THE UNIVERSITY OF QUEENSLAND

AUSTRALIA

Optimizing performance in Gaussian Processes

Michael Ciccotosto-Camp

Supervisor: Fred (Farbod) Roosta
Co-Supervisors: Andries Potgieter
Yan Zhao

Bachelor of Mathematics (Honours)
June 2022

The University of Queensland
School of Mathematics and Physics

Contents

Acknowledgements

I would like to deeply thank my supervisor Dr. Masoud Kamgarpour for his advice and all of his time spent with me. I consider myself lucky and am glad to have been his student for my honours year. I would also like to thank my co-supervisor Dr. Anna Puskás for the same reasons. A special thanks to Dr. Valentin Buciumas for his time spent teaching me while he was at The University of Queensland.

Symbols and Notation

Matrices are capitalized bold face letters while vectors are lowercase bold face letters.

Syntax	*Meaning*
$\triangleq$	An equality which acts as a statement
$\lvert\boldsymbol{A}\rvert$	The determinate of a matrix.
$\langle\cdot,\cdot\rangle_{\mathcal{H}}$	The inner product with respect to the Hilbert space $\mathcal{H}$, sometimes abbreviated as $\langle\cdot,\cdot\rangle$ if the Hilbert space is clear from context.
$\lVert\cdot\rVert_{\mathcal{V}}$	The norm of a vector with respect to the vector space $\mathcal{V}$, sometimes abbreviated as $\lVert\cdot\rVert$ if the vector space is clear from context.
$\boldsymbol{x}^{\mathsf{T}}, \boldsymbol{X}^{\mathsf{T}}$	The transpose operator.
$\boldsymbol{x}^{*}, \boldsymbol{X}^{*}$	The hermitian operator.
$\boldsymbol{a}. * \boldsymbol{b}$ or $\boldsymbol{A}. * \boldsymbol{B}$	Element-wise vector (matrix) multiplication, similar to Matlab.
$\propto$	Proportional to.
∇ or $\nabla_{\boldsymbol{f}}$	The partial derivative (with respect to $\boldsymbol{f}$).
∇	The Hessian.
$\sim$	Distributed according to, example $x \sim \mathcal{N}(0,1)$
$\mathbf{0}$ or $\mathbf{0}_n$ or $\mathbf{0}_{n\times m}$	The zero vector (matrix) of appropriate length (size) or the zero vector of length n or the zero matrix with dimensions $n \times m$.
$\mathbf{1}$ or $\mathbf{1}_n$ or $\mathbf{1}_{n\times m}$	The one vector (matrix) of appropriate length (size) or the one vector of length n or the one matrix with dimensions $n \times m$.
$\mathbb{1}_{n\times m}$	The matrix with ones along the diagonal and zeros on off diagonal elements.

$\boldsymbol{A}_{(\cdot,\cdot)}$	Index slicing to extract a submatrix from the elements of $\boldsymbol{A} \in \mathbb{R}^{n\times m}$, similar to indexing slicing from the python and Matlab programming languages. Each parameter can receive a single value or a 'slice' consisting of a start and an end value separated by a semicolon. The first and second parameter describe what row and columns should be selected, respectively. A single value means that only values from the single specified row/column should be selected. A slice tells us that all rows/columns between the provided range should be selected. Additionally if now start and end values are specified in the slice then all rows/columns should be selected. For example, the slice $\boldsymbol{A}_{(1:3,j:j')}$ is the submatrix $\mathbb{R}^{3\times(j'-j+1)}$ matrix containing the first three rows of $\boldsymbol{A}$ and columns j to j'. As another example, $\boldsymbol{A}_{(:,j)}$ is the j^{th} column of $\boldsymbol{A}$.
$\boldsymbol{A}^{\dagger}$	Denotes the unique psuedo inverse or Moore-Penore inverse of $\boldsymbol{A}$.
$\mathbb{C}$	The complex numbers.
C	The classes in a classification problem.
$\text{cholesky}\left(\boldsymbol{A}\right)$	A function to compute the Cholesky decomposition of the matrix $\boldsymbol{A}$, where $\boldsymbol{L}\boldsymbol{L}^{\intercal} = \boldsymbol{A}$.
$\text{cov}\left(\boldsymbol{f}\right)$	Gaussian process posterior covariance.
d	The number of features in the data set.
D	The dimension of the feature space of the feature mapping constructed in the Random Fourier Feature method.
$\mathcal{D}$	The dataset, $\mathcal{D} = \{(\boldsymbol{x}_i, y_i)\}_{i=1}^{n}$.
$\text{diag}\left(\boldsymbol{w}\right)$	Vector argument, a diagonal matrix containing the elements of vector $\boldsymbol{w}$.
$\text{diag}\left(\boldsymbol{W}\right)$	Matrix argument, a vector containing the diagonal elements of the matrix $\boldsymbol{W}$.
$\mathbb{E}$ or $\mathbb{E}_{q(x)}\left[z(x)\right]$	Expectation, or expectation of $z(x)$ where $x \sim q(x)$.
$\mathcal{GP}$	Gaussian process $f \sim \mathcal{GP}\left(m(\boldsymbol{x}), k\left(\boldsymbol{x}, \boldsymbol{x}'\right)\right)$, the function f is distributed as a Guassian process with mean function $m\left(\boldsymbol{x}\right)$ and covariance function $k\left(\boldsymbol{x}, \boldsymbol{x}'\right)$.

$k(\cdot,\cdot)$	A covariance or kernel matrix.
$\boldsymbol{K}_{\boldsymbol{W}\boldsymbol{W}'}$	For two data sets $\boldsymbol{W} = [\boldsymbol{w}_1, \boldsymbol{w}_2, \ldots, \boldsymbol{w}_n]^\intercal \in \mathbb{R}^{n \times d}$ and $\boldsymbol{W}' = [\boldsymbol{w}'_1, \boldsymbol{w}'_2, \ldots, \boldsymbol{w}'_m]^\intercal \in \mathbb{R}^{n' \times d}$ the matrix $\boldsymbol{K}_{\boldsymbol{W}\boldsymbol{W}'} \in \mathbb{R}^{n \times n'}$ has elements $(\boldsymbol{K}_{\boldsymbol{W}\boldsymbol{W}'})_{i,j} = k\left(\boldsymbol{w}_i, \boldsymbol{w}'_j\right)$.
lin-solve $(\boldsymbol{A}, \boldsymbol{B})$	A function used to solve $\boldsymbol{X} = \boldsymbol{A}^{-1}\boldsymbol{B}$ in the linear system $\boldsymbol{A}\boldsymbol{X} = \boldsymbol{B}$.
$\mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ or $\mathcal{N}(\boldsymbol{x} \mid \boldsymbol{\mu}, \boldsymbol{\Sigma})$	(the variable $\boldsymbol{x}$ has a) Multivariate Gaussian distribution with mean vector $\boldsymbol{\mu}$ and covariance $\boldsymbol{\Sigma}$.
n and n_*	The number of training (and tests) cases.
N	The dimenion of the feature space.
$\mathbb{N}$	The natural numbers, $\mathbb{N} = \{1, 2, 3, \ldots\}$.
$\mathcal{O}(\cdot)$	Big-O notation. If a function $f \in \mathcal{O}(g)$ then the absolute value of $f(x)$ is at most a positive multiple of $g(x)$ for all sufficiently large values of x.
$y \mid x$ and $p(x \mid y)$	A conditonal random variable y given x and its probability density.
$\boldsymbol{Q}, \boldsymbol{V}$	Typically used to denote a matrix with orthonormal structure.
$\mathbb{R}$	The real numbers.
$\operatorname{tr}(\boldsymbol{A})$	The trace of a matrix.
$\mathbb{V}$ or $\mathbb{V}_{q(x)}[z(x)]$	Variance, the variance of $z(x)$ when $x \sim q(x)$.
$\mathcal{X}$	Input space.
$\boldsymbol{X}$	The $n \times d$ matrix of training inputs.
$\boldsymbol{X}_*$	The $n_* \times d$ matrix of test inputs.
$\boldsymbol{x}_i$	The i^{th} training input.
$\mathbb{Z}$	The integers, $\mathbb{Z} = \{\ldots, -2, -1, 0, 1, 2, \ldots\}$.

1. The Nystrom Method

In chapter **??** we saw that GP regression and classification relied on a Gram matrix (see definition **??**) to produce predictions. Unfortunately, from a computational perspective, constructing the Gram matrix for a data set $\mathcal{D} = \{(\boldsymbol{x}_i, y_i)\}_{i=1}^{n}$ brings about a nasty bottle neck owed by the $\mathcal{O}\left(n^2\right)$ kernel evaluations. Even before the rise of ML, there has been a lot of research devoted to creating numerical methods that quickly construct a low rank approximation of large matrices, $\boldsymbol{A}$, which ordinarily are a computational burdened to build exactly. These methods are centered around the idea of capturing the columns space of the matrix that best describes the the action of $\boldsymbol{A}$ as an operator. For lack of a better explanation, Mahoney gives a fantastic summary of why the column space is of much importance in these approximation techniques

> *"To understand why sampling columns (or rows) from a matrix is of interest,recall that matrices are "about" their columns and rows that is, linear combinations are taken with respect to them; one all but understands a given matrix if one understands its column space, row space, and null spaces; and understanding the subspace structure of a matrix sheds a great deal of light on the linear transformation that the matrix represents."*
> [MWM11, page 13]

Moreover, this class of algorithms lend very nice forms when $\boldsymbol{A}$ possess positive definite structure, which is exactly the case for our Gram matrix.

1.1. **The Nystrom Method.**

Attempting to compute an entire kernel matrix can prove to be quite a computational headache, prompting us to seek estimative alternatives. The approximation techniques studied in this chapter have been spurred on by the John-Lindenstrauss lemma stated in lemma 1.

Lemma 1 (John-Lindenstrauss)**.** *Given $0 < \varepsilon < 0$, any set of n points, X, in a high dimensional Euclidean space can be embedded into a $\ell-$dimensional Euclidean space where $\ell = \mathcal{O}\left(\ln(n)\right)$ via some linear map $\boldsymbol{\Omega} \in \mathbb{R}^{n\times\ell}$ which satisfies*

$$(1-\varepsilon)\left\|\boldsymbol{u}-\boldsymbol{v}\right\|^2 \leq \left\|\boldsymbol{\Omega u}-\boldsymbol{\Omega v}\right\|^2 \leq \varepsilon\left\|\boldsymbol{u}-\boldsymbol{v}\right\|^2$$

for any $\boldsymbol{u}, \boldsymbol{v} \in X$ [MWM11, page 15].

The John-Lindenstrauss lemma tells us that $\boldsymbol{QQ}^*\boldsymbol{A}$ will serve as a good approximation to some matrix $\boldsymbol{A}$ where $\boldsymbol{QQ}^*$, in some sense, projects onto some rank-k subspace of $\boldsymbol{A}$'s column space. This is because if $\boldsymbol{QQ}^*$ closesly matches the behavior of $\boldsymbol{\Omega}$ from the lemma then the pair-wise distances between points before and after applying $\boldsymbol{QQ}^*$ should remain fairly similar. To state this a little more explicitly, for a matrix $\boldsymbol{A}$ and a positive error tolerance ε we seek a matrix $\boldsymbol{Q} \in \mathbb{R}^{n\times k_\varepsilon}$ with orthonormal columns such that

$$\left\|\boldsymbol{A}-\boldsymbol{QQ}^*\boldsymbol{A}\right\|_F \leq \varepsilon$$

which can be expressed a more short hand notation as

$$\boldsymbol{A} \simeq \boldsymbol{QQ}^*\boldsymbol{A}. \tag{1}$$

This is commonly called the *fixed precision approximation problem*. Although, to simplify algorithmic development, a value of k is specified in advanced (instead of ε, thus removing k's dependence on ε) which is instead given the name *fixed rank problem*. Within the fixed rank problem framework, when $\boldsymbol{A}$ is hermitian, the matrix $\boldsymbol{Q}\boldsymbol{Q}^*$ acts as a good projection for both the columns and row space of $\boldsymbol{A}$ so that we have both $\boldsymbol{A} \simeq \boldsymbol{Q}\boldsymbol{Q}^*\boldsymbol{A}$ and $\boldsymbol{A} \simeq \boldsymbol{A}\boldsymbol{Q}\boldsymbol{Q}^*$ so that

$$\boldsymbol{A} \simeq \boldsymbol{Q}\boldsymbol{Q}^* \left(\boldsymbol{A}\right) \simeq \boldsymbol{Q}\boldsymbol{Q}^*\boldsymbol{A}\boldsymbol{Q}\boldsymbol{Q}^*. \tag{2}$$

Furthermore, if $\boldsymbol{A}$ is positive semi-definite we can improve the quality of our approximation of our approximation at almost no additional cost [Hal11, page 32]. Using the approximation from 1

$$\begin{aligned}
\boldsymbol{A} &\simeq \boldsymbol{Q}\left(\boldsymbol{Q}^*\boldsymbol{A}\boldsymbol{Q}\right)\boldsymbol{Q}^* \\
&= \boldsymbol{Q}\left(\boldsymbol{Q}^*\boldsymbol{A}\boldsymbol{Q}\right)\left(\boldsymbol{Q}^*\boldsymbol{A}\boldsymbol{Q}\right)^{\dagger}\left(\boldsymbol{Q}^*\boldsymbol{A}\boldsymbol{Q}\right)\boldsymbol{Q}^* \\
&\simeq \left(\boldsymbol{A}\boldsymbol{Q}\right)\left(\boldsymbol{Q}^*\boldsymbol{A}\boldsymbol{Q}\right)^{\dagger}\left(\boldsymbol{Q}^*\boldsymbol{A}\right). \qquad (3)
\end{aligned}$$

This is known as the Nystrom method. A general Nystrom framework is presented in Algorithm TODO.

References

[Ras06] Carl Edward and Williams Rasmussen Christopher K. I, *Gaussian processes for machine learning / Carl Edward Rasmussen, Christopher K.I. Williams.*, Adaptive computation and machine learning, MIT Press, Cambridge, Mass., 2006 (eng).

[HHF73] H. Howard Frisinger, *Aristotle's legacy in meteorology*, Bulletin of the American Meteorological Society **54** (1973), no. 3, 198–204.

[Yul27] G. Udny Yule, *On a Method of Investigating Periodicities in Disturbed Series, with Special Reference to Wolfer's Sunspot Numbers*, Philosophical transactions of the Royal Society of London. Series A, Containing papers of a mathematical or physical character **226** (1927), no. 636-646, 267–298 (eng).

[Box08] George E. P. and Jenkins Box Gwilym M and Reinsel, *Time series analysis : forecasting and control / George E.P. Box, Gwilym M. Jenkins, Gregory C. Reinsel.*, 4th ed., Wiley series in probability and statistics, John Wiley, Hoboken, N.J., 2008 (eng).

[VdW19] Mark Van der Wilk, *Sparse Gaussian process approximations and applications*, University of Cambridge, 2019.

[Cao18] Yanshuai Cao, *Scaling Gaussian Processes*, University of Toronto (Canada), 2018.

[SD22] Matías and Estévez Salinero-Delgado José and Pipia, *Monitoring Cropland Phenology on Google Earth Engine Using Gaussian Process Regression*, Remote Sensing **14** (2022), no. 1, DOI 10.3390/rs14010146.

[Pot13] Andries and Lawson Potgieter Kenton and Huete, *Determining crop acreage estimates for specific winter crops using shape attributes from sequential MODIS imagery*, International Journal of Applied Earth Observation and Geoinformation **23** (2013), DOI 10.1016/j.jag.2012.09.009.

[Mur12] Kevin P. Murphy, *Machine learning : a probabilistic perspective / Kevin P. Murphy.*, Adaptive computation and machine learning, MIT Press, Cambridge, MA, 2012 (eng).

[Ber96] Z.G. Sheftel Berezansky G.F, *Functional analysis. Volume 1 / Y.M. Berezansky, Z.G. Sheftel, G.F. Us ; translated from the Russian by Peter V. Malyshev.*, 1st ed. 1996., Operator Theory: Advances and Applications, 85, Basel ; Boston ; Berlin : BirkhaIuser Verlag, Basel ; Boston ; Berlin, 1996 (eng).

[Tre97] Lloyd N. (Lloyd Nicholas) and Bau Trefethen David, *Numerical linear algebra / Lloyd N. Trefethen, David Bau.*, SIAM Society for Industrial and Applied Mathematics, Philadelphia, 1997 (eng).

[Dem97] James W Demmel, *Applied numerical linear algebra / James W. Demmel.*, Society for Industrial and Applied Mathematics, Philadelphia, Pa., 1997 (eng).

[Ste08] Ingo and Christmann Steinwart Andreas, *Support Vector Machines*, 1st ed. 2008., Information Science and Statistics, Springer New York, New York, NY, 2008 (eng).

[Ber03] Alain and Thomas-Agnan Berlinet Christine, *Reproducing Kernel Hilbert Spaces in Probability and Statistics*, Springer, SpringerLink (Online service), Boston, MA, 2003 (eng).

[Ste99] Michael L Stein, *Interpolation of Spatial Data Some Theory for Kriging / by Michael L. Stein.*, 1st ed. 1999., Springer Series in Statistics, Springer New York : Imprint: Springer, New York, NY, 1999 (eng).

[Bos92] Bernhard and Guyon Boser Isabelle and Vapnik, *A training algorithm for optimal margin classifiers*, Proceedings of the fifth annual workshop on computational learning theory, 1992, pp. 144–152 (eng).

[Cor95] Corinna Cortes, *Support-Vector Networks*, Machine learning **20** (1995), no. 3, 273 (eng).

[Kro14] Dirk P and C.C. Chan Kroese Joshua, *Statistical Modeling and Computation by Dirk P. Kroese, Joshua C.C. Chan.*, 1st ed. 2014., Springer New York : Imprint: Springer, New York, NY, 2014 (eng).

[Fle00] R Fletcher, *Practical Methods of Optimization*, John Wiley and Sons, Incorporated, New York, 2000 (eng).

[Bis06] Christopher M Bishop, *Pattern recognition and machine learning / Christopher M. Bishop.*, Information science and statistics, Springer, New York, 2006 (eng).

[Spi90] David J and Lauritzen Spiegelhalter Steffen L, *Sequential updating of conditional probabilities on directed graphical structures*, Networks **20** (1990), no. 5, 579–605.

[MWM11] Michael W. Mahoney, *Randomized algorithms for matrices and data*, CoRR **abs/1104.5557** (2011).

[Hal11] Nathan and Martinsson Halko Per-Gunnar and Tropp, *Finding structure with randomness: Probabilistic algorithms for constructing approximate matrix decompositions*, SIAM review **53** (2011), no. 2, 217–288.

[Pre92] William H. (William Henry) Press, *Numerical recipes in C : the art of scientific computing / William H. Press ...* [*et al.*], 2nd ed., Cambridge University Press, Cambridge, 1992 (eng).

[Wan] Guorong and Wei Wang Yimin and Qiao, *Generalized Inverses: Theory and Computations*, Developments in Mathematics, vol. 53, Springer Singapore, Singapore (eng).

[Gre97] Anne Greenbaum, *Iterative methods for solving linear systems Anne Greenbaum.*, Frontiers in applied mathematics ; 17, Society for Industrial and Applied Mathematics SIAM, 3600 Market Street, Floor 6, Philadelphia, PA 19104, Philadelphia, Pa., 1997 (eng).

[Cho07] Sou-Cheng (Terrya) Choi, *Iterative methods for singular linear equations and least -squares problems*, ProQuest Dissertations Publishing, 2007 (eng).

[CHO11] Sou-Cheng T and PAIGE CHOI Christopher C and SAUNDERS, *MINRES-QLP: A KRYLOV SUBSPACE METHOD FOR INDEFINITE OR SINGULAR SYMMETRIC SYSTEMS*, SIAM journal on scientific computing **33** (2011), no. 3-4, 1810–1836 (eng).

[Rah08] Ali and Recht Rahimi Benjamin, *Random Features for Large-Scale Kernel Machines*, Advances in Neural Information Processing Systems, 2008.

[Pot21] Andres and Wu Potapczynski Luhuan and Biderman, *Bias-Free Scalable Gaussian Processes via Randomized Truncations* (2021) (eng).

[Hah33] Hans Hahn, *S. Bochner, Vorlesungen über Fouriersche Integrale: Mathematik und ihre Anwendungen, Bd. 12.) Akad. Verlagsges., Leipzig 1932, VIII. u. 229S. Preis brosch. RM 14,40, geb. RM16*, Monatshefte für Mathematik **40** (1933), no. 1, A27–A27 (ger).

[Liu21] Fanghui and Huang Liu Xiaolin and Chen, *Random Features for Kernel Approximation: A Survey on Algorithms, Theory, and Beyond*, IEEE transactions on pattern analysis and machine intelligence **PP** (2021) (eng).

[HAe16] Haim Avron etal, *Quasi-Monte Carlo Feature Maps for Shift-Invariant Kernels*, Journal of Machine Learning Research **17** (2016), no. 120, 1-38.

[DJSaJS15] Danica J. Sutherland and Jeff Schneider, *On the Error of Random Fourier Features*, 2015.

[Yu16] Felix X and Suresh Yu Ananda Theertha and Choromanski, *Orthogonal Random Features* (2016) (eng).

[Bro91] Peter J and Davis Brockwell Richard A, *Time Series: Theory and Methods*, Second Edition., Springer Series in Statistics, Springer New York, SpringerLink (Online service), New York, NY, 1991 (eng).

[Cho17] Krzysztof and Rowland Choromanski Mark and Weller, *The Unreasonable Effectiveness of Structured Random Orthogonal Embeddings* (2017) (eng).

[FaA76] Fino and Algazi, *Unified Matrix Treatment of the Fast Walsh-Hadamard Transform*, IEEE transactions on computers **C-25** (1976), no. 11, 1142–1146 (eng).

[And15] Alexandr and Indyk Andoni Piotr and Laarhoven, *Practical and Optimal LSH for Angular Distance* (2015) (eng).

[Cho20] Krzysztof and Likhosherstov Choromanski Valerii and Dohan, *Rethinking Attention with Performers* (2020) (eng).

[Boj16] Mariusz and Choromanska Bojarski Anna and Choromanski, *Structured adaptive and random spinners for fast machine learning computations* (2016) (eng).