Doc. Math. J. DMV 575

A New Version of the Fast Gauss Transform

LESLIE GREENGARD AND XIAOBAI SUN

ABSTRACT. The evaluation of the sum of N Gaussians at M points in space arises as a computational task in diffusion, fluid dynamics, finance, and, more generally, in mollification. The work required for direct evaluation grows like the product NM, rendering large-scale calculations impractical. We present an improved version of the fast Gauss transform [L. Greengard and J. Strain, SIAM J. Sci. Stat. Comput. 12, 79 (1991)], which evaluates the sum of N Gaussians at M arbitrarily distributed points in O(N+M) work, where the constant of proportionality depends only on the precision required. The new scheme is based on a diagonal form for translating Hermite expansions and is significantly faster than previous versions.

1991 Mathematics Subject Classification: 65R10, 44A35, 35K05 Keywords and Phrases: diffusion, fast algorithms, Gauss transform

1 Introduction

Many problems in mathematics and its applications involve the Gauss transform

$$G_{\delta}f(x) = (\pi\delta)^{-d/2} \int_{\Gamma} e^{-|x-y|^2/\delta} f(y) \, dy \qquad (\delta > 0)$$
 (1)

of a function f, where Γ is some subset of \mathbf{R}^d . This is, of course, the exact solution to the Cauchy problem

$$\begin{array}{rcl} u_t(x,t) & = & \Delta u(x,t), & t > 0 \\ u(x,0) & = & f(x), & x \in \mathbf{R}^d \end{array}$$

at time $t = \delta/4$ and corresponds to a *mollification* of the function f. Similar transforms occur in solving initial/boundary value problems for the heat equation by means of potential theory [3, 10, 11] and in nonparametric statistics [4, 17].

In the present paper, we will focus our attention on the $\emph{discrete Gauss transform}$

$$G(x) = \sum_{j=1}^{N} q_j e^{-|x-s_j|^2/\delta} , \qquad (2)$$

where the coefficients q_j and "source" locations s_j are given, and we wish to evaluate the expression (2) at a large number of "target" points x_j .

If the number of target points is denoted by M, we can define the rectangular transform matrix G by the formula

$$\mathbf{G}_{ij} = e^{-|x_i - s_j|^2/\delta}. (3)$$

Direct application of this matrix to the vector $q = (q_1, \ldots, q_N)^T$. requires O(NM) work, which makes large scale calculations prohibitively expensive.

To overcome this obstacle, Greengard and Strain developed a fast Gauss transform [9], which requires only O(N+M) work, with a constant prefactor which depends on the physical dimension d and the desired precision. The amount of memory required is also proportional to N+M, so that the algorithm is asymptotically optimal in terms of both work and storage. In this scheme, the sources and targets can be placed anywhere; methods based on the fast Fourier transform (FFT), by contrast, are restricted to a regular grid and require $O(N \log N)$ operations. For the case where the variance δ is not constant:

$$G(x) = \sum_{j=1}^{N} q_j e^{-|x-s_j|^2/\delta_j},$$
(4)

a generalization of the fast Gauss transform has been developed by Strain [18], but we will consider only the simpler case (2) here.

The fast Gauss transform is an analysis-based fast algorithm. Like the closely related fast multipole methods for the Laplace and Helmholtz equations [1, 5, 7, 13, 8, 14, 15, 16], it achieves a speedup in computation by using approximation theory to attain a specified, albeit arbitrarily high, precision. The FFT, on the other hand, is exact in exact arithmetic. It is an algebra-based fast algorithm which uses symmetry properties to reduce the computational work.

2 The original fast Gauss transform

The starting point for the fast algorithms of [9, 18] is the generating function for Hermite polynomials [2, 12]

$$e^{2xs-s^2} = \sum_{n=0}^{\infty} \frac{s^n}{n!} H_n(x) ,$$

where

$$H_n(x) = (-1)^n e^{x^2} D^n e^{-x^2}$$
 $x \in \mathbf{R}$

and D = d/dx. A small amount of algebra leads to the expansion

$$e^{-(x-s)^2/\delta} = \sum_{n=0}^{\infty} \frac{1}{n!} \left(\frac{s-s_0}{\sqrt{\delta}} \right)^n h_n \left(\frac{x-s_0}{\sqrt{\delta}} \right),$$

where

$$h_n(x) = (-1)^n D^n e^{-x^2}$$

and s_0 is an arbitrary point.

This formula describes the Gaussian field $e^{-(x-s)^2/\delta}$ at the target x due to the source at s as an Hermite expansion centered at s_0 . The higher dimensional analog of (5) is obtained using multi-index notation. Let x and s lie in d-dimensional Euclidean space \mathbf{R}^d , and consider the Gaussian

$$e^{-|x-s|^2} = e^{-(x_1-s_1)^2 - \dots - (x_d-s_d)^2}$$

For any multi-index $\alpha = (\alpha_1, \alpha_2, \dots, \alpha_d)$ and any $x \in \mathbf{R}^d$, we define

$$|\alpha| = \alpha_1 + \alpha_2 + \dots + \alpha_d$$

$$\alpha! = \alpha_1! \alpha_2! \dots \alpha_d!$$

$$x^{\alpha} = x_1^{\alpha_1} x_2^{\alpha_2} \dots x_d^{\alpha_d}$$

$$D^{\alpha} = \partial_1^{\alpha_1} \partial_2^{\alpha_2} \dots \partial_d^{\alpha_d}$$

where ∂_i is differentiation with respect to the *i*th coordinate in \mathbf{R}^d . If p is an integer, we say $\alpha \geq p$ if $\alpha_i \geq p$ for $1 \leq i \leq d$.

The multidimensional Hermite polynomials and Hermite functions are defined by

$$H_{\alpha}(x) = H_{\alpha_{1}}(x_{1}) \dots H_{\alpha_{d}}(x_{d})$$

$$h_{\alpha}(x) = e^{-|x|^{2}} H_{\alpha}(x) = h_{\alpha_{1}}(x_{1}) \dots h_{\alpha_{d}}(x_{d})$$
(5)

where $|x|^2 = x_1^2 + \ldots + x_d^2$. The Hermite expansion of a Gaussian in \mathbf{R}^d is then simply

$$e^{-|x-s|^2} = \sum_{\alpha \ge 0} \frac{(x-s_0)^{\alpha}}{\alpha!} h_{\alpha}(s-s_0).$$
 (6)

LEMMA 2.1 ([9], 1991) Let N_B sources s_j lie in a box B with center s_B and side length $\sqrt{\delta}$. Then the Gaussian field due to the sources in B,

$$G(x) = \sum_{j=1}^{N_B} q_j e^{-|x-s_j|^2/\delta},$$
 (7)

is equal to a single Hermite expansion about s_B :

$$G(x) = \sum_{\alpha \ge 0} A_{\alpha} h_{\alpha} \left(\frac{x - s_B}{\sqrt{\delta}} \right).$$

The coefficients A_{α} are given by

$$A_{\alpha} = \frac{1}{\alpha!} \sum_{j=1}^{N_B} q_j \left(\frac{s_j - s_B}{\sqrt{\delta}} \right)^{\alpha} . \tag{8}$$

The error $E_H(p)$ due to truncating the series after p^d terms satisfies the bound:

$$|E_H(p)| = |\sum_{\alpha > p} A_{\alpha} h_{\alpha} \left(\frac{x - s_B}{\sqrt{\delta}} \right)| \le 2.75^d Q_B \left(\frac{1}{p!} \right)^{d/2} \left(\frac{1}{2} \right)^{(p+1)d/2}$$
 (9)

where $Q_B = \sum |q_j|$.

Lemma 2.2 ([9], 1991) Let N_B sources s_j lie in a box B with center s_B and side length $\sqrt{\delta}$ and let x be a target point in a box C with center x_C . Then the corresponding Hermite expansion

$$G(x) = \sum_{\alpha \ge 0} A_{\alpha} h_{\alpha} \left(\frac{x - s_B}{\sqrt{\delta}} \right).$$

can be expanded as a Taylor series of the form

$$G(x) = \sum_{\beta > 0} B_{\beta} \left(\frac{x - x_C}{\sqrt{\delta}} \right)^{\beta}.$$

The coefficients B_{β} are given by

$$B_{\beta} = \frac{(-1)^{|\beta|}}{\beta!} \sum_{\alpha > 0} A_{\alpha} h_{\alpha+\beta} \left(\frac{s_B - x_C}{\sqrt{\delta}} \right). \tag{10}$$

The error $E_T(p)$ due to truncating the series after p^d terms satisfies the bound:

$$|E_T(p)| = |\sum_{\beta \ge p} B_\beta \left(\frac{x - x_C}{\sqrt{\delta}}\right)^\beta | \le 2.75^d Q_B \left(\frac{1}{p!}\right)^{d/2} \left(\frac{1}{2}\right)^{(p+1)d/2}$$
 (11)

These are the only tools required to construct a simple fast algorithm for the evaluation of

$$G(x_i) = \sum_{i=1}^{N} q_j e^{-|x_i - s_j|^2/\delta}$$
(12)

for $1 \leq i \leq M$, using O(M+N) work. By shifting the origin and rescaling δ if necessary, we can assume (as a convenient normalization) that the sources s_j and targets x_i all lie in the unit box $B_0 = [0,1]^d$.

ALGORITHM

STEP 1. Subdivide B_0 into smaller boxes with sides of length $\sqrt{\delta}$ parallel to the axes. Assign each source s_j to the box B in which it lies and each target x_i to the box C in which it lies. The source boxes B and the target boxes C may, of course, be the same.

STEP 2. Given ϵ , use Lemma 2.1 to create an Hermite expansion for each source box B with p^d terms satisfying:

$$G(x) = \sum_{B} \sum_{s_{j} \in B} q_{j} e^{-|x-s_{j}|^{2}/\delta}$$
$$= \sum_{B} \sum_{\alpha \leq p} A_{\alpha}(B) h_{\alpha} \left(\frac{x-s_{B}}{\sqrt{\delta}}\right) + O(\epsilon)$$

where

$$A_{\alpha}(B) = \frac{1}{\alpha!} \sum_{s_j \in B} q_j \left(\frac{s_j - s_B}{\sqrt{\delta}} \right)^{\alpha}. \tag{13}$$

The amount of work required for this step is of the order p^dN .

Consider now a fixed target box C. For each $x_j \in C$, we need to evaluate the total field due to sources in all boxes of type B. Because of the exponential decay of the Gaussian field, however, it is easy to verify that, if we include only the sources in the nearest $(2r+1)^d$ boxes, we incur an error bounded by Qe^{-r^2} , where $Q = \sum_{j=1}^N |q_j|$. Given a desired precision ϵ , we can always choose r so that this truncation error is bounded by $Q\epsilon$. With r=4, for example, we get single precision accuracy ($\epsilon=10^{-7}$) and with r=6, we get double precision ($\epsilon=10^{-14}$). We denote the nearest $(2r+1)^d$ boxes as the interaction region for box C, denoted by IR(C).

STEP 3. For each target box C, use Lemma 2.2 to transform all Hermite expansions in source boxes within the $interaction\ region$ into a single Taylor expansion. Thus, we approximate G(x) in C by

$$G(x) = \sum_{B} \sum_{s_{j} \in B} q_{j} e^{-|x-s_{j}|^{2}/\delta}$$
$$= \sum_{\beta < p} C_{\beta} \left(\frac{x-x_{C}}{\sqrt{\delta}}\right)^{\beta} + O(\epsilon)$$

where

$$C_{\beta} = \frac{(-1)^{|\beta|}}{\beta!} \sum_{B \in IR(C)} \sum_{\alpha \le p} A_{\alpha}(B) h_{\alpha+\beta} \left(\frac{s_B - x_C}{\sqrt{\delta}}\right) , \qquad (14)$$

and the coefficients $A_{\alpha}(B)$ are given by (13). Because of the product form (5) of $h_{\alpha+\beta}$, the computation of the p^d coefficients C_{β} involves only $O(d\,p^{d+1})$ operations for each box B. Therefore, a total of $O((2r+1)^d\,d\,p^{d+1})$ work per target box C is required. Finally, evaluating the appropriate Taylor series for each target x_i requires $O(p^dM)$ work. Hence this algorithm has net CPU requirements of the order

$$O((2r+1)^d\,d\,p^{d+1}N_{box}) + O(p^dN) + O(p^dM)\ ,$$

where the number of boxes N_{box} is bounded by $min(\delta^{-d/2}, N+M)$. The work is cleanly decoupled into three parts; $O(p^dN)$ to form Hermite expansions, $O(p^dM)$ to evaluate Taylor series, and a constant term depending on the number of box-box interactions and the cost of transforming Hermite expansions into Taylor series.

REMARK: A proper implementation of the fast Gauss transform is a bit more complex. For example, if a box contains only a few sources, it is more efficient to compute their influence directly than to use expansions.

Suppose now that the source boxes are denoted by B_1, B_2, \ldots, B_S , that the target boxes are denoted by C_1, C_2, \ldots, C_T , that N_j sources lie in box B_j , that M_j targets lie in box C_j , and that the points are ordered so that

Then the approximation G_{ϵ} to the discrete Gauss transform matrix (3) can be written in the factored form

$$\mathbf{G}_{\epsilon} = \mathbf{D} \cdot \mathbf{E} \cdot \mathbf{F}. \tag{15}$$

Here, **F** is a block diagonal matrix of dimension $S \times S$. The *j*th diagonal block $\mathbf{F}(j) \in \mathbf{R}^{p^d \times N_j}$ satisfies

$$\mathbf{F}(j)_{n,m} = \frac{1}{\alpha_n!} \left(\frac{s_m - s_{B_j}}{\sqrt{\delta}} \right)^{\alpha_n},$$

where s_{B_j} is the center of box B_j and the p^d Hermite expansion coefficients are ordered in some fashion from $n=0,\ldots,p^d$. **D** is very similar. It is a block diagonal matrix of dimension $T\times T$, with the jth diagonal block $\mathbf{D}(j)\in\mathbf{R}^{M_j\times p^d}$ satisfying

$$\mathbf{D}(j)_{n,m} = \frac{(-1)^{\beta_m}}{\beta_m!} \left(\frac{x_n - x_{C_j}}{\sqrt{\delta}}\right)^{\beta_m}$$

where x_{C_j} is the center of box C_j , and the p^d Taylor expansion coefficients are ordered from $n = 0, ..., p^d$ in the same fashion as the Hermite series. Note that, if the sources and targets coincide, then **D** is the transpose of **F**.

The mapping \mathbf{E} is a sparse block matrix of dimension $T \times S$, with up to $(2r+1)^d$ nonzero entries per row. The nonzero entries $\mathbf{E}(ij)$ are matrices of dimension $p^d \times p^d$, corresponding to a conversion of the Hermite series for box S_j into a Taylor series for box T_i , assuming S_j is in the interaction region $IR(T_i)$. The matrix entries are dense.

$$\mathbf{E}(ij)_{nm} = h_{\alpha_n + \beta_m} \left(\frac{s_{B_j} - x_{C_i}}{\sqrt{\delta}} \right).$$

Given this notation, Step 2 of the fast Gauss transform described above corresponds to multiplying the vector $\{q_1, q_2, \ldots, q_N\}$ by \mathbf{F} . Step 3 of the fast Gauss transform corresponds to multiplying the output of Step 2 by \mathbf{E} to create all the Taylor expansions. The result is then multiplied by \mathbf{D} to evaluate the Taylor series at all target locations.

REMARK: The factorization (15) reveals the structure of G_{ϵ} . When δ is large enough, only one box is created and the rank of G_{ϵ} is bounded by p^d (the order of the factor **E**). When δ is very small, the dimensions of **E** grow, but it becomes sparse and structured.

3 Diagonal form for translation operators

Our new version of the fast Gauss transform is based on replacing Hermite and Taylor expansions with an expansion in terms of exponentials (plane waves). The starting point is the Fourier relation

$$e^{-|x-s|^2/\delta} = \left(\frac{1}{2\sqrt{\pi}}\right)^d \int_{\mathbf{R}^d} e^{-|k|^2/4} e^{ik \cdot (x-s)/\sqrt{\delta}} dk$$
 (16)

which is easily seen to satisfy the estimate

$$\left| e^{-\frac{|x-s|^2}{\delta}} - \left(\frac{1}{2\sqrt{\pi}}\right)^d \int_{|k| \le K} e^{-\frac{|k|^2}{4}} e^{i\frac{k \cdot (x-s)}{\sqrt{\delta}}} dk \right| \le \begin{cases} e^{-\frac{K^2}{4}} & \text{for } d = 1, 2\\ Ke^{-\frac{K^2}{4}} & \text{for } d = 3. \end{cases}$$

Setting K=7.5, the truncation error from ignoring high frequency contributions is approximately 10^{-7} . Ssetting K=12, the truncation error is approximately 10^{-14} . It still remains to discretize the Fourier integral in (16) within the range determined by K. The trapezoidal rule is particularly appropriate here since it is rapidly convergent for functions which have decayed at the boundary. Note, however, that the integrand is more and more oscillatory as x-s grows. Fortunately, we only need accurate quadrature when s is within the interaction region of x, so that $|x-s|/\sqrt{\delta} \le 5$ for seven digit precision and $|x-s|/\sqrt{\delta} \le 7$ for fourteen digit precision. It is easy to verify that p=12 equispaced modes in the interval [0,7.5] are sufficient to reduce the quadrature error to 10^{-7} when $|x-s|/\sqrt{\delta} \le 5$ and that p=24 equispaced modes in the interval [0,12] are sufficient to reduce the quadrature error to 10^{-14} when $|x-s|/\sqrt{\delta} \le 7$.

Thus, for a source box B with center s_B , we replace the Hermite series of Lemma 2.1 with

$$\begin{split} G(x) &= \sum_{s_j \in B} q_j e^{-|x-s_j|^2/\delta} \\ &= \sum_{\beta \leq p} C_\beta e^{i\frac{K\beta \cdot (x-s_B)}{p\sqrt{\delta}}} + O(\epsilon), \end{split}$$

where

$$C_{\beta} = \left(\frac{K}{2p\sqrt{\pi}}\right)^{d} e^{-|\beta|^{2}K^{2}/(4p^{2})} \sum_{j=1}^{N_{B}} q_{j} e^{-i\frac{K\beta \cdot (s_{j} - s_{B})}{p\sqrt{\delta}}}.$$

There are two reasons to prefer this form. First, the translation operator described in Lemma 2.2 becomes diagonal.

COROLLARY 3.1 Let N_B sources s_j lie in a box B with center s_B and side length $\sqrt{\delta}$ and let x be a target point in a box C with center x_C . Then the plane wave expansion

$$G(x) = \sum_{\beta < p} C_{\beta} e^{i\frac{K\beta \cdot (x_i - s_B)}{p\sqrt{\delta}}} + O(\epsilon),$$

can be expanded about x_C as

$$G(x) = \sum_{\beta < p} D_{\beta} e^{i \frac{K\beta \cdot (x_i - x_C)}{p\sqrt{\delta}}} + O(\epsilon).$$

The coefficients D_{β} are given by

$$D_{\beta} = C_{\beta} e^{i\frac{K\beta \cdot (x_C - s_B)}{p\sqrt{\delta}}}.$$
 (17)

In terms of matrix factorization, we have

$$\mathbf{G}_{\epsilon} = \mathbf{D}' \cdot \mathbf{E}' \cdot \mathbf{F}'. \tag{18}$$

In this formulation, the diagonal blocks of \mathbf{F}' and \mathbf{D}' are given by

$$\mathbf{F}'(j)_{n,m} = \left(\frac{K}{2p\sqrt{\pi}}\right)^{d/2} e^{\frac{-|\beta_n|K}{2p}} e^{-i\frac{K\beta_n \cdot (s_m - s_{B_j})}{p\sqrt{\delta}}}.$$

$$\mathbf{D}'(j)_{n,m} = \left(\frac{K}{2p\sqrt{\pi}}\right)^{d/2} e^{\frac{-|\beta_n|_K}{2p}} e^{i\frac{K\beta_m \cdot (x_n - x_{C_j})}{p\sqrt{\delta}}}.$$

As in the original algorithm, note that if the sources and targets coincide, then \mathbf{D}' is the adjoint of \mathbf{F}' . The nonzero entries $\mathbf{E}'(ij)$ are now diagonal matrices of dimension $p^d \times p^d$, with entries defined in (17). The net cost of all translations per target box is reduced from $O((2r+1)^d d p^{d+1})$ work to $O((2r+1)^d p^d)$ work.

The second (and more important) reason to prefer the new form is that the number of translations can be dramatically reduced. We describe the modification to the algorithm in the one-dimensional case. For this, imagine that we are sweeping across all boxes from left to right and that, at present, a target box C_j has accumulated all plane wave expansions from source boxes within its interaction region (Fig. 1(a)). The net expansion can be shifted to the center of C_{j+1} using Corollary 3.1. By adding in the contribution from the box marked by + and subtracting the contribution from the box marked by -, we have the correct plane wave expansion for box C_{j+1} (Fig. 1(b)). Thus, (2r+1) translations are replaced by three. In d-dimensions, the cost $O((2r+1)^d p^d)$ work can be reduced to $O(3d p^d)$, by sweeping across each dimension separately.

4 Conclusions

We have presented a new version of the fast Gauss transform, which uses plane wave expansions to diagonalize the translation of information between boxes. The

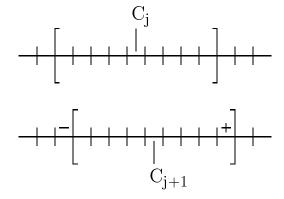


Figure 1: After shifting the expansion from box C_j to box C_{j+1} , one needs only to subtract the contribution from the box marked — and add the contribution from the box marked +. (The interaction regions are indicated by the square brackets).

approach is similar to the new diagonal forms used in fast multipole methods for the Laplace and Helmholtz equations [8, 6, 13]. When the present improvements have been incorporated into existing fast Gauss transform codes, the resulting scheme should provide a powerful kernel for one, two and three-dimensional calculations.

References

- [1] J. Carrier, L. Greengard, and V. Rokhlin, A Fast Adaptive Multipole Algorithm for Particle Simulations, Siam J. Sci. Stat. Comput., 9 (1988), pp. 669–686.
- [2] H. Dym and H. P. McKean, Fourier Series and Integrals, Academic Press, San Diego, 1972.
- [3] A. Friedman, Partial Differential Equations of Parabolic Type, Prentice-Hall, New Jersey, 1964.
- [4] S. Geman and C. Hwang, Nonparametric Maximum Likelihood Estimation by the Method of Sieves, Ann. Statist., 10 (1982), pp. 401–414.
- [5] L. Greengard, Fast algorithms for classical physics, Science, 265 (1994), pp. 909-xxx.
- [6] L. Greengard, J. Huang, V. Rokhlin and S. Wandzura, Accelerating fast multipole methods for low frequency scattering, CMCL Report 1998-003, New York University.
- [7] L. Greengard and V. Rokhlin, A fast algorithm for particle simulations, J. Comput. Phys., 73 (1987), pp. 325–348.

- [8] L. Greengard and V. Rokhlin, A new version of the fast multipole method for the Laplace equation in three dimensions, Acta Numerica, 6 (1987), pp. 229–269.
- [9] L. Greengard and J. Strain, The Fast Gauss Transform, SIAM J. Sci. Stat. Comput., 12 (1991), pp. 79–94.
- [10] L. Greengard and J. Strain, A Fast Algorithm for the Evaluation of Heat Potentials, Comm. on Pure and Appl. Math., 43 (1990), pp. 949–963.
- [11] R. B. Guenther and J. W. Lee, Partial Differential Equations of Mathematical Physics and Integral Equations, Prentice-Hall, Englewood Cliffs, New Jersey, 1988.
- [12] E. Hille, A Class of Reciprocal Functions, Ann. Math., 27 (1926), pp. 427–464.
- [13] T. Hrycak and V. Rokhlin, An improved fast multipole algorithm for potential fields, Department of Computer Science Research Report 1089, Yale University, 1995.
- [14] V. Rokhlin, Rapid solution of integral equations of classical potential theory,
 J. Comput. Phys. 60 (1985), pp. 187–207.
- [15] V. Rokhlin, Rapid solution of integral equations of scattering theory in two dimensions, J. Comput. Phys., 86 (1990), pp. 414-439.
- [16] V. Rokhlin, Diagonal forms of translation operators for the Helmholtz equation in three dimensions, Appl. and Comput. Harmonic Analysis 1 (1993), pp. 82–93.
- [17] B. W. Silverman, Density Estimation for Statistics and Data Analysis, Chapman and Hall, London, 1986.
- [18] J. Strain, The Fast Gauss Transform with variable scales, SIAM J. Sci. Stat. Comput. 12 (1991), pp. 1131–1139.

Leslie Greengard Courant Institute New York University New York, NY 10012-1110 USA greengard@cims.nyu.edu Xiaobai Sun Department of Computer Science Duke University Durham, NC 27708-0129 USA xiaobai@cs.duke.edu