Overlapping Domain Cover for Scalable and Accurate Regression Kernel Machines

Mohamed Elhoseiny
m.elhoseiny@cs.rutgers.edu
Ahmed Elgammal
elgammal@cs.rutgers.edu

Computer Science Department
Rutgers University
New Jersey, USA

Table 1: Contrast against most relevant methods

	[5]	FIC/PIC [7]	NN [1]	ODC (ours)
Accurate	No for high input dimension	Limited	Yes	Yes
Efficient	No	Yes	No	Yes
Scalable to arbitrary input dimension	No (2D)	Yes	Yes	Yes
Consistent on Boundaries	Yes	No	Yes	Yes
supported kernel machines	GPR	GPR	TGP	GPR, TGP, IWTGP and others
Easy to parallelize	No	No	Yes	Yes

Motivation. Recent advances in structure regression encouraged researchers to adopt it for formulating various problems with high dimensional output spaces, such as segmentation, detection, and image reconstruction, as regression problems. However, the computational complexity of the state-of-the-art regression algorithms limits their applicability for big data. In particular, kernel-based regression algorithms such as Ridge Regression [3], Gaussian Process Regression (GPR) [6], and the Twin Gaussian Processes (TGP) [1] require inversion of kernel matrices ($O(N^3)$, where N is the number of the training points), which limits their applicability for big data. We refer to these non-scalable versions of GPR and TGP as full-GPR and full-TGP, respectively.

The problems of the existing approximation approaches, detailed and justified in the paper, motivated us to develop an approach that satisfies the properties listed in table 1. The table also shows which of these properties are satisfied for the relevant methods. Khandekar et. al. [4] discussed properties and benefits of overlapping clusters for minimizing the conductance from spectral perspective. These properties of overlapping clusters also motivate studying scalable local prediction based on overlapping kernel machines; see figure 1.

Our Contribution. In summary, the main question, we address in this paper, is how local kernel machines with overlapping training data could help speedup the computations and gain accurate predictions. We achieved considerable speedup and good performance on GPR, TGP, and IWTGP (Importance Weighted TGP) applied to 3D pose estimation datasets. To the best of our knowledge, our framework is the first to achieve quadratic prediction complexity for TGP. The ODC concept is also novel in the context of kernel machines and is shown here to be successfully applicable to multiple kernel-machines. We also theoretically justified the idea behind our method and build on it to propose an ODC framework that reduces the complexity of TGP regression from cubic to quadratic. As a part of the framework, we proposed Assign&Balance K-Means algorithm, a version of K-means clustering that generates equal size clusters and we showed that it better than RPC used previously for GPR; see details in the main paper. We validated and analyzed our method on three human pose estimation datasets and interesting findings are discussed.

ODC Framework Overview We define the ODC as a collection of overlapping subsets of the training points, denoted by subdomains, such that they are as spatially coherent as possible. During training, an ODC is computed such that each subdomain overlaps with the neighboring subdomains. Then, a local prediction model (kernel machine) is created for each subdomain and the computations that does not depend on the test data are factored out and precomputed (e.g. inversion of matrices). The nature of the ODC generation makes these kernel machines consistent in the overlapped regions, which are the boundaries since we constraint the subdomains to be coherent. On prediction, the output is calculated as a reduction function of the predictions on the closed subdomain(s).

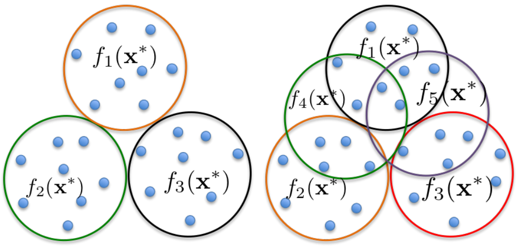

Figure 1: 24 points, Left: 3 disjoint kernel machines of 8 points, Right: 5 Overlapping kernel machines of 8 points. $f_i(\mathbf{x}^*)$ is the i^{th} kernel machine prediction for $\mathbf{x}^*$ test point.

Table 2: Error & Time on Poser and Human Eva datasets (Intel core-i7 2.6GHZ), M = 800

		Poser			HumanEva		
		Error (deg)	Training Time	Prediction Time	Error (mm)	Training Time	Prediction Time
TGP	NN [1]	5.43	-	188.99 sec	**38.1**	-	6364 sec
	ODC ($p = 0.9, t = 1, K' = 1$)-Ekmeans	**5.40**	(3.7 +25.1) sec	**16.5** sec	**38.9**	(2001 + 45.4) sec	**298** sec
	ODC ($p = 0, t = 1, K' = 1$)-Ekmeans	7.60	(3.9 + 1.33) sec	14.8 sec	41.87	(240 + 4.9) sec	257 sec
	ODC ($p = 0.9, t = 1, K' = 1$)-RPC	5.60	(0.23 +41.6) sec	15.8 sec	39.9	(0.45 + 49.1) sec	277 sec
	ODC ($p = 0, t = 1, K' = 1$)-RPC	7.70	(0.15 + 1.7) sec	13.89 sec	42.32	(0.19 + 5.2) sec	242 sec
GPR	NN	6.77	-	24 sec	54.8	-	618 sec
	ODC ($p = 0.9, t = 1, K' = 1$)-Ekmeans	**6.27**	(3.7 +11.1) sec	**0.56** sec	**49.3**	(2001 + 42.85)sec	**79** sec
	ODC($p = 0.0, t = 1, K' = 1$)-Ekmeans	7.54	(3.9 + 1.38 sec)	0.35 sec	49.6	(240 + 6.4) sec	48 sec
	ODC ($p = 0.9, t = 1, K' = 1$)-RPC	6.45	(0.23 +17.3) sec	0.52 sec	52.8	(0.49 + 46.06) sec	64 sec
	ODC ($p = 0.0, t = 1, K' = 1$)-RPC = [2]	7.46	(0.15 + 1.5) sec	0.27 sec	54.6	(0.26 + 4.6) sec	44 sec
	FIC [7]	7.63	(- + 20.63)	0.3106	68.36	-	102 sec

Given a set of input data $X = \{\mathbf{x}_1, \cdots, \mathbf{x}_N\}$, our prediction framework firstly generates a set of non-overlapping equal-size partitions, $C = \{C_1, \cdots, C_K\}$, such that $\cup_i C_i = X$, $|C_i| = N/K$. Then, the ODC is defined based on them as $\mathcal{D} = \{D_1, \cdots, D_K\}$, such that $|D_i| = M \forall i$, $D_i = C_i \cup O_i, \forall i$. O_i a the set of points that overlaps with the other partitions, i.e., $O_i = \{x : x \in \{\cup_{j \neq i} C_j\}\}$, such that $|O_i| = p \cdot M$, $|C_i| = (1-p) \cdot M$, $0 \leq p \leq 1$ is the ratio of points in each overlapping subdomain, D_i, that belongs to/overlaps with partitions, other than its own, C_i.

An ODC could be specified by two parameters, M and p, which are the number of points in each subdomain and the ratio of overlap respectively; this is since $K = N/(1-p)M$. As p goes to 0, the generated ODC reduces to the set of non-overlapping clusters. Similarly, as p approaches $1 - 1/M$, the ODC reduces to generating a cluster at each point with maximum overlap with other clusters, i.e., $K = N$, $|C_i| = 1$, and $|O_i| = M - 1$. *Our main claim is two fold. First, precomputing local kernel machines (e.g. GPR, TGP, IWTGP) during training on the ODC significantly increase the speedup on prediction time. Second, given a fixed M and N, as p increases, local prediction performance increases, theoretically supported by Lemma 4.1.* Detailed about training and prediction could be found in the main paper.

Lemma 4.1. Under ODC notion, as the overlap p increases, the closer the nearest model to an arbitrary test point and the more likely that model get trained on a big neighborhood of the test point; see the proof in the Supplementary Materials (SM).

Experiments. We validated our framework on Poser, HumanEva, and Human3.6M datasets for human pose estimation task. Table 2 shows comparison between our method and the baseline approximation methods on Poser and HumanEva datasets; details could be found in the paper. We also tried full TGP and GPR on Poser and Human Eva Datasets. Full TGP error is 5.35 for Poser and 40.3 for Human Eva. Full GPR error is 6.10 for Poser and 59.62 for Human Eva. The results indicate that ODC achieves either better or competitive to the full models. Based on our comprehensive experiments on HumanEva and Poser datasets, we conducted an experiment on Human3.6M dataset with TGP kernel machine, where $M = 1390$, $t = 1$, $p = 0.6, K' = 1$, Ekmeans for clustering. We achieved a speedup of $41.7X$ on prediction time using our ODC framework compared with NN-scheme, i.e., 7 days if NN-scheme is used versus 4.03 hours in our case. More experiments and details are in the paper.

[1] Liefeng Bo and Cristian Sminchisescu. Twin gaussian processes for structured prediction. *Int. J. Comput. Vision*, 87(1-2):28–52, March 2010. ISSN 0920-5691.

[2] Krzysztof Chalupka, Christopher K. I. Williams, and Iain Murray. A framework for evaluating approximation methods for gaussian process regression. *JMLR*, 14(1), February 2013.

[3] A. E. Hoerl and R. W. Kennard. Ridge Regression: Biased Estimation for Nonorthogonal Problems. *Technometrics*, 1970.

[4] Rohit Khandekar, Guy Kortsarz, and Vahab Mirrokni. On the advantage of overlapping clusters for minimizing conductance. *Algorithmica*, 69(4):844–863, 2014.

[5] Chiwoo Park, Jianhua Z. Huang, and Yu Ding. Domain decomposition approach for fast gaussian process regression of large spatial data sets. *Journal of Machine Learning Research*, 12:1697–1728, 2011.

[6] Carl Edward Rasmussen and Christopher K. I. Williams. *Gaussian Processes for Machine Learning (Adaptive Computation and Machine Learning)*. The MIT Press, 2005. ISBN 026218253X.

[7] Edward Snelson and Zoubin Ghahramani. Sparse gaussian processes using pseudo-inputs. In *NIPS*, 2006.