A J-Symmetric Quasi-Newton Method for Minimax Problems

Azam Asl* Haihao Lu†

February 4, 2022

Abstract

Minimax problems have gained tremendous attentions across the optimization and machine learning community recently. In this paper, we introduce a new quasi-Newton method for minimax problems, which we call J-symmetric quasi-Newton method. The method is obtained by exploiting the J-symmetric structure of the second-order derivative of the objective function in minimax problem. We show that the Hessian estimation (as well as its inverse) can be updated by a rank-2 operation, and it turns out that the update rule is a natural generalization of the classic Powell symmetric Broyden (PSB) method from minimization problems to minimax problems. In theory, we show that our proposed quasi-Newton algorithm enjoys local Q-superlinear convergence to a desirable solution under standard regularity conditions. Furthermore, we introduce a trust-region variant of the algorithm that enjoys global R-superlinear convergence. Finally, we present numerical experiments that verify our theory and show the effectiveness of our proposed algorithms compared to Broyden's method and the extragradient method on three classes of minimax problems.

1 Introduction.

Our problem of interest in this paper is the minimax problem (a.k.a. saddle-point problem)

$$\min_{x\in\mathbb{R}^n}\max_{w\in\mathbb{R}^m} L(x,w)\ , \tag{1}$$

where $L(x,w)$ is a smooth objective in both x and w, and we call x the primal variable and w the dual variable. Minimax problem is one of the most important classes of optimization problems, with a long research history and wide applications. The earliest motivation for minimax problems may come from the Lagrangian form of constrained optimization problems [7]. Another major application of minimax problem is zero sum games [40]. More recently, minimax problem (1) has regained significant attentions across the optimization and machine learning communities, mainly due to their applications in machine learning, such as generative adversarial networks (GANs) [27], reinforcement learning [17], robust training [34], image processing [15], and applications in classic constrained optimization, such as linear programming [3].

Here, we develop a quasi-Newton method for minimax problem (1). Quasi-Newton method is a successful optimization method for minimization problems [38]. While Newton's method enjoys the fast local quadratic convergence, the iteration cost to access the Hessian and to solve the linear equation can be prohibitive

*University of Chicago Booth School of Business.
†University of Chicago Booth School of Business.

when solving large instances. Instead, quasi-Newton method constructs an approximate Hessian (more often constructs an approximate inverse Hessian) and updates it with a low-rank operation at each iteration, which can significantly reduce the iteration cost. Under proper regularity conditions, one can show that quasi-Newton method has local superlinear convergence and global linear convergence. Some famous quasi-Newton updates for minimization problems include BFGS formula [26], DFP formula [19], PSB formula [43], etc. The low cost-per-iteration and superlinear eventual convergence make quasi-Newton method a highly efficient algorithm. It is widely used in practice and listed as one of the ten algorithms with the greatest impact on the development and practice of science and engineering in the early 21st century [1].

Surprisingly, there has been very limited research on quasi-Newton methods for minimax problems. As a special case of nonlinear equations or as a special case of variational inequalities, one can adapt quasi-Newton methods for these problems to solve minimax problems (1). In particular, Broyden's ("good" or "bad") methods [9] are quasi-Newton methods for solving generic nonlinear equations, and we can use them to solve the KKT system of (1). In 1990s, Burke and Qian proposed a variable metric proximal point method for monotone variational inequality [13, 12], where they effectively introduce a proximal point variant of quasi-Newton method and use Broyden's formula to update the second-order term. Broyden's method and Burke and Qian's method target at a much larger class of problems, and *do not* utilize the structure of minimax problems. In contrast, in this paper, we propose a new quasi-Newton method specialized for minimax problems, which utilize the structure of the second-order derivative of minimax problems. The utilization of such structures has the following advantages compared to existing quasi-Newton methods and first-order methods:

- Many classic quasi-Newton methods, such as BFGS formula and DFP formula, target at minimization problems and construct symmetric and positive definite approximations of Hessian. These methods do not directly work for minimax problems, where the second-order derivative is no longer symmetric.
- Broyden's formula targets at finding root of nonlinear equations, and does not require any structure on the Jacobian estimation. While it is very general, it ignores the meaningful information of the Jacobian structure in minimax problems and it is numerically unstable even when solving simple bilinear minimax problems (as shown in Section 5). Furthermore, it is unclear how to properly initialize the Jacobian estimation of Broyden's method for minimax problems, which may lead to numerical issues.
- Compared with first-order methods, such as EGM, quasi-Newton method enjoys a local superlinear convergence rate and the convergence speed does not heavily rely on the condition number of the problem.

Throughout the paper, we assume the objective function $L(x, w)$ is third-order differentiable. For notational convenience, we denote $z = (x, w) \in \mathbb{R}^{m+n}$ as the primal-dual solution pair, $F(z) = [\nabla_x L(x, w), -\nabla_w L(x, w)]$ as the gradient (more precisely gradient for the primal and negative gradient for the dual) of $L(x, w)$. $F(z)$ is the cornerstone of first-order methods for minimax problems. For example, the gradient descent ascent (GDA) method has an iteration update $z_{k+1} = z_k - sF(z_k)$, the proximal point method (PPM) has an iteration update $z_{k+1} = z_k - sF(z_{k+1})$, and the extragradient method (EGM) has an iteration update $z'_k = z_k - sF(z_k), z_{k+1} = z_k - sF(z'_k)$.

When turning to second-order methods, we denote

$$\nabla F(z) = \begin{bmatrix} \nabla_{xx} L(x, w) & \nabla_{xw} L(x, w) \\ -\nabla_{xw} L(x, w)^T & -\nabla_{ww} L(x, w) \end{bmatrix} \tag{2}$$

as the Jacobian of $F(z)$. Then the standard Newton's method has an iteration update

$$z_{k+1} = z_k - \nabla F(z_k)^{-1} F(z_k) \ .$$

We here focus on quasi-Newton method with an iteration update

$$z_{k+1} = z_k - B_k^{-1} F(z_k) \ ,$$

where B_k is an approximation of $\nabla F(z_k)$. A key observation is that $\nabla F(z)$ defined in (2) is symmetric on the main diagonal terms and skew-symmetric on the anti-diagonal terms. This type of matrix is called J-symmetric in the related literature [33]. A J-symmetric matrix has many desirable numerical properties, see, for example, [6, Theorems 3.6 and 3.7] and [5, Lemma 1.1]. J-symmetric matrix naturally appears and has been used in numerical analysis and applied mathematics. For example, [5] introduces a J-symmetric system as a preconditioner for Krylov subspace methods for solving nonlinear equations. [47] uses J-symmetric matrices as preconditioner when solving discrete Navier-Stokes equations in incompressible fluid mechanics. In particular, when the minimax problem is convex-concave, $\nabla F(z)$ has positive semidefinite main diagonal terms. The J-symmetric structure is the cornerstone of our quasi-Newton update, and the utilization of J-symmetric structure is the major novelty of our approach over existing literature.

The major contributions of our work can be summarised as following:

- We introduce a new quasi-Newton update for minimax problems that comes from the J-symmetric structure of the Jacobian of the minimax objective. We show that we can efficiently update the Jacobian estimation as well as its inverse in our method via a rank-2 update. It turns out the update rule is a natural generalization of Powell's symmetric Broyden (PSB) update from minimization problems to minimax problems.
- We prove that the proposed unit-step quasi-Newton method enjoys local Q-superlinear convergence towards an stationary point of the minimax problem via the *bounded deterioration* technique. Furthermore, we propose a trust-region variant of the proposed quasi-Newton method and prove its global R-superlinear convergence. The convergence results do not require convexity-concavity of objective function in the minimax problem.
- We present preliminary numerical experiments, which verifies our theory and showcases that our proposed methods are more stable and faster compared to Broyden's update when solving minimax problems. They also enjoy faster convergence compared to first-order method such as EGM.

1.1 Applications of Minimax Problems.

We here briefly discuss three applications of minimax problems.

(Linear equality-constrained convex optimization.) Consider a constrained optimization problem of the form

$$\begin{aligned} \min_x \ & f(x) \\ \text{s.t. } & Ax = b \ . \end{aligned} \tag{3}$$

This type of problem is the subproblem in sequential quadratic programming [38] and arises in computational physics [6]. The Lagrangian is $L(x, w) = f(x) + w^T(Ax - b)$, where w is the Lagrange multiplier.

(Inequality-constrained convex optimization.) Consider a generic constrained convex optimization problem

$$\begin{aligned} \min_x \ & f(x) \\ \text{s.t. } & g(x) \geq 0 \ . \end{aligned}$$

Introducing the Lagrangian multiplier w yields

$$\min_x \max_{w \geq 0} f(x) - w^T g(x) \ .$$

Notice that there exists $w \geq 0$ constraint for the dual variable. We can instead consider a logarithmic-barrier formulation with barrier parameter μ:

$$\min_x \max_w L(x, w; \mu) := f(x) - w^T g(x) + \mu \sum_i \log w_i \ .$$

The solution to the above minimax problem identifies an optimal solution to the original minimization problem when $\mu \to 0$. Indeed, as long as μ is chosen properly, it provides an approximated solution.

(Generative Adversarial Network.) Generative Adversarial Network (GAN) [27] is a recent development in machine learning, which has many applications in image processing such as producing realistic images [29], quality super-resolution [31] and image-to-image translation [48]. A GAN is a minimax problem of form (1) which is the equilibrium condition of a zero-sum two-player game. The two players are the generator (parameterized by G) and the discriminator (parameterized by D):

$$\min_G \max_D \mathop{\mathbb{E}}_{s \sim p} [\log D(s)] + \mathop{\mathbb{E}}_{e \sim q} [\log(1 - D(G(e)))] \ , \tag{4}$$

where p is the data distribution and q is the latent distribution. The generator produces a sample and the discriminator decides whether they are real or fake data. The goal is to learn the best generator which can produce realistic images [24]. Notice that G and D are usually represented as the parameters of neural networks, thus (4) is a nonconvex nonconcave minimax problem.

1.2 Related Literature.

Minimax optimization. Minimax optimization (1) has long history and wide applications. The early work on minimax optimization focus on a more general problem, monotone variational inequalities. The two classical algorithms for monotone variational inequality/minimax problems are perhaps proximal point method (PPM) proposed by Rockafellar [45] and extragradient method (EGM) proposed by Korpelevich [30] in 1970s. Later on, Nemirovski [37] proposes mirror prox algorithm, which generalizes EGM with Bregman divergence and builds up the connection between EGM and PPM.

Motivated by machine learning applications, there is a renewed recent interest on developing efficient first-order algorithms for minimax problems. [18] studies an Optimistic Gradient Descent Ascent (OGDA) with applications in GAN. [35] presents an interesting observation that OGDA approximates PPM on bilinear problems. [32] proposes a high-resolution ODE framework that can characterize different primal-dual algorithms. [28] studies the landscape of PPM and presents examples showing that the classic algorithms such as PPM, EGM, gradient descent ascent, and alternating gradient descent ascent may converge to a limit circle on a simple nonconvex-nonconcave example. See [28] for a thorough literature review on the recent development of minimax problems. Compared to these first-order methods, our focus is on quasi-Newton methods, and our theoretical results do not rely on the convexity of the objective.

Quasi-Newton methods. Quasi-Newton methods are alternatives to the classical Newton's method. Instead of computing the Newton's direction by solving linear equations using the Jacobian, quasi-Newton methods often formulate an approximate inverse Jacobian and use a low-rank operation to update the inverse. Quasi-Newton methods are computationally more efficient than Newton's method, thus can solve much larger instances. Due to their high impact on minimization problems, quasi-Newton methods have appeared in the updated list of "10 algorithms with the greatest influence on the development and practice of science and engineering in the 20th century" given by Nick Higham in 2016 [1].

The first quasi-Newton method was developed by W.C. Davidon in 1959 and later published in [19] in 1991. Instead of computing the inverse Hessian at every iteration, Davidon's method obtains a good approximation of it using gradient differences. Soon after, Fletcher and Powell realized the efficiency of this method. They studied and popularized Davidon's original formula and established its convergence for convex quadratic functions [23]. This method became known as DFP method. BFGS [26] is perhaps the most popular quasi-Newton method [38]. It was discovered by Broyden, Fletcher, Goldfarb and Shanno independently in 1970s. Soon after, Broyden, Dennis and Moré proved the first local and superlinear convergence results for BFGS, DFP as well as other quasi-Newton methods [11]. Later on, Powell [44] presented the first global convergence result of BFGS with an inexact Armijo-Wolfe line search for a general class of smooth convex optimization problems. [14] extended Powell's result to a broader class of quasi-Newton methods.

Quasi-Newton methods for minimax problems. While minimax problems and quasi-Newton methods are both well studied individually, there are fairly limited works on quasi-Newton methods for minimax problems. Notice that one can solve minimax problems by finding a root of a corresponding nonlinear equation, thus one can use the classical Broyden's (good and bad) algorithms [9, 10] for minimax problems. Another line of early research is to use proximal quasi-Newton methods for monotone variational inequalities proposed in [16, 13, 12] to solve convex-concave minimax problems. However, both Broyden's methods and the proximal quasi-Newton methods target at a more general class of problems, without considering the special structure of the minimax problems. As a result, these algorithms may not always be stable even when solving simple bilinear minimax problems as we see in our numerical experiments. More recently, [2, 22] proposed different quasi-Newton methods for minimax problems. However, neither of them show the convergence rate of their algorithms. In contrast to these works, we introduce a new quasi-Newton method for minimax problem and present its local/global superlinear rate.

Trust-region method. Trust-region method is another classic algorithm in numerical optimization. It first defines a region around the current best solution, and then creates a quadratic model that can approximate the objective function in the region and takes a step by solving a subproblem based on this quadratic model. Quasi-Newton methods are often used together with trust-region method [38]. Unlike a line-search method, which picks the direction first and then looks for an acceptable stepsize along that direction, a trust-region method first picks the stepsize and then looks for an acceptable direction within that region.

There are different methods for solving the trust-region subproblem. The simplest way is to move along the negative gradient direction to a point within the trust-region. The resulting solution is called Cauchy point. While Cauchy point is cheap to calculate, it may perform poorly in some cases. A famous approach to avoid this issue is the dogleg method. The dogleg method was originally introduced by Powell as hybrid method in [42]. The dogleg point refers to a point on the boundary of the trust-region that is a linear combination of the Cauchy point and the minimizer of the quadratic model, and it is used only when the Cauchy point is strictly inside the trust-region and the minimizer of the quadratic model is strictly outside the trust-region. See [38] for more details of the trust-region method.

1.3 Notations.

Throughout this paper, the norm $\|\cdot\|$ denotes the ℓ_2 norm for a vector or the operator norm (i.e., the $\ell_{2,2}$ norm) for a matrix, unless specified. The norm $\|\cdot\|_F$ refers to the Frobenius norm for a matrix. As a common notation in quasi-Newton method, s_k denotes the potential step at iteration k. When s_k is a sufficient decrease step and we accept it, we have $s_k = z_{k+1} - z_k$. Otherwise, we reject it (equivalently, we take a null step and set $z_{k+1} = z_k$). We use $y_k = F(z_k + s_k) - F(z_k)$ to denote the gradient difference between two consecutive points. We use $J \in \mathbb{R}^{(n+m)\times(n+m)}$ to represent the following block diagonal square matrix:

$$J = \begin{bmatrix} I_{n\times n} & 0 \\ 0 & -I_{m\times m} \end{bmatrix} .$$

2 J-symmetric Update.

In this section we present our J-symmetric update for minimax problems. The major idea is to construct the estimated Jacobian by utilizing the J-symmetric structure in $\nabla F(z)$. We begin by introducing the following notations for notational convenience:

$$D(z) = \nabla_{xx} L(z) \ , C(z) = -\nabla_{ww} L(z) \ , A(z) = \nabla_{xw} L(z)^T \ .$$

Then the Jacobian defined in (2) can be rewritten as

$$\nabla F(z) = \begin{bmatrix} D(z) & A^T(z) \\ -A(z) & C(z) \end{bmatrix} ,$$

where the main diagonal terms are symmetric and the main off-diagonal terms are anti-symmetric. This structure is called J-symmetric [5, 6]. Recall that matrix $J = \begin{bmatrix} I_{n\times n} & 0 \\ 0 & -I_{m\times m} \end{bmatrix}$. It is easy to check it holds for a J-symmetric matrix M that

$$M = JM^T J \text{ and } JM = M^T J \ .$$

The general scheme of the quasi-Newton method consists of iteration updates of the following form

$$z_{k+1} = z_k - B_k^{-1} F(z_k) \ , \tag{5}$$

where B_k denotes the approximation to the current Jacobian $\nabla F(z_k)$, and we hope to obtain a better and better approximation over time. In particular, we seek update rules from B_k to B_{k+1} such that:

(a) B_{k+1} is a good approximation to $\nabla F(z_{k+1})$.

(b) B_{k+1} is a J-symmetric matrix.

(c) B_{k+1} is not too far away from B_k.

(d) There is an efficient way for computing B_{k+1} from B_k by a low rank update.

A common requirement to satisfy (a) is that B_{k+1} should satisfy the secant condition

$$y_k = B_{k+1} s_k \ . \tag{6}$$

The secant condition imposes only $n + m$ constraints on B_{k+1} and even after taking into consideration the required J-symmetric structure, we are still left with many degrees of freedom to pick B_{k+1}. In addition, we select B_{k+1} such that it is the closest matrix to B_k in Frobenius norm. In summary, B_{k+1} is given by solving the following minimization problem:

$$\begin{aligned} \min_B \ & \frac{1}{2}\|B - B_k\|_F^2 \\ \text{s.t. } & Bs_k - y_k = 0 \\ & D = D^T, \quad C = C^T \quad \text{and} \quad B = \begin{bmatrix} D & A^T \\ -A & C \end{bmatrix} . \end{aligned} \tag{7}$$

The last line imposes the J-symmetric structure on B. Notice that the constraint set is a convex set and the objective is strongly convex, thus (7) is a convex optimization problem with a unique solution, and furthermore:

Proposition 2.1. *The unique solution to the constrained optimization problem* (7) *is given by*

$$B_{k+1} = B_k + \frac{Js_k(y_k - B_k s_k)^T J}{s_k^T s_k} + \frac{(y_k - B_k s_k)s_k^T}{s_k^T s_k} - \frac{(Js_k)^T(y_k - B_k s_k)Js_k s_k^T}{(s_k^T s_k)^2} \ , \tag{8}$$

which is a rank-2 update.

Proof. Define $E = B - B_k$ and $r = y_k - B_k s_k$. It is easy to see then the minimization problem (7) is equivalent to the following after changing variables

$$\min_E \ \frac{1}{2}\|E\|_F^2 \tag{9}$$

$$\text{s.t. } Es_k - r = 0 \tag{10}$$

$$D = D^T, \quad C = C^T \quad \text{and} \quad E = \begin{bmatrix} D & A^T \\ -A & C \end{bmatrix} \ . \tag{11}$$

J-symmetry constraint (11) is equivalent to that $E + E^T$ is a block diagonal matrix, and $E - E^T$ is a block anti-diagonal matrix. We dualize these two constraints and let $\Gamma_A, \Gamma_D \in \mathbb{R}^{(m+n)\times(m+n)}$ be the Lagrange multipliers of the condition involving $E + E^T$ and $E - E^T$, respectively. Then Γ_A is a block anti-diagonal matrix and Γ_D is a block diagonal matrix. Let $\lambda \in \mathbb{R}^{m+n}$ be the Lagrange multiplier corresponding to the secant condition (10). Then, the Lagrangian can be written as:

$$\Phi(E;\ \lambda, \Gamma_D, \Gamma_A) = \frac{1}{2}\operatorname{Tr}\left(EE^T\right) + \lambda^T(Es_k - r) + \operatorname{Tr}\left(\Gamma_D(E - E^T)\right) + \operatorname{Tr}\left(\Gamma_A(E + E^T)\right) \ .$$

Since $\lambda^T(Es_k - r) = \operatorname{Tr}\left((Es_k - r)\lambda^T\right)$, then

$$\Phi(E;\ \lambda, \Gamma_D, \Gamma_A) = \frac{1}{2}\operatorname{Tr}\left(EE^T\right) + \operatorname{Tr}\left((Es_k - r)\lambda^T\right) + \operatorname{Tr}\left(\Gamma_D(E - E^T)\right) + \operatorname{Tr}\left(\Gamma_A(E + E^T)\right) \ .$$

The KKT condition requires $\partial\Phi/\partial E = 0$, whereby

$$E = -\left(\lambda s_k^T + \Gamma_D^T - \Gamma_D + \Gamma_A^T + \Gamma_A\right) \ . \tag{12}$$

Furthermore, we decompose λs_k^T as following

$$\lambda = \begin{bmatrix} \lambda_1 \\ \lambda_2 \end{bmatrix} \ , \quad s_k = \begin{bmatrix} s_1 \\ s_2 \end{bmatrix}, \quad \lambda s_k^T = \begin{bmatrix} \lambda_1 s_1^T & \lambda_1 s_2^T \\ \lambda_2 s_1^T & \lambda_2 s_2^T \end{bmatrix} \ , \tag{13}$$

where $\lambda_1,\ s_1 \in \mathbb{R}^n$ and $\lambda_2,\ s_2 \in \mathbb{R}^m$. Note that the J-symmetry constraint (11) requires the diagonal blocks in (12) are symmetric, so

$$\begin{bmatrix} \lambda_1 s_1^T & 0 \\ 0 & \lambda_2 s_2^T \end{bmatrix} + \Gamma_D^T - \Gamma_D = \begin{bmatrix} s_1 \lambda_1^T & 0 \\ 0 & s_2 \lambda_2^T \end{bmatrix} + \Gamma_D - \Gamma_D^T \ ,$$

thus

$$\Gamma_D^T - \Gamma_D = -\frac{1}{2}\begin{bmatrix} \lambda_1 s_1^T - s_1\lambda_1^T & 0 \\ 0 & \lambda_2 s_2^T - s_2 \lambda_2^T \end{bmatrix} \ . \tag{14}$$

Furthermore, the block anti-diagonal matrix in (12) is skew-symmetric, so

$$\begin{bmatrix} 0 & \lambda_1 s_2^T \\ \lambda_2 s_1^T & 0 \end{bmatrix} + \Gamma_A^T + \Gamma_A = -\left(\begin{bmatrix} 0 & s_1 \lambda_2^T \\ s_2 \lambda_1^T & 0 \end{bmatrix} + \Gamma_A + \Gamma_A^T\right) \ ,$$

thus

$$\Gamma_A^T + \Gamma_A = -\frac{1}{2}\begin{bmatrix} 0 & \lambda_1 s_2^T + s_1\lambda_2^T \\ \lambda_2 s_1^T + s_2\lambda_1^T & 0 \end{bmatrix} . \tag{15}$$

Substituting (14) and (15) back into (12) and noticing (13), we obtain:

$$E = -\frac{1}{2}\begin{bmatrix} \lambda_1 s_1^T + s_1\lambda_1^T & \lambda_1 s_2^T - s_1\lambda_2^T \\ \lambda_2 s_1^T - s_2\lambda_1^T & \lambda_2 s_2^T + s_2\lambda_2^T \end{bmatrix} = -\frac{1}{2}(\lambda s_k^T + Js_k\lambda^T J) . \tag{16}$$

The rest of the proof is to compute the multiplier λ. Substituting (16) into the secant condition (10), we obtain

$$(\lambda s_k^T + Js_k\lambda^T J)s_k = -2r ,$$

and since both $s_k^T s_k$ and $\lambda^T J s_k$ are scalars, it holds that

$$\lambda = -\frac{1}{s_k^T s_k}\Big(2r + (s_k^T J\lambda)Js_k\Big) . \tag{17}$$

Multiplying both sides with $s_k^T J$, we arrive at

$$s_k^T J\lambda = -\frac{2s_k^T Jr + (s_k^T J\lambda)s_k^T JJs_k}{s_k^T s_k} ,$$

which can be further simplified to $s_k^T J\lambda = -s_k^T Jr/(s_k^T s_k)$ by using $J^2 = I$. Substituting this into (17), we obtain

$$\lambda = \frac{s_k^T Jr}{(s_k^T s_k)^2}Js_k - \frac{2}{s_k^T s_k}r .$$

Now, by substituting λ into (16) we obtain

$$E = -\frac{1}{2}\Big(\frac{s_k^T Jr}{(s_k^T s_k)^2}Js_k s_k^T - \frac{2}{s_k^T s_k}rs_k^T + \frac{s_k^T Jr}{(s_k^T s_k)^2}Js_k s_k^T JJ - \frac{2}{s_k^T s_k}Js_k r^T J\Big) .$$

By noticing $J^2 = I$, we conclude that the unique solution of the problem (9)-(11) is given by:

$$E = \frac{1}{s_k^T s_k}rs_k^T + \frac{1}{s_k^T s_k}Js_k r^T J - \frac{s_k^T Jr}{(s_k^T s_k)^2}Js_k s_k^T .$$

Finally by substituting $E = B_{k+1} - B_k$ into this equation we obtain the unique solution of problem (7) as:

$$B_{k+1} = B_k + \frac{1}{s_k^T s_k}(r - \frac{s_k^T Jr}{s_k^T s_k}Js_k)s_k^T + \frac{Js_k r^T J}{s_k^T s_k} .$$

This equations reveals the update is a rank-2 update. By changing the order and plugging in $r = y_k - B_k s_k$ we arrive at (8). □

Next, we show that the inverse update of (8) can also be obtained by a low rank update via Sherman-Woodbury identity.

Proposition 2.2. *Let $r = y_k - B_k s_k$, $H_k = B_k^{-1}$ and $H_{k+1} = B_{k+1}^{-1}$. The inverse update of* (8) *is*

$$H_{k+1} = Q^{-1} - \frac{Q^{-1}Js_k(Jr)^T Q^{-1}}{s_k^T s_k + (Jr)^T Q^{-1}Js_k} , \quad \text{where} \qquad Q^{-1} = H_k - \frac{H_k JP_k Jrs_k^T H_k}{s_k^T s_k + s_k^T H_k JP_k Jr} , \tag{18}$$

and

$$P_k = I - \frac{s_k s_k^T}{s_k^T s_k} . \tag{19}$$

Proof. Define $a = r - s_k^T JrJs_k/(s_k^T s_k) = JP_kJr$ and $Q = B_k + as_k^T/(s_k^T s_k)$, then from (8) we have that

$$B_{k+1} = Q + \frac{Js_k(Jr)^T}{s_k^T s_k} \ .$$

From one application of Sherman-Woodbury to Q, we obtain Q^{-1} and from another application to B_{k+1} we obtain H_{k+1} in (18). $\square$

Algorithm 1 describes the basic J-symmetric quasi-Newton method. We initialize with a solution z_0 and an inverse Jacobian estimation H_0. For every iteration, we calculate the direction s_k, update the iterates, compute the difference in $F(z)$, and finally update the inverse Jacobian estimation via (18). The algorithm is similar to any quasi-Newton method, and the key is the inverse Jacobian update rule (18). We will present the local Q-superlinear convergence of Algorithm 1 in the next section, and present the global R-superlinear convergence of a variant of Algorithm 1 in Section 4. Next, Algorithm 2 presents a simple line-search version

Algorithm 1 Unit-step J-symmetric Quasi-Newton Algorithm (J-symm)

Initialize with solution $z_0 \in \mathbb{R}^{m+n}$ and inverse Jacobian estimation $H_0 \in \mathbb{R}^{(m+n)\times(m+n)}$
for $k = 1, 2, 3, \ldots,$ **do**
  $s_k = -H_k F(z_k)$
  $z_{k+1} = z_k + s_k$
  $y_k = F(z_{k+1}) - F(z_k)$
  update H_{k+1} via (18)
end for

of the above algorithm. More specifically, after computing the J-symmetric direction $s_k = -H_k F(z_k)$ as in Algorithm 1, we test how much improvement we can obtain by taking the step. If we see enough improvement, we take this step, and otherwise we halve the stepsize t_k, as one does in a backtracking line-search. Notice that we start with $t_k = 1$ at each iteration, thus Algorithm 2 recovers Algorithm 1 if we see sufficient improvements every iteration with $t_k = 1$. Unfortunately, we do not have theoretical guarantees on this line search scheme, but numerical experiments in Section 5 showcases the benefits of the line-search scheme over other schemes.

Algorithm 2 J-symmetric Quasi-Newton Algorithm with Line Search (J-symm-LS)

Initialize with solution $z_0 \in \mathbb{R}^{m+n}$, inverse Jacobian estimation $H_0 \in \mathbb{R}^{(m+n)\times(m+n)}$ and linear-search parameter $c_1 \in (0, 1/2)$.
for $k = 1, 2, 3, \ldots,$ **do**
  $s_k = -H_k F(z_k)$
  $t_k = 1$
  while $\|F(z_k)\| - \|F(z_k + t_k s_k)\| < c_1 \|F(z_k)\|$ **do**
    $t_k = t_k/2$
  end while
  $z_{k+1} = z_k + t_k s_k$
  $y_k = F(z_{k+1}) - F(z_k)$
  update H_{k+1} via (18)
end for

In the end of this section, we discuss the connections between our method and Powell symmetric Broyden (PSB) update, and comment on the instability of Broyden's update.

The traditional minimization problem can be viewed as a special case of minimax problem (1) when the dual dimension is eliminated (namely $m = 0$). In such a case, our quasi-Newton update (8) recovers Powell symmetric Broyden update

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

Indeed, PSB update is known to be the unique minimizer of:

$$\begin{aligned} \min_B \ & \frac{1}{2}\|B - B_k\|_F^2 \\ \text{s.t. } & Bs_k - y_k = 0 \ , \\ & B = B^T \ . \end{aligned}$$

Therefore, the J-symmetric update (8) is a direct generalization of PSB update (PSB).

Next, let us look at Broyden's update for minimization problems. Broyden rank-1 update

$$B_{k+1} = B_k + \frac{(y_k - B_k s_k)s_k^T}{s_k^T s_k} \ , \tag{Broyden}$$

also known as good Broyden update, is the unique minimizer of the above minimization problem without the symmetry constraint

$$\begin{aligned} \min_B \ & \frac{1}{2}\|B - B_k\|_F^2 \\ \text{s.t. } & Bs_k - y_k = 0 \ . \end{aligned}$$

The inverse Hessian estimation in Broyden's update can be written as

$$H_{k+1} = H_k + \frac{(s_k - H_k y_k)s_k^T H_k}{s_k^T H_k y_k} \ .$$

Notice that Broyden's update can be numerically unstable, because there is no guarantee that the denominator $s_k^T H_k y_k$ is far away from 0. Indeed, avoiding such numerical instability is a major task in the historical development of quasi-Newton methods. According to a survey by Dennis and Moré [21], the motivation which led to the derivation of PSB update and in fact later on to a whole new class of quasi-Newton methods using Powell's technique, was due to the fact that Symmetric Rank-1 (SR1) update has a similar numerical instability issue. A similar issue could happen in BFGS formula, where the update is

$$B_{k+1} = B_k - \frac{B_k s_k {s_k}^T B_k}{{s_k}^T B_k s_k} + \frac{y_k {y_k}^T}{{y_k}^T s_k} \ . \tag{BFGS}$$

The advantage of BFGS versus Broyden's method is that one can guarantee the denominator $s_k^T y_k > 0$ by imposing the Wolfe condition [4]. In fact our experiment with Broyden method shows that it can be unstable even when applied to simple bilinear problems. Additionally, another major drawback of Broyden method is that unlike BFGS, it is not self-correcting. B_k in Broyden method depends on each B_j with $j \le k$, and it might carry along irrelevant information for a long time [21].

3 Local Q-Superlinear Convergence of the J-symmetric Update.

In this section we present the local Q-superlinear convergence of Algorithm 1. In the local setting, we assume the initial solution z_0 and initial estimation of the Jacobian B_0 to be chosen from a close neighbourhood of z^* and $\nabla F(z^*)$, respectively. Here we assume

Assumption 3.1. *(Assumptions for Local Superlinear Convergence)*

(a) There exists a minimax solution z^* such that $F(z^*) = 0$, and $\nabla F(z^*)$ is invertible with $\gamma = \|\nabla F^{-1}(z^*)\|$.

(b) There exist a nonzero open ball of radius ϵ centered at z^*, $B_\epsilon(z^*) := \{z \| \|z - z^*\| < \epsilon\}$, such that for any $z \in B_\epsilon(z^*)$, it holds that:

$$\|\nabla F(z) - \nabla F(z^*)\| \leq \Lambda \|z - z^*\| \ . \tag{20}$$

In the local convergence, we consider a ball $B_\epsilon(z^*)$ around a minimax solution z^*. Assumption 3.1 (a) assumes the non-singularity of $\nabla F(z^*)$, and (b) assumes Lipschitz continuity of $\nabla F(z^*)$ inside $B_\epsilon(z^*)$. These assumptions are quite weak and only require the Jacobian of the solution $\nabla F(z^*)$ to be invertible and Lipschitz continuous in a neighbourhood. The local superlinear convergence of Algorithm 1 is formalized in the next theorem:

Theorem 3.2. *Consider Algorithm 1 for solving minimax problem* (1)*. Suppose there exists an optimal minimax solution z^* that satisfies Assumption 3.1. Then for any given $0 < r < 1$, there exists positive constants $\bar{\epsilon}$ and δ such that for any $z_0 \in \{\|z_0 - z^*\| < \bar{\epsilon}\}$ and $B_0 \in \{\|B_0 - \nabla F(z^*)\|_F < \delta\}$, it holds that:*
***(a)**. The sequence $\{z_k\}$ generated by Algorithm 1 is well defined and converges to z^*, and $\{\|B_k\|\}$ and $\{\|B_k^{-1}\|\}$ are uniformly bounded for any $k \geq 0$. Additionally,*

$$\|z_{k+1} - z^*\| \leq r \|z_k - z^*\| \ . \tag{21}$$

***(b)**. The iterates $\{z_k\}$ enjoys Q-superlinear convergence towards z^*.*

Our local analysis in Theorem 3.2 is based on the *bounded deterioration* technique and is similar to the analysis given in [11]. To establish Theorem 3.2, we first present two lemmas which are used in the proof.

Lemma 3.3. *Suppose Assumption 3.1 holds. Then, it holds for any small enough $\epsilon > 0$ and $u, v \in B_\epsilon(z^*)$:*

(a)

$$\|F(v) - F(u) - \nabla F(z^*)(v - u)\| \leq \Lambda \max\{\|v - z^*\|, \|u - z^*\|\} \|v - u\| \ . \tag{22}$$

(b) There exists $\rho > 0$ such that

$$\frac{\|v - u\|}{\rho} \leq \|F(v) - F(u)\| \leq \rho \|v - u\| \ . \tag{23}$$

Proof. (a). Denote $T(z) = F(z) - \nabla F(z^*) z$, then $T(z)$ is differentiable by noticing $F(z)$ is differentiable and $\nabla T(z) = \nabla F(z) - \nabla F(z^*)$. By Taylor expansion at u, we obtain

$$T(v) = T(u) + \int_0^1 \nabla T\Big(u + t(v - u)\Big)(v - u) dt \ ,$$

thus,

$$\|T(v) - T(u)\| \leq \sup_{0 \leq t \leq 1} \|\nabla T\big(u + t(v - u)\big)\| \|v - u\| \ .$$

Substituting $T(z)$ to the above inequality, we obtain

$$\begin{aligned}\|F(v) - F(u) - \nabla F(z^*)(v - u)\| &\leq \sup_{0 \leq t \leq 1} \|\nabla F\big(u + t(v - u)\big) - \nabla F(z^*)\| \|v - u\| \\ &\leq \sup_{0 \leq t \leq 1} \Lambda \|u + t(v - u) - z^*\| \|v - u\| \\ &= \Lambda \max\{\|v - z^*\|, \|u - z^*\|\} \|v - u\| \ ,\end{aligned}$$

where the second inequality uses (20).

(b). It follows from (22) by triangle inequality that

$$\|F(v) - F(u)\| \leq \Lambda \max\{\|v - z^*\|, \|u - z^*\|\}\|v - u\| + \|\nabla F(z^*)(v - u)\| \leq \big(\Lambda\epsilon + \|\nabla F(z^*)\|\big)\|v - u\| \ .$$

Furthermore, let σ be the smallest singularvalue of $\nabla F(z^*)$, then $\sigma > 0$ as $\nabla F(z^*)$ is full rank, whereby it holds for any u, v that

$$\sigma\|v - u\| \leq \|\nabla F(z^*)(v - u)\| \ .$$

Therefore, it follows from (22) that

$$\|F(v) - F(u)\| \geq \|\nabla F(z^*)(v - u)\| - \Lambda \max\{\|v - z^*\|, \|u - z^*\|\}\|v - u\| \geq (\sigma - \Lambda\epsilon)\|v - u\| \ .$$

Now suppose $\epsilon < \sigma/\Lambda$ and setting $\rho = \max\{1/(\sigma - \Lambda\epsilon)\ ,\ \Lambda\epsilon + \|\nabla F(z^*)\|\}$, we arrive at (23). □

The following lemma presents an equivalent representation of (8) which we use later.

Lemma 3.4. *Consider the B_k update rule* (8). *Then it holds that*

$$B_{k+1} = JP_kJB_kP_k + \frac{y_k s_k^T}{s_k^T s_k} + \frac{Js_k y_k^T J}{s_k^T s_k} P_k \ , \tag{24}$$

where P_k is the projection matrix defined in (19). *Furthermore, we have*

$$\|B_{k+1} - \nabla F(z^*)\|_F \leq \sqrt{(1 - \theta_{1,k}^2)(1 - \theta_{2,k}^2)}\|B_k - \nabla F(z^*)\|_F + (1 + \sqrt{n + m - 1})\frac{\|y_k - \nabla F(z^*)s_k\|}{\|s_k\|} \ , \tag{25}$$

where

$$\theta_{1,k} = \frac{\|JP_kJ(B_k - \nabla F(z^*))s_k\|}{\|s_k\|\|JP_kJ(B_k - \nabla F(z^*))\|_F} \ , \quad \text{and} \quad \theta_{2,k} = \frac{\|(B_k - \nabla F(z^*))^T Js_k\|}{\|s_k\|\|B_k - \nabla F(z^*)\|_F} \ . \tag{26}$$

Proof. First note that P_k is the projection matrix onto the $m + n - 1$ dimension subspace which is perpendicular to s_k, thus

$$\|P_k\| = 1 \ , \quad \|P_k\|_F = \sqrt{m + n - 1} \ . \tag{27}$$

Let O be any J-symmetric matrix with proper size and let $M = B_k - O$ and $\bar{M} = B_{k+1} - O$, then we claim the following holds:

$$\bar{M} = JP_kJMP_k + \frac{(y_k - Os_k)s_k^T}{s_k^T s_k} + \frac{Js_k(y_k - Os_k)^T J}{s_k^T s_k} P_k \ . \tag{28}$$

This is because from (8) we have

$$\begin{aligned} B_{k+1} - O = \ & B_k - O + \frac{Js_k(y_k - B_ks_k + Os_k - Os_k)^T J}{s_k^T s_k} + \frac{(y_k - B_ks_k + Os_k - Os_k)s_k^T}{s_k^T s_k} - \\ & \frac{(Js_k)^T(y_k - B_ks_k + Os_k - Os_k)Js_ks_k^T}{(s_k^T s_k)^2} \ . \end{aligned}$$

Substituting $B_{k+1} - O = \bar{M}$ and $B_k - O = M$ we obtain

$$
\begin{aligned}
\bar{M} = \; & M - \frac{Js_k s_k^T M^T J}{s_k^T s_k} + \frac{Js_k(y_k - Os_k)^T J}{s_k^T s_k} - \frac{Ms_k s_k^T}{s_k^T s_k} + \frac{(y_k - Os_k)s_k^T}{s_k^T s_k} + \frac{(Js_k)^T Ms_k Js_k s_k^T}{(s_k^T s_k)^2} - \frac{(Js_k)^T(y_k - Os_k)Js_k s_k^T}{(s_k^T s_k)^2} \\
= \; & M - \frac{Js_k s_k^T M^T J}{s_k^T s_k} - \frac{Ms_k s_k^T}{s_k^T s_k} + \frac{(Js_k)^T Ms_k Js_k s_k^T}{(s_k^T s_k)^2} + \frac{Js_k(y_k - Os_k)^T J}{s_k^T s_k} + \frac{(y_k - Os_k)s_k^T}{s_k^T s_k} - \frac{(Js_k)^T(y_k - Os_k)Js_k s_k^T}{(s_k^T s_k)^2} \\
= \; & M - \frac{Js_k s_k^T JM}{s_k^T s_k} - \frac{Ms_k s_k^T}{s_k^T s_k} + \frac{Js_k(Js_k)^T Ms_k s_k^T}{(s_k^T s_k)^2} + \frac{Js_k(y_k - Os_k)^T J}{s_k^T s_k} + \frac{(y_k - Os_k)s_k^T}{s_k^T s_k} - \frac{(Js_k)^T(y_k - Os_k)Js_k s_k^T}{(s_k^T s_k)^2} \\
= \; & JP_k JMP_k + \frac{(y_k - Os_k)s_k^T}{s_k^T s_k} + \frac{Js_k(y_k - Os_k)^T J}{s_k^T s_k} P_k \;,
\end{aligned}
$$

where the second equality comes from rearrangement. The third equality uses the fact that M is J-symmetric, thus we have $M^T J = JM$, and the fact that $(Js_k)^T Ms_k$ is a scalar thus we have $(Js_k)^T Ms_k Js_k s_k^T = Js_k(Js_k)^T Ms_k s_k^T$. The last equality uses $J^2 = I$ and (19) thus $JP_k J = I - J(s_k s_k^T)J/(s_k^T s_k)$, and therefore the sum of the first four terms in the third line is exactly $JP_k JMP_k$. Additionally, in the final term in the same line, $(Js_k)^T(y_k - Os_k)$ is a scalar, so we can use $(Js_k)^T(y_k - Os_k)Js_k s_k^T = Js_k(y_k - Os_k)^T Js_k s_k^T$. By factoring out $\big(Js_k(y_k - Os_k)^T J\big)/(s_k^T s_k)$ from this term and the fifth term and recalling (19) we arrive at (28).

Utilizing (28) and setting O equal to the zero we obtain (24) and therefore conclude that the update rule (8) is equivalent to (24).

To show (25), we start by bounding the first term in (28) as following:

$$
\begin{aligned}
\|JP_k JMP_k\|_F^2 &= \Big\|JP_k JM\big(I - \frac{s_k s_k^T}{s_k^T s_k}\big)\Big\|_F^2 \\
&= \|JP_k JM\|_F^2 - 2\frac{\|JP_k JMs_k\|^2}{s_k^T s_k} + \frac{\|JP_k JMs_k\|^2}{s_k^T s_k} \\
&= \Big(1 - \frac{\|JP_k JMs_k\|^2}{\|s_k\|^2 \|JP_k JM\|_F^2}\Big)\|JP_k JM\|_F^2 \;,
\end{aligned}
$$

where the first equality uses (19) and the second equality follows directly from the definition of the Frobenius norm. Furthermore,

$$
\begin{aligned}
\|JP_k JM\|_F^2 = \|M^T JP_k J\|_F^2 &= \Big\|M^T\big(I - \frac{Js_k(Js_k)^T}{s_k^T s_k}\big)\Big\|_F^2 \\
&= \|M^T\|_F^2 - 2\frac{(Js_k)^T MM^T Js_k}{s_k^T s_k} + \frac{\|M^T Js_k\|^2}{s_k^T s_k} \\
&= \|M\|_F^2 - \frac{\|M^T Js_k\|^2}{\|s_k\|^2} \\
&= \Big(1 - \frac{\|M^T Js_k\|^2}{\|s_k\|^2 \|M\|_F^2}\Big)\|M\|_F^2 \;,
\end{aligned}
$$

where the second equality uses $JP_k J = I - J(s_k s_k^T)J/(s_k^T s_k)$ and the third equality follows directly from the definition of the Frobenius norm. For the remainder of this proof we set $O = \nabla F(z^*)$, so we obtain

$$
\theta_{1,k} = \frac{\|JP_k JMs_k\|}{\|s_k\| \|JP_k JM\|_F} \quad \text{and,} \qquad \theta_{2,k} = \frac{\|M^T Js_k\|}{\|s_k\| \|M\|_F} \;.
$$

By Cauchy-Schwarz inequality we know $\|JP_k JMs_k\|/(\|s_k\| \|JP_k JM\|) \leq 1$ and since induced l_2 norm is less than Frobenius norm, we conclude: $0 < \theta_{1,k} \leq 1$. Similarly, we obtain $\|M^T Js_k\|/(\|Js_k\| \|M^T\|_F) \leq 1$ and since$\|Js_k\| = \|s_k\|$ and $\|M^T\|_F = \|M\|_F$, we conclude $0 < \theta_{2,k} \leq 1$. Hence, we can safely take square root from both sides and arrive at:

$$
\|JP_k JMP_k\|_F = \sqrt{(1 - \theta_{1,k}^2)(1 - \theta_{2,k}^2)}\|M\|_F \;. \tag{29}
$$

We obtain the following equality for the norm of the second term (recall $O = \nabla F(z^*)$) in (28):

$$\left\| \frac{\big(y_k - \nabla F(z^*) s_k\big) s_k^T}{s_k^T s_k} \right\|_F = \frac{\sqrt{\mathrm{Tr}\Big(\big(y_k - \nabla F(z^*) s_k\big) s_k^T s_k \big(y_k - \nabla F(z^*) s_k\big)^T\Big)}}{\|s_k\|^2} = \frac{\|y_k - \nabla F(z^*) s_k\|}{\|s_k\|} \; . \tag{30}$$

Finally from the application of the inequality $\|AB\|_F \leq \|A\| \|B\|_F$ (see Lemma A.2 in the appendix), to the third term in (28) we obtain:

$$\begin{aligned}\left\| \frac{J s_k \big(y_k - \nabla F(z^*) s_k\big)^T J}{s_k^T s_k} P_k \right\|_F &\leq \|P_k\|_F \frac{\|J s_k \big(y_k - \nabla F(z^*) s_k\big)^T J\|}{\|s_k\|^2} \\ &\leq \|P_k\|_F \frac{\|J s_k\| \|\big(y_k - \nabla F(z^*) s_k\big)^T J\|}{\|s_k\|^2} = \sqrt{n+m-1} \frac{\|y_k - \nabla F(z^*) s_k\|}{\|s_k\|} \; , \end{aligned} \tag{31}$$

where we use the fact that $\|Jq\| = \|q\|$ for any vector q of the appropriate size and (27) in the final equality. Combining (29),(30) and (31), and then substituting $\bar{M} = B_{k+1} - \nabla F(z^*)$ and $M = B_k - \nabla F(z^*)$, we obtain (25). ☐

Proposition 3.5. *Suppose Assumption 3.1 holds. Recall that $\gamma = \|\nabla F^{-1}(z^*)\|$. For any given $z_k \in B_\epsilon(z^*)$ and invertible J-symmetric matrix B_k such that $\|B_k^{-1}\| < 2\gamma$, we have $z_{k+1} \in B_\epsilon(z^*)$, where z_{k+1} is obtained from (5). Moreover, if B_{k+1} is obtained from (8), we have*

$$\begin{aligned}\|B_{k+1} - \nabla F(z^*)\|_F \; &\leq \sqrt{(1-\theta_{1,k}^2)(1-\theta_{2,k}^2)} \|B_k - \nabla F(z^*)\|_F \\ &\quad + \Lambda(1 + \sqrt{n+m-1}) \max\{\|z_{k+1} - z^*\|, \|z_k - z^*\|\} \; . \end{aligned} \tag{32}$$

Proof. Starting from (5) we have

$$\|z_{k+1} - z_k\| = \|B_k^{-1} F(z_k)\| \leq \|B_k^{-1}\| \|F(z_k)\| \leq 2\gamma \|F(z_k)\| \; .$$

Since $F(z^*) = 0$ it follows from (23) that $\|F(z_k)\| \leq \rho \|z_k - z^*\|$, so, $\|z_{k+1} - z_k\| \leq 2\rho\gamma \|z_k - z^*\|$. By further restricting z_k such that

$$\|z_k - z^*\| < \min\left\{ \epsilon/2 \; , \; \frac{\epsilon/2}{2\rho\gamma} \right\} \; ,$$

we obtain $\|z_{k+1} - z^*\| \leq \|z_{k+1} - z_k\| + \|z_k - z^*\| < \epsilon$, and therefore it holds that $z_{k+1} \in B_\epsilon(z^*)$. Applying (22), we obtain

$$\begin{aligned}\|F(z_{k+1}) - F(z_k) - \nabla F(z^*)(z_{k+1} - z_k)\| &\leq \Lambda \max\{\|z_{k+1} - z^*\|, \|z_k - z^*\|\} \|z_{k+1} - z_k\| \\ \frac{\|y_k - \nabla F(z^*) s_k\|}{\|s_k\|} &\leq \Lambda \max\{\|z_{k+1} - z^*\|, \|z_k - z^*\|\} \; . \end{aligned}$$

We arrive at (32) by substituting the above inequality into (25). ☐

Now we are ready to prove Theorem 3.2:

Proof of Theorem 3.2. Set

$$\delta = \frac{r}{\gamma(r+1)\Big(\frac{1-r}{1+\sqrt{m+n-1}} + 2\Big)} \; , \tag{33}$$

and

$$\bar{\epsilon} = \min\left\{ \frac{(1-r)\delta}{\Lambda(1 + \sqrt{m+n-1})} \; , \; \epsilon \right\} \; , \tag{34}$$

then it holds that:

$$\gamma(r+1)\big(\Lambda\bar{\epsilon}+2\delta\big) \le r \ . \tag{35}$$

We prove part **(a)** by induction. We begin with $k=0$.
From $\|B_0 - \nabla F(z^*)\|_F < \delta$ we know $\|B_0 - \nabla F(z^*)\| < \delta < 2\delta$, and recall $\|\nabla F^{-1}(z^*)\| = \gamma$. Notice that (35) implies $\gamma 2\delta < r/(r+1) < 1$. So, we can apply Banach Perturbation Lemma (see Lemma A.1 in the appendix) to the matrices $\nabla F(z^*)$ and B_0, and obtain

$$\|B_0^{-1}\| \le \frac{\gamma}{1 - r/(1+r)} = \gamma(r+1) \ . \tag{36}$$

To prove (21), recall that $F(z^*) = 0$, and since $\|z_0 - z^*\| < \bar{\epsilon}$, Lemma 3.3 applies. From (5) we have

$$\begin{aligned}
\|z_1 - z^*\| &= \|B_0^{-1}F(z_0) - (z_0 - z^*)\| \\
&= \|B_0^{-1}F(z_0) - B_0^{-1}\nabla F(z^*)(z_0 - z^*) + B_0^{-1}\nabla F(z^*)(z_0 - z^*) - (z_0 - z^*)\| \\
&\le \|B_0^{-1}\|\Big(\|F(z_0) - F(z^*) - \nabla F(z^*)(z_0 - z^*)\| + \|\nabla F(z^*) - B_0\|\|z_0 - z^*\|\Big) \\
&\le \gamma(r+1)(\Lambda\bar{\epsilon}+2\delta)\|z_0 - z^*\| \ ,
\end{aligned}$$

where in the final inequality we use (36) and (22). By applying (35) to this inequality we obtain

$$\|z_1 - z^*\| \le r\|z_0 - z^*\| \ . \tag{37}$$

This implies $\|z_1 - z^*\| < \bar{\epsilon} \le \epsilon$ and hence $z_1 \in B_\epsilon(z^*)$. Now we prove the claims for $k = K$, assuming (36) and (37) hold for $k = 0, \ldots, K-1$. Notice that (36) implies $\|B_{K-1}^{-1}\| \le 2\gamma$ and hence we can apply Proposition 3.5 and via (32) together with $\|z_K - z^*\| \le r\|z_{K-1} - z^*\|$ conclude that

$$\|B_K - \nabla F(z^*)\|_F \ \le \ \|B_{K-1} - \nabla F(z^*)\|_F + \Lambda\big(1+\sqrt{n+m-1}\big)\|z_{K-1} - z^*\| \ .$$

Summing up from $k=0$ to $k = K-1$ we have

$$\|B_K - \nabla F(z^*)\|_F \ \le \|B_0 - \nabla F(z^*)\|_F + \Lambda\big(1+\sqrt{n+m-1}\big)\bar{\epsilon}\frac{1-r^K}{1-r} \ . \tag{38}$$

From (34) we have $\Lambda\big(1+\sqrt{n+m-1}\big)\bar{\epsilon}/(1-r) \le \delta$ and recalling $\|B_0 - \nabla F(z^*)\|_F < \delta$, we conclude:

$$\|B_K - \nabla F(z^*)\|_F \ < 2\delta \ .$$

Using this inequality and $\gamma = \|\nabla F^{-1}(z^*)\|$, via Banach Perturbation Lemma and with the same exact proof as we did for $k=0$, we conclude

$$\|B_K^{-1}\| \le \gamma(r+1) \ . \tag{39}$$

Now let us prove (21) for $k = K$. Notice

$$\begin{aligned}
\|z_{K+1} - z^*\| &= \|B_K^{-1}F(z_K) - (z_K - z^*)\| \\
&\le \|B_K^{-1}\|\Big(\|F(z_K) - F(z^*) - \nabla F(z^*)(z_K - z^*)\| + \|\nabla F(z^*) - B_K\|\|z_K - z^*\|\Big) \\
&\le \gamma(r+1)\Big(\Lambda\bar{\epsilon}+2\delta\Big)\|z_K - z^*\| \ .
\end{aligned}$$

Thus, we get $\|z_{K+1} - z^*\| \le r\|z_K - z^*\|$, which completes the proof of part **(a)** by induction.

Next we move to part **(b)** to show (5) is Q-superlinearly convergent. As a result of part **(a)**, Proposition 3.5 applies for all k. In (32) define $\bar{\theta}_k = (\theta_{1,k} + \theta_{2,k})/2$ and since $\sqrt{(1-\theta_{1,k}^2)(1-\theta_{2,k}^2)} \le 1 - \bar{\theta}_k$, together with (21), we deduce

$$\bar{\theta}_k\|B_k - \nabla F(z^*)\|_F \le \|B_k - \nabla F(z^*)\|_F - \|B_{k+1} - \nabla F(z^*)\|_F + \Lambda(1+\sqrt{n+m-1})\|z_k - z^*\| \ .$$

Summing up for $k = 0, \dots, \infty$ we obtain $\sum_{k=0}^{\infty} \bar{\theta}_k \|B_k - \nabla F(z^*)\|_F$ in the L.H.S. and since we know that the R.H.S. is bounded above (see (38)) we conclude

$$\lim_{k\to\infty} \bar{\theta}_k \|B_k - \nabla F(z^*)\|_F = \frac{1}{2} \lim_{k\to\infty} (\theta_{1,k} + \theta_{2,k}) \|B_k - \nabla F(z^*)\|_F = 0 \ .$$

Since both $\theta_{1,k}$ and $\theta_{2,k}$ are positive, we conclude: $\lim_{k\to\infty} \theta_{1,k} \|B_k - \nabla F(z^*)\|_F = 0$ and $\lim_{k\to\infty} \theta_{2,k} \|B_k - \nabla F(z^*)\|_F = 0$. Substituting $\theta_{2,k}$ from (26) followed by replacing $\left(B_k - \nabla F(z^*)\right)^T J = J\left(B_k - \nabla F(z^*)\right)$ we obtain

$$\lim_{k\to\infty} \frac{\left\| J\big(B_k - \nabla F(z^*)\big) s_k \right\|}{\|s_k\|} = 0 \ ,$$

and recalling that $\|Jq\| = \|q\|$ for any vector $q \in \mathbb{R}^{n+m}$, we conclude

$$\lim_{k\to\infty} \frac{\left\| \big(B_k - \nabla F(z^*)\big) s_k \right\|}{\|s_k\|} = 0 \ .$$

This is Dennis-Moré characterization identity for Q-superlinear convergence (see Theorem A.4 in the appendix) and therefore the proof is finished. $\square$

4 A Globally Convergent J-symmetric quasi-Newton Method.

The previous section establishes the local superlinear convergence of the J-Symmetric quasi-Newton method. In this section, we present a trust-region J-symmetric quasi-Newton method (Algorithm 3), and show its global superlinear convergence guarantees. To present our algorithm, we first introduce a merit function minimization problem:

$$\min_{z\in\mathbb{R}^{n+m}} \frac{1}{2}\|F(z)\|^2 \ . \tag{40}$$

Then it is straight-forward to see that the global minimizers to (40) are exactly the same as the saddle points to (1). Furthermore, we define $m_k(s)$ as the quadratic model of the merit function at z_k:

$$m_k(s) := \frac{1}{2}\|F(z_k)\|^2 + g_k^T s + \frac{1}{2} s^T B_k^T B_k s \ , \tag{41}$$

where

$$g_k = \nabla F(z_k)^T F(z_k) \ ,$$

is the gradient of the merit function and B_k is the estimation of the Jacobian $\nabla F(z_k)$ (see the below update rule (47)). Then, $m_k(s)$ is an approximated second order expansion of the merit function around z_k. Here we would like to highlight that (i) while the calculation of g_k involves $\nabla F(z_k)$, it can be performed efficiently by using fast Hessian-vector product for many applications [41, 46]; (ii) similar to other quasi-Newton methods, we can store B_k^{-1} in memory and update B_{k+1}^{-1} by a low-rank operation. As a result, calculating the minimizer of the quadratic model (41) only involves matrix-vector multiplication, in contrast to Newton's method which involves solving linear equations. In other words, it has the same order of cost-per-iteration as a first-order method in general.

Algorithm 3 presents our Trust-region J-Symmetric Algorithm. We initialize with solution z_0, Jacobian estimation B_0, trust-region radius upper bound R_0, initial trust-region radius $\Delta_0 \in (0,\ R_0]$, and valid step (sufficient decrease) parameter $\zeta \in (0, 10^{-3})$. In the k-th iteration of the algorithm, there are three potential valid steps (i) the quasi-Newton step p_k^B, (ii) the Cauchy point step p_k^C, and (iii) the dogleg step p_k^D, as defined below:

(Quasi-Newton step p_k^B). The quasi-Newton point p_k^B is defined as the global minimizer of $m_k(s)$, namely

$$p_k^B = -B_k^{-1}(B_k^{-1})^T g_k \ . \tag{42}$$

(Cauchy point step p_k^C). The Cauchy point is defined as the minimizer of $m_k(s)$ over the trust-region along the negative gradient direction:

$$p_k^C := -\tau_k(\Delta_k/\|g_k\|)g_k \ ,$$

where $\tau_k := \arg\min_{0\le\tau\le 1} \ m_k\Big(\tau\Delta_k g_k/\|g_k\|\Big)$. The Cauchy point has the following closed-form solution

$$p_k^C = -\min\Big\{\|g_k\|^2/({g_k}^T B_k^T B_k g_k) \ , \ \Delta_k/\|g_k\|\Big\}g_k \ . \tag{43}$$

(Dogleg step p_k^D). When the Cauchy point is strictly inside the trust-region ($\|p_k^C\| < \Delta_k$) and the Quasi-Newton step p_k^B is strictly outside the trust-region ($\|p_k^B\| > \Delta_k$), that is when $p_k^C = -\Big(\|g_k\|^2/({g_k}^T B_k^T B_k g_k)\Big)g_k$, we then define the dogleg point as

$$p_k^D = p_k^C + \alpha(p_k^B - p_k^C) \ , \tag{44}$$

where $\alpha \in (0,1)$ is the unique solution that satisfies $\|p_k^C + \alpha(p_k^B - p_k^C)\| = \Delta_k$.

To update the iterate solution, we first calculate the quasi-Newton step p_k^B. If p_k^B is inside the trust-region, we take the quasi-Newton step. Otherwise, we calculate the Cauchy point step p_k^C. If the Cauchy point step is on the boundary of the trust-region, we take the Cauchy point step, otherwise, we compute and take the dogleg step p_k^D. In summary, we set the step s_k as

$$s_k = \begin{cases} p_k^B & \text{if } \|p_k^B\| \le \Delta_k \ , \\ p_k^C & \text{if } \|p_k^B\| > \Delta_k \text{ and } \|p_k^C\| = \Delta_k \ , \\ p_k^D & \text{if } \|p_k^B\| > \Delta_k \text{ and } \|p_k^C\| < \Delta_k \ . \end{cases} \tag{45}$$

Then it is obvious that the step s_k is always within the trust-region, namely, $\|s_k\| \le \Delta_k$.

Next, we compute the ratio between the actual decay and the predicted decay of the merit function ρ_k as

$$\rho_k := \frac{\|F(z_k)\|^2/2 - \|F(z_k + s_k)\|^2/2}{m_k(0) - m_k(s_k)} \ . \tag{46}$$

If the ratio ρ_k is reasonably large (i.e., $\rho_k > 0.5$), the step s_k provides sufficient decay on the merit function, and we safely expand the trust-region radius (recall R_0 is the maximal trust-region radius specified by the user):

$$\Delta_{k+1} = \min\{2\Delta_k, R_0\} \ ,$$

otherwise, we reduce the trust-region radius:

$$\Delta_{k+1} = \Delta_k/2 \ .$$

Moreover, if ρ_k is not too small (i.e., $\rho_k \ge \zeta \in (0, 10^{-3})$), we update the iterate solution by accepting the step $z_{k+1} = z_k + s_k$, and call it a *valid step*; otherwise we reject the update and take a *null step* by setting $z_{k+1} = z_k$.

Finally, we update the Jacobian estimation B_{k+1} by a slightly modified version of (8) in order to guarantee the non-singularity of B_{k+1}:

$$B_{k+1} = B_k + \beta_k \frac{Js_k(y_k - B_k s_k)^T J + (y_k - B_k s_k)s_k^T}{s_k^T s_k} - \beta_k^2 \frac{(Js_k)^T(y_k - B_k s_k)Js_k s_k^T}{(s_k^T s_k)^2} \ , \tag{47}$$

Algorithm 3 J-symmetric Quasi-Newton Method with Trust-region (J-symm-Tr)

Initialize with solution $z_0 \in \mathbb{R}^{m+n}$, Jacobian estimation $B_0 \in \mathbb{R}^{(m+n)\times(m+n)}$, maximum allowed trust-region radius $R_0 > 0$, initial trust-region radius $\Delta_0 \in (0,\ R_0]$, parameter $\hat{\beta} = 0.9$, sufficient decrease threshold $\zeta \in (0, 10^{-3})$ and iteration counter $k = 0$.
for $k = 1, 2, 3, \ldots,$ **do**
  compute p_k^B via (42)
  if $\|p_k^B\| \le \Delta_k$ **then**
    $s_k = p_k^B$
  else
    compute p_k^C via (43)
    if $\|p_k^C\| = \Delta_k$ **then**
      $s_k = p_k^C$
    else
      compute p_k^D via (44)
      set $s_k = p_k^D$
    end if
  end if
  evaluate ρ_k from (46)
  if $\rho_k \le 0.5$ **then**
    $\Delta_{k+1} = \Delta_k/2$
  else
    $\Delta_{k+1} = \min\{2\Delta_k\ ,\ R_0\}$
  end if
  if $\rho_k \ge \zeta$ **then**
    $z_{k+1} = z_k + s_k$
  else
    $z_{k+1} = z_k$
  end if
  $y_k = F(z_k + s_k) - F(z_k)$
  update B_{k+1} via (47) with β_k uniformly randomly chosen from $[1-\hat{\beta}, 1+\hat{\beta}]$
  $k = k + 1$
end for

where for any given $\hat{\beta} \in (0,1)$ we pick $1 - \hat{\beta} \le \beta_k \le 1 + \hat{\beta}$ such that B_{k+1} is nonsingular for any k. Indeed, suppose B_k is nonsingular, then there only exists finite number of β_k such that B_{k+1} is singular, thus B_{k+1} is nonsingular with probability 1 if we randomly pick β_k uniformly from the range $[1-\hat{\beta}, 1+\hat{\beta}]$. This strategy dates back to Powell [43].

In the rest of this section, we present the global convergence and local superlinear convergence of Algorithm 3. First, we define the level set of the merit function as $S = \{z \mid \|F(z)\|^2/2 \le \|F(z_0)\|^2/2\}$, and the extended level set as

$$S(R_0) := \{z + s \mid \|s\| < R_0 \text{ for some } z \in S\}\ .$$

The following assumptions are needed to develop the global convergence results of Algorithm 3:

Assumption 4.1. *(Assumptions for Global Convergence)*

(a) For any $R_0 > 0$, $F(z)$ and $\nabla F(z)$ are Lipschitz continuous in $S(R_0)$ namely, there exist constants γ_1 and γ_2 such that it holds for any $z, z+s \in S(R_0)$ that

$$\|F(z) - F(z+s)\| \le \gamma_1\|s\| \quad \textit{and} \quad \|\nabla F(z) - \nabla F(z+s)\| \le \gamma_2\|s\|\ .$$

(b) *There exists at least one z^* such that $F(z^*) = 0$. Furthermore, $\nabla F(z^*)$ is invertible for all saddle point z^*, and there exists γ such that $\gamma \geq \|\nabla F^{-1}(z^*)\|$.*

(c) *The sequence of vectors $\{s_k\}$ is uniformly linearly independent* [1].

We here examine Assumption 4.1. Part (a) impose regularity conditions on the function L (or equivalently on the function F). Since $F(z)$ is twice continuously differentiable and if the level set S is bounded, then (a) automatically holds. Part (b) assumes the existence of (at least one) saddle point z^*, and furthermore, the saddle point z^* is non-degenerate (i.e., $\nabla F(z^*)$ is invertible). Part (c) implies that every $n+m$ consecutive steps in the sequence $\{s_k/\|s_k\|\}$ span the entire $\mathbb{R}^{n+m}$. The non-degenerate assumption (b) and the uniformly linearly independent assumption (c) are the classic assumptions for obtaining the global convergence of a quasi-Newton method for a minimization problem. As an example see [38, Theorem 6.2] which requires such conditions in order for SR1 update to generate a good Hessian approximation. We here extend them to minimax problems.

Our main theoretical results are presented in the following two theorems:

Theorem 4.2. *Consider Algorithm 3 to solve the minimax problem* (1). *Under Assumption 4.1, it holds that the sequence $\{g_k\}$ generated by Algorithm 3 converges to 0, that is*

$$\lim_{k\to\infty} \|g_k\| = 0 \ . \tag{48}$$

Theorem 4.2 states that under Assumption 4.1 Algorithm 3 generate iterates such that the gradient g_k converges to 0. As a direct consequence of Theorem 4.2, we know that if $\nabla F(z_k)$ is nonsingular and bounded, then $F(z_k)$ converges to 0 by noticing $g_k = \nabla F(z_k)^T F(z_k)$. Furthermore, if the saddle point solution z^* is unique, Theorem 4.2 implies $z_k \to z^*$. Similar arguments appear in Powell's hybrid algorithm for minimization problem [42].

Now we assume $\{z_k\}$ converges to a stationary solution, then the next theorem states that (1) $\{B_k\}$ must converge to $\nabla F(z^*)$, namely B_k eventually provides a good approximation of the Jacobian; (2) states that Algorithm 3 is R-superlinearly convergent, which showcases the global convergence property of Algorithm 3.

Theorem 4.3 assumes that $\{s_k\}$ converges to zero. This is a reasonable assumption since $z_k \to z^*$ and equivalently $\|z_{k+1} - z_k\| \to 0$, so we know that the valid step subsequence of $\{s_k\}$ converges to 0. To guarantee $\{s_k\} \to 0$ we can modify Algorithm 3 such that when s_k is a null step with a norm greater than previous valid step norm, we reset $\|s_k\|$ to be equal to that norm. This modification does not change any of our results and the only extra cost is to recompute $F(z_k + s_k)$ which is needed to compute y_k. In order to keep Algorithm 3 simple, we drop this modification and instead assume $\{s_k\} \to 0$.

Theorem 4.3. *Consider Algorithm 3 to solve the minimax problem* (1). *Suppose Assumption 4.1 holds, $\{z_k\}$ converges to a saddle point z^* such that $F(z^*) = 0$ and $\{s_k\}$ converges to zero, then it holds that*

1. *$\{B_k\}$ converges to $\nabla F(z^*)$.*

2. *Algorithm 3 is R-superlinearly convergent to z^*.*

In the remainder of this section, we present proofs for the above two theorems. We start with presenting three simple facts:

Fact 4.4. *As a direct consequence of Assumption 4.1 we have $\|\nabla F(z)\| \leq \gamma_1$ and $\|\nabla^2 F(z)\| \leq \gamma_2$.*

[1] see Definition A.5 in the appendix for a formal definition of uniform linear independence.

Fact 4.5. *Let $D_0 = \|z_0 - z^*\|$. Then for any $z \in S(R_0)$, $\|F(z)\|$ is upper-bounded as*

$$\|F(z)\| \leq \gamma_1(D_0 + R_0) \ .$$

Fact 4.6. *Denote $\mu = \gamma_2\gamma_1(D_0 + R_0) + \gamma_1^2$. Then it holds for any $z \in S$,and $\|s\| \leq R_0$ that*

$$\|\nabla F(z)^T F(z) - \nabla F(z+s)^T F(z+s)\| \leq \mu\|s\| \ . \tag{49}$$

Proof. It holds that

$$\begin{aligned}
&\|\nabla F(z)^T F(z) - \nabla F(z+s)^T F(z+s)\| \\
&= \left\| \Big(\nabla F(z) - \nabla F(z+s)\Big)^T F(z) + \nabla F(z+s)^T \Big(F(z) - F(z+s)\Big) \right\| \\
&\leq \gamma_2 \|F(z)\|\|s\| + \gamma_1^2\|s\| \\
&\leq \Big(\gamma_2\gamma_1(D_0 + R_0) + \gamma_1^2\Big)\|s\| \ ,
\end{aligned}$$

the first equality comes from adding and subtracting $\nabla F(z+s)^T F(z)$, the following inequality comes from the Lipschitz-continuity of $F(z)$ and $\nabla F(z)$, and finally the last inequality uses Fact 4.5. ☐

The proof of theorem 4.2 heavily relies on the following two propositions. Proposition 4.7 shows that $\|B_k\|$ is always upper-bounded. Proposition 4.8 shows that $m_k(s)$ has sufficient decay in Algorithm 3.

Proposition 4.7. *Suppose Assumption 4.1 holds. Then there exists ν_2 such that it holds for any $k \geq 0$ that $\|B_k\| \leq \nu_2$.*

Proposition 4.8. *Algorithm 3 generates steps s_k such that for all k we have:*

$$m_k(0) - m_k(s_k) \geq \frac{\|g_k\|}{2} \min\left\{\Delta_k \ , \ \frac{\|g_k\|}{\nu_2^2}\right\} \ .$$

Remark 4.9. *As a direct consequence of Proposition 4.8, Algorithm 3 is a nonincreasing algorithm in $\|F(z_k)\|$, namely, $\|F(z_{k+1})\| \leq \|F_k\|$ for all iterate k. This is because (i) if the k-th step is a null step, then $z_{k+1} = z_k$ thus it is a nonincreasing step; (ii) if the k-th step is a valid step, then*

$$\|F(z_k)\|^2/2 - \|F(z_{k+1})\|^2/2 \geq \zeta\big(m_k(0) - m_k(s_k)\big) \geq 0 \ ,$$

where the last inequality is from Proposition 4.8. Furthermore, if $g_k \neq 0$, then Proposition 4.8 shows that a valid step of Algorithm 3 provides sufficient decay in the merit function $\|F(z)\|^2/2$. This observation is the cornerstone of the convergence results of Algorithm 3.

To show Proposition 4.7 and Proposition 4.8, we first establish two simple lemmas to better understand the update rule of B_k.

Lemma 4.10. *Let*

$$Q_k = I - \beta_k \frac{s_k s_k^T}{s_k^T s_k} \ . \tag{50}$$

Then it holds that $\|Q_k\| \leq 1$. Furthermore, under Assumption 4.1, there exists a constant $\theta \in (0, 1)$ and an index K such that for $k \geq K$ we have:

$$\left\| \prod_{j=k+1}^{k+n+m} Q_j \right\| \leq \theta \ .$$

Proof. Notice that it holds for any vector $v \in \mathbb{R}^{m+n}$ that $\|Q_k v\|^2 = \|v\|^2 - \beta_k(2-\beta_k)(v^T s_k)^2/\|s_k\|^2$, and since $0 < 1-\hat{\beta} \leq \beta_k \leq 1+\hat{\beta} < 2$, then, $\|Q_k v\| \leq \|v\|$. Hence, $\|Q_k\| \leq 1$. Furthermore, the existence of such K and $\theta \in (0,1)$ is from Theorem A.6 in the appendix, following the uniform linear independence assumption of $\{s_k/\|s_k\|\}$ and $|1-\beta_k| \leq \hat{\beta}$. $\square$

Lemma 4.11. *For any J-symmetric matrix $O \in \mathbb{R}^{(n+m)\times(n+m)}$, let $M_k = B_k - O$ and $M_{k+1} = B_{k+1} - O$, then it holds that*

$$M_{k+1} = JQ_kJM_kQ_k + \beta_k \frac{(y_k - Os_k)s_k^T}{s_k^T s_k} + \beta_k \frac{Js_k(y_k - Os_k)^T J}{s_k^T s_k} Q_k \ , \tag{51}$$

and in particular, we obtain the following equivalent representation of (47) *by letting $O = 0$:*

$$B_{k+1} = JQ_kJB_kQ_k + \beta_k \frac{y_k s_k^T}{s_k^T s_k} + \beta_k \frac{Js_k y_k^T J}{s_k^T s_k} Q_k \ , \tag{52}$$

where Q_k is defined in (50).

Proof. Notice that we can write (47) as

$$\begin{aligned} B_{k+1} - O = \ & B_k - O + \frac{Js_k(y_k - B_ks_k + Os_k - Os_k)^T J}{(s_k^T s_k)/\beta_k} + \frac{(y_k - B_ks_k + Os_k - Os_k)s_k^T}{(s_k^T s_k)/\beta_k} - \\ & \frac{(Js_k)^T(y_k - B_ks_k + Os_k - Os_k)Js_ks_k^T}{\left((s_k^T s_k)/\beta_k\right)^2} \ . \end{aligned}$$

The rest of the proof follows the same steps of Lemma 3.4 by replacing $(s_k^T s_k)/\beta_k$ with $s_k^T s_k$. $\square$

Now we are ready to prove Proposition 4.7 and Proposition 4.8.

Proof of Proposition 4.7. First, notice that $\|y_k\| = \|F(z_k + s_k) - F(z_k)\| \leq \gamma_1 \|s_k\|$, where we utilize the fact that F is γ_1-Lipschitz continuous in $S(R_0)$, $z_k \in S$ and $\|s_k\| \leq R_0$. By expanding B_{j+1} using (52) for $j = k+n+m$ we obtain

$$\begin{aligned} \|B_{j+1}\| &= \left\| JQ_jJB_jQ_j + \beta_j \frac{y_j s_j^T}{s_j^T s_j} + \beta_j \frac{Js_j y_j^T J}{s_j^T s_j} Q_j \right\| \\ &\leq \|JQ_jJ\|\|B_jQ_j\| + \beta_j \frac{\|y_j\|}{\|s_j\|} + \beta_j \frac{\|Jy_j\|}{\|s_j\|} \leq \|B_jQ_j\| + 4\gamma_1 \ , \end{aligned}$$

where the first inequality uses Cauchy-Schwarz followed by $\|Js_j\| = \|s_j\|$ and $\|Q_j\| \leq 1$, the second inequality uses $\|J\| = 1$, $\|Q_j\| \leq 1$, $\beta_j < 2$, $\|Jy_j\| = \|y_j\|$, and $\|y_j\| \leq \gamma_1 \|s_j\|$. Expanding B_j in the R.H.S. of the inequality $\|B_{j+1}\| \leq \|B_jQ_j\| + 4\gamma_1$ recursively for $n+m-1$ times, using (52) and in the same way as we did for B_{j+1}, we obtain:

$$\|B_{k+n+m+1}\| \leq \|B_{k+1}Q_{k+1} \dots Q_{k+n+m-1}Q_{k+n+m}\| + 4(n+m)\gamma_1 \ .$$

It follows from Lemma 4.10 that there exists a constant $\theta \in (0,1)$ and index K such that $\|\prod_{j=k+1}^{k+n+m} Q_j\| \leq \theta$ for any $k \geq K$, thus

$$\|B_{k+n+m+1}\| \leq \theta \|B_{k+1}\| + 4(n+m)\gamma_1 \ .$$

We now apply Lemma A.7 in the appendix and conclude since $4(n+m)\gamma_1$ is upper-bounded, then so is $\{\|B_k\|\}$ for $k \geq K+n+m+1$. Thus there exists ν_2 such that $\{\|B_k\|\} \leq \nu_2$ for all k. $\square$

Proof of Proposition 4.8. We prove the lemma by the following two steps:
(a) $m_k(p_k^C) \geq m_k(s_k)$,
(b) $m_k(0) - m_k(p_k^C) \geq \frac{\|g_k\|}{2} \min\{\Delta_k \ , \ \frac{\|g_k\|}{\nu_2^2}\}$.
To show part (a), it follows from (45) that the step s_k takes values from $\{p_k^B, p_k^C, p_k^D\}$. Notice that p_k^B is the global minimizer of $m_k(s)$, thus $m_k(p_k^C) \geq m_k(p_k^B)$. Then we just need to show that $m_k(p_k^D) \leq m_k(p_k^C)$ under the condition that $\|p_k^B\| > \Delta_k$ and $\|p_k^C\| < \Delta_k$, in which case we have $p_k^C = -\|g_k\|^2/({g_k}^T B_k^T B_k g_k) g_k$. Recall that $p_k^D = p_k^C + \alpha(p_k^B - p_k^C)$, thus we just need to show that $h(\alpha) := m_k\big(p_k^C + \alpha(p_k^B - p_k^C)\big)$ is monotonically nonincreasing in α, by noticing $h(0) = m_k(p_k^C)$. This is because $h(\alpha)$ is differentiable with derivative

$$
\begin{aligned}
h'(\alpha) &= g_k^T(p_k^B - p_k^C) + (p_k^B - p_k^C)^T B_k^T B_k p_k^C + \alpha(p_k^B - p_k^C)^T B_k^T B_k (p_k^B - p_k^C) \\
&= (p_k^B - p_k^C)^T \Big(g_k + B_k^T B_k p_k^B - B_k^T B_k p_k^B + B_k^T B_k p_k^C + \alpha B_k^T B_k (p_k^B - p_k^C)\Big) \\
&= (p_k^B - p_k^C)^T \Big(g_k + B_k^T B_k p_k^B - (1-\alpha) B_k^T B_k (p_k^B - p_k^C)\Big) \\
&= (p_k^B - p_k^C)^T \Big(0 - (1-\alpha) B_k^T B_k (p_k^B - p_k^C)\Big) \\
&\leq 0 \ ,
\end{aligned}
$$

where the first equality comes from substituting $s_k = p_k^C + \alpha(p_k^B - p_k^C)$ into the definition of m_k in (41), the second equality and the third equality come from rearrangement, the fourth equality is a result of the definition of $p_k^B = -B_k^{-1}(B_k^{-1})^T g_k$, and finally the inequality comes from noticing that $B_k^T B_k$ is positive definite and $0 < \alpha < 1$. This shows part (a).

To show part (b), recall the formulation of Cauchy point p_k^C (43). If $p_k^C = -\Delta_k g_k / \|g_k\|$, then it must hold that

$$
\Delta_k / \|g_k\| \leq \|g_k\|^2 / ({g_k}^T B_k^T B_k g_k) \ . \tag{53}
$$

Therefore,

$$
\begin{aligned}
m_k(0) - m_k(p_k^C) &= \Delta_k \|g_k\| - \frac{1}{2} \frac{\Delta_k^2}{\|g_k\|^2} g_k^T B_k^T B_k g_k \\
&= \Delta_k \|g_k\| - \frac{1}{2} \Delta_k \|g_k\| \frac{\Delta_k g_k^T B_k^T B_k g_k}{\|g_k\|^3} \\
&\geq \frac{1}{2} \|g_k\| \Delta_k \ ,
\end{aligned}
$$

where the inequality is from (53).

Otherwise, we have $p_k^C = -\|g_k\|^2/({g_k}^T B_k^T B_k g_k) g_k$, thus

$$
\begin{aligned}
m_k(0) - m_k(p_k^C) &= \frac{\|g_k\|^2}{{g_k}^T B_k^T B_k g_k} \|g_k\|^2 - \frac{1}{2} \Big(\frac{\|g_k\|^2}{{g_k}^T B_k^T B_k g_k}\Big)^2 g_k^T B_k^T B_k g_k \\
&= \frac{\|g_k\|^4}{2 {g_k}^T B_k^T B_k g_k} \geq \frac{\|g_k\|^2}{2\|B_k\|^2} \geq \frac{\|g_k\|^2}{2\nu_2^2} \ ,
\end{aligned}
$$

where the last inequality uses Proposition 4.7.

Putting the last two inequalities together we conclude the claim in part (b). □

Next, we present the proof of Theorem 4.2. The proof of Theorem 4.2 is inspired by [38, Theorems 4.5-4.6].

Proof of Theorem 4.2. We prove the theorem in two steps:
(a). First we show $\liminf_{k \to \infty} \|g_k\| = 0$.

(b). Next we prove $\lim_{k\to\infty}\|g_k\| = 0$.
We first prove **(a).** by contradiction. Suppose that there exists $\epsilon > 0$ and a positive index K such that $\|g_k\| \geq \epsilon$ for all $k \geq K$. It follows from the mean value theorem on the one-dimensional function $f(t) = \|F(z_k + ts_k)\|^2/2$ that

$$\|F(z_k+s_k)\|^2/2 = \|F(z_k)\|^2/2 + F(z_k+ts_k)^T\nabla F(z_k+ts_k)s_k,$$

for some $0 < t < 1$. By the definition of $m_k(s_k)$, we obtain

$$\begin{aligned}\left|m_k(s_k) - \|F(z_k+s_k)\|^2/2\right| &= \left|F(z_k)^T\nabla F(z_k)s_k + (s_k^T B_k^T B_k s_k)/2 - F(z_k+ts_k)^T\nabla F(z_k+ts_k)s_k\right| \\ &\leq (\mu+\nu_2^2/2)\|s_k\|^2 \leq (\mu+\nu_2^2/2)\Delta_k^2 \ ,\end{aligned} \tag{54}$$

where the first inequality comes from Fact 4.6 and Proposition 4.7, and the second inequality uses $s_k \leq \Delta_k$. Thus, it holds for any $k \geq K$ that

$$|\rho_k - 1| = \frac{\left|m_k(s_k) - \|F(z_k+s_k)\|^2/2\right|}{|m_k(0) - m_k(s_k)|} \leq \frac{(\mu+\nu_2^2/2)\Delta_k^2}{\frac{\epsilon}{2}\min\left\{\Delta_k \ , \ \frac{\epsilon}{\nu_2^2}\right\}} \ , \tag{55}$$

where the inequality comes from the Proposition 4.8 and (54) by noticing $\|g_k\| \geq \epsilon$.

Define $\bar\Delta := \min\left\{\frac{\epsilon/2}{\mu+\nu_2^2/2}, R_0\right\}$ and then $\bar\Delta < \frac{\epsilon}{\nu_2^2}$. We here first show by induction that the trust-region radius is lower bounded:

$$\Delta_k \geq \min\left\{\Delta_K \ , \ \frac{\bar\Delta}{4}\right\} \text{ for any } k \geq K \ . \tag{56}$$

Apparently, (56) holds for $k = K$. Now suppose (56) holds for k. If $\Delta_k < \bar\Delta/2$, (so then $\Delta_k = \min\left\{\Delta_k \ , \ \frac{\epsilon}{\nu_2^2}\right\}$), it follows from (55) that

$$|\rho_k - 1| \leq \frac{(\mu+\nu_2^2/2)\Delta_k^2}{\frac{\epsilon}{2}\Delta_k} \leq \frac{\Delta_k}{\bar\Delta} < \frac{1}{2} \ .$$

Therefore, we have $\rho_k > 1/2$, and as a result, the trust-region increases in the next iteration: $\Delta_{k+1} = \min\{2\Delta_k, R_0\} \geq \Delta_k$, thus (56) holds for $k+1$ by induction. Otherwise, we have $\Delta_k \geq \bar\Delta/2$, and it follows from the fact that the trust-region in one iteration can only contract by a factor of 2 that $\Delta_{k+1} \geq \bar\Delta/4$, thus (56) holds for $k+1$. Combining the above two cases, we prove (56) by induction.

Next, if we have an infinite increasing subsequence $\{k^i\} \subseteq \{K, K+1, K+2...\}$ such that $\rho_{k^i} > 1/2$, then we deduce from (46) that

$$\|F(z_{k^i})\|^2/2 - \|F(z_{k^i+1})\|^2/2 \geq \rho_{k^i}\left(m_{k^i}(0) - m_{k^i}(s_{k^i})\right) \geq \frac{\epsilon}{4}\min\left\{\Delta_{k^i} \ , \ \frac{\epsilon}{\nu_2^2}\right\} \geq \frac{\epsilon}{4}\min\left\{\Delta_K \ , \ \bar\Delta/4 \ , \ \frac{\epsilon}{\nu_2^2}\right\} \ ,$$

where the second inequality uses Proposition 4.8 and $\|g_{k^i}\| \geq \epsilon$, and the last inequality is from (56). Therefore, noticing $\|F(z_k)\|^2/2$ is monotonically nonincreasing, and summing up the above inequality, we have

$$\|F(z_K)\|^2/2 - \|F(z_{k^i+1})\|^2/2 \geq \sum_{j=1}^{i}\|F(z_{k^j})\|^2/2 - \|F(z_{k^j+1})\|^2/2 \geq \frac{i\epsilon}{4}\min\left\{\Delta_K \ , \ \bar\Delta/4 \ , \ \frac{\epsilon}{\nu_2^2}\right\}.$$

This cannot happen for a large enough i since $\|F(z_{k^i+1})\|^2/2 \geq 0$ is lower bounded.

Otherwise, if there is no such infinite subsequence $\{k^i\}$, then there exists $K' \geq K$ such that $\rho_k \leq 1/2$ for all $k \geq K'$. As a result, the trust-region radius contracts at each iteration after K', thus $\lim_{k\to\infty}\Delta_k = 0$, which contradicts with (56). Combining the above two cases, we conclude that our original assumption cannot hold and therefore $\liminf_{k\to\infty}\|g_k\| = 0$.

Now we turn to **(b).** We first present the high-level ideas of the proof. We will show **(b).** by contradiction. Suppose **(b).** does not hold, namely, there exists $\epsilon > 0$ and an infinite increasing subsequence $\{t_i\}_{i=1}^{\infty}$ of $\{1, 2, ...\}$ such that $\|g_{t_i}\| \geq \epsilon$. Then we will show that there exists a constant $C > 0$ and an increasing subsequence $\{u_i\}_{i=1}^{\infty}$ of $\{t_i\}_{i=1}^{\infty}$ such that $\|F(z_{u_i})\|^2/2 - \|F(z_{u_{i+1}})\|^2/2 \geq C$. Thus by the monotonicity of $\|F(z_k)\|^2/2$, we have $\|F(z_{u_i})\|^2/2 \to -\infty$ as $i \to \infty$, which contradicts with the fact that $\|F(z_{u_i})\|^2/2 \geq 0$. In the rest of this proof we construct the sequence $\{u_i\}_{i=1}^{\infty}$ by induction.

For initialization, we set $u_1 = t_1$. Next, for a given u_i, we show how to build u_{i+1}. Consider the point z_{u_i} and a close ball $\mathcal{B}(z_{u_i}, R) = \{z \mid \|z - z_{u_i}\| \leq R\}$ with center z_{u_i} and radius R, where

$$R := \min\{\epsilon/(2\mu) \ , \ R_0\} \ .$$

Notice that for any $z_k \in \mathcal{B}(z_{u_i}, R)$, we have from (49) that

$$\|g_k - g_{u_i}\| \leq \ \mu\|z_k - z_{u_i}\| \leq \mu R \leq \epsilon/2 \ ,$$

and thus

$$\|g_k\| \geq \|g_{u_i}\| - \|g_{u_i} - g_k\| \geq \epsilon - \epsilon/2 = \epsilon/2 \ .$$

Recall from **(a).** that there is a subsequence of $\{\|g_k\|\}_{k=0}^{\infty}$ that converges to 0, thus there exists at least one solution in the sequence $\{z_k\}_{k \geq u_i}$ that leaves $\mathcal{B}(z_{u_i}, R)$. Let z_{l+1} be the first of such iterates. Then $\|z_k - z_{u_i}\| \leq R$ for $k = u_i + 1, ..., l$, and it holds that

$$\begin{aligned}
\|F(z_{u_i})\|^2/2 - \|F(z_{l+1})\|^2/2 &= \sum_{k=u_i}^{l} \|F(z_k)\|^2/2 - \|F(z_{k+1})\|^2/2 \\
&= \sum_{k=u_i,...,l, z_k \neq z_{k+1}} \|F(z_k)\|^2/2 - \|F(z_{k+1})\|^2/2 \\
&\geq \zeta \sum_{k=u_i,...,l, z_k \neq z_{k+1}} m_k(0) - m_k(s_k) \\
&\geq \frac{\zeta}{2} \sum_{k=u_i,...,l, z_k \neq z_{k+1}} \|g_k\| \min\left\{\Delta_k \ , \ \frac{\|g_k\|}{\nu_2^2}\right\} \\
&\geq \frac{\zeta\epsilon}{4} \sum_{k=u_i,...,l, z_k \neq z_{k+1}} \min\left\{\Delta_k \ , \ \frac{\epsilon}{2\nu_2^2}\right\} \ ,
\end{aligned} \tag{57}$$

where the second equality considers only the valid steps (namely ignores the null steps), the first inequality utilizes the criteria of a valid step, the second inequality uses Proposition 4.8, and the third inequality is implied from $\|g_k\| \geq \epsilon/2$ since $z_k \in \mathcal{B}(z_{u_i}, R)$.

Next we present a lower bound for (57). There are only two possibilities:

(i). Suppose there exists k in the summation in the R.H.S. of (57) that $\Delta_k > \epsilon/(2\nu_2^2)$, then we obtain by noticing $\|F(z_k)\|^2/2$ is monotonically nonincreasing that

$$\|F(z_{u_i})\|^2/2 - \|F(z_{l+1})\|^2/2 \geq \frac{\zeta\epsilon}{4}\frac{\epsilon}{2\nu_2^2} \ .$$

(ii). Otherwise, $\Delta_k \leq \epsilon/(2\nu_2^2)$ for all of k in the summation in the R.H.S. of (57). Notice that z_{l+1} is outside $\mathcal{B}(z_{u_i}, R)$, thus

$$R \leq \|z_{l+1} - z_{u_i}\| \leq \sum_{k=u_i}^{l} \|z_{k+1} - z_k\| = \sum_{k=u_i,...,l, z_k \neq z_{k+1}} \|z_{k+1} - z_k\| \leq \sum_{k=u_i,...,l, z_k \neq z_{k+1}} \Delta_k \ ,$$

where the last inequality uses $\|z_{k+1}-z_k\| = \|s_k\| \le \Delta_k$ for a valid step. Therefore, it holds from (57) that

$$\|F(z_{u_i})\|^2/2 - \|F(z_{l+1})\|^2/2 \ge \frac{\zeta\epsilon}{4} \sum_{k=u_i,\ldots,l, z_k \ne z_{k+1}} \Delta_k \ge \frac{\zeta\epsilon}{4} R = \frac{\zeta\epsilon}{4} \min\{\epsilon/(2\mu)\ ,\ R_0\}\ .$$

Combining (i) and (ii), we arrive at

$$\|F(z_{u_i})\|^2/2 - \|F(z_{l+1})\|^2/2 \ge \frac{\zeta\epsilon}{4} \min\left\{\epsilon/(2\mu)\ ,\ R_0\ ,\ \epsilon/(2\nu_2^2)\right\}\ .$$

Now choose u_{i+1} to be the first index in the infinite sequence $\{t_i\}_i^\infty$ such that $u_{i+1} \ge l+1$, then such u_{i+1} exists, because $\{t_i\}_i^\infty$ has infinite values and

$$\|F(z_{u_i})\|^2/2 - \|F(z_{u_{i+1}})\|^2/2 \ge \|F(z_{u_i})\|^2/2 - \|F(z_{l+1})\|^2/2 \ge \frac{\zeta\epsilon}{4} \min\left\{\epsilon/(2\mu)\ ,\ R_0\ ,\ \epsilon/(2\nu_2^2)\right\}\ ,$$

where the first inequality uses the monotonicity of $\|F(z_k)\|^2/2$. As a result, let $C = (\zeta\epsilon/4)\min\left\{\epsilon/(2\mu), R_0, \epsilon/(2\nu_2^2)\right\} > 0$, then we have that $\|F(z_{u_i})\|^2/2 \le \|F(z_{u_1})\|^2/2 - (i-1)C \to -\infty$ when $i \to \infty$, which contradicts with the fact that $\|F(z_{u_i})\|^2/2 \ge 0$. This finishes the proof by contradiction. □

Finally, we prove Theorem 4.3:

Proof of Theorem 4.3 part (1). Define $\delta_k := \sup_{0\le t\le 1} \|\nabla F(z_k + ts_k) - \nabla F(z^*)\|$. Then it follows from $z_k \to z^*$, $s_k \to 0$ and the continuity of $\nabla F(z)$ that $\delta_k \to 0$. Let $T(z) = F(z) - \nabla F(z^*)z$, then $\nabla T(z) = \nabla F(z) - \nabla F(z^*)$. Thus, it holds that

$$\begin{aligned}\|y_k - \nabla F(z^*)s_k\| &= \|F(z_k+s_k) - F(z_k) - \nabla F(z^*)s_k\| = \|T(z_k+s_k) - T(z_k)\| \\ &\le \sup_{0\le t\le 1} \|\nabla T\big(z_k + ts_k\big)\|\|s_k\| = \delta_k\|s_k\|\ .\end{aligned} \tag{58}$$

The rest of the proof is very similar to the proof of Proposition 4.7. Applying Lemma 4.11 with $O = \nabla F(z^*)$ we expand M_{j+1} using (51) for $j = k+n+m$ in the following way:

$$\begin{aligned}\|M_{j+1}\| &= \left\| JQ_jJM_jQ_j + \beta_j \frac{\big(y_j - \nabla F(z^*)s_j\big)s_j^T}{s_j^Ts_j} + \beta_j \frac{Js_j\big(y_j - \nabla F(z^*)s_j\big)^TJ}{s_j^Ts_j} Q_j \right\| \\ &\le \|JQ_jJ\|\|M_jQ_j\| + 2\beta_j\delta_j \le \|M_jQ_j\| + 4\delta_j\ ,\end{aligned}$$

where the first inequality utilizes $\|Q_j\| \le 1$, (58) and the fact $\|Jq\| = \|q\|$ for any vector q of the appropriate size, the second inequality comes from $\|J\| = 1$, $\|Q_j\| \le 1$ and $\beta_j < 2$. Expanding M_j recursively for $n+m-1$ times in the R.H.S. of the inequality $\|M_{j+1}\| \le \|M_jQ_j\| + 4\delta_j$, using (51) and in the same way as we did for M_{j+1} we obtain:

$$\|M_{k+n+m+1}\| \le \|M_{k+1}Q_{k+1}\ldots Q_{k+n+m-1}Q_{k+n+m}\| + 4\sum_{j=k+1}^{k+m+n} \delta_j\ .$$

It follows from Lemma 4.10 that there exists a constant $\theta \in (0,1)$ and index K such that $\|\prod_{j=k+1}^{k+n+m} Q_j\| \le \theta$ for any $k \ge K$, thus

$$\|M_{k+n+m+1}\| \le \theta\|M_{k+1}\| + 4\sum_{j=k+1}^{k+m+n} \delta_j\ .$$

We now apply Lemma A.7 in the appendix and conclude since $4\sum_{j=k+1}^{k+m+n}\delta_j \to 0$, then $\|M_k\| \to 0$ as $k \to \infty$. Recalling $M_k = B_k - \nabla F(z^*)$, the proof is complete. $\square$

To prove part (2) of Theorem 4.3, we first present two lemmas.

Lemma 4.12. *Under the assumptions stated in Theorem 4.3 part (1), it holds that $\|B_k^{-1}\|$ is upper-bounded, that is, there exists a positive value ν_1 such that*

$$\|B_k^{-1}\| \le \nu_1 \ .$$

Proof. Recall from the construction that B_k is invertible and from Assumption 4.1 that $\nabla F(z^*)$ is nonsingular with $\|\nabla F^{-1}(z^*)\| \le \gamma$. Notice that Theorem 4.3 part (1) shows that $\|B_k - \nabla F(z^*)\| \to 0$, whereby there exists K such that $\|B_k - \nabla F(z^*)\| \le 1/(2\gamma)$ for any $k \ge K$. We can now apply Banach Perturbation Lemma (Lemma A.1 in the appendix) to the matrices $\nabla F(z^*)$ and B_k (note $\gamma/(2\gamma) = 0.5 < 1$) and obtain:

$$\|B_k^{-1}\| \le \frac{\gamma}{1-0.5} \le 2\gamma \ ,$$

and as a result it holds for all k:

$$\|B_k^{-1}\| \le \max\left\{\|B_0^{-1}\|, \|B_1^{-1}\|, ..., \|B_{K-1}^{-1}\|, 2\gamma\right\} := \nu_1 \ .$$

$\square$

The next Lemma provides a lower bound on the amount of predicted decrease:

Lemma 4.13. *Under the assumptions stated in Theorem 4.3 part (1), it holds for any k that*

$$m_k(0) - m_k(s_k) \ge \frac{\|s_k^2\|}{2} \min\left\{\frac{1}{\nu_1^2} \ , \ \frac{1}{\nu_1^4\nu_2^2}\right\} \ . \tag{59}$$

Proof. It follows from (42) and Lemma 4.12 that $\|p_k^B\| \le \nu_1^2\|g_k\|$. Furthermore, recall that in Algorithm 3 we set $s_k = p_k^B$ if $\|p_k^B\| \le \Delta_k$ and otherwise we have $\|s_k\| = \Delta_k$, thus in either case we have $\|s_k\| \le \|p_k^B\|$. Hence, it holds that $\|g_k\|/\|s_k\| \ge 1/\nu_1^2$. Then, it follows from Proposition 4.8 that

$$\begin{aligned}
m_k(0) - m_k(s_k) &\ge \frac{\|g_k\|}{2} \min\left\{\Delta_k \ , \ \frac{\|g_k\|}{\nu_2^2}\right\} \\
&= \|s_k\|\frac{\|g_k\|}{2} \min\left\{\frac{\Delta_k}{\|s_k\|} \ , \ \frac{\|g_k\|}{\|s_k\|\nu_2^2}\right\} \\
&\ge \|s_k\|\frac{\|g_k\|}{2} \min\left\{1 \ , \ \frac{1}{\nu_1^2\nu_2^2}\right\} \\
&\ge \frac{\|s_k\|^2}{2} \min\left\{\frac{1}{\nu_1^2} \ , \ \frac{1}{\nu_1^4\nu_2^2}\right\} ,
\end{aligned}$$

where the second inequality is from $\Delta_k/\|s_k\| \ge 1$ and $\|g_k\|/\|s_k\| \ge 1/\nu_1^2$, and the third inequality uses $\|g_k\|/\|s_k\| \ge 1/\nu_1^2$ again. $\square$

Proof of Theorem 4.3 part (2). We give proofs for the following two claims:
(a). There exists a constant K such that it holds for all $k \ge K$ that $\|p_k^B\| \le \Delta_k$, so $s_k = p_k^B$ and $z_{k+1} = z_k + s_k$. This shows that after a finite number of steps, we always take quasi-Newton step, and the quasi-Newton step is a valid step.

(b). The rate of the convergence of Algorithm 3 is R-superlinear, that is, $\lim_{k\to\infty} \|z_k - z^*\|^{1/k} = 0$.
To see **(a).** we begin by defining

$$\eta_k = \sup_{0\le t\le 1} \| \nabla F(z_k + ts_k) - B_k\| \ .$$

Since $z_k \to z^*$, $s_k \to 0$ and as we showed in part (1), $\|B_k - \nabla F(z^*)\| \to 0$, as $k \to \infty$, and since $\nabla F(z)$ is continuous around z^*, we conclude $\eta_k \to 0$.

Consider now the one-dimensional function $f(t) = \|F(z_k + ts_k)\|^2/2$, and then by second-order mean value theorem we know there exists $0 < t < 1$ such that

$$\begin{aligned}\|F(z_k + s_k)\|^2/2 =&\|F(z_k)\|^2/2 + F(z_k)^T\nabla F(z_k)s_k+ \\ & s_k^T \left(\nabla F(z_k + ts_k)^T\nabla F(z_k + ts_k)\right) s_k/2 + \nabla^2 F(z_k + ts_k) \left(F(z_k + ts_k)/2, s_k, s_k\right) \ ,\end{aligned} \tag{60}$$

where $\nabla^2 F(z_k+ts_k)$ is a 3-dimensional tensor and $F(z_k+ts_k)\left(F(z_k + ts_k), s_k, s_k\right)$ refers to the tensor-vector product. Furthermore, notice that $F(z)$ is γ_1-Lipschitz and $\nabla F(z)$ is γ_2-Lipschitz, thus

$$\nabla^2 F(z_k + ts_k)\Big(F(z_k + ts_k), s_k, s_k\Big)/2 \le \gamma_2\|F(z_k + ts_k)\|\|s_k\|^2/2 \le \gamma_2\gamma_1 \left(\|z_k - z_*\| + \|s_k\|\right) \|s_k\|^2/2 \ .$$

Substituting this inequality to (60) and recalling $g_k^T = F(z_k)^T\nabla F(z_k)$ we obtain

$$\|F(z_k+s_k)\|^2/2 \le \|F(z_k)\|^2/2+g_k^T s_k+s_k^T\nabla F(z_k+ts_k)^T\nabla F(z_k+ts_k)s_k/2+\gamma_2\gamma_1 \left(\|z_k - z_*\| + \|s_k\|\right) \|s_k\|^2/2.$$

Denote by $A_k = \nabla F(z_k + ts_k) - B_k$ and clearly $\|A_k\| \le \eta_k$. It holds by recalling the definition of $m_k(s_k)$ that

$$\begin{aligned}&\|F(z_k + s_k)\|^2/2 - m_k(s_k)\\ \le&s_k^T\nabla F(z_k + ts_k)^T\nabla F(z_k + ts_k)s_k/2 + \gamma_2\gamma_1 \left(\|z_k - z_*\| + \|s_k\|\right) \|s_k\|^2/2 - s_k^T B_k^T B_k s_k/2\\ =&s_k^T\Big((B_k + A_k)^T(B_k + A_k)\Big)s_k/2 + \gamma_2\gamma_1 \left(\|z_k - z_*\| + \|s_k\|\right) \|s_k\|^2/2 - s_k^T B_k^T B_k s_k/2\\ =&s_k^T\Big(B_k^T A_k + A_k^T B_k + A_k^T A_k\Big)s_k/2 + \gamma_2\gamma_1 \left(\|z_k - z_*\| + \|s_k\|\right) \|s_k\|^2/2\\ \le&\Big(2\eta_k\nu_2 + \eta_k^2 + \gamma_2\gamma_1 \left(\|z_k - z_*\| + \|s_k\|\right)\Big)\|s_k\|^2/2 \ ,\end{aligned}$$

where the final inequality uses $\|B_k\| \le \nu_2$ and $\|A_k\| \le \eta_k$. Recalling the definition of ρ_k and combining (59) and this inequality we arrive at

$$|1 - \rho_k| = \frac{|\|F(z_k + s_k)\|^2/2 - m_k(s_k)|}{m_k(0) - m_k(s_k)} \le \frac{2\eta_k\nu_2 + \eta_k^2 + \gamma_2\gamma_1 \left(\|z_k - z_*\| + \|s_k\|\right)}{\min\left\{\frac{1}{\nu_1^2} \ , \ \frac{1}{\nu_1^4\nu_2^2}\right\}} \ .$$

We have $z_k \to z^*$, $s_k \to 0$ and $\eta_k \to 0$ as $k \to \infty$. Thus, in the R.H.S. of the above inequality the numerator goes to 0 and the denominator is a constant. Therefore, there exists K_1 such that $|1 - \rho_k| \le 0.5$, thus $\rho_k > 0.5$ for all $k \ge K_1$. This means that for $k \ge K_1$, we always expand the trust-region radius: $\Delta_{k+1} = \min\{2\Delta_k, R_0\}$. As such, there exists K_2 such that $\Delta_k = R_0$ for all $k \ge K_2$. Furthermore, it follows from Lemma 4.12 that $\|p_k^B\| \le \nu_1^2\|g_k\|$, thus $\|p_k^B\| \to 0$ as $k \to \infty$, whereby there exists K_3 such that $\|p_k^B\| \le R_0$ for all $k \ge K_3$. Let $K = \max\{K_2, K_3\}$, then we have $\|p_k^B\| \le R_0 = \Delta_k$ for all $k \ge K$. This finishes the proof of **(a).** and we conclude eventually all steps are valid and they are quasi-Newton steps.

To see **(b).** notice that R-superlinear convergence studies the eventual behavior of the algorithm as the iteration count $k \to \infty$. It follows from part **(a).** that eventually (i.e., when $k \ge K$) we always take quasi-Newton step, i.e., $s_k = p_k^B$ and the step is a valid step, i.e., $z_{k+1} = z_k + s_k = z_k + p_k^B$. It then follows from the definition of p_k^B and g_k that $s_k = p_k^B = -B_k^{-1}B_k^{-T}\nabla F(z_k)^T F(z_k)$. Reusing the notation, denote by $A_k = \nabla F(z_k) - B_k$ and clearly $\|A_k\| \le \eta_k$, further let $N_k = B_{k+1} - B_k$, so, $\|N_k\| \to 0$ as $k \to \infty$. We have

$$
\begin{aligned}
\|F(z_{k+1})\| &= \|F(z_k) + B_{k+1}s_k\| \\
&= \|F(z_k) - B_{k+1}B_k^{-1}B_k^{-T}\nabla F(z_k)^T F(z_k)\| \\
&\le \|I - B_{k+1}B_k^{-1}B_k^{-T}\nabla F(z_k)^T\|\|F(z_k)\| \\
&= \left\| I - (B_k + N_k)B_k^{-1}\Big((B_k + A_k)B_k^{-1}\Big)^T \right\| \|F(z_k)\| \\
&= \|I - (I + N_k B_k^{-1})(I + A_k B_k^{-1})^T\|\|F(z_k)\| \\
&\le \left(\|B_k^{-1}\|\|N_k\| + \|B_k^{-1}\|\|A_k\| + \|N_k\|\|A_k\|\|B_k^{-1}\|^2\right)\|F(z_k)\| \\
&\le \nu_1\left(\|N_k\| + \eta_k + \|N_k\|\eta_k\nu_1\right)\|F(z_k)\| \ ,
\end{aligned}
$$

where the first equality comes from the secant condition (6), the first and the second inequalities utilize Cauchy-Schwarz inequality, and the last inequality uses $\|B_k^{-1}\| \le \nu_1$ and $\|A_k\| \le \eta_k$. Since both $\eta_k \to 0$ and $\|N_k\| \to 0$ as $k \to \infty$, then, for any arbitrary $0 < \epsilon < 1$, there exists iteration $\bar{K} \ge K$ such that $\nu_1(\|N_k\| + \eta_k + \|N_k\|\eta_k\nu_1) \le \epsilon$ for all $k \ge \bar{K}$. Therefore, it holds for $k \ge \bar{K}$ that $\|F(z_k)\| \le \epsilon^{k-\bar{K}}\|F(z_{\bar{K}})\|$, and

$$
\lim_{k\to\infty}\|F(z_k)\|^{1/k} \le \lim_{k\to\infty}\left(\epsilon^{k-\bar{K}}\|F(z_{\bar{K}})\|\right)^{1/k} = \lim_{k\to\infty}\epsilon\Big(\frac{\|F(z_{\bar{K}})\|}{\epsilon^{\bar{K}}}\Big)^{1/k} = \epsilon \ .
$$

Notice that the above inequality holds for any $0 < \epsilon < 1$. Together with $\|F(z_k)\| \ge 0$, we conclude that $\lim_{k\to\infty}\|F(z_k)\|^{1/k} = 0$. Furthermore, recall that $F(z^*) = 0$, $\nabla F(z^*)$ is nonsingular and $\nabla F(z)$ is continuous around z^*, thus we obtain $\lim_{k\to\infty}\|z_k - z^*\|^{1/k} = 0$, which finishes the proof. □

5 Numerical Experiments.

In this section we present numerical experiments of J-symmetric quasi-Newton algorithms and compare it with classical algorithms for minimax problems. We present three sets of experiments: quadratic convex-concave minimax problems, analytic center, and a nonconvex-nonconcave minimax problem. The source code is available from `https://github.com/azamasl/Jsymm`.

5.1 Quadratic Minimax Problem.

We consider quadratic convex-concave minimax problems of the form

$$
L(x, w) = \frac{1}{2}(x - x^*)^T D(x - x^*) + (w - w^*)^T A(x - x^*) - \frac{1}{2}(w - w^*)^T C(w - w^*) \ , \tag{61}
$$

where C and D are positive semidefinite matrices. Quadratic functions comprise an important class of problems since any function around its optimal solution (x^*, y^*) behaves similar to a quadratic minimax problem (61), as such, the following experiment shows the asymptotic behavior of different algorithms for general minimax problems.

In this experiment, we generate synthetic data with $D \in \mathbb{R}^{500\times500}$, $C \in \mathbb{R}^{500\times500}$ and $A \in \mathbb{R}^{500\times500}$ in the following way. Entries of A are drawn randomly from a normal distribution $\mathcal{N}(0, 1/\sqrt{500})$. To generate a random positive definite matrix D, we first create a random matrix $S \in \mathbb{R}^{500\times500}$ with entries drawn from

$\mathcal{N}(0, 1/\sqrt{500})$, and then we symmetrise the matrix by $S = (S + S^T)/2$. Next, we shift the matrix using a scaled identity matrix to make it positive definite $S = S + (|\lambda_{\min}| + 1)I$, where $\lambda_{\min}$ is the minimal eigenvalue of S (which is usually negative), and the identity matrix guarantees the 1-strong-convexity-strong-concavity of the matrix S. Finally, we set $D = \alpha S$, with α taking values from $\{0, 10^{-4}, 10^{-2}, 1\}$. We set the matrix C by the same procedure and with a different random seed. The value α measures the scale ratio between the diagonal terms and the off-diagonal terms, and it turns out to be the critical parameter to characterize the performance of different algorithms. When $\alpha = 0$, the problem is a bilinear convex-concave minimax problem. When $\alpha > 0$, the problem is strongly-convex-strongly-concave.

We compare the behaviors of the following five algorithms:

- EGM: The extra-gradient algorithm [30, 37] with stepsize $1/||\nabla F(z)||$;
- Broyden: Broyden's good method [9, 10] with a fixed stepsize 0.01 first and stepsize 1 after $\|F(z)\| \leq 0.1$;
- J-symm: J-symmetric quasi-Newton Algorithm (Algorithm 1) with a fixed stepsize 0.01 first and stepsize 1 after $\|F(z)\| \leq 0.1$;
- J-symm-LS: J-symmetric quasi-Newton Algorithm with line-search (Algorithm 2);
- J-symm-Tr: J-symmetric quasi-Newton Algorithm with trust-region (Algorithm 3).

We initialize the Jacobian estimation $H_0 = I$ for J-symm, J-symm-LS, and J-symm-Tr. Notice that $H_0 = I$ would make the first step of Broyden's method go to infinity, thus we initialize the Jacobian estimation H_0 for Broyden's method as a diagonal matrix with each entry coming from uniform distribution $U(0, 1)$. For Broyden's method and J-symm, we start with stepsize 0.01 first, and then 1 after $\|F(z)\| \leq 0.1$. The reason is because a large initial stepsize (i.e., 1) for both methods quickly blows up the solutions due to the bad initial estimation of the Jacobian.

Figure 1 plots $\|F(z)\|$ in log scale versus the number of iterations for the five algorithms and different α values $\alpha \in \{0, 10^{-4}, 10^{-2}, 1\}$. A few observations in sequence: Firstly, as the value of α increases, all five methods have better performance. Secondly, J-symm (black line) always enjoys superlinear convergence for all four settings, which verifies the result in theorem 3.2. In contrast, Broyden's methods do not converge with small α, which is consistent with the instability of Broyden's method. Thirdly, J-symm-LS (magenta line) has the best performance in this set of experiments due to its adaptive stepsize choice. Finally, J-symm-Tr (blue line) turns out to be too conservative, in particular for the case with small α value.

5.2 Analytical Center of Polytope.

Analytic center is one way to define the geometric center of a polytope, and it has important applications in barrier methods. Consider a polytope given by linear inequalities:

$$a_i^T x \leq b_i \ , \qquad i = 1, \dots m \ ,$$

where $x \in \mathbb{R}^n$, $a_i \in \mathbb{R}^n$, and $b_i \in \mathbb{R}$ for $i = 1, \dots, m$. The analytic center of the polytope is the minimizer of the following problem [8, p. 141]:

$$\min_x \quad -\sum_{i=1}^{m} \log(b_i - a_i^T x) \ . \tag{62}$$

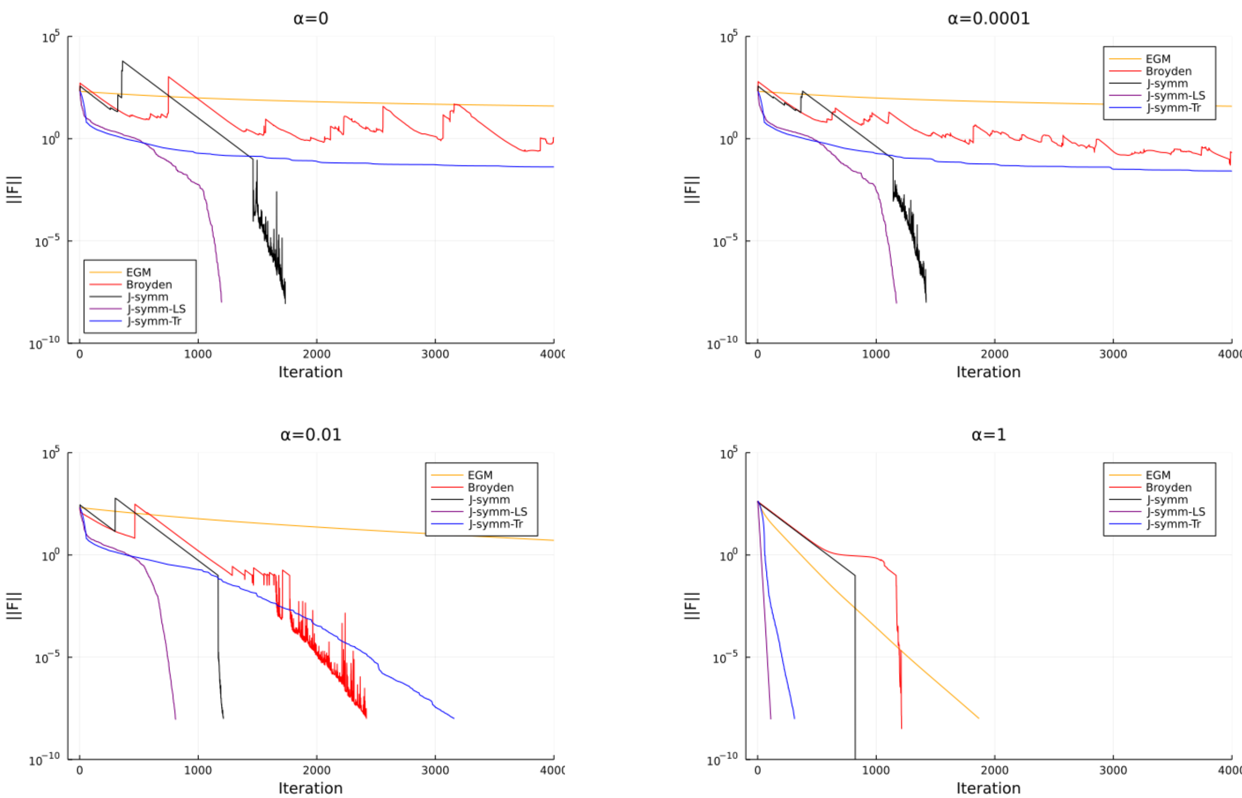

Figure 1: Plots showing $\|F(z)\|$ in log scale versus the number of iterations of EGM (yellow), Broyden's method (red), J-symmetric method with fixed stepsize (black), J-symmetric method with line-search (magenta) and J-symmetric method with trust-region (blue) for solving the quadratic convex-concave problems (61). The four figures are with $\alpha = 0$ (top-left), $\alpha = 10^{-4}$ (top-right), $\alpha = 10^{-2}$ (bottom-left) and $\alpha = 1$ (bottom-right), respectively.

The classical algorithm for finding the analytic center (62) is infeasible-start Newton method [25]. Here we focus on quasi-Newton methods, which avoid linear equation solving and can be used for larger instances. Notice that it can be nontrivial to identify a feasible solution to (62). We instead consider an equivalent formulation of (62):

$$\begin{aligned} \min_{x,y} \quad & -\sum_{i=1}^{m} \log y_i \\ \text{s.t.} \quad & y = b - Ax \ , \end{aligned}$$

where $A = [a_1, \ldots, a_m]^T$ and $b = [b_1, \ldots, b_m]$, and then we dualize the linear constraints to consider the minimax problem:

$$\min_{x,y} \max_{w} \ L(x, y, w) = -\sum_{i=1}^{m} \log y_i + w^T(Ax - b + y) \ . \tag{63}$$

Now we have a minimax problem of the form (1), and we can apply our J-symmetric methods.

In the numerical experiments, we generate $A \in \mathbb{R}^{100\times 500}$ and $b \in \mathbb{R}^{100}$ randomly from standard normal distribution. For quasi-Newton methods, we initialize the Hessian estimation $H_0 = I$. We terminate the algorithms when $\|F(z)\| \leq 10^{-4}$ (the value is chosen to avoid potential numerical issues caused by the log terms).

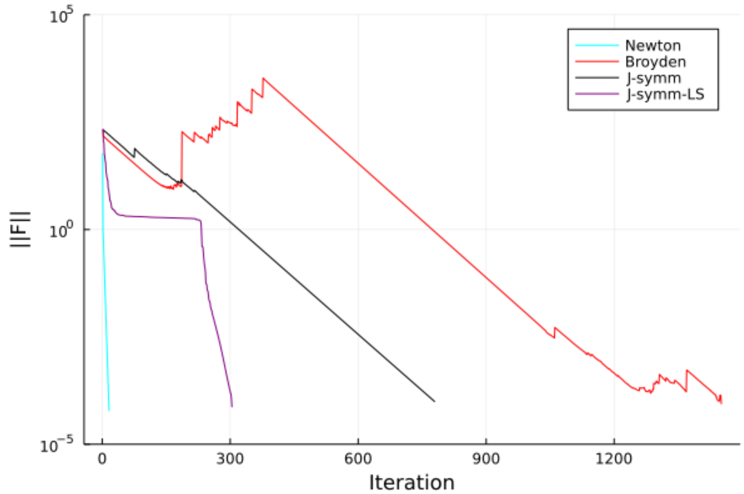

Figure 2: Plot showing $\|F(z)\|$ in log scale versus the number of iterations of Newton method (cyan), Broyden's method (red), J-symmetric method with fixed stepsize (black), J-symmetric method with line-search (magenta) for solving the analytic center problem (63).

Figure 2 compares the result of four algorithms: Newton's method, Broyden's method with fixed stepsize 0.02, J-symmetric quasi-Newton method (J-symm, Algorithm 1) with fixed stepsize 0.02, and J-symmetric quasi-Newton with line-search (J-symm-LS, Algorithm 2). The stepsize 0.02 is tuned to avoid blowing up. There is no doubt that Newton method has the fastest convergence rate (indeed it terminates only with 14 iterations). We can see that J-symm-LS stays in a plateau for a while to construct meaningful Jacobian estimation, and then enjoys a local superlinear convergence. Generally speaking, J-symm and Broyden's method both exhibit linear convergence with a similar rate, and J-symm obtains solutions with a certain tolerance earlier than Broyden's method.

5.3 A Nonconvex-Nonconcave Example

Many algorithms for minimax problems assume the problem to be convex-concave. For example, the proximal quasi-Newton methods proposed by Burke and Qian [13, 12] only work for monotone operators, thus do not work for nonconvex-nonconcave case. Furthermore, it is a well-known fact that many classical first-order methods fail to converge when applied to nonconvex-nonconcave problems [28]. In contrast, our J-symm algorithms and their analysis do not rely on convexity assumptions.

In this section, we examine the behaviors of our quasi-Newton algorithms on a two-dimension nonconvex-nonconcave example:

$$\min_x \min_y L(x,y) = (x^2-1)(x^2-9) + xAy - (y^2-1)(y^2-9) \ , \tag{64}$$

where A is a scalar measuring the interaction term in the minimax problem. (64) is perhaps the simplest non-trivial nonconvex-nonconcave example, and it has been used in [28] to illustrate the landscape of first-order methods for minimax problems.

Figure 3 presents the trajectories of five algorithms, EGM, Broyden's method, J-symmetric quasi-Newton method with a fixed stepsize (Algorithm 1), J-symmetric quasi-Newton method with line-search (Algorithm 2), and J-symmetric quasi-Newton method with trust-region (Algorithm 3), for solving (64). For each

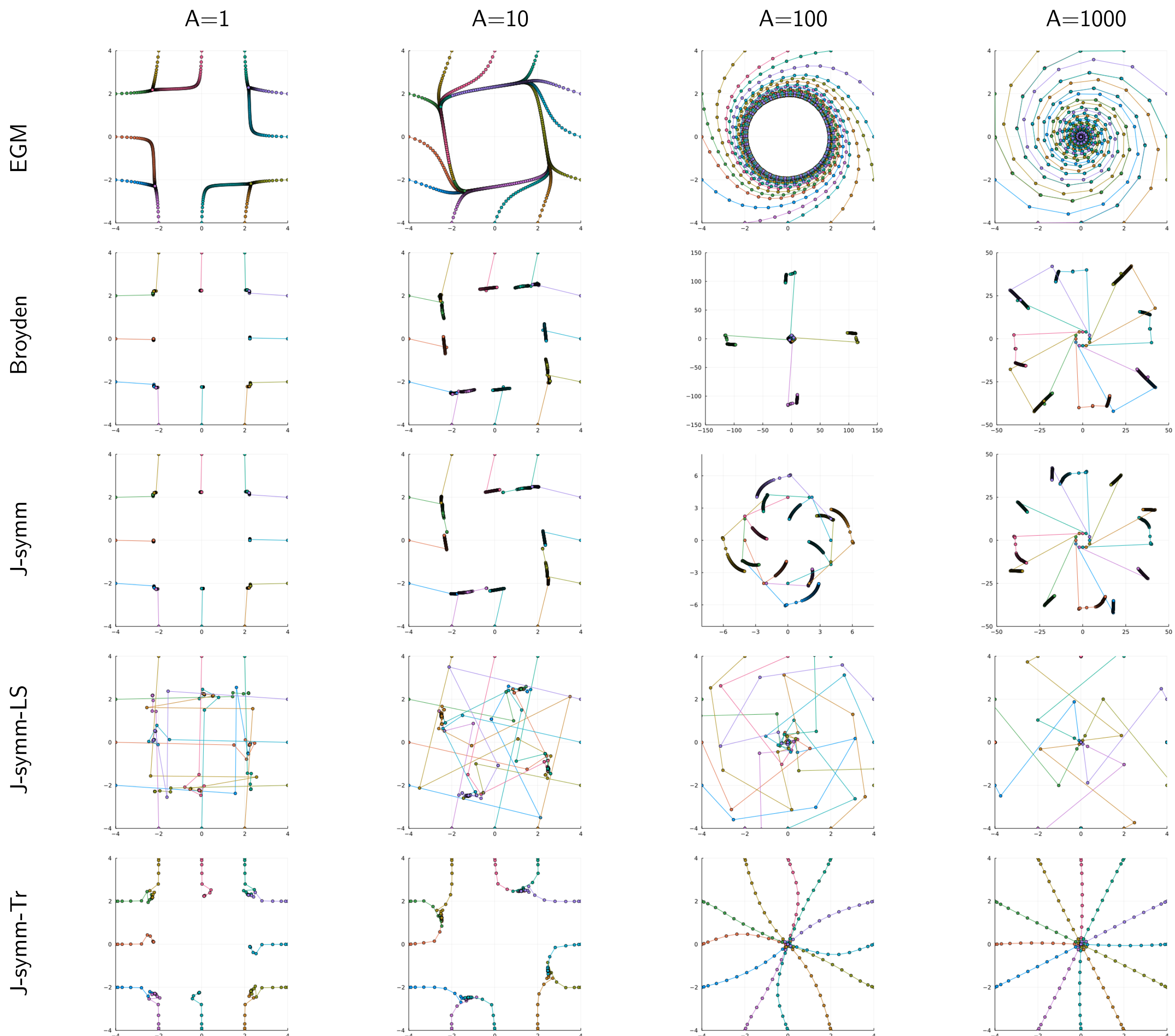

Figure 3: Plots showing the trajectories of five algorithms from twelve different initial solutions.

algorithm, we consider four interaction levels $A \in \{1, 10, 100, 1000\}$, and we start with twelve different initial solutions $(-4\ ,\ -2), (-4\ ,\ 0), (-4\ ,\ 2), (-2\ ,\ -4), (-2\ ,\ 4), (0\ ,\ -4), (0\ ,\ 4), (2\ ,\ -4), (2\ ,\ 4), (4\ ,\ -2), (4\ ,\ 0)$ and $(4\ ,\ 2)$.

When the interaction term A is small (i.e. $A = 1$ as shown in the first column), all five methods converge to local minimax solutions. Furthermore, we can clearly see that quasi-Newton methods have faster convergence compared to first-order methods such as EGM. When the interaction is medium (i.e. $A = 10$ as in the second column), EGM converges to an attractive limit circle, while all four quasi-Newton methods converge quickly to some solutions. It turns out that Algorithm 3 converges to a local minimizer of $\|F(z)\|$, i.e., $g_k = \nabla F(z_k)F(z_k) = 0$, which is consistent with our Theorem 4.2. As interaction A increases, we move to the third column (i.e., $A = 100$). While EGM still converges to a limit circle, quasi-Newton methods quickly converge to some solution. In particular, both J-symm-LS and J-symm-Tr converge to the unique global first-order Nash equilibrium $(0, 0)$ within a few iterations. J-symm and Broyden's method both converge to local solutions but compared to J-symm, Broyden's method is less stable since it moves further away in the beginning. Lastly, when the interaction term is sufficiently large (i.e., $A = 1000$ as in the fourth column), EGM converges to the unique stationary point $(0, 0)$, too. Again, J-symm-Tr converges to $(0, 0)$ within a few steps. However, while for some initial solutions J-symm-LS has rapid convergence to $(0, 0)$, for others (namely those with one dimension equal to 0, such as $(0, 4)$), it converges extremely slowly. This is because the line-search step chooses a very small stepsize. J-symm and Broyden's method both take a large step first, and move back slowly afterwards. The initial large step is because the stepsize choice 0.01 is too large initially for the case when $A = 1000$. The slow convergence is because the step size 0.01 is small once we construct a reasonable Jacobian.

Overall, in contrast to first-order methods quasi-Newton methods can avoid the undesirable limit circle for this nonconvex-nonconcave example. The trust-region method J-symm-Tr shows its advantages over the others, and indeed it is the only method with global theoretical guarantees. J-symm-LS performs well in most of the cases, but it may have slow convergence when the interaction term is large. Broyden's method and J-symm have similar behaviors, but J-symm may be more stable in the medium interaction regimes.

Acknowledgment

The authors would like to thank Jinwen Yang for checking the proofs in the paper.

References

[1] *Nicholas J. Higham on the top 10 algorithms in applied mathematics*, `https://press.princeton.edu/ideas/nicholas-higham-on-the-top-10-algorithms-in-applied-mathematics`, Accessed: 2022-01-10.

[2] Fatemeh Abdi and Fatemeh Shakeri, *A globally convergent BFGS method for pseudo-monotone variational inequality problems*, Optimization Methods and Software **34** (2019), no. 1, 25–36.

[3] David Applegate, Mateo Díaz, Oliver Hinder, Haihao Lu, Miles Lubin, Brendan O'Donoghue, and Warren Schudy, *Practical large-scale linear programming using primal-dual hybrid gradient*, arXiv preprint arXiv:2106.04756 (2021).

[4] Azam Asl and Michael L. Overton, *Analysis of the gradient method with an Armijo–Wolfe line search on a class of non-smooth convex functions*, Optimization Methods and Software **35** (2020), no. 2, 223–242.

[5] Michele Benzi and Gene H. Golub, *A preconditioner for generalized saddle point problems*, SIAM Journal on Matrix Analysis and Applications **26** (2004), no. 1, 20–41. MR 2112850

[6] Michele Benzi, Gene H. Golub, and Jörg Liesen, *Numerical solution of saddle point problems*, Acta numerica **14** (2005), 1–137.

[7] Dimitri P. Bertsekas, *Constrained optimization and lagrange multiplier methods*, Academic press, 2014.

[8] Stephen Boyd and Lieven Vandenberghe, *Convex optimization*, Cambridge University Press, March 2004.

[9] Charles G. Broyden, *A class of methods for solving nonlinear simultaneous equations*, Mathematics of computation **19** (1965), no. 92, 577–593.

[10] ———, *The convergence of single-rank quasi-Newton methods*, Math. Comp. **24** (1970), 365–382. MR 279993

[11] Charles G. Broyden, John E. Dennis, and Jorge J. Moré, *On the Local and Superlinear Convergence of Quasi-Newton Methods*, IMA Journal of Applied Mathematics **12** (1973), no. 3, 223–245.

[12] James V. Burke and Maijian Qian, *On the Superlinear Convergence of the Variable Metric Proximal Point Algorithm Using Broyden and BFGS Matrix Secant Updating*, Mathematical Programming **88** (1997), 157–181.

[13] ———, *A Variable Metric Proximal Point Algorithm for Monotone Operators*, SIAM Journal on Control and Optimization **37** (1999), no. 2, 353–375.

[14] Richard H. Byrd, Jorge Nocedal, and Ya Xiang Yuan, *Global convergence of a class of quasi-Newton methods on convex problems*, SIAM J. Numer. Anal. **24** (1987), no. 5, 1171–1190. MR 909072

[15] Antonin Chambolle and Thomas Pock, *A first-order primal-dual algorithm for convex problems with applications to imaging*, Journal of mathematical imaging and vision **40** (2011), no. 1, 120–145.

[16] Xiaojun Chen and Masao Fukushima, *Proximal quasi-Newton methods for nondifferentiable convex optimization*, Mathematical Programming. A Publication of the Mathematical Programming Society **85** (1999), no. 2, Ser. A, 313–334. MR 1700141

[17] Bo Dai, Albert Shaw, Lihong Li, Lin Xiao, Niao He, Zhen Liu, Jianshu Chen, and Le Song, *SBEED: Convergent reinforcement learning with nonlinear function approximation*, Proceedings of the 35th International Conference on Machine Learning (Jennifer Dy and Andreas Krause, eds.), Proceedings of Machine Learning Research, vol. 80, PMLR, 10–15 Jul 2018, pp. 1125–1134.

[18] Constantinos Daskalakis, Andrew Ilyas, Vasilis Syrgkanis, and Haoyang Zeng, *Training GANs with optimism*, International Conference on Learning Representations, 2018.

[19] William C. Davidon, *Variable metric method for minimization*, SIAM Journal on Optimization **1** (1991), no. 1, 1–17. MR 1094786

[20] John E. Dennis and Jorge J. Moré, *A characterization of superlinear convergence and its application to quasi-newton methods*, Mathematics of Computation **28** (1974), no. 126, 549–560.

[21] ———, *Quasi-newton methods, motivation and theory*, SIAM Review **19** (1977), no. 1, 46–89.

[22] Montacer Essid, Esteban Tabak, and Giulio Trigila, *An implicit gradient-descent procedure for minimax problems*, 2019.

[23] Richard Fletcher and Michael J. D. Powell, *A rapidly convergent descent method for minimization*, The Computer Journal **6** (1963/64), 163–168. MR 152116

[24] Gauthier Gidel, Hugo Berard, Gaëtan Vignoud, Pascal Vincent, and Simon Lacoste-Julien, *A variational inequality perspective on generative adversarial networks*, 2020.

[25] J. L. Goffin and F. Sharifi-Mokhtarian, *Primal–dual–infeasible newton approach for the analytic center deep-cutting plane method*, Journal of Optimization Theory and Applications **101** (1999), no. 1, 35–58.

[26] Donald Goldfarb, *A family of variable-metric methods derived by variational means*, Mathematics of Computation **24** (1970), no. 109, 23–26.

[27] Ian J. Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio, *Generative adversarial networks*, 2014.

[28] Benjamin Grimmer, Haihao Lu, Pratik Worah, and Vahab Mirrokni, *The landscape of the proximal point method for nonconvex-nonconcave minimax optimization*, arXiv preprint arXiv:2006.08667 (2020).

[29] Tero Karras, Timo Aila, Samuli Laine, and Jaakko Lehtinen, *Progressive growing of gans for improved quality, stability, and variation*, 2018.

[30] Korpelevich., *The extragradient method for finding saddle points and other problems*, Matecon **12** (1976).

[31] Christian Ledig, Lucas Theis, Ferenc Huszar, Jose Caballero, Andrew P. Aitken, Alykhan Tejani, Johannes Totz, Zehan Wang, and Wenzhe Shi, *Photo-realistic single image super-resolution using a generative adversarial network*, CoRR **abs/1609.04802** (2016).

[32] Haihao Lu, *An $o(s^r)$-resolution ode framework for understanding discrete-time algorithms and applications to the linear convergence of minimax problems*, 2021.

[33] D. Steven Mackey, Niloufer Mackey, and Françoise Tisseur, *Structured tools for structured matrices*, Electronic Journal of Linear Algebra **10** (2003), 106–145. MR 2001979

[34] Aleksander Madry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, and Adrian Vladu, *Towards deep learning models resistant to adversarial attacks*, International Conference on Learning Representations, 2018.

[35] Aryan Mokhtari, Asuman Ozdaglar, and Sarath Pattathil, *A unified analysis of extra-gradient and optimistic gradient methods for saddle point problems: Proximal point approach*, International Conference on Artificial Intelligence and Statistics, 2020.

[36] Jorge J. Moré and John A. Trangenstein, *On the Global Convergence of Broyden's Method*, Mathematics of Computation **30** (1976), no. 135, 523–540.

[37] Arkadi Nemirovski, *Prox-method with rate of convergence O(1/t) for variational inequalities with lipschitz continuous monotone operators and smooth convex-concave saddle point problems*, SIAM Journal on Optimization **15** (2004), no. 1, 229–251.

[38] Jorge Nocedal and Stephen J. Wright, *Numerical optimization*, 2nd ed., Springer, New York, 2006.

[39] James M. Ortega and Werner C. Rheinboldt, *Iterative solution of nonlinear equations in several variables*, Classics in Applied Mathematics, Society for Industrial and Applied Mathematics, 1970.

[40] Martin J. Osborne and Ariel Rubinstein, *A course in game theory*, MIT press, 1994.

[41] Barak A. Pearlmutter, *Fast exact multiplication by the hessian*, Neural computation **6** (1994), no. 1, 147–160.

[42] Michael J. D. Powell, *A hybrid method for nonlinear equations*, Numerical methods for nonlinear algebraic equations (Proc. Conf., Univ. Essex, Colchester, 1969), 1970, pp. 87–114. MR 0343589

[43] ———, *A new algorithm for unconstrained optimization*, Nonlinear Programming (Proc. Sympos., Univ. of Wisconsin, Madison, Wis., 1970), Academic Press, New York, 1970, pp. 31–65. MR 0272162

[44] ———, *Some global convergence properties of a variable metric algorithm for minimization without exact line searches*, Nonlinear Programming (Providence), Amer. Math. Soc., 1976, SIAM-AMS Proc., Vol. IX, pp. 53–72.

[45] R. Tyrrell Rockafellar, *Monotone operators and the proximal point algorithm*, SIAM Journal on Control and Optimization **14** (1976), no. 5, 877–898. MR 410483

[46] Nicol N. Schraudolph, *Fast curvature matrix-vector products for second-order gradient descent*, Neural computation **14** (2002), no. 7, 1723–1738.

[47] Avram Sidi, *A Zero-Cost Preconditioning for a Class of Indefinite Linear Systems*, WSEAS Transactions on Mathematics **2** (2003).

[48] Jun-Yan Zhu, Taesung Park, Phillip Isola, and Alexei A. Efros, *Unpaired image-to-image translation using cycle-consistent adversarial networks*, 2017 IEEE International Conference on Computer Vision (ICCV), 2017, pp. 2242–2251.

A Existing Definitions and Results Used in the Proofs.

Lemma A.1 (Banach Perturbation Lemma)**.** *([39, page 45]) Consider square matrices $A, B \in \mathbb{R}^{d\times d}$. Suppose that A is invertible with $\|A^{-1}\| \le a$. If $\|A - B\| \le b$ and $ab < 1$, then B is also invertible and*

$$\|B^{-1}\| \le \frac{a}{1-ab} \ .$$

Lemma A.2. *([21, Eq. (1.2)]) Consider square matrices $A, B \in \mathbb{R}^{d\times d}$. Then*

$$\|AB\|_F \le \min\{\ \|A\|_F\|B\|\ ,\ \|A\|\|B\|_F\ \} \ .$$

Definition A.3 (R-superlinear and Q-superlinear Convergence Rates [36])**.** *We say the sequence $\{z_k\}$ is converging to z^* R-superlinearly, if*

$$\lim_{k\to\infty} \|z_k - z^*\|^{1/k} = 0 \ ,$$

and $\{z_k\}$ is converging to z^* Q-superlinearly, if there exists a sequence $\{q_k\}$ converging to zero such that

$$\lim_{k\to\infty} \frac{\|z_{k+1} - z^*\|}{\|z_k - z^*\|} \le q_k \ .$$

Theorem A.4 (Dennis-Moré Q-superlinear Characterization Identity)**.** *([20, Theorem 2.2]) Let the mapping F be differentiable in the open convex set $\mathbb{D}$ and assume that for some $z^* \in \mathbb{D}$, ∇F is continuous at z^* and $\nabla F(z^*)$ is invertible. Let $\{B_k\}$ be a sequence of invertible matrices and suppose $\{z_k\}$, with $z_{k+1} = z_k - B_k^{-1}F(z_k)$, remains in $\mathbb{D}$ and converges to z^*. Then $\{z_k\}$ converges Q-superlinearly to z^* and $F(z^*) = 0$ iff*

$$\lim_{k\to\infty} \frac{\left\|\big(B_k - \nabla F(z^*)\big)(z_{k+1} - z_k)\right\|}{\|z_{k+1} - z_k\|} = 0 \ .$$

Definition A.5 (Uniform Linear Independence)**.** *([36, Definition 5.1.]) A sequence of unit vectors $\{u_j\}$ in $\mathbb{R}^{n+m}$ is uniformly linearly independent if there is $\beta > 0$, $k_0 \geq 0$ and $t \geq n+m$, such that for $k \geq k_0$ and $\|x\| = 1$, we have:*

$$\max\Big\{\big|\langle x\ ,\ u_j\rangle\big| : \quad j = k+1, \ldots, k+t\Big\} \geq \beta\ .$$

Theorem A.6. *([36, Theorem 5.3.]) Let $\{u_k\}$ be a sequence of unit vectors in $\mathbb{R}^{n+m}$. Then the following options are equivalent.*

- *The sequence $\{u_k\}$ is uniformly linearly independent.*
- *For any $\hat{\beta} \in [0,1)$ there is a constant $\theta \in (0,1)$ such that if $|\beta_j - 1| \leq \hat{\beta}$ then:*

$$\Big\|\prod_{j=k+1}^{k+t}\big(I - \beta_j u_j u_j^T\big)\Big\| \leq \theta, \quad \text{for} \quad k \geq k_0 \quad \text{and} \quad t \geq n+m\ .$$

Lemma A.7. *([36, Lemma 5.5.]) Let $\{\phi_k\}$ and $\{\delta_k\}$ be sequences of nonnegative numbers such that $\phi_{k+t} \leq \theta\phi_k + \delta_k$ for some fixed integer $t \geq 1$ and $\theta \in (0,1)$. If $\{\delta_k\}$ is bounded then $\{\phi_k\}$ is also bounded, and if in addition, $\{\delta_k\}$ converges to zero, then $\{\phi_k\}$ converges to zero.*