Let

- $\hat{\theta} = \hat{\theta}(X)$ be the MLE, where X is the observed data to estimate θ,
- θ^* be the minimizer of

$$\mathrm{KL}(p, g) \quad = \quad \mathrm{E}_p\left[\log\left(\frac{p(X)}{g(X \mid \theta)}\right)\right]$$

  with repect to θ,
- x^* is a future observation and

$$R_n \quad = \quad \int p(x^*) \log g\left(x^* \mid \hat{\theta}\right) dx^*.$$

An naive estimator of $\mathrm{E}[R_n]$ is

$$\frac{1}{n}\sum_{i=1}^{n} \log g\left(x_i \mid \hat{\theta}\right).$$

But this is a biased estimator, since X^* should be independent of $\hat{\theta}$, if we assume data are independent.

The Taylor expansion yields

$$\begin{aligned}
\frac{1}{n}\sum_{i=1}^{n} \log g\left(x_i \mid \hat{\theta}\right) \quad &= \quad \frac{1}{n}\sum_{i=1}^{n} \log g(x_i \mid \theta^*) + \frac{\partial \frac{1}{n}\sum_{i=1}^{n} \log g(x_i \mid \theta^*)}{\partial \theta^T}\left(\hat{\theta} - \theta^*\right) \\
&\quad + \frac{1}{2}\left(\hat{\theta} - \theta^*\right)^T \underbrace{\frac{\partial^2 \frac{1}{n}\sum_{i=1}^{n} \log g(x_i \mid \theta^*)}{\partial \theta \partial \theta^T}}_{\to -H(\theta^*)}\left(\hat{\theta} - \theta^*\right) + \text{Remainder},
\end{aligned}$$

where $H(\theta^*) = -\mathrm{E}\left[\frac{\partial^2 \log g(x^* \mid \theta^*)}{\partial \theta \partial \theta^T}\right]$. Similarly, Taylor expansion yields

$$\log g\left(x^* \mid \hat{\theta}\right) \quad = \quad \log g(x^* \mid \theta^*) + \frac{\partial \log g(x^* \mid \theta^*)}{\partial \theta^T}\left(\hat{\theta} - \theta^*\right) + \frac{1}{2}\left(\hat{\theta} - \theta^*\right)^T \frac{\partial^2 \log g(x^* \mid \theta^*)}{\partial \theta \partial \theta^T}\left(\hat{\theta} - \theta^*\right) + \text{Remainder}.$$

We plug in the expansion to R_n and obtain

$$\begin{aligned}
R_n \quad &= \quad \int p(x^*) \log g(x^* \mid \theta^*) dx^* + \int p(x^*) \frac{\partial \log g(x^* \mid \theta^*)}{\partial \theta^T} dx^* \left(\hat{\theta} - \theta^*\right) \\
&\quad + \frac{1}{2}\left(\hat{\theta} - \theta^*\right)^T \underbrace{\int p(x^*) \frac{\partial^2 \log g(x^* \mid \theta^*)}{\partial \theta \partial \theta^T} dx^*}_{= -H(\theta^*)}\left(\hat{\theta} - \theta^*\right) + \text{Remainder}.
\end{aligned}$$

Since θ^* also minimizes $\int p(x) \log g(x \mid \theta) dx$, we should have

$$0 = \frac{\partial \int p(x) \log g(x \mid \theta) dx}{\partial \theta} \quad = \quad \int p(x^*) \frac{\partial \log g(x^* \mid \theta^*)}{\partial \theta} dx^*.$$

Thus, we can write R_n as

$$R_n \quad = \quad \int p(x^*) \log g(x^*, \theta^*) dx^* - \frac{1}{2n}\sqrt{n}\left(\hat{\theta} - \theta^*\right)^T H(\theta^*) \sqrt{n}\left(\hat{\theta} - \theta^*\right) + \text{Remainder}.$$

Under some regularity conditions, the MLE $\hat{\theta}$ satisfies

$$V_n = \sqrt{n}\left(\hat{\theta} - \theta^*\right) \quad \approx \quad N\left(0, H^{-1}\mathcal{I}H^{-1}\right),$$

where $\mathcal{I}$ is the Fisher information. Hence, using V_n we obtain

$$\frac{1}{n}\sum_{i=1}^{n}\log g\left(x_i \mid \hat{\theta}\right) = \frac{1}{n}\sum_{i=1}^{n}\log g\left(x_i \mid \theta^*\right) + \frac{1}{\sqrt{n}}\frac{\partial \frac{1}{n}\sum_{i=1}^{n}\log g\left(x_i \mid \theta^*\right)}{\partial \theta^T}V_n - \frac{1}{2n}V_n^T H V_n + \text{Remainder},$$

and

$$R_n = \int p\left(x^*\right)\log g\left(x^* \mid \theta^*\right)dx^* - \frac{1}{2n}V_n^T H V_n + \text{Remainder}.$$

Thus, the bias becomes

$$\begin{aligned}
\mathrm{E}\left(\frac{1}{n}\sum_{i=1}^{n}\log g\left(x_i \mid \hat{\theta}\right) - \mathrm{E}\left[R_n\right]\right) &= \underbrace{\mathrm{E}\left\{\frac{1}{n}\sum_{i=1}^{n}\left[\log g\left(x_i \mid \theta^*\right) - \int p\left(x^*\right)\log g\left(x^* \mid \theta^*\right)dx^*\right]\right\}}_{\text{term 1}} \\
&\quad + \frac{1}{\sqrt{n}}\underbrace{\mathrm{E}\left\{\frac{\partial \frac{1}{n}\sum_{i=1}^{n}\log g\left(x_i \mid \theta^*\right)}{\partial \theta^T}V_n\right\}}_{\text{term 2}} + \text{Remainder}.
\end{aligned}$$

1. For the first term, $\int p\left(x^*\right)\log g\left(x^* \mid \theta^*\right)dx^*$ is just the expected value of $\log g\left(x_i \mid \theta^*\right)$. Hence, the expected value of the first term is 0.

2. For the second term, the Taylor expansion yields

$$0 = \frac{\partial \frac{1}{n}\sum_{i=1}^{n}\log g\left(x_i \mid \hat{\theta}\right)}{\partial \theta} = \frac{\partial \frac{1}{n}\sum_{i=1}^{n}\log g\left(x_i \mid \theta^*\right)}{\partial \theta} + \frac{\partial^2 \frac{1}{n}\sum_{i=1}^{n}\log g\left(x_i \mid \theta^*\right)}{\partial \theta \partial \theta^T}\left(\hat{\theta} - \theta^*\right) + \text{Remainder}.$$

Hence,

$$\begin{aligned}
\frac{\partial \frac{1}{n}\sum_{i=1}^{n}\log g\left(x_i \mid \theta^*\right)}{\partial \theta} &= -\frac{\partial^2 \frac{1}{n}\sum_{i=1}^{n}\log g\left(x_i \mid \theta^*\right)}{\partial \theta \partial \theta^T}\left(\hat{\theta} - \theta^*\right) + \text{Remainder} \\
&= H\left(\hat{\theta} - \theta^*\right) + \text{Remainder},
\end{aligned}$$

and

$$\begin{aligned}
\mathrm{E}\left\{\frac{\partial \frac{1}{n}\sum_{i=1}^{n}\log g\left(x_i \mid \theta^*\right)}{\partial \theta^T}V_n\right\} &= \mathrm{E}\left\{\left(\hat{\theta} - \theta^*\right)^T H V_n\right\} + \text{Remainder} \\
&= \frac{1}{\sqrt{n}}\mathrm{E}\left\{V_n^T H V_n\right\} + \text{Remainder} \\
&= \frac{1}{\sqrt{n}}\mathrm{E}\left\{\mathrm{tr}\left[V_n^T \frac{\partial^2 \frac{1}{n}\sum_{i=1}^{n}\log g\left(x_i \mid \theta^*\right)}{\partial \theta \partial \theta^T}V_n\right]\right\} + \text{Remainder} \\
&= \frac{1}{\sqrt{n}}\mathrm{tr}\left\{H\mathrm{E}\left[V_n V_n^T\right]\right\} + \text{Remainder} \\
&= \frac{1}{\sqrt{n}}\mathrm{tr}\left\{\mathcal{I}H^{-1}\right\} + \text{Remainder}.
\end{aligned}$$

This means that the bias is

$$\mathrm{E}\left(\frac{1}{n}\sum_{i=1}^{n}\log g\left(x_i \mid \hat{\theta}\right) - \mathrm{E}\left[R_n\right]\right) = \frac{1}{n}\mathrm{tr}\left\{\mathcal{I}H^{-1}\right\} + \text{Remainder}.$$

A bias corrected estimator is

$$\frac{1}{n}\sum_{i=1}^{n}\log g\left(x_i, \hat{\theta}\right) - \frac{1}{n}\mathrm{tr}\left\{\mathcal{I}H^{-1}\right\}.$$