CS181 Practice Questions: Linear Regression, Continued

1. **Posterior Weight Distribution Using Bayes' Rule for Linear Gaussian Systems**

Some background: In section (2.3.3), Bishop derives the following facts about linear Gaussian systems: assuming we have a marginal distribution on $\boldsymbol{x}$ and a conditional distribution on $\boldsymbol{y}$ given by

$$p(\boldsymbol{x}) = \mathcal{N}(\boldsymbol{x} \mid \boldsymbol{\mu}, \boldsymbol{\Lambda}^{-1}) \tag{2.113}$$

$$p(\boldsymbol{y} \mid \boldsymbol{x}) = \mathcal{N}(\boldsymbol{y} \mid \boldsymbol{A}\boldsymbol{x} + \boldsymbol{b}, \boldsymbol{L}^{-1}) \tag{2.114}$$

then

$$p(\boldsymbol{x} \mid \boldsymbol{y}) = \mathcal{N}(\boldsymbol{x} \mid \boldsymbol{\Sigma}(\boldsymbol{A}^\top \boldsymbol{L}(\boldsymbol{y} - \boldsymbol{b}) + \boldsymbol{\Lambda}\boldsymbol{\mu}), \boldsymbol{\Sigma}) \tag{2.116}$$

where

$$\boldsymbol{\Sigma} = (\boldsymbol{\Lambda} + \boldsymbol{A}^\top \boldsymbol{L}\boldsymbol{A})^{-1}. \tag{2.117}$$

Now, we know from (3.10) in Bishop that the regression likelihood can be written as

$$p(\boldsymbol{t} \mid \boldsymbol{w}) = \prod_{n=1}^{N} \mathcal{N}(t_n \mid \boldsymbol{w}^\top \phi(\boldsymbol{x}_n), \beta^{-1}),$$

where $\beta = \frac{1}{\sigma^2}$. If the prior distribution on $\boldsymbol{w}$ is given by $p(\boldsymbol{w}) = \mathcal{N}(\boldsymbol{w} \mid \boldsymbol{m}_0, \boldsymbol{S}_0)$, derive that the posterior distribution $p(\boldsymbol{w} \mid \boldsymbol{t})$ is given by

$$p(\boldsymbol{w} \mid \boldsymbol{t}) = \mathcal{N}(\boldsymbol{w} \mid \boldsymbol{m}_N, \boldsymbol{S}_N)$$

where

$$\begin{aligned} \boldsymbol{m}_N &= \boldsymbol{S}_N(\boldsymbol{S}_0^{-1}\boldsymbol{m}_0 + \beta\boldsymbol{\Phi}^\top \boldsymbol{t}) \\ \boldsymbol{S}_N^{-1} &= \boldsymbol{S}_0^{-1} + \beta\boldsymbol{\Phi}^\top\boldsymbol{\Phi} \end{aligned}$$

in the following way:

(a) Write down the likelihood in the form of a multivariate Gaussian.

(b) Explain what we need to substitute for $\boldsymbol{x}, \boldsymbol{y}, \boldsymbol{\mu}, \boldsymbol{\Lambda}, \boldsymbol{A}, \boldsymbol{b}, \boldsymbol{L}$ (respectively) in equations (2.113)-(2.117) to derive the posterior.

(a) The likelihood can be written $p(\boldsymbol{t} \mid \boldsymbol{w}) = \mathcal{N}(\boldsymbol{t} | \boldsymbol{\Phi}\boldsymbol{w}, \beta^{-1}\boldsymbol{I})$.

(b) Substituting $\boldsymbol{x} = \boldsymbol{w}, \boldsymbol{y} = \boldsymbol{t}, \boldsymbol{\mu} = \boldsymbol{m}_0, \boldsymbol{\Lambda} = \boldsymbol{S}_0^{-1}, \boldsymbol{A} = \boldsymbol{\Phi}, \boldsymbol{b} = \boldsymbol{0}, \boldsymbol{L} = \beta \boldsymbol{I}$, then using (2.117) we first get that $\boldsymbol{\Sigma} = \boldsymbol{S}_N = (\boldsymbol{S}_0^{-1} + \beta \boldsymbol{\Phi}^\top \boldsymbol{\Phi})^{-1}$, and plugging into (2.116) we get that $\boldsymbol{m}_N = \boldsymbol{S}_N(\beta \boldsymbol{\Phi}^\top \boldsymbol{I}(\boldsymbol{t} - \boldsymbol{0}) + \boldsymbol{S}_0^{-1}\boldsymbol{m}_0) = \boldsymbol{S}_N(\boldsymbol{S}_0^{-1}\boldsymbol{m}_0 + \beta \boldsymbol{\Phi}^\top \boldsymbol{t})$.

2. **Posterior Weight Distribution By Completing the Square (Bishop 3.7)**

We know from (3.10) in Bishop that the regression likelihood can be written as

$$\begin{aligned} p(\boldsymbol{t} \mid \boldsymbol{w}) &= \prod_{n=1}^{N} \mathcal{N}(t_n \mid \boldsymbol{w}^\top \phi(\boldsymbol{x}_n), \beta^{-1}) \\ &\propto \exp\left(-\frac{\beta}{2}(\boldsymbol{t} - \boldsymbol{\Phi}\boldsymbol{w})^\top(\boldsymbol{t} - \boldsymbol{\Phi}\boldsymbol{w})\right), \end{aligned}$$

where $\beta = \frac{1}{\sigma^2}$ and in the second line above we have ignored the Gaussian normalizing constants. By completing the square, show that with a prior distribution on $\boldsymbol{w}$ given by $p(\boldsymbol{w}) = \mathcal{N}(\boldsymbol{w} \mid \boldsymbol{m}_0, \boldsymbol{S}_0)$, the posterior distribution $p(\boldsymbol{w} \mid \boldsymbol{t})$ is given by

$$p(\boldsymbol{w} \mid \boldsymbol{t}) = \mathcal{N}(\boldsymbol{w} \mid \boldsymbol{m}_N, \boldsymbol{S}_N)$$

where

$$\begin{aligned} \boldsymbol{m}_N &= \boldsymbol{S}_N(\boldsymbol{S}_0^{-1}\boldsymbol{m}_0 + \beta\boldsymbol{\Phi}^\top \boldsymbol{t}) \\ \boldsymbol{S}_N^{-1} &= \boldsymbol{S}_0^{-1} + \beta\boldsymbol{\Phi}^\top\boldsymbol{\Phi} \end{aligned}$$

Multiplying the likelihood and prior, we get

$$\begin{aligned} p(\boldsymbol{w} \mid \boldsymbol{t}) &\propto p(\boldsymbol{t} \mid \boldsymbol{w}) p(\boldsymbol{w}) \\ &\propto \exp\left(-\frac{\beta}{2}(\boldsymbol{t} - \boldsymbol{\Phi}\boldsymbol{w})^\top(\boldsymbol{t} - \boldsymbol{\Phi}\boldsymbol{w})\right) \exp\left(-\frac{1}{2}(\boldsymbol{w} - \boldsymbol{m}_0)^\top \boldsymbol{S}_0^{-1}(\boldsymbol{w} - \boldsymbol{m}_0)\right) && (1) \\ &= \exp\Big(-\frac{1}{2}\boldsymbol{w}^\top(\boldsymbol{S}_0^{-1} + \beta\boldsymbol{\Phi}^\top\boldsymbol{\Phi})\boldsymbol{w} - \beta\boldsymbol{t}^\top\boldsymbol{\Phi}\boldsymbol{w} - \beta\boldsymbol{w}^\top\boldsymbol{\Phi}^\top\boldsymbol{t} + \beta\boldsymbol{t}^\top\boldsymbol{t} - \boldsymbol{m}_0^\top\boldsymbol{S}_0^{-1}\boldsymbol{w} && (2) \\ &\qquad - \boldsymbol{w}^\top\boldsymbol{S}_0^{-1}\boldsymbol{m}_0 + \boldsymbol{m}_0^\top\boldsymbol{S}_0^{-1}\boldsymbol{m}_0\Big) \\ &= \exp\Big(-\frac{1}{2}\boldsymbol{w}^\top(\boldsymbol{S}_0^{-1} + \beta\boldsymbol{\Phi}^\top\boldsymbol{\Phi})\boldsymbol{w} - (\boldsymbol{S}_0^{-1}\boldsymbol{m}_0 + \beta\boldsymbol{\Phi}^\top\boldsymbol{t})^\top\boldsymbol{w} - \boldsymbol{w}^\top(\boldsymbol{S}_0^{-1}\boldsymbol{m}_0 + \beta\boldsymbol{\Phi}^\top\boldsymbol{t}) && (3) \\ &\qquad + \beta\boldsymbol{t}^\top\boldsymbol{t} + \boldsymbol{m}_0^\top\boldsymbol{S}_0^{-1}\boldsymbol{m}_0\Big) \\ &= \exp\left(-\frac{1}{2}(\boldsymbol{w} - \boldsymbol{m}_N)^\top\boldsymbol{S}_N^{-1}(\boldsymbol{w} - \boldsymbol{m}_N)\right) \exp\left(-\frac{1}{2}(\beta\boldsymbol{t}^\top\boldsymbol{t} + \boldsymbol{m}_0^\top\boldsymbol{S}_0^{-1}\boldsymbol{m}_0 - \boldsymbol{m}_N^\top\boldsymbol{S}_N^{-1}\boldsymbol{m}_N)\right), && (4) \end{aligned}$$

where the first exponential in (4) has the form of the desired (unnormalized) Gaussian, and the second exponential in (4) doesn't involve $\boldsymbol{w}$, and so can be absorbed into the Gaussian's normalization constant.

To derive (4) from (3), note that plugging in the definitions of $\boldsymbol{m}_N$ and $\boldsymbol{S}_N^{-1}$ into the first exponential in (4) and multiplying through gives us all the terms in (3) involving $\boldsymbol{w}$, except with an additional $\boldsymbol{m}_N^\top \boldsymbol{S}_N^{-1} \boldsymbol{m}_N$ term. We subtract this off in the second exponential, and also collect there the other terms not involving $\boldsymbol{w}$, which are absorbed into the normalization.

3. **Predictive Distribution**

Bishop notes in section (3.2.2) that if our prior distribution on $\boldsymbol{w}$ is

$$p(\boldsymbol{w}) = \mathcal{N}(\boldsymbol{w} \,|\, \mathbf{0}, \alpha^{-1}\boldsymbol{I}),$$

and if we assume again that our likelihood involves an inverse variance parameter β, then the predictive distribution of t for a new datapoint $\boldsymbol{x}$ is given by

$$p(t \,|\, \boldsymbol{t}, \alpha, \beta) = \int p(t \,|\, \boldsymbol{w}, \beta) p(\boldsymbol{w} \,|\, \boldsymbol{t}, \alpha, \beta) \, d\boldsymbol{w} \tag{3.57}$$

Does the equation in (3.57) make any independence assumptions about the variables involved? If so, which?

Yes, it assumes that t is independent of $\boldsymbol{t}$ (and α) given $\boldsymbol{w}$. This can be seen, for instance, by writing out $p(t, \boldsymbol{w} \,|\, \boldsymbol{t}, \alpha, \beta)$ and using the chain rule of probability.

4. **Deriving Lasso Regularization with Lagrange Multipliers**

Show that minimization of the unregularized sum-of-squares error function given by

$$E_D(\boldsymbol{w}) = \frac{1}{2}\sum_{n=1}^{N}(t_n - \boldsymbol{w}^{\mathsf{T}}\phi(\boldsymbol{x}_n))^2,$$

subject to the constraint

$$\sum_{j=1}^{M}|w_j| \le \eta,$$

is equivalent to minimizing the regularized error function

$$\frac{1}{2}\sum_{n=1}^{N}(t_n - \boldsymbol{w}^{\mathsf{T}}\phi(\boldsymbol{x}_n))^2 + \frac{\lambda}{2}\sum_{j=1}^{M}|w_j|$$

Rewrite the constraint as

$$\sum_{j=1}^{M}|w_j| - \eta \le 0$$

We get the Lagrangian function

$$L(\boldsymbol{w}, \lambda) = \frac{1}{2}\sum_{n=1}^{N}(t_n - \boldsymbol{w}^{\mathsf{T}}\phi(\boldsymbol{x}_n))^2 + \frac{\lambda}{2}\sum_{j=1}^{M}(|w_j| - \eta)$$

where we introduce the factor of $1/2$ in front of the second term for convenience. We see immediately that the above function is equal to the regularized error function plus the terms of η which do not depend on $\boldsymbol{w}$. Therefore, minimizing the Lagrangian with respect to $\boldsymbol{w}$ will give the same $\boldsymbol{w}^*$ as minimizing the regularized error function.

5. **Connection between Priors and Regularization**

Consider the Bayesian linear regression model given in Bishop 3.3.1. The prior is given by

$$p(\boldsymbol{w}|\alpha) = \mathcal{N}(\boldsymbol{w}|\mathbf{0}, \alpha^{-1}\boldsymbol{I}),$$

where α is the precision parameter that controls the variance of the Gaussian prior. The likelihood can be written as

$$p(\boldsymbol{t} \mid \boldsymbol{w}) = \prod_{n=1}^{N} \mathcal{N}(t_n \mid \boldsymbol{w}^\mathsf{T}\phi(\boldsymbol{x}_n), \beta^{-1}),$$

Using the fact that the posterior is the product of the prior and the likelihood, show that maximizing the log posterior (i.e. $\ln p(\boldsymbol{w} \mid \boldsymbol{t}) = \ln p(\boldsymbol{w}|\alpha) + \ln p(\boldsymbol{t} \mid \boldsymbol{w})$) is equivalent to minimizing the regularized error term given by $E_D(\boldsymbol{w}) + \lambda E_W(\boldsymbol{w})$ with

$$\begin{aligned} E_D(\boldsymbol{w}) &= \frac{1}{2}\sum_{n=1}^{N}(t_n - \boldsymbol{w}^\mathsf{T}\phi(\boldsymbol{x}_n))^2 \\ E_W(\boldsymbol{w}) &= \frac{\lambda}{2}\boldsymbol{w}^\mathsf{T}\boldsymbol{w} \end{aligned}$$

Do this by writing $\ln p(\boldsymbol{w} \mid \boldsymbol{t})$ as a function of $E_D(\boldsymbol{w})$ and $E_W(\boldsymbol{w})$, dropping constant terms if necessary.
Conclude that maximizing this posterior is equivalent to minimizing the regularized error term given by $E_D(\boldsymbol{w}) + \lambda E_W(\boldsymbol{w})$.
(Hint: take $\lambda = \alpha/\beta$)

Expanding the posterior gives

$$\begin{aligned} p(\boldsymbol{w} \mid \boldsymbol{t}) &= p(\boldsymbol{w}|\alpha) + p(\boldsymbol{t} \mid \boldsymbol{w}) \implies \\ \ln p(\boldsymbol{w} \mid \boldsymbol{t}) &= \ln p(\boldsymbol{w}|\alpha) + \ln p(\boldsymbol{t} \mid \boldsymbol{w}) \\ &= \sum_{n=1}^{N} \ln \mathcal{N}(\boldsymbol{t}_n \mid \boldsymbol{w}^\mathsf{T}\phi(\boldsymbol{x}_n), \beta^{-1}) + \ln \mathcal{N}(\boldsymbol{w} \mid \mathbf{0}, \alpha^{-1}\boldsymbol{I}) \\ &= \sum_{n=1}^{N} -\frac{\beta}{2}(t_n - \boldsymbol{w}^\mathsf{T}\phi(\boldsymbol{x}_n))^2 + const - \frac{1}{2}\alpha\boldsymbol{w}^\mathsf{T}\boldsymbol{w} + const \\ &= -\beta E_D(\boldsymbol{w}) - \alpha E_W(\boldsymbol{w}) \end{aligned}$$

Taking $\lambda = \alpha/\beta$, note that minimizing this function is equivalent to maximizing its negative, and that multiplying by a constant positive factor ($1/\beta$) has no effect on the extrema, and we have the result.

6. **Bayesian Updates in Linear Regression, Bishop 3.8**

Suppose we have the standard Bayesian linear regression model and we have already observed N data points, so the posterior distribution is the same as the one derived in problem 2,

$$p(\boldsymbol{w} \mid \boldsymbol{t}) = \mathcal{N}(\boldsymbol{w} \mid \boldsymbol{m}_N, \boldsymbol{S}_N)$$

where

$$\begin{aligned}\boldsymbol{m}_N &= \boldsymbol{S}_N(\boldsymbol{S}_0^{-1}\boldsymbol{m}_0 + \beta\boldsymbol{\Phi}^\mathsf{T}\boldsymbol{t}) \\ \boldsymbol{S}_N^{-1} &= \boldsymbol{S}_0^{-1} + \beta\boldsymbol{\Phi}^\mathsf{T}\boldsymbol{\Phi}\end{aligned}$$

Suppose we observe a new data point $(\boldsymbol{x}_{N+1}, t_{N+1})$. Show that the resulting posterior distribution is of the same form with $\boldsymbol{m}_{N+1}$ and $\boldsymbol{S}_{N+1}$.

We proceed by completing the square, as in problem 2.

The prior is of the form

$$p(\boldsymbol{w}) = \mathcal{N}(\boldsymbol{w} \mid \boldsymbol{m}_N, \boldsymbol{S}_N),$$

and the likelihood is given by

$$p(t_{N+1} \mid \boldsymbol{x}_{N+1}, \boldsymbol{w}) = (\beta/2\pi)^{1/2} \exp\left(-\beta/2(t_{N+1} - \boldsymbol{w}^\mathsf{T}\phi_{N+1})^2\right)$$

where $\phi_{N+1} = \phi(\boldsymbol{x}_{N+1})$. Then the posterior is

$$p(\boldsymbol{w} \mid \boldsymbol{t}) \propto \exp\left(-\frac{1}{2}(\boldsymbol{w} - \boldsymbol{m}_N)^\mathsf{T}\boldsymbol{S}_N^{-1}(\boldsymbol{w} - \boldsymbol{m}_N) - \frac{1}{2}\beta(t_{N+1} - \boldsymbol{w}^\mathsf{T}\phi_{N+1})^2\right)$$

Completing the square in the exponential just as we did in problem 2, we get

$$\begin{aligned}&(\boldsymbol{w} - \boldsymbol{m}_N)^\mathsf{T}\boldsymbol{S}_N^{-1}(\boldsymbol{w} - \boldsymbol{m}_N) + \beta(t_{N+1} - \boldsymbol{w}^\mathsf{T}\phi_{N+1})^2 \\ &= \boldsymbol{w}^\mathsf{T}\boldsymbol{S}_N^{-1}\boldsymbol{w} - 2\boldsymbol{w}^\mathsf{T}\boldsymbol{S}_N^{-1}\boldsymbol{m}_N + \beta\boldsymbol{w}^\mathsf{T}\phi_{N+1}^\mathsf{T}\phi_{N+1}\boldsymbol{w} - 2\beta\boldsymbol{w}^\mathsf{T}\phi_{N+1}t_{N+1} + \mathit{const} \\ &= \boldsymbol{w}^\mathsf{T}(\boldsymbol{S}_N^{-1} + \beta\phi_{N+1}\phi_{N+1}^\mathsf{T})\boldsymbol{w} - 2\boldsymbol{w}^\mathsf{T}(\boldsymbol{S}_N^{-1}\boldsymbol{m}_N + \beta\phi_{N+1}t_{N+1}) + \mathit{const}\end{aligned}$$

Plugging this back in to the expression for the posterior we get

$$p(\boldsymbol{w} \mid \boldsymbol{t}) = \mathcal{N}(\boldsymbol{w} \mid \boldsymbol{m}_{N+1}, \boldsymbol{S}_{N+1})$$

where

$$\begin{aligned}\boldsymbol{m}_{N+1} &= \boldsymbol{S}_{N+1}(\boldsymbol{S}_N^{-1}\boldsymbol{m}_N + \beta\phi_{N+1}\boldsymbol{t}_{N+1}) \\ \boldsymbol{S}_{N+1}^{-1} &= \boldsymbol{S}_N^{-1} + \beta\phi_{N+1}\phi_{N+1}^\mathsf{T}\end{aligned}$$