Optimal Distribution Regression

Zoltán Szabó

Joint work with

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

Gatsby Unit, Research Talk
May 23, 2016

Outline

- Context.
- Problem formulation.
- Consistency guarantees.

Context

Motivation

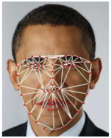
Regression on labelled bags:

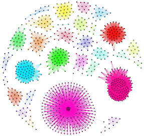
- ML: multi-instance learning [Haussler, 1999, Gärtner et al., 2002].
- Statistics: point estimation tasks.

Motivation

Regression on labelled bags:

- Bag examples: aerosol prediction,





  - time-series modelling: user = set of time-series,
  - network analysis: group of people = bag of friendship graphs,
  - NLP: corpus = bag of documents.

One-page summary

- Question: How many samples/bag?
- Contributions:
  1. General bags: vectors, graphs, time series, texts, ...
  2. Computational-statistical tradeoff analysis.
  3. Minimax optimality.
  4. Well-specified & misspecified case.

Problem formulation

Regression on labelled bags

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

Regression on labelled bags

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

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

Regression on labelled bags

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

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

- Prediction:

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

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

Regression on labelled bags

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

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

- Prediction:

$$\hat{y}\left(\hat{P}\right) = \mathbf{g}^T(\mathbf{G} + \ell\lambda\mathbf{I})^{-1}\mathbf{y},$$
$$\mathbf{g} = \left[K\left(\mu_{\hat{P}}, \mu_{\hat{P}_i}\right)\right], \mathbf{G} = \left[K\left(\mu_{\hat{P}_i}, \mu_{\hat{P}_j}\right)\right], \mathbf{y} = [y_i].$$

Challenges

1. Inner product of distributions: $K\left(\mu_{\hat{P}_i}, \mu_{\hat{P}_j}\right) = ?$
2. How many samples/bag?

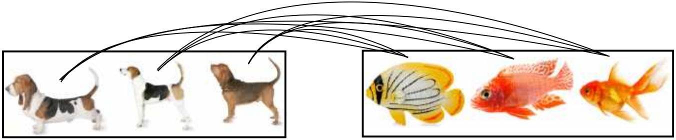
Regression on labelled bags: similarity

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

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

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

Intuition:



Regression on labelled bags: similarity

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

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

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

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

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

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

Regression on labelled bags: performance measure

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_ρ? Possible?

Consistency guarantee, optimal rate (well-specified case)

Well-specified case: optimal rate with P_i-s

Having access to P_i-s what rate can be achieved?

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

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

b – size of the input space, c – smoothness of f_ρ.

Well-specified case: result – briefly

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

Well-specified case: result – briefly

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

Our result

- If $2 \le a$, then $f_{\hat{\mathbf{z}}}^{\lambda}$ has optimal rate.

Well-specified case: result – briefly

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

Our result

- If $2 \leq a$, then $f_{\hat{\mathbf{z}}}^{\lambda}$ has optimal rate.
- In fact, $a = \frac{b(c+1)}{bc+1} < 2$ is enough.
- Consequence: regression with set kernel is consistent.

Well-specified case: computational & statistical tradeoff

Let $N = \tilde{\mathcal{O}}\left(\ell^a\right)$.

Our result

- If $\frac{b(c+1)}{bc+1} \leq a$, then $\mathcal{R}\left(f_{\hat{\mathbf{z}}}^{\lambda}\right) - \mathcal{R}\left(f_{\rho}\right) = \mathcal{O}\left(\ell^{-\frac{bc}{bc+1}}\right)$.

Well-specified case: computational & statistical tradeoff

Let $N = \tilde{\mathcal{O}}\left(\ell^a\right)$.

Our result

- If $\frac{b(c+1)}{bc+1} \leq a$, then $\mathcal{R}\left(f_{\hat{\mathbf{z}}}^{\lambda}\right) - \mathcal{R}\left(f_{\rho}\right) = \mathcal{O}\left(\ell^{-\frac{bc}{bc+1}}\right)$.
- If $a \leq \frac{b(c+1)}{bc+1}$, then $\mathcal{R}\left(f_{\hat{\mathbf{z}}}^{\lambda}\right) - \mathcal{R}\left(f_{\rho}\right) = \mathcal{O}\left(\ell^{-\frac{ac}{c+1}}\right)$.

Meaning:

- smaller a: computational saving, but reduced statistical efficiency.

Well-specified case: computational & statistical tradeoff

Let $N = \tilde{\mathcal{O}}(\ell^a)$.

Our result

- If $\frac{b(c+1)}{bc+1} \leq a$, then $\mathcal{R}\left(f_{\hat{\mathbf{z}}}^{\lambda}\right) - \mathcal{R}\left(f_{\rho}\right) = \mathcal{O}\left(\ell^{-\frac{bc}{bc+1}}\right)$.
- If $a \leq \frac{b(c+1)}{bc+1}$, then $\mathcal{R}\left(f_{\hat{\mathbf{z}}}^{\lambda}\right) - \mathcal{R}\left(f_{\rho}\right) = \mathcal{O}\left(\ell^{-\frac{ac}{c+1}}\right)$.

Meaning:

- smaller a: computational saving, but reduced statistical efficiency.
- $c \mapsto \frac{b(c+1)}{bc+1}$ decreasing: easier problems $\Rightarrow$ smaller bags.

Consistency guarantee: misspecified case

Misspecified case: results – briefly

Relevant setting: $f_\rho \in L^2 \backslash \mathcal{H}$. Results:

1. Generally: 'richness' of $\mathcal{H}$ is realizable.
2. If f_ρ is s-smooth, then we also get rates.

Misspecified case: generally

Let

- $N = \tilde{\mathcal{O}}(\ell)$,
- $\ell \to \infty$, $\lambda \to 0$, $\lambda\sqrt{\ell} \to \infty$.

Our result (consistency)

$\mathcal{R}\left(f_{\hat{\mathbf{z}}}^{\lambda}\right) - \mathcal{R}\left(f_{\rho}\right) \to \inf_{f \in \mathcal{H}} \left\| f - f_{\rho} \right\|_{L^2}$.

Misspecified case: s-smooth

Let $N = \tilde{O}\left(\ell^{2a}\right)$. f_ρ: s-smooth, $s \in (0,1]$.

Our result (computational & statistical tradeoff)

- If $\frac{s+1}{s+2} \leq a$, then $\mathcal{R}\left(f_{\hat{\mathbf{z}}}^{\lambda}\right) - \mathcal{R}\left(f_\rho\right) = \mathcal{O}\left(\ell^{-\frac{2s}{s+2}}\right)$.

Misspecified case: s-smooth

Let $N = \tilde{O}\left(\ell^{2a}\right)$. f_ρ: s-smooth, $s \in (0,1]$.

Our result (computational & statistical tradeoff)

- If $\frac{s+1}{s+2} \leq a$, then $\mathcal{R}\left(f_{\hat{\mathbf{z}}}^{\lambda}\right) - \mathcal{R}\left(f_\rho\right) = \mathcal{O}\left(\ell^{-\frac{2s}{s+2}}\right)$.
- If $a \leq \frac{s+1}{s+2}$, then $\mathcal{R}\left(f_{\hat{\mathbf{z}}}^{\lambda}\right) - \mathcal{R}\left(f_\rho\right) = \mathcal{O}\left(\ell^{-\frac{2sa}{s+1}}\right)$.

Meaning:

- Smaller a: computational saving, but reduced statistical efficiency.

Misspecified case: s-smooth

Let $N = \tilde{O}\left(\ell^{2a}\right)$. f_ρ: s-smooth, $s \in (0,1]$.

Our result (computational & statistical tradeoff)

- If $\frac{s+1}{s+2} \leq a$, then $\mathcal{R}\left(f_{\hat{\mathbf{z}}}^{\lambda}\right) - \mathcal{R}\left(f_\rho\right) = \mathcal{O}\left(\ell^{-\frac{2s}{s+2}}\right)$.
- If $a \leq \frac{s+1}{s+2}$, then $\mathcal{R}\left(f_{\hat{\mathbf{z}}}^{\lambda}\right) - \mathcal{R}\left(f_\rho\right) = \mathcal{O}\left(\ell^{-\frac{2sa}{s+1}}\right)$.

Meaning:

- Smaller a: computational saving, but reduced statistical efficiency.
- Sensible choice: $a \leq \frac{s+1}{s+2} \leq \frac{2}{3} \Rightarrow 2a \leq \frac{4}{3} < 2$!

Misspecified case: s-smooth

Let $N = \tilde{O}\left(\ell^{2a}\right)$. f_ρ: s-smooth, $s \in (0, 1]$.

Our result (computational & statistical tradeoff)

- If $\frac{s+1}{s+2} \le a$, then $\mathcal{R}\left(f_{\hat{\mathbf{z}}}^{\lambda}\right) - \mathcal{R}\left(f_\rho\right) = \mathcal{O}\left(\ell^{-\frac{2s}{s+2}}\right)$.
- If $a \le \frac{s+1}{s+2}$, then $\mathcal{R}\left(f_{\hat{\mathbf{z}}}^{\lambda}\right) - \mathcal{R}\left(f_\rho\right) = \mathcal{O}\left(\ell^{-\frac{2sa}{s+1}}\right)$.

Meaning:

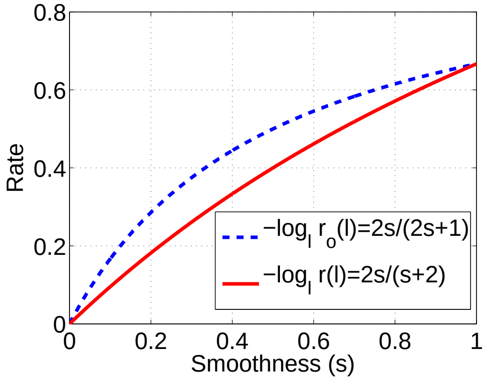
- Smaller a: computational saving, but reduced statistical efficiency.
- Sensible choice: $a \le \frac{s+1}{s+2} \le \frac{2}{3} \Rightarrow 2a \le \frac{4}{3} < 2$!
- $s \mapsto \frac{2s}{s+2}$ is increasing: easier task = better rate.
  - $s \to 0$: arbitrary slow rate. $s = 1$: $\mathcal{O}(\ell^{-\frac{2}{3}})$ speed.

Misspecified case: optimality

- Our rate: $r(\ell) = \ell^{-\frac{2s}{s+2}}$.
- One-stage sampled optimal rate: $r_o(\ell) = \ell^{-\frac{2s}{2s+1}}$ [Steinwart et al., 2009],
  - s-smoothness + eigendecay constraint,
  - $\mathcal{D}$: compact metric, $Y = \mathbb{R}$.

Misspecified case: optimality

- Our rate: $r(\ell) = \ell^{-\frac{2s}{s+2}}$.
- One-stage sampled optimal rate: $r_o(\ell) = \ell^{-\frac{2s}{2s+1}}$ [Steinwart et al., 2009],
  - s-smoothness + eigendecay constraint,
  - $\mathcal{D}$: compact metric, $Y = \mathbb{R}$.



s-smoothness: intuition

- Assumption: $f_\rho \in Im(C^s)$, $s \in (0,1]$. C = 'uncentered covariance'.

s-smoothness: intuition

- Assumption: $f_\rho \in Im(C^s)$, $s \in (0, 1]$. $C =$ 'uncentered covariance'.
- Imagine: $C \in \mathbb{R}^{d \times d}$ is a symmetric matrix,

$$C = U \Lambda U^T$$

s-smoothness: intuition

- Assumption: $f_\rho \in Im(C^s)$, $s \in (0,1]$. $C =$ 'uncentered covariance'.
- Imagine: $C \in \mathbb{R}^{d\times d}$ is a symmetric matrix,

$$C = U\Lambda U^T, \qquad Cv = \sum_{n=1}^{d} \lambda_n \langle u_n, v\rangle u_n.$$

s-smoothness: intuition

- Assumption: $f_\rho \in Im(C^s)$, $s \in (0,1]$. C = 'uncentered covariance'.
- Imagine: $C \in \mathbb{R}^{d\times d}$ is a symmetric matrix,

$$C = U\Lambda U^T, \qquad Cv = \sum_{n=1}^{d} \lambda_n \langle u_n, v\rangle u_n.$$

- General C:

$$C(v) = \sum_n \lambda_n \langle u_n, v\rangle u_n,$$

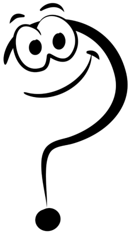
$$C^s(v) = \sum_n \lambda_n^s \langle u_n, v\rangle u_n,$$

$$Im(C^s) = \left\{ \sum_n c_n u_n : \sum_n c_n^2 \lambda_n^{-2s} < \infty \right\}.$$

Larger $s \Rightarrow$ faster decay of the c_n Fourier coefficients.

Summary

- Task: regression with labelled bags.
- Results:
  - consistency guarantees,
  - well-specified & misspecified case,
  - minimax rates.



Altun, Y. and Smola, A. (2006).
Unifying divergence minimization and statistical inference via convex duality.
In *Conference on Learning Theory (COLT)*, pages 139–153.

Berlinet, A. and Thomas-Agnan, C. (2004).
Reproducing Kernel Hilbert Spaces in Probability and Statistics.
Kluwer.

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

Gärtner, T., Flach, P. A., Kowalczyk, A., and Smola, A. (2002).
Multi-instance kernels.
In *International Conference on Machine Learning (ICML)*, pages 179–186.

Haussler, D. (1999).

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

- Smola, A., Gretton, A., Song, L., and Schölkopf, B. (2007).
A Hilbert space embedding for distributions.
In *Algorithmic Learning Theory (ALT)*, pages 13–31.

- Steinwart, I., Hush, D. R., and Scovel, C. (2009).
Optimal rates for regularized least squares regression.
In *Conference on Learning Theory (COLT)*.