CSE 546: Machine Learning **Lecture 7**

Gradient Descent and Stochastic 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 (or there will be some oscillatory or weird behavior). A natural starting choice is some factor (say 2) smaller than when things start to diverge.

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 3.1. *(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:

$$G\left(\overline{w}_T\right) - G(w_*) \leq \frac{RB}{\sqrt{T}} \text{ where } \overline{w}_T = \frac{1}{T}\sum_t w_t$$

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. □

4 Stochastic Gradient Descent

Suppose we want to minimize $G(w)$, where $G(w)$ is of the form:

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

We could use gradient descent. One practical difficulty 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.

Theorem 4.1. *(SGD) Suppose that for all (x, y) and w we have that:*

$$\|\nabla \ell((x, y), w)\| \leq B$$

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:

$$\mathbb{E}[G(\overline{w}_T)] - G(w_*) \leq \frac{RB}{\sqrt{T}} \text{ where } \overline{w}_T = \frac{1}{T}\sum_t w_t$$

where the expectation is over the random points (x_i, y_i) drawn in our algorithm.

Proof. Suppose that (x_i, y_i) are drawn at timestep t. Let us define sampled loss function at time t to be:

$$\ell_t(w) = \ell((x_i, y_i), w)$$

where

Just as in the non-smooth case, we have that:

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

using the definition of B.

Due to the random sampling at time t (which is uncorrelated with the history of samples before time t), we have:

$$E[\nabla \ell_t(w_t) | \text{history before } t] = \nabla G(w_t)$$

By taking an expectation with respect to sample at time t, we have:

$$\mathbb{E}[\|w_{t+1} - w_*\|^2 | \text{history before } t] \leq \|w_t - w_*\|^2 - 2\eta \nabla G(w_t) \cdot (w_t - w_*) + \eta^2 B^2$$

(here we condition on the history up to time t).

By taking unconditional expectations,

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

and so:

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

where the last step uses our choice of η.

The proof is completed using convexity. □