

Fast Linear Convergence of Randomized BFGS

Peter Richtárik



Beyond First Order Methods in ML Systems

ICML 2020

July 17, 2020



Dmitry Kovalev
PhD Student



Robert M. Gower
Assistant Professor



Alexander Rogozin
MS Student



Fast Linear Convergence of Randomized BFGS

Dmitry Kovalev¹, Robert M. Gower², Peter Richtárik¹, and Alexander Rogozin^{1,3}

¹King Abdullah University of Science and Technology (KAUST), Thuwal, Saudi Arabia

²Facebook AI Research and Télécom Paris, Paris, France *

³Moscow Institute of Physics and Technology, Dolgoprudny, Russia

February 9, 2020

(Revised: February 26, 2020)

Abstract

Since the late 1950's when quasi-Newton methods first appeared, they have become one of the most widely used and efficient algorithmic paradigms for unconstrained optimization. Despite their immense practical success, there is little theory that shows why these methods are so efficient. We provide a semi-local rate of convergence for the randomized BFGS method which can be significantly better than that of gradient descent, finally giving theoretical evidence supporting the superior empirical performance of the method.

Contents

1 Introduction	2
1.1 Background	3
1.2 Notation and definitions	3
2 Convergence Results	4
2.1 Local linear convergence for self-concordant functions	4
2.2 Local linear convergence for smooth and strongly convex functions	5
2.3 Superlinear convergence	5
3 Examples and Applications	6
3.1 Invertible \mathbf{S}	6
3.2 One column \mathbf{S}	6
3.3 Generalized linear models	6
3.4 Linear programs with box constraints	8
4 Numerical Experiments	9
4.1 Synthetic quadratic problem	9
4.2 Classification problem on real data	10
5 Conclusion, Consequences and Future Work	14
A Proof of Theorem 2.1	17
A.1 Properties of self-concordant functions	17
A.2 The distance of the iterates	18
A.3 The distance of the quasi-Newton matrix	19
A.4 Detailed proof of Theorem 2.1	21
B Proof of Theorem 2.5	21
B.1 Getting ready for the proof	22
B.2 The Proof	24
C Proof of Theorem 2.6	25

*Currently on leave from Télécom Paris



Dmitry Kovalev, Robert M. Gower, P. R. and Alexander Rogozin

Fast Linear Convergence of Randomized BFGS

arXiv:2002.11337, February 2020

Well behaved:

1. Self-concordant
2. Strongly convex, Lipschitz gradient, Lipschitz Hessian

$\min_{x \in \mathbb{R}^d}$

$f(x)$

$\stackrel{\text{def}}{=}$

$$\frac{1}{n} \sum_{i=1}^n f_i(x)$$

parameters / features
(Allowed to be large)

training data points
(Assumed to be of reasonable size)

The Problem

The Problem

$$\min_{x \in \mathbb{R}^d} f(x)$$

Detour: Randomized 2nd Order Methods in the Big n Regime



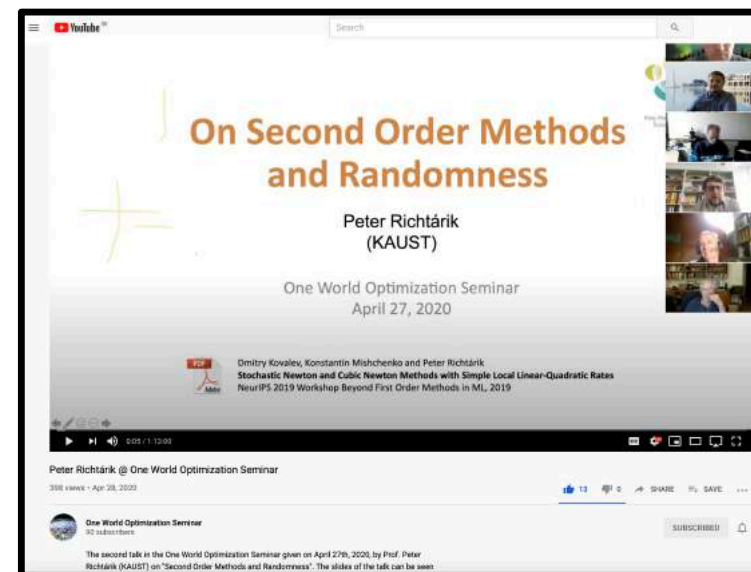
Dmitry Kovalev, Konstantin Mishchenko and P.R.

Stochastic Newton and cubic Newton methods with simple local linear-quadratic rates

NeurIPS 2019 Workshop: Beyond First Order Methods in ML (arXiv:1912.01597, 2019)

First “2nd Order SGD” method which works even when sampling 1 datapoint in each iteration

Unlike all first order methods, enjoys (local) linear rate **independent of condition number!**



April 27, 2020 talk at the “One World Optimization Seminar”

Talk Outline

I. Quasi-Newton Methods

II. Randomized BFGS for Matrix Inversion



Robert M. Gower and P. R.

Randomized Quasi-Newton Updates are Linearly Convergent Matrix Inversion Algorithms

SIAM Journal on Matrix Analysis and Applications 38(4):1380-1409, 2017

III. Randomized BFGS for Optimization



Dmitry Kovalev, Robert M. Gower, P. R. and Alexander Rogozin

Fast Linear Convergence of Randomized BFGS

arXiv:2002.11337, February 2020



Part I

Quasi-Newton Methods

From Gradient Descent to Newton's Method

$$f(x_k + h) \approx T_k(h) \stackrel{\text{def}}{=} f(x_k) + \langle \nabla f(x_k), h \rangle + \frac{1}{2} \langle \mathbf{B}_k^{-1} h, h \rangle$$

$\mathbf{B}_k^{-1} \approx \nabla^2 f(x_k)$

$$x_{k+1} = x_k + \arg \min_{h \in \mathbb{R}^d} T_k(h) = x_k - \mathbf{B}_k \nabla f(x_k)$$

Better approximation
of the Hessian



$$\mathbf{B}_k = \alpha_k \mathbf{I}$$

Gradient Descent

$$\mathbf{B}_k \approx (\nabla^2 f(x_k))^{-1}$$

Quasi-Newton Methods

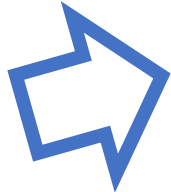
$$\mathbf{B}_k = (\nabla^2 f(x_k))^{-1}$$

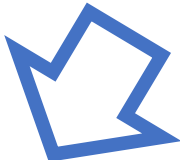
Newton's Method

Quasi-Newton Methods: Secant Equation for Convex Quadratics

$$f(x) = \frac{1}{2}x^\top \mathbf{H}x + b^\top x + c$$

$\mathbf{H} \succ 0$


$$\begin{aligned}\nabla f(x) &= \mathbf{H}x + b \\ \nabla^2 f(x) &\equiv \mathbf{H}\end{aligned}$$


$$\nabla^2 f(x)(u - v) = \nabla f(u) - \nabla f(v)$$

$$u - v = (\nabla^2 f(x))^{-1} (\nabla f(u) - \nabla f(v))$$

$$(\nabla f(u) - \nabla f(v))^\top (\nabla^2 f(x))^{-1} = (u - v)^\top$$

$$(\nabla f(u) - \nabla f(v))^{\top} (\nabla^2 f(x))^{-1} = (u - v)^{\top}$$

$$u = x_{k+1}, v = x_k, x = x_{k+1}$$

Secant Equation

$$(\nabla f(x_{k+1}) - \nabla f(x_k))^{\top} \mathbf{B}_{k+1} = (x_{k+1} - x_k)^{\top}$$

Diagram illustrating the Secant Equation:

- Known vector** (yellow box) points to $(\nabla f(x_{k+1}) - \nabla f(x_k))^{\top}$.
- Unknown matrix** (green box) points to \mathbf{B}_{k+1} . Below it, it states: "Desire: $\mathbf{B}_{k+1} \approx (\nabla^2 f(x_{k+1}))^{-1}$ ".
- Known vector** (orange box) points to $(x_{k+1} - x_k)^{\top}$.

This can be seen as a system of linear equations with the unknown \mathbf{B}_{k+1}

Generally, there will be multiple solutions. Which one to choose?

“Solving” the Secant Equation

Closest matrix to the current one satisfying the secant equation and symmetry

\mathbf{B}_{k+1}

$$= \arg \min_{\mathbf{B} \in \mathbb{R}^{d \times d}} \|\mathbf{B} - \mathbf{B}_k\|_{F(\mathbf{W})}$$

Weighted Frobenius norm

$$\|\mathbf{X}\|_{F(\mathbf{W})} \stackrel{\text{def}}{=} \|\mathbf{W}^{1/2} \mathbf{X} \mathbf{W}^{1/2}\|_F$$
$$\mathbf{W} \succ 0$$

subject to

$$(\nabla f(x_{k+1}) - \nabla f(x_k))^{\top} \mathbf{B} = (x_{k+1} - x_k)^{\top}$$

$$\mathbf{B} = \mathbf{B}^{\top}$$

Enforcing symmetry (optional)

Secant equation

Broyden-Fletcher-Goldfarb-Shanno (1970)

J. Inst. Maths Applies (1970) 6, 76-90

The Convergence of a Class of Double-rank Minimization Algorithms

1. General Considerations

C. G. BROYDEN
Computing Centre, University of Essex,
Wivenhoe Park, Colchester, Essex

[Received 7 March 1969 and in revised form 19 May 1969]

This paper presents a more detailed analysis of a class of minimization algorithms, which includes as a special case the DFP (Davidon-Fletcher-Powell) method, than has previously appeared. Only quadratic functions are considered but particular attention is paid to the magnitude of successive errors and their dependence upon the initial matrix. On the basis of this a possible explanation of some of the observed characteristics of the class is tentatively suggested.

1. Introduction

PROBABLY the best-known algorithm for determining the unconstrained minimum of a function of many variables, where explicit expressions are available for the first partial derivatives, is that of Davidon (1959) as modified by Fletcher & Powell (1963). This algorithm has many virtues. It is simple and does not require at any stage the solution of linear equations. It minimizes a quadratic function exactly in a finite number of steps and this property makes convergence of this algorithm rapid, when applied to more general functions, in the neighbourhood of the solution. It is, at least in theory, stable since the iteration matrix H_k , which transforms the i th gradient into the i th step direction, may be shown to be positive definite.

In practice the algorithm has been generally successful, but it has exhibited some puzzling behaviour. Broyden (1967) noted that H_k does not always remain positive definite, and attributed this to rounding errors. Pearson (1968) found that for some problems the solution was obtained more efficiently if H_1 was reset to a positive definite matrix, often the unit matrix, at intervals during the computation. Bard (1968) noted that H_k could become singular, attributed this to rounding error and suggested the use of suitably chosen scaling factors as a remedy.

In this paper we analyse the more general algorithm given by Broyden (1967), of which the DFP algorithm is a special case, and determine how for quadratic functions the choice of an arbitrary parameter affects convergence. We investigate how the successive errors depend, again for quadratic functions, upon the initial choice of iteration matrix paying particular attention to the cases where this is either the unit matrix or a good approximation to the inverse Hessian. We finally give a tentative explanation of some of the observed experimental behaviour in the case where the function to be minimized is not quadratic.

2. Basic Theory

Define a quadratic function $F(x)$ by

$$F(x) = \frac{1}{2}x^T Ax - b^T x + c, \quad (2.1)$$

76

A new approach to variable metric algorithms

R. Fletcher

Atomic Energy Research Establishment, Harwell, Didcot, Berkshire

An approach to variable metric algorithms has been investigated in which the linear search sub-problem is no longer necessary. The property of quadratic termination has been explained in one of monotonic convergence of the eigenvalues of the approximating matrix to the inverse Hessian. A covers class of updating formulae which possess this property has been established, and a strategy has been indicated for choosing a member of the class so as to keep the approximation away from both singularity and suboptimality. A FORTRAN program has been tested extensively with encouraging results.
(Received October 1969)

1. Motivation

This paper deals with the problem of minimizing a function $F(x)$ of n variables $x^T = (x_1, x_2, \dots, x_n)^T$ assuming that the gradient vector $\nabla_x F = g(x)$ is available explicitly, but that the Hessian G is not ($G_{ij} = \partial^2 F / \partial x_i \partial x_j$). Superscript T is used to denote transposition. A type of method which has achieved considerable success in solving this problem is the variable metric method (VMM) due to Davidon (1959), and simplified by Fletcher and Powell (1963). The main feature of the VMM is that an approximation H of G^{-1} is kept, and is updated at each iteration using the formula

$$H^* = H + \frac{\delta\delta^T}{\delta^T\gamma} - \frac{H\gamma\gamma^T H}{\gamma^T H \gamma} \quad (1)$$

where $\delta = x^* - x$ and $\gamma = g^* - g$ are the changes in x and g made on that iteration, and superscript T denotes values appropriate to the next iteration. The correction δ is taken as a multiple α of a 'direction of search' $s = -Hg$ chosen by analogy with Newton's method, so that

$$\delta = \alpha s = -\alpha Hg. \quad (2)$$

The multiple α is taken as the value of λ which minimizes $F(x + \lambda s)$, that is the function is minimized locally along the direction of search. The method has a number of important properties, for instance if the approximating matrix H is initially chosen to be positive definite, then this property is retained by subsequent approximations. Also if the function to be minimized is a positive definite quadratic form, then the iteration terminates in at most n iterations. Moreover Powell (1968) report, to be published but recently produced a convergence proof for a more general class of functions.

The algorithm has, however, some inconvenient features. The main one is the need to solve the sub-problem of finding α at each iteration (the 'linear search'). This is usually done by evaluating the function and gradient for a number of different values of λ and interpolating according to some strategy, until a sufficiently accurate minimum is obtained. Thus a considerable extra computing effort is required, above that for

calculating γ and updating H . (Computing effort is most readily measured by the number of times F and g have to be evaluated.) A further disadvantage is that the linear search is hazardous to program because of the many special circumstances which can arise (for instance the minimum may not exist at all). This can lead to worst to understated program errors; at best to a proliferation of different programs for implementing the VMM, giving rise to incompatibility in results. The linear search can also often be a disadvantage when constraints are present, because then the minimum along the line may not be feasible, even though no constraints limit the position of the ultimate solution. In this context, the flexibility of being able to generate directions of search, other than by $s = -Hg$, would also be convenient.

It is important therefore to consider whether the linear search subproblem can be dispensed with. The importance of the linear search is that it furnishes a property which enables finite termination to be proved for quadratic functions. The first point to examine therefore is whether this termination can be proved for variable metric algorithms not requiring linear searches, and based upon updating formulae other than (1). Now quadratic termination can be proved by requiring that the successive matrices H satisfy a 'hereditary property' when the function is quadratic: that is not only must H^* satisfy $H^* \gamma = \delta$ (a natural property because $G^{-1} \gamma = \delta$), but also $H^* \delta = \delta$ where γ and δ are a pair of differences from an earlier iteration. It is quite easy to show that there is only one formula for which hereditary properties can be proved without relying upon linear searches, and for which the correction is of rank 2 in the space of δ and γ . This formula is

$$H^* = H + \frac{(\delta - H\gamma)(\delta - H\gamma)^T}{\gamma^T(\delta - H\gamma)} \quad (3)$$

in which the correction has degenerated to be of rank 1, and which has attracted a lot of attention in recent years. The formula was discovered by a number of workers, a list of references being given by Powell (1969). Although the formula does remove the need to solve the linear

A Family of Variable-Metric Methods Derived by Variational Means

By Donald Goldfarb

Abstract. A new rank-two variable-metric method is derived using Greenstadt's variational approach [Math. Comp., this issue]. Like the Davidon-Fletcher-Powell (DFP) variable-metric method, the new method preserves the positive-definiteness of the approximating matrix. Together with Greenstadt's method, the new method gives rise to a one-parameter family of variable-metric methods that includes the DFP and rank-one methods as special cases. It is equivalent to Broyden's one-parameter family [Math. Comp., v. 21, 1967, pp. 368-381]. Choices for the inverse of the weighting matrix in the variational approach are given that lead to the derivation of the DFP and rank-one methods directly.

In the preceding paper [6], Greenstadt derives two variable-metric methods, using a classical variational approach. Specifically, two iterative formulas are developed for updating the matrix H_k (i.e., the inverse of the variable metric), where H_k is an approximation to the inverse Hessian $G^{-1}(x_k)$ of the function being minimized.*

Using the iteration formula

$$H_{k+1} = H_k + E_k$$

to provide revised estimates to the inverse Hessian at each step, Greenstadt solves for the correction term E_k that minimizes the norm

$$N(E_k) = \text{Tr}(WE_kWE_k^T)$$

subject to the conditions

$$E_k^T = E_k$$

and

(2)

$$E_k y_k = g_k - H_k y_k.$$

W is a positive-definite symmetric matrix and Tr denotes the trace.

The first condition is a symmetry condition which ensures that all iterates H_k will be symmetric as long as the initial estimate H_1 is chosen to be symmetric. The second condition ensures that the updated matrix H_{k+1} satisfies the equation

$$H_{k+1} y_k = g_k$$

and hence, that the method is of the "quasi-Newton" type [1].

Received June 30, 1969, revised August 4, 1969.

AMS Subject Classifications. Primary 30, Secondary 10.

Key Words and Phrases. Unconstrained optimization, variable-metric, variational methods, Davidon method, rank-one formulae.

*The reader is referred to Greenstadt's paper [6] for a more detailed discussion of variable-metric methods and for definitions of some of the terms used here.

MATHEMATICS OF COMPUTATION, VOLUME 24, NUMBER 111, JULY, 1970

Conditioning of Quasi-Newton Methods for Function Minimization

By D. F. Shanno

Abstract. Quasi-Newton methods accelerate the steepest-descent technique for function minimization by using computational history to generate a sequence of approximations to the inverse of the Hessian matrix. This paper presents a class of approximating matrices as a function of a scalar parameter. The problem of optimal conditioning of these matrices under an appropriate norm as a function of the scalar parameter is investigated. A set of computational results verifies the superiority of the new methods arising from conditioning considerations to known methods.

1. Introduction. Newton's method for minimizing a function $f(x)$, x an n -vector, is to generate a sequence of points,

$$(1) \quad x^{(k+1)} = x^{(k)} - \alpha^{(k)} [J^{(k)}]^{-1} g^{(k)},$$

where $g^{(k)} = \nabla f(x^{(k)})$, $J^{(k)} = [\partial^2 f / \partial x_i \partial x_j]$ the Hessian matrix of f evaluated at $x^{(k)}$, and $\alpha^{(k)}$ is an appropriately chosen scalar. Quasi-Newton methods use an initial estimate and computational history to generate an estimate $H^{(k)}$ to $[J^{(k)}]^{-1}$ at each step rather than performing the computational work of evaluating and inverting $J^{(k)}$. The sequence (1) then becomes

$$(2) \quad x^{(k+1)} = x^{(k)} - \alpha^{(k)} H^{(k)} g^{(k)}.$$

Here $\alpha^{(k)}$ is chosen to minimize f along $-H^{(k)}g^{(k)}$. Some well-known techniques of this type are the Fletcher-Powell modification of Davidon's method [1], [2], Broyden methods [3], [10], the Barnes-Rosen method [4], [5], and Goldfarb's method [11].

The Fletcher-Powell and Barnes-Rosen methods share the computational feature that, if $f(x)$ is a positive definite quadratic form, the sequence (2) converges in n iterations. This feature is also true of Broyden's method defined in [10], but not of those devised in [3] (see [6]).

Further, the Fletcher-Powell technique guarantees that the matrix, $H^{(k)}$, will always be positive semidefinite, expediting the search for $\alpha^{(k)}$ at each step.

This paper will develop a family of matrices, $H^{(k)}$, as a function of a scalar parameter, ϵ , all of which can be shown to possess the quadratic convergence property of the Fletcher-Powell and Barnes-Rosen techniques. It will further be shown that both the Fletcher-Powell and Barnes-Rosen matrices are special cases of this parametric family, and that positivity depends only on proper choice of the parameter.

A problem which arises in connection with quasi-Newton methods occurs when the smallest eigenvalue of $H^{(k)}$ goes to zero. This is the so-called conditioning problem.

Received March 14, 1969, revised January 22, 1970.

AMS Subject Classifications. Primary 65J05; Secondary 65J08.

Key Words and Phrases. Function minimization, quasi-Newton methods, variable metric methods, gradient search, steepest-descent methods, stability of search methods, conditioning of search methods, Hessian matrix, inverse approximations.

Copyright © 1971, American Mathematical Society

647

B

F

G

S

Issues with Theoretical Analysis of Quasi-Newton Methods

Virtually all previous analyses rely on (assumed or proved) bounds of the type:

$$\hat{\mu}\mathbf{I} \preceq \mathbf{B}_k^{-1} \preceq \hat{L}\mathbf{I} \quad \forall k$$

The analysis then proceeds similarly to analysis of GD

Rate depends on the condition number $\frac{\hat{L}}{\hat{\mu}}$

- **Can be astronomical!** Much worse (by many orders of magnitude!) than the condition number of GD.
- Analysis does not benefit from what QN methods are all about: “better estimation of the inverse Hessian”.

Issues with Theoretical Analysis of Quasi-Newton Methods

**Despite 50+ years of history,
theoretical understanding of Quasi-Newton
methods is very weak!**

Part II

Randomized BFGS for Matrix Inversion



Robert M. Gower and P. R.
Randomized Quasi-Newton Updates are Linearly Convergent Matrix Inversion Algorithms
SIAM Journal on Matrix Analysis and Applications 38(4):1380-1409, 2017

RANDOMIZED QUASI-NEWTON UPDATES ARE LINEARLY CONVERGENT MATRIX INVERSION ALGORITHMS*

ROBERT M. GOWER[†] AND PETER RICHTÁRIK[‡]

Abstract. We develop and analyze a broad family of stochastic/randomized algorithms for calculating an approximate inverse matrix. We also develop specialized variants maintaining symmetry or positive definiteness of the iterates. All methods in the family converge globally and linearly (i.e., the error decays exponentially), with explicit rates. In special cases, we obtain stochastic block variants of several quasi-Newton updates, including bad Broyden (BB), good Broyden (GB), Powell-symmetric-Broyden (PSB), Davidson-Fletcher-Powell (DFP), and Broyden-Fletcher-Goldfarb-Shanno (BFGS). Ours are the first stochastic versions of these updates shown to converge to an inverse of a fixed matrix. Through a dual viewpoint we uncover a fundamental link between quasi-Newton updates and approximate inverse preconditioning. Further, we develop an adaptive variant of randomized block BFGS, where we modify the distribution underlying the stochasticity of the method throughout the iterative process to achieve faster convergence. By inverting several matrices from varied applications, we demonstrate that adaptive randomized BFGS (AdaRBFGS) is highly competitive when compared to the Newton-Schulz method, a minimal residual method and direct inversion method based on a Cholesky decomposition. In particular, on large-scale problems our method outperforms the standard methods by orders of magnitude at calculating an approximate inverse. Development of efficient methods for estimating the inverse of very large matrices is a much needed tool for preconditioning and variable metric optimization methods in the advent of the big data era.

Key words. matrix inversion, stochastic methods, iterative methods, quasi-Newton, BFGS, stochastic convergence

AMS subject classifications. 15A09, 90C53, 68W20, 65N75, 65F35, 65Y20, 68Q25, 68W40

DOI. 10.1137/16M1062053

1. Introduction. Matrix inversion is a standard tool in numerics that is needed, for instance, in computing a projection matrix or a Schur complement, which are commonplace calculations. When only an approximate inverse is required, then iterative methods are the methods of choice, for they can terminate the iterative process when the desired accuracy is reached. This can be far more efficient than using a direct method. Calculating an approximate inverse is a much needed tool in preconditioning [33], and, if the output is guaranteed to be positive definite, then it can be used to design variable metric optimization methods. Furthermore, iterative methods can make use of an initial estimate of the inverse when available.

The driving motivation of this work is the need to develop algorithms capable of computing an approximate inverse of very large matrices, where standard techniques take an exorbitant amount of time or simply fail. In particular, we develop a family of randomized/stochastic methods for inverting a matrix, with specialized variants maintaining symmetry or positive definiteness of the iterates. All methods in the family converge globally (i.e., from any starting point) and linearly (i.e., the error decays exponentially). We give an explicit expression for the convergence rate.

*Received by the editors February 19, 2016; accepted for publication (in revised form) by M. P. Friedlander September 19, 2017; published electronically November 14, 2017.
<http://www.siam.org/journals/simax/38-4/M106205.html>

Funding: The work of the second author was supported by the EPSRC grant EP/K02325X/1, *Accelerated Coordinate Descent Methods for Big Data Optimization*, and the EPSRC Fellowship EP/N005538/1, *Randomized Algorithms for Extreme Converse Optimization*.

[†]School of Mathematics, The Maxwell Institute for Mathematical Sciences, University of Edinburgh, Edinburgh EH9 3FD, UK (gowerrobert@gmail.com, peter.richtarik@ed.ac.uk).

RANDOMIZED QUASI-NEWTON UPDATES ARE LINEARLY CONVERGENT MATRIX INVERSION ALGORITHMS*

ROBERT M. GOWER[†] AND PETER RICHTÁRIK[‡]

Abstract. We develop and analyze a broad family of stochastic/randomized algorithms for calculating an approximate inverse matrix. We also develop specialized variants maintaining symmetry or positive definiteness of the iterates. All methods in the family converge globally and linearly (i.e., the error decays exponentially), with explicit rates. In special cases, we obtain stochastic block variants of several quasi-Newton updates, including bad Broyden (BB), good Broyden (GB), Powell-symmetric-Broyden (PSB), Davidson-Fletcher-Powell (DFP), and Broyden-Fletcher-Goldfarb-Shanno (BFGS). Ours are the first stochastic versions of these updates shown to converge to an inverse of a fixed matrix. Through a dual viewpoint we uncover a fundamental link between quasi-Newton updates and approximate inverse preconditioning. Further, we develop an adaptive variant of randomized block BFGS, where we modify the distribution underlying the stochasticity of the method throughout the iterative process to achieve faster convergence. By inverting several matrices from varied applications, we demonstrate that adaptive randomized BFGS (AdaRBFGS) is highly competitive when compared to the Newton-Schulz method, a minimal residual method and direct inversion method based on a Cholesky decomposition. In particular, on large-scale problems our method outperforms the standard methods by orders of magnitude at calculating an approximate inverse. Development of efficient methods for estimating the inverse of very large matrices is a much needed tool for preconditioning and variable metric optimization methods in the advent of the big data era.

Key words. matrix inversion, stochastic methods, iterative methods, quasi-Newton, BFGS, stochastic convergence

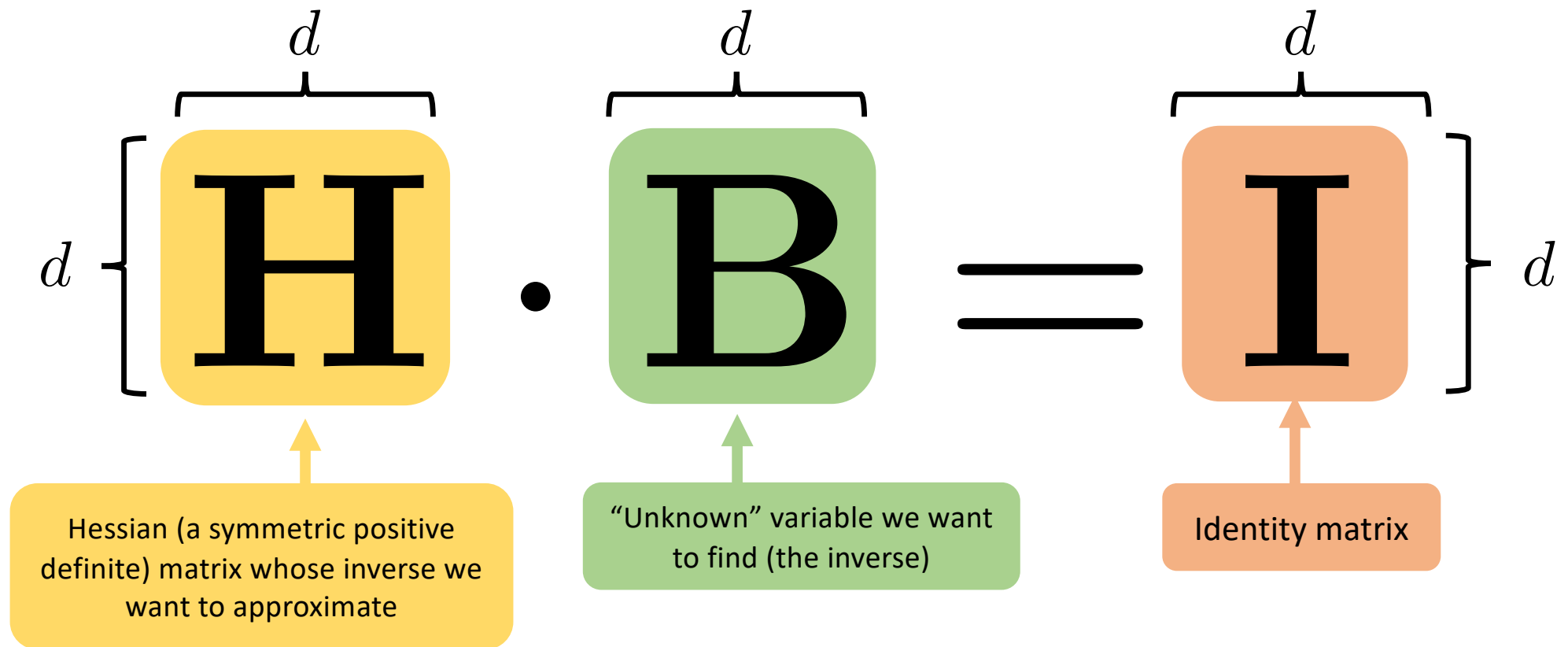
AMS subject classifications. 15A09, 90C53, 68W20, 65N75, 65F35, 65Y20, 68Q25, 68W40

DOI. 10.1137/16M1062053

Matrix Inversion: The Problem

Approximate the inverse of $\mathbf{H} \in \mathbb{S}_{++}^d$

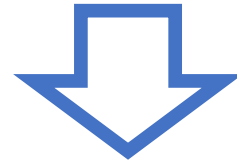
Linear Algebra Formulation of Matrix Inversion



Unique solution: $\mathbf{B} = \mathbf{H}^{-1}$

Sketched System aka Random Secant Equation

$$\mathbf{H} \cdot \mathbf{B} = \mathbf{I}$$



Random Secant Equation

$$\mathbf{S}^\top \cdot \mathbf{H} \cdot \mathbf{B} = \mathbf{S}^\top \cdot \mathbf{I}$$

Random matrix

$$\mathbf{S} \in \mathbb{R}^{d \times \tau}$$

Classical Secant Equation

$$(\nabla f(x_{k+1}) - \nabla f(x_k))^\top \cdot \mathbf{B} = (x_{k+1} - x_k)^\top$$

Classical vs Random Secant Equation

Classical

$$(\nabla f(x_{k+1}) - \nabla f(x_k))^\top \cdot \mathbf{B} = (x_{k+1} - x_k)^\top$$

Random

$$\mathbf{S}^\top \cdot \mathbf{H} \cdot \mathbf{B} = \mathbf{S}^\top \cdot \mathbf{I}$$



1 equation per column of \mathbf{B}

τ equations per column of \mathbf{B}

$$\mathbf{S} \in \mathbb{R}^{d \times \tau} \quad \tau \in \{1, 2, \dots\}$$



$\mathbf{B} = \mathbf{H}^{-1}$ may not be a solution

$\mathbf{B} = \mathbf{H}^{-1}$ is a solution



Does not need access to \mathbf{H}

Needs access to \mathbf{H}



Equations are
deterministic and adaptive

Equations are
random and stationary

Three Equivalent Formulations of Randomized BFGS

RBFGS: Primal Formulation (Sketch & Project)

Closest matrix to the current one satisfying the **random** secant equation and symmetry

Current estimate of the inverse Hessian

$$\mathbf{B}_k \approx \mathbf{H}^{-1}$$

$$\mathbf{B}_{k+1} = \arg \min_{\mathbf{B} \in \mathbb{R}^{d \times d}} \|\mathbf{B} - \mathbf{B}_k\|_F(\mathbf{H})$$

subject to

Enforcing symmetry

$$\mathbf{s}_k^\top \mathbf{H} \mathbf{B} = \mathbf{s}_k^\top$$

$$\mathbf{B} = \mathbf{B}^\top$$

Random secant equation

Weighted Frobenius norm

$$\|\mathbf{X}\|_{F(\mathbf{W})} \stackrel{\text{def}}{=} \|\mathbf{W}^{1/2} \mathbf{X} \mathbf{W}^{1/2}\|_F$$
$$\mathbf{W} = \mathbf{H}$$

RBFGS: Dual Formulation (Constrain & Approximate)

Closest matrix to the inverse Hessian
belonging to a certain **random** affine space

Inverse Hessian

$$\mathbf{B}_{k+1} = \arg \min_{\mathbf{B} \in \mathbb{R}^{d \times d}, \mathbf{Y} \in \mathbb{R}^{d \times \tau}} \|\mathbf{B} - \mathbf{H}^{-1}\|_{F(\mathbf{H})}$$

subject to

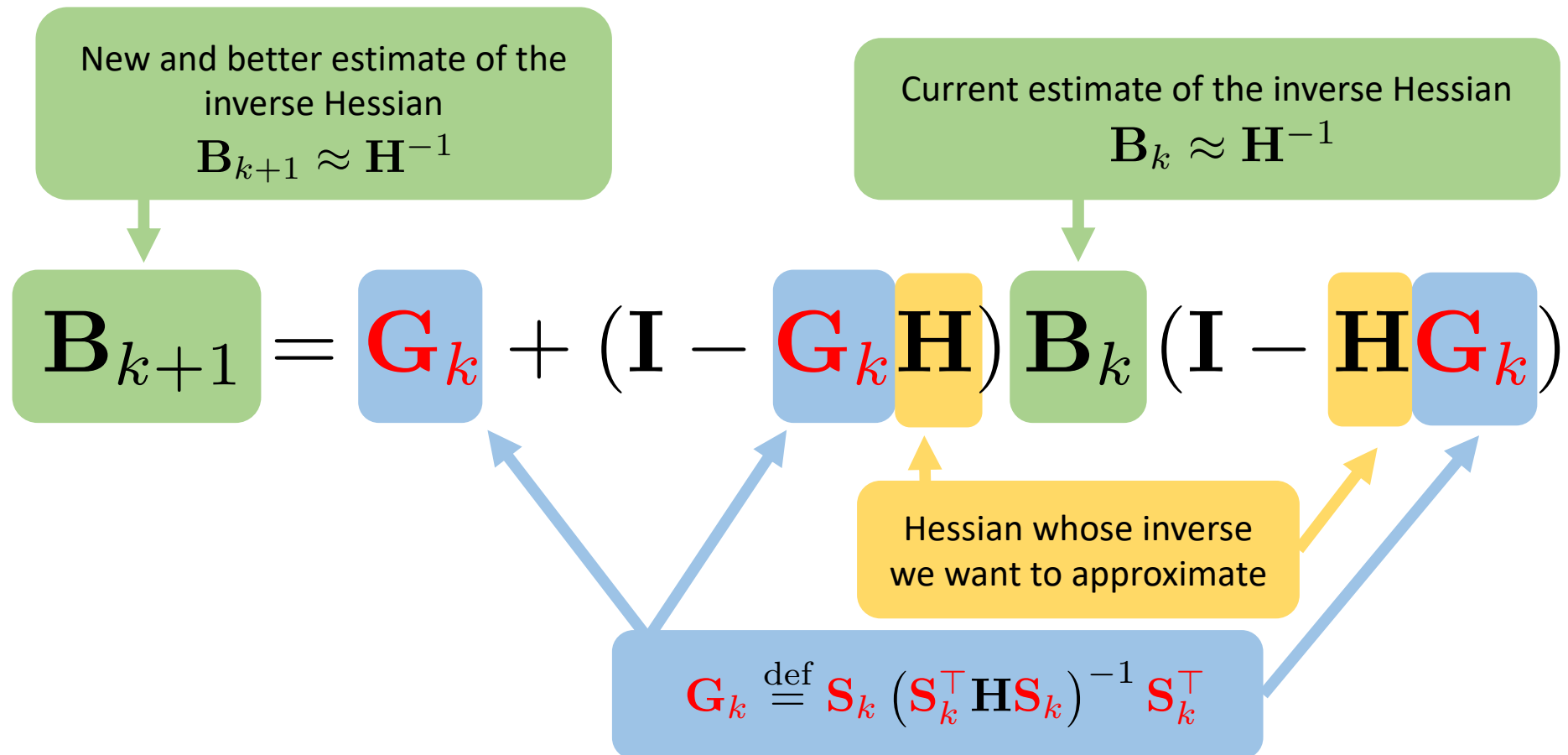
$$\mathbf{B} = \mathbf{B}_k + \mathbf{Y} \mathbf{S}_k^\top + \mathbf{S}_k \mathbf{Y}^\top$$

Current estimate of the inverse Hessian

$$\mathbf{B}_k \approx \mathbf{H}^{-1}$$

Symmetric rank-**2** τ update

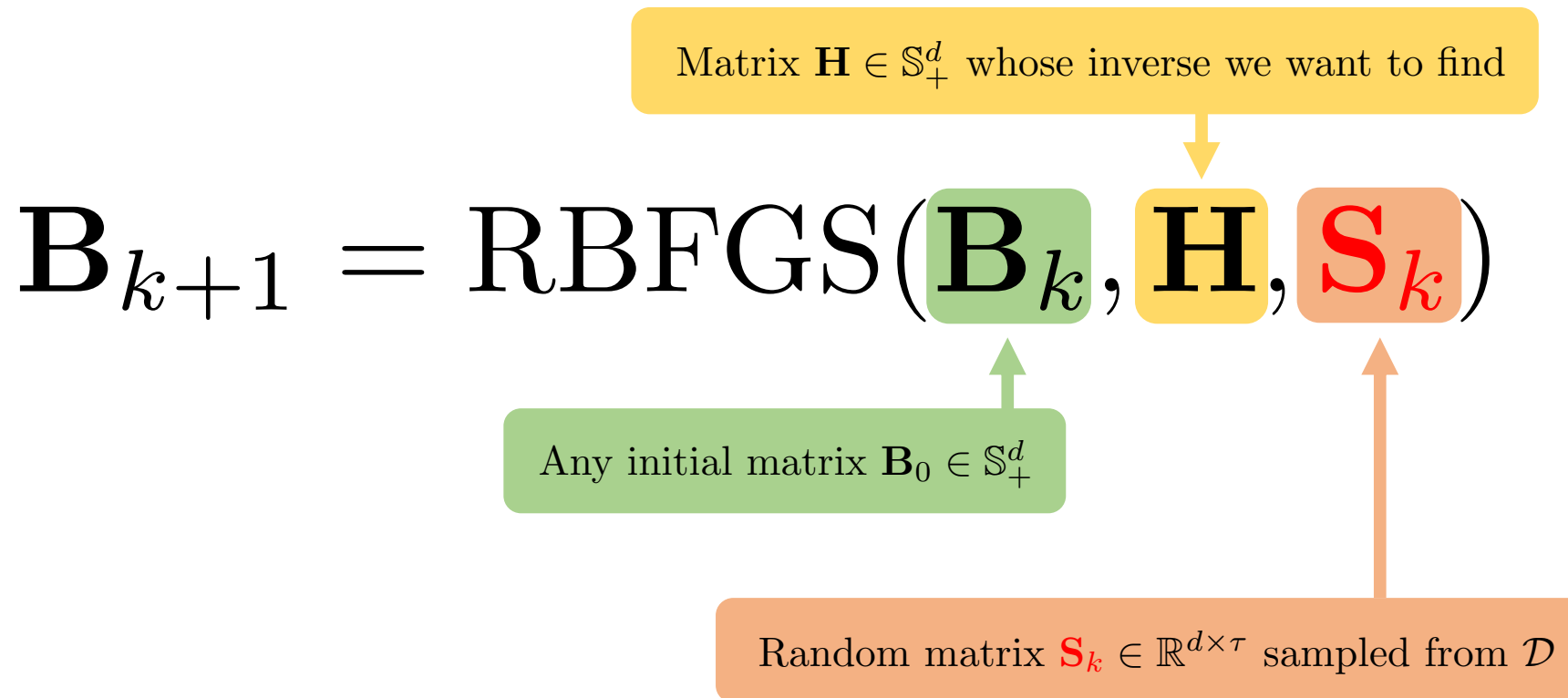
RBFGS: Explicit Solution



Three Equivalent Formulations of RBFGS

<p>Primal (Sketch & Project)</p>	<p>Closest matrix to the current one satisfying the random secant equation and symmetry</p> <p>Current estimate of the inverse Hessian $B_k \approx H^{-1}$</p> <p>$B_{k+1} = \arg \min_{B \in \mathbb{R}^{d \times d}} \ B - B_k\ _{F(H)}$</p> <p>subject to $S_k^T H B = S_k^T$</p> <p>Enforcing symmetry $B = B^T$</p> <p>Random secant equation</p> <p>Weighted Frobenius norm $\ X\ _{F(W)} \stackrel{\text{def}}{=} \ W^{1/2} X W^{1/2}\ _F$, $W = H$</p>
<p>Dual (Constrain & Approximate)</p>	<p>Closest matrix to the inverse Hessian belonging to a certain random affine space</p> <p>Inverse Hessian</p> <p>$B_{k+1} = \arg \min_{B \in \mathbb{R}^{d \times d}, Y \in \mathbb{R}^{d \times r}} \ B - H^{-1}\ _{F(H)}$</p> <p>subject to $B = B_k + Y S_k^T + S_k Y^T$</p> <p>Current estimate of the inverse Hessian $B_k \approx H^{-1}$</p> <p>Symmetric rank-$2r$ update</p>
<p>Explicit Solution</p>	<p>New and better estimate of the inverse Hessian $B_{k+1} \approx H^{-1}$</p> <p>Current estimate of the inverse Hessian $B_k \approx H^{-1}$</p> <p>$B_{k+1} = G_k + (I - G_k H) B_k (I - H G_k)$</p> <p>Hessian whose inverse we want to approximate</p> <p>$G_k \stackrel{\text{def}}{=} S_k (S_k^T H S_k)^{-1} S_k^T$</p>

Randomized BFGS for Matrix Inversion




Convergence Rate of RBFGS

RBFGS: Convergence Rate

Theorem (Gower-R, 2017)

$$\mathbb{E} \left[\left\| \mathbf{B}_k - \mathbf{H}^{-1} \right\|_{F(\mathbf{H})}^2 \right] \leq (1 - \rho)^k \left\| \mathbf{B}_0 - \mathbf{H}^{-1} \right\|_{F(\mathbf{H})}^2$$


$$\rho \stackrel{\text{def}}{=} \lambda_{\min} \left(\mathbb{E}_{\mathbf{S} \sim \mathcal{D}} \left[\mathbf{H}^{1/2} \mathbf{S} (\mathbf{S}^\top \mathbf{H} \mathbf{S})^{-1} \mathbf{S}^\top \mathbf{H}^{1/2} \right] \right)$$
$$0 \leq \rho \leq \frac{\tau}{d} \quad \mathbf{S} \in \mathbb{R}^{d \times \tau}$$

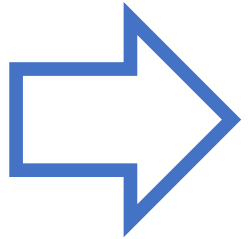
Part III

Randomized BFGS for Optimization



Dmitry Kovalev, Robert M. Gower, P. R. and Alexander Rogozin
Fast Linear Convergence of Randomized BFGS
arXiv:2002.11337, 2020

From Matrix Inversion to Optimization

Compute \mathbf{H}^{-1}  $\min_{x \in \mathbb{R}^d} f(x)$

$$x_* = \arg \min_x f(x)$$

Algorithm: RBFGS for Optimization

Any initial matrix $\mathbf{B}_0 \in \mathbb{S}_{++}^d$

Goal: $\mathbf{B}_k \approx (\nabla f(x_k))^{-1}$

Random matrix $\mathbf{S}_k \in \mathbb{R}^{d \times \tau}$
sampled from \mathcal{D}

$$x_{k+1} = x_k - \mathbf{B}_k \nabla f(x_k)$$

$$\mathbf{B}_{k+1} = \text{RBFGS}(\mathbf{B}_k, \mathbf{H}_k, \mathbf{S}_k)$$

RBFGS was (semi) heuristically applied to optimization in



Robert M. Gower, Donald Goldfarb and P. R.
Stochastic Block BFGS: Squeezing More Curvature out of Data
ICML 2016

Now the matrix is changing!

$$\mathbf{H}_k = \nabla^2 f(x_k)$$



Three Theorems

#	Result	Assumptions
Theorem 1	Local linear convergence $x_k \rightarrow x_*, \mathbf{B}_k \rightarrow (\nabla^2 f(x_*))^{-1}$	<ul style="list-style-type: none">• Self concordance
Theorem 2	Local linear convergence $f(x_k) \rightarrow f(x_*), \mathbf{B}_k \rightarrow (\nabla^2 f(x_*))^{-1}$	<ul style="list-style-type: none">• Strong convexity• Lipschitz gradient• Lipschitz Hessian
Theorem 3	Superlinear convergence with probability 1 $\sqrt{f(x_k) - f(x_*)} \rightarrow 0$	

RBFGS = First quasi-Newton method whose (local) linear rate is (in some regimes) provably better than that of gradient descent!

Theorem 1

$$\|x_0 - x_*\|_{\mathbf{H}_*} \leq \frac{1}{2} \min \left\{ \frac{1}{2}, 1 - \sqrt{\frac{1 - \rho}{1 - \frac{2\rho}{3}}}, \frac{4 - 2\rho}{9\rho d + 10} \right\}$$

If f is self-concordant and x_0 is close enough to x_* , then

$$\mathbb{E} [\Phi_k] \leq \left(1 - \frac{\rho}{2}\right)^k \Phi_0$$

$$\mathbf{H}_* = \nabla^2 f(x_*)$$

$$\Phi_k \stackrel{\text{def}}{=} \frac{3}{7\rho} \|\mathbf{B}_k - \mathbf{H}_*^{-1}\|_{F(\mathbf{H}_*)}^2 + \|x_k - x_*\|_{\mathbf{H}_*}$$

$$\rho \stackrel{\text{def}}{=} \inf_{x \in \mathbb{R}^d} \lambda_{\min} \left(\mathbb{E}_{\mathbf{S} \sim \mathcal{D}} \left[\mathbf{H}_x^{1/2} \mathbf{S} (\mathbf{S}^\top \mathbf{H}_x \mathbf{S})^{-1} \mathbf{S}^\top \mathbf{H}_x^{1/2} \right] \right)$$

$$\mathbf{H}_x = \nabla^2 f(x)$$

RBFGS can be Better than GD

$$0 < l \leq \phi_i''(t) \leq u \text{ for all } t \in \mathbb{R}$$

$$\phi_i'''(t) \leq C \text{ for all } t \in \mathbb{R}$$

Generalized linear models: $f(x) = \frac{1}{n} \sum_{i=1}^n \phi_i(a_i^\top x)$

$$\mathbf{A} = [a_1, \dots, a_n] \in \mathbb{R}^{d \times n}$$

full row rank

Definition (SVD Sketch)

$$\mathbf{A} = \mathbf{U} \mathbf{\Sigma} \mathbf{V}^\top$$

$\mathbf{U} \in \mathbb{R}^{d \times d}$ $\mathbf{\Sigma} \in \mathbb{R}^{d \times d}$ $\mathbf{V} \in \mathbb{R}^{n \times d}$

\mathcal{D} is defined by:

$$\text{Prob} \left(\mathbf{S}_k = \frac{1}{\Sigma_{ii}} \mathbf{U}_{:i} \right) = \frac{1}{d} \quad \text{for } i = 1, 2, \dots, d$$

Corollary of Theorem 1

RBFGS with SVD sketch converges with rate $(1 - \frac{\rho}{2})^k$, where $\rho \geq \frac{l}{u} \frac{1}{d}$

GD converges with rate $(1 - \rho_{\text{GD}})^k$, where $\rho_{\text{GD}} = \frac{l}{u} \frac{\sigma_{\min}^2(\mathbf{A})}{\sigma_{\max}^2(\mathbf{A})}$

Unlike GD, RBFGS has rate independent of the conditioning of the matrix **A**.

Theorem 2

$$\begin{aligned}\|\nabla f(x) - \nabla f(y)\| &\leq L_1 \|x - y\| \\ \|\nabla^2 f(x) - \nabla^2 f(y)\| &\leq L_2 \|x - y\|\end{aligned}$$

$$f(x_0) - f(x_*) \leq \frac{1}{4} \left[\frac{\sqrt{2L_1}L_2}{\mu^2} + \frac{32\sqrt{2}dL_1^{5/2}L_2}{\rho\mu^4} \right]^{-2}$$

If f is (L_1, L_2) -smooth, μ -strongly convex and x_0 is close enough to x_* , then

$$\mathbb{E} [\Psi_k] \leq \left(1 - \frac{\rho}{2}\right)^k \Psi_0$$

$$\mathbf{H}_* = \nabla^2 f(x_*)$$

$$\Psi_k \stackrel{\text{def}}{=} \frac{4\sqrt{2}L_1^{5/2}}{\mu L_2 \rho} \|\mathbf{B}_k - \mathbf{H}_*^{-1}\|_{F(\mathbf{H}_*)}^2 + \sqrt{f(x_k) - f(x_*)}$$

$$\rho \stackrel{\text{def}}{=} \inf_{\substack{x: f(x) \leq f(x_0) \\ \text{NEW}}} \lambda_{\min} \left(\mathbb{E}_{\mathbf{S} \sim \mathcal{D}} \left[\mathbf{H}_x^{1/2} \mathbf{S} (\mathbf{S}^\top \mathbf{H}_x \mathbf{S})^{-1} \mathbf{S}^\top \mathbf{H}_x^{1/2} \right] \right)$$

$\mathbf{H}_x = \nabla^2 f(x)$

Experiments

Convex Quadratic with Hilbert Hessian

Condition number

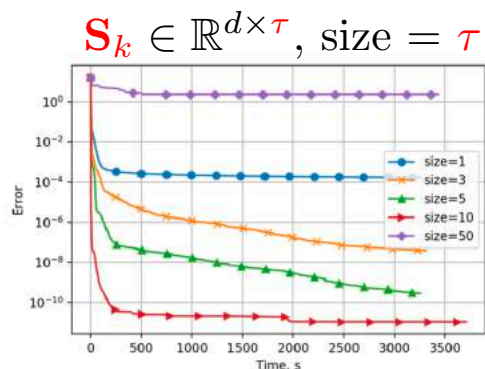
$$\kappa = O((1 + \sqrt{2})^{4d}) \approx 8.5 \times 10^{16}$$

Hilbert Matrix

$$\mathbf{A}_{ij} = \frac{1}{i + j - 1}$$

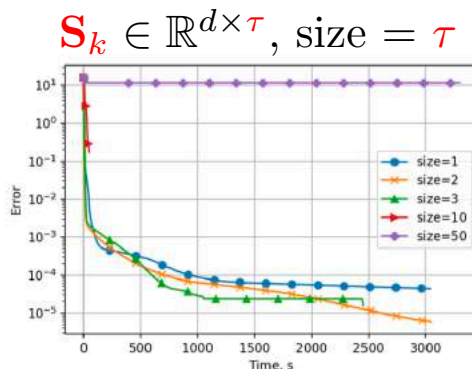
$$\min_{x \in \mathbb{R}^d} \left\{ f(x) = \frac{1}{2} x^\top (\mathbf{A}^\top \mathbf{A}) x \right\}$$

$$d = 10,000$$



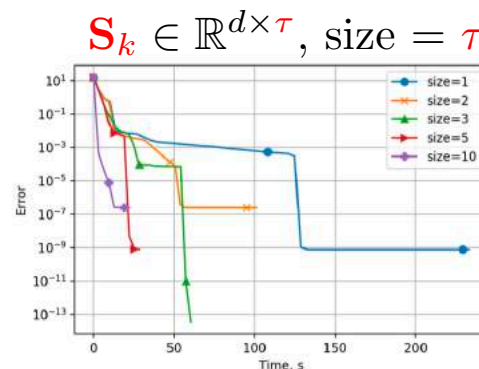
(a) gauss

$$\mathbf{S}_k \sim \mathcal{D}_{\text{gauss}}$$



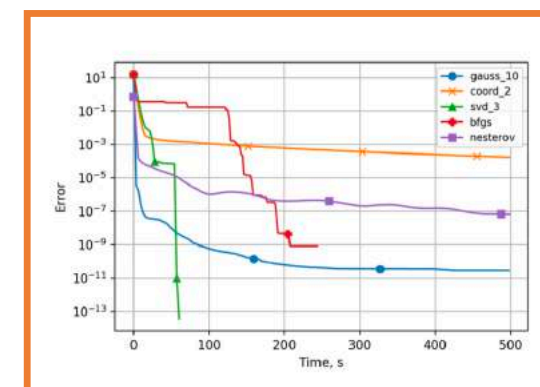
(b) coord

$$\mathbf{S}_k \sim \mathcal{D}_{\text{coord}}$$



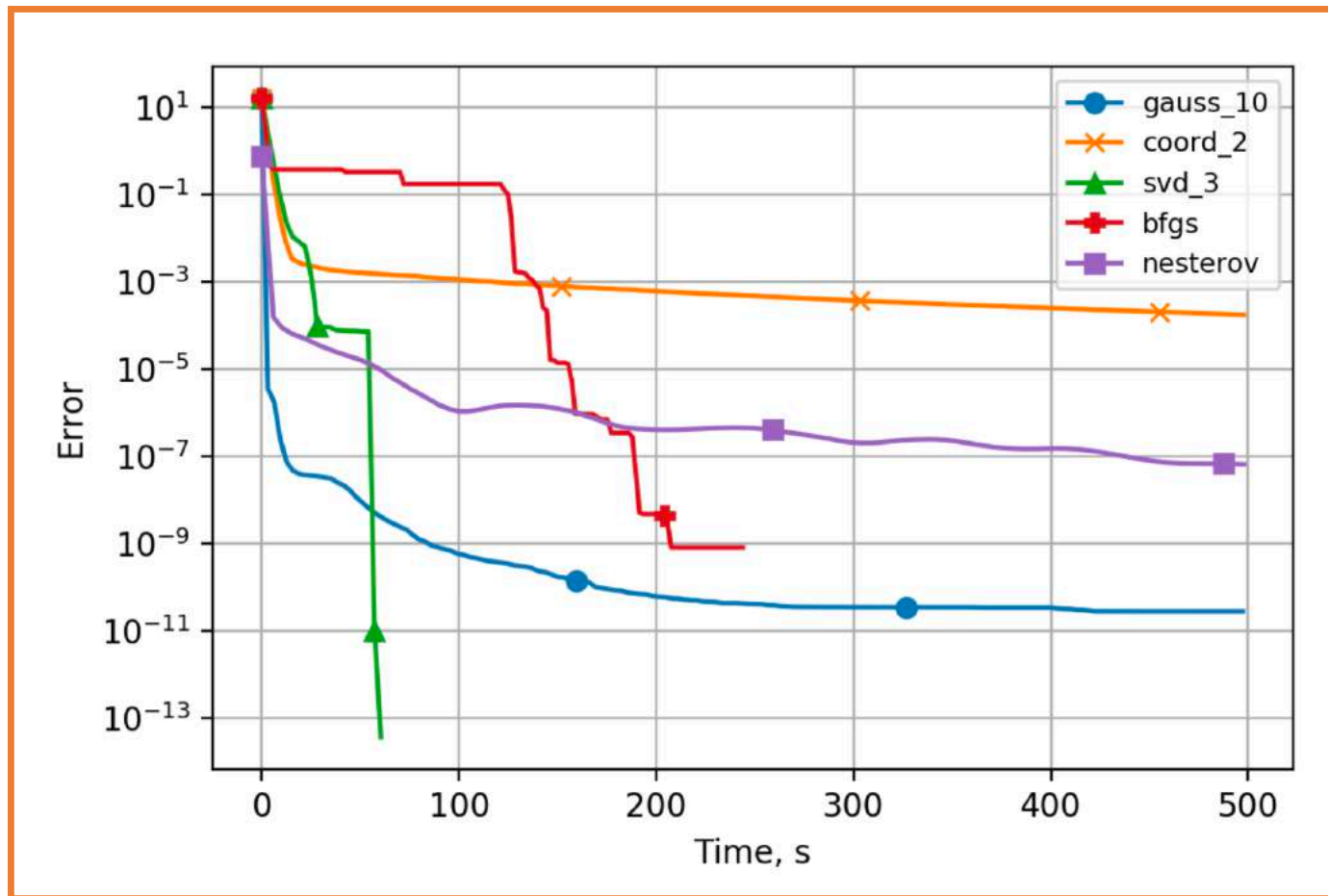
(c) svd

$$\mathbf{S}_k \sim \mathcal{D}_{\text{svd}}$$



(d) methods compared

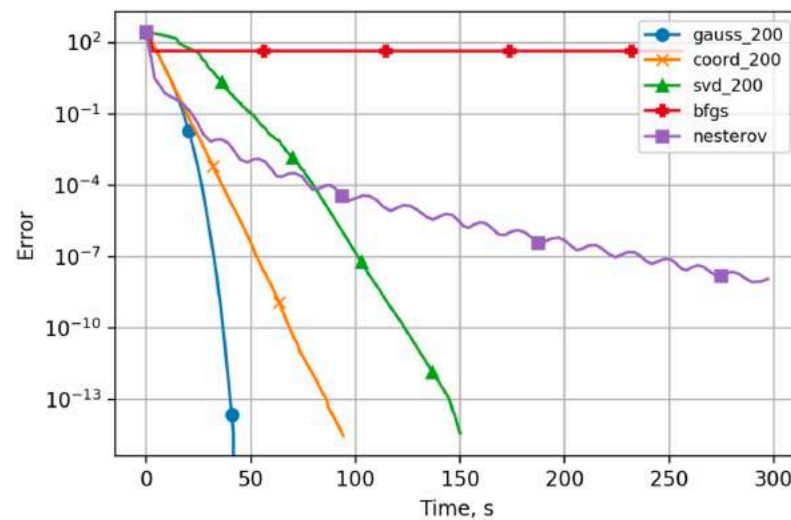
Convex Quadratic with Hilbert Hessian



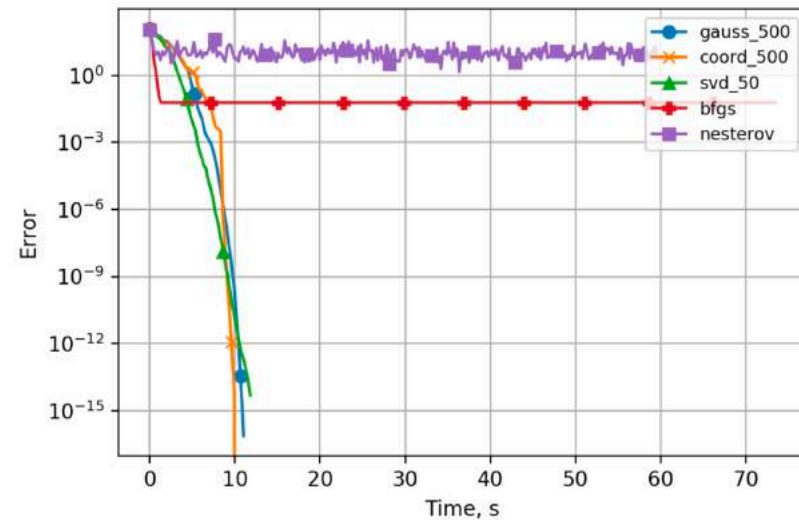
Chang & Lin (2011)

Binary Classification on LIBSVM Data

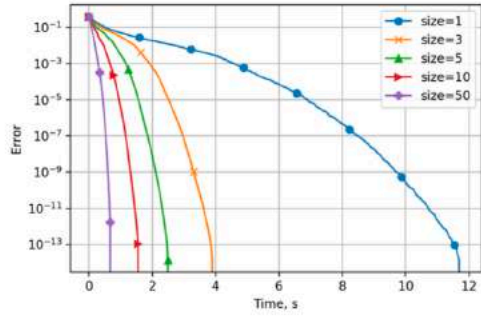
$$\min_{x \in \mathbb{R}^d} \left\{ f(x) = \frac{1}{n} \sum_{i=1}^n \log(1 + \exp(-b_i \langle a_i, x \rangle)) + \frac{\lambda}{2} \|x\|_2^2 \right\}$$



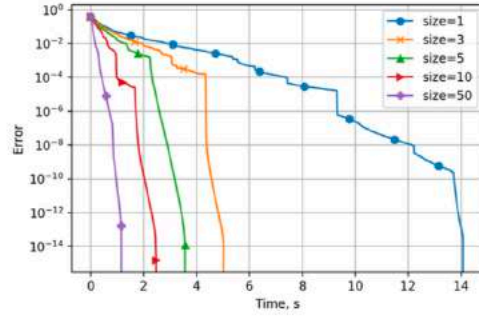
gisette
 $\lambda = 10^{-1}$
 $n = 6,000$; $d = 5,000$
 $\kappa = 1.2 \times 10^4$



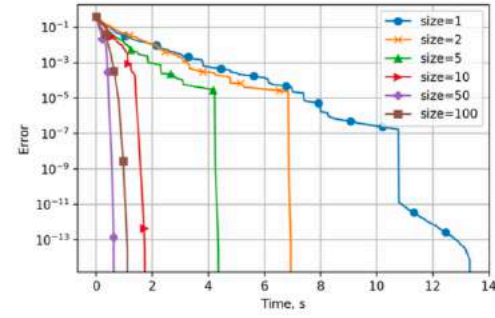
colon-cancer
 $\lambda = 10^{-1}$
 $n = 62$; $d = 2,000$
 $\kappa = 9.6 \times 10^3$



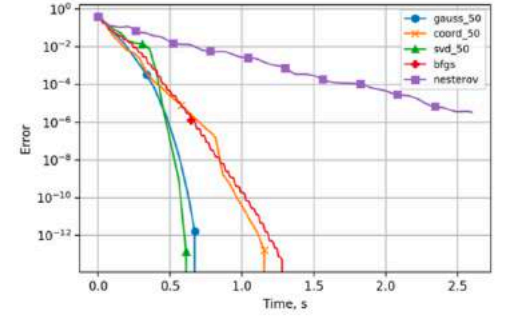
(a) gauss



(b) coord

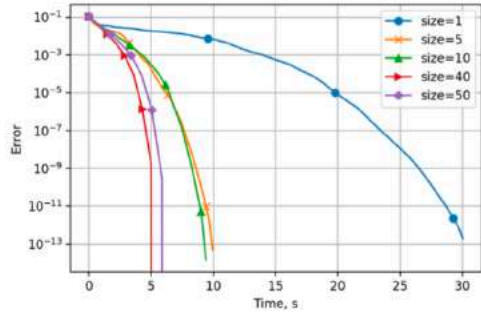


(c) svd

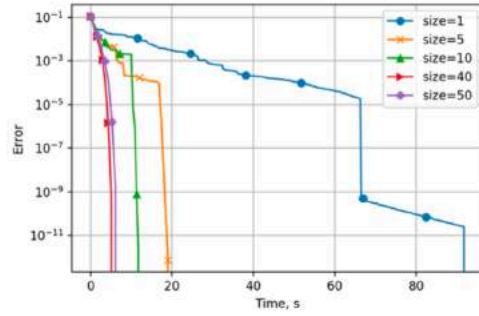


(d) methods compared

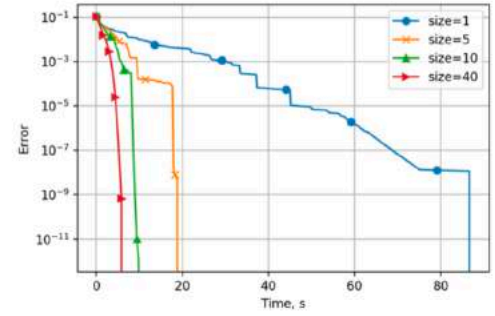
Figure 3: a9a; $\lambda = 10^{-3}$; $n = 29, 159$; $d = 123$; $\kappa = 3.5 \cdot 10^3$



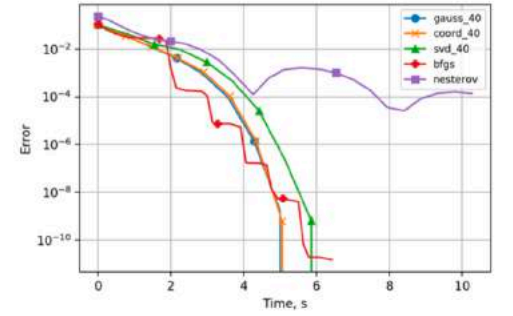
(a) gauss



(b) coord



(c) svd



(d) methods compared

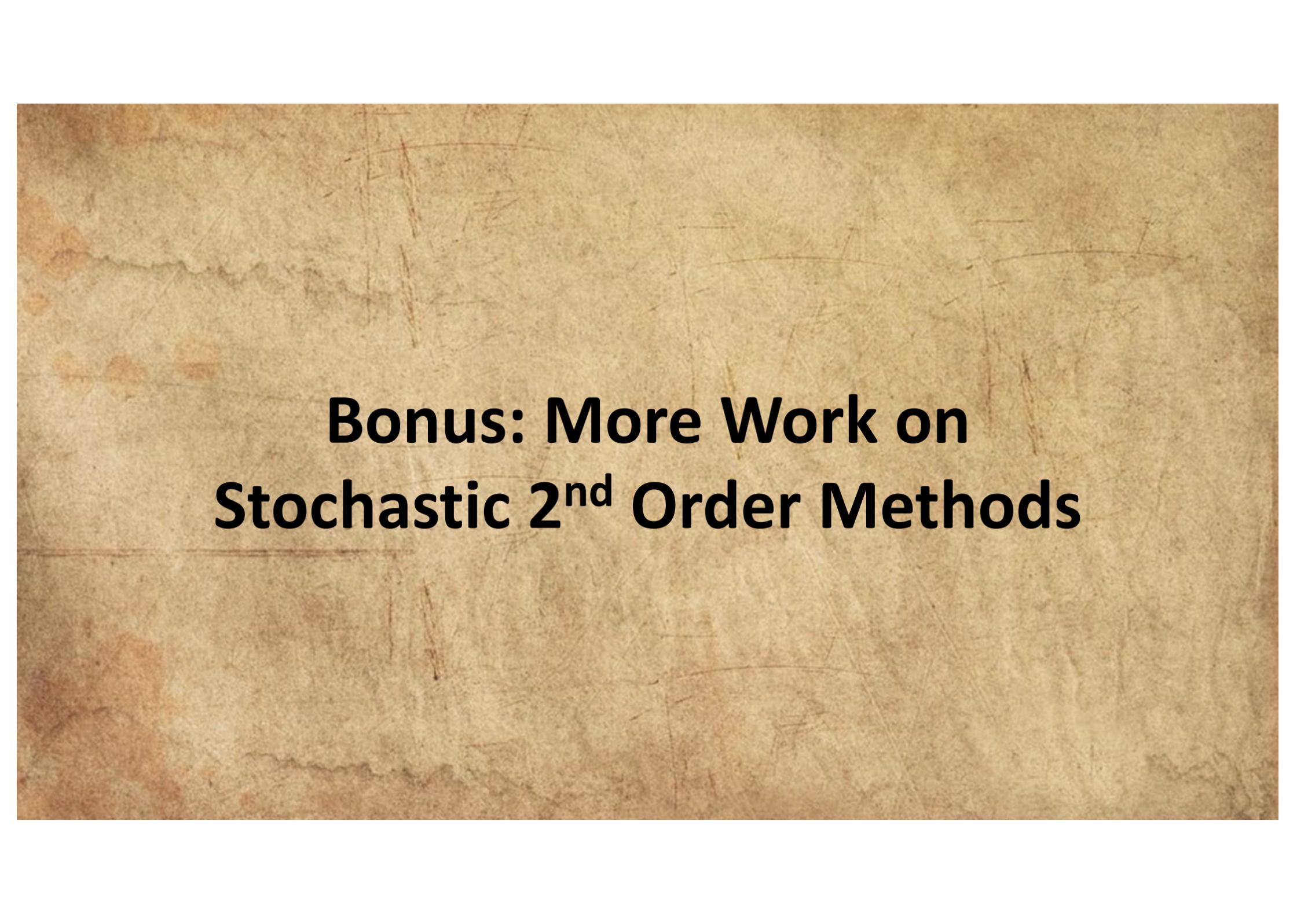
Figure 4: covtype; $\lambda = 10^{-3}$; $n = 581, 012$; $d = 54$; $\kappa = 1.9 \cdot 10^3$



Summary

Summary

- Randomized BFGS was introduced
 - by Gower & R (arXiv 2/2016; SIMAX 2017) for **matrix inversion**
 - by Gower-Goldfarb-R (arXiv 3/2016, ICML 2016) for **optimization**
- We established **local linear convergence rate of RBFGS**
 - Theorem 1: self-concordant functions
 - Theorem 2: smooth and strongly convex functions
 - Theorem 3: superlinear convergence
- **First analysis of any quasi-Newton method (RBFGS) which shows improvement on GD**
 - **Novel Lyapunov style analysis**
 - **Convergence of inverse Hessian estimates (theoretical benefits!)**
 - Convergence of itearates & function values



**Bonus: More Work on
Stochastic 2nd Order Methods**

Big d Regime



Zheng Qu, Peter Richtárik, Martin Takáč and Olivier Fercoq
SDNA: Stochastic dual Newton ascent for empirical risk minimization
ICML 2016

Handles big n regime by taking (randomized) subspace Newton steps in the dual.

Superlinear speedup in minibatch size.



Robert M. Gower, Donald Goldfarb and Peter Richtárik
Stochastic block BFGS: squeezing more curvature out of data
ICML 2016

Work used to motivate this talk



Robert M. Gower and Peter Richtárik
Randomized quasi-Newton updates are linearly convergent matrix inversion algorithms
SIAM Journal on Matrix Analysis and Applications 38(4):1380-1409, 2017



Robert M. Gower, Filip Hanzely, Peter Richtárik and Sebastian Stich
Accelerated stochastic matrix inversion: general theory and speeding up BFGS rules for faster second-order optimization
NeurIPS 2018

First accelerated quasi-Newton matrix inversion rules

Big d Regime



Nikita Doikov and Peter Richtárik
Randomized Block Cubic Newton Method
ICML 2018



Robert M. Gower, Dmitry Kovalev, Felix Lieder and Peter Richtárik
RSN: Randomized Subspace Newton
NeurIPS 2019



Filip Hanzely, Nikita Doikov, Peter Richtárik and Yurii Nesterov
Stochastic Subspace Cubic Newton Method
arXiv:2002.09526, 2020 (ICML 2020)



Dmitry Kovalev, Robert M. Gower, Peter Richtárik and Alexander Rogozin
Fast Linear Convergence of Randomized BFGS
arXiv:2002.11337, 2020

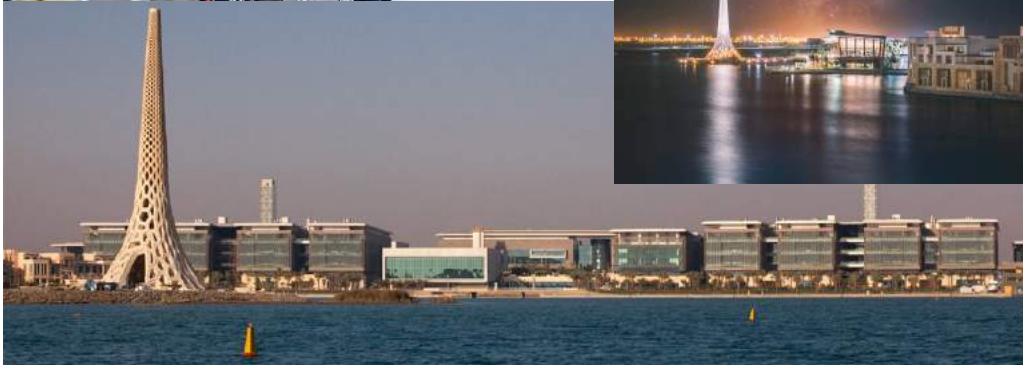
Best rates for
(randomized) subspace
Newton methods

Work presented in this
talk

Three Slides About KAUST



Started: 2009
Graduate-Only



Optimization and Machine Learning Lab



Openings: research scientists, postdocs, PhD & MS students, interns

Research Scientists

[El Houcine Bergou](#) (from Toulouse)

[Laurent Condat](#) (from Grenoble)

Postdocs

[Mher Safaryan](#) (from Yerevan)

[Zhize Li](#) (from Tsinghua)

[Adil Salim](#) (from Télécom ParisTech)

[Xun Qian](#) (from Hong Kong)

PhD Students

[Dmitry Kovalev](#) (from MIPT)

[Elnur Gasanov](#) (from MIPT)

[Samuel Horváth](#) (from Comenius)

[Alibek Sailanbayev](#) (from MIPT)

[Konstantin Mishchenko](#) (from ENS Paris-Saclay)

[Filip Hanzely](#) (from Edinburgh)

MS Students

[Egor Shulgin](#) (from MIPT)

[Alyazeed Basyoni](#) (from CMU)

[Slavomír Hanzely](#) (from Comenius)

Research Interns

[Rustem Islamov](#) (from MIPT)

[Othmane Sebbouh](#) (from École Polytechnique)

[Ahmed Khaled](#) (from Cairo)



The End