Distribution Regression with Minimax-Optimal Guarantee

Zoltán Szabó (Gatsby Unit, UCL)

Joint work with

○ Bharath K. Sriperumbudur (Department of Statistics, PSU),
○ Barnabás Póczos (ML Department, CMU),
○ Arthur Gretton (Gatsby Unit, UCL)

MASCOT-NUM, Toulouse
March 25, 2016
Example: sustainability

- **Goal**: aerosol prediction $\Rightarrow$ air pollution $\Rightarrow$ climate.

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

Multi-instance learning:
- [Haussler, 1999, Gärtner et al., 2002] (set kernel):
  - sensible methods in regression: few,
    1. restrictive technical conditions,
    2. super-high resolution satellite image: would be needed.
Contributions:

1. **Practical:** state-of-the-art accuracy (aerosol).
2. **Theoretical:**
   - General bags: graphs, time series, texts, …
   - Consistency of set kernel in regression (17-year-old open problem).
   - How many samples/bag?
Contributions:

1. Practical: state-of-the-art accuracy (aerosol).
2. Theoretical:
   - General bags: graphs, time series, texts, …
   - Consistency of set kernel in regression (17-year-old open problem).
   - How many samples/bag?
   - AISTATS-2015 (oral – 6.11%) → JMLR in revision.
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, ...
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.
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: \( \hat{P}. \)
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:**

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

\( \mu_{\hat{P}_i} \): feature of \( \hat{P}_i \)
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:**
  $$f_{\hat{z}}^\lambda = \arg \min_{f \in \mathcal{H}(K)} \frac{1}{\ell} \sum_{i=1}^\ell \left[ f(\mu_{\hat{P}_i}) - y_i \right]^2 + \lambda \|f\|_{\mathcal{H}}^2.$$

- **Prediction:**
  $$\hat{y}(\hat{P}) = g^T (G + \ell\lambda I)^{-1} y,$$
  $$g = [K(\mu_{\hat{P}}, \mu_{\hat{P}_i})], \ G = [K(\mu_{\hat{P}_i}, \mu_{\hat{P}_j})], \ y = [y_i].$$
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:
  \[
  f^\lambda_{\hat{z}} = \arg \min_{f \in H(K)} \frac{1}{\ell} \sum_{i=1}^\ell \left[f(\mu_{\hat{P}_i}) - y_i\right]^2 + \lambda \|f\|^2_{H(K)}.
  \]
- Prediction:
  \[
  \hat{y}(\hat{P}) = g^T (G + \ell \lambda I)^{-1} y,
  \]
  \[
  g = [K(\mu_{\hat{P}}, \mu_{\hat{P}_i})],
  G = [K(\mu_{\hat{P}_i}, \mu_{\hat{P}_j})],
  y = [y_i].
  \]

Challenges

1. Inner product of distributions: \( K(\mu_{\hat{P}_i}, \mu_{\hat{P}_j}) = ? \)
2. How many samples/bag?
Let us define an inner product on distributions $\tilde{K}(P, Q)$:

Set kernel: $A = \{a_i\}_{i=1}^N$, $B = \{b_j\}_{j=1}^N$.

$$\tilde{K}(A, B) = \frac{1}{N^2} \sum_{i,j=1}^{N} k(a_i, b_j) = \left\langle \frac{1}{N} \sum_{i=1}^{N} \varphi(a_i), \frac{1}{N} \sum_{j=1}^{N} \varphi(b_j) \right\rangle.$$

Remember:
Let us define an inner product on distributions $\tilde{K}(P, Q)$:

1. Set kernel: $A = \{a_i\}_{i=1}^N, B = \{b_j\}_{j=1}^N$.

   $$\tilde{K}(A, B) = \frac{1}{N^2} \sum_{i,j=1}^{N} k(a_i, b_j) = \left\langle \frac{1}{N} \sum_{i=1}^{N} \varphi(a_i), \frac{1}{N} \sum_{j=1}^{N} \varphi(b_j) \right\rangle.$$  
   
   feature of bag $A$

2. Taking 'limit' [Berlinet and Thomas-Agnan, 2004, Altun and Smola, 2006, Smola et al., 2007]: $a \sim P, b \sim Q$

   $$\tilde{K}(P, Q) = \mathbb{E}_{a,b} k(a, b) = \left\langle \mathbb{E}_a \varphi(a), \mathbb{E}_b \varphi(b) \right\rangle.$$  
   
   feature of distribution $P=\mu_P$

Example (Gaussian kernel): $k(a, b) = e^{-\|a-b\|_2^2/(2\sigma^2)}$. 
Given: $\mathcal{D}$ set.

- Kernel: $k(a, b) = \langle \varphi(a), \varphi(b) \rangle_{\mathcal{F}}$, $\mathcal{F}$: Hilbert space.
Given: $\mathcal{D}$ set.

- Kernel: $k(a, b) = \langle \varphi(a), \varphi(b) \rangle_{\mathcal{F}}$, $\mathcal{F}$: Hilbert space.
- RKHS: $H \subset \mathbb{R}^D$ Hilbert space, $\delta_b(f) = f(b)$ is continuous ($\forall b$).
Given: $\mathcal{D}$ set.

- Kernel: $k(a, b) = \langle \varphi(a), \varphi(b) \rangle_{\mathcal{F}}$, $\mathcal{F}$: Hilbert space.
- RKHS: $H \subset \mathbb{R}^D$ Hilbert space, $\delta_b(f) = f(b)$ is continuous ($\forall b$).
- Reproducing kernel of an $H \subset \mathbb{R}^D$ Hilbert space,
  - $k(\cdot, b) \in H$, 

**RKHS definition(s)**

Zoltán Szabó  Distribution Regression with Minimax-Optimal Guarantee
Given: $\mathcal{D}$ set.

- Kernel: $k(a, b) = \langle \varphi(a), \varphi(b) \rangle_{\mathcal{F}}$, $\mathcal{F}$: Hilbert space.
- RKHS: $H \subset \mathbb{R}^D$ Hilbert space, $\delta_b(f) = f(b)$ is continuous ($\forall b$).
- Reproducing kernel of an $H \subset \mathbb{R}^D$ Hilbert space,
  1. $k(\cdot, b) \in H$,
  2. $\langle f, k(\cdot, b) \rangle_H = f(b)$. Note: $k(a, b) = \langle k(\cdot, a), k(\cdot, b) \rangle_H$. 
RKHS definition(s)

Given: $\mathcal{D}$ set.

- Kernel: $k(a, b) = \langle \varphi(a), \varphi(b) \rangle_{\mathcal{F}}$, $\mathcal{F}$: Hilbert space.
- RKHS: $H \subset \mathbb{R}^\mathcal{D}$ Hilbert space, $\delta_b(f) = f(b)$ is continuous $(\forall b)$.
- Reproducing kernel of an $H \subset \mathbb{R}^\mathcal{D}$ Hilbert space,
  1. $k(\cdot, b) \in H$,
  2. $\langle f, k(\cdot, b) \rangle_H = f(b)$. Note: $k(a, b) = \langle k(\cdot, a), k(\cdot, b) \rangle_H$.
- $k : \mathcal{D} \times \mathcal{D} \to \mathbb{R}$ sym. is pd. if $G = [k(x_i, x_j)]_{i,j=1}^n \succeq 0$. 

Zoltán Szabó  Distribution Regression with Minimax-Optimal Guarantee
Other valid similarities

Recall: \( K(P, Q) = \langle \mu_P, \mu_Q \rangle \).

<table>
<thead>
<tr>
<th>( K_G )</th>
<th>( K_e )</th>
<th>( K_C )</th>
</tr>
</thead>
<tbody>
<tr>
<td>( e^{-\frac{|\mu_P - \mu_Q|^2}{2\theta^2}} )</td>
<td>( e^{-\frac{|\mu_P - \mu_Q|^2}{2\theta^2}} )</td>
<td>( \left(1 + \frac{|\mu_P - \mu_Q|^2}{\theta^2}\right)^{-1} )</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>( K_t )</th>
<th>( K_i )</th>
</tr>
</thead>
<tbody>
<tr>
<td>( \left(1 + |\mu_P - \mu_Q|^\theta\right)^{-1} )</td>
<td>( \left(|\mu_P - \mu_Q|^2 + \theta^2\right)^{-\frac{1}{2}} )</td>
</tr>
</tbody>
</table>

Functions of \( \|\mu_P - \mu_Q\| \Rightarrow \) computation: similar to set kernel.
Regression on labelled bags: baseline

Quality of estimator, baseline:

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

\[ f_\rho = \text{best regressor}. \]

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

Assume (for a moment): \( f_\rho \in \mathcal{H}(K). \)
Known [Caponnetto and De Vito, 2007]: best/achieved rate

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

\( b \) – size of the input space, \( c \) – smoothness of \( f_\rho \).
Our result: how many samples/bag

- Known [Caponnetto and De Vito, 2007]: best/achieved rate

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

- \( b \) – 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 best achievable rate.
Our result: how many samples/bag

- Known [Caponnetto and De Vito, 2007]: best/achieved rate

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

- \(b\) – 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 \(f_\lambda^\ast\) attains the best achievable rate.
- In fact, \(a = \frac{b(c+1)}{bc+1} < 2\) is enough.
- Consequence: regression with set kernel is consistent.
Why can we get consistency/rates? – intuition

- Convergence of the mean embedding:
  \[ \| \mu_P - \mu_{\hat{P}} \|_H = O \left( \frac{1}{\sqrt{N}} \right). \]

- Hölder property of \( K \) (\( 0 < L, 0 < h \leq 1 \)):
  \[ \| K(\cdot, \mu_P) - K(\cdot, \mu_{\hat{P}}) \|_{\mathcal{H}} \leq L \| \mu_P - \mu_{\hat{P}} \|_H^h. \]

- \( f_{\hat{z}}^\lambda \) depends 'nicely' on \( \mu_{\hat{P}} \). [39 pages]
Misspecified setting ($f_\rho \in L^2 \setminus \mathcal{H}$):

- Consistency: convergence to $\inf_{f \in \mathcal{H}} \| f - f_\rho \|_{L^2}$.
- Smoothness on $f_\rho$: computational & statistical tradeoff.
Vector-valued output:

- $Y$: separable Hilbert space $\Rightarrow K(\mu_P, \mu_Q) \in \mathcal{L}(Y)$.
- Prediction on a test bag $\hat{P}$:

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

$$g = [K(\mu_{\hat{P}}, \mu_{\hat{P}_i})], G = [K(\mu_{\hat{P}_i}, \mu_{\hat{P}_j})], y = [y_i].$$

Specifically: $Y = \mathbb{R} \Rightarrow \mathcal{L}(Y) = \mathbb{R}$; $Y = \mathbb{R}^d \Rightarrow \mathcal{L}(Y) = \mathbb{R}^{d \times d}$. 
We perform on par with the state-of-the-art, hand-engineered method.

Zhuang Wang, Liang Lan, Slobodan Vucetic. IEEE Transactions on Geoscience and Remote Sensing, 2012: 7.5 – 8.5 (±0.1 – 0.6):
  • hand-crafted features.

Our prediction accuracy: 7.81 (±1.64).
  • no expert knowledge.

Code in ITE: #2 on mloss,

https://bitbucket.org/szzoli/ite/
Summary

- Problem: distribution regression.
- Contribution:
  - computational & statistical tradeoff analysis,
  - specifically, the set kernel is consistent: 17-year-old open question,
  - minimax optimal rate is achievable: sub-quadratic bag size.
- Details (JMLR in revision):

  http://arxiv.org/abs/1411.2066
Thank you for the attention!

**Acknowledgments:** This work was supported by the Gatsby Charitable Foundation, and by NSF grants IIS1247658 and IIS1250350. A part of the work was carried out while Bharath K. Sriperumbudur was a research fellow in the Statistical Laboratory, Department of Pure Mathematics and Mathematical Statistics at the University of Cambridge, UK.


Convolution kernels on discrete structures.
Technical report, Department of Computer Science, University of California at Santa Cruz.
