Mathematical Foundations of Deep Neural Networks, M1407.001200
E. Ryu
Spring 2024

Homework 9
Due 5pm, Wednesday, May 20, 2024

Problem 1: *Anomaly detection via AE.* In this problem, you will use an autoencoder to perform anomaly detection between the MNIST and the Kuzushiji(崩し字)-MNIST (KMNIST) [1] datasets. KMNIST contains handwritten Japanese characters. Download the starter code `anomaly_detection.py` and implement the following steps. In step 1, load the MNIST and KMNIST datasets, and split the MNIST training dataset into "training" and "validation" sets. (Together with the "test" set you will have three datasets in total.) In step 2, define the AE model. In step 3, instantiate the model and select the Adam optimizer. In step 4, train the AE with the training data $X_1, \dots, X_N$ with loss

$$\ell(\theta, \varphi) = \sum_{i=1}^{N} \|X_i - D_\varphi(E_\theta(X_i))\|^2,$$

where E_θ is the encoder and D_φ is the decoder. Do not use the validation set in this stage. In step 5, define the score function

$$s(X) = \|X - D_\varphi(E_\theta(X))\|^2$$

and calculate the mean and standard deviation of

$$\{s(Y_i)\}_{i=1}^{M}$$

where $Y_1, \dots, Y_M$ are the validation data. Define a threshold to be mean + 3 standard deviations, and define inputs with score function value exceeding this threshold to be anomalies. In step 6, check how many of the MNIST images within the test set are classified as anomalies and report the type I error rate. In step 7, check how many of the KMNIST images are classified as non-anomalies and report the type II error rate.

Figure 1: KMNIST images

Solution. See `anomaly_detection_sol.py`. ■

References

[1] T. Clanuwat, M. Bober-Irizar, A. Kitamoto, A. Lamb, K. Yamamoto, and D. Ha, Deep learning for classical Japanese literature, *NeurIPS ML for Creativity Workshop*, 2018.

Problem 2: *1D flow to Gaussian.* Consider the flow

$$f_\theta(x) = \sum_{i=1}^{n} e^{w_i}(\Phi_{\mu_i,\exp(\tau_i)}(x) - 0.5),$$

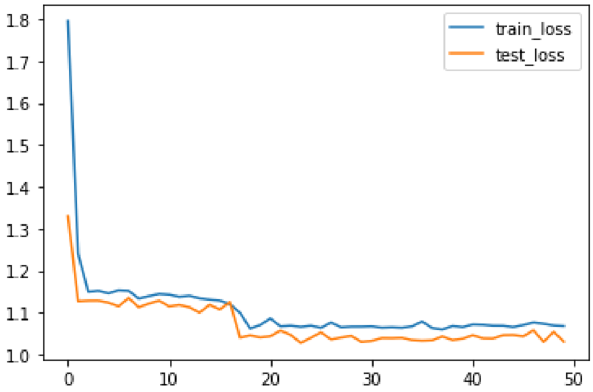
where $\theta = (w_1, \dots, w_n, \mu_1, \dots, \mu_n, \tau_1, \dots, \tau_n)$ and

$$\Phi_{\mu,\sigma}(x) = \frac{1}{\sigma\sqrt{2\pi}} \int_{-\infty}^{x} \exp\left(-\frac{1}{2}\left(\frac{s-\mu}{\sigma}\right)^2\right) ds.$$

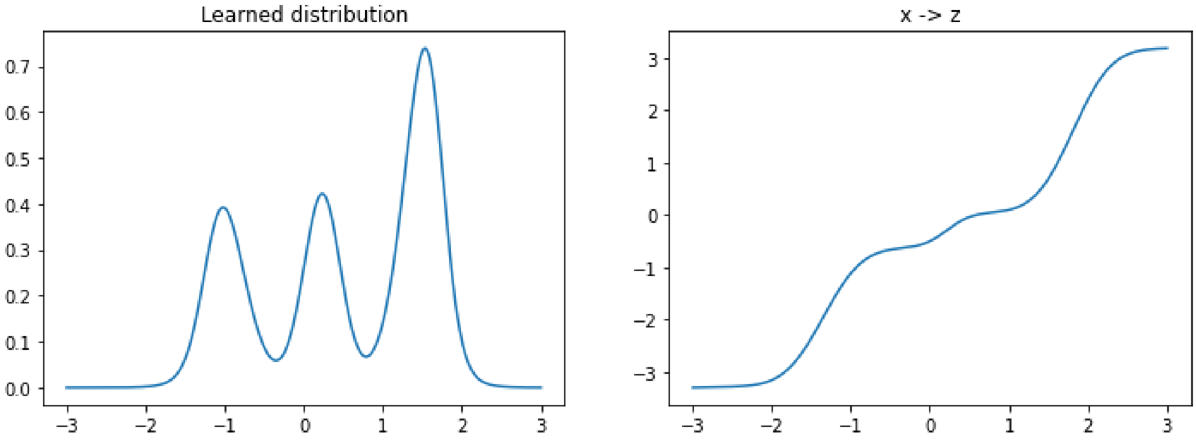
Note that $f_\theta \colon \mathbb{R} \to \mathbb{R}$. Download the starter code `normalizingFlow1d.py` and fit the flow model with $n = 5$ and $p_Z \sim \mathcal{N}(0,1)$.

Remark. Since p_Z is an unbounded distribution, we do not require $w_1, \dots, w_n$ to be normalized.

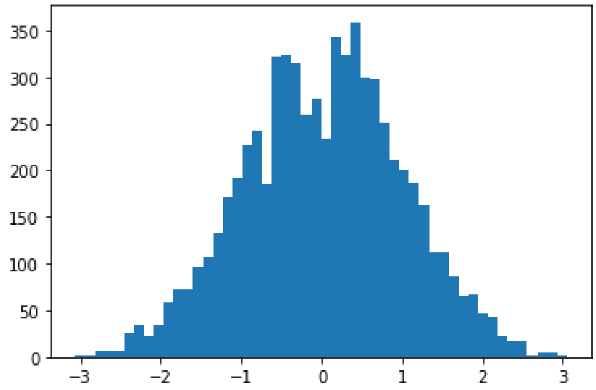
Solution. The Python code for the problem is included in the `normalizingFlow1d_sol.py` file. The true distribution of x is concatenation of three gaussian distribution $\mathcal{N}(-1, 0.25^2)$, $\mathcal{N}(0.2, 0.25^2)$, $\mathcal{N}(1.5, 0.25^2)$ with sampling ratio 1:1:2. In the solution code we used $n = 5$, epoch $= 50$, learning rate $= 5e - 2$. Following are the results.



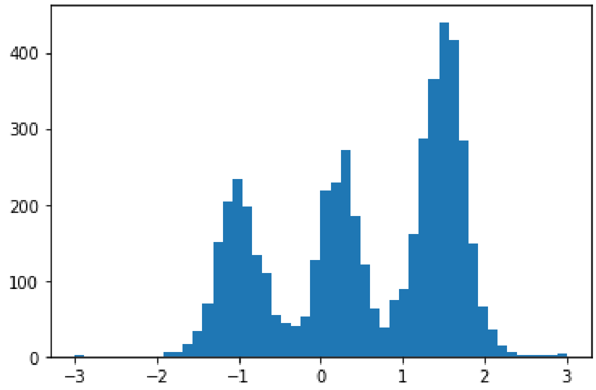
Through 50 epochs, learned distribution seems to converge to the real distribution of x. Bottom right image shows the function f_θ that sends x to $z \sim \mathcal{N}(0,1)$



The distribution of z using x and f_θ seems to follow $p_Z \sim \mathcal{N}(0, 1)$,



and the distribution of x using $z \sim \mathcal{N}(0, 1)$ and f_θ^{-1} seems to follow the true distribution of x.



■

Problem 3: *Affine coupling layer with permutations.* Consider the affine coupling layer defined as follows. Let $\Omega \subseteq \{1, \dots, n\}$ and $0 < |\Omega| < n$. Define $\Omega^{\mathsf{C}} = \{1, \dots, n\} \backslash \Omega$. For $x \in \mathbb{R}^n$, define

$$x_\Omega \in \mathbb{R}^{|\Omega|}, \qquad x_{\Omega^{\mathsf{C}}} \in \mathbb{R}^{n-|\Omega|}$$

to be the sub-vectors of x with the indices within Ω and Ω^{C} selected. Define z_Ω and $z_{\Omega^{\mathsf{C}}}$ analogously for $z \in \mathbb{R}^n$. The affine coupling layer is

$$\begin{aligned} z_\Omega &= x_\Omega \\ z_{\Omega^{\mathsf{C}}} &= e^{s_\theta(x_\Omega)} \odot x_{\Omega^{\mathsf{C}}} + t_\theta(x_\Omega), \end{aligned}$$

where $s_\theta \colon \mathbb{R}^{|\Omega|} \to \mathbb{R}^{n-|\Omega|}$ and $t_\theta \colon \mathbb{R}^{|\Omega|} \to \mathbb{R}^{n-|\Omega|}$. Show that

$$\log \left| \frac{\partial z}{\partial x} \right| = \mathbf{1}_{n-|\Omega|}^{\mathsf{T}} s_\theta(x_\Omega).$$

Clarification. We are not assuming $|\Omega| = n/2$.
Hint. Find a permutation σ such that

$$\frac{\partial z}{\partial x} = P_{\sigma^{-1}} \begin{bmatrix} I & 0 \\ * & \operatorname{diag}(e^{s_\theta(x_\Omega)}) \end{bmatrix} P_\sigma.$$

Solution. To compute $\frac{\partial z}{\partial x}$, first calculate with indices separated using Ω.

- $\frac{\partial z_\Omega}{\partial x_\Omega} = I_{|\Omega|}$.
- $\frac{\partial z_\Omega}{\partial x_{\Omega^{\mathsf{C}}}} = 0$.
- $\frac{\partial z_{\Omega^{\mathsf{C}}}}{\partial x_{\Omega^{\mathsf{C}}}} = \operatorname{diag}(e^{s_\theta(x_\Omega)})$.

Next, consider the permutation $\sigma = (\omega_1, ..., \omega_{|\Omega|}, \tilde{\omega_1}, ..., \tilde{\omega_{|\Omega^{\mathsf{C}}|}})$ where $\Omega = \{\omega_1, ..., \omega_{|\Omega|}\}$, $\Omega^{\mathsf{C}} = \{\tilde{\omega_{|\Omega^{\mathsf{C}}|}}\}$. Then, $\frac{\partial z}{\partial x}$ is :

$$\frac{\partial z}{\partial x} = P_{\sigma^{-1}} \begin{bmatrix} \frac{\partial z_\Omega}{\partial x_\Omega} & \frac{\partial z_\Omega}{\partial x_{\Omega^{\mathsf{C}}}} \\ \frac{\partial z_{\Omega^{\mathsf{C}}}}{\partial x_\Omega} & \frac{\partial z_{\Omega^{\mathsf{C}}}}{\partial x_{\Omega^{\mathsf{C}}}} \end{bmatrix} P_\sigma = P_{\sigma^{-1}} \begin{bmatrix} I & 0 \\ * & \operatorname{diag}(e^{s_\theta(x_\Omega)}) \end{bmatrix} P_\sigma.$$

Finally, the determinant of $\frac{\partial z}{\partial x}$ is

$$\left| \frac{\partial z}{\partial x} \right| = \left| P_{\sigma^{-1}} \begin{bmatrix} I & 0 \\ * & \operatorname{diag}(e^{s_\theta(x_\Omega)}) \end{bmatrix} P_\sigma \right| = \left| \operatorname{diag}(e^{s_\theta(x_\Omega)}) \right| = e^{\mathbf{1}_{(n-|\Omega|)}^{\mathsf{T}} s_\theta(x_\Omega)}$$

since $|P_\sigma| = |P_\sigma^{\mathsf{T}}| \in \{1, -1\}$.
Thus,

$$\log \left| \frac{\partial z}{\partial x} \right| = \mathbf{1}_{(n-|\Omega|)}^{\mathsf{T}} s_\theta(x_\Omega).$$

■

Problem 4: D_{KL} *of continuous random variables.* The KL-divergence between continuous random variables $X \sim f$ and $Y \sim g$, where f and g are probability density functions in $\mathbb{R}^d$, is

$$D_{\mathrm{KL}}\left(X \| Y\right) = \int_{\mathbb{R}^d} f(x) \log\left(\frac{f(x)}{g(x)}\right)\, dx.$$

(a) Show that

$$D_{\mathrm{KL}}\left(X \| Y\right) \geq 0.$$

(b) Show that if $X = (X_1, \ldots, X_d)$ is a continuous random variable such that $X_1, \ldots, X_d$ are independent and $Y = (Y_1, \ldots, Y_d)$ is a continuous random variable such that $Y_1, \ldots, Y_d$ are independent, then

$$D_{\mathrm{KL}}(X \| Y) = D_{\mathrm{KL}}(X_1 \| Y_1) + \cdots + D_{\mathrm{KL}}(X_d \| Y_d).$$

Solution.

(a) Since $x \mapsto -\log x$ is a convex function, Jensen's inequality can be used as

$$\begin{aligned} D_{\mathrm{KL}}\left(X \| Y\right) &= -\int_{\mathbb{R}^d} f(x) \log\left(\frac{g(x)}{f(x)}\right)\, dx \\ &\geq -\log\left(\int_{\mathbb{R}^d} f(x) \frac{g(x)}{f(x)} dx\right) \\ &= -\log\left(\int_{\mathbb{R}^d} g(x)\, dx\right) = 0. \end{aligned}$$

(b) If X and Y are a continuous random variable with d independent variables, $f((x_1, \ldots, x_d)) = f_1(x_1) f_2(x_2) \ldots f_d(x_d)$ and $g((x_1, \ldots, x_d)) = g_1(x_1) g_2(x_2) \ldots g_d(x_d)$. Therefore,

$$\begin{aligned} & D_{\mathrm{KL}}\left(X \| Y\right) \\ &= \int_{\mathbb{R}^d} f(x) \log\left(\frac{f(x)}{g(x)}\right)\, dx \\ &= \int_{\mathbb{R}} \cdots \int_{\mathbb{R}} f_1(x_1) \ldots f_d(x_d) \log\left(\frac{f_1(x_1) \ldots f_d(x_d)}{g_1(x_1) \ldots g_d(x_d)}\right)\, dx_1 \ldots dx_d \\ &= \int_{\mathbb{R}} \cdots \int_{\mathbb{R}} f_1(x_1) \ldots f_d(x_d) \left(\log\left(\frac{f_1(x_1)}{g_1(x_1)}\right) + \cdots + \left(\frac{f_d(x_d)}{g_d(x_d)}\right)\right)\, dx_1 \ldots dx_d \\ &= \int_{\mathbb{R}} f_1(x_1) \log\left(\frac{f_1(x_1)}{g_1(x_1)}\right)\, dx_1 \underbrace{\int_{\mathbb{R}} f_2(x_2)\, dx_2}_{=1} \cdots \underbrace{\int_{\mathbb{R}} f_d(x_d) dx_d}_{=1} \\ &\quad + \cdots + \underbrace{\int_{\mathbb{R}} f_1(x_1) dx_1}_{=1} \cdots \underbrace{\int_{\mathbb{R}} f_{d-1}(x_{d-1})\, dx_{d-1}}_{=1} \int_{\mathbb{R}} f_d(x_d) \log\left(\frac{f_d(x_d)}{g_d(x_d)}\right)\, dx_d \\ &= \int_{\mathbb{R}} f_1(x_1) \log\left(\frac{f_1(x_1)}{g_1(x_1)}\right)\, dx_1 + \cdots + \int_{\mathbb{R}} f_d(x_d) \log\left(\frac{f_d(x_d)}{g_d(x_d)}\right)\, dx_d \\ &= D_{\mathrm{KL}}(X_1 \| Y_1) + \cdots + D_{\mathrm{KL}}(X_d \| Y_d). \end{aligned}$$

■

Problem 5: D_{KL} *of Gaussian random variables.* Let $\mathcal{N}(\mu, \Sigma)$ denote the Gaussian distribution with mean μ and covariance Σ. So if $X \sim \mathcal{N}(\mu, \Sigma)$, then

$$\mathbb{E}[X] = \mu, \qquad \mathbb{E}[(X-\mu)(X-\mu)^\intercal] = \Sigma.$$

Show that

$$D_{\mathrm{KL}}\left(\mathcal{N}(\mu_0, \Sigma_0) \| \mathcal{N}(\mu_1, \Sigma_1)\right) = \frac{1}{2}\left(\operatorname{tr}\left(\Sigma_1^{-1}\Sigma_0\right) + (\mu_1 - \mu_0)^\intercal \Sigma_1^{-1}(\mu_1 - \mu_0) - d + \log\left(\frac{\det \Sigma_1}{\det \Sigma_0}\right)\right),$$

where d is the underlying dimension of the random variables $\mathcal{N}(\mu_0, \Sigma_0)$ and $\mathcal{N}(\mu_1, \Sigma_1)$. Assume Σ_0 and Σ_1 are positive definite.

Solution. Let $P \sim \mathcal{N}(\mu_0, \Sigma_0)$, $Q \sim \mathcal{N}(\mu_1, \Sigma_1)$. From previous homework,

$$\begin{aligned}
p_P(x) &= \frac{1}{\sqrt{(2\pi)^d \det \Sigma_0}} \exp\Big(-\frac{1}{2}(x-\mu_0)^\intercal \Sigma^{-1}(x-\mu_0)\Big), \\
p_Q(x) &= \frac{1}{\sqrt{(2\pi)^d \det \Sigma_1}} \exp\Big(-\frac{1}{2}(x-\mu_1)^\intercal \Sigma^{-1}(x-\mu_1)\Big).
\end{aligned}$$

Then,

$$\begin{aligned}
& D_{\mathrm{KL}}(P\|Q) \\
&= \mathbb{E}_P[\log P - \log Q] \\
&= \frac{1}{2}\mathbb{E}_P\Big[-\log\det\Sigma_0 - (x-\mu_0)^\intercal \Sigma_0^{-1}(x-\mu_0) + \log\det\Sigma_1 + (x-\mu_1)^\intercal\Sigma_1^{-1}(x-\mu_1)\Big] \\
&= \frac{1}{2}\log\det\frac{\Sigma_1}{\Sigma_0} + \frac{1}{2}\mathbb{E}_P\Big[-(x-\mu_0)^\intercal\Sigma_0^{-1}(x-\mu_0) + (x-\mu_1)^\intercal\Sigma_1^{-1}(x-\mu_1)\Big] \\
&= \frac{1}{2}\log\det\frac{\Sigma_1}{\Sigma_0} + \frac{1}{2}\mathbb{E}_P\Big[-\operatorname{Tr}\Big((x-\mu_0)^\intercal\Sigma_0^{-1}(x-\mu_0)\Big) + \operatorname{Tr}\Big((x-\mu_1)^\intercal\Sigma_1^{-1}(x-\mu_1)\Big)\Big] \\
&= \frac{1}{2}\log\det\frac{\Sigma_1}{\Sigma_0} + \frac{1}{2}\mathbb{E}_P\Big[-\operatorname{Tr}\Big(\Sigma_0^{-1}(x-\mu_0)(x-\mu_0)^\intercal\Big) + \operatorname{Tr}\Big(\Sigma_1^{-1}(x-\mu_1)(x-\mu_1)^\intercal\Big)\Big].
\end{aligned}$$

Since trace of scalar is equal to scalar, and multiplication inside trace is commutative.
By using the fact that

$$\begin{aligned}
\mathbb{E}_P\Big[-\operatorname{Tr}\Big(\Sigma_0^{-1}(x-\mu_0)(x-\mu_0)^\intercal\Big)\Big] &= -\operatorname{Tr}\Big(\Sigma_0^{-1}\mathbb{E}_P\Big[(x-\mu_0)(x-\mu_0)^\intercal\Big]\Big) \\
&= -\operatorname{Tr}(\Sigma_0^{-1}\Sigma_0) = -\operatorname{Tr}(I_d) \\
&= -d,
\end{aligned}$$

D_{KL} can be arranged as

$$\begin{aligned}
& D_{\mathrm{KL}}(P\|Q) \\
&= \frac{1}{2}\log\det\frac{\Sigma_1}{\Sigma_0} - \frac{d}{2} + \frac{1}{2}\mathbb{E}_P\Big[\operatorname{Tr}\Big(\Sigma_1^{-1}(x-\mu_1)(x-\mu_1)^\intercal\Big)\Big] \\
&= \frac{1}{2}\log\det\frac{\Sigma_1}{\Sigma_0} - \frac{d}{2} + \frac{1}{2}\mathbb{E}_P\Big[\operatorname{Tr}\Big(\Sigma_1^{-1}\Big(xx^\intercal - 2x\mu_1 + \mu_1\mu_1^\intercal\Big)\Big)\Big] \\
&= \frac{1}{2}\log\left(\frac{\det\Sigma_1}{\det\Sigma_0}\right) - \frac{d}{2} + \frac{1}{2}\operatorname{Tr}\Big(\Sigma_1^{-1}\Big(\Sigma_0 + \mu_0\mu_0^\intercal - 2\mu_0\mu_1 + \mu_1\mu_1^\intercal\Big)\Big) \\
&= \frac{1}{2}\Bigg(\log\left(\frac{\det\Sigma_1}{\det\Sigma_0}\right) - d + \frac{1}{2}\operatorname{Tr}\Big(\Sigma_1^{-1}\Big(\Sigma_0 + \mu_0\mu_0^\intercal - 2\mu_0\mu_1 + \mu_1\mu_1^\intercal\Big)\Big)\Bigg) \\
&= \frac{1}{2}\Bigg(\operatorname{Tr}(\Sigma_1^{-1}\Sigma_0) + (\mu_1-\mu_0)^\intercal\Sigma_1^{-1}(\mu_1-\mu_0) - d + \log\left(\frac{\det\Sigma_1}{\det\Sigma_0}\right)\Bigg).
\end{aligned}$$

■

Problem 6: *When maximizing a lower bound is tight.* Consider the optimization problem

$$\underset{\theta\in\Theta}{\text{maximize}} \quad f(\theta).$$

Informally assume f is an intractable function, i.e., evaluating $f(\theta)$ is difficult. However, assume there exists a decomposition

$$f(\theta) = g(\theta, \phi) + h(\theta, \phi) \qquad \forall\, \phi \in \Phi,$$

where g is tractable, i.e., evaluating $g(\theta, \phi)$ is easy, $h(\theta, \phi) \geq 0$ for all $\theta \in \Theta$ and $\phi \in \Phi$, and for any $\theta \in \Theta$ there exists a $\phi \in \Phi$ such that $h(\theta, \phi) = 0$, i.e., for any $\theta \in \Theta$,

$$\min_{\phi\in\Phi} h(\theta, \phi) = 0$$

and the minimum is attained. Now we consider the following problem with the tractable objective function

$$\underset{\theta\in\Theta,\,\phi\in\Phi}{\text{maximize}} \quad g(\theta, \phi).$$

Show that the two optimization problems are equivalent in the sense that

$$\operatorname{argmax} f = \{\theta \,|\, (\theta, \phi) \in \operatorname{argmax} g\}.$$

Hint. Use the fact that

$$\sup_{\theta,\phi} g(\theta, \phi) = \sup_{\theta} \left(\sup_{\phi} g(\theta, \phi) \right).$$

Remark. Training variational autoencoders involves maximizing the variational lower bound (VLB/ELBO). If the encoder network is infinitely expressive (if the encoder network can represent any function), maximizing the VLB is equivalent to maximizing the log-likelihood. This problem abstracts the explanation of why that is the case.

Solution. First, we will show $\operatorname{argmax} f \subseteq \{\theta \,|\, (\theta, \phi) \in \operatorname{argmax} g\}$. Let $\theta^* \in \operatorname{argmax} f$. Then $h(\theta, \phi) \geq 0$ implies that $g(\theta, \phi) \leq f(\theta^*)$ for all $(\theta, \phi) \in \Theta \times \Phi$. Therefore, $\sup_{\theta,\phi} g(\theta, \phi) \leq f(\theta^*)$. And let's take $\phi^* \in \Phi$ such that $h(\theta^*, \phi^*) = 0$.

$$f(\theta^*) \geq \sup_{\theta,\phi} g(\theta, \phi) \geq g(\theta^*, \phi^*) = f(\theta^*)$$

Hence $g(\theta^*, \phi^*) = \sup_{\theta,\phi} g(\theta, \phi)$ which implies $\operatorname{argmax} f \subseteq \{\theta \,|\, (\theta, \phi) \in \operatorname{argmax} g\}$.

Second, we will show $\{\theta \,|\, (\theta, \phi) \in \operatorname{argmax} g\} \subseteq \operatorname{argmax} f$. For given $\theta \in \Theta$, there exists $\phi_\theta \in \Phi$ such that $h(\theta, \phi_\theta) = 0$. Since $h(\theta, \phi) \geq 0$, $\sup_\phi g(\theta, \phi) = g(\theta, \phi_\theta)$. Let $(\theta^*, \phi^*) \in \operatorname{argmax} g$.

$$g(\theta^*, \phi^*) = \sup_{\theta,\phi} g(\theta, \phi) = \sup_{\theta} \sup_{\phi} g(\theta, \phi) = \sup_{\theta} g(\theta, \phi_\theta) = \sup_{\theta} f(\theta)$$

Hence $\{\theta \,|\, (\theta, \phi) \in \operatorname{argmax} g\} \subseteq \operatorname{argmax} f$. ■