CS2281 **Supplementary Notes**

Optimization 1: Gradient Descent

Instructor: Sham Kakade

1 Gradient Descent and Stochastic Gradient Descent

Suppose we want to solve:

$$\min_w G(w)$$

In many machine learning problems, we have that $G(w)$ is of the form:

$$G(w) = \frac{1}{n}\sum_i \ell((x_i, y_i), w)$$

Gradient descent: Gradient descent (GD) is one of the simplest of algorithms:

$$w_{t+1} = w_t - \eta_t \nabla G(w_t)$$

Note that if we are at a 0 gradient point, then we do not move. For this reason, gradient descent tends to be somewhat robust in practice.

Stochastic gradient descent: One practically difficult is that computing the gradient itself can be costly, particularly when n is large. An alternative algorithm is *stochastic gradient descent* (SGD).

This algorithms is as follows.

1. Sample a point i at random
2. Update the parameter:
$$w_{t+1} = w_t - \eta_t \nabla \ell((x_i, y_i), w_t)$$
and return to step 1.

Note that, in expectation, we are moving in the direction of the gradient. Typically, with SGD, we have to take a little care with the rate at which we decrease the learning rate to ensure convergence of the algorithm. If we decrease the learning rate too quickly, we may not converge. If we decrease it too slowly, then we may be slowing down convergence.

2 Setting the learning rate

Two things to keep in mind:

1. In practice, things like dimensional analysis give us good heuristics to set the learning rate. In particular, consider how the learning rate should scale if you change the problem parameters.

2. For non-convex problems, often setting the learning rate based on the insights from the convex case work well.

3. Also, for SGD (for both the convex and non-convex case), if we set the learning too large, the parameters will diverge. A natural starting choice is some factor (say 10) smaller than when things start to diverge.

3 Convergence rates of GD

Throughout suppose that G is convex and that:

$$w_* \in \arg\min_w G(w)$$

. We now look at standard analyses of GD.

3.1 The Smooth Case

Suppose G is upper bounded by a quadratic function. In particular, assume that:

A function G is L smooth if

$$G(w') \leq G(w) + \nabla G(w) \cdot (w' - w) + \frac{L}{2}\|w - w'\|^2$$

Under this assumption, we consider the update rule:

$$w_{t+1} = w_t - \eta \nabla G(w_t)$$

for a constant learning rate.

How should we choose η? Since G is L-smooth, a natural idea is to simply choose η to minimize the upper bound. This leads to a setting of $\eta = \frac{1}{L}$. Let us analyze the algorithm under this rate.

The following theorem shows gradient descent converges at rate of $1/t$.

Theorem 3.1. *Suppose G is L-smooth and that $\eta = \frac{1}{L}$, then*

$$G(w_t) - G(w_*) \leq \frac{L\|w_0 - w_*\|^2}{t}$$

Proof. By smoothness:

$$\begin{aligned} G(w_{t+1}) - G(w_t) &\leq G(w_t - \eta \nabla G(w_t)) - G(w_t) \\ &\leq -\eta \|\nabla G(w_t)\|^2 - \frac{L\eta^2}{2}\|\nabla G(w_t)\|^2 \\ &= -\frac{1}{2L}\|\nabla G(w_t)\|^2 \end{aligned}$$

Define:

$$\Delta_t = G(w_t) - G(w_*)$$

Our goal is to show that $\Delta_t < \frac{L\|w_0 - w_*\|^2}{t}$.

We have shown that:

$$\Delta_{t+1} \leq \Delta_t - \frac{1}{2L}\|\nabla G(w_t)\|^2$$

Also, by convexity, we have that:

$$\Delta_t \leq \nabla G(w_t) \cdot (w_t - w_*) \leq \|\nabla G(w_t)\| \|w_t - w_*\| .$$

Hence,

$$\Delta_{t+1} \leq \Delta_t - \frac{1}{2L\|w_t - w_*\|^2}\Delta_t^2$$

Also, it is not difficult to show that $\|w_t - w_*\|^2$ decreases with t. Thus,

$$\Delta_{t+1} \leq \Delta_t - \frac{1}{2L\|w_1 - w_*\|^2}\Delta_t^2$$

where we have w_1 in the above expression. The above implies the claimed rate (which one can show through induction). □

3.2 The Smooth and Strongly Convex Case

The most standard analysis of gradient descent is for a function G which is both upper and lower bounded by quadratic functions.

A function G is μ strongly convex if

$$G(w') \geq G(w) + \nabla G(w) \cdot (w' - w) + \frac{\mu}{2}\|w - w'\|^2$$

Note that every convex function is 0-strongly convex.

Let us again consider the gradient descent algorithm:

$$w_{t+1} = w_t - \eta \nabla G(w_t)$$

for a constant learning rate.

The following theorem shows gradient descent converges very rapidly if G is both strongly convex and smooth.

Theorem 3.2. *Suppose G is μ-strongly convex and L-smooth. If $\eta = \frac{1}{L}$, then*

$$\|w_t - w_*\| \leq \left(1 - \frac{\mu}{L}\right)^t \|w_0 - w_*\|$$

In particular, in $\frac{L}{\mu}\log(\|w_0 - w_*\|/\epsilon)$ iterations our distance to the optimal point is $O(\epsilon)$.

First, the following lemma is helpful:

Lemma 3.3. *(A Gradient Bound under Smoothness) First, let us show that:*

$$\|\nabla G(w)\|^2 \leq 2L(G(w) - G(w_*))$$

Proof. To see this, observe that, by smoothness and the optimality of w_*, we have

$$G(w_*) \leq G\left(w - \frac{1}{L}\nabla G(w)\right) \leq G(w) - \frac{1}{L}\|\nabla G(w)\|^2 + \frac{1}{2L}\|\nabla G(w)\|^2 \leq G(w) - \frac{1}{2L}\|\nabla G(w)\|^2$$

which proves the claim. □

Now we are ready to prove our theorem.

Proof. Note that by strong convexity we have:

$$\nabla G(w) \cdot (w - w_*) \geq G(w) - G(w_*) + \frac{\mu}{2}\|w - w_*\|^2$$

Using these, we have that:

$$\begin{aligned}
\|w_{t+1} - w_*\|^2 &= \|w_t - \eta\nabla G(w_t) - w_*\|^2 \\
&= \|w_t - w_*\|^2 - 2\eta\nabla G(w_t) \cdot (w_t - w_*) + \eta^2\|\nabla G(w_t)\|^2 \\
&\leq \|w_t - w_*\|^2 - 2\eta\left(G(w) - G(w_*) + \frac{\mu}{2}\|w_t - w_*\|^2\right) + \eta^2\|\nabla G(w_t)\|^2 \\
&\leq \|w_t - w_*\|^2 - 2\eta(G(w) - G(w_*)) - \eta\mu\|w_t - w_*\|^2 + 2\eta^2 L(G(w) - G(w_*)) \\
&\leq \|w_t - w_*\|^2 - \eta\mu\|w_t - w_*\|^2 + 2\eta(\eta L - 1)(G(w) - G(w_*)) \\
&\leq (1 - \frac{\mu}{L})\|w_t - w_*\|^2
\end{aligned}$$

where we used the setting of η in the last step. □

3.3 Non-smooth optimization and (sub-)gradient descent

The the sub-gradient update rule is again:

$$w_{t+1} = w_t - \eta\nabla G(w_t)$$

where $\nabla G(w_t)$ is the *sub-gradient* at w_t.

We say that $\nabla G(w)$ is a *sub-gradient* at w if it satisfies, for all w', that:

$$G(w') \geq G(w) + \nabla G(w) \cdot (w' - w)$$

For non-differentiable convex functions, the sub-gradient is a natural concept to work with.

Theorem 3.4. *(The non-smooth case) Suppose that for all w we have that:*

$$\|\nabla G(w)\| \leq B$$

Also, suppose that we know a bound on our starting distance, i.e. $\|w_0 - w_*\| \leq R$. Set $\eta = \frac{R}{B}\sqrt{\frac{2}{T}}$, then we have that:

$$G\left(\frac{1}{T}\sum_t w_t\right) - G(w_*) \leq \frac{RB}{\sqrt{T}}$$

Proof. First, note that the We have that:

$$\begin{aligned}
\|w_{t+1} - w_*\|^2 &= \|w_t - \nabla G(w_t) - w_*\|^2 \\
&= \|w_t - w_*\|^2 - 2\eta\nabla G(w_t) \cdot (w_t - w_*) + \eta^2\|\nabla G(w_t)\|^2 \\
&\leq \|w_t - w_*\|^2 - 2\eta\nabla G(w_t) \cdot (w_t - w_*) + \eta^2 B^2
\end{aligned}$$

using the definition of B.

Hence,

$$\nabla G(w_t) \cdot (w_t - w_*) = \frac{1}{2\eta}\|w_t - w_*\|^2 - \|w_{t+1} - w_*\|^2 + \frac{\eta}{2}B^2$$

and so:

$$\begin{aligned}
\frac{1}{T}\sum_{t=1}^{T} \nabla G(w_t) \cdot (w_t - w_*) &= \frac{1}{2\eta}\left(\|w_1 - w_*\|^2 - \|w_{T+1} - w_*\|^2\right) + \frac{\eta T}{2}B^2 \\
&\leq \frac{\|w_1 - w_*\|^2}{2\eta} + \frac{\eta T}{2}B^2 \\
&\leq \frac{RB}{\sqrt{T}}
\end{aligned}$$

where the last step uses our choice of η.

The proof is completed since:

$$G\left(\frac{1}{T}\sum_t w_t\right) \leq \frac{1}{T}\sum_t G(w_t) \leq \frac{1}{T}\sum_{t=1}^{T} \nabla G(w_t) \cdot (w_t - w_*)$$

where both steps follow from convexity. □