**Lecture 4 - The Gradient Method**

**Objective:** find an optimal solution of the problem

\[
\min\{f(x) : x \in \mathbb{R}^n\}.
\]

The iterative algorithms that we will consider are of the form

\[
x_{k+1} = x_k + t_k d_k, \quad k = 0, 1, \ldots
\]

- \(d_k\) - 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 \(0 \neq d \in \mathbb{R}^n\) is called a descent direction of \(f\) at \(x\) if the directional derivative \(f'(x; d)\) is negative, meaning that

\[
f'(x; d) = \nabla f(x)^T d < 0.
\]
Lemma: Let \( f \) be a continuously differentiable function over \( \mathbb{R}^n \), and let \( x \in \mathbb{R}^n \). Suppose that \( d \) is a descent direction of \( f \) at \( x \). Then there exists \( \varepsilon > 0 \) such that

\[
f(x + td) < f(x)
\]

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

Proof.

\( \triangleright \) Since \( f'(x; d) < 0 \), it follows from the definition of the directional derivative that

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

\( \triangleright \) Therefore, \( \exists \varepsilon > 0 \) such that

\[
\frac{f(x + td) - f(x)}{t} < 0
\]

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 $x_0 \in \mathbb{R}^n$ arbitrarily.

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

(a) pick a descent direction $d_k$.
(b) find a stepsize $t_k$ satisfying $f(x_k + t_k d_k) < f(x_k)$.
(c) set $x_{k+1} = x_k + t_k d_k$.
(d) if a stopping criteria is satisfied, then STOP and $x_{k+1}$ 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 $x_k + t d_k$:
  \[ t_k \in \arg\min_{t \geq 0} f(x_k + t d_k). \]

- **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
  \[ f(x_k) - f(x_k + t_k d_k) < -\alpha t_k \nabla f(x_k)^T d_k. \]
  set $t_k := \beta t_k$

**Sufficient Decrease Property:**

\[ f(x_k) - f(x_k + t_k d_k) \geq -\alpha t_k \nabla f(x_k)^T d_k. \]

\(^1\)also referred to as Armijo

How do you know it will terminate? Ans: Lemma 4.3
Exact Line Search for Quadratic Functions

\[ f(x) = x^T Ax + 2b^T x + c \]

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

\[
\min_{t \geq 0} f(x + td).
\]

In class
The Gradient Method - Taking the Direction of Minus the Gradient

▶ In the gradient method $d_k = -\nabla f(x_k)$.
▶ This is a descent direction as long as $\nabla f(x^k) \neq 0$ since

$$f'(x_k; -\nabla f(x_k)) = -\nabla f(x_k)^T \nabla f(x_k) = -\|\nabla f(x_k)\|^2 < 0.$$ 

▶ 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 $x \in \mathbb{R}^n$ be a non-stationary point ($\nabla f(x) \neq 0$). Then an optimal solution of

$$\min_d \{f'(x; d) : \|d\| = 1\} \quad (1)$$

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

Proof. In class
The Gradient Method

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

Initialization: pick $x_0 \in \mathbb{R}^n$ 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

$$g(t) = f(x_k - t \nabla f(x_k)).$$

(b) set $x_{k+1} = x_k - t_k \nabla f(x_k)$.

(c) if $\|\nabla f(x_{k+1})\| \leq \varepsilon$, then STOP and $x_{k+1}$ is the output.
Numerical Example

\[ \min x^2 + 2y^2 \]

\[ x_0 = (2; 1), \varepsilon = 10^{-5}, \text{exact line search.} \]

13 iterations until convergence.
The Zig-Zag Effect

Lemma. Let \( \{x_k\}_{k \geq 0} \) 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 \( k = 0, 1, 2, \ldots \)

\[
(x_{k+2} - x_{k+1})^T (x_{k+1} - x_k) = 0.
\]

Proof.

- \( x_{k+1} - x_k = -t_k \nabla f(x_k), x_{k+2} - x_{k+1} = -t_{k+1} \nabla f(x_{k+1}). \)
- Therefore, we need to prove that \( \nabla f(x_k)^T \nabla f(x_{k+1}) = 0. \)
- \( t_k \in \arg\min_{t \geq 0} \{g(t) \equiv f(x_k - t \nabla f(x_k))\} \)
- Hence, \( g'(t_k) = 0. \)
- \( -\nabla f(x_k)^T \nabla f(x_k - t_k \nabla f(x_k)) = 0. \)
- \( \nabla f(x_k)^T \nabla f(x_{k+1}) = 0. \)
Numerical Example - Constant Stepsize, $\bar{t} = 0.1$

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

$x_0 = (2; 1), \varepsilon = 10^{-5}, \bar{t} = 0.1.$

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, $\bar{t} = 10$

\[ \min x^2 + 2y^2 \]

\[ x_0 = (2; 1), \epsilon = 10^{-5}, \bar{t} = 10 \ldots \]

<table>
<thead>
<tr>
<th>iter_number</th>
<th>norm_grad</th>
<th>fun_val</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1783.488716</td>
<td>476806.000000</td>
</tr>
<tr>
<td>2</td>
<td>656209.693339</td>
<td>56962873606.0000</td>
</tr>
<tr>
<td>3</td>
<td>256032703.004797</td>
<td>8318300807190406.0000</td>
</tr>
<tr>
<td>...</td>
<td>...</td>
<td>...</td>
</tr>
<tr>
<td>119</td>
<td>NaN</td>
<td>NaN</td>
</tr>
</tbody>
</table>

The sequence diverges:(

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

See HW 1
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

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

$L$ is called the **Lipschitz constant**.

- If $\nabla f$ 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 $C^{1,1}_L(\mathbb{R}^n)$ or just $C^{1,1}_L$.
- **Linear functions** - Given $a \in \mathbb{R}^n$, the function $f(x) = a^T x$ is in $C^{1,1}_0$.
- **Quadratic functions** - Let $A$ be a symmetric $n \times n$ matrix, $b \in \mathbb{R}^n$ and $c \in \mathbb{R}$. Then the function $f(x) = x^T Ax + 2b^T x + c$ is a $C^{1,1}$ function. The smallest Lipschitz constant of $\nabla f$ is $2\|A\|_2$ – 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. $f \in C^{1,1}_L(\mathbb{R}^n)$.
2. $\|\nabla^2 f(x)\| \leq L$ for any $x \in \mathbb{R}^n$.

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 \( \{x_k\}_{k \geq 0} \) be the sequence generated by GM for solving

\[
\min_{x \in \mathbb{R}^n} f(x)
\]

with one of the following stepsize strategies:

- constant stepsize \( \bar{t} \in (0, \frac{2}{L}) \).
- 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 \( f(x) > m \) for all \( x \in \mathbb{R}^n \).

Then

1. for any \( k \), \( f(x_{k+1}) < f(x_k) \) unless \( \nabla f(x_k) = 0 \).
2. \( \nabla f(x_k) \to 0 \) 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
\vdots
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

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

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

holds.

Proof.

- Denote $m = \lambda_{\min}(A)$ and $M = \lambda_{\max}(A)$.
- The eigenvalues of the matrix $A + MmA^{-1}$ are $\lambda_i(A) + \frac{Mm}{\lambda_i(A)}$.
- 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 $A + MmA^{-1}$ are smaller than $(M + m)$.
- $A + MmA^{-1} \preceq (M + m)I$.
- $x^T Ax + Mm(x^T A^{-1} x) \leq (M + m)(x^T x)$,
- Therefore,

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

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

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

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

$$
f(x_{k+1}) \leq \left( \frac{M - m}{M + m} \right)^2 f(x_k),
$$

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

**Proof.**

$$
x_{k+1} = x_k - t_k d_k,
$$

where $t_k = \frac{d_k^T d_k}{2d_k^T Ad_k}, d_k = 2Ax_k$. 

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

What about constant step size? (Also in HW1)
Proof of Rate of Convergence Contd.

\[ f(x_{k+1}) = x_{k+1}^T A x_{k+1} = (x_k - t_k d_k)^T A (x_k - t_k d_k) \]
\[ = x_k^T A x_k - 2 t_k d_k^T A x_k + t_k^2 d_k^T A d_k \]
\[ = x_k^T A x_k - t_k d_k^T d_k + t_k^2 d_k^T A d_k. \]

Plugging in the expression for \( t_k \)

\[ f(x_{k+1}) = x_k^T A x_k - \frac{1}{4} \frac{(d_k^T d_k)^2}{d_k^T A d_k} \]
\[ = x_k^T A x_k \left( 1 - \frac{1}{4} \frac{(d_k^T d_k)^2}{(d_k^T A d_k)(d_k^T A^{-1} A x_k)} \right) \]
\[ = \left( 1 - \frac{(d_k^T d_k)^2}{(d_k^T A d_k)(d_k^T A^{-1} d_k)} \right) f(x_k). \]

By Kantorovich:

\[ f(x_{k+1}) \leq \left( 1 - \frac{4 Mm}{(M + m)^2} \right) f(x_k) = \left( \frac{M - m}{M + m} \right)^2 f(x_k) = \left( \frac{\kappa(A) - 1}{\kappa(A) + 1} \right)^2 f(x_k), \]
**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

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

- Matrices (or quadratic functions) with large condition number are called **ill-conditioned**.
- Matrices with small condition number are called **well-conditioned**.
- A **large** condition number implies a **large** number of iterations of the gradient method.
- A **small** condition number implies a **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.

\[
\nabla f(x) = \begin{pmatrix} -400x_1(x_2 - x_1^2) - 2(1 - x_1) \\ 200(x_2 - x_1^2) \end{pmatrix},
\]

\[
\nabla^2 f(x) = \begin{pmatrix} -400x_2 + 1200x_1^2 + 2 & -400x_1 \\ -400x_1 & 200 \end{pmatrix}.
\]

\[
\nabla^2 f(1, 1) = \begin{pmatrix} 802 & -400 \\ -400 & 200 \end{pmatrix}
\]

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

\[ Ax = b \]

where \( A \succ 0 \) and we assume that \( x \) is indeed the solution of the system \( (x = A^{-1}b) \).
- Suppose that the right-hand side is perturbed \( b + \Delta b \). What can be said on the solution of the new system \( x + \Delta x \)?
- \( \Delta x = A^{-1} \Delta b \).
- Result (derivation in class):

\[
\frac{\|\Delta x\|}{\|x\|} \leq \kappa(A) \frac{\|\Delta b\|}{\|b\|}
\]
Numerical Example

consider the ill-condition matrix:

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

```matlab
>> A=[1+1e-5,1;1,1+1e-5];
>> cond(A)
ans =
    2.00000999998795e+005
```

We have

```matlab
>> A\[1;1]
ans =
    0.499997500018278
    0.49999750006722
```

However,

```matlab
>> A\[1.1;1]
ans =
    1.0e+003 *
    5.000524997400047
   -4.99475002650021
```
Scaled Gradient Method

Consider the minimization problem

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

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

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

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

$$y_{k+1} = y_k - t_k S^T \nabla f(Sy_k).$$ 

Multiplying the latter equality by $S$ from the left, and using the notation $x_k = Sy_k$:

$$x_{k+1} = x_k - t_k SS^T \nabla f(x_k).$$ 

Defining $D = SS^T$, we obtain the scaled gradient method:

$$x_{k+1} = x_k - t_k D \nabla f(x_k).$$
Scaled Gradient Method

- $D \succ 0$, so the direction $-D \nabla f(x_k)$ is a descent direction:

$$f'(x_k; -D \nabla f(x_k)) = -\nabla f(x_k)^T D \nabla f(x_k) < 0,$$

We also allow different scaling matrices at each iteration.

**Scaled Gradient Method**

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

**Initialization:** pick $x_0 \in \mathbb{R}^n$ arbitrarily.

**General step:** for any $k = 0, 1, 2, \ldots$ execute the following steps:

(a) pick a scaling matrix $D_k \succ 0$.

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

$$g(t) = f(x_k - tD_k \nabla f(x_k)).$$

(c) set $x_{k+1} = x_k - t_k D_k \nabla f(x_k)$.

(c) if $\|\nabla f(x_{k+1})\| \leq \varepsilon$, then STOP and $x_{k+1}$ is the output.
Choosing the Scaling Matrix $D_k$

- The scaled gradient method with scaling matrix $D$ is equivalent to the gradient method employed on the function $g(y) = f(D^{1/2}y)$.
- Note that the gradient and Hessian of $g$ are given by
  \[
  \nabla g(y) = D^{1/2}f(D^{1/2}y) = D^{1/2}f(x),
  \nabla^2 g(y) = D^{1/2}\nabla^2 f(D^{1/2}y)D^{1/2} = D^{1/2}\nabla^2 f(x)D^{1/2}.
  \]
- The objective is usually to pick $D_k$ so as to make $D_k^{1/2}\nabla^2 f(x_k)D_k^{1/2}$ as well-conditioned as possible.
- A well known choice (Newton’s method): $D_k = (\nabla^2 f(x_k))^{-1}$. See HW2
- **diagonal scaling:** $D_k$ is picked to be diagonal. For example,
  \[
  (D_k)_{ii} = \left(\frac{\partial^2 f(x_k)}{\partial x_i^2}\right)^{-1}.
  \]
- Diagonal scaling can be very effective when the decision variables are of different magnitudes.
The Gauss-Newton Method

- Nonlinear least squares problem:

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

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

- Denote:

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

- Then the problem becomes:

\[
\min \|F(x)\|^2.
\]
The Gauss-Newton Method

Given the \( k \)th iterate \( x_k \), the next iterate is chosen to minimize the sum of squares of the linearized terms, that is,

\[
x_{k+1} = \arg\min_{x \in \mathbb{R}^n} \left\{ \sum_{i=1}^{m} \left[ f_i(x_k) + \nabla f_i(x_k)^T (x - x_k) - c_i \right]^2 \right\}.
\]

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

\[
\min \| A_k x - b_k \|^2,
\]

where

\[
A_k = \begin{pmatrix}
\nabla f_1(x_k)^T \\
\nabla f_2(x_k)^T \\
\vdots \\
\nabla f_m(x_k)^T
\end{pmatrix} = J(x_k)
\]

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

\[
b_k = \begin{pmatrix}
\nabla f_1(x_k)^T x_k - f_1(x_k) + c_1 \\
\nabla f_2(x_k)^T x_k - f_2(x_k) + c_2 \\
\vdots \\
\nabla f_m(x_k)^T x_k - f_m(x_k) + c_m
\end{pmatrix} = J(x_k) x_k - F(x_k)
\]
The Gauss-Newton Method

The Gauss-Newton method can thus be written as:

\[ x_{k+1} = (J(x_k)^T J(x_k))^{-1} J(x_k)^T b_k. \]

The gradient of the objective function \( f(x) = \|F(x)\|^2 \) is

\[ \nabla f(x) = 2J(x)^T F(x) \]

The GN method can be rewritten as follows:

\[ x_{k+1} = (J(x_k)^T J(x_k))^{-1} J(x_k)^T (J(x_k)x_k - F(x_k)) \]
\[ = x_k - (J(x_k)^T J(x_k))^{-1} J(x_k)^T F(x_k) \]
\[ = x_k - \frac{1}{2} (J(x_k)^T J(x_k))^{-1} \nabla f(x_k), \]

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

\[ D_k = \frac{1}{2} (J(x_k)^T J(x_k))^{-1}. \]
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:** $\varepsilon$ - tolerance parameter.

**Initialization:** pick $x_0 \in \mathbb{R}^n$ arbitrarily.

**General step:** for any $k = 0, 1, 2, \ldots$ execute the following steps:

(a) Set $d_k = -(J(x_k)^T J(x_k))^{-1} J(x_k)^T F(x_k)$.

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

$$h(t) = g(x_k + t d_k).$$

(c) set $x_{k+1} = x_k + t_k d_k$.

(c) if $\|\nabla f(x_{k+1})\| \leq \varepsilon$, then STOP and $x_{k+1}$ is the output.
Fermat-Weber Problem

**Fermat-Weber Problem:** Given \( m \) points in \( \mathbb{R}^n : a_1, \ldots, a_m \) – also called “anchor point” – and \( m \) weights \( \omega_1, \omega_2, \ldots, \omega_m > 0 \), find a point \( x \in \mathbb{R}^n \) that minimizes the weighted distance of \( x \) to each of the points \( a_1, \ldots, a_m \):

\[
\min_{x \in \mathbb{R}^n} \left\{ f(x) \equiv \sum_{i=1}^{m} \omega_i \|x - a_i\| \right\}.
\]

- The objective function is not differentiable at the anchor points \( a_1, \ldots, a_m \).
- One of the simplest instances of **facility location** problems.
Weiszfeld’s Method (1937)

- Start from the stationarity condition $\nabla f(x) = 0$.

$$\sum_{i=1}^{m} \omega_i \frac{x-a_i}{\|x-a_i\|} = 0.$$  

$$\left(\sum_{i=1}^{m} \frac{\omega_i}{\|x-a_i\|}\right) x = \sum_{i=1}^{m} \frac{\omega_i a_i}{\|x-a_i\|},$$

$$x = \frac{1}{\sum_{i=1}^{m} \omega_i \|x-a_i\|} \sum_{i=1}^{m} \frac{\omega_i a_i}{\|x-a_i\|}.$$  

- The stationarity condition can be written as $x = T(x)$, where $T$ is the operator

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

- Weiszfeld’s method is a fixed point method:

$$x_{k+1} = T(x_k).$$

$^2$We implicitly assume here that $x$ is not an anchor point.
Weiszfeld’s Method as a Gradient Method

Weiszfeld’s Method
Initialization: pick \( x_0 \in \mathbb{R}^n \) such that \( x \neq a_1, a_2, \ldots, a_m \).
General step: for any \( k = 0, 1, 2, \ldots \) compute:

\[
x_{k+1} = T(x_k) = \frac{1}{\sum_{i=1}^{m} \frac{\omega_i}{\|x_k - a_i\|}} \sum_{i=1}^{m} \frac{\omega_i a_i}{\|x_k - a_i\|}.
\]

Weiszfeld’s method is a gradient method since

\[
x_{k+1} = \frac{1}{\sum_{i=1}^{m} \frac{\omega_i}{\|x_k - a_i\|}} \sum_{i=1}^{m} \frac{\omega_i a_i}{\|x_k - a_i\|} = x_k - \frac{1}{\sum_{i=1}^{m} \frac{\omega_i}{\|x_k - a_i\|}} \nabla f(x_k).
\]

A gradient method with a special choice of stepsize: \( t_k = \frac{1}{\sum_{i=1}^{m} \frac{\omega_i}{\|x_k - a_i\|}} \).