Question 2

Part a

Under i.i.d assumptions we can show that:

$$P(w|\mu=0,b)=\prod_{i=1}^{m}P(w_i|\mu=0,b)=$$
$$\prod_{i=1}^{m}\frac{1}{2b}exp(-\frac{|w_i|}{b})=$$
$$(2b)^{-m}exp(-\frac{\sum_{i=1}^{m}|w_i|}{b})$$

Part b

The lasso regression problem can be written as follows:

$$\hat{w}_{LASSO}=argmin_w\ ||Xw-y||^2+\lambda||w||_1$$
$$=argmin_w\ \sum_{i=1}^{m}(x_i^Tw-y_i)^2+\lambda||w||_1$$

We can show that:

$$p(\{(x_i,y_i\}_{i=1}^{m}|w,\mu=0,b)=\Pi_{i=1}^{m}p((x_i,y_i)|w)=\Pi_{i=1}^{m}p(y_i|x_i,w)$$
$$=\Pi_{i=1}^{m}\frac{1}{\sqrt{2\pi}}exp(-\frac{(x_i^Tw-y_i)^2}{2})$$
$$=(2\pi)^{-\frac{m}{2}}exp(-\frac{1}{2}\sum_{i=1}^{m}(x_i^Tw-y_i)^2)$$

Using our knowledge of the noise $\epsilon_i\sim\mathcal{N}(0,1)$

We can now show that $\hat{w}_{MAP}=\hat{w}_{LASSO}$ using bayes theroem:

$$\hat{w}_{MAP}\triangleq argmax_w\ p(w|\{(x_i,y_i)\}_{i=1}^{m},\mu=0,b)$$
$$=argmax_w\ p(\{(x_i,y_i)\}_{i=1}^{m}|w,\mu=0,b)p(w|\mu=0,b)$$
$$=argmax_w\ ln(p(\{(x_i,y_i)\}_{i=1}^{m}|w,\mu=0,b)p(w|\mu=0,b))$$
$$=argmax_w\ -\frac{m}{2}\cdot ln(2\pi)-\frac{1}{2}\sum_{i=1}^{m}(x_i^Tw-y_i)^2-m\cdot ln(2b)-\frac{\sum_{i=1}^{m}|w_i|}{b}$$
$$=argmin_w\ \sum_{i=1}^{m}(x_i^Tw-y_i)^2+\frac{1}{b}\sum_{i=1}^{m}|w_i|$$
$$=argmin_w\ ||Xw-y||^2+\lambda||w||_1$$
$$=\hat{w}_{LASSO}$$

Where the suitable parameter λ is $\frac{1}{b}$.

Part c

The intuition is that by looking at the figure we can see that the cdf is more "concentrated" around 0, than the normal-distributed case.
Translated to the regression (Least Squares) problem, using the above formulation, we can see that this

means that most weight w_i are probably close to 0, and fewer dominent weights are more distant from 0. This is exactly the sparsity we are talking about, as the lasso regressor fits a weights vector w, in which most of the weights are very close to 0, and only a few are different from 0.