— Solution notes —

First assignment in Machine learning 1 – 2024 – Paper 1

1 Multivariate Calculus (Recommended timeline: TBD)

In this exercise, you are going to compute several gradients. Simplify your answers as much as possible, *and use index-notation for all your derivations.* Consider $\nabla_{\boldsymbol{x}} f(\boldsymbol{x})$ as the same with $\frac{df}{d\boldsymbol{x}}$.

Compute the following:

(*a*) $\nabla_{\boldsymbol{x}} \boldsymbol{\sigma}(\boldsymbol{x})$ with $\boldsymbol{x} \in \mathbb{R}^m$ where $\boldsymbol{\sigma}$ denotes the Sigmoid function applied element-wise. [1 point]

Answer:

$\nabla_{\boldsymbol{x}} = \begin{bmatrix} \cdots & \frac{\partial \boldsymbol{\sigma}(\boldsymbol{x})_i}{\partial x_i} & \cdots \end{bmatrix} = \begin{bmatrix} \cdots & \sigma(x_i)(1-\sigma(x_i)) & \cdots \end{bmatrix} = [\boldsymbol{\sigma}(\boldsymbol{x}) \circ (\mathbf{1} - \boldsymbol{\sigma}(\boldsymbol{x}))]$ ($\circ$ denotes the Hadamard product). Using our conventions it should actually be $\text{diag}(\boldsymbol{\sigma}(\boldsymbol{x}) \circ (\mathbf{1} - \boldsymbol{\sigma}(\boldsymbol{x})))$

(*b*) $\frac{d}{d\boldsymbol{w}} \boldsymbol{f}$ with $\boldsymbol{f} = \boldsymbol{X}\boldsymbol{w}$ with $\boldsymbol{X} \in \mathbb{R}^{n \times n}$ and $\boldsymbol{w} \in \mathbb{R}^n$ [1 point]

Answer:

$$\boldsymbol{\nabla}_{\boldsymbol{w}} \boldsymbol{f} = \begin{bmatrix} \frac{\partial f_1}{\partial \boldsymbol{w}} & \cdots & \frac{\partial f_n}{\partial \boldsymbol{w}} \end{bmatrix}^T = \begin{bmatrix} \boldsymbol{x}_1 & \cdots & \boldsymbol{x}_n \end{bmatrix}^T = \boldsymbol{X}$$

(*c*) $\frac{d}{d\boldsymbol{w}} f$ with $f = \boldsymbol{w}^T \boldsymbol{X} \boldsymbol{w}$ with $\boldsymbol{X} \in \mathbb{R}^{n \times n}$ and $\boldsymbol{w} \in \mathbb{R}^n$ [1 point]

Answer:

$$\frac{d}{d\boldsymbol{w}} \boldsymbol{f} = [\cdots, \frac{\partial}{\partial w_k} \boldsymbol{f}, \cdots] = [\cdots, \sum_i w_i X_{ik} + \sum_j X_{kj} w_j, \cdots] = \boldsymbol{w}^T (X + X^T)$$

(*d*) $\frac{d}{d\boldsymbol{x}} \boldsymbol{\varsigma}(\boldsymbol{x}), \boldsymbol{x} \in \mathbb{R}^n$ where $\boldsymbol{\varsigma}(\boldsymbol{x})_i = \frac{\exp x_i}{\sum_{j=1}^n \exp x_j}$. Try to write it in matrix form making use of the diag function. [2 points]

Answer:

$$\nabla_{\boldsymbol{x}} \boldsymbol{\varsigma}(\boldsymbol{x}) = \begin{bmatrix} \cdots & \frac{\partial \boldsymbol{\varsigma}(\boldsymbol{x})_i}{\partial \boldsymbol{x}} & \cdots \end{bmatrix}^T$$

$$\frac{\partial \boldsymbol{\varsigma}(\boldsymbol{x})_i}{\partial \boldsymbol{x}} = \begin{bmatrix} \vdots \\ \frac{\varsigma(\boldsymbol{x})_i}{\partial x_i} \\ \vdots \\ \frac{\partial \boldsymbol{\varsigma}(\boldsymbol{x})_i}{\partial x_j} \\ \vdots \end{bmatrix} = \begin{bmatrix} \vdots \\ \boldsymbol{\varsigma}(\boldsymbol{x})_i (1 - \boldsymbol{\varsigma}(\boldsymbol{x})_i) \\ \vdots \\ \boldsymbol{\varsigma}(\boldsymbol{x})_i \cdot -\boldsymbol{\varsigma}(\boldsymbol{x})_j \\ \vdots \end{bmatrix} = \begin{bmatrix} \vdots \\ \boldsymbol{\varsigma}(\boldsymbol{x})_i (1 - \boldsymbol{\varsigma}(\boldsymbol{x})_i) \\ \vdots \\ \boldsymbol{\varsigma}(\boldsymbol{x})_i \cdot (0 - \boldsymbol{\varsigma}(\boldsymbol{x})_j) \\ \vdots \end{bmatrix}$$

Written other ways:

$$\frac{\partial \boldsymbol{\varsigma}(\boldsymbol{x})_i}{\partial \boldsymbol{x}_j} = \boldsymbol{\sigma}(\boldsymbol{x})_i (\delta - \boldsymbol{\sigma}(\boldsymbol{x})_j) = \boldsymbol{\sigma}(\boldsymbol{x})_i (I_{ij} - \boldsymbol{\sigma}(\boldsymbol{x})_j) = \boldsymbol{\sigma}(\boldsymbol{x})_i (\mathbb{I}(i=j) - \boldsymbol{\sigma}(\boldsymbol{x})_j)$$

where δ is the Kronecker delta, I the identity matrix and $\mathbb{I}$ the indicator function.

Therefore,

In matrix form:

$$\frac{\partial \boldsymbol{\varsigma}(\boldsymbol{x})}{\partial \boldsymbol{x}} = \text{diag}(\boldsymbol{\varsigma}(\boldsymbol{x})) - \boldsymbol{\varsigma}(\boldsymbol{x})\boldsymbol{\varsigma}(\boldsymbol{x})^T$$

(*e*) $\frac{d}{d\boldsymbol{\theta}}\|X\boldsymbol{\theta} - \boldsymbol{y}\|_2^2$, with $X \in \mathbb{R}^{n\times d}$, $\boldsymbol{\theta} \in \mathbb{R}^d$, and $\boldsymbol{y} \in \mathbb{R}^n$. Can you set this to zero and solve for $\boldsymbol{\theta}$? Congratulations, you derived the linear regression closed form on your own! [2 points]

Answer:

$$\begin{aligned}\frac{d}{d\boldsymbol{\theta}}\|X\boldsymbol{\theta} - \boldsymbol{y}\|_2^2 &= \frac{d}{d\boldsymbol{\theta}}(X\boldsymbol{\theta} - \boldsymbol{y})^T(X\boldsymbol{\theta} - \boldsymbol{y}) \\ &= \frac{d}{d\boldsymbol{\theta}}\boldsymbol{\theta}^T X^T X\boldsymbol{\theta} - 2\boldsymbol{\theta}^T X^T\boldsymbol{y} + \boldsymbol{y}^T\boldsymbol{y} \\ &= 2X^T X\boldsymbol{\theta} - 2X^T\boldsymbol{y}\end{aligned}$$

Solving for $\boldsymbol{\theta}$ is simple if we assume that $X^T X$ is invertible:

$$\boldsymbol{\theta} = (X^T X)^{-1}X^T\boldsymbol{y}$$

(*f*) **BONUS:** When does the formula that you just obtained fail? Explain how L2 regularization (i.e., adding the penalty term $\lambda\|\boldsymbol{\theta}\|_2^2$, where $\lambda > 0$ is a hyperparameter) will fix this issue. [1 point]

Answer: We cannot use the formula if $X^T X$ is not invertible.

If we add the penalty term $\lambda\|\boldsymbol{\theta}\|_2^2$, then our equation will become:

$$\begin{aligned}\frac{d}{d\boldsymbol{\theta}}\|X\boldsymbol{\theta} - \boldsymbol{y}\|_2^2 + \lambda\|\boldsymbol{\theta}\|_2^2 &= \frac{d}{d\boldsymbol{\theta}}(X\boldsymbol{\theta} - \boldsymbol{y})^T(X\boldsymbol{\theta} - \boldsymbol{y}) + \lambda\boldsymbol{\theta}^T\boldsymbol{\theta} \\ &= \frac{d}{d\boldsymbol{\theta}}\boldsymbol{\theta}^T X^T X\boldsymbol{\theta} - 2\boldsymbol{\theta}^T X^T\boldsymbol{y} + \boldsymbol{y}^T\boldsymbol{y} + \lambda\boldsymbol{\theta}^T\boldsymbol{\theta} \\ &= 2X^T X\boldsymbol{\theta} - 2X^T\boldsymbol{y} + 2\lambda\boldsymbol{\theta}\end{aligned}$$

Solving for $\boldsymbol{\theta}$ would mean

$$(X^T X + \lambda I)\boldsymbol{\theta} = X^T\boldsymbol{y}$$

All we have to do is to show that $(X^T X + \lambda I)$ is invertible, for every matrix $X \in \mathbb{R}^{n\times d}$.

For this, we can use the following results:

(*i*) If $\boldsymbol{A}$ is positive semi-definite and $\boldsymbol{B}$ is positive definite, then $\boldsymbol{C} = \boldsymbol{A} + \boldsymbol{B}$ is positive definite.
Proof: A is positive semi-definite $\iff \boldsymbol{v}^T A\boldsymbol{v} \geq 0, \forall \boldsymbol{v} \neq \boldsymbol{0}$.

B is positive definite $\iff \boldsymbol{v}^T B\boldsymbol{v} > 0, \forall \boldsymbol{v} \neq \boldsymbol{0}$.

Summing the two inequalities will result in:

$\boldsymbol{v}^T(A + B)\boldsymbol{v} > 0, \forall \boldsymbol{v} \neq \boldsymbol{0} \implies A + B$ is positive definite.

(*ii*) If $\boldsymbol{C}$ is positive definite, then $\boldsymbol{C}$ is invertible.
Proof: C is positive definite $\iff \boldsymbol{v}^T C\boldsymbol{v} > 0, \forall \boldsymbol{v} \neq \mathbf{0}$.

If we take $\boldsymbol{v}$ as an eigenvector of C, then the inequality will become $\boldsymbol{v}^T\lambda'\boldsymbol{v} > 0 \iff \lambda'\|\boldsymbol{v}\|_2^2 > 0$, where λ' is an eigenvalue of C.

This means that every eigenvalue λ' is non-zero, which means that C is invertible.

First, let's show that X^TX is positive semi-definite and λI is positive definite.

X^TX is positive semi-definite $\iff \boldsymbol{v}^TX^TX\boldsymbol{v} \geq 0, \forall \boldsymbol{v} \in \mathbb{R}^d, \boldsymbol{v} \neq \mathbf{0} \iff (X\boldsymbol{v})^T(X\boldsymbol{v}) \geq 0 \iff \|X\boldsymbol{v}\|_2^2 \geq 0$, which is true $\forall \boldsymbol{v} \in \mathbb{R}^d$ and $X \in \mathbb{R}^{n\times d}$.

λI is positive definite $\iff \boldsymbol{v}^T\lambda I\boldsymbol{v} > 0, \forall \boldsymbol{v} \in \mathbb{R}^d, \boldsymbol{v} \neq \mathbf{0} \iff \|\boldsymbol{v}\|_2^2 > 0$, which is true $\forall \boldsymbol{v} \in \mathbb{R}^d, \boldsymbol{v} \neq \mathbf{0}$.

Then, if we apply (i) and (ii), we obtain that $X^TX + \lambda I$ is invertible, $\forall X \in \mathbb{R}^{n\times d}$, meaning that:

$$\boldsymbol{\theta} = (X^TX + \lambda I)^{-1}X^T\boldsymbol{y}.$$

.

First assignment in Machine learning 1 – 2024 – Paper 1

2 Full analysis of a distribution: Exponential distribution (Recommended timeline: TBD)

The Poisson process is a model for series of discrete events where an average time between events is known, but the exact time at which they occur is random. It is also assumed that the process is memoryless or Markovian, i.e. the occurrence of a new event is independent of the previous events.

In this exercise, we are interested in analyzing the exponential distribution, which is a continuous probability distribution commonly used to model the time between events in a Poisson process. There are many processes that are modeled this way. For example, it is often used to estimate the lifespan of electronic devices, where failure rates do not change as the device ages. In manufacturing, the time between the completion of successive products on an assembly line can be modeled using an exponential distribution, when the production rate is constant. Even the amount of money customers spend in one trip to the supermarket follows an exponential distribution!

Formally, an exponential distribution with rate parameter $\lambda > 0$ can be modeled as such:

$$p(x) = \begin{cases} \lambda e^{-\lambda x}, & x \geq 0 \\ 0, & x < 0 \end{cases}$$

The aim of this exercise is to familiarize you with arbitrary distributions. Note that by no means the following questions are the only things you might want to know about a distribution, but rather serve as a starting point for further research. Using these insights, answer the following questions:

Useful results:

- The cumulative distribution function (cdf) of the exponential distribution is:

$$P(X \leq x) = \begin{cases} 1 - e^{-\lambda x}, & x \geq 0 \\ 0, & x < 0 \end{cases}$$

(*a*) If $X \sim \text{Exp}(\lambda)$, prove that $\mathbb{E}[X] = \frac{1}{\lambda}$. If you learned Calculus in English, remember "ultraviolet voodoo". If you speak Spanish, remember "un día vi una vaca vestida de uniforme". [2 points]

Answer: If $X \sim \text{Exp}(\lambda)$, then

$$\mathbb{E}[X] = \int_{-\infty}^{\infty} xp(x)dx = \int_{0}^{\infty} \lambda x e^{-\lambda x} dx = \int_{0}^{\infty} \lambda x (-\frac{e^{-\lambda x}}{\lambda})' dx$$

Applying Integration by Parts, we obtain

$$\mathbb{E}[X] = \int_0^\infty \lambda x(-\frac{e^{-\lambda x}}{\lambda})'dx$$
$$= \left[-xe^{-\lambda x}\right]_0^\infty + \int_0^\infty e^{-\lambda x}dx = \left[-xe^{-\lambda x}\right]_0^\infty + \int_0^\infty (-\frac{e^{-\lambda x}}{\lambda})'dx$$
$$= \left[-xe^{-\lambda x}\right]_0^\infty + \left[-\frac{e^{-\lambda x}}{\lambda}\right]_0^\infty = \lim_{x\to\infty}(-xe^{-\lambda x}) - 0 + \lim_{x\to\infty} -\frac{e^{-\lambda x}}{\lambda} - (-\frac{1}{\lambda})$$

The second limit is 0 by substituting directly x with ∞. For the first limit, however, we need to apply L'Hôpital's rule. Thus:

$$\lim_{x\to\infty}(-xe^{-\lambda x}) = \lim_{x\to\infty} -\frac{x}{e^{\lambda x}} = \lim_{x\to\infty} -\frac{1}{\lambda e^{\lambda x}} = 0$$

In the end, substituting these results in our equality, we obtain

$$\mathbb{E}[X] = \frac{1}{\lambda}$$

(*b*) Students arrive to the ML1 lecture in the morning according to an exponential distribution with an average time between arrivals of 2 minutes. What is the probability that the time between two consecutive arrivals is less than 1 minutes?

[1 point]

Answer: Using a) and the fact that the average time between arrivals is 2 minutes, we get that $\lambda = \frac{1}{2} = 0.5$.

The probability that the time between two consecutive arrivals is less than 1 minute can be determined using the cdf of the exponential distribution (i.e., $P(X \leq 1)$ with $\lambda = 0.5$).

Therefore, we obtain:

$$P(x \leq 1) = 1 - e^{-0.5} \simeq 0.3934$$

(*c*) Consider a dataset $\boldsymbol{x} = [x_1, x_2, \cdots, x_N]$ which are independent and identically distributed (i.i.d.) non-negative random variables from an exponential distribution with the rate parameter λ. Derive the log-likelihood function for the parameter λ given the dataset. [1 point]

Answer: Knowing that $x_i \geq 0, \forall 1 \leq i \leq N$, we can write the likelihood as follows:

$$p(\boldsymbol{x}|\lambda) = \prod_{i=1}^N p(x_i) = \prod_{i=1}^N \lambda e^{-\lambda x_i} = \lambda^N e^{-\lambda \sum_{i=1}^N x_i}$$

Next, we take the logarithm to obtain the log-likelihood:

$$\log p(\boldsymbol{x}|\lambda) = N \log \lambda - \lambda \sum_{i=1}^N x_i$$

(*d*) One common method to estimate the parameters of the assumed probability distribution is called maximum likelihood optimization. In this approach, we wish to find the parameters of the distribution (in our case λ) that will maximize the likelihood function. Find the maximum likelihood estimator λ_{ML} for the likelihood function calculated in part c). [2 points]

Answer: We begin by taking the derivative of the log-likelihood with respect to the model parameter λ and setting it to zero:

$$\begin{aligned}\frac{d \log p(\boldsymbol{x}|\lambda)}{d\lambda} &= \frac{d}{d\lambda}\left(N \log \lambda - \lambda \sum_{i=1}^{N} x_i\right) \\ &= \frac{N}{\lambda} - \sum_{i=1}^{N} x_i\end{aligned}$$

Setting this to zero gives:

$$\frac{N}{\lambda} - \sum_{i=1}^{N} x_i = 0 \Rightarrow \lambda = \frac{1}{N}\sum_{i=1}^{N} x_i$$

This result makes sense as λ is the mean of our distribution, so the ML solution for λ corresponds to the empirical mean.

(*e*) In general it is difficult to find a closed form expression for the posterior distribution of our model parameters because of the integral in the evidence. Rather than finding the full posterior distribution, we can find a point estimate of the model parameters. A common point estimate is the maximum a posteriori estimation, or the MAP, which estimates the model parameters as the mode of the posterior distribution, i.e.

$$\lambda_{\text{MAP}} = \arg\max_{\lambda} p(\lambda \mid \boldsymbol{x}).$$

Assume that the prior for the parameter λ is given by the Gamma distribution with hyperparameters α_1 and α_2:

$$p(\lambda|\alpha_1, \alpha_2) = \frac{\alpha_2^{\alpha_1}}{\Gamma(\alpha_1)}\lambda^{\alpha_1 - 1} e^{-\alpha_2 \lambda},$$

where Γ denotes the gamma function. Show that we can find λ_{MAP} by optimizing:

$$\lambda_{\text{MAP}} = \arg\max_{\lambda} \; (N + \alpha_1 - 1) \log \lambda - (\alpha_2 + \sum_{i=1}^{N} x_i)\lambda.$$

[2 points]

Hint: Show first that $\lambda_{\text{MAP}} = \arg\max_{\lambda} \log p(\boldsymbol{x} \mid \lambda) + \log p(\lambda)$.

Answer: The fact that $\lambda_{\text{MAP}} = \arg\max_\lambda \log p(\boldsymbol{x} \mid \lambda) + \log p(\lambda)$ follows directly from Bayes' rule.

We know from question *c*) that $\log p(\boldsymbol{x}|\lambda) = N \log \lambda - \lambda \sum_{i=1}^N x_i$. We observe that

$$\begin{aligned}
\lambda_{\text{MAP}} &= \arg\max_\lambda \ (\log p(\boldsymbol{x} \mid \lambda) + \log p(\lambda)) \\
&= \arg\max_\lambda \ \left(N \log \lambda - \lambda \sum_{i=1}^N x_i + \log(\frac{\alpha_2^{\alpha_1}}{\Gamma(\alpha_1)} \lambda^{\alpha_1 - 1} e^{-\alpha_2 \lambda}) \right) \\
&= \arg\max_\lambda \ \left(N \log \lambda - \lambda \sum_{i=1}^N x_i + (\alpha_1 - 1) \log \lambda - \alpha_2 \lambda \right) \\
&= \arg\max_\lambda \ \left((N + \alpha_1 - 1) \log \lambda - (\alpha_2 + \sum_{i=1}^N x_i) \lambda \right)
\end{aligned}$$

(*f*) Find the MAP estimator λ_{MAP}. [1 point]

Answer: Taking the derivative with respect to λ gives us

$$\frac{d}{d\lambda} \left((N + \alpha_1 - 1) \log \lambda - (\alpha_2 + \sum_{i=1}^N x_i) \lambda \right) = \frac{N + \alpha_1 - 1}{\lambda} - (\alpha_2 + \sum_{i=1}^N x_i).$$

By setting the derivative equal to zero, it follows that

$$\frac{N + \alpha_1 - 1}{\lambda} = \sum_{i=1}^N x_i + \alpha_2 \implies \lambda_{\text{MAP}} = \frac{N + \alpha_1 - 1}{\alpha_2 + \sum_{i=1}^N x_i}.$$

(*g*) In the case of a Exponential distribution with a Gamma prior, the resulting posterior distribution can be derived analytically. The resulting distribution is a Gamma distribution Gamma(α_1', α_2').

Show that the posterior distribution is indeed a Gamma distribution. [2 points]

Hint: The resulting distribution follows $\alpha_1' = N + \alpha_1$ and $\alpha_2' = (\sum_{i=1}^N x_i) + \alpha_2$.

Answer:

$$p(\lambda|\boldsymbol{x}) \propto \frac{\alpha_2^{\alpha_1}}{\Gamma(\alpha_1)} \lambda^{\alpha_1 - 1} e^{-\alpha_2 \lambda} \prod_{i=1}^N \lambda e^{-\lambda x_i}$$

$$p(\lambda|\boldsymbol{x}) \propto \frac{\alpha_2^{\alpha_1}}{\Gamma(\alpha_1)} \lambda^{\alpha_1 - 1} e^{-\alpha_2 \lambda} \lambda^N e^{-\lambda \sum_{i=1}^N x_i}$$

$$p(\lambda|\boldsymbol{x}) \propto \frac{\alpha_2^{\alpha_1}}{\Gamma(\alpha_1)} \lambda^{\alpha_1 - 1} e^{-\alpha_2 \lambda} \lambda^N e^{-\lambda \sum_{i=1}^N x_i}$$

$$p(\lambda|\boldsymbol{x}) \propto \lambda^{N + \alpha_1 - 1} e^{-\lambda(\alpha_2 + \sum_{i=1}^N x_i)}$$

$$p(\lambda|\boldsymbol{x}) \propto \lambda^{\alpha_1' - 1} e^{-\lambda \alpha_2'}$$

Any correct alternative solution will receive the maximum amount of points!

First assignment in Machine learning 1 – 2024 – Paper 1

3 General Multiple Outputs Linear Regression (Recommended timeline: TBD)

So far, all linear regression models assumed that the target t is a single target. In a more general case, however, we may wish to predict multiple targets $\mathbf{t}$.

One possibility is to perform an independent linear regression for each component of the target vector $\mathbf{t}$ by introducing a different set of basis functions for each component. In other words, if the target is a K-dimensional vector, we would perform a separate linear regression for each component t_i, $i \in 1, \ldots, K$ of the target vector $\mathbf{t}$.

The other more common approach is to use the same set of basis function to model the target vector directly in the following form:

$$\mathbf{y}(\mathbf{x}, \mathbf{W}) = \mathbf{W}^{\mathrm{T}} \boldsymbol{\phi}(\mathbf{x}),$$

where $\mathbf{y}$ is the model prediction, $\mathbf{x}$ is a M-dimensional input vector, and $\mathbf{W}$ is matrix of parameters. $\boldsymbol{\phi}(\mathbf{x})$ is an M-dimensional vector with elements $\phi_j(\mathbf{x})$, and $\phi_0(\mathbf{x}) = 1$ as usual. Assume that the conditional distribution of the target vector to be an Gaussian of the form:

$$p\left(\mathbf{t} \middle| \mathbf{W}, \boldsymbol{\Sigma}\right) = \mathcal{N}\left(\mathbf{t} \middle| \mathbf{y}(\mathbf{x}, \mathbf{W}), \boldsymbol{\Sigma}\right),$$

where $\boldsymbol{\Sigma}$ is the covariance matrix, and $\mathbf{y}(\mathbf{x}, \mathbf{W}) = \mathbf{W}^{\mathrm{T}} \boldsymbol{\phi}(\mathbf{x})$. Assume that we have N observations $\mathbf{t_1}, \ldots, \mathbf{t_N}$, which can be combined into a matrix $\mathbf{T}$ of size $N \times K$, such that the n^{th} row is given by $\mathbf{t}_n^{\mathrm{T}}$.

(a) What are the dimensions of the parameter matrix $\mathbf{W}$? [1 point]

Answer: Since the matrix transformation is supposed to map an M-dimensional input vector to a K-dimensional output, the matrix $\mathbf{W}$ has to have a shape $M \times K$.

(b) Write down the log-likehood. [1 point]

Answer: Assuming that we have $\mathbf{t_1}, \ldots, \mathbf{t_N}$ observations for the input vectors $\mathbf{x_1}, \ldots, \mathbf{x_N}$, we can first write the likelihood as follows:

$$\begin{aligned} p(\mathbf{T}|\mathbf{X}, \mathbf{W}, \boldsymbol{\Sigma}) &= \prod_{i=1}^{N} \mathcal{N}\left(\mathbf{t}_i \middle| \mathbf{y}(\mathbf{x}, \mathbf{W}), \boldsymbol{\Sigma}\right) \\ &= \prod_{i=1}^{N} \frac{1}{(2\pi)^{K/2} \left|\boldsymbol{\Sigma}\right|^{1/2}} \exp\left(-\frac{1}{2}(\mathbf{t}_i - \mathbf{W}^{\mathrm{T}}\phi(\mathbf{x_i}))^{\mathrm{T}} \boldsymbol{\Sigma}^{-1} (\mathbf{t}_i - \mathbf{W}^{\mathrm{T}}\phi(\mathbf{x}_i))\right) \end{aligned}$$

Next we take the logarithm:

$$\log p(\mathbf{T}|\mathbf{X},\mathbf{W},\mathbf{\Sigma}) = \sum_{i=1}^{N} \log \frac{1}{(2\pi)^{M/2}\,|\mathbf{\Sigma}|^{1/2}} \exp\left(-\frac{1}{2}(\mathbf{t}_i - \mathbf{W}^{\mathrm{T}}\phi(\mathbf{x_i}))^{\mathrm{T}}\mathbf{\Sigma}^{-1}(\mathbf{t}_i - \mathbf{W}^{\mathrm{T}}\phi(\mathbf{x}_i))\right)$$
$$= \frac{N}{2}\log|\mathbf{\Sigma}^{-1}| - \frac{1}{2}\sum_{i=1}^{N}(\mathbf{t}_i - \mathbf{W}^{\mathrm{T}}\phi(\mathbf{x_i}))^{\mathrm{T}}\mathbf{\Sigma}^{-1}(\mathbf{t}_i - \mathbf{W}^{\mathrm{T}}\phi(\mathbf{x}_i)) + \text{const.},$$

where we made use of the fact that matrix inverses and determinants commute, i.e.

$$|\mathbf{A}^{-1}| = \frac{1}{|\mathbf{A}|}.$$

(*c*) Find the maximum likelihood solution $\mathbf{W}_{\mathrm{ML}}$ in the terms of feature matrix $\mathbf{\Phi}$ and the target matrix $\mathbf{T}$, and show that it is independent of the covariance matrix $\mathbf{\Sigma}$.

Hint: Make use of the following derivation identity:

$$\frac{\partial}{\partial \mathbf{A}}(\mathbf{x} - \mathbf{A}\mathbf{s})^T\mathbf{W}(\mathbf{x} - \mathbf{A}\mathbf{s}) = -2\mathbf{W}(\mathbf{x} - \mathbf{A}\mathbf{s})\mathbf{s}^T$$

[2 points]

Answer: We begin by taking the derivative of the log-likelihood with respect to the matrix $\mathbf{W}$ and setting it to zero. Since the first two terms are constants w.r.t the matrix $\mathbf{W}$, they are equal to zero by default. The other terms can be easily calculated using known identities from vector calculus:

$$0 = -\sum_{i=1}^{N}\mathbf{\Sigma}^{-1}(\mathbf{t}_i - \mathbf{W}^{\mathrm{T}}\phi(\mathbf{x_i}))\phi(\mathbf{x_i})^{\mathrm{T}}$$

Next we multiply by $\mathbf{\Sigma}$ and replace the sum with to obtain the expression in terms of of feature matrix $\mathbf{\Phi}$ and the target matrix $\mathbf{T}$:

$$\mathbf{\Phi}^{\mathrm{T}}\mathbf{\Phi}\mathbf{W} = \mathbf{\Phi}^{\mathrm{T}}\mathbf{T}$$

Solving for $\mathbf{W}$ gives:

$$\mathbf{W}_{\mathrm{ML}} = (\mathbf{\Phi}^{\mathrm{T}}\mathbf{\Phi})^{-1}\mathbf{\Phi}^{\mathrm{T}}\mathbf{T}$$

We have obtained a very similar result to the case when we had only a single output, the only difference being that instead of a target vector $\mathbf{t}$ we now have a target matrix $\mathbf{T}$.

Index Notation solution

First, let $\mathbf{W}_k = \mathbf{W}_{[:,k]}$ be the k-th column of $\mathbf{W}$. Then, let $\boldsymbol{\phi}_i = \phi(\boldsymbol{x}_i)$ and $\Phi_{ij} = \phi_j(\boldsymbol{x}_i)$, i.e. $\boldsymbol{\phi}_i$ is the i-th row of the matrix Φ.

$$\log p(\mathbf{T}|\mathbf{X},\mathbf{W},\Sigma) = \frac{N}{2}\log|\Sigma^{-1}| - \frac{1}{2}\sum_{i}^{N}\sum_{k,l}^{K}\left((t_{ik}-\mathbf{W}_k^T\boldsymbol{\phi}_i)[\Sigma^{-1}]_{kl}(t_{il}-\mathbf{W}_l^T\boldsymbol{\phi}_i)\right) + \text{const.}$$

$$\begin{aligned}
\frac{\partial}{\partial W_{ab}}\log p(\mathbf{T}|\mathbf{X},\mathbf{W},\Sigma) &= -\frac{1}{2}\sum_{i}^{N}\sum_{k,l}^{K}\frac{\partial}{\partial W_{ab}}\left((t_{ik}-\mathbf{W}_k^T\boldsymbol{\phi}_i)[\Sigma^{-1}]_{kl}(t_{il}-\mathbf{W}_l^T\boldsymbol{\phi}_i)\right)\\
&= -\frac{1}{2}\sum_{i}^{N}\sum_{k,l}^{K}[\Sigma^{-1}]_{kl}(t_{il}-\mathbf{W}_l^T\boldsymbol{\phi}_i)\frac{\partial}{\partial W_{ab}}\left(t_{ik}-\mathbf{W}_k^T\boldsymbol{\phi}_i\right)\\
&\quad -\frac{1}{2}\sum_{i}^{N}\sum_{k,l}^{K}[\Sigma^{-1}]_{kl}(t_{ik}-\mathbf{W}_k^T\boldsymbol{\phi}_i)\frac{\partial}{\partial W_{ab}}\left(t_{il}-\mathbf{W}_l^T\boldsymbol{\phi}_i\right)\\
&= -\frac{1}{2}\sum_{i}^{N}\sum_{k,l}^{K}[\Sigma^{-1}]_{kl}\left(-(t_{il}-\mathbf{W}_l^T\boldsymbol{\phi}_i)\frac{\partial}{\partial W_{ab}}\left(\sum_c \mathbf{W}_{ck}\Phi_{ic}\right)\right)\\
&\quad -\frac{1}{2}\sum_{i}^{N}\sum_{k,l}^{K}[\Sigma^{-1}]_{kl}\left(-(t_{ik}-\mathbf{W}_k^T\boldsymbol{\phi}_i)\frac{\partial}{\partial W_{ab}}\left(\sum_d \mathbf{W}_{dl}\Phi_{id}\right)\right)\\
&= \frac{1}{2}\sum_{i}^{N}\sum_{k,l}^{K}[\Sigma^{-1}]_{kl}\left((t_{il}-\mathbf{W}_l^T\boldsymbol{\phi}_i)\left(\sum_c \delta_{ac}\delta_{bk}\Phi_{ic}\right)\right)\\
&\quad +\frac{1}{2}\sum_{i}^{N}\sum_{k,l}^{K}[\Sigma^{-1}]_{kl}\left((t_{ik}-\mathbf{W}_k^T\boldsymbol{\phi}_i)\left(\sum_d \delta_{ad}\delta_{bl}\Phi_{id}\right)\right)\\
&= \frac{1}{2}\sum_{i}^{N}\sum_{k,l}^{K}[\Sigma^{-1}]_{kl}\left((t_{il}-\mathbf{W}_l^T\boldsymbol{\phi}_i)\delta_{bk}\Phi_{ia} + (t_{ik}-\mathbf{W}_k^T\boldsymbol{\phi}_i)\delta_{bl}\Phi_{ia}\right)\\
&= \frac{1}{2}\sum_{i}^{N}\sum_{k,l}^{K}[\Sigma^{-1}]_{kl}(t_{il}-\mathbf{W}_l^T\boldsymbol{\phi}_i)\delta_{bk}\Phi_{ia} + \frac{1}{2}\sum_{i}^{N}\sum_{k,l}^{K}[\Sigma^{-1}]_{kl}(t_{ik}-\mathbf{W}_k^T\boldsymbol{\phi}_i)\delta_{bl}\Phi_{ia}\\
&= \frac{1}{2}\sum_{i}^{N}\sum_{l}^{K}[\Sigma^{-1}]_{bl}(t_{il}-\mathbf{W}_l^T\boldsymbol{\phi}_i)\Phi_{ia} + \frac{1}{2}\sum_{i}^{N}\sum_{k}^{K}[\Sigma^{-1}]_{kb}(t_{ik}-\mathbf{W}_k^T\boldsymbol{\phi}_i)\Phi_{ia}\\
&= \frac{1}{2}\sum_{i}^{N}\sum_{k}^{K}\left([\Sigma^{-1}]_{bk}+[\Sigma^{-1}]_{kb}\right)\left[t_{ik}-\mathbf{W}_k^T\boldsymbol{\phi}_i\right]\Phi_{ia}\\
&= \frac{1}{2}\sum_{i}^{N}\sum_{k}^{K}\left([\Sigma^{-1}]_{bk}+[\Sigma^{-1}]_{kb}\right)\left[\boldsymbol{t}_i-\mathbf{W}^T\boldsymbol{\phi}_i\right]_k\Phi_{ia}
\end{aligned}$$

by using the fact the convariance matrix is symmetric:

$$\begin{aligned}
&= \sum_{i}^{N}\sum_{k}^{K}[\Sigma^{-1}]_{bk}[\boldsymbol{t}_i-\mathbf{W}^T\boldsymbol{\phi}_i]_k\Phi_{ia}\\
&= \sum_{i}^{N}[\Sigma^{-1}]_{b,:}\left(\boldsymbol{t}_i-\mathbf{W}^T\boldsymbol{\phi}_i\right)\Phi_{ia}\\
&= \sum_{i}^{N}[\Sigma^{-1}\left(\boldsymbol{t}_i-\mathbf{W}^T\boldsymbol{\phi}_i\right)\boldsymbol{\phi}_i^T]_{ba}\\
\frac{\partial}{\partial W}\log p(\mathbf{T}|\mathbf{X},\mathbf{W},\Sigma) &= \sum_{i}^{N}\left(\Sigma^{-1}\left(\boldsymbol{t}_i-\mathbf{W}^T\boldsymbol{\phi}_i\right)\boldsymbol{\phi}_i^T\right)^T
\end{aligned}$$

∎

To find the arg-max, we set the derivative to zero and solve the resulting equation:

$$\begin{aligned}
\frac{\partial}{\partial W} \log p(\mathbf{T}|\mathbf{X}, \mathbf{W}, \Sigma) &= 0 \\
\sum_i^N \left(\Sigma^{-1} \left(\boldsymbol{t}_i - \mathbf{W}^T \boldsymbol{\phi}_i\right) \boldsymbol{\phi}_i^T\right)^T &= 0 \\
\sum_i^N \Sigma^{-1} \left(\boldsymbol{t}_i - \mathbf{W}^T \boldsymbol{\phi}_i\right) \boldsymbol{\phi}_i^T &= 0
\end{aligned}$$

which resembles the equation found using the matrix identities.

(d) Show that the maximum likelihood solution for $\mathbf{\Sigma}$ is given by:

$$\mathbf{\Sigma} = \frac{1}{N} \sum_{n=1}^{N} \left(\mathbf{t}_n - \mathbf{W}_{\mathrm{ML}}^{\mathrm{T}} \phi(\mathbf{x}_n)\right) \left(\mathbf{t}_n - \mathbf{W}_{\mathrm{ML}}^{\mathrm{T}} \phi(\mathbf{x}_n)\right)^{\mathrm{T}}$$

[2 points]

Substitute $\mathbf{\Omega} := \mathbf{\Sigma}^{-1}$ and differentiate with respect to $\mathbf{\Omega}$. Moreover, make use of the identities $\frac{d}{d\mathbf{A}} \log |\mathbf{A}| = (\mathbf{A^{-1}})^{\mathbf{T}}$ and $\frac{d}{d\mathbf{X}} \mathbf{a^T X^T b} = \mathbf{b a^T}$.

Answer: Using the identities, we observe that

$$\begin{aligned}
\frac{d}{d\mathbf{\Omega}} \log p(\mathbf{T}|\mathbf{X}, \mathbf{W}, \mathbf{\Omega}) &= \frac{d}{d\mathbf{\Omega}} \frac{N}{2} \log |\Omega| - \frac{1}{2} \sum_{i=1}^{N} (\mathbf{t}_i - \mathbf{W}^{\mathrm{T}} \phi(\mathbf{x_i}))^{\mathrm{T}} \Omega (\mathbf{t}_i - \mathbf{W}^{\mathrm{T}} \phi(\mathbf{x}_i)) \\
&= \frac{N}{2} (\mathbf{\Omega^{-1}})^{\mathbf{T}} - \frac{1}{2} \sum_{i=1}^{N} (\mathbf{t}_i - \mathbf{W}^{\mathrm{T}} \phi(\mathbf{x_i}))(\mathbf{t}_i - \mathbf{W}^{\mathrm{T}} \phi(\mathbf{x}_i))^{\mathrm{T}}.
\end{aligned}$$

Setting this derivative equal to zero and substituting back in $\mathbf{\Sigma}$ gives us

$$\frac{N}{2} \mathbf{\Sigma^T} = \frac{1}{2} \sum_{i=1}^{N} (\mathbf{t}_i - \mathbf{W}^{\mathrm{T}} \phi(\mathbf{x_i}))(\mathbf{t}_i - \mathbf{W}^{\mathrm{T}} \phi(\mathbf{x}_i))^{\mathrm{T}},$$

and thus

$$\mathbf{\Sigma}_{\mathrm{MLE}} = \frac{1}{N} \sum_{i=1}^{N} (\mathbf{t}_i - \mathbf{W}^{\mathrm{T}} \phi(\mathbf{x_i}))(\mathbf{t}_i - \mathbf{W}^{\mathrm{T}} \phi(\mathbf{x}_i))^{\mathrm{T}}.$$

First assignment in Machine learning 1 – 2024 – Paper 1

4 Counting Fish (Recommended timeline: TBD)

Imagine you are in an aquarium, and you notice that each fish has a tag with a number attached in a fin. You see number 3, then 6, then 3 again, and finally 9. After asking an employee, they tell you that each fish has a consecutive number assigned, and that the sequence starts at one.

(*a*) Intuitively, how many fishes do you think there are after observing these four samples? [1 point]

Answer: There must be at least 9 fishes. Any number smaller or equal to 8 is wrong. Any number greater or equal to 9 is okay.

(*b*) What is the most likely estimate? State your assumptions clearly. [1 point]

Answer: Since we saw a fish with the tag 3 twice, we know that we are taking samples with replacement (so we can assume independence). Thus, it is reasonable to assume that each sample X_i follows a **discrete** uniform distribution between 1 and M.

$$\begin{aligned} P(X_1 = x1, \dots, X_n = x_n) &= P(X_1 = x_1) \dots P(X_n = x_n) \\ &= [x_1 \leq M] \frac{1}{M} \dots [x_n \leq M] \frac{1}{M} \\ &= [x_{\max} \leq M] \frac{1}{M^n} \end{aligned}$$

This number is going to be zero if $M < x_{\max}$. If $M \geq x_{\max}$, we get a curve that decreases with M. Thus, the maximum is attained when $M = x_{\max}$. No need to take any derivatives, but the students should plot the likelihood or argue as above.

(*c*) Is this estimation unbiased, or does it overestimate / underestimate the true number of fishes? If the estimation is biased, compute the bias. [2 points]

Answer: This estimator always **underestimates** the true value, that will always be greater of equal than the biggest sample. Let us prove it formally.

The estimator is $X_{\max}$, so we need to take its expected value. In order to do so, we need the probability mass function (pmf), but it is easier to start with the cumulative distribution function (cdf).

$$\begin{aligned} P(X_{\max} \leq x) &= P(X_1 \leq x, \dots, X_n \leq x) \\ &= P(X_1 \leq x) \dots P(X_n \leq x) \\ &= \left(\frac{x}{M}\right)^n [1 \leq x] \end{aligned}$$

Then,

$$\begin{aligned} P(X_{\max} = x) &= P(X_{\max} \leq x) - P(X_{\max} \leq x - 1) \\ &= \frac{x^n - (x-1)^n}{M^n} [1 \leq x] \end{aligned}$$

We are now ready to compute the expected value of our estimator:

$$\begin{aligned}
\mathbb{E}[X_{\max}] &= \sum_{x=1}^{M} xP(X_{\max} = x) \\
&= \frac{1}{M^n} \sum_{x=1}^{M} x^{n+1} - x(x-1)^n \\
&= \frac{M^{n+1} - (1^n + 2^n + \cdots + (M-1)^n)}{M^n} \\
&= M - \frac{1^n + 2^n + \cdots + (M-1)^n}{M^n}
\end{aligned}$$

In order to get the third equality, we exploit that the sum is (almost) telescopic. Since the second term (bias) is negative, we can conclude that the estimator underestimates the true value.

Fun fact: the British used a similar technique to estimate how many tanks Nazis were building per month. This estimation was significantly better than the figures that the British spies were reporting.