CSE 547/Stat 548: Machine Learning for Big Data **Lecture**

Dimension Free Optimization and Non-Convex Optimization

Instructor: Sham Kakade

1 Non-convex optimization and Black-Box Oracle Complexity

Suppose we are trying to minimize the function $F(w)$. What can we hope to achieve with a method which provides us with gradients of F? In particular, we can think of having an oracle which when provided with w as input, returns $\nabla F(w)$.

A basic question would be what might we hope to achieve and how many gradient computations are needed to achieve this? In the non-convex setting, the most minimal thing we might hope for is to (quickly?) converge to a stationary point, i.e. a point where the gradient is near to 0 (or near to 0). Note this does not necessarily imply that we are even at a local minima, which is a far more subtle issue.

Regardless, we will now review some basic "dimension free" results for how we can find such stationary points.

Smoothness

Let us say a function $F : \mathbb{R}^d \to \mathbb{R}$ is L *-smooth* if

$$\|\nabla F(w) - \nabla F(w')\| \leq L\|w - w'\| \,,$$

where the norm is the Euclidean norm. In other words, the derivatives of F do not change too quickly. If the Hessian exists, then smoothness implies the Hessian is bounded.

Smoothness implies the following:

$$F(w + \Delta) \leq F(w) + \nabla F(w)^\top \Delta + \frac{L}{2}\|\Delta\|^2 \,.$$

In other words, it gives us an (upper) bound on the error in Taylor's theorem. (Taylor's theorem plus the intermediate value theorem implies the previous inequality).

1.1 Gradient Descent converges to (first-order) Stationary Points

Gradient descent, with a constant learning rate, is the algorithm:

$$w^{(k+1)} = w^{(k)} - \eta \cdot \nabla F(w^{(k)})$$

Here, we do *not* assume that F is convex. Also, we do not need to assume that F is twice differentiable.

Theorem 1.1. *(GD finds Stationary Points) Let F_* be the minimal function value (i.e. the value at the global minima). Using $\eta = 1/L$, Gradient descent will find a $w^{(k)}$ that is "almost" a stationary point in a bounded (and polynomial) number of steps. Precisely,*

$$\min_{k<K} \|\nabla F(w^{(k)})\|^2 \leq \frac{2L(F(w^{(0)}) - F_*)}{K} \,.$$

(Note that $\|\nabla F(w^{(k)})\|$ may not be decreasing at every step.)

Proof. Smoothness implies that:

$$F(w^{(k+1)}) \leq F(w^{(k)}) - \eta\|\nabla F(w^{(k)})\|^2 + \frac{1}{2}\eta^2 L\|\nabla F(w^{(k)})\|^2$$

Our setting of η implies:

$$F(w^{(k+1)}) \leq F(w^{(k)}) - \frac{1}{2L}\|\nabla F(w^{(k)})\|^2$$

Using that the min is less than the average and by summing over k,

$$\begin{aligned}
\min_{0\leq k<K} \|\nabla F(w^{(k)})\|^2 &\leq \frac{1}{K}\sum_{t=0}^{K-1} \|\nabla F(w^{(k)})\|^2 \\
&\leq \frac{2L}{K}\sum_{t=0}^{K-1} \Big(F(w^{(k)}) - F(w^{(k+1)})\Big) \\
&= \frac{2L}{K}\Big(F(w^{(0)}) - F(w^{(K)})\Big) \\
&\leq \frac{2L}{K}\Big(F(w^{(0)}) - F_*\Big)
\end{aligned}$$

which completes the proof. □

1.2 Gradient Descent, plus a little noise, converges to (second order) Stationary Points

See the readings.

1.3 SGD finds Stationary Points

For SGD, we provide the argument due to [Ghadimi and Lan(2013)]

Assume we have an N sized training set $\mathcal{T}$.

Define:

$$F(w) = \frac{1}{N}\sum_{(x,y)\in\mathcal{T}} \ell(w,(x,y))$$

Gradient descent, with a constant learning rate, is the algorithm:

1. Initialize at some $w^{(0)}$.
2. Sample (x, y) uniformly at random from the set $\mathcal{T}$
3. Update the parameters:
$$w^{(k+1)} = w^{(k)} - \eta_k \cdot \nabla\ell(w^{(k)},(x,y))$$
and go back to 2.

Here, we do *not* assume that F is convex. Also, we do not need to assume that F is twice differentiable.

Theorem 1.2. *Let us run SGD for K steps. Suppose our gradient is bounded as follows: $\nabla \ell(w,(x,y)) \leq B$ for all w and examples (x,y). Assume our (constant) learning rate is $\eta_k = \eta = c/\sqrt{K}$, where $c = \sqrt{\frac{2(F(w^{(0)})-F_*)}{LB^2}}$). We have that:*

$$\min_{k<K} \mathbb{E}[\|\nabla F(w^{(k)})\|^2] \leq B\sqrt{\frac{2(F(w^{(0)})-F_*)L}{K}}$$

where the expectation is with respect to the random sampling in our algorithm.

It is interesting to compare the complexity of SGD with GD. Importantly, note the convergence rate of SGD does not depend on N.

Remark: The above bound implicitly assumes we know the end iteration K in advance. Alternatively, we could adaptive set $\eta_k = O(1/\sqrt{k})$ to obtain the same bound (up to constant factors). The proof is simpler when we know K in advance.

Proof. Denote the sampled gradient at iteration k by $\widehat{\nabla F(w^{(k)})}$. From smoothness of F and the gradient descent update rule, we get,

$$\begin{aligned}
\mathbb{E}F(w^{(k+1)}) &= \mathbb{E}F(w^{(k)} + w^{(k+1)} - w^{(k)}) \\
&\leq \mathbb{E}\left[F(w^{(k)}) + \nabla F(w^{(k)})^\top (w^{(k+1)} - w^{(k)}) + \frac{L}{2}\|w^{(k+1)} - w^{(k)}\|^2\right] \\
&= \mathbb{E}\left[F(w^{(k)}) - \eta \nabla F(w^{(k)})^\top \widehat{\nabla F(w^{(k)})} + \eta^2 \frac{L}{2}\|\widehat{\nabla F(w^{(k)})}\|^2\right] \\
&\leq \mathbb{E}[F(w^{(k)})] - \eta E\|\nabla F(w^{(k)})\|^2 + \eta^2 \frac{LB^2}{2}
\end{aligned}$$

Rearranging gives:

$$E\|\nabla F(w^{(k)})\|^2 \leq \frac{1}{\eta}\left(\mathbb{E}[F(w^{(k)})] - \mathbb{E}[F(w^{(k+1)})]\right) + \eta\frac{LB^2}{2}$$

Summing over k gives:

$$\begin{aligned}
\min_{0\leq k<K} E\|\nabla F(w^{(k)})\|^2 &\leq \frac{1}{K}\sum_{t=0}^{K-1} E\|\nabla F(w^{(k)})\|^2 \\
&\leq \frac{1}{K\eta}\left(\mathbb{E}[F(w^{(0)})] - \mathbb{E}[F(w^{(K)})]\right) + \eta\frac{LB^2}{2} \\
&\leq \frac{1}{K\eta}\left(F(w^{(0)}) - F_*\right) + \eta\frac{LB^2}{2}
\end{aligned}$$

and our choice of η leads to the result. Note that our choice of η is the one which minimizes this upper bound. □

1.4 Adaptive Gradient Methods

This is an argument due Krishna Pillutla.

Let us consider the gradient descent iteration $w^{(k+1)} = w^{(k)} - \eta_k \nabla F(w^{(k)})$. In this section, we shall analyze the effect of setting step-sizes as $\eta_k = C/\sqrt{\sum_{j=0}^{k} \|\nabla F(w^{(j)})\|^2}$, where C is a constant.

Theorem 1.3. *Suppose F is L-smooth and bounded from below by F_*. Then, gradient descent with adaptive step-sizes $\eta_k = C/\sqrt{\sum_{j=0}^{k} \|\nabla F(w^{(j)})\|^2}$ produces a sequence of iterates $\{w^{(k)}\}_{k \geq 0}$ such that*

$$\min_{j \leq k} \|\nabla F(w^{(j)})\|^2 \leq \frac{4}{C^2} \cdot \frac{(F(w^{(0)}) - F_*)^2}{k+1}$$

provided $C \leq \frac{\|\nabla F(w^{(0)})\|}{L}$.

Proof. Define $\Delta_k := F(w^{(k)}) - F_*$. From smoothness of F and the gradient descent update rule, we get,

$$\begin{aligned} \Delta_{k+1} &\leq \Delta_k + \nabla F(w^{(k)})^\top (w^{(k+1)} - w^{(k)}) + \frac{L}{2}\|w^{(k+1)} - w^{(k)}\|^2 \\ &= \Delta_k - \|\nabla F(w^{(k)})\|^2 \left(\eta_k - \frac{L}{2}\eta_k^2\right). \end{aligned}$$

If the gradient is non-zero, the method produces a stricts decrease in the objective value if $\eta_k < 2/L$. Moreover, if $\eta_k \leq 1/L$, we have that $(\eta_k - \frac{L}{2}\eta_k^2) \geq \frac{\eta_k}{2}$. Note that the condition on C ensures this for all k. And so, we get

$$\|\nabla F(w^{(k)})\|^2 \leq \frac{2}{\eta_k} \left(\Delta_k - \Delta_{k+1}\right).$$

Summing up, and noting that $0 \leq \Delta_k \leq \Delta_0$, we get

$$\sum_{j=0}^{k} \|\nabla F(w^{(j)})\|^2 \leq 2 \left(\frac{\Delta_0}{\eta_0} + \sum_{j=1}^{k} \Delta_j \left(\frac{1}{\eta_j} - \frac{1}{\eta_{j-1}} \right) - \frac{\Delta_{k+1}}{\eta_k} \right) \leq \frac{2}{\eta_k} \Delta_0.$$

Plugging in the rule to set η_k,

$$\sum_{j=0}^{k} \|\nabla f(w^{(j)})\|^2 \leq \frac{4}{C^2} \Delta_0^2.$$

Now divide by $k+1$ and note that the minimum is no larger than the average to complete the proof. □

References

[Ghadimi and Lan(2013)] Saeed Ghadimi and Guanghui Lan. Stochastic first- and zeroth-order methods for nonconvex stochastic programming. *SIAM Journal on Optimization*, 23(4):2341–2368, 2013.