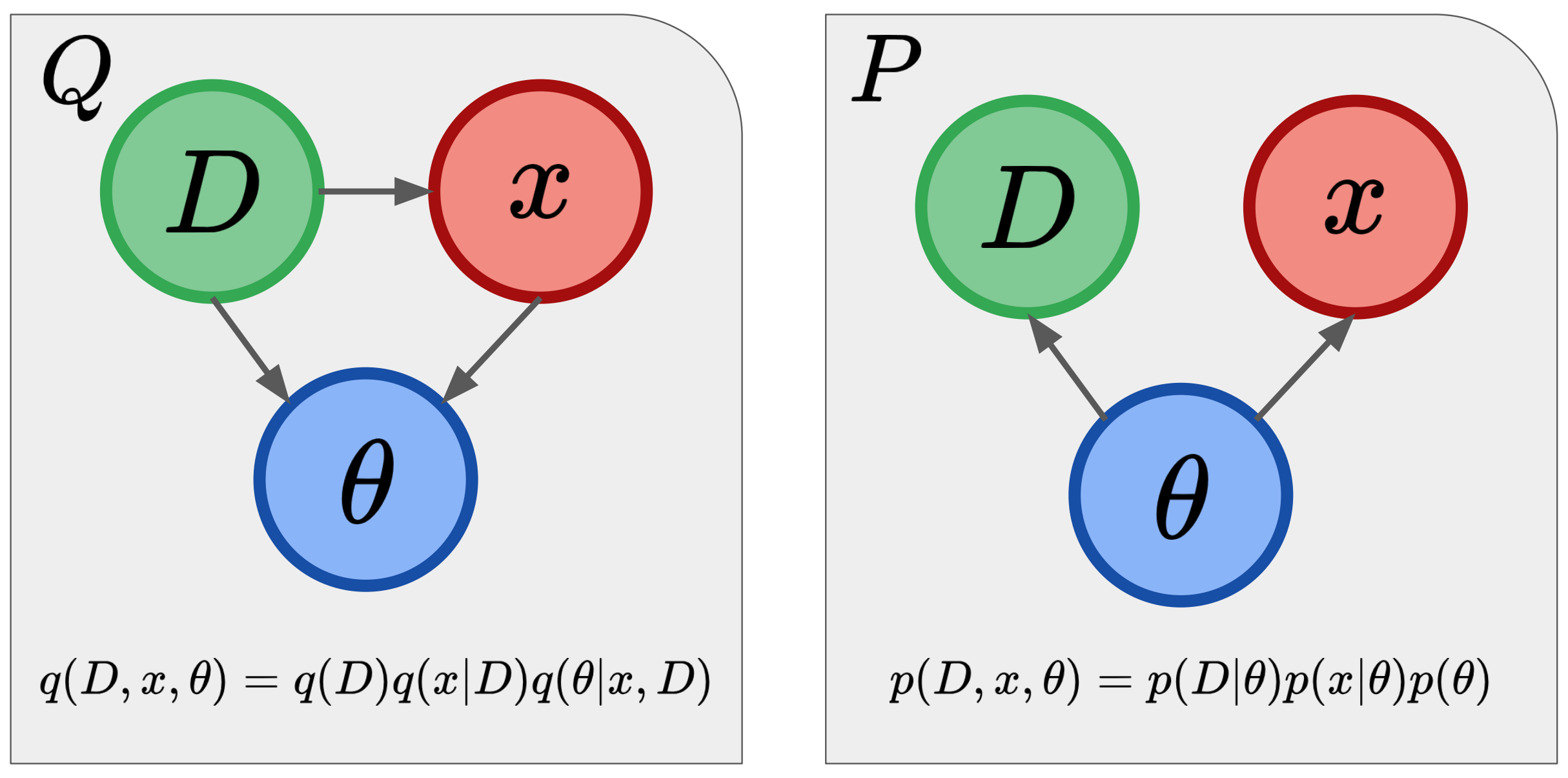
Variational Prediction

Alexander A. Alemi, Ben Poole
alemi@google.com
Google Research

arXiv:2307.07568

Bayesian inference offers benefits over maximum likelihood, but it also comes with computational costs. Computing the posterior is typically intractable, as is marginalizing that posterior to form the posterior predictive distribution. In this paper, we present variational prediction, a technique for directly learning a variational approximation to the posterior predictive distribution using a variational bound. This approach can provide good predictive distributions without test time marginalization costs. We demonstrate Variational Prediction on an illustrative toy example.

The usual Bayesian generative modelling assumption (P) versus what we really want (Q).



$$q(D,x,\theta) = q(D)q(x|D)q(\theta|x,D)$$

$$p(D,x,\theta) = p(D|\theta)p(x|\theta)p(\theta)$$

This objective is a bound on the marginal likelihood:

$$\left\langle \log \frac{q(x,\theta|D)}{p(x,\theta|D)} \right\rangle = \left\langle \log \frac{q(x|D)q(\theta|x,D)}{p(x|\theta)p(\theta|D)} \right\rangle = \left\langle \log \frac{q(x|D)q(\theta|x,D)}{p(x|\theta)p(D|\theta)p(\theta)} \right\rangle + \log p(D) \geq 0$$

The objective drives our variational predictive distribution closer to the true Bayesian predictive distribution:

$$\left\langle \log \frac{q(x,\theta|D)}{p(x,\theta|D)} \right\rangle = \left\langle \log \frac{q(x|D)}{p(x|D)} \right\rangle + \left\langle \log \frac{q(\theta|x,D)}{p(\theta|x,D)} \right\rangle \geq \left\langle \log \frac{q(x|D)}{p(x|D)} \right\rangle \geq 0$$

It naturally extends to conditional modelling:

$$\left\langle \log \frac{q(y|x,D)q(x|D)q(\theta|y,x,D)}{p(y|x,\theta)p(x)p(D|\theta)p(\theta)} \right\rangle_q + \log p(D) \geq \left\langle \log \frac{q(y|x,D)}{p(y|x,D)} \right\rangle_q$$

Taking the KL divergence between the real world (Q) and the desired world (P) gives us the ***Variational Prediction*** objective:

$$\left\langle \log \frac{\overbrace{q(x|D)}^{\text{Variational Predictive}}\overbrace{q(\theta|x,D)}^{\text{Variational Augmented Posterior}}}{\underbrace{p(x|\theta)}_{\text{Likelihood'}}\underbrace{p(D|\theta)}_{\text{Likelihood}}\underbrace{p(\theta)}_{\text{Prior}}} \right\rangle_q$$

For computational ease, we used a MAML inspired implicit variational augmented posterior:

$$q(\theta|y,x,D) = q(\theta'|D)$$
$$\theta' = \theta - \lambda\nabla\left\langle -\beta \log p(y|x,D) + \log \frac{q(\theta|D)}{p(\theta)} \right\rangle_{q(\theta|D)}$$

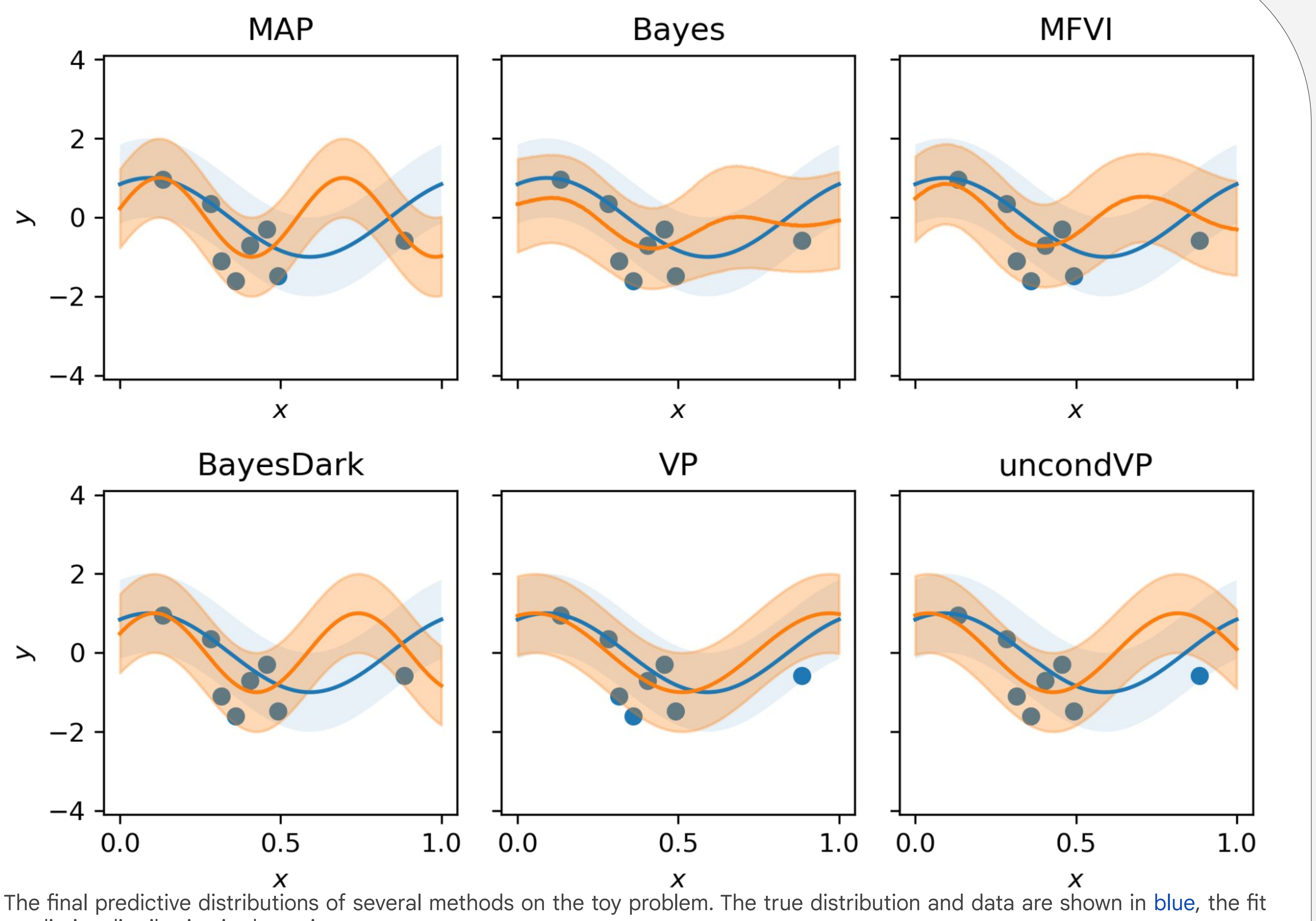
```
def variational_prediction(predictive_model, aug_posterior, likelihood, prior, data):
    # first we sample from the predictive model
    x = predictive_model.sample()
    # then we compute the approximate posterior induced by that sample
    approx_posterior = aug_posterior(x)
    # now we proceed as we would normally with and ELBO by sampling a parameter value
    theta = approx_posterior.sample()
    # and then computing all of the relevant log probabilities.
    return (predictive_model.log_prob(x) + approx_posterior.log_prob(theta)
            - likelihood.log_prob(x) - likelihood.log_prob(data) - prior.log_prob(theta))
```

Toy Model

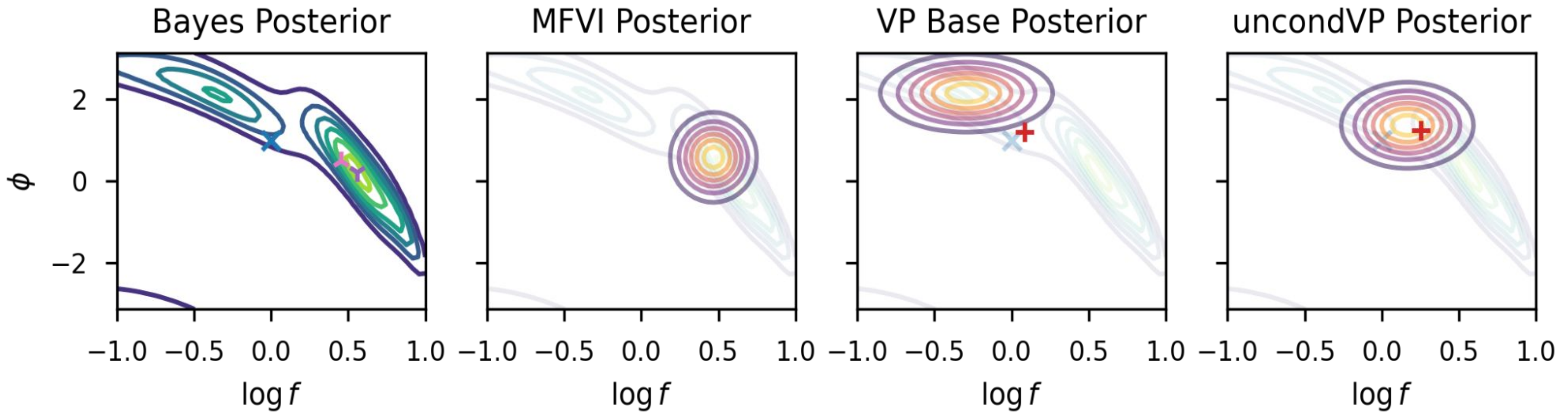
To demonstrate the Variational Prediction method we tried it out on a simple toy problem. The generative process was:

$$x \sim \mathcal{U}(0,1)$$
$$y \sim \mathcal{N}(\mu(x),1)$$
$$\mu(x) = \sin(2\pi f x + \phi)$$

For the same dataset of 8 data points, to the right we show the learned predictive distributions. Below we show the corresponding parameter distributions.



The final predictive distributions of several methods on the toy problem. The true distribution and data are shown in blue, the fit predictive distribution is shown in orange.



The exact Bayesian posterior for the data shown in blue in fig. 2 is shown on the left. The × marks the true parameter values. The ⋎ marks the MAP parameters, ⅄ marks the BayesDark parameters. The second and last column show the MFVI and uncondVP learned approximate posteriors, respectively. For the VP example, we learned a conditional posterior which is hard to visualize, shown is the base unconditional posterior that is modified with MAML. For the VP and uncondVP solutions the corresponding learned predictive distribution parameters are indicated by +.

As a fun aside, in a single page paper, Besag (1989) shared the curious formula that appeared on a candidate's exam:

$$p(x|D) = \frac{p(x|\theta)p(\theta|D)}{p(\theta|x,D)}$$

Which highlights how one can determine the Bayesian posterior predictive distribution without marginalization if one knows the likelihood, posterior and an *augmented posterior.*

Variational Prediction uses a learned variational augmented posterior to learn how to approximate the predictive distribution.