Homework 2 Solutions

February 2019

1 Problem 1:

This exercise makes use of Jupyter, a "computational notebook" application which allows the user to embed Python code (or any other language) in a document, which can be reset and run, with output stored in the document. The document can also be edited to contain illustrative text, using a simple markdown language. To gain familiarity with Jupyter, please take a look at this tutorial (`http://bi1.caltech.edu/code/t0b_jupyter_notebooks.html`).
Recall the SignSGD Algorithm from the class,

Algorithm 1 SignSGD Algorithm

1: **Input:** Learning rate δ, current point x_k
2: $\tilde{g}_k \leftarrow \text{stochasticGradient}(x_k)$
3: $x_{k+1} \leftarrow x_k - \delta \cdot \text{Sign}(\tilde{g}_k)$.

Please refer to [1] for more information about the algorithm and its analysis. The goal of this exercise is for you to come up with a simple example of an optimization, for which the Sign SGD diverges or fails to achieve a minimum. Please justify your example by simulating your proposed optimization using Jupyter Notebook and include the code and the plots in your submission.

Solution: Define the function $f : \mathbb{R}^D \to \mathbb{R}$ to be $f(\mathbf{x}) = \frac{1}{2}\|\mathbf{x}\|_2^2$. In this case, the stochastic gradient would be $\mathbf{x}_i + \mathbf{z}_i$. We added the noise vector $\mathbf{z}_i$ to model the *stochastic* gradient descent. Therefore in SGD, the updates are

$$-\eta_i \cdot (\mathbf{x}_i + \mathbf{z}_i) \ , \tag{1}$$

which are unbiased and thus, we expect the SGD to converge. But this is not the case in the SignSGD, since the updates are while in the SGD the updates take the form of

$$-\eta_i \cdot \text{Sign}(\mathbf{x}_i + \mathbf{z}_i) \ , \tag{2}$$

which might not be biased, depending on the noise distribution. This might cause the SignSGD to oscillate, especially when we are close to optimal.
A naive case in which SGD converges but SignSGD doesn't, is where the step sizes are fixed, and the error is large. So if we are close to the optimal point $\hat{\mathbf{x}} = 0$, we might have $\text{Sign}(\mathbf{x}_i + \mathbf{z}_i) = -sign(\mathbf{x}_i)$, in the case of a large noise. Therefore, the update occurs in the wrong direction, with a large magnitude. That's why SignSGD might oscillate near the optimal solution.

For instance, let the noise be Bernoulli$(.5) - .5$, and the data points are zero-mean Gaussian vectors with covariance $\mathbb{I}$. For a fixed step size, you will observe that the SGD will converge while the SignSGD oscillates.

In order to grade this problem, the student gets the point if the code is well written in Jupyter notebook, plots are drawn and the example is in accordance with the points made above.

2 Problem 2

Show that SGA with some step size η fails to converge to the stationary point $(0, 0)$ when run on the function $U(\theta, \omega) = \theta\omega$ from any initial point θ_0, ω_0 where both θ_0, ω_0 are non-zero.

Solution: The SGA dynamics have the form:

$$\theta_{t+1} = \theta_t - \eta \nabla_\theta U(\theta_t, \omega_t),$$

$$\omega_{t+1} = \omega_t + \eta \nabla_\omega U(\theta_t, \omega_t).$$

Let d_t be the distance from $(\omega_t, \theta_t$ to $(0, 0)$, i.e. $d_t = \sqrt{x_t^2 + y_t^2}$. Applying the dynamics, we have

$$d_{t+1} = \sqrt{x_{t+1}^2 + y_{t+1}^2} = \sqrt{(\theta_t - \eta\omega_t)^2 + (\omega_t + \eta\theta_t)^2} = \sqrt{(1 + \eta^2)(\theta_t^2 + \omega_t^2)} > d_t$$

Since the iterates are getting farther and farther away from the stationary point, they will never converge.

References

[1] Jeremy Bernstein, Yu-Xiang Wang, Kamyar Azizzadenesheli, and Anima Anandkumar. signsgd: compressed optimisation for non-convex problems. *arXiv preprint arXiv:1802.04434*, 2018.