CSE 546: Machine Learning **Lecture 11**

SGD and Generalization

Instructor: Sham Kakade

1 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 1.1. *(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. □

2 Online Learning and Generalization

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

$$L(w) = \mathbb{E}\ell((x, y), w)$$

given only samples of (x, y) pairs, which are sampled according to some distribution D.

Let us consider *stochastic gradient descent* (SGD) where we only touch each data point *once*.

This algorithms is as follows.

1. Sample a point (x, y) from the distribution D.
2. Update the parameter:
$$w_{t+1} = w_t - \eta_t\nabla\ell((x, y), w_t)$$
and return to step 1.

Note the distinction here that we use a fresh samples for each update (and we never reuse a sample).

Corollary 2.1. *(SGD) Suppose that for all (x, y) 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:

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

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

Note the above provides a *generalization* bound, in that it bounds the risk on the *true* loss.

2.1 Example: linear regression and generalization

Suppose we want to minimize the square loss:

$$L(w) = \mathbb{E}(y - w^\top x)^2$$

given only samples of (x, y) pairs, which are sampled according to some distribution D.

Suppose we have the following upper bound that $\|x\| \leq X_+$. Also, suppose that $\|w_*\| \leq W_+$. Also assume that $y \leq X_+ W_+$ (note that this is a minor assumption since $w^\top x \leq W_+ X_+$, so it must be the case that y is upper bounded by this).

This algorithms is as follows.

1. Sample a point (x, y) from the distribution D.
2. Update the parameter:
$$w_{t+1} = \text{Proj}_{W_+}(w_t + \eta_t(y - w^\top x)x)$$
and return to step 1.

Note the distinction here that we use a fresh samples for each update (and we never reuse a sample).

Corollary 2.2. *(SGD) Suppose the previous assumptions are satisfied. We have that:*

$$\|\nabla \ell((x, y), w)\| \leq W_+ X_+^2 = B$$

Also, suppose w_0 so that $\|w_0 - w_*\| \leq W_+ = R$. Set $\eta = \frac{1}{X_+^2}\sqrt{\frac{2}{T}}$, then we have that:

$$\mathbb{E}[L(\overline{w}_T)] - L(w_*) \leq \frac{(W_+ X_+)^2}{\sqrt{T}} \text{ where } \overline{w}_T = \frac{1}{T}\sum_t w_t$$

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

Note the above provides a *generalization* bound, in that it bounds the risk on the *true* loss.