Discussion 5

Note: *Your TA will probably not cover all the problems on this worksheet. The discussion worksheets are not designed to be finished within an hour. They are deliberately made slightly longer so they can serve as resources you can use to practice, reinforce, and build upon concepts discussed in lectures, discussions, and homework.*

This Week's Cool AI Demo/Video:
Deep Dream: `https://www.youtube.com/watch?v=DgPaCWJL7XI`

Explanation for Deep Dream: `https://youtu.be/ta5fdaqDT3M?si=TfGM_QnBBHoJ4sbl&t=2590`

1 Maximum Likelihood Estimation with Laplace Noise

Suppose we have a regression model with parameters $w \in \mathbb{R}^D$ and observations (x_n, t_n) for $n = 1, \ldots, N$. We assume that the observed targets are generated as

$$t_n = x_n^\top w + \epsilon_n,$$

where the noise terms ϵ_n are i.i.d. from a Laplace distribution:

$$p(\epsilon) = \frac{1}{2b} \exp\left(-\frac{|\epsilon|}{b}\right).$$

(a) Write down the likelihood $p(\mathcal{D}|w)$ for the parameters w given the dataset $\mathcal{D} = \{(x_n, t_n)\}_{n=1}^N$.

(b) What is the log-likelihood? Simplify your answer as much as possible.

(c) Show that maximizing the log-likelihood is equivalent to minimizing the **mean absolute error (MAE)**, defined as

$$\text{MAE}(w) = \frac{1}{N}\sum_{n=1}^{N} |t_n - x_n^\top w|.$$

(d) Now consider instead a standard regression model with *Gaussian* noise. Let $t = (t_1, \ldots, t_N)^\top \in \mathbb{R}^N$ denote the vector of observed targets, and let $X \in \mathbb{R}^{N \times D}$ be the design matrix whose rows are the feature vectors $x_n^\top$. Now, we assume our observations follow the model:

$$t_n = x_n^\top w + \varepsilon_n, \qquad \varepsilon_n \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0, \sigma^2).$$

The likelihood of our data under Gaussian noise is

$$p(t \mid X, w) = \prod_{n=1}^{N} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(t_n - x_n^\top w)^2}{2\sigma^2}\right).$$

Place an independent Laplace prior on the coefficients w using $\mu = 0$ and $b = 2\sigma^2/\lambda$ such that

$$p(w) \;=\; \prod_{d=1}^{D} \frac{\lambda}{4\sigma^2}\, \exp\big(-\frac{\lambda}{2\sigma^2}|w_d|\big).$$

Recall that the **Maximum a Posteriori (MAP) estimate** chooses the parameter w that maximizes the posterior density:

$$\hat{w}_{\text{MAP}} = \arg\max_{w}\; p(w \mid t, X) \;\;=\;\; \arg\max_{w}\; p(t \mid X, w)\, p(w).$$

Show that this MAP estimate is equivalent to the solution to the Lasso (L_1 regularized) regression problem.

(e) Briefly contrast the two estimators:

1. **Laplace-noise MLE**: Minimizes the MAE (L_1 loss on residuals)
2. **Gaussian-noise + Laplace-prior MAP (Lasso)**: Minimizes squared error with an L_1 penalty on coefficients

2 Bias-Variance Trade-off

As we saw in lecture, the expected error for a model f_w at a test point x can be decomposed as:

$$\mathbb{E}\left[(t - f_w(x))^2\right] = \underbrace{\mathbb{E}\left[(t - h(x))^2\right]}_{(i)} + \underbrace{\left(h(x) - \mathbb{E}[f_w(x)]\right)^2}_{(ii)} + \underbrace{\mathbb{E}\left[\left(\mathbb{E}[f_w(x)] - f_w(x)\right)^2\right]}_{(iii)}, \tag{1}$$

where $f_w(x)$ is the prediction of our model, parameterized by w, trained on a particular dataset.

We set t to be the true label for the test point, and assume it comes from some underlying ground-truth function h such that

$$t = h(x) + \epsilon,$$

where ϵ is random noise inherent in the system. Assume $\mathbb{E}[\epsilon] = 0$ and $\text{Var}(\epsilon) = \sigma^2$.

(a) From the decomposition above, consider the following symbols:

- t: the true test label
- w: the weights of the fitted model
- x: the test input

Which of these are random variables, and which are not? Briefly explain.

(b) Rewrite equation (1) in terms of σ^2, b (the **model's bias**), and v (the **model's variance**).

(c) Suppose you're tasked with predicting housing prices using a regression model.

(i) If you move from linear regression to a neural network model with many more parameters, how do you expect the bias and variance to change?

(ii) What if you keep linear regression but add a ridge (L_2) regularization term. How does this affect bias and variance?