Understanding Deep Learning Equations

February 23, 2025

Chapter 1

Introduction

Chapter 2

Supervised Learning

$$\mathbf{y} = \mathbf{f}[\mathbf{x}]. \tag{2.1}$$

$$\mathbf{y} = \mathbf{f}[\mathbf{x}, \boldsymbol{\phi}]. \tag{2.2}$$

$$\hat{\boldsymbol{\phi}} = \underset{\boldsymbol{\phi}}{\mathrm{argmin}}\Big[L\left[\boldsymbol{\phi}\right]\Big]. \tag{2.3}$$

$$\begin{aligned} y &= \mathrm{f}[x, \boldsymbol{\phi}] \\ &= \phi_0 + \phi_1 x. \end{aligned} \tag{2.4}$$

$$\begin{aligned} L[\boldsymbol{\phi}] &= \sum_{i=1}^{I} \left(\mathrm{f}[x_i, \boldsymbol{\phi}] - y_i\right)^2 \\ &= \sum_{i=1}^{I} \left(\phi_0 + \phi_1 x_i - y_i\right)^2. \end{aligned} \tag{2.5}$$

$$\begin{aligned} \hat{\boldsymbol{\phi}} &= \underset{\boldsymbol{\phi}}{\mathrm{argmin}}\Big[L[\boldsymbol{\phi}]\Big] \\ &= \underset{\boldsymbol{\phi}}{\mathrm{argmin}}\left[\sum_{i=1}^{I} \left(\mathrm{f}[x_i, \boldsymbol{\phi}] - y_i\right)^2\right] \\ &= \underset{\boldsymbol{\phi}}{\mathrm{argmin}}\left[\sum_{i=1}^{I} \left(\phi_0 + \phi_1 x_i - y_i\right)^2\right]. \end{aligned} \tag{2.6}$$

Chapter 3

Shallow neural networks

$$\begin{aligned} y &= \mathrm{f}[x, \boldsymbol{\phi}] \\ &= \phi_0 + \phi_1 \mathrm{a}[\theta_{10} + \theta_{11}x] + \phi_2 \mathrm{a}[\theta_{20} + \theta_{21}x] + \phi_3 \mathrm{a}[\theta_{30} + \theta_{31}x]. \end{aligned} \tag{3.1}$$

$$\mathrm{a}[z] = \mathrm{ReLU}[z] = \begin{cases} 0 & z < 0 \\ z & z \geq 0 \end{cases}. \tag{3.2}$$

$$\begin{aligned} h_1 &= \mathrm{a}[\theta_{10} + \theta_{11}x] \\ h_2 &= \mathrm{a}[\theta_{20} + \theta_{21}x] \\ h_3 &= \mathrm{a}[\theta_{30} + \theta_{31}x], \end{aligned} \tag{3.3}$$

$$y = \phi_0 + \phi_1 h_1 + \phi_2 h_2 + \phi_3 h_3. \tag{3.4}$$

$$h_d = \mathrm{a}[\theta_{d0} + \theta_{d1}x], \tag{3.5}$$

$$y = \phi_0 + \sum_{d=1}^{D} \phi_d h_d. \tag{3.6}$$

$$\begin{aligned} h_1 &= \mathrm{a}[\theta_{10} + \theta_{11}x] \\ h_2 &= \mathrm{a}[\theta_{20} + \theta_{21}x] \\ h_3 &= \mathrm{a}[\theta_{30} + \theta_{31}x] \\ h_4 &= \mathrm{a}[\theta_{40} + \theta_{41}x], \end{aligned} \tag{3.7}$$

$$\begin{aligned} y_1 &= \phi_{10} + \phi_{11}h_1 + \phi_{12}h_2 + \phi_{13}h_3 + \phi_{14}h_4 \\ y_2 &= \phi_{20} + \phi_{21}h_1 + \phi_{22}h_2 + \phi_{23}h_3 + \phi_{24}h_4. \end{aligned} \tag{3.8}$$

$$\begin{aligned} h_1 &= \mathrm{a}[\theta_{10} + \theta_{11}x_1 + \theta_{12}x_2] \\ h_2 &= \mathrm{a}[\theta_{20} + \theta_{21}x_1 + \theta_{22}x_2] \\ h_3 &= \mathrm{a}[\theta_{30} + \theta_{31}x_1 + \theta_{32}x_2], \end{aligned} \tag{3.9}$$

$$y = \phi_0 + \phi_1 h_1 + \phi_2 h_2 + \phi_3 h_3. \tag{3.10}$$

$$h_d = \mathrm{a}\left[\theta_{d0} + \sum_{i=1}^{D_i} \theta_{di} x_i\right], \tag{3.11}$$

$$y_j = \phi_{j0} + \sum_{d=1}^{D} \phi_{jd} h_d, \tag{3.12}$$

$$\mathrm{HardSwish}[z] = \begin{cases} 0 & z < -3 \\ z(z+3)/6 & -3 \leq z \leq 3 \\ z & z > 3 \end{cases}. \tag{3.13}$$

$$\mathrm{ReLU}[\alpha \cdot z] = \alpha \cdot \mathrm{ReLU}[z]. \tag{3.14}$$

$$\mathrm{heaviside}[z] = \begin{cases} 0 & z < 0 \\ 1 & z \geq 0 \end{cases} \qquad \mathrm{rect}[z] = \begin{cases} 0 & z < 0 \\ 1 & 0 \leq z \leq 1 \\ 0 & z > 1 \end{cases}. \tag{3.15}$$

Chapter 4

Deep neural networks

$$
\begin{aligned}
h_1 &= \mathrm{a}[\theta_{10} + \theta_{11}x] \\
h_2 &= \mathrm{a}[\theta_{20} + \theta_{21}x] \\
h_3 &= \mathrm{a}[\theta_{30} + \theta_{31}x],
\end{aligned} \tag{4.1}
$$

$$
y = \phi_0 + \phi_1 h_1 + \phi_2 h_2 + \phi_3 h_3. \tag{4.2}
$$

$$
\begin{aligned}
h'_1 &= \mathrm{a}[\theta'_{10} + \theta'_{11}y] \\
h'_2 &= \mathrm{a}[\theta'_{20} + \theta'_{21}y] \\
h'_3 &= \mathrm{a}[\theta'_{30} + \theta'_{31}y],
\end{aligned} \tag{4.3}
$$

$$
y' = \phi'_0 + \phi'_1 h'_1 + \phi'_2 h'_2 + \phi'_3 h'_3. \tag{4.4}
$$

$$
\begin{aligned}
h'_1 &= \mathrm{a}[\theta'_{10} + \theta'_{11}y] = \mathrm{a}[\theta'_{10} + \theta'_{11}\phi_0 + \theta'_{11}\phi_1 h_1 + \theta'_{11}\phi_2 h_2 + \theta'_{11}\phi_3 h_3] \\
h'_2 &= \mathrm{a}[\theta'_{20} + \theta'_{21}y] = \mathrm{a}[\theta'_{20} + \theta'_{21}\phi_0 + \theta'_{21}\phi_1 h_1 + \theta'_{21}\phi_2 h_2 + \theta'_{21}\phi_3 h_3] \\
h'_3 &= \mathrm{a}[\theta'_{30} + \theta'_{31}y] = \mathrm{a}[\theta'_{30} + \theta'_{31}\phi_0 + \theta'_{31}\phi_1 h_1 + \theta'_{31}\phi_2 h_2 + \theta'_{31}\phi_3 h_3],
\end{aligned} \tag{4.5}
$$

$$
\begin{aligned}
h'_1 &= \mathrm{a}[\psi_{10} + \psi_{11}h_1 + \psi_{12}h_2 + \psi_{13}h_3] \\
h'_2 &= \mathrm{a}[\psi_{20} + \psi_{21}h_1 + \psi_{22}h_2 + \psi_{23}h_3] \\
h'_3 &= \mathrm{a}[\psi_{30} + \psi_{31}h_1 + \psi_{32}h_2 + \psi_{33}h_3],
\end{aligned} \tag{4.6}
$$

$$\begin{aligned} h_1 &= \mathrm{a}[\theta_{10}+\theta_{11}x] \\ h_2 &= \mathrm{a}[\theta_{20}+\theta_{21}x] \\ h_3 &= \mathrm{a}[\theta_{30}+\theta_{31}x], \end{aligned} \tag{4.7}$$

$$\begin{aligned} h'_1 &= \mathrm{a}[\psi_{10}+\psi_{11}h_1+\psi_{12}h_2+\psi_{13}h_3] \\ h'_2 &= \mathrm{a}[\psi_{20}+\psi_{21}h_1+\psi_{22}h_2+\psi_{23}h_3] \\ h'_3 &= \mathrm{a}[\psi_{30}+\psi_{31}h_1+\psi_{32}h_2+\psi_{33}h_3], \end{aligned} \tag{4.8}$$

$$y' = \phi'_0+\phi'_1h'_1+\phi'_2h'_2+\phi'_3h'_3. \tag{4.9}$$

$$\begin{aligned} y' = \;& \phi'_0+\phi'_1\mathrm{a}\left[\psi_{10}+\psi_{11}\mathrm{a}[\theta_{10}+\theta_{11}x]+\psi_{12}\mathrm{a}[\theta_{20}+\theta_{21}x]+\psi_{13}\mathrm{a}[\theta_{30}+\theta_{31}x]\right] \\ & +\phi'_2\mathrm{a}[\psi_{20}+\psi_{21}\mathrm{a}[\theta_{10}+\theta_{11}x]+\psi_{22}\mathrm{a}[\theta_{20}+\theta_{21}x]+\psi_{23}\mathrm{a}[\theta_{30}+\theta_{31}x]] \\ & +\phi'_3\mathrm{a}[\psi_{30}+\psi_{31}\mathrm{a}[\theta_{10}+\theta_{11}x]+\psi_{32}\mathrm{a}[\theta_{20}+\theta_{21}x]+\psi_{33}\mathrm{a}[\theta_{30}+\theta_{31}x]], \end{aligned} \tag{4.10}$$

$$\begin{bmatrix} h_1 \\ h_2 \\ h_3 \end{bmatrix} = \mathbf{a}\left[\begin{bmatrix} \theta_{10} \\ \theta_{20} \\ \theta_{30} \end{bmatrix} + \begin{bmatrix} \theta_{11} \\ \theta_{21} \\ \theta_{31} \end{bmatrix} x\right], \tag{4.11}$$

$$\begin{bmatrix} h'_1 \\ h'_2 \\ h'_3 \end{bmatrix} = \mathbf{a}\left[\begin{bmatrix} \psi_{10} \\ \psi_{20} \\ \psi_{30} \end{bmatrix} + \begin{bmatrix} \psi_{11} & \psi_{12} & \psi_{13} \\ \psi_{21} & \psi_{22} & \psi_{23} \\ \psi_{31} & \psi_{32} & \psi_{33} \end{bmatrix} \begin{bmatrix} h_1 \\ h_2 \\ h_3 \end{bmatrix}\right], \tag{4.12}$$

$$y' = \phi'_0 + \begin{bmatrix} \phi'_1 & \phi'_2 & \phi'_3 \end{bmatrix} \begin{bmatrix} h'_1 \\ h'_2 \\ h'_3 \end{bmatrix}, \tag{4.13}$$

$$\begin{aligned} \mathbf{h} &= \mathbf{a}\left[\boldsymbol{\theta}_0+\boldsymbol{\theta}x\right] \\ \mathbf{h}' &= \mathbf{a}\left[\boldsymbol{\psi}_0+\boldsymbol{\Psi}\mathbf{h}\right] \\ y' &= \phi'_0+\boldsymbol{\phi}'\mathbf{h}', \end{aligned} \tag{4.14}$$

$$\begin{aligned}
\mathbf{h}_1 &= \mathbf{a}[\boldsymbol{\beta}_0 + \boldsymbol{\Omega}_0\mathbf{x}] \\
\mathbf{h}_2 &= \mathbf{a}[\boldsymbol{\beta}_1 + \boldsymbol{\Omega}_1\mathbf{h}_1] \\
\mathbf{h}_3 &= \mathbf{a}[\boldsymbol{\beta}_2 + \boldsymbol{\Omega}_2\mathbf{h}_2] \\
&\vdots \\
\mathbf{h}_K &= \mathbf{a}[\boldsymbol{\beta}_{K-1} + \boldsymbol{\Omega}_{K-1}\mathbf{h}_{K-1}] \\
\mathbf{y} &= \boldsymbol{\beta}_K + \boldsymbol{\Omega}_K\mathbf{h}_K.
\end{aligned} \tag{4.15}$$

$$\mathbf{y} = \boldsymbol{\beta}_K + \boldsymbol{\Omega}_K\mathbf{a}\left[\boldsymbol{\beta}_{K-1} + \boldsymbol{\Omega}_{K-1}\mathbf{a}\left[\ldots\boldsymbol{\beta}_2 + \boldsymbol{\Omega}_2\mathbf{a}\left[\boldsymbol{\beta}_1 + \boldsymbol{\Omega}_1\mathbf{a}\left[\boldsymbol{\beta}_0 + \boldsymbol{\Omega}_0\mathbf{x}\right]\right]\ldots\right]\right]. \tag{4.16}$$

$$N_r = \left(\frac{D}{D_i} + 1\right)^{D_i(K-1)} \cdot \sum_{j=0}^{D_i}\binom{D}{j}. \tag{4.17}$$

$$\text{ReLU}\Big[\boldsymbol{\beta}_1 + \lambda_1 \cdot \boldsymbol{\Omega}_1\text{ReLU}\left[\boldsymbol{\beta}_0 + \lambda_0 \cdot \boldsymbol{\Omega}_0\mathbf{x}\right]\Big] = \lambda_0\lambda_1 \cdot \text{ReLU}\left[\frac{1}{\lambda_0\lambda_1}\boldsymbol{\beta}_1 + \boldsymbol{\Omega}_1\text{ReLU}\left[\frac{1}{\lambda_0}\boldsymbol{\beta}_0 + \boldsymbol{\Omega}_0\mathbf{x}\right]\right], \tag{4.18}$$

Chapter 5

Loss functions

$$
\begin{aligned}
\hat{\boldsymbol{\phi}} &= \underset{\boldsymbol{\phi}}{\operatorname{argmax}}\left[\prod_{i=1}^{I} Pr(\mathbf{y}_i|\mathbf{x}_i)\right] \\
&= \underset{\boldsymbol{\phi}}{\operatorname{argmax}}\left[\prod_{i=1}^{I} Pr(\mathbf{y}_i|\boldsymbol{\theta}_i)\right] \\
&= \underset{\boldsymbol{\phi}}{\operatorname{argmax}}\left[\prod_{i=1}^{I} Pr(\mathbf{y}_i|\mathbf{f}[\mathbf{x}_i,\boldsymbol{\phi}])\right].
\end{aligned}
\tag{5.1}
$$

$$
Pr(\mathbf{y}_1,\mathbf{y}_2,\ldots,\mathbf{y}_I|\mathbf{x}_1,\mathbf{x}_2,\ldots,\mathbf{x}_I) = \prod_{i=1}^{I} Pr(\mathbf{y}_i|\mathbf{x}_i).
\tag{5.2}
$$

$$
\begin{aligned}
\hat{\boldsymbol{\phi}} &= \underset{\boldsymbol{\phi}}{\operatorname{argmax}}\left[\prod_{i=1}^{I} Pr(\mathbf{y}_i|\mathbf{f}[\mathbf{x}_i,\boldsymbol{\phi}])\right] \\
&= \underset{\boldsymbol{\phi}}{\operatorname{argmax}}\left[\log\left[\prod_{i=1}^{I} Pr(\mathbf{y}_i|\mathbf{f}[\mathbf{x}_i,\boldsymbol{\phi}])\right]\right] \\
&= \underset{\boldsymbol{\phi}}{\operatorname{argmax}}\left[\sum_{i=1}^{I} \log\Bigl[Pr(\mathbf{y}_i|\mathbf{f}[\mathbf{x}_i,\boldsymbol{\phi}])\Bigr]\right].
\end{aligned}
\tag{5.3}
$$

$$
\begin{aligned}
\hat{\boldsymbol{\phi}} &= \underset{\boldsymbol{\phi}}{\operatorname{argmin}}\left[-\sum_{i=1}^{I} \log\Bigl[Pr(\mathbf{y}_i|\mathbf{f}[\mathbf{x}_i,\boldsymbol{\phi}])\Bigr]\right] \\
&= \underset{\boldsymbol{\phi}}{\operatorname{argmin}}\Bigl[L[\boldsymbol{\phi}]\Bigr],
\end{aligned}
\tag{5.4}
$$

$$\hat{\mathbf{y}} = \underset{\mathbf{y}}{\operatorname{argmax}}\Big[Pr(\mathbf{y}|\mathbf{f}[\mathbf{x},\hat{\boldsymbol{\phi}}])\Big]. \tag{5.5}$$

$$\hat{\boldsymbol{\phi}} = \underset{\boldsymbol{\phi}}{\operatorname{argmin}}\Big[L[\boldsymbol{\phi}]\Big] = \underset{\boldsymbol{\phi}}{\operatorname{argmin}}\left[-\sum_{i=1}^{I}\log\Big[Pr(\mathbf{y}_i|\mathbf{f}[\mathbf{x}_i,\boldsymbol{\phi}])\Big]\right]. \tag{5.6}$$

$$Pr(y|\mu,\sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}}\exp\left[-\frac{(y-\mu)^2}{2\sigma^2}\right]. \tag{5.7}$$

$$Pr(y|\mathrm{f}[\mathbf{x},\boldsymbol{\phi}],\sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}}\exp\left[-\frac{(y-\mathrm{f}[\mathbf{x},\boldsymbol{\phi}])^2}{2\sigma^2}\right]. \tag{5.8}$$

$$\begin{aligned}
L[\boldsymbol{\phi}] &= -\sum_{i=1}^{I}\log\left[Pr(y_i|\mathrm{f}[\mathbf{x}_i,\boldsymbol{\phi}],\sigma^2)\right] \\
&= -\sum_{i=1}^{I}\log\left[\frac{1}{\sqrt{2\pi\sigma^2}}\exp\left[-\frac{(y_i-\mathrm{f}[\mathbf{x}_i,\boldsymbol{\phi}])^2}{2\sigma^2}\right]\right].
\end{aligned} \tag{5.9}$$

$$\begin{aligned}
\hat{\boldsymbol{\phi}} &= \underset{\boldsymbol{\phi}}{\operatorname{argmin}}\left[-\sum_{i=1}^{I}\log\left[\frac{1}{\sqrt{2\pi\sigma^2}}\exp\left[-\frac{(y_i-\mathrm{f}[\mathbf{x}_i,\boldsymbol{\phi}])^2}{2\sigma^2}\right]\right]\right] \\
&= \underset{\boldsymbol{\phi}}{\operatorname{argmin}}\left[-\sum_{i=1}^{I}\left(\log\left[\frac{1}{\sqrt{2\pi\sigma^2}}\right]-\frac{(y_i-\mathrm{f}[\mathbf{x}_i,\boldsymbol{\phi}])^2}{2\sigma^2}\right)\right] \\
&= \underset{\boldsymbol{\phi}}{\operatorname{argmin}}\left[-\sum_{i=1}^{I}-\frac{(y_i-\mathrm{f}[\mathbf{x}_i,\boldsymbol{\phi}])^2}{2\sigma^2}\right] \\
&= \underset{\boldsymbol{\phi}}{\operatorname{argmin}}\left[\sum_{i=1}^{I}(y_i-\mathrm{f}[\mathbf{x}_i,\boldsymbol{\phi}])^2\right],
\end{aligned} \tag{5.10}$$

$$L[\boldsymbol{\phi}] = \sum_{i=1}^{I}\big(y_i-\mathrm{f}[\mathbf{x}_i,\boldsymbol{\phi}]\big)^2. \tag{5.11}$$

$$\hat{y} = \underset{y}{\operatorname{argmax}}\left[Pr(y|\mathrm{f}[\mathbf{x},\hat{\boldsymbol{\phi}}],\sigma^2)\right]. \tag{5.12}$$

$$\hat{\boldsymbol{\phi}}, \hat{\sigma}^2 = \underset{\boldsymbol{\phi},\sigma^2}{\operatorname{argmin}}\left[-\sum_{i=1}^{I}\log\left[\frac{1}{\sqrt{2\pi\sigma^2}}\exp\left[-\frac{(y_i - \mathrm{f}[\mathbf{x}_i,\boldsymbol{\phi}])^2}{2\sigma^2}\right]\right]\right]. \tag{5.13}$$

$$\begin{aligned}\mu &= \mathrm{f}_1[\mathbf{x},\boldsymbol{\phi}]\\ \sigma^2 &= \mathrm{f}_2[\mathbf{x},\boldsymbol{\phi}]^2,\end{aligned} \tag{5.14}$$

$$\hat{\boldsymbol{\phi}} = \underset{\boldsymbol{\phi}}{\operatorname{argmin}}\left[-\sum_{i=1}^{I}\left(\log\left[\frac{1}{\sqrt{2\pi \mathrm{f}_2[\mathbf{x}_i,\boldsymbol{\phi}]^2}}\right] - \frac{(y_i - \mathrm{f}_1[\mathbf{x}_i,\boldsymbol{\phi}])^2}{2\mathrm{f}_2[\mathbf{x}_i,\boldsymbol{\phi}]^2}\right)\right]. \tag{5.15}$$

$$Pr(y|\lambda) = \begin{cases}1-\lambda & y=0\\ \lambda & y=1\end{cases}, \tag{5.16}$$

$$Pr(y|\lambda) = (1-\lambda)^{1-y}\cdot\lambda^{y}. \tag{5.17}$$

$$\mathrm{sig}[z] = \frac{1}{1+\exp[-z]}. \tag{5.18}$$

$$Pr(y|\mathbf{x}) = (1-\mathrm{sig}[\mathrm{f}[\mathbf{x},\boldsymbol{\phi}]])^{1-y}\cdot\mathrm{sig}[\mathrm{f}[\mathbf{x},\boldsymbol{\phi}]]^{y}. \tag{5.19}$$

$$L[\boldsymbol{\phi}] = \sum_{i=1}^{I} -(1-y_i)\log\Big[1-\mathrm{sig}[\mathrm{f}[\mathbf{x}_i,\boldsymbol{\phi}]]\Big] - y_i\log\Big[\mathrm{sig}[\mathrm{f}[\mathbf{x}_i,\boldsymbol{\phi}]]\Big]. \tag{5.20}$$

$$Pr(y=k) = \lambda_k. \tag{5.21}$$

$$\mathrm{softmax}_k[\mathbf{z}] = \frac{\exp[z_k]}{\sum_{k'=1}^{K}\exp[z_{k'}]}, \tag{5.22}$$

$$Pr(y=k|\mathbf{x}) = \mathrm{softmax}_k\Big[\mathbf{f}[\mathbf{x},\boldsymbol{\phi}]\Big]. \tag{5.23}$$

$$\begin{aligned} L[\boldsymbol{\phi}] &= -\sum_{i=1}^{I} \log\left[\text{softmax}_{y_i}\Big[\mathbf{f}\left[\mathbf{x}_i, \boldsymbol{\phi}\right]\Big]\right] \\ &= -\sum_{i=1}^{I}\left(\mathrm{f}_{y_i}\left[\mathbf{x}_i, \boldsymbol{\phi}\right] - \log\left[\sum_{k'=1}^{K}\exp\left[\ \mathrm{f}_{k'}\left[\mathbf{x}_i, \boldsymbol{\phi}\right]\right]\right]\right), \end{aligned} \tag{5.24}$$

$$Pr(\mathbf{y}|\mathbf{f}[\mathbf{x}, \boldsymbol{\phi}]) = \prod_{d} Pr(y_d|\mathbf{f}_d[\mathbf{x}, \boldsymbol{\phi}]), \tag{5.25}$$

$$L[\boldsymbol{\phi}] = -\sum_{i=1}^{I}\log\Big[Pr(\mathbf{y}_i|\mathbf{f}[\mathbf{x}_i, \boldsymbol{\phi}])\Big] = -\sum_{i=1}^{I}\sum_{d}\log\Big[Pr(y_{id}|\mathbf{f}_d[\mathbf{x}_i, \boldsymbol{\phi}])\Big]. \tag{5.26}$$

$$D_{KL}\big[q||p\big] = \int_{-\infty}^{\infty} q(z)\log\big[q(z)\big]dz - \int_{-\infty}^{\infty} q(z)\log\big[p(z)\big]dz. \tag{5.27}$$

$$q(y) = \frac{1}{I}\sum_{i=1}^{I}\delta[y - y_i], \tag{5.28}$$

$$\begin{aligned} \hat{\boldsymbol{\theta}} &= \underset{\boldsymbol{\theta}}{\text{argmin}}\left[\int_{-\infty}^{\infty} q(y)\log\big[q(y)\big]dy - \int_{-\infty}^{\infty} q(y)\log\big[Pr(y|\boldsymbol{\theta})\big]dy\right] \\ &= \underset{\boldsymbol{\theta}}{\text{argmin}}\left[-\int_{-\infty}^{\infty} q(y)\log\big[Pr(y|\boldsymbol{\theta})\big]dy\right], \end{aligned} \tag{5.29}$$

$$\begin{aligned} \hat{\boldsymbol{\theta}} &= \underset{\theta}{\text{argmin}}\left[-\int_{-\infty}^{\infty}\left(\frac{1}{I}\sum_{i=1}^{I}\delta[y - y_i]\right)\log\big[Pr(y|\boldsymbol{\theta})\big]dy\right] \\ &= \underset{\boldsymbol{\theta}}{\text{argmin}}\left[-\frac{1}{I}\sum_{i=1}^{I}\log\big[Pr(y_i|\boldsymbol{\theta})\big]\right] \\ &= \underset{\boldsymbol{\theta}}{\text{argmin}}\left[-\sum_{i=1}^{I}\log\big[Pr(y_i|\boldsymbol{\theta})\big]\right]. \end{aligned} \tag{5.30}$$

$$\hat{\boldsymbol{\phi}} = \underset{\boldsymbol{\phi}}{\text{argmin}}\left[-\sum_{i=1}^{I}\log\big[Pr(y_i|\mathbf{f}[\mathbf{x}_i, \boldsymbol{\phi}])\big]\right]. \tag{5.31}$$

$$\mathrm{sig}[z] = \frac{1}{1+\exp[-z]}. \tag{5.32}$$

$$L = -(1-y)\log\Big[1-\mathrm{sig}[\mathrm{f}[\mathbf{x},\boldsymbol{\phi}]]\Big] - y\log\Big[\mathrm{sig}[\mathrm{f}[\mathbf{x},\boldsymbol{\phi}]]\Big], \tag{5.33}$$

$$Pr(y|\mu,\kappa) = \frac{\exp\big[\kappa\cos[y-\mu]\big]}{2\pi\cdot\mathrm{Bessel}_0[\kappa]}, \tag{5.34}$$

$$Pr(y|\lambda,\mu_1,\mu_2,\sigma_1^2,\sigma_2^2) = \frac{\lambda}{\sqrt{2\pi\sigma_1^2}}\exp\left[\frac{-(y-\mu_1)^2}{2\sigma_1^2}\right] + \frac{1-\lambda}{\sqrt{2\pi\sigma_2^2}}\exp\left[\frac{-(y-\mu_2)^2}{2\sigma_2^2}\right], \tag{5.35}$$

$$Pr(y=k) = \frac{\lambda^k e^{-\lambda}}{k!}. \tag{5.36}$$

Chapter 6

Fitting models

$$\hat{\boldsymbol{\phi}} = \underset{\boldsymbol{\phi}}{\operatorname{argmin}}\Big[L[\boldsymbol{\phi}]\Big]. \tag{6.1}$$

$$\frac{\partial L}{\partial \boldsymbol{\phi}} = \begin{bmatrix} \frac{\partial L}{\partial \phi_0} \\ \frac{\partial L}{\partial \phi_1} \\ \vdots \\ \frac{\partial L}{\partial \phi_N} \end{bmatrix}. \tag{6.2}$$

$$\boldsymbol{\phi} \longleftarrow \boldsymbol{\phi} - \alpha \cdot \frac{\partial L}{\partial \boldsymbol{\phi}}, \tag{6.3}$$

$$\begin{aligned} y &= \mathrm{f}[x, \boldsymbol{\phi}] \\ &= \phi_0 + \phi_1 x. \end{aligned} \tag{6.4}$$

$$\begin{aligned} L[\boldsymbol{\phi}] = \sum_{i=1}^{I} \ell_i &= \sum_{i=1}^{I} \left(\mathrm{f}[x_i, \boldsymbol{\phi}] - y_i\right)^2 \\ &= \sum_{i=1}^{I} \left(\phi_0 + \phi_1 x_i - y_i\right)^2, \end{aligned} \tag{6.5}$$

$$\frac{\partial L}{\partial \boldsymbol{\phi}} = \frac{\partial}{\partial \boldsymbol{\phi}} \sum_{i=1}^{I} \ell_i = \sum_{i=1}^{I} \frac{\partial \ell_i}{\partial \boldsymbol{\phi}}, \tag{6.6}$$

$$\frac{\partial \ell_i}{\partial \boldsymbol{\phi}} = \begin{bmatrix} \frac{\partial \ell_i}{\partial \phi_0} \\ \frac{\partial \ell_i}{\partial \phi_1} \end{bmatrix} = \begin{bmatrix} 2(\phi_0 + \phi_1 x_i - y_i) \\ 2x_i(\phi_0 + \phi_1 x_i - y_i) \end{bmatrix}. \tag{6.7}$$

$$\mathrm{f}[x, \boldsymbol{\phi}] = \sin[\phi_0 + 0.06 \cdot \phi_1 x] \cdot \exp\left(-\frac{(\phi_0 + 0.06 \cdot \phi_1 x)^2}{32.0}\right). \tag{6.8}$$

$$L[\boldsymbol{\phi}] = \sum_{i=1}^{I} \left(\mathrm{f}[x_i, \boldsymbol{\phi}] - y_i\right)^2. \tag{6.9}$$

$$\boldsymbol{\phi}_{t+1} \longleftarrow \boldsymbol{\phi}_t - \alpha \cdot \sum_{i \in \mathcal{B}_t} \frac{\partial \ell_i[\boldsymbol{\phi}_t]}{\partial \boldsymbol{\phi}}, \tag{6.10}$$

$$\begin{aligned} \mathbf{m}_{t+1} &\leftarrow \beta \cdot \mathbf{m}_t + (1-\beta) \sum_{i \in \mathcal{B}_t} \frac{\partial \ell_i[\boldsymbol{\phi}_t]}{\partial \boldsymbol{\phi}} \\ \boldsymbol{\phi}_{t+1} &\leftarrow \boldsymbol{\phi}_t - \alpha \cdot \mathbf{m}_{t+1}, \end{aligned} \tag{6.11}$$

$$\begin{aligned} \mathbf{m}_{t+1} &\leftarrow \beta \cdot \mathbf{m}_t + (1-\beta) \sum_{i \in \mathcal{B}_t} \frac{\partial \ell_i[\boldsymbol{\phi}_t - \alpha\beta \cdot \mathbf{m}_t]}{\partial \boldsymbol{\phi}} \\ \boldsymbol{\phi}_{t+1} &\leftarrow \boldsymbol{\phi}_t - \alpha \cdot \mathbf{m}_{t+1}, \end{aligned} \tag{6.12}$$

$$\begin{aligned} \mathbf{m}_{t+1} &\leftarrow \frac{\partial L[\boldsymbol{\phi}_t]}{\partial \boldsymbol{\phi}} \\ \mathbf{v}_{t+1} &\leftarrow \left(\frac{\partial L[\boldsymbol{\phi}_t]}{\partial \boldsymbol{\phi}}\right)^2. \end{aligned} \tag{6.13}$$

$$\boldsymbol{\phi}_{t+1} \leftarrow \boldsymbol{\phi}_t - \alpha \cdot \frac{\mathbf{m}_{t+1}}{\sqrt{\mathbf{v}_{t+1}} + \epsilon}, \tag{6.14}$$

$$\begin{aligned} \mathbf{m}_{t+1} &\leftarrow \beta \cdot \mathbf{m}_t + (1-\beta) \frac{\partial L[\boldsymbol{\phi}_t]}{\partial \boldsymbol{\phi}} \\ \mathbf{v}_{t+1} &\leftarrow \gamma \cdot \mathbf{v}_t + (1-\gamma) \left(\frac{\partial L[\boldsymbol{\phi}_t]}{\partial \boldsymbol{\phi}}\right)^2, \end{aligned} \tag{6.15}$$

$$\begin{aligned}\tilde{\mathbf{m}}_{t+1} &\leftarrow \frac{\mathbf{m}_{t+1}}{1-\beta^{t+1}}\\ \tilde{\mathbf{v}}_{t+1} &\leftarrow \frac{\mathbf{v}_{t+1}}{1-\gamma^{t+1}}.\end{aligned} \tag{6.16}$$

$$\boldsymbol{\phi}_{t+1} \leftarrow \boldsymbol{\phi}_t - \alpha\cdot\frac{\tilde{\mathbf{m}}_{t+1}}{\sqrt{\tilde{\mathbf{v}}_{t+1}}+\epsilon}. \tag{6.17}$$

$$\begin{aligned}\mathbf{m}_{t+1} &\leftarrow \beta\cdot\mathbf{m}_t + (1-\beta)\sum_{i\in\mathcal{B}_t}\frac{\partial\ell_i[\boldsymbol{\phi}_t]}{\partial\boldsymbol{\phi}}\\ \mathbf{v}_{t+1} &\leftarrow \gamma\cdot\mathbf{v}_t + (1-\gamma)\left(\sum_{i\in\mathcal{B}_t}\frac{\partial\ell_i[\boldsymbol{\phi}_t]}{\partial\boldsymbol{\phi}}\right)^2,\end{aligned} \tag{6.18}$$

$$\mathbf{H}[\boldsymbol{\phi}] = \begin{bmatrix} \frac{\partial^2 L}{\partial\phi_0^2} & \frac{\partial^2 L}{\partial\phi_0\partial\phi_1} & \cdots & \frac{\partial^2 L}{\partial\phi_0\partial\phi_N}\\ \frac{\partial^2 L}{\partial\phi_1\partial\phi_0} & \frac{\partial^2 L}{\partial\phi_1^2} & \cdots & \frac{\partial^2 L}{\partial\phi_1\partial\phi_N}\\ \vdots & \vdots & \ddots & \vdots\\ \frac{\partial^2 L}{\partial\phi_N\partial\phi_0} & \frac{\partial^2 L}{\partial\phi_N\partial\phi_1} & \cdots & \frac{\partial^2 L}{\partial\phi_N^2}\end{bmatrix}. \tag{6.19}$$

$$\mathbf{H}[\boldsymbol{\phi}] = \begin{bmatrix} \frac{\partial^2 L}{\partial\phi_0^2} & \frac{\partial^2 L}{\partial\phi_0\partial\phi_1}\\ \frac{\partial^2 L}{\partial\phi_1\partial\phi_0} & \frac{\partial^2 L}{\partial\phi_1^2}\end{bmatrix}, \tag{6.20}$$

$$Pr(y=1|x) = \mathrm{sig}[\phi_0+\phi_1 x], \tag{6.21}$$

$$\mathrm{sig}[z] = \frac{1}{1+\exp[-z]}. \tag{6.22}$$

$$\mathrm{f}[x,\boldsymbol{\phi}] = \phi_0 + \phi_1\mathrm{a}[\theta_{10}+\theta_{11}x] + \phi_2\mathrm{a}[\theta_{20}+\theta_{21}x] + \phi_3\mathrm{a}[\theta_{30}+\theta_{31}x]. \tag{6.23}$$

Chapter 7

Gradients and initialization

$$
\begin{aligned}
\mathbf{h}_1 &= \mathbf{a}[\boldsymbol{\beta}_0 + \boldsymbol{\Omega}_0\mathbf{x}] \\
\mathbf{h}_2 &= \mathbf{a}[\boldsymbol{\beta}_1 + \boldsymbol{\Omega}_1\mathbf{h}_1] \\
\mathbf{h}_3 &= \mathbf{a}[\boldsymbol{\beta}_2 + \boldsymbol{\Omega}_2\mathbf{h}_2] \\
\mathbf{f}[\mathbf{x}, \boldsymbol{\phi}] &= \boldsymbol{\beta}_3 + \boldsymbol{\Omega}_3\mathbf{h}_3,
\end{aligned} \tag{7.1}
$$

$$
L[\boldsymbol{\phi}] = \sum_{i=1}^{I} \ell_i. \tag{7.2}
$$

$$
\boldsymbol{\phi}_{t+1} \longleftarrow \boldsymbol{\phi}_t - \alpha \sum_{i \in \mathcal{B}_t} \frac{\partial \ell_i[\boldsymbol{\phi}_t]}{\partial \boldsymbol{\phi}}, \tag{7.3}
$$

$$
\frac{\partial \ell_i}{\partial \boldsymbol{\beta}_k} \quad \text{and} \quad \frac{\partial \ell_i}{\partial \boldsymbol{\Omega}_k}, \tag{7.4}
$$

$$
\mathrm{f}[x, \boldsymbol{\phi}] = \beta_3 + \omega_3 \cdot \cos\Bigl[\beta_2 + \omega_2 \cdot \exp\bigl[\beta_1 + \omega_1 \cdot \sin[\beta_0 + \omega_0 \cdot x]\bigr]\Bigr], \tag{7.5}
$$

$$
\ell_i = (\mathrm{f}[x_i, \boldsymbol{\phi}] - y_i)^2, \tag{7.6}
$$

$$
\frac{\partial \ell_i}{\partial \beta_0}, \quad \frac{\partial \ell_i}{\partial \omega_0}, \quad \frac{\partial \ell_i}{\partial \beta_1}, \quad \frac{\partial \ell_i}{\partial \omega_1}, \quad \frac{\partial \ell_i}{\partial \beta_2}, \quad \frac{\partial \ell_i}{\partial \omega_2}, \quad \frac{\partial \ell_i}{\partial \beta_3}, \quad \text{and} \quad \frac{\partial \ell_i}{\partial \omega_3}. \tag{7.7}
$$

$$
\begin{aligned}
\frac{\partial \ell_i}{\partial \omega_0} &= -2\left(\beta_3+\omega_3\cdot\cos\Big[\beta_2+\omega_2\cdot\exp\big[\beta_1+\omega_1\cdot\sin[\beta_0+\omega_0\cdot x_i]\big]\Big]-y_i\right)\\
&\quad\cdot\omega_1\omega_2\omega_3\cdot x_i\cdot\cos[\beta_0+\omega_0\cdot x_i]\cdot\exp\Big[\beta_1+\omega_1\cdot\sin[\beta_0+\omega_0\cdot x_i]\Big]\\
&\quad\cdot\sin\Big[\beta_2+\omega_2\cdot\exp\Big[\beta_1+\omega_1\cdot\sin[\beta_0+\omega_0\cdot x_i]\Big]\Big].
\end{aligned}
\tag{7.8}
$$

$$
\begin{aligned}
f_0 &= \beta_0+\omega_0\cdot x_i\\
h_1 &= \sin[f_0]\\
f_1 &= \beta_1+\omega_1\cdot h_1\\
h_2 &= \exp[f_1]\\
f_2 &= \beta_2+\omega_2\cdot h_2\\
h_3 &= \cos[f_2]\\
f_3 &= \beta_3+\omega_3\cdot h_3\\
\ell_i &= (f_3-y_i)^2.
\end{aligned}
\tag{7.9}
$$

$$
\frac{\partial \ell_i}{\partial f_3},\quad \frac{\partial \ell_i}{\partial h_3},\quad \frac{\partial \ell_i}{\partial f_2},\quad \frac{\partial \ell_i}{\partial h_2},\quad \frac{\partial \ell_i}{\partial f_1},\quad \frac{\partial \ell_i}{\partial h_1},\quad \text{and}\quad \frac{\partial \ell_i}{\partial f_0}.
\tag{7.10}
$$

$$
\frac{\partial \ell_i}{\partial f_3} = 2(f_3-y_i).
\tag{7.11}
$$

$$
\frac{\partial \ell_i}{\partial h_3} = \frac{\partial f_3}{\partial h_3}\frac{\partial \ell_i}{\partial f_3}.
\tag{7.12}
$$

$$
\begin{aligned}
\frac{\partial \ell_i}{\partial f_2} &= \frac{\partial h_3}{\partial f_2}\left(\frac{\partial f_3}{\partial h_3}\frac{\partial \ell_i}{\partial f_3}\right)\\
\frac{\partial \ell_i}{\partial h_2} &= \frac{\partial f_2}{\partial h_2}\left(\frac{\partial h_3}{\partial f_2}\frac{\partial f_3}{\partial h_3}\frac{\partial \ell_i}{\partial f_3}\right)\\
\frac{\partial \ell_i}{\partial f_1} &= \frac{\partial h_2}{\partial f_1}\left(\frac{\partial f_2}{\partial h_2}\frac{\partial h_3}{\partial f_2}\frac{\partial f_3}{\partial h_3}\frac{\partial \ell_i}{\partial f_3}\right)\\
\frac{\partial \ell_i}{\partial h_1} &= \frac{\partial f_1}{\partial h_1}\left(\frac{\partial h_2}{\partial f_1}\frac{\partial f_2}{\partial h_2}\frac{\partial h_3}{\partial f_2}\frac{\partial f_3}{\partial h_3}\frac{\partial \ell_i}{\partial f_3}\right)\\
\frac{\partial \ell_i}{\partial f_0} &= \frac{\partial h_1}{\partial f_0}\left(\frac{\partial f_1}{\partial h_1}\frac{\partial h_2}{\partial f_1}\frac{\partial f_2}{\partial h_2}\frac{\partial h_3}{\partial f_2}\frac{\partial f_3}{\partial h_3}\frac{\partial \ell_i}{\partial f_3}\right).
\end{aligned}
\tag{7.13}
$$

$$\begin{aligned}\frac{\partial \ell_i}{\partial \beta_k} &= \frac{\partial f_k}{\partial \beta_k}\frac{\partial \ell_i}{\partial f_k}\\ \frac{\partial \ell_i}{\partial \omega_k} &= \frac{\partial f_k}{\partial \omega_k}\frac{\partial \ell_i}{\partial f_k}.\end{aligned} \tag{7.14}$$

$$\frac{\partial f_k}{\partial \beta_k} = 1 \qquad \text{and} \qquad \frac{\partial f_k}{\partial \omega_k} = h_k. \tag{7.15}$$

$$\frac{\partial f_0}{\partial \beta_0} = 1 \qquad \text{and} \qquad \frac{\partial f_0}{\partial \omega_0} = x_i. \tag{7.16}$$

$$\begin{aligned}\mathbf{f}_0 &= \boldsymbol{\beta}_0 + \boldsymbol{\Omega}_0\mathbf{x}_i\\ \mathbf{h}_1 &= \mathbf{a}[\mathbf{f}_0]\\ \mathbf{f}_1 &= \boldsymbol{\beta}_1 + \boldsymbol{\Omega}_1\mathbf{h}_1\\ \mathbf{h}_2 &= \mathbf{a}[\mathbf{f}_1]\\ \mathbf{f}_2 &= \boldsymbol{\beta}_2 + \boldsymbol{\Omega}_2\mathbf{h}_2\\ \mathbf{h}_3 &= \mathbf{a}[\mathbf{f}_2]\\ \mathbf{f}_3 &= \boldsymbol{\beta}_3 + \boldsymbol{\Omega}_3\mathbf{h}_3\\ \ell_i &= \mathrm{l}[\mathbf{f}_3, y_i],\end{aligned} \tag{7.17}$$

$$\frac{\partial \ell_i}{\partial \mathbf{f}_2} = \frac{\partial \mathbf{h}_3}{\partial \mathbf{f}_2}\frac{\partial \mathbf{f}_3}{\partial \mathbf{h}_3}\frac{\partial \ell_i}{\partial \mathbf{f}_3}. \tag{7.18}$$

$$\frac{\partial \ell_i}{\partial \mathbf{f}_1} = \frac{\partial \mathbf{h}_2}{\partial \mathbf{f}_1}\frac{\partial \mathbf{f}_2}{\partial \mathbf{h}_2}\left(\frac{\partial \mathbf{h}_3}{\partial \mathbf{f}_2}\frac{\partial \mathbf{f}_3}{\partial \mathbf{h}_3}\frac{\partial \ell_i}{\partial \mathbf{f}_3}\right) \tag{7.19}$$

$$\frac{\partial \ell_i}{\partial \mathbf{f}_0} = \frac{\partial \mathbf{h}_1}{\partial \mathbf{f}_0}\frac{\partial \mathbf{f}_1}{\partial \mathbf{h}_1}\left(\frac{\partial \mathbf{h}_2}{\partial \mathbf{f}_1}\frac{\partial \mathbf{f}_2}{\partial \mathbf{h}_2}\frac{\partial \mathbf{h}_3}{\partial \mathbf{f}_2}\frac{\partial \mathbf{f}_3}{\partial \mathbf{h}_3}\frac{\partial \ell_i}{\partial \mathbf{f}_3}\right). \tag{7.20}$$

$$\frac{\partial \mathbf{f}_3}{\partial \mathbf{h}_3} = \frac{\partial}{\partial \mathbf{h}_3}\left(\boldsymbol{\beta}_3 + \boldsymbol{\Omega}_3\mathbf{h}_3\right) = \boldsymbol{\Omega}_3^T. \tag{7.21}$$

$$\begin{aligned}\frac{\partial \ell_i}{\partial \boldsymbol{\beta}_k} &= \frac{\partial \mathbf{f}_k}{\partial \boldsymbol{\beta}_k}\frac{\partial \ell_i}{\partial \mathbf{f}_k}\\ &= \frac{\partial}{\partial \boldsymbol{\beta}_k}\left(\boldsymbol{\beta}_k + \boldsymbol{\Omega}_k\mathbf{h}_k\right)\frac{\partial \ell_i}{\partial \mathbf{f}_k}\\ &= \frac{\partial \ell_i}{\partial \mathbf{f}_k},\end{aligned} \tag{7.22}$$

$$\begin{aligned}\frac{\partial \ell_i}{\partial \boldsymbol{\Omega}_k} &= \frac{\partial \mathbf{f}_k}{\partial \boldsymbol{\Omega}_k}\frac{\partial \ell_i}{\partial \mathbf{f}_k}\\ &= \frac{\partial}{\partial \boldsymbol{\Omega}_k}\left(\boldsymbol{\beta}_k+\boldsymbol{\Omega}_k\mathbf{h}_k\right)\frac{\partial \ell_i}{\partial \mathbf{f}_k}\\ &= \frac{\partial \ell_i}{\partial \mathbf{f}_k}\mathbf{h}_k^T. \end{aligned} \tag{7.23}$$

$$\begin{aligned}\mathbf{f}_0 &= \boldsymbol{\beta}_0+\boldsymbol{\Omega}_0\mathbf{x}_i\\ \mathbf{h}_k &= \mathbf{a}[\mathbf{f}_{k-1}] && k\in\{1,2,\ldots,K\}\\ \mathbf{f}_k &= \boldsymbol{\beta}_k+\boldsymbol{\Omega}_k\mathbf{h}_k. && k\in\{1,2,\ldots,K\}\end{aligned} \tag{7.24}$$

$$\begin{aligned}\frac{\partial \ell_i}{\partial \boldsymbol{\beta}_k} &= \frac{\partial \ell_i}{\partial \mathbf{f}_k} && k\in\{K,K-1,\ldots,1\}\\ \frac{\partial \ell_i}{\partial \boldsymbol{\Omega}_k} &= \frac{\partial \ell_i}{\partial \mathbf{f}_k}\mathbf{h}_k^T && k\in\{K,K-1,\ldots,1\}\\ \frac{\partial \ell_i}{\partial \mathbf{f}_{k-1}} &= \mathbb{I}[\mathbf{f}_{k-1}>0]\odot\left(\boldsymbol{\Omega}_k^T\frac{\partial \ell_i}{\partial \mathbf{f}_k}\right), && k\in\{K,K-1,\ldots,1\}\end{aligned} \tag{7.25}$$

$$\begin{aligned}\frac{\partial \ell_i}{\partial \boldsymbol{\beta}_0} &= \frac{\partial \ell_i}{\partial \mathbf{f}_0}\\ \frac{\partial \ell_i}{\partial \boldsymbol{\Omega}_0} &= \frac{\partial \ell_i}{\partial \mathbf{f}_0}\mathbf{x}_i^T.\end{aligned} \tag{7.26}$$

$$\begin{aligned}\mathbf{f}_k &= \boldsymbol{\beta}_k+\boldsymbol{\Omega}_k\mathbf{h}_k\\ &= \boldsymbol{\beta}_k+\boldsymbol{\Omega}_k\mathbf{a}[\mathbf{f}_{k-1}],\end{aligned} \tag{7.27}$$

$$\begin{aligned}\mathbf{h} &= \mathbf{a}[\mathbf{f}],\\ \mathbf{f}' &= \boldsymbol{\beta}+\boldsymbol{\Omega}\mathbf{h}\end{aligned} \tag{7.28}$$

$$\begin{aligned}
\mathbb{E}[f'_i] &= \mathbb{E}\left[\beta_i + \sum_{j=1}^{D_h} \Omega_{ij} h_j\right] \\
&= \mathbb{E}\left[\beta_i\right] + \sum_{j=1}^{D_h} \mathbb{E}\left[\Omega_{ij} h_j\right] \\
&= \mathbb{E}\left[\beta_i\right] + \sum_{j=1}^{D_h} \mathbb{E}\left[\Omega_{ij}\right] \mathbb{E}\left[h_j\right] \\
&= 0 + \sum_{j=1}^{D_h} 0 \cdot \mathbb{E}\left[h_j\right] = 0,
\end{aligned} \tag{7.29}$$

$$\begin{aligned}
\sigma^2_{f'_i} &= \mathbb{E}[f'^2_i] - \mathbb{E}[f'_i]^2 \\
&= \mathbb{E}\left[\left(\beta_i + \sum_{j=1}^{D_h} \Omega_{ij} h_j\right)^2\right] - 0 \\
&= \mathbb{E}\left[\left(\sum_{j=1}^{D_h} \Omega_{ij} h_j\right)^2\right] \\
&= \sum_{j=1}^{D_h} \mathbb{E}\left[\Omega^2_{ij}\right] \mathbb{E}\left[h^2_j\right] \\
&= \sum_{j=1}^{D_h} \sigma^2_\Omega \mathbb{E}\left[h^2_j\right] = \sigma^2_\Omega \sum_{j=1}^{D_h} \mathbb{E}\left[h^2_j\right],
\end{aligned} \tag{7.30}$$

$$\sigma^2_{f'_i} = \sigma^2_\Omega \sum_{j=1}^{D_h} \frac{\sigma^2_f}{2} = \frac{1}{2} D_h \sigma^2_\Omega \sigma^2_f. \tag{7.31}$$

$$\sigma^2_\Omega = \frac{2}{D_h}, \tag{7.32}$$

$$\sigma^2_\Omega = \frac{2}{D_{h'}}, \tag{7.33}$$

$$\sigma^2_\Omega = \frac{4}{D_h + D_{h'}}. \tag{7.34}$$

$$
\begin{aligned}
y \;=\; & \phi_0 + \phi_1 \mathrm{a}\Big[\psi_{01} + \psi_{11}\mathrm{a}[\theta_{01} + \theta_{11}x] + \psi_{21}\mathrm{a}[\theta_{02} + \theta_{12}x]\Big] \\
& + \phi_2 \mathrm{a}\Big[\psi_{02} + \psi_{12}\mathrm{a}[\theta_{01} + \theta_{11}x] + \psi_{22}\mathrm{a}[\theta_{02} + \theta_{12}x]\Big],
\end{aligned} \tag{7.35}
$$

$$
\ell_i = (y_i - \mathrm{f}[\mathbf{x}_i, \boldsymbol{\phi}])^2. \tag{7.36}
$$

$$
\ell_i = -(1 - y_i)\log\Big[1 - \mathrm{sig}\big[\mathrm{f}[\mathbf{x}_i, \boldsymbol{\phi}]\big]\Big] - y_i \log\Big[\mathrm{sig}\big[\mathrm{f}[\mathbf{x}_i, \boldsymbol{\phi}]\big]\Big], \tag{7.37}
$$

$$
\mathrm{sig}[z] = \frac{1}{1 + \exp[-z]}. \tag{7.38}
$$

$$
\frac{\partial \mathbf{z}}{\partial \mathbf{h}} = \boldsymbol{\Omega}^T, \tag{7.39}
$$

$$
\mathrm{Heaviside}[z] = \begin{cases} 0 & z < 0 \\ 1 & z \geq 0 \end{cases}, \tag{7.40}
$$

$$
\mathrm{rect}[z] = \begin{cases} 0 & z < 0 \\ 1 & 0 \leq z \leq 1 \\ 0 & z > 1 \end{cases}. \tag{7.41}
$$

$$
\frac{\partial \ell}{\partial \boldsymbol{\Omega}} = \frac{\partial \ell}{\partial \mathbf{f}} \mathbf{h}^T. \tag{7.42}
$$

$$
\mathrm{a}[z] = \mathrm{ReLU}[z] = \begin{cases} \alpha \cdot z & z < 0 \\ z & z \geq 0 \end{cases}, \tag{7.43}
$$

$$
y = \exp\big[\exp[x] + \exp[x]^2\big] + \sin[\exp[x] + \exp[x]^2]. \tag{7.44}
$$

$$
\begin{aligned}
f_1 &= \exp[x] \\
f_2 &= f_1^2 \\
f_3 &= f_1 + f_2 \\
f_4 &= \exp[f_3] \\
f_5 &= \sin[f_3] \\
y &= f_4 + f_5.
\end{aligned} \tag{7.45}
$$

$$\frac{\partial y}{\partial f_5}, \frac{\partial y}{\partial f_4}, \frac{\partial y}{\partial f_3}, \frac{\partial y}{\partial f_2}, \frac{\partial y}{\partial f_1} \text{ and } \frac{\partial y}{\partial x}, \tag{7.46}$$

$$\frac{\partial f_1}{\partial x}, \frac{\partial f_2}{\partial x}, \frac{\partial f_3}{\partial x}, \frac{\partial f_4}{\partial x}, \frac{\partial f_5}{\partial x}, \text{ and } \frac{\partial y}{\partial x}, \tag{7.47}$$

$$b = \mathrm{ReLU}[a] = \begin{cases} 0 & a < 0 \\ a & a \geq 0 \end{cases}, \tag{7.48}$$

Chapter 8

Measuring performance

$$\mu[x] = \mathbb{E}_y[y[x]] = \int y[x]Pr(y|x)dy, \tag{8.1}$$

$$\begin{aligned} L[x] &= \big(\mathrm{f}[x,\boldsymbol{\phi}] - y[x]\big)^2 \\ &= \Big(\big(\mathrm{f}[x,\boldsymbol{\phi}] - \mu[x]\big) + \big(\mu[x] - y[x]\big)\Big)^2 \\ &= \big(\mathrm{f}[x,\boldsymbol{\phi}] - \mu[x]\big)^2 + 2\big(\mathrm{f}[x,\boldsymbol{\phi}] - \mu[x]\big)\big(\mu[x] - y[x]\big) + \big(\mu[x] - y[x]\big)^2, \end{aligned} \tag{8.2}$$

$$\begin{aligned} \mathbb{E}_y\big[L[x]\big] &= \mathbb{E}_y\Big[\big(\mathrm{f}[x,\boldsymbol{\phi}]-\mu[x]\big)^2 + 2\big(\mathrm{f}[x,\boldsymbol{\phi}]-\mu[x]\big)\big(\mu[x]-y[x]\big) + \big(\mu[x]-y[x]\big)^2\Big] \\ &= \big(\mathrm{f}[x,\boldsymbol{\phi}]-\mu[x]\big)^2 + 2\big(\mathrm{f}[x,\boldsymbol{\phi}] - \mu[x]\big)\big(\mu[x]-\mathbb{E}_y\left[y[x]\right]\big) + \mathbb{E}_y\left[(\mu[x]-y[x])^2\right] \\ &= \big(\mathrm{f}[x,\boldsymbol{\phi}]-\mu[x]\big)^2 + 2\big(\mathrm{f}[x,\boldsymbol{\phi}]-\mu[x]\big)\cdot 0 + \mathbb{E}_y\left[\big(\mu[x]-y[x]\big)^2\right] \\ &= \big(\mathrm{f}[x,\boldsymbol{\phi}] - \mu[x]\big)^2 + \sigma^2, \end{aligned} \tag{8.3}$$

$$\mathrm{f}_\mu[x] = \mathbb{E}_\mathcal{D}\Big[\mathrm{f}\big[x,\boldsymbol{\phi}[\mathcal{D}]\big]\Big]. \tag{8.4}$$

$$\begin{aligned} &\big(\mathrm{f}[x,\boldsymbol{\phi}[\mathcal{D}]]-\mu[x]\big)^2 \\ &\quad= \Big(\big(\mathrm{f}[x,\boldsymbol{\phi}[\mathcal{D}]]-\mathrm{f}_\mu[x]\big) + \big(\mathrm{f}_\mu[x] - \mu[x]\big)\Big)^2 \\ &\quad= \big(\mathrm{f}[x,\boldsymbol{\phi}[\mathcal{D}]]-\mathrm{f}_\mu[x]\big)^2 + 2\big(\mathrm{f}[x,\boldsymbol{\phi}[\mathcal{D}]]-\mathrm{f}_\mu[x]\big)\big(\mathrm{f}_\mu[x]-\mu[x]\big) + \big(\mathrm{f}_\mu[x]-\mu[x]\big)^2. \end{aligned} \tag{8.5}$$

$$\mathbb{E}_\mathcal{D}\Big[\big(\mathrm{f}[x,\boldsymbol{\phi}[\mathcal{D}]] - \mu[x]\big)^2\Big] = \mathbb{E}_\mathcal{D}\Big[\big(\mathrm{f}[x,\boldsymbol{\phi}[\mathcal{D}]] - \mathrm{f}_\mu[x]\big)^2\Big] + \big(\mathrm{f}_\mu[x] - \mu[x]\big)^2, \tag{8.6}$$

$$\mathbb{E}_{\mathcal{D}}\Big[\mathbb{E}_y[L[x]]\Big] = \underbrace{\mathbb{E}_{\mathcal{D}}\Big[\big(\mathrm{f}[x,\boldsymbol{\phi}[\mathcal{D}]] - \mathrm{f}_\mu[x]\big)^2\Big]}_{\text{variance}} + \underbrace{\big(\mathrm{f}_\mu[x]-\mu[x]\big)^2}_{\text{bias}} + \underbrace{\sigma^2}_{\text{noise}}. \tag{8.7}$$

$$\mathrm{Vol}[r] = \frac{r^D\pi^{D/2}}{\Gamma[D/2+1]}, \tag{8.8}$$

Chapter 9

Regularization

$$\begin{aligned}\hat{\boldsymbol{\phi}} &= \underset{\boldsymbol{\phi}}{\operatorname{argmin}}\big[L[\boldsymbol{\phi}]\big] \\ &= \underset{\boldsymbol{\phi}}{\operatorname{argmin}}\left[\sum_{i=1}^{I}\ell_i[\mathbf{x}_i,\mathbf{y}_i]\right],\end{aligned} \tag{9.1}$$

$$\hat{\boldsymbol{\phi}} = \underset{\boldsymbol{\phi}}{\operatorname{argmin}}\left[\sum_{i=1}^{I}\ell_i[\mathbf{x}_i,\mathbf{y}_i] + \lambda\cdot \mathrm{g}[\boldsymbol{\phi}]\right], \tag{9.2}$$

$$\hat{\boldsymbol{\phi}} = \underset{\boldsymbol{\phi}}{\operatorname{argmax}}\left[\prod_{i=1}^{I}Pr(\mathbf{y}_i|\mathbf{x}_i,\boldsymbol{\phi})\right]. \tag{9.3}$$

$$\hat{\boldsymbol{\phi}} = \underset{\boldsymbol{\phi}}{\operatorname{argmax}}\left[\prod_{i=1}^{I}Pr(\mathbf{y}_i|\mathbf{x}_i,\boldsymbol{\phi})Pr(\boldsymbol{\phi})\right]. \tag{9.4}$$

$$\hat{\boldsymbol{\phi}} = \underset{\boldsymbol{\phi}}{\operatorname{argmin}}\left[\sum_{i=1}^{I}\ell_i[\mathbf{x}_i,\mathbf{y}_i] + \lambda\sum_{j}\phi_j^2\right], \tag{9.5}$$

$$\frac{d\boldsymbol{\phi}}{dt} = -\frac{\partial L}{\partial \boldsymbol{\phi}}. \tag{9.6}$$

$$\boldsymbol{\phi}_{t+1} = \boldsymbol{\phi}_t - \alpha\frac{\partial L[\boldsymbol{\phi}_t]}{\partial \boldsymbol{\phi}}, \tag{9.7}$$

$$\tilde{L}_{GD}[\boldsymbol{\phi}] = L[\boldsymbol{\phi}] + \frac{\alpha}{4}\left\|\frac{\partial L}{\partial \boldsymbol{\phi}}\right\|^2. \tag{9.8}$$

$$\begin{aligned}\tilde{L}_{SGD}[\boldsymbol{\phi}] &= \tilde{L}_{GD}[\boldsymbol{\phi}] + \frac{\alpha}{4B}\sum_{b=1}^{B}\left\|\frac{\partial L_b}{\partial \boldsymbol{\phi}} - \frac{\partial L}{\partial \boldsymbol{\phi}}\right\|^2 \\ &= L[\boldsymbol{\phi}] + \frac{\alpha}{4}\left\|\frac{\partial L}{\partial \boldsymbol{\phi}}\right\|^2 + \frac{\alpha}{4B}\sum_{b=1}^{B}\left\|\frac{\partial L_b}{\partial \boldsymbol{\phi}} - \frac{\partial L}{\partial \boldsymbol{\phi}}\right\|^2.\end{aligned} \tag{9.9}$$

$$L = \frac{1}{I}\sum_{i=1}^{I}\ell_i[\mathbf{x}_i, y_i] \qquad \text{and} \qquad L_b = \frac{1}{|\mathcal{B}|}\sum_{i\in\mathcal{B}_b}\ell_i[\mathbf{x}_i, y_i]. \tag{9.10}$$

$$Pr(\boldsymbol{\phi}|\{\mathbf{x}_i, \mathbf{y}_i\}) = \frac{\prod_{i=1}^{I} Pr(\mathbf{y}_i|\mathbf{x}_i, \boldsymbol{\phi})Pr(\boldsymbol{\phi})}{\int \prod_{i=1}^{I} Pr(\mathbf{y}_i|\mathbf{x}_i, \boldsymbol{\phi})Pr(\boldsymbol{\phi})d\boldsymbol{\phi}}, \tag{9.11}$$

$$Pr(\mathbf{y}|\mathbf{x}, \{\mathbf{x}_i, \mathbf{y}_i\}) = \int Pr(\mathbf{y}|\mathbf{x}, \boldsymbol{\phi})Pr(\boldsymbol{\phi}|\{\mathbf{x}_i, \mathbf{y}_i\})d\boldsymbol{\phi}. \tag{9.12}$$

$$\boldsymbol{\phi} \longleftarrow (1-\lambda')\boldsymbol{\phi} - \alpha\frac{\partial L}{\partial \boldsymbol{\phi}}, \tag{9.13}$$

$$\boldsymbol{\phi}_1 = \boldsymbol{\phi}_0 + \alpha\cdot\mathbf{g}[\boldsymbol{\phi}_0], \tag{9.14}$$

$$\frac{d\boldsymbol{\phi}}{dt} = \mathbf{g}[\boldsymbol{\phi}]. \tag{9.15}$$

$$\frac{d\boldsymbol{\phi}}{dt} \approx \mathbf{g}[\boldsymbol{\phi}] + \alpha\mathbf{g}_1[\boldsymbol{\phi}] + \ldots, \tag{9.16}$$

$$\begin{aligned}\boldsymbol{\phi}[\alpha] &\approx \boldsymbol{\phi} + \alpha\frac{d\boldsymbol{\phi}}{dt} + \frac{\alpha^2}{2}\frac{d^2\boldsymbol{\phi}}{dt^2}\bigg|_{\boldsymbol{\phi}=\boldsymbol{\phi}_0} \\ &\approx \boldsymbol{\phi} + \alpha\left(\mathbf{g}[\boldsymbol{\phi}] + \alpha\mathbf{g}_1[\boldsymbol{\phi}]\right) + \frac{\alpha^2}{2}\left(\frac{\partial\mathbf{g}[\boldsymbol{\phi}]}{\partial\boldsymbol{\phi}}\frac{d\boldsymbol{\phi}}{dt} + \alpha\frac{\partial\mathbf{g}_1[\boldsymbol{\phi}]}{\partial\boldsymbol{\phi}}\frac{d\boldsymbol{\phi}}{dt}\right)\bigg|_{\boldsymbol{\phi}=\boldsymbol{\phi}_0} \\ &= \boldsymbol{\phi} + \alpha\left(\mathbf{g}[\boldsymbol{\phi}] + \alpha\mathbf{g}_1[\boldsymbol{\phi}]\right) + \frac{\alpha^2}{2}\left(\frac{\partial\mathbf{g}[\boldsymbol{\phi}]}{\partial\boldsymbol{\phi}}\mathbf{g}[\boldsymbol{\phi}] + \alpha\frac{\partial\mathbf{g}_1[\boldsymbol{\phi}]}{\partial\boldsymbol{\phi}}\mathbf{g}[\boldsymbol{\phi}]\right)\bigg|_{\boldsymbol{\phi}=\boldsymbol{\phi}_0} \\ &\approx \boldsymbol{\phi} + \alpha\mathbf{g}[\boldsymbol{\phi}] + \alpha^2\left(\mathbf{g}_1[\boldsymbol{\phi}] + \frac{1}{2}\frac{\partial\mathbf{g}[\boldsymbol{\phi}]}{\partial\boldsymbol{\phi}}\mathbf{g}[\boldsymbol{\phi}]\right)\bigg|_{\boldsymbol{\phi}=\boldsymbol{\phi}_0},\end{aligned} \tag{9.17}$$

$$\mathbf{g}_1[\boldsymbol{\phi}] = -\frac{1}{2}\frac{\partial \mathbf{g}[\boldsymbol{\phi}]}{\partial \boldsymbol{\phi}}\mathbf{g}[\boldsymbol{\phi}]. \tag{9.18}$$

$$\begin{aligned}\frac{d\boldsymbol{\phi}}{dt} &\approx \mathbf{g}[\boldsymbol{\phi}] + \alpha\mathbf{g}_1[\boldsymbol{\phi}] \\ &= -\frac{\partial L}{\partial \boldsymbol{\phi}} - \frac{\alpha}{2}\left(\frac{\partial^2 L}{\partial \boldsymbol{\phi}^2}\right)\frac{\partial L}{\partial \boldsymbol{\phi}}.\end{aligned} \tag{9.19}$$

$$L_{GD}[\boldsymbol{\phi}] = L[\boldsymbol{\phi}] + \frac{\alpha}{4}\left\|\frac{\partial L}{\partial \boldsymbol{\phi}}\right\|^2, \tag{9.20}$$

$$Pr(\boldsymbol{\phi}) = \prod_{j=1}^{J} \text{Norm}_{\phi_j}[0, \sigma^2_{\boldsymbol{\phi}}], \tag{9.21}$$

$$\boldsymbol{\phi} \longleftarrow (1-\lambda)\boldsymbol{\phi} - \alpha\frac{\partial L}{\partial \boldsymbol{\phi}}, \tag{9.22}$$

$$\tilde{L}[\boldsymbol{\phi}] = L[\boldsymbol{\phi}] + \frac{\lambda}{2\alpha}\sum_k \phi_k^2, \tag{9.23}$$

Chapter 10

Convolutional networks

$$\mathbf{f}\big[\mathbf{t}[\mathbf{x}]\big] = \mathbf{f}[\mathbf{x}]. \tag{10.1}$$

$$\mathbf{f}\big[\mathbf{t}[\mathbf{x}]\big] = \mathbf{t}\big[\mathbf{f}[\mathbf{x}]\big]. \tag{10.2}$$

$$z_i = \omega_1 x_{i-1} + \omega_2 x_i + \omega_3 x_{i+1}, \tag{10.3}$$

$$\begin{aligned} h_i &= \mathrm{a}\left[\beta + \omega_1 x_{i-1} + \omega_2 x_i + \omega_3 x_{i+1}\right] \\ &= \mathrm{a}\left[\beta + \sum_{j=1}^{3} \omega_j x_{i+j-2}\right], \end{aligned} \tag{10.4}$$

$$h_i = \mathrm{a}\left[\beta_i + \sum_{j=1}^{D} \omega_{ij} x_j\right]. \tag{10.5}$$

$$h_{ij} = \mathrm{a}\left[\beta + \sum_{m=1}^{3} \sum_{n=1}^{3} \omega_{mn} x_{i+m-2,j+n-2}\right], \tag{10.6}$$

Chapter 11

Residual networks

$$
\begin{aligned}
\mathbf{h}_1 &= \mathbf{f}_1[\mathbf{x}, \boldsymbol{\phi}_1] \\
\mathbf{h}_2 &= \mathbf{f}_2[\mathbf{h}_1, \boldsymbol{\phi}_2] \\
\mathbf{h}_3 &= \mathbf{f}_3[\mathbf{h}_2, \boldsymbol{\phi}_3] \\
\mathbf{y} &= \mathbf{f}_4[\mathbf{h}_3, \boldsymbol{\phi}_4], \qquad (11.1)
\end{aligned}
$$

$$
\mathbf{y} = \mathbf{f}_4\Bigl[\mathbf{f}_3\Bigl[\mathbf{f}_2\bigl[\mathbf{f}_1[\mathbf{x}, \boldsymbol{\phi}_1], \boldsymbol{\phi}_2\bigr], \boldsymbol{\phi}_3\Bigr], \boldsymbol{\phi}_4\Bigr]. \qquad (11.2)
$$

$$
\frac{\partial \mathbf{y}}{\partial \mathbf{f}_1} = \frac{\partial \mathbf{f}_2}{\partial \mathbf{f}_1}\frac{\partial \mathbf{f}_3}{\partial \mathbf{f}_2}\frac{\partial \mathbf{f}_4}{\partial \mathbf{f}_3}. \qquad (11.3)
$$

$$
\begin{aligned}
\mathbf{h}_1 &= \mathbf{x} + \mathbf{f}_1[\mathbf{x}, \boldsymbol{\phi}_1] \\
\mathbf{h}_2 &= \mathbf{h}_1 + \mathbf{f}_2[\mathbf{h}_1, \boldsymbol{\phi}_2] \\
\mathbf{h}_3 &= \mathbf{h}_2 + \mathbf{f}_3[\mathbf{h}_2, \boldsymbol{\phi}_3] \\
\mathbf{y} &= \mathbf{h}_3 + \mathbf{f}_4[\mathbf{h}_3, \boldsymbol{\phi}_4], \qquad (11.4)
\end{aligned}
$$

$$
\begin{aligned}
\mathbf{y} = \mathbf{x} &+ \mathbf{f}_1[\mathbf{x}] \qquad (11.5) \\
&+ \mathbf{f}_2\bigl[\mathbf{x} + \mathbf{f}_1[\mathbf{x}]\bigr] \\
&+ \mathbf{f}_3\Bigl[\mathbf{x} + \mathbf{f}_1[\mathbf{x}] + \mathbf{f}_2\bigl[\mathbf{x} + \mathbf{f}_1[\mathbf{x}]\bigr]\Bigr] \\
&+ \mathbf{f}_4\Bigl[\mathbf{x} + \mathbf{f}_1[\mathbf{x}] + \mathbf{f}_2\bigl[\mathbf{x} + \mathbf{f}_1[\mathbf{x}]\bigr] + \mathbf{f}_3\Bigl[\mathbf{x} + \mathbf{f}_1[\mathbf{x}] + \mathbf{f}_2\bigl[\mathbf{x} + \mathbf{f}_1[\mathbf{x}]\bigr]\Bigr]\Bigr],
\end{aligned}
$$

$$
\frac{\partial \mathbf{y}}{\partial \mathbf{f}_1} = \mathbf{I} + \frac{\partial \mathbf{f}_2}{\partial \mathbf{f}_1} + \left(\frac{\partial \mathbf{f}_3}{\partial \mathbf{f}_1} + \frac{\partial \mathbf{f}_2}{\partial \mathbf{f}_1}\frac{\partial \mathbf{f}_3}{\partial \mathbf{f}_2}\right) + \left(\frac{\partial \mathbf{f}_4}{\partial \mathbf{f}_1} + \frac{\partial \mathbf{f}_2}{\partial \mathbf{f}_1}\frac{\partial \mathbf{f}_4}{\partial \mathbf{f}_2} + \frac{\partial \mathbf{f}_3}{\partial \mathbf{f}_1}\frac{\partial \mathbf{f}_4}{\partial \mathbf{f}_3} + \frac{\partial \mathbf{f}_2}{\partial \mathbf{f}_1}\frac{\partial \mathbf{f}_3}{\partial \mathbf{f}_2}\frac{\partial \mathbf{f}_4}{\partial \mathbf{f}_3}\right), \qquad (11.6)
$$

$$\begin{aligned} m_h &= \frac{1}{|\mathcal{B}|}\sum_{i\in\mathcal{B}} h_i \\ s_h &= \sqrt{\frac{1}{|\mathcal{B}|}\sum_{i\in\mathcal{B}}(h_i - m_h)^2}, \end{aligned} \tag{11.7}$$

$$h_i \leftarrow \frac{h_i - m_h}{s_h + \epsilon} \qquad\qquad \forall i \in \mathcal{B}, \tag{11.8}$$

$$h_i \leftarrow \gamma h_i + \delta \qquad\qquad \forall i \in \mathcal{B}. \tag{11.9}$$

$$\begin{aligned} f_1 &= \mathbb{E}[z_i] & f_5 &= \sqrt{f_4 + \epsilon} \\ f_{2i} &= z_i - f_1 & f_6 &= 1/f_5 \\ f_{3i} &= f_{2i}^2 & f_{7i} &= f_{2i} \times f_6 \\ f_4 &= \mathbb{E}[f_{3i}] & z_i' &= f_{7i} \times \gamma + \delta, \end{aligned} \tag{11.10}$$

Chapter 12

Transformers

$$\mathbf{f}[\mathbf{x}] = \mathbf{ReLU}[\boldsymbol{\beta} + \boldsymbol{\Omega}\mathbf{x}], \tag{12.1}$$

$$\mathbf{v}_m = \boldsymbol{\beta}_v + \boldsymbol{\Omega}_v \mathbf{x}_m, \tag{12.2}$$

$$\mathbf{sa}_n[\mathbf{x}_1, \ldots, \mathbf{x}_N] = \sum_{m=1}^{N} a[\mathbf{x}_m, \mathbf{x}_n]\mathbf{v}_m. \tag{12.3}$$

$$\begin{aligned} \mathbf{q}_n &= \boldsymbol{\beta}_q + \boldsymbol{\Omega}_q \mathbf{x}_n \\ \mathbf{k}_m &= \boldsymbol{\beta}_k + \boldsymbol{\Omega}_k \mathbf{x}_m, \end{aligned} \tag{12.4}$$

$$\begin{aligned} a[\mathbf{x}_m, \mathbf{x}_n] &= \mathrm{softmax}_m \left[\mathbf{k}_\bullet^T \mathbf{q}_n\right] \\ &= \frac{\exp\left[\mathbf{k}_m^T \mathbf{q}_n\right]}{\sum_{m'=1}^{N} \exp\left[\mathbf{k}_{m'}^T \mathbf{q}_n\right]}, \end{aligned} \tag{12.5}$$

$$\begin{aligned} \mathbf{V}[\mathbf{X}] &= \boldsymbol{\beta}_v \mathbf{1}^T + \boldsymbol{\Omega}_v \mathbf{X} \\ \mathbf{Q}[\mathbf{X}] &= \boldsymbol{\beta}_q \mathbf{1}^T + \boldsymbol{\Omega}_q \mathbf{X} \\ \mathbf{K}[\mathbf{X}] &= \boldsymbol{\beta}_k \mathbf{1}^T + \boldsymbol{\Omega}_k \mathbf{X}, \end{aligned} \tag{12.6}$$

$$\mathbf{Sa}[\mathbf{X}] = \mathbf{V}[\mathbf{X}] \cdot \mathbf{Softmax}\Big[\mathbf{K}[\mathbf{X}]^T \mathbf{Q}[\mathbf{X}]\Big], \tag{12.7}$$

$$\mathbf{Sa}[\mathbf{X}] = \mathbf{V}\cdot\mathbf{Softmax}\Big[\mathbf{K}^T\mathbf{Q}\Big]. \tag{12.8}$$

$$\mathbf{Sa}[\mathbf{X}] = \mathbf{V}\cdot\mathbf{Softmax}\left[\frac{\mathbf{K}^T\mathbf{Q}}{\sqrt{D_q}}\right]. \tag{12.9}$$

$$\begin{aligned}\mathbf{V}_h &= \boldsymbol{\beta}_{vh}\mathbf{1}^T + \boldsymbol{\Omega}_{vh}\mathbf{X}\\ \mathbf{Q}_h &= \boldsymbol{\beta}_{qh}\mathbf{1}^T + \boldsymbol{\Omega}_{qh}\mathbf{X}\\ \mathbf{K}_h &= \boldsymbol{\beta}_{kh}\mathbf{1}^T + \boldsymbol{\Omega}_{kh}\mathbf{X}.\end{aligned} \tag{12.10}$$

$$\mathbf{Sa}_h[\mathbf{X}] = \mathbf{V}_h\cdot\mathbf{Softmax}\left[\frac{\mathbf{K}_h^T\mathbf{Q}_h}{\sqrt{D_q}}\right], \tag{12.11}$$

$$\mathbf{MhSa}[\mathbf{X}] = \boldsymbol{\Omega}_c\Big[\mathbf{Sa}_1[\mathbf{X}]^T, \mathbf{Sa}_2[\mathbf{X}]^T, \ldots, \mathbf{Sa}_H[\mathbf{X}]^T\Big]^T. \tag{12.12}$$

$$\begin{aligned}\mathbf{X} &\leftarrow \mathbf{X} + \mathbf{MhSa}[\mathbf{X}]\\ \mathbf{X} &\leftarrow \mathbf{LayerNorm}[\mathbf{X}]\\ \mathbf{x}_n &\leftarrow \mathbf{x}_n + \mathbf{mlp}[\mathbf{x}_n] \qquad \forall\, n \in \{1,\ldots,N\}\\ \mathbf{X} &\leftarrow \mathbf{LayerNorm}[\mathbf{X}],\end{aligned} \tag{12.13}$$

$$\begin{aligned}Pr(\text{It takes great courage to let yourself appear weak}) \quad =& \\ Pr(\text{It}) \times Pr(\text{takes}|\text{It}) \times Pr(\text{great}|\text{It takes}) \times Pr(\text{courage}|\text{It takes great}) \times& \\ Pr(\text{to}|\text{It takes great courage}) \times Pr(\text{let}|\text{It takes great courage to}) \times& \\ Pr(\text{yourself}|\text{It takes great courage to let}) \times& \\ Pr(\text{appear}|\text{It takes great courage to let yourself}) \times& \\ Pr(\text{weak}|\text{It takes great courage to let yourself appear}).&\end{aligned} \tag{12.14}$$

$$Pr(t_1, t_2, \ldots, t_N) = Pr(t_1)\prod_{n=2}^{N} Pr(t_n|t_1, \ldots, t_{n-1}). \tag{12.15}$$

$$\mathbf{Sa}[\mathbf{X}] = \mathbf{V}\cdot\mathbf{Softmax}\left[\frac{\mathbf{K}^T\mathbf{Q}}{\sqrt{D_q}}\right], \tag{12.16}$$

$$
\begin{aligned}
\mathbf{V} &= \boldsymbol{\beta}_v \mathbf{1}^T + \boldsymbol{\Omega}_v \mathbf{X} \\
\mathbf{Q} &= \boldsymbol{\beta}_q \mathbf{1}^T + \boldsymbol{\Omega}_q (\mathbf{X} + \boldsymbol{\Pi}) \\
\mathbf{K} &= \boldsymbol{\beta}_k \mathbf{1}^T + \boldsymbol{\Omega}_k (\mathbf{X} + \boldsymbol{\Pi}).
\end{aligned} \tag{12.17}
$$

$$
\mathbf{Sa}[\mathbf{XP}] = \mathbf{Sa}[\mathbf{X}]\mathbf{P}. \tag{12.18}
$$

$$
y_i = \text{softmax}_i[\mathbf{z}] = \frac{\exp[z_i]}{\sum_{j=1}^{5} \exp[z_j]}, \tag{12.19}
$$

$$
a[\mathbf{x}_m, \mathbf{x}_n] = \text{softmax}_m \left[\mathbf{k}_\bullet^T \mathbf{q}_n\right] = \frac{\exp\left[\mathbf{k}_m^T \mathbf{q}_n\right]}{\sum_{m'=1}^{N} \exp\left[\mathbf{k}_{m'}^T \mathbf{q}_n\right]}. \tag{12.20}
$$

Chapter 13

Graph neural networks

$$\begin{aligned}\mathbf{X}' &= \mathbf{X}\mathbf{P}\\ \mathbf{A}' &= \mathbf{P}^T\mathbf{A}\mathbf{P},\end{aligned}\tag{13.1}$$

$$Pr(y=1|\mathbf{X},\mathbf{A}) = \mathrm{sig}\left[\beta_K + \boldsymbol{\omega}_K\mathbf{H}_K\mathbf{1}/N\right],\tag{13.2}$$

$$Pr(y^{(n)}=1|\mathbf{X},\mathbf{A}) = \mathrm{sig}\left[\beta_K + \boldsymbol{\omega}_K\mathbf{h}_K^{(n)}\right].\tag{13.3}$$

$$Pr(y^{(mn)}=1|\mathbf{X},\mathbf{A}) = \mathrm{sig}\left[\mathbf{h}^{(m)T}\mathbf{h}^{(n)}\right].\tag{13.4}$$

$$\begin{aligned}\mathbf{H}_1 &= \mathbf{F}[\mathbf{X},\mathbf{A},\boldsymbol{\phi}_0]\\ \mathbf{H}_2 &= \mathbf{F}[\mathbf{H}_1,\mathbf{A},\boldsymbol{\phi}_1]\\ \mathbf{H}_3 &= \mathbf{F}[\mathbf{H}_2,\mathbf{A},\boldsymbol{\phi}_2]\\ \vdots &= \vdots\\ \mathbf{H}_K &= \mathbf{F}[\mathbf{H}_{K-1},\mathbf{A},\boldsymbol{\phi}_{K-1}],\end{aligned}\tag{13.5}$$

$$\mathbf{H}_{k+1}\mathbf{P} = \mathbf{F}[\mathbf{H}_k\mathbf{P},\mathbf{P}^T\mathbf{A}\mathbf{P},\boldsymbol{\phi}_k].\tag{13.6}$$

$$y = \mathrm{sig}\left[\beta_K + \boldsymbol{\omega}_K\mathbf{H}_K\mathbf{1}/N\right] = \mathrm{sig}\left[\beta_K + \boldsymbol{\omega}_K\mathbf{H}_K\mathbf{P}\mathbf{1}/N\right],\tag{13.7}$$

$$\mathbf{agg}[n,k] = \sum_{m\in\mathrm{ne}[n]} \mathbf{h}_k^{(m)}, \tag{13.8}$$

$$\mathbf{h}_{k+1}^{(n)} = \mathbf{a}\left[\boldsymbol{\beta}_k + \boldsymbol{\Omega}_k \cdot \mathbf{h}_k^{(n)} + \boldsymbol{\Omega}_k \cdot \mathbf{agg}[n,k]\right]. \tag{13.9}$$

$$\begin{aligned}\mathbf{H}_{k+1} &= \mathbf{a}\left[\boldsymbol{\beta}_k\mathbf{1}^T + \boldsymbol{\Omega}_k\mathbf{H}_k + \boldsymbol{\Omega}_k\mathbf{H}_k\mathbf{A}\right]\\ &= \mathbf{a}\left[\boldsymbol{\beta}_k\mathbf{1}^T + \boldsymbol{\Omega}_k\mathbf{H}_k(\mathbf{A}+\mathbf{I})\right],\end{aligned} \tag{13.10}$$

$$\begin{aligned}\mathbf{H}_1 &= \mathbf{a}\left[\boldsymbol{\beta}_0\mathbf{1}^T + \boldsymbol{\Omega}_0\mathbf{X}(\mathbf{A}+\mathbf{I})\right]\\ \mathbf{H}_2 &= \mathbf{a}\left[\boldsymbol{\beta}_1\mathbf{1}^T + \boldsymbol{\Omega}_1\mathbf{H}_1(\mathbf{A}+\mathbf{I})\right]\\ \vdots &= \vdots\\ \mathbf{H}_K &= \mathbf{a}\left[\boldsymbol{\beta}_{K-1}\mathbf{1}^T + \boldsymbol{\Omega}_{K-1}\mathbf{H}_{k-1}(\mathbf{A}+\mathbf{I})\right]\\ \mathrm{f}[\mathbf{X},\mathbf{A},\boldsymbol{\Phi}] &= \mathrm{sig}\left[\beta_K + \boldsymbol{\omega}_K\mathbf{H}_K\mathbf{1}/N\right],\end{aligned} \tag{13.11}$$

$$\mathrm{f}[\mathbf{X},\mathbf{A},\boldsymbol{\Phi}] = \mathbf{sig}\left[\beta_K\mathbf{1}^T + \boldsymbol{\omega}_K\mathbf{H}_K\right], \tag{13.12}$$

$$\mathbf{H}_{k+1} = \mathbf{a}\Big[\boldsymbol{\beta}_k\mathbf{1}^T + \boldsymbol{\Omega}_k\mathbf{H}_k(\mathbf{A}+\mathbf{I})\Big]. \tag{13.13}$$

$$\mathbf{H}_{k+1} = \mathbf{a}\Big[\boldsymbol{\beta}_k\mathbf{1}^T + \boldsymbol{\Omega}_k\mathbf{H}_k(\mathbf{A}+(1+\epsilon_k)\mathbf{I})\Big]. \tag{13.14}$$

$$\begin{aligned}\mathbf{H}_{k+1} &= \mathbf{a}\left[\boldsymbol{\beta}_k\mathbf{1}^T + \boldsymbol{\Omega}_k\mathbf{H}_k\mathbf{A} + \boldsymbol{\Psi}_k\mathbf{H}_k\right]\\ &= \mathbf{a}\left[\boldsymbol{\beta}_k\mathbf{1}^T + \begin{bmatrix}\boldsymbol{\Omega}_k & \boldsymbol{\Psi}_k\end{bmatrix}\begin{bmatrix}\mathbf{H}_k\mathbf{A}\\ \mathbf{H}_k\end{bmatrix}\right]\\ &= \mathbf{a}\left[\boldsymbol{\beta}_k\mathbf{1}^T + \boldsymbol{\Omega}'_k\begin{bmatrix}\mathbf{H}_k\mathbf{A}\\ \mathbf{H}_k\end{bmatrix}\right],\end{aligned} \tag{13.15}$$

$$\mathbf{H}_{k+1} = \begin{bmatrix}\mathbf{a}\left[\boldsymbol{\beta}_k\mathbf{1}^T + \boldsymbol{\Omega}_k\mathbf{H}_k\mathbf{A}\right]\\ \mathbf{H}_k\end{bmatrix}. \tag{13.16}$$

$$\mathbf{agg}[n] = \frac{1}{|\mathrm{ne}[n]|} \sum_{m \in \mathrm{ne}[n]} \mathbf{h}_m, \tag{13.17}$$

$$\mathbf{H}_{k+1} = \mathbf{a}\Big[\boldsymbol{\beta}_k \mathbf{1}^T + \boldsymbol{\Omega}_k \mathbf{H}_k (\mathbf{A}\mathbf{D}^{-1} + \mathbf{I})\Big]. \tag{13.18}$$

$$\mathbf{agg}[n] = \sum_{m \in \mathrm{ne}[n]} \frac{\mathbf{h}_m}{\sqrt{|\mathrm{ne}[n]||\mathrm{ne}[m]|}}, \tag{13.19}$$

$$\mathbf{H}_{k+1} = \mathbf{a}\Big[\boldsymbol{\beta}_k \mathbf{1}^T + \boldsymbol{\Omega}_k \mathbf{H}_k (\mathbf{D}^{-1/2}\mathbf{A}\mathbf{D}^{-1/2} + \mathbf{I})\Big]. \tag{13.20}$$

$$\mathbf{agg}[n] = \max_{m \in \mathrm{ne}[n]} \big[\mathbf{h}_m\big], \tag{13.21}$$

$$\mathbf{H}'_k = \boldsymbol{\beta}_k \mathbf{1}^T + \boldsymbol{\Omega}_k \mathbf{H}_k. \tag{13.22}$$

$$s_{mn} = \mathrm{a}\left[\boldsymbol{\phi}_k^T \begin{bmatrix} \mathbf{h}'_m \\ \mathbf{h}'_n \end{bmatrix}\right]. \tag{13.23}$$

$$\mathbf{H}_{k+1} = \mathbf{a}\Big[\mathbf{H}'_k \cdot \mathbf{Softmask}[\mathbf{S}, \mathbf{A} + \mathbf{I}]\Big], \tag{13.24}$$

$$\mathbf{h}_n \leftarrow \mathbf{f}\left[\mathbf{x}_n, \mathbf{x}_{m \in \mathrm{ne}[n]}, \mathbf{e}_{e \in \mathrm{nee}[n]}, \mathbf{h}_{m \in \mathrm{ne}[n]}, \boldsymbol{\phi}\right], \tag{13.25}$$

$$\mathbf{h}_{k+1}^{(n)} = \mathbf{mlp}\left[(1 + \epsilon_k)\, \mathbf{h}_k^{(n)} + \sum_{m \in \mathrm{ne}[n]} \mathbf{h}_k^{(m)}\right]. \tag{13.26}$$

$$\mathbf{A}_1 = \begin{bmatrix} 0 & 1 & 1 & 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 1 & 1 & 1 & 0 \\ 1 & 0 & 0 & 0 & 0 & 1 & 1 \\ 0 & 1 & 0 & 0 & 0 & 1 & 1 \\ 0 & 1 & 0 & 0 & 0 & 0 & 1 \\ 0 & 1 & 1 & 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 1 & 1 & 0 & 0 \end{bmatrix} \quad \text{and} \quad \mathbf{A}_2 = \begin{bmatrix} 0 & 0 & 1 & 1 & 0 & 0 & 1 \\ 0 & 0 & 1 & 1 & 1 & 0 & 0 \\ 1 & 1 & 0 & 0 & 0 & 0 & 0 \\ 1 & 1 & 0 & 0 & 1 & 1 & 1 \\ 0 & 1 & 0 & 1 & 0 & 0 & 1 \\ 0 & 0 & 0 & 1 & 0 & 0 & 1 \\ 1 & 0 & 0 & 1 & 1 & 1 & 0 \end{bmatrix}.$$

$$\mathrm{sig}\left[\beta_K + \boldsymbol{\omega}_K \mathbf{H}_K \mathbf{1}\right] = \mathrm{sig}\left[\beta_K + \boldsymbol{\omega}_K \mathbf{H}_K \mathbf{P}\mathbf{1}\right], \tag{13.27}$$

$$\begin{aligned}\mathbf{H}_{k+1} &= \mathrm{GraphLayer}[\mathbf{H}_k, \mathbf{A}] \\ &= \mathbf{a}\left[\boldsymbol{\beta}_k \mathbf{1}^T + \boldsymbol{\Omega}_k \begin{bmatrix}\mathbf{H}_k \\ \mathbf{H}_k\mathbf{A}\end{bmatrix}\right],\end{aligned} \tag{13.28}$$

$$\mathrm{GraphLayer}[\mathbf{H}_k, \mathbf{A}]\mathbf{P} = \mathrm{GraphLayer}[\mathbf{H}_k\mathbf{P}, \mathbf{P}^T\mathbf{A}\mathbf{P}], \tag{13.29}$$

$$\mathbf{agg}[n] = \frac{1}{1 + |\mathrm{ne}[n]|}\left(\mathbf{h}_n + \sum_{m\in \mathrm{ne}[n]} \mathbf{h}_m\right). \tag{13.30}$$

Chapter 14

Unsupervised learning

$$L[\boldsymbol{\phi}] = -\sum_{i=1}^{I} \log\Big[Pr(\mathbf{x}_i|\boldsymbol{\phi})\Big]. \tag{14.1}$$

$$IS = \exp\left[\frac{1}{I}\sum_{i=1}^{I} D_{KL}\Big[Pr(y|\mathbf{x}_i^*)||Pr(y)\Big]\right], \tag{14.2}$$

$$Pr(y) = \frac{1}{I}\sum_{i=1}^{I} Pr(y|\mathbf{x}_i^*). \tag{14.3}$$

Chapter 15

Generative adversarial networks

$$x_j^* = \mathrm{g}[z_j, \theta] = z_j + \theta, \tag{15.1}$$

$$\hat{\boldsymbol{\phi}} = \underset{\boldsymbol{\phi}}{\mathrm{argmin}} \left[\sum_i -(1-y_i)\log\Bigl[1-\mathrm{sig}[\mathrm{f}[\mathbf{x}_i,\boldsymbol{\phi}]]\Bigr] - y_i \log\Bigl[\mathrm{sig}[\mathrm{f}[\mathbf{x}_i,\boldsymbol{\phi}]]\Bigr]\right], \tag{15.2}$$

$$\hat{\boldsymbol{\phi}} = \underset{\boldsymbol{\phi}}{\mathrm{argmin}} \left[\sum_j -\log\Bigl[1-\mathrm{sig}[\mathrm{f}[\mathbf{x}_j^*,\boldsymbol{\phi}]]\Bigr] - \sum_i \log\Bigl[\mathrm{sig}[\mathrm{f}[\mathbf{x}_i,\boldsymbol{\phi}]]\Bigr]\right], \tag{15.3}$$

$$\hat{\boldsymbol{\theta}} = \underset{\boldsymbol{\theta}}{\mathrm{argmax}} \left[\min_{\boldsymbol{\phi}} \left[\sum_j -\log\Bigl[1-\mathrm{sig}[\mathrm{f}[\mathbf{g}[\mathbf{z}_j,\boldsymbol{\theta}],\boldsymbol{\phi}]]\Bigr] - \sum_i \log\Bigl[\mathrm{sig}[\mathrm{f}[\mathbf{x}_i,\boldsymbol{\phi}]]\Bigr]\right]\right]. \tag{15.4}$$

$$\begin{aligned} L[\boldsymbol{\phi}] &= \sum_j -\log\Bigl[1-\mathrm{sig}[\mathrm{f}[\mathbf{g}[\mathbf{z}_j,\boldsymbol{\theta}],\boldsymbol{\phi}]]\Bigr] - \sum_i \log\Bigl[\mathrm{sig}[\mathrm{f}[\mathbf{x}_i,\boldsymbol{\phi}]]\Bigr] \\ L[\boldsymbol{\theta}] &= \sum_j \log\Bigl[1-\mathrm{sig}[\mathrm{f}[\mathbf{g}[\mathbf{z}_j,\boldsymbol{\theta}],\boldsymbol{\phi}]]\Bigr], \end{aligned} \tag{15.5}$$

$$\begin{aligned} L[\boldsymbol{\phi}] &= -\frac{1}{J}\sum_{j=1}^{J}\Bigl(\log\Bigl[1-\mathrm{sig}[\mathrm{f}[\mathbf{x}_j^*,\boldsymbol{\phi}]]\Bigr]\Bigr) - \frac{1}{I}\sum_{i=1}^{I}\Bigl(\log\Bigl[\mathrm{sig}[\mathrm{f}[\mathbf{x}_i,\boldsymbol{\phi}]]\Bigr]\Bigr) \\ &\approx -\mathbb{E}_{\mathbf{x}^*}\Bigl[\log\Bigl[1-\mathrm{sig}[\mathrm{f}[\mathbf{x}^*,\boldsymbol{\phi}]]\Bigr]\Bigr] - \mathbb{E}_{\mathbf{x}}\Bigl[\log\Bigl[\mathrm{sig}[\mathrm{f}[\mathbf{x},\boldsymbol{\phi}]]\Bigr]\Bigr] \\ &= -\int Pr(\mathbf{x}^*)\log\Bigl[1-\mathrm{sig}[\mathrm{f}[\mathbf{x}^*,\boldsymbol{\phi}]]\Bigr]d\mathbf{x}^* - \int Pr(\mathbf{x})\log\Bigl[\mathrm{sig}[\mathrm{f}[\mathbf{x},\boldsymbol{\phi}]]\Bigr]d\mathbf{x}, \end{aligned} \tag{15.6}$$

$$Pr(\text{real}|\tilde{\mathbf{x}}) = \text{sig}\big[\text{f}[\tilde{\mathbf{x}}, \boldsymbol{\phi}]\big] = \frac{Pr(\tilde{\mathbf{x}}|\text{real})}{Pr(\tilde{\mathbf{x}}|\text{generated}) + Pr(\tilde{\mathbf{x}}|\text{real})} = \frac{Pr(\mathbf{x})}{Pr(\mathbf{x}^*) + Pr(\mathbf{x})}, \tag{15.7}$$

$$\begin{aligned} L[\boldsymbol{\phi}] &= -\int Pr(\mathbf{x}^*)\log\Big[1-\text{sig}[\text{f}[\mathbf{x}^*,\boldsymbol{\phi}]]\Big]d\mathbf{x}^* - \int Pr(\mathbf{x})\log\Big[\text{sig}[\text{f}[\mathbf{x},\boldsymbol{\phi}]]\Big]d\mathbf{x} \qquad (15.8)\\ &= -\int Pr(\mathbf{x}^*)\log\left[1-\frac{Pr(\mathbf{x})}{Pr(\mathbf{x}^*)+Pr(\mathbf{x})}\right]d\mathbf{x}^* - \int Pr(\mathbf{x})\log\left[\frac{Pr(\mathbf{x})}{Pr(\mathbf{x}^*)+Pr(\mathbf{x})}\right]d\mathbf{x}\\ &= -\int Pr(\mathbf{x}^*)\log\left[\frac{Pr(\mathbf{x}^*)}{Pr(\mathbf{x}^*)+Pr(\mathbf{x})}\right]d\mathbf{x}^* - \int Pr(\mathbf{x})\log\left[\frac{Pr(\mathbf{x})}{Pr(\mathbf{x}^*)+Pr(\mathbf{x})}\right]d\mathbf{x}. \end{aligned}$$

$$\begin{aligned} &D_{JS}\Big[Pr(\mathbf{x}^*)\;||\;Pr(\mathbf{x})\Big] \qquad (15.9)\\ &\quad= \frac{1}{2}D_{KL}\left[Pr(\mathbf{x}^*)\,\middle|\middle|\,\frac{Pr(\mathbf{x}^*)+Pr(\mathbf{x})}{2}\right] + \frac{1}{2}D_{KL}\left[Pr(\mathbf{x})\,\middle|\middle|\,\frac{Pr(\mathbf{x}^*)+Pr(\mathbf{x})}{2}\right]\\ &\quad= \underbrace{\frac{1}{2}\int Pr(\mathbf{x}^*)\log\left[\frac{2Pr(\mathbf{x}^*)}{Pr(\mathbf{x}^*)+Pr(\mathbf{x})}\right]d\mathbf{x}^*}_{\text{quality}} + \underbrace{\frac{1}{2}\int Pr(\mathbf{x})\log\left[\frac{2Pr(\mathbf{x})}{Pr(\mathbf{x}^*)+Pr(\mathbf{x})}\right]d\mathbf{x}}_{\text{coverage}}. \end{aligned}$$

$$D_w\Big[Pr(x)||q(x)\Big] = \min_{\mathbf{P}}\left[\sum_{i,j}P_{ij}\cdot|i-j|\right], \tag{15.10}$$

$$\begin{array}{rclll} \sum_j P_{ij} &=& Pr(x=i) & & \text{initial distribution of } Pr(x)\\ \sum_i P_{ij} &=& q(x=j) & & \text{initial distribution of } q(x)\\ P_{ij} &\geq& 0 & & \text{non-negative masses.} \end{array} \tag{15.11}$$

$$D_w\Big[Pr(x)||q(x)\Big] = \max_{\mathbf{f}}\left[\sum_i Pr(x=i)f_i - \sum_j q(x=j)f_j\right], \tag{15.12}$$

$$|f_{i+1} - f_i| < 1. \tag{15.13}$$

$$D_w\Big[Pr(\mathbf{x}), q(\mathbf{x})\Big] = \min_{\pi[\bullet,\bullet]}\left[\iint \pi(\mathbf{x}_1,\mathbf{x}_2)\cdot||\mathbf{x}_1-\mathbf{x}_2||d\mathbf{x}_1d\mathbf{x}_2\right], \tag{15.14}$$

$$D_w\Big[Pr(\mathbf{x}), q(\mathbf{x})\Big] = \max_{\text{f}[\mathbf{x}]}\left[\int Pr(\mathbf{x})\text{f}[\mathbf{x}]d\mathbf{x} - \int q(\mathbf{x})\text{f}[\mathbf{x}]d\mathbf{x}\right], \tag{15.15}$$

$$
\begin{aligned}
L[\boldsymbol{\phi}] &= \sum_j \mathrm{f}[\mathbf{x}_j^*, \boldsymbol{\phi}] - \sum_i \mathrm{f}[\mathbf{x}_i, \boldsymbol{\phi}] \\
&= \sum_j \mathrm{f}[\mathbf{g}[\mathbf{z}_j, \boldsymbol{\theta}], \boldsymbol{\phi}] - \sum_i \mathrm{f}[\mathbf{x}_i, \boldsymbol{\phi}],
\end{aligned} \tag{15.16}
$$

$$
\left| \frac{\partial \mathrm{f}[\mathbf{x}, \boldsymbol{\phi}]}{\partial \mathbf{x}} \right| < 1. \tag{15.17}
$$

$$
\mathbf{b} = \left[Pr(x\!=\!1), Pr(x\!=\!2), Pr(x\!=\!3), Pr(x\!=\!4), q(x\!=\!1), q(x\!=\!2), q(x\!=\!3), q(x\!=\!4)\right]^T \tag{15.18}
$$

$$
Pr(z) = \begin{cases} 0 & z < 0 \\ 1 & 0 \leq z \leq 1 \\ 0 & z > 1 \end{cases}, \quad \text{and} \quad Pr(z) = \begin{cases} 0 & z < a \\ 1 & a \leq z \leq a+1 \\ 0 & z > a \end{cases}. \tag{15.19}
$$

$$
D_{kl} = \log\left[\frac{\sigma_2}{\sigma_1}\right] + \frac{\sigma_1^2 + (\mu_1 - \mu_2)^2}{2\sigma_2^2} - \frac{1}{2}, \tag{15.20}
$$

$$
D_w = (\mu_1 - \mu_2)^2 + \sigma_1 + \sigma_2 - 2\sqrt{\sigma_1 \sigma_2}, \tag{15.21}
$$

Chapter 16

Normalizing flows

$$Pr(x|\boldsymbol{\phi}) \quad = \quad \left|\frac{\partial \mathrm{f}[z,\boldsymbol{\phi}]}{\partial z}\right|^{-1}\cdot Pr(z), \tag{16.1}$$

$$\begin{aligned}\hat{\boldsymbol{\phi}} \quad &= \quad \underset{\boldsymbol{\phi}}{\mathrm{argmax}}\left[\prod_{i=1}^{I} Pr(x_i|\boldsymbol{\phi})\right] \\ &= \quad \underset{\boldsymbol{\phi}}{\mathrm{argmin}}\left[\sum_{i=1}^{I} -\log\Big[Pr(x_i|\boldsymbol{\phi})\Big]\right] \\ &= \quad \underset{\boldsymbol{\phi}}{\mathrm{argmin}}\left[\sum_{i=1}^{I} \log\left[\left|\frac{\partial \mathrm{f}[z_i,\boldsymbol{\phi}]}{\partial z_i}\right|\right] - \log\big[Pr(z_i)\big]\right],\end{aligned} \tag{16.2}$$

$$Pr(\mathbf{x}|\boldsymbol{\phi}) = \left|\frac{\partial \mathbf{f}[\mathbf{z},\boldsymbol{\phi}]}{\partial \mathbf{z}}\right|^{-1}\cdot Pr(\mathbf{z}), \tag{16.3}$$

$$\mathbf{x} = \mathbf{f}[\mathbf{z},\boldsymbol{\phi}] = \mathbf{f}_K\bigg[\mathbf{f}_{K-1}\Big[\ldots\mathbf{f}_2\big[\mathbf{f}_1[\mathbf{z},\boldsymbol{\phi}_1],\boldsymbol{\phi}_2\big],\ldots\boldsymbol{\phi}_{K-1}\Big],\boldsymbol{\phi}_K\bigg]. \tag{16.4}$$

$$\mathbf{z} = \mathbf{f}^{-1}[\mathbf{x},\boldsymbol{\phi}] = \mathbf{f}_1^{-1}\bigg[\mathbf{f}_2^{-1}\Big[\ldots\mathbf{f}_{K-1}^{-1}\big[\mathbf{f}_K^{-1}[\mathbf{x},\boldsymbol{\phi}_K],\boldsymbol{\phi}_{K-1}\big],\ldots\boldsymbol{\phi}_2\Big],\boldsymbol{\phi}_1\bigg]. \tag{16.5}$$

$$\frac{\partial \mathbf{f}[\mathbf{z},\boldsymbol{\phi}]}{\partial \mathbf{z}} = \frac{\partial \mathbf{f}_K[\mathbf{f}_{K-1},\boldsymbol{\phi}_K]}{\partial \mathbf{f}_{K-1}}\cdot\frac{\partial \mathbf{f}_{K-1}[\mathbf{f}_{K-2},\boldsymbol{\phi}_{K-1}]}{\partial \mathbf{f}_{K-2}}\ldots\frac{\partial \mathbf{f}_2[\mathbf{f}_1,\boldsymbol{\phi}_2]}{\partial \mathbf{f}_1}\cdot\frac{\partial \mathbf{f}_1[\mathbf{z},\boldsymbol{\phi}_1]}{\partial \mathbf{z}}, \tag{16.6}$$

$$\left|\frac{\partial \mathbf{f}[\mathbf{z},\boldsymbol{\phi}]}{\partial \mathbf{z}}\right| = \left|\frac{\partial \mathbf{f}_K[\mathbf{f}_{K-1},\boldsymbol{\phi}_K]}{\partial \mathbf{f}_{K-1}}\right|\cdot\left|\frac{\partial \mathbf{f}_{K-1}[\mathbf{f}_{K-2},\boldsymbol{\phi}_{K-1}]}{\partial \mathbf{f}_{K-2}}\right|\ldots\left|\frac{\partial \mathbf{f}_2[\mathbf{f}_1,\boldsymbol{\phi}_2]}{\partial \mathbf{f}_1}\right|\cdot\left|\frac{\partial \mathbf{f}_1[\mathbf{z},\boldsymbol{\phi}_1]}{\partial \mathbf{z}}\right|. \tag{16.7}$$

$$
\begin{aligned}
\hat{\boldsymbol{\phi}} &= \underset{\boldsymbol{\phi}}{\operatorname{argmax}}\left[\prod_{i=1}^{I} Pr(\mathbf{z}_i)\cdot\left|\frac{\partial\mathbf{f}[\mathbf{z}_i,\boldsymbol{\phi}]}{\partial\mathbf{z}_i}\right|^{-1}\right]\\
&= \underset{\boldsymbol{\phi}}{\operatorname{argmin}}\left[\sum_{i=1}^{I}\log\left[\left|\frac{\partial\mathbf{f}[\mathbf{z}_i,\boldsymbol{\phi}]}{\partial\mathbf{z}_i}\right|\right]-\log\bigl[Pr(\mathbf{z}_i)\bigr]\right],
\end{aligned}
\tag{16.8}
$$

$$
\boldsymbol{\Omega} = \mathbf{PL}(\mathbf{U}+\mathbf{D}),
\tag{16.9}
$$

$$
\mathbf{f}[\mathbf{h}] = \Bigl[\mathrm{f}[h_1,\boldsymbol{\phi}],\mathrm{f}[h_2,\boldsymbol{\phi}],\ldots\mathrm{f}[h_D,\boldsymbol{\phi}]\Bigr]^T.
\tag{16.10}
$$

$$
\left|\frac{\partial\mathbf{f}[\mathbf{h}]}{\partial\mathbf{h}}\right| = \prod_{d=1}^{D}\left|\frac{\partial\mathrm{f}[h_d]}{\partial h_d}\right|.
\tag{16.11}
$$

$$
\mathrm{f}[h,\boldsymbol{\phi}] = \left(\sum_{k=1}^{b-1}\phi_k\right) + (hK-b+1)\phi_b,
\tag{16.12}
$$

$$
\begin{aligned}
\mathbf{h}'_1 &= \mathbf{h}_1\\
\mathbf{h}'_2 &= \mathbf{g}\Bigl[\mathbf{h}_2,\boldsymbol{\phi}[\mathbf{h}_1]\Bigr].
\end{aligned}
\tag{16.13}
$$

$$
\begin{aligned}
\mathbf{h}_1 &= \mathbf{h}'_1\\
\mathbf{h}_2 &= \mathbf{g}^{-1}\Bigl[\mathbf{h}'_2,\boldsymbol{\phi}[\mathbf{h}_1]\Bigr].
\end{aligned}
\tag{16.14}
$$

$$
h'_d = \mathrm{g}\Bigl[h_d,\boldsymbol{\phi}[\mathbf{h}_{1:d-1}]\Bigr].
\tag{16.15}
$$

$$
\begin{aligned}
h'_1 &= \mathrm{g}\Bigl[h_1,\boldsymbol{\phi}\Bigr]\\
h'_2 &= \mathrm{g}\Bigl[h_2,\boldsymbol{\phi}[h_1]\Bigr]\\
h'_3 &= \mathrm{g}\Bigl[h_3,\boldsymbol{\phi}[h_{1:2}]\Bigr]\\
h'_4 &= \mathrm{g}\Bigl[h_4,\boldsymbol{\phi}[h_{1:3}]\Bigr].
\end{aligned}
\tag{16.16}
$$

$$
\begin{aligned}
h_1 &= \mathrm{g}^{-1}\Big[h_1', \boldsymbol{\phi}\Big] \\
h_2 &= \mathrm{g}^{-1}\Big[h_2', \boldsymbol{\phi}[h_1]\Big] \\
h_3 &= \mathrm{g}^{-1}\Big[h_3', \boldsymbol{\phi}[h_{1:2}]\Big] \\
h_4 &= \mathrm{g}^{-1}\Big[h_4', \boldsymbol{\phi}[h_{1:3}]\Big].
\end{aligned} \tag{16.17}
$$

$$
\begin{aligned}
\mathbf{h}_1' &= \mathbf{h}_1 + \mathbf{f}_1[\mathbf{h}_2, \boldsymbol{\phi}_1] \\
\mathbf{h}_2' &= \mathbf{h}_2 + \mathbf{f}_2[\mathbf{h}_1', \boldsymbol{\phi}_2],
\end{aligned} \tag{16.18}
$$

$$
\begin{aligned}
\mathbf{h}_2 &= \mathbf{h}_2' - \mathbf{f}_2[\mathbf{h}_1', \boldsymbol{\phi}_2] \\
\mathbf{h}_1 &= \mathbf{h}_1' - \mathbf{f}_1[\mathbf{h}_2, \boldsymbol{\phi}_1].
\end{aligned} \tag{16.19}
$$

$$
\mathrm{dist}\Big[\mathrm{f}[z'], \mathrm{f}[z]\Big] < \beta \cdot \mathrm{dist}\Big[z', z\Big] \qquad\qquad \forall\, z, z', \tag{16.20}
$$

$$
y = z + \mathrm{f}[z] \tag{16.21}
$$

$$
\begin{aligned}
\log\left[\left|\mathbf{I} + \frac{\partial \mathbf{f}[\mathbf{h}, \boldsymbol{\phi}]}{\partial \mathbf{h}}\right|\right] &= \mathrm{trace}\left[\log\left[\mathbf{I} + \frac{\partial \mathbf{f}[\mathbf{h}, \boldsymbol{\phi}]}{\partial \mathbf{h}}\right]\right] \\
&= \sum_{k=1}^{\infty} \frac{(-1)^{k-1}}{k} \mathrm{trace}\left[\frac{\partial \mathbf{f}[\mathbf{h}, \boldsymbol{\phi}]}{\partial \mathbf{h}}\right]^k,
\end{aligned} \tag{16.22}
$$

$$
\begin{aligned}
\mathrm{trace}[\mathbf{A}] &= \mathrm{trace}\left[\mathbf{A}\mathbb{E}\left[\boldsymbol{\epsilon}\boldsymbol{\epsilon}^T\right]\right] \\
&= \mathrm{trace}\left[\mathbb{E}\left[\mathbf{A}\boldsymbol{\epsilon}\boldsymbol{\epsilon}^T\right]\right] \\
&= \mathbb{E}\left[\mathrm{trace}\left[\mathbf{A}\boldsymbol{\epsilon}\boldsymbol{\epsilon}^T\right]\right] \\
&= \mathbb{E}\left[\mathrm{trace}\left[\boldsymbol{\epsilon}^T\mathbf{A}\boldsymbol{\epsilon}\right]\right] \\
&= \mathbb{E}\left[\boldsymbol{\epsilon}^T\mathbf{A}\boldsymbol{\epsilon}\right],
\end{aligned} \tag{16.23}
$$

$$
\begin{aligned}
\mathrm{trace}[\mathbf{A}] &= \mathbb{E}\left[\boldsymbol{\epsilon}^T\mathbf{A}\boldsymbol{\epsilon}\right] \\
&\approx \frac{1}{I}\sum_{i=1}^{I} \boldsymbol{\epsilon}_i^T\mathbf{A}\boldsymbol{\epsilon}_i.
\end{aligned} \tag{16.24}
$$

$$\hat{\boldsymbol{\phi}} = \underset{\boldsymbol{\phi}}{\operatorname{argmin}} \left[\mathrm{KL} \left[\frac{1}{I} \sum_{i=1}^{I} \delta\big[\mathbf{x} - \mathrm{f}[\mathbf{z}_i, \boldsymbol{\phi}]\big] \bigg|\bigg| q(\mathbf{x}) \right] \right]. \tag{16.25}$$

$$\hat{\boldsymbol{\phi}} = \underset{\boldsymbol{\phi}}{\operatorname{argmin}} \left[\mathrm{KL} \left[\frac{1}{I} \sum_{i=1}^{I} \delta[\mathbf{x} - \mathbf{x}_i] \bigg|\bigg| Pr(\mathbf{x}_i, \boldsymbol{\phi}) \right] \right]. \tag{16.26}$$

$$Pr(z) = \frac{1}{\sqrt{2\pi}} \exp\left[\frac{-z^2}{2}\right], \tag{16.27}$$

$$x = \mathrm{f}[z] = \frac{1}{1 + \exp[-z]}. \tag{16.28}$$

$$\boldsymbol{\Omega}_1 = \begin{bmatrix} 2 & 0 & 0 & 0 \\ 0 & -5 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 2 \end{bmatrix} \qquad \boldsymbol{\Omega}_2 = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 2 & 4 & 0 & 0 \\ 1 & -1 & 2 & 0 \\ 4 & -2 & -2 & 1 \end{bmatrix}. \tag{16.29}$$

$$Pr(\mathbf{x}) = \mathrm{Pr}(\mathbf{z}) \cdot \left| \frac{\partial \mathbf{f}[\mathbf{z}]}{\partial \mathbf{z}} \right|^{-1}. \tag{16.30}$$

$$\mathrm{LReLU}[z] = \begin{cases} 0.1z & z < 0 \\ z & z \geq 0 \end{cases}. \tag{16.31}$$

$$\mathbf{f}[\mathbf{z}] = \Big[\mathrm{LReLU}[z_1], \mathrm{LReLU}[z_2], \ldots, \mathrm{LReLU}[z_D]\Big]^T. \tag{16.32}$$

$$\mathbf{h}' = \mathrm{f}[h, \boldsymbol{\phi}] = \sqrt{[Kh - b + 1]\,\phi_b} + \sum_{k=1}^{b-1} \sqrt{\phi_k}, \tag{16.33}$$

Chapter 17

Variational autoencoders

$$Pr(\mathbf{x}) = \int Pr(\mathbf{x}, \mathbf{z}) d\mathbf{z}. \tag{17.1}$$

$$Pr(\mathbf{x}) = \int Pr(\mathbf{x}|\mathbf{z}) Pr(\mathbf{z}) d\mathbf{z}. \tag{17.2}$$

$$\begin{aligned} Pr(z = n) &= \lambda_n \\ Pr(x|z = n) &= \text{Norm}_x\big[\mu_n, \sigma_n^2\big]. \end{aligned} \tag{17.3}$$

$$\begin{aligned} Pr(x) &= \sum_{n=1}^{N} Pr(x, z = n) \\ &= \sum_{n=1}^{N} Pr(x|z = n) \cdot Pr(z = n) \\ &= \sum_{n=1}^{N} \lambda_n \cdot \text{Norm}_x\big[\mu_n, \sigma_n^2\big]. \end{aligned} \tag{17.4}$$

$$Pr(\mathbf{z}) = \text{Norm}_{\mathbf{z}}[\mathbf{0}, \mathbf{I}]. \tag{17.5}$$

$$Pr(\mathbf{x}|\mathbf{z}, \boldsymbol{\phi}) = \text{Norm}_{\mathbf{x}}\Big[\mathbf{f}[\mathbf{z}, \boldsymbol{\phi}], \sigma^2\mathbf{I}\Big]. \tag{17.6}$$

$$\begin{aligned}
Pr(\mathbf{x}|\boldsymbol{\phi}) &= \int Pr(\mathbf{x},\mathbf{z}|\boldsymbol{\phi})d\mathbf{z} \\
&= \int Pr(\mathbf{x}|\mathbf{z},\boldsymbol{\phi})\cdot Pr(\mathbf{z})d\mathbf{z} \\
&= \int \text{Norm}_{\mathbf{x}}\Big[\mathbf{f}[\mathbf{z},\boldsymbol{\phi}],\sigma^2\mathbf{I}\Big]\cdot\text{Norm}_{\mathbf{z}}\left[\mathbf{0},\mathbf{I}\right]d\mathbf{z}.
\end{aligned} \tag{17.7}$$

$$\hat{\boldsymbol{\phi}} = \underset{\boldsymbol{\phi}}{\text{argmax}}\left[\sum_{i=1}^{I}\log\Big[Pr(\mathbf{x}_i|\boldsymbol{\phi})\Big]\right], \tag{17.8}$$

$$Pr(\mathbf{x}_i|\boldsymbol{\phi}) = \int \text{Norm}_{\mathbf{x}_i}[\mathbf{f}[\mathbf{z},\boldsymbol{\phi}],\sigma^2\mathbf{I}]\cdot\text{Norm}_{\mathbf{z}}[\mathbf{0},\mathbf{I}]d\mathbf{z}. \tag{17.9}$$

$$g[\mathbb{E}[y]] \geq \mathbb{E}\big[g[y]\big]. \tag{17.10}$$

$$\log\big[\mathbb{E}[y]\big] \geq \mathbb{E}\big[\log[y]\big], \tag{17.11}$$

$$\log\left[\int Pr(y)ydy\right] \geq \int Pr(y)\log[y]dy. \tag{17.12}$$

$$\log\left[\int Pr(y)h[y]dy\right] \geq \int Pr(y)\log[h[y]]dy. \tag{17.13}$$

$$\begin{aligned}
\log[Pr(\mathbf{x}|\boldsymbol{\phi})] &= \log\left[\int Pr(\mathbf{x},\mathbf{z}|\boldsymbol{\phi})d\mathbf{z}\right] \\
&= \log\left[\int q(\mathbf{z})\frac{Pr(\mathbf{x},\mathbf{z}|\boldsymbol{\phi})}{q(\mathbf{z})}d\mathbf{z}\right],
\end{aligned} \tag{17.14}$$

$$\log\left[\int q(\mathbf{z})\frac{Pr(\mathbf{x},\mathbf{z}|\boldsymbol{\phi})}{q(\mathbf{z})}d\mathbf{z}\right] \geq \int q(\mathbf{z})\log\left[\frac{Pr(\mathbf{x},\mathbf{z}|\boldsymbol{\phi})}{q(\mathbf{z})}\right]d\mathbf{z}, \tag{17.15}$$

$$\text{ELBO}[\boldsymbol{\theta},\boldsymbol{\phi}] = \int q(\mathbf{z}|\boldsymbol{\theta})\log\left[\frac{Pr(\mathbf{x},\mathbf{z}|\boldsymbol{\phi})}{q(\mathbf{z}|\boldsymbol{\theta})}\right]d\mathbf{z}. \tag{17.16}$$

$$
\begin{aligned}
\text{ELBO}[\boldsymbol{\theta},\boldsymbol{\phi}] &= \int q(\mathbf{z}|\boldsymbol{\theta}) \log\left[\frac{Pr(\mathbf{x},\mathbf{z}|\boldsymbol{\phi})}{q(\mathbf{z}|\boldsymbol{\theta})}\right] d\mathbf{z} \\
&= \int q(\mathbf{z}|\boldsymbol{\theta}) \log\left[\frac{Pr(\mathbf{z}|\mathbf{x},\boldsymbol{\phi})Pr(\mathbf{x}|\boldsymbol{\phi})}{q(\mathbf{z}|\boldsymbol{\theta})}\right] d\mathbf{z} \\
&= \int q(\mathbf{z}|\boldsymbol{\theta}) \log\big[Pr(\mathbf{x}|\boldsymbol{\phi})\big] d\mathbf{z} + \int q(\mathbf{z}|\boldsymbol{\theta}) \log\left[\frac{Pr(\mathbf{z}|\mathbf{x},\boldsymbol{\phi})}{q(\mathbf{z}|\boldsymbol{\theta})}\right] d\mathbf{z} \\
&= \log\big[Pr(\mathbf{x}|\boldsymbol{\phi})\big] + \int q(\mathbf{z}|\boldsymbol{\theta}) \log\left[\frac{Pr(\mathbf{z}|\mathbf{x},\boldsymbol{\phi})}{q(\mathbf{z}|\boldsymbol{\theta})}\right] d\mathbf{z} \\
&= \log\big[Pr(\mathbf{x}|\boldsymbol{\phi})\big] - \mathrm{D}_{KL}\Big[q(\mathbf{z}|\boldsymbol{\theta})\Big|\Big|Pr(\mathbf{z}|\mathbf{x},\boldsymbol{\phi})\Big]. \qquad (17.17)
\end{aligned}
$$

$$
\begin{aligned}
\text{ELBO}[\boldsymbol{\theta},\boldsymbol{\phi}] &= \int q(\mathbf{z}|\boldsymbol{\theta}) \log\left[\frac{Pr(\mathbf{x},\mathbf{z}|\boldsymbol{\phi})}{q(\mathbf{z}|\boldsymbol{\theta})}\right] d\mathbf{z} \\
&= \int q(\mathbf{z}|\boldsymbol{\theta}) \log\left[\frac{Pr(\mathbf{x}|\mathbf{z},\boldsymbol{\phi})Pr(\mathbf{z})}{q(\mathbf{z}|\boldsymbol{\theta})}\right] d\mathbf{z} \\
&= \int q(\mathbf{z}|\boldsymbol{\theta}) \log\left[Pr(\mathbf{x}|\mathbf{z},\boldsymbol{\phi})\right] d\mathbf{z} + \int q(\mathbf{z}|\boldsymbol{\theta}) \log\left[\frac{Pr(\mathbf{z})}{q(\mathbf{z}|\boldsymbol{\theta})}\right] d\mathbf{z} \\
&= \int q(\mathbf{z}|\boldsymbol{\theta}) \log\big[Pr(\mathbf{x}|\mathbf{z},\boldsymbol{\phi})\big] d\mathbf{z} - \mathrm{D}_{KL}\Big[q(\mathbf{z}|\boldsymbol{\theta})\Big|\Big|Pr(\mathbf{z})\Big], \qquad (17.18)
\end{aligned}
$$

$$
Pr(\mathbf{z}|\mathbf{x},\boldsymbol{\phi}) = \frac{Pr(\mathbf{x}|\mathbf{z},\boldsymbol{\phi})Pr(\mathbf{z})}{Pr(\mathbf{x}|\boldsymbol{\phi})}, \qquad (17.19)
$$

$$
q(\mathbf{z}|\mathbf{x},\boldsymbol{\theta}) = \text{Norm}_{\mathbf{z}}\Big[\mathbf{g}_{\boldsymbol{\mu}}[\mathbf{x},\boldsymbol{\theta}], \mathbf{g}_{\boldsymbol{\Sigma}}[\mathbf{x},\boldsymbol{\theta}]\Big], \qquad (17.20)
$$

$$
\text{ELBO}[\boldsymbol{\theta},\boldsymbol{\phi}] = \int q(\mathbf{z}|\mathbf{x},\boldsymbol{\theta}) \log\big[Pr(\mathbf{x}|\mathbf{z},\boldsymbol{\phi})\big] d\mathbf{z} - \mathrm{D}_{KL}\Big[q(\mathbf{z}|\mathbf{x},\boldsymbol{\theta})\Big|\Big|Pr(\mathbf{z})\Big], \qquad (17.21)
$$

$$
\mathbb{E}_{\mathbf{z}}\big[\mathrm{a}[\mathbf{z}]\big] = \int \mathrm{a}[\mathbf{z}] q(\mathbf{z}|\mathbf{x},\boldsymbol{\theta}) d\mathbf{z} \approx \frac{1}{N}\sum_{n=1}^{N} \mathrm{a}[\mathbf{z}_n^*], \qquad (17.22)
$$

$$
\text{ELBO}[\boldsymbol{\theta},\boldsymbol{\phi}] \approx \log\big[Pr(\mathbf{x}|\mathbf{z}^*,\boldsymbol{\phi})\big] - \mathrm{D}_{KL}\Big[q(\mathbf{z}|\mathbf{x},\boldsymbol{\theta})\Big|\Big|Pr(\mathbf{z})\Big]. \qquad (17.23)
$$

$$
\mathrm{D}_{KL}\Big[q(\mathbf{z}|\mathbf{x},\boldsymbol{\theta})\Big|\Big|Pr(\mathbf{z})\Big] = \frac{1}{2}\left(\text{Tr}[\boldsymbol{\Sigma}] + \boldsymbol{\mu}^T\boldsymbol{\mu} - D_{\mathbf{z}} - \log\Big[\det[\boldsymbol{\Sigma}]\Big]\right). \qquad (17.24)
$$

$$\mathbf{z}^* = \boldsymbol{\mu} + \boldsymbol{\Sigma}^{1/2}\boldsymbol{\epsilon}^*, \tag{17.25}$$

$$\begin{aligned} Pr(\mathbf{x}) &= \int Pr(\mathbf{x}|\mathbf{z})Pr(\mathbf{z})d\mathbf{z} \\ &= \mathbb{E}_{\mathbf{z}}\Bigl[Pr(\mathbf{x}|\mathbf{z})\Bigr] \\ &= \mathbb{E}_{\mathbf{z}}\Bigl[\text{Norm}_{\mathbf{x}}[\mathbf{f}[\mathbf{z},\boldsymbol{\phi}],\sigma^2\mathbf{I}]\Bigr]. \end{aligned} \tag{17.26}$$

$$Pr(\mathbf{x}) \approx \frac{1}{N}\sum_{n=1}^{N} Pr(\mathbf{x}|\mathbf{z}_n). \tag{17.27}$$

$$\begin{aligned} Pr(\mathbf{x}) &= \int Pr(\mathbf{x}|\mathbf{z})Pr(\mathbf{z})d\mathbf{z} \\ &= \int \frac{Pr(\mathbf{x}|\mathbf{z})Pr(\mathbf{z})}{q(\mathbf{z})}q(\mathbf{z})d\mathbf{z} \\ &= \mathbb{E}_{q(\mathbf{z})}\left[\frac{Pr(\mathbf{x}|\mathbf{z})Pr(\mathbf{z})}{q(\mathbf{z})}\right] \\ &\approx \frac{1}{N}\sum_{n=1}^{N}\frac{Pr(\mathbf{x}|\mathbf{z}_n)Pr(\mathbf{z}_n)}{q(\mathbf{z}_n)}, \end{aligned} \tag{17.28}$$

$$L_{\text{new}} = -\text{ELBO}[\boldsymbol{\theta},\boldsymbol{\phi}] + \lambda_1\mathbb{E}_{Pr(\mathbf{x})}\Bigl[\text{r}_1\bigl[q(\mathbf{z}|\mathbf{x},\boldsymbol{\theta})\bigr]\Bigr] + \lambda_2\text{r}_2\bigl[q(\mathbf{z}|\boldsymbol{\theta})\bigr]. \tag{17.29}$$

$$\text{ELBO}[\boldsymbol{\theta},\boldsymbol{\phi}] \approx \log\bigl[Pr(\mathbf{x}|\mathbf{z}^*,\boldsymbol{\phi})\bigr] - \beta\cdot\text{D}_{KL}\Bigl[q(\mathbf{z}|\mathbf{x},\boldsymbol{\theta})\Bigl|\Bigr|Pr(\mathbf{z})\Bigr], \tag{17.30}$$

$$g\bigl[\mathbb{E}[y]\bigr] \leq \mathbb{E}\bigl[g[y]\bigr]. \tag{17.31}$$

$$\text{D}_{KL}\Bigl[q(\mathbf{z}|\mathbf{x})\Bigl|\Bigr|Pr(\mathbf{z}|\mathbf{x},\boldsymbol{\phi})\Bigr] = \int q(\mathbf{z}|\mathbf{x})\log\left[\frac{q(\mathbf{z}|\mathbf{x})}{Pr(\mathbf{z}|\mathbf{x},\boldsymbol{\phi})}\right]d\mathbf{z}. \tag{17.32}$$

$$\frac{\partial}{\partial\boldsymbol{\phi}}\mathbb{E}_{Pr(x|\boldsymbol{\phi})}\bigl[\text{f}[x]\bigr], \tag{17.33}$$

$$\begin{aligned} \frac{\partial}{\partial\boldsymbol{\phi}}\mathbb{E}_{Pr(x|\boldsymbol{\phi})}\bigl[\text{f}[x]\bigr] &= \mathbb{E}_{Pr(x|\boldsymbol{\phi})}\left[\text{f}[x]\frac{\partial}{\partial\boldsymbol{\phi}}\log\bigl[Pr(x|\boldsymbol{\phi})\bigr]\right] \\ &\approx \frac{1}{I}\sum_{i=1}^{I}\text{f}[x_i]\frac{\partial}{\partial\boldsymbol{\phi}}\log\bigl[Pr(x_i|\boldsymbol{\phi})\bigr]. \end{aligned} \tag{17.34}$$

Chapter 18

Diffusion models

$$\begin{aligned}\mathbf{z}_1 &= \sqrt{1-\beta_1}\cdot\mathbf{x}+\sqrt{\beta_1}\cdot\boldsymbol{\epsilon}_1 \\ \mathbf{z}_t &= \sqrt{1-\beta_t}\cdot\mathbf{z}_{t-1}+\sqrt{\beta_t}\cdot\boldsymbol{\epsilon}_t \qquad \forall\, t\in 2,\ldots,T,\end{aligned} \tag{18.1}$$

$$\begin{aligned}q(\mathbf{z}_1|\mathbf{x}) &= \text{Norm}_{\mathbf{z}_1}\left[\sqrt{1-\beta_1}\mathbf{x},\beta_1\mathbf{I}\right] \\ q(\mathbf{z}_t|\mathbf{z}_{t-1}) &= \text{Norm}_{\mathbf{z}_t}\left[\sqrt{1-\beta_t}\mathbf{z}_{t-1},\beta_t\mathbf{I}\right] \qquad \forall\, t\in\{2,\ldots,T\}.\end{aligned} \tag{18.2}$$

$$q(\mathbf{z}_{1\ldots T}|\mathbf{x}) = q(\mathbf{z}_1|\mathbf{x})\prod_{t=2}^{T} q(\mathbf{z}_t|\mathbf{z}_{t-1}). \tag{18.3}$$

$$\begin{aligned}\mathbf{z}_1 &= \sqrt{1-\beta_1}\cdot\mathbf{x}+\sqrt{\beta_1}\cdot\boldsymbol{\epsilon}_1 \\ \mathbf{z}_2 &= \sqrt{1-\beta_2}\cdot\mathbf{z}_1+\sqrt{\beta_2}\cdot\boldsymbol{\epsilon}_2.\end{aligned} \tag{18.4}$$

$$\begin{aligned}\mathbf{z}_2 &= \sqrt{1-\beta_2}\left(\sqrt{1-\beta_1}\cdot\mathbf{x}+\sqrt{\beta_1}\cdot\boldsymbol{\epsilon}_1\right)+\sqrt{\beta_2}\cdot\boldsymbol{\epsilon}_2 \\ &= \sqrt{1-\beta_2}\left(\sqrt{1-\beta_1}\cdot\mathbf{x}+\sqrt{1-(1-\beta_1)}\cdot\boldsymbol{\epsilon}_1\right)+\sqrt{\beta_2}\cdot\boldsymbol{\epsilon}_2 \\ &= \sqrt{(1-\beta_2)(1-\beta_1)}\cdot\mathbf{x}+\sqrt{1-\beta_2-(1-\beta_2)(1-\beta_1)}\cdot\boldsymbol{\epsilon}_1+\sqrt{\beta_2}\cdot\boldsymbol{\epsilon}_2.\end{aligned} \tag{18.5}$$

$$\mathbf{z}_2 = \sqrt{(1-\beta_2)(1-\beta_1)}\cdot\mathbf{x}+\sqrt{1-(1-\beta_2)(1-\beta_1)}\cdot\boldsymbol{\epsilon}, \tag{18.6}$$

$$\mathbf{z}_t = \sqrt{\alpha_t}\cdot\mathbf{x}+\sqrt{1-\alpha_t}\cdot\boldsymbol{\epsilon}, \tag{18.7}$$

$$q(\mathbf{z}_t|\mathbf{x}) = \mathrm{Norm}_{\mathbf{z}_t}\Big[\sqrt{\alpha_t}\cdot\mathbf{x}, (1-\alpha_t)\mathbf{I}\Big]. \tag{18.8}$$

$$\begin{aligned} q(\mathbf{z}_t) &= \iint q(\mathbf{z}_{1\ldots t}, \mathbf{x}) d\mathbf{z}_{1\ldots t-1} d\mathbf{x} \\ &= \iint q(\mathbf{z}_{1\ldots t}|\mathbf{x}) Pr(\mathbf{x}) d\mathbf{z}_{1\ldots t-1} d\mathbf{x}, \end{aligned} \tag{18.9}$$

$$q(\mathbf{z}_t) = \int q(\mathbf{z}_t|\mathbf{x}) Pr(\mathbf{x}) d\mathbf{x}. \tag{18.10}$$

$$q(\mathbf{z}_{t-1}|\mathbf{z}_t) = \frac{q(\mathbf{z}_t|\mathbf{z}_{t-1}) q(\mathbf{z}_{t-1})}{q(\mathbf{z}_t)}. \tag{18.11}$$

$$\begin{aligned} q(\mathbf{z}_{t-1}|\mathbf{z}_t, \mathbf{x}) &= \frac{q(\mathbf{z}_t|\mathbf{z}_{t-1}, \mathbf{x}) q(\mathbf{z}_{t-1}|\mathbf{x})}{q(\mathbf{z}_t|\mathbf{x})} \\ &\propto q(\mathbf{z}_t|\mathbf{z}_{t-1}) q(\mathbf{z}_{t-1}|\mathbf{x}) \\ &= \mathrm{Norm}_{\mathbf{z}_t}\left[\sqrt{1-\beta_t}\cdot\mathbf{z}_{t-1}, \beta_t\mathbf{I}\right] \mathrm{Norm}_{\mathbf{z}_{t-1}}\left[\sqrt{\alpha_{t-1}}\cdot\mathbf{x}, (1-\alpha_{t-1})\mathbf{I}\right] \\ &\propto \mathrm{Norm}_{\mathbf{z}_{t-1}}\left[\frac{1}{\sqrt{1-\beta_t}}\mathbf{z}_t, \frac{\beta_t}{1-\beta_t}\mathbf{I}\right] \mathrm{Norm}_{\mathbf{z}_{t-1}}\left[\sqrt{\alpha_{t-1}}\cdot\mathbf{x}, (1-\alpha_{t-1})\mathbf{I}\right] \end{aligned} \tag{18.12}$$

$$\mathrm{Norm}_{\mathbf{v}}[\mathbf{A}\mathbf{w}, \mathbf{B}] \propto \mathrm{Norm}_{\mathbf{w}}\Big[\big(\mathbf{A}^T\mathbf{B}^{-1}\mathbf{A}\big)^{-1}\mathbf{A}^T\mathbf{B}^{-1}\mathbf{v}, \big(\mathbf{A}^T\mathbf{B}^{-1}\mathbf{A}\big)^{-1}\Big], \tag{18.13}$$

$$\begin{aligned} \mathrm{Norm}_{\mathbf{w}}[\mathbf{a}, \mathbf{A}]\cdot\mathrm{Norm}_{\mathbf{w}}[\mathbf{b}, \mathbf{B}] \quad &\propto \\ &\mathrm{Norm}_{\mathbf{w}}\Big[\big(\mathbf{A}^{-1}+\mathbf{B}^{-1}\big)^{-1}(\mathbf{A}^{-1}\mathbf{a}+\mathbf{B}^{-1}\mathbf{b}), \big(\mathbf{A}^{-1}+\mathbf{B}^{-1}\big)^{-1}\Big], \end{aligned} \tag{18.14}$$

$$q(\mathbf{z}_{t-1}|\mathbf{z}_t, \mathbf{x}) = \mathrm{Norm}_{\mathbf{z}_{t-1}}\left[\frac{(1-\alpha_{t-1})}{1-\alpha_t}\sqrt{1-\beta_t}\mathbf{z}_t + \frac{\sqrt{\alpha_{t-1}}\beta_t}{1-\alpha_t}\mathbf{x}, \frac{\beta_t(1-\alpha_{t-1})}{1-\alpha_t}\mathbf{I}\right]. \tag{18.15}$$

$$\begin{aligned} Pr(\mathbf{z}_T) &= \mathrm{Norm}_{\mathbf{z}_T}[\mathbf{0}, \mathbf{I}] \\ Pr(\mathbf{z}_{t-1}|\mathbf{z}_t, \boldsymbol{\phi}_t) &= \mathrm{Norm}_{\mathbf{z}_{t-1}}\Big[\mathbf{f}_t[\mathbf{z}_t, \boldsymbol{\phi}_t], \sigma_t^2\mathbf{I}\Big] \\ Pr(\mathbf{x}|\mathbf{z}_1, \boldsymbol{\phi}_1) &= \mathrm{Norm}_{\mathbf{x}}\Big[\mathbf{f}_1[\mathbf{z}_1, \boldsymbol{\phi}_1], \sigma_1^2\mathbf{I}\Big], \end{aligned} \tag{18.16}$$

$$Pr(\mathbf{x}, \mathbf{z}_{1\ldots T}|\boldsymbol{\phi}_{1\ldots T}) = Pr(\mathbf{x}|\mathbf{z}_1, \boldsymbol{\phi}_1) \prod_{t=2}^{T} Pr(\mathbf{z}_{t-1}|\mathbf{z}_t, \boldsymbol{\phi}_t) \cdot Pr(\mathbf{z}_T). \tag{18.17}$$

$$Pr(\mathbf{x}|\boldsymbol{\phi}_{1\ldots T}) = \int Pr(\mathbf{x}, \mathbf{z}_{1\ldots T}|\boldsymbol{\phi}_{1\ldots T}) d\mathbf{z}_{1\ldots T}. \tag{18.18}$$

$$\hat{\boldsymbol{\phi}}_{1\ldots T} = \underset{\boldsymbol{\phi}_{1\ldots T}}{\operatorname{argmax}} \left[\sum_{i=1}^{I} \log\Bigl[Pr(\mathbf{x}_i|\boldsymbol{\phi}_{1\ldots T}) \Bigr] \right]. \tag{18.19}$$

$$\begin{aligned}
\log\left[Pr(\mathbf{x}|\boldsymbol{\phi}_{1\ldots T})\right] &= \log\left[\int Pr(\mathbf{x}, \mathbf{z}_{1\ldots T}|\boldsymbol{\phi}_{1\ldots T}) d\mathbf{z}_{1\ldots T}\right] \\
&= \log\left[\int q(\mathbf{z}_{1\ldots T}|\mathbf{x}) \frac{Pr(\mathbf{x}, \mathbf{z}_{1\ldots T}|\boldsymbol{\phi}_{1\ldots T})}{q(\mathbf{z}_{1\ldots T}|\mathbf{x})} d\mathbf{z}_{1\ldots T}\right] \\
&\geq \int q(\mathbf{z}_{1\ldots T}|\mathbf{x}) \log\left[\frac{Pr(\mathbf{x}, \mathbf{z}_{1\ldots T}|\boldsymbol{\phi}_{1\ldots T})}{q(\mathbf{z}_{1\ldots T}|\mathbf{x})}\right] d\mathbf{z}_{1\ldots T}.
\end{aligned} \tag{18.20}$$

$$\mathrm{ELBO}\bigl[\boldsymbol{\phi}_{1\ldots T}\bigr] = \int q(\mathbf{z}_{1\ldots T}|\mathbf{x}) \log\left[\frac{Pr(\mathbf{x}, \mathbf{z}_{1\ldots T}|\boldsymbol{\phi}_{1\ldots T})}{q(\mathbf{z}_{1\ldots T}|\mathbf{x})}\right] d\mathbf{z}_{1\ldots T}. \tag{18.21}$$

$$\begin{aligned}
\log\left[\frac{Pr(\mathbf{x}, \mathbf{z}_{1\ldots T}|\boldsymbol{\phi}_{1\ldots T})}{q(\mathbf{z}_{1\ldots T}|\mathbf{x})}\right] &= \log\left[\frac{Pr(\mathbf{x}|\mathbf{z}_1, \boldsymbol{\phi}_1)\prod_{t=2}^{T} Pr(\mathbf{z}_{t-1}|\mathbf{z}_t, \boldsymbol{\phi}_t) \cdot Pr(\mathbf{z}_T)}{q(\mathbf{z}_1|\mathbf{x})\prod_{t=2}^{T} q(\mathbf{z}_t|\mathbf{z}_{t-1})}\right] \\
&= \log\left[\frac{Pr(\mathbf{x}|\mathbf{z}_1, \boldsymbol{\phi}_1)}{q(\mathbf{z}_1|\mathbf{x})}\right] + \log\left[\frac{\prod_{t=2}^{T} Pr(\mathbf{z}_{t-1}|\mathbf{z}_t, \boldsymbol{\phi}_t)}{\prod_{t=2}^{T} q(\mathbf{z}_t|\mathbf{z}_{t-1})}\right] + \log\Bigl[Pr(\mathbf{z}_T)\Bigr].
\end{aligned} \tag{18.22}$$

$$q(\mathbf{z}_t|\mathbf{z}_{t-1}) = q(\mathbf{z}_t|\mathbf{z}_{t-1}, \mathbf{x}) = \frac{q(\mathbf{z}_{t-1}|\mathbf{z}_t, \mathbf{x}) q(\mathbf{z}_t|\mathbf{x})}{q(\mathbf{z}_{t-1}|\mathbf{x})}, \tag{18.23}$$

$$\begin{aligned}
&\log\left[\frac{Pr(\mathbf{x}, \mathbf{z}_{1\ldots T}|\boldsymbol{\phi}_{1\ldots T})}{q(\mathbf{z}_{1\ldots T}|\mathbf{x})}\right] \\
&\quad = \log\left[\frac{Pr(\mathbf{x}|\mathbf{z}_1, \boldsymbol{\phi}_1)}{q(\mathbf{z}_1|\mathbf{x})}\right] + \log\left[\frac{\prod_{t=2}^{T} Pr(\mathbf{z}_{t-1}|\mathbf{z}_t, \boldsymbol{\phi}_t) \cdot q(\mathbf{z}_{t-1}|\mathbf{x})}{\prod_{t=2}^{T} q(\mathbf{z}_{t-1}|\mathbf{z}_t, \mathbf{x}) \cdot q(\mathbf{z}_t|\mathbf{x})}\right] + \log\Bigl[Pr(\mathbf{z}_T)\Bigr] \\
&\quad = \log\left[Pr(\mathbf{x}|\mathbf{z}_1, \boldsymbol{\phi}_1)\right] + \log\left[\frac{\prod_{t=2}^{T} Pr(\mathbf{z}_{t-1}|\mathbf{z}_t, \boldsymbol{\phi}_t)}{\prod_{t=2}^{T} q(\mathbf{z}_{t-1}|\mathbf{z}_t, \mathbf{x})}\right] + \log\left[\frac{Pr(\mathbf{z}_T)}{q(\mathbf{z}_T|\mathbf{x})}\right] \\
&\quad \approx \log\left[Pr(\mathbf{x}|\mathbf{z}_1, \boldsymbol{\phi}_1)\right] + \sum_{t=2}^{T} \log\left[\frac{Pr(\mathbf{z}_{t-1}|\mathbf{z}_t, \boldsymbol{\phi}_t)}{q(\mathbf{z}_{t-1}|\mathbf{z}_t, \mathbf{x})}\right],
\end{aligned} \tag{18.24}$$

$$
\begin{aligned}
&\mathrm{ELBO}\big[\boldsymbol{\phi}_{1\ldots T}\big] \qquad (18.25)\\
&\quad= \int q(\mathbf{z}_{1\ldots T}|\mathbf{x}) \log\left[\frac{Pr(\mathbf{x},\mathbf{z}_{1\ldots T}|\boldsymbol{\phi}_{1\ldots T})}{q(\mathbf{z}_{1\ldots T}|\mathbf{x})}\right] d\mathbf{z}_{1\ldots T}\\
&\quad\approx \int q(\mathbf{z}_{1\ldots T}|\mathbf{x}) \left(\log\left[Pr(\mathbf{x}|\mathbf{z}_1,\boldsymbol{\phi}_1)\right] + \sum_{t=2}^{T}\log\left[\frac{Pr(\mathbf{z}_{t-1}|\mathbf{z}_t,\boldsymbol{\phi}_t)}{q(\mathbf{z}_{t-1}|\mathbf{z}_t,\mathbf{x})}\right]\right) d\mathbf{z}_{1\ldots T}\\
&\quad= \mathbb{E}_{q(\mathbf{z}_1|\mathbf{x})}\Big[\log\left[Pr(\mathbf{x}|\mathbf{z}_1,\boldsymbol{\phi}_1)\right]\Big] - \sum_{t=2}^{T}\mathbb{E}_{q(\mathbf{z}_t|\mathbf{x})}\Big[\mathrm{D}_{KL}\Big[q(\mathbf{z}_{t-1}|\mathbf{z}_t,\mathbf{x})\Big|\Big|Pr(\mathbf{z}_{t-1}|\mathbf{z}_t,\boldsymbol{\phi}_t)\Big]\Big],
\end{aligned}
$$

$$
Pr(\mathbf{x}|\mathbf{z}_1,\boldsymbol{\phi}_1) = \mathrm{Norm}_{\mathbf{x}}\Big[\mathbf{f}_1[\mathbf{z}_1,\boldsymbol{\phi}_1],\sigma_1^2\mathbf{I}\Big], \qquad (18.26)
$$

$$
\begin{aligned}
Pr(\mathbf{z}_{t-1}|\mathbf{z}_t,\boldsymbol{\phi}_t) &= \mathrm{Norm}_{\mathbf{z}_{t-1}}\Big[\mathbf{f}_t[\mathbf{z}_t,\boldsymbol{\phi}_t],\sigma_t^2\mathbf{I}\Big] \qquad (18.27)\\
q(\mathbf{z}_{t-1}|\mathbf{z}_t,\mathbf{x}) &= \mathrm{Norm}_{\mathbf{z}_{t-1}}\left[\frac{(1-\alpha_{t-1})}{1-\alpha_t}\sqrt{1-\beta_t}\,\mathbf{z}_t + \frac{\sqrt{\alpha_{t-1}}\beta_t}{1-\alpha_t}\mathbf{x}, \frac{\beta_t(1-\alpha_{t-1})}{1-\alpha_t}\mathbf{I}\right].
\end{aligned}
$$

$$
\begin{aligned}
D_{KL}\Big[q(\mathbf{z}_{t-1}|\mathbf{z}_t,\mathbf{x})\Big|\Big|Pr(\mathbf{z}_{t-1}|\mathbf{z}_t,\boldsymbol{\phi}_t)\Big] &= \qquad (18.28)\\
&\frac{1}{2\sigma_t^2}\left\|\frac{(1-\alpha_{t-1})}{1-\alpha_t}\sqrt{1-\beta_t}\,\mathbf{z}_t + \frac{\sqrt{\alpha_{t-1}}\beta_t}{1-\alpha_t}\mathbf{x} - \mathbf{f}_t[\mathbf{z}_t,\boldsymbol{\phi}_t]\right\|^2 + C.
\end{aligned}
$$

$$
\begin{aligned}
L[\boldsymbol{\phi}_{1\ldots T}] &= \sum_{i=1}^{I}\Bigg(\overbrace{-\log\Big[\mathrm{Norm}_{\mathbf{x}_i}\big[\mathbf{f}_1[\mathbf{z}_{i1},\boldsymbol{\phi}_1],\sigma_1^2\mathbf{I}\big]\Big]}^{\text{reconstruction term}} \qquad (18.29)\\
&+\sum_{t=2}^{T}\frac{1}{2\sigma_t^2}\Bigg\|\underbrace{\frac{1-\alpha_{t-1}}{1-\alpha_t}\sqrt{1-\beta_t}\,\mathbf{z}_{it} + \frac{\sqrt{\alpha_{t-1}}\beta_t}{1-\alpha_t}\mathbf{x}_i}_{\text{target, mean of } q(\mathbf{z}_{t-1}|\mathbf{z}_t,\mathbf{x})} - \underbrace{\mathbf{f}_t[\mathbf{z}_{it},\boldsymbol{\phi}_t]}_{\text{predicted } \mathbf{z}_{t-1}}\Bigg\|^2\Bigg),
\end{aligned}
$$

$$
\mathbf{z}_t = \sqrt{\alpha_t}\cdot\mathbf{x} + \sqrt{1-\alpha_t}\cdot\boldsymbol{\epsilon}. \qquad (18.30)
$$

$$
\mathbf{x} = \frac{1}{\sqrt{\alpha_t}}\cdot\mathbf{z}_t - \frac{\sqrt{1-\alpha_t}}{\sqrt{\alpha_t}}\cdot\boldsymbol{\epsilon}. \qquad (18.31)
$$

$$
\begin{aligned}
&\frac{(1-\alpha_{t-1})}{1-\alpha_t}\sqrt{1-\beta_t}\mathbf{z}_t + \frac{\sqrt{\alpha_{t-1}}\beta_t}{1-\alpha_t}\mathbf{x} \\
&\qquad = \frac{(1-\alpha_{t-1})}{1-\alpha_t}\sqrt{1-\beta_t}\mathbf{z}_t + \frac{\sqrt{\alpha_{t-1}}\beta_t}{1-\alpha_t}\left(\frac{1}{\sqrt{\alpha_t}}\mathbf{z}_t - \frac{\sqrt{1-\alpha_t}}{\sqrt{\alpha_t}}\boldsymbol{\epsilon}\right) \\
&\qquad = \frac{(1-\alpha_{t-1})}{1-\alpha_t}\sqrt{1-\beta_t}\mathbf{z}_t + \frac{\beta_t}{1-\alpha_t}\left(\frac{1}{\sqrt{1-\beta_t}}\mathbf{z}_t - \frac{\sqrt{1-\alpha_t}}{\sqrt{1-\beta_t}}\boldsymbol{\epsilon}\right),
\end{aligned}
\tag{18.32}
$$

$$
\begin{aligned}
&\frac{(1-\alpha_{t-1})}{1-\alpha_t}\sqrt{1-\beta_t}\mathbf{z}_t + \frac{\sqrt{\alpha_{t-1}}\beta_t}{1-\alpha_t}\mathbf{x} \\
&\qquad = \left(\frac{(1-\alpha_{t-1})\sqrt{1-\beta_t}}{1-\alpha_t} + \frac{\beta_t}{(1-\alpha_t)\sqrt{1-\beta_t}}\right)\mathbf{z}_t - \frac{\beta_t}{\sqrt{1-\alpha_t}\sqrt{1-\beta_t}}\boldsymbol{\epsilon} \\
&\qquad = \left(\frac{(1-\alpha_{t-1})(1-\beta_t)}{(1-\alpha_t)\sqrt{1-\beta_t}} + \frac{\beta_t}{(1-\alpha_t)\sqrt{1-\beta_t}}\right)\mathbf{z}_t - \frac{\beta_t}{\sqrt{1-\alpha_t}\sqrt{1-\beta_t}}\boldsymbol{\epsilon} \\
&\qquad = \frac{(1-\alpha_{t-1})(1-\beta_t)+\beta_t}{(1-\alpha_t)\sqrt{1-\beta_t}}\mathbf{z}_t - \frac{\beta_t}{\sqrt{1-\alpha_t}\sqrt{1-\beta_t}}\boldsymbol{\epsilon} \\
&\qquad = \frac{1-\alpha_t}{(1-\alpha_t)\sqrt{1-\beta_t}}\mathbf{z}_t - \frac{\beta_t}{\sqrt{1-\alpha_t}\sqrt{1-\beta_t}}\boldsymbol{\epsilon} \\
&\qquad = \frac{1}{\sqrt{1-\beta_t}}\mathbf{z}_t - \frac{\beta_t}{\sqrt{1-\alpha_t}\sqrt{1-\beta_t}}\boldsymbol{\epsilon},
\end{aligned}
\tag{18.33}
$$

$$
\begin{aligned}
L[\boldsymbol{\phi}_{1\ldots T}] \quad = \quad & \sum_{i=1}^{I}\Biggl(-\log\Bigl[\mathrm{Norm}_{\mathbf{x}_i}\left[\mathbf{f}_1[\mathbf{z}_{i1},\boldsymbol{\phi}_1],\sigma_1^2\mathbf{I}\right]\Bigr] \\
& +\sum_{t=2}^{T}\frac{1}{2\sigma_t^2}\left\|\left(\frac{1}{\sqrt{1-\beta_t}}\mathbf{z}_{it} - \frac{\beta_t}{\sqrt{1-\alpha_t}\sqrt{1-\beta_t}}\boldsymbol{\epsilon}_{it}\right) - \mathbf{f}_t[\mathbf{z}_{it},\boldsymbol{\phi}_t]\right\|^2\Biggr).
\end{aligned}
\tag{18.34}
$$

$$
\mathbf{f}_t[\mathbf{z}_t,\boldsymbol{\phi}_t] = \frac{1}{\sqrt{1-\beta_t}}\mathbf{z}_t - \frac{\beta_t}{\sqrt{1-\alpha_t}\sqrt{1-\beta_t}}\mathbf{g}_t[\mathbf{z}_t,\boldsymbol{\phi}_t].
\tag{18.35}
$$

$$
\begin{aligned}
&L[\boldsymbol{\phi}_{1\ldots T}] = \\
&\quad \sum_{i=1}^{I} -\log\Bigl[\mathrm{Norm}_{\mathbf{x}_i}\left[\mathbf{f}_1[\mathbf{z}_{i1},\boldsymbol{\phi}_1],\sigma_1^2\mathbf{I}\right]\Bigr] + \sum_{t=2}^{T}\frac{\beta_t^2}{(1-\alpha_t)(1-\beta_t)2\sigma_t^2}\Bigl\|\mathbf{g}_t[\mathbf{z}_{it},\boldsymbol{\phi}_t] - \boldsymbol{\epsilon}_{it}\Bigr\|^2.
\end{aligned}
\tag{18.36}
$$

$$
L[\boldsymbol{\phi}_{1\ldots T}] = \sum_{i=1}^{I}\frac{1}{2\sigma_1^2}\Bigl\|\mathbf{x}_i - \mathbf{f}_1[\mathbf{z}_{i1},\boldsymbol{\phi}_1]\Bigr\|^2 + \sum_{t=2}^{T}\frac{\beta_t^2}{(1-\alpha_t)(1-\beta_t)2\sigma_t^2}\Bigl\|\mathbf{g}_t[\mathbf{z}_{it},\boldsymbol{\phi}_t] - \boldsymbol{\epsilon}_{it}\Bigr\|^2 + C_t
\tag{18.37}
$$

$$\frac{1}{2\sigma_1^2}\left\|\mathbf{x}_i-\mathbf{f}_1[\mathbf{z}_{i1},\boldsymbol{\phi}_1]\right\|^2=\frac{1}{2\sigma_1^2}\left\|\frac{\beta_1}{\sqrt{1-\alpha_1}\sqrt{1-\beta_1}}\mathbf{g}_1[\mathbf{z}_{i1},\boldsymbol{\phi}_1]-\frac{\beta_1}{\sqrt{1-\alpha_1}\sqrt{1-\beta_1}}\boldsymbol{\epsilon}_{i1}\right\|^2 \tag{18.38}$$

$$L[\boldsymbol{\phi}_{1\ldots T}]=\sum_{i=1}^{I}\sum_{t=1}^{T}\frac{\beta_t^2}{(1-\alpha_t)(1-\beta_t)2\sigma_t^2}\left\|\mathbf{g}_t[\mathbf{z}_{it},\boldsymbol{\phi}_t]-\boldsymbol{\epsilon}_{it}\right\|^2, \tag{18.39}$$

$$\begin{aligned} L[\boldsymbol{\phi}_{1\ldots T}] &= \sum_{i=1}^{I}\sum_{t=1}^{T}\left\|\mathbf{g}_t[\mathbf{z}_{it},\boldsymbol{\phi}_t]-\boldsymbol{\epsilon}_{it}\right\|^2 \\ &= \sum_{i=1}^{I}\sum_{t=1}^{T}\left\|\mathbf{g}_t\Big[\sqrt{\alpha_t}\cdot\mathbf{x}_i+\sqrt{1-\alpha_t}\cdot\boldsymbol{\epsilon}_{it},\boldsymbol{\phi}_t\Big]-\boldsymbol{\epsilon}_{it}\right\|^2, \end{aligned} \tag{18.40}$$

$$\mathbf{z}_{t-1}=\hat{\mathbf{z}}_{t-1}+\sigma_t^2\frac{\partial\log\big[\Pr(c|\mathbf{z}_t)\big]}{\partial\mathbf{z}_t}+\sigma_t\boldsymbol{\epsilon}. \tag{18.41}$$

$$\mathbf{x}_t=\sqrt{1-\beta_t}\cdot\mathbf{x}_{t-1}+\sqrt{\beta_t}\cdot\boldsymbol{\epsilon}_t, \tag{18.42}$$

$$z=a\cdot\epsilon_1+b\cdot\epsilon_2, \tag{18.43}$$

$$\begin{aligned} \mathbb{E}[z] &= 0 \\ \mathrm{Var}[z] &= a^2+b^2, \end{aligned} \tag{18.44}$$

$$\mathbf{z}_3 = \sqrt{(1-\beta_3)(1-\beta_2)(1-\beta_1)}\cdot\mathbf{x}+\sqrt{1-(1-\beta_3)(1-\beta_2)(1-\beta_1)}\cdot\boldsymbol{\epsilon}', \tag{18.45}$$

$$\mathrm{Norm}_{\mathbf{v}}\left[\mathbf{A}\mathbf{w},\mathbf{B}\right]\propto\mathrm{Norm}_{\mathbf{w}}\Big[(\mathbf{A}^T\mathbf{B}^{-1}\mathbf{A})^{-1}\mathbf{A}^T\mathbf{B}^{-1}\mathbf{v},(\mathbf{A}^T\mathbf{B}^{-1}\mathbf{A})^{-1}\Big]. \tag{18.46}$$

$$\mathrm{Norm}_{\mathbf{x}}[\mathbf{a},\mathbf{A}]\mathrm{Norm}_{\mathbf{x}}[\mathbf{b},\mathbf{B}] \propto \mathrm{Norm}_{\mathbf{x}}\Big[(\mathbf{A}^{-1}+\mathbf{B}^{-1})^{-1}(\mathbf{A}^{-1}\mathbf{a}+\mathbf{B}^{-1}\mathbf{b}),(\mathbf{A}^{-1}+\mathbf{B}^{-1})^{-1}\Big]. \tag{18.47}$$

$$D_{KL}\Big[\text{Norm}_{\mathbf{w}}[\mathbf{a},\mathbf{A}]\Big|\Big|\text{Norm}_{\mathbf{w}}[\mathbf{b},\mathbf{B}]\Big] \;=\; \frac{1}{2}\left(\text{tr}\left[\mathbf{B}^{-1}\mathbf{A}\right]-d+(\mathbf{a}-\mathbf{b})^T\mathbf{B}^{-1}(\mathbf{a}-\mathbf{b})+\log\left[\frac{|\mathbf{B}|}{|\mathbf{A}|}\right]\right). \tag{18.48}$$

$$\sqrt{\frac{\alpha_t}{\alpha_{t-1}}}=\sqrt{1-\beta_t}. \tag{18.49}$$

$$\frac{(1-\alpha_{t-1})(1-\beta_t)+\beta_t}{(1-\alpha_t)\sqrt{1-\beta_t}}=\frac{1}{\sqrt{1-\beta_t}}. \tag{18.50}$$

Chapter 19

Reinforcement learning

$$G_t = \sum_{k=0}^{\infty} \gamma^k r_{t+k+1}. \tag{19.1}$$

$$v[s_t|\pi] = \mathbb{E}\Big[G_t|s_t, \pi\Big]. \tag{19.2}$$

$$q[s_t, a_t|\pi] = \mathbb{E}\Big[G_t|s_t, a_t, \pi\Big]. \tag{19.3}$$

$$v^*[s_t] = \max_{\pi}\Big[\mathbb{E}\Big[G_t|s_t, \pi\Big]\Big]. \tag{19.4}$$

$$q^*[s_t, a_t] = \max_{\pi}\Big[\mathbb{E}\Big[G_t|s_t, a_t, \pi\Big]\Big]. \tag{19.5}$$

$$\pi[a_t|s_t] \leftarrow \underset{a_t}{\operatorname{argmax}}\Big[q^*[s_t, a_t]\Big]. \tag{19.6}$$

$$v[s_t] = \sum_{a_t} \pi[a_t|s_t] q[s_t, a_t]. \tag{19.7}$$

$$q[s_t, a_t] = r[s_t, a_t] + \gamma \cdot \sum_{s_{t+1}} Pr(s_{t+1}|s_t, a_t) v[s_{t+1}]. \tag{19.8}$$

$$v[s_t] = \sum_{a_t} \pi[a_t|s_t] \left(r[s_t, a_t] + \gamma \cdot \sum_{s_{t+1}} Pr(s_{t+1}|s_t, a_t) v[s_{t+1}] \right). \tag{19.9}$$

$$q[s_t, a_t] = r[s_t, a_t] + \gamma \cdot \sum_{s_{t+1}} Pr(s_{t+1}|s_t, a_t) \left(\sum_{a_{t+1}} \pi[a_{t+1}|s_{t+1}] q[s_{t+1}, a_{t+1}] \right). \tag{19.10}$$

$$v[s_t] \leftarrow \sum_{a_t} \pi[a_t|s_t] \left(r[s_t, a_t] + \gamma \cdot \sum_{s_{t+1}} Pr(s_{t+1}|s_t, a_t) v[s_{t+1}] \right), \tag{19.11}$$

$$\pi[a_t|s_t] \leftarrow \underset{a_t}{\operatorname{argmax}} \left[r[s_t, a_t] + \gamma \cdot \sum_{s_{t+1}} Pr(s_{t+1}|s_t, a_t) v[s_{t+1}] \right]. \tag{19.12}$$

$$\pi[a|s] \leftarrow \underset{a}{\operatorname{argmax}} \Big[q[s, a] \Big]. \tag{19.13}$$

$$q[s_t, a_t] \leftarrow q[s_t, a_t] + \alpha \Big(r[s_t, a_t] + \gamma \cdot q[s_{t+1}, a_{t+1}] - q[s_t, a_t] \Big), \tag{19.14}$$

$$q[s_t, a_t] \leftarrow q[s_t, a_t] + \alpha \Big(r[s_t, a_t] + \gamma \cdot \max_a \big[q[s_{t+1}, a] \big] - q[s_t, a_t] \Big), \tag{19.15}$$

$$L[\boldsymbol{\phi}] = \left(r[\mathbf{s}_t, a_t] + \gamma \cdot \max_a \Big[q[\mathbf{s}_{t+1}, a, \boldsymbol{\phi}] \Big] - q[\mathbf{s}_t, a_t, \boldsymbol{\phi}] \right)^2, \tag{19.16}$$

$$\boldsymbol{\phi} \leftarrow \boldsymbol{\phi} + \alpha \left(r[\mathbf{s}_t, a_t] + \gamma \cdot \max_a \Big[q[\mathbf{s}_{t+1}, a, \boldsymbol{\phi}] \Big] - q[\mathbf{s}_t, a_t, \boldsymbol{\phi}] \right) \frac{\partial q[\mathbf{s}_t, a_t, \boldsymbol{\phi}]}{\partial \boldsymbol{\phi}}. \tag{19.17}$$

$$\boldsymbol{\phi} \leftarrow \boldsymbol{\phi} + \alpha \left(r[\mathbf{s}_t, a_t] + \gamma \cdot \max_a \Big[q[\mathbf{s}_{t+1}, a, \boldsymbol{\phi}^-] \Big] - q[\mathbf{s}_t, a_t, \boldsymbol{\phi}] \right) \frac{\partial q[\mathbf{s}_t, a_t, \boldsymbol{\phi}]}{\partial \boldsymbol{\phi}}. \tag{19.18}$$

$$q[s_t, a_t] \leftarrow q[s_t, a_t] + \alpha \Big(r[s_t, a_t] + \gamma \cdot \max_a \big[q[s_{t+1}, a] \big] - q[s_t, a_t] \Big) \tag{19.19}$$

$$\begin{aligned} q_1[s_t, a_t] \quad &\leftarrow \quad q_1[s_t, a_t] + \alpha \Big(r[s_t, a_t] + \gamma \cdot q_2 \left[s_{t+1}, \underset{a}{\operatorname{argmax}} \Big[q_1[s_{t+1}, a] \Big] \right] - q_1[s_t, a_t] \Big) \\ q_2[s_t, a_t] \quad &\leftarrow \quad q_2[s_t, a_t] + \alpha \Big(r[s_t, a_t] + \gamma \cdot q_1 \left[s_{t+1}, \underset{a}{\operatorname{argmax}} \Big[q_2[s_{t+1}, a] \Big] \right] - q_2[s_t, a_t] \Big). \end{aligned} \tag{19.20}$$

$$\begin{aligned}\boldsymbol{\phi}_1 \leftarrow \boldsymbol{\phi}_1+\alpha\biggl(r[\mathbf{s}_t,a_t]+\gamma\cdot q\biggl[\mathbf{s}_{t+1},\underset{a}{\mathrm{argmax}}\Bigl[q[\mathbf{s}_{t+1},a,\boldsymbol{\phi}_1]\Bigr],\boldsymbol{\phi}_2\biggr]-q[\mathbf{s}_t,a_t,\boldsymbol{\phi}_1]\biggr)\frac{\partial q[\mathbf{s}_t,a_t,\boldsymbol{\phi}_1]}{\partial \boldsymbol{\phi}_1}\\ \boldsymbol{\phi}_2 \leftarrow \boldsymbol{\phi}_2+\alpha\biggl(r[\mathbf{s}_t,a_t]+\gamma\cdot q\biggl[\mathbf{s}_{t+1},\underset{a}{\mathrm{argmax}}\Bigl[q[\mathbf{s}_{t+1},a,\boldsymbol{\phi}_2]\Bigr],\boldsymbol{\phi}_1\biggr]-q[\mathbf{s}_t,a_t,\boldsymbol{\phi}_2]\biggr)\frac{\partial q[\mathbf{s}_t,a_t,\boldsymbol{\phi}_2]}{\partial \boldsymbol{\phi}_2}.\end{aligned} \tag{19.21}$$

$$Pr(\boldsymbol{\tau}|\boldsymbol{\theta}) = Pr(\mathbf{s}_1)\prod_{t=1}^{T}\pi[a_t|\mathbf{s}_t,\boldsymbol{\theta}]Pr(\mathbf{s}_{t+1}|\mathbf{s}_t,a_t). \tag{19.22}$$

$$\boldsymbol{\theta} = \underset{\boldsymbol{\theta}}{\mathrm{argmax}}\Bigl[\mathbb{E}_{\boldsymbol{\tau}}\Bigl[r[\boldsymbol{\tau}]\Bigr]\Bigr] = \underset{\boldsymbol{\theta}}{\mathrm{argmax}}\Bigl[\int Pr(\boldsymbol{\tau}|\boldsymbol{\theta})r[\boldsymbol{\tau}]d\boldsymbol{\tau}\Bigr], \tag{19.23}$$

$$\begin{aligned}\boldsymbol{\theta} &\leftarrow \boldsymbol{\theta}+\alpha\cdot\frac{\partial}{\partial\boldsymbol{\theta}}\int Pr(\boldsymbol{\tau}|\boldsymbol{\theta})r[\boldsymbol{\tau}]d\boldsymbol{\tau}\\ &= \boldsymbol{\theta}+\alpha\cdot\int\frac{\partial Pr(\boldsymbol{\tau}|\boldsymbol{\theta})}{\partial\boldsymbol{\theta}}r[\boldsymbol{\tau}]d\boldsymbol{\tau}.\end{aligned} \tag{19.24}$$

$$\begin{aligned}\boldsymbol{\theta} &\leftarrow \boldsymbol{\theta}+\alpha\cdot\int\frac{\partial Pr(\boldsymbol{\tau}|\boldsymbol{\theta})}{\partial\boldsymbol{\theta}}r[\boldsymbol{\tau}]d\boldsymbol{\tau}\\ &= \boldsymbol{\theta}+\alpha\cdot\int Pr(\boldsymbol{\tau}|\boldsymbol{\theta})\frac{1}{Pr(\boldsymbol{\tau}|\boldsymbol{\theta})}\frac{\partial Pr(\boldsymbol{\tau}|\boldsymbol{\theta})}{\partial\boldsymbol{\theta}}r[\boldsymbol{\tau}]d\boldsymbol{\tau}\\ &\approx \boldsymbol{\theta}+\alpha\cdot\frac{1}{I}\sum_{i=1}^{I}\frac{1}{Pr(\boldsymbol{\tau}_i|\boldsymbol{\theta})}\frac{\partial Pr(\boldsymbol{\tau}_i|\boldsymbol{\theta})}{\partial\boldsymbol{\theta}}r[\boldsymbol{\tau}_i].\end{aligned} \tag{19.25}$$

$$\frac{\partial\log[\mathrm{f}[z]]}{\partial z} = \frac{1}{f[z]}\frac{\partial\mathrm{f}[z]}{\partial z}, \tag{19.26}$$

$$\boldsymbol{\theta}\leftarrow\boldsymbol{\theta}+\alpha\cdot\frac{1}{I}\sum_{i=1}^{I}\frac{\partial\log\bigl[Pr(\boldsymbol{\tau}_i|\boldsymbol{\theta})\bigr]}{\partial\boldsymbol{\theta}}r[\boldsymbol{\tau}_i]. \tag{19.27}$$

$$\begin{aligned}\log[Pr(\boldsymbol{\tau}|\boldsymbol{\theta})] &= \log\Bigl[Pr(\mathbf{s}_1)\prod_{t=1}^{T}\pi[a_t|\mathbf{s}_t,\boldsymbol{\theta}]Pr(\mathbf{s}_{t+1}|\mathbf{s}_t,a_t)\Bigr] \qquad (19.28)\\ &= \log\bigl[Pr(\mathbf{s}_1)\bigr]+\sum_{t=1}^{T}\log\bigl[\pi[a_t|\mathbf{s}_t,\boldsymbol{\theta}]\bigr]+\sum_{t=1}^{T}\log\bigl[Pr(\mathbf{s}_{t+1}|\mathbf{s}_t,a_t)\bigr],\end{aligned}$$

$$\boldsymbol{\theta} \;\leftarrow\; \boldsymbol{\theta} + \alpha \cdot \frac{1}{I}\sum_{i=1}^{I}\sum_{t=1}^{T}\frac{\partial \log\bigl[\pi[a_{it}|\mathbf{s}_{it},\boldsymbol{\theta}]\bigr]}{\partial \boldsymbol{\theta}} r[\boldsymbol{\tau}_i], \tag{19.29}$$

$$r[\boldsymbol{\tau}_i] = \sum_{t=1}^{T} r_{i,t+1} = \sum_{k=1}^{t} r_{i,k+1} + \sum_{k=t}^{T} r_{i,k+1}, \tag{19.30}$$

$$\boldsymbol{\theta} \;\leftarrow\; \boldsymbol{\theta} + \alpha \cdot \frac{1}{I}\sum_{i=1}^{I}\sum_{t=1}^{T}\frac{\partial \log\bigl[\pi[a_{it}|\mathbf{s}_{it},\boldsymbol{\theta}]\bigr]}{\partial \boldsymbol{\theta}} \sum_{k=t}^{T} r_{i,k+1}. \tag{19.31}$$

$$r[\boldsymbol{\tau}_{it}] = \sum_{k=t+1}^{T} \gamma^{k-t-1} r_{i,k+1}, \tag{19.32}$$

$$\boldsymbol{\theta} \leftarrow \boldsymbol{\theta} + \alpha \cdot \gamma^t \frac{\partial \log\bigl[\pi_{a_{it}}[\mathbf{s}_{it},\boldsymbol{\theta}]\bigr]}{\partial \boldsymbol{\theta}} r[\boldsymbol{\tau}_{it}] \qquad\qquad \forall\; i,t, \tag{19.33}$$

$$\boldsymbol{\theta} \leftarrow \boldsymbol{\theta} + \alpha \cdot \frac{1}{I}\sum_{i=1}^{I}\sum_{t=1}^{T}\frac{\partial \log\bigl[\pi_{a_{it}}[\mathbf{s}_{it},\boldsymbol{\theta}]\bigr]}{\partial \boldsymbol{\theta}} \left(r[\boldsymbol{\tau}_{it}] - b\right). \tag{19.34}$$

$$\mathbb{E}_{\boldsymbol{\tau}}\left[\sum_{t=1}^{T}\frac{\partial \log\bigl[\pi_{a_{it}}[\mathbf{s}_{it},\boldsymbol{\theta}]\bigr]}{\partial \boldsymbol{\theta}} \cdot b\right] = 0, \tag{19.35}$$

$$b = \sum_{i} \frac{\sum_{t=1}^{T}\left(\partial \log\bigl[\pi_{a_{it}}[\mathbf{s}_{it},\boldsymbol{\theta}]\bigr]/\partial \boldsymbol{\theta}\right)^2 r[\boldsymbol{\tau}_{it}]}{\sum_{t=1}^{T}\left(\partial \log\bigl[\pi_{a_{it}}[\mathbf{s}_{it},\boldsymbol{\theta}]\bigr]/\partial \boldsymbol{\theta}\right)^2}. \tag{19.36}$$

$$b = \frac{1}{I}\sum_{i} r[\boldsymbol{\tau}_i]. \tag{19.37}$$

$$\boldsymbol{\theta} \leftarrow \boldsymbol{\theta} + \alpha \cdot \frac{1}{I}\sum_{i=1}^{I}\sum_{t=1}^{T}\frac{\partial \log\bigl[\pi_{a_{it}}[\mathbf{s}_{it},\boldsymbol{\theta}]\bigr]}{\partial \boldsymbol{\theta}} \left(r[\boldsymbol{\tau}_{it}] - b[\mathbf{s}_{it}]\right). \tag{19.38}$$

$$L[\boldsymbol{\phi}] = \sum_{i=1}^{I}\sum_{t=1}^{T}\left(v[\mathbf{s}_{it},\boldsymbol{\phi}] - \sum_{j=t}^{T} r_{i,j+1}\right)^2. \tag{19.39}$$

$$r[\boldsymbol{\tau}_{it}] \approx r_{i,t+1} + \gamma\cdot v[\mathbf{s}_{i,t+1},\boldsymbol{\phi}]. \tag{19.40}$$

$$\boldsymbol{\theta} \leftarrow \boldsymbol{\theta} + \alpha\cdot\frac{1}{I}\sum_{i=1}^{I}\sum_{t=1}^{T}\frac{\partial \log\bigl[Pr(a_{it}|\mathbf{s}_{it},\boldsymbol{\theta})\bigr]}{\partial\boldsymbol{\theta}}\Bigl(r_{i,t+1} + \gamma\cdot v[\mathbf{s}_{i,t+1},\boldsymbol{\phi}] - v[\mathbf{s}_{i,t},\boldsymbol{\phi}]\Bigr). \tag{19.41}$$

$$L[\boldsymbol{\phi}] = \sum_{i=1}^{I}\sum_{t=1}^{T}\left(r_{i,t+1} + \gamma\cdot v[\mathbf{s}_{i,t+1},\boldsymbol{\phi}] - v[\mathbf{s}_{i,t},\boldsymbol{\phi}]\right)^2. \tag{19.42}$$

$$\pi'[a_t|s_t] \leftarrow \underset{a_t}{\operatorname{argmax}}\Biggl[r[s_t,a_t] + \gamma\cdot\sum_{s_{t+1}} Pr(s_{t+1}|s_t,a_t)v[s_{t+1}|\pi]\Biggr]. \tag{19.43}$$

$$\begin{aligned} v[s_t|\pi] &\leq q\Bigl[s_t,\pi'[a_t|s_t]\Bigm|\pi\Bigr] \\ &= \mathbb{E}_{\pi'}\Bigl[r_{t+1} + \gamma\cdot v[s_{t+1}|\pi]\Bigr]. \end{aligned} \tag{19.44}$$

$$\pi[a|s] = \frac{\exp\bigl[q[s,a]/\tau\bigr]}{\sum_{a'}\exp\bigl[q[s,a']/\tau\bigr]}. \tag{19.45}$$

$$\mathrm{f}\bigl[q[s,a]\bigr] = r[s,a] + \gamma\cdot\max_{a}\bigl[q[s',a]\bigr]. \tag{19.46}$$

$$\Bigl|\Bigl|\mathrm{f}\bigl[q_1[s,a]\bigr] - \mathrm{f}\bigl[q_2[s,a]\bigr]\Bigr|\Bigr|_\infty < \Bigl|\Bigl|q_1[s,a] - q_2[s,a]\Bigr|\Bigr|_\infty \qquad \forall\ q_1,q_2. \tag{19.47}$$

$$\mathbb{E}_{\boldsymbol{\tau}}\left[\frac{\partial}{\partial\boldsymbol{\theta}}\log\bigl[Pr(\boldsymbol{\tau}|\boldsymbol{\theta})\bigr]b\right] = 0, \tag{19.48}$$

$$a' = a - c(b-\mu_b). \tag{19.49}$$

$$\mathbb{E}_{\boldsymbol{\tau}}\Big[g[\boldsymbol{\theta}](r[\boldsymbol{\tau}_t]-b)\Big], \tag{19.50}$$

$$g[\theta] = \sum_{t=1}^{T} \frac{\partial \log\big[Pr(a_t|\mathbf{s}_t, \boldsymbol{\theta})]\big]}{\partial \boldsymbol{\theta}}, \tag{19.51}$$

$$r[\boldsymbol{\tau}_t] = \sum_{k=t}^{T} r_k. \tag{19.52}$$

$$b = \frac{\mathbb{E}[g[\boldsymbol{\tau}]^2]r[\boldsymbol{\tau}]}{\mathbb{E}[g[\boldsymbol{\tau}]^2]}. \tag{19.53}$$

Chapter 20

Why does deep learning work?

Chapter 21

Ethics