CS 49/149: 21st Century Algorithms (Fall 2018)

Date: 2nd October, 2018
Topic: Gradient Descent on Smooth and Strongly Convex Objective Functions
Scribe: Edward Yao

1 "Smooth" functions and gradient descent

It turns out that with stronger assumptions on the objective functions we are optimizing with respect to, we can obtain stronger bounds on the rates of convergence of the gradient descent algorithm we explored last time in class. In the analysis last class, we showed that we could obtain a an error under ε with gradient descent (for a reminder, see the appendix) if we ran the algorithm for $T = \frac{||x_1 - x_*||_2 * \rho^2}{\varepsilon^2}$ time (where x_1 is the initial point the gradient descent starts at and x_* is the optimum point minimizing the objective function). We will now investigate one such example where we can make a stronger assumption on the objective function, which will enable us to obtain a faster rate of convergence for gradient descent.

Definition: L-smooth in the euclidean norm
A function f is L-smooth in the euclidean norm if for all x, y, $||\nabla f(x) - \nabla f(y)||_2 < L \cdot ||x - y||_2$.

Theorem 1. If a function f is L-smooth, $\forall x, y : f(y) \leq f(x) + (y - x)^T \nabla f(x) + \frac{L}{2}||x - y||_2^2$

Proof. Define function $g(t) = f(x + t(y - x))$ for $0 \leq t \leq 1$. Observe that because of the way we have constructed this function, $g(0) = f(x)$, and $g(1) = f(y)$. Then, we can take the derivative with respect to t of both sides of the equality to get

$$\frac{dg(t)}{dt} = (y - x)^T \nabla f(x + t(y - x))$$

We can then begin analysis of our function:

$$\begin{aligned}
f(y) - f(x) &= g(1) - g(0) \\
&= \int_0^1 g'(t) dt \\
&= \int_0^1 (y - x)^T \nabla f(x + t(y - x)) dt \\
&= \int_0^1 (y - x)^T [\nabla f(x + t(y - x)) - \nabla f(x)] dt + \int_0^1 (y - x)^T \nabla f(x) dt \\
&= \int_0^1 (y - x)^T [\nabla f(x + t(y - x)) - \nabla f(x)] dt + (y - x)^T \nabla f(x) \\
\implies f(y) - f(x) - (y - x)^T \nabla f(x) &= \int_0^1 (y - x)^T \nabla f(x + t(y - x)) dt \\
&\leq \int_0^1 ||y - x||_2 \cdot ||\nabla f(x) + t(y - x)) - \nabla f(x)||_2 dt \qquad \text{(Cauchy-Shwartz)}
\end{aligned}$$

$$
\begin{aligned}
&\leq \frac{L}{2}\int_0^1 ||y-x||_2 \cdot t||y-x||_2 dt && \text{(smoothness)}\\
&\leq \frac{L}{2}||y-x||_2^2 \int_0^1 tdt\\
&\leq \frac{L}{2}||y-x||_2^2
\end{aligned}
$$

□ ✍

Question: *Show that* $||\nabla f(x)||_2 \leq \rho \iff |f(x)-f(y)| \leq \rho \cdot ||x-y||_2$ *(known as the Lipschitz condition)*

Understanding these properties for smooth functions, let us now turn to analyzing gradient descent in the case where our objective function is a smooth function.

Gradient descent for smooth functions Recall that the update rule for gradient descent is $x_{t+1} = x_t - \eta\nabla f(x)$. Let us consider several facts that arise from our previous theorem, which shall help us in our convergence analysis.

1. Setting $x = x_*$ (i.e. the optimal value of x that minimizes the objective function) and $y = x_t$ (i.e. the value of x which gradient descent holds at time t), we know that $f(x_t) \leq f(x_*) + \frac{L}{2}||x_t - x_*||$, written another way as $error(t) \leq \frac{L}{2}D_t^2$, where ($D_t$ is $||x_t - x_*||$, i.e. the distance from our current point to the optimum at time t).

2. Setting $x = x_*$ and $y = x_{t+1}$, we get that

$$
\begin{aligned}
f(x_{t+1}) &\leq f(x_t) + (x_{t+1}-x_t)^T\nabla f(x_t) + \frac{L}{2}||x_{t+1}-x_t||_2^2\\
&= f(x_t) - \eta||\nabla f(x_t)||_2^2 + \frac{\eta^2 \cdot L}{2}||\nabla f(x_t)||_2^2\\
&= f(x_t) - ||\nabla f(x_t)||_2^2(\eta - \frac{\eta^2 \cdot L}{2})\\
&= f(x_t) - \frac{1}{2L}||\nabla f(x_t)||_2^2 && \text{(setting } \eta = \tfrac{1}{L}\text{)}
\end{aligned}
$$

   Note that $||\nabla f(x_t)||_2^2$ is a measure of distance to a saddle point (since the gradient of a vector function being zero does not guarantee local optimality).

With these two equations in our hands, we can now turn to analysis of of gradient descent. In particular, we would like to find how many steps are needed until our algorithm reaches a local optimum/saddle point. Adding together two equations we get from properties of convexity and smoothness,

$$
\begin{array}{rl}
f(x_t) - f(x_*) = error(t) \leq (x_t - x_*)^T\nabla(f(x_t)) & \text{(convexity)}\\
+ \quad f(x_{t+1}) - f(x_t) \leq (x_{t+1}-x_t)^T\nabla f(x_t) + \frac{L}{2}||x_{t+1}-x_t||_2^2 & \text{(smoothness)}\\
\hline
err(t+1) \leq (x_{t+1}-x_*)^T\nabla f(x_t) + \frac{L}{2}||x_{t+1}-x_t||_2^2 &
\end{array}
$$

Now, analyzing the last equation that came from the addition of the smoothness and convexity properties,

$$\begin{aligned}
err(t+1) &\leq (x_{t+1}-x_*)^T \nabla f(x_t) + \frac{L}{2}||x_{t+1}-x_t||_2^2 \\
&= \frac{1}{n}(x_{t+1}-x_*)^T(x_t - x_{t+1}) + \frac{L}{2}||x_{t+1}-x_t||_2^2 && \text{(gradient descent)} \\
&= \frac{1}{2n}(||x_t - x_*||_2^2 - ||x_{t+1}-x_*||_2^2 - ||x_{t+1}-x_t||_2^2) + \frac{L}{2}(||x_{t+1}-x_t||_2^2) && \text{(cosine identity)} \\
&= \frac{1}{2n}(D_t^2 - D_{t+1}^2) - ||x_{t+1}-x_t||_2^2 \cdot (\frac{1}{2\eta} \cdot \frac{L}{2}) \\
&= \frac{L}{2}(D_t^2 - D_{t+1}^2) && (\text{if } \eta = \tfrac{1}{L})
\end{aligned}$$

If we run this algorithm for T steps, then $\sum_{i=1}^{T} err(t+1) \leq \frac{L}{2}(D_1^2 - D_f^2) \leq \frac{LD^2}{2}$, which implies that $f(x_t) - f(x_*) \leq \frac{LD^2}{2T}$ where $D = ||x_1 - x^*||_2^2$.

Recall that we also know for smooth functions that $f(x_{t+1}) \leq f(x_t) - \frac{1}{2L}||\nabla f(x)||_2^2$. Plugging in $T = \frac{LD^2}{2\varepsilon}$, we get that $f(x_t) - f(x_*) \leq \varepsilon$. ✍

Question: *We have shown in this analysis that the value of the objective function drops to under ε. Show that D_t, i.e. $||x_t - x_*||$, is also dropping over time.*

✍

Question: *This analysis was done in the context of unconstrained gradient descent. Show also that the same bounds hold for projected gradient descent.*

Examples of smoothness: Let us examine the function $f(x) = \frac{1}{2}x^T Q x$, where $Q \in \mathbb{R}^{n \times m}$. Then, we can see that $(\nabla f(x))_j = \frac{\partial f}{\partial j} = Q_{jj}x_j + \sum_{i \neq j} Q_{ij}x_i = (x^T Q) \implies \nabla f(x) = Q^T x$.
How smooth is this function?

$$\begin{aligned}
&||Q^T x - Q^T y||_2 \leq L||x-y||_2 \\
\implies &||Q^T(x-y)||_2 \leq L||x-y||_2 && (u = x - y) \\
&L = \max_{u \neq 0, u \in \mathbb{R}^n} \frac{||Q^T u||_2}{||u_2||_2} = \sqrt{\lambda_{\max}(QQ^T)}
\end{aligned}$$

To see this, we can expand out the numerator above:

$$||Q^T u||_2 = u^T Q Q^T u$$

If QQ^T had a largest value λ,

$$\begin{aligned} &\exists v, QQ^T v \leq \lambda v \\ &v^T(QQ^T v) \leq \lambda v^T v \\ \implies &\frac{||Q^T v||_2^2}{||v||_2^2} \leq \lambda \end{aligned}$$

According to gradient descent rules, for an objective function $f(x) = \frac{1}{2}x^T Qx$, our update rule should look like $x_{t+1} = x_t - \frac{1}{\delta}x_t^T Q$, where $\delta = \sqrt{\lambda_{\max}(QQ^T)}$. If we run this algorithm for about $O(\frac{1}{\varepsilon})$ iterations, we can get about ε-close to the global optimum (since this is a convex function). ✍

Question: *How smooth is* $f(x) = ||Ax - b||_2^2$*, where* $A \in \mathbb{R}^{m \times n}, b \in \mathbb{R}^m, x \in \mathbb{R}^n$*?*

2 Strongly convex functions

We saw in the previous section that if we know that our objective function is not only convex but also $L-$smooth, we can run the algorithm for a time that scales linearly with L, instead of quadratically with ρ. Another way that we can guarantee faster convergence than just on normal convex function is if we know the objective function is strongly convex and it is smooth.

Definition: Strongly Convex
A function $f : \mathbb{R}^n \to \mathbb{R}$ is $l-$strongly convex if $\forall x, y : f(y) \geq f(x) + (y - x)^T \nabla f(x) + \frac{l}{2}||y - x||_2^2$

The motivation for examining strong convexity is as follows: if a function f is both $l-$strongly convex and $L-$smooth, we can bound it between two quantities based on the properties of the functions, i.e.

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

which should help our analysis.

We now move to the analysis of f such that f is $l-$ strongly convex and $L-$smooth. From the definition of $l-$strongly convex, $err(t) = f(x_t) - f(x_*) \leq (x_t - x_*)^T \nabla f(x_t) - \frac{l}{2}||x_t - x_*||_2^2$. Much the same as in the analysis of the smooth function, we add the two properties from f being strongly convex and f being smooth together:

$$\begin{aligned} error(t) &\leq (x_t - x_*)^T \nabla f(x_t) - \frac{l}{2}D_t^2 && \text{(strong convexity)} \\ + \quad f(x_{t+1}) - f(x_t) &\leq (x_{t+1} - x_t)^T \nabla f(x_t) + \frac{L}{2}||x_{t+1} - x_t||_2^2 && \text{(smoothness)} \\ \hline err(t+1) &\leq \frac{L}{2}(D_t^2 - D_{t+1}^2) - \frac{l}{2}D_t^2 \end{aligned}$$

We know that this above quantity is lower bounded by 0 while the algorithm is still running. Thus,

$$\begin{aligned} 0 \leq err(t+1) &\leq \frac{L}{2}(D_t^2 - D_{t+1}^2) - \frac{l}{2}D_t^2 \\ \implies \frac{L}{2}D_{t+1}^2 &\leq \left(\frac{L-l}{2}\right) \cdot D_t^2 \\ \implies D_{t+1}^2 &\leq (1 - \frac{l}{L}) \cdot D_t^2 \end{aligned}$$

If we run gradient descent on this function for T stepsm we can see that $D_T^2 \leq (1 - \frac{l}{L})^T D_1^2$. Combine this with the fact that $f(x_t) - f(x_*) \leq \frac{L}{2}D_t^2$ for $L-$ smooth functions, and we get that $err(T) = f(x_T) - f(x_*) \leq \frac{L}{2}(1 - \frac{l}{L})^T \cdot D^2 = \varepsilon$. Solving this for T, we get T scales with the $\log(\frac{1}{\varepsilon})$, which means this function has "linear convergence" (though the mathematical form has a $\log$ in it).

The results here make sense - with even stronger restrictions on the objective functions, we get an even tighter bound on the number of steps needed for gradient descent to converge to within ε of the optimum objective function value.

In practice, we care a lot about the quantity $\frac{l}{L}$, which we call the condition number. Also note that for these bounds to even work, $l < L$. Otherwise, intuitively, the sandwiching we are doing of the function between two quantities cannot happen because the top bound may be smaller than the bottom bound.

3 Appendix: Useful Reminders

Cauchy-Shwartz inequality:

$$a^T b \leq ||a||_2 \cdot ||b||_2$$

Cosine identity:

$$a^T b = \frac{1}{2}(||a+b||_2^2 - ||a||_2^2 - ||b||_2^2)$$

```
UNCONSTRAINED GRADIENT DESCENT
  • $x_1$ := arbitrary point
  • run for iterations $t = 1, \ldots, T$:
      – $x_{t+1} = x_t - \eta_t \nabla f(x_t)$
```