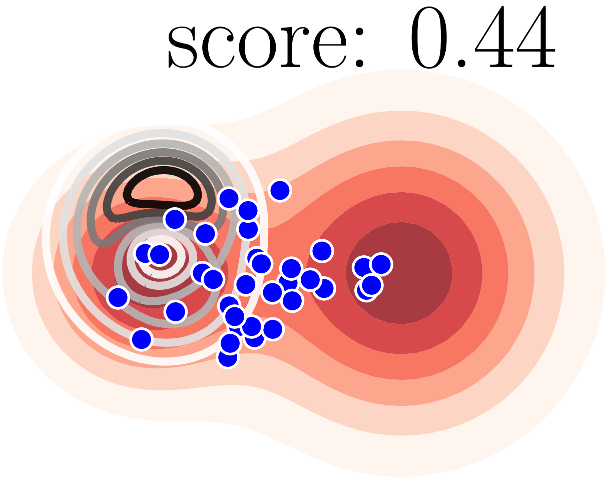
Proposal: Model Criticism with the Stein Witness

score: 0.45



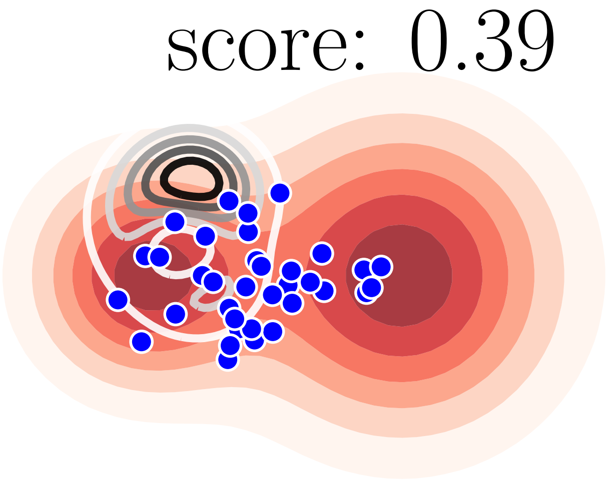
$$\text{score}(\mathbf{v}) = \frac{\text{witness}^2(\mathbf{v})}{\text{noise}(\mathbf{v})}.$$

Proposal: Model Criticism with the Stein Witness



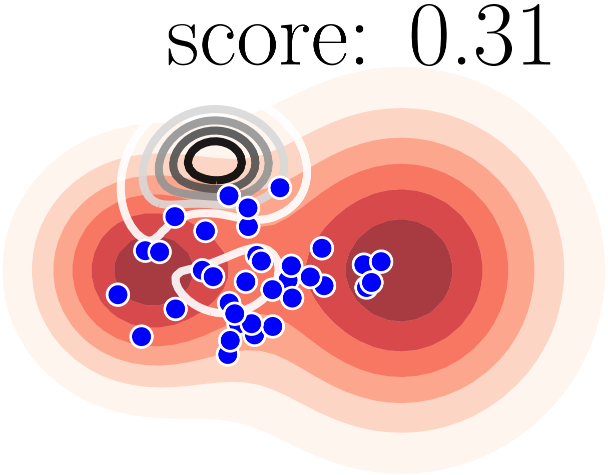
$$\text{score}(\mathbf{v}) = \frac{\text{witness}^2(\mathbf{v})}{\text{noise}(\mathbf{v})}.$$

Proposal: Model Criticism with the Stein Witness



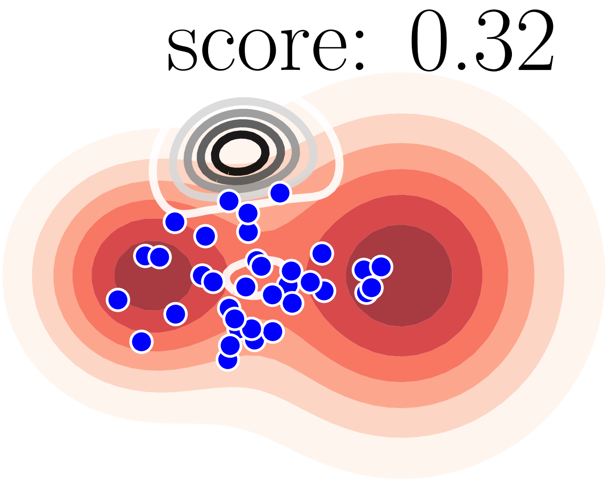
$$\text{score}(\mathbf{v}) = \frac{\text{witness}^2(\mathbf{v})}{\text{noise}(\mathbf{v})}.$$

Proposal: Model Criticism with the Stein Witness



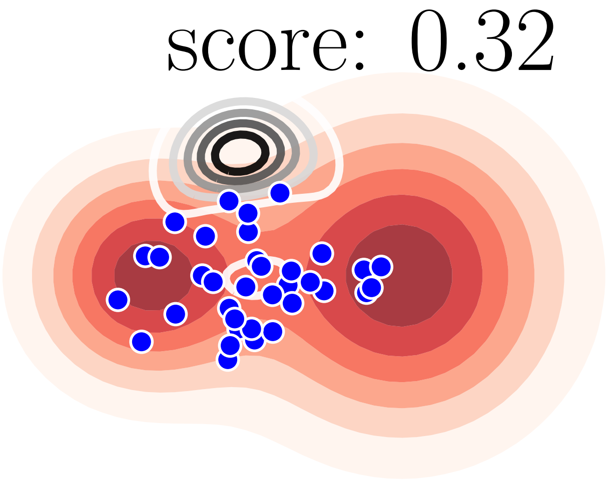
$$\mathrm{score}(\mathbf{v}) = \frac{\mathrm{witness}^2(\mathbf{v})}{\mathrm{noise}(\mathbf{v})}.$$

Proposal: Model Criticism with the Stein Witness



$$\text{score}(\mathbf{v}) = \frac{\text{witness}^2(\mathbf{v})}{\text{noise}(\mathbf{v})}.$$

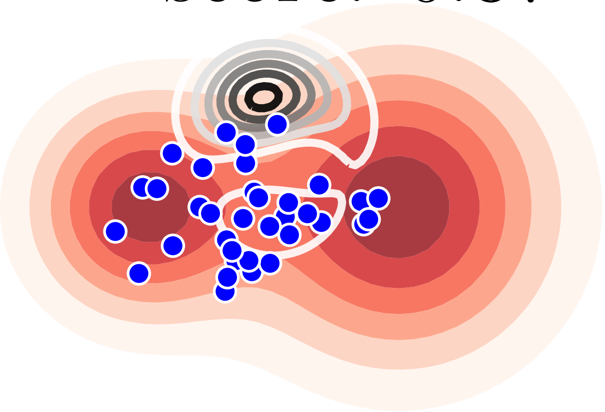
Proposal: Model Criticism with the Stein Witness



$$\text{score}(\mathbf{v}) = \frac{\text{witness}^2(\mathbf{v})}{\text{noise}(\mathbf{v})}.$$

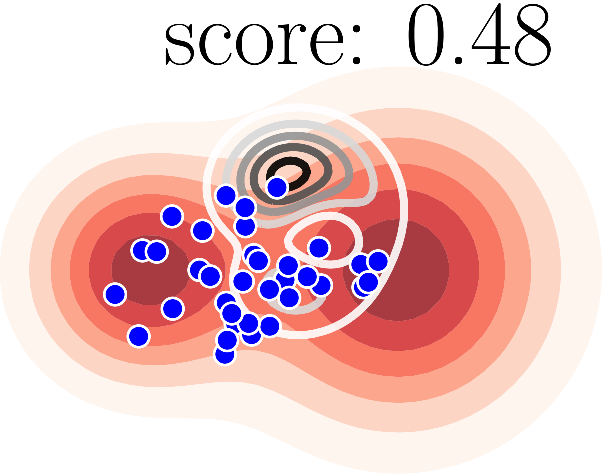
Proposal: Model Criticism with the Stein Witness

score: 0.37



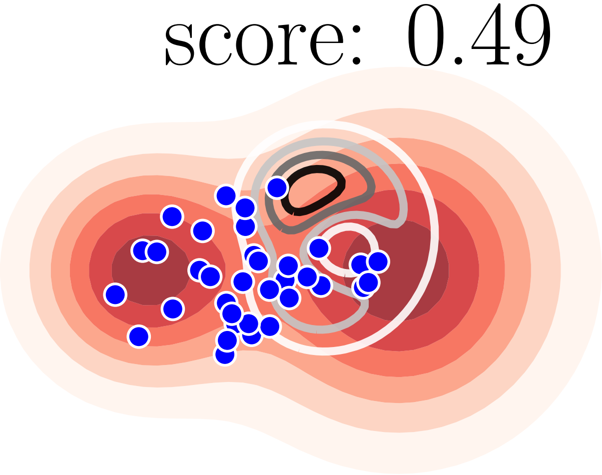
$$\text{score}(\mathbf{v}) = \frac{\text{witness}^2(\mathbf{v})}{\text{noise}(\mathbf{v})}.$$

Proposal: Model Criticism with the Stein Witness



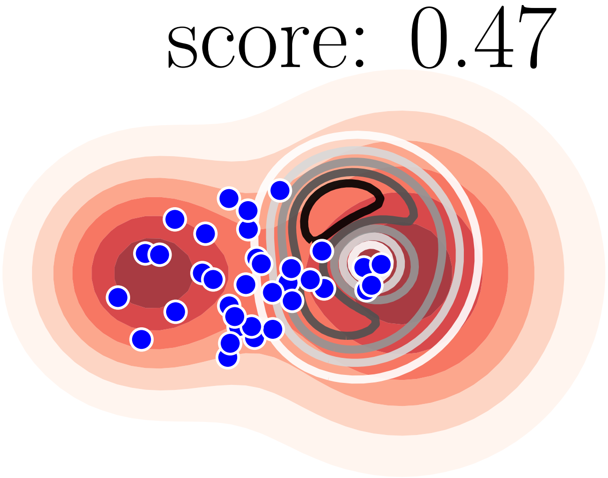
$$\text{score}(\mathbf{v}) = \frac{\text{witness}^2(\mathbf{v})}{\text{noise}(\mathbf{v})}.$$

Proposal: Model Criticism with the Stein Witness



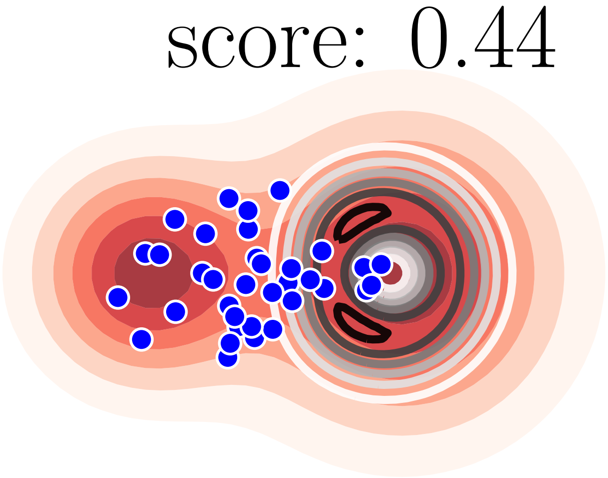
$$\text{score}(\mathbf{v}) = \frac{\text{witness}^2(\mathbf{v})}{\text{noise}(\mathbf{v})}.$$

Proposal: Model Criticism with the Stein Witness



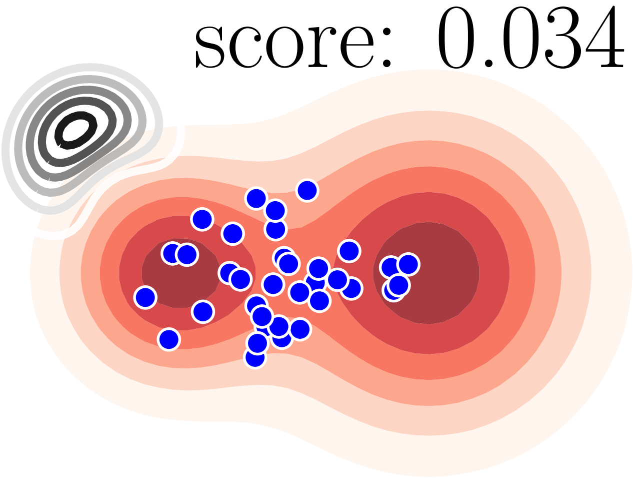
$$\text{score}(\mathbf{v}) = \frac{\text{witness}^2(\mathbf{v})}{\text{noise}(\mathbf{v})}.$$

Proposal: Model Criticism with the Stein Witness



$$\text{score}(\mathbf{v}) = \frac{\text{witness}^2(\mathbf{v})}{\text{noise}(\mathbf{v})}.$$

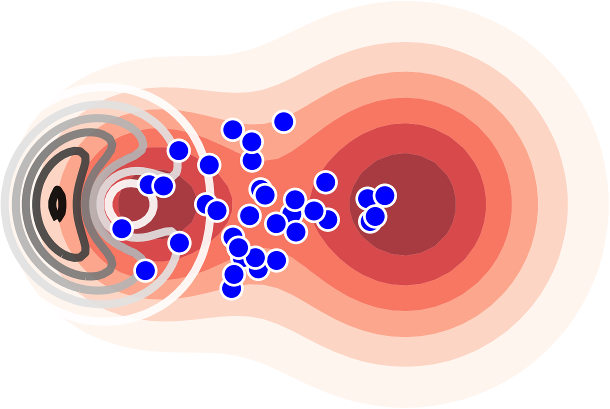
Proposal: Model Criticism with the Stein Witness



$$\text{score}(\mathbf{v}) = \frac{\text{witness}^2(\mathbf{v})}{\text{noise}(\mathbf{v})}.$$

Proposal: Model Criticism with the Stein Witness

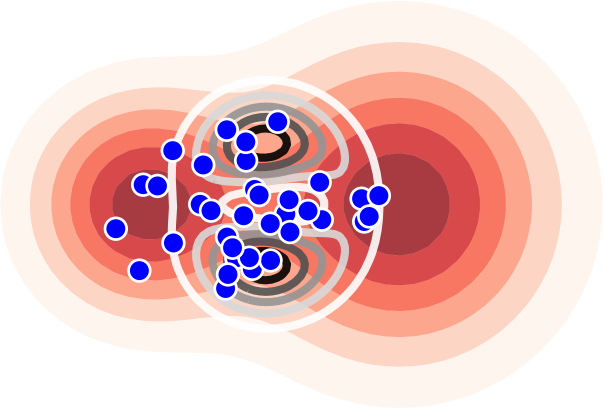
score: 0.37



$$\text{score}(\mathbf{v}) = \frac{\text{witness}^2(\mathbf{v})}{\text{noise}(\mathbf{v})}.$$

Proposal: Model Criticism with the Stein Witness

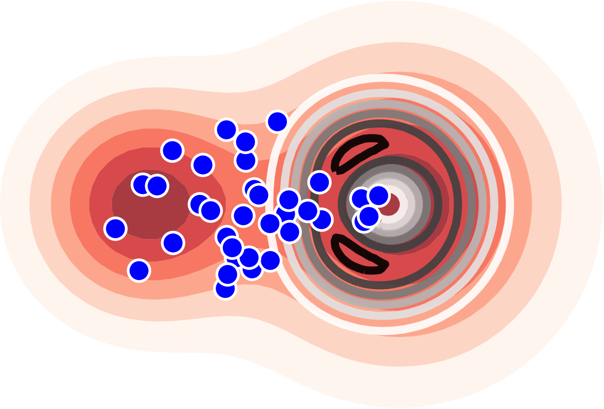
score: 0.16



$$\text{score}(\mathbf{v}) = \frac{\text{witness}^2(\mathbf{v})}{\text{noise}(\mathbf{v})}.$$

Proposal: Model Criticism with the Stein Witness

score: 0.44



$$\text{score}(\mathbf{v}) = \frac{\text{witness}^2(\mathbf{v})}{\text{noise}(\mathbf{v})}.$$

Theory

1. What is $T_p k_{\mathbf{v}}$?
2. Test statistic
3. Distributions of the test statistic, test threshold.
4. What does $\mathbf{v}^* = \arg\max_{\mathbf{v}} \mathrm{score}(\mathbf{v})$ do theoretically?

(1) What is $T_p k_{\mathbf{v}}$?

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

(1) What is $T_p k_{\mathbf{v}}$?

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

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

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

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

(1) What is $T_p k_{\mathbf{v}}$?

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

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

Normalizer cancels

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

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

(1) What is $T_p k_{\mathbf{v}}$?

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

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

Normalizer cancels

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

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

Proof:

$$\mathbb{E}_{\mathbf{y}\sim p}\left[(T_p k_{\mathbf{v}})(\mathbf{y})\right]$$

(1) What is $T_p k_{\mathbf{v}}$?

Recall witness$(\mathbf{v}) = \mathbb{E}_{\mathbf{x}\sim q}(T_p k_{\mathbf{v}})(\mathbf{x})$ − ~~$\mathbb{E}_{\mathbf{y}\sim p}(T_p k_{\mathbf{v}})(\mathbf{y})$~~

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

Normalizer cancels

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

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

Proof:

$$\mathbb{E}_{\mathbf{y}\sim p}\left[(T_p k_{\mathbf{v}})(\mathbf{y})\right] = \int_{-\infty}^{\infty}\left[(T_p k_{\mathbf{v}})(\mathbf{y})\right] p(\mathbf{y})d\mathbf{y}$$

(1) What is $T_p k_{\mathbf{v}}$?

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

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

Normalizer cancels

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

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

Proof:

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

(1) What is $T_p k_{\mathbf{v}}$?

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

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

Normalizer cancels

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

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

Proof:

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

(1) What is $T_p k_{\mathbf{v}}$?

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

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

Normalizer cancels

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

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

Proof:

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

(1) What is $T_p k_{\mathbf{v}}$?

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

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

Normalizer cancels

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

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

Proof:

$$\begin{aligned}
\mathbb{E}_{\mathbf{y}\sim p}\left[(T_p k_{\mathbf{v}})(\mathbf{y})\right] &= \int_{-\infty}^{\infty}\left[\frac{1}{\cancel{p(\mathbf{y})}}\frac{d}{d\mathbf{y}}[k_{\mathbf{v}}(\mathbf{y})p(\mathbf{y})]\right]\cancel{p(\mathbf{y})}\,d\mathbf{y} \\
&= \int_{-\infty}^{\infty}\frac{d}{d\mathbf{y}}[k_{\mathbf{v}}(\mathbf{y})p(\mathbf{y})]\,d\mathbf{y} \\
&= [k_{\mathbf{v}}(\mathbf{y})p(\mathbf{y})]_{\mathbf{y}=-\infty}^{\mathbf{y}=\infty}
\end{aligned}$$

(1) What is $T_p k_{\mathbf{v}}$?

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

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

Normalizer cancels

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

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

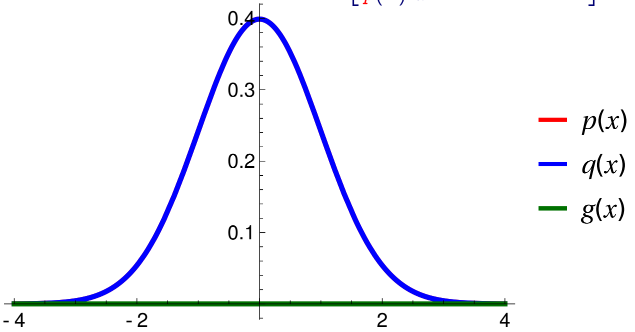
Proof:

$$\begin{aligned}
\mathbb{E}_{\mathbf{y}\sim p}\left[(T_p k_{\mathbf{v}})(\mathbf{y})\right] &= \int_{-\infty}^{\infty} \left[\frac{1}{\cancel{p(\mathbf{y})}} \frac{d}{d\mathbf{y}} [k_{\mathbf{v}}(\mathbf{y}) p(\mathbf{y})]\right] \cancel{p(\mathbf{y})}\, d\mathbf{y} \\
&= \int_{-\infty}^{\infty} \frac{d}{d\mathbf{y}} [k_{\mathbf{v}}(\mathbf{y}) p(\mathbf{y})]\, d\mathbf{y} \\
&= [k_{\mathbf{v}}(\mathbf{y}) p(\mathbf{y})]_{\mathbf{y}=-\infty}^{\mathbf{y}=\infty} \\
&= 0
\end{aligned}$$

(assume $\lim_{|\mathbf{y}|\to\infty} k_{\mathbf{v}}(\mathbf{y}) p(\mathbf{y})$)

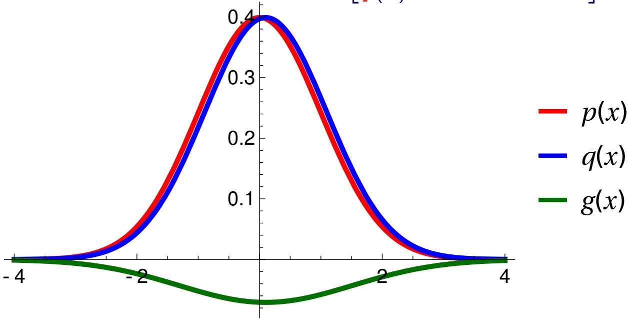
(2) Proposal: The Finite Set Stein Discrepancy (FSSD)

- Recall Stein witness: $\mathbf{g}(\mathbf{v}) := \mathbb{E}_{\mathbf{x}\sim q}\left[\frac{1}{p(\mathbf{x})}\frac{d}{d\mathbf{x}}[k_{\mathbf{v}}(\mathbf{x})p(\mathbf{x})]\right]$.



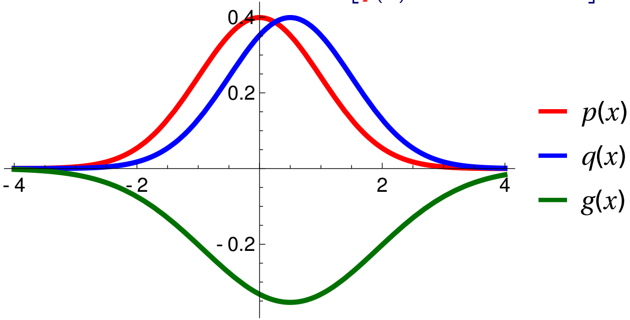
(2) Proposal: The Finite Set Stein Discrepancy (FSSD)

- Recall Stein witness: $\mathbf{g}(\mathbf{v}) := \mathbb{E}_{\mathbf{x}\sim q}\left[\frac{1}{p(\mathbf{x})}\frac{d}{d\mathbf{x}}[k_{\mathbf{v}}(\mathbf{x})p(\mathbf{x})]\right]$.



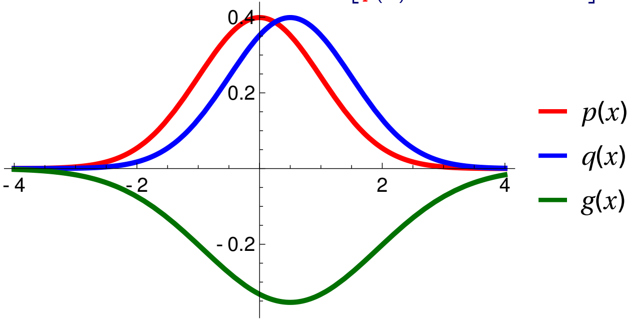
(2) Proposal: The Finite Set Stein Discrepancy (FSSD)

- Recall Stein witness: $\mathbf{g}(\mathbf{v}) := \mathbb{E}_{\mathbf{x}\sim q}\left[\frac{1}{p(\mathbf{x})}\frac{d}{d\mathbf{x}}[k_{\mathbf{v}}(\mathbf{x})p(\mathbf{x})]\right]$.



(2) Proposal: The Finite Set Stein Discrepancy (FSSD)

- Recall Stein witness: $\mathbf{g}(\mathbf{v}) := \mathbb{E}_{\mathbf{x}\sim q}\left[\frac{1}{p(\mathbf{x})}\frac{d}{d\mathbf{x}}[k_{\mathbf{v}}(\mathbf{x})p(\mathbf{x})]\right]$.



- **FSSD statistic**: Evaluate g^2 at J test locations $V = \{\mathbf{v}_1, \ldots, \mathbf{v}_J\}$.
- Population FSSD

$$\mathrm{FSSD}^2 = \frac{1}{dJ}\sum_{j=1}^{J}\|\mathbf{g}(\mathbf{v}_j)\|_2^2.$$

- Unbiased estimator $\widehat{\mathrm{FSSD}^2}$ computable in $\mathcal{O}(d^2 Jn)$ time. (d = input dimension)

(2) FSSD is a Discrepancy Measure

- $\mathrm{FSSD}^2 = \frac{1}{dJ}\sum_{j=1}^{J} \|\mathbf{g}(\mathbf{v}_j)\|_2^2$.

Theorem 1 (FSSD is a discrepancy measure).

Main conditions:

1. *(Nice kernel) Kernel k is C_0-universal, and real analytic e.g., Gaussian kernel.*
2. *(Vanishing boundary)* $\lim_{\|\mathbf{x}\|\to\infty} p(\mathbf{x}) k_{\mathbf{v}}(\mathbf{x}) = \mathbf{0}$.
3. *(Avoid "blind spots") Locations $\mathbf{v}_1, \ldots, \mathbf{v}_J \sim \eta$ which has a density.*

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

$$\mathrm{FSSD}^2 = 0 \iff p = q.$$

Summary: Evaluating the witness at random locations is sufficient to detect the discrepancy between p, q.

(2) What Are "Blind Spots"?

$$\text{Recall } \mathbf{g}(\mathbf{v}) := \mathbb{E}_{\mathbf{x}\sim q}\left[\frac{1}{p(\mathbf{x})}\frac{d}{d\mathbf{x}}[k_{\mathbf{v}}(\mathbf{x})p(\mathbf{x})]\right].$$

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

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

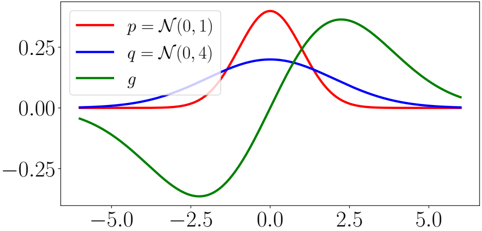
- If $v = 0$, then $\text{FSSD}^2 = g^2(v) = 0$ regardless of σ_q^2.
- If $g \neq 0$, and k is real analytic, $R = \{v \mid g(v) = 0\}$ (blind spots) has 0 Lebesgue measure.
- So, if $v \sim$ a distribution with a density, then $v \notin R$.

(2) What Are "Blind Spots"?

$$\text{Recall } \mathbf{g}(\mathbf{v}) := \mathbb{E}_{\mathbf{x}\sim q}\left[\frac{1}{p(\mathbf{x})}\frac{d}{d\mathbf{x}}[k_{\mathbf{v}}(\mathbf{x})p(\mathbf{x})]\right].$$

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

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



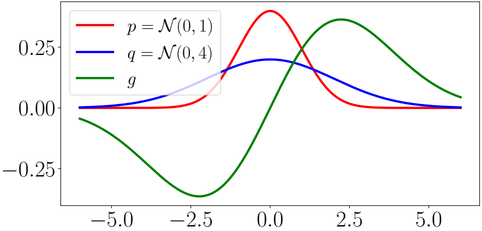
- If $v = 0$, then $\text{FSSD}^2 = g^2(v) = 0$ regardless of σ_q^2.
- If $g \neq 0$, and k is real analytic, $R = \{v \mid g(v) = 0\}$ (blind spots) has 0 Lebesgue measure.
- So, if $v \sim$ a distribution with a density, then $v \notin R$.

(2) What Are "Blind Spots"?

$$\text{Recall } \mathbf{g}(\mathbf{v}) := \mathbb{E}_{\mathbf{x}\sim q}\left[\frac{1}{p(\mathbf{x})}\frac{d}{d\mathbf{x}}[k_{\mathbf{v}}(\mathbf{x})p(\mathbf{x})]\right].$$

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

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



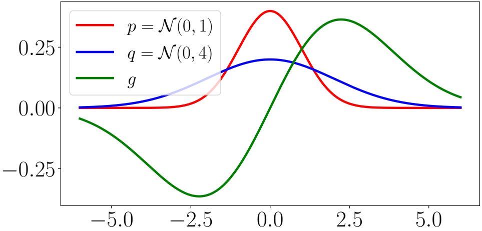
- If $v = 0$, then $\mathrm{FSSD}^2 = g^2(v) = 0$ regardless of σ_q^2.
- If $g \neq 0$, and k is real analytic, $R = \{v \mid g(v) = 0\}$ (blind spots) has 0 Lebesgue measure.
- So, if $v \sim$ a distribution with a density, then $v \notin R$.

(2) What Are "Blind Spots"?

$$\text{Recall } \mathbf{g}(\mathbf{v}) := \mathbb{E}_{\mathbf{x}\sim q}\left[\frac{1}{p(\mathbf{x})}\frac{d}{d\mathbf{x}}[k_{\mathbf{v}}(\mathbf{x})p(\mathbf{x})]\right].$$

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

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



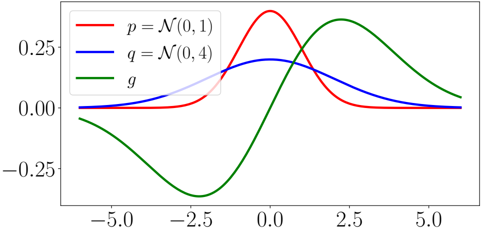
- If $v = 0$, then $\text{FSSD}^2 = g^2(v) = 0$ regardless of σ_q^2.
- If $g \neq 0$, and k is real analytic, $R = \{v \mid g(v) = 0\}$ (blind spots) has 0 Lebesgue measure.
- So, if $v \sim$ a distribution with a density, then $v \notin R$.

(2) What Are "Blind Spots"?

$$\text{Recall } \mathbf{g}(\mathbf{v}) := \mathbb{E}_{\mathbf{x}\sim q}\left[\frac{1}{p(\mathbf{x})}\frac{d}{d\mathbf{x}}[k_{\mathbf{v}}(\mathbf{x})p(\mathbf{x})]\right].$$

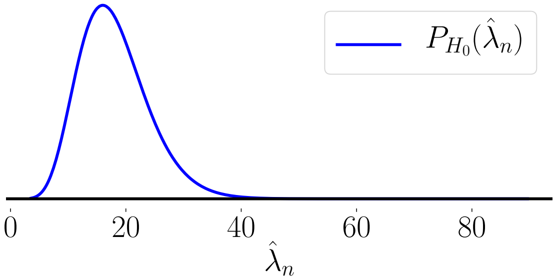
Consider $p = \mathcal{N}(0,1)$ and $q = \mathcal{N}(0, \sigma_q^2)$. Use unit-width Gaussian kernel.

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

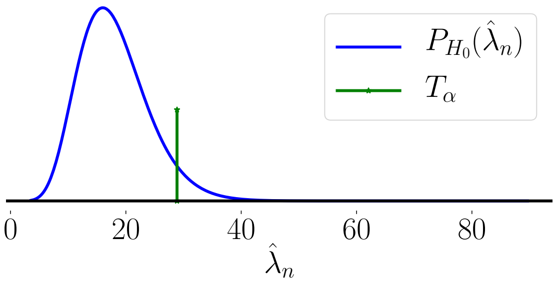


- If $v = 0$, then $\mathrm{FSSD}^2 = g^2(v) = 0$ regardless of σ_q^2.
- If $g \neq 0$, and k is real analytic, $R = \{v \mid g(v) = 0\}$ (blind spots) has 0 Lebesgue measure.
- So, if $v \sim$ a distribution with a density, then $v \notin R$.

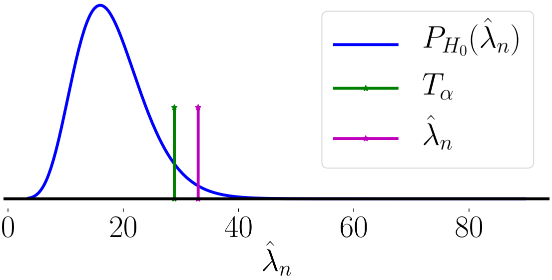
(3) Asymptotic Distributions of $\hat{\lambda}_n := n\widehat{\mathrm{FSSD}^2}$



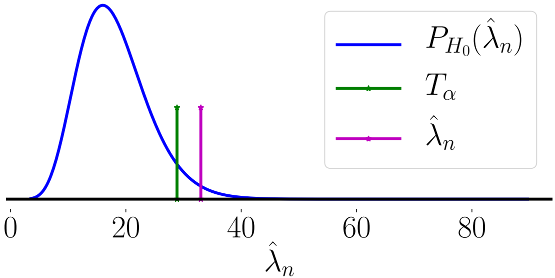
(3) Asymptotic Distributions of $\hat{\lambda}_n := n\widehat{\mathrm{FSSD}^2}$



(3) Asymptotic Distributions of $\hat{\lambda}_n := n\widehat{\text{FSSD}^2}$



(3) Asymptotic Distributions of $\hat{\lambda}_n := n\widehat{\mathrm{FSSD}^2}$

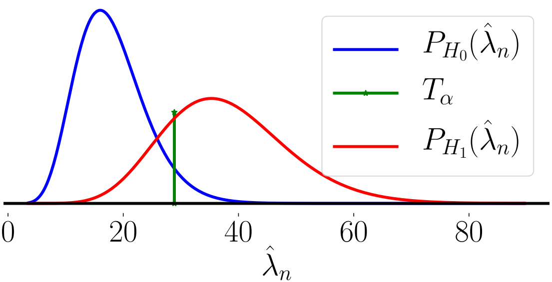


- Under $H_0 : p = q$, asymptotically

$$\hat{\lambda}_n := n\widehat{\mathrm{FSSD}^2} \xrightarrow{d} \sum_{i=1}^{dJ} (Z_i^2 - 1)\omega_i,$$

  - $\{\omega_i\}_{i=1}^{dJ}$ are non-negative, computable quantities. $Z_1, \ldots, Z_{dJ} \overset{i.i.d.}{\sim} \mathcal{N}(0, 1)$

(3) Asymptotic Distributions of $\hat{\lambda}_n := n\widehat{\mathrm{FSSD}^2}$

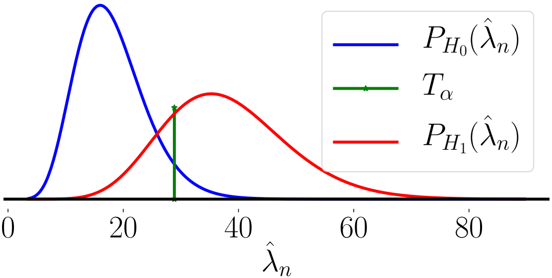


- Under $H_0 : p = q$, asymptotically

$$\hat{\lambda}_n := n\widehat{\mathrm{FSSD}^2} \xrightarrow{d} \sum_{i=1}^{dJ} (Z_i^2 - 1)\omega_i,$$

  - $\{\omega_i\}_{i=1}^{dJ}$ are non-negative, computable quantities.
    $Z_1, \ldots, Z_{dJ} \overset{i.i.d.}{\sim} \mathcal{N}(0, 1)$
- Under $H_1 : p \neq q$, asymptotically $\sqrt{n}(\widehat{\mathrm{FSSD}^2} - \mathrm{FSSD}^2) \xrightarrow{d} \mathcal{N}(0, \sigma_{H_1}^2)$.

(3) Asymptotic Distributions of $\hat{\lambda}_n := n\widehat{\text{FSSD}^2}$



- Under $H_0 : p = q$, asymptotically

$$\hat{\lambda}_n := n\widehat{\text{FSSD}^2} \xrightarrow{d} \sum_{i=1}^{dJ} (Z_i^2 - 1)\omega_i,$$

  - $\{\omega_i\}_{i=1}^{dJ}$ are non-negative, computable quantities.
    $Z_1, \ldots, Z_{dJ} \overset{i.i.d.}{\sim} \mathcal{N}(0, 1)$
- Under $H_1 : p \neq q$, asymptotically $\sqrt{n}(\widehat{\text{FSSD}^2} - \text{FSSD}^2) \xrightarrow{d} \mathcal{N}(0, \sigma^2_{H_1})$.

witness$^2(V)$ → noise(V)

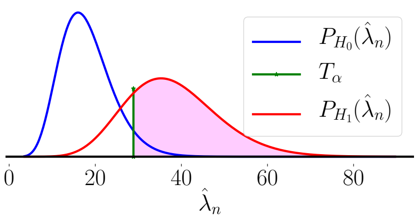
(4) What Does $\arg\max_{\mathbf{v}} \mathrm{score}(\mathbf{v})$ Do?

Proposition 1 (Asymptotic test power).

For large n, the test power $\mathbb{P}(\text{reject } H_0 \mid H_1 \text{ true}) =$

$$\mathbb{P}_{H_1}(n\widehat{\mathrm{FSSD}^2} > T_\alpha)$$
$$\approx \Phi\left(\sqrt{n}\frac{\mathrm{FSSD}^2}{\sigma_{H_1}} - \frac{T_\alpha}{\sqrt{n}\sigma_{H_1}}\right),$$

where $\Phi =$ CDF of $\mathcal{N}(0,1)$.



- For large n, the 2^{nd} term dominates.

$$\arg\max_{V,\sigma_k^2} \mathbb{P}_{H_1}(n\widehat{\mathrm{FSSD}^2} > T_\alpha) \approx \arg\max_{V,\sigma_k^2}\left[\frac{\widehat{\mathrm{FSSD}^2}}{\widehat{\sigma_{H_1}}} = \mathrm{score}(V,\sigma_k^2)\right].$$

Maximize $\mathrm{score}(V,\sigma_k^2) \iff$ Maximize test power

- In practice, split $\{\mathbf{x}_i\}_{i=1}^n$ into independent training/test sets. Optimize on tr. Goodness-of-fit test on te.

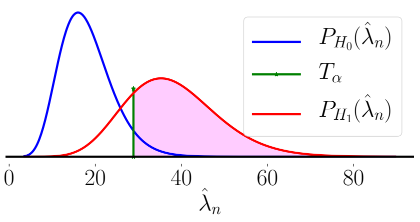
(4) What Does $\arg\max_{\mathbf{v}} \text{score}(\mathbf{v})$ Do?

Proposition 1 (Asymptotic test power).

For large n, the test power $\mathbb{P}(\text{reject } H_0 \mid H_1 \text{ true}) =$

$$\mathbb{P}_{H_1}(n\widehat{\text{FSSD}^2} > T_\alpha) \approx \Phi\left(\sqrt{n}\frac{\text{FSSD}^2}{\sigma_{H_1}} - \frac{T_\alpha}{\sqrt{n}\sigma_{H_1}}\right),$$

where $\Phi =$ CDF of $\mathcal{N}(0,1)$.



- For large n, the 2^{nd} term dominates.

$$\arg\max_{V,\sigma_k^2} \mathbb{P}_{H_1}(n\widehat{\text{FSSD}^2} > T_\alpha) \approx \arg\max_{V,\sigma_k^2} \left[\frac{\widehat{\text{FSSD}^2}}{\widehat{\sigma_{H_1}}} = \text{score}(V,\sigma_k^2)\right].$$

Maximize $\text{score}(V,\sigma_k^2) \iff$ Maximize test power

- In practice, split $\{\mathbf{x}_i\}_{i=1}^n$ into independent training/test sets. Optimize on tr. Goodness-of-fit test on te.

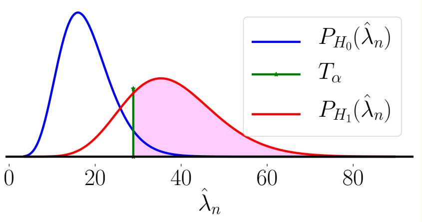
(4) What Does $\arg\max_{\mathbf{v}} \text{score}(\mathbf{v})$ Do?

Proposition 1 (Asymptotic test power).

For large n, the test power $\mathbb{P}(\text{reject } H_0 \mid H_1 \text{ true}) =$

$$\mathbb{P}_{H_1}(n\widehat{\text{FSSD}^2} > T_\alpha) \approx \Phi\left(\sqrt{n}\frac{\text{FSSD}^2}{\sigma_{H_1}} - \frac{T_\alpha}{\sqrt{n}\sigma_{H_1}}\right),$$

where $\Phi =$ CDF of $\mathcal{N}(0,1)$.



- For large n, the 2^{nd} term dominates.

$$\arg\max_{V,\sigma_k^2} \mathbb{P}_{H_1}(n\widehat{\text{FSSD}^2} > T_\alpha) \approx \arg\max_{V,\sigma_k^2} \left[\frac{\widehat{\text{FSSD}^2}}{\widehat{\sigma_{H_1}}} = \text{score}(V, \sigma_k^2)\right].$$

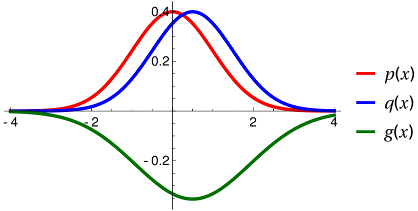
$$\boxed{\text{Maximize score}(V, \sigma_k^2) \iff \text{Maximize test power}}$$

- In practice, split $\{\mathbf{x}_i\}_{i=1}^n$ into independent training/test sets. Optimize on tr. Goodness-of-fit test on te.

Related Works

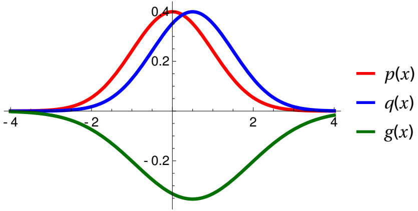
Kernel Stein Discrepancy (KSD) [Liu et al., 2016, Chwialkowski et al., 2016]

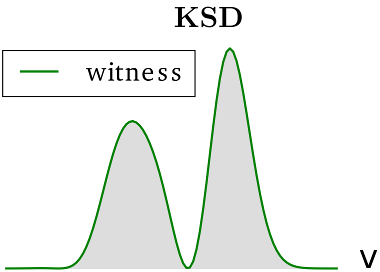
- Recall Stein witness:
  $$\mathbf{g}(\mathbf{v}) := \mathbb{E}_{\mathbf{x}\sim q}\left[\frac{1}{p(\mathbf{x})}\frac{d}{d\mathbf{x}}\left[k_{\mathbf{v}}(\mathbf{x})p(\mathbf{x})\right]\right].$$



Kernel Stein Discrepancy (KSD) [Liu et al., 2016, Chwialkowski et al., 2016]

- Recall Stein witness:
  $\mathbf{g}(\mathbf{v}) := \mathbb{E}_{\mathbf{x}\sim q}\left[\frac{1}{p(\mathbf{x})}\frac{d}{d\mathbf{x}}[k_{\mathbf{v}}(\mathbf{x})p(\mathbf{x})]\right].$



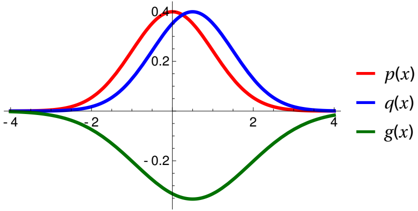


$\text{KSD}^2 = ||\mathbf{g}||^2_{\text{RKHS}}$ (RKHS norm).

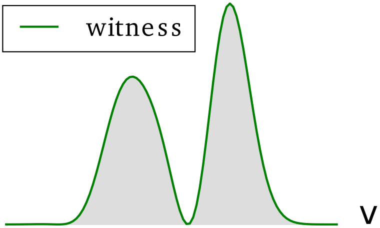
Good when the difference between p, q is spatially diffuse.

Kernel Stein Discrepancy (KSD) [Liu et al., 2016, Chwialkowski et al., 2016]

- Recall Stein witness:
  $\mathbf{g}(\mathbf{v}) := \mathbb{E}_{\mathbf{x}\sim q}\left[\frac{1}{p(\mathbf{x})}\frac{d}{d\mathbf{x}}[k_{\mathbf{v}}(\mathbf{x})p(\mathbf{x})]\right].$



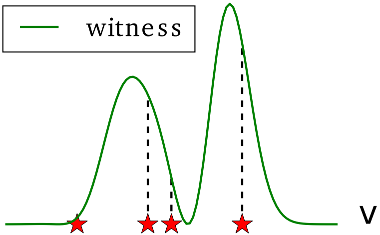
KSD



$\mathrm{KSD}^2 = \|\mathbf{g}\|^2_{\mathrm{RKHS}}$ (RKHS norm).

Good when the difference between p, q is spatially diffuse.

Proposed FSSD



$\mathrm{FSSD}^2 = \frac{1}{dJ}\sum_{j=1}^{J}\|\mathbf{g}(\mathbf{v}_j)\|_2^2.$

Good when the difference between p, q is local.

Kernel Stein Discrepancy (KSD)

Closed-form expression for KSD: [Liu et al., 2016, Chwialkowski et al., 2016]

$$\mathrm{KSD}^2 = \|\mathbf{g}\|^2_{\mathrm{RKHS}} = \overbrace{\mathbb{E}_{\mathbf{x}\sim q}\mathbb{E}_{\mathbf{y}\sim q}}^{\text{double sums}} h_p(\mathbf{x}, \mathbf{y})$$

where

$$\begin{aligned} h_p(\mathrm{x}, \mathrm{y}) := & \left[\partial_{\mathrm{x}} \log p(\mathrm{x})\right] k(\mathrm{x}, \mathrm{y}) \left[\partial_{\mathrm{y}} \log p(\mathrm{y})\right] \\ & + \left[\partial_{\mathrm{y}} \log p(\mathrm{y})\right] \partial_{\mathrm{x}} k(\mathrm{x}, \mathrm{y}) \\ & + \left[\partial_{\mathrm{x}} \log p(\mathrm{x})\right] \partial_{\mathrm{y}} k(\mathrm{x}, \mathrm{y}) \\ & + \partial_{\mathrm{x}} \partial_{\mathrm{y}} k(\mathrm{x}, \mathrm{y}) \end{aligned}$$

and k is a kernel.

- ✗ The “double sums” make it $\mathcal{O}(d^2 n^2)$. Slow.

Kernel Stein Discrepancy (KSD)

Closed-form expression for KSD: [Liu et al., 2016, Chwialkowski et al., 2016]

$$\mathrm{KSD}^2 = \|\mathbf{g}\|^2_{\mathrm{RKHS}} = \overbrace{\mathbb{E}_{\mathbf{x}\sim q}\mathbb{E}_{\mathbf{y}\sim q}}^{\text{double sums}} h_p(\mathbf{x}, \mathbf{y})$$

where

$$\begin{aligned} h_p(\mathbf{x}, \mathbf{y}) := & \left[\partial_{\mathbf{x}} \log p(\mathbf{x})\right] k(\mathbf{x}, \mathbf{y}) \left[\partial_{\mathbf{y}} \log p(\mathbf{y})\right] \\ & + \left[\partial_{\mathbf{y}} \log p(\mathbf{y})\right] \partial_{\mathbf{x}} k(\mathbf{x}, \mathbf{y}) \\ & + \left[\partial_{\mathbf{x}} \log p(\mathbf{x})\right] \partial_{\mathbf{y}} k(\mathbf{x}, \mathbf{y}) \\ & + \partial_{\mathbf{x}} \partial_{\mathbf{y}} k(\mathbf{x}, \mathbf{y}) \end{aligned}$$

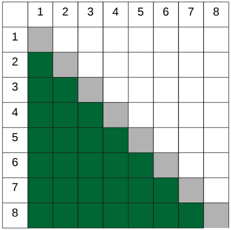
and k is a kernel.

- ✗ The "double sums" make it $\mathcal{O}(d^2 n^2)$. Slow.

Linear-Time Kernel Stein Discrepancy (LKS)

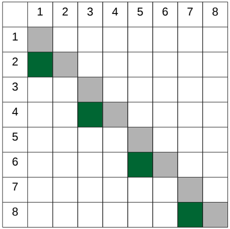
- [Liu et al., 2016] also proposed a linear version of KSD.
- For $\{\mathbf{x}_i\}_{i=1}^n \sim q$, KSD test statistic is

$$\frac{2}{n(n-1)} \sum_{i<j} h_p(\mathbf{x}_i, \mathbf{x}_j).$$



- LKS test statistic is a "running average"

$$\frac{2}{n} \sum_{i=1}^{n/2} h_p(\mathbf{x}_{2i-1}, \mathbf{x}_{2i}).$$



- Both unbiased. LKS has $\mathcal{O}(d^2 n)$ runtime. Same as proposed FSSD.
- ✗ LKS has high variance. Poor test power.

Simulation Settings

- Gaussian kernel $k(\mathbf{x}, \mathbf{v}) = \exp\left(-\frac{\|\mathbf{x}-\mathbf{v}\|_2^2}{2\sigma_k^2}\right)$

	Method	Description
1	**FSSD-opt**	Proposed. With optimization. $J = 5$.
2	FSSD-rand	Proposed. Random test locations.
3	**KSD**	Quadratic-time kernel Stein discrepancy [Liu et al., 2016, Chwialkowski et al., 2016]
4	LKS	Linear-time running average version of KSD.
5	MMD-opt	MMD two-sample test [Gretton et al., 2012]. With optimization.
6	ME-test	Mean Embeddings two-sample test [Jitkrittum et al., 2016]. With optimization.

- Two-sample tests need to draw sample from p.
- Tests with optimization use 20% of the data.
- Significance level $\alpha = 0.05$.

Simulation Settings

- Gaussian kernel $k(\mathbf{x}, \mathbf{v}) = \exp\left(-\frac{\|\mathbf{x}-\mathbf{v}\|_2^2}{2\sigma_k^2}\right)$

	Method	Description
1	**FSSD-opt**	Proposed. With optimization. $J = 5$.
2	FSSD-rand	Proposed. Random test locations.
3	**KSD**	Quadratic-time kernel Stein discrepancy [Liu et al., 2016, Chwialkowski et al., 2016]
4	LKS	Linear-time running average version of KSD.
5	MMD-opt	MMD two-sample test [Gretton et al., 2012]. With optimization.
6	ME-test	Mean Embeddings two-sample test [Jitkrittum et al., 2016]. With optimization.

- Two-sample tests need to draw sample from p.
- Tests with optimization use 20% of the data.
- Significance level $\alpha = 0.05$.

Simulation Settings

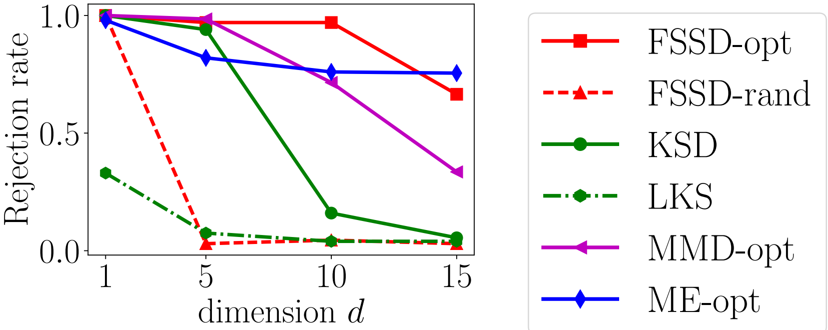
- Gaussian kernel $k(\mathbf{x}, \mathbf{v}) = \exp\left(-\frac{\|\mathbf{x}-\mathbf{v}\|_2^2}{2\sigma_k^2}\right)$

	Method	Description
1	**FSSD-opt**	Proposed. With optimization. $J = 5$.
2	FSSD-rand	Proposed. Random test locations.
3	**KSD**	Quadratic-time kernel Stein discrepancy [Liu et al., 2016, Chwialkowski et al., 2016]
4	LKS	Linear-time running average version of KSD.
5	MMD-opt	MMD two-sample test [Gretton et al., 2012]. With optimization.
6	ME-test	**M**ean **E**mbeddings two-sample test [Jitkrittum et al., 2016]. With optimization.

- Two-sample tests need to draw sample from p.
- Tests with optimization use 20% of the data.
- Significance level $\alpha = 0.05$.

Gaussian Vs. Laplace

- p = Gaussian. q = Laplace. Same mean and variance. High-order moments differ.
- Sample size $n = 1000$.



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

Gaussian-Bernoulli Restricted Boltzmann Machine (RBM)

- $p(\mathbf{x})$ is the marginal of

$$p(\mathbf{x}, \mathbf{h}) = \frac{1}{Z} \exp\left(\mathbf{x}^\top \mathbf{B}\mathbf{h} + \mathbf{b}^\top \mathbf{x} + \mathbf{c}^\top \mathbf{x} - \frac{1}{2}\|\mathbf{x}\|^2\right),$$

  where $\mathbf{x} \in \mathbb{R}^{50}$, $\mathbf{h} \in \{\pm 1\}^{40}$ is latent. Randomly pick $\mathbf{B}, \mathbf{b}, \mathbf{c}$.
- $q(\mathbf{x}) = p(\mathbf{x})$ with i.i.d. $\mathcal{N}(0, \sigma_{per})$ noise added to all entries of $\mathbf{B}$.
- Sample size $n = 1000$.

Gaussian-Bernoulli Restricted Boltzmann Machine (RBM)

- $p(\mathbf{x})$ is the marginal of
$$p(\mathbf{x}, \mathbf{h}) = \frac{1}{Z} \exp\left(\mathbf{x}^\top \mathbf{B}\mathbf{h} + \mathbf{b}^\top \mathbf{x} + \mathbf{c}^\top \mathbf{x} - \frac{1}{2}\|\mathbf{x}\|^2\right),$$
where $\mathbf{x} \in \mathbb{R}^{50}$, $\mathbf{h} \in \{\pm 1\}^{40}$ is latent. Randomly pick $\mathbf{B}, \mathbf{b}, \mathbf{c}$.
- $q(\mathbf{x}) = p(\mathbf{x})$ with i.i.d. $\mathcal{N}(0, \sigma_{per})$ noise added to all entries of $\mathbf{B}$.
- Sample size $n = 1000$.

Gaussian-Bernoulli Restricted Boltzmann Machine (RBM)

- $p(\mathbf{x})$ is the marginal of

$$p(\mathbf{x}, \mathbf{h}) = \frac{1}{Z} \exp\left(\mathbf{x}^\top \mathbf{B} \mathbf{h} + \mathbf{b}^\top \mathbf{x} + \mathbf{c}^\top \mathbf{x} - \frac{1}{2} \|\mathbf{x}\|^2 \right),$$

  where $\mathbf{x} \in \mathbb{R}^{50}$, $\mathbf{h} \in \{\pm 1\}^{40}$ is latent. Randomly pick $\mathbf{B}, \mathbf{b}, \mathbf{c}$.
- $q(\mathbf{x}) = p(\mathbf{x})$ with i.i.d. $\mathcal{N}(0, \sigma_{per})$ noise added to all entries of $\mathbf{B}$.
- Sample size $n = 1000$.

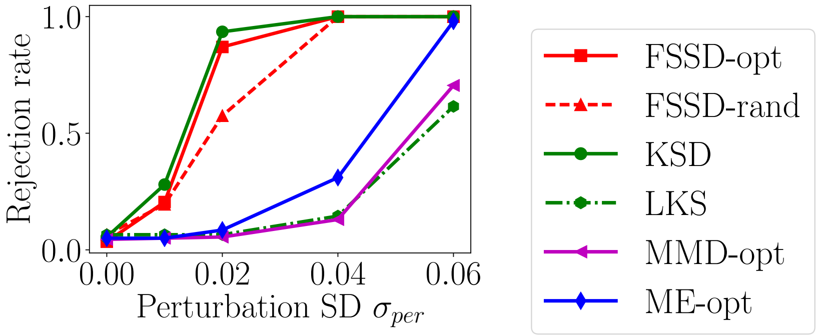
Gaussian-Bernoulli Restricted Boltzmann Machine (RBM)

- $p(\mathbf{x})$ is the marginal of

$$p(\mathbf{x}, \mathbf{h}) = \frac{1}{Z} \exp\left(\mathbf{x}^\top \mathbf{B}\mathbf{h} + \mathbf{b}^\top \mathbf{x} + \mathbf{c}^\top \mathbf{x} - \frac{1}{2}\|\mathbf{x}\|^2 \right),$$

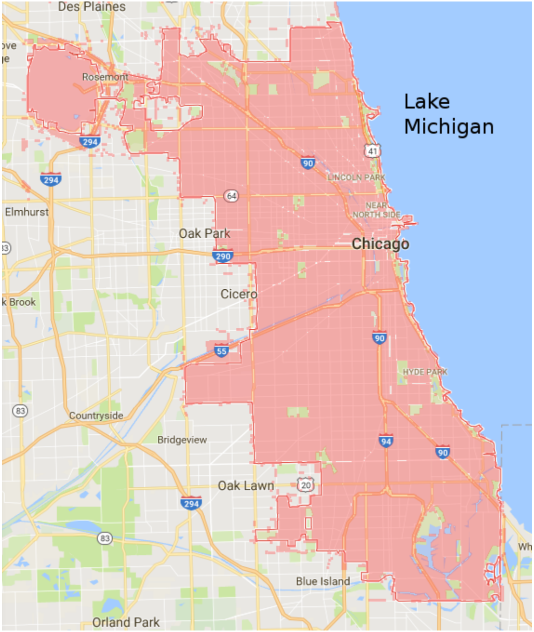
where $\mathbf{x} \in \mathbb{R}^{50}$, $\mathbf{h} \in \{\pm 1\}^{40}$ is latent. Randomly pick $\mathbf{B}, \mathbf{b}, \mathbf{c}$.

- $q(\mathbf{x}) = p(\mathbf{x})$ with i.i.d. $\mathcal{N}(0, \sigma_{per})$ noise added to all entries of $\mathbf{B}$.
- Sample size $n = 1000$.

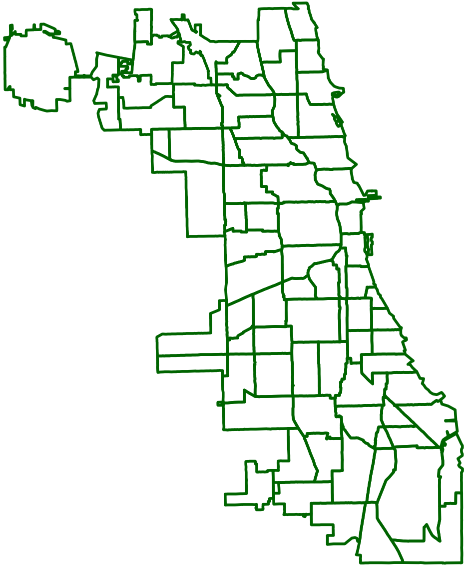


KSD ($\mathcal{O}(n^2)$), FSSD-opt ($\mathcal{O}(n)$) comparable. LKS has low power.

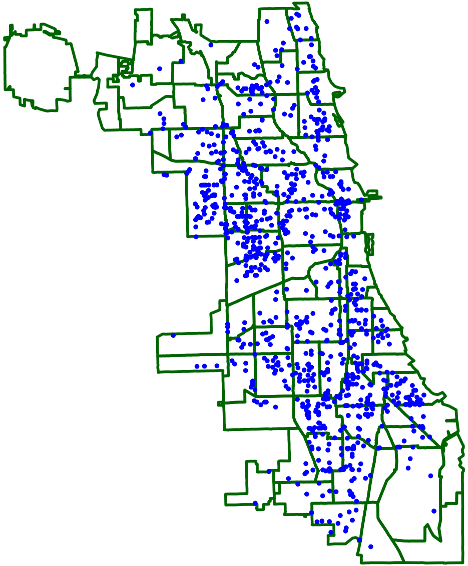
Interpretable Test Locations: Chicago Crime



Interpretable Test Locations: Chicago Crime

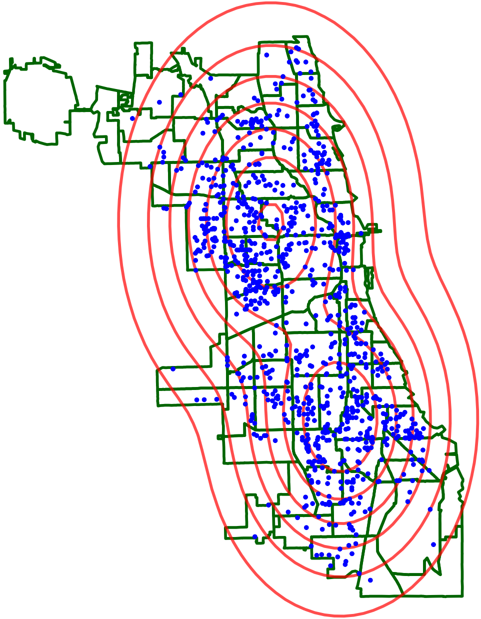


Interpretable Test Locations: Chicago Crime



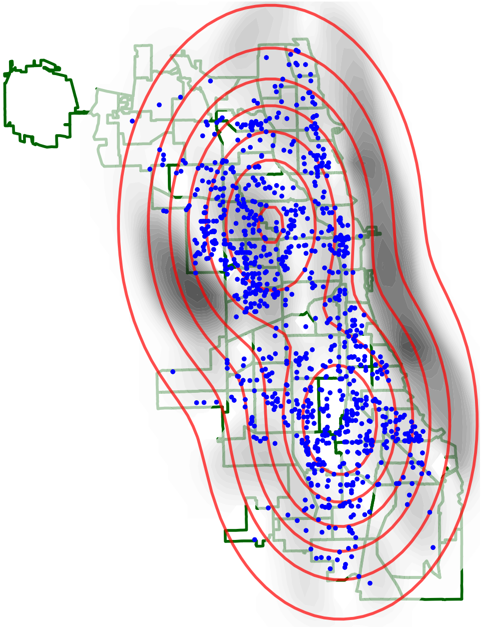
- $n = 11957$ robbery events in Chicago in 2016.
  - lat/long coordinates = sample from q.
- Model spatial density with Gaussian mixtures.

Interpretable Test Locations: Chicago Crime



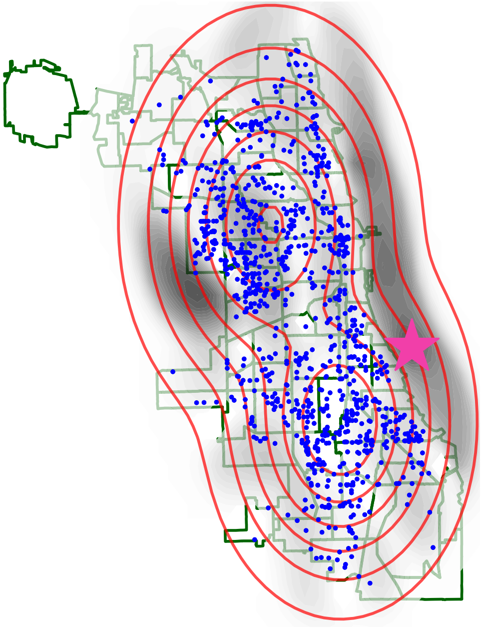
Model p = 2-component Gaussian mixture.

Interpretable Test Locations: Chicago Crime



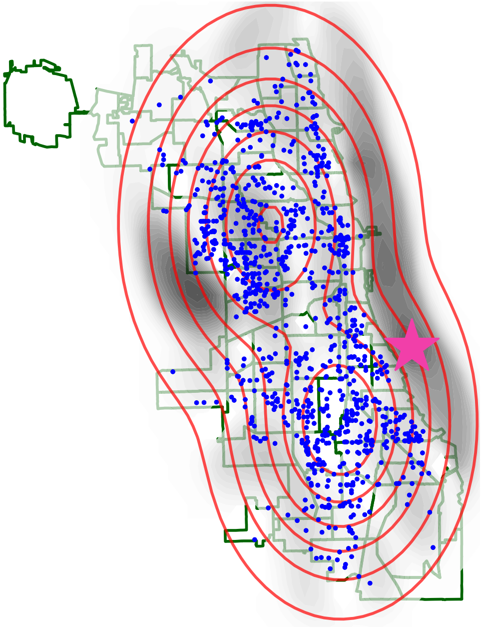
Score surface

Interpretable Test Locations: Chicago Crime



★ = optimized **v**.

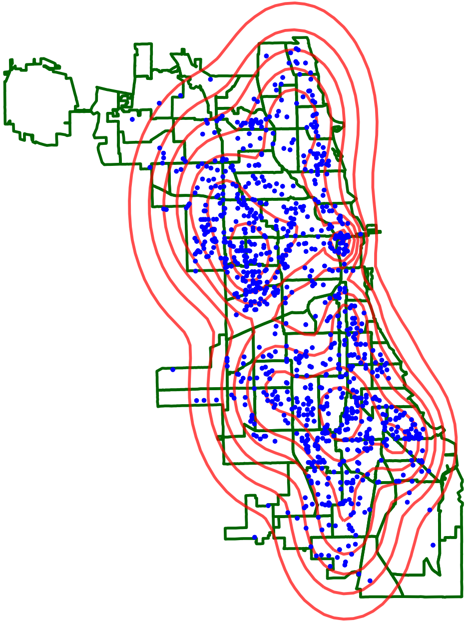
Interpretable Test Locations: Chicago Crime



★ = optimized $\mathbf{v}$.

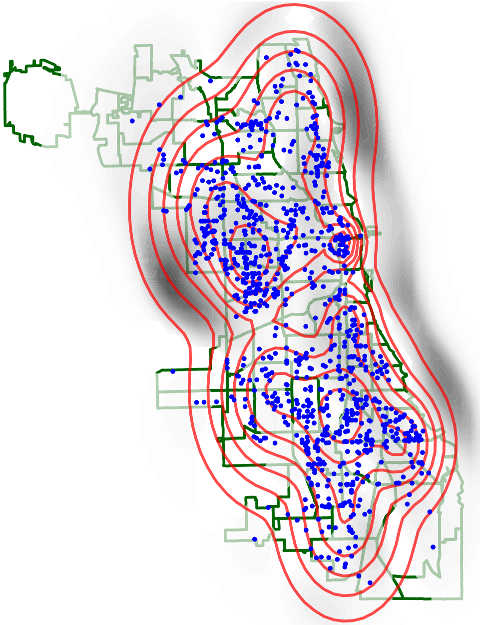
No robbery in Lake Michigan.

Interpretable Test Locations: Chicago Crime



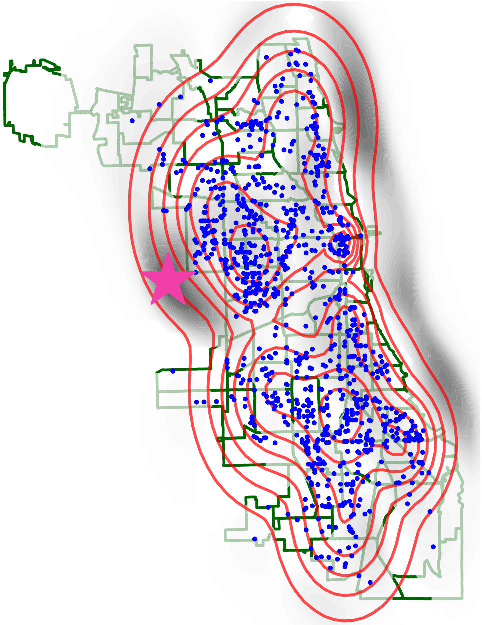
Model p = 10-component Gaussian mixture.

Interpretable Test Locations: Chicago Crime



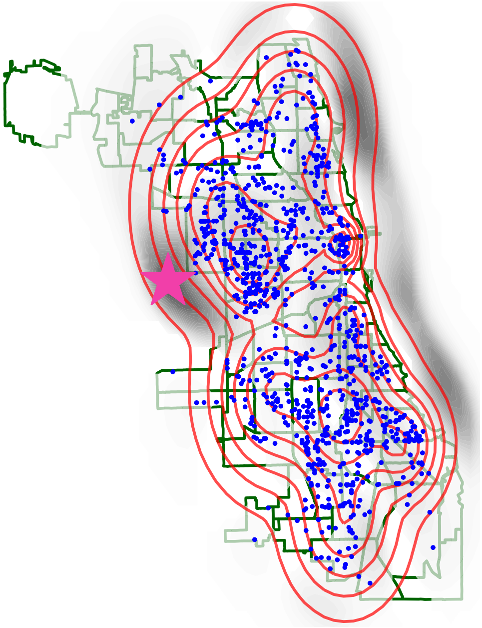
Capture the right tail better.

Interpretable Test Locations: Chicago Crime



Still, does not capture the left tail.

Interpretable Test Locations: Chicago Crime



Still, does not capture the left tail.

Learned test locations are interpretable.

Bahadur Slope and Bahadur Efficiency

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

$$H_0 \colon \theta = \mathbf{0},$$
$$H_1 \colon \theta \neq \mathbf{0}.$$

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

Bahadur slope

$$c(\theta) := -2 \operatorname*{plim}_{n\to\infty} \frac{\log\left(1 - F(T_n)\right)}{n},$$

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

$$H_0 : \theta = \mathbf{0},$$
$$H_1 : \theta \neq \mathbf{0}.$$

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

Bahadur slope

$$c(\theta) := -2 \operatorname*{plim}_{n\to\infty} \frac{\log\left(1 - F(T_n)\right)}{n},$$

where $F(t)$ = CDF of T_n under H_0.

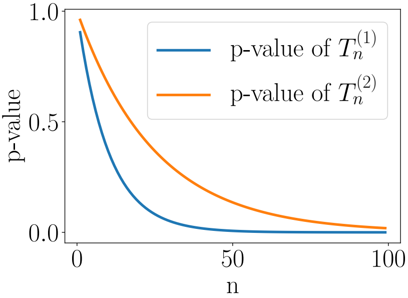
- Bahadur efficiency = ratio of slopes of two tests.

Bahadur Slope and Bahadur Efficiency

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

$$H_0: \boldsymbol{\theta} = \mathbf{0},$$
$$H_1: \boldsymbol{\theta} \neq \mathbf{0}.$$

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



Bahadur slope

$$c(\boldsymbol{\theta}) := -2 \operatorname*{plim}_{n\to\infty} \frac{\log\left(1 - F(T_n)\right)}{n},$$

where $F(t) = $ CDF of T_n under H_0.

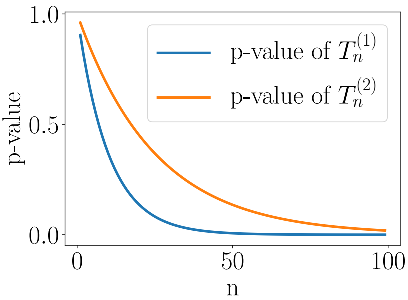
- Bahadur efficiency = ratio of slopes of two tests.

Bahadur Slope and Bahadur Efficiency

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

$$H_0 \colon \theta = \mathbf{0},$$
$$H_1 \colon \theta \neq \mathbf{0}.$$

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



Bahadur slope

$$c(\theta) := -2 \operatorname*{plim}_{n\to\infty} \frac{\log(1 - F(T_n))}{n},$$

where $F(t) =$ CDF of T_n under H_0.

- Bahadur efficiency = ratio of slopes of two tests.

Gaussian Mean Shift Problem

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

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

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

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

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

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

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

$$\frac{c^{(\mathrm{FSSD})}(\mu_q, v, \sigma_k^2)}{c^{(\mathrm{LKS})}(\mu_q, \kappa^2)} > 2.$$

Gaussian Mean Shift Problem

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

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

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

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

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

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

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

$$\frac{c^{(\mathrm{FSSD})}(\mu_q, v, \sigma_k^2)}{c^{(\mathrm{LKS})}(\mu_q, \kappa^2)} > 2.$$

Gaussian Mean Shift Problem

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

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

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

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

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

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

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

$$\frac{c^{(\mathrm{FSSD})}(\mu_q, v, \sigma_k^2)}{c^{(\mathrm{LKS})}(\mu_q, \kappa^2)} > 2.$$

Gaussian Mean Shift Problem

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

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

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

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

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

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

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

$$\frac{c^{(\mathrm{FSSD})}(\mu_q, v, \sigma_k^2)}{c^{(\mathrm{LKS})}(\mu_q, \kappa^2)} > 2.$$

Conclusion

- Proposed The Finite Set Stein Discrepancy (**FSSD**).
- Goodness-of-fit test based on FSSD is
  1. nonparametric,
  2. linear-time,
  3. tunable (parameters automatically tuned).
  4. interpretable.

A Linear-Time Kernel Goodness-of-Fit Test
Wittawat Jitkrittum, Wenkai Xu, Zoltán Szabó, Kenji Fukumizu, Arthur Gretton
NIPS 2017 (best paper award)
Python code: `https://github.com/wittawatj/kgof`

Questions?

Thank you

Illustration: Score Surface

- Consider $J = 1$ location.
- $\text{score}(\mathbf{v}) = \frac{\widehat{\text{FSSD}^2}(\mathbf{v})}{\widehat{\sigma_{H_1}}(\mathbf{v})}$ (gray), p in wireframe, $\{\mathbf{x}_i\}_{i=1}^n \sim q$ in purple, ★ = best $\mathbf{v}$.

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

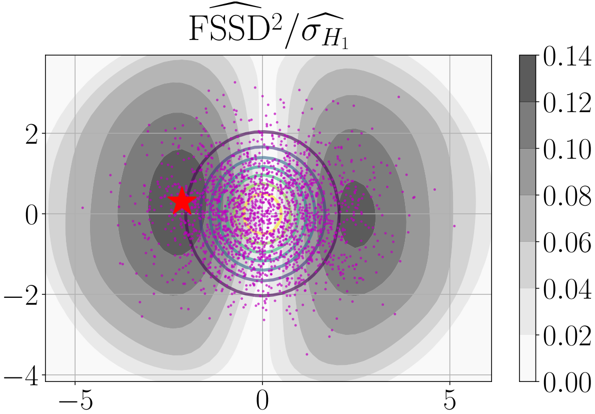
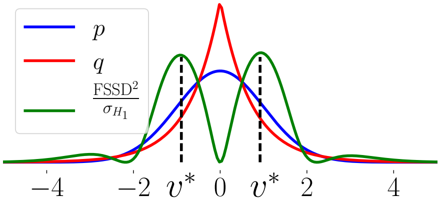
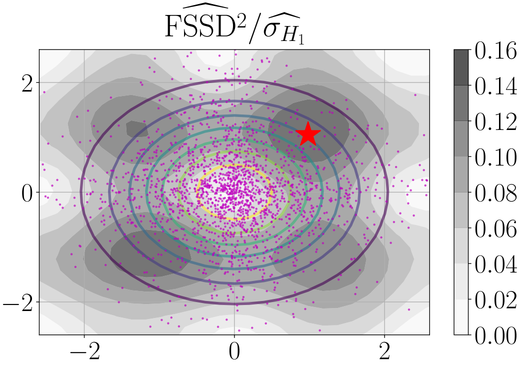


Illustration: Score Surface

- Consider $J = 1$ location.
- $\text{score}(\mathbf{v}) = \frac{\widehat{\text{FSSD}^2}(\mathbf{v})}{\widehat{\sigma_{H_1}}(\mathbf{v})}$ (gray), p in wireframe, $\{\mathbf{x}_i\}_{i=1}^n \sim q$ in purple, ★ = best $\mathbf{v}$.

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





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 $n\widehat{\mathrm{FSSD}^2}$. Gaussian kernel (bandwidth $= \sigma_k^2$).

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

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

$$S^2 = \frac{\mu_q^2 \left(\kappa^2 + 2\sigma_q^2\right) + \left(\sigma_q^2 - 1\right)^2}{\left(\kappa^2 + 2\sigma_q^2\right) \sqrt{\frac{2\sigma_q^2}{\kappa^2} + 1}}.$$

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 $n\widehat{\mathrm{FSSD}^2}$. Gaussian kernel (bandwidth $= \sigma_k^2$).

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

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

$$S^2 = \frac{\mu_q^2\left(\kappa^2+2\sigma_q^2\right) + \left(\sigma_q^2-1\right)^2}{\left(\kappa^2+2\sigma_q^2\right)\sqrt{\frac{2\sigma_q^2}{\kappa^2}+1}}.$$

FSSD is a Discrepancy Measure

Theorem 3.

Let $V = \{\mathbf{v}_1, \ldots, \mathbf{v}_J\} \subset \mathbb{R}^d$ be drawn i.i.d. from a distribution η 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) $\|g\|_{\mathcal{F}}^2 < \infty$.*
3. *(Finite Fisher divergence) $\mathbb{E}_{\mathbf{x}\sim q}\|\nabla_{\mathbf{x}} \log \frac{p(\mathbf{x})}{q(\mathbf{x})}\|^2 < \infty$.*
4. *(Vanishing boundary) $\lim_{\|\mathbf{x}\|\to\infty} p(\mathbf{x})\mathbf{g}(\mathbf{x}) = \mathbf{0}$.*

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

$$\mathrm{FSSD}^2 = 0 \text{ if and only if } p = q.$$

- Gaussian kernel $k(\mathbf{x}, \mathbf{v}) = \exp\left(-\frac{\|\mathbf{x}-\mathbf{v}\|_2^2}{2\sigma_k^2}\right)$ works.
- In practice, $J = 1$ or $J = 5$.

Asymptotic Distributions of $\widehat{\mathrm{FSSD}^2}$

- Recall $\xi(\mathbf{x}, \mathbf{v}) := \frac{1}{p(\mathbf{x})}\partial_{\mathbf{x}}[k(\mathbf{x}, \mathbf{v})p(\mathbf{x})] \in \mathbb{R}^d$.
- $\tau(\mathbf{x}) :=$ vertically stack $\xi(\mathbf{x}, \mathbf{v}_1), \ldots \xi(\mathbf{x}, \mathbf{v}_J) \in \mathbb{R}^{dJ}$. Feature vector of $\mathbf{x}$.
- Mean feature: $\mu := \mathbb{E}_{\mathbf{x}\sim q}[\tau(\mathbf{x})]$.
- $\Sigma_r := \mathrm{cov}_{\mathbf{x}\sim r}[\tau(\mathbf{x})] \in \mathbb{R}^{dJ\times dJ}$ for $r \in \{p, q\}$

Proposition 2 (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 Σ_p.

1. *Under $H_0 : p = q$, asymptotically $n\widehat{\mathrm{FSSD}^2} \xrightarrow{d} \sum_{i=1}^{dJ}(Z_i^2 - 1)\omega_i$.*
   - *Easy to simulate to get p-value.*
   - *Simulation cost independent of n.*
2. *Under $H_1 : p \neq q$, we have $\sqrt{n}(\widehat{\mathrm{FSSD}^2} - \mathrm{FSSD}^2) \xrightarrow{d} \mathcal{N}(0, \sigma^2_{H_1})$ where $\sigma^2_{H_1} := 4\mu^\top\Sigma_q\mu$. Implies $\mathbb{P}$(reject H_0) $\to 1$ as $n \to \infty$.*

But, how to estimate Σ_p? No sample from p!

- Theorem: Using $\hat{\Sigma}_q$ (computed with $\{\mathbf{x}_i\}_{i=1}^n \sim q$) still leads to a consistent test.

Asymptotic Distributions of $\widehat{\mathrm{FSSD}^2}$

- Recall $\xi(\mathbf{x}, \mathbf{v}) := \frac{1}{p(\mathbf{x})} \partial_{\mathbf{x}} [k(\mathbf{x}, \mathbf{v}) p(\mathbf{x})] \in \mathbb{R}^d$.
- $\tau(\mathbf{x}) :=$ vertically stack $\xi(\mathbf{x}, \mathbf{v}_1), \ldots \xi(\mathbf{x}, \mathbf{v}_J) \in \mathbb{R}^{dJ}$. Feature vector of $\mathbf{x}$.
- Mean feature: $\mu := \mathbb{E}_{\mathbf{x} \sim q}[\tau(\mathbf{x})]$.
- $\Sigma_r := \mathrm{cov}_{\mathbf{x} \sim r}[\tau(\mathbf{x})] \in \mathbb{R}^{dJ \times dJ}$ for $r \in \{p, q\}$

Proposition 2 (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 Σ_p.

1. *Under $H_0 : p = q$, asymptotically $n\widehat{\mathrm{FSSD}^2} \xrightarrow{d} \sum_{i=1}^{dJ} (Z_i^2 - 1)\omega_i$.*
   - *Easy to simulate to get p-value.*
   - *Simulation cost independent of n.*
2. *Under $H_1 : p \neq q$, we have $\sqrt{n}(\widehat{\mathrm{FSSD}^2} - \mathrm{FSSD}^2) \xrightarrow{d} \mathcal{N}(0, \sigma_{H_1}^2)$ 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 Σ_p? No sample from p!

- Theorem: Using $\hat{\Sigma}_q$ (computed with $\{\mathbf{x}_i\}_{i=1}^n \sim q$) still leads to a consistent test.

Asymptotic Distributions of $\widehat{\text{FSSD}^2}$

- Recall $\xi(\mathbf{x}, \mathbf{v}) := \frac{1}{p(\mathbf{x})}\partial_{\mathbf{x}}[k(\mathbf{x}, \mathbf{v})p(\mathbf{x})] \in \mathbb{R}^d$.
- $\boldsymbol{\tau}(\mathbf{x}) :=$ vertically stack $\xi(\mathbf{x}, \mathbf{v}_1), \ldots \xi(\mathbf{x}, \mathbf{v}_J) \in \mathbb{R}^{dJ}$. Feature vector of $\mathbf{x}$.
- Mean feature: $\boldsymbol{\mu} := \mathbb{E}_{\mathbf{x}\sim q}[\boldsymbol{\tau}(\mathbf{x})]$.
- $\Sigma_r := \text{cov}_{\mathbf{x}\sim r}[\boldsymbol{\tau}(\mathbf{x})] \in \mathbb{R}^{dJ\times dJ}$ for $r \in \{p, q\}$

Proposition 2 (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 Σ_p.

1. *Under $H_0 : p = q$, asymptotically $n\widehat{\text{FSSD}^2} \overset{d}{\to} \sum_{i=1}^{dJ}(Z_i^2 - 1)\omega_i$.*
   - *Easy to simulate to get p-value.*
   - *Simulation cost independent of n.*
2. *Under $H_1 : p \neq q$, we have $\sqrt{n}(\widehat{\text{FSSD}^2} - \text{FSSD}^2) \overset{d}{\to} \mathcal{N}(0, \sigma_{H_1}^2)$ where $\sigma_{H_1}^2 := 4\boldsymbol{\mu}^\top\Sigma_q\boldsymbol{\mu}$. Implies $\mathbb{P}$(reject H_0) $\to 1$ as $n \to \infty$.*

But, how to estimate Σ_p? No sample from p!

- Theorem: Using $\hat{\Sigma}_q$ (computed with $\{\mathbf{x}_i\}_{i=1}^n \sim q$) still leads to a consistent test.

Asymptotic Distributions of $\widehat{\mathrm{FSSD}^2}$

- Recall $\xi(\mathbf{x}, \mathbf{v}) := \frac{1}{p(\mathbf{x})}\partial_{\mathbf{x}}[k(\mathbf{x}, \mathbf{v})p(\mathbf{x})] \in \mathbb{R}^d$.
- $\tau(\mathbf{x}) :=$ vertically stack $\xi(\mathbf{x}, \mathbf{v}_1), \ldots \xi(\mathbf{x}, \mathbf{v}_J) \in \mathbb{R}^{dJ}$. Feature vector of $\mathbf{x}$.
- Mean feature: $\mu := \mathbb{E}_{\mathbf{x}\sim q}[\tau(\mathbf{x})]$.
- $\Sigma_r := \mathrm{cov}_{\mathbf{x}\sim r}[\tau(\mathbf{x})] \in \mathbb{R}^{dJ\times dJ}$ for $r \in \{p, q\}$

Proposition 2 (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 Σ_p.

1. *Under $H_0 : p = q$, asymptotically $n\widehat{\mathrm{FSSD}^2} \xrightarrow{d} \sum_{i=1}^{dJ}(Z_i^2 - 1)\omega_i$.*
   - *Easy to simulate to get p-value.*
   - *Simulation cost independent of n.*
2. *Under $H_1 : p \neq q$, we have $\sqrt{n}(\widehat{\mathrm{FSSD}^2} - \mathrm{FSSD}^2) \xrightarrow{d} \mathcal{N}(0, \sigma_{H_1}^2)$ where $\sigma_{H_1}^2 := 4\mu^\top\Sigma_q\mu$. Implies $\mathbb{P}$(reject H_0) $\to 1$ as $n \to \infty$.*

But, how to estimate Σ_p? No sample from p!

- **Theorem**: Using $\hat{\Sigma}_q$ (computed with $\{\mathbf{x}_i\}_{i=1}^n \sim q$) still leads to a consistent test.

Bahadur Slopes of FSSD and LKS

Theorem 4.

The Bahadur slope of $n\widehat{\mathrm{FSSD}^2}$ *is*

$$c^{(\mathrm{FSSD})} := \mathrm{FSSD}^2/\omega_1,$$

where ω_1 *is the maximum eigenvalue of* $\Sigma_p := \mathrm{cov}_{\mathbf{x}\sim p}[\tau(\mathbf{x})]$.

Theorem 5.

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

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

where h_p *is the U-statistic kernel of the KSD statistic.*

Bahadur Slopes of FSSD and LKS

Theorem 4.

The Bahadur slope of $n\widehat{\mathrm{FSSD}^2}$ *is*

$$c^{(\mathrm{FSSD})} := \mathrm{FSSD}^2/\omega_1,$$

where ω_1 *is the maximum eigenvalue of* $\Sigma_p := \mathrm{cov}_{\mathbf{x}\sim p}[\tau(\mathbf{x})]$.

Theorem 5.

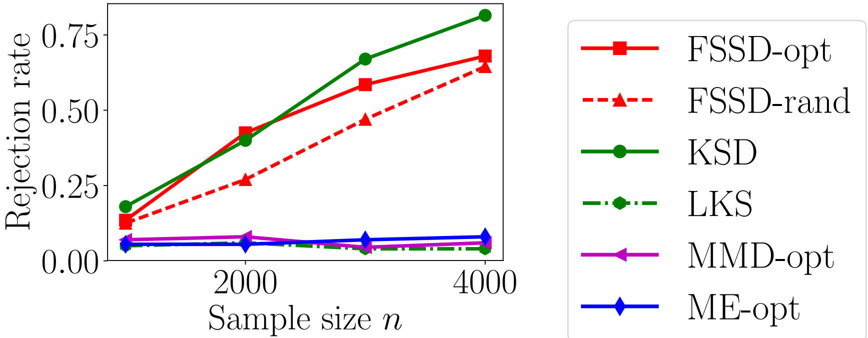
The Bahadur slope of the linear-time kernel Stein (LKS) statistic $\sqrt{n}\widehat{S_l^2}$ *is*

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

where h_p *is the U-statistic kernel of the KSD statistic.*

Harder RBM Problem

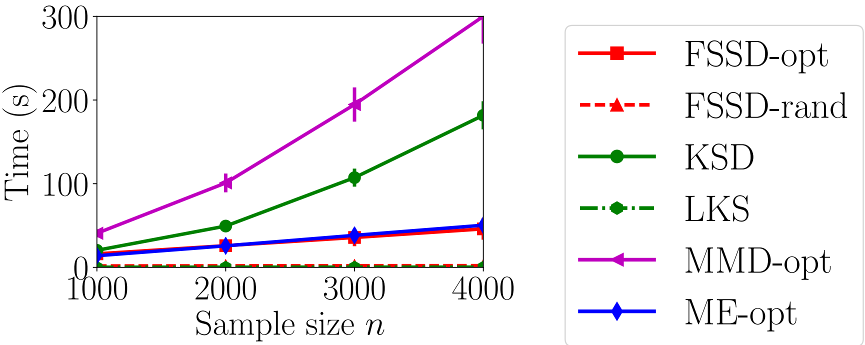
- Perturb only one entry of $\mathbf{B} \in \mathbb{R}^{50\times 40}$ (in the RBM).
- $B_{1,1} \leftarrow B_{1,1} + \mathcal{N}(0, \sigma_{per}^2 = 0.1^2)$.



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



- 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.

References I

Bahadur, R. R. (1960).
Stochastic comparison of tests.
The Annals of Mathematical Statistics, 31(2):276–295.

Chwialkowski, K., Strathmann, H., and Gretton, A. (2016).
A kernel test of goodness of fit.
In *ICML*, pages 2606–2615.

Gretton, A., Borgwardt, K. M., Rasch, M. J., Schölkopf, B., and Smola, A. (2012).
A Kernel Two-Sample Test.
JMLR, 13:723–773.

Jitkrittum, W., Szabó, Z., Chwialkowski, K. P., and Gretton, A. (2016).
Interpretable Distribution Features with Maximum Testing Power.
In *NIPS*, pages 181–189.

References II

Liu, Q., Lee, J., and Jordan, M. (2016).
A Kernelized Stein Discrepancy for Goodness-of-fit Tests.
In *ICML*, pages 276–284.