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 $t \in [0, 1]$.

Solution: We use Taylor's theorem with the choice $x = x_k$ and $y = x_{k+1} = x_k - \alpha \nabla f(x_k)$. Observe that $y - x = -\alpha \nabla f(x_k)$. Since $\nabla^2 f(x) \preceq LI$ for any x, this means:

$$\begin{aligned} f(x_{k+1}) &\leq f(x_k) - \alpha \langle \nabla f(x_k), \nabla f(x_k) \rangle + \frac{\alpha^2 L}{2} \left\|\nabla f(x_k)\right\|_2^2 \\ &= f(x_k) - \alpha \left\|\nabla f(x_k)\right\|_2^2 + \frac{\alpha^2 L}{2} \left\|\nabla f(x_k)\right\|_2^2 \,. \end{aligned}$$

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

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

Solution: Define the function $g(\alpha) = -\alpha + L\alpha^2/2$. By setting $g'(\alpha) = 0$, we obtain that the minimizer is $\alpha_\star = 1/L$ which attains the value $g(\alpha_\star) = -\frac{1}{2L}$.

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

Solution: Define $\varepsilon = \frac{2L(f(x_0)-f_\star)}{k}$. Suppose that condition (b) does not hold. This means that $\left\| \nabla f(x_\ell) \right\|_2^2 > \varepsilon$ for all $\ell = 0, ..., k-1$. Unroll the recursion $f(x_{k+1}) \le f(x_k) - \frac{1}{2L} \left\| \nabla f(x_k) \right\|_2^2$ down to $f(x_0)$ to conclude that:

$$f(x_k) \le f(x_0) - \frac{1}{2L} \sum_{\ell=0}^{k-1} \left\| \nabla f(x_\ell) \right\|_2^2 .$$

Subtracting $f_\star$ on both sides yields

$$\begin{aligned} f(x_k) - f_\star &\le f(x_0) - f_\star - \frac{1}{2L} \sum_{\ell=0}^{k-1} \left\| \nabla f(x_\ell) \right\|_2^2 \\ &< f(x_0) - f_\star - \frac{1}{2L} k\varepsilon \\ &= 0 . \end{aligned}$$

Since the LHS is ≥ 0 by definition, we have sandwiched $0 \le f(x_k) - f_\star \le 0$ from which we conclude $f(x_k) = f_\star$ which is condition (a). Hence either (b) holds, or if it does not then (a) holds.

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

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

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

Solution: Subtracting $f_\star$ from the descent inequality above we have:

$$\begin{aligned} f(x_{k+1}) - f_\star &\le f(x_k) - f_\star - \frac{1}{2L} \left\| \nabla f(x_k) \right\|_2^2 \\ &\le f(x_k) - f_\star - \frac{m}{L}(f(x_k) - f_\star) \\ &= \left(1 - \frac{m}{L} \right) (f(x_k) - f_\star) . \end{aligned}$$

The first claim now follows by unrolling the recursion down to $f(x_0) - f_\star$.

To obtain the bound on k, use the hint to bound:

$$f(x_k) - f_\star \le \left(1 - \frac{m}{L} \right)^k (f(x_0) - f_\star) \le e^{-(m/L)k} (f(x_0) - f_\star)$$

Now set the RHS $\le \varepsilon$ and solve for k.

(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 $m = \lambda_{\min}(A)$.

Solution: We recall that $\nabla f(x) = Ax - b$ and $\nabla^2 f(x) = A$. Hence the fact that we can take $L = \lambda_{\max}(A)$ is immediate.

For the other inequality, we first note that $x_\star = A^{-1}b$ and hence $x - x_\star = x - A^{-1}b = A^{-1}(Ax - b) = A^{-1}\nabla f(x)$. Since f is quadratic, its second order Taylor expansion is exact. Therefore, recalling that $\nabla f(x_\star) = 0$,

$$\begin{aligned} f(x_k) &= f_\star + \frac{1}{2}(x_k - x_\star)^\top A(x_k - x_\star) \\ &= f_\star + \frac{1}{2}\nabla f(x_k)^\top A^{-1}AA^{-1}\nabla f(x_k) \\ &= f_\star + \frac{1}{2}\nabla f(x_k)^\top A^{-1}\nabla f(x_k) \\ &\leq f_\star + \frac{\lambda_{\max}(A^{-1})}{2}\left\|\nabla f(x_k)\right\|_2^2 \\ &= f_\star + \frac{1}{2\lambda_{\min}(A)}\left\|\nabla f(x_k)\right\|_2^2 . \end{aligned}$$

Rearranging the last inequality yields the desired inequality.

(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 α_k as:

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

Show that:

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

Solution: Using the calculation for $\nabla f(x)$ and $\nabla f^2(x)$ from the previous question, we observe that:

$$f(x_k - \alpha\nabla f(x_k)) = f(x_k) - \alpha\left\|\nabla f(x_k)\right\|^2 + \frac{\alpha^2}{2}\nabla f(x_k)^\top A\nabla f(x_k) .$$

As before, we define $g(\alpha) = -\alpha\left\|\nabla f(x_k)\right\|^2 + \frac{\alpha^2}{2}\nabla f(x_k)^\top A\nabla f(x_k)$. Setting $g'(\alpha) = 0$ and solving for $\alpha_\star$ yields the claimed solution.

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 $S = \{(x_1, y_1), \dots (x_n, y_n)\}$ 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

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

Where clip is the function

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

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

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

and the *empirical risk* of w as

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

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

Solution: It is not convex. Drawing the function shows that the line from $(-1, 1)$ to $(1, 0)$ lies below the graph of the clip function.

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

Solution: The margin of the normalized hyperplane is defined as

$$\min_{1 \leq i \leq n} y_i(w^\top x) \, .$$

If $R_S[w] = 0$, then since $\mathrm{clip}(z) \geq 0$ this quantity is greater than or equal to 1 for all $1 \leq i \leq n$.

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

Solution:

$$\mathbb{E}\left[\frac{1}{n}\sum_{i=1}^{n} \mathrm{loss}(w^\top x_i, y_i)\right] = \frac{1}{n}\sum_{i=1}^{n} \mathbb{E}\left[\mathrm{loss}(w^\top x_i, y_i)\right] = \frac{1}{n}\sum_{i=1}^{n} R[w] = R[w]$$

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

Solution:

$$\begin{aligned}
\mathrm{Var}(R_S[w]) &= \mathbb{E}\left[(R_S[w] - R[w])^2\right] \\
&= \frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n}\mathbb{E}(\mathrm{loss}(w^\top x_i, y_i) - R[w])(\mathrm{loss}(w^\top x_j, y_j) - R[w]) \\
&= \frac{1}{n^2}\sum_{i=1}^{n}\mathbb{E}\left[(\mathrm{loss}(w^\top x_i, y_i) - R[w])^2\right] \\
&= \frac{1}{n}\mathbb{E}\left[(\mathrm{loss}(w^\top x, y) - R[w])^2\right] \\
&\leq \frac{1}{n}
\end{aligned}$$

Here, the first line is the definition of variance, the second line expands the square, the third line follows because (x_i, y_i) and (x_j, y_j) are independent. The fourth line follows because the (x_i, y_i) are identically distributed. The last line follows because the clip loss is nonnegative and bounded above by 1.

Alternate proof of first 4 steps:

$$\begin{aligned}
\mathrm{Var}(R_S[w]) &= \mathrm{Var}(\frac{1}{n}\sum_{i=1}^{n}\mathrm{loss}(w^\top x_i, y_i)) \\
&= \frac{1}{n^2}\mathrm{Var}(\sum_{i=1}^{n}\mathrm{loss}(w^\top x_i, y_i)) \\
&= \frac{1}{n^2}\sum_{i=1}^{n}\mathrm{Var}(\mathrm{loss}(w^\top x_i, y_i), \text{ by i.i.d} \\
&= \frac{1}{n}\mathrm{Var}(\mathrm{loss}(w^\top x, y))
\end{aligned}$$

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

Solution: Yes. Consider the case when $n = 1$. Then it is possible to classify the single data point correctly while classifying all of the opposite class incorrectly.