Math 4250 Minihomework: The gradient and the hessian

In this minihomework, we are going to reintroduce the gradient and the Hessian matrix from multivariable calculus, and pair them with some ideas from linear algebra. In order to do this, we're going to introduce a (possibly) new idea: the difference between the maximum (or $\max$) of a function and the arguments of the maxima (or argmax) of the function.

Definition. *If $f\colon \mathcal{X} \to \mathbb{R}$, and $\mathcal{S} \subset \mathcal{X}$ is a set of inputs to f, we define*

$$\max_{\mathcal{S}} f(x) := \textit{ the unique } \textbf{\textit{value}} \ f(x_0) \textit{ so that } x_0 \in \mathcal{S} \textit{ and } f(x_0) \geq f(x) \textit{ for all } x \in \mathcal{S}.$$

We define

$$\operatorname*{argmax}_{\mathcal{S}} f(x) := \textit{ the set } \{x_0 \in \mathcal{S} \textit{ s.t. } f(x_0) = \max_{\mathcal{S}} f(x)\}.$$

The distinction between argmax and $\max$ is much like the distinction between critical *points* of a function $f\colon \mathbb{R} \to \mathbb{R}$, which are x's (or inputs), and critical *values* of a function $f\colon \mathbb{R} \to \mathbb{R}$ which are y's (or outputs). Argmax is a core operation in machine learning, and it's a standard library function in the software libraries `numpy` and `tensorflow`.

We note that $\max_{\mathcal{S}} f(x)$ may not exist[a], but it is unique if it exists. If $\max_{\mathcal{S}} f(x)$ exists, then $\operatorname{argmax}_{\mathcal{S}} f(x)$ is nonempty. In general, there is no reason to believe that a nonempty $\operatorname{argmax}_{\mathcal{S}} f(x)$ contains only a single point.[b]

1. (20 points) Suppose we have a function $f\colon \mathbb{R}^n \to \mathbb{R}$. Recall that

   Definition. *The* gradient vector $\nabla f(\vec{x}) := (\frac{\partial f}{\partial x_1}(\vec{x}) \cdots \frac{\partial f}{\partial x_n}(\vec{x}))^T$*, and the* directional derivative

   $$(D_{\vec{v}} f)(\vec{x}) := \lim_{h \to 0} \frac{f(\vec{x} + h\vec{v}) - f(\vec{x})}{h}.$$

   It's a theorem we prove in multivariable calculus that

   $$D_{\vec{v}} f(\vec{x}) = \langle \vec{v}, \nabla f(\vec{x}) \rangle \qquad (\bigstar)$$

   Use ($\bigstar$) and the fact that $\langle \vec{x}, \vec{y} \rangle = \|\vec{x}\| \|\vec{y}\| \cos\theta$ (where θ is the angle between $\vec{x}$ and $\vec{y}$) to prove the following two facts:

---

[a] For instance $\max_{(0,1)} x^2$ is undefined, because x^2 comes arbitrarily close to 1 on this open interval, but never reaches it.

[b] Unless we have some additional hypothesis about the function! For example, we learned in first-year calculus that if $f\colon \mathbb{R} \to \mathbb{R}$ and $f''(x) < 0$, then there is exactly one point where $f(x)$ is maximized and so $\operatorname{argmax}_{\mathbb{R}} f(x)$ exists and contains only one point.

(1) (10 points) "The norm of the gradient is the largest rate of ascent of the function f", or

$$\max_{\{\vec{v} \text{ s.t. } \|\vec{v}\|=1\}} D_{\vec{v}} f(\vec{x}) = \|\nabla f(\vec{x})\|$$

Solution:

(2) (10 points) "The gradient points in the direction of steepest ascent", or

$$\underset{\{\vec{v} \text{ s.t. } \|\vec{v}\|=1\}}{\operatorname{argmax}} D_{\vec{v}} f(\vec{x}) = \left\{ \frac{\nabla f(\vec{x})}{\|\nabla f(\vec{x})\|} \right\}$$

Note: You must show both that $\frac{\nabla f(\vec{x})}{\|\nabla f(\vec{x})\|} \in \operatorname{argmax}_{\{\vec{v} \text{ s.t. } \|\vec{v}\|=1\}} D_{\vec{v}} f(\vec{x})$ *and* that no other $\vec{v} \in \operatorname{argmax}_{\{\vec{v} \text{ s.t. } \|\vec{v}\|=1\}} D_{\vec{v}} f(\vec{x})$.

Solution:

2. (10 points) Suppose that $f\colon \mathbb{R}^n \to \mathbb{R}$. Recall that $\mathcal{H}f$ is the symmetric $n \times n$ Hessian matrix of f defined (at each point $\vec{x}$) by

$$(\mathcal{H}f(\vec{x}))_{ij} = \frac{\partial^2 f}{\partial x_i \partial x_j}(\vec{x})$$

and that for any symmetric $n \times n$ matrix A, $Q_A(\vec{v}, \vec{w}) = \sum_{ij} A_{ij}\vec{v}_i\vec{w}_j$.

Assuming (★), prove that

$$D_{\vec{w}}(D_{\vec{v}}f(\vec{x})) = Q_{\mathcal{H}f}(\vec{w}, \vec{v}) \qquad (\clubsuit)$$

Hint: We can express the right-hand side of (♣) as a double sum. Can you express the left hand side of (♣) as a double sum as well?

Solution:

3. (20 points) We learned a form of Taylor's theorem in Calculus II; for a smooth enough function $f(x)$ near a point a, we have

$$f(a+x) \simeq f(a) + f'(a)x + \frac{1}{2}f''(a)x^2$$

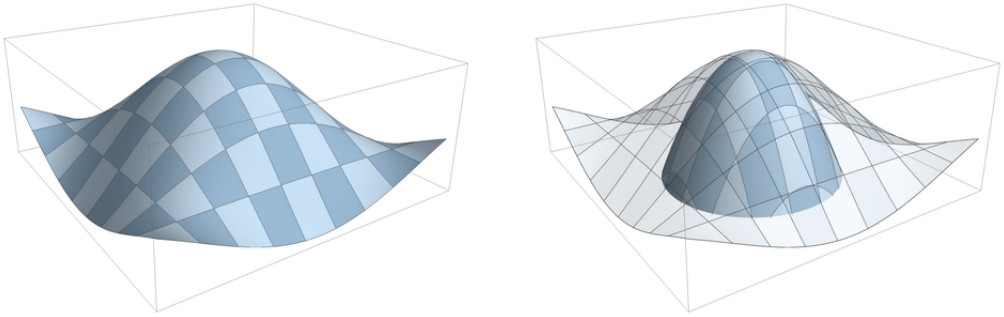
and the quadratic polynomial in x on the right-hand side is the "best"[c] quadratic approximation to $f(x)$ near a. In general, we have

Theorem (Multivariable Taylor's Theorem)**.** *If $f\colon \mathbb{R}^n \to \mathbb{R}$ is a C^3 function near $\vec{a}$, we have*

$$\begin{aligned} f(\vec{a}+\vec{x}) &\simeq f(\vec{a}) + (D_{\vec{x}}f)(\vec{a}) + \frac{1}{2}((D_{\vec{x}}(D_{\vec{x}}f))(\vec{a}) \\ &\simeq f(\vec{a}) + \langle \vec{x}, (\nabla f)(\vec{a})\rangle + \frac{1}{2}\langle \vec{x}, \mathcal{H}f(\vec{a})\vec{x}\rangle \end{aligned}$$

and the quadratic polynomial in $\vec{x}$ on the right hand side is the "best" quadratic approximation to $f(x)$ near $\vec{a}$.

(1) (10 points) The pictures below show a plot of the function $f(x,y) = \cos x \cos y$ and the best quadratic approximation $p(x,y)$ of $f(x,y)$ near $(0,0)$:



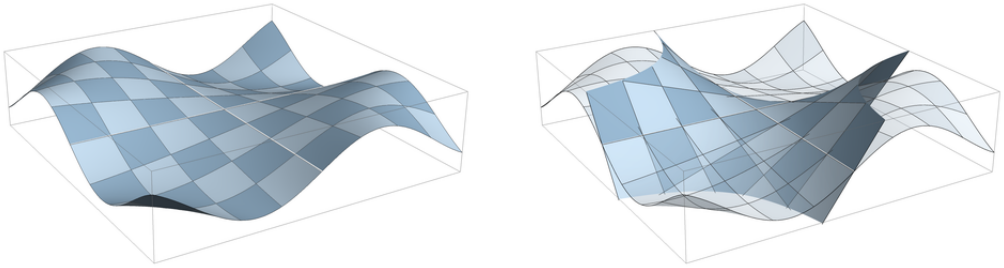
Use the multivariable Taylor theorem to find the function $p(x,y)$, which should be a quadratic polynomial in x and y.

Solution:

[c] For the moment, we're not going to get into the question of what it means to be the best approximation. Suffice it to say that (at one point), you learned theorems about the remainder term in Taylor's theorem.

Solution:

(2) (10 points) The pictures below show a plot of the function $f(x, y) = \sin(xy)$ and the best quadratic approximation $p(x, y)$ of $f(x, y)$ near $(0, 0)$:



Use the multivariable Taylor theorem to find the function $p(x, y)$, which should be a quadratic polynomial in x and y.

Solution: