Statistical Modeling and Inference – Problem Set #3

Niti Mishra · Miquel Torrens · Bálint Ván

October 26th, 2015

Solution to proposed exercises.

Exercise 1

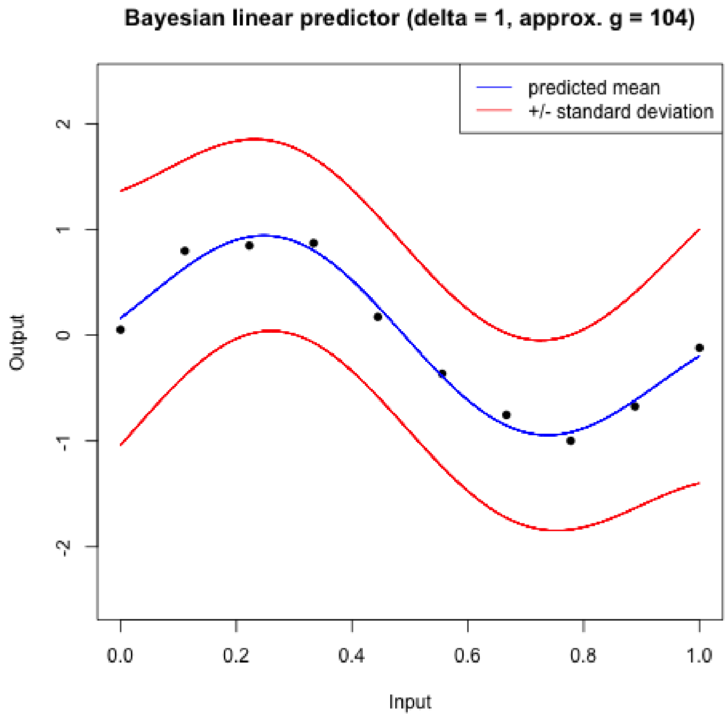
Part 1

The exercise requires two extra parameters not defined, which are δ and g.

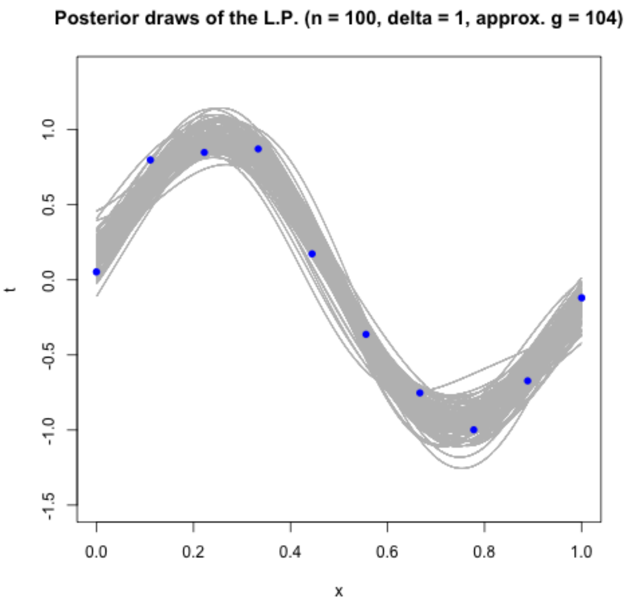
We keep δ as the degree of freedom. We will try to find g in the following manner:

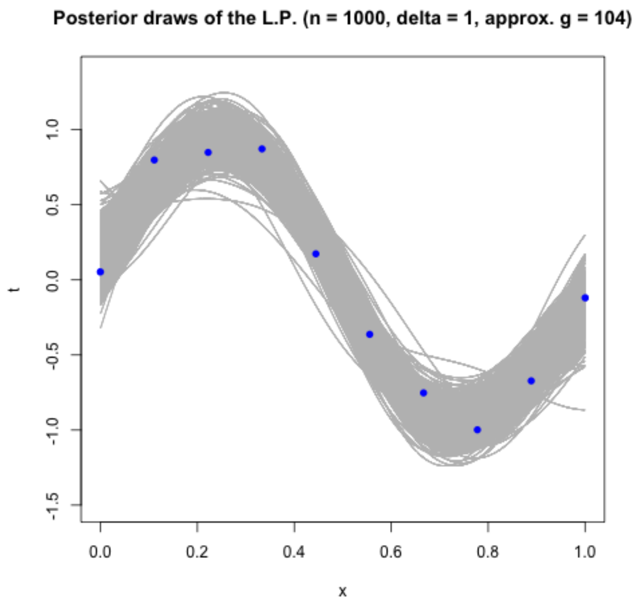
- To start, we compute the sample variance of the model with only the intercept.
- We plug the result in out first Bayesian iteration as a proxy for g, given a specific δ.
- We obtain an updated result on g after running the model and we use this result to re-run the model trying to find an updated g.
- We keep running models using this updating method until we detect that g stabilizes, keeping that stable g for the ultimate model.

The model shall look as follows:



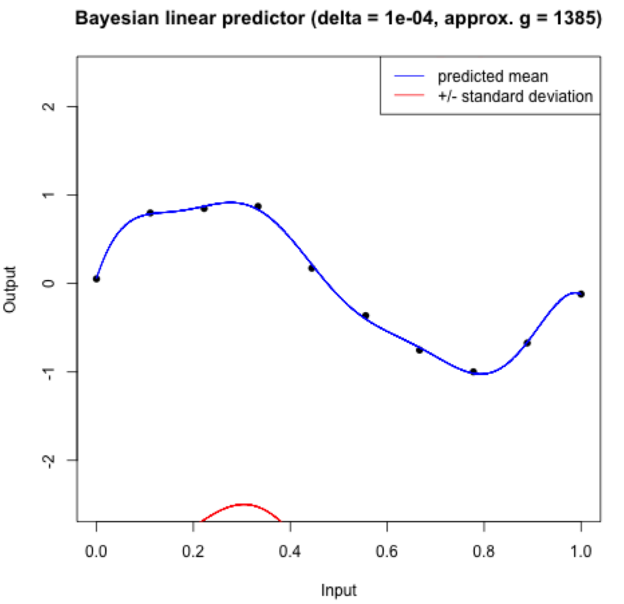
For clarity we will plot the results of posterior draws using $\delta = 1$, as used in the previous simulation exercise, mostly for comparability.

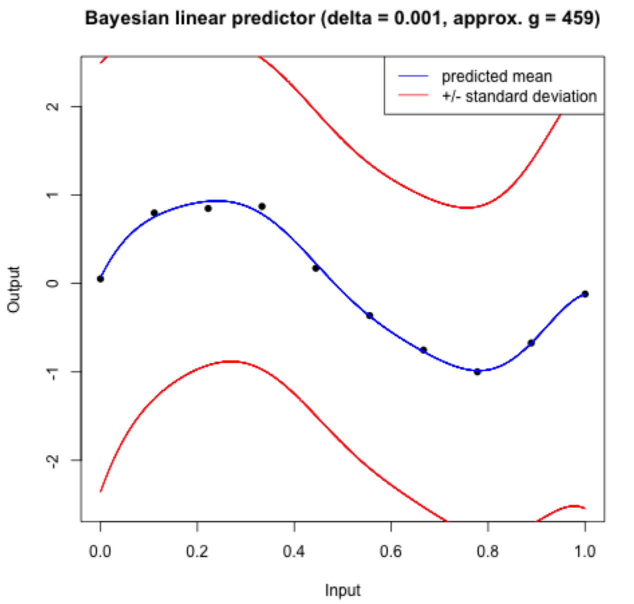


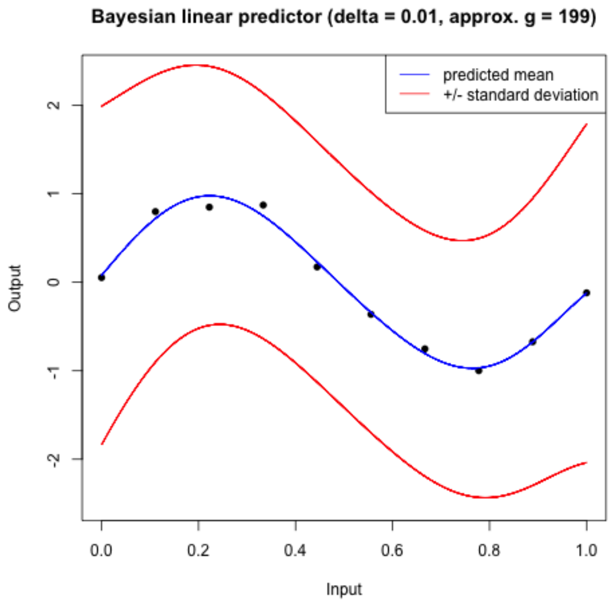


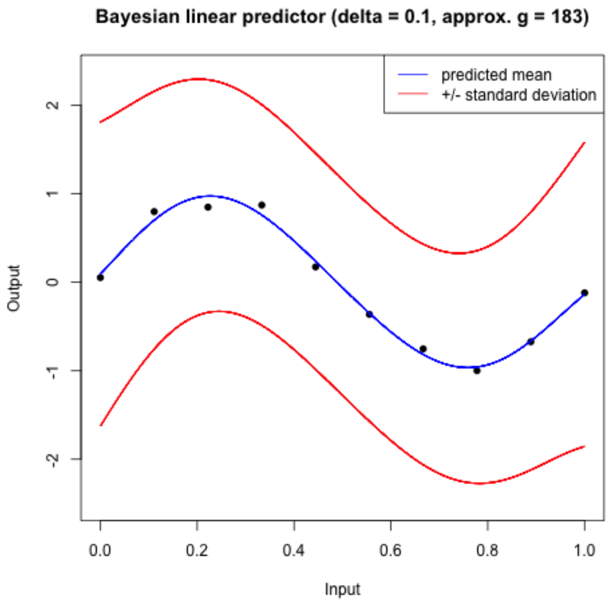
Part 2

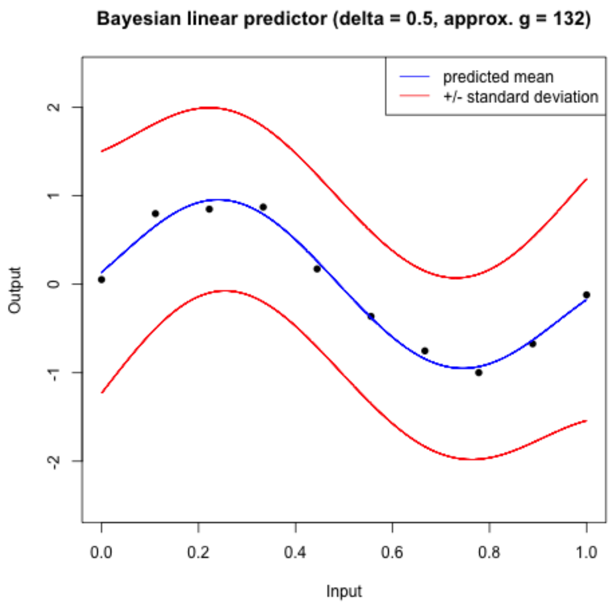
Now we explore how the models are sensible to δ. Note that we fix the value of δ, and that g is then determined by the aforementioned algorithm. To avoid plotting the simulations with its random component, we plot the mean of the predicted model, with its standard deviation side by side:

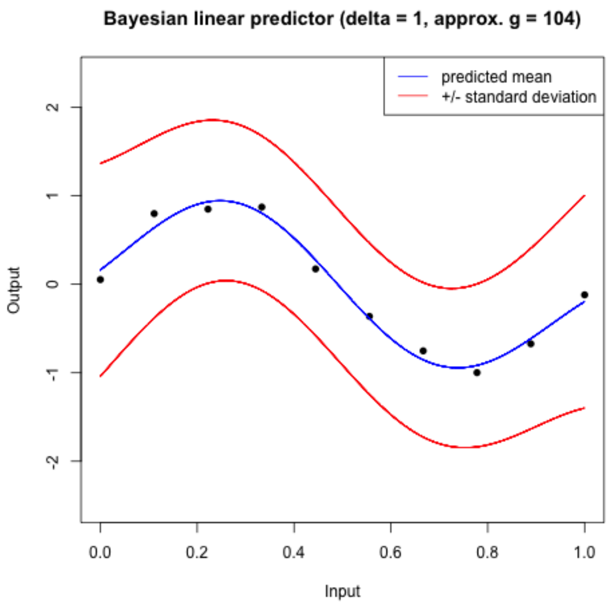


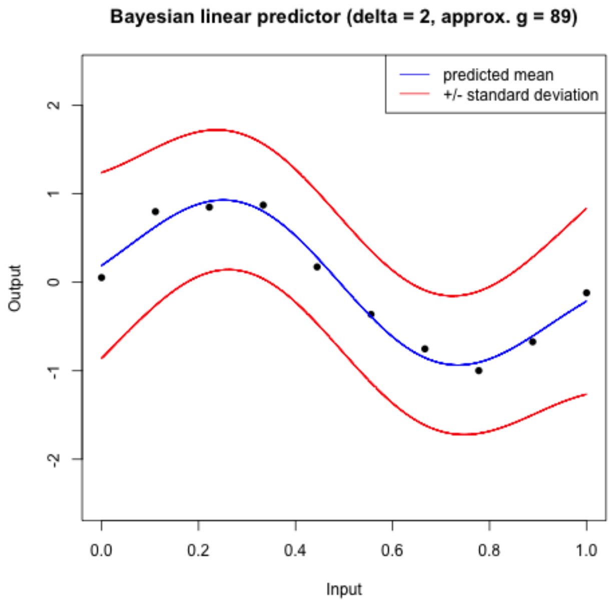


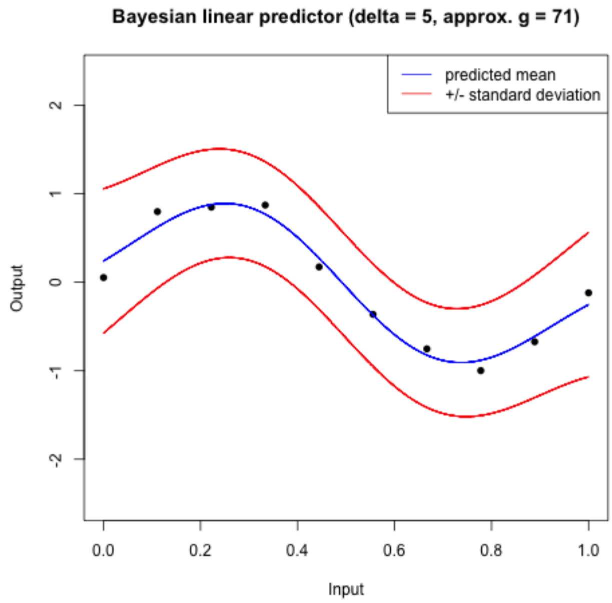

Bayesian linear predictor (delta = 0.01, approx. g = 199)
predicted mean
+/- standard deviation
Output
-2
-1
0
1
2
0.0
0.2
0.4
0.6
0.8
1.0
Input

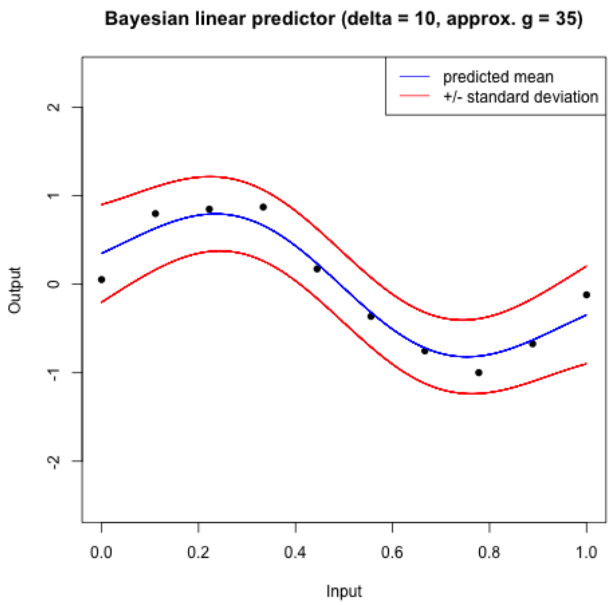

Bayesian linear predictor (delta = 0.1, approx. g = 183)
predicted mean
+/- standard deviation
Output
-2
-1
0
1
2
0.0
0.2
0.4
0.6
0.8
1.0
Input

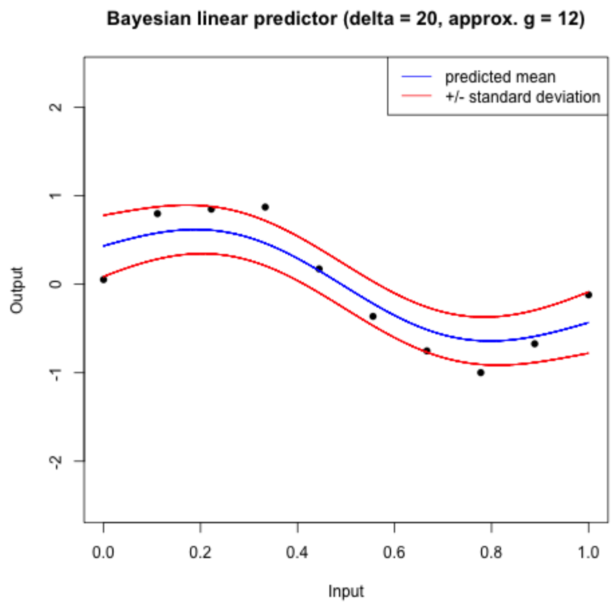

Bayesian linear predictor (delta = 0.5, approx. g = 132)
predicted mean
+/- standard deviation
Output
-2
-1
0
1
2
0.0
0.2
0.4
0.6
0.8
1.0
Input

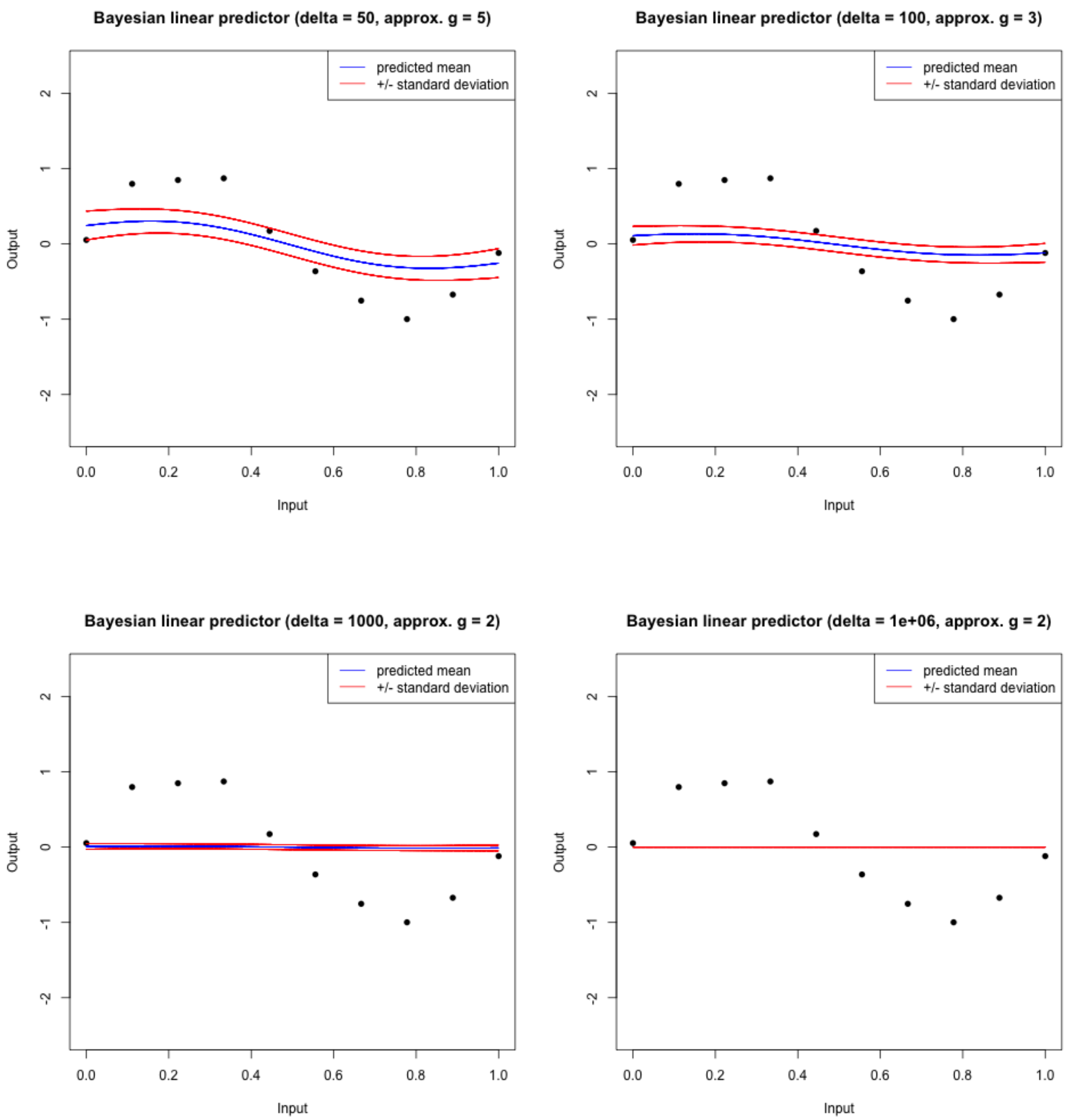

Bayesian linear predictor (delta = 1, approx. g = 104)
predicted mean
+/- standard deviation
Output
-2
-1
0
1
2
0.0
0.2
0.4
0.6
0.8
1.0
Input


Bayesian linear predictor (delta = 2, approx. g = 89)
predicted mean
+/- standard deviation
Output
Input


Bayesian linear predictor (delta = 5, approx. g = 71)
predicted mean
+/- standard deviation
Output
Input


Bayesian linear predictor (delta = 10, approx. g = 35)
predicted mean
+/- standard deviation
Output
Input


Bayesian linear predictor (delta = 20, approx. g = 12)
predicted mean
+/- standard deviation
Output
Input



Note that as δ increases (and g decreases), the prediction becomes flatter and confidence bands become wider, which is consistent with the definition of the parameter δ, a measure of prior precision. Higher prior precision, which implies a smaller g, translates to flatter predicted means with narrower confidence bandwidths (high confidence on the prior, observed data less important). Small δ on the other hand leads to overfitting (data very relevant) with wider confidence bandwidths (more uncertainty).

Exercise 2

Part 1

We minimize the function:

$$f(\mu) = (\mu - a)^2 + \lambda|\mu|$$

Given $a, \lambda > 0$ on μ^+. Hence,

$$\begin{aligned}\frac{\partial f(\mu)}{\partial \mu^+} = 0 \quad &\Leftrightarrow \quad 2(\mu^+ - a) + \lambda = 0 \\ &\Leftrightarrow \quad \mu = \left(a - \frac{\lambda}{2}\right)^+\end{aligned}$$

Also, $f''(\mu^+) = 2$ so this is indeed a minimum.

Part 2

The solution is the following:

$$\begin{aligned}
w_{MAP} &= \arg\max \log p(\mathbf{w}|\mathbf{t}) \\
&= \arg\max \log p(\mathbf{t}|\mathbf{w}) + \log p(\mathbf{w}) \\
&\propto \arg\max \log \prod_n p(\mathbf{t}_n|\mathbf{w}) + \log\exp\left\{-\frac{\delta}{2}\sum_i |w_i|\right\} \\
&\propto \arg\max \sum_n \log \mathcal{N}(\mathbf{w}|q^{-1}\mathbf{I}) + \log\exp\left\{-\frac{\delta}{2}\sum_i |w_i|\right\} \\
&\propto \arg\max \sum_n \log\exp\left\{-\frac{1}{2}(\mathbf{t}_n - \mathbf{w})^T q(\mathbf{t}_n - \mathbf{w})\right\} + \log\exp\left\{-\frac{\delta}{2}\sum_i |w_i|\right\} + C \\
&= \arg\max \sum_n -\frac{1}{2}(\mathbf{t}_n - \mathbf{w})^T q(\mathbf{t}_n - \mathbf{w}) - \frac{\delta}{2}\sum_i |w_i| + C \\
&= \arg\min \sum_n q(\mathbf{t}_n - \mathbf{w})^T(\mathbf{t}_n - \mathbf{w}) + \delta\sum_i |w_i| + C \\
&= \arg\min q\sum_n\sum_i (t_{ni} - w_i)^2 + \delta\sum_i |w_i| + C
\end{aligned}$$

We now we maximize this expression with respect to w_i:

$$\frac{\partial}{\partial w_i} = 0 \quad \Leftrightarrow \quad -2q\sum_n (t_{ni} - w_i) + \delta\frac{w_i}{|w_i|} = 0$$

Given that we are working on the side of w_i^+ this simplifies to:

$$\begin{aligned}\frac{\partial}{\partial w_i^+} = 0 \quad &\Leftrightarrow \quad -2q\sum_n (t_{ni} - w_i) + \delta = 0 \\ &\Leftrightarrow \quad w_{MAP} = \frac{1}{N}\left(\sum_n t_{ni} - \frac{\delta}{2}q^{-1}\right)^+\end{aligned}$$