CS 189 Introduction to Machine Learning
Fall 2018 DIS7

1 Understanding first order methods

In lecture we have encountered two first order methods so far: The gradient descent method and the stochastic gradient method. Both of them only use first order information (the gradient), hence the name. In this question we want to further our understanding as to why they work. Let $f : \mathbb{R}^d \to \mathbb{R}$ be a convex function that is minimized uniquely at $\mathbf{w}^\star$.

Remember the update for gradient descent on an arbitrary function $f(w)$

$$\mathbf{w}^{(k)} = \mathbf{w}^{(k-1)} - \alpha_k \nabla f(\mathbf{w}^{(k-1)}) \tag{1}$$

for some stepsize $\alpha > 0$ (in general the stepsize might vary in each iteration as seen in the homeworks).

The corresponding stochastic gradient method (with constant stepsize) does not take a step in the direction of $\nabla f(\mathbf{w}^{(k-1)})$ but a random direction $G(\mathbf{w}^{(k-1)}, \xi_{k-1})$, where ξ_{k-1} is a random variable which encodes the randomness of the gradient and is independent from $\mathbf{w}^{(k-1)}$. G is a random vector-valued function which is in expectation equal to the gradient ∇f, i.e. $\mathbb{E}_\xi[G(\mathbf{w}, \xi)] = \nabla f(\mathbf{w})$ for all $\mathbf{w}$ where the expectation is over the stochasticity of the gradient. Therefore, we also have for the random variables $\mathbf{w}^{(k-1)}$ that $\mathbb{E}_{\mathbf{w}^{(k-1)}}\mathbb{E}_{\xi_{k-1}}[G(\mathbf{w}^{(k-1)}, \xi_{k-1})] = \mathbb{E}_{\mathbf{w}^{(k-1)}}\nabla f(\mathbf{w}^{(k-1)})$.

Then, the update reads

$$\mathbf{w}^{(k)} = \mathbf{w}^{(k-1)} - \alpha_k G(\mathbf{w}^{(k-1)}, \xi_{k-1}), \tag{2}$$

where ξ_k are i.i.d. random variables.

In this problem, we want to focus on the convergence for (small enough) *constant stepsize* of this method, so $\alpha_k = \alpha$ for all k.

(a) In many practical cases in machine learning applications, the loss f that we want to minimize can be decomposed into a sum of functions, that is $f(\mathbf{w}) = \frac{1}{n}\sum_{i=1}^n f_i(\mathbf{w})$. This holds in particular for problems involving an average over the training data, which is for example the case when we want to find the maximum likelihood estimator given i.i.d. data in a generative probabilistic model.

In this case, we can define a random direction by just drawing an index ξ uniformly at random from $\{1, \ldots, n\}$ and setting $G(\mathbf{w}, \xi) = \nabla f_\xi(\mathbf{w})$. Show that this choice of stochastic gradient is unbiased, i.e. $\mathbb{E}_\xi[G(\mathbf{w}, \xi)] = \nabla f(\mathbf{w})$, where the expectation is over the random draw.

(b) Let us denote the random index at iteration k by ξ_k. Given SGD updates $\mathbf{w}^{(k)} = \mathbf{w}^{(k-1)} - \alpha_k \nabla f_{\xi_k}(\mathbf{w}^{(k-1)})$, one key step in any convergence proof (including the one in homework) is to

use the so-called law of iterated expectation (or *tower property*)

$$\mathbb{E}[\langle G(\mathbf{w}^{(k)}, \xi_k), \mathbf{w}^{(k)} - \mathbf{w}^\star \rangle] = \mathbb{E}_{\xi_1,\ldots,\xi_{k-1}}[\mathbb{E}_{\xi_k}[\langle G(\mathbf{w}^{(k)}, \xi_k), \mathbf{w}^{(k)} - \mathbf{w}^\star \rangle | \xi_1, \ldots, \xi_{k-1}]]. \quad (3)$$

In order to see that this is true, let us consider the following abstract general setting: Given a joint discrete probability distribution P on a finite number of vectors $\{\boldsymbol{\ell}_1, \ldots, \boldsymbol{\ell}_n\}$ and scalars $\{k_1, \ldots, k_m\}$. Now let $\boldsymbol{\ell}$, K be r.v. drawn from the distribution. Recall the conditional probability $P(\boldsymbol{\ell}|K = k)$ and for fixed $k \in \{k_1, \ldots, k_m\}$ the definition of $\mathbb{E}(\boldsymbol{\ell}|k) := \sum_{i=1}^n \boldsymbol{\ell}_i P(\boldsymbol{\ell} = \boldsymbol{\ell}_i | K = k)$. When k is now random too, show that by taking another expectation over K we obtain

$$\mathbb{E}_K[\mathbb{E}_{\boldsymbol{\ell}}(\boldsymbol{\ell}|K)] = \mathbb{E}\boldsymbol{\ell}$$

using Bayes rule. How can we use this for proving (3)?

(c) In this question we want to see how the noise of the stochastic gradient $\boldsymbol{\epsilon} = G(\mathbf{w}, \xi) - \nabla f(\mathbf{w})$ affects the convergence of SGD compared to GD with noiseless gradients.

Here, we assume $\boldsymbol{\epsilon}$ are independent and zero-mean. As you will see in the homework, for "nice" functions, the convergence of SGD for *constant stepsize* obeys

$$\mathbb{E}\|\mathbf{w}^{(k)} - \mathbf{w}^\star\|^2 \leq \beta^k \mathbb{E}[\|\mathbf{w}^{(0)} - \mathbf{w}^\star\|_2^2] + c\alpha B^2. \quad (4)$$

with $B^2 = \mathbb{E}\|\boldsymbol{\epsilon}\|^2$ and the constants $\beta < 1$ and c depend on the loss function. Recall the gradient descent convergence rate you derived in HW 5, assuming a suitably chosen constant stepsize:

$$\|\mathbf{w}^{(k)} - \mathbf{w}^\star\|_2^2 \leq \beta^k \|\mathbf{w}^{(0)} - \mathbf{w}^\star\|_2^2 \quad (5)$$

where β depends on the minimum and maximum eigenvalues of the Hessian.

Describe what the repercussions are of using noisy gradients by comparing the two bounds (4) and (5). Where does SGD converge to for constant stepsize? Does the squared error of the estimator $\|\mathbf{w}^{(k)} - \mathbf{w}^\star\|^2$ strictly decrease with each iteration?

2 Computational complexity of SGD vs. GD

Let us consider a simple least squares problem, where $\mathbf{A} \in \mathbb{R}^{n \times d}, \mathbf{w}, \mathbf{y} \in \mathbb{R}^d$ and we are interested in optimizing the function

$$F(\mathbf{w}) = \frac{1}{2n}\|\mathbf{A}\mathbf{w} - \mathbf{y}\|_2^2 = \frac{1}{n}\sum_{i=1}^n \frac{1}{2}(\mathbf{a}_i^\top \mathbf{w} - y_i)^2.$$

(a) Write down the gradient descent update. From HW 5 Problem 3 we know that:

$$d_k \leq \left(\frac{\kappa - 1}{\kappa + 1}\right)^{2k} d_0$$

where, $d_k = \|\mathbf{w}^{(k)} - \mathbf{w}^\star\|_2^2$ and $\kappa = \frac{\lambda_{max}(\mathbf{A}^\top \mathbf{A})}{\lambda_{min}(\mathbf{A}^\top \mathbf{A})}$ denotes the condition number of $\mathbf{A}^\top \mathbf{A}$. Show that $dn\kappa \log\big(1/\epsilon\big)$ is the time complexity of computing an ϵ optimal solution.

(b) Show that the regularization term in ridge regression improves the time complexity of computing an ϵ optimal solution?

(c) Write down the stochastic gradient descent update. In future HW you will show that the following bound holds for the SGD iterate $\mathbf{w}^{(k)}$ at time step k updated using a constant stepsize

$$\Delta_k \leq (1 - 2\alpha m)^k \Delta_0 + \alpha \frac{B^2}{m}$$

where, $\Delta_k = \mathbb{E}\left[\|\mathbf{w}^{(k)} - \mathbf{w}^\star\|_2^2\right]$.

Show that $\frac{d}{\epsilon}\log\big(1/\epsilon\big)$ is the time complexity of computing an ϵ optimal solution. How does this compare with the complexity of gradient descent in section (a)? Hint: You may choose α as a function of B and ϵ.