Outline

1 Optimization without constraints
   - Optimization scheme
   - Linear search methods
   - Gradient descent
   - Conjugate gradient
   - Newton method
   - Quasi-Newton methods

2 Optimization under constraints
   - Lagrange
   - Equality constraints
   - Inequality constraints
   - Dual problem - Resolution by duality
   - Numerical methods
     - Penalty functions
     - Projected gradient: equality constraints
     - Projected gradient: inequality constraints

3 Conclusion
Outline

1. Optimization without constraints
   - Optimization scheme
   - Linear search methods
   - Gradient descent
   - Conjugate gradient
   - Newton method
   - Quasi-Newton methods

2. Optimization under constraints
   - Lagrange
   - Equality constraints
   - Inequality constraints
   - Dual problem - Resolution by duality
   - Numerical methods
     - Penalty functions
     - Projected gradient: equality constraints
     - Projected gradient: inequality constraints

3. Conclusion
General numerical optimization scheme

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

1. **Init**: by \(x^0\), an initial guess of a minimum \(x^*\).
2. **Recursion**: until a convergence criterion is satisfied at \(x^n\)
   - at \(x^n\), determine of a search direction \(d^n \in \mathbb{R}^d\),
   - **linear search**: find \(x^{n+1}\) along the semi-line \(x^n + t d^n\), \(t \in \mathbb{R}^+\); amounts to minimizing \(\phi(t) = f(x^n + t d^n)\) in \(t > 0\).
Remarks:

- It is important to notice that $f$ can’t be plotted ($d$ large). An optimization scheme is short-sighted i.e. only has access to local knowledge on $f$.
- The information available will determine the method to use:
  - order 0: only $f(x^n)$ available,
  - 1st order: $\nabla f(x^n)$ also known
  - 2nd order: $\nabla^2 f(x^n)$ known (or estimated)
- Stop criteria are on $\|\nabla f(x^n)\|$, on the relative norm of the last step $\frac{\|x^{n+1} - x^n\|}{\|x^n\|}$, etc.
- The linear search may simply be an approximate minimization.
Linear search methods

Linear search

We look for a local minimum of $\phi$:

$$\min_{t>0} \phi(t) = f(x^n + t d^n)$$

Equivalently, and to simplify notations, we can assume that $f : \mathbb{R} \to \mathbb{R}$ and look for a minimum of $f$.

There exist many methods, according to the assumptions on $f$: convex, unimodular, $C^1$, $C^2$, etc.
Newton-Raphson method

Assumes $f$ is $C^2$. Principle:

- approximate $f$ by its second order expansion around $x^n$, 
  \[ f(x) = f(x^n) + f'(x^n)(x - x^n) + \frac{1}{2} f''(x^n)(x - x^n)^2 + o(x - x^n)^2 \]
- take as $x^{n+1}$ the min of the quadratic approx. of $f$.

\[ x^{n+1} = x^n - \frac{f'(x^n)}{f''(x^n)} \]
Equivalently, amounts to finding a zero of $f'(x)$.

$$f'(x) = f'(x^n) + f''(x^n)(x - x^n) + o(x - x^n)$$

$$x^{n+1} = x^n - \frac{f'(x^n)}{f''(x^n)}$$
Secant method

Assumes $f$ is only $C^1$.

Principle:

- same as the Newton-Raphson method, but $f''(x^n)$ is approximated by $\frac{f'(x^n) - f'(x^{n-1})}{x^n - x^{n-1}}$
- this yields $x^{n+1} = x^n - \frac{x^n - x^{n-1}}{f'(x^n) - f'(x^{n-1})} f'(x^n)$
- Standard to find the zero of a function when its derivative is unknown.
Wolfe’s method

Assumes $f$ is only $C^1$.

Principle:
- approximate linear search
- proposes an $x^{n+1}$ that “sufficiently” decreases $|f'(x^{n+1})|$ w.r.t. $|f'(x^n)|$
- and that also significantly decreases $f(x^{n+1})$ w.r.t. $f(x^n)$
- The search of $x^{n+1}$ is done by dichotomy in an interval $[a = x^n, b]$. 
Let $0 < m_1 < \frac{1}{2} < m_2 < 1$ be two parameters. The point $x$ is acceptable as $x^{n+1}$ iff

$$f(x) \leq f(x^0) + m_1(x - x^0)f'(x_0)$$
$$f'(x) \geq m_2 f'(x^0)$$
Gradient descent (steepest descent)

Back to the general case: \( f : \mathbb{R}^d \rightarrow \mathbb{R} \)

Principle:
- Performs the linear search along the steepest descent direction

\[
d^n = -\nabla f(x^n)
\]

- the optimal step \( t^* \) minimizes \( \phi(t) = f(x^n + t \, d^n) \), so

\[
\phi'(t^*) = \nabla f(x^n + t^* \, d^n)^t \, d^n = 0
\]

- the descent stops when the new gradient \( \nabla f(x^{n+1}) \) becomes orthogonal to the current descent direction \( d^n = -\nabla f(x^n) \).
Gradient descent

Properties:

- Easy to implement, requires only first order information on $f$.
- Slow convergence. Performs poorly even on simple functions like quadratic forms!
- In practice, a fast suboptimal descent step is preferred.
Example: on "Rosenbrock’s banana"

\[ f(x) = 100(x_2 - x_1^2)^2 + (1 - x_1)^2 \]
Conjugate gradient

Overview:

- A first order method, simple variation of the gradient descent,
- designed to perform well on quadratic forms.
- Idea = tilt the next search direction to better aim at the minimum of the quadratic form.
We assume $A$ is a positive symmetric matrix and

$$f(x) = \frac{1}{2} x^t A x + b^t x$$

and $$\nabla f(x) = A x + b$$

Principle:
- start at $x^0$, $d^0 = -\nabla f(x^0) \triangleq -g^0$,
- at $x^n$, instead of $d^n = -\nabla f(x^n) \triangleq -g^n$,
  look for the minimum of $f$ in the affine space

$$\mathcal{W}_{n+1} = x^0 + \text{sp}\{d^0, d^1, \ldots, d^{n-1}, g^n\}$$

Lemma

$x^{n+1}$ is the min of $f$ in $\mathcal{W}_{n+1} \Rightarrow g^{n+1} \triangleq \nabla f(x^{n+1}) \perp \mathcal{W}_{n+1}$
We assume $A$ is a positive symmetric matrix and

$$f(x) = \frac{1}{2} x^t A x + b^t x \quad \text{and} \quad \nabla f(x) = A x + b$$

Principle:

- start at $x^0$, $d^0 = -\nabla f(x^0) \triangleq -g^0$,
- at $x^n$, instead of $d^n = -\nabla f(x^n) \triangleq -g^n$,

look for the minimum of $f$ in the affine space

$$\mathcal{W}_{n+1} = x^0 + sp\{d^0, d^1, ..., d^{n-1}, g^n\}$$

Lemma

$x^{n+1}$ is the min of $f$ in $\mathcal{W}_{n+1}$ \implies \quad g^{n+1} \triangleq \nabla f(x^{n+1}) \perp \mathcal{W}_{n+1}$
Lemma

$x^n$ is the minimum of $f$ in $\mathcal{W}_n$, from $x^n$, direction $d^n$ points to the minimum $x^{n+1}$ in $\mathcal{W}_{n+1}$ iff

$$(d^n)^t A d^i = 0 \quad \text{for} \quad 0 \leq i \leq n - 1$$

The direction $d^n$ is said to be conjugate to all the previous $d^i$.

Proof:

$$x^{n+1} = x^n + t d^n$$
$$g^{n+1} \triangleq \nabla f(x^{n+1}) = A x^{n+1} + b = g^n + t A d^n$$

From the previous lemma $g^{n+1} \perp \mathcal{W}_{n+1}$ and $g^n \perp \mathcal{W}_n$, so

$$(g^{n+1})^t g^n = \|g^n\|^2 + t (d^n)^t A g^n = 0 \quad \Rightarrow \quad t \neq 0$$
$$(g^{n+1})^t d^i = (g^n)^t d^i + t (d^n)^t A d^i = 0 \quad \text{for} \quad 0 \leq i \leq n - 1$$
Lemma

$x^n$ is the minimum of $f$ in $\mathcal{W}_n$, from $x^n$, direction $d^n$ points to the minimum $x^{n+1}$ in $\mathcal{W}_{n+1}$ iff

$$(d^n)^t A d^i = 0 \quad \text{for} \quad 0 \leq i \leq n - 1$$

The direction $d^n$ is said to be conjugate to all the previous $d^i$.

Proof:

\[ x^{n+1} = x^n + td^n \]

\[ g^{n+1} \triangleq \nabla f(x^{n+1}) = Ax^{n+1} + b = g^n + tAd^n \]

From the previous lemma $g^{n+1} \perp \mathcal{W}_{n+1}$ and $g^n \perp \mathcal{W}_n$, so

\[ (g^{n+1})^t g^n = \|g^n\|^2 + t(d^n)^t A g^n = 0 \quad \Rightarrow \quad t \neq 0 \]

\[ (g^{n+1})^t d^i = (g^n)^t d^i + t(d^n)^t A d^i = 0 \quad \text{for} \quad 0 \leq i \leq n - 1 \]
**Question**: How to find direction $d^n$, conjugate to all previous $d^i$?

Notice $g^{i+1} - g^i = A(x^{i+1} - x^i) \propto A d^i$, so

$$(d^n)^t A d^i = 0 \Rightarrow (d^n)^t g^{i+1} = (d^n)^t g^i = \text{cst}$$

Since the $g^i$ form an orthogonal family, one has

$$d^n \propto \sum_{i=0}^{n} \frac{g^i}{\|g^i\|^2} \Rightarrow d^n = -g^n + c_n d^{n-1}$$

**Answer**: steepest slope, slightly corrected by previous descent direction.
Example: on Rosenbrock’s banana
Expressions of the correction coefficient $c_n$:

- $c_n = \frac{\|g^n\|^2}{\|g^{n-1}\|^2}$  \quad \text{Fletcher & Reeves (1964)}$

- $c_n = \frac{(g^n - g^{n-1})^t g^n}{\|g^{n-1}\|^2}$ \quad \text{Polak & Ribi`ere (1971)}$

- $c_n = \frac{(g^n)^t A d^{n-1}}{\|d^{n-1}\|^2_A}$

Properties:

- converges in $d$ steps for a quadratic form $f : \mathbb{R}^d \to \mathbb{R}$
- same complexity as the gradient method!
- Works well on non quadratic forms if the Hessian doesn’t change much between $x^n$ and $x^{n+1}$
- Caution: $d^n$ may not be a descent direction... In this case, reset to $-g^n$. 
Expressions of the correction coefficient $c_n$:

- $c_n = \frac{\|g^n\|^2}{\|g^{n-1}\|^2}$  \hspace{1em} Fletcher & Reeves (1964)
- $c_n = \frac{(g^n - g^{n-1})^t g^n}{\|g^{n-1}\|^2}$  \hspace{1em} Polak & Ribi` ere (1971)
- $c_n = \frac{(g^n)^t A d^{n-1}}{\|d^{n-1}\|^2_A}$

Properties:

- converges in $d$ steps for a quadratic form $f : \mathbb{R}^d \rightarrow \mathbb{R}$
- same complexity as the gradient method!
- Works well on non quadratic forms if the Hessian doesn’t change much between $x^n$ and $x^{n+1}$
- Caution: $d^n$ may not be a descent direction... In this case, reset to $-g^n$. 
Newton method

Principle:
- Replace $f$ by its second order approximation at $x^n$

$$
\phi(x) = f(x^n) + \nabla f(x^n)^t (x - x^n) + \frac{1}{2} (x - x^n)^t \nabla^2 f(x^n) (x - x^n)
$$

- take as $x^{n+1}$ the min of $\phi(x)$

$$
\nabla \phi(x) = \nabla f(x^n) + \nabla^2 f(x^n) (x - x^n)
$$

which amounts to solving the linear system

$$
\nabla^2 f(x^n) (x^{n+1} - x^n) = -\nabla f(x^n)
$$
Example: on Rosenbrock’s banana
Comments:

- + faster convergence (1 step for quadratic functions!), but expensive: requires second order information on $f$
- yields a stationary point of $f$: one still has to check that it is a minimum
- in practice, try $d^n = -[\nabla^2 f(x^n)]^{-1}\nabla f(x^n)$ as descent direction, and perform a linear search
- - no guarantee that $d^n$ is an admissible descent direction...
- - no guarantee that $x^{n+1}$ is a better point than $x^n$...
- - $\nabla^2 f(x^n)$ may be singular, or badly conditioned...
- the Levenberg-Marquardt regularization suggests to solve
  \[
  [\nabla^2 f(x^n) + \mu \mathbf{1}] \ d^n = -\nabla f(x^n)
  \]
Quasi-Newton methods

Principle:

- Take advantage of the efficiency of the Newton method...
- ... when the Hessian $\nabla^2 f(x)$ is unavailable!
- **Idea:** *approximate* $[\nabla^2 f(x^n)]^{-1}$ by matrix $K_n$ in

\[
    x^{n+1} = x^n - [\nabla^2 f(x^n)]^{-1} \nabla f(x^n)
\]

- More precisely, explore direction $d^n = -K_n \nabla f(x^n)$ from $x^n$. 
Consider the second order Taylor expansion of $f$ at $x^n$

$$f(x) = f(x^n) + \nabla f(x^n)^t (x - x^n)$$
$$+ \frac{1}{2}(x - x^n)^t \nabla^2 f(x^n) (x - x^n) + o(\|x - x^n\|^2)$$

$$\nabla f(x) = \nabla f(x^n) + \nabla^2 f(x^n) (x - x^n) + o(\|x - x^n\|)$$

The estimate $K_n$ of the inverse Hessian must satisfy the quasi-Newton equation (QNE)

$$x^{n+1} - x^n = K_{n+1} [\nabla f(x^{n+1}) - \nabla f(x^n)]$$

Notice that this should be $K_n$... but $K_n$ is used to find $x^{n+1}$, so we impose the relation be satisfied at the next step.
Quasi-Newton methods

Quasi-Newton equation

Consider the second order Taylor expansion of $f$ at $x^n$

$$f(x) = f(x^n) + \nabla f(x^n)^t (x - x^n) + \frac{1}{2} (x - x^n)^t \nabla^2 f(x^n) (x - x^n) + o(\|x - x^n\|^2)$$

$$\nabla f(x) = \nabla f(x^n) + \nabla^2 f(x^n) (x - x^n) + o(\|x - x^n\|)$$

The estimate $K_n$ of the inverse Hessian must satisfy the quasi-Newton equation (QNE)

$$x^{n+1} - x^n = K_{n+1} [\nabla f(x^{n+1}) - \nabla f(x^n)]$$

Notice that this should be $K_n$... but $K_n$ is used to find $x^{n+1}$, so we impose the relation be satisfied at the next step.
All quasi-Newton Methods recursively build the $K_n$ by

$$K_{n+1} = K_n + C_n$$

where the correction $C_n$ is adjusted to satisfy the QNE.

Notations:

$$u^n = x^n - x^{n-1}$$
$$v^n = g^n - g^{n-1}$$

QNE:

$$u^{n+1} = K_{n+1} v^{n+1}$$
Correspondence $C_n$ of rank 1

$$K_{n+1} = K_n + \frac{w^n(w^n)^t}{(w^n)^t v^{n+1}}$$

where $w^n = u^{n+1} - K_n v^{n+1}$

- If initialized with $K_0 = I$, $K_n$ converges in $d$ steps to the true $A^{-1}$ for a quadratic form.

**DFP** (Davidon, Fletcher, Powell) correction of rank 2

$$K_{n+1} = K_n + \frac{u^{n+1}(u^{n+1})^t}{(u^{n+1})^t v^{n+1}} - \frac{K_n v^{n+1}(v^{n+1})^t K_n}{(v^{n+1})^t K_n v^{n+1}}$$

- Converges in $d$ steps to the true $A^{-1}$ for a quadratic form.
- Descent directions are conjugate w.r.t. $A$.
- Coincides with the conjugate gradient method.
Correction $C_n$ of rank 1

\[ K_{n+1} = K_n + \frac{w^n (w^n)^t}{(w^n)^t v^{n+1}} \]

where \( w^n = u^{n+1} - K_n v^{n+1} \)

- If initialized with $K_0 = I$, $K_n$ converges in $d$ steps to the true $A^{-1}$ for a quadratic form.

DFP (Davidon, Fletcher, Powell) correction of rank 2

\[ K_{n+1} = K_n + \frac{u^{n+1} (u^{n+1})^t}{(u^{n+1})^t v^{n+1}} - \frac{K_n v^{n+1} (v^{n+1})^t K_n}{(v^{n+1})^t K_n v^{n+1}} \]

- Converges in $d$ steps to the true $A^{-1}$ for a quadratic form.
- Descent directions are conjugate w.r.t. $A$.
- Coincides with the conjugate gradient method.
- **BFGS** (Broyden, Fletcher, Goldfarb, Shanno, 1970), correction of rank 3

\[
K_{n+1} = K_n - \frac{u^{n+1}(v^{n+1})^t K_n + K_n v^{n+1}(u^{n+1})^t}{(u^{n+1})^t v^{n+1}} \\
+ \left(1 + \frac{(v^{n+1})^t K_n v^{n+1}}{(u^{n+1})^t v^{n+1}}\right) \frac{u^{n+1}(u^{n+1})^t}{(u^{n+1})^t v^{n+1}}
\]

Considered as the best Quasi-Newton method.

- In practice, one should check that \(-K_n g^n\) is a descent direction, *i.e.* \(-(g^n)^t K_n g^n < 0\), otherwise reinitialize by \(K_n = I\).
**BFGS** (Broyden, Fletcher, Goldfarb, Shanno, 1970), correction of rank 3

\[ K_{n+1} = K_n - \frac{u^{n+1}(v^{n+1})^t K_n + K_n v^{n+1}(u^{n+1})^t}{(u^{n+1})^t v^{n+1}} \]

\[ + \left( 1 + \frac{(v^{n+1})^t K_n v^{n+1}}{(u^{n+1})^t v^{n+1}} \right) \frac{u^{n+1}(u^{n+1})^t}{(u^{n+1})^t v^{n+1}} \]

Considered as the best Quasi-Newton method.

In practice, one should check that \(-K_n g^n\) is a descent direction, *i.e.* \(-(g^n)^t K_n g^n < 0\), otherwise reinitialize by \(K_n = I\).
Example: on Rosenbrock’s banana
Outline

1. Optimization without constraints
   - Optimization scheme
   - Linear search methods
   - Gradient descent
   - Conjugate gradient
   - Newton method
   - Quasi-Newton methods

2. Optimization under constraints
   - Lagrange
   - Equality constraints
   - Inequality constraints
   - Dual problem - Resolution by duality
   - Numerical methods
     - Penalty functions
     - Projected gradient: equality constraints
     - Projected gradient: inequality constraints

3. Conclusion
Joseph Louis count of Lagrange

- *Giuseppe Ludovico di Lagrangia*, Italian mathematician, born in Turin (1736)
- founder of the Academy of Turin (1758)
- called by Euler to the Academy of Berlin
- director of the French Academy of Sciences (1788)
- survived the French revolution (b.c.w. Condorcet, ...)
- resting at the Pantheon (1813)
Among his contributions:

- the calculus of variations,
- the Taylor-Lagrange formula,
- the least action principle in mechanics,
- some results on the 3 bodies problem, (the Lagrange points)
- ...
- and the notion of Lagrangian!
Equality constraints

\[
\min_x f(x) \quad \text{s.t.} \quad \theta_j(x) = 0, \quad 1 \leq j \leq m
\]

\(\mathcal{D} : \quad \theta(x) = [\theta_1(x), \ldots, \theta_m(x)]^t = 0\)

defines a manifold of dimension \(d - m\) in \(\mathbb{R}^d\)

\(\nabla \theta_j(x^0)^t (x - x^0) = 0\) : tangent hyperplane to \(\theta_j(x) = 0\) at point \(x^0\)

\(\nabla \theta(x^0)^t (x - x^0) = 0\) : tangent space to \(\mathcal{D}\) at \(x^0\)
Definition

In domain $\mathcal{D} = \{ x \in \mathbb{R}^d : \theta(x) = 0 \}$, the point $x^0$ is regular iff the gradients $\nabla \theta_j(x^0)$ of the $m$ constraints are linearly independent.

Lemma

If $x^0$ is regular, every (unit) direction $d$ in the tangent space is admissible, i.e. can be obtained as the limit of $\frac{x^n-x^0}{\|x^n-x^0\|}$, with $\lim_{n} x^n = x^0$ and $x^n \in \mathcal{D}$.
Theorem

Let $x^*$ be a regular point of $\mathcal{D}$, if $x^*$ is a local extremum of $f$ in $\mathcal{D}$, then there exists a unique vector $\lambda^* \in \mathbb{R}^m$ of Lagrange multipliers such that

$$\nabla f(x^*) + \sum_{j=1}^{m} \lambda_j^* \nabla \theta_j(x^*) = 0$$
Proof:

- Project $\nabla f(x^*)$ on $sp\{\nabla \theta_1(x^*), ..., \nabla \theta_m(x^*)\}$

$$\nabla f(x^*) = \sum_{j=1}^{m} -\lambda_j^* \nabla \theta_j(x^*) + u$$

- $u$ belongs to the tangent space to $D$ at $x^*$
- progressing along $-u$ decreases $f$ and doesn’t change $\theta$
To solve $\min_x f(x)$ s.t. $\theta(x) = 0$,

1. build the Lagrangian

$$L(x, \lambda) = f(x) + \sum_j \lambda_j \theta_j(x)$$

2. find a stationary point $(x^*, \lambda^*)$ of the Lagrangian, i.e. a zero of $\nabla L(x, \lambda)$

$$\nabla_x L(x, \lambda) = \nabla f(x) + \sum_j \lambda_j \nabla \theta_j(x)$$
$$\nabla_\lambda L(x, \lambda) = \theta(x)$$

i.e. $d + m$ (non-linear) equations, with $d + m$ unknowns.
**Example**

**Problem:** find the radius $x_1$ and the height $x_2$ of a cooking pan in order to minimize its surface, s.t. the capacity of the pan is 1 litre.

\[
\begin{align*}
  f(x) &= \pi x_1^2 + 2\pi x_1 x_2 \\
  \theta(x) &= \pi x_1^2 x_2 - 1
\end{align*}
\]
Equality constraints
Solution: Lagrangian $L(x, \lambda) = f(x) + \lambda \theta(x)$

\[
\frac{\partial L(x, \lambda)}{\partial x_1} = 2\pi x_1 + 2\pi x_2 + \lambda 2\pi x_1 x_2 = 0 \\
\frac{\partial L(x, \lambda)}{\partial x_2} = 2\pi x_1 + \lambda \pi x_1^2 = 0
\]

We obtain $x_1^* = x_2^* = -\frac{2}{\lambda}$.

Finally, $\theta(x^*) = 0$ gives the value of $\lambda$ to plug:

$\lambda^* = -\frac{\pi^{1/3}}{2}$, so $x_1^* = x_2^* = \pi^{-1/3}$. 
Another interpretation

Consider the unconstrained problem, where \( \lambda \) is fixed

\[
\min_x L(x, \lambda) = f(x) + \lambda \theta(x)
\]

\( f \) and \( L \) have the same local minima in \( D = \{x : \theta(x) = 0\} \).

Let \( x^*(\lambda) \) be a local minimum of \( L(x, \lambda) \) in \( \mathbb{R}^d \).

If \( x^*(\lambda) \in D \), then it is also a local min of \( f \).

So one just has to adjust \( \lambda \) to get this property.
Second order conditions

Theorem

Let \((x^*, \lambda^*)\) be a stationary point of \(L(x, \lambda)\), and consider the Hessian of the Lagrangian

\[
\nabla^2_x L(x^*, \lambda^*) = \nabla^2 f(x^*) + \sum_{j=1}^{m} \lambda_j^* \nabla^2 \theta_j(x^*)
\]

- **NC**: \(x^*\) is a local min of \(f\) on \(D\) \(\Rightarrow\) \(\nabla^2_x L(x^*, \lambda^*)\) is a positive quadratic form on the tangent space at \(x^*\), i.e. the kernel of matrix \(\nabla \theta(x^*)^t\).

- **SC**: \(\nabla^2_x L(x^*, \lambda^*)\) is strictly positive on the tangent space \(\Rightarrow\) \(x^*\) is a local min of \(f\) on \(D\)
Equality constraints
Inequality constraints

\[ \min_x f(x) \text{ s.t. } \theta_j(x) \leq 0, \ 1 \leq j \leq m \]

- \( D : \theta(x) = [\theta_1(x), \ldots, \theta_m(x)]^t \leq 0 \) defines a volume in \( \mathbb{R}^d \) limited by \( m \) manifolds of dimension \( d - 1 \).
- At point \( x \), constraint \( \theta_j \) is active iff \( \theta_j(x) = 0 \).
  \[ \mathcal{A}(x) = \{ j : \theta_j(x) = 0 \} = \text{active set at } x. \]
- One could have simultaneously equality and inequality constraints (not done here for a matter of clarity).
  Equality constraints are always active.
- \( \bigcap_{j \in \mathcal{A}(x^0)} \{ x : \nabla \theta_j(x^0)^t (x - x^0) = 0 \} \)
  defines the tangent space to \( D \) at \( x^0 \).
Let $x^0 \in D$, we look for directions $d \in \mathbb{R}^d$ that keep us inside domain $D$: $x^0 + \epsilon \cdot d \in D$.

**Definition**

Direction $d$ is **admissible** from $x^0$ iff $\exists (x^n)_{n>0}$ in $D$ such that

\[
\lim_{n} x^n = x^0 \quad \text{and} \quad \lim_{n} \frac{x^n - x^0}{\|x^n - x^0\|} = \frac{d}{\|d\|}
\]
- Admissible directions at $x^0$ form a cone $C(x^0)$.
- This cone is not necessarily convex...

$C(x^0)$ can be determined from the $\nabla \theta_j(x^0)$ of the active constraints.
Theorem

If $x^0$ is a regular point, i.e. the gradients of the active constraints at $x^0$ are linearly independent, then $C(x^0)$ is the convex cone given by

$$C(x^0) = \{ u \in \mathbb{R}^d : \nabla \theta_j(x^0)^t u \leq 0, \ j \in A(x^0) \}$$

Interpretation: an admissible displacement must not increase the value of $\theta_j(x^0)$ for an already active constraint, it can only decrease it or leave it unchanged.
For $v_1, \ldots, v_J \in \mathbb{R}^d$, consider cone $C = \{ u : u^t v_1 \leq 0, \ldots, u^t v_J \leq 0 \}$.

**Farkas-Minkowski lemma**

Let $g \in \mathbb{R}^d$, one has the equivalence

- $\forall u \in C$, $g^t u \leq 0$,
- $C$ is included in the half-space $\{ u : g^t u \leq 0 \}$,
- $g$ belongs to the dual cone $C' = \{ w : \forall u \in C, w^t u \leq 0 \}$
- $g = \sum_{j=1}^{J} \alpha_j v_j$ where $\alpha_j \geq 0$ for all $j$
1st order optimality conditions

**Theorem (Karush-Kuhn-Tucker conditions)**

Let $x^*$ be a regular point of domain $D$. If $x^*$ is a local minimum of $f$ in $D$, there exists a unique set of **generalized Lagrange multipliers** $\lambda^*_j$ for $j \in A(x^*)$ such that

$$\nabla f(x^*) + \sum_{j \in A(x^*)} \lambda^*_j \nabla \theta_j(x^*) = 0 \quad \text{and} \quad \lambda^*_j \geq 0, \ j \in A(x^*)$$

**Remarks:**

- Similar to the case of equality constraints: here only active constraints are considered.
- The positivity condition is new: translates the fact that one side of the manifold is permitted.
Theorem (Karush-Kuhn-Tucker conditions)

Let $x^*$ be a regular point of domain $D$. If $x^*$ is a local minimum of $f$ in $D$, there exists a unique set of generalized Lagrange multipliers $\lambda^*_j$ for $j \in A(x^*)$ such that

$$
\nabla f(x^*) + \sum_{j \in A(x^*)} \lambda^*_j \nabla \theta_j(x^*) = 0 \quad \text{and} \quad \lambda^*_j \geq 0, \ j \in A(x^*)
$$

Remarks:

- Similar to the case of equality constraints: here only active constraints are considered.
- The positivity condition is new: translates the fact that one side of the manifold is permitted.
Proof:

- take any admissible direction:
  \[ d \in C(x^*) = \{ u : u^t \nabla \theta_j(x^*) \leq 0, \ j \in A(x^*) \} \]
- progressing along \( d \) doesn’t decrease \( f \):
  \[ [-\nabla f(x^*)]^t d \leq 0 \]
- this means that \( g = -\nabla f(x^*) \) belongs to the dual cone \( C(x^*)' \), so by Farkas lemma

\[
-\nabla f(x^*) = \sum_{j \in A(x^*)} \lambda_j^* \nabla \theta_j(x^*) \quad \text{and} \quad \lambda_j^* \geq 0
\]
Corollary

The Karush-Kuhn-Tucker conditions are equivalent to

\[ \nabla f(x^*) + \sum_{j=1}^{m} \lambda_j^* \nabla \theta_j(x^*) = 0 \quad \text{and} \quad \lambda_j^* \geq 0, \quad 1 \leq j \leq m \]

with the extra complementarity condition

\[ \sum_{j=1}^{m} \lambda_j^* \theta_j(x^*) = 0 \]

- This entails \( \lambda_j^* = 0 \) for an inactive constraint \( \theta_j \) at \( x^* \).
- To be usable, requires to know/guess the set of active constraints at the optimum.
- \( \mathcal{A}(x^*) \) known, leaves a set of non-linear equations + positivity constraints.
**Problem:** Minimize distance from point $P$ to the red segment

$$
\min_x f(x) = (x_1 - 1)^2 + (x_2 - 2)^2
$$

subject to

$$
\theta_1(x) = x_1 - x_2 - 1 = 0
$$

$$
\theta_2(x) = x_1 + x_2 - 2 \leq 0
$$

$$
\theta_3(x) = -x_1 \leq 0
$$

$$
\theta_4(x) = -x_2 \leq 0
$$
Objective: cancel the gradient of the Lagrangian

\[ \frac{\partial L(x, \lambda)}{\partial x_1} = x_1 - 1 + \lambda_1 + \lambda_2 - \lambda_3 = 0 \]

\[ \frac{\partial L(x, \lambda)}{\partial x_2} = x_2 - 2 - \lambda_1 + \lambda_2 - \lambda_4 = 0 \]

equality \[ \frac{\partial L(x, \lambda)}{\partial \lambda_1} = x_1 - x_2 - 1 = 0 \]

inequalities \[ \lambda_2(x_1 + x_2 - 2) = 0, \quad \lambda_2 \geq 0 \]
\[ -\lambda_3 x_1 = 0, \quad \lambda_3 \geq 0 \]
\[ -\lambda_4 x_2 = 0, \quad \lambda_4 \geq 0 \]

1st guess: \( A(x^*) = \{1\}, \text{ i.e. only } \theta_1 \text{ active at the optimum.} \)
Complementarity \( \Rightarrow \lambda_2^* = \lambda_3^* = \lambda_4^* = 0. \)
This yields \( x^* = (2, 1) \) which violates \( \theta_2(x) \leq 0. \)
Objective: cancel the gradient of the Lagrangian

\[
\frac{\partial L(x, \lambda)}{\partial x_1} = x_1 - 1 + \lambda_1 + \lambda_2 - \lambda_3 = 0
\]
\[
\frac{\partial L(x, \lambda)}{\partial x_2} = x_2 - 2 - \lambda_1 + \lambda_2 - \lambda_4 = 0
\]

Equality
\[
\frac{\partial L(x, \lambda)}{\partial \lambda_1} = x_1 - x_2 - 1 = 0
\]

Inequalities
\[
\lambda_2(x_1 + x_2 - 2) = 0, \quad \lambda_2 \geq 0
\]
\[
-\lambda_3 x_1 = 0, \quad \lambda_3 \geq 0
\]
\[
-\lambda_4 x_2 = 0, \quad \lambda_4 \geq 0
\]

2nd guess: \( \mathcal{A}(x^*) = \{1, 2\} \), i.e. \( \theta_2 \) is added to the active set.
Complementarity \( \Rightarrow \lambda^*_3 = \lambda^*_4 = 0 \).
This yields \( x^* = \left( \frac{3}{2}, \frac{1}{2} \right) \) which belongs to \( \mathcal{D} \).
[ For simplicity we consider the case of inequality constraints. ]

Idea: under some conditions, a stationary point \((x^*, \lambda^*)\) of the Lagrangian, i.e. 
\[
\nabla L(x^*, \lambda^*) = \nabla f(x^*) + \sum_i \lambda_i^* \nabla \theta(x^*) = 0
\]
corresponds to a saddle point of the Lagrangian, i.e.
\[
\inf_x L(x, \lambda^*) = L(x^*, \lambda^*) = \sup_\lambda L(x^*, \lambda)
\]

So the resolution amounts to finding such saddle points, and then extract \(x^*\).
Saddle points

Definition

\((x^*, \lambda^*)\) is a **saddle point** of \(L\) in \(D_x \times D_\lambda\) iff

\[
\sup_{\lambda \in D_\lambda} L(x^*, \lambda) = L(x^*, \lambda^*) = \inf_{x \in D_x} L(x, \lambda^*)
\]
Lemma

If \((x^*, \lambda^*)\) is a saddle point of \(L\) in \(D_x \times D_\lambda\), then

\[
\sup_{\lambda \in D_\lambda} \inf_{x \in D_x} L(x, \lambda) = L(x^*, \lambda^*) = \inf_{x \in D_x} \sup_{\lambda \in D_\lambda} L(x, \lambda)
\]

Proof

- one always has \(\sup_{\lambda} \inf_{x} L(x, \lambda) \leq \inf_{x} \sup_{\lambda} L(x, \lambda)\)
- the difference is called the duality gap, generally \(> 0\)
- from the def. of a saddle point, one has

\[
\sup_{\lambda} L(x^*, \lambda) = L(x^*, \lambda^*) = \inf_{x} L(x, \lambda^*)
\]

then

\[
\inf_{x} \left[ \sup_{\lambda} L(x, \lambda) \right] \leq \sup_{\lambda} L(x^*, \lambda)
\]

\[
\inf_{x} L(x, \lambda^*) \leq \sup_{\lambda} \left[ \inf_{x} L(x, \lambda) \right]
\]

- Morality: one can look for \(\lambda^*\) first, and then for \(x^*\)...
Lemma

If \((x^*, \lambda^*)\) is a saddle point of \(L\) in \(D_x \times D_\lambda\), then

\[
\sup_{\lambda \in D_\lambda} \inf_{x \in D_x} L(x, \lambda) = L(x^*, \lambda^*) = \inf_{x \in D_x} \sup_{\lambda \in D_\lambda} L(x, \lambda)
\]

Proof

- one always has \(\sup_{\lambda} \inf_{x} L(x, \lambda) \leq \inf_{x} \sup_{\lambda} L(x, \lambda)\)
  the difference is called the duality gap, generally \(> 0\)
- from the def. of a saddle point, one has
  \[
  \sup_{\lambda} L(x^*, \lambda) = L(x^*, \lambda^*) = \inf_{x} L(x, \lambda^*)
  \]
  then
  \[
  \inf_{x} \left[ \sup_{\lambda} L(x, \lambda) \right] \leq \sup_{\lambda} L(x^*, \lambda)
  \]
  \[
  \inf_{x} L(x, \lambda^*) \leq \sup_{\lambda} \left[ \inf_{x} L(x, \lambda) \right]
  \]
- Morality: one can look for \(\lambda^*\) first, and then for \(x^*\)...
Lemma

If \((x^*, \lambda^*)\) is a saddle point of \(L\) in \(D_x \times D_\lambda\), then

\[
\sup_{\lambda \in D_\lambda} \inf_{x \in D_x} L(x, \lambda) = L(x^*, \lambda^*) = \inf_{x \in D_x} \sup_{\lambda \in D_\lambda} L(x, \lambda)
\]

Proof

- one always has \(\sup_{\lambda} \inf_{x} L(x, \lambda) \leq \inf_{x} \sup_{\lambda} L(x, \lambda)\)
  - the difference is called the *duality gap*, generally \(> 0\)
- from the def. of a saddle point, one has
  \[
  \sup_{\lambda} L(x^*, \lambda) = L(x^*, \lambda^*) = \inf_{x} L(x, \lambda^*)
  \]

then

\[
\inf_{x} \left[ \sup_{\lambda} L(x, \lambda) \right] \leq \sup_{\lambda} L(x^*, \lambda)
\]

\[
\inf_{x} L(x, \lambda^*) \leq \sup_{\lambda} \left[ \inf_{x} L(x, \lambda) \right]
\]

- Morality: one can look for \(\lambda^*\) first, and then for \(x^*\)...
**Lemma**

If \((x^*, \lambda^*)\) is a saddle point of \(L\) in \(\mathcal{D}_x \times \mathcal{D}_\lambda\), then

\[
\sup_{\lambda \in \mathcal{D}_\lambda} \inf_{x \in \mathcal{D}_x} L(x, \lambda) = L(x^*, \lambda^*) = \inf_{x \in \mathcal{D}_x} \sup_{\lambda \in \mathcal{D}_\lambda} L(x, \lambda)
\]

**Proof**

- one always has \(\sup_{\lambda} \inf_{x} L(x, \lambda) \leq \inf_{x} \sup_{\lambda} L(x, \lambda)\)
  - the difference is called the *duality gap*, generally \(\geq 0\)
- from the def. of a saddle point, one has
  \[
  \sup_{\lambda} L(x^*, \lambda) = L(x^*, \lambda^*) = \inf_{x} L(x, \lambda^*)
  \]
  then

\[
\inf_{x} [\sup_{\lambda} L(x, \lambda)] \leq \sup_{\lambda} L(x^*, \lambda) \\
\inf_{x} L(x, \lambda^*) \leq \sup_{\lambda} [\inf_{x} L(x, \lambda)]
\]

- Morality: one can look for \(\lambda^*\) first, and then for \(x^*\)...
Saddle points of the Lagrangian

**Theorem**

If \((x^*, \lambda^*)\) is a saddle point of the Lagrangian \(L\) in \(\mathbb{R}^d \times \mathbb{R}^m_+\), then \(x^*\) is a solution of the primal problem \((P)\)

\[
\min_{x} f(x) \quad \text{s.t.} \quad \theta_i(x) \leq 0, \quad 1 \leq i \leq m
\]

**Proof**

- From \(L(x^*, \lambda) \leq L(x^*, \lambda^*), \quad \forall \lambda \in D_\lambda = \mathbb{R}^m_+\)

\[
f(x^*) + \sum_i \lambda_i \theta_i(x^*) \leq f(x^*) + \sum_i \lambda_i^* \theta_i(x^*)
\]

\[
\sum_i (\lambda_i - \lambda_i^*) \theta_i(x^*) \leq 0
\]

whence \(\theta_i(x^*) \leq 0\) by \(\lambda_i \to +\infty\): \(x^*\) satisfies constraints
Saddle points of the Lagrangian

Theorem

If \((x^*, \lambda^*)\) is a saddle point of the Lagrangian \(L\) in \(\mathbb{R}^d \times \mathbb{R}_+^m\), then \(x^*\) is a solution of the primal problem \((P)\)

\[
(P) \quad \min_x f(x) \quad \text{s.t.} \quad \theta_i(x) \leq 0, \quad 1 \leq i \leq m
\]

Proof

From \(L(x^*, \lambda) \leq L(x^*, \lambda^*), \quad \forall \lambda \in D_{\lambda} = \mathbb{R}_+^m\)

\[
f(x^*) + \sum_i \lambda_i \theta_i(x^*) \leq f(x^*) + \sum_i \lambda_i^* \theta_i(x^*)
\]

\[
\sum_i (\lambda_i - \lambda_i^*) \theta_i(x^*) \leq 0
\]

whence \(\theta_i(x^*) \leq 0\) by \(\lambda_i \rightarrow +\infty\): \(x^*\) satisfies constraints
Moreover, \( \sum_i -\lambda_i^* \theta_i(x^*) \leq 0 \), by \( \lambda_i = 0 \), and so \( \sum_i \lambda_i^* \theta_i(x^*) = 0 \) (complementarity condition)

From \( L(x^*, \lambda^*) \leq L(x, \lambda^*) \), \( \forall x \in \mathbb{R}^d \)

\[
f(x^*) + \sum_i \lambda_i^* \theta_i(x^*) \leq f(x) + \sum_i \lambda_i^* \theta_i(x)
\]

so for all admissible \( x \), i.e. such that \( \theta_i(x) \leq 0 \), \( 1 \leq i \leq m \)

\[
f(x^*) \leq f(x)
\]

Summary:
saddle points of the Lagrangian, when they exist, give solutions to the optimization problem.
But they don’t always exist...
Moreover, \( \sum_i -\lambda_i^* \theta_i(x^*) \leq 0 \), by \( \lambda_i = 0 \),
and so \( \sum_i \lambda_i^* \theta_i(x^*) = 0 \) (complementarity condition)

From \( L(x^*, \lambda^*) \leq L(x, \lambda^*) \), \( \forall x \in \mathbb{R}^d \)

\[
f(x^*) + \sum_i \lambda_i^* \theta_i(x^*) \leq f(x) + \sum_i \lambda_i^* \theta_i(x)
\]
so for all admissible \( x \), i.e. such that \( \theta_i(x) \leq 0, 1 \leq i \leq m \)

\[
f(x^*) \leq f(x)
\]

Summary:
saddle points of the Lagrangian, when they exist, give solutions to the optimization problem.

*But they don’t always exist...*
Moreover,  \( \sum_i -\lambda^*_i \theta_i(x^*) \leq 0 \), by \( \lambda_i = 0 \), and so  \( \sum_i \lambda^*_i \theta_i(x^*) = 0 \) (complementarity condition).

From  \( L(x^*, \lambda^*) \leq L(x, \lambda^*) \),  \( \forall x \in \mathbb{R}^d \)

\[
f(x^*) + \sum_i \lambda^*_i \theta_i(x^*) \leq f(x) + \sum_i \lambda^*_i \theta_i(x)
\]

so for all admissible  \( x \), i.e. such that  \( \theta_i(x) \leq 0 \),  \( 1 \leq i \leq m \)

\[
f(x^*) \leq f(x)
\]

**Summary:**
saddle points of the Lagrangian, when they exist, give solutions to the optimization problem.

*But they don’t always exist...*
Existence of saddle points

**Theorem**

If $f$ and the constraints $\theta_i$ are **convex** functions of $x$ in $\mathbb{R}^d$, and if $x^* = \arg \min_x f(x)$ in $\{x : \theta_i(x) \leq 0, 1 \leq i \leq m\}$ is regular then $x^*$ corresponds to a saddle point $(x^*, \lambda^*)$ of the Lagrangian

Proof: from Kuhn-Tucker, derive the saddle point property

- $L(x^*, \lambda) = f(x^*) + \sum_i \lambda_i \theta_i(x^*)$
  
  $\leq f(x^*) = f(x^*) + \sum_i \lambda_i \theta_i(x^*) = L(x^*, \lambda^*)$

  using admissibility of $x^*$, positivity of $\lambda_i$ and complementarity

- $L(x, \lambda^*) = f(x) + \sum_i \lambda_i^* \theta_i(x)$ is a convex function of $x$

  From the stationarity of $L$, one has

  $\nabla_x L(x^*, \lambda^*) = \nabla f(x^*) + \sum_i \lambda_i^* \nabla \theta_i(x^*) = 0$

  sufficient to show that $x^*$ is a minimum of the convex function $L(x, \lambda^*)$
Existence of saddle points

Theorem

If \( f \) and the constraints \( \theta_i \) are convex functions of \( x \) in \( \mathbb{R}^d \), and if \( x^* = \arg \min_x f(x) \) in \( \{ x : \theta_i(x) \leq 0, 1 \leq i \leq m \} \) is regular then \( x^* \) corresponds to a saddle point \( (x^*, \lambda^*) \) of the Lagrangian

**Proof:** from Kuhn-Tucker, derive the saddle point property

\[
L(x^*, \lambda) = f(x^*) + \sum_i \lambda_i \theta_i(x^*) \\
\leq f(x^*) = f(x^*) + \sum_i \lambda_i \theta_i(x^*) = L(x^*, \lambda^*)
\]

using admissibility of \( x^* \), positivity of \( \lambda_i \) and complementarity

\[
L(x, \lambda^*) = f(x) + \sum_i \lambda_i^* \theta_i(x)
\]

is a convex function of \( x \)

From the stationarity of \( L \), one has

\[
\nabla_x L(x^*, \lambda^*) = \nabla f(x^*) + \sum_i \lambda_i^* \nabla \theta_i(x^*) = 0
\]

sufficient to show that \( x^* \) is a minimum of the convex function \( L(x, \lambda^*) \)
Existence of saddle points

Theorem

If \( f \) and the constraints \( \theta_i \) are \textbf{convex} functions of \( x \) in \( \mathbb{R}^d \), and if \( x^* = \arg\min_x f(x) \) in \( \{x : \theta_i(x) \leq 0, 1 \leq i \leq m\} \) is regular then \( x^* \) corresponds to a saddle point \( (x^*, \lambda^*) \) of the Lagrangian

Proof: from Kuhn-Tucker, derive the saddle point property

\[ L(x^*, \lambda) = f(x^*) + \sum_i \lambda_i \theta_i(x^*) \leq f(x^*) = f(x^*) + \sum_i \lambda_i \theta_i(x^*) = L(x^*, \lambda^*) \]

using admissibility of \( x^* \), positivity of \( \lambda_i \) and complementarity

\[ L(x, \lambda^*) = f(x) + \sum_i \lambda_i^* \theta_i(x) \]

is a convex function of \( x \)

From the stationarity of \( L \), one has

\[ \nabla_x L(x^*, \lambda^*) = \nabla f(x^*) + \sum_i \lambda_i^* \nabla \theta_i(x^*) = 0 \]

sufficient to show that \( x^* \) is a minimum of the convex function \( L(x, \lambda^*) \)
Summary: Provided the Lagrangian has saddle points

- Solutions to \((P)\) \(\min_x f(x)\) s.t. \(\theta_i(x) \leq 0, 1 \leq i \leq m\)
  are the 1st argument of a saddle point \((x^*, \lambda^*)\) of the Lagrangian \(L(x, \lambda)\)

- If \(\lambda^*\) were known, amounts to solving an unconstrained problem
  \[x^* = \arg \min_x L(x, \lambda^*)\]

- How to find such a \(\lambda^*\) ?
  
  One has \(L(x^*, \lambda^*) = \max_{\lambda \in \mathbb{R}_+^m} \min_x L(x, \lambda),\)
  
  so \(\lambda^*\) should be a solution of the dual problem
  
  \[(D) \max_{\lambda} g(\lambda), \quad \text{s.t.} \quad \lambda \in \mathbb{R}_+^m, \quad \text{where} \quad g(\lambda) = \min_x L(x, \lambda)\]
Dual problem

Summary: Provided the Lagrangian has saddle points

- Solutions to \((P)\) \(\min_x f(x) \text{ s.t. } \theta_i(x) \leq 0, \ 1 \leq i \leq m\)
  are the 1st argument of a saddle point \((x^*, \lambda^*)\) of the Lagrangian \(L(x, \lambda)\)

- If \(\lambda^*\) were known, amounts to solving an unconstrained problem

\[
x^* = \arg\min_x L(x, \lambda^*)
\]

- How to find such a \(\lambda^*\) ?
  One has \(L(x^*, \lambda^*) = \max_{\lambda \in \mathbb{R}^m_+} \min_x L(x, \lambda)\), so \(\lambda^*\) should be a solution of the dual problem
  \[
  (D) \quad \max_{\lambda} g(\lambda), \text{ s.t. } \lambda \in \mathbb{R}^m_+, \text{ where } g(\lambda) = \min_x L(x, \lambda)
  \]
Dual problem

Summary: Provided the Lagrangian has saddle points

- Solutions to \((P)\) \(\min_x f(x)\) s.t. \(\theta_i(x) \leq 0, 1 \leq i \leq m\) are the 1st argument of a saddle point \((x^*, \lambda^*)\) of the Lagrangian \(L(x, \lambda)\)

- If \(\lambda^*\) were known, amounts to solving an **unconstrained** problem

\[ x^* = \arg \min_x L(x, \lambda^*) \]

- How to find such a \(\lambda^*\)?
  
  One has \(L(x^*, \lambda^*) = \max_{\lambda \in \mathbb{R}^m_+} \min_x L(x, \lambda)\), so \(\lambda^*\) should be a solution of the **dual problem**

\[ (D) \max_{\lambda} g(\lambda), \text{ s.t. } \lambda \in \mathbb{R}^m_+, \text{ where } g(\lambda) = \min_x L(x, \lambda) \]
Under some conditions, it is equivalent to solve the (P) or (D):

**Theorem**

- If the $\theta_i$ are continuous over $\mathbb{R}^d$, and
  $\forall \lambda \in \mathbb{R}_+^m$, $x^*(\lambda) = \arg \min_x L(x, \lambda)$ is unique, and
  $x^*(\lambda)$ is a continuous function of $\lambda$
  then $\lambda^*$ solves (D) \Rightarrow \ x^*(\lambda^*)$ solves (P)

- If (P) has at least one solution $x^*$, $f$ and the $\theta_i$ are convex and $x^*$ is regular, then (D) has at least a solution $\lambda^*$.

**Remark**

(D) is still an optimization problem under constraints...
... but constraints $\lambda \in \mathbb{R}_+^m$ are much simpler to handle!
Under some conditions, it is equivalent to solve the (P) or (D):

**Theorem**

- If the $\theta_i$ are continuous over $\mathbb{R}^d$, and
  $\forall \lambda \in \mathbb{R}^m_+$, $x^*(\lambda) = \arg \min_x L(x, \lambda)$ is unique, and
  $x^*(\lambda)$ is a continuous function of $\lambda$
  then $\lambda^*$ solves (D) $\Rightarrow x^*(\lambda^*)$ solves (P)

- If (P) has at least one solution $x^*$, $f$ and the $\theta_i$ are convex and $x^*$ is regular, then (D) has at least a solution $\lambda^*$.

**Remark**

(D) is still an optimization problem under constraints...
... but constraints $\lambda \in \mathbb{R}^m_+$ are much simpler to handle!
Example

Minimize a quadratic function under a quadratic constraint in $\mathbb{R}$

- $\min_x (x - x_0)^2 \quad \text{s.t.} \quad (x - x_1)^2 - d \leq 0$ with $d > 0$, $x_1 > x_0$
- convex, regular case... unique saddle point of the Lagrangian

![Graph showing the minimization of a quadratic function under a quadratic constraint](image)
Optimization without constraints

Dual problem - Resolution by duality

- \( L(x, \lambda) = (x - x_0)^2 + \lambda[(x - x_1)^2 - d] \)

- Compute \( g(\lambda) = \min_{x \in \mathbb{R}} L(x, \lambda) \)
  \[ \nabla_x L(x, \lambda) = 2(x - x_0) + 2\lambda(x - x_1) = 0 \quad \Rightarrow \quad x^*(\lambda) = \frac{x_0 + \lambda x_1}{1 + \lambda} \]

- \( g(\lambda) = (x_1 - x_0)^2 \frac{\lambda}{1 + \lambda} - \lambda d \)

- Solve (D) : \( \max_{\lambda \geq 0} g(\lambda) \)
  \[ g'(\lambda) = \frac{(x_1 - x_0)^2}{(1 + \lambda)^2} - d = 0 \]
  \[ \lambda^* = \frac{x_1 - x_0}{\sqrt{d}} - 1 \text{ if } \geq 0, \text{ otherwise } \lambda^* = 0 \text{ (constraint is inactive)} \]

- When \( \lambda^* > 0 \), \( x^*(\lambda^*) = x_1 - \sqrt{d} \)
  otherwise, for \( \lambda^* = 0 \), \( x^*(\lambda^*) = x_0 \)
\[ L(x, \lambda) = (x - x_0)^2 + \lambda[(x - x_1)^2 - d] \]

- Compute \( g(\lambda) = \min_{x \in \mathbb{R}} L(x, \lambda) \)
  \[ \nabla_x L(x, \lambda) = 2(x - x_0) + 2\lambda(x - x_1) = 0 \quad \Rightarrow \quad x^*(\lambda) = \frac{x_0 + \lambda x_1}{1 + \lambda} \]

- \( g(\lambda) = (x_1 - x_0)^2 \frac{\lambda}{1 + \lambda} - \lambda d \)

- Solve (D) : \( \max_{\lambda \geq 0} g(\lambda) \)
  \[ g'(\lambda) = \frac{(x_1 - x_0)^2}{(1 + \lambda)^2} - d = 0 \]
  \[ \lambda^* = \frac{x_1 - x_0}{\sqrt{d}} - 1 \text{ if } \geq 0, \text{ otherwise } \lambda^* = 0 \text{ (constraint is inactive)} \]

- When \( \lambda^* > 0 \), \( x^*(\lambda^*) = x_1 - \sqrt{d} \)
  otherwise, for \( \lambda^* = 0 \), \( x^*(\lambda^*) = x_0 \)
\[ L(x, \lambda) = (x - x_0)^2 + \lambda[(x - x_1)^2 - d] \]

- Compute \( g(\lambda) = \min_{x \in \mathbb{R}} L(x, \lambda) \)
  \[ \nabla_x L(x, \lambda) = 2(x - x_0) + 2\lambda(x - x_1) = 0 \Rightarrow x^*(\lambda) = \frac{x_0 + \lambda x_1}{1 + \lambda} \]
  \[ g(\lambda) = (x_1 - x_0)^2 \frac{\lambda}{1 + \lambda} - \lambda d \]

- Solve (D) : \( \max_{\lambda \geq 0} g(\lambda) \)
  \[ g'(\lambda) = \frac{(x_1 - x_0)^2}{(1 + \lambda)^2} - d = 0 \]
  \[ \lambda^* = \frac{x_1 - x_0}{\sqrt{d}} - 1 \text{ if } \geq 0, \text{ otherwise } \lambda^* = 0 \text{ (constraint is inactive)} \]

- When \( \lambda^* > 0, \quad x^*(\lambda^*) = x_1 - \sqrt{d} \)
  otherwise, for \( \lambda^* = 0, \quad x^*(\lambda^*) = x_0 \)
Plot of the Lagrangian
Case where $\lambda^* > 0$, i.e. $x_1 - x_0 > \sqrt{d}$
(here $x_0 = 1$, $x_1 = 3$, $d = 1$, $\lambda^* = 1$)
Plot of the Lagrangian

Case where $\lambda^* = 0$, i.e. $x_1 - x_0 \leq \sqrt{d}$

(here $x_0 = 1$, $x_1 = 1.5$, $d = 1$)
Numerical methods

Same principle as for the unconstrained case, with 2 extra difficulties

- constraints limit the choice of admissible directions,
- progressing along an admissible direction may meet the boundary of $\mathcal{D}$. 
Penalty functions

Also called **Lagrangian relaxation**

**Exterior points method**: for equality constraints $\theta(x) = 0$

- Principle = penalize non-admissible solutions.
- Let $\psi(x) \geq 0$ and $\psi(x) = 0$ exactly on $D$, for example $\psi(x) = \|\theta(x)\|^2$
- Consider the unconstrained problem

  $$\min_{x} F(x) = f(x) + c_k \psi(x), \quad c_k > 0$$

  and let $c_k$ go to $+\infty$. 
**Interior points method**: better suited to inequalities $\theta(x) \leq 0$

- Principle = completely forbid non-admissible solutions, penalize those that get close to the boundaries of $\mathcal{D}$.
- Let $\psi(x) \geq 0$ and $\psi(x) \to +\infty$ when $\theta_j(x) \to 0_-$, for example $\psi(x) = -\sum_j \frac{1}{\theta_j(x)}$
- then same as exterior points method: $\min_x F(x) = f(x) + c_k \psi(x)$...
Projected gradient: equality constraints

**Principle:** project $-\nabla f(x^n)$ on the tangent space to constraints

**Lemma**

Let $C = [C_1, \ldots, C_m] \in \mathbb{R}^{d \times m}$ be the matrix formed by $m$ linearly independent (column) vectors $C_j$ of $\mathbb{R}^d$. In $\mathbb{R}^r$, the projection on $sp\{C_1, \ldots, C_m\}$ is given by

$$\pi_C(x) = Px \quad \text{with} \quad P = C(C^t C)^{-1}C^t$$

**Proof:** This amounts to solving the quadratic problem

$$\min_{\alpha \in \mathbb{R}^m} \|x - C\alpha\|^2$$

**Remark:** The projection on $sp\{C_1, \ldots, C_m\}^\perp = \{x : C^t x = 0\}$ is given by matrix $Q = I - P$
Projected gradient: equality constraints

**Principle:** project $-\nabla f(x^n)$ on the tangent space to constraints

**Lemma**

Let $C = [C_1, \ldots, C_m] \in \mathbb{R}^{d \times m}$ be the matrix formed by $m$ linearly independent (column) vectors $C_j$ of $\mathbb{R}^d$. In $\mathbb{R}^r$, the projection on $sp\{C_1, \ldots, C_m\}$ is given by

$$\pi_C(x) = Px \quad \text{with} \quad P = C(C^tC)^{-1}C^t$$

**Proof:** This amounts to solving the quadratic problem

$$\min_{\alpha \in \mathbb{R}^m} \|x - C\alpha\|^2$$

**Remark:** The projection on $sp\{C_1, \ldots, C_m\}^\perp = \{x : C^t x = 0\}$ is given by matrix $Q = I - P$
Affine equality constraints

- Replace $\min_x f(x)$ s.t. $\theta(x) = C^t x - c = 0$, by $\min_x F(x)$ with $F(x) = f[\pi_D(x)]$.
- These two functions coincide on $D = \{x : \theta(x) = 0\} = \{x : x = \pi_D(x)\}$
For \( x^0 \in D \), one has \( \pi_D(x) = x^0 + Q(x - x^0) \), so

\[
\min_x F(x) = f[x^0 + Q(x - x^0)] \\
\nabla F(x) = Q \nabla f[x^0 + Q(x - x^0)] \\
\nabla^2 F(x) = Q \nabla f[x^0 + Q(x - x^0)] Q
\]

\( \nabla F(x^n) \) is the projection of \( \nabla f(x^n) \) on \( D = \{ x : C^t x = 0 \} \).

Iterations starting with \( x^0 \in D \) stay in \( D \).
Non linear equality constraints,

- project $\nabla f(x)$ on the tangent space $sp\{\nabla_1 \theta(x), ..., \nabla_m \theta(x)\}^\perp$ ...

- ... then project $x^{n+1}$ on $\mathcal{D}$. 
Projected gradient: affine inequality constraints

- Similar to the case of equality constraints, but only active constraints are considered.
- Some constraints may become active/inactive during the linear search...
- Stop when the Kuhn-Tucker conditions are met.
Optimization without constraints

Optimization under constraints

Conclusion

Numerical methods

dual cone of admissible directions: stop decision zone
Outline

1. Optimization without constraints
   - Optimization scheme
   - Linear search methods
   - Gradient descent
   - Conjugate gradient
   - Newton method
   - Quasi-Newton methods

2. Optimization under constraints
   - Lagrange
   - Equality constraints
   - Inequality constraints
   - Dual problem - Resolution by duality
   - Numerical methods
     - Penalty functions
     - Projected gradient: equality constraints
     - Projected gradient: inequality constraints

3. Conclusion
When facing a constrained optimization problem...

...one reflex: **build the Lagrangian!**

\[ L(x, \lambda) = f(x) + \sum_j \lambda_j \theta_j(x) \]

- solve \( \nabla L(x, \lambda) = 0 \) to find a candidate optimum \( (x^*, \lambda^*) \)
- check the positivity of its Hessian \( \nabla_x^2 L(x^*, \lambda^*) \) to check if \( x^* \) is a min, a max or a saddle point of \( f \).
Conclusion

When facing a constrained optimization problem...

...one reflex: **build the Lagrangian!**

\[ L(x, \lambda) = f(x) + \sum_j \lambda_j \theta_j(x) \]

- solve \( \nabla L(x, \lambda) = 0 \) to find a candidate optimum \((x^*, \lambda^*)\)
- check the positivity of its Hessian \( \nabla^2_x L(x^*, \lambda^*) \) to check if \( x^* \) is a min, a max or a saddle point of \( f \).
Conclusion

When facing a constrained optimization problem...
...one reflex: **BUILD THE LAGRANGIAN**!

\[ L(x, \lambda) = f(x) + \sum_j \lambda_j \theta_j(x) \]

- solve \( \nabla L(x, \lambda) = 0 \) to find a candidate optimum \((x^*, \lambda^*)\)
- check the positivity of its Hessian \( \nabla_x^2 L(x^*, \lambda^*) \) to check if \( x^* \) is a min, a max or a saddle point of \( f \).
Conclusion

When facing a constrained optimization problem...

...one reflex: **BUILD THE LAGRANGIAN**!

\[
L(x, \lambda) = f(x) + \sum_j \lambda_j \theta_j(x)
\]

- solve \( \nabla L(x, \lambda) = 0 \) to find a candidate optimum \((x^*, \lambda^*)\)
- check the positivity of its Hessian \( \nabla^2_x L(x^*, \lambda^*) \) to check if \( x^* \)
  is a min, a max or a saddle point of \( f \).
Conclusion

When facing a constrained optimization problem...

...one reflex: **build the Lagrangian**!

\[ L(x, \lambda) = f(x) + \sum_j \lambda_j \theta_j(x) \]

- solve \( \nabla L(x, \lambda) = 0 \) to find a candidate optimum \((x^*, \lambda^*)\)
- check the positivity of its Hessian \( \nabla^2_x L(x^*, \lambda^*) \) to check if \( x^* \)
  is a min, a max or a saddle point of \( f \).