Variational Auto-Regressive Gaussian Processes for Continual Learning

Sanyam Kapoor, Theofanis Karaletsos, Thang D. Bui

NEW YORK UNIVERSITY

FACEBOOK AI

THE UNIVERSITY OF SYDNEY

Continual Learning

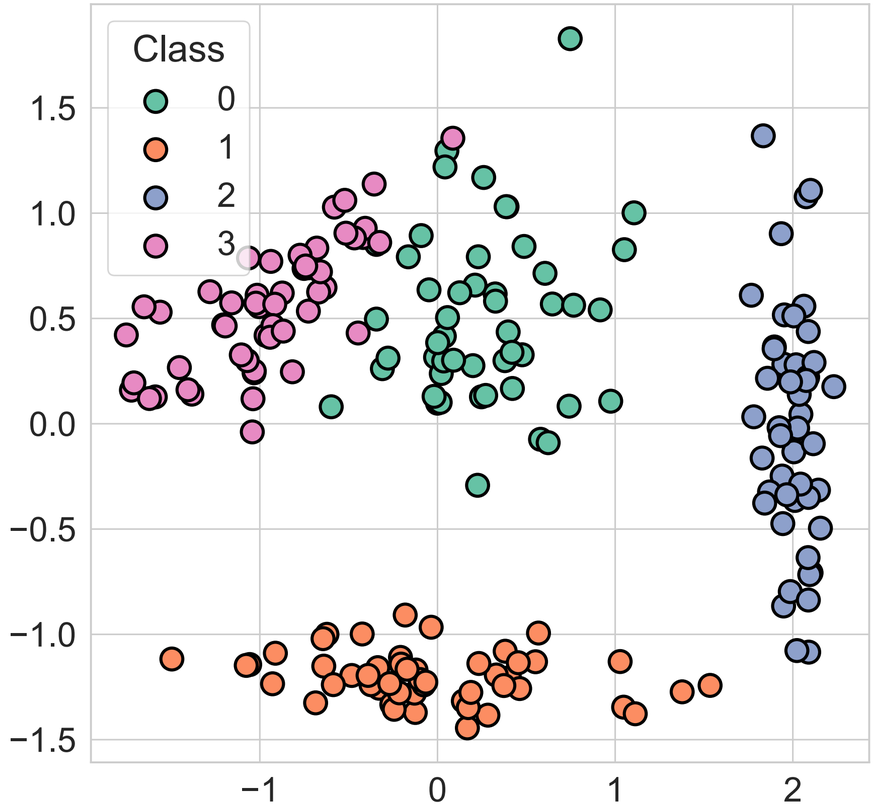
- The holy-grail of *intelligent* systems - adapt under new experiences, without compromising old learnings.
- Most machine learning methods, however, have been successful under the strong assumption of *i.i.d*. data.
- Assumption violated for many important applications like on-device learning on smartphones, on-premise patient diagnostics in hospitals, etc.
- Failures manifest as *catastrophic forgetting*.

Tackling Catastrophic Forgetting

- Bayes' theorem already provides a sound computational framework for continual learning.
- We propose **VAR-GPs** that:
  - model continual learning via sparse Gaussian processes,
  - use a novel auto-regressive parametrization to preserve old information.

A Toy Continual Learning Problem

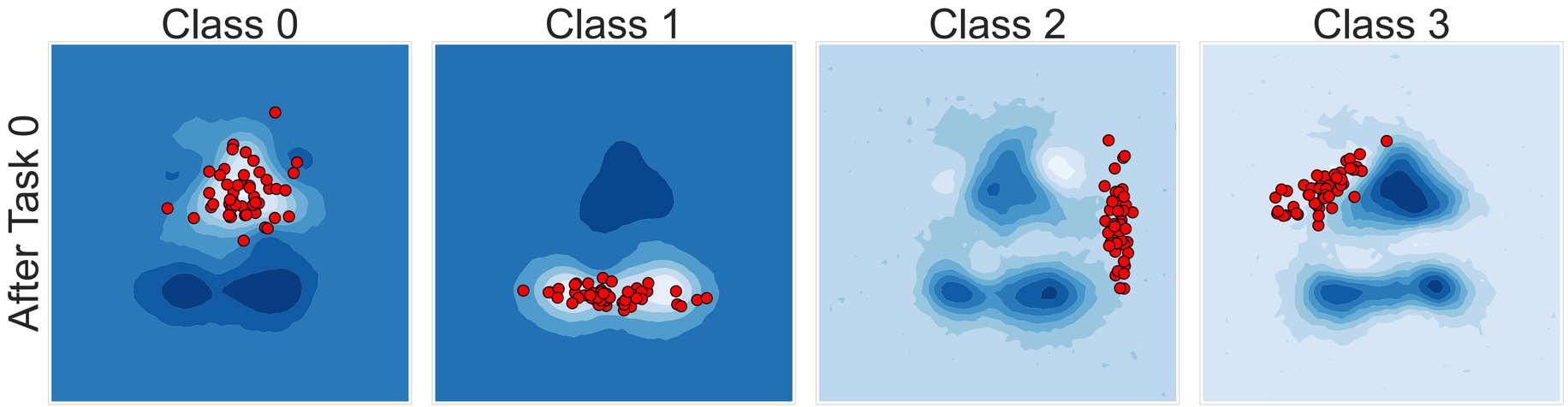
- Consider a four-way classification problem.
- To mimic a continual learning setting, we split the learning into two tasks:
  - Task 0 - Observing *only* classes 0 and 1,
  - Task 1 - Observing *only* classes 2 and 3.



Visualizing Learning the First Task

- We visualize how a typical learning algorithm would behave. Brighter regions represent a higher *posterior predictive probability*.

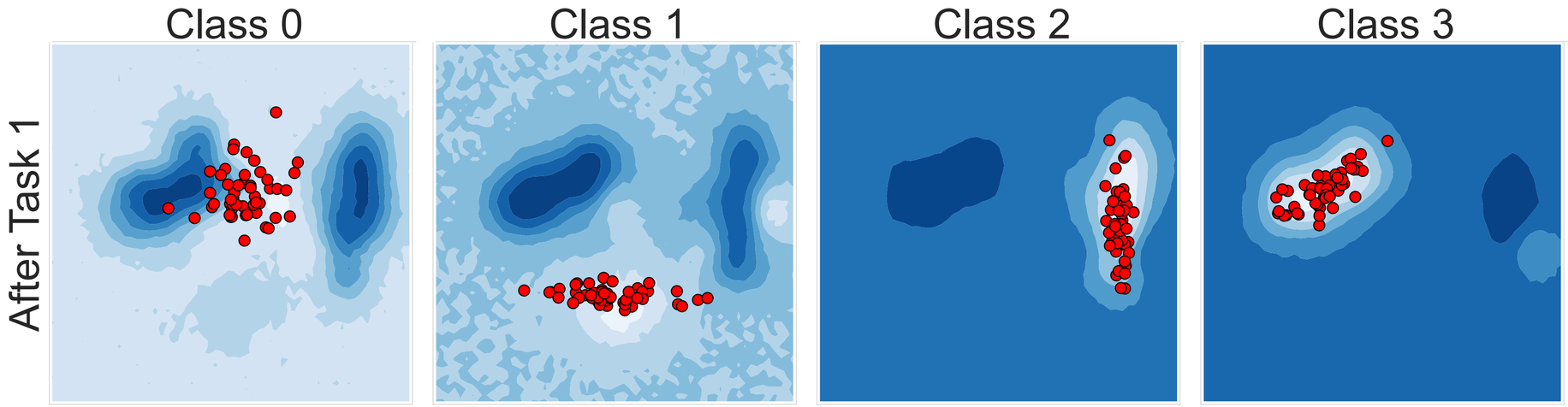
- Task 0 is learned well; Task 1 shows high uncertainty, as expected.



Visualizing Catastrophic Forgetting

- Continuing the learning, Task 1 is a success, i.e high probability regions *only* near the observed data.

- Classes 0 and 1, however, show high predictive probability, i.e. *almost* all information from Task 0 is lost.



Sparse Variational GPs

- We imagine a set of learnable *inducing points* $\{\mathbf{u} = f(\mathbf{Z}), \mathbf{Z}\}$, that partially explain the complete function, i.e. $f = \{f_{\neq \mathbf{u}}, \mathbf{u}\}$.

- The full prior model for observations $\{\mathbf{X}, \mathbf{y}\}$ is then,

$$p(\mathbf{y}, f, \theta \mid \mathbf{X}) = p(\mathbf{y} \mid f(\mathbf{X}))p(f_{\neq \mathbf{u}} \mid \mathbf{u}, \theta)p(\mathbf{u} \mid \mathbf{Z}, \theta)p(\theta).$$

- The objective of inference is to find an *approximate* posterior $q(f, \theta)$ via maximizing the evidence lower bound (ELBO).

Hensman, Matthews, and Ghahramani. Scalable Variational Gaussian Process Classification. In AISTATS 2015.

Inference in SVGPs

- We posit the variational posterior as,

$$q(f, \theta) = p(f_{\neq \mathbf{u}} \mid \mathbf{u}, \theta) q(\mathbf{u}) q(\theta).$$

- The evidence lower bound (ELBO) is then given by,

$$F(q, \theta) = \sum_{i=1}^{N} \mathbb{E}_{q(f,\theta)} \left[\log p \left(y_i \mid f(\mathbf{X}_i) \right) \right] - \mathrm{KL} \left[q(\theta) \mid\mid p(\theta) \right] - \mathbb{E}_{q(\theta)} \left[\mathrm{KL} \left[q(\mathbf{u}) \mid\mid p(\mathbf{u} \mid \theta) \right] \right]$$

- Notably, for inference, we:
  - can now use stochastic maximization of ELBO via mini-batching.
  - regularize the prior over kernel hyperparameters to stabilize learning.

Learning the First Task

- For a set of T tasks, we augment the notation with task numbers $\{1,2,\ldots,T\}$.
- Learning the first task is *exactly* equivalent to the usual ELBO.
- For a dataset $\{\mathbf{X}^{(1)}, \mathbf{y}^{(1)}\}$ of size N_1, we reproduce the ELBO using inducing points $\{\mathbf{u}_1, \mathbf{Z}_1\}$ as,

$$F(q_1, \theta) = \sum_{i=1}^{N_1} \mathbb{E}_{q_1(f,\theta)} \left[\log p \left(y_i^{(1)} \mid f\left(\mathbf{X}_i^{(1)} \right) \right) \right] - \mathrm{KL}\left[q_1(\theta) \mid\mid p(\theta) \right] - \mathbb{E}_{q_1(\theta)} \left[\mathrm{KL}\left[q(\mathbf{u}_1) \mid\mid p(\mathbf{u}_1 \mid \theta) \right] \right]$$

Approximate Running Joint Model

- For continual learning, we build upon the SVGP model, treating the *approximate posterior* so far as the new prior.
- Introducing new inducing points $\{\mathbf{u}_t, \mathbf{Z}_t\}$, such that $f = \{f_{\neq \mathbf{u}_{\leq t}}, \mathbf{u}_t, \mathbf{u}_{<t}\}$,

$$p(\mathbf{y}^{(t)}, f, \theta \mid \mathbf{X}^{(t)}, \mathrm{D}^{(<t)}) \approx \prod_{i=1}^{N_t} p(y_i^{(t)} \mid f, \mathbf{x}_i^{(t)})$$
$$p(f_{\neq \mathbf{u}_{\leq t}} \mid \mathbf{u}_{\leq t}, \theta)$$
$$p(\mathbf{u}_t \mid \mathbf{Z}_t, \mathbf{u}_{<t}, \theta)$$
$$q(\mathbf{u}_{<t} \mid \mathbf{Z}_{<t}, \theta)$$
$$q_{t-1}(\theta) \ .$$

Learning Subsequent Tasks

- We propose an *auto-regressive variational posterior*,

$$q_t(f, \theta) = p(f_{\neq \mathbf{u}_{\leq t}} \mid \mathbf{u}_{\leq t}, \theta) q(\mathbf{u}_t \mid \mathbf{Z}_t, \mathbf{u}_{<t}, \mathbf{Z}_{<t}, \theta) q(\mathbf{u}_{<t} \mid \mathbf{Z}_{<t}, \theta) q_t(\theta)$$

- and arrive at a Generalized ELBO for Continual Learning,

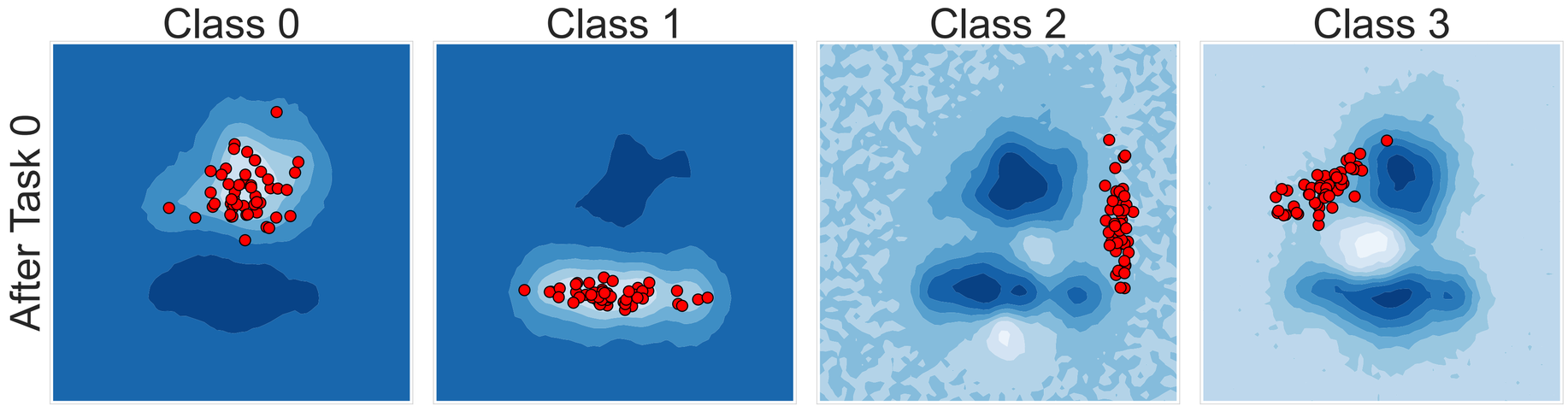
$$F(q_t) = \sum_{i=1}^{N_t} \mathbb{E}_{q_t(f,\theta)} \left[\log p(y_i^{(t)} \mid f, \mathbf{x}_i^{(t)}) \right] - \mathrm{KL} \left[q_t(\theta) || q_{t-1}(\theta) \right] - \mathbb{E}_{q_t(\theta) q(\mathbf{u}_{<t} | \mathbf{Z}_{<t}, \theta)} \left[\mathfrak{D}_t \right]$$

- This leads to a natural interpretation — we maximize the likelihood of data on the current task, and balance learning against past tasks.

Visualizing Learning with VAR-GPs

- This time, we visualize how VAR-GPs behave on the toy problem. Again, brighter regions represent a higher *posterior predictive probability*.

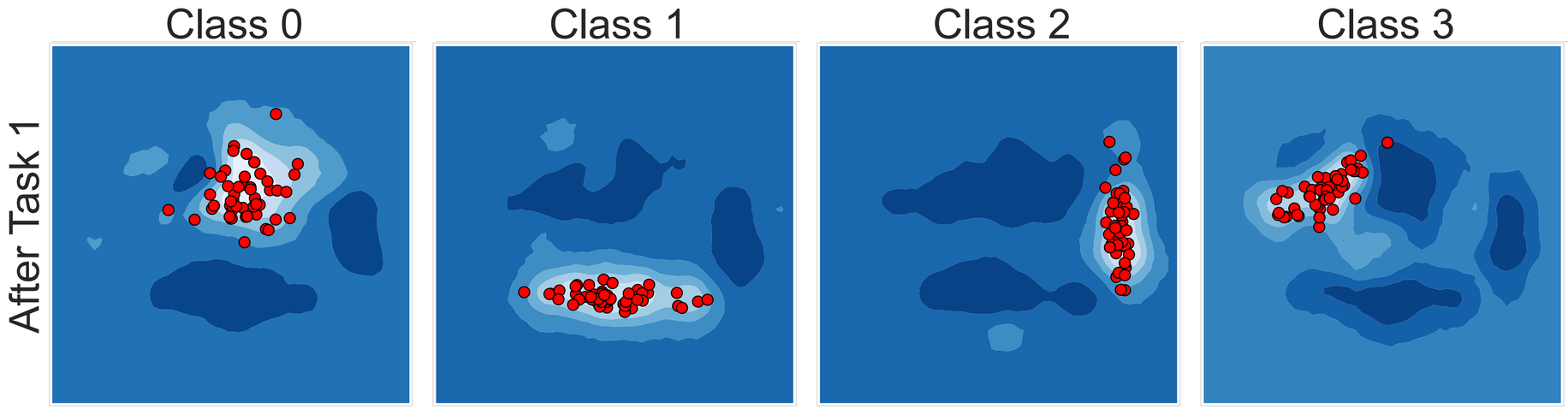
- Task 0 is learned well; Task 1 shows high uncertainty, as expected.



Catastrophic Forgetting Alleviated

- Continuing the learning, Task 1 (observing only Classes 2 and 3) is a success.

- Moreover, almost all high probability regions from Task 0 are preserved, i.e. we alleviate catastrophic forgetting.



Summary and Outlook

- Our *Generalized Variational Lower Bound* provides a principled objective for continual learning.
- VAR-GPs have a fruitful connection to *Expectation Propagation* (EP) and *Orthogonal Inducing Points* in Gaussian Processes.
- VAR-GPs scale cubically w.r.t number of tasks, which is unfavorable for long runs of continual learning.
- We hope VAR-GPs stimulate further research in bringing favorable properties of GPs to continual learning.

Resources

perhapsbay.es/vargp

perhapsbay.es/vargp-code