Ridge - Regul° - exact:

We now minimize

$$\mathcal{L} = \frac{1}{N}\sum_n^N (w x_n - t_n)^2 + \lambda \underbrace{\|w\|_2^2}_{= w^T \cdot w}$$

$$D\mathcal{L} = \frac{2}{N}\sum_n^N (w x_n - t_n) x_n + 2\lambda \vec{w}$$

(We check $D\|w\|^2 = \vec{w}$: $\vec{D}(w_1^2 + w_2^2 + \dots + w_D^2) = \begin{pmatrix} 2w_1 \\ 2w_2 \\ \vdots \\ 2w_D \end{pmatrix}$)

$w^* = \text{argmin}\ \mathcal{L}$

$\Rightarrow$ w^* is such that $\vec{D}_{\vec{w}}\mathcal{L} = \vec{0}$

$$\Rightarrow \frac{2}{N}\underbrace{X^T(Xw - T)}_{\text{shape } (D,N)(N,D)(D,1) - (N,1))} + 2\lambda \underbrace{w}_{\text{shape } (D,1)} = 0$$

$\Rightarrow (D,N)(N,1) \to (D,1)$

$$\Rightarrow X^T(Xw - T) + N\lambda w = 0$$

$$w = (X^T X + N\lambda)^{-1} X^T T$$

↳ (here is an implicit $I_{D,D}$ matrix)

Again, the hard step is inverting $X^T X$, which is just $O(D^3)$ (+ $O(ND^2)$ to run things)

When $D = 1$,

$$w = \frac{X^T T}{N\lambda + X^T X} = \frac{\sum_{n=1}^N x_n t_n}{N\lambda + \sum_n^N x_n^2}$$

If $\text{Var}(x) \gg \lambda$, large w is permitted

For large λ, $w \to 0$. This has to be compared with $\frac{1}{N}\sum_n^N x_n^2$, which is Variance$(x_n)$, essentially. $(= \langle x \rangle)$