1.1) $P(y|x,w) = \mathcal{N}(\vec{w}^T\vec{x}, \sigma) = \frac{1}{\sqrt{2\pi\sigma}} e^{\left[-\frac{(y-\vec{w}^T\vec{x})^2}{2\sigma}\right]}$

1.2) $P(w|D) = \frac{P(w)\cdot P(D|w)}{P(D)} = \frac{P(w)}{P(D)} \cdot \prod_{i=0}^{N} \mathcal{N}(\vec{w}^T\vec{x}, \sigma)$

1.3) $\hat{w} = \arg_w \max \; p(\vec{w}|D)$

$= \arg_w \min \; -\log P(\vec{w}|D)$

$= \arg_w \min \; -\log \frac{P(\vec{w})\cdot P(D|\vec{w})}{P(D)}$

$= \arg_w \min\left(-\log P(\vec{w}) - \log P(D|\vec{w}) + \cancel{\log P(D)}\right)$

* log of the products we calculated in 1.2 becomes a summation in 1.3

$= \arg_w \min\left(-\log \prod_{i=0}^{N} \mathcal{N}(\vec{w}^T\vec{x}, \sigma) - \log(P(\vec{w}))\right)$

$= \arg_w \min\left(-\sum_{i=0}^{N}\left(\cancel{\log(2\pi\sigma)^{-\frac{1}{2}}} + \log e^{-\frac{(y-\vec{w}^T\vec{x})^2}{2\sigma}}\right) - \log(P(\vec{w}))\right)$

$= \arg_w \min\left(-\sum_{i=0}^{N} -\frac{(y-\vec{w}^T\vec{x})^2}{2\sigma} - \log(P(w))\right)$

$= \arg_w \min\left(\cancel{\frac{1}{2\sigma}} \sum_{i=0}^{N}(y-\vec{w}^T\vec{x})^2 - \log \mathcal{N}(0, \alpha\vec{I})\right)$

$= \arg_w \min\left(\sum_{i=0}^{N}(y-\vec{w}^T\vec{x})^2 - \log\left(\cancel{\frac{1}{\sqrt{2\pi\alpha I^2}}} \cdot e^{\frac{1}{2\alpha I}\vec{w}^T\vec{w}}\right)\right)$

$= \arg_w \min\left(\sum_{i=0}^{N}(y-\vec{w}^T\vec{x})^2 - \frac{1}{2\alpha I} w^T w\right)$

$\approx \min\left(\sum(\vec{w}^T\vec{x}-y)^2 - \lambda\|w\|_2^2\right)$ which is RLR

1.4) because $P(D)$ drops out of min/max equations,

$$\arg\max_w p(D|w) = \arg\min_w -\log p(D|w) = \arg\min_w -\log P(w|D) - P(w)$$

$$= \arg\min_w \left(\sum (y - \vec{w}^T \vec{x})^2 \right)$$

which is the linear regression equation without the $P(w)$ which is the regularization term so therefor this is the unregularized linear regression equation.

2.1) $\vec{w}_{k+1}^T \vec{w}_{opt} \geq \vec{w}_k^T \vec{w}_{opt} + \gamma ||\vec{w}_{opt}||$

$$\left(\vec{w}_{k+1} - \vec{w}_k\right)^T \vec{w}_{opt} \geq \gamma ||\vec{w}_{opt}||$$

$$\left(\vec{w}_k + y_k \vec{x}_k - \vec{w}_k\right)^T \vec{w}_{opt} \geq \gamma ||\vec{w}_{opt}||$$

$$(y_k x_k)^T \vec{w}_{opt} \geq \gamma ||\vec{w}_{opt}||$$

Since γ is the smallest distance of any data point from the seperating plane w, we know that:

$$\left|(y_k x_k)^T \vec{w}_{opt}\right| \geq \gamma ||\vec{w}_{opt}||$$

and we know the left hand side of our equation will always be positive because y_k and x_k will always have the same sign and w_{opt} will always be positive because the classes are linearly seperable so we can then prove

$$(y_k x_k)^T \vec{w}_{opt} \geq \gamma ||\vec{w}_{opt}||$$

2.2 $$||\vec{w}_{k+1}||^2 = \vec{w}_{k+1}{}^T \vec{w}_{k+1} = (\vec{w}_k + y_k x_k)^T (w_k + y_k x_k)$$

$$= \vec{w}_k{}^2 + 2 w_k y_k x_k + y_k{}^2 x_k{}^2$$

$$\cancel{||w_k||^2} + 2 w_k y_k x_k + y_k{}^2 x_k{}^2 \leq \cancel{||w_k||^2} + 1$$

$$2 w_k y_k x_k + y_k{}^2 \cdot x_k{}^2 \leq 1$$

This has to be negative based on the problem definition

This is in range $\{-1, 1\}$ so at worst it will be $1^2 = 1$

x^2 is normalized to equal 1

in the end we have a negative number plus the product of 2 things that each max at 1 so their product also maxes at 1 which means the whole left side maxes out at less than or equal to 1

2.3) $\|w_{k+1}\| \leq \sqrt{M}$ RHS

$\|w_{k+1}\|^2 \leq M$

$\|W_k\|^2 + 1 \leq M$

$\|W_0\|^2 = 0$

$\|W_1\|^2 = 1$

$\|W_2\|^2 = 2$

$\|W_3\|^2 = 3$

from the result of 2.2, $\|W_k\|^2$ will increment by 1 each time if we start with weights of zero then $\|W_k\|^2 + 1$ has to be less than or equal to the number of mistakes.

$$\cancel{w_{opt}} \cdot M \cdot \min_x \frac{w_{opt}^T x}{\cancel{w_{opt}}} \leq \|w_{k+1}\| \cdot w_{opt}$$

$$M \cdot \min_x w_{opt}^T x \leq \|w_{k+1}\| w_{opt} \leq$$

$$M \cdot \min_x w_{opt}^T x \leq w_k^T w_{opt} + \gamma \|w_{opt}\|$$

$$M \cdot \min_x w_{opt}^T x \leq w_k^T w_{opt} + \min \frac{w_{opt} x}{\cancel{w_{opt}}} \cdot w_{opt}$$

2.3 cont')

$$w_{k+1}^T w_{opt} \geq w_k^T w_{opt} + \gamma \|w_{opt}\|$$

$$\|w_{k+1}\| \cancel{\|w_{opt}\|} \geq \|w_k\| \cancel{\|w_{opt}\|} + \gamma \cancel{\|w_{opt}\|}$$

$$\|w_{k+1}\| - \|w_k\| \geq \gamma$$

telescoping sum: $w_{k+1} - w_k + w_k - w_{k-1} + w_{k-1} - w_{k-2} \ldots w_1 - w_0$

$$\|w_{k+1}\| - \|\underbrace{w_0}_{0}\| \geq \gamma$$

$$\|w_{k+1}\| \geq \gamma \quad \text{LHS}$$

2.4) $M \times \gamma \leq \|w_{k+1}\| \leq \sqrt{M}$

$$M \times \gamma \leq \sqrt{M}$$

$$M^2 \times \gamma^2 \leq M$$

$$M^2 \leq \frac{M}{\gamma^2}$$

$$\frac{M^2}{M} \leq \gamma^{-2}$$

$$M \leq \gamma^{-2}$$

2.5) if 2 classes are not linearly seperable, this would cause 2.1's conclusion to not be provable which would mean we also could not prove 2.3 so we could not prove that the preceptron converges

3) $\hat{w} = \arg_w \max \sum_{i=1}^{N} y_i \vec{w}^T \vec{x} + \lambda (\vec{w}^T \vec{w} - 1)$

$\hat{w} = \arg_w \max \sum_{i=1}^{N} y_i \vec{w}^T \vec{x} + \lambda (||w||^2 - 1)$

$$\frac{df}{dw} = 0$$

$$\frac{df}{dw} = \sum_{i=1}^{N} y_i x_i + 2\lambda |\vec{w}| = 0$$

$$\sum_{i=1}^{N} y_i x_i = -2\lambda |\vec{w}|$$

$$\frac{\sum_{i=1}^{N} y_i x_i}{-2\lambda} = \hat{w}$$

$$\sum_{i=1}^{N} y_i x_i \propto \hat{w}$$

$$\hat{w} \propto \sum_{i: x_i \in C_1} y_i x_i + \sum_{j: x_j \in C_{-1}} y_j x_j$$

$$\hat{w} \propto \sum_{i: x_i \in C_1} x_i - \sum_{j: x_j \in C_{-1}} x_j$$