EECS 127/227AT **Fall 2020**
Solutions for 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

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

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 $s[k] \in \{1, \ldots, m\}$ 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 $\{1, \ldots, m\}$, 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.

Solution:
For SGD, if x_i is the point drawn at iteration $k \in \mathbb{N}$, we can write the update rule as

$$\begin{aligned}\theta_{k+1} &= \theta_k - \frac{1}{2}\eta \nabla_{\theta_k}(x_i^\top \theta_k - y_i)^2 \\ &= \theta_k - \eta(x_i^\top \theta_k - y_i)x_i.\end{aligned}$$

To analyze the convergence, we consider the dynamics of the error terms $e_k^{(i)} := x_i^\top \theta_k - y_i$, where $e_k^{(i)}$ denotes the error associated with data point i at time k, $1 \le i \le m$. For any $1 \le j \le m$, we have

$$\theta_{k+1}^\top x_j - y_j = \theta_k^\top x_j - y_j - \eta(x_i^\top \theta_k - y_i) x_i^\top x_j,$$

which shows that (because $\{x_i\}_{i=1}^m$ is a mutually-orthogonal set of vectors) we have

$$e_{k+1}^{(i)} = \begin{cases} e_k^{(i)} & \text{if } x_i \text{ is not drawn at time } k, \\ (1 - \eta \|x_i\|_2^2) e_k^{(i)} & \text{if } x_i \text{ is drawn at time } k. \end{cases}$$

If $0 < \eta < 2/\max_i \|x_i\|_2^2$ then we have $-1 < 1 - \eta\|x_i\|_2^2 < 1$ for all i. Hence, if every point in the set $\{x_i\}_{i=1}^m$ is drawn infinitely often, all the error terms go to zero, so the objective function converges to zero, which is its optimal value.

To show that θ_k also converges, we also consider the dynamics of θ_k. Assume the initialization of θ is

$$\theta_0 = \sum_{i=1}^m \alpha_i x_i + \sum_{i=1}^{n-m} \beta_i z_i,$$

where $\{z_i\}_{i=1}^{n-m}$ is a set of vectors orthogonal to $\{x_i\}_{i=1}^m$ which span $\mathbb{R}^n$ along with $\{x_i\}_{i=1}^m$. Then the update rule for θ_k shows that α_i converges to its optimal value for all $i = 1, \dots, m$, whereas β_i remains at their initial value for all $i = 1, \dots, n-m$. Therefore, θ_k also converges (to a point that in general depends on the initial point θ_0).

(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?

Solution:
Assume that $\{x_i\}_{i=1}^m$ is not an orthogonal set of vectors. If x_i is drawn at time k, the error term for point j becomes:

$$\theta_{k+1}^\top x_j - y_j = \theta_k^\top x_j - y_j - \eta(x_i^\top \theta_k - y_i) x_i^\top x_j,$$

which we can write as

$$e_{k+1}^{(j)} = e_k^{(j)} - \eta e_k^{(i)} \langle x_i, x_j \rangle.$$

For the gradient terms to vanish eventually, we have to have

$$e_k^{(i)}, e_k^{(j)} \to 0 \quad \text{for every } i, j : \langle x_i, x_j \rangle \neq 0.$$

However, it is possible that the error terms $e_k^{(i)}$, $e_k^{(j)}$ cannot be both zero due to inconsistency in y_i and y_j; for example, when $x_j = 2x_i$ but $y_j \neq 2y_i$. Therefore it may happen that the stochastic gradient method with fixed step size does not converge, no matter how small the step size is chosen. In general, to ensure convergence of such an algorithm requires diminishing the step size over time.

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

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

Remark: When a function $g : \mathbb{R}^n \to \mathbb{R}^n$ satisfies the condition $(g(y) - g(x))^T (y - x) \geq 0$ 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.

Solution:

Suppose f is convex. Then $\mathrm{dom}(f)$ is convex. Also, given any $x, y \in \mathrm{dom}(f)$, we have

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

and

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

Adding these two inequalities gives

$$0 \geq (\nabla f(x) - \nabla f(y))^T (y - x),$$

which is the same as (1).

Conversely, suppose the differentiable function f is such that $\mathrm{dom}(f)$ is a convex set and the condition in (1) holds for all $x, y \in \mathrm{dom}(f)$. Given $x, y \in \mathrm{dom}(f)$, we have $x + t(y - x) \in \mathrm{dom}(f)$ for all t in an open interval containing $[0, 1]$, because $\mathrm{dom}(f)$ is open and convex. Define

$$g(t) := f(x + t(y - x)),$$

for t in such an open interval containing $[0, 1]$. Then

$$g'(t) = \nabla f(x + t(y - x))^T (y - x), \;\; t \in [0, 1].$$

We have

$$\begin{aligned} f(y) - f(x) &= g(1) - g(0) \\ &= \int_0^1 g'(t) dt \\ &= \int_0^1 \nabla f(x + t(y - x))^T (y - x) dt \\ &= \int_0^1 (\nabla f(x + t(y - x)) - \nabla f(x))^T (y - x) dt + \nabla f(x)^T (y - x) \\ &\geq \nabla f(x)^T (y - x), \end{aligned}$$

where we have used (1) in the last step (for the pair $x, x + t(y - x) \in \mathrm{dom}(f)$). Since this holds for all $x, y \in \mathrm{dom}(f)$ and since $\mathrm{dom}(f)$ is a convex set, we conclude that f is convex.

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

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

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

Solution:

The argument is more or less identical to that in the preceding sub-part of this part of the question, with the obvious changes (strict inequality instead of weak inequality). For convenience, we give the full argument as an optional addendum at the end of this question.

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 $\nabla^2 f(x) = A$.

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

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

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 $x, y \in \text{dom}(f)$ and $\lambda \in [0,1]$ we have

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

Solution:
We have

$$\|x-y\|_2^2 = \|x\|_2^2 + \|y\|_2^2 - 2x^T y,$$

and

$$\|\lambda x + (1-\lambda)y\|_2^2 = \lambda^2\|x\|_2^2 + (1-\lambda)^2\|y\|_2^2 + 2\lambda(1-\lambda)x^T y.$$

Adding $\lambda(1-\lambda)$ times the former to the latter gives

$$\lambda(1-\lambda)\|x-y\|_2^2 + \|\lambda x + (1-\lambda)y\|_2^2 = \lambda\|x\|_2^2 + (1-\lambda)\|y\|_2^2. \tag{4}$$

Now, (3) is equivalent to

$$h(\lambda x + (1-\lambda)y) + \frac{m}{2}\|\lambda x + (1-\lambda)y\|_2^2 \leq \lambda h(x) + \frac{m}{2}\lambda\|x\|_2^2 + (1-\lambda)h(y) + \frac{m}{2}(1-\lambda)\|y\|_2^2 - \frac{m}{2}\lambda(1-\lambda)\|x-y\|_2^2,$$

the truth of which for all $x, y \in \text{dom}(h)$ is, in view of (4), equivalent to the convexity of h. This proves what is desired.

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

$$(\nabla f(y) - \nabla f(x))^T (y-x) \geq m\|x-y\|_2^2. \tag{5}$$

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 $x, y \in \text{dom}(g)$ 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 (5) is saying that ∇f is strongly monotone.

Solution:
Suppose f is m-strongly convex. Then, in particular, f is convex, so $\text{dom}(f)$ is a convex set. Let

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

with $\text{dom}(h) := \text{dom}(f)$. Then h is convex. By the first sub-part of the first part of this question, we have

$$(\nabla h(y) - \nabla h(x))^T (y-x) \geq 0,$$

for all $x, y \in \text{dom}(h) = \text{dom}(f)$. Since

$$\nabla h(x) = \nabla f(x) - mx, \;\; \nabla h(y) = \nabla f(y) - my,$$

this gives (5), as desired.

Conversely, suppose $\text{dom}(f)$ is convex and (5) holds for all $x, y \in \text{dom}(f)$. If we define

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

for all $z \in \mathrm{dom}(f)$, with $\mathrm{dom}(h) := \mathrm{dom}(f)$, then we have

$$\nabla h(z) = \nabla f(z) - mz,$$

and so from (5) we have

$$(\nabla h(y) - \nabla h(x))^T (y - x) \geq 0,$$

from which it follows, since this holds for all $x \neq y \in \mathrm{dom}(h) = \mathrm{dom}(f)$, that h is convex (as shown in the first sub-part of the first part of this question) and hence that f is m-strongly convex.

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.

Solution:

Suppose f is strictly convex. Then $\mathrm{dom}(f)$ is convex. Also, given any $x \neq y \in \mathrm{dom}(f)$, we have

$$f(y) - f(x) > \nabla f(x)^T (y - x),$$

and

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

Adding these two inequalities gives

$$0 > (\nabla f(x) - \nabla f(y))^T (y - x),$$

which is the same as (2).

Conversely, suppose the differentiable function f is such that $\mathrm{dom}(f)$ is a convex set and the condition in (2) holds for all $x \neq y \in \mathrm{dom}(f)$. Given $x \neq y \in \mathrm{dom}(f)$, we have $x + t(y - x) \in \mathrm{dom}(f)$ for all t in an open interval containing $[0, 1]$, because $\mathrm{dom}(f)$ is open and convex. Define

$$g(t) := f(x + t(y - x)),$$

for t in such an open interval containing $[0, 1]$. Then

$$g'(t) = \nabla f(x + t(y - x))^T (y - x), \;\; t \in [0, 1].$$

We have

$$\begin{aligned} f(y) - f(x) &= g(1) - g(0) \\ &= \int_0^1 g'(t) dt \\ &= \int_0^1 \nabla f(x + t(y - x))^T (y - x) dt \\ &= \int_0^1 (\nabla f(x + t(y - x)) - \nabla f(x))^T (y - x) dt + \nabla f(x)^T (y - x) \\ &> \nabla f(x)^T (y - x), \end{aligned}$$

where we have used (2) in the last step (for the pair $x, x + t(y - x) \in \mathrm{dom}(f)$, with $t \in (0, 1)$). Since this holds for all $x \neq y \in \mathrm{dom}(f)$ and since $\mathrm{dom}(f)$ is a convex set, we conclude that f is strictly convex.

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 $\mathrm{dom}(f)$ (note that this means $\mathrm{dom}(f)$ must be an open set). Given $L > 0$, f is said to be L-*smoooth* if for all $x, y \in \mathrm{dom}(f)$ we have

$$\|\nabla f(y) - \nabla f(x)\|_2 \le L\|x - y\|_2. \tag{6}$$

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

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

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

Solution:
We have

$$(\nabla f(y) - \nabla f(x))^T (y - x) \le \|\nabla f(y) - \nabla f(x)\|_2 \|y - x\|_2 \le L\|y - x\|_2^2,$$

where the first inequality is an application of the Cauchy-Schwarz inequality and the second inequality is from the definition of L-smoothness.

(b) Let $f : \mathbb{R}^n \to \mathbb{R}$ be a differentiable function with domain $\mathrm{dom}(f)$. Assume that $\mathrm{dom}(f)$ is a convex set. Show that f satisfies (7) for all $x, y \in \mathrm{dom}(f)$ iff the function

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

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

Solution:
Note that

$$\nabla h(x) = Lx - \nabla f(x).$$

For any $x, y \in \mathrm{dom}(h) = \mathrm{dom}(f)$, since

$$\begin{aligned}(\nabla h(x) - \nabla h(y))^T (y - x) &= (Lx - \nabla f(x) - Ly + \nabla f(y))^T (y - x)\\ &= (\nabla f(y) - \nabla f(x))^T (y - x) - L\|y - x\|_2^2,\end{aligned}$$

we see that (7) is equivalent

$$(\nabla h(y) - \nabla h(x))^T (y - x) \ge 0,$$

which we know from the first sub-part of the first part of this question is equivalent to h being convex. This proves what was desired.

(c) Let $f : \mathbb{R}^n \to \mathbb{R}$ be a differentiable function. Assume that $\mathrm{dom}(f)$ is a convex set Show that f that satisfies (7) for all $x, y \in \mathrm{dom}(f)$ iff it satisfies

$$f(y) \le f(x) + \nabla f(x)^T (y - x) + \frac{L}{2}\|y - x\|_2^2, \tag{8}$$

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

Solution:
Let

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

with $\mathrm{dom}(h) = \mathrm{dom}(f)$, and note that

$$\nabla h(x) = Lx - \nabla f(x).$$

In the preceding part of this question we have established that h is convex. Hence, for all $x \neq y \in \mathrm{dom}(f) = \mathrm{dom}(h)$, we have

$$h(y) \geq h(x) + \nabla h(x)^T(y - x).$$

Substituting for h and ∇h in terms of f and ∇f respectively, this gives

$$\frac{L}{2}\|y\|_2^2 - f(y) \geq \frac{L}{2}\|x\|_2^2 - f(x) + (Lx - \nabla f(x))^T (y - x).$$

Rearranging this gives (8).

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 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 $\mathrm{dom}(g)$, 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 $x, y \in \mathrm{dom}(g)$. Thus a differentiable function $f : \mathbb{R}^n \to \mathbb{R}$ is L-smooth precisely when $\nabla f : \mathbb{R}^n \to \mathbb{R}^n$, with $\mathrm{dom}(\nabla f) := \mathrm{dom}(f)$, 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 $\mathrm{dom}(f)$ be a convex set and $\lambda_{\max}(\nabla^2 f(x)) \leq L$ for all $x \in \mathrm{dom}(f)$. Thus this is an equivalent characterization of L-smoothness for twice differentiable functions.