Comparing samples from two distributions

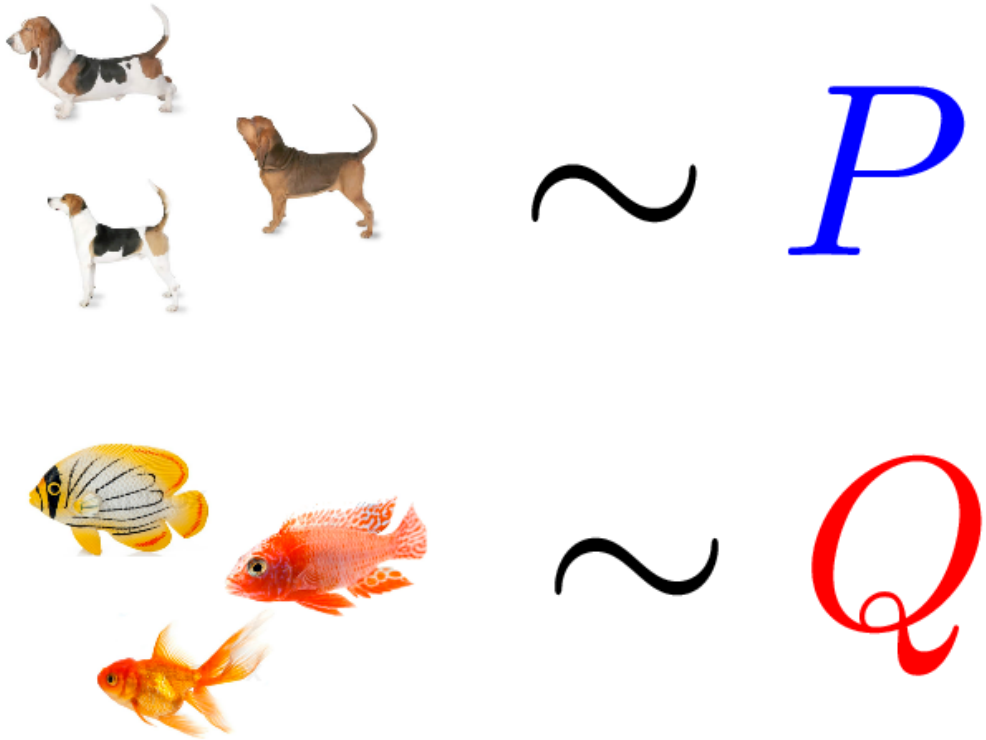
Arthur Gretton

Gatsby Computational Neuroscience Unit,
University College London

Dagstuhl 2016

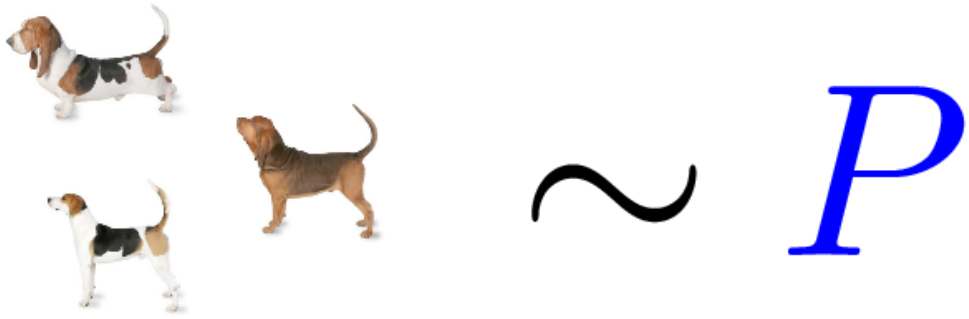
Overview

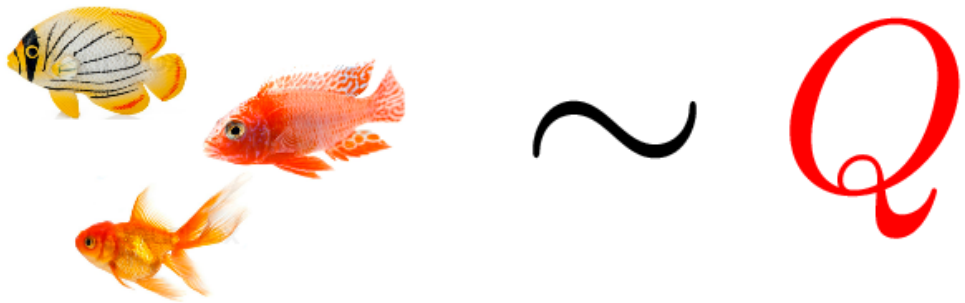
- **Have:** Two collections of samples X, Y from unknown distributions P and Q.
- **Goal:** Learn distinguishing features that indicate how P and Q differ.



Overview

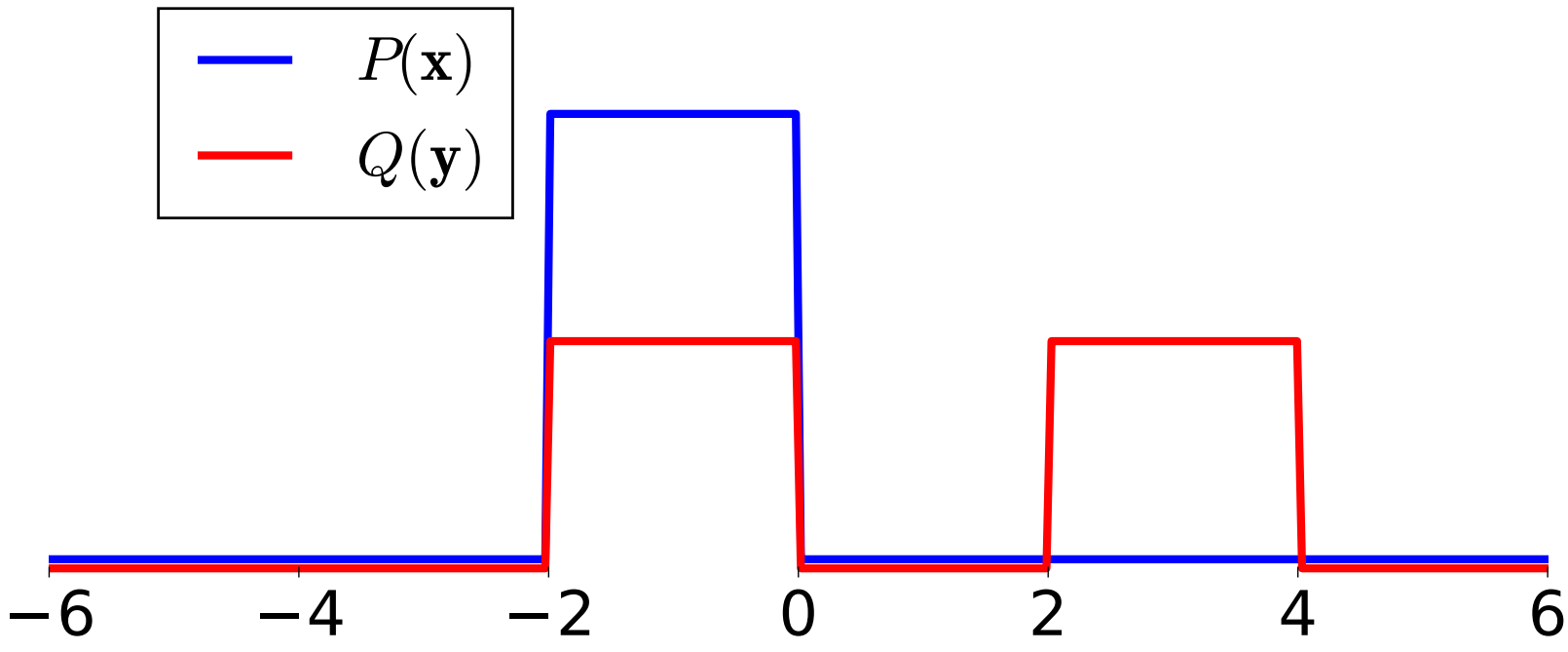
- **Have:** Two collections of samples X, Y from unknown distributions P and Q.
- **Goal:** Learn distinguishing features that indicate how P and Q differ.





The maximum mean discrepancy

Are P and Q different?



The maximum mean discrepancy

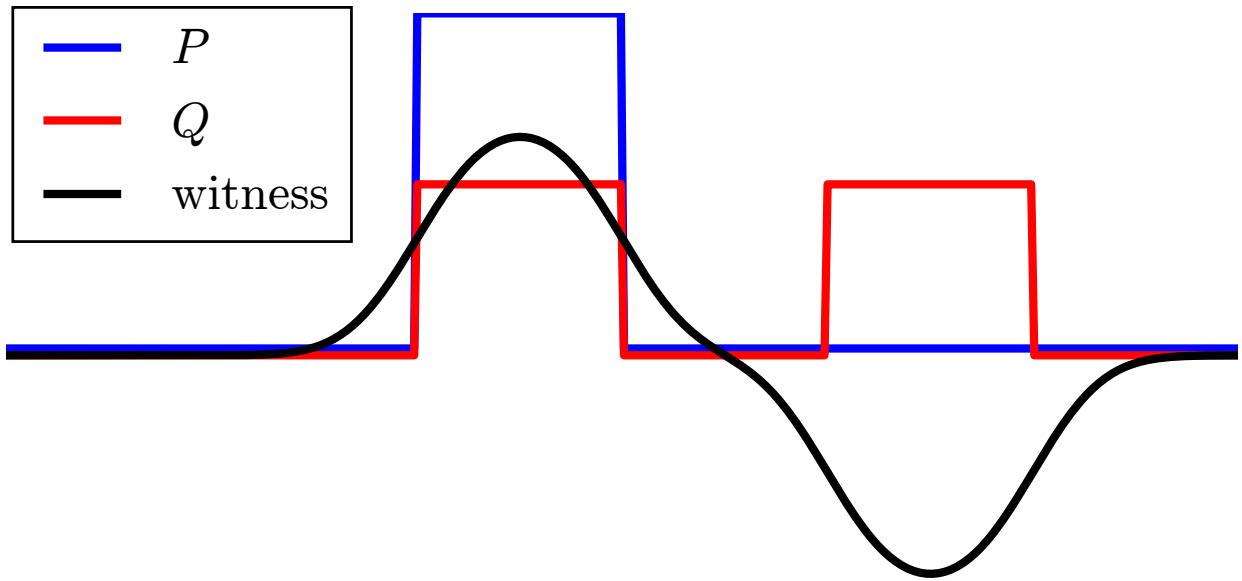
- Maximum mean discrepancy: (G., Borgwardt, Rasch, Schoelkopf, Smola, JMLR 2012).

$$MMD(P, Q; \mathcal{F}) := \sup_{\|f\|_{\mathcal{F}} \leq 1} [E_{\Pr} \mathbf{f}(\mathrm{x}) - E_Q \mathbf{f}(\mathrm{y})]$$

- When $\mathcal{F}$ is an RKHS,

$$f(v) \propto \mu_P(\mathbf{v}) - \mu_Q(v) \qquad \mu_P(\mathbf{v}) := \int k(v, x) dP(v)$$

and $MMD(P, Q; \mathcal{F}) := \|\mu_P(\mathbf{v}) - \mu_Q(v)\|_{\mathcal{F}}$.



The maximum mean discrepancy

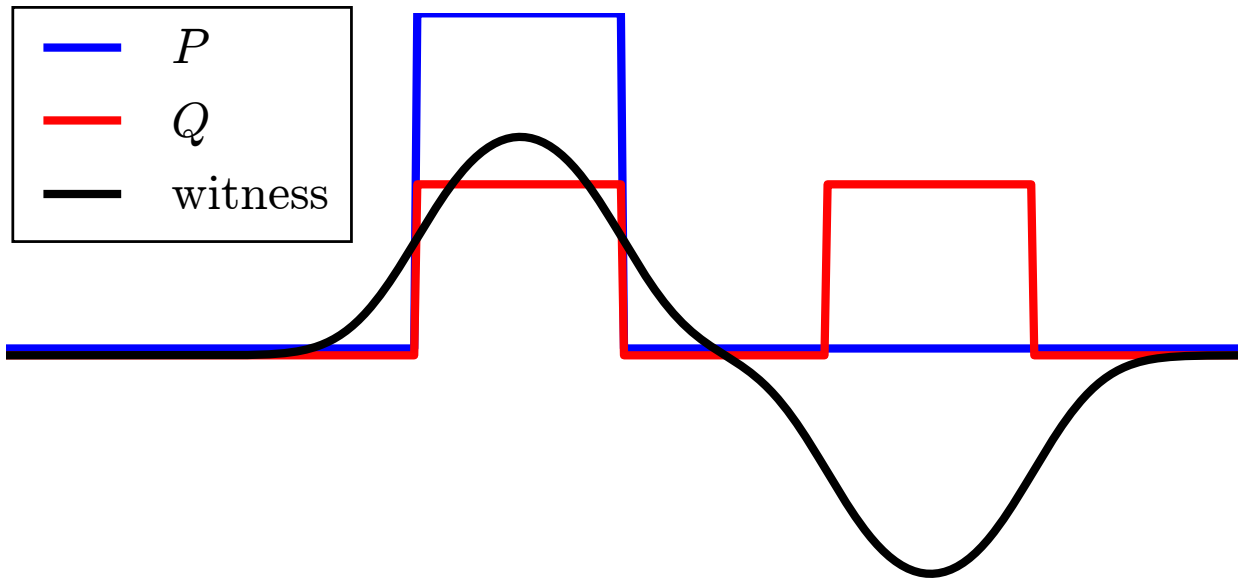
- Maximum mean discrepancy: (G., Borgwardt, Rasch, Schoelkopf, Smola, JMLR 2012).

$$MMD(P, Q; \mathcal{F}) := \sup_{||f||_{\mathcal{F}} \leq 1} [E_{\Pr} \mathbf{f}(\mathrm{x}) - E_Q \mathbf{f}(\mathrm{y})]$$

- When $\mathcal{F}$ is an RKHS,

$$f(v) \propto \mu_P(\mathbf{v}) - \mu_Q(v) \qquad \mu_P(\mathbf{v}) := \int k(v, x) dP(v)$$

and $MMD(P, Q; \mathcal{F}) := \|\mu_P(\mathbf{v}) - \mu_Q(v)\|_{\mathcal{F}}$.

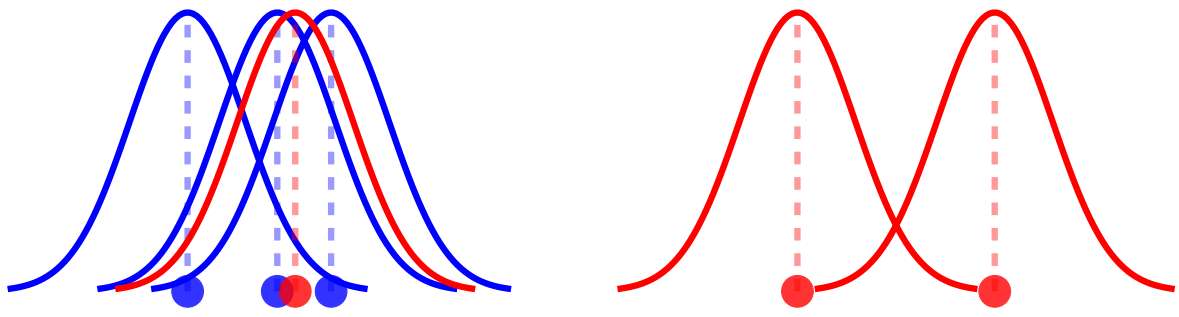


Maximum mean discrepancy on sample

- Observe $\mathsf{X} = \{\mathbf{x}_1, \ldots, \mathbf{x}_n\} \sim P$. $\mathsf{Y} = \{\mathbf{y}_1, \ldots, \mathbf{y}_n\} \sim Q$.
- Mean embeddings: $\hat{\mu}_P(\mathbf{v}) := \frac{1}{m}\sum_{i=1}^{m} k(x_i, v)$ and $\hat{\mu}_Q(v)$.
- witness$(\mathbf{v}) \propto \hat{\mu}_P(\mathbf{v}) - \hat{\mu}_Q(\mathbf{v})$

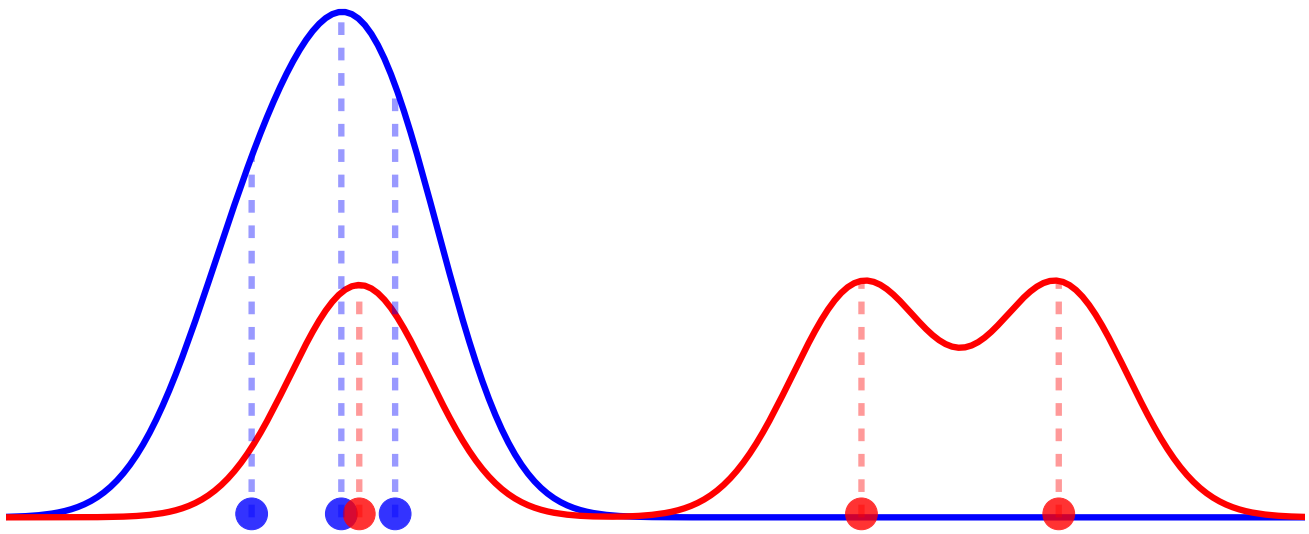
Maximum mean discrepancy on sample

- Observe $\mathsf{X} = \{\mathbf{x}_1, \ldots, \mathbf{x}_n\} \sim P$. $\mathsf{Y} = \{\mathbf{y}_1, \ldots, \mathbf{y}_n\} \sim Q$.
- Mean embeddings: $\hat{\mu}_P(\mathbf{v}) := \frac{1}{m}\sum_{i=1}^{m} k(x_i, v)$ and $\hat{\mu}_Q(v)$.
- $\text{witness}(\mathbf{v}) \propto \hat{\mu}_P(\mathbf{v}) - \hat{\mu}_Q(\mathbf{v})$



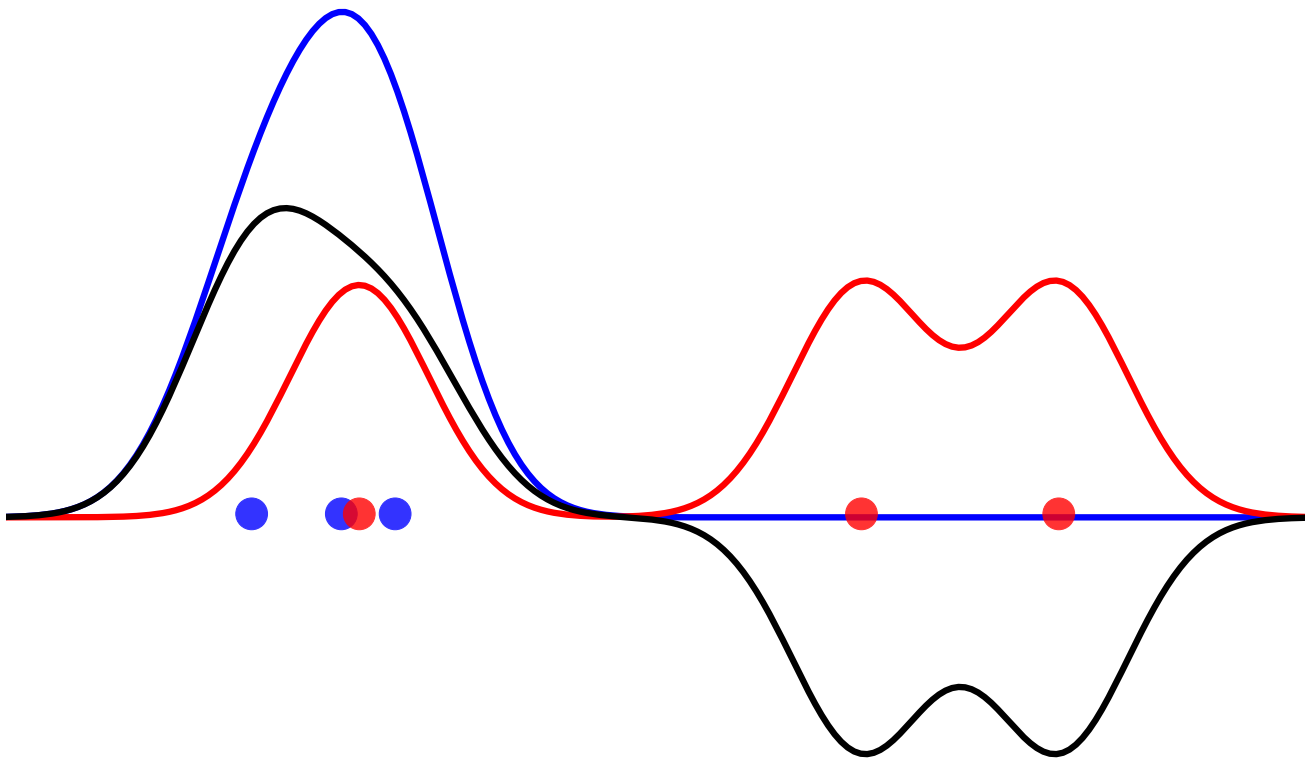
Maximum mean discrepancy on sample

- Observe $\mathsf{X} = \{\mathbf{x}_1, \ldots, \mathbf{x}_n\} \sim P$. $\mathsf{Y} = \{\mathbf{y}_1, \ldots, \mathbf{y}_n\} \sim Q$.
- Mean embeddings: $\hat{\mu}_P(\mathbf{v}) := \frac{1}{m} \sum_{i=1}^{m} k(x_i, v)$ and $\hat{\mu}_Q(v)$.
- $\text{witness}(\mathbf{v}) \propto \hat{\mu}_P(\mathbf{v}) - \hat{\mu}_Q(\mathbf{v})$



Maximum mean discrepancy on sample

- Observe $\mathsf{X} = \{\mathbf{x}_1, \ldots, \mathbf{x}_n\} \sim P$. $\mathsf{Y} = \{\mathbf{y}_1, \ldots, \mathbf{y}_n\} \sim Q$.
- Mean embeddings: $\hat{\mu}_P(\mathbf{v}) := \frac{1}{m} \sum_{i=1}^{m} k(x_i, v)$ and $\hat{\mu}_Q(v)$.
- witness$(\mathbf{v}) \propto \hat{\mu}_P(\mathbf{v}) - \hat{\mu}_Q(\mathbf{v})$



Asymptotics of MMD

- An unbiased empirical estimate of MMD (quadratic cost):

$$\widehat{\mathrm{MMD}}^2 = \frac{1}{n(n-1)} \sum_{i=1}^{n} \sum_{j \neq i} \underbrace{k(x_i, x_j) + k(y_i, y_j) - k(x_i, y_j) - k(y_i, x_j)}_{h((x_i, y_i), (x_j, y_j))}$$

- MMD "far from zero" vs "close to zero" - threshold? **One answer:** asymptotic distribution of $\widehat{\mathrm{MMD}}^2$

Asymptotics of MMD

- An unbiased empirical estimate of MMD (quadratic cost):

$$\widehat{\mathrm{MMD}}^2 = \frac{1}{n(n-1)} \sum_{i=1}^{n} \sum_{j \neq i} \underbrace{k(x_i, x_j) + k(y_i, y_j) - k(x_i, y_j) - k(y_i, x_j)}_{h((x_i, y_i), (x_j, y_j))}$$

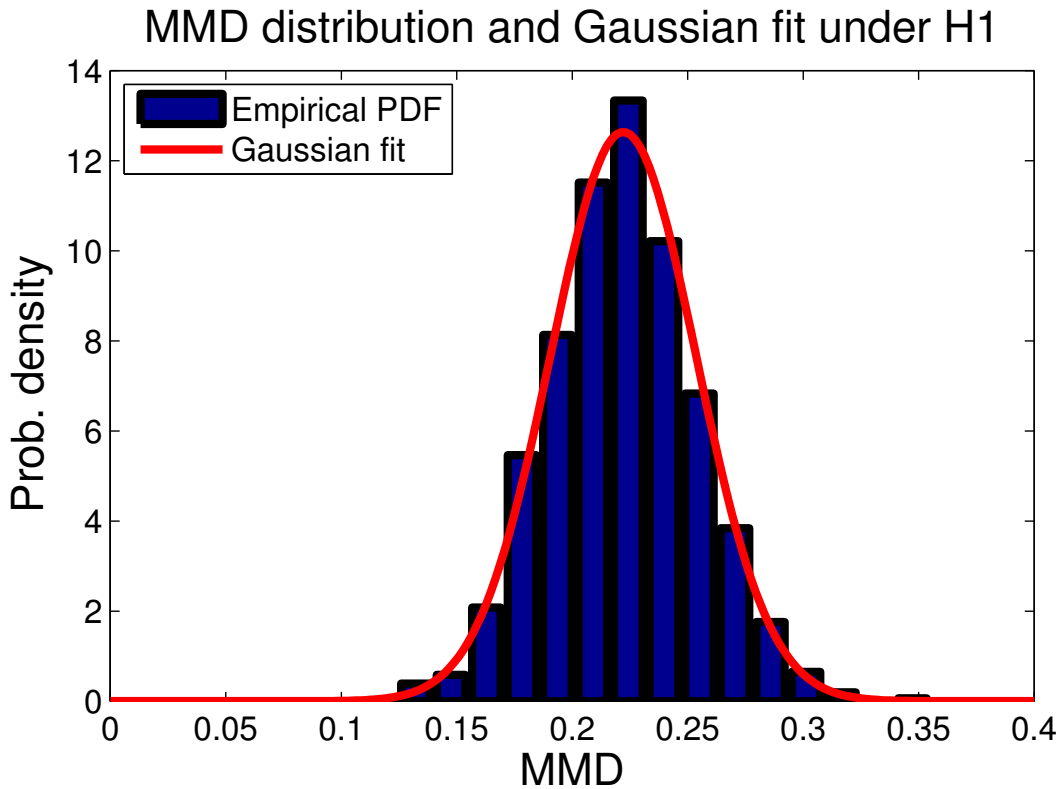
- MMD "far from zero" vs "close to zero" - threshold? **One answer:** asymptotic distribution of $\widehat{\mathrm{MMD}}^2$

Asymptotics of MMD

- When $P \neq Q$, statistic is asymptotically normal,

$$\frac{\widehat{\mathrm{MMD}}^2 - \mathrm{MMD}(P, Q)}{\sqrt{V_n(P, Q)}} \xrightarrow{D} \mathcal{N}(0, 1),$$

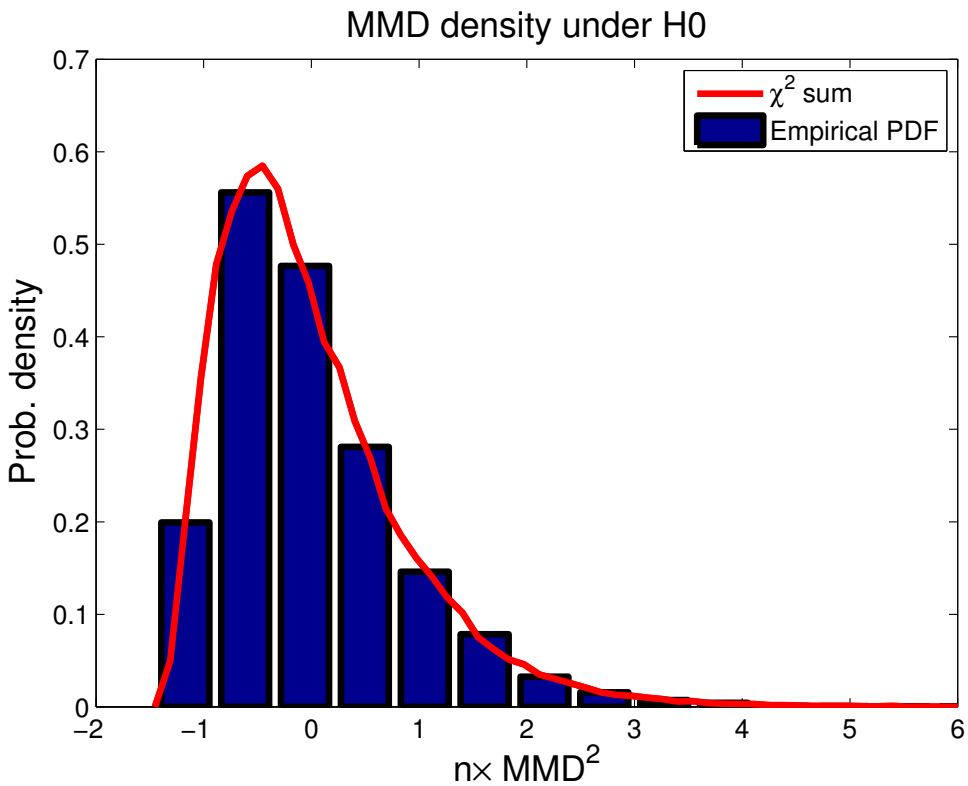
where $V_n(P, Q) = O\left(n^{-1}\right)$.



Asymptotics of MMD

Where $P = Q$, statistic has asymptotic distribution

$$n\widehat{\mathrm{MMD}}^2 \sim \sum_{l=1}^{\infty} \lambda_l \left[z_l^2 - 2 \right]$$

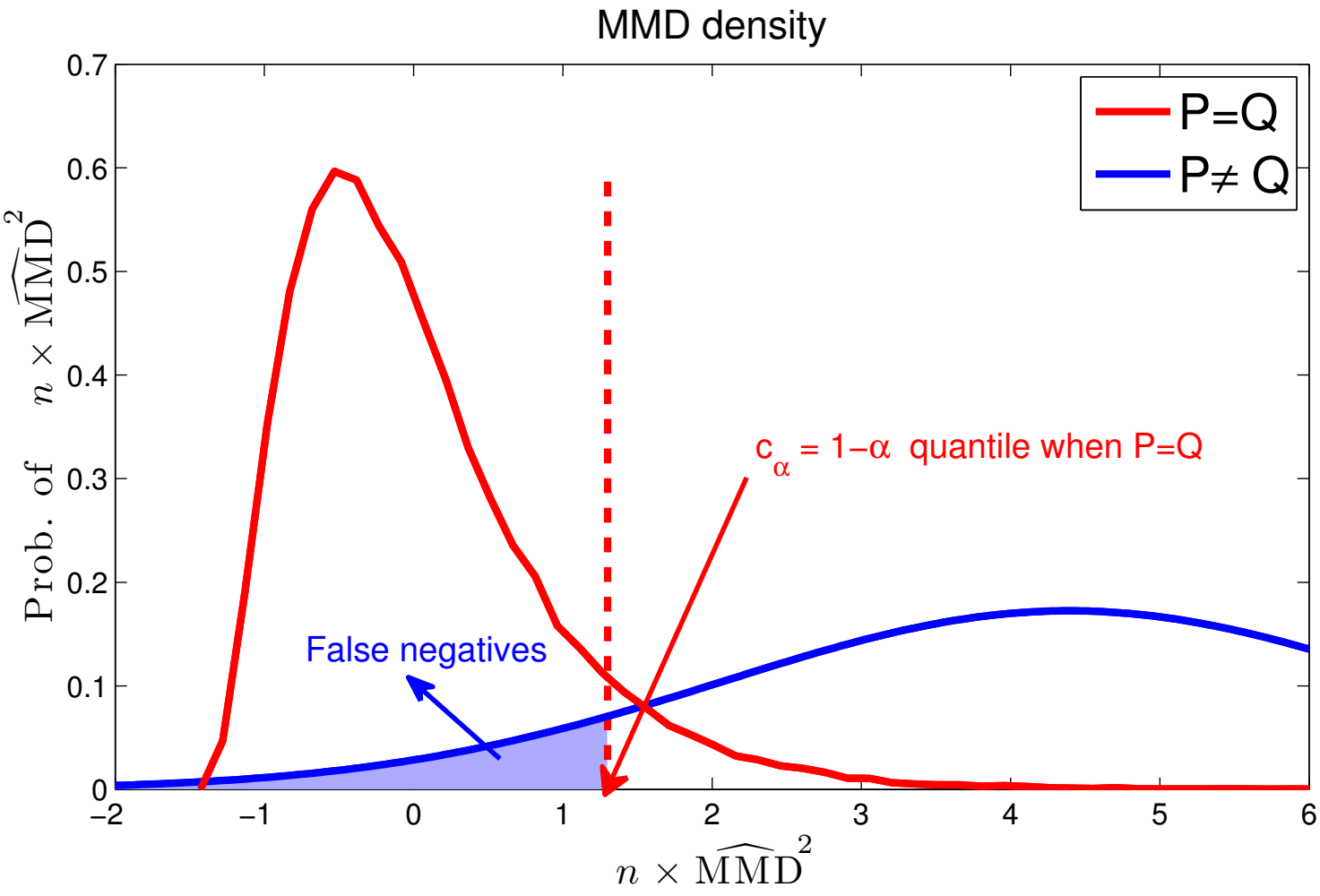


where

$$\lambda_i \psi_i(x') = \int_{\mathcal{X}} \underbrace{\tilde{k}(x, x')}_{\text{centred}} \psi_i(x) dP(x)$$

$$z_l \sim \mathcal{N}(0, 2) \quad \text{i.i.d.}$$

A statistical test



Optimizing test power

- The power of our test ($\Pr_1$ denotes probability under $\mathcal{H}_1$):

$$\Pr_1\left(n\widehat{\mathrm{MMD}}^2 > \hat{c}_\alpha\right)$$

$$= \Pr_1\left(\frac{\widehat{\mathrm{MMD}}^2 - \mathrm{MMD}^2(P,Q)}{\sqrt{V_n(P,Q)}} > \frac{\hat{c}_\alpha/n - \mathrm{MMD}^2(P,Q)}{\sqrt{V_n(P,Q)}}\right)$$

$$\to 1 - \Phi\left(\frac{c_\alpha}{n\sqrt{V_n(P,Q)}} - \frac{\mathrm{MMD}^2(P,Q)}{\sqrt{V_n(P,Q)}}\right)$$

where

  - Φ is the CDF of the standard normal distribution.
  - $\hat{c}_\alpha$ is an estimate of the $1-\alpha$ quantile c_α of the null distribution.

- To maximize test power, maximize

$$\frac{\mathrm{MMD}^2(P,Q)}{\sqrt{V_n(P,Q)}}$$

(Sutherland, Tung, Strathmann, De, Ramdas, Smola, G., in review for ICLR 2017)

Optimizing test power

- The power of our test ($\Pr_1$ denotes probability under $\mathcal{H}_1$):

$$\Pr_1\left(n\widehat{\mathrm{MMD}}^2 > \hat{c}_\alpha\right)$$

$$= \Pr_1\left(\frac{\widehat{\mathrm{MMD}}^2 - \mathrm{MMD}^2(P,Q)}{\sqrt{V_n(P,Q)}} > \frac{\hat{c}_\alpha/n - \mathrm{MMD}^2(P,Q)}{\sqrt{V_n(P,Q)}}\right)$$

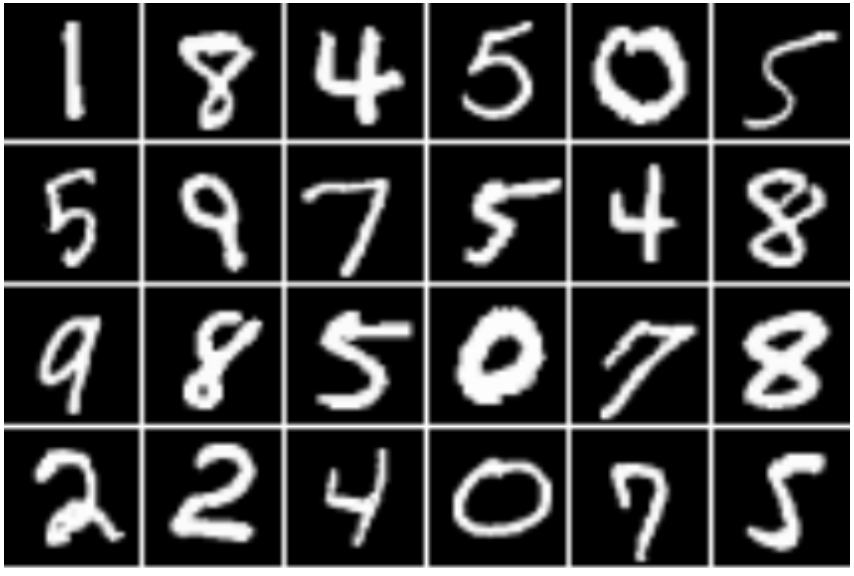
$$\to 1 - \Phi\left(\frac{c_\alpha}{n\sqrt{V_n(P,Q)}} - \frac{\mathrm{MMD}^2(P,Q)}{\sqrt{V_n(P,Q)}}\right)$$

  where
  - Φ is the CDF of the standard normal distribution.
  - $\hat{c}_\alpha$ is an estimate of the $1-\alpha$ quantile c_α of the null distribution.
- To maximize test power, maximize

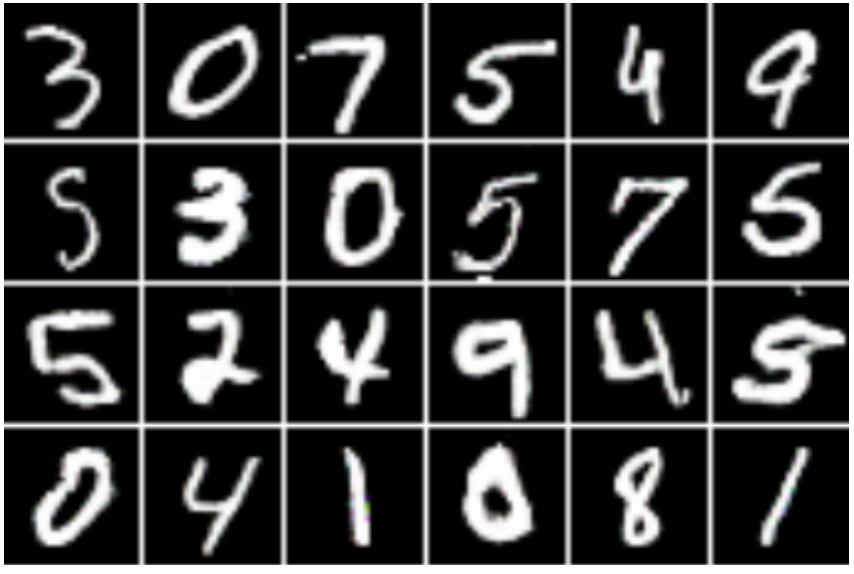
$$\frac{\mathrm{MMD}^2(P,Q)}{\sqrt{V_n(P,Q)}}$$

(Sutherland, Tung, Strathmann, De, Ramdas, Smola, G., in review for ICLR 2017)

Benchmarking generative adversarial networks



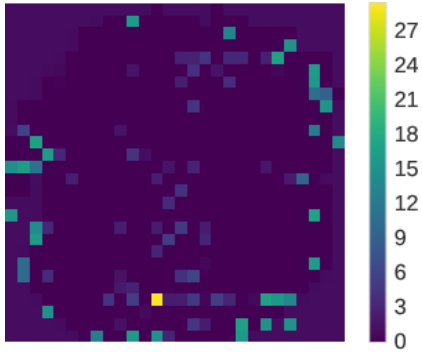
MNIST samples



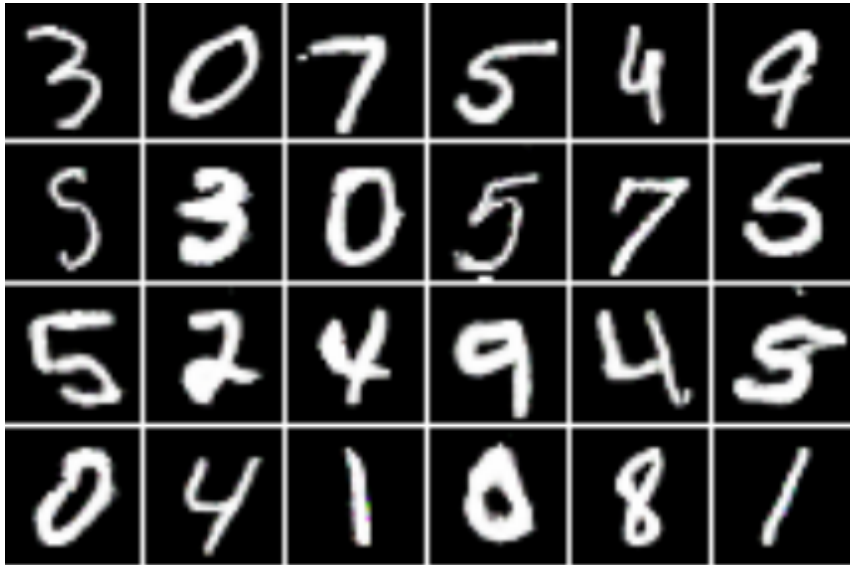
Samples from a GAN

Benchmarking generative adversarial networks

MNIST samples



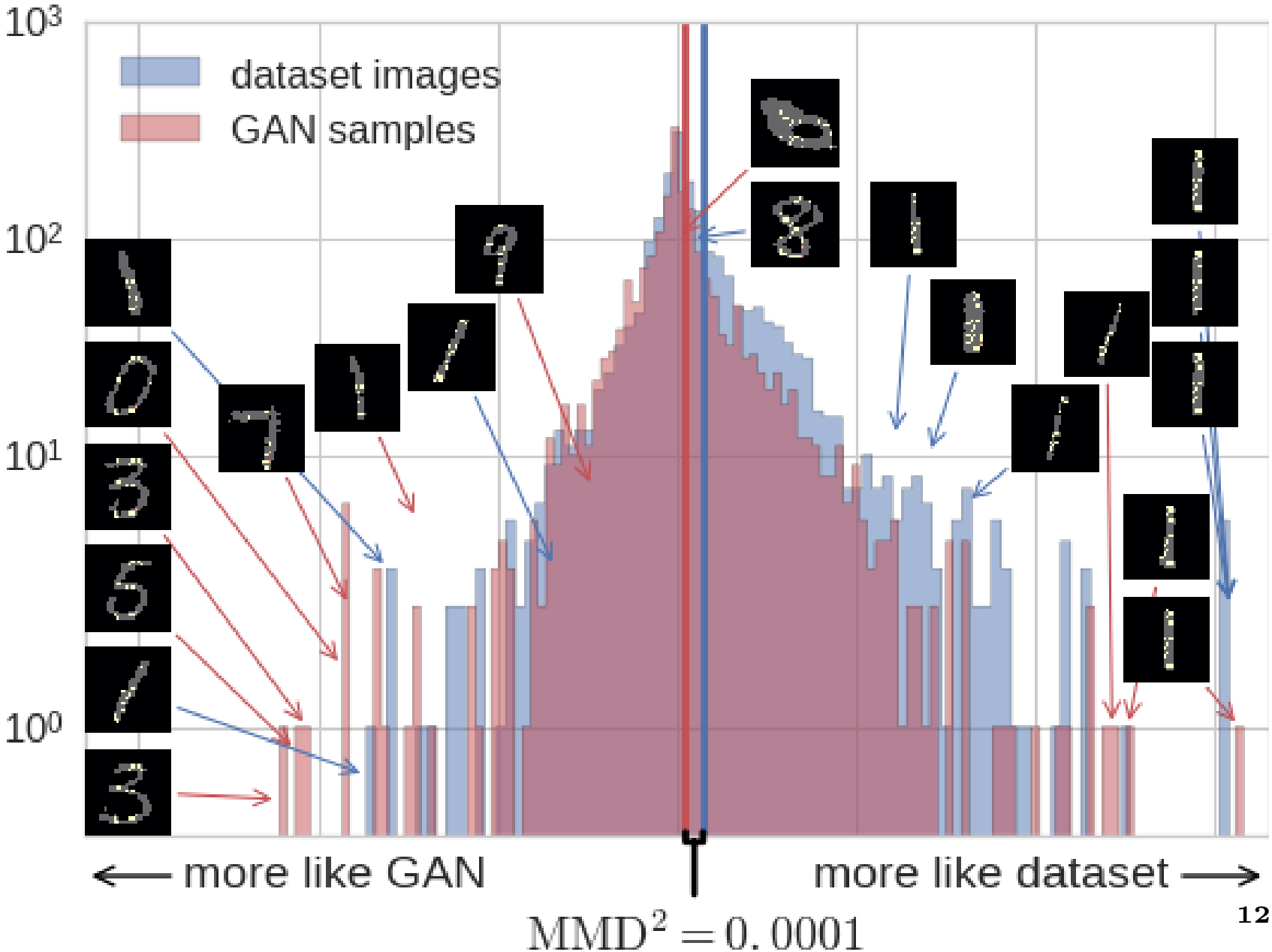
ARD map



Samples from a GAN

- Power for **optimzed ARD kernel**: 1.00 at $\alpha = 0.01$
- Power for optimized RBF kernel: 0.57 at $\alpha = 0.01$

Benchmarking generative adversarial networks



Thoughts and open questions

Shrunken mean embedding estimates (Muandet et al. 2016)

- **First shrinkage estimate:** for some $f^* \in \mathcal{F}$, observe $\mathsf{X} = \{\mathbf{x}_1, \ldots, \mathbf{x}_n\} \sim P$.

$$\hat{\mu}_\alpha := \alpha f^* + (1-\alpha)\hat{\mu} \qquad \hat{\mu} = \frac{1}{n}\sum_{i=1}^{n} k(x_i, \cdot)$$

where

$$\mathbb{E}\left\|\hat{\mu}_\alpha - \mu_P\right\|^2 < \mathbb{E}\left\|\hat{\mu} - \mu_P\right\|^2, \quad \alpha \in \left(0, \frac{2\mathbb{E}\left\|\hat{\mu} - \mu_P\right\|^2}{\mathbb{E}\left\|\hat{\mu} - \mu_P\right\|^2 + \left\|f^* - \mu_P\right\|^2}\right)$$

- **Second shrinkage estimator:**

$$\check{\mu}_\lambda := \sum_{i=1}^{n} \frac{\gamma_i}{\gamma_i + \lambda} \langle \hat{\mu}, \phi_i \rangle \phi_i \qquad \mathbb{E}_P\left(k(X,\cdot) \otimes k(X,\cdot)\right) = \sum_{i=1}^{\infty} \gamma_i \phi_i \otimes \phi_i.$$

(suppresses high frequencies more). Related to **kernel Fisher discriminant**, used in testing (Harchaoui, Bach, Moulines 2007).
Alternatively: Bayesian estimate of μ (Flaxman et al., 2016)

Thoughts and open questions

Shrunken mean embedding estimates (Muandet et al. 2016)

- **First shrinkage estimate:** for some $f^* \in \mathcal{F}$, observe $\mathsf{X} = \{\mathbf{x}_1, \ldots, \mathbf{x}_n\} \sim P$.

$$\hat{\mu}_\alpha := \alpha f^* + (1-\alpha)\hat{\mu} \qquad \hat{\mu} = \frac{1}{n}\sum_{i=1}^{n} k(x_i, \cdot)$$

where

$$\mathbb{E}\left\|\hat{\mu}_\alpha - \mu_P\right\|^2 < \mathbb{E}\left\|\hat{\mu} - \mu_P\right\|^2, \quad \alpha \in \left(0, \frac{2\mathbb{E}\left\|\hat{\mu} - \mu_P\right\|^2}{\mathbb{E}\left\|\hat{\mu} - \mu_P\right\|^2 + \left\|f^* - \mu_P\right\|^2}\right)$$

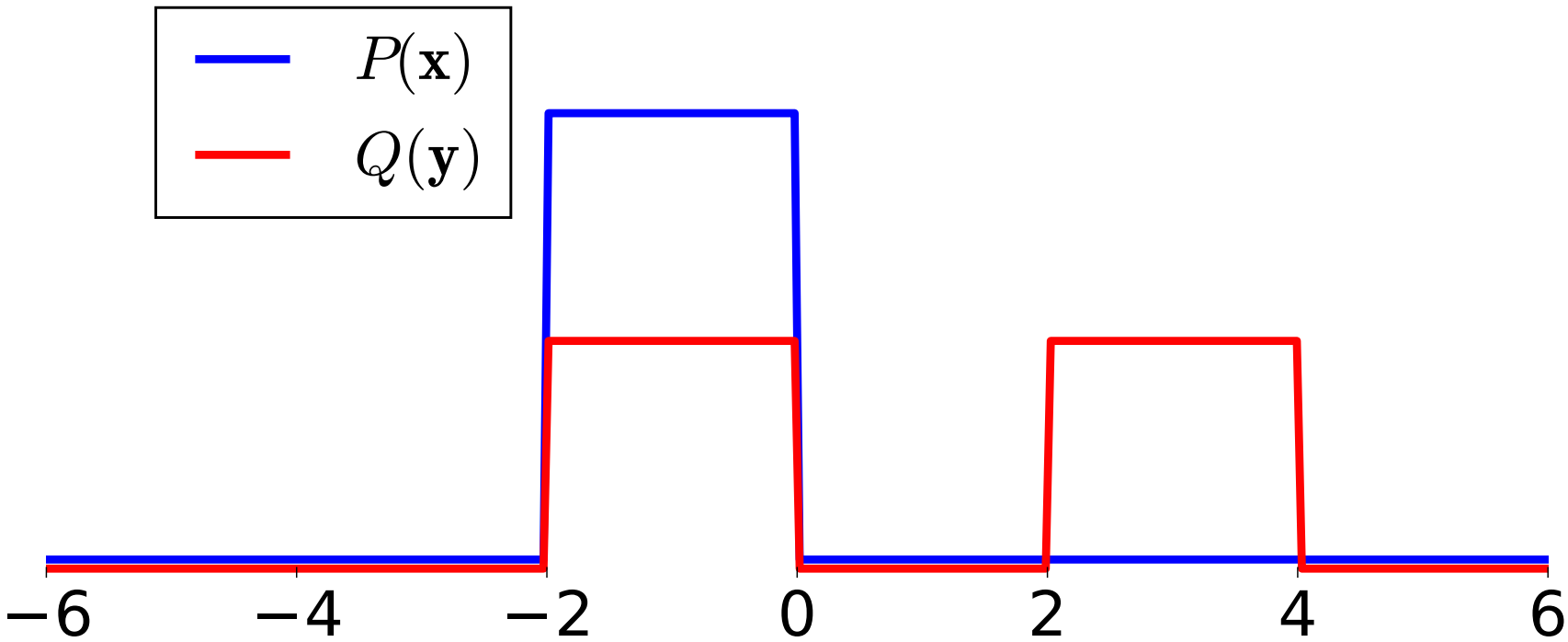
- **Second shrinkage estimator:**

$$\check{\mu}_\lambda := \sum_{i=1}^{n} \frac{\gamma_i}{\gamma_i + \lambda} \langle \hat{\mu}, \phi_i \rangle \phi_i \qquad \mathbb{E}_P\left(k(X, \cdot) \otimes k(X, \cdot)\right) = \sum_{i=1}^{\infty} \gamma_i \phi_i \otimes \phi_i.$$

(suppresses high frequencies more). Related to **kernel Fisher discriminant**, used in testing (Harchaoui, Bach, Moulines 2007).
Alternatively: Bayesian estimate of μ (Flaxman et al., 2016)

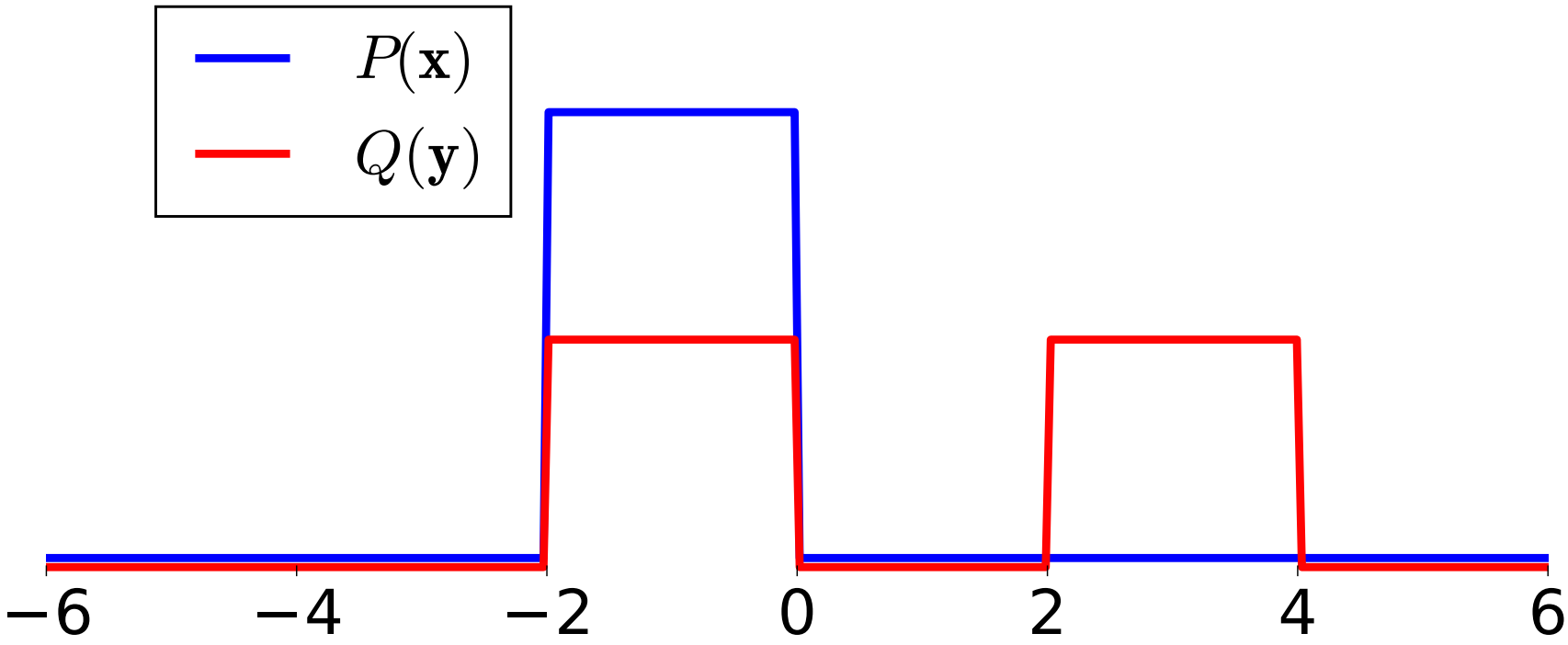
Distinguishing Feature(s)

Where is the best location to observe the difference of $P(\mathbf{x})$ and $Q(\mathbf{y})$? (Jitkrittum, Szabo, Chwialkowski, G., NIPS 2016)



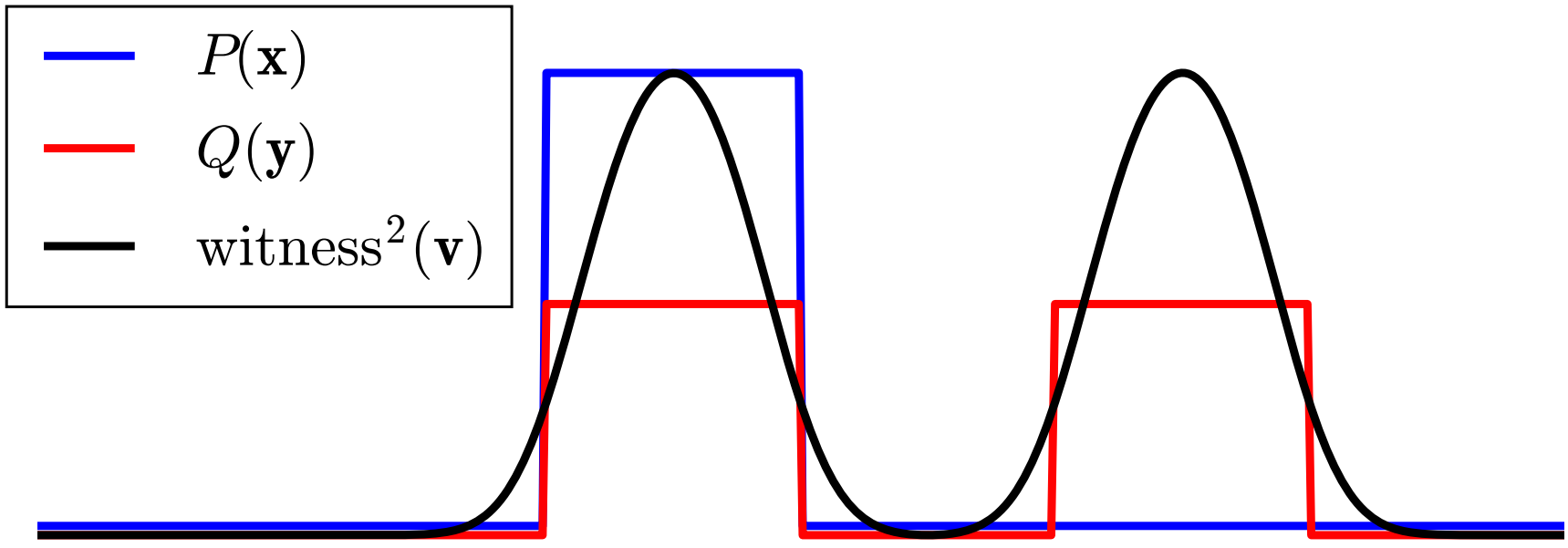
Distinguishing Feature(s)

Where is the best location to observe the difference of $P(\mathbf{x})$ and $Q(\mathbf{y})$? (Jitkrittum, Szabo, Chwialkowski, G., NIPS 2016)



- **Why:** best location = distinguishing feature.

Maximum of the Witness Function



- Two equal maxima
- Are they equally good?

Variance of witness function

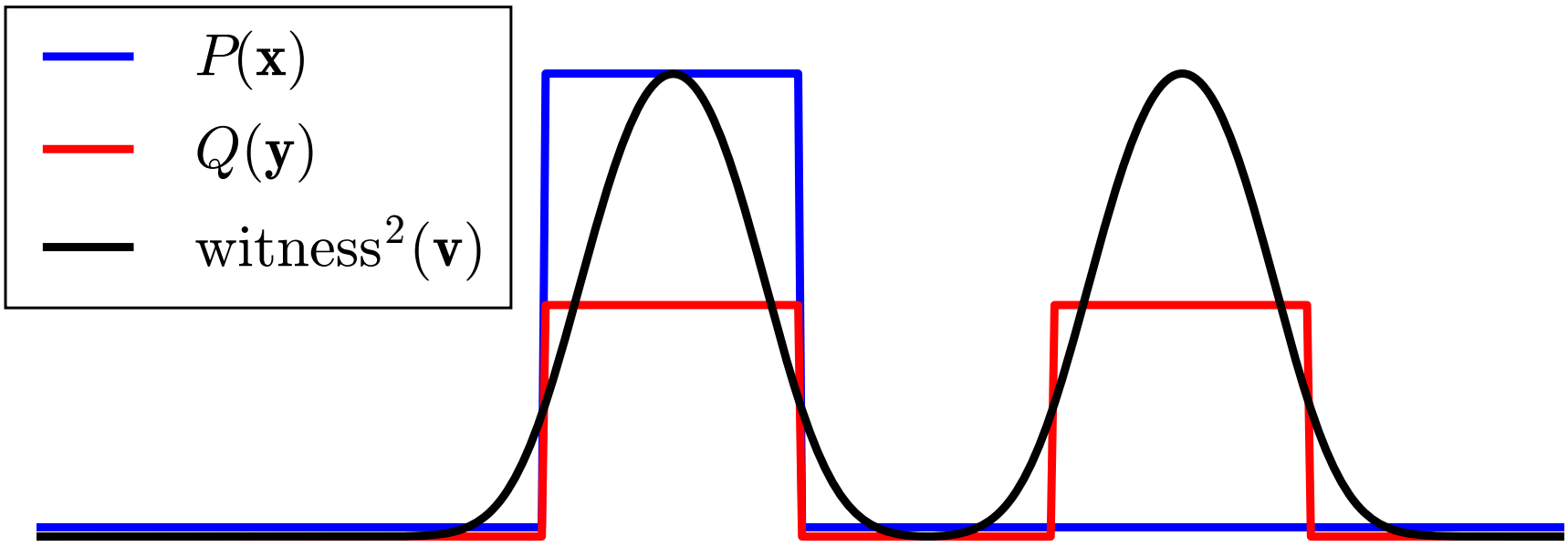
- Variance at $\mathbf{v}$ = variance of X at $\mathbf{v}$ + variance of Y at $\mathbf{v}$.
- ME Statistic: $\hat{\lambda}_n(\mathbf{v}) := n \frac{\text{witness}^2(\mathbf{v})}{\text{variance of } \mathbf{v}}$.

Variance of witness function

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

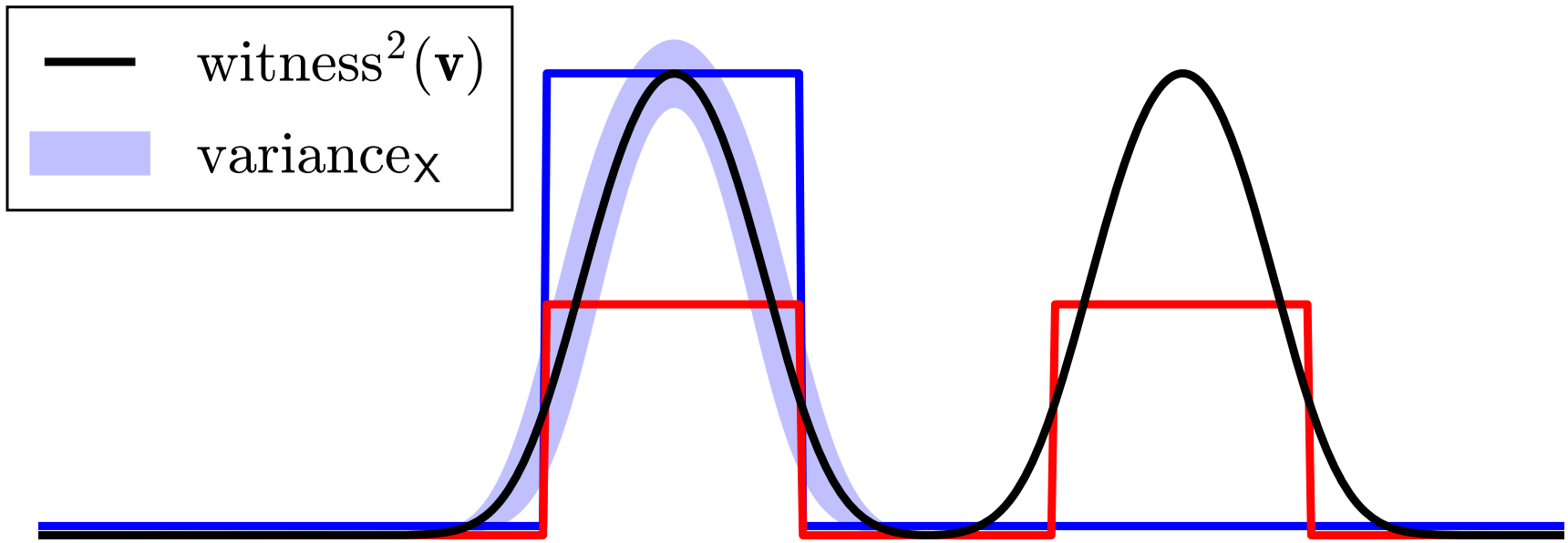
Variance of witness function

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



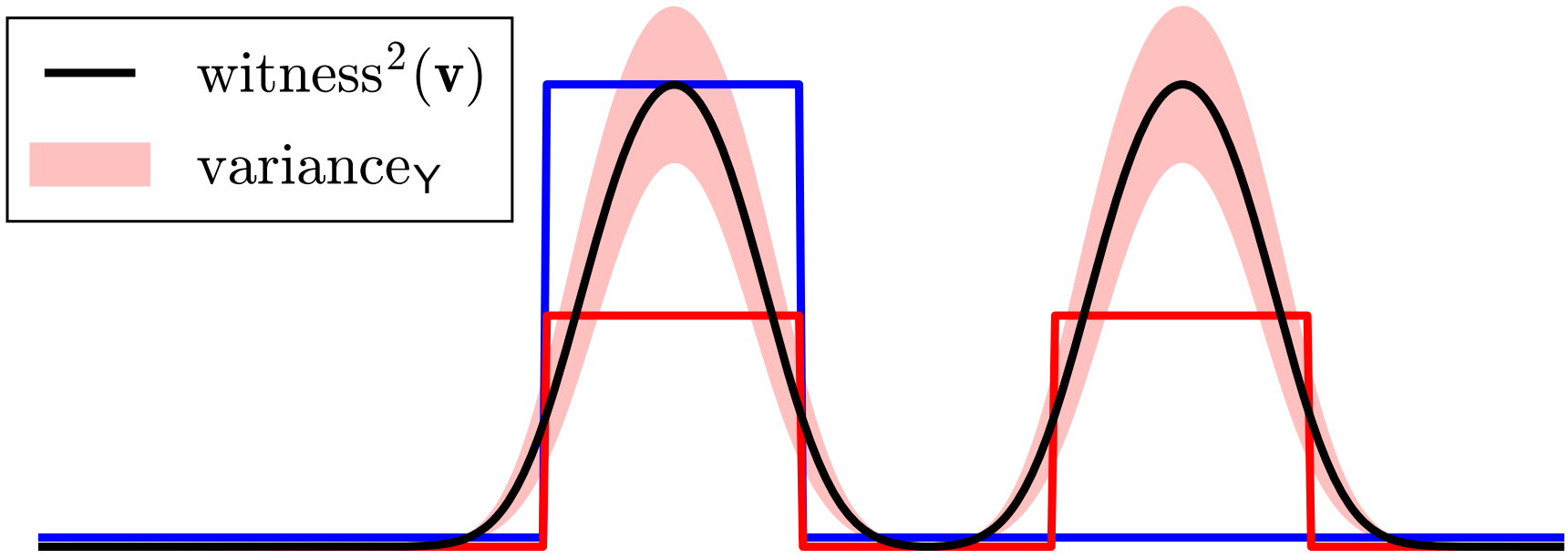
Variance of witness function

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



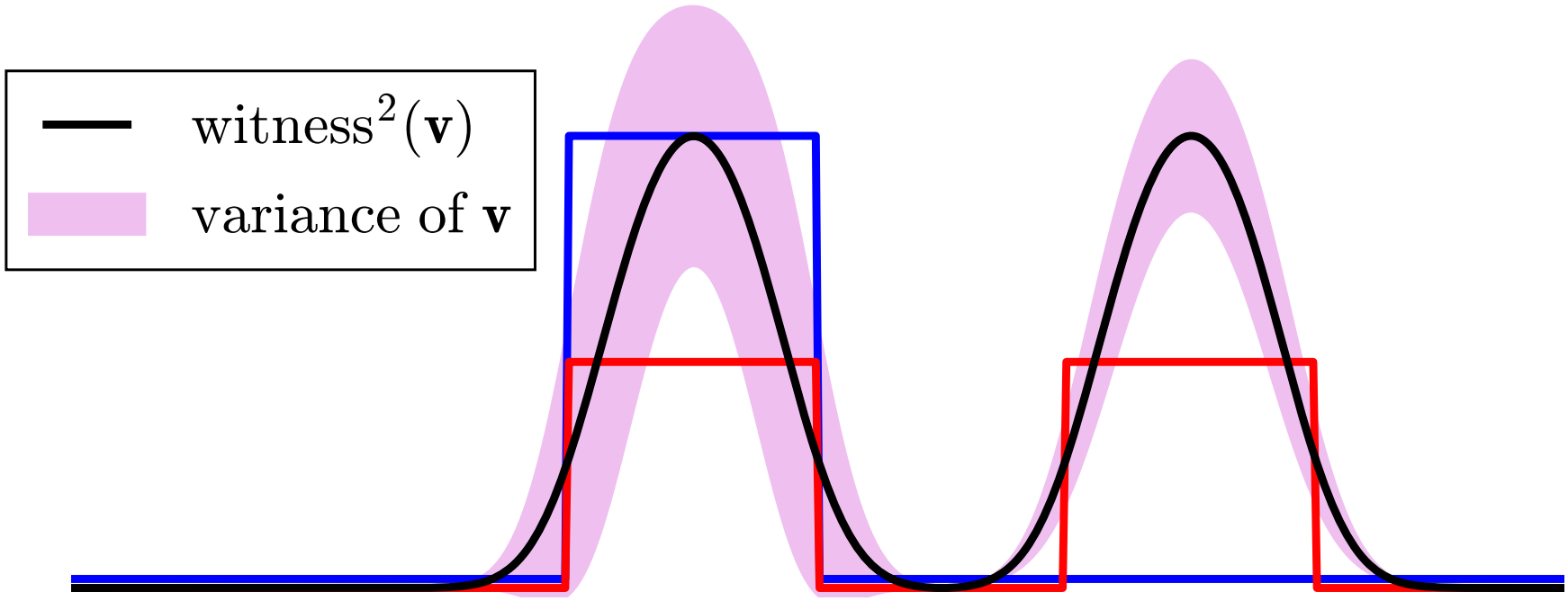
Variance of witness function

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



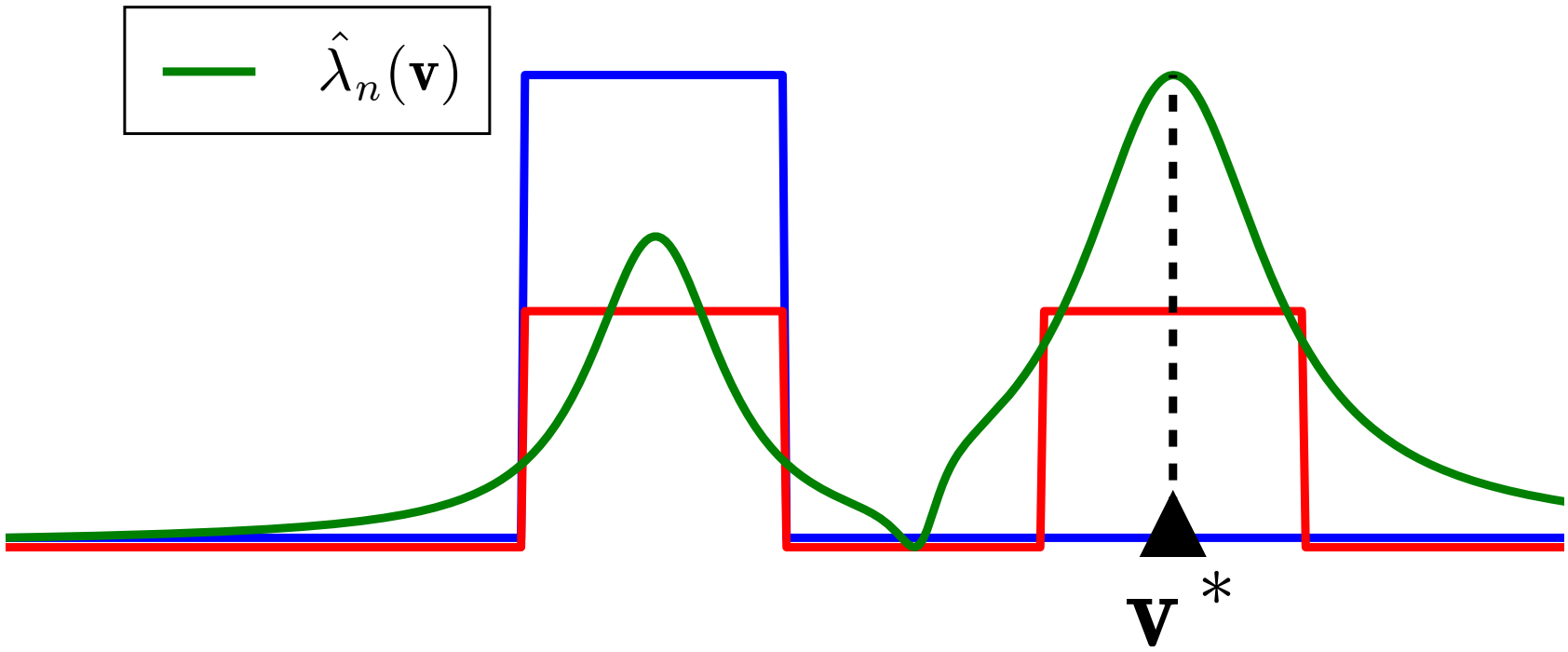
Variance of witness function

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



Variance of witness function

- Variance at $\mathbf{v}$ = variance of X at $\mathbf{v}$ + variance of Y at $\mathbf{v}$.
- 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 $\hat{\lambda}_n$.

Full ME Test Statistic

- Let $\mathcal{V} = \{\mathbf{v}_1, \ldots, \mathbf{v}_J\}$ be the J test locations.
- Define $\mathbf{z}_i = \begin{pmatrix} k(\mathbf{x}_i, \mathbf{v}_1) - k(\mathbf{y}_i, \mathbf{v}_1) \\ \vdots \\ k(\mathbf{x}_i, \mathbf{v}_J) - k(\mathbf{y}_i, \mathbf{v}_J) \end{pmatrix} \in \mathbb{R}^J$.
- Let $\overline{\mathbf{z}}_n := \frac{1}{n}\sum_{i=1}^n \mathbf{z}_i = \begin{pmatrix} \hat{\mu}_P(\mathbf{v}_1) - \hat{\mu}_Q(\mathbf{v}_1) \\ \vdots \\ \hat{\mu}_P(\mathbf{v}_J) - \hat{\mu}_Q(\mathbf{v}_J) \end{pmatrix}$.
- Let $\mathbf{S}_n := \frac{1}{n-1}\sum_{i=1}^n (\mathbf{z}_i - \overline{\mathbf{z}}_n)(\mathbf{z}_i - \overline{\mathbf{z}}_n)^\top \in \mathbb{R}^{J\times J}$.
- Equivalently,
  $(\mathbf{S}_n)_{ij} = \widehat{\mathrm{cov}}_{\mathbf{x}}[k(\mathbf{x}, \mathbf{v}_i), k(\mathbf{x}, \mathbf{v}_j)] + \widehat{\mathrm{cov}}_{\mathbf{y}}[k(\mathbf{y}, \mathbf{v}_i), k(\mathbf{y}, \mathbf{v}_j)]$.
- Then, the statistic

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

  where $\gamma_n > 0$ is a regularization parameter.

Full ME Test Statistic

- Let $\mathcal{V} = \{\mathbf{v}_1, \dots, \mathbf{v}_J\}$ be the J test locations.
- Define $\mathbf{z}_i = \begin{pmatrix} k(\mathbf{x}_i, \mathbf{v}_1) - k(\mathbf{y}_i, \mathbf{v}_1) \\ \vdots \\ k(\mathbf{x}_i, \mathbf{v}_J) - k(\mathbf{y}_i, \mathbf{v}_J) \end{pmatrix} \in \mathbb{R}^J$.
- Let $\overline{\mathbf{z}}_n := \frac{1}{n}\sum_{i=1}^n \mathbf{z}_i = \begin{pmatrix} \hat{\mu}_P(\mathbf{v}_1) - \hat{\mu}_Q(\mathbf{v}_1) \\ \vdots \\ \hat{\mu}_P(\mathbf{v}_J) - \hat{\mu}_Q(\mathbf{v}_J) \end{pmatrix}$.
- Let $\mathbf{S}_n := \frac{1}{n-1}\sum_{i=1}^n (\mathbf{z}_i - \overline{\mathbf{z}}_n)(\mathbf{z}_i - \overline{\mathbf{z}}_n)^\top \in \mathbb{R}^{J \times J}$.
- Equivalently,
  $(\mathbf{S}_n)_{ij} = \widehat{\mathrm{cov}}_{\mathbf{x}}[k(\mathbf{x}, \mathbf{v}_i), k(\mathbf{x}, \mathbf{v}_j)] + \widehat{\mathrm{cov}}_{\mathbf{y}}[k(\mathbf{y}, \mathbf{v}_i), k(\mathbf{y}, \mathbf{v}_j)]$.
- Then, the statistic

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

  where $\gamma_n > 0$ is a regularization parameter.

Lower Bound on Test Power

- Let $\mathcal{K}$ be a kernel class such that $\sup_{k\in\mathcal{K}} \sup_{(\mathbf{x},\mathbf{y})\in\mathcal{X}^2} |k(\mathbf{x},\mathbf{y})| \leq B$.
- Let $\mathbb{V}$ be a collection in which each element is a set of J test locations.
- Assume $\tilde{c} := \sup_{\mathcal{V}\in\mathbb{V},k\in\mathcal{K}} \|\Sigma^{-1}\|_F < \infty$.

Proposition

For large n, the test power $\mathbb{P}_{H_1}\left(\hat{\lambda}_n \geq T_\alpha\right)$ of the ME test satisfies $\mathbb{P}_{H_1}\left(\hat{\lambda}_n \geq T_\alpha\right) \geq L(\lambda_n)$ where

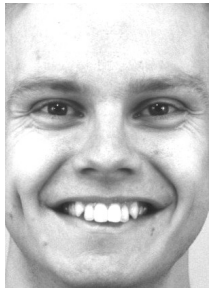
$$L(\lambda_n) := 1 - 2e^{-\xi_1(\lambda_n - T_\alpha)^2/n} - 2e^{-\frac{[\gamma_n(\lambda_n - T_\alpha)(n-1)-\xi_2 n]^2}{\xi_3 n(2n-1)^2}} - 2e^{-[(\lambda_n - T_\alpha)/3 - \bar{c}_3 n\gamma_n]^2\gamma_n^2/\xi_4},$$

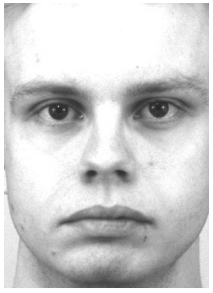
and $\bar{c}_3, \xi_1, \ldots \xi_4$ are positive constants depending on only B, J and $\tilde{c}$. For large n, $L(\lambda_n)$ is increasing in λ_n.

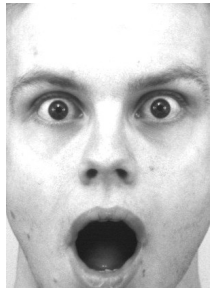
- $\lambda_n := n\mu^\top\Sigma^{-1}\mu$ is the population counterpart of $\hat{\lambda}_n$.
- $\mu = \mathbb{E}_{\mathbf{xy}}[\mathbf{z}_1]$ and $\Sigma = \mathbb{E}_{\mathbf{xy}}[(\mathbf{z}_1 - \mu)(\mathbf{z}_1 - \mu)^\top]$.

Distinguishing Positive/Negative Emotions

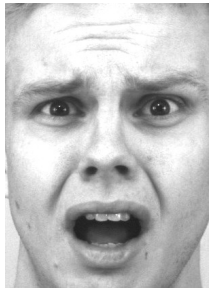
$+$:

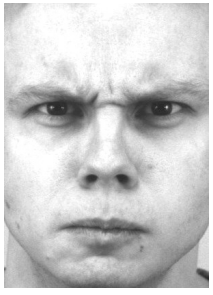

happy

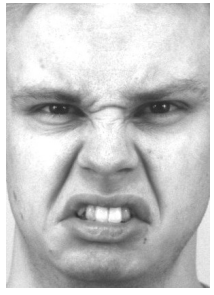

neutral


surprised

$-$:


afraid

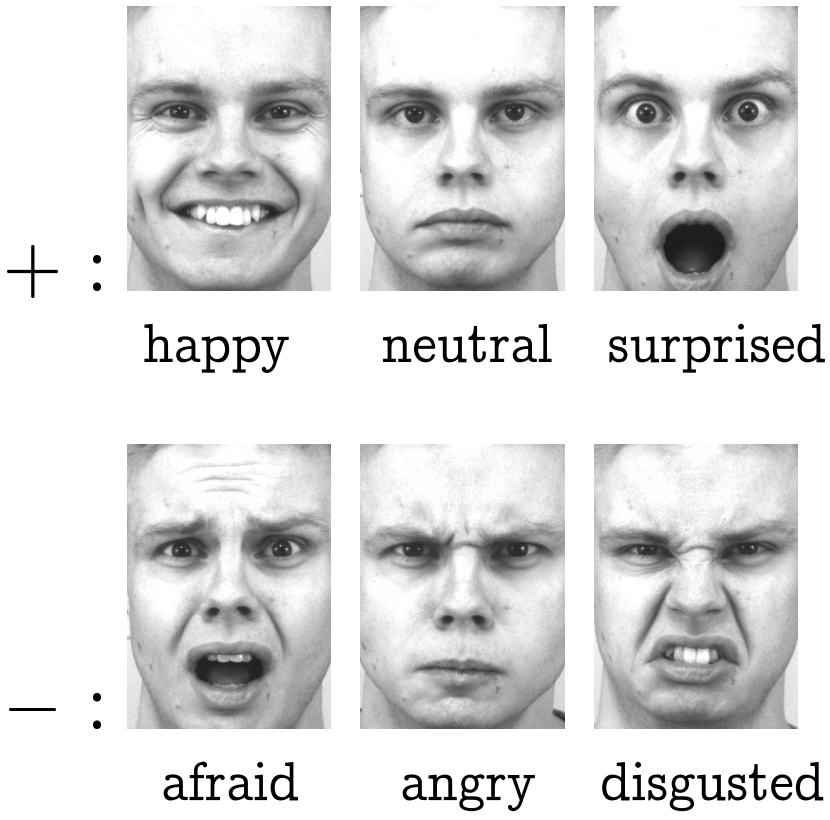

angry

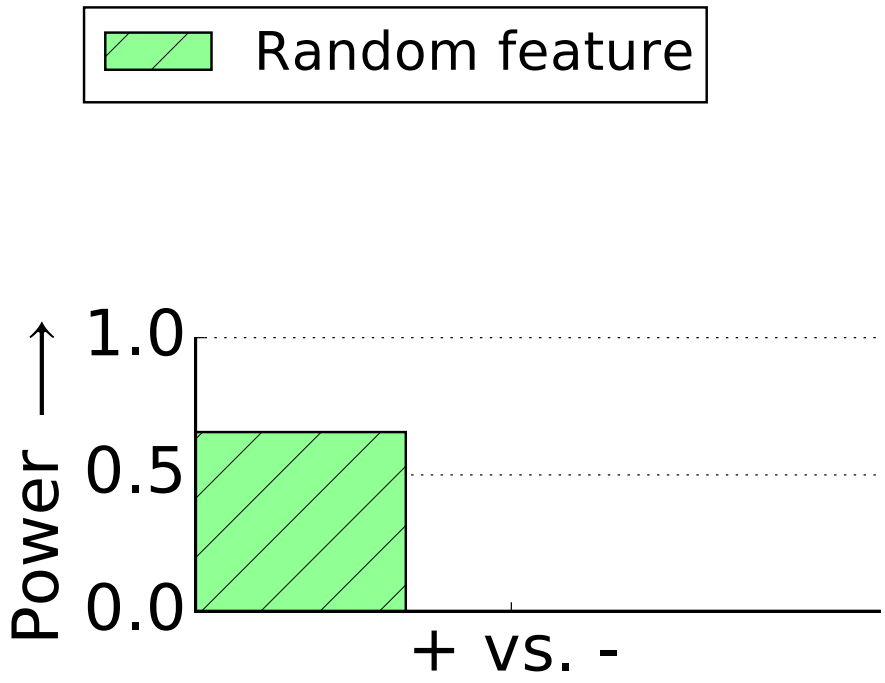

disgusted

- 35 females and 35 males (Lundqvist et al., 1998).
- $48 \times 34 = 1632$ dimensions. Pixel features.
- Sample size: 402.

- The proposed test achieves **maximum test power** in **time** $O(n)$.
- **Informative features:** differences at the nose, and smile lines.

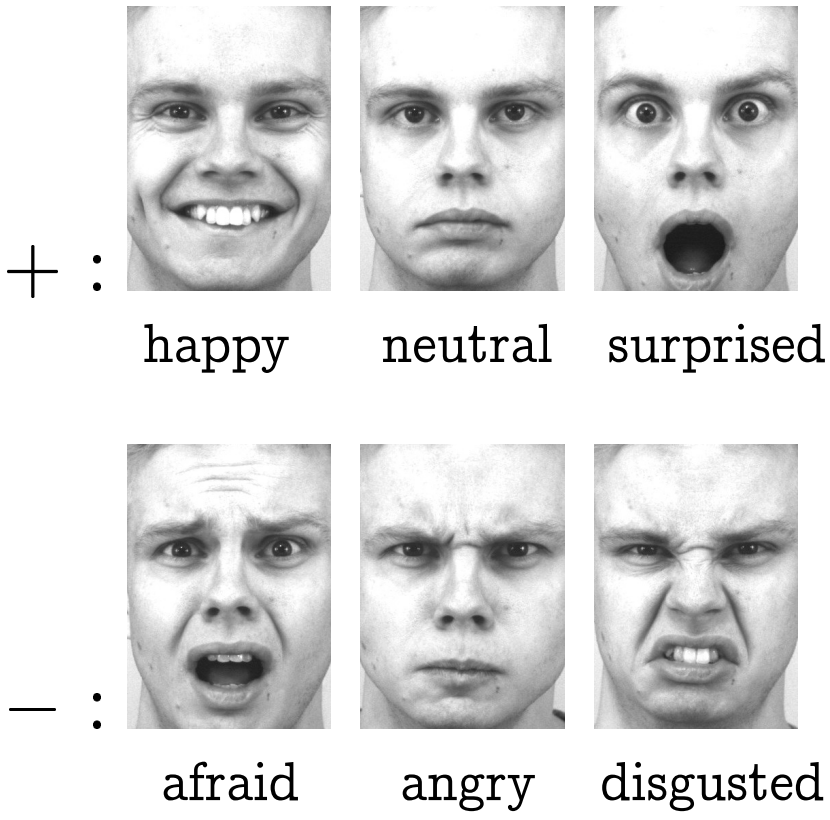
Distinguishing Positive/Negative Emotions

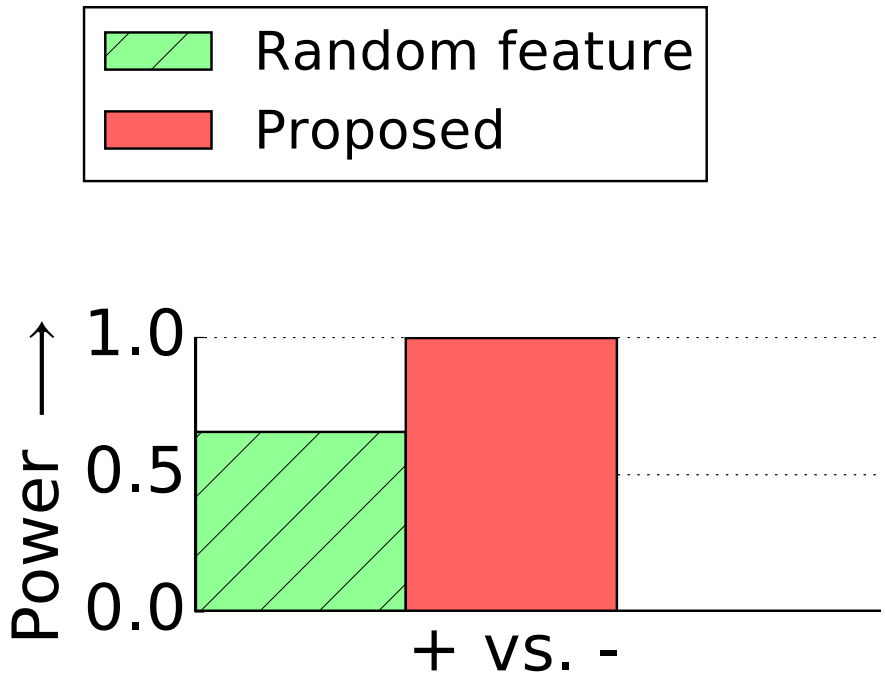




- The proposed test achieves **maximum test power** in **time** $O(n)$.
- **Informative features:** differences at the nose, and smile lines.

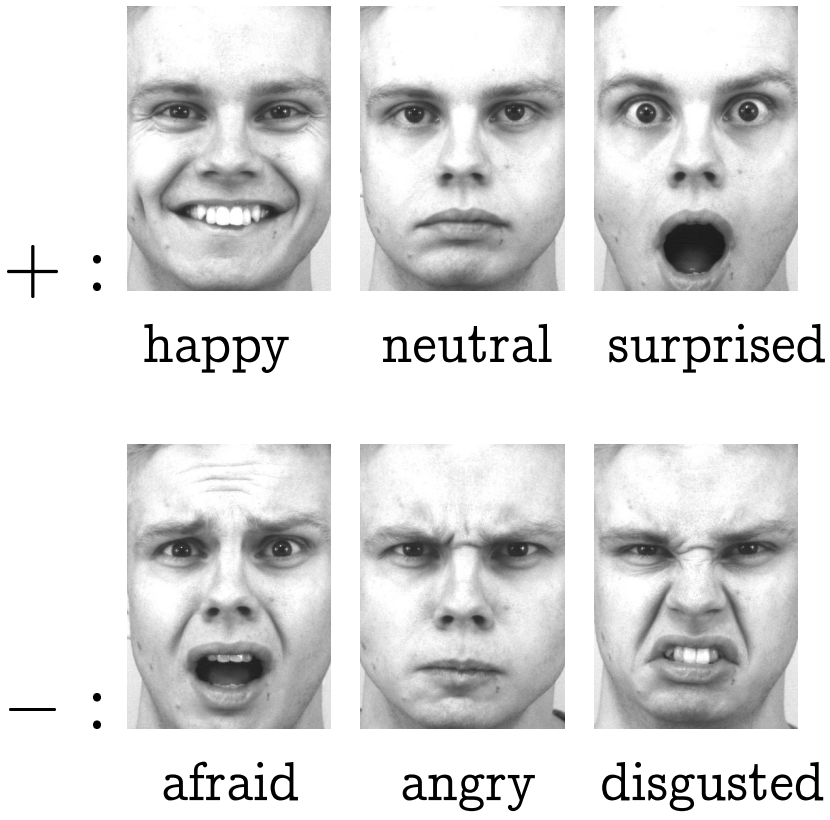
Distinguishing Positive/Negative Emotions

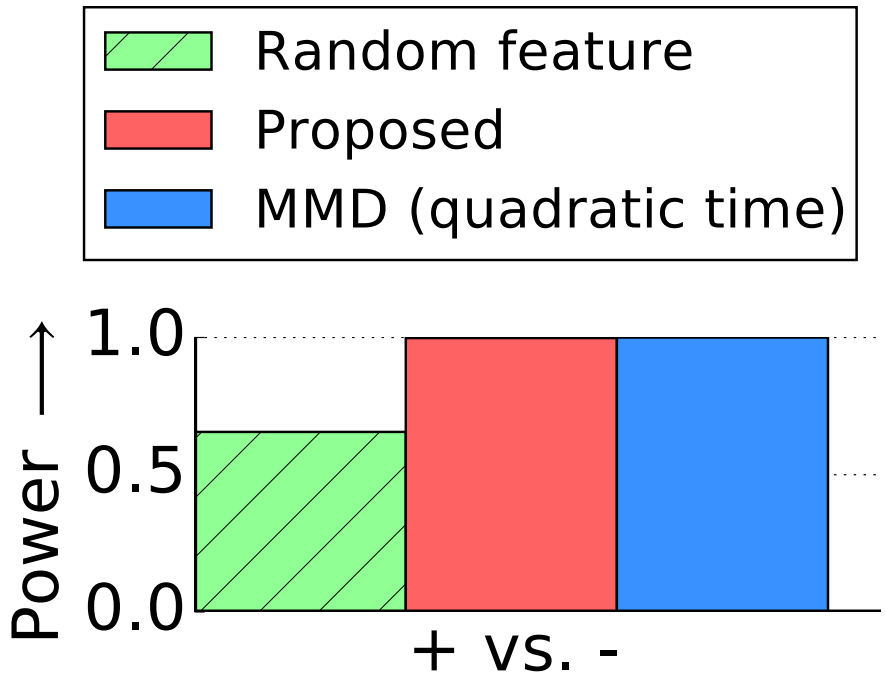




- The proposed test achieves **maximum test power** in **time** $O(n)$.
- **Informative features:** differences at the nose, and smile lines.

Distinguishing Positive/Negative Emotions

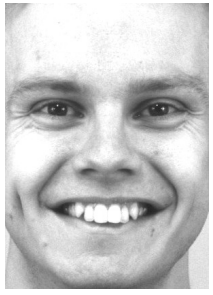


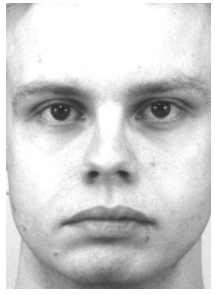


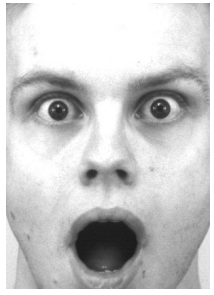
- The proposed test achieves **maximum test power** in **time** $O(n)$.
- Informative features: differences at the nose, and smile lines.

Distinguishing Positive/Negative Emotions

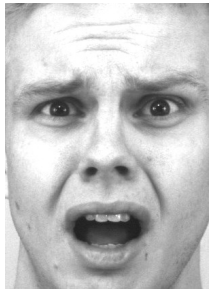
$+$:

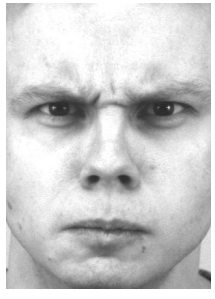

happy

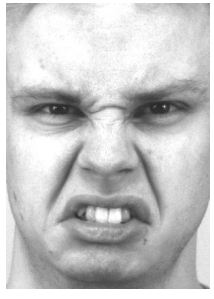

neutral

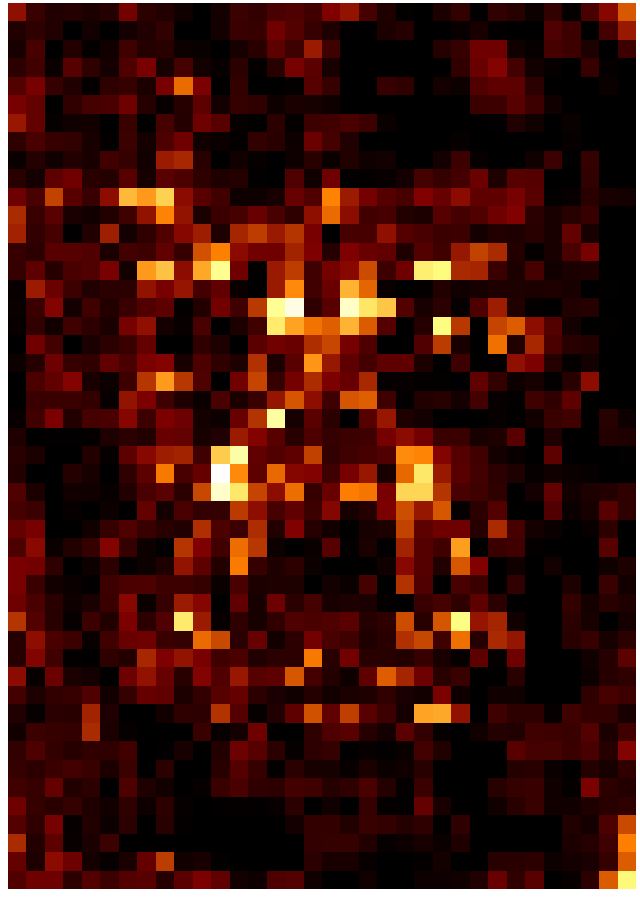

surprised

$-$:


afraid


angry

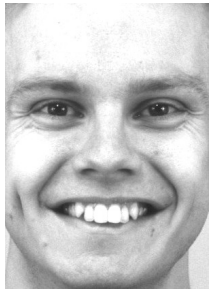

disgusted

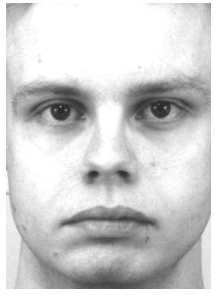

Learned feature

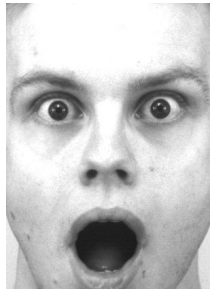
- The proposed test achieves **maximum test power** in **time** $O(n)$.
- **Informative features**: differences at the nose, and smile lines.

Distinguishing Positive/Negative Emotions

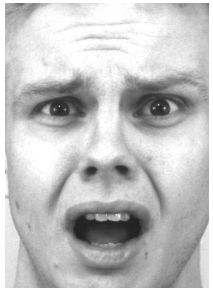
$+$:

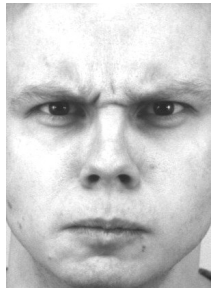

happy

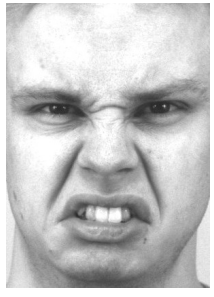

neutral


surprised

$-$:


afraid


angry

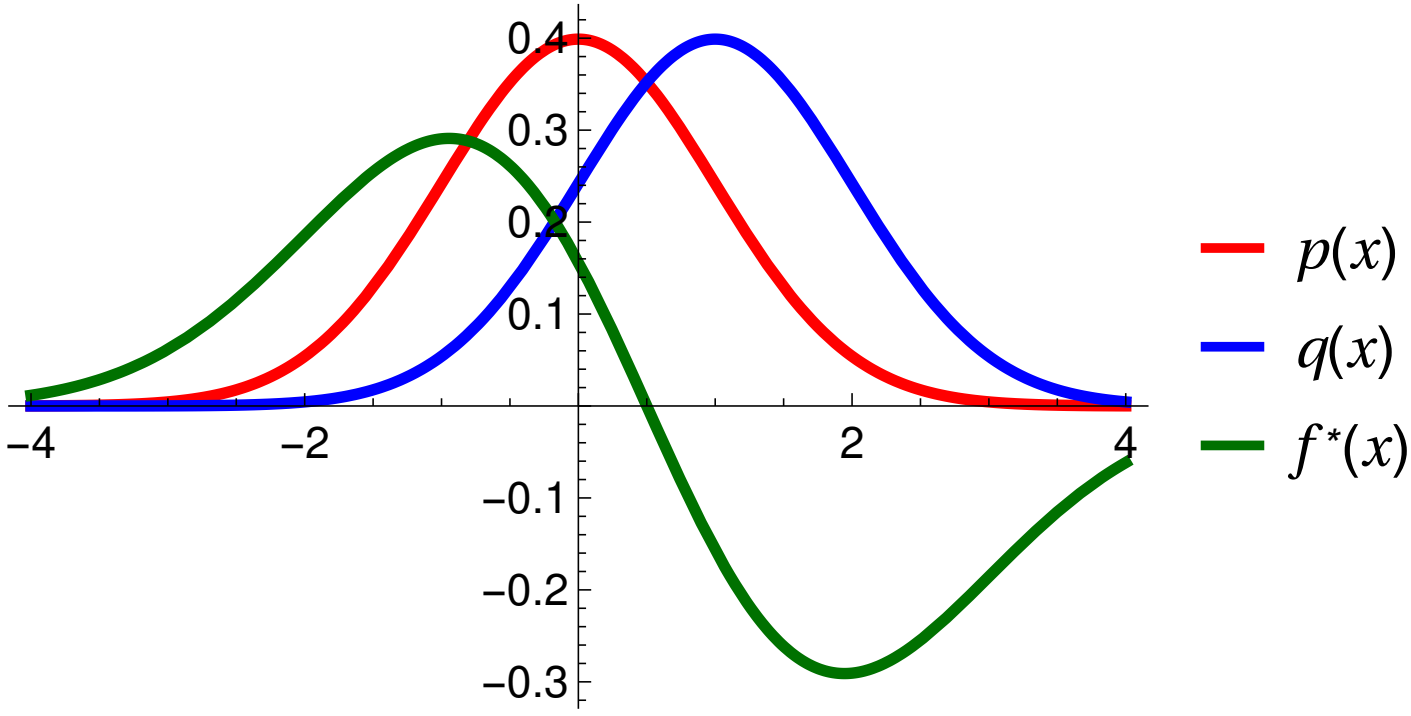

disgusted


Learned feature

- The proposed test achieves **maximum test power** in **time** $O(n)$.
- **Informative features**: differences at the nose, and smile lines.

Statistical model criticism

$$MMD(P, Q, \mathcal{F}) = \sup_{||f||_{\mathcal{F}} \leq 1} [E_Q f - E_p f]$$



Can we compute MMD with samples from Q and a **model** P?

Problem: usualy can't compute $E_p f$ in closed form.

Stein idea

To get rid of $E_p f$ in

$$\sup_{||f||_{\mathcal{F}} \leq 1} \left[E_q f - E_p f \right]$$

we define the **Stein operator**

$$T_p f = \partial_x f + f \left(\partial_x \log p \right)$$

Then (Oates, Girolami, Chopin, 2016),

$$E_P \, T_P f = 0$$

subject to appropriate boundary conditions.

Stein idea: proof sketch

Consider the class

$$G = \{\partial_x f + f \partial_x (\log p) \mid f \in \mathcal{F}\}$$

Given $g \in G$, then (integration by parts)

$$\begin{aligned} E_p g(X) &= E_p \left[\partial_x f(X) + (\partial_x \log p(X)) f(X)\right] \\ &= \int \partial_x f(x) p(x) + f(x) \partial_x p(x) dx \\ &= \int_{-\infty}^{\infty} \partial_x (f(x) p(x)) dx \\ &= f(x) p(x) \Big|_{x=-\infty}^{x=\infty} = 0 \end{aligned}$$

Stein idea: proof sketch

Consider the class

$$G = \{\partial_x f + f \partial_x (\log p) \mid f \in \mathcal{F}\}$$

Given $g \in G$, then (integration by parts)

$$\begin{aligned} E_p g(X) &= E_p \left[\partial_x f(X) + (\partial_x \log p(X)) f(X)\right] \\ &= \int \partial_x f(x) p(x) + f(x) \partial_x p(x) \, dx \\ &= \int_{-\infty}^{\infty} \partial_x (f(x) p(x)) \, dx \\ &= f(x) p(x) \Big|_{x=-\infty}^{x=\infty} = 0 \end{aligned}$$

Maximum Stein Discrepancy

Stein operator

$$T_p f = \partial_x f + f \partial_x (\log p)$$

Maximum Stein Discrepancy (MSD)

$$MSD(p, q, \mathcal{F}) = \sup_{\|g\|_{\mathcal{F}} \leq 1} E_q T_p g - E_p T_p g$$

Maximum Stein Discrepancy

Stein operator

$$T_p f = \partial_x f + f \partial_x (\log p)$$

Maximum Stein Discrepancy (MSD)

$$MSD(p, q, \mathcal{F}) = \sup_{\|g\|_{\mathcal{F}} \leq 1} E_q T_p g - \cancel{E_p T_p g}$$

Maximum Stein Discrepancy

Stein operator

$$T_p f = \partial_x f + f \partial_x (\log p)$$

Maximum Stein Discrepancy (MSD)

$$MSD(p, q, \mathcal{F}) = \sup_{||g||_{\mathcal{F}} \leq 1} E_q T_p g - \cancel{E_p T_p g} = \sup_{||g||_{\mathcal{F}} \leq 1} E_q T_p g$$

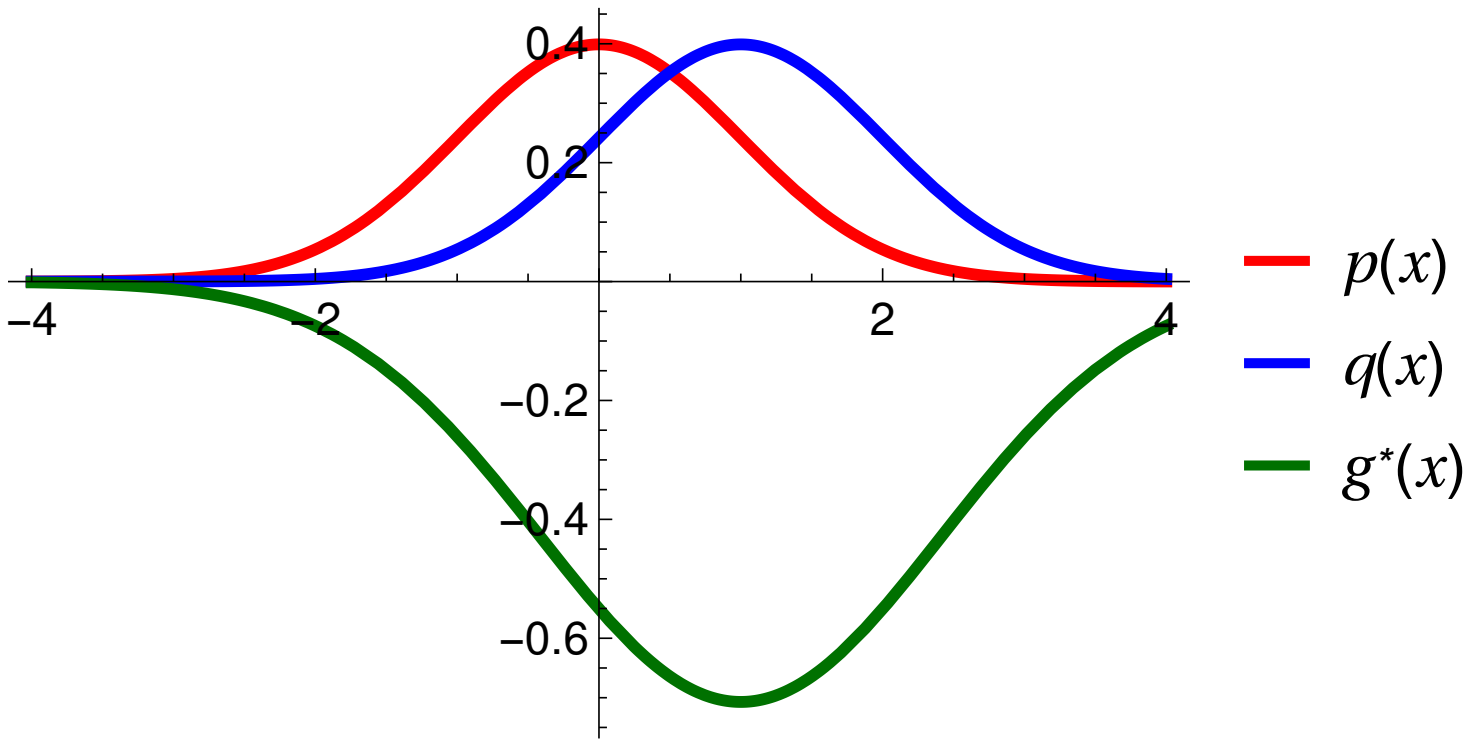
Maximum Stein Discrepancy

Stein operator

$$T_p f = \partial_x f + f \partial_x (\log p)$$

Maximum Stein Discrepancy (MSD)

$$MSD(p, q, \mathcal{F}) = \sup_{\|g\|_{\mathcal{F}} \leq 1} E_q T_p g - \cancel{E_p T_p g} = \sup_{\|g\|_{\mathcal{F}} \leq 1} E_q T_p g$$



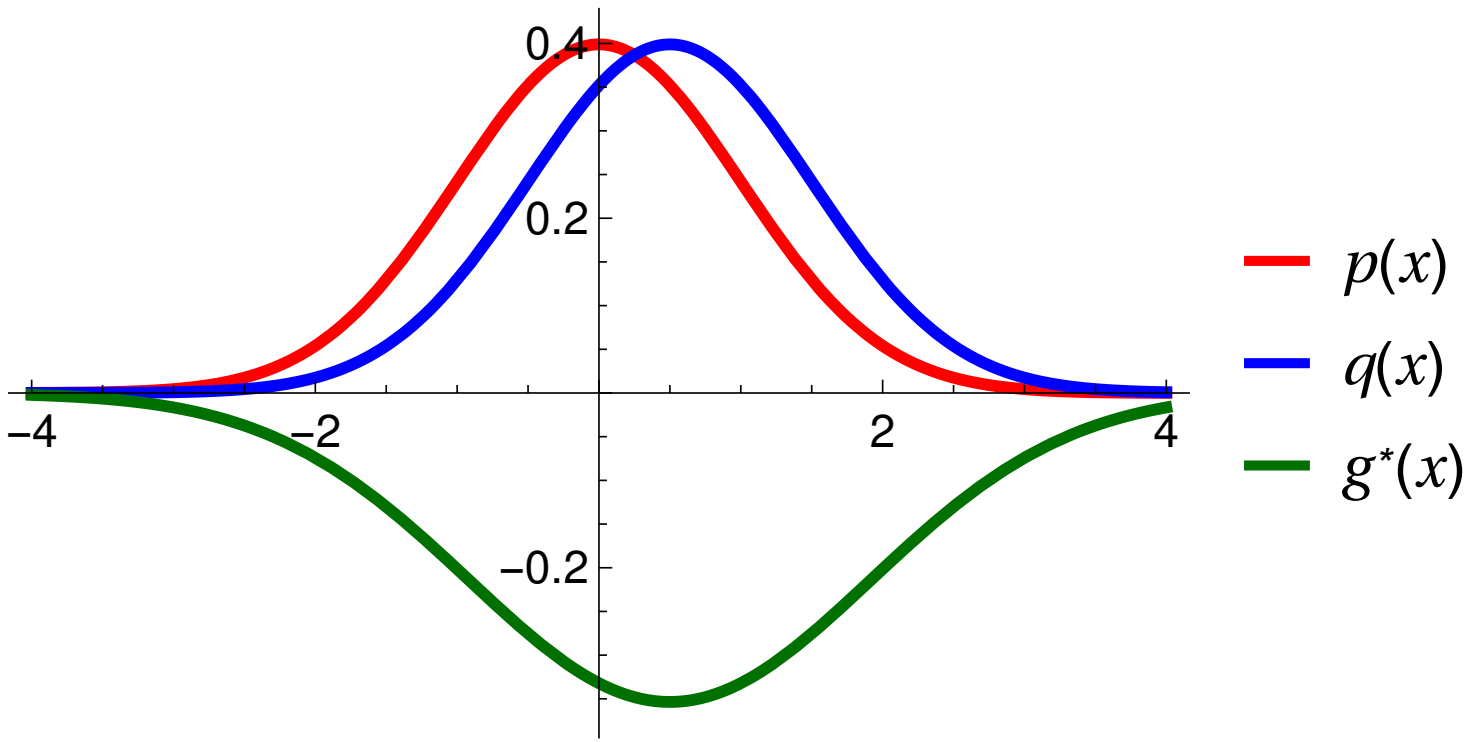
Maximum Stein Discrepancy

Stein operator

$$T_p f = \partial_x f + f \partial_x (\log p)$$

Maximum Stein Discrepancy (MSD)

$$MSD(p, q, \mathcal{F}) = \sup_{\|g\|_{\mathcal{F}} \leq 1} E_q T_p g - \cancel{E_p T_p g} = \sup_{\|g\|_{\mathcal{F}} \leq 1} E_q T_p g$$



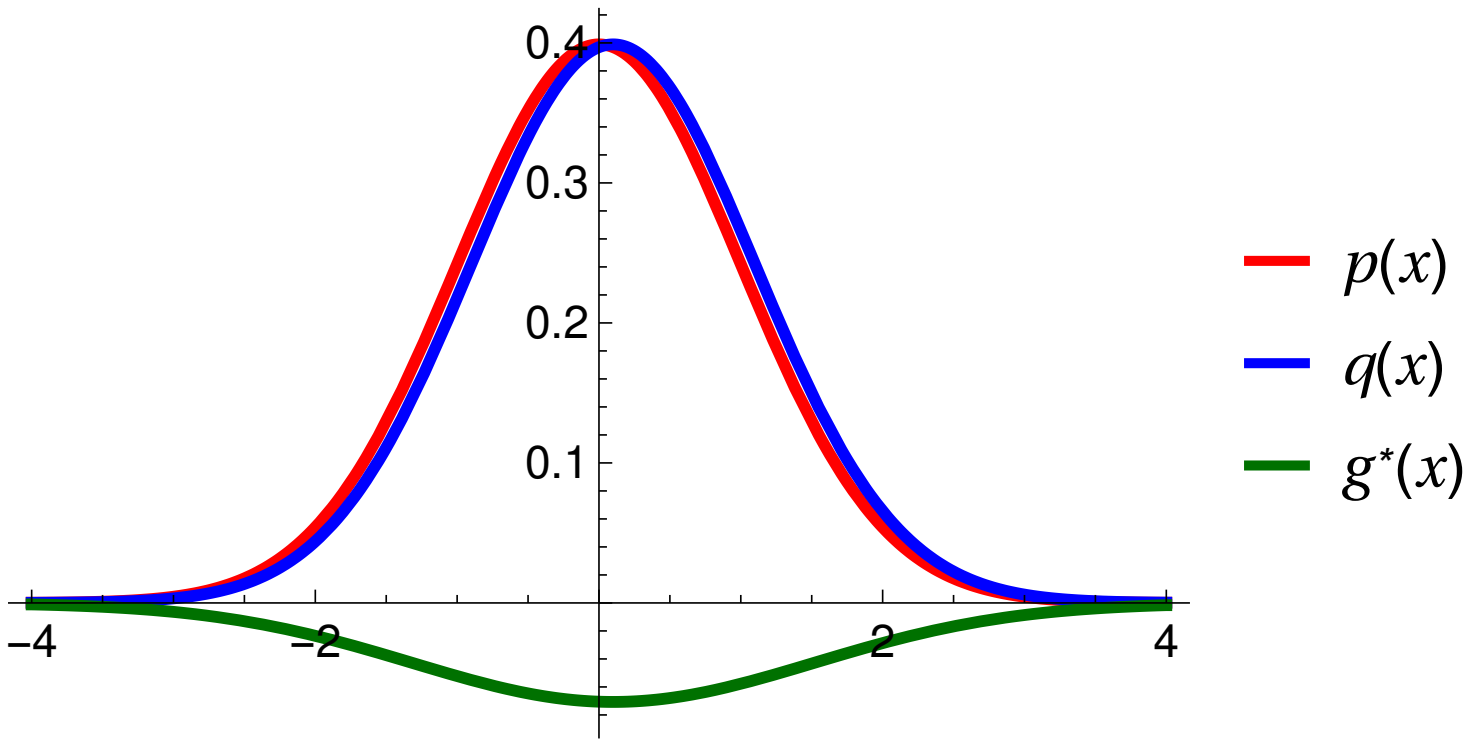
Maximum Stein Discrepancy

Stein operator

$$T_p f = \partial_x f + f \partial_x (\log p)$$

Maximum Stein Discrepancy (MSD)

$$MSD(p, q, \mathcal{F}) = \sup_{\|g\|_{\mathcal{F}} \leq 1} E_q T_p g - \cancel{E_p T_p g} = \sup_{\|g\|_{\mathcal{F}} \leq 1} E_q T_p g$$



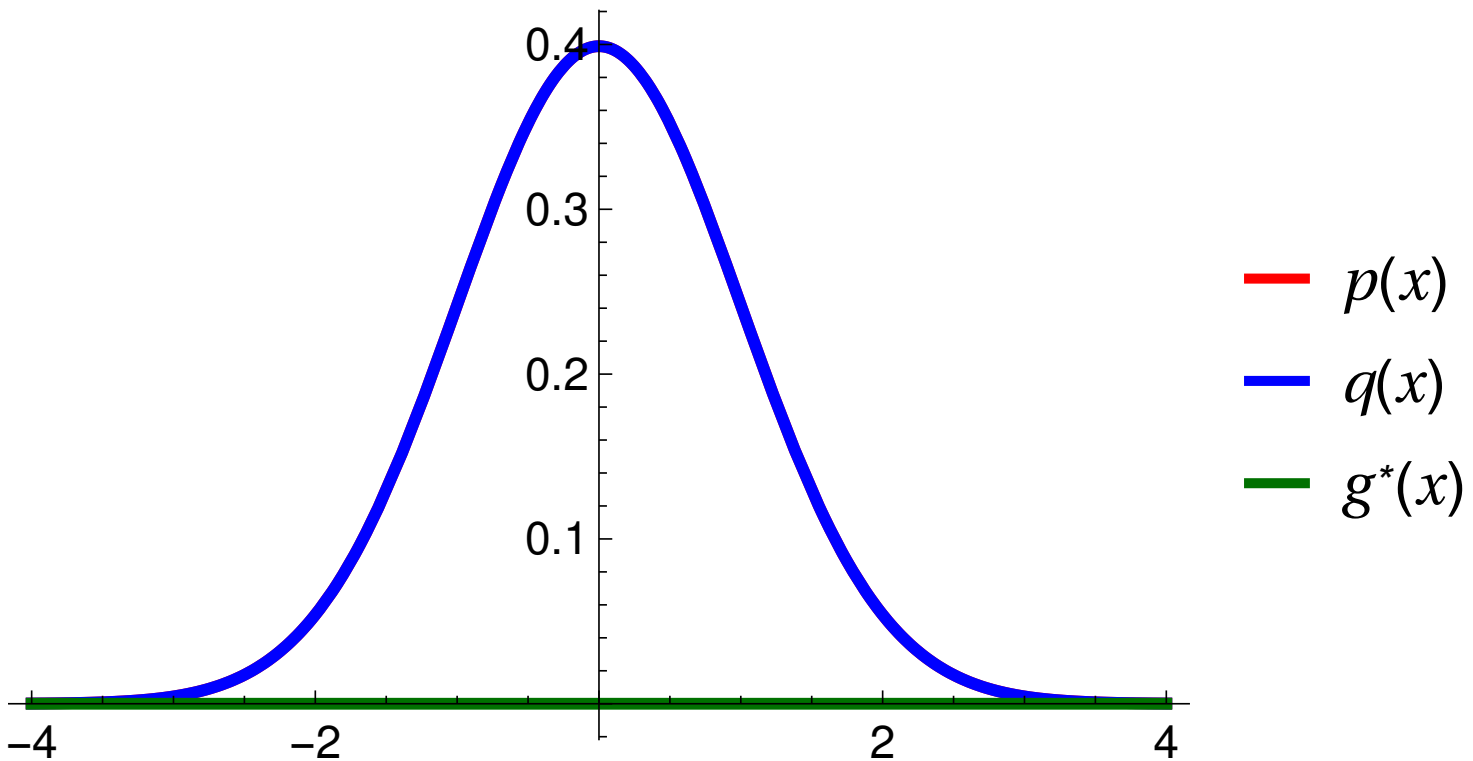
Maximum Stein Discrepancy

Stein operator

$$T_p f = \partial_x f + f \partial_x (\log p)$$

Maximum Stein Discrepancy (MSD)

$$MSD(p, q, \mathcal{F}) = \sup_{\|g\|_{\mathcal{F}} \leq 1} E_q T_p g - \cancel{E_p T_p g} = \sup_{\|g\|_{\mathcal{F}} \leq 1} E_q T_p g$$



Maximum stein discrepancy

Closed-form expression for MSD: given $Z, Z' \sim q$, then (Chwialkowski, Strathmann, G., ICML 2016)

$$\mathrm{MSD}(p, q, \mathcal{F}) = E_q h_p(Z, Z')$$

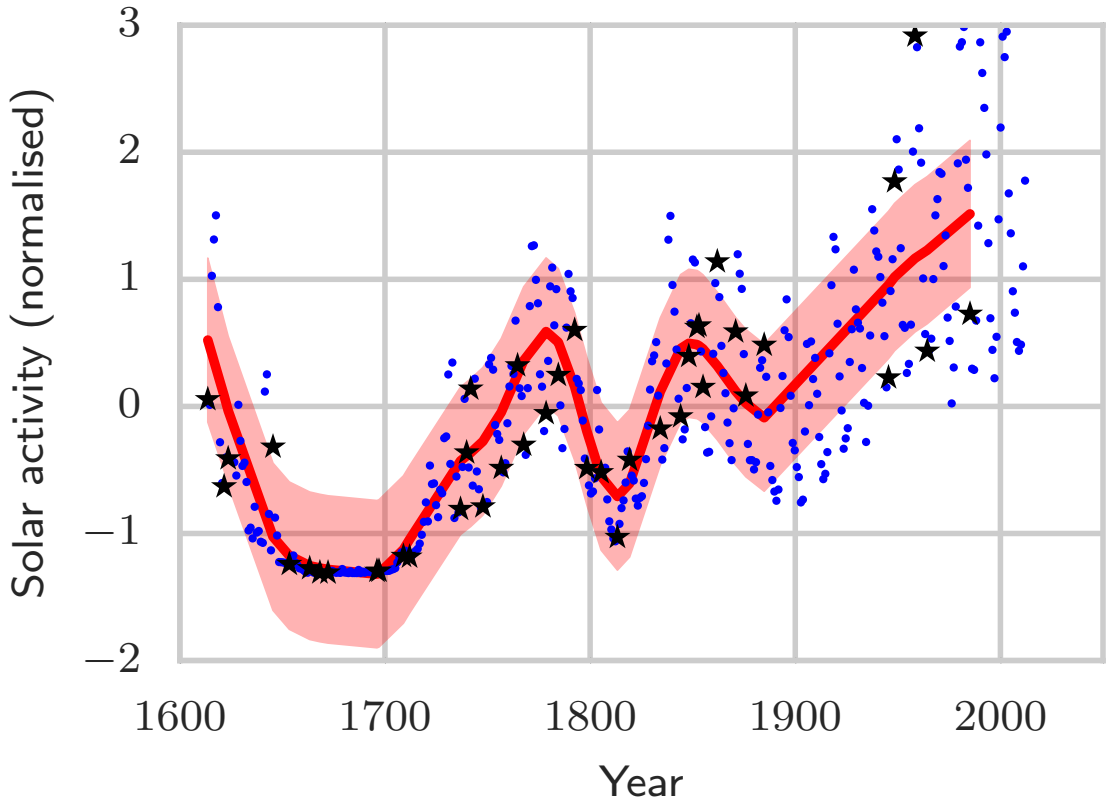
where

$$\begin{aligned} h_p(x, y) := &\partial_x \log p(x) \partial_x \log p(y) k(x, y) \\ &+ \partial_y \log p(y) \partial_x k(x, y) \\ &+ \partial_x \log p(x) \partial_y k(x, y) \\ &+ \partial_x \partial_y k(x, y) \end{aligned}$$

and k is RKHS kernel for $\mathcal{F}$

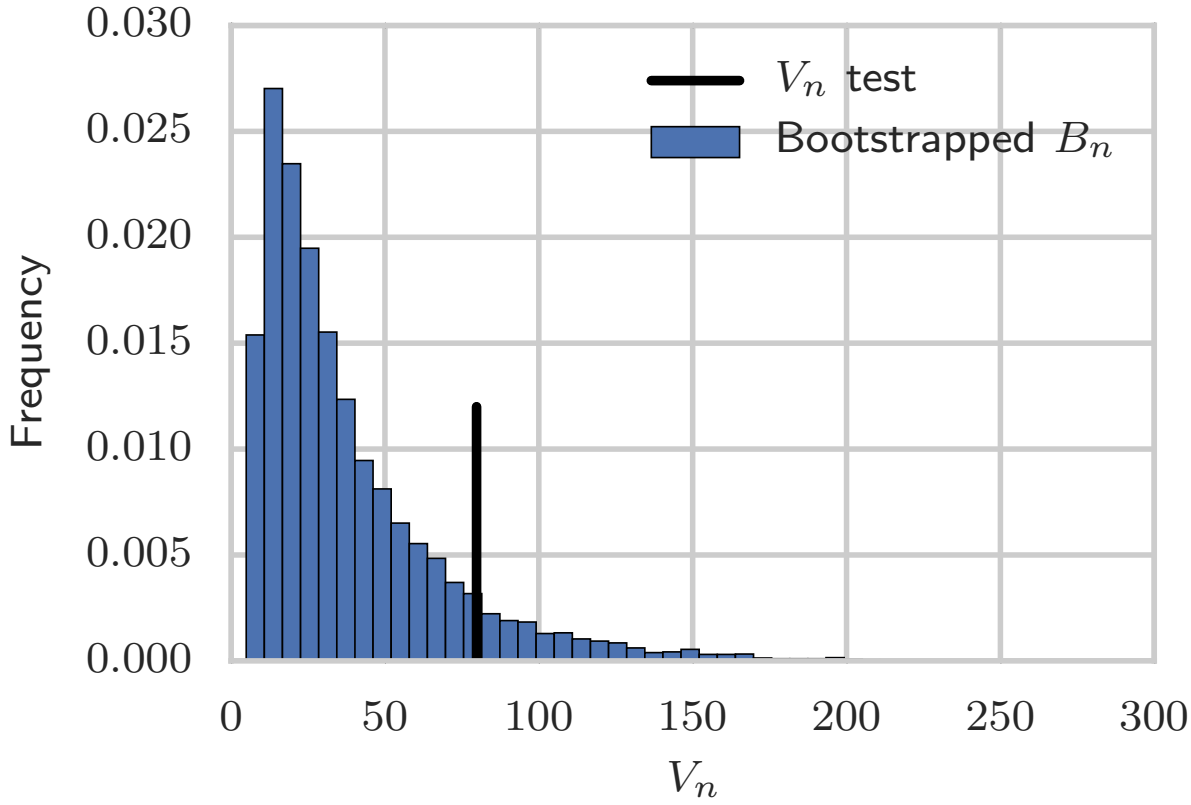
Only depends on kernel and $\partial_x \log p(x)$. Do not need to normalize p, or sample from it.

Statistical model criticism



Test the hypothesis that a Gaussian process **model**, learned from **data** $\star$, is a good fit for the test data (example from Lloyd and Ghahramani, 2015)

Statistical model criticism



Test the hypothesis that a Gaussian process **model**, learned from **data** $\star$, is a good fit for the test data

Co-authors

Students and postdocs:

- Kacper Chwialkowski (at Voleon)
- Wittawat Jitkrittum
- Heiko Strathmann
- Dougal Sutherland

Collaborators

- Kenji Fukumizu
- Krikamol Muandet
- Bernhard Schoelkopf
- Bharath Sriperumbudur
- Zoltan Szabo

Questions?