Performance guarantees for kernel-based learning on probability distributions

Zoltán Szabó (Gatsby Unit, UCL)

Max Planck Institute for Intelligent Systems
Tübingen

March 16, 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.
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.
Given:

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

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

Estimator:

\[
w_\lambda \hat{z} = \arg \min_w \frac{1}{\ell} \sum_{i=1}^\ell \left[ \langle w, \psi(\hat{P}_i) \rangle - y_i \right]^2 + \lambda \|w\|^2.
\]
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:**
  \[
  w^\lambda_{\hat{z}} = \arg\min_{w} \frac{1}{\ell} \sum_{i=1}^{\ell} \left[ \langle w, \psi(\hat{P}_i) \rangle - y_i \right]^2 + \lambda \| w \|^2 .
  \]

- **Prediction:**
  \[
  \hat{y}(\hat{P}) = g^T (K + \ell \lambda I)^{-1} y ,
  \]
  \[
  g = [K(\hat{P}_i, \hat{P})] , \quad K = [K(\hat{P}_i, \hat{P}_j)] , \quad y = [y_i] .
  \]
  \[
  := \langle \psi(\hat{P}_i), \psi(\hat{P}_j) \rangle
  \]
Let us define an inner product on distributions \( K(P, Q) \):

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

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

Remember:
Let us define an inner product on distributions \([K(P, Q)]\):

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

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

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

   \[
   K(P, Q) = \mathbb{E}_{a,b} k(a, b) = \left\langle \mathbb{E}_a \varphi(a), \mathbb{E}_b \varphi(b) \right\rangle.
   \]

Example (Gaussian kernel): \(k(a, b) = e^{-\|a-b\|_2^2/(2\sigma^2)}\).
Regression on labelled bags: baseline

Quality of estimator, baseline:

\[ \mathcal{R}(w) = \mathbb{E}_{(\psi(Q), y) \sim \rho} \left[ \langle w, \psi(Q) \rangle - y \right]^2, \]

\( w_\rho = \text{best regressor}. \)

How many samples/bag to get the accuracy of \( w_\rho \)? Possible?
Our result: how many samples/bag

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

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

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

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

\[ R(w_2^\lambda) - R(w_\rho) = O\left(\ell^{-\frac{bc}{bc+1}}\right), \]

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

Our result

- If \( 2 \leq a \), then \( w_2^\lambda \) attains the best achievable rate.
Our result: how many samples/bag

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

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

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

**Our result**

- If \( 2 \leq a \), then \( w^\lambda_z \) attains the best achievable rate.
- In fact, \( a = \frac{b(c+1)}{bc+1} < 2 \) is enough.
- Consequence: regression with set kernel is consistent.
- The same result holds for Hölder K-s: Gaussian
  [Christmann and Steinwart, 2010], . . .
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/
Related results
Kernel EP [UAI-2015]:
- distribution regression phrasing,
- learn the message-passing operator for 'tricky' factors.
Kernel EP [UAI-2015]:
- distribution regression phrasing,
- learn the message-passing operator for 'tricky' factors.
- extends Infer.NET; speed $\Leftarrow$ RFF.
Distribution regression with random Fourier features

- **Kernel EP [UAI-2015]:**
  - distribution regression phrasing,
  - learn the message-passing operator for 'tricky' factors.
  - extends Infer.NET; speed $\leftrightarrow$ RFF.

- **Random Fourier features [NIPS-2015 (spotlight - 3.65%)]:**
  - exponentially tighter guarantee.
Applications, with Gatsby students

- Bayesian manifold learning [NIPS-2015]:
  - App.: climate data $\rightarrow$ weather station location.

- Fast, adaptive sampling method based on RFF [NIPS-2015]:
  - App.: approximate Bayesian computation, hyperparameter inference.
Applications, with Gatsby students

- Bayesian manifold learning [NIPS-2015]:
  - App.: climate data $\rightarrow$ weather station location.

- Fast, adaptive sampling method based on RFF [NIPS-2015]:
  - App.: approximate Bayesian computation, hyperparameter inference.

- Interpretable 2-sample testing [ICML-2016 submission]:
  - App.:
    - random $\rightarrow$ smart features,
    - discriminative for doc. categories, emotions.
  - empirical process theory (VC subgraphs).
Regression on

- bags/distributions:
  - minimax optimality,
  - set kernel is consistent.

- random Fourier features: exponentially tighter bounds.

Several applications (with open source code).

**Acknowledgments:** This work was supported by the Gatsby Charitable Foundation.


Multi-instance kernels.

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

A Hilbert space embedding for distributions.