1 A K-Sided Die

1.1

$$\begin{aligned}
\ln p(\boldsymbol{\theta} \mid \mathcal{D}) &= \ln \prod_{k=1}^{K} \theta_k^{N_k+\alpha_k-1} && (1)\\
&= \sum_{k=1}^{K} \ln \theta_k^{N_k+\alpha_k-1} && (2)\\
&= \sum_{k=1}^{K} (N_k+\alpha_k-1) \ln \theta_k && (3)
\end{aligned}$$

1.2

$$\begin{aligned}
\ell(\boldsymbol{\theta}, \lambda, \boldsymbol{\mu}) &= \ln p(\boldsymbol{\theta} \mid \mathcal{D}) + \lambda\big(\sum_{k=1}^{K} \theta_k - 1\big) + \sum_{k=1}^{K} \mu_k \theta_k && (4)\\
&= \sum_{k=1}^{K} (N_k+\alpha_k-1) \ln \theta_k + \lambda\big(\sum_{k=1}^{K} \theta_k - 1\big) + \sum_{k=1}^{K} \mu_k \theta_k && (5)
\end{aligned}$$

1.3

KKT conditions (for all $k \in \{1, \dots, K\}$ if not specified):

$$\begin{aligned}
&\sum_{k=1}^{K} \theta_k - 1 = 0\\
&\mu_k \geq 0\\
&\theta_k \geq 0\\
&\mu_k \theta_k = 0
\end{aligned}$$

1.4

$$\begin{aligned}
\frac{\partial}{\partial \theta_k} \ell(\boldsymbol{\theta}, \lambda, \boldsymbol{\mu}) &= \frac{\partial}{\partial \theta_k} \Big[\sum_{k'=1}^{K} (N_{k'}+\alpha_{k'}-1) \ln \theta_{k'} + \lambda\big(\sum_{k'=1}^{K} \theta_{k'} - 1\big) + \sum_{k'=1}^{K} \mu_{k'} \theta_{k'} \Big] && (6)\\
&= (N_k+\alpha_k-1) \frac{\partial}{\partial \theta_k} \ln \theta_k + \lambda \frac{\partial}{\partial \theta_k} \theta_k + \frac{\partial}{\partial \theta_k} \mu_k \theta_k && (7)\\
&= \frac{1}{\theta_k}(N_k+\alpha_k-1) + \lambda + \mu_k && (8)\\
&\stackrel{!}{=} 0 && (9)
\end{aligned}$$

We see that $\theta_k \neq 0$, so $\theta_k > 0$, which means we get $\mu_k = 0$ in order to meet the KKT conditions.

$$\begin{aligned}
-\lambda &= \frac{1}{\theta_k}(N_k+\alpha_k-1) && (10)\\
\theta_k &= \frac{N_k+\alpha_k-1}{-\lambda} && (11)
\end{aligned}$$

Summing this over all k should equal 1:

$$\sum_{k=1}^{K} \theta_k = 1 = \sum_{k=1}^{K} \frac{N_k + \alpha_k - 1}{-\lambda} \tag{12}$$

$$= -\frac{1}{\lambda} \sum_{k=1}^{K} (N_k + \alpha_k - 1) \tag{13}$$

$$-\lambda = \sum_{k=1}^{K} (N_k + \alpha_k - 1) \tag{14}$$

Combining this with Eq. (11) gives:

$$\theta_k = \frac{N_k + \alpha_k - 1}{\sum_{k=1}^{K} (N_k + \alpha_k - 1)} \tag{15}$$

$$\boldsymbol{\theta}_{\text{MAP}} = \frac{\mathbf{N} + \boldsymbol{\alpha} - \mathbf{1}}{\sum_{k=1}^{K} (N_k + \alpha_k - 1)} \tag{16}$$

2 Maximum Margin Classifier

2.1

In this case we are minimizing, so we get minus signs for the constraints.

$$L(R, \beta, \boldsymbol{\xi}, \{\lambda_n\}, \{\mu_n\}) = \frac{1}{2}\beta^2 + C \sum_{n=1}^{N} \xi_n - \sum_{n=1}^{N} \lambda_n [t_n(\beta \|x_n\| - R) - 1 + \xi_n] - \sum_{n=1}^{N} \mu_n \xi_n \tag{17}$$

2.2

KKT conditions (for all $n \in \{1, \ldots, N\}$ if not specified):

$$\begin{aligned}
&t_n(\beta \|x_n\| - R) - 1 + \xi_n \geq 0 \\
&\lambda \geq 0 \\
&\xi_n \geq 0 \\
&\mu_n \geq 0 \\
&\lambda_n [t_n(\beta \|x_n\| - R) - 1 + \xi_n] = 0 \\
&\mu_n \xi_n = 0
\end{aligned}$$

Which together make up a total of $6N$ conditions.

2.3

$$\frac{\partial}{\partial R} L(R, \beta, \boldsymbol{\xi}, \{\lambda_n\}, \{\mu_n\}) = \frac{\partial}{\partial R} \Big[-\sum_{n=1}^{N} \lambda_n [t_n(\beta \|x_n\| - R) - 1 + \xi_n] \Big] \tag{18}$$

$$= -\sum_{n=1}^{N} \lambda_n t_n \frac{\partial}{\partial R} (\beta \|x_n\| - R) \tag{19}$$

$$= \sum_{n=1}^{N} \lambda_n t_n \stackrel{!}{=} 0 \tag{20}$$

$$\frac{\partial}{\partial \beta} L(R, \beta, \boldsymbol{\xi}, \{\lambda_n\}, \{\mu_n\}) = \frac{\partial}{\partial \beta} \frac{1}{2}\beta^2 - \frac{\partial}{\partial \beta} \sum_{n=1}^{N} \lambda_n [t_n(\beta \|x_n\| - R) - 1 + \xi_n] \tag{21}$$

$$= \beta - \sum_{n=1}^{N} \lambda_n t_n \frac{\partial}{\partial \beta}(\beta \|x_n\| - R) \tag{22}$$

$$= \beta - \sum_{n=1}^{N} \lambda_n t_n \|x_n\| \stackrel{!}{=} 0 \tag{23}$$

$$\beta = \sum_{n=1}^{N} \lambda_n t_n \|x_n\| \tag{24}$$

$$\frac{\partial}{\partial \xi_n} L(R, \beta, \boldsymbol{\xi}, \{\lambda_n\}, \{\mu_n\}) = \frac{\partial}{\partial \xi_n} \Big[C \sum_{n'=1}^{N} \xi_{n'} - \sum_{n'=1}^{N} \lambda_{n'} [t_{n'}(\beta \|x_{n'}\| - R) - 1 + \xi_{n'}] - \sum_{n'=1}^{N} \mu_{n'} \xi_{n'} \Big] \tag{25}$$

$$= C - \lambda_n - \mu_n \stackrel{!}{=} 0 \tag{26}$$

$$\lambda_n = C - \mu_n \tag{27}$$

Using the results we get:

$$L = \frac{1}{2}\beta^2 + C \sum_{n=1}^{N} \xi_n - \sum_{n=1}^{N} \lambda_n [t_n(\beta \|x_n\| - R) - 1 + \xi_n] - \sum_{n=1}^{N} \mu_n \xi_n \tag{28}$$

$$= \frac{1}{2}\beta^2 + C \sum_{n=1}^{N} \xi_n - \sum_{n=1}^{N} \lambda_n t_n \beta \|x_n\| + \sum_{n=1}^{N} \lambda_n t_n R + \sum_{n=1}^{N} \lambda_n - \sum_{n=1}^{N} \lambda_n \xi_n - \sum_{n=1}^{N} \mu_n \xi_n \tag{29}$$

$$= \frac{1}{2}\beta^2 + \sum_{n=1}^{N} (C - \mu_n) \xi_n - \sum_{n=1}^{N} \lambda_n t_n \beta \|x_n\| + \sum_{n=1}^{N} \lambda_n t_n R + \sum_{n=1}^{N} \lambda_n - \sum_{n=1}^{N} \lambda_n \xi_n \tag{30}$$

$$= \frac{1}{2}\big(\sum_{n=1}^{N} \lambda_n t_n \|x_n\|\big)^2 + \sum_{n=1}^{N} \lambda_n \xi_n - \sum_{n=1}^{N} \lambda_n t_n \big(\sum_{n=1}^{N} \lambda_n t_n \|x_n\|\big) \|x_n\| + \sum_{n=1}^{N} \lambda_n - \sum_{n=1}^{N} \lambda_n \xi_n \tag{31}$$

$$= \frac{1}{2}\big(\sum_{n=1}^{N} \lambda_n t_n \|x_n\|\big)^2 - \sum_{n=1}^{N} \lambda_n t_n \big(\sum_{n=1}^{N} \lambda_n t_n \|x_n\|\big) \|x_n\| + \sum_{n=1}^{N} \lambda_n \tag{32}$$

$$= \sum_{n=1}^{N} \lambda_n - \frac{1}{2} \sum_{n=1}^{N} \sum_{m=1}^{N} \lambda_n \lambda_m t_n t_m \|x_n\| \|x_m\| \tag{33}$$

Constraints (for all $n \in \{1, \ldots, N\}$ if not specified):

$$0 \leq \lambda_n \leq C \tag{34}$$

$$\sum_{n=1}^{N} \lambda_n t_n = 0 \tag{35}$$

2.4

$$\kappa(x_n, x_m) = \|x_n\| \|x_m\| \tag{36}$$

2.5

[Bishop p.334] $\lambda_n < C$ holds for points that lie on the margin, since this implies $\mu_n > 0$ from Eq. (27), which requires $\xi = 0$ from the KKT conditions. On each side of the decision boundary we need at least one point on the margin in order to achieve a maximum margin classifier. The minimum number of λ_n for which $0 \leq \lambda_n \leq C$ is therefore two.

2.6

The new point x^* is classified by evaluating:

$$\beta\|x^*\| - R = \sum_{n=1}^{N} \lambda_n t_n \|x_n\| \|x^*\| - R \tag{37}$$

$$= \sum_{n=1}^{N} \lambda_n t_n k(x_n, x^*) - R \tag{38}$$

If this gives a negative result, the point lies inside the circle, and if it gives a positive result, the point lies outside the circle.

2.7

The KKT conditions imply that:

$$\lambda_n = 0 \quad \text{for} \quad t_n(\beta\|x_n\| - R) - 1 + \xi_n > 0 \tag{39}$$

$$\mu_n = 0 \quad \text{for} \quad \xi_n > 0 \tag{40}$$

and

$$\text{if} \quad \lambda_n > 0 \quad \text{then} \quad t_n(\beta\|x_n\| - R) = 1 - \xi_n \tag{41}$$

$$\text{if} \quad \mu_n > 0 \quad \text{then} \quad \xi_n = 0 \tag{42}$$

So, data points that lie inside the margin will have $\lambda_n > 0$ (and $\mu_n = 0$). Data points that lie outside the margin will have $\mu > 0$ (and $\lambda_n = 0$). Points that lie on the margin will have $\lambda_n > 0$ and $\mu > 0$.

2.8

The optimal values for $\{\mu_i\}$ depend on λ_i through:

$$\mu_i = C - \lambda_i \tag{43}$$

The optimal values are given by:

$$\mu_i^* = C - \lambda_i^* \tag{44}$$

For $\lambda_i^* > 0$ and $\mu_n^* > 0$:

$$R^* = \beta^* \|x_i\| - \frac{1}{t_i}, \tag{45}$$

where

$$\beta^* = \sum_{i=1}^{N} \lambda_i^* t_i \|x_i\| \tag{46}$$

For $\lambda_i^* = 0$ and $\mu_n^* > 0$:

$$\xi_i^* = 0 \tag{47}$$

For $\lambda_i^* > 0$ and $\mu_n^* = 0$:

$$\xi_i^* = 1 - t_i(\beta^* \|x_i\| - R^*) \tag{48}$$

2.9

If we use an RBF kernel instead, we could separate the data, which we could not do with our kernel. Geometrically, the decision boundary will not resemble a circle anymore, but will look like a more complex polynomial in two dimensions.