Topic(s): gradients, local approximations to a function, tangent planes

Consider $f : \mathbb{R}^n \to \mathbb{R}$. The **gradient** of f is defined to be

$$\nabla f = \begin{bmatrix} f_{x_1} \\ f_{x_2} \\ \vdots \\ f_{x_n} \end{bmatrix}.$$

For $f: \mathbb{R} \to \mathbb{R}$ and x near a, linear approximation is given by

$$f(x) \approx f(a) + f'(a)(x-a).$$

The gradient of f is a vector-valued function from $\mathbb{R}^n$ to $\mathbb{R}^n$. For $\mathbf{x}$ near $\mathbf{a} \in \mathbb{R}^n$, the **linear approximation** to f is

$$\begin{aligned} f(\mathbf{x}) &\approx f(\mathbf{a}) + (\nabla f(\mathbf{a})) \cdot (\mathbf{x} - \mathbf{a}) \\ &= f(\mathbf{a}) + f_{x_1}(\mathbf{a})(x_1 - a_1) + f_{x_2}(\mathbf{a})(x_2 - a_2) + \cdots + f_{x_n}(\mathbf{a})(x_n - a_n), \end{aligned}$$

displacements in $x_1, x_2, \ldots, x_n$.

where $\mathbf{x} = \begin{bmatrix} x_1 \\ \vdots \\ x_n \end{bmatrix}$ and $\mathbf{a} = \begin{bmatrix} a_1 \\ \vdots \\ a_n \end{bmatrix}$.

For example, if $n = 2$, then $f(x, y) \approx f(a, b) + f_x(a, b)(x - a) + f_y(a, b)(y - b)$.

Example 1. Consider the function $f(x, y) = x^3 + 2x^2y + 4xy^2 - y^3$ near the point $\mathbf{a} = (2, 0)$. Use linear approximation to approximate $f(2.5, -0.5)$, $f(2.05, -0.05)$, and $f(2.005, -0.005)$.

We see that $f_x = 3x^2 + 4xy + 4y^2$ and $f_y = 2x^2 + 8xy - 3y^2$. So,

$$f(\vec{a}) = 8, \quad f_x(\vec{a}) = 12, \quad f_y(\vec{a}) = 8.$$

Hence,

$$\begin{aligned} f(2.5, -0.5) &\approx 8 + 12(2.5-2) + 8(-0.5-0) = \boxed{10} \\ f(2.05, -0.05) &\approx 8 + 12(2.05-2) + 8(-0.05-0) = \boxed{8.2} \\ f(2.005, -0.005) &\approx 8 + 12(2.005-2) + 8(-0.005-0) = \boxed{8.02} \end{aligned}$$

<u>Note</u> The actual values are

$$\begin{aligned} f(2.5, -0.5) &= 12 \\ f(2.05, -0.05) &= 8.2155 \\ f(2.005, -0.005) &= 8.02015 \end{aligned}$$

The linear approximations get better as $\vec{x}$ gets closer to $\vec{a}$.

Theorem 11.2.1. Let $f : \mathbb{R}^2 \to \mathbb{R}$ be a function and suppose that $\nabla f(a, b) \neq \mathbf{0}$.

1. The gradient $\nabla f(a, b)$ is *perpendicular* to the level set of f that goes through (a, b); more precisely, the gradient is perpendicular to the tangent line to the level curve. The gradient points in the direction of maximal increase for f for (x, y) moving away from (a, b).

2. The mathematical statement of the statement above is: the equation of the line tangent to the level curve of f passing through (a, b) is

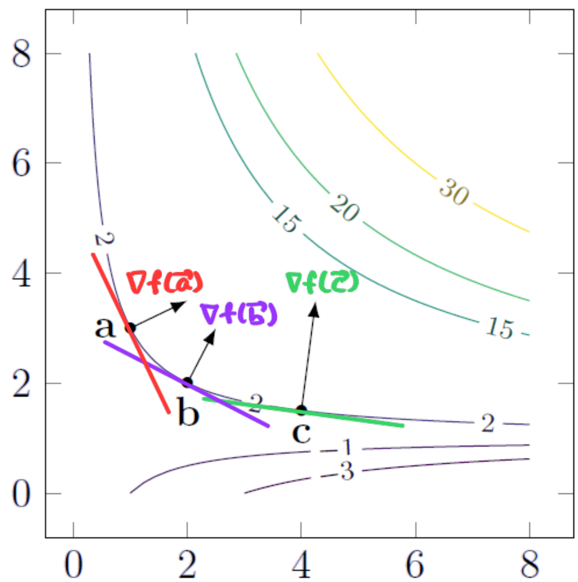
$(\nabla f(\vec{a}))\cdot(\vec{x}-\vec{a})=0$

$$\nabla f(a, b) \cdot \begin{bmatrix} x - a \\ y - b \end{bmatrix} = 0.$$

More explicitly, this equation is

$$f_x(a, b)(x - a) + f_y(a, b)(y - b) = 0.$$

Example 2. Consider the function $f(x, y) = xy - x$. The contour plot is shown below; in particular, let us consider the level curve at 2. The gradients at $\mathbf{a} = (1, 3)$, $\mathbf{b} = (2, 2)$, and $\mathbf{c} = \left(4, \frac{3}{2}\right)$ are shown.



We compute

$$\nabla f = \begin{bmatrix} f_x \\ f_y \end{bmatrix} = \begin{bmatrix} y-1 \\ x \end{bmatrix}.$$

Hence,

$$\nabla f(\vec{a}) = \begin{bmatrix} 2 \\ 1 \end{bmatrix}$$

$$\nabla f(\vec{b}) = \begin{bmatrix} 1 \\ 2 \end{bmatrix}$$

$$\nabla f(\vec{c}) = \begin{bmatrix} 1/2 \\ 4 \end{bmatrix}$$

Example 3. For $f(x, y) = \sqrt{1 + xy}$, use linear approximation to estimate the value of $f(1.1, -0.2)$.

$\vec{x} = \begin{bmatrix} 1.1 \\ -0.2 \end{bmatrix}$ is near $\vec{a} = \begin{bmatrix} 1 \\ 0 \end{bmatrix}$, and $f(\vec{a}) = 1$. The gradient is

$$\nabla f = \begin{bmatrix} \frac{y}{2\sqrt{1+xy}} \\ \frac{x}{2\sqrt{1+xy}} \end{bmatrix},$$

and so, $\nabla f(\vec{a}) = \begin{bmatrix} 0 \\ \frac{1}{2} \end{bmatrix}$. Thus,

$$f(\vec{x}) \approx f(\vec{a}) + \nabla f(\vec{a}) \cdot (\vec{x} - \vec{a}) = 1 + \begin{bmatrix} 0 \\ \frac{1}{2} \end{bmatrix} \cdot \begin{bmatrix} 0.1 \\ -0.2 \end{bmatrix} = 0.9.$$

<u>Note</u> The actual value is $f(\vec{x}) = \sqrt{1 + (1.1)(-0.2)} = \sqrt{0.78} = 0.88318.$

Theorem 11.2.2. For a function $f : \mathbb{R}^3 \to \mathbb{R}$ and $\mathbf{a}$ for which $\nabla f(\mathbf{a}) \neq \mathbf{0}$, the gradient vector is perpendicular to the plane tangent to the level set of f through $\mathbf{a}$. In particular, this tangent plane has the equation

$$\nabla f(a_1, a_2, a_3) \cdot \begin{bmatrix} x - a_1 \\ y - a_2 \\ z - a_3 \end{bmatrix} = 0.$$

If $\vec{x}$ is on the tangent plane,
$\nabla f(\vec{a}) \cdot (\vec{x} - \vec{a}) = 0$

As a special case, the graph of a function $h : \mathbb{R}^2 \to \mathbb{R}$ is the surface S with equation $z = h(x, y)$ that is the level set $f = 0$ of $f(x, y, z) = z - h(x, y)$ whose gradient $(-h_x, -h_y, 1)$ never vanishes (since the third entry is always nonzero). The tangent plane to S at $(a, b, h(a, b))$ then has the equation

$$\underbrace{\begin{bmatrix} -h_x(a, b) \\ -h_y(a, b) \\ 1 \end{bmatrix}}_{\nabla f(a,b,h(a,b))} \cdot \begin{bmatrix} x - a \\ y - b \\ z - h(a, b) \end{bmatrix} = 0,$$

which is equivalent to

$$z = h(a, b) + h_x(a, b)(x - a) + h_y(a, b)(y - b).$$

Example 4. Consider the ellipse defined by $4x^2 + 9y^2 = 36$. Find the tangent line to the ellipse at the point $\left(\frac{3\sqrt{2}}{2}, \sqrt{2}\right)$. $= \vec{a}$

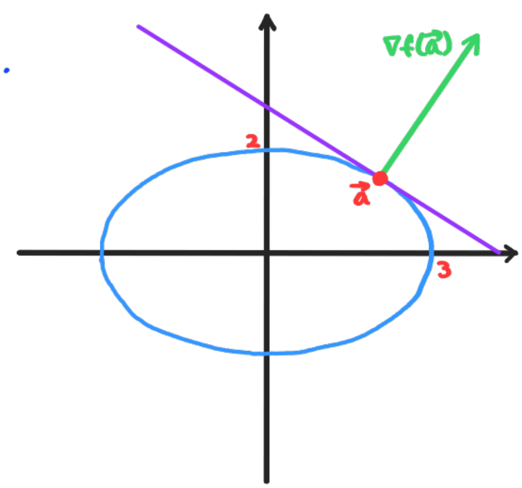
Level set of $f(x,y) = 4x^2 + 9y^2$ at 36

We see that $\nabla f = \begin{bmatrix} 8x \\ 18y \end{bmatrix}$ and $\nabla f(\vec{a}) = \begin{bmatrix} 12\sqrt{2} \\ 18\sqrt{2} \end{bmatrix}$.

$$\begin{bmatrix} 12\sqrt{2} \\ 18\sqrt{2} \end{bmatrix} \cdot \begin{bmatrix} x - \frac{3\sqrt{2}}{2} \\ y - \sqrt{2} \end{bmatrix} = 0$$

yields $12\sqrt{2}x - 36 + 18\sqrt{2}y - 36 = 0$, which reduces to

$$\boxed{2x + 3y = 6\sqrt{2}}.$$



Example 5. Consider the sphere S given by the equation $x^2 + y^2 + z^2 = 14$. Find an equation for the tangent plane to S through the point $(3, 2, 1)$. $= \vec{a}$

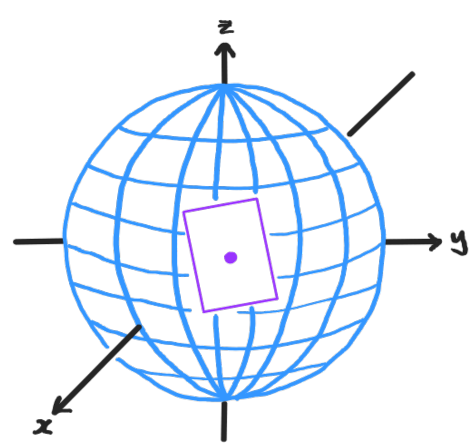
Level set of $f(x,y,z) = x^2 + y^2 + z^2$ at 14

Since $\nabla f = \begin{bmatrix} 2x \\ 2y \\ 2z \end{bmatrix}$, $\nabla f(\vec{a}) = \begin{bmatrix} 6 \\ 4 \\ 2 \end{bmatrix}$.

$$\begin{bmatrix} 6 \\ 4 \\ 2 \end{bmatrix} \cdot \begin{bmatrix} x - 3 \\ y - 2 \\ z - 1 \end{bmatrix} = 0$$

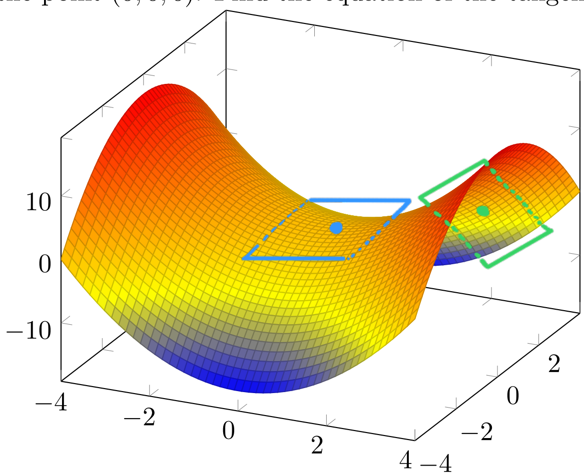
gives us $6(x-3) + 4(y-2) + 2(z-1) = 0$, which reduces to

$$\boxed{3x + 2y + z = 14}.$$



Level set of $f(x,y,z) = z-(x^2-y^2)$ at 0

Example 6. Consider the surface S defined by $z = x^2 - y^2$. Find the equation of the tangent plane to S at the point $(0, 0, 0)$. Find the equation of the tangent plane to S at the point $(2, 1, 3)$. $= \vec{a}$



$$\nabla f = \begin{bmatrix} -2x \\ 2y \\ 1 \end{bmatrix} \Rightarrow \nabla f(\vec{0}) = \begin{bmatrix} 0 \\ 0 \\ 1 \end{bmatrix} \text{ and } \nabla f(\vec{a}) = \begin{bmatrix} -4 \\ 2 \\ 1 \end{bmatrix}.$$

Tangent plane at $\vec{0}$:

$$\begin{bmatrix} 0 \\ 0 \\ 1 \end{bmatrix} \cdot \begin{bmatrix} x \\ y \\ z \end{bmatrix} = 0 \Rightarrow \boxed{z = 0}.$$

Tangent plane at $\vec{a}$:

$$\begin{bmatrix} -4 \\ 2 \\ 1 \end{bmatrix} \cdot \begin{bmatrix} x-2 \\ y-1 \\ z-3 \end{bmatrix} = 0 \Rightarrow \boxed{-4x + 2y + z = -3}.$$

One of the main reasons to study multivariable calculus is to optimize multivariable functions. However, realistic problems of this type cannot be solved exactly; so we need a numerical way to approximate the answer. One powerful method of doing this is the **gradient descent**. The main idea is

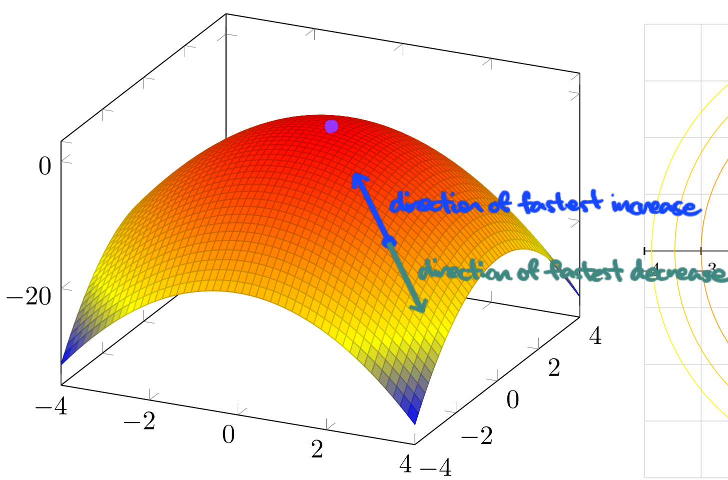
1. start at some point (x, y)
2. move away from (x, y) in the direction in which f decreases the fastest
3. rinse, lather, and repeat

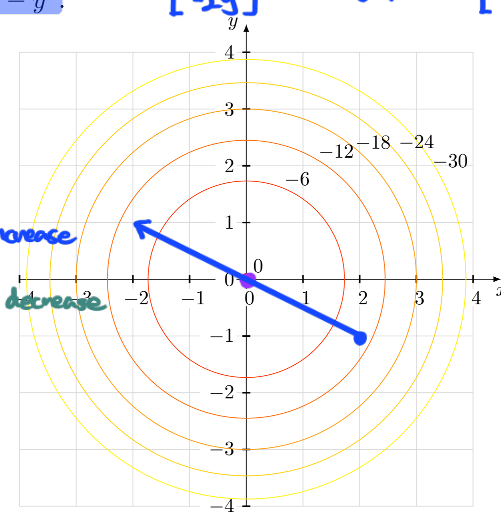
This process will end once we arrive at a local minimum (if there is one). We can modify the "decreases" to "increases" in the second step to find a local maximum; this process is called the **gradient ascent**.

Theorem 11.3.2. Let $f : \mathbb{R}^n \to \mathbb{R}$ be a function and $\mathbf{a} \in \mathbb{R}^n$ is a point at which the gradient $\nabla f(\mathbf{a})$ is non-zero. Then, the unit vector associated with the gradient, $\dfrac{\nabla f(\mathbf{a})}{\|\nabla f(\mathbf{a})\|}$, is the direction in which f increases most rapidly at $\mathbf{a}$. Similarly, the opposite unit vector, $-\dfrac{\nabla f(\mathbf{a})}{\|\nabla f(\mathbf{a})\|}$, is the direction in which f decreases most rapidly at $\mathbf{a}$.

Example 7. Consider the surface defined by $\underbrace{z}_{f(x,y)} = -x^2 - y^2$.

$$\nabla f = \begin{bmatrix} -2x \\ -2y \end{bmatrix} \Rightarrow \nabla f(2,-1) = \begin{bmatrix} -4 \\ 2 \end{bmatrix}$$





Each step of gradient descent (resp. ascent) is moving from $\mathbf{a}$ to

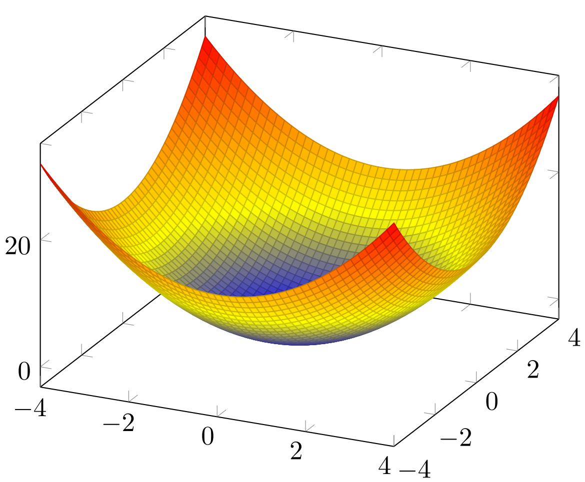
$$\mathbf{a} + t\nabla f(\mathbf{a}),$$

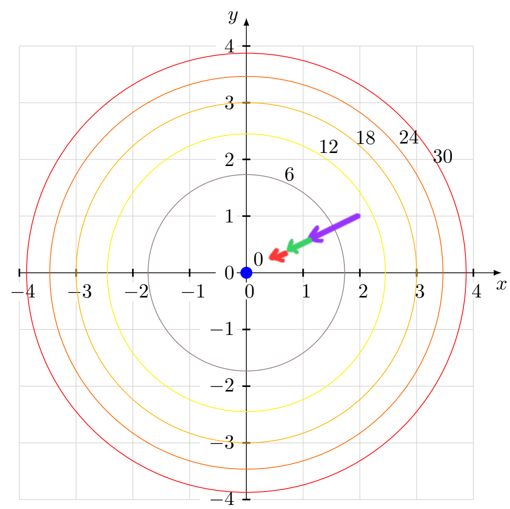
where t is a small negative (resp. positive) number. In order to use this algorithm, we need to decide two things:

↳ $t\nabla f(\vec{a})$ is in a decreasing direction.

1. the first $\mathbf{a}$ – the starting point of our gradient descent
2. t – how far do we go at each step? In the context of machine learning, t is called the **learning rate**.

Example 8. Consider the surface defined by $z = x^2 + y^2$. Start at $(2, 1)$ and apply three steps of gradient descent with $t = -0.2$.





We see that $\nabla f = \begin{bmatrix} 2x \\ 2y \end{bmatrix}$.

Step 1: $\begin{bmatrix} 2 \\ 1 \end{bmatrix} + (-0.2)\begin{bmatrix} 4 \\ 2 \end{bmatrix} = \begin{bmatrix} 1.2 \\ 0.6 \end{bmatrix}$

Step 2: $\begin{bmatrix} 1.2 \\ 0.6 \end{bmatrix} + (-0.2)\begin{bmatrix} 2.4 \\ 1.2 \end{bmatrix} = \begin{bmatrix} 0.72 \\ 0.36 \end{bmatrix}$

Step 3: $\begin{bmatrix} 0.72 \\ 0.36 \end{bmatrix} + (-0.2)\begin{bmatrix} 1.44 \\ 0.72 \end{bmatrix} = \begin{bmatrix} 0.432 \\ 0.216 \end{bmatrix}$

Example 9. Consider the function $f : \mathbb{R}^2 \to \mathbb{R}$ given by $f(x, y) = x^2 - y^2$.

(a) Calculate the first two steps of gradient descent using $t = -0.1$ and starting at $\begin{bmatrix} 2 \\ 0 \end{bmatrix}$, as well as starting at $\begin{bmatrix} 2 \\ 0.6 \end{bmatrix}$. Plot these; do you notice any difference?

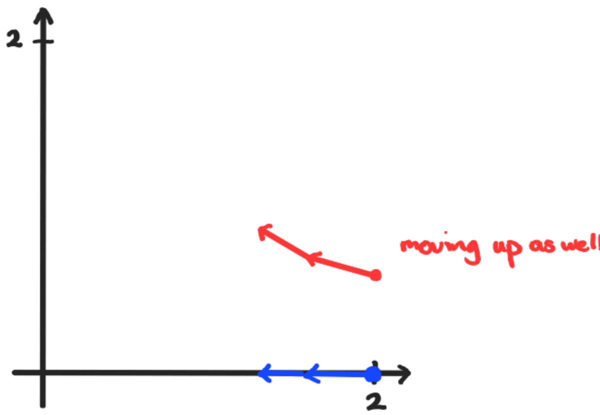
$$\nabla f = \begin{bmatrix} 2x \\ -2y \end{bmatrix}.$$

① $\begin{bmatrix} 2 \\ 0 \end{bmatrix} + (-0.1)\begin{bmatrix} 4 \\ 0 \end{bmatrix} = \begin{bmatrix} 1.6 \\ 0 \end{bmatrix}$

② $\begin{bmatrix} 1.6 \\ 0 \end{bmatrix} + (-0.1)\begin{bmatrix} 3.2 \\ 0 \end{bmatrix} = \begin{bmatrix} 1.28 \\ 0 \end{bmatrix}$

① $\begin{bmatrix} 2 \\ 0.6 \end{bmatrix} + (-0.1)\begin{bmatrix} 4 \\ -1.2 \end{bmatrix} = \begin{bmatrix} 1.6 \\ 0.72 \end{bmatrix}$

② $\begin{bmatrix} 1.6 \\ 0.72 \end{bmatrix} + (-0.1)\begin{bmatrix} 3.2 \\ -1.44 \end{bmatrix} = \begin{bmatrix} 1.28 \\ 0.864 \end{bmatrix}$



(b) For general $\mathbf{a} = \begin{bmatrix} a \\ b \end{bmatrix}$, where do we wind up after a step of gradient descent with $t = -0.1$ starting at $\mathbf{a}$? Your answer should be a vector whose entries are expressed in terms of a and b.

$$\begin{bmatrix} a \\ b \end{bmatrix} + (-0.1)\begin{bmatrix} 2a \\ -2b \end{bmatrix} = \begin{bmatrix} 0.8a \\ 1.2b \end{bmatrix}$$

(c) Using your formula in part (b), if you start at a general $\mathbf{a} = \begin{bmatrix} a \\ b \end{bmatrix}$, feed the procedure into itself repeatedly to say where one winds up after 2 steps and after 3 steps (in terms of a and b).

$$\begin{bmatrix} a \\ b \end{bmatrix} \to \begin{bmatrix} 0.8a \\ 1.2b \end{bmatrix} \to \begin{bmatrix} (0.8)^2 a \\ (1.2)^2 b \end{bmatrix} \to \begin{bmatrix} (0.8)^3 a \\ (1.2)^3 b \end{bmatrix}$$

(d) Explain why iterating gradient descent repeatedly will converge to $\mathbf{0}$ (*not* a local minimum for f, but rather a saddle point) when we start at any point $\mathbf{a}$ with $b = 0$, but will always diverge when $b \neq 0$. (Hint: for any number $0 < c < 1$, the powers c^n converge to 0 as n grows. Use this with $c = 0.8$)

Starting at $\begin{bmatrix} a \\ b \end{bmatrix}$, we get to $\begin{bmatrix} (0.8)^n a \\ (1.2)^n b \end{bmatrix}$ after n steps.

As $n \to \infty$, $(0.8)^n \to 0$ and $(1.2)^n \to \infty$.

Hence, if $b = 0$, gradient descent converges to $\vec{0}$.

However, if $b \neq 0$, gradient descent "falls off" towards $+\infty$ or $-\infty$ in the y-direction, depending on the sign of b.