INFR10086 Machine Learning (MLG) Semester 1, 2022/23

Coursework 1: Solutions

1 Marking scheme

- Below are the reference answers, but most questions do not have unique answers.
- Do not penalize comments that are out of scope. Do correct them if they are wrong.
- Be generous but rigorous.
- Raise a question if you are unsure how to mark.

2 Answers

1. a)

$$\begin{aligned}\mathrm{Var}[x] &= \mathbb{E}[(x - \mathbb{E}[x])^2] = \mathbb{E}[x^2 - 2x\mathbb{E}[x] + \mathbb{E}[x]^2] && (1)\\ &= \mathbb{E}[x^2] - 2\mathbb{E}[x]\mathbb{E}[x] + \mathbb{E}[x]^2 && (2)\\ &= \mathbb{E}[x^2] - \mathbb{E}[x]^2 && (3)\end{aligned}$$

b)

$$\begin{aligned}\mathbb{E}[-\log p(x)] &= \mathbb{E}\left[\frac{1}{2}\log(2\pi\sigma^2) + \frac{1}{2\sigma^2}(x-\mu)^2\right] && (4)\\ &= \frac{1}{2}\log(2\pi\sigma^2) + \frac{1}{2\sigma^2}\mathbb{E}[x^2 - 2x\mu + \mu^2] && (5)\\ &= \frac{1}{2}\log(2\pi\sigma^2) + \frac{1}{2\sigma^2}(\mathbb{E}[x^2] - 2\mu\mu + \mu^2) && (6)\\ &= \frac{1}{2}\log(2\pi\sigma^2) + \frac{1}{2\sigma^2}(\mathbb{E}[x^2] - \mu^2) && (7)\\ &= \frac{1}{2}\log(2\pi\sigma^2) + \frac{1}{2\sigma^2}\sigma^2 && (8)\\ &= \frac{1}{2}\log(2\pi\sigma^2) + \frac{1}{2} && (9)\end{aligned}$$

c) Yes, the entropy can be negative. It does sound a bit weird to require negative bits to send a message, but this can happen.

2. a) No, convex functions do not necessarily have unique minimizers. For example, as we will see, $f(x) = \max(0, x)$ is convex in x. Any point in less than or equal to 0 is a minimizer of f.

b) Yes, a strongly convex function has a unique minimizer. Assume x_1^* and x_2^* are both minimizers. If f is μ-strongly convex, then for any x and y,

$$f(x) \geq f(y) + (x-y)^\top \nabla f(y) + \frac{\mu}{2}\|x-y\|^2. \quad (10)$$

If we choose $x = x_1^*$ and $y = x_2^*$, we have

$$f(x_1^*) \geq f(x_2^*) + (x_1^* - x_2^*)^\top \nabla f(x_2^*) + \frac{\mu}{2}\|x_1^* - x_2^*\|^2. \tag{11}$$

Because x_2^* is a minimizer, $\nabla f(x_2^*) = 0$. In addition, $f(x_1^*) = f(x_2^*)$ because both are minimizers. We have

$$\|x_1^* - x_2^*\|^2 \leq 0, \tag{12}$$

so $x_1^* = x_2^*$.

c) The gradient at $x = 1$ is

$$\frac{\partial f}{\partial x}(1) = 2. \tag{13}$$

Given that $x_{t-1} = 1$,

$$x_t = x_{t-1} - \eta_t \frac{\partial f}{\partial x}(x_{t-1}) = 1 - 2\eta_t. \tag{14}$$

If we set, say, $\eta_t = 2$, $x_t = -3$, and $f(x_t) = 9 \geq 1 = f(x_{t-1})$. No, gradient descent does not guarantee to reduce the objective.

d) If $a + b$ attains the maximum on the left-hand side, Because

$$a \leq \max(a, c) \tag{15}$$
$$b \leq \max(b, d) \tag{16}$$

we have

$$a + b \leq \max(a, c) + \max(b + d). \tag{17}$$

Similarly, if $c + d$ attains the maximum on the left-hand side,

$$c + d \leq \max(a, c) + \max(b + d). \tag{18}$$

Based on both cases, we can conclude

$$\max(a + b, c + d) \leq \max(a, c) + \max(b + d). \tag{19}$$

To show $h(x) = \max(f(x), g(x))$ is convex in x, we proceed with the definition. For $0 \leq \alpha \leq 1$,

$$\begin{aligned}
h(\alpha x + (1-\alpha)y) &= \max(f(\alpha x + (1-\alpha)y), g(\alpha x + (1-\alpha)y)) && (20)\\
&\leq \max(\alpha f(x) + (1-\alpha)f(y), \alpha g(x) + (1-\alpha)g(y)) && (21)\\
&\leq \max(\alpha f(x) + \alpha g(y)) + \max((1-\alpha)f(y) + (1-\alpha)g(y)) && (22)\\
&\leq \alpha \max(f(x) + g(y)) + (1-\alpha)\max(f(y) + g(y)) && (23)\\
&\leq \alpha h(x) + (1-\alpha)h(y). && (24)
\end{aligned}$$

By the definition of convexity, h is convex in x.

3. a) Because

$$\frac{\partial^2 f}{\partial s^2} = 2 \geq 0, \tag{25}$$

$f(s) = s^2$ is convex in s.

b) We can see that $\ell_{\text{new}} = (\max(0, 1-s))^2$ where $s = yw^\top \phi(x)$. Suppose $(\max(0, 1-s))^2$ is convex in s (which we will prove later). We use the fact that $g(x) = f(Ax+b)$ is convex if f is convex. Because $yw^\top \phi(x)$ is affine in w, we can conclude that $\ell_{\text{new}} = (\max(0, 1-yw^\top \phi(x))^2$ is convex in w. To show that $(\max(0, 1-s))^2$ is convex in s, we can see that it is a composition of two convex functions. If g is convex and non-decreasing and f is convex,

$$g(f(\alpha x + (1-\alpha)y)) \leq g(\alpha f(x) + (1-\alpha)f(y)) \tag{26}$$
$$\leq \alpha g(f(x)) + (1-\alpha)g(f(y)). \tag{27}$$

Now, s^2 for $s \geq 0$ is convex and increasing and $\max(0, 1-s)$ is convex in s. We can conclude that $(\max(0, 1-s))^2$ is convex in s.

c) No, ℓ_{new} is not Lipschitz, because s^2 is not Lipschitz in s.

d)

$$\nabla_w \ell_{\text{new}} = \begin{cases} 2(yw^\top \phi(x) - 1)(y\phi(x)) & \text{if } yw^\top \phi(x) \leq 1 \\ 0 & \text{otherwise} \end{cases} \tag{28}$$

e) When $yw^\top \phi(x) > 1$, both functions are 0. When $0 \leq yw^\top \phi(x) \leq 1$, the zero-one loss is 0 but ℓ_{new} is positive. When $yw^\top \phi(x) < 0$, the zero-one loss is 1 but ℓ_{new} is larger than 1.

4. a)

$$\frac{\partial \ell}{\partial \hat{y}} = -2(y - \hat{y}) \tag{29}$$

b) i) Since both $f(z) = 0$ and $g(z) = z$ are convex in z, ReLU $= \max(f(z), g(z)) = \max(0, z)$ is convex in z.

ii) ReLu is not differentiable at 0. Any number between 0 and 1 is a subgradient at that point.

iii) If f is L_f-Lipschitz and g is L_g-Lipschitz, we first show that $h(x) = \max(f(x), g(x))$ is $\max(L_f, L_g)$-Lipschitz. In other words, we need to show for all x and y,

$$|\max(f(x), g(x)) - \max(f(y), g(y))| \leq \max(L_f, L_g)\|x - y\|. \tag{30}$$

Because of Lipschitz continuity,

$$\max(f(x), g(x)) \leq \max(f(y) + L_f\|x-y\|, g(y) + L_g\|x-y\|) \tag{31}$$
$$\leq \max(f(y) + g(y)) + \max(L_f\|x-y\|, L_g\|x-y\|) \tag{32}$$
$$\leq \max(f(y) + g(y)) + \max(L_f, L_g)\|x-y\| \tag{33}$$

We have

$$|\max(f(x), g(x)) - \max(f(y) + g(y))| \leq \max(L_f, L_g)\|x - y\|. \tag{34}$$

Now, $f(z) = 0$ is 0-Lipschitz and $g(z) = z$ is 1-Lipschitz. ReLU is $\max(0, 1)$-Lipschitz.

5. The decision boundaries learned from a discriminative approach and a generative approach is shown in Figure 1. The generative approach models how the data is distributed, not just the decision boundry. The generative approach requires samples to fit the generative model well, and also requires us to assume the right distribution. The discriminative approach does not assume how the data is distributed, and depends more on the samples close to the decision boundary than those far from the decision boundary.

In this particular example, the points with positive labels clearly do not follow a Gaussian distribution. We can also expect the means of the Gaussians to change when we rearrange points far from the decision

boundary. For the discriminative approach, the decision bounary is only determined by the points close to the decision bounary.

This example is not to show that the generative approach is necessarily worse than the discriminative approach. The two have different goals: one is to learn to separate, and the other is to learn the overall shape of the data.

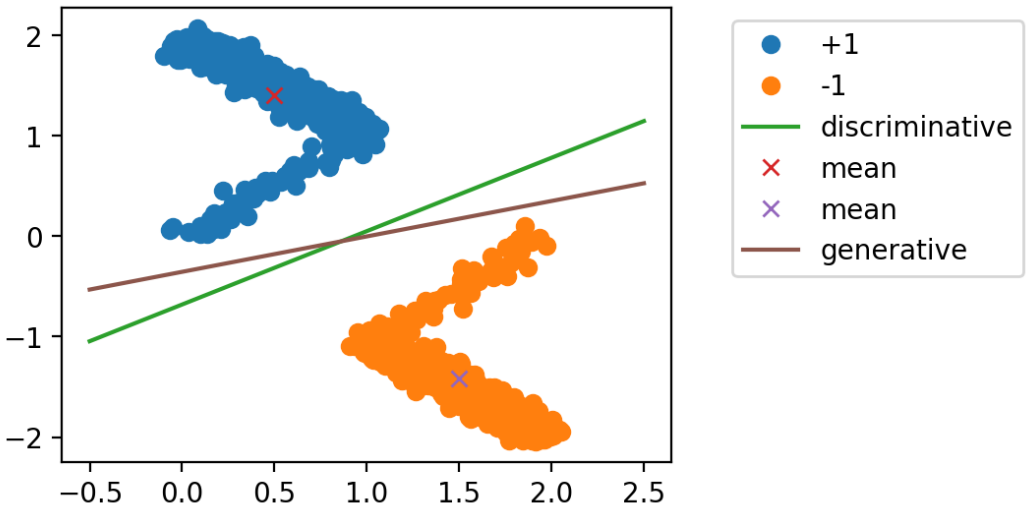

Figure 1: A comparison of decision boundaries learned by a discriminative approach and a generative approach.