Distribution Regression and Beyond

Zoltán Szabó

PhD Open Day, LSE
Oct 14, 2021
(afternoon)

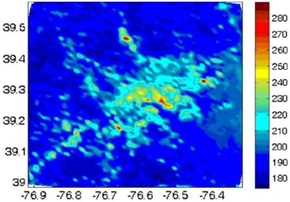
Motivating example

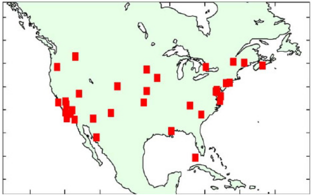
- **Goal**: aerosol prediction.

Motivating example

- **Goal**: aerosol prediction.

- Prediction using labelled bags:
  - bag := multi-spectral satellite measurements over an area,
  - label := local aerosol value.

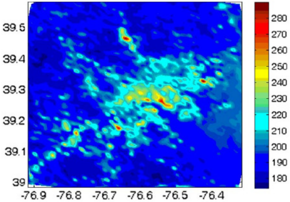


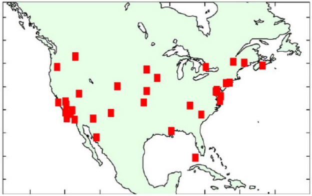


Motivating example

- **Goal**: aerosol prediction.

- Prediction using labelled bags:
  - bag := multi-spectral satellite measurements over an area,
  - label := local aerosol value.

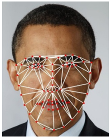


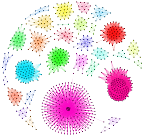


Needed

similarity of bags (or probability distributions)!

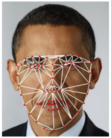
More generally: objects in the bags

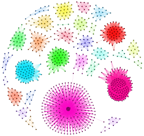




- Examples:
  - time-series modelling: user = set of time-series,
  - computer vision: image = collection of patch vectors,
  - NLP: corpus = bag of documents,
  - network analysis: group of people = bag of friendship graphs, ...

More generally: objects in the bags





- Examples:
  - time-series modelling: user = set of time-series,
  - computer vision: image = collection of patch vectors,
  - NLP: corpus = bag of documents,
  - network analysis: group of people = bag of friendship graphs, ...
- Wider context (statistics): point estimation tasks.

From similarity on $\mathbb{R}^d$

- On $\mathbb{R}^d$: we have a natural measure of similarity $\Rightarrow \|\cdot\|$, $\sphericalangle$.

$$\mathbb{R} \ni \langle \mathbf{x}, \mathbf{y} \rangle := \sum_{i=1}^{d} x_i y_i, \quad \mathbf{x}, \mathbf{y} \in \mathbb{R}^d.$$

From similarity on $\mathbb{R}^d$

- On $\mathbb{R}^d$: we have a natural measure of similarity $\Rightarrow \|\cdot\|, \sphericalangle$.

$$\mathbb{R} \ni \langle \mathbf{x}, \mathbf{y} \rangle := \sum_{i=1}^{d} x_i y_i, \quad \mathbf{x}, \mathbf{y} \in \mathbb{R}^d.$$

- Generalized inner product on objects in $\mathcal{X}$ (a.k.a. kernel):

$$\mathbb{R} \ni k(x, y) := \langle \underbrace{\varphi(x)}_{\text{feature of } x}, \varphi(y) \rangle_{\mathcal{H}}, \quad x, y \in \mathcal{X}.$$

From similarity on $\mathbb{R}^d$

- On $\mathbb{R}^d$: we have a natural measure of similarity $\Rightarrow \|\cdot\|, \sphericalangle$.

$$\mathbb{R} \ni \langle \mathbf{x}, \mathbf{y} \rangle := \sum_{i=1}^{d} x_i y_i, \quad \mathbf{x}, \mathbf{y} \in \mathbb{R}^d.$$

- Generalized inner product on objects in $\mathcal{X}$ (a.k.a. kernel):

$$\mathbb{R} \ni k(x, y) := \langle \underbrace{\varphi(x)}_{\text{feature of } x}, \varphi(y) \rangle_{\mathcal{H}}, \quad x, y \in \mathcal{X}.$$

Notes

- Examples: $k_P(\mathbf{x}, \mathbf{y}) = (\langle \mathbf{x}, \mathbf{y} \rangle + \gamma)^p$, $k_G(\mathbf{x}, \mathbf{y}) = e^{-\gamma \|\mathbf{x}-\mathbf{y}\|_2^2}$.
- One can choose $\varphi(x) = k(\cdot, x)$.
- RKHS: $\mathcal{H}_k = \overline{\left\{\sum_{i=1}^{n} \alpha_i k(\cdot, x_i)\right\}} \xrightarrow{\text{spec}}$ polynomials, splines, Fourier analysis, ...

Kernels exist on various objects

Few examples:

- strings
  [Watkins, 1999, Lodhi et al., 2002, Leslie et al., 2002, Kuang et al., 2004, Leslie and Kuang, 2004, Saigo et al., 2004, Cuturi and Vert, 2005],
- time series [Rüping, 2001, Cuturi et al., 2007, Cuturi, 2011, Király and Oberhauser, 2019],
- trees [Collins and Duffy, 2001, Kashima and Koyanagi, 2002],
- groups and specifically rankings [Cuturi et al., 2005, Jiao and Vert, 2016],
- sets [Haussler, 1999, Gärtner et al., 2002],
- various generative models [Jaakkola and Haussler, 1999, Tsuda et al., 2002, Seeger, 2002, Jebara et al., 2004],
- fuzzy domains [Guevara et al., 2017], or
- graphs
  [Kondor and Lafferty, 2002, Gärtner et al., 2003, Kashima et al., 2003, Borgwardt and Kriegel, 2005, Shervashidze et al., 2009, Vishwanathan et al., 2010, Kondor and Pan, 2016, Draief et al., 2018, Bai et al., 2020, Borgwardt et al., 2020].

Similarity of bags (or probability distributions)

- Characteristic function:

$$\mathbb{P} \mapsto c(\mathbb{P}) = \int_{\mathbb{R}^d} e^{i\langle \cdot, \mathbf{x} \rangle} \mathrm{d}\mathbb{P}(\mathbf{x}).$$

[other examples: $\mathbb{I}_{(-\infty,\cdot)}(x)$, $e^{\langle \cdot, x \rangle}$]

Similarity of bags (or probability distributions)

- Characteristic function:

$$\mathbb{P} \mapsto c(\mathbb{P}) = \int_{\mathbb{R}^d} e^{i\langle \cdot, \mathbf{x} \rangle} \mathrm{d}\mathbb{P}(\mathbf{x}).$$

[other examples: $\mathbb{I}_{(-\infty,\cdot)}(x)$, $e^{\langle \cdot, x \rangle}$]

- Mean embedding with $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$:

$$\mathbb{P} \mapsto \mu_k(\mathbb{P}) = \int_{\mathcal{X}} \varphi(x) \mathrm{d}\mathbb{P}(x).$$

Similarity of bags (or probability distributions)

- Characteristic function:

$$\mathbb{P} \mapsto c(\mathbb{P}) = \int_{\mathbb{R}^d} e^{i\langle \cdot, \mathbf{x}\rangle} \mathrm{d}\mathbb{P}(\mathbf{x}).$$

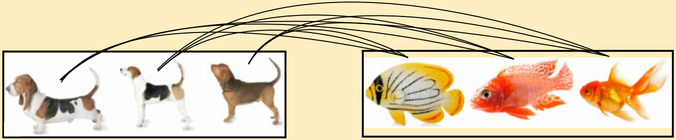
[other examples: $\mathbb{I}_{(-\infty,\cdot)}(x)$, $e^{\langle \cdot, x\rangle}$]

- Mean embedding with $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$:

$$\mathbb{P} \mapsto \mu_k(\mathbb{P}) = \int_{\mathcal{X}} \varphi(x) \mathrm{d}\mathbb{P}(x).$$

Induced similarity: set kernel [Haussler, 1999, Gärtner et al., 2002]

$$\langle \mu_k(\mathbb{P}_N), \mu_k(\mathbb{Q}_M)\rangle_{\mathcal{H}_k} = \frac{1}{NM} \sum_{n=1}^{N} \sum_{m=1}^{M} k(x_n, x'_m).$$



μ_k, MMD and friends

Applications:

- two-sample testing [Baringhaus and Franz, 2004, Székely and Rizzo, 2004, Székely and Rizzo, 2005, Borgwardt et al., 2006, Harchaoui et al., 2007, Gretton et al., 2012, Jitkrittum et al., 2016], and its differential private variant [Raj et al., 2019]; independence [Gretton et al., 2008, Pfister et al., 2018, Jitkrittum et al., 2017a] and goodness-of-fit testing [Jitkrittum et al., 2017b, Balasubramanian et al., 2017], causal discovery [Mooij et al., 2016, Pfister et al., 2018],
- domain adaptation [Zhang et al., 2013], -generalization [Blanchard et al., 2017], change-point detection [Harchaoui and Cappé, 2007], post selection inference [Yamada et al., 2018],
- kernel Bayesian inference [Song et al., 2011, Fukumizu et al., 2013], approximate Bayesian computation [Park et al., 2016], probabilistic programming [Schölkopf et al., 2015], model criticism [Lloyd et al., 2014, Kim et al., 2016],
- topological data analysis [Kusano et al., 2016],
- distribution classification [Muandet et al., 2011, Lopez-Paz et al., 2015], distribution regression [Szabó et al., 2016, Zaheer et al., 2017, Law et al., 2018, Fang et al., 2019, Mücke, 2021],
- generative adversarial networks [Dziugaite et al., 2015, Li et al., 2015, Binkowski et al., 2018], understanding the dynamics of complex dynamical systems [Klus et al., 2018, Klus et al., 2019], . . .

Aerosol prediction = regression on labelled bags

- Given:
  - labelled bags: $\hat{\mathbf{z}} = \left\{\left(\hat{\mathbb{P}}_i, y_i\right)\right\}_{i=1}^{\ell}$, $\hat{\mathbb{P}}_i$: bag from $\mathbb{P}_i$, $N := |\hat{\mathbb{P}}_i|$.
  - test bag: $\hat{\mathbb{P}}$.

Aerosol prediction = regression on labelled bags

- Given:
  - labelled bags: $\hat{\mathbf{z}} = \left\{\left(\hat{\mathbb{P}}_i, y_i\right)\right\}_{i=1}^{\ell}$, $\hat{\mathbb{P}}_i$: bag from $\mathbb{P}_i$, $N := |\hat{\mathbb{P}}_i|$.
  - test bag: $\hat{\mathbb{P}}$.
- Estimator:

$$f_{\hat{\mathbf{z}}}^{\lambda} = \underset{f \in \mathcal{H}_K}{\arg\min} \frac{1}{\ell} \sum_{i=1}^{\ell} \left[f\left(\underbrace{\mu_k(\hat{\mathbb{P}}_i)}_{\text{feature of the i-th bag}} \right) - y_i \right]^2 + \lambda \left\| f \right\|_{\mathcal{H}_K}^2 .$$

Aerosol prediction = regression on labelled bags

- Given:
  - labelled bags: $\hat{\mathbf{z}} = \left\{\left(\hat{\mathbb{P}}_i, y_i\right)\right\}_{i=1}^{\ell}$, $\hat{\mathbb{P}}_i$: bag from $\mathbb{P}_i$, $N := |\hat{\mathbb{P}}_i|$.
  - test bag: $\hat{\mathbb{P}}$.
- Estimator:

$$f_{\hat{\mathbf{z}}}^{\lambda} = \underset{f \in \mathcal{H}_K}{\arg\min} \frac{1}{\ell} \sum_{i=1}^{\ell} \left[f\left(\underbrace{\mu_k(\hat{\mathbb{P}}_i)}_{\text{feature of the i-th bag}} \right) - y_i \right]^2 + \lambda \|f\|_{\mathcal{H}_K}^2 .$$

- Prediction:

$$\hat{y}(\hat{\mathbb{P}}) = \mathbf{g}^T(\mathbf{G} + \ell\lambda\mathbf{I}_\ell)^{-1}\mathbf{y},$$

$$\mathbf{g} = \left[K\left(\mu_k(\hat{\mathbb{P}}), \mu_k(\hat{\mathbb{P}}_i)\right)\right], \mathbf{G} = \left[K\left(\mu_k(\hat{\mathbb{P}}_i), \mu_k(\hat{\mathbb{P}}_j)\right)\right], \mathbf{y} = [y_i].$$

Aerosol prediction = regression on labelled bags

- Given:
  - labelled bags: $\hat{\mathbf{z}} = \left\{\left(\hat{\mathbb{P}}_i, y_i\right)\right\}_{i=1}^{\ell}$, $\hat{\mathbb{P}}_i$: bag from $\mathbb{P}_i$, $N := |\hat{\mathbb{P}}_i|$.
  - test bag: $\hat{\mathbb{P}}$.
- Estimator:

$$f_{\hat{\mathbf{z}}}^{\lambda} = \operatorname*{arg\,min}_{f \in \mathcal{H}_K} \frac{1}{\ell} \sum_{i=1}^{\ell} \Big[f\big(\underbrace{\mu_k(\hat{\mathbb{P}}_i)}_{\text{feature of the i-th bag}} \big) - y_i \Big]^2 + \lambda \|f\|_{\mathcal{H}_K}^2 .$$

- Prediction:

$$\hat{y}(\hat{\mathbb{P}}) = \mathbf{g}^T (\mathbf{G} + \ell\lambda \mathbf{I}_\ell)^{-1} \mathbf{y},$$

$$\mathbf{g} = \left[K\left(\mu_k(\hat{\mathbb{P}}), \mu_k(\hat{\mathbb{P}}_i) \right) \right], \mathbf{G} = \left[K\left(\mu_k(\hat{\mathbb{P}}_i), \mu_k(\hat{\mathbb{P}}_j) \right) \right], \mathbf{y} = [y_i].$$

Challenge

Consistent? How many samples per bag?

Performance measure, baseline

Quality of estimator, baseline:

$$\mathcal{R}(f) = \mathbb{E}_{(\mu_k(\mathbb{P}),y)\sim\rho}[f(\mu_k(\mathbb{P})) - y]^2,$$
$$f_\rho = \text{regression function (assume now: } f_\rho \in \mathcal{H}_K).$$

How many samples/bag to achieve the accuracy of f_ρ? Possible?

Performance measure, baseline

Quality of estimator, baseline:

$$\mathcal{R}(f) = \mathbb{E}_{(\mu_k(\mathbb{P}),y)\sim\rho}[f(\mu_k(\mathbb{P})) - y]^2,$$

$$f_\rho = \text{regression function (assume now: } f_\rho \in \mathcal{H}_K).$$

How many samples/bag to achieve the accuracy of f_ρ? Possible?

Blanket assumptions:

1. $\mathcal{X}$: separable topological; k: bounded & continuous.
2. y: bounded; Y: separable Hilbert.
3. K: bounded, Hölder continuous.

Result [Szabó et al., 2016]: Hölder exponent = 1 below

- Known [Caponnetto and De Vito, 2007]: optimal rate

$$\mathcal{R}(f_{\mathbf{z}}^{\lambda}) - \mathcal{R}(f_{\rho}) = \mathcal{O}_{p}\left(\ell^{-\frac{bc}{bc+1}}\right),$$

b – size of the input space, c – smoothness of f_{ρ}.

Result [Szabó et al., 2016]: Hölder exponent = 1 below

- Known [Caponnetto and De Vito, 2007]: optimal rate

$$\mathcal{R}(f_{\mathbf{z}}^{\lambda}) - \mathcal{R}(f_{\rho}) = \mathcal{O}_p\left(\ell^{-\frac{bc}{bc+1}}\right),$$

b – size of the input space, c – smoothness of f_{ρ}.

- Let $N = \tilde{\mathcal{O}}(\ell^a)$. N: size of the bags. ℓ: number of bags.

Our result

- If $2 \leq a$, then $f_{\hat{\mathbf{z}}}^{\lambda}$ attains the optimal rate.

Result [Szabó et al., 2016]: Hölder exponent = 1 below

- Known [Caponnetto and De Vito, 2007]: optimal rate

$$\mathcal{R}(f_{\mathbf{z}}^{\lambda}) - \mathcal{R}(f_{\rho}) = \mathcal{O}_p\left(\ell^{-\frac{bc}{bc+1}}\right),$$

  b – size of the input space, c – smoothness of f_{ρ}.

- Let $N = \tilde{\mathcal{O}}(\ell^a)$. N: size of the bags. ℓ: number of bags.

Our result

- If $2 \leq a$, then $f_{\hat{\mathbf{z}}}^{\lambda}$ attains the optimal rate.
- In fact, $a = \frac{b(c+1)}{bc+1} < 2$ is enough.

Result [Szabó et al., 2016]: Hölder exponent = 1 below

- Known [Caponnetto and De Vito, 2007]: optimal rate

$$\mathcal{R}(f_{\mathbf{z}}^{\lambda}) - \mathcal{R}(f_{\rho}) = \mathcal{O}_p\left(\ell^{-\frac{bc}{bc+1}}\right),$$

b – size of the input space, c – smoothness of f_{ρ}.

- Let $N = \tilde{\mathcal{O}}(\ell^a)$. N: size of the bags. ℓ: number of bags.

Our result

- If $2 \leq a$, then $f_{\hat{\mathbf{z}}}^{\lambda}$ attains the optimal rate.
- In fact, $a = \frac{b(c+1)}{bc+1} < 2$ is enough.
- Similar result holds for the misspecified setting.

Result [Szabó et al., 2016]: Hölder exponent = 1 below

- Known [Caponnetto and De Vito, 2007]: optimal rate

$$\mathcal{R}(f_{\mathbf{z}}^{\lambda}) - \mathcal{R}(f_{\rho}) = \mathcal{O}_p\left(\ell^{-\frac{bc}{bc+1}}\right),$$

b – size of the input space, c – smoothness of f_{ρ}.

- Let $N = \tilde{\mathcal{O}}(\ell^a)$. N: size of the bags. ℓ: number of bags.

Our result

- If $2 \le a$, then $f_{\hat{\mathbf{z}}}^{\lambda}$ attains the optimal rate.
- In fact, $a = \frac{b(c+1)}{bc+1} < 2$ is enough.
- Similar result holds for the misspecified setting.
- Set kernel is consistent in regression (17-year-old open): ✓

Result [Szabó et al., 2016]: Hölder exponent = 1 below

- Known [Caponnetto and De Vito, 2007]: optimal rate

$$\mathcal{R}(f_{\mathbf{z}}^{\lambda}) - \mathcal{R}(f_{\rho}) = \mathcal{O}_p\left(\ell^{-\frac{bc}{bc+1}}\right),$$

b – size of the input space, c – smoothness of f_{ρ}.

- Let $N = \tilde{\mathcal{O}}(\ell^a)$. N: size of the bags. ℓ: number of bags.

Our result

- If $2 \leq a$, then $f_{\hat{\mathbf{z}}}^{\lambda}$ attains the optimal rate.
- In fact, $a = \frac{b(c+1)}{bc+1} < 2$ is enough.
- Similar result holds for the misspecified setting.
- Set kernel is consistent in regression (17-year-old open): ✓
- [Fang et al., 2019] (shorter proof; log-improvement), [Zaheer et al., 2017] (deep net specialization), [Mücke, 2021] (SGD), ...

Various research questions

- Adaptivity & reduced memory footprint; spectral methods [Neubauer et al., 1996, Blanchard and Mücke, 2018, Lin et al., 2020]:
  - Kernel ridge regression: $\mathbf{y} = [y_i]_{i\in[n]}$

$$f_{\mathbf{z}}^{\lambda}(x) = \sum_{i\in[n]} \alpha_i K(x, x_i),\ \boldsymbol{\alpha} = (\mathbf{K} + n\lambda\mathbf{I}_n)^{-1}\mathbf{y}.$$

Various research questions

- Adaptivity & reduced memory footprint; spectral methods [Neubauer et al., 1996, Blanchard and Mücke, 2018, Lin et al., 2020]:
  - Kernel ridge regression: $\mathbf{y} = [y_i]_{i\in[n]}$

$$f_{\mathbf{z}}^{\lambda}(x) = \sum_{i\in[n]} \alpha_i K(x, x_i),\ \boldsymbol{\alpha} = (\mathbf{K} + n\lambda\mathbf{I}_n)^{-1}\mathbf{y}.$$

  - This falls with $g_\lambda(\sigma) = \frac{1}{\sigma+\lambda}$ under the umbrella

$$f_{\mathbf{z}}^{\lambda}(x) = \sum_{i\in[n]} \alpha_i K(x, x_i),\ \boldsymbol{\alpha} = \frac{1}{n} g_\lambda\left(\frac{\mathbf{K}}{n}\right)\mathbf{y}.$$

- Scaling:
  - RFF [Rahimi and Recht, 2007] $\xrightarrow{\text{exp. boost}}$ [Sriperumbudur and Szabó, 2015].

Various research questions

- Adaptivity & reduced memory footprint ; spectral methods [Neubauer et al., 1996, Blanchard and Mücke, 2018, Lin et al., 2020]:
  - Kernel ridge regression: $\mathbf{y} = [y_i]_{i\in[n]}$

$$f_{\mathbf{z}}^{\lambda}(x) = \sum_{i\in[n]} \alpha_i K(x, x_i),\ \boldsymbol{\alpha} = (\mathbf{K} + n\lambda \mathbf{I}_n)^{-1}\mathbf{y}.$$

  - This falls with $g_\lambda(\sigma) = \frac{1}{\sigma+\lambda}$ under the umbrella

$$f_{\mathbf{z}}^{\lambda}(x) = \sum_{i\in[n]} \alpha_i K(x, x_i),\ \boldsymbol{\alpha} = \frac{1}{n} g_\lambda\left(\frac{\mathbf{K}}{n}\right)\mathbf{y}.$$

- Scaling :
  - RFF [Rahimi and Recht, 2007] $\xrightarrow{\text{exp. boost}}$ [Sriperumbudur and Szabó, 2015].
- Novel applications .

Thank you for the attention!

Bai, L., Cui, L., Rossi, L., Xu, L., Bai, X., and Hancock, E. (2020).
Local-global nested graph kernels using nested complexity traces.
Pattern Recognition Letters, 134:87–95.

Balasubramanian, K., Li, T., and Yuan, M. (2017).
On the optimality of kernel-embedding based goodness-of-fit tests.
Technical report.
(https://arxiv.org/abs/1709.08148).

Baringhaus, L. and Franz, C. (2004).
On a new multivariate two-sample test.
Journal of Multivariate Analysis, 88:190–206.

Binkowski, M., Sutherland, D., Arbel, M., and Gretton, A. (2018).
Demystifying MMD GANs.

In *International Conference on Learning Representations (ICLR)*.

Blanchard, G., Deshmukh, A. A., Dogan, U., Lee, G., and Scott, C. (2017).
Domain generalization by marginal transfer learning.
Technical report.
(https://arxiv.org/abs/1711.07910).

Blanchard, G. and Mücke, N. (2018).
Optimal rates for regularization of statistical inverse learning problems.
Foundations of Computational Mathematics, 18(4):971–1013.

Borgwardt, K., Ghisu, E., Llinares-López, F., O'Bray, L., and Riec, B. (2020).
Graph kernels: State-of-the-art and future challenges.
Foundations and Trends in Machine Learning, 13(5-6):531–712.

Borgwardt, K., Gretton, A., Rasch, M. J., Kriegel, H.-P., Schölkopf, B., and Smola, A. J. (2006).
Integrating structured biological data by kernel maximum mean discrepancy.
Bioinformatics, 22:e49–57.

Borgwardt, K. M. and Kriegel, H.-P. (2005).
Shortest-path kernels on graphs.
In *International Conference on Data Mining (ICDM)*, pages 74–81.

Caponnetto, A. and De Vito, E. (2007).
Optimal rates for regularized least-squares algorithm.
Foundations of Computational Mathematics, 7:331–368.

Collins, M. and Duffy, N. (2001).
Convolution kernels for natural language.
In *Advances in Neural Information Processing Systems (NIPS)*, pages 625–632.

Cuturi, M. (2011).

Fast global alignment kernels.
In *International Conference on Machine Learning (ICML)*, pages 929–936.

Cuturi, M., Fukumizu, K., and Vert, J.-P. (2005).
Semigroup kernels on measures.
Journal of Machine Learning Research, 6:1169–1198.

Cuturi, M. and Vert, J.-P. (2005).
The context-tree kernel for strings.
Neural Networks, 18(8):1111–1123.

Cuturi, M., Vert, J.-P., Birkenes, O., and Matsui, T. (2007).
A kernel for time series based on global alignments.
In *International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, pages 413–416.

Draief, M., Kutzkov, K., Scaman, K., and Vojnovic, M. (2018).
KONG: Kernels for ordered-neighborhood graphs.
Technical report.

(https://arxiv.org/abs/1805.10014).

Dziugaite, G. K., Roy, D. M., and Ghahramani, Z. (2015).
Training generative neural networks via maximum mean discrepancy optimization.
In *Conference on Uncertainty in Artificial Intelligence (UAI)*, pages 258–267.

Fang, Z., Guo, Z.-C., and Zhou, D.-X. (2019).
Optimal learning rates for distribution regression.
Journal of Complexity.
(101426,https://doi.org/10.1016/j.jco.2019.101426).

Fukumizu, K., Song, L., and Gretton, A. (2013).
Kernel Bayes' rule: Bayesian inference with positive definite kernels.
Journal of Machine Learning Research, 14:3753–3783.

Gärtner, T., Flach, P., Kowalczyk, A., and Smola, A. (2002).
Multi-instance kernels.

In *International Conference on Machine Learning (ICML)*, pages 179–186.

Gärtner, T., Flach, P., and Wrobel, S. (2003).
On graph kernels: Hardness results and efficient alternatives.
Learning Theory and Kernel Machines, pages 129–143.

Gretton, A., Borgwardt, K., Rasch, M., Schölkopf, B., and Smola, A. (2012).
A kernel two-sample test.
Journal of Machine Learning Research, 13(25):723–773.

Gretton, A., Fukumizu, K., Teo, C. H., Song, L., Schölkopf, B., and Smola, A. (2008).
A kernel statistical test of independence.
In *Advances in Neural Information Processing Systems (NIPS)*, pages 585–592.

Guevara, J., Hirata, R., and Canu, S. (2017).
Cross product kernels for fuzzy set similarity.

In *International Conference on Fuzzy Systems (FUZZ-IEEE)*, pages 1–6.

Harchaoui, Z., Bach, F., and Moulines, E. (2007).
Testing for homogeneity with kernel Fisher discriminant analysis.
In *Advances in Neural Information Processing Systems (NIPS)*, pages 609–616.

Harchaoui, Z. and Cappé, O. (2007).
Retrospective multiple change-point estimation with kernels.
In *IEEE/SP Workshop on Statistical Signal Processing*, pages 768–772.

Haussler, D. (1999).
Convolution kernels on discrete structures.
Technical report, University of California at Santa Cruz.
(http://cbse.soe.ucsc.edu/sites/default/files/convolutions.pdf).

Jaakkola, T. S. and Haussler, D. (1999).

Exploiting generative models in discriminative classifiers.
In *Advances in Neural Information Processing Systems (NIPS)*, pages 487–493.

Jebara, T., Kondor, R., and Howard, A. (2004).
Probability product kernels.
Journal of Machine Learning Research, 5:819–844.

Jiao, Y. and Vert, J.-P. (2016).
The Kendall and Mallows kernels for permutations.
In *International Conference on Machine Learning (ICML)*, volume 37, pages 2982–2990.

Jitkrittum, W., Szabó, Z., Chwialkowski, K., and Gretton, A. (2016).
Interpretable distribution features with maximum testing power.
In Lee, D. D., Sugiyama, M., Luxburg, U. V., Guyon, I., and Garnett, R., editors, *Advances in Neural Information*

Processing Systems (NIPS), pages 181–189, Barcelona, Spain. Curran Associates, Inc.
(full oral presentation = top 1.84%).

Jitkrittum, W., Szabó, Z., and Gretton, A. (2017a).
An adaptive test of independence with analytic kernel embeddings.
In Precup, D. and Teh, Y. W., editors, *International Conference on Machine Learning (ICML)*, volume 70, pages 1742–1751, Sydney, Australia. PMLR.
(25.46% acceptance rate).

Jitkrittum, W., Xu, W., Szabó, Z., Fukumizu, K., and Gretton, A. (2017b).
A linear-time kernel goodness-of-fit test.
In Guyon, I., Luxburg, U. V., Bengio, S., Wallach, H., Fergus, R., Vishwanathan, S., and Garnett, R., editors, *Advances in Neural Information Processing Systems (NIPS)*, pages 261–270, Long Beach, CA, U.S. Curran Associates, Inc.
(Best Paper Award = in top 3 out of 3240 submissions).

Kashima, H. and Koyanagi, T. (2002).
Kernels for semi-structured data.
In *International Conference on Machine Learning (ICML)*, pages 291–298.

Kashima, H., Tsuda, K., and Inokuchi, A. (2003).
Marginalized kernels between labeled graphs.
In *International Conference on Machine Learning (ICML)*, pages 321–328.

Kim, B., Khanna, R., and Koyejo, O. (2016).
Examples are not enough, learn to criticize! criticism for interpretability.
In *Advances in Neural Information Processing Systems (NIPS)*, pages 2280–2288.

Király, F. J. and Oberhauser, H. (2019).
Kernels for sequentially ordered data.
Journal of Machine Learning Research, 20:1–45.

Klus, S., Bittracher, A., Schuster, I., and Schütte, C. (2019).
A kernel-based approach to molecular conformation analysis.
The Journal of Chemical Physics, 149:244109.

Klus, S., Schuster, I., and Muandet, K. (2018).
Eigendecompositions of transfer operators in reproducing kernel Hilbert spaces.
Technical report.
(https://arxiv.org/abs/1712.01572).

Kondor, R. and Pan, H. (2016).
The multiscale Laplacian graph kernel.
In *Advances in Neural Information Processing Systems (NIPS)*, pages 2982–2990.

Kondor, R. I. and Lafferty, J. (2002).
Diffusion kernels on graphs and other discrete input.
In *International Conference on Machine Learning (ICML)*, pages 315–322.

Kuang, R., Ie, E., Wang, K., Wang, K., Siddiqi, M., Freund, Y., and Leslie, C. (2004).
Profile-based string kernels for remote homology detection and motif extraction.
Journal of Bioinformatics and Computational Biology, 13(4):527–550.

Kusano, G., Fukumizu, K., and Hiraoka, Y. (2016).
Persistence weighted Gaussian kernel for topological data analysis.
In *International Conference on Machine Learning (ICML)*, pages 2004–2013.

Law, H. C. L., Sutherland, D., Sejdinovic, D., and Flaxman, S. (2018).
Bayesian approaches to distribution regression.
International Conference on Artificial Intelligence and Statistics (AISTATS), 84:1167–1176.

Leslie, C., Eskin, E., and Noble, W. S. (2002).

The spectrum kernel: A string kernel for SVM protein classification.
Biocomputing, pages 564–575.

Leslie, C. and Kuang, R. (2004).
Fast string kernels using inexact matching for protein sequences.
Journal of Machine Learning Research, 5:1435–1455.

Li, Y., Swersky, K., and Zemel, R. (2015).
Generative moment matching networks.
In *International Conference on Machine Learning (ICML)*, pages 1718–1727.

Lin, J., Rudi, A., Rosasco, L., and Cevher, V. (2020).
Optimal rates for spectral algorithms with least-squares regression over Hilbert spaces.
Applied and Computational Harmonic Analysis, 48(3):868–890.

Lloyd, J. R., Duvenaud, D., Grosse, R., Tenenbaum, J., and Ghahramani, Z. (2014).

Automatic construction and natural-language description of nonparametric regression models.
In *AAAI Conference on Artificial Intelligence*, pages 1242–1250.

Lodhi, H., Saunders, C., Shawe-Taylor, J., Cristianini, N., and Watkins, C. (2002).
Text classification using string kernels.
Journal of Machine Learning Research, 2:419–444.

Lopez-Paz, D., Muandet, K., Schölkopf, B., and Tolstikhin, I. (2015).
Towards a learning theory of cause-effect inference.
International Conference on Machine Learning (ICML), 37:1452–1461.

Mooij, J., Peters, J., Janzing, D., Zscheischler, J., and Schölkopf, B. (2016).
Distinguishing cause from effect using observational data: Methods and benchmarks.
Journal of Machine Learning Research, 17:1–102.

Muandet, K., Fukumizu, K., Dinuzzo, F., and Schölkopf, B. (2011).
Learning from distributions via support measure machines.
In *Advances in Neural Information Processing Systems (NIPS)*, pages 10–18.

Mücke, N. (2021).
Stochastic gradient descent meets distribution regression.
In *International Conference on Artificial Intelligence and Statistics (AISTATS)*, pages 2143–2151.

Neubauer, A., Engl, H. W., and Hanke, M. (1996).
Regularization of Inverse Problems.
Springer.

Park, M., Jitkrittum, W., and Sejdinovic, D. (2016).
K2-ABC: Approximate Bayesian computation with kernel embeddings.
In *International Conference on Artificial Intelligence and Statistics (AISTATS)*, volume 51, pages 398–407.

Pfister, N., Bühlmann, P., Schölkopf, B., and Peters, J. (2018).
Kernel-based tests for joint independence.
Journal of the Royal Statistical Society: Series B (Statistical Methodology), 80(1):5–31.

Rahimi, A. and Recht, B. (2007).
Random features for large-scale kernel machines.
In *Advances in Neural Information Processing Systems (NIPS)*, pages 1177–1184.

Raj, A., Law, H. C. L., Sejdinovic, D., and Park, M. (2019).
A differentially private kernel two-sample test.
In *European Conference on Machine Learning and Principles and Practice of Knowledge Discovery in Databases (ECML-PKDD)*.

Rüping, S. (2001).
SVM kernels for time series analysis.
Technical report, University of Dortmund.

(http://www.stefan-rueping.de/publications/rueping-2001-a.pdf).

Saigo, H., Vert, J.-P., Ueda, N., and Akutsu, T. (2004).
Protein homology detection using string alignment kernels.
Bioinformatics, 20(11):1682–1689.

Schölkopf, B., Muandet, K., Fukumizu, K., Harmeling, S., and Peters, J. (2015).
Computing functions of random variables via reproducing kernel Hilbert space representations.
Statistics and Computing, 25(4):755–766.

Seeger, M. (2002).
Covariance kernels from Bayesian generative models.
In *Advances in Neural Information Processing Systems (NIPS)*, pages 905–912.

Shervashidze, N., Vishwanathan, S. V. N., Petri, T., Mehlhorn, K., and Borgwardt, K. M. (2009).
Efficient graphlet kernels for large graph comparison.

In *International Conference on Artificial Intelligence and Statistics (AISTATS)*, pages 488–495.

Song, L., Gretton, A., Bickson, D., Low, Y., and Guestrin, C. (2011).
Kernel belief propagation.
In *International Conference on Artificial Intelligence and Statistics (AISTATS)*, pages 707–715.

Sriperumbudur, B. K. and Szabó, Z. (2015).
Optimal rates for random Fourier features.
In Cortes, C., Lawrence, N. D., Lee, D. D., Sugiyama, M., and Garnett, R., editors, *Advances in Neural Information Processing Systems (NIPS)*, pages 1144–1152, Montréal, Canada. Curran Associates, Inc.
(contributed equally; spotlight presentation – 3.65% acceptance rate).

Szabó, Z., Sriperumbudur, B. K., Póczos, B., and Gretton, A. (2016).

Learning theory for distribution regression.
Journal of Machine Learning Research, 17(152):1–40.

Székely, G. and Rizzo, M. (2004).
Testing for equal distributions in high dimension.
InterStat, 5:1249–1272.

Székely, G. and Rizzo, M. (2005).
A new test for multivariate normality.
Journal of Multivariate Analysis, 93:58–80.

Tsuda, K., Kin, T., and Asai, K. (2002).
Marginalized kernels for biological sequences.
Bioinformatics, 18:268–275.

Vishwanathan, S. N., Schraudolph, N., Kondor, R., and Borgwardt, K. (2010).
Graph kernels.
Journal of Machine Learning Research, 11:1201–1242.

Watkins, C. (1999).
Dynamic alignment kernels.

In *Advances in Neural Information Processing Systems (NIPS)*, pages 39–50.

Yamada, M., Umezu, Y., Fukumizu, K., and Takeuchi, I. (2018).
Post selection inference with kernels.
In *International Conference on Artificial Intelligence and Statistics (AISTATS)*, volume 84, pages 152–160.

Zaheer, M., Kottur, S., Ravanbakhsh, S., Póczos, B., Salakhutdinov, R., and Smola, A. (2017).
Deep sets.
In *Advances in Neural Information Processing Systems (NIPS)*, pages 3394–3404.

Zhang, K., Schölkopf, B., Muandet, K., and Wang, Z. (2013).
Domain adaptation under target and conditional shift.
Journal of Machine Learning Research, 28(3):819–827.