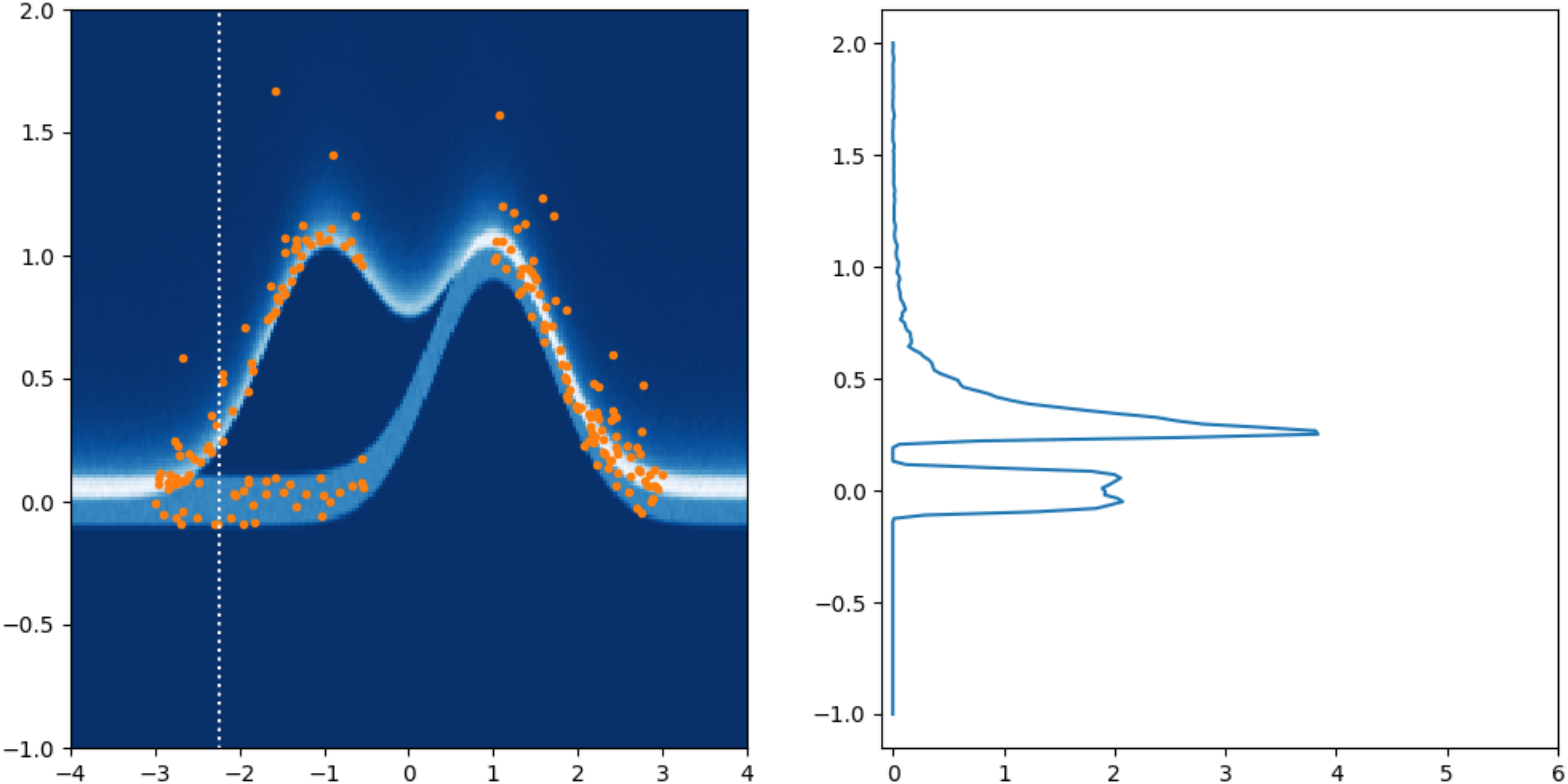
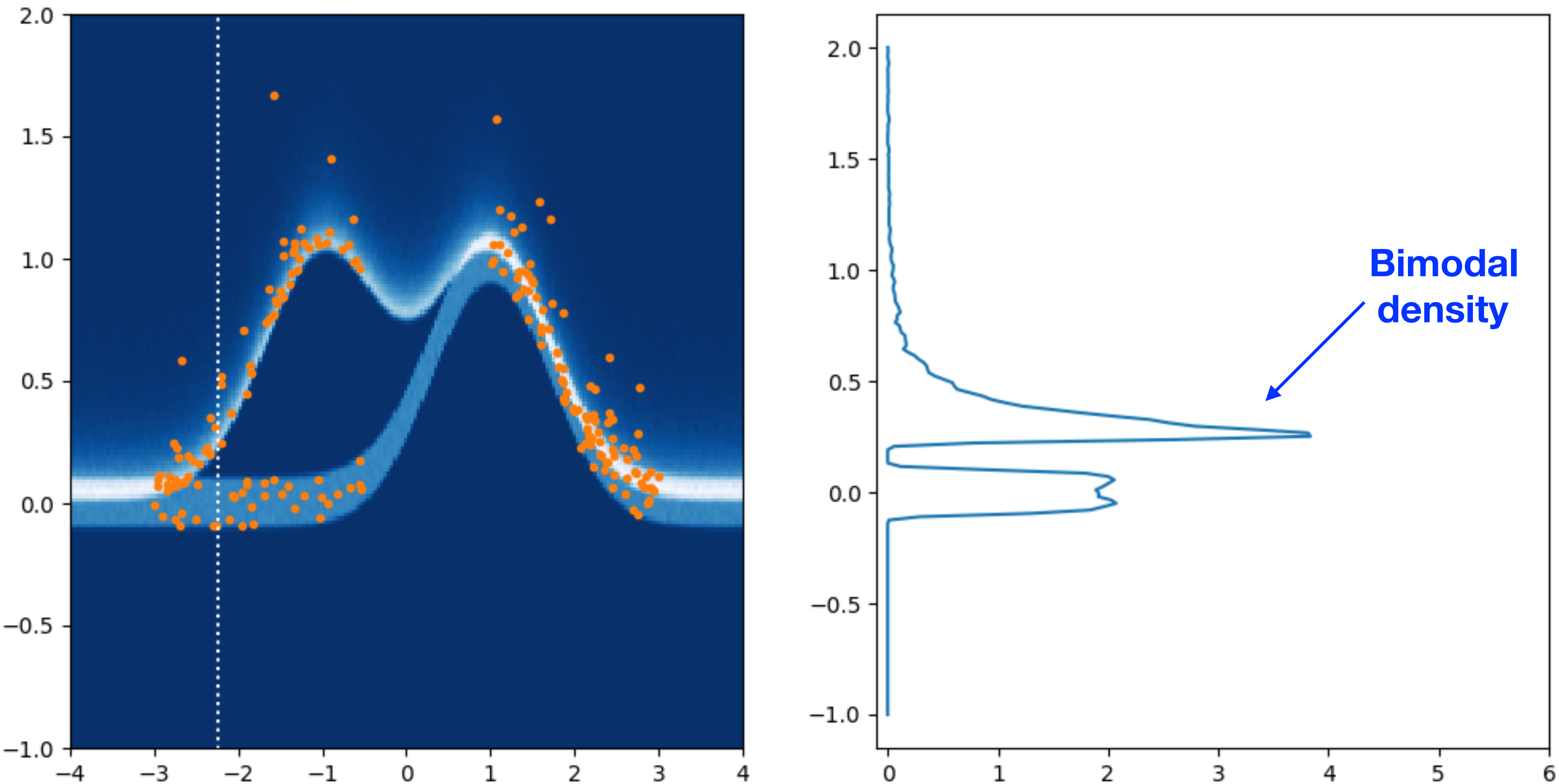
Deep Gaussian Processes with Importance-Weighted Variational Inference

Hugh Salimbeni
Vincent Dutordoir, James Hensman, Marc P Deisenroth

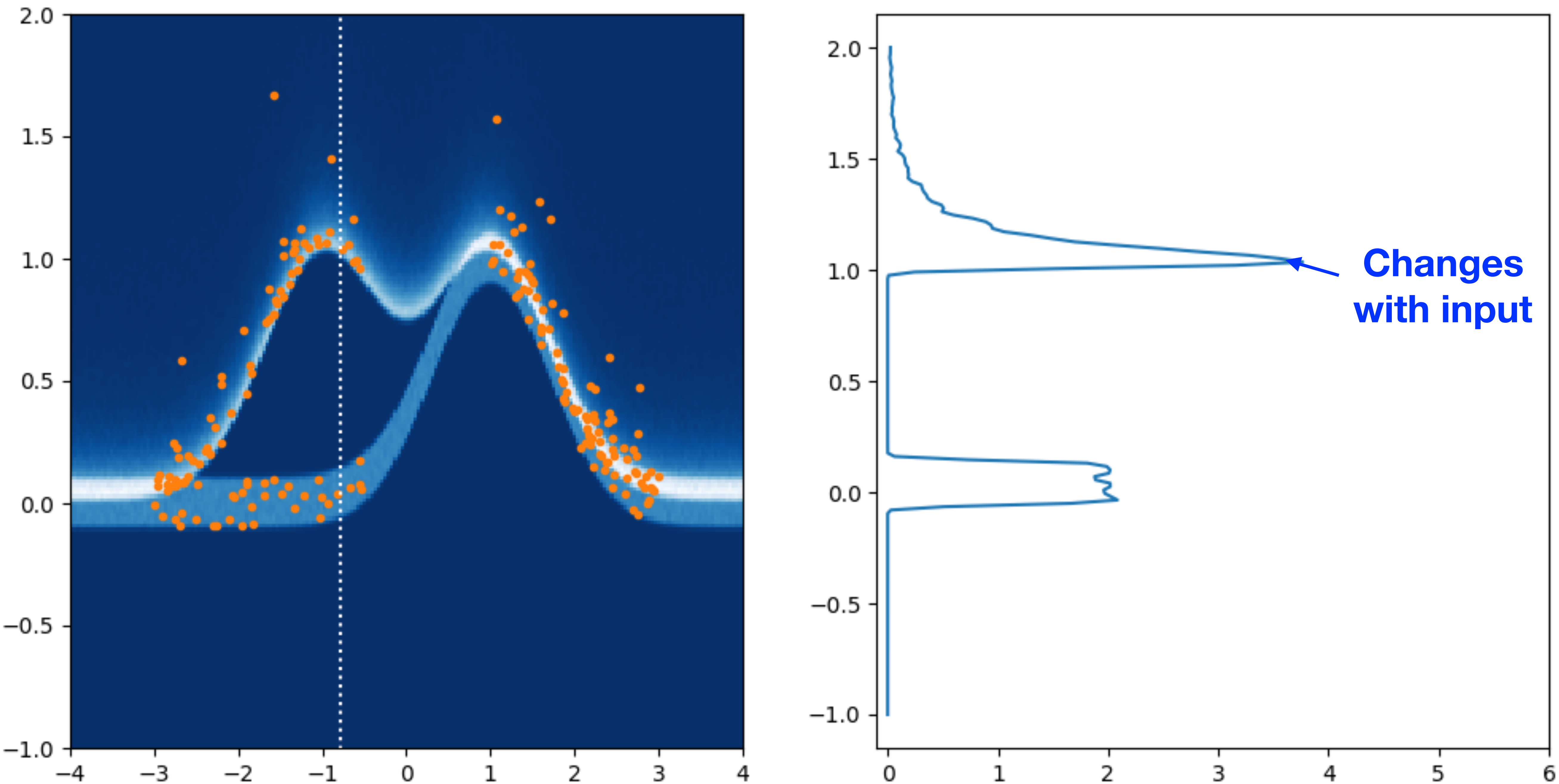
Problem setting



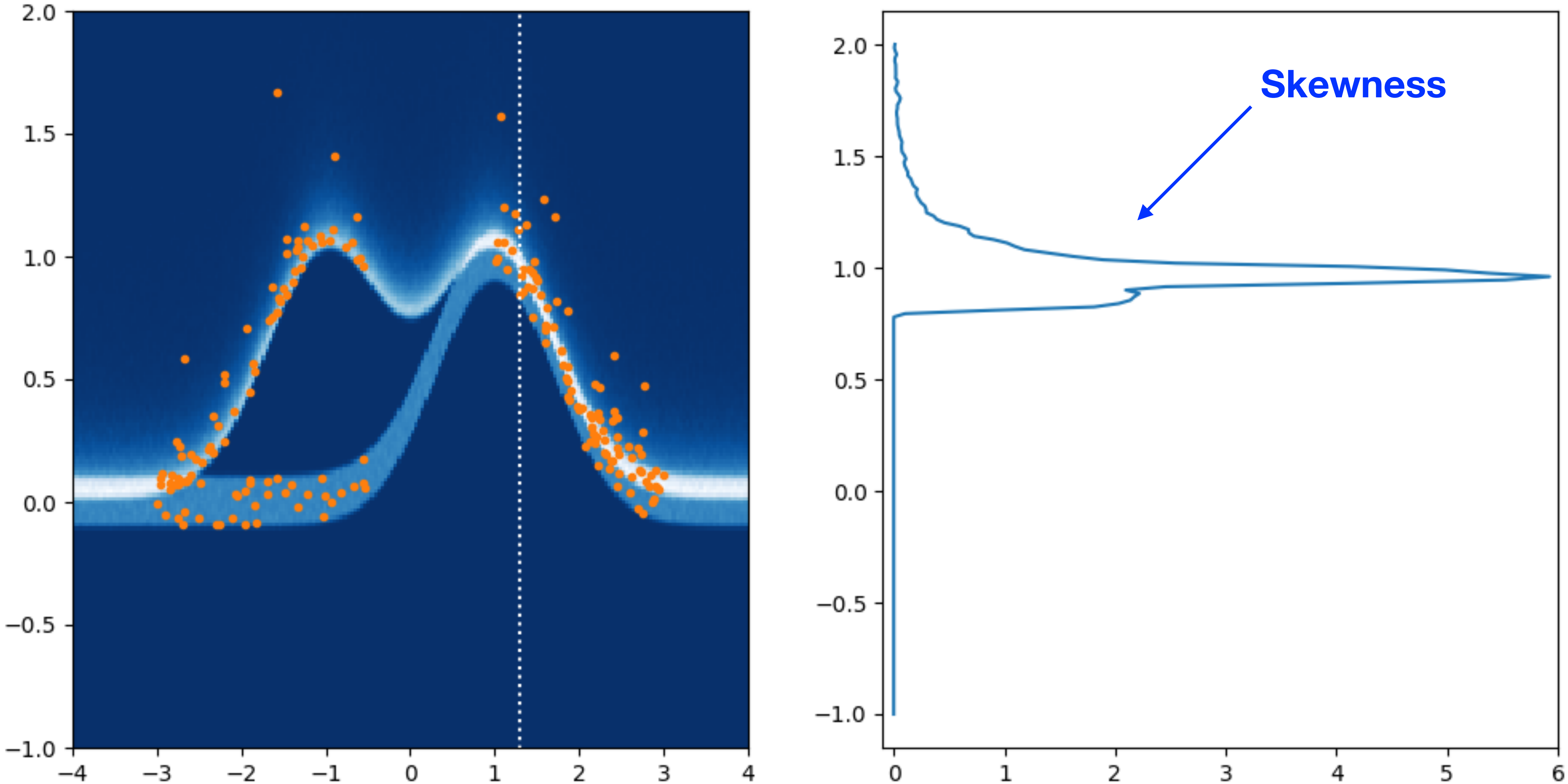
Problem setting



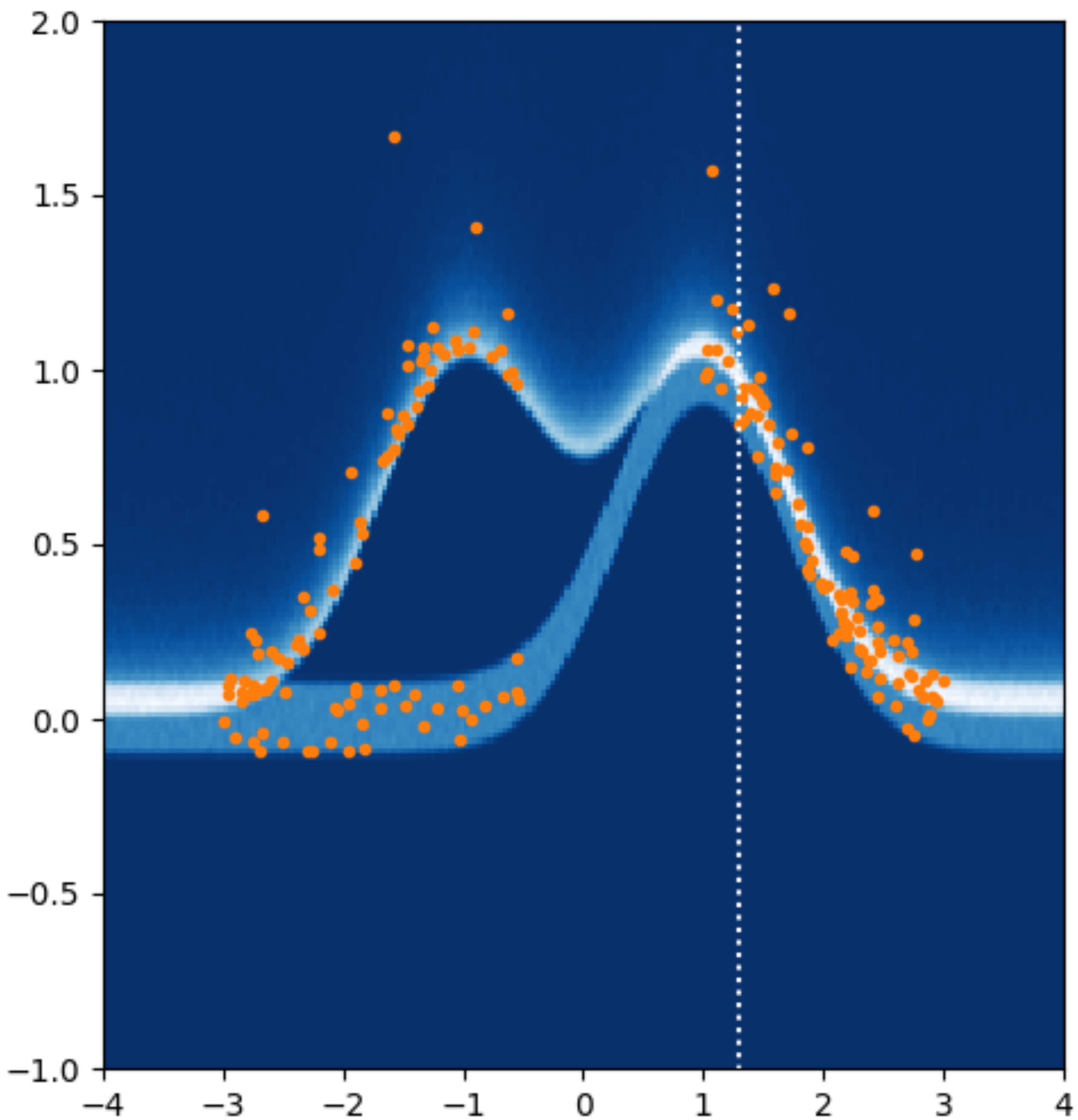
Problem setting

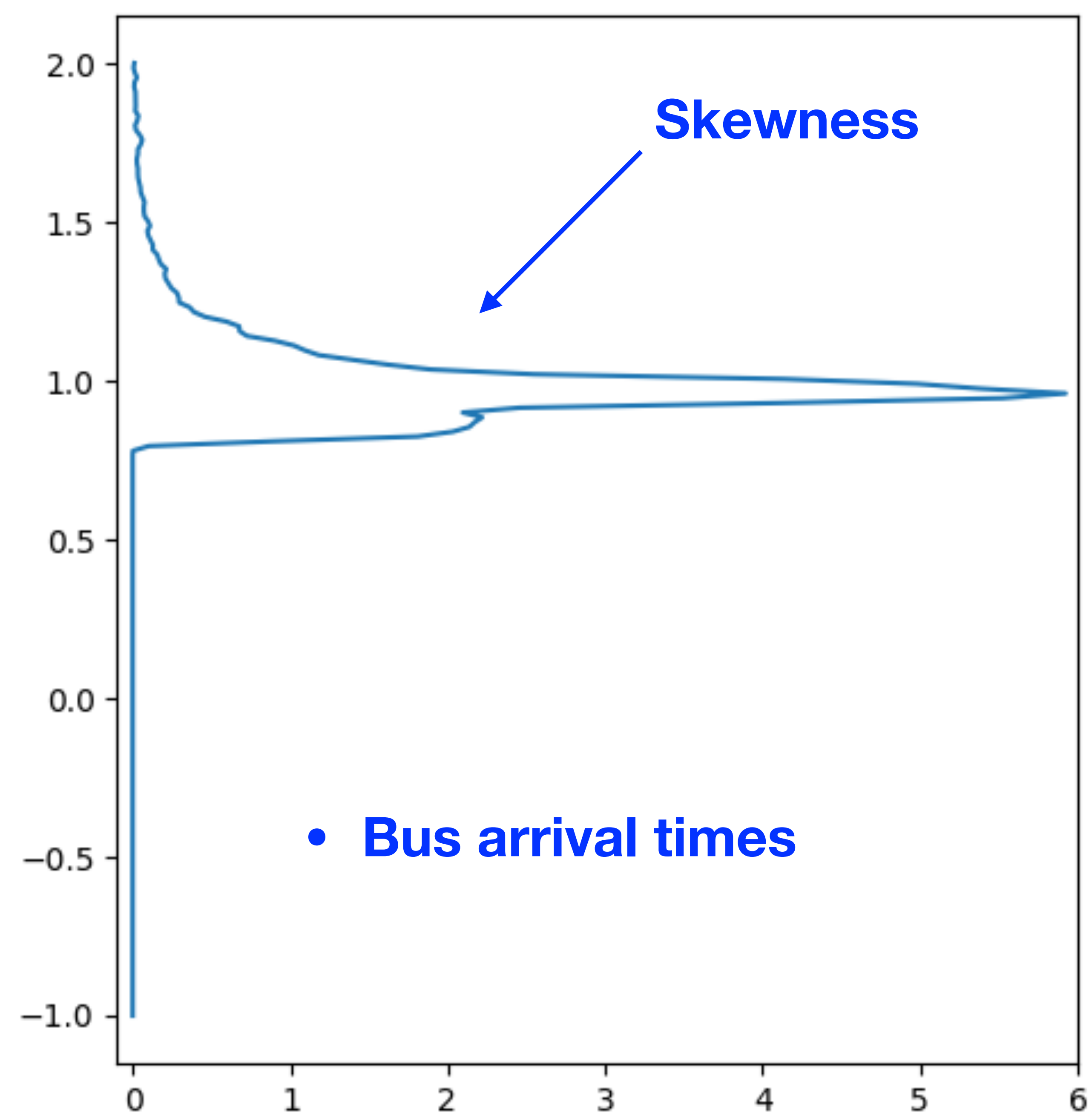


Problem setting

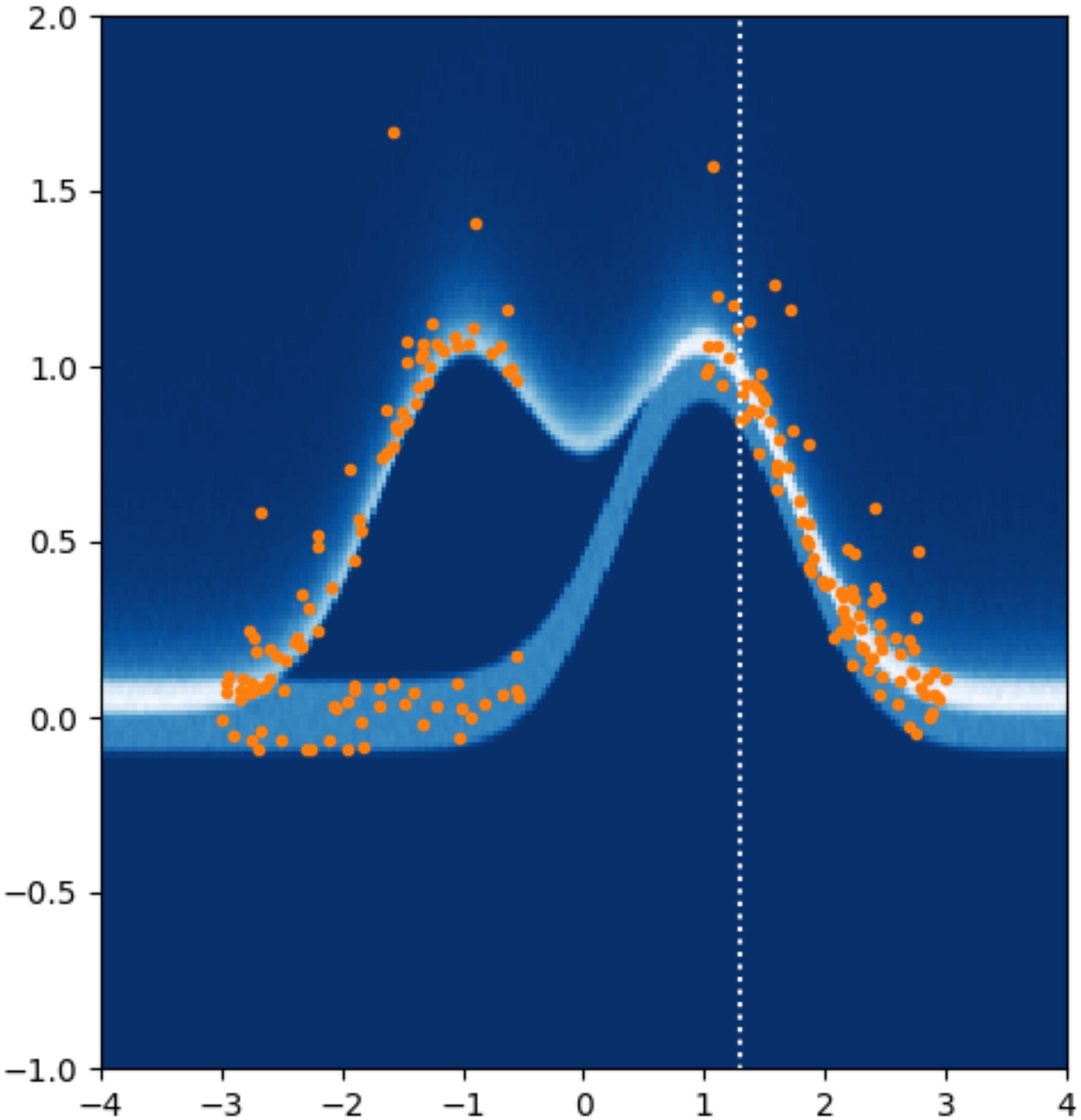


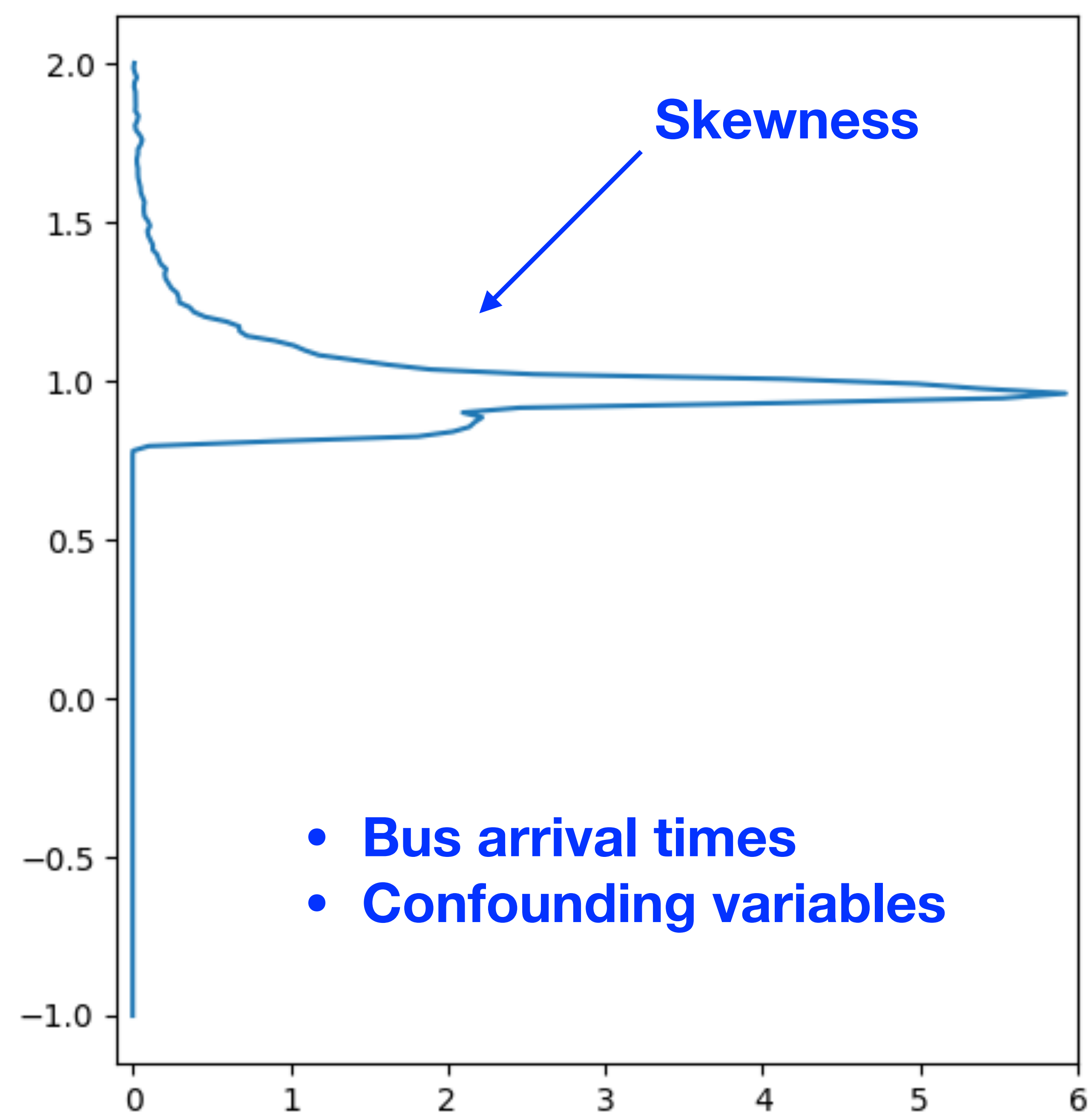
Problem setting



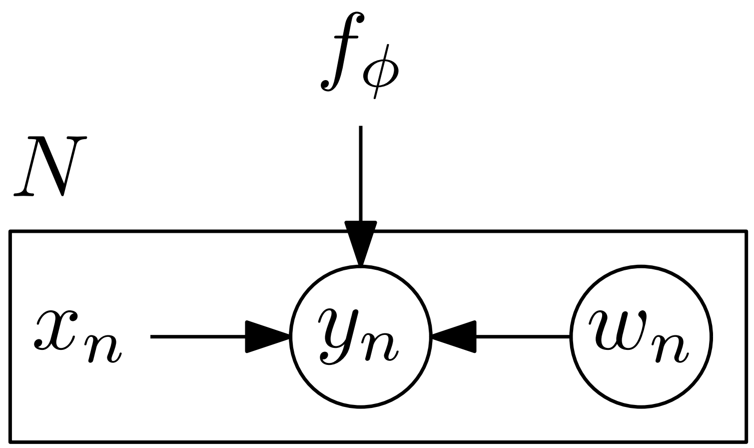


Problem setting

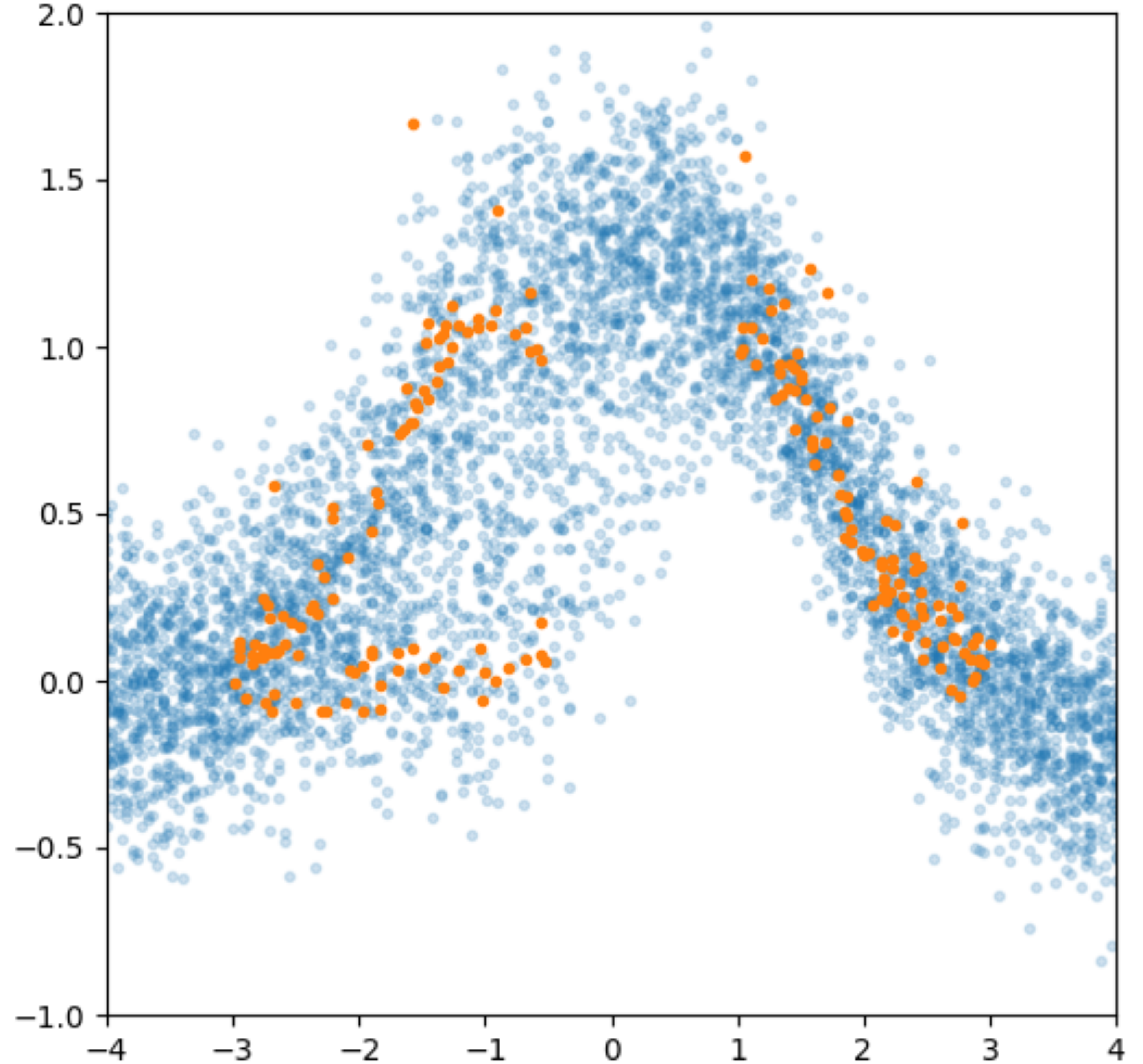




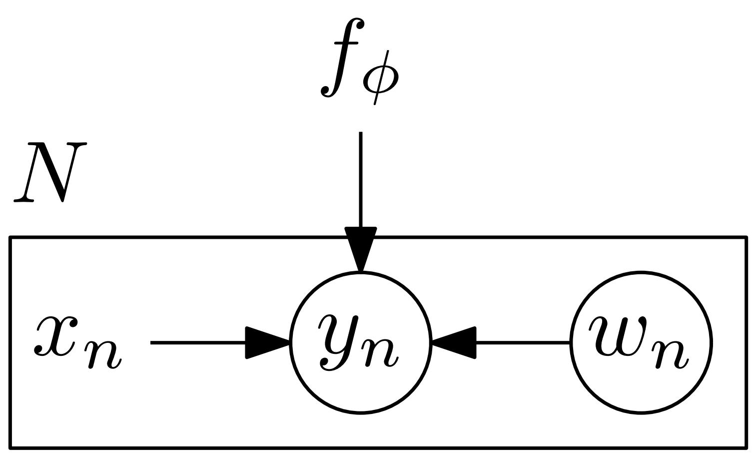
A possible approach



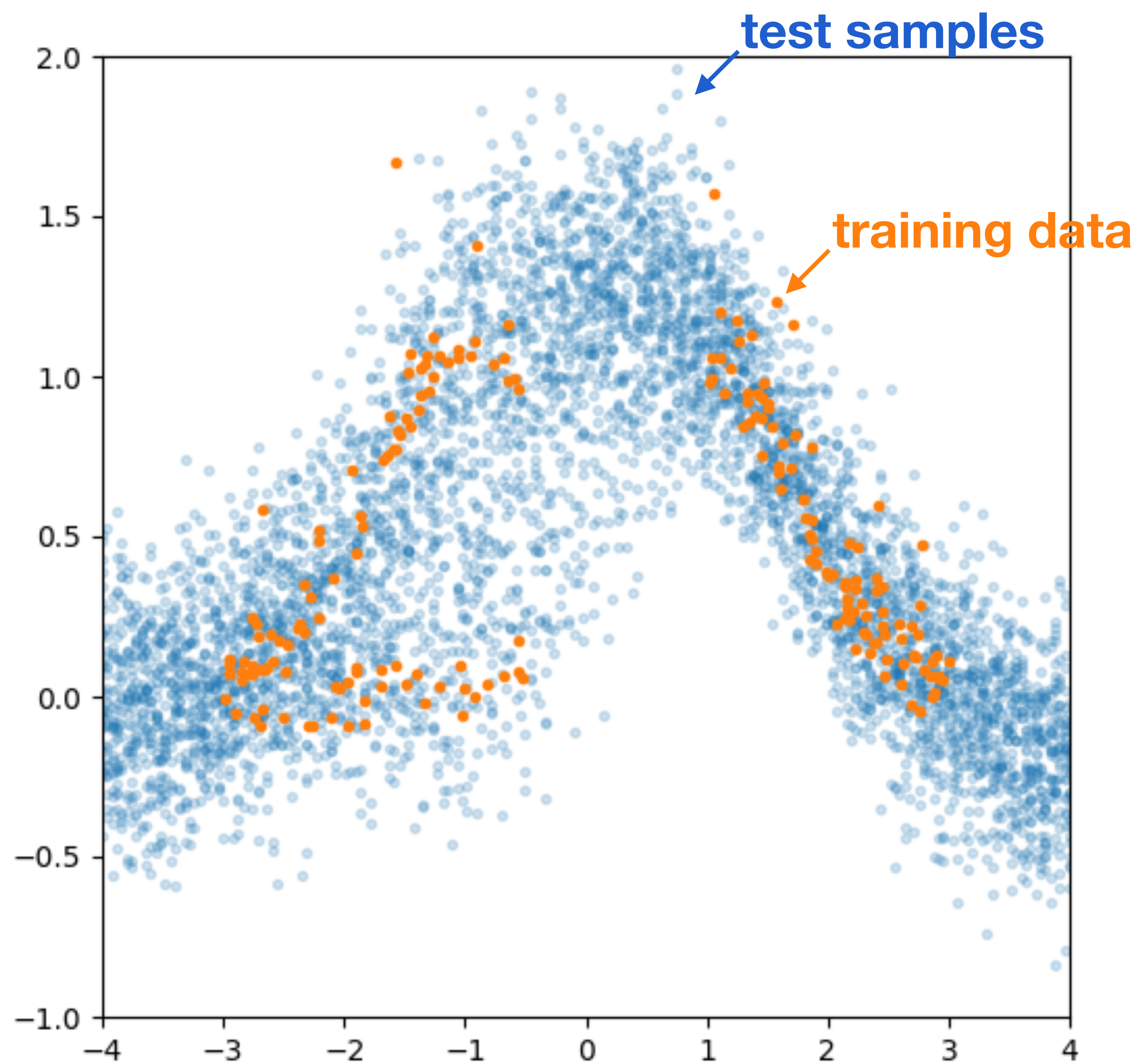
$$y_n = \mathcal{N}(f_\phi([x_n, w_n]), \sigma^2)$$
$$w_n \sim \mathcal{N}(0, 1)$$



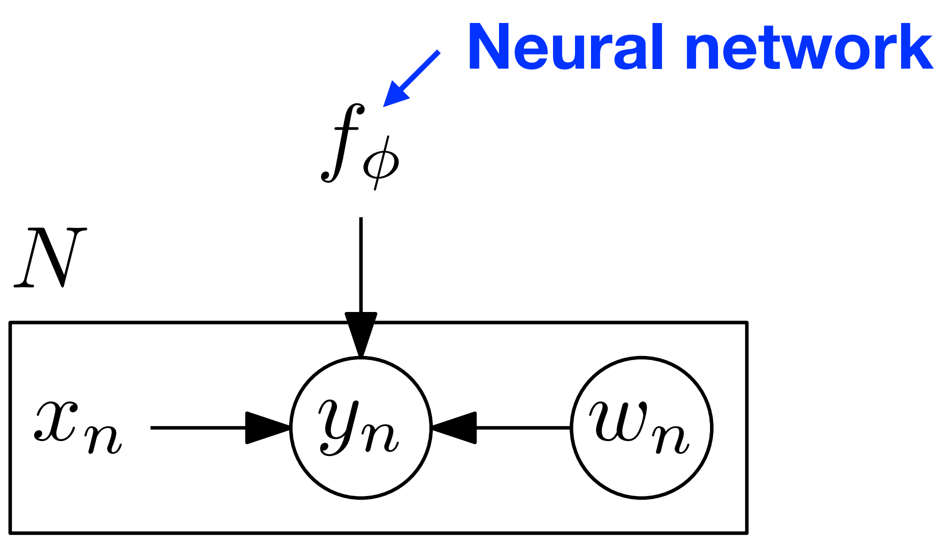
A possible approach



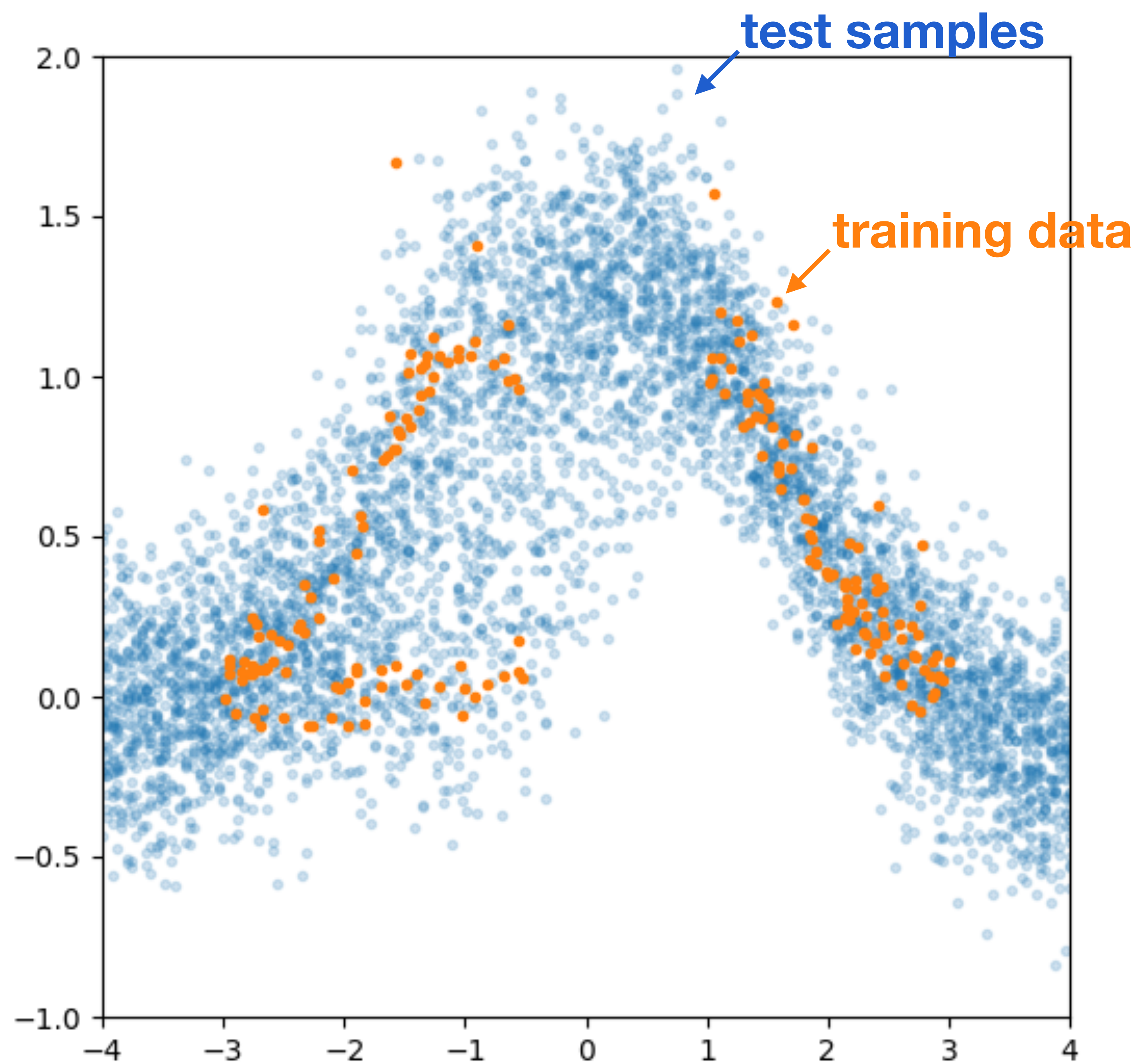
$$y_n = \mathcal{N}(f_\phi([x_n, w_n]), \sigma^2)$$
$$w_n \sim \mathcal{N}(0, 1)$$



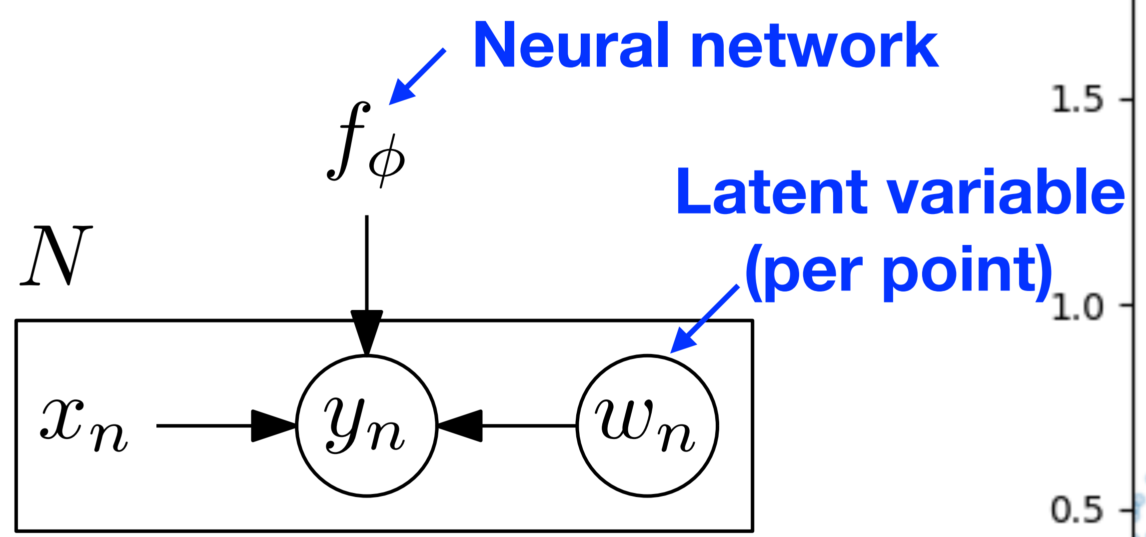
A possible approach



$$y_n = \mathcal{N}(f_\phi([x_n, w_n]), \sigma^2)$$
$$w_n \sim \mathcal{N}(0, 1)$$

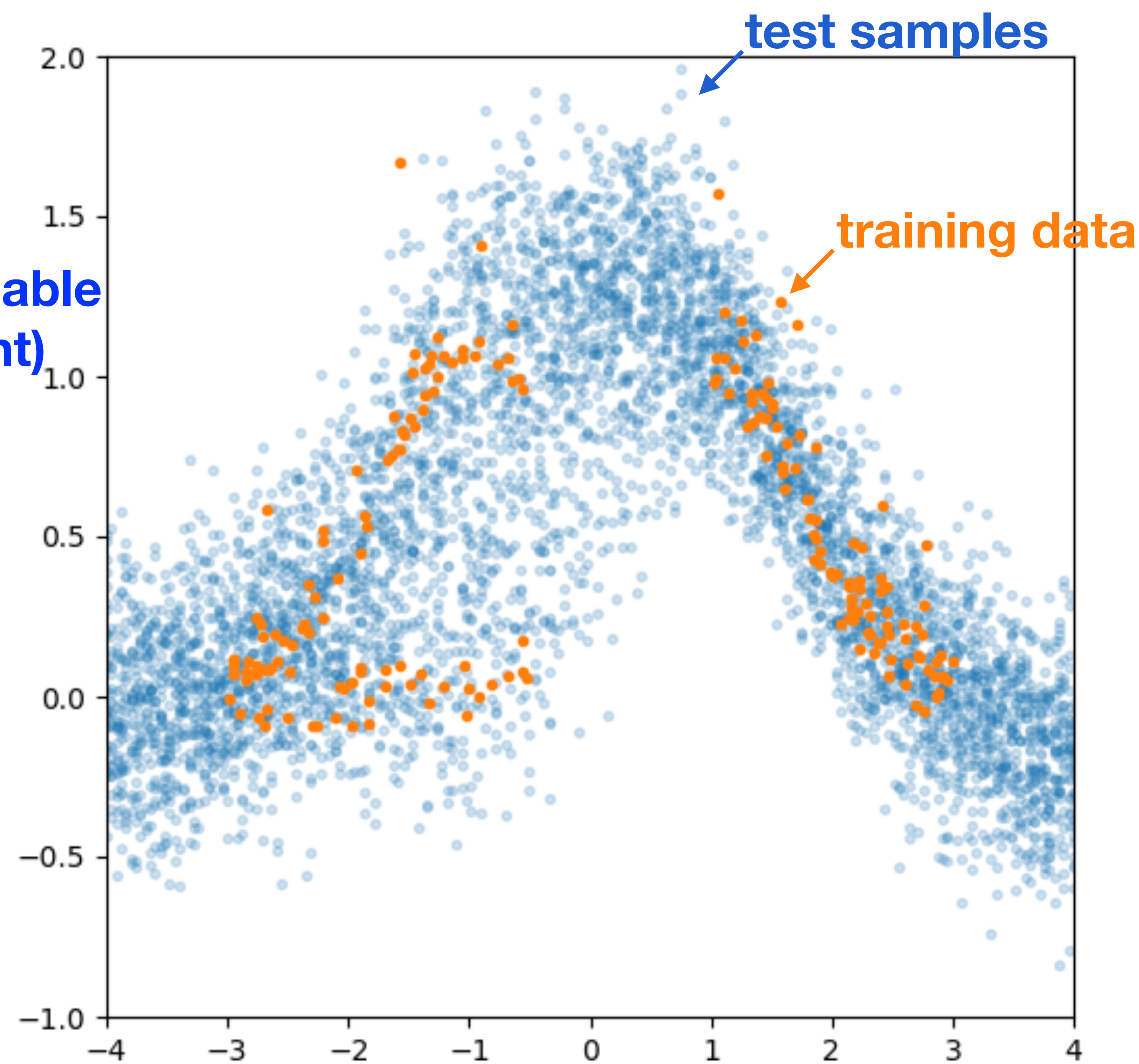


A possible approach

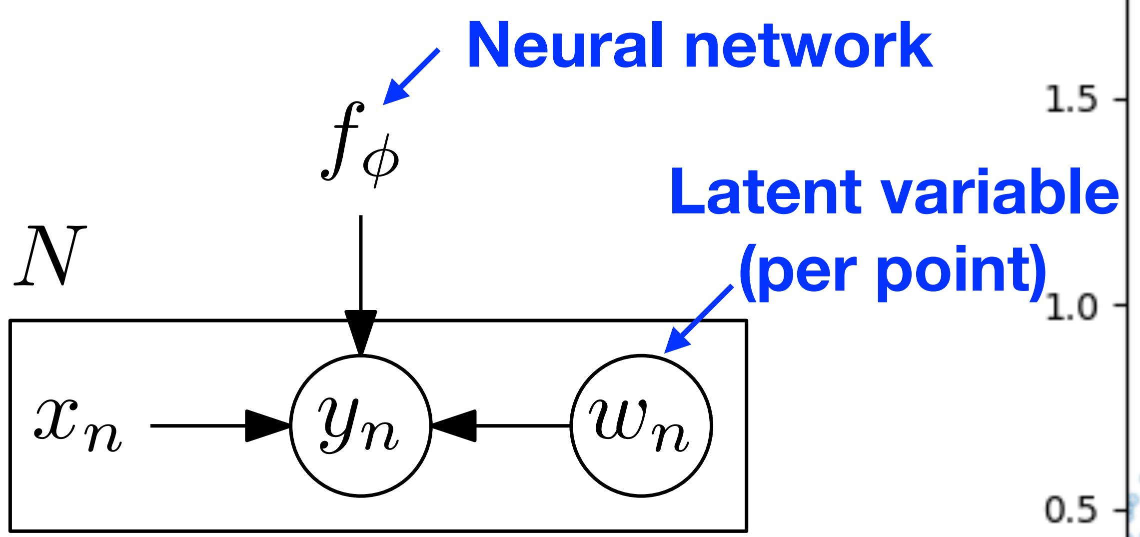


$$y_n = \mathcal{N}(f_\phi([x_n, w_n]), \sigma^2)$$

$$w_n \sim \mathcal{N}(0, 1)$$



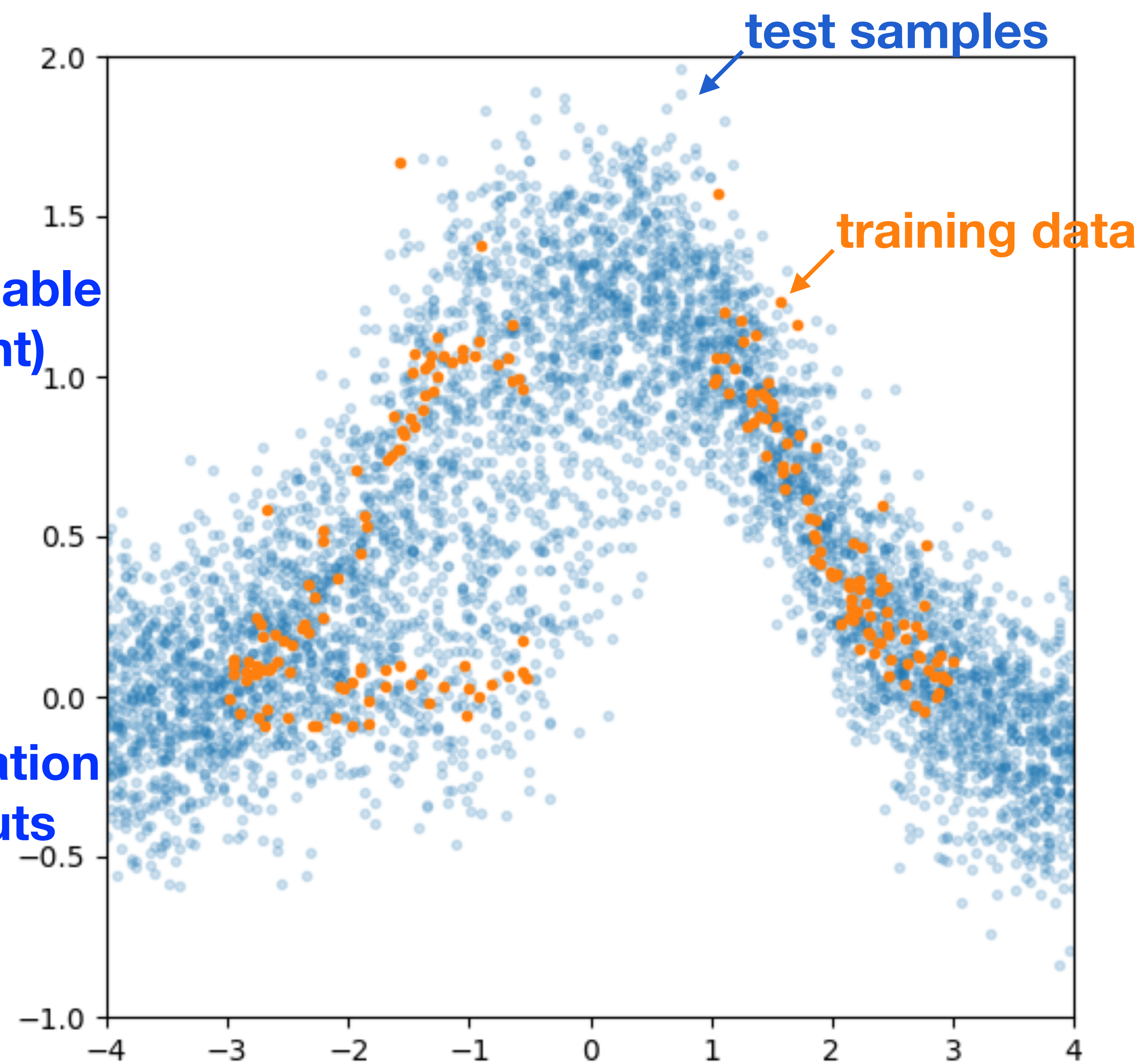
A possible approach



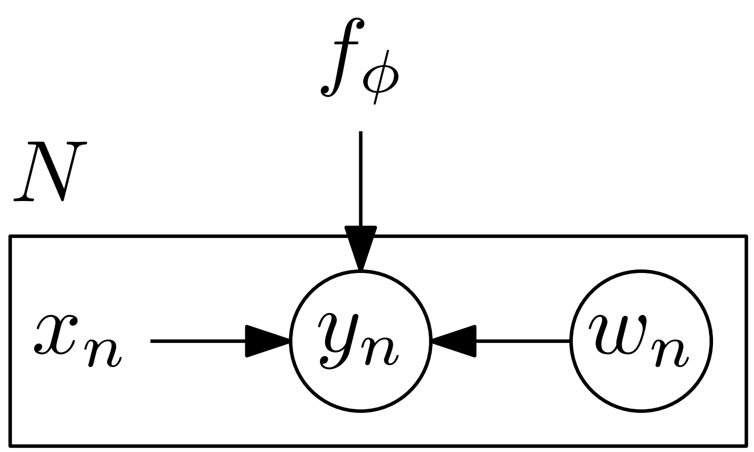
$$y_n = \mathcal{N}(f_\phi([x_n, w_n]), \sigma^2)$$

$$w_n \sim \mathcal{N}(0, 1)$$

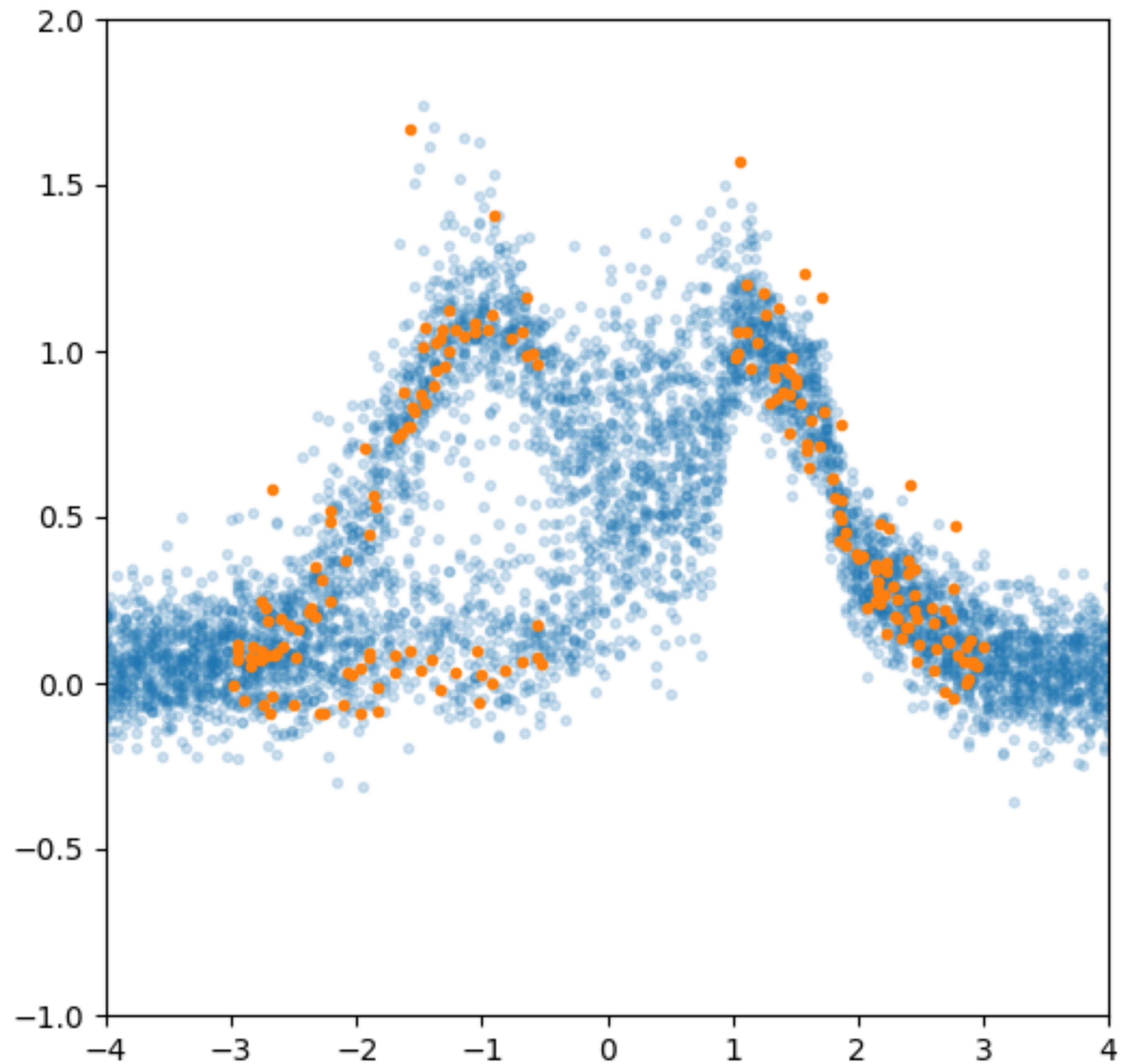
Concatenation with inputs



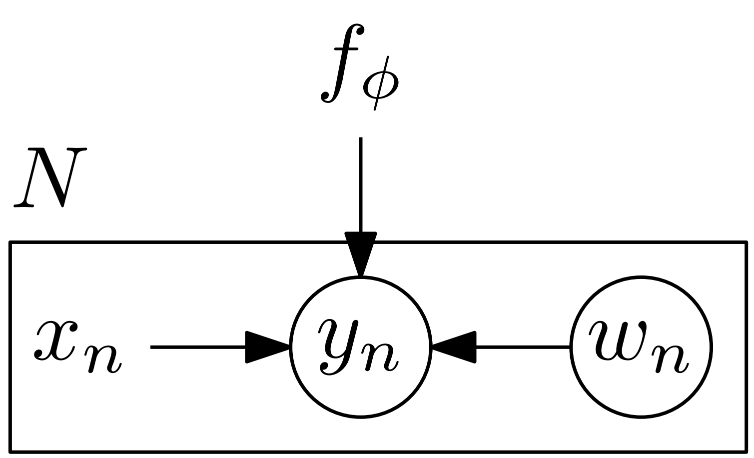
A possible approach



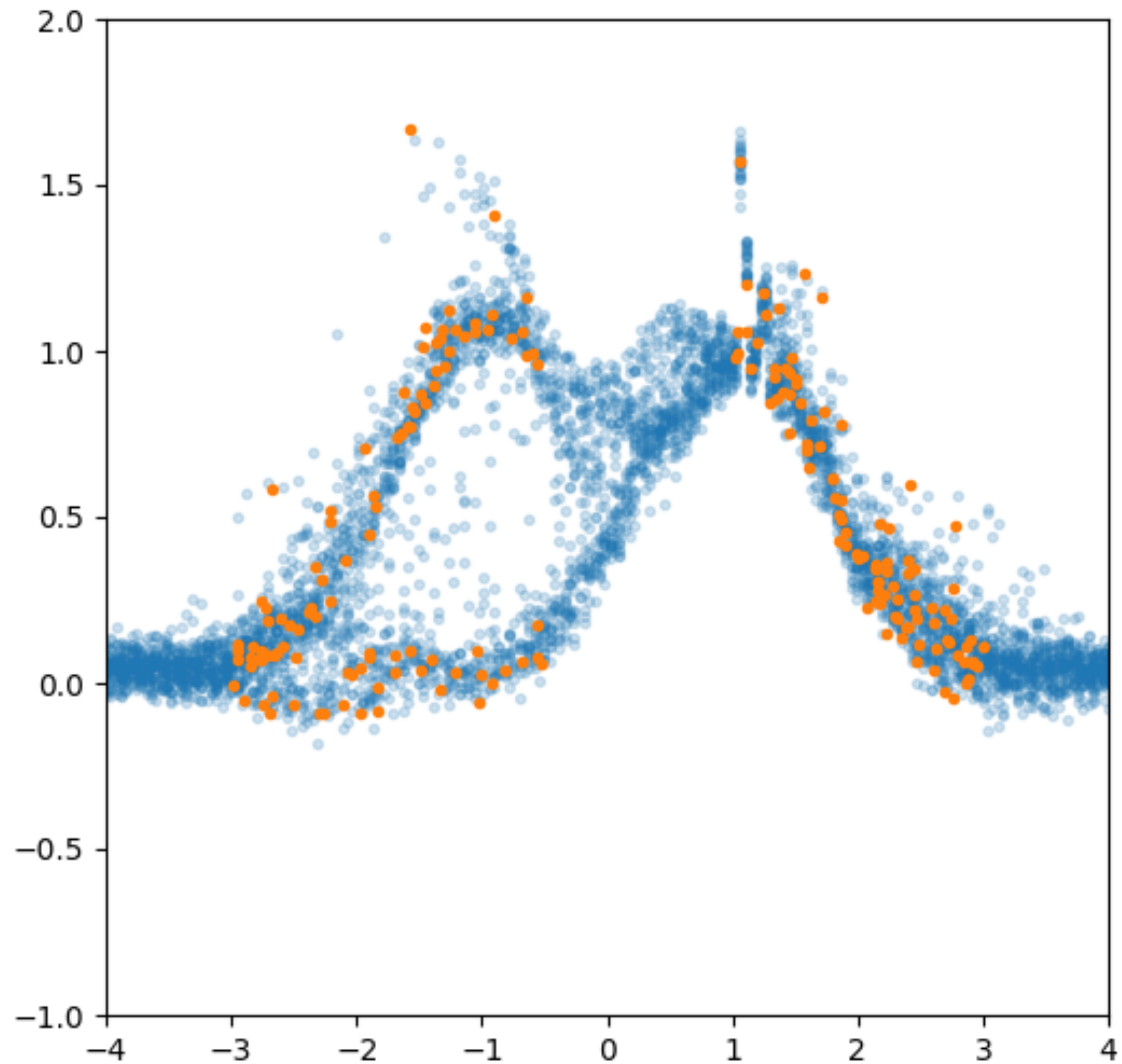
$$y_n = \mathcal{N}(f_\phi([x_n, w_n]), \sigma^2)$$
$$w_n \sim \mathcal{N}(0, 1)$$



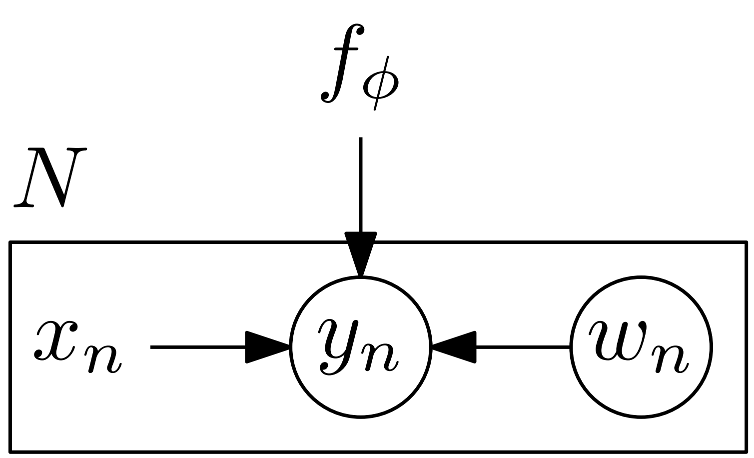
A possible approach



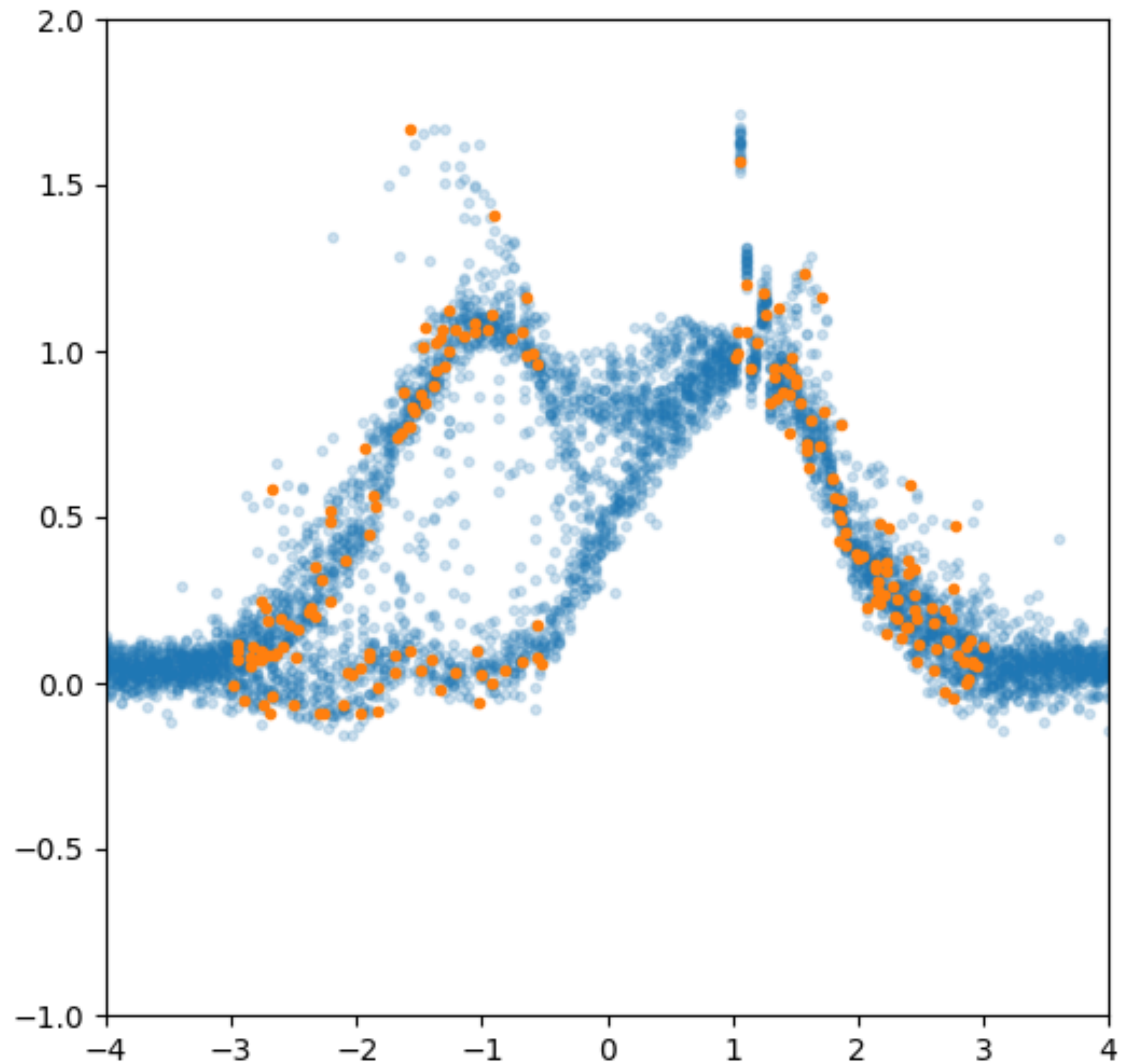
$$y_n = \mathcal{N}(f_\phi([x_n, w_n]), \sigma^2)$$
$$w_n \sim \mathcal{N}(0, 1)$$



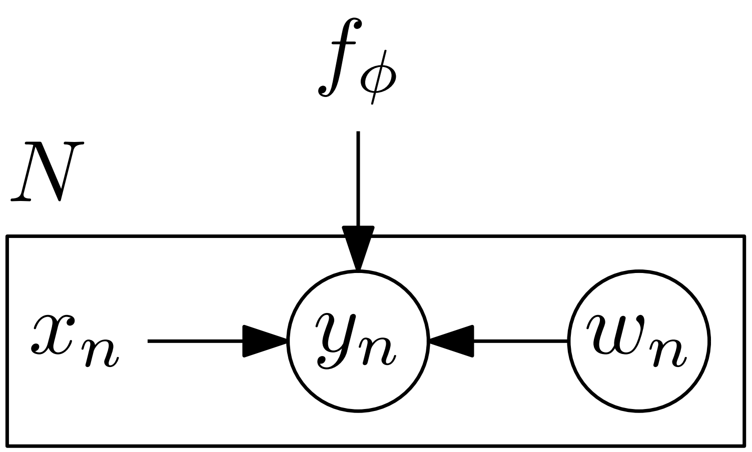
A possible approach



$$y_n = \mathcal{N}(f_\phi([x_n, w_n]), \sigma^2)$$
$$w_n \sim \mathcal{N}(0, 1)$$

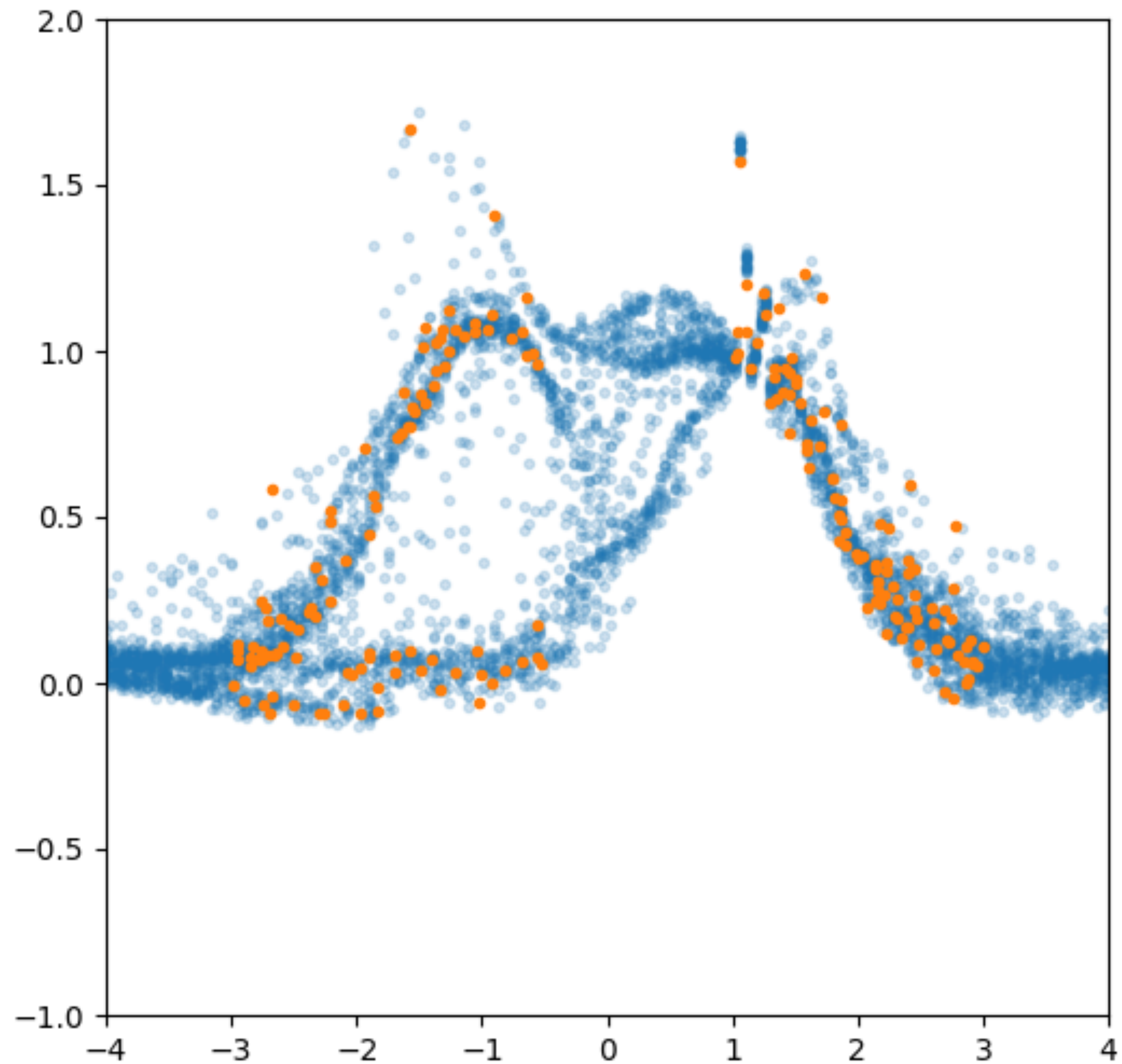


A possible approach

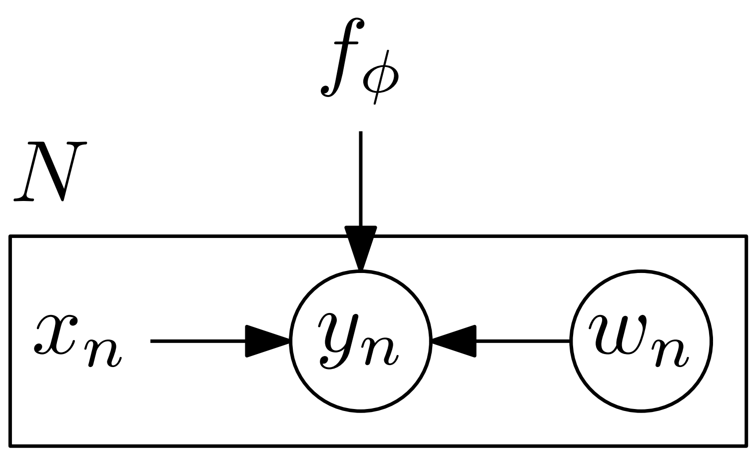


$$y_n = \mathcal{N}(f_\phi([x_n, w_n]), \sigma^2)$$

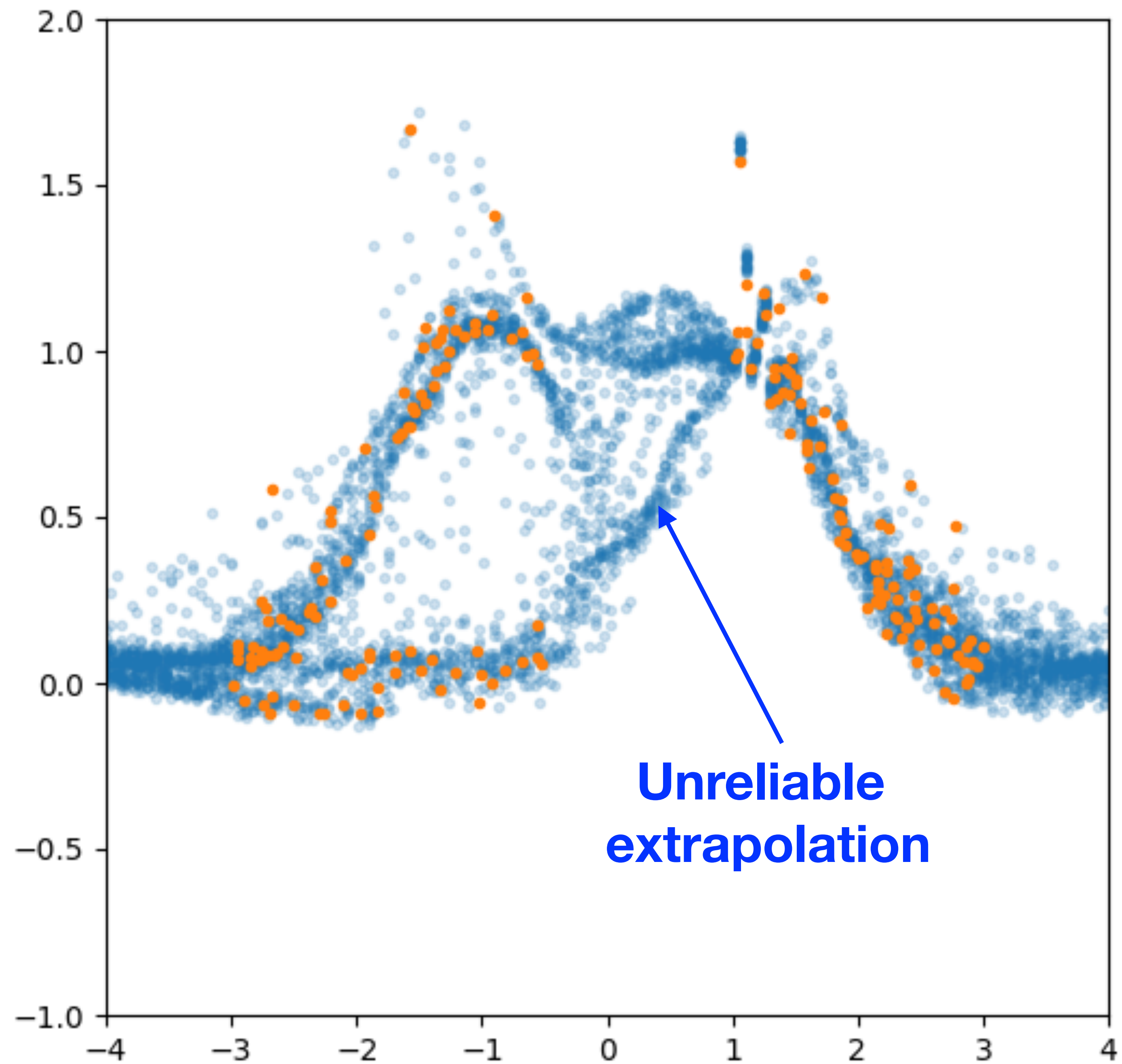
$$w_n \sim \mathcal{N}(0, 1)$$



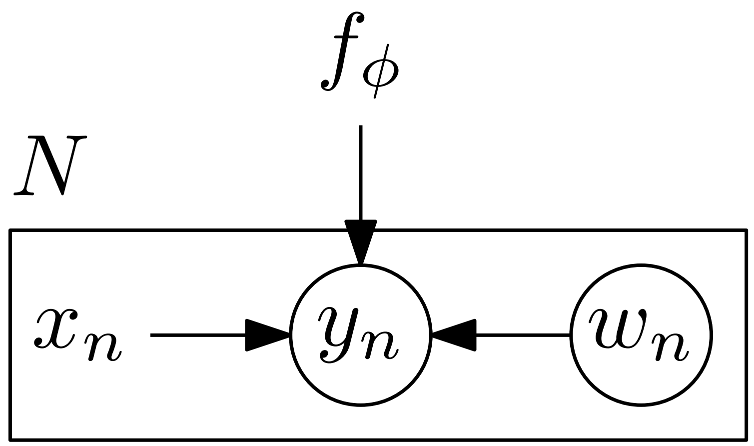
A possible approach



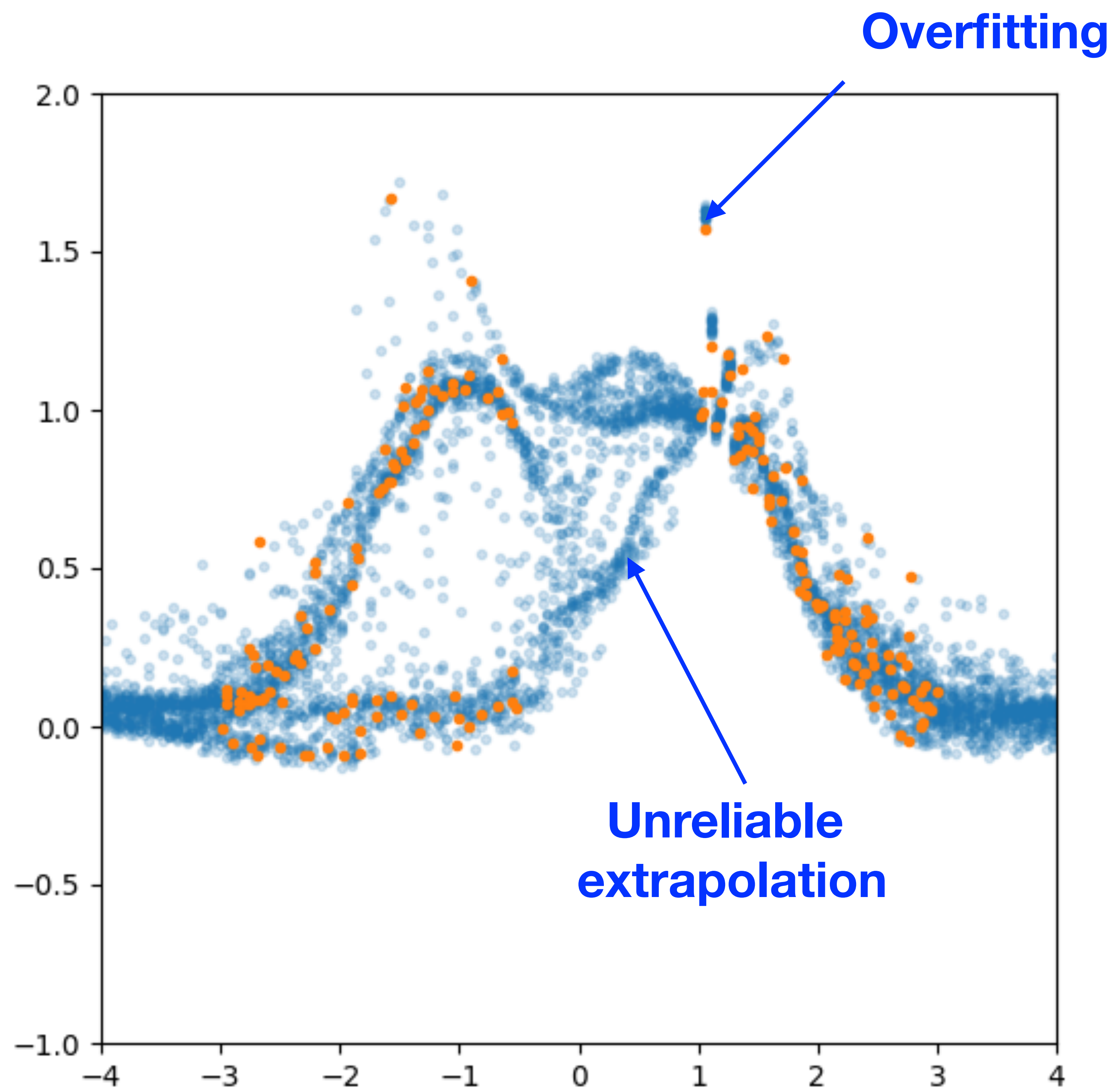
$$y_n = \mathcal{N}(f_\phi([x_n, w_n]), \sigma^2)$$
$$w_n \sim \mathcal{N}(0, 1)$$



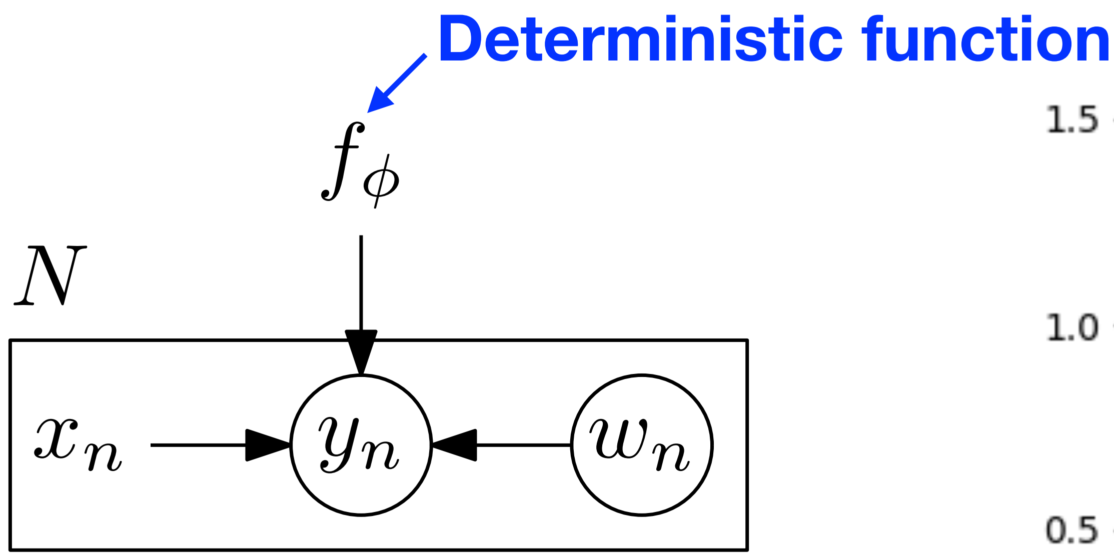
A possible approach



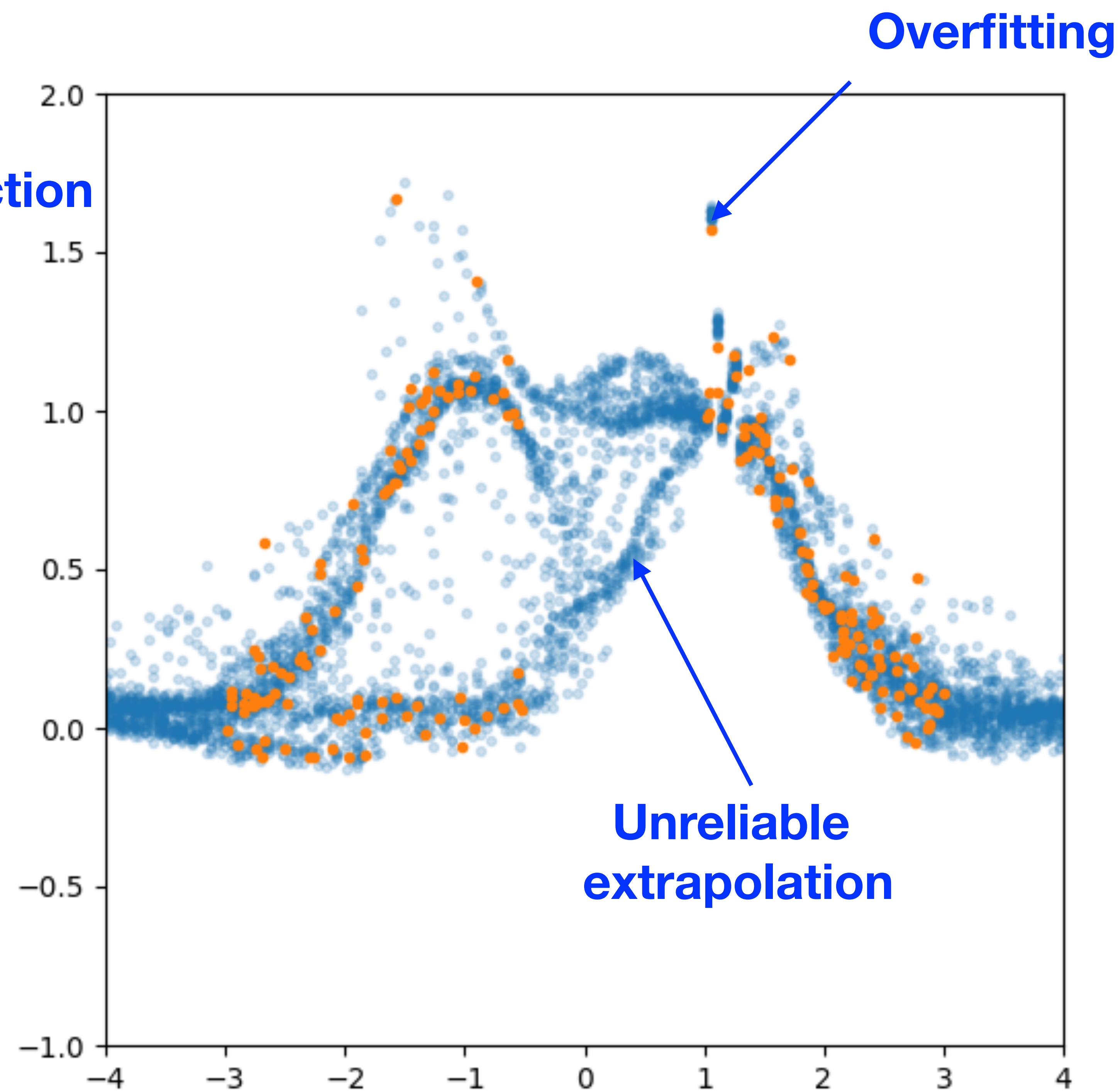
$$y_n = \mathcal{N}(f_\phi([x_n, w_n]), \sigma^2)$$
$$w_n \sim \mathcal{N}(0, 1)$$



A possible approach

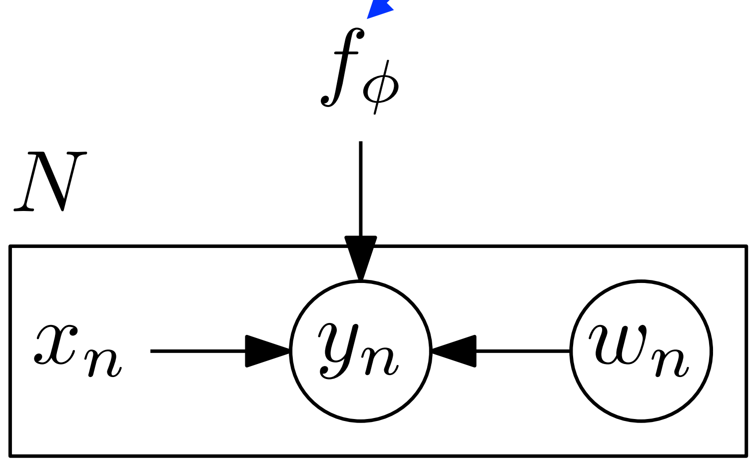


$$y_n = \mathcal{N}(f_\phi([x_n, w_n]), \sigma^2)$$
$$w_n \sim \mathcal{N}(0, 1)$$



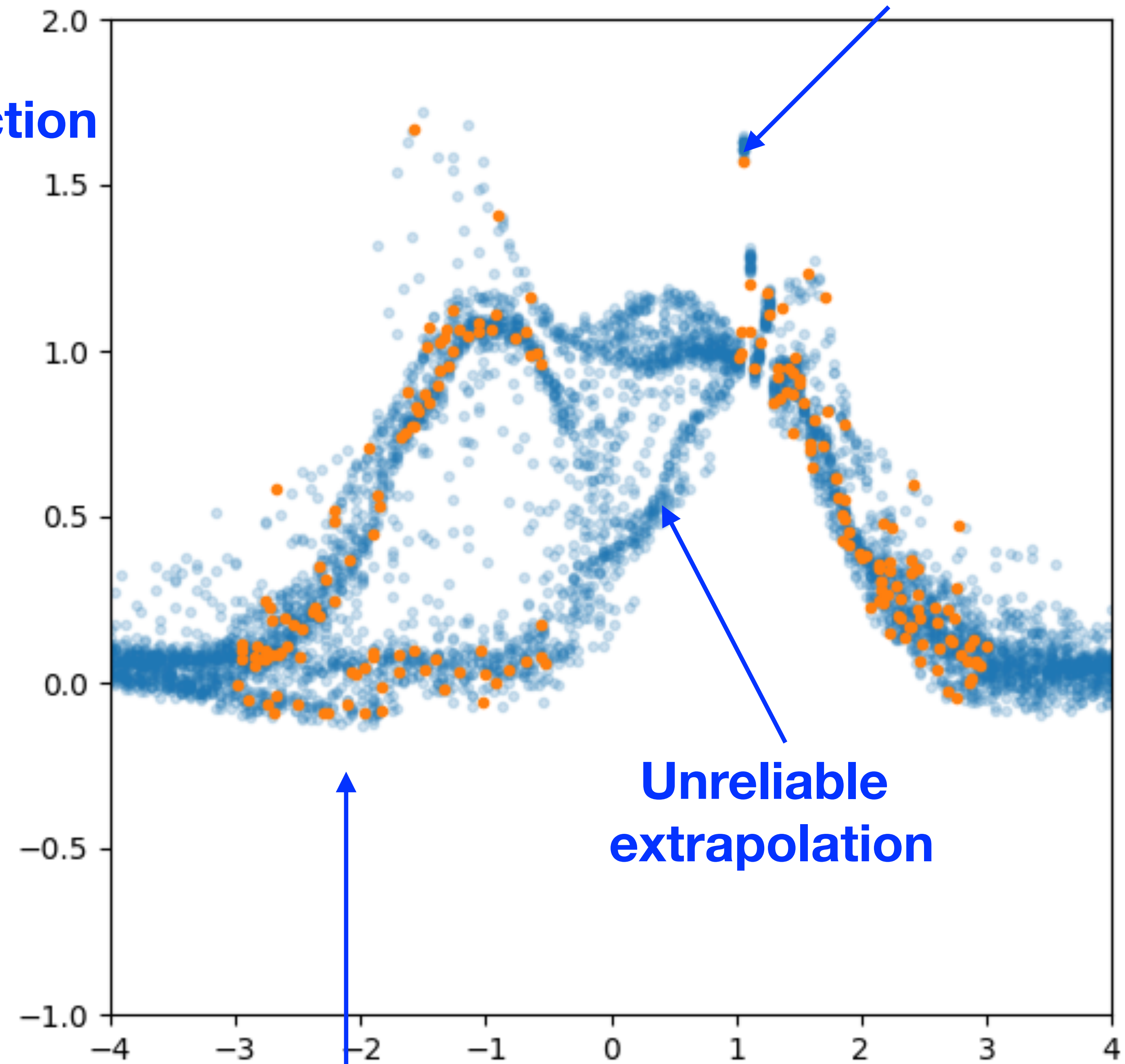
A possible approach

Deterministic function



$$y_n = \mathcal{N}(f_\phi([x_n, w_n]), \sigma^2)$$
$$w_n \sim \mathcal{N}(0,1)$$

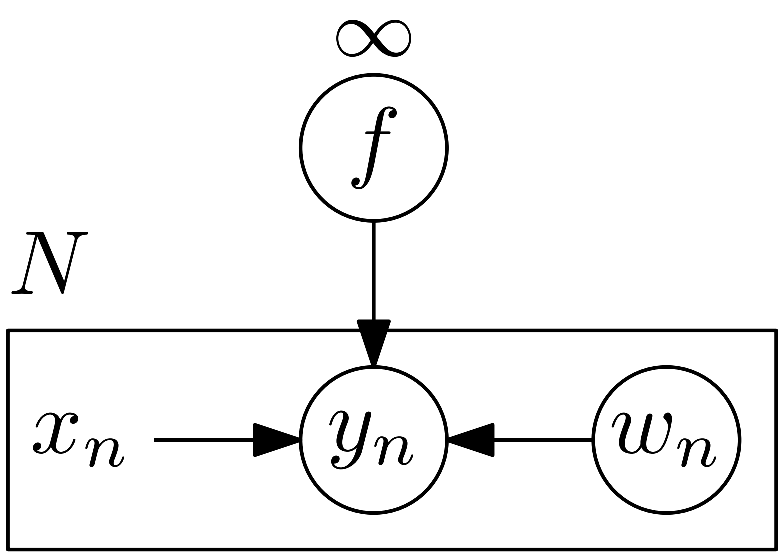
Overfitting



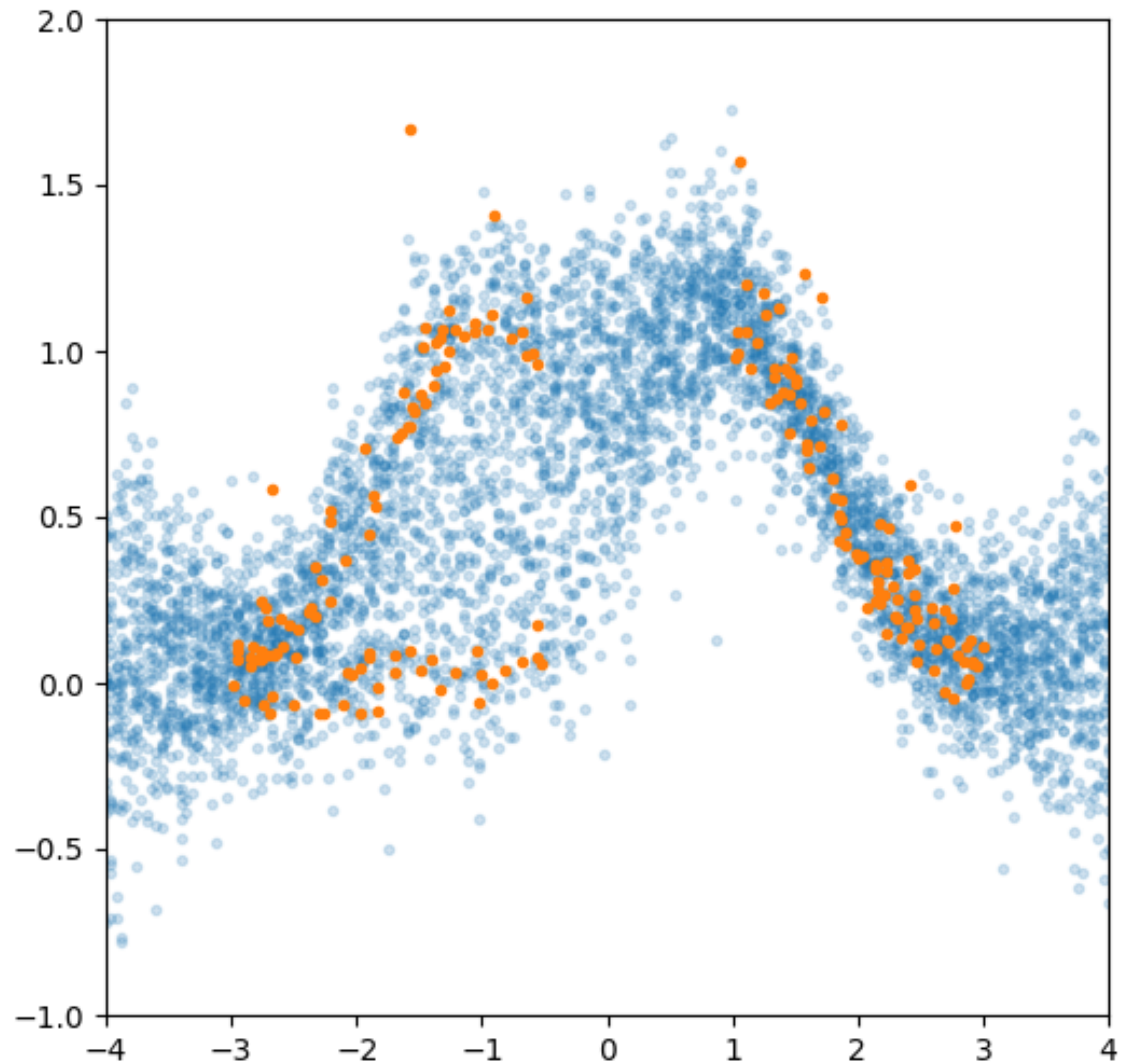
Unreliable extrapolation

Small number of examples per input x_n

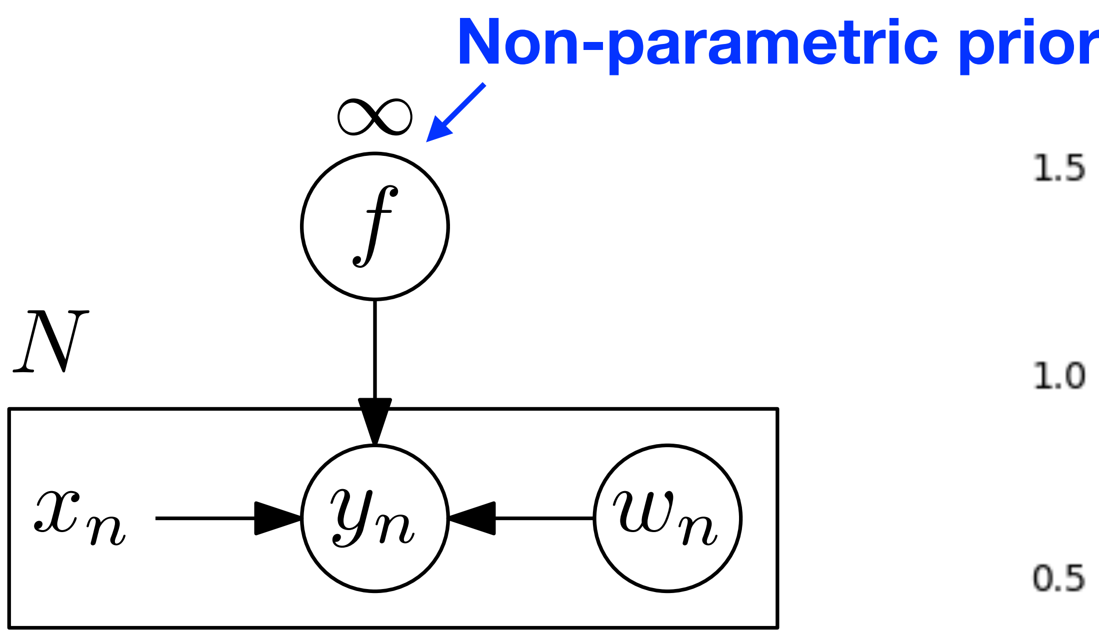
Another possible approach



$$y_n = \mathcal{N}(f([x_n, w_n]), \sigma^2)$$
$$w_n \sim \mathcal{N}(0, 1)$$
$$f \sim \mathcal{GP}(\mu, k)$$



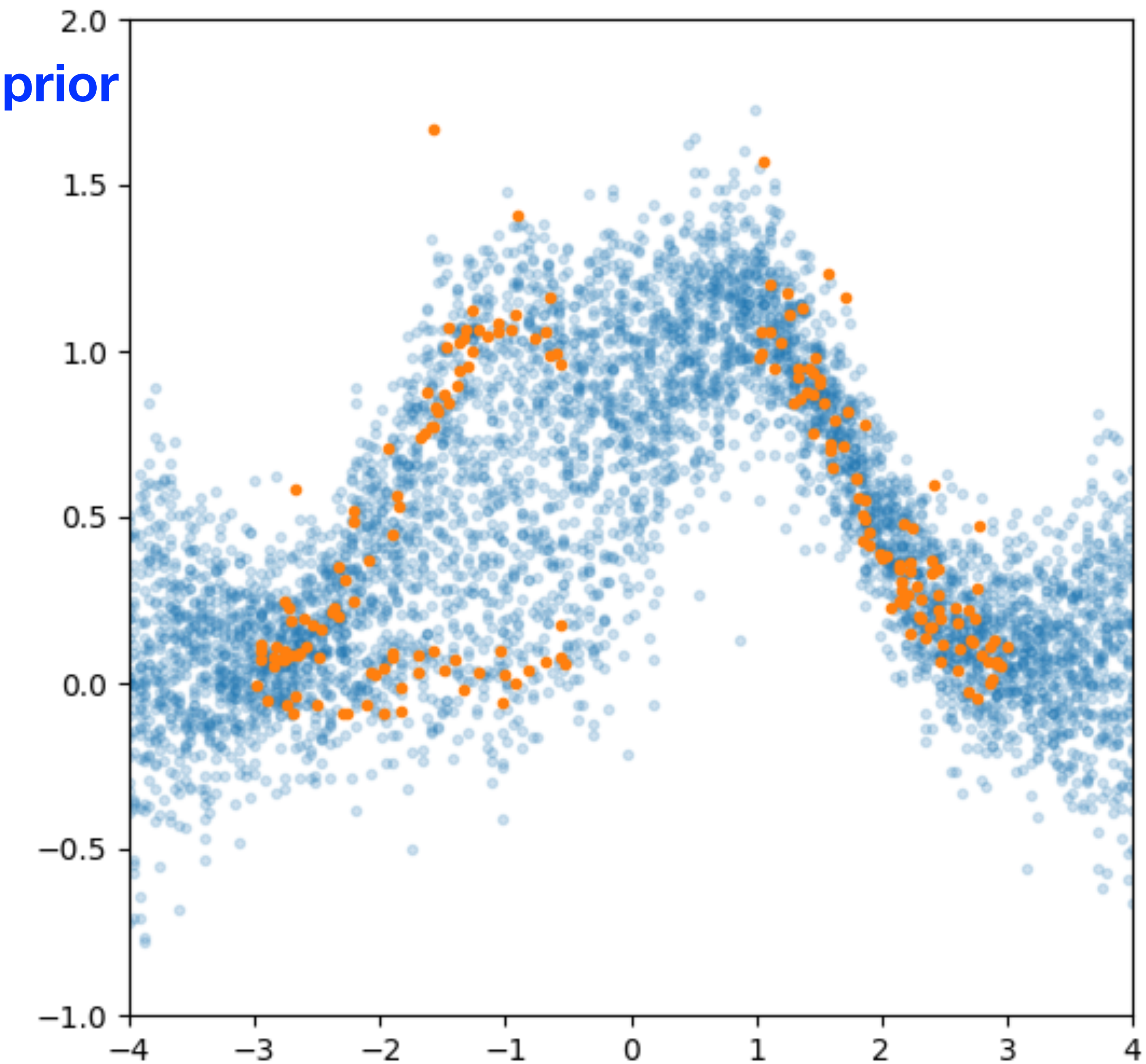
Another possible approach



$$y_n = \mathcal{N}(f([x_n, w_n]), \sigma^2)$$

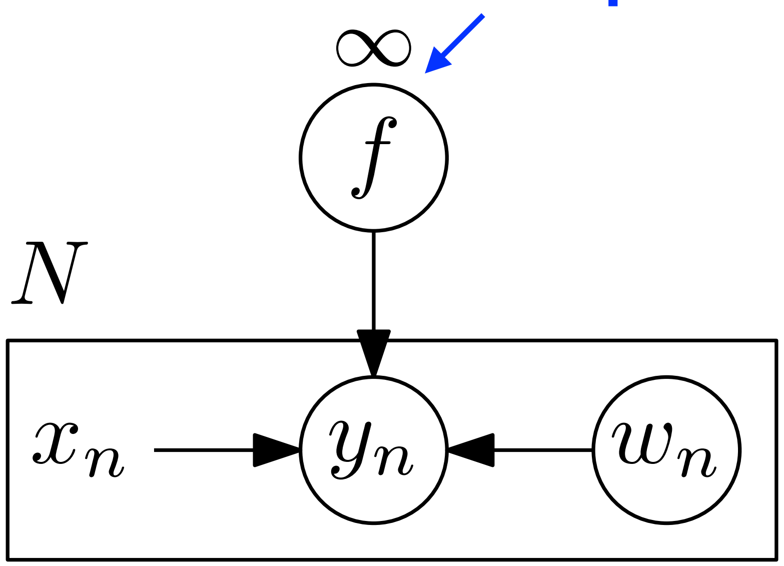
$$w_n \sim \mathcal{N}(0, 1)$$

$$\boxed{f \sim \mathcal{GP}(\mu, k)}$$



Another possible approach

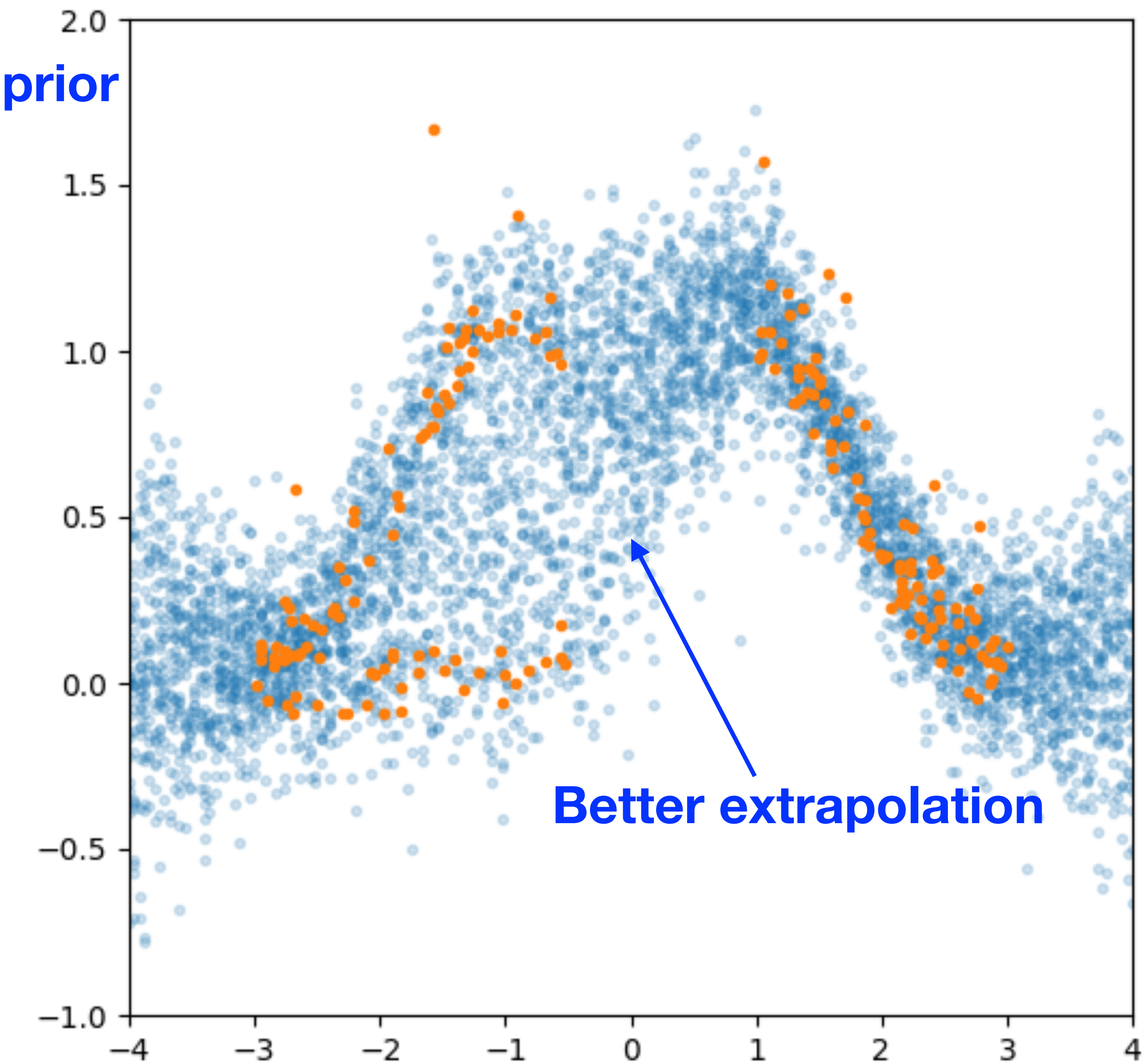
Non-parametric prior



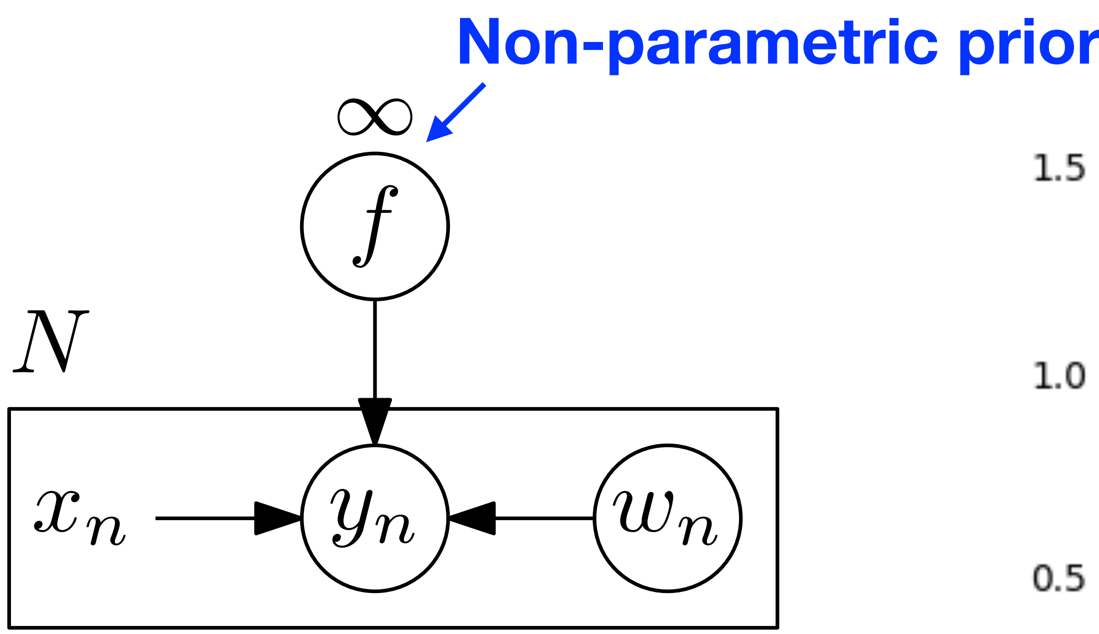
$$y_n = \mathcal{N}(f([x_n, w_n]), \sigma^2)$$

$$w_n \sim \mathcal{N}(0, 1)$$

$$\boxed{f \sim \mathcal{GP}(\mu, k)}$$



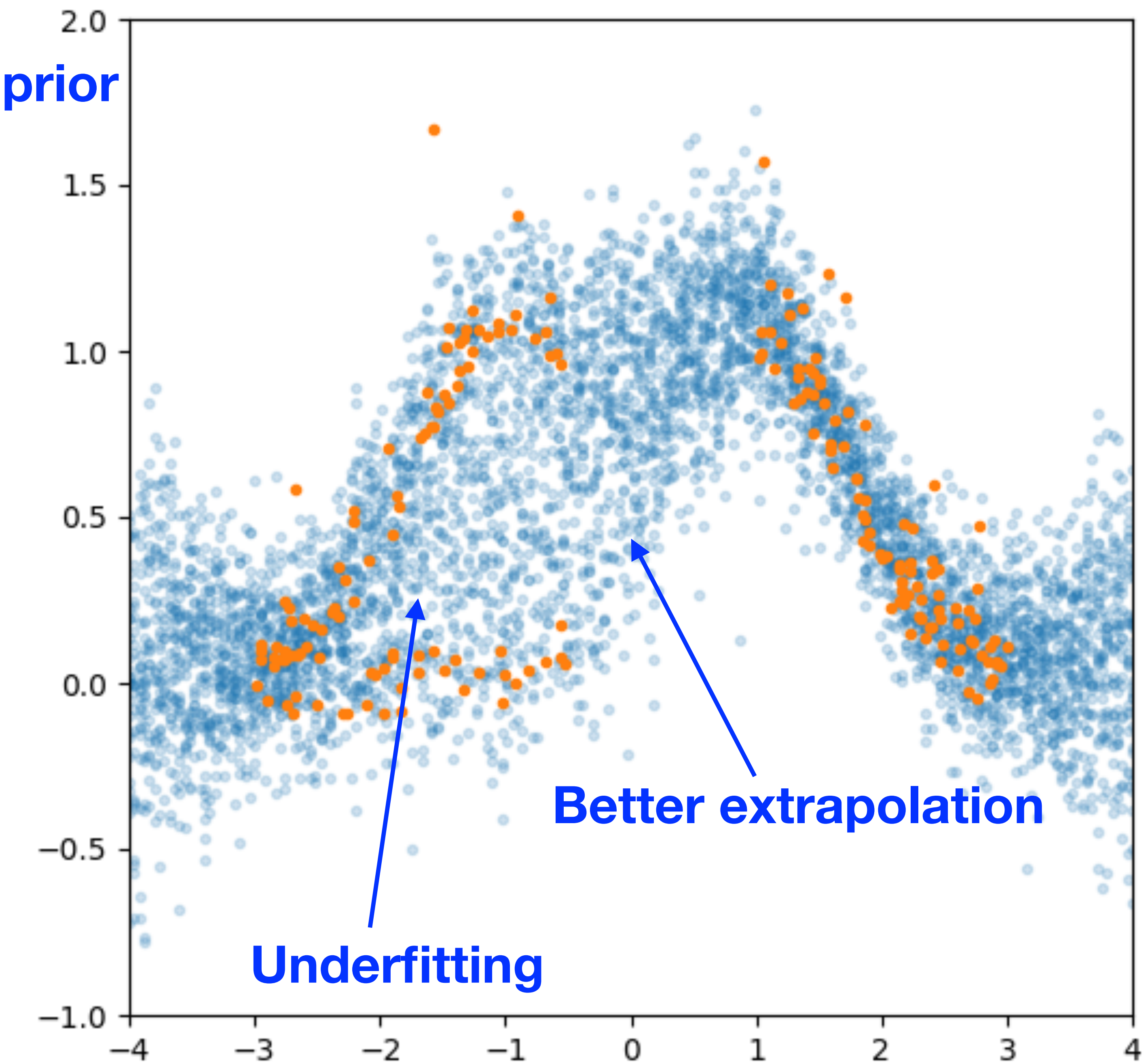
Another possible approach



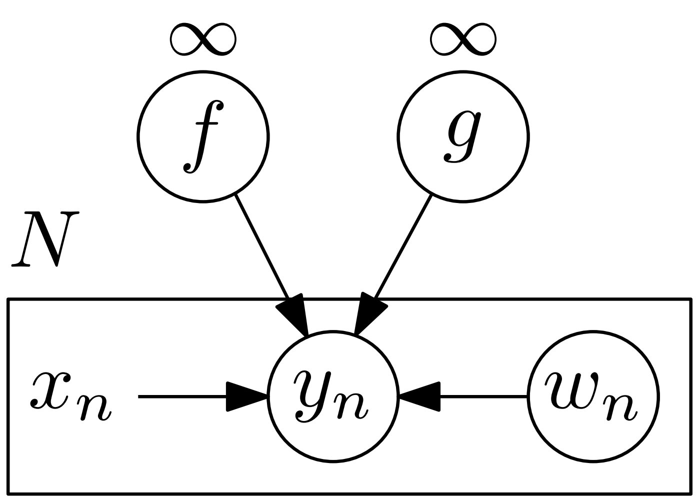
$$y_n = \mathcal{N}(f([x_n, w_n]), \sigma^2)$$

$$w_n \sim \mathcal{N}(0, 1)$$

$$f \sim \mathcal{GP}(\mu, k)$$



Our model

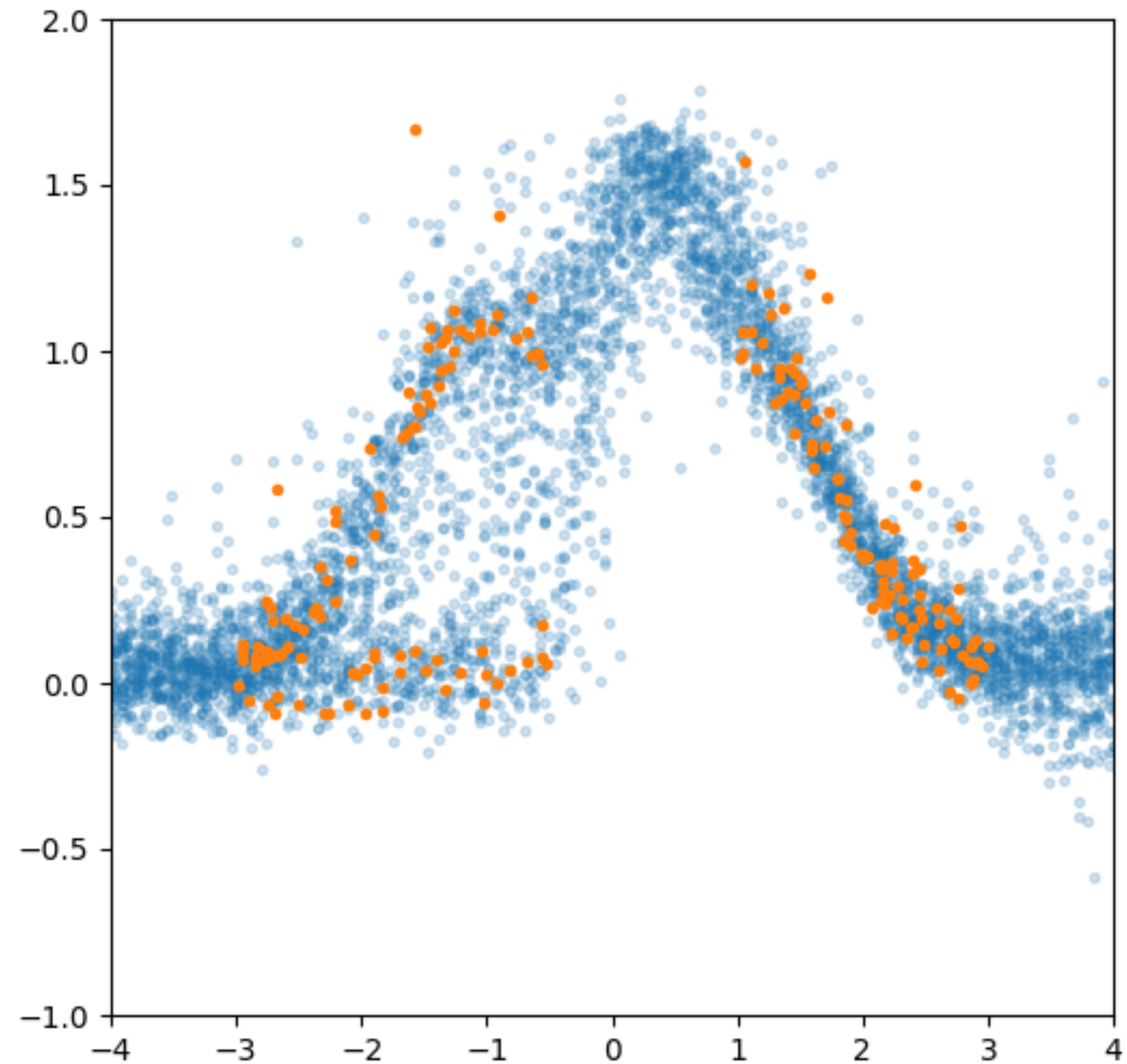


$$y_n = \mathcal{N}(f(g([x_n, w_n])), \sigma^2)$$

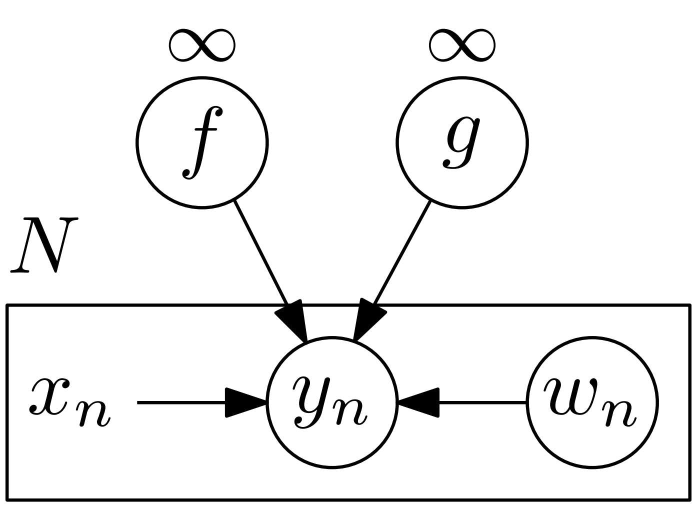
$$w_n \sim \mathcal{N}(0, 1)$$

$$f \sim \mathcal{GP}(\mu_1, k_1)$$

$$g \sim \mathcal{GP}(\mu_2, k_2)$$



Our model

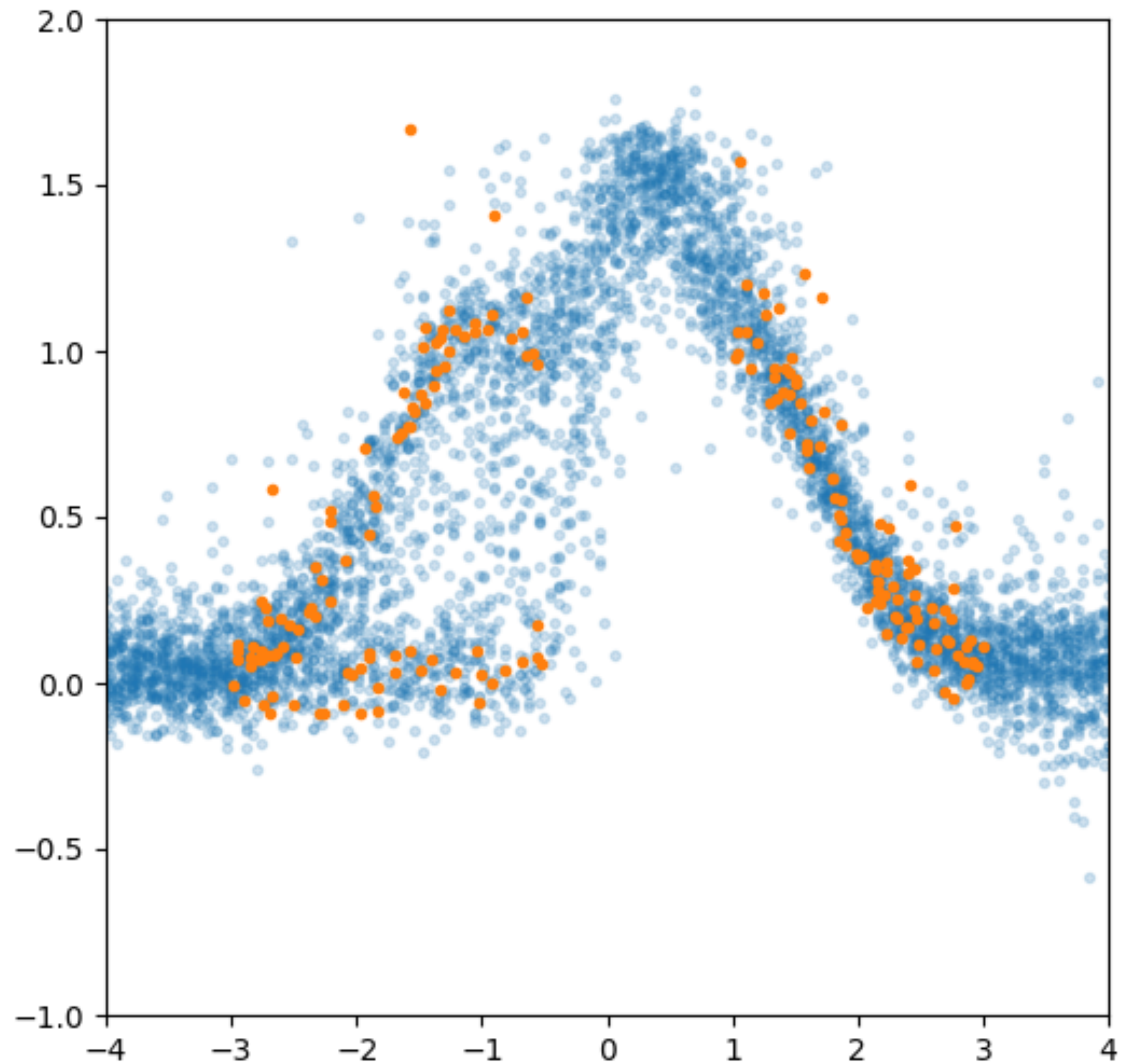


$$y_n = \mathcal{N}(f(g([x_n, w_n])), \sigma^2)$$

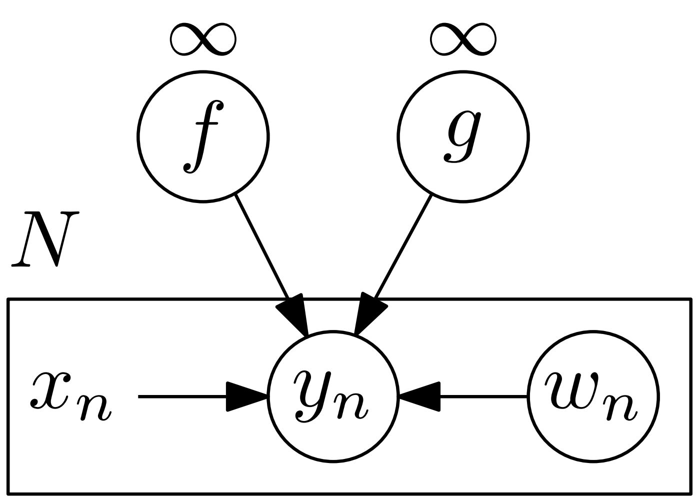
$$w_n \sim \mathcal{N}(0, 1)$$

$$f \sim \mathcal{GP}(\mu_1, k_1)$$

$$g \sim \mathcal{GP}(\mu_2, k_2)$$



Our model

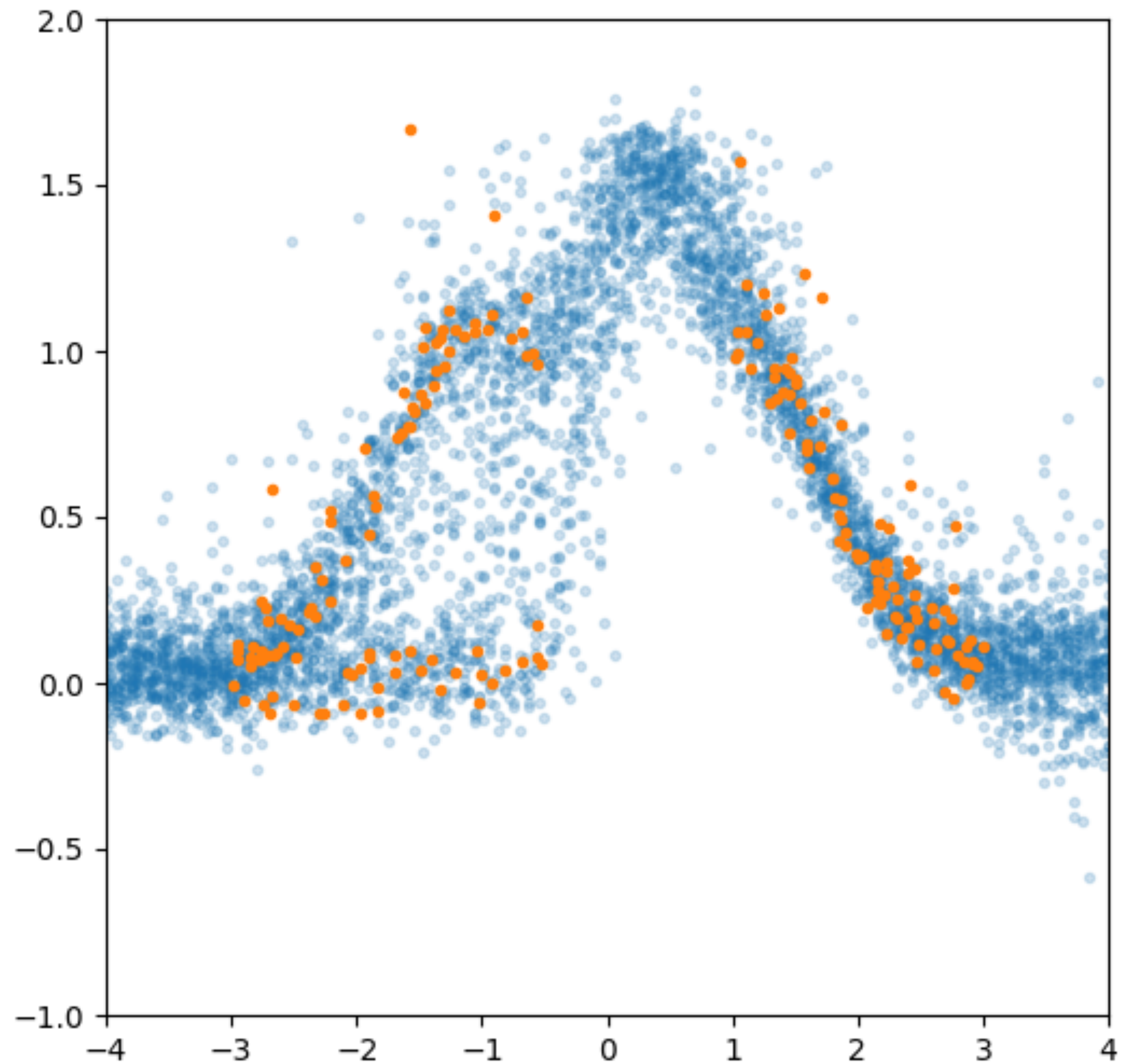


$$y_n = \mathcal{N}(f(g([x_n, w_n])), \sigma^2)$$

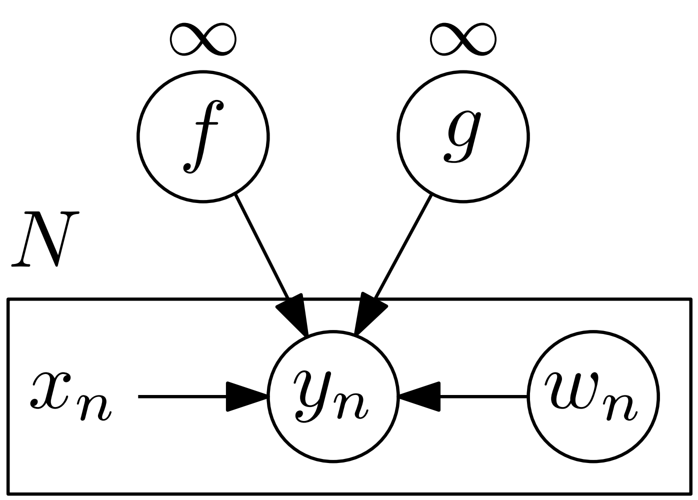
$$w_n \sim \mathcal{N}(0, 1)$$

$$f \sim \mathcal{GP}(\mu_1, k_1)$$

$$g \sim \mathcal{GP}(\mu_2, k_2)$$



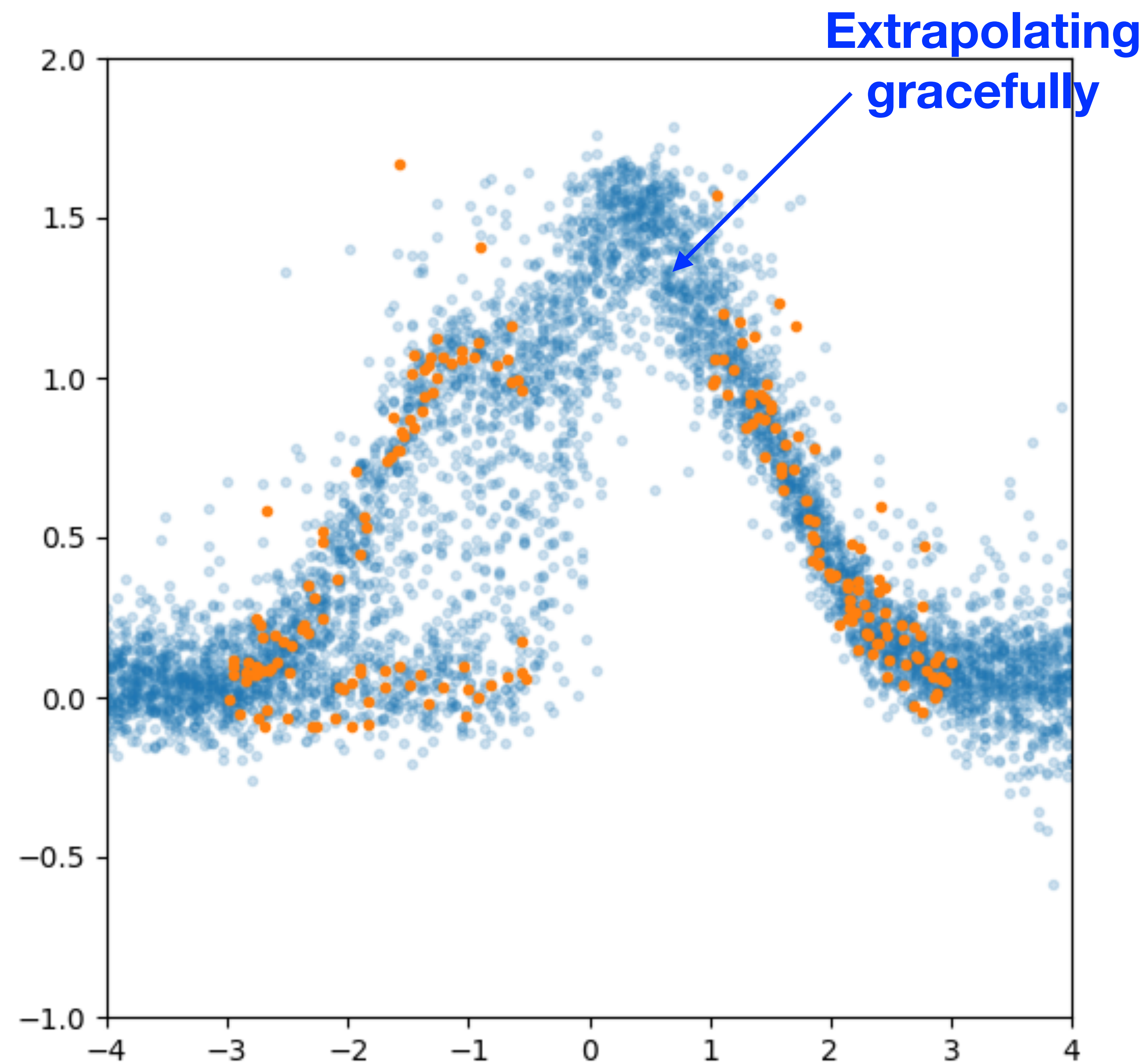
Our model



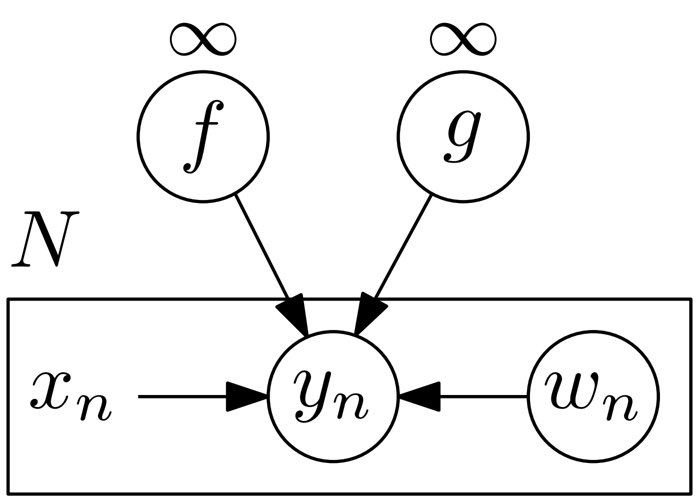
$$y_n = \mathcal{N}(\boxed{f(g}([x_n, w_n])), \sigma^2)$$

$$w_n \sim \mathcal{N}(0, 1)$$

$$\boxed{\begin{aligned} f &\sim \mathcal{GP}(\mu_1, k_1) \\ g &\sim \mathcal{GP}(\mu_2, k_2) \end{aligned}}$$



Our model

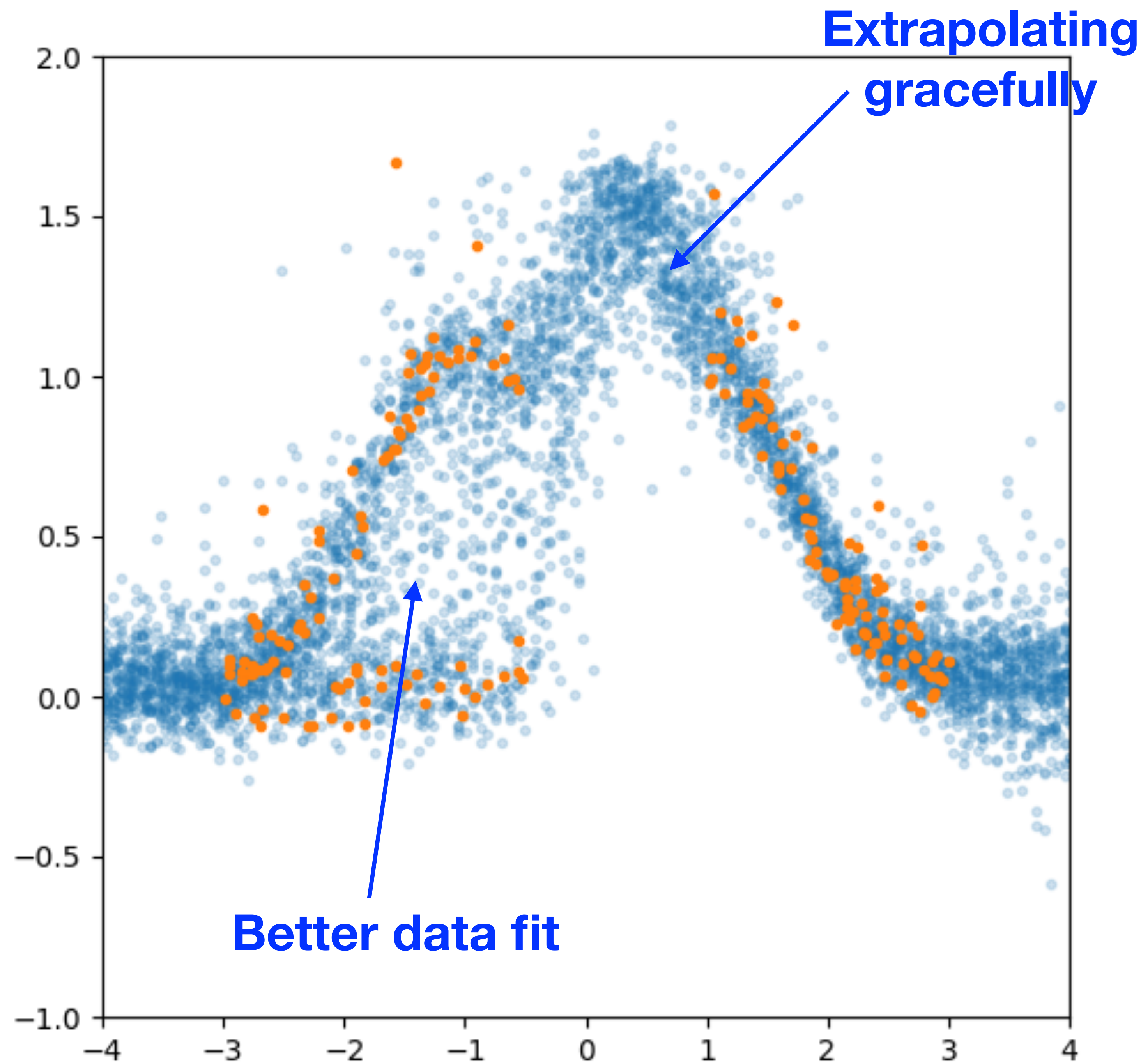


$$y_n = \mathcal{N}(f(g([x_n, w_n])), \sigma^2)$$

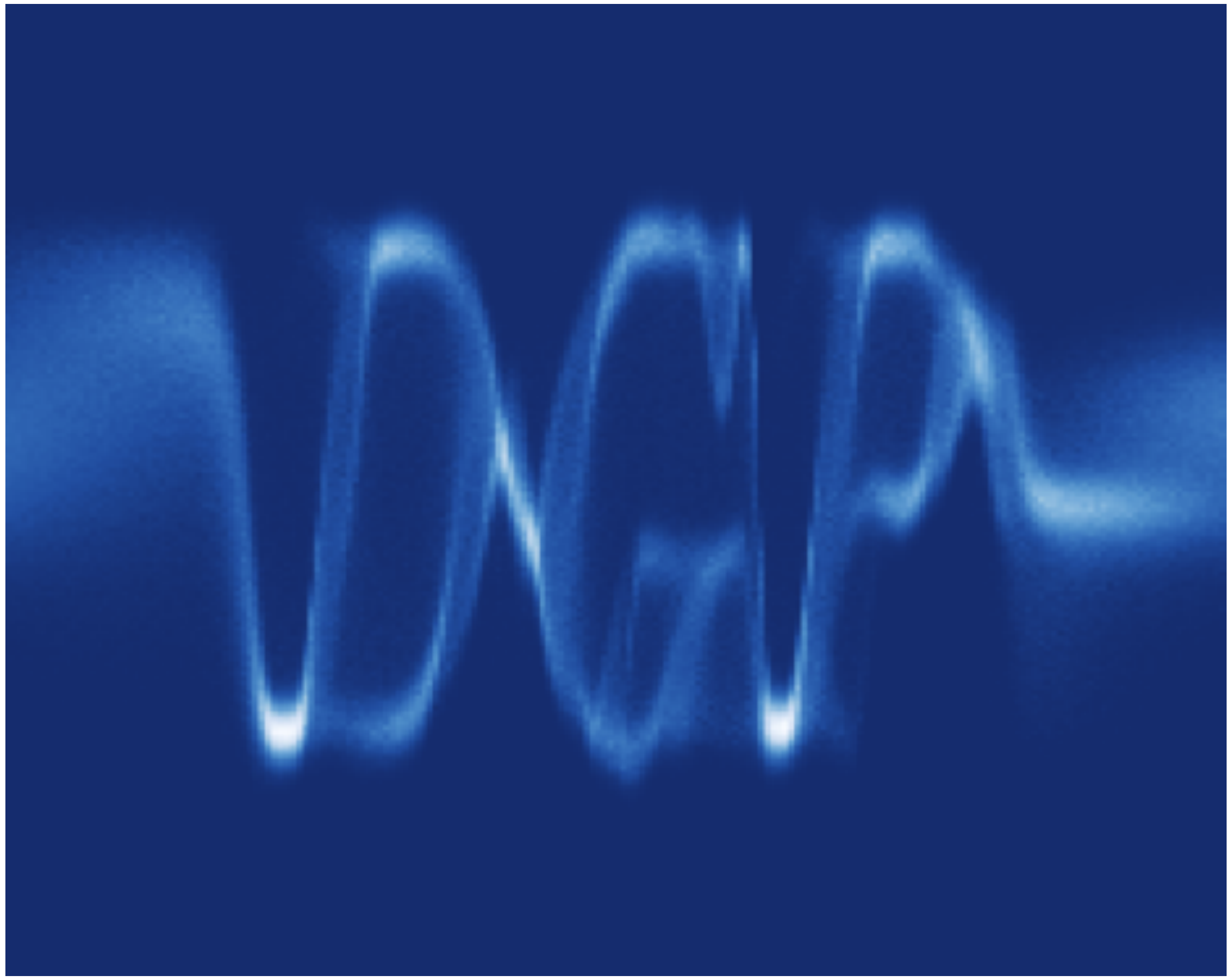
$$w_n \sim \mathcal{N}(0, 1)$$

$$f \sim \mathcal{GP}(\mu_1, k_1)$$

$$g \sim \mathcal{GP}(\mu_2, k_2)$$

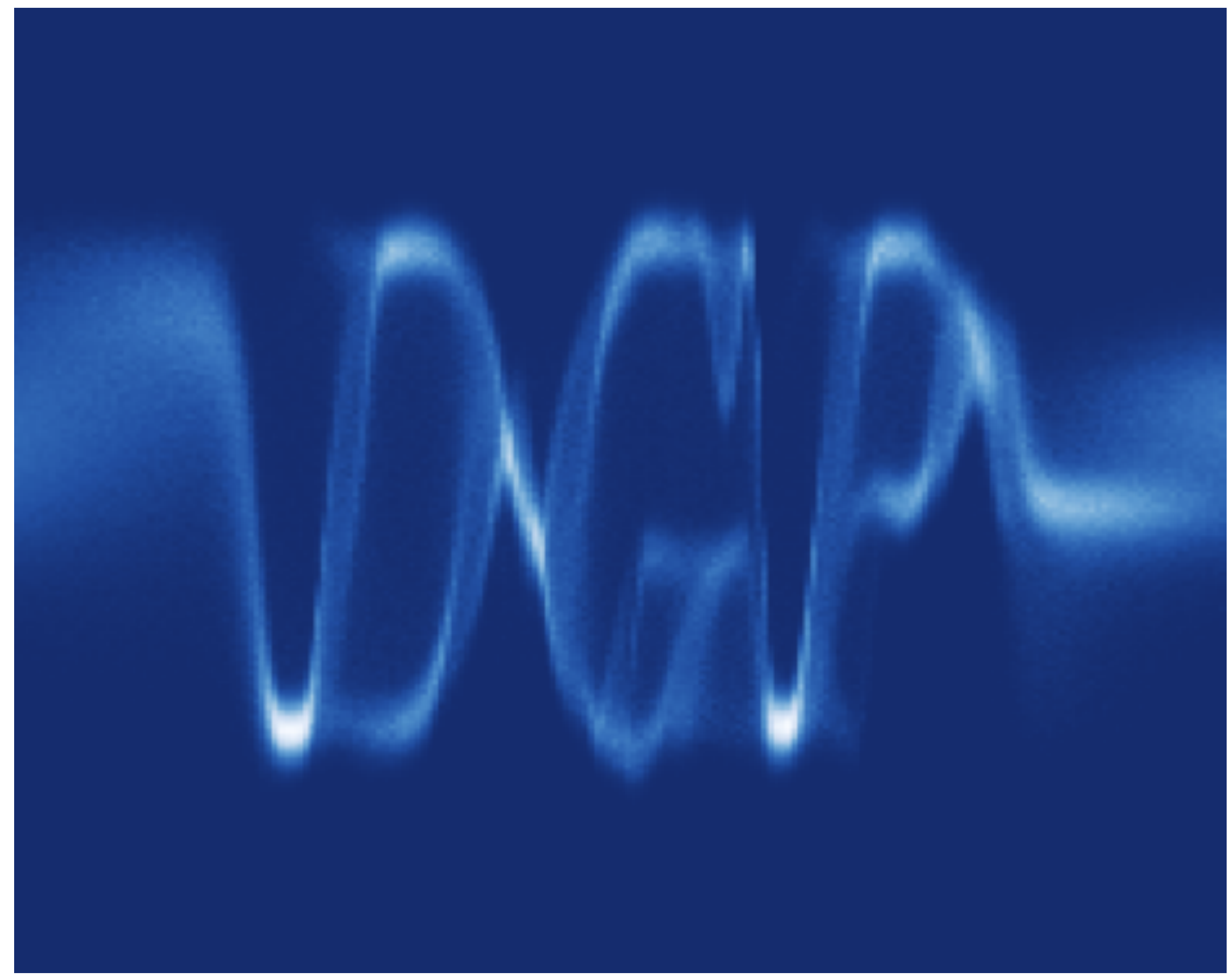


Contributions



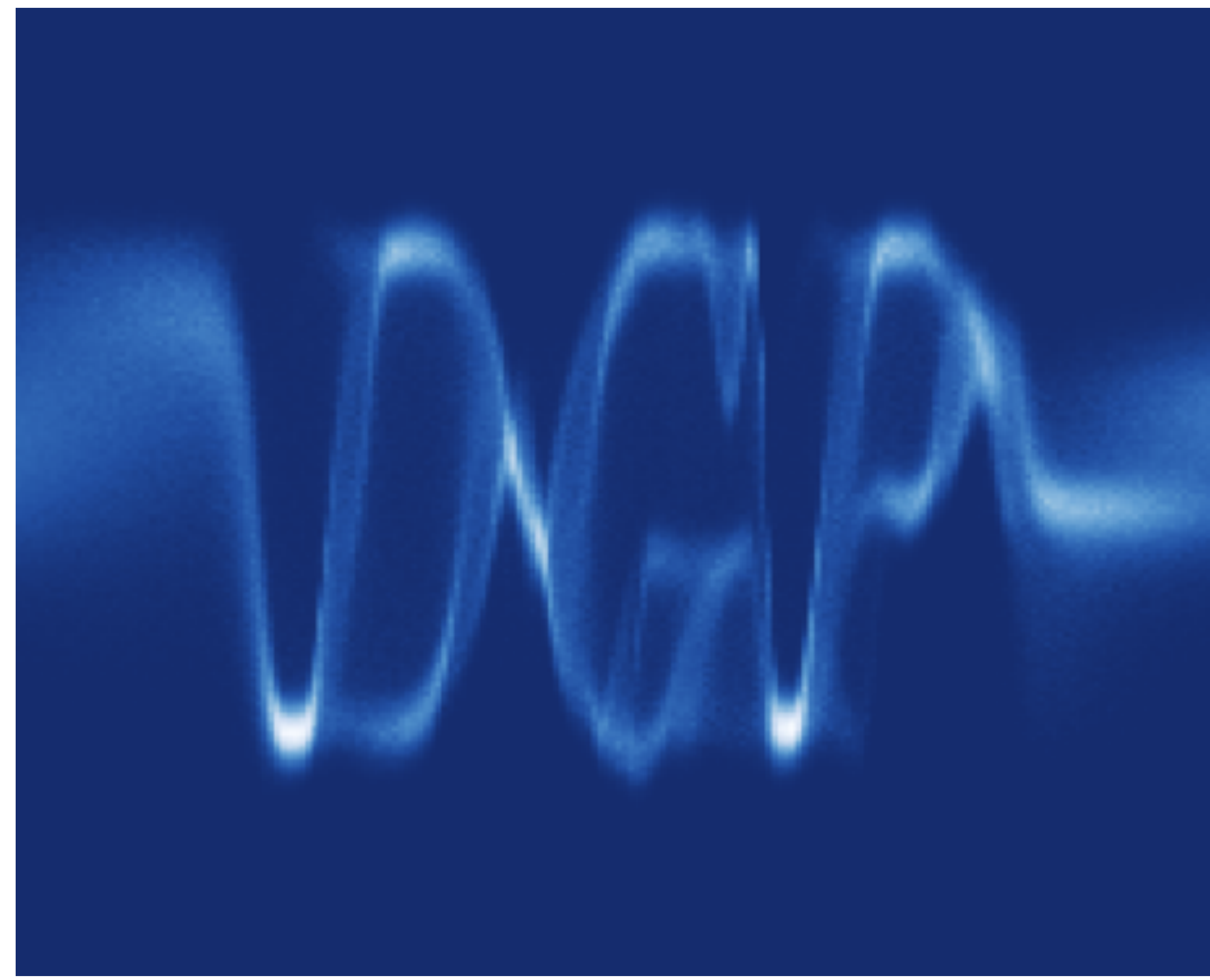
Contributions

- **New architecture** - latent variables by concatenation, not addition



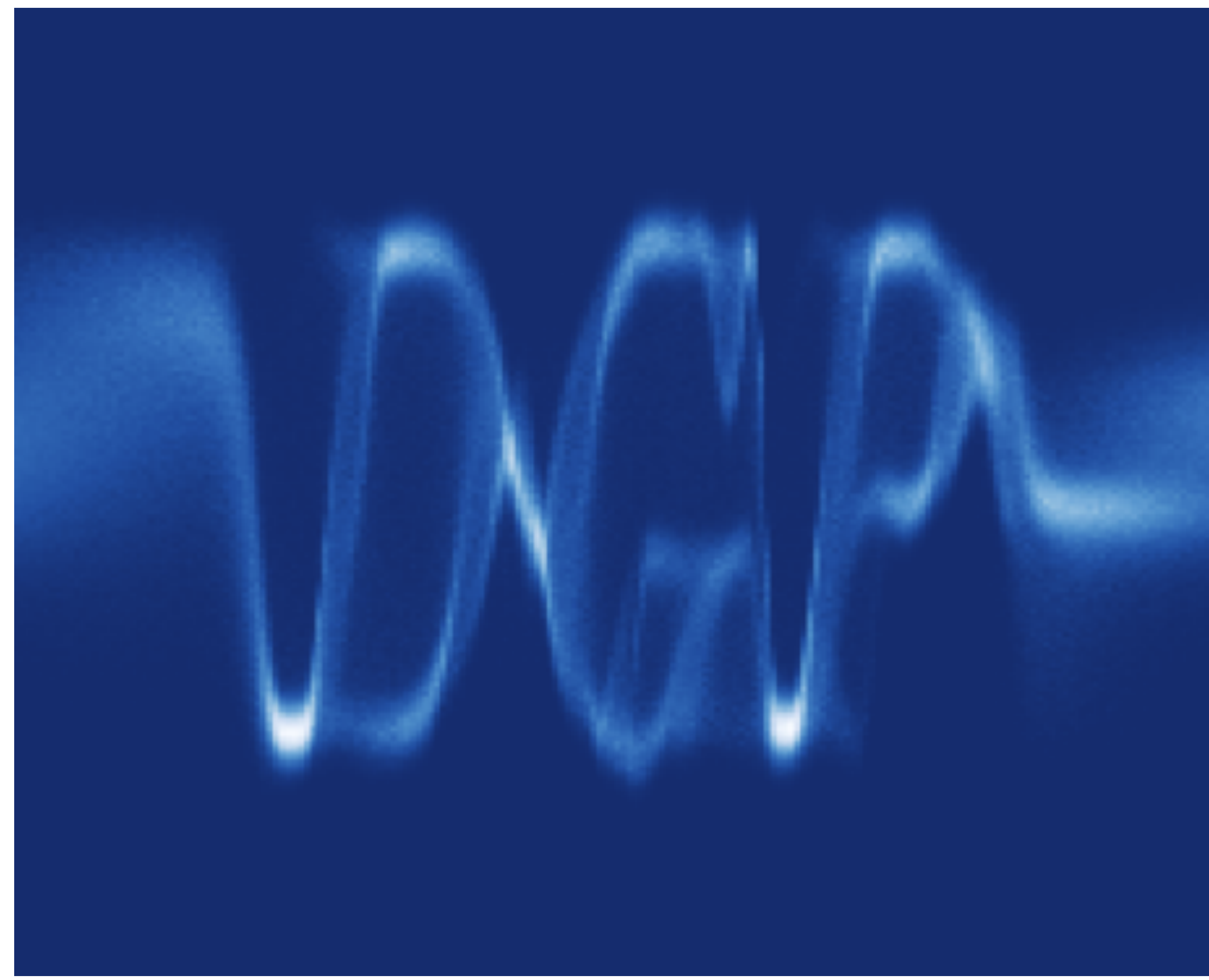
Contributions

- **New architecture** - latent variables by concatenation, not addition
- **Importance-weighted** variational inference, exploiting analytic results

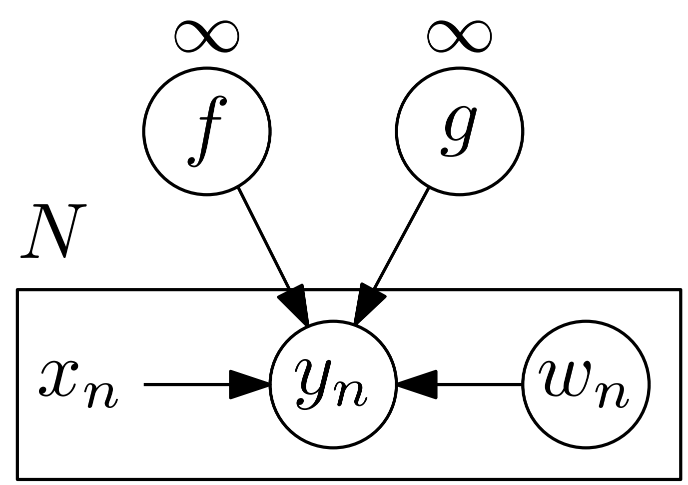


Contributions

- **New architecture** - latent variables by concatenation, not addition
- **Importance-weighted** variational inference, exploiting analytic results
- Provide an extensive empirical comparison with all **41 UCI regression datasets**



A few details



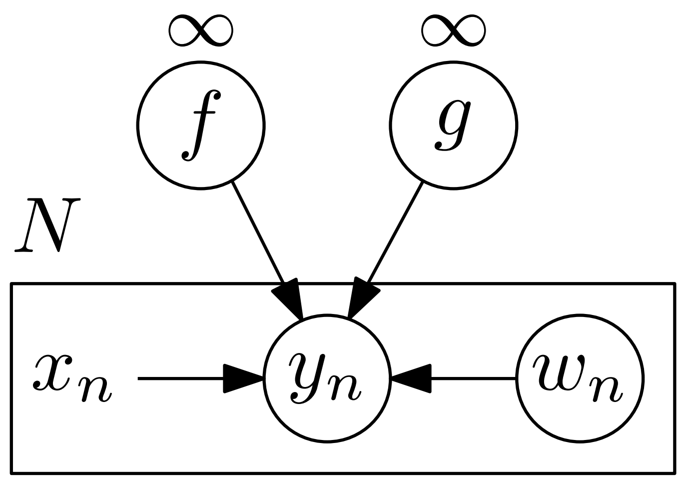
$$y_n = \mathcal{N}(f(g([x_n, w_n])), \sigma^2)$$

$$w_n \sim \mathcal{N}(0, 1)$$

$$f \sim \mathcal{GP}(\mu_1, k_1)$$

$$g \sim \mathcal{GP}(\mu_2, k_2)$$

A few details



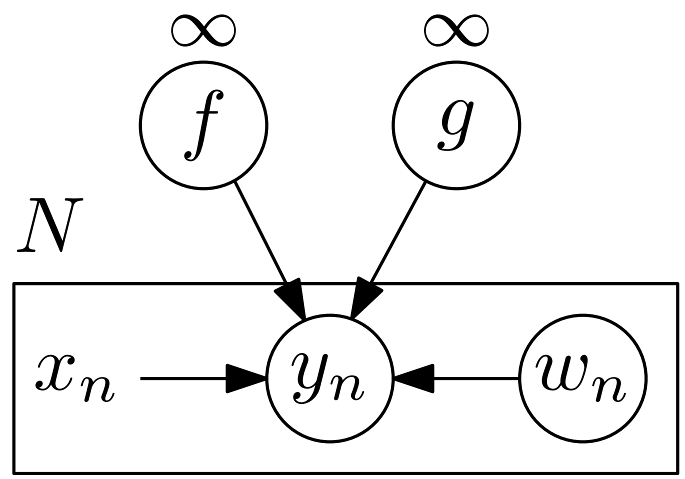
$$y_n = \mathcal{N}(f(g([x_n, w_n])), \sigma^2)$$

Importance weighting (Gaussian proposal) → $$w_n \sim \mathcal{N}(0, 1)$$

$$f \sim \mathcal{GP}(\mu_1, k_1)$$

$$g \sim \mathcal{GP}(\mu_2, k_2)$$

A few details



$$y_n = \mathcal{N}(f(g([x_n, w_n])), \sigma^2)$$

Importance weighting (Gaussian proposal) →

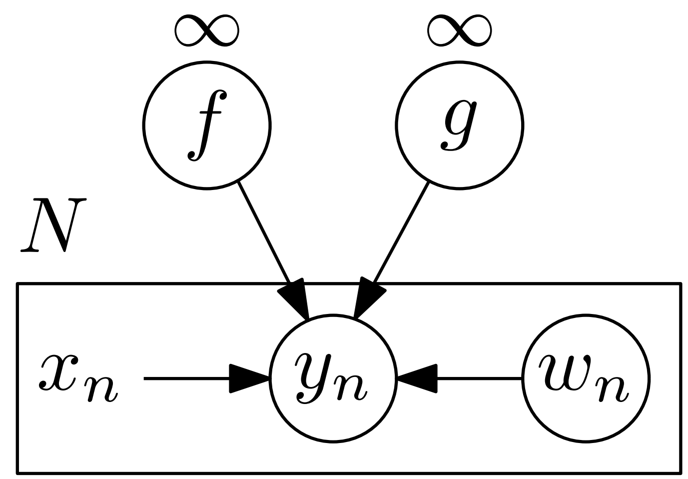
$$w_n \sim \mathcal{N}(0, 1)$$

$$f \sim \mathcal{GP}(\mu_1, k_1)$$

$$g \sim \mathcal{GP}(\mu_2, k_2)$$

← **Variational inference (sparse GP posterior)**

A few details



$$y_n = \mathcal{N}(f(g([x_n, w_n])), \sigma^2)$$

Importance weighting (Gaussian proposal) →

$$w_n \sim \mathcal{N}(0, 1)$$

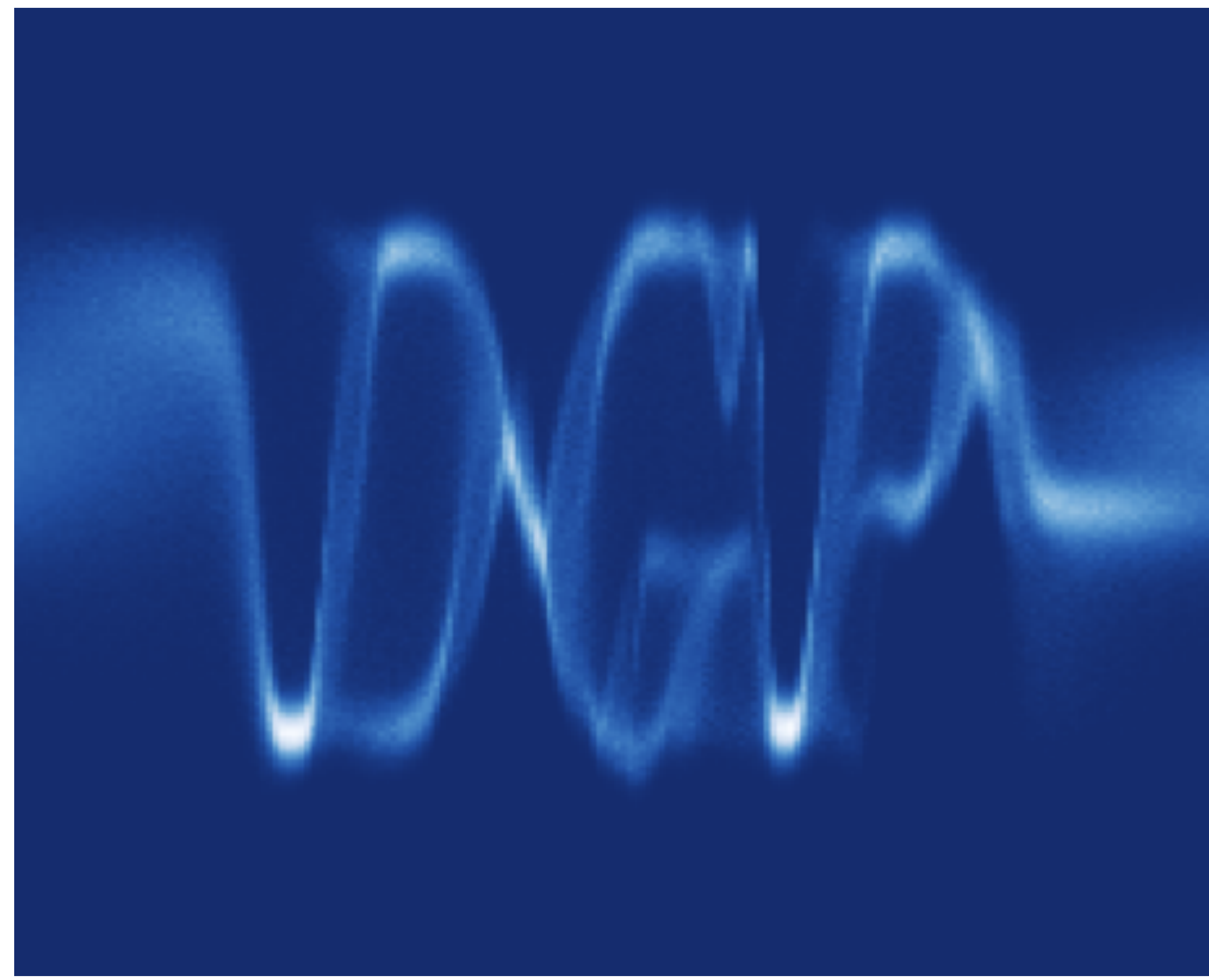
$$f \sim \mathcal{GP}(\mu_1, k_1)$$

$$g \sim \mathcal{GP}(\mu_2, k_2)$$

← **Variational inference (sparse GP posterior)**

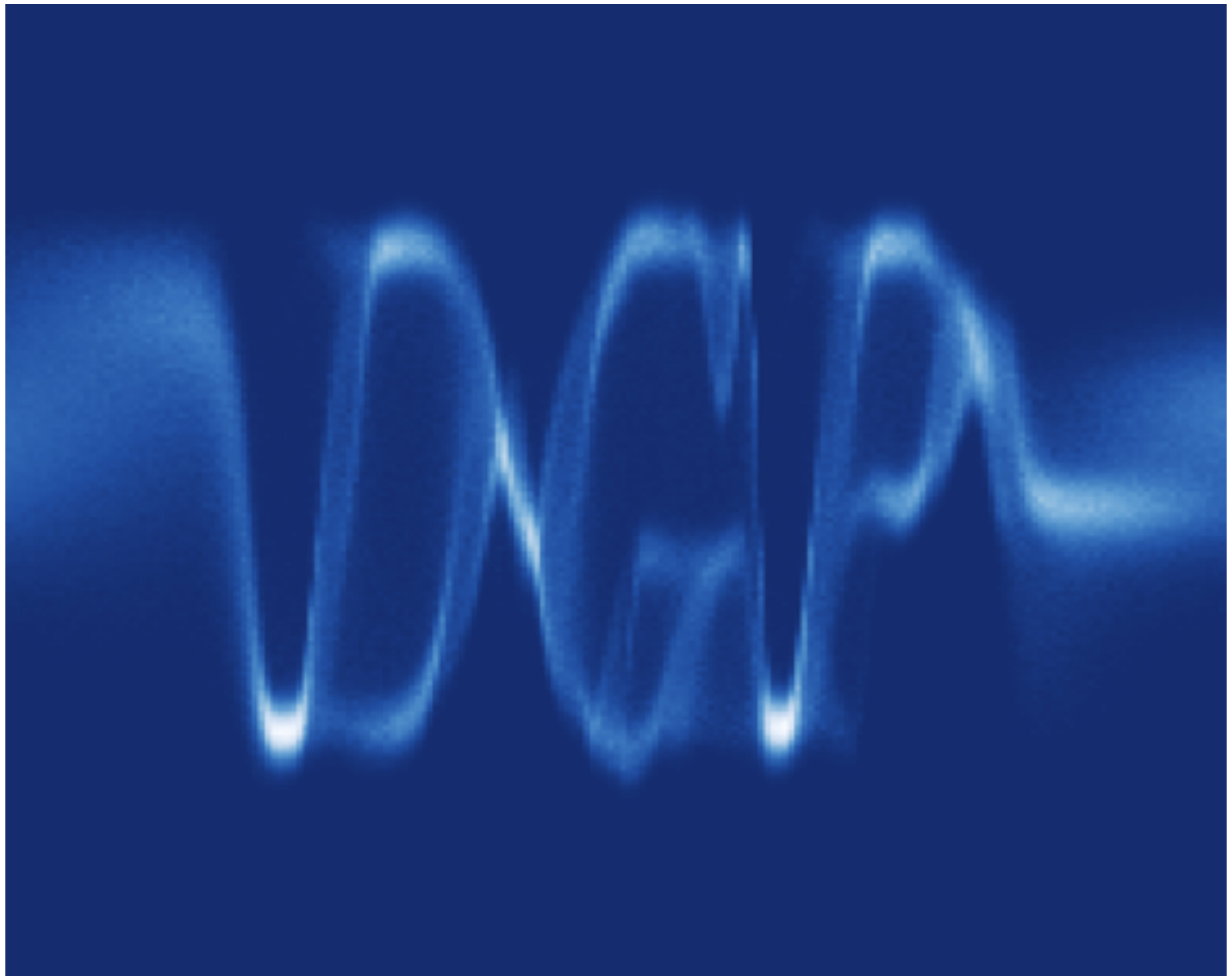
Our approach exploits analytic results, leading to a tighter bound

Results



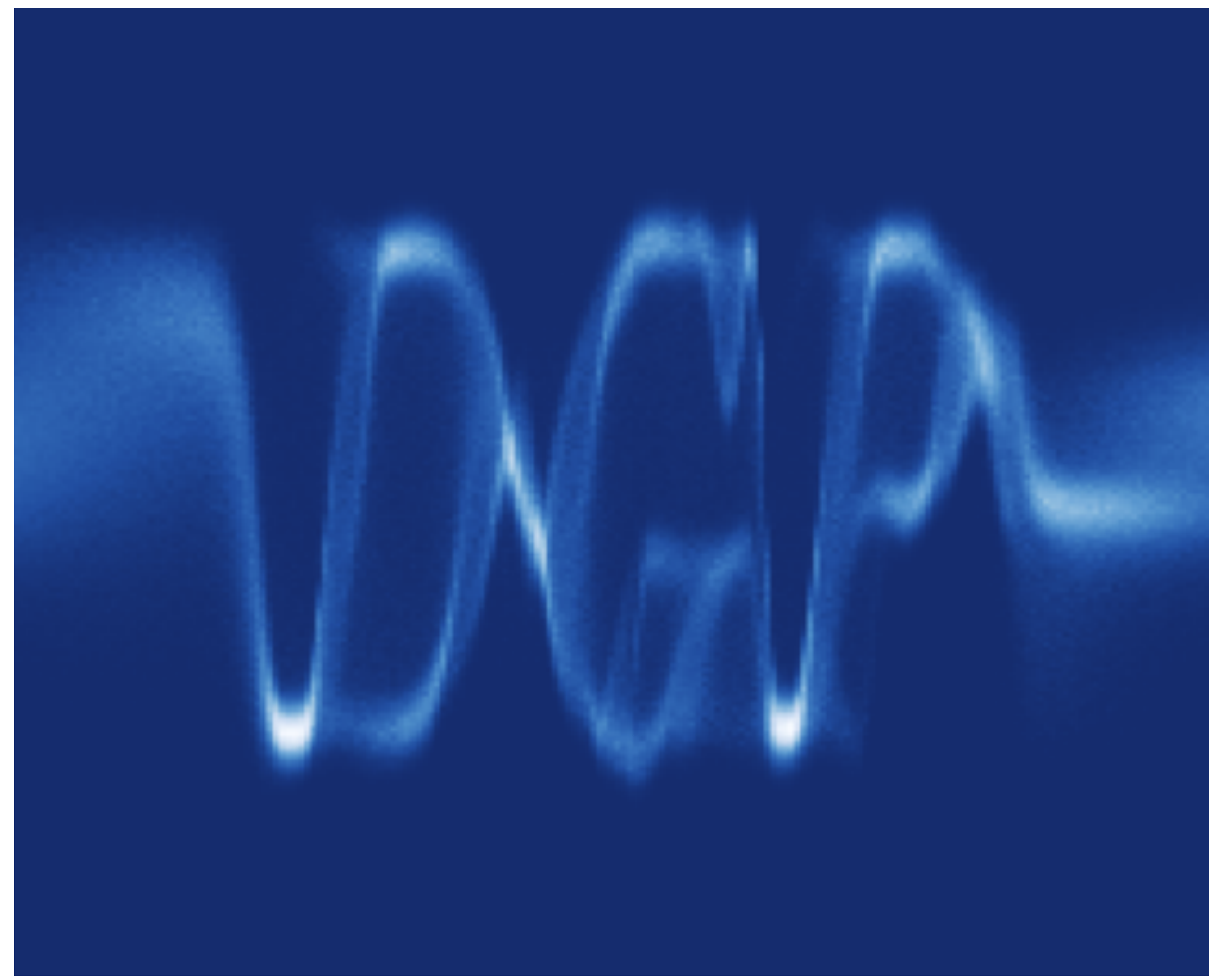
Results

- **Latent variables** in the DGP are highly beneficial



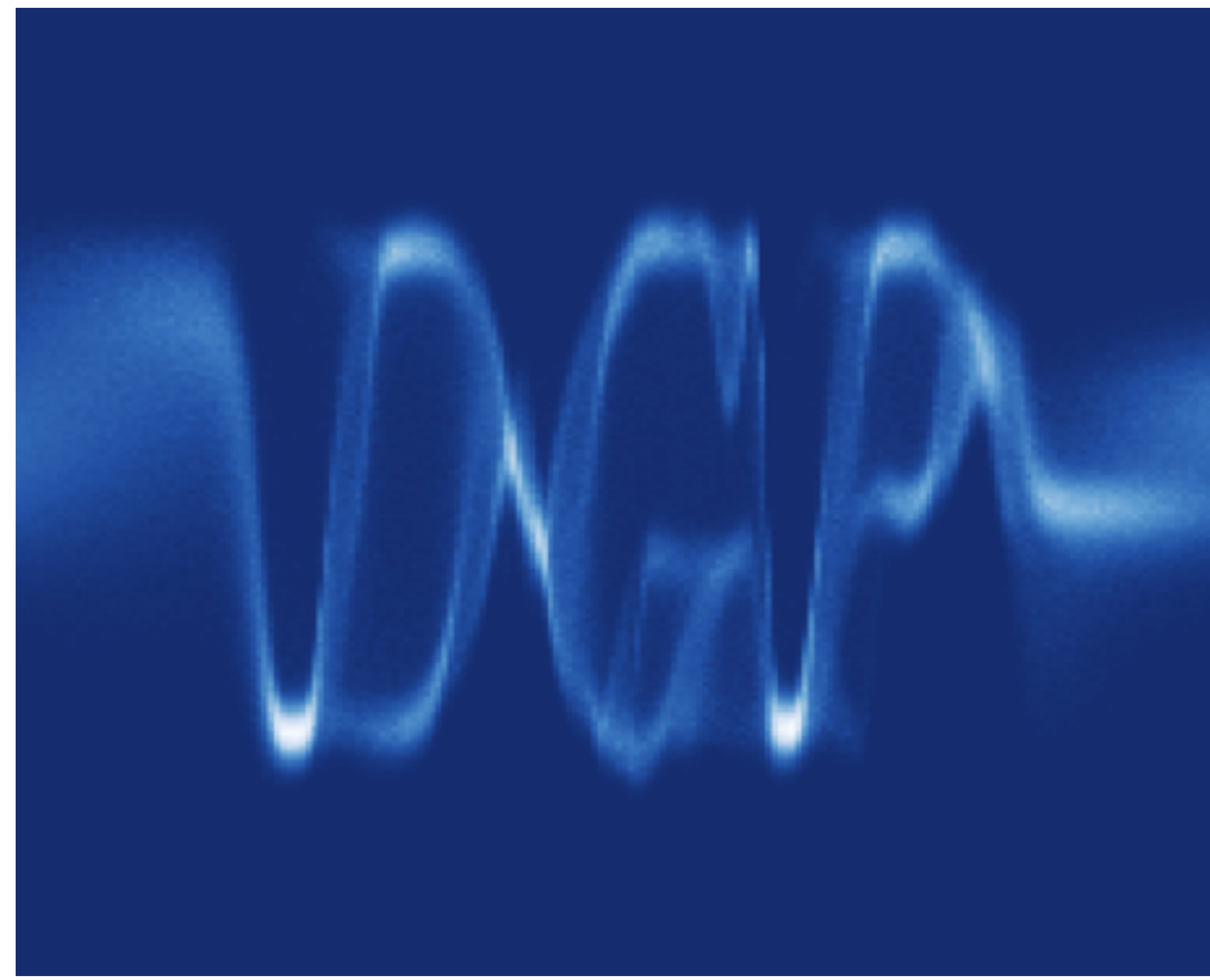
Results

- **Latent variables** in the DGP are highly beneficial
- Sometimes **depth** is enough. Sometimes **latent variables** are enough. Some datasets need **both**.



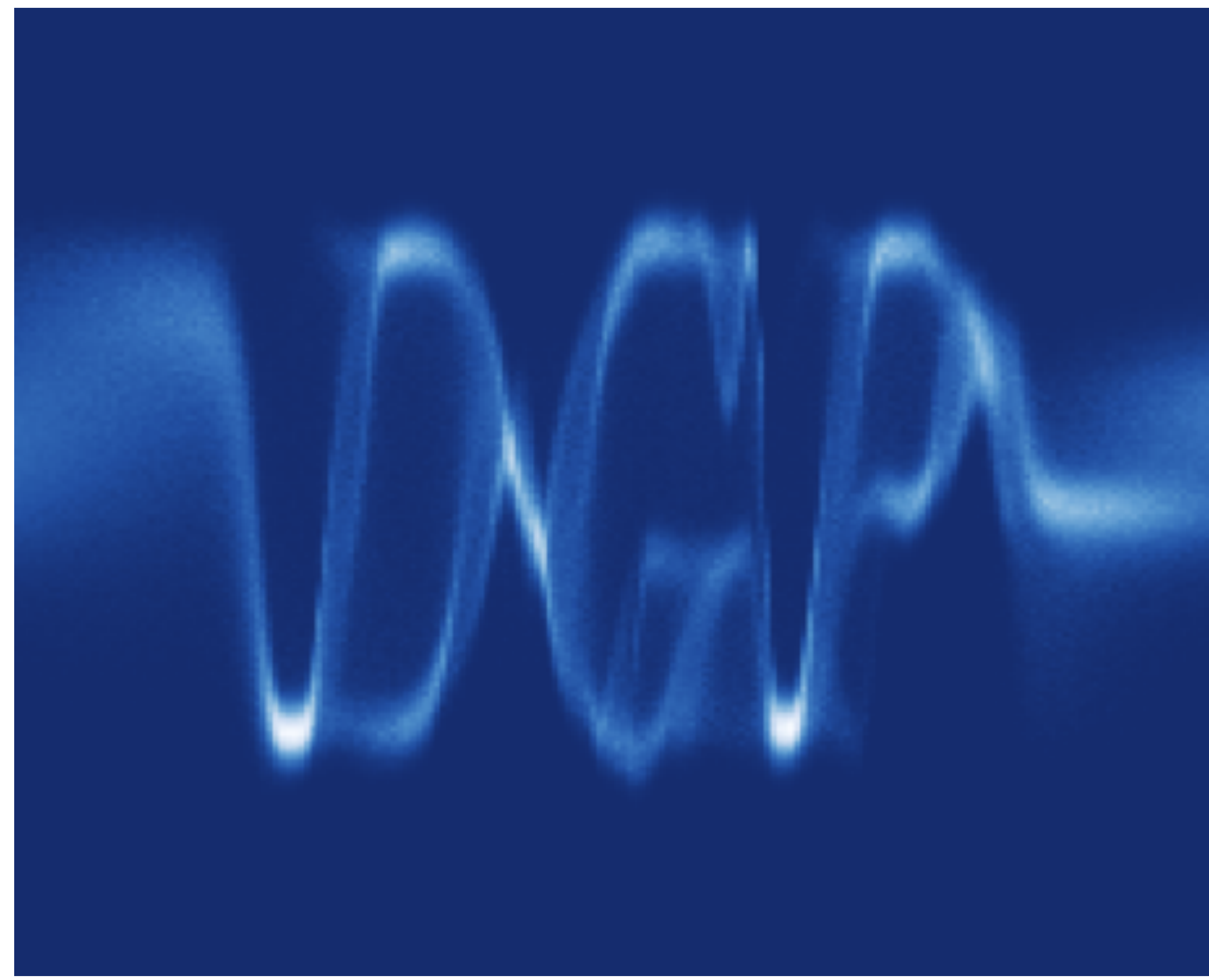
Results

- **Latent variables** in the DGP are highly beneficial
- Sometimes **depth** is enough. Sometimes **latent variables** are enough. Some datasets need **both**.
- **Importance-weighted** VI outperforms VI



Results

- **Latent variables** in the DGP are highly beneficial
- Sometimes **depth** is enough. Sometimes **latent variables** are enough. Some datasets need **both**.
- **Importance-weighted** VI outperforms VI



hughsalimbeni / **DGPs_with_IWVI**

hughsalimbeni / **bayesian_benchmarks**

Thanks for listening

Poster #218

- **New architecture**
- **Importance-weighted**
- **41 datasets**