1.1

$$\mathbf{y}_1 = \exp\big(\mathcal{G}(\mathbf{x}_2)\big) \circ \mathbf{x}_1 + \mathcal{F}(\mathbf{x}_2)$$

$$\mathbf{y}_1 - \mathcal{F}(\mathbf{x}_2) = \exp\big(\mathcal{G}(\mathbf{x}_2)\big) \circ \mathbf{x}_1$$

$$\mathbf{x}_1 = \big(\mathbf{y}_1 - \mathcal{F}(\mathbf{x}_2)\big) / \exp\big(\mathcal{G}(\mathbf{x}_2)\big)$$

$$\mathbf{y}_2 = \exp(\mathbf{s}) \circ \mathbf{x}_2$$

$$\mathbf{x}_2 = \mathbf{y}_2 / \exp(\mathbf{s})$$

1.2

$$\frac{\partial \mathbf{y}}{\partial \mathbf{x}} = \begin{bmatrix} \frac{\partial \mathbf{y}_1}{\partial \mathbf{x}_1} & \frac{\partial \mathbf{y}_1}{\partial \mathbf{x}_2} \\ \frac{\partial \mathbf{y}_2}{\partial \mathbf{x}_1} & \frac{\partial \mathbf{y}_2}{\partial \mathbf{x}_2} \end{bmatrix}$$

Where

$$\frac{\partial \mathbf{y}_1}{\partial \mathbf{x}_1} = \exp(\mathcal{G}(\mathbf{x}_2))$$

$$\frac{\partial \mathbf{y}_1}{\partial \mathbf{x}_2} = \exp(\mathcal{G}(\mathbf{x}_2)) \circ \frac{\partial \mathcal{G}}{\partial \mathbf{x}_2} \circ \mathbf{x}_1 + \frac{\partial \mathcal{F}}{\partial \mathbf{x}_2}$$

$$\frac{\partial \mathbf{y}_2}{\partial \mathbf{x}_1} = 0$$

$$\frac{\partial \mathbf{y}_2}{\partial \mathbf{x}_2} = \exp(\mathbf{s})$$

1.3

$$\det\left(\frac{\partial \mathbf{y}}{\partial \mathbf{x}}\right) = \exp(\mathcal{G}(\mathbf{x}_2)) \circ \exp(\mathbf{s})$$

2.1

First, notice $\tilde{a}$ can only equal to 0 or 1.

Then, note that from equation (4), we can represent $f(\tilde{a}) = \tilde{a}$

Therefore, $g[\theta, \tilde{a}] = \tilde{a} \frac{\partial}{\partial \theta} \log p(a = \tilde{a}|\theta)$

For $\tilde{a} = 0$, $g[\theta, \tilde{a}] = 0$

For $\tilde{a} = 1$:

Note that since $\tilde{a} = 1$, we know $p(a = \tilde{a}|\theta) = \mu$ from equation (2). Also note that the derivative of sigmoid is sigmoid*(1-sigmoid)

$$g[\theta, \tilde{a}] = \frac{\partial}{\partial \theta}(\log \mu) = \frac{1}{\mu}\frac{\partial \mu}{\partial \theta} = \frac{1}{\mu}\mathbf{x}\sigma' = \frac{1}{\mu}\mathbf{x}\mu(1-\mu) = \mathbf{x}(1-\mu)$$

To combine both cases of $\tilde{a}$ we can express it as: $g[\theta, \tilde{a}] = \tilde{a}\mathbf{x}(1-\mu)$

2.2

Using the hint given, and we know both μ and $\mathbf{x}$ being constants.

$$Var(g[\theta, \tilde{a}]_1) = Var\big(\tilde{a}\mathbf{x}(1-\mu)\big) = \mathbf{x}^2(1-\mu)^2 Var(\tilde{a}) = \mathbf{x}^2\mu(1-\mu)^3$$

2.3

Note that when μ approaches 1, the variance from 2.2 will be very close to 0. With a very low variance, the model will be learning slowly because the model will be learning the same thing over and over since the weights does not change as much.

3.1

First, we find $\nabla_{\boldsymbol{w_t}} J$

$$\nabla_{\boldsymbol{w_t}} J = \frac{\partial}{\partial \boldsymbol{w}_t}\left(\frac{1}{2}\|\boldsymbol{T}^{\pi}\boldsymbol{X}\bar{\boldsymbol{w}}_{\boldsymbol{t}} - \boldsymbol{X}\boldsymbol{w_t}\|_2^2\right)$$

$$= -\boldsymbol{X}^{\mathrm{T}}(\boldsymbol{T}^{\pi}\boldsymbol{X}\bar{\boldsymbol{w}}_{\boldsymbol{t}} - \boldsymbol{X}\boldsymbol{w_t}) = -\boldsymbol{X}^{\mathrm{T}}((\boldsymbol{r} + \gamma\boldsymbol{P}^{\pi})\boldsymbol{X}\bar{\boldsymbol{w}}_{\boldsymbol{t}} - \boldsymbol{X}\boldsymbol{w_t})$$

So, the update rule would be:

$$\boldsymbol{w_{t+1}} \leftarrow \boldsymbol{w_t} + \alpha\boldsymbol{X}^{\mathrm{T}}((\boldsymbol{r} + \gamma\boldsymbol{P}^{\pi})\boldsymbol{X}\bar{\boldsymbol{w}}_{\boldsymbol{t}} - \boldsymbol{X}\boldsymbol{w_t})$$