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 $\Rightarrow \|\cdot\|$, $\sphericalangle$.

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

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

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

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

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

where $\varphi(x)$ is a feature of $x$. 

Notes:

Examples:

- $k_P(x, y) = (\langle x, y \rangle + \gamma)_p$
- $k_G(x, y) = e^{-\gamma \|x - y\|^2}$

One can choose $\varphi(x) = k(\cdot, x)$.

RKHS: $\mathcal{H}_k = \{ \sum_{i=1}^{d} \alpha_i k(\cdot, x_i) \}$: polynomials, splines, Fourier analysis, ...
From similarity on $\mathbb{R}^d$

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

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

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

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

Notes

- Examples: $k_P(x, y) = (\langle x, y \rangle + \gamma)^p$, $k_G(x, y) = e^{-\gamma \|x-y\|^2_2}$.

- One can choose $\varphi(x) = k(\cdot, x)$.

- RKHS: $\mathcal{H}_k = \left\{ \sum_{i=1}^{n} \alpha_i k(\cdot, x_i) \right\}^{\text{spec}} \rightarrow$ polynomials, splines, Fourier analysis, ...
Kernels exist on various objects

Few examples:

- strings
- 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
Similarity of bags (or probability distributions)

- Characteristic function:

\[
P \mapsto c(P) = \int_{\mathbb{R}^d} e^{i \langle \cdot, x \rangle} dP(x).
\]

[other examples: \( \mathbb{I}_{(-\infty, \cdot)}(x), e^{\langle \cdot, x \rangle} \)]
Similarity of bags (or probability distributions)

- Characteristic function:

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

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

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

\[
\mathbb{P} \mapsto \mu_k(\mathbb{P}) = \int_{\mathcal{X}} \varphi(x) d\mathbb{P}(x).
\]
Similarity of bags (or probability distributions)

- Characteristic function:
  \[ P \mapsto c(P) = \int_{\mathbb{R}^d} e^{i \langle \cdot, x \rangle} dP(x). \]

- Other examples: \( \mathbb{I}_{(-\infty, \cdot)}(x), e^{\langle \cdot, x \rangle} \)

- Mean embedding with \( k: \mathcal{X} \times \mathcal{X} \to \mathbb{R} \):
  \[ P \mapsto \mu_k(P) = \int_{\mathcal{X}} \varphi(x) dP(x). \]

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

\[ \langle \mu_k(P_N), \mu_k(Q_M) \rangle_{\mathcal{H}_k} = \frac{1}{NM} \sum_{n=1}^{N} \sum_{m=1}^{M} k(x_n, x'_m). \]
Applications:

- **two-sample testing**

- **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 $\equiv$ regression on labelled bags

- **Given:**
  - labelled bags: $\hat{z} = \{(\hat{P}_i, y_i)\}_{i=1}^{\ell}$, $\hat{P}_i$: bag from $P_i$, $N := |\hat{P}_i|$.
  - test bag: $\hat{P}$.
Aerosol prediction = regression on labelled bags

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

- **Estimator:**

\[
f_{\hat{z}}^\lambda = \arg \min_{f \in \mathcal{H}_K} \frac{1}{\ell} \sum_{i=1}^\ell \left[ f\left( \mu_k(\hat{P}_i) \right) - y_i \right]^2 + \lambda \|f\|_{\mathcal{H}_K}^2.
\]

feature of the i-th bag

Zoltán Szabó  Distribution Regression and Beyond
Aerosol prediction = regression on labelled bags

- Given:
  - labelled bags: \( \hat{z} = \{(\hat{P}_i, y_i)\}_{i=1}^\ell \), \( \hat{P}_i \): bag from \( P_i \), \( N := |\hat{P}_i| \).
  - test bag: \( \hat{P} \).
- Estimator:
  \[
  f_\lambda^\hat{z} = \arg \min_{f \in \mathcal{H}_K} \frac{1}{\ell} \sum_{i=1}^\ell \left[ f\left( \mu_k(\hat{P}_i) \right) - y_i \right]^2 + \lambda \| f \|_{\mathcal{H}_K}^2.
  \]
  feature of the i-th bag
- Prediction:
  \[
  \hat{y}(\hat{P}) = g^T (G + \ell \lambda I_\ell)^{-1} y,
  \]
  \[
  g = \left[ K\left( \mu_k(\hat{P}), \mu_k(\hat{P}_i) \right) \right],
  G = \left[ K\left( \mu_k(\hat{P}_i), \mu_k(\hat{P}_j) \right) \right],
  y = [y_i].
  \]
Aerosol prediction = regression on labelled bags

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

- Estimator:
  \[
  f^{\lambda}_{\hat{z}} = \arg \min_{f \in \mathcal{H}_K} \frac{1}{\ell} \sum_{i=1}^\ell \left[ f \left( \mu_k(\hat{P}_i) \right) - y_i \right]^2 + \lambda \| f \|_{\mathcal{H}_K}^2.
  \]
  feature of the i-th bag

- Prediction:
  \[
  \hat{y}(\hat{P}) = g^T (G + \ell \lambda I_\ell)^{-1} y,
  \]
  \[
  g = \left[ K \left( \mu_k(\hat{P}), \mu_k(\hat{P}_i) \right) \right], G = \left[ K \left( \mu_k(\hat{P}_i), \mu_k(\hat{P}_j) \right) \right], y = [y_i].
  \]

Challenge

Consistent? How many samples per bag?
Quality of estimator, baseline:

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

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

How many samples/bag to achieve the accuracy of \( f_\rho \)? Possible?
Quality of estimator, baseline:

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

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

How many samples/bag to achieve the accuracy of \( f_\rho \)? Possible?

Blanket assumptions:

1. \( X \): separable topological; \( k \): bounded & continuous.
2. \( y \): bounded; \( Y \): separable Hilbert.
3. \( K \): bounded, Hölder continuous.
-known result: Hölder exponent = 1 below

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

\[ \mathcal{R}(f^\lambda) - \mathcal{R}(f_\rho) = \mathcal{O}_p \left( \ell^{-\frac{bc}{bc+1}} \right), \]

\(b\) – size of the input space, \(c\) – smoothness of \(f_\rho\).
Result [Szabó et al., 2016]: Hölder exponent \( = 1 \) below

- Known [Caponnetto and De Vito, 2007]: optimal rate
  \[
  \mathcal{R}(f^\lambda_z) - \mathcal{R}(f_\rho) = \mathcal{O}_p \left( \ell^{-\frac{bc}{bc+1}} \right),
  \]
  \( b \) – size of the input space, \( c \) – smoothness of \( f_\rho \).
- Let \( N = \tilde{\mathcal{O}}(\ell^a) \). \( N \): size of the bags. \( \ell \): number of bags.

Our result

- If \( 2 \leq a \), then \( f^\lambda_2 \) attains the optimal rate.
Result [Szabó et al., 2016]: Hölder exponent = 1 below

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

\[
\mathcal{R}(f^\lambda_z) - \mathcal{R}(f^\rho) = \mathcal{O}_p \left( \ell^{-\frac{bc}{bc+1}} \right),
\]

- \(b\) – size of the input space, \(c\) – smoothness of \(f^\rho\).
- Let \(N = \tilde{O}(\ell^a)\). \(N\): size of the bags. \(\ell\): number of bags.

Our result

- If \(2 \leq a\), then \(f^\lambda_z\) 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

\[ \mathcal{R}(f^\lambda_z) - \mathcal{R}(f_\rho) = \mathcal{O}_p \left( \ell^{-\frac{bc}{bc+1}} \right), \]

\[ b \] – size of the input space, \[ c \] – smoothness of \[ f_\rho \].

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

Our result

- If \[ 2 \leq a \], then \[ f^\lambda_z \] 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

\[
\mathcal{R}(f^\lambda_z) - \mathcal{R}(f^\lambda_\rho) = \mathcal{O}_p \left( \ell^{-\frac{bc}{bc+1}} \right),
\]

- \(b\) – size of the input space, \(c\) – smoothness of \(f^\rho\).
- Let \(N = \tilde{O}(\ell^a)\). \(N\): size of the bags. \(\ell\): number of bags.

Our result

- If \(2 \leq a\), then \(f^\lambda_z\) 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): \(\checkmark\)
Result [Szabó et al., 2016]: Hölder exponent $= 1$ below

- Known [Caponnetto and De Vito, 2007]: optimal rate
  \[ R(f_\lambda^z) - R(f_\rho) = O_p \left( \ell^{-\frac{bc}{bc+1}} \right), \]
  
  $b$ – size of the input space, $c$ – smoothness of $f_\rho$.

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

**Our result**

- If $2 \leq a$, then $f_2^\lambda$ 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: $y = [y_i]_{i \in [n]}$

  $f_{\lambda}^z(x) = \sum_{i \in [n]} \alpha_i K(x, x_i), \alpha = (K + n\lambda I_n)^{-1} y.$
Various research questions

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

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

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

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

- **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: $y = [y_i]_{i \in [n]}$
    
    $$f_z^\lambda(x) = \sum_{i \in [n]} \alpha_i K(x, x_i), \quad \alpha = (K + n\lambda I_n)^{-1}y.$$  

    This falls with $g_\lambda(\sigma) = \frac{1}{\sigma + \lambda}$ under the umbrella
    
    $$f_z^\lambda(x) = \sum_{i \in [n]} \alpha_i K(x, x_i), \quad \alpha = \frac{1}{n} g_\lambda \left( \frac{K}{n} \right) y.$$  

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

Zoltán Szabó  
Distribution Regression and Beyond
Thank you for the attention!


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


Fast global alignment kernels.


Exploiting generative models in discriminative classifiers.


Interpretable distribution features with maximum testing power.


*International Conference on Artificial Intelligence and Statistics (AISTATS), 84*:1167–1176.

The spectrum kernel: A string kernel for SVM protein classification.

*Biocomputing*, pages 564–575.


Automatic construction and natural-language description of nonparametric regression models.

Text classification using string kernels.

Towards a learning theory of cause-effect inference.

Distinguishing cause from effect using observational data: Methods and benchmarks.


Szabó, Z., Sriperumbudur, B. K., Póczos, B., and Gretton, A. (2016).
Learning theory for distribution regression.  

Testing for equal distributions in high dimension.  

A new test for multivariate normality.  

Marginalized kernels for biological sequences.  

Graph kernels.  

Dynamic alignment kernels.
