Fast High-dimensional Kernel Summations Using the Monte Carlo Multipole Method

Dongryeol Lee (dongryel@cc.gatech.edu), Georgia Institute of Technology
Alexander Gray (agray@cc.gatech.edu), Georgia Institute of Technology

Problem

In this paper, we propose new computational techniques for efficiently approximating the following sum for each *query point* $q_i \in \mathcal{Q}$:

$$\Phi(q_i, \mathcal{R}) = \sum_{r_j \in \mathcal{R}} e^{-||q_i - r_j||^2/(2h^2)} \quad (1)$$

where $\mathcal{R}$ is the *reference set*; each *reference point* is associated with a Gaussian function with a smoothing parameter h (the 'bandwidth'). This form of summation is ubiquitous in many statistical learning methods:

- Kernel density estimation
- Kernel regression
- Gaussian process regression
- Radial basis function networks
- Spectral clustering
- Support vector machines
- Kernel PCA

Cross-validation in all of these methods require evaluating Equation 1 for multiple values of h. Kernel density estimation, for example, requires $|\mathcal{R}|$ density estimate based on $|\mathcal{R}| - 1$ points, yielding a brute-force computational cost scaling **quadratically** (that is $O(|\mathcal{R}|^2)$).

Error Bounds

Due to expensive computational cost, many algorithms approximate the Gaussian kernel sums at the expense of reduced precision, two of which is shown below:

Definition 0.1. *An algorithm guarantees* ϵ **absolute error bound**, *if for each exact value* $\Phi(q_i, \mathcal{R})$ *for* $q_i \in \mathcal{Q}$, *it computes* $\widetilde{\Phi}(q_i, \mathcal{R})$ *such that* $\left|\widetilde{\Phi}(q_i, \mathcal{R}) - \Phi(q_i, \mathcal{R})\right| \leq \epsilon$.

Definition 0.2. *An algorithm guarantees* ϵ **relative error bound**, *if for each exact value* $\Phi(q_i, \mathcal{R})$ *for* $q_i \in \mathcal{Q}$, *it computes* $\widetilde{\Phi}(q_i, \mathcal{R}) \in \mathbb{R}$ *such that* $\left|\widetilde{\Phi}(q_i, \mathcal{R}) - \Phi(q_i, \mathcal{R})\right| \leq \epsilon \left|\Phi(q_i, \mathcal{R})\right|$.

Our Contributions

We propose a better Gaussian summation algorithm that overcomes the curse of dimensionality:

- Extends an earlier work [8] to guarantee "per-query" estimates using the following new error bound criterion:

  Definition 0.3. *An algorithm guarantees* $(1 - \alpha)$ **probabilistic** ϵ **relative error bound**, *if for each exact value* $\Phi(q_i, \mathcal{R})$ *for* $q_i \in \mathcal{Q}$, *it computes* $\widetilde{\Phi}(q_i, \mathcal{R}) \in \mathbb{R}$, *such that with at least probability* $0 < 1 - \alpha < 1$, $\left|\widetilde{\Phi}(q_i, \mathcal{R}) - \Phi(q_i, \mathcal{R})\right| \leq \epsilon \left|\Phi(q_i, \mathcal{R})\right|$.

- A new tree structure called *subspace tree* for reducing the computational cost of each distance computation inspired by an earlier work [11].
- Extensive experimental results up to 89 dimensional datasets for the first time.

Algorithm

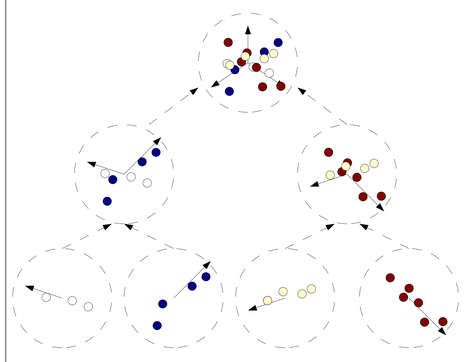
MCMM(Q, R, β)
 if CANSUMMARIZEEXACT(Q, R, ϵ)
  SUMMARIZEEXACT(Q, R)
 else if CANSUMMARIZEMC(Q, R, ϵ, β)
  SUMMARIZEMC(Q, R, ϵ, β)
 else
  if Q is a leaf node
   if R is a leaf node
    MCMMBASE(Q, R)
   else
    MCMM $\left(Q, R^L, \frac{\beta}{2}\right)$, MCMM $\left(Q, R^R, \frac{\beta}{2}\right)$
  else
   if R is a leaf node
    MCMM(Q^L, R, β), MCMM(Q^R, R, β)
   else
    MCMM $\left(Q^L, R^L, \frac{\beta}{2}\right)$, MCMM $\left(Q^L, R^R, \frac{\beta}{2}\right)$
    MCMM $\left(Q^R, R^L, \frac{\beta}{2}\right)$, MCMM $\left(Q^R, R^R, \frac{\beta}{2}\right)$

Subspace Tree

We use principal component analysis for building a subspace for each node.

- Leaf node: computed using PCABASE which can use the exact PCA [2] or a stochastic one [1].
- Internal node: the subspaces of the child nodes are approximately merged using the MERGESUBSPACES function [7].

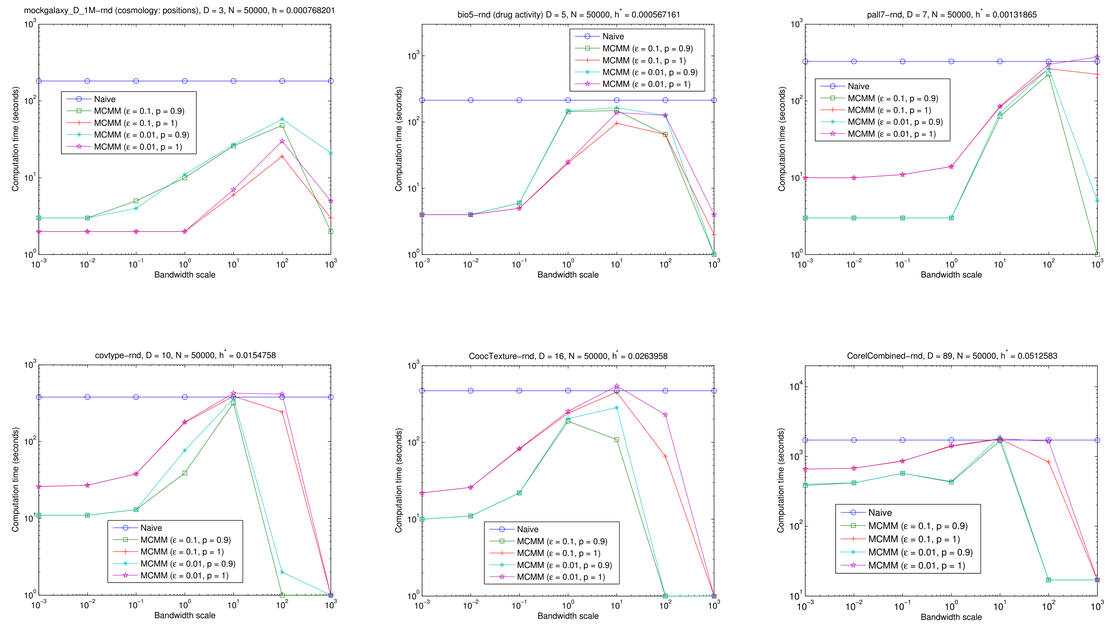
BUILDPCATREE$(\mathcal{P})$
 if CANPARTITION$(\mathcal{P})$
  $\{\mathcal{P}^L, \mathcal{P}^R\} \leftarrow$ PARTITIONSET$(\mathcal{P})$
  $N \leftarrow$ empty node
  $N^L \leftarrow$ BUILDPCATREE$(\mathcal{P}^L)$
  $N^R \leftarrow$ BUILDPCATREE$(\mathcal{P}^R)$
  $N.S \leftarrow$ MERGESUBSPACES$(N^L.S, N^R.S)$
 else
  $N \leftarrow$ BUILDPCATREEBASE$(\mathcal{P})$
  $N.S \leftarrow$ PCABASE$(\mathcal{P})$
  $N.\mathcal{P}_{proj} \leftarrow$ PROJECT$(\mathcal{P}, N.S)$
 return N



Experimental Results and Conclusion

We evaluated our algorithm [a] on six real-world datasets (scaled to fit in $[0, 1]^D$ hypercube) for kernel density estimation at every query point with a range of bandwidths, from 3 orders of magnitude smaller than to three orders larger than optimal, according to the least-squares cross-validation score [12].

- On low dimensional datasets (below 6 dimensions), the algorithm using series-expansion based bounds (MCMM algorithm with $p = 1$ value) gives two to three times speedup compared to our approach that uses Monte Carlo sampling.
- From 7 dimensions and beyond, our probabilistic algorithm ($p = 0.9$) consistently performs better than the algorithm using exact bounds by at least a factor of two.
- Our new method has the minimum time requirement for cross-validation over all bandwidth scales over the past work [3, 4, 6, 5, 10, 9]. We have already developed an improved version of our algorithm.



References

[1] P. Drineas, R. Kannan, and M. Mahoney. Fast monte carlo algorithms for matrices iii: Computing a compressed approximate matrix decomposition, 2004.

[2] G. Golub. *Matrix Computations, Third Edition*. The Johns Hopkins University Press, 1996.

[3] A. Gray and A. W. Moore. N-Body Problems in Statistical Learning. In Todd K. Leen, Thomas G. Dietterich, and Volker Tresp, editors, *Advances in Neural Information Processing Systems 13 (December 2000)*. MIT Press, 2001.

[4] Alexander G. Gray and Andrew W. Moore. Nonparametric Density Estimation: Toward Computational Tractability. In *SIAM International Conference on Data Mining 2003*, 2003.

[5] Alexander G. Gray and Andrew W. Moore. Rapid Evaluation of Multiple Density Models. In *Artificial Intelligence and Statistics 2003*, 2003.

[6] Alexander G. Gray and Andrew W. Moore. Very Fast Multivariate Kernel Density Estimation via Computational Geometry. In *Joint Statistical Meeting 2003*, 2003. to be submitted to JASA.

[7] Peter Hall, David Marshall, and Ralph Martin. Merging and splitting eigenspace models. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 22(9):1042–1049, 2000.

[8] Michael Holmes, Alexander Gray, and Charles Isbell. Ultrafast monte carlo for statistical summations. In J.C. Platt, D. Koller, Y. Singer, and S. Roweis, editors, *Advances in Neural Information Processing Systems 20*, pages 673–680. MIT Press, Cambridge, MA, 2008.

[9] Dongryeol Lee and Alexander Gray. Faster gaussian summation: Theory and experiment. In *Proceedings of the Twenty-second Conference on Uncertainty in Artificial Intelligence*. 2006.

[10] Dongryeol Lee, Alexander Gray, and Andrew Moore. Dual-tree fast gauss transforms. In Y. Weiss, B. Schölkopf, and J. Platt, editors, *Advances in Neural Information Processing Systems 18*, pages 747–754. MIT Press, Cambridge, MA, 2006.

[11] A. W. Moore, J. Schneider, and K. Deng. Efficient locally weighted polynomial regression predictions. In D. Fisher, editor, *Proceedings of the Fourteenth International Conference on Machine Learning*, pages 196–204, San Francisco, 1997. Morgan Kaufmann.

[12] B. W. Silverman. *Density Estimation for Statistics and Data Analysis*. Chapman and Hall/CRC, 1986.

[a] Times which include preprocessing costs are measured in CPU seconds on an AMD Opteron 3.0 GHz machine with 8 Gb of main memory and 1 Mb of CPU cache. All the codes that we have written and obtained are written in C and C++, and was compiled under `-O3` flags on Linux kernel 2.6.9. We have limited all datasets to 50K points.