Amortized Monte Carlo Integration

Adam Goliński*, Frank Wood, Tom Rainforth*

11/06/19

UNIVERSITY OF OXFORD

$$\mathbb{E}_{p(x|y)}[f(x;\theta)]$$

Adam Goliński*, Frank Wood, Tom Rainforth*

$$\mathbb{E}_{p(x|y)}[f(x;\theta)]$$

AMCI = novel estimator +

$$\mathbb{E}_{p(x|y)}[f(x;\theta)]$$

AMCI = novel estimator + amortization objectives

UNIVERSITY OF OXFORD

Importance Sampling (IS)

$$\mathbb{E}_{\pi(x)}[f(x)] \approx \widehat{\mu} := \frac{1}{N}\sum_{n=1}^{N} f\left(x_n\right) w_n$$

$$\text{where } x_n \sim q(x), w_n = \frac{\pi\left(x_n\right)}{q\left(x_n\right)}$$

$$\mathbb{E}[\widehat{\mu}] = \mathbb{E}_{\pi(x)}[f(x)]$$

$$\text{Var}[\widehat{\mu}] = \frac{\text{Var}[f(x_1)w_1]}{N}$$

UNIVERSITY OF OXFORD

Importance Sampling (IS)

When $f(x) \geq 0$ and $q(x) \propto \pi(x) f(x)$

yields an exact estimate using a single sample

Importance Sampling (IS)

When $f(x) \geq 0$ and $q(x) \propto \pi(x) f(x)$

yields an exact estimate using a single sample

$$q^*(x) \propto f(x)\pi(x)$$

$$\implies q^*(x) = \frac{f(x)\pi(x)}{\int f(x)\pi(x)\mathrm{d}x} = \frac{f(x)\pi(x)}{\mathbb{E}_{\pi(x)}[f(x)]}$$

$$f(x_1)w_1 = f(x_1)\frac{\pi(x_1)}{q^*(x_1)} = f(x_1)\frac{\pi(x_1)}{\frac{f(x_1)\pi(x_1)}{\mathbb{E}_{\pi(x)}[f(x)]}} = \mathbb{E}_{\pi(x)}[f(x)]$$

Self-Normalized Importance Sampling (SNIS)

$$\mathbb{E}_{p(x|y)}[f(x)]$$

UNIVERSITY OF OXFORD

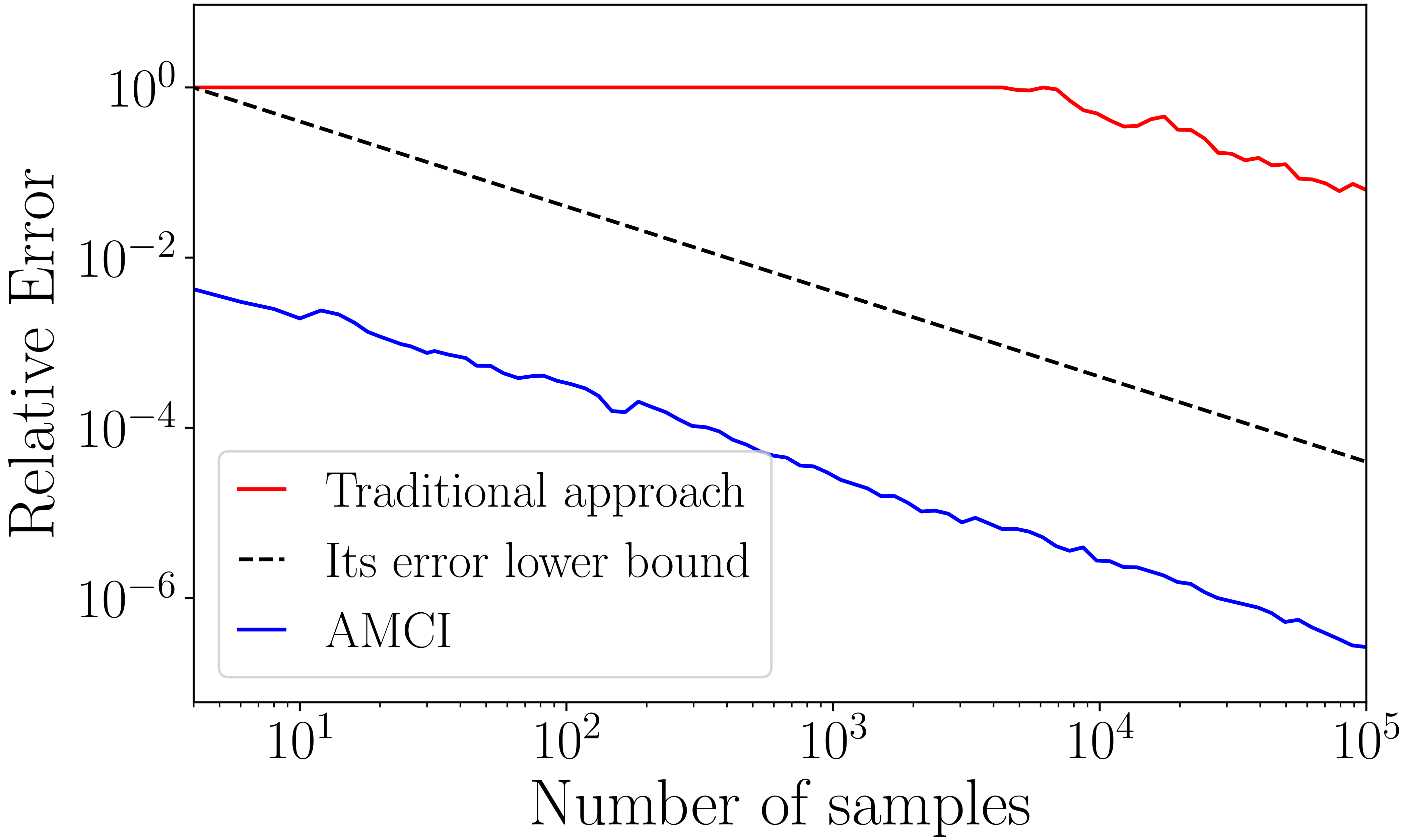
Self-Normalized Importance Sampling (SNIS)

$$\mathbb{E}_{p(x|y)}[f(x)] = \frac{\mathbb{E}_{p(x)}[f(x)p(y|x)]}{\mathbb{E}_{p(x)}[p(y|x)]}$$

$$\approx \frac{\frac{1}{N}\sum_{n=1}^{N} f(x_n)\, w_n}{\frac{1}{N}\sum_{n=1}^{N} w_n}$$

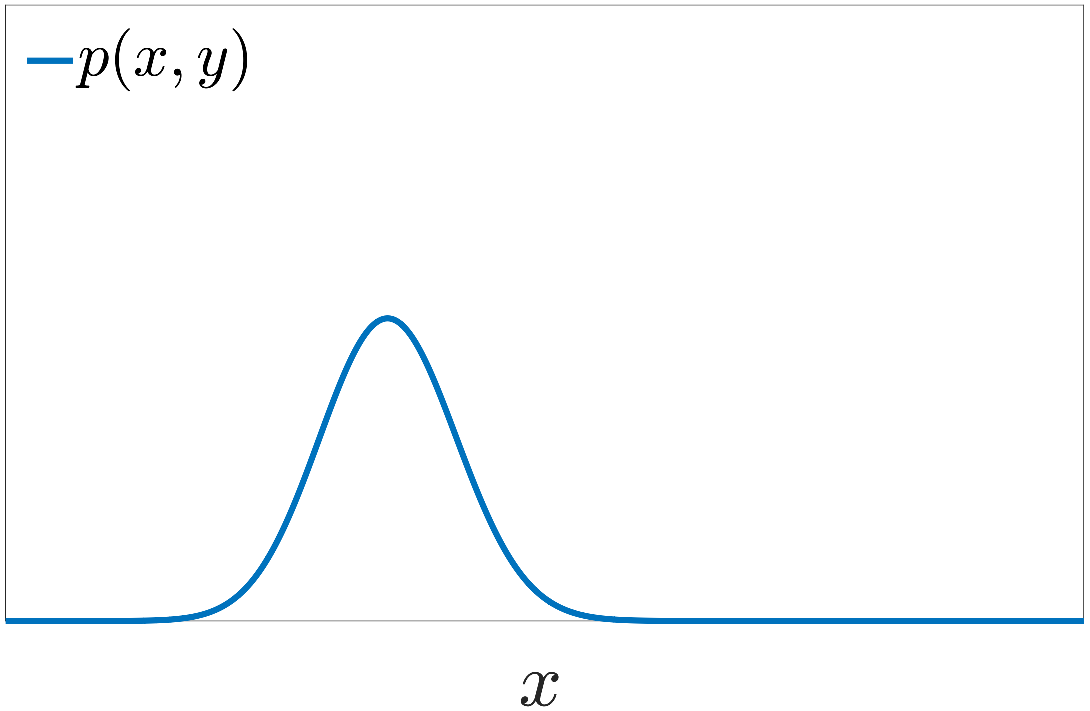
$$x_n \sim q(x)$$

$$w_n = \frac{p(x_n, y)}{q(x_n)}$$

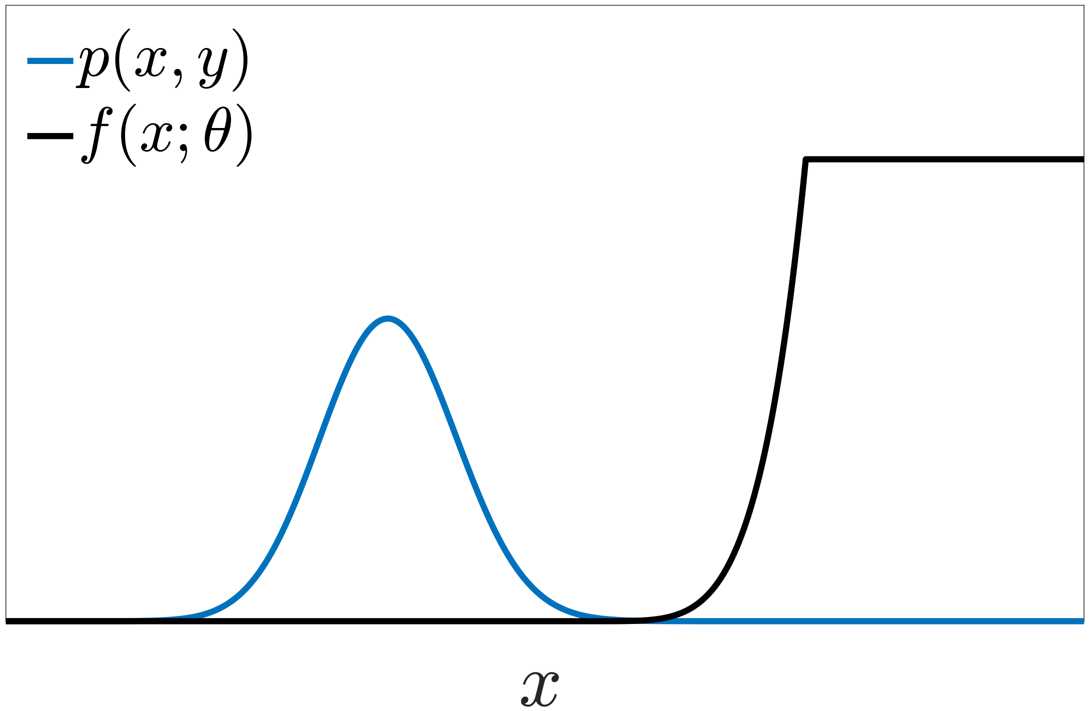

Relative Error
10^{0}
10^{-2}
10^{-4}
10^{-6}
Traditional approach
Its error lower bound
AMCI
10^{1}
10^{2}
10^{3}
10^{4}
10^{5}
Number of samples

$$\mathbb{E}_{p(x|y)}[f(x)] = \frac{\mathbb{E}_{p(x)}[f(x)p(y|x)]}{\mathbb{E}_{p(x)}[p(y|x)]}$$

$$\approx \frac{\frac{1}{N}\sum_{n=1}^{N} f\left(x_n\right) w_n}{\frac{1}{N}\sum_{n=1}^{N} w_n}$$

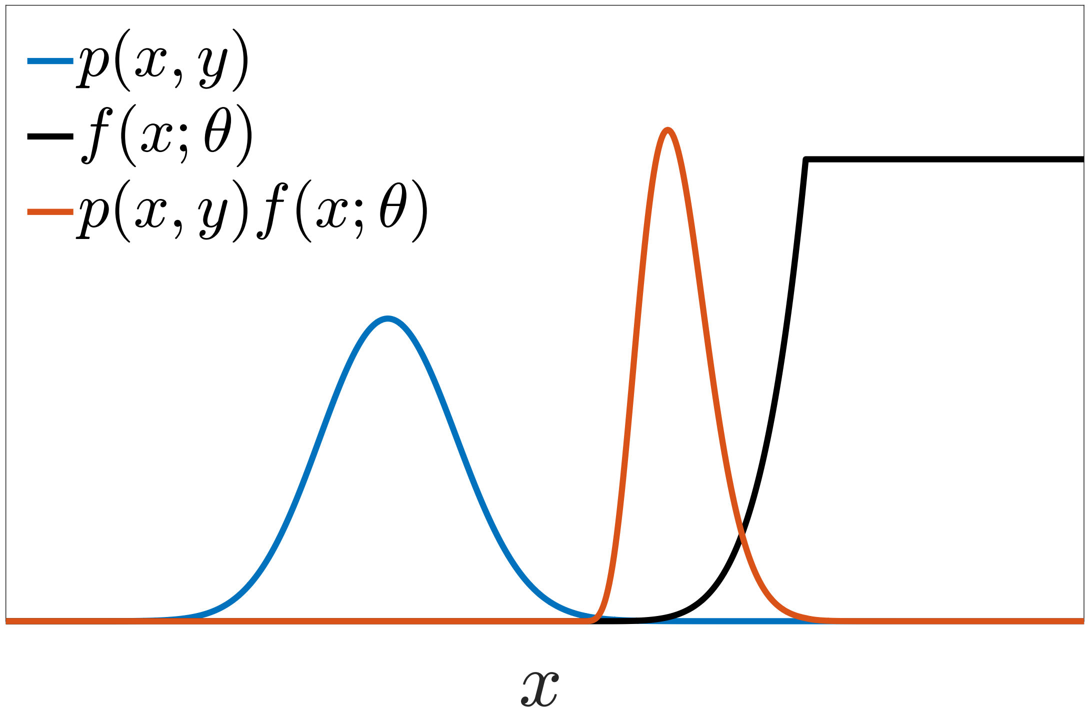


$$\mathbb{E}_{p(x|y)}[f(x)] = \frac{\mathbb{E}_{p(x)}[f(x)p(y|x)]}{\mathbb{E}_{p(x)}[p(y|x)]}$$
$$\approx \frac{\frac{1}{N}\sum_{n=1}^{N} f(x_n)\, w_n}{\frac{1}{N}\sum_{n=1}^{N} w_n}$$



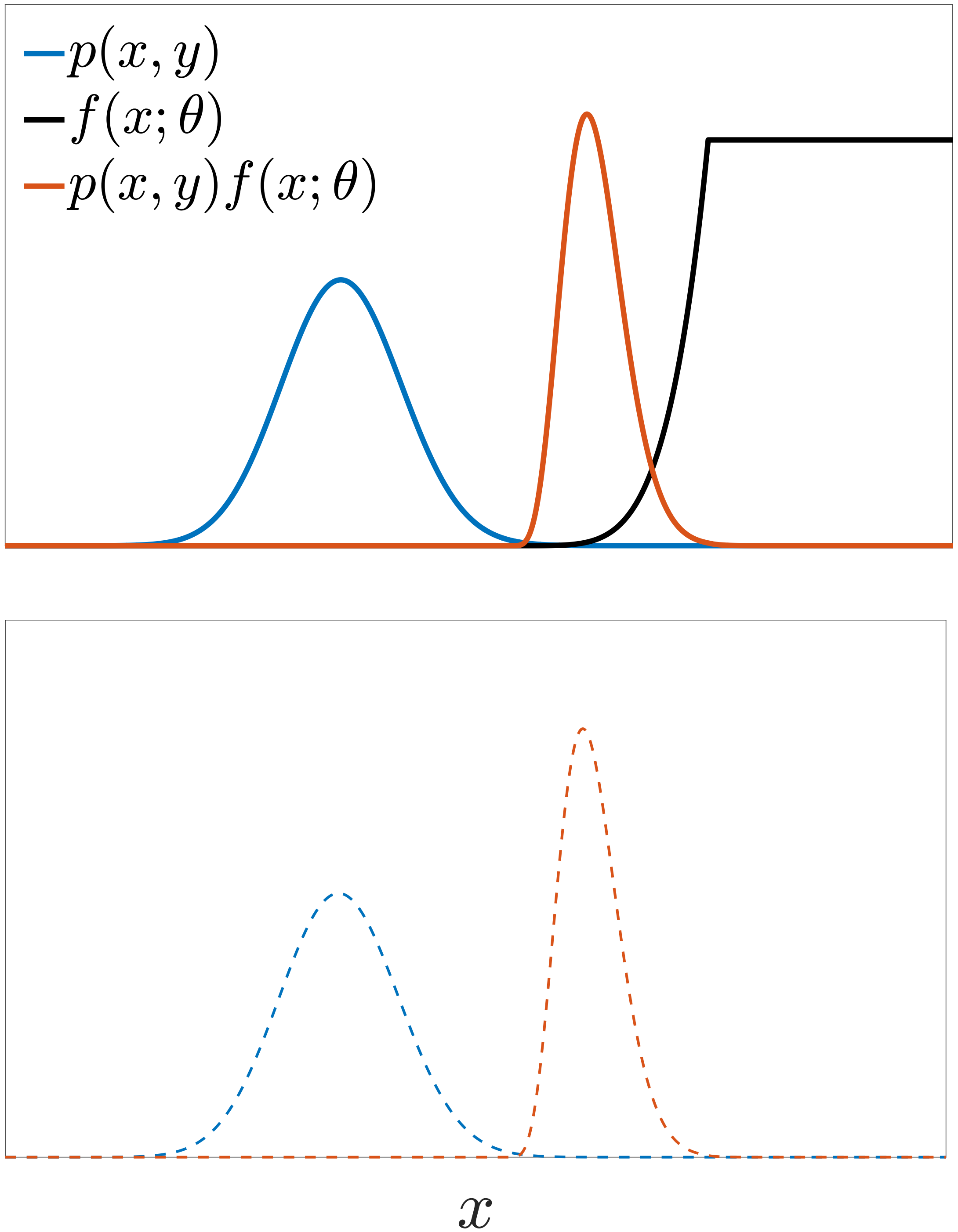
$$\mathbb{E}_{p(x|y)}[f(x)] = \frac{\mathbb{E}_{p(x)}[f(x)p(y|x)]}{\mathbb{E}_{p(x)}[p(y|x)]}$$

$$\approx \frac{\frac{1}{N}\sum_{n=1}^{N} f(x_n)\, w_n}{\frac{1}{N}\sum_{n=1}^{N} w_n}$$

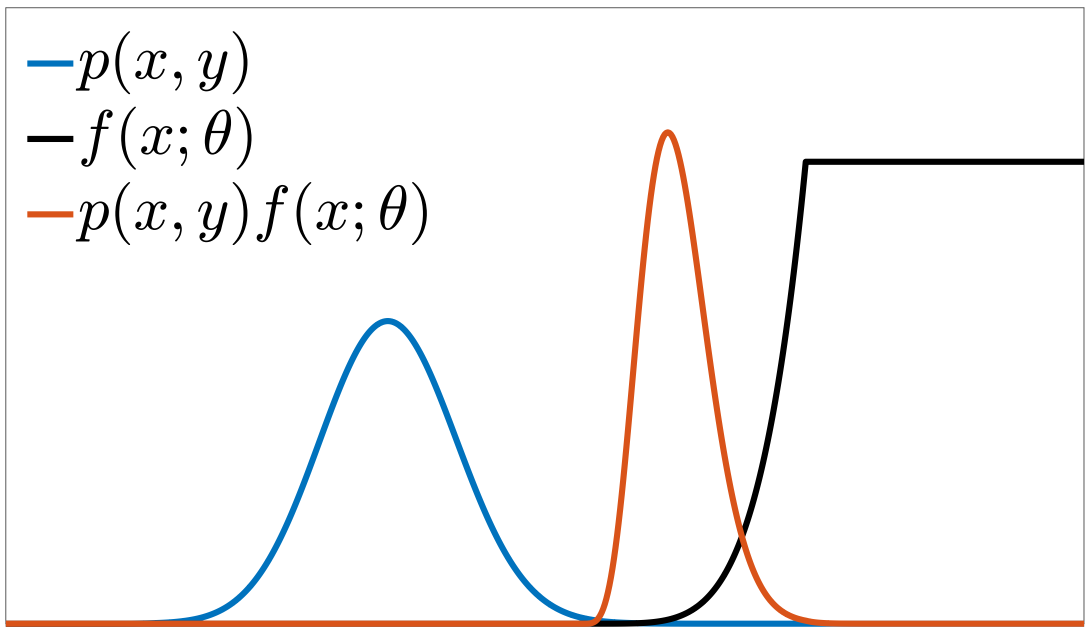


$$\mathbb{E}_{p(x|y)}[f(x)] = \frac{\mathbb{E}_{p(x)}[f(x)p(y|x)]}{\mathbb{E}_{p(x)}[p(y|x)]}$$

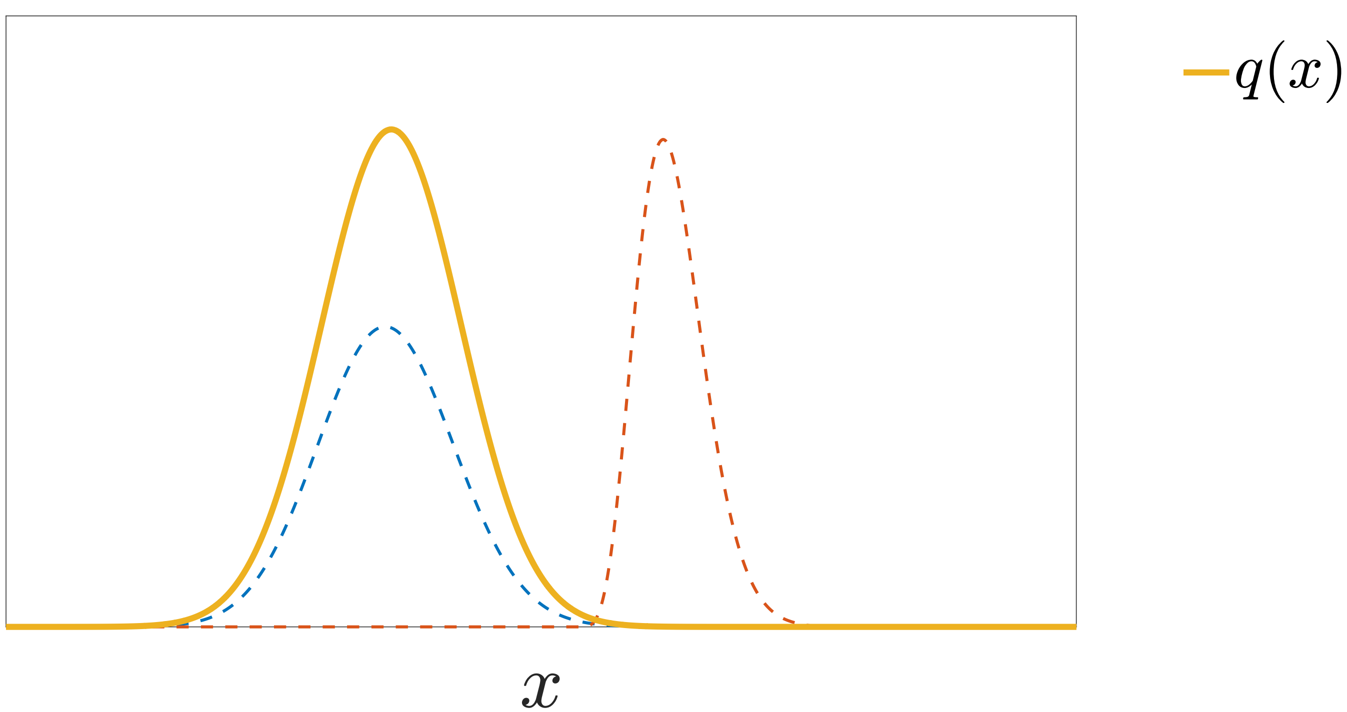
$$\approx \frac{\frac{1}{N}\sum_{n=1}^{N} f(x_n)\, w_n}{\frac{1}{N}\sum_{n=1}^{N} w_n}$$

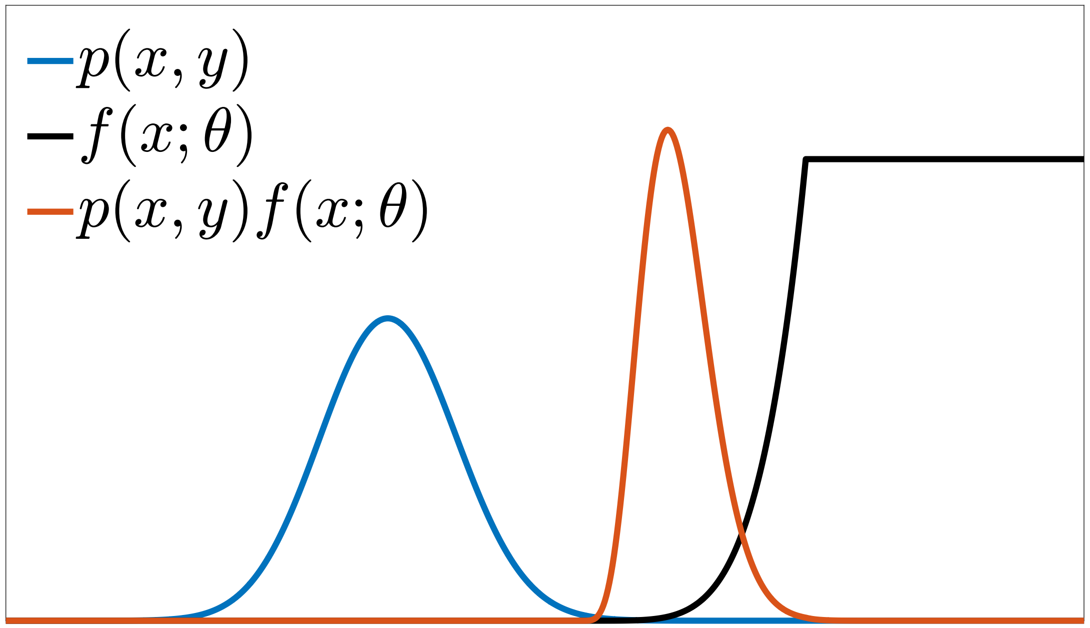


$$\mathbb{E}_{p(x|y)}[f(x)] = \frac{\mathbb{E}_{p(x)}[f(x)p(y|x)]}{\mathbb{E}_{p(x)}[p(y|x)]}$$
$$\approx \frac{\frac{1}{N}\sum_{n=1}^{N} f(x_n)\, w_n}{\frac{1}{N}\sum_{n=1}^{N} w_n}$$

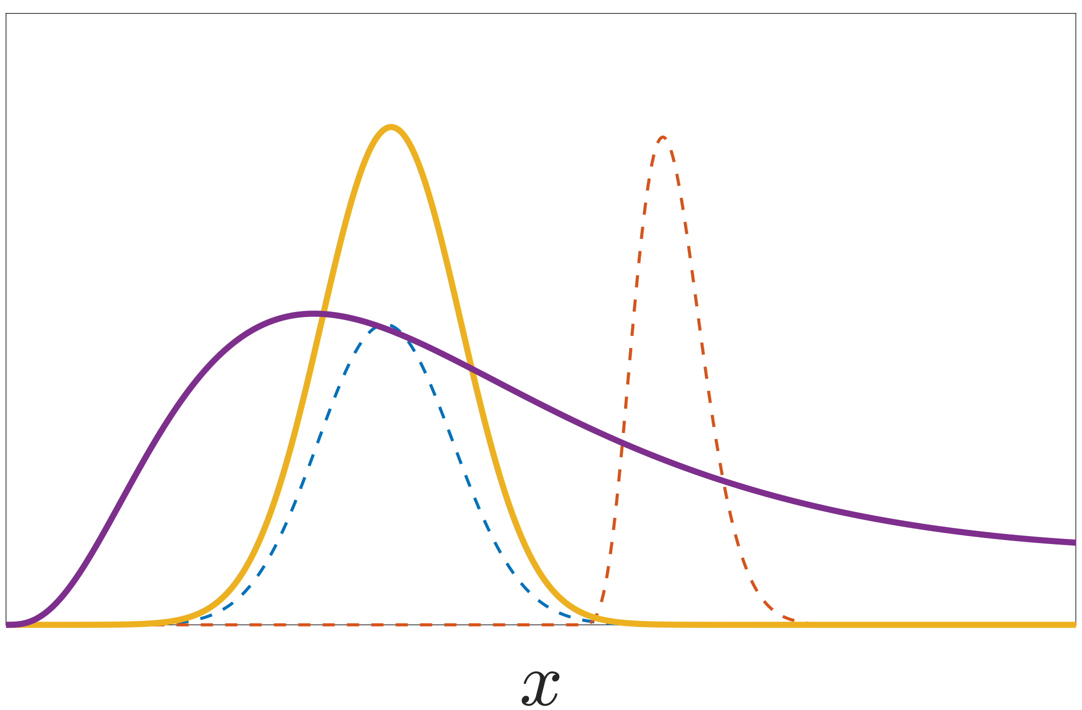


$$\mathbb{E}_{p(x|y)}[f(x)] = \frac{\mathbb{E}_{p(x)}[f(x)p(y|x)]}{\mathbb{E}_{p(x)}[p(y|x)]}$$
$$\approx \frac{\frac{1}{N}\sum_{n=1}^{N} f(x_n)\, w_n}{\frac{1}{N}\sum_{n=1}^{N} w_n}$$



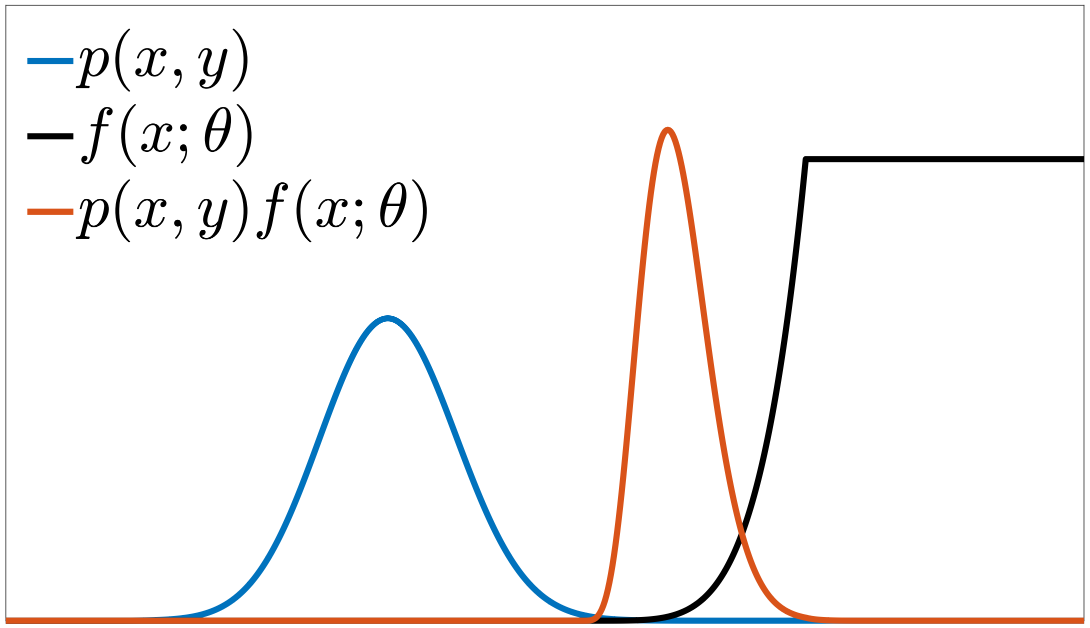


$$\mathbb{E}_{p(x|y)}[f(x)] = \frac{\mathbb{E}_{p(x)}[f(x)p(y|x)]}{\mathbb{E}_{p(x)}[p(y|x)]} \approx \frac{\frac{1}{N}\sum_{n=1}^{N} f(x_n)\, w_n}{\frac{1}{N}\sum_{n=1}^{N} w_n}$$



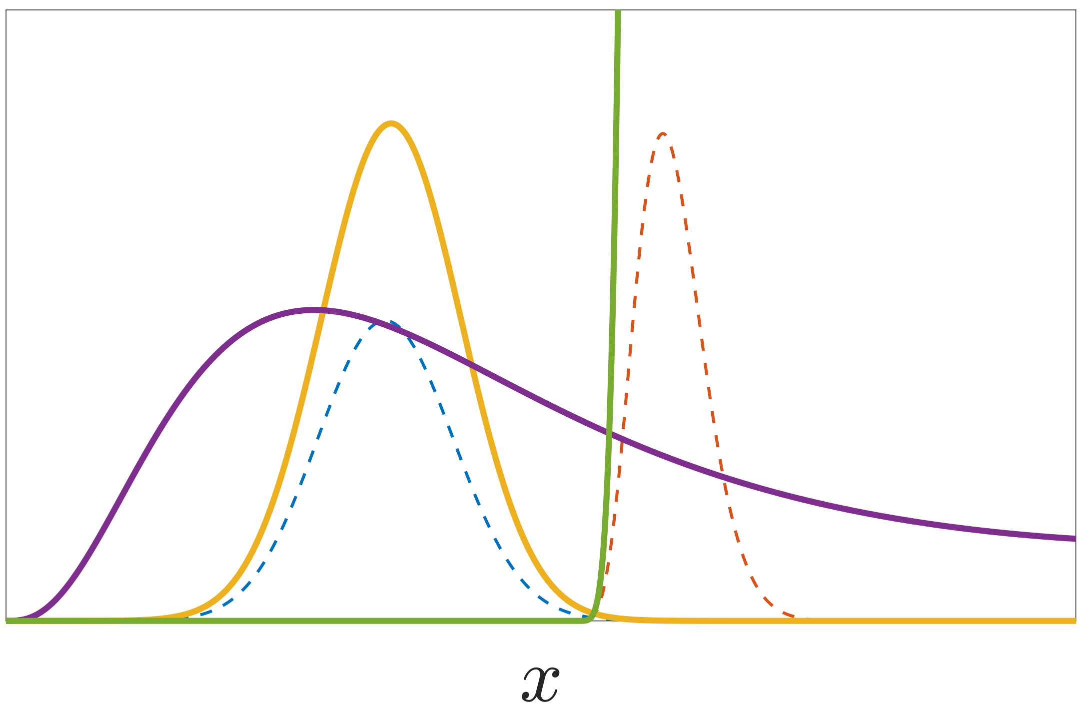
$q(x)$

$w(x) = \frac{p(x,y)}{q(x)}$



$$\mathbb{E}_{p(x|y)}[f(x)] = \frac{\mathbb{E}_{p(x)}[f(x)p(y|x)]}{\mathbb{E}_{p(x)}[p(y|x)]}$$

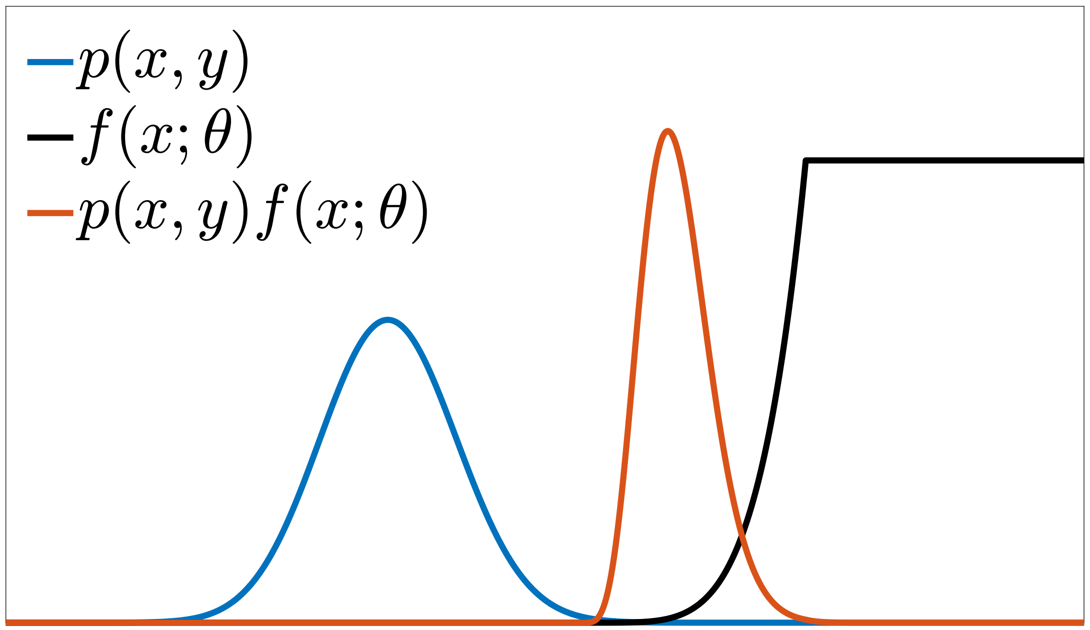
$$\approx \frac{\frac{1}{N}\sum_{n=1}^{N} f(x_n)\, w_n}{\frac{1}{N}\sum_{n=1}^{N} w_n}$$



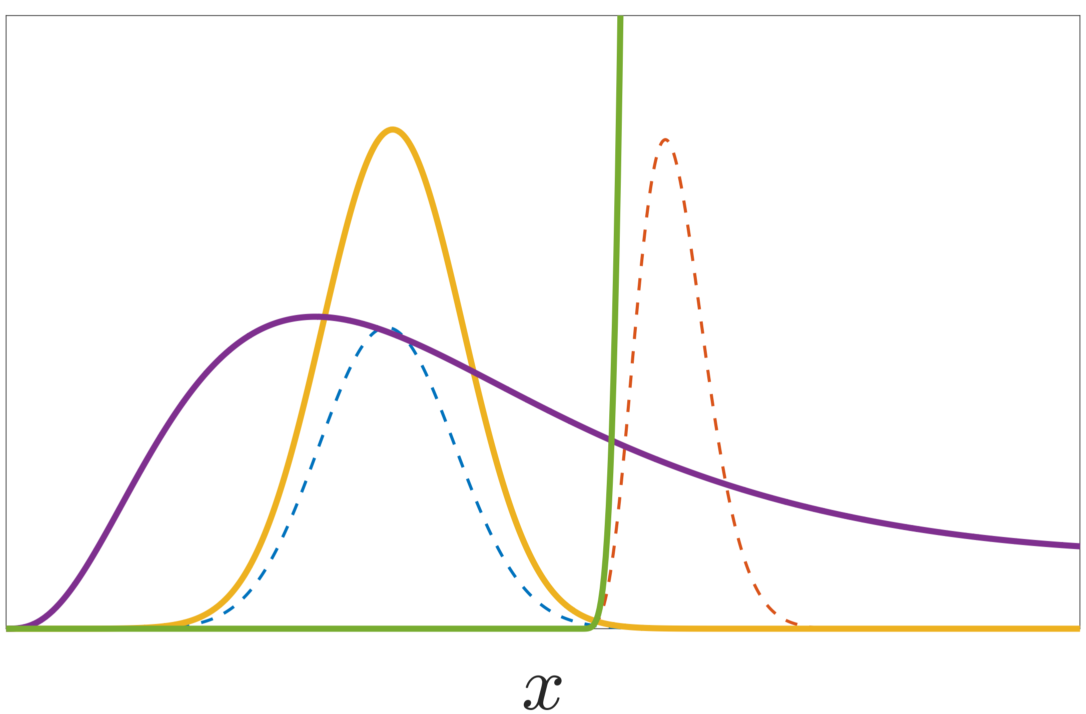
$q(x)$

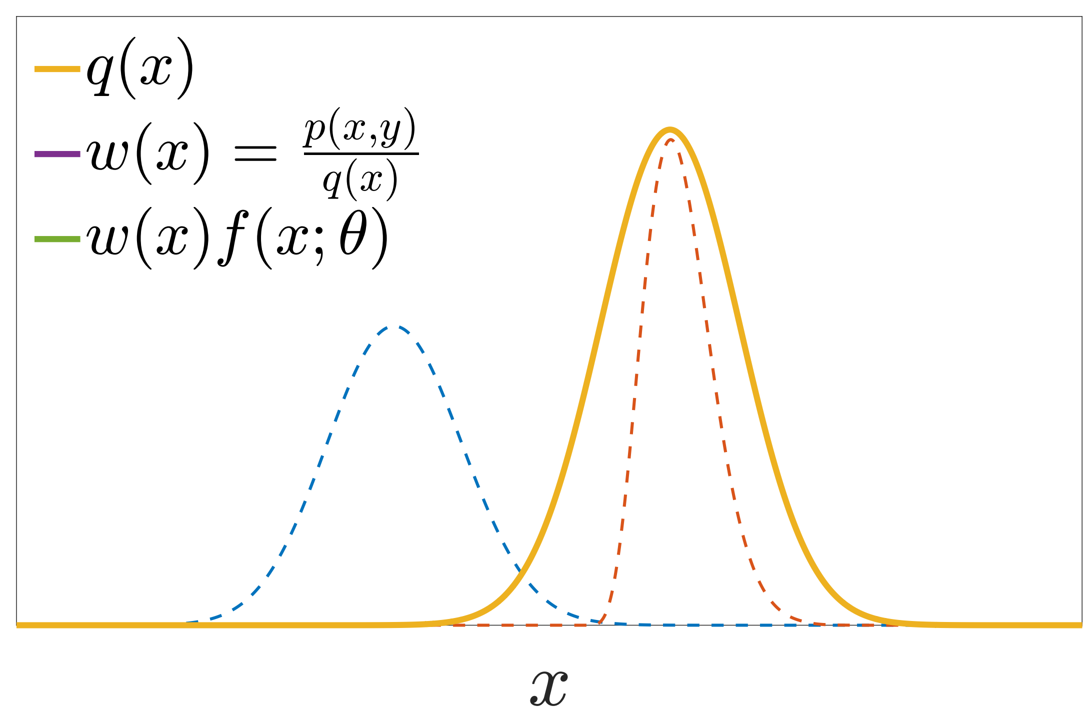
$w(x) = \frac{p(x,y)}{q(x)}$

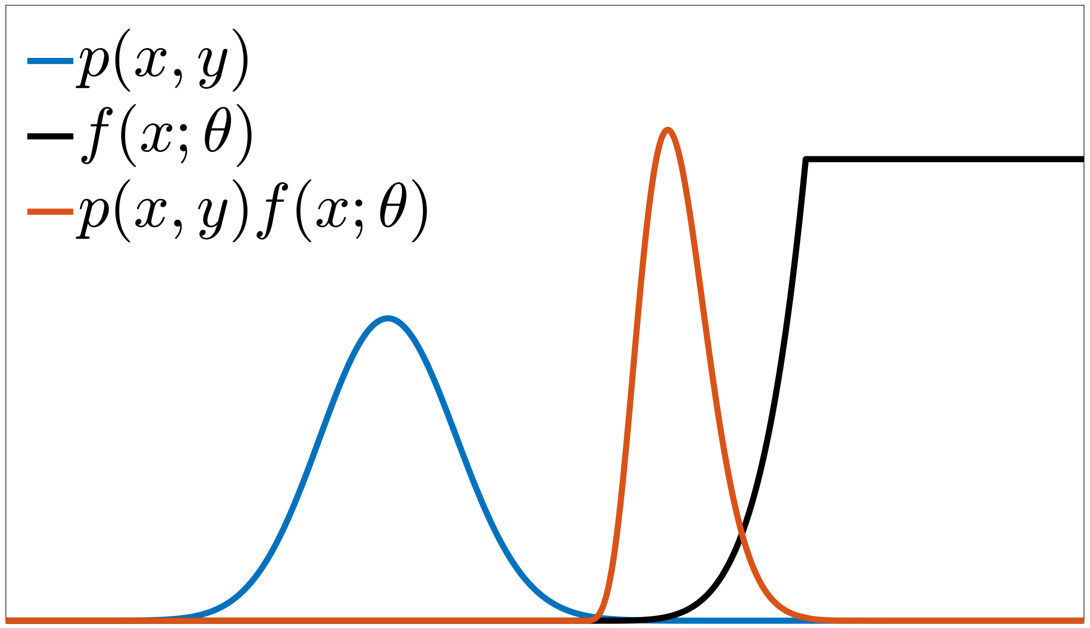
$w(x) f(x; \theta)$



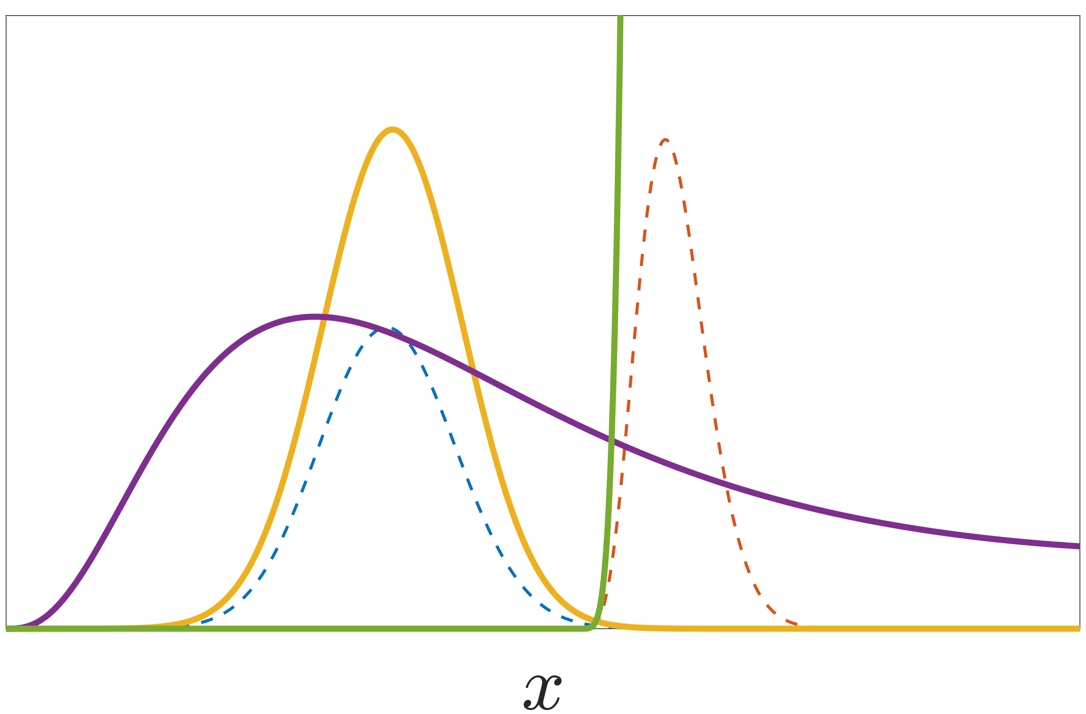
$$\mathbb{E}_{p(x|y)}[f(x)] = \frac{\mathbb{E}_{p(x)}[f(x)p(y|x)]}{\mathbb{E}_{p(x)}[p(y|x)]} \approx \frac{\frac{1}{N}\sum_{n=1}^{N} f(x_n)\, w_n}{\frac{1}{N}\sum_{n=1}^{N} w_n}$$

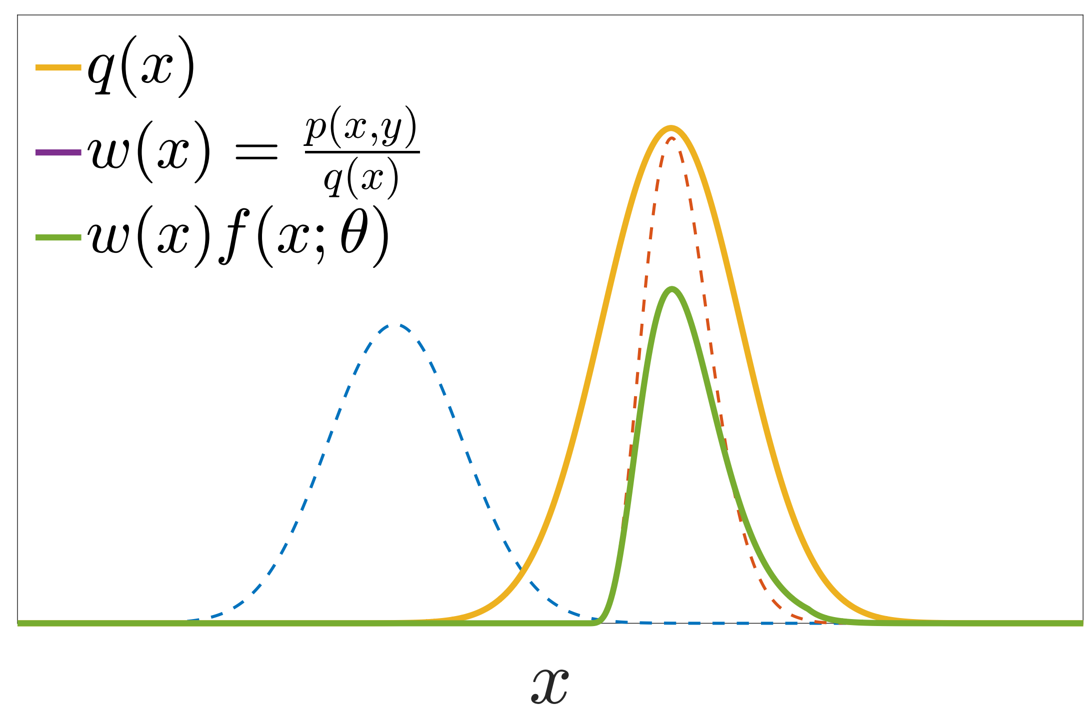


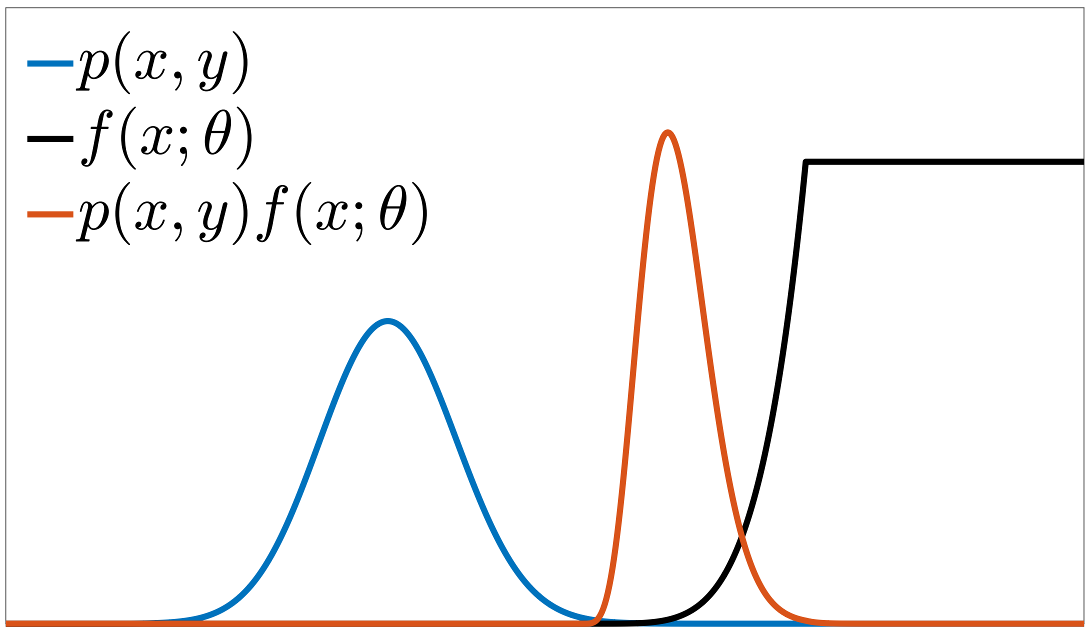




$$\mathbb{E}_{p(x|y)}[f(x)] = \frac{\mathbb{E}_{p(x)}[f(x)p(y|x)]}{\mathbb{E}_{p(x)}[p(y|x)]} \approx \frac{\frac{1}{N}\sum_{n=1}^{N} f(x_n)\, w_n}{\frac{1}{N}\sum_{n=1}^{N} w_n}$$

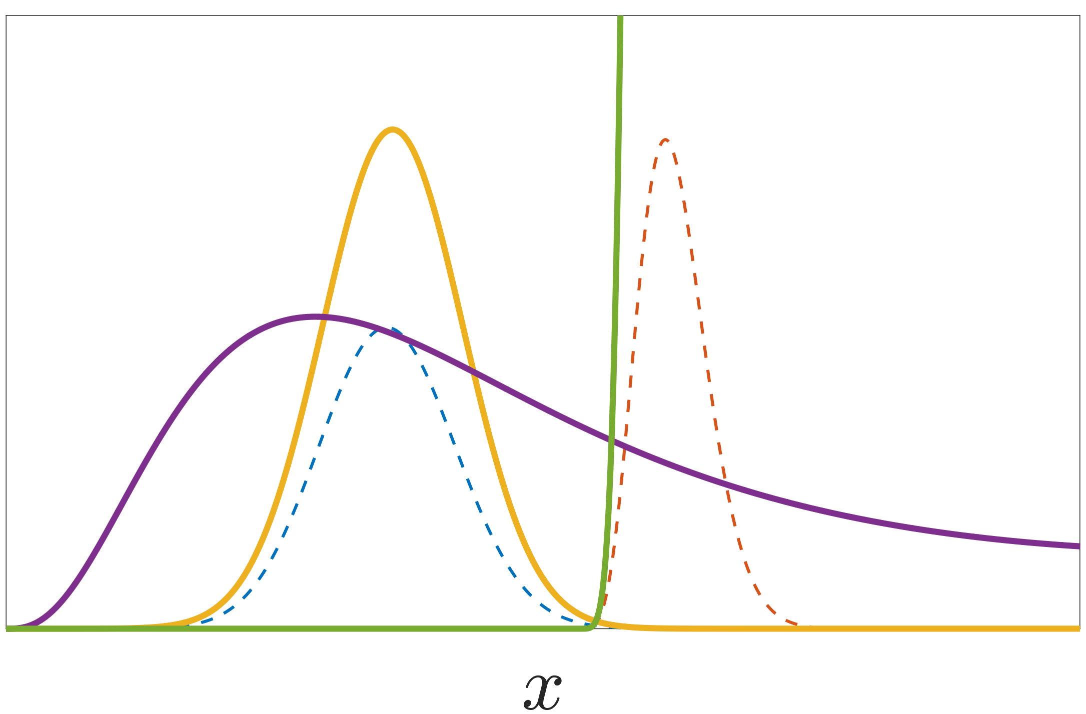


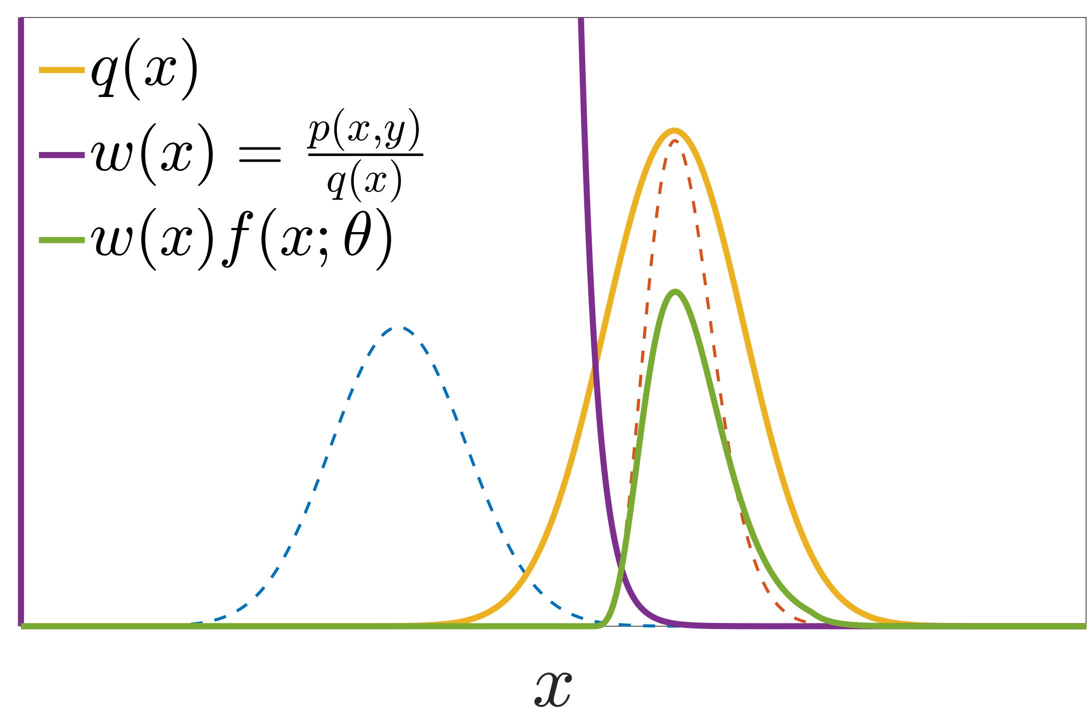




$$\mathbb{E}_{p(x|y)}[f(x)] = \frac{\mathbb{E}_{p(x)}[f(x)p(y|x)]}{\mathbb{E}_{p(x)}[p(y|x)]}$$

$$\approx \frac{\frac{1}{N}\sum_{n=1}^{N} f(x_n)\, w_n}{\frac{1}{N}\sum_{n=1}^{N} w_n}$$





Self-Normalized Importance Sampling (SNIS)

$$\text{Mean Squared Error} \geq \frac{1}{N} \left(\mathbb{E}_{p(x|y)} \left[\left| f(x) - \mathbb{E}_{p(x|y)}[f(x)] \right| \right] \right)^2$$

Solution: Use Multiple Proposals Targeting Different Aspects of the Problem

The AMCI Estimator

$$\mathbb{E}_{p(x|y)}[f(x;\theta)] = \frac{\mathbb{E}_{p(x)}[f(x;\theta)p(y|x)]}{\mathbb{E}_{p(x)}[p(y|x)]}$$

UNIVERSITY OF OXFORD

The AMCI Estimator

$$f^+(x;\theta) = \max(f(x;\theta), 0)$$
$$f^-(x;\theta) = -\min(f(x;\theta), 0)$$
$$f(x;\theta) = f^+(x;\theta) - f^-(x;\theta)$$

$$\mathbb{E}_{p(x|y)}[f(x;\theta)] = \frac{\mathbb{E}_{p(x)}[f(x;\theta)p(y|x)]}{\mathbb{E}_{p(x)}[p(y|x)]}$$

The AMCI Estimator

$$f^{+}(x;\theta) = \max(f(x;\theta), 0)$$
$$f^{-}(x;\theta) = -\min(f(x;\theta), 0)$$
$$f(x;\theta) = f^{+}(x;\theta) - f^{-}(x;\theta)$$

$$\begin{aligned}\mathbb{E}_{p(x|y)}[f(x;\theta)] &= \frac{\mathbb{E}_{p(x)}[f(x;\theta)p(y|x)]}{\mathbb{E}_{p(x)}[p(y|x)]} \\ &= \frac{\mathbb{E}_{p(x)}[f^{+}(x;\theta)p(y|x)] - \mathbb{E}_{p(x)}[f^{-}(x;\theta)p(y|x)]}{\mathbb{E}_{p(x)}[p(y|x)]}\end{aligned}$$

The AMCI Estimator

$$f^+(x;\theta) = \max(f(x;\theta), 0)$$
$$f^-(x;\theta) = -\min(f(x;\theta), 0)$$
$$f(x;\theta) = f^+(x;\theta) - f^-(x;\theta)$$

$$\mathbb{E}_{p(x|y)}[f(x;\theta)] = \frac{\mathbb{E}_{p(x)}[f(x;\theta)p(y|x)]}{\mathbb{E}_{p(x)}[p(y|x)]}$$
$$= \frac{\mathbb{E}_{p(x)}[f^+(x;\theta)p(y|x)] - \mathbb{E}_{p(x)}[f^-(x;\theta)p(y|x)]}{\mathbb{E}_{p(x)}[p(y|x)]}$$
$$= \frac{\mathbb{E}_{q_1^+(x;y,\theta)}\left[\frac{f^+(x;\theta)p(y|x)}{q_1^+(x;y,\theta)}\right] - \mathbb{E}_{q_1^-(x;y,\theta)}\left[\frac{f^-(x;\theta)p(y|x)}{q_1^-(x;y,\theta)}\right]}{\mathbb{E}_{q_2(x;y)}\left[\frac{p(y|x)}{q_2(x;y)}\right]}$$

UNIVERSITY OF OXFORD

The AMCI Estimator

$$f^+(x;\theta) = \max(f(x;\theta), 0)$$
$$f^-(x;\theta) = -\min(f(x;\theta), 0)$$
$$f(x;\theta) = f^+(x;\theta) - f^-(x;\theta)$$

$$\begin{aligned}
\mathbb{E}_{p(x|y)}[f(x;\theta)] &= \frac{\mathbb{E}_{p(x)}[f(x;\theta)p(y|x)]}{\mathbb{E}_{p(x)}[p(y|x)]} \\
&= \frac{\mathbb{E}_{p(x)}[f^+(x;\theta)p(y|x)] - \mathbb{E}_{p(x)}[f^-(x;\theta)p(y|x)]}{\mathbb{E}_{p(x)}[p(y|x)]} \\
&= \frac{\mathbb{E}_{q_1^+(x;y,\theta)}\left[\frac{f^+(x;\theta)p(y|x)}{q_1^+(x;y,\theta)}\right] - \mathbb{E}_{q_1^-(x;y,\theta)}\left[\frac{f^-(x;\theta)p(y|x)}{q_1^-(x;y,\theta)}\right]}{\mathbb{E}_{q_2(x;y)}\left[\frac{p(y|x)}{q_2(x;y)}\right]} \\
&=: \frac{E_1^+ - E_1^-}{E_2}
\end{aligned}$$

UNIVERSITY OF OXFORD

The AMCI Estimator

$$E_2 = \mathbb{E}_{p(x)}\left[p(y|x)\right]$$

$$E_1^+ = \mathbb{E}_{p(x)}\left[\max(f(x;\theta),0)p(y|x)\right]$$

$$E_1^- = \mathbb{E}_{p(x)}\left[-\min(f(x;\theta),0)p(y|x)\right]$$

The AMCI Estimator

$$E_2 = \mathbb{E}_{p(x)}\left[p(y|x)\right] \approx \hat{E}_2 = \frac{1}{M}\sum_{m=1}^{M}\frac{p(x_m, y)}{q_2(x_m; y)}$$

$$E_1^+ = \mathbb{E}_{p(x)}\left[\max(f(x;\theta), 0)p(y|x)\right] \approx \hat{E}_1^+ = \frac{1}{N}\sum_{n=1}^{N}\frac{f^+(x_n^+;\theta)p(x_n^+, y)}{q_1^+(x_n^+; y, \theta)}$$

$$E_1^- = \mathbb{E}_{p(x)}\left[-\min(f(x;\theta), 0)p(y|x)\right] \approx \hat{E}_1^- = \frac{1}{K}\sum_{k=1}^{K}\frac{f^-(x_k^-;\theta)p(x_k^-, y)}{q_1^-(x_k^-; y, \theta)}$$

UNIVERSITY OF OXFORD

AMCI Can Produce Perfect Estimates with a Single Sample from Each Proposal!

Theorem 1. *If* $q_2(x;y) \propto p(x,y)$, $q_1^+(x;y,\theta) \propto f^+(x;\theta)p(x,y)$, *and* $q_1^-(x;y,\theta) \propto f^-(x;\theta)p(x,y)$, *then the AMCI estimator* $(\hat{E}_1^+ - \hat{E}_1^-)/\hat{E}_2$ *is an exact estimator for* $\mathbb{E}_{p(x|y)}[f(x;\theta)]$ *even if* $N = K = M = 1$

Amortized Inference

Amortized Inference

$$D_{KL}[p(x|y)||q_2(x;y,\eta)]$$

Amortized Inference

$$\mathcal{J}_{q_2}(\eta) = \mathbb{E}_{p(y)}\left[D_{KL}[p(x|y)||q_2(x;y,\eta)]\right]$$

B. Paige, F. Wood. Inference networks for SMC in graphical models. ICML, 2016.

UNIVERSITY OF OXFORD

Amortized Inference

$$\begin{aligned} \mathcal{J}_{q_2}(\eta) &= \mathbb{E}_{p(y)}\left[D_{KL}[p(x|y) \| q_2(x; y, \eta)]\right] \\ &= \mathbb{E}_{p(x,y)}[-\log q_2(x; y, \eta)] + \text{const wrt } \eta \end{aligned}$$

B. Paige, F. Wood. Inference networks for SMC in graphical models. ICML, 2016.

Amortizing Monte Carlo Integration

Amortizing Monte Carlo Integration

To amortize over function parameters, we introduce a pseudo prior $p(\theta)$

Amortizing Monte Carlo Integration

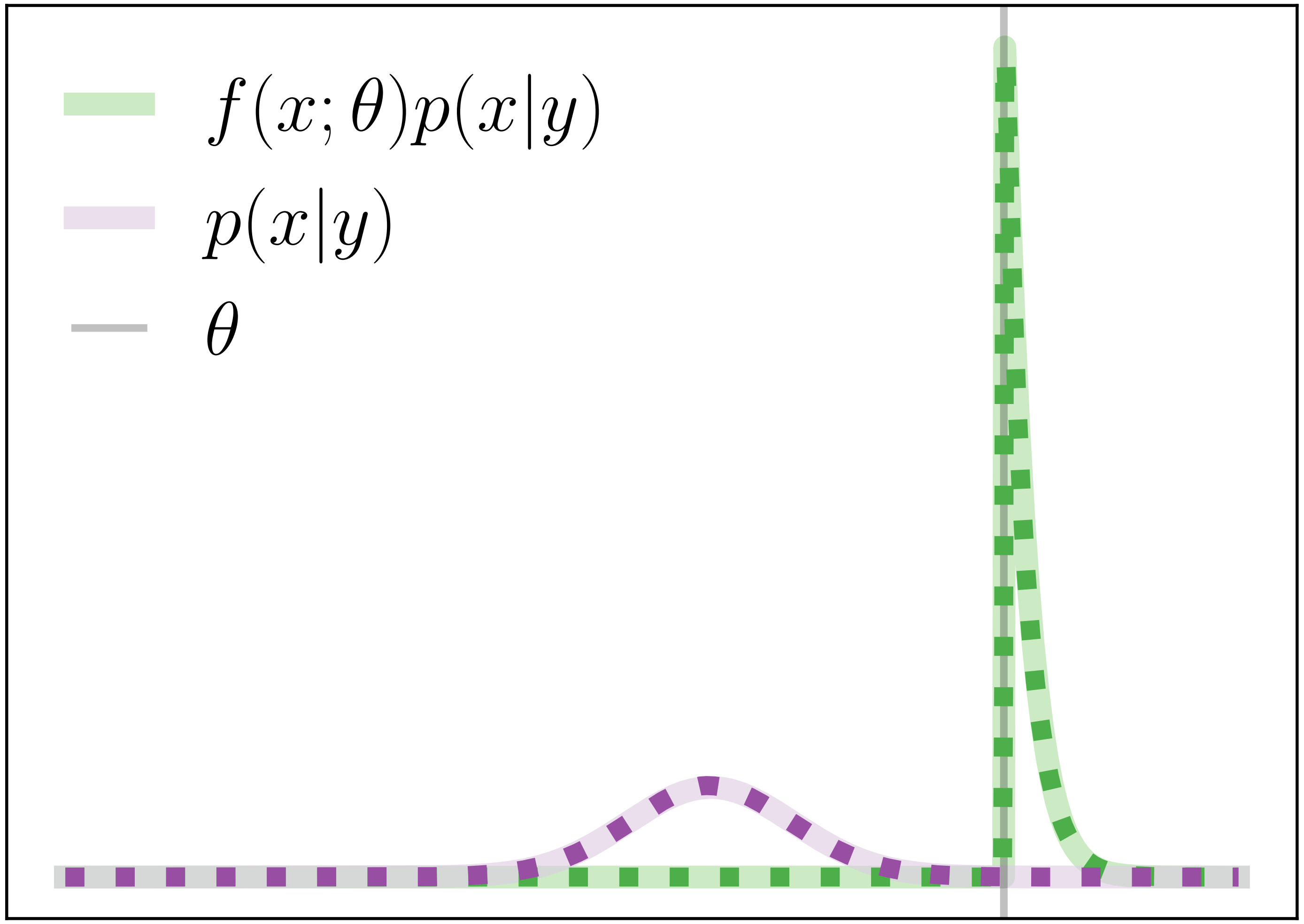
To amortize over function parameters, we introduce a pseudo prior $p(\theta)$

$$q_1^{\pm} \propto f^{\pm}(x;\theta)p(x,y)$$

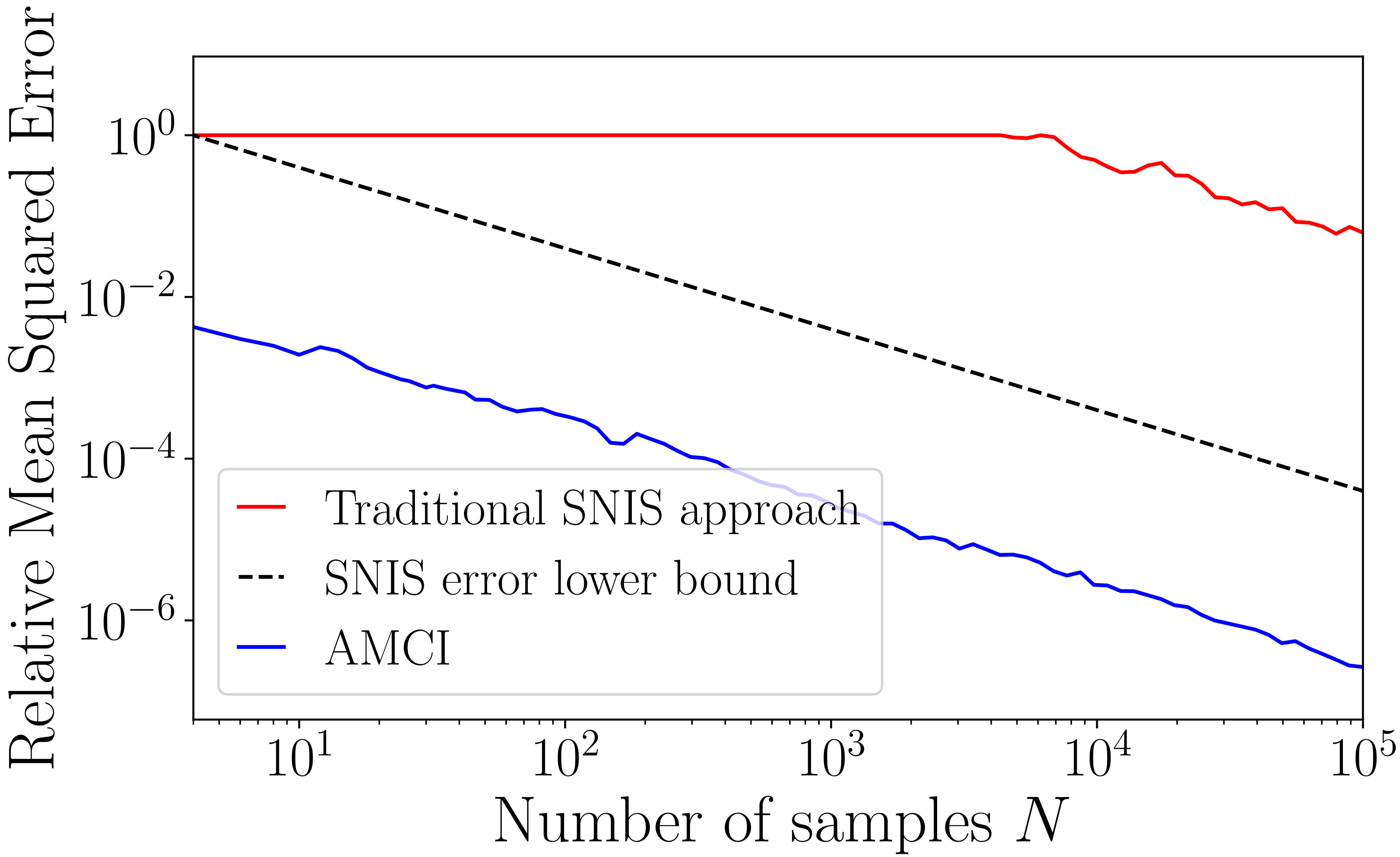
$$\mathcal{J}_{q_1^{\pm}}(\eta) = \mathbb{E}_{p(x,y)p(\theta)}\left[-f^{\pm}(x;\theta)\log q_1^{\pm}(x;y,\theta,\eta)\right] + \text{const}$$

Experiments: Gaussian Tail Integral

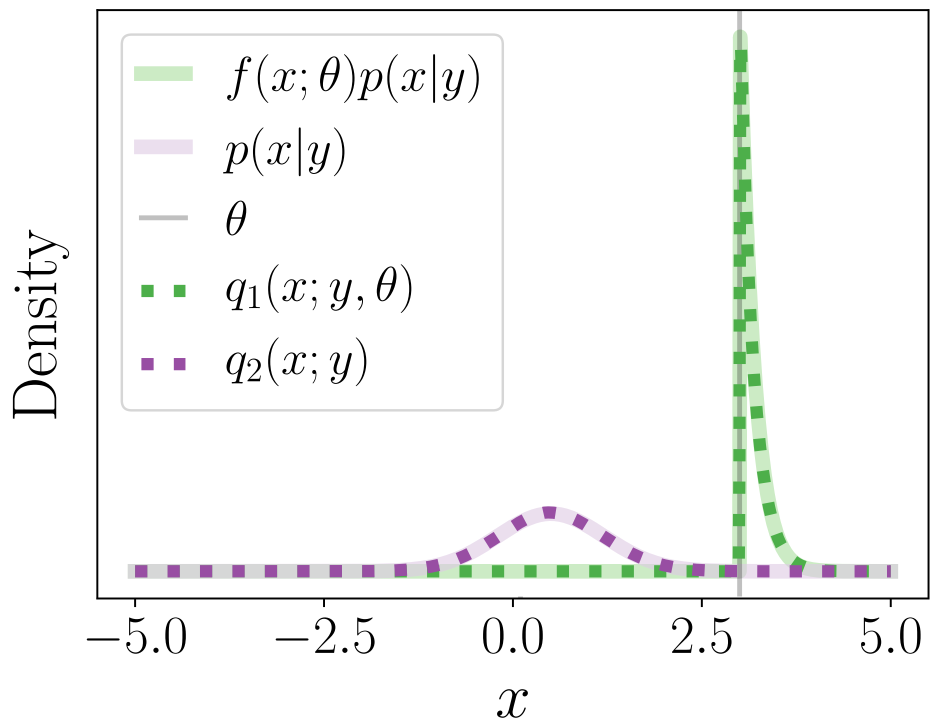
$$p(x) = \mathcal{N}(x; 0, 1)$$
$$p(y|x) = \mathcal{N}(y; x, 1)$$
$$f(x;\theta) = \mathbb{1}_{x>\theta}$$



Experiments: Gaussian Tail Integral

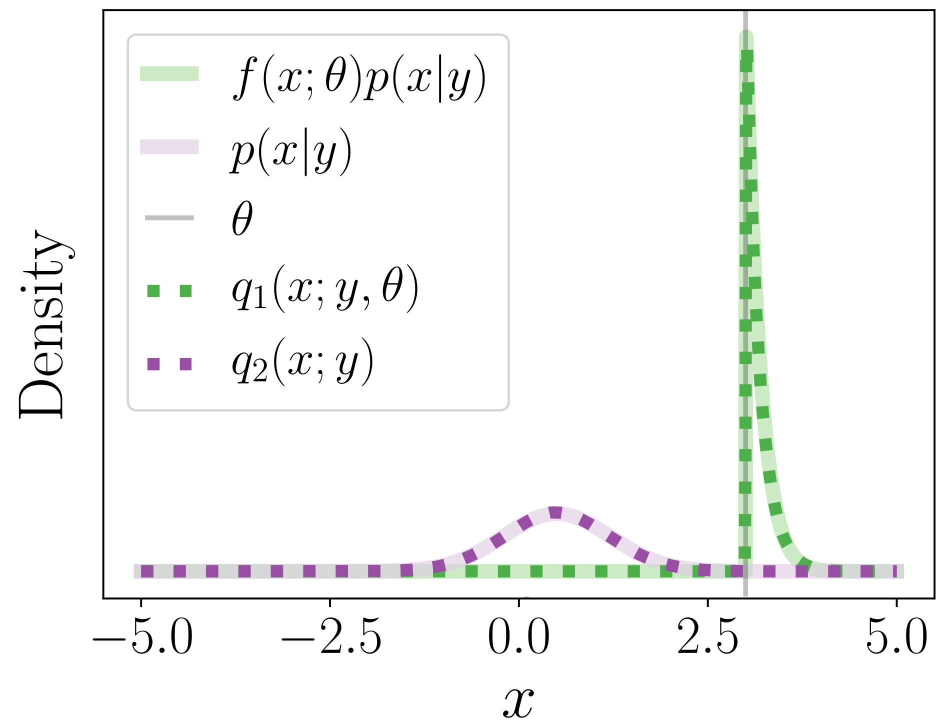


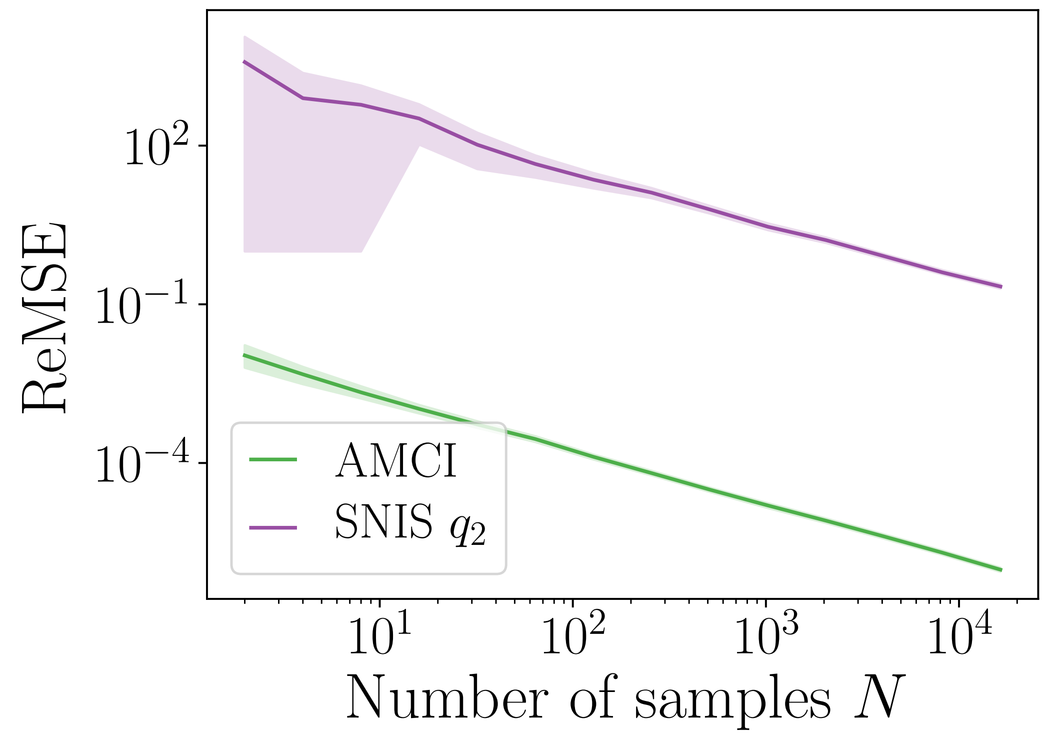
When Does AMCI Work Well in Practice?



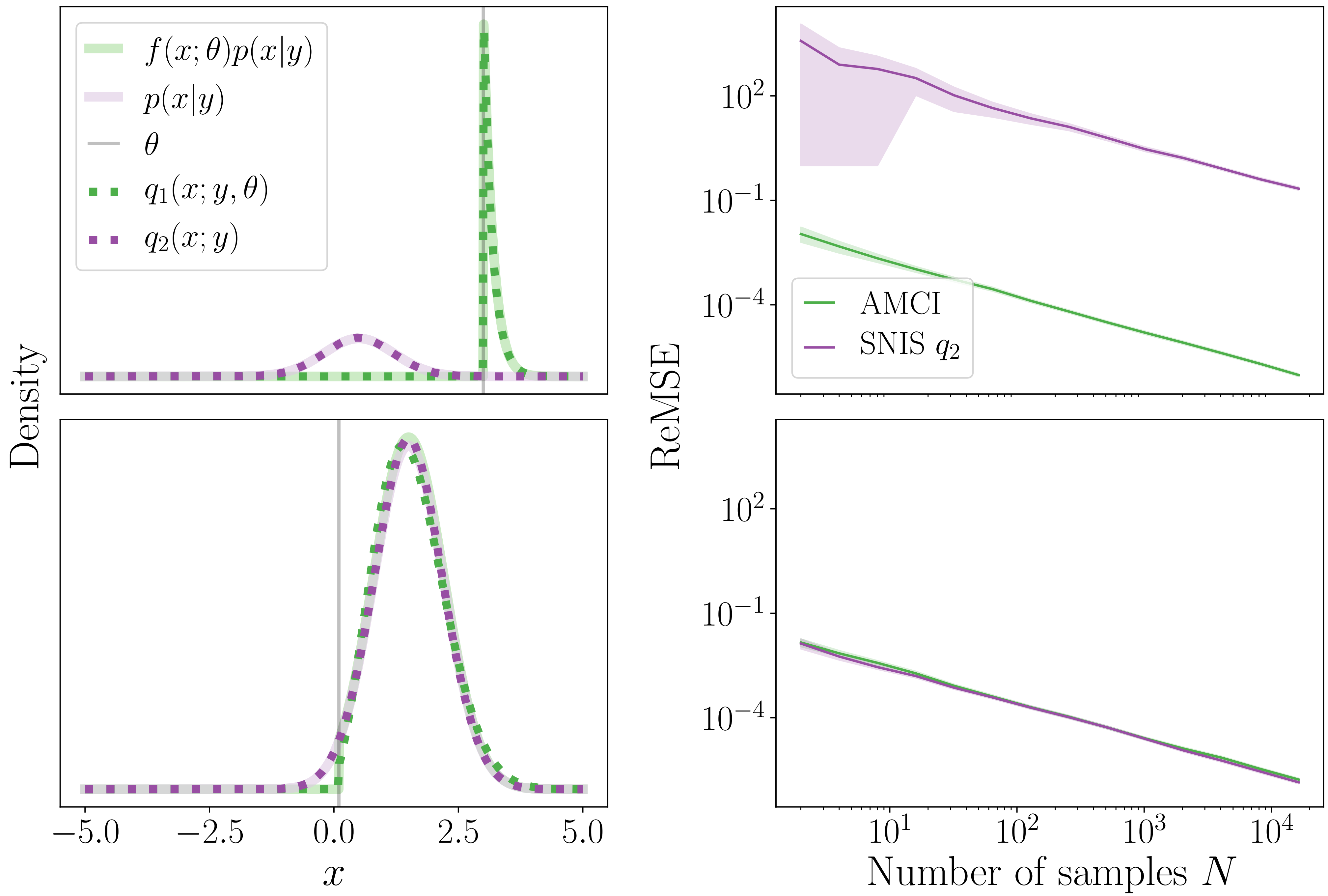
UNIVERSITY OF OXFORD

When Does AMCI Work Well in Practice?





When Does AMCI Work Well in Practice?



Recap

- Amortized inference focuses on approximating the posterior $p(x|y)$, later using this to estimate expectation(s) $\mathbb{E}_{p(x|y)}[f(x;\theta)]$
- This pipeline is inefficient if $f(x;\theta)$ is known upfront
- AMCI instead targets $\mathbb{E}_{p(x|y)}[f(x;\theta)]$ directly, allowing amortization over datasets y and/or function parameters θ
- It can give exact estimates for any expectation with only a single sample from each of three separate amortized proposals
- It can empirically outperform the theoretically optimal self-normalized importance sampler, even in non-amortized settings

UNIVERSITY OF OXFORD

Adam Goliński
@adam_golinski

Frank Wood
@frankdonaldwood

Tom Rainforth
@tom_rainforth

Come see us at poster #999
today at 6.30pm!

When Does AMCI Work Well in Practice?

κ is a measure of relative performance of top and bottom estimators

For AMCI, we can control κ, for SNIS we cannot

$$\text{MSE} \approx \frac{\sigma_2^2}{E_2^2} \cdot \left((\kappa - \text{Corr}[\xi_1, \xi_2])^2 + 1 - \text{Corr}[\xi_1, \xi_2]^2\right)$$

Intractability in the naively adjusted objective

$$E_1(y) := \mathbb{E}_{p(x)}[f(x)p(y|x)]$$

$$g(x|y) := \frac{f(x)p(x,y)}{E_1(y)}$$

$$\begin{aligned}\mathcal{J}'_{q_1}(\eta) &= \mathbb{E}_{p(y)}\left[D_{KL}\left(g(x|y)\|q_1(x;y,\eta)\right)\right] \\ &= \mathbb{E}_{p(y)}\left[-\int_{\mathcal{X}} \frac{f(x)p(x,y)}{E_1(y)} \log q_1(x;y,\eta)\mathrm{d}x\right] + \text{const}\end{aligned}$$

UNIVERSITY OF OXFORD

Intractability in the naively adjusted objective

$$E_1(y) := \mathbb{E}_{p(x)}[f(x)p(y|x)]$$

$$g(x|y) := \frac{f(x)p(x,y)}{E_1(y)}$$

$$\mathcal{J}'_{q_1}(\eta) = \mathbb{E}_{p(y)}\left[D_{KL}\left(g(x|y) \| q_1(x;y,\eta)\right)\right]$$

$$= \mathbb{E}_{p(y)}\left[-\int_{\mathcal{X}} \frac{f(x)p(x,y)}{E_1(y)} \log q_1(x;y,\eta)\mathrm{d}x\right] + \text{const}$$

$$h(y) \propto p(y)E_1(y)$$

$$\mathcal{J}_{q_1}(\eta) = \mathbb{E}_{h(y)}\left[D_{KL}\left(g(x|y) \| q_1(x;y,\eta)\right)\right]$$

$$= \frac{1}{\text{const}} \mathbb{E}_{p(x,y)}\left[-f(x) \log q_1(x;y,\eta)\right] + \text{const}$$

Experiments: Cancer Treatment Planning