Machine Learning Foundation Homework 4

B04705003 Tzu-Chuan, Lin

Problem 1. See Figure 1.

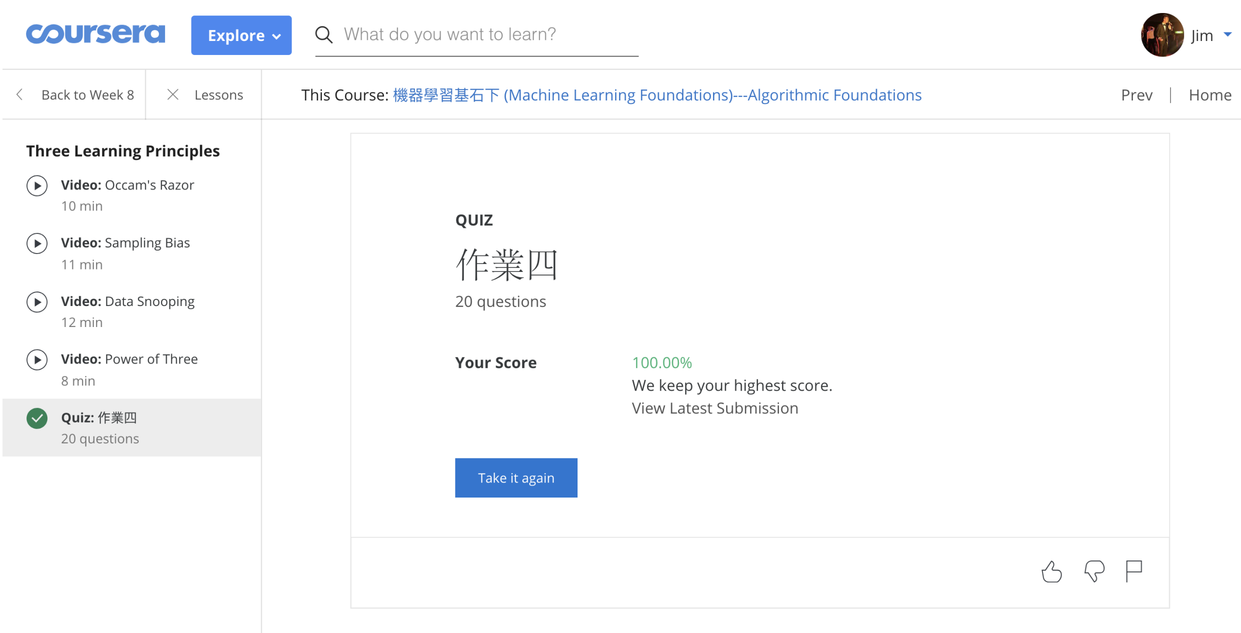

Figure 1: Problem 1

Problem 2. Compute $\nabla_{\mathbf{w}} E_{aug}(\mathbf{w})$.

$$\nabla_{\mathbf{w}} E_{aug}(\mathbf{w}) = \nabla_{\mathbf{w}} E_{in}(\mathbf{w}) + \frac{2\lambda}{N}\mathbf{w}$$
$$\Rightarrow -\eta\nabla_{\mathbf{w}} E_{aug}(\mathbf{w}) = -\eta\nabla_{\mathbf{w}} E_{in}(\mathbf{w}) - \frac{2\eta\lambda}{N}\mathbf{w}$$

Therefore, the update rule becomes:

$$\mathbf{w}_{t+1} \leftarrow \mathbf{w}_t - \eta\nabla_{\mathbf{w}} E_{in}(\mathbf{w}_t) - \frac{2\eta\lambda}{N}\mathbf{w}_t$$
$$\Rightarrow \mathbf{w}_{t+1} \leftarrow (1 - \frac{2\eta\lambda}{N})\mathbf{w}_t - \eta\nabla_{\mathbf{w}} E_{in}(\mathbf{w}_t)$$

Problem 3.

1. When $\mathcal{D}_{val}^{(1)} = \{(1,0)\}$: $g_1^-(x) = \frac{1}{\rho+1}x + \frac{1}{\rho+1}$
   Thus,
$$e_1 = \frac{4}{(\rho+1)^2}$$

2. When $\mathcal{D}_{val}^{(2)} = \{(\rho,1)\}$: $g_2^-(x) = 0$
   Thus,
$$e_2 = 1$$

3. When $\mathcal{D}_{val}^{(3)} = \{(-1,0)\}$:
   (a) Assume $\rho \neq 1$: $g_3^-(x) = \frac{1}{\rho-1}x - \frac{1}{\rho-1}$
   Thus,
$$e_3 = \frac{4}{(1-\rho)^2}$$
   (b) Assume $\rho = 1$ and we choose $[b_1, a_1] = X^{\dagger}\mathbf{y}$ as our optimal solution: $g_3^-(x) = \frac{1}{4}x + \frac{1}{4}$
   Thus,
$$e_3 = (-1 \cdot \frac{1}{4} + \frac{1}{4} - 0)^2 = 0$$

In conclusion:
$$E_{loo}(\rho) = \begin{cases} \frac{1}{3}(\frac{4}{(\rho+1)^2} + 1 + \frac{4}{(1-\rho)^2}), \text{if } \rho \neq 1 \\ \frac{1}{3}(\frac{4}{(\rho+1)^2} + 1 + 0), \text{if } \rho = 1 \end{cases}$$

Problem 4. See Algorithm 1.

Data: $\mathcal{D} = set(X, \mathbf{y}) \cup set(\tilde{X} = \sqrt{\lambda}\mathbf{I}_{K=d+1}, \tilde{\mathbf{y}} = \mathbf{0}) =$
$\{(\mathbf{x}_1, y_1), (\mathbf{x}_2, y_2), ..., (\mathbf{x}_N, y_N), (\tilde{\mathbf{x}}_1, \tilde{y}_1), ..., (\tilde{\mathbf{x}}_{K=d+1}, \tilde{y}_{K=d+1})\}$
$\mathbf{w} = \mathbf{0}$;
begin
  for $t \leftarrow 1$ **to** T **do**
    $(\mathbf{x}, y) \sim \mathcal{D}$;
    $\mathbf{w} \leftarrow \mathbf{w} - 2\eta(\mathbf{w}^T\mathbf{x} - y)\mathbf{x}$;
  end
end

Algorithm 1: SGD with Virtual Examples

Actually, the update rule is derived from $\nabla_{\mathbf{w}} err$:
$$err(\mathbf{w}, \mathbf{x}, y) = (\mathbf{w}^T\mathbf{x} - y)^2$$
$$\Rightarrow \nabla_{\mathbf{w}} err = 2(\mathbf{w}^T\mathbf{x} - y)\mathbf{x}$$

Comparing to the update rule in Question 3, we need to take expectation over stochastic gradient estimator:

$$\begin{aligned}
\mathbb{E}_{(\mathbf{x},y)\sim\mathcal{D}}[\nabla_{\mathbf{w}} err(\mathbf{w},\mathbf{x},y)] &= \frac{N}{N+K}\,\mathbb{E}_{(\mathbf{x},y)\sim set(X,\mathbf{y})}[\nabla_{\mathbf{w}} err] + \frac{K}{N+K}\,\mathbb{E}_{(\mathbf{x},y)\sim set(\tilde{X},\tilde{\mathbf{y}})}[\nabla_{\mathbf{w}} err] \\
&= \frac{N}{N+K}\nabla E_{in}(\mathbf{w}) + \frac{K}{N+K}\,\mathbb{E}_{(\mathbf{x},y)\sim set(\tilde{X},\tilde{\mathbf{y}})}[2(\mathbf{w}^T\mathbf{x})\mathbf{x}] \text{ (because } \tilde{\mathbf{y}} = \mathbf{0}) \\
&= \frac{N}{N+K}\nabla E_{in}(\mathbf{w}) + \frac{K}{N+K}\frac{1}{K}\sum_{i=1}^{K} 2\lambda \begin{bmatrix} 0 \\ \dots \\ \mathbf{w}_i \\ \dots \\ 0 \end{bmatrix} \text{ (because } \tilde{X} = \sqrt{\lambda}\mathbf{I}_{K=d+1}) \\
&= \frac{N}{N+K}\nabla E_{in}(\mathbf{w}) + \frac{2\lambda}{N+K}\mathbf{w}
\end{aligned}$$

Note that when $\eta_{SGD} = \frac{\eta_{GD}(N+K)}{N}$, the update rule of Question 3 and the expected update rule of Question 12(with Virtual Examples) becomes the same:

$$\begin{aligned}
\mathbf{w}_{t+1} &\leftarrow \mathbf{w}_t - \eta_{SGD}\,\mathbb{E}_{(\mathbf{x},y)\sim\mathcal{D}}[\nabla_{\mathbf{w}} err(\mathbf{w}_t,\mathbf{x},y)] \\
\Rightarrow \mathbf{w}_{t+1} &\leftarrow \mathbf{w}_t - \eta_{GD}(\nabla E_{in}(\mathbf{w}_t) + \frac{2\lambda}{N}\mathbf{w}_t) \\
\Rightarrow \mathbf{w}_{t+1} &\leftarrow (1 - \frac{2\eta_{GD}\lambda}{N})\mathbf{w}_t - \eta_{GD}\nabla E_{in}(\mathbf{w}_t)
\end{aligned}$$

Therefore, we can say that adding virtual examples will expectedly result the same behavior with Question 3(Gradient Descent).

Problem 5(Bonus). First, we would need to solve the minimization problem to get w^*:

$$\min_w \mathbb{E}_{x\sim U(0,2\pi)}[(wx - \sin(ax))^2]$$

Expand this objective function:

$$\begin{aligned}
\mathbb{E}_{x\sim U(0,2\pi)}[(wx - \sin(ax))^2] &= \int_0^{2\pi} \frac{1}{2\pi}(\sin^2(ax) - 2wx\sin(ax) + w^2x^2)dx \\
&= \frac{1}{2\pi}(\int_0^{2\pi} \sin^2(ax)dx - 2w\int_0^{2\pi} x\sin(ax)dx + w^2\int_0^{2\pi} x^2dx)
\end{aligned}$$

Notice that this function is convex, we can take gradient equal to 0 and get the minimizer:

$$\frac{d\,\mathbb{E}}{dw} = 0$$
$$\Rightarrow \frac{1}{2\pi}(-2\int_0^{2\pi} x\sin(ax)dx + 2w\int_0^{2\pi} x^2 dx) = 0$$
$$\Rightarrow -2(\frac{-1}{a}(x\cos(ax) - \frac{\sin(ax)}{a})\Big|_0^{2\pi}) + 2w(\frac{8\pi^3}{3}) = 0$$
$$\Rightarrow \frac{2}{a}(2\pi\cos(2\pi a) - \frac{\sin(2\pi a)}{a}) + \frac{16\pi^3 w}{3} = 0$$
$$\Rightarrow \frac{4\pi\cos(2\pi a)}{a} - \frac{2\sin(2\pi a)}{a^2} + \frac{16\pi^3 w}{3} = 0$$
$$\Rightarrow w = \frac{3}{16\pi^3}(-\frac{4\pi\cos(2\pi a)}{a} + \frac{2\sin(2\pi a)}{a^2})$$

Therefore, the deterministic noise of each x is:

$$\Big|\frac{3}{16\pi^3}(-\frac{4\pi\cos(2\pi a)}{a} + \frac{2\sin(2\pi a)}{a^2})x - \sin(ax)\Big|$$