ECE901 Homework 1

Instructions: Due March 15, 2018. Please sent via email at `dimitris@papail.io`. The homework is to be solved individually, and not in groups. If you have any questions please contact the Instructor. The answers should be submitted as a single pdf file using LaTeX typesetting.

Exercise 1. *Let us consider the following linear regression problem*

$$\min_w f(w) = \min_w \sum_{i=1}^{n} (x_i^T w - y_i)^2.$$

1. *Relate the smoothness (β) and strong convexity (λ) parameters of $f(w)$ with the maximum and minimum eigenlavues λ_1 and λ_d of the matrix XX^T where $X = [x_1, \ldots, x_n]$ is $d \times n$ data matrix. Reminder: λ-strong convexity implies that $\langle \nabla f(w), w - w^* \rangle \geq \lambda/2 \|w - w^*\|^2$.*

2. *If you were to run Gradient Descent to minimize a function as above for $\lambda_n = \frac{\lambda_1}{\sqrt{d}}$ compared to one where $\lambda_n = \frac{\lambda_1}{d}$, when would you expect GD to converge faster and why? Please argue in terms of convergence bounds.*

3. *Can you give an example of a data matrix X that yields $\lambda_1 = \lambda_n = 1$?*

Exercise 2. *Let $f_i(w)$ be a β-smooth function, for all $i = 1, \ldots, n$, and let*

$$f(w) = \frac{1}{n} \sum_{i=1}^{n} f_i(w)$$

be a λ-strongly convex function. Assume that each of the n functions $f_1(w), f_2(w), \ldots, f_n(w)$ have multiple global minima. Moreover these n functions share at least one global minimum w^* with $f(w)$. That is there exist a w^* that is the gloval minimum of all functions $f_1(w), \ldots, f_n(w)$ and their sum $f(w)$.

1. *Prove that SGD with constant stepsize $\gamma = c\frac{\lambda}{\beta^2}$, achieves the following convergence rates:*

$$E\|w_T - w^*\|^2 \leq e^{-c' \cdot (\lambda^2/\beta^2) \cdot T} E\|w_0 - w^*\|^2,$$

   for some constants $c, c' > 0$. Hint: $(1 - x) \leq e^{-x}$, for $x \geq 0$.

2. *Compare the above bound with the one you would get for general λ-strongly convex functions with bounded stochastic gradients (e.g., $E\|\nabla f_{s_k}(w)\|^2 \leq M^2$ for all w). What do you observe?*

3. *Give an example of a function $f(w) = \frac{1}{n}\sum_{i=1}^{n} f_i(w)$ that shares at least one global minimum with the n functions f_i, but f_i can have infinite more global minima.*

Exercise 3. *You decide to train a neural network on CIFAR10 with SGD, and the overall loss function*

$$f(w) = \frac{1}{n}\sum_{i=1}^{n} f_i(w)$$

is non-convex. However, you know that the following probabilistic guarantee holds true, for all iterations $t = 1, \ldots, T$ and all samples s_t that are sampled uniformly in $1, \ldots, n$

$$\langle \nabla f_{s_t}(w_t), w_t - w^* \rangle = c \cdot \|w_t - w^*\|^2, \text{ with probability } 1/2 + \delta/2$$
$$\langle \nabla f_{s_t}(w_t), w_t - w^* \rangle = -c \cdot \|w_t - w^*\|^2, \text{ with probability } 1/2 - \delta/2$$

That is, the stochastic gradient is on-average positively correlated with the "correct" direction a little more than half of the times, however the rest of the times the stochastic gradient is pointing toward the wrong direction. Assuming that

$$\|w_0 - w^*\|^2 = D \text{ and } E\|\nabla f_{s_k}(w)\|^2 \leq M^2$$

provide an upper bound on the number of iterations T (as a function of $c, D, \delta, \epsilon, M$) after which SGD with constant stepsize (which you need to determine) will converge to a solution that satisfies

$$E\|w_T - w^*\|^2 = \epsilon.$$

Exercise 4. *You decide to solve a large non-linear regression problem with Gradient Descent where the data points are stored in a distributed way across machines. The function $f(w)$ you decided to minimize is β-smooth but non-convex. When you run GD in parallel over the distributed network of computers, you realize that communicating the gradient between nodes (e.g., the parameter server and the compute nodes) is too expensive. To save some communication, you decide to quantize the gradient vector down to its sign. That is, the algorithm now takes the form:*

$$w_{k+1} = w_k - \gamma \operatorname{sign}(\nabla f(w_k)).$$

Since this is only a binary version of GD, is it possible to show that it converges?

1. *Show that $\|\nabla f(w_k)\|_1 \leq \frac{f(w_k) - f(w_{k+1})}{\gamma} + \beta\gamma d$, where d is the dimension of $\operatorname{sign}(\nabla f(w_k))$.*
2. *Establish convergence rates for $\min_{t \in \{0,1,\ldots,T\}} \|\nabla f(w_t)\|_1$ as a function of T, β, $f(w_0) - f^*$ and d.*

Exercise 5. *A lot of recent research in algorithmic machine learning is studying the effect of implicit regularization when running gradient-based algorithms during training. Regularization is an algorithmic principle that helps avoid overfitting during training, and is sometimes explicitly enforced by adding norm penalties on the learning objective. However, researchers have observed that gradient-based algorithms seem to incur an implicit regularization without any explicit norm penalties on the cost functions.*

Consider the simple linear regressions setup, where we want to minimize an under-determined least squares problem,

$$\min_w f(w) = \min_w \|X^T w - y\|^2$$

where X is a $d \times n$ full column-rank data matrix with $n < d$, and $w \in \mathbb{R}^d$ and $y \in \mathbb{R}^n$. Show that when a gradient descent based model

$$w_{k+1} = w_k - \gamma_k \nabla f(w_k)$$

converges to a solution such that $\|X^T w_T - y\|^2 = 0$, for some $T \geq 1$, then this is also a solution to the following least norm (or regularized) problem

$$\min \|w\|^2$$
$$\text{s.t.: } y = X^T w$$

Hint: The gradient of $\|X^T w - y\|^2$ with respect to w is a linear combination of the columns of X.