Expected information gain in marginals

Consider the marginal posterior and marginal prior

$$p(\boldsymbol{\theta}|\boldsymbol{y}, \boldsymbol{d}) = \int_H \underbrace{p(\boldsymbol{\theta}, \boldsymbol{\eta}|\boldsymbol{y}, \boldsymbol{d})}_{\text{joint posterior}} \, \mathrm{d}\boldsymbol{\eta}, \quad p(\boldsymbol{\theta}) = \int_H \underbrace{p(\boldsymbol{\theta}, \boldsymbol{\eta})}_{\text{joint prior}} \, \mathrm{d}\boldsymbol{\eta}$$

Information gain in $\boldsymbol{\theta}$ from a single observation $\boldsymbol{y}$:

$$u_{\boldsymbol{\theta}}(\boldsymbol{y}, \boldsymbol{d}) = \mathrm{D_{KL}}\left[\overbrace{p(\boldsymbol{\theta}|\boldsymbol{y}, \boldsymbol{d})}^{\substack{\text{marginal}\\\text{posterior}}} \,\Big\|\, \overbrace{p(\boldsymbol{\theta})}^{\substack{\text{marginal}\\\text{prior}}} \right]$$

Expected information gain – experimental outcome is unknown beforehand:

$$U_{\theta}(\boldsymbol{d}) = \mathbb{E}_{\boldsymbol{y}}[u_{\theta}(\boldsymbol{y}, \boldsymbol{d})] = \int_{\boldsymbol{Y}} u_{\theta}(\boldsymbol{y}, \boldsymbol{d}) \, p(\boldsymbol{y}|\boldsymbol{d}) \, \mathrm{d}\boldsymbol{y}$$

Expectation is taken over the prior predictive of the data

Expected information gain in marginals

Consider the marginal posterior and marginal prior

$$\color{blue}{p(\boldsymbol{\theta}|\boldsymbol{y}, \boldsymbol{d})} = \int_{\boldsymbol{H}} \underbrace{p(\boldsymbol{\theta}, \boldsymbol{\eta}|\boldsymbol{y}, \boldsymbol{d})}_{\text{joint posterior}} \, \mathrm{d}\boldsymbol{\eta}, \quad \color{red}{p(\boldsymbol{\theta})} = \int_{\boldsymbol{H}} \underbrace{p(\boldsymbol{\theta}, \boldsymbol{\eta})}_{\text{joint prior}} \, \mathrm{d}\boldsymbol{\eta}$$

Information gain in $\boldsymbol{\theta}$ from a single observation $\boldsymbol{y}$:

$$u_{\boldsymbol{\theta}}(\boldsymbol{y}, \boldsymbol{d}) = \mathrm{D}_{\mathrm{KL}}\left[\overbrace{\color{blue}{p(\boldsymbol{\theta}|\boldsymbol{y}, \boldsymbol{d})}}^{\substack{\text{marginal}\\\text{posterior}}} \, \| \, \overbrace{\color{red}{p(\boldsymbol{\theta})}}^{\substack{\text{marginal}\\\text{prior}}} \right]$$

Expected information gain – experimental outcome is unknown beforehand:

$$U_{\boldsymbol{\theta}}(\boldsymbol{d}) = \mathbb{E}_{\boldsymbol{y}}[u_{\boldsymbol{\theta}}(\boldsymbol{y}, \boldsymbol{d})] = \int_{\boldsymbol{Y}} u_{\boldsymbol{\theta}}(\boldsymbol{y}, \boldsymbol{d}) \, p(\boldsymbol{y}|\boldsymbol{d}) \, \mathrm{d}\boldsymbol{y}$$

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

$$
\begin{aligned}
U_\theta(\boldsymbol{d}) &= \mathbb{E}_{\boldsymbol{y}}\left[\mathrm{D}_{\mathrm{KL}}\left[p(\boldsymbol{\theta}|\boldsymbol{y},\boldsymbol{d}) \,\middle|\middle|\, p(\boldsymbol{\theta})\right]\right] \\
&= \int_{\mathcal{Y}}\int_{\Theta} \log\left[\frac{p(\boldsymbol{\theta}|\boldsymbol{y},\boldsymbol{d})}{p(\boldsymbol{\theta})}\right] p(\boldsymbol{\theta}|\boldsymbol{y},\boldsymbol{d})\,\mathrm{d}\boldsymbol{\theta}\, p(\boldsymbol{y}|\boldsymbol{d})\,\mathrm{d}\boldsymbol{y} \\
&= \int_{\mathcal{Y}}\int_{\Theta} \left\{ \ln\Big[\underbrace{p(\boldsymbol{y}|\boldsymbol{\theta},\boldsymbol{d})}_{\substack{\text{conditional}\\\text{likelihood}}}\Big] - \ln\Big[\underbrace{p(\boldsymbol{y}|\boldsymbol{d})}_{\substack{\text{marginal}\\\text{likelihood}}}\Big] \right\} p(\boldsymbol{y}|\boldsymbol{\theta},\boldsymbol{d})p(\boldsymbol{\theta})\mathrm{d}\boldsymbol{\theta}\mathrm{d}\boldsymbol{y}
\end{aligned}
$$

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

$$
\begin{aligned}
U_\theta(\boldsymbol{d}) &= \mathbb{E}_{\boldsymbol{y}}\left[\mathrm{D}_{\mathrm{KL}}\left[p(\boldsymbol{\theta}|\boldsymbol{y},\boldsymbol{d})\,\middle|\middle|\,p(\boldsymbol{\theta})\right]\right] \\
&= \int_Y \int_\Theta \log\left[\frac{p(\boldsymbol{\theta}|\boldsymbol{y},\boldsymbol{d})}{p(\boldsymbol{\theta})}\right] p(\boldsymbol{\theta}|\boldsymbol{y},\boldsymbol{d})\,\mathrm{d}\boldsymbol{\theta}\,p(\boldsymbol{y}|\boldsymbol{d})\,\mathrm{d}\boldsymbol{y} \\
&= \int_Y \int_\Theta \left\{\ln\Big[\underbrace{p(\boldsymbol{y}|\boldsymbol{\theta},\boldsymbol{d})}_{\substack{\text{conditional}\\ \text{likelihood}}}\Big] - \ln\Big[\underbrace{p(\boldsymbol{y}|\boldsymbol{d})}_{\substack{\text{marginal}\\ \text{likelihood}}}\Big]\right\} p(\boldsymbol{y}|\boldsymbol{\theta},\boldsymbol{d})p(\boldsymbol{\theta})\mathrm{d}\boldsymbol{\theta}\mathrm{d}\boldsymbol{y}
\end{aligned}
$$

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

$$
\begin{aligned}
U_\theta(\boldsymbol{d}) &= \mathbb{E}_{\boldsymbol{y}}\left[\mathrm{D}_{\mathrm{KL}}\left[p(\boldsymbol{\theta}|\boldsymbol{y},\boldsymbol{d})\,\middle|\middle|\,p(\boldsymbol{\theta})\right]\right] \\
&= \int_{\boldsymbol{Y}}\int_{\boldsymbol{\Theta}} \log\left[\frac{p(\boldsymbol{\theta}|\boldsymbol{y},\boldsymbol{d})}{p(\boldsymbol{\theta})}\right] p(\boldsymbol{\theta}|\boldsymbol{y},\boldsymbol{d})\,\mathrm{d}\boldsymbol{\theta}\,p(\boldsymbol{y}|\boldsymbol{d})\,\mathrm{d}\boldsymbol{y} \\
&= \int_{\boldsymbol{Y}}\int_{\boldsymbol{\Theta}} \left\{\ln\Big[\underbrace{p(\boldsymbol{y}|\boldsymbol{\theta},\boldsymbol{d})}_{\substack{\text{conditional}\\ \text{likelihood}}}\Big] - \ln\Big[\underbrace{p(\boldsymbol{y}|\boldsymbol{d})}_{\substack{\text{marginal}\\ \text{likelihood}}}\Big]\right\} p(\boldsymbol{y}|\boldsymbol{\theta},\boldsymbol{d})p(\boldsymbol{\theta})\mathrm{d}\boldsymbol{\theta}\mathrm{d}\boldsymbol{y}
\end{aligned}
$$

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

$$
\begin{aligned}
U_\theta(\boldsymbol{d}) &= \mathbb{E}_{\boldsymbol{y}}\left[\mathrm{D}_{\mathrm{KL}}\left[p(\boldsymbol{\theta}|\boldsymbol{y},\boldsymbol{d})\,\middle|\middle|\,p(\boldsymbol{\theta})\right]\right] \\
&= \int_{\boldsymbol{Y}}\int_{\boldsymbol{\Theta}} \log\left[\frac{p(\boldsymbol{\theta}|\boldsymbol{y},\boldsymbol{d})}{p(\boldsymbol{\theta})}\right] p(\boldsymbol{\theta}|\boldsymbol{y},\boldsymbol{d})\,\mathrm{d}\boldsymbol{\theta}\,p(\boldsymbol{y}|\boldsymbol{d})\,\mathrm{d}\boldsymbol{y} \\
&\approx \frac{1}{N}\sum_{i=1}^{N}\left\{\ln\Big[\underbrace{p(\boldsymbol{y}^{(i)}|\boldsymbol{\theta}^{(i)},\boldsymbol{d})}_{\text{conditional likelihood}}\Big] - \ln\Big[\underbrace{p(\boldsymbol{y}^{(i)}|\boldsymbol{d})}_{\text{marginal likelihood}}\Big]\right\}, \qquad \begin{aligned}\boldsymbol{\theta}^{(i)} &\sim p(\boldsymbol{\theta}) \\ \boldsymbol{y}^{(i)} &\sim p(\boldsymbol{y}|\boldsymbol{\theta}^{(i)},\boldsymbol{d})\end{aligned}
\end{aligned}
$$

Estimating the expected information gain (importance sampling)

$$U_\theta(\boldsymbol{d}) \approx \frac{1}{N}\sum_{i=1}^{N}\left\{\ln\Big[\underbrace{p(\boldsymbol{y}^{(i)}|\boldsymbol{\theta}^{(i)},\boldsymbol{d})}_{\substack{\text{conditional}\\ \text{likelihood}}}\Big] - \ln\Big[\underbrace{p(\boldsymbol{y}^{(i)}|\boldsymbol{d})}_{\substack{\text{marginal}\\ \text{likelihood}}}\Big]\right\}, \qquad \begin{array}{l}\boldsymbol{\theta}^{(i)} \sim p(\boldsymbol{\theta})\\ \boldsymbol{y}^{(i)} \sim p(\boldsymbol{y}|\boldsymbol{\theta}^{(i)},\boldsymbol{d})\end{array}$$

Importance sampling estimates for the conditional and marginal likelihood:

$$\underbrace{p(\boldsymbol{y}^{(i)}|\boldsymbol{\theta}^{(i)},\boldsymbol{d})}_{\substack{\text{conditional}\\ \text{likelihood}}} = \int_{\boldsymbol{H}} p(\boldsymbol{y}^{(i)}|\boldsymbol{\theta}^{(i)},\boldsymbol{\eta},\boldsymbol{d})\,p(\boldsymbol{\eta}|\boldsymbol{\theta}^{(i)},\boldsymbol{d})\,\mathrm{d}\boldsymbol{\eta}$$

$$\underbrace{p(\boldsymbol{y}^{(i)}|\boldsymbol{d})}_{\substack{\text{marginal}\\ \text{likelihood}}} = \int_{\boldsymbol{\Theta}}\int_{\boldsymbol{H}} p(\boldsymbol{y}^{(i)}|\boldsymbol{\theta},\boldsymbol{\eta},\boldsymbol{d})\,p(\boldsymbol{\theta},\boldsymbol{\eta}|\boldsymbol{d})\,\mathrm{d}\boldsymbol{\eta}\,\mathrm{d}\boldsymbol{\theta}$$

Biasing distributions

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

Biasing weights

- $w_{\text{cond}}^{(i,k)} = p(\boldsymbol{\eta}^{(i,k)}|\boldsymbol{\theta}^{(i)})/q_{\text{cond}}^{(i)}(\boldsymbol{\eta}^{(i,k)})$
- $w_{\text{marg}}^{(i,j)} = p(\boldsymbol{\theta}^{(i,j)},\boldsymbol{\eta}^{(i,j)})/q_{\text{marg}}^{(i)}(\boldsymbol{\theta}^{(i,j)},\boldsymbol{\eta}^{(i,j)})$

Estimating the expected information gain (importance sampling)

$$U_\theta(\boldsymbol{d}) \approx \frac{1}{N}\sum_{i=1}^{N}\Big\{\ln\big[\underbrace{p(\boldsymbol{y}^{(i)}|\boldsymbol{\theta}^{(i)},\boldsymbol{d})}_{\text{conditional likelihood}}\big] - \ln\big[\underbrace{p(\boldsymbol{y}^{(i)}|\boldsymbol{d})}_{\text{marginal likelihood}}\big]\Big\}, \qquad \begin{array}{l}\boldsymbol{\theta}^{(i)} \sim p(\boldsymbol{\theta}) \\ \boldsymbol{y}^{(i)} \sim p(\boldsymbol{y}|\boldsymbol{\theta}^{(i)},\boldsymbol{d})\end{array}$$

Importance sampling estimates for the conditional and marginal likelihood:

$$\underbrace{p(\boldsymbol{y}^{(i)}|\boldsymbol{\theta}^{(i)},\boldsymbol{d})}_{\text{conditional likelihood}} \approx \frac{1}{M_2}\sum_{k=1}^{M_2} p(\boldsymbol{y}^{(i)}|\boldsymbol{\theta}^{(i)},\boldsymbol{\eta}^{(i,k)},\boldsymbol{d})\, w_{\text{cond}}^{(i,k)}$$

$$\underbrace{p(\boldsymbol{y}^{(i)}|\boldsymbol{d})}_{\text{marginal likelihood}} \approx \frac{1}{M_1}\sum_{j=1}^{M_1} p(\boldsymbol{y}^{(i)}|\boldsymbol{\theta}^{(i,j)},\boldsymbol{\eta}^{(i,j)},\boldsymbol{d})\, w_{\text{marg}}^{(i,j)}$$

Biasing distributions

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

Biasing weights

- $w_{\text{cond}}^{(i,k)} = p(\boldsymbol{\eta}^{(i,k)}|\boldsymbol{\theta}^{(i)}) / q_{\text{cond}}^{(i)}(\boldsymbol{\eta}^{(i,k)})$
- $w_{\text{marg}}^{(i,j)} = p(\boldsymbol{\theta}^{(i,j)},\boldsymbol{\eta}^{(i,j)}) / q_{\text{marg}}^{(i)}(\boldsymbol{\theta}^{(i,j)},\boldsymbol{\eta}^{(i,j)})$

Computational setup

The computational problem in a nutshell:

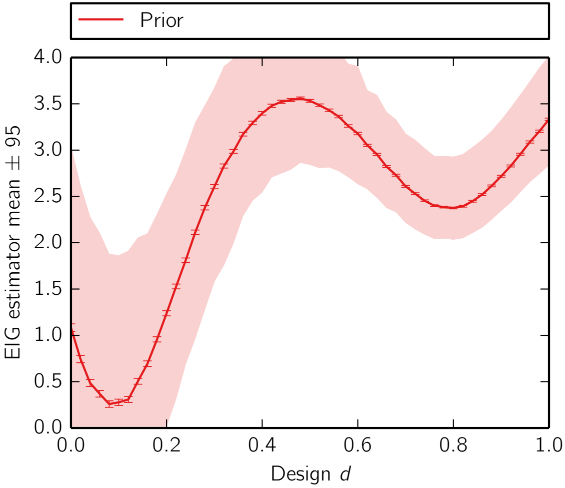
- We must consider many realizations of the data $\boldsymbol{y}^{(i)}$
- For each realization, we have a posterior distribution. Compute:
  - Complete posterior normalizing constant *(marginal likelihood)*
  - Partial posterior normalizing constant *(conditional likelihood)*
- Hence, *nested* Monte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

Computational setup

The computational problem in a nutshell:

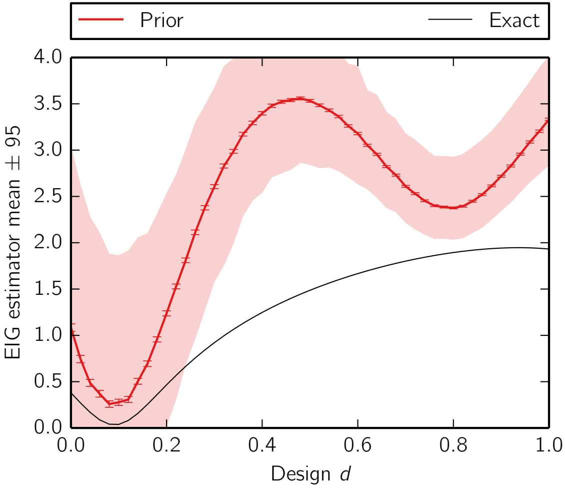
- We must consider many realizations of the data $\boldsymbol{y}^{(i)}$
- For each realization, we have a posterior distribution. Compute:
  - Complete posterior normalizing constant *(marginal likelihood)*
  - Partial posterior normalizing constant *(conditional likelihood)*
- Hence, *nested* Monte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



- $N = M_1 = M_2 = 316 \quad \rightarrow \quad 10^5$ samples at each design
- Bias is particularly misleading for the case of **focused** design...

Sampling distribution of the naïve estimator



- $N = M_1 = M_2 = 316 \quad \rightarrow \quad 10^5$ samples at each design
- Bias is particularly misleading for the case of **focused** design...

Estimator bias and variance

- Use Δ–method to analyze bias and variance of EIG estimator

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

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

  - $A(\boldsymbol{d}), B(\boldsymbol{d}), D(\boldsymbol{d}), E(\boldsymbol{d}) \sim \mathbb{E}_{\boldsymbol{y}}$[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 Δ–method to analyze bias and variance of EIG estimator

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

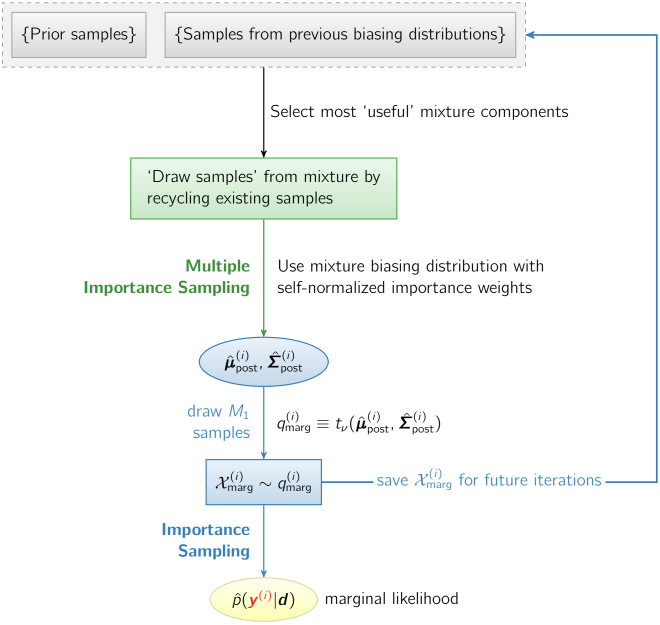
$$\mathbb{V}[\hat{U}(\boldsymbol{d})] \approx \frac{C(\boldsymbol{d})}{N} + \frac{D(\boldsymbol{d})}{NM_1} + \frac{E(\boldsymbol{d})}{NM_2}$$

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

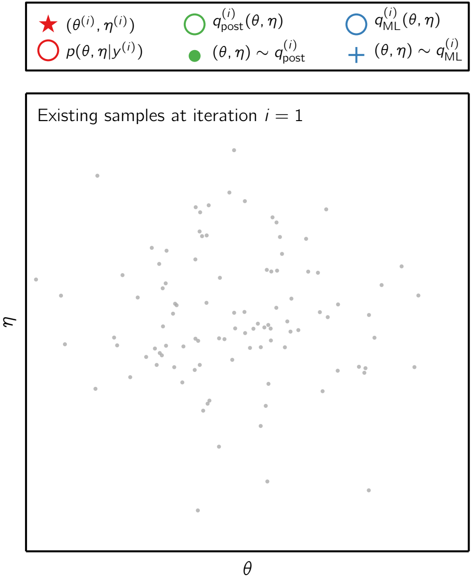
- Need to find biasing distributions $q^{(i)}_{\text{marg}}$ and $q^{(i)}_{\text{cond}}$ to reduce variance

- **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

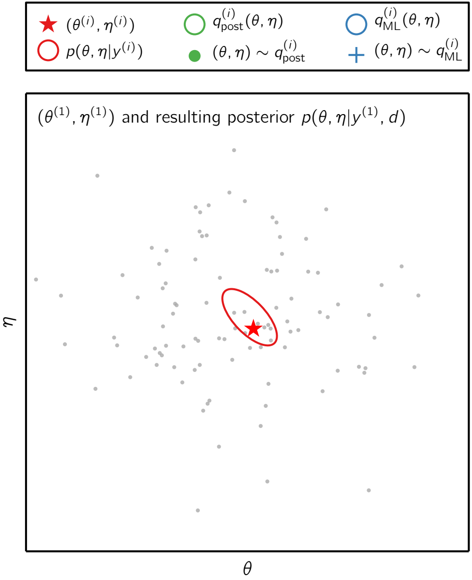
Layered multiple importance sampling



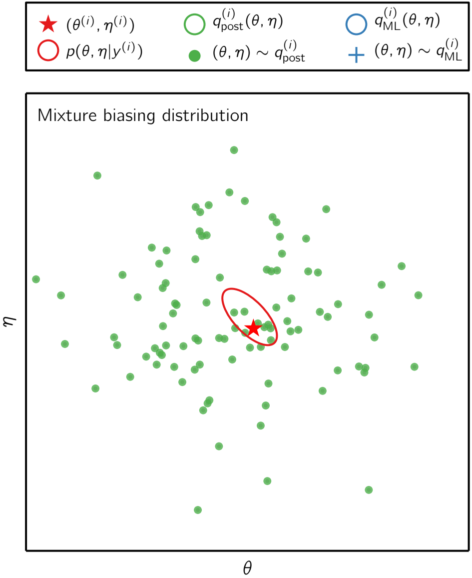
Layered multiple importance sampling



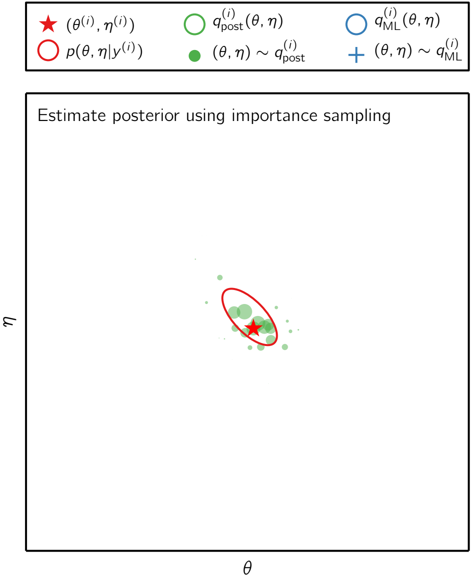
Layered multiple importance sampling



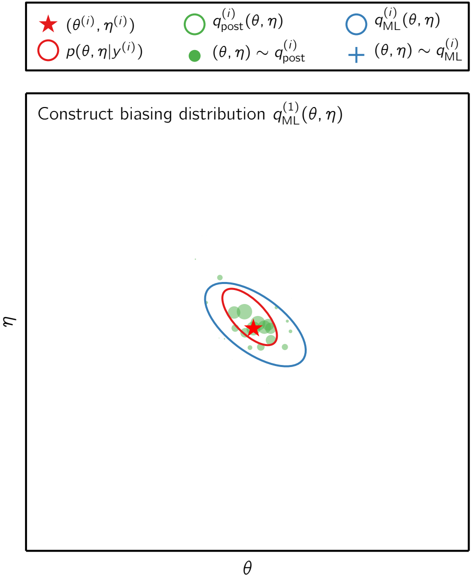
Layered multiple importance sampling



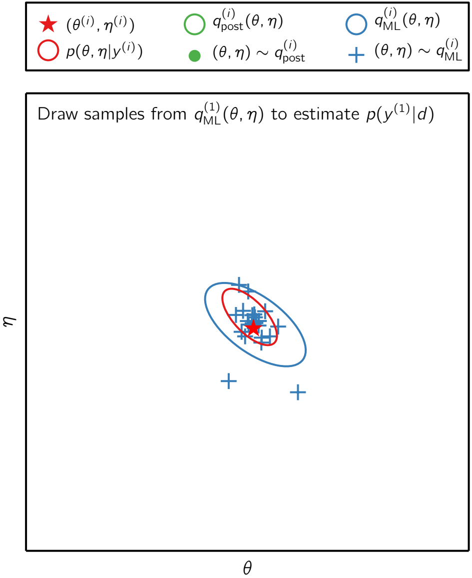
Layered multiple importance sampling



Layered multiple importance sampling



Layered multiple importance sampling



Layered multiple importance sampling