Representing and comparing probabilities with kernels: Part 3

Arthur Gretton

Gatsby Computational Neuroscience Unit, University College London

MLSS Madrid, 2018
Training GANs with MMD
What is a Generative Adversarial Network (GAN)?

- **Generator (student)**
- **Critic (teacher)**

  - Task: critic must teach generator to draw images (here dogs)
What is a Generative Adversarial Network (GAN)?
What is a Generative Adversarial Network (GAN)?
What is a Generative Adversarial Network (GAN)?
Why is classification not enough?

Classification **not** enough!
Need to compare **sets**
(otherwise student can just produce the same dog over and over)
MMD for GAN critic

Can you use **MMD as a critic** to train GANs?

From ICML 2015:

---

**Generative Moment Matching Networks**

---

Yujia Li¹
Kevin Swersky¹
Richard Zemel¹,²

¹Department of Computer Science, University of Toronto, Toronto, ON, CANADA
²Canadian Institute for Advanced Research, Toronto, ON, CANADA

---

From UAI 2015:

---

**Training generative neural networks via Maximum Mean Discrepancy optimization**

---

Gintare Karolina Dziugaite
University of Cambridge

Daniel M. Roy
University of Toronto

Zoubin Ghahramani
University of Cambridge
MMD for GAN critic

Can you use MMD as a critic to train GANs?

Need better image features.
How to improve the critic witness

- Add convolutional features!
- The critic (teacher) also needs to be trained.
- How to regularise?

MMD GAN Li et al., [NIPS 2017]
Coulomb GAN Unterthiner et al., [ICLR 2018]
WGAN-GP

Wasserstein GAN  Arjovsky et al. [ICML 2017]
WGAN-GP  Gukrajani et al. [NIPS 2017]
**WGAN-GP**

**Wasserstein GAN** Arjovsky et al. [ICML 2017]

**WGAN-GP** Gukrajani et al. [NIPS 2017]

- Given a generator $G_\theta$ with parameters $\theta$ to be trained. Samples $Y \sim G_\theta(Z)$ where $Z \sim R$

- Given critic features $h_\psi$ with parameters $\psi$ to be trained. $f_\psi$ a linear function of $h_\psi$. 
WGAN-GP

Wasserstein GAN  Arjovsky et al. [ICML 2017]
WGAN-GP  Gukrajani et al. [NIPS 2017]

- Given a generator $G_\theta$ with parameters $\theta$ to be trained.

  Samples $Y \sim G_\theta(Z)$ where $Z \sim R$

- Given critic features $h_\psi$ with parameters $\psi$ to be trained. $f_\psi$ a linear function of $h_\psi$.

WGAN-GP gradient penalty:

$$\max_{\psi} \mathbb{E}_{X \sim P} f_\psi(X) - \mathbb{E}_{Z \sim R} f_\psi(G_\theta(Z)) + \lambda \mathbb{E}_{\tilde{X}} \left( \| \nabla_{\tilde{X}} f_\theta(\tilde{X}) \| - 1 \right)^2$$

where

$$\tilde{X} = \gamma x_i + (1 - \gamma) G_\psi(z_j)$$

$\gamma \sim \mathcal{U}([0, 1])$, $x_i \in \{x_l\}_{l=1}^m$, $z_j \in \{z_l\}_{l=1}^n$
The (W)MMD

Train MMD critic features with the witness function gradient penalty
Binkowski, Sutherland, Arbel, G. [ICLR 2018], Bellemare et al. [2017] for energy distance:

$$
\max_{\psi} \text{MMD}^2(h_{\psi}(X), h_{\psi}(G_\theta(Z))) + \lambda E_{\tilde{X}} \left( \| \nabla_{\tilde{X}} f_{\psi}(\tilde{X}) \| - 1 \right)^2
$$

where

$$
f_{\psi}(\cdot) = \frac{1}{m} \sum_{i=1}^{m} k(h_{\psi}(x_i), \cdot) - \frac{1}{n} \sum_{j=1}^{n} k(h_{\psi}(G_\theta(z_j))), \cdot)
$$

Remark by Bottou et al. (2017): gradient penalty modifies the function class. So critic is not an MMD in RKHS $\mathcal{F}$.
MMD for GAN critic: revisited

From ICLR 2018:

DEMYSTIFYING MMD GANs

Mikołaj Binkowski*
Department of Mathematics
Imperial College London
mikbinkowski@gmail.com

Dougal J. Sutherland, Michael Arbel & Arthur Gretton
Gatsby Computational Neuroscience Unit
University College London
{dougal, michael.n.arbel, arthur.gretton}@gmail.com
MMD for GAN critic: revisited

Samples are better!
MMD for GAN critic: revisited

Samples are better!

Can we do better still?
Convergence issues for WGAN-GP penalty

WGAN-GP style gradient penalty may not converge near solution
Nagarajan and Kolter [NIPS 2017], Mescheder et al. [ICML 2018], Balduzzi et al. [ICML 2018]

The Dirac-GAN

\[ P = \delta_0 \quad Q = \delta_\theta \quad f_\psi(x) = \psi \cdot x \]

Figure from Mescheder et al. [ICML 2018]
Convergence issues for WGAN-GP penalty

WGAN-GP style gradient penalty may not converge near solution
Nagarajan and Kolter [NIPS 2017], Mescheder et al. [ICML 2018], Balduzzi et al. [ICML 2018]

The Dirac-GAN

\[ P = \delta_0 \quad Q = \delta_{\theta} \quad f_{\psi}(x) = \psi \cdot x \]
A better gradient penalty

- New MMD GAN witness regulariser (just accepted, NIPS 2018)
  Arbel, Sutherland, Binkowski, G. [NIPS 2018]
- Based on semi-supervised learning regulariser Bousquet et al. [NIPS 2004]
- Related to Sobolev GAN Mroueh et al. [ICLR 2018]
A better gradient penalty

- New MMD GAN witness regulariser (just accepted, NIPS 2018)
  Arbel, Sutherland, Binkowski, G. [NIPS 2018]
- Based on semi-supervised learning regulariser Bousquet et al. [NIPS 2004]
- Related to Sobolev GAN Mroueh et al. [ICLR 2018]
A better gradient penalty

- New MMD GAN witness regulariser (just accepted, NIPS 2018)  
  Arbel, Sutherland, Binkowski, G. [NIPS 2018]
- Based on semi-supervised learning regulariser Bousquet et al. [NIPS 2004]
- Related to Sobolev GAN Mroueh et al. [ICLR 2018]

Modified witness function:

\[
\widehat{MMD} := \sup_{\| f \|_S \leq 1} \left[ \mathbb{E}_P f(X) - \mathbb{E}_Q f(Y) \right]
\]

where

\[
\| f \|_S^2 = \| f \|_{L_2(P)}^2 + \| \nabla f \|_{L_2(P)}^2 + \lambda \| f \|_k^2
\]

- L₂ norm control
- Gradient control
- RKHS smoothness
A better gradient penalty

- New MMD GAN witness regulariser (just accepted, NIPS 2018)
  Arbel, Sutherland, Binkowski, G. [NIPS 2018]
- Based on **semi-supervised learning** regulariser Bousquet et al. [NIPS 2004]
- Related to **Sobolev GAN** Mroueh et al. [ICLR 2018]

Modified witness function:

\[
\overline{\text{MMD}} := \sup_{\|f\|_S \leq 1} \left[ \mathbb{E}_P f(X) - \mathbb{E}_Q f(Y) \right]
\]

where

\[
\|f\|^2_S = \|f\|^2_{L_2(P)} + \|\nabla f\|^2_{L_2(P)} + \lambda \|f\|^2_k
\]

**Problem:** not computationally feasible: \( O(n^3) \) per iteration.
A better gradient penalty

- New MMD GAN witness regulariser (just accepted, NIPS 2018)
  Arbel, Sutherland, Binkowski, G. [NIPS 2018]
- Based on semi-supervised learning regulariser Bousquet et al. [NIPS 2004]
- Related to Sobolev GAN Mroueh et al. [ICLR 2018]

The scaled MMD:

\[
SMMD = \sigma_{k,P,\lambda} \ MMD
\]

where

\[
\sigma_{k,P,\lambda} = \left( \lambda + \int k(x, x) \, dP(x) + \sum_{i=1}^{d} \int \partial_i \partial_{i+d} k(x, x) \, dP(x) \right)^{-1/2}
\]

Replace expensive constraint with cheap upper bound:

\[
\|f\|_S^2 \leq \sigma_{k,P,\lambda}^{-1} \|f\|_k^2
\]
A better gradient penalty

- New MMD GAN witness regulariser (just accepted, NIPS 2018)
  Arbel, Sutherland, Binkowski, G. [NIPS 2018]
- Based on semi-supervised learning regulariser Bousquet et al. [NIPS 2004]
- Related to Sobolev GAN Mroueh et al. [ICLR 2018]

The scaled MMD:

\[ SMMD = \sigma_{k,P,\lambda} \ MMD \]

where

\[ \sigma_{k,P,\lambda} = \left( \lambda + \int k(x, x) dP(x) + \sum_{i=1}^{d} \int \partial_i \partial_{i+d} k(x, x) dP(x) \right)^{-1/2} \]

Replace expensive constraint with cheap upper bound:

\[ \|f\|_S^2 \leq \sigma_{k,P,\lambda}^{-1} \|f\|_k^2 \]

Idea: rather than regularise the critic or witness function, regularise features directly
Evaluation and experiments
The inception score? Salimans et al. [NIPS 2016]

Based on the classification output $p(y|x)$ of the inception model Szegedy et al. [ICLR 2014],

$$E_x \exp KL(P(y|X)||P(y)).$$

High when:

- predictive label distribution $P(y|x)$ has low entropy (good quality images)
- label entropy $P(y)$ is high (good variety).
Evaluation of GANs

The inception score?  Salimans et al. [NIPS 2016]

Based on the classification output $p(y|x)$ of the inception model Szegedy et al. [ICLR 2014],

$$E_X \exp KL(P(y|X) \| P(y)).$$

High when:

- predictive label distribution $P(y|x)$ has low entropy (good quality images)
- label entropy $P(y)$ is high (good variety).

Problem: relies on a trained classifier! Can’t be used on new categories (celeb, bedroom...).
Evaluation of GANs

The Frechet inception distance?  Heusel et al. [NIPS 2017]

Fits Gaussians to features in the inception architecture (pool3 layer):

\[
FID(P, Q) = \|\mu_P - \mu_Q\|^2 + \text{tr}(\Sigma_P) + \text{tr}(\Sigma_Q) - 2\text{tr}\left((\Sigma_P \Sigma_Q)^{\frac{1}{2}}\right)
\]

where \(\mu_P\) and \(\Sigma_P\) are the feature mean and covariance of \(P\)
Evaluation of GANs

The Frechet inception distance? Heusel et al. [NIPS 2017]

Fits Gaussians to features in the inception architecture (pool3 layer):

\[
FID(P, Q) = \|\mu_P - \mu_Q\|^2 + \text{tr}(\Sigma_P) + \text{tr}(\Sigma_Q) - 2\text{tr}\left((\Sigma_P \Sigma_Q)^{\frac{1}{2}}\right)
\]

where \(\mu_P\) and \(\Sigma_P\) are the feature mean and covariance of \(P\)

Problem: bias. For finite samples can consistently give incorrect answer.

- Bias demo,
  CIFAR-10 train vs test
Evaluation of GANs

The FID can give the wrong answer in theory.

Assume $m$ samples from $P$ and $n \to \infty$ samples from $Q$.

Given two alternatives:

$$P_1 \sim N(0, (1 - m^{-1})^2) \quad P_2 \sim N(0, 1) \quad Q \sim N(0, 1).$$

Clearly,

$$FID(P_1, Q) = \frac{1}{m^2} > FID(P_2, Q) = 0$$

Given $m$ samples from $P_1$ and $P_2$,

$$FID(\hat{P}_1, Q) < FID(\hat{P}_2, Q).$$
Evaluation of GANs

The FID can give the wrong answer in theory.
Assume \( m \) samples from \( P \) and \( n \to \infty \) samples from \( Q \).

Given two alternatives:

\[
P_1 \sim \mathcal{N}(0, (1 - m^{-1})^2) \quad P_2 \sim \mathcal{N}(0, 1) \quad Q \sim \mathcal{N}(0, 1).
\]

Clearly,

\[
FID(P_1, Q) = \frac{1}{m^2} > FID(P_2, Q) = 0
\]

Given \( m \) samples from \( P_1 \) and \( P_2 \),

\[
FID(\hat{P}_1, Q) < FID(\hat{P}_2, Q)
\]
Evaluation of GANs

The FID can give the **wrong answer in theory**.

Assume $m$ samples from $P$ and $n \to \infty$ samples from $Q$.

Given two alternatives:

\[
P_1 \sim \mathcal{N}(0, (1 - m^{-1})^2) \quad P_2 \sim \mathcal{N}(0, 1) \quad Q \sim \mathcal{N}(0, 1).
\]

Clearly,

\[
FID(P_1, Q) = \frac{1}{m^2} > FID(P_2, Q) = 0
\]

Given $m$ samples from $P_1$ and $P_2$,

\[
FID(\hat{P}_1, Q) < FID(\hat{P}_2, Q).
\]
Evaluation of GANs

The FID can give the wrong answer in theory.
Assume $m$ samples from $P$ and $n \to \infty$ samples from $Q$.
Given two alternatives:

\[ P_1 \sim \mathcal{N}(0, (1 - m^{-1})^2) \quad P_2 \sim \mathcal{N}(0, 1) \quad Q \sim \mathcal{N}(0, 1). \]

Clearly,

\[ \text{FID}(P_1, Q) = \frac{1}{m^2} > \text{FID}(P_2, Q) = 0 \]

Given $m$ samples from $P_1$ and $P_2$,

\[ \text{FID}(\hat{P}_1, Q) < \text{FID}(\hat{P}_2, Q). \]
Evaluation of GANs

The FID can give the wrong answer in practice.

Let $d = 2048$, and define

\[ P_1 = \text{relu}(\mathcal{N}(0, I_d)) \quad P_2 = \text{relu}(\mathcal{N}(1, .8\Sigma + .2I_d)) \quad Q = \text{relu}(\mathcal{N}(1, I_d)) \]

where $\Sigma = \frac{4}{d} CC^T$, with $C$ a $d \times d$ matrix with iid standard normal entries.

For a random draw of $C$:

\[ \text{FID}(P_1, Q) \approx 1123.0 > 1114.8 \approx \text{FID}(P_2, Q) \]

With $m = 50000$ samples,

\[ \text{FID}(\hat{P}_1, Q) \approx 1133.7 < 1136.2 \approx \text{FID}(\hat{P}_2, Q) \]

At $m = 100000$ samples, the ordering of the estimates is correct.

This behavior is similar for other random draws of $C$. 
Evaluation of GANs

The FID can give the wrong answer in practice.

Let \( d = 2048 \), and define

\[
\begin{align*}
P_1 &= \text{relu}(\mathcal{N}(0, I_d)) & P_2 &= \text{relu}(\mathcal{N}(1, .8\Sigma + .2I_d)) & Q &= \text{relu}(\mathcal{N}(1, I_d))
\end{align*}
\]

where \( \Sigma = \frac{4}{d} CC^T \), with \( C \) a \( d \times d \) matrix with iid standard normal entries.

For a random draw of \( C \):

\[
FID(P_1, Q) \approx 1123.0 > 1114.8 \approx FID(P_2, Q)
\]

With \( m = 50000 \) samples,

\[
FID(\hat{P}_1, Q) \approx 1133.7 < 1136.2 \approx FID(\hat{P}_2, Q)
\]

At \( m = 100000 \) samples, the ordering of the estimates is correct.

This behavior is similar for other random draws of \( C \).
Evaluation of GANs

The FID can give the wrong answer in practice.
Let $d = 2048$, and define

$$P_1 = \text{relu}(\mathcal{N}(0, I_d)) \quad P_2 = \text{relu}(\mathcal{N}(1, .8\Sigma+.2I_d)) \quad Q = \text{relu}(\mathcal{N}(1, I_d))$$

where $\Sigma = \frac{4}{d} CC^T$, with $C$ a $d \times d$ matrix with iid standard normal entries.

For a random draw of $C$:

$$FID(P_1, Q) \approx 1123.0 > 1114.8 \approx FID(P_2, Q)$$

With $m = 50000$ samples,

$$FID(\hat{P}_1, Q) \approx 1133.7 < 1136.2 \approx FID(\hat{P}_2, Q)$$

At $m = 100000$ samples, the ordering of the estimates is correct. This behavior is similar for other random draws of $C$. 

Evaluation of GANs

The FID can give the wrong answer in practice.

Let $d = 2048$, and define

\[
P_1 = \text{relu}(\mathcal{N}(0, I_d)) \quad P_2 = \text{relu}(\mathcal{N}(1, 0.8\Sigma + 0.2I_d)) \quad Q = \text{relu}(\mathcal{N}(1, I_d))
\]

where $\Sigma = \frac{4}{d} CC^T$, with $C$ a $d \times d$ matrix with iid standard normal entries.

For a random draw of $C$:

\[
FID(P_1, Q) \approx 1123.0 > 1114.8 \approx FID(P_2, Q)
\]

With $m = 50000$ samples,

\[
FID(\hat{P}_1, Q) \approx 1133.7 < 1136.2 \approx FID(\hat{P}_2, Q)
\]

At $m = 100000$ samples, the ordering of the estimates is correct.

This behavior is similar for other random draws of $C$. 


The kernel inception distance (KID)

The Kernel inception distance Binkowski, Sutherland, Arbel, G. [ICLR 2018]

Measures similarity of the samples’ representations in the inception architecture (pool3 layer)

MMD with kernel

\[ k(x, y) = \left( \frac{1}{d} x^\top y + 1 \right)^3. \]

- Checks match for feature means, variances, skewness
- Unbiased: eg CIFAR-10 train/test
The kernel inception distance (KID)

The Kernel inception distance Binkowski, Sutherland, Arbel, G. [ICLR 2018]

Measures similarity of the samples’ representations in the inception architecture (pool3 layer)

MMD with kernel

\[ k(x, y) = \left( \frac{1}{d} x^\top y + 1 \right)^3. \]

- Checks match for feature means, variances, skewness
- Unbiased: eg CIFAR-10 train/test

...“but isn’t KID is computationally costly?”
The kernel inception distance (KID)

The Kernel inception distance \(\text{Binkowski, Sutherland, Arbel, G.} \ [\text{ICLR 2018}]\)

Measures similarity of the samples’ representations in the inception architecture (pool3 layer)

\(\text{MMD with kernel}\)

\[
k(x, y) = \left(\frac{1}{d} x^\top y + 1\right)^3.
\]

- Checks match for feature means, variances, skewness
- Unbiased : eg CIFAR-10 train/test

...“but isn’t KID is computationally costly?”

“Block” KID implementation is cheaper than FID: see paper (or use Tensorflow implementation)!
The kernel inception distance (KID)

The Kernel Inception distance Binkowski, Sutherland, Arbel, G. [ICLR 2018]

Measures similarity of the samples’ representations in the inception architecture (pool3 layer)

**MMD with kernel**

\[ k(x, y) = \left( \frac{1}{d} x^\top y + 1 \right)^3. \]

- Checks match for feature means, variances, skewness
- **Unbiased**: eg CIFAR-10 train/test

Also used for automatic learning rate adjustment: if \( KID(\hat{P}_{t+1}, Q) \) not significantly better than \( KID(\hat{P}_t, Q) \) then reduce learning rate.

[Bounliphone et al. ICLR 2016]

Related: “An empirical study on evaluation metrics of generative adversarial networks”, Xu et al. [arxiv, June 2018]
Benchmarks for comparison (all from ICLR 2018)

SPECTRAL NORMALIZATION
FOR GENERATIVE ADVERSARIAL NETWORKS

Takeru Miyato¹, Toshiki Kataoka¹, Masanori Koyama², Yuichi Yoshida³
{miyato, kataoka}@preferred.jp
koyama.masanori1@gmail.com
yoshida@nii.ac.jp
¹Preferred Networks, Inc. ²Ritsumeikan University ³National Institute of Informatics

Sobolev GAN

Youssef Mroueh¹, Chun-Liang Li²*, Tom Sercu¹**, Anant Raj⁹* & Yu Cheng¹
† IBM Research AI
ο Carnegie Mellon University
◊ Max Planck Institute for Intelligent Systems
* denotes Equal Contribution
{mroueh, chengu}@us.ibm.com, chunliai@cs.cmu.edu, tom.sercu@ibm.com, anant.raj@tuebingen.mpg.de

DEMYSTIFYING MMD GANs

Mikołaj Bifikowski*
Department of Mathematics
Imperial College London
mikbinkowski@gmail.com

Demystifying MMD GANs

Dougal J. Sutherland; Michael Arbel & Arthur Gretton
Gatsby Computational Neuroscience Unit
University College London
dougal.sutherland, michael.n.arbel, arthur.gretton@gmail.com

Boundary-Seeking
Generative Adversarial Networks

R Devon Hjelm*
MILA, University of Montréal, IVADO
erroneus@gmail.com

Boundary-Seeking
Generative Adversarial Networks

Tong Che
MILA, University of Montréal
tong.che@umontreal.ca

Boundary-Seeking
Generative Adversarial Networks

Kyunghyun Cho
New York University,
CIFAR Azrieli Global Scholar
kyunghyun.cho@nyu.edu

Boundary-Seeking
Generative Adversarial Networks

Yoshua Bengio
MILA, University of Montréal, CIFAR, IVADO
yoshua.bengio@umontreal.ca
Results: what does MMD buy you?

- Critic features from DCGAN: an $f$-filter critic has $f$, $2f$, $4f$, and $8f$ convolutional filters in layers 1-4. LSUN $64 \times 64$.

MMD GAN samples, $f = 64$, $\text{FID}=32$, $\text{KID}=3$

WGAN samples, $f = 64$, $\text{FID}=41$, $\text{KID}=4$
Results: what does MMD buy you?

- **Critic features from DCGAN**: an $f$-filter critic has $f$, $2f$, $4f$ and $8f$ convolutional filters in layers 1-4. LSUN $64 \times 64$.

MMD GAN samples, $f = 16$, FID=86, KID=9

WGAN samples, $f = 16$, $f = 64$, FID=293, KID=37/71
The kernel inception distance (KID)

Faster training: performance scores vs generator iterations on MNIST
Results: celebrity faces $160 \times 160$

KID (FID)
scores:

- Sobolev GAN: 14 (20)
- SN-GAN: 18 (28)
- Old MMD GAN: 13 (21)
- SMMD GAN: 6 (12)

202,599 face images, resized and cropped to $160 \times 160$
Results: imagenet $64 \times 64$

KID (FID) scores:

- **BGAN:** 47 (44)
- **SN-GAN:** 44 (48)
- **SMMD GAN:** 35 (37)

ILSVRC2012 (ImageNet) dataset, 1,281,167 images, resized to $64 \times 64$. Around 20,000 classes.
Results: imagenet 64×64

KID (FID) scores:

- **BGAN:** 47 (44)
- **SN-GAN:** 44 (48)
- **SMMD GAN:** 35 (37)

ILSVRC2012 (ImageNet) dataset, 1 281 167 images, resized to 64 × 64. Around 20 000 classes.
Results: imagenet $64 \times 64$

**KID (FID) scores:**

- **BGAN:** 47 (44)
- **SN-GAN:** 44 (48)
- **SMMD GAN:** 35 (37)

ILSVRC2012 (ImageNet) dataset, 1 281 167 images, resized to $64 \times 64$. Around 20 000 classes.
Summary

- MMD critic gives state-of-the-art performance for GAN training (FID and KID)
  - use convolutional input features
  - train with new gradient regulariser
- Faster training, simpler critic network
- Reasons for good performance:
  - Unlike WGAN-GP, MMD loss still a valid critic when features not optimal
  - Kernel features do some of the “work”, so simpler $h_\psi$ features possible.
  - Better gradient/feature regulariser gives better critic


Code for new SMMD: https://github.com/MichaelArbel/Scaled-MMD-GAN
Testing against a probabilistic model
Statistical model criticism

\[ MMD(P, Q) = \|f^*\|^2 = \sup_{\|f\|_{\mathcal{F}} \leq 1} [E_Q f - E_P f] \]

\( f^*(x) \) is the witness function

Can we compute MMD with samples from \( Q \) and a model \( P \)?

**Problem:** usually can’t compute \( E_P f \) in closed form.
To get rid of $E_p f$ in

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

we define the **Stein operator**

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

Then

$$E_P T_P f = 0$$

subject to appropriate boundary conditions. (Oates, Girolami, Chopin, 2016)
Stein idea: proof

\[
E_p [ T_p f ] = \int \left[ \frac{1}{p(x)} \frac{d}{dx} (f(x)p(x)) \right] p(x) \, dx
\]

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

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

\[
= 0
\]
Stein idea: proof

\[
E_p [T_p f] = \int \left[ \frac{1}{p(x)} \frac{d}{dx} (f(x)p(x)) \right] p(x) dx
\]

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

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

\[= 0\]
Stein idea: proof

\[ E_p \left[ T_pf \right] = \int \left[ \frac{1}{p(x)} \frac{d}{dx} (f(x)p(x)) \right] p(x) \, dx \]
\[ \int \left[ \frac{d}{dx} (f(x)p(x)) \right] \, dx \]

= \left[ f(x)p(x) \right]_{-\infty}^{\infty}

= 0
Stein idea: proof

\[
E_p \left[ T_p f \right] = \int \left[ \frac{1}{p(x)} \frac{d}{dx} (f(x)p(x)) \right] p(x) \, dx \\
\int \left[ \frac{d}{dx} (f(x)p(x)) \right] \, dx \\
= [f(x)p(x)]_{-\infty}^{\infty} \\
= 0
\]
Stein idea: proof

\[ E_p [T_p f] = \int \left[ \frac{1}{p(x)} \frac{d}{dx} (f(x)p(x)) \right] p(x) \, dx \]
\[ = \int \left[ \frac{d}{dx} (f(x)p(x)) \right] \, dx \]
\[ = [f(x)p(x)]_{-\infty}^{\infty} \]
\[ = 0 \]
Kernel Stein Discrepancy

Stein operator

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

Kernel Stein Discrepancy (KSD)

\[ KSD(p, q, \mathcal{F}) = \sup \quad E_q T_p g - E_p T_p g \]

\[ \|g\|_{\mathcal{F}} \leq 1 \]
Kernel Stein Discrepancy

Stein operator

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

Kernel Stein Discrepancy (KSD)

\[ KSD(p, q, \mathcal{F}) = \sup_{\|g\|_{\mathcal{F}} \leq 1} E_q T_p g - E_p T_p g = \sup_{\|g\|_{\mathcal{F}} \leq 1} E_q T_p g \]
Kernel Stein Discrepancy

Stein operator

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

Kernel Stein Discrepancy (KSD)

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

![Graph showing p(x), q(x), and g*(x) with KSD values at different points.](image-url)
Kernel Stein Discrepancy

Stein operator

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

Kernel Stein Discrepancy (KSD)

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

![Graph showing distributions](image-url)
Kernel stein discrepancy

Closed-form expression for KSD: given $Z, Z' \sim q$, then

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

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

where

$$h_p(x, y) := \partial_x \log p(x) \partial_x \log p(y) k(x, y)$$

$$+ \partial_y \log p(y) \partial_x k(x, y)$$

$$+ \partial_x \log p(x) \partial_y k(x, y)$$

$$+ \partial_x \partial_y k(x, y)$$

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

Chicago crime data
Chicago crime data

Model is Gaussian mixture with two components.
Statistical model criticism

Chicago crime data

Model is Gaussian mixture with two components

Stein witness function
Statistical model criticism

Chicago crime data

Model is Gaussian mixture with ten components.
Chicago crime data

Model is Gaussian mixture with ten components

Stein witness function

Code: https://github.com/karlnapf/kernel_goodness_of_fit
Kernel stein discrepancy

Further applications:

- Evaluation of approximate MCMC methods.
  (Chwialkowski, Strathmann, G., ICML 2016; Gorham, Mackey, ICML 2017)

What kernel to use?

- The inverse multiquadric kernel,

\[ k(x, y) = \left( c + \|x - y\|_2^2 \right)^\beta \]

for \( \beta \in (-1, 0) \).
Testing statistical dependence
# Dependence testing

- **Given:** Samples from a distribution $P_{XY}$
- **Goal:** Are $X$ and $Y$ independent?

<table>
<thead>
<tr>
<th>X</th>
<th>Y</th>
</tr>
</thead>
<tbody>
<tr>
<td><img src="dog.jpg" alt="Dog Image" /></td>
<td>A large animal who slings slobber, exudes a distinctive houndy odor, and wants nothing more than to follow his nose.</td>
</tr>
<tr>
<td><img src="dog.jpg" alt="Dog Image" /></td>
<td>Their noses guide them through life, and they're never happier than when following an interesting scent.</td>
</tr>
<tr>
<td><img src="cat.jpg" alt="Cat Image" /></td>
<td>A responsive, interactive pet, one that will blow in your ear and follow you everywhere.</td>
</tr>
</tbody>
</table>

Text from dogtime.com and petfinder.com
MMD as a dependence measure?

Could we use MMD?

\[ MMD(P_{XY}, P_X P_Y, \mathcal{H}_\kappa) \]

- We don’t have samples from \( Q := P_X P_Y \), only pairs \( \{(x_i, y_i)\}_{i=1}^n \) i.i.d. \( P_{XY} \)
  - Solution: simulate \( Q \) with pairs \( (x_i, y_j) \) for \( j \neq i \).

- What kernel \( \kappa \) to use for the RKHS \( \mathcal{H}_\kappa \)?
MMD as a dependence measure?

Could we use MMD?

\[ MMD(P_{XY}, P_X P_Y, \mathcal{H}_\kappa) \]

- We don’t have samples from \( Q := P_X P_Y \), only pairs
  \( \{ (x_i, y_i) \}_{i=1}^n \overset{\text{i.i.d.}}{\sim} P_{XY} \)
  - Solution: simulate \( Q \) with pairs \((x_i, y_j)\) for \( j \neq i \).

- What kernel \( \kappa \) to use for the RKHS \( \mathcal{H}_\kappa \)?
MMD as a dependence measure?

Could we use MMD?

\[
MMD(P_{XY}, P_X P_Y, \mathcal{H}_\kappa)
\]

- We don’t have samples from \( Q := P_X P_Y \), only pairs 
  \[ \{(x_i, y_i)\}_{i=1}^n \overset{\text{i.i.d.}}{\sim} P_{XY} \]
  - Solution: simulate \( Q \) with pairs \((x_i, y_j)\) for \( j \neq i \).

- What kernel \( \kappa \) to use for the RKHS \( \mathcal{H}_\kappa \)?
MMD as a dependence measure

Kernel \( k \) on images with feature space \( \mathcal{F} \),

\[
k(x, y)
\]

Kernel \( l \) on captions with feature space \( \mathcal{G} \),

\[
l(A, B)
\]
MMD as a dependence measure

Kernel $k$ on images with feature space $\mathcal{F}$,

$$k(\text{dog}, \text{cat})$$

Kernel $l$ on captions with feature space $\mathcal{G}$,

$$l(\text{A large animal who slings slobber, ...}, \text{A responsive, interactive pet, ...})$$

Kernel $\kappa$ on image-text pairs: are images and captions similar?

$$\kappa(\text{dog}, \text{cat}) = k(\text{dog}, \text{cat}) \times l(\text{A large animal who slings slobber, ...}, \text{A responsive, interactive pet, ...})$$
MMD as a dependence measure

- **Given:** Samples from a distribution $P_{XY}$
- **Goal:** Are $X$ and $Y$ independent?

$$MMD^2(\hat{P}_{XY}, \hat{P}_X \hat{P}_Y, \mathcal{H}_\kappa) := \frac{1}{n^2} \text{trace}(K L)$$

($K$, $L$ column centered)
MMD as a dependence measure

- **Given:** Samples from a distribution $P_{XY}$
- **Goal:** Are $X$ and $Y$ independent?

$$MMD^2(\hat{P}_{XY}, \hat{P}_X \hat{P}_Y, \mathcal{H}_\kappa) := \frac{1}{n^2} \text{trace}(KL)$$

Text from dogtime.com and petfinder.com
MMD as a dependence measure

Two questions:

- Why the product kernel? Many ways to combine kernels - why not eg a sum?
- Is there a more interpretable way of defining this dependence measure?
Illustration: dependence ≠ correlation

- **Given:** Samples from a distribution $P_{XY}$
- **Goal:** Are $X$ and $Y$ dependent?

Correlation: 0.88
Illustration: dependence $\neq$ correlation

- **Given:** Samples from a distribution $P_{XY}$
- **Goal:** Are $X$ and $Y$ dependent?

![Scatter plot with correlation 0.07](image-url)
Illustration: dependence ≠ correlation

- **Given:** Samples from a distribution $P_{XY}$
- **Goal:** Are $X$ and $Y$ dependent?

![Correlation: 0.00](image)

The scatter plot shows a circular pattern, indicating no linear dependence, hence a correlation of 0.00.
Finding covariance with smooth transformations

Illustration: two variables with no correlation but strong dependence.
Finding covariance with smooth transformations

Illustration: two variables with no correlation but strong dependence.
Finding covariance with smooth transformations

Illustration: two variables with no correlation but strong dependence.
Define two spaces, one for each witness

**Function in \( \mathcal{F} \)**

\[
f(x) = \sum_{j=1}^{\infty} f_j \varphi_j(x)
\]

**Feature map**

\[
\varphi(x) = \begin{bmatrix}
\varphi_1(x) \\
\varphi_2(x) \\
\varphi_3(x) \\
\vdots
\end{bmatrix}
\]

**Kernel for RKHS \( \mathcal{F} \) on \( \mathcal{X} \):**

\[
k(x, x') = \langle \varphi(x), \varphi(x') \rangle_{\mathcal{F}}
\]

**Function in \( \mathcal{G} \)**

\[
g(y) = \sum_{j=1}^{\infty} g_j \phi_j(y)
\]

**Feature map**

\[
\phi(y) = \begin{bmatrix}
\phi_1(y) \\
\phi_2(y) \\
\phi_3(y) \\
\vdots
\end{bmatrix}
\]

**Kernel for RKHS \( \mathcal{G} \) on \( \mathcal{Y} \):**

\[
l(x, x') = \langle \phi(y), \phi(y') \rangle_{\mathcal{G}}
\]
The constrained covariance

The constrained covariance is

$$\text{COCO}(P_{XY}) = \sup f \| f \|_{\mathcal{F}} \leq 1 \quad \text{cov}[f(x)g(y)] \quad \| g \|_{\mathcal{G}} \leq 1$$
The constrained covariance

The constrained covariance is

\[
\text{COCO}(P_{XY}) = \sup_{\|f\|_F \leq 1, \|g\|_g \leq 1} \text{cov} \left[ \left( \sum_{j=1}^{\infty} f_j \varphi_j(x) \right) \left( \sum_{j=1}^{\infty} g_j \phi_j(y) \right) \right]
\]
The constrained covariance

The constrained covariance is

$$\text{COCO}(P_{XY}) = \sup_{\|f\|_\mathcal{F} \leq 1, \|g\|_\mathcal{G} \leq 1} E_{xy} \left[ \left( \sum_{j=1}^{\infty} f_j \varphi_j(x) \right) \left( \sum_{j=1}^{\infty} g_j \phi_j(y) \right) \right]$$

Fine print: feature mappings $\varphi(x)$ and $\phi(y)$ assumed to have zero mean.
The constrained covariance

The constrained covariance is

\[
\text{COCO}(P_{XY}) = \sup_{\|f\|_\mathcal{F} \leq 1} E_{xy} \left[ \left( \sum_{j=1}^{\infty} f_j \varphi_j(x) \right) \left( \sum_{j=1}^{\infty} g_j \phi_j(y) \right) \right]
\]

\[
\text{COCO}(P_{XY}) = \sup_{\|f\|_\mathcal{F} \leq 1, \|g\|_{\mathcal{G}} \leq 1} E_{xy} \left[ \left( \sum_{j=1}^{\infty} f_j \varphi_j(x) \right) \left( \sum_{j=1}^{\infty} g_j \phi_j(y) \right) \right]
\]

Fine print: feature mappings \( \varphi(x) \) and \( \phi(y) \) assumed to have zero mean.

Rewriting:

\[
E_{xy}[f(x)g(y)]
= \begin{bmatrix} f_1 \\ f_2 \\ \vdots \end{bmatrix}^\top \text{E}_{xy} \left( \begin{bmatrix} \varphi_1(x) \\ \varphi_2(x) \\ \vdots \end{bmatrix} \begin{bmatrix} \phi_1(y) & \phi_2(y) & \cdots \end{bmatrix} \right) \begin{bmatrix} g_1 \\ g_2 \\ \vdots \end{bmatrix}
\]

\[
E_{xy}[f(x)g(y)] = \begin{bmatrix} f_1 \\ f_2 \\ \vdots \end{bmatrix}^\top \text{E}_{xy} \left( \begin{bmatrix} \varphi_1(x) \\ \varphi_2(x) \\ \vdots \end{bmatrix} \begin{bmatrix} \phi_1(y) & \phi_2(y) & \cdots \end{bmatrix} \right) \begin{bmatrix} g_1 \\ g_2 \\ \vdots \end{bmatrix}
\]

\[
\text{C}_{\varphi(x)\phi(y)}
\]
The constrained covariance

The constrained covariance is

$$\text{COCO}(P_{XY}) = \sup \left[ \frac{1}{E_{xy}}\left( \sum_{j=1}^{\infty} f_j \varphi_j(x) \right) \left( \sum_{j=1}^{\infty} g_j \phi_j(y) \right) \right]$$

subject to

$$\|f\|_\mathcal{F} \leq 1$$

$$\|g\|_g \leq 1$$

**Fine print:** feature mappings $\varphi(x)$ and $\phi(y)$ assumed to have zero mean.

**Rewriting:**

$$E_{xy}[f(x)g(y)]$$

$$= \left[ \begin{array}{c} f_1 \\ f_2 \\ \vdots \end{array} \right]^\top \text{E}_{xy} \left( \begin{array}{c} \varphi_1(x) \\ \varphi_2(x) \\ \vdots \end{array} \right) \left[ \begin{array}{c} \phi_1(y) \\ \phi_2(y) \\ \vdots \end{array} \right]$$

$$C_{\varphi(x)\phi(y)}$$

**COCO:** max singular value of feature covariance $C_{\varphi(x)\phi(y)}$
Computing COCO in practice

Given sample \( \{(x_i, y_i)\}_{i=1}^n \sim P_{XY} \), what is empirical \( \widehat{COCO} \)?
Computing COCO in practice

Given sample \( \{(x_i, y_i)\}_{i=1}^{n} \overset{\text{i.i.d.}}{\sim} P_{XY} \), what is empirical \( \widehat{\text{COCO}} \)?

\( \widehat{\text{COCO}} \) is largest eigenvalue \( \gamma_{\max} \) of

\[
\begin{bmatrix}
0 & \frac{1}{n} KL \\
\frac{1}{n} LK & 0
\end{bmatrix}
\begin{bmatrix}
\alpha \\
\beta
\end{bmatrix}
= \gamma
\begin{bmatrix}
K & 0 \\
0 & L
\end{bmatrix}
\begin{bmatrix}
\alpha \\
\beta
\end{bmatrix}.
\]

\( K_{ij} = k(x_i, x_j) \) and \( L_{ij} = l(y_i, y_j) \).

Fine print: kernels are computed with empirically centered features \( \varphi(x) - \frac{1}{n} \sum_{i=1}^{n} \varphi(x_i) \) and \( \phi(y) - \frac{1}{n} \sum_{i=1}^{n} \phi(y_i) \).

G., Smola., Bousquet, Herbrich, Belitski, Augath, Murayama, Pauls, Schoelkopf, and Logothetis, AISTATS’05
Computing COCO in practice

Given sample \( \{(x_i, y_i)\}_{i=1}^n \overset{\text{i.i.d.}}{\sim} P_{XY} \), what is empirical \( \widehat{\text{COCO}} \)?

\( \text{COCO} \) is largest eigenvalue \( \gamma_{\text{max}} \) of

\[
\begin{bmatrix}
0 & \frac{1}{n} KL \\
\frac{1}{n} LK & 0
\end{bmatrix}
\begin{bmatrix}
\alpha \\
\beta
\end{bmatrix}
= \gamma
\begin{bmatrix}
K & 0 \\
0 & L
\end{bmatrix}
\begin{bmatrix}
\alpha \\
\beta
\end{bmatrix}.
\]

\( K_{ij} = k(x_i, x_j) \) and \( L_{ij} = l(y_i, y_j) \).

Witness functions (singular vectors):

\[
f(x) \propto \sum_{i=1}^n \alpha_i k(x_i, x) \quad g(y) \propto \sum_{i=1}^n \beta_i l(y_i, y)
\]

Fine print: kernels are computed with empirically centered features \( \varphi(x) = \frac{1}{n} \sum_{i=1}^n \varphi(x_i) \) and \( \phi(y) = \frac{1}{n} \sum_{i=1}^n \phi(y_i) \).

G., Smola., Bousquet, Herbrich, Belitski, Augath, Murayama, Pauls, Schoelkopf, and Logothetis, AISTATS’05
Empirical COCO: proof (1)

The Lagrangian is

\[ \mathcal{L}(f, g, \lambda, \gamma) = \frac{1}{n} \sum_{i=1}^{n} [f(x_i)g(y_i)] - \frac{\lambda}{2} \left( \|f\|_F^2 - 1 \right) - \frac{\gamma}{2} \left( \|g\|_G^2 - 1 \right). \]

\text{covariance} \quad \text{smoothness constraints}

Fine print: \( f(x_i)g(y_i) \) centered to have zero empirical mean.
Empirical COCO: proof (1)

The Lagrangian is

\[
\mathcal{L}(f, g, \lambda, \gamma) = \frac{1}{n} \sum_{i=1}^{n} [f(x_i)g(y_i)] - \frac{\lambda}{2} \left( \|f\|_F^2 - 1 \right) - \frac{\gamma}{2} \left( \|g\|_G^2 - 1 \right).
\]

\(\underbrace{\text{covariance}}_{\text{smoothness constraints}}\)

Fine print: \(f(x_i)g(y_i)\) centered to have zero empirical mean.

Assume (cf representer theorem):

\[
f = \sum_{i=1}^{n} \alpha_i \varphi(x_i) \quad g = \sum_{i=1}^{n} \beta_i \psi(y_i)
\]

for centered \(\varphi(x_i), \psi(y_i)\).
Empirical COCO: proof (1)

The Lagrangian is

\[ \mathcal{L}(f, g, \lambda, \gamma) = \frac{1}{n} \sum_{i=1}^{n} [f(x_i)g(y_i)] - \frac{\lambda}{2} \left( \|f\|_{\mathcal{F}}^2 - 1 \right) - \frac{\gamma}{2} \left( \|g\|_{\mathcal{G}}^2 - 1 \right). \]

Covariance \hspace{1cm} Smoothness constraints

Fine print: \( f(x_i)g(y_i) \) centered to have zero empirical mean.

Assume (cf representer theorem):

\[ f = \sum_{i=1}^{n} \alpha_i \varphi(x_i) \quad g = \sum_{i=1}^{n} \beta_i \psi(y_i) \]

for centered \( \varphi(x_i), \psi(y_i) \).

First step is smoothness constraint:

\[ \|f\|_{\mathcal{F}}^2 - 1 = \langle f, f \rangle_{\mathcal{F}} - 1 \]
The Lagrangian is

\[ \mathcal{L}(f, g, \lambda, \gamma) = \frac{1}{n} \sum_{i=1}^{n} [f(x_i)g(y_i)] - \frac{\lambda}{2} \left( \|f\|_{\mathcal{F}}^2 - 1 \right) - \frac{\gamma}{2} \left( \|g\|_{\mathcal{G}}^2 - 1 \right). \]

**Fine print:** \(f(x_i)g(y_i)\) centered to have zero empirical mean.

**Assume** (cf representer theorem):

\[ f = \sum_{i=1}^{n} \alpha_i \varphi(x_i) \quad g = \sum_{i=1}^{n} \beta_i \psi(y_i) \]

for centered \(\varphi(x_i), \phi(y_i)\).

**First step is smoothness constraint:**

\[ \|f\|_{\mathcal{F}}^2 - 1 = \langle f, f \rangle_{\mathcal{F}} - 1 \]

\[ = \left\langle \sum_{i=1}^{n} \alpha_i \varphi(x_i), \sum_{i=1}^{n} \alpha_i \varphi(x_i) \right\rangle_{\mathcal{F}} - 1 \]
Empirical COCO: proof (1)

The Lagrangian is

\[
\mathcal{L}(f, g, \lambda, \gamma) = \frac{1}{n} \sum_{i=1}^{n} [f(x_i)g(y_i)] - \frac{\lambda}{2} \left( \|f\|^2_{\mathcal{F}} - 1 \right) - \frac{\gamma}{2} \left( \|g\|^2_{\mathcal{G}} - 1 \right).
\]

\[
\text{covariance} \quad \text{smoothness constraints}
\]

Fine print: \( f(x_i)g(y_i) \) centered to have zero empirical mean.

Assume (cf representer theorem):

\[
f = \sum_{i=1}^{n} \alpha_i \varphi(x_i) \quad g = \sum_{i=1}^{n} \beta_i \psi(y_i)
\]

for centered \( \varphi(x_i), \phi(y_i) \).

First step is smoothness constraint:

\[
\|f\|^2_{\mathcal{F}} - 1 = \langle f, f \rangle_{\mathcal{F}} - 1
\]

\[
= \left\langle \sum_{i=1}^{n} \alpha_i \varphi(x_i), \sum_{i=1}^{n} \alpha_i \varphi(x_i) \right\rangle_{\mathcal{F}} - 1
\]

\[
= \alpha^\top K \alpha - 1
\]
Second step is covariance:

\[
\frac{1}{n} \sum_{i=1}^{n} [f(x_i) g(y_i)] = \frac{1}{n} \sum_{i=1}^{n} \langle f, \varphi(x_i) \rangle_{\mathcal{F}} \langle g, \varphi(y_i) \rangle_{\mathcal{G}} \\
= \frac{1}{n} \sum_{i=1}^{n} \left\langle \sum_{\ell=1}^{n} \alpha_{\ell} \varphi(x_{\ell}), \varphi(x_i) \right\rangle_{\mathcal{F}} \langle g, \varphi(y_i) \rangle_{\mathcal{G}} \\
= \frac{1}{n} \alpha^T KL \beta
\]
Second step is covariance:

\[
\frac{1}{n} \sum_{i=1}^{n} [f(x_i)g(y_i)] = \frac{1}{n} \sum_{i=1}^{n} \langle f, \varphi(x_i) \rangle_{\mathcal{F}} \langle g, \varphi(y_i) \rangle_{\mathcal{G}}
\]

\[
= \frac{1}{n} \sum_{i=1}^{n} \left\langle \sum_{\ell=1}^{n} \alpha_{\ell} \varphi(x_{\ell}), \varphi(x_i) \right\rangle_{\mathcal{F}} \langle g, \varphi(y_i) \rangle_{\mathcal{G}}
\]

\[
= \frac{1}{n} \alpha^\top KL \beta
\]
Proof sketch (2)

Second step is covariance:

\[
\frac{1}{n} \sum_{i=1}^{n} [f(x_i)g(y_i)] = \frac{1}{n} \sum_{i=1}^{n} \langle f, \phi(x_i) \rangle_F \langle g, \phi(y_i) \rangle_G \\
= \frac{1}{n} \sum_{i=1}^{n} \left\langle \sum_{\ell=1}^{n} \alpha_\ell \phi(x_\ell), \phi(x_i) \right\rangle_F \langle g, \phi(y_i) \rangle_G \\
= \frac{1}{n} \alpha^\top KL \beta
\]

where \( K_{ij} = k(x_i, x_j) = \langle \phi(x_i), \phi(x_j) \rangle_F \) \quad \( L_{ij} = l(y_i, y_j) \).
Proof sketch (2)

Second step is covariance:

\[
\frac{1}{n} \sum_{i=1}^{n} [f(x_i)g(y_i)] = \frac{1}{n} \sum_{i=1}^{n} \langle f, \varphi(x_i) \rangle_{\mathcal{F}} \langle g, \varphi(y_i) \rangle_{\mathcal{G}}
\]

\[
= \frac{1}{n} \sum_{i=1}^{n} \left< \sum_{\ell=1}^{n} \alpha_{\ell} \varphi(x_{\ell}), \varphi(x_i) \right> \mathcal{F} \langle g, \varphi(y_i) \rangle_{\mathcal{G}}
\]

\[
= \frac{1}{n} \alpha^\top K L \beta
\]

where \( K_{ij} = k(x_i, x_j) = \langle \varphi(x_i), \varphi(x_j) \rangle_{\mathcal{F}} \quad L_{ij} = l(y_i, y_j). \)

The **Lagranian** is now:

\[
\mathcal{L}(f, g, \lambda, \gamma) = \frac{1}{n} \alpha^\top K L \beta - \frac{\lambda}{2} \left( \alpha^\top K \alpha - 1 \right) - \frac{\gamma}{2} \left( \beta^\top L \beta - 1 \right)
\]
What is a large dependence with COCO?

Density takes the form:

\[ P_{XY} \propto 1 + \sin(\omega x) \sin(\omega y) \]

Which of these is the more “dependent”?
Finding covariance with smooth transformations

Case of $\omega = 1$:
Finding covariance with smooth transformations

Case of $\omega = 2$:

- Correlation: 0.02
- COCO: 0.07

- Correlation: 0.54
- COCO: 0.07
Finding covariance with smooth transformations

Case of $\omega = 3$:

- Correlation: 0.03
- COCO: 0.04
Finding covariance with smooth transformations

Case of $\omega = 4$:

- Correlation: 0.05

- Correlation: 0.25  COCO: 0.02
Finding covariance with smooth transformations

Case of $\omega = ??$:

[Graphs showing correlation and COCO values]
Finding covariance with smooth transformations

Case of $\omega = 0$: uniform noise! (shows bias)
Dependence largest when at “low” frequencies

- As dependence is encoded at higher frequencies, the smooth mappings \( f, g \) achieve lower linear dependence.
- Even for independent variables, COCO will not be zero at finite sample sizes, since some mild linear dependence will be found by \( f, g \) (bias).
- This bias will decrease with increasing sample size.
Can we do better than COCO?

A second example with zero correlation.

**First** singular value of feature covariance $C_{\varphi(x)\varphi(y)}$:
Can we do better than COCO?

A second example with zero correlation.

Second singular value of feature covariance $C_{\varphi(x)\phi(y)}$:
Can we do better than COCO?

A second example with zero correlation.

Second singular value of feature covariance \( C_{\varphi(x)\varphi(y)} \):

\[
\begin{array}{c|c|c}
\hline
\text{Correlation: 0.00} & \text{Correlation: 0.37} & \text{COCO}_2: 0.06 \\
\hline
\end{array}
\]
The Hilbert-Schmidt Independence Criterion

Writing the $i$th singular value of the feature covariance $C_{\varphi(x)\varphi(y)}$ as

$$\gamma_i := COCO_i(P_{XY}; \mathcal{F}, \mathcal{G}),$$

define Hilbert-Schmidt Independence Criterion (HSIC)

$$HSIC^2(P_{XY}; \mathcal{F}, \mathcal{G}) = \sum_{i=1}^{\infty} \gamma_i^2.$$
The Hilbert-Schmidt Independence Criterion

Writing the $i$th singular value of the feature covariance $C_{\varphi(x)\phi(y)}$ as

$$\gamma_i := COCO_i(P_{XY}; \mathcal{F}, \mathcal{G}),$$

define **Hilbert-Schmidt Independence Criterion (HSIC)**

$$HSIC^2(P_{XY}; \mathcal{F}, \mathcal{G}) = \sum_{i=1}^{\infty} \gamma_i^2.$$ 

G., Bousquet , Smola., and Schoelkopf, ALT05; G.,, Fukumizu, Teo., Song., Schoelkopf., and Smola, NIPS 2007,.

**HSIC is MMD with product kernel!**

$$HSIC^2(P_{XY}; \mathcal{F}, \mathcal{G}) = MMD^2(P_{XY}, P_X P_Y; \mathcal{H}_\kappa)$$

where $\kappa((x, y), (x', y')) = k(x, x')l(y, y').$
Asymptotics of HSIC under independence

- Given sample \( \{(x_i, y_i)\}_{i=1}^{n} \overset{\text{i.i.d.}}{\sim} P_{XY} \), what is empirical \( \widehat{\text{HSIC}} \)?

- **Empirical HSIC** (biased)

  \[
  \widehat{\text{HSIC}} = \frac{1}{n^2} \text{trace}(KL)
  \]

  \( K_{ij} = k(x_i, x_j) \) and \( L_{ij} = l(y_i y_j) \)  

  (\( K \) and \( L \) computed with empirically centered features)

- **Statistical testing:** given \( P_{XY} = P_X P_Y \), what is the threshold \( c_\alpha \) such that \( P(\widehat{\text{HSIC}} > c_\alpha) < \alpha \) for small \( \alpha \)?

- **Asymptotics of \( \widehat{\text{HSIC}} \) when \( P_{XY} = P_X P_Y \):**

  \[
  n \widehat{\text{HSIC}} \xrightarrow{D} \sum_{l=1}^{\infty} \lambda_l z_l^2, \quad z_l \sim \mathcal{N}(0, 1)\text{i.i.d.}
  \]

  where \( \lambda_l \psi_l(z_j) = \int h_{ijqr} \psi_l(z_i) dF_{i,q,r} \),  

  \( h_{ijqr} = \frac{1}{4!} \sum_{(i,j,q,r)} k_{tu} l_{tu} + k_{tu} l_{vw} - 2k_{tu} l_{tv} \)
Asymptotics of HSIC under independence

- Given sample \( \{(x_i, y_i)\}_{i=1}^{n} \overset{\text{i.i.d.}}{\sim} P_{XY} \), what is empirical \( \hat{\text{HSIC}} \)?
- **Empirical HSIC** (biased)

\[
\hat{\text{HSIC}} = \frac{1}{n^2} \text{trace}(KL)
\]

\( K_{ij} = k(x_i, x_j) \) and \( L_{ij} = l(y_i y_j) \) (\( K \) and \( L \) computed with empirically centered features)

- **Statistical testing:** given \( P_{XY} = P_X P_Y \), what is the threshold \( c_\alpha \) such that \( P(\hat{\text{HSIC}} > c_\alpha) < \alpha \) for small \( \alpha \)?
- **Asymptotics of \( \hat{\text{HSIC}} \) when \( P_{XY} = P_X P_Y \):**

\[
n \hat{\text{HSIC}} \xrightarrow{D} \sum_{l=1}^{\infty} \lambda_l z_l^2, \quad z_l \sim \mathcal{N}(0, 1) \text{i.i.d.}
\]

where \( \lambda_l \psi_l(z_j) = \int h_{ijqr} \psi_l(z_i) dF_{i,q,r}, \quad h_{ijqr} = \frac{1}{4!} \sum_{(i,j,q,r)}^{(i,j,q,r)} k_{tu} l_{tu} + k_{tu} l_{vw} - 2 k_{tu} l_{tv} \)
Asymptotics of HSIC under independence

- Given sample \( \{(x_i, y_i)\}_{i=1}^n \) i.i.d. \( P_{XY} \), what is empirical \( \widehat{HSIC} \)?
- Empirical HSIC (biased)

\[
\widehat{HSIC} = \frac{1}{n^2} \text{trace}(KL)
\]

\( K_{ij} = k(x_i, x_j) \) and \( L_{ij} = l(y_i y_j) \)

- Statistical testing: given \( P_{XY} = P_X P_Y \), what is the threshold \( c_\alpha \) such that \( \text{P}(\widehat{HSIC} > c_\alpha) < \alpha \) for small \( \alpha \)?

- Asymptotics of \( \widehat{HSIC} \) when \( P_{XY} = P_X P_Y \):

\[
n \widehat{HSIC} \xrightarrow{D} \sum_{l=1}^{\infty} \lambda_l z^2_l, \quad z_l \sim \mathcal{N}(0, 1) \text{i.i.d.}
\]

where \( \lambda_l \psi_l(z_j) = \int h_{ijqr} \psi_l(z_i) dF_{i,q,r}, \quad h_{ijqr} = \frac{1}{4!} \sum_{(i,j,q,r)} k_{tu} l_{tu} + k_{tu} l_{vw} - 2k_{tu} l_{tu}
\]
Asymptotics of HSIC under independence

- Given sample \( \{ (x_i, y_i) \}_{i=1}^n \overset{\text{i.i.d.}}{\sim} P_{XY} \), what is empirical \( \hat{\text{HSIC}} \)?

- **Empirical HSIC** (biased)

  \[
  \hat{\text{HSIC}} = \frac{1}{n^2} \text{trace}(KL)
  \]

  \( K_{ij} = k(x_i, x_j) \) and \( L_{ij} = l(y_i y_j) \) \((K \text{ and } L \text{ computed with empirically centered features})\)

- **Statistical testing**: given \( P_{XY} = P_X P_Y \), what is the threshold \( c_\alpha \) such that \( P(\hat{\text{HSIC}} > c_\alpha) < \alpha \) for small \( \alpha \)?

- **Asymptotics of \( \hat{\text{HSIC}} \)** when \( P_{XY} = P_X P_Y \):

  \[
  n \hat{\text{HSIC}} \xrightarrow{D} \sum_{l=1}^\infty \lambda_l z_l^2, \quad z_l \sim N(0, 1) \text{i.i.d.}
  \]

  where \( \lambda_l \psi_l(z_j) = \int h_{ijqr} \psi_l(z_i) dF_{i,q,r} \),

  \[
  h_{ijqr} = \frac{1}{4!} \sum_{(i,j,q,r)} k_{tu} l_{tu} + k_{tu} l_{vw} - 2k_{tu} l_{tv}
  \]
A statistical test

Given \( P_{XY} = P_X P_Y \), what is the threshold \( c_\alpha \) such that \( P(\text{HSIC} > c_\alpha) < \alpha \) for small \( \alpha \) (prob. of false positive)?

Original time series:

\[ X_1 \ X_2 \ X_3 \ X_4 \ X_5 \ X_6 \ X_7 \ X_8 \ X_9 \ X_{10} \]
\[ Y_1 \ Y_2 \ Y_3 \ Y_4 \ Y_5 \ Y_6 \ Y_7 \ Y_8 \ Y_9 \ Y_{10} \]

Permutation:

\[ X_1 \ X_2 \ X_3 \ X_4 \ X_5 \ X_6 \ X_7 \ X_8 \ X_9 \ X_{10} \]
\[ Y_7 \ Y_3 \ Y_9 \ Y_2 \ Y_4 \ Y_8 \ Y_5 \ Y_1 \ Y_6 \ Y_{10} \]

Null distribution via permutation
- Compute HSIC for \( \{x_i, y_{\pi(i)}\}_{i=1}^{n} \) for random permutation \( \pi \) of indices \( \{1, \ldots, n\} \). This gives HSIC for independent variables.
- Repeat for many different permutations, get empirical CDF
- Threshold \( c_\alpha \) is \( 1 - \alpha \) quantile of empirical CDF
A statistical test

- Given $P_{XY} = P_X P_Y$, what is the threshold $c_\alpha$ such that $P(\widehat{\text{HSIC}} > c_\alpha) < \alpha$ for small $\alpha$ (prob. of false positive)?

- Original time series:

  \[
  X_1 \; X_2 \; X_3 \; X_4 \; X_5 \; X_6 \; X_7 \; X_8 \; X_9 \; X_{10}
  \]
  \[
  Y_1 \; Y_2 \; Y_3 \; Y_4 \; Y_5 \; Y_6 \; Y_7 \; Y_8 \; Y_9 \; Y_{10}
  \]

- Permutation:

  \[
  X_1 \; X_2 \; X_3 \; X_4 \; X_5 \; X_6 \; X_7 \; X_8 \; X_9 \; X_{10}
  \]
  \[
  Y_7 \; Y_3 \; Y_9 \; Y_2 \; Y_4 \; Y_8 \; Y_5 \; Y_1 \; Y_6 \; Y_{10}
  \]

- Null distribution via permutation
  - Compute HSIC for $\{x_i, y_{\pi(i)}\}_{i=1}^n$ for random permutation $\pi$ of indices $\{1, \ldots, n\}$. This gives HSIC for independent variables.
  - Repeat for many different permutations, get empirical CDF
  - Threshold $c_\alpha$ is $1 - \alpha$ quantile of empirical CDF
A statistical test

Given $P_{XY} = P_X P_Y$, what is the threshold $c_\alpha$ such that $P(\text{HSIC} > c_\alpha) < \alpha$ for small $\alpha$ (prob. of false positive)?

Original time series:

\[
\begin{align*}
X_1 & \quad X_2 & \quad X_3 & \quad X_4 & \quad X_5 & \quad X_6 & \quad X_7 & \quad X_8 & \quad X_9 & \quad X_{10} \\
Y_1 & \quad Y_2 & \quad Y_3 & \quad Y_4 & \quad Y_5 & \quad Y_6 & \quad Y_7 & \quad Y_8 & \quad Y_9 & \quad Y_{10}
\end{align*}
\]

Permutation:

\[
\begin{align*}
X_1 & \quad X_2 & \quad X_3 & \quad X_4 & \quad X_5 & \quad X_6 & \quad X_7 & \quad X_8 & \quad X_9 & \quad X_{10} \\
Y_7 & \quad Y_3 & \quad Y_9 & \quad Y_2 & \quad Y_4 & \quad Y_8 & \quad Y_5 & \quad Y_1 & \quad Y_6 & \quad Y_{10}
\end{align*}
\]

Null distribution via permutation

- Compute HSIC for $\{x_i, y_{\pi(i)}\}_{i=1}^{n}$ for random permutation $\pi$ of indices $\{1, \ldots, n\}$. This gives HSIC for independent variables.
- Repeat for many different permutations, get empirical CDF
- Threshold $c_\alpha$ is $1 - \alpha$ quantile of empirical CDF
# Application: dependence detection across languages

## Testing task: detect dependence between English and French text

<table>
<thead>
<tr>
<th>X</th>
<th>Y</th>
</tr>
</thead>
<tbody>
<tr>
<td>Honourable senators, I have a question for the Leader of the Government in the Senate</td>
<td>Honorables sénateurs, ma question s’adresse au leader du gouvernement au Sénat</td>
</tr>
<tr>
<td>No doubt there is great pressure on provincial and municipal governments</td>
<td>Les ordres de gouvernements provinciaux et municipaux subissent de fortes pressions</td>
</tr>
<tr>
<td>In fact, we have increased federal investments for early childhood development.</td>
<td>Au contraire, nous avons augmenté le financement fédéral pour le développement des jeunes</td>
</tr>
</tbody>
</table>

Text from the aligned hansards of the 36th parliament of Canada, [https://www.isi.edu/natural-language/download/hansard/](https://www.isi.edu/natural-language/download/hansard/)
Application: dependence detection across languages

Testing task: detect dependence between English and French text

$k$-spectrum kernel, $k = 10$, sample size $n = 10$

$$\text{HSIC} = \frac{1}{n^2} \text{trace}(KL)$$

(K and L column centered)
Application: Dependence detection across languages

Results (for $\alpha = 0.05$)

- k-spectrum kernel: average Type II error 0
- Bag of words kernel: average Type II error 0.18

Settings: Five line extracts, averaged over 300 repetitions, for "Agriculture" transcripts. Similar results for Fisheries and Immigration transcripts.
Testing higher order interactions
Detecting higher order interaction

How to detect V-structures with pairwise weak individual dependence?
Detecting higher order interaction

How to detect V-structures with pairwise weak individual dependence?
Detecting higher order interaction

How to detect V-structures with pairwise weak individual dependence?

\[ X \perp Y, Y \perp Z, X \perp Z \]

- \( X, Y \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0, 1) \)
- \( Z \mid X, Y \sim \text{sign}(XY)\exp\left(\frac{1}{\sqrt{2}}\right) \)

**Fine print:** Faithfulness violated here!
V-structure discovery

Assume $X \perp\!\!\!\!\!\!\perp Y$ has been established.

V-structure can then be detected by:

- **Consistent CI test:** $H_0 : X \perp\!\!\!\!\!\!\perp Y \mid Z$ [Fukumizu et al. 2008, Zhang et al. 2011]

- **Factorisation test:** $H_0 : (X, Y) \perp\!\!\!\!\!\!\perp Z \lor (X, Z) \perp\!\!\!\!\!\!\perp Y \lor (Y, Z) \perp\!\!\!\!\!\!\perp X$
  (multiple standard two-variable tests)

How well do these work?
Detecting higher order interaction

Generalise earlier example to $p$ dimensions

$X \perp Y, Y \perp Z, X \perp Z$

- $X_1$ vs $Y_1$
- $Y_1$ vs $Z_1$
- $X_1$ vs $Z_1$
- $X_1 \times Y_1$ vs $Z_1$

$X, Y \overset{i.i.d.}{\sim} \mathcal{N}(0, 1)$

$Z | X, Y \sim \text{sign}(XY) \exp\left(\frac{1}{\sqrt{2}}\right)$

$X_2:p, Y_2:p, Z_2:p \overset{i.i.d.}{\sim} \mathcal{N}(0, I_{p-1})$

Fine print: Faithfulness violated here!
V-structure discovery

CI test for $X \perp Y \mid Z$ from Zhang et al. (2011), and a factorisation test, $n = 500$
Lancaster interaction measure

Lancaster interaction measure of \((X_1, \ldots, X_D) \sim P\) is a signed measure \(\Delta P\) that vanishes whenever \(P\) can be factorised non-trivially.

\[
D = 2: \quad \Delta_L P = P_{XY} - P_X P_Y
\]
Lancaster interaction measure

Lancaster interaction measure of \((X_1, \ldots, X_D) \sim P\) is a signed measure \(\Delta P\) that vanishes whenever \(P\) can be factorised non-trivially.

\[
D = 2 : \quad \Delta_L P = P_{XY} - P_X P_Y \\
D = 3 : \quad \Delta_L P = P_{XYZ} - P_X P_{YZ} - P_Y P_{XZ} - P_Z P_{XY} + 2P_X P_Y P_Z
\]
Lancaster interaction measure

Lancaster interaction measure of \((X_1, \ldots, X_D) \sim P\) is a signed measure \(\Delta P\) that vanishes whenever \(P\) can be factorised non-trivially.

\[D = 2: \quad \Delta_L P = P_{XY} - P_X P_Y\]
\[D = 3: \quad \Delta_L P = P_{XYZ} - P_X P_{YZ} - P_Y P_{XZ} - P_Z P_{XY} + 2P_X P_Y P_Z\]

\[\Delta_L P = P_{XYZ}\]

\(-P_X P_{YZ}\)  \(-P_Y P_{XZ}\)  \(-P_Z P_{XY}\)  \(+2P_X P_Y P_Z\)

\(X\)  \(Y\)  \(Z\)  \(X\)  \(Y\)  \(Z\)  \(X\)  \(Y\)  \(Z\)
Lancaster interaction measure

Lancaster interaction measure of \((X_1, \ldots, X_D) \sim P\) is a signed measure \(\Delta P\) that vanishes whenever \(P\) can be factorised non-trivially.

\[\begin{align*}
D = 2 : & \quad \Delta_L P = P_{XY} - P_X P_Y \\
D = 3 : & \quad \Delta_L P = P_{XYZ} - P_X P_{YZ} - P_Y P_{XZ} - P_Z P_{XY} + 2P_X P_Y P_Z
\end{align*}\]

Case of \(P_X \indep P_{YZ}\)
Lancaster interaction measure

Lancaster interaction measure of \((X_1, \ldots, X_D) \sim P\) is a signed measure \(\Delta P\) that vanishes whenever \(P\) can be factorised non-trivially.

\[
\begin{align*}
D = 2 : & \quad \Delta_L P = P_{XY} - P_X P_Y \\
D = 3 : & \quad \Delta_L P = P_{XYZ} - P_X P_{YZ} - P_Y P_{XZ} - P_Z P_{XY} + 2P_X P_Y P_Z
\end{align*}
\]

\[(X, Y) \perp Z \lor (X, Z) \perp Y \lor (Y, Z) \perp X \implies \Delta_L P = 0.\]

...so what might be missed?
Lancaster interaction measure

Lancaster interaction measure of \((X_1, \ldots, X_D) \sim P\) is a signed measure \(\Delta P\) that vanishes whenever \(P\) can be factorised non-trivially.

\[
D = 2 : \quad \Delta_L P = P_{XY} - P_X P_Y \\
D = 3 : \quad \Delta_L P = P_{XYZ} - P_X P_{YZ} - P_Y P_{XZ} - P_Z P_{XY} + 2P_X P_Y P_Z
\]

\(\Delta_L P = 0 \not\Rightarrow (X, Y) \perp \!
\!
\!\perp Z \vee (X, Z) \perp \!
\!
\!\perp Y \vee (Y, Z) \perp \!
\!
\!\perp X\)

Example:

<table>
<thead>
<tr>
<th>(P(0,0,0))</th>
<th>(P(0,0,1))</th>
<th>(P(1,0,0))</th>
<th>(P(1,0,1))</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.2</td>
<td>0.1</td>
<td>0.1</td>
<td>0.1</td>
</tr>
<tr>
<td>(P(0,1,0))</td>
<td>(P(0,1,1))</td>
<td>(P(1,1,0))</td>
<td>(P(1,1,1))</td>
</tr>
<tr>
<td>0.1</td>
<td>0.1</td>
<td>0.1</td>
<td>0.2</td>
</tr>
</tbody>
</table>
Construct a test by estimating \( \| \mu_\kappa (\Delta_L P) \|_{\mathcal{H}_\kappa}^2 \), where \( \kappa = k \otimes l \otimes m \):

\[
\| \mu_\kappa (P_{XYZ} - P_{XY} P_Z - \cdots) \|_{\mathcal{H}_\kappa}^2 = \\
\langle \mu_\kappa P_{XYZ}, \mu_\kappa P_{XYZ} \rangle_{\mathcal{H}_\kappa} - 2 \langle \mu_\kappa P_{XYZ}, \mu_\kappa P_{XY} P_Z \rangle_{\mathcal{H}_\kappa} \cdots
\]
A kernel test statistic using Lancaster Measure

<table>
<thead>
<tr>
<th>$\nu \setminus \nu'$</th>
<th>$P_{XYZ}$</th>
<th>$P_{XY}P_{Z}$</th>
<th>$P_{XZ}P_{Y}$</th>
<th>$P_{YZ}P_{X}$</th>
<th>$P_{X}P_{Y}P_{Z}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$P_{XYZ}$</td>
<td>$(K \circ L \circ M)_{++}$</td>
<td>$((K \circ L) M)_{++}$</td>
<td>$((K \circ M) L)_{++}$</td>
<td>$((M \circ L) K)_{++}$</td>
<td>$\text{tr}(K_+ \circ L_+ \circ M_+)$</td>
</tr>
<tr>
<td>$P_{XY}P_{Z}$</td>
<td>$(K \circ L)<em>{++} M</em>{++}$</td>
<td>$(MKL)_{++}$</td>
<td>$(KLM)_{++}$</td>
<td>$(KL)<em>{++} M</em>{++}$</td>
<td>$(KL)<em>{++} M</em>{++}$</td>
</tr>
<tr>
<td>$P_{XZ}P_{Y}$</td>
<td>$(K \circ M)<em>{++} L</em>{++}$</td>
<td>$(KML)_{++}$</td>
<td>$(KM)<em>{++} L</em>{++}$</td>
<td>$L_{++} K_{++}$</td>
<td>$(LM)<em>{++} K</em>{++}$</td>
</tr>
<tr>
<td>$P_{YZ}P_{X}$</td>
<td></td>
<td>$(L \circ M)<em>{++} K</em>{++}$</td>
<td>$K_{++} L_{++} M_{++}$</td>
<td>$K_{++} L_{++} M_{++}$</td>
<td>$K_{++} L_{++} M_{++}$</td>
</tr>
<tr>
<td>$P_{X}P_{Y}P_{Z}$</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Table: $V$-statistic estimators of $\langle \mu_\kappa \nu, \mu_\kappa \nu' \rangle_{\mathcal{H}_\kappa}$ (without terms $P_X P_Y P_Z$). $H$ is centering matrix $I - n^{-1}$

**Lancaster interaction statistic:** Sejdinovic, G, Bergsma, NIPS13

$$\|\mu_\kappa (\Delta_L P)\|^2_{\mathcal{H}_\kappa} = \frac{1}{n^2} \left( HKH \circ H \circ L \circ H \circ R \circ M \circ H \right)_{++}.$$
## A kernel test statistic using Lancaster Measure

<table>
<thead>
<tr>
<th>( \nu \backslash \nu' )</th>
<th>( P_{XYZ} )</th>
<th>( P_{XY}P_Z )</th>
<th>( P_{XZ}P_Y )</th>
<th>( P_{YZ}P_X )</th>
<th>( P_XP_YP_Z )</th>
</tr>
</thead>
<tbody>
<tr>
<td>( P_{XYZ} )</td>
<td>((K \circ L \circ M)_{++})</td>
<td>(((K \circ L) M)_{++})</td>
<td>(((K \circ M) L)_{++})</td>
<td>(((M \circ L) K)_{++})</td>
<td>( \text{tr}(K_+ \circ L_+ \circ M_+) )</td>
</tr>
<tr>
<td>( P_{XY}P_Z )</td>
<td>((K \circ L)<em>{++} M</em>{++})</td>
<td>((M K L)_{++})</td>
<td>((K L M)_{++})</td>
<td>((K L)<em>{++} M</em>{++})</td>
<td></td>
</tr>
<tr>
<td>( P_{XZ}P_Y )</td>
<td>((K \circ M)<em>{++} L</em>{++})</td>
<td>((K M L)_{++})</td>
<td>((K M)<em>{++} L</em>{++})</td>
<td></td>
<td></td>
</tr>
<tr>
<td>( P_{YZ}P_X )</td>
<td>((L \circ M)<em>{++} K</em>{++})</td>
<td>((L M)<em>{++} K</em>{++})</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>( P_XP_YP_Z )</td>
<td>( K_{++} L_{++} M_{++})</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Table: \( V \)-statistic estimators of \( \langle \mu_\kappa \nu, \mu_\kappa \nu' \rangle_{\mathcal{H}_\kappa} \) (without terms \( P_XP_YP_Z \)). \( H \) is centering matrix \( I - n^{-1} \)

**Lancaster interaction statistic:** Sejdinovic, G, Bergsma, NIPS13

\[
|\mu_\kappa (\Delta LP)|_{\mathcal{H}_\kappa}^2 = \frac{1}{n^2} \left( H K H \circ H L H \circ H M H \right)_{++}.
\]

Empirical joint central moment in the feature space
V-structure discovery

Lancaster test, CI test for $X \perp Y \mid Z$ from Zhang et al. (2011), and a factorisation test, $n = 500$
Interaction measure valid for all $D$:

\[
\Delta S P = \sum_{\pi} (-1)^{|\pi|-1} (|\pi| - 1)! J_\pi P
\]

- For a partition $\pi$, $J_\pi$ associates to the joint the corresponding factorisation, e.g., $J_{13|2|4} P = X_1 X_3 X_2 X_4$. 
Interaction for $D \geq 4$

Interaction measure valid for all $D$:

(Streitberg, 1990)

$$\Delta SP = \sum_{\pi}(-1)^{|\pi|-1}(|\pi|-1)!J_{\pi}P$$

- For a partition $\pi$, $J_{\pi}$ associates to the joint the corresponding factorisation, e.g., $J_{13|2|4}P = P_{X_1 X_3} P_{X_2} P_{X_4}$. 
Interaction for $D \geq 4$

- Interaction measure valid for all $D$:

  $\Delta_s P = \sum_{\pi} (-1)^{|\pi|-1} (|\pi| - 1)! J_\pi P$

- For a partition $\pi$, $J_\pi$ associates to the joint the corresponding factorisation, e.g., $J_{13|2|4} P = P_{X_1 X_3} P_{X_2} P_{X_4}$.

![Bell numbers growth graph](image)
Co-authors

From Gatsby:
- Mikolaj Binkowski
- Kacper Chwialkowski
- Wittawat Jitkrittum
- Heiko Strathmann
- Dougal Sutherland
- Wenkai Xu

External collaborators:
- Kenji Fukumizu
- Bernhard Schoelkopf
- Dino Sejdinovic
- Bharath Sriperumbudur
- Alex Smola
- Zoltan Szabo
Questions?