Problem 1 [100 marks]. Consider the ℓ_2-penalized mean square error (MSE) $J = \frac{1}{2M}\|\mathbf{X}\mathbf{w} - \mathbf{t}\|_2^2 + \lambda\|\mathbf{w}\|_2^2$, where $\mathbf{w} \in \mathbb{R}^{d+1}$ is the weight vector (with an augmented feature) and M is the number of samples.

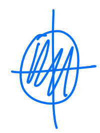
a) [10 marks] What are the dimensions of $\mathbf{X}$ and $\mathbf{t}$, respectively? ***Hint:*** No explanation needed.

$X \in \mathbb{R}^{M \times (d+1)}$ $\qquad$ $t \in \mathbb{R}^{M}$

b) [10 marks] Transform the soft ℓ_2-penalty in the form of a hard constraint. Write the formula(s). Draw the feasible set of the constraint (i.e., the region that satisfies the constraint). ***Hint:*** You may assume $\mathbf{w} \in \mathbb{R}^2$ and use C to represent the constant involved. No proof/explanation needed.

minimize $\frac{1}{2M}\|Xw - t\|_2^2$

subject to $\|w\|_2^2 \leq C$



c) [5 marks] Generally speaking, is the resulting solution sparse? And Why?

No.

The feasible set is round (no angle / edges / ...)

The minimum is unlikely to be on some axis.

d) [15 marks] Derive $\frac{\partial J}{\partial \mathbf{w}}$. Please provide derivation steps. ***Hints:*** You may use either 1) scalar calculus and organize partial derivatives in the vector form, or 2) the matrix calculus identities

- If $\mathbf{A}$ is not a function of $\mathbf{x}$, then $\nabla_{\mathbf{x}}\mathbf{A}\mathbf{x} = \mathbf{A}^\top$
- If $\mathbf{A}$ is not a function of $\mathbf{x}$ and $\mathbf{A}$ is symmetric, then $\nabla_{\mathbf{x}}\mathbf{x}^\top\mathbf{A}\mathbf{x} = 2\mathbf{A}\mathbf{x}$

Please also note that $\|\mathbf{x}\|_2^2 = \mathbf{x}^\top\mathbf{x} = \mathbf{x}^\top\mathbf{I}\mathbf{x}$, where $\mathbf{I}$ is an identity matrix.

$$J = \frac{1}{2M}(w^T X^T - t^T)(Xw - t) + \lambda w^T w$$

$$= \frac{1}{2M}(w^T X^T X w - 2 w^T X^T t + t^T t) + \lambda w^T w$$

$$\frac{\partial J}{\partial w} = \frac{1}{2M}(2 X^T X w - 2 \cdot X^T t) + 2\lambda w$$

$$= \left(\frac{1}{M} X^T X + 2\lambda I\right) w - \frac{1}{M} X^T t$$

e) [5 marks] Describe a gradient-based optimization approach in pseudo-code. ***Hint:*** The gradient has been computed above; no need to repeat. Any variant (e.g., full-batch/mini-batch) is okay.

Loop until convergence:

$$w \leftarrow w - \alpha \cdot \frac{\partial J}{\partial w}$$

f) [10 marks] Alternatively, we may solve the problem by closed-form solution. Compute the closed-form solution for the ℓ_2-penalized MSE. Provide a few derivation steps.

Set $\frac{\partial J}{\partial w} = 0$

$$\left(\frac{1}{M} X^T X + 2\lambda I\right) w - \frac{1}{M} X^T t = 0$$

$$\left(X^T X + 2\lambda M I\right) w = X^T t$$

$$w = \left(X^T X + 2\lambda M I\right)^{-1} X^T t$$

g) [15 marks] Prove that minimizing ℓ_2-penalized MSE is equivalent to max *a posterior* (MAP) estimation. ***Hint:*** Please provide the general principle of MAP estimation (5 marks), appropriate probabilistic assumptions (5 marks), and derivation steps (5 marks).

For a univariate Gaussian distribution, $p(x) \propto \exp\{-\gamma x^2\}$ for some positive γ. Handling constants is not needed for this question.

General principle: Assume prior $p(\underline{w})$, likelihood $p(\mathcal{D}|w)$

MAP:

$$\hat{w}_{MAP} = \underset{w}{\operatorname{argmax}}\ p(w|\mathcal{D})$$

$$= \underset{w}{\operatorname{argmax}}\ p(w) \cdot p(\mathcal{D}|w)$$

$$= \underset{w}{\operatorname{argmax}}\ \log p(w) + \log p(\mathcal{D}|w)$$

For ℓ_2-penalized MSE assume

$$p(w_i) \propto \exp\{-\gamma_1 w_i^2\} \quad \text{for } i = 0, \cdots, d$$

$$p(t^{(m)} | x^{(m)}; w) \propto \exp\{-\gamma_2 (t^{(m)} - w^T x^{(m)})^2\}$$

MAP:

$$\underset{w}{\text{maximize}}\ \log \prod_{i=0}^{d} \exp\{-\gamma_1 w_i^2\} + \log \prod_{m=1}^{M} \exp\{-\gamma_2 (t^{(m)} - w^T x^{(m)})^2\}$$

$$\Longleftrightarrow \underset{w}{\text{maximize}}\ -\gamma_1 \sum_{i=0}^{d} w_i^2 - \gamma_2 \sum_{m=1}^{M} (t^{(m)} - w^T x^{(m)})^2$$

$$\Longleftrightarrow \underset{w}{\text{minimize}}\ \underbrace{\gamma_1 \|w\|_2^2}_{\ell_2\text{-penalty}} + \underbrace{\gamma_2 \|t^{(m)} - w^T x^{(m)}\|_2^2}_{\text{MSE}}$$

h) [5 marks] We know that the mean square error (without ℓ_2-penalty) yields an unbiased estimate. Why do we prefer ℓ_2-penalized MSE in some cases? ***Hint:*** One or a few sentences suffice.

The unbiased estimate assume $t = w^T x + \varepsilon$ for some unknown constant w.

But this assumption may not be true.

We need bias-variance tradeoff

i) [10 marks] Fill in the blanks with the word "overfitting" or "underfitting." No explanation needed.

	More overfitting or underfitting?
Increase λ	(a) Underfitting
Exclude some features (i.e., decrease d)	(b) underfitting

j) [5 marks] Consider labeled data $\mathcal{D}_{\text{label}}$ and test data $\mathcal{D}_{\text{test}}$. Explain why we cannot tune λ with $\mathcal{D}_{\text{test}}$. One or a few sentences suffice.

Because $\mathcal{D}_{test}$ mimicks the deployment, which could only be accessed once in practice.

Tuning λ on $\mathcal{D}_{test}$ leads to over-optimistic performance estimate.

k) [10 marks] Present a correct approach to tune the hyperparameter. Explain how to handle the datasets and provide pseudo code.

Split $\mathcal{D}_{label} = \mathcal{D}_{train} \cup \mathcal{D}_{val}$

For candidate λ.

Train $h^*_\lambda = \underset{h \in \mathcal{H}}{\operatorname{argmin}} J_\lambda(h^*, \mathcal{D}_{train})$

Validate by $Err(h^*_\lambda, \mathcal{D}_{val})$

Pick $\lambda^* = \underset{\lambda}{\operatorname{argmin}} Err(h^*_\lambda, \mathcal{D}_{val})$

Report test performance $Err(h^*_{\lambda^*}, \mathcal{D}_{test})$

END

Scrap paper. May be detached. Additional paper is available upon request as appropriate.
May be used as answer sheets if you

- Print your name and ID on every answer sheet (including additional sheets) submitted (1 bonus mark)
- Mark your solution and corresponding problem number clearly
- Submit the sheet by the end of the exam