Skoltech
Skolkovo Institute of Science and Technology

Jury Member Report – Doctor of Philosophy thesis.

Name of Candidate: Vladimir Fanaskov
PhD Program: Mathematics and Mechanics
Title of Thesis: Statistical inference and machine learning in numerical linear algebra
Supervisor: Associate Professor Aslan Kasimov

Name of the Reviewer: Prof. Evgeny Burnaev

I confirm the absence of any conflict of interest	**Date:** 29 August 2022

Reviewer's Report
Vladimir's thesis is on the development and applications of machine learning and Bayesian methods to problems of numerical linear algebra. The main focus is on the solution of large sparse linear systems via iterative techniques. The typical source of such systems is local (e.g., finite difference, finite element, finite volume) PDE discretization, so the choice of the systems to solve arises from practical engineering problems. Since any iterative method is essentially a consistent and convergent fixed-point iteration, the work can be alternatively described as a study of how contraction maps that approximate the solution to the linear system can be constructed using machine learning and Bayesian statistics. To achieve this goal the author considers a diverse set of techniques including belief propagation algorithm (Chapter 2), hierarchical Bayesian modeling (Chapter 3), variational approximation (Chapter 4), unsupervised learning (Chapters 6 and 7), convolutional neural networks (Chapter 7), reinforcement learning (Chapter 8). New and improved algorithms for the solutions of linear systems are proposed and benchmarked against the state-of-the-art techniques including preconditioned Krylov subspace methods and geometric multigrid. The main text is split into two parts and an appendix. The first and the second parts are on Bayesian statistics and machine learning methods respectively. Alternatively, the first part is on the analytic construction of fixed points, whereas the second part is on the numerical construction, i.e., via learning and/or optimization. The appendix contains the proofs of the results given in the main text as well as supplementary information on iterative methods and PDE used to showcase the performance.

The first part contains three chapters on belief propagation, probabilistic numerical approaches, and "hidden representation" (this term is introduced by the author of the thesis). Below I describe the content and original contributions of these three chapters.

The chapter on Belief propagation contains an explanation of how Pearl's belief propagation applied to the Gaussian Markov random field can be generalized to non-symmetric linear systems. To achieve this author explores the connection between walks on the graph and the solution to the linear system given by the convergent Neumann series. He also proposes a block version of the same algorithm using generalized belief propagation and considers both iterative techniques in the context of the multigrid method. For all algorithms, sufficient condition for convergence and consistency is established. The presented results appear to be on par or better than the state-of-the-art relaxation techniques (Jacobi iteration, color relaxation schemes, ILU, polynomial relaxations) as shown on the array of test problems.

The chapter on probabilistic numerical approaches is mostly on uncertainty quantification. The author starts with the general description of the reinterpretation of Krylov subspace methods as a statistical inference that was proposed in prior works and shows how to construct prior distribution to have a posterior distribution with a meaningful covariance matrix (i.e., well-calibrated uncertainty). To achieve this, Vladimir employs a hierarchical Bayesian model and empirical uncertainty calibration. It is shown in this part, that based on statistics (uncertainty quality measures) proposed in previous works, the current approach comes with much better uncertainty calibration both for conjugate gradient (symmetric case) and GMRES (symmetric and non-symmetric cases). In addition, it is shown how probabilistic projection methods can be applied to PDE-constrained optimization problems.
In the last chapter of the first part author described his own more general approach to uncertainty calibration that is more flexible than the current probabilistic projection techniques described in the previous chapter. In short, this approach is based on the so-called indifference principle to assign probabilities combined with symmetry transformations.

First, for a given numerical algorithm one needs to find transformations that leave the exact solution invariant but perturb the approximate solution. The author shows that it is easy to come up with these transformations for many numerical algorithms including interpolation, differentiation, finding dominant eigenvalue, and solution of ODEs. After the transformations are found, one can use the indifference principle to assign probabilities and use the resulting uncertainty to characterize the error of the method in probabilistic terms. The whole approach is developed for the case of linear systems with the use of multivariate normal models and variational inference. Empirical results are also given, including the algorithm that speeds up classical iterations using well-calibrated uncertainty.
The second part contains chapters on a generalization of BPX, neural multigrid architectures, and the use of reinforcement learning for the online optimization of iterative methods.

In Chapter 6 Vladimir describes how to construct multilevel preconditioners using unsupervised learning. He introduces a general black-box optimization technique that leverages stochastic trace approximation of Gelfand's formula with Richardson iterations. This allows to sidestep estimation of lower eigenvalue and at the same time is equivalent to the optimization of the condition number of the preconditioned system. It is shown empirically that the new loss function is better than the usual one used in optimal circulant preconditioners which leads to the increase in the condition number after training. The scheme is applied to generalized BPX preconditioners proposed by the author. The chapter is concluded by numerical benchmarks on a large set of physically-relevant linear systems. Results indicate that learning of optimal preconditioners is feasible and leads to improved condition numbers.

The next chapter is on similar black-box techniques but for multigrid solvers. This time the loss is a stochastic trace approximation of the spectral radius of the error propagation matrix. This kind of optimization is not new per se, but the author introduced several novelties. First, Vladimir proposed a new convolution-based architecture that seamlessly blends U-net operator-free neural networks with multigrid solvers. The main problem is that current ML frameworks work poorly with sparse matrices especially when they need to be recomputed on coarse grids during each iteration. Since the new architecture completely avoids recomputing sparse matrices, it leads to an order of magnitude faster training and allows for more flexibility in the sizes of projection and relaxation stencils. Second, Vladimir explains how serialization can be used to achieve practical architectures that generalize on finer meshes after training on a coarse grid.

Chapter 8 is on how to use k-armed bandits to accelerate iterative methods. Here the author motivates the problem when the linear system needs to be solved repeatedly with the different right-hand sides. This problem can be used for online optimization with reinforcement learning. However, it is shown that naive applications of algorithms such as epsilon-greedy k-armed bandit fall short to achieve optimal relaxation parameters. Vladimir replaces this algorithm with the variant that uses restarted power iterations and shows that it has much better performance. The chapter contains a number of numerical experiments demonstrating the feasibility of the approach, but lacks analytical results on convergence or its rate.

Overall I found the thesis to be well-structured and to contain many novel ideas. The results are theoretically and/or numerically justified, and most of them are published in several high-ranking journals such as SIAM Journal on Scientific Computing, Statistics and Computing, Journal of Computational and Applied Mathematics and reported at two leading international conferences International Joint Conference on Neural Networks, SIAM Conference on Applied Linear Algebra. The practical applications of the proposed algorithms are mainly to PDEs from continuum mechanics and are considered throughout the text.

A few minor issues I suggest addressing before the defense are as follows.

First, the belief propagation for nonsymmetrical systems in the second chapter lacks statistical interpretation. Surely, it is impossible to construct a normal Markov random field with a non-symmetric matrix. I suggest adding a brief discussion of the interpretation or the lack of it.

Second, Chapter 7 is a little light on considered architectures. It seems that the main advantage of the method is flexibility, so I suggest performing more experiments for architectures with more aggressive coarsening and larger interpolation/relaxation stencils.

Third, Chapter 8 contains incomplete results and lacks theoretical justifications. I understand that in general, it is a difficult task to provide guarantees for online optimization of the solver for a sufficiently general linear system, but I suggest the author give a short discussion with possible lines of attack on the problem, or some back-of-the-envelope estimations, or heuristics explanations on why algorithms are going to converge. The later ones are present to some extent but I believe it is better to have them gathered in a separate section.

Provisional Recommendation

I recommend that the candidate should defend the thesis by means of a formal thesis defense.