1. $\text{Sig}(z) = \dfrac{1}{1+e^{-z}}$

$z \to -\infty$: $e^{-z} \to \infty$, $\dfrac{1}{1+e^{-z}} \to 0$

$z = 0$: $e^{0} = 1$, $\dfrac{1}{1+1} = 0.5$

$z = \infty$: $e^{-z} \to 0$, $\dfrac{1}{1+e^{-z}} = 1$

3. van Mises : $Pr(y \mid \mu, \kappa) = \dfrac{e^{\kappa \cos(y-\mu)}}{2\pi \,\text{Bessel}_0(\kappa)}$

μ: mean direction

κ: concentration. Predict $\mu := f(\underline{x} \mid \phi)$

$Pr(y \mid f(x \mid \phi), \kappa)$

$$L(\phi) = \sum_{i}^{I} -\log Pr(y_i \mid \mu, \kappa)$$

$$= -\sum_{i}^{I} \kappa \cos(y-\mu) - \log(2\pi \,\text{Bessel}_0(\kappa))$$

not dep on ϕ

$$\hat{\underline{\phi}} = \underset{\phi}{\text{argmin}} \; -\sum_{i}^{I} \kappa \cos(y - f(x \mid \phi)) - \log(2\pi \,\text{Bess}_0(\kappa))$$

$$\therefore L(\phi) = -\sum_{i}^{I} \kappa \cos(y - f(x \mid \phi))$$

For inference, take the most likely value of μ, which is what the model already outputs.

5. $Pr(y|\mu, k) = A \dfrac{e^{k\cos(y-\mu_1)}}{2\pi \,\text{Bessel}_0(k)} + (1-A) \dfrac{e^{k\cos(y-\mu_2)}}{2\pi \,\text{Bessel}_0(k)}$

$\underline{f}(x|\phi)$ needs to output:

$f_1 = \mu_1$
$f_2 = \mu_2$
$f_3 = A$ (mixing coefficient)

7. $\underline{y} \in \mathbb{R}^{10}$

$$Pr(\underline{y}|f(\underline{x}, \underline{\phi})) = \prod_i^I \prod_d^{10} \frac{e^{-(y_d - \mu_d)^2/2\sigma^2}}{\sqrt{2\pi\sigma^2}} \qquad \mu_d = f_D(\underline{x}|\underline{\phi})$$

$$\log Pr(\underline{y}|f(\underline{x}, \underline{\phi})) = \sum_i^I \sum_d^{10} -\frac{1}{2}\log(2\pi\sigma^2) - \frac{(y_d - \mu_d)^2}{2\sigma^2}$$

$$\underset{\phi}{\text{argmin}} \underbrace{-\log Pr(y|f(x,\phi))}_{NLL} = \underset{\phi}{\text{argmin}} \sum_i^I \sum_d^{10} (y_d - \mu_d)^2$$

9. The height will be a smaller magnitude than the weight, so the loss will be more focused on weight.
→ rescale the inputs
→ learn two seperate variances for both predictions.