Generalization of Quasi-Newton Methods: Application to Robust Symmetric Multisecant Updates

Damien Scieur[1,*] **Lewis Liu**[2,*] **Thomas Pumir**[3] **Nicolas Boumal**[4]
[*] Equal contribution

[1] Samsung SAIT AI Lab, Montréal [2] MILA and DIRO, Université de Montréal

[3] Princeton University and the Voleon Group [4] Institute of Mathematics, EPFL

Abstract

Quasi-Newton (qN) techniques approximate the Newton step by estimating the Hessian using the so-called secant equations. Some of these methods compute the Hessian using several secant equations but produce non-symmetric updates. Other quasi-Newton schemes, such as BFGS, enforce symmetry but cannot satisfy more than one secant equation. We propose a new type of quasi-Newton symmetric update using several secant equations in a least-squares sense. Our approach generalizes and unifies the design of quasi-Newton updates and satisfies provable robustness guarantees.

1 Introduction

We consider second-order methods for unconstrained minimization of a smooth, possibly non-convex function $f\colon \mathbb{R}^d \to \mathbb{R}$. Despite a locally quadratic convergence rate, the well-known Newton method iteration

$$\mathbf{x}_{k+1} = \mathbf{x}_k - \left[\nabla^2 f(\mathbf{x}_k)\right]^{-1} \nabla f(\mathbf{x}_k) \tag{1}$$

is not suitable for large-scale problems, in part because it requires solving a $d \times d$ linear system involving the Hessian at every iteration. To address this issue, quasi-Newton algorithms replace the update rule (1) by

$$\begin{aligned} \mathbf{x}_{k+1} &= \mathbf{x}_k - \mathbf{B}_k^{-1} \nabla f(\mathbf{x}_k) \quad \text{or} \\ \mathbf{x}_{k+1} &= \mathbf{x}_k - \mathbf{H}_k \ \ \nabla f(\mathbf{x}_k), \end{aligned} \tag{2}$$

where $\mathbf{B}_k \approx \nabla^2 f(\mathbf{x}_k)$ and $\mathbf{H}_k \approx \left[\nabla^2 f(\mathbf{x}_k)\right]^{-1}$ are approximations of the Hessian and its inverse (respectively) at $\mathbf{x}_k$. Choosing the right approximation has drawn considerable attention in the optimization literature, notably the DFP update [Davidon, 1959], Broyden method [Broyden, 1965], SR1 update [Byrd et al., 1996] and the well-known BFGS method [Broyden, 1970], [Fletcher, 1970], [Goldfarb, 1970] [Shanno, 1970]. In general, those methods estimate a matrix $\mathbf{B}_k$ or $\mathbf{H}_k$ satisfying the *secant* equation

$$\begin{aligned} \nabla f(\mathbf{x}_k) - \nabla f(\mathbf{x}_{k-1}) &= \mathbf{B}_k(\mathbf{x}_k - \mathbf{x}_{k-1}) \text{ or} \\ \mathbf{H}_k(\nabla f(\mathbf{x}_k) - \nabla f(\mathbf{x}_{k-1})) &= \quad \mathbf{x}_k - \mathbf{x}_{k-1}, \end{aligned} \tag{3}$$

then perform the quasi-Newton step (2). It is also possible to satisfy *several* secant equations. For instance, the multisecant Type-I and Type-II Broyden methods [Fang and Saad, 2009] find a *non-symmetric* matrix $\mathbf{B}_k$ or $\mathbf{H}_k$ satisfying a block of secants: for a memory size m and for $i = k - m + 1 \dots k$,

$$\begin{aligned} \nabla f(\mathbf{x}_i) - \nabla f(\mathbf{x}_{i-1}) &= \mathbf{B}_k[\mathbf{x}_i - \mathbf{x}_{i-1}] \text{ or} \\ \mathbf{H}_k[\nabla f(\mathbf{x}_i) - \nabla f(\mathbf{x}_{i-1})] &= \quad \mathbf{x}_i - \mathbf{x}_{i-1}. \end{aligned}$$

By contrast, other methods like BFGS and DFP enforce the symmetry of the update, but they satisfy only *one* secant equation, in which case Powell [1986] showed their high dependence in the step size. Indeed, while BFGS and DFP enjoy an optimal convergence rate on quadratics using exact line-search [Nocedal and Wright, 1999], Powell [1986] showed that with a

Proceedings of the 24[th] International Conference on Artificial Intelligence and Statistics (AISTATS) 2021, San Diego, California, USA. PMLR: Volume 130.

unitary step size, these updates converge particularly slowly on a simple quadratic function with just two variables. Moreover, it was also observed that BFGS updates are sensitive to gradient noise, and designing quasi-Newton methods for stochastic algorithms is still a challenge [Byrd et al., 2016, Bollapragada et al., 2018, 2019, Berahas et al., 2020].

Unfortunately, except for quadratic functions [Schnabel, 1983], it is usually impossible to find a symmetric matrix that satisfies more than one secant equation. Gower et al. [2016] adopted Hessian-vector products instead of the secant equations. Moreover, line search has been shown to be computationally expensive. Finally, the stabilisation procedure for stochastic BFGS usually requires a growing batch size to reduce the gradient noise, making it unpractical in many applications.

In this paper, we tackle those problems by proposing a symmetric multisecant update, that satisfies the secant equations in a least-squares sense. We show their optimality on quadratics *with unitary step size*, and prove their robustness to gradient noise, making them good candidates in the context of stochastic optimization.

1.1 Notation

We use boldface small letters, like $\mathbf{x}$, to refer to vectors and boldface capital letters, like $\mathbf{A}$, for matrices. We use d to refer to the *dimension* of the problem, and m for the *memory* of the algorithm (we will see later that m is the number of secant equations). For a function $f\colon \mathbb{R}^d \to \mathbb{R}$, its gradient and Hessian at $\mathbf{x}$ are denoted by $\nabla f(\mathbf{x})$ and $\nabla^2 f(\mathbf{x})$ respectively. Consistently with the notations in the literature, we use $\mathbf{H}$ to denote an approximation of the *inverse* of the Hessian, while we use $\mathbf{B}$ to denote an approximation of the Hessian. We denote the usual *Frobenius* norm as $\|\cdot\|$. Moreover, for any square matrix $\mathbf{A} \in \mathbb{R}^{d\times d}$ and any positive definite matrix $\mathbf{W} \in \mathbb{R}^{d\times d}$, we define the norm $\|\mathbf{A}\|_{\mathbf{W}}$ as

$$\|\mathbf{A}\|_{\mathbf{W}} = \|\mathbf{W}^{\frac{1}{2}}\mathbf{A}\mathbf{W}^{\frac{1}{2}}\|. \tag{4}$$

We often use the matrices $\mathbf{X}$, $\mathbf{G} \in \mathbb{R}^{d\times m+1}$, that concatenates the iterates and their gradients as follow,

$$\mathbf{X} = [\mathbf{x}_i, \dots, \mathbf{x}_{i+m}], \quad \mathbf{G} = [\nabla f(\mathbf{x}_i), \dots, \nabla f(\mathbf{x}_{i+m})].$$

Also, we define $\mathbf{C}$, and $\Delta\mathbf{X}$ and $\Delta\mathbf{G}$ as

$$\Delta\mathbf{X} = \mathbf{X}\mathbf{C}, \qquad \Delta\mathbf{G} = \mathbf{G}\mathbf{C},$$

where $\mathbf{C} \in \mathbb{R}^{m+1\times m}$ is a matrix of rank $m-1$ such that $\mathbf{1}_{m+1}^T\mathbf{C} = 0$, $\mathbf{1}_{m+1}$ being a vector of size $m+1$ full of ones. Typically, $\mathbf{C}$ is the column-difference matrix

$$\mathbf{C} = \begin{bmatrix} -1 & 0 & 0 & \dots \\ 1 & -1 & 0 & \dots \\ 0 & 1 & -1 & \dots \\ & \ddots & \ddots & \ddots \\ & & & 1 \ -1 \\ & & & 0 \quad 1 \end{bmatrix}.$$

1.2 Related work

The idea of updating an approximation of the Hessian or its inverse can be traced back to Davidon [1959, 1991] with the DFP update. Several updates, such as the Broyden method [Broyden, 1965] or the BFGS method [Broyden, 1970, Fletcher, 1970, Goldfarb, 1970, Shanno, 1970] have been proposed since then. Notably, Dembo et al. [1982], Dembo and Steihaug [1983] proposed to approximately invert the Hessian using a Conjugate Gradient method. Limited memory BFGS (L-BFGS) [Liu and Nocedal, 1989], where a limited number of vectors are stored for the approximation of the Hessian, has proven to be a powerful type of quasi-Newton method. The use of multisecant equations has also been used in a different context by Gower and Gondzio [2014] and Hennig [2015], and their connection with Anderson Acceleration [Anderson, 1965] was studied by [Fang and Saad, 2009]. This connection, combined with recent results on Anderson Acceleration [Toth and Kelley, 2015, Walker and Ni, 2011, Rohwedder and Schneider, 2011, Scieur et al., 2016, 2018], especially in the stochastic [Scieur et al., 2017] and non-smooth [Zhang et al., 2018] settings, may indicates that multisecant methods also enjoy some good theoretical properties. To scale up second-order methods, recent works focus on stochastic quasi-Newton methods. The use of stochastic quasi-Newton updates has been investigated by Schraudolph et al. [2007], Mokhtari and Ribeiro [2015], Moritz et al. [2016], Byrd et al. [2016] and Gower et al. [2016], while approximating the Hessian through sampling methods has been proposed by Erdogdu and Montanari [2015], Xu et al. [2016] and Agarwal et al. [2017], among others. In contrast to Gower et al. [2016] and Jahani et al. [2020], our approach never compute the exact Hessian. We now present two popular quasi-Newton updates: the BFGS method, and the multi-secant Broyden method. They will serve as a basis to motivate the needs of generalization of quasi-Newton updates.

1.2.1 Single secant DFP/BFGS updates

The BFGS update finds a symmetric matrix $\mathbf{H}_k$ that satisfies the secant equation (3). Among the many possible solutions, it selects the one closest to $\mathbf{H}_{k-1}$ in

a weighted Frobenius norm (4), specifically,

$$\begin{aligned}\mathbf{H}_k = &\operatorname*{argmin}_{\mathbf{H}=\mathbf{H}^T} \|\mathbf{H} - \mathbf{H}_{k-1}\|_{\mathbf{W}} \\ &\text{s.t.} \quad \mathbf{H}(\nabla f(\mathbf{x}_k) - \nabla f(\mathbf{x}_{k-1})) = \mathbf{x}_k - \mathbf{x}_{k-1}.\end{aligned} \tag{5}$$

where $\mathbf{W}$ is *any* positive definite matrix such that $\mathbf{W}(\nabla f(\mathbf{x}_k) - \nabla f(\mathbf{x}_{k-1})) = \mathbf{x}_k - \mathbf{x}_{k-1}$ [Nocedal and Wright, 1999, §8.1] — a similar claim holds for the update formula of $\mathbf{B}_k$, known as DFP, whose update reads

$$\begin{aligned}\mathbf{B}_k = &\operatorname*{argmin}_{\mathbf{B}=\mathbf{B}^T} \|\mathbf{B} - \mathbf{B}_{k-1}\|_{\mathbf{W}^{-1}} \\ &\text{s.t.} \quad \mathbf{B}(\mathbf{x}_k - \mathbf{x}_{k-1}) = \nabla f(\mathbf{x}_k) - \nabla f(\mathbf{x}_{k-1}).\end{aligned} \tag{6}$$

The matrix is then inverted using the Woodbury matrix identity. In the two update rules, the matrices $\mathbf{W}$ and $\mathbf{W}^{-1}$ are used implicitly, i.e., we do not need to form $\mathbf{W}$ to evaluate $\mathbf{H}_k$ nor $\mathbf{B}_k$.

Solving (5) repeatedly, BFGS builds a sequence $\mathbf{H}_1, \mathbf{H}_2, \ldots$ of matrices such that each $\mathbf{H}_k$ satisfies the kth secant equation. While it may satisfy the $k-1$ other secants approximately, the update rule offers no such guarantees. The same holds for the DFP update.

1.2.2 Multi-secant Broyden updates

In the case of Broyden updates, we seek a matrix $\mathbf{B}$ for the type-I, or $\mathbf{H}$ for the type-II, that satisfies the secant equations only, without any restriction on the symmetry of the estimate. The update of the standard Broyden method reads, for $i = k-m, \ldots, k$,

$$\begin{aligned}\mathbf{B}_k = &\operatorname*{argmin}_{\mathbf{B}} \|\mathbf{B} - \mathbf{B}_{k-m}\| \\ &\text{s.t.} \quad \mathbf{B}(\mathbf{x}_i - \mathbf{x}_{i-1}) = \nabla f(\mathbf{x}_i) - \nabla f(\mathbf{x}_{i-1}), \\ \mathbf{H}_k = &\operatorname*{argmin}_{\mathbf{H}} \|\mathbf{H} - \mathbf{H}_{k-m}\| \\ &\text{s.t.} \quad \mathbf{H}\big(\nabla f(\mathbf{x}_i) - \nabla f(\mathbf{x}_{i-1})\big) = \mathbf{x}_i - \mathbf{x}_{i-1}.\end{aligned} \tag{7}$$

As for the DFP update, the matrix $\mathbf{B}_k$ can also be inverted cheaply. In [Fang and Saad, 2009], the authors show how to extend this update to the case where we want to satisfy more than one secant equation. However, its solution is generally not symmetric.

1.3 Contributions

Quasi-Newton methods approximate the Hessian. The previous section shows they do this in very different ways that seem incompatible given the work of Schnabel [1983]. Despite their differences, they share similarities, such as the idea of secant equations. This leads to the following questions:

> *Is it possible to design a generalized framework for quasi-Newton updates encompassing Broyden's, DFP and BFGS schemes?*
>
> *Can Symmetric and Multisecant techniques be combined into a single update?*

Our work proposes a positive answer to these questions trough the following contributions.

- We propose a general framework that models and generalizes previous quasi-Newton updates.
- We derive new quasi-Newton update rules (Algorithm 1), which are symmetric and take into account *several secant equations*. The bottleneck is an (economic size) Singular Value Decomposition (SVD), whose complexity is linear in the dimension of the problem d and quadratic in the memory size m (this term is minor since we assume $m \ll d$), therefore comparable to other quasi-Newton methods.
- We show the optimality of the convergence rate of any multisecant quasi-Newton update built using our framework, on quadratic functions *without line search*. This improves over the BFGS and DFP updates as they are inefficient with unitary step size on quadratics [Powell, 1986], and suboptimal if exact line-search is not used.
- We introduce novel *robust updates*, that provably reduce the sensitivity to the noise of our quasi-Newton schemes. This robustness property is a direct consequence of considering several secant equations at once.

Organization of the paper In Section 2 we list the desirable properties of quasi-Newton schemes, and end with a generic quasi-Newton update. The choice of its parameters, like the loss/regularization functions, the preconditioner, the number of secants or the initialization leads to different, existing methods but also to potentially new ones. Then, Section 3 proposes a novel quasi-Newton scheme (Algorithm 1) based on our framework, combining the ideas of DFP/BFGS and multisecant Broyden methods. This algorithm has the advantage of presenting a regularization term, which controls the stability of the update.

2 Generalization of Quasi-Newton

We have seen in the previous section two different quasi-Newton (qN) updates: one that focuses on the *symmetry* of the estimate, the other on the number of satisfied *secant equations*. The idea presented in Sections 2.1 and 2.2, if taken separately, are not novel. However, we propose a framework that unifies existing and new qN schemes in Section 2.3. This generalization gives a novel view on how qN updates are

Algorithm 1 Type-I Symmetric Multisecant step

(See Appendix A for the type-II version)

Input: Function f and gradient ∇f, initial approximation of the Hessian $\mathbf{B}_{\text{ref}}$, maximum memory m (can be ∞), relative regularization parameter $\bar{\lambda}$.

1: Compute $g_0 = \nabla f(x_0)$ and perform the initial step

$$\mathbf{x}_1 = \mathbf{x}_0 - \mathbf{B}_0^{-1}\boldsymbol{g}_0$$

2: **for** $t = 1, 2, \ldots$ **do**

3: Form the matrices $\Delta\mathbf{X}$ and $\Delta\mathbf{G}$ (see Section 1.1) using the m last pairs $(\mathbf{x}_i, \nabla f(\mathbf{x}_i))$.

4: Compute the quasi-Newton direction $\mathbf{d}$ as

$$\mathbf{d}_t = -\mathbf{Z}_\star^{-1} g_t,$$

see (Inv-RSP) with $\mathbf{A} = \Delta\mathbf{X}$, $\mathbf{D} = \Delta\mathbf{G}$, $\mathbf{Z}_{\text{ref}} = \mathbf{B}_{\text{ref}}$, $\lambda = \bar{\lambda}\|\mathbf{A}\|$.

5: Perform an approximate-line search

$$\mathbf{x}_{t+1} = \mathbf{x}_t + h_t\mathbf{d}_t, \quad h_t \approx \arg\min_h f\big(\mathbf{x}_t + h_t\mathbf{d}_t\big).$$

6: **end for**

built. We also provide a unified convergence result on quadratics in Section 2.5

2.1 Generalized (Multi-)Secant Equations

The central part of qN methods is the secant equation. The idea follows from the linearization of the gradient of the objective function. Indeed, consider the function $f(\mathbf{x})$, assumed to be smooth, strongly convex and twice differentiable. The linearization of its gradient around the minimum $\mathbf{x}_\star$ satisfies

$$\nabla f(\mathbf{x}) \approx \underbrace{\nabla f(\mathbf{x}_\star)}_{=0} + \nabla^2 f(\mathbf{x}_\star)(\mathbf{x} - \mathbf{x}_\star). \tag{8}$$

After a "Newton step", we get

$$\mathbf{x} - [\nabla^2 f(\mathbf{x}^\star)]^{-1}\nabla f(\mathbf{x}) \approx \mathbf{x}_\star.$$

Unfortunately, we do not have access to the matrix $\nabla^2 f(\mathbf{x}^\star)$ as we do not know $\mathbf{x}_\star$. Moreover, solving the linear system $[\nabla^2 f(\mathbf{x}^\star)]^{-1}\nabla f(\mathbf{x})$ may be costly when d is large.

To overcome such issues, consider a sequence $\{\mathbf{x}_0, \ldots, \mathbf{x}_m\}$ of points at which we have computed the gradients. Then, (8) can be stated as

$$\mathbf{G} = \nabla^2 f(\mathbf{x}_\star)(\mathbf{X} - \mathbf{X}_\star),$$

where $\mathbf{X}_\star = \mathbf{x}_\star \mathbf{1}_{m+1}^T$, i.e., the matrix concatenating $m+1$ copies of the vector $\mathbf{x}_\star$. Matrices $\mathbf{X}$ and $\mathbf{G}$ are defined in Section 1.1.

Ideally, the estimate $\mathbf{B}$ of the Hessian, or the estimate of its inverse $\mathbf{H}$, has to satisfy the condition

$$\mathbf{G} = \mathbf{B}(\mathbf{X} - \mathbf{X}_\star) \quad \text{or} \quad \mathbf{H}\mathbf{G} = (\mathbf{X} - \mathbf{X}_\star).$$

However, the dependency on $\mathbf{x}_\star$ makes the problem of estimating $\mathbf{B}$ or $\mathbf{H}$ intractable. To remove this problematic dependency, consider a matrix $\mathbf{C} \in \mathbb{R}^{m+1\times m}$ of rank m such that $\mathbf{1}_{m+1}^T\mathbf{C} = 0$ (see Section 1.1 for an example). After multiplying by $\mathbf{x}_\star$ on the right, we simplify $\mathbf{X}_\star\mathbf{C} = 0$ and we obtain the *multisecant equations*

$$\Delta\mathbf{G} = \mathbf{B}\Delta\mathbf{X}, \quad \text{or} \quad \mathbf{H}\Delta\mathbf{G} = \Delta\mathbf{X}, \tag{9}$$

where $\Delta\mathbf{X}$ and $\Delta\mathbf{G}$ are defined in Section 1.1. In the specific case where we have only one secant equation, (9) corresponds exactly to the standard secant equation in (5). In the case where $\mathbf{C}$ is the column-difference operator, we obtain the multisecant equations usually used in multisecant Broyden methods.

2.2 Regularization and Constraints

The matrices $\mathbf{B}$ (Broyden Type-I and DFP updates) and $\mathbf{H}$ (Broyden Type-II and BFGS) are selected so as to minimize the distances w.r.t. the reference matrices, called $\mathbf{B}_{\text{ref}}$ and $\mathbf{H}_{\text{ref}}$ respectively, as shown in (7). In the case where there is only a sequence of single secant equations, the reference matrix is taken as being the previous estimate, with an arbitrary initialization. In the case of a multisecant update, the reference matrix is arbitrary. Moreover, in the case of DFP and BFGS, we have in addition a *symmetry* constraint, restraining even more the search space for the estimate of the Hessian. For simplicity, we will consider only the type-I update here, i.e., the estimate $\mathbf{B}$. The formulation for estimate $\mathbf{H}$ can be easily derived by swapping $\Delta\mathbf{G}$ and $\Delta\mathbf{X}$.

The intuition behind the regularization term is due to the number of degrees of freedom in the problem. The secant equation $\mathbf{B}\Delta\mathbf{X} = \Delta\mathbf{G}$ defines the behavior of the operator $\mathbf{B}$, mapping from $\mathbf{span}\{\Delta\mathbf{X}\}$ to $\mathbf{span}\{\Delta\mathbf{G}\}$. However, the dimension of these two spans is as most $m < d$. This means we have to define the behavior of $\mathbf{B}$ *outside* $\mathbf{span}\{\Delta\mathbf{X}\}$ and $\mathbf{span}\{\Delta\mathbf{G}\}$, i.e., from $\mathbf{span}\{\Delta\mathbf{X}\}^\perp$ to $\mathbf{span}\{\Delta\mathbf{G}\}^\perp$.

Since $\mathbf{B}$ outside the span is not driven by the secant equations, we have to define an operator $\mathbf{B}_{\text{ref}}$, characterizing the default behavior of $\mathbf{B}$ outside the span of secant equations. This means, that in the case where $\mathbf{B}$ satisfies exactly the secant equations, $\mathbf{B}$ reads

$$\mathbf{B} = [\Delta\mathbf{G}\Delta\mathbf{X}^\dagger] + \Theta(\mathbf{I} - \mathbf{P}),$$

where $\mathbf{P}$ is the projector to the span of $\Delta\mathbf{X}$, $\Delta\mathbf{X}^\dagger$ is a pseudo-inverse of $\Delta\mathbf{X}$, and Θ depends on $\mathbf{B}_{\text{ref}}$ and constraints (different Θ lead to different qN updates). In this way, $\mathbf{B}$ satisfies the secant equation, since multiplying $\mathbf{B}$ by $\Delta\mathbf{X}$ gives $\Delta\mathbf{G}$,

$$\mathbf{B}\Delta\mathbf{X} = \Delta\mathbf{G}\Delta\mathbf{X}^\dagger\Delta\mathbf{X} + \Theta(\mathbf{I} - \mathbf{P})\Delta\mathbf{X}.$$

We have $\mathbf{P}\Delta\mathbf{X} = \Delta\mathbf{X}$, thus $(\mathbf{I} - \mathbf{P})\Delta\mathbf{X} = 0$ (by construction of $\mathbf{P}$). Moreover, $\Delta\mathbf{G}\Delta\mathbf{X}^\dagger\Delta\mathbf{X} = \Delta\mathbf{G}$ by definition of the pseudo-inverse.

The way $\mathbf{B}$ behaves outside the span is thus driven by Θ, which depends on the regularization, the initialization $\mathbf{B}_{\text{ref}}$ and the constraints. To make a parallel with machine learning problems, Θ can be seen as the "generalization" (or "out-of-sample") term. We give example choices for Θ in Appendix E.6.

Consider the regularisation function $\mathcal{R}(\cdot, \mathbf{B}_{\text{ref}})$, assumed to be strictly-convex, whose minimum is attained at $\mathbf{B}_{\text{ref}}$, and the convex constraint set $\mathcal{C}$. We can write the qN update estimation problem as

$$\min_{\mathbf{B}\in\mathcal{C}} \mathcal{R}(\mathbf{B}, \mathbf{B}_{\text{ref}}) \quad \text{subject to } \mathbf{B}\Delta\mathbf{X} = \Delta\mathbf{G}. \tag{10}$$

This approach generalizes the way we define qN updates. Indeed, for instance, we recover DFP by setting $\mathcal{R} = \|\mathbf{B} - \mathbf{B}_{\text{ref}}\|_{\mathbf{W}^{-1}}$, $\mathcal{C} = \mathbb{S}^{d\times d}$ (the set of symmetric matrices), $m = 1$ and $\mathbf{B}_{\text{ref}} = \mathbf{B}_{k-1}$ in (10). We also recover the Type-I Broyden method by setting $\mathcal{R} = \|\mathbf{B} - \mathbf{B}_{\text{ref}}\|$ and $\mathcal{C} = \mathbb{R}^{d\times d}$.

2.3 Generalized QN Update

A natural extension, given the updates of DFP/BFGS and multisecant Broyden, would be the symmetric multi-secant update. This update would read, for an arbitrary regularization function,

$$\min_{\mathbf{B}\in\mathbb{S}^{d\times d}} \mathcal{R}(\mathbf{B}, \mathbf{B}_{\text{ref}}) \quad \text{subject to } \mathbf{B}\Delta\mathbf{X} = \Delta\mathbf{G}.$$

In the case where $m > 1$, this multisecant technique seems promising as it combines the advantages of multisecant Broyden and symmetric updates.

Assuming $\Delta\mathbf{X}, \Delta\mathbf{G}$ have full column rank, these equations always have a solution $\mathbf{B}$. However, there exists a *symmetric* solution *if and only if* $\Delta\mathbf{X}^T\Delta\mathbf{G}$ is symmetric [Schnabel, 1983, Henk Don, 1987].

When $\Delta\mathbf{X}^T\Delta\mathbf{G}$ is symmetric, Schnabel [1983] derived a multisecant BFGS update rule. This assumption indeed holds for quadratic objectives, but not for general objective functions when $m \geq 2$, that is, when we consider more than one secant condition [Schnabel, 1983, Example 3.1]. Hence, a naive extension of symmetric quasi-Newton update leads to infeasible problems.

To tackle the problem of infeasible updates, we can relax the constraint on the secant equations by a *loss function* $\mathcal{L}(\cdot, \Delta\mathbf{X}, \Delta\mathbf{G})$. We finally end up with the *generalized (type-I and type-II) qN update*

$$\mathbf{B}_k = \lim_{\lambda\to 0} \operatorname*{arg\,min}_{\mathbf{B}\in\mathcal{C}} \mathcal{L}(\mathbf{B}, \Delta\mathbf{X}, \Delta\mathbf{G}) + \frac{\lambda}{2}\mathcal{R}(\mathbf{B}, \mathbf{B}_{\text{ref}}) \tag{GQN-I}$$

$$\mathbf{H}_k = \lim_{\lambda\to 0} \operatorname*{arg\,min}_{\mathbf{H}\in\mathcal{C}} \mathcal{L}(\mathbf{H}, \Delta\mathbf{G}, \Delta\mathbf{X}) + \frac{\lambda}{2}\mathcal{R}(\mathbf{H}, \mathbf{H}_{\text{ref}}) \tag{GQN-II}$$

where we assume that $\mathcal{L}$ and $\mathcal{R}$ are strictly convex, and sufficiently simple to have an explicit formula for $\mathbf{H}_k$. The limits here simply state that we first minimize the loss function, then with the remaining degrees of freedom we minimize the regularization term. In the case where the update (10) is feasible, (GQN-I)/(GQN-II) and (10) are equivalent.

2.4 Preconditioning

As shown for instance in DFP and BFGS, it is common to use a preconditioner to reduce the dependence of the update to the units of the Hessian. We give here the example for type-II update. The type-I follows immediately by considering $\mathbf{W}^{-1}$ instead of $\mathbf{W}$.

The idea of preconditioning is, instead of considering $\mathbf{H}$, to set

$$\mathbf{M} = \mathbf{W}^{(1-\alpha)}\mathbf{H}\mathbf{W}^{\alpha},$$

where $\mathbf{W}$ ideally has the same units as the *Hessian* of the function f. For example, in BFGS, $\mathbf{W}$ is *any* matrix such that $\mathbf{W}\Delta\mathbf{X} = \Delta\mathbf{G}$, which always exists in the case where $\Delta\mathbf{X}$ and $\Delta\mathbf{G}$ are vectors. Ideally, the preconditioner cancels the units in the update rules, i.e., $\mathbf{W}$ has to have the same units as the Hessian.

In the case where we consider a preconditioner,

$$\mathbf{M}\mathbf{W}^{-\alpha}\Delta\mathbf{X} = \mathbf{W}^{1-\alpha}\Delta\mathbf{G}, \quad \mathbf{M}_{\text{ref}} = \mathbf{W}^{\alpha-1}\mathbf{H}_{\text{ref}}\mathbf{W}^{-\alpha}.$$

We now have the *type-II Preconditioned Generalized Quasi-Newton* update

$$\operatorname*{arg\,min}_{\mathbf{M}\in\tilde{\mathcal{C}}} \quad \mathcal{L}(\mathbf{M}, \mathbf{W}^{-\alpha}\Delta\mathbf{X}, \mathbf{W}^{(1-\alpha)}\Delta\mathbf{G}) + \frac{\lambda}{2}\mathcal{R}(\mathbf{M}, \mathbf{M}_{\text{ref}}) \tag{PGQN-II}$$

where $\tilde{\mathcal{C}} = \mathbf{W}^{(1-\alpha)}\mathcal{C}\mathbf{W}^{\alpha}$, i.e., the image of the constraint after application of the preconditioner. To retrieve the update $\mathbf{H}$, it suffices to solve

$$\mathbf{H} = \mathbf{W}^{-(1-\alpha)}\mathbf{M}\mathbf{W}^{-\alpha}.$$

2.5 Rate of Convergence on Quadratics

Our theorem below shows that generalized qN methods (GQN-I) and (GQN-II) are optimal on quadratics under mild assumptions, in the sense that their performance is comparable to conjugate gradients.

Theorem 1. *Consider any multisecant quasi-Newton method* (GQN-II) *with unitary step size and* $m = \infty$*,*

$$\boldsymbol{x}_{k+1} = \boldsymbol{x}_k - \boldsymbol{H}_k\nabla f(\boldsymbol{x}_k) \tag{11}$$

where f is the quadratic form $(\boldsymbol{x}-\boldsymbol{x}_\star)^T\frac{\boldsymbol{Q}}{2}(\boldsymbol{x}-\boldsymbol{x}_\star)$ for some $Q\succ 0$, and $\boldsymbol{H}$ satisfies exactly the secant equations. If the update (11) *is a preconditioned first-order method, i.e., there exists a symmetric positive definite matrix $\tilde{\boldsymbol{H}}$ independent of k such that*

$$\boldsymbol{x}_{k+1}\in\boldsymbol{x}_0+\tilde{\boldsymbol{H}}\mathbf{span}\{\nabla f(\boldsymbol{x}_0),\dots,\nabla f(\boldsymbol{x}_k)\}$$

then $\boldsymbol{x}_k=\boldsymbol{x}_\star$ if $k\geq d+1$; for smaller k the method satisfies the rate

$$\|\nabla f(\boldsymbol{x}_k)\|\leq\mathcal{O}\Big(\tfrac{1-\sqrt{\kappa}}{1+\sqrt{\kappa}}\Big)^k\|\nabla f(\boldsymbol{x}_0)\|,$$

Where κ is the inverse of the condition number of $\tilde{\boldsymbol{H}}\boldsymbol{Q}$.

Proof. See Appendix E for a detailed proof. □

Notice that, for instance, the multisecant Broyden updates (7) or the multisecant BFGS update [Schnabel, 1983] satisfies the assumptions of Theorem 1 if $\mathbf{B}_{\text{ref}}$ or $\mathbf{H}_{\text{ref}}$ are symmetric positive definite matrices (see Appendix E.6). For all these methods, we have $\tilde{\mathbf{H}}=\mathbf{H}_{\text{ref}}$ (or $\mathbf{B}_{\text{ref}}^{-1}$). This indicates that the initialization is crucial, since a good initial approximation of $\mathbf{Q}^{-1}$ drastically reduces the condition number κ.

We have now a generic form of qN update, but it raises some important questions. Specifically, which practical losses and regularization functions should we use, and what happens if λ does not go to zero? The next section addresses the first point by giving an example that extends (limited memory) DFP and multi-secant Broyden methods. Then, we analyse the robustness of the method when λ is non-zero.

3 Robust Symmetric Multisecant Updates

This section proposes a novel quasi-Newton scheme, that extends the BFGS and multisecant Broyden method into the type-II Symmetric Multisecant Update (12) below, solving the problem (PGQN-II) in the special case where the loss and the regularization are Frobenius norms. For simplicity, we do not consider any preconditioner here. The method reads

$$\mathbf{H}_k=\underset{\mathbf{H}=\mathbf{H}^T}{\arg\min}\,\|\mathbf{H}\Delta\mathbf{G}-\Delta\mathbf{X}\|^2+\frac{\lambda}{2}\|\mathbf{H}-\mathbf{H}_{\text{ref}}\|^2 \quad (12)$$

and its type-I counterpart is $\mathbf{B}_k^{-1}$, where

$$\mathbf{B}_k=\underset{\mathbf{B}=\mathbf{B}^T}{\arg\min}\,\|\mathbf{B}\Delta\mathbf{X}-\Delta\mathbf{G}\|^2+\frac{\lambda}{2}\|\mathbf{B}-\mathbf{B}_{\text{ref}}\|^2 \quad (13)$$

Explicit Formula We now solve problem (12) efficiently. This is an extension of the *symmetric Procrustes problem* from [Higham, 1988]. Indeed, Higham [1988] solves the problem

$$\min_{\mathbf{Z}=\mathbf{Z}^T}\|\mathbf{ZA}-\mathbf{D}\|,$$

where $\mathbf{A}$ and $\mathbf{D}$ are $\mathbb{R}^{d\times m}$ matrices, where $m>d$. In our case, we have $m\ll d$, and an extra regularization term, that makes the update formula more complicated. Fortunately, the matrix-vector multiplication $\mathbf{Zv}$ can still be done efficiently even in our case, the bottleneck being the computation of the SVD of a thin matrix. The next theorem details the explicit formula to compute $\mathbf{H}_k$ (and its inverse if one wants to use a type-I method).

Theorem 2. *Consider the Regularized Symmetric Procrustes* (RSP) *problem*

$$\boldsymbol{Z}_\star=\underset{\boldsymbol{Z}=\boldsymbol{Z}^T}{\arg\min}\,\|\boldsymbol{ZA}-\boldsymbol{D}\|^2+\frac{\lambda}{2}\|\boldsymbol{Z}-\boldsymbol{Z}_{ref}\|^2, \quad \text{(RSP)}$$

where $\boldsymbol{Z}_{ref}$ is symmetric (otherwise, take the symmetric part of $\boldsymbol{Z}_{ref}$), $\boldsymbol{Z}$, $\boldsymbol{Z}_{ref}\in\mathbb{R}^{d\times d}$, and $\boldsymbol{A}$, $\boldsymbol{D}\in\mathbb{R}^{d\times m}$, $m\leq d$. Then, the solution $\boldsymbol{Z}_\star$ is given by

$$\boldsymbol{Z}_\star=\boldsymbol{V}_1\boldsymbol{Z}_1\boldsymbol{V}_1^T+\boldsymbol{V}_1\boldsymbol{Z}_2+\boldsymbol{Z}_2^T\boldsymbol{V}_1^T+(\boldsymbol{I}-\boldsymbol{P})\boldsymbol{Z}_{ref}(\boldsymbol{I}-\boldsymbol{P}) \quad \text{(Sol-RSP)}$$

where

$$\begin{aligned}[\boldsymbol{U},\Sigma,\boldsymbol{V}_1]&=\boldsymbol{SVD}(\boldsymbol{A}^T,\texttt{'econ'}),\ \textit{(economic SVD)}\\ \boldsymbol{Z}_1&=\boldsymbol{S}\odot\Big[\boldsymbol{V}_1^T\Big(\boldsymbol{A}\boldsymbol{D}^T+\boldsymbol{D}\boldsymbol{A}^T+\lambda\boldsymbol{Z}_{ref}\Big)\boldsymbol{V}_1\Big],\\ \boldsymbol{S}&=\frac{1}{\Sigma^2\boldsymbol{1}\boldsymbol{1}^T+\boldsymbol{1}\boldsymbol{1}^T\Sigma^2+\lambda\boldsymbol{1}\boldsymbol{1}^T},\\ \boldsymbol{P}&=\boldsymbol{V}_1\boldsymbol{V}_1^T,\\ \boldsymbol{Z}_2&=(\Sigma^2+\lambda\boldsymbol{I})^{-1}\boldsymbol{V}_1^T(\boldsymbol{A}\boldsymbol{D}^T+\lambda\boldsymbol{Z}_{ref})(\boldsymbol{I}-\boldsymbol{P}).\end{aligned}$$

The fraction in $\boldsymbol{S}$ stands for the element-wise inversion (Hadamard inverse), and the notation $\odot$ stands for the element-wise product (Hadamard product). The inverse $\boldsymbol{Z}_\star^{-1}$ reads

$$\begin{aligned}\boldsymbol{Z}_\star^{-1}&=\boldsymbol{E}\Big(\boldsymbol{Z}_1-\boldsymbol{Z}_2\boldsymbol{Z}_{ref}^{-1}\boldsymbol{Z}_2^T\Big)^{-1}\boldsymbol{E}^T+(\boldsymbol{I}-\boldsymbol{P})\boldsymbol{Z}_{ref}^{-1}(\boldsymbol{I}-\boldsymbol{P})\\ \boldsymbol{E}&=\boldsymbol{V}_1-(\boldsymbol{I}-\boldsymbol{P})\boldsymbol{Z}_{ref}^{-1}\boldsymbol{Z}_2^T.\end{aligned} \quad \text{(Inv-RSP)}$$

Proof. See Appendix F for a detailed proof. □

The type-I update uses the matrix $\mathbf{Z}_\star^{-1}$, using $\mathbf{A}=\Delta\mathbf{X}$ and $\mathbf{D}=\Delta\mathbf{G}$. The type-II uses instead $\mathbf{Z}_\star$, with $\mathbf{A}=\Delta\mathbf{G}$ and $\mathbf{D}=\Delta\mathbf{X}$. See Appendix I for an efficient Matlab implementation of the type-I and type-II updates.

The next proposition shows the complexity of performing one matrix-vector multiplication with $\mathbf{Z}_\star$ and its

inverse. The bottleneck of the method is the SVD of a $\mathbb{R}^{m\times d}$ matrix, whose complexity is $O(m^2 d)$, thus linear in the dimension.

Proposition 1. *The complexity of evaluating $\boldsymbol{Z}_\star \boldsymbol{v}$ and $\boldsymbol{Z}_\star^{-1}\boldsymbol{v}$ is $O(m^2 d)$, assuming $m \ll d$ and that the complexity of $\boldsymbol{Z}_{ref}\boldsymbol{v}$ and $\boldsymbol{Z}_{ref}^{-1}\boldsymbol{v}$ is at most $O(m^2 d)$.*

Robustness The symmetric multisecant update can be used in two different modes, one that lets $\lambda \to 0$, the other, biased but more robust, that sets $\lambda > 0$.

The update formula is slightly simpler when $\lambda = 0$. However, due to the presence of matrix inversion, this may lead to instability issues in some cases, similarly to the BFGS method when

$$(\mathbf{x}_{k+1} - \mathbf{x}_k)^T(\nabla f(\mathbf{x}_{k+1}) - \nabla f(\mathbf{x}_k)) \approx 0,$$

i.e., when the step and difference of gradients are close to being orthogonal. In BFGS, such issues are tackled by a filtering step, discarding the update if the scalar product goes below some threshold. Unfortunately, when the gradient is corrupted by some noise, the impact on the BFGS update can be huge.

In the case where $\lambda > 0$, we can show that our update is robust when $\mathbf{A}$ and $\mathbf{D}$ are corrupted.

Proposition 2. *Let $\boldsymbol{Z}_\star(\lambda)$ be defined as the solution of* (RSP) *for some λ, and $\boldsymbol{Z}_\star(0) = \lim_{\lambda\to 0} \boldsymbol{Z}_\lambda$. Let $\tilde{\boldsymbol{A}}$, $\tilde{\boldsymbol{D}}$ be a corrupted version of $\boldsymbol{A}$ and $\boldsymbol{D}$ where*

$$\|\boldsymbol{A} - \tilde{\boldsymbol{A}}\| \leq \delta_{\boldsymbol{A}}, \quad \|\boldsymbol{D} - \tilde{\boldsymbol{D}}\| \leq \delta_{\boldsymbol{D}}.$$

Finally, let $\tilde{\boldsymbol{Z}}_\star(\lambda)$ be the solution of (Sol-RSP) *using $\tilde{\boldsymbol{A}}$ and $\tilde{\boldsymbol{C}}$. Then, we have*

$$\|\tilde{\boldsymbol{Z}}_\star(\lambda) - \boldsymbol{Z}_\star(0)\| \leq \underbrace{\|\boldsymbol{Z}_\star(\lambda) - \boldsymbol{Z}_\star(0)\|}_{\textit{Bias}} + \underbrace{\|\tilde{\boldsymbol{Z}}_\star(\lambda) - \boldsymbol{Z}_\star(\lambda)\|}_{\textit{Stability}},$$

where

$$\|\boldsymbol{Z}_\star(\lambda) - \boldsymbol{Z}_\star(0)\| \leq \frac{\lambda\|\boldsymbol{Z}_\star(0) - \boldsymbol{Z}_{ref}\|}{\sigma_{\min}^2(\boldsymbol{A}) + \lambda}, \tag{14}$$

$$\|\tilde{\boldsymbol{Z}}_\star(\lambda) - \boldsymbol{Z}_\star(\lambda)\| \leq \mathcal{O}\left(\frac{\delta_{\boldsymbol{A}} + \delta_{\boldsymbol{D}}}{\lambda}\right). \tag{15}$$

Proof. See Appendix G for a detailed proof. □

Proposition 2 suggests that λ should satisfy a trade-off to achieve the best performing approximation. Notice that when $\lambda = 0$ in the noise-less case, we recover the optimal $\mathbf{Z}_\star$, and when $\lambda \to \infty$, we have $\mathbf{Z}_\star = \mathbf{Z}_{\text{ref}}$.

Our result is called *robust* as we can bound the maximum perturbation without restriction on its magnitude. This is *not* the case in [Higham, 1988], whose main assumption is $\delta_{\mathbf{A}} \leq \sigma_{\min}(\mathbf{A})$ (which is extremely restrictive), where $\sigma_{\min}$ is the smallest non-zero singular value of $\mathbf{A}$.

Since the singular values of $\mathbf{A}$ are, in practice, often small, it is always recommended to set a small λ: we will show latter, in the numerical experiments, that even for quadratic functions (i.e., in the "perturbation-free regime"), a small value of λ drastically changes the final result, as this makes the method robust to numerical noise.

Scaling of λ. The parameter λ has to be scaled w.r.t. the problem input. It is clear, from Theorem 2, that the role of λ is to regularize the matrix inversion by lower-bounding the eigenvalues of the inverted matrix. Therefore, we advise to set $\lambda = \bar{\lambda}\|\mathbf{A}^T\mathbf{A}\|_2$, i.e., proportional to $\|\mathbf{A}^T\mathbf{A}\|_2$. This way, assuming $\sigma_{\min}$ small, the conditioning of $(\mathbf{A}^T\mathbf{A}+\lambda\mathbf{I})^{-1}$ is upper-bounded by $1 + 1/\bar{\lambda}$.

4 Numerical Experiment

This section compares our symmetric multisecant algorithms to existing methods in the literature. We present in this section only a few experiments concerning stochastic-related experiments: We first compare the quality of the estimate of the Hessian (and its inverse). Then, we compare the speed of convergence when using this estimate to estimate the Newton-step in the case where the gradient is stochastic.

Hessian Recovery Consider the problem of recovering the inverse of a symmetric Hessian $\mathbf{Q}^{-1}$ of a quadratic function, that satisfies

$$\mathbf{Q}^{-1}\Delta\mathbf{G} = \Delta\mathbf{X}, \quad \mathbf{Q} = \mathbf{Q}^T.$$

However, we have only access to $\tilde{\Delta\mathbf{G}}$, a corrupted version of $\Delta\mathbf{G}$. This notably happens when the oracle provides stochastic gradients.

In our case, we consider the worst-case ℓ_2 corruption

$$\tilde{\Delta\mathbf{G}} = \mathbf{U}_{\Delta\mathbf{G}} \max\{\Sigma_{\Delta\mathbf{G}} - \epsilon \cdot \sigma_1(\Delta\mathbf{G}),\, 0\}\mathbf{V}_{\Delta\mathbf{G}}^T,$$

where $\mathbf{U}_{\Delta\mathbf{G}}\Sigma_{\Delta\mathbf{G}}\mathbf{V}_{\Delta\mathbf{G}}^T$ is the SVD of $\Delta\mathbf{G}$, and ϵ is the relative perturbation intensity. When $\epsilon = 1$, the matrix $\tilde{\Delta\mathbf{G}}$ is full of zeros.

We estimate $\mathbf{Q}^{-1}$ using different techniques, that we compare using the relative residual error

$$\text{error}(\mathbf{Q}_{\text{est}}^{-1}) = \|\mathbf{Q}_{\text{est}}^{-1}\Delta\mathbf{G} - \Delta\mathbf{X}\|/\|\Delta\mathbf{X}\|.$$

Note that, in our error function, we use the noise-free version of $\Delta\mathbf{G}$.

Our baseline is the diagonal estimate, corresponding to the inverse of the Lipchitz constant of $\mathbf{Q}$, typically used as a step size in the gradient method. We

compare ℓ-BFGS, Multisecant Broyden updates [Fang and Saad, 2009] and our Type-1 and Type-2 multisecant algorithms, solving respectively (Inv-RSP) and (Sol-RSP) with $\mathbf{A} = \Delta\tilde{\mathbf{G}}$, $\mathbf{D} = \Delta\mathbf{X}$, $\mathbf{B}_0 = \mathbf{H}_0^{-1} = \|\mathbf{Q}\|$. The number of secant equations is 50 and the dimension of the problem is 250. The results are reported in Figure 1. In this experiment, we used a worse-case noise to better show the differences between the methods. Stochastic noise also works, but makes the graph more dense and harder to read.

Optimization problem We aim to solve

$$\min_{\mathbf{x}\in\mathbb{R}^d} f(\mathbf{x}) \stackrel{\text{def}}{=} \frac{1}{N}\sum_{i=0}^{N} \ell(\mathbf{a}_i^T\mathbf{x}, \mathbf{b}_i) + \frac{\tau}{2}\|\mathbf{x}\|^2, \qquad (16)$$

where $\ell(\cdot,\cdot)$ is a loss function. The pair $(\mathbf{A}, \mathbf{b})$ is a dataset, where $\mathbf{a}_i \in \mathbb{R}^d$ is a data point composed by d features, and b_i is the label of the i^{th} data point.

Here, we present the specific cases where ℓ is either a quadratic loss or a logistic loss, on the Madelon [Guyon et al., 2008] dataset.

- **Quadratic loss, deterministic gradient.** See Figure 3.
- **Logistic loss, deterministic gradient.** See Figure 4.
- **Quadratic loss, stochastic gradient.** See Figure 5. We use SAGA [Defazio et al., 2014] to obtain the stochastic estimates of the gradient, with a batch size of 64.

We have other experiments on other datasets in Appendix H. We also show the evolution of the spectrum of $\mathbf{H}_k$ and $\mathbf{B}_k^{-1}$ in Figure 6, Appendix H.

5 Discussion and Future Directions

We briefly discuss our contributions and propose possible improvements. Although our approach performs sufficiently well to be competitive with current qN updates, the authors believe the method can be improved in several aspects.

Contrary to BFGS, the update (13) (resp. (12)) does not guarantee its positive-definiteness when applied to a smooth and strongly convex function. However, for large enough λ the matrix is p.s.d. given that $\mathbf{H}_{\text{ref}}$ (resp. $\mathbf{B}_{\text{ref}}$) is also positive-definite. Also, it is possible to project a small matrix in (Inv-RSP) (resp. (Sol-RSP)) to ensure positive definiteness. We discuss this in more details in Appendix B. The ideal way would be to solve the symmetric Procrustes problem with a semi-definite constraint, but this is still considered as an open problem [Higham, 1988].

A direct consequence of the non positive-definiteness is the lack of robustness guarantees for the Type-I method, that inverts a matrix that is possibly not positive definite. Therefore, it is probably impossible to bound the smallest eigenvalue, unless we use the robust projection trick in Appendix B. Surprisingly however, in our experiments the Type-I method seems to be the most stable among all updates.

Moreover, we considered here a plain method with *no preconditioner*. In BFGS and DFP updates, the preconditioner $\mathbf{W}$ is *any* matrix such that $\mathbf{W}\Delta\mathbf{X} = \Delta\mathbf{G}$ where $\Delta\mathbf{X}$ and $\Delta\mathbf{G}$ are *vectors*. This matrix is used implicitly in the update: all occurrences of $\mathbf{W}\Delta\mathbf{X}$ are replaced by $\Delta\mathbf{G}$, in a way that $\mathbf{W}$ disappears. We cannot use a similar trick here, since such matrices do not exist in general when $\Delta\mathbf{X}$ and $\Delta\mathbf{G}$ are matrices [Schnabel, 1983]. We propose in Appendix C possible options to include such preconditioners that may potentially improve the method.

It is also possible to consider a general qN step, that takes the direction $\mathbf{HGv}$ (or $\mathbf{B}^{-1}\mathbf{Gv}$), where $\mathbf{v}$ is a vector that sums to one, instead of taking the direction computed with the latest gradient, $\mathbf{H}\nabla f(\mathbf{x}_k)$. In the special case where $\mathbf{v}$ is full of zeros but one as the last element, this reduces to the standards qN step. We discuss this strategy in Appendix D, and we suspect this technique may reduce even more the impact of the noise on the qN step if $\mathbf{v}$ is chosen to be the averaging vector $\mathbf{1}_m/m$, for instance.

The complexity of the method is somewhat worse than current qN methods: $O(m^2d)$ instead of $O(md)$. The authors believe it may be possible to reduce the complexity by a factor m by using a low-rank SVD update [Brand, 2006] and by changing our direct formulas in Theorem 2 into recursive ones.

Another interesting direction is the study of the the matrix $\mathbf{C}$ that forms $\Delta\mathbf{X}$ and $\Delta\mathbf{G}$. We suspect that, in the case where those matrices are corrupted, choosing the right $\mathbf{C}$ may affect the stability of the method. For instance, it is possible to design $\mathbf{C}$ to set more weight on some selected secant equations that may be more recent, or that contain less noise.

We proposed a novel method with distinct theoretical properties, including symmetry, optimality on quadratics with *unitary step size*, and robustness, and which performs encouragingly well in practice. In view of the new questions that multisecant methods raise, we hope our work can add to efforts for the design of possibly other, better-performing quasi-Newton schemes.

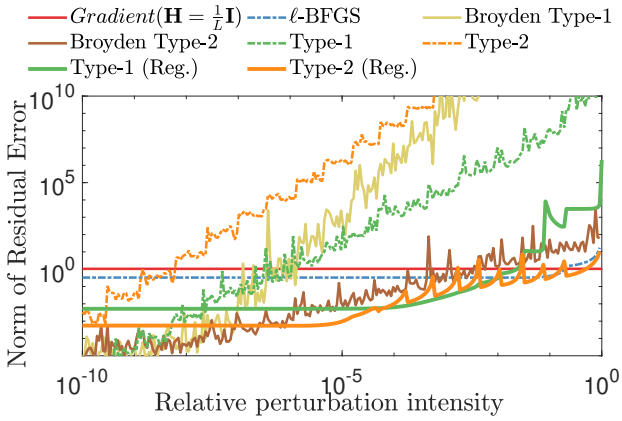

Figure 1: Comparison of different methods to estimate a symmetric matrix. We see that symmetric multisecant methods perform well in a small-noise regime, but quickly get out of control for larger perturbations. This is not the case for their regularized counterpart ($\lambda = 10^{-10}$), clearly showing a more stable behavior. BFGS performs poorly compared to multisecant algorithms, since it can only satisfy one secant equation at a time. The type-II Broyden method seems stable, but does not recover a symmetric matrix.

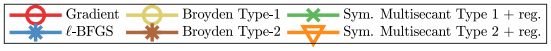

Figure 2: Legend for the numerical experiment. The methods proposed in this paper are `Sym. Multisecant Type I` (resp. `II`) `+reg`. The regularization parameter is set to $10^{-8}\|\Delta\mathbf{X}\|$ (resp. $10^{-8}\|\Delta\mathbf{G}\|$).

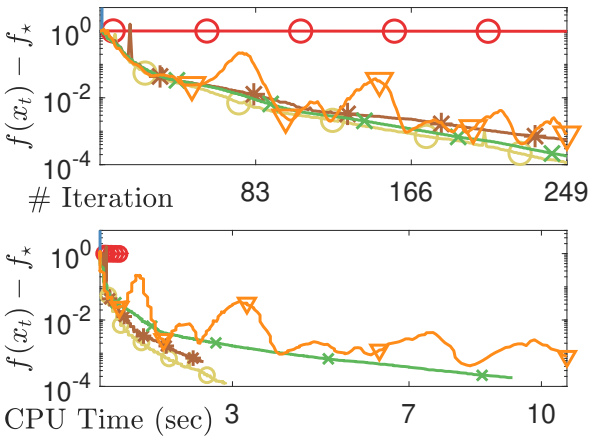

Figure 3: Solving a quadratic regression using the Madelon Dataset, where the regularization is set such that the condition number is equal to 10^{10}. Except for gradient descent, all methods use an *unitary step size* and use all previous iterates and gradients. As expected, all multisecant methods have similar rates of convergence, as they are optimal - which means that their rate of convergence is similar to that of conjugate gradients. On the other side, as there is no line-search, the BFGS algorithm diverges [Powell, 1986].

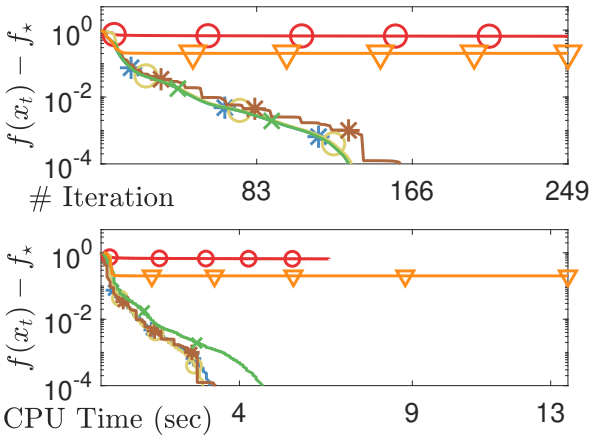

Figure 4: Solving a logistic regression using the Madelon Dataset, where the regularization is set such that the condition number is equal to 10^{10}. Except for gradient descent, all methods use an approximate line-search, and use a limited memory of last 25 secant equations. Except for the multisecant Type-II method, all methods converge at the same speed.

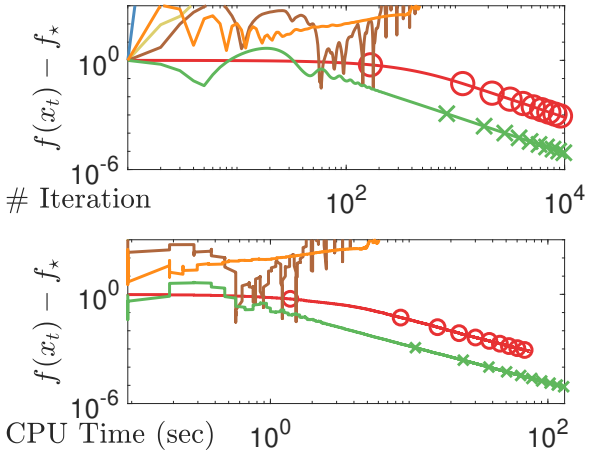

Figure 5: Comparison of the stability of qN methods with stochastic gradients on Madelon dataset. We report the function value of the average of the iterates. The batch size is 64. Since the function is stochastic, we used only unitary step sizes. The memory is 25, and the relative regularization $\bar{\lambda} = 10^{-2}$. The condition number is 10^3. $\ell-$BFGS and Broyden methods are divergent in this situation. With unitary step sizes, the regularized symmetric multisecant Type-I method is faster than stochastic gradient.

References

N. Agarwal, B. Bullins, and E. Hazan. Second-order Stochastic Optimization for Machine Learning in Linear Time. *J. Mach. Learn. Res.*, 18(1):4148–4187, Jan. 2017. ISSN 1532-4435.

D. G. Anderson. Iterative procedures for nonlinear integral equations. *Journal of the ACM (JACM)*, 12(4):547–560, 1965.

D. Ballabio, F. Grisoni, V. Consonni, and R. Todeschini. Integrated qsar models to predict acute oral systemic toxicity. *Molecular informatics*, 38(8-9):1800124, 2019.

A. S. Berahas, R. Bollapragada, and J. Nocedal. An investigation of newton-sketch and subsampled newton methods. *Optimization Methods and Software*, pages 1–20, 2020.

R. Bollapragada, D. Mudigere, J. Nocedal, H.-J. M. Shi, and P. T. P. Tang. A progressive batching l-bfgs method for machine learning. *arXiv preprint arXiv:1802.05374*, 2018.

R. Bollapragada, R. H. Byrd, and J. Nocedal. Exact and inexact subsampled newton methods for optimization. *IMA Journal of Numerical Analysis*, 39(2):545–578, 2019.

M. Brand. Fast low-rank modifications of the thin singular value decomposition. *Linear Algebra and its Applications*, 415:20–30, 05 2006. doi: 10.1016/j.laa.2005.07.021.

C. G. Broyden. A class of methods for solving nonlinear simultaneous equations. *Mathematics of computation*, 19(92):577–593, 1965.

C. G. Broyden. The Convergence of a Class of Double-Rank Minimization Algorithms. *Journal of the Institute of Mathematics and Its Applications*, 6:76–90, 09 1970. doi: 10.1093/imamat/6.3.222.

R. H. Byrd, H. F. Khalfan, and R. B. Schnabel. Analysis of a symmetric rank-one trust region method. *SIAM J. on Optimization*, 6(4):1025–1039, Apr. 1996. ISSN 1052-6234. doi: 10.1137/S1052623493252985. URL https://doi.org/10.1137/S1052623493252985.

R. H. Byrd, S. L. Hansen, J. Nocedal, and Y. Singer. A Stochastic Quasi-Newton Method for Large-scale Optimization. *SIAM Journal on Optimization*, 26:1008–1031, 2016.

S. A. Danziger, S. J. Swamidass, J. Zeng, L. R. Dearth, Q. Lu, J. H. Chen, J. Cheng, V. P. Hoang, H. Saigo, R. Luo, et al. Functional census of mutation sequence spaces: the example of p53 cancer rescue mutants. *IEEE/ACM transactions on computational biology and bioinformatics*, 3(2):114–125, 2006.

W. Davidon. Variable metric method for minimization. *Technical Report ANL 5990 (revised), Argonne National Laboratory, Argonne, Il*, 1959.

W. Davidon. Variable metric method for minimization. *SIAM Journal on Optimization*, 1:1–17, 1991.

A. Defazio, F. Bach, and S. Lacoste-Julien. Saga: A fast incremental gradient method with support for non-strongly convex composite objectives. In *Advances in neural information processing systems*, pages 1646–1654, 2014.

R. Dembo, S. Eisenstat, and T. Steihaug. Inexact Newton Methods. *SIAM Journal on Numerical Analysis*, 19(2):400–408, 1982. doi: 10.1137/0719025. URL https://doi.org/10.1137/0719025.

R. S. Dembo and T. Steihaug. Truncated-Newton algorithms for large-scale unconstrained optimization. *Mathematical Programming*, 26(2):190–212, Jun 1983. ISSN 1436-4646. doi: 10.1007/BF02592055. URL https://doi.org/10.1007/BF02592055.

M. A. Erdogdu and A. Montanari. Convergence rates of sub-sampled newton methods. In *Proceedings of the 28th International Conference on Neural Information Processing Systems - Volume 2*, NIPS'15, pages 3052–3060, Cambridge, MA, USA, 2015. MIT Press. URL http://dl.acm.org/citation.cfm?id=2969442.2969580.

H.-r. Fang and Y. Saad. Two classes of multisecant methods for nonlinear acceleration. *Numerical Linear Algebra with Applications*, 16(3):197–221, 2009.

R. Fletcher. A new approach to variable metric algorithms. *The Computer Journal*, 13(3):317?322, 1970.

D. Goldfarb. A family of variable-metric methods derived by variational means. *Mathematics of Computation*, 24(109):23?26, 1970.

R. M. Gower and J. Gondzio. Action constrained quasi-newton methods. *arXiv preprint arXiv:1412.8045*, 2014.

R. M. Gower, D. Goldfarb, and P. Richtárik. Stochastic Block BFGS: Squeezing More Curvature out of Data. In *ICML*, 2016.

I. Guyon, S. Gunn, M. Nikravesh, and L. A. Zadeh. *Feature extraction: foundations and applications*, volume 207. Springer, 2008.

F. Henk Don. On the symmetric solutions of a linear matrix equation. *Linear Algebra and its Applications*, 93:1–7, 07 1987. doi: 10.1016/S0024-3795(87)90308-9.

P. Hennig. Probabilistic interpretation of linear solvers. *SIAM Journal on Optimization*, 25(1):234–260, 2015. doi: 10.1137/140955501. URL https://doi.org/10.1137/140955501.

N. J. Higham. The symmetric Procrustes problem. *BIT*, 28, 03 1988. doi: 10.1007/BF01934701.

M. Jahani, M. Nazari, R. Tappenden, A. S. Berahas, and M. Takáč. Sonia: A symmetric blockwise truncated optimization algorithm. *arXiv preprint arXiv:2006.03949*, 2020.

N. Kushmerick. Learning to remove internet advertisements. In *Proceedings of the third annual conference on Autonomous Agents*, pages 175–181, 1999.

D. C. Liu and J. Nocedal. On the limited memory bfgs method for large scale optimization. *Mathematical Programming*, 45(1):503–528, Aug 1989. ISSN 1436-4646. doi: 10.1007/BF01589116. URL https://doi.org/10.1007/BF01589116.

A. Mokhtari and A. Ribeiro. Global Convergence of Online Limited Memory BFGS. *Journal of Machine Learning Research*, 16:3151–3181, 2015. URL http://jmlr.org/papers/v16/mokhtari15a.html.

P. Moritz, R. Nishihara, and M. Jordan. A Linearly-Convergent Stochastic L-BFGS Algorithm. In A. Gretton and C. C. Robert, editors, *Proceedings of the 19th International Conference on Artificial Intelligence and*

Statistics, volume 51 of *Proceedings of Machine Learning Research*, pages 249–258, Cadiz, Spain, 09–11 May 2016. PMLR. URL http://proceedings.mlr.press/v51/moritz16.html.

J. Nocedal and S. Wright. *Numerical optimization, Second Edition*. Springer Verlag, 1999.

M. J. Powell. How bad are the BFGS and DFP methods when the objective function is quadratic? *Math. Program.*, 34:34–47, 1986.

T. Rohwedder and R. Schneider. An analysis for the diis acceleration method used in quantum chemistry calculations. *Journal of mathematical chemistry*, 49(9):1889, 2011.

M. Schmidt. minfunc: unconstrained differentiable multivariate optimization in matlab, 2005. URL https://www.cs.ubc.ca/~schmidtm/Software/minFunc.html.

R. B. Schnabel. Quasi-newton methods using multiple secant equations. Technical report, University of Colorado Boulder, Computer Science Department, 1983.

N. N. Schraudolph, J. Yu, and S. Gunter. A Stochastic Quasi-Newton Method for Online Convex Optimization. In M. Meila and X. Shen, editors, *Proceedings of the Eleventh International Conference on Artificial Intelligence and Statistics*, volume 2 of *Proceedings of Machine Learning Research*, pages 436–443, San Juan, Puerto Rico, 21–24 Mar 2007. PMLR. URL http://proceedings.mlr.press/v2/schraudolph07a.html.

D. Scieur, A. d'Aspremont, and F. Bach. Regularized nonlinear acceleration. In *Advances In Neural Information Processing Systems*, pages 712–720, 2016.

D. Scieur, F. Bach, and A. d'Aspremont. Nonlinear acceleration of stochastic algorithms. In *Advances in Neural Information Processing Systems*, pages 3982–3991, 2017.

D. Scieur, E. Oyallon, A. d'Aspremont, and F. Bach. Online regularized nonlinear acceleration. *arXiv preprint arXiv:1805.09639*, 2018.

D. Shanno. Conditioning of Quasi-Newton Methods for Function Minimization. *Mathematics of Computing*, 24:647–656, 07 1970. doi: 10.1090/S0025-5718-1970-0274029-X.

A. Toth and C. Kelley. Convergence analysis for anderson acceleration. *SIAM Journal on Numerical Analysis*, 53 (2):805–819, 2015.

H. F. Walker and P. Ni. Anderson acceleration for fixed-point iterations. *SIAM Journal on Numerical Analysis*, 49(4):1715–1735, 2011.

M. A. Woodbury. Inverting modified matrices. *Memorandum Rept. 42, Statistical Research Group, Princeton University, Princeton, NJ*, 1950.

P. Xu, J. Yang, F. Roosta-Khorasani, C. Ré, and M. W. Mahoney. Sub-sampled newton methods with non-uniform sampling. In D. D. Lee, M. Sugiyama, U. V. Luxburg, I. Guyon, and R. Garnett, editors, *Advances in Neural Information Processing Systems 29*, pages 3000–3008. Curran Associates, Inc., 2016. URL http://papers.nips.cc/paper/6037-sub-sampled-newton-methods-with-non-uniform-sampling.pdf.

J. Zhang, B. O'Donoghue, and S. Boyd. Globally convergent type-i anderson acceleration for non-smooth fixed-point iterations. *arXiv preprint arXiv:1808.03971*, 2018.