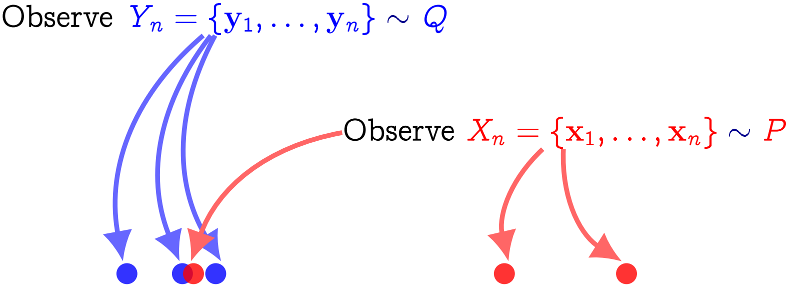
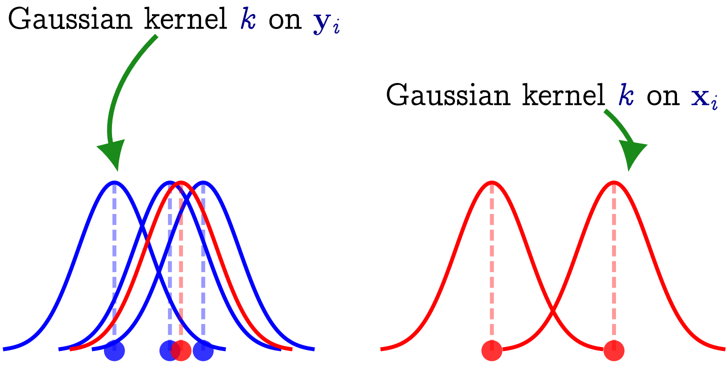
Linear-time, interpretable two-sample test

Recall from part 1: the MMD witness (Gretton et al., 2012)

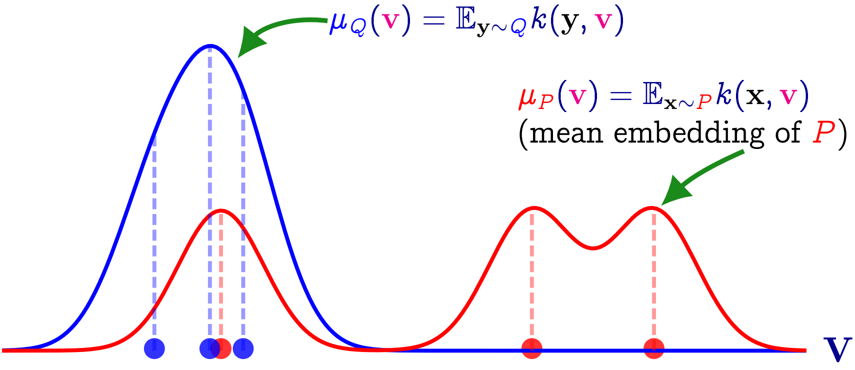
Recall from part 1: the MMD witness (Gretton et al., 2012)



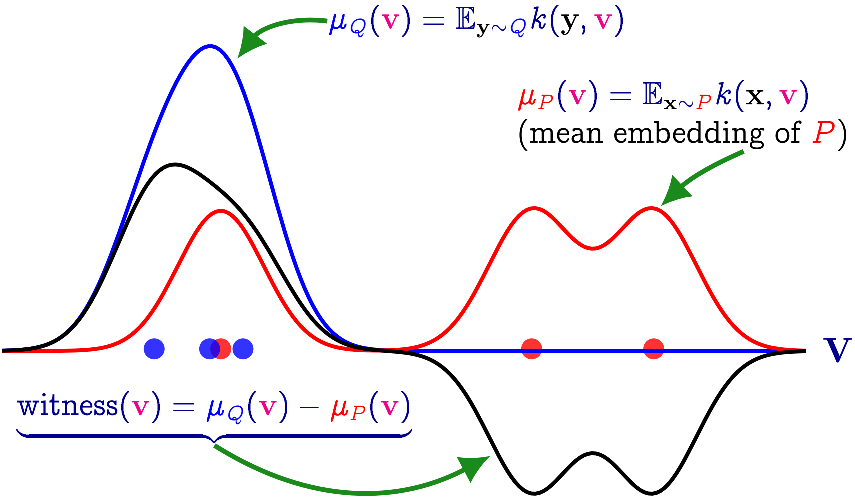
Recall from part 1: the MMD witness (Gretton et al., 2012)



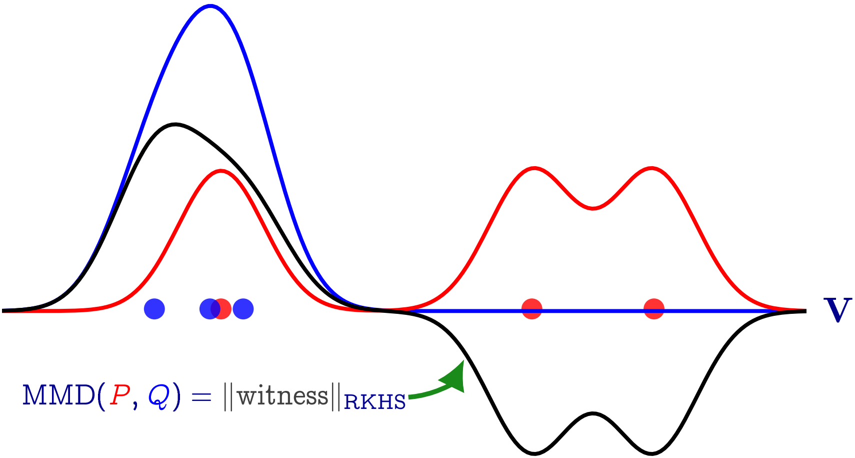
Recall from part 1: the MMD witness (Gretton et al., 2012)



Recall from part 1: the MMD witness (Gretton et al., 2012)

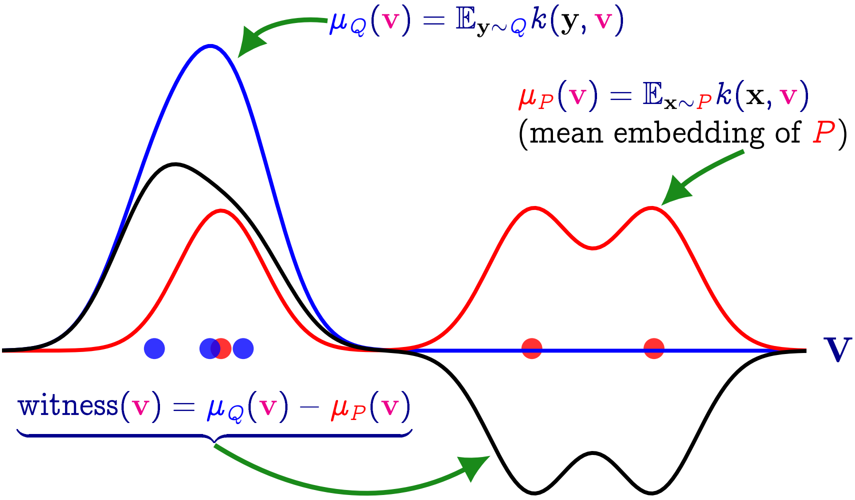


Recall from part 1: the MMD witness (Gretton et al., 2012)



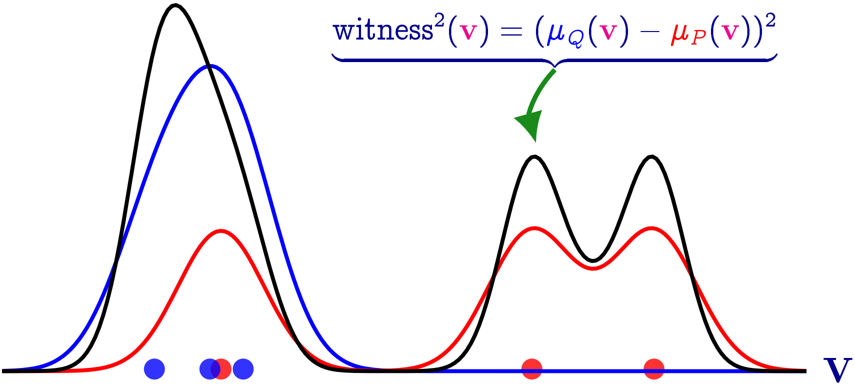
$\mathrm{MMD}(P, Q) = \|\text{witness}\|_{\mathrm{RKHS}}$

The Unnormalized Mean Embeddings statistic (Chwialkowski et al., 2015)



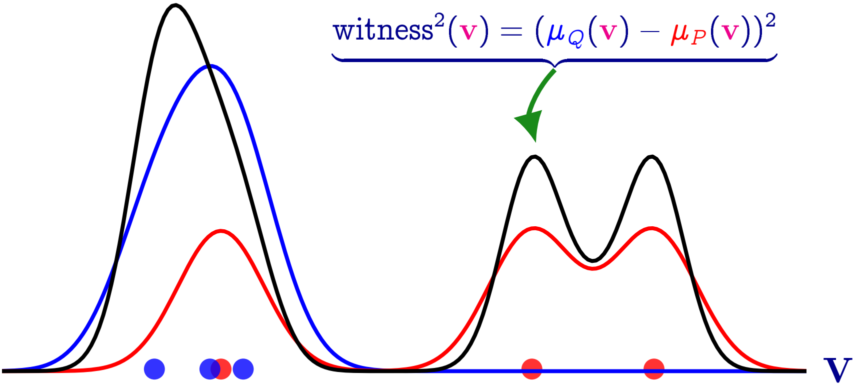
The Unnormalized Mean Embeddings statistic (Chwialkowski et al., 2015)

$$\underbrace{\text{witness}^2(\mathbf{v}) = (\mu_Q(\mathbf{v}) - \mu_P(\mathbf{v}))^2}$$



The Unnormalized Mean Embeddings statistic (Chwialkowski et al., 2015)

$$\underbrace{\text{witness}^2(\mathbf{v}) = (\mu_Q(\mathbf{v}) - \mu_P(\mathbf{v}))^2}$$



- Given J **test locations** $V := \{\mathbf{v}_j\}_{j=1}^J$, ($V$ gives interpretability later)

$$\text{UME}^2(P, Q) = \frac{1}{J}\sum_{j=1}^{J}[\mu_P(\mathbf{v}_j) - \mu_Q(\mathbf{v}_j)]^2.$$

The Unnormalized Mean Embeddings (UME) statistic

$$\mathrm{UME}^2(P, Q) = \frac{1}{J}\sum_{j=1}^{J}[\mu_P(\mathbf{v}_j) - \mu_Q(\mathbf{v}_j)]^2 = \frac{1}{J}\sum_{j=1}^{J}\mathrm{witness}^2(\mathbf{v}_j).$$

Proposition (Chwialkowski et al., NeurIPS 2015)

Main assumptions:

1. *Nice kernel k (characteristic, real analytic).*
2. $\{\mathbf{v}_j\}_{j=1}^{J}$ *drawn from a distribution that covers the whole domain.*

$$\boxed{\mathrm{UME}^2(P, Q) = 0 \textit{ iff } P = Q.}$$

- Key: Evaluating witness^2 is enough to detect the difference.
- Runtime complexity: $\mathcal{O}(Jn)$. J is constant.

The Unnormalized Mean Embeddings (UME) statistic

$$\mathrm{UME}^2(P, Q) = \frac{1}{J}\sum_{j=1}^{J}[\mu_P(\mathbf{v}_j) - \mu_Q(\mathbf{v}_j)]^2 = \frac{1}{J}\sum_{j=1}^{J}\mathrm{witness}^2(\mathbf{v}_j).$$

Proposition (Chwialkowski et al., NeurIPS 2015)

Main assumptions:

1. *Nice kernel k (characteristic, real analytic).*
2. $\{\mathbf{v}_j\}_{j=1}^{J}$ *drawn from a distribution that covers the whole domain.*

$$\boxed{\mathrm{UME}^2(P, Q) = 0 \textit{ iff } P = Q.}$$

- **Key**: Evaluating witness^2 is enough to detect the difference.
- Runtime complexity: $\mathcal{O}(Jn)$. J is constant.

Normalized ME (NME) statistic (Chwialkowski et al., 2015, Jitkrittum et al., 2016)

- Null distribution P_{H_0} of UME is complicated.
  - Weighted sum of correlated chi-squares. No closed form.
- **Idea**: decorrelate the J terms in the sum.

$$\mathrm{UME}^2(P, Q) = \mathbf{t}^\top \mathbf{t} \text{ where } \mathbf{t} \in \mathbb{R}^J$$

Normalized ME (NME)

$$\mathrm{NME}^2(P, Q) = \mathbf{t}^\top \mathbf{C}^{-1} \mathbf{t}$$

where $\mathbf{C}$ = covariance of the J terms ($J \times J$ matrix).

- $\mathbf{t}, \mathbf{C}$ depend on samples from P, Q and test locations $\{\mathbf{v}_j\}_{j=1}^J$.
- Runtime complexity: $\mathcal{O}(J^3 + J^2 n + Jdn)$. Linear in n.

- Null distribution P_{H_0} of UME is complicated.
  - Weighted sum of correlated chi-squares. No closed form.
- **Idea**: decorrelate the J terms in the sum.

$$\mathrm{UME}^2(P, Q) = \mathbf{t}^\top \mathbf{t} \text{ where } \mathbf{t} \in \mathbb{R}^J$$

Normalized ME (NME)

$$\mathrm{NME}^2(P, Q) = \mathbf{t}^\top \mathbf{C}^{-1} \mathbf{t}$$

where $\mathbf{C}$ = covariance of the J terms ($J \times J$ matrix).

- $\mathbf{t}, \mathbf{C}$ depend on samples from P, Q and test locations $\{\mathbf{v}_j\}_{j=1}^J$.
- Runtime complexity: $\mathcal{O}(J^3 + J^2 n + Jdn)$. Linear in n.

Normalized ME (NME) statistic (Chwialkowski et al., 2015, Jitkrittum et al., 2016)

- Null distribution P_{H_0} of UME is complicated.
  - Weighted sum of correlated chi-squares. No closed form.
- **Idea**: decorrelate the J terms in the sum.

$$\mathrm{UME}^2(P, Q) = \mathbf{t}^\top \mathbf{t} \text{ where } \mathbf{t} \in \mathbb{R}^J$$

Normalized ME (NME)

$$\mathrm{NME}^2(P, Q) = \mathbf{t}^\top \mathbf{C}^{-1} \mathbf{t}$$

where $\mathbf{C}$ = covariance of the J terms ($J \times J$ matrix).

- $\mathbf{t}, \mathbf{C}$ depend on samples from P, Q and test locations $\{\mathbf{v}_j\}_{j=1}^J$.
- Runtime complexity: $\mathcal{O}(J^3 + J^2 n + Jdn)$. Linear in n.

Normalized ME (NME) statistic (Chwialkowski et al., 2015, Jitkrittum et al., 2016)

- Null distribution P_{H_0} of UME is complicated.
  - Weighted sum of correlated chi-squares. No closed form.
- **Idea**: decorrelate the J terms in the sum.

$$\mathrm{UME}^2(P, Q) = \mathbf{t}^\top \mathbf{t} \text{ where } \mathbf{t} \in \mathbb{R}^J$$

Normalized ME (NME)

$$\mathrm{NME}^2(P, Q) = \mathbf{t}^\top \mathbf{C}^{-1} \mathbf{t}$$

where $\mathbf{C} =$ covariance of the J terms ($J \times J$ matrix).

- $\mathbf{t}, \mathbf{C}$ depend on samples from P, Q and test locations $\{\mathbf{v}_j\}_{j=1}^J$.
- Runtime complexity: $\mathcal{O}(J^3 + J^2 n + Jdn)$. Linear in n.

Normalized ME (NME) statistic (Chwialkowski et al., 2015, Jitkrittum et al., 2016)

- Null distribution P_{H_0} of UME is complicated.
  - Weighted sum of correlated chi-squares. No closed form.
- **Idea**: decorrelate the J terms in the sum.

$$\mathrm{UME}^2(P, Q) = \mathbf{t}^\top \mathbf{t} \text{ where } \mathbf{t} \in \mathbb{R}^J$$

Normalized ME (NME)

$$\mathrm{NME}^2(P, Q) = \mathbf{t}^\top \mathbf{C}^{-1} \mathbf{t}$$

where $\mathbf{C} =$ covariance of the J terms ($J \times J$ matrix).

- $\mathbf{t}, \mathbf{C}$ depend on samples from P, Q and test locations $\{\mathbf{v}_j\}_{j=1}^J$.
- Runtime complexity: $\mathcal{O}(J^3 + J^2 n + Jdn)$. Linear in n.

Asymptotic distributions of NME

Proposition (Chwialkowski et al., 2015, Jitkrittum et al., 2016)

As sample size $n \to \infty$,

1. *When $P = Q$, $n\widehat{\mathrm{NME}}^2$ follows χ^2_J (chi-square).*
2. *When $P \neq Q$, the test power goes to 1.*

Proposition (Jitkrittum et al., 2016)

Choosing $\{\mathbf{v}_j\}_{j=1}^J$ by maximizing $\widehat{\mathrm{NME}}^2$ will maximize (a lower bound on) the test power. ▸ see lower bound

Optimized locations $\{\mathbf{v}_j\}_{j=1}^J$ are interpretable.
Indicate where P, Q differ most.

Asymptotic distributions of NME

Proposition (Chwialkowski et al., 2015, Jitkrittum et al., 2016)

As sample size $n \to \infty$,

1. *When* $P = Q$, $n\widehat{\mathrm{NME}^2}$ *follows* χ^2_J *(chi-square).*
2. *When* $P \neq Q$, *the test power goes to 1.*

Proposition (Jitkrittum et al., 2016)

Choosing $\{\mathbf{v}_j\}_{j=1}^J$ *by maximizing* $\widehat{\mathrm{NME}^2}$ *will maximize (a lower bound on) the test power.* see lower bound

Optimized locations $\{\mathbf{v}_j\}_{j=1}^J$ are interpretable.
Indicate where P, Q differ most.

Asymptotic distributions of NME

Proposition (Chwialkowski et al., 2015, Jitkrittum et al., 2016)

As sample size $n \to \infty$,

1. *When* $P = Q$, $n\widehat{\mathrm{NME}^2}$ *follows* χ^2_J *(chi-square).*
2. *When* $P \neq Q$, *the test power goes to 1.*

Proposition (Jitkrittum et al., 2016)

Choosing $\{\mathbf{v}_j\}_{j=1}^J$ *by maximizing* $\widehat{\mathrm{NME}^2}$ *will maximize (a lower bound on) the test power.* ▸ *see lower bound*

Optimized locations $\{\mathbf{v}_j\}_{j=1}^J$ are interpretable.
Indicate where P, Q differ most.

Illustration: a good location $\mathbf{v}$ for NME

- Use $J = 1$ location.

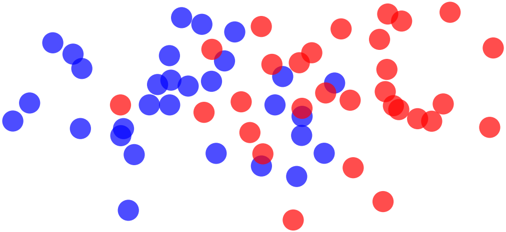


Illustration: a good location $\mathbf{v}$ for NME

- Use $J = 1$ location. Let $\text{score}(\mathbf{v}) := \widehat{\text{NME}^2}$.

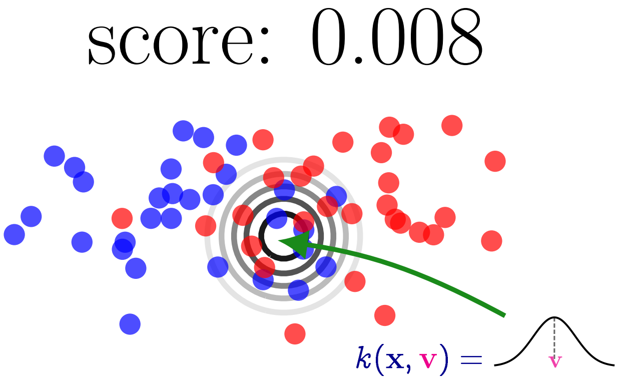


Illustration: a good location $\mathbf{v}$ for NME

- Use $J = 1$ location. Let $\text{score}(\mathbf{v}) := \widehat{\text{NME}^2}$.

score: 1.6

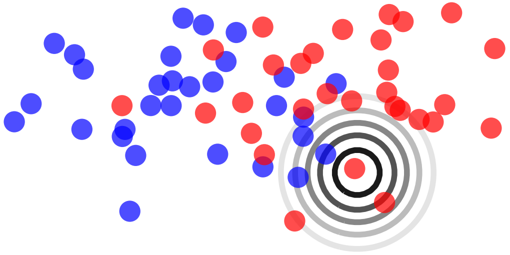


Illustration: a good location $\mathbf{v}$ for NME

- Use $J = 1$ location. Let $\text{score}(\mathbf{v}) := \widehat{\text{NME}^2}$.

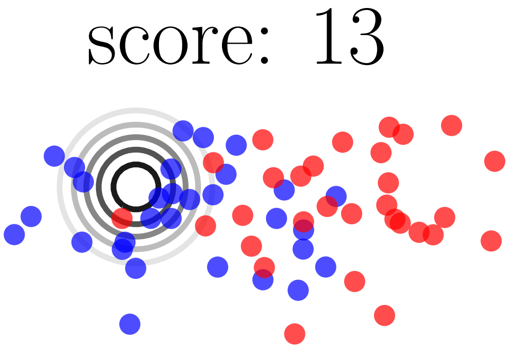


Illustration: a good location $\mathbf{v}$ for NME

- Use $J = 1$ location. Let $\text{score}(\mathbf{v}) := \widehat{\text{NME}^2}$.

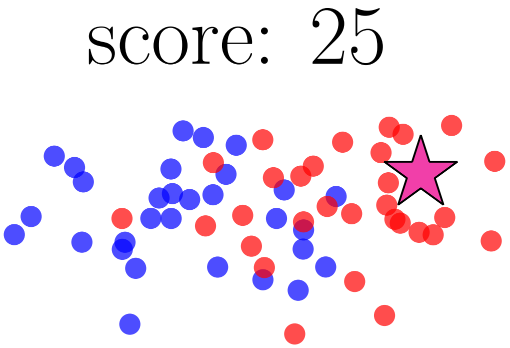
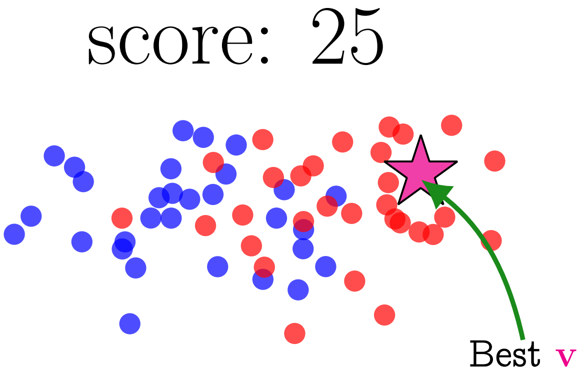


Illustration: a good location $\mathbf{v}$ for NME

- Use $J = 1$ location. Let $\mathrm{score}(\mathbf{v}) := \widehat{\mathrm{NME}^2}$.



- Best $\mathbf{v}$ reveals where P and Q differ most.
- Maximizes the probability of detecting differences between P and Q.

NME: distinguishing positive/negative emotions

P :

happy neutral surprised

Q:

afraid angry disgusted

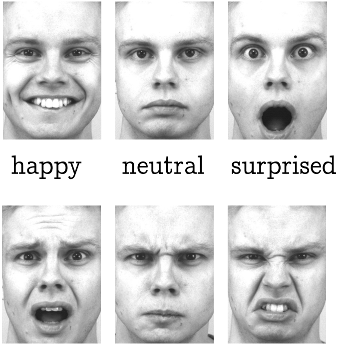
- 35 females and 35 males (Lundqvist et al., 1998).
- $48 \times 34 = 1632$ dimensions. Pixel features.
- $n = 201$.

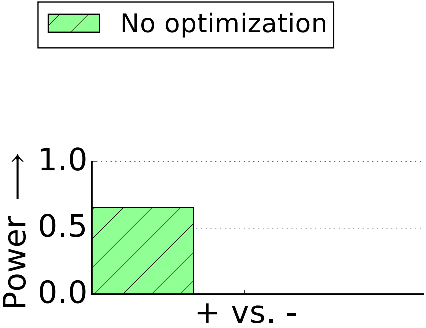
- Test power comparable to the state-of-the-art MMD test.
- Informative features: differences at the nose, and smile lines.

NME: distinguishing positive/negative emotions

P : happy, neutral, surprised

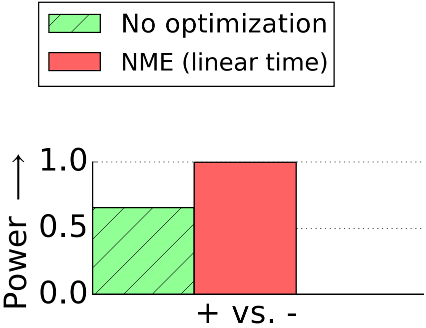
Q: afraid, angry, disgusted





- Test power comparable to the state-of-the-art MMD test.
- Informative features: differences at the nose, and smile lines.

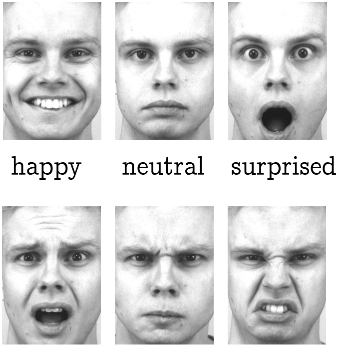
NME: distinguishing positive/negative emotions



- Test power comparable to the state-of-the-art MMD test.
- Informative features: differences at the nose, and smile lines.

NME: distinguishing positive/negative emotions

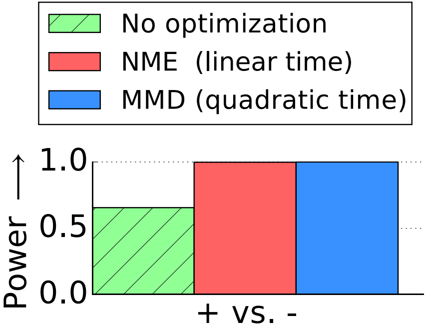
P :



happy neutral surprised

Q:

afraid angry disgusted



- Test power comparable to the state-of-the-art MMD test.
- Informative features: differences at the nose, and smile lines.

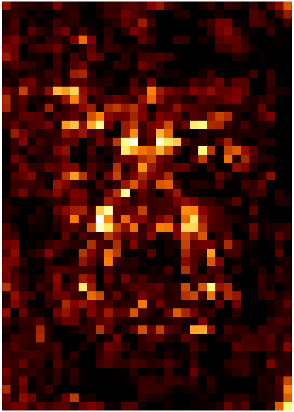
NME: distinguishing positive/negative emotions

P :

happy neutral surprised

Q:

afraid angry disgusted



Learned $\mathbf{v}$

- Test power comparable to the state-of-the-art MMD test.
- **Informative features**: differences at the nose, and smile lines.

NME: distinguishing positive/negative emotions

P :

happy neutral surprised

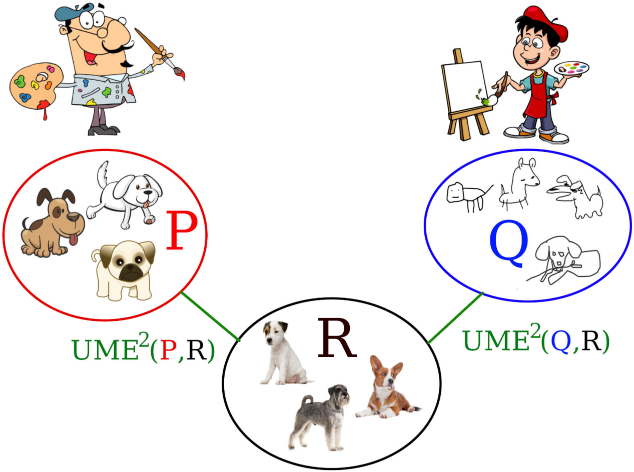
Q:

afraid angry disgusted

Learned $\mathbf{v}$

- Test power comparable to the state-of-the-art MMD test.
- **Informative features**: differences at the nose, and smile lines.

Extension: model comparison by relative UME



- Both models P, Q can be wrong.
- **Goal**: pick the better one.

A model comparison test (Jitkrittum et al., 2018)

- P, Q : two candidate generative models that can be sampled.
- R : true distribution (unknown).
- Observe $X_n \overset{i.i.d.}{\sim} P$, $Y_n \overset{i.i.d.}{\sim} Q$, and $Z_n \overset{i.i.d.}{\sim} R$. Three sets.

$$H_0 : \mathrm{UME}_V^2(P, R) - \mathrm{UME}_V^2(Q, R) \leq 0$$
$$H_1 : \mathrm{UME}_V^2(P, R) - \mathrm{UME}_V^2(Q, R) > 0$$

- Statistic: $\hat{S}_n = \widehat{\mathrm{UME}}_V^2(P, R) - \widehat{\mathrm{UME}}_V^2(Q, R)$.
- Reject H_0 if $\hat{S}_n$ is too large.

> Optimize V by maximizing power of relative UME test.
> V shows where Q is better than P.

A model comparison test (Jitkrittum et al., 2018)

- P, Q : two candidate generative models that can be sampled.
- R : true distribution (unknown).
- Observe $X_n \overset{i.i.d.}{\sim} P$, $Y_n \overset{i.i.d.}{\sim} Q$, and $Z_n \overset{i.i.d.}{\sim} R$. Three sets.

$$H_0: \mathrm{UME}_V^2(P,R) - \mathrm{UME}_V^2(Q,R) \le 0$$
$$H_1: \mathrm{UME}_V^2(P,R) - \mathrm{UME}_V^2(Q,R) > 0$$

- Statistic: $\hat{S}_n = \widehat{\mathrm{UME}}_V^2(P,R) - \widehat{\mathrm{UME}}_V^2(Q,R)$.
- Reject H_0 if $\hat{S}_n$ is too large.

Optimize V by maximizing power of relative UME test.
V shows where Q is better than P.

A model comparison test (Jitkrittum et al., 2018)

- P, Q : two candidate generative models that can be sampled.
- R : true distribution (unknown).
- Observe $X_n \overset{i.i.d.}{\sim} P$, $Y_n \overset{i.i.d.}{\sim} Q$, and $Z_n \overset{i.i.d.}{\sim} R$. Three sets.

$$H_0: \mathrm{UME}^2_V(P, R) - \mathrm{UME}^2_V(Q, R) \leq 0$$
$$H_1: \mathrm{UME}^2_V(P, R) - \mathrm{UME}^2_V(Q, R) > 0$$

- Statistic: $\hat{S}_n = \widehat{\mathrm{UME}}^2_V(P, R) - \widehat{\mathrm{UME}}^2_V(Q, R)$.
- Reject H_0 if $\hat{S}_n$ is too large.

Optimize V by maximizing power of relative UME test.
V shows where Q is better than P.

A model comparison test (Jitkrittum et al., 2018)

- P, Q : two candidate generative models that can be sampled.
- R : true distribution (unknown).
- Observe $X_n \overset{i.i.d.}{\sim} P$, $Y_n \overset{i.i.d.}{\sim} Q$, and $Z_n \overset{i.i.d.}{\sim} R$. Three sets.

$$H_0 : \mathrm{UME}^2_V(P, R) - \mathrm{UME}^2_V(Q, R) \leq 0$$
$$H_1 : \mathrm{UME}^2_V(P, R) - \mathrm{UME}^2_V(Q, R) > 0$$

- Statistic: $\hat{S}_n = \widehat{\mathrm{UME}}^2_V(P, R) - \widehat{\mathrm{UME}}^2_V(Q, R)$.
- Reject H_0 if $\hat{S}_n$ is too large.

Optimize V by maximizing power of relative UME test.
V shows where Q is better than P.

A model comparison test (Jitkrittum et al., 2018)

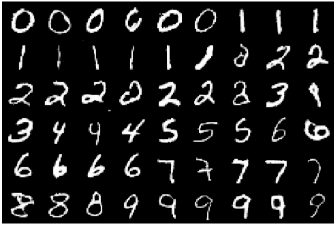
- P, Q : two candidate generative models that can be sampled.
- R : true distribution (unknown).
- Observe $X_n \overset{i.i.d.}{\sim} P$, $Y_n \overset{i.i.d.}{\sim} Q$, and $Z_n \overset{i.i.d.}{\sim} R$. Three sets.

$$H_0 : \mathrm{UME}^2_V(P, R) - \mathrm{UME}^2_V(Q, R) \leq 0$$
$$H_1 : \mathrm{UME}^2_V(P, R) - \mathrm{UME}^2_V(Q, R) > 0$$

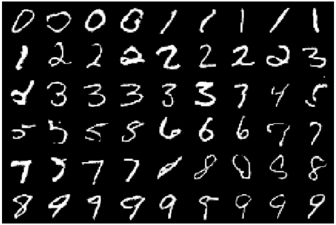
- Statistic: $\hat{S}_n = \widehat{\mathrm{UME}}^2_V(P, R) - \widehat{\mathrm{UME}}^2_V(Q, R)$.
- Reject H_0 if $\hat{S}_n$ is too large.

Optimize V by maximizing power of relative UME test.
V shows where Q is better than P.

Where does each GAN do better?



Q = LSGAN [Mao et al., 2017]

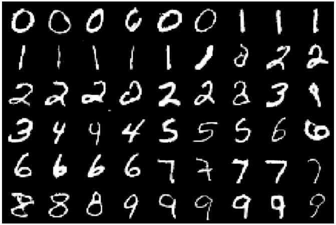


P = GAN

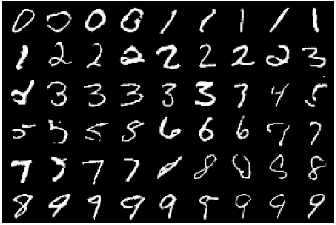
[Goodfellow et al., 2014]

- R = real MNIST images.
- Set V = 40 (real) images of digit $i = 0, \ldots, 9$.
- Q is better at "1" and "5". P is slightly better at "3". **Interpretable**.

Where does each GAN do better?



Q = LSGAN [Mao et al., 2017]

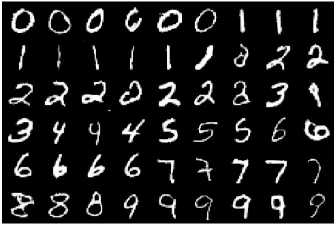


P = GAN

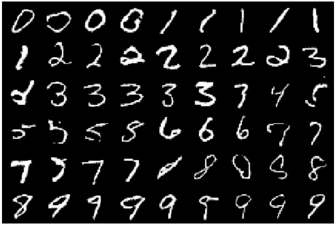
[Goodfellow et al., 2014]

- R = real MNIST images.
- Set V = 40 (real) images of digit $i = 0, \ldots, 9$.
- Q is better at "1" and "5". P is slightly better at "3". Interpretable.

Where does each GAN do better?

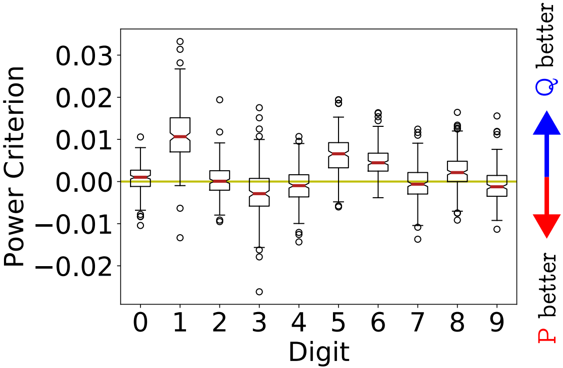


Q = LSGAN [Mao et al., 2017]



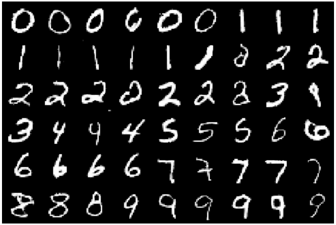
P = GAN

[Goodfellow et al., 2014]

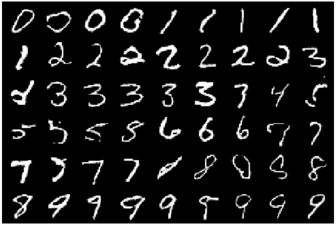


- R = real MNIST images.
- Set V = 40 (real) images of digit $i = 0, \ldots, 9$.
- Q is better at “1” and “5”. P is slightly better at “3”. Interpretable.

Where does each GAN do better?

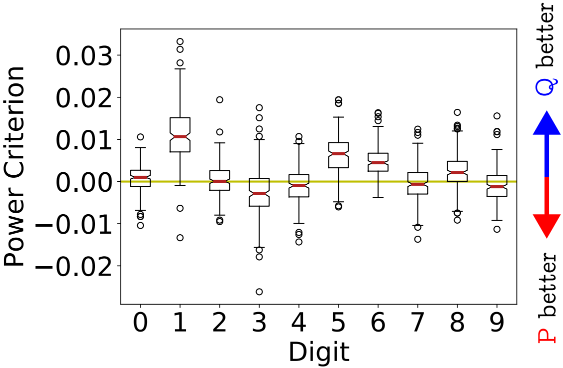


Q = LSGAN [Mao et al., 2017]



P = GAN

[Goodfellow et al., 2014]



- R = real MNIST images.
- Set V = 40 (real) images of digit $i = 0, \ldots, 9$.
- Q is better at "1" and "5". P is slightly better at "3". **Interpretable**.

(k = Gaussian kernel on top of features from a CNN classifier.)

Testing Explicit Models with Kernel Stein Discrepancy

Goodness-of-fit Testing

$\boxed{\text{Two-sample testing (so far)}}$

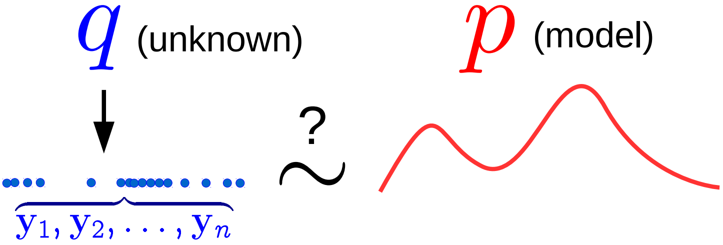
q (unknown) $\qquad$ p (unknown)

$$\underbrace{\mathbf{y}_1, \mathbf{y}_2, \ldots, \mathbf{y}_n} \overset{?}{\sim} \underbrace{\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n}$$

Test goal: Do data follow the model p?

Goodness-of-fit Testing

$\boxed{\text{Goodness-of-fit testing}}$



Test goal: Do data follow the model p?

- p is an explicit density function known up to the normalizer e.g., a restricted Boltzmann machine.
- **Important**: no sample from p.

Recall the MMD (part 1)

Integral probability metric form of MMD:

$$\mathrm{MMD}(p, q; \mathcal{F}) = \sup_{\|f\|_{\mathcal{F}} \leq 1} [\mathbf{E}_q f - \mathbf{E}_p f],$$

where $\mathcal{F}$ = RKHS defined by a kernel k.

Can we compute MMD with samples and a density p?

- **Problem 1:** usually can't compute $\mathbf{E}_p f$ in closed form.
- **Problem 2:** cannot sample from p. Also statistically inefficient.

Recall the MMD (part 1)

Integral probability metric form of MMD:

$$\mathrm{MMD}(p, q; \mathcal{F}) = \sup_{\|f\|_{\mathcal{F}} \leq 1} [\mathbf{E}_q f - \mathbf{E}_p f],$$

where $\mathcal{F}$ = RKHS defined by a kernel k.

Can we compute MMD with samples and a **density** p?

- **Problem 1:** usually can't compute $\mathbf{E}_p f$ in closed form.
- **Problem 2:** cannot sample from p. Also statistically inefficient.

Recall the MMD (part 1)

Integral probability metric form of MMD:

$$\mathrm{MMD}(p, q; \mathcal{F}) = \sup_{\|f\|_{\mathcal{F}} \leq 1} [\mathbf{E}_q f - \mathbf{E}_p f],$$

where $\mathcal{F}$ = RKHS defined by a kernel k.

Can we compute MMD with samples and a **density** p?

- **Problem 1:** usually can't compute $\mathbf{E}_p f$ in closed form.
- **Problem 2:** cannot sample from p. Also statistically inefficient.

Stein Idea

To get rid of $\mathbf{E}_p f$ in

$$\sup_{\|f\|_{\mathcal{F}}\leq 1} [\mathbf{E}_q \quad f - \mathbf{E}_p \quad f],$$

we define the (1-D) **Stein operator**

$$[T_p f](x) = \frac{1}{p(x)} \frac{d}{dx} (f(x) p(x)).$$

Then, $\mathbf{E}_p T_p f = 0$ subject to appropriate boundary conditions.

Stein Idea

To get rid of $\mathbf{E}_p f$ in

$$\sup_{\|f\|_{\mathcal{F}} \leq 1} [\mathbf{E}_q T_p f - \mathbf{E}_p T_p f],$$

we define the (1-D) **Stein operator**

$$[T_p f](x) = \frac{1}{p(x)} \frac{d}{dx} (f(x) p(x)) .$$

Then, $\mathbf{E}_p T_p f = 0$ subject to appropriate boundary conditions.

Stein Idea

To get rid of $\mathbf{E}_p f$ in

$$\sup_{\|f\|_{\mathcal{F}} \leq 1} [\mathbf{E}_q T_p f - \mathbf{E}_p T_p f],$$

we define the (1-D) **Stein operator**

$$[T_p f](x) = \frac{1}{p(x)} \frac{d}{dx} (f(x) p(x)).$$

Then, $\mathbf{E}_p T_p f = 0$ subject to appropriate boundary conditions.

Stein Idea

To get rid of $\mathbf{E}_p f$ in

$$\sup_{\|f\|_{\mathcal{F}} \leq 1} [\mathbf{E}_q T_p f - \mathbf{E}_p T_p f],$$

we define the (1-D) **Stein operator**

$$[T_p f](x) = \frac{1}{p(x)} \frac{d}{dx} (f(x) p(x)) .$$

Then, $\mathbf{E}_p T_p f = 0$ subject to appropriate boundary conditions.

Stein Idea

To get rid of $\mathbf{E}_p f$ in

$$\sup_{\|f\|_{\mathcal{F}} \leq 1} [\mathbf{E}_q T_p f - \mathbf{E}_p T_p f],$$

we define the (1-D) **Stein operator**

$$[T_p f](x) = \frac{1}{p(x)} \frac{d}{dx} (f(x) p(x)) .$$

Then, $\mathbf{E}_p T_p f = 0$ subject to appropriate boundary conditions.

Proof [Gorham and Mackey (NeurIPS 15), Oates, Girolami, Chopin (JRSS B 2016)].

$$\mathbf{E}_p [T_p f] = \int \left[\frac{1}{p(x)} \frac{d}{dx} (f(x) p(x)) \right] p(x) dx$$

$$= \int \left[\frac{d}{dx} (f(x) p(x)) \right] dx$$

$$= [f(x) p(x)]_{-\infty}^{\infty} = 0$$

Stein Idea

To get rid of $\mathbf{E}_p f$ in

$$\sup_{\|f\|_{\mathcal{F}} \le 1} [\mathbf{E}_q T_p f - \mathbf{E}_p T_p f],$$

we define the (1-D) **Stein operator**

$$[T_p f](x) = \frac{1}{p(x)} \frac{d}{dx} (f(x) p(x)) \,.$$

Then, $\mathbf{E}_p T_p f = 0$ subject to appropriate boundary conditions.

Proof [Gorham and Mackey (NeurIPS 15), Oates, Girolami, Chopin (JRSS B 2016)].

$$\mathbf{E}_p [T_p f] = \int \left[\frac{1}{\cancel{p(x)}} \frac{d}{dx} (f(x) p(x)) \right] \cancel{p(x)} dx$$

$$= \int \left[\frac{d}{dx} (f(x) p(x)) \right] dx$$

$$= [f(x) p(x)]_{-\infty}^{\infty} = 0$$

Stein Idea

To get rid of $\mathbf{E}_p f$ in

$$\sup_{\|f\|_{\mathcal{F}} \leq 1} [\mathbf{E}_q T_p f - \mathbf{E}_p T_p f],$$

we define the (1-D) **Stein operator**

$$[T_p f](x) = \frac{1}{p(x)} \frac{d}{dx} (f(x)p(x)).$$

Then, $\mathbf{E}_p T_p f = 0$ subject to appropriate boundary conditions.

Proof [Gorham and Mackey (NeurIPS 15), Oates, Girolami, Chopin (JRSS B 2016)].

$$\begin{aligned}
\mathbf{E}_p [T_p f] &= \int \left[\frac{1}{\cancel{p(x)}} \frac{d}{dx} (f(x)p(x)) \right] \cancel{p(x)} dx \\
&= \int \left[\frac{d}{dx} (f(x)p(x)) \right] dx \\
&= [f(x)p(x)]_{-\infty}^{\infty} = 0
\end{aligned}$$

Stein Idea

To get rid of $\mathbf{E}_p f$ in

$$\sup_{\|f\|_{\mathcal{F}} \le 1} [\mathbf{E}_q T_p f - \mathbf{E}_p T_p f],$$

we define the (1-D) **Stein operator**

$$[T_p f](x) = \frac{1}{p(x)} \frac{d}{dx} (f(x) p(x)) .$$

Then, $\mathbf{E}_p T_p f = 0$ subject to appropriate boundary conditions.

Proof [Gorham and Mackey (NeurIPS 15), Oates, Girolami, Chopin (JRSS B 2016)].

$$\begin{aligned}
\mathbf{E}_p [T_p f] &= \int \left[\frac{1}{\cancel{p(x)}} \frac{d}{dx} (f(x) p(x)) \right] \cancel{p(x)} dx \\
&= \int \left[\frac{d}{dx} (f(x) p(x)) \right] dx \\
&= [f(x) p(x)]_{-\infty}^{\infty} = 0
\end{aligned}$$

$\square$

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

- Stein operator: $T_p f = \frac{1}{p(x)} \frac{d}{dx} \left(f(x) p(x) \right).$

Kernel Stein Discrepancy (KSD)

$$\mathrm{KSD}_p(q) = \sup_{\|f\|_{\mathcal{F}} \leq 1} \mathbb{E}_q T_p f - \mathbb{E}_p T_p f$$

where

$$g(v) := \mathbb{E}_{\mathrm{x} \sim q} \left[\frac{1}{p(\mathrm{x})} \frac{d}{d\mathrm{x}} [k(\mathrm{x}, v) p(\mathrm{x})] \right].$$

- Known as the Stein witness function. (This will come back later!)

Chwialkowski, Strathmann, G., (ICML 2016)
Liu, Lee, Jordan (ICML 2016)

▸ full derivation

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

- Stein operator: $T_p f = \frac{1}{p(x)} \frac{d}{dx} \left(f(x) p(x) \right).$

Kernel Stein Discrepancy (KSD)

$$\mathrm{KSD}_p(q) = \sup_{\|f\|_{\mathcal{F}} \leq 1} \mathbb{E}_q T_p f - \cancel{\mathbb{E}_p T_p f}$$

where

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

- Known as the Stein witness function. (This will come back later!)

Chwialkowski, Strathmann, G., (ICML 2016)
Liu, Lee, Jordan (ICML 2016)

▸ full derivation

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

- Stein operator: $T_p f = \frac{1}{p(x)} \frac{d}{dx} \left(f(x) p(x) \right).$

Kernel Stein Discrepancy (KSD)

$$\begin{aligned} \mathrm{KSD}_p(q) &= \sup_{\|f\|_{\mathcal{F}} \le 1} \mathbb{E}_q T_p f - \cancel{\mathbb{E}_p T_p f} \\ &= \sup_{\|f\|_{\mathcal{F}} \le 1} \mathbb{E}_q T_p f \end{aligned}$$

where

$$g(v) := \mathbb{E}_{\mathrm{x} \sim q} \left[\frac{1}{p(\mathrm{x})} \frac{d}{d\mathrm{x}} [k(\mathrm{x}, v) p(\mathrm{x})] \right].$$

- Known as the Stein witness function. (This will come back later!)

Chwialkowski, Strathmann, G., (ICML 2016)
Liu, Lee, Jordan (ICML 2016)

full derivation

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

- Stein operator: $T_p f = \frac{1}{p(x)} \frac{d}{dx} \left(f(x) p(x)\right).$

Kernel Stein Discrepancy (KSD)

$$
\begin{aligned}
\mathrm{KSD}_p(q) &= \sup_{\|f\|_{\mathcal{F}} \leq 1} \mathbb{E}_q T_p f - \cancel{\mathbb{E}_p T_p f} \\
&= \sup_{\|f\|_{\mathcal{F}} \leq 1} \mathbb{E}_q T_p f \\
\text{(closed-form sup)} &= \|g\|_{\mathcal{F}},
\end{aligned}
$$

where

$$
g(v) := \mathbb{E}_{\mathrm{x} \sim q} \left[\frac{1}{p(\mathrm{x})} \frac{d}{d\mathrm{x}} [k(\mathrm{x}, v) p(\mathrm{x})] \right].
$$

- Known as the Stein witness function. (This will come back later!)

Chwialkowski, Strathmann, G., (ICML 2016)
Liu, Lee, Jordan (ICML 2016)

full derivation

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

- Stein operator: $T_p f = \frac{1}{p(x)} \frac{d}{dx} \left(f(x) p(x) \right).$

Kernel Stein Discrepancy (KSD)

$$
\begin{aligned}
\mathrm{KSD}_p(q) &= \sup_{\|f\|_{\mathcal{F}} \leq 1} \mathbb{E}_q T_p f - \cancel{\mathbb{E}_p T_p f} \\
&= \sup_{\|f\|_{\mathcal{F}} \leq 1} \mathbb{E}_q T_p f \\
\text{(closed-form sup)} &= \|g\|_{\mathcal{F}},
\end{aligned}
$$

where

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

- Known as the **Stein witness function**. (This will come back later!)

Chwialkowski, Strathmann, G., (ICML 2016)
Liu, Lee, Jordan (ICML 2016)

▸ full derivation

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

- Stein operator: $T_p f = \frac{1}{p(x)} \frac{d}{dx} \left(f(x) p(x) \right)$. (normalizer cancels)

Kernel Stein Discrepancy (KSD)

$$\begin{aligned}
\mathrm{KSD}_p(q) &= \sup_{\|f\|_{\mathcal{F}} \leq 1} \mathbb{E}_q T_p f - \cancel{\mathbb{E}_p T_p f} \\
&= \sup_{\|f\|_{\mathcal{F}} \leq 1} \mathbb{E}_q T_p f \\
\text{(closed-form sup)} &= \|g\|_{\mathcal{F}},
\end{aligned}$$

where

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

- Known as the **Stein witness function**. (This will come back later!)

Chwialkowski, Strathmann, G., (ICML 2016)
Liu, Lee, Jordan (ICML 2016)

▸ full derivation

Kernel Stein Discrepancy: population expression

Test statistic when $x \in \mathbb{R}^d$, given *independent* $y, y' \sim q$,

$$\mathrm{KSD}_p^2(q) = \|g\|^2_{\mathcal{F}^d} = \mathbb{E}_{y \sim q}\mathbb{E}_{y' \sim q} h_p(y, y'),$$

where

$$\begin{aligned} h_p(x, x') &= \mathbf{s}_p(x)^\top \mathbf{s}_p(x') k(x, x') \\ &\quad + \mathbf{s}_p(x)^\top \nabla_{x'} k(x, x') \\ &\quad + \mathbf{s}_p(x')^\top \nabla_x k(x, x') \\ &\quad + \mathrm{tr}\left[\nabla_x \nabla_{x'} k(x, x')\right] \end{aligned}$$

- $\mathbf{s}_p(x) \in \mathbb{R}^d = \nabla_x \log p(x)$ (score function of p)

Theorem (Chwialkowski et al. (ICML 2016))

Assume appropriate boundary conditions. If kernel is C_0-universal and Q satisfies $\mathbb{E}_{x\sim q}\left\|\nabla\left(\log\frac{p(x)}{q(x)}\right)\right\|^2 < \infty$, then $\mathrm{KSD}_p^2(q) = 0$ iff $p = q$.

Kernel Stein Discrepancy: population expression

Test statistic when $x \in \mathbb{R}^d$, given *independent* $y, y' \sim q$,

$$\mathrm{KSD}_p^2(q) = \|g\|_{\mathcal{F}^d}^2 = \mathbb{E}_{y \sim q} \mathbb{E}_{y' \sim q} h_p(y, y'),$$

where

$$\begin{aligned} h_p(x, x') &= \mathbf{s}_p(x)^\top \mathbf{s}_p(x') k(x, x') \\ &\quad + \mathbf{s}_p(x)^\top \nabla_{x'} k(x, x') \\ &\quad + \mathbf{s}_p(x')^\top \nabla_x k(x, x') \\ &\quad + \mathrm{tr}\left[\nabla_x \nabla_{x'} k(x, x')\right] \end{aligned}$$

- $\mathbf{s}_p(x) \in \mathbb{R}^d = \nabla_x \log p(x)$ (score function of p)

Theorem (Chwialkowski et al. (ICML 2016))

Assume appropriate boundary conditions. If kernel is C_0-universal and Q satisfies $\mathbb{E}_{x \sim q} \left\| \nabla \left(\log \frac{p(x)}{q(x)} \right) \right\|^2 < \infty$, then $\mathrm{KSD}_p^2(q) = 0$ iff $p = q$.

KSD: Empirical statistic and asymptotics

Given: $\{y_i\}_{i=1}^{n} \overset{i.i.d.}{\sim} q$, a differentiable density p.

- Empirical statistic:

$$\widehat{\mathrm{KSD}_p^2}(q) := \frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n} h_p(y_i, y_j).$$

- Runtime complexity: $\mathcal{O}(d^2 n^2)$.

Asymptotics:

1. When $p = q$, $\widehat{\mathrm{KSD}_p^2}(q) \xrightarrow{d}$ infinite weighted sum of chi-squared variables.
2. When $p \neq q$, $\widehat{\mathrm{KSD}_p^2}(q) \xrightarrow{d}$ a Gaussian.

Testing:

- Get test threshold via wild bootstrap.
- Permutation test not applicable. Have only one set of samples.

▸ wild bootstrap detail

KSD: Empirical statistic and asymptotics

Given: $\{y_i\}_{i=1}^{n} \overset{i.i.d.}{\sim} q$, a differentiable density p.

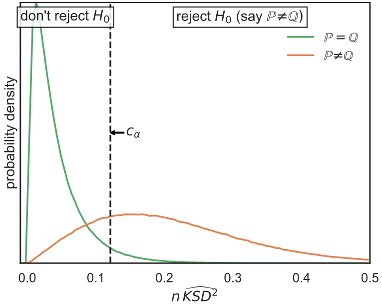
- Empirical statistic:

$$\widehat{\mathrm{KSD}}_p^2(q) := \frac{1}{n^2} \sum_{i=1}^{n} \sum_{j=1}^{n} h_p(y_i, y_j).$$

- Runtime complexity: $\mathcal{O}(d^2 n^2)$.

Asymptotics:

1. When $p = q$, $\widehat{\mathrm{KSD}}_p^2(q) \xrightarrow{d}$ infinite weighted sum of chi-squared variables.
2. When $p \neq q$, $\widehat{\mathrm{KSD}}_p^2(q) \xrightarrow{d}$ a Gaussian.



Testing:

- Get test threshold via wild bootstrap.
- Permutation test not applicable. Have only one set of samples.

wild bootstrap detail

KSD: Empirical statistic and asymptotics

Given: $\{y_i\}_{i=1}^n \overset{i.i.d.}{\sim} q$, a differentiable density p.

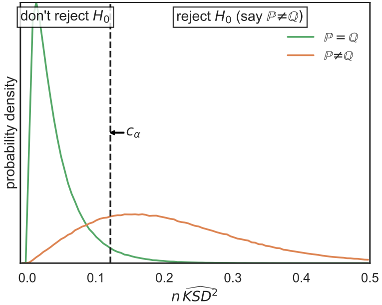
- Empirical statistic:

$$\widehat{\mathrm{KSD}}_p^2(q) := \frac{1}{n^2} \sum_{i=1}^n \sum_{j=1}^n h_p(y_i, y_j).$$

- Runtime complexity: $\mathcal{O}(d^2 n^2)$.

Asymptotics:

1. When $p = q$, $\widehat{\mathrm{KSD}}_p^2(q) \xrightarrow{d}$ infinite weighted sum of chi-squared variables.
2. When $p \neq q$, $\widehat{\mathrm{KSD}}_p^2(q) \xrightarrow{d}$ a Gaussian.



Testing:

- Get test threshold via wild bootstrap.
- Permutation test not applicable. Have only one set of samples.

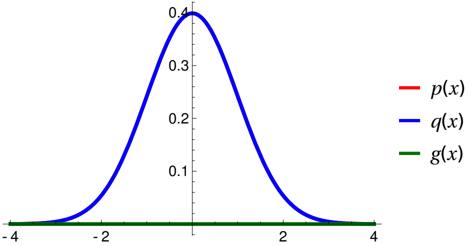
▸ wild bootstrap detail

Linear-time, interpretable Goodness-of-fit Test

The Finite Set Stein Discrepancy (FSSD) (Jitkrittum et al., 2017)

Recall Stein witness:

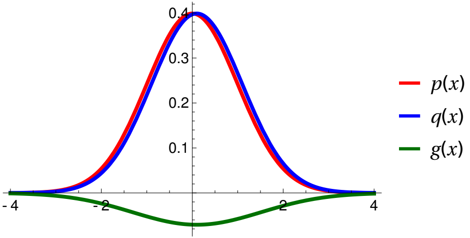
$$\mathbf{g}(\mathbf{v}) := \mathbb{E}_{\mathbf{x}\sim q}\left[\frac{1}{p(\mathbf{x})}\frac{d}{d\mathbf{x}}[k(\mathbf{x},\mathbf{v})p(\mathbf{x})]\right].$$



The Finite Set Stein Discrepancy (FSSD) (Jitkrittum et al., 2017)

Recall Stein witness:

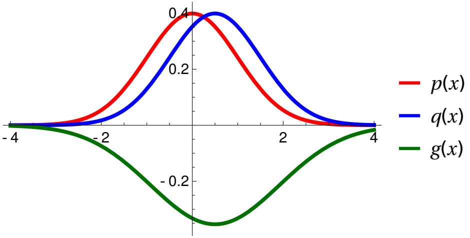
$$\mathbf{g}(\mathbf{v}) := \mathbb{E}_{\mathbf{x}\sim q}\left[\frac{1}{p(\mathbf{x})}\frac{d}{d\mathbf{x}}[k(\mathbf{x},\mathbf{v})p(\mathbf{x})]\right].$$



The Finite Set Stein Discrepancy (FSSD) (Jitkrittum et al., 2017)

Recall Stein witness:

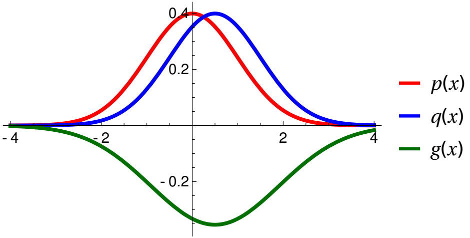
$$\mathbf{g}(\mathbf{v}) := \mathbb{E}_{\mathbf{x}\sim q}\left[\frac{1}{p(\mathbf{x})}\frac{d}{d\mathbf{x}}[k(\mathbf{x},\mathbf{v})p(\mathbf{x})]\right].$$



The Finite Set Stein Discrepancy (FSSD) (Jitkrittum et al., 2017)

Recall Stein witness:

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



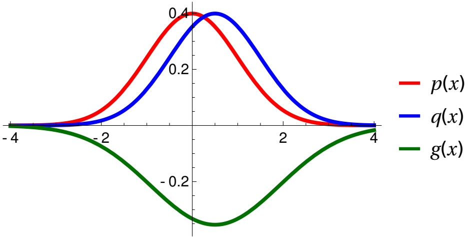
- **FSSD statistic**: Evaluate g^2 at J test locations $V = \{\mathbf{v}_1, \ldots, \mathbf{v}_J\}$.
- FSSD is to KSD as UME is to MMD.

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

The Finite Set Stein Discrepancy (FSSD) (Jitkrittum et al., 2017)

Recall Stein witness:

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



- **FSSD statistic**: Evaluate g^2 at J test locations $V = \{\mathbf{v}_1, \ldots, \mathbf{v}_J\}$.
- FSSD is to KSD as UME is to MMD.

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

Theorem (Jitkrittum et al., NeurIPS 2017)

Assume same conditions as KSD, and a real analytic kernel. Assume V drawn from a distribution that covers the domain. Then,

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

FSSD: Empirical statistic and asymptotics

- Estimate $\widehat{\mathrm{FSSD}^2}$ with samples $\{y_i\}_{i=1}^n \overset{i.i.d.}{\sim} q$.
- Runtime complexity: $\mathcal{O}(d^2 Jn)$. Linear in n.

Asymptotics:

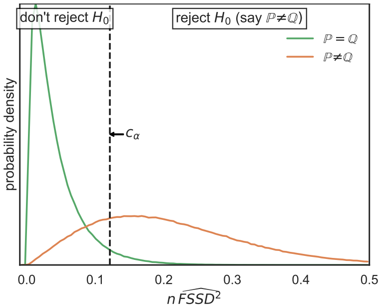
1. When $p = q$, $\widehat{\mathrm{FSSD}^2} \xrightarrow{d}$ finite weighted sum of chi-squared variables.
2. When $p \neq q$, $\widehat{\mathrm{FSSD}^2} \xrightarrow{d}$ a Gaussian.

Testing:

- Weights = eigenvalues of a $dJ \times dJ$ covariance matrix.
- Test threshold = empirical $(1 - \alpha)$-quantile.

FSDD: Empirical statistic and asymptotics

- Estimate $\widehat{\mathrm{FSSD}^2}$ with samples $\{y_i\}_{i=1}^n \overset{i.i.d.}{\sim} q$.
- Runtime complexity: $\mathcal{O}(d^2 Jn)$. Linear in n.



Asymptotics:

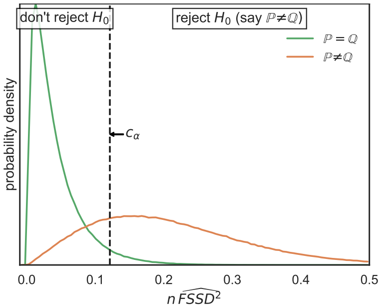
1. When $p = q$, $\widehat{\mathrm{FSSD}^2} \xrightarrow{d}$ finite weighted sum of chi-squared variables.
2. When $p \neq q$, $\widehat{\mathrm{FSSD}^2} \xrightarrow{d}$ a Gaussian.

Testing:

- Weights = eigenvalues of a $dJ \times dJ$ covariance matrix.
- Test threshold = empirical $(1-\alpha)$-quantile.

FSSD: Empirical statistic and asymptotics

- Estimate $\widehat{\mathrm{FSSD}^2}$ with samples $\{y_i\}_{i=1}^n \overset{i.i.d.}{\sim} q$.
- Runtime complexity: $\mathcal{O}(d^2 Jn)$. Linear in n.



Asymptotics:

1. When $p = q$, $\widehat{\mathrm{FSSD}^2} \xrightarrow{d}$ finite weighted sum of chi-squared variables.
2. When $p \neq q$, $\widehat{\mathrm{FSSD}^2} \xrightarrow{d}$ a Gaussian.

Testing:

- Weights = eigenvalues of a $dJ \times dJ$ covariance matrix.
- Test threshold = empirical $(1-\alpha)$-quantile.

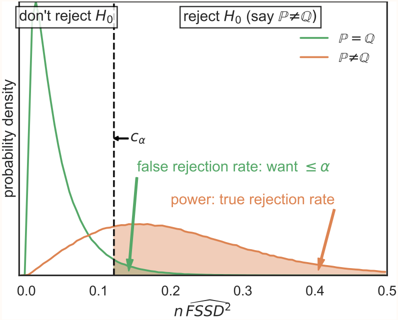
Find test locations by maximizing power

Proposition (Asymptotic power of FSSD^2 [Jitkrittum et al., 2017])

For large n, the test power

$$\mathbb{P}(\textit{reject } H_0 \mid H_1 \textit{ true}) \approx \Phi\left(\sqrt{n}\frac{\mathrm{FSSD}^2}{\sigma_{H_1}} - \frac{c_\alpha}{\sqrt{n}\sigma_{H_1}}\right),$$

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



- For large n, 1^{st} term $\sqrt{n}\frac{\mathrm{FSSD}^2}{\sigma_{H_1}}$ dominates. Similar to MMD.

$$\text{(maximize test power)} \quad \arg\max_V \text{power} \approx \arg\max_V \frac{\widehat{\mathrm{FSSD}^2}}{\widehat{\sigma_{H_1}}}$$

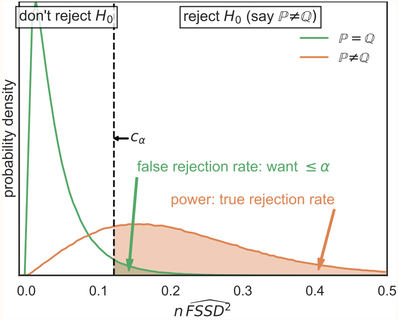
Find test locations by maximizing power

Proposition (Asymptotic power of FSSD^2 [Jitkrittum et al., 2017])

For large n, the test power

$$\mathbb{P}(\textit{reject } H_0 \mid H_1 \textit{ true}) \approx \Phi\left(\sqrt{n}\frac{\mathrm{FSSD}^2}{\sigma_{H_1}} - \frac{c_\alpha}{\sqrt{n}\sigma_{H_1}}\right),$$

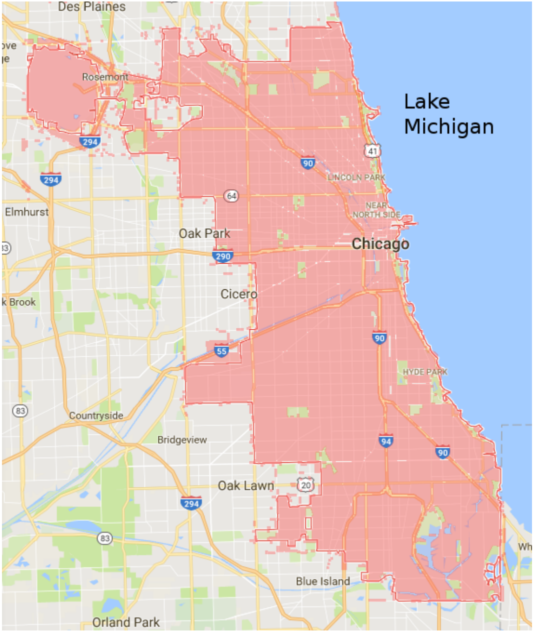
where Φ = CDF of $\mathcal{N}(0,1)$.



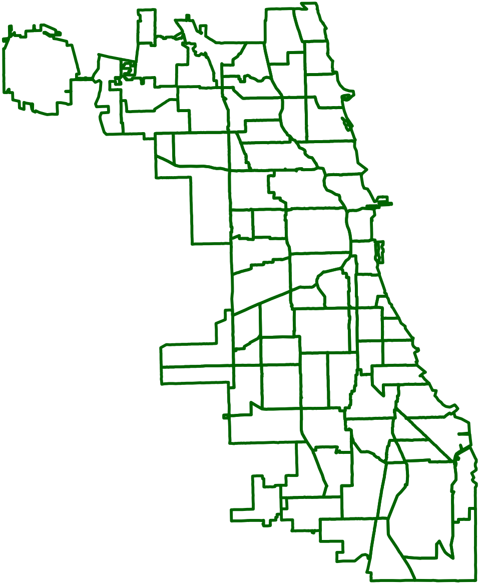
- For large n, 1^{st} term $\sqrt{n}\frac{\mathrm{FSSD}^2}{\sigma_{H_1}}$ dominates. Similar to MMD.

$$\text{(maximize test power)} \quad \arg\max_V \text{power} \approx \arg\max_V \frac{\widehat{\mathrm{FSSD}^2}}{\widehat{\sigma_{H_1}}}$$

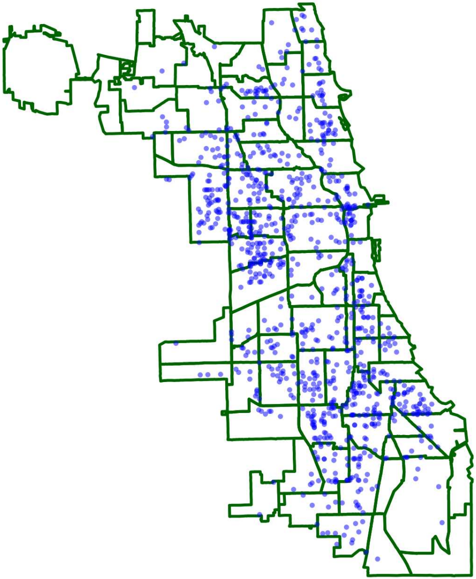
Interpretable Test Locations: Chicago Crime



Interpretable Test Locations: Chicago Crime

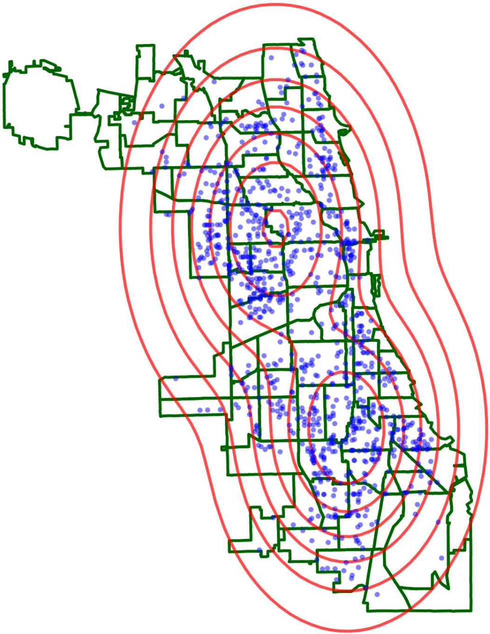


Interpretable Test Locations: Chicago Crime



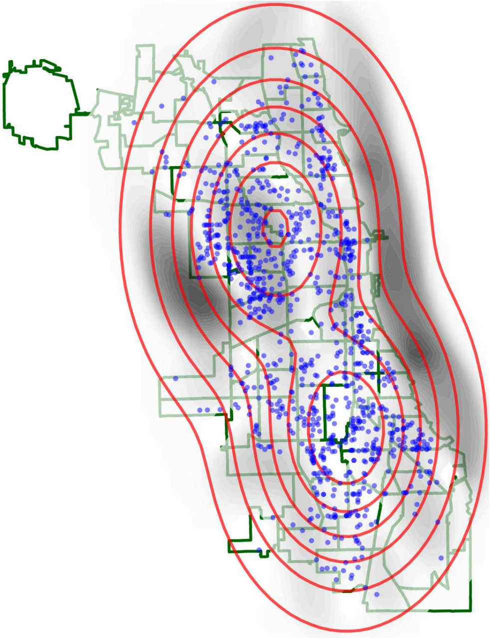
- $n = 11957$ robbery events in Chicago in 2016.
  - lat/long coordinates $=$ sample from q.
- Model spatial density with Gaussian mixtures.

Interpretable Test Locations: Chicago Crime



Model p = 2-component Gaussian mixture.

Interpretable Test Locations: Chicago Crime



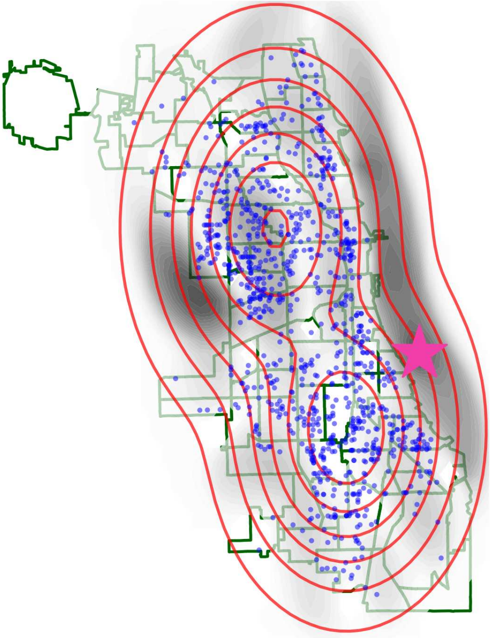
Score surface

$$\color{green}{\text{score}}(\mathbf{v}) := \frac{\widehat{\mathrm{FSD}^2}}{\widehat{\sigma_{H_1}}}$$

(power criterion)

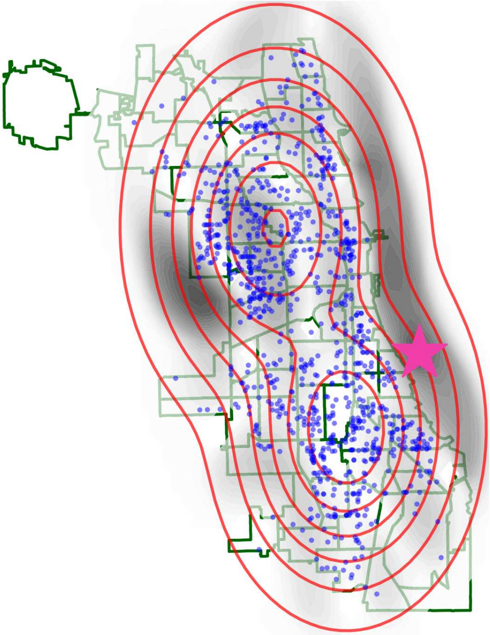
- Dark = high mismatch between p and q.

Interpretable Test Locations: Chicago Crime



★ = optimized $\mathbf{v}$.

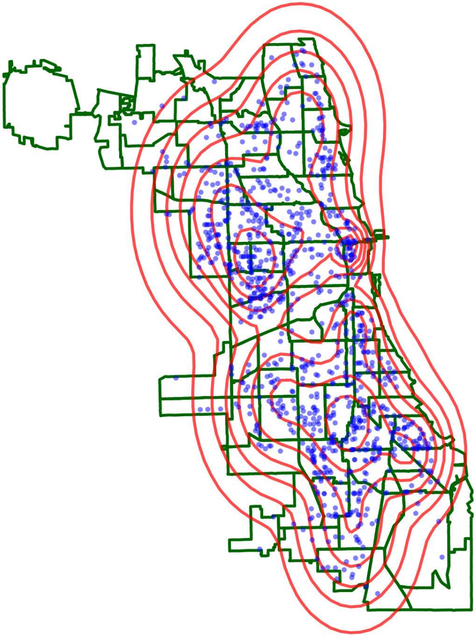
Interpretable Test Locations: Chicago Crime



★ = optimized $\mathbf{v}$.

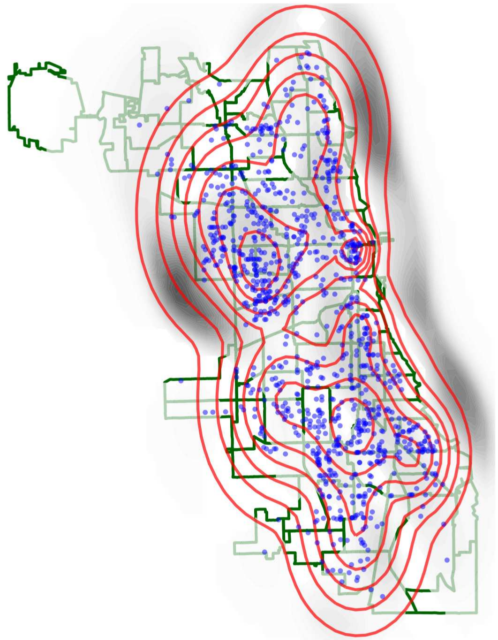
No robbery in Lake Michigan.

Interpretable Test Locations: Chicago Crime



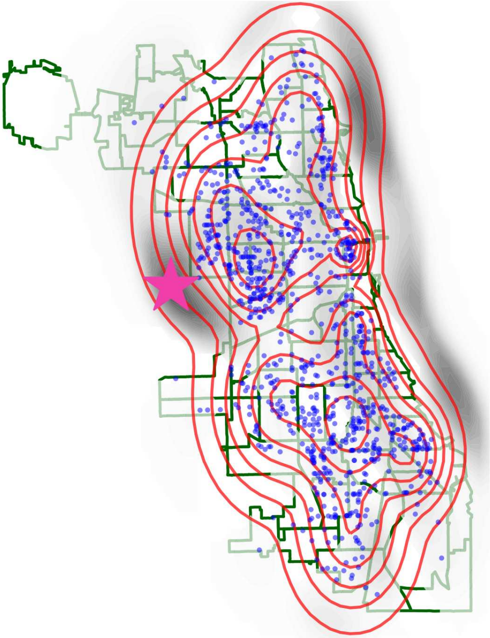
Model p = 10-component Gaussian mixture.

Interpretable Test Locations: Chicago Crime



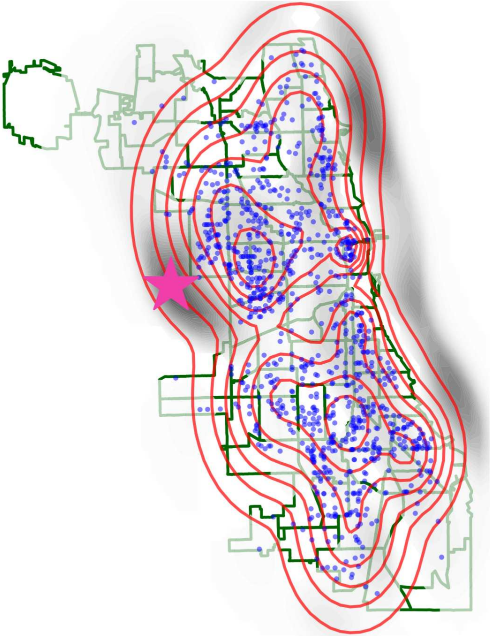
Capture the right tail better.

Interpretable Test Locations: Chicago Crime



Still, does not capture the left tail.

Interpretable Test Locations: Chicago Crime



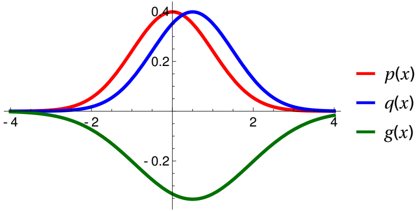
Still, does not capture the left tail.

Learned test locations are interpretable.

KSD vs. FSSD

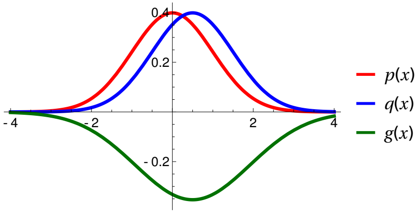
- Recall Stein witness:

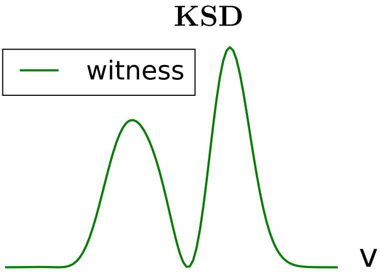
$$\mathbf{g}(\mathbf{v}) := \mathbb{E}_{\mathbf{x}\sim q}\left[\frac{1}{p(\mathbf{x})}\frac{d}{d\mathbf{x}}[k_{\mathbf{v}}(\mathbf{x})p(\mathbf{x})]\right].$$



KSD vs. FSSD

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



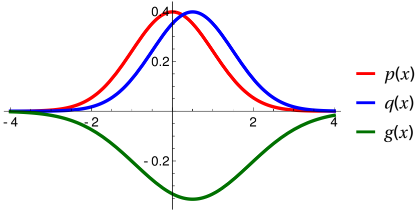


$\text{KSD}^2 = \|\mathbf{g}\|^2_{\text{RKHS}}$ (RKHS norm).

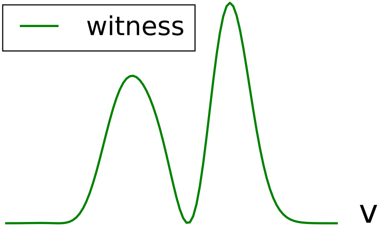
Good when the difference between p, q is spatially diffuse.

KSD vs. FSSD

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



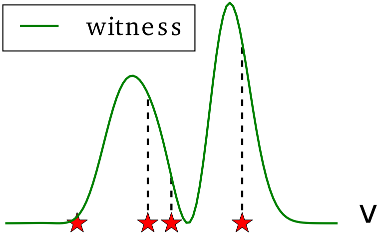
KSD



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

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

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.

Conclusion

- **Part 1: Divergence measures**
  - Integral probability metrics
  - ϕ-divergences (f-divergences)

- **Part 2: Statistical hypothesis testing**
  - Using integral probability metrics (MMD)
  - Relation of testing and classification
  - Learned features for powerful tests

- **Part 3: Linear-time features and model criticism**
  - Interpretable, linear time features for testing (UME)
  - Stein's method for model evaluation (KSD)

References and further reading

- UME/NME
  - Chwialkowski et al., NeurIPS 2015. NME with random locations.
  - Jitkrittum et al., NeurIPS 2016. NME with optimized locations.
  - Scetbon and Varoquaux, NeurIPS 2019. Extension of UME/NME with L1 norm.
- Kernel Stein Discrepancy
  - Chwialkowski et al., ICML 2016 and Liu et al., ICML 2016. KSD testing.
  - Oates et al., RSS 2016 and Gorham et al., NeurIPS 2015. MCMC convergence check.
  - Liu and Wang, NeurIPS 2016. Stein variational gradient descent.
  - Barp et al., NeurIPS 2019. For model fitting.
- FSSD. Jitkrittum et al., NeurIPS 2017 (best paper).
- Relative tests
  - Bounliphone et al., ICLR 2016. Relative MMD. For 2 models.
  - Jitkrittum et al., NeurIPS 2018. Relative UME, FSSD. For 2 models
  - Lim et al., NeurIPS 2019. Relative KSD, MMD. For > 2 models.

Questions?

Thank you

Appendix

Outline

1. Appendix: UME, NME
2. Appendix: Relative UME
3. Appendix: Kernel Stein Discrepancy
4. Appendix: FSSD

Normalized ME (NME) Statistic (Chwialkowski et al., 2015, Jitkrittum et al., 2016)

- Let $\psi(\mathbf{x}) := \frac{1}{\sqrt{J}}\left(k(\mathbf{x}, \mathbf{v}_1), \ldots, k(\mathbf{x}, \mathbf{v}_J)\right)^\top \in \mathbb{R}^J$. Equivalently,

$$\mathrm{UME}^2(P, Q) = \|\mathbb{E}_{\mathbf{x}\sim P}\psi(\mathbf{x}) - \mathbb{E}_{\mathbf{y}\sim Q}\psi(\mathbf{y})\|_2^2.$$

- Covariance matrix $\mathbf{C} := \mathrm{cov}_{\mathbf{x}\sim P}[\psi_V(\mathbf{x})] + \mathrm{cov}_{\mathbf{y}\sim Q}[\psi_V(\mathbf{y})] \in \mathbb{R}^{J\times J}$.

$$\mathrm{NME}^2(P, Q) = \left[\mathbb{E}_{\mathbf{x}\sim P}\psi_V(\mathbf{x}) - \mathbb{E}_{\mathbf{y}\sim Q}\psi_V(\mathbf{y})\right]^\top \mathbf{C}^{-1}\left[\mathbb{E}_{\mathbf{x}\sim P}\psi_V(\mathbf{x}) - \mathbb{E}_{\mathbf{y}\sim Q}\psi_V(\mathbf{y})\right]$$

- $\mathbf{S}^{-1}$ decorrelates the J terms. Simpler null distribution.
- $\Longrightarrow$ **Normalized ME (NME) statistic.**

Normalized ME (NME) Statistic (Chwialkowski et al., 2015, Jitkrittum et al., 2016)

- Let $\psi(\mathbf{x}) := \frac{1}{\sqrt{J}}\left(k(\mathbf{x}, \mathbf{v}_1), \ldots, k(\mathbf{x}, \mathbf{v}_J)\right)^\top \in \mathbb{R}^J$. Equivalently,

$$\mathrm{UME}^2(P, Q) = \|\mathbb{E}_{\mathbf{x}\sim P}\psi(\mathbf{x}) - \mathbb{E}_{\mathbf{y}\sim Q}\psi(\mathbf{y})\|_2^2.$$

- Covariance matrix $\mathbf{C} := \mathrm{cov}_{\mathbf{x}\sim P}[\psi_V(\mathbf{x})] + \mathrm{cov}_{\mathbf{y}\sim Q}[\psi_V(\mathbf{y})] \in \mathbb{R}^{J\times J}$.

$$\mathrm{NME}^2(P, Q) = \left[\mathbb{E}_{\mathbf{x}\sim P}\psi_V(\mathbf{x}) - \mathbb{E}_{\mathbf{y}\sim Q}\psi_V(\mathbf{y})\right]^\top \mathbf{C}^{-1}\left[\mathbb{E}_{\mathbf{x}\sim P}\psi_V(\mathbf{x}) - \mathbb{E}_{\mathbf{y}\sim Q}\psi_V(\mathbf{y})\right]$$

- $\mathbf{S}^{-1}$ decorrelates the J terms. Simpler null distribution.
- $\Longrightarrow$ **Normalized ME (NME) statistic.**

Normalized ME (NME) Statistic (Chwialkowski et al., 2015, Jitkrittum et al., 2016)

- Let $\psi(\mathbf{x}) := \frac{1}{\sqrt{J}}\left(k(\mathbf{x},\mathbf{v}_1),\ldots,k(\mathbf{x},\mathbf{v}_J)\right)^\top \in \mathbb{R}^J$. Equivalently,

$$\mathrm{UME}^2(P,Q) = \|\mathbb{E}_{\mathbf{x}\sim P}\psi(\mathbf{x}) - \mathbb{E}_{\mathbf{y}\sim Q}\psi(\mathbf{y})\|_2^2.$$

- Covariance matrix $\mathbf{C} := \mathrm{cov}_{\mathbf{x}\sim P}[\psi_V(\mathbf{x})] + \mathrm{cov}_{\mathbf{y}\sim Q}[\psi_V(\mathbf{y})] \in \mathbb{R}^{J\times J}$.

$$\boxed{\mathrm{NME}^2(P,Q) = \left[\mathbb{E}_{\mathbf{x}\sim P}\psi_V(\mathbf{x}) - \mathbb{E}_{\mathbf{y}\sim Q}\psi_V(\mathbf{y})\right]^\top \mathbf{C}^{-1}\left[\mathbb{E}_{\mathbf{x}\sim P}\psi_V(\mathbf{x}) - \mathbb{E}_{\mathbf{y}\sim Q}\psi_V(\mathbf{y})\right]}$$

- $\mathbf{S}^{-1}$ decorrelates the J terms. Simpler null distribution.
- $\Longrightarrow$ **Normalized ME (NME) statistic.**

Normalized ME (NME) Statistic (Chwialkowski et al., 2015, Jitkrittum et al., 2016)

- Let $\psi(\mathbf{x}) := \frac{1}{\sqrt{J}}\left(k(\mathbf{x}, \mathbf{v}_1), \ldots, k(\mathbf{x}, \mathbf{v}_J)\right)^\top \in \mathbb{R}^J$. Equivalently,

$$\mathrm{UME}^2(P, Q) = \|\mathbb{E}_{\mathbf{x}\sim P}\psi(\mathbf{x}) - \mathbb{E}_{\mathbf{y}\sim Q}\psi(\mathbf{y})\|_2^2.$$

- Covariance matrix $\mathbf{C} := \mathrm{cov}_{\mathbf{x}\sim P}[\psi_V(\mathbf{x})] + \mathrm{cov}_{\mathbf{y}\sim Q}[\psi_V(\mathbf{y})] \in \mathbb{R}^{J\times J}$.

$$\boxed{\mathrm{NME}^2(P, Q) = \left[\mathbb{E}_{\mathbf{x}\sim P}\psi_V(\mathbf{x}) - \mathbb{E}_{\mathbf{y}\sim Q}\psi_V(\mathbf{y})\right]^\top \mathbf{C}^{-1}\left[\mathbb{E}_{\mathbf{x}\sim P}\psi_V(\mathbf{x}) - \mathbb{E}_{\mathbf{y}\sim Q}\psi_V(\mathbf{y})\right]}$$

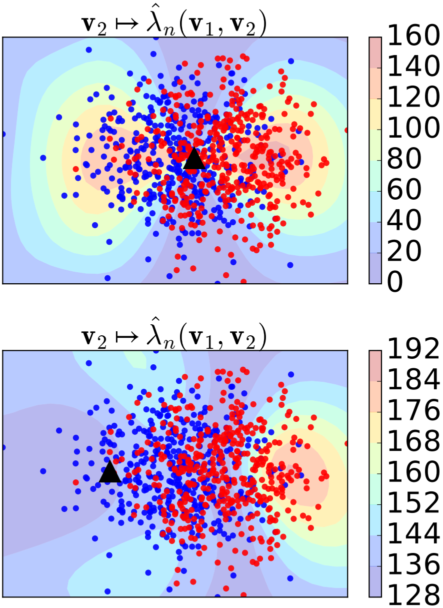
- $\mathbf{S}^{-1}$ decorrelates the J terms. Simpler null distribution.
- $\implies$ **Normalized ME (NME) statistic.**

Illustration of NME: Two Informative Features

- 2D problem.

$$P : \mathcal{N}([0,0], I)$$
$$Q : \mathcal{N}([1,0], I)$$

- $J = 2$ features.
- Fix $\mathbf{v}_1$ to ▲.
- Contour plot of $\mathbf{v}_2 \mapsto \hat{\lambda}_n(\{\mathbf{v}_1, \mathbf{v}_2\})$.
- $\{\mathbf{v}_1, \mathbf{v}_2\}$ chosen to reveal the difference of P and Q.



Full NME Test Statistic. $J = 1$

- Let $\mathcal{V} = \{\mathbf{v}_1, \ldots, \mathbf{v}_J\}$ be the J test locations.
- Let $\overline{\mathbf{z}}_n := \begin{pmatrix} \hat{\mu}_P(\mathbf{v}_1) - \hat{\mu}_Q(\mathbf{v}_1) \\ \vdots \\ \hat{\mu}_P(\mathbf{v}_J) - \hat{\mu}_Q(\mathbf{v}_J) \end{pmatrix} \in \mathbb{R}^J$.
- Let $(\mathbf{S}_n)_{ij} := \widehat{\mathrm{cov}}_{\mathbf{x}}[k(\mathbf{x}, \mathbf{v}_i), k(\mathbf{x}, \mathbf{v}_j)] + \widehat{\mathrm{cov}}_{\mathbf{y}}[k(\mathbf{y}, \mathbf{v}_i), k(\mathbf{y}, \mathbf{v}_j)] \in \mathbb{R}^{J \times J}$.
- Then, the statistic

$$\hat{\lambda}_n := n \overline{\mathbf{z}}_n^\top \left(\mathbf{S}_n + \gamma_n I\right)^{-1} \overline{\mathbf{z}}_n,$$

  where $\gamma_n > 0$ is a regularization parameter.
- When $J = 1$,

$$\hat{\lambda}_n = n \frac{[\hat{\mu}_P(\mathbf{v}) - \hat{\mu}_Q(\mathbf{v})]^2}{\gamma_\mathrm{n} + \mathrm{var}_{\mathbf{x}}[k(\mathbf{x}, \mathbf{v})] + \mathrm{var}_{\mathbf{y}}[k(\mathbf{y}, \mathbf{v})]}.$$

- Computing $\hat{\lambda}_n$: $\mathcal{O}(J^3 + J^2 n + Jdn)$.
- Optimization of $\mathcal{V}$: $\mathcal{O}(J^3 + J^2 dn)$.

Illustration: NME Statistic. $J = 1$

- Variance of $\mathbf{v}$ = variance of $\mathbf{v}$ from X + variance of $\mathbf{v}$ from Y.
- ME Statistic: $\hat{\lambda}_n(\mathbf{v}) := n \frac{\text{witness}^2(\mathbf{v})}{\text{variance of } \mathbf{v}}$.

Illustration: NME Statistic. $J = 1$

- Variance of $\mathbf{v}$ = variance of $\mathbf{v}$ from X + variance of $\mathbf{v}$ from Y.
- ME Statistic: $\hat{\lambda}_n(\mathbf{v}) := n \frac{\text{witness}^2(\mathbf{v})}{\text{variance of } \mathbf{v}}$.

Illustration: NME Statistic. $J = 1$

- Variance of $\mathbf{v}$ = variance of $\mathbf{v}$ from X + variance of $\mathbf{v}$ from Y.
- ME Statistic: $\hat{\lambda}_n(\mathbf{v}) := n \frac{\text{witness}^2(\mathbf{v})}{\text{variance of } \mathbf{v}}$.

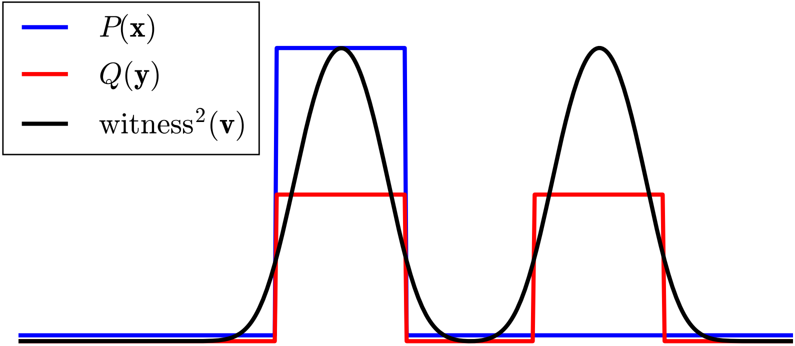


Illustration: NME Statistic. $J = 1$

- Variance of $\mathbf{v}$ = variance of $\mathbf{v}$ from X + variance of $\mathbf{v}$ from Y.
- ME Statistic: $\hat{\lambda}_n(\mathbf{v}) := n\frac{\text{witness}^2(\mathbf{v})}{\text{variance of } \mathbf{v}}$.

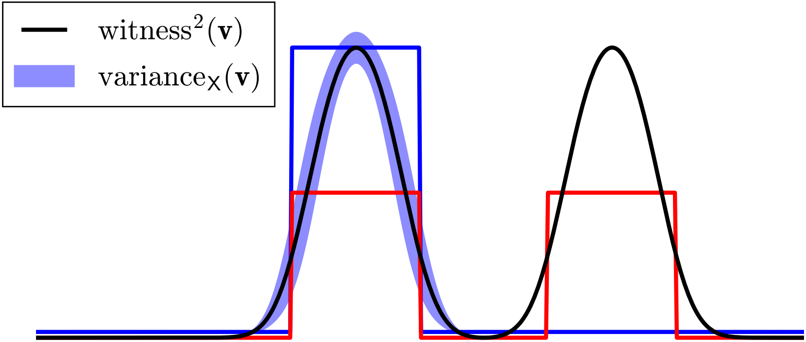


Illustration: NME Statistic. $J = 1$

- Variance of $\mathbf{v}$ = variance of $\mathbf{v}$ from X + variance of $\mathbf{v}$ from Y.
- ME Statistic: $\hat{\lambda}_n(\mathbf{v}) := n \frac{\text{witness}^2(\mathbf{v})}{\text{variance of } \mathbf{v}}$.

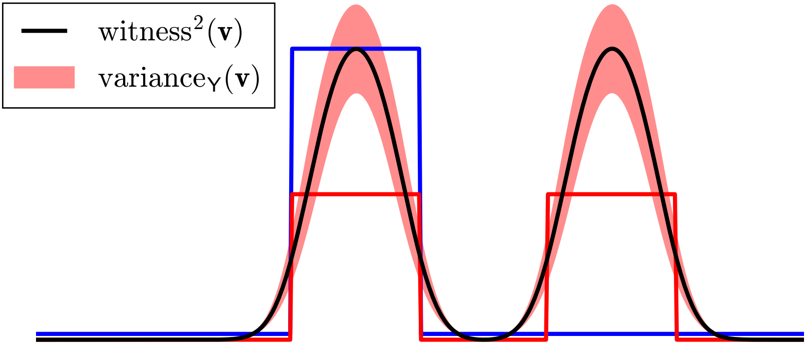


Illustration: NME Statistic. $J = 1$

- Variance of $\mathbf{v}$ = variance of $\mathbf{v}$ from X + variance of $\mathbf{v}$ from Y.
- ME Statistic: $\hat{\lambda}_n(\mathbf{v}) := n \frac{\text{witness}^2(\mathbf{v})}{\text{variance of } \mathbf{v}}$.

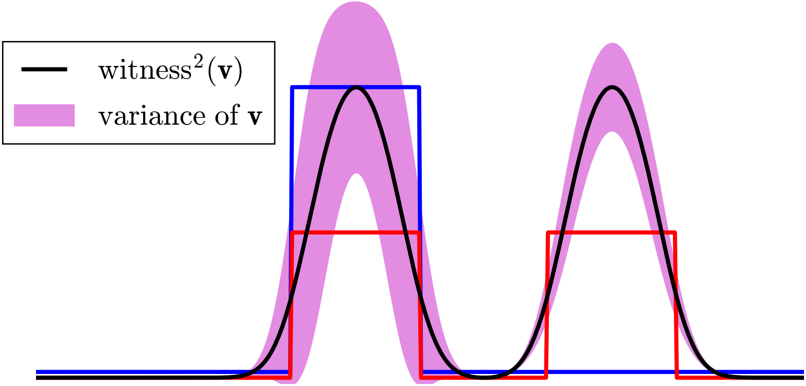


Illustration: NME Statistic. $J = 1$

- Variance of $\mathbf{v}$ = variance of $\mathbf{v}$ from X + variance of $\mathbf{v}$ from Y.
- ME Statistic: $\hat{\lambda}_n(\mathbf{v}) := n \frac{\text{witness}^2(\mathbf{v})}{\text{variance of } \mathbf{v}}$.

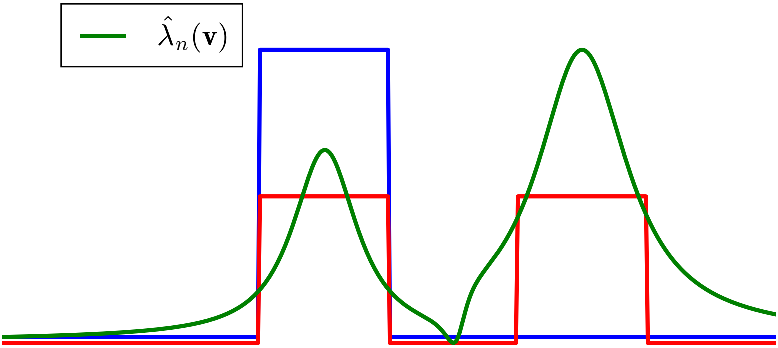
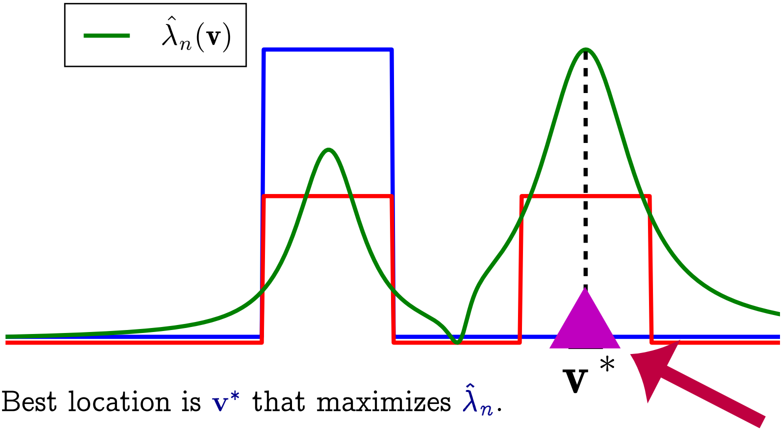


Illustration: NME Statistic. $J = 1$

- Variance of $\mathbf{v}$ = variance of $\mathbf{v}$ from X + variance of $\mathbf{v}$ from Y.
- ME Statistic: $\hat{\lambda}_n(\mathbf{v}) := n \frac{\text{witness}^2(\mathbf{v})}{\text{variance of } \mathbf{v}}$.



- Best location is $\mathbf{v}^*$ that maximizes $\hat{\lambda}_n$.

A Lower Bound on the Test Power of NME

Proposition (Jitkrittum et al., 2016)

The power $\mathbb{P}_{H_1}(\hat{\lambda}_n > T_\alpha) \geq L(\lambda_n) =$

$$1 - 2e^{-\xi_1(\lambda_n - T_\alpha)^2/n} - 2e^{-\frac{[\gamma_n(\lambda_n - T_\alpha)(n-1) - \xi_2 n]^2}{\xi_3 n(2n-1)^2}} - 2e^{-\frac{[(\lambda_n - T_\alpha)/3 - \bar{c}_3 n \gamma_n]^2 \gamma_n^2}{\xi_4}}$$

where

- $\lambda_n = n\mathrm{NME}^2(P, Q)$. *Population quantity.*
- $\gamma_n, \xi_1, \ldots, \xi_4 > 0$ *are constants.*

For large n, $L(\lambda_n)$ is an increasing function of λ_n.

$$\text{Best parameters} = \arg\max L(\lambda_n) = \arg\max \lambda_n.$$

- Optimize (gradient ascent) on a held-out set (estimated λ_n). Test on a separate set.

back to NME

A Lower Bound on the Test Power of NME

Proposition (Jitkrittum et al., 2016)

The power $\mathbb{P}_{H_1}(\hat{\lambda}_n > T_\alpha) \geq L(\lambda_n) =$

$$1 - 2e^{-\xi_1(\lambda_n - T_\alpha)^2/n} - 2e^{-\frac{[\gamma_n(\lambda_n - T_\alpha)(n-1) - \xi_2 n]^2}{\xi_3 n(2n-1)^2}} - 2e^{-\frac{[(\lambda_n - T_\alpha)/3 - \bar{c}_3 n \gamma_n]^2 \gamma_n^2}{\xi_4}}$$

where

- $\lambda_n = n\mathrm{NME}^2(P, Q)$. *Population quantity.*
- $\gamma_n, \xi_1, \ldots, \xi_4 > 0$ *are constants.*

For large n, $L(\lambda_n)$ is an increasing function of λ_n.

$$\boxed{\text{Best parameters} = \arg\max L(\lambda_n) = \arg\max \lambda_n.}$$

- Optimize (gradient ascent) on a held-out set (estimated λ_n). Test on a separate set.

back to NME

A Lower Bound on the Test Power of NME

Proposition (Jitkrittum et al., 2016)

The power $\mathbb{P}_{H_1}(\hat{\lambda}_n > T_\alpha) \geq L(\lambda_n) =$

$$1 - 2e^{-\xi_1(\lambda_n - T_\alpha)^2/n} - 2e^{-\frac{[\gamma_n(\lambda_n - T_\alpha)(n-1) - \xi_2 n]^2}{\xi_3 n(2n-1)^2}} - 2e^{-\frac{[(\lambda_n - T_\alpha)/3 - \bar{c}_3 n \gamma_n]^2 \gamma_n{}^2}{\xi_4}}$$

where

- $\lambda_n = n\mathrm{NME}^2(P, Q)$. *Population quantity.*
- $\gamma_n, \xi_1, \ldots, \xi_4 > 0$ *are constants.*

For large n, $L(\lambda_n)$ is an increasing function of λ_n.

$$\boxed{\text{Best parameters} = \arg\max L(\lambda_n) = \arg\max \lambda_n.}$$

- Optimize (gradient ascent) on a held-out set (estimated λ_n). Test on a separate set.

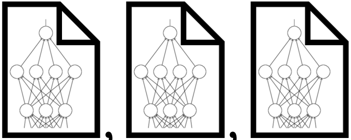
back to NME

Bayesian Inference Vs. Deep Learning Papers

Papers on Bayesian inference

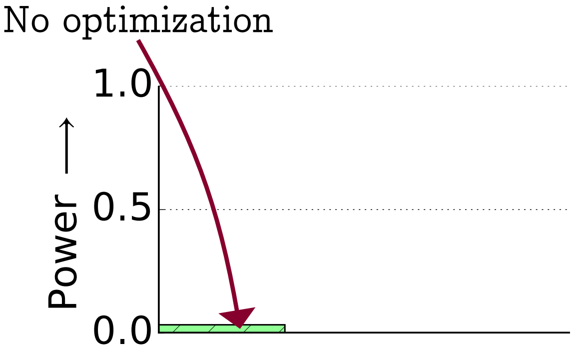
$X = \{$  $, \ldots\} \sim P$

Papers on deep learning

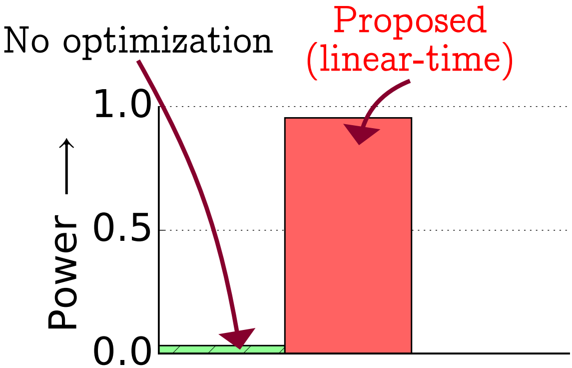
$Y = \{$  $, \ldots\} \sim Q$

- NIPS papers (1988-2015)
- Sample size $n = 216$.
- Random 2000 nouns (dimensions). TF-IDF representation.

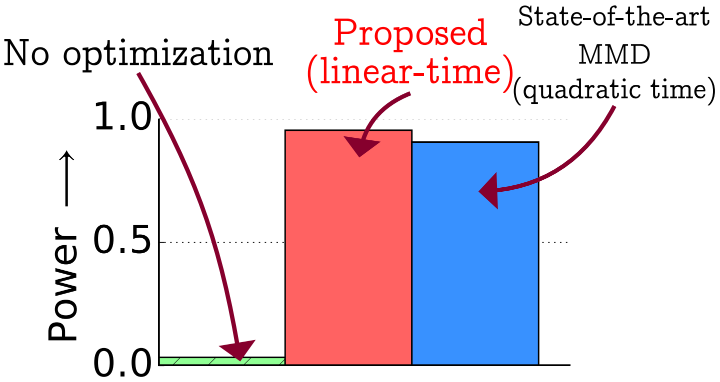
Bayesian Inference Vs. Deep Learning Papers



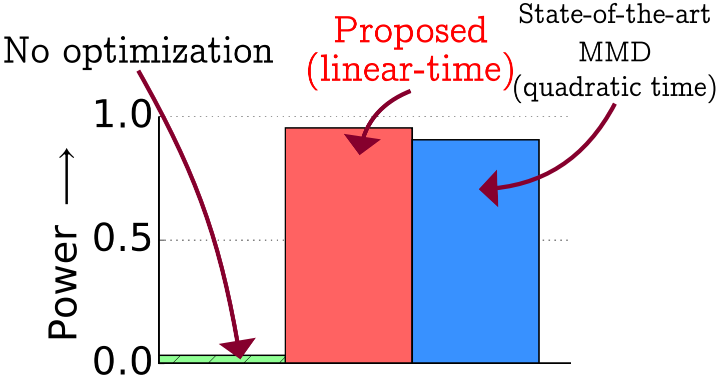
Bayesian Inference Vs. Deep Learning Papers



Bayesian Inference Vs. Deep Learning Papers



Bayesian Inference Vs. Deep Learning Papers



Learned informative feature (a new document):

infer, Bayes, Monte Carlo, adaptor, motif,
haplotype, ECG, covariance, Boltzmann

Outline

1 Appendix: UME, NME

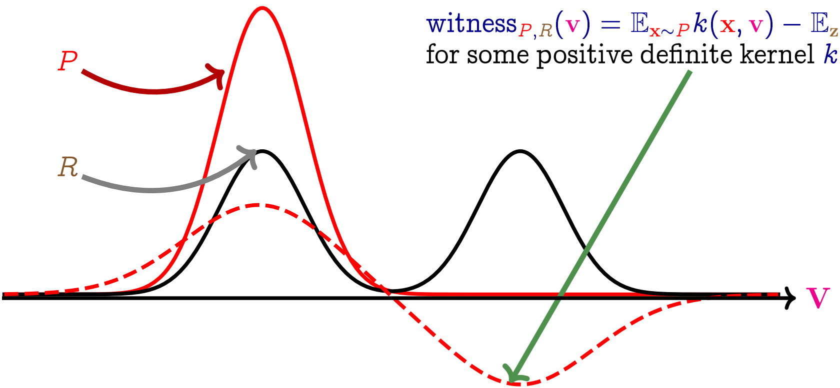
2 Appendix: Relative UME

3 Appendix: Kernel Stein Discrepancy

4 Appendix: FSSD

Rel-UME: Difference of Two Witness Functions

Recall the witness function between P and R:

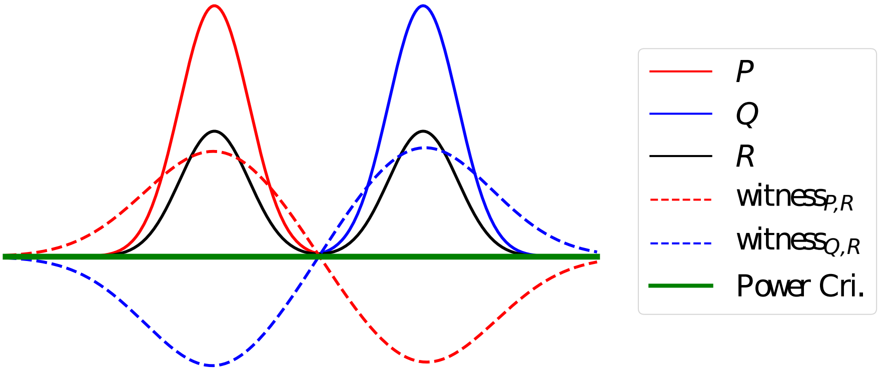


$\text{witness}_{P,R}(\mathbf{v}) = \mathbb{E}_{\mathbf{x}\sim P} k(\mathbf{x}, \mathbf{v}) - \mathbb{E}_{\mathbf{z}\sim R} k(\mathbf{z}, \mathbf{v})$
for some positive definite kernel $k(\mathbf{x}, \mathbf{v})$.

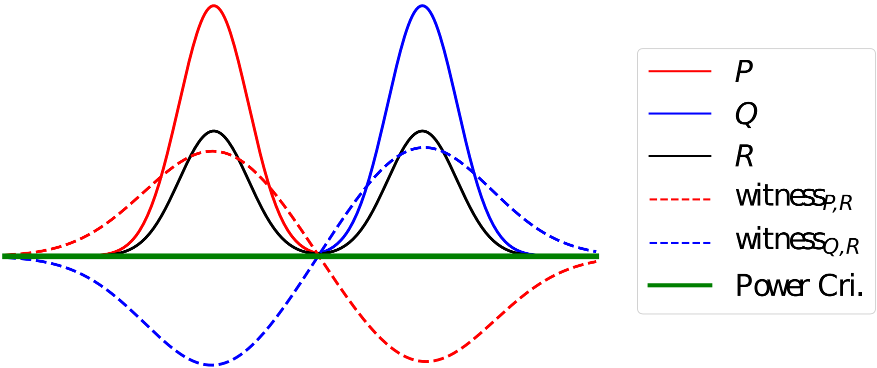
Assume only one test location $\mathbf{v}$. Recall

$$\text{UME}^2_{\mathbf{v}}(P, R) = \text{witness}^2_{P,R}(\mathbf{v}) = \left(\mu_P(\mathbf{v}) - \mu_R(\mathbf{v})\right)^2$$

Rel-UME: Difference of Two Witness Functions



Rel-UME: Difference of Two Witness Functions

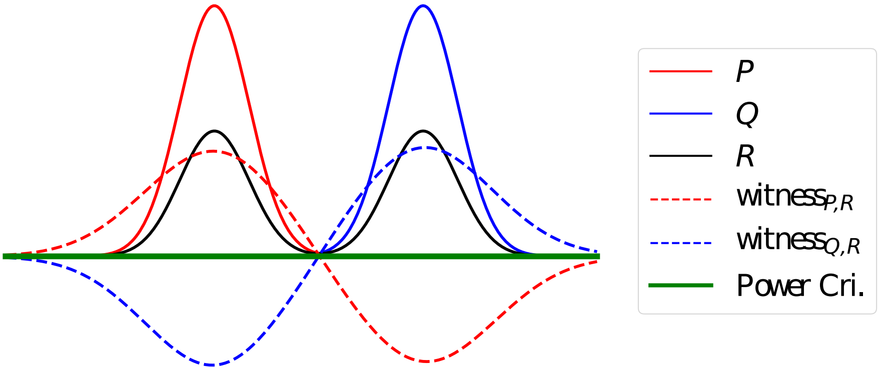


- Power criterion$(\mathbf{v}) = f(\mathbf{v})$ is a function such that maximizing it corresponds to maximizing the test power.

$$f(\mathbf{v}) = \frac{\text{witness}^2_{P,R}(\mathbf{v}) - \text{witness}^2_{Q,R}(\mathbf{v})}{\text{standard deviation}_{P,Q,R}(\mathbf{v})} = \frac{U_P^2 - U_Q^2}{\sqrt{4(\zeta_P^2 - 2\zeta_{PQ} + \zeta_Q^2)}}$$

- $f(\mathbf{v}) > 0 \implies Q$ is better in the region around $\mathbf{v}$
- $f(\mathbf{v}) < 0 \implies P$ is better in the region around $\mathbf{v}$

Rel-UME: Difference of Two Witness Functions

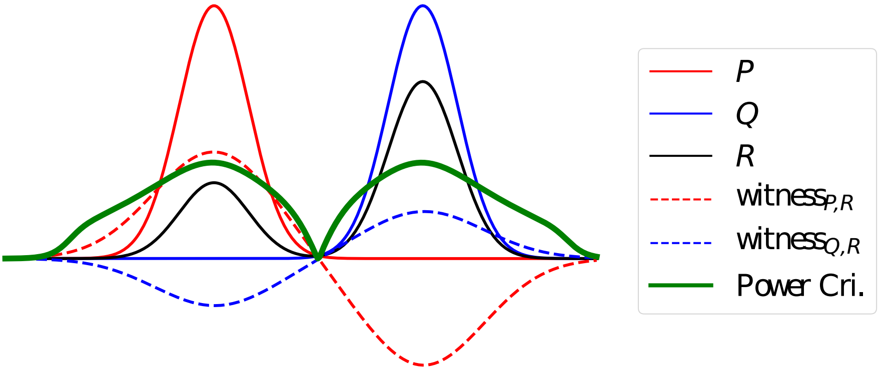


- Power criterion$(\mathbf{v}) = f(\mathbf{v})$ is a function such that maximizing it corresponds to maximizing the test power.

$$f(\mathbf{v}) = \frac{\text{witness}^2_{P,R}(\mathbf{v}) - \text{witness}^2_{Q,R}(\mathbf{v})}{\text{standard deviation}_{P,Q,R}(\mathbf{v})} = \frac{U_P^2 - U_Q^2}{\sqrt{4(\zeta_P^2 - 2\zeta_{PQ} + \zeta_Q^2)}}$$

- $f(\mathbf{v}) > 0 \implies Q$ is better in the region around $\mathbf{v}$
- $f(\mathbf{v}) < 0 \implies P$ is better in the region around $\mathbf{v}$

Rel-UME: Difference of Two Witness Functions



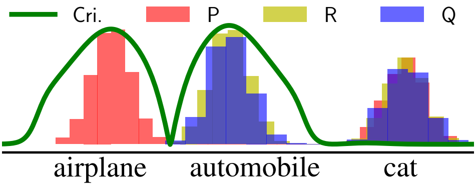
- Power criterion$(\mathbf{v}) = f(\mathbf{v})$ is a function such that maximizing it corresponds to maximizing the test power.

$$f(\mathbf{v}) = \frac{\text{witness}^2_{P,R}(\mathbf{v}) - \text{witness}^2_{Q,R}(\mathbf{v})}{\text{standard deviation}_{P,Q,R}(\mathbf{v})} = \frac{U_P^2 - U_Q^2}{\sqrt{4(\zeta_P^2 - 2\zeta_{PQ} + \zeta_Q^2)}}$$

- $f(\mathbf{v}) > 0 \implies Q$ is better in the region around $\mathbf{v}$
- $f(\mathbf{v}) < 0 \implies P$ is better in the region around $\mathbf{v}$

Experiment on CIFAR10

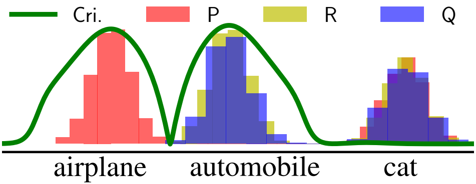
- $P = \{\text{airplane, cat}\}$,
  $Q = \{\text{automobile, cat}\}$
- (true) $R = \{\text{automobile, cat}\}$



- Gaussian kernel on 2048 features extracted by the Inception-v3 network at the pool3 layer.

Experiment on CIFAR10

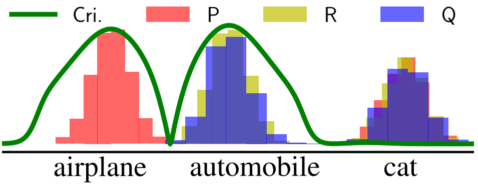
- $P = \{\text{airplane, cat}\}$,
  $Q = \{\text{automobile, cat}\}$
- (true) $R = \{\text{automobile, cat}\}$



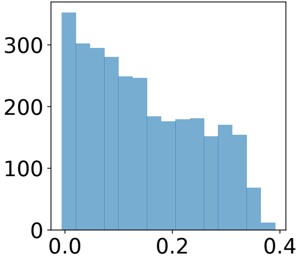
- Gaussian kernel on 2048 features extracted by the Inception-v3 network at the pool3 layer.

Experiment on CIFAR10

- $P = \{\text{airplane, cat}\}$,
  $Q = \{\text{automobile, cat}\}$
- (true) $R = \{\text{automobile, cat}\}$



- Gaussian kernel on 2048 features extracted by the Inception-v3 network at the pool3 layer.

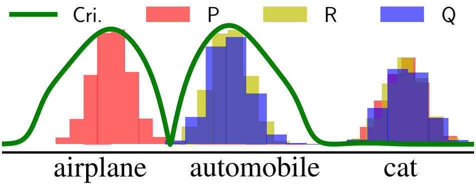


Histogram of power criterion values $f(\mathbf{v})$ evaluated at $\mathbf{v} = \{\text{airplane, automobile, cat}\}$.

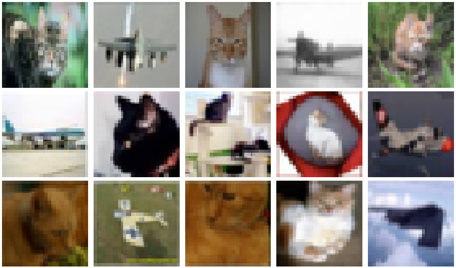
- All non-negative. $\Longrightarrow$ Q is equally good or better than P everywhere.

Experiment on CIFAR10

- $P = \{\text{airplane, cat}\}$,
  $Q = \{\text{automobile, cat}\}$
- (true) $R = \{\text{automobile, cat}\}$



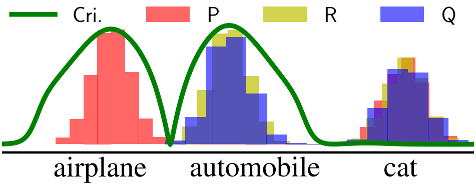
- Gaussian kernel on 2048 features extracted by the Inception-v3 network at the pool3 layer.



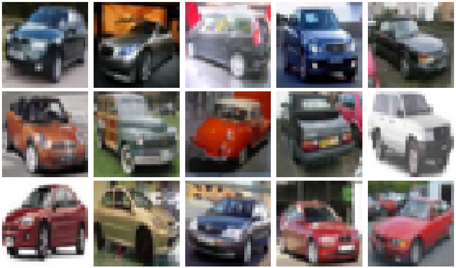
Images $\mathbf{v}$ with the lowest values of $f(\mathbf{v}) \approx 0$. $\Longrightarrow$ P, Q perform equally well in these regions.

Experiment on CIFAR10

- P = {airplane, cat},
  Q = {automobile, cat}
- (true) R = {automobile, cat}



- Gaussian kernel on 2048 features extracted by the Inception-v3 network at the pool3 layer.



Images $\mathbf{v}$ with the highest values of $f(\mathbf{v}) > 0$. $\Longrightarrow$ Q is better than P in these regions.

Outline

1 Appendix: UME, NME

2 Appendix: Relative UME

3 Appendix: Kernel Stein Discrepancy

4 Appendix: FSSD

Stein operator is linear

Re-write Stein operator as:

$$
\begin{aligned}
[T_p f](x) &= \frac{1}{p(x)} \frac{d}{dx} \left(f(x) p(x) \right) \\
&= \frac{1}{p(x)} \left[p(x) \frac{df}{dx}(x) + f(x) \frac{dp}{dx}(x) \right] \\
&= f(x) \frac{d}{dx} \log p(x) + \frac{d}{dx} f(x)
\end{aligned}
$$

Stein features in $\mathcal{F}$

$$
\begin{aligned}
[T_p f](x) &= \left(\frac{d}{dx} \log p(x) \right) f(x) + \frac{d}{dx} f(x) \\
&=: \langle f, \underbrace{\xi(x)}_{\text{Stein features}} \rangle_{\mathcal{F}}
\end{aligned}
$$

where $\mathbf{E}_{x \sim p} \xi(x) = 0$.

go to KSD

Stein operator is linear

Re-write Stein operator as:

$$
\begin{aligned}
[T_p f](x) &= \frac{1}{p(x)} \frac{d}{dx}\left(f(x)p(x)\right) \\
&= \frac{1}{p(x)}\left[p(x)\frac{df}{dx}(x) + f(x)\frac{dp}{dx}(x)\right] \\
&= f(x)\frac{d}{dx}\log p(x) + \frac{d}{dx}f(x)
\end{aligned}
$$

Stein features in $\mathcal{F}$

$$
\begin{aligned}
[T_p f](x) &= \left(\frac{d}{dx}\log p(x)\right) f(x) + \frac{d}{dx} f(x) \\
&=: \langle f, \underbrace{\xi(x)}_{\text{Stein features}} \rangle_{\mathcal{F}}
\end{aligned}
$$

where $\mathbf{E}_{x\sim p}\xi(x) = 0$.

go to KSD

The kernel trick for derivatives

Reproducing property for the derivative: for differentiable $k(x, x')$,

$$\frac{d}{dx} f(x) = \left\langle f, \frac{d}{dx} \varphi(x) \right\rangle_{\mathcal{F}}$$

The kernel trick for derivatives

Reproducing property for the derivative: for differentiable $k(x, x')$,

$$\frac{d}{dx} f(x) = \left\langle f, \frac{d}{dx} \varphi(x) \right\rangle_{\mathcal{F}}$$

Using kernel derivative trick in (a),

$$\begin{aligned}
[T_p f](x) &= \left(\frac{d}{dx} \log p(x) \right) f(x) + \frac{d}{dx} f(x) \\
&= \left\langle f, \left(\frac{d}{dx} \log p(x) \right) \varphi(x) + \frac{d}{dx} \varphi(x) \right\rangle_{\mathcal{F}} \\
&=: \langle f, \xi(x) \rangle_{\mathcal{F}} .
\end{aligned}$$

Kernel Stein Discrepancy: Derivation

- Can be shown that $[T_p f](x) = \langle f, \xi(x) \rangle_{\mathcal{F}}$. where
  - $\xi(x) = \left(\frac{d}{dx} \log p(x)\right) \varphi(x) + \frac{d}{dx} \varphi(x)$,
  - $\varphi(x) =$ feature map associated with k

Closed-form expression for KSD:

$$\begin{aligned} \mathrm{KSD}_p(q) &= \sup_{\|f\|_{\mathcal{F}} \leq 1} \mathbf{E}_{y \sim q} [T_p f](y) \\ &= \sup_{\|f\|_{\mathcal{F}} \leq 1} \mathbf{E}_{y \sim q} \langle f, \xi(y) \rangle_{\mathcal{F}} \\ &\underset{(a)}{=} \sup_{\|f\|_{\mathcal{F}} \leq 1} \langle f, \mathbf{E}_{y \sim q} \xi(y) \rangle_{\mathcal{F}} \underset{(b)}{=} \|\mathbf{E}_{y \sim q} \xi(y)\|_{\mathcal{F}}. \end{aligned}$$

Chwialkowski, Strathmann, G., (ICML 2016) Liu, Lee, Jordan (ICML 2016)

Kernel Stein Discrepancy: Derivation

- Can be shown that $[T_p f](x) = \langle f, \xi(x) \rangle_{\mathcal{F}}$. where
  - $\xi(x) = \left(\frac{d}{dx} \log p(x)\right) \varphi(x) + \frac{d}{dx}\varphi(x)$,
  - $\varphi(x) =$ feature map associated with k

Closed-form expression for KSD:

$$
\begin{aligned}
\mathrm{KSD}_p(q) &= \sup_{\|f\|_{\mathcal{F}} \le 1} \mathbf{E}_{y\sim q} [T_p f](y) \\
&= \sup_{\|f\|_{\mathcal{F}} \le 1} \mathbf{E}_{y\sim q} \langle f, \xi(y) \rangle_{\mathcal{F}} \\
&\underset{(a)}{=} \sup_{\|f\|_{\mathcal{F}} \le 1} \langle f, \mathbf{E}_{y\sim q}\xi(y) \rangle_{\mathcal{F}} \underset{(b)}{=} \|\mathbf{E}_{y\sim q}\xi(y)\|_{\mathcal{F}}.
\end{aligned}
$$

Kernel Stein Discrepancy: Derivation

- Can be shown that $[T_p f](x) = \langle f, \xi(x) \rangle_{\mathcal{F}}$. where
  - $\xi(x) = \left(\frac{d}{dx} \log p(x)\right) \varphi(x) + \frac{d}{dx}\varphi(x)$,
  - $\varphi(x) =$ feature map associated with k

Closed-form expression for KSD:

$$\begin{aligned}
\mathrm{KSD}_p(q) &= \sup_{\|f\|_{\mathcal{F}} \le 1} \mathbf{E}_{y\sim q} [T_p f](y) \\
&= \sup_{\|f\|_{\mathcal{F}} \le 1} \mathbf{E}_{y\sim q} \langle f, \xi(y) \rangle_{\mathcal{F}} \\
&\underset{(a)}{=} \sup_{\|f\|_{\mathcal{F}} \le 1} \langle f, \mathbf{E}_{y\sim q} \xi(y) \rangle_{\mathcal{F}} \underset{(b)}{=} \|\mathbf{E}_{y\sim q} \xi(y)\|_{\mathcal{F}}.
\end{aligned}$$

- At (b), we have $f^* = \mathbf{E}_{y\sim q}\xi(y)$ as the $\arg\sup$.

Kernel Stein Discrepancy: Derivation

- Can be shown that $[T_p f](x) = \langle f, \xi(x) \rangle_{\mathcal{F}}$. where
  - $\xi(x) = \left(\frac{d}{dx} \log p(x)\right) \varphi(x) + \frac{d}{dx} \varphi(x)$,
  - $\varphi(x) =$ feature map associated with k

Closed-form expression for KSD:

$$\begin{aligned}
\mathrm{KSD}_p(q) &= \sup_{\|f\|_{\mathcal{F}} \le 1} \mathbf{E}_{y \sim q} [T_p f](y) \\
&= \sup_{\|f\|_{\mathcal{F}} \le 1} \mathbf{E}_{y \sim q} \langle f, \xi(y) \rangle_{\mathcal{F}} \\
&\underset{(a)}{=} \sup_{\|f\|_{\mathcal{F}} \le 1} \langle f, \mathbf{E}_{y \sim q} \xi(y) \rangle_{\mathcal{F}} \underset{(b)}{=} \|\mathbf{E}_{y \sim q} \xi(y)\|_{\mathcal{F}}.
\end{aligned}$$

- At (b), we have $f^* = \mathbf{E}_{y \sim q} \xi(y)$ as the $\arg\sup$.

Caution: (a) requires a condition for the Riesz theorem to hold,

$$\mathbf{E}_{x \sim q} \left(\frac{d}{dx} \log p(x)\right)^2 < \infty.$$

Chwialkowski, Strathmann, G., (ICML 2016) Liu, Lee, Jordan (ICML 2016)

KSD: Empirical statistic and asymptotics

- **Given**: $\{y_i\}_{i=1}^n \overset{i.i.d.}{\sim} q$, a differentiable density p.

The empirical statistic:

$$\widehat{\mathrm{KSD}_p^2}(q) := \frac{1}{n^2} \sum_{i=1}^{n} \sum_{j=1}^{n} h_p(y_i, y_j).$$

KSD: Empirical statistic and asymptotics

- **Given**: $\{y_i\}_{i=1}^n \overset{i.i.d.}{\sim} q$, a differentiable density p.

The empirical statistic:

$$\widehat{\mathrm{KSD}_p^2}(q) := \frac{1}{n^2} \sum_{i=1}^n \sum_{j=1}^n h_p(y_i, y_j).$$

Asymptotic distribution when $p \neq q$:

$$\sqrt{n}\left(\widehat{\mathrm{KSD}_p^2}(q) - \mathrm{KSD}_p^2(q)\right) \xrightarrow{d} \mathcal{N}(0, \sigma_{h_p}^2) \qquad \sigma_{h_p}^2 = 4\mathrm{Var}_y[\mathbf{E}_{y'}[h_p(y, y')]].$$

KSD: Empirical statistic and asymptotics

- **Given**: $\{y_i\}_{i=1}^n \overset{i.i.d.}{\sim} q$, a differentiable density p.

The empirical statistic:

$$\widehat{\mathrm{KSD}_p^2}(q) := \frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n} h_p(y_i, y_j).$$

Asymptotic distribution when $p = q$:

$$n\widehat{\mathrm{KSD}_p^2}(q) \sim \sum_{\ell=1}^{\infty} \lambda_\ell Z_\ell^2 \text{ where } Z_\ell \sim \mathcal{N}(0,1) \quad \text{i.i.d.},$$

$$\lambda_i \psi_i(x') = \int_{\mathcal{X}} h_p(x, x')\psi_i(x)\, dp(x).$$

Get test threshold via wild bootstrap.

Wild bootstrap test for KSD [Chwialkowski et al. ICML 2016)]

Generate samples $B_1, \ldots, B_m$ by wild bootstrap

1. For $l = 1, \ldots, m$:
   1. Draw i.i.d. $W_1, \ldots, W_n$ (-1/+1) where $P(W_i = 1) = P(W_1 = -1) = 1/2$.
   2. $B_l := \frac{1}{n^2} \sum_{i=1}^{n} \sum_{j=1}^{n} W_i W_j h_p(y_i, y_j)$
2. Threshold = $(1 - \alpha)$-quantile from $\{B_1, \ldots, B_m\}$
3. Reject H_0 if $\widehat{\mathrm{KSD}^2_p}(q) = \frac{1}{n^2} \sum_{i=1}^{n} \sum_{j=1}^{n} h_p(y_i, y_j)$ is larger than the threshold.

Proposition ([Chwialkowski et al. ICML 2016)])

- *When $p = q$, $B_1, \ldots, B_m$ are samples from the null distribution as $n \to \infty$.*
- *When $p \neq q$, $B_1, \ldots, B_m$ converge to 0. $\widehat{\mathrm{KSD}^2_p}(q)$ converges to $\mathrm{KSD}^2_p(q) > 0$.*

◂ return to KSD

Wild bootstrap test for KSD [Chwialkowski et al. ICML 2016)]

Generate samples $B_1, \ldots, B_m$ by wild bootstrap

1. For $l = 1, \ldots, m$:
   1. Draw i.i.d. $W_1, \ldots, W_n$ (-1/+1) where $P(W_i = 1) = P(W_1 = -1) = 1/2$.
   2. $B_l := \frac{1}{n^2} \sum_{i=1}^{n} \sum_{j=1}^{n} W_i W_j h_p(y_i, y_j)$
2. Threshold $= (1 - \alpha)$-quantile from $\{B_1, \ldots, B_m\}$
3. Reject H_0 if $\widehat{\mathrm{KSD}^2_p}(q) = \frac{1}{n^2} \sum_{i=1}^{n} \sum_{j=1}^{n} h_p(y_i, y_j)$ is larger than the threshold.

Proposition ([Chwialkowski et al. ICML 2016)])

- *When $p = q$, $B_1, \ldots, B_m$ are samples from the null distribution as $n \to \infty$.*
- *When $p \neq q$, $B_1, \ldots, B_m$ converge to 0. $\widehat{\mathrm{KSD}^2_p}(q)$ converges to $\mathrm{KSD}^2_p(q) > 0$.*

◂ return to KSD

Wild bootstrap test for KSD [Chwialkowski et al. ICML 2016)]

Generate samples $B_1, \ldots, B_m$ by wild bootstrap

1. For $l = 1, \ldots, m$:
   1. Draw i.i.d. $W_1, \ldots, W_n$ (-1/+1) where $P(W_i = 1) = P(W_1 = -1) = 1/2$.
   2. $B_l := \frac{1}{n^2} \sum_{i=1}^n \sum_{j=1}^n W_i W_j h_p(y_i, y_j)$
2. Threshold $= (1-\alpha)$-quantile from $\{B_1, \ldots, B_m\}$
3. Reject H_0 if $\widehat{\mathrm{KSD}_p^2}(q) = \frac{1}{n^2} \sum_{i=1}^n \sum_{j=1}^n h_p(y_i, y_j)$ is larger than the threshold.

Proposition ([Chwialkowski et al. ICML 2016)])

- *When $p = q$, $B_1, \ldots, B_m$ are samples from the null distribution as $n \to \infty$.*
- *When $p \neq q$, $B_1, \ldots, B_m$ converge to 0. $\widehat{\mathrm{KSD}_p^2}(q)$ converges to $\mathrm{KSD}_p^2(q) > 0$.*

◂ return to KSD

Wild bootstrap test for KSD [Chwialkowski et al. ICML 2016)]

Generate samples $B_1, \ldots, B_m$ by wild bootstrap

1. For $l = 1, \ldots, m$:
   1. Draw i.i.d. $W_1, \ldots, W_n$ (-1/+1) where $P(W_i = 1) = P(W_1 = -1) = 1/2$.
   2. $B_l := \frac{1}{n^2} \sum_{i=1}^n \sum_{j=1}^n W_i W_j h_p(y_i, y_j)$
2. Threshold $= (1-\alpha)$-quantile from $\{B_1, \ldots, B_m\}$
3. Reject H_0 if $\widehat{\mathrm{KSD}_p^2}(q) = \frac{1}{n^2} \sum_{i=1}^n \sum_{j=1}^n h_p(y_i, y_j)$ is larger than the threshold.

Proposition ([Chwialkowski et al. ICML 2016)])

- *When $p = q$, $B_1, \ldots, B_m$ are samples from the null distribution as $n \to \infty$.*
- *When $p \neq q$, $B_1, \ldots, B_m$ converge to 0. $\widehat{\mathrm{KSD}_p^2}(q)$ converges to $\mathrm{KSD}_p^2(q) > 0$.*

return to KSD

KSD for discrete-valued variables

Discrete domains: $\mathcal{X} = \{1, \ldots, L\}^D$ with $L \in \mathbb{N}$.

The population KSD (discrete):

$$\mathrm{KSD}_p^2(Q) = \mathbf{E}_{x,x' \sim q} h_p(x, x')$$

where

$$\begin{aligned} h_p(x, x') = \mathbf{s}_p(x)^\top \mathbf{s}_p(x') k(x, x') - \mathbf{s}_p(x)^\top k_2(x, x') \\ - \mathbf{s}_p(x')^\top k_1(x, x') + \mathrm{tr}\left[k_{12}(x, x')\right] \end{aligned}$$

$k_1(x, x') = \Delta_x^{-1} k(x, x')$, Δ_x^{-1} is cyclic backwards difference on x,
$\mathbf{s}_p(x) = \frac{\Delta p(x)}{p(x)}$

Ranganath et al. (NeurIPS 2016), Yang et al. (ICML 2018)

KSD for discrete-valued variables

Discrete domains: $\mathcal{X} = \{1, \ldots, L\}^D$ with $L \in \mathbb{N}$.

The population KSD (discrete):

$$\mathrm{KSD}_p^2(Q) = \mathbf{E}_{x,x' \sim q} h_p(x, x')$$

where

$$h_p(x, x') = \mathbf{s}_p(x)^\top \mathbf{s}_p(x') k(x, x') - \mathbf{s}_p(x)^\top k_2(x, x') \\ - \mathbf{s}_p(x')^\top k_1(x, x') + \operatorname{tr}\left[k_{12}(x, x')\right]$$

$k_1(x, x') = \Delta_x^{-1} k(x, x')$, Δ_x^{-1} is cyclic backwards difference on x, $\mathbf{s}_p(x) = \frac{\Delta p(x)}{p(x)}$

A discrete kernel: $k(x, x') = \exp\left(-d_H(x, x')\right)$, where $d_H(x, x') = D^{-1} \sum_{d=1}^{D} \mathbb{I}(x_d \neq x'_d)$.

Ranganath et al. (NeurIPS 2016), Yang et al. (ICML 2018)

KSD for discrete-valued variables

Discrete domains: $\mathcal{X} = \{1, \dots, L\}^D$ with $L \in \mathbb{N}$.

The population KSD (discrete):

$$\mathrm{KSD}_p^2(Q) = \mathbf{E}_{x,x' \sim q} h_p(x, x')$$

where

$$\begin{aligned} h_p(x, x') = \mathbf{s}_p(x)^\top \mathbf{s}_p(x') k(x, x') - \mathbf{s}_p(x)^\top k_2(x, x') \\ - \mathbf{s}_p(x')^\top k_1(x, x') + \mathrm{tr}\left[k_{12}(x, x')\right] \end{aligned}$$

$k_1(x, x') = \Delta_x^{-1} k(x, x')$, Δ_x^{-1} is cyclic backwards difference on x, $\mathbf{s}_p(x) = \frac{\Delta p(x)}{p(x)}$

A discrete kernel: $k(x, x') = \exp\left(-d_H(x, x')\right)$, where $d_H(x, x') = D^{-1} \sum_{d=1}^{D} \mathbb{I}(x_d \neq x'_d)$.

$\mathrm{KSD}_p^2(Q) = 0$ iff $P = Q$ if

- Gram matrix over all the configurations in $\mathcal{X}$ is strictly positive definite,
- $P > 0$ and $Q > 0$.

Ranganath et al. (NeurIPS 2016), Yang et al. (ICML 2018)

Outline

1 Appendix: UME, NME

2 Appendix: Relative UME

3 Appendix: Kernel Stein Discrepancy

4 Appendix: FSSD

FSSD is a Discrepancy Measure

Theorem

Let $\mathcal{X}$ be a connected open set in $\mathbb{R}^d$. Assume

1. *(Nice RKHS) Kernel $k\colon \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ 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) $\lim_{\|\mathbf{x}\|\to\infty} p(\mathbf{x})\mathbf{g}(\mathbf{x}) = \mathbf{0}$.*

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. Then, for any $J \geq 1$,

- *If $p = q$, $\mathrm{FSSD}^2 = 0$.*
- *If $p \neq q$, η-almost surely, $\mathrm{FSSD}^2 > 0$.*

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

FSSD is a Discrepancy Measure

Theorem

Let $\mathcal{X}$ be a connected open set in $\mathbb{R}^d$. Assume

1. *(Nice RKHS) Kernel $k\colon \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ 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) $\lim_{\|\mathbf{x}\|\to\infty} p(\mathbf{x})\mathbf{g}(\mathbf{x}) = \mathbf{0}$.*

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. Then, for any $J \geq 1$,

- *If $p = q$, $\mathrm{FSSD}^2 = 0$.*
- *If $p \neq q$, η-almost surely, $\mathrm{FSSD}^2 > 0$.*

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

FSSD is a Discrepancy Measure

Theorem

Let $\mathcal{X}$ be a connected open set in $\mathbb{R}^d$. Assume

1. *(Nice RKHS) Kernel $k\colon \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ 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) $\lim_{\|\mathbf{x}\|\to\infty} p(\mathbf{x})\mathbf{g}(\mathbf{x}) = \mathbf{0}$.*

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. Then, for any $J \geq 1$,

- *If $p = q$, $\mathrm{FSSD}^2 = 0$.*
- *If $p \neq q$, η-almost surely, $\mathrm{FSSD}^2 > 0$.*

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

What Are "Blind Spots" in the Stein Witness?

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

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

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

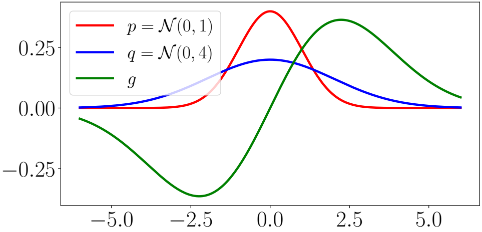
- If $v = 0$, then $\mathrm{FSSD}^2 = g^2(v) = 0$ regardless of σ_q^2.
- If $g \neq 0$, and k is real analytic, $R = \{v \mid g(v) = 0\}$ (blind spots) has 0 Lebesgue measure.
- So, if $v \sim$ a distribution with a density, then $v \notin R$.

What Are “Blind Spots” in the Stein Witness?

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

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

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



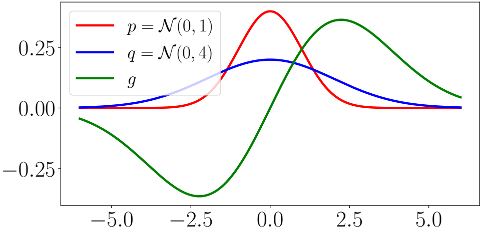
- If $v = 0$, then $\mathrm{FSSD}^2 = g^2(v) = 0$ regardless of σ_q^2.
- If $g \neq 0$, and k is real analytic, $R = \{v \mid g(v) = 0\}$ (blind spots) has 0 Lebesgue measure.
- So, if $v \sim$ a distribution with a density, then $v \notin R$.

What Are "Blind Spots" in the Stein Witness?

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

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

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



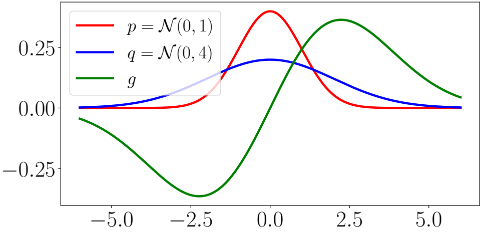
- If $v = 0$, then $\mathrm{FSSD}^2 = g^2(v) = 0$ regardless of σ_q^2.
- If $g \neq 0$, and k is real analytic, $R = \{v \mid g(v) = 0\}$ (blind spots) has 0 Lebesgue measure.
- So, if $v \sim$ a distribution with a density, then $v \notin R$.

What Are "Blind Spots" in the Stein Witness?

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

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

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



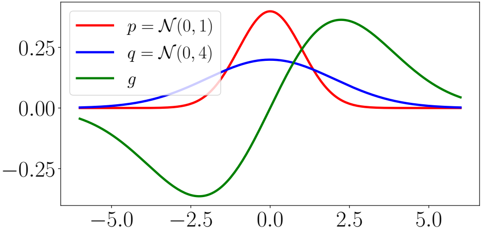
- If $v = 0$, then $\mathrm{FSSD}^2 = g^2(v) = 0$ regardless of σ_q^2.
- If $g \neq 0$, and k is real analytic, $R = \{v \mid g(v) = 0\}$ (blind spots) has 0 Lebesgue measure.
- So, if $v \sim$ a distribution with a density, then $v \notin R$.

What Are "Blind Spots" in the Stein Witness?

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

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 $v \notin R$.