1. (a) We can decompose the inner-product of $w^\star$ and $w^{(k)}$ to get the relation

$$\langle w^\star, w^{(k)}\rangle = \langle w^\star, w^{(k-1)} + y^{(k)}x^{(k)}\rangle = \langle w^\star, w^{(k-1)}\rangle + \langle w^\star, y^{(k)}x^{(k)}\rangle \geq \langle w^\star, w^{(k-1)}\rangle + 1$$

which by induction implies that

$$\langle w^\star, w^{(k)}\rangle \geq k.$$

Also

$$\begin{aligned} &\|w^{(k)}\|^2 = \|w^{(k-1)} + y^{(k)}x^{(k)}\|^2 \leq \|w^{(k-1)}\|^2 + \|x^{(k)}\|^2 \leq \|w^{(k)}\|^2 + R^2 \\ &\implies \|w^{(k)}\|^2 \leq kR^2 \end{aligned}$$

Cauchy-Schwarz gives us that $\langle w^\star, w^{(k)}\rangle \leq \|w^\star\|\|w^{(k)}\|$. Combining all three relations we get

$$\begin{aligned} &k^2 \leq \|w^\star\|^2\|w^{(k)}\|^2 \leq B^2(kR^2) \\ &\implies k \leq (RB)^2 \end{aligned}$$

(b) Pick any $d \geq m$ and let B be some orthonormal basis for $\mathbb{R}^d$. Let $\{x_i\}$ be a sequence of unique elements in B and associate with each a label $y_i = 1$. Immediately $\max \|x_i\| \leq 1$. If we let $w^\star = \sum_i x_i$ we see that

$$\begin{aligned} &\|w^\star\|^2 = \sum_{i,j} x_i^T x_j = \sum_i x_i^T x_i = m \\ \text{and} \quad &y_i x_i^T w^\star = x_i^T x_i = 1. \end{aligned}$$

If we initialize our perceptron to $w^{(0)} = 0$ and run it on $\{x_i\}$, at each update k we will add x_k to $w^{(k-1)}$ such that $w^{(k)}$ will correctly classify x_i for any $i \leq k$ and misclassify any $i > k$. Thus it will take exactly m steps for the perceptron to converge.

2. (a) Applying stochastic gradient descent to the loss function $\ell(w) = \sum\limits_{i\in\mathcal{M}} - y_i(w^T x_i)$ gives

$$w^{(t+1)} = w^{(t)} - \eta\nabla(\ell(w))_i = w^{(t)} + \eta y_i x_i$$

where $\nabla_w \ell(w)_i$ reprsents the contribution of x_i to the gratident of the loss and where (in the last expression) we set x_i to zero if $i \notin \mathcal{M}$ for some randomly chosen sample. This the the same formula as that of the modified perceptron.

(b) Note that any separating plane specified by x_i is also specified by ηx_i. Take two perceptrons $w^{(t+1)} = w^{(t)} + y_i x_i$ and $v^{(t+1)} = v^{(t)} + \eta y_i x_i$ and let $w^{(0)} = v^{(0)} = 0$. Then $\eta w^{(1)} = \eta y_{i_1} x_{i_1} = v^{(1)}$. If we assume that $\eta w^{(t)} = v^{(t)}$ then

$$\eta w^{(t+1)} = \eta w^{(t)} + \eta y_{i_t} x_{i_t} = v^{(t)} + \eta y_{i_t} x_{i_t} = v^{(t+1)}.$$

By induction, we see that $v(t)$ is a scalar multiple of $w^{(t)}$ at every step t. Thus they allways separate the same points and both perceptions will converge after the same number of iterations.

3. (a)

$$\begin{aligned}\nabla_w J(w,x) &= -\sum_{i=1}^{n} \alpha_i(x)[y_i(1-h_w(x_i))x_i - (1-y_i)h_w(x_i)]x_i \\ &= \sum_{i=1}^{n} \alpha_i(x)(h_w(x_i)-y_i)x_i\end{aligned}$$

(b)

$$H = \sum_{i=1}^{n} \alpha_i(x)x_i \frac{\partial}{\partial w} h_w(x_i) = \sum_{i=1}^{n} \alpha_i(x)x_i[h_w(x_i)(1-h_w(x_i))]x_i^T$$

H is PSD since $\alpha_i(x)h_w(x_i)(1-h_w(x_i)) > 0$ and so it's a sum of PSD matrices.

(c) Given a query point x and some step size η the gradient descent update rule is

$$w^{(k+1)} = w^{(k)} - \eta \sum_{i=1}^{n} \alpha_i(x)(h_w(x_i)-y_i)x_i$$

(d) This is not a parametric method. We are not presupposing a specific model with parameters to estimate. Our loss function will always estimate differring w's depending on the query point we select.

4. (a) Likelihood and log-likelihood are:

$$\begin{aligned}\mathcal{L}(w) &= \prod_i p(y_i|x_i) = \frac{1}{y!}(e^{-e^{w^T x}})e^{w^T xy} \\ \log \mathcal{L}(w) &= \sum_i -\log(y!) - e^{w^T x} + yw^T x\end{aligned}$$

The corresponding optimization problem involves finding

$$\hat{w} = \arg\max_w \log \mathcal{L}(w) = \arg\max_w \sum_i -e^{w^T x} + yw^T x$$

The gradient of $\log \mathcal{L}(w)$ is

$$\nabla_w \log \mathcal{L}(w) = \sum_i -e^{w^T x_i}x_i + y_i x_i = 0$$

such that a gradient descent update step can be expresed as

$$w^{(k)} = w^{(k-1)} - \eta \sum_i (e^{(w^{(k-1)})^T x_i}x_i + y_i x_i)$$

(b) i. The gaussian distribution can be rewritten as

$$f(y) = e^{\log f(y)} = e^{-\frac{y^2 - y\mu + \mu^2}{2\sigma^2} - \frac{1}{2}\log(2\pi\sigma^2)}$$

Thus we can $f(y)$ as a GLM by using the following assignments

$$\theta = \mu,\ \phi = \sigma^2,\ a(\phi) = \phi,\ b(\phi) = 2\phi,\ c(y,\phi) = -\frac{y^2}{2\phi} - \frac{1}{2}\log(2\pi\phi)$$

ii. The Bernoulli distribution can be written as a pmf (restricted to $\{0, 1\}$) where $f(y) = p^y p^{(y-1)}$. This expresison can be rewritten as

$$f(y) = e^{y \log p + (1-y)\log(1-p)} = e^{y \log \frac{p}{1-p} + \log(1-p)}$$

The corresponding GLM parameters and equations arebsal

$$\theta = \log \frac{p}{1-p},\ \phi = 0,\ a(\phi) = 1,\ b(\theta) = \log \frac{1}{1+e^{\theta}},\ c(y, \phi) = 0$$

iii. Distribution

$$f(y) = \frac{1}{y!} e^{-\mu} \mu^y = e^{y \log \mu - \mu - \log y!}$$

GLM components

$$\theta = \log \mu,\ \phi = 0,\ a(\phi) = 1,\ b(\theta) = e^{\theta},\ c(y, \theta) = -\log y!$$

5. Let $\tilde{Y} = [Y^T\ 0]^T$ and $\tilde{X} = [X^T\ \sqrt{\lambda} I]^T$ and minimize the squared error:

$$\begin{aligned} \min\ \|\tilde{Y} - \tilde{X}\theta\|^2 &= \min\ \|\tilde{Y}\|^2 - 2\tilde{Y}\tilde{X}\theta + \|\tilde{X}\theta\|^2 \\ &= \min\ \|Y\|^2 - 2Y^T X\theta + \theta^T(X^T X + \lambda I)\theta \\ &= \min\ \|Y\|^2 - 2Y^T X\theta + \|X\theta\|^2 + \lambda\|\theta\|^2 \\ &= \min\ \|Y - X\theta\|^2 + \lambda\|\theta\|^2 \end{aligned}$$

Minimizing the last expression solves the ridge regression problem.

6. Given that our samples are i.i.d. we can represent their joint distribution as a product such that

$$\begin{aligned} &\mathcal{L}(\mu, \sigma^2) = \log P(D) = \sum_i \{-\frac{(x_i - \mu)^2}{2\sigma^2} - \log(\sqrt{2}\pi\sigma^2)\} = \sum_i \\ &\frac{\partial}{\partial u}\mathcal{L} = \sum_i -\frac{(x_i - \mu)^2}{\sigma^2} = 0 \\ &\frac{\partial}{\partial \sigma^2} = -\frac{N}{2\sigma^2} + \frac{1}{2\sigma^2}\sum_i (xi - \mu)^2 \end{aligned}$$

Setting both derivatives equal to zero gives us

$$\hat{\mu} = \frac{1}{N}\sum_i x_i \quad \text{and} \quad \hat{\sigma}^2 = \frac{1}{N}\sum_i (x_i - \hat{\mu})^2$$