CS7643: Deep Learning
Fall 2017
Problem Set 1 – Solutions

Instructor: Dhruv Batra
TAs: Michael Cogswell, Abhishek Das, Zhaoyang Lv
Discussions: http://piazza.com/gatech/fall2017/cs7643

Due: Friday, Sep 8, 11:55pm

1.
$$\mathbf{w}^{(t+1)} = \underset{\mathbf{w}}{\operatorname{argmin}}\, f(\mathbf{w}^{(t)}) + \langle \mathbf{w} - \mathbf{w}^{(t)}, \nabla f(\mathbf{w}^{(t)}) \rangle + \frac{\lambda}{2} \left\| \mathbf{w} - \mathbf{w}^{(t)} \right\|^2 \tag{1}$$

Taking derivative with respect to $\mathbf{w}$ and equating it to 0 yields

$$\mathbf{w}^{(t+1)} = \mathbf{w}^{(t)} - \frac{1}{\lambda} \nabla f(\mathbf{w}^{(t)}) \tag{2}$$

Gradient descent minimizes the quadratic approximation at every time step.

Comparing (2) with the update equation, $\eta = \frac{1}{\lambda}$.

2. See section 14.1.1 from the book "Understanding Machine Learning: From Theory to Algorithms" by Shai Shalev-Schwartz and Shai Ben-David (relevant section attached).

3. Same as above.

4. No, SGD does not guarantee to decrease the objective function at every iteration.
Let $w^{(0)} = 0$. If the second subfunction $\frac{1}{2}(w+1)^2$ is sampled, $w^{(1)} = w^{(0)} - \eta(w^{(0)} + 1) = -\eta$. Since η is positive, $f(w^{(1)}) > f(w^{(0)})$.

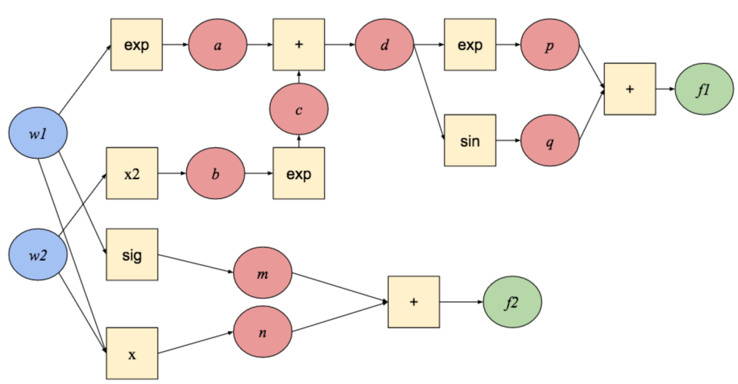

Figure 1: Computation graph figure for Q5, submitted by Kumar Ashis Pati

5. a) $f(1,2) = (7.802 \times 10^{24}, 2.731)$

b) $\frac{\partial \mathbf{f}}{\partial \mathbf{w}} \approx \left[\frac{\mathbf{f}(1+\delta,2)-\mathbf{f}(1,2)}{0.01} \; \frac{\mathbf{f}(1,2+\delta)-\mathbf{f}(1,2)}{0.01} \right]$

$\frac{\partial \mathbf{f}}{\partial \mathbf{w}} \approx \begin{bmatrix} 2.161 \times 10^{25} & 1.571 \times 10^{27} \\ 2.19 & 1.00 \end{bmatrix}$

c, d) $\frac{\partial \mathbf{f}}{\partial \mathbf{w}} = \begin{bmatrix} 2.121 \times 10^{25} & 8.520 \times 10^{27} \\ 2.197 & 1.000 \end{bmatrix}$

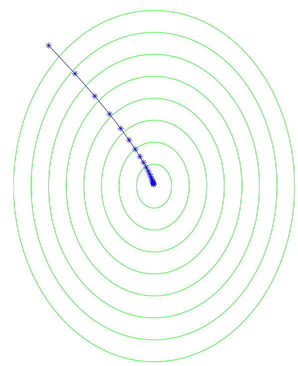

Figure 14.1 An illustration of the gradient descent algorithm. The function to be minimized is $1.25(x_1+6)^2+(x_2-8)^2$.

14.1.1 Analysis of GD for Convex-Lipschitz Functions

To analyze the convergence rate of the GD algorithm, we limit ourselves to the case of convex-Lipschitz functions (as we have seen, many problems lend themselves easily to this setting). Let $\mathbf{w}^\star$ be any vector and let B be an upper bound on $\|\mathbf{w}^\star\|$. It is convenient to think of $\mathbf{w}^\star$ as the minimizer of $f(\mathbf{w})$, but the analysis that follows holds for every $\mathbf{w}^\star$.

We would like to obtain an upper bound on the suboptimality of our solution with respect to $\mathbf{w}^\star$, namely, $f(\bar{\mathbf{w}})-f(\mathbf{w}^\star)$, where $\bar{\mathbf{w}}=\frac{1}{T}\sum_{t=1}^T \mathbf{w}^{(t)}$. From the definition of $\bar{\mathbf{w}}$, and using Jensen's inequality, we have that

$$\begin{aligned} f(\bar{\mathbf{w}})-f(\mathbf{w}^\star) &= f\left(\tfrac{1}{T}\sum_{t=1}^T \mathbf{w}^{(t)}\right)-f(\mathbf{w}^\star) \\ &\le \frac{1}{T}\sum_{t=1}^T\left(f(\mathbf{w}^{(t)})\right)-f(\mathbf{w}^\star) \\ &= \frac{1}{T}\sum_{t=1}^T\left(f(\mathbf{w}^{(t)})-f(\mathbf{w}^\star)\right). \end{aligned} \tag{14.2}$$

For every t, because of the convexity of f, we have that

$$f(\mathbf{w}^{(t)})-f(\mathbf{w}^\star) \;\le\; \langle \mathbf{w}^{(t)}-\mathbf{w}^\star, \nabla f(\mathbf{w}^{(t)})\rangle. \tag{14.3}$$

Combining the preceding we obtain

$$f(\bar{\mathbf{w}})-f(\mathbf{w}^\star) \;\le\; \frac{1}{T}\sum_{t=1}^T \langle \mathbf{w}^{(t)}-\mathbf{w}^\star, \nabla f(\mathbf{w}^{(t)})\rangle.$$

To bound the right-hand side we rely on the following lemma:

LEMMA 14.1 *Let $\mathbf{v}_1, \ldots, \mathbf{v}_T$ be an arbitrary sequence of vectors. Any algorithm with an initialization $\mathbf{w}^{(1)} = 0$ and an update rule of the form*

$$\mathbf{w}^{(t+1)} = \mathbf{w}^{(t)} - \eta \mathbf{v}_t \tag{14.4}$$

satisfies

$$\sum_{t=1}^{T} \langle \mathbf{w}^{(t)} - \mathbf{w}^\star, \mathbf{v}_t \rangle \leq \frac{\|\mathbf{w}^\star\|^2}{2\eta} + \frac{\eta}{2} \sum_{t=1}^{T} \|\mathbf{v}_t\|^2. \tag{14.5}$$

In particular, for every $B, \rho > 0$, if for all t we have that $\|\mathbf{v}_t\| \leq \rho$ and if we set $\eta = \sqrt{\frac{B^2}{\rho^2 T}}$, then for every $\mathbf{w}^\star$ with $\|\mathbf{w}^\star\| \leq B$ we have

$$\frac{1}{T} \sum_{t=1}^{T} \langle \mathbf{w}^{(t)} - \mathbf{w}^\star, \mathbf{v}_t \rangle \leq \frac{B\,\rho}{\sqrt{T}}.$$

Proof Using algebraic manipulations (completing the square), we obtain:

$$\begin{aligned}
\langle \mathbf{w}^{(t)} - \mathbf{w}^\star, \mathbf{v}_t \rangle &= \frac{1}{\eta} \langle \mathbf{w}^{(t)} - \mathbf{w}^\star, \eta \mathbf{v}_t \rangle \\
&= \frac{1}{2\eta} (-\|\mathbf{w}^{(t)} - \mathbf{w}^\star - \eta \mathbf{v}_t\|^2 + \|\mathbf{w}^{(t)} - \mathbf{w}^\star\|^2 + \eta^2 \|\mathbf{v}_t\|^2) \\
&= \frac{1}{2\eta} (-\|\mathbf{w}^{(t+1)} - \mathbf{w}^\star\|^2 + \|\mathbf{w}^{(t)} - \mathbf{w}^\star\|^2) + \frac{\eta}{2} \|\mathbf{v}_t\|^2,
\end{aligned}$$

where the last equality follows from the definition of the update rule. Summing the equality over t, we have

$$\sum_{t=1}^{T} \langle \mathbf{w}^{(t)} - \mathbf{w}^\star, \mathbf{v}_t \rangle = \frac{1}{2\eta} \sum_{t=1}^{T} \left(-\|\mathbf{w}^{(t+1)} - \mathbf{w}^\star\|^2 + \|\mathbf{w}^{(t)} - \mathbf{w}^\star\|^2 \right) + \frac{\eta}{2} \sum_{t=1}^{T} \|\mathbf{v}_t\|^2. \tag{14.6}$$

The first sum on the right-hand side is a telescopic sum that collapses to

$$\|\mathbf{w}^{(1)} - \mathbf{w}^\star\|^2 - \|\mathbf{w}^{(T+1)} - \mathbf{w}^\star\|^2.$$

Plugging this in Equation (14.6), we have

$$\begin{aligned}
\sum_{t=1}^{T} \langle \mathbf{w}^{(t)} - \mathbf{w}^\star, \mathbf{v}_t \rangle &= \frac{1}{2\eta} (\|\mathbf{w}^{(1)} - \mathbf{w}^\star\|^2 - \|\mathbf{w}^{(T+1)} - \mathbf{w}^\star\|^2) + \frac{\eta}{2} \sum_{t=1}^{T} \|\mathbf{v}_t\|^2 \\
&\leq \frac{1}{2\eta} \|\mathbf{w}^{(1)} - \mathbf{w}^\star\|^2 + \frac{\eta}{2} \sum_{t=1}^{T} \|\mathbf{v}_t\|^2 \\
&= \frac{1}{2\eta} \|\mathbf{w}^\star\|^2 + \frac{\eta}{2} \sum_{t=1}^{T} \|\mathbf{v}_t\|^2,
\end{aligned}$$

where the last equality is due to the definition $\mathbf{w}^{(1)} = 0$. This proves the first part of the lemma (Equation (14.5)). The second part follows by upper bounding $\|\mathbf{w}^\star\|$ by B, $\|\mathbf{v}_t\|$ by ρ, dividing by T, and plugging in the value of η. $\square$

Lemma 14.1 applies to the GD algorithm with $\mathbf{v}_t = \nabla f(\mathbf{w}^{(t)})$. As we will show later in Lemma 14.7, if f is ρ-Lipschitz, then $\|\nabla f(\mathbf{w}^{(t)})\| \le \rho$. We therefore satisfy the lemma's conditions and achieve the following corollary:

COROLLARY 14.2 *Let f be a convex, ρ-Lipschitz function, and let $\mathbf{w}^\star \in \operatorname{argmin}_{\{\mathbf{w}:\|\mathbf{w}\|\le B\}} f(\mathbf{w})$. If we run the GD algorithm on f for T steps with $\eta = \sqrt{\frac{B^2}{\rho^2 T}}$, then the output vector $\bar{\mathbf{w}}$ satisfies*

$$f(\bar{\mathbf{w}}) - f(\mathbf{w}^\star) \;\le\; \frac{B\,\rho}{\sqrt{T}}.$$

Furthermore, for every $\epsilon > 0$, to achieve $f(\bar{\mathbf{w}}) - f(\mathbf{w}^\star) \le \epsilon$, it suffices to run the GD algorithm for a number of iterations that satisfies

$$T \ge \frac{B^2\rho^2}{\epsilon^2}.$$

14.2 Subgradients

The GD algorithm requires that the function f be differentiable. We now generalize the discussion beyond differentiable functions. We will show that the GD algorithm can be applied to nondifferentiable functions by using a so-called subgradient of $f(\mathbf{w})$ at $\mathbf{w}^{(t)}$, instead of the gradient.

To motivate the definition of subgradients, recall that for a convex function f, the gradient at $\mathbf{w}$ defines the slope of a tangent that lies below f, that is,

$$\forall \mathbf{u}, \quad f(\mathbf{u}) \ge f(\mathbf{w}) + \langle \mathbf{u} - \mathbf{w}, \nabla f(\mathbf{w})\rangle. \tag{14.7}$$

An illustration is given on the left-hand side of Figure 14.2.

The existence of a tangent that lies below f is an important property of convex functions, which is in fact an alternative characterization of convexity.

LEMMA 14.3 *Let S be an open convex set. A function $f : S \to \mathbb{R}$ is convex iff for every $\mathbf{w} \in S$ there exists $\mathbf{v}$ such that*

$$\forall \mathbf{u} \in S, \quad f(\mathbf{u}) \ge f(\mathbf{w}) + \langle \mathbf{u} - \mathbf{w}, \mathbf{v}\rangle. \tag{14.8}$$

The proof of this lemma can be found in many convex analysis textbooks (e.g., (Borwein & Lewis 2006)). The preceding inequality leads us to the definition of subgradients.

DEFINITION 14.4 (Subgradients) A vector $\mathbf{v}$ that satisfies Equation (14.8) is called a *subgradient* of f at $\mathbf{w}$. The set of subgradients of f at $\mathbf{w}$ is called the *differential set* and denoted $\partial f(\mathbf{w})$.

An illustration of subgradients is given on the right-hand side of Figure 14.2. For scalar functions, a subgradient of a convex function f at w is a slope of a line that touches f at w and is not above f elsewhere.