b

TTIC 31230 Fundamentals of Deep Learning

SGD Problems.

Problem 1: Moving Averages. Consider a sequence of vectors x_0, x_1, ... and two running averages y_t and z_t defined by as follows for $0 < \beta < 1$ and $\gamma > 0$.

$$\begin{aligned} y_0 &= 0 \\ y_{t+1} &= \beta y_t + (1-\beta)x_t \end{aligned}$$

$$\begin{aligned} z_0 &= 0 \\ z_{t+1} &= \beta z_t + \gamma x_t \end{aligned}$$

(a) Suppose that the values x_t are drawn IID from a distribution with mean vector $\overline{x} = E\ x_t$. Give values for

$$\overline{y} = \lim_{t \to \infty} E\ y_t$$

and

$$\overline{z} = \lim_{t \to \infty} E\ z_t$$

as functions of β, γ and $\overline{x}$
Hint: Solve for $E\ y_{t+1}$ as a function of $E\ y_t$ and assume that a limiting expectation exists.
(b) Express z_t as a function of y_t, β and γ.

Problem 2. Variance of an exponential moving average. For two independent random variables x and y and a weighted sum $s = ax + by$ we have

$$\sigma_s^2 = a^2\sigma_x^2 + b^2\sigma_y^2$$

Now consider a runing average for computing $\hat{\mu}_1, \ldots, \hat{\mu}_t$ from $x_1, \ldots, x_t$

$$\hat{\mu}_0 = 0$$

$$\hat{\mu}_t = \left(1 - \frac{1}{N}\right)\hat{\mu}_{t-1} + \frac{1}{N}x_t$$

(a) Assume that the values of x_t are independent and identically distributed with variance σ_x^2. We now have that $\hat{\mu}_t$ is a random variable depending on the draws of x_t. The random variable $\hat{\mu}_t$ has a variance $\sigma_{\hat{\mu},t}^2$. Assume that as $t \to \infty$ we have that $\sigma_{\hat{\mu},t}^2$ converges to a limit (it does). Solve for this limit $\sigma_{\hat{\mu},\infty}^2$. Your solution should yield that for $N = 1$ we have $\sigma_{\hat{\mu},\infty}^2 = \sigma_x^2$ (a sanity check).

(b) Compare your answer to (a) with the variance of an average of N values of x_t defined by

$$\hat{\mu} = \frac{1}{N} \sum_{t=1}^{N} x_t$$

Problem 3. Reformulating Momentum as a Exponential Moving Average. Consider the following update equation.

$$\begin{aligned} y_0 &= 0 \\ y_t &= \left(1 - \frac{1}{N}\right) y_{t-1} + x_t \end{aligned}$$

(a) Assume that y_t converges to a limit, i.e., that $\lim_{t\to\infty} y_t$ exists. If the input sequence is constant with $x_t = c$ for all $t \geq 1$, what is $\lim_{t\to\infty}\ y_t$? Give a derivation of your answer (Hint: you do not need to compute a closed form solution for y_t).

(b) y_t is an exponential moving average of what quantity?

(c) Express y_t as a function of μ_t where μ_t is defined by

$$\begin{aligned} \mu_0 &= 0 \\ \mu_t &= \left(1 - \frac{1}{N}\right) \mu_{t-1} + \frac{1}{N} x_t \end{aligned}$$

Problem 4. Bias Correction Consider the following update equation for computing $y_1, \ldots, y_t$ from $x_1, \ldots, x_t$.

$$y_t = \left(1 - \frac{1}{\min(t, N)}\right) y_{t-1} + \frac{1}{\min(t, N)}\, x_t$$

If $x_t = c$ for all $t \geq 1$ give a closed form solution for y_t.

Problem 5. This problem is on interaction of learning rate and scaling of the loss function.

(a) Consider vanilla SGD on cross entropy loss for classification with batch size 1 and no moment in which case we have

$$\Phi_{t+1} = \Phi_t - \eta \nabla_\Phi \ln P_\Phi(y|x)$$

Now suppose someone uses log base 2 (to get loss in bits) and uses the update

$$\Phi_{t+1} = \Phi_t - \eta' \nabla_\Phi \log_2 P_\Phi(y|x)$$

Suppose that we find that leatning rate η works well for the natural log version (with loss in nats). What value of η' should be used in the second version with loss measured in bits? You can use the relation that $\log_b z = \ln z / \ln b$.

(b) Now consider the following simplified version of RMSprop where for each parameter $\Phi[i]$ we have

$$\Phi_{t+1}[i] = \Phi_t[i] - \frac{\eta}{\sigma_i} \nabla_\Phi \mathcal{L}_\Phi(x_t, y_t)$$

where σ_i is exactly the standard deviation of ith component of the gradient as defined by

$$\begin{aligned} \mu_i &= E_{x,y}\left[\nabla_{\Phi[i]}\; \mathcal{L}_\Phi(x,y)\right] \\ \sigma_i &= \sqrt{E_{x,y}\left[\left(\nabla_{\Phi[i]}\; \mathcal{L}_\Phi(x,y) - \mu_i\right)^2\right]} \end{aligned}$$

If we replace $\mathcal{L}$ by $2\mathcal{L}$ what learning rate η' should we use with loss $2\mathcal{L}$ to get the same temperature?

Problem 6. Adaptive SGD. This problem considers the question of whether the convergence theorem hold for adaptive methods — in the limit as the learning rate goes to zero do adaptive methods converge to a local minimum of the loss.

Consider a generalization of RMSProp where we allow an arbitrary adaptation with with different learning rates for different parameter values. More specifically consider the SGD update equation

$$(1) \quad \Phi_{t+1} = \Phi_t - \eta\left(A(\Phi_t, x_t, y_t) \odot \nabla_\Phi \mathcal{L}(\Phi_t, x_t, y_t)\right)$$

where $\langle x_t, y_t \rangle$ is the tth training pair, $A(\Phi_t, x_t, y_t)$ is an adaptation vector, and $\odot$ is the Haddamard product $(x \odot y)[i] = x[i]\; y[i]$.

Consider the special case given by

$$\begin{aligned} A(\Phi, x, y)[i] &= \frac{1}{\sqrt{s(\Phi, x, y)} + \epsilon} \\ s(\Phi, x, y) &= \frac{1}{d} ||\nabla_\Phi\; \mathcal{L}(\Phi, x, y)||^2 \end{aligned}$$

where d is the dimension of Φ.

(a) For the given interpretation of $A(\Phi, x, y)$, let Φ^* be a parameter setting that is a stationary point of the update equation (1) in the sense that expected update over a random draw from the population is zero. Write this stationary condition on Φ^* explicitly as an expectation equaling zero under the given interpretation of $A(\Phi, x, y)$.

(b) Is Φ^* as defined in part (a) a stationary point of the original loss — a point where the expected gradient of $\mathcal{L}(\Phi^*, x, y)$ is equal to zero?

(c) Do these observations have implications for the adaptive methods described in this class. Explain your answer.

Problem 7. This problem is on a non-standard form of adaptive learning rates. In general when we consider the significance of a change Δx to a number x it is reasonable to consider the change as a percentage of x. For example, a baseline annual raise in salary is often a percentage raise when different employees have significantly different salaries. So we might consider the following "multiplicative update SGD" which we will write here for batch size 1.

$$\Phi^{t+1}[i] = \Phi^t[i] - \eta \ \max(\epsilon, |\Phi^t[i]|) \ \ \hat{g}(\Phi, x_t, y_t)[i] \tag{1}$$

where $\hat{g}(\Phi, x, y)$ abbreviates the gradient $\nabla_\Phi \mathcal{L}(\Phi, x, y)$ where $\mathcal{L}(\Phi, x, y)$ is the loss for the training point (x, y) at parameter setting Φ, and where and $\hat{g}(\Phi, x, y)[i]$ is the ith component of the gradient. For $|\Phi^t[i]| >> \epsilon$ this is a multiplicative update. Multiplicative updates have a long history and rich theory for mixtures of experts prior to the deep revolution. However, I do not know of a citation for the above multiplicative variant of SGD (let me know if you find one later). The parameter ϵ allows a weight to flip sign — to pass through zero more easily. Recall that a stationary point is a parameter setting where the total gradient is zero.

$$\sum_{(x,y)\sim \text{Train}} \nabla_\Phi \ \mathcal{L}(x, y) = 0 \tag{2}$$

(a) At a stationary point of the loss function, is the expected update of equation (4) over a random draw of (x_t, y_t) always equal to zero. In other words, is a stationary point of the loss function also a stationary point of the update equation?

(b) Consider an adaptive algorithm which makes the update proportional to the loss. i.e.,

$$\Phi^{t+1} = \Phi^t - \eta \ \mathcal{L}(\Phi, x_t, y_t) \ \hat{g}^t \tag{3}$$

Is a stationary point of the loss function always a stationary point of the update defined by (6)? Justify your answer.

You can assume that there exists a training set of two points (x_1, y_1) and (x_2, y_2) and a stationary point of the loss function Φ with $\mathcal{L}(\Phi, x_1, y_1) \neq \mathcal{L}(\Phi, x_2, y_2)$ and $\nabla_\Phi(\Phi, x_1, y_1) \neq \nabla_\Phi(\Phi, x_2, y_2)$.

Problem 8. This problem is on interaction of learning rate and scaling of the loss function.

(a) Consider vanilla SGD on cross entropy loss for classification with batch size 1 and no moment in which case we have

$$\Phi_{t+1} = \Phi_t - \eta \nabla_\Phi \ln P_\Phi(y|x)$$

Now suppose someone uses log base 2 (to get loss in bits) and uses the update

$$\Phi_{t+1} = \Phi_t - \eta' \nabla_\Phi \log_2 P_\Phi(y|x)$$

Suppose that we find that leatning rate η works well for the natural log version (with loss in nats). What value of η' should be used in the second version with loss measured in bits? You can use the relation that $\log_b z = \ln z / \ln b$.

(b) Now consider the following simplified version of RMSprop where for each parameter $\Phi[i]$ we have

$$\Phi_{t+1}[i] = \Phi_t[i] - \frac{\eta}{\sigma_i} \nabla_\Phi \mathcal{L}_\Phi(x_t, y_t)$$

where σ_i is exactly the standard deviation of ith component of the gradient as defined by

$$\begin{aligned} \mu_i &= E_{x,y}\left[\nabla_{\Phi[i]}\ \mathcal{L}_\Phi(x,y)\right] \\ \sigma_i &= \sqrt{E_{x,y}\left[\left(\nabla_{\Phi[i]}\ \mathcal{L}_\Phi(x,y) - \mu_i\right)^2\right]} \end{aligned}$$

If we replace $\mathcal{L}$ by $2\mathcal{L}$ what learning rate η' should we use with loss $2\mathcal{L}$ to get the same temperature?

Problem 9. This problem is on a non-standard form of adaptive learning rates. In general when we consider the significance of a change Δx to a number x it is reasonable to consider the change as a percentage of x. For example, a baseline annual raise in salary is often a percentage raise when different employees have significantly different salaries. So we might consider the following "multiplicative update SGD" which we will write here for batch size 1.

$$\Phi^{t+1}[i] = \Phi^t[i] - \eta\ \max(\epsilon, |\Phi^t[i]|)\ \hat{g}(\Phi, x_t, y_t)[i] \tag{4}$$

where $\hat{g}(\Phi, x, y)$ abbreviates the gradient $\nabla_\Phi \mathcal{L}(\Phi, x, y)$ where $\mathcal{L}(\Phi, x, y)$ is the loss for the training point (x, y) at parameter setting Φ, and where and $\hat{g}(\Phi, x, y)[i]$ is the ith component of the gradient. For $|\Phi^t[i]| >> \epsilon$ this is a multiplicative update. Multiplicative updates have a long history and rich theory for mixtures of experts prior to the deep revolution. However, I do not know of a citation for the above multiplicative variant of SGD (let me know if you find one later). The parameter ϵ allows a weight to flip sign — to pass through zero more easily. Recall that a stationary point is a parameter setting where the total gradient is zero.

$$\sum_{(x,y)\sim \text{Train}} \nabla_\Phi\ \mathcal{L}(x,y) = 0 \tag{5}$$

(a) At a stationary point of the loss function, is the expected update of equation (4) over a random draw of (x_t, y_t) always equal to zero. In other words, is a stationary point of the loss function also a stationary point of the update equation?

(b) Consider an adaptive algorithm which makes the update proportional to the loss. i.e.,

$$\Phi^{t+1} = \Phi^t - \eta\, \mathcal{L}(\Phi, x_t, y_t)\, \hat{g}^t \tag{6}$$

Is a stationary point of the loss function always a stationary point of the update defined by (6)? Justify your answer.

You can assume that there exists a training set of two points (x_1, y_1) and (x_2, y_2) and a stationary point of the loss function Φ with $\mathcal{L}(\Phi, x_1, y_1) \neq \mathcal{L}(\Phi, x_2, y_2)$ and $\nabla_\Phi(\Phi, x_1, y_1) \neq \nabla_\Phi(\Phi, x_2, y_2)$.