Linear Regression - Deterministic Representation
Model: $f : \mathbf{x} \to y, \;\; f(\mathbf{x}) = \mathbf{w}^T \mathbf{x}$

Training data: $\{(\mathbf{x_1}, y_1), \ldots, (\mathbf{x_N}, y_N)\}$, or $\mathbf{X} = \begin{bmatrix} - \mathbf{x_1}^T - \\ \vdots \\ - \mathbf{x_N}^T - \end{bmatrix}$ and $\mathbf{y} = [y_1, \ldots, y_N]^T$

Evaluation through residual sum of squares (no regularization):
$J(\mathbf{w}) = RSS(\mathbf{w}) = \sum_{n=1}^{N} (y_n - \mathbf{w}^T \mathbf{x_n})^2 = (\mathbf{y} - \mathbf{X}\mathbf{w})^T (\mathbf{y} - \mathbf{X}\mathbf{w})$

Evaluation through residual sum of squares (l2-norm regularization):

$$\begin{aligned} J(\mathbf{w}) &= RSS(\mathbf{w}) + \lambda \sum_{d=1}^{D} w_d^2 \\ &= RSS(\mathbf{w}) + \lambda \|\mathbf{w}\|_2^2 \\ &= (\mathbf{y} - \mathbf{X}\mathbf{w})^T (\mathbf{y} - \mathbf{X}\mathbf{w}) + \lambda \|\mathbf{w}\|_2^2 \\ &= \mathbf{y}^T \mathbf{y} - 2\mathbf{w}^T (\mathbf{X}^T \mathbf{y}) + \mathbf{w}^T (\mathbf{X}^T \mathbf{X}) \mathbf{w} + \lambda \mathbf{w}^T \mathbf{w} \\ &= \mathbf{y}^T \mathbf{y} - 2\mathbf{w}^T (\mathbf{X}^T \mathbf{y}) + \mathbf{w}^T (\mathbf{X}^T \mathbf{X} + \lambda \mathbf{I}_{D \times D}) \mathbf{w} \end{aligned}$$

Optimization (analytical/closed-form solution):
$\nabla J(\mathbf{w}) = 0 \Rightarrow -2\mathbf{X}^T \mathbf{y} + 2(\mathbf{X}^T \mathbf{X} + \lambda \mathbf{I}_{D \times D}) \mathbf{w} = 0 \Rightarrow \mathbf{w}^* = (\mathbf{X}^T \mathbf{X} + \lambda \mathbf{I}_{D \times D})^{-1} \mathbf{X}^T \mathbf{y}$

Linear Regression - Noisy observation model
Assuming Gaussian noise and Gaussian prior on the weights
Model: $y = \mathbf{w}^T\mathbf{x} + \epsilon, \;\; \epsilon \sim \mathcal{N}(0, \sigma^2), \; w_d \sim \mathcal{N}(0, \tau^2)$

Model Posterior (given the observed data and the prior belief):

$$\begin{aligned}
p(\mathbf{w}|\mathcal{D}) &= p(\mathcal{D}|\mathbf{w})p(\mathbf{w}) \\
&= \prod_{n=1}^{N} p(y_n|\mathbf{x_n}, \mathbf{w}) \prod_{d=1}^{D} p(w_d) \\
&= \prod_{n=1}^{N} \mathcal{N}(\mathbf{w}^T\mathbf{x_n}, \sigma^2) \prod_{d=1}^{D} \mathcal{N}(0, \tau^2) \\
&= \prod_{n=1}^{N} \frac{1}{\sigma\sqrt{2\pi}} \exp\left[-\frac{(y_n - \mathbf{w}^T\mathbf{x_n})^2}{2\sigma^2}\right] \cdot \prod_{d=1}^{D} \frac{1}{\tau\sqrt{2\pi}} \exp\left(-\frac{w_d^2}{2\tau^2}\right)
\end{aligned}$$

Performing logarithmic transformation

$$\begin{aligned}
\log p(\mathbf{w}|\mathcal{D}) &= \log\left\{\prod_{n=1}^{N} \frac{1}{\sigma\sqrt{2\pi}} \exp\left[-\frac{(y_n - \mathbf{w}^T\mathbf{x_n})^2}{2\sigma^2}\right]\right\} + \log\left\{\prod_{d=1}^{D} \frac{1}{\sigma\sqrt{2\tau}} \exp\left[-\frac{w_d^2}{2\tau^2}\right]\right\} \\
&= \sum_{n=1}^{N} \log\left(\frac{1}{\sigma\sqrt{2\pi}}\right) - \sum_{n=1}^{N} \frac{(y_n - \mathbf{w}^T\mathbf{x_n})^2}{2\sigma^2} + \sum_{d=1}^{D} \log\left(\frac{1}{\tau\sqrt{2\pi}}\right) - \sum_{d=1}^{D} \frac{w_d^2}{2\tau^2} \\
&= -N\log\sigma - N\log(2\pi)^{1/2} - \sum_{n=1}^{N} \frac{(y_n - \mathbf{w}^T\mathbf{x_n})^2}{2\sigma^2} - D\log\tau - D\log(2\pi)^{1/2} - \sum_{d=1}^{D} \frac{w_d^2}{2\tau^2} \\
&= -\frac{N}{2}\log\sigma^2 - \frac{1}{2\sigma^2} \sum_{n=1}^{N} (y_n - \mathbf{w}^T\mathbf{x_n})^2 + -\frac{D}{2}\log\tau^2 - \frac{1}{2\tau^2} \sum_{d=1}^{D} w_d^2 + \text{const} \\
&= -\frac{1}{2}\left[N\log\sigma^2 + \frac{1}{\sigma^2}(\mathbf{y} - \mathbf{X}\mathbf{w})^T(\mathbf{y} - \mathbf{X}\mathbf{w}) + D\log\tau^2 + \frac{1}{\tau^2}\mathbf{w}^T\mathbf{w}\right] + \text{const} \\
&= -\frac{1}{2}\left[N\log\sigma^2 + D\log\tau^2 + \frac{1}{\sigma^2}\mathbf{y}^T\mathbf{y} - 2\frac{1}{\sigma^2}\mathbf{w}^T(\mathbf{X}^T\mathbf{y}) + \frac{1}{\sigma^2}\mathbf{w}^T(\mathbf{X}^T\mathbf{X})\mathbf{w} + \frac{1}{\tau^2}\mathbf{w}^T\mathbf{w}\right] + \text{const}
\end{aligned}$$

Optimization (analytical/closed-form solution):

$$\nabla_{\mathbf{w}}(\log p(\mathbf{w}|\mathcal{D})) = 0 \Rightarrow$$

$$-2\frac{1}{\sigma^2}\mathbf{X}^T\mathbf{y} + 2\frac{1}{\sigma^2}(\mathbf{X}^T\mathbf{X})\mathbf{w} + 2\frac{1}{\tau^2}\mathbf{w} = 0 \Rightarrow$$

$$\left(\frac{1}{\sigma^2}\mathbf{X}^T\mathbf{X} + \frac{1}{\tau^2}\mathbf{I}_{D\times D}\right)\mathbf{w} - \frac{1}{\sigma^2}\mathbf{X}^T\mathbf{y} = 0$$

$$\left(\mathbf{X}^T\mathbf{X} + \frac{\sigma^2}{\tau^2}\mathbf{I}_{D\times D}\right)\mathbf{w} - \mathbf{X}^T\mathbf{y} = 0$$

$$\mathbf{w}^* = (\mathbf{X}^T\mathbf{X} + \frac{\sigma^2}{\tau^2}\mathbf{I}_{D\times D})^{-1}\mathbf{X}^T\mathbf{y}$$