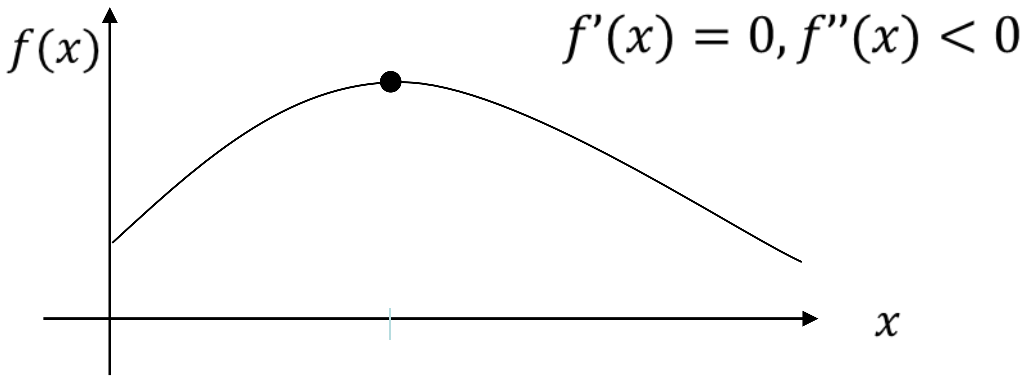
# Unit: Optimality Conditions and Karush Kuhn Tucker Theorem

# Goals

1. What is the Gradient of a function? What are its properties?
2. How can it be used to find a linear approximation of a nonlinear function?
3. Given a continuously differentiable function, which equations are fulfilled for local optima in the following cases?
   1. Unconstrained
   2. Equality Constraints
   3. Inequality Constraints
4. How can this be used to find Pareto fronts analytically?
5. How to state conditions for locally efficient in multiobjective optimization?

# Optimality conditions for differentiable problems

- Given a point on a continuous differentiable function. A sufficient condition for a point $x \in \mathbb{R}$ to be a local maximum is for instance *f'(x)=0, f''(x)<0*:




- *Necessary* conditions, can be used to restrict the set of candidate solutions. (*f'(x)=0)*
- If *sufficient* conditions are met, this implies the solution is locally (Pareto) optimal, so it provides us with verified solutions.

# Recall: Derivatives

$\partial^- f(x)/\partial x = \lim_{\Delta \downarrow 0} \frac{f(x)-f(x-\Delta)}{\Delta}$ (left sided derivative)

$\partial^+ f(x)/\partial x = \lim_{\Delta \downarrow 0} \frac{f(x+\Delta)-f(x)}{\Delta}$ (right sided derivative)

If for some $x$ it holds $\partial^- f(x)/\partial x = \partial^+ f(x)/\partial x$, then we simply write $\partial f(x)/\partial x$ (derivative).

$$
\begin{aligned}
\partial cx/\partial x &= c \\
\partial c/\partial x &= 0 \\
\partial x^p/\partial x &= px^{p-1} \\
\partial \exp(x)/\partial x &= \exp(x) \\
\partial u(v(x))/\partial x &= \partial u/\partial x(v(x))\partial v/\partial x(x) \text{ chain rule} \\
\partial u(x)v(x)/\partial x &= \partial u/\partial x(x)v(x) + \partial v/\partial x(x)u(x) \text{ product rule} \\
\partial \ln(x)/\partial x &= 1/x
\end{aligned}
$$

These rules will be used in the examples that follow.

# Recall: Partial Derivatives

$$\partial^{-} f(x_1,\ldots,x_i,\ldots,x_n)/\partial x_i = \lim_{\Delta \downarrow 0} \frac{f(x_1,\ldots,x_i,\ldots,x_n) - f(x_1,\ldots,x_i - \Delta,\ldots,x_n)}{\Delta}$$

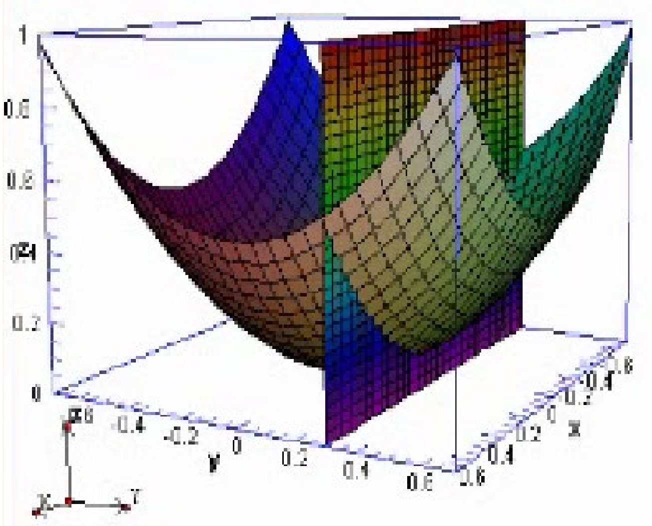
$$\partial^{+} f(x_1,\ldots,x_i,\ldots,x_n)/\partial x_i = \lim_{\Delta \downarrow 0} \frac{f(x_1,\ldots,x_i + \Delta,\ldots,x_n) - f(x_1,\ldots,x_i,\ldots,x_n)}{\Delta}$$

If for some $x_i$ it holds $\partial^{-} f(\mathbf{x})/\partial x_i = \partial^{+} f(\mathbf{x})/\partial x_i$ then we simply write $\partial f(\mathbf{x})/\partial x_i$ (partial derivative).

Example:

$$\frac{\partial (x_1)^2 + 3(x_2)^2 + 4x_1x_2}{\partial x_1} = 2x_1 + 4x_2 \text{ ($x_2$ is treated as a constant)}$$

What about $\frac{\partial (x_1)^2 + 3(x_2)^2 + 4x_1x_2}{\partial x_2}$ ?



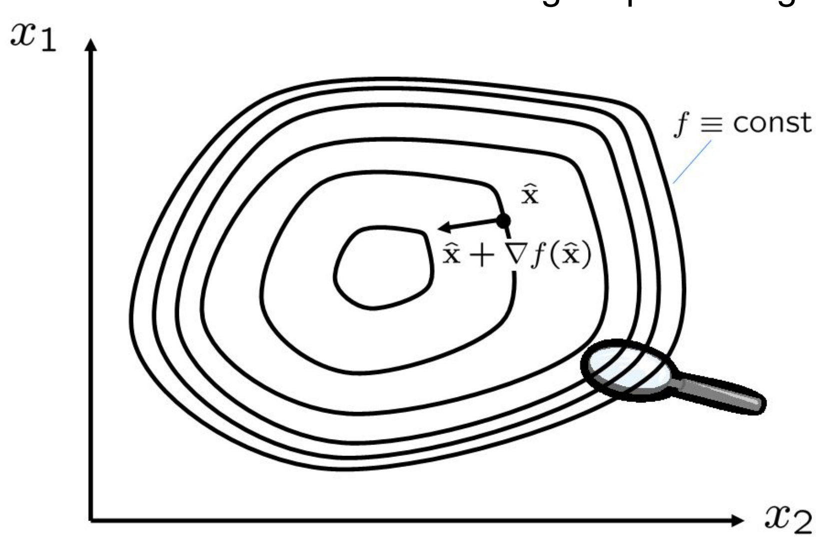
https://www.khanacademy.org/math/multivariable-calculus/partial_derivatives_topic/partial_derivatives/v/partial-derivatives

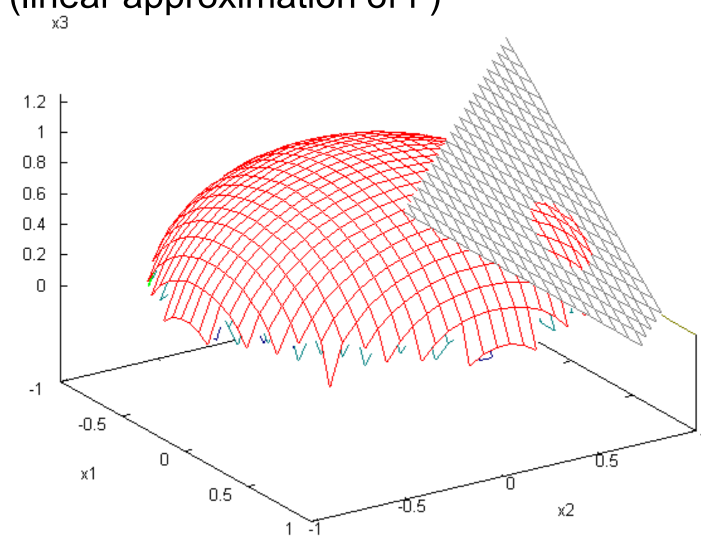
# Gradient / Linear Taylor Approximations

$$\nabla \overbrace{f(\mathbf{x}_0)}^{\text{Gradient of } f \text{ at } \boldsymbol{x}_0} = \left(\frac{\partial f}{\partial x_1}(\mathbf{x}_0), \ldots, \frac{\partial f}{\partial x_n}(\mathbf{x}_0)\right)^\top$$

Continuously differentiable functions can locally be approximated by tangent planes, i.e.

$$\lim_{\mathbf{x} \to \mathbf{x}_0} f(\mathbf{x}) - [\underbrace{f(\mathbf{x}_0) + \nabla f(\mathbf{x}_0)(\mathbf{x} - \mathbf{x}_0)}_{\text{Tangent-plane height at (linear approximation of f )}}] = 0$$



# Example 1-Dimension

In one dimension the gradient is given by $\partial f(x)/\partial x(x) = f'(x)$.

$\lim_{\mathbf{x}\to\mathbf{x}_0} f(\mathbf{x}) - [f(\mathbf{x}_0) + \nabla f(\mathbf{x}_0)(\mathbf{x} - \mathbf{x}_0)] = 0$ specializes to $\lim_{x\to x_0} f(x) - [f(x_0) + f'(\mathbf{x}_0)(x - x_0)] = 0$

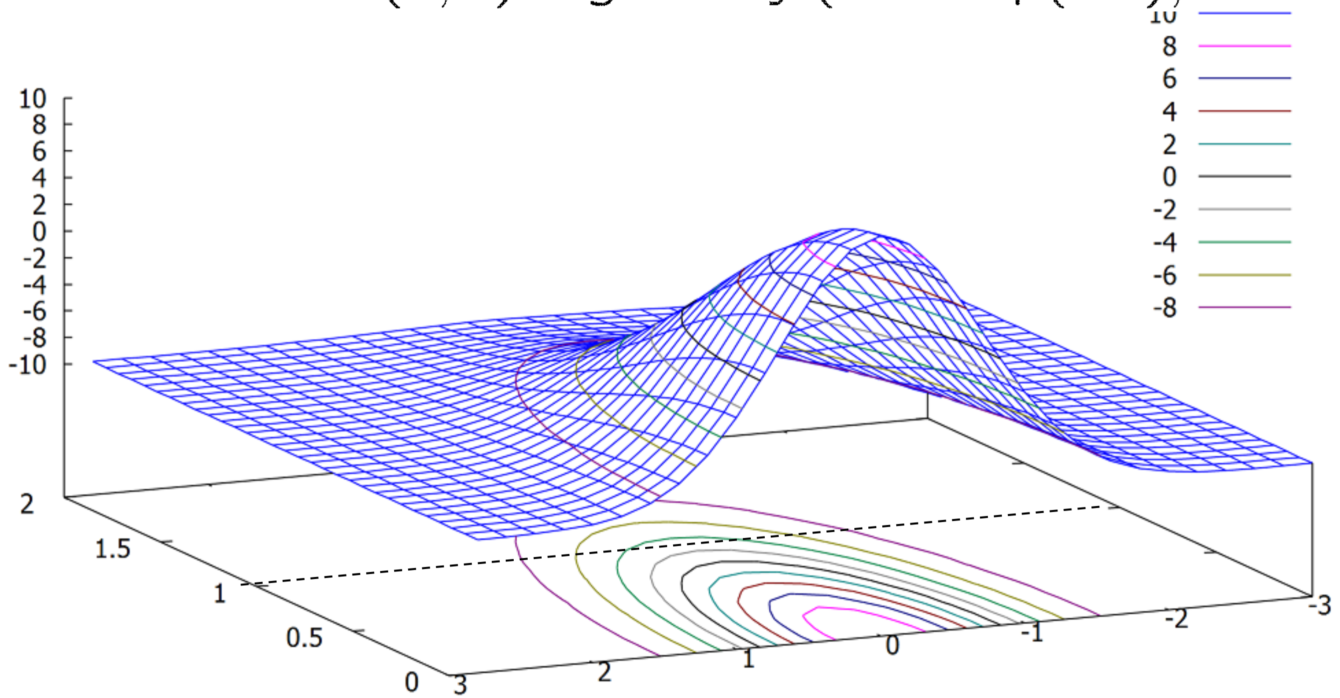
This means that locally:
$f(x) \approx f(x_0) + f'(x_0)(x - x_0)$
Now, the left hand side can be written as
$\underbrace{f(x_0) - f'(x_0)x_0}_{=c} + \underbrace{f'(x_0)}_{=m} x$
and hence $f(x)$ is locally well approximated by a linear function of the form

$$f(x) \approx mx + c$$

# Gradient computation: Example

$$f(\mathbf{x}) = 20\,\mathrm{exp}(-(x_1)^2 - (x_2)^2)$$
$$\partial f/\partial x_1(\mathbf{x}) = -20 * 2x_1\,\mathrm{exp}(-(x_1)^2 - (x_2)^2) \quad \text{chain rule}$$
$$\partial f/\partial x_2(\mathbf{x}) = -20 * 2x_2\,\mathrm{exp}(-(x_1)^2 - (x_2)^2)$$
$$\nabla f(\mathbf{x}) = \begin{pmatrix} -40x_1\,\mathrm{exp}(-(x_1)^2 - (x_2)^2) \\ -40x_2\,\mathrm{exp}(-(x_1)^2 - (x_2)^2) \end{pmatrix}$$

Gradient vector at $(1, 1)$ is given by $(-40\,\mathrm{exp}(-2), -40\,\mathrm{exp}(-2))^\top$:

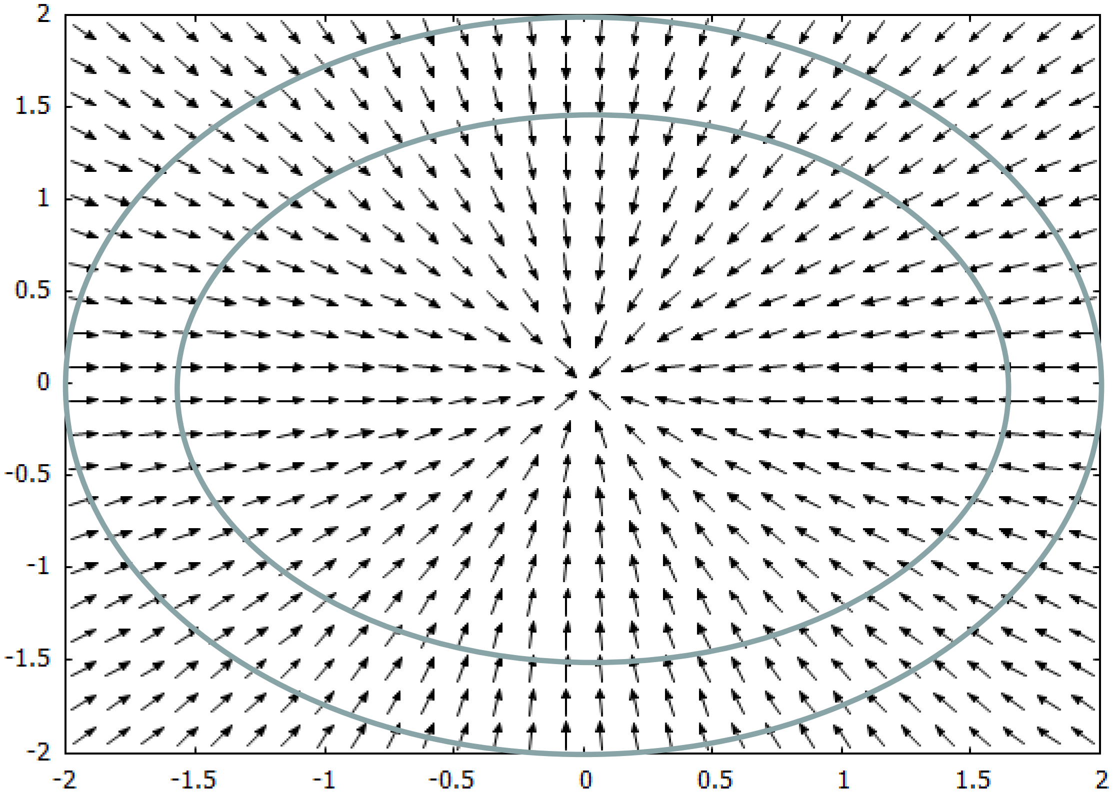



Program: wxMaxima:
load(draw);
draw3d(explicit(20*exp(-x^2-y^2)-10,x,0,2,y,-3,3), contour_levels = 15, contour = both, surface_hide = true);

# Gradient properties

**Theorem:** The gradient $\nabla f(\mathbf{x})$ is perpendicular (orthogonal) to the local tangent (line, plane) of the level curve $L_{=}(f(\mathbf{x}))$.

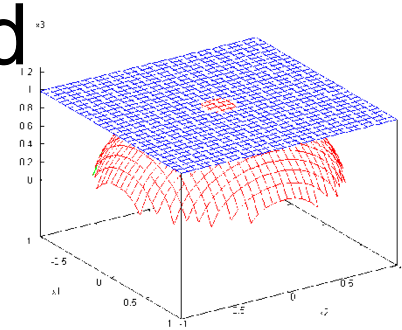
The plot shows the Gradient at different points of the function (Gradient field) $f(x_1,x_2) = 20 \exp(-(x_1)^2 - (x_2)^2)$




Program: wxMaxima:

```
f1(x1,x2):=20*exp(-x1^2-x2^2);
gx1f1(x1,x2):=diff(f1(x1,x2),x1,1);
gx2f1(x1,x2):=diff(f1(x1,x2),x2,1);
load(dfdraw);
drawdf([gx1f1(x1,x2),gx2f1(x1,x2)],[x1,-2,2],[x2,-2,2]);
```

# Single objective, unconstrained



Continuous unconstrained

$$f(\mathbf{x}) \to \min, \quad \mathbf{x} \in \mathbb{R}^n$$

Optimality condition: x is a local minimizer (maximizer), iff

$\nabla f(x) = 0, \nabla^2 f(x)$ positive (negative) semidefinite.

$$\nabla f(\mathbf{x}) = \left(\frac{\partial f}{\partial x_1}, \ldots, \frac{\partial f}{\partial x_n}\right)^\top(\mathbf{x})$$

$$\nabla^2 f(\mathbf{x}) = \left[\frac{\partial^2 f}{\partial x_i \partial x_j}\right](\mathbf{x})_{i=1,\ldots,n, j=1,\ldots,n} \quad \frac{\partial^2 f}{\partial x_i \partial x_j}(\mathbf{x}) = \frac{\frac{\partial f}{\partial x_i}}{\partial x_j}(\mathbf{x})$$

A matrix is positive (semi-)definite if all eigenvalues are positive (non-negative)

Often, as in the case $x^2 + y^2 \to \min$, a lower bound can be obtained and used to argue whether a (stationary) point is a local/global optimum.

# Single objective, unconstrained (example)

Neccesary condition for optimality (differentiable, unconstrained problem):
$\nabla f(\mathbf{x}) = (\frac{\partial f}{\partial x_1}, \ldots, \frac{\partial f}{\partial x_n})^\top = (0, \ldots, 0)^\top$

Example:
$f(x_1, x_2) = 1.1(x_1)^2 + (x_2)^2$
$\partial f / \partial x_1 = 2.2 x_1 = 0$
$\partial f / \partial x_2 = 2 x_2 = 0$
$\Rightarrow x_1 = 0, x_2 = 0.$

Sufficient condition max.:
$\nabla^2 f(\mathbf{x})$ positive definite
$\nabla^2 f(\mathbf{x}) = \begin{pmatrix} 2.2 & 0 \\ 0 & 2 \end{pmatrix}.$
Eigenvalues are:
$\lambda_1 = 2.2$
$\lambda_2 = 2$
$\Rightarrow \nabla^2 f(x)$ is positive definite in $x$
$\Rightarrow \mathbf{x}$ is a local minimizer.

# Constraints (equalities)

$$f(\mathbf{x}) \to \text{min}, \;\; \text{s.t.} \;\; g_1(\mathbf{x}) = 0, \ldots, g_m(\mathbf{x}) = 0$$

All functions are continuously differentiable.

A necessary condition for $\mathbf{x}^*$ to be a local extremum is given, if there exists multipliers $\lambda_1$, ..., $\lambda_{m+1}$ with at least one $\lambda_i \neq 0$ for $i = 1, \ldots, m+1$, such that:

$$\lambda_1 \nabla f(\mathbf{x}^*) + \sum_{i=1}^{m} \lambda_{i+1} \nabla g_i(\mathbf{x}^*) = \mathbf{0}$$

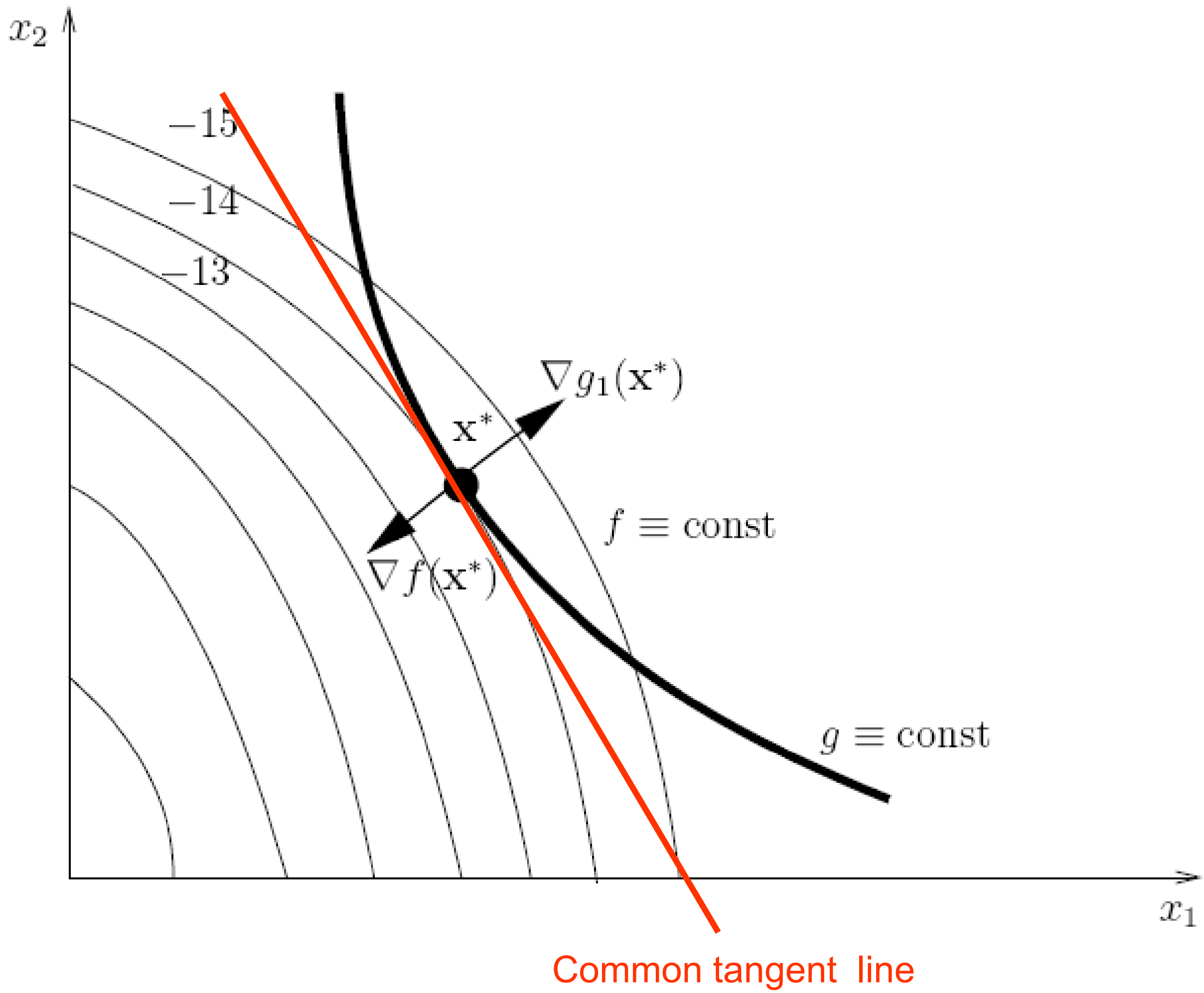
.

The Langrange multipliers $\lambda_i$ are named after Lagrange (1736-1813), who discovered this theorem, but could not prove it. It took 100 years before the proof was found.

Show that this yields m+n equations with m+n+1 unknowns.

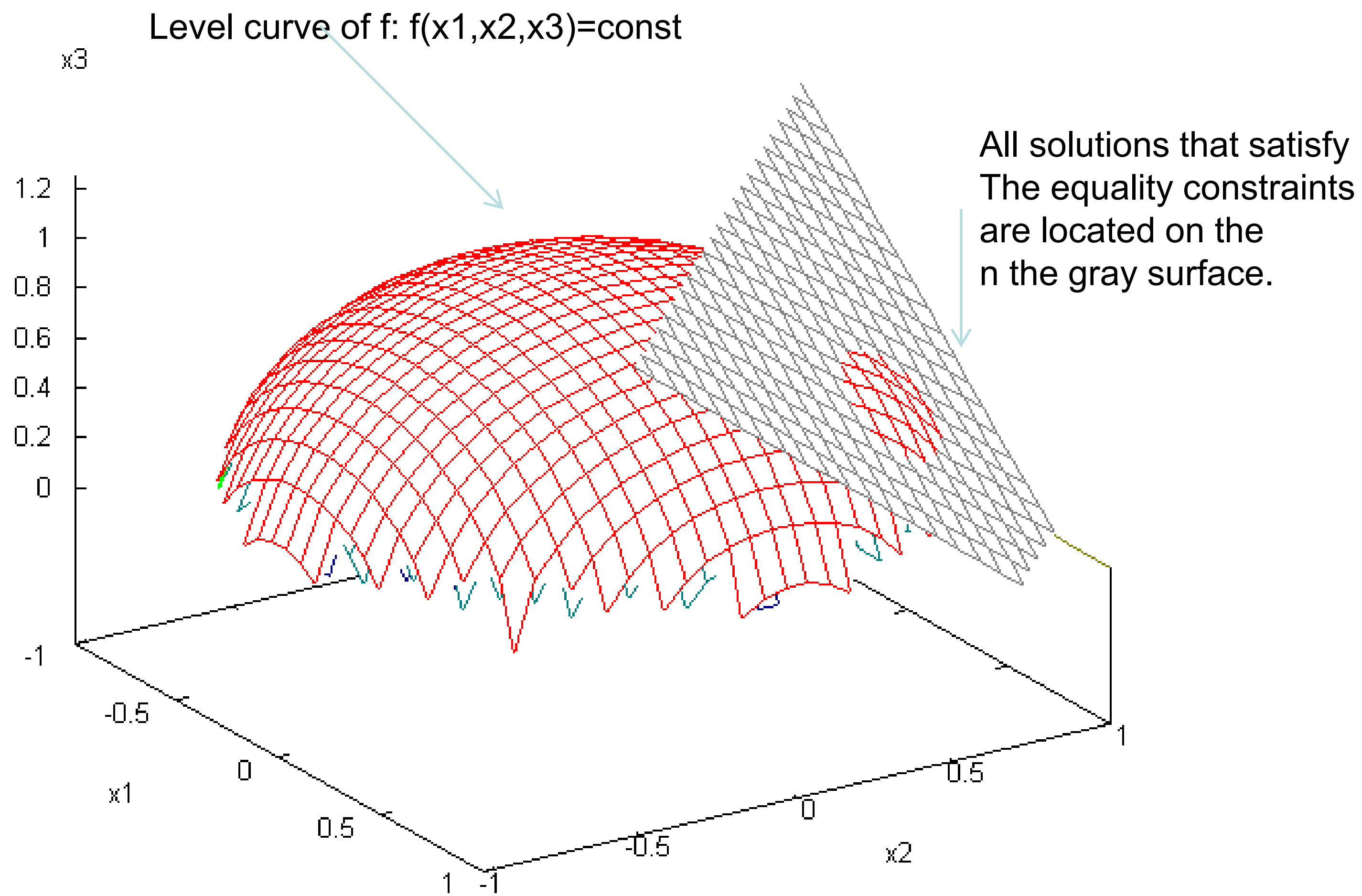
A standard approach is to try $\lambda_1 = 0$ and $\lambda_1 = 1$ (Lagrange multiplier rule).

# Constraints (equalities) - interpretation

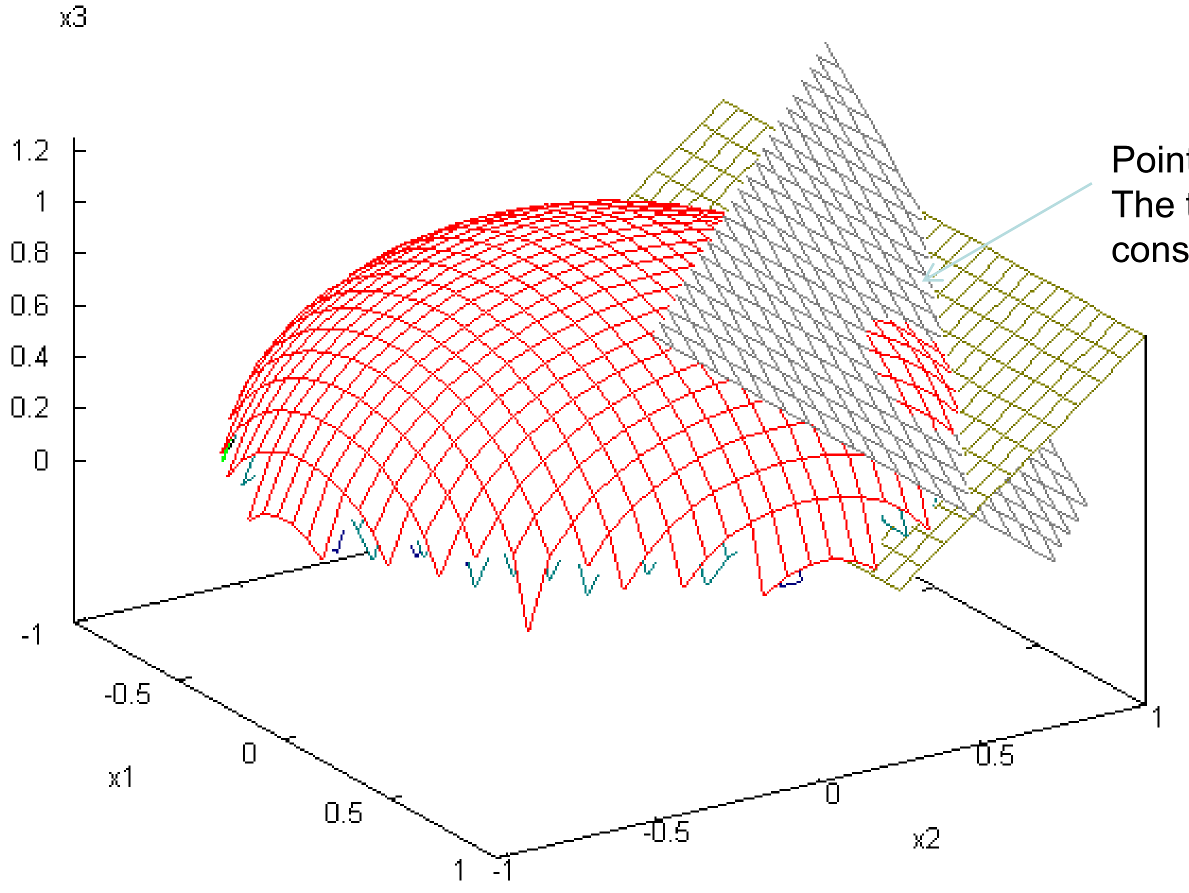



Note that $\nabla f(\mathbf{x})$ is perpendicular to the level curves.

# Example: One equality constraint, three dimensions



# Example: 2 Equality constraints, three dimensions




Points on the intersection of
The two planes satisfy both
constraints.

# Constraints (inequalities)

$f(\mathbf{x}) \to$ min, s.t. $g_1(\mathbf{x}) \leq 0, \ldots, g_m(\mathbf{x}) \leq 0$, all functions are continuously differentiable.

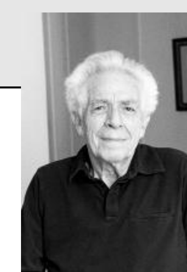
The Karush Kuhn Tucker conditions are said to hold for $\mathbf{x}^*$, if there exists multipliers $\lambda_1 \geq 0$, ..., $\lambda_{m+1} \geq 0$ and at least one $\lambda_i > 0$ for $i = 1, \ldots, m+1$, such that:

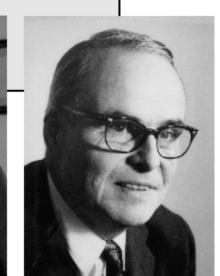
(1) $\lambda_1 \nabla f(\mathbf{x}^*) + \sum_{i=1}^{m} \lambda_{i+1} \nabla g_i(\mathbf{x}^*) = \mathbf{0}$.
(2) $\lambda_{i+1} g_i(\mathbf{x}^*) = 0, i = 1, \ldots, m$

**KKT Theorem - Neccessary conditions for smooth, convex programming:** Assume the objective and all constraint functions are convex in some $\epsilon$-neighborhood of $\mathbf{x}^*$, if $\mathbf{x}^*$ is a local minimum, then there exists $\lambda_1, \ldots, \lambda_{m+1}$ such that KKT conditions are fulfilled.

Harold W. Kuhn
US-American
Mathematician
1924-2014

Albert William Tucker
Canadian
Mathematician,
1905-1995





# Recall: Polyhedral cones

Def.: Polyhedral cone: A polyhedral cone in $\mathbb{R}^m$is determined by a number of $k$ direction vectors $\mathbf{d_1} \in \mathbb{R}^m$, ... $\mathbf{d}_k \in \mathbb{R}^m$ (cone generators). It is the set that comprises all positive linear combinations of these vectors: $C = \{\mathbf{y} \in \mathbb{R}^m \quad | \text{ exists } \lambda_1 \geq 0, \ldots, \lambda_k \geq 0 : \mathbf{y} = \sum_{i=1}^{k} \lambda_i \mathbf{d}_i\}$

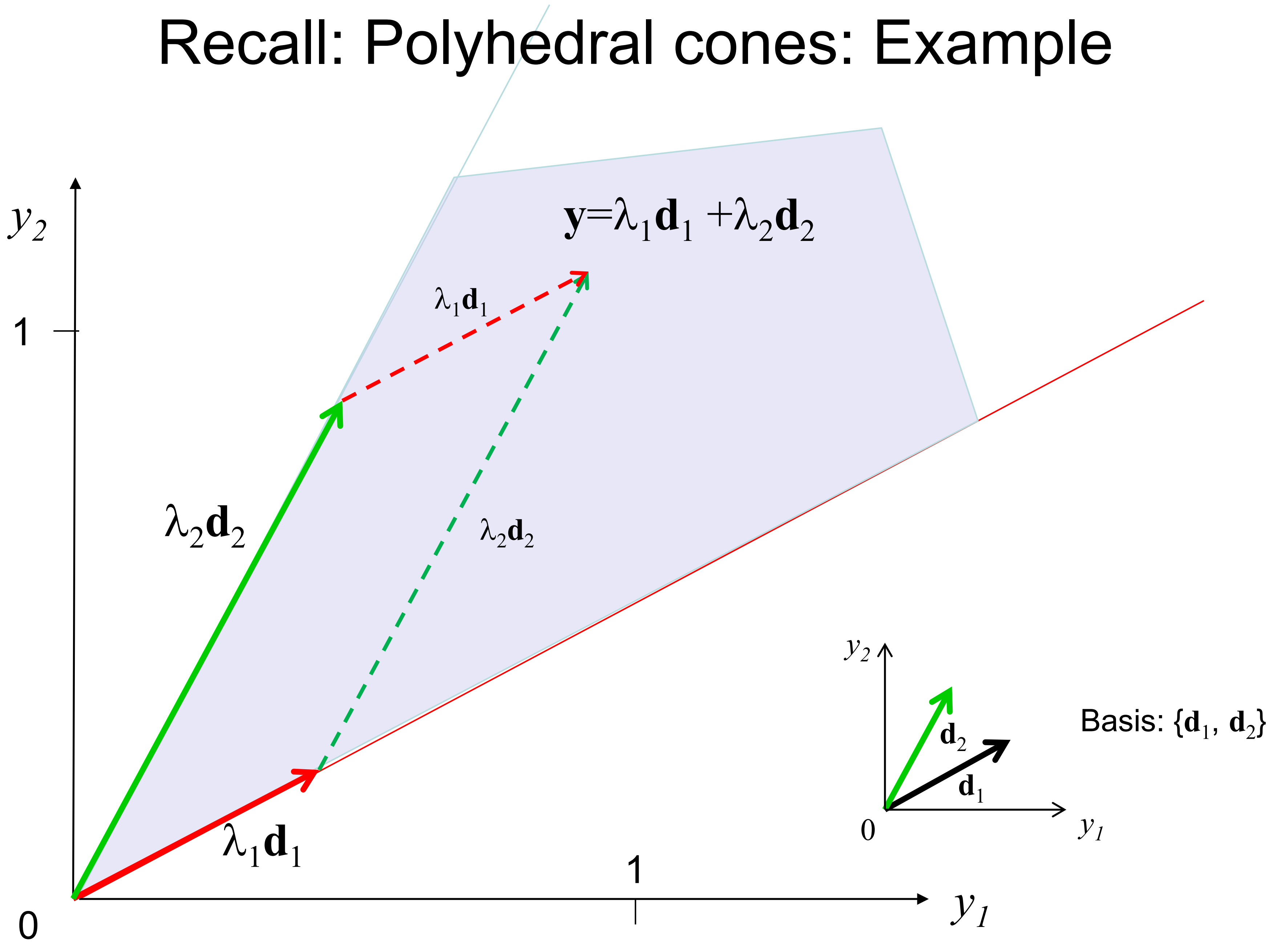
Example: Is $\mathbf{y} \in \mathbf{R}^m$ inside a polyhedral cone given by some linear independent directions $\mathbf{d}_1, \ldots, \mathbf{d}_m$?

Answer: Solve linear equation system:

$$\begin{aligned} y_1 &= \lambda_1 d_{11} + \lambda_2 d_{12} + \ldots + \lambda_m d_{1m} \\ &\vdots \\ y_m &= \lambda_1 d_{m1} + \lambda_2 d_{m2} + \ldots + \lambda_m d_{mm} \end{aligned}$$

If the solution vector $(\lambda_1, \ldots, \lambda_m) \geq 0$ then $\mathbf{y}$ lies inside the cone.

# Recall: Polyhedral cones: Example



# Constraint (inequality)

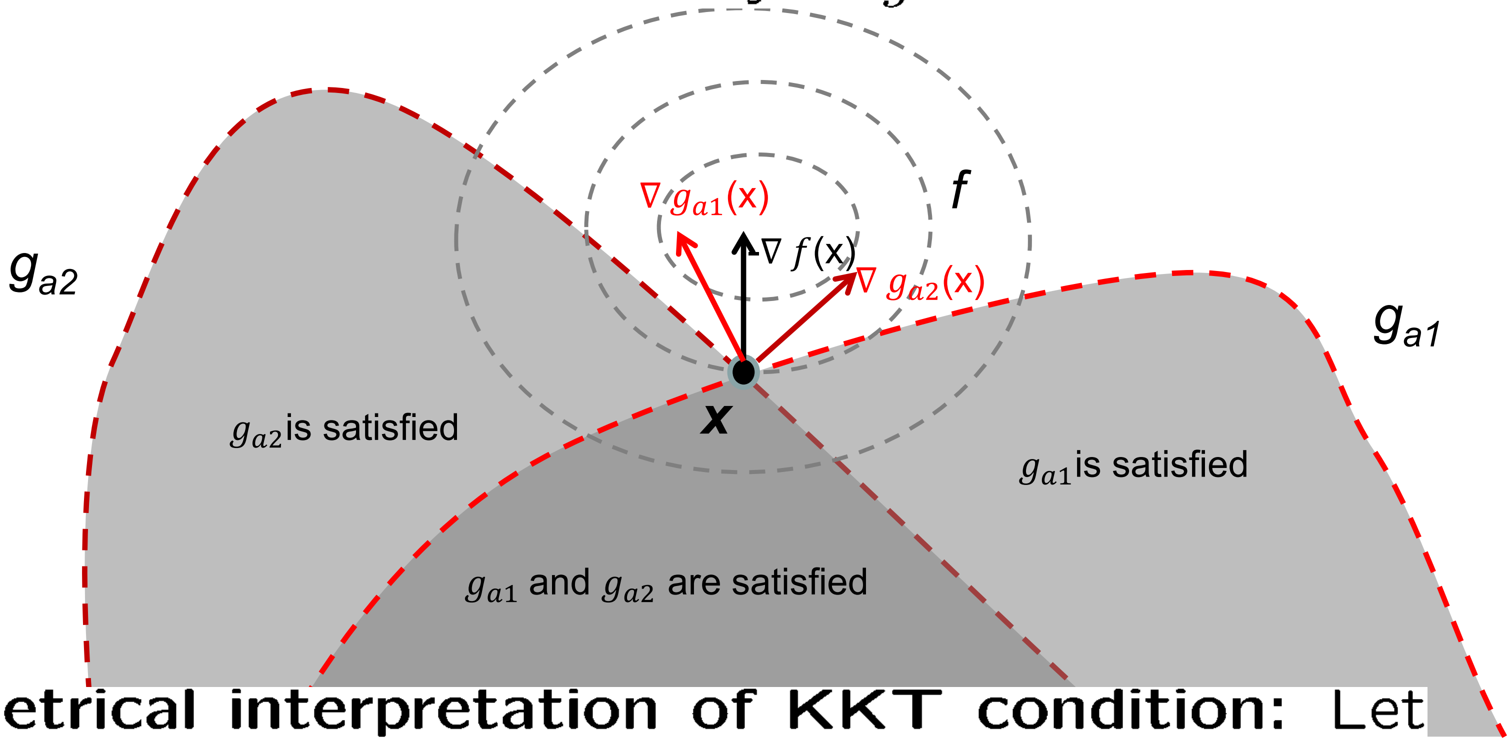
As in the case of Lagrange multiplier, we get $m+n$ non-linear equations, the solution of which results in candidate solutions.

> The KKT conditions are sufficient for optimiality, provided $\lambda_1 = 1$. In this case $\mathbf{x}^*$ is a local minimum.
>
> Note that if $\mathbf{x}^*$ is in the interior of the feasible region (a Slater point), all $g_i(\mathbf{x}) < 0$ and thus $\lambda_1 > 0$.

[Brinkhuis, Tikhomirov, 2005]

# Geometrical interpretation KKT conditions

A constraint function $g$ is called **active** in $\mathbf{x}$ if $g(\mathbf{x}) = 0$, i.e. $\mathbf{x}$ is located at the boundary of $g$.




**Geometrical interpretation of KKT condition:** Let $a_1, \ldots, a_k$ denote the indexes for the constraint functions that are active in $\mathbf{x}$. Then $-\nabla f(\mathbf{x})$ lies in the cone spanned by $\nabla g_{a_1}(\mathbf{x}), \ldots, \nabla g_{a_k}(\mathbf{x})$.

(Recall: Convex polyhedral cone: $\{\mathbf{y} | \ \mathbf{y} = \sum_{i=1}^{n} \lambda_i \mathbf{d}_i\}$ for $\lambda_i \geq 0$ and $\mathbf{d}_i$ a set of vectors 'spanning' the cone.)

# Multiobjective Optimization [cf. Miettinnen '99]

**Fritz John neccessary conditions**

A neccessary condition for $\mathbf{x}^*$ to be a locally efficient point is that there exists vectors $\lambda_1, \ldots, \lambda_k$ and $\upsilon_1, \ldots, \upsilon_m$ such that

(0) $\lambda \succ \mathbf{0}, \upsilon \succ \mathbf{0}$
(1) $\sum_{i=1}^{k} \lambda_i \nabla f_i(\mathbf{x}^*) - \sum_{i=1}^{m} \upsilon_i \nabla g_i(\mathbf{x}^*) = \mathbf{0}$.
(2) $\upsilon_i g_i(\mathbf{x}^*) = 0, i = 1, \ldots, m$

**Karush Kuhn Tucker sufficient conditions for a solution to be Pareto optimal:** Let $\mathbf{x}^*$ be a feasible point. Assume that all objective functions are locally convex and all constraint functions are locally concave, and the Fritz John conditions hold in $\mathbf{x}^*$, then $\mathbf{x}^*$ is a local efficient point.

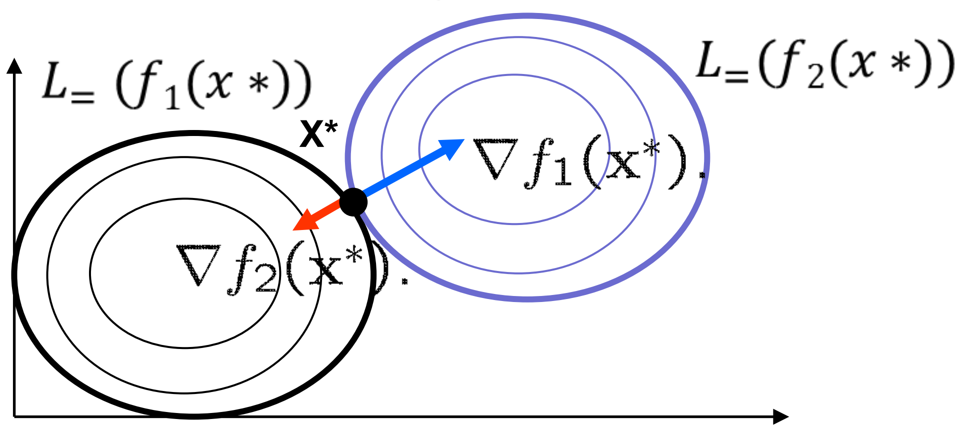
# Unconstrained Multiobjective Optimization

In the unconstrained case Fritz John neccessary conditions reduce to

There exist numbers $\lambda_1, \ldots, \lambda_k \in \mathbb{R}$, such that

(1) $\lambda \succ \mathbf{0}$
(2) $\sum_{i=1}^{k} \lambda_i \nabla f_i(\mathbf{x}^*) = \mathbf{0}$.

$\mathbf{x}^*$ is optimum for some linear scalarization with some weights $\lambda_1, \ldots, \lambda_k$.




In 2-dimensional spaces this criterion reduces to the observation, that either one of the objectives has a zero gradient ( neccesary condition for ideal points) or the gradients are parallel.

# Strategy: Solve multiobjective optimization problems by level set continuation

Recall: KKT conditions, unconstrained case: For efficient $x$ there exists $\lambda_1, \ldots, \lambda_m$, such that $\lambda \succ \mathbf{0}$ and $\sum_{i=1}^{m} \lambda_i \nabla f_i(\mathbf{x}^*) = \mathbf{0}$.

Strategy: Solve equation system with $m+n$ unknowns and $n+1$ equations:

$$\sum_{i=1}^{m} \lambda_i \nabla f_i(\mathbf{x}) = 0$$

$$\sum_{i=1}^{m} \lambda_i = 1$$

$$\lambda_i \in \mathbb{R}_0^+, i = 1, \ldots, m$$

Yields $m-1$ dimensional manifold for regular problems. Strategy: Find one point using optimization and extend surface by predictor-corrector method (Hillermeier 2001).

# Strategy: Find efficient points using determinant

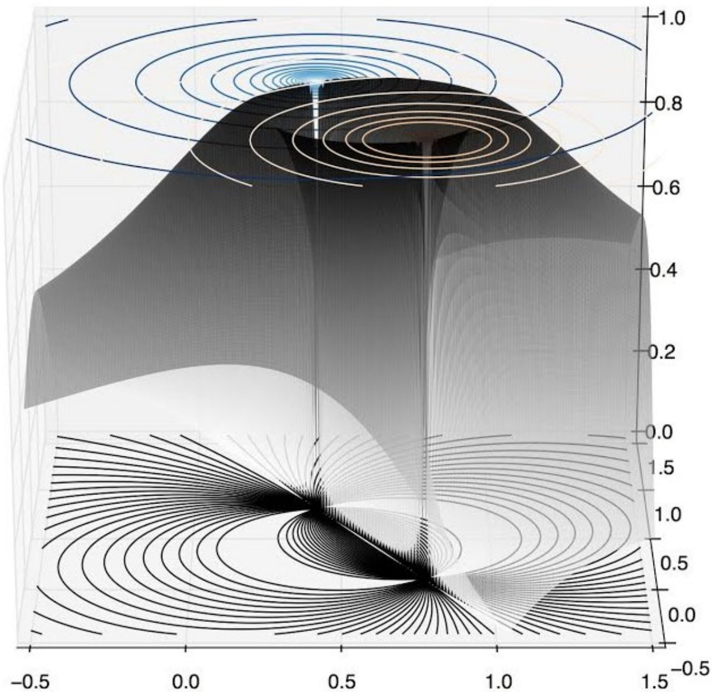
- $\exists\, \lambda$: $\lambda \nabla f_1(x) + \nabla f_2(x) = 0$

- Linearly dependent ⇒
  $$\det([\nabla f_1, \nabla f_2]) = 0$$

- Efficient points can be found by searching for points with

  $$\left(\det(\nabla f_1, \nabla f_2)\right)^2 \rightarrow min$$

  (necessary condition)



# Take home messages

1. Gradient is a vector of first order partial derivatives that is perpendicular to level curves; Hessian contains second order partial derivatives.
2. Local linearization yields optimality condition; in single objective case ‘gradient zero’ and positive/negative definite Hessian.
3. Lagrange multiplier rule can be used to solve constrained optimization problems with equality constraints.
4. KKT conditions generalize it to inequality constraints; negative gradient points in cone spanned by active constraints.
5. KKT conditions for multiobjective optimization require for interior points to be optimal that they have gradients which point in exactly the opposite directions.
6. KKT conditions define equation system the solution of which is an at most m-1 dimensional manifold

# References


- Kuhn, Harold W., and Albert W. Tucker. "Nonlinear programming." *Proceedings of the second Berkeley symposium on mathematical statistics and probability*. Vol. 5. 1951.
- Miettinen, Kaisa. *Nonlinear multiobjective optimization*. Vol. 12. Springer, 1999.
- Miettinen, Kaisa. "Some methods for nonlinear multi-objective optimization."*Evolutionary Multi-Criterion Optimization*. Springer Berlin Heidelberg, 2001.
- Hillermeier, C. (2001). Generalized homotopy approach to multiobjective optimization. *Journal of Optimization Theory and Applications*, *110*(3), 557-583.
- Schütze, O., Coello Coello, C. A., Mostaghim, S., Talbi, E. G., & Dellnitz, M. (2008). Hybridizing evolutionary strategies with continuation methods for solving multi-objective problems. *Engineering Optimization*, *40*(5), 383-402.