3 Information Measures

Kullback-Leibler divergence

$$\mathrm{KL}_s(A,B) := D_{\mathrm{KL}}(\mathcal{P}(s|A)||\mathcal{P}(s|B)) = \int \mathrm{d}s\, \mathcal{P}(s|A)\, \ln\left(\frac{\mathcal{P}(s|A)}{\mathcal{P}(s|B)}\right)$$

amount of extra information on s contained in A with respect to B

Units: $\begin{cases} \text{nit} = \text{nat} & \text{if } \ln \text{ is used} \\ \text{bit} = \text{shannon} & \text{if } \log_2 \text{ is used} \end{cases}$, $1\,\text{nit} = {}^1/_{\ln 2}\,\text{bit} \approx 1.44\,\text{bit}$

$$\begin{aligned}
\textbf{Information} \text{ or surprise: } \mathcal{H}(s|I) &= -\log \mathcal{P}(s|I) \\
\text{Product rule: } \mathcal{P}(d,s|I) &= \mathcal{P}(d|s,I)\,\mathcal{P}(s|I) \\
&= \mathcal{P}(s|d,I)\,\mathcal{P}(d|I) \\
\Rightarrow \mathcal{H}(d,s|I) &= \mathcal{H}(d|s,I) + \mathcal{H}(s|I) \\
&= \mathcal{H}(s|d,I) + \mathcal{H}(d|I) \Rightarrow \text{ information is additive}
\end{aligned}$$

Information Gain

Kullback-Leibler divergence

$$\begin{aligned}\mathrm{KL}_s(A,B) &= D_{\mathrm{KL}}(\mathcal{P}(s|A)||\mathcal{P}(s|B)) = \int \mathrm{d}s\, \mathcal{P}(s|A)\,\ln\left(\frac{\mathcal{P}(s|A)}{\mathcal{P}(s|B)}\right)\\ &= \left\langle \ln\left(\frac{\mathcal{P}(s|A)}{\mathcal{P}(s|B)}\right)\right\rangle_{(s|A)} = \langle \mathcal{H}(s|B) - \mathcal{H}(s|A)\rangle_{(s|A)}\end{aligned}$$

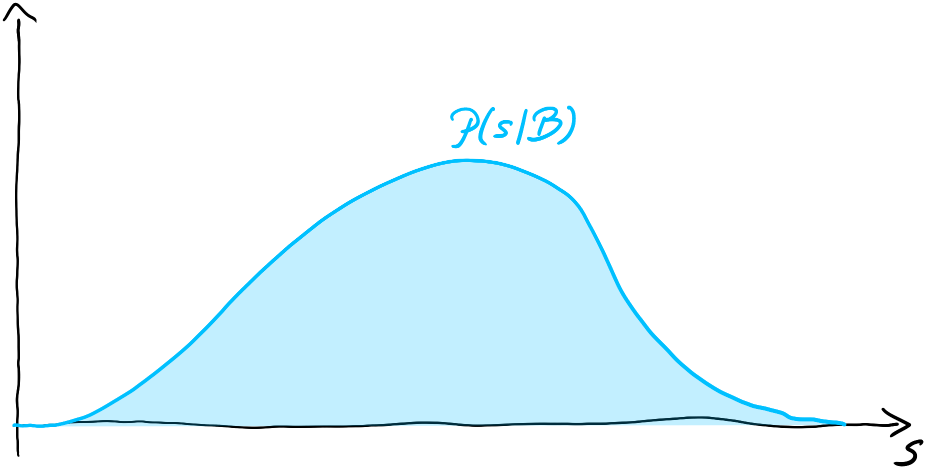
measures expected information gain on s while updating from knowledge B to A.
$\mathcal{P}(s|A)$-weighing favours regions, where $\ln \mathcal{P}(s|A) = -\mathcal{H}(s|A)$ is largest.

Example: learning result of n tosses of $\overbrace{\text{a fair coin}}^{=I}$, $d^* \in \{0,1\}^n$
prior $P(d|I) = 2^{-n}$, posterior $P(d|d^*, I) = \delta_{d,d^*}$

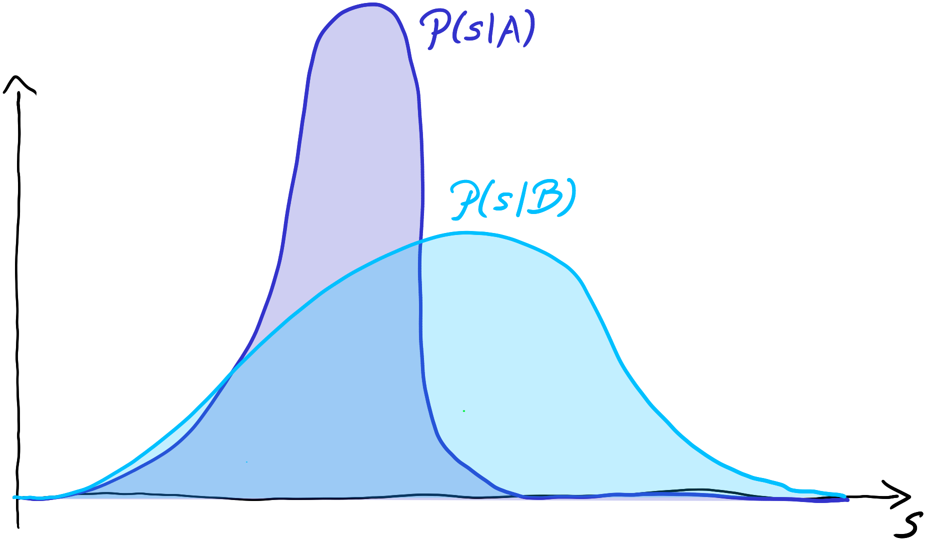
$$\begin{aligned}\frac{\mathrm{KL}_d((d^*,I),I)}{\mathrm{bit}} &= \sum_d P(d|d^*,I)\log_2\left(\frac{P(d|d^*,I)}{P(d|I)}\right) = \sum_d \delta_{d,d^*}\log_2\left(\frac{\delta_{d,d^*}}{2^{-n}}\right)\\ &= \log_2(2^n) = n\,\log_2(2) = n\end{aligned}$$

n tosses of a fair coin provide n bits information on the outcome

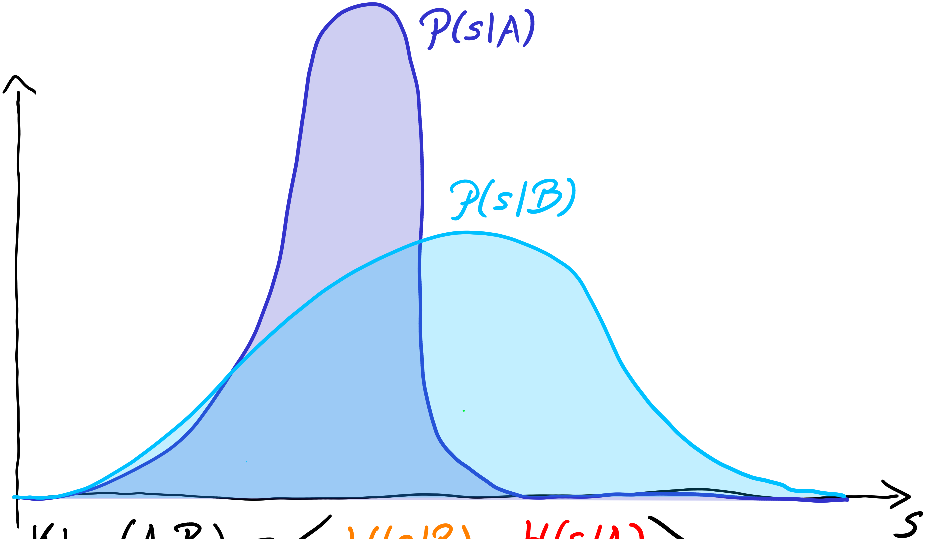
Information Gain



Information Gain

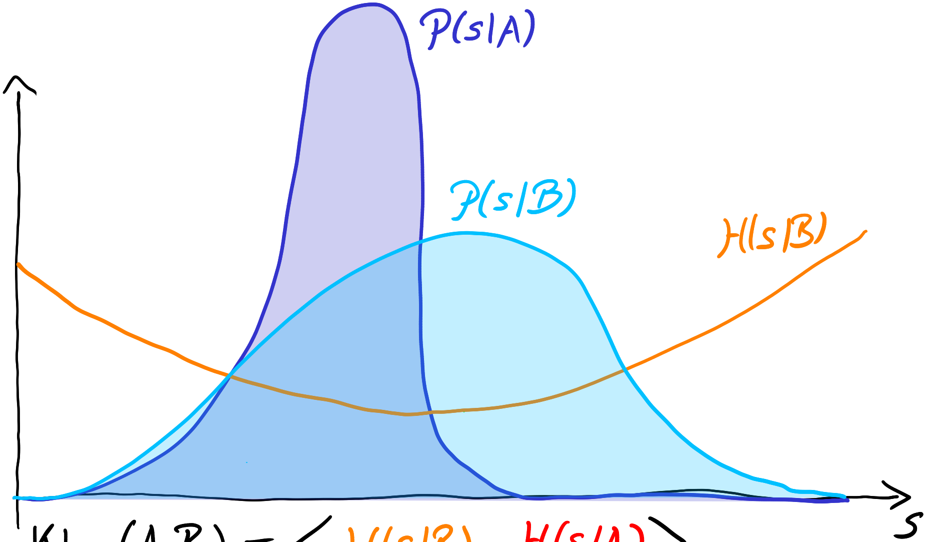


Information Gain



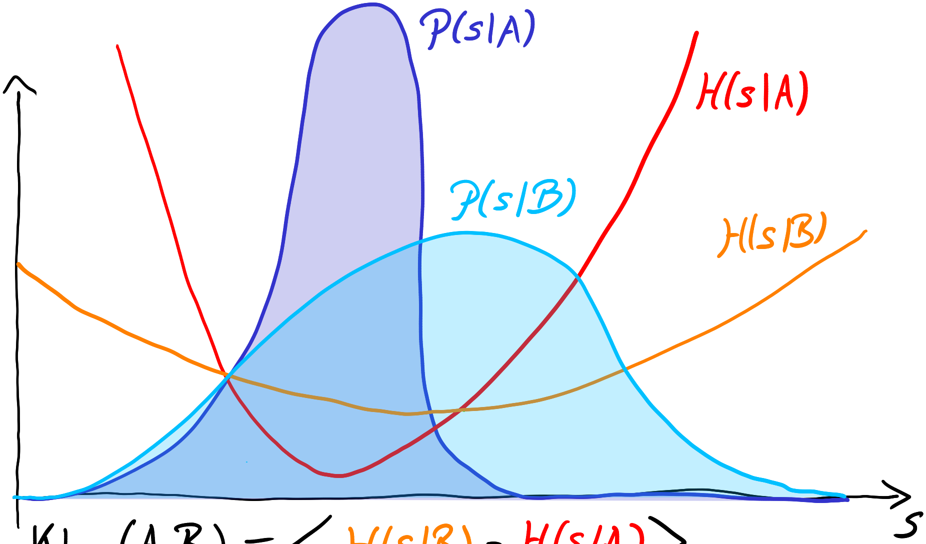
$$KL_s(A,B) = \left\langle H(s|B) - H(s|A) \right\rangle_{(s|A)}$$

Information Gain



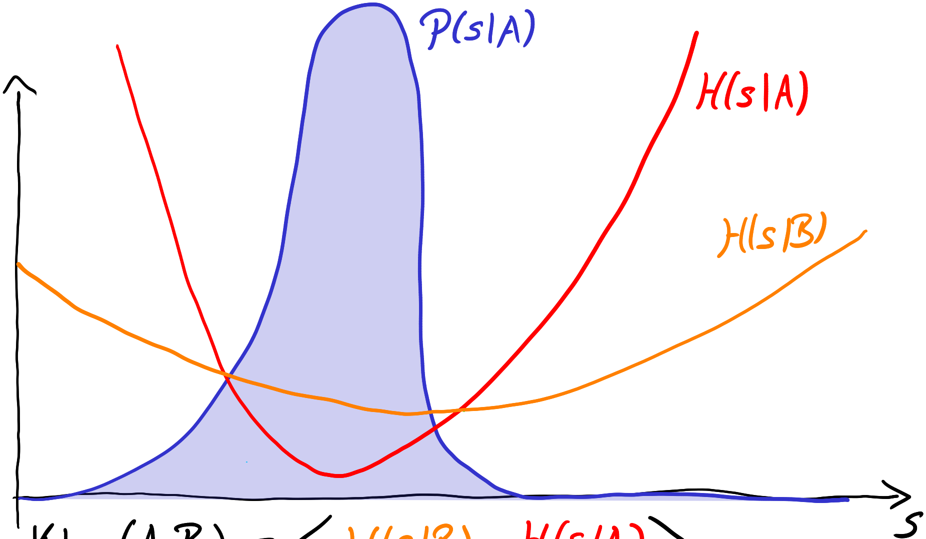
$$KL_s(A,B) = \left\langle H(s|B) - H(s|A) \right\rangle_{(s|A)}$$

Information Gain



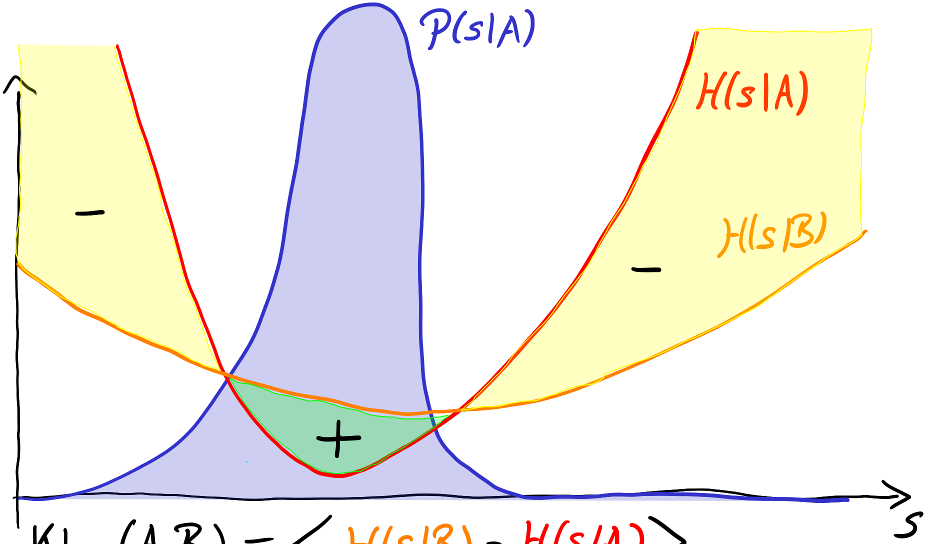
$$KL_s(A,B) = \left\langle H(s|B) - H(s|A) \right\rangle_{(s|A)}$$

Information Gain



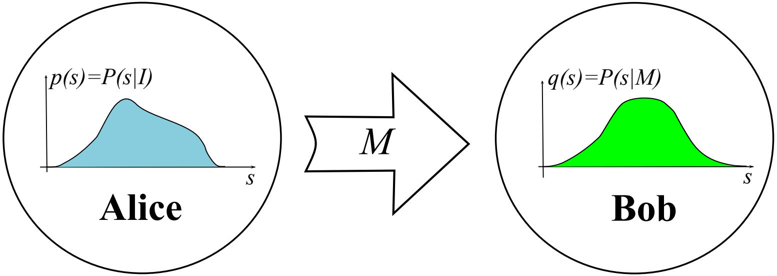
$$KL_s(A,B) = \langle H(s|B) - H(s|A) \rangle_{(s|A)}$$

Information Gain



$$KL_s(A,B) = \left\langle H(s|B) - H(s|A) \right\rangle_{(s|A)}$$

Optimal Coding



Alice chooses message M that minimizes Bob's expected surprise

$$\mathrm{KL}_s(I,M) = \langle \mathcal{H}(s|M) - \mathcal{H}(s|I) \rangle_{(s|I)}$$

and the amount of information Bob needs to update from M to Alice knowlege I.

Independence

$$x \perp y \mid I \Leftrightarrow \mathcal{P}(x, y|I) = \mathcal{P}(x|I)\,\mathcal{P}(y|I); \text{ assume } x \perp y \mid A,\ x \perp y \mid B$$

$$\begin{aligned}
\mathrm{KL}_{(x,y)}(A, B) &= \int \mathrm{d}x \int \mathrm{d}y\, \mathcal{P}(x, y|A)\, \ln\left(\frac{\mathcal{P}(x, y|A)}{\mathcal{P}(x, y|B)}\right) \\
&= \int \mathrm{d}x \int \mathrm{d}y\, \mathcal{P}(x|A)\,\mathcal{P}(y|A)\, \ln\left(\frac{\mathcal{P}(x|A)\,\mathcal{P}(y|A)}{\mathcal{P}(x|B)\,\mathcal{P}(y|B)}\right) \\
&= \int \mathrm{d}x\, \mathcal{P}(x|A) \int \mathrm{d}y\, \mathcal{P}(y|A) \left[\ln\left(\frac{\mathcal{P}(x|A)}{\mathcal{P}(x|B)}\right) + \ln\left(\frac{\mathcal{P}(y|A)}{\mathcal{P}(y|B)}\right)\right] \\
&= \int \mathrm{d}x\, \mathcal{P}(x|A) \ln\left(\frac{\mathcal{P}(x|A)}{\mathcal{P}(x|B)}\right) + \int \mathrm{d}y\, \mathcal{P}(y|A) \ln\left(\frac{\mathcal{P}(y|A)}{\mathcal{P}(y|B)}\right) \\
&= \mathrm{KL}_x(A, B) + \mathrm{KL}_y(A, B)
\end{aligned}$$

KL is additive for independent quantities.

Mutual Information

$$\begin{aligned}
\mathrm{MI}_{(x,y)}(I) &= D_{\mathrm{KL}}(\mathcal{P}(x,y|I)||\mathcal{P}(x|I)\mathcal{P}(y|I)) \\
&= \int \mathrm{d}x \int \mathrm{d}y\, \mathcal{P}(x,y|I)\, \ln\left(\frac{\mathcal{P}(x,y|I)}{\mathcal{P}(x|I)\mathcal{P}(y|I)}\right) \\
&= \langle \mathcal{H}(x|I) + \mathcal{H}(y|I) - \mathcal{H}(x,y|I)\rangle_{(x,y|I)} \;\geq 0 \\
\text{since } \mathcal{H}(x,y|I) &= \mathcal{H}(x|I) + \mathcal{H}(y|x,I) = \mathcal{H}(y|I) + \mathcal{H}(x|y,I)
\end{aligned}$$

$$\begin{aligned}
\mathrm{MI}_{(x,y)}(I) &= \langle \mathcal{H}(x|I) + \mathcal{H}(y|I) - \mathcal{H}(x,y|I)\rangle_{(x,y|I)} \\
&= \langle \mathcal{H}(x|I) - \mathcal{H}(x|y,I)\rangle_{(x,y|I)} \\
&= \langle \mathcal{H}(y|I) - \mathcal{H}(y|x,I)\rangle_{(x,y|I)} \;\geq 0
\end{aligned}$$

MI expresses the reduction of expected surprises on one variable by learning the other one.
$\mathrm{MI}_{(x,y)}(I) = 0$ for independent quantities.
MI senses relations between quantities.

Bayesian Updating

$$I \rightarrow (d, I),\, \mathcal{P}(s|I) \rightarrow \mathcal{P}(s|d,I) = \tfrac{\mathcal{P}(d|s,I)}{\mathcal{P}(d|I)} \mathcal{P}(s|I)$$

$$\begin{aligned}
\mathrm{KL}_s((d,I),I) &= \langle \mathcal{H}(s|I) - \mathcal{H}(s|d,I) \rangle_{(s|d,I)} \\
&= \int \mathrm{d}s\, \mathcal{P}(s|d,I) \ln \left(\frac{\mathcal{P}(s|d,I)}{\mathcal{P}(s|I)} \right) \\
&= \int \mathrm{d}s\, \mathcal{P}(s|d,I) \ln \left(\frac{\mathcal{P}(d|s,I)}{\mathcal{P}(d|I)} \right) \\
&= \langle \mathcal{H}(d|I) - \mathcal{H}(d|s,I) \rangle_{(s|d,I)}
\end{aligned}$$

Information gain on s by data d
How much data is less surprising if signal is known on (posterior) average.

Divergence

KL divergence is asymmetric distance measure, depends on direction

KL divergence is symmetric for small distances:

$p(s) = q(s) + \varepsilon(s);\ \varepsilon(s) \ll q(s), p(s)\ \forall s;\ \int \mathrm{d}s\, \varepsilon(s) = 0$

$$
\begin{aligned}
D_{\mathrm{KL}}(p||q) &= \int \mathrm{d}s\, p(s) \log\left(\frac{p(s)}{q(s)}\right) = \int \mathrm{d}s\ (q(s) + \varepsilon(s)) \log\left(1 + \frac{\varepsilon(s)}{q(s)}\right) \\
&= \int \mathrm{d}s\ \left\{ (q(s) + \varepsilon(s)) \left[\frac{\varepsilon(s)}{q(s)} - \frac{1}{2}\left(\frac{\varepsilon(s)}{q(s)}\right)^2\right] + \mathcal{O}(\varepsilon^3) \right\} \\
&= \int \mathrm{d}s\ \left[\varepsilon(s) + \frac{(\varepsilon(s))^2}{2q(s)} + \mathcal{O}(\varepsilon^3)\right] = 0 + \int \mathrm{d}s \frac{[p(s) - q(s)]^2}{2q(s)} + \mathcal{O}(\varepsilon^3) \\
&= \int \mathrm{d}s \frac{[p(s) - q(s)]^2}{2\sqrt{p(s)\, q(s)}} + \mathcal{O}(\varepsilon^3)
\end{aligned}
$$

${}^1\!/\!\sqrt{pq} \approx {}^1\!/\!p \approx {}^1\!/\!q$ seems to be “metric” in space of probabilities→ “information geometry”

WARNING: Original KL is not a distance! ϟ

Fisher Information Metric

Probabilities are parameterized in terms of conditional parameters, $\mathcal{P}(s|\theta)$.
Expansion in terms of those leads to **Fisher information metric** g^{ij}.

$$
\begin{aligned}
\theta &= (\theta_1, \ldots \theta_n) =: (\theta_i)_i \in \mathbb{R}^n \\
\theta' &= \theta + \varepsilon \\
\mathcal{P}(s|\theta') &= \mathcal{P}(s|\theta) + \frac{\partial \mathcal{P}(s|\theta)}{\partial \theta_i}\varepsilon_i + \mathcal{O}(\varepsilon^2), \text{ sum convention} \\
\mathrm{KL}_s(\theta', \theta) &= \underbrace{\mathrm{KL}_s(\theta, \theta)}_{=0} + \underbrace{\left.\frac{\partial \mathrm{KL}_s(\theta', \theta)}{\partial \theta'_i}\right|_{\theta'=\theta}}_{=0} \varepsilon_i + \frac{1}{2} \underbrace{\left.\frac{\partial^2 \mathrm{KL}_s(\theta', \theta)}{\partial \theta'_i \partial \theta'_j}\right|_{\theta'=\theta}}_{=g^{ij}} \varepsilon_i \varepsilon_j + \mathcal{O}(\varepsilon^3) \\
&= \frac{1}{2}\varepsilon_i g^{ij} \varepsilon_j + \mathcal{O}(\varepsilon^3)
\end{aligned}
$$

Measures expected information gain in limit of small update $\varepsilon = \theta' - \theta$.
Used to characterize sensitivity of future experiments.

Fisher Information Metric

$$
\begin{aligned}
g^{ij} &= \frac{\partial^2}{\partial\theta'_i\partial\theta'_j}\int \mathrm{d}s\,\mathcal{P}(s|\theta')\,\ln\frac{\mathcal{P}(s|\theta')}{\mathcal{P}(s|\theta)}\bigg|_{\theta'=\theta} = \frac{\partial}{\partial\theta'_i}\int \mathrm{d}s\left[\frac{\partial\mathcal{P}(s|\theta')}{\partial\theta'_j}\,\ln\frac{\mathcal{P}(s|\theta')}{\mathcal{P}(s|\theta)} + \frac{\partial\mathcal{P}(s|\theta')}{\partial\theta'_j}\right]\bigg|_{\theta'=\theta} \\
&= \frac{\partial}{\partial\theta'_i}\int \mathrm{d}s\left[\ln\frac{\mathcal{P}(s|\theta')}{\mathcal{P}(s|\theta)}+1\right]\frac{\partial\mathcal{P}(s|\theta')}{\partial\theta'_j}\bigg|_{\theta'=\theta} \\
&= \int \mathrm{d}s\left\{\frac{1}{\mathcal{P}(s|\theta')}\frac{\partial\mathcal{P}(s|\theta')}{\partial\theta'_i}\frac{\partial\mathcal{P}(s|\theta')}{\partial\theta'_j} + \left[\ln\frac{\mathcal{P}(s|\theta')}{\mathcal{P}(s|\theta)}+1\right]\frac{\partial^2\mathcal{P}(s|\theta')}{\partial\theta'_i\partial\theta'_j}\right\}\bigg|_{\theta'=\theta} \\
&= \int \mathrm{d}s\left[\frac{1}{\mathcal{P}(s|\theta)}\frac{\partial\mathcal{P}(s|\theta)}{\partial\theta_i}\frac{\partial\mathcal{P}(s|\theta)}{\partial\theta_j} + \frac{\partial^2\mathcal{P}(s|\theta)}{\partial\theta_i\partial\theta_j}\right] \\
&= \int \mathrm{d}s\,\mathcal{P}(s|\theta)\,\frac{\partial\ln\mathcal{P}(s|\theta)}{\partial\theta_i}\frac{\partial\ln\mathcal{P}(s|\theta)}{\partial\theta_j} + \underbrace{\frac{\partial^2}{\partial\theta_i\partial\theta_j}\underbrace{\int \mathrm{d}s\,\mathcal{P}(s|\theta)}_{=1}}_{=0} = \left\langle\frac{\partial\mathcal{H}(s|\theta)}{\partial\theta_i}\frac{\partial\mathcal{H}(s|\theta)}{\partial\theta_j}\right\rangle_{(s|\theta)}
\end{aligned}
$$

Fisher Information Metric

$$
\begin{aligned}
g^{ij} &= \left\langle \frac{\partial \mathcal{H}(s|\theta)}{\partial \theta_i} \frac{\partial \mathcal{H}(s|\theta)}{\partial \theta_j} \right\rangle_{(s|\theta)} = \int \mathrm{d}s \, \frac{\partial \ln \mathcal{P}(s|\theta)}{\partial \theta_i} \frac{\partial \mathcal{P}(s|\theta)}{\mathcal{P}(s|\theta)\, \partial \theta_j} \mathcal{P}(s|\theta) \\
&= \int \mathrm{d}s \, \frac{\partial \ln \mathcal{P}(s|\theta)}{\partial \theta_i} \frac{\partial \mathcal{P}(s|\theta)}{\partial \theta_j} \\
&= \frac{\partial}{\partial \theta_j} \int \mathrm{d}s \, \mathcal{P}(s|\theta) \frac{\partial \ln \mathcal{P}(s|\theta)}{\partial \theta_i} - \int \mathrm{d}s \, \mathcal{P}(s|\theta) \frac{\partial^2 \ln \mathcal{P}(s|\theta)}{\partial \theta_i \partial \theta_j} \\
&= \frac{\partial}{\partial \theta_j} \int \mathrm{d}s \, \frac{\partial \mathcal{P}(s|\theta)}{\partial \theta_i} + \left\langle \frac{\partial^2 \mathcal{H}(s|\theta)}{\partial \theta_i \partial \theta_j} \right\rangle_{(s|\theta)} = \underbrace{\frac{\partial^2}{\partial \theta_i \partial \theta_j} \underbrace{\int \mathrm{d}s \, \mathcal{P}(s|\theta)}_{=1}}_{=0} + \left\langle \frac{\partial^2 \mathcal{H}(s|\theta)}{\partial \theta_i \partial \theta_j} \right\rangle_{(s|\theta)}
\end{aligned}
$$

$$
\boxed{g^{ij} = \left\langle \frac{\partial \mathcal{H}(s|\theta)}{\partial \theta_i} \frac{\partial \mathcal{H}(s|\theta)}{\partial \theta_j} \right\rangle_{(s|\theta)} = \left\langle \frac{\partial^2 \mathcal{H}(s|\theta)}{\partial \theta_i \partial \theta_j} \right\rangle_{(s|\theta)}}
$$

End