1 Norms and Regularizers

MAP maximises $\log p(\mathcal{D}|\mathbf{w}) + \log p(\mathbf{w})$, which is the same as minimising $-\log p(\mathcal{D}|\mathbf{w}) - \log p(\mathbf{w})$. The first term gives the error function, so, as in the question, the regulariser is $-\log p(\mathbf{w})$, up to an additive constant that does not change the minimiser.

Problem 1. With covariance $\beta^{-1}\mathbf{I}$, the determinant is $|\beta^{-1}\mathbf{I}| = \beta^{-M}$ and the inverse covariance is $\beta\mathbf{I}$, so

$$p(\mathbf{w}) = \left(\frac{\beta}{2\pi}\right)^{M/2} \exp\left(-\frac{\beta}{2}(\mathbf{w}-\boldsymbol{\mu})^\top(\mathbf{w}-\boldsymbol{\mu})\right), \qquad \log p(\mathbf{w}) = \frac{M}{2}\log\frac{\beta}{2\pi} - \frac{\beta}{2}\|\mathbf{w}-\boldsymbol{\mu}\|_2^2.$$

The first term is a constant, so the regulariser is

$$\boxed{\frac{\beta}{2}\|\mathbf{w}-\boldsymbol{\mu}\|_2^2 = \frac{\beta}{2}\sum_{m=1}^{M}(w_m-\mu_m)^2.}$$

This is an L2 penalty that pulls $\mathbf{w}$ towards $\boldsymbol{\mu}$ instead of towards $\mathbf{0}$. Setting $\boldsymbol{\mu} = \mathbf{0}$ recovers the usual L2 regulariser with $\lambda = \beta$.

Problem 2. The log of the product is the sum of the logs:

$$\log p(\mathbf{w}) = \sum_{m=1}^{M}\left(-\log(2b) - \frac{|w_m|}{b}\right) = -M\log(2b) - \frac{1}{b}\sum_{m=1}^{M}|w_m|.$$

The first term is a constant, so the regulariser is

$$\boxed{\frac{1}{b}\sum_{m=1}^{M}|w_m| = \frac{1}{b}\|\mathbf{w}\|_1,}$$

an L1 penalty (as in the lasso) with strength $\lambda = 1/b$. A smaller scale b concentrates the prior more tightly around zero and gives a stronger penalty. Unlike the L2 penalty, the L1 penalty tends to set some weights exactly to zero; for an explanation, see this answer on Cross Validated.

Problem 3. The density of $\mathcal{U}(w_m|0,1)$ is 1 on $[0,1]$ and 0 elsewhere. By independence,

$$p(\mathbf{w}) = \prod_{m=1}^{M} p(w_m) = \begin{cases} 1 & \text{if } 0 \le w_m \le 1 \text{ for every } m, \\ 0 & \text{otherwise,} \end{cases} \qquad \log p(\mathbf{w}) = \begin{cases} 0 & \text{if } 0 \le w_m \le 1 \text{ for every } m, \\ -\infty & \text{otherwise.} \end{cases}$$

The regulariser $-\log p(\mathbf{w})$ is 0 inside the box $[0,1]^M$ and $+\infty$ outside it. MAP estimation therefore maximises the likelihood subject to the constraint $0 \le w_m \le 1$ for every m. Inside the box, the prior does not prefer any weight vector over another, so there is no shrinkage: if the likelihood has a unique maximiser and it lies inside the box, the MAP estimate equals it.

2 Connection between Naive Bayes and Logistic Regression

Clarification of the assumptions. The features are independent *conditional on the class* t, and each conditional distribution $p(x^m|t=j)$ is Gaussian. The features need not be identically distributed: their means μ_j^m and variances v^m may depend on m. Nor need they be independent or Gaussian after averaging over the class. The shared-variance assumption means that each v^m is the same for classes 0 and 1. Throughout the derivation, assume $v^m > 0$ and $0 < \pi < 1$, so the densities and log odds are well defined.

Problem 1. Expand the sum over $t \in \{0,1\}$ in the denominator, use the product rule $p(\mathbf{x},t) = p(\mathbf{x}|t)p(t)$, and divide the numerator and the denominator by $p(\mathbf{x}|t=1)p(t=1)$:

$$\begin{aligned} p(t=1|\mathbf{x}) &= \frac{p(\mathbf{x}|t=1)p(t=1)}{p(\mathbf{x},t=0)+p(\mathbf{x},t=1)} \\ &= \frac{p(\mathbf{x}|t=1)p(t=1)}{p(\mathbf{x}|t=0)p(t=0)+p(\mathbf{x}|t=1)p(t=1)} \\ &= \frac{1}{1+\dfrac{p(\mathbf{x}|t=0)p(t=0)}{p(\mathbf{x}|t=1)p(t=1)}} \\ &= \frac{1}{1+\exp\left(\log\dfrac{p(t=0)}{p(t=1)}+\log\dfrac{p(\mathbf{x}|t=0)}{p(\mathbf{x}|t=1)}\right)}, \end{aligned}$$

where the last line writes the ratio as exp log of itself and uses $\log(ab) = \log a + \log b$.

Problem 2. Substitute the naive Bayes model. The normalising factors $1/\sqrt{2\pi v^m}$ are the same in the numerator and the denominator, because the variances are shared, so they cancel:

$$\begin{aligned} \log\frac{p(\mathbf{x}|t=0)}{p(\mathbf{x}|t=1)} &= \log\frac{\prod_m \frac{1}{\sqrt{2\pi v^m}}\exp\left(-\frac{(x^m-\mu_0^m)^2}{2v^m}\right)}{\prod_m \frac{1}{\sqrt{2\pi v^m}}\exp\left(-\frac{(x^m-\mu_1^m)^2}{2v^m}\right)} \\ &= \log\prod_m \frac{\exp\left(-\frac{(x^m-\mu_0^m)^2}{2v^m}\right)}{\exp\left(-\frac{(x^m-\mu_1^m)^2}{2v^m}\right)} \\ &= \sum_{m=1}^{M} \log\frac{\exp\left(-\frac{(x^m-\mu_0^m)^2}{2v^m}\right)}{\exp\left(-\frac{(x^m-\mu_1^m)^2}{2v^m}\right)}. \end{aligned}$$

Problem 3. Combine the exponentials and cancel log with exp:

$$\log\frac{p(\mathbf{x}|t=0)}{p(\mathbf{x}|t=1)} = \sum_m \frac{(x^m-\mu_1^m)^2-(x^m-\mu_0^m)^2}{2v^m}.$$

Expand the squares. The $(x^m)^2$ terms cancel, because the two classes share the variance v^m:

$$(x^m-\mu_1^m)^2-(x^m-\mu_0^m)^2 = -2x^m\mu_1^m+(\mu_1^m)^2+2x^m\mu_0^m-(\mu_0^m)^2 = 2(\mu_0^m-\mu_1^m)\,x^m+(\mu_1^m)^2-(\mu_0^m)^2.$$

Dividing by $2v^m$ and summing over m,

$$\log \frac{p(\mathbf{x}|t=0)}{p(\mathbf{x}|t=1)} = \sum_m \left(\frac{\mu_0^m - \mu_1^m}{v^m} x^m + \frac{(\mu_1^m)^2 - (\mu_0^m)^2}{2v^m} \right).$$

Problem 4. Substituting $\log \frac{p(t=0)}{p(t=1)} = \log \frac{1-\pi}{\pi}$ and the result of Problem 3 into the expression from Problem 1 gives

$$p_{\mathrm{NB}}(t=1|\mathbf{x}) = \frac{1}{1 + \exp\left(\log \frac{1-\pi}{\pi} + \sum_m \frac{(\mu_1^m)^2 - (\mu_0^m)^2}{2v^m} + \sum_m \frac{\mu_0^m - \mu_1^m}{v^m} x^m \right)}.$$

Compare this with

$$p_{\mathrm{LR}}(t=1|\mathbf{x}) = \frac{1}{1 + \exp\left(-w^0 - \sum_{m=1}^M w^m x^m \right)}.$$

The two expressions agree for every $\mathbf{x}$ when the terms inside the exponentials agree. Matching the constant terms and the coefficients of each x^m, and moving the minus sign across, gives

$$\boxed{w^0 = -\log \frac{1-\pi}{\pi} - \sum_m \frac{(\mu_1^m)^2 - (\mu_0^m)^2}{2v^m}, \qquad w^m = -\frac{\mu_0^m - \mu_1^m}{v^m} = \frac{\mu_1^m - \mu_0^m}{v^m}.}$$

Two remarks:

- The shared variances are what make the log odds linear in $\mathbf{x}$. With class-specific variances, the $(x^m)^2$ terms in Problem 3 do not cancel, and the log odds are quadratic in $\mathbf{x}$.
- The two models have the same form but are fitted differently. Naive Bayes estimates π, μ_j^m and v^m by maximising the joint likelihood $p(\mathbf{x}, t)$, and the weights then follow from the formulas above. Logistic regression fits $\mathbf{w}$ directly by maximising the conditional likelihood $p(t|\mathbf{x})$. On the same data they generally give different weights: the displayed formulas describe the logistic representation induced by a fitted naive Bayes model, rather than an identity between the two separately fitted models.

Bonus. The sample-size comparison in the question describes a possible trade-off, not a universal model-selection rule. Naive Bayes can perform well with few examples, while logistic regression may benefit from more data when the generative assumptions are inaccurate. If the shared-variance Gaussian naive Bayes model is correctly specified, consistent fits of both models can converge to the same Bayes classifier. Sample size alone therefore does not determine which method performs better; the validity of the assumptions and the choice of regularisation also matter.