Tailoring Density Estimation via Reproducing Kernel Moment Matching

Le Song[1] *Xinhua Zhang*[1] Alex Smola[1]
Arthur Gretton[2] Bernhard Schölkopf[2]

[1]Statistical Machine Learning Program, NICTA, Canberra, Australia

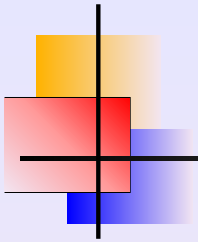
[2]Max Planck Institute for Biological Cybernetics, Tübingen, Germany

International Conference on Machine Learning
Helsinki, Finland, July 2008

THE AUSTRALIAN NATIONAL UNIVERSITY

Max-Planck-Institut für biologische Kybernetik

NICTA

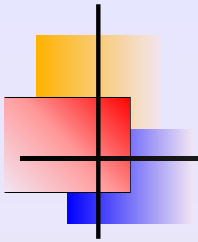


Outline

Motivation of our algorithm and estimation bounds

Formulation of our algorithm: quadratic programming

Experimental results



Motivation: Tailoring density estimation

- Density estimation is often NOT the ultimate goal
  - E.g. interested in expectations of a random variable (r.v.), or functions of the r.v.
  - E.g. parameter estimation for graphical models (gradient)
- Hence, not clear whether maximum likelihood is ideal

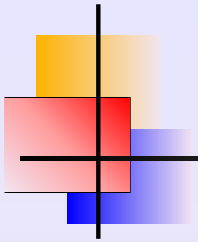
Full density estimation by MLE, for *arbitrary* functions

v.s.

Focus on approximating the expectation of a set of functions known a priori

Given a distribution p and function set $\mathcal{F}$, find distribution $\tilde{p}$, s.t.

$$|\mathbb{E}_{x\sim\tilde{p}}[f(x)] - \mathbb{E}_{x\sim p}[f(x)]| < \varepsilon \qquad \forall f \in \mathcal{F}$$



Motivation: Tailoring density estimation

- Density estimation is often NOT the ultimate goal
  - E.g. interested in expectations of a random variable (r.v.), or functions of the r.v.
  - E.g. parameter estimation for graphical models (gradient)
- Hence, not clear whether maximum likelihood is ideal

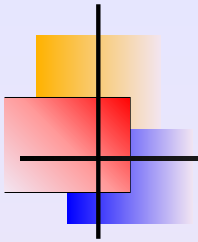
Full density estimation by MLE, for *arbitrary* functions

v.s.

Focus on approximating the expectation of a set of functions known a priori

Given a distribution p and function set $\mathcal{F}$, find distribution $\tilde{p}$, s.t.

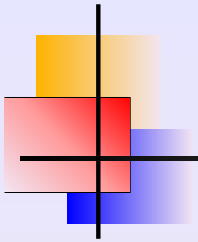
$$\sup_{f \in \mathcal{F}} |\mathbb{E}_{x \sim \tilde{p}}[f(x)] - \mathbb{E}_{x \sim p}[f(x)]| < \varepsilon$$



Idea of Touchstone functions classes

Similar spirit

- Weak convergence of probability measures
  - Probability measure $(\mu_n)_{n\geq 1}$ converges weakly to μ if
    $$\int f(x)\mu_n(\mathrm{d}x) \to \int f(x)\mu(\mathrm{d}x) \qquad n \to \infty$$
    $\forall\ f$ which is real valued, continuous and bounded on $\mathbb{R}^d$.
- Independence criteria [Rényi59]
  - for sufficiently rich function classes, the function correlation or cross-covariance serves as an independence test
- Density estimation [Shawe-Taylor and Dolia 07]:
  - Loss measured on a set of randomly drawn touchstone functions



Choice of function class: RKHS embeddings of distribution

Pick: $\mathcal{F} := \{f \in \mathcal{H} : \|f\|_{\mathcal{H}} \leq 1\}$

Given: distribution $p(x)$, kernel $k(\cdot, \cdot) \Rightarrow$ RKHS $\mathcal{H}$

$$\sup_{\|f\|_{\mathcal{H}} \leq 1} \left| \mathop{\mathbb{E}}_{x \sim p}[f(x)] - \mathop{\mathbb{E}}_{x \sim \tilde{p}}[f(x)] \right|$$

Choice of function class: RKHS embeddings of distribution

Pick: $\mathcal{F} := \{f \in \mathcal{H} : \|f\|_{\mathcal{H}} \leq 1\}$

Given: distribution $p(x)$, kernel $k(\cdot,\cdot) \Rightarrow$ RKHS $\mathcal{H}$

$$\sup_{\|f\|_{\mathcal{H}} \leq 1} \left| \mathop{\mathbb{E}}_{x\sim p}[f(x)] - \mathop{\mathbb{E}}_{x\sim \tilde{p}}[f(x)] \right| = \| \underbrace{\mathop{\mathbb{E}}_{x\sim p}[k(x,\cdot)]}_{:=\mu[p]} - \underbrace{\mathop{\mathbb{E}}_{x\sim \tilde{p}}[k(x,\cdot)]}_{:=\mu[\tilde{p}]} \|_{\mathcal{H}}$$

Key idea: Embed p, $\tilde{p}$ into the RKHS by kernel mean map:

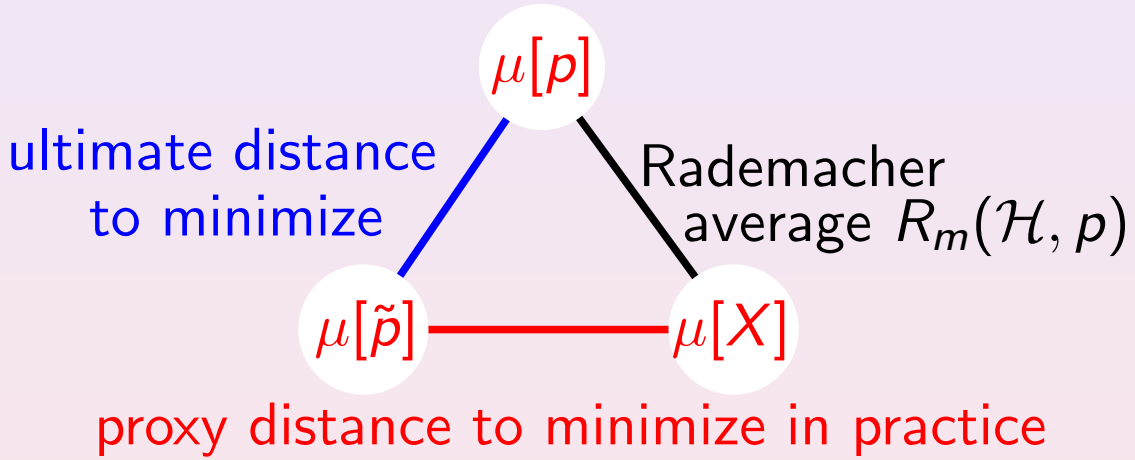
$$\boxed{\mu[p] := \mathbb{E}_{x\sim p}[k(x,\cdot)] = \int_{\mathcal{X}} k(x,\cdot)p(x)\mathrm{d}x}$$

Naturally expects: $\tilde{p} \approx p \quad \Leftrightarrow \quad \|\mu[\tilde{p}] - \mu[p]\|_{\mathcal{H}}$ is small

Estimation bounds

$$\text{real density } p \Rightarrow \begin{cases} \mu[p] = \mathbb{E}_{x\sim p}[k(x,\cdot)] \\ \text{sample } X \Rightarrow \mu[X] := \frac{1}{n}\sum_{i=1}^{n} k(x_i,\cdot) \end{cases}$$

$$\text{estimated density } \tilde{p} \Rightarrow \mu[\tilde{p}] = \mathbb{E}_{x\sim \tilde{p}}[k(x,\cdot)]$$



With probability $1 - \exp(-\epsilon^2 m R_m^{-2}/2)$, we have:

$$\|\mu[p] - \mu[\tilde{p}]\|_{\mathcal{H}} \leq 2R_m(\mathcal{H}, p) + \|\mu[X] - \mu[\tilde{p}]\|_{\mathcal{H}} + \epsilon$$

Formulation: Quadratic Programming

Suppose $\tilde{p} = \sum_{i=1}^{m} \alpha_i \underbrace{p_i}_{\text{fixed}}$, where α is in m-dim probability simplex Δ_m

$$\underset{\tilde{p}}{\text{minimize}} \quad \|\mu[\tilde{p}] - \mu[X]\|_{\mathcal{H}}$$

$$\Updownarrow$$

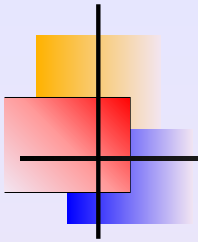
$$\boxed{\underset{\alpha \in \Delta_m}{\text{minimize}} \quad \frac{1}{2}\alpha^\top \mathbf{Q}\alpha - \mathbf{l}^\top \alpha}$$

where

$$\mathbf{Q}_{ij} = \langle \mu[p_i], \mu[p_j] \rangle_{\mathcal{H}} = \mathbb{E}_{x \sim p_i, x' \sim p_j}[k(x, x')]$$

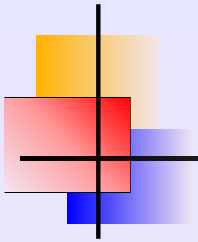
$$\mathbf{l}_i = \langle \mu[X], \mu[p_i] \rangle_{\mathcal{H}} = \frac{1}{n} \sum_{s=1}^{n} \mathbb{E}_{x \sim p_i}[k(x_s, x)]$$

Closed form formulae exist for $\mathbf{Q}_{ij}$ and $\mathbf{l}_i$.



Experimental results

- UCI dataset
- Application 1: Message passing compression
- Application 2: Image retrieval and categorization



UCI dataset

- To show
  - Outperforms in terms of estimating the expectation of functions which are in the RKHS
  - Does not outperform other algorithms in log likelihood
- Algorithms under comparison:
  - Kernel Moment Matching (KMM) $\leftarrow$ ours
  - Gaussian Mixture Model (GMM)
  - Parzen window (PZ)
  - Reduced Set Density Estimation (RSDE)
- What to compare and how
  1. Randomly generate function f from RKHS
  2. $1/2$ data for density estimation $\tilde{p}$, and $\mathbb{E}_{x\sim\tilde{p}}[f(x)]$.
  3. $1/2$ data to compute the empirical average of f
  4. Report relative discrepancy

Result of function expectation estimation on UCI datasets

Data	Polynomials ($d = 3$)				RBF Functions			
	PZ	GMM	RSDE	POL3	PZ	GMM	RSDE	RBF
covertype	0.418	0.539	1.240	0.412	0.073	0.023	0.071	0.020
ionosphere	0.615	0.664	1.659	0.626	0.120	0.024	0.142	0.022
sonar	0.691	0.745	2.558	0.673	0.857	0.030	0.873	0.029
australian	0.832	0.837	1.031	0.833	0.089	0.028	0.106	0.024
specft	0.922	0.878	1.265	0.867	0.903	0.067	0.904	0.062
wdbc	0.519	0.612	1.362	0.512	0.482	0.027	0.456	0.023
wine	0.679	0.718	2.782	0.682	0.471	0.040	0.545	0.039
satimage	0.260	0.281	1.230	0.256	0.307	0.028	0.359	0.026
segment	0.590	0.572	1.021	0.588	0.053	0.025	0.247	0.022
vehicle	0.496	0.478	1.686	0.493	0.095	0.028	0.325	0.027
svmguide2	0.866	0.782	2.603	0.729	0.798	0.019	0.808	0.018
vowel	0.348	0.394	1.741	0.352	0.028	0.019	0.111	0.018
housing	0.393	0.421	0.890	0.391	0.044	0.027	0.091	0.025
bodyfat	1.029	1.017	1.200	1.015	0.430	0.038	0.432	0.037
abalone	0.629	0.636	3.308	0.628	0.049	0.044	0.294	0.043
Total Win:	8	2	0	12	0	0	0	15

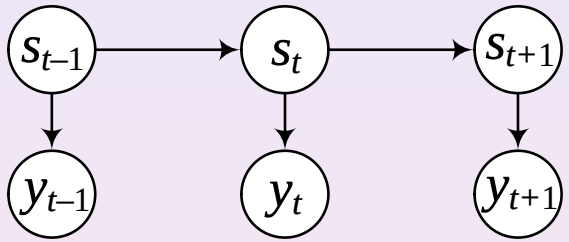
Result of function expectation estimation on UCI datasets

Data	Linear Functions				Polynomials ($d = 2$)			
	PZ	GMM	RSDE	LIN	PZ	GMM	RSDE	POL2
covertype	2.003	2.003	10.280	2.003	0.185	0.194	0.396	0.150
ionosphere	2.006	2.006	17.995	2.006	0.159	0.232	0.383	0.169
sonar	2.000	2.000	12.288	2.000	0.971	0.354	0.933	0.242
australian	2.000	2.000	14.217	2.000	0.369	0.380	0.587	0.380
specft	2.000	2.000	3.594	2.000	0.891	0.515	0.522	0.488
wdbc	2.004	2.004	16.447	2.004	0.209	0.233	0.406	0.166
wine	2.017	2.017	9.489	2.017	0.822	0.236	1.027	0.211
satimage	2.000	2.000	27.561	2.000	0.146	0.126	0.533	0.122
segment	2.003	2.003	23.388	2.003	0.258	0.245	0.803	0.263
vehicle	2.005	2.005	26.331	2.005	0.126	0.135	0.780	0.119
svmguide2	2.005	2.005	7.248	2.005	3.468	0.247	3.341	0.183
vowel	2.000	2.000	12.913	2.000	0.131	0.150	0.642	0.131
housing	2.000	2.000	7.668	2.000	0.117	0.126	0.399	0.121
bodyfat	2.000	2.000	7.295	2.000	0.288	0.243	0.595	0.242
abalone	2.005	2.005	17.010	2.005	0.105	0.101	0.234	0.103
Total Win:	15	15	0	15	6	0	0	15

Application 1: Message passing compression

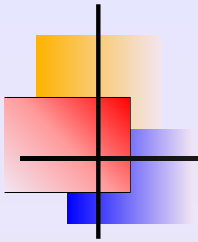
- Task: given a temporal system

$$s_t := f(s_{t-1}) + \xi,$$
$$y_t := g(s_t) + \zeta, \text{ where}$$
$$\xi \sim \tfrac{1}{5}\sum_{i=1}^{5} \mathcal{N}(\mu_i, \sigma^2)$$
$$\zeta \sim \mathcal{N}(0, \sigma^2)$$



  - Filtering: compute $p(s_T|Y_T)$, where $Y_T := (y_1, \ldots, y_T)$.
- Idea: recursively estimate the filtering density

$$p(s_{t+1}|Y_T) = \int p(s_{t+1}|s_t)p(s_t|Y_T)\mathrm{d}s_t$$
$$= \mathop{\mathbb{E}}_{s_t \sim \underbrace{p(s_t|Y_T)}_{\text{density estimated}}} [\ \underbrace{p(s_{t+1}|s_t)}_{\text{function in RKHS}}\ ]$$



Experimental results

Root mean square error and standard deviation of the filtering results before and after particle compression.

Particle #	PF	GMM	KMM
100	0.683±0.114	0.558±0.084	0.546±0.072
500	0.679±0.111	0.556±0.076	0.530±0.070
1000	0.685±0.111	0.556±0.082	0.526±0.070

Application 2: Image retrieval

- Task: Given an image database D and an image p, retrieve from D a set of images similar to p.

- Idea:
  - On each image, perform density estimation over the feature distribution
  - Retrieve by ranking Dissimilarity(p, q) over $q \in D$
  - Dissimilarity measure: Earth Mover's Distance (EMD) [Rubner et al, 98], applied on mixture of Gaussians

Side note: EMD is not in the RKHS in general.

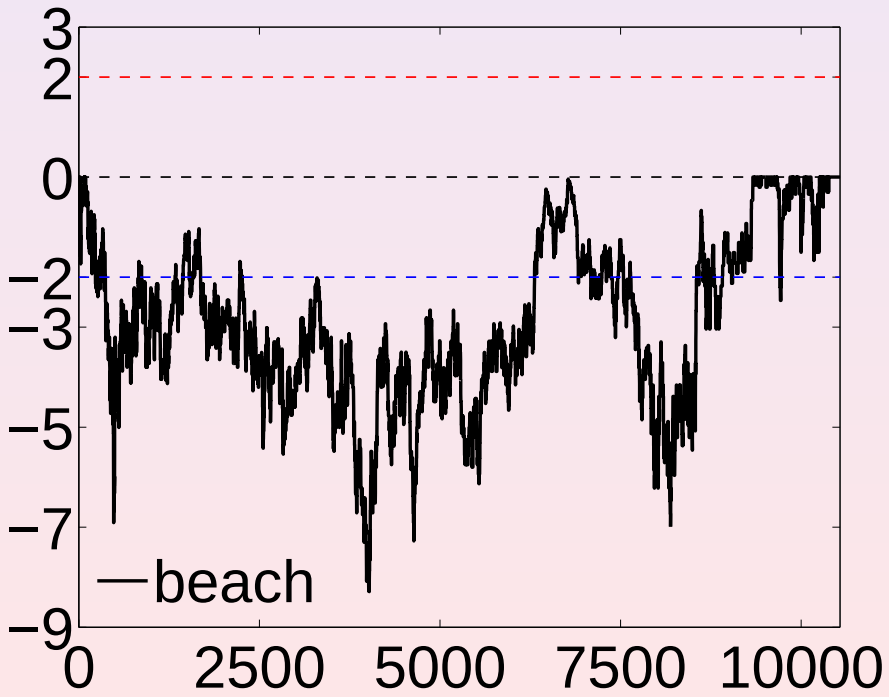
Experimental results

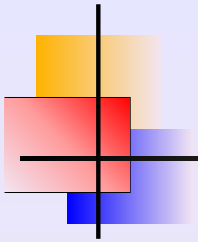
Horizontal axis: number of retrieved images

Vertical axis: signed log p-value of paired sign test.

> 2: GMM retrieved more correct images with significance < 0.01.

< -2: KMM retrieved more correct images with significance < 0.01.





Conclusion and discussion

- Proposed a density estimation algorithm tailored for a particular function class, solvable by simple QP
- Proved uniform convergence guarantees for approximating function expectations
- Experimental results show that KMM better approximates the expectation of functions in the RKHS, though no longer maximizing likelihood.
- Closely connected to expectation propagation
- Future directions:
  - Apply KMM to nonparametric loopy belief propagation in graphical model inference
  - Online data stream compression by solving online QP