INFR10086 Machine Learning (MLG) Semester 2, 2025/6

Coursework

Instructions

- Due date: 9 March, Monday, at 12pm
- The submission is through Gradescope `https://www.gradescope.com/courses/1224477/assignments/7455000`.
- It's best to typeset your answers, but it is fine to submit hand-written answers.

Questions

1. In this question, we will look at the relationship between the hinge loss and support vector machines (SVM).

   Given a data set $\{(\mathbf{x}_1, y_1), \dots, (\mathbf{x}_n, y_n)\}$, soft-margin SVM is defined as the following optimization problem

$$\begin{aligned} \min \quad & \frac{1}{2}\|\mathbf{w}\|^2 + C\sum_{i=1}^{n} \xi_i \\ \text{s.t.} \quad & 1 - y_i \mathbf{w}^\top \mathbf{x}_i \le \xi_i \text{ for } i = 1, \dots, n \\ & \xi_i \ge 0 \text{ for } i = 1, \dots, n \end{aligned} \tag{1}$$

   where C is a hyperparameter.[1] The hinge loss for a single sample (x, y) is defined as $\max(0, 1 - y\mathbf{w}^\top \mathbf{x})$. Minimizing the hinge loss on the entire data set becomes

$$\frac{\lambda}{2}\|\mathbf{w}\|^2 + \sum_{i=1}^{n} \max(0, 1 - y_i \mathbf{w}^\top \mathbf{x}_i), \tag{2}$$

   where λ is another hyperparameter.

   (a) Show that $\xi_i = \max(0, 1 - y\mathbf{w}^\top \mathbf{x}_i)$. In other words, the slack variable ξ is the hinge loss for the data point $(\mathbf{x}_i, y_i)$, and thus soft-margin SVM is equivalent to minimizing the hinge loss.

   [5 marks]

[1] A hyperparameter is not something we optimize as part of the optimization problem, but rather a constant of our choice. In practice, we typically try a few a hyperparameters and choose the best.

From the definition, we see that $\xi_i \geq 1 - y_i w^\top x_i$ and $\xi_i \geq 0$, so it is easy to conclude that $\xi_i \geq \max(0, 1 - y_i w^\top x_i)$. Because ξ_i is in the objective, any ξ_i that satisfies $\xi_i > \max(0, 1 - y_i w^\top x_i)$ can be improved until $\xi_i = \max(0, 1 - y_i w^\top x_i)$ is satisfied. The rest is just wrangling with the constants and let $C = 1/\lambda$.

(b) Find the Lagrangian L of soft-margin SVM and solve $\nabla_{\mathbf{w}} L = 0$. In particular, if α_i is the Lagrange multiplier for $1 - y_i \mathbf{w}^\top \mathbf{x}_i \leq \xi_i$, then

$$\mathbf{w} = \sum_{i=1}^{n} \alpha_i y_i \mathbf{x}_i. \tag{3}$$

The optimal $\mathbf{w}$ is a linear combination of data points from the data set.[2]

[5 marks]

Based on the optimization problem, the Lagrangian can be derived as

$$L = \frac{1}{2}\|\mathbf{w}\|^2 + C\sum_{i=1}^{n} \xi + \sum_{i=1}^{n} \alpha_i (1 - y_i \mathbf{w}^\top \mathbf{x}_i - \xi_i) + \sum_{i=1}^{n} \beta_i \xi_i \tag{4}$$

where $\alpha_i \geq 0$ and $\beta_i \geq 0$ are Lagrange multipliers for $i = 1, \dots, n$. It we take the derivative with respect to $\mathbf{w}$, we have

$$\nabla_{\mathbf{w}} L = 2\mathbf{w} + \sum_{i=1}^{n} \alpha_i (-y_i \mathbf{x}_i). \tag{5}$$

Solving $\nabla_{\mathbf{w}} L = 0$ gives

$$\mathbf{w} = \sum_{i=1}^{n} \alpha_i y_i \mathbf{x}_i. \tag{6}$$

(c) Because the optimal solution needs to satisfy complementary slackness, show that when

$$1 - y_i \mathbf{w}^\top \mathbf{x}_i < 0 \tag{7}$$

then $\alpha_i = 0$.

[5 marks]

Based on complementary slackness, the optimal solution needs to satisfy

$$\alpha_i (1 - y_i \mathbf{w}^\top \mathbf{x}_i - \xi_i) = 0. \tag{8}$$

This leads to the following two cases.

[2] This result is itself important and has a name called the representer theorem.

1) $\alpha_i = 0$ while $1 - y_i \mathbf{w}^\top \mathbf{x}_i - \xi_i < 0$

2) $\alpha_i > 0$ while $1 - y_i \mathbf{w}^\top \mathbf{x}_i - \xi_i = 0$

When $1 - y_i \mathbf{w}^\top \mathbf{x}_i < 0$, we fall under the first case because $\xi_i \geq 0$, and that means $\alpha_i = 0$.

(d) Use the above and conclude that any point $(\mathbf{x}_i, y_i)$ such that

$$1 - y_i \mathbf{w}^\top \mathbf{x}_i < 0 \tag{9}$$

are not part of the optimal $\mathbf{w}$. In particular, the optimal solution does not change if some of these points are removed from the data set.

[5 marks]

From complementary slackness, we know that a point $(\mathbf{x}_i, y_i)$ that satisfies $1 - y_i \mathbf{w}^\top \mathbf{x}_i < 0$ has an $\alpha_i = 0$. Because $\mathbf{w} = \sum_{i=1}^n \alpha_i y_i \mathbf{x}_i$, any point $(\mathbf{x}_i, y_i)$ that has $\alpha_i = 0$ is not involved in the optimal $\mathbf{w}$.

2. In this question, we will prove a convergence result for gradient descent.

(a) A function f is L-Lipschitz if

$$f(\mathbf{x}) - f(\mathbf{y}) \leq L\|\mathbf{x} - \mathbf{y}\| \tag{10}$$

for all $\mathbf{x}$ and $\mathbf{y}$. A function f is L-smooth if its gradient is L-Lipschitz. Show that when a function is both convex and L-smooth, then

$$f(\mathbf{y}) \leq f(\mathbf{x}) + \nabla f(\mathbf{x})^\top (\mathbf{y} - \mathbf{x}) + L\|\mathbf{x} - \mathbf{y}\|^2. \tag{11}$$

You will need to use the Cauchy-Shwarz inequality, which states that $\mathbf{x}^\top \mathbf{y} \leq \|\mathbf{x}\|\|\mathbf{y}\|$ for all $\mathbf{x}$ and $\mathbf{y}$.

[5 marks]

There are two potential solutions: one uses the convexity early on but arrives at a looser bound, while the other does not rely on convexity and is tighter.

1) The first proof is significantly easier to derive. We start with convexity and use the supporting hyperplane property, where

$$f(\mathbf{y}) - f(\mathbf{x}) \leq (\mathbf{y} - \mathbf{x})^\top \nabla f(\mathbf{y}) \tag{12}$$

We simply subtract $(\mathbf{y}-\mathbf{x})^\top \nabla f(\mathbf{x})$ from both sides and get

$$
\begin{aligned}
f(\mathbf{y}) - f(\mathbf{x}) - (\mathbf{y}-\mathbf{x})^\top \nabla f(\mathbf{x}) &\leq (\mathbf{y}-\mathbf{x})^\top (\nabla f(\mathbf{y}) - \nabla f(\mathbf{x})) && (13)\\
&\leq \|\mathbf{y}-\mathbf{x}\| \|\nabla f(\mathbf{y}) - \nabla f(\mathbf{x})\| && (14)\\
&\leq L\|\mathbf{y}-\mathbf{x}\|^2, && (15)
\end{aligned}
$$

where we first use Cauchy-Schwarz and then use the Lipschitz-ness of the gradient.

2) The second proof requires a lot more work. We start with the property of directional derivative, where the derivative along a direction v is the dot product with the gradient, i.e.,

$$\nabla f(\mathbf{x})^\top \mathbf{v} = \lim_{t\to 0} \frac{f(\mathbf{x}+t\mathbf{v}) - f(\mathbf{x})}{t}. \tag{16}$$

Since we are interested in $f(\mathbf{x})^\top(\mathbf{y}-\mathbf{x})$, we replace $\mathbf{v}$ with $\mathbf{y}-\mathbf{x}$ to have

$$\nabla f(\mathbf{x})^\top (\mathbf{y}-\mathbf{x}) = \lim_{t\to 0} \frac{f(\mathbf{x}+t(\mathbf{y}-\mathbf{x})) - f(\mathbf{x})}{t} = g'(0), \tag{17}$$

where $g(t) = f(\mathbf{x}+t(\mathbf{y}-\mathbf{x}))$. We can further relax the constant 0 to any arbitrary point by replacing $\mathbf{x}$ with $\mathbf{x}+\alpha(\mathbf{y}-\mathbf{x})$ and have

$$
\begin{aligned}
\nabla f(\mathbf{x}+\alpha(\mathbf{y}-\mathbf{x}))^\top (\mathbf{y}-\mathbf{x}) &= \lim_{t\to 0} \frac{f(\mathbf{x}+(\alpha+t)(\mathbf{y}-\mathbf{x})) - f(\mathbf{x}+\alpha(\mathbf{y}-\mathbf{x}))}{t} && (18)\\
&= \lim_{t\to 0} \frac{g(\alpha+t) - g(\alpha)}{t} = g'(\alpha), && (19)
\end{aligned}
$$

where $g(t) = f(\mathbf{x}+t(\mathbf{y}-\mathbf{x}))$. What's nice about g is that $f(\mathbf{x}) = g(0)$ and $f(\mathbf{y}) = g(1)$.

Now we need to connect all the ingredients, starting from

$$
\begin{aligned}
f(\mathbf{y}) - f(\mathbf{x}) = g(1) - g(0) &= \int_0^1 g'(t)dt && (20)\\
&= \int_0^1 (\mathbf{y}-\mathbf{x})^\top \nabla f(\mathbf{x}+t(\mathbf{y}-\mathbf{x}))dt, && (21)
\end{aligned}
$$

where the last line is just the Fundamental Theorem of Calculus. The rest is pretty much the same as the first proof. We subtract $(\mathbf{y}-\mathbf{x})^\top \nabla f(\mathbf{x})$ from both sides and get

$$
\begin{aligned}
f(\mathbf{y}) - f(\mathbf{x}) - (\mathbf{y}-\mathbf{x})^\top \nabla f(\mathbf{x}) &= \int_0^1 (\mathbf{y}-\mathbf{x})^\top [\nabla f(\mathbf{x} + t(\mathbf{y}-\mathbf{x})) - \nabla f(\mathbf{x})] dt \tag{22} \\
&\leq \int_0^1 \|\mathbf{y}-\mathbf{x}\| \|\nabla f(\mathbf{x} + t(\mathbf{y}-\mathbf{x})) - \nabla f(\mathbf{x})\| dt \tag{23} \\
&\leq \int_0^1 \|\mathbf{y}-\mathbf{x}\| L \|\mathbf{x} + t(\mathbf{y}-\mathbf{x}) - \mathbf{x}\| dt \tag{24} \\
&= \int_0^1 \|\mathbf{y}-\mathbf{x}\| Lt \|\mathbf{y}-\mathbf{x}\| dt \tag{25} \\
&= L\|\mathbf{y}-\mathbf{x}\|^2 \int_0^1 t dt \tag{26} \\
&= \frac{L}{2}\|\mathbf{y}-\mathbf{x}\|^2 \leq L\|\mathbf{y}-\mathbf{x}\|^2 \tag{27}
\end{aligned}
$$

Even though the proof is a lot more involved, this proof does not require convexity and the constant is slightly better. We will stick with the constant in the coursework, though you might see a better constant in proofs elsewhere.

(b) Consider doing gradient descent

$$\mathbf{x}_t = \mathbf{x}_{t-1} - \eta_t \nabla f(\mathbf{x}_{t-1}) \tag{28}$$

for $t = 1, \dots, k$, on a convex and L-smooth function f.

i. Start with equation (11), use the definition of gradient descent, and show that

$$f(\mathbf{x}_{t-1}) - f(\mathbf{x}_t) \geq \frac{1}{4L} \|\nabla f(\mathbf{x}_{t-1})\|^2, \tag{29}$$

when $\eta_t = \frac{1}{2L}$. Note that the norm is always nonnegative, and our objective, in this case, always decreases. This result is commonly known as the descent lemma.

[2 marks]

We start from equation (11) and let $y = x_t$ and $x = x_{t-1}$ to get

$$
\begin{aligned}
f(\mathbf{x}_{t-1}) - f(\mathbf{x}_t) &\geq (\mathbf{x}_{t-1} - \mathbf{x}_t)^\top \nabla f(\mathbf{x}_{t-1}) - L\|\mathbf{x}_{t-1} - \mathbf{x}\|^2 \tag{30} \\
&= \eta_t \nabla f(\mathbf{x}_{t-1})^\top \nabla f(\mathbf{x}_{t-1}) - L\eta_t^2 \|\nabla f(\mathbf{x}_{t-1})\|^2 \tag{31} \\
&= (\eta_t - L\eta_t^2)\|\nabla f(\mathbf{x}_{t-1})\|^2. \tag{32}
\end{aligned}
$$

We want the improvement in objective $f(\mathbf{x}_{t-1}) - f(\mathbf{x}_t)$ to be as large as possible. Since $\eta_t - L\eta_t^2 = -L(\eta_t - \frac{1}{2L})^2 + \frac{1}{4L}$, we can see that this is a concave parabola and has a maximum $\frac{1}{4L}$ at $\frac{1}{2L}$. In fact, any $\eta_t < \frac{1}{L}$ would guarantee to improve the objective; hence the name, the descent lemma. It's also quite amazing that this result does *not* require convexity.

ii. Expand $\|\mathbf{x}_t - \mathbf{x}^*\|^2$ and use the definition of gradient descent to show

$$\nabla f(\mathbf{x}_{t-1})^\top (\mathbf{x}_{t-1} - \mathbf{x}^*) = L(\|\mathbf{x}_{t-1} - \mathbf{x}^*\|^2 - \|\mathbf{x}_t - \mathbf{x}^*\|^2) + \frac{1}{4L}\|\nabla f(\mathbf{x}_{t-1})\|^2, \quad (33)$$

where $\mathbf{x}^*$ is the optimal solution.

[2 marks]

We use the definition of gradient descent and expand

$$\|\mathbf{x}_t - \mathbf{x}^*\|^2 = \|\mathbf{x}_{t-1} - \eta_t \nabla f(\mathbf{x}_{t-1}) - \mathbf{x}^*\|^2 \quad (34)$$

$$= \|\mathbf{x}_{t-1} - \mathbf{x}^*\|^2 - 2\eta_t \nabla f(\mathbf{x}_{t-1})^\top (\mathbf{x}_{t-1} - \mathbf{x}^*) + \eta_t^2 \|\nabla f(\mathbf{x}_{t-1})\|^2. \quad (35)$$

Rearranging the terms, we get

$$\nabla f(\mathbf{x}_{t-1})^\top (\mathbf{x}_{t-1} - \mathbf{x}^*) = \frac{1}{2\eta_t}(\|\mathbf{x}_{t-1} - \mathbf{x}^*\|^2 - \|\mathbf{x}_t - \mathbf{x}^*\|^2) + \frac{\eta_t}{2}\|\nabla f(\mathbf{x}_{t-1})\|^2 \quad (36)$$

$$= L(\|\mathbf{x}_{t-1} - \mathbf{x}^*\|^2 - \|\mathbf{x}_t - \mathbf{x}^*\|^2) + \frac{1}{4L}\|\nabla f(\mathbf{x}_{t-1})\|^2. \quad (37)$$

iii. Use the descent lemma and show

$$\nabla f(\mathbf{x}_{t-1})^\top (\mathbf{x}_{t-1} - \mathbf{x}^*) \le L(\|\mathbf{x}_{t-1} - \mathbf{x}^*\|^2 - \|\mathbf{x}_t - \mathbf{x}^*\|^2) + f(\mathbf{x}_{t-1}) - f(\mathbf{x}_t). \quad (38)$$

[1 mark]

This is literally using the descent lemma on the last term to get

$$\nabla f(\mathbf{x}_{t-1})^\top (\mathbf{x}_{t-1} - \mathbf{x}^*) = L(\|\mathbf{x}_{t-1} - \mathbf{x}^*\|^2 - \|\mathbf{x}_t - \mathbf{x}^*\|^2) + \frac{1}{4L}\|\nabla f(\mathbf{x}_{t-1})\|^2 \quad (39)$$

$$\le L(\|\mathbf{x}_{t-1} - \mathbf{x}^*\|^2 - \|\mathbf{x}_t - \mathbf{x}^*\|^2) + f(\mathbf{x}_{t-1}) - f(\mathbf{x}_t). \quad (40)$$

iv. Use convexity and equation (38) to show

$$f(\mathbf{x}_{t-1}) - f(\mathbf{x}^*) \le L(\|\mathbf{x}_{t-1} - \mathbf{x}^*\|^2 - \|\mathbf{x}_t - \mathbf{x}^*\|^2) + f(\mathbf{x}_{t-1}) - f(\mathbf{x}_t). \quad (41)$$

In particular,

$$f(\mathbf{x}_t) - f(\mathbf{x}^*) \leq L(\|\mathbf{x}_{t-1} - \mathbf{x}^*\|^2 - \|\mathbf{x}_t - \mathbf{x}^*\|^2). \tag{42}$$

[1 mark]

Based on convexity, we have

$$f(\mathbf{x}^*) \geq f(\mathbf{x}_{t-1}) + (\mathbf{x}^* - \mathbf{x}_{t-1})^\top \nabla f(\mathbf{x}_{t-1}). \tag{43}$$

Rearranging the terms, we can further get

$$\begin{aligned} f(\mathbf{x}_{t-1}) - f(\mathbf{x}^*) &\leq (\mathbf{x}_{t-1} - \mathbf{x}^*)^\top \nabla f(\mathbf{x}_{t-1}) & (44) \\ &\leq L(\|\mathbf{x}_{t-1} - \mathbf{x}^*\|^2 - \|\mathbf{x}_t - \mathbf{x}^*\|^2) + f(\mathbf{x}_{t-1}) - f(\mathbf{x}_t) & (45) \end{aligned}$$

Finally, dropping $f(\mathbf{x}_{t-1})$ from both sides, we have

$$f(\mathbf{x}_t) - f(\mathbf{x}^*) \leq L(\|\mathbf{x}_{t-1} - \mathbf{x}^*\|^2 - \|\mathbf{x}_t - \mathbf{x}^*\|^2). \tag{46}$$

We are now ready for telescoping sum.

v. Take the sum of t from 1 to k on both sides of equation (42) (and divide it by k) to get

$$\frac{1}{k}\sum_{t=1}^{k} f(\mathbf{x}_t) - f(\mathbf{x}^*) \leq \frac{L}{k}(\|\mathbf{x}_0 - \mathbf{x}^*\|^2 - \|\mathbf{x}_k - \mathbf{x}^*\|^2) \leq \frac{L}{k}\|\mathbf{x}_0 - \mathbf{x}^*\|^2. \tag{47}$$

[1 mark]

We can list all the iterates and sum them

$$\begin{aligned} f(\mathbf{x}_k) - f(\mathbf{x}^*) &\leq L(\|\mathbf{x}_{k-1} - \mathbf{x}^*\|^2 - \|\mathbf{x}_k - \mathbf{x}^*\|^2) & (48) \\ f(\mathbf{x}_{k-1}) - f(\mathbf{x}^*) &\leq L(\|\mathbf{x}_{k-2} - \mathbf{x}^*\|^2 - \|\mathbf{x}_{k-1} - \mathbf{x}^*\|^2) & (49) \\ &\vdots & (50) \\ f(\mathbf{x}_1) - f(\mathbf{x}^*) &\leq L(\|\mathbf{x}_0 - \mathbf{x}^*\|^2 - \|\mathbf{x}_1 - \mathbf{x}^*\|^2) & (51) \end{aligned}$$

The terms cancel each other, and this is known as the telescoping sum. We end up with

$$\begin{aligned} \sum_{t=1}^{k} f(\mathbf{x}_t) - kf(\mathbf{x}^*) &\leq L(\|\mathbf{x}_0 - \mathbf{x}^*\|^2 - \|\mathbf{x}_k - \mathbf{x}^*\|^2) & (52) \\ &\leq L\|\mathbf{x}_0 - \mathbf{x}^*\|^2, & (53) \end{aligned}$$

which is the desired result once we divide both sides by k.

vi. Because of the descent lemma, show that

$$f(\mathbf{x}_k) \leq \frac{1}{k}\sum_{t=1}^{k} f(\mathbf{x}_t). \tag{54}$$

[1 mark]

The descent lemma guarantees improvement for every gradient step, so we have

$$\begin{aligned} f(\mathbf{x}_k) &= \frac{1}{k}[f(\mathbf{x}_k) + f(\mathbf{x}_k) + \cdots + f(\mathbf{x}_k)] && (55) \\ &\leq \frac{1}{k}[f(\mathbf{x}_1) + f(\mathbf{x}_2) + \cdots + f(\mathbf{x}_k)] && (56) \\ &= \frac{1}{k}\sum_{t=1}^{k} f(\mathbf{x}_t) && (57) \end{aligned}$$

vii. Finally, putting everything together, we have

$$f(\mathbf{x}_k) - f(\mathbf{x}^*) \leq \frac{L}{k}\|\mathbf{x}_0 - \mathbf{x}^*\|^2, \tag{58}$$

after running k steps of gradient descent on a convex and L-smooth function f.

[1 mark]

This is literally applying the previous result to the first term to get

$$f(\mathbf{x}_k) - f(\mathbf{x}^*) \leq \frac{1}{k}\sum_{t=1}^{k} f(\mathbf{x}_t) - f(\mathbf{x}^*) \leq \frac{L}{k}\|\mathbf{x}_0 - \mathbf{x}^*\|^2. \tag{59}$$

viii. Given the above, what is the convergence rate of gradient descent on a convex and L-smooth function?

[1 mark]

Since we have $f(\mathbf{x}_k) - f(\mathbf{x}^*) = O(1/k)$, by definition, the convergence rate is sublinear.

3. In programming, we have various programming constructs to work with, such as if statements and for loops. In this question, we will look at learning the following if statement.

```
**if** $\mathbf{x}$ has property $\zeta$ **then**
  **return** $f(\mathbf{x})$
```

```
else
  return $g(\mathbf{x})$
end if
```

We can let $z = +1$ when $\mathbf{x}$ has property ζ, and let $z = -1$ otherwise. The if statement can then be rewritten into

$$\text{cond}(\mathbf{x}, f, g) = \begin{cases} f(\mathbf{x}) & \text{if } p(z = +1|\mathbf{x}) \geq 0.5 \\ g(\mathbf{x}) & \text{if } p(z = +1|\mathbf{x}) < 0.5 \end{cases} \tag{60}$$

Suppose the result of cond is send immediately to a loss function L, i.e., computing $L(\text{cond}(x, f, g), y)$ for a labeled sample (x, y).

Because we do not know ahead of time whether z is going to be $+1$ or -1, we can only measure the loss in expectation

$$\mathbb{E}_z[L(\text{cond}(\mathbf{x}, f, g), y)] = p(z = +1|\mathbf{x})L(f(\mathbf{x}), y) + p(z = -1|\mathbf{x})L(g(\mathbf{x}), y). \tag{61}$$

(a) Suppose we parameterize $p(z|\mathbf{x})$ as

$$p(z = +1|\mathbf{x}) = \frac{1}{1 + \exp(-\mathbf{w}^\top \mathbf{x})}. \tag{62}$$

Show that

$$\nabla_{\mathbf{w}} \mathbb{E}_z[L(\text{cond}(\mathbf{x}, f, g), y)] = p(z = +1|\mathbf{x})p(z = -1|\mathbf{x})(L(f(\mathbf{x}), y) - L(g(\mathbf{x}), y))\mathbf{x}. \tag{63}$$

In particular, learning $\mathbf{w}$ only requires computing $f(\mathbf{x})$ and $g(\mathbf{x})$ but does not require backpropagation through f and g.

[5 marks]

Based on chain rule, we have

$$\begin{aligned} \nabla_{\mathbf{w}} p(z = +1|\mathbf{x}) &= \frac{-1}{1 + \exp(-\mathbf{w}^\top \mathbf{x})} \frac{\exp(-\mathbf{w}^\top \mathbf{x})}{1 + \exp(-\mathbf{w}^\top \mathbf{x})}(-\mathbf{x}) && (64) \\ &= p(z = +1|\mathbf{x})p(z = -1|\mathbf{x})\mathbf{x} && (65) \\ \nabla_{\mathbf{w}} p(z = -1|\mathbf{x}) &= \nabla_{\mathbf{w}}(1 - p(z = +1|\mathbf{x})) = -p(z = +1|\mathbf{x})p(z = -1|\mathbf{x})\mathbf{x} && (66) \end{aligned}$$

By the definition of $\mathbb{E}_z[L(\text{cond}(\mathbf{x}, f, g), y)]$,

$$\nabla_{\mathbf{w}} \mathbb{E}_z[L(\text{cond}(\mathbf{x}, f, g), y)] = p(z = +1|\mathbf{x})p(z = -1|\mathbf{x})(L(f(\mathbf{x}), y) - L(g(\mathbf{x}), y))\mathbf{x}. \tag{67}$$

(b) If f and g happen to have the same return type, we have

$$\mathbb{E}_z[\text{cond}(\mathbf{x}, f, g)] = p(z = +1|\mathbf{x})f(\mathbf{x}) + p(z = -1|\mathbf{x})g(\mathbf{x}). \tag{68}$$

When L is convex in the first argument, show that

$$L(\mathbb{E}_z[\text{cond}(\mathbf{x}, f, g)], y) \leq \mathbb{E}_z[L(\text{cond}(\mathbf{x}, f, g), y)]. \tag{69}$$

[5 marks]

Jensen's inequality says that

$$f(\mathbb{E}[x]) \leq \mathbb{E}[f(x)], \tag{70}$$

when f is convex. We simply apply Jensen's inequality and get

$$L(\mathbb{E}_z[\text{cond}(\mathbf{x}, f, g)], y) \leq \mathbb{E}_z[L(\text{cond}(\mathbf{x}, f, g), y)]. \tag{71}$$

(c) Discuss the pros and cons of measuring $L(\mathbb{E}_z[\text{cond}(\mathbf{x}, f, g)], y)$ or $\mathbb{E}_z[L(\text{cond}(\mathbf{x}, f, g), y)]$, in particular, when are they equal and which one we should use when they are equal.

[5 marks]

There are cases such that the two are equal, for example, when L is linear or when z is always +1 or -1. In practice, it is unlikely that the two are equal.

In terms of the pros and cons, there are many other issues. Computing $\mathbb{E}_z[L(\text{cond}(\mathbf{x}, f, g), y)]$ can be expensive. In fact, when we have many branches in a large program, $\mathbb{E}_z[L(\text{cond}(\mathbf{x}, f, g), y)]$ can scale exponentially with the number of if statements. Computing $\mathbb{E}_z[\text{cond}(\mathbf{x}, f, g)]$ can be local within each if statement.

However, the loss function might not be nice. For example, when L only expects, say, one-hot vectors, we cannot really evaluate $L(\mathbb{E}_z[\text{cond}(\mathbf{x}, f, g)], y)$, because $\mathbb{E}_z[\text{cond}(\mathbf{x}, f, g)]$ is not guaranteed to be one-hot.