CS 189 Introduction to Machine Learning
Fall 2017 DIS6

1 Gradient descent for simple functions

(a) Recall Taylor's theorem for twice differentiable functions of vectors, which holds for all $x, y \in \mathbb{R}^d$:

$$f(y) = f(x) + \nabla f(x)^\top (y - x) + \frac{1}{2}(y - x)^\top \nabla^2 f(\tilde{x})(y - x),$$

for some $\tilde{x}$. Show that the function f is convex if $\nabla^2 f(x)$ is positive semidefinite for all $x \in \mathbb{R}^d$.

Solution: All of these problems are best understood using plots.

Note that a convex function is one which always lies above the tangent at any point. Mathematically, that looks like

$$f(y) \geq f(x) + \nabla f(x)^\top (y - x),$$

for all $x, y \in \mathbb{R}^d$. Thus, it suffices to prove that $\frac{1}{2}(y - x)^\top \nabla^2 f(\tilde{x})(y - x) \geq 0$ for all $\tilde{x}, x, y$.

By the definition of positive semidefiniteness of a matrix A, we have $v^\top A v \geq 0$ for all vectors v. Choosing $A = \nabla^2 f(\tilde{x})$ completes the proof.

(b) Let $L \geq 0$. Consider the function of one variable $f(x) = \frac{L}{2}x^2$. Show that it is convex.

Solution: From the above discussion, it suffices to show that $\nabla^2 f(x)$ is PSD for all x. Since we are in one dimension, $\nabla^2 f(x) = L \geq 0$.

(c) Derive the gradient descent update where we use a step-size of γ and start at some point $x^{(0)} \neq 0$.

Solution: The gradient is given by $f'(x) = Lx$. The gradient descent update is therefore given by

$$\begin{aligned} x^{(i+1)} &= x^{(i)} - \gamma \nabla f(x^{(i)}) \\ &= (1 - \gamma L) x^{(i)}. \end{aligned}$$

(d) What does the behavior look like for the above setting and the choices $\gamma \in \{1/L, 2/L\}$?

Solution: Plugging the choice $\gamma = 1/L$ into the update, we see that $x^{(1)} = 0$, and so we reach the optimal solution in just one step.

On the other hand, the choice $\gamma = 2/L$ yields $x^{(1)} = -x^{(0)}$, and so we oscillate between the points $x^{(0)}$ and $-x^{(0)}$ forever.

(e) Consider the above setup and assume we use a step size $\gamma \in \left[0, \frac{2}{L}\right)$. Also assume that $\gamma \neq 1/L$. How many steps does it take for us to converge to within ε of the optimum (as a function of the tuple $(\gamma, L, |x^{(0)}|, \varepsilon)$)?

Solution: Iterating the gradient descent update, we have

$$x^{(i)} = (1 - \gamma L)^i x^{(0)}.$$

We know that the optimum is at 0, and so we would like $|x^{(i)}| \leq \varepsilon$. Thus, we need

$$(1 - \gamma L)^i \leq \varepsilon / |x^{(0)}|.$$

Simplifying, we see that setting $i \geq \frac{\log \frac{\varepsilon}{|x^{(0)}}}{\log(1-\gamma L)}$ suffices.

A better way to see the scaling of the problem is to use the inequality $1 - t \leq e^{-t}$, which holds for all scalar t. Thus, it suffices to have

$$\left(e^{-\gamma L}\right)^i \leq \varepsilon / |x^{(0)}|,$$

and simplifying yields that $i \geq \frac{1}{\gamma L} \log(|x^{(0)}/\varepsilon)$ is sufficient.

(f) How do your answers above change if $f(x) = \frac{L}{2}(x - c)^2$ for some constant c?

Solution: The gradient changes, but the behavior of the algorithm does not since the optimum also changes to being at $x = c$.

(g) Let $L \geq m \geq 0$. Now consider the function of two variables $f(x) = \frac{L}{2}x_1^2 + \frac{m}{2}x_2^2$. Show that the function is convex by computing its Hessian $\nabla^2 f(x)$.

Solution: In order to compute the Hessian, we can compute the "gradient" of the gradient. We have $\frac{\partial f}{\partial x_1} = Lx_1$, and $\frac{\partial f}{\partial x_2} = Lx_2$. Differentiating once more, we have

$$\frac{\partial^2 f}{(\partial x_1)^2} = L$$
$$\frac{\partial^2 f}{(\partial x_2)^2} = m$$
$$\frac{\partial^2 f}{\partial x_1 \partial x_2} = 0$$
$$\frac{\partial^2 f}{\partial x_2 \partial x_1} = 0.$$

Thus, the Hessian is given by the diagonal matrix $\begin{bmatrix} L & 0 \\ 0 & m \end{bmatrix}$, which is clearly positive semidefinite.

(h) With the setup of the previous part, let us say we started at the point $(0,5)$. What is the maximum step-size that results in convergence? How would your answer change if we started at the point $(5,0)$?

Solution: When we are at the point $(0,5)$, we have already found the minimizer in the x_1 direction, and must minimize the x_2 direction. We thus have a quadratic problem in one variable that is very similar to the parts above, where we are minimizing the function $f(x_2) = \frac{m}{2}x_2^2$. Thus, using the above parts, we know that having a step-size less than $2/m$ is sufficient to guarantee convergence.

Starting from $(5,0)$, we are minimizing along the other direction, corresponding to the function $f(x_1) = \frac{L}{2}x_1^2$. Thus, choosing a step-size less than $2/L$ is necessary and sufficient to guarantee convergence.

(i) Derive closed form expressions for the iterations if we start at the point (a,b), and run gradient descent with step-size γ. Start by writing out the result of the first iteration as $A\begin{bmatrix}a\\b\end{bmatrix}$ for some matrix A.

Solution: As we derived above, the gradient of the function is given by

$$\nabla f(x) = \begin{bmatrix}Lx_1\\mx_2\end{bmatrix},$$

and so the first iterate is given by

$$x_1^{(1)} = (1-\gamma L)x_1^{(0)}$$
$$x_2^{(1)} = (1-\gamma m)x_2^{(0)}.$$

Writing this in matrix form, we have

$$x^{(1)} = \begin{bmatrix}(1-\gamma L) & 0\\0 & (1-\gamma m)\end{bmatrix}x^{(0)}.$$

Denoting the matrix by A, the ith iterate therefore takes the form $x^{(i)} = A^i x^{(0)}$.

(j) Now consider the function of one variable $f(x) = L|x|$. Is this function convex? Discuss how performing gradient descent with a fixed step-size performs on this function.

Solution: This function convex when $L \geq 0$ and concave otherwise. Assuming $L > 0$, the function lies above its tangent at every point. Note that there are multiple tangents at the point 0, called "subgradients" but this is just a technicality.

However, performing gradient descent with a constant step-size takes us to within $\gamma/2L$ of the optimum 0, and then we oscillate about the optimum. To see this, say we are at a point $x^{(i)} = L\gamma/2$, and run one step of gradient descent. This takes us to the point $x^{(i+1)} = x^{(i)} - \gamma L = -L\gamma/2$, since the gradient at $L\gamma/2$ is L. We therefore bounce around the optimum indefinitely, and can only converge to within a neighborhood $L\gamma/2$, unless we get lucky and $|x^{(0)}|$ is an exact multiple of $L\gamma$.