# Informative Features for Model Comparison


**Wittawat Jitkrittum**

Max Planck Institute for Intelligent Systems

wittawat.com

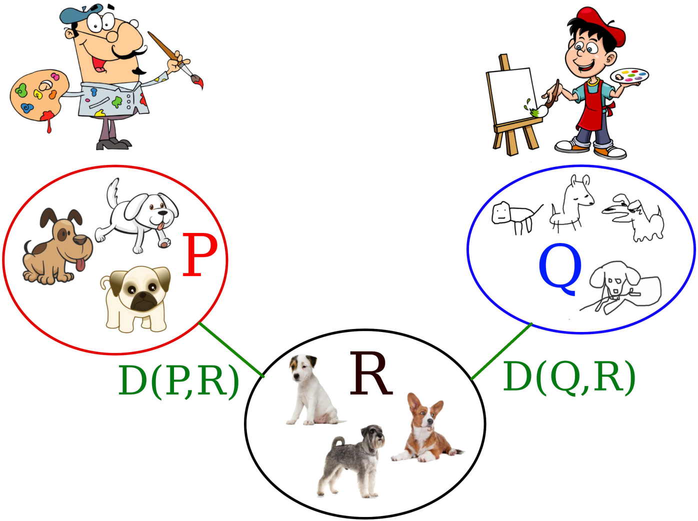
Heishiro Kanagawa, Patsorn Sangkloy, James Hays,
Bernhard Schölkopf, Arthur Gretton

Swiss Data Science Center

29 October 2019

# Model Comparison




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

# Outline

1. Problem setting
2. Motivations for the proposed test
3. Hypothesis testing 101
4. The Unnormalized Mean Embeddings (UME) statistic (3-sample test)
   1. Asymptotic distributions
   2. Interpretability
5. Experiments
6. The Finite Set Stein Discrepancy (FSSD) statistic (2 density models and 1 set of samples)

# Problem Setting

- $P$, $Q$ : candidate generative models that can be sampled e.g., GANs.
- $R$ : data generating distribution (unknown).
- Observe $X_n \overset{i.i.d.}{\sim} P$, $Y_n \overset{i.i.d.}{\sim} Q$, and $Z_n \overset{i.i.d.}{\sim} R$ be three sets of samples, each of size $n$.

$$H_0: P \text{ and } Q \text{ equally model } R$$
$$H_1: Q \text{ models } R \text{ better.}$$

- Formulate as

$$H_0: D(P, R) - D(Q, R) = 0$$
$$H_1: D(P, R) - D(Q, R) > 0,$$

for some distance $D$.

- Statistic: $\hat{S}_n = \widehat{D}(P, R) - \widehat{D}(Q, R)$. Large, positive $\implies$ $Q$ is better.

# Problem Setting

- $P$, $Q$ : candidate generative models that can be sampled e.g., GANs.
- $R$ : data generating distribution (unknown).
- Observe $X_n \overset{i.i.d.}{\sim} P$, $Y_n \overset{i.i.d.}{\sim} Q$, and $Z_n \overset{i.i.d.}{\sim} R$ be three sets of samples, each of size $n$.

$H_0$: $P$ and $Q$ equally model $R$

$H_1$: $Q$ models $R$ better.

- Formulate as

$$H_0: D(P,R) - D(Q,R) = 0$$
$$H_1: D(P,R) - D(Q,R) > 0,$$

for some distance $D$.

- Statistic: $\hat{S}_n = \widehat{D}(P,R) - \widehat{D}(Q,R)$. Large, positive $\implies$ $Q$ is better.

# Problem Setting

- $P$, $Q$ : candidate generative models that can be sampled e.g., GANs.
- $R$ : data generating distribution (unknown).
- Observe $X_n \overset{i.i.d.}{\sim} P$, $Y_n \overset{i.i.d.}{\sim} Q$, and $Z_n \overset{i.i.d.}{\sim} R$ be three sets of samples, each of size $n$.

$$H_0 : P \text{ and } Q \text{ equally model } R$$
$$H_1 : Q \text{ models } R \text{ better.}$$

- Formulate as

$$H_0 : D(P, R) - D(Q, R) = 0$$
$$H_1 : D(P, R) - D(Q, R) > 0,$$

  for some distance $D$.
- Statistic: $\hat{S}_n = \widehat{D}(P, R) - \widehat{D}(Q, R)$. Large, positive $\implies$ $Q$ is better.

# Problem Setting

- $P$, $Q$ : candidate generative models that can be sampled e.g., GANs.
- $R$ : data generating distribution (unknown).
- Observe $X_n \overset{i.i.d.}{\sim} P$, $Y_n \overset{i.i.d.}{\sim} Q$, and $Z_n \overset{i.i.d.}{\sim} R$ be three sets of samples, each of size $n$.

$$H_0: P \text{ and } Q \text{ equally model } R$$
$$H_1: Q \text{ models } R \text{ better.}$$

- Formulate as

$$H_0: D(P, R) - D(Q, R) = 0$$
$$H_1: D(P, R) - D(Q, R) > 0,$$

for some distance $D$.

- Statistic: $\hat{S}_n = \widehat{D}(P, R) - \widehat{D}(Q, R)$. Large, positive $\Longrightarrow$ $Q$ is better.

# Problem Setting

- $P$, $Q$ : candidate generative models that can be sampled e.g., GANs.
- $R$ : data generating distribution (unknown).
- Observe $X_n \overset{i.i.d.}{\sim} P$, $Y_n \overset{i.i.d.}{\sim} Q$, and $Z_n \overset{i.i.d.}{\sim} R$ be three sets of samples, each of size $n$.

$$H_0\colon P \text{ and } Q \text{ equally model } R$$
$$H_1\colon Q \text{ models } R \text{ better.}$$

- Formulate as

$$H_0\colon D(P, R) - D(Q, R) = 0$$
$$H_1\colon D(P, R) - D(Q, R) > 0,$$

for some distance $D$.

- Statistic: $\hat{S}_n = \widehat{D}(P, R) - \widehat{D}(Q, R)$. Large, positive $\implies$ $Q$ is better.

# Motivations

**A common approach:**
Compare $\widehat{D}(P, R)$ and $\widehat{D}(Q, R)$ estimated from samples (e.g., FID).
If $\widehat{D}(Q, R) < \widehat{D}(P, R)$, conclude that $Q$ is better than $P$.

**Problems:**

1. Noisy decision. $\widehat{D}$ is random. $\rightarrow$ Statistical testing accounts for this.
2. Not interpretable. A scalar $\widehat{D}$ is not informative enough.

$Q$ = LSGAN [Mao et al., 2017] $P$ = GAN [Goodfellow et al., 2014]

- 1's from $Q$ are better. But 3's from $P$ are better.
- Our new interpretable test can output this information.

# Motivations

**A common approach:**
Compare $\widehat{D}(P, R)$ and $\widehat{D}(Q, R)$ estimated from samples (e.g., FID).
If $\widehat{D}(Q, R) < \widehat{D}(P, R)$, conclude that $Q$ is better than $P$.

**Problems:**

1. Noisy decision. $\widehat{D}$ is random. $\rightarrow$ Statistical testing accounts for this.
2. Not interpretable. A scalar $\widehat{D}$ is not informative enough.

$Q$ = LSGAN [Mao et al., 2017] $P$ = GAN [Goodfellow et al., 2014]

- 1's from $Q$ are better. But 3's from $P$ are better.
- Our new interpretable test can output this information.

# Motivations

**A common approach:**
Compare $\widehat{D}(P, R)$ and $\widehat{D}(Q, R)$ estimated from samples (e.g., FID).
If $\widehat{D}(Q, R) < \widehat{D}(P, R)$, conclude that $Q$ is better than $P$.

**Problems:**

1. Noisy decision. $\widehat{D}$ is random. $\rightarrow$ Statistical testing accounts for this.
2. Not interpretable. A scalar $\widehat{D}$ is not informative enough.

$Q$ = LSGAN [Mao et al., 2017] $P$ = GAN [Goodfellow et al., 2014]

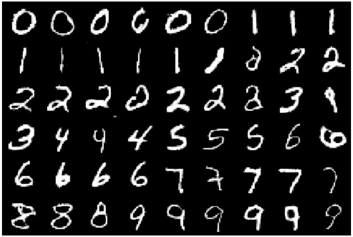
- 1's from $Q$ are better. But 3's from $P$ are better.
- Our new interpretable test can output this information.

# Motivations

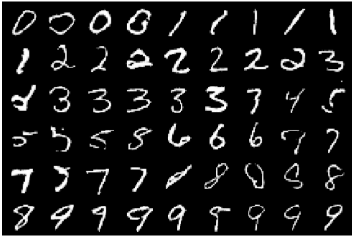
**A common approach:**
Compare $\widehat{D}(P, R)$ and $\widehat{D}(Q, R)$ estimated from samples (e.g., FID).
If $\widehat{D}(Q, R) < \widehat{D}(P, R)$, conclude that $Q$ is better than $P$.

**Problems:**

1. Noisy decision. $\widehat{D}$ is random. → Statistical testing accounts for this.
2. Not interpretable. A scalar $\widehat{D}$ is not informative enough.



$Q$ = LSGAN [Mao et al., 2017]



$P$ = GAN [Goodfellow et al., 2014]

- 1's from $Q$ are better. But 3's from $P$ are better.
- Our new interpretable test can output this information.

# Review: Hypothesis Testing

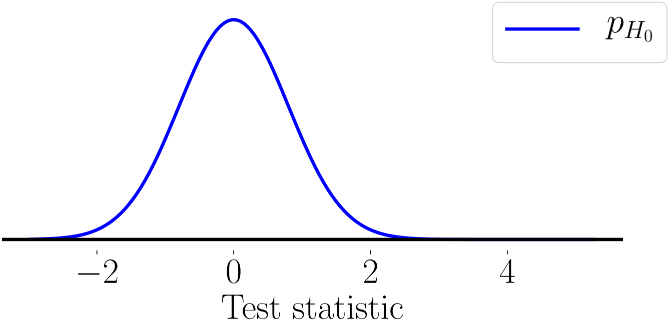
$$H_0 : D(P, R) - D(Q, R) = 0$$
$$H_1 : D(P, R) - D(Q, R) > 0.$$

Test statistic: $\hat{S}_n = \widehat{D}(P, R) - \widehat{D}(Q, R)$

# Review: Hypothesis Testing

$$H_0 : D(P, R) - D(Q, R) = 0$$
$$H_1 : D(P, R) - D(Q, R) > 0.$$

Test statistic: $\hat{S}_n = \widehat{D}(P, R) - \widehat{D}(Q, R)$

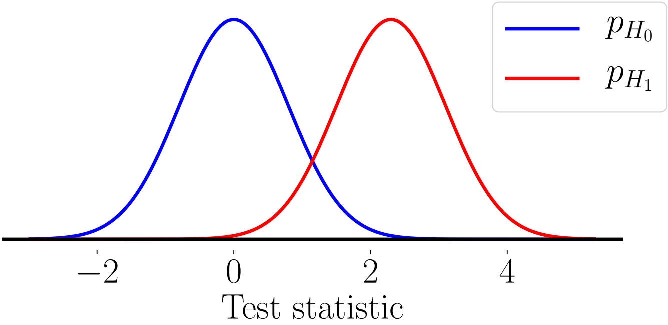



- Null distribution $p_{H_0}(\hat{S}_n)$ = distribution of $\hat{S}_n$ when $H_0$ is true.

# Review: Hypothesis Testing

$$H_0: D(P, R) - D(Q, R) = 0$$
$$H_1: D(P, R) - D(Q, R) > 0.$$

Test statistic: $\hat{S}_n = \widehat{D}(P, R) - \widehat{D}(Q, R)$

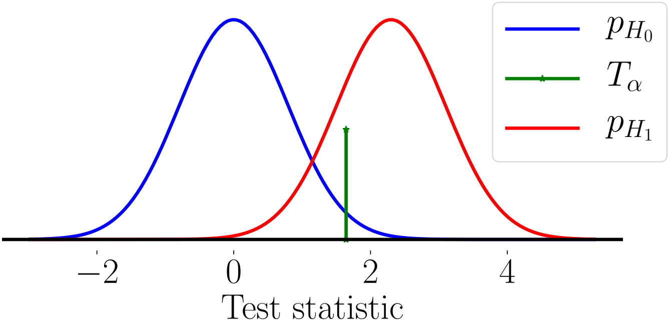



- Null distribution $p_{H_0}(\hat{S}_n)$ = distribution of $\hat{S}_n$ when $H_0$ is true.

# Review: Hypothesis Testing

$$H_0 : D(P, R) - D(Q, R) = 0$$
$$H_1 : D(P, R) - D(Q, R) > 0.$$

Test statistic: $\hat{S}_n = \widehat{D}(P, R) - \widehat{D}(Q, R)$

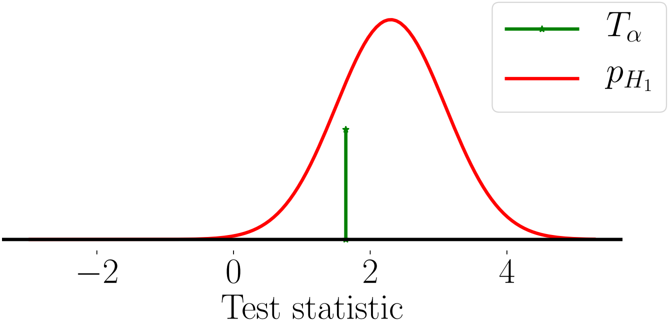



- Null distribution $p_{H_0}(\hat{S}_n)$ = distribution of $\hat{S}_n$ when $H_0$ is true.
- $T_\alpha = (1 - \alpha)$-quantile of $p_{H_0}$. Need to know $p_{H_0}$.

# Review: Hypothesis Testing

$$H_0: D(P, R) - D(Q, R) = 0$$
$$H_1: D(P, R) - D(Q, R) > 0.$$

Test statistic: $\hat{S}_n = \widehat{D}(P, R) - \widehat{D}(Q, R)$

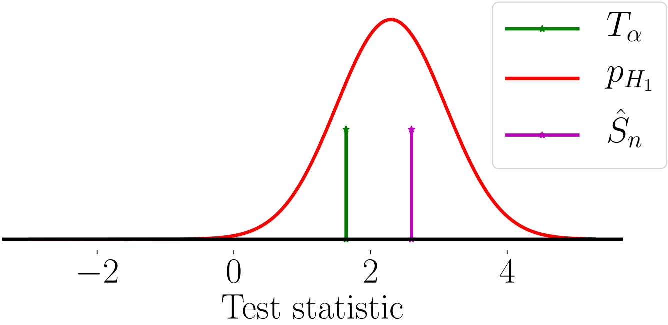



- Null distribution $p_{H_0}(\hat{S}_n) =$ distribution of $\hat{S}_n$ when $H_0$ is true.
- $T_\alpha = (1 - \alpha)$-quantile of $p_{H_0}$. Need to know $p_{H_0}$.

# Review: Hypothesis Testing

$$H_0: D(P, R) - D(Q, R) = 0$$
$$H_1: D(P, R) - D(Q, R) > 0.$$

Test statistic: $\hat{S}_n = \widehat{D}(P, R) - \widehat{D}(Q, R)$

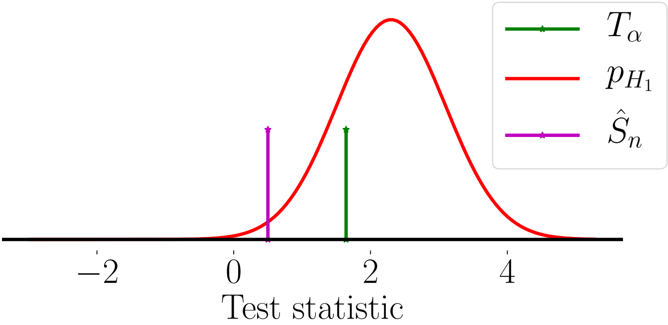



- Null distribution $p_{H_0}(\hat{S}_n)$ = distribution of $\hat{S}_n$ when $H_0$ is true.
- $T_\alpha = (1-\alpha)$-quantile of $p_{H_0}$. Need to know $p_{H_0}$.
- Test: Reject $H_0$ when $\hat{S}_n > T_\alpha$. False rejection rate of $H_0$ is $\alpha$.

# Review: Hypothesis Testing

$$H_0: D(P, R) - D(Q, R) = 0$$
$$H_1: D(P, R) - D(Q, R) > 0.$$

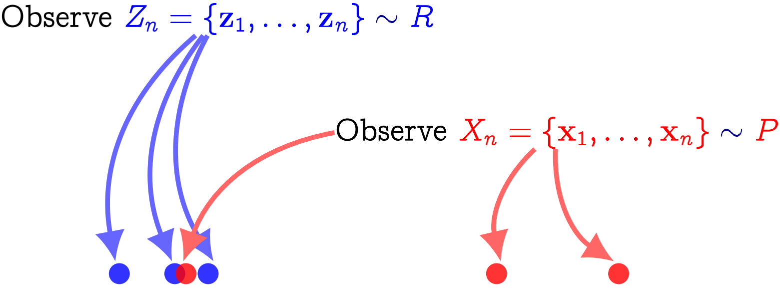
Test statistic: $\hat{S}_n = \widehat{D}(P, R) - \widehat{D}(Q, R)$



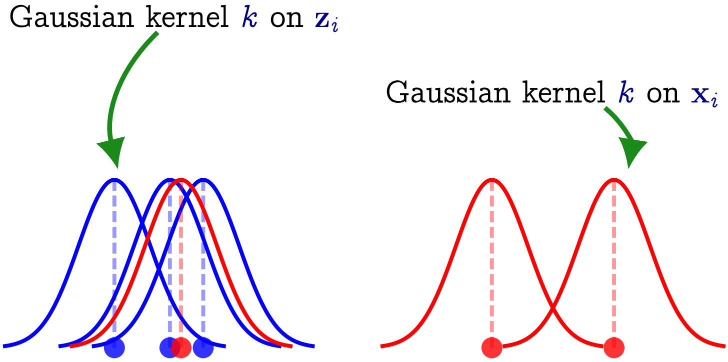

- Null distribution $p_{H_0}(\hat{S}_n)$ = distribution of $\hat{S}_n$ when $H_0$ is true.
- $T_\alpha = (1 - \alpha)$-quantile of $p_{H_0}$. Need to know $p_{H_0}$.
- Test: Reject $H_0$ when $\hat{S}_n > T_\alpha$. False rejection rate of $H_0$ is $\alpha$.

# The Witness Function (Gretton et al., 2012)

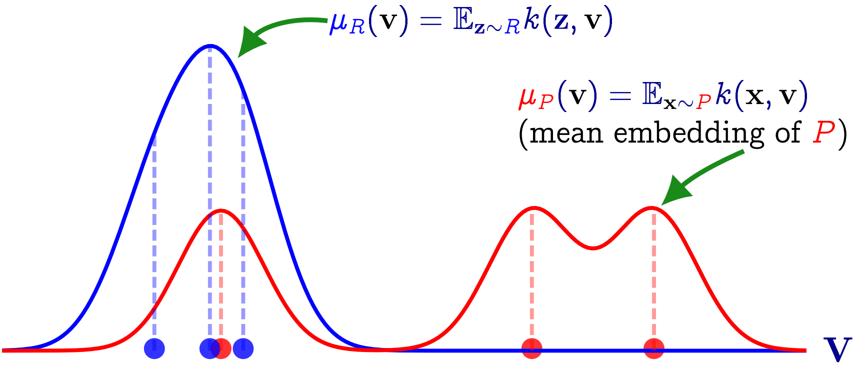
# The Witness Function (Gretton et al., 2012)



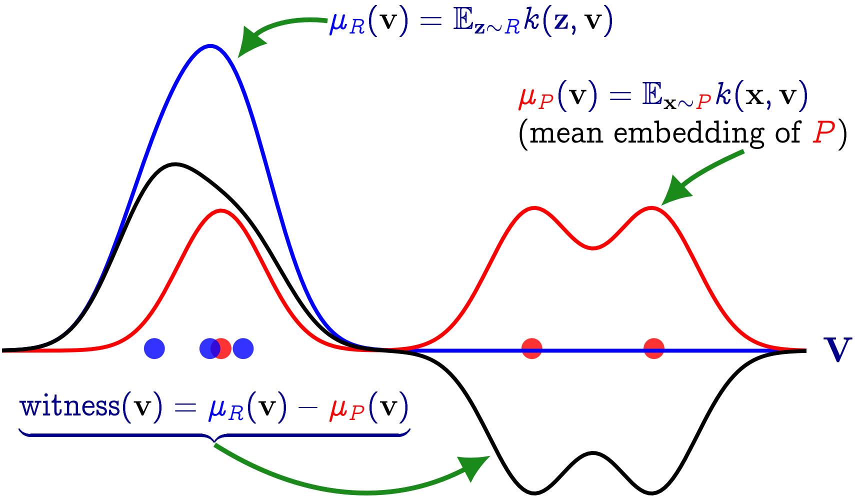
# The Witness Function (Gretton et al., 2012)



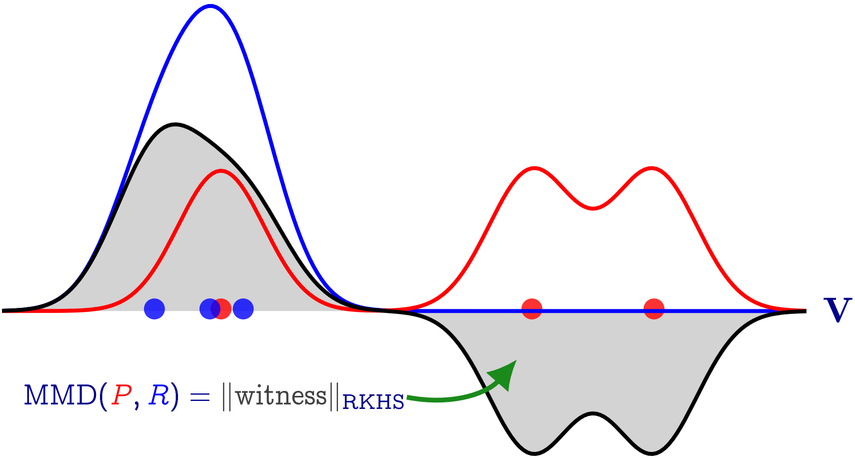
# The Witness Function (Gretton et al., 2012)



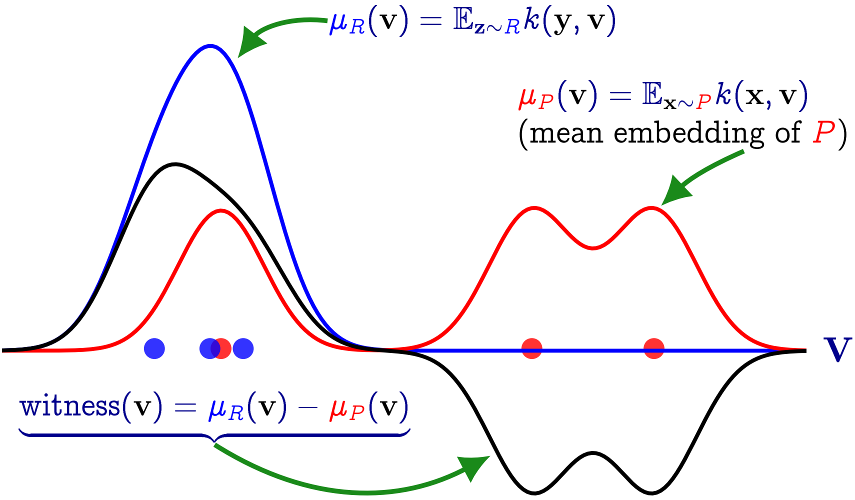
# The Witness Function (Gretton et al., 2012)



# The Witness Function (Gretton et al., 2012)

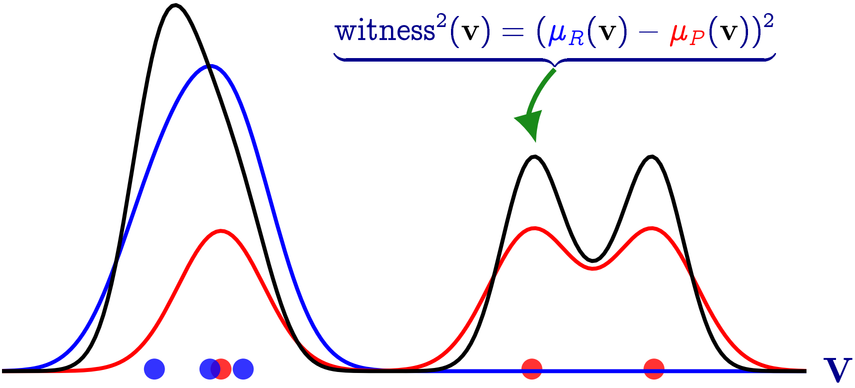


# The Unnormalized Mean Embeddings Statistic (Chwialkowski et al., 2015)

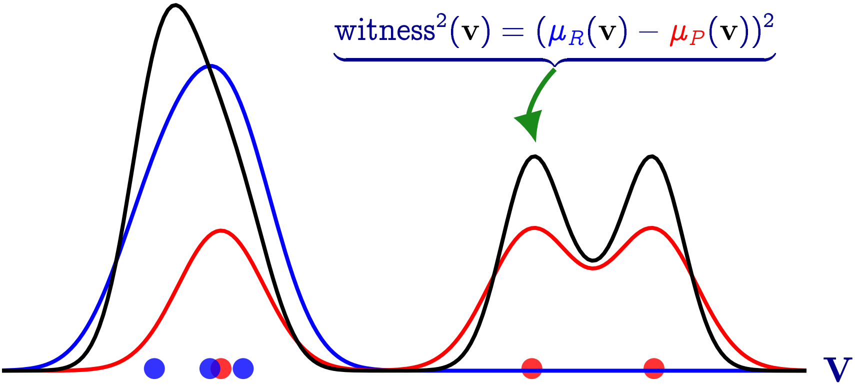


# The Unnormalized Mean Embeddings Statistic (Chwialkowski et al., 2015)

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



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




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

$$\text{UME}_V^2(P, R) = \frac{1}{J}\sum_{j=1}^{J} \text{witness}^2(\mathbf{v}_j) = U_P^2.$$

- $\text{UME}_V^2$ will be $D$ for model comparison.

# The Unnormalized Mean Embeddings (UME) Statistic

$$\mathrm{UME}^2_V(P,R) = U_P^2 = \frac{1}{J}\sum_{j=1}^{J}(\mu_P(\mathbf{v}_j) - \mu_R(\mathbf{v}_j))^2.$$

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

*Assume*

1. *Kernel $k$ is real analytic, integrable, and characteristic;*
2. *$V$ is drawn from $\eta$, a distribution with a density.*

*Then, for any $J > 0$, any $P$ and $R$,*

$$\boxed{\mathrm{UME}^2_V(P,R) = 0 \text{ iff } P = R,}$$

*$\eta$-almost surely.*

- **Key**: Evaluating witness[2] is enough to detect the difference (in theory).
- Runtime complexity: $\mathcal{O}(Jn)$. $J$ is small.

# The Unnormalized Mean Embeddings (UME) Statistic

$$\mathrm{UME}^2_V(P,R) = U_P^2 = \frac{1}{J}\sum_{j=1}^{J}(\mu_P(\mathbf{v}_j) - \mu_R(\mathbf{v}_j))^2.$$

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

*Assume*

1. *Kernel $k$ is real analytic, integrable, and characteristic;*
2. *$V$ is drawn from $\eta$, a distribution with a density.*

*Then, for any $J > 0$, any $P$ and $R$,*

$$\boxed{\mathrm{UME}^2_V(P,R) = 0 \text{ iff } P = R,}$$

*$\eta$-almost surely.*

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

# The Unnormalized Mean Embeddings (UME) Statistic

$$\mathrm{UME}^2_V(P,R) = U_P^2 = \frac{1}{J}\sum_{j=1}^{J}(\mu_P(\mathbf{v}_j) - \mu_R(\mathbf{v}_j))^2.$$

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

*Assume*

1. *Kernel $k$ is real analytic, integrable, and characteristic;*
2. *$V$ is drawn from $\eta$, a distribution with a density.*

*Then, for any $J > 0$, any $P$ and $R$,*

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

*$\eta$-almost surely.*

- Key: Evaluating witness[2] is enough to detect the difference (in theory).
- Runtime complexity: $\mathcal{O}(Jn)$. $J$ is small.

# The Unnormalized Mean Embeddings (UME) Statistic

$$\mathrm{UME}_V^2(P,R) = U_P^2 = \frac{1}{J}\sum_{j=1}^{J}(\mu_P(\mathbf{v}_j) - \mu_R(\mathbf{v}_j))^2.$$

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

*Assume*

1. *Kernel $k$ is real analytic, integrable, and characteristic;*
2. *$V$ is drawn from $\eta$, a distribution with a density.*

*Then, for any $J > 0$, any $P$ and $R$,*

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

*$\eta$-almost surely.*

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

# Asymptotic Distribution of $\widehat{\mathrm{UME}^2}_V(P,R) = \widehat{U_P^2}$

- Define $\psi_V(\mathbf{y}) := \frac{1}{\sqrt{J}}\left(k(\mathbf{y},\mathbf{v}_1),\ldots,k(\mathbf{y},\mathbf{v}_J)\right)^\top \in \mathbb{R}^J$.
- Let $\psi_V^P := \mathbb{E}_{\mathbf{x}\sim P}[\psi_V(\mathbf{x})] \in \mathbb{R}^J$. Let $C_V^P := \mathrm{cov}_{\mathbf{x}\sim P}[\psi_V(\mathbf{x})] \in \mathbb{R}^{J\times J}$.

**Proposition (Asymptotic distribution of $\widehat{U_P^2}$)**

*If $P \neq R$, for any $V$, as $n \to \infty$*

$$\sqrt{n}\left[\widehat{\mathrm{UME}_V^2}(P,R) - \mathrm{UME}_V^2(P,R)\right] \xrightarrow{d} \mathcal{N}(0, 4\zeta_P^2),$$

*where $\zeta_P^2 := (\psi_V^P - \psi_V^R)^\top (C_V^P + C_V^R)(\psi_V^P - \psi_V^R) > 0$.*

> **Main point:** When $P \neq R$, $\widehat{\mathrm{UME}_V^2}(P,R)$ is asymptotically normally distributed. Simple.

- But we will need the distribution of $\widehat{S}_n = \widehat{\mathrm{UME}_V^2}(P,R) - \widehat{\mathrm{UME}_V^2}(Q,R)$ which is ... ?

# Asymptotic Distribution of $\widehat{\mathrm{UME}^2}_V(P,R) = \widehat{U_P^2}$

- Define $\psi_V(\mathbf{y}) := \frac{1}{\sqrt{J}}\left(k(\mathbf{y},\mathbf{v}_1),\ldots,k(\mathbf{y},\mathbf{v}_J)\right)^\top \in \mathbb{R}^J$.
- Let $\psi_V^P := \mathbb{E}_{\mathbf{x}\sim P}[\psi_V(\mathbf{x})] \in \mathbb{R}^J$. Let $C_V^P := \mathrm{cov}_{\mathbf{x}\sim P}[\psi_V(\mathbf{x})] \in \mathbb{R}^{J\times J}$.

**Proposition (Asymptotic distribution of $\widehat{U_P^2}$)**

*If $P \neq R$, for any $V$, as $n \to \infty$*

$$\sqrt{n}\left[\widehat{\mathrm{UME}^2_V}(P,R) - \mathrm{UME}^2_V(P,R)\right] \xrightarrow{d} \mathcal{N}(0, 4\zeta_P^2),$$

*where* $\zeta_P^2 := (\psi_V^P - \psi_V^R)^\top (C_V^P + C_V^R)(\psi_V^P - \psi_V^R) > 0$.

**Main point:** When $P \neq R$, $\widehat{\mathrm{UME}^2_V}(P,R)$ is asymptotically normally distributed. Simple.

- But we will need the distribution of $\widehat{S}_n = \widehat{\mathrm{UME}^2_V}(P,R) - \widehat{\mathrm{UME}^2_V}(Q,R)$ which is … ?

# Asymptotic Distribution of $\widehat{\mathrm{UME}^2}_V(P,R) = \widehat{U_P^2}$

- Define $\psi_V(\mathbf{y}) := \frac{1}{\sqrt{J}}\left(k(\mathbf{y},\mathbf{v}_1),\ldots,k(\mathbf{y},\mathbf{v}_J)\right)^\top \in \mathbb{R}^J$.
- Let $\psi_V^P := \mathbb{E}_{\mathbf{x}\sim P}[\psi_V(\mathbf{x})] \in \mathbb{R}^J$. Let $C_V^P := \mathrm{cov}_{\mathbf{x}\sim P}[\psi_V(\mathbf{x})] \in \mathbb{R}^{J\times J}$.

**Proposition (Asymptotic distribution of $\widehat{U_P^2}$)**

*If $P \neq R$, for any $V$, as $n \to \infty$*

$$\sqrt{n}\left[\widehat{\mathrm{UME}^2_V}(P,R) - \mathrm{UME}^2_V(P,R)\right] \xrightarrow{d} \mathcal{N}(0, 4\zeta_P^2),$$

*where* $\zeta_P^2 := (\psi_V^P - \psi_V^R)^\top (C_V^P + C_V^R)(\psi_V^P - \psi_V^R) > 0$.

**Main point:** When $P \neq R$, $\widehat{\mathrm{UME}^2_V}(P,R)$ is asymptotically normally distributed. Simple.

- But we will need the distribution of $\widehat{S}_n = \widehat{\mathrm{UME}^2_V}(P,R) - \widehat{\mathrm{UME}^2_V}(Q,R)$ which is ... ?

# Asymptotic Distribution of $\widehat{\mathrm{UME}^2}_V(P,R) = \widehat{U_P^2}$

- Define $\psi_V(\mathbf{y}) := \frac{1}{\sqrt{J}}(k(\mathbf{y},\mathbf{v}_1),\ldots,k(\mathbf{y},\mathbf{v}_J))^\top \in \mathbb{R}^J$.
- Let $\psi_V^P := \mathbb{E}_{\mathbf{x}\sim P}[\psi_V(\mathbf{x})] \in \mathbb{R}^J$. Let $C_V^P := \mathrm{cov}_{\mathbf{x}\sim P}[\psi_V(\mathbf{x})] \in \mathbb{R}^{J\times J}$.

**Proposition (Asymptotic distribution of $\widehat{U_P^2}$)**

*If $P \neq R$, for any $V$, as $n \to \infty$*

$$\sqrt{n}\left[\widehat{\mathrm{UME}^2_V}(P,R) - \mathrm{UME}^2_V(P,R)\right] \xrightarrow{d} \mathcal{N}(0, 4\zeta_P^2),$$

*where* $\zeta_P^2 := (\psi_V^P - \psi_V^R)^\top (C_V^P + C_V^R)(\psi_V^P - \psi_V^R) > 0.$

**Main point:** When $P \neq R$, $\widehat{\mathrm{UME}^2_V}(P,R)$ is asymptotically normally distributed. Simple.

- But we will need the distribution of $\widehat{S}_n = \widehat{\mathrm{UME}^2_V}(P,R) - \widehat{\mathrm{UME}^2_V}(Q,R)$ which is ... ?

# Asymptotic Distribution of $\widehat{\mathrm{UME}^2}_V(P,R) = \widehat{U_P^2}$

- Define $\psi_V(\mathbf{y}) := \frac{1}{\sqrt{J}}(k(\mathbf{y},\mathbf{v}_1),\ldots,k(\mathbf{y},\mathbf{v}_J))^\top \in \mathbb{R}^J$.
- Let $\psi_V^P := \mathbb{E}_{\mathbf{x}\sim P}[\psi_V(\mathbf{x})] \in \mathbb{R}^J$. Let $C_V^P := \mathrm{cov}_{\mathbf{x}\sim P}[\psi_V(\mathbf{x})] \in \mathbb{R}^{J\times J}$.

**Proposition (Asymptotic distribution of $\widehat{U_P^2}$)**

*If $P \neq R$, for any $V$, as $n \to \infty$*

$$\sqrt{n}\left[\widehat{\mathrm{UME}^2_V}(P,R) - \mathrm{UME}^2_V(P,R)\right] \xrightarrow{d} \mathcal{N}(0, 4\zeta_P^2),$$

*where* $\zeta_P^2 := (\psi_V^P - \psi_V^R)^\top (C_V^P + C_V^R)(\psi_V^P - \psi_V^R) > 0.$

**Main point:** When $P \neq R$, $\widehat{\mathrm{UME}^2_V}(P,R)$ is asymptotically normally distributed. Simple.

- But we will need the distribution of $\widehat{S}_n = \widehat{\mathrm{UME}^2_V}(P,R) - \widehat{\mathrm{UME}^2_V}(Q,R)$ which is ... ?

# $\widehat{\mathrm{UME}^2_V}(P, R)$ and $\widehat{\mathrm{UME}^2_V}(Q, R)$ are Correlated

- Write $U_P^2 = \mathrm{UME}^2(P, R)$ and $U_Q^2 = \mathrm{UME}^2(Q, R)$.
- Let $S := U_P^2 - U_Q^2$. So $H_0 : S = 0$ and $H_1 : S > 0$.

**Proposition (Joint distribution of $\widehat{U_P^2}$ and $\widehat{U_Q^2}$)**

*Assume that $P, Q$ and $R$ are all distinct. Under mild conditions, for any $V$,*

1. $\sqrt{n}\left(\begin{pmatrix} \widehat{U_P^2} \\ \widehat{U_Q^2} \end{pmatrix} - \begin{pmatrix} U_P^2 \\ U_Q^2 \end{pmatrix}\right) \xrightarrow{d} \mathcal{N}\left(\mathbf{0}, 4\begin{pmatrix} \zeta_P^2 & \zeta_{PQ} \\ \zeta_{PQ} & \zeta_Q^2 \end{pmatrix}\right)$.
2. $\sqrt{n}\left(\widehat{S}_n - S\right) \xrightarrow{d} \mathcal{N}\left(0, 4(\zeta_P^2 - 2\zeta_{PQ} + \zeta_Q^2)\right)$.

So, asymptotic null distribution is normal. Easy to get $T_\alpha$.

- [1] $\rightarrow$ use theory of multivariate U-statistics
- [2] $\rightarrow$ continuous mapping theorem. Follows from [1].

## $\widehat{\mathrm{UME}^2_V}(P,R)$ and $\widehat{\mathrm{UME}^2_V}(Q,R)$ are Correlated

- Write $U_P^2 = \mathrm{UME}^2(P,R)$ and $U_Q^2 = \mathrm{UME}^2(Q,R)$.
- Let $S := U_P^2 - U_Q^2$. So $H_0 : S = 0$ and $H_1 : S > 0$.

**Proposition (Joint distribution of $\widehat{U_P^2}$ and $\widehat{U_Q^2}$)**

*Assume that $P, Q$ and $R$ are all distinct. Under mild conditions, for any $V$,*

1. $\sqrt{n}\left(\begin{pmatrix}\widehat{U_P^2}\\ \widehat{U_Q^2}\end{pmatrix} - \begin{pmatrix}U_P^2\\ U_Q^2\end{pmatrix}\right) \xrightarrow{d} \mathcal{N}\left(\mathbf{0}, 4\begin{pmatrix}\zeta_P^2 & \zeta_{PQ}\\ \zeta_{PQ} & \zeta_Q^2\end{pmatrix}\right)$.
2. $\sqrt{n}\left(\widehat{S}_n - S\right) \xrightarrow{d} \mathcal{N}\left(0, 4(\zeta_P^2 - 2\zeta_{PQ} + \zeta_Q^2)\right)$.

So, asymptotic null distribution is normal. Easy to get $T_\alpha$.

- [1] $\rightarrow$ use theory of multivariate U-statistics
- [2] $\rightarrow$ continuous mapping theorem. Follows from [1].

# $\widehat{\mathrm{UME}}^2_V(P,R)$ and $\widehat{\mathrm{UME}}^2_V(Q,R)$ are Correlated

- Write $U_P^2 = \mathrm{UME}^2(P,R)$ and $U_Q^2 = \mathrm{UME}^2(Q,R)$.
- Let $S := U_P^2 - U_Q^2$. So $H_0 : S = 0$ and $H_1 : S > 0$.

**Proposition (Joint distribution of $\widehat{U_P^2}$ and $\widehat{U_Q^2}$)**

*Assume that $P, Q$ and $R$ are all distinct. Under mild conditions, for any $V$,*

1. $$\sqrt{n}\left(\begin{pmatrix}\widehat{U_P^2}\\ \widehat{U_Q^2}\end{pmatrix} - \begin{pmatrix}U_P^2\\ U_Q^2\end{pmatrix}\right) \xrightarrow{d} \mathcal{N}\left(\mathbf{0}, 4\begin{pmatrix}\zeta_P^2 & \zeta_{PQ}\\ \zeta_{PQ} & \zeta_Q^2\end{pmatrix}\right).$$
2. $$\sqrt{n}\left(\widehat{S}_n - S\right) \xrightarrow{d} \mathcal{N}\left(0, 4(\zeta_P^2 - 2\zeta_{PQ} + \zeta_Q^2)\right).$$

So, asymptotic null distribution is normal. Easy to get $T_\alpha$.

- [1] $\rightarrow$ use theory of multivariate U-statistics
- [2] $\rightarrow$ continuous mapping theorem. Follows from [1].

# $\widehat{\mathrm{UME}^2_V}(P,R)$ and $\widehat{\mathrm{UME}^2_V}(Q,R)$ are Correlated

- Write $U_P^2 = \mathrm{UME}^2(P,R)$ and $U_Q^2 = \mathrm{UME}^2(Q,R)$.
- Let $S := U_P^2 - U_Q^2$. So $H_0 : S = 0$ and $H_1 : S > 0$.

**Proposition (Joint distribution of $\widehat{U_P^2}$ and $\widehat{U_Q^2}$)**

*Assume that $P, Q$ and $R$ are all distinct. Under mild conditions, for any $V$,*

1. $\sqrt{n}\left(\begin{pmatrix}\widehat{U_P^2}\\ \widehat{U_Q^2}\end{pmatrix} - \begin{pmatrix}U_P^2\\ U_Q^2\end{pmatrix}\right) \xrightarrow{d} \mathcal{N}\left(\mathbf{0}, 4\begin{pmatrix}\zeta_P^2 & \zeta_{PQ}\\ \zeta_{PQ} & \zeta_Q^2\end{pmatrix}\right).$
2. $\sqrt{n}\left(\widehat{S}_n - S\right) \xrightarrow{d} \mathcal{N}\left(0, 4(\zeta_P^2 - 2\zeta_{PQ} + \zeta_Q^2)\right).$

So, asymptotic null distribution is normal. Easy to get $T_\alpha$.

- [1] $\rightarrow$ use theory of multivariate U-statistics
- [2] $\rightarrow$ continuous mapping theorem. Follows from [1].

## $\widehat{\mathrm{UME}}^2_V(P,R)$ and $\widehat{\mathrm{UME}}^2_V(Q,R)$ are Correlated

- Write $U_P^2 = \mathrm{UME}^2(P,R)$ and $U_Q^2 = \mathrm{UME}^2(Q,R)$.
- Let $S := U_P^2 - U_Q^2$. So $H_0 : S = 0$ and $H_1 : S > 0$.

**Proposition (Joint distribution of $\widehat{U_P^2}$ and $\widehat{U_Q^2}$)**

*Assume that $P, Q$ and $R$ are all distinct. Under mild conditions, for any $V$,*

1. $$\sqrt{n}\left(\begin{pmatrix}\widehat{U_P^2}\\ \widehat{U_Q^2}\end{pmatrix} - \begin{pmatrix}U_P^2\\ U_Q^2\end{pmatrix}\right) \xrightarrow{d} \mathcal{N}\left(\mathbf{0}, 4\begin{pmatrix}\zeta_P^2 & \zeta_{PQ}\\ \zeta_{PQ} & \zeta_Q^2\end{pmatrix}\right).$$
2. $$\sqrt{n}\left(\widehat{S}_n - S\right) \xrightarrow{d} \mathcal{N}\left(0, 4(\zeta_P^2 - 2\zeta_{PQ} + \zeta_Q^2)\right).$$

So, asymptotic null distribution is normal. Easy to get $T_\alpha$.

- [1] $\rightarrow$ use theory of multivariate U-statistics
- [2] $\rightarrow$ continuous mapping theorem. Follows from [1].

# Choose Test Locations $V = \{\mathbf{v}_j\}_{j=1}^J$ in Practice

- Pick $V$ so as to maximize the test power.
- $H_0 : U_P^2 - U_Q^2 = 0$ vs. $H_1 : U_P^2 - U_Q^2 > 0$ (i.e., $Q$ is better).

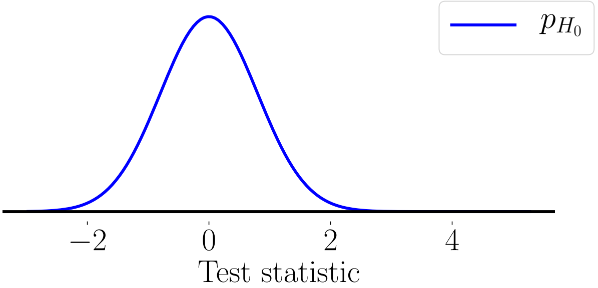
---

# Choose Test Locations $V = \{\mathbf{v}_j\}_{j=1}^J$ in Practice

- Pick $V$ so as to maximize the test power.
- $H_0 : U_P^2 - U_Q^2 = 0$ vs. $H_1 : U_P^2 - U_Q^2 > 0$ (i.e., $Q$ is better).

---

Under $H_0 : U_P^2 - U_Q^2 = 0$,

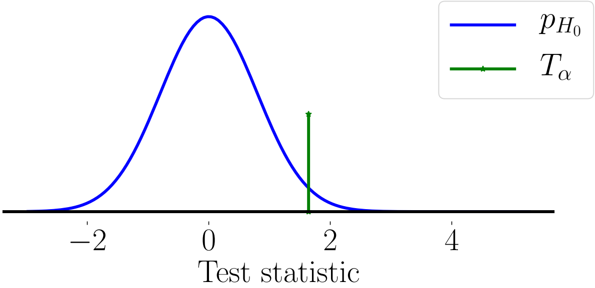


# Choose Test Locations $V = \{\mathbf{v}_j\}_{j=1}^J$ in Practice

- Pick $V$ so as to maximize the test power.
- $H_0 : U_P^2 - U_Q^2 = 0$ vs. $H_1 : U_P^2 - U_Q^2 > 0$ (i.e., $Q$ is better).

---

Under $H_0 : U_P^2 - U_Q^2 = 0$,

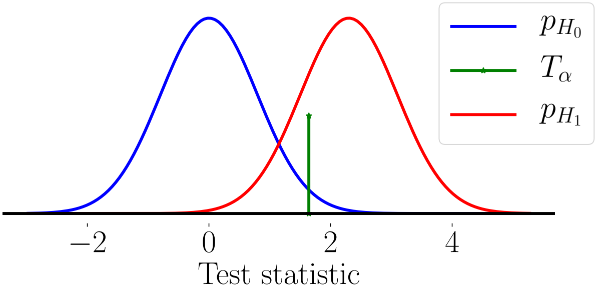


# Choose Test Locations $V = \{\mathbf{v}_j\}_{j=1}^J$ in Practice

- Pick $V$ so as to maximize the test power.
- $H_0 : U_P^2 - U_Q^2 = 0$ vs. $H_1 : U_P^2 - U_Q^2 > 0$ (i.e., $Q$ is better).

---

Under $H_1 : U_P^2 - U_Q^2 > 0$,

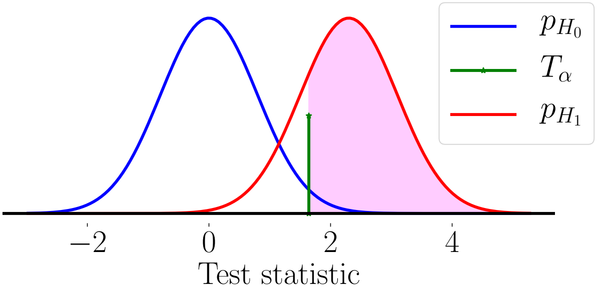


# Choose Test Locations $V = \{\mathbf{v}_j\}_{j=1}^J$ in Practice

- Pick $V$ so as to maximize the test power.
- $H_0 : U_P^2 - U_Q^2 = 0$ vs. $H_1 : U_P^2 - U_Q^2 > 0$ (i.e., $Q$ is better).

---

Test power $= \mathbb{P}(\text{reject } H_0 \mid H_1 \text{ true}) = \mathbb{P}(\text{Decide } Q \text{ better} \mid Q \text{ better})$

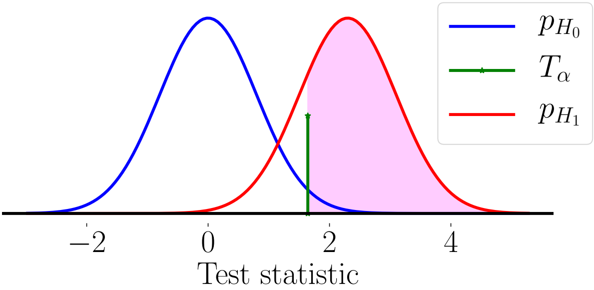


# Choose Test Locations $V = \{\mathbf{v}_j\}_{j=1}^{J}$ in Practice

- Pick $V$ so as to maximize the test power .
- $H_0 : U_P^2 - U_Q^2 = 0$ vs. $H_1 : U_P^2 - U_Q^2 > 0$ (i.e., $Q$ is better).

---

Test power $= \mathbb{P}(\text{reject } H_0 \mid H_1 \text{ true}) = \mathbb{P}(\text{Decide } Q \text{ better} \mid Q \text{ better})$



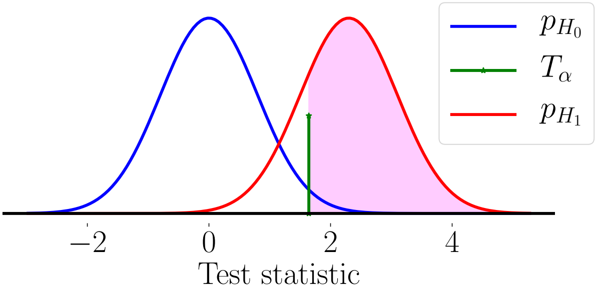

- Split the data into tr and te. Optimize $V$ on tr. Test on te.
-

# Choose Test Locations $V = \{\mathbf{v}_j\}_{j=1}^J$ in Practice

- Pick $V$ so as to maximize the test power.
- $H_0 : U_P^2 - U_Q^2 = 0$ vs. $H_1 : U_P^2 - U_Q^2 > 0$ (i.e., $Q$ is better).

---

Test power $= \mathbb{P}(\text{reject } H_0 \mid H_1 \text{ true}) = \mathbb{P}(\text{Decide } Q \text{ better} \mid Q \text{ better})$



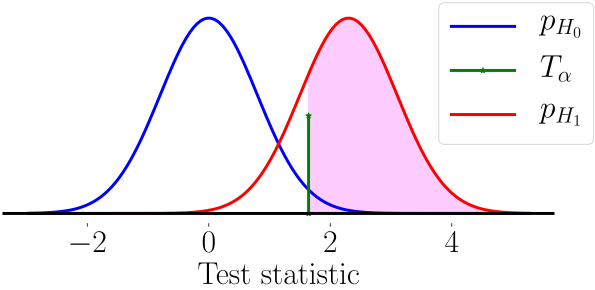

- Split the data into tr and te. Optimize $V$ on tr. Test on te.
- Optimized $V$ show where $Q$ is better than $P$.

# Choose Test Locations $V = \{\mathbf{v}_j\}_{j=1}^J$ in Practice

- Pick $V$ so as to maximize the test power .
- $H_0 : U_P^2 - U_Q^2 = 0$ vs. $H_1 : U_P^2 - U_Q^2 > 0$ (i.e., $Q$ is better).

---

Test power $= \mathbb{P}(\text{reject } H_0 \mid H_1 \text{ true}) = \mathbb{P}(\text{Decide } Q \text{ better} \mid Q \text{ better})$

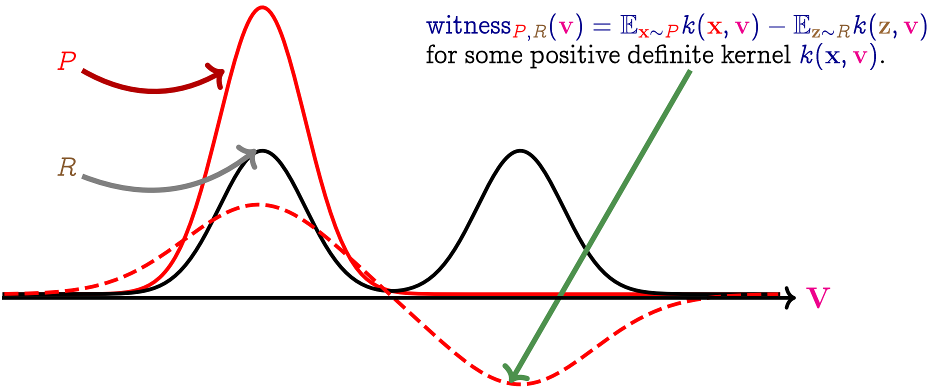



- Split the data into tr and te. Optimize $V$ on tr. Test on te.
- Optimized $V$ show where $Q$ is better than $P$.
- For large $n$, $\arg\max_V \text{power} = \arg\max_V f(V)$ where $f = \frac{\text{mean of } p_{H_1}}{\text{std of } p_{H_1}}$.

  Call $f$ the power criterion .

# Rel-UME: Difference of Two Witness Functions

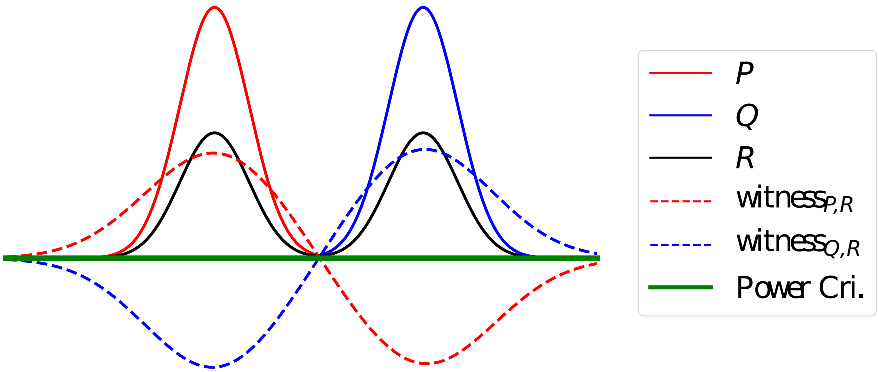
Recall the witness function between $P$ and $R$:



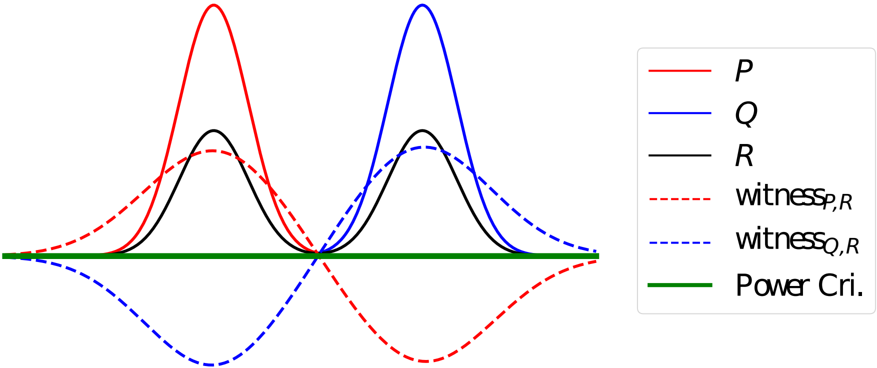

Assume only one test location $\mathbf{v}$. Recall

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

# Rel-UME: Difference of Two Witness Functions



# Rel-UME: Difference of Two Witness Functions

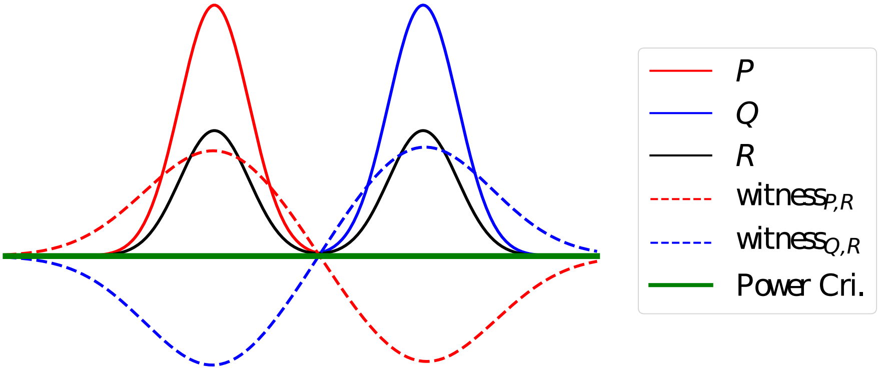



- Power criterion$(\mathbf{v}) = f(\mathbf{v})$ is a function such that maximizing it corresponds to maximizing the test power.

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

- $f(\mathbf{v}) > 0 \implies Q$ 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

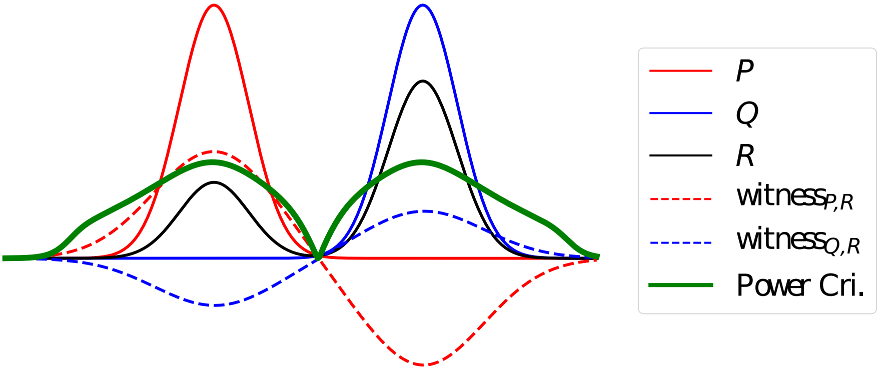



- Power criterion$(\mathbf{v}) = f(\mathbf{v})$ is a function such that maximizing it corresponds to maximizing the test power.

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

- $f(\mathbf{v}) > 0 \implies Q$ 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

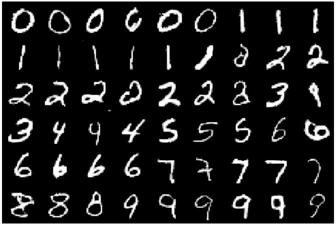



- Power criterion($\mathbf{v}$) $= f(\mathbf{v})$ is a function such that maximizing it corresponds to maximizing the test power.

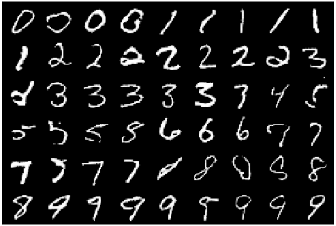
$$f(\mathbf{v}) = \frac{\text{witness}^2_{P,R}(\mathbf{v}) - \text{witness}^2_{Q,R}(\mathbf{v})}{\text{standard deviation}_{P,Q,R}(\mathbf{v})} = \frac{U_P^2 - U_Q^2}{\sqrt{4(\zeta_P^2 - 2\zeta_{PQ} + \zeta_Q^2)}}$$

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

# Where Does Each GAN Do Better?



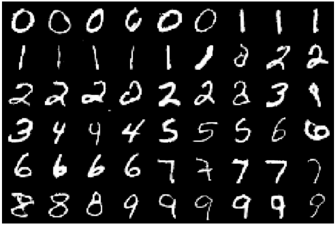
$Q$ = LSGAN [Mao et al., 2017]



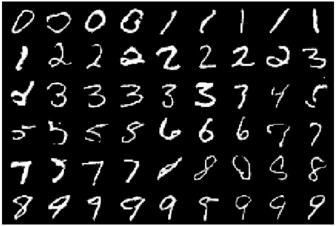
$P$ = GAN [Goodfellow et al., 2014]

- Set $V$ = 40 (real) images of digit $i = 0, \ldots, 9$.
- Evaluate power criterion with $n = 2000$.
- $Q$ 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]

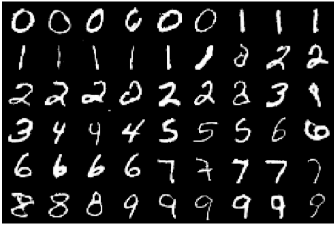


$P$ = GAN

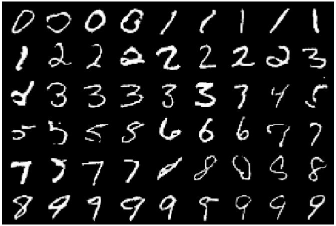
[Goodfellow et al., 2014]

- Set $V$ = 40 (real) images of digit $i = 0, \ldots, 9$.
- Evaluate power criterion with $n = 2000$.
- $Q$ is better at "1" and "5". $P$ is slightly better at "3". **Interpretable.**

# Where Does Each GAN Do Better?

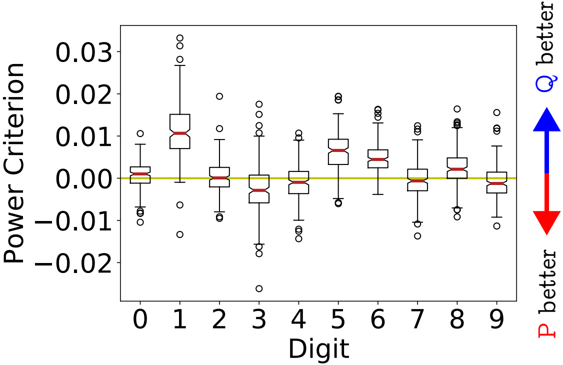


$Q$ = LSGAN [Mao et al., 2017]



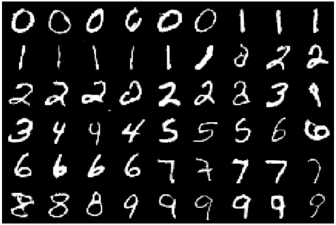
$P$ = GAN

[Goodfellow et al., 2014]

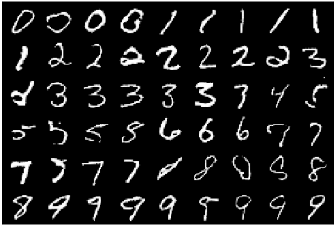



- Set $V$ = 40 (real) images of digit $i = 0, \ldots, 9$.
- Evaluate power criterion with $n = 2000$.
- $Q$ is better at "1" and "5". $P$ is slightly better at "3". **Interpretable.**

# Where Does Each GAN Do Better?

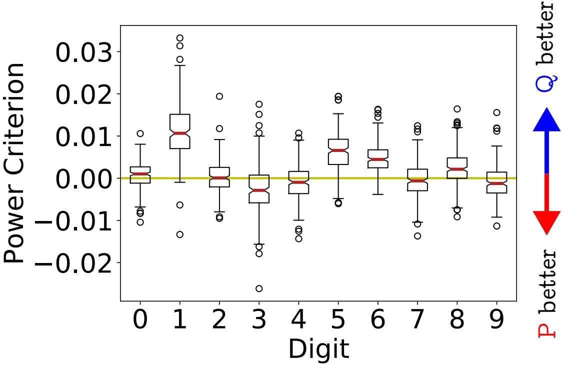


$Q$ = LSGAN [Mao et al., 2017]



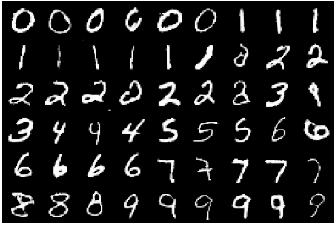
$P$ = GAN

[Goodfellow et al., 2014]

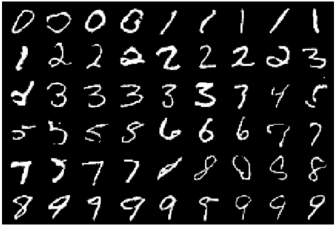



- Set $V$ = 40 (real) images of digit $i = 0, \ldots, 9$.
- Evaluate power criterion with $n = 2000$.
- $Q$ is better at "1" and "5". $P$ is slightly better at "3". **Interpretable.**

# Where Does Each GAN Do Better?

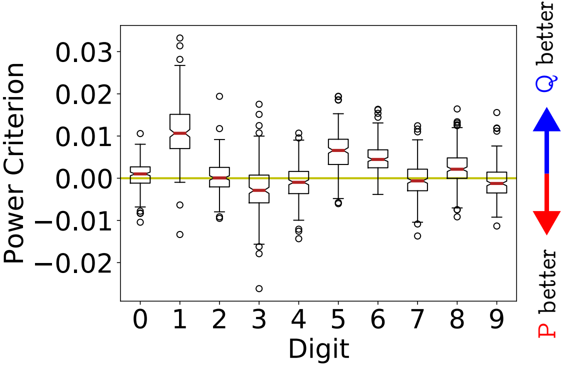


$Q$ = LSGAN [Mao et al., 2017]



$P$ = GAN

[Goodfellow et al., 2014]

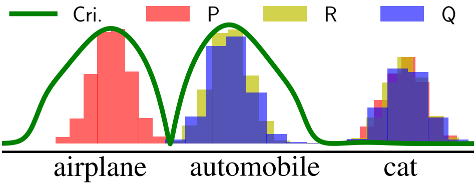



- Set $V$ = 40 (real) images of digit $i = 0, \ldots, 9$.
- Evaluate power criterion with $n = 2000$.
- $Q$ is better at "1" and "5". $P$ is slightly better at "3". **Interpretable.**

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

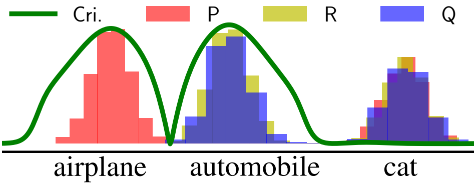
# Experiment on CIFAR10

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




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

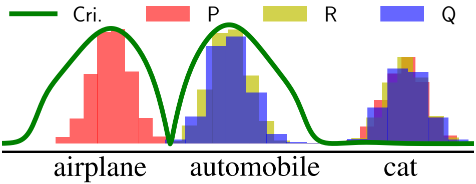
# Experiment on CIFAR10



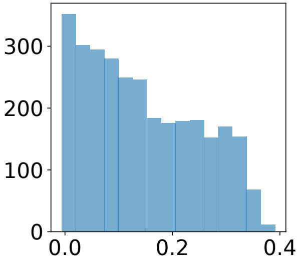

- $P = \{\text{airplane, cat}\}$,
  $Q = \{\text{automobile, cat}\}$
- (true) $R = \{\text{automobile, cat}\}$
- Gaussian kernel on 2048 features extracted by the Inception-v3 network at the pool3 layer.

# 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.

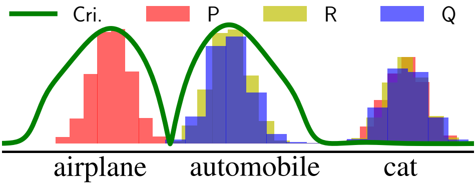



Histogram of power criterion values $f(\mathbf{v})$ evaluated at $\mathbf{v} = \{\text{airplane, automobile, cat}\}$.

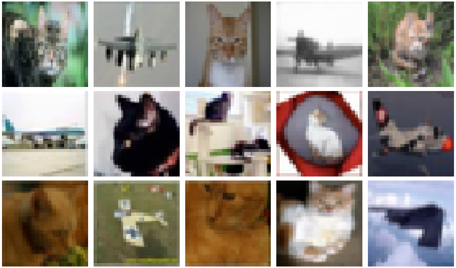
- All non-negative. $\Longrightarrow$ $Q$ is equally good or better than $P$ everywhere.

# Experiment on CIFAR10

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



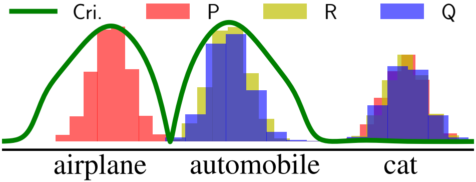

- Gaussian kernel on 2048 features extracted by the Inception-v3 network at the pool3 layer.



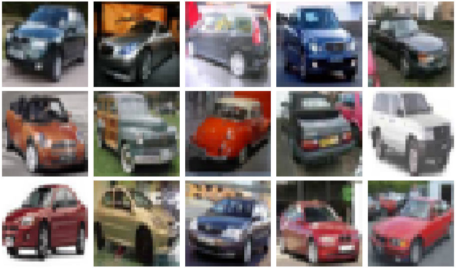
Images $\mathbf{v}$ with the lowest values of $f(\mathbf{v}) \approx 0$. $\implies$ $P, Q$ perform equally well in these regions.

# Experiment on CIFAR10

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



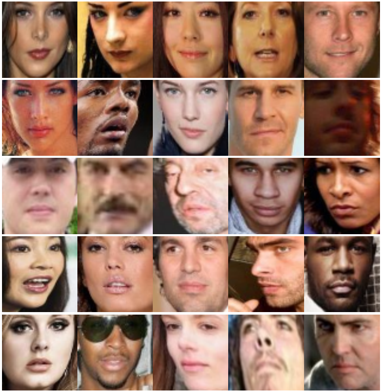

- Gaussian kernel on 2048 features extracted by the Inception-v3 network at the pool3 layer.



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

# Experiment on CelebA

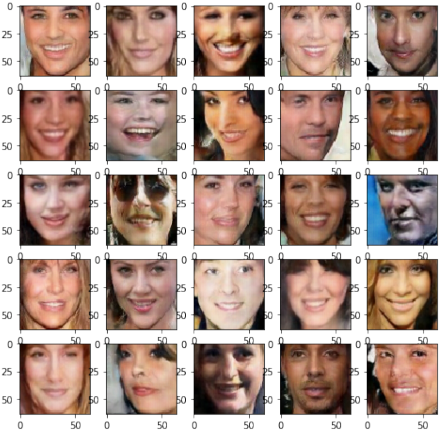
Real smiling faces (RS)



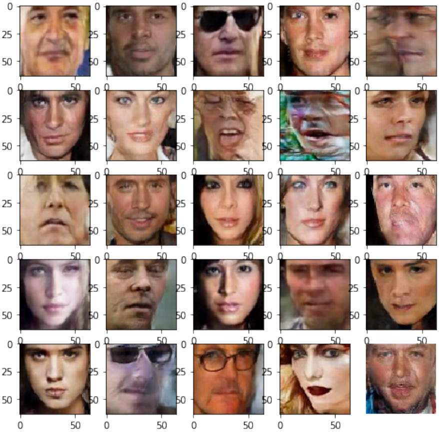
Real non-smiling faces (NS)

- Two datasets for training two models.
- Center-cropped CelebA images to $64 \times 64$ pixels.

# Experiment on CelebA



Model for smiling faces (S)



Model for non-smiling faces (N)

- Trained with DCGAN. Get two models.

# Experiment on CelebA

- Report avg rejection rate (e.g., rate of claiming $Q$ is better).
- Fréchet Inception Distance (FID) (Heusel et al., 2017). Not a test. If $\mathrm{FID}(P, R) > \mathrm{FID}(Q, R)$, claim $Q$ is better.
- RS = real smiling images. RN = real non-smiling images.
- RM = mixture of RS and RN

| Case | $P$ | $Q$ | $R$ | Truth | Rel-UME J10 | Rel-UME J40 | Rel-MMD | FID | FID diff. |
|---|---|---|---|---|---|---|---|---|---|
| 1. | S | S | RS | Not rej | 0.0 | 0.0 | 0.0 | 0.53 | -0.045 $\pm$ 0.52 |
| 2. | RS | RS | RS | Not rej | 0.0 | 0.0 | 0.03 | 0.7 | 0.04 $\pm$ 0.19 |
| 3. | S | N | RN | Rej | 0.57 | 1.0 | 1.0 | 1.0 | 5.25 $\pm$ 0.75 |
| 4. | S | N | RM | Not rej | 0.0 | 0.0 | 0.0 | 0.0 | -4.55 $\pm$ 0.82 |

- FID claims $Q$ is better when the two models are equally good. Not account for uncertainty.
- All have high test power when $Q$ is indeed better.

# Experiment on CelebA

- Report avg rejection rate (e.g., rate of claiming $Q$ is better).
- Fréchet Inception Distance (FID) (Heusel et al., 2017). Not a test. If $\mathrm{FID}(P, R) > \mathrm{FID}(Q, R)$, claim $Q$ is better.
- **RS** = real smiling images. **RN** = real non-smiling images.
- **RM** = mixture of RS and RN

| Case | $P$ | $Q$ | $R$ | Truth | Rel-UME J10 | Rel-UME J40 | Rel-MMD | FID | FID diff. |
|---|---|---|---|---|---|---|---|---|---|
| 1. | S | S | RS | Not rej | 0.0 | 0.0 | 0.0 | 0.53 | -0.045 $\pm$ 0.52 |
| 2. | RS | RS | RS | Not rej | 0.0 | 0.0 | 0.03 | 0.7 | 0.04 $\pm$ 0.19 |
| 3. | S | N | RN | Rej | 0.57 | 1.0 | 1.0 | 1.0 | 5.25 $\pm$ 0.75 |
| 4. | S | N | RM | Not rej | 0.0 | 0.0 | 0.0 | 0.0 | -4.55 $\pm$ 0.82 |

- FID claims $Q$ is better when the two models are equally good. Not account for uncertainty.
- All have high test power when $Q$ is indeed better.

# Experiment on CelebA

- Report avg rejection rate (e.g., rate of claiming $Q$ is better).
- Fréchet Inception Distance (FID) (Heusel et al., 2017). Not a test. If $\mathrm{FID}(P, R) > \mathrm{FID}(Q, R)$, claim $Q$ is better.
- **RS** = real smiling images. **RN** = real non-smiling images.
- **RM** = mixture of RS and RN

| Case | $P$ | $Q$ | $R$ | Truth | Rel-UME J10 | Rel-UME J40 | Rel-MMD | FID | FID diff. |
|---|---|---|---|---|---|---|---|---|---|
| 1. | S | S | RS | Not rej | 0.0 | 0.0 | 0.0 | 0.53 | -0.045 $\pm$ 0.52 |
| 2. | RS | RS | RS | Not rej | 0.0 | 0.0 | 0.03 | 0.7 | 0.04 $\pm$ 0.19 |
| 3. | S | N | RN | Rej | 0.57 | 1.0 | 1.0 | 1.0 | 5.25 $\pm$ 0.75 |
| 4. | S | N | RM | Not rej | 0.0 | 0.0 | 0.0 | 0.0 | -4.55 $\pm$ 0.82 |

- FID claims $Q$ is better when the two models are equally good. Not account for uncertainty.
- All have high test power when $Q$ is indeed better.

# Experiment on CelebA

- Report avg rejection rate (e.g., rate of claiming $Q$ is better).
- Fréchet Inception Distance (FID) (Heusel et al., 2017). Not a test. If $\text{FID}(P, R) > \text{FID}(Q, R)$, claim $Q$ is better.
- **RS** = real smiling images. **RN** = real non-smiling images.
- **RM** = mixture of RS and RN

| Case | $P$ | $Q$ | $R$ | Truth | Rel-UME J10 | Rel-UME J40 | Rel-MMD | FID | FID diff. |
|---|---|---|---|---|---|---|---|---|---|
| 1. | S | S | RS | Not rej | 0.0 | 0.0 | 0.0 | 0.53 | -0.045 $\pm$ 0.52 |
| 2. | RS | RS | RS | Not rej | 0.0 | 0.0 | 0.03 | 0.7 | 0.04 $\pm$ 0.19 |
| 3. | S | N | RN | Rej | 0.57 | 1.0 | 1.0 | 1.0 | 5.25 $\pm$ 0.75 |
| 4. | S | N | RM | Not rej | 0.0 | 0.0 | 0.0 | 0.0 | -4.55 $\pm$ 0.82 |

- FID claims $Q$ is better when the two models are equally good. Not account for uncertainty.
- All have high test power when $Q$ is indeed better.

# Experiment on CelebA

- Report avg rejection rate (e.g., rate of claiming $Q$ is better).
- Fréchet Inception Distance (FID) (Heusel et al., 2017). Not a test. If $\mathrm{FID}(P, R) > \mathrm{FID}(Q, R)$, claim $Q$ is better.
- **RS** = real smiling images. **RN** = real non-smiling images.
- **RM** = mixture of RS and RN

| Case | $P$ | $Q$ | $R$ | Truth | Rel-UME J10 | Rel-UME J40 | Rel-MMD | FID | FID diff. |
|---|---|---|---|---|---|---|---|---|---|
| 1. | S | S | RS | Not rej | 0.0 | 0.0 | 0.0 | 0.53 | -0.045 $\pm$ 0.52 |
| 2. | RS | RS | RS | Not rej | 0.0 | 0.0 | 0.03 | 0.7 | 0.04 $\pm$ 0.19 |
| 3. | S | N | RN | Rej | 0.57 | 1.0 | 1.0 | 1.0 | 5.25 $\pm$ 0.75 |
| 4. | S | N | RM | Not rej | 0.0 | 0.0 | 0.0 | 0.0 | -4.55 $\pm$ 0.82 |

- FID claims $Q$ is better when the two models are equally good. Not account for uncertainty.
- All have high test power when $Q$ is indeed better.

# Problem Setting 2

- $p$, $q$ : probability density functions up to the normalizer
- $r$ : unknown data generating density (unknown).
- Observe $Z_n \overset{i.i.d.}{\sim} R$ and have explicit $p$, $q$.

$$H_0 \colon p \text{ and } q \text{ equally model } r$$
$$H_1 \colon q \text{ models } r \text{ better.}$$

- Formulate as

$$H_0 \colon D(p, r) - D(q, r) = 0$$
$$H_1 \colon D(p, r) - D(q, r) > 0,$$

  for some distance $D$.
- Statistic: $\hat{S}_n = \widehat{D}(p, r) - \widehat{D}(q, r)$. Large, positive $\implies$ $Q$ is better.

# Problem Setting 2

- $p$, $q$ : probability density functions up to the normalizer
- $r$ : unknown data generating density (unknown).
- Observe $Z_n \overset{i.i.d.}{\sim} R$ and have explicit $p$, $q$.

$$H_0\colon p \text{ and } q \text{ equally model } r$$
$$H_1\colon q \text{ models } r \text{ better.}$$

- Formulate as

$$H_0\colon D(p,r) - D(q,r) = 0$$
$$H_1\colon D(p,r) - D(q,r) > 0,$$

  for some distance $D$.
- Statistic: $\hat{S}_n = \widehat{D}(p,r) - \widehat{D}(q,r)$. Large, positive $\implies$ $Q$ is better.

# Problem Setting 2

- $p, q$ : probability density functions up to the normalizer
- $r$ : unknown data generating density (unknown).
- Observe $Z_n \overset{i.i.d.}{\sim} R$ and have explicit $p$, $q$.

$$H_0 \colon p \text{ and } q \text{ equally model } r$$
$$H_1 \colon q \text{ models } r \text{ better.}$$

- Formulate as

$$H_0 \colon D(p, r) - D(q, r) = 0$$
$$H_1 \colon D(p, r) - D(q, r) > 0,$$

  for some distance $D$.
- Statistic: $\hat{S}_n = \widehat{D}(p, r) - \widehat{D}(q, r)$. Large, positive $\implies$ $Q$ is better.

# Problem Setting 2

- $p$, $q$ : probability density functions up to the normalizer
- $r$ : unknown data generating density (unknown).
- Observe $Z_n \overset{i.i.d.}{\sim} R$ and have explicit $p$, $q$.

$$H_0: p \text{ and } q \text{ equally model } r$$
$$H_1: q \text{ models } r \text{ better.}$$

- Formulate as

$$H_0: D(p,r) - D(q,r) = 0$$
$$H_1: D(p,r) - D(q,r) > 0,$$

for some distance $D$.

- Statistic: $\hat{S}_n = \widehat{D}(p,r) - \widehat{D}(q,r)$. Large, positive $\implies$ $Q$ is better.

# Problem Setting 2

- $p$, $q$ : probability density functions up to the normalizer
- $r$ : unknown data generating density (unknown).
- Observe $Z_n \overset{i.i.d.}{\sim} R$ and have explicit $p$, $q$.

$$H_0: p \text{ and } q \text{ equally model } r$$
$$H_1: q \text{ models } r \text{ better.}$$

- Formulate as

$$H_0: D(p, r) - D(q, r) = 0$$
$$H_1: D(p, r) - D(q, r) > 0,$$

  for some distance $D$.
- Statistic: $\hat{S}_n = \widehat{D}(p, r) - \widehat{D}(q, r)$. Large, positive $\implies$ $Q$ is better.

## The Finite Set Stein Discrepancy (FSSD) (NeurIPS 2017 Best Paper)

Recall $\text{witness}(\mathbf{v}) = \mathbb{E}_{\mathbf{z}\sim r}[k_{\mathbf{v}}(\mathbf{z})] - \mathbb{E}_{\mathbf{x}\sim p}[k_{\mathbf{v}}(\mathbf{x})]$

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

## The Finite Set Stein Discrepancy (FSSD) (NeurIPS 2017 Best Paper)

Recall $\text{witness}(\mathbf{v}) = \mathbb{E}_{\mathbf{z}\sim r}[k_{\mathbf{v}}(\mathbf{z})] - \mathbb{E}_{\mathbf{x}\sim p}[k_{\mathbf{v}}(\mathbf{x})]$

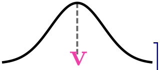
**Problem**: No sample from $p$. Cannot estimate $\mathbb{E}_{\mathbf{x}\sim p}[k_{\mathbf{v}}(\mathbf{x})]$.

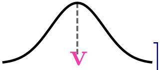
$$\text{(Stein) witness}(\mathbf{v}) = \mathbb{E}_{\mathbf{z}\sim r}[\quad T_p k_{\mathbf{v}}(\mathbf{z}) \quad] - \mathbb{E}_{\mathbf{x}\sim p}[\quad T_p k_{\mathbf{v}}(\mathbf{x}) \quad]$$

# The Finite Set Stein Discrepancy (FSSD) (NeurIPS 2017 Best Paper)

Recall $\text{witness}(\mathbf{v}) = \mathbb{E}_{\mathbf{z}\sim r}[k_{\mathbf{v}}(\mathbf{z})] - \mathbb{E}_{\mathbf{x}\sim p}[k_{\mathbf{v}}(\mathbf{x})]$

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

(Stein) $\text{witness}(\mathbf{v}) = \mathbb{E}_{\mathbf{z}\sim r}[\, T_p$ 

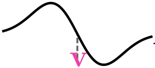
$] - \mathbb{E}_{\mathbf{x}\sim p}[\, T_p$ 

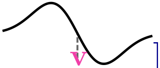
$]$

# The Finite Set Stein Discrepancy (FSSD) (NeurIPS 2017 Best Paper)

Recall witness$(\mathbf{v}) = \mathbb{E}_{\mathbf{z}\sim r}[k_{\mathbf{v}}(\mathbf{z})] - \mathbb{E}_{\mathbf{x}\sim p}[k_{\mathbf{v}}(\mathbf{x})]$

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

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

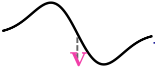
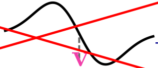
$] - \mathbb{E}_{\mathbf{x}\sim p}[$ 

$]$

## The Finite Set Stein Discrepancy (FSSD) (NeurIPS 2017 Best Paper)

Recall $\text{witness}(\mathbf{v}) = \mathbb{E}_{\mathbf{z}\sim r}[k_{\mathbf{v}}(\mathbf{z})] - \mathbb{E}_{\mathbf{x}\sim p}[k_{\mathbf{v}}(\mathbf{x})]$

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

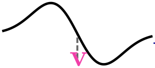
(Stein) $\text{witness}(\mathbf{v}) = \mathbb{E}_{\mathbf{z}\sim r}[$  $] - \mathbb{E}_{\mathbf{x}\sim p}[$  $]$

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

## The Finite Set Stein Discrepancy (FSSD) (NeurIPS 2017 Best Paper)

Recall $\mathrm{witness}(\mathbf{v}) = \mathbb{E}_{\mathbf{z}\sim r}[k_{\mathbf{v}}(\mathbf{z})] - \mathbb{E}_{\mathbf{x}\sim p}[k_{\mathbf{v}}(\mathbf{x})]$

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

(Stein) $\mathrm{witness}(\mathbf{v}) = \mathbb{E}_{\mathbf{z}\sim r}[$ 

$]$

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

## The Finite Set Stein Discrepancy (FSSD) (NeurIPS 2017 Best Paper)

Recall $\text{witness}(\mathbf{v}) = \mathbb{E}_{\mathbf{z}\sim r}[k_{\mathbf{v}}(\mathbf{z})] - \mathbb{E}_{\mathbf{x}\sim p}[k_{\mathbf{v}}(\mathbf{x})]$

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

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

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

## The Finite Set Stein Discrepancy (FSSD) (NeurIPS 2017 Best Paper)

Recall $\text{witness}(\mathbf{v}) = \mathbb{E}_{\mathbf{z}\sim r}[k_{\mathbf{v}}(\mathbf{z})] - \mathbb{E}_{\mathbf{x}\sim p}[k_{\mathbf{v}}(\mathbf{x})]$

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

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

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

- UME defined with this new Stein witness function is called the Finite-Set Stein Discrepancy (Jitkrittum et al., 2017).

# The Finite Set Stein Discrepancy (FSSD) (NeurIPS 2017 Best Paper)

Recall $\text{witness}(\mathbf{v}) = \mathbb{E}_{\mathbf{z}\sim r}[k_{\mathbf{v}}(\mathbf{z})] - \mathbb{E}_{\mathbf{x}\sim p}[k_{\mathbf{v}}(\mathbf{x})]$

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

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

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

- UME defined with this new Stein witness function is called the Finite-Set Stein Discrepancy (Jitkrittum et al., 2017).
- $T_p$ is called a **Stein operator**.

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

which is independent of the normalizer of $p$.

## The Finite Set Stein Discrepancy (FSSD) (NeurIPS 2017 Best Paper)

Recall $\text{witness}(\mathbf{v}) = \mathbb{E}_{\mathbf{z}\sim r}[k_{\mathbf{v}}(\mathbf{z})] - \mathbb{E}_{\mathbf{x}\sim p}[k_{\mathbf{v}}(\mathbf{x})]$

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

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

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

- UME defined with this new Stein witness function is called the Finite-Set Stein Discrepancy (Jitkrittum et al., 2017).
- $T_p$ is called a **Stein operator**.

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

  which is independent of the normalizer of $p$.
- Can construct Rel-FSSD test similarly: optimize $V$ to show where $Q$ is better, asymptotic normality, etc.

# FSSD is a Proper Discrepancy Measure

- $\mathrm{FSSD}^2(p, r) = \frac{1}{dJ} \sum_{j=1}^{J} \|\mathbf{g}_{p,r}(\mathbf{v}_j)\|_2^2$ where $\mathbf{g}_{p,r}(\mathbf{v}) = \mathbb{E}_{\mathbf{z}\sim r} \left[ \frac{1}{p(\mathbf{z})} \frac{d}{d\mathbf{z}} [k_{\mathbf{v}}(\mathbf{z}) p(\mathbf{z})] \right]$ (Stein witness).

**Theorem (FSSD is a discrepancy measure (Jitkrittum et al., 2017))**

*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 $J \geq 1$, $\eta$-almost surely,*

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

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

# FSSD is a Proper Discrepancy Measure

- $\mathrm{FSSD}^2(p, r) = \frac{1}{dJ} \sum_{j=1}^{J} \|\mathbf{g}_{p,r}(\mathbf{v}_j)\|_2^2$ where $\mathbf{g}_{p,r}(\mathbf{v}) = \mathbb{E}_{\mathbf{z}\sim r}\left[\frac{1}{p(\mathbf{z})}\frac{d}{d\mathbf{z}}[k_{\mathbf{v}}(\mathbf{z})p(\mathbf{z})]\right]$ (Stein witness).

**Theorem (FSSD is a discrepancy measure (Jitkrittum et al., 2017))**

*Main conditions:*

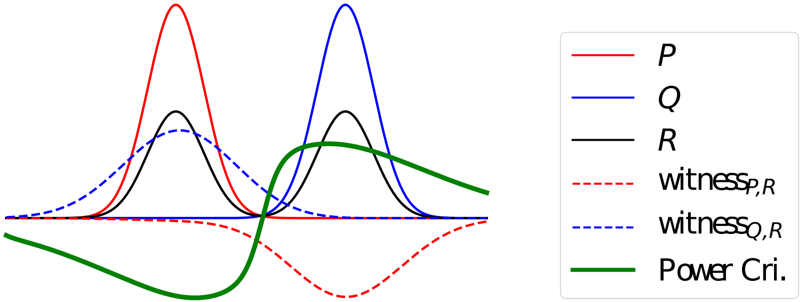
1. *(Nice kernel) Kernel $k$ is $C_0$-universal, and real analytic e.g., Gaussian kernel.*
2. *(Vanishing boundary)* $\lim_{\|\mathbf{x}\|\to\infty} p(\mathbf{x})k_{\mathbf{v}}(\mathbf{x}) = \mathbf{0}$.
3. *(Avoid "blind spots") Locations $\mathbf{v}_1, \ldots, \mathbf{v}_J \sim \eta$ which has a density.*

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

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

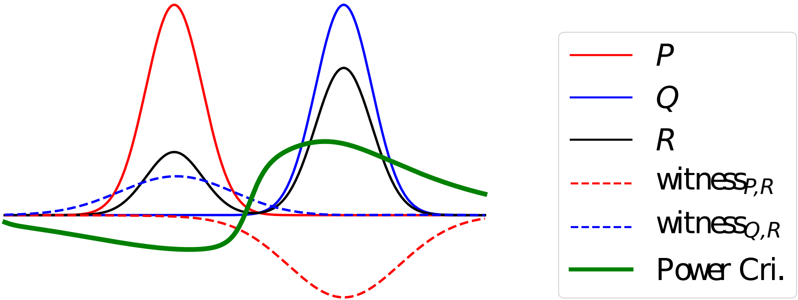
**Summary**: Evaluating the witness at random locations is sufficient to detect the discrepancy between $p, r$.

# Relative FSSD Witness Function



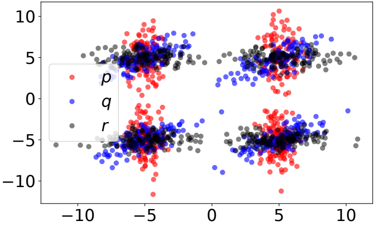

- Unlike UME which cares about probability mass, FSSD cares about shape of density functions.
- In FSSD, $p$, $q$ are represented by $\nabla_{\mathbf{x}} \log p(\mathbf{x})$ and $\nabla_{\mathbf{y}} \log q(\mathbf{y})$ (instead of samples).

# Relative FSSD Witness Function



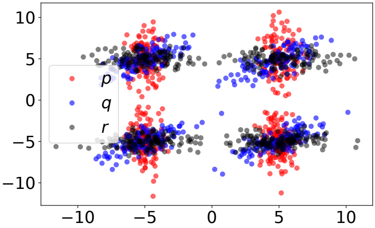

- Unlike UME which cares about probability mass, FSSD cares about shape of density functions.
- In FSSD, $p$, $q$ are represented by $\nabla_{\mathbf{x}} \log p(\mathbf{x})$ and $\nabla_{\mathbf{y}} \log q(\mathbf{y})$ (instead of samples).

# Experiment: 2d Blobs

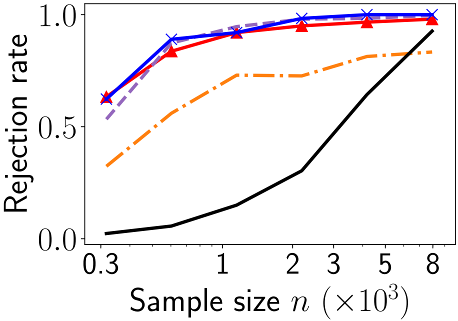



- Problem in $\mathbb{R}^2$. Difference in small scale relative to the global structure.
- $q$ is closer to $r$. So, $H_1$ is true.

# Experiment: 2d Blobs

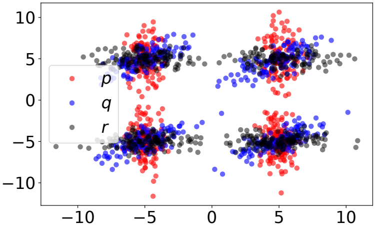



- Problem in $\mathbb{R}^2$. Difference in small scale relative to the global structure.
- $q$ is closer to $r$. So, $H_1$ is true.

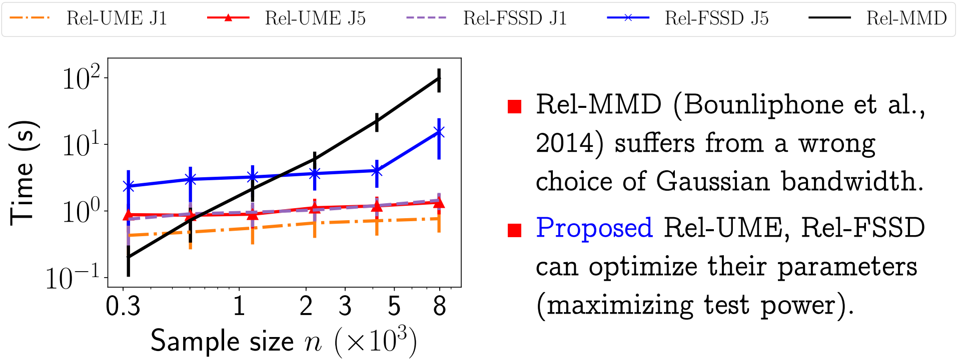




- Rel-MMD (Bounliphone et al., 2014) suffers from a wrong choice of Gaussian bandwidth.
- Proposed Rel-UME, Rel-FSSD can optimize their parameters (maximizing test power).

# Experiment: 2d Blobs




- Problem in $\mathbb{R}^2$. Difference in small scale relative to the global structure.
- $q$ is closer to $r$. So, $H_1$ is true.




- Rel-MMD (Bounliphone et al., 2014) suffers from a wrong choice of Gaussian bandwidth.
- Proposed Rel-UME, Rel-FSSD can optimize their parameters (maximizing test power).

# Summary

Propose a model comparison test Relative UME:

- **Statistical testing**: account for randomness of the distance
- **Linear-time**: runtime complexity = $O(n)$
- **Interpretable**: tells where $Q$ is better $P$ (vice versa)

Another variant Relative FSSD: $P$, $Q$ are explicit (unnormalized) density functions. No need to sample.

Informative Features for Model Comparison
W. Jitkrittum, H. Kanagawa, P. Sangkloy, J. Hays, B. Schölkopf, A. Gretton
NeurIPS 2018
**Python code**: `https://github.com/wittawatj/kernel-mod`

# Questions?

Thank you

# Rewriting UME

- $V := \{\mathbf{v}_1, \ldots, \mathbf{v}_J\} = J$ test locations

$$\mathrm{UME}_V^2(P, R) = \frac{1}{J}\sum_{j=1}^{J}(\mu_P(\mathbf{v}_j) - \mu_R(\mathbf{v}_j))^2$$

# Rewriting UME

- $V := \{\mathbf{v}_1, \ldots, \mathbf{v}_J\} = J$ test locations

$$\begin{aligned}\mathrm{UME}^2_V(P, R) &= \frac{1}{J}\sum_{j=1}^{J}(\mu_P(\mathbf{v}_j) - \mu_R(\mathbf{v}_j))^2 \\ &= \frac{1}{J}\left\| \begin{pmatrix} \mu_P(\mathbf{v}_1) \\ \vdots \\ \mu_P(\mathbf{v}_J) \end{pmatrix} - \begin{pmatrix} \mu_R(\mathbf{v}_1) \\ \vdots \\ \mu_R(\mathbf{v}_J) \end{pmatrix} \right\|_2^2\end{aligned}$$

# Rewriting UME

- $V := \{\mathbf{v}_1, \ldots, \mathbf{v}_J\} = J$ test locations

$$\mathrm{UME}_V^2(P, R) = \frac{1}{J}\sum_{j=1}^{J}(\mu_P(\mathbf{v}_j) - \mu_R(\mathbf{v}_j))^2$$

$$= \frac{1}{J}\left\| \begin{pmatrix} \mu_P(\mathbf{v}_1) \\ \vdots \\ \mu_P(\mathbf{v}_J) \end{pmatrix} - \begin{pmatrix} \mu_R(\mathbf{v}_1) \\ \vdots \\ \mu_R(\mathbf{v}_J) \end{pmatrix} \right\|_2^2$$

$$= \frac{1}{J}\left\| \mathbb{E}_{\mathbf{x}\sim P} \begin{pmatrix} k(\mathbf{x}, \mathbf{v}_1) \\ \vdots \\ k(\mathbf{x}, \mathbf{v}_J) \end{pmatrix} - \mathbb{E}_{\mathbf{z}\sim R} \begin{pmatrix} k(\mathbf{z}, \mathbf{v}_1) \\ \vdots \\ k(\mathbf{z}, \mathbf{v}_J) \end{pmatrix} \right\|_2^2$$

# Rewriting UME

- $V := \{\mathbf{v}_1, \ldots, \mathbf{v}_J\} =$ $J$ test locations

$$\mathrm{UME}_V^2(P, R) = \frac{1}{J}\sum_{j=1}^{J}(\mu_P(\mathbf{v}_j) - \mu_R(\mathbf{v}_j))^2$$

$$= \frac{1}{J}\left\| \begin{pmatrix} \mu_P(\mathbf{v}_1) \\ \vdots \\ \mu_P(\mathbf{v}_J) \end{pmatrix} - \begin{pmatrix} \mu_R(\mathbf{v}_1) \\ \vdots \\ \mu_R(\mathbf{v}_J) \end{pmatrix} \right\|_2^2$$

$$= \frac{1}{J}\left\| \mathbb{E}_{\mathbf{x}\sim P} \begin{pmatrix} k(\mathbf{x}, \mathbf{v}_1) \\ \vdots \\ k(\mathbf{x}, \mathbf{v}_J) \end{pmatrix} - \mathbb{E}_{\mathbf{z}\sim R} \begin{pmatrix} k(\mathbf{z}, \mathbf{v}_1) \\ \vdots \\ k(\mathbf{z}, \mathbf{v}_J) \end{pmatrix} \right\|_2^2$$

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

$$\mathrm{UME}_V^2(P, R) = \|\mathbb{E}_{\mathbf{x}\sim P}[\psi_V(\mathbf{x})] - \mathbb{E}_{\mathbf{z}\sim R}[\psi_V(\mathbf{z})]\|_2^2.$$

# Rewriting UME

- $V := \{\mathbf{v}_1, \dots, \mathbf{v}_J\} =$ $J$ test locations

$$\mathrm{UME}_V^2(P, R) = \frac{1}{J}\sum_{j=1}^{J}(\mu_P(\mathbf{v}_j) - \mu_R(\mathbf{v}_j))^2$$

$$= \frac{1}{J}\left\| \begin{pmatrix} \mu_P(\mathbf{v}_1) \\ \vdots \\ \mu_P(\mathbf{v}_J) \end{pmatrix} - \begin{pmatrix} \mu_R(\mathbf{v}_1) \\ \vdots \\ \mu_R(\mathbf{v}_J) \end{pmatrix} \right\|_2^2$$

$$= \frac{1}{J}\left\| \mathbb{E}_{\mathbf{x}\sim P} \begin{pmatrix} k(\mathbf{x}, \mathbf{v}_1) \\ \vdots \\ k(\mathbf{x}, \mathbf{v}_J) \end{pmatrix} - \mathbb{E}_{\mathbf{z}\sim R} \begin{pmatrix} k(\mathbf{z}, \mathbf{v}_1) \\ \vdots \\ k(\mathbf{z}, \mathbf{v}_J) \end{pmatrix} \right\|_2^2$$

Let $\psi_V(\mathbf{x}) := \frac{1}{\sqrt{J}}(k(\mathbf{x}, \mathbf{v}_1), \dots, k(\mathbf{x}, \mathbf{v}_J))^\top \in \mathbb{R}^J$. Equivalently,

$$\mathrm{UME}_V^2(P, R) = \|\mathbb{E}_{\mathbf{x}\sim P}[\psi_V(\mathbf{x})] - \mathbb{E}_{\mathbf{z}\sim R}[\psi_V(\mathbf{z})]\|_2^2.$$

- Empirical $\widehat{\mathrm{UME}^2}(P, R) =$ replace $\mathbb{E}$'s above with $\frac{1}{n}\sum_{i=1}^{n}$.

## $\widehat{\mathrm{UME}^2_V}(P, R)$ and $\widehat{\mathrm{UME}^2_V}(Q, R)$ are Correlated

- Write $U_P^2 = \mathrm{UME}^2(P, R)$ and $U_Q^2 = \mathrm{UME}^2(Q, R)$.
- Let $S := U_P^2 - U_Q^2$. So $H_0 : S = 0$ and $H_1 : S > 0$.
- Let $C_V^S := \mathrm{cov}_{\boldsymbol{y} \sim S}[\psi_V(\boldsymbol{y})]$ where $S \in \{P, Q, R\}$.
- Let $\boldsymbol{M} := \begin{pmatrix} \psi_V^P - \psi_V^R & \mathbf{0} \\ \mathbf{0} & \psi_W^Q - \psi_W^R \end{pmatrix}$.
- Let $\begin{pmatrix} \zeta_P^2 & \zeta_{PQ} \\ \zeta_{PQ} & \zeta_Q^2 \end{pmatrix} := \boldsymbol{M}^\top \begin{pmatrix} C_V^P + C_V^R & C_V^R \\ (C_V^R)^\top & C_W^Q + C_W^R \end{pmatrix} \boldsymbol{M}$

**Proposition (Joint distribution of $\widehat{U_P^2}$ and $\widehat{U_Q^2}$)**

*Assume that $P, Q$ and $R$ are all distinct. Under mild conditions,*

1. $\sqrt{n}\left(\begin{pmatrix} \widehat{U_P^2} \\ \widehat{U_Q^2} \end{pmatrix} - \begin{pmatrix} U_P^2 \\ U_Q^2 \end{pmatrix}\right) \xrightarrow{d} \mathcal{N}\left(\mathbf{0}, 4\begin{pmatrix} \zeta_P^2 & \zeta_{PQ} \\ \zeta_{PQ} & \zeta_Q^2 \end{pmatrix}\right)$;
2. $\sqrt{n}\left(\widehat{S}_n - S\right) \xrightarrow{d} \mathcal{N}\left(0, 4(\zeta_P^2 - 2\zeta_{PQ} + \zeta_Q^2)\right)$.

So, asymptotic null distribution is normal. Easy to get $T_\alpha$.

# $\widehat{\mathrm{UME}^2_V}(P, R)$ and $\widehat{\mathrm{UME}^2_V}(Q, R)$ are Correlated

- Write $U^2_P = \mathrm{UME}^2(P, R)$ and $U^2_Q = \mathrm{UME}^2(Q, R)$.
- Let $S := U^2_P - U^2_Q$. So $H_0 : S = 0$ and $H_1 : S > 0$.
- Let $C^S_V := \mathrm{cov}_{\boldsymbol{y} \sim S}[\psi_V(\boldsymbol{y})]$ where $S \in \{P, Q, R\}$.
- Let $\boldsymbol{M} := \begin{pmatrix} \psi^P_V - \psi^R_V & \mathbf{0} \\ \mathbf{0} & \psi^Q_W - \psi^R_W \end{pmatrix}$.
- Let $\begin{pmatrix} \zeta^2_P & \zeta_{PQ} \\ \zeta_{PQ} & \zeta^2_Q \end{pmatrix} := \boldsymbol{M}^\top \begin{pmatrix} C^P_V + C^R_V & C^R_V \\ (C^R_V)^\top & C^Q_W + C^R_W \end{pmatrix} \boldsymbol{M}$

**Proposition (Joint distribution of $\widehat{U^2_P}$ and $\widehat{U^2_Q}$)**

*Assume that $P, Q$ and $R$ are all distinct. Under mild conditions,*

1. $\sqrt{n}\left(\begin{pmatrix} \widehat{U^2_P} \\ \widehat{U^2_Q} \end{pmatrix} - \begin{pmatrix} U^2_P \\ U^2_Q \end{pmatrix}\right) \xrightarrow{d} \mathcal{N}\left(\mathbf{0}, 4\begin{pmatrix} \zeta^2_P & \zeta_{PQ} \\ \zeta_{PQ} & \zeta^2_Q \end{pmatrix}\right)$;
2. $\sqrt{n}\left(\widehat{S}_n - S\right) \xrightarrow{d} \mathcal{N}\left(0, 4(\zeta^2_P - 2\zeta_{PQ} + \zeta^2_Q)\right)$.

So, asymptotic null distribution is normal. Easy to get $T_\alpha$.

# $\widehat{\mathrm{UME}^2_V}(P,R)$ and $\widehat{\mathrm{UME}^2_V}(Q,R)$ are Correlated

- Write $U_P^2 = \mathrm{UME}^2(P,R)$ and $U_Q^2 = \mathrm{UME}^2(Q,R)$.
- Let $S := U_P^2 - U_Q^2$. So $H_0 : S = 0$ and $H_1 : S > 0$.
- Let $C_V^S := \mathrm{cov}_{\boldsymbol{y}\sim S}[\psi_V(\boldsymbol{y})]$ where $S \in \{P, Q, R\}$.
- Let $\boldsymbol{M} := \begin{pmatrix} \psi_V^P - \psi_V^R & \mathbf{0} \\ \mathbf{0} & \psi_W^Q - \psi_W^R \end{pmatrix}$.
- Let $\begin{pmatrix} \zeta_P^2 & \zeta_{PQ} \\ \zeta_{PQ} & \zeta_Q^2 \end{pmatrix} := \boldsymbol{M}^\top \begin{pmatrix} C_V^P + C_V^R & C_V^R \\ (C_V^R)^\top & C_W^Q + C_W^R \end{pmatrix} \boldsymbol{M}$

**Proposition (Joint distribution of $\widehat{U_P^2}$ and $\widehat{U_Q^2}$)**

*Assume that $P, Q$ and $R$ are all distinct. Under mild conditions,*

1. $\sqrt{n}\left(\begin{pmatrix} \widehat{U_P^2} \\ \widehat{U_Q^2} \end{pmatrix} - \begin{pmatrix} U_P^2 \\ U_Q^2 \end{pmatrix}\right) \xrightarrow{d} \mathcal{N}\left(\mathbf{0}, 4\begin{pmatrix} \zeta_P^2 & \zeta_{PQ} \\ \zeta_{PQ} & \zeta_Q^2 \end{pmatrix}\right)$;
2. $\sqrt{n}\left(\widehat{S}_n - S\right) \xrightarrow{d} \mathcal{N}\left(0, 4(\zeta_P^2 - 2\zeta_{PQ} + \zeta_Q^2)\right)$.

So, asymptotic null distribution is normal. Easy to get $T_\alpha$.

## $\widehat{\mathrm{UME}}^2_V(P,R)$ and $\widehat{\mathrm{UME}}^2_V(Q,R)$ are Correlated

- Write $U_P^2 = \mathrm{UME}^2(P,R)$ and $U_Q^2 = \mathrm{UME}^2(Q,R)$.
- Let $S := U_P^2 - U_Q^2$. So $H_0 : S = 0$ and $H_1 : S > 0$.
- Let $C_V^S := \mathrm{cov}_{\boldsymbol{y}\sim S}[\psi_V(\boldsymbol{y})]$ where $S \in \{P, Q, R\}$.
- Let $\boldsymbol{M} := \begin{pmatrix} \psi_V^P - \psi_V^R & \mathbf{0} \\ \mathbf{0} & \psi_W^Q - \psi_W^R \end{pmatrix}$.
- Let $\begin{pmatrix} \zeta_P^2 & \zeta_{PQ} \\ \zeta_{PQ} & \zeta_Q^2 \end{pmatrix} := \boldsymbol{M}^\top \begin{pmatrix} C_V^P + C_V^R & C_V^R \\ (C_V^R)^\top & C_W^Q + C_W^R \end{pmatrix} \boldsymbol{M}$

**Proposition (Joint distribution of $\widehat{U_P^2}$ and $\widehat{U_Q^2}$)**

*Assume that $P, Q$ and $R$ are all distinct. Under mild conditions,*

1. $\sqrt{n}\left(\begin{pmatrix} \widehat{U_P^2} \\ \widehat{U_Q^2} \end{pmatrix} - \begin{pmatrix} U_P^2 \\ U_Q^2 \end{pmatrix}\right) \xrightarrow{d} \mathcal{N}\left(\mathbf{0}, 4\begin{pmatrix} \zeta_P^2 & \zeta_{PQ} \\ \zeta_{PQ} & \zeta_Q^2 \end{pmatrix}\right)$;
2. $\sqrt{n}\left(\widehat{S}_n - S\right) \xrightarrow{d} \mathcal{N}\left(0, 4(\zeta_P^2 - 2\zeta_{PQ} + \zeta_Q^2)\right)$.

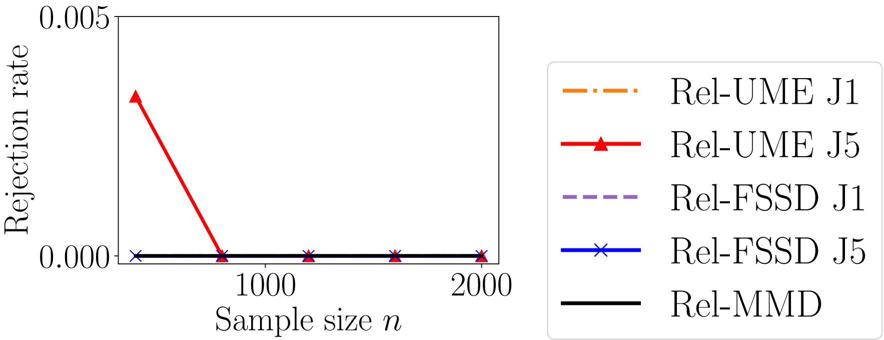
So, asymptotic null distribution is normal. Easy to get $T_\alpha$.

# Experiment: Mean Shift

- Model 1: $p = \mathcal{N}([0.5, 0, \ldots, 0], \boldsymbol{I})$. Model 2: $q = \mathcal{N}([1, 0, \ldots 0], \boldsymbol{I})$
- Data distribution $r = \mathcal{N}(\boldsymbol{0}, \boldsymbol{I})$. Defined on $\mathbb{R}^{50}$.
- Set $\alpha = 0.05$. Should not reject $H_0$.

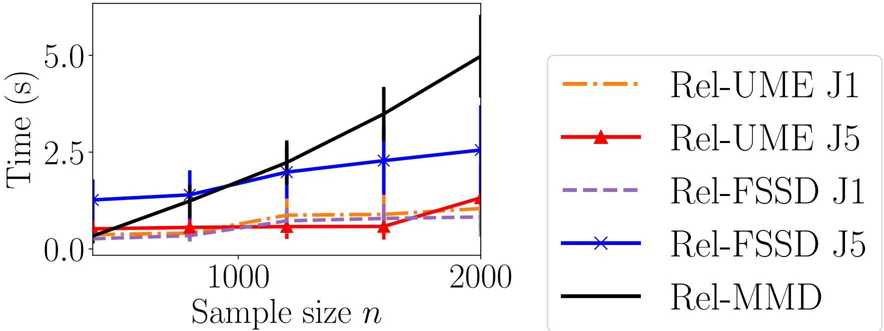
# Experiment: Mean Shift

- Model 1: $p = \mathcal{N}([0.5, 0, \ldots, 0], \boldsymbol{I})$. Model 2: $q = \mathcal{N}([1, 0, \ldots 0], \boldsymbol{I})$
- Data distribution $r = \mathcal{N}(\mathbf{0}, \boldsymbol{I})$. Defined on $\mathbb{R}^{50}$.
- Set $\alpha = 0.05$. Should not reject $H_0$.



# Experiment: Mean Shift

- Model 1: $p = \mathcal{N}([0.5, 0, \ldots, 0], \boldsymbol{I})$. Model 2: $q = \mathcal{N}([1, 0, \ldots 0], \boldsymbol{I})$
- Data distribution $r = \mathcal{N}(\mathbf{0}, \boldsymbol{I})$. Defined on $\mathbb{R}^{50}$.
- Set $\alpha = 0.05$. Should not reject $H_0$.




- MMD runs in $O(n^2)$ time.
- Proposed Rel-UME and Rel-FSSD run in $O(n)$.

# Experiment: Gaussian-Bernoulli Restricted Boltzmann Machine

- $p, q, r$ are all RBM models. $d = 20$ dimensions. $n = 2000$.
- $g_{\mathbf{B},\mathbf{b},\mathbf{c}}(\mathbf{x}) := \frac{1}{Z} \sum_{\mathbf{h}} \exp\left(\mathbf{x}^\top \mathbf{B}\mathbf{h} + \mathbf{b}^\top \mathbf{x} + \mathbf{c}^\top \mathbf{h} - \frac{1}{2}\|\mathbf{x}\|^2\right)$ where $\mathbf{h} \in \{-1, 1\}^5$.
- Define $r(\mathbf{x}) := g_{\mathbf{B},\mathbf{b},\mathbf{c}}(\mathbf{x})$ for some randomly drawn $\mathbf{B}, \mathbf{b}, \mathbf{c}$.
- Let $p(\mathbf{x}) := g_{\mathbf{B}^p,\mathbf{b},\mathbf{c}}(\mathbf{x})$, and $q(\mathbf{x}) := g_{\mathbf{B}^q,\mathbf{b},\mathbf{c}}(\mathbf{x})$.
- $\mathbf{B}^p = \mathbf{B}$ but with $\epsilon$ added to its first entry $B_{1,1}$
- $\mathbf{B}^q = \mathbf{B}$ but with 0.3 added to its first entry $B_{1,1}$
- If $\epsilon > 0.3$, $q$ is better. Should reject $H_0$.

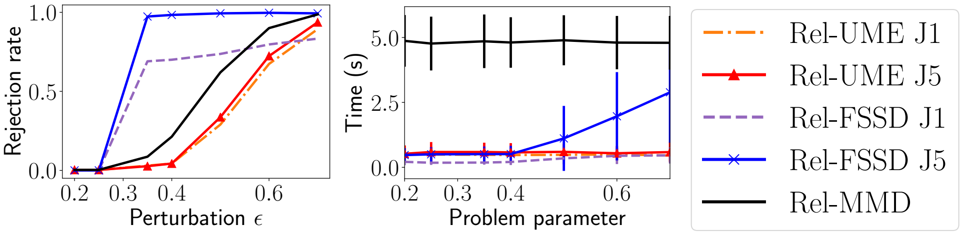
## Experiment: Gaussian-Bernoulli Restricted Boltzmann Machine

- $p, q, r$ are all RBM models. $d = 20$ dimensions. $n = 2000$.
- $g_{\mathbf{B},\mathbf{b},\mathbf{c}}(\mathbf{x}) := \frac{1}{Z}\sum_{\mathbf{h}} \exp\left(\mathbf{x}^\top \mathbf{B}\mathbf{h} + \mathbf{b}^\top \mathbf{x} + \mathbf{c}^\top \mathbf{h} - \frac{1}{2}\|\mathbf{x}\|^2\right)$ where $\mathbf{h} \in \{-1, 1\}^5$.
- Define $r(\mathbf{x}) := g_{\mathbf{B},\mathbf{b},\mathbf{c}}(\mathbf{x})$ for some randomly drawn $\mathbf{B}, \mathbf{b}, \mathbf{c}$.
- Let $p(\mathbf{x}) := g_{\mathbf{B}^p,\mathbf{b},\mathbf{c}}(\mathbf{x})$, and $q(\mathbf{x}) := g_{\mathbf{B}^q,\mathbf{b},\mathbf{c}}(\mathbf{x})$.
- $\mathbf{B}^p = \mathbf{B}$ but with $\epsilon$ added to its first entry $B_{1,1}$
- $\mathbf{B}^q = \mathbf{B}$ but with 0.3 added to its first entry $B_{1,1}$
- If $\epsilon > 0.3$, $q$ is better. Should reject $H_0$.

## Experiment: Gaussian-Bernoulli Restricted Boltzmann Machine

- $p, q, r$ are all RBM models. $d = 20$ dimensions. $n = 2000$.
- $g_{\mathbf{B},\mathbf{b},\mathbf{c}}(\mathbf{x}) := \frac{1}{Z}\sum_{\mathbf{h}} \exp\left(\mathbf{x}^\top \mathbf{B}\mathbf{h} + \mathbf{b}^\top \mathbf{x} + \mathbf{c}^\top \mathbf{h} - \frac{1}{2}\|\mathbf{x}\|^2\right)$ where $\mathbf{h} \in \{-1, 1\}^5$.
- Define $r(\mathbf{x}) := g_{\mathbf{B},\mathbf{b},\mathbf{c}}(\mathbf{x})$ for some randomly drawn $\mathbf{B}, \mathbf{b}, \mathbf{c}$.
- Let $p(\mathbf{x}) := g_{\mathbf{B}^p,\mathbf{b},\mathbf{c}}(\mathbf{x})$, and $q(\mathbf{x}) := g_{\mathbf{B}^q,\mathbf{b},\mathbf{c}}(\mathbf{x})$.
- $\mathbf{B}^p = \mathbf{B}$ but with $\epsilon$ added to its first entry $B_{1,1}$
- $\mathbf{B}^q = \mathbf{B}$ but with 0.3 added to its first entry $B_{1,1}$
- If $\epsilon > 0.3$, $q$ is better. Should reject $H_0$.

# Experiment: Gaussian-Bernoulli Restricted Boltzmann Machine




- Models and and true distribution are very close. Difficult.
- FSSD has access to the density. Higher power than UME, MMD (rely on samples).

# What is $T_p k_{\mathbf{v}}$?

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

# What is $T_p k_{\mathbf{v}}$?

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

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

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

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

# What is $T_p k_{\mathbf{v}}$?

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

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

Normalizer cancels

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

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

# What is $T_p k_{\mathbf{v}}$?

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

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

Normalizer cancels

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

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

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

(assume $\lim_{|\mathbf{x}|\to\infty} k(\mathbf{v},\mathbf{x})p(\mathbf{x})$)

# References I


Interpretable Distribution Features with Maximum Testing Power
Wittawat Jitkrittum, Zoltán Szabó, Kacper Chwialkowski, Arthur Gretton
NIPS 2016 (oral)
**Paper/code**: https://github.com/wittawatj/interpretable-test

A Linear-Time Kernel Goodness-of-Fit Test
Wittawat Jitkrittum, Wenkai Xu, Zoltán Szabó, Kenji Fukumizu, Arthur Gretton
NIPS 2017 (oral, best paper)
**Paper/code**: https://github.com/wittawatj/kernel-gof