MATH 462 LECTURE NOTES

ADAM M. OBERMAN

1. Gradient descent for empirical losses

Starting from the definition of the empirical loss.

$$\widehat{L}(w) = \frac{1}{m}\sum_{i=1}^{m} \ell(h_w(X_i), y_i) \tag{EL}$$

Then define the gradient vector using the following notation

$$g(w) = \nabla_w \widehat{L}(w) = \frac{1}{m}\sum_{i=1}^{m} \partial_h \ell(h_w(x_i), w)\nabla_w h_w(x_i) \tag{G}$$

Then the gradient descent algorithm is given by the following.

Definition 1.1. Gradient descent with learning rate $\alpha > 0$ for (EL) is given by

$$w_{t+1} = w_t - \alpha g(w_t), \tag{GD}$$

where $g(w_t) = \nabla_w \widehat{L}(w_t)$ given by (G)

Remark 1.2. Note, by the chain rule,

$$g(w) = \nabla_w \widehat{L}(w) = \frac{1}{m}\sum_{i=1}^{m} \partial_h \ell(h_w(x_i), w)\nabla_w h_w(x_i)$$

So the gradient of the loss is the weighted sum of the model gradients, weighted by the loss derivative.

Example 1.3. Consider a very simple quadratic problem for (EL),

$$\widehat{L}(w) = \frac{1}{m}\sum_{i=1}^{m} \frac{(w - y_i)^2}{2} \tag{1}$$

Then it is an exercise to show the following:

(1) The minimizer w^* of (EL) is given by $\bar{y} = \frac{1}{m}\sum_i y_i$.
(2) $\nabla \widehat{L}(w) = \widehat{L}'(w) = w - \bar{y}$.
(3) Applying gradient descent (GD) leads to the recursion $w_{t+1} - w^* = (1-\alpha)(w_t - w^*)$

In particular, conclude that

$$(w_t - w^*) = (1-\alpha)^k (w_0 - w^*)$$

In particular,

(1) Gradient descent (GD) with $\alpha = 1$ converges in one step.
(2) Gradient descent (GD) with $\alpha = 1/2$ satisfies $|w_t - \bar{y}| \le (1/2)^t |w_0 - \bar{y}|$

Date: November 2, 2022.

Remark 1.4. In the example above, the general behaviour of GD for strongly convex losses is apparent. In general, the error converges at a geometric rate: $e_k \leq \gamma^t e_0$. When γ is bounded away from 1, the rate is fairly good, and the problems are called well-conditioned. However, even with the optimal learning rate, there are ill-conditioned problems where γ can be arbitrarily close to 1, which is very slow convergence. (In the convex case, there can be an even slower convergence rate, of $1/t$.

Exercise 1.1. Define the hypothesis class, dataset, and which reduces (EL) to (1). (Hint: take the quadratic loss, the constant model, $h_w = w$, $w \in \mathbb{R}$, and $S_m = \{y_1, \ldots, y_m\}$.)

2. Stochastic Gradient Descent

Now for stochastic gradient descent, we use an approximation of the gradient.

Definition 2.1. Choose a minibatch size N and let I be consist of N indices randomly chosen from $1, \ldots, m$. Define the minibatch gradient

$$\widehat{g}(w) = \widehat{g}(w, I) = \frac{1}{N} \sum_{i \in I} \nabla_w \ell(h_w(X_i), y_i) \tag{2}$$

Note we can write

$$\widehat{g}(w) = g(w) + \widehat{e}$$

where $\widehat{e}$ is a mean zero error term.

Exercise 2.1. Prove that if I is a random minibatch, then the expected value (over the choice of the minibatch) of the error vector $\widehat{e}$ is zero.

Definition 2.2. Stochastic gradient descent with learning rate α_t for (EL) is given by

$$w_{t+1} = w_t - \alpha \widehat{g}(w_t),$$

where $g(w_t)$ is a stochastic gradient, given by (2).

2.1. Classical and minibatch SGD.

The typical minibatch gradients are

1. Classical SGD, is where we use $N = 1$, one data point at a time.
2. Deep learning minibatches: partition the dataset randomly into equal fractions of size N, then go through the data exact once, an epoch, before repeating.

The typical learning rates are

1. Constant learning rate $\alpha_t = \alpha$. Then the error saturates after a while.
2. One over t learning rate $\alpha_t = \frac{\alpha_0}{t + t_0}$
3. Deep Learning stepwise decreasing learning rate. Here $\alpha = \alpha_0$, for $T = m$ iterations, or one batch. Then we decrease α by constant proportion, each T iterations.

2.2. TODO: Example again.

Reconsider the simple quadratic problem (1), this time with a stochastic gradient.

$$\widehat{L}(w) = \frac{1}{m} \sum_{i=1}^{m} (w - y_i)^2$$

Now, with a stochastic gradient,

$$\widehat{g}(w) = \frac{1}{N} \sum (w - y_i) = w - \widehat{y}$$

we get a stochastic approximation to the average.

Write $\widehat{y} = \bar{y} + e_t$, where the error, e_t is random, and mean zero. Then we have

$$w_{t+1} - w^* = (1 - \alpha_t)(w_t - w^*) + \alpha_t e_t$$

so we get an error term, which can accumulate. By carefully choosing a decreasing learning rate, we balance out the errors.

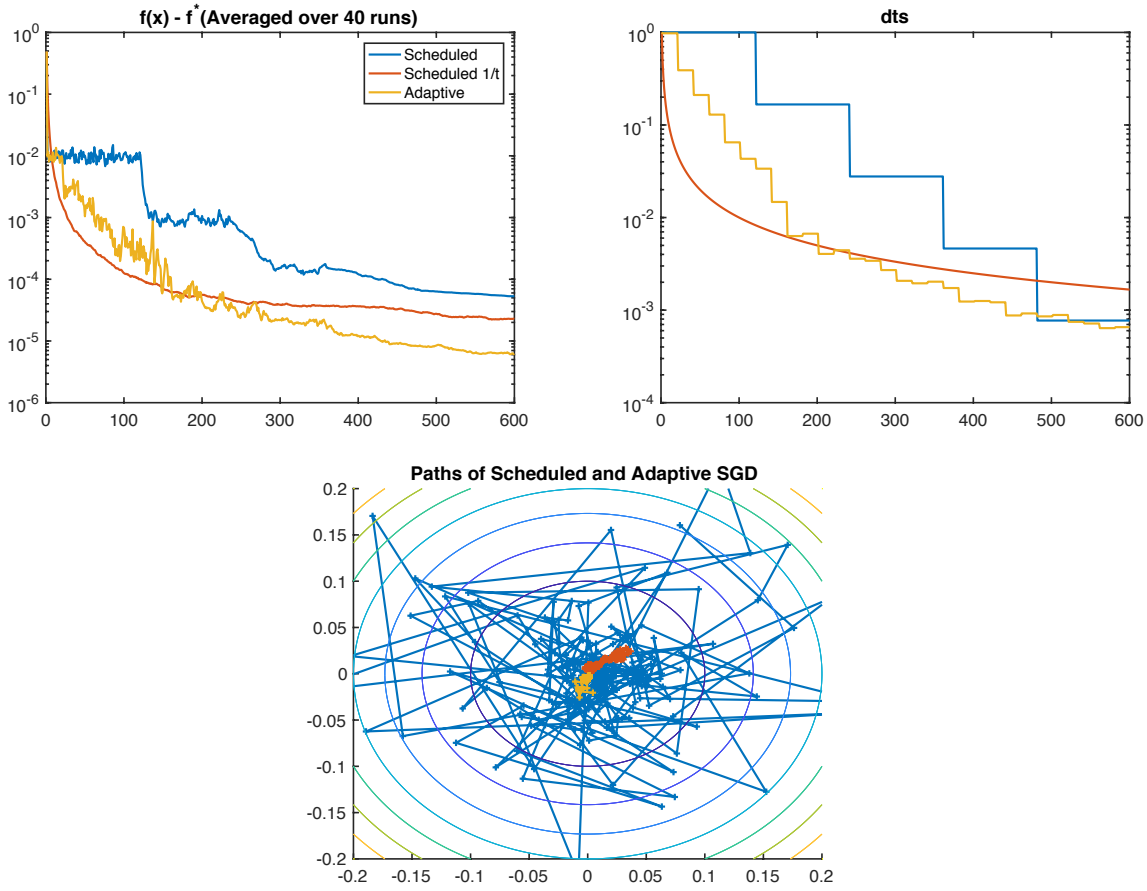

FIGURE 1. Illustration of SGD with a two dimensional vector. Blue curves: piecewise constant learning rate. The loss saturates quickly, then drops when the learning rate decreases. Even when the loss is nearly constant, the vector is changing, but bounces around the level sets of the loss. Red curves: a tuned $1/t$ learning rate. When the two constants in the learning rate are tuned, the loss decreases steadily, at a $1/t$ rate. Yellow curves: a method which tries to adapt the learning rate (ignore this).

3. Analysis

3.1. **Proof of descent for Gradient Descent.** Gradient descent is called a descent algorithm, because the loss decreases at every iteration, provided the learning rate is small enough. In this section, we give the idea of the proof, in the one dimensional case. The more general case is addressed in convex optimization textbooks (e.g. boyd).

To use this result, we use a special for of Taylor approximation, with an exact remainder.

Theorem 3.1 (Taylor with remainder)**.** *Let $f : \mathbb{R} \to \mathbb{R}$ be twice differentiatiable, with f'' continuous. Then*

$$f(x+h) = f(x) + hf'(x) + \frac{h^2}{2} f''(\eta) \tag{3}$$

for some η between x and $x+h$.

Theorem 3.2 (Gradient descent decreases the loss)**.** *Suppose $\widehat{L}$ is twice differentiable with and f'' is continuous, and bounded,*

$$|\widehat{L}''(w)| \leq L, \quad \text{for all } w$$

Then if we apply (GD) *with*

$$0 < \alpha \leq \frac{2}{L} \tag{4}$$

the loss does not increase,

$$\widehat{L}(w_{t+1}) \leq \widehat{L}(w_t)$$

moreover, the loss decreases, unless w_t is a critical point.

Proof. Apply (3) to $\widehat{L}$

$$\widehat{L}(w_t + h) - \widehat{L}(w_t) = h\widehat{L}'(w_t) + \frac{h^2}{2}\widehat{L}''(\eta)$$

for some η (which depends on h).

Apply the equation above with $h = -\alpha\widehat{L}'(w_t)$ to obtain

$$\widehat{L}(w_t + h) - \widehat{L}(w_t) = -\alpha(\widehat{L}'(w_t))^2 \left(1 - \alpha\frac{\widehat{L}''(\eta)}{2}\right)$$

By the assumption (4)

$$\left(1 - \alpha\frac{\widehat{L}''(\eta)}{2}\right) \geq 0$$

so we have

$$\widehat{L}(w_t + h) \leq \widehat{L}(w_t)$$

which is the desired result. □

3.2. **Expected descent for SGD.** For SGD, a given step may increase the loss. However, on average, assuming the stepsize is small enough, a given step is non-decreasing in the loss.