MATH 6001. Nonlinear Optimization in Machine Learning.
Assignment 4.

There are 3 problems, each problem is worth 5 points, total is 15 points.

1. Consider the minibatch SGD Algorithm 5.1 in the Lecture Notes applied to the standard objective function

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

in machine learning, with mini–batch size $m \leq n$. The iteration is given by

$$x_{k+1} = x_k - \frac{\eta}{m}\sum_{i\in\mathcal{B}_t} \nabla f_i(x_k)\ ,$$

in which $\mathcal{B}_t \subset \{1, 2, ..., n\}$ is an i.i.d sequence of size–m minibatches uniformly sampled from the index set $[n] = \{1, 2, ..., n\}$ and $\eta > 0$ is the learning rate. That is to say, $\mathcal{B}_t$ is sampled uniformly from all size m–subsets of the index set $[n] = \{1, 2, ..., n\}$ and for different t the choice of $\mathcal{B}_t$ are independent of each other. Show that the stochastic gradient term is an unbiased estimator of the standard gradient $\nabla f = \frac{1}{n}\sum\limits_{i=1}^{n} \nabla f_i(x)$. That is

$$\mathbf{E}\left(\frac{1}{m}\sum_{i\in\mathcal{B}_t} \nabla f_i(x)\right) = \nabla f = \frac{1}{n}\sum_{i=1}^{n} \nabla f_i(x) \tag{0.1}$$

(Hint: Each index $i \in \{1, 2, ..., n\}$ has been repeated C_{n-1}^{m-1} times in all C_n^m possible size m minibatches sampled from $\{1, 2, ..., n\}$ uniformly. Therefore

$$\mathbf{E}\left(\frac{1}{m}\sum_{i\in\mathcal{B}_t} \nabla f_i(x)\right) = \frac{1}{mC_n^m}\sum_{i=1}^{n} C_{n-1}^{m-1}\nabla f_i(x) = \frac{\frac{(n-1)!}{(m-1)!(n-m)!}}{m\frac{n!}{m!(n-m)!}}\sum_{i=1}^{n}\nabla f_i(x) = \frac{1}{n}\sum_{i=1}^{n}\nabla f_i(x)\,.$$

)

2. Following problem 1, calculate the covariance matrix of the stochastic gradient term

$$\mathbf{Cov}\left(\frac{1}{m}\sum_{i\in\mathcal{B}_t} \nabla f_i(x)\right) \equiv \mathbf{E}\left(\frac{1}{m}\sum_{i\in\mathcal{B}_t} \nabla f_i(x) - \nabla f(x)\right)\left(\frac{1}{m}\sum_{i\in\mathcal{B}_t} \nabla f_i(x) - \nabla f(x)\right)^T$$

and show that

$$\mathbf{Cov}\left(\frac{1}{m}\sum_{i\in\mathcal{B}_t} \nabla f_i(x)\right) = \left(\frac{1}{m} - \frac{1}{n}\right)\Sigma_0(x)\ , \tag{0.2}$$

such that

$$\Sigma_0(x) = \frac{1}{n-1}\sum_{i=1}^{n}(\nabla f(x) - \nabla f_i(x))(\nabla f(x) - \nabla f_i(x))^T\ .$$

In particular the equation (0.2) indicates quantitatively the effect of batchsize on the covariance noise magnitude of SGD training. This has been validated to affect the generalization of deep learning models trained by SGD. See Keskar, N.S., Mudigere, D., Nocedal, J., Smelyanskiy, M., Tang, P.T.P., On large–batch training for deep learning: generalization gap and sharp minima. *ICLR*, 2017.

(Hint: Rewrite the stochastic gradient term as

$$\frac{1}{m}\sum_{i\in\mathcal{B}_t}\nabla f_i(x)=\frac{1}{m}\sum_{i=1}^{m}\nabla f_{\zeta_i}(x)\ ,$$

where $\zeta=(\zeta_1,...,\zeta_m)$ is such that ζ_i is picked with equal probability from $\{1,2,...,n\}\backslash\{\zeta_j: 1\le j\le i-1\}$ for all $1\le i\le m$. Then

$$\begin{aligned}&\mathbf{E}\left(\frac{1}{m}\sum_{i\in\mathcal{B}_t}\nabla f_i(x)\right)\left(\frac{1}{m}\sum_{i\in\mathcal{B}_t}\nabla f_i(x)\right)^T\\&=\mathbf{E}\left(\frac{1}{m}\sum_{i=1}^{m}\nabla f_{\zeta_i}(x)\right)\left(\frac{1}{m}\sum_{i=1}^{m}\nabla f_{\zeta_i}(x)\right)^T\\&=\frac{1}{m^2}\sum_{1\le i,j\le m}\mathbf{E}\nabla f_{\zeta_i}(x)\nabla f_{\zeta_j}^T(x)\ .\end{aligned}$$

If $i=j$ the above expectation is

$$\mathbf{E}\nabla f_{\zeta_i}(x)\nabla f_{\zeta_j}^T(x)=\frac{1}{n}\sum_{k=1}^{n}\nabla f_k(x)\nabla f_k^T(x)\ .$$

If $i\neq j$ then the above expectation is

$$\mathbf{E}\nabla f_{\zeta_i}(x)\nabla f_{\zeta_j}^T(x)=\frac{1}{n}\sum_{k=1}^{n}\mathbf{E}(\nabla f_{\zeta_j}(x)|\zeta_i=k)\nabla f_k^T(x)=\frac{1}{n(n-1)}\sum_{j\neq k}\nabla f_j(x)\nabla f_k^T(x)\ .$$

We then make use of (0.1), to see that we have

$$
\begin{aligned}
&\mathbf{Cov}\left(\frac{1}{m}\sum_{i\in\mathcal{B}_t}\nabla f_i(x)\right)\\
&\equiv \mathbf{E}\left(\frac{1}{m}\sum_{i\in\mathcal{B}_t}\nabla f_i(x)-\nabla f(x)\right)\left(\frac{1}{m}\sum_{i\in\mathcal{B}_t}\nabla f_i(x)-\nabla f(x)\right)^T\\
&=\mathbf{E}\left(\frac{1}{m}\sum_{i\in\mathcal{B}_t}\nabla f_i(x)\right)\left(\frac{1}{m}\sum_{i\in\mathcal{B}_t}\nabla f_i(x)\right)^T-\nabla f(x)\nabla f^T(x)\\
&=\frac{1}{m^2}\sum_{1\le i,j\le m}\mathbf{E}\nabla f_{\zeta_i}(x)\nabla f_{\zeta_j}^T(x)-\nabla f(x)\nabla f^T(x)\\
&=\frac{1}{m^2}m\frac{1}{n}\sum_{k=1}^n\nabla f_k(x)\nabla f_k^T(x)+\frac{1}{m^2}m(m-1)\frac{1}{n(n-1)}\sum_{j\neq k}\nabla f_j(x)\nabla f_k^T(x)-\frac{1}{n^2}\sum_{j,k=1}^n\nabla f_j(x)\nabla f_k^T(x)\\
&=\left(\frac{1}{mn}-\frac{m-1}{mn(n-1)}\right)\sum_{k=1}^n\nabla f_k(x)\nabla f_k^T(x)+\left(\frac{m-1}{mn(n-1)}-\frac{1}{n^2}\right)\sum_{j,k=1}^n\nabla f_j(x)\nabla f_k^T(x)\\
&=\left(\frac{1}{m}-\frac{1}{n}\right)\frac{1}{n-1}\sum_{k=1}^n\nabla f_k(x)\nabla f_k^T(x)+\left(\frac{n(m-1)}{m(n-1)}-1\right)\nabla f(x)\nabla f^T(x)\\
&=\left(\frac{1}{m}-\frac{1}{n}\right)\frac{1}{n-1}\sum_{k=1}^n\nabla f_k(x)\nabla f_k^T(x)-\frac{n-1}{n}\left(1-\frac{n(m-1)}{m(n-1)}\right)\frac{n}{n-1}\nabla f(x)\nabla f^T(x)\\
&=\left(\frac{1}{m}-\frac{1}{n}\right)\left(\frac{1}{n-1}\sum_{k=1}^n\nabla f_k(x)\nabla f_k^T(x)-\frac{n}{n-1}\nabla f(x)\nabla f^T(x)\right)\ .
\end{aligned}
$$

This gives (0.2), provided that one can easily check $\Sigma_0(x)=\frac{1}{n-1}\sum_{i=1}^n(\nabla f(x)-\nabla f_i(x))(\nabla f(x)-\nabla f_i(x))^T=\frac{1}{n-1}\sum_{k=1}^n\nabla f_k(x)\nabla f_k^T(x)-\frac{n}{n-1}\nabla f(x)\nabla f^T(x)$.

See Hu, W., Li, C.J., Li, L., Liu, J., On the diffusion approximation of nonconvex stochastic gradient descent. *Annals of Mathematical Science and Applications*, Vol. **4**, No. 1 (2019), pp. 3-32.)

3. Show that when f is strongly convex then the *curvature condition* holds, i.e., for any $x,y\in\mathrm{Dom}(f)$ and $x\neq y$ we have

$$(x-y)^T(\nabla f(x)-\nabla f(y))>0\ .$$

(Hint: Let us say f is strongly m–convex. Then by Taylor's formula

$$\nabla f(x)-\nabla f(y)=\int_0^1\nabla^2 f(y+t(x-y))(x-y)dt\ ,$$

which implies that

$$(x-y)^T(\nabla f(x)-\nabla f(y))=\int_0^1(x-y)^T\nabla^2 f(y+t(x-y))(x-y)dt\geq m\|x-y\|^2>0$$

as desired.)