This paper describes objective technical results and analysis. Any subjective views or opinions that might be expressed in the paper do not necessarily represent the views of the U.S. Department of Energy or the United States Government.

Gaussian process regression and Bayesian optimization for computational materials science applications

Anh Tran

Sandia National Laboratories

April 12–15, 2022
Atlanta, GA
SIAM Conference on Uncertainty Quantification
SIAM UQ 22

boratories is a multimission laboratory managed and operated by National Technology & Engineering Solutions of Sandia, LL
Honeywell International Inc., for the U.S. Department of Energy's National Nuclear Security Administration under contract DE

Acknowledgment

Joint work with

- SNL: Tim Wildey, Mike Eldred, Bart G van Bloemen Waanders, Kathryn Maupin, John Mitchell, Laura Swiler, Julien Tranchida, Aidan Thompson, Theron Rodgers, Hojun Lim, David Montes de Oca Zapiain
- ORNL: Hoang Tran
- Georgia Tech: Yan Wang, Stefano Travaglino, Wei Sun
- Others: Scott McCann (Xilinx), John Furlan (GIW), Krishnan Pagalthivarthi (GIW), Robert Visintainer (GIW)

Funded by

- NSF
- DOE/Office of Science/ASCR
- Sandia ASC and LDRD Program

The views expressed in the article do not necessarily represent the views of the U.S. Department of Energy or the United States Government. Sandia National Laboratories is a multimission laboratory managed and operated by National Technology and Engineering Solutions of Sandia, LLC., a wholly owned subsidiary of Honeywell International, Inc., for the U.S. Department of Energy's National Nuclear Security Administration under contract DE-NA-0003525.

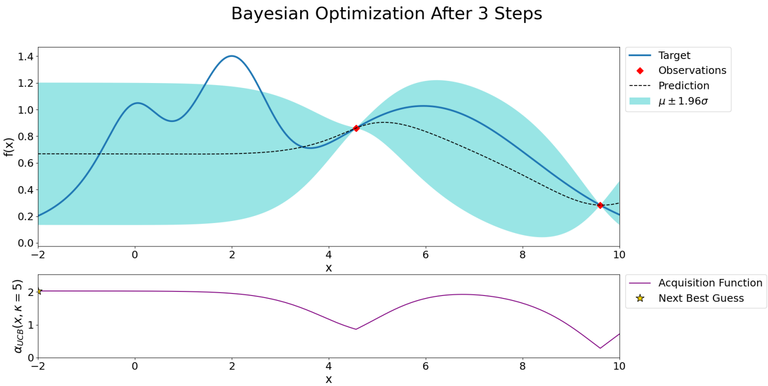
Gaussian process / Bayesian optimization

- Demo
- Introduction
- Fundamentals
- Acquisition
- Constraints
- Parallel
- Multi-fidelity
- Multi-objective
- Mixed-integer
- Big Data
- High-dimensional
- Analysis
- Connection to deep learning

ICME applications

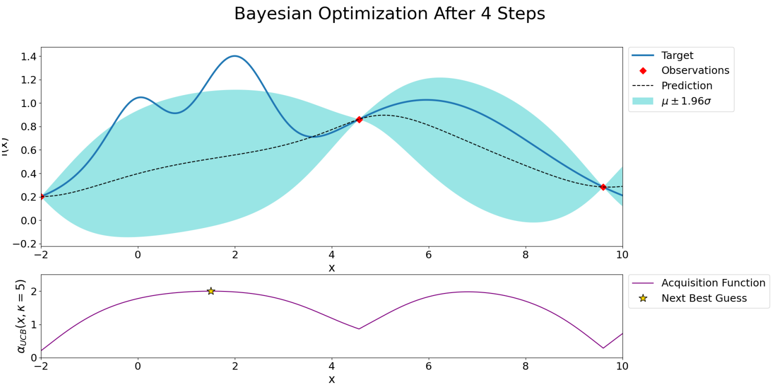
Conclusion

Bayesian optimization (animation)



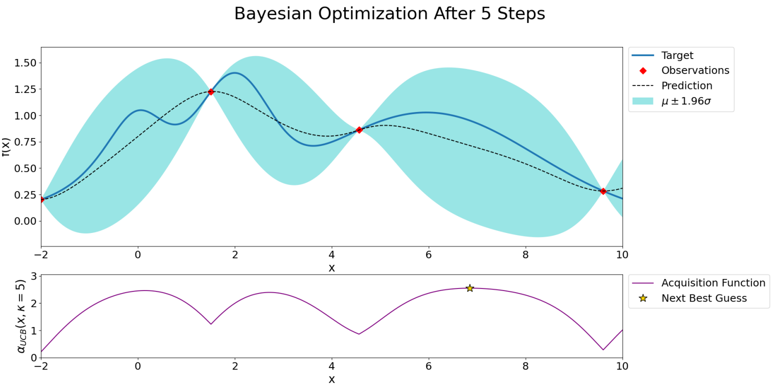
Bayesian optimization - Iteration 3

Bayesian optimization (animation)



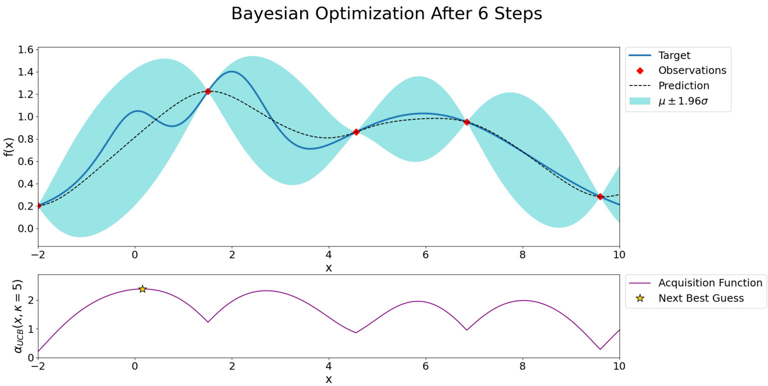
Bayesian optimization - Iteration 4

Bayesian optimization (animation)



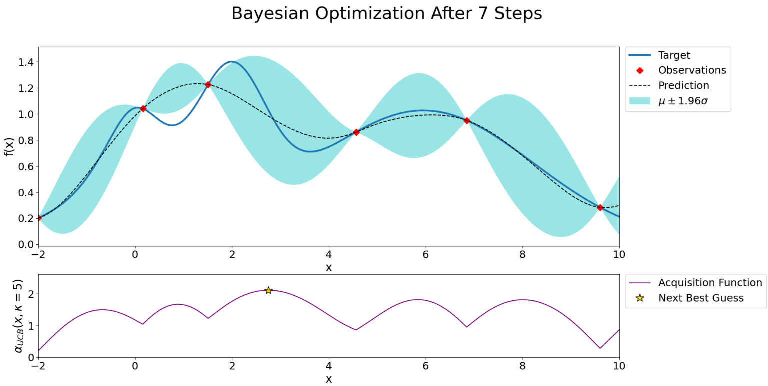
Bayesian optimization - Iteration 5

Bayesian optimization (animation)



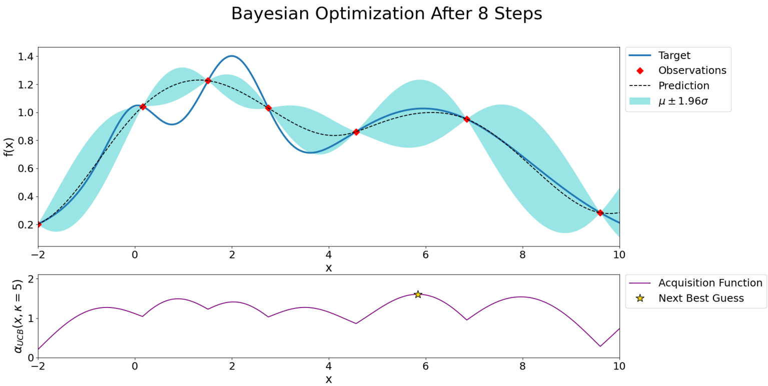
Bayesian optimization - Iteration 6

Bayesian optimization (animation)



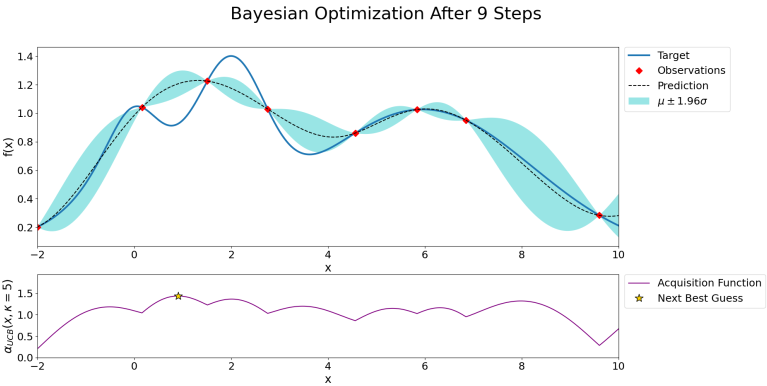
Bayesian optimization - Iteration 7

Bayesian optimization (animation)



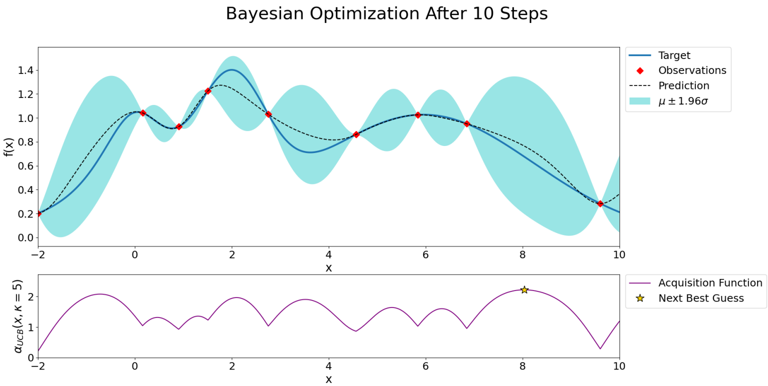
Bayesian optimization - Iteration 8

Bayesian optimization (animation)



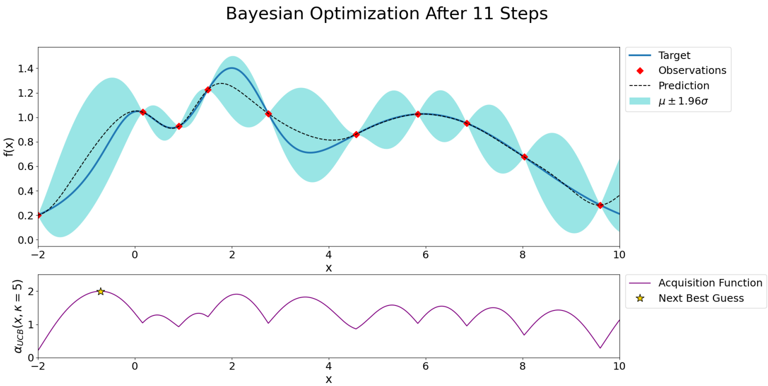
Bayesian optimization - Iteration 9

Bayesian optimization (animation)



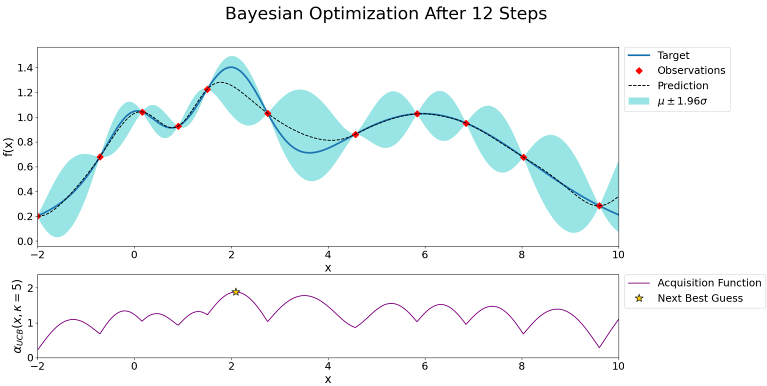
Bayesian optimization - Iteration 11

Bayesian optimization (animation)



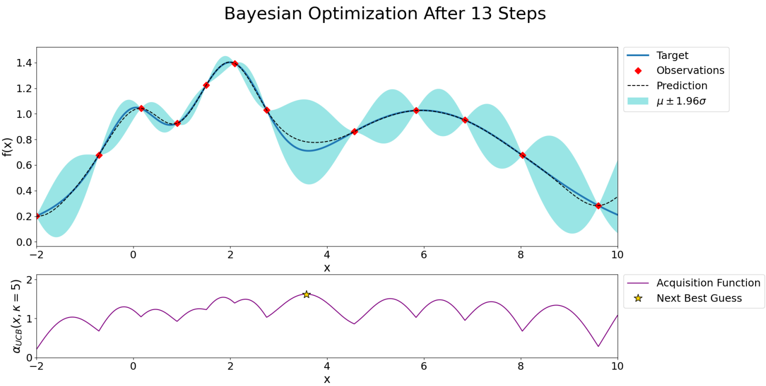
Bayesian optimization - Iteration 11

Bayesian optimization (animation)



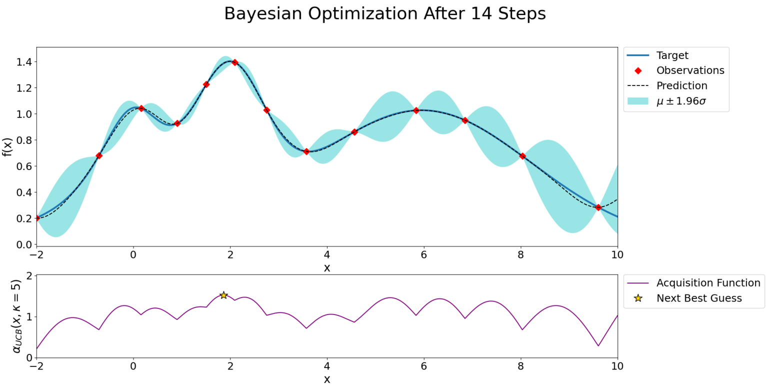
Bayesian optimization - Iteration 12

Bayesian optimization (animation)



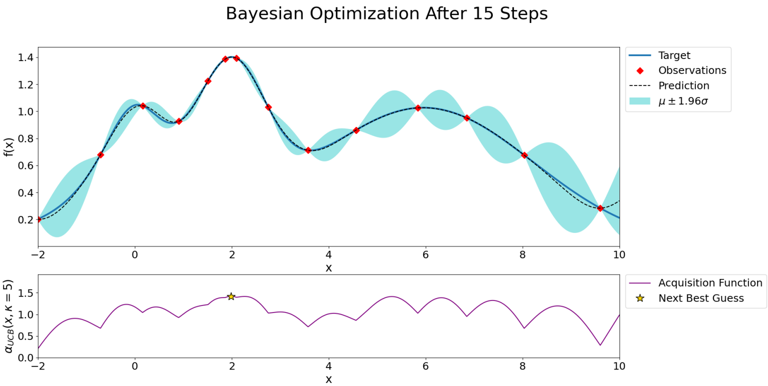
Bayesian optimization - Iteration 13

Bayesian optimization (animation)



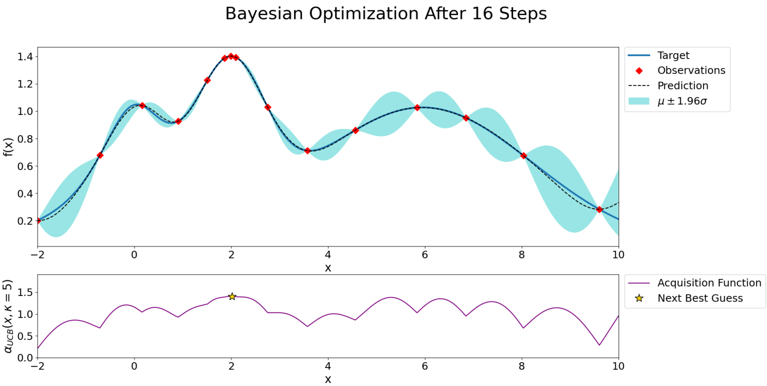
Bayesian optimization - Iteration 14

Bayesian optimization (animation)



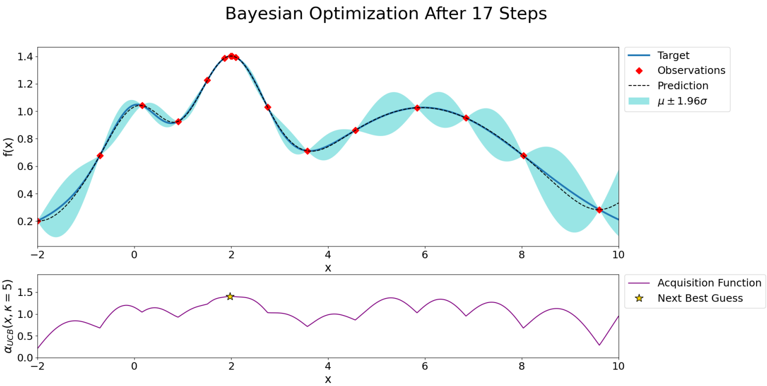
Bayesian optimization - Iteration 15

Bayesian optimization (animation)



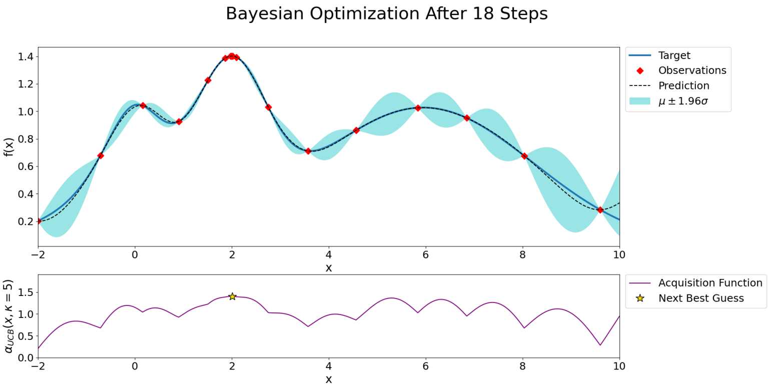
Bayesian optimization - Iteration 16

Bayesian optimization (animation)



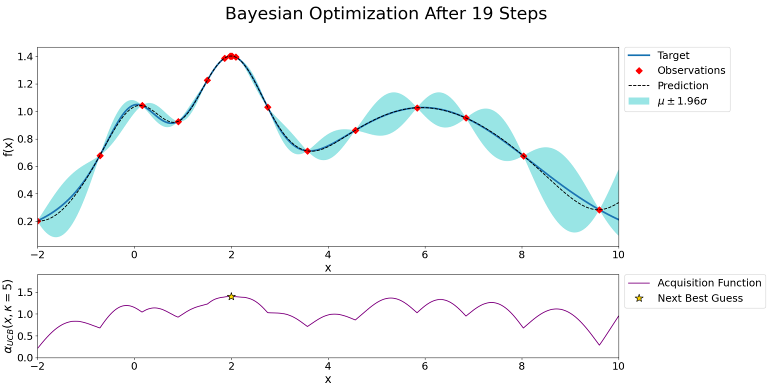
Bayesian optimization - Iteration 17

Bayesian optimization (animation)



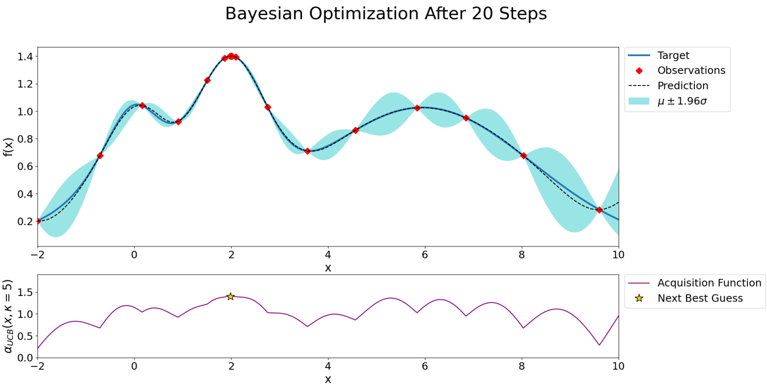
Bayesian optimization - Iteration 18

Bayesian optimization (animation)



Bayesian optimization - Iteration 19

Bayesian optimization (animation)



Bayesian optimization - Iteration 20

Advantages/Disadvantages

Bayesian optimization in a nutshell

Bayesian optimization = Gaussian process + sampling strategy

Advantages:

- optimize with uncertainty consideration (e.g. noisy observations)
- active machine learning (balance exploration-exploitation)
- derivative free (avoid computing Jacobian)
- global optimization (convergence in probability to global optimum)
- good convergence rate (provable asymptotic regret)

Disadvantages:

- high-dimensionality
- scalability: computational bottleneck $\mathcal{O}(n^3)$

Bayesian optimization features

very versatile (open for methodological extensions)

- acquisition functions: PI, EI, UCB, Thompson sampling, entropy-based, KG, or combination among these
- constrained on objectives (known + unknown constraints)
- multi-objective (Pareto frontier/optimal, domination)
- multi-output
- multi-fidelity
- batch parallelization → asynchronous parallel
- stochastic, heteroscedastic
- time-series (forecasting, e.g. causal kernel)
- mixed-integer, e.g. discrete/categorical
- scalable to Big Data
- latent variable model
- gradient-enhanced
- high-dimensional (with low effective dimensionality or separable)
- physics-constrained: monotonic, discontinuous, symmetric, bounded
- outlier: student-t distribution
- non-stationary kernels

Classical GP: Fundamentals

Let $\mathcal{D}_n = \{\mathbf{x}_i, y_i\}_{i=1}^{n}$ denote the set of observations and $\mathbf{x}$ denote an arbitrary test points

$$\mu_n(\mathbf{x}) = \mu_0(\mathbf{x}) + \mathbf{k}(\mathbf{x})^T (\mathbf{K} + \sigma^2 \mathbf{I})^{-1} (\mathbf{y} - \mathbf{m}) \tag{1}$$

$$\sigma_n^2(\mathbf{x}) = k(\mathbf{x}, \mathbf{x}) - \mathbf{k}(\mathbf{x})^T (\mathbf{K} + \sigma^2 \mathbf{I})^{-1} \mathbf{k}(\mathbf{x}) \tag{2}$$

where $\mathbf{k}(\mathbf{x})$ is a vector of covariance terms between $\mathbf{x}$ and $\mathbf{x}_{1:n}$.

Classical GP: Fundamentals

- assuming stationary kernel → $k(\mathbf{x}, \mathbf{x}')$ only depends on $r = ||\mathbf{x} - \mathbf{x}'||$
- the covariance matrix: symmetric positive-semidefinite matrix made up of pairwise inner products

$$\mathbf{K}_{ij} = k(\mathbf{x}_i, \mathbf{x}_j) = k(\mathbf{x}_j, \mathbf{x}_i) = \mathbf{K}_{ji} \tag{3}$$

- kernel choice: assuming unknown function is smooth to some degree

Matérn kernels:

$$\mathbf{K}_{i,j} = k(\mathbf{x}_i, \mathbf{x}_j) = \theta_0^2 \frac{2^{1-\nu}}{\Gamma(\nu)} (\sqrt{2\nu} r)^{\nu} K_{\nu}(\sqrt{2\nu} r), \tag{4}$$

K_ν is a modified Bessel fuction of the second kind and order ν.
Common kernels:

- $\nu = 1/2 : k_{\text{Matérn}1}(\mathbf{x}, \mathbf{x}') = \theta_0^2 \exp{(-r)}$
- $\nu = 3/2 : k_{\text{Matérn}3}(\mathbf{x}, \mathbf{x}') = \theta_0^2 \exp{(-\sqrt{3}r)}(1 + \sqrt{3}r)$,
- $\nu = 5/2 : k_{\text{Matérn}5}(\mathbf{x}, \mathbf{x}') = \theta_0^2 \exp{(-\sqrt{5}r)} \left(1 + \sqrt{5}r + \frac{5}{3}r^2\right)$,
- $\nu \to \infty : k_{\text{sq-exp}}(\mathbf{x}, \mathbf{x}') = \theta_0^2 \exp\left(-\frac{r^2}{2}\right)$

Log (marginal) likelihood function:

$$\log p(\mathbf{y}|\mathbf{x}_{1:n}, \theta) = - \underbrace{\frac{n}{2}\log{(2\pi)}}_{\text{data likelihood} \downarrow \text{ as } n\uparrow} - \underbrace{\frac{1}{2}\log|\mathbf{K}^{\theta} + \sigma^2\mathbf{I}|}_{\substack{\text{“complexity” term} \\ \text{smoother covariance matrix}}} - \underbrace{\frac{1}{2}(\mathbf{y} - \mathbf{m}_\theta)^T(\mathbf{K}^\theta + \sigma^2\mathbf{I})^{-1}(\mathbf{y} - \mathbf{m}_\theta)}_{\substack{\text{“data-fit” term} \\ \text{how well model fits data}}} \tag{5}$$

Classical GP: A Bayesian perspective

Mostly follow Quiñonero-Candela and Hansen 2004; Quiñonero-Candela and Rasmussen 2005.

Denote training $\mathbf{f}$, testing $\mathbf{f}_*$, the joint GP prior is

$$p(\mathbf{f}, \mathbf{f}_*) = \mathcal{N}\left(\begin{bmatrix} \mathbf{m} \\ \mathbf{m} \end{bmatrix}, \begin{bmatrix} \mathbf{K}_{\mathbf{f},\mathbf{f}} & \mathbf{K}_{*,\mathbf{f}} \\ \mathbf{K}_{\mathbf{f},*} & \mathbf{K}_{*,*} \end{bmatrix} \right). \tag{6}$$

By Bayes' rule

$$\begin{aligned} p(\mathbf{f}_*|\mathbf{y}) &= \int p(\mathbf{f}, \mathbf{f}_*|\mathbf{y}) d\mathbf{f} \\ &= \frac{1}{p(\mathbf{y})} \int p(\mathbf{y}|\mathbf{f})\ p(\mathbf{f}, \mathbf{f}_*) d\mathbf{f} \\ &= \mathcal{N}(\mathbf{m} + \mathbf{K}_{*,\mathbf{f}}[\mathbf{K}_{\mathbf{f},\mathbf{f}} + \sigma^2 \mathbf{I}]^{-1}(\mathbf{y} - \mathbf{m}), \mathbf{K}_{*,*} - \mathbf{K}_{*,\mathbf{f}}[\mathbf{K}_{\mathbf{f},\mathbf{f}} + \sigma^2 \mathbf{I}]^{-1} \mathbf{K}_{\mathbf{f},*}), \end{aligned} \tag{7}$$

Log (marginal) likelihood function:

$$\begin{aligned} \log p(\mathbf{y}|\mathbf{X}) &= \log \int p(\mathbf{y}|\mathbf{f}) p(\mathbf{f}|\mathbf{X}) d\mathbf{f} \\ &= -\frac{n}{2} \log(2\pi) - \frac{1}{2} \log |\mathbf{K}_{\mathbf{f},\mathbf{f}} + \sigma^2 \mathbf{I}| \\ &\quad -\frac{1}{2} (\mathbf{y} - \mathbf{m})^\top (\mathbf{K}_{\mathbf{f},\mathbf{f}} + \sigma^2 \mathbf{I})^{-1} (\mathbf{y} - \mathbf{m}). \end{aligned} \tag{8}$$

Classical GP: A Bayesian perspective

A conditional of a Gaussian is also Gaussian.

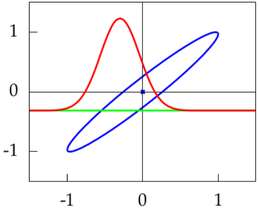


Photo courtesy of from Lawrence 2016.

If

$$P(\mathbf{x},\mathbf{y}) = \mathcal{N}\left(\begin{bmatrix}\mu_{\mathbf{x}}\\ \mu_{\mathbf{y}}\end{bmatrix}, \begin{bmatrix}A & C\\ C^\top & B\end{bmatrix}\right) \tag{9}$$

then

$$P(\mathbf{x}|\mathbf{y}) = \mathcal{N}(\mu_{\mathbf{x}} + CB^{-1}(y-\mu_{\mathbf{y}}), A - CB^{-1}C^\top) \tag{10}$$

(cf. App. A, Quiñonero-Candela and Rasmussen 2005).

Acquisition function: How to pick the next point(s)

- how to pick the next point: exploitation (if $\sigma_A^2 = \sigma_B^2$ but $\mu_A > \mu_B$ then choose A) or exploration (if $\mu_A = \mu_B$ but $\sigma_A^2 > \sigma_B^2$ then choose A). If
- different flavors:
  1. probability of improvement (PI) Mockus 1982
     $$\alpha_{\mathsf{PI}}(\mathbf{x}) = \Phi(\gamma(\mathbf{x})), \tag{11}$$
     where
     $$\gamma(\mathbf{x}) = \frac{\mu(\mathbf{x}) - f(\mathbf{x}_{\mathsf{best}})}{\sigma(\mathbf{x})}, \tag{12}$$
  2. expected improvement (EI) scheme Huang et al. 2006; Mockus 1975
     $$\alpha_{\mathsf{EI}}(\mathbf{x}) = \sigma(\mathbf{x})[\gamma(\mathbf{x})\Phi(\gamma(\mathbf{x})) + \phi(\gamma(\mathbf{x})] \tag{13}$$
  3. upper confidence bound (UCB) schemeSrinivas et al. 2009, 2012
     $$\alpha_{\mathsf{UCB}}(\mathbf{x}) = \mu(\mathbf{x}) + \kappa\sigma(\mathbf{x}), \tag{14}$$
     where κ is a hyper-parameter describing the exploitation-exploration balance.
  4. pure exploration*:
     - maximal MSE $\sigma^2(\mathbf{x}) \Leftrightarrow$ maximal entropy: $\frac{1}{2}\log\left[2\pi\sigma^2(\mathbf{x})\right] + \frac{1}{2}$
     - maximal IMSE $\int_{\mathbf{x}\in\mathcal{X}} \sigma^2(\mathbf{x})$

QRAK taxonomy for constrained optimization problem

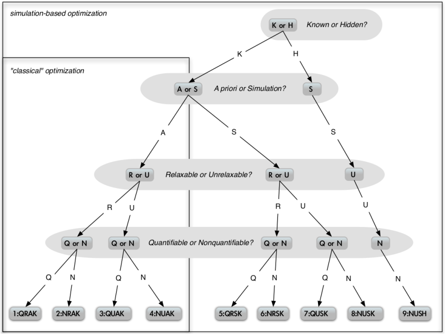


Photo courtesy of Digabel and Wild 2015. Tree-based view of the QRAK taxonomy of constraints.

Constraints that are either not known beforehand or have to assessed through simulations are called unknown.

Constrained problems: known constraints

Problem statement

optimize $f(\mathbf{x})$ subject to $\lambda(\mathbf{x}) \leq \mathbf{c}$, $\lambda(\cdot)$ computationally cheap

known constraints:

- known before evaluation
- typically physics-based, e.g. total composition $\geq 100\%$
- formulated as inequality constraints $\lambda(\mathbf{x}) \leq \mathbf{c}$, λ is computationally cheap
- directly penalize the acquisition function $\alpha = 0$ when constraints are violated, i.e. $\lambda(\mathbf{x}) \not\leq \mathbf{c}$

  $$\alpha^{\text{known}}_{\text{constrained}}(\mathbf{x}) = \alpha(\mathbf{x}) I_{\text{known}}(\mathbf{x}) \tag{15}$$

  where $I_{\text{known}}(\mathbf{x})$ is the indicator function

  $$I_{\text{known}}(\mathbf{x}) = \begin{cases} 1, & \lambda(\mathbf{x}) \leq \mathbf{c} \\ 0, & \lambda(\mathbf{x}) \not\leq \mathbf{c} \end{cases} \tag{16}$$

- can be conveniently ignored to become unknown constraints if the model is aware of the constraints violation, i.e. returns error

Constrained problems: unknown constraints

Problem statement

optimize $f(\mathbf{x})$ where $f(\mathbf{x})$ may or may not exist

unknown constraints:

- can convert known → unknown but not vice versa
- form a probabilistic binary classifier to predict the probability mass function of passing unknown constraint at $\mathbf{x}$, i.e. kNN, AdaBoost, RandomForest, GP, etc.
- penalize the acquisition function based on the predicted feasibility from GP classifier

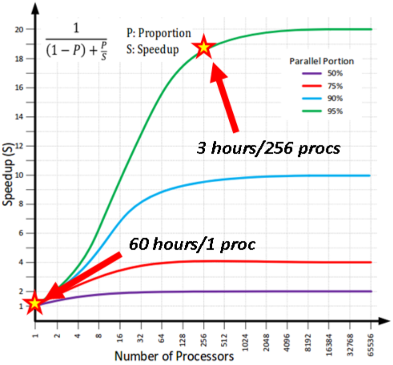
$$\alpha_{\text{constrained}}^{\text{unknown}}(\mathbf{x}) = \begin{cases} \alpha(\mathbf{x}), & \text{with } \Pr(\text{clf}(\mathbf{x}) = 1) \\ 0, & \text{with } \Pr(\text{clf}(\mathbf{x}) = 0) \end{cases} \tag{17}$$

- our approach:
  - use another GP to learn when $f(\mathbf{x})$ does not exist
  - optimize the conditioned acquisition function $\mathbb{E}[\alpha_{\text{constrained}}^{\text{unknown}}(\mathbf{x})] = \alpha(\mathbf{x})\Pr_{\text{unknown}}(\text{clf}(\mathbf{x}) = 1)$

Batch parallel on HPC

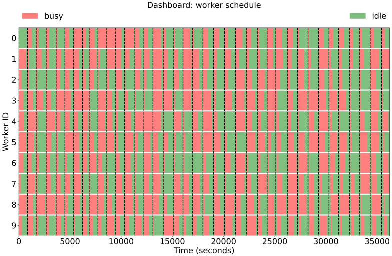
Might as well be beneficial when computing resource is insufficient; examples:

- $P = 0.95 \rightarrow$ SpeedUp $\approx$ 20 times
- CFD simulation takes 3 hours to finish with 256 procs $\rightarrow$ 20 cases/60 hours
- or 60 hours (2.5 days) with 1 proc for 1 case $\rightarrow$ 256 cases/60 hours
- fixed computational budget: 256 $\times$60 CPU hours
- observation: parallelizing optimization can provide more observations than parallelizing the code

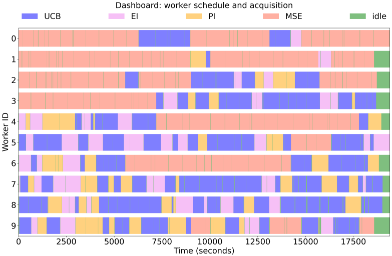


Amdahl's law for parallelization.

Asynchronous parallel on HPC



Batch-sequential parallel



Multi-α asynchronous parallel

Multi-fidelity

Kennedy and O'Hagan Kennedy and O'Hagan 2000: auto-regressive model based on a first-order auto-regressive relation between model output of different levels of fidelity.

- s-levels of variable-fidelity model $y_t(\mathbf{x})_{t=1}^{s}$
- $y_1(\mathbf{x})$: cheapest, $y_s(\mathbf{x})$: most expensive
- auto-regressive model:

$$y_t(\mathbf{x}) = \rho_{t-1} y_{t-1}(\mathbf{x}) + \delta_t(\mathbf{x}) \tag{18}$$

- Markov property: assuming that given $y_{t-1}(\mathbf{x})$, we can learn nothing about $y_t(\mathbf{x})$ from any other model output $y_{t-1}(\mathbf{x}')$, for $\mathbf{x} \neq \mathbf{x}'$

$$\mathrm{Cov}[y_t(\mathbf{x}), y_{t-1}(\mathbf{x}') | y_{t-1}(\mathbf{x})] = 0, \quad \forall \mathbf{x} \neq \mathbf{x}' \tag{19}$$

Multi-fidelity

- model the lowest fidelity y_1 as a classical GP
- model the discrepancies δ_t's as GPs
- for two levels of fidelity: $\blacksquare_c$ = cheap, $\blacksquare_e$ = expensive
- covariance vector and covariance matrix

$$k(\mathbf{x}) = \begin{pmatrix} \rho k_c(\mathbf{x}) & k_e(\mathbf{x}) \end{pmatrix}, \tag{20}$$

$$\mathbf{K} = \begin{pmatrix} \sigma_c^2 \mathbf{K}_c & \rho \sigma_c^2 \mathbf{K}_c(\mathbf{x}_c, \mathbf{x}_s) \\ \rho \sigma_c^2 \mathbf{K}_c(\mathbf{x}_s, \mathbf{x}_c) & \rho^2 \sigma_c^2 \mathbf{K}_c(\mathbf{x}_s, \mathbf{x}_s) + \sigma_d^2 \mathbf{K}_d(\mathbf{x}_s, \mathbf{x}_s) \end{pmatrix} \tag{21}$$

Selection of level of fidelity

Question: Fix a sampling location $\mathbf{x}^*$, what level of fidelity should be selected to query?

Compare computational cost vs. benefit:

- $1 \leq t \leq s$: level of fidelity
- if $\mathbf{x}^*$ is queried, how much uncertainty is reduced?
- at what cost?
- our approach: balance computational cost vs. gain (reduction of uncertainty)

$$t^* = \operatorname*{argmin}_t \left(C_t \int_{\mathcal{X}} \sigma^2(\mathbf{x}) d\mathbf{x} \right), \tag{22}$$

- promote high-fidelity if the cost is similar: If $C_{t^*}|\mathcal{D}^{(t^*)}| \geq C_s|\mathcal{D}^{(s)}|$ then choose s.

Multi-objective

Let:

- $\mathbf{x} = \{x_i\}_{i=1}^{d} \in \mathcal{X} \subseteq \mathbb{R}^d$ be input in d-dimensional space,
- $\mathbf{y} = \{y_j\}_{j=1}^{s}$ as s outputs.

$$\underset{\mathbf{x}\in\mathcal{X}}{\operatorname{argmax}}(f_1(\mathbf{x}), \cdots, f_s(\mathbf{x})) \tag{23}$$

subjected to $\mathbf{c}(\mathbf{x}) \leq \mathbf{0}$.
Pareto definition: $\mathbf{x}_1$ is said to dominate $\mathbf{x}_2$, denoted as $\mathbf{x}_1 \preceq \mathbf{x}_2$, if and only if $\forall 1 \leq j \leq s$, such that $y_j(\mathbf{x}_1) \leq y_j(\mathbf{x}_2)$, and $\exists 1 \leq j \leq s$, such that $y_j(\mathbf{x}_1) < y_j(\mathbf{x}_2)$.
Scalarization: multi-objective → single-objective

1. weighted Tchebycheff with ℓ_∞: $y = \max_{1\leq i\leq s} w_i(y_i(\mathbf{x}) - z_i^*)$,
2. weighted sum with ℓ_1: $y = \sum_{i=1}^{s} w_i y_i(\mathbf{x})$,
3. augmented Tchebycheff with $\ell_1 + \ell_\infty$:
   $y = \max_{1\leq i\leq s} w_i(y_i(\mathbf{x}) - z_i^*) + \rho \sum_{i=1}^{s} w_i y_i(\mathbf{x})$,

z_i^* denotes the inferred ideal i-th objective value; normalized weights: $0 \leq w_i \leq 1$, $\sum_{i=1}^{m} w_i = 1$; $0 < \rho < 1$ positive constant.

Multi-objective

Hypervolume approach:

- hypervolume indicator, aka $\mathcal{S}$-metric
- strictly monotonic
- complexity $\mathcal{O}(n \log n + n^{d/2} \log n)$
- $d = 3$: lower and upper bounds $\mathcal{O}(n \log n)$ Beume et al. 2009
- and any other sorts of approximation ...

A near-optimal approach:

- Low-dimensional output $(d \leq 3)$: hypervolume estimator
- High-dimensional output $(d > 3)$: Tchebycheff decomposition

Mixed-integer

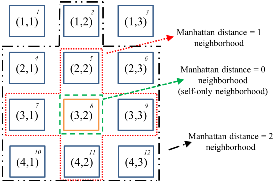
Main idea:

- decompose to a set of continuous and discrete variables $\mathbf{x} = (\mathbf{x}_d, \mathbf{x}_c)$
- enumerate $\mathbf{x}_d$ and build a local GP for each $\mathbf{x}_d$
- form a Gaussian mixture prediction with adaptively weighted average
- applicable when $|\mathbf{x}_c| \gg |\mathbf{x}_d|$, i.e. not combinatorial optimization problems
- Gaussian mixture model predictions for posterior mean and variance:

$$\hat{\mu} = \sum_{\ell^* \in \mathcal{B}(\ell)} w_{\ell^*} \left(\hat{\mu}^{(\ell^*)} + \underbrace{\bar{\mu}^{(\ell)} - \bar{\mu}^{(\ell^*)}}_{\substack{\text{bias correction term} \\ \mathbb{E}[\hat{\mu}]=\bar{\mu}^{(\ell)}}} \right) \quad (24)$$

$$\hat{\sigma}^2 = \sum_{\ell^* \in \mathcal{B}(\ell)} w_{\ell^*}^2 \sigma_{(\ell^*)}^2 \quad (25)$$

- weighted average estimation, weights depends on (1) cluster distances, (2) original cluster predictions
- theoretical bounds for weighted average prediction
- asymptotic behavior when $n \to \infty$


(1,1) 1
(1,2) 2
(1,3) 3
(2,1) 4
(2,2) 5
(2,3) 6
(3,1) 7
(3,2) 8
(3,3) 9
(4,1) 10
(4,2) 11
(4,3) 12
Manhattan distance = 1 neighborhood
Manhattan distance = 0 neighborhood (self-only neighborhood)
Manhattan distance = 2 neighborhood

Neighborhood $\mathcal{B}(\ell)$ of a local GP ℓ with $\mathbf{x}_d = (3, 2)$, defined by thresholding a similarity measure of discrete tuples

Mixed-integer

A special case: Wasserstein distance (Earth mover's distance). Assume the query point $\mathbf{x} = (\mathbf{x}_d, \mathbf{x}_c)$, where $\mathbf{x}_d$ corresponds to ℓ-th cluster.

$$w_{\ell^*} \propto \left[\sigma_l^2 + W_2\left(\mathcal{N}(y^{(\ell^*)}, \sigma^2_{(\ell^*)}), \mathcal{N}(y^{(\ell)}, \sigma^2_{(\ell)})\right)\right]^{-1}. \tag{26}$$

$$W_2\left(\mathcal{N}(y^{(\ell^*)}, \sigma^2_{(\ell^*)}), \mathcal{N}(y^{(\ell)}, \sigma^2_{(\ell)})\right) = \left\| y^{(\ell)} - y^{(\ell^*)} \right\|^2 + \left\| \sqrt{\sigma^2_{(\ell)}} - \sqrt{\sigma^2_{(\ell^*)}} \right\|^2 \tag{27}$$

Weighted linear average prediction

The largest weight is associated with the ℓ-th cluster.

Asymptotic analysis $n \to \infty$

$\lim_{n\to\infty} w_l \to \infty$, as $\sigma_l \to 0$ and $W_2(\cdot_l, \cdot_l) = 0$.

Interpretation: If data is abundant, then the proposed approach converge asymptotically to a single local GP prediction.

Sparse variational GP

Low-rank approximation for $\mathbf{K}_{\mathbf{f},\mathbf{f}}$

Low-rank approximation $\mathbf{K} \approx \widetilde{\mathbf{K}} = \mathbf{K}_{n \times m} \mathbf{K}_{m \times m}^{-1} \mathbf{K}_{m \times n}$ (cf. Section 8.1 Rasmussen 2006) and scales as $\mathcal{O}(nm^2 + m^3)$ instead of $\mathcal{O}(n^3)$.
For $n \gg m$, this method scales as $\mathcal{O}(nm^2)$.

Following Quiñonero-Candela and Rasmussen 2005; Quiñonero-Candela, Rasmussen, and Williams 2007, Chalupka, Williams, and Murray 2013, Vanhatalo et al. 2012, 2013.
Cost complexity:

- local GP: $\mathcal{O}(m^3)$
- sparse GP: $\mathcal{O}(nm^2)$
- classical GP (Cholesky decomposition): $\mathcal{O}\left(\frac{1}{3} n^3\right)$
- classical GP (LU decomposition): $\mathcal{O}\left(\frac{2}{3} n^3\right)$
- classical GP (QR decomposition): $\mathcal{O}\left(\frac{4}{3} n^3\right)$

Sparse variational GP

Follows Frigola, Chen, and Rasmussen 2014 and Rasmussen's corresponding slides. By Bayes' rule,

$$p(\mathbf{f}|\mathbf{y},\theta) = \frac{p(\mathbf{y}|\mathbf{f})p(\mathbf{f}|\theta)}{p(\mathbf{y}|\theta)} \Leftrightarrow p(\mathbf{y}|\theta) = \frac{p(\mathbf{y}|\mathbf{f})p(\mathbf{f}|\theta)}{p(\mathbf{f}|\mathbf{y},\theta)}. \tag{28}$$

The idea: approximate the (computationally intractable) $p(\mathbf{f}|\mathbf{y},\theta)$ by a (computationally tractable) parameterized variational $q(\mathbf{f})$. For any $q(\mathbf{f})$,

$$p(\mathbf{y}|\theta) = \frac{p(\mathbf{y}|\mathbf{f})p(\mathbf{f}|\theta)}{p(\mathbf{f}|\mathbf{y},\theta)} \frac{q(\mathbf{f})}{q(\mathbf{f})} \Leftrightarrow \log p(\mathbf{y}|\theta) = \log \frac{p(\mathbf{y}|\mathbf{f})p(\mathbf{f}|\theta)}{q(\mathbf{f})} + \log \frac{q(\mathbf{f})}{p(\mathbf{f}|\mathbf{y},\theta)}. \tag{29}$$

Apply $\int q(\mathbf{f})d\mathbf{f}$ to both sides

$$\underbrace{\log p(\mathbf{y}|\theta)}_{\text{marginal likelihood}} = \underbrace{\int q(\mathbf{f}) \log \frac{p(\mathbf{y}|\mathbf{f})p(\mathbf{f}|\theta)}{q(\mathbf{f})} d\mathbf{f}}_{\text{Evidence Lower BOund}} + \underbrace{\int q(\mathbf{f}) \log \frac{q(\mathbf{f})}{p(\mathbf{f}|\mathbf{y},\theta)} d\mathbf{f}}_{KL(q(\mathbf{f})||p(\mathbf{f}|\mathbf{y},\theta))} \tag{30}$$

Turn our attention to maximizing the variational ELBO (or equivalently minimizing the KL divergence) instead of maximizing the log marginal likelihood.

High-dimensional: Active subspace method

Formulations are derived by Constantine 2015; Constantine, Dow, and Wang 2014

Ideas:

- approximate high-dimensional function $f : \mathcal{X} \subset \mathbb{R}^D \to \mathbb{R}$
- perform SVD on covariance of gradient vector with descending eigenvalues

$$\mathbb{E}[\nabla f(\mathbf{x}) \nabla f(\mathbf{x})^\top] = \mathbf{W}\mathrm{Diag}[\lambda_1, \dots, \lambda_D]\mathbf{W}^\top \tag{31}$$

$$\mathrm{Diag}[\lambda_1, \dots, \lambda_D] = \mathrm{Diag}[\lambda_1, \dots, \lambda_d] \bigoplus \mathrm{Diag}[\lambda_{d+1}, \dots, \lambda_D], \quad \mathbf{W} = [\mathbf{W}_1 \ \mathbf{W}_2] \tag{32}$$

- rotate the inputs $\mathbf{W}_1 \in \mathbb{R}^{D \times d}, \mathbf{W}_2 \in \mathbb{R}^{D \times (D-d)}$

$$f(\mathbf{x}) = f(\mathbf{W}\mathbf{W}^\top \mathbf{x}) = f(\mathbf{W}_1 \mathbf{W}_1^\top \mathbf{x} + \mathbf{W}_2 \mathbf{W}_2^\top \mathbf{x}) = f(\mathbf{W}_1 \mathbf{y} + \mathbf{W}_2 \mathbf{z}) \tag{33}$$

- if $\mathbf{z}$ invariant in an inactive subspace $\lambda_{d+1} = \cdots = \lambda_D = 0$, then $f(\mathbf{x}) = f(\mathbf{W}_1 \mathbf{y})$: reduce from D to d
- work great if gradients are readily available
- but what if gradients are not available? estimation by GP? constrained manifold optimization for $\mathbf{W}_1^\top$ besides the original optimization?

High-dimensional: Gaussian random projection

Mostly follow Wang et al. 2013, 2016. Main idea:

- choose (wisely) and optimize over $\mathcal{Z} \subset \mathbb{R}^d$
- embed and project onto high-dimensional space as $\mathbf{x} \leftarrow p_{\mathcal{X}}(\mathbf{A}z)$
- $\mathbf{A} \in \mathbb{R}^{D \times d}$: tall-and-skinny random matrix with standard normal element

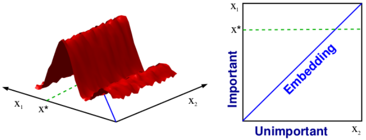
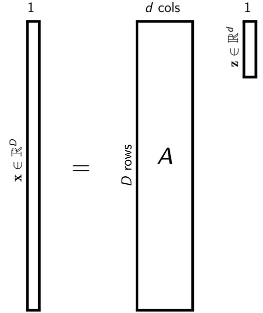


Photo courtesy of Wang et al ibid. Optimizing a 2d function (with 1d active subspace) via random embedding.

REMBO algorithm ibid. with deviation from BO highlighted.

```
generate a random matrix $\mathbf{A} \in \mathbb{R}^{D \times d} : a_{ij} \sim \mathcal{N}(0, 1)$
choose the bounded region set $\mathcal{Z} \subset \mathbb{R}^d$
$\mathcal{D}_0 \leftarrow \varnothing$
for $i = 1, 2, \cdots$ do
    locate next sampling point $\mathbf{z}_{i+1} \leftarrow \mathrm{argmax}_{\mathbf{z} \in \mathcal{Z}}\, a(\mathbf{z}) \in \mathbb{R}^d$
    query $\mathcal{D}_{i+1} \leftarrow \mathcal{D}_i \cup \{z_{i+1}, f(p_{\mathcal{X}}(\mathbf{A}z_{i+1}))\}$
    update GP
end for
```

High-dimensional: Gaussian random projection



A random embedding or a random projection $\mathbf{x} = \mathbf{A}\mathbf{z}$ is built as a corollary from the Johnson-Lindenstrauss lemma, where $\mathbf{A}$ is a random normal matrix.

Lemma (Johnson-Lindenstrauss)

Given n points $\{\mathbf{x}_i\}_{i=1}^n$, each of which is in $\mathbb{R}^D$, $\mathbf{A} \sim \mathcal{MN}_{D\times d}(0, \mathbf{I}, \mathbf{I})$, and let $\mathbf{z} \in \mathbb{R}^d$ defined as $\mathbf{z} = \mathbf{A}^\top \mathbf{x}$. Then, if $d \geq \frac{9 \log n}{\varepsilon^2 - \varepsilon^3}$, for some $\varepsilon \in \left(0, \frac{1}{2}\right)$, then with probability at least $\frac{1}{2}$, all pairwise distances are preserved, i.e. for all i, j, we have

$$(1-\varepsilon)\|\mathbf{x}_i - \mathbf{x}_j\|_2^2 \leq \|\mathbf{z}_i - \mathbf{z}_j\|_2^2 \leq (1+\varepsilon)\|\mathbf{x}_i - \mathbf{x}_j\|_2^2 \tag{34}$$

Compared to active subspace method: also linear and does not require gradient and the rotation matrix $\mathbf{W}^\top$.
There are alternative approaches, e.g. additive GP.

Convergence rate analysis

Regret of action $\mathbf{x}_t$:

$$r_t = |f(\mathbf{x}^*) - f(\mathbf{x}_t)| > 0, \tag{35}$$

where $\mathbf{x}^* = \mathrm{argmax}_{\mathbf{x}\in\mathcal{X}} f(\mathbf{x})$.
Aim to minimize the cumulative regret at the horizon T

$$R_T = \sum_{t<T} r_t. \tag{36}$$

No-regret in infinite horizon: $\lim_{T\to\infty} r_T = \lim_{T\to\infty} \frac{R_T}{T} = 0$
→ motivation for sublinear bounds of R_T, or more precisely, $\mathcal{O}(R_T) \leq \mathcal{O}(T)$.

Convergence rate analysis

$$\mathbf{x}_t = \operatorname*{argmax}_{\mathbf{x}\in\mathcal{D}} \mu_{t-1}(\mathbf{x}) + \beta_t^{1/2}\sigma_{t-1}(\mathbf{x}) \tag{37}$$

For α_{UCB} with Matérn kernel: see Srinivas et al. 2009, 2012; tighter bounds for UCB in noiseless environment, see De Freitas, Smola, and Zoghi 2012.

Theorem ($\mathcal{O}(\sqrt{T})$ Srinivas et al. 2009)

Let $\delta \in (0,1)$*, and* $\beta_t = 2\log\left(\frac{|D|t^2\pi^2}{6\delta}\right)$*, then*

$$Pr\left(R_T \leq \sqrt{C_1 T \beta_T \gamma_T}\right) \geq 1-\delta, \tag{38}$$

where $C_1 = \frac{8}{\log(1+\sigma^{-2})}$.

Convergence rate analysis

Proof.

Sketch of proof Srinivas et al. 2009

1. pick $\delta \in (0,1)$, set $\beta_t = 2\log\left(\frac{|D|\pi_t}{\delta}\right)$, then

$$\Pr\left(|f(\mathbf{x}) - \mu_{t-1}(\mathbf{x})| \leq \beta_t^{1/2}\sigma_{t-1}(\mathbf{x})\right) \geq 1-\delta \tag{39}$$

2. bound r_t of action $\mathbf{x}_t$

$$r_t \leq 2\beta_t^{1/2}\sigma_{t-1}(\mathbf{x}_t) \tag{40}$$

3. associate information gain with posterior variance $\mathbb{I}(\mathbf{y}_T;\mathbf{f}_T) = \frac{1}{2}\sum_{t=1}^{T}\log(1+\sigma^{-2}\sigma_{t-1}^2(\mathbf{x}_t))$
4. $C_1 = \frac{8}{\log(1+\sigma^{-2})}$:

$$\Pr\left(R_T^2/T \leq \sum_{t=1}^{T} r_t^2 \leq \beta_T C_1 \mathbb{I}(\mathbf{y}_T;\mathbf{f}_T) \leq C_1\beta_T\gamma_T\right) \geq 1-\delta. \tag{41}$$

□

Convergence rate analysis

- α_{UCB} with noisy GP: $r_t = \mathcal{O}\left(\frac{1}{\sqrt{t}}\right)$.
- α_{UCB} noiseless setting, see De Freitas, Smola, and Zoghi 2012: $r_t = \mathcal{O}\left(\mathrm{e}^{-\frac{\tau t}{(\ln t)^{d/4}}}\right)$.
- α_{EI}, see Bull 2011:

$$r_t = \begin{cases} \mathcal{O}(t^{-\nu/d}(\log t)^{\alpha}), & \nu \leq 1 \\ \mathcal{O}(t^{-1/d}), & \nu > 1 \end{cases} \tag{42}$$

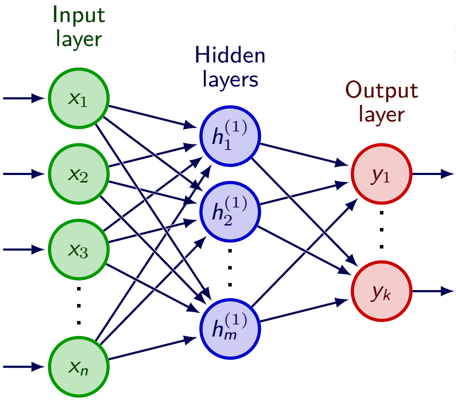
- batch parallel with batch size K α_{BUCB}, see Desautels, Krause, and Burdick 2014:

$$r_t^K = \mathcal{O}\left(C\sqrt{\frac{\log(tK)}{tK}\gamma_{tK}}\right) \tag{43}$$

- batch parallel with batch size K $\alpha_{\text{UCB-PE}}$, see Contal et al. 2013

$$r_t^K = \mathcal{O}\left(\sqrt{\frac{\log(t)}{tK}\gamma_{tK}}\right) \tag{44}$$

Connection to deep learning



Pioneered by Neal 1996: 1 hidden layer with infinite number of nodes, i.e. $m \to \infty$
For every output node y_i, $1 \leq i \leq k$,

$$y_i(x) = b_i^1 + \sum_{j=1}^{m} W_{ij}^1 h_j^{(1)}(x) \quad (45)$$

For every hidden node $h_i^{(1)}$, $1 \leq i \leq m$,

$$h_i^{(1)}(x) = \phi\left(b_i^0 + \sum_{j=1}^{n} W_{ij}^0 x_j\right), \quad (46)$$

Connection to deep learning

Weights and bias are i.i.d $\Rightarrow$ y_i is Gaussian (by Central Limit Theorem)

$$\begin{aligned} \mathbf{K}(x, x') &\equiv \mathbb{E}\left[y_i(x) y_i(x')\right] \\ &= \sigma_b^2 + \sigma_w^2 \mathbb{E}\left[h_i^{(1)}(x), h_i^{(1)}(x')\right] \\ &\equiv \sigma_b^2 + \sigma_w^2 C(x, x') \end{aligned} \tag{47}$$

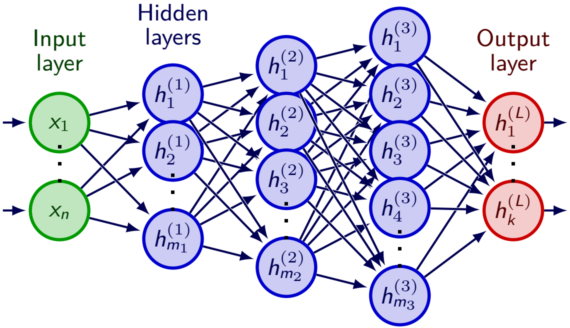
Single-layer, infinite width: y_i, y_j: joint Gaussian, zero covariance, and independent

$$y_i \sim \mathcal{GP}(\mu, \mathbf{K}) \tag{48}$$

For $\phi(x) = max(0, x)$, i.e. ReLU, the equivalent kernel is arccosine (cf. Cho and Saul 2009).

Connection to deep learning

More general results available from Lee et al. 2018 (cf. Appendix C), as $m_L \to \infty, \cdot, m_1 \to \infty$, i.e. multi-layer, infinite width, NN is still equivalent to a GP.



At the last layer L,

$$\lim_{m_L \to \infty, \dots, m_1 \to \infty} p(h^{(L)} | x) = \mathcal{GP}\left(h^{(L)}; 0, (G \circ (F \circ G)) (K^0) \right) \tag{49}$$

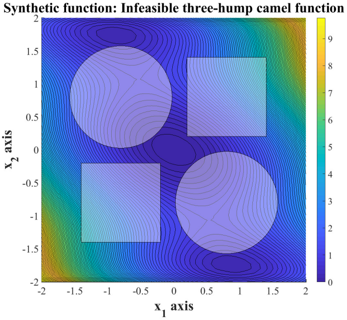
Gaussian process / Bayesian optimization

ICME applications
- Benchmark functions (numerical)
- Flip-chip BGA package design (FEM)
- Heart valve optimization (FEM)
- Pump design optimization (CFD)
- Inverse problems in process-structure (kinetic Monte Carlo)
- Inverse problems in composition-property (DFT + MD)
- Inverse problems in structure-property (CPFEM)

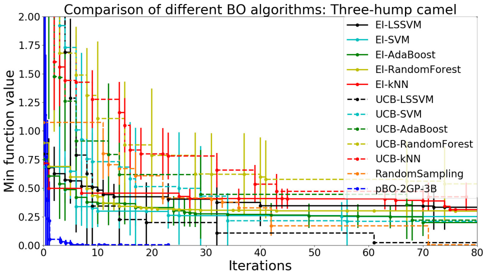
Conclusion

2d three-hump camel

(joint work w/ Yan Wang)



2d three-hump camel.

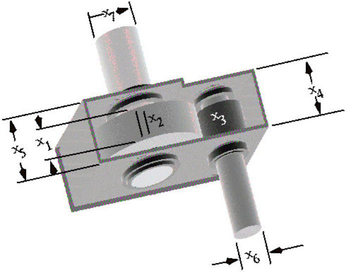


Convergence comparison with different classifiers.

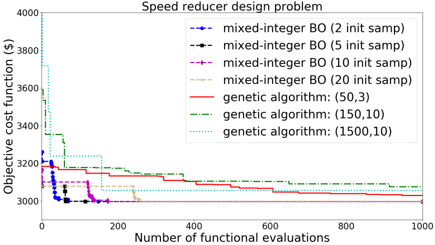
Speed reducer design optimization

(1d+6d) (mixed-integer)

(joint work w/ Yan Wang)



Speed reducer design



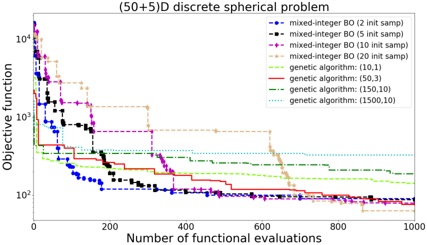
Comparison against GA.

High-dimensional discrete sphere function

(5d+50d) (mixed-integer)

(joint work w/ Yan Wang)

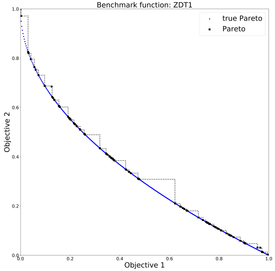
$f(\mathbf{x}^{(d)}, \mathbf{x}^{(c)}) =$
$f(x_1, \cdots, x_n, x_{n+1}, \cdots, x_m) =$
$\prod_{i=1}^{n} |x_i| \left(\sum_{j=n+1}^{m} x_j^2 \right)$ where
$1 \leq x_i \leq 2 (1 \leq i \leq n)$ are n integer variables and
$-5.12 \leq x_j \leq 5.12 (n+1 \leq j \leq m)$ are $m - n$ continuous variables.



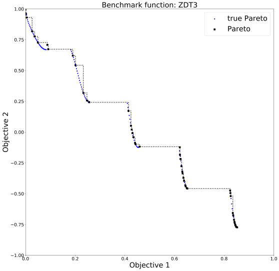
Comparison against GA.

Multi-objective: 2 objectives

(joint work w/ Mike Eldred)



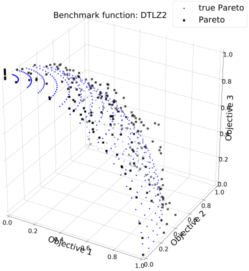
ZDT1.



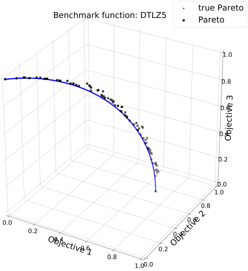
ZDT3.

Multi-objective: 3 objectives

(joint work w/ Mike Eldred)



DTLZ2.



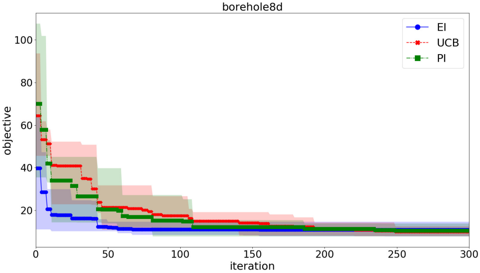
DTLZ5.

Multi-fidelity: borehole8d

(joint work w/ Scott McCann, Tim Wildey)

$$f_H(\boldsymbol{x}) = \frac{2\pi x_3(x_4 - x_6)}{\log(x_2/x_1)\left(1 + \frac{2x_7x_3}{\log(x_2/x_1)x_1^2x_8} + \frac{x_3}{x_5}\right)}, \tag{50}$$

$$f_L(\boldsymbol{x}) = \frac{5x_3(x_4 - x_6)}{\log(x_2/x_1)\left(1.5 + \frac{2x_7x_3}{\log(x_2/x_1)x_1^2x_8} + \frac{x_3}{x_5}\right)}. \tag{51}$$



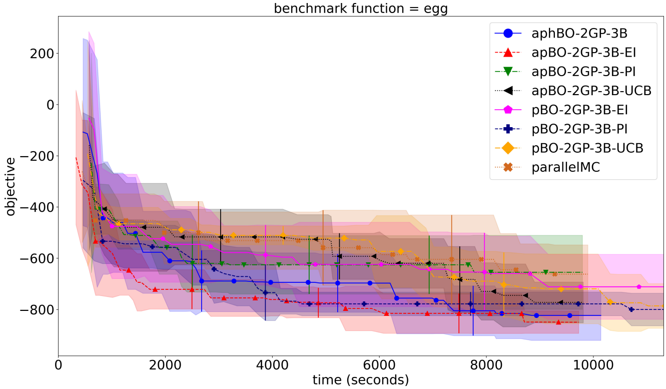
Borehole function (8d) - 2 levels of fidelity.

Asynchronous parallel

(joint work w/ Mike Eldred)

$$f(\mathbf{x}) = \frac{1}{0.839}\left[1.1 - \sum_{i=1}^{4} \alpha_i \exp\left(-\sum_{j=4}^{3} A_{ij}(x_j - P_{ij})^2\right)\right], \tag{52}$$

Hart4 function, $t \sim \mathcal{U}[30, 900]$ on $\mathcal{X} = [0, 1]^4$.

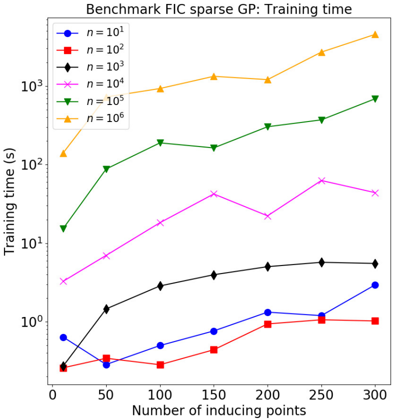


Asynchronous parallel vs. batch-parallel on egg function.

Sparse GP for Big Data

(joint work w/ Bart G van Bloemen Waanders)

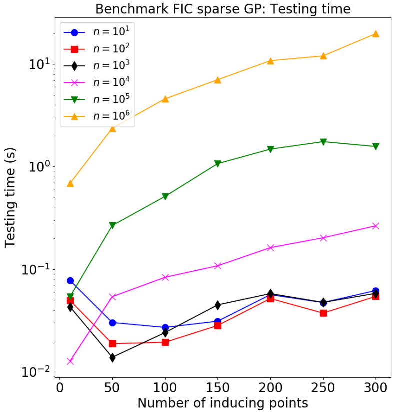
- Intel Xeon Platinum 8160 CPU @ 2.10GHz
- 24 cores, 48 threads
- RHEL 7.1 (Maipo)
- 180 GB of memory
- sphere function $y = \left(\sum_{i=1}^{3} x_i\right)^2$, $\mathcal{X} = [-1, 1]^3$
- training data points: $n \in \{10^1, 10^2, \ldots, 10^6\}$
- number of inducing points: $m \in \{10, 50, 100, \ldots, 300\}$
- GPstuff with SuitSparse toolbox on MATLAB
- $m = 300$, $n = 10^6$ takes ∼48 minutes



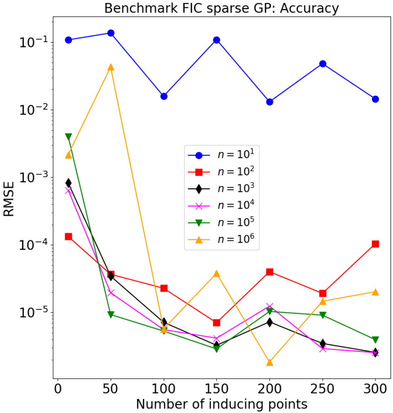
Benchmark of training time.

Sparse GP for Big Data

(joint work w/ Bart G van Bloemen Waanders)



Benchmark of testing time.



Benchmark of accuracy.

High-dimensional (with low effective dimensionality): Gaussian random projection

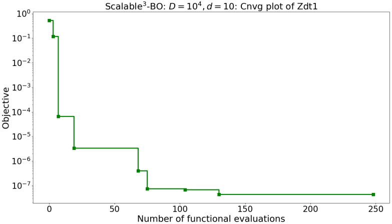
(joint work w/ Bart G van Bloemen Waanders)

The modified ZDT1 function, which is defined on $[-1, 1]^D$, is

$$f_2(\mathbf{x}) = g\left(1 - \sqrt{\frac{x_1^2}{g}}\right), \qquad (53)$$

where $g = 1 + 9\left(\sum_{i=2}^{D} \frac{x_i}{D-1}\right)^2$.

- (non-unique) global minimizer $\mathbf{x}^* = [1, 0, \ldots, 0]$
- $f_2(\mathbf{x}^*) = 0$
- $D = 10^4$
- $d = 10$
- $d_e = 2$

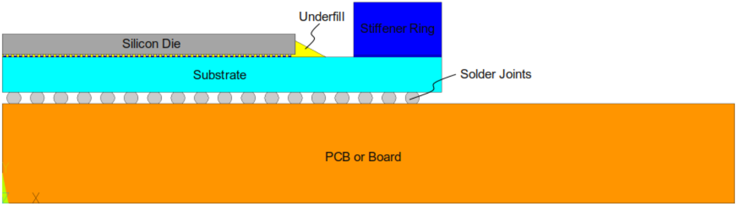


Convergence plot with $D = 10,000$, $d = 10$.

Flip-chip BGA package design (FEM)

(joint work w/ Scott McCann (Xilinx))

FE model geometry

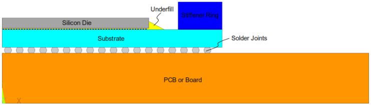


- 2.5D FE on (ANSYS) APDL: half symmetry to reduce comp.
- evaluate component warpage at 20°C and 200°C, and the strain energy density to predict the fatigue life of the solder joints during thermal cycling
- two levels of fidelity: varies mesh density parameter
- average comp. time: 0.4 CPU hr for low-fidelity, $\sim$ 1 CPU hr for high-fidelity

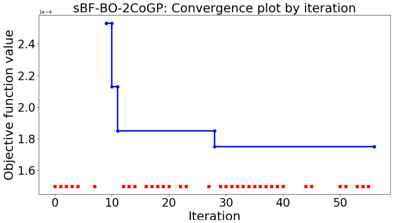
Flip-chip BGA package design (FEM)

(joint work w/ Scott McCann (Xilinx))

FE model

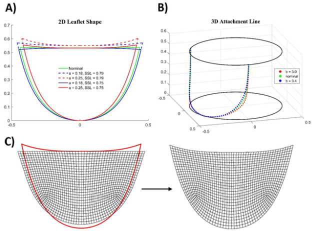


Conv. plot at high-fidelity

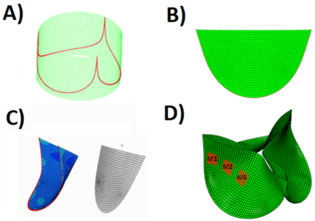


Heart valve optimization (FEM)

(joint work w/ Yan Wang, Wei Sun)

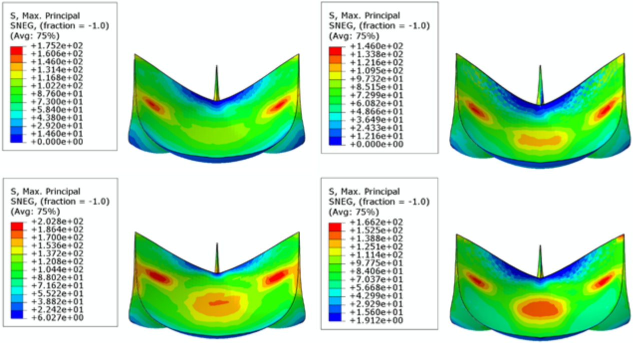


(A) Parameterization of 2D leaflet geometry; (B) 3D attachment edge shape; (C) Template leaflet mesh and nodes transformation.



(A) 3D suturing line; (B) 2D attachment edge; (C) 2D-to-3D transformation; (D) Node and element mid-leaflet sets.

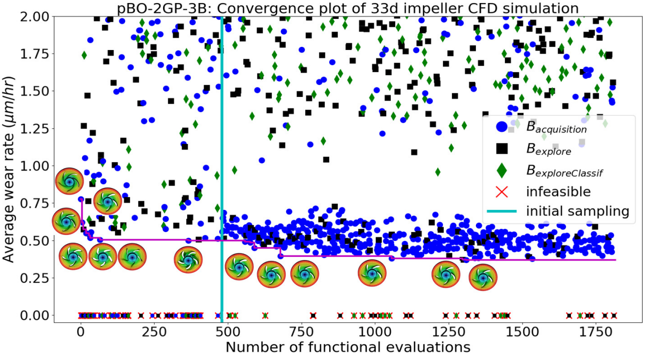
Heart valve optimization (FEM)



Comparison of nominal (left) and optimized (right) designs for bovine (top) and porcine (bottom) leaflet materials under diastolic pressurization.

Impeller design optimization using CFD

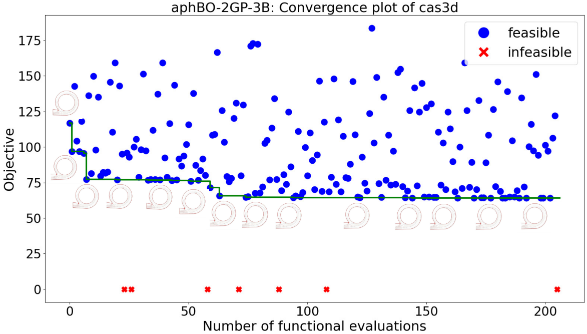
(joint work w/ GIW Industries)



Design evolution of 33d slurry pump impeller using a solid-liquid multi-phase CFD package.

Casing design optimization using CFD

(joint work w/ GIW Industries)



Design evolution of 14d slurry pump casing using a solid-liquid multi-phase CFD package.

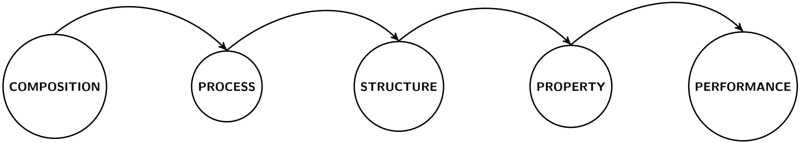
Inverse problems in process-structure

(joint work w/ Laura Swiler, John Mitchell, Tim Wildey, Theron Rodgers)

Reference

Anh Tran et al. (2020a). "An active-learning high-throughput microstructure calibration framework for process-structure linkage in materials informatics". In: *Acta Materialia* 194, pp. 80–92

- process: $\mathbf{x} + \delta, \delta \sim \mathcal{U}[\underline{\delta}, \overline{\delta}]$ – controllable within a tolerance δ
- (micro)structure – spatio-temporal noisy, questionable microstructure representations (physics-based vs. data-driven), image (i.e. high-dimensional), limited/scarce data
- property: $y = f(\mathbf{x}) + \varepsilon, \varepsilon \sim \mathcal{N}(0, \sigma^2)$ – noisy observations



Inverse problems in process-structure

(joint work w/ Laura Swiler, John Mitchell, Tim Wildey, Theron Rodgers)

A formal problem statement:

- there exists a forward tool $f(\cdot)$ to predict microstructure, $u = f(x)$ (represented as images)
- given a target u^* (represented as images)
- task: find x^* such that $f(x^*) = u^* \approx u$

$\approx$ is defined in the sense of statistical equivalence for microstructures, $p_{\mathcal{D}}$ is the p.d.f. of statistical microstructure descriptors $\mathcal{D}$, i.e.

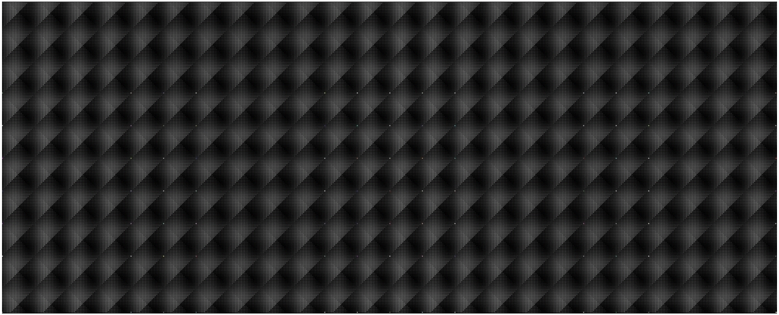
$$p_{\mathcal{D}} : \Omega \to L^1 : p_{\mathcal{D}}(u^*) \approx p_{\mathcal{D}}(u) \tag{54}$$

$$d\Big(p_{\mathcal{D}}(u^*), p_{\mathcal{D}}(u)\Big) \leq \text{TOL} \tag{55}$$

<u>Hint:</u> quantitatively differentiate microstructures using statistical microstructure descriptors

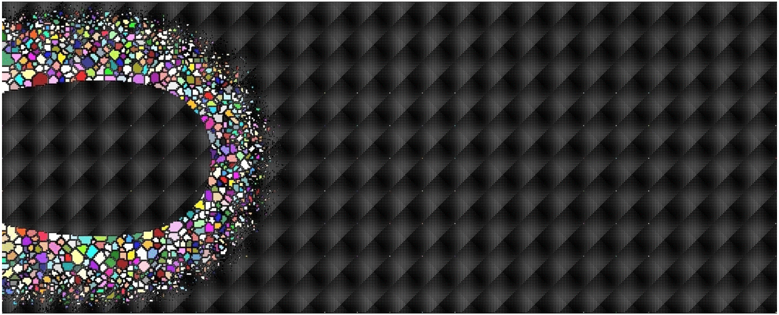
Inverse problems in process-structure

(joint work w/ Laura Swiler, John Mitchell, Tim Wildey, Theron Rodgers)



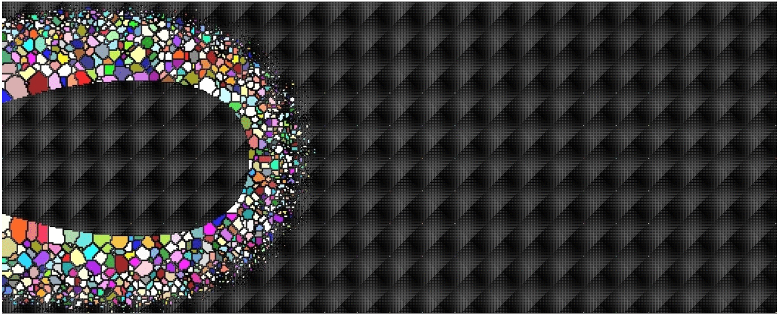
Inverse problems in process-structure

(joint work w/ Laura Swiler, John Mitchell, Tim Wildey, Theron Rodgers)



Inverse problems in process-structure

(joint work w/ Laura Swiler, John Mitchell, Tim Wildey, Theron Rodgers)

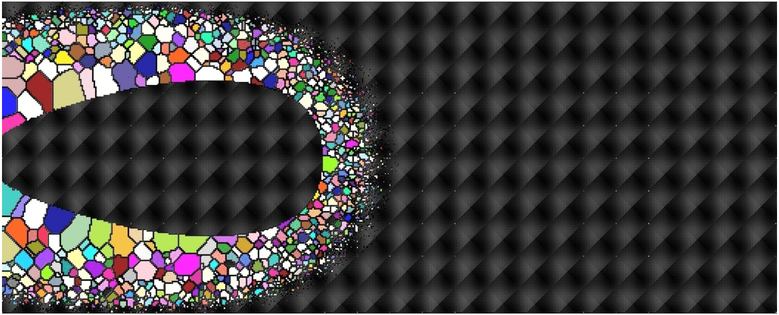


Inverse problems in process-structure

(joint work w/ Laura Swiler, John Mitchell, Tim Wildey, Theron Rodgers)

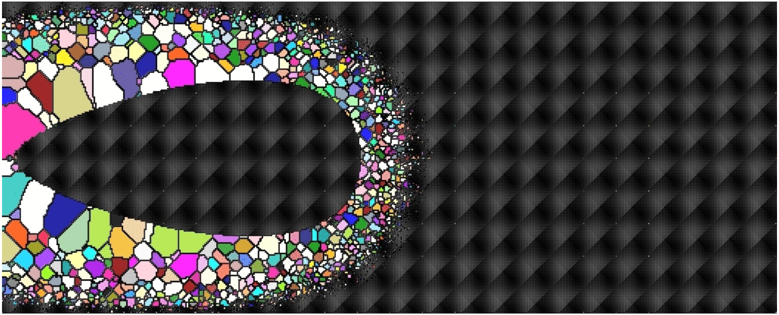
Inverse problems in process-structure

(joint work w/ Laura Swiler, John Mitchell, Tim Wildey, Theron Rodgers)



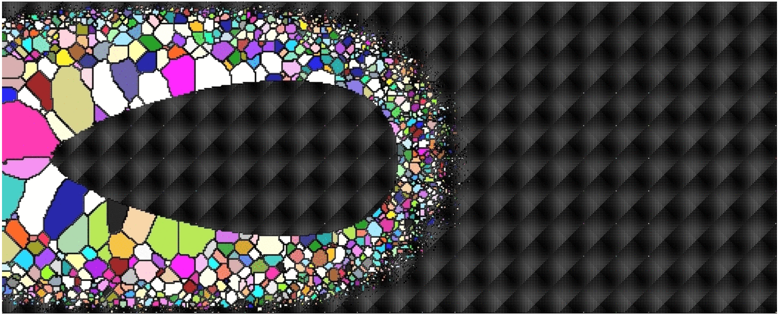
Inverse problems in process-structure

(joint work w/ Laura Swiler, John Mitchell, Tim Wildey, Theron Rodgers)



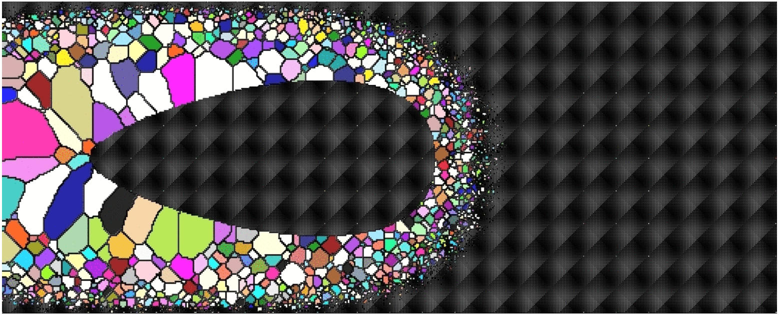
Inverse problems in process-structure

(joint work w/ Laura Swiler, John Mitchell, Tim Wildey, Theron Rodgers)



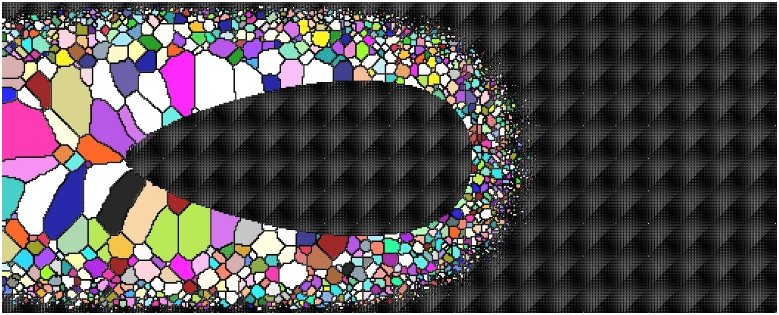
Inverse problems in process-structure

(joint work w/ Laura Swiler, John Mitchell, Tim Wildey, Theron Rodgers)



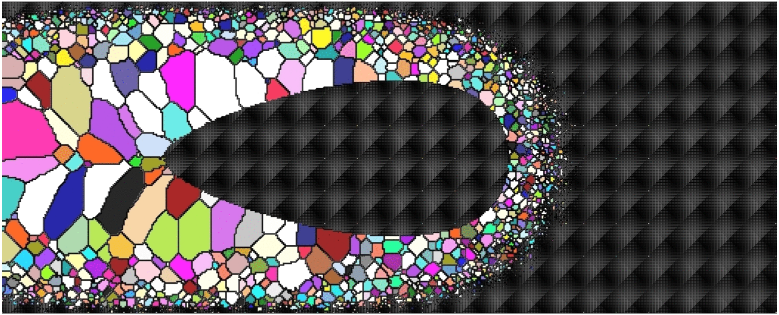
Inverse problems in process-structure

(joint work w/ Laura Swiler, John Mitchell, Tim Wildey, Theron Rodgers)



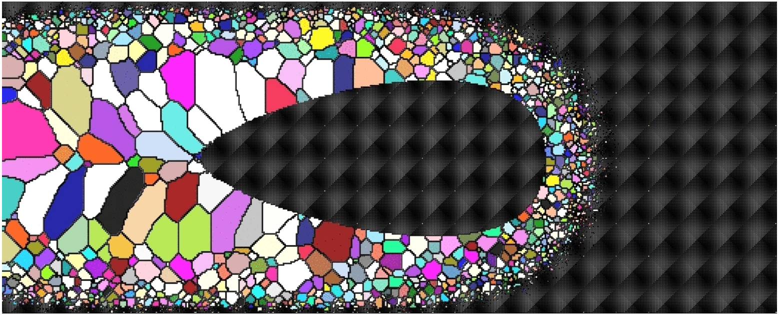
Inverse problems in process-structure

(joint work w/ Laura Swiler, John Mitchell, Tim Wildey, Theron Rodgers)



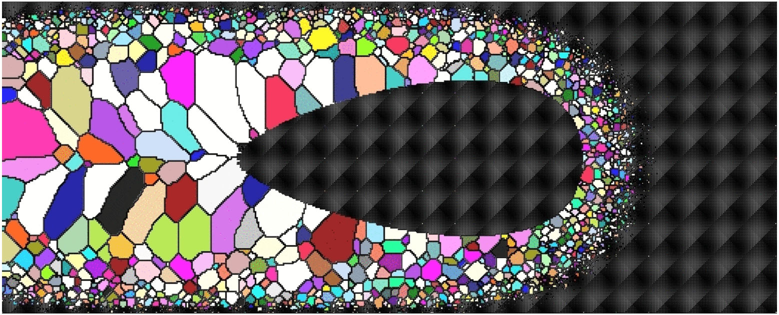
Inverse problems in process-structure

(joint work w/ Laura Swiler, John Mitchell, Tim Wildey, Theron Rodgers)



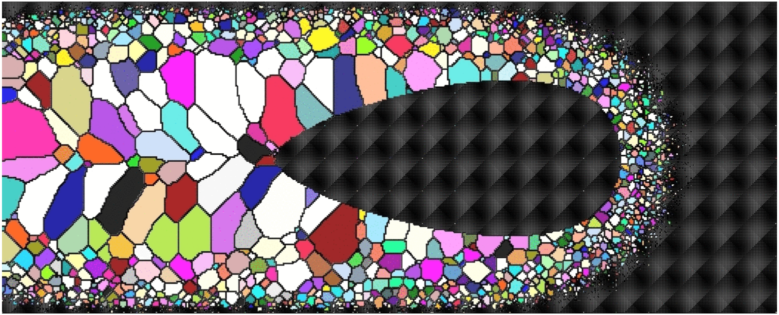
Inverse problems in process-structure

(joint work w/ Laura Swiler, John Mitchell, Tim Wildey, Theron Rodgers)



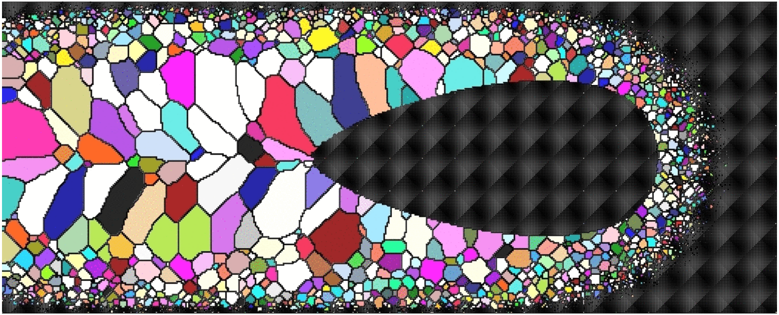
Inverse problems in process-structure

(joint work w/ Laura Swiler, John Mitchell, Tim Wildey, Theron Rodgers)



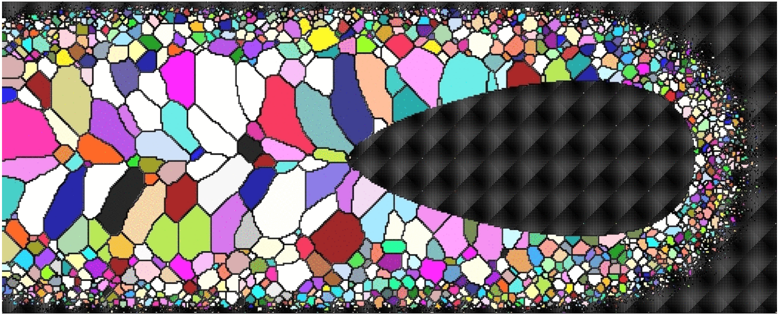
Inverse problems in process-structure

(joint work w/ Laura Swiler, John Mitchell, Tim Wildey, Theron Rodgers)



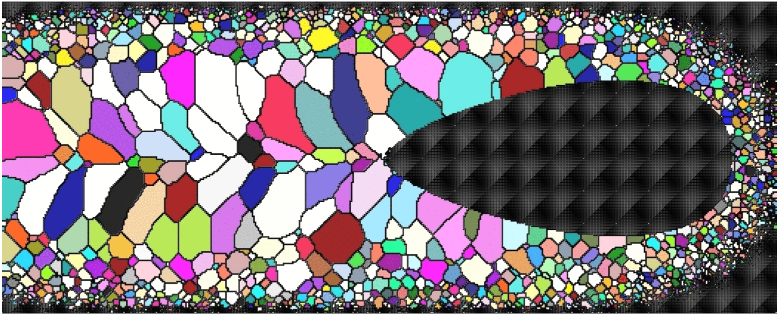
Inverse problems in process-structure

(joint work w/ Laura Swiler, John Mitchell, Tim Wildey, Theron Rodgers)



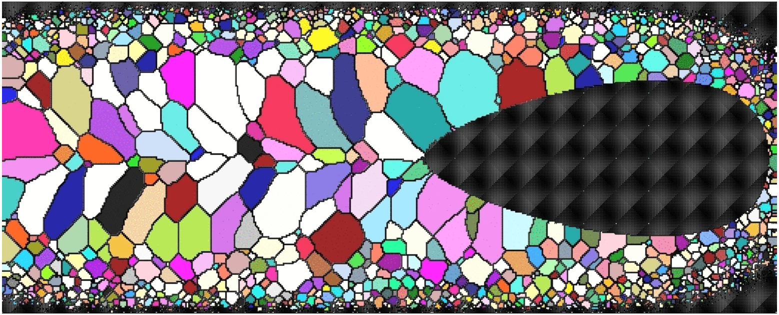
Inverse problems in process-structure

(joint work w/ Laura Swiler, John Mitchell, Tim Wildey, Theron Rodgers)



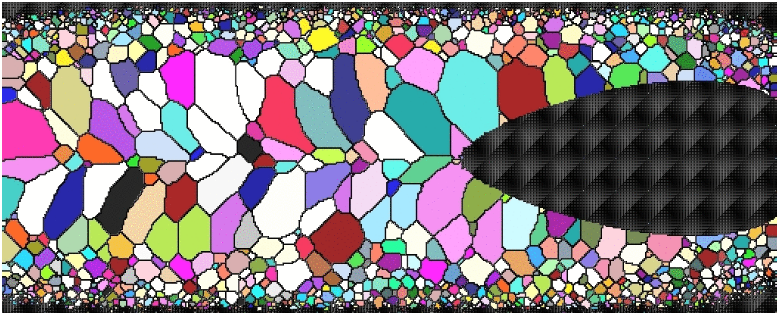
Inverse problems in process-structure

(joint work w/ Laura Swiler, John Mitchell, Tim Wildey, Theron Rodgers)



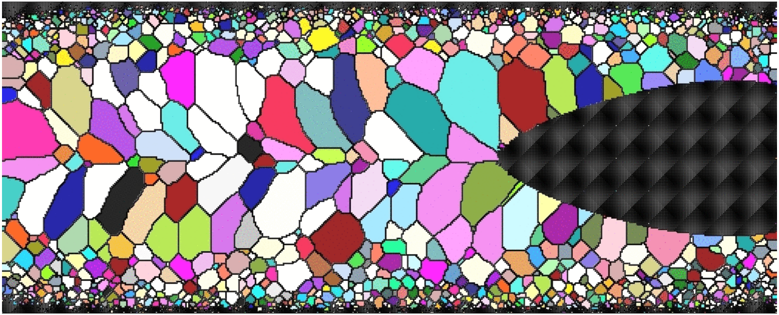
Inverse problems in process-structure

(joint work w/ Laura Swiler, John Mitchell, Tim Wildey, Theron Rodgers)



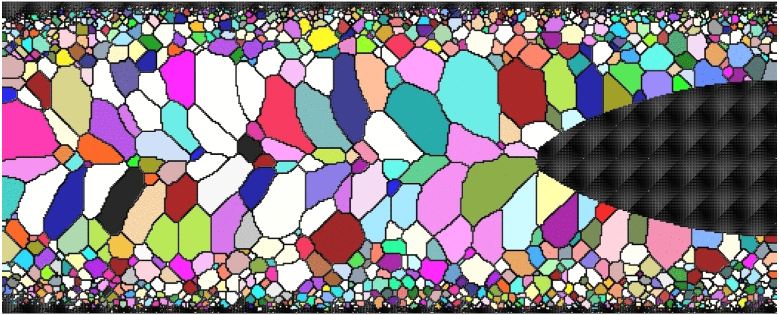
Inverse problems in process-structure

(joint work w/ Laura Swiler, John Mitchell, Tim Wildey, Theron Rodgers)



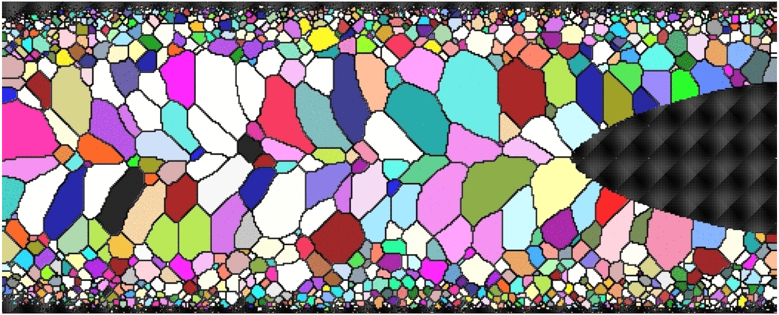
Inverse problems in process-structure

(joint work w/ Laura Swiler, John Mitchell, Tim Wildey, Theron Rodgers)



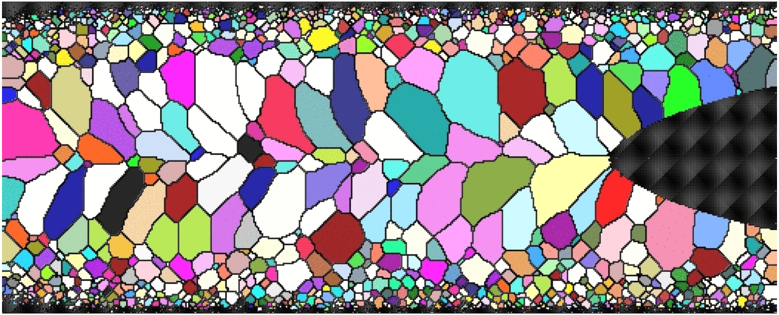
Inverse problems in process-structure

(joint work w/ Laura Swiler, John Mitchell, Tim Wildey, Theron Rodgers)



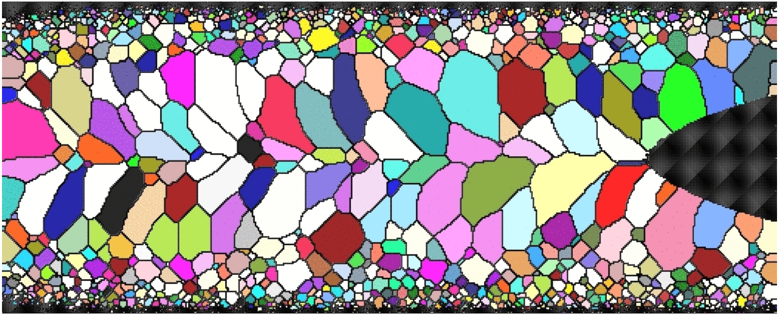
Inverse problems in process-structure

(joint work w/ Laura Swiler, John Mitchell, Tim Wildey, Theron Rodgers)



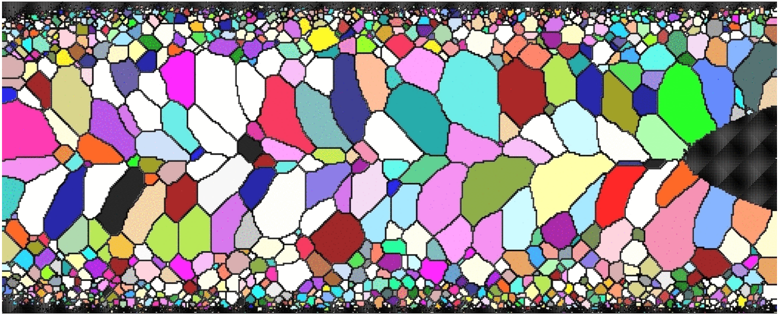
Inverse problems in process-structure

(joint work w/ Laura Swiler, John Mitchell, Tim Wildey, Theron Rodgers)



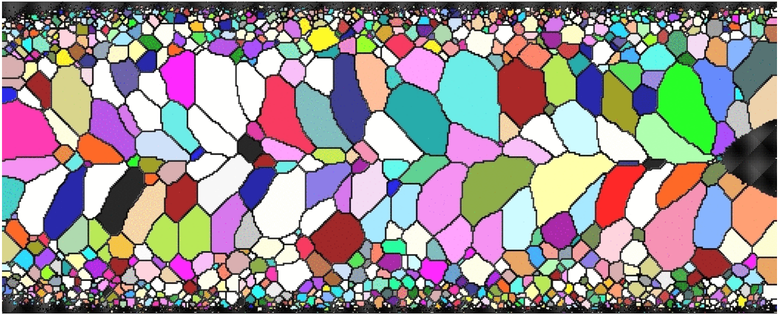
Inverse problems in process-structure

(joint work w/ Laura Swiler, John Mitchell, Tim Wildey, Theron Rodgers)



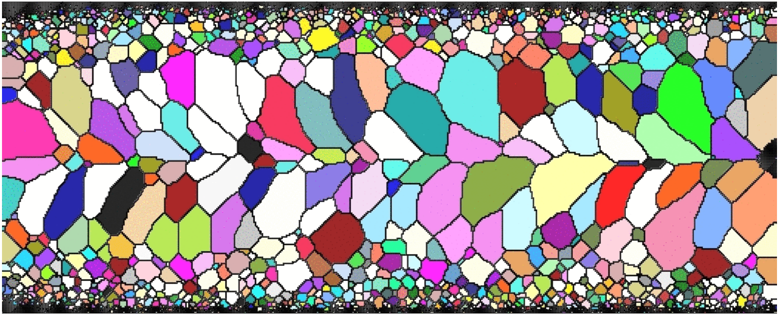
Inverse problems in process-structure

(joint work w/ Laura Swiler, John Mitchell, Tim Wildey, Theron Rodgers)



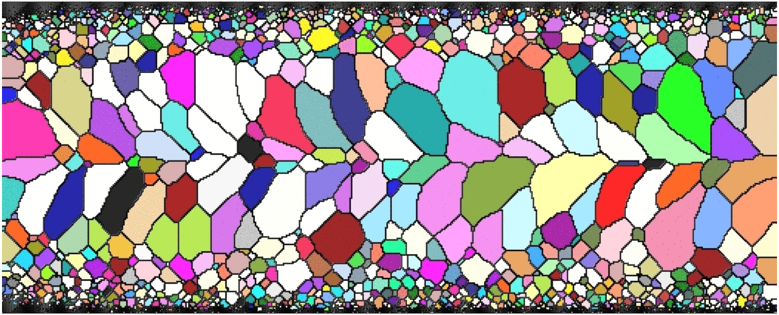
Inverse problems in process-structure

(joint work w/ Laura Swiler, John Mitchell, Tim Wildey, Theron Rodgers)



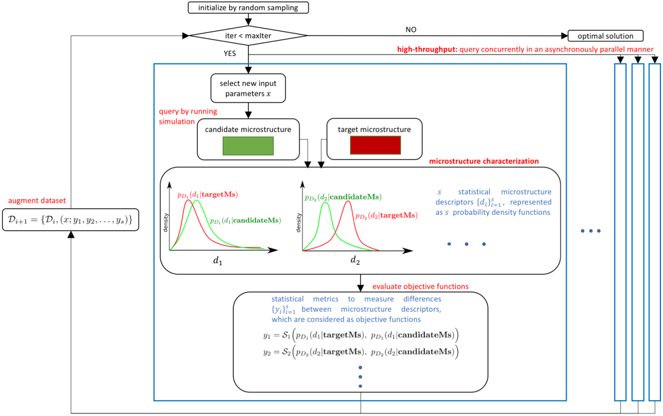
Inverse problems in process-structure

(joint work w/ Laura Swiler, John Mitchell, Tim Wildey, Theron Rodgers)



Inverse problems in process-structure

(joint work w/ Laura Swiler, John Mitchell, Tim Wildey, Theron Rodgers)


initialize by random sampling
iter < maxIter
NO
optimal solution
YES
high-throughput: query concurrently in an asynchronously parallel manner
select new input parameters x
query by running simulation
candidate microstructure
target microstructure
microstructure characterization
$p_{D_1}(d_1|\mathbf{targetMs})$
$p_{D_1}(d_1|\mathbf{candidateMs})$
density
d_1
$p_{D_2}(d_2|\mathbf{candidateMs})$
$p_{D_2}(d_2|\mathbf{targetMs})$
density
d_2
augment dataset
evaluate objective functions

$\mathcal{D}_{i+1} = \{\mathcal{D}_i, (x; y_1, y_2, \dots, y_s)\}$

s statistical microstructure descriptors $\{d_i\}_{i=1}^s$, represented as s probability density functions

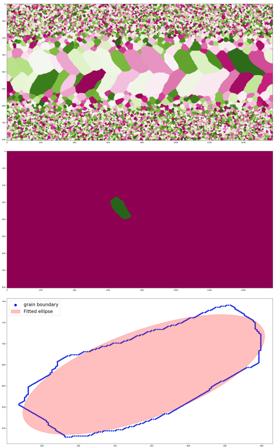
statistical metrics to measure differences $\{y_i\}_{i=1}^s$ between microstructure descriptors, which are considered as objective functions

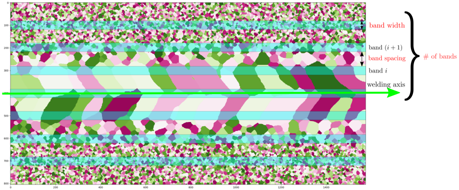
$$y_1 = \mathcal{S}_1\Big(p_{D_1}(d_1|\mathbf{targetMs}),\ p_{D_1}(d_1|\mathbf{candidateMs})\Big)$$
$$y_2 = \mathcal{S}_2\Big(p_{D_2}(d_2|\mathbf{targetMs}),\ p_{D_2}(d_2|\mathbf{candidateMs})\Big)$$

An asynchronous parallel Bayesian optimization workflow for inverse problems in process-structure linkage.

Inverse problems in process-structure

(joint work w/ Laura Swiler, John Mitchell, Tim Wildey, Theron Rodgers)

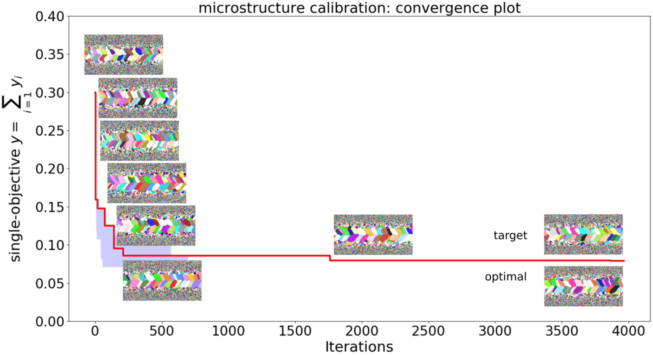




Collecting local + global statistical microstructure descriptors given a microstructure.

Inverse problems in process-structure

(joint work w/ Laura Swiler, John Mitchell, Tim Wildey, Theron Rodgers)



Reverse engineering an AM specimen through kinetic Monte Carlo (Sandia/SPPARKS).

Inverse problems in composition-property

(joint work w/ Julien Tranchida, Aidan Thompson, Tim Wildey)

Reference

Active learning from chemical composition space to material property
Anh Tran et al. (2020b). "Multi-fidelity machine-learning with uncertainty quantification and Bayesian optimization for materials design: Application to ternary random alloys". In: *The Journal of Chemical Physics* 153 (7), p. 074705.

Main ideas:

- Forward models:
  - MD-MLIAP: low-fidelity (low accuracy, low cost)
  - DFT: high-fidelity (high accuracy, high cost)
- Exploit correlation between low- and high-fidelity models
- Input: chemical composition
- Output/QoI: bulk modulus B_0
- What chemical composition would optimize the QoI?

Inverse problems in composition-property

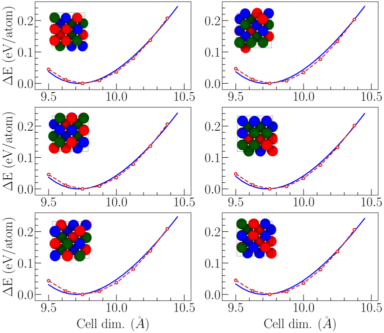
(joint work w/ Julien Tranchida, Aidan Thompson, Tim Wildey)

Ab-initio:

- DFT implemented in Quantum ESPRESSO
- high cost + high accuracy → high-fidelity

MD:

- MD with ML interatomic potential (SNAP)
- orders of magnitude faster
- low cost + low accuracy → low-fidelity



EOS calculations for 6 configs. red line: DFT; blue line: MD + SNAP

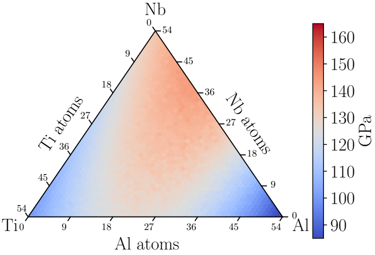
Birch-Murnaghan polynomials for B_0:

$$E(V) = E_0 + \frac{9V_0 B_0}{16}\left\{B_0'\left[\left(\frac{V_0}{V}\right)^{\frac{3}{2}} - 1\right]^3 + \left[\left(\frac{V_0}{V}\right)^{\frac{3}{2}} - 1\right]^2\left[6 - 4\left(\frac{V_0}{V}\right)^{\frac{3}{2}}\right]\right\} \quad (56)$$

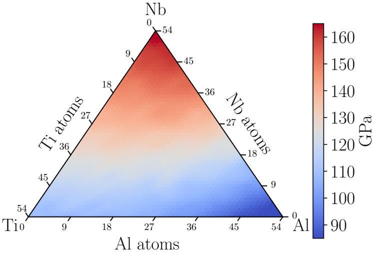
Inverse problems in composition-property

(joint work w/ Julien Tranchida, Aidan Thompson, Tim Wildey)

$R^2 = 0.7122$: not exactly the same



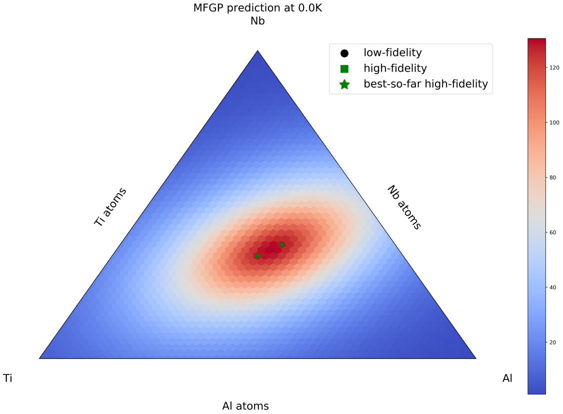
Low-fidelity: MD with SNAP potential.



Multi-fidelity GP ≈ high-fidelity: DFT.

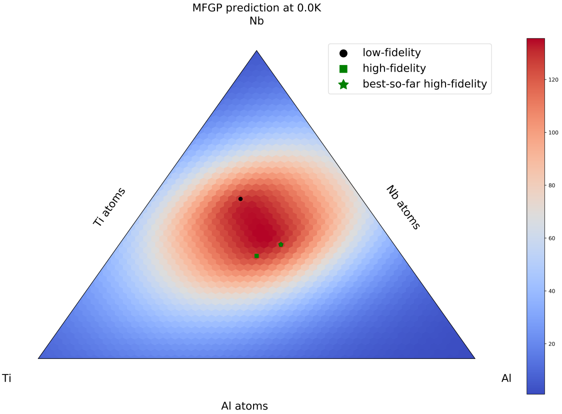
Inverse problems in composition-property

(joint work w/ Julien Tranchida, Aidan Thompson, Tim Wildey)



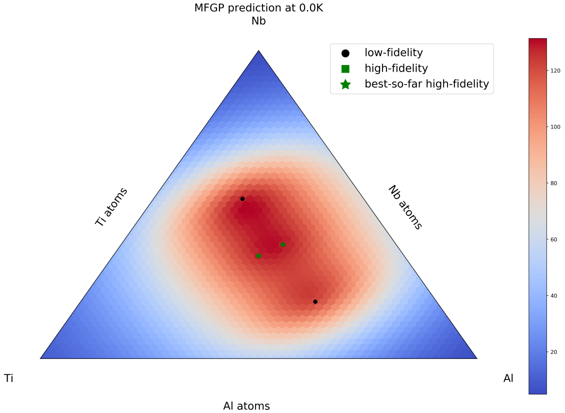
Inverse problems in composition-property

(joint work w/ Julien Tranchida, Aidan Thompson, Tim Wildey)



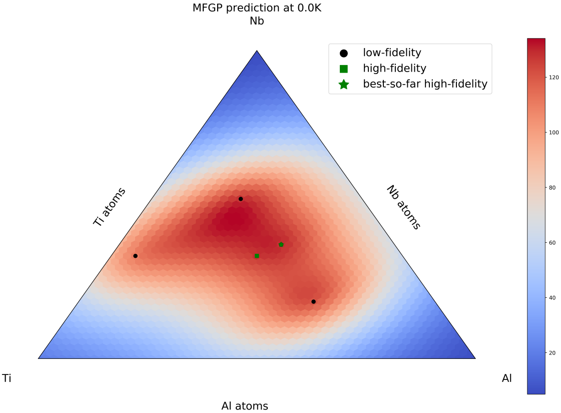
Inverse problems in composition-property

(joint work w/ Julien Tranchida, Aidan Thompson, Tim Wildey)



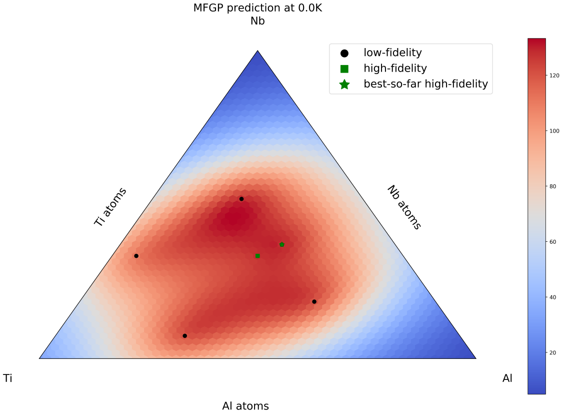
Inverse problems in composition-property

(joint work w/ Julien Tranchida, Aidan Thompson, Tim Wildey)



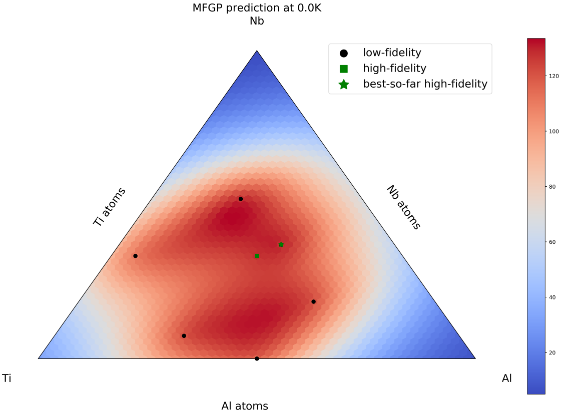
Inverse problems in composition-property

(joint work w/ Julien Tranchida, Aidan Thompson, Tim Wildey)



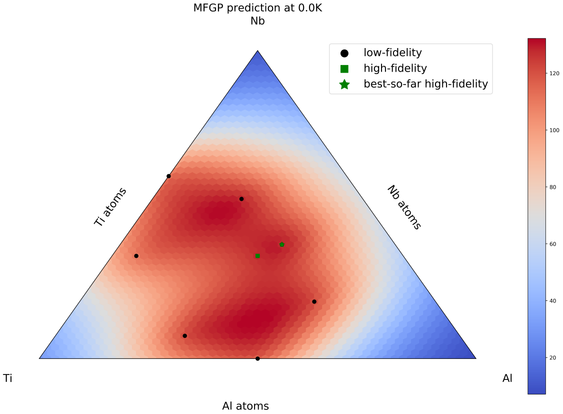
Inverse problems in composition-property

(joint work w/ Julien Tranchida, Aidan Thompson, Tim Wildey)



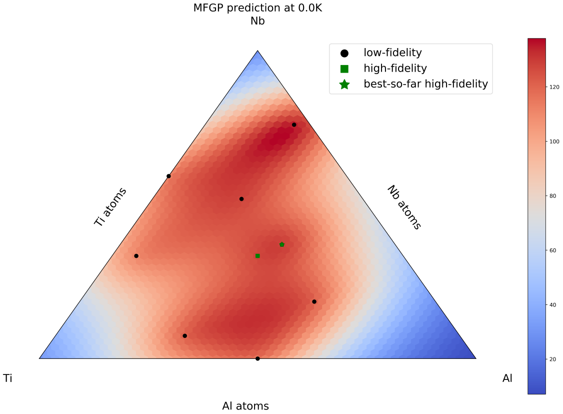
Inverse problems in composition-property

(joint work w/ Julien Tranchida, Aidan Thompson, Tim Wildey)



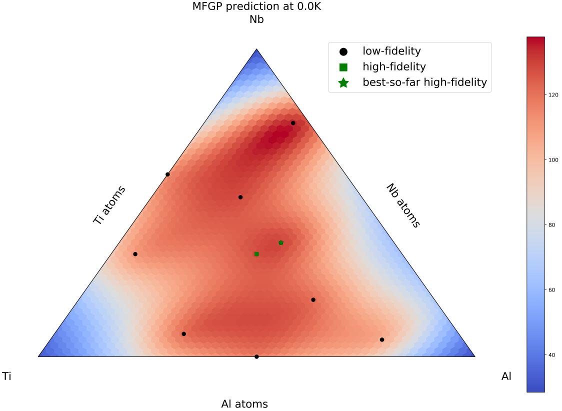
Inverse problems in composition-property

(joint work w/ Julien Tranchida, Aidan Thompson, Tim Wildey)



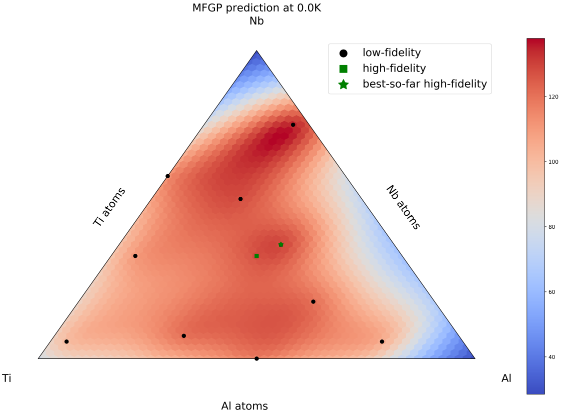
Inverse problems in composition-property

(joint work w/ Julien Tranchida, Aidan Thompson, Tim Wildey)



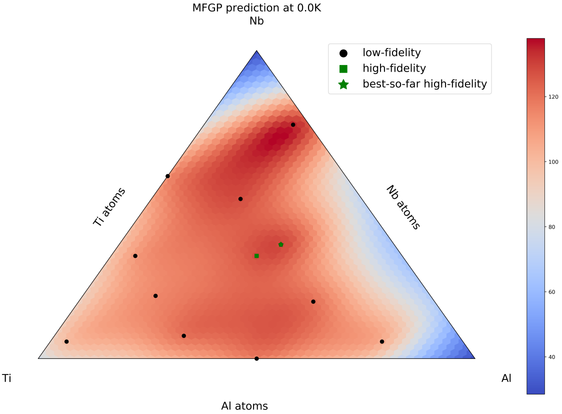
Inverse problems in composition-property

(joint work w/ Julien Tranchida, Aidan Thompson, Tim Wildey)



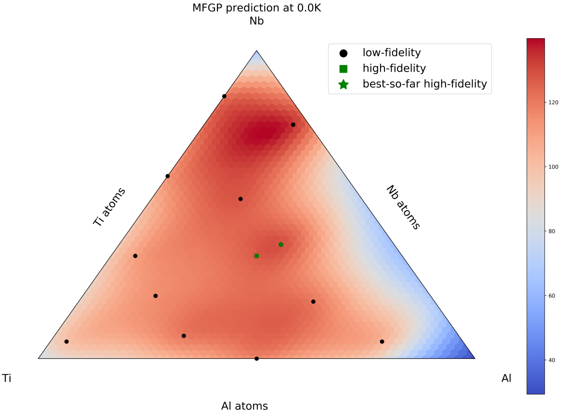
Inverse problems in composition-property

(joint work w/ Julien Tranchida, Aidan Thompson, Tim Wildey)



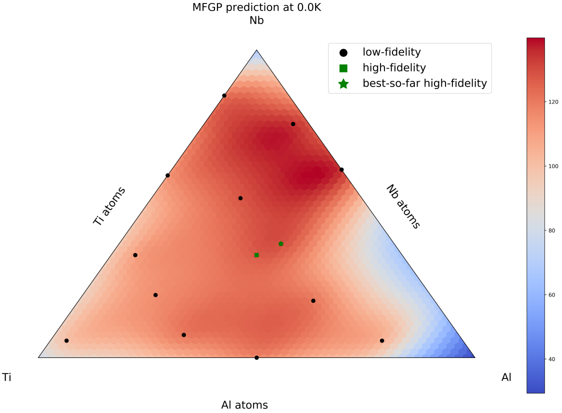
Inverse problems in composition-property

(joint work w/ Julien Tranchida, Aidan Thompson, Tim Wildey)



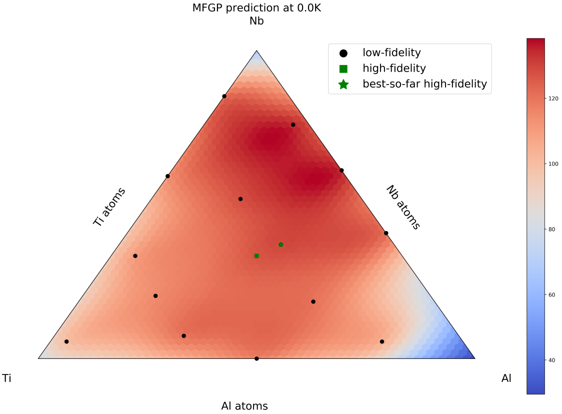
Inverse problems in composition-property

(joint work w/ Julien Tranchida, Aidan Thompson, Tim Wildey)



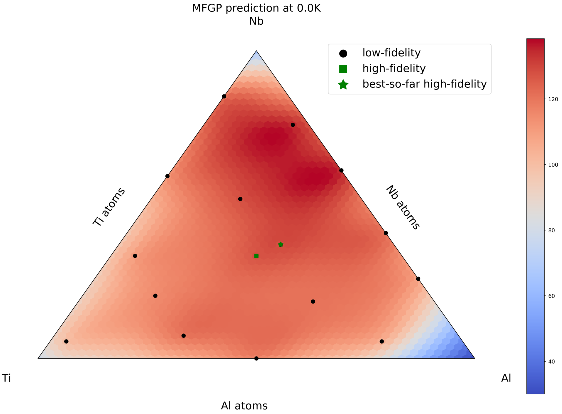
Inverse problems in composition-property

(joint work w/ Julien Tranchida, Aidan Thompson, Tim Wildey)



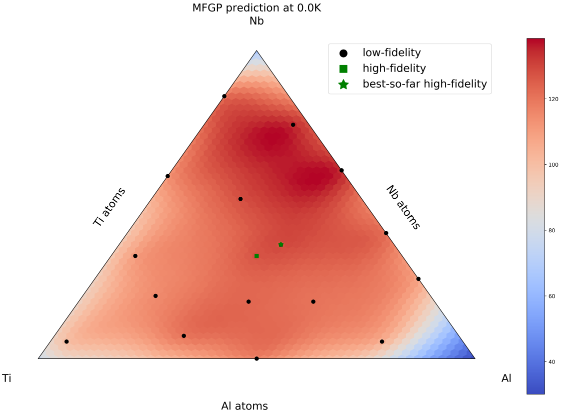
Inverse problems in composition-property

(joint work w/ Julien Tranchida, Aidan Thompson, Tim Wildey)



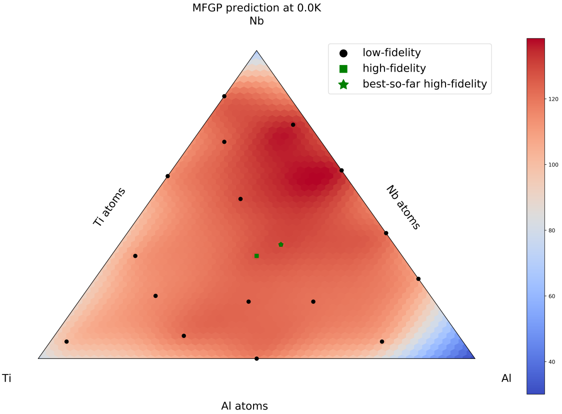
Inverse problems in composition-property

(joint work w/ Julien Tranchida, Aidan Thompson, Tim Wildey)



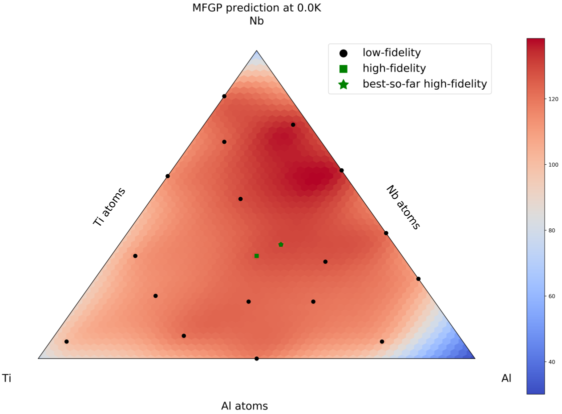
Inverse problems in composition-property

(joint work w/ Julien Tranchida, Aidan Thompson, Tim Wildey)



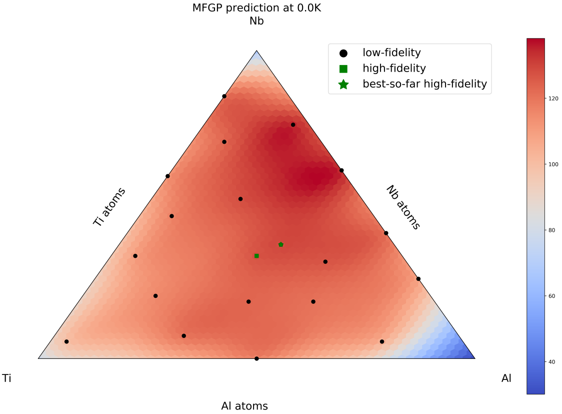
Inverse problems in composition-property

(joint work w/ Julien Tranchida, Aidan Thompson, Tim Wildey)



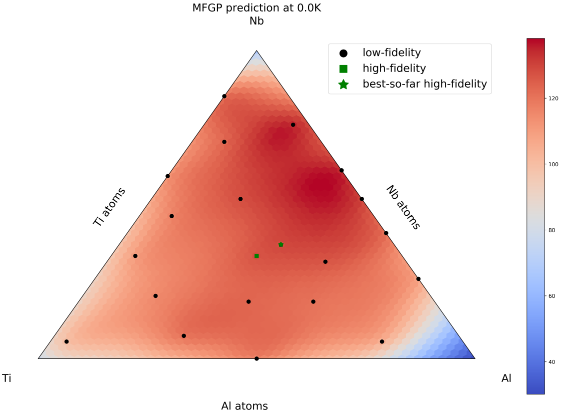
Inverse problems in composition-property

(joint work w/ Julien Tranchida, Aidan Thompson, Tim Wildey)



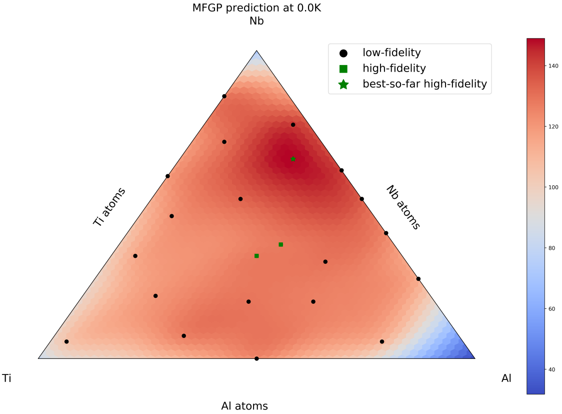
Inverse problems in composition-property

(joint work w/ Julien Tranchida, Aidan Thompson, Tim Wildey)



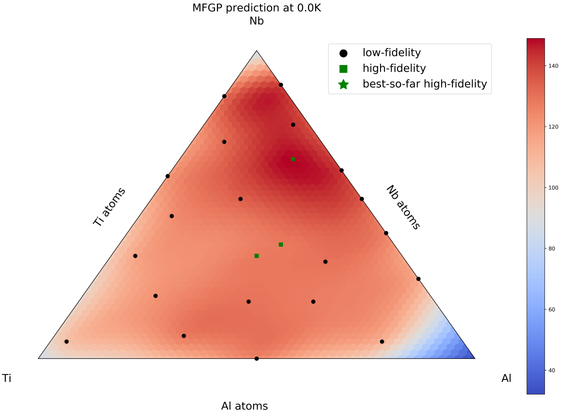
Inverse problems in composition-property

(joint work w/ Julien Tranchida, Aidan Thompson, Tim Wildey)



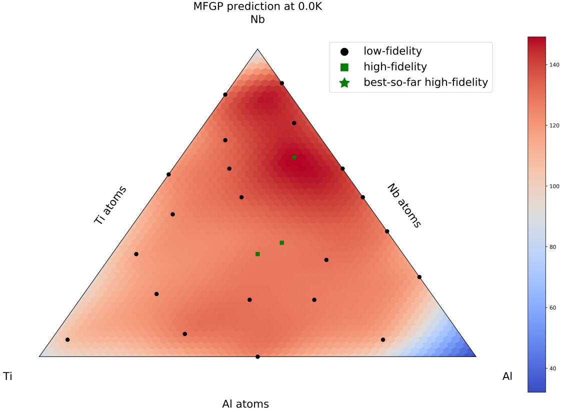
Inverse problems in composition-property

(joint work w/ Julien Tranchida, Aidan Thompson, Tim Wildey)



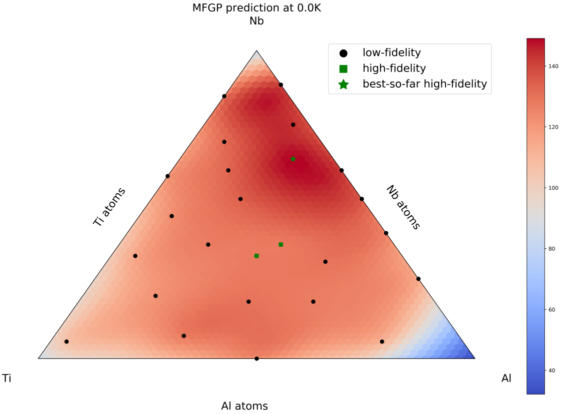
Inverse problems in composition-property

(joint work w/ Julien Tranchida, Aidan Thompson, Tim Wildey)



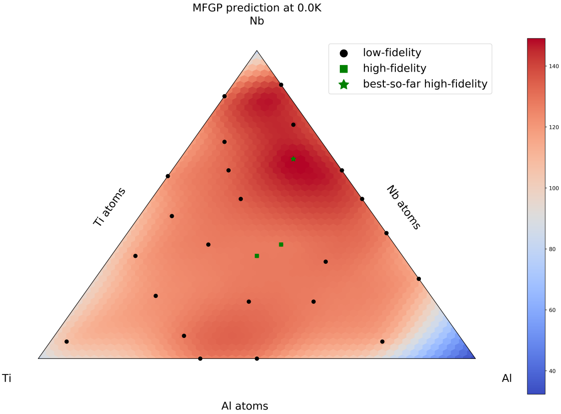
Inverse problems in composition-property

(joint work w/ Julien Tranchida, Aidan Thompson, Tim Wildey)



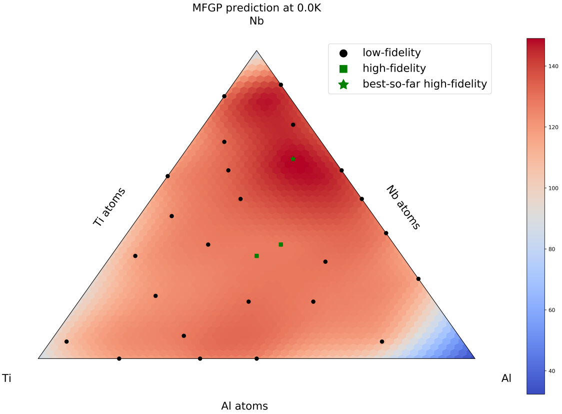
Inverse problems in composition-property

(joint work w/ Julien Tranchida, Aidan Thompson, Tim Wildey)



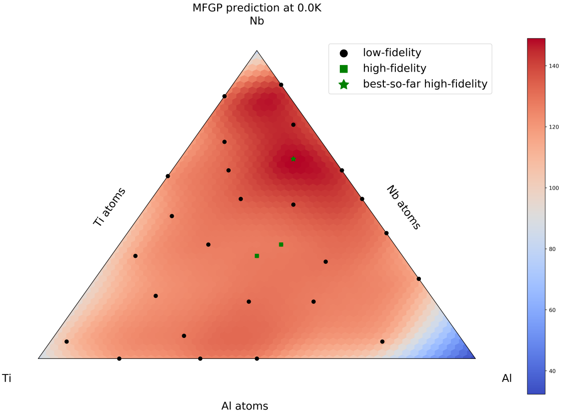
Inverse problems in composition-property

(joint work w/ Julien Tranchida, Aidan Thompson, Tim Wildey)



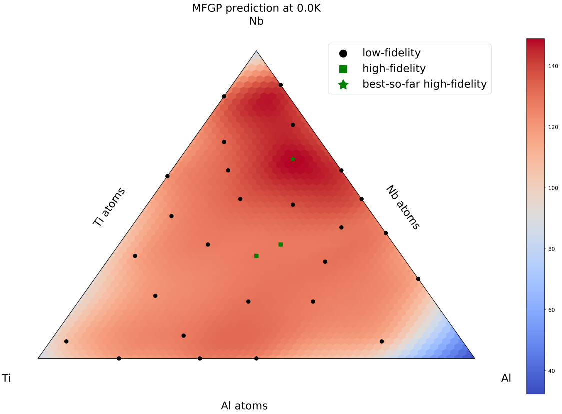
Inverse problems in composition-property

(joint work w/ Julien Tranchida, Aidan Thompson, Tim Wildey)



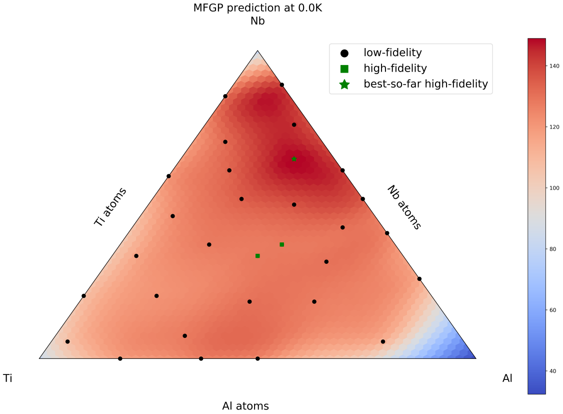
Inverse problems in composition-property

(joint work w/ Julien Tranchida, Aidan Thompson, Tim Wildey)



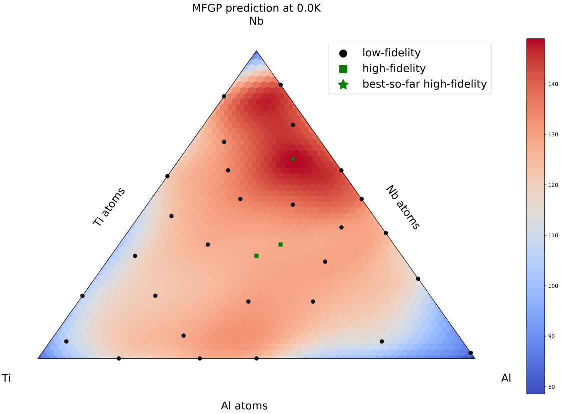
Inverse problems in composition-property

(joint work w/ Julien Tranchida, Aidan Thompson, Tim Wildey)



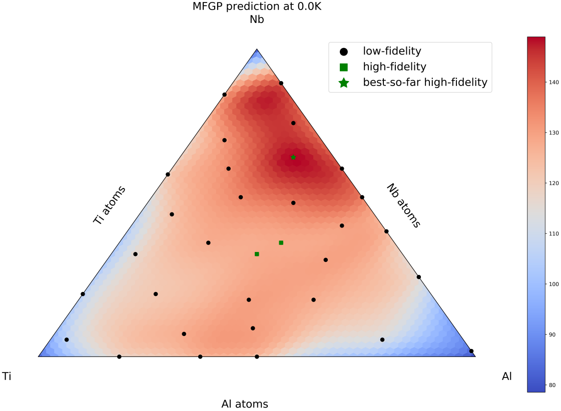
Inverse problems in composition-property

(joint work w/ Julien Tranchida, Aidan Thompson, Tim Wildey)



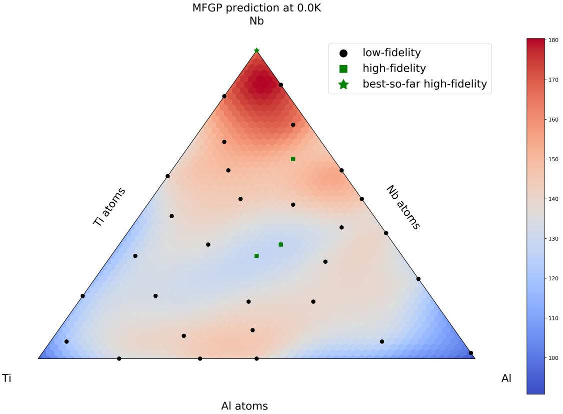
Inverse problems in composition-property

(joint work w/ Julien Tranchida, Aidan Thompson, Tim Wildey)



Inverse problems in composition-property

(joint work w/ Julien Tranchida, Aidan Thompson, Tim Wildey)



Inverse problems in structure-property

(joint work w/ Tim Wildey)

Reference

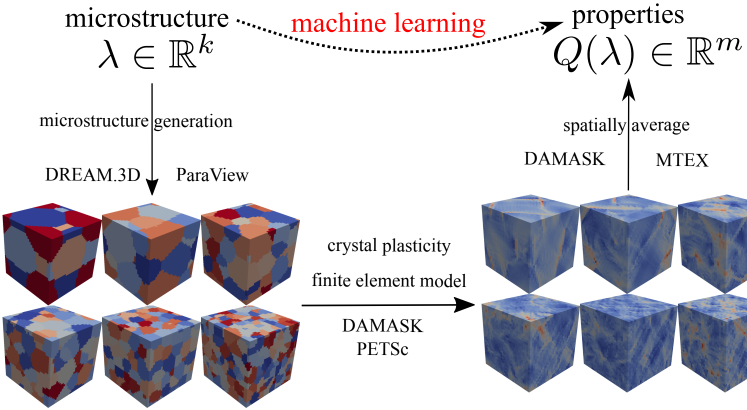
Anh Tran and Tim Wildey (2020). "Solving stochastic inverse problems for property-structure linkages using data-consistent inversion and machine learning". In: *JOM* 73, pp. 72–89

Main ideas:

- require some statistical treatment for stochastic microstructure, due to the inherent randomness
- parameterize *deterministic* λ as microstructure features, e.g. average grain size, Weibull parameters, etc.
- sample N microstructure RVE (DREAM.3D)
- run crystal plasticity over RVE ensemble (DAMASK)
- collect $Q(\lambda)$ as quantities of interest
- approximate $Q(\cdot)$ by machine learning, e.g. heteroscedastic GP

Inverse problems in structure-property

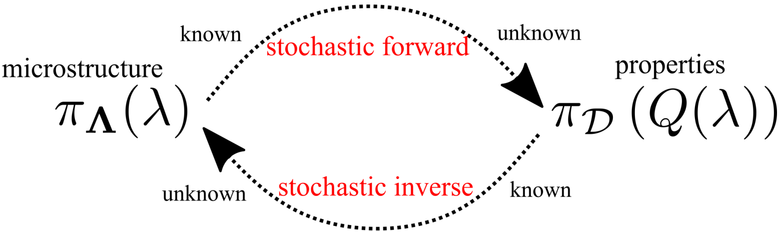
(joint work w/ Tim Wildey)



Microstructure-homogenized properties map over an ensemble of microstructures with a heteroscedastic GP.

Inverse problems in structure-property

(joint work w/ Tim Wildey)



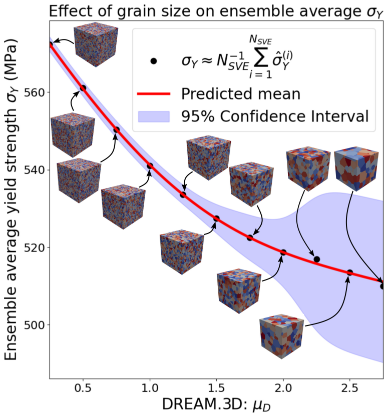
Stochastic forward vs. stochastic inverse problems in structure-property context.

- stochastic forward: given uncertain input $\lambda \to$ uncertain output $Q(\lambda)$
- stochastic inverse: given uncertain output $Q(\lambda) \to$ uncertain input λ

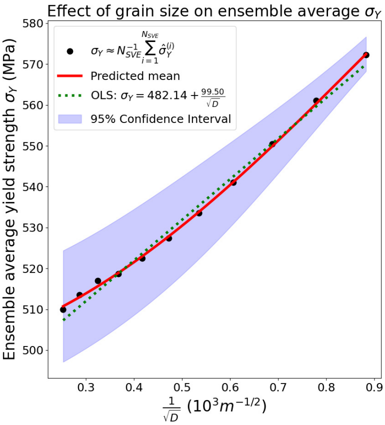
Inverse problems in structure-property

(joint work w/ Tim Wildey)

Ensemble average yield stress via Monte Carlo with different grain sizes



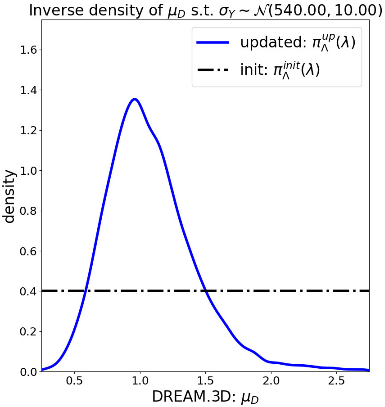
Comparison: GP (ML/UQ) and the Hall-Petch (ordinary least square)



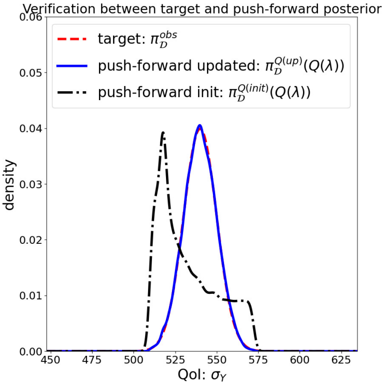
Inverse problems in structure-property

(joint work w/ Tim Wildey)

Initial density and updated density: normal case



Comparison: Distributions of materials properties



Conclusion

Takeaway message

Gaussian process is a versatile machine learning, uncertainty quantification, and optimization toolbox for ICME applications.

This talk: two parts

- theoretical / computational aspects of Gaussian process and Bayesian optimization
  - constrained (known + unknown)
  - batch-sequential and asynchronous parallel
  - multi-objective
  - multi-fidelity
  - Big Data, high-dimensional
- ICME applications
  - density functional theory: Quantum ESPRESSO
  - molecular dynamics: LAMMPS
  - kinetic Monte Carlo: SPPARKS
  - crystal plasticity finite element: DREAM.3D + DAMASK

Thank you for your time and listening.

References

Methodology:

- Anh Tran et al. (2022). "aphBO-2GP-3B: a budgeted asynchronous parallel multi-acquisition functions for constrained Bayesian optimization on high-performing computing architecture". In: *Structural and Multidisciplinary Optimization* 65.4, pp. 1–45
- Anh Tran (Aug. 2021). "Scalable^3-BO: Big Data meets HPC - A scalable asynchronous parallel high-dimensional Bayesian optimization framework on supercomputers". In: *Proceedings of the ASME 2021 IDETC/CIE*. vol. Volume 1: 41th Computers and Information in Engineering Conference. International Design Engineering Technical Conferences and Computers and Information in Engineering Conference. American Society of Mechanical Engineers
- Anh Tran et al. (Aug. 2020c). "srMO-BO-3GP: A sequential regularized multi-objective constrained Bayesian optimization for design applications". In: *Proceedings of the ASME 2020 IDETC/CIE*. vol. Volume 1: 40th Computers and Information in Engineering Conference. International Design Engineering Technical Conferences and Computers and Information in Engineering Conference. American Society of Mechanical Engineers
- Anh Tran, Tim Wildey, and Scott McCann (2020). "sMF-BO-2CoGP: A sequential multi-fidelity constrained Bayesian optimization for design applications". In: *Journal of Computing and Information Science in Engineering* 20.3, pp. 1–15
- Anh Tran, Tim Wildey, and Scott McCann (Aug. 2019). "sBF-BO-2CoGP: A sequential bi-fidelity constrained Bayesian optimization for design applications". In: *Proceedings of the ASME 2019 IDETC/CIE*. vol. Volume 1: 39th Computers and Information in Engineering Conference. International Design Engineering Technical Conferences and Computers and Information in Engineering Conference. V001T02A073. American Society of Mechanical Engineers
- Anh Tran, Minh Tran, and Yan Wang (2019). "Constrained mixed-integer Gaussian mixture Bayesian optimization and its applications in designing fractal and auxetic metamaterials". In: *Structural and Multidisciplinary Optimization* 59 (6), pp. 2131–2154
- Anh Tran et al. (2019a). "pBO-2GP-3B: A batch parallel known/unknown constrained Bayesian optimization with feasibility classification and its applications in computational fluid dynamics". In: *Computer Methods in Applied Mechanics and Engineering* 347, pp. 827–852

Applications:

- Anh Tran and Tim Wildey (2020). "Solving stochastic inverse problems for property-structure linkages using data-consistent inversion and machine learning". In: *JOM* 73, pp. 72–89
- Anh Tran et al. (2020b). "Multi-fidelity machine-learning with uncertainty quantification and Bayesian optimization for materials design: Application to ternary random alloys". In: *The Journal of Chemical Physics* 153 (7), p. 074705
- Anh Tran et al. (2020a). "An active-learning high-throughput microstructure calibration framework for process-structure linkage in materials informatics". In: *Acta Materialia* 194, pp. 80–92
- Stefano Travaglino et al. (2020). "Computational optimization study of transcatheter aortic valve leaflet design using porcine and bovine leaflets". In: *Journal of Biomechanical Engineering* 142 (1)
- Anh Tran et al. (2019b). "WearGP: A computationally efficient machine learning framework for local erosive wear predictions via nodal Gaussian processes". In: *Wear* 422, pp. 9–26
- Anh Tran, Lijuan He, and Yan Wang (2018). "An efficient first-principles saddle point searching method based on distributed kriging metamodels". In: *ASCE-ASME Journal of Risk and Uncertainty in Engineering Systems, Part B: Mechanical Engineering* 4.1, p. 011006

References I

Bauer, Matthias, Mark van der Wilk, and Carl Edward Rasmussen (2016). “Understanding probabilistic sparse Gaussian process approximations”. In: *Advances in neural information processing systems* 29, pp. 1533–1541.

Beume, Nicola et al. (2009). “On the complexity of computing the hypervolume indicator”. In: *IEEE Transactions on Evolutionary Computation* 13.5, pp. 1075–1082.

Bonilla, Edwin V, Karl Krauth, and Amir Dezfouli (2019). “Generic inference in latent Gaussian process models.”. In: *Journal of Machine Learning Research* 20.117, pp. 1–63.

Bull, Adam D (2011). “Convergence rates of efficient global optimization algorithms”. In: *Journal of Machine Learning Research* 12.Oct, pp. 2879–2904.

References II

- Burt, David R., Carl Edward Rasmussen, and Mark van der Wilk (2020). "Convergence of Sparse Variational Inference in Gaussian Processes Regression". In: *Journal of Machine Learning Research* 21.131, pp. 1–63.
- Chalupka, Krzysztof, Christopher KI Williams, and Iain Murray (2013). "A framework for evaluating approximation methods for Gaussian process regression". In: *Journal of Machine Learning Research* 14.Feb, pp. 333–350.
- Cho, Youngmin and Lawrence Saul (2009). "Kernel methods for deep learning". In: *Advances in neural information processing systems* 22.
- Constantine, Paul G (2015). *Active subspaces: Emerging ideas for dimension reduction in parameter studies*. SIAM.
- Constantine, Paul G, Eric Dow, and Qiqi Wang (2014). "Active subspace methods in theory and practice: applications to kriging surfaces". In: *SIAM Journal on Scientific Computing* 36.4, A1500–A1524.

References III

- Contal, Emile et al. (2013). "Parallel Gaussian process optimization with upper confidence bound and pure exploration". In: *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*. Springer, pp. 225–240.
- De Freitas, Nando, Alex Smola, and Masrour Zoghi (2012). "Exponential regret bounds for Gaussian process bandits with deterministic observations". In: *arXiv preprint arXiv:1206.6457*.
- Desautels, Thomas, Andreas Krause, and Joel W Burdick (2014). "Parallelizing exploration-exploitation tradeoffs in Gaussian process bandit optimization". In: *The Journal of Machine Learning Research* 15.1, pp. 3873–3923.
- Digabel, Sébastien Le and Stefan M Wild (2015). "A Taxonomy of Constraints in Simulation-Based Optimization". In: *arXiv preprint arXiv:1505.07881*.
- Frigola, Roger, Yutian Chen, and Carl Edward Rasmussen (2014). "Variational Gaussian Process State-Space Models". In: *Advances in Neural Information Processing Systems*. Vol. 27. Curran Associates, Inc.

References IV

- Huang, Deng et al. (2006). "Global optimization of stochastic black-box systems via sequential kriging meta-models". In: *Journal of Global Optimization* 34.3, pp. 441–466.
- Kennedy, Marc C and Anthony O'Hagan (2000). "Predicting the output from a complex computer code when fast approximations are available". In: *Biometrika* 87.1, pp. 1–13.
- Lawrence, Neil D (2016). "Introduction to gaussian processes". In: *MLSS* 8, p. 504. URL: inverseprobability.com/talks/slides/gp_mlss16.pdf.
- Lee, Jaehoon et al. (2018). "Deep Neural Networks as Gaussian Processes". In: *ICLR*.
- Li, Mu, James Tin-Yau Kwok, and Baoliang Lü (2010). "Making large-scale Nyström approximation possible". In: *ICML 2010-Proceedings, 27th International Conference on Machine Learning*, p. 631.

References V

- Martinsson, Per-Gunnar and Joel A Tropp (2020). "Randomized numerical linear algebra: Foundations and algorithms". In: *Acta Numerica* 29, pp. 403–572.
- Matthews, Alexander G de G et al. (2016). "On sparse variational methods and the Kullback-Leibler divergence between stochastic processes". In: *Artificial Intelligence and Statistics*. PMLR, pp. 231–239.
- Mockus, Jonas (1975). "On Bayesian methods for seeking the extremum". In: *Optimization Techniques IFIP Technical Conference*. Springer, pp. 400–404.
- — (1982). "The Bayesian approach to global optimization". In: *System Modeling and Optimization*, pp. 473–481.
- Neal, Radford M (1996). "Priors for infinite networks". In: *Bayesian Learning for Neural Networks*. Springer, pp. 29–53.
- Quiñonero-Candela, Joaquin and Lars Kai Hansen (2004). "Learning with uncertainty-Gaussian processes and relevance vector machines". In: *Technical University of Denmark, Copenhagen*.

References VI

- Quiñonero-Candela, Joaquin and Carl Edward Rasmussen (2005). "A unifying view of sparse approximate Gaussian process regression". In: *Journal of Machine Learning Research* 6.Dec, pp. 1939–1959.
- Quiñonero-Candela, Joaquin, Carl Edward Rasmussen, and Christopher KI Williams (2007). "Approximation methods for Gaussian process regression". In: *Large-scale kernel machines*, pp. 203–224.
- Rasmussen, Carl Edward (2006). *Gaussian processes in machine learning*. MIT Press.
- Srinivas, Niranjan et al. (2009). "Gaussian process optimization in the bandit setting: No regret and experimental design". In: *arXiv preprint arXiv:0912.3995*.
- Srinivas, Niranjan et al. (2012). "Information-theoretic regret bounds for Gaussian process optimization in the bandit setting". In: *IEEE Transactions on Information Theory* 58.5, pp. 3250–3265.

References VII

- Titsias, Michalis (2009a). "Variational learning of inducing variables in sparse Gaussian processes". In: *Artificial Intelligence and Statistics*, pp. 567–574.
- Titsias, Michalis K (2009b). "Variational model selection for sparse Gaussian process regression". In: *Report, University of Manchester, UK*.
- Tran, Anh (Aug. 2021). "Scalable^3-BO: Big Data meets HPC - A scalable asynchronous parallel high-dimensional Bayesian optimization framework on supercomputers". In: *Proceedings of the ASME 2021 IDETC/CIE*. Vol. Volume 1: 41th Computers and Information in Engineering Conference. International Design Engineering Technical Conferences and Computers and Information in Engineering Conference. American Society of Mechanical Engineers.
- Tran, Anh, Lijuan He, and Yan Wang (2018). "An efficient first-principles saddle point searching method based on distributed kriging metamodels". In: *ASCE-ASME Journal of Risk and Uncertainty in Engineering Systems, Part B: Mechanical Engineering* 4.1, p. 011006.

References VIII

- Tran, Anh, Minh Tran, and Yan Wang (2019). “Constrained mixed-integer Gaussian mixture Bayesian optimization and its applications in designing fractal and auxetic metamaterials”. In: *Structural and Multidisciplinary Optimization* 59 (6), pp. 2131–2154.
- Tran, Anh and Tim Wildey (2020). “Solving stochastic inverse problems for property-structure linkages using data-consistent inversion and machine learning”. In: *JOM* 73, pp. 72–89.
- Tran, Anh, Tim Wildey, and Scott McCann (Aug. 2019). “sBF-BO-2CoGP: A sequential bi-fidelity constrained Bayesian optimization for design applications”. In: *Proceedings of the ASME 2019 IDETC/CIE*. Vol. Volume 1: 39th Computers and Information in Engineering Conference. International Design Engineering Technical Conferences and Computers and Information in Engineering Conference. V001T02A073. American Society of Mechanical Engineers.
- — (2020). “sMF-BO-2CoGP: A sequential multi-fidelity constrained Bayesian optimization for design applications”. In: *Journal of Computing and Information Science in Engineering* 20.3, pp. 1–15.

References IX

- Tran, Anh et al. (2019a). "pBO-2GP-3B: A batch parallel known/unknown constrained Bayesian optimization with feasibility classification and its applications in computational fluid dynamics". In: *Computer Methods in Applied Mechanics and Engineering* 347, pp. 827–852.
- Tran, Anh et al. (2019b). "WearGP: A computationally efficient machine learning framework for local erosive wear predictions via nodal Gaussian processes". In: *Wear* 422, pp. 9–26.
- Tran, Anh et al. (2020a). "An active-learning high-throughput microstructure calibration framework for process-structure linkage in materials informatics". In: *Acta Materialia* 194, pp. 80–92.
- Tran, Anh et al. (2020b). "Multi-fidelity machine-learning with uncertainty quantification and Bayesian optimization for materials design: Application to ternary random alloys". In: *The Journal of Chemical Physics* 153 (7), p. 074705.

References X

- Tran, Anh et al. (Aug. 2020c). "srMO-BO-3GP: A sequential regularized multi-objective constrained Bayesian optimization for design applications". In: *Proceedings of the ASME 2020 IDETC/CIE*. Vol. Volume 1: 40th Computers and Information in Engineering Conference. International Design Engineering Technical Conferences and Computers and Information in Engineering Conference. American Society of Mechanical Engineers.
- Tran, Anh et al. (2022). "aphBO-2GP-3B: a budgeted asynchronous parallel multi-acquisition functions for constrained Bayesian optimization on high-performing computing architecture". In: *Structural and Multidisciplinary Optimization* 65.4, pp. 1–45.
- Travaglino, Stefano et al. (2020). "Computational optimization study of transcatheter aortic valve leaflet design using porcine and bovine leaflets". In: *Journal of Biomechanical Engineering* 142 (1).
- Vanhatalo, Jarno et al. (2012). "Bayesian modeling with Gaussian processes using the GPstuff toolbox". In: *arXiv preprint arXiv:1206.5754*.

References XI

Vanhatalo, Jarno et al. (2013). “GPstuff: Bayesian modeling with Gaussian processes”. In: *Journal of Machine Learning Research* 14.Apr, pp. 1175–1179.

Wang, Ziyu et al. (2013). “Bayesian optimization in high dimensions via random embeddings”. In: AAAI Press/International Joint Conferences on Artificial Intelligence.

Wang, Ziyu et al. (2016). “Bayesian optimization in a billion dimensions via random embeddings”. In: *Journal of Artificial Intelligence Research* 55, pp. 361–387.

Williams, Christopher and Matthias Seeger (2001). “Using the Nyström method to speed up kernel machines”. In: *Advances in neural information processing systems* 13, pp. 682–688.

Sparse variational GP

- $p(\cdot)$: true pdf
- $q(\cdot)$: approximate pdf

Assume the fully independent training conditional (FITC) Quiñonero-Candela and Rasmussen 2005; Quiñonero-Candela, Rasmussen, and Williams 2007, augment the joint model $p(\mathbf{f}_*, \mathbf{f})$ as

$$p(\mathbf{f}_*, \mathbf{f}) = \int p(\mathbf{f}_*, \mathbf{f}, \mathbf{u}) d\mathbf{u} = \int p(\mathbf{f}_*, \mathbf{f}|\mathbf{u}) p(\mathbf{u}) d\mathbf{u}, \tag{57}$$

$\mathbf{u}$: inducing variables at m locations $\mathbf{X}_\mathbf{u}$. The training and testing conditionals are

$$p(\mathbf{f}|\mathbf{u}) = \mathcal{N}(\mathbf{m} + \mathbf{K}_{\mathbf{f},\mathbf{u}} \mathbf{K}_{\mathbf{u},\mathbf{u}}^{-1} (\mathbf{u} - \mathbf{m}),\ \mathbf{K}_{\mathbf{f},\mathbf{f}} - \mathbf{Q}_{\mathbf{f},\mathbf{f}}), \tag{58}$$

and

$$p(\mathbf{f}_*|\mathbf{u}) = \mathcal{N}(\mathbf{m} + \mathbf{K}_{*,\mathbf{u}} \mathbf{K}_{\mathbf{u},\mathbf{u}}^{-1} (\mathbf{u} - \mathbf{m}),\ \mathbf{K}_{*,*} - \mathbf{Q}_{*,*}), \tag{59}$$

where

$$\mathbf{Q}_{\mathbf{a},\mathbf{b}} := \mathbf{K}_{\mathbf{a},\mathbf{u}} \mathbf{K}_{\mathbf{u},\mathbf{u}}^{-1} \mathbf{K}_{\mathbf{u},\mathbf{b}}. \tag{60}$$

The likelihood and inducing priors remain the same, i.e. $p(\mathbf{y}|\mathbf{f}) = \mathcal{N}(\mathbf{f}, \sigma^2 \mathbf{I})$, and $p(\mathbf{u}) = \mathcal{N}(\mathbf{m}, \mathbf{K}_{\mathbf{u},\mathbf{u}})$.

Sparse variational GP

FITC training prior based on the inducing priors is modified as

$$q(\mathbf{f}|\mathbf{u}) = \prod_{i=1}^{n} p(\mathbf{f}_i|\mathbf{u}) = \mathcal{N}(\mathbf{m} + \mathbf{K}_{\mathbf{f},\mathbf{u}}\mathbf{K}_{\mathbf{u},\mathbf{u}}^{-1}(\mathbf{u} - \mathbf{m}), \mathrm{Diag}[\mathbf{K}_{\mathbf{f},\mathbf{f}} - \mathbf{Q}_{\mathbf{f},\mathbf{f}}]) \qquad (61)$$

and keeping the testing prior the same

$$q(\mathbf{f}_*|\mathbf{u}) = p(\mathbf{f}_*|\mathbf{u}) = \mathcal{N}(\mathbf{m} + \mathbf{K}_{*,\mathbf{u}}\mathbf{K}_{\mathbf{u},\mathbf{u}}^{-1}(\mathbf{u} - \mathbf{m}),\ \mathbf{K}_{*,*} - \mathbf{Q}_{*,*}), \qquad (62)$$

the effective prior under the FITC assumption is

$$q(\mathbf{f}, \mathbf{f}_*) = \mathcal{N}\left(\begin{bmatrix} \mathbf{m} \\ \mathbf{m} \end{bmatrix}, \begin{bmatrix} \mathbf{Q}_{\mathbf{f},\mathbf{f}} - \mathrm{Diag}[\mathbf{Q}_{\mathbf{f},\mathbf{f}} - \mathbf{K}_{\mathbf{f},\mathbf{f}}] & \mathbf{Q}_{\mathbf{f},*} \\ \mathbf{Q}_{*,\mathbf{f}} & \mathbf{K}_{*,*} \end{bmatrix} \right), \qquad (63)$$

which implies the testing distribution as

$$\begin{array}{rcl} q(\mathbf{f}_*|\mathbf{y}) & = & \mathcal{N}(\mathbf{m} + \mathbf{Q}_{*,\mathbf{f}}(\mathbf{Q}_{\mathbf{f},\mathbf{f}} + \Lambda)^{-1}(\mathbf{y} - \mathbf{m}), \mathbf{K}_{*,*} - \mathbf{Q}_{*,\mathbf{f}}(\mathbf{Q}_{\mathbf{f},\mathbf{f}} + \Lambda)^{-1}\mathbf{Q}_{\mathbf{f},*}) \\ & = & \mathcal{N}(\mathbf{m} + \mathbf{K}_{*,\mathbf{u}}\Sigma\mathbf{K}_{\mathbf{u},\mathbf{f}}\Lambda^{-1}(\mathbf{y} - \mathbf{m}), \mathbf{K}_{*,*} - \mathbf{Q}_{*,*} + \mathbf{K}_{*,\mathbf{u}}\Sigma\mathbf{K}_{\mathbf{u},*}) \end{array}, \qquad (64)$$

where $\Sigma = [\mathbf{K}_{\mathbf{u},\mathbf{u}} + \mathbf{K}_{\mathbf{u},\mathbf{f}}\Lambda^{-1}\mathbf{K}_{\mathbf{f},\mathbf{u}}]^{-1}$ and $\Lambda = \mathrm{Diag}[\mathbf{K}_{\mathbf{f},\mathbf{f}} - \mathbf{Q}_{\mathbf{f},\mathbf{f}} + \sigma^2\mathbf{I}]$.

Sparse variational GP

The marginal likelihood conditioned on the inducing inputs is therefore

$$q(\mathbf{y}|\mathbf{X_u}) = \int\int p(\mathbf{y}|\mathbf{f})q(\mathbf{f}|\mathbf{u})p(\mathbf{u}|\mathbf{X_u})d\mathbf{u}d\mathbf{f} = \int p(\mathbf{y}|\mathbf{f})q(\mathbf{f}|\mathbf{X_u})d\mathbf{f}, \tag{65}$$

which implies the log marginal likelihood as

$$\log q(\mathbf{y}|\mathbf{X_u}) = -\frac{n}{2}\log(2\pi) - \frac{1}{2}\log|\mathbf{Q_{f,f}} + \Lambda| - \frac{1}{2}(\mathbf{y}-\mathbf{m})^\top[\mathbf{Q_{f,f}}+\Lambda]^{-1}(\mathbf{y}-\mathbf{m}), \tag{66}$$

where $\Lambda = \text{Diag}[\mathbf{K_{f,f}} - \mathbf{Q_{f,f}}] + \sigma^2\mathbf{I}$.
Cost complexity: $\mathcal{O}(nm^2)$ Li, Kwok, and Lü 2010; Williams and Seeger 2001. (Note: do not multiply matrices directly – cf. Section 14.3 Martinsson and Tropp 2020).

Variational inference

Mostly follow Titsias 2009a; Titsias 2009b and Bonilla, Krauth, and Dezfouli 2019.
Definition of conditionally independent condition:

$$p(\mathbf{f}|\mathbf{u},\mathbf{y}) = p(\mathbf{f}|\mathbf{u}), \tag{67}$$

which implies $p(\mathbf{f},\mathbf{u}|\mathbf{y}) = p(\mathbf{f}|\mathbf{u},\mathbf{y})p(\mathbf{u}|\mathbf{y}) \approx q(\mathbf{f},\mathbf{u}) = p(\mathbf{f}|\mathbf{u})q(\mathbf{u})$, where $q(\mathbf{u})$ is the approximate variational posterior. Main tool: Jensen's inequality.

$$\begin{aligned}
\log q(\mathbf{y}|\mathbf{X_u}) \quad &= \quad \log \int\int p(\mathbf{y}|\mathbf{f})q(\mathbf{f}|\mathbf{u})p(\mathbf{u}|\mathbf{X_u}) \times \tfrac{q(\mathbf{u},\mathbf{f})}{q(\mathbf{u},\mathbf{f})} d\mathbf{u}d\mathbf{f} \\
&\geq \quad \int\int q(\mathbf{u},\mathbf{f}) \log \tfrac{p(\mathbf{y}|\mathbf{f})q(\mathbf{f}|\mathbf{u})p(\mathbf{u}|\mathbf{X_u})}{q(\mathbf{u},\mathbf{f})} d\mathbf{u}d\mathbf{f} \\
&= \quad \int\int p(\mathbf{f}|\mathbf{u})q(\mathbf{u}) \log \tfrac{p(\mathbf{y}|\mathbf{f})\cancel{q(\mathbf{f}|\mathbf{u})}p(\mathbf{u}|\mathbf{X_u})}{\cancel{p(\mathbf{f}|\mathbf{u})}q(\mathbf{u})} d\mathbf{u}d\mathbf{f} \\
&= \quad \int q(\mathbf{u}) \left\{ \int p(\mathbf{f}|\mathbf{u}) \log p(\mathbf{y}|\mathbf{f}) d\mathbf{f} + \log \tfrac{p(\mathbf{u}|\mathbf{X_u})}{q(\mathbf{u})} \right\} d\mathbf{u} \\
&= \quad \int q(\mathbf{u}) \left\{ \log G(\mathbf{u},\mathbf{y}) + \log \tfrac{p(\mathbf{u}|\mathbf{X_u})}{q(\mathbf{u})} \right\} d\mathbf{u} \\
&= \quad \int q(\mathbf{u}) \left\{ \log \tfrac{G(\mathbf{u},\mathbf{y})p(\mathbf{u}|\mathbf{X_u})}{q(\mathbf{u})} \right\} d\mathbf{u} := \mathcal{F}_V(\mathbf{X_u},\mathbf{u}),
\end{aligned} \tag{68}$$

$$\begin{aligned}
\log G(\mathbf{u},\mathbf{y}) \quad &= \quad \int p(\mathbf{f}|\mathbf{u}) \log p(\mathbf{y}|\mathbf{f}) d\mathbf{f} \\
&= \quad \int p(\mathbf{f}|\mathbf{u}) \left\{ -\tfrac{n}{2} \log(2\pi\sigma^2) - \tfrac{1}{2\sigma^2} \operatorname{Tr}\left[\mathbf{y}\mathbf{y}^\top - 2\mathbf{y}\mathbf{f}^\top + \mathbf{f}\mathbf{f}^\top \right] \right\} d\mathbf{f} \\
&= \quad -\tfrac{n}{2} \log(2\pi\sigma^2) - \tfrac{1}{2\sigma^2} \operatorname{Tr}\left[\mathbf{y}\mathbf{y}^\top - 2\mathbf{y}\alpha^\top + \alpha\alpha^\top + \mathbf{Q}_{\mathbf{f},\mathbf{f}} - \mathbf{K}_{\mathbf{f},\mathbf{f}} \right] \\
&= \quad \mathcal{N}(\mathbf{y}|\alpha, \sigma^2\mathbf{I}) - \tfrac{1}{2\sigma^2} \operatorname{Tr}[\operatorname{Cov}(\alpha)],
\end{aligned} \tag{69}$$

Variational inference

where $\alpha = \mathbf{f}|\mathbf{u}$, with

$$\mathbb{E}[\alpha] = \mathbb{E}[\mathbf{f}|\mathbf{u}] = \mathbf{m} + \mathbf{K}_{\mathbf{f},\mathbf{u}}\mathbf{K}_{\mathbf{u},\mathbf{u}}^{-1}(\mathbf{u} - \mathbf{m}) \tag{70}$$

and

$$\mathsf{Cov}[\alpha] = \mathsf{Cov}[\mathbf{f}|\mathbf{u}] = \mathbf{K}_{\mathbf{f},\mathbf{f}} - \mathbf{Q}_{\mathbf{f},\mathbf{f}} = \mathbf{K}_{\mathbf{f},\mathbf{f}} - \mathbf{K}_{\mathbf{f},\mathbf{u}}\mathbf{K}_{\mathbf{u},\mathbf{u}}^{-1}\mathbf{K}_{\mathbf{u},\mathbf{f}}. \tag{71}$$

Reverse Jensen's inequality to maximize the variational evidence lower bound $\mathcal{F}_V(\mathbf{X}_\mathbf{u}, \mathbf{u})$ w.r.t. $q(\mathbf{u})$

$$\begin{aligned}
\mathcal{F}_V(\mathbf{X}_\mathbf{u}, \mathbf{u}) &= \int q(\mathbf{u}) \left\{ \log \frac{G(\mathbf{u},\mathbf{y}) p(\mathbf{u}|\mathbf{X}_\mathbf{u})}{q(\mathbf{u})} \right\} d\mathbf{u} \\
&\leq \int \log G(\mathbf{u}, \mathbf{y}) p(\mathbf{u}|\mathbf{X}_\mathbf{u}) d\mathbf{u} \\
&= \log[\mathcal{N}(\mathbf{y}|\mathbf{m}, \sigma^2\mathbf{I} + \mathbf{Q}_{\mathbf{f},\mathbf{f}})] - \frac{1}{2\sigma^2} \mathsf{Tr}\left[\mathbf{K}_{\mathbf{f},\mathbf{f}} - \mathbf{K}_{\mathbf{f},\mathbf{u}}\mathbf{K}_{\mathbf{u},\mathbf{u}}^{-1}\mathbf{K}_{\mathbf{u},\mathbf{f}}\right] =: \mathcal{F}_V(\mathbf{X}_\mathbf{u})
\end{aligned} \tag{72}$$

Train sparse GP by maximizing $\mathcal{F}_V(\mathbf{X}_\mathbf{u})$. See also Vanhatalo et al. 2012, 2013, Bauer, Wilk, and Rasmussen 2016; Burt, Rasmussen, and Wilk 2020, Matthews et al. 2016.