Kernel Matrix Compression with Proxy Points

Xin Ye[1] Jianlin Xia[2] Lexing Ying[3]

[1]Department of Computer Science and Engineering
University of Minnesota

[2]Department of Mathematics
Purdue University

[3]Department of Mathematics
Stanford University

2018 Conference on Fast Direct Solvers
November 9, 2018

University of Minnesota
Driven to Discover℠

Outline

1. Introduction
   - Background
   - Review of compression methods
   - Proxy point method
2. Proxy point selection via contour integration
   - Model problem
   - Approximation error analysis
   - Optimal proxy points
3. Hybrid method
   - Dissect the proxy point method
   - Approximation error analysis

Kernel matrix compression

For a kernel function $k(x, y)$ and two well separated sets X and Y, find the low-rank approximation

$$\underset{(m\times n)}{K^{X,Y}} := \left(k(x_i, y_j)\right)_{x_i\in X, y_j\in Y} \approx \underset{(m\times r)}{U} \cdot \underset{(r\times n)}{V}$$

Kernel matrix compression

For a kernel function $k(x, y)$ and two well separated sets X and Y, find the low-rank approximation

$$\underset{(m\times n)}{K^{X,Y}} := (k(x_i, y_j))_{x_i\in X, y_j\in Y} \approx \underset{(m\times r)}{U} \cdot \underset{(r\times n)}{V}$$

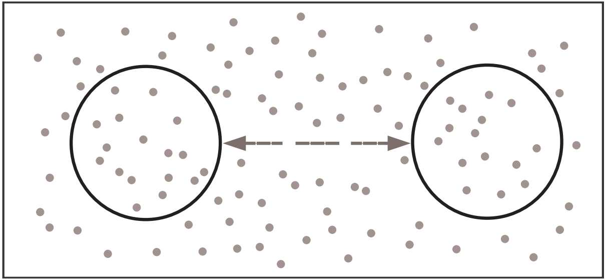
Where this problm often appears:

- Numerical solution to PDE/IE
- Cauchy/Toeplitz/Vandermonde systems
- Kernel method in machine learning
- N-body problem
- ...

Kernel matrix compression

For a kernel function $k(x, y)$ and two well separated sets X and Y, find the low-rank approximation

$$\underset{(m\times n)}{K^{X,Y}} := (k(x_i, y_j))_{x_i\in X, y_j\in Y} \approx \underset{(m\times r)}{U} \cdot \underset{(r\times n)}{V}$$



Different compression methods

- Algebraic method
  - Singular value decomposition (SVD)
  - Rank-revealing factorizations: SRRQR [Gu, Eisenstat 96], SRRLU [Miranian, Gu 03], ID [Cheng, et al. 05]. . .
  - Randomized compression [Frieze, et al. 04][Halko, et al. 11]

The algorithms deal with the matrix purely algebraically regardless of how it is generated.

Different compression methods

- Algebraic method
  - Singular value decomposition (SVD)
  - Rank-revealing factorizations: SRRQR [Gu, Eisenstat 96], SRRLU [Miranian, Gu 03], ID [Cheng, et al. 05]...
  - Randomized compression [Frieze, et al. 04][Halko, et al. 11]

  The algorithms deal with the matrix purely algebraically regardless of how it is generated.
- Analytical method
  - Multipole expansion [Greengard, Rokhlin 87]
  - Spherical harmonic expansion [Sun, Pitsianis 01]
  - Chebyshev interpolation [Fong, Darve 09]
  - Taylor expansion [Cai, Xia 16]
  - ...

  The resulting low-rank approximation usually lacks the structure preserving feature.

Proxy point method

To compress the kernel matrix $K^{X,Y}$

Proxy point method

To compress the kernel matrix $K^{X,Y}$

SRRQR/ID

$$K^{X,Y} \approx P \begin{pmatrix} I \\ E \end{pmatrix} K^{\tilde{X},Y} := UK^{\tilde{X},Y}$$

U column basis, $\tilde{X}$ representative points.

Proxy point method

To compress the kernel matrix $K^{X,Y}$

SRRQR/ID

$$K^{X,Y} \approx P \begin{pmatrix} I \\ E \end{pmatrix} K^{\tilde{X},Y} := UK^{\tilde{X},Y}$$

U column basis, $\tilde{X}$ representative points.

Proxy point method

1. Pick proxy surface Γ and proxy points $Z \subset \Gamma$
2. Compress $K^{X,Z}$ with SRRQR: $K^{X,Z} \approx UK^{\tilde{X},Z}$
3. Then $K^{X,Y} \approx UK^{\tilde{X},Y}$

Proxy point method

Appealing features:

- Fast and accurate
  $|Z|$ can be much smaller than $|Y|$ while still keep very small approximation error.
- Structure preserving
  Benefits hierarchical matrix techniques.

Proxy point method

Appealing features:

- Fast and accurate
  $|Z|$ can be much smaller than $|Y|$ while still keep very small approximation error.
- Structure preserving
  Benefits hierarchical matrix techniques.

Unanswered questions:

- Why can we use the proxy surface and proxy points?
  (In some cases, this can be answered by potential theory/Green's identity.)
- Where to pick them? How many?

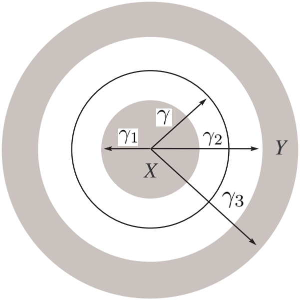
Model problem

- The kernel function is

$$k(x, y) = \frac{1}{(x - y)^d}, \quad d \in \mathbb{Z}^+.$$

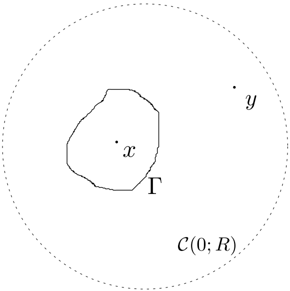
- Two sets of points satisfy

$$X = \{x_j\}_{j=1}^m \subset \mathcal{D}(0; \gamma_1), \quad Y = \{y_j\}_{j=1}^n \subset \mathcal{A}(0; \gamma_2, \gamma_3).$$



Introducing the proxy surface

For an $x \in X$ and $y \in Y$, draw a closed curve Γ between them.



We can show with Cauchy integral theorem:

$$k(x, y) = \frac{1}{2\pi i} \int_\Gamma \frac{k(x, z)}{y - z} \mathrm{d}z.$$

Introducing the proxy surface

With a quadrature rule $\{(z_j, \omega_j)\}_{j=1}^N$ on Γ:

$$k(x,y) \approx k_N(x,y) = \frac{1}{2\pi i}\sum_{j=1}^{N} \omega_j \frac{k(x,z_j)}{y-z_j} = \sum_{j=1}^{N} k(x,z_j)\frac{\omega_j}{2\pi i(y-z_j)}$$

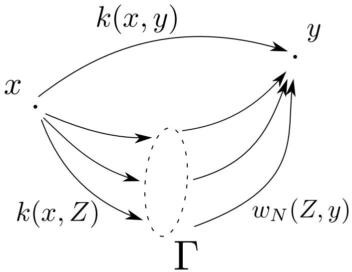
$$:= \sum_{j=1}^{N} k(x,z_j) w_N(z_j,y) = K^{x,Z} W_N^{Z,y}.$$

Introducing the proxy surface

With a quadrature rule $\{(z_j, \omega_j)\}_{j=1}^N$ on Γ:

$$k(x,y) \approx k_N(x,y) = \frac{1}{2\pi i} \sum_{j=1}^{N} \omega_j \frac{k(x,z_j)}{y-z_j} = \sum_{j=1}^{N} k(x,z_j) \frac{\omega_j}{2\pi i(y-z_j)}$$

$$:= \sum_{j=1}^{N} k(x,z_j) w_N(z_j,y) = K^{x,Z} W_N^{Z,y}.$$



Approximation error analysis

Assume $\Gamma = \mathcal{C}(0;\gamma)$ is a circle ($|x| < \gamma < |y|$) and the N-point composite trapezoidal rule is used, define

$$\varepsilon_N(x,y) = \left[k_N(x,y) - k(x,y)\right]/k(x,y)$$

Approximation error analysis

Assume $\Gamma = \mathcal{C}(0;\gamma)$ is a circle ($|x| < \gamma < |y|$) and the N-point composite trapezoidal rule is used, define

$$\varepsilon_N(x,y) = \left[k_N(x,y) - k(x,y)\right]/k(x,y)$$

Theorem (approximation error bound)

There exists an $N_1 > 0$ such that for any $N > N_1$, the error is bounded by

$$|\varepsilon_N(x,y)| \le \frac{1}{|y/\gamma|^N - 1} + \frac{C}{|\gamma/x|^N - 1}$$

where C is a constant dependent on N, d and $|y/x|$.

Note: N_1 is independent of γ.

Approximation error analysis

Assume $\Gamma = \mathcal{C}(0;\gamma)$ is a circle ($|x| < \gamma < |y|$) and the N-point composite trapezoidal rule is used, define

$$\varepsilon_N(x,y) = \left[k_N(x,y) - k(x,y)\right]/k(x,y)$$

Theorem (optimal γ)

If the error bound is viewed as a real function in γ on the interval $(|x|,|y|)$, then there exists $N_2 > 0$ such that if $N > N_2$,

1. the function has a unique minimizer γ^*,
2. the minimum decays as $\mathcal{O}(|y/x|^{-N/2})$.

Note: γ^* is dependent on N, d and $|y/x|$.

Approximation error analysis

Now for a block

$$K^{X,Y} \approx K_N^{X,Y} = K^{X,Z} W_N^{Z,Y},$$

note that all entry-wise results still hold if $|x|$ and $|y|$ are replaced by γ_1 and γ_2.

Approximation error analysis

Now for a block

$$K^{X,Y} \approx K_N^{X,Y} = K^{X,Z} W_N^{Z,Y},$$

note that all entry-wise results still hold if $|x|$ and $|y|$ are replaced by γ_1 and γ_2.

Corollary (block error bound)

With $\gamma \in (\gamma_1, \gamma_2)$, the F-norm relative approximation error is bounded by

$$\frac{\|K_N^{X,Y} - K^{X,Y}\|_F}{\|K^{X,Y}\|_F} \leq \frac{1}{(\gamma_2/\gamma)^N - 1} + \frac{C}{(\gamma/\gamma_1)^N - 1}$$

where C is as defined as before with $|y/x|$ replaced by γ_2/γ_1.

Similarly there exists an optimal γ^*.

Case 1: $d = 1$

In this case, the kernel function is $k(x, y) = 1/(x - y)$ which is associated with Toeplitz and Cauchy-like matrices.

Proposition

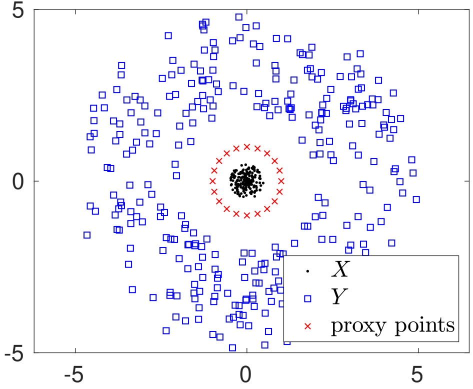
When $d = 1$, for any $N > 0$ and $\gamma \in (\gamma_1, \gamma_2)$, the approximation error is bounded by

$$\frac{\|K_N^{X,Y} - K^{X,Y}\|_F}{\|K^{X,Y}\|_F} \leq \frac{1}{(\gamma/\gamma_1)^N - 1} + \frac{1}{(\gamma_2/\gamma)^N - 1}.$$

If viewed as a function in γ, this upper bound has a unique minimizer $\gamma^* = \sqrt{\gamma_1 \gamma_2}$ and the optimal upper bound is $2/\left((\gamma_2/\gamma_1)^{N/2} - 1\right)$.

Case 1: $d = 1$

A simple numerical test: $m = 200$, $n = 300$, $\gamma_1 = 0.5$, $\gamma_2 = 2$ and $\gamma_3 = 5$, pick X and Y uniformly from their corresponding regions.



Case 1: $d = 1$

A simple numerical test: $m = 200$, $n = 300$, $\gamma_1 = 0.5$, $\gamma_2 = 2$ and $\gamma_3 = 5$, pick X and Y uniformly from their corresponding regions.

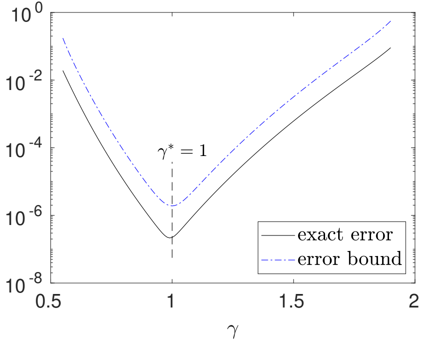


Figure: Varying γ.

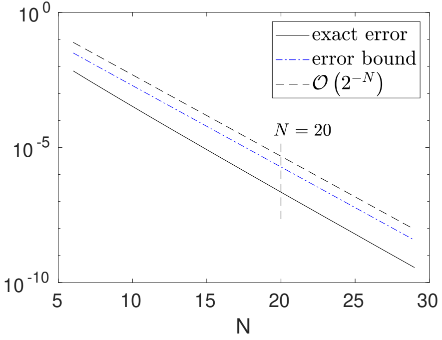


Figure: Varying N.

Case 2: $d > 1$

- Nothing explicit can be obtained in this case.

Case 2: $d > 1$

- Nothing explicit can be obtained in this case.
- We can turn to our previous theorems for help:

Case 2: $d > 1$

- Nothing explicit can be obtained in this case.
- We can turn to our previous theorems for help:
    - γ^* and the optimal bound are only dependent on N, d and γ_2/γ_1.

Case 2: $d > 1$

- Nothing explicit can be obtained in this case.
- We can turn to our previous theorems for help:
  - γ^* and the optimal bound are only dependent on N, d and γ_2/γ_1.
  - They are independent of m, n.

Case 2: $d > 1$

- Nothing explicit can be obtained in this case.
- We can turn to our previous theorems for help:
  - γ^* and the optimal bound are only dependent on N, d and γ_2/γ_1.
  - They are independent of m, n.
- Pick $X_0 \subset \mathcal{D}(0;\gamma_1)$ and $Y_0 \subset \mathcal{A}(0;\gamma_2,\gamma_3)$, then

$$E_N^0(\gamma) := \frac{\|K_N^{X_0,Y_0} - K^{X_0,Y_0}\|_F}{\|K^{X_0,Y_0}\|_F} \quad \text{and} \quad E_N(\gamma) := \frac{\|K_N^{X,Y} - K^{X,Y}\|_F}{\|K^{X,Y}\|_F}$$

  are expected to have similar behavior when γ varies in (γ_1,γ_2), thus $E_N^0(\gamma)$ can be used to approximate γ^*.

Case 2: $d > 1$

- Nothing explicit can be obtained in this case.
- We can turn to our previous theorems for help:
  - γ^* and the optimal bound are only dependent on N, d and γ_2/γ_1.
  - They are independent of m, n.
- Pick $X_0 \subset \mathcal{D}(0;\gamma_1)$ and $Y_0 \subset \mathcal{A}(0;\gamma_2,\gamma_3)$, then

$$E_N^0(\gamma) := \frac{\|K_N^{X_0,Y_0} - K^{X_0,Y_0}\|_F}{\|K^{X_0,Y_0}\|_F} \quad \text{and} \quad E_N(\gamma) := \frac{\|K_N^{X,Y} - K^{X,Y}\|_F}{\|K^{X,Y}\|_F}$$

  are expected to have similar behavior when γ varies in (γ_1,γ_2), thus $E_N^0(\gamma)$ can be used to approximate γ^*.
- Computing $E_N^0(\gamma)$ is cheap if $|X_0||Y_0|$ is small.

Case 2: $d > 1$

Numerical test:

- We set $|X_0| = |Y_0| = l$ and let $l = 1, 2, 3$.
- Always have $\gamma_1 \in X_0$ and $\gamma_2 \in Y_0$ ($x = \gamma_1$ and $y = \gamma_2$ correspond to the worst case of approximation error).

Case 2: $d > 1$

Numerical test:

- We set $|X_0| = |Y_0| = l$ and let $l = 1, 2, 3$.
- Always have $\gamma_1 \in X_0$ and $\gamma_2 \in Y_0$ ($x = \gamma_1$ and $y = \gamma_2$ correspond to the worst case of approximation error).

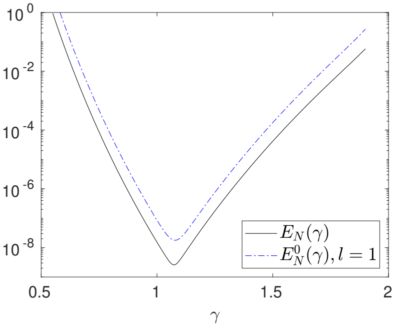


Figure: $d = 2$.

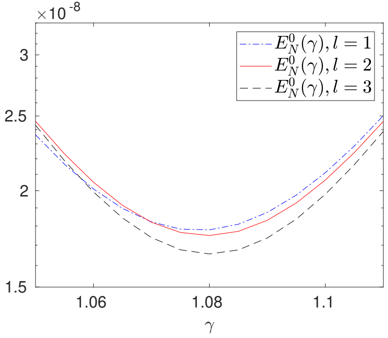


Figure: $d = 2$, zoom in at critical point.

Case 2: $d > 1$

Numerical test:

- We set $|X_0| = |Y_0| = l$ and let $l = 1, 2, 3$.
- Always have $\gamma_1 \in X_0$ and $\gamma_2 \in Y_0$ ($x = \gamma_1$ and $y = \gamma_2$ correspond to the worst case of approximation error).

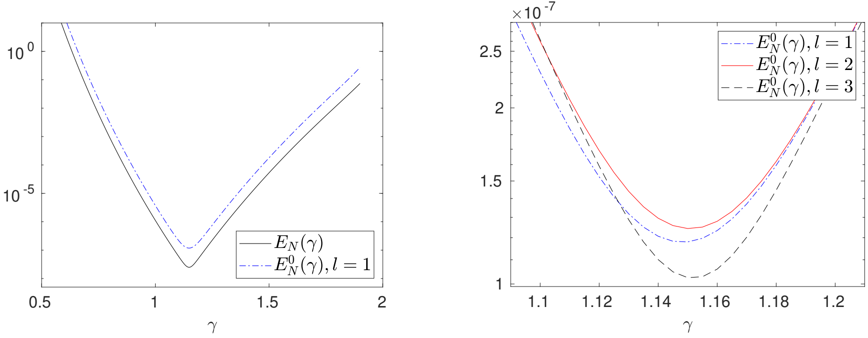


Figure: $d = 3$.

Figure: $d = 3$, zoom in at critical point.

Dissect the proxy point method

What we've got so far is an analytical compression method (CI) for a kernel matrix

$$K^{X,Y} \approx K_N^{X,Y} = K^{X,Z} W_N^{Z,Y}.$$

- Approximation error bounds.
- Optimal choose for γ^*.

Dissect the proxy point method

What we've got so far is an analytical compression method (CI) for a kernel matrix

$$K^{X,Y} \approx K_N^{X,Y} = K^{X,Z} W_N^{Z,Y}.$$

- Approximation error bounds.
- Optimal choose for γ^*.

Proxy point method can be viewed as a hybrid method by combining CI and ID:

$$\begin{aligned} K^{X,Y} &\approx K_N^{X,Y} = K^{X,Z} W_N^{Z,Y} && \text{(by CI on } K^{X,Y}\text{)},\\ &\approx U K^{\tilde{X},Z} W_N^{Z,Y} && \text{(by ID on } K^{X,Z}\text{)},\\ &= U K_N^{\tilde{X},Y} \approx U K^{\tilde{X},Y} && \text{(by CI on } K^{\tilde{X},Y}\text{)}. \end{aligned}$$

Approximation error bound

Theorem (error bound)

The compression error τ_{CI} for the analytical step is the optimal error bound, the relative tolerance (in F-norm) used in ID is τ_{ID} and the constant in SRRQR is $f > 1$ and the compression rank is $r < N$. Then a rank-r approximation of the kernel matrix $K^{X,Y}$ by the hybrid method satisfies

$$\|K^{X,Y} - UK^{\tilde{X},Y}\|_F \leq (C_{\mathrm{CI}}\tau_{\mathrm{CI}} + C_{\mathrm{ID}}\tau_{\mathrm{ID}})\,\|K^{X,Y}\|_F$$

where

$$C_{\mathrm{CI}} = 1 + \sqrt{r + (m-r)rf^2}\sqrt{1 - \frac{(m-r)(\gamma_2-\gamma_1)^{2d}}{m(\gamma_1+\gamma_3)^{2d}}},$$

$$C_{\mathrm{ID}} = \frac{\gamma^*(\gamma_1+\gamma_3)^d}{(\gamma_2-\gamma^*)(\gamma^*-\gamma_1)^d}.$$

Remarks

- The cost of the process is $\mathcal{O}(mNr)$.
- The compression accuracy can be conveniently controlled by this result.
  - In most cases, $C_{\mathrm{CI}} \sim \mathcal{O}(\sqrt{m})$ and $C_{\mathrm{ID}} \sim \mathcal{O}(1)$.

Remarks

- The cost of the process is $\mathcal{O}(mNr)$.
- The compression accuracy can be conveniently controlled by this result.
  - In most cases, $C_{\text{CI}} \sim \mathcal{O}(\sqrt{m})$ and $C_{\text{ID}} \sim \mathcal{O}(1)$.
- It explains some heuristics for proxy point method.
  - As long as the set Y is within the annulus region, the approximation error bound is independent of $|Y|$ or where they are.
  - $N = |Z|$ can be very small regardless of $|X|$ and $|Y|$. By our analysis, it is only dependent on γ_2/γ_1 (separation of two sets).

Conclusion

- We rigorously justified the use of proxy points via contour integration, presented the corresponding error analysis and discussed how to achieve optimal performance.
- Apply the results to proxy point method understood as a hybrid method, obtained a clear connection between the approximation error and how proxy points are picked.
- This can be applied to hierarchical techniques for certain types of matrices and potentially reduce the construction cost to be below linear.
- We are currently working on similar analysis for other kernels and geometries.

References

X. Ye, J. Xia, and L. Ying, Analytical compression via proxy point selection and contour integration, to be submitted, 2018.

Thank you!