1 Preliminary

The Kullback-Leibler (KL) divergence between densities $q(y)$ and $\hat{p}(y)$ is defined as

$$D_{KL}(q||\hat{p}) = \int q(y) \log \frac{q(y)}{\hat{p}(y)} dy. \tag{1}$$

Suppose $q(y)$ is a deterministic "target" distribution and $\hat{p}(y)$ is an estimate of $q(y)$, e.g., a probability statement derived from the output of a neural network. We have a (possibly infinite) ensemble of such estimators. Expectation with respect to this ensemble is indicated by the operator $\mathbb{E}_\Omega$ where Ω refers to all estimators.

2 Average Model $\overline{p}(y)$

It is intuitive to assume that the average model $\overline{p}(y)$ is an arithmetic mean of $\hat{p}(y)$, however, we first prove that $\overline{p}(y)$ can be a (normalized) geometric mean of the densities $\hat{p}(y)$. Define $\overline{p}$ to the following average distribution

$$\overline{p} = \arg \min_{a:\int a(y)dy=1} \mathbb{E}_\Omega[D_{KL}(a||\hat{p})] = \arg \min_{a:\int a(y)dy=1} ED_{KL}(a||\hat{p}) \tag{2}$$

where $\overline{p}$ has the smallest average distance to all estimators with the constraint $\int a(y)dy = 1$. By introducing a Lagrange multiplier μ for the constraint $\int a(y)dy = 1$ and taking the function derivative[1] to $a(y)$,

$$\begin{aligned}
\int \frac{\delta ED_{KL}}{\delta \overline{p}} \phi(y) dy &= \left[\frac{d}{d\epsilon} \left[ED_{KL}[\overline{p} + \epsilon\phi] + \mu(1 - \int (\overline{p} + \epsilon\phi) dy) \right] \right]_{\epsilon=0} && (3)\\
&= \left[\frac{d}{d\epsilon} \mathbb{E}_\Omega[D_{KL}(\overline{p} + \epsilon\phi||\hat{p})] \right]_{\epsilon=0} + \left[\frac{d}{d\epsilon} \mu(1 - \int (\overline{p} + \epsilon\phi) dy) \right]_{\epsilon=0} && (4)\\
&= \left[\frac{d}{d\epsilon} \mathbb{E}_\Omega[\int (\overline{p} + \epsilon\phi) \log \frac{\overline{p} + \epsilon\phi}{\hat{p}} dy] \right]_{\epsilon=0} - \mu \int \phi dy && (5)\\
&= \left[\frac{d}{d\epsilon} \int (\overline{p} + \epsilon\phi) \mathbb{E}_\Omega[\log \frac{\overline{p} + \epsilon\phi}{\hat{p}}] dy \right]_{\epsilon=0} - \mu \int \phi dy && (6)\\
&= \left[\int (\phi \mathbb{E}_\Omega[\log \frac{\overline{p} + \epsilon\phi}{\hat{p}}] + (\overline{p} + \epsilon\phi) \frac{\phi}{\overline{p}}) dy \right]_{\epsilon=0} - \mu \int \phi dy && (7)\\
&= \int (\phi \mathbb{E}_\Omega[\log \frac{\overline{p}}{\hat{p}}] + \phi) dy - \mu \int \phi dy && (8)\\
&= \int (\mathbb{E}_\Omega[\log \frac{\overline{p}}{\hat{p}}] + 1 - \mu) \phi(y) dy && (9)\\
\frac{\delta ED_{KL}}{\delta \overline{p}} &= \mathbb{E}_\Omega[\log \frac{\overline{p}}{\hat{p}}] + 1 - \mu = \log \overline{p} - \mathbb{E}_\Omega[\log \hat{p}] + 1 - \mu && (10)
\end{aligned}$$

where $\phi(y)$ is an arbitrary function (ϕ for short). The quantity $\epsilon\phi$ is called the variation of $\overline{p}$. Note that we exchange the order of $\int$ and $\mathbb{E}_\Omega$ since the expectation $\mathbb{E}_\Omega$ is defined on $\hat{p}$ instead of $\overline{p}$. We also exchange the order of $\int$ and $\frac{\delta}{\delta\epsilon}$ according to the Lebesgue's dominated convergence theorem [2]. By setting $\frac{\delta ED_{KL}}{\delta \overline{p}}$ to zero (i.e., Equation (10)), we easily obtain the average model

$$\overline{p}(y) = \frac{1}{Z} \exp \left[\mathbb{E}_\Omega[\log \hat{p}(y)] \right] \tag{11}$$

where Z a normalization constant independent of y.

[1] https://en.wikipedia.org/wiki/Functional_derivative

[2] You may assume that the sufficient conditions hold in our case, though it has NOT yet been rigorously proved.

3 Bias

The bias is defined as the distance $D_{KL}(q, \overline{p})$ between the average model and the target distribution.

$$Bias = D_{KL}(q, \overline{p}) \tag{12}$$

Substituting Equation (11) into (12), we obtain

$$\begin{aligned} Bias = \int q \log \frac{q}{\overline{p}} dy &= \int q \log q dy - \int q \log \frac{1}{Z} \exp\left(\mathbb{E}_\Omega[\log \hat{p}]\right) && (13) \\ &= \int q \log q dy + \int q \log Z dy - \int q \mathbb{E}_\Omega[\log \hat{p}] dy && (14) \\ &= \mathbb{E}_\Omega[\int q \log q dy] + \int q \log Z dy - \mathbb{E}_\Omega[\int q \log \hat{p} dy] && (15) \\ &= \mathbb{E}_\Omega[\int q \log \frac{q}{\hat{p}} dy] + \log Z && (16) \\ &= \mathbb{E}_\Omega[D_{KL}(q||\hat{p})] + \log Z && (17) \end{aligned}$$

Here we utilize $\mathbb{E}[c] = c$ if c is a constant. The expected value of an integral is an iterated integral, and the normal mathematical rules for interchange of integrals apply to (15).

If you are uncomfortable with $\mathbb{E}_\Omega[\int q \log \hat{p} dy] = \int q \mathbb{E}_\Omega[\log \hat{p}] dy$, the expectation formulation is easier to understand

$$\mathbb{E}_\Omega[\int q \log \hat{p} dy] = \mathbb{E}_\Omega\left[\mathbb{E}_q[\log \hat{p}]\right] = \mathbb{E}_q\left[\mathbb{E}_\Omega[\log \hat{p}]\right]. \tag{18}$$

4 Variance

The variance is defined as the expected distance $\mathbb{E}_\Omega[D_{KL}(\overline{p}||\hat{p})]$ between the average model and every single estimator

$$Variance = \mathbb{E}_\Omega[D_{KL}(\overline{p}||\hat{p})] = -\mathbb{E}_\Omega[\int \overline{p} \log \frac{\hat{p}}{\overline{p}} dy] = -\int \overline{p} \mathbb{E}_\Omega[\log \frac{\hat{p}}{\overline{p}}] dy. \tag{19}$$

Recalling Equation (11),

$$\log Z = \mathbb{E}_\Omega[\log \hat{p}] - \log \overline{p} = \mathbb{E}_\Omega[\log \hat{p}] - \mathbb{E}_\Omega[\log \overline{p}] = \mathbb{E}_\Omega[\log \frac{\hat{p}}{\overline{p}}]. \tag{20}$$

Since $\log Z$ is a constant, $\mathbb{E}_\Omega[\log \frac{\hat{p}}{\overline{p}}]$ is also a constant independent of y. Considering that $\int \overline{p} dy = 1$, we have

$$\log Z = \log Z \int \overline{p} dy = \mathbb{E}_\Omega[\log \frac{\hat{p}}{\overline{p}}] \int \overline{p} dy = \int \overline{p} \mathbb{E}_\Omega[\log \frac{\hat{p}}{\overline{p}}] dy. \tag{21}$$

Combining Equation (19) and (21),

$$Variance = -\log Z. \tag{22}$$

5 Error

Here we present two ways to prove the decomposition of Bias/Variance for KL divergence.

5.1 Bottom-up

Using Equation (17) and (22),

$$Error = \mathbb{E}_\Omega[D_{KL}(q||\hat{p})] = Bias - \log Z = Bias + Variance \tag{23}$$

5.2 Top-down

$$\begin{aligned}
Error &= \mathbb{E}_{\Omega}[D_{KL}(q||\hat{p})] && (24)\\
&= \mathbb{E}_{\Omega}[\int q \log \frac{q}{\hat{p}} dy] && (25)\\
&= \mathbb{E}_{\Omega}[\int (q \log q - q \log \hat{p}) dy] && (26)\\
&= \mathbb{E}_{\Omega}[\int (q \log q - q \log \hat{p}) dy] - \int q \log \overline{p} dy + \int q \log \overline{p} dy && (27)\\
&= \int q \log q dy - \mathbb{E}_{\Omega}[\int q \log \hat{p} dy] - \int q \log \overline{p} dy + \int q \log \overline{p} dy && (28)\\
&= (\int q \log q dy - \int q \log \overline{p} dy) + (\int q \log \overline{p} dy - \mathbb{E}_{\Omega}[\int q \log \hat{p} dy]) && (29)\\
&= D_{KL}(q||\overline{p}) + (\mathbb{E}_{\Omega}[\int q \log \overline{p} dy] - \mathbb{E}_{\Omega}[\int q \log \hat{p} dy]) && (30)\\
&= D_{KL}(q||\overline{p}) + \mathbb{E}_{\Omega}[\int q \log \frac{\overline{p}}{\hat{p}} dy] && (31)\\
&= D_{KL}(q||\overline{p}) + \int q \mathbb{E}_{\Omega}[\log \frac{\overline{p}}{\hat{p}}] dy && (32)\\
&= D_{KL}(q||\overline{p}) + \int \overline{p} \mathbb{E}_{\Omega}[\log \frac{\overline{p}}{\hat{p}}] dy && (33)\\
&= D_{KL}(q||\overline{p}) + \mathbb{E}_{\Omega}[\int \overline{p} \log \frac{\overline{p}}{\hat{p}} dy] && (34)\\
&= D_{KL}(q||\overline{p}) + \mathbb{E}_{\Omega}[D_{KL}(\overline{p}||\hat{p})] && (35)\\
&= Bias + Variance && (36)
\end{aligned}$$

For Equation (30), we use the result of Equation (11) that

$$\mathbb{E}_{\Omega}[\log \overline{p}(y)] = \mathbb{E}_{\Omega}[\log \frac{1}{Z} + \mathbb{E}_{\Omega}[\log \hat{p}(y)]] = \log \frac{1}{Z} + \mathbb{E}_{\Omega}[\log \hat{p}(y)] = \log \overline{p}(y) \quad (37)$$

Then, we have

$$\mathbb{E}_{\Omega}[\int q \log \overline{p} dy] = \int q \mathbb{E}_{\Omega}[\log \overline{p}] dy = \int q \log \overline{p} dy. \quad (38)$$

For Equation (32) and (33), we use the result of Equation (21) that $\mathbb{E}_{\Omega}[\log \frac{\hat{p}}{\overline{p}}]$ is a constant and $\int c \cdot q(y) dy = \int c \cdot \overline{p}(y) dy = c$.