Almost multisecant quasi-Newton method

Mokhwa Lee MOKHWA.LEE@STONYBROOK.EDU
Yifan Sun YIFAN.0.SUN@GMAIL.COM
Stony Brook University

Abstract

Quasi-Newton (QN) methods provide an alternative to second-order techniques for solving minimization problems by approximating curvature. This approach reduces computational complexity as it relies solely on first-order information, and satisfying the secant condition. This paper focuses on multi-secant (MS) extensions of QN for convex optimization problems, which enhances the Hessian approximation at low cost. Specifically, we use a low-rank perturbation strategy to construct an almost-secant QN method that maintains positive definiteness of the Hessian estimate, which in turn helps ensure constant descent (and reduces method divergence). Our results show that careful tuning of the updates greatly improve stability and effectiveness of multisecant updates.

1. Introduction

We consider the unconstrained minimization problem

$$\underset{x}{\text{minimize}} \quad f(x) \tag{1}$$

where $f : \mathbb{R}^n \to \mathbb{R}$ is convex, in $\mathcal{C}^2$, and bounded below. Newton's method iteratively solves the linear system of order n to get a search direction p_k,

$$\nabla^2 f(x_k) p_k = \nabla f(x_k)$$

where $\nabla^2 f(x_k)$ is the Hessian and $\nabla f(x_k)$ is the gradient. In this case, the next iterate is updated as

$$x_{k+1} = x_k - \alpha p_k$$

where $\alpha > 0$ is a step length parameter. However, when dealing with large-scale problems, getting the Hessian matrix and solving (1) is not computationally scalable. For this reason, Quasi-Newton (QN) methods, like BFGS, are introduced and become good substitutes which efficiently approximate the Hessian with simple operations performed on successive gradient vectors.

Specifically, we investigate a series of multisecant quasi-Newton methods for minimizing (1), via repeated iterations

$$x_{k+1} = x_k - \alpha H_k^{-1} \nabla f(x_k) \tag{2}$$

where H_k serves as a Hessian approximation of f at x_k and satisfies multiple secant conditions

$$H_k(x_i - x_j) = \nabla f(x_i) - \nabla f(x_j) \tag{3}$$

for some subset of $i \neq j \in \{k, k-1, ..., k-p+1\}$ where p is the number of previous information taken into account. In high dimensional cases, where $n > 2p$, such updates are nonunique because the number of variables to define H_k is more than the number of constraints.

Methods of this type are referred to as *multisecant quasi-Newton methods*, because they approximate the Hessian through satisfying the multiple secant equations in (3). The main advantage of such methods is that they exploit second order information using only first order oracles, and do not in general require matrix inversion. In addition, limited memory versions exist which significantly reduce storage limits. Thus, they are often superior to gradient methods in smooth, but very ill-conditioned problems.

Perhaps the most well-known family of single-secant quasi-Newton methods are Broyden's method [1], Powell's method, Davidson-Fletcher-Powell (DFP) [10], and BFGS named after the concurrent works of Broyden [1], Fletcher [2], Goldfarb [3], and Shanno [4]. The *multi-secant* extensions were first explored not long later; [6] for Broyden's method, and [7] for extensions of Broyden's, Powell's method, DFP, and BFGS updates. Gay and Schnabel [8] provided an improved version of Broyden's method for inverse Hessian update. These methods also attempt to progressively include desired features, such as 1. fast and cheap updates, 2. symmetry, and 3. positive definiteness. However, the addition of these features is much less straightforward in the multisecant case; for this reason, multisecant methods are primarily used to solve quadratic systems, where symmetric positive semidefinite updates of multi-secant DFP and BFGS are easier to guarantee. *However, for general convex optimization problems, multi-secant quasi-Newton methods do not ensure descent.*

Later, a generalized framework [11] of the Broyden's method was also provided in which a block of secant conditions can be satisfied at each iteration which gives the flexibility to the rank of update on the inverse Hessian. Fang and Saad [9] also proposed the generalization of Broyden's and Multisecant family with several successful techniques for handling QN-type problems. More recently, closely related works include Gao et al. [14], Liu et al. [13], and Mokhtari[12]. These are higher rank update schemes that use only first-order information, and are shown to achieve q-superlinear convergence, at least in the local sense.

In this work, we explore various techniques to impose symmetric and positive semidefinite updates in multisecant DFP and BFGS through carefully tuned perturbations, for ill-conditioned non-quadratic problems. We compare these techniques against the perturbation methods presented in the seminal work [7].

2. Preliminaries

2.1. Single-secant quasi-Newton methods

The well-known single-secant quasi-Newton methods are DFP [10] and BFGS [1–4] which maintain B_k to be symmetric or positive semidefinite:

$$B_{k+1} = B_k + \frac{(y_k - Bs_k)y_k^T + y_k(y_k - B_k s_k)^T}{y_k^T s_k} - \frac{y_k(y_k - B_k s_k)^T s_k y_k^T}{(y_k^T s_k)^2}$$

(Davidon, Fletcher, Powell, 1991)

$$B_{k+1} = B_k + \frac{y_k y_k^T}{y_k^T s_k} - \frac{B_k s_k s_k^T B_k}{(s_k^T B_k s_k)} \qquad \text{(Broyden, Fletcher, Goldfarb, Shanno,1970)}$$

where the Hessian approximation update B_{k+1} satisfy the (single)secant condition

$$B_{k+1} \underbrace{(x_{k+1} - x_k)}_{s_k} \approx \underbrace{\nabla f(x_{k+1}) - \nabla f(x_k)}_{y_k}. \tag{4}$$

The secant condition is derived from the Taylor's second order expansion and its differential

$$\nabla f(x_{k+1}) \approx \nabla f(x_k) + \nabla^2 f(x_k)(x_{k+1} - x_k).$$

where $B_{k+1} \approx \nabla^2 f(x_{k+1})$. If we restrict B_{k+1} to be symmetric, then the equation (4) has $\frac{n(n+1)}{2} - n$ degrees of freedom where $n \geq 1$. If $n = 1$, then (4) has a unique solution, however, $n > 1$ case explains why there are many variations of quasi-Newton methods. After computing B_{k+1}, each quasi-Newton method will update x_{k+1} at each iteration

$$x_{k+1} = x_k - \alpha B_k^{-1} \nabla f(x_k).$$

To guarantee that each step taken is in a descent direction, the following

$$-\nabla f_k^T B_k^{-1} \nabla f_k < 0 \tag{5}$$

should be satisfied. If B_{k+1} is not positive semidefinite, (5) is no longer satisfied and hence the algorithm will not be guaranteed to monotonically decrease at each iteration. Therefore, maintaining positive semidefinite Hessian approximation B_{k+1} is an important key for quasi-Newton methods.

2.2. multisecant quasi-Newton methods

Schnabel [7] explained four typical multisecant quasi-Newton methods. Firstly, we consider two choices for s_i and y_i: the "curve-hugging" version for $i = k, ..., k - p + 1$ such that

$$s_i = x_{i+1} - x_i, \quad y_i = \nabla f(x_{i+1}) - \nabla f(x_i)$$

and the "anchored at most recent" version for $i = k - 1, ..., k - p$ such that

$$s_i = x_{k+1} - x_i, \quad y_i = \nabla f(x_{k+1}) - \nabla f(x_i).$$

Basically, both are interpolating the same previous point and this is explained well in Schnabel's paper. For the simplicity, we will use the former 'curve-hugging' version from now on.

We want to extend single-secant version to *multi-secant* by considering p previous points where $p > 1$, more than one column vectors for s_k and y_k. We create matrix version of iterative and derivative difference matrices, S_k and Y_k respectively, by

$$S_k = \begin{bmatrix} | & | & & | \\ s_{k-p} & s_{k-p+1} & \dots & s_k \\ | & | & & | \end{bmatrix}, \qquad Y_k = \begin{bmatrix} | & | & & | \\ y_{k-p} & y_{k-p+1} & \dots & y_k \\ | & | & & | \end{bmatrix}$$

where $s_i = x_{i+1} - x_i$ and $y_i = \nabla f(x_{i+1}) - \nabla f(x_i)$. Then, we can define multisecant condition

$$B_{k+1} S_k = Y_k \tag{6}$$

which interpolates p number of previous iterates. Given the matrices S_k and Y_k, (6) is an under-determined problem because the number of constraints is less than the number of variables that should be defined for $B_k \in \mathbb{R}^{n\times n}$. Following multisecant DFP and multisecant BFGS updates are under the assumption that Y^TS is symmetric (and positive semidefinite).

$$\begin{aligned} B_{k+1} = B_k &+ (Y_k - B_kS_k)(Y_k^TS_k)^{-1}Y_k^T + Y_k(Y_k^TS_k)^{-1}(Y_k - B_kS_k)^T \\ &- Y_k(Y_k^TS_k)^{-1}(Y_k - B_kS_k)^TS_k(Y_k^TS_k)^{-1}Y_k^T. \end{aligned} \tag{MS DFP}$$

$$B_{k+1} = B_k + Y_k(Y_k^TS_k)^{-1}Y_k^T - B_kS_k(S_k^TB_kS_k)^{-1}S_k^TB_k \tag{MS BFGS}$$

However, the assumption that Y^TS is symmetric and/or positive semidefinite is not true for general convex functions f. In fact, outside of f being a quadratic function, it is usually untrue. Specifically, if S^TY is not symmetric (or positive semidefinite) then it is impossible to both satisfy (6) *and* have B_{k+1} be symmetric (or positive semidefintie).

3. An almost-multi-secant method

We first summarize all the existing MS quasi-Newton methods as

$$B_{k+1} = B_k - D_1W^{-1}D_2^T$$

for some D_1, D_2, W. (Note that W is not usually symmetric nor positive semidefinite.) The natural perturbation to enforce symmetry and positive semidefiniteness is to

$$B_{k+1} = B_k - \frac{D_1W^{-1}D_2^T + (D_1W^{-1}D_2^T)^T}{2} + \mu I$$

where μ is the smallest positive value needed to ensure that $B_{k+1} \succeq 0$. That is, defining

$$\bar\Delta = -\frac{1}{2}\begin{bmatrix} D_1 & D_2 \end{bmatrix}\begin{bmatrix} 0 & W_k^{-1} \\ W_k^{-T} & 0 \end{bmatrix}\begin{bmatrix} D_1^T \\ D_2^T \end{bmatrix} \in \mathbb{R}^{n\times n}$$

then the goal is to find $\mu = \max\{0, -\lambda_{\min}(B_k + \bar\Delta)\}$.

Note that the multisecant condition $B_{k+1}S_k = Y_k$ may not be exact when we perturb B_{k+1}, and this is the reason of being an 'almost multisecant' scheme. However, in general, finding $\lambda_{\min}(B_k)$ may not be computationally cheap. The obvious approach is to use a fast power method or Lanczos method, but there is no reason to assume that B_k is sparse, nor low rank after n iterations. Therefore, we assume that this operation is prohibitive, or at least can only be used rarely.

We therefore approximate $\mu = \max\{0, -\lambda_{\min}(\bar\Delta)\}$. This can be simply done by computing the eigenvalue of a tiny $2p \times 2p$ matrix by exploiting the Schur complement property. More mathematical details are written in the Appendix. Note that $\mu I + \bar\Delta$ is the Schur-complement of

$$H = \frac{1}{2}\begin{bmatrix} 2\mu I & D_1 & D_2 \\ D_1^T & 0 & W_k \\ D_2^T & W_k^T & 0 \end{bmatrix} \prec 0$$

where H is not PSD no matter how large μ is because of zeros in its diagonal. Therefore, we add a nontrivial diagonal block A whose Schur complement reduces to Δ. Let

$$\begin{aligned} \Delta &= -\tfrac{1}{2}\begin{bmatrix} D_1 & D_2 \end{bmatrix}\begin{bmatrix} 0 & W_k^{-1} \\ W_k^{-T} & 0 \end{bmatrix}\begin{bmatrix} D_1^T \\ D_2^T \end{bmatrix} + \mu I \in \mathbb{R}^{n\times n} \\ H_1 &= 2\mu I + A - \begin{bmatrix} D_1 & D_2 \end{bmatrix}\begin{bmatrix} cI & F \\ F^T & cI \end{bmatrix}^{-1}\begin{bmatrix} D_1^T \\ D_2^T \end{bmatrix} \in \mathbb{R}^{n\times n} \\ H_2 &= \begin{bmatrix} cI & F \\ F^T & cI \end{bmatrix} - \begin{bmatrix} D_1^T \\ D_2^T \end{bmatrix}(A + 2\mu I)^{-1}\begin{bmatrix} D_1 & D_2 \end{bmatrix} \in \mathbb{R}^{4p\times 4p} \end{aligned}$$

Then, for the right choice of A and F (details are in the Appendix A), $\Delta = \bar{\Delta} + \mu$ and H_1 is PSD if and only if H_2 is PSD. Since H_2 is a much smaller $(2p \times 2p)$ matrix, finding μ large enough such that H_2 is PSD can be done much more efficiently. The Figure 3 in the Appendix C sustains the argument that Δ is positive semidefinite if and only if H_2 is positive semidefinite.

4. Numerical Results

Quadratic Problem We define a quadratic problem with $A \in \mathbb{R}^{p \times n}$, $x_0 = .001 \times \bar{1}$, $\eta \sim N(0, 1)$ and $b = Ax_0 + \eta$ where

$$\min_{x \in \mathbb{R}^p} f(x) = \min_{x \in \mathbb{R}^p} \frac{||Ax - b||_2^2}{2p}$$

Logistic Regression Problem We define Logistic Regression problem with b is a binary vector and σ is the sigmoid function, $\sigma(x) = \frac{1}{1+e^{-x}}$ where

$$\min_{\theta \in \mathbb{R}^n} f(\theta) = \min_{\theta \in \mathbb{R}^n} -\frac{1}{p} \sum_{i=1}^{p} log(\sigma(b_i a_i^T \theta))$$

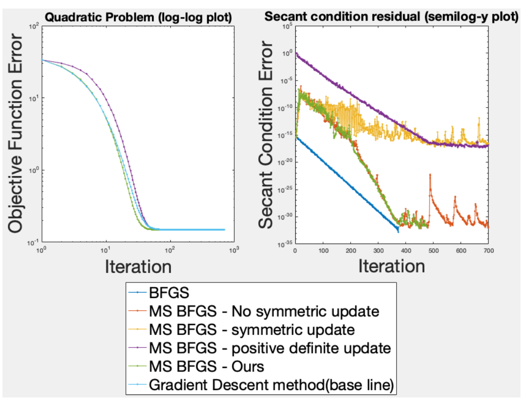

Figure 1: Quadratic Problem

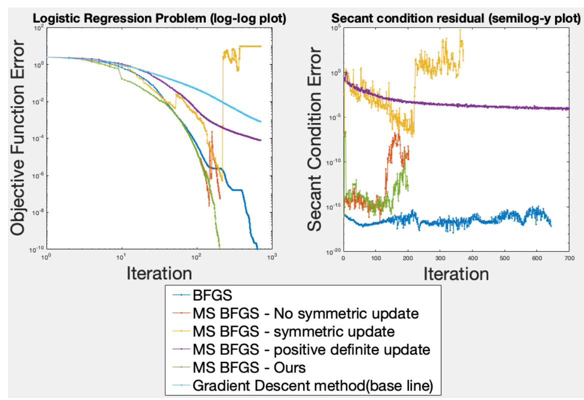

Figure 2: Logistic Regression Problem

In the above simulation results, Quadratic problem's loss value is monotonically decreasing for every multisecant case because $B \succeq 0$ which satisfies $S^T BS = S^T Y$ and maintains $-B^{-1}\nabla f(x)$ be descent direction. On the other hand, Logistic regression problem monotonically decrease only if B is positive definite by controlling μ. In this case, $f(x)$ is not quadratic and $Y^T S$ is not symmetric which shows that the secant condition is not fully satisfied (exact) in Figure 2.

5. Conclusion

We aimed at improving the approximation of the Hessian matrix while keeping computational costs low. More precisely, we employ a strategy involving low-rank perturbations to create an almost-secant quasi-Newton approach, ensuring that the estimated Hessian remains positive definite by the Schur-Complement theorem. This, in turn, contributes to maintaining a consistent descent in solving a minimization problem, thereby reducing the risk of method divergence. Our findings demonstrate that meticulously adjusting the update process by getting the right value μ enhances the stability and efficiency of multisecant quasi-Newton updates.

References

Broyden, Charles George. ”The convergence of a class of double-rank minimization algorithms 1. general considerations.” IMA Journal of Applied Mathematics 6.1 (1970): 76-90.

Fletcher, Roger. ”A new approach to variable metric algorithms.” The computer journal 13.3 (1970): 317-322.

Goldfarb, Donald. ”A family of variable-metric methods derived by variational means.” Mathematics of computation 24.109 (1970): 23-26.

Shanno, David F. ”Conditioning of quasi-Newton methods for function minimization.” Mathematics of computation 24.111 (1970): 647-656.

Broyden, Charles G. ”A class of methods for solving nonlinear simultaneous equations.” Mathematics of computation 19.92 (1965): 577-593.

Gay, D. Schnabel, R. (1977). “Solving systems of non-linear equations by Broydan’s method with projected updates.” NBER Working Papers, National Bureau of Economic Research, Inc, 0169

Schnabel, Robert B (1983). “Quasi-Newton Methods Using Multiple Secant Equations.” COLORADO UNIV AT BOULDER DEPT OF COMPUTER SCIENCE

Gay, David M., and Robert B. Schnabel. ”Solving systems of nonlinear equations by Broyden’s method with projected updates.” Nonlinear Programming 3. Academic Press, 1978. 245-281.

Fang, Haw-ren, and Yousef Saad. ”Two classes of multisecant methods for nonlinear acceleration.” Numerical Linear Algebra with Applications 16.3 (2009): 197-221.

Davidon, William C. ”Variable metric method for minimization.” SIAM Journal on Optimization 1.1 (1991): 1-17.

Eyert, V. ”A comparative study on methods for convergence acceleration of iterative vector sequences.” Journal of Computational Physics 124.2 (1996): 271-285.

Mokhtari, Aryan, Mark Eisen, and Alejandro Ribeiro. ”IQN: An incremental quasi-Newton method with local superlinear convergence rate.” SIAM Journal on Optimization 28.2 (2018): 1670-1698.

Liu, Chengchang, Cheng Chen, and Luo Luo. ”Symmetric Rank-k Methods.” arXiv preprint arXiv:2303.16188 (2023).

Gao, Wenbo, and Donald Goldfarb. ”Block BFGS methods.” SIAM Journal on Optimization 28.2 (2018): 1205-1231.

Appendix A.

Lemma 1 *Define $c = c_1 = c_2$ and*

$$P = (cI - c^{-1}FF^T)^{-1}, \qquad Q = (cI - c^{-1}F^TF)^{-1}$$

Pick $F = c_3USV^T$ where $W^{-1} = U\Sigma V^T$ is the full SVD of W^{-1}, and S is a diagonal matrix satisfying

$$\Sigma = (S^2 - c^2I)^{-1}S. \tag{7}$$

Pick $c_3 = \frac{c\epsilon}{c+\|W\|_2}$ for some $\epsilon \in (0,1)$. Then the inverse

$$\begin{bmatrix} P & -F(c^2I - F^TF)^{-1} \\ -(c^2I - F^TF)^{-1}F^T & Q \end{bmatrix} = \begin{bmatrix} cI & F \\ F^T & cI \end{bmatrix}^{-1}$$

Then the following three statements are equivalently true.

1. $\|F\|_2 \le c$
2. $\begin{bmatrix} cI & F \\ F^T & cI \end{bmatrix}$ *is PSD*
3. P *and* Q *exists and are also PSD*

Proof

Recall the inverse of a 2x2 block matrix can be written as

$$\begin{bmatrix} cI & F \\ F^T & cI \end{bmatrix}^{-1} = \begin{bmatrix} (cI - c^{-1}FF^T)^{-1} & -F(c^2I - F^TF)^{-1} \\ -(c^2I - F^TF)^{-1}F^T & (cI - c^{-1}F^TF)^{-1} \end{bmatrix}.$$

Thus gives the correct construction of P and Q. Then, in the off diagonal terms,

$$\begin{aligned} -F(c^2I - F^TF)^{-1} &= -c_3USV^T(c^2VV^T - c_3^2VS^2V^T)^{-1} \\ &= c_3US(c_3^2S^2 - c^2I)^{-1}V^T \end{aligned}$$

and

$$\begin{aligned} -(c^2I - F^TF)^{-1}F^T &= -(c^2I - c_3^2VS^2V^T)^{-1}c_3VSU^T \\ &= V(c_3^2S^2 - c^2I)^{-1}c_3SU^T \end{aligned}$$

Note that

$$\Sigma = (S^2 - c^2I)^{-1}S \iff (S^2 - c^2I)\Sigma = S \iff S_{ii}^2\Sigma_i - S_{ii} - \Sigma_i c^2 = 0 \text{ for } \forall i$$

and from the quadratic formula, we have the singular values of F as

$$S_{ii} = \frac{1 + \sqrt{1 + 4\Sigma_i^2c^2}}{2\Sigma_i} \le \frac{1 + \sqrt{1} + \sqrt{4\Sigma_i^2c^2}}{2\Sigma_i} = \frac{1}{\Sigma_i} + c$$

or, $\frac{1}{c_3}\|F\|_2 \leq c + \|W\|_2$ *because* $\|F\|_2 = c_3\|S\|_2$ *and* $\|F\|_2 \leq (c + \|W\|_2)c_3 = c\epsilon \leq c$. *Thus 1. is true under our assignment of* c_3.

Next, we can simply prove the second and third properties by the Schur-complement

$$\begin{aligned}\begin{bmatrix} cI & F \\ F^T & cI \end{bmatrix} \succeq 0 \quad &\iff \quad cI \succeq 0 \text{ and } cI - \frac{1}{c}FF^T = \frac{1}{c}(c^2 I - FF^T) \succeq 0 \\ &\iff \quad cI \succeq 0 \text{ and } cI - \frac{1}{c}F^T F = \frac{1}{c}(c^2 I - F^T F) \succeq 0\end{aligned}$$

if and only if

$$c^2 - \|F\|_2^2 \geq 0 \iff \|F\|_2 \leq c$$

■

Appendix B.

Lemma 2 *For any choice of positive number c, The matrix*

$$A := \Delta - 2\mu I + \begin{bmatrix} D_1 & D_2 \end{bmatrix} \begin{bmatrix} cI & F \\ F^T & cI \end{bmatrix}^{-1} \begin{bmatrix} D_1^T \\ D_2^T \end{bmatrix}$$

is positive semidefinite.

Proof *Based on our construction,* A *can be written as*

$$\begin{aligned} A &= \begin{bmatrix} D_1 & D_2 \end{bmatrix} \begin{bmatrix} (cI - c^{-1}FF^T)^{-1} & -F(c^2 I - F^T F)^{-1} - W^{-1} \\ -(c^2 I - F^T F)^{-1} F^T - W^{-T} & (cI - c^{-1}F^T F)^{-1} \end{bmatrix} \begin{bmatrix} D_1^T \\ D_2^T \end{bmatrix} \\ &= \begin{bmatrix} D_1 U & D_2 V \end{bmatrix} \underbrace{\begin{bmatrix} (cI - c^{-1}c_3^2 S^2)^{-1} & c_3 S(c_3^2 S^2 - c^2 I)^{-1} - \Sigma \\ c_3(c_3^2 S^2 - c^2 I)^{-1} S - \Sigma & (cI - c^{-1}c_3^2 S^2)^{-1} \end{bmatrix}}_{=:B} \begin{bmatrix} U^T D_1^T \\ V^T D_2^T \end{bmatrix} \end{aligned}$$

where $W^{-1} = U\Sigma V^T$ *and* $F = c_3 USV^T$. *We are left to show if* B *is PSD. Note that we may partition* B *into 4 blocks of diagonal matrices, which means there exists a permutation* PBP^T *which is block diagonal, with 2x2 symmetric blocks*

$$B_{ii} = \begin{bmatrix} \frac{1}{c - \frac{1}{c}c_3^2 S_{ii}^2} & \frac{c_3 S_{ii}}{c_3^2 S_{ii}^2 - c^2} - \Sigma_{ii} \\ \frac{c_3 S_{ii}}{c_3^2 S_{ii}^2 - c^2} - \Sigma_{ii} & \frac{1}{c - \frac{1}{c}c_3^2 S_{ii}^2}. \end{bmatrix}$$

The (1,1) and (2,2) blocks can be shown to be positive since

$$c_3 S_{ii} \leq \frac{c}{c + \|W\|_2}(\|W\|_2 + c) = c. \tag{8}$$

Therefore, B_{ii} *is PSD iff the (2,1) element has magnitude smaller than both diagonal elements; that is,*

$$B_{ii} \succeq 0 \iff \frac{1}{c - \frac{1}{c}c_3^2 S_{ii}^2} \geq \frac{c_3 S_{ii}}{c_3^2 S_{ii}^2 - c^2} - \Sigma_{ii}.$$

Since (8)*, this is equivalent to*

$$c \geq -c_3 S_{ii} - \underbrace{(c^2 - c_3^2 S_{ii}^2)}_{\geq 0} \Sigma_{ii}$$

which is true since the right hand side is negative.

■

Appendix C.

Theorem 1 *Consider W a nonsymmetric matrix, and*

$$\Delta = \mu I - \frac{1}{2} \begin{bmatrix} D_1 & D_2 \end{bmatrix} \begin{bmatrix} 0 & W^{-1} \\ W^{-T} & 0 \end{bmatrix} \begin{bmatrix} D_1^T \\ D_2^T \end{bmatrix}.$$

Then Δ is PSD if and only if

$$H_2 = \begin{bmatrix} cI & F \\ F^T & cI \end{bmatrix} - \begin{bmatrix} D_1^T \\ D_2^T \end{bmatrix} (A + 2\mu I)^{-1} \begin{bmatrix} D_1 & D_2 \end{bmatrix}. \tag{9}$$

is PSD, for

$$A = \begin{bmatrix} D_1 & D_2 \end{bmatrix} \begin{bmatrix} P & -(c^2 I - F^T F)^{-1} F^T - W^{-1} \\ -F(c^2 I - F^T F)^{-1} - W^{-T} & Q \end{bmatrix} \begin{bmatrix} D_1^T \\ D_2^T \end{bmatrix}$$

and $F = \frac{c\epsilon}{c + \|W\|_2} V S U^T$ where $W^{-1} = U \Sigma V^T$ is the SVD of W^{-1}, and S is a diagonal matrix satisfying (7).

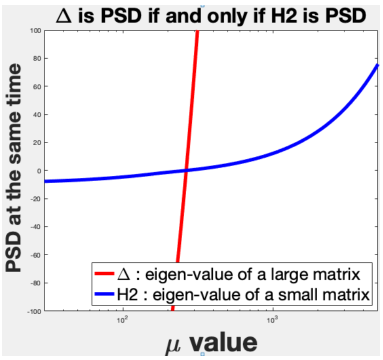

Figure 3: **Picking μ:** $\Delta \succ 0$ (large matrix) if and only if $H_2 \succ 0$ (small matrix).

Proof *Consider the matrix*

$$H = \begin{bmatrix} 2\mu I + A & D_1 & D_2 \\ D_1^T & cI & F \\ D_2^T & F^T & cI \end{bmatrix}.$$

where c is a nonnegative scalar, F is a $2p \times 2p$ matrix (yet undefined), and A is some (unspecified) symmetric matrix. Then the two Schur complements of H are H_1 and H_2:

$$\begin{aligned} H_1 &:= 2\mu I + A - \begin{bmatrix} D_1 & D_2 \end{bmatrix} \begin{bmatrix} cI & F \\ F^T & cI \end{bmatrix}^{-1} \begin{bmatrix} D_1^T \\ D_2^T \end{bmatrix} \in \mathbb{R}^{n\times n} \\ H_2 &= \underbrace{\begin{bmatrix} cI & F \\ F^T & cI \end{bmatrix}}_{H_3} - \begin{bmatrix} D_1^T \\ D_2^T \end{bmatrix} (A + 2\mu I)^{-1} \begin{bmatrix} D_1 & D_2 \end{bmatrix} \in \mathbb{R}^{4m\times 4m}. \end{aligned}$$

Then,

$$H_1 \text{ is PSD and } \begin{bmatrix} c_1 I & F \\ F^T & c_2 I \end{bmatrix} \text{ is PSD and invertible}$$

if and only if

$$H_2 \text{ is PSD and } A + 2\mu I \text{ is PSD and invertible.}$$

From Lemma 1, we see that the proposed construction of A and F is indeed valid for setting $\Delta = H_1$; moreover, for any value of $c > 0$, A and H_3 are both PSD. Thus, Δ is PSD if and only if H_2 is PSD.

■

Note that while we have pushed the certification of PSD from our original $n \times n$ matrix Δ to that of a smaller $2p \times 2p$ matrix in (9), the inversion $(A + 2\mu I)^{-1} \in \mathbb{R}^{n\times n}$ still seems daunting. However, note that

$$A + 2\mu I = \begin{bmatrix} D_1 & D_2 \end{bmatrix} B \begin{bmatrix} D_1^T \\ D_2^T \end{bmatrix} + 2\mu I$$

for

$$B = \begin{bmatrix} (cI - c^{-1}FF^T)^{-1} & -(c^2 I - F^T F)^{-1} F^T - W^{-1} \\ -F(c^2 I - F^T F)^{-1} - W^{-T} & (cI - c^{-1}F^T F)^{-1} \end{bmatrix} \in \mathbb{R}^{4m\times 4m}$$

is a diagonal-plus-low-rank matrix, and its inverse can be efficiently computed using another Woodbury inversion

$$(A + 2\mu I)^{-1} = \frac{1}{2\mu} I - \frac{1}{2\mu} \begin{bmatrix} D_1 & D_2 \end{bmatrix} \left(2\mu B^{-1} + \begin{bmatrix} D_1^T \\ D_2^T \end{bmatrix} \begin{bmatrix} D_1 & D_2 \end{bmatrix} \right)^{-1} \begin{bmatrix} D_1^T \\ D_2^T \end{bmatrix}.$$