This is a set of notes from in-class discussion (APPM 4600) on methods for non-linear systems of equations. We will be assuming that we have a set of n equations in n real variables, which can be written as $F(x) = 0$ for $x \in \mathbb{R}^n$. As is the case for one non-linear equation, we will make smoothness assumptions on this $F(x)$ sufficient for the method to converge for an initial guess x_0 near a root r. For each method, we will center our discussion to answer two key questions:

1. Under what assumptions on $F(x)$ will the method converge for x_0 near the root?
2. What is the cost per iteration (as a function of n), and how quickly does the method converge?

Recall that, for general iterative methods, answering this second question gives us an estimate of the overall cost (and so the time it will take) to solve $F(x) = 0$.

1 Systems of non-linear equations and rootfinding

In applications, it is often the case that we have not one, but many (potentially non-linear) equations in many variables. The m equations are m conditions our variables $x_1, x_2, \ldots, x_n$ have to satisfy simultaneously; geometrically you can think of it as intersecting m "surfaces" in n dimensional space.

For example, let's say we want to find the intersections (if they exist) between two circles of radius 1, centered at $(1, 0)$ and $(2, 1)$, respectively. If we plot this (as we did in class), we find that there are two solutions for this problem: $\mathbf{r}_1 = (1, 1)$ and $\mathbf{r}_2 = (2, 0)$. We can also write the two conditions using the equation for a circle to find the system:

$$\begin{aligned}(x_1 - 1)^2 + x_2^2 &= 1 \\ (x_1 - 2)^2 + (x_2 - 1)^2 &= 1\end{aligned}$$

If we want to turn this into a rootfinding problem, we can define a function $F(\mathbf{x})$ such that the solution of this system is a root of F, that is, $F(\mathbf{r}) = \mathbf{0}$. In this case,

$$F(\mathbf{x}) = \begin{bmatrix} (x_1 - 1)^2 + x_2^2 - 1 \\ (x_1 - 2)^2 + (x_2 - 1)^2 - 1 \end{bmatrix}$$

So, in general, we can always define $F : \mathbb{R}^n \to \mathbb{R}^m$ such that our system of equations is equivalent to the rootfinding problem for F. The question we can now ask is: out of the methods we studied for the one-dimensional rootfinding problem, which ones would work in this case? If so, what (about their performance) stays the same? What changes?

Remark 1.1 *From now on, we will be assuming that $m = n$; that is, we have as many equations as we have variables. There are ways to use what we learn here and apply it to the cases $m > n$ and $m < n$, but we will not be going over them in this class.*

The first thing we can say is that bisection method is not applicable, as it uses ideas that only work in 1 dimension. However, the other 3 methods we discussed (FPI, Newton, Secant) will generalize to the multivariate case (with some modifications for the case of Secant).

2 Fixed Point Iteration

We write down a direct generalization of the Fixed Point Iteration, which we have previously discussed for scalar rootfinding problems. The idea remains the same: given $F(\mathbf{x}) = \mathbf{0}$, we come up with $G : \mathbb{R}^n \to \mathbb{R}^n$ such that $\mathbf{r}$ is a root of $F(\mathbf{x})$ if and only if $G(\mathbf{r}) = \mathbf{r}$. This is a very general idea. We can at least consider methods where G is of the form:

$$G(x) = x + \mathbf{S}(x)F(x) \tag{2.1}$$

where matrix $\mathbf{S}(x)$ is invertible (non-singular) on a neighborhood $B_\varepsilon(r)$ of the root. This of course includes the cases where $\mathbf{S}$ does not depend on x, and where it is multiplication by a scalar ($\mathbf{S} = s\mathbf{I}$).

2.1 Error analysis and convergence

In order to analyze the convergence of FPI, we need to extend the ideas we discussed for one nonlinear equation. Same as in that case, we can take iterate x_n and use Taylor expansion around the fixed point $\mathbf{r}$. We get the following equation:

$$\begin{aligned} G(\mathbf{x_n}) &= G(\mathbf{r}) + \mathbf{J}_G(\mathbf{r})(\mathbf{x}_n - \mathbf{r}) + O(||\mathbf{x_n} - \mathbf{r}||^2) \\ G(\mathbf{x_n}) - G(\mathbf{r}) &= \mathbf{J}_G(\mathbf{r})(\mathbf{x}_n - \mathbf{r}) + O(||\mathbf{x_n} - \mathbf{r}||^2) \\ \mathbf{e_{n+1}} &= \mathbf{J}_G(\mathbf{r})\mathbf{e_n} + O(||\mathbf{e_n}||^2) \end{aligned}$$

Where $\mathbf{J}_G(\mathbf{r})$ is the $n \times n$ Jacobian matrix of first derivatives, with (i, j) entries $\frac{\partial g_i}{\partial x_j}(\mathbf{r})$. This tells us that for $\mathbf{x_n}$ close to $\mathbf{r}$, Fixed Point Iteration from $\mathbf{x_0}$ gives us, ignoring high-order terms,

$$\mathbf{x_n} \simeq \mathbf{r} + \mathbf{J}_G^n \mathbf{e}_0$$

One way to interpret this is that, for $\mathbf{x_0}$ close to $\mathbf{r}$, we can use the linear model $L_G(\mathbf{x}) = G(\mathbf{r}) + \mathbf{J}_G(\mathbf{r})(\mathbf{x} - \mathbf{r})$ and the behavior of the Fixed Point Iteration for G will be the same as for L_G.

2.2 Fixed Point Iteration for Linear G(x)

So, say $G(\mathbf{x}) = \mathbf{Jx} - \mathbf{b}$. A Fixed Point $\mathbf{r}$ of G is such that $\mathbf{r} = \mathbf{Jr} - \mathbf{b}$. In other words, it is the solution to the linear system of equations

$$(\mathbf{J} - \mathbf{I})\mathbf{x} = \mathbf{b}$$

Let's say we begin the FPI from an initial guess $\mathbf{x_0}$. We have that:

$$\begin{aligned} G(\mathbf{x_k}) - G(\mathbf{r}) &= (\mathbf{Jx_k} - \mathbf{b}) - (\mathbf{Jr} - \mathbf{b}) \\ &= \mathbf{J}(\mathbf{x_k} - \mathbf{r}) \\ \mathbf{e_{k+1}} &= \mathbf{Je_k} = \mathbf{J}^k\mathbf{e_0} \end{aligned}$$

So, we need to ask: under what conditions does $\mathbf{J}^k\mathbf{e_0}$ go to zero as $k \to \infty$?

Say $\mathbf{J}$ is such that there is a basis of $\mathbb{R}^n$ (or $\mathbb{C}^n$) made of eigenvectors of $\mathbf{J}$. Let those eigenvectors be $\mathbf{v}_j$ with corresponding eigenvalues $\lambda_j \in \mathbb{C}$. If $\mathbf{e_0} = \alpha_1 \mathbf{v}_1 + \cdots + \alpha_n \mathbf{v}_n$, we have that:

$$\mathbf{J}\mathbf{e_0} = \sum_{j=1}^{n} \alpha_j \lambda_j \mathbf{v}_j$$
$$\mathbf{J}^2\mathbf{e_0} = \sum_{j=1}^{n} \alpha_j \lambda_j^2 \mathbf{v}_j$$
$$\vdots$$
$$\mathbf{J}^k\mathbf{e_0} = \sum_{j=1}^{n} \alpha_j \lambda_j^k \mathbf{v}_j$$

A sufficient condition for this to go to zero as $k \to \infty$ is that either α_j is zero or $|\lambda_j| < 1$ (and generally we do not want to ask stuff of $\mathbf{x_n} - \mathbf{r}$, as we do not know what $\mathbf{r}$ is!). So, one condition we can ask to ensure convergence of FPI is that $|\lambda_j| < 1$ for all j.

Theorem 2.1 (Fixed Point Iteration convergence (Linear case)) *Let $G(\mathbf{x}) = \mathbf{J}\mathbf{x} - \mathbf{b}$, and $\rho(\mathbf{J}) = \max(|\lambda_j|) < 1$ (this is known as the spectral radius of $\mathbf{J}$). Then, the Fixed Point Iteration converges linearly with rate at worst k for any initial guess $\mathbf{x_0}$.*

Examples: Linear transformations around a fixed point

When is a linear function contractive? The condition that makes G a contractive map with respect to a given norm $||\cdot||$ in the linear case is if there exists $0 \leq L < 1$ such that for all $\mathbf{x}, \mathbf{y}$, $||G(\mathbf{x}) - G(\mathbf{y})|| = ||\mathbf{J}(\mathbf{x} - \mathbf{y})|| \leq L||\mathbf{x} - \mathbf{y}||$. We can define

$$||\mathbf{J}|| = \max_{\mathbf{z} \neq \mathbf{0}} \frac{||\mathbf{J}\mathbf{z}||}{||\mathbf{z}||}$$

this is known as an operator norm. G is contractive if $||\mathbf{J}|| \leq k < 1$. If this is true, then FPI converges from any initial guess $\mathbf{x_0}$. While we will not cover it in this class, there is a way to relate this condition to the condition that $\rho(\mathbf{J}) = \max(|\lambda_j|) < 1$.

2.3 Fixed Point Theorem(s) for NonLinear G(x)

For non-linear $G(\mathbf{x})$ that is at least C^1 near r, we can now revisit the Taylor expansion around a fixed point $\mathbf{r}$:

$$G(\mathbf{x_n}) - G(\mathbf{r}) = \mathbf{J}_G(\mathbf{r})(\mathbf{x}_n - \mathbf{r}) + O(||\mathbf{x_n} - \mathbf{r}||^2)$$
$$\mathbf{e_{n+1}} = \mathbf{J}_G(\mathbf{r})\mathbf{e_n} + O(||\mathbf{e_n}||^2)$$

What this says is that, close enough to the root, G is contractive if $\mathbf{J}_G(\mathbf{r})$ is contractive. So, if we assume that $\rho(\mathbf{J}_G(r)) \leq k < 1$ or that $||\mathbf{J}(r)|| \leq k < 1$, then this condition ensures that FPI converges at least linearly (with rate k or smaller) for $|x_0 - r| < \delta$. We can write theorems analogous to those we had for the scalar case:

Theorem 2.2 *Assume that $G(\mathbf{x})$ is continuous in a closed and bounded subset $D \subset \mathbb{R}^n$. Then:*

(Existence) *If $G(\mathbf{x}) \in D$ for all $\mathbf{x} \in D$, then there is at least one fixed point $\mathbf{r}$ in D.*

(Uniqueness) *If, in addition, $\mathbf{J}_G(\mathbf{x})$ exists in D and there exists a positive constant $k < 1$ such that $\rho(\mathbf{J}(\mathbf{x})) \leq k$ for all $\mathbf{x} \in D$, then there is* exactly one *fixed point in D (the solution is unique).*

(Fixed Point iteration performance) *If the assumptions for uniqueness are met, then for any $\mathbf{x_0}$ in D, the fixed point iteration will converge* at least linearly *(it could be faster) to the unique solution $\mathbf{r}$.*

And we also have

Theorem 2.3 (Order and rate of convergence of Fixed Point Iteration) *Let $G(\mathbf{x})$ a function that is at least twice continuously differentiable on a neighborhood of the fixed point $\mathbf{r}$. Then, if $\rho(\mathbf{J}_G(\mathbf{r})) < 1$, then there exists a δ such that if $||\mathbf{x_0} - \mathbf{r}|| < \delta$, then the FPI converges linearly with rate $\rho(\mathbf{J}_g(\mathbf{r}))$ (or better).*

Note that the condition for both theorems could be changed to $||\mathbf{J}_G(r)|| < 1$, or most generally, to ask that there exists an $0 \leq L < 1$ such that, $\forall \mathbf{x}, \mathbf{y}$ such that $||\mathbf{x} - \mathbf{r}|| < \varepsilon, ||\mathbf{y} - \mathbf{r}|| < \varepsilon$,

$$||G(\mathbf{x}) - G(\mathbf{y})|| \leq L||\mathbf{x} - \mathbf{y}|| \tag{2.2}$$

(That $G(x)$ is a contractive map). To see how G being contractive is what we need, we once again can consider the sequence of iterates $\mathbf{x_{k+1}} = G(\mathbf{x_k})$, and show that

$$||\mathbf{e_{k+1}}|| = ||\mathbf{x_{k+1}} - \mathbf{r}|| = ||G(\mathbf{x_k}) - G(\mathbf{r})|| \leq L||\mathbf{x_k} - \mathbf{r}|| \leq \cdots \leq L^{k+1}||\mathbf{e_0}|| \tag{2.3}$$

And so, as we take the limit as $k \to \infty$, the error will go to zero. This also tells us that:

$$\limsup_{k\to\infty} \frac{||\mathbf{e_{k+1}}||}{||\mathbf{e_k}||} \leq L \tag{2.4}$$

So, convergence is linear (or better), with rate *at most L*.

3 Newton method

We once again extend the idea from the scalar Newton method: to use the linearization of our function $F(x)$ around our current guess x_k, and set it equal to 0. The linearization now involves the Jacobian of $F(x)$, of course:

$$L(\mathbf{x_{k+1}}) = F(\mathbf{x_k}) + J_F(\mathbf{x_k})(\mathbf{x_{k+1}} - \mathbf{x_k}) = 0$$
$$J_F(\mathbf{x_k})(\mathbf{x_{k+1}} - \mathbf{x_k}) = -F(\mathbf{x_k})$$
$$\mathbf{x_{k+1}} = \mathbf{x_k} - J_F(\mathbf{x_k})^{-1}F(\mathbf{x_k})$$

An important note on computing the Newton step: Even though $J_F(\mathbf{x_k})^{-1}$ appears in the formula, *DO NOT NEED TO COMPUTE THE INVERSE.* This is heavily discouraged for two reasons: not only is it unnecessary, but it is expensive and may be less stable (lead to more loss of precision). Instead, always compute $\mathbf{p_k} = \mathbf{x_{k+1}} - \mathbf{x_k}$ as the solution of the linear system given by $J_F(\mathbf{x_k})(\mathbf{x_{k+1}} - \mathbf{x_k}) = -F(\mathbf{x_k})$. For example, on python, $\mathbf{p_k} = np.linalg.solve(J_F(\mathbf{x_k}), -F(\mathbf{x_k}))$.

In other words: the Newton step is $\mathbf{x_{k+1}} = \mathbf{x_k} + \mathbf{p_k}$, where the step $\mathbf{p_k}$ is the solution to a system of n linear equations in n variables. Unless we know something special about our Jacobian, this means we will have to solve it using Gauss Elimination. This implies $O(n^3)$ cost per iteration.

As is the case for scalar Newton, given smooth $F(x)$ (we usually assume $F \in C^2(B_\delta(r))$ and $J_F(x)$ invertible in that neighborhood), we can show that there exists a neighborhood $B_\varepsilon(r)$ of r (with ε potentially smaller than δ) such that for x_0 in that neighborhood, Newton converges *quadratically.* This has the same issue as scalar Newton: we don't a priori know what this neighborhood is. We state the result for convergence of the Newton method without proof, but note that versions of the proofs for the scalar case can be extended to achieve it.

Theorem 3.1 *Let $F(\mathbf{x})$ be such that it is twice continuously differentiable and $J_F(\mathbf{x})$ is invertible on a neighborhood D of the root. Then, there exists $\delta > 0$ such that for all $\mathbf{x_0} \in B_\delta(r)$, Newton converges to* $\mathbf{r}$ quadratically*. Recall this means that for k sufficiently large,*

$$\frac{||\mathbf{x_{k+1}} - \mathbf{r}||_2}{||\mathbf{x_k} - \mathbf{r}||_2^2} \leq M \neq 0$$

where $B_\delta(\mathbf{r}) = \{\mathbf{x} \mid ||\mathbf{x} - \mathbf{r}||_2 < \delta\}$.

Note that this is exactly the same result as in the scalar case, except that it asks that the Jacobian be invertible on a neighborhood of the root. One thing you can experiment with is what happens if the Jacobian is invertible except at the root. Does Newton retain quadratic convergence?

3.1 The problems with Newton

Now, Newton is a fantastic method for rootfinding due to its quadratic convergence. However, there are a number of well-known issues with it, and these are greatly accentuated in systems of equations, especially for large n:

- Quadratic convergence is only guaranteed for a neighborhood of r. It may take the iteration a number of steps to get into this basin of quadratic convergence, *if it converges at all.* If we do NOT have confidence in our application that x_0 is close to a solution, it is dangerous to use Newton alone. We must use some other method to *get close* to r, or we must guide the Newton iteration. A family of methods to look into to guide Newton are *Line-search algorithms.* We can also, of course, use *hybrid* methods.

- Evaluating the $n \times n$ Jacobian matrix accurately may be, for some applications, really expensive and impractical. This is usually the case when we don't have a closed formula for $\mathbf{F}(\mathbf{x})$ (e.g. it is the state of some physical system).

- Even if we have $\mathbf{J}_F(x_k)$ readily available, we must solve a linear system for it. This is generally expensive. Unless our Jacobian is really special, we usually use Gaussian Elimination, which is $O(n^3)$.

4 Quasi-Newton methods

Quasi-Newton methods arise from the goal to find Newton-like methods that do not suffer from some of the issues Newton does, namely: the need to evaluate the Jacobian once per timestep, and the high cost per iteration of the corresponding linear system solve. We want a method that:

1. Is **superlinearly convergent** for x_0 close to r. This ensures small number of iterations once we are in the basin of superlinear convergence.

2. Does *not* evaluate a Jacobian or incur in $O(n^3)$ cost per timestep. We allow *at most* one Jacobian evaluation and one $O(n^3)$ cost (e.g. LU factorization) at the *beginning* of our algorithm.

4.1 Lazy and Approximate Newton methods

The first thing we tried is to fix the initial Jacobian and commit to it for the entirety of the iteration. This is informally known as ”Lazy Newton” (more formally known as the *chord iteration*). The step then becomes:

$$\mathbf{x}_{k+1} = \mathbf{x}_k - \mathbf{J}_F(\mathbf{x}_0)^{-1}F(x_k) \tag{4.1}$$

The advantage we gain is: given an LU or another such factorization of $J_F(x_0)$, we can calculate this step in at most $O(n^2)$ cost (2 triangular solves). However, because we have given up on this matrix changing, we can show that this is a Fixed Point Iteration where the matrix S does not change. In practice, this means convergence of the chord iteration is *linear, at best.*

Other ideas

- **Shaminskii method:** We can generalize the chord iteration by updating the Jacobian *every m iterations* instead of committing to the initial Jacobian. This improves convergence (potentially making it superlinear), but the cost per timestep goes up, so it might not be worth it.

- **Inaccurate or approximate Newton:** We can use an innaccurate or sparsified version of the Jacobian to compute the Newton step. This might, in some instances, alleviate the cost of evaluating the Jacobian and somewhat reduce the cost of the Newton step.

  For these methods, what we must show is that the innacurate Newton step $q_k \simeq -\tilde{\mathbf{J}}_F(x_k)^{-1}F(x_k)$ is close enough to the exact Newton step $p_k = -\mathbf{J}_F(x_k)^{-1}F(x_k)$. An exercise for the reader is to show that if $\|p_k - q_k\| \leq \eta\|F(x_k)\|$, then there exist $C, M > 0$ constants such that:

  $$\|e_{k+1}\| \leq C\|e_k\|^2 + \eta M\|e_k\|$$

  What this implies in practice is that the convergence of inexact Newton is quadratic *until the error is proportional to a multiple of η*. If we ask for more precision, it slows down to linear convergence (and would eventually plateau).

4.2 Quasi-Newton method: Broyden

The methods proposed above don’t satisfy both of our asks, or they do only under certain conditions. Quasi-Newton methods like **Broyden** (there is a whole zoo of them, especially for optimization), on the other hand, are the real deal: they retain superlinear convergence while providing formulas that allow us to compute the ”Quasi-Newton step” in at most $O(n^2)$ cost, if not faster. For this reason, they are widely used in modern rootfinding and smooth optimization routines and packages.

The ideas behind Quasi-Newton methods go back to the work pioneered by Broyden and his collaborators in the 60s (the method we discuss below was described in 1965). They are, in a sense, an extension of the secant method. Assume we have an initial guess $\mathbf{x}_0$, and an initial matrix $\mathbf{B}_0$

(usually $\mathbf{B}_0 \simeq J_F(x_0)$, but this isnt' required by the method). We then take one "Quasi-Newton step":

$$x_1 = x_0 - \mathbf{B}_0^{-1} F(x_0)$$

We now want to update the matrix from $\mathbf{B}_0$ to $\mathbf{B}_1$. What we want from $\mathbf{B}_1$ is:

1. To satisfy the **Secant equation** for x_1 and x_0. That is,

$$F(x_1) - F(x_0) = \mathbf{B}_1(x_1 - x_0) \tag{4.2}$$
$$\Delta F_0 = \mathbf{B}_1 \Delta x_0 \tag{4.3}$$

   where we denote $\Delta F_0 = F(x_1) - F(x_0)$, $\Delta x_0 = x_1 - x_0$. This should remind us of the slope of the secant line in one dimension.

2. **Rank 1 update:** $\mathbf{B}_1 = \mathbf{B}_0 + \mathbf{u}\mathbf{v}^T$. We can justify this by arguing that these matrices are imitating Jacobian matrices, which are continuous. $\mathbf{B}_1$ should thus be a minimal update of $\mathbf{B}_0$.

3. **Best rank 1 update:** Out of all possible rank 1 updates, $\mathbf{B}_1$ should be the closest to $\mathbf{B}_0$ in Frobenius norm. That is:

$$\mathbf{B}_1 = \text{argmin}\|\mathbf{B_1} - \mathbf{B_0}\|_F^2 = \text{argmin}\|\mathbf{u}\|_2^2\|\mathbf{v}\|_2^2 \tag{4.4}$$

   where the minimum is taken over $\mathbf{B}_1 = \mathbf{B}_0 + \mathbf{u}\mathbf{v}^T$.

4.3 Derivation of Broyden

We can use the 3 conditions above to derive the formula for Broyden. This is included in these notes for completion, but is beyond our syllabus. In practice, all that we need to implement is the formula obtained at the end of this subsection.

We substitute the rank 1 update formula into the secant equation. This gives us:

$$(\mathbf{B}_0 + \mathbf{u}\mathbf{v}^T)\Delta\mathbf{x}_0 = \Delta\mathbf{F}_0 \tag{4.5}$$
$$\mathbf{u}(\mathbf{v}^T \Delta x_0) = \Delta\mathbf{F}_0 - \mathbf{B}_0\Delta\mathbf{x}_0 = \mathbf{r}_0 \tag{4.6}$$
$$\mathbf{u} = \frac{1}{(\mathbf{v}^T \Delta x_0)}\mathbf{r}_0 \tag{4.7}$$

where $\mathbf{r}_0 = \Delta\mathbf{F}_0 - \mathbf{B}_0\Delta\mathbf{x}_0$ is a residual vector for the Secant Equation *applied to* $\mathbf{B}_0$. We note two things: if this residual is zero, this formula tells us to keep using $\mathbf{B}_0$. If it is non-zero, it gives us a formula for $\mathbf{u}$ as a function of $\mathbf{v}$.

The only condition we have not used is that the update should be minimal Frobenius norm:

$$\|\mathbf{B}_1 - \mathbf{B}_0\|_F^2 = \|\mathbf{u}\mathbf{v}^T\|_F^2 = \|\frac{\mathbf{r}_0}{(\mathbf{v}^T\Delta\mathbf{x}_0)}\mathbf{v}^T\|_F^2 \tag{4.8}$$

The Frobenius norm squared is the sum of all the matrix entries squared. For an exterior product $\mathbf{u}\mathbf{v}^T$ (rank 1 matrix), this is simply the product of the norms of each vector squared (exercise: show this). So, we must minimize:

$$\min_v \| \frac{\mathbf{r}_0}{(\mathbf{v}^T \Delta \mathbf{x}_0)} \mathbf{v}^T \|_F^2 = \min_v \| \frac{\mathbf{r}_0}{(\mathbf{v}^T \Delta \mathbf{x}_0)} \|_2^2 \|\mathbf{v}\|_2^2 \tag{4.9}$$

$$= \|\mathbf{r}_0\|_2^2 \min_v \frac{1}{|\mathbf{v}^T \Delta \mathbf{x}_0|^2} \|\mathbf{v}\|_2^2 \tag{4.10}$$

$$= \|\mathbf{r}_0\|_2^2 \min_v \frac{1}{\|\mathbf{v}\|_2^2 \|\Delta \mathbf{x}_0|\|_2^2 \cos^2\theta} \|\mathbf{v}\|_2^2 \tag{4.11}$$

$$= \frac{\|\mathbf{r}_0\|_2^2}{\|\Delta \mathbf{x}_0|\|_2^2} \min_v \frac{1}{\cos^2\theta} \tag{4.12}$$

Where θ is the angle between $\mathbf{v}$ and Δx_0. Clearly, this is minimized when this cosine is maximized, meaning $\theta = 0$, and $\mathbf{v} = \alpha \Delta x_0$ for some non-zero α. Putting everything together:

$$\mathbf{B}_1 = \mathbf{B}_0 + \frac{\mathbf{r}_0}{(\Delta \mathbf{x}_0)^T \Delta \mathbf{x}_0} \Delta \mathbf{x}_0^T \tag{4.13}$$

And so, the **Broyden update formula** reads:

$$\mathbf{B}_{k+1} = \mathbf{B}_k + \frac{\mathbf{r}_k}{(\Delta \mathbf{x}_k)^T \Delta \mathbf{x}_k} \Delta \mathbf{x}_k^T \tag{4.14}$$

where $\Delta \mathbf{F}_k = \mathbf{F}(\mathbf{x}_{k+1}) - \mathbf{F}(\mathbf{x}_k)$, $\Delta \mathbf{x}_k = \mathbf{x}_{k+1} - \mathbf{x}_k$ and $\mathbf{r}_k = \Delta \mathbf{F}_k - \mathbf{B}_k \Delta \mathbf{x}_k$.

Using the Sherman-Morrison formula, we can produce a formula *for the inverse update*, which is what is actually implemented in Broyden methods:

$$\mathbf{B}_{k+1}^{-1} = \mathbf{B}_k^{-1} + \frac{\Delta \mathbf{x}_k - \mathbf{B}_k^{-1} \Delta \mathbf{F}_k}{(\Delta \mathbf{x}_k)^T \mathbf{B}_k^{-1} \Delta \mathbf{F}_k} ((\mathbf{B}_k^{-1})^T \Delta \mathbf{x}_k)^T \tag{4.15}$$

We note that the most expensive operation in this rank 1 update formula for the inverse involves computing $\mathbf{B}_k^{-1} \Delta \mathbf{F}_k$ and $(\mathbf{B}_k^{-1})^T \Delta \mathbf{x}_k$.

4.4 Implementation details for Broyden

We assume that we have a function that, given a vector v, computes $\mathbf{B}_0^{-1}$ (that is, solves the system $\mathbf{B}_0 \mathbf{x} = \mathbf{b}$) in at most $O(n^2)$ work (e.g. we have computed an LU decomposition of $\mathbf{B}_0$). The pseudocode for Broyden will read as follows:

```
Given  $\mathbf{x}_0$,   $\mathbf{B}_0$ :
$k = 0$;
$np_k = 1$;
while  $(np_k \geq \varepsilon$  and  $k \leq k_{max})$ :
    $\mathbf{p}_k = -(\mathbf{B}_0^{-1} + \mathbf{U}\mathbf{V}^T)\mathbf{F}(\mathbf{x}_k)$
    $np_k = \|\mathbf{p}_k\|_\infty$
    $\mathbf{x}_{k+1} = \mathbf{x}_k + \mathbf{p}_k$

    $\mathbf{y}_k = (\mathbf{B}_0^{-1} + \mathbf{U}\mathbf{V}^T)\Delta\mathbf{F}_k$
    $\mathbf{z}_k = ((\mathbf{B}_0^{-1})^T + \mathbf{V}\mathbf{U}^T)\Delta\mathbf{x}_k$

    $\mathbf{U}(:, k+1) = \dfrac{\Delta\mathbf{x}_k - \mathbf{y}_k}{(\Delta\mathbf{x}_k)^T\mathbf{y}_k}$
    $\mathbf{V}(:, k+1) = \mathbf{z}_k$
    $k = k + 1$;
```

The key to implementing Broyden and other Quasi-Newton methods in practice, as can be seen from this pseudocode, is to write the rank 1 update to $\mathbf{B}_k^{-1}$ as a rank $k+1$ update to $\mathbf{B}_0^{-1}$. We store this as $\mathbf{U}\mathbf{V}^T$ for $n \times (k+1)$ matrices $\mathbf{U}, \mathbf{V}$

Key results

Broyden has the following properties:

- Broyden still is only guaranteed to converge superlinearly for $\mathbf{x}_0$ in a neighborhood of the root. This also depends on our choice of $\mathbf{B}_0$. Intuitively, the closer it is to $J_F(x_0)$, the more it will behave like Newton.

- Like Newton, this method needs to be guided to become more robust. We can use the same hybrid methods or *line-search algorithms* that work for Newton.

- Broyden's cost per iteration amounts to two solves for $\mathbf{B_0}$ plus $O(nk)$ work to apply the rank k update for the k-th iteration. Worst case scenario, this is $O(n^2)$. If we want to live dangerously, we can try $B_0 = sI$, which is $O(n)$.

- We don't have to compute the Jacobian during the iteration. This can be a major savings in computational cost.

5 Summary

Similar to our summary for linear solvers, we compile a table summarizing what we know about each of our methods to solve nonlinear systems of n equations in n variables $F(x) = 0$, with $J_F(x)$ the $n \times n$ Jacobian Matrix. Once again, we can ask:

- For what $F(x)$ and what initial values x_0 is this guaranteed to work?
- For iterative methods, what is the cost per iteration?
- For iterative methods, what do we know about convergence?

Let's recall what the step for each of the iterative methods looks like:

- **Fixed Point Iteration:** $x_{k+1} = G(x_k)$, where typically $G(x) = x - \mathbf{S}(x)F(x)$ for $\mathbf{S}(x)$ non-singular around the root.
- **Newton:** $x_{k+1} = x_k - \mathbf{J_F}^{-1}(x_k)F(x_k)$, where $\mathbf{J_F^{-1}}(x)$ non-singular around the root.
- **Lazy Newton:** $x_{k+1} = x_k - \mathbf{J_F}^{-1}(x_0)F(x_k)$, with similar assumptions as Newton.
- **Broyden:** $x_{k+1} = x_k - \mathbf{B_k}^{-1}F(x_k)$, with similar assumptions as Newton.

Method	Assumptions	Cost per Iteration	Convergence
Fixed Point	Contractive in $B_\varepsilon(r)$ x_0 near r	$F(x_k)$ eval + applying $\mathbf{S}$	Linear (at least)
Newton	$F \in C^2(B_\varepsilon(r))$ x_0 near r	$F(x_k), \mathbf{J}_F(x_k)$ eval Solve $\mathbf{J}_F(x_k)p_k = -F(x_k)$ $O(n^3)$	Quadratic!
Lazy Newton	Newton	Solve $\mathbf{J}_F(x_0)p_k = -F(x_k)$ $O(n^2)$	Linear
Approximate Newton	Newton + $\|q_k - p_k\|$ control	Approx solve $\mathbf{J}_F(x_k)p_k = -F(x_k)$ (complexity depends)	Quadratic until approx error
Broyden	Newton	Solve $\mathbf{B}_0\mathbf{p} = \mathbf{F}$ $O(n^2)$	Superlinear

6 Methods for smooth optimization and their relationship to rootfinding

A huge area of application of non-linear systems of equations is smooth unconstrained optimization. That is, we are trying to solve a problem of the form $\min q(\mathbf{x})$ for $q : \mathbb{R}^n \to \mathbb{R}$, and by that, we mean to find at least a local minimizer of this function. For this to exist, we just need this function to be locally convex. In other words, we need to find a critical point $\mathbf{r}$:

$$\nabla q(\mathbf{r}) = 0 \tag{6.1}$$

and the Hessian matrix (matrix of second derivatives) should be Symmetric Positive Definite (SPD) in a neighborhood of the critical point, so that it is a local minima.

We may notice that the problem of finding a critical point is, by itself, typically a system of non-linear equations. We may think to use the methods discussed above, as well as a generalization of the steepest descent algorithm, to solve this problem. At first, the fact that we don't want *any critical point, but a particular kind* (a local max or a saddle point won't do) sounds like a complication, but it is in fact a boon, because it *gives us a way to guide the steepest descent, Newton and Quasi-Newton iterations.* This makes them more reliable methods.

As we discussed in class, relating systems of non-linear equations (rootfinding) with smooth optimization gives us a powerful relationship that can help us solve these two types of problems. If the problem we want to solve is of the form $F(\mathbf{x}) = 0$, we can use an optimization solver (e.g. gradient descent) to find the minimum of the function

$$q(\mathbf{x}) = \sum_{j=1}^{n} F_j(\mathbf{x})^2 = ||F(\mathbf{x})||_2^2$$

we derived a formula for the gradient of this function: $\nabla q(\mathbf{x}) = \mathbf{J_F}(\mathbf{x})^T F(\mathbf{x})$. So, much as is the case for fixed point, if we have a rootfinding problem, we can solve it using optimization solvers, and if we have an optimization problem, we can solve it using rootfinding methods.

6.1 Steepest (Gradient) descent method

From multivariate calculus we know that given a function $q(\mathbf{x})$, we can produce a contour plot and then, for a given input x_0, we can ask: what is the direction of steepest ascent? What is the direction of steepest descent?

In other words, we can calculate the directional derivative of moving away from $\mathbf{x}_0$ in the direction of $\mathbf{p}$. If $q(x)$ is differentiable, we find that this derivative is equal to $\nabla q(x_0)^T \mathbf{d}$. In other words, if we define $\phi(\alpha) = q(\mathbf{x}_0 + \alpha \mathbf{p})$, then $\phi'(0) = \nabla q(\mathbf{x}_0)^T \mathbf{d}$. We then conclude that the direction that leads to the steepest descent (most negative value) is $\mathbf{p} = -\nabla q(\mathbf{x}_0)$.

Once we have chosen a direction to move in, and this is a direction in which *the value of q decreases, at least for some range of α*, we would like to pick α such that this descent is optimal. However, it is impractical to pick the best possible α.

Backtracking line-search algorithm

Instead, we start with $\alpha = 1$, and check what is known as "sufficient descent" conditions. That is, we check if our function value has gone down by a sufficient amount (Armijo condition), and we check that the slope of the tangent line at the new iterate is smaller in absolute value (Wolfe condition) by a sufficient amount. If so, we accept this α. If not, we reject and cut α by a factor of 0.5. We cut α until we either accept the step or we reach a maximum number of attempts (at which point we accept the tiny step).

In class and in our homework, we see that this implementation of gradient descent (with a backtracking linesearch) reliably converges to a local minimum (or to a root if we are applying this method to rootfinding) *linearly.* The rate can, however, be quite slow (close to 1), and for challenging functions, the iterations might ”zig-zag” to the solution.

We include below an introduction to applying other rootfinding methods to smooth optimization:

6.2 Newton method

The full Newton step for the problem of finding a critical point becomes:

$$\mathbf{x}_{k+1} = \mathbf{x}_k - \mathbf{H}q(\mathbf{x}_k)^{-1}\nabla q(\mathbf{x_k}) \tag{6.2}$$

where $\mathbf{H}q(\mathbf{x}_k)$ is the Hessian matrix. That is, we take a step in the direction p_k where

$$\mathbf{p}_k = -\mathbf{H}q(\mathbf{x}_k)^{-1}\nabla q(\mathbf{x_k}) \tag{6.3}$$

To make Newton more rubust, we implement the exact same backtracking line-search algorithm we used for steepest descent. That is, the Newton step becomes:

$$\mathbf{x}_{k+1} = \mathbf{x}_k + \alpha_k \mathbf{p}_k \tag{6.4}$$

where α_k is determined by the backtracking line-search to provide sufficient descent. This makes Newton converge from further away. Once our iterates get into the ”basin of quadratic convergence”, the method generally accepts $\alpha = 1$ and quickly converges to the solution to high accuracy.

6.3 Quasi-Newton methods

There are a number of Quasi-Newton methods specialized to smooth optimization. Broyden will not do very well, and for one key reason: it is not designed to make $\mathbf{B}_k$ an SPD matrix. Even if $\mathbf{B}_0$ is SPD, the updates are not guaranteed to remain SPD. Intuitively: the matrix we are imitating is now the Hessian. If $\mathbf{B}_k$ stops being SPD, the associated quadratic model for $f(\mathbf{x})$ is no longer locally convex, and our algorithm stops descending.

Here are some famous Quasi-Newton methods for optimization. They all involve either rank 1 or rank 2 updates for the Hessian-like matrix $\mathbf{B}_k$ every iteration:

- **BFGS** (Broyden, Fletcher, Goldfarb, Shanno)
- **DFP** (Davidon, Fletcher, Powell)
- **SR1** (Symmetric rank one)

Because they all involve low rank updates, they can all be implemented using exactly the same ideas that we used to implement Broyden, plus of course the same line-search algorithm to improve global convergence.

Out of these, BFGS and DFP are the most competitive; they each have their advantages in terms of performance. There are large-scale implementations of BFGS known as ”limited memory BFGS” that tackle enormous optimization problems.