Q1

	Larger C	Smaller C
Model capacity (large/small?)	Large	Low
Overfitting/Underfitting?	Overfitting	Underfitting
Bias variance (how/low?)	Low bias & High variance	High bias & Low variance

Q2

$$
\begin{aligned}
p(w|\mathcal{D}) &\propto p(w)p(w|\mathcal{D}) \\
&= p(w)\prod_{i=1}^{m} p(t^{(i)}|x^{(i)}, w) \\
&= \frac{1}{\sqrt{2\pi}\sigma_w}\exp\{-\frac{1}{2\sigma_w^2}w^2\}\prod_{i=1}^{m}\frac{1}{\sqrt{2\pi}\sigma_\epsilon}\exp\{-\frac{1}{2\sigma_\epsilon^2}(wx^{(i)}-t^{(i)})^2\} \\
&= \frac{1}{\sqrt{2\pi}\sigma_w}\exp\{-\frac{1}{2\sigma_w^2}w^2\}(\frac{1}{\sqrt{2\pi}\sigma_\epsilon})^m\exp\{-\frac{1}{2\sigma_\epsilon^2}\sum_{i=1}^{m}(wx^{(i)}-t^{(i)})^2\} \\
&\propto \exp\{-\frac{1}{2}(\frac{w^2}{\sigma_w^2}+\frac{\sum_{i=1}^{m}(wx^{(i)}-t^{(i)})^2}{\sigma_\epsilon^2})\} \\
&\quad \text{(Dropping constant term inside exp.)} \\
&= \exp\{-\frac{1}{2}(\frac{w^2}{\sigma_w^2}+\frac{\sum_{i=1}^{m}(x^{(i)})^2w^2-2\sum_{i=1}^{m}t^{(i)}x^{(i)}w+\sum_{i=1}^{m}(t^{(i)})^2}{\sigma_\epsilon^2})\} \\
&\quad \text{(Rearrange.)} \\
&\propto \exp\{-\frac{1}{2\sigma_w^2\sigma_\epsilon^2}\Big((\sigma_\epsilon^2+\sigma_w^2\sum_{i=1}^{m}(x^{(i)})^2)w^2-2\sigma_w^2\sum_{i=1}^{m}t^{(i)}x^{(i)}w\Big)\} \\
&\quad \text{(Dropping constant term inside exp.)} \\
&\propto \exp\{-\frac{\sigma_\epsilon^2+\sigma_w^2\sum_{i=1}^{m}(x^{(i)})^2}{2\sigma_w^2\sigma_\epsilon^2}\big(w^2-2\frac{\sigma_w^2\sum_{i=1}^{m}t^{(i)}x^{(i)}}{\sigma_\epsilon^2+\sigma_w^2\sum_{i=1}^{m}(x^{(i)})^2}w\big)\} \\
&\quad \text{(Dropping constant term inside exp.)} \\
&\propto \exp\{-\frac{\sigma_\epsilon^2+\sigma_w^2\sum_{i=1}^{m}(x^{(i)})^2}{2\sigma_w^2\sigma_\epsilon^2}\big(w-\frac{\sigma_w^2\sum_{i=1}^{m}t^{(i)}x^{(i)}}{\sigma_\epsilon^2+\sigma_w^2\sum_{i=1}^{m}(x^{(i)})^2}\big)^2\} \\
&\quad \text{(Dropping constant term inside exp.)}
\end{aligned}
$$

We can see that $p(w|\mathcal{D})$ is also of the form of a Gaussian distribution, where $\sigma_{post}^2 = \frac{\sigma_w^2\sigma_\epsilon^2}{\sigma_\epsilon^2+\sigma_w^2\sum_{i=1}^{m}(x^{(i)})^2}$ and $\mu_{post} = \frac{\sigma_w^2\sum_{i=1}^{m}t^{(i)}x^{(i)}}{\sigma_\epsilon^2+\sigma_w^2\sum_{i=1}^{m}(x^{(i)})^2}$

Q3

We know that $p(\boldsymbol{w}|\mathcal{D}) \propto p(\boldsymbol{w})p(\mathcal{D}|\boldsymbol{w})$.

For maximization, it is easy to see that

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

Following the lecture notes given a dataset $\mathcal{D}$ and a linear regression model, we know that

$$p(\mathcal{D}|\boldsymbol{w}) = \prod_{i=1}^{m} \frac{1}{\sigma\sqrt{2\pi}} \exp\big\{ -\frac{1}{2\sigma^2}(t^{(i)} - \boldsymbol{w}^\top \boldsymbol{x}^{(i)})^2 \big\}$$

and

$$\log p(\mathcal{D}|\boldsymbol{w}) = -\frac{1}{2\sigma^2} \sum_{i=1}^{m} (t^{(i)} - \boldsymbol{w}^\top \boldsymbol{x}^{(i)})^2 + \text{const} \tag{1}$$

Assume each w_i in $\boldsymbol{w}$ is drawn from $\text{Laplace}(0, b)$, denoted as

$$p(w_i) = \frac{1}{2b} \exp\{-\frac{|w_i|}{b}\}$$

We have

$$p(\boldsymbol{w}) = \prod_{i=1}^{m} \frac{1}{2b} \exp\{-\frac{|w_i|}{b}\}$$

and

$$\begin{aligned} \log p(\boldsymbol{w}) &= -\frac{1}{b} \sum_{i=1}^{m} |w_i| + \text{const} \\ &= -\frac{1}{b} ||\boldsymbol{w}||_1 + \text{const} \end{aligned} \tag{2}$$

From (1) and (2), we have

$$\begin{aligned} &\log p(\boldsymbol{w}) + \log p(\mathcal{D}|\boldsymbol{w}) \\ = &-\frac{1}{2\sigma^2} \sum_{i=1}^{m} (t^{(i)} - \boldsymbol{w}^\top \boldsymbol{x}^{(i)})^2 - \frac{1}{b} ||\boldsymbol{w}||_1 + \text{const} \end{aligned}$$

Therefore, the Laplace prior gives us an linear regression model with L_1 regularization.