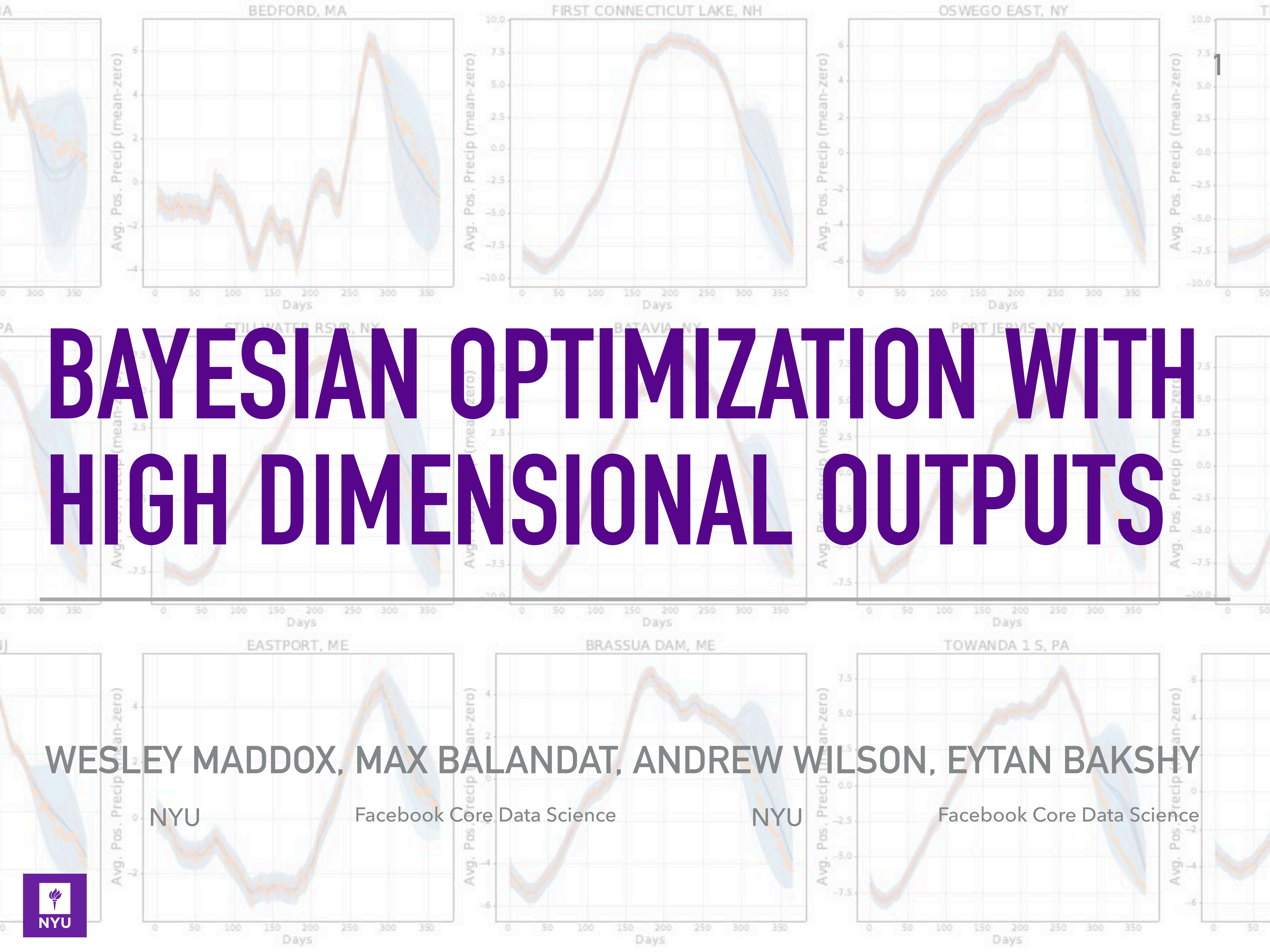
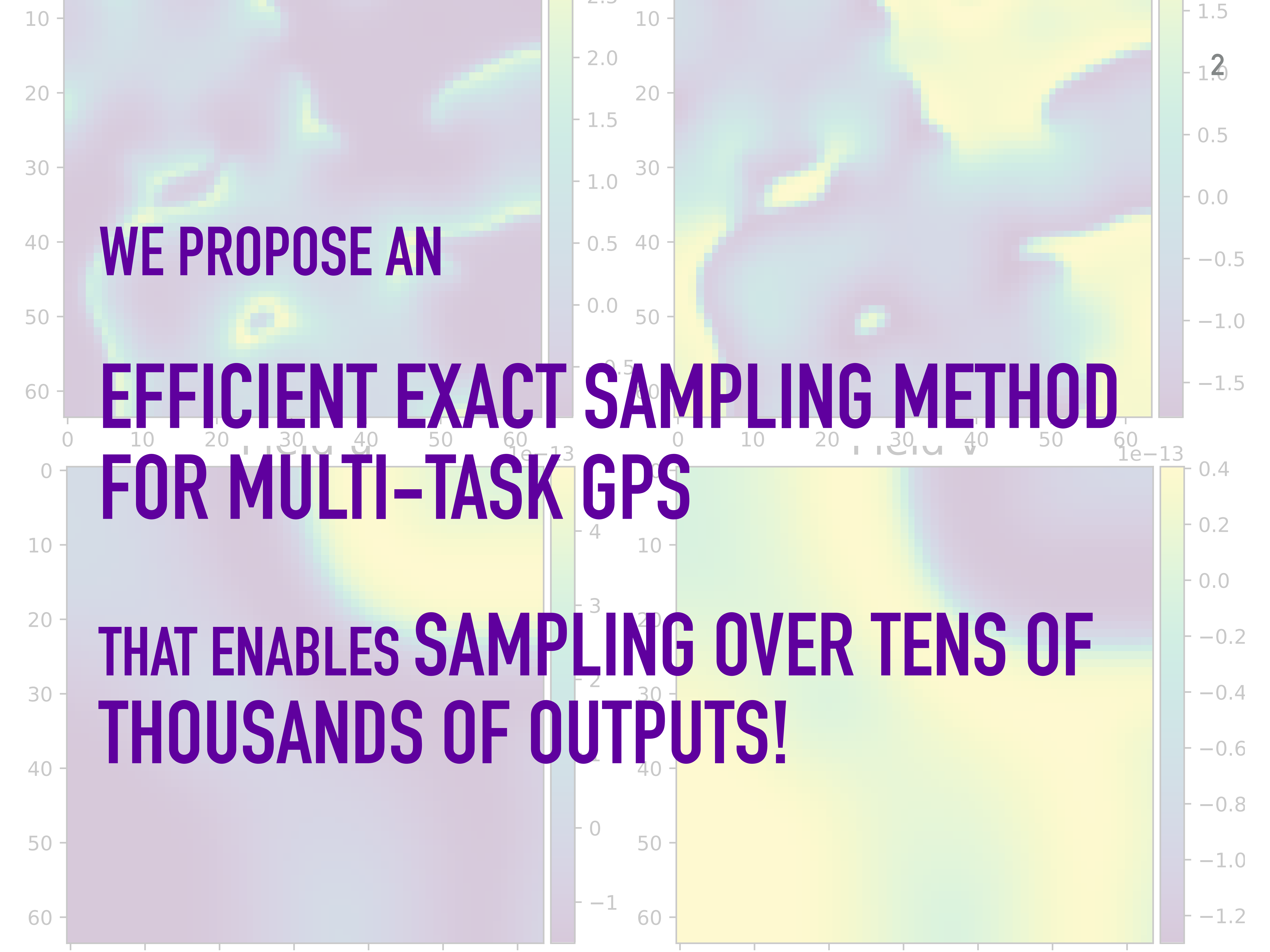


BAYESIAN OPTIMIZATION WITH HIGH DIMENSIONAL OUTPUTS

WESLEY MADDOX, MAX BALANDAT, ANDREW WILSON, EYTAN BAKSHY

NYU Facebook Core Data Science NYU Facebook Core Data Science



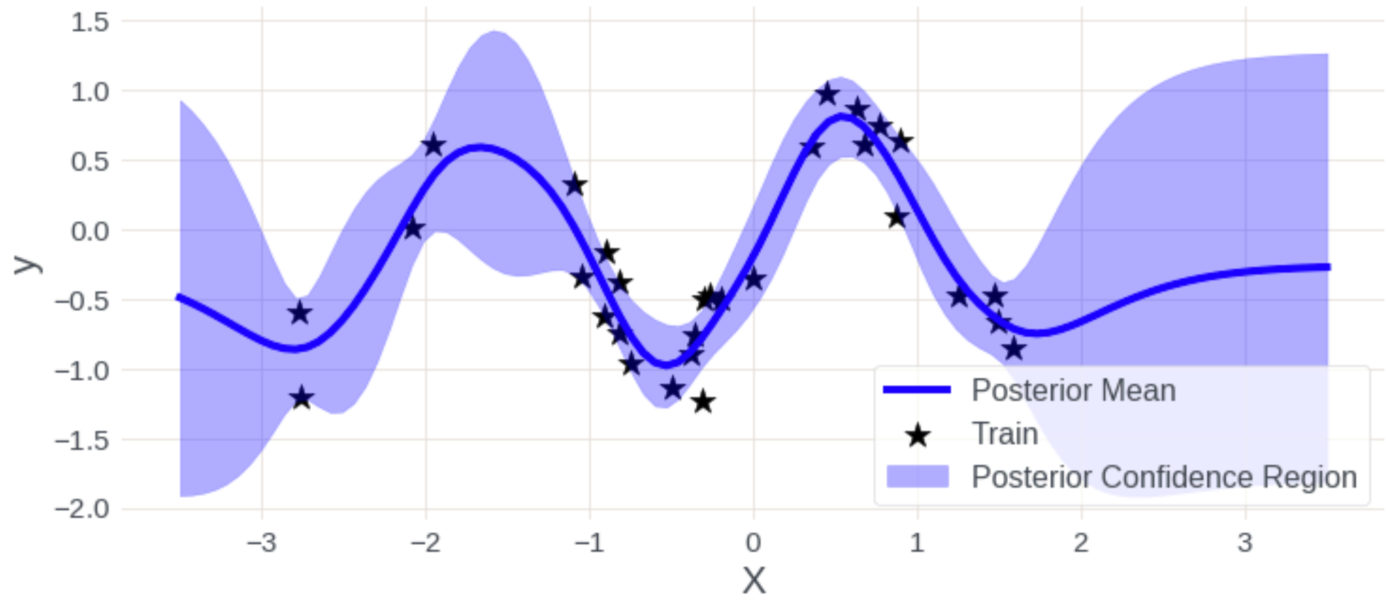
WE PROPOSE AN

EFFICIENT EXACT SAMPLING METHOD FOR MULTI-TASK GPS

THAT ENABLES **SAMPLING OVER TENS OF THOUSANDS OF OUTPUTS!**

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)

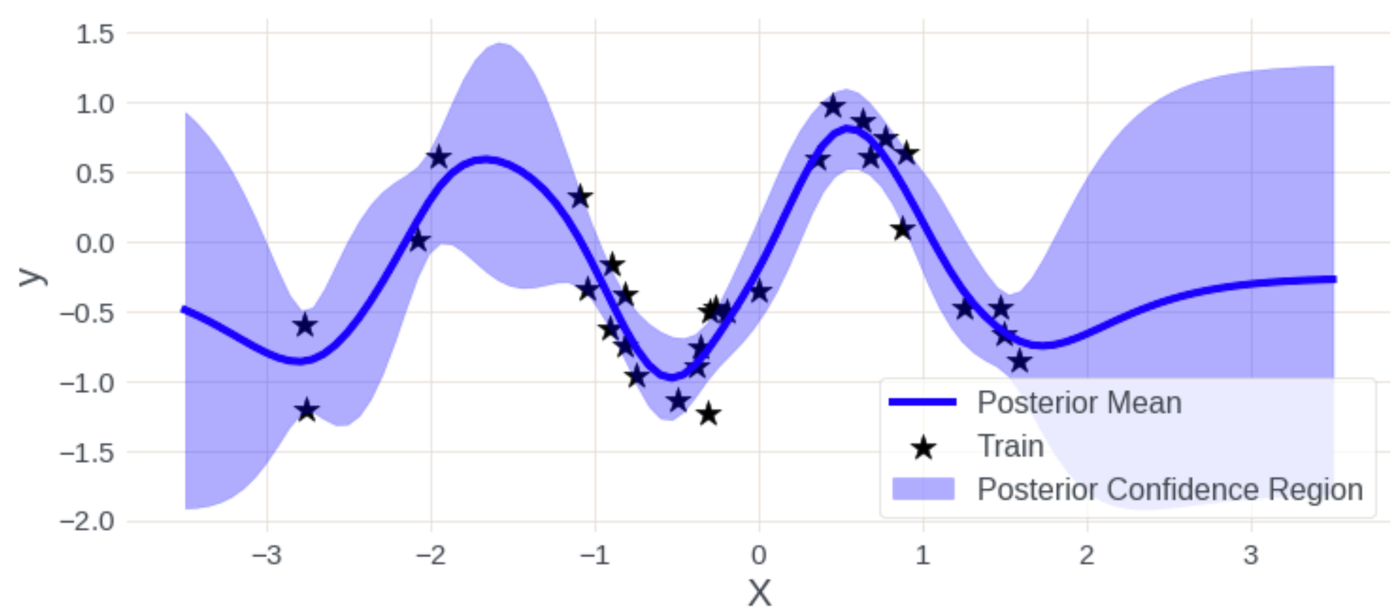


GAUSSIAN PROCESSES: PREDICTION

- The predictive distribution is given by:

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

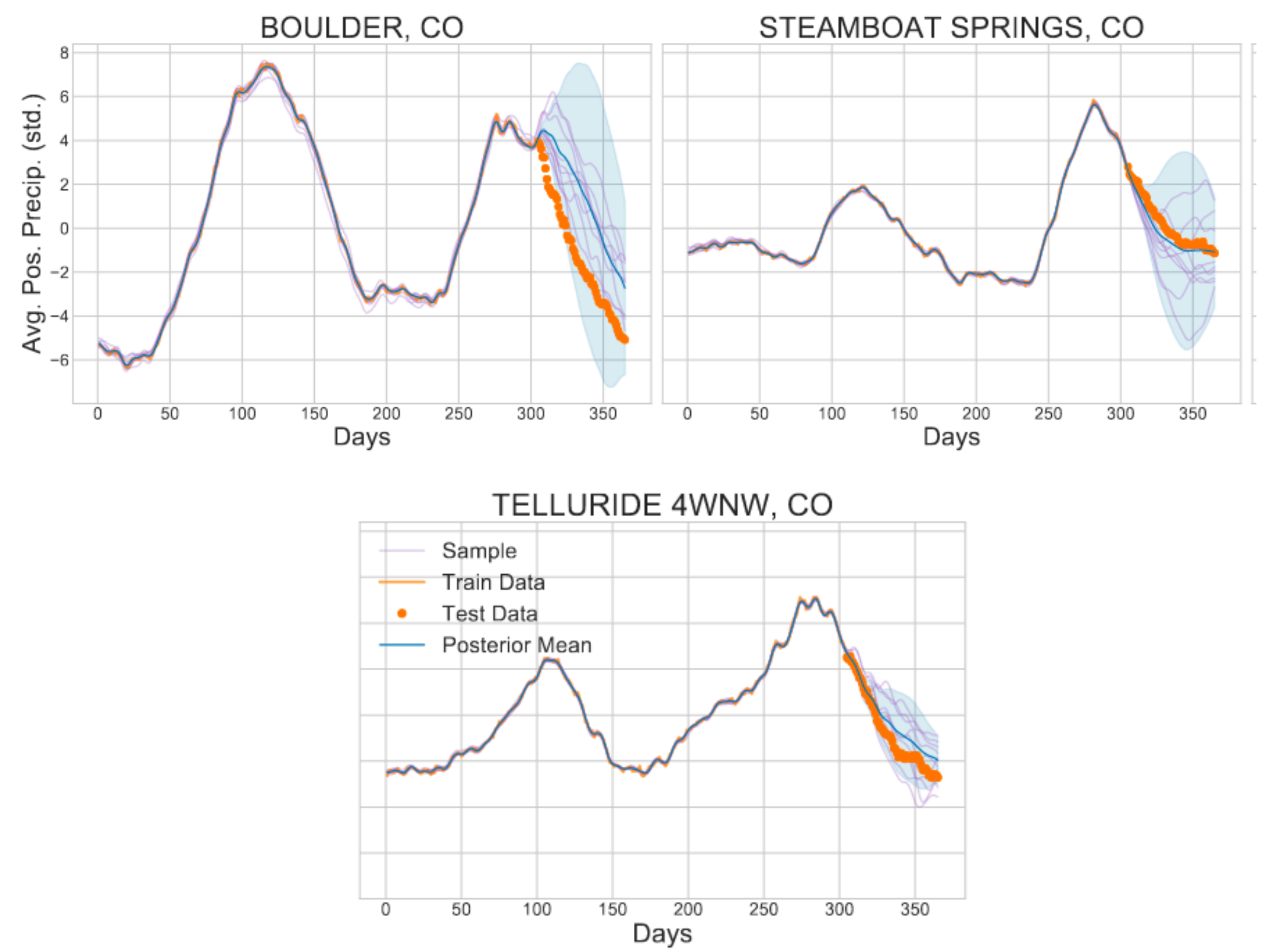
$$\mu_{f|\mathcal{D}} = K_{\mathbf{x}^* X}(K_{XX} + \sigma^2 I)^{-1}\mathbf{y},$$
$$\Sigma_{f|\mathcal{D}} = K_{\mathbf{x}^*\mathbf{x}^*} - K_{\mathbf{x}^* X}(K_{XX} + \sigma^2 I)^{-1} K_{X\mathbf{x}^*}.$$



MULTI TASK GAUSSIAN PROCESSES

- Model multiple outputs that are related
  - Typically separate data covariance from task covariance
  - $\mathrm{vec}(y) \sim \mathcal{N}(0, K_{XX} \otimes K_T).$

Nt x nt matrix

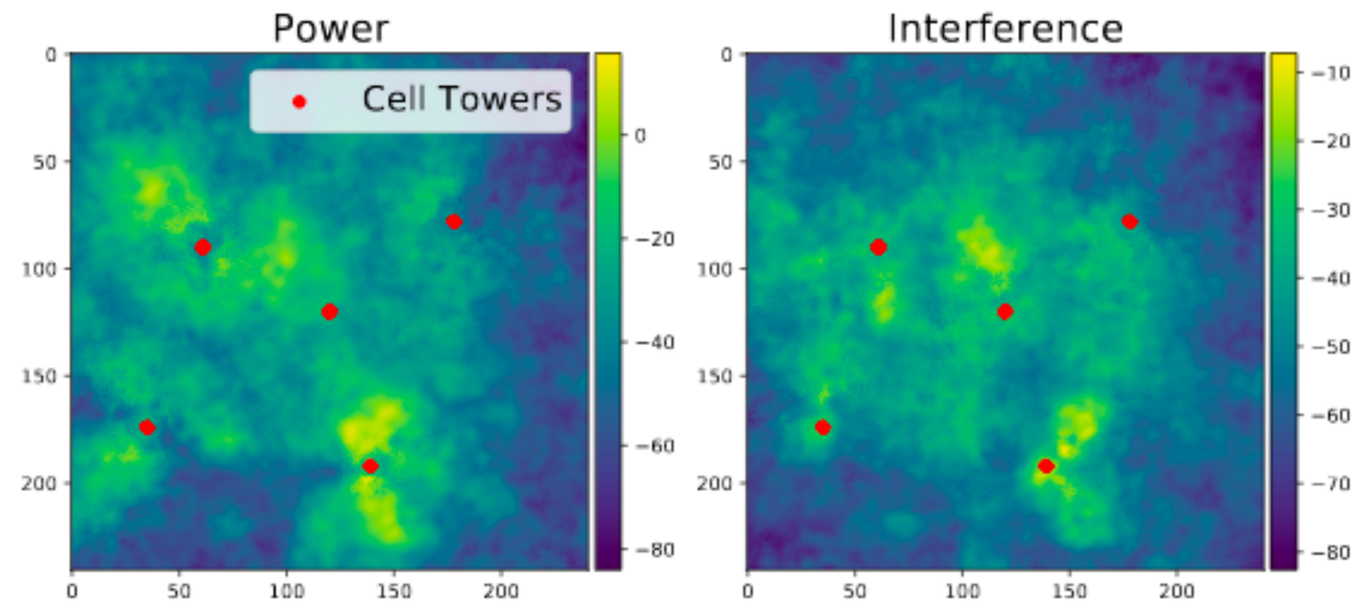


MULTI TASK GAUSSIAN PROCESSES

- Model multiple outputs that are related
  - Typically separate data covariance from task covariance
  - $\mathrm{vec}(y) \sim \mathcal{N}(0, K_{XX} \otimes K_T).$

Nt x nt matrix

(50*5000) x (50*5000)



Cell tower interference: given location + angle of towers, how can we model power and interference?

50 x 50 x 2 tensors **(5000 outputs)**

Posterior is not Kronecker structured

MULTI TASK GAUSSIAN PROCESSES: PREDICTION

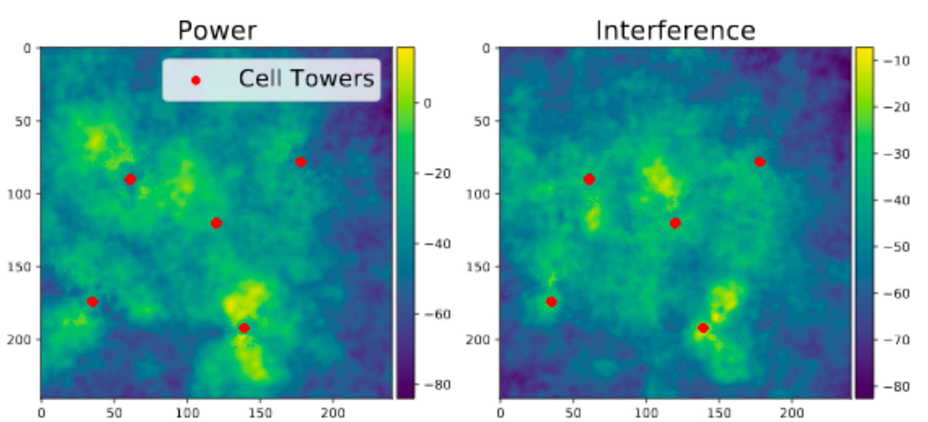
Predictive distribution is closed form: $p(f^*|X^*, X, y) = \mathcal{N}\left(\mu^*, \Sigma^*\right)$

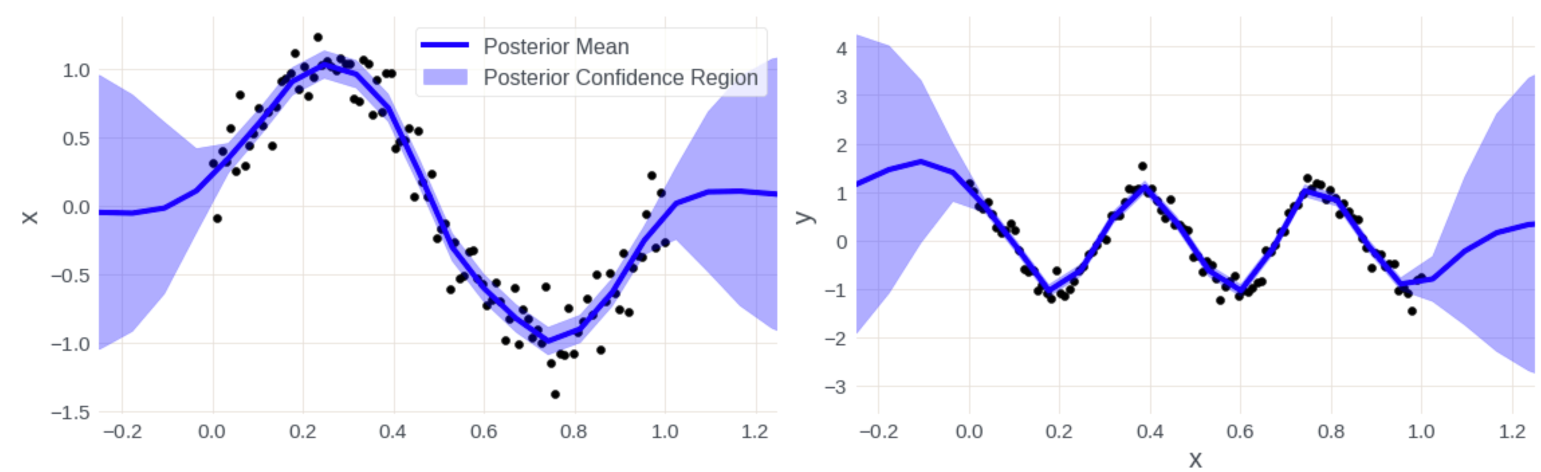
$$\mu^* = (K_{x^*,X} \otimes K_T)(K_{XX} \otimes K_T + \sigma^2 I_{nT})^{-1} y$$

$$\Sigma^* = (K_{x^*,x^*} \otimes K_T) - (K_{x^*,X} \otimes K_T)(K_{XX} \otimes K_T + \sigma^2 I_{nT})^{-1}(K_{x^*,X}^\top \otimes K_T)$$

This matrix is no longer Kronecker structured, and it gets really big!

50 data points. 5000 outputs ==> Σ^* is (50*5000) x (50*5000)





MATHERON'S RULE FOR SAMPLING GAUSSIAN PROCESS POSTERIORS

Can sample from conditional Gaussian random variables via Matheron's rule (1970s)

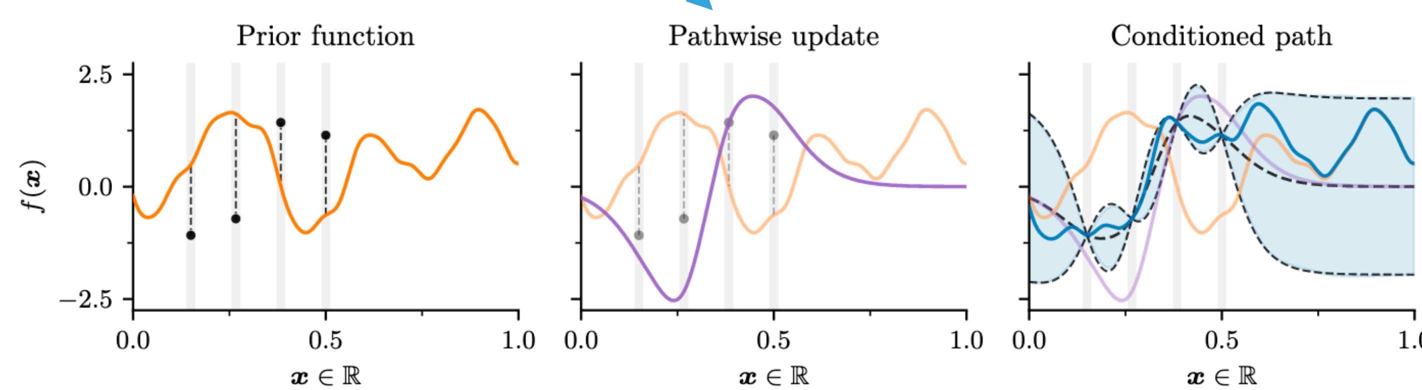
$$x|Y = y \overset{d}{=} x + \mathrm{Cov}(x, y)(\mathrm{Cov}(y, y)^{-1})(y - Y).$$

For Gaussian processes, this becomes

$$f^*|(Y = y) \overset{d}{=} f^* + K_{x_{\text{test}}X}(K_{XX} + \sigma^2 I)^{-1}(y - Y - \epsilon)$$

Steps:

1) Draw (f^*, Y) from joint prior
2) Draw iid noise epsilon
3) Compute equation



From "Efficiently sampling functions from Gaussian Process posteriors," Wilson et al, ICML, 2020

MATHERON'S RULE: MULTITASK SETTING

Can sample from conditional Gaussian random variables via Matheron's rule (1970s)

$$f^*|(Y = y) \stackrel{d}{=} f^* + K_{x_{\text{test}}X}(K_{XX} + \sigma^2 I)^{-1}(y - Y - \epsilon)$$

Prior function comes from $(f^*, Y) \sim \mathcal{N}(0, K_{(x_{\text{test}}X),(x_{\text{test}}X)})$,

Which is structured (e.g. efficient sampling)

$$\begin{aligned} K_{\text{mt},(x_{\text{test}}X),(x_{\text{test}}X)} &= K_{(x_{\text{test}}X),(x_{\text{test}}X)} \otimes K_T = \tilde{R}\tilde{R}^\top \otimes LL^\top \\ &= (\tilde{R} \otimes L)(\tilde{R} \otimes L)^\top \end{aligned}$$

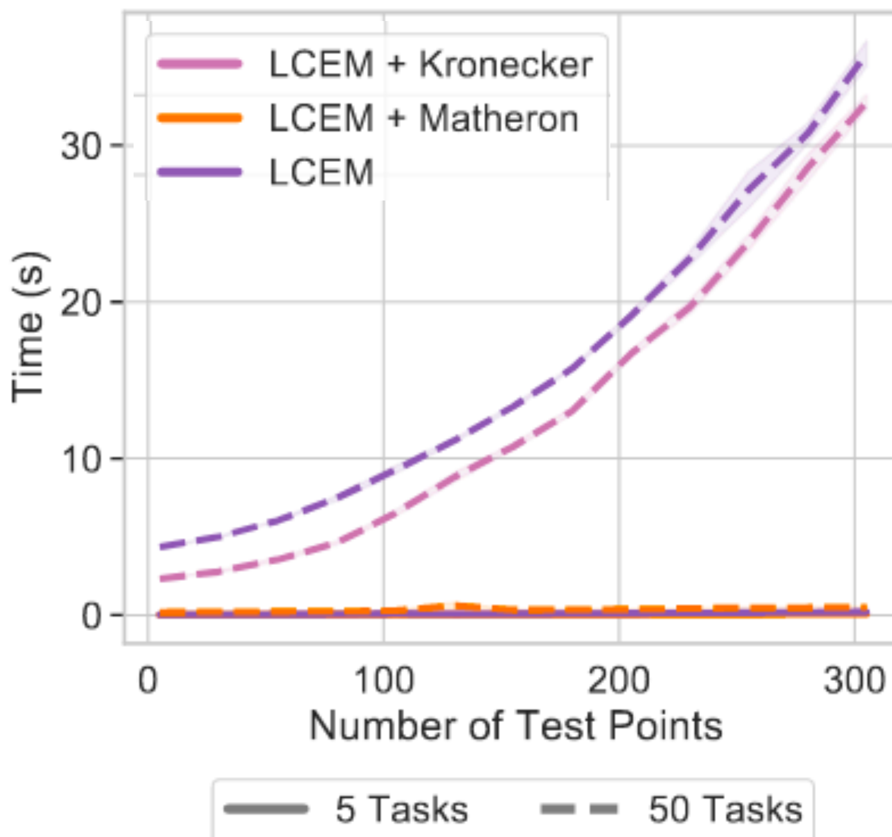
Pathwise update is a structured solve and a Kronecker MVM.

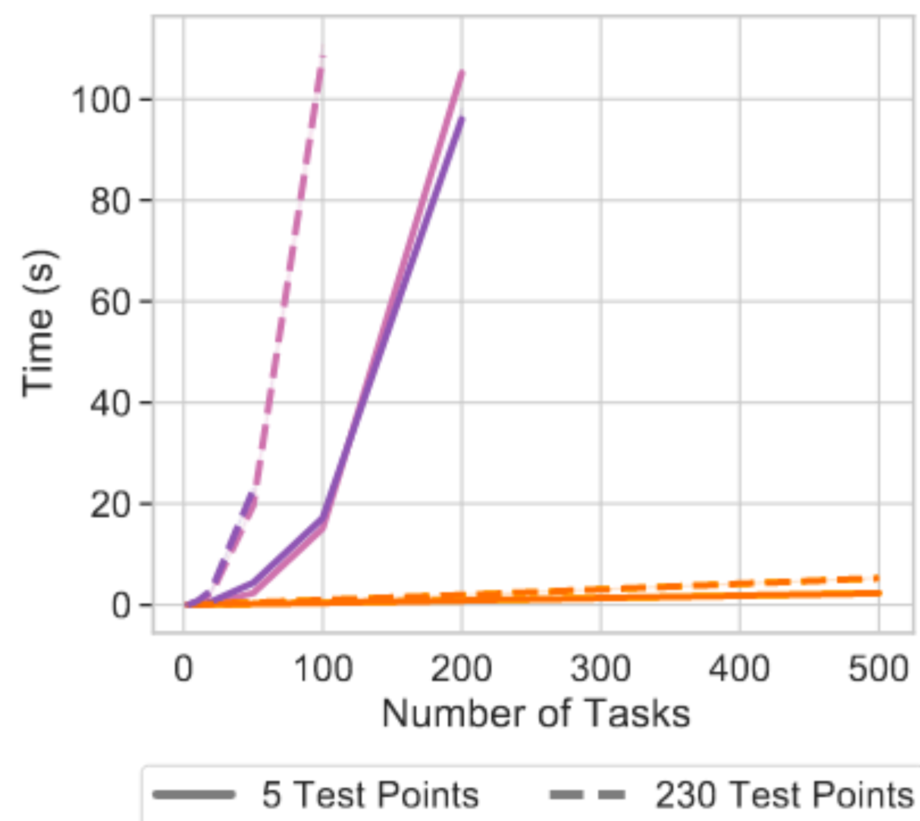
$$\begin{aligned} &(\otimes_{i=}^d K_i + \sigma^2 I)^{-1} z = \\ &\otimes_{i=1}^d Q_i (\otimes_{i=1}^d \Lambda_i + \sigma^2 I)^{-1} \otimes_{i=1}^d Q_i^\top z. \end{aligned}$$

Posterior sampling is O(n^3 + t^3) time rather than O($n^3 t^3$) time.

EMPIRICAL RESULTS

LCEM: contextual multi-task GP



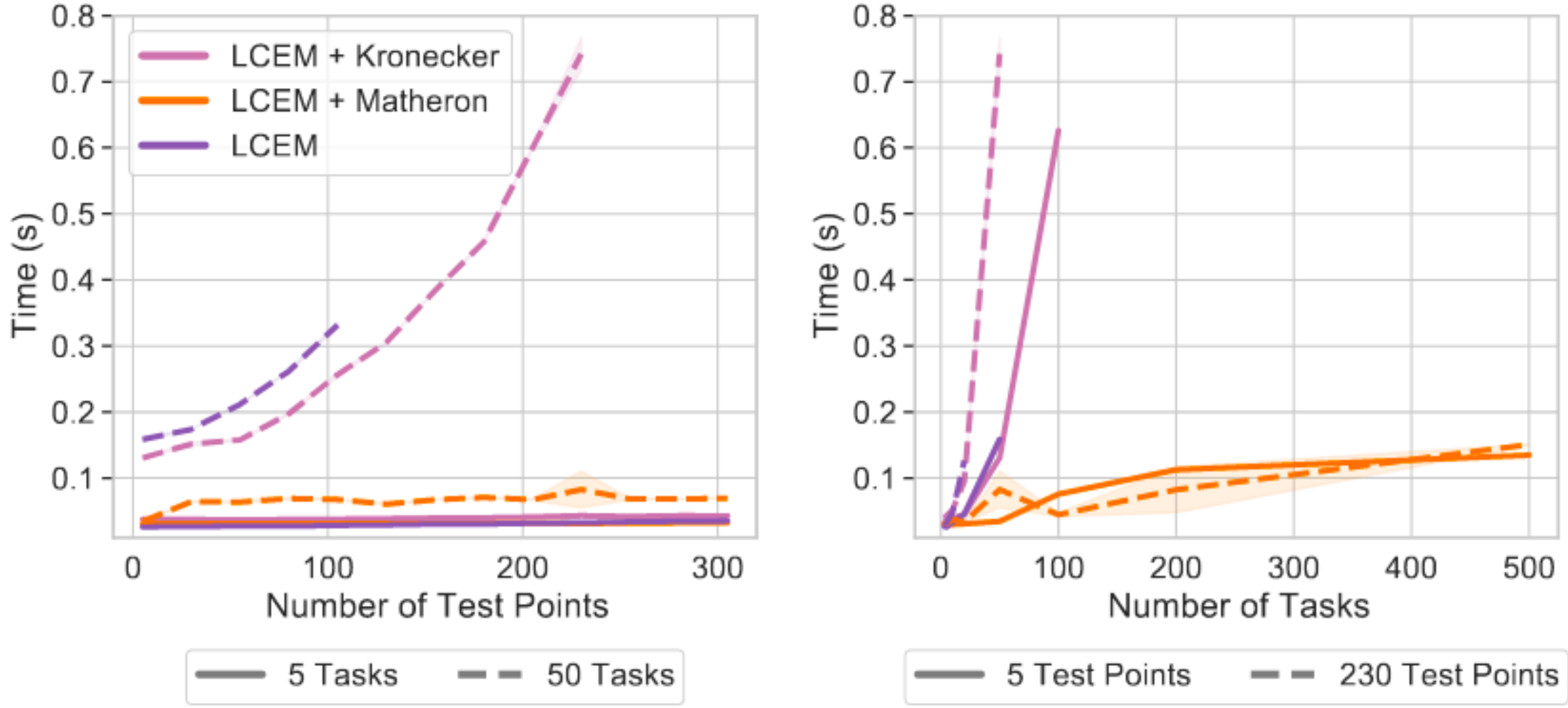


Fixed training points, results on CPU.

Using Matheron's rule allows efficient posterior sampling to many tasks and test points

EMPIRICAL RESULTS

LCEM: contextual multi-task GP

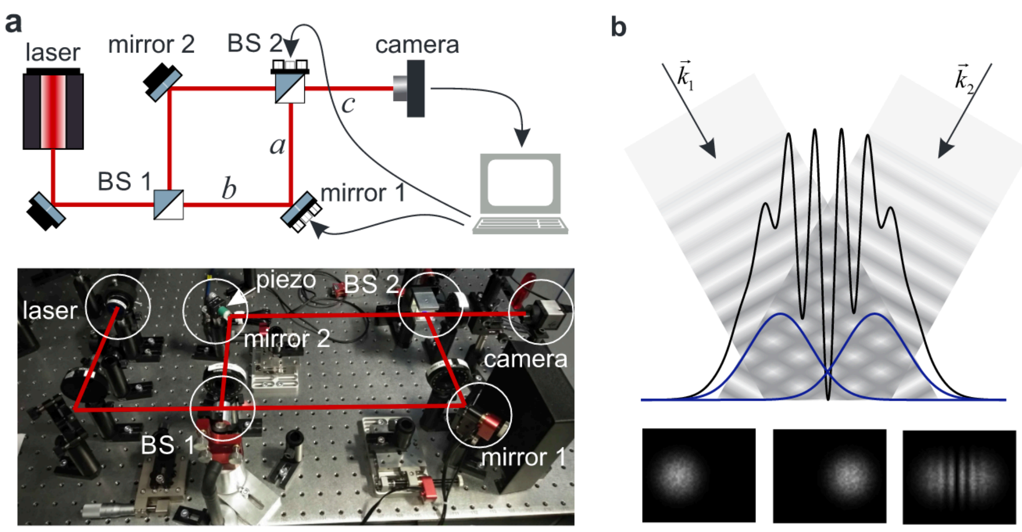


Fixed training points, results on GPU (less memory).

Using Matheron's rule allows efficient posterior sampling to many tasks and test points

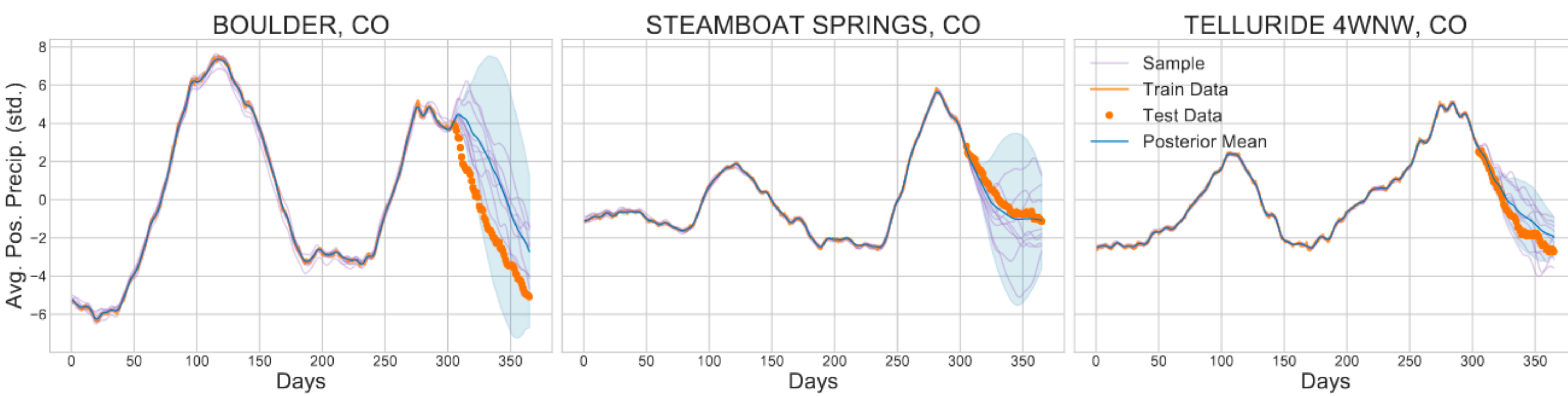
APPLICATIONS OF GAUSSIAN PROCESSES

- Tuning expensive models
- (Bayesian optimization)



From “Interferobot,” Sorokin et al, NeurIPS, 2020

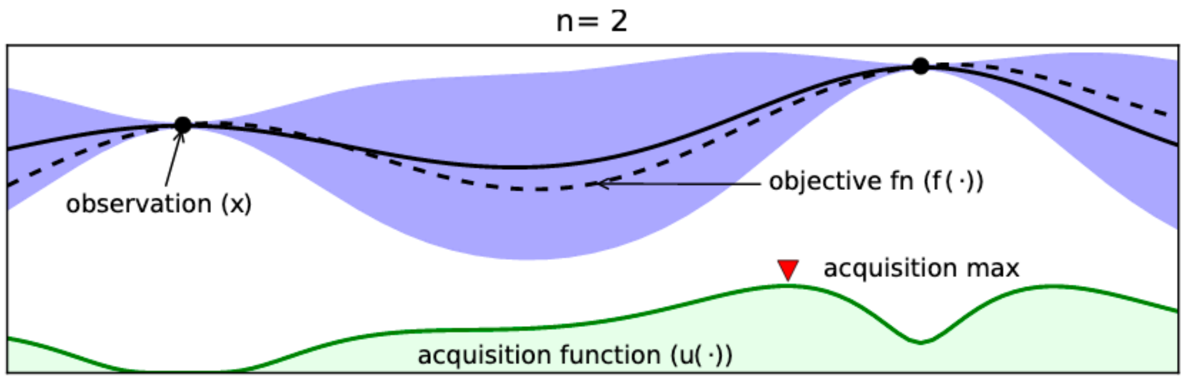
- Continual (e.g. time series) modeling



From Benton et al, NeurIPS, 2019

BAYESIAN OPTIMIZATION INTRO

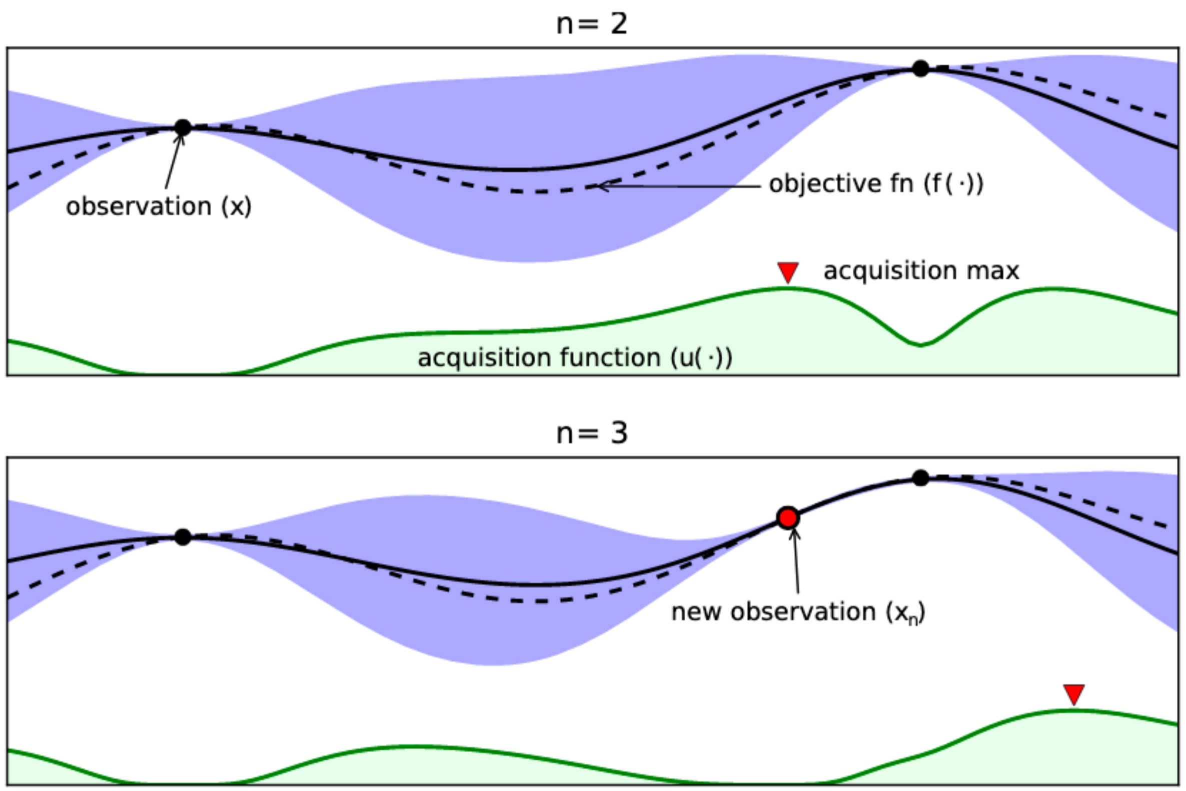
- Goal: $\max_x f(x)$
  - F is costly to evaluate
  - Make minimal assumptions about problem
  - X is low-dimensional
- Approach:
  - Build a probabilistic *surrogate* model
  - Suggest new points by optimizing an *acquisition function* on the surrogate



From Shahriari et al, '16

BAYESIAN OPTIMIZATION INTRO

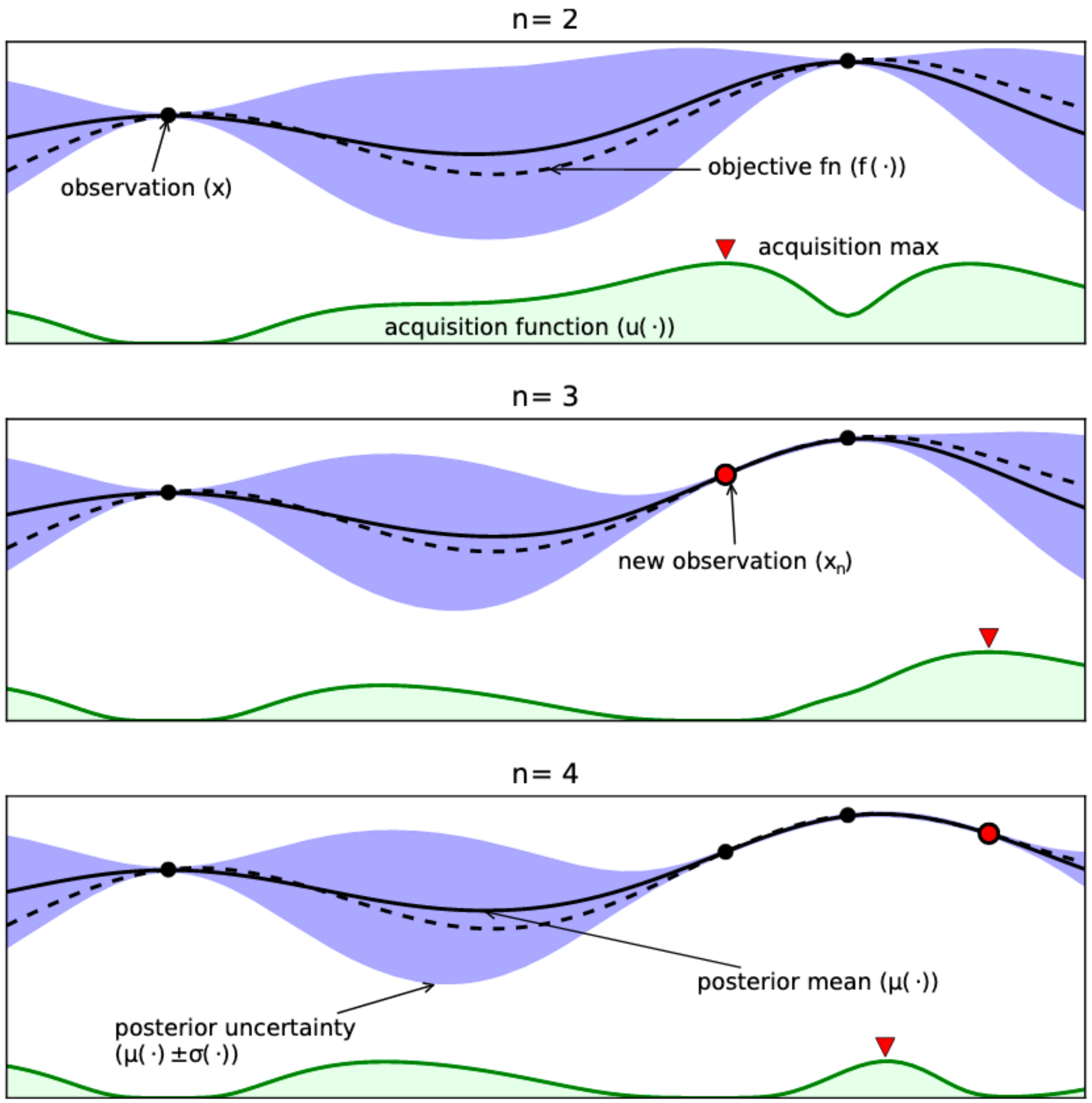
- Goal: $\max_x f(x)$
  - F is costly to evaluate
  - Make minimal assumptions about problem
  - X is low-dimensional
- Approach:
  - Build a probabilistic *surrogate* model
  - Suggest new points by optimizing an *acquisition function* on the surrogate



From Shahriari et al, '16

BAYESIAN OPTIMIZATION INTRO

- Goal: $\max_x f(x)$
  - F is costly to evaluate
  - Make minimal assumptions about problem
  - X is low-dimensional
- Approach:
  - Build a probabilistic *surrogate* model
  - Suggest new points by optimizing an *acquisition function* on the surrogate

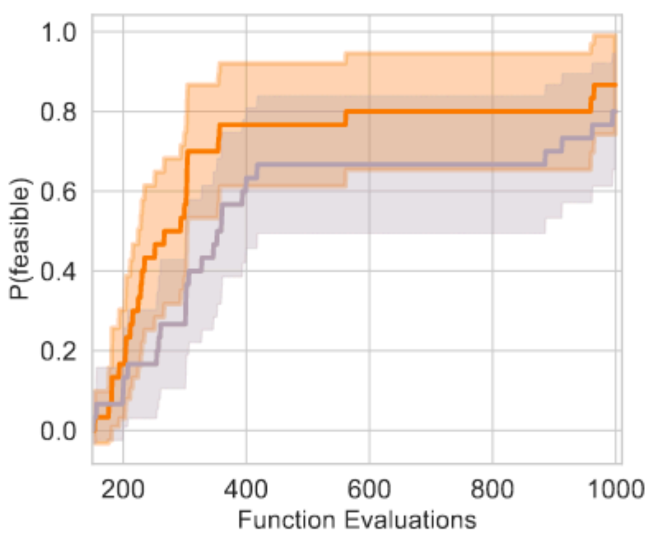


From Shahriari et al, '16

LARGE SCALE CONSTRAINED BAYESIAN OPTIMIZATION

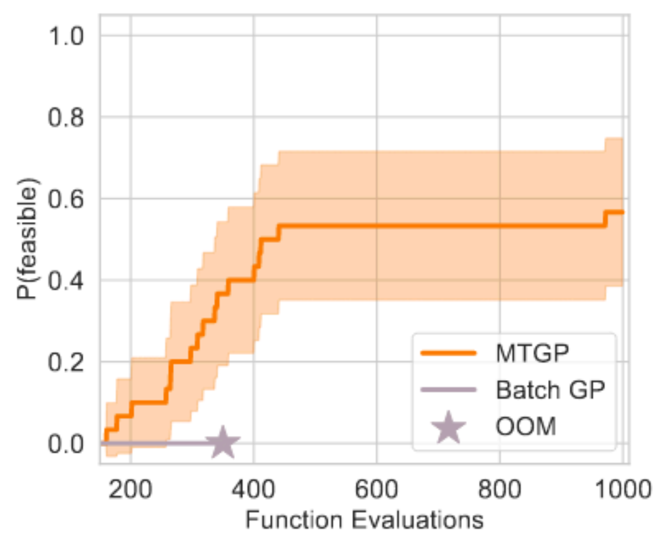
Goal: optimize function subject to black box constraints (model these as well)

$$\arg\min_x f(x) \quad \text{s.t.} \quad c_i(x) \leq 0 \quad \forall i \in \{1, \cdots, m\}.$$



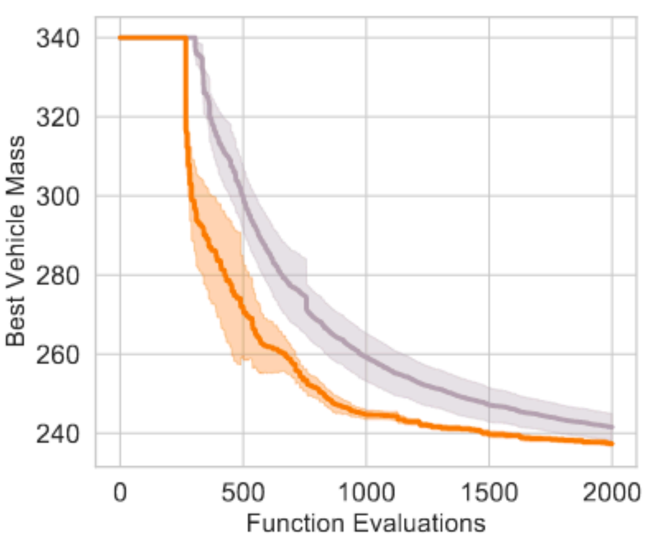
Feasibility, Lunar Lander, $m = 50$

51 Tasks!



Feasibility, Lunar Lander, $m = 100$

101 Tasks!



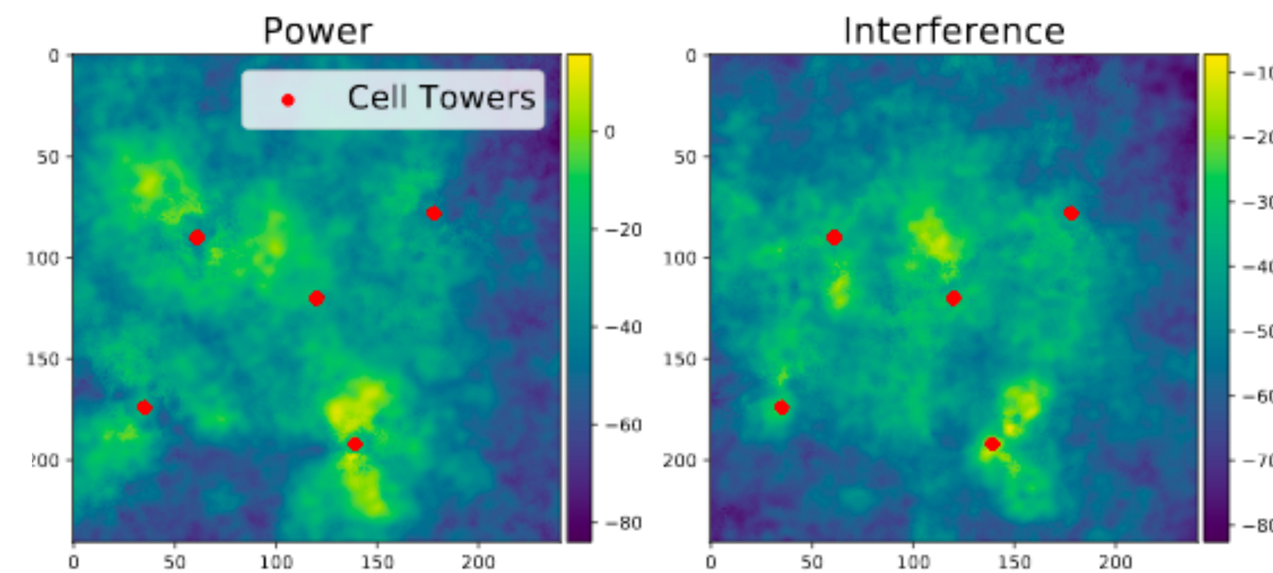
Vehicle mass, MOPTA08, $m = 68$

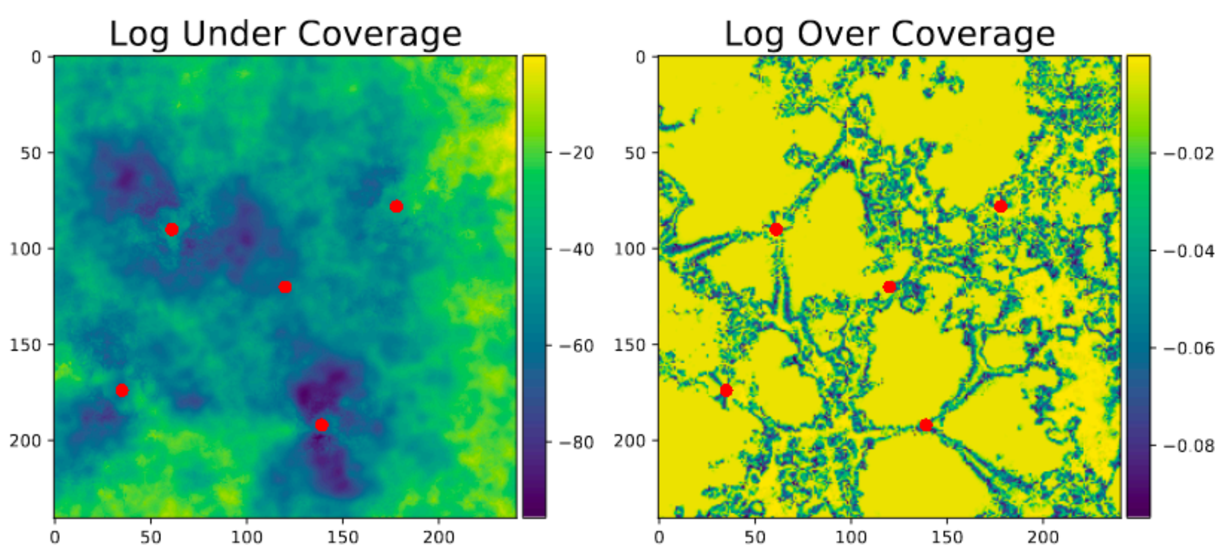
69 Tasks!

Using Scalable Constrained Bayesian Optimization, Eriksson & Poloczek, UAI, '20

LARGE SCALE COMPOSITE BAYESIAN OPTIMIZATION

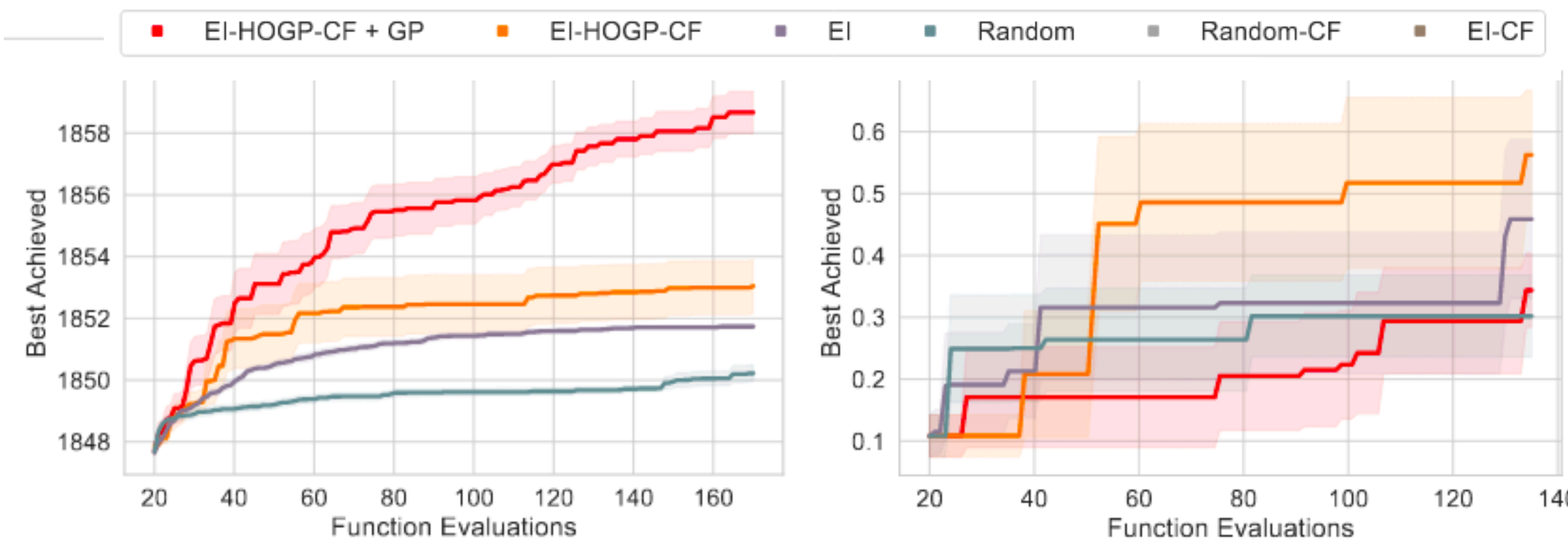
Objective is a nonlinear function of the responses:





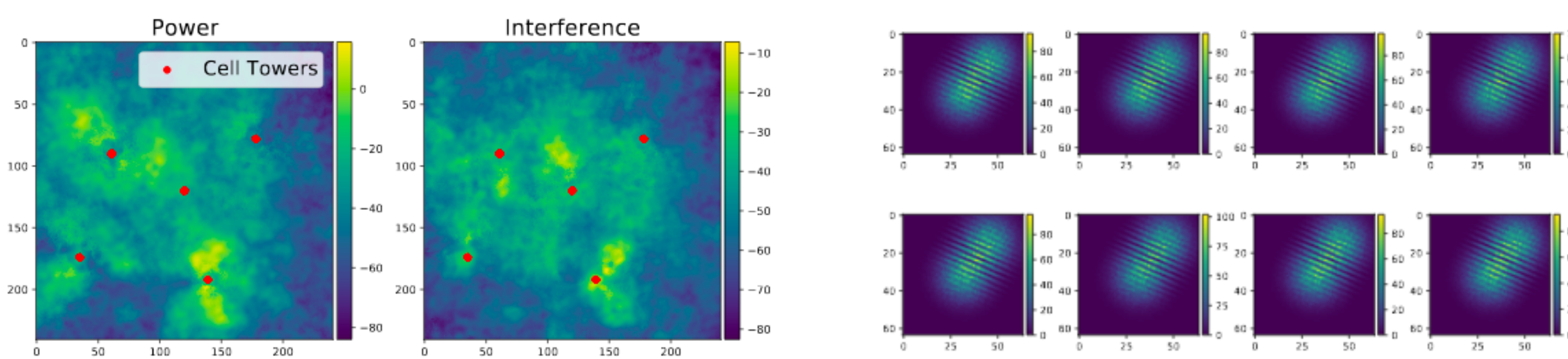
Final objective value

LARGE SCALE COMPOSITE BAYESIAN OPTIMIZATION

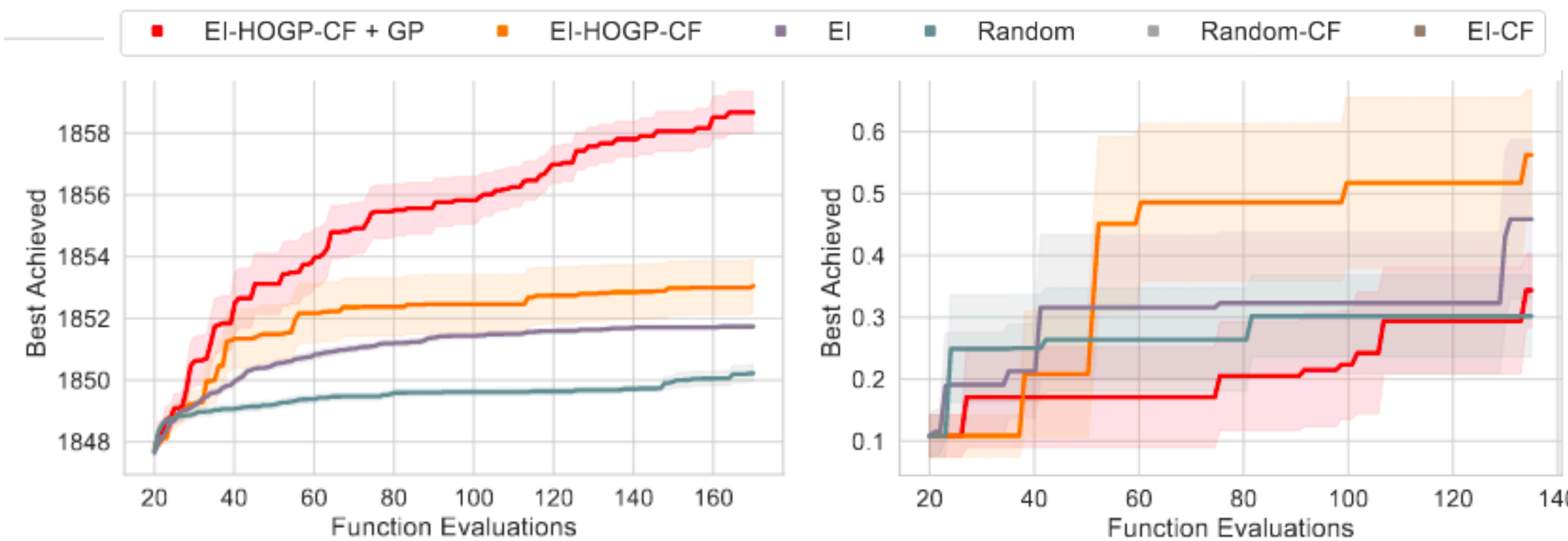


(c) Cell-tower coverage ($t = 2 \times 50 \times 50, d = 30$).

(d) Optics ($t = 16 \times 64 \times 64, d = 4$).

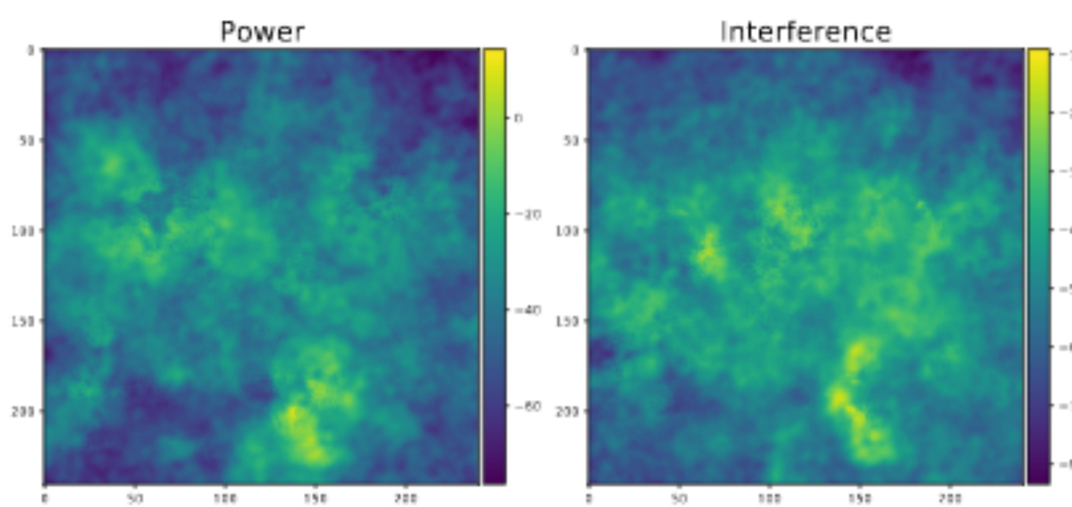


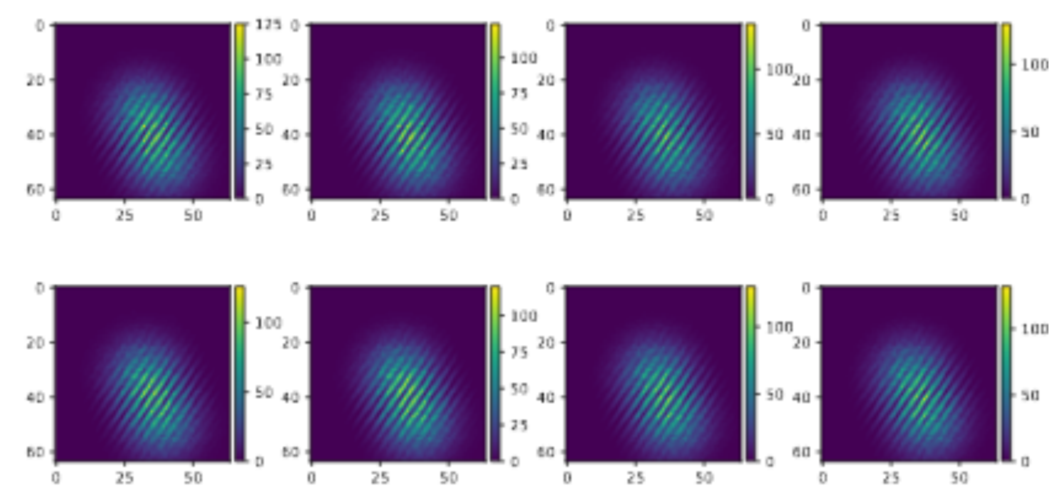
LARGE SCALE COMPOSITE BAYESIAN OPTIMIZATION



(c) Cell-tower coverage ($t = 2 \times 50 \times 50, d = 30$).

(d) Optics ($t = 16 \times 64 \times 64, d = 4$).





PAPER AT: HTTPS://ARXIV.ORG/ABS/2106.12997

CODE AT: BOTORCH.ORG

Thanks

Contact: wjm363 at nyu.edu