EECS 127/227AT **Fall 2020**
Discussion 12

1 Stochastic Gradient Method: A Simple Case

Given a differentiable function $f : \mathbb{R}^n \to \mathbb{R}$ with domain $\mathbb{R}^n$ whose minimum we seek to find, we could use the gradient descent algorithm

$$\theta_{k+1} = \theta_k - \eta \nabla f(\theta_k),$$

with fixed step size $\eta > 0$, starting from an initial condition $\theta_0 \in \mathbb{R}^n$. As we have seen, of course, there is no guarantee that this algorithm converges, and even if it does it may only converge to a local minimum of the function.

One issue with the gradient descent algorithm is the complexity of computing the gradient at each time step. If the function could be decomposed as a summation of multiple functions

$$f(\theta) = \sum_{l=1}^{m} f_l(\theta),$$

for each of which the gradient is easily computable, then we can use the *stochastic gradient* method. For instance, the squared-error-loss function which shows up in the least squares problem is well-suited for minimization with the stochastic gradient method. Here our problem is

$$\min_{\theta \in \mathbb{R}^n} \frac{1}{2}\|X\theta - y\|_2^2 = \frac{1}{2}\sum_{i=1}^{m}(x_i^\top \theta - y_i)^2,$$

where ${x_i}^\top$ is the ith row of $X \in \mathbb{R}^{m\times n}$, and $y \in \mathbb{R}^m$ (recall that the rows of X are the transposes of the *feature vectors* and the entries of y are the corresponding *responses*). We can write this objective function as

$$f(\theta) = \sum_{i=1}^{m} f_i(\theta),$$

with

$$f_i(\theta) := \frac{1}{2}(x_i^\top \theta - y_i)^2, \quad \text{for } i = 1, \dots, m.$$

Then the stochastic gradient method gives the update rule

$$\theta_{k+1} = \theta_k - \eta_k \nabla f_{s[k]}(\theta_k),$$

where η_k is the step size (also called the learning rate) at time $k \in \mathbb{N}$, and $s[k] \in \{1, \dots, m\}$ is the index of the component function chosen at time k in order to decide the update. The value of $s[k]$ is usually chosen by drawing a number at random from the set $\{1, \dots, m\}$, or by randomly shuffling this set and going over it sequentially in cyclic order.

(a) Assume $\{x_i\}_{i=1}^m$ is a set of mutually orthogonal vectors. Find a fixed step size η so that the stochastic gradient method converges to a solution of the least squares problem.

(b) If we no longer assume $\{x_i\}_{i=1}^m$ is orthogonal, can we still find a fixed step size small enough that the stochastic gradient method converges?

2 Convexity and strong convexity

In this question we will explore the concept of *strong convexity*, which is one the of standard conditions on convex function under which many convergence theorems about algorithms are proved.

(a) Let $f : \mathbb{R}^n \to \mathbb{R}$ be a differentiable function with domain $\mathrm{dom}(f)$. Note that requiring that f is differentiable automatically implies that we are assuming that $\mathrm{dom}(f)$ is an open set.

i. Show that f is convex iff it holds that $\mathrm{dom}(f)$ is a convex set and for all $x, y \in \mathrm{dom}(f)$ we have

$$\left(\nabla f(y) - \nabla f(x)\right)^T (y - x) \geq 0. \tag{1}$$

Remark: When a function $g : \mathbb{R}^n \to \mathbb{R}^n$ satisfies the condition $\left(g(y) - g(x)\right)^T (y - x) \geq 0$ for all $x, y \in \mathrm{dom}(g)$, we say that that g is *monotone*. Note that this is consistent with the use of the term "monotone" to refer to a function $g : \mathbb{R} \to \mathbb{R}$ that is monotonically increasing (although one often uses the term in this case to also apply to a function $g : \mathbb{R} \to \mathbb{R}$ that is monotonically decreasing). Thus the condition in (1) is saying that ∇f is monotone.

ii. Recall that a function $f : \mathbb{R}^n \to \mathbb{R}$ with domain $\mathrm{dom}(f)$ is said to be strictly convex if $\mathrm{dom}(f)$ is a convex set and for all $x \neq y \in \mathrm{dom}(f)$ and $\lambda \in (0, 1)$ we have

$$f(\lambda x + (1 - \lambda) y) < \lambda f(x) + (1 - \lambda) f(y).$$

Show that f is strictly convex iff it holds that $\mathrm{dom}(f)$ is a convex set and for all $x \neq y \in \mathrm{dom}(f)$ we have

$$\left(\nabla f(y) - \nabla f(x)\right)^T (y - x) > 0. \tag{2}$$

Remark: When a function $g : \mathbb{R}^n \to \mathbb{R}^n$ satisfies the condition $\left(g(y) - g(x)\right)^T (y - x) > 0$ for all $x \neq y \in \mathrm{dom}(g)$ we say that that g is *strictly monotone*.

Example: Let $A \in \mathbb{S}^n$. Consider the quadratic function $f : \mathbb{R}^n \to \mathbb{R}$ given by $f(x) := \frac{1}{2} x^T A x$, with $\mathrm{dom}(f) = \mathbb{R}^n$. Then $\nabla f(x) = Ax$ and $\nabla^2 f(x) = A$.

Thus, f is convex iff A is positive semidefinite. f is strictly convex iff A is positive definite.

For $x, y \in \mathbb{R}^n$ we have

$$\left(\nabla f(y) - \nabla f(x)\right)^T (y - x) = (y - x)^T A^T (y - x) = (y - x)^T A (y - x).$$

This expression nonnegative for all $x, y \in \mathbb{R}^n$ iff A is positive semidefinite. It is positive for all $x \neq y \in \mathbb{R}^n$ iff A is positive definite.

Thus this example is consistent with the results that have been been proved in this part of the question.

Example: It is important to realize that strict convexity of a function does not imply that its Hessian needs to be positive definite everywhere. For example, consider $f : \mathbb{R} \to \mathbb{R}$, with domain $\mathbb{R}$, given by $f(x) = x^4$. Then $f'(x) = 4x^3$ and $f''(x) = 12x^2$. Note that $f''(0) = 0$. Nevertheless, f is strictly convex. This can be checked from the definition, or by observing that for all $x \neq y \in \mathbb{R}$ we have

$$(f'(y) - f'(x))(y - x) = 4(y^3 - x^3)(y - x) > 0.$$

(b) Let $m > 0$. A function $f : \mathbb{R}^n \to \mathbb{R}$ is called m-*strongly convex* if the function

$$h(x) := f(x) - \frac{m}{2}\|x\|_2^2,$$

with $\mathrm{dom}(h) := \mathrm{dom}(f)$, is convex.

Remark: Suppose f is twice differentiable. Then the convexity of h is equivalent to requiring that $\lambda_{\min}(\nabla^2 f(x)) \geq m$ for all $x \in \mathrm{dom}(f)$. Thus having this property and a convex domain is an equivalent characterization of m-strong convexity for twice differentiable functions.

Example: Let $A \in \mathbb{S}^n$. Consider the quadratic function $f : \mathbb{R}^n \to \mathbb{R}$ given by $f(x) := \frac{1}{2}x^T Ax$, with $\mathrm{dom}(f) = \mathbb{R}^n$. Then $\nabla f(x) = Ax$ and $\nabla^2 f(x) = A$.

For $x, y \in \mathbb{R}^n$ we have

$$(\nabla f(y) - \nabla f(x))^T (y - x) = (y - x)^T A^T (y - x) = (y - x)^T A(y - x).$$

Thus, in this example, f is m-strongly convex iff $\lambda_{\min}(A) \geq m$.

i. Show that $f : \mathbb{R}^n \to \mathbb{R}$ is m-strongly convex iff for all $x, y \in \mathrm{dom}(f)$ and $\lambda \in [0, 1]$ we have

$$f(\lambda x + (1 - \lambda)y) \leq \lambda f(x) + (1 - \lambda)f(y) - \frac{m}{2}\lambda(1 - \lambda)\|x - y\|_2^2. \tag{3}$$

ii. Let $f : \mathbb{R}^n \to \mathbb{R}$ be a differentiable function with domain $\mathrm{dom}(f)$ (note that this means $\mathrm{dom}(f)$ must be an open set). Given $m > 0$, show that f is m-strongly convex iff it holds that $\mathrm{dom}(f)$ is a convex set and for all $x, y \in \mathrm{dom}(f)$ we have

$$(\nabla f(y) - \nabla f(x))^T (y - x) \geq m\|x - y\|_2^2. \tag{4}$$

Remark: When a function $g : \mathbb{R}^n \to \mathbb{R}^n$ satisfies the condition $(g(y) - g(x))^T (y - x) > m\|x - y\|^2$ for all $x, y \in \mathrm{dom}(g)$ we say that that g is *strongly monotone* or *coercive* (confusingly, the term "coercive" is also used in a different sense, which we will encounter later). Thus the condition in (4) is saying that ∇f is strongly monotone.

Addendum (optional):

We give a complete proof for the second sub-part of the first part of this question. This will show up in the solutions document.

3 Convexity and smoothness

In this question we will explore the concept of L-*smoothness*, which is another one of the standard conditions on convex function under which many convergence theorems about algorithms are proved.

Let $f : \mathbb{R}^n \to \mathbb{R}$ be a differentiable function with domain $\operatorname{dom}(f)$ (note that this means $\operatorname{dom}(f)$ must be an open set). Given $L > 0$, f is said to be L-*smoooth* if for all $x, y \in \operatorname{dom}(f)$ we have

$$\|\nabla f(y) - \nabla f(x)\|_2 \leq L\|x - y\|_2. \tag{5}$$

(a) Let $f : \mathbb{R}^n \to \mathbb{R}$ be a differentiable function with domain $\operatorname{dom}(f)$ that is L-smooth for some $L > 0$. Show that we have

$$\left(\nabla f(y) - \nabla f(x)\right)^T (y - x) \leq L\|y - x\|_2^2, \tag{6}$$

for all $x, y \in \operatorname{dom}(f)$.

(b) Let $f : \mathbb{R}^n \to \mathbb{R}$ be a differentiable function with domain $\operatorname{dom}(f)$. Assume that $\operatorname{dom}(f)$ is a convex set. Show that f satisfies (6) for all $x, y \in \operatorname{dom}(f)$ iff the function

$$h(x) := \frac{L}{2}\|x\|_2^2 - f(x),$$

with $\operatorname{dom}(h) := \operatorname{dom}(f)$ is a convex function.

(c) Let $f : \mathbb{R}^n \to \mathbb{R}$ be a differentiable function. Assume that $\operatorname{dom}(f)$ is a convex set Show that f that satisfies (6) for all $x, y \in \operatorname{dom}(f)$ iff it satisfies

$$f(y) \leq f(x) + \nabla f(x)^T(y - x) + \frac{L}{2}\|y - x\|_2^2, \tag{7}$$

for all $x, y \in \operatorname{dom}(f)$.

Example: Let $A \in \mathbb{S}^n$. Consider the quadratic function $f : \mathbb{R}^n \to \mathbb{R}$ given by $f(x) := \frac{1}{2}x^T Ax$, with $\operatorname{dom}(f) = \mathbb{R}^n$. Then $\nabla f(x) = Ax$ and $\nabla^2 f(x) = A$.

For $x, y \in \mathbb{R}^n$ we have

$$\left(\nabla f(y) - \nabla f(x)\right)^T (y - x) = (y - x)^T A^T (y - x) = (y - x)^T A(y - x).$$

Thus f is L-smooth iff $\lambda_{\max}(A) \leq L$. Note that

$$\frac{L}{2}\|x\|_2^2 - \frac{1}{2}x^T Ax = \frac{1}{2}x^T(LI - A)x$$

defines a convex function iff $L \geq \lambda_{\max}(A)$.

Remark: Let $g : \mathbb{R}^n \to \mathbb{R}^m$ with domain $\operatorname{dom}(g)$, and let $L > 0$. Then g is said to be *Lipschitz with Lipschitz constant* L if we have

$$\|g(y) - g(x)\|_2 \leq L\|y - x\|_2,$$

for all $x, y \in \operatorname{dom}(g)$. Thus a differentiable function $f : \mathbb{R}^n \to \mathbb{R}$ satisfies (6) precisely when $\nabla f : \mathbb{R}^n \to \mathbb{R}^n$, with $\operatorname{dom}(\nabla f) := \operatorname{dom}(f)$, is Lipschitz with Lipschitz constant L.

Remark: If $f : \mathbb{R}^n \to \mathbb{R}$ is twice differentiable then the convexity of $\frac{L}{2}\|x\|_2^2 - f(x)$ is equivalent to requiring that $\operatorname{dom}(f)$ be a convex set and $\lambda_{\max}(\nabla^2 f(x)) \leq L$ for all $x \in \operatorname{dom}(f)$. Thus this is an equivalent characterization of L-smoothness for twice differentiable functions.