Student ID: Student Name (print):

Problem 1: Philosophical discussion [10 marks]

A student was enrolled in CMPUT 466/566 but did not attend lectures. The student decided to prepare for the exam by doing written assignments, because it was reasonable to assume the problems are iid in exams and assignments.

With the help of the reference solutions provided by the instructor, the student was able to solve all problems in the written assignments, thus having a high expectation on the exam.

Unfortunately, it later turned out that the student did not perform well in the exam, although the problems were indeed iid compared with written assignments.

(a) [5 marks] Explain why this happened from the perspective of machine learning (or human learning, where *learning* is in the sense of "machine learning").
(b) [5 marks] Give one suggestion to the student on how to better estimate the performance in the exam with these iid written assignments.

Hint: A few words suffice. Longer solutions may not lead to higher marks. A longer solution with a wrong statement will lead to mark deduction.

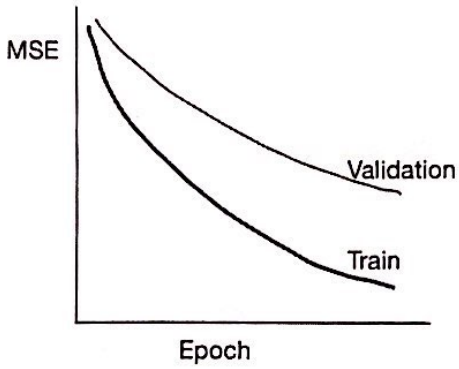
(a) Overfitted to written assignments

(b) Split written assignments into 2 parts. One for validation.

Problem 2: Regularization [15 marks]

Suppose we see a learning curve as below, where the horizontal axis is the number of epochs (iterations over all data samples), and the vertical axis is the mean square error, which is our training objective (excluding the regularization term) and also the measure of success.

(a) [5 marks] If we put more regularization, e.g., increasing λ in 3(b), how would the training curve respond (going up, going down, remaining the same, unknown, etc) in a typical scenario?
(b) [5 marks] If we put more regularization, e.g., increasing λ in 3(b), how would the validation curve respond in a typical scenario?
(c) [5 marks] If we would like to obtain a better machine learning model, shall we put more or less regularization?



(a) Going up

(b) Unknown

(c) Unknown,

(No explanation needed for this problem.)

Student ID: Student Name (print):

Problem 3: Max *a posterior* parameter estimation [30 marks]

(a) [10 marks] Suppose we have a probabilistic machine learning model $p_\theta(\mathcal{D})$, where $\mathcal{D}$ is the dataset and θ is model parameters. Give the generic formula for max *a posteriori* (MAP) inference (i.e., write an equation like $\widehat{\theta}_{\text{MAP}} = \cdots$).

(b) [10 marks] In the lectures, we know that a Gaussian prior $\mathcal{N}(0, \sigma^2)$ for MAP estimation is equivalent to ℓ_2-penalized mean square error (MSE) for linear regression, i.e., the training objective of MAP estimation is $J = J_{\text{MSE}} + \lambda \|\boldsymbol{w}\|_2^2$ for some λ. Prove this.

(c) [10 marks] In the lectures, we also know that there is some correspondence between σ and λ. Express λ in σ, i.e., $\lambda = f(\sigma)$ for some function f. What is f? Provide derivations.

Hint:

- You may make your assumptions, but state your assumptions explicitly and explain every additional variable, function, probability, etc.
- *Mean square error (MSE)* penalizes 1/2 * the mean (over all data samples) of the square of predicted value and the true value. Dropping constants (1/2 or the number of data samples in the denominator) may result in an error up to a multiplicative constant in (c).
- A Gaussian distribution $x \sim \mathcal{N}(\mu, \sigma^2)$ means that $p(x) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left\{-\frac{1}{2}(\frac{x-\mu}{\sigma})^2\right\}$
- The ℓ_2-norm of a vector $\boldsymbol{v}$ is the sum of the square of every element in $\boldsymbol{v}$.
- We usually work with log probabilities rather than the original probabilities.
- For (b), you need to derive MLE within MAP, so that you can solve (c) quantitatively.

(a) $\hat{\theta}_{MAP} = \arg\max\ p(\theta) \cdot p(\mathcal{D}|\theta)$ where $p(\theta)$ is the prior

$p(\mathcal{D}|\theta) = p_\theta(\mathcal{D})$ is the likelihood

(b) Assume the parameters $w \in \mathbb{R}^{d+1}$ has prior w_i iid $N(0, \sigma_w)$

Assume data are generated by $y^{(m)} = w^T x^{(m)} + \varepsilon^{(m)}$, $m = 1 \ldots M$

where $\varepsilon^{(m)}$ iid $N(0, \sigma_\varepsilon)$

Student ID: Student Name (print):

(Intentionally blank page for the student to provide solutions to Problem 3)

$$\hat{w}_{MAP} = \arg\max \; p(w) \cdot p(\mathcal{D} \mid w)$$

$$= \arg\max \prod_{i=0}^{d} p(w_i) \prod_{m=1}^{M} p(y^{(m)} \mid x^{(m)}, w)$$

$$= \arg\max \log \prod_{i=1}^{d} p(w_i) \prod_{m=1}^{M} p(y^{(m)} \mid x^{(m)}, w)$$

$$= \arg\max \sum_{i=1}^{d} \log p(w_i) + \sum_{m=1}^{M} \log p(y^{(m)} \mid x^{(m)}, w)$$

$$= \arg\max \sum_{i=1}^{d} \log \exp\left\{ -\frac{w_i^2}{2\sigma_w^2} \right\} + \sum_{m=1}^{M} \log \exp\left\{ -\frac{(w^T x^{(m)} - y^{(m)})^2}{2\sigma_\varepsilon^2} \right\}$$

$$= \arg\max \sum_{m=1}^{M} \left(-\frac{1}{2\sigma_\varepsilon^2} (y^{(m)} - w^T x^{(m)})^2 \right) + \sum_{i=1}^{d} \left(-\frac{1}{2\sigma_w^2} w_i^2 \right)$$

$$= \arg\max -\frac{M}{\sigma_\varepsilon^2} \left[\sum_{m=1}^{M} \frac{1}{2M} (y^{(m)} - w^T x^{(m)})^2 + \frac{\sigma_\varepsilon^2}{2M\sigma_w^2} \sum_{i=1}^{d} w_i^2 \right]$$

$$= \arg\min \frac{1}{2M} \sum_{m=1}^{M} (y^{(m)} - w^T x^{(m)})^2 + \frac{\sigma_\varepsilon^2}{2M\sigma_w^2} \|w\|_2^2$$

$$= \arg\min J_{MSE} + \lambda \|w\|_2^2$$

(C) $\lambda = \dfrac{\sigma_\varepsilon^2}{2M\sigma_w^2}$

Mark deduction: Didn't distinguish σ_ε, σ_w.

Student ID: Student Name (print):

Problem 4: ℓ_1-penalty for MSE [35 marks]

(a) [20 marks] We know that mean square error (MSE) is convex for linear regression. Prove that ℓ_1-penalized MSE for linear regression is also convex.

(b) [10 marks] Give an optimization algorithm that solves the above problem.

(c) [5 marks] It is known that ℓ_1-penalized MSE will lead to a sparse solution for linear regression. Discuss the main pitfall/drawback/challenge of your algorithm in terms of the sparsity of your solution (what and why).

Hint:

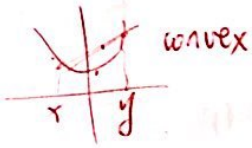
- Dropping constants in MSE is OK for this problem because we only need to prove convexity. And obviously, a convex function multiplied by a constant is still convex (no need to prove).
- *ℓ_1-penalty* is a regularization term of $\lambda\|w\|_1$ where w is the parameters and λ is a hyperparameter. $\|v\|_1$ is the *ℓ_1-norm of a vector* v, which is the sum of the absolute value of every element in the vector v.
- Roadmap to (a)
  1. [10 marks] First prove that the sum of two convex functions (with the same domain) is a convex function.
  2. [10 marks] Then, show that ℓ_1-penalized MSE is indeed convex. If a student cannot solve (a.1), the student may directly solve (a.2) assuming (a.1) is known.

  Note: Not following the roadmap is also acceptable, but detailed marking would also change.
- Suggestion for (b)
  - If you give a closed-form solution, please provide detailed derivations.
  - If you give a gradient-based solution, please derive the gradient and present a pseudo-code algorithm.

(a.1): Let f and g be convex functions on a convex domain S.

~~for convex~~

$\forall x, y \in S, \forall \lambda \in (0,1)$



f convex $\Rightarrow$ $f(\lambda x + (1-\lambda)y) \le \lambda f(x) + (1-\lambda) f(y)$ ①

g convex $\Rightarrow$ $g(\lambda x + (1-\lambda)y) \le \lambda g(x) + (1-\lambda) g(y)$ ②

① + ②:

$$f(\lambda x + (1-\lambda)y) + g(\lambda x + (1-\lambda)y) \le \lambda f(x) + (1-\lambda) f(y) + \lambda g(x) + (1-\lambda) g(y)$$

$\Rightarrow$

$$(f+g)(\lambda x + (1-\lambda)y) \le \lambda (f+g)(x) + (1-\lambda)\cdot(f+g)(y)$$

$\Rightarrow$ $f+g$ is convex.

Mark deduction

Gradient-based solution is wrong, because the functions may not be differentiable

Student ID: Student Name (print):

(Intentionally blank page for the student to provide solutions to Problem 4)

(a.2) $J = J_{MSE} + \|w\|_1$

We know J_{MSE} is convex. we will show $\|w\|_1$ is convex

$\forall\ u, v \in \mathbb{R}^{d+1}$, $\forall \lambda \in (0,1)$

$$\|\lambda u + (1-\lambda) v\| \le \|\lambda u\| + \|(1-\lambda) v\| \quad \text{(Triangle inequality)}$$
$$= \lambda \|u\| + (1-\lambda)\|v\|$$

$\Rightarrow \|\cdot\|$ is convex

Mark deduction: $|w_i|$ convex in w_i $\Rightarrow \|w\|$ convex is w.
This is incorrect, because $\|w\|$ may not be jointly convex for $w_0, \cdots, w_d$.

(b) No closed-form solution, because of non-differentiability.

Gradient-based solution:

$$\nabla_w J = \nabla_w J_{MSE} + \nabla_w \lambda \|w\|_1$$
$$= \frac{1}{M} X^T (Xw - y) + \lambda \cdot \text{sign}(w)$$

where $\text{sign}(w) = \begin{pmatrix} \text{sign}(w_0) \\ \vdots \\ \text{sign}(w_d) \end{pmatrix}$ $\quad \text{sign}(w_i) = \begin{cases} 1 & \text{if } w_i > 0 \\ 0 & \text{if } w_i = 0 \\ -1 & \text{if } w_i < 0 \end{cases}$

Gradient Descent

For epoch / batch

$$w^{(new)} = w^{(old)} - \alpha \nabla_w J(w)\Big|_{w = w^{(old)}}$$

where α is the learning rate

Mark deduction

Closed-form solution is wrong

Student ID: Student Name (print):

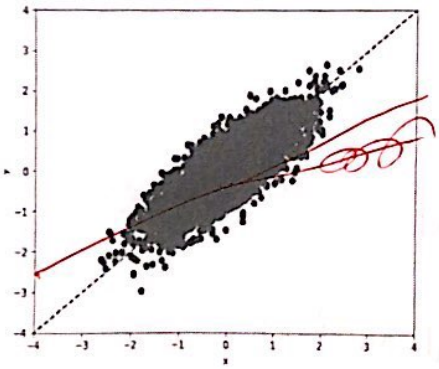
(c) Due to the nature of gradient, we may not get exact 0 for any of w_i. The resulting ~~result~~ w may not be sparse.

Problem 5: Coding Assignment 1 [10 marks]

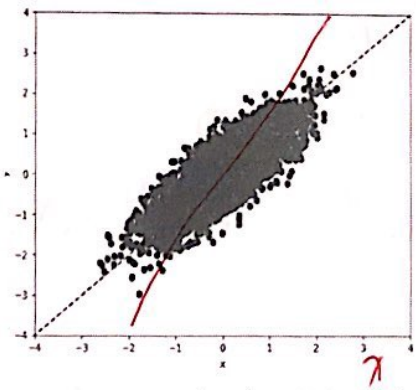
In Problem 1 of Coding Assignment 1, we explored x2y regression and y2x regression. Notice that, in both plots below, the horizontal axis is the x-axis and the vertical axis is y-axis.

(a) Draw qualitatively the line of the x2y regression model in the left plot.

(b) Draw qualitatively the line of the y2x regression model in the right plot.



x2y regression (predicting y from x)



y2x regression (predicting y from x)

END OF MIDTERM EXAM