
Open-Set Recognition with Gaussian Mixture Variational Autoencoders: Supplementary Material

Alexander Cao

Department of Industrial Engineering and Management Sciences
Northwestern University
Evanston, IL 60208
a-cao@u.northwestern.edu

Yuan Luo

Department of Preventive Medicine
Northwestern University
Chicago, IL 60611
yuan.luo@northwestern.edu

Diego Klabjan

Department of Industrial Engineering and Management Sciences
Northwestern University
Evanston, IL 60208
d-klabjan@northwestern.edu

1 Neural network assumptions

We call a neural network f_τ an n -headed neural network if

1. $f_\tau : \mathbb{R}^m \rightarrow \prod_{i=1}^n \mathbb{R}^s$, i.e. it maps b to (a_1, a_2, \dots, a_n) with $a_i \in \mathbb{R}^s$,
2. for each i , $1 \leq i \leq n$, we have $a_i = f_{\ell_i}^i \circ f_{\ell_i-1}^i \circ \dots \circ f_{t+1}^i \circ f_t \circ f_{t-1} \circ \dots \circ f_1(b)$ for an integer t not depending on i , $\ell_i \geq t + 1$, and each f_j, f_j^i is a typical neural network single layer parameterized by a matrix and a bias vector, and it includes an activation function. Vector τ corresponds to all these parameters.

In GMVAE, neural networks corresponding to q_{ϕ_z}, q_{ϕ_w} are 2-headed neural networks (mean and covariance) with ϕ_z, ϕ_w denoting all of the respective parameters. Probability p_θ is a 1 or 2-headed network with parameters θ , and p_β for $\beta = (\beta_{K_1}, \beta_{K_2}, \dots, \beta_{K_C})$ consists of a $(2 \sum_{c=1}^C K_c)$ -headed neural network.

Assumption 1. In each network $q_{\phi_z}, q_{\phi_w}, p_\theta$, and p_β , the last layer in each head $f_{\ell_i}^i$ has an identity activation function.

Assumption 2. Neural network $p_{\beta'}$ for $\beta' = (\beta_{K_1}, \dots, \beta_{K_{c+1}}, \dots, \beta_{K_C})$ consists of p_β with simply two additional heads, while all other network architectures are the same.

Lemma 1. Under Assumption 1 for an n -headed network, we have that given any $\bar{a} = (\bar{a}_1, \dots, \bar{a}_n)$, there exists $\tau = \tau(\bar{a})$ such that $f_\tau(b) = \bar{a}$ for every b .

Proof. Let \bar{a} be given. We define τ to consist of 0 matrices and biases for each layer except $f_{\ell_i}^i$. In $f_{\ell_i}^i$, the matrix is 0 but the bias is \bar{a}_i . Since $f_{\ell_i}^i$ has the identity activation, it follows $f_\tau(b) = \bar{a}$ for every b . \square

2 Proof of Proposition 1

Proposition 1. *Let us assume that $x \in \mathcal{X}$ is distributed as $x \sim p_{\text{data}} = \mathcal{B}(\mu_x)$, $C = 1$, and Assumption 1 holds. Then the optimal GMVAE loss is constant with respect to K . In fact, we have that $\min -\mathbb{E}_{\mathcal{X}}[\mathcal{L}(K)] = -\mathbb{E}_{\mathcal{X}}[\log p_{\text{data}}]$ for every $K \geq 1$ and a globally optimal solution reads*

$$\left. \begin{aligned} \mu(x; \phi_z^*) &= \mu_{c=1,k}(w; \beta^*) = \mu_z \\ \sigma^2(x; \phi_z^*) &= \sigma_{c=1,k}^2(w; \beta^*) = \sigma_z^2 \\ \mu(x, y; \phi_w^*) &= \bar{0} \\ \sigma^2(x, y; \phi_w^*) &= \bar{1} \\ \mu(z; \theta^*) &= \mu_x \end{aligned} \right\} \quad (1)$$

for any constant vectors μ_z, σ_z .

Proof. Note that $(\phi_z^*, \phi_w^*, \beta^*, \theta^*)$ exist due to Assumption 1 and Lemma 1. First, we show that (θ^*, β^*) given in (1) maximize the log likelihood $\mathbb{E}_{\mathcal{X}}[\log p_{\theta, \beta}(x|y=1)]$ and results in $p_{\theta^*, \beta^*}(x|y=1) = p_{\text{data}}$. We have

$$KL(p_{\text{data}} \| p_{\theta, \beta}(x|y=1)) = \mathbb{E}_{\mathcal{X}}[\log p_{\text{data}}] - \mathbb{E}_{\mathcal{X}}[\log p_{\theta, \beta}(x|y=1)]$$

and thus maximizing $\mathbb{E}_{\mathcal{X}}[\log p_{\theta, \beta}(x|y=1)]$ is equivalent to minimizing $KL(p_{\text{data}} \| p_{\theta, \beta}(x|y=1))$. The global minimum of $KL(p_{\text{data}} \| p_{\theta, \beta}(x|y=1))$ is clearly when $p_{\text{data}} = p_{\theta, \beta}(x|y=1)$. This is indeed the case for (θ^*, β^*) , since

$$\begin{aligned} p_{\theta^*, \beta^*}(x|y=1) &= \int_{w, z, v} p_{\beta^*, \theta^*}(x, v, w, z|y=1) dw dz dv \\ &= \int_{w, z, v} p_{\theta^*}(x|z) p_{\beta^*}(z|w, y=1, v) p(v|y=1) p(w) dw dz dv \\ &= \int_{w, z, v} p_{\text{data}} p_{\beta^*}(z|w, y=1, v) p(v|y=1) p(w) dw dz dv \\ &= p_{\text{data}} \end{aligned} \quad (2)$$

because of GMVAE's generative model factorization and (1). Now we have

$$\begin{aligned} \mathbb{E}_{\mathcal{X}}[\log p_{\text{data}}] &= \mathbb{E}_{\mathcal{X}}[\log p_{\theta^*, \beta^*}(x|y=1)] \\ &= \mathbb{E}_{\mathcal{X}} \left[\mathbb{E}_{q_{\phi^*}(v, w, z|x, y=1)} \left[\log \frac{p_{\theta^*, \beta^*}(x, z, w, v|y=1)}{q_{\phi^*}(v, w, z|x, y=1)} \right] \right] \\ &+ \mathbb{E}_{\mathcal{X}} \left[\mathbb{E}_{q_{\phi^*}(v, w, z|x, y=1)} \left[\log \frac{q_{\phi^*}(v, w, z|x, y=1)}{p_{\theta^*, \beta^*}(z, w, v|x, y=1)} \right] \right] \\ &= \mathbb{E}_{\mathcal{X}}[\mathcal{L}(K; \phi_z^*, \phi_w^*, \beta^*, \theta^*)] + \mathbb{E}_{\mathcal{X}}[\text{VG}(\phi_z^*, \phi_w^*, \beta^*, \theta^*)] \end{aligned} \quad (3)$$

$$= \mathbb{E}_{\mathcal{X}}[\mathcal{L}(K; \phi_z^*, \phi_w^*, \beta^*, \theta^*)] + \mathbb{E}_{\mathcal{X}}[\text{VG}(\phi_z^*, \phi_w^*, \beta^*, \theta^*)] \quad (4)$$

where $\text{VG}(\phi_z^*, \phi_w^*, \beta^*, \theta^*)$ corresponds to (3). We next show that $\text{VG}(\phi_z^*, \phi_w^*, \beta^*, \theta^*) = 0$. This together with the facts that maximized $\mathbb{E}_{\mathcal{X}}[\mathcal{L}(K; \phi_z, \phi_w, \beta, \theta)]$ corresponds with minimized $\mathbb{E}_{\mathcal{X}}[\text{VG}(\phi_z, \phi_w, \beta, \theta)]$, and $\text{VG}(\phi_z, \phi_w, \beta, \theta) \geq 0$ (it is a KL divergence), shows optimality.

From (1) we have that $p_{\theta^*}(x|z) = p_{\text{data}}(x)$ for all x and z and thus with (2) we have

$$\begin{aligned} p_{\theta^*, \beta^*}(z, w, v|x, y=1) &= \frac{p_{\theta^*}(x|z, w, v, y=1) p_{\beta^*}(z, w, v|y=1)}{p_{\theta^*, \beta^*}(x|y=1)} \\ &= \frac{p_{\theta^*}(x|z) p_{\beta^*}(z, w, v|y=1)}{p_{\text{data}}(x)} \\ &= p_{\beta^*}(z, w, v|y=1). \end{aligned} \quad (5)$$

The reconstruction term $p_{\theta}(x|z, w, v, y=1) = p_{\theta}(x|z)$ for every θ because in GMVAE, data reconstruction depends only on z and is independent of w and v (see §3.1 of the paper).

Also from Bayes' and GMVAE's generative model factorization, we have the following simplification

$$p_{\beta^*}(v|z, w, y=1) = \frac{p_{\beta^*}(z|w, y=1, v) p(v|y=1) p(w)}{p_{\beta^*}(z, w|y=1)}$$

$$\begin{aligned}
&= \frac{p_{\beta^*}(z|w, y=1, v)p(v|y=1)p(w)}{p_{\beta^*}(z|w, y=1)p(w|y=1)} \\
&= \frac{p_{\beta^*}(z|w, y=1, v)p(v|y=1)}{\sum_{v'} p_{\beta^*}(z|w, y=1, v')p(v'|y=1)} \tag{6} \\
&= p(v|y=1) \tag{7}
\end{aligned}$$

where (1) is only used in the last line. Substituting (5) into $\text{VG}(\phi_z^*, \phi_w^*, \beta^*, \theta^*)$ we obtain

$$\begin{aligned}
&\text{VG}(\phi_z^*, \phi_w^*, \beta^*, \theta^*) \\
&= \mathbb{E}_{q_{\phi^*}(v, w, z|x, y=1)} \left[\log \frac{q_{\phi^*}(v, w, z|x, y=1)}{p_{\theta^*, \beta^*}(z, w, v|x, y=1)} \right] \\
&= \mathbb{E}_{q_{\phi^*}(v, w, z|x, y=1)} \left[\log \frac{q_{\phi^*}(v, w, z|x, y=1)}{p_{\beta^*}(z, w, v|y=1)} \right] \\
&= \mathbb{E}_{p_{\beta^*}(v|z, w, y=1)q_{\phi_w^*}(w|x, y=1)q_{\phi_z^*}(z|x)} \left[\log \frac{p_{\beta^*}(v|z, w, y=1)q_{\phi_w^*}(w|x, y=1)q_{\phi_z^*}(z|x)}{p_{\beta^*}(z|w, y=1, v)p(w)p(v|y=1)} \right] \\
&= \mathbb{E}_{q_{\phi_w^*}(w|x, y=1)q_{\phi_z^*}(z|x)} \left[\log q_{\phi_z^*}(z|x) - \sum_{j=1}^K p_{\beta^*}(v=j|z, w, y=1) \log p_{\beta^*}(z|w, y=1, v=j) \right] \\
&+ KL(q_{\phi_w^*}(w|x, y=1)||p(w)) \\
&+ \mathbb{E}_{q_{\phi_w^*}(w|x, y=1)q_{\phi_z^*}(z|x)} [KL(p_{\beta^*}(v|z, w, y=1)||p(v|y=1))] \\
&= 0
\end{aligned}$$

due to (1) and (7). To complete the proof, simply note that negating (4) yields $-\mathbb{E}_{\mathcal{X}}[\mathcal{L}(K; \phi_z^*, \phi_w^*, \beta^*, \theta^*)] = -\mathbb{E}_{\mathcal{X}}[\log p_{\text{data}}]$. \square

3 Proof of Proposition 2

Lemma 2. For every $\delta > 0$ and μ , there exists σ^2 such that if $f(z)$ is the pdf of a d -dimensional Normal random vector with mean μ and diagonal covariance σ^2 then

$$f(z) \leq \delta \quad \text{for every } z.$$

Proof. Let $u = (\frac{1}{\delta}(2\pi)^{-d/2})^{1/d}$ and $\sigma = (u, \dots, u)$. We have

$$f(z) = \prod_i \frac{1}{\sigma_i \sqrt{2\pi}} \exp \left\{ -\frac{1}{2\sigma_i^2} (z_i - \mu_i)^2 \right\} \leq \prod_i \frac{1}{\sigma_i \sqrt{2\pi}} = \delta. \quad \square$$

Proposition 2. Let us assume $C = 1$, Assumptions 1 and 2 hold, and that $p(v|y=1)$ is uniform in the appropriate dimension. We have

$$\min \{-\mathbb{E}_{\mathcal{X}}[\mathcal{L}(K; \phi_z, \phi_w, \beta, \theta)]\} - \min \{-\mathbb{E}_{\mathcal{X}}[\mathcal{L}(K+1; \phi_z, \phi_w, \beta, \theta)]\} \geq \epsilon_K$$

where $-\log 2 \leq \log(K/(K+1)) \leq \epsilon_K$ for all K .

Proof. We show that for every solution $(\phi'_z, \phi'_w, \beta', \theta')$ to $\min \mathbb{E}_{\mathcal{X}}[-\mathcal{L}(K; \phi_z, \phi_w, \beta, \theta)]$, there exists a corresponding solution $(\phi_z^*, \phi_w^*, \beta^*, \theta^*)$ such that

$$-\mathbb{E}_{\mathcal{X}}[\mathcal{L}(K; \phi'_z, \phi'_w, \beta', \theta')] = -\mathbb{E}_{\mathcal{X}}[\mathcal{L}(K+1; \phi_z^*, \phi_w^*, \beta^*, \theta^*)] + \epsilon_K.$$

Let us assume that $(\phi'_z, \phi'_w, \beta', \theta')$ minimizes $-\mathbb{E}_{\mathcal{X}}[\mathcal{L}(K; \phi_z, \phi_w, \beta, \theta)]$. Then we can choose

$$\begin{aligned}
\phi_z^* &= \phi'_z \\
\phi_w^* &= \phi'_w \\
\theta^* &= \theta'
\end{aligned} \tag{8}$$

which is a valid choice by Assumption 2, and have β^* such that

$$p_{\beta^*}(z|w, y=1, v) = p_{\beta'}(z|w, y=1, v) \quad \text{for all } v \leq K \tag{9}$$

$$p_{\beta^*}(z|w, y = 1, v = K + 1) \leq \delta \quad \text{for every } z, w \quad (10)$$

for any fixed $0 < \delta < 1/e$. Conditions (9) and (10) are always possible due to Assumptions 1 and 2 and Lemmas 1 and 2. In essence, we choose β^* such that the first K subcluster generative distributions are the same as the case β' but we take the $(K + 1)$ -th subcluster generative distribution to map all points w to the same Normal distribution with large enough covariance.

Inserting (9) and (10) into (6) and combined with uniform priors, we get that

$$p_{\beta^*}(v = K + 1|z, w, y = 1) = \frac{p_{\beta^*}(z|w, y = 1, v = K + 1)}{\sum_{j=1}^K p_{\beta'}(z|w, y = 1, v = j) + p_{\beta^*}(z|w, y = 1, v = K + 1)} \quad (11)$$

and

$$\begin{aligned} p_{\beta^*}(v = k|z, w, y = 1) &= \frac{p_{\beta'}(z|w, y = 1, v = k)}{\sum_{j=1}^K p_{\beta'}(z|w, y = 1, v = j) + p_{\beta^*}(z|w, y = 1, v = K + 1)} \\ &\leq \frac{p_{\beta'}(z|w, y = 1, v = k)}{\sum_{j=1}^K p_{\beta'}(z|w, y = 1, v = j)} = p_{\beta'}(v = k|z, w, y = 1) \end{aligned} \quad (12)$$

for all $k \leq K$. The absolute difference between the two posteriors for $k \leq K$ in (12) is bounded by a factor of δ as follows:

$$\begin{aligned} &\left| p_{\beta^*}(v = k|z, w, y = 1) - p_{\beta'}(v = k|z, w, y = 1) \right| \\ &= \left| \frac{p_{\beta'}(z|w, y = 1, v = k)}{\sum_{j=1}^K p_{\beta'}(z|w, y = 1, v = j) + p_{\beta^*}(z|w, y = 1, v = K + 1)} - \frac{p_{\beta'}(z|w, y = 1, v = k)}{\sum_{j=1}^K p_{\beta'}(z|w, y = 1, v = j)} \right| \\ &= \frac{p_{\beta^*}(z|w, y = 1, v = K + 1)p_{\beta'}(z|w, y = 1, v = k)}{\left(\sum_{j=1}^K p_{\beta'}(z|w, y = 1, v = j) + p_{\beta^*}(z|w, y = 1, v = K + 1) \right)} \\ &\quad \times \frac{1}{\sum_{j=1}^K p_{\beta'}(z|w, y = 1, v = j)} \\ &\leq \delta \frac{p_{\beta'}(z|w, y = 1, v = k)}{\left(\sum_{j=1}^K p_{\beta'}(z|w, y = 1, v = j) \right)^2} \\ &= \delta A(z, w, v = k). \end{aligned} \quad (13)$$

Now we calculate ϵ_K given by

$$\mathbb{E}_{\mathcal{X}}[-\mathcal{L}(K; \phi'_z, \phi'_w, \beta', \theta')] - \mathbb{E}_{\mathcal{X}}[-\mathcal{L}(K + 1; \phi_z^*, \phi_w^*, \beta^*, \theta^*)] = \epsilon_K$$

Because of (8), ϵ_K simplifies to

$$\begin{aligned} \epsilon_K &= \\ &- \mathbb{E}_{\mathcal{X}} \left[\mathbb{E}_{q_{\phi_w^*}(w|x, y=1)q_{\phi_z^*}(z|x)} \left[\sum_{j=1}^K p_{\beta'}(v = j|z, w, y = 1) \log p_{\beta'}(z|w, y = 1, v = j) \right] \right] \\ &+ \mathbb{E}_{\mathcal{X}} \left[\mathbb{E}_{q_{\phi_w^*}(w|x, y=1)q_{\phi_z^*}(z|x)} \left[\sum_{j=1}^{K+1} p_{\beta^*}(v = j|z, w, y = 1) \log p_{\beta^*}(z|w, y = 1, v = j) \right] \right] \\ &+ \mathbb{E}_{\mathcal{X}} \left[\mathbb{E}_{q_{\phi_w^*}(w|x, y=1)q_{\phi_z^*}(z|x)} [KL(p_{\beta'}(v|z, w, y = 1)||p_K(v|y = 1))] \right] \\ &- \mathbb{E}_{\mathcal{X}} \left[\mathbb{E}_{q_{\phi_w^*}(w|x, y=1)q_{\phi_z^*}(z|x)} [KL(p_{\beta^*}(v|z, w, y = 1)||p_{K+1}(v|y = 1))] \right] \\ &= \epsilon_K^{(1)} + \epsilon_K^{(2)} \end{aligned}$$

where $p_K(v|y=1)$ indicates that v is K -dimensional, and $\epsilon_K^{(1)}$ are the first two terms while $\epsilon_K^{(2)}$ are the last two terms.

We first analyze $\epsilon_K^{(1)}$. For brevity, we combine the expectations and simply write $\mathbb{E}[\cdot]$. Together with (9), (11), and (13), we get

$$\begin{aligned}
\left| \epsilon_K^{(1)} \right| &= \left| -\mathbb{E} \left[\sum_{j=1}^K p_{\beta'}(v=j|z, w, y=1) \log p_{\beta'}(z|w, y=1, v=j) \right] \right. \\
&\quad + \mathbb{E} \left[\sum_{j=1}^K p_{\beta^*}(v=j|z, w, y=1) \log p_{\beta'}(z|w, y=1, v=j) \right] \\
&\quad \left. + \mathbb{E} [p_{\beta^*}(v=K+1|z, w, y=1) \log p_{\beta^*}(z|w, y=1, v=K+1)] \right| \\
&= \left| \mathbb{E} \left[\sum_{j=1}^K \log p_{\beta'}(z|w, y=1, v=j) (p_{\beta^*}(v=j|z, w, y=1) - p_{\beta'}(v=j|z, w, y=1)) \right] \right| \\
&\quad + \mathbb{E} \left[\frac{p_{\beta^*}(z|w, y=1, v=K+1) \log p_{\beta^*}(z|w, y=1, v=K+1)}{\sum_{j=1}^K p_{\beta'}(z|w, y=1, v=j) + p_{\beta^*}(z|w, y=1, v=K+1)} \right] \Big| \\
&\leq \delta \cdot \mathbb{E} \left[\sum_{j=1}^K \left| \log p_{\beta'}(z|w, y=1, v=j) \right| A(z, w, v=j) \right] \\
&\quad + |\delta(\log \delta)| \mathbb{E} \left[\frac{1}{\sum_{j=1}^K p_{\beta'}(z|w, y=1, v=j)} \right] = o(1), \tag{14}
\end{aligned}$$

where the last inequality follows from $|x \log x|$ being increasing for $x \leq 1/e$ and in $o(1)$ we consider $\delta \rightarrow 0$.

Next we study $\epsilon_K^{(2)}$. For shorthand, let us define

$$\begin{aligned}
&\log((K+1)p_{\beta^*}(v=K+1|z, w, y=1)) \\
&= \log \left(\frac{(K+1)p_{\beta^*}(z|w, y=1, v=K+1)}{\sum_{j=1}^K p_{\beta'}(z|w, y=1, v=j) + p_{\beta^*}(z|w, y=1, v=K+1)} \right) \\
&= \log p_{\beta^*}(z|w, y=1, v=K+1) + B(z, w)
\end{aligned}$$

and note that

$$\begin{aligned}
|B(z, w)| &= \left| \log \left(\frac{(K+1)}{\sum_{j=1}^K p_{\beta'}(z|w, y=1, v=j) + p_{\beta^*}(z|w, y=1, v=K+1)} \right) \right| \\
&\leq \max \left\{ \left| \log \left(\frac{(K+1)}{\sum_{j=1}^K p_{\beta'}(z|w, y=1, v=j)} \right) \right|, \right. \\
&\quad \left. \left| \log \left(\frac{(K+1)}{\sum_{j=1}^K p_{\beta'}(z|w, y=1, v=j) + 1/e} \right) \right| \right\} \\
&= C(z, w).
\end{aligned}$$

We have

$$\begin{aligned}
&\epsilon_K^{(2)} \\
&= \mathbb{E} \left[\sum_{j=1}^K p_{\beta'}(v=j|z, w, y=1) \log(K p_{\beta'}(v=j|z, w, y=1)) \right]
\end{aligned}$$

$$\begin{aligned}
& - \mathbb{E} \left[\sum_{j=1}^K p_{\beta^*}(v = j|z, w, y = 1) \log((K+1)p_{\beta^*}(v = j|z, w, y = 1)) \right] \\
& - \mathbb{E} [p_{\beta^*}(v = K+1|z, w, y = 1) \log((K+1)p_{\beta^*}(v = K+1|z, w, y = 1))] \\
& = \mathbb{E} \left[\sum_{j=1}^K (\log K)p_{\beta'}(v = j|z, w, y = 1) - (\log(K+1))p_{\beta^*}(v = j|z, w, y = 1) \right] \\
& + \mathbb{E} \left[\sum_{j=1}^K p_{\beta'}(v = j|z, w, y = 1) \log p_{\beta'}(v = j|z, w, y = 1) \right. \\
& \quad \left. - p_{\beta^*}(v = j|z, w, y = 1) \log p_{\beta^*}(v = j|z, w, y = 1) \right] \\
& - \mathbb{E} [p_{\beta^*}(v = K+1|z, w, y = 1) \log((K+1)p_{\beta^*}(v = K+1|z, w, y = 1))] \\
& \geq \log(K) - (\log(K+1))\mathbb{E} \left[\sum_{j=1}^K p_{\beta^*}(v = j|z, w, y = 1) \right] \\
& + \mathbb{E} \left[\sum_{j=1}^K (p_{\beta'}(v = j|z, w, y = 1) - p_{\beta^*}(v = j|z, w, y = 1)) \log(p_{\beta'}(v = j|z, w, y = 1)) \right] \quad (15) \\
& - \left| \mathbb{E} \left[\frac{p_{\beta^*}(z|w, y = 1, v = K+1) \log p_{\beta^*}(z|w, y = 1, v = K+1)}{\sum_{j=1}^K p_{\beta'}(z|w, y = 1, v = j) + p_{\beta^*}(z|w, y = 1, v = K+1)} \right] \right| \\
& - \left| \mathbb{E} \left[\left(\frac{p_{\beta^*}(z|w, y = 1, v = K+1)}{\sum_{j=1}^K p_{\beta'}(z|w, y = 1, v = j) + p_{\beta^*}(z|w, y = 1, v = K+1)} \right) B(z, w) \right] \right| \\
& \geq \log(K) - \log(K+1) \\
& - \delta \cdot \mathbb{E} \left[\sum_{j=1}^K A(z, w, v = j) \left| \log(p_{\beta'}(v = j|z, w, y = 1)) \right| \right] \quad (16) \\
& - \delta (\log \delta) \mathbb{E} \left[\frac{1}{\sum_{j=1}^K p_{\beta'}(z|w, y = 1, v = j)} \right] \quad (17) \\
& - \delta \cdot \mathbb{E} \left[\frac{1}{\sum_{j=1}^K p_{\beta'}(z|w, y = 1, v = j)} C(z, w) \right] \\
& = \log \frac{K}{K+1} + o(1).
\end{aligned}$$

In (15) we use (12), in (16) we rely on (13), and in (17) we use (14) again.

To summarize, we have $\epsilon_K \geq -|\epsilon_K^{(1)}| + \epsilon_K^{(2)} \geq -o(1) + o(1) + \log \frac{K}{K+1} = \log \frac{K}{K+1} + o(1)$. Thus $\epsilon_K \geq \log \frac{K}{K+1}$. \square