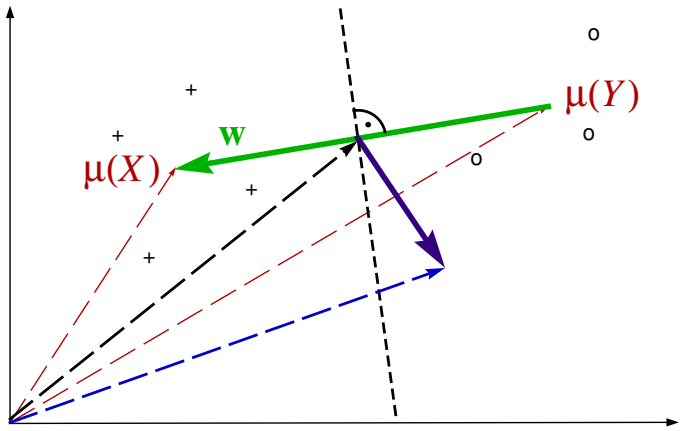
Kernel Means

Bernhard Schölkopf

joint work with

Karsten Borgwardt, Kenji Fukumizu, Arthur Gretton, Jiayuan Huang, Quoc Le, Malte Rasch, Alex Smola, Le Song, Xiaohai Sun

An example of a kernel algorithm, revisited



Schölkopf and Smola (2002)

$\mathcal{X}$ compact subset of a separable metric space, k positive definite kernel on $\mathcal{X}$ with RKHS $\mathcal{H}$.

Positive class $X := \{x_1, \ldots, x_m\} \subset \mathcal{X}$
Negative class $Y := \{y_1, \ldots, y_n\} \subset \mathcal{X}$

Means in the RKHS $\mu(X) = \frac{1}{m} \sum_{i=1}^{m} k(x_i, \cdot)$ and $\mu(Y) = \frac{1}{n} \sum_{i=1}^{n} k(y_i, \cdot)$.

Get a problem if $\mu(X) = \mu(Y)$. When does this happen?

$k(x, x') = \langle x, x' \rangle$:	the means coincide
$k(x, x') = (\langle x, x' \rangle + 1)^d$:	all empirical moments up to order d coincide
k strictly positive definite (e.g., Gaussian):	$X = Y$.

The mean “remembers” each point that contributed to it.

The mean map

$$\mu \colon X = (x_1, \ldots, x_m) \mapsto \frac{1}{m} \sum_{i=1}^{m} k(x_i, \cdot)$$

satisfies

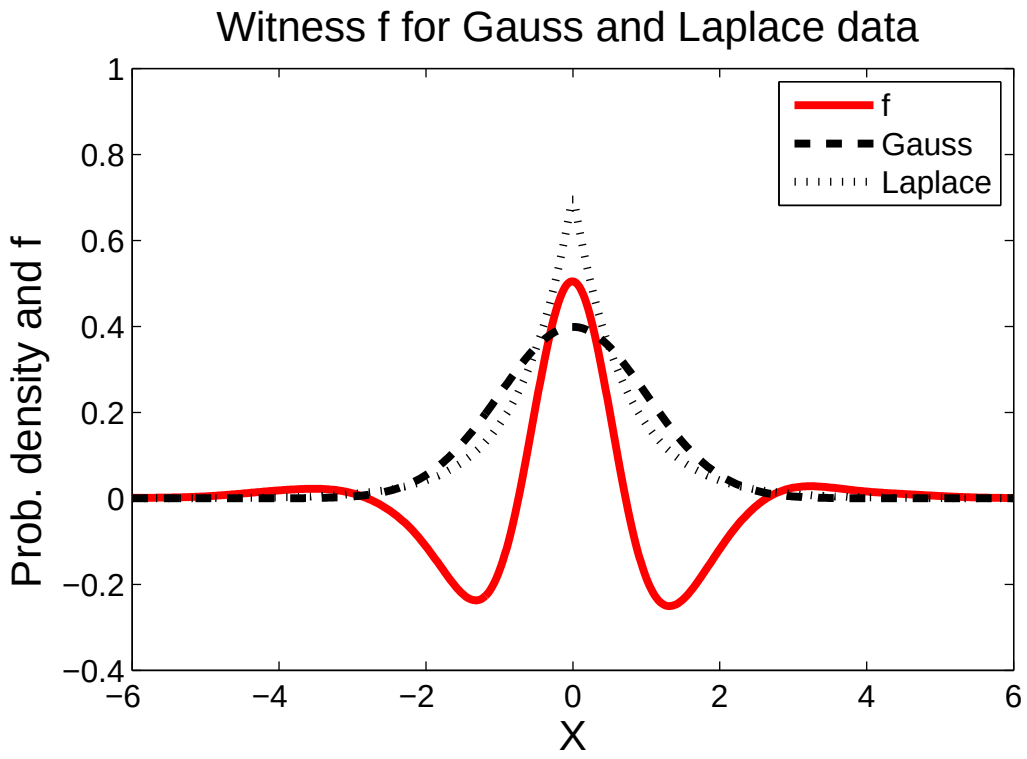
$$\langle \mu(X), f \rangle = \left\langle \frac{1}{m} \sum_{i=1}^{m} k(x_i, \cdot), f \right\rangle = \frac{1}{m} \sum_{i=1}^{m} f(x_i)$$

and

$$\|\mu(X) - \mu(Y)\| = \sup_{\|f\| \leq 1} |\langle \mu(X) - \mu(Y), f \rangle| = \sup_{\|f\| \leq 1} \left| \frac{1}{m} \sum_{i=1}^{m} f(x_i) - \frac{1}{n} \sum_{i=1}^{n} f(y_i) \right|.$$

Note: distance in the RKHS = solution of a high-dimensional optimization problem.

Witness function $f = \frac{\mu(X)-\mu(Y)}{\|\mu(X)-\mu(Y)\|}$, thus $f(x) \propto \langle \mu(X) - \mu(Y), k(x, .) \rangle$):



This function is in the RKHS of a Gaussian kernel, but not in the RKHS of the linear kernel.

The mean map for measures

p, q Borel probability measures,

$\mathbf{E}_{x,x' \sim p}[k(x, x')],\ \mathbf{E}_{x,x' \sim q}[k(x, x')] < \infty$. ($\|k(x, .)\| \leq M < \infty$ is sufficient)

Define

$$\mu : p \mapsto \mathbf{E}_{x \sim p}[k(x, \cdot)].$$

Note

$$\langle \mu(p), f \rangle = \mathbf{E}_{x \sim p}[f(x)]$$

and

$$\|\mu(p) - \mu(q)\| = \sup_{\|f\| \leq 1} \left| \mathbf{E}_{x \sim p}[f(x)] - \mathbf{E}_{x \sim q}[f(x)] \right|.$$

Recall that in the finite sample case, for strictly p.d. kernels, μ was injective — how about now?

Theorem 1 *[Fortet and Mourier (1953); Dudley (2002)]*

$$p = q \iff \sup_{f \in C(\mathcal{X})} |\mathbf{E}_{x \sim p}(f(x)) - \mathbf{E}_{x \sim q}(f(x))| = 0,$$

where $C(\mathcal{X})$ is the space of continuous bounded functions on $\mathcal{X}$.

Replace $C(\mathcal{X})$ by the unit ball in an RKHS that is dense in $C(\mathcal{X})$ — universal kernel (Steinwart, 2002) — e.g., Gaussian.

Theorem 2 *If k is universal, then*

$$p = q \Longleftrightarrow \|\mu(p) - \mu(q)\| = 0.$$

- μ is invertible on its image $\mathcal{M} = \{\mu(p) \mid p \text{ is a probability distribution}\}$ (the "marginal polytope", Wainwright and Jordan (2003))
- generalization of the *moment generating function* of a RV x with distribution p:

$$M_p(.) = \mathbf{E}_{x\sim p}\left[e^{\langle x,\,\cdot\,\rangle}\right].$$

Uniform convergence bounds

Let X be an i.i.d. m-sample from p. The discrepancy

$$\|\mu(p) - \mu(X)\| = \sup_{\|f\|\leq 1}\left|\mathbf{E}_{x\sim p}[f(x)] - \frac{1}{m}\sum_{i=1}^{m} f(x_i)\right|$$

can be bounded using uniform convergence methods (Smola et al., 2007).

Application 1: Two-sample problem

Gretton et al. (2007)

X, Y i.i.d. m-samples from p, q, respectively.

$$\|\mu(p)-\mu(q)\|^2 = \mathbf{E}_{x,x'\sim p}\left[k(x,x')\right]-2\mathbf{E}_{x\sim p,y\sim q}\left[k(x,y)\right]+\mathbf{E}_{y,y'\sim q}\left[k(y,y')\right] = \mathbf{E}_{x,x'\sim p,y,y'\sim q}\left[h((x,y),(x',y'))\right]$$

with

$$h((x,y),(x',y')) := k(x,x') - k(x,y') - k(y,x') + k(y,y').$$

Define

$$D(p,q)^2 := \mathbf{E}_{x,x'\sim p,y,y'\sim q} h((x,y),(x',y'))$$
$$\hat{D}(X,Y)^2 := \frac{1}{m(m-1)} \sum_{i\neq j} h((x_i,y_i),(x_j,y_j)).$$

$\hat{D}(X,Y)^2$ is an unbiased estimator of $D(p,q)^2$.

It's easy to compute, and works on structured data.

Theorem 3 *Assume k is bounded.*
$\hat{D}(X,Y)^2$ converges to $D(p,q)^2$ in probability with rate $\mathcal{O}(m^{-\frac{1}{2}})$.

This *could* be used as a basis for a test, but uniform convergence bounds are often loose..

Theorem 4 *We assume* $\mathbf{E}\left(h^2\right) < \infty$. *When* $p \neq q$, *then* $\sqrt{m}(\hat{D}(X,Y)^2 - D(p,q)^2)$ *converges in distribution to a zero mean Gaussian with variance*

$$\sigma_u^2 = 4\left(\mathbf{E}_z\left[(\mathbf{E}_{z'}h(z,z'))^2\right] - \left[\mathbf{E}_{z,z'}(h(z,z'))\right]^2\right).$$

When $p = q$, *then* $m(\hat{D}(X,Y)^2 - D(p,q)^2) = m\hat{D}(X,Y)^2$ *converges in distribution to*

$$\sum_{l=1}^{\infty} \lambda_l \left[q_l^2 - 2\right], \tag{1}$$

where $q_l \sim \mathcal{N}(0,2)$ *i.i.d.,* λ_i *are the solutions to the eigenvalue equation*

$$\int_x \tilde{k}(x,x')\psi_i(x)dp(x) = \lambda_i\psi_i(x'),$$

and $\tilde{k}(x_i,x_j) := k(x_i,x_j) - \mathbf{E}_x k(x_i,x) - \mathbf{E}_x k(x,x_j) + \mathbf{E}_{x,x'}k(x,x')$ *is the centred RKHS kernel.*

Application 2: Dependence Measures

Assume that (x, y) are drawn from p_{xy}, with marginals p_x, p_y.
Want to know whether p_{xy} factorizes.

Bach and Jordan (2002); Fukumizu et al. (2004): kernel generalized variance
Gretton et al. (2005a,b): kernel constrained covariance, HSIC

Main idea (Jacod and Protter, 2000; Rényi, 1959):

x and y independent $\iff \forall$ bounded continuous functions f, g, we have $\mathrm{Cov}(f(x), g(y)) = 0$.

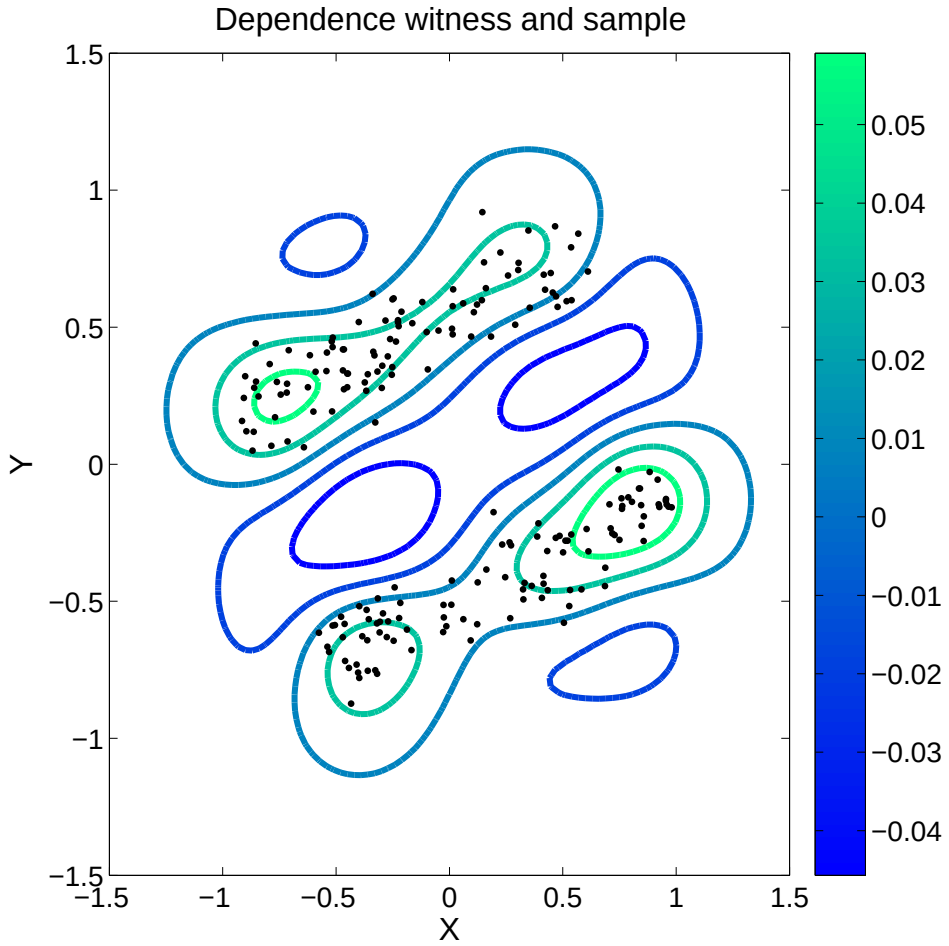
k kernel on $\mathcal{X} \times \mathcal{Y}$.

$$\mu(p_{xy}) := \mathbf{E}_{(x,y)\sim p}\left[k((x, y), \cdot)\right]$$
$$\mu(p_x \times p_y) := \mathbf{E}_{x\sim p_x, y\sim p_y}\left[k((x, y), \cdot)\right].$$

Use $\Delta := \|\mu(p_{xy}) - \mu(p_x \times p_y)\|$ as a measure of dependence.

For $k((x, y), (x', y')) = k_x(x, x')k_y(y, y')$, can show Δ^2 equals the Hilbert-Schmidt norm of the covariance operator between the two RKHSs (HSIC), with empirical estimate $m^{-2}\,\mathrm{tr}\, HK_xHK_y$, where $H = I - \mathbf{1}/m$ (Gretton et al., 2005a; Smola et al., 2007).

Witness function of the equivalent optimisation problem:



5

Cf. the talk of Xiaohai Sun, who uses the *conditional* cross-covariance operator (Fukumizu et al., 2004) to learn causal structures (cf. his talk on Saturday before lunch)

Application 3: Covariate Shift Correction and Local Learning

training set $X = \{(x_1, y_1), \ldots, (x_m, y_m)\}$ drawn from p, test set $X' = \{(x'_1, y'_1), \ldots, (x'_n, y'_n)\}$ from $p' \neq p$. Assume $p_{y|x} = p'_{y|x}$.

Shimodaira (2000): reweight training set

Minimize

$$\Delta := \left\| \sum_{i=1}^{m} \beta_i k(x_i, \cdot) - \mu(X') \right\| \quad \text{subject to} \quad \beta_i \geq 0, \quad \sum_i \beta_i = 1.$$

In practice, minimize $\Delta^2 + \lambda \|\beta\|_2^2$. This is equivalent to the QP:

$$\begin{aligned} &\underset{\beta}{\text{minimize}} \ \frac{1}{2} \beta^\top (K + \lambda \mathbf{1}) \beta - \beta^\top l \\ &\text{subject to } \beta_i \geq 0 \text{ and } \sum_i \beta_i = 1, \end{aligned}$$

where $K_{ij} := k(x_i, x_j)$, $l_i = \langle k(x_i, \cdot), \mu(X') \rangle$.

Experiments show that in underspecified situations (e.g., large kernel widths), this helps (Huang et al., 2007).

$X' = \{x'\}$ leads to a local sample weighting scheme.

Application 4: Measure estimation and dataset squashing

Dudík et al. (2004); Balakrishnan and Schonfeld (2006); Altun and Smola (2006); Smola et al. (2007)

Given a sample X, minimize

$$\|\mu(X) - \mu(p)\|$$

over a convex combination of measures p_i,

$$p = \sum_i \alpha_i p_i, \quad \alpha_i \geq 0, \quad \sum_i \alpha_i = 1.$$

This can be written as a convex QP with objective function

$$\|\mu(X) - \mu(p)\|^2 = \alpha^\top Q \alpha + 1_m^\top K 1_m - 2\alpha^\top L 1_m,$$

where

$$\begin{aligned} L_{ij} :=& \mathbf{E}_{x \sim p_i}[k(x, x_j)] \\ Q_{ij} :=& \mathbf{E}_{x \sim p_i, x' \sim p_j}[k(x, x')] \\ K_{ij} =& k(x_i, x_j) \\ 1_m :=& (1/m, \ldots, 1/m)^\top \in \mathbb{R}^m. \end{aligned}$$

In practice, use

$$\alpha^\top [Q + \lambda I]\alpha - 2\alpha^\top L 1_m$$

Some cases where Q and L can be computed in closed form (Smola et al., 2007):

- Gaussian p_i and k
- Dirac measures $p_i = \delta_{x_i}$ (dataset squashing, DuMouchel et al. (1999))
- X test set, Dirac measures $p_i = \delta_{y_i}$ centered on the training points Y — covariate shift, see above

References

Y. Altun and A.J. Smola. Unifying divergence minimization and statistical inference via convex duality. In H.U. Simon and G. Lugosi, editors, *Proc. Annual Conf. Computational Learning Theory*, LNCS, pages 139–153. Springer, 2006.

F. R. Bach and M. I. Jordan. Kernel independent component analysis. *J. Mach. Learn. Res.*, 3: 1–48, 2002.

N. Balakrishnan and D. Schonfeld. A maximum entropy kernel density estimator with applications to function interpolation and texture segmentation. In *SPIE Proceedings of Electronic*

Imaging: Science and Technology. Conference on Computational Imaging IV, San Jose, CA, 2006.

M. Dudík, S. Phillips, and R.E. Schapire. Performance guarantees for regularized maximum entropy density estimation. In *Proc. Annual Conf. Computational Learning Theory*. Springer Verlag, 2004.

R. M. Dudley. *Real analysis and probability*. Cambridge University Press, Cambridge, UK, 2002.

W. DuMouchel, C. Volinsky, C. Cortes, D. Pregibon, and T. Johnson. Squashing flat files flatter. In *International Conference on Knowledge Discovery and Data Mining (KDD)*, 1999.

R. Fortet and E. Mourier. Convergence de la réparation empirique vers la réparation théorique. *Ann. Scient. École Norm. Sup.*, 70:266–285, 1953.

K. Fukumizu, F. R. Bach, and M. I. Jordan. Dimensionality reduction for supervised learning with reproducing kernel hilbert spaces. *Journal of Machine Learning Research*, 5:73–99, 2004.

A. Gretton, K. Borgwardt, M. Rasch, B. Schölkopf, and A. J. Smola. A kernel method for the two-sample-problem. In B. Schölkopf, J. Platt, and T. Hofmann, editors, *Advances in Neural Information Processing Systems 19*, volume 19. The MIT Press, Cambridge, MA, 2007.

A. Gretton, O. Bousquet, A.J. Smola, and B. Schölkopf. Measuring statistical dependence

with Hilbert-Schmidt norms. In S. Jain, H. U. Simon, and E. Tomita, editors, *Proceedings Algorithmic Learning Theory*, pages 63–77, Berlin, Germany, 2005a. Springer-Verlag.

A. Gretton, R. Herbrich, A. Smola, O. Bousquet, and B. Schölkopf. Kernel methods for measuring independence. *J. Mach. Learn. Res.*, 6:2075–2129, 2005b.

J. Huang, A.J. Smola, A. Gretton, K. Borgwardt, and B. Schölkopf. Correcting sample selection bias by unlabeled data. In B. Schölkopf, J. Platt, and T. Hofmann, editors, *Advances in Neural Information Processing Systems 19*, volume 19. The MIT Press, Cambridge, MA, 2007.

J. Jacod and P. Protter. *Probability Essentials*. Springer, New York, 2000.

A. Rényi. On measures of dependence. *Acta Math. Acad. Sci. Hungar.*, 10:441–451, 1959.

B. Schölkopf and A. Smola. *Learning with Kernels*. MIT Press, Cambridge, MA, 2002.

H. Shimodaira. Improving predictive inference under convariance shift by weighting the log-likelihood function. *Journal of Statistical Planning and Inference*, 90, 2000.

A. J. Smola, A. Gretton, K. Borgwardt, L. Song, and B. Schölkopf. A hilbert space embedding for distributions. In *Proc. Intl. Conf. Algorithmic Learning Theory*, 2007.

I. Steinwart. On the influence of the kernel on the consistency of support vector machines. *J. Mach. Learn. Res.*, 2:67–93, 2002.

M. J. Wainwright and M. I. Jordan. Graphical models, exponential families, and variational inference. Technical Report 649, UC Berkeley, Department of Statistics, September 2003.

7

8

9