Expected information gain in marginals

Consider the marginal posterior and marginal prior

\[ p(\theta|y, d) = \int_H p(\theta, \eta|y, d) \, d\eta, \quad p(\theta) = \int_H p(\theta, \eta) \, d\eta \]

Information gain in \( \theta \) from a single observation \( y \):

\[ u_\theta(y, d) = D_{KL}[p(\theta|y, d) || p(\theta)] \]

Expected information gain – experimental outcome is unknown beforehand:

\[ U_\theta(d) = \mathbb{E}_y[u_\theta(y, d)] = \int_Y u_\theta(y, d) \, p(y|d) \, dy \]

Expectation is taken over the prior predictive of the data
Expected information gain in marginals

Consider the marginal posterior and marginal prior

\[ p(\theta|y, d) = \int_H p(\theta, \eta|y, d) \, d\eta, \quad p(\theta) = \int_H p(\theta, \eta) \, d\eta \]

Information gain in \( \theta \) from a single observation \( y \):

\[ u_\theta(y, d) = D_{KL} \left[ \underbrace{p(\theta|y, d)}_{\text{marginal posterior}} \left\| \underbrace{p(\theta)}_{\text{marginal prior}} \right. \right] \]

**Expected information gain** – experimental outcome is unknown beforehand:

\[ U_\theta(d) = \mathbb{E}_y[u_\theta(y, d)] = \int_Y u_\theta(y, d) \, p(y|d) \, dy \]

Expectation is taken over the prior predictive of the data.
Estimating the expected information gain

**Challenge:** No closed-form expression for expected information gain

▶ Nonlinear computational models
▶ Nontrivial priors and noise models
▶ Avoid large approximations (mean-field, Gaussian, etc.)

**Solution:** use a Monte Carlo estimate

\[
U_\theta(d) = \mathbb{E}_y \left[ D_{KL} \left[ p(\theta|y,d) \parallel p(\theta) \right] \right] \\
= \int_Y \int_\Theta \log \left[ \frac{p(\theta|y,d)}{p(\theta)} \right] p(\theta|y,d) \, d\theta \, p(y|d) \, dy \\
= \int_Y \int_\Theta \left\{ \ln \left[ p(y|\theta,d) \right] - \ln \left[ p(y|d) \right] \right\} p(y|\theta,d)p(\theta) \, d\theta \, dy
\]
Estimating the expected information gain

**Challenge:** No closed-form expression for expected information gain

- Nonlinear computational models
- Nontrivial priors and noise models
- Avoid large approximations (mean-field, Gaussian, etc.)

**Solution:** use a Monte Carlo estimate

\[
U_\theta(d) = \mathbb{E}_y \left[ D_{KL} \left[ p(\theta|y, d) \mid \mid p(\theta) \right] \right]
\]

\[
= \int_Y \int_\Theta \log \left( \frac{p(\theta|y, d)}{p(\theta)} \right) p(\theta|y, d) \, d\theta \, p(y|d) \, dy
\]

\[
= \int_Y \int_\Theta \left\{ \ln \left[ p(y|\theta, d) \right] - \ln \left[ p(y|d) \right] \right\} p(y|\theta, d) p(\theta) \, d\theta \, dy
\]
Estimating the expected information gain

**Challenge:** No closed-form expression for expected information gain

- Nonlinear computational models
- Nontrivial priors and noise models
- Avoid large approximations (mean-field, Gaussian, etc.)

**Solution:** use a Monte Carlo estimate

\[
U_\theta(d) = \mathbb{E}_y \left[ D_{KL} \left[ p(\theta|y, d) \parallel p(\theta) \right] \right]
\]

\[
= \int_Y \int_\Theta \log \left[ \frac{p(\theta|y, d)}{p(\theta)} \right] p(\theta|y, d) \, d\theta \, p(y|d) \, dy
\]

\[
= \int_Y \int_\Theta \left\{ \ln \left[ \frac{p(y|\theta, d)}{p(y|d)} \right] - \ln \left[ p(y|d) \right] \right\} p(y|\theta, d) p(\theta) \, d\theta \, dy
\]
Estimating the expected information gain

**Challenge:** No closed-form expression for expected information gain

- Nonlinear computational models
- Nontrivial priors and noise models
- Avoid large approximations (mean-field, Gaussian, etc.)

**Solution:** use a Monte Carlo estimate

\[
U_{\theta}(d) = \mathbb{E}_y \left[ D_{KL} [p(\theta|y, d) \| p(\theta)] \right] \\
= \int_Y \int_{\Theta} \log \left( \frac{p(\theta|y, d)}{p(\theta)} \right) p(\theta|y, d) \, d\theta \, p(y|d) \, dy \\
\approx \frac{1}{N} \sum_{i=1}^{N} \left\{ \ln \left[ p(y^{(i)}|\theta^{(i)}, d) \right] - \ln \left[ p(y^{(i)}|d) \right] \right\}, \quad \theta^{(i)} \sim p(\theta), \quad y^{(i)} \sim p(y|\theta^{(i)}, d)
\]
Estimating the expected information gain (importance sampling)

\[ U_\theta(d) \approx \frac{1}{N} \sum_{i=1}^{N} \left\{ \ln [p(y^{(i)}|\theta^{(i)}, d)] - \ln [p(y^{(i)}|d)] \right\}, \quad \theta^{(i)} \sim p(\theta), \quad y^{(i)} \sim p(y|\theta^{(i)}, d) \]

Importance sampling estimates for the conditional and marginal likelihood:

\[ p(y^{(i)}|\theta^{(i)}, d) = \int_H p(y^{(i)}|\theta^{(i)}, \eta, d) p(\eta|\theta^{(i)}, d) d\eta \]

\[ p(y^{(i)}|d) = \int_{\Theta} \int_H p(y^{(i)}|\theta, \eta, d) p(\theta, \eta|d) d\eta d\theta \]

Biasing distributions

- \( \eta^{(i,k)} \sim q^{(i)}_{\text{cond}} \)
- \( \theta^{(i,j)}, \eta^{(i,j)} \sim q^{(i)}_{\text{marg}} \)

Biasing weights

- \( w^{(i,k)}_{\text{cond}} = \frac{p(\eta^{(i,k)}|\theta^{(i)})}{q^{(i)}_{\text{cond}}(\eta^{(i,k)})} \)
- \( w^{(i,j)}_{\text{marg}} = \frac{p(\theta^{(i,j)}, \eta^{(i,j)})}{q^{(i)}_{\text{marg}}(\theta^{(i,j)}, \eta^{(i,j)})} \)
Estimating the expected information gain (importance sampling)

\[ U_\theta(d) \approx \frac{1}{N} \sum_{i=1}^{N} \left\{ \ln \left[ p(y^{(i)}|\theta^{(i)}, d) \right] - \ln \left[ p(y^{(i)}|d) \right] \right\}, \quad \theta^{(i)} \sim p(\theta) \]

\[ y^{(i)} \sim p(y|\theta^{(i)}, d) \]

Importance sampling estimates for the conditional and marginal likelihood:

\[ p(y^{(i)}|\theta^{(i)}, d) \approx \frac{1}{M_2} \sum_{k=1}^{M_2} p(y^{(i)}|\theta^{(i)}, \eta^{(i,k)}, d) w_{\text{cond}}^{(i,k)} \]

\[ p(y^{(i)}|d) \approx \frac{1}{M_1} \sum_{j=1}^{M_1} p(y^{(i)}|\theta^{(i,j)}, \eta^{(i,j)}, d) w_{\text{marg}}^{(i,j)} \]

Biasing distributions

- \( \eta^{(i,k)} \sim q_{\text{cond}}^{(i)} \)
- \( \theta^{(i,j)}, \eta^{(i,j)} \sim q_{\text{marg}}^{(i)} \)

Biasing weights

- \( w_{\text{cond}}^{(i,k)} = p(\eta^{(i,k)}|\theta^{(i)})/q_{\text{cond}}^{(i)}(\eta^{(i,k)}) \)
- \( w_{\text{marg}}^{(i,j)} = p(\theta^{(i,j)}, \eta^{(i,j)})/q_{\text{marg}}^{(i)}(\theta^{(i,j)}, \eta^{(i,j)}) \)
The computational problem in a nutshell:

▶ We must consider many realizations of the data $y^{(i)}$
▶ For each realization, we have a posterior distribution. Compute:
  ▶ Complete posterior normalizing constant (*marginal likelihood*)
  ▶ Partial posterior normalizing constant (*conditional likelihood*)
▶ Hence, *nested* Monte Carlo sampling

▶ Naïve approach: use prior as a biasing distribution [Ryan 2003]
  ▶ Can also use surrogates [Huan & M 2013] to allow very large sample sizes, but surrogate leaves an asymptotic bias
The computational problem in a nutshell:

- We must consider many realizations of the data $y^{(i)}$
- For each realization, we have a posterior distribution. Compute:
  - Complete posterior normalizing constant (*marginal likelihood*)
  - Partial posterior normalizing constant (*conditional likelihood*)
- Hence, *nested* Monte Carlo sampling

- Naïve approach: use prior as a biasing distribution [Ryan 2003]
  - Can also use surrogates [Huan & M 2013] to allow very large sample sizes, but surrogate leaves an asymptotic bias

- **How bad could prior sampling be?**
Sampling distribution of the naïve estimator

\[ \text{EIG estimator mean ± 95} \]

\[ N = M_1 = M_2 = 316 \rightarrow 10^5 \text{ samples at each design} \]

\[ \text{Bias is particularly misleading for the case of } \textbf{focused} \text{ design...} \]
Sampling distribution of the naïve estimator

▶ $N = M_1 = M_2 = 316 \rightarrow 10^5$ samples at each design
▶ Bias is particularly misleading for the case of focused design…
Estimator bias and variance

- Use $\Delta$–method to analyze bias and variance of EIG estimator

$$
\mathbb{E}[^\hat{U} (d)] \approx U(d) + \underbrace{\frac{A(d)}{M_1} - \frac{B(d)}{M_2}}_{\text{bias}}
$$

$$
\nabla[^\hat{U} (d)] \approx \frac{C(d)}{N} + \frac{D(d)}{NM_1} + \frac{E(d)}{NM_2}
$$

- $A(d), B(d), D(d), E(d) \sim \mathbb{E}_y[\text{variance of inner estimators}]$

- Need to find biasing distributions $q_{\text{marg}}^{(i)}$ and $q_{\text{cond}}^{(i)}$ to reduce variance
Estimator bias and variance

- Use $\Delta$-method to analyze bias and variance of EIG estimator

$$
\mathbb{E}[\hat{U}(d)] \approx U(d) + \frac{A(d)}{M_1} - \frac{B(d)}{M_2}
$$

$$
\nabla[\hat{U}(d)] \approx \frac{C(d)}{N} + \frac{D(d)}{NM_1} + \frac{E(d)}{NM_2}
$$

- $A(d), B(d), D(d), E(d) \sim \mathbb{E}_y[\text{variance of inner estimators}]$

- Need to find biasing distributions $q_{\text{marg}}^{(i)}$ and $q_{\text{cond}}^{(i)}$ to reduce variance

- **Idea:**
  - An individual normalizing constant is difficult to estimate. Collectively, the problem is easier!
  - Recycle existing samples/model evaluations to obtain increasingly better biasing distributions, as the outer loop proceeds
  - Preserve consistency of the estimator
Layered multiple importance sampling

{Prior samples}  {Samples from previous biasing distributions}

Select most ‘useful’ mixture components

‘Draw samples’ from mixture by recycling existing samples

Multiple Importance Sampling

Use mixture biasing distribution with self-normalized importance weights

\[ \hat{\mu}_{post}^{(i)}, \hat{\Sigma}_{post}^{(i)} \]

draw \( M_1 \) samples

\[ q_{marg}^{(i)} \equiv t_\nu(\hat{\mu}_{post}^{(i)}, \hat{\Sigma}_{post}^{(i)}) \]

\[ \mathcal{X}_{marg}^{(i)} \sim q_{marg}^{(i)} \]

save \( \mathcal{X}_{marg}^{(i)} \) for future iterations

Importance Sampling

\[ \hat{\rho}(y^{(i)} | d) \] marginal likelihood
Layered multiple importance sampling

Existing samples at iteration $i = 1$

- $\mathbf{\theta}^{(i)}, \eta^{(i)}$
- $p(\mathbf{\theta}, \eta | y^{(i)})$
- $\mathbf{\theta}, \eta \sim q_{\text{post}}^{(i)}$
- $\mathbf{\theta}, \eta \sim q_{\text{ML}}^{(i)}$

$p(\mathbf{\theta}, \eta | y^{(i)}) \sim q_{\text{post}}^{(i)}(\mathbf{\theta}, \eta)$

$q_{\text{post}}^{(i)}(\mathbf{\theta}, \eta) \sim q_{\text{ML}}^{(i)}(\mathbf{\theta}, \eta)$

$\mathbf{\theta}, \eta \sim q_{\text{ML}}^{(i)}(\mathbf{\theta}, \eta)$

$p(\mathbf{\theta}, \eta | y^{(i)})$
Layered multiple importance sampling

\[
(\theta^{(i)}, \eta^{(i)}) \quad \circ \quad q_{\text{post}}^{(i)}(\theta, \eta) \quad \circ \quad q_{\text{ML}}^{(i)}(\theta, \eta) \\
\circ \quad p(\theta, \eta|y^{(i)}) \quad \bullet \quad (\theta, \eta) \sim q_{\text{post}}^{(i)} \quad + \quad (\theta, \eta) \sim q_{\text{ML}}^{(i)}
\]

\[
(\theta^{(1)}, \eta^{(1)}) \quad \text{and resulting posterior } p(\theta, \eta|y^{(1)}, d)
\]
Layered multiple importance sampling

\[ \begin{align*}
\star & (\theta^{(i)}, \eta^{(i)}) & q_{\text{post}}^{(i)}(\theta, \eta) & q_{\text{ML}}^{(i)}(\theta, \eta) \\
\circ & p(\theta, \eta|y^{(i)}) & (\theta, \eta) \sim q_{\text{post}}^{(i)} & (\theta, \eta) \sim q_{\text{ML}}^{(i)}
\end{align*} \]

Mixture biasing distribution

\[
(\theta, \eta) \sim q_{\text{post}}^{(i)}
\]
Layered multiple importance sampling

\[
\begin{align*}
(\theta(i), \eta(i)) & \quad q_{post}^{(i)}(\theta, \eta) & \quad q_{ML}^{(i)}(\theta, \eta) \\
p(\theta, \eta|y(i)) & \quad (\theta, \eta) \sim q_{post}^{(i)} & \quad (\theta, \eta) \sim q_{ML}^{(i)}
\end{align*}
\]

Estimate posterior using importance sampling
Layered multiple importance sampling

Construct biasing distribution $q^{(1)}_{ML}(\theta, \eta)$
Layered multiple importance sampling

\[
\begin{align*}
\star (\theta^{(i)}, \eta^{(i)}) & \quad q_{\text{post}}^{(i)}(\theta, \eta) & \quad q_{\text{ML}}^{(i)}(\theta, \eta) \\
\circ \ p(\theta, \eta|y^{(i)}) & \quad (\theta, \eta) \sim q_{\text{post}}^{(i)} & \quad (\theta, \eta) \sim q_{\text{ML}}^{(i)}
\end{align*}
\]

Draw samples from \(q_{\text{ML}}^{(1)}(\theta, \eta)\) to estimate \(p(y^{(1)}|d)\)
Layered multiple importance sampling

\[
\begin{align*}
\star & \quad (\theta^{(i)}, \eta^{(i)}) \\
\bigcirc & \quad q^{(i)}_{\text{post}}(\theta, \eta) \\
\bigcirc & \quad p(\theta, \eta | y^{(i)}) \\
\bigcirc & \quad (\theta, \eta) \sim q^{(i)}_{\text{post}} \\
\bigcirc & \quad + \quad (\theta, \eta) \sim q^{(i)}_{\text{ML}}
\end{align*}
\]

Store samples for future iterations

\[
(\theta(i), \eta(i))
\]

\[
p(\theta, \eta | y(i))
\]

\[
q(i)_{\text{post}}
\]

\[
(\theta, \eta) \sim q(i)_{\text{post}}
\]

\[
q(i)_{\text{ML}}
\]

\[
(\theta, \eta) \sim q(i)_{\text{ML}}
\]