A variational derivation of a class of BFGS-like methods

Michele Pavon[a]

[a]Dipartimento di Matematica "Tullio Levi-Civita", Università di Padova, via Trieste 63, 35121 Padova, Italy.

ARTICLE HISTORY
Compiled September 21, 2018

ABSTRACT
We provide a maximum entropy derivation of a new family of BFGS-like methods. Similar results are then derived for block BFGS methods. This also yields an independent proof of a result of Fletcher 1991 and its generalisation to the block case.

KEYWORDS
Quasi-Newton method, BFGS method, maximum entropy problem, block BFGS.

1. Introduction

Suppose $f:\mathbb{R}^n\rightarrow\mathbb{R}$ is a C^2 function to be minimized. Then Newton's iteration is

$$x_{k+1}=x_k-[H(x_k)]^{-1}\nabla f(x_k),\quad k\in\mathcal{N},\tag{1}$$

where $H(x_k)=\nabla^2 f(x_k)$ is the Hessian of f at the point x_k. In quasi-Newton methods, one employs instead an approximation B_k of $H(x_k)$ to avoid the costly operations of computing, storing and inverting the Hessian (B_0 is often taken to be the identity I_n). These methods appear to perform well even in nonsmooth optimization, see [1]. Instead of (1), one uses

$$x_{k+1}=x_k-\alpha_k B_k^{-1}\nabla f(x_k),\quad \alpha_k>0,\quad k\in\mathcal{N},\tag{2}$$

with α_k chosen by a line search, imposing the *secant* equation

$$y_k=B_{k+1}s_k,\tag{3}$$

where

$$y_k:=\nabla f(x_k+s_k)-\nabla f(x_k),\quad s_k:=\Delta x_k=x_{k+1}-x_k.$$

Email: pavon@math.unipd.it

The secant condition is motivated by the expansion

$$\nabla f(x_k + s_k) \approx \nabla f(x_k) + H(x_k)s_k. \tag{4}$$

For $n > 1$, B_{k+1} satisfying (3) is underdetermined. Various methods are used to find a symmetric B_{k+1} that satisfies the secant equation (3) and is closest in some metric to the current approximation B_k. In several methods, B_{k+1} or its inverse is a rank one or two update of the previous estimate [2].

Since for a strongly convex function the Hessian $H(x_k)$ is a symmetric positive definite matrix, we can think of its approximation B_k as a covariance of a zero-mean, multivariate Gaussian distribution. Recall that in the case of two zero-mean multivariate normal distributions p, q with nonsingular $n \times n$ covariance matrixes P, Q, respectively, the relative entropy (divergence, Kullback-Leibler index) can be derived in closed form

$$\mathbb{D}(p||q) = \int \log \frac{p(x)}{q(x)} p(x)dx = \frac{1}{2}\left[\log\det\left(P^{-1}Q\right) + tr(Q^{-1}P) - n\right].$$

Since P^{-1} and Q^{-1} are the natural parameters of the Gaussian distributions, we write

$$\mathbb{D}(P^{-1}||Q^{-1}) = \frac{1}{2}\left[\log\det\left(P^{-1}Q\right) + \operatorname{trace}(Q^{-1}P) - n\right] \tag{5}$$

2. A maximum entropy problem

Consider minimizing $\mathbb{D}(B^{-1}||B_k^{-1})$ over symmetric, positive definite B subject to the secant equation

$$B^{-1}y_k = s_k. \tag{6}$$

In [3], Fletcher indeed showed that the solution to this variational problem is provided by the BFGS iterate thereby providing a variational characterization for it alternative to Goldfarb's classical one [4], [2, Section 6.1]. We take a different approach leading to a family of BFGS-like methods.

First of all, observe that $B^{-1}y_k$ must be the given vector s_k. Thus, it seems reasonable that B_{k+1}^{-1} should approximate B_k^{-1} only in directions different from y_k. We are then led to consider the following new problem

$$\min_{\{B=B^T, B>0\}} \mathbb{D}(B^{-1}||P_k^T B_k^{-1} P_k) \tag{7}$$

subject to (6), where P_k is a rank $n-1$ matrix satisfying $P_k y_k = 0$, subject to the secant equation (6). One possible choice for P_k is the orthogonal projection

$$P_k = I_n - \frac{y_k y_k^T}{y_k^T y_k} = I_n - \Pi_{y_k}.$$

Since $P_k B_k^{-1} P_k$ is singular, however, (7) does not make sense. Thus, to regularize the problem, we replace P_k with the nonsingular, positive definite matrix $P_k^\epsilon = P_k + \epsilon I_n$.

The Lagrangian for this problem is

$$\mathcal{L}(B,\lambda) = \frac{1}{2}\left[\log\det\left(B^{-1}(P_k^\epsilon)^{-1}B_k P_k^\epsilon\right) + tr\left(P_k^\epsilon B_k^{-1} P_k^\epsilon B\right) - n\right] + \lambda_k^T[Bs_k - y_k] =$$
$$\frac{1}{2}\left[\log\det\left(B^{-1}B_k\right) + \frac{1}{2}\log\det\left((P_k^\epsilon)^{-2}\right) + tr\left(P_k^\epsilon B_k^{-1} P_k^\epsilon B\right) - n\right] + \lambda_k^T[Bs_k - y_k].$$

Observe that the term

$$\frac{1}{2}\log\det\left((P_k^\epsilon)^{-2}\right)$$

does not depend on B and therefore plays no role in the variational analysis. To compute the first variation of $\mathcal{L}$ in direction δB, we first recall a simple result. Consider the map J defined on nonsingular, $n\times n$ matrices M by $J(M) = \log|\det[M]|$. Let $\delta J(M;\delta M)$ denote the directional derivative of J in direction $\delta M \in \mathbb{R}^{n\times n}$. We then have the following result :

Lemma 2.1. [5, Lemma 2] *If M is nonsingular then, for any $\delta M \in \mathbb{R}^{n\times n}$,*

$$\delta J(M;\delta M) = \text{trace}[M^{-1}\delta M].$$

Observe also that any positive definite matrix B is an interior point in the cone $\mathcal{C}$ of positive semidefinite matrices in any symmetric direction $\delta B \in \mathbb{R}^{n\times n}$. Imposing $\delta\mathcal{L}(B,\lambda;\delta B) = 0$ for all such δB, we get, in view of Lemma 2.1,

$$\text{trace}\left[\left(-(B_{k+1}^\epsilon)^{-1} + P_k^\epsilon B_k^{-1} P_k^\epsilon + 2s_k\lambda_k^T\right)\delta B\right] = 0, \quad \forall \delta B,$$

which gives

$$(B_{k+1}^\epsilon)^{-1} = P_k^\epsilon B_k^{-1} P_k^\epsilon + 2s_k\lambda_k^T. \tag{8}$$

As $\epsilon \searrow 0$, we get the iteration

$$B_{k+1}^{-1} = P_k B_k^{-1} P_k + 2s_k\lambda_k^T. \tag{9}$$

Since $P_k y_k = 0$, in order to satisfy the secant equation

$$B_{k+1}^{-1} y_k = s_k.$$

it suffices to choose the multiplier λ_k so that

$$2\lambda_k^T y_k = 1.$$

We need, however, to also guarantee symmetry and positive definiteness of the solution. We are then led to choose λ_k as

$$\lambda_k = \frac{s_k}{2y_k^T s_k}. \tag{10}$$

Finally, notice that, under the *curvature* assumption

$$y_k^T s_k > 0, \tag{11}$$

if $B_k > 0$, indeed B_{k+1} in (9) is symmetric, positive definite justifying the previous calculations. We have therefore established the following result.

Theorem 2.2. *Assume $B_k > 0$ and $y_k^T s_k > 0$. A solution B^* of*

$$\min_{\{B=B^T, B>0\}} \mathbb{D}(B^{-1} || P_k^T B_k^{-1} P_k),$$

subject to constraint (6)*, in the regularized sense described above, is given by*

$$(B^*)^{-1} = \left(I_n - \frac{y_k y_k^T}{y_k^T y_k}\right) B_k^{-1} \left(I_n - \frac{y_k y_k^T}{y_k^T y_k}\right) + \frac{s_k s_k^T}{y_k^T s_k}. \tag{12}$$

3. BFGS-like methods

From Theorem 2.2, we get the following quasi-Newton iteration:

$$\begin{aligned} x_{k+1} &= x_k - \alpha_k B_k^{-1} \nabla f(x_k), \quad x_0 = \bar{x}, & (13)\\ B_{k+1}^{-1} &= \left(I_n - \frac{y_k y_k^T}{y_k^T y_k}\right) B_k^{-1} \left(I_n - \frac{y_k y_k^T}{y_k^T y_k}\right) + \frac{s_k s_k^T}{y_k^T s_k}, \quad B_0 = I_n. & (14) \end{aligned}$$

Note that, for limited-memory iterations, this method has the same storage requirement as standard limited-memory BFGS, say $(s_j, y_j), j = k, k-1, \ldots, k-m+1$. Now let $v_k \in \mathbb{R}^n$ be any vector not orthogonal to y_k. Then

$$P_k(v_k) := \frac{y_k v_k^T}{y_k^T v_k} \tag{15}$$

is an oblique projection onto y_k. Employing $P_k(v_k)$ and its transpose in place of Π_{y_k} in (7) and performing the variational analysis after regularisation, we get a BFGS-like iteration

$$B_{k+1}^{-1} = (I_n - P_k(v_k))^T B_k^{-1} (I_n - P_k(v_k)) + \frac{s_k s_k^T}{y_k^T s_k} \tag{16}$$

In particular, if $v_k = s_k$, the corresponding oblique projection is

$$P_k(s_k) = \frac{y_k s_k^T}{y_k^T s_k}.$$

In such case, (16) is just the standard (BFGS) iteration for the inverse approximate Hessian

$$B_{k+1}^{-1} = \left(I_n - \frac{y_k s_k^T}{y_k^T s_k}\right)^T B_k^{-1} \left(I_n - \frac{y_k s_k^T}{y_k^T s_k}\right) + \frac{s_k s_k^T}{y_k^T s_k}. \tag{17}$$

Here $T_k = I_n - P_k(s_k)$ is a rank $n-1$ matrix satisfying $T_k y_k = 0$ as is $I - \Pi_{y_k}$. We now get an alternative derivation of Fletcher's result [3].

Corollary 3.1. *Assume $B_k > 0$ and $y_k^T s_k > 0$. A solution B^* of*

$$\min_{\{B=B^T, B>0\}} \mathbb{D}(B^{-1}||B_k^{-1}),$$

subject to constraint (6) *is given by the standard* (BFGS) *iteration* (17).

Proof. We show that in the limit, as $\epsilon \searrow 0$, $\mathbb{D}(B^{-1}||B_k^{-1})$ and $\mathbb{D}\left(B^{-1}||\left(I_n - \frac{y_k s_k^T}{y_k^T s_k} + \epsilon I_n\right)^T B_k^{-1}\left(I_n - \frac{y_k s_k^T}{y_k^T s_k} + \epsilon I_n\right)\right)$ only differ by terms not depending on B. Indeed,

$$\begin{aligned}
&\mathbb{D}\left(B^{-1}||\left(I_n - \frac{y_k s_k^T}{y_k^T s_k} + \epsilon I_n\right)^T B_k^{-1}\left(I_n - \frac{y_k s_k^T}{y_k^T s_k} + \epsilon I_n\right)\right)\\
&= \frac{1}{2}\left\{\log\det\left(B^{-1}B_k\right) + \log\det\left[\left(I_n - \frac{y_k s_k^T}{y_k^T s_k} + \epsilon I_n\right)^{-1}\left(I_n - \frac{y_k s_k^T}{y_k^T s_k} + \epsilon I_n\right)^{-T}\right]\right.\\
&\left.+ \operatorname{trace}\left[\left((1+\epsilon)I_n - \frac{y_k s_k^T}{y_k^T s_k}\right)^T B_k^{-1}\left((1+\epsilon)I_n - \frac{y_k s_k^T}{y_k^T s_k}\right)B\right] - n\right\}
\end{aligned}$$

Note that, by the circulant property of the trace,

$$\operatorname{trace}\left[-\frac{s_k y_k^T}{y_k^T s_k} B_k^{-1}(1+\epsilon)B\right] = \operatorname{trace}\left[-B\frac{s_k y_k^T}{y_k^T s_k} B_k^{-1}(1+\epsilon)\right]$$

It now suffices to observe that, for symmetric matrices B satisfying (6) $Bs_k = y_k$, the products

$$B\frac{s_k y_k^T}{y_k^T s_k} = \frac{y_k s_k^T}{y_k^T s_k}B = \frac{y_k y_k^T}{y_k^T s_k}$$

are independent of B. □

Iterations (13)-(14) and (13)-(16) are expected to enjoy the same convergence properties as the canonical BFGS method [2, Chapter 6]. They can, in principle, be applied also to nonsmooth cases along the lines of [1] with an exact line search to compute α_k at each step.

4. Block BFGS-like methods

In some large dimensional problems, it is prohibitive to calculate the full gradient at each iteration. Consider for instance *deep neural networks*. A deep network consists of a nested composition of a linear transformation and a nonlinear one σ. In the learning phase of a deep network, one compares the predictions $y(x, \xi^i)$ for the input sample ξ^i with the actual output y^i. This is done through a cost function $f_i(x)$, e.g.

$$f_i(x) = \|y^i - y(x; \xi^i)\|^2.$$

The goal is to learn the *weights* x through minimization of the empirical loss function

$$f(x) = \frac{1}{N}\sum_{i=1}^{N} f_i(x).$$

In modern datasets, N can be in the millions and therefore calculation of the full gradient $\frac{1}{N}\sum_{i=1}^{N}\nabla f_i(x)$ at each iteration to perform gradient descent is unfeasible. One can then resort to *stochastic gradients* by sampling uniformly from the set $\{1, \ldots, N\}$ the index i_k where to compute the gradient at iteration k. In alternative, one can also average the gradient over a set of randomly chosen samples called a "mini-batch". In [6], a so-called block BFGS was proposed. Let S_k be a *sketching matrix* of directions [6] and let $\mathcal{T} \subset [N]$. Rather than taking differences of random gradients, one computes the action of the sub-sampled Hessian on S_k as

$$Y_k := \frac{1}{|\mathcal{T}|}\sum_{i\in\mathcal{T}} \nabla^2 f_i(x_k) S_k$$

To update B_k^{-1}, we can now consider the problem

$$\min_{\{B=B^T, B>0\}} \mathbb{D}\left(B^{-1} || P_k^T B_k^{-1} P_k\right) \tag{18}$$

where $I - P_k$ projects onto the space spanned by the columns of Y_k, subject to the block-secant equation

$$B^{-1} Y_k = S_k. \tag{19}$$

Again, one possible choice for S_k is $I - \Pi_{Y_k}$ where $\Pi_{Y_k} = Y_k(Y_k^T Y_k)^{-1} Y_k^T$ is the orthogonal projection. The same variational argument as in Section 2 leads to the iteration

$$B_{k+1}^{-1} = (I - \Pi_{Y_k})\, B_k^{-1}\, (I - \Pi_{Y_k}) + S_k(S_k^T Y_k)^{-1} S_k^T. \tag{20}$$

Another choice for P_k is the oblique projection $I - Y_k(S_k^T Y_k)^{-1} S_k^T$ leading to the iteration in [6]

$$B_{k+1}^{-1} = \left(I - Y_k(S_k^T Y_k)^{-1} S_k^T\right)^T B_k^{-1} \left(I - Y_k(S_k^T Y_k)^{-1} S_k^T\right) + S_k(S_k^T Y_k)^{-1} S_k^T. \tag{21}$$

We then obtain a variational characterisation of the iteration (21) alternative to the one of [6, Appendix A] and generalizing Fletcher [3].

Corollary 4.1. *Assume $B_k > 0$ and $S_k^T Y_k > 0$. A solution B^* of*

$$\min_{\{B=B^T, B>0\}} \mathbb{D}(B^{-1}||B_k^{-1}),$$

subject to constraint (19) is given by B_{k+1} in (21).

The proof is analogous to the proof of Corollary 3.1.

5. Numerical Experiments

The algorithm (13)-(14) has the form:

procedure BFGS-LIKE($f, Gf, x_0, tolerance$)
    $B \leftarrow I_d$ $\triangleright$ d is the dimension of x_0 and I_d is the identity in R^d
    $x \leftarrow x_0$
    for $n = 1, ..., MaxIterations$ **do**
        $y \leftarrow Gf(x)$
        if $||y|| < tolerance$ **then**
            break
        $SearchDirection \leftarrow -By$
        $\alpha \leftarrow LineSearch(f, GF, x, SearchDirection)$
        $\Delta x \leftarrow \alpha\, SearchDirection$
        $S \leftarrow I_d - \frac{yy^T}{y^T y}$
        $B \leftarrow S^T B S + \frac{\Delta x \Delta x^T}{y^T dx}$
        $x \leftarrow x + \Delta x$
    return x

Algorithm 1: BFGS-like algorithm (13)-(14)

While the effectiveness of the BFGS-like algorithms introduced in Section 3 needs to be tested on a significant number of large scale benchmark problems, we provide below two examples where the BFGS-like algorithm (13)-(14) appears to perform better than standard BFGS. Consider the strictly convex function f on $\mathbb{R}^2$

$$f(x_1, x_2) = e^{x_1-1} + e^{-x_2+1} + (x_1 - x_2)^2$$

whose minimum point is $x^* \approx (0.8, 1.2)$. Take as starting point: $(5, -7)$. Figure 1 illustrates the decay of the error $||x^n - x^*||_2$ over 50 iterations for the classical BFGS and for algorithm (13)-(14).

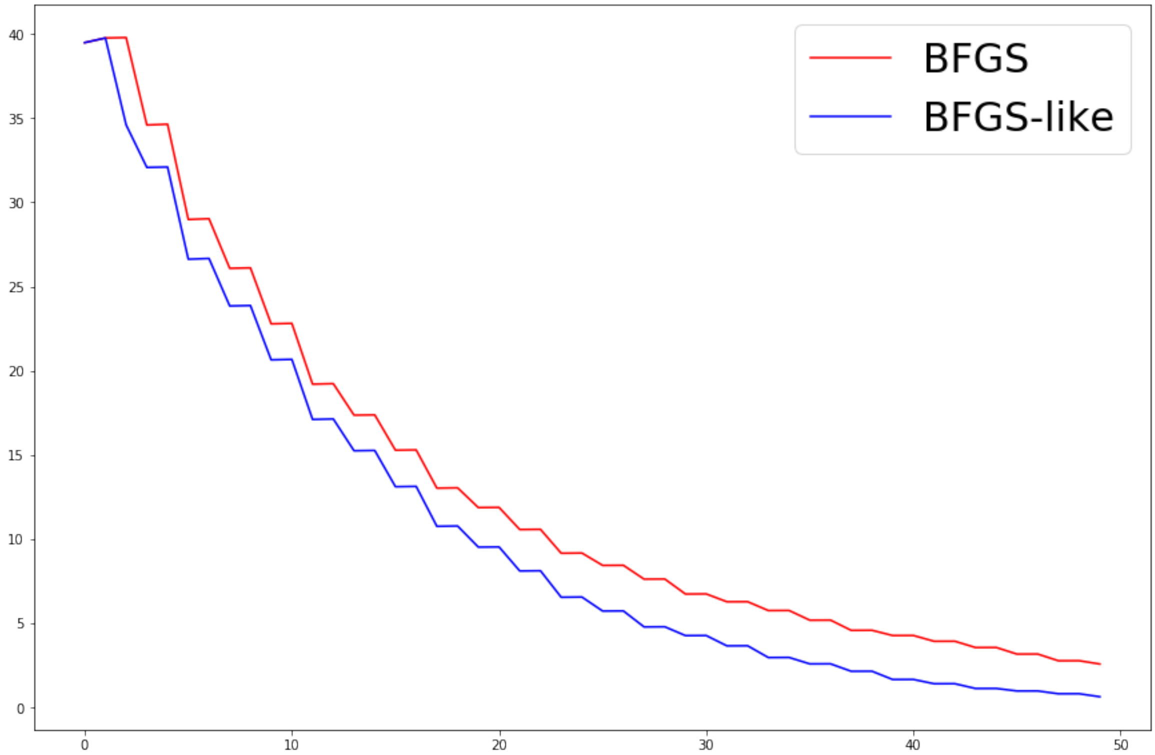

Figure 1. Plot of $||x^n - x^*||_2$ for each iteration n

Consider now the (nonconvex) Generalized Rosenbrock function in 10 dimensions:

$$f(x) = \sum_{i=1}^{9} \left[100 \left(x_{i+1} - x_i^2 \right)^2 + (x_i - 1)^2 \right], \quad -30 \le x_i \le 30, \ i = 1, 2, \ldots, 10.$$

It has an absolute minumum at $x_i^* = 1, i = 1, \ldots, 10$ and $f(x^*) = 0$. Taking as initial point $x_0 = (0, 0, \ldots, 0)$ the origin, both methods get stuck in a local minimum, see Figure 2.

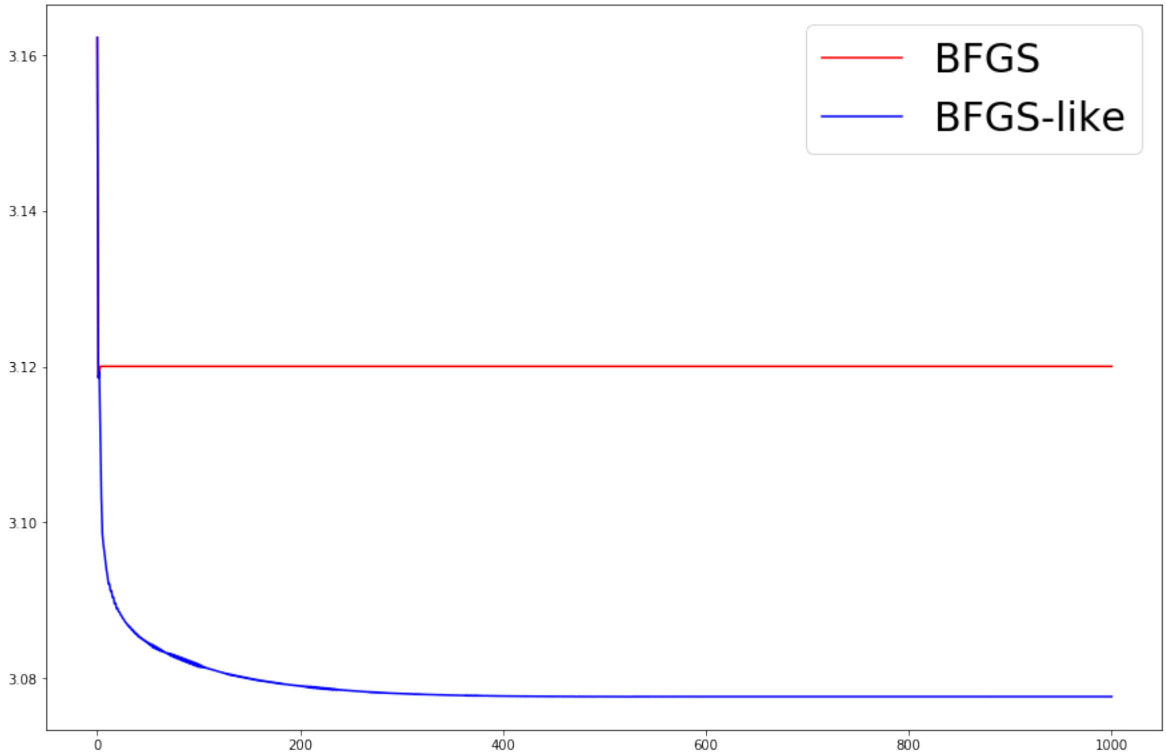

Figure 2. Plot of $||x^n - x^*||_2$ for each iteration n

Instead, initiating the recursions at $x_0 = (0.9, 0.9, \ldots, 0.9)$, both algorithms converge to the absolute minimum (Figure 3 depicts 100 iterations). After a few initial steps, BFGS-like appears to perform better than BFGS.

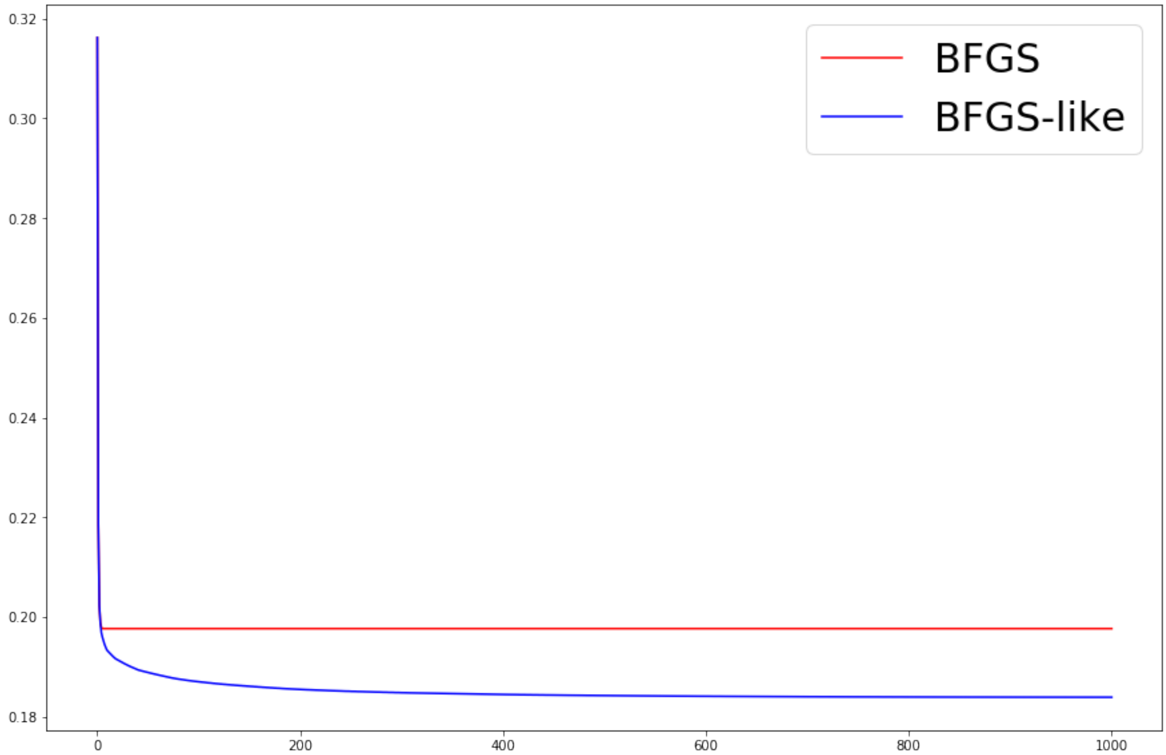

Figure 3. Plot of $||x^n - x^*||_2$ for each iteration n

6. Closing comments

We have proposed a new family of BFGS-like iterations of which (13)-(14) is a most natural one. The entropic variational derivation provides theoretical support for these methods and a new proof of Fletcher's classical derivation [3]. Further study is needed to exploit the flexibility afforded by this new family (the vector v_k determining the oblique projection in (15) appears as a "free parameter"). Similar results have been established for block BFGS. A few numerical experiments seem to indicate that (13)-(14) may perform better in some problems than standard BFGS.

Acknowledgments

This paper was written during a stay at the Courant Institute of Mathematical Sciences of the New York University whose hospitality is gratefully acknowledged. In particular, I would like to thank Michael Overton and Esteban Tabak for useful conversations and for pointing out some relevant literature. I would also like to thank Montacer Essid for kindly providing the code and the numerical examples of Section 5.

Funding

Supported in part by the Università di Padova Research Project CPDA 140897.

References

[1] A.S. Lewis and M.L. Overton, Nonsmooth Optimization via Quasi-Newton Methods *Math. Programming* **141** (2013), pp. 135-163.

[2] J. Nocedal and S. J. Wright, *Nonlinear Optimization*, 2nd edn. Springer, New York, 2006.

[3] R. Fletcher, A New Variational Result for Quasi-Newton Formulae, *SIAM J. Optimiz.*, 1991, **1**, No. 1 : pp. 18-21.

[4] D. Goldfarb, A family of variable metric methods derived by variational means, *Math. Comp.*, **24**, (1970), pp. 23-26.

[5] A. Ferrante and M. Pavon, Matrix Completion *à la* Dempster by the Principle of Parsimony, *IEEE Trans. Information Theory*, **57**, Issue 6, June 2011, 3925-3931.

[6] W. Gao and D. Goldfarb, Block BFGS Methods, preprint arXiv:1609.00318.