This handout includes space for every question that requires a written response. Please feel free to use it to handwrite your solutions (legibly, please). If you choose to typeset your solutions, the `README.md` for this assignment includes instructions to regenerate this handout with your typeset LaTeX solutions.

1.a

Since $g'(z) = g(z)(1 - g(z))$ and $h(x) = g(\theta^T x)$, it follows that $\partial h(x)/\partial \theta_k = h(x)(1 - h(x))x_k$.
Letting $h_\theta(x^{(i)}) = g(\theta^T x^{(i)}) = 1/(1 + \exp(-\theta^T x^{(i)}))$, we have

$$\frac{\partial \log h_\theta(x^{(i)})}{\partial \theta_k} = \frac{\partial \log h_\theta(x^{(i)})}{\partial h_\theta(x^{(i)})} \frac{\partial h_\theta(x^{(i)})}{\partial \theta_k} = \frac{1}{h_\theta(x^{(i)})} \frac{\partial g(\theta^T x^{(i)})}{\partial \theta_k} = \frac{x_k^{(i)} g(\theta^T x^{(i)})(1 - g(\theta^T x^{(i)}))}{h_\theta(x^{(i)})} = x_k^{(i)}(1 - h_\theta(x^{(i)}))$$

$$\frac{\partial \log(1 - h_\theta(x^{(i)}))}{\partial \theta_k} = -\frac{1}{(1 - h_\theta(x^{(i)}))} \frac{\partial h_\theta(x^{(i)})}{\partial \theta_k} = -\frac{1}{(1 - h_\theta(x^{(i)}))} \frac{\partial g(\theta^T x^{(i)})}{\partial \theta_k} = -\frac{x_k^{(i)} g(\theta^T x^{(i)})(1 - g(\theta^T x^{(i)}))}{(1 - h_\theta(x^{(i)}))} = -x_k^{(i)} h_\theta(x^{(i)})$$

Substituting into our equation for $J(\theta)$, we have

$$\frac{\partial J(\theta)}{\partial \theta_k} = \frac{1}{n} \sum_{i=1}^{n} y^{(i)} \frac{\partial log(h_\theta(x^{(i)}))}{\partial \theta_k} + (1 - y^{(i)}) \frac{\partial log(1 - h_\theta(x^{(i)}))}{\partial \theta_k} =$$

$$= \frac{1}{n} \sum_{i=1}^{n} y^{(i)} x_k^{(i)}(1 - h_\theta(x^{(i)})) - (1 - y^{(i)}) x_k^{(i)} h_\theta(x^{(i)}) = -\frac{1}{n} \sum_{i=1}^{n} (y^{(i)} - h_\theta(x^{(i)})) x_k^{(i)}$$

Consequently, the (k, l) entry of the Hessian is given by

$$H_{kl} = \frac{\partial^2 J(\theta)}{\partial \theta_k \partial \theta_l} = -\frac{1}{n} \sum_{i=1}^{n} \frac{\partial (y^{(i)} - h_\theta(x^{(i)})) x_k^{(i)}}{\partial \theta_l} = \frac{1}{n} \sum_{i=1}^{n} x_k^{(i)} \frac{\partial h_\theta(x^{(i)})}{\partial \theta_l} = \frac{1}{n} \sum_{i=1}^{n} h_\theta(x^{(i)})(1 - h_\theta(x^{(i)})) x_k^{(i)} x_l^{(i)}$$

Using the fact that $X_{ij} = x_i x_j$ if and only if $X = xx^T$, we have

$$H = \frac{1}{n} \sum_{i=1}^{n} h_\theta(x^{(i)})(1 - h_\theta(x^{(i)})) x^{(i)} x^{(i)T} =$$

To prove that H is positive semi-definite, show $z^T H z \geq 0$ for all $z \in \mathbb{R}^d$.

$$z^T H z = z^T \left(\frac{1}{n} \sum_{i=1}^{n} h_\theta(x^{(i)})(1 - h_\theta(x^{(i)})) x^{(i)} x^{(i)T} \right) z =$$

$$= \frac{1}{n} \sum_{i=1}^{n} h_\theta(x^{(i)})(1 - h_\theta(x^{(i)}))(z^T x^{(i)} x^{(i)T} z) =$$

$$= \frac{1}{n} \sum_{i=1}^{n} h_\theta(x^{(i)})(1 - h_\theta(x^{(i)}))(z^T x^{(i)})^2$$

The scalar $(z^T x^{(i)})^2 \geq 0$, and $h^{(i)}(1 - h^{(i)}) \geq 0$, therefore for $\forall z \in \mathbb{R}^d$, $z^T H z \geq 0$ and $H \succeq 0$.

1.c

For shorthand, we let $\mathcal{H} = \{\phi, \Sigma, \mu_0, \mu_1\}$ denote the parameters for the problem. Since the given formulae are conditioned on y, use Bayes rule to get:

$$p(y=1|x;\mathcal{H}) = \frac{p(x|y=1;\mathcal{H})p(y=1;\mathcal{H})}{p(x;\mathcal{H})} = \frac{p(x|y=1;\mathcal{H})p(y=1;\mathcal{H})}{p(x|y=1;\mathcal{H})p(y=1;\mathcal{H}) + p(x|y=0;\mathcal{H})p(y=0;\mathcal{H})} =$$
$$= \frac{\phi \exp(-\frac{1}{2}(x-\mu_1)^T\Sigma^{-1}(x-\mu_1))}{\phi \exp(-\frac{1}{2}(x-\mu_1)^T\Sigma^{-1}(x-\mu_1)) + (1-\phi)\exp(-\frac{1}{2}(x-\mu_0)^T\Sigma^{-1}(x-\mu_0))} =$$
$$= \frac{1}{1 + \frac{(1-\phi)}{\phi}\exp(\frac{1}{2}((x-\mu_1)^T\Sigma^{-1}(x-\mu_1) - (x-\mu_0)^T\Sigma^{-1}(x-\mu_0))}$$

Let's simplify the expression in exponent, assuming Σ is symmetric and positive definite and therefore Σ^{-1} is also symmetric:

$$\frac{1}{2}(x-\mu_1)^T\Sigma^{-1}(x-\mu_1) - (x-\mu_0)^T\Sigma^{-1}(x-\mu_0) = \frac{1}{2}(\cancel{x^T\Sigma^{-1}x} - 2x^T\Sigma^{-1}\mu_1 + \mu_1^T\Sigma\mu_1 - \cancel{x^T\Sigma^{-1}x} + 2x^T\Sigma^{-1}\mu_0 - \mu_0^T\Sigma^{-1}\mu_0) =$$
$$= -x^T\Sigma^{-1}(\mu_1 - \mu_0) + \frac{1}{2}(\mu_1^T\Sigma^{-1}\mu_1 - \mu_0^T\Sigma^{-1}\mu_0) =$$

Let's set $\theta = \Sigma^{-1}(\mu_1 - \mu_0)$ and $\theta_0 = -\frac{1}{2}\mu_1^T\Sigma^{-1}\mu_1 + \frac{1}{2}\mu_0^T\Sigma^{-1}\mu_0 + \log(\frac{\phi}{1-\phi}) = \log(\frac{\phi}{1-\phi}) - \frac{1}{2}(\mu_1+\mu_0)^T\Sigma^{-1}(\mu_1-\mu_0)$, so the final expression will be:

$$p(y=1|x) = \frac{1}{1+\exp(-(\theta^T x + \theta_0))}$$

1.d

First, derive the expression for the log-likelihood of the training data:

$$\begin{aligned}\ell(\phi, \mu_0, \mu_1, \Sigma) &= \log \prod_{i=1}^{n} p(x^{(i)}|y^{(i)}; \mu_0, \mu_1, \Sigma) p(y^{(i)}; \phi) \\ &= \sum_{i=1}^{n} \log p(x^{(i)}|y^{(i)}; \mu_0, \mu_1, \Sigma) + \sum_{i=1}^{n} \log p(y^{(i)}; \phi) = \\ &= \sum_{i=1}^{n} \Big(y^{(i)} \log p(x^{(i)}|y^{(i)} = 1; \mu_0, \mu_1, \Sigma) + (1 - y^{(i)}) \log p(x^{(i)}|y^{(i)} = 0; \mu_0, \mu_1, \Sigma) \Big) + \\ &\sum_{i=1}^{n} \Big(y^{(i)} \log \phi + (1 - y^{(i)}) \log(1 - \phi) \Big)\end{aligned}$$

Now, the likelihood is maximized by setting the derivative (or gradient) with respect to each of the parameters to zero.

For ϕ:

$$\frac{\partial \ell}{\partial \phi} = \sum_{i=1}^{n} \frac{\partial (y^{(i)} \log \phi + (1 - y^{(i)}) \log(1 - \phi))}{\partial \phi} = \sum_{i=1}^{n} \left(\frac{y^{(i)}}{\phi} - \frac{(1 - y^{(i)})}{(1 - \phi)} \right)$$

Setting this equal to zero and solving for ϕ gives the maximum likelihood estimate.

$$\sum_{i=1}^{n} \left(\frac{y^{(i)}}{\phi} - \frac{(1 - y^{(i)})}{(1 - \phi)} \right) = \frac{1}{\phi} \sum_{i=1}^{n} y^{(i)} - \frac{1}{1 - \phi} \sum_{i=1}^{n} (1 - y^{(i)}) = 0$$

Let's assume n is a total number of $y^{(i)}$ in the dataset, n_0 is a number of $y^{(i)} = 0$ and n_1 is a number of $y^{(i)} = 1$, $n_1 = \sum_{i=1}^{n} 1\{y^{(i)} = 1\}$, $n = n_0 + n_1$, then

$$\frac{n_1}{\phi} - \frac{n_0}{(1 - \phi)} = \frac{n_1}{\phi} - \frac{n - n_1}{(1 - \phi)} = 0$$

$$n_1(1 - \phi) = (n - n_1)\phi$$

$$n_1 - n_1\phi = n\phi - n_1\phi$$

$$\phi = \frac{n_1}{n} = \frac{1}{n} \sum_{i=1}^{n} y^{(i)} = \frac{1}{n} \sum_{i=1}^{n} 1\{y^{(i)} = 1\}$$

For μ_0:

Hint: Remember that Σ (and thus Σ^{-1}) is symmetric. Let's skip part of ℓ depending on ϕ. Also, the part that depend on μ_1 is crossed, since in it there are no components depending on μ_0:

$$\begin{aligned}\nabla_{\mu_0} \ell &= \nabla_{\mu_0} \sum_{i=1}^{n} \Big(y^{(i)} \log p(x^{(i)}|y^{(i)} = 1; \mu_0, \mu_1, \Sigma) + (1 - y^{(i)}) \log p(x^{(i)}|y^{(i)} = 0; \mu_0, \mu_1, \Sigma) \Big) \\ &= \sum_{i=1}^{n} y^{(i)} \nabla_{\mu_0} \log \left(\frac{\exp(-\frac{1}{2}(x^{(i)} - \mu_1)^T \Sigma^{-1} (x^{(i)} - \mu_1))}{(2\pi)^{d/2} |\Sigma|^{1/2}} \right) + \\ &+ \sum_{i=1}^{n} (1 - y^{(i)}) \nabla_{\mu_0} \log \left(\frac{\exp(-\frac{1}{2}(x^{(i)} - \mu_0)^T \Sigma^{-1} (x^{(i)} - \mu_0))}{(2\pi)^{d/2} |\Sigma|^{1/2}} \right)\end{aligned}$$

Let's further simplify it:

$$\nabla_{\mu_0}\ell = \sum_{i=1}^{n}(1-y^{(i)})\nabla_{\mu_0}\left(-\frac{1}{2}(x^{(i)}-\mu_0)^T\Sigma^{-1}(x^{(i)}-\mu_0) - \frac{d}{2}\log(2\pi) - \frac{1}{2}\log|\Sigma|)\right) =$$
$$= \sum_{i=1}^{n}(1-y^{(i)})\nabla_{\mu_0}\left(-\frac{1}{2}(x^{(i)}-\mu_0)^T\Sigma^{-1}(x^{(i)}-\mu_0)\right) =$$
$$= \sum_{i=1}^{n}(1-y^{(i)})\Sigma^{-1}(x^{(i)}-\mu_0)$$

Since $(1-y^{(i)})$ is a scalar, and Σ^{-1} is a constant symmetric matrix, we could re-write the result as:

$$\nabla_{\mu_0}\ell = \Sigma^{-1}\sum_{i=1}^{n}(1-y^{(i)})(x^{(i)}-\mu_0)$$

Setting this gradient to zero gives the maximum likelihood estimate for μ_0.

$$\sum_{i=1}^{n}(1-y^{(i)})(x^{(i)}-\mu_0) = 0$$
$$\sum_{i=1}^{n}1\{y^{(i)}=0\}\mu_0 = \sum_{i=1}^{n}1\{y^{(i)}=0\}x^{(i)}$$
$$\mu_0 = \frac{\sum_{i=1}^{n}1\{y^{(i)}=0\}x^{(i)}}{\sum_{i=1}^{n}1\{y^{(i)}=0\}}$$

For μ_1:

Hint: Remember that Σ (and thus Σ^{-1}) is symmetric. Similar reasoning as above works for $\nabla_{\mu_1}\ell$ with only difference that it's valid for $y^{(i)}=1$ instead of $y^{(i)}=0$:

$$\nabla_{\mu_1}\ell = \Sigma^{-1}\sum_{i=1}^{n}y^{(i)}(x^{(i)}-\mu_1)$$

Setting this gradient to zero gives the maximum likelihood estimate for μ_1.

$$\sum_{i=1}^{n}y^{(i)}(x^{(i)}-\mu_1) = 0$$
$$\sum_{i=1}^{n}1\{y^{(i)}=1\}\mu_1 = \sum_{i=1}^{n}1\{y^{(i)}=1\}x^{(i)}$$
$$\mu_1 = \frac{\sum_{i=1}^{n}1\{y^{(i)}=1\}x^{(i)}}{\sum_{i=1}^{n}1\{y^{(i)}=1\}}$$

For Σ, we find the gradient with respect to $S=\Sigma^{-1}$ rather than Σ just to simplify the derivation (note that $|S| = \frac{1}{|\Sigma|}$). You should convince yourself that the maximum likelihood estimate S_n found in this way would correspond to the actual maximum likelihood estimate Σ_n as $S_n^{-1} = \Sigma_n$.

Hint: You may need the following identities:

$$\nabla_S|S| = |S|(S^{-1})^T$$
$$\nabla_S b_i^T S b_i = \nabla_S tr\left(b_i^T S b_i\right) = \nabla_S tr\left(S b_i b_i^T\right) = b_i b_i^T$$

$$\begin{aligned}
\nabla_S \ell &= \sum_{i=1}^{n} 1\{y^{(i)}=0\}\left(\nabla_S(-\frac{1}{2}(x^{(i)}-\mu_0)^T S(x^{(i)}-\mu_0)) - \nabla_S \frac{d}{2}\log(2\pi) + \frac{1}{2}\nabla_S \log|S|\right) + \\
&+ \sum_{i=1}^{n} 1\{y^{(i)}=1\}\left(\nabla_S(-\frac{1}{2}(x^{(i)}-\mu_1)^T S(x^{(i)}-\mu_1)) - \nabla_S \frac{d}{2}\log(2\pi) + \frac{1}{2}\nabla_S \log|S|\right) = \\
&= \sum_{i=1}^{n} 1\{y^{(i)}=0\}\left(-\frac{1}{2}(x^{(i)}-\mu_0)(x^{(i)}-\mu_0)^T + \frac{|S|(S^{-1})}{2|S|}\right) + \\
&+ \sum_{i=1}^{n} 1\{y^{(i)}=1\}\left(-\frac{1}{2}(x^{(i)}-\mu_1)(x^{(i)}-\mu_1)^T + \frac{|S|(S^{-1})}{2|S|}\right) = \\
&= \frac{n}{2}S^{-1} - \frac{1}{2}\sum_{i=1}^{n}(x^{(i)}-\mu_{y^{(i)}})(x^{(i)}-\mu_{y^{(i)}})^T
\end{aligned}$$

Next, substitute $\Sigma = S^{-1}$. Setting this gradient to zero gives the required maximum likelihood estimate for Σ.

$$\begin{aligned}
\frac{n}{2}S^{-1} &= \frac{1}{2}\sum_{i=1}^{n}(x^{(i)}-\mu_{y^{(i)}})(x^{(i)}-\mu_{y^{(i)}})^T \\
\Sigma &= \frac{1}{n}\sum_{i=1}^{n}(x^{(i)}-\mu_{y^{(i)}})(x^{(i)}-\mu_{y^{(i)}})^T
\end{aligned}$$

1.f

For dataset 1 we can see that GDA separates the classes with an almost vertical line. It's not very effective. Logistic regression line better separates the classes. In the plot the classes are distributed in a non-gaussian way and there are outliers, that distort mean and covariance estimates. For such distribution, logistic regression is a more appropriate choice.

1.g

For dataset 2 GDA performs much better, similar to logistic regression. This is a consequence of data distribution as in dataset 2 the fields are roughly equal in area, clearly separated - have small intersection.

1.h

If we make $x_{(i)}$ to be distributed more normally in dataset 1 and eliminate outliers, then GDA will work better.

2.c

Let's use Bayes theorem:

$$p(t=1|y=1,x) = \frac{p(y=1|t=1,x)\ p(t=1|x)}{p(y=1|x)}$$

We assume that if $t=0$ then $y=1$ never occurs:

$$p(y=1|t=0,x) = 0$$

We also assume that if $t=1$ then $y=1$ with some non zero probability. Rewriting the denominator of the first formula:

$$\begin{aligned} p(y=1|x) &= p(y=1|t=1,x)\ p(t=1|x) + p(y=1|t=0,x)\ p(t=0|x) = \\ &= p(y=1|t=1,x)\ p(t=1|x) + 0 \cdot p(t=0|x) = \\ &= p(y=1|t=1,x)\ p(t=1|x) \end{aligned}$$

Therefore the first formula becomes:

$$p(t=1|y=1,x) = \frac{p(y=1|t=1,x)\ p(t=1|x)}{p(y=1|x)} = \frac{p(y=1|t=1,x)\ p(t=1|x)}{p(y=1|t=1,x)\ p(t=1|x)} = 1$$

2.d

Assuming $p(y = 1|x) = \alpha\ p(t = 1|x)$ and $p(y = 1|t = 0, x) = 0$ let's use the law of total probability:

$$\begin{aligned} p(y = 1|x) &= p(y = 1|t = 1, x)\ p(t = 1|x) + p(y = 1|t = 0, x)\ p(t = 0|x) = \\ &= \alpha\ p(t = 1|x) + 0 \cdot p(t = 0|x) = \\ &= \alpha\ p(t = 1|x) \end{aligned}$$

Therefore:

$$p(t = 1|x) = \frac{1}{\alpha}\ p(y = 1|x)$$

2.e

We have as an input:

$$h(x) = p(y = 1|x)$$

We also know that:

$$p(y = 1|x) = \alpha \; p(t = 1|x)$$

Therefore:

$$h(x) = \alpha \; p(t = 1|x)$$

Our assumption is that $p(t^{(i)} = 1|x^{(i)}) \in \{0, 1\}$, therefore:

$$h(x) = \begin{cases} \alpha, & \text{if } t = 1 \\ 0, & \text{if } t = 0 \end{cases} \tag{1}$$

Let's compute $\mathbb{E}$

$$\mathbb{E}[h(x)|y = 1] = \int h(x) \; p(x|y = 1)dx \tag{2}$$

taking (1) into account, the expression (2) becomes the average of α over x points where $y = 1$, therefore:

$$\mathbb{E}[h(x)|y = 1] = \alpha \; p(t = 1|y = 1)$$

From previous result $p(t = 1|y = 1, x) = 1$, consequently:

$$\mathbb{E}[h(x)|y = 1] = \alpha \cdot 1 = \alpha$$

Finally

$$\alpha = \mathbb{E}[h(x^{(i)})|y^{(i)} = 1]$$

If we assume V_+ as a set of examples where $y^{(i)} = 1$, then:

$$\alpha \approx \frac{1}{|V_+|} \sum_{x^{(i)} \in V_+} h(x^{(i)})$$

3.ai

Let's create a trivial classifier $f(x)$ that always predicts the negative class label 0, regardless of x. It correctly classifies all the negative examples with label 0 but incorrectly classifies positive examples with label 1. The dataset has ρ positive examples and $1-\rho$ negative examples. There are n examples in the dataset. So:

$$A = \frac{\#\text{ correct predictions}}{\#\text{ all examples}} = \frac{(1-\rho)\cdot n}{n} = 1-\rho$$

3.aii

We have the following definition of accuracy:

$$A = \frac{TP+TN}{TP+TN+FP+FN} \tag{3}$$

The number of positive examples in the data set could be expressed as $TP+FN$. We sum up only the positive examples. The following subsets make up all the positive examples in the dataset:

$$TP \triangleq \#\text{ positive examples with a correct (positive) prediction}$$
$$FN \triangleq \#\text{ positive examples with a incorrect (negative) prediction}$$

The overall amount of examples is expressed as $TP+TN+FP+FN=n$. The number of positive examples n_+ could be expressed as follows:

$$\#\text{ positive examples} = n_+ = \rho\ n = \rho\ (TP+TN+FP+FN)$$

From the other side:

$$n_+ = TP+FN$$

Therefore:

$$n_+ = TP+FN = \rho\ (TP+TN+FP+FN)$$
$$\rho = \frac{TP+FN}{TP+TN+FP+FN}$$

Similarly to that, let's write the formula for # negative examples, n_-:

$$n_- = TN+FP = (1-\rho)\ (TP+TN+FP+FN)$$
$$1-\rho = \frac{TN+FP}{TP+TN+FP+FN}$$

We take into account that:

$$A_0 = \frac{TN}{TN+FP} => TN = A_0(TN+FP)$$
$$A_1 = \frac{TP}{TP+FN} => TP = A_1(TP+FN)$$

Let's write the expression for A taking these formulae into account, also remember expressions with ρ and $1-\rho$ written earlier:

$$A = \frac{TP+TN}{TP+TN+FP+FN} = \frac{A_1(TP+FN)+A_0(TN+FP)}{TP+TN+FP+FN} =$$
$$= \frac{A_1\rho(TP+TN+FP+FN)+A_0(1-\rho)(TP+TN+FP+FN)}{TP+TN+FP+FN} =$$
$$= A_1\rho + A_0(1-\rho)$$

3.aiii

We define the balanced accuracy as:

$$\overline{A} = \frac{1}{2}(A_0 + A_1)$$

The mentioned trivial classifier predicts 0 for any input x, therefore we correctly predict all negative examples, therefore the following equalities for negative examples are correct:

$$FN = n_+ = \#\text{ positive examples predicted as negative}$$
$$TN = n_- = \#\text{ negative examples}$$

For positive part of the dataset we have the following:

$$FP = 0$$
$$TP = 0$$

Therefore, provided that there is at least one negative example in the dataset, the balanced accuracy becomes:

$$\overline{A} = \frac{1}{2}(A_0 + A_1) = \frac{1}{2}(\frac{TN}{TN + FP} + \frac{TP}{TP + FN}) = \frac{1}{2}(\frac{TN}{TN + 0} + \frac{0}{0 + FN}) = \frac{1}{2}(1 + 0) = \frac{1}{2}$$

3.c

The balanced accuracy definition:

$$\overline{A} = \frac{1}{2}(A_0 + A_1)$$

We assume that $\rho < \frac{1}{2}$ and $\kappa = \frac{\rho}{1-\rho}$, $\frac{1}{\kappa} \in \mathbb{Z}$

Let's write the expressions for positive, negative examples and κ in the dataset D as:

$$\#\text{positive examples} = n_+ = \rho\ n$$
$$\#\text{negative examples} = n_- = (1 - \rho)\ n$$
$$\kappa = \frac{\rho}{1 - \rho} = \frac{n_+}{n_-}$$

Let's create dataset $D^{'}$ as:

$$\#\text{negative examples in dataset } D' = n_-^{'} = n_- = (1 - \rho)\ n$$
$$\#\text{positive examples in dataset } D' = n_+^{'} = \frac{n_+}{\kappa} = \frac{\rho\ n}{\kappa} = n_-$$

Overall amount of entries in dataset D' equals to $2n_-$ examples with n_- negative examples and n_- positives examples. Taking this into account, let's write the balanced accuracy for dataset D:

$$n_+ = TP + FN$$
$$n_- = TN + FP$$
$$\overline{A} = \frac{1}{2}(\frac{TP}{n_=} + \frac{TN}{n_+})$$

Let's write the balanced accuracy $A^{'}$ for dataset D', where TP' and FN' are true positives and false negatives for dataset D':

$$
\begin{aligned}
n_{+}^{'} &= n_{-} = TP' + FN' \\
n_{-}^{'} &= n_{-} = TN + FP \\
TP' &= \frac{TP}{\kappa} \\
\overline{A'} &= \frac{1}{2}(\frac{TN}{n_{-}^{'}} + \frac{TP'}{n_{+}^{'}}) = \frac{1}{2}(\frac{TN}{n_{-}} + \frac{TP}{\kappa\, n_{-}}) = \frac{1}{2}(\frac{TN}{n_{-}} + \frac{TP}{n_{+}}) = \overline{A}
\end{aligned}
$$

Let's derive the average empirical loss for logistic regression on dataset D'. The loss per sample:

$$\ell(x^{(i)}, y^{(i)}) = -[y^{(i)} \log h_\theta(x^{(i)}) + (1 - y^{(i)}) \log(1 - h_\theta(x^{(i)})]$$

The total loss $J_{total}(\theta)$:

$$J_{total}(\theta) = \sum_{i=1}^{n} \ell(x^{(i)}, y^{(i)}) = \sum_{i=1, y^{(i)}=0}^{n_-} \ell(x^{(i)}, y^{(i)}) + \frac{1}{\kappa} \sum_{i=1, y^{(i)}=1}^{n_-} \ell(x^{(i)}, y^{(i)})$$

Given that:

$$\omega^{(i)} = \begin{cases} 1, & \text{if } y^{(i)} = 0 \\ \frac{1}{\kappa}, & \text{if } y^{(i)} = 1 \end{cases}$$

The total loss becomes:

$$J_{total}(\theta) = \sum_{i=1}^{n} \omega^{(i)} \ell(x^{(i)}, y^{(i)})$$

The average loss $\overline{L}$ will be:

$$J(\theta) = \frac{1}{n} \sum_{i=1}^{n} \omega^{(i)} \ell(x^{(i)}, y^{(i)})$$

As we know, $n_- = (1 - \rho)n$ and $\kappa = \frac{\rho}{1-\rho}$. Let's make an expression for $\rho(\kappa)$:

$$
\begin{aligned}
\kappa &= \frac{\rho}{1 - \rho} \\
\kappa - \kappa\rho - \rho &= 0 \\
\kappa &= \rho(\kappa + 1) \\
\rho &= \frac{\kappa}{\kappa + 1}
\end{aligned}
$$

Therefore, taking into account that $n = 2n_- = 2(1 - \rho)n$

$$
\begin{aligned}
J(\theta) &= \frac{1}{2n(1-\rho)} \sum_{i=1}^{n} \omega^{(i)} \ell(x^{(i)}, y^{(i)}) = \frac{1}{2n(1 - \frac{k}{k+1})} \sum_{i=1}^{n} \omega^{(i)} \ell(x^{(i)}, y^{(i)}) = \\
&= \frac{k+1}{2n} \sum_{i=1}^{n} \omega^{(i)} \ell(x^{(i)}, y^{(i)}) = \frac{k+1}{2n} \sum_{i=1}^{n} \omega^{(i)} \ell(x^{(i)}, y^{(i)}) = \\
&= -\frac{k+1}{2n} \sum_{i=1}^{n} \omega^{(i)} [y^{(i)} \log h_\theta(x^{(i)}) + (1 - y^{(i)}) \log(1 - h_\theta(x^{(i)})]
\end{aligned}
$$