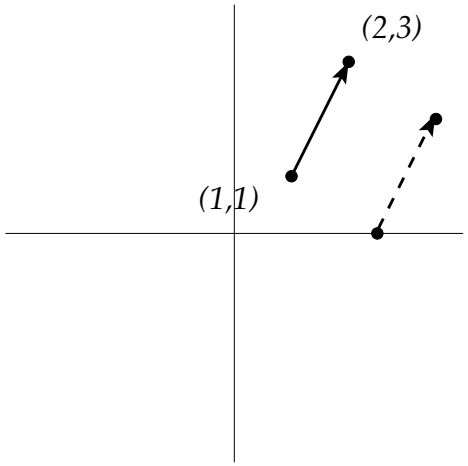
12 Gradients and optimization

12.1 Vectors

Think of a vector as an arrow drawn from one point in the plane or three dimensional space to another. The arrow from $(1, 1)$ to $(2, 3)$ is shown in the figure. The only tricky thing about the definition is that we don't care where the arrow is drawn, we only care about its magnitude (length) and direction. So for example the dashed arrow represents the same vector, started at the point $(5/2, 0)$ instead of $(1, 1)$. In other words, the vector represents the *move* from the beginning to the end of the arrow, regardless of the absolute location of the beginning point.

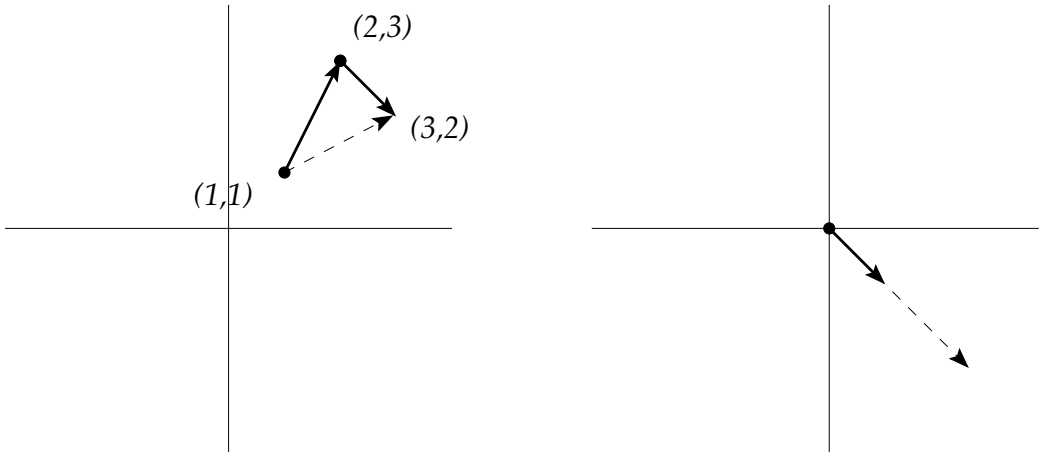


The vector of unit length in the x-direction is called $\hat{\mathbf{i}}$, the vector of unit length in the y-direction is called $\hat{\mathbf{j}}$, and, if we're in three dimensions, the vector of unit length in the z-direction is called $\hat{\mathbf{k}}$. A vector that goes a units in the x-direction and b units in the y-direction is denoted $a\hat{\mathbf{i}} + b\hat{\mathbf{j}}$. It's called that because you can add vectors and multiply them by real numbers (see definition below). For example, the vector in the picture should be written $\hat{\mathbf{i}} + 2\hat{\mathbf{j}}$.

> **Definition of adding vectors.** First make one move, then make the other. You can do this by sliding one of the arrows (don't rotate it!) so it starts where the other one ends, then following them both. If you add $a\hat{\mathbf{i}} + b\hat{\mathbf{j}}$ to $c\hat{\mathbf{i}} + d\hat{\mathbf{j}}$ you get $(a + c)\hat{\mathbf{i}} + (b + d)\hat{\mathbf{j}}$.

Definition of multiplying a vector by a real number. Don't change the direction, just multiply the length. As a formula: multiply $a\hat{\mathbf{i}} + b\hat{\mathbf{j}}$ by c you get $ac\hat{\mathbf{i}} + bc\hat{\mathbf{j}}$. This easy formula hides an important fact: if you mutliply both the $\hat{\mathbf{i}}$ and $\hat{\mathbf{j}}$ coefficients by the same real number, the direction doesn't change. That's why the two vectors in the right-hand figure below are on top of each other.

The left-hand side of the figure below shows the vector $\hat{\mathbf{i}} + 2\hat{\mathbf{j}}$ being added, tip to tail, to the vector $\hat{\mathbf{i}} - \hat{\mathbf{j}}$. The result is the vector $2\hat{\mathbf{i}} + \hat{\mathbf{j}}$ show by the dotted arrow. In the right-hand figure, the vector $\hat{\mathbf{i}} - \hat{\mathbf{j}}$ is multiplied by the real number $\sqrt{6}$ which is a little under 21/2.



The length of a vector can be computed by the Pythagorean Theorem. The length of $a\hat{\mathbf{i}} + b\hat{\mathbf{j}}$ is $\sqrt{a^2 + b^2}$. For example, the vector $\hat{\mathbf{i}} + 2\hat{\mathbf{j}}$ which appears in the previous figures has length $\sqrt{5}$. The length of the vector $\mathbf{v}$ is denoted $|\mathbf{v}|$. A *unit vector* is any vector whose length is 1. Often we want to know a unit vector in a given direction: what vector, having the same direction as $\mathbf{v}$, has length 1? Answer: divide $\mathbf{v}$ by $|\mathbf{v}|$ (that is, multiply $\mathbf{v}$ by the reciprocal of its length). Self-check: what is the unit vector in the direction of our favorite example vector, $\mathbf{v} = \hat{\mathbf{i}} + 2\hat{\mathbf{j}}$? The answer is posted in a link on Canvas (first student who actually wants to look at it, tell me and I'll activate the link).

The dot product

The *dot product* of the vectors $a\hat{\mathbf{i}} + b\hat{\mathbf{j}}$ and $c\hat{\mathbf{i}} + d\hat{\mathbf{j}}$ is defined to be the number (it's not a vector!) $ac + bd$. You'll see next class why this quantity is important. The last thing you need to know is a fact: the dot product of two vectors $\mathbf{v}$ and $\mathbf{w}$ is equal to the product of the lengths times the cosine of the angle $\alpha(\mathbf{v}, \mathbf{w})$ between them:

$$\mathbf{v} \cdot \mathbf{w} = |\mathbf{v}||\mathbf{w}| \cos \alpha(\mathbf{v}, \mathbf{w}) \tag{12.1}$$

Again, there is something important hidden in the content of this formula. You already know one way of computing the dot product: multiply corresponding components and add them. The formula gives you another way. The first way is algebraic. The second way is completely geometric: you could do it by seeing only the picture. The dot product theorem says that these two computations produce the same result. Take a minute to register this, because it will come up in applicaitons, problem sets and, yes, exams.

Parallel vectors

Vectors in the same direction are called *parallel*. How do you tell whether $\mathbf{v} = a\hat{\mathbf{i}} + b\hat{\mathbf{j}}$ is parallel to $\mathbf{w} = c\hat{\mathbf{i}} + d\hat{\mathbf{j}}$? This is the same as saying you can multiply one vector by real number to get the other. This is the same as asking when the fraction c/a is equal to d/b. To test this you crossmultiply, arriving at the condition

$$ad - bc = 0\,. \tag{12.2}$$

Three or more dimensions (optional paragraph)

In three dimensions a generic vector will be the sum of three components: $a\hat{\mathbf{i}} + b\hat{\mathbf{j}} + c\hat{\mathbf{k}}$. The basic definitions are still the same. A vector $\mathbf{v}$ is still defined as having a length and a direction. Both the algebraic and the geometric formulae for the dot product look analogous to they way they looked in two dimensions and give the same answer. Addition of vectors and multiplication of a vector by a real number still have both an algebraic and a geometric definition that give the same result. In fact, you define vectors in any dimension. You can't visualize it, and you run out of letters after $\hat{\mathbf{i}}, \hat{\mathbf{j}}, \hat{\mathbf{k}}$, but all the math still works the way it did in two dimensions. Our treatment is very minimal: we will stick to two dimensions. If vector calculus intrigues you then consider taking Math 114.

12.2 The gradient

Let z be a function of x and y. Think of this for now as the elevation at a point x units east and y units north of a central point. Pick a point (x_0, y_0), let $a = (\partial z/\partial x)(x_0, y_0)$ and let $b = \partial z/\partial y)(x_0, y_0)$. Using these we can figure out the rate of elevation increase for a hiker traveling on the path $(x(t), y(t))$. By the multivariate chain rule, if the hiker is at the position (x_0, y_0) at some time t_0, then the rate of increase of the hiker's elevation at time t_0 will be $ax'(t) + by'(t)$ evaluated at $t = t_0$.

Here's the important point. If we calculate a and b just once, we can figure out the rate of elevation gain of any hiker traveling with any speed in the x- and y-directions. The vector $a\hat{\mathbf{i}} + g\hat{\mathbf{j}}$ is called the *gradient* of z at the point (x_0, y_0) and is denoted $\nabla z(x_0, y_0)$ or just $|\nabla z|$. This definition is given in a box in the middle of page 833 in Section 14.5 of the textbook:

$$\nabla z(x_0, y_0) = \frac{\partial z}{\partial x}(x_0, y_0)\ \hat{\mathbf{i}} + \frac{\partial z}{\partial y}(x_0, y_0)\ \hat{\mathbf{j}}\,.$$

This leads to the idea of the *directional derivative*: what is the rate of elevation gain per unit traveled in any direction? The key here is "per unit traveled". The unit vector $\mathbf{w}$ in the direction making an angle of θ with the positive x-direction is $(\cos\theta)\hat{\mathbf{i}} + (\sin\theta)\hat{\mathbf{j}}$. Therefore, a hiker traveling at unit speed in this direction gains elevation at the rate of $a\cos\theta + b\sin\theta$. That's the dot product $\nabla z \cdot \mathbf{w}$. THIS IS THE MAIN REASON WE COVER VECTORS AND DOT PRODUCTS IN THIS COURSE.

Here are some conclusions you can draw from all of this. Let $L = |\nabla z(x_0, y_0)|$ be the length of the gradient vector of z at the point (x_0, y_0). Now consider all directions the hiker could possibly be traveling: which one maximizes the rate of elevation gain? Let α be the angel between the gradient vector and the hiker's direction in the x-y plane. We have just seen that the rate of elevation gain per unit motion in the direction $\mathbf{w}$ is $\nabla z \cdot \mathbf{w}$. The length of ∇z is L and the lngth of $\mathbf{w}$ is 1, so by formula (12.1), the dot product is $L\cos\alpha$. This cosine is at most 1 and is maximized when the angle is zero, in other words, when the hiker's direction is parallel to the gradient vector. In that case the directional derivative is L. If the hiker is going in a direction making an angle α with the gradient then the rate of elevation gain per unit distance traveled is $L\cos\alpha$. If α is a right angle then this rate is zero. We can summarize these observations in a theorem, which constitutes more or less the "Properties of the directional derivative" stated in a box on page 834.

Gradient Theorem:

> (i) The direction of greatest increase of a function $z(x, y)$ at a point (x_0, y_0) is the direction of its gradient vector $\nabla z(x_0, y_0)$. The rate of increase per unit distance traveled in that direction is the length of the gradient vector which is given by
>
> $$L = \sqrt{\frac{\partial z}{\partial x}(x_0, y_0)^2 + \frac{\partial z}{\partial y}(x_0, y_0)^2}\,.$$
>
> (ii) In general the directional derivative in a direction making angle α with respect to the gradient direction is equalt to $L\cos\alpha$.
>
> (iii) In particular, when α is a right angle, we see that the rate of elevation increase in direction α is zero.

This theorem is, more or less what's in the box on page 834 entitled "Properties of the directional derivative". Mull it over for a minute. By computing partial derivatives, we can stake out the‘ direction of maximum ascent, and it will have the property that the direction of zero elevation gain is at right angles to it (also, the direction of maximal descent is exactly opposite). Remember level curves? Along these, the elevation is constant. Therefore, traveling in these directions makes the rate of elevation gain zero. We see that the tangent to the level curve must be in the zero gain direction, that is, perpendicular to the gradient. This is shown in Figure 14.31 on page 835. A real life illustration is shown in the picture on page 831 of the textbook. A contour map shows contours of an actual mountainside in Yosemite National Park. These are perpendicular to the directions of steepest ascent and descent. You can see this because streams typically flow in the directions of steepest descent. The streams and the level contours are marked on the map and do, indeed, look perpendicular.

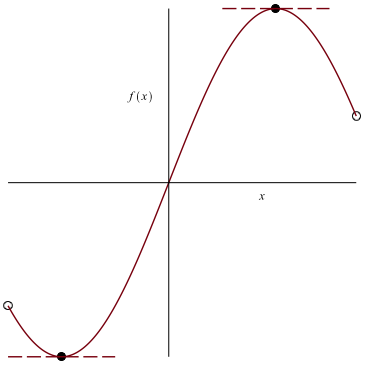
Some rules for computing

We won't need a lot of rules for computing gradients because we'll always be able to compute them by hand but it is good to look them over once. They're collected in a box on page 836 of the textbook. Basically all the rules that work for derivatives work for gradients because in each component separately (the $\hat{\mathbf{i}}$ component, etc.) the gradient is a kind of all-encompassing partial derivative, and partial derivatives obey these laws.

12.3 Optimization

One-paragraph review of univariate optimization

Here is a brief review of optimization in one variable (also known as solving max-min problems) which you can skip if you don't need. Suppose you want to find the maximum of a differentiable function f on an interval $[a, b]$. If the maximum occurs at a point x in the interior of the interval then $f'(x)$ must be zero (if $f'(x) > 0$ then a point just to the right of x will have a higher value of f, whereas if $f'(x) < 0$ then a point just to the left of x will have a higher value of f). Therefore, to find the maximum, you list all the critical points (points where $f' = 0$ and both the endpoints, and see where among these points f has the greatest value. It is the same with minima. List all the critical points of f and the endpoints, and determine the least value of f among this list of points.



The maximum and minimum of f on the interval shown must occur either at the critical points (solid dots) or at the endpoints (open dots). In fact the maximum value (upper dashed line) and minimum value (lower dashed line) occur at critical points in this case.

Optimization along a curve

Now switch gears and consider a function $f(x, y)$ of two variables. There are two kinds of optimization problems that commonly occur. One is to find the maximum or minimum of f on a curve. The second is to find the maximum or minimum of f over a region in the plane.

Conceptually, optimization along a curve is easy: read f "as you go along the curve";

find the critical points where the derivative of the readout is zero; the maximum will have to occur at one of these places; check them all. Computationally, the tricky part is to describe the curve in equations, then use those equations to compute the derivative along the curve.

The description of a curve γ can take one of three forms. It could be given by some function $y = g(x)$. It could be given parametrically by $((x(t), y(t))$. Finally, and most commonly, γ could be given implicitly, meaning it is the solution set to the equation $H(x, y) = 0$ for some function H. We treat these in the order: parametric, function, implicit, because each computation relies on the previous one.

Parametric case: the derivative along $(x(t), (y(t))$.

If the curve γ is paramterized as $(x(t), y(t))$, then the derivative of f along γ is just $\nabla f \cdot \mathbf{v}$ where $\mathbf{v}$ is the velocity vector $x'(t)\hat{\mathbf{i}} + y'(t)\hat{\mathbf{j}}$. In this case, finding the points where the derivative of f along γ vanishes boils down to solving

$$x'(t)\frac{\partial f}{\partial x} + y'(t)\frac{\partial f}{\partial y} = 0\,. \tag{12.3}$$

Self-check: what does it mean that the derivative of f along the curve $(x(t), y(t))$ is given by (12.3)? This formula computes the rate of change of what with respect to what?

Function case: the derivative along $y = g(x)$.

If γ is paramtrized by $y = g(x)$ then you can use the parametric description $x = x, y = g(x)$ so that this equation becomes

$$\frac{\partial f}{\partial x} + g'(x)\frac{\partial f}{\partial y} = 0\,. \tag{12.4}$$

Self-check: again, this is the rate of change of what with respect to what?

Implicit case: the derivative along $H(x, y) = 0$.

Finally, suppose that γ is given implicitly by $H(x, y) = 0$. Recall that we know how to find the slope dy/dx of the tangent line to the level curve $H(x, y) = 0$. By implicit differentiation, we computed $dy/dx = -H_x/H_y$. Therefore we can apply equation (12.4) with $g'(t) = -H_x/H_y$. We get $\partial f/\partial x - (H_x/H_y)\partial f/\partial y = 0$, which simplifies slightly to

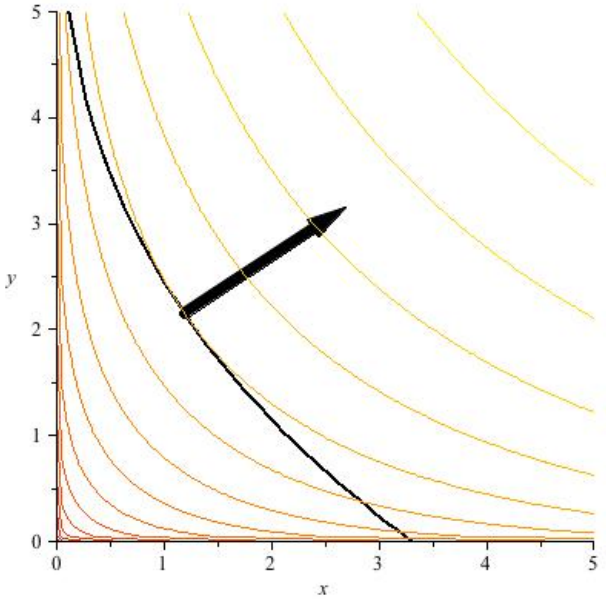
$$H_y\frac{\partial f}{\partial x} - H_x\frac{\partial f}{\partial y} = 0\,. \tag{12.5}$$

IMPORTANT GEOMETRIC INTERPRETATION OF (12.5):

The gradient of H is $H_x\hat{\mathbf{i}} + H_y\hat{\mathbf{j}}$. The gradient of f is $f_x\hat{\mathbf{i}} + f_y\hat{\mathbf{j}}$. The test for these to be parallel is given by applying (12.2) to these two vectors. This results precisely in (12.5). In other words:

> The critical points of f along a level curve of H are those points where the gradients of f and H are parallel.

PICTORIAL EXAMPLE: The figure shows a black constraint curve, $H(x, y) = 0$, along with contours for another function $f(x, y)$. The maximum of f along the curve $H(x, y) = 0$ is the place where the level curves, when you move from higher to lower, just hit the black curve. At this point, the curves are tangent and the gradients are parallel. The single arrow represents the directions of both gradients.



12.4 Optimization over a region

If the maximum of f on the region R occurs at an interior point, then both partial derivatives must vanish there. Why? If one of the partials, say $\partial f/\partial x$ is positive, then the value of the function is just to the right is greater. If, say $\partial f/\partial y$ is negative, then the value of f just below is greater. And so forth. Asking that both partial derivatives vanish is the same as asking that the gradient vanishes. Therefore, we have the following procedure.

> To find the maximum, find all the places inside R where the gradient vanishes, compute f at these places and take the maximum among these, and compare to the maximum of f on the boundary of R.

Note: the last part, finding the maximum on the boundary of f, unless this boundary is very simple, relies on knowing how to do constrained optimization.

EXAMPLE: What is the maximum of the function $x + 2y$ on the region R where the unit circle intersects the first quadrant?

SOLUTION: The gradient of $f(x) = x + 2y$ is the vector $\hat{\mathbf{i}} + 2\hat{\mathbf{j}}$. This never vanishes so the maximum of f is NOT in the interior of the unit circle. The boundary of the region R is made of three pieces: the line segment on the x-axis from the origin to $(1, 0)$; the line segment on the y-axis from the origin to $(0, 1)$; the arc of the unit circle in the first quadrant. The maximum of f on each line segment occurs at the endpoint away from the origin, with values 1 and 2 respectively.

To find the maximum of $x + 2y$ on the arc $x^2 + y^2 = 1$, we compute the gradient of $x^2 + y^2$ which is $2x\hat{\mathbf{i}} + 2y\hat{\mathbf{j}}$. This is parallel to $\nabla f = \hat{\mathbf{i}} + 2\hat{\mathbf{j}}$ when the cross-multiple $2(2x) - 1(2y) = 0$. This happens when $y = 2x$. Plugging in $x^2 + (2x)^2 = 1$, we find that $x = 1/\sqrt{5}$ and $y = 2/\sqrt{5}$. There, $f(x, y) = x + 2y = 1/\sqrt{5} + 4/\sqrt{5} = 5/\sqrt{5} = \sqrt{5} \approx 2.23606$. This beats all the other maxima, therefore the global maximum of $x + 2y$ in the unit circle in the first quadrant is $\sqrt{5}$ and is achieved at $(1/\sqrt{5}, 2/\sqrt{5})$.

EXAMPLE: Where is the maximum of $f(x, y) = x/(1+x^2+y^2)$ on the disk of radius 2? The critical points on the interior are where both partial derivatives of f vanish. The partial derivatives, when expressed with denominator $(1 + x^2 + y^2)^2$ have respective numerators $-2xy$ and $1+y^2-x^2$. Setting these equal to zero gives the points $(\pm 1, 0)$; here $f(x, y) = \pm 1/2$. On the boundary, the denominator is 5 so f can be no more thna $2/5$, therefore the overall maximum is $1/2$ at the point $(1, 0)$.

Application

Let's go back to the pizza and FroYo example from Unit 11.4, but without numbers. Let $H(x, y)$ be the utility of a consumer who gets x ounces of pizza and y pints of FroYo. Let $f(x, y)$ be the cost to me of producing x ounces of pizza and y pints of Froyo. For my ten dollar family bargain, I need to offer a pair that is on the curve $H(x, y) = c$ because that's what Burger Chef is offering and I will lose customers if my pizza-FroYo combo is less desirable than theirs. But my function f is different from Burger Chef's because my prodution line is different. Question: what bundle should I offer?

In mathematical terms, What value of (x, y) on the curve $H(x, y) = c$ minimizes $f(x, y)$? We just saw the answer to that: it is either an endpoint of the curve or a place where ∇f is parallel to ∇H. Let's interpret the parallel gradients in economic terms. Parallel gradients at a point occur when the tangent lines to the level curves are the same at that point. These tangents tell me the marginal rate of substitution. Remember the FroYo example. The tangent to $H(x, y) = 30$ at the point $(60, 1/2)$ tells me the marginal rate of substitution. Consumers at this point are indifferent between another ounce of pizza and another $1/120$ point of FroYo. The tangent to the level curve of f at this point tells me the rate of substitution for costs: how many extra pints of FroYo can I make from the cost savings on each fewer ounce of pizza? If the two slopes are not the same, then I can slide along the customers indifference curve one direction or the other, decreasing my costs while maintaining the same customers. The only way I can be at the minimim cost point on the consumers' indifference curve is to be at a point where the slopes are parallel.

Example: Using the numbers $H(x, y) = xy$ from the original pizza and FroYo example, suppose my cost function is a simple linear function: it coses 10 cents to produce each ounce of pizza and \$1 for each pint of FroYo. Thus $f(x, y) = (0.1)x + y$. The gradient of a linear function is constant: $\nabla f = (1/10)\hat{\mathbf{i}} + \hat{\mathbf{j}}$. The gradient of H is $y\hat{\mathbf{i}} + x\hat{\mathbf{j}}$. These are parallel when $y - x/10 = 0$. At what point on the curve $H(x, y) = 30$ does this occur? We solve

$$\begin{aligned} x &= 10y \\ xy &= 30 \end{aligned}$$

to get $y = \sqrt{3}$ and $x = 10\sqrt{3}$. Look up the approximate value $\sqrt{3} = 1.732\ldots$ on your cheatsheet. In other words, the optimum combo meal for me to sell is (roughly) a 17 and a third ounce pizza and a pint and three quarters of FroYo.