Lecture 4 - The Gradient Method

Objective: find an optimal solution of the problem

$$\min\{f(\mathbf{x}) : \mathbf{x} \in \mathbb{R}^n\}.$$

The iterative algorithms that we will consider are of the form

$$\mathbf{x}_{k+1} = \mathbf{x}_k + t_k \mathbf{d}_k, k = 0, 1, \ldots$$

- $\mathbf{d}_k$ - direction.
- t_k - stepsize.

We will limit ourselves to descent directions.

Definition. Let $f : \mathbb{R}^n \to \mathbb{R}$ be a continuously differentiable function over $\mathbb{R}^n$. A vector $\mathbf{0} \neq \mathbf{d} \in \mathbb{R}^n$ is called a descent direction of f at $\mathbf{x}$ if the directional derivative $f'(\mathbf{x}; \mathbf{d})$ is negative, meaning that

$$f'(\mathbf{x}; \mathbf{d}) = \nabla f(\mathbf{x})^T \mathbf{d} < 0.$$

The Descent Property of Descent Directions

Lemma: Let f be a continuously differentiable function over $\mathbb{R}^n$, and let $\mathbf{x} \in \mathbb{R}^n$. Suppose that $\mathbf{d}$ is a descent direction of f at $\mathbf{x}$. Then there exists $\varepsilon > 0$ such that

$$f(\mathbf{x} + t\mathbf{d}) < f(\mathbf{x})$$

for any $t \in (0, \varepsilon]$.

Proof.

- Since $f'(\mathbf{x}; \mathbf{d}) < 0$, it follows from the definition of the directional derivative that

$$\lim_{t \to 0^+} \frac{f(\mathbf{x} + t\mathbf{d}) - f(\mathbf{x})}{t} = f'(\mathbf{x}; \mathbf{d}) < 0.$$

- Therefore, $\exists \varepsilon > 0$ such that

$$\frac{f(\mathbf{x} + t\mathbf{d}) - f(\mathbf{x})}{t} < 0$$

for any $t \in (0, \varepsilon]$, which readily implies the desired result.

See Lemma 4.3 for a stronger version of this result.

Schematic Descent Direction Method

Initialization: pick $\mathbf{x}_0 \in \mathbb{R}^n$ arbitrarily.
General step: for any $k = 0, 1, 2, \ldots$ set

(a) pick a descent direction $\mathbf{d}_k$.

(b) find a stepsize t_k satisfying $f(\mathbf{x}_k + t_k\mathbf{d}_k) < f(\mathbf{x}_k)$.

(c) set $\mathbf{x}_{k+1} = \mathbf{x}_k + t_k\mathbf{d}_k$.

(d) if a stopping criteria is satisfied, then STOP and $\mathbf{x}_{k+1}$ is the output.

Of course, many details are missing in the above schematic algorithm:

- What is the starting point?
- How to choose the descent direction?
- What stepsize should be taken?
- What is the stopping criteria?

Stepsize Selection Rules

- **constant stepsize** - $t_k = \bar{t}$ for any k.
- **exact stepsize** - t_k is a minimizer of f along the ray $\mathbf{x}_k + t\mathbf{d}_k$:

$$t_k \in \underset{t \geq 0}{\operatorname{argmin}}\, f(\mathbf{x}_k + t\mathbf{d}_k).$$

- **backtracking**[1] - The method requires three parameters: $s > 0, \alpha \in (0,1), \beta \in (0,1)$. Here we start with an initial stepsize $t_k = s$. While

$$f(\mathbf{x}_k) - f(\mathbf{x}_k + t_k\mathbf{d}_k) < -\alpha t_k \nabla f(\mathbf{x}_k)^T \mathbf{d}_k.$$

set $t_k := \beta t_k$

How do you know it will terminate? Ans: Lemma 4.3

Sufficient Decrease Property:

$$f(\mathbf{x}_k) - f(\mathbf{x}_k + t_k\mathbf{d}_k) \geq -\alpha t_k \nabla f(\mathbf{x}_k)^T \mathbf{d}_k.$$

[1] also referred to as Armijo

Exact Line Search for Quadratic Functions

$f(\mathbf{x}) = \mathbf{x}^T\mathbf{A}\mathbf{x} + 2\mathbf{b}^T\mathbf{x} + c$ where $\mathbf{A}$ is an $n \times n$ positive definite matrix, $\mathbf{b} \in \mathbb{R}^n$ and $c \in \mathbb{R}$. Let $\mathbf{x} \in \mathbb{R}^n$ and let $\mathbf{d} \in \mathbb{R}^n$ be a descent direction of f at $\mathbf{x}$. The objective is to find a solution to

$$\min_{t \geq 0} f(\mathbf{x} + t\mathbf{d}).$$

In class

The Gradient Method - Taking the Direction of Minus the Gradient

- In the gradient method $\mathbf{d}_k = -\nabla f(\mathbf{x}_k)$.
- This is a descent direction as long as $\nabla f(\mathbf{x}^k) \neq 0$ since

$$f'(\mathbf{x}_k; -\nabla f(\mathbf{x}_k)) = -\nabla f(\mathbf{x}_k)^T \nabla f(\mathbf{x}_k) = -\|\nabla f(\mathbf{x}_k)\|^2 < 0.$$

- In addition for being a descent direction, minus the gradient is also the steepest direction method.

Lemma: Let f be a continuously differentiable function and let $\mathbf{x} \in \mathbb{R}^n$ be a non-stationary point ($\nabla f(\mathbf{x}) \neq \mathbf{0}$). Then an optimal solution of

$$\min_{\mathbf{d}} \{f'(\mathbf{x}; \mathbf{d}) : \|\mathbf{d}\| = 1\} \tag{1}$$

is $\mathbf{d} = -\nabla f(\mathbf{x}) / \|\nabla f(\mathbf{x})\|$.

Proof. In class

The Gradient Method

The Gradient Method

Input: $\varepsilon > 0$ - tolerance parameter.

Initialization: pick $\mathbf{x}_0 \in \mathbb{R}^n$ arbitrarily.
General step: for any $k = 0, 1, 2, \ldots$ execute the following steps:

(a) pick a stepsize t_k by a line search procedure on the function

$$g(t) = f(\mathbf{x}_k - t\nabla f(\mathbf{x}_k)).$$

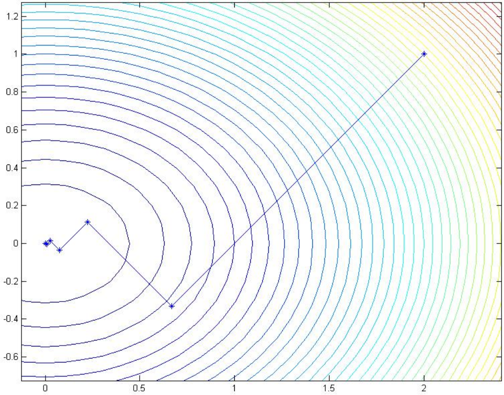
(b) set $\mathbf{x}_{k+1} = \mathbf{x}_k - t_k \nabla f(\mathbf{x}_k)$.

(c) if $\|\nabla f(\mathbf{x}_{k+1})\| \le \varepsilon$, then STOP and $\mathbf{x}_{k+1}$ is the output.

Numerical Example

$$\min x^2 + 2y^2$$

$\mathbf{x}_0 = (2;1), \varepsilon = 10^{-5}$, exact line search.



13 iterations until convergence.

The Zig-Zag Effect

Lemma. Let $\{\mathbf{x}_k\}_{k\geq 0}$ be the sequence generated by the gradient method with exact line search for solving a problem of minimizing a continuously differentiable function f. Then for any $k = 0, 1, 2, \ldots$

$$(\mathbf{x}_{k+2} - \mathbf{x}_{k+1})^T(\mathbf{x}_{k+1} - \mathbf{x}_k) = 0.$$

Proof.

- $\mathbf{x}_{k+1} - \mathbf{x}_k = -t_k \nabla f(\mathbf{x}_k), \mathbf{x}_{k+2} - \mathbf{x}_{k+1} = -t_{k+1} \nabla f(\mathbf{x}_{k+1})$.
- Therefore, we need to prove that $\nabla f(\mathbf{x}_k)^T \nabla f(\mathbf{x}_{k+1}) = 0$.
- $t_k \in \underset{t\geq 0}{\operatorname{argmin}}\{g(t) \equiv f(\mathbf{x}_k - t\nabla f(\mathbf{x}_k))\}$
- Hence, $g'(t_k) = 0$.
- $-\nabla f(\mathbf{x}_k)^T \nabla f(\mathbf{x}_k - t_k \nabla f(\mathbf{x}_k)) = 0$.
- $\nabla f(\mathbf{x}_k)^T \nabla f(\mathbf{x}_{k+1}) = 0$.

Numerical Example - Constant Stepsize, $\bar{t} = 0.1$

$$\min x^2 + 2y^2$$

$\mathbf{x}_0 = (2;1), \varepsilon = 10^{-5}, \bar{t} = 0.1.$

```
iter_number =   1 norm_grad = 4.000000 fun_val = 3.280000
iter_number =   2 norm_grad = 2.937210 fun_val = 1.897600
iter_number =   3 norm_grad = 2.222791 fun_val = 1.141888
     :                  :                    :
iter_number =  56 norm_grad = 0.000015 fun_val = 0.000000
iter_number =  57 norm_grad = 0.000012 fun_val = 0.000000
iter_number =  58 norm_grad = 0.000010 fun_val = 0.000000
```

- quite a lot of iterations...

Q: what is the problem here?

Numerical Example - Constant Stepsize, $\bar{t} = 10$

$$\min x^2 + 2y^2$$

$\mathbf{x}_0 = (2; 1), \varepsilon = 10^{-5}, \bar{t} = 10$..

```
iter_number =   1 norm_grad = 1783.488716 fun_val = 476806.000000
iter_number =   2 norm_grad = 656209.693339 fun_val = 56962873606.00
iter_number =   3 norm_grad = 256032703.004797 fun_val = 83183008071
     :                    :                                     :
iter_number = 119 norm_grad = NaN fun_val = NaN
```

- The sequence diverges:(

See HW 1

- Important question: how can we choose the constant stepsize so that convergence is guaranteed?

Lipschitz Continuity of the Gradient

Definition Let f be a continuously differentiable function over $\mathbb{R}^n$. We say that f has a Lipschitz gradient if there exists $L \geq 0$ for which

$$\|\nabla f(\mathbf{x}) - \nabla f(\mathbf{y})\| \leq L\|\mathbf{x} - \mathbf{y}\| \text{ for any } \mathbf{x}, \mathbf{y} \in \mathbb{R}^n.$$

L is called the Lipschitz constant.

- If ∇f is Lipschitz with constant L, then it is also Lipschitz with constant $\tilde{L}$ for all $\tilde{L} \geq L$.
- The class of functions with Lipschitz gradient with constant L is denoted by $C_L^{1,1}(\mathbb{R}^n)$ or just $C_L^{1,1}$.
- **Linear functions** - Given $\mathbf{a} \in \mathbb{R}^n$, the function $f(\mathbf{x}) = \mathbf{a}^T\mathbf{x}$ is in $C_0^{1,1}$.
- **Quadratic functions** - Let $\mathbf{A}$ be a symmetric $n \times n$ matrix, $\mathbf{b} \in \mathbb{R}^n$ and $c \in \mathbb{R}$. Then the function $f(\mathbf{x}) = \mathbf{x}^T\mathbf{A}\mathbf{x} + 2\mathbf{b}^T\mathbf{x} + c$ is a $C^{1,1}$ function. The smallest Lipschitz constant of ∇f is $2\|\mathbf{A}\|_2$ – why? In class

Equivalence to Boundedness of the Hessian

Theorem. Let f be a twice continuously differentiable function over $\mathbb{R}^n$. Then the following two claims are equivalent:

1. $f \in C_L^{1,1}(\mathbb{R}^n)$.
2. $\|\nabla^2 f(\mathbf{x})\| \le L$ for any $\mathbf{x} \in \mathbb{R}^n$.

Proof on pages 73,74 of the book

Example: $f(x) = \sqrt{1+x^2} \in C^{1,1}$

In class

Use this result for HW1, Problem 1(i).

Convergence of the Gradient Method

Theorem. Let $\{\mathbf{x}_k\}_{k\geq 0}$ be the sequence generated by GM for solving

$$\min_{\mathbf{x}\in\mathbb{R}^n} f(\mathbf{x})$$

with one of the following stepsize strategies:

- constant stepsize $\bar{t} \in \left(0, \frac{2}{L}\right)$.
- exact line search.
- backtracking procedure with parameters $s > 0$ and $\alpha, \beta \in (0, 1)$.

Assume that

- $f \in C_L^{1,1}(\mathbb{R}^n)$.
- f is bounded below over $\mathbb{R}^n$, that is, there exists $m \in \mathbb{R}$ such that $f(\mathbf{x}) > m$ for all $\mathbf{x} \in \mathbb{R}^n$).

Then

1. for any k, $f(\mathbf{x}_{k+1}) < f(\mathbf{x}_k)$ unless $\nabla f(\mathbf{x}_k) = \mathbf{0}$.
2. $\nabla f(\mathbf{x}_k) \to 0$ as $k \to \infty$.

Theorem 4.25 in the book.

Two Numerical Examples - Backtracking

$$\min x^2 + 2y^2$$

$\mathbf{x}_0 = (2; 1), s = 2, \alpha = 0.25, \beta = 0.5, \varepsilon = 10^{-5}$.

```
iter_number =   1 norm_grad = 2.000000 fun_val = 1.000000
iter_number =   2 norm_grad = 0.000000 fun_val = 0.000000
```

- fast convergence (also due to lack!)
- no real advantage to exact line search.

ANOTHER EXAMPLE:

$\min 0.01x^2 + y^2$, $s = 2, \alpha = 0.25, \beta = 0.5, \varepsilon = 10^{-5}$.

```
iter_number =   1 norm_grad = 0.028003 fun_val = 0.009704
iter_number =   2 norm_grad = 0.027730 fun_val = 0.009324
iter_number =   3 norm_grad = 0.027465 fun_val = 0.008958
      :                    :                    :
iter_number = 201 norm_grad = 0.000010 fun_val = 0.000000
```

Important Question: Can we detect key properties of the objective function that imply slow/fast convergence?

Kantorovich Inequality

By the spectral theorem, can WLOG assume A is diagonal.

Lemma. Let $\mathbf{A}$ be a positive definite $n \times n$ matrix. Then for any $\mathbf{0} \neq \mathbf{x} \in \mathbb{R}^n$ the inequality

$$\frac{\mathbf{x}^T\mathbf{x}}{(\mathbf{x}^T\mathbf{A}\mathbf{x})(\mathbf{x}^T\mathbf{A}^{-1}\mathbf{x})} \geq \frac{4\lambda_{\max}(\mathbf{A})\lambda_{\min}(\mathbf{A})}{(\lambda_{\max}(\mathbf{A}) + \lambda_{\min}(\mathbf{A}))^2}$$

holds.

Proof.

- Denote $m = \lambda_{\min}(\mathbf{A})$ and $M = \lambda_{\max}(\mathbf{A})$.
- The eigenvalues of the matrix $\mathbf{A} + Mm\mathbf{A}^{-1}$ are $\lambda_i(\mathbf{A}) + \frac{Mm}{\lambda_i(\mathbf{A})}$.
- The maximum of the 1-D function $\varphi(t) = t + \frac{Mm}{t}$ over $[m, M]$ is attained at the endpoints m and M with a corresponding value of $M + m$.
- Thus, the eigenvalues of $\mathbf{A} + Mm\mathbf{A}^{-1}$ are smaller than $(M + m)$.
- $\mathbf{A} + Mm\mathbf{A}^{-1} \preceq (M + m)\mathbf{I}$.
- $\mathbf{x}^T\mathbf{A}\mathbf{x} + Mm(\mathbf{x}^T\mathbf{A}^{-1}\mathbf{x}) \leq (M + m)(\mathbf{x}^T\mathbf{x})$,
- Therefore,

$$(\mathbf{x}^T\mathbf{A}\mathbf{x})[Mm(\mathbf{x}^T\mathbf{A}^{-1}\mathbf{x})] \leq \frac{1}{4}\left[(\mathbf{x}^T\mathbf{A}\mathbf{x}) + Mm(\mathbf{x}^T\mathbf{A}^{-1}\mathbf{x})\right]^2 \leq \frac{(M + m)^2}{4}(\mathbf{x}^T\mathbf{x})^2,$$

Gradient Method for Minimizing $\mathbf{x}^T\mathbf{A}\mathbf{x}$

Theorem. Let $\{\mathbf{x}_k\}_{k\geq 0}$ be the sequence generated by the gradient method with exact linesearch for solving the problem

$$\min_{\mathbf{x}\in\mathbb{R}^n} \mathbf{x}^T\mathbf{A}\mathbf{x} \quad (\mathbf{A}\succ \mathbf{0}).$$

Then for any $k = 0, 1, \ldots$:

$$f(\mathbf{x}_{k+1}) \leq \left(\frac{M-m}{M+m}\right)^2 f(\mathbf{x}_k),$$

where $M = \lambda_{\max}(\mathbf{A}), m = \lambda_{\min}(\mathbf{A})$.

Don't we lose too much generality just to consider f(x)=x' A x? (Take HW1 seriously and you may answer this question.)

Proof.

- $$\mathbf{x}_{k+1} = \mathbf{x}_k - t_k\mathbf{d}_k,$$

  where $t_k = \frac{\mathbf{d}_k^T\mathbf{d}_k}{2\mathbf{d}_k^T\mathbf{A}\mathbf{d}_k}, \mathbf{d}_k = 2\mathbf{A}\mathbf{x}_k$.

What about constant step size? (Also in HW1)

Proof of Rate of Convergence Contd.

- $$\begin{aligned} f(\mathbf{x}_{k+1}) &= \mathbf{x}_{k+1}^T\mathbf{A}\mathbf{x}_{k+1} = (\mathbf{x}_k - t_k\mathbf{d}_k)^T\mathbf{A}(\mathbf{x}_k - t_k\mathbf{d}_k) \\ &= \mathbf{x}_k^T\mathbf{A}\mathbf{x}_k - 2t_k\mathbf{d}_k^T\mathbf{A}\mathbf{x}_k + t_k^2\mathbf{d}_k^T\mathbf{A}\mathbf{d}_k \\ &= \mathbf{x}_k^T\mathbf{A}\mathbf{x}_k - t_k\mathbf{d}_k^T\mathbf{d}_k + t_k^2\mathbf{d}_k^T\mathbf{A}\mathbf{d}_k. \end{aligned}$$

- Plugging in the expression for t_k

$$\begin{aligned} f(\mathbf{x}_{k+1}) &= \mathbf{x}_k^T\mathbf{A}\mathbf{x}_k - \frac{1}{4}\frac{(\mathbf{d}_k^T\mathbf{d}_k)^2}{\mathbf{d}_k^T\mathbf{A}\mathbf{d}_k} \\ &= \mathbf{x}_k^T\mathbf{A}\mathbf{x}_k\left(1 - \frac{1}{4}\frac{(\mathbf{d}_k^T\mathbf{d}_k)^2}{(\mathbf{d}_k^T\mathbf{A}\mathbf{d}_k)(\mathbf{x}_k^T\mathbf{A}\mathbf{A}^{-1}\mathbf{A}\mathbf{x}_k)}\right) \\ &= \left(1 - \frac{(\mathbf{d}_k^T\mathbf{d}_k)^2}{(\mathbf{d}_k^T\mathbf{A}\mathbf{d}_k)(\mathbf{d}_k^T\mathbf{A}^{-1}\mathbf{d}_k)}\right)f(\mathbf{x}_k). \end{aligned}$$

- By Kantorovich:

$$f(\mathbf{x}_{k+1}) \le \left(1 - \frac{4Mm}{(M+m)^2}\right)f(\mathbf{x}_k) = \left(\frac{M-m}{M+m}\right)^2 f(\mathbf{x}_k) = \left(\frac{\kappa(\mathbf{A})-1}{\kappa(\mathbf{A})+1}\right)^2 f(\mathbf{x}_k),$$

The Condition Number

Definition. Let $\mathbf{A}$ be an $n \times n$ positive definite matrix. Then the condition number of $\mathbf{A}$ is defined by

$$\kappa(\mathbf{A}) = \frac{\lambda_{\max}(\mathbf{A})}{\lambda_{\min}(\mathbf{A})}.$$

- matrices (or quadratic functions) with large condition number are called ill-conditioned.
- matrices with small condition number are called well-conditioned.
- large condition number implies large number of iterations of the gradient method.
- small condition number implies small number of iterations of the gradient method.
- For a non-quadratic function, the asymptotic rate of convergence of $\mathbf{x}_k$ to a stationary point $\mathbf{x}^*$ is usually determined by the condition number of $\nabla^2 f(\mathbf{x}^*)$.

A Severely Ill-Condition Function - Rosenbrock

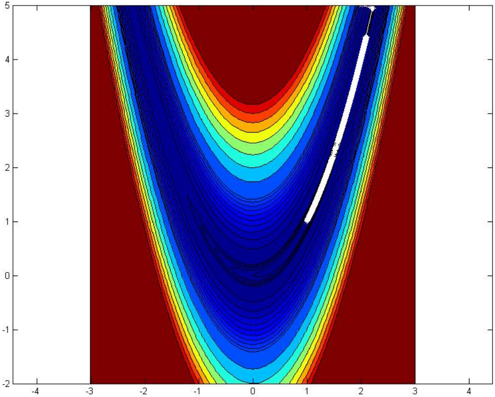
$$\min \left\{ f(x_1, x_2) = 100(x_2 - x_1^2)^2 + (1 - x_1)^2 \right\}.$$

- optimal solution:$(x_1, x_2) = (1, 1)$, optimal value: 0.
-
$$\begin{aligned} \nabla f(\mathbf{x}) &= \begin{pmatrix} -400x_1(x_2 - x_1^2) - 2(1 - x_1) \\ 200(x_2 - x_1^2) \end{pmatrix}, \\ \nabla^2 f(\mathbf{x}) &= \begin{pmatrix} -400x_2 + 1200x_1^2 + 2 & -400x_1 \\ -400x_1 & 200 \end{pmatrix}. \end{aligned}$$
-
$$\nabla^2 f(1, 1) = \begin{pmatrix} 802 & -400 \\ -400 & 200 \end{pmatrix}$$

  condition number: 2508

Solution of the Rosenbrock Problem with the Gradient Method

$\mathbf{x}_0 = (2; 5), s = 2, \alpha = 0.25, \beta = 0.5, \varepsilon = 10^{-5}$, backtracking stepsize selection.



6890(!!!) iterations.

Sensitivity of Solutions to Linear Systems

- Suppose that we are given the linear system

$$\mathbf{A}\mathbf{x} = \mathbf{b}$$

where $\mathbf{A} \succ \mathbf{0}$ and we assume that $\mathbf{x}$ is indeed the solution of the system ($\mathbf{x} = \mathbf{A}^{-1}\mathbf{b}$).

- Suppose that the right-hand side is perturbed $\mathbf{b} + \Delta\mathbf{b}$. What can be said on the solution of the new system $\mathbf{x} + \Delta\mathbf{x}$?
- $\Delta\mathbf{x} = \mathbf{A}^{-1}\Delta\mathbf{b}$.
- Result (derivation In class):

$$\frac{\|\Delta\mathbf{x}\|}{\|\mathbf{x}\|} \leq \kappa(\mathbf{A})\frac{\|\Delta\mathbf{b}\|}{\|\mathbf{b}\|}$$

Numerical Example

- consider the ill-condition matrix:

$$\mathbf{A} = \begin{pmatrix} 1+10^{-5} & 1 \\ 1 & 1+10^{-5} \end{pmatrix}$$

```
>> A=[1+1e-5,1;1,1+1e-5];
>> cond(A)
ans =
    2.000009999998795e+005
```

- We have

```
>> A\[1;1]
ans =
   0.499997500018278
   0.499997500006722
```

- However,

```
>> A\[1.1;1]
ans =
  1.0e+003 *
   5.000524997400047
  -4.999475002650021
```

Scaled Gradient Method

- Consider the minimization problem

$$\text{(P)} \quad \min\{f(\mathbf{x}) : \mathbf{x} \in \mathbb{R}^n\}.$$

- For a given nonsingular matrix $\mathbf{S} \in \mathbb{R}^{n \times n}$, we make the linear change of variables $\mathbf{x} = \mathbf{S}\mathbf{y}$, and obtain the equivalent problem

$$\text{(P')} \quad \min\{g(\mathbf{y}) \equiv f(\mathbf{S}\mathbf{y}) : \mathbf{y} \in \mathbb{R}^n\}.$$

- Since $\nabla g(\mathbf{y}) = \mathbf{S}^T \nabla f(\mathbf{S}\mathbf{y}) = \mathbf{S}^T \nabla f(\mathbf{x})$, the gradient method for (P') is

$$\mathbf{y}_{k+1} = \mathbf{y}_k - t_k \mathbf{S}^T \nabla f(\mathbf{S}\mathbf{y}_k).$$

- Multiplying the latter equality by $\mathbf{S}$ from the left, and using the notation $\mathbf{x}_k = \mathbf{S}\mathbf{y}_k$:

$$\mathbf{x}_{k+1} = \mathbf{x}_k - t_k \mathbf{S}\mathbf{S}^T \nabla f(\mathbf{x}_k).$$

- Defining $\mathbf{D} = \mathbf{S}\mathbf{S}^T$, we obtain the scaled gradient method:

$$\mathbf{x}_{k+1} = \mathbf{x}_k - t_k \mathbf{D} \nabla f(\mathbf{x}_k).$$

Scaled Gradient Method

- $\mathbf{D} \succ \mathbf{0}$, so the direction $-\mathbf{D}\nabla f(\mathbf{x}_k)$ is a descent direction:

$$f'(\mathbf{x}_k; -\mathbf{D}\nabla f(\mathbf{x}_k)) = -\nabla f(\mathbf{x}_k)^T \mathbf{D}\nabla f(\mathbf{x}_k) < 0,$$

We also allow different scaling matrices at each iteration.

Scaled Gradient Method

Input: $\varepsilon > 0$ - tolerance parameter.
Initialization: pick $\mathbf{x}_0 \in \mathbb{R}^n$ arbitrarily.
General step: for any $k = 0, 1, 2, \ldots$ execute the following steps:

(a) pick a scaling matrix $\mathbf{D}_k \succ \mathbf{0}$.

(b) pick a stepsize t_k by a line search procedure on the function

$$g(t) = f(\mathbf{x}_k - t\mathbf{D}_k \nabla f(\mathbf{x}_k)).$$

(c) set $\mathbf{x}_{k+1} = \mathbf{x}_k - t_k \mathbf{D}_k \nabla f(\mathbf{x}_k)$.

(c) if $\|\nabla f(\mathbf{x}_{k+1})\| \le \varepsilon$, then STOP and $\mathbf{x}_{k+1}$ is the output.

Choosing the Scaling Matrix $\mathbf{D}_k$

- The scaled gradient method with scaling matrix $\mathbf{D}$ is equivalent to the gradient method employed on the function $g(\mathbf{y}) = f(\mathbf{D}^{1/2}\mathbf{y})$.
- Note that the gradient and Hessian of g are given by

$$\begin{aligned} \nabla g(\mathbf{y}) &= \mathbf{D}^{1/2} f(\mathbf{D}^{1/2}\mathbf{y}) = \mathbf{D}^{1/2} f(\mathbf{x}), \\ \nabla^2 g(\mathbf{y}) &= \mathbf{D}^{1/2}\nabla^2 f(\mathbf{D}^{1/2}\mathbf{y})\mathbf{D}^{1/2} = \mathbf{D}^{1/2}\nabla^2 f(\mathbf{x})\mathbf{D}^{1/2}. \end{aligned}$$

.

- The objective is usually to pick $\mathbf{D}_k$ so as to make $\mathbf{D}_k^{1/2}\nabla^2 f(\mathbf{x}_k)\mathbf{D}_k^{1/2}$ as well-conditioned as possible.
- A well known choice (Newton's method): $\mathbf{D}_k = (\nabla^2 f(\mathbf{x}_k))^{-1}$. See HW2
- diagonal scaling: $\mathbf{D}_k$ is picked to be diagonal. For example,

$$(\mathbf{D}_k)_{ii} = \left(\frac{\partial^2 f(\mathbf{x}_k)}{\partial x_i^2}\right)^{-1}.$$

- Diagonal scaling can be very effective when the decision variables are of different magnitudes.

The Gauss-Newton Method

- Nonlinear least squares problem:

$$\text{(NLS):} \quad \min_{\mathbf{x}\in\mathbb{R}^n} \left\{ g(\mathbf{x}) \equiv \sum_{i=1}^{m} (f_i(\mathbf{x}) - c_i)^2 \right\}.$$

$f_1, \ldots, f_m$ are continuously differentiable over $\mathbb{R}^n$ and $c_1, \ldots, c_m \in \mathbb{R}$.

- Denote:

$$F(\mathbf{x}) = \begin{pmatrix} f_1(\mathbf{x}) - c_1 \\ f_2(\mathbf{x}) - c_2 \\ \vdots \\ f_m(\mathbf{x}) - c_m \end{pmatrix},$$

- Then the problem becomes:

$$\min \|F(\mathbf{x})\|^2.$$

The Gauss-Newton Method

Given the kth iterate $\mathbf{x}_k$, the next iterate is chosen to minimize the sum of squares of the linearized terms, that is,

$$\mathbf{x}_{k+1} = \operatorname*{argmin}_{\mathbf{x}\in\mathbb{R}^n} \left\{ \sum_{i=1}^{m} \left[f_i(\mathbf{x}_k) + \nabla f_i(\mathbf{x}_k)^T(\mathbf{x}-\mathbf{x}_k) - c_i \right]^2 \right\}.$$

- The general step actually consists of solving the linear LS problem

  $$\min \|\mathbf{A}_k\mathbf{x} - \mathbf{b}_k\|^2,$$

  where

  $$\mathbf{A}_k = \begin{pmatrix} \nabla f_1(\mathbf{x}_k)^T \\ \nabla f_2(\mathbf{x}_k)^T \\ \vdots \\ \nabla f_m(\mathbf{x}_k)^T \end{pmatrix} = J(\mathbf{x}_k)$$

  is the so-called Jacobian matrix, assumed to have full column rank.

  $$\mathbf{b}_k = \begin{pmatrix} \nabla f_1(\mathbf{x}_k)^T\mathbf{x}_k - f_1(\mathbf{x}_k) + c_1 \\ \nabla f_2(\mathbf{x}_k)^T\mathbf{x}_k - f_2(\mathbf{x}_k) + c_2 \\ \vdots \\ \nabla f_m(\mathbf{x}_k)^T\mathbf{x}_k - f_m(\mathbf{x}_k) + c_m \end{pmatrix} = J(\mathbf{x}_k)\mathbf{x}_k - F(\mathbf{x}_k)$$

The Gauss-Newton Method

- The Gauss-Newton method can thus be written as:

$$\mathbf{x}_{k+1} = (J(\mathbf{x}_k)^T J(\mathbf{x}_k))^{-1} J(\mathbf{x}_k)^T \mathbf{b}_k.$$

- The gradient of the objective function $f(\mathbf{x}) = \|F(\mathbf{x})\|^2$ is

$$\nabla f(\mathbf{x}) = 2J(\mathbf{x})^T F(\mathbf{x})$$

- The GN method can be rewritten as follows:

$$\begin{aligned} \mathbf{x}_{k+1} &= (J(\mathbf{x}_k)^T J(\mathbf{x}_k))^{-1} J(\mathbf{x}_k)^T (J(\mathbf{x}_k)\mathbf{x}_k - F(\mathbf{x}_k)) \\ &= \mathbf{x}_k - (J(\mathbf{x}_k)^T J(\mathbf{x}_k))^{-1} J(\mathbf{x}_k)^T F(\mathbf{x}_k) \\ &= \mathbf{x}_k - \frac{1}{2}(J(\mathbf{x}_k)^T J(\mathbf{x}_k))^{-1} \nabla f(\mathbf{x}_k), \end{aligned}$$

- that is, it is a scaled gradient method with a special choice of scaling matrix:

$$\mathbf{D}_k = \frac{1}{2}(J(\mathbf{x}_k)^T J(\mathbf{x}_k))^{-1}.$$

The Damped Gauss-Newton Method

The Gauss-Newton method does not incorporate a stepsize, which might cause it to diverge. A well known variation of the method incorporating stepsizes is the damped Gauss-newton Method.

Damped Gauss-Newton Method

Input: ε - tolerance parameter.

Initialization: pick $\mathbf{x}_0 \in \mathbb{R}^n$ arbitrarily.
General step: for any $k = 0, 1, 2, \dots$ execute the following steps:

(a) Set $\mathbf{d}_k = -(J(\mathbf{x}_k)^T J(\mathbf{x}_k))^{-1} J(\mathbf{x}_k)^T F(\mathbf{x}_k)$.

(b) Set t_k by a line search procedure on the function

$$h(t) = g(\mathbf{x}_k + t\mathbf{d}_k).$$

(c) set $\mathbf{x}_{k+1} = \mathbf{x}_k + t_k \mathbf{d}_k$.

(c) if $\|\nabla f(\mathbf{x}_{k+1})\| \le \varepsilon$, then STOP and $\mathbf{x}_{k+1}$ is the output.

Fermat-Weber Problem

Fermat-Weber Problem: Given m points in $\mathbb{R}^n$: $\mathbf{a}_1, \ldots, \mathbf{a}_m$ – also called "anchor point" – and m weights $\omega_1, \omega_2, \ldots, \omega_m > 0$, find a point $\mathbf{x} \in \mathbb{R}^n$ that minimizes the weighted distance of $\mathbf{x}$ to each of the points $\mathbf{a}_1, \ldots, \mathbf{a}_m$:

$$\min_{\mathbf{x} \in \mathbb{R}^n} \left\{ f(\mathbf{x}) \equiv \sum_{i=1}^{m} \omega_i \|\mathbf{x} - \mathbf{a}_i\| \right\}.$$

- The objective function is not differentiable at the anchor points $\mathbf{a}_1, \ldots, \mathbf{a}_m$.
- One of the simplest instances of facility location problems.

Weiszfeld's Method (1937)

- Start from the stationarity condition $\nabla f(\mathbf{x}) = \mathbf{0}$.[2]
- $\sum_{i=1}^m \omega_i \frac{\mathbf{x}-\mathbf{a}_i}{\|\mathbf{x}-\mathbf{a}_i\|} = \mathbf{0}$.
- $\left(\sum_{i=1}^m \frac{\omega_i}{\|\mathbf{x}-\mathbf{a}_i\|}\right)\mathbf{x} = \sum_{i=1}^m \frac{\omega_i \mathbf{a}_i}{\|\mathbf{x}-\mathbf{a}_i\|}$,
- $\mathbf{x} = \frac{1}{\sum_{i=1}^m \frac{\omega_i}{\|\mathbf{x}-\mathbf{a}_i\|}} \sum_{i=1}^m \frac{\omega_i \mathbf{a}_i}{\|\mathbf{x}-\mathbf{a}_i\|}$.
- The stationarity condition can be written as $\mathbf{x} = T(\mathbf{x})$, where T is the operator

$$T(\mathbf{x}) \equiv \frac{1}{\sum_{i=1}^m \frac{\omega_i}{\|\mathbf{x}-\mathbf{a}_i\|}} \sum_{i=1}^m \frac{\omega_i \mathbf{a}_i}{\|\mathbf{x}-\mathbf{a}_i\|}.$$

- Weiszfeld's method is a fixed point method:

$$\mathbf{x}_{k+1} = T(\mathbf{x}_k).$$

[2] We implicitly assume here that $\mathbf{x}$ is not an anchor point.

Weiszfeld's Method as a Gradient Method

Weiszfeld's Method

Initialization: pick $\mathbf{x}_0 \in \mathbb{R}^n$ such that $\mathbf{x} \neq \mathbf{a}_1, \mathbf{a}_2, \ldots, \mathbf{a}_m$.

General step: for any $k = 0, 1, 2, \ldots$ compute:

$$\mathbf{x}_{k+1} = T(\mathbf{x}_k) = \frac{1}{\sum_{i=1}^m \frac{\omega_i}{\|\mathbf{x}_k - \mathbf{a}_i\|}} \sum_{i=1}^m \frac{\omega_i \mathbf{a}_i}{\|\mathbf{x}_k - \mathbf{a}_i\|}.$$

- Weiszfeld's method is a gradient method since

$$\begin{aligned}
\mathbf{x}_{k+1} &= \frac{1}{\sum_{i=1}^m \frac{\omega_i}{\|\mathbf{x}_k - \mathbf{a}_i\|}} \sum_{i=1}^m \frac{\omega_i \mathbf{a}_i}{\|\mathbf{x}_k - \mathbf{a}_i\|} \\
&= \mathbf{x}_k - \frac{1}{\sum_{i=1}^m \frac{\omega_i}{\|\mathbf{x}_k - \mathbf{a}_i\|}} \sum_{i=1}^m \omega_i \frac{\mathbf{x}_k - \mathbf{a}_i}{\|\mathbf{x}_k - \mathbf{a}_i\|} \\
&= \mathbf{x}_k - \frac{1}{\sum_{i=1}^m \frac{\omega_i}{\|\mathbf{x}_k - \mathbf{a}_i\|}} \nabla f(\mathbf{x}_k).
\end{aligned}$$

- A gradient method with a special choice of stepsize: $t_k = \frac{1}{\sum_{i=1}^m \frac{\omega_i}{\|\mathbf{x}_k - \mathbf{a}_i\|}}$.