Master-MIRI
Topics on Optimization and Machine Learning (TOML)

José M. Barceló Ordinas
Departament d'Arquitectura de Computadors
(UPC)

Descent Methods for unconstrained minimization

- **Objective:** find the optimal point $\mathbf{x^* \in R^n}$, that minimizes an objective function $f_0(x)$ → then the optimal value is $\mathbf{p^*=f_0(x^*)}$

The objective is to produce a minimizing sequence $x^{(k)}$, k=0,... such that:

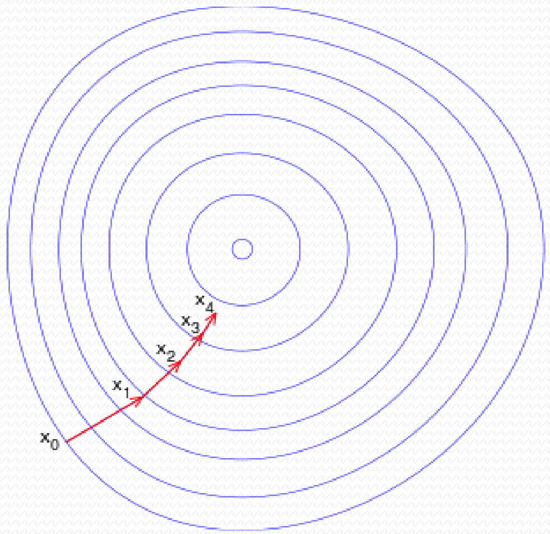
$$\mathbf{x^{(k+1)} = x^{(k)} + t\,\Delta x,}$$

Where:

$\Delta x = d^{(k)} \in R^n$ is the **search direction** (vector) or **step**

$t \geq 0$ is the **step size**

k is the **iteration time**



Descent Methods for unconstrained minimization

The objective is to produce a minimizing sequence $x^{(k)}$, k=0,... such that:

$$\mathbf{x^{(k+1)} = x^{(k)} + t\, \Delta x,}$$

Where:

$\Delta x = d^{(k)} \in R^n$ is the **search direction** (vector) or **step**

$t \geq 0$ is the **step size**

k is the **iteration time**

- Any descent method should satisfy:

$$f(x^{(k+1)}) < f(x^{(k)})$$

except when $x^{(k)}$ is optimal (<u>with certain accuracy:</u> $|x^{(k+1)}-x^{(k)}| \leq \varepsilon$), and then a descent method is characterized by:

$$\nabla f(x^{(k)})^T \Delta x = \nabla f(x^{(k)})^T d^{(k)} < 0$$

i.e., must form an acute angle (<90º) with the negative gradient

- The rate of convergence tell us how fast the method approaches the optimal value. <u>It can diverge</u>.

Descent Methods for unconstrained minimization

General Descent Method:

$$\mathbf{x^{(k+1)} = x^{(k)} + t\,\Delta x,}$$

i. Determine the descent direction $\Delta x = d^{(k)}$,
ii. Line search: Choose the step size t
iii. Update k++, calculate $x^{(k+1)}$, and return to i)

Gradient Descent Method:

$$\mathbf{x^{(k+1)} = x^{(k)} + \rho^{(k)} d^{(k)} = x^{(k)} - t}\nabla f(x^{(k)})$$

i. Choose an initial value $x^{(0)}$
ii. The descent direction $d^{(k)}$ is the **negative gradient**:
$$\mathbf{\Delta x = d^{(k)} = -\nabla f(x^{(k)})}$$
iii. Line search: Choose the step size $t = \boldsymbol{\rho^{(k)}}$
iv. Update k++, calculate $x^{(k+1)}$, and return to i) if stop criterion is not fulfil
v. Stop Criterion: $||\nabla f(x^{(k)})||_2 \leq \varepsilon$ with $\varepsilon > 0$

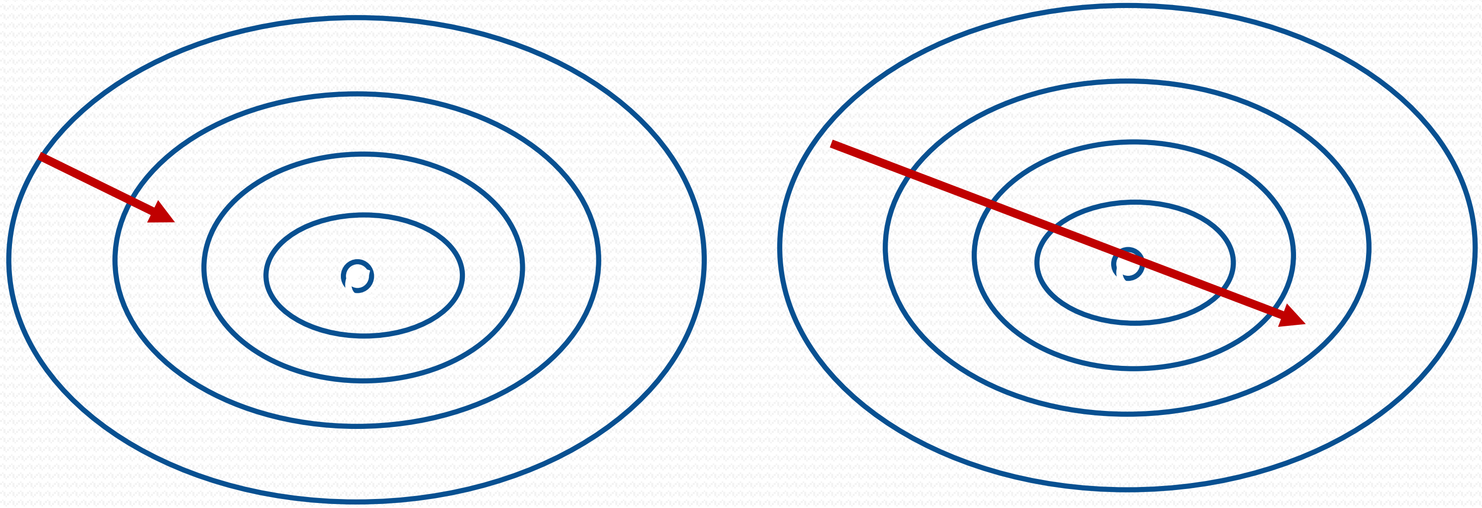
- **Descent Methods for unconstrained minimization**

Selection of the Step Size (Line Search):

i. **Exact Line Search:** choose $t=\rho^{(k)}$ such that minimizes f along the ray $\{x+t\Delta x \mid t \geq 0\}$

$$t = \rho^{(k)} = \operatorname{argmin}_{s>0} f(x^{(k)} + s \cdot d^{(k)})$$

However, sometimes the cost of this optimization problem is high and approximate methods are used (e.g. Backtracking)



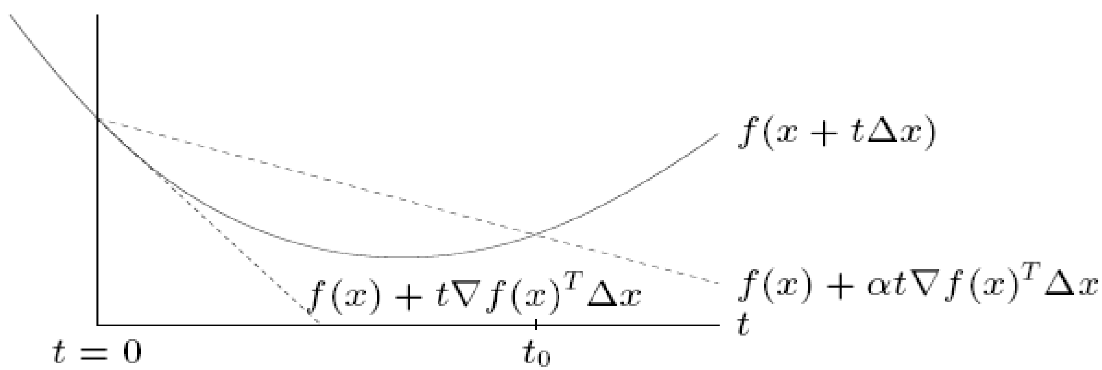
Descent Methods for unconstrained minimization

Selection of the Step Size (Line Search):

i. **Backtracking Line Search (Wolfe Condition)**: choose $t=\rho^{(k)}$ such that approximately minimizes f along the ray $\{x+t\Delta x \mid t\geq 0\}$.

- given a descent direction $\mathbf{\Delta x= d^{(k)} = -\nabla f(x)}$ for f at $x \in$ dom f, $\alpha \in (0,0.5)$ and $\beta \in (0,1)$ and t=1
- **while** $f(x^{(k)}+td^{(k)}) > f(x^{(k)}) + \alpha t \nabla f(x^{(k)})^T d^{(k)}$ (stop criterion) → $t:=\beta t$

- If α is chosen to be between 0.01-0.3, we want to decrease f between 1%-30% of the prediction based on linear interpolation. β is typically chosen 0.1 (very crude search) or 0.8 (less crude search)



Descent Methods for unconstrained minimization

Steepest Descent Method:

$$\mathbf{x^{(k+1)} = x^{(k)} + t\, \Delta x_{sd}}$$

i. Choose an initial value $x^{(0)}$

ii. The descent direction $\mathbf{\Delta x_{sd}}$ makes the **directional derivative** ($\nabla f(x^{(k)})^T v$) as negative as possible

ii. Line search: Choose the step size t using Exact line search or backtracking line search.

iii. Stop Criterion: quit if $||\nabla f(x^{(k)})||_2 \leq \varepsilon$ with $\varepsilon > 0$

iv. Update: $x^{(k+1)} = x^{(k)} + t\, \Delta x$

$\Delta x_{sd} = \text{argmin} \{\nabla f(x^{(k)})^T v \mid \ ||v|| \leq 1 \}$ is the normalized steepest descent direction with respect norm $||\cdot||$.

- If the norm is the Euclidean norm → $\Delta x_{sd} = -\nabla f(x)$ and the steepest descent method is the gradient method.
- Other norms result in other steepest descent methods, e.g. $l_1 = ||\cdot||_1$ produces the **coordinate descent method**

Descent Methods for unconstrained minimization

- The **condition number** measures how much a function f changes in proportion to small changes of the argument x. A problem with small condition number values is said **well-conditioned**, while a problem with large condition numbers is said to be **ill-conditioned**.

K= xf'(x)/f(x) (if 1-dim), K=||x|| ||J||/||f|| if more than 1-dim

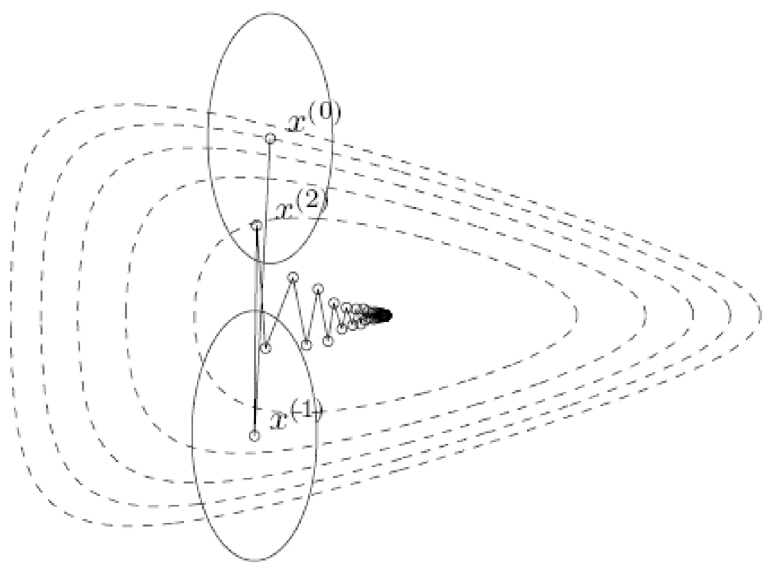


Figure 9.12 Steepest descent method, with quadratic norm $\|\cdot\|_{P_2}$.

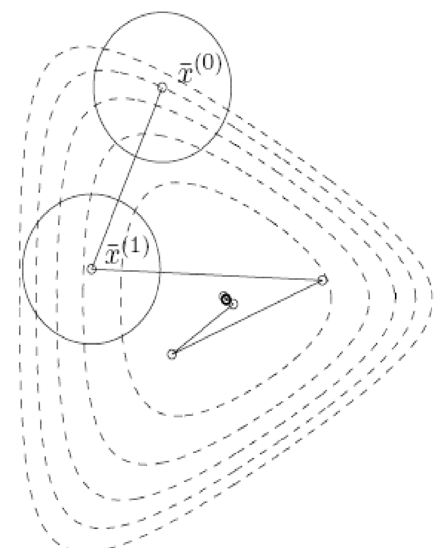


Figure 9.14 The iterates of steepest descent with norm $\|\cdot\|_{P_1}$, after the change of coordinates. This change of coordinates reduces the condition number of the sublevel sets, and so speeds up convergence.

- **Descent Methods for unconstrained minimization**

Newton's Method:

i. Choose an initial value $x^{(0)}$
ii. Define the Newton decrement: $\lambda^2=\nabla f(x^{(k)})^T\nabla^2 f(x^{(k)})^{-1}\nabla f(x^{(k)})$
iii. Define the Newton step: $\Delta x = - \nabla^2 f(x^{(k)})^{-1} \nabla f(x^{(k)})$
iv. Stopping Criterion: quit if $\lambda^2/2\leq \varepsilon$ with $\varepsilon >0$
ii. Line search: choose step size t using backtracking line search
iii. Update: $x^{(k+1)} = x^{(k)}+t\Delta x$

Good estimate when x is near x*, since the Newton step is a minimizer of second-order approximation:

$$f(x+v) = f(x) + \nabla f(x)^T v + ½\, v^T \nabla^2 f(x)\, v$$

Before using Newton's method, the rate of convergence should be checked (it is quadratic) → proof of quadratic convergence

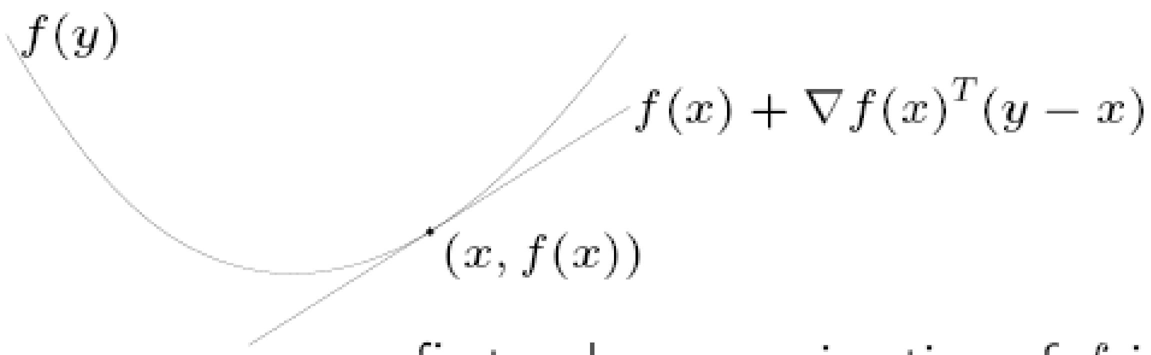
Subgradient methods

We have always assumed that f(x) is differentiable, that is that $\nabla f(x)$ exist for all $x \in X$. What happens if f(x) is not differentiable ?

For example, $f(x) = |x|$, $x \in R$ is not differentiable at x=0, however, the function is convex and the minimum is at x=0, How do we apply the Gradient Descent method if we can not obtain the gradient ?

Remember that the first order condition states for convex differentiable functions that:

> $\mathbf{f(y) \geq f(x)+\nabla f(x)^T(y-x)}$ (first order Taylor approximation of f near x) for all x, $y \in$ dom f (remember topic 12)



first-order approximation of f is global underestimator

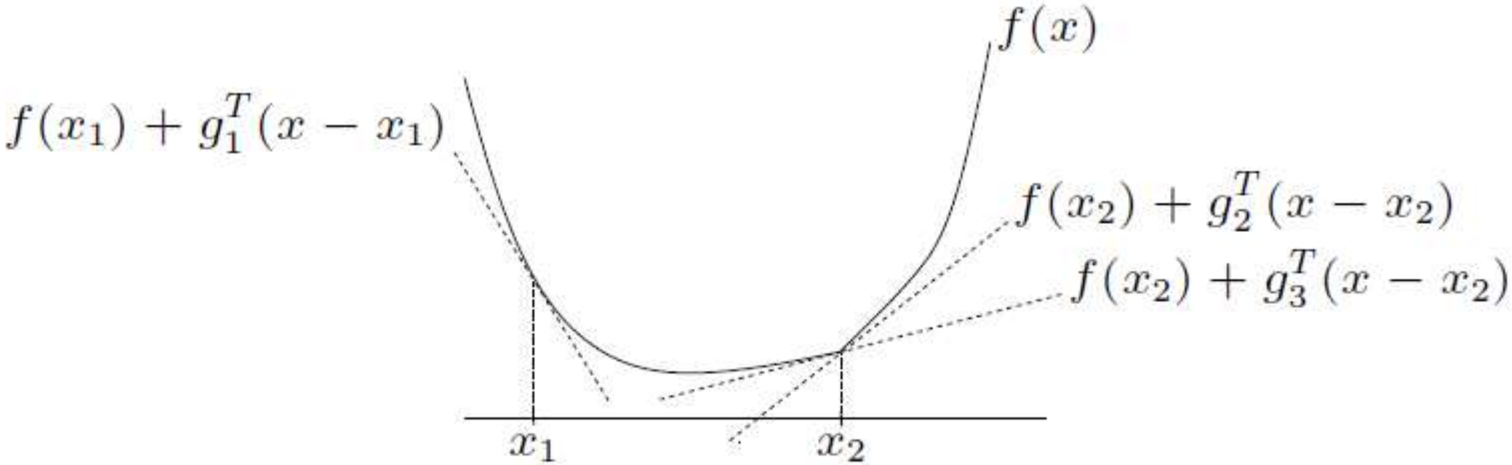
- **Subgradient methods**

The function g is a subgradient of f if:

$$f(y) \geq f(x) + g(x)^T (y-x) \text{ for all } y \in \text{dom } f$$

e.g. g_1, g_2 and g_3 are subgradients of f(x)

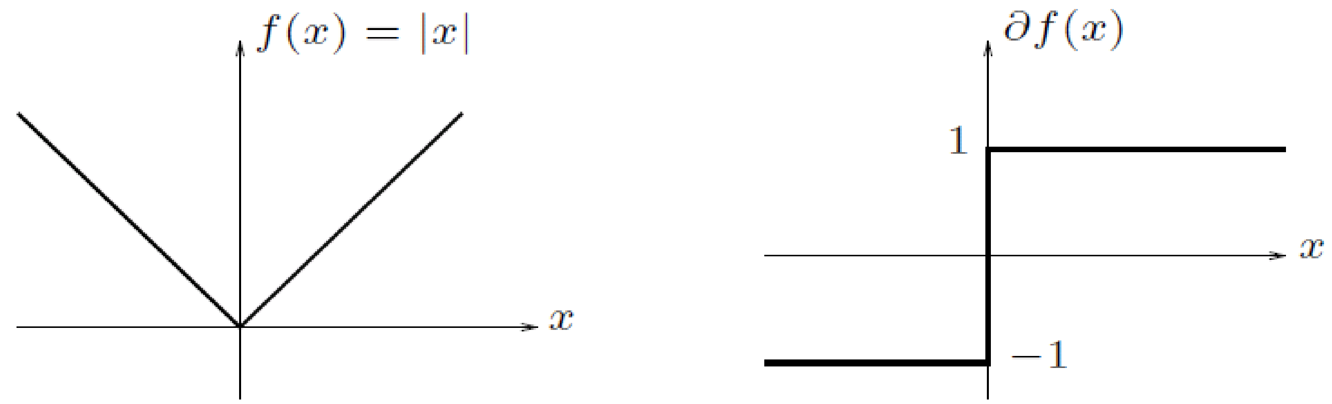
- if f(x) is convex, it has at least one subgradient at every point in the relint of the domain
- if f(x) is convex and differentiable then $\nabla f(x)$ is a subgradient of f at x.



Subgradient methods

The set of all subgradients of f(x) is called the subdifferential of f at x and is denoted as $\partial f(x)$.

e.g. $f(x) = |x|$ → $\partial f(x) = -1$ if $x<0$
$= 1$ if $x>0$



There are a method called "subgradient method" to solve numerically

Subgradient methods

There are descend methods called "subgradient method" to solve iteratively optimization problems in which the function is not differentiable but in which it is possible to find a subgradient.

In this case:

The objective is to produce a minimizing sequence $x^{(k)}$, k=1,... such that:

$$\mathbf{x^{(k+1)} = x^{(k)} - t\, g^{(k)}(x^{(k)})},$$

Where:

$g^{(k)}(x^{(k)})$ is a **subgradient** of **f** at $\mathbf{x^{(k)}}$

$t \geq 0$ is the **step size**

k is the **iteration time**

If $-g^{(k)}$ is not a descent direction of f at $x^{(k)}$, we maintain an f_{best} that keeps track of the lowest objective function value found so far:

$$\mathbf{f^{(k)}_{best} = \{ f^{(k-1)}_{best}, f(x^{(k)})\}}$$

Descent Methods with equality constrains

Let us assume a minimization problem with equality constraints:

minimize $f(x)$

subject to $Ax = b$

and substitute the objective function by its Taylor second-order approximation near x:

minimize $f(x+v)=f(x)+ \nabla f(x)^T v + (½)\, v^T \nabla^2 f(x)\, v = r + q^T v + (½)\, v^T P v$

subject to $A(x+v) = b \rightarrow Ax+Av=b \rightarrow Av=0$

var x, v

quadratic COP that can be solved analytically.

Define Δx_{nt}, **the Newton Step at x**, as the solution of the former COP, it is to say the increment to x to solve the problem when the quadratic approximation is used in place of f.

Descent Methods with equality constrains

The KKT conditions (remember last slide of topic 14) for quadratic problems with equality constraint is:

minimize **$(1/2)\, x^T P x + q^T x + r$ → $f(x) + \nabla f(x)^T w + (½)\, w^T \nabla^2 f(x)\, w$**

subject to **$Ax=b$ → $Ax + Aw = b$**

where Δx_{nt} is the step and w is the associated optimal variable of the dual problem

$$\begin{bmatrix} P & A^T \\ A & 0 \end{bmatrix}\begin{bmatrix} x^* \\ \upsilon^* \end{bmatrix}=\begin{bmatrix} -q \\ b \end{bmatrix} \longrightarrow \begin{bmatrix} \nabla^2 f(x) & A^T \\ A & 0 \end{bmatrix}\begin{bmatrix} \Delta x_{nt} \\ w \end{bmatrix}=\begin{bmatrix} -\nabla f(x) \\ 0 \end{bmatrix}$$

Newton's Method with equality constraints:

i. Choose an initial value $x^{(0)}$, such that $x^{(0)} \in \mathrm{dom}\, f$ with $Ax=b$ and choose tolerance $\varepsilon > 0$

ii. Compute the Newton Δx_{nt} step and decrement

For Δx_{nt} solve the above linear equation system

$\lambda^2 = \nabla f(x^{(k)})^T\ \nabla^2 f(x^{(k)})^{-1}\ \nabla f(x^{(k)})$

i. Stopping Criterion: quit if $\lambda^2/2 \leq \varepsilon$ with $\varepsilon > 0$

ii. Line search: choose step size t using backtracking line search

iii. Update: $x^{(k+1)} = x^{(k)} + t\, \Delta x_{nt}$

Interior-Point Methods for COP

$$\begin{array}{ll} \textbf{minimize} & f_0(x) \\ \textbf{subject to} & f_i(x) \leq 0 \quad i=1,\ldots,m \\ & Ax = b \end{array}$$

That satisfies KKT conditions:

i. **Primal constraints:** $f_i(x^*) \leq 0, \quad i=1,\ldots,m$
ii. **Primal constraints:** $Ax^* = b,$
iii. **Dual constraints:** $\lambda_i^* \geq 0 \quad i=1,\ldots,m$
iv. **Complementary slackness:** $\lambda_i^* f_i(x^*)=0 \quad i=1,\ldots,m$
v. **Gradient of Lagrangian vanishes:**

$$\nabla_x L(x,\lambda^*,\nu^*) = \nabla f_0(x^*) + \Sigma_{i=1,..m} \lambda^*_i \nabla f_i(x^*) + A^T\nu^* = 0$$

Interior-point methods solve the COP problem (or KKT problem) by applying Newton's method to a sequence of equality constraint problems or sequence of modified versions of the KKT conditions.

Interior-Point Methods – Logarithmic barrier

Objective: formulate the inequality constrained COP as an equality constrained problem to which Newton's method can be applied.

minimize $f_0(x)$

subject to $f_i(x) \leq 0 \quad i=1,\ldots,m$

$Ax = b$

Can be re-written as:

minimize $f_0(x) + \Sigma_{i=1,..m} I\_(f_i(x))$

subject to $Ax = b$

Where:

$$I\_(u) = \begin{cases} 0 & u \leq 0 \\ \infty & u > 0 \end{cases}$$

Whose objective function is not differentiable (Newton's method can not be applied)

Interior-Point Methods – Logarithmic barrier

Approximate the indicator function I_- as:

$$I_-(u) = -(1/t)\log(-u) \quad \text{with dom } I_- = -R_{++}$$

where t>0 sets the accuracy of the approximation and we note that I_- is convex and non-decreasing and takes value ∞ if u>0.

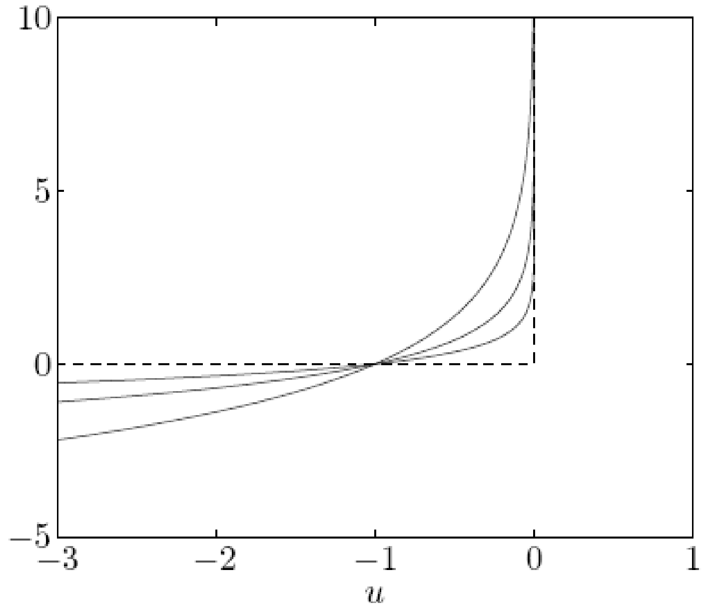


Figure 11.1 The dashed lines show the function $I_-(u)$, and the solid curves show $\widehat{I}_-(u) = -(1/t)\log(-u)$, for $t = 0.5,\ 1,\ 2$. The curve for $t = 2$ gives the best approximation.

- **Interior-Point Methods – Logarithmic barrier**

Since:

$$I\_(u) = -(1/t)\log(-u) \quad \text{with dom } I\_ = -R_{++}$$

Then, now:

$$\textbf{minimize} \quad f_0(x) + \Sigma_{i=1,..m} I\_(f_i(x))$$
$$\textbf{subject to} \quad Ax = b$$

can be re-written as:

$$\textbf{minimize} \quad f_0(x) + \Sigma_{i=1,..m} -(1/t)\log(-f_i(x))$$
$$\textbf{subject to} \quad Ax = b$$

and Newton's method can be applied since the objective function is differentiable

Interior-Point Methods – Logarithmic barrier

The function **$\phi(x) = -\Sigma_{i=1,..m} \log(-f_i(x))$** is called **Log barrier**.

- As t grows, the approximation improves
- As t grows, $f_0 + (1/t)\ \phi(x)$ is difficult to minimize using Newton's method

The gradient and Hessian of $\phi(x)$ are:

$$\nabla \phi(x) = \sum_{i=1}^{m} \frac{-1}{f_i(x)} \nabla f_i(x)$$

$$\nabla^2 \phi(x) = \sum_{i=1}^{m} \frac{1}{f_i(x)^2} \nabla f_i(x) \nabla f_i(x)^T + \sum_{i=1}^{m} \frac{-1}{f_i(x)} \nabla^2 f_i(x)$$

Interior-Point Methods – Central Path

The **central path** is defined as the set of **central points** $x^*(t)$, $t>0$, that solve the minimization problem:

minimize $\quad f_0(x) + (1/t)\,\phi(x)$
subject to $\quad Ax=b$

The **central point x*(t)** is characterized by being strictly feasible, it is to say:

$Ax^*(t)=b$

$f_i(x^*(t)) < 0$

The **Lagrangian** is $L(x,v) = f_0(x) + 1/t\,\phi(x) + v^T(Ax-b)$, and

The **Lagrange function** is $q(x,v) = \min_{x\in X} L(x,v) = \min_{x\in X} \{ f_0(x) + 1/t\,\phi(x) + v^T(Ax-b) \}$

We say that the **centrality condition** holds if (calculate the gradient of the Lagrangian with respect x):

$$
\begin{aligned}
0 &= \nabla f_0(x^*(t)) + 1/t\,\nabla\phi(x^*(t)) + A^T v \\
&= \nabla f_0(x^*(t)) + \Sigma_{i=1,..m} (-1)/(t\, f_i(x^*(t))\ \nabla f_i(x^*(t)) + A^T v = \\
&= \nabla f_0(x^*(t)) + \Sigma_{i=1,..m} \lambda_i \nabla f_i(x^*(t)) + A^T v
\end{aligned}
$$

Interior-Point Methods – Central Path

Observe that we can interpret the former equation as coming from a Langragian such as: $L(x,\lambda,\nu) = f_0(x) + \Sigma_{i=1,..m} \lambda_i f_i(x) + \nu^T(Ax-b)$ where $\lambda^*(t)=-1/(tf_i(x^*(t))$

Then, $(\lambda^*(t),\nu^*(t))$ are dual feasible and the dual function $q(\lambda^*(t),\nu^*(t))$ is:

$$q(\lambda^*(t),\nu^*(t)) = f_0(x^*(t)) + \Sigma_{i=1,..m} \lambda_i^*(t) f_i(x^*(t)) + \nu^*(t)^T(Ax^*(t)-b)$$

$$= f_0(x^*(t)) + \Sigma_{i=1,..m} 1/t = f_0(x^*(t)) - m/t$$

where m is the number of inequalities, i=1,..., m

Finally, the duality gap tell us that:

$q(\lambda^*(t),\nu^*(t)) = d^* \leq p^* \rightarrow f_0(x^*(t)) - m/t - p^* \leq 0 \rightarrow$ $\mathbf{f_0(x^*(t)) - p^* \leq m/t}$

which tell us that **x*(t) converges to p* as t → ∞**

Interior-Point Methods – Central Path

KKT interpretation: We can also interpret the central path conditions as a continuous deformation of the KKT optimality conditions: a point x is equal to $x^*(t)$ if and only if there exists λ, v such that:

i. **Primal constraints:** $f_i(x) \leq 0$, i=1,...,m
ii. **Primal constraints:** $Ax = b$
iii. **Dual constraints:** $\lambda_i \geq 0$ i=1,...,m
iv. **Complementary slackness:** $-\lambda_i f_i(x) = 1/t$ i=1,...,m
v. **Gradient of Lagrangian vanishes:**

$$\nabla_x L(x,\lambda,\nu) = \nabla f_0(x) + \Sigma_{i=1,..m} \lambda_i \nabla f_i(x) + A^T v = 0$$

Note that the only difference is in the slackness condition in which $-\lambda_i f_i(x)=1/t$ instead of $\lambda_i f_i(x) = 0$.

In fact as $\mathbf{t \rightarrow \infty, \lambda_i f_i(x) \rightarrow 0}$, for all i=1,…,m and $\lambda(t)$ and v(t) **almost satisfy** the KKT conditions.

Interior-Point Methods – The Barrier Method

The Barrier Method SUMT (Sequential Unconstrained Minimization Technique)

Given strictly feasible x, $t=t^{(0)}$, $\mu>1$, tolerance $\varepsilon>0$

i. **Centering step** or **outer iteration**:
Compute x*(t) by minimizing $f_0+1/t\ \phi$, subject to Ax=b, starting at x.

ii. **Update:** x=x(t)

iii. **Stopping criterion:** quit if $m/t<\varepsilon$ (m/t is the duality gap)

iv. **Increase t** as $t=\mu t$

At each step, we compute the central point x*(t) starting from the previous computed central point. The algorithm also computes λ*(t) and ν*(t)

We refer to the Newton iterations or steps executed during the centering step as **inner iterations**. At each inner step, we have a primal feasible point; but we have a dual feasible point only at the end of each outer (centering) step.