16.1 Stochastic gradient descent

Consider minimizing the average of a bunch of functions:

$$min_x \frac{1}{n} \sum_{i=1}^{n} f_i(x)$$

Recall that gradient descent goes as follows:

$$x^{(k)} = x^{(k-1)} - t_k . \frac{1}{n} \sum_{i=1}^{n} \nabla f_i(x^{k-1})$$

The stochastic gradient descent replaces the averaged gradient with the gradient of a randomly chosen function:

$$x^{(k)} = x^{(k-1)} - t_k . \nabla f_{ik}(x^{k-1})$$

where ik is randomly drawn from 1, ..., n. Note that:

$$E[\nabla f_{ik}(x^{(k-1)})] = \frac{1}{n} \sum_{i=1}^{n} \nabla f_i(x^{(k-1)})$$

So the gradient used is an unbiased , albeit higher variance estimate of the gradient used for GD. To reduce the variance, mini-batch stochastic gradient descent can be used, in which we sample $I_k \in [n]$ with —I_k| = $bfunctionsandusetheaverageoftheirgradients$: $E[\nabla f_{ik}(x^{(k-1)})] = \frac{1}{n} \sum_{i \in I_k} b \nabla f_i(x^{(k-1)})$Note that this is again an unbiased estimate of the gradient, b times more expensive than SGD but having 1/bth the variance.

16.2 SGD for Risk Minimization

Recall that in learning we face the problem of minimizing the risk function

$$L_D(w) = E_{z \sim D}[l(w, z)].$$

We have seen the method of empirical risk minimization, where we minimize the empirical risk, $L_S(w)$, as an estimate to minimizing $L_D(w)$. SGD allows us to take a different approach and minimize $L_D(w)$ directly. Since we do not know D, we cannot simply calculate $\nabla L_D(w(t))$ and minimize it with the GD method. With SGD, however, all we need is to find an unbiased estimate of the gradient of $L_D(w)$, that is, a random vector whose conditional

expected value is $\nabla L_D(w(t))$. We shall now see how such an estimate can be easily constructed. For simplicity, let us first consider the case of differentiable loss functions. Hence the risk function L_D is also differentiable. The construction of the random vector v_t will be as follows: First, sample zD. Then, define v_t to be the gradient of the function l(w,z) with respect to w, at the point $w^{(t)}$. Then, by the linearity of the gradient we have

$$E[v_t|w^{(t)}] = E[\nabla l(w^{(t)}, z)] = \nabla E[l(w^{(t)}, z)] = \nabla L_D(w^{(t)}).$$

The gradient of the loss function l(w, z) at $w^{(t)}$ is therefore an unbiased estimate of the gradient of the risk function $L_D(w^{(t)})$ and is easily constructed by sampling a single fresh example $z \sim D$ at each iteration t. The same argument holds for nondifferentiable loss functions. We simply let v_t be a subgradient of l(w,z) at $w^{(t)}$. Then, for every u we have

$$l(u, z) - l(w^{(t)}, z) \geq u - w^{(t)}, v_t.$$

Taking expectation on both sides with respect to $z \sim D$and conditioned on the value of $w^{(t)}$ we obtain

$$\begin{aligned} L_D(u) - L_D(w^{(t)}) &= E[l(u, z) - l(w^{(t)}, z)|w^{(t)}] \\ &\geq E[u - w^{(t)}, v_t|w^{(t)})] \\ &= u - w^{(t)}, E[v_t|w^{(t)}]. \end{aligned}$$

It follows that $E[v_t|w^{(t)}]$ is a subgradient of $L_D(w)$ at $w^{(t)}$.

16.3 Strongly Convex Functions

In this section we show a variant of SGD that enjoys a faster convergence rate for problems in which the objective function is strongly convex If f is -strongly convex then for every w,u and v f(w) we have, $w - u, v \geq f(w) - f(u) + \lambda||w - u||^2$ Let $\nabla(t) = E[v_t|w^{(t)}]$. Since f is strongly convex and $\nabla(t)$ is in the subgradient set of f at $w^{(t)}$ we have that

$$w^{(t)} - w^*, \nabla^t \geq f(w^{(t)}) - f(w^*) + \lambda||w^{(t)} - w^*||^2$$

Next, we show that

$$w^{(t)} - w^*, \nabla^{(t)} \leq \frac{E[||w^{(t)} - w^*|^2 - ||w^{(t+1)} - w^*||^2}{2\eta_t} + \frac{\eta_t \rho^2}{2}$$

Since $w^{(t+1)}$ is the projection of $w^{(t+1)}$ onto H, and $w^* \in H$. We have that

$$||w^{(t+1)} - w^*||^2 \geq ||w^{(t+1)} - w^*||^2$$

Therefore,

$$\begin{aligned} ||w^{(t)} - w^*||^2 - ||w^{(t+1)} - w^*||^2 &\geq ||w^{(t)} - w^*||^2 - ||w^{(t+1/2)} - w^*||^2 \\ &= 2\eta_t w^{(t)} w^*, v_t - \eta_t^2 ||v_t||^2 \end{aligned}$$

Taking expectation of both sides, rearranging, and using the assumption $E[||v_t||_2] \leq \rho^2$

$$\sum_{1}^{T} E[f(w^{(t)})] - f(w^*))$$

$$\leq E[\sum_{i=1}^{T} \frac{||w^{(t)} - w^*||^2 - ||w^{(t+1)} - w^*||^2}{2\eta_t} - \frac{\lambda}{2}||w^{(t)} - w^*||^2] + \frac{\rho^2}{2}\sum_{i=1}^{T}\eta_t$$

Next, we use the definition $\eta_t = 1/(\lambda_t)$ and note that the first sum on the right-hand side of the equation collapses to - $\lambda T||w^{(T+1)} - w^*||^2 \leq 0$. Thus,

$$\sum_{1}^{T} E[f(w^{(t)})] - f(w^*)) \leq \frac{\rho^2}{2\lambda}\sum_{i=1}^{T}\frac{1}{t} \leq \frac{\rho^2}{2\lambda}(1 + log(T))$$

16.4 Comparison with Gradient Descent

1. Stochastic Gradient Descent
   (a) per iteration cos tO(d)
   (b) number of iterations$O(1/\epsilon)$
   (c) total cost $O(d/\epsilon)$
2. Gradient Descent
   (a) per iteration cost O(nd)
   (b) number of iterations $O(1/\epsilon)$
   (c) total cost $O(nd/\epsilon)$

SGD can be faster for largenand low accuracy