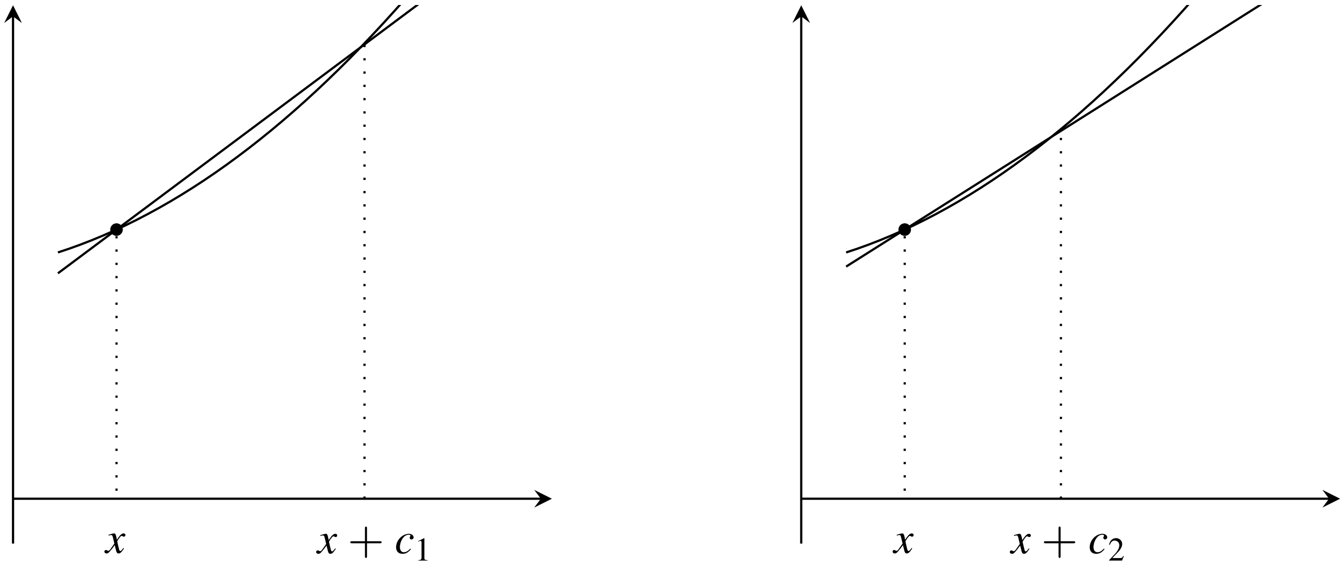
FINDING THE DERIVATIVE



- We fix a constant $c > 0$ and draw a straight line through the points $(x, f(x))$ and $(x + c, f(x + c))$. The slope of that line is

$$\frac{f(x+c) - f(x)}{c}$$

- Now make c smaller and smaller: Choose $c_1 > c_2 > \ldots$, for example $c_n = \frac{1}{n}$.
- We then ask what happens as c gets infinitely small, i.e. we try to find the limit

$$\lim_{n \to \infty} \frac{f(x + c_n) - f(x)}{c_n}$$

- If f is differentiable, this limit exists, and its slope is exactly that of the best possible linear approximation. That is, the limit is $f'(x)$.
- If the limit does not exist, f is not differentiable at x.

Not examinable.

SUMMARY

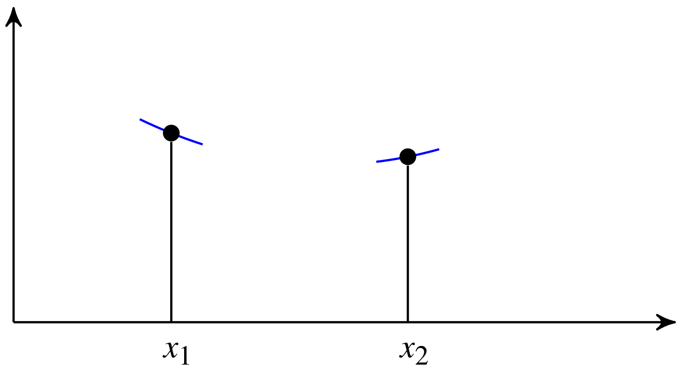
The derivative of a function f at a point x is the the slope of the locally best linear approximation to f around x.

If you are not familiar with calculus, keep in mind:

- The derivative $f'(x)$ exists if f is "sufficiently smooth" at x.
- Sign: The derivative is positive if f increases at x, negative if it decreases, and 0 if f is a maximum or minimum.
- Magnitude: The absolute value $|f'(x)|$ is the larger the more rapidly f changes around x.

BACK TO OPTIMIZATION

Recall that we had asked: How can would we find a minimum if we had access to the entire function in a small neighborhood around points $x_1, x_2, \ldots$ that we are allowed to choose?



- If we can compute the derivatives $f'(x_1)$ and $f'(x_2)$, we have (the slope of) linear approximations to f at both points that are locally exact.
- That is: We can substitute the derivatives for the two short blue lines in the figure.
- We can tell from the sign of the derivative in which direction the function decreases.
- We also know that $f'(x) = 0$ if x is a minimum.

Minimization Strategy

Basic idea

Start with some point x_0. Compute the derivative $f'(x_0)$ at x. Then:

- "Move downhill": Choose some $c > 0$, and set $x_1 = x_0 + c$ if $f'(x_0) < 0$ and $x_1 = x_0 - c$ if $f'(x_0) > 0$.
- Compute $f'(x_1)$. If it is 0 (possibly a minimum), stop.
- Otherwise, move downhill from x_1, etc.

Observations

- Since the sign of f' is determined by whether f increases or decreases, we can summarize the case distinction above by setting

$$x_1 = x_0 - \text{sign}(f'(x_0)) \cdot c$$

- If f changes rapidly, it may be a good strategy to make a large step (choose a large c), since we presumably are still far from the minimum. If f changes slowly, c should be small.
- One way of doing so is to choose c as the magnitude of f', since $|f'|$ has exactly this property. In that case:

$$x_1 = x_0 - \text{sign}(f'(x_0)) \cdot |f'(x_0)| = x_0 - f'(x_0)$$

The algorithm obtained by replying this step repeatedly is called **gradient descent**.

Gradient Descent

Gradient descent searches for a minimum of a differentiable function f.

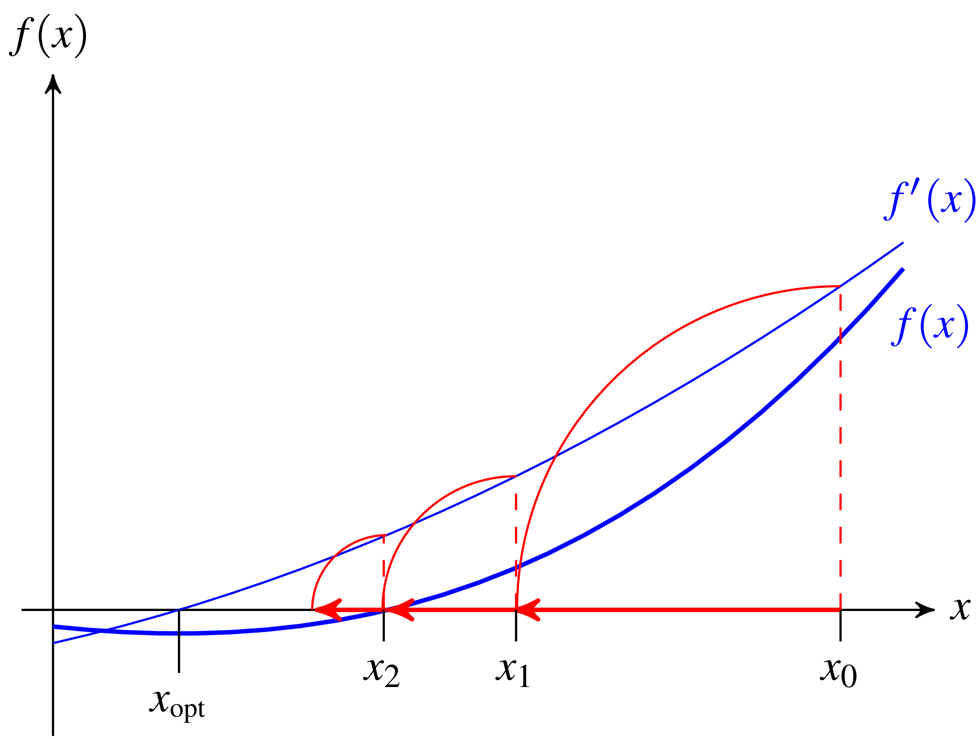
Algorithm

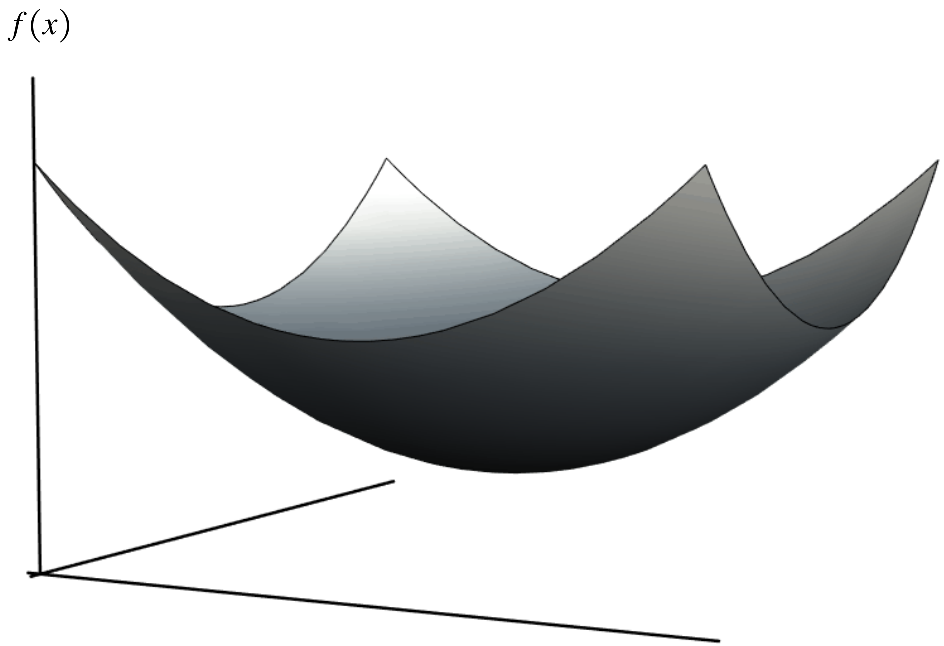
Start with some point $x_0 \in \mathbb{R}$ and fix a precision $\varepsilon > 0$.

Repeat for $n = 1, 2, \ldots$:

1. Check whether $|f'(x_n)| < \varepsilon$. If so, report the solution $x^* := x_n$ and terminate.
2. Otherwise, set

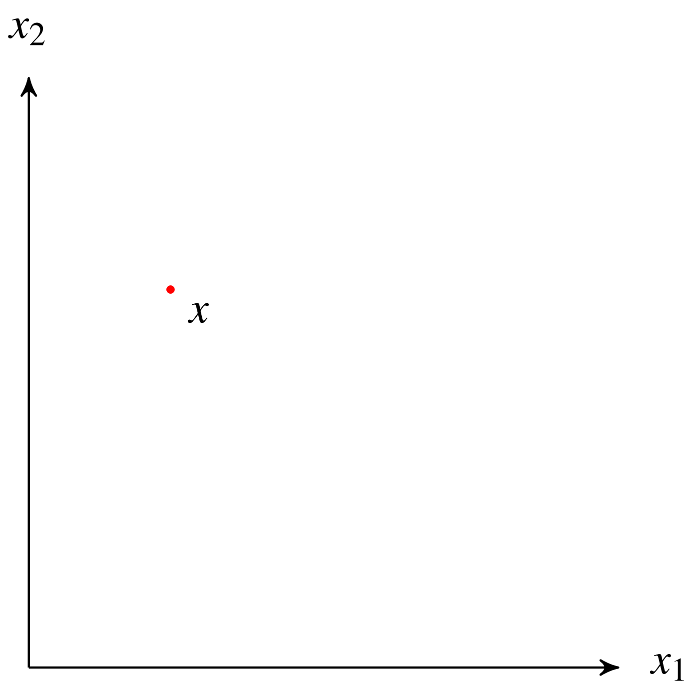
$$x_{n+1} := x_n - f'(x_n)$$

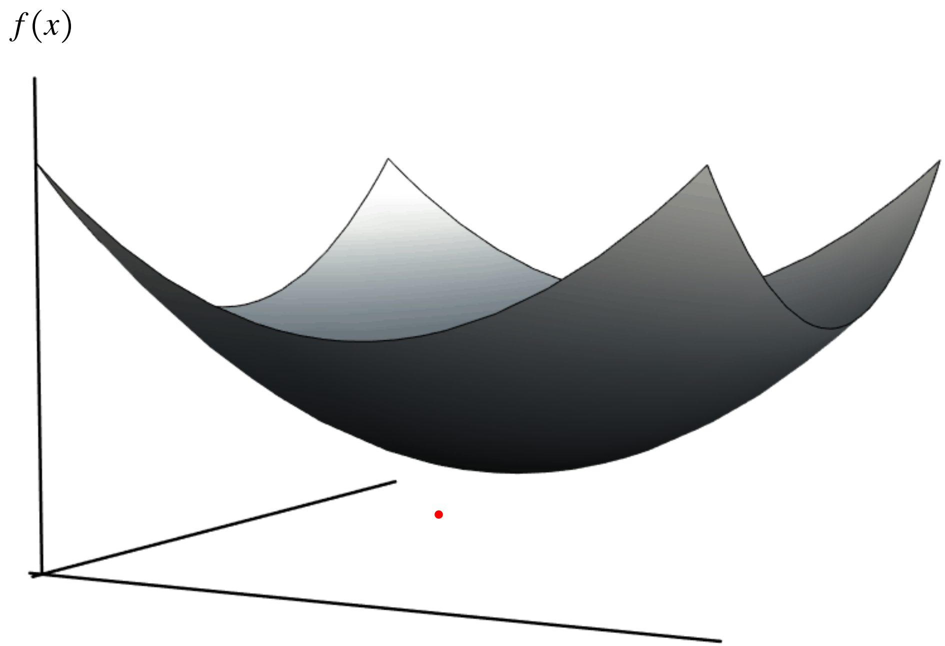




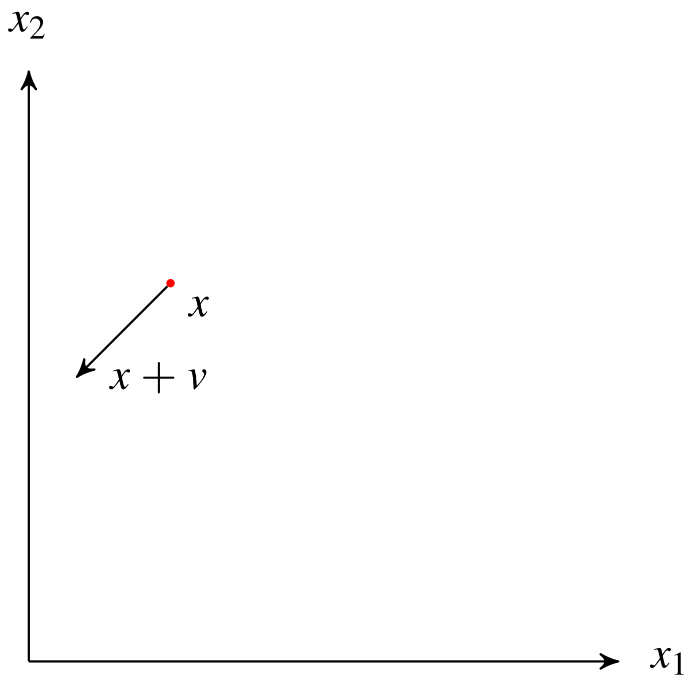
- We now ask how to define a derivative in multiple dimensions.
- Consider a function $f : \mathbb{R}^d \to \mathbb{R}$. What is the derivative of f at a point x?
- For simplicity, we assume $d = 2$ (so that we can plot the function).

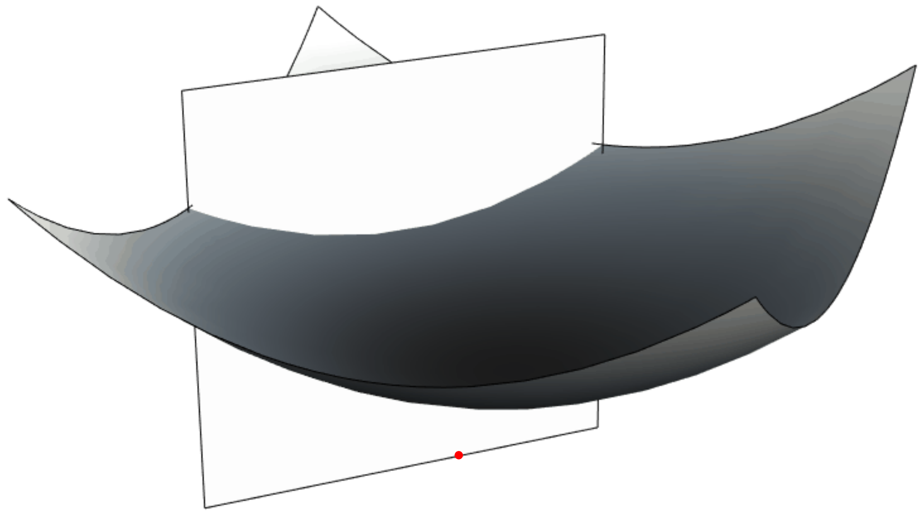
Derivatives in Multiple Dimensions



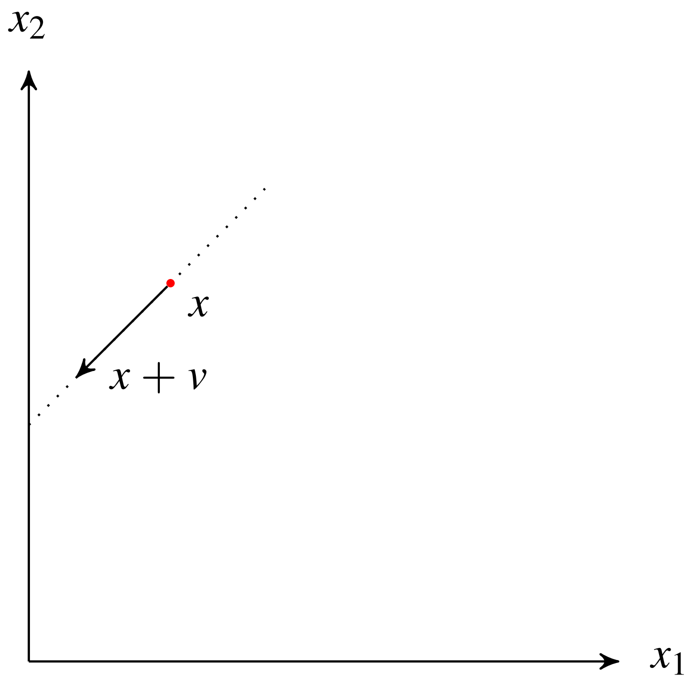


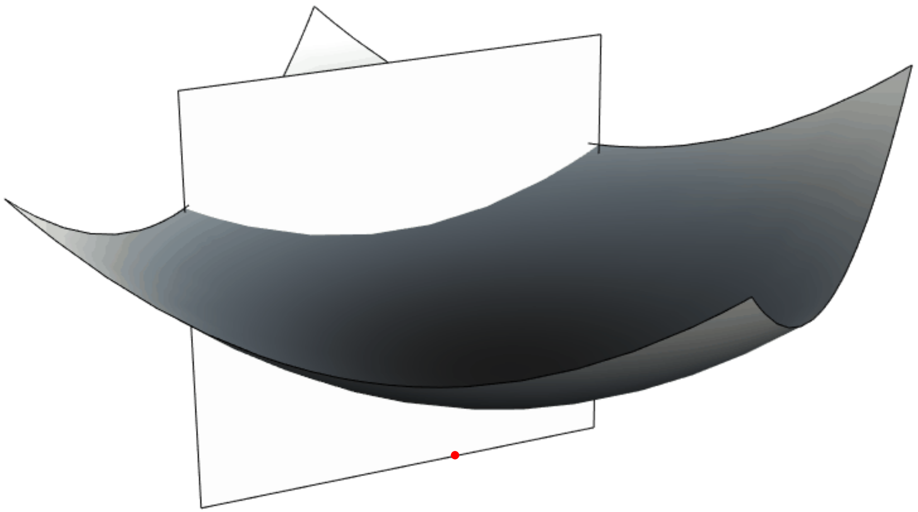
- We fix a point $x = (x_1, x_2)$ in $\mathbb{R}^2$, marked red above.
- We will try to turn this into a 1-dimensional problem, so that we can use the definition of a derivative we already know.





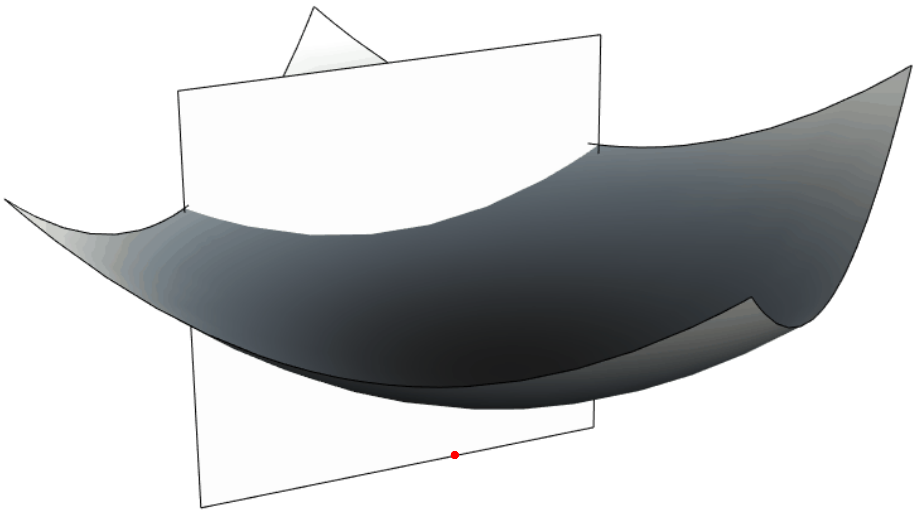
- To make the problem 1-dimensional, fix some vector $v \in \mathbb{R}_2$, and draw a line through x in direction of v.
- Then intersect f with a plane given by this line: In the coordinate system of f, choose the plane that contains the line and is orthogonal to $\mathbb{R}^2$.
- The plane contains the point x.
- Note we can do that even if $d > 2$. We still obtain a plane.

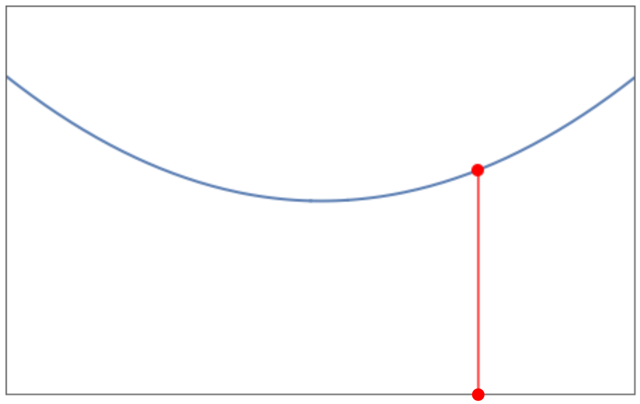




- To make the problem 1-dimensional, fix some vector $v \in \mathbb{R}_2$, and draw a line through x in direction of v.
- Then intersect f with a plane given by this line: In the coordinate system of f, choose the plane that contains the line and is orthogonal to $\mathbb{R}^2$.
- The plane contains the point x.
- Note we can do that even if $d > 2$. We still obtain a plane.

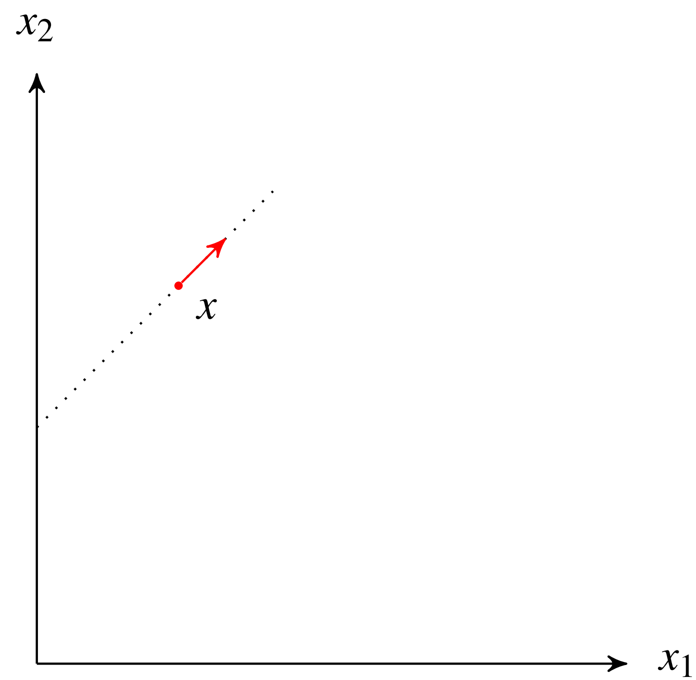
REDUCING TO ONE DIMENSION





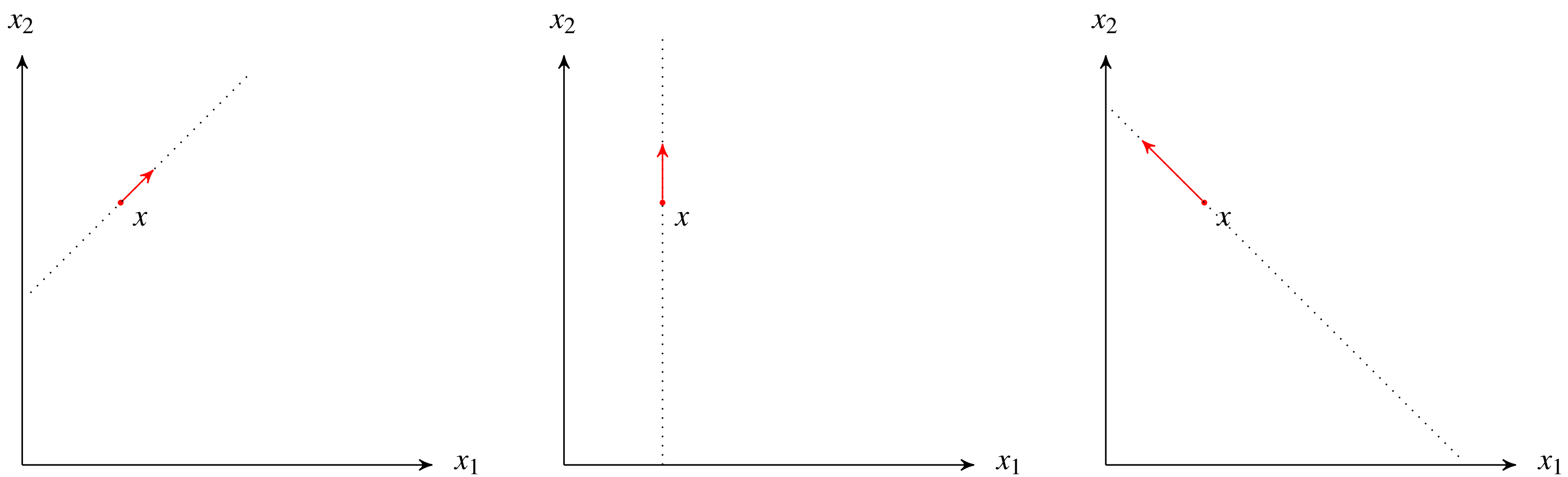
- The intersection of f with the plane is a 1-dimensional function f_H, and x corresponds to a point x_H in its domain.
- We can now compute the derivative f'_H of f_H at x_H. The idea is to use this as the derivative of f at x.

Back to Multiple Dimensions



- In the domain of f, we draw a vector from x *in direction of* H such that:
  1. The vector is oriented to point in the direction in which f_H increases.
  2. Its length is the value of the derivative $f'_H(x)$.
- That completely determines the vector (shown in red above).
- There is one problem still to be solved: f_H depends on H, that is, on the direction of the vector v. Which direction should we use?

THE GRADIENT



- We now rotate the plane H around x. For each position of the plane, we get a new derivative $f'_H(x)$, and a new red vector.
- We choose the plane for which f'_H is largest:

$$H^* := \arg \max_{\text{all rotations of } H} f'_H(x)$$

  Provided that f_H is differentiable for all H, one can show that this is always unique (or $f'_H(x)$ is zero for all H).
- We then define the vector

$$\nabla f(x) := \text{ vector given by } H^* \text{ as above}$$

  The vector $\nabla f(x)$ is called the **gradient** of f at x.

PROPERTIES OF THE GRADIENT

The gradient $\nabla f(x)$ of $f : \mathbb{R}^d \to \mathbb{R}$ at a point $x \in \mathbb{R}^d$ is a vector in the domain $\mathbb{R}^d$ in the direction in which f most rapidly increases at x.

- The length of the gradient measures steepness: The more rapidly f increases at x, the larger $\|\nabla f(x)\|$.
- The gradient has length 0 if x is a maximum or minimum of f. A constant function has gradient of length 0 at every point x.
- The gradient operation is linear:

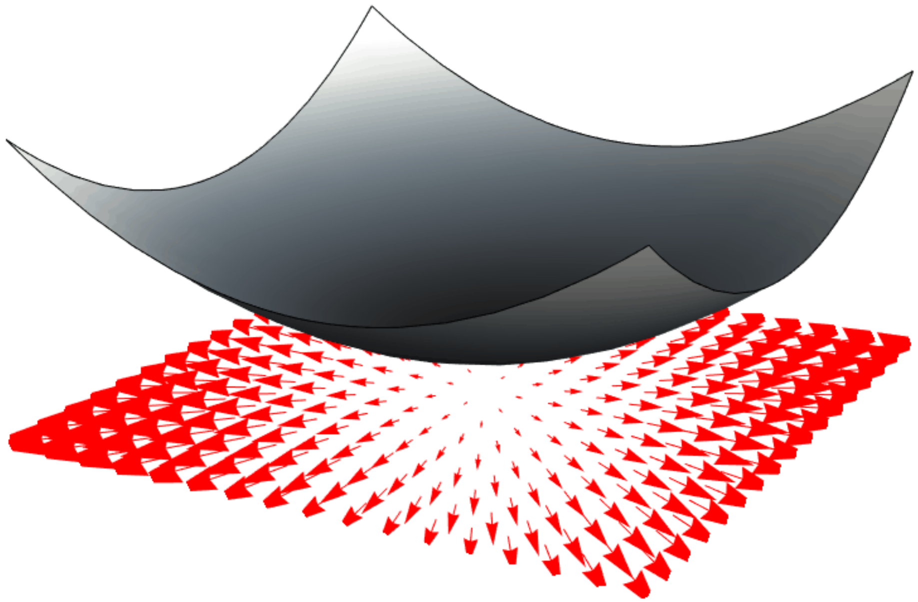
$$\nabla(\alpha f(x) + \beta g(x)) = \alpha \nabla f(x) + \beta \nabla g(x)$$

Gradients and Contour Lines

- Recall that a contour line (or contour set) of f is a set of points along which f remains constant,
$$C[f,c] := \{x \in \mathbb{R}^d \,|\, f(x) = c\} \qquad \text{for some } c \in \mathbb{R}.$$
- One can show that if $C[f,c]$ contains x, the gradient at x is orthogonal to the contour:
$$\nabla f(x) \perp C[f,c] \qquad \text{if } x \in C[f,c]\;.$$
- Intuition: The gradient points in the direction of maximal *local* change, whereas $C[f,c]$ is a direction in which there is no change. Locally, these two are orthogonal.

Gradients are orthogonal to contour lines.

Gradients and Contour Lines



- For this parabolic function, all contour lines are concentric circles around the minimum.
- The picture above shows the gradients plotted at various points in the plane.

$$f : \mathbb{R}^d \to \mathbb{R}$$

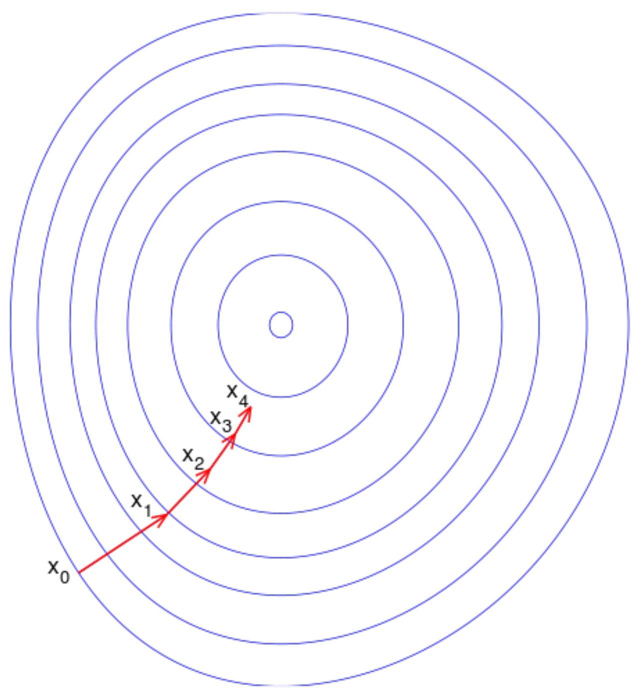
Algorithm

Start with some point $x_0 \in \mathbb{R}$ and fix a precision $\varepsilon > 0$.

Repeat for $n = 1, 2, \ldots$:

1. Check whether $\|\nabla f(x_n)\| < \varepsilon$. If so, report the solution $x^* := x_n$ and terminate.
2. Otherwise, set

$$x_{n+1} := x_n - \nabla f(x_n)$$



Gradient Descent

$$f : \mathbb{R}^d \to \mathbb{R}$$

Algorithm

Start with some point $x_0 \in \mathbb{R}$ and fix a precision $\varepsilon > 0$.

Repeat for $n = 1, 2, \dots$:

1. Check whether $\|\nabla f(x_n)\| < \varepsilon$. If so, report the solution $x^* := x_n$ and terminate.
2. Otherwise, set
$$x_{n+1} := x_n - \alpha(n) \nabla f(x_n)$$

Here, $\alpha(n) > 0$ is a coefficient that may depend on n. It is called the **step size** in optimization, or the **learning rate** in machine learning.