CS 189 Introduction to Machine Learning
Fall 2018 DIS6

# 1 Convergence Behavior of Gradient Descent

Let $f : \mathbb{R}^d \to \mathbb{R}$ be a differentiable function. Let $f_\star = \min_x f(x)$ and suppose that $f_\star$ is finite (i.e. $f_\star > -\infty$). In this question, we will look at the convergence of gradient descent under several different assumptions on the function $f$. Recall that gradient descent starts by choosing an $x_0 \in \mathbb{R}^d$ and iterates:

$$x_{k+1} = x_k - \alpha_k \nabla f(x_k) \,,$$

where $\{\alpha_k\}_{k \geq 0}$ is a sequence of step sizes.

(a) Suppose that $f$ is twice differentiable and that the Hessian $\nabla^2 f(x)$ satisfies the uniform upper bound $\nabla^2 f(x) \preceq LI$ for all $x \in \mathbb{R}^d$. Suppose we use a fixed step size $\alpha_k \equiv \alpha$. Show that for all $k$:

$$f(x_{k+1}) \leq f(x_k) + \left(-\alpha + \frac{L\alpha^2}{2}\right) \left\|\nabla f(x_k)\right\|_2^2 .$$

*Hint:* Recall that Taylor's theorem states that for all $x, y \in \mathbb{R}^d$, we have:

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

with $\tilde{x} = tx + (1-t)y$ for some $t \in [0, 1]$.

(b) Minimize the right hand side of the bound above to show that for an appropriate choice of $\alpha$, we have,

$$f(x_{k+1}) \leq f(x_k) - \frac{1}{2L} \left\|\nabla f(x_k)\right\|_2^2 .$$

(c) After $k$ iterations, show that either (a) $f(x_k) = f_\star$ or (b) $\min_{0 \leq \ell \leq k-1} \left\|\nabla f(x_\ell)\right\|_2^2 \leq \frac{2L(f(x_0) - f_\star)}{k}$.

(d) Now suppose furthermore that $f$ satisfies the condition $\frac{1}{2}\left\|\nabla f(x)\right\|^2 \geq m(f(x) - f_\star)$ for all $x \in \mathbb{R}^d$. Show that we now have:

$$f(x_k) - f_\star \leq \left(1 - \frac{m}{L}\right)^k (f(x_0) - f_\star) \,.$$

Conclude that at most $k = \frac{L}{m} \log\big((f(x_0) - f_\star)/\varepsilon\big)$ iterations are sufficient to achieve $f(x_k) - f_\star \leq \varepsilon$.

 

(e) Let $A$ be a symmetric positive definite matrix. Show that the function $f(x) = \frac{1}{2}x^\top Ax - x^\top b$ satisfies $\nabla^2 f(x) \preceq LI$ and $\frac{1}{2}\left\|\nabla f(x)\right\|^2 \geq m(f(x) - f_\star)$ with $L = \lambda_{\max}(A)$ and $m = \lambda_{\min}(A)$.

(f) Consider again the function from the last part, $f(x) = \frac{1}{2}x^\top Ax - x^\top b$. Suppose now that instead of using a fixed step size $\alpha_k \equiv \alpha$, we want to use exact line search. Specifically, we want to set $\alpha_k$ as:

$$\alpha_k = \arg\min_\alpha f(x_k - \alpha \nabla f(x_k)) \,.$$

Show that:

$$\alpha_k = \frac{\left\|\nabla f(x_k)\right\|^2}{\nabla f(x_k)^\top A \nabla f(x_k)} \,.$$

# 2 Clip Loss

In lecture, you saw the example of different loss functions like the squared-error loss and the hinge-loss. This question explores a different loss function.

Let $S = \{(x_1, y_1), \ldots (x_n, y_n)\}$ be a set of $n$ points sampled i.i.d. from a distribution $\mathcal{D}$. This is the training set with $x_i \in \mathbb{R}^d$ being the features and $y_i \in \{-1, 1\}$ being the labels.

We are thinking about a linear classifier that is going to look at the sign of $w^\top x$ to make a decision as to whether the label is $+1$ or $-1$.

Define the *clip loss* of a linear classifier $w \in \mathbb{R}^d$ as

$$\text{loss}(w^\top x, y) = \text{clip}(yw^\top x)$$

Where clip is the function

$$\text{clip}(z) = \begin{cases} 1 & \text{if } z < 0 \\ 0 & \text{if } z \geq 1 \\ 1 - z & \text{otherwise.} \end{cases}$$

For any $d$-dimensional vector $w$, define the *risk* of $w$ as

$$R[w] = \mathbb{E}_{(x,y)\sim\mathcal{D}}[\text{loss}(w^\top x, y)] \,,$$

and the *empirical risk* of $w$ as

$$R_S[w] = \frac{1}{n}\sum_{i=1}^{n} \text{loss}(w^\top x_i, y_i) \,.$$

(a) Draw the clip loss function. **Is the function** clip **convex?** Justify your answer.

(b) **Prove that if** $R_S[w] = 0$ **and** $\|w\|_2 = 1$**, then the hyperplane defined by** $w$ **has a classification margin** $\geq 1$ **on this training set.**

(c) **Prove that** $\mathbb{E}_S[R_S[w]] = R[w]$**.** Here, the outer expectation is being taken over the randomly drawn training set.

(d) **Prove that** $\mathrm{Var}(R_S[w]) \leq \frac{1}{n}$**.**

(e) **Is it possible to have an** $S$ **and** $w$ **such that** $R_S[w] = 0$**, but** $R[w] > 0$**?** Justify your answer.