CONDITIONING VARIATIONAL GAUSSIAN PROCESSES FOR ONLINE DECISION-MAKING

WESLEY MADDOX, SAM STANTON, ANDREW GORDON WILSON

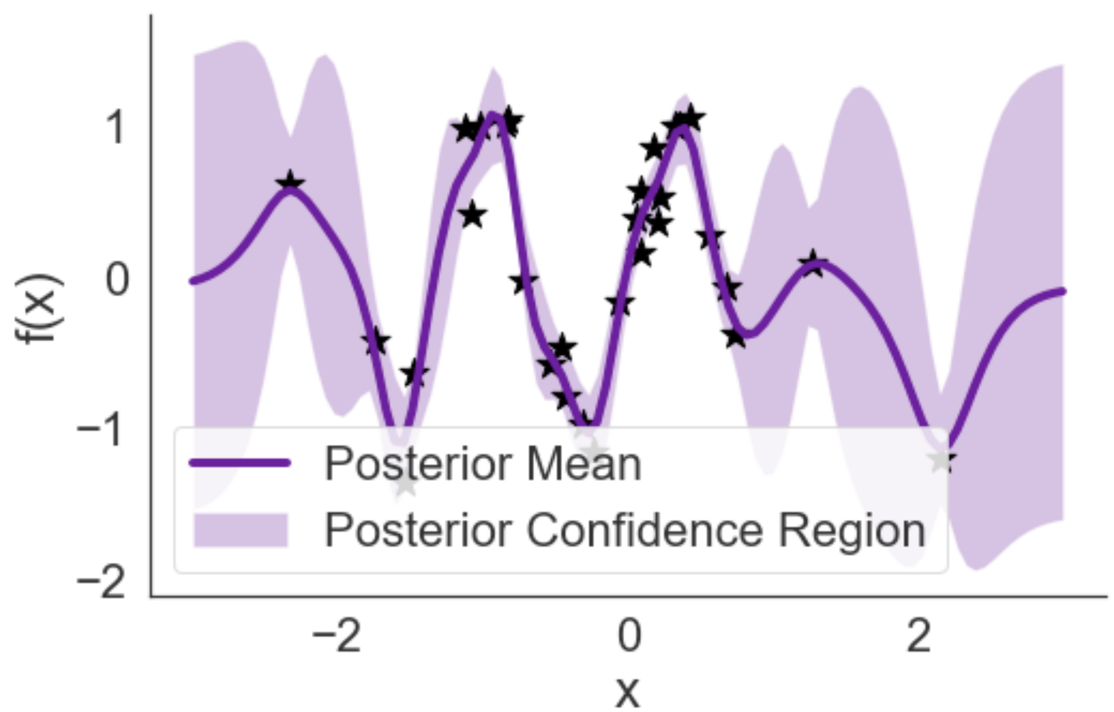
NYU

TAKEAWAYS

- We propose online variational conditioning (OVC) which enables SVGPs to condition on new data like exact GPs do.
- OVC enables online learning for SVGPs as well as advanced acquisitions in Bayes opt (e.g. lookahead acquisitions and knowledge gradient).

GAUSSIAN PROCESSES

- Nonparametric models over functions
  - Extend multivariate gaussians to function spaces
    - Latent function $f \sim \mathcal{GP}(\mu_\theta(x), k_\theta(x, x'))$ $y \sim \mathcal{N}(f, \sigma^2 I)$
- Predictive distribution is closed form (for regression)

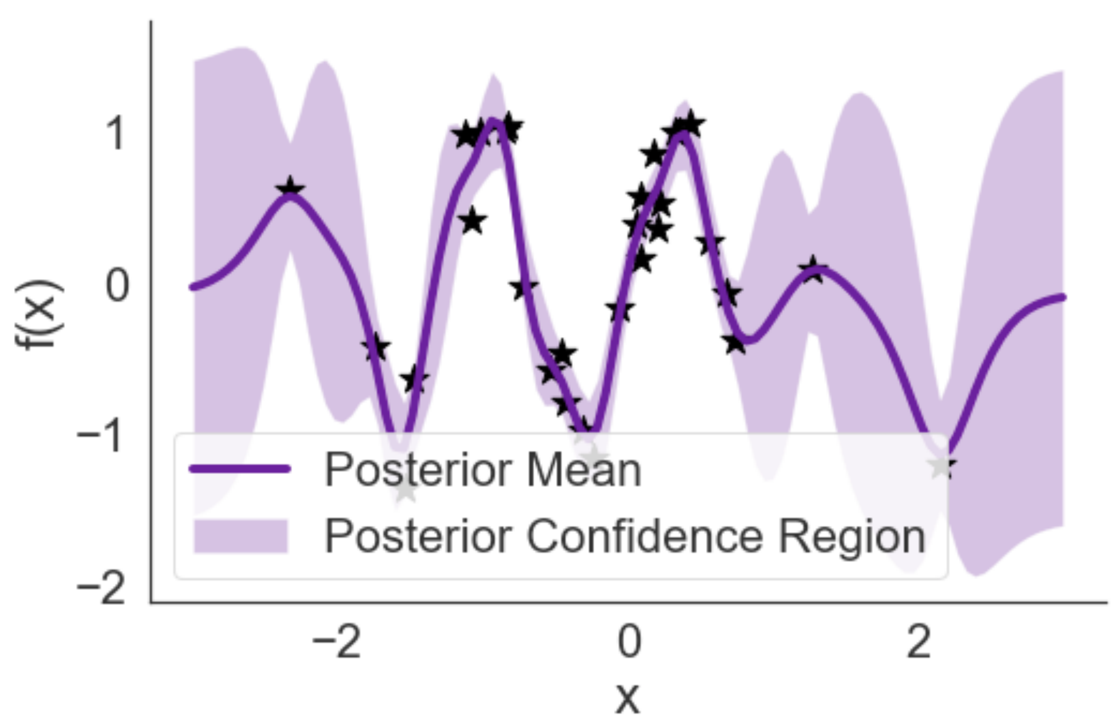


GAUSSIAN PROCESSES: PREDICTION

- The predictive distribution is given by:

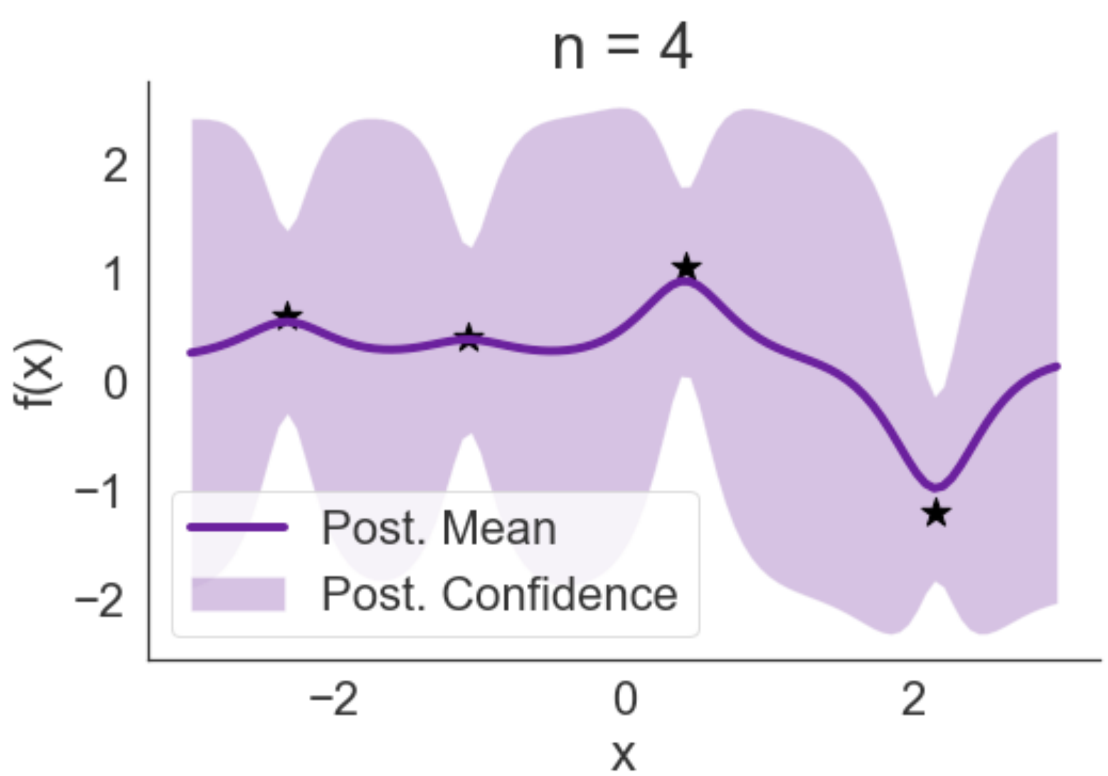
$$p(f^*|X^*, X, y) = \mathcal{N}(\mu_{f|\mathcal{D}}, \Sigma_{f|\mathcal{D}})$$

$$\mu_{f|\mathcal{D}} = K_{\mathbf{x}^* X}(K_{XX} + \sigma^2 I)^{-1}\mathbf{y},$$
$$\Sigma_{f|\mathcal{D}} = K_{\mathbf{x}^*\mathbf{x}^*} - K_{\mathbf{x}^* X}(K_{XX} + \sigma^2 I)^{-1} K_{X\mathbf{x}^*}.$$



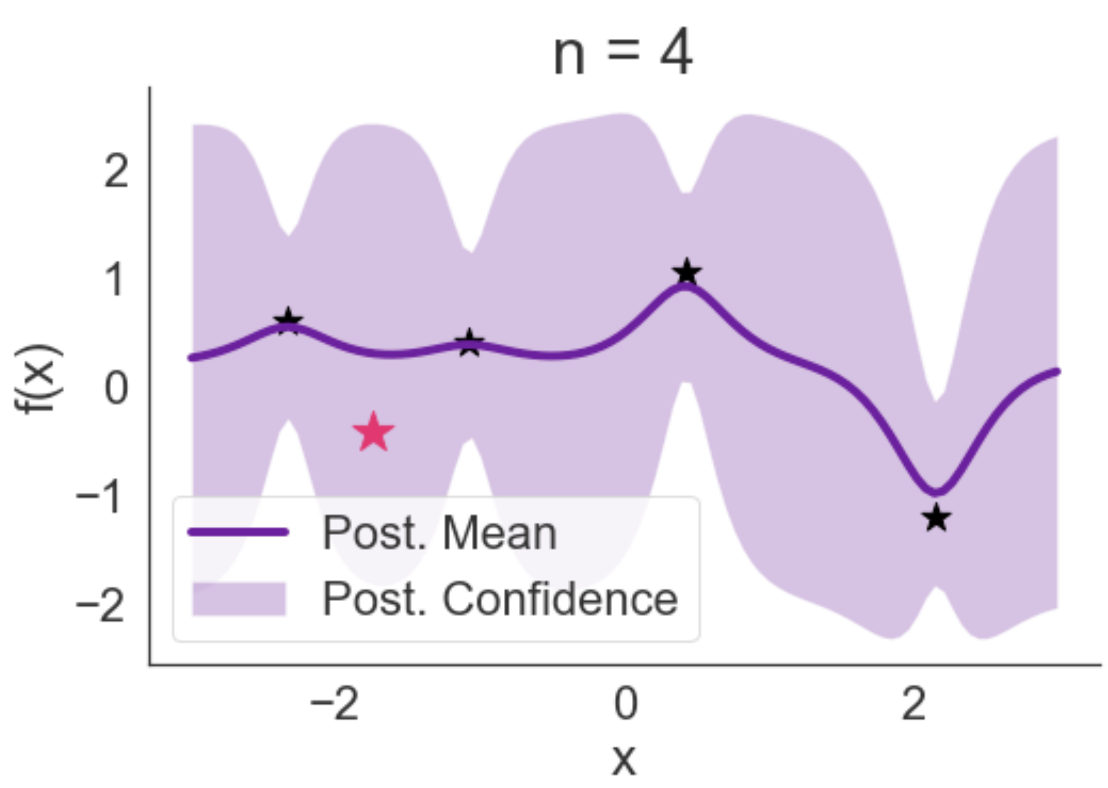
MAKING GPS STREAMING

- Add new data points to a GP model



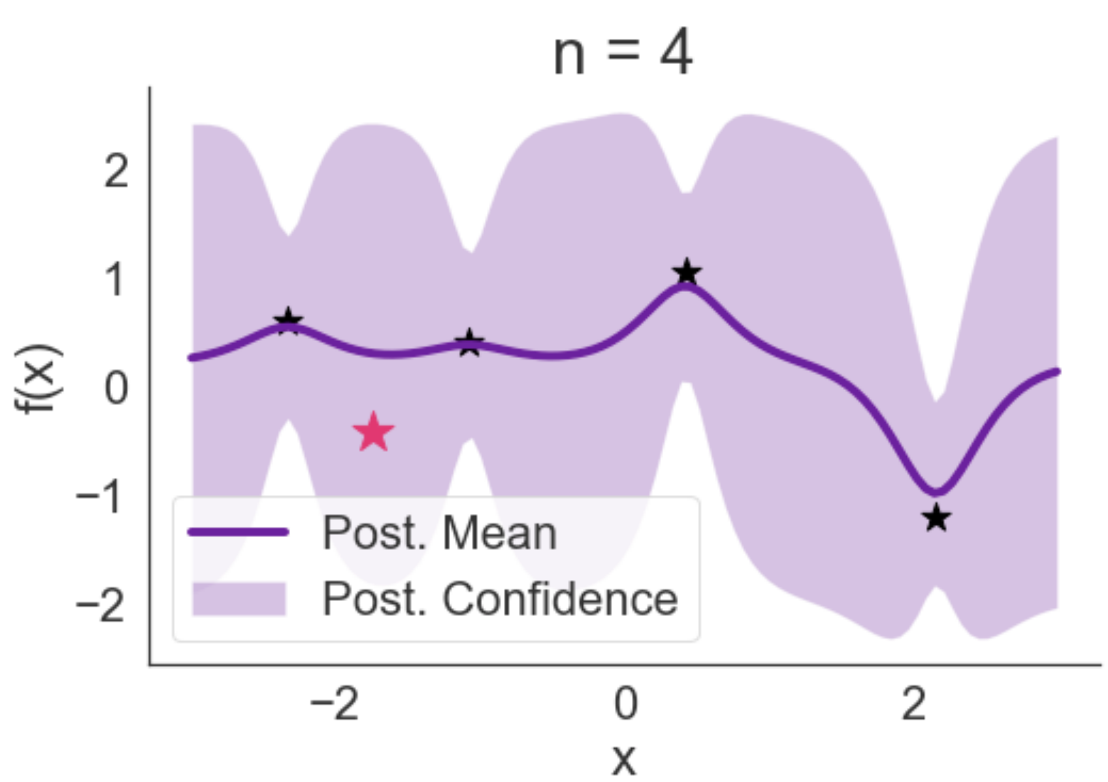
MAKING GPS STREAMING

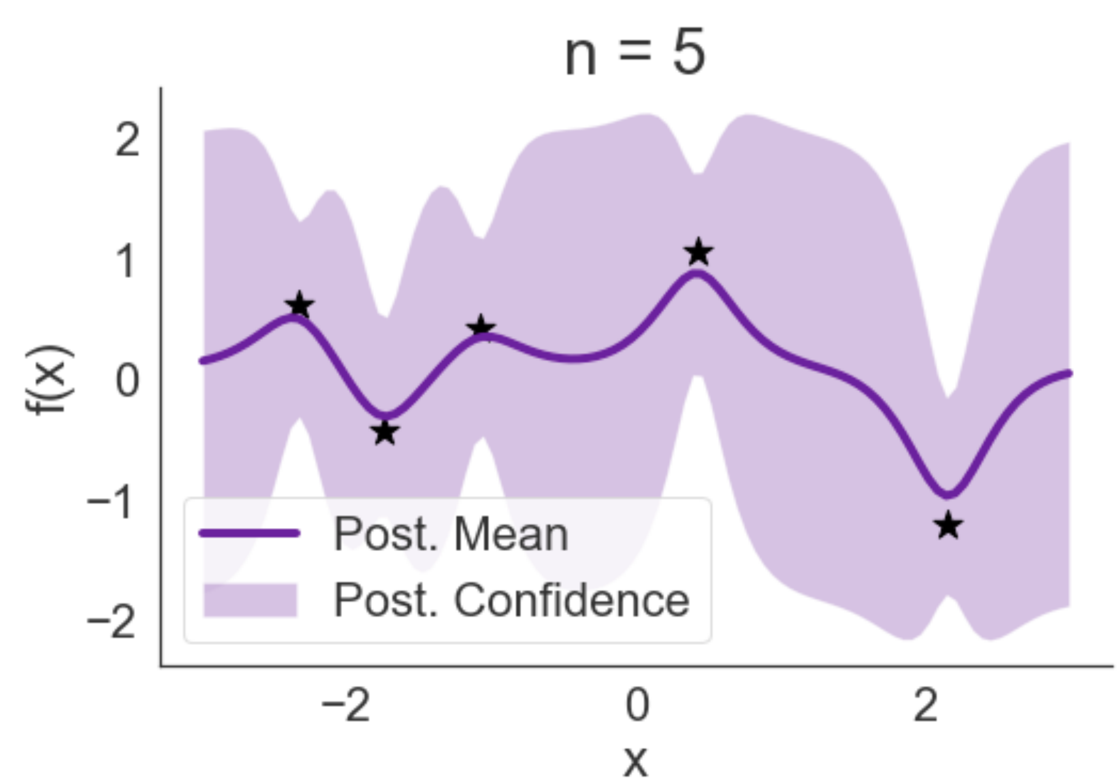
- Add new data points to a GP model



MAKING GPS STREAMING

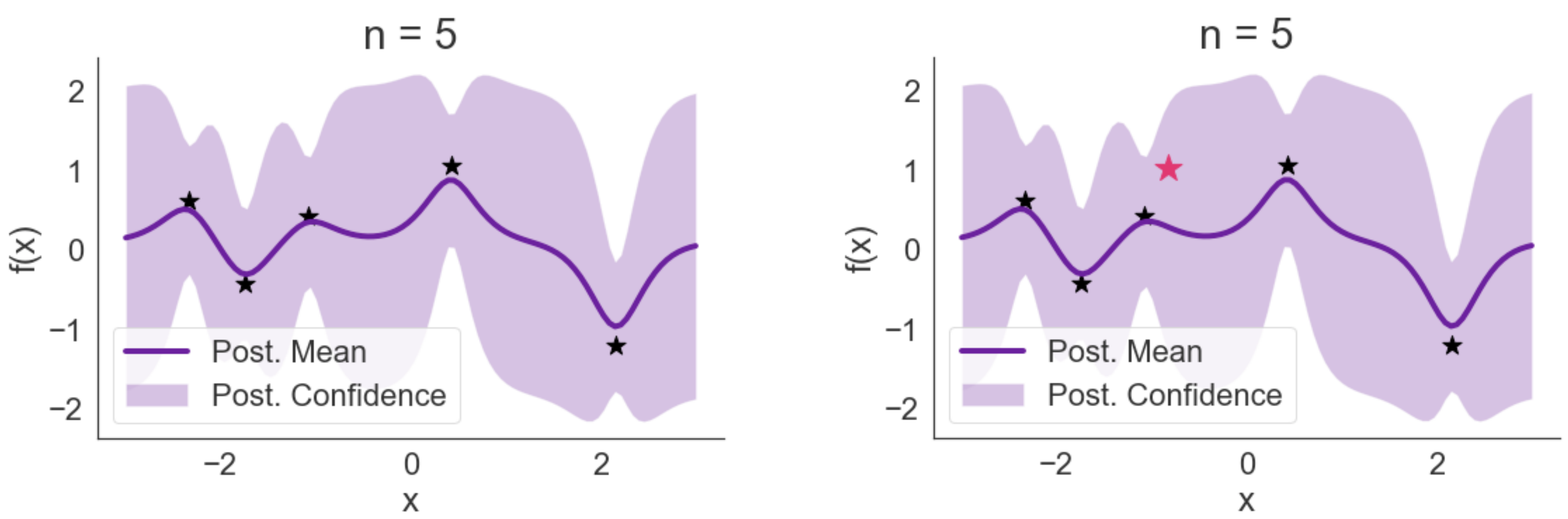
- Add new data points to a GP model





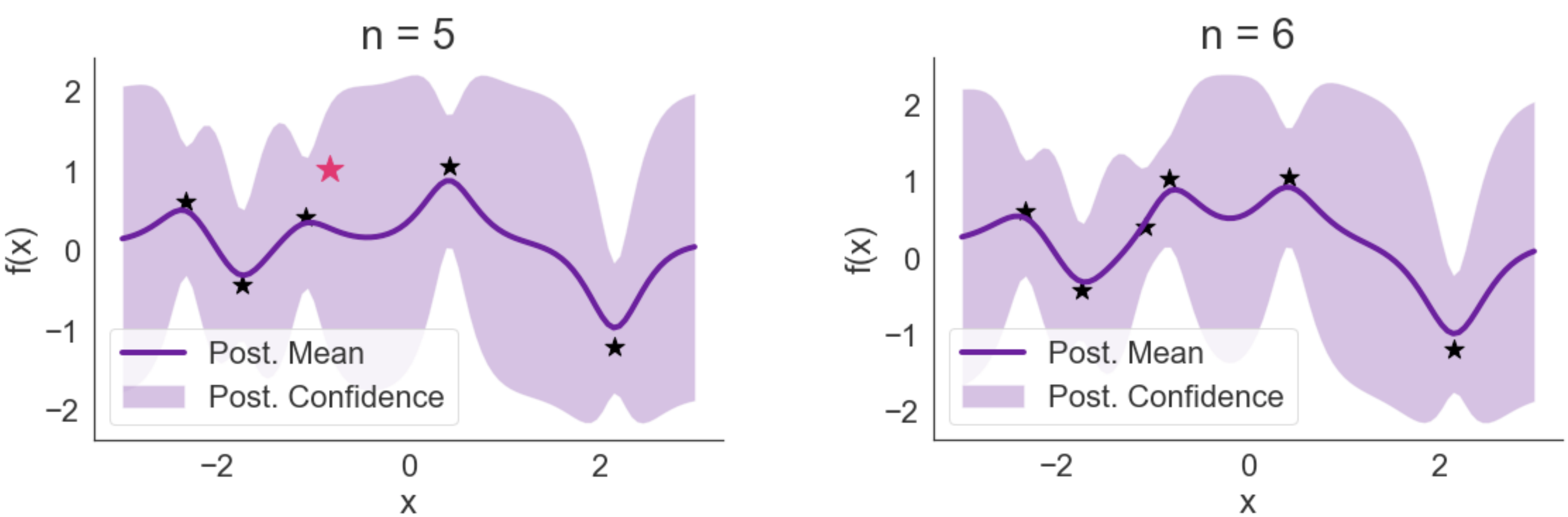
MAKING GPS STREAMING

- Add new data points to a GP model



MAKING GPS STREAMING

- Add new data points to a GP model



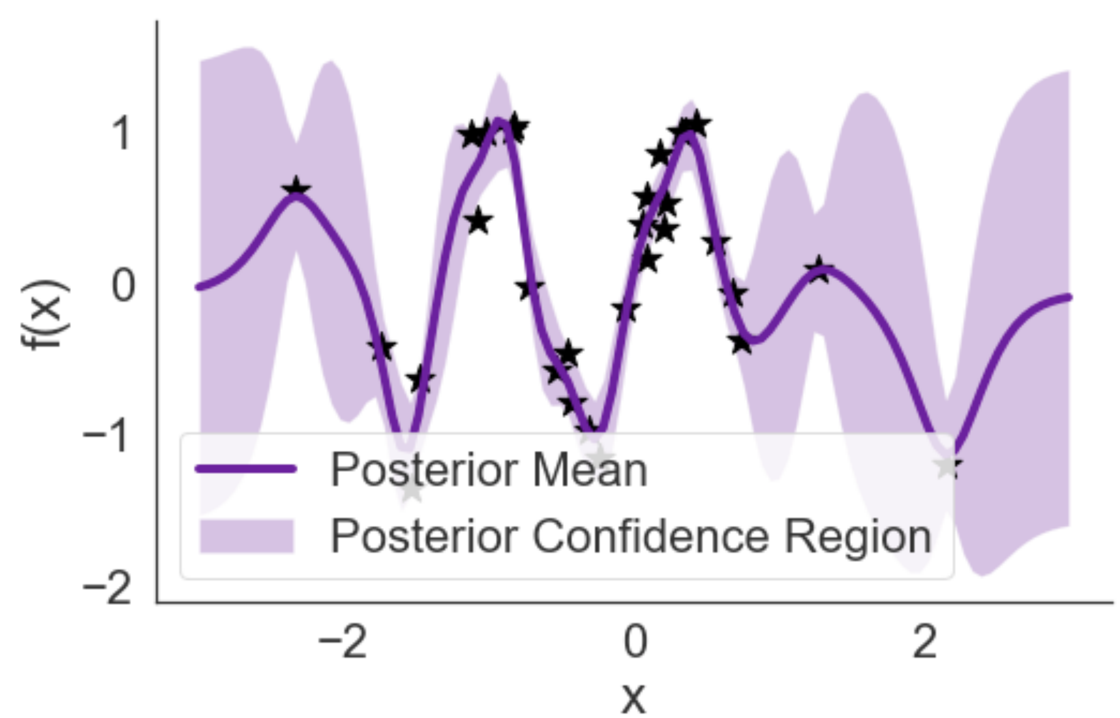
GAUSSIAN PROCESSES: UPDATING THE PREDICTIVE

- The predictive distribution is given by:

$$p(f^*|X^*, X, y) = \mathcal{N}(\mu_{f|\mathcal{D}}, \Sigma_{f|\mathcal{D}})$$

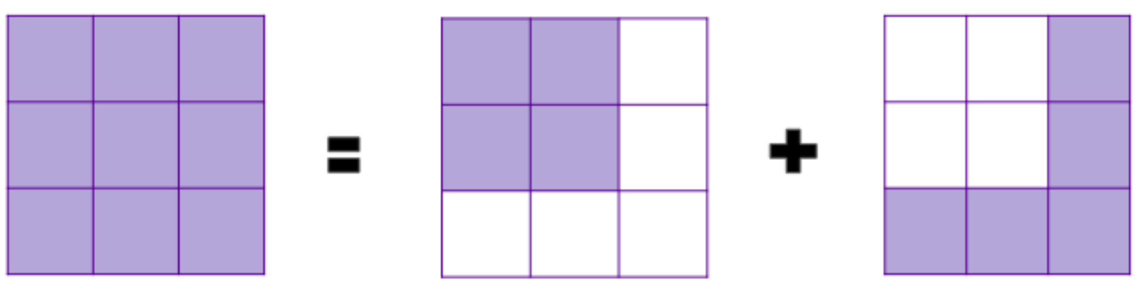
We need to update these terms

$$\mu_{f|\mathcal{D}} = K_{\mathbf{x}^* X}(K_{XX} + \sigma^2 I)^{-1}\mathbf{y},$$
$$\Sigma_{f|\mathcal{D}} = K_{\mathbf{x}^*\mathbf{x}^*} - K_{\mathbf{x}^* X}(K_{XX} + \sigma^2 I)^{-1}K_{X\mathbf{x}^*}.$$



GAUSSIAN PROCESSES: UPDATING THE PREDICTIVE

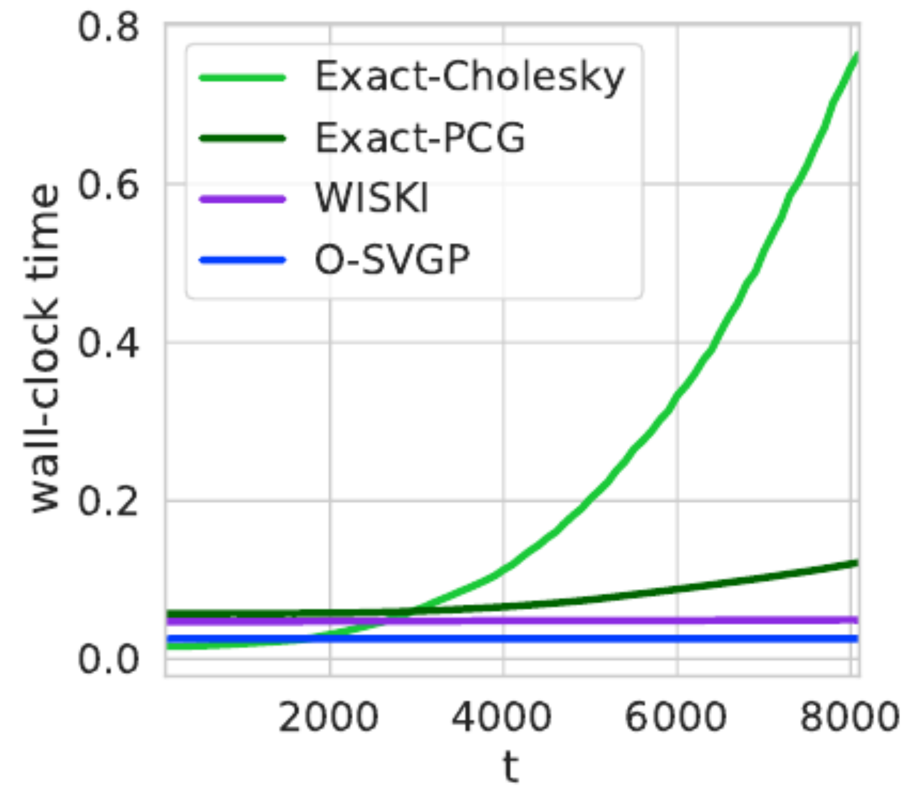
- But the training data covariance grows over time!!



$(\mathbf{K}_{X'X'} + \sigma^2 I)$ $(\mathbf{K}_{XX} + \sigma^2 I)$ $\begin{bmatrix} \mathbf{0} & k(X, \mathbf{x}') \\ k(\mathbf{x}', X) & k(\mathbf{x}', \mathbf{x}') + \sigma^2 \end{bmatrix}$

Low-rank updates cost at least **O(kn)** even if using low-rank decompositions (e.g. Lanczos).

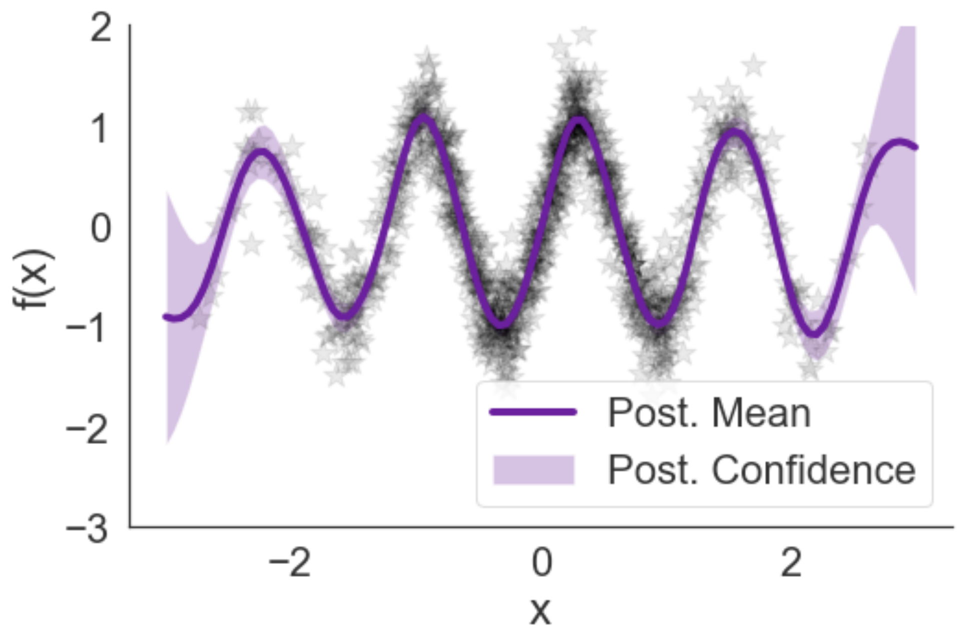
E.g. Jiang et al, NeurIPS, '20, Osborne, '10

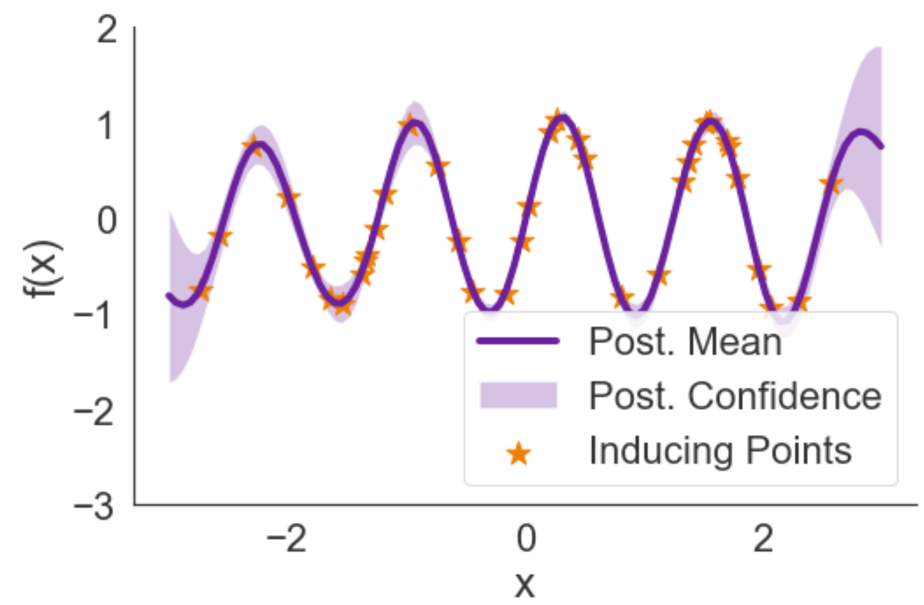


Time per iteration as more data points are observed.

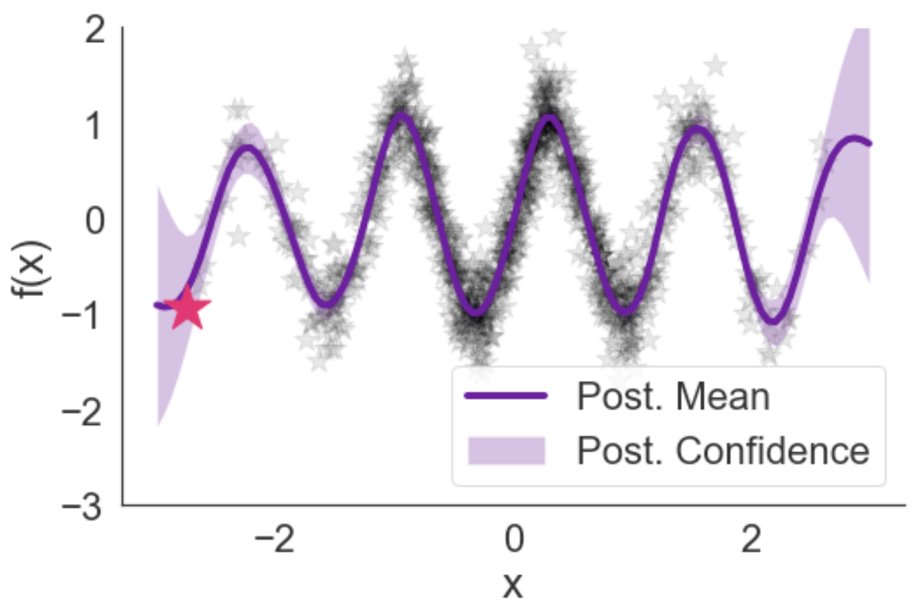
From Stanton et al, AISTATS, '21

APPROXIMATE INFERENCE TO THE RESCUE, ALMOST





Stochastic variational GPs (Titsias, '09, Hensman et al, '13, '15) condition the GP on **"inducing points"** and optimize the ELBO wrt to the inducing points and their corresponding variational distribution, $q(u) = \mathcal{N}(\mathbf{m}, \mathbf{S})$.



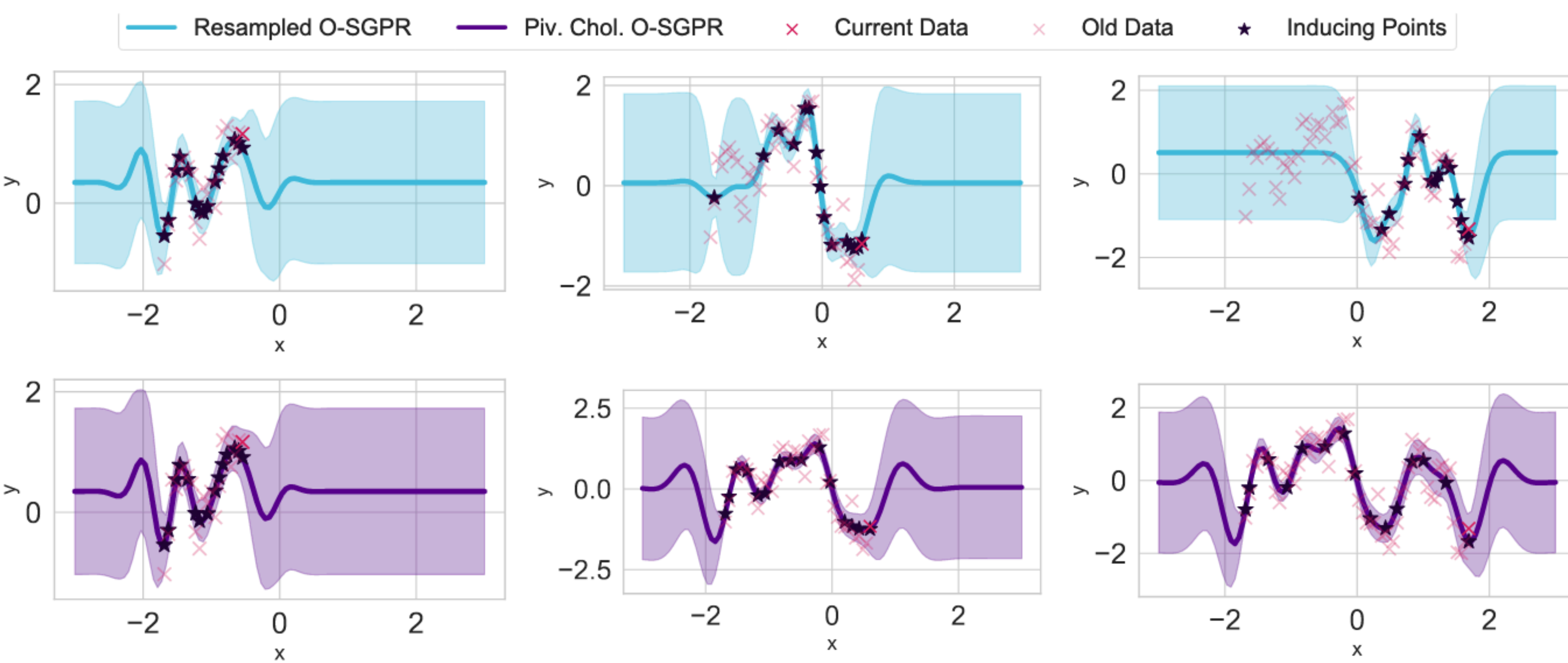
?

But it's not easy to update a SVGP wrt a new data pt

Bui et al, '17 provide one attempt

APPROXIMATE INFERENCE TO THE RESCUE!

- Update inducing points via a pivoted cholesky on the previous inducing points and the new data



(a) $t = 20$ (b) $t = 40$ (c) $t = 60$

APPROXIMATE INFERENCE TO THE RESCUE!

- Parameterize the variational distribution as

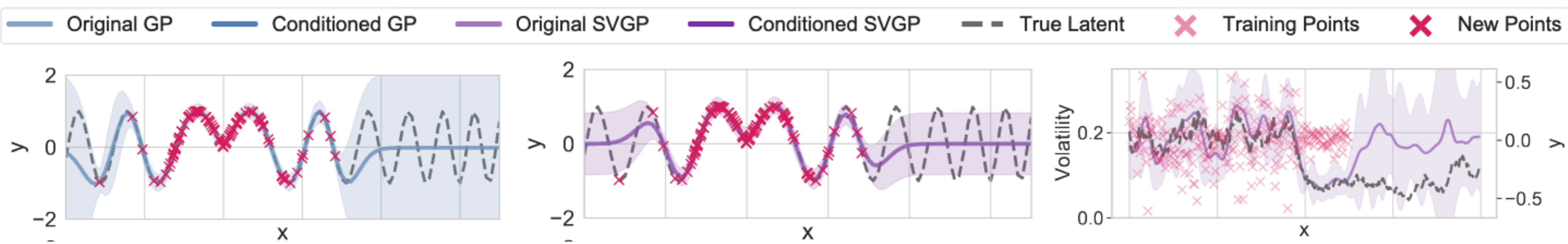
$$\mathbf{m} = K_{uu}(K_{uu} + C)^{-1}c \qquad \mathsf{S} = K_{uu}(K_{uu} + C)^{-1}K_{uu}$$

And optimize wrt c, C instead

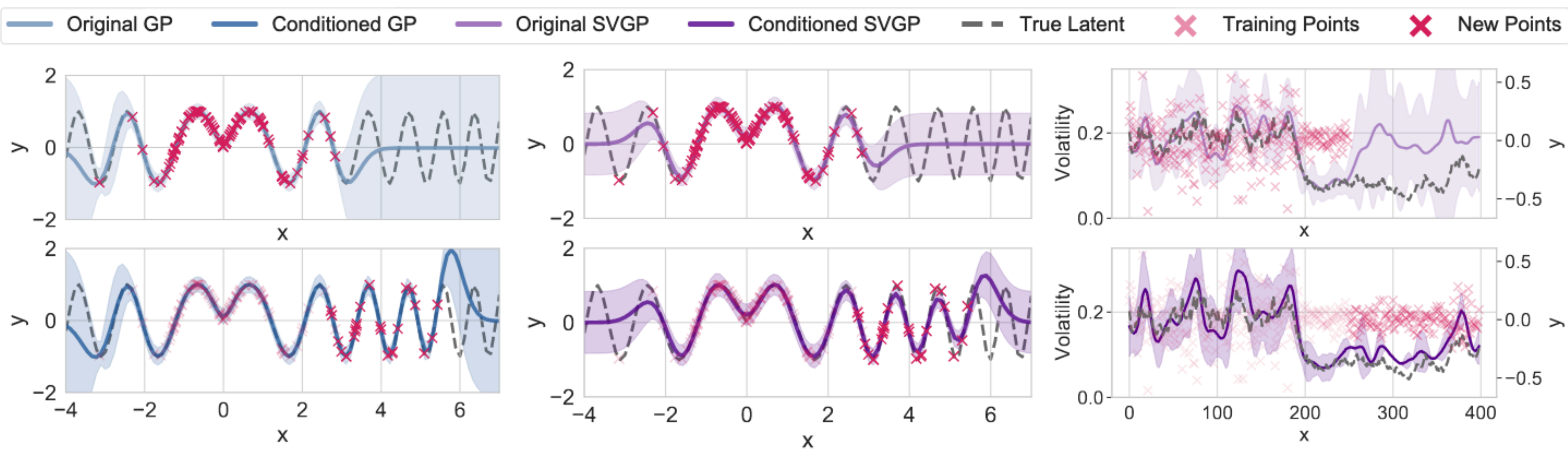
- Enables closed form updates to the variational distribution

$$\mathbf{c}_t = K_{\mathbf{u}\hat{\mathbf{v}}}\Sigma_{\hat{\mathbf{y}}}^{-1}\hat{\mathbf{y}} = K_{\mathbf{u}\mathbf{v}}\Sigma_{\mathbf{y}_t}^{-1}\mathbf{y}_t + K_{\mathbf{u}\mathbf{u}'}K'^{-1}_{\mathbf{u}'\mathbf{u}'}\mathbf{c}_t,$$
$$C_t = K_{\mathbf{u}\hat{\mathbf{v}}}\Sigma_{\hat{\mathbf{y}}}^{-1}K_{\hat{\mathbf{v}}\mathbf{u}} = K_{\mathbf{u}\mathbf{v}}\Sigma_{\mathbf{y}_t}^{-1}K_{\mathbf{v}\mathbf{u}} + K_{\mathbf{u}\mathbf{u}'}(K'^{-1}_{\mathbf{u}'\mathbf{u}'}C_{t-1}K'^{-1}_{\mathbf{u}'\mathbf{u}'})K_{\mathbf{u}\mathbf{u}'}.$$

CONDITIONED MODEL!



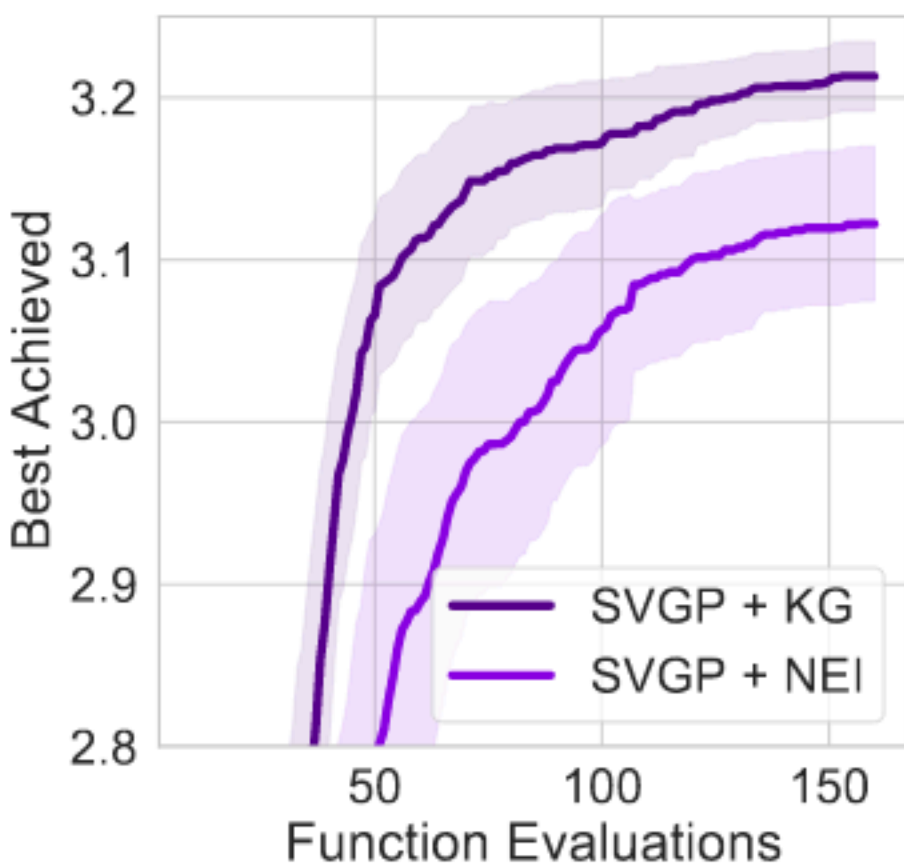
CONDITIONED MODEL!



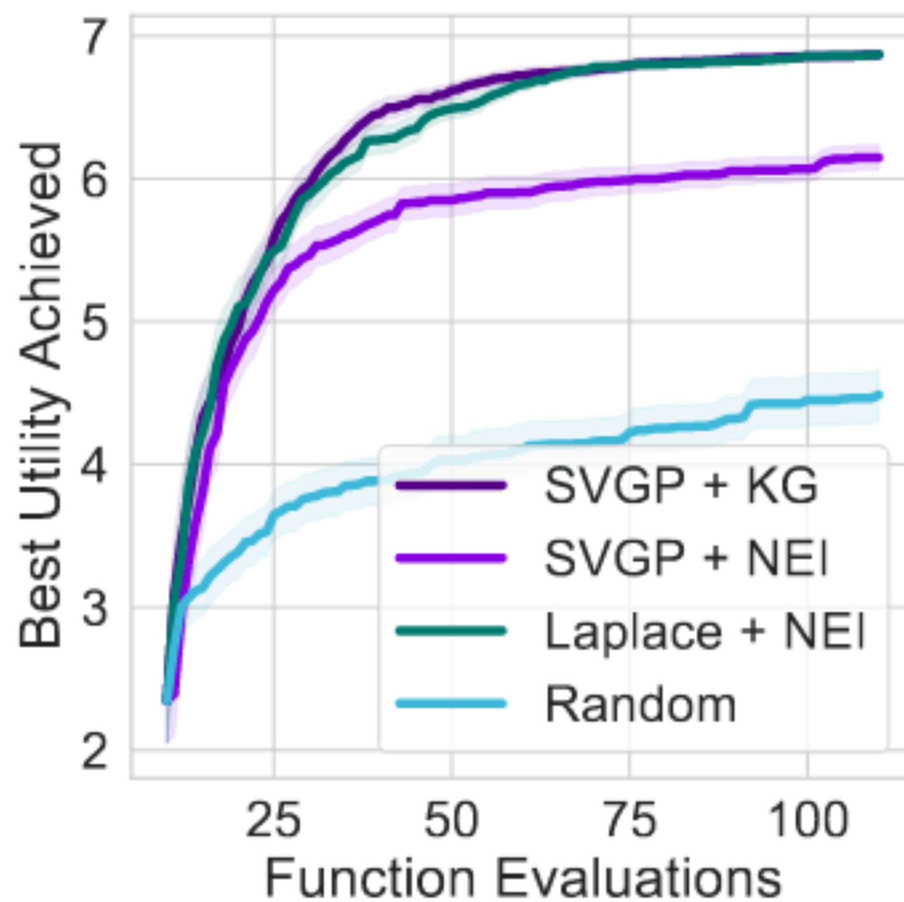
(a) Exact GP (b) SVGP + OVC (c) SVGP + OVC volatility model

NON GAUSSIAN BAYESIAN OPTIMIZATION PROBLEMS

- Enables "lookahead" bayesian optimization (like the **Knowledge gradient**) on non-Gaussian responses

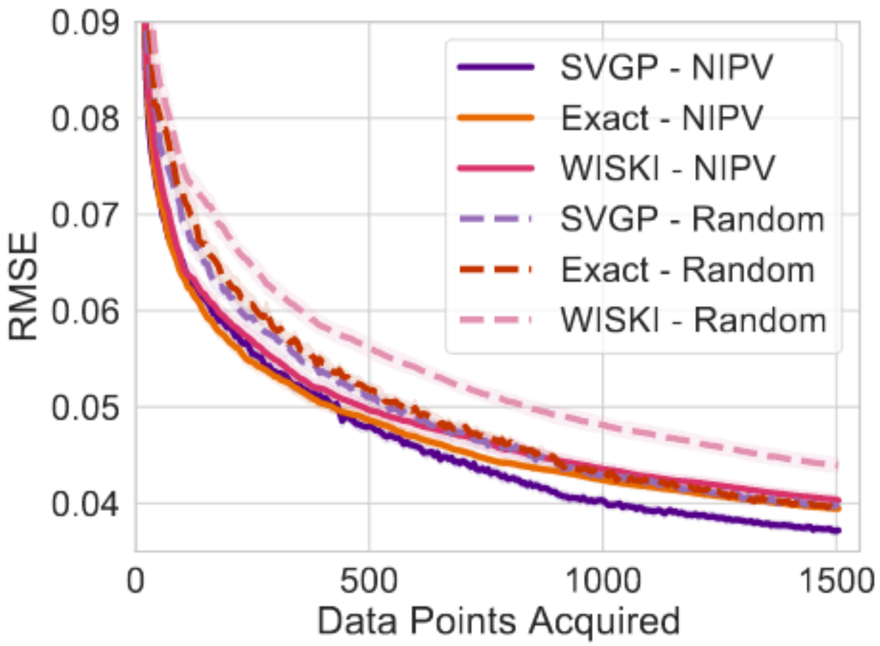


(c) Poisson-Hartmann6

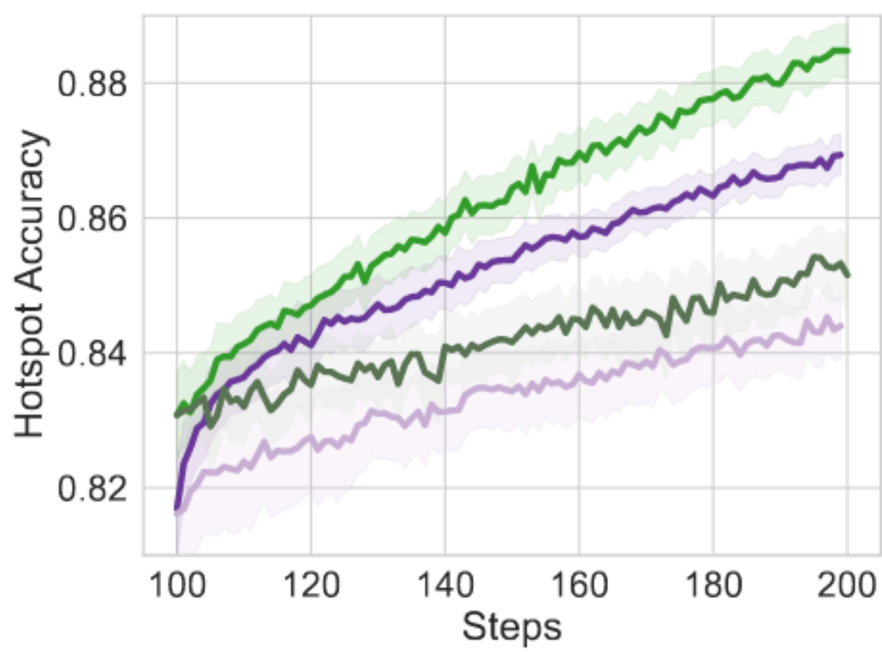


(d) Preference Learning

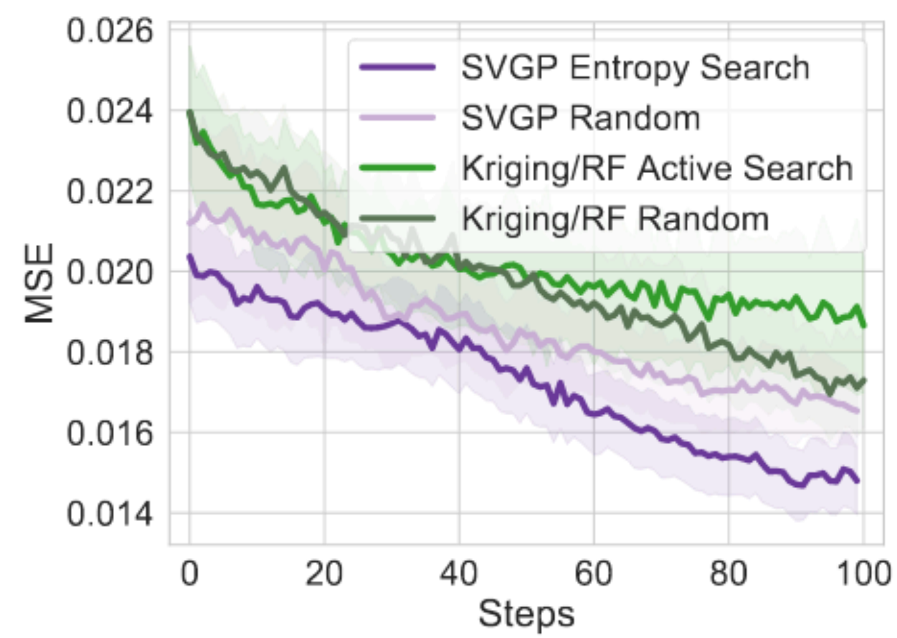
ACTIVE LEARNING



(a) Malaria incidence in Nigeria



(b) Hotspot prediction accuracy

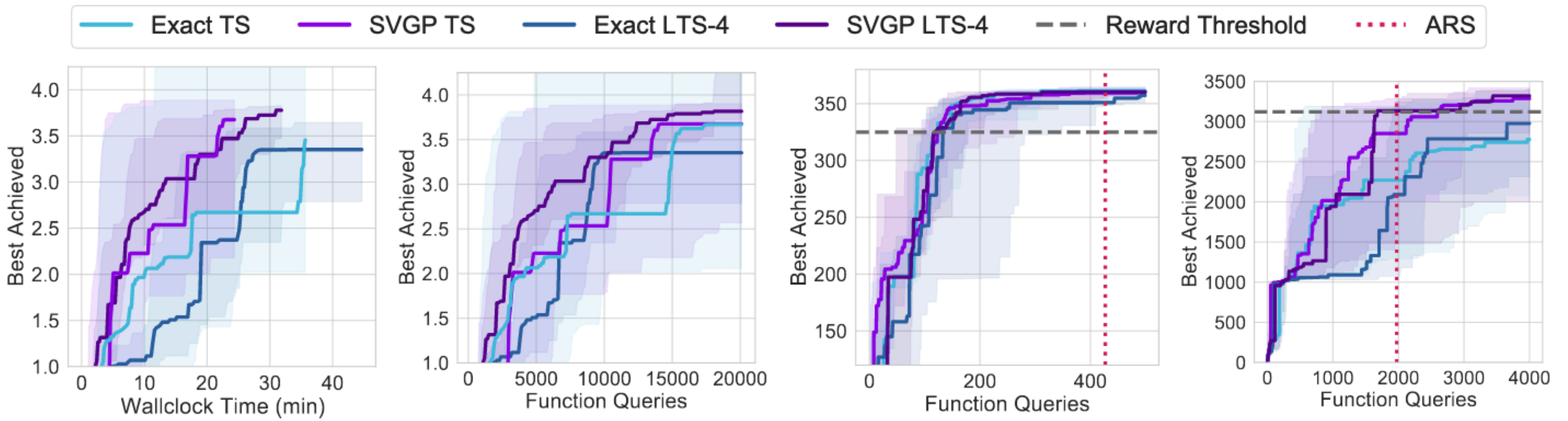


(c) Prevalence modelling

Regression problem

Binomial observations!

HIGH DIMENSIONAL BO / REINFORCEMENT LEARNING



(a) Wall-Clock time, rover, $d = 60$

(b) rover, $d = 60$

(c) swimmer, $d = 16$

(d) hopper, $d = 33$

OVC is faster than sub-sampled exact GPs

Even on mujoco problems!

And performs better due to numerical stability

Optimization methods: TrBO (Eriksson et al, NeurIPS, '19) for rover, LaMCTS + TrBO (Wang et al, NeurIPS, '20) for swimmer / hopper.

PAPER AT: HTTPS://NIPS.CC/CONFERENCES/2021/SCHEDULEMULTITRACK?EVENT=26863

CODE AT: HTTPS://GITHUB.COM/WJMADDOX/VOLATILITYGP

Coming soon to GPyTorch and BoTorch

Thanks

Contact: wjm363 at nyu.edu