Optimal Experimental Design for Large-Scale Bayesian Inverse Problems via Multi-PDE-Constrained Optimization

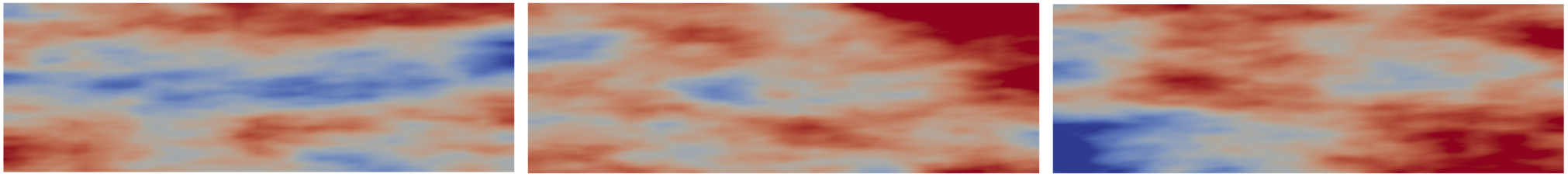
Umberto Villa[1] and Omar Ghattas[2]

[1]Washington University, Dept of Electrical & Systems Engineering

[2]The Oden Institute for Computational Engineering and Sciences

[2]Departments of Geological Sciences & Mechanical Engineering

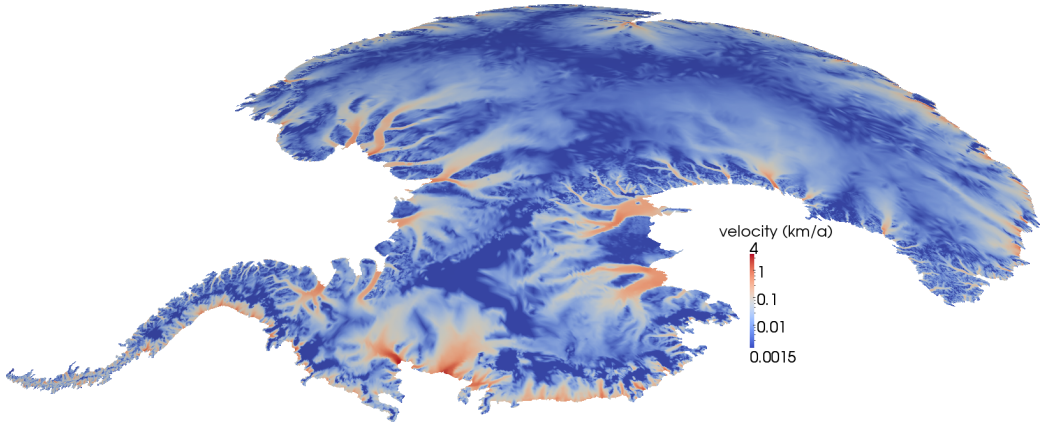
[2]The University of Texas at Austin



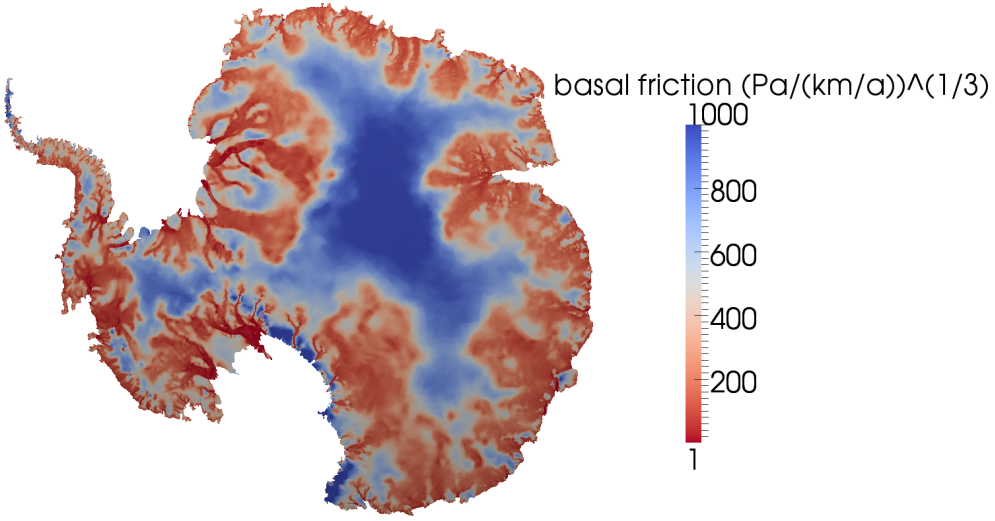
Motivation

- PDE models (and other) often depend on some **uncertain input parameters (fields)**, whose values cannot be directly measured
- **Bayesian inversion** seeks to estimate the parameters and **quantify the uncertainty** in the inversion, given observations of the outputs
- But for many problems, **data acquisition is very expensive**; how do we choose optimal experiments?
  - what is measured, where, and how often; which experiments to run, etc.
- **Optimal experimental design** (OED): find **optimal data acquisition strategy** to minimize the uncertainty in the inferred parameters or maximize the information gained about the model
- **Significant challenges**:
  - the parameter to be inferred is often a **spatially correlated field**
  - leads to **high dimensional** parameter space after discretization
  - the forward model is often **expensive to solve** (PDEs)
  - OED problem may have **combinatorial complexity**

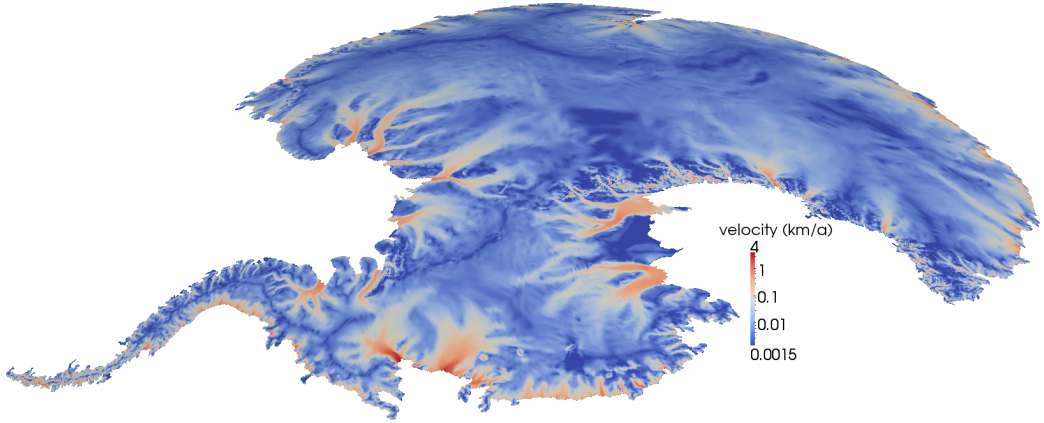
Bayesian inversion for basal friction field in Antarctica



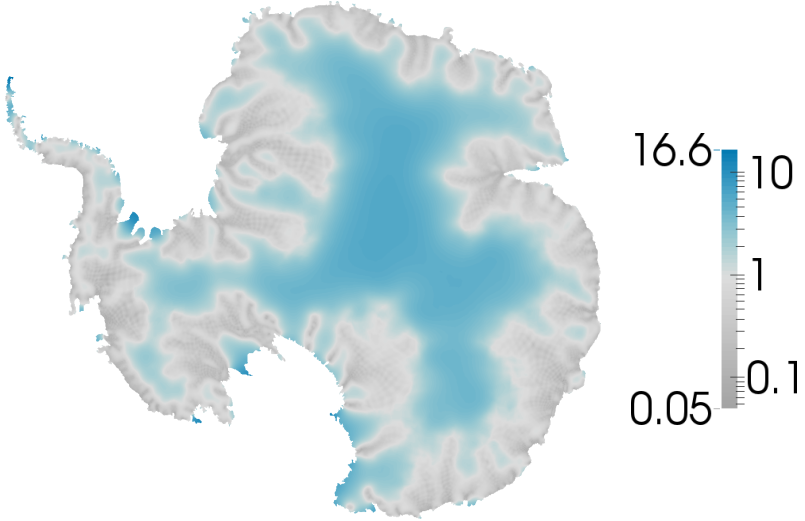
InSAR-based ice surface velocity observations



Inferred mean of basal friction field



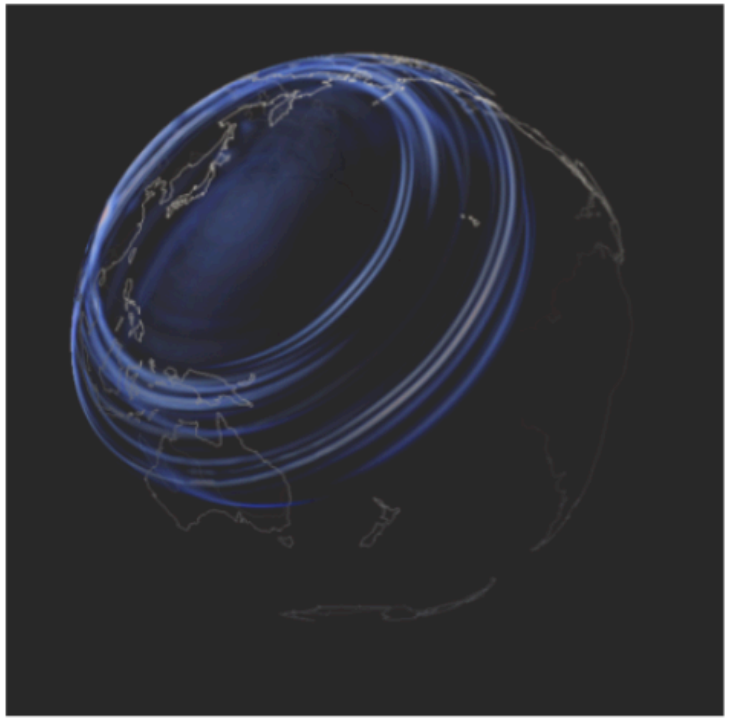
Reconstructed ice surface velocity field (based on inferred mean of basal friction field)



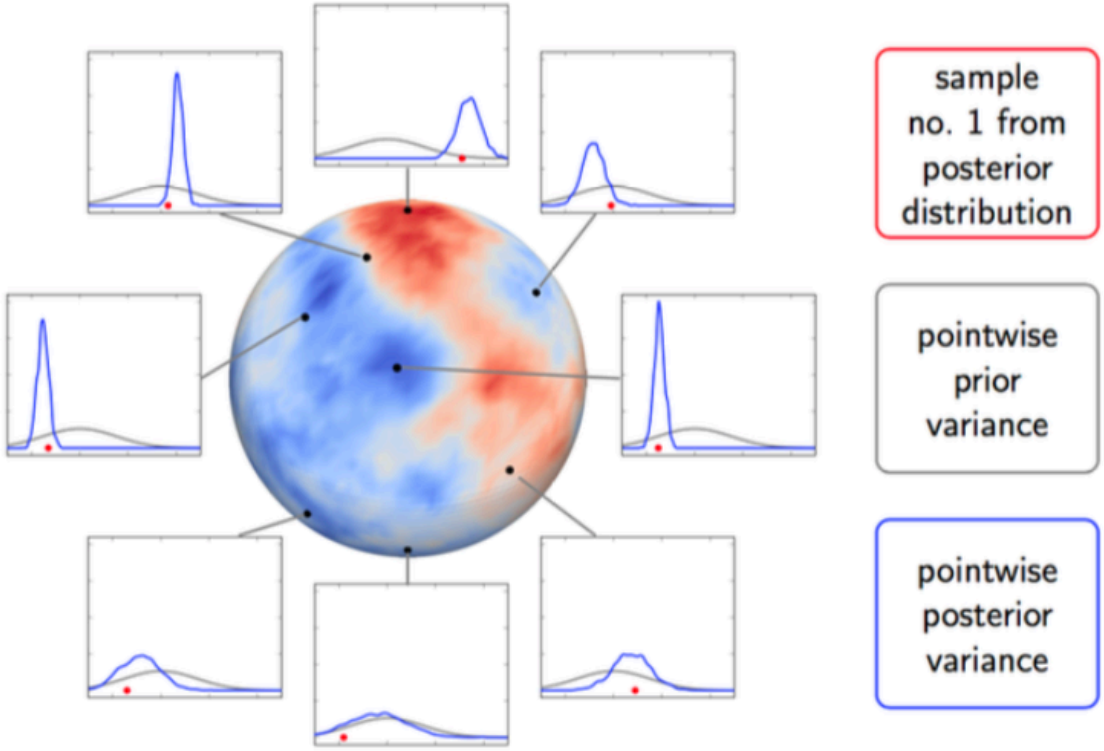
Inferred uncertainty in basal friction field (standard deviation of Laplace approx of posterior of log basal friction)

Joint work with Tobin Isaac (Georgia Tech), Noemi Petra (UC-Merced), Georg Stadler (NYU)

Bayesian global seismic inversion



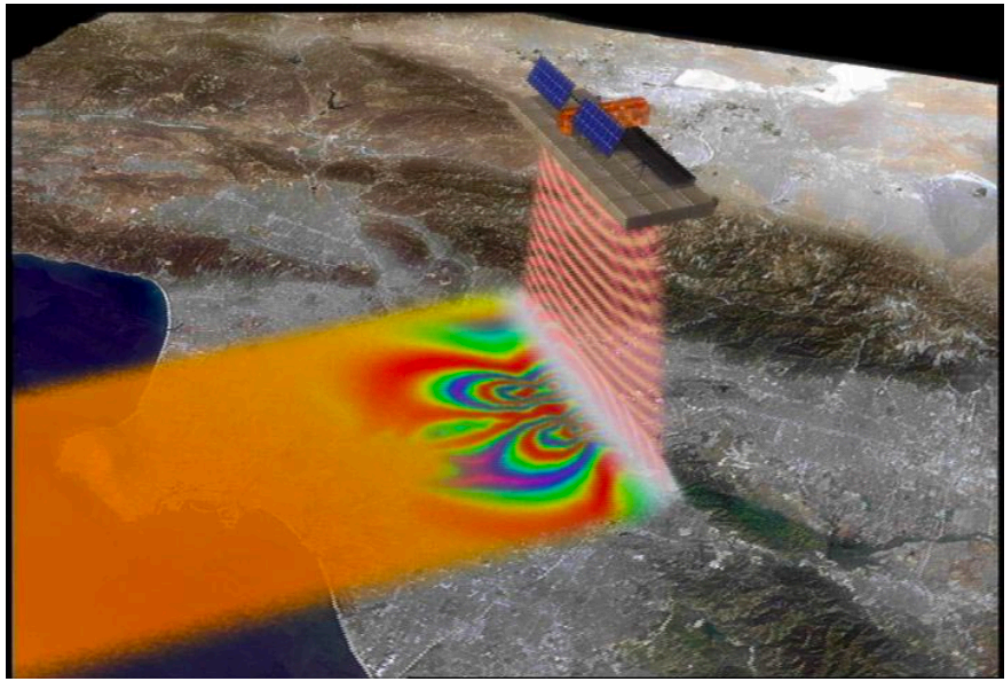
Parallel adaptive DG wave propagation

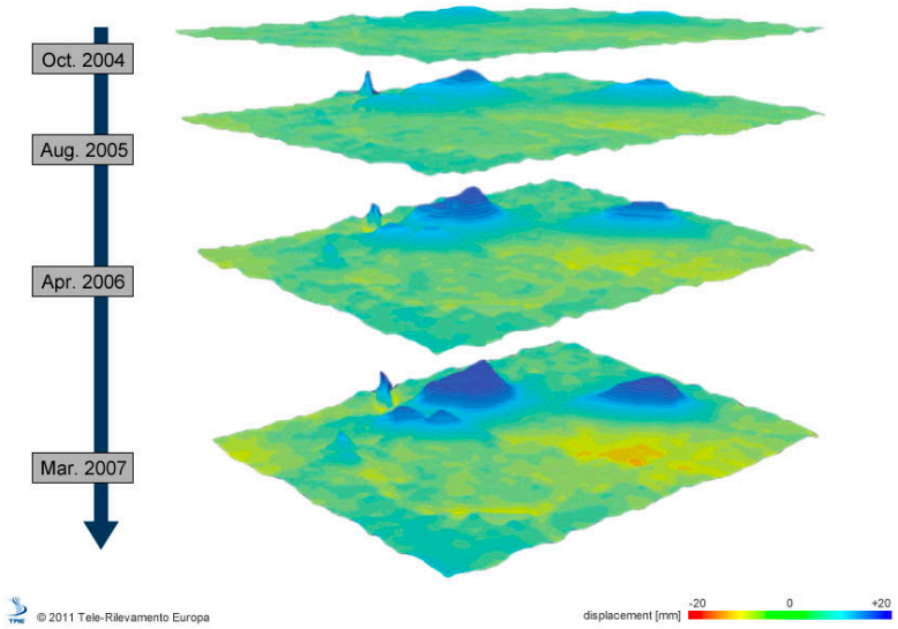


Prior and posterior seismic velocity marginals

Joint work with Tan Bui-Thanh (UT Austin), Carsten Burstedde (Bonn), Georg Stadler (NYU), Lucas Wilcox (Naval Postgraduate School)

Bayesian poroelastic inversion for management of subsurface fluid injection



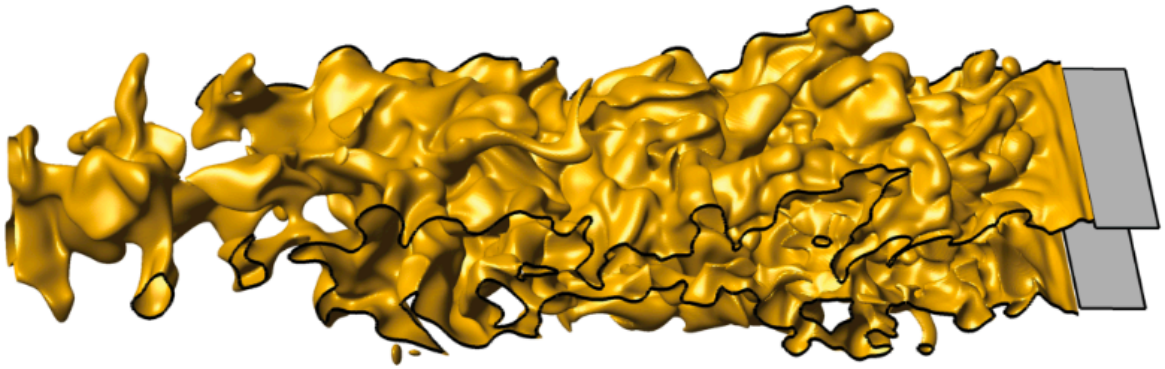


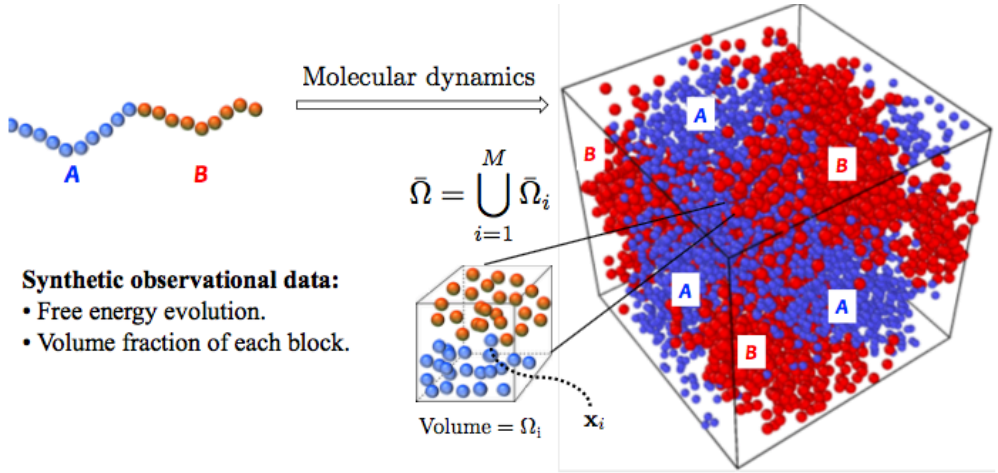
Use observations (InSAR, GPS) of surface deformation induced by either wastewater injection or production of reservoir or aquifer, in addition to well pressure measurements, to infer subsurface permeability and elastic properties. Forward predict and then ultimately optimize injection or extraction processes to avoid induced seismicity or preserve reservoir integrity.

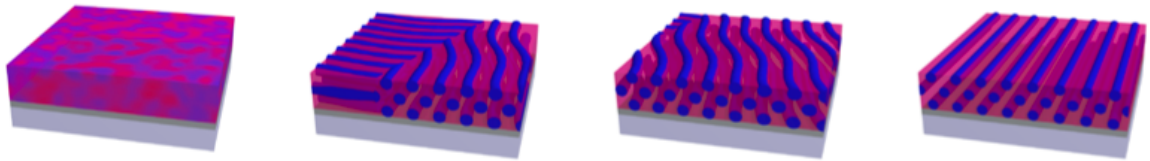
Joint work with Amal Alghamdi, Ann Chen, Marc Hesse, Georg Stadler

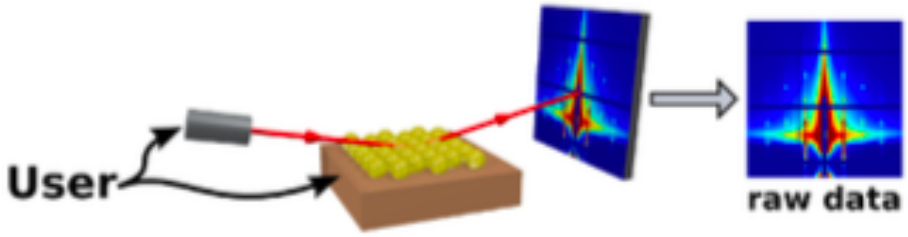
Calibration of reduced models using DNS “data”: Which DNS experiments to run?

Turbulence/reduced chemical mechanism model for turbulent combustion, calibrated using DNS data (w/ Bob Moser, Todd Oliver, Umberto Villa, Peng Chen, UT Austin)









Inference of a mesoscopic nonlocal Cahn-Hilliard model parameters of block copolymers from ”DNS” molecular dynamics simulations and X-ray scattering experiments (w/Frank Alexander (Brookhaven), Tinsley Oden, Karen Willcox (UT Austin), Danial Faghihi (Buffalo))

Goals

Devise **scalable & efficient** algorithms for **OED problems** governed by **Bayesian inverse problems** that are in turn governed by **forward PDE** models with uncertain **parameter fields** [*an outer loop around an outer loop*]

- "Scalable" means the cost, measured in **# of PDE solves**, is **independent** of the **parameter** (m), **data** (y), and **experimental design** (w) dimensions
- Efficient means the constant is "small" (difficult to achieve)

Optimality Criteria

- **Bayes risk** seeks to minimize **expected error between true & inferred parameter**

$$\int_{\mathcal{Y}} \frac{1}{2} \|m_{\mathrm{MAP}}(y,w) - m_y\|^2 \pi(y) dy \mapsto \min$$

- **A-optimal** seeks to minimize the **expectation of the trace of posterior covariance** (i.e. the *average variance* of m)

$$\int_{\mathcal{Y}} \mathrm{trace}\left[\mathcal{C}_{\mathrm{post}}(y,w)\right] \pi(y) dy \mapsto \min$$

- **Expected information Gain** (EIG) seeks to maximize **the expected Kullback-Leibler divergence from the prior to the posterior** :

$$\int_{\mathcal{Y}} D_{KL}(\mu_{\mathrm{post}}(\cdot|y;w) \| \mu_{\mathrm{pr}}) \, \pi(y) \, dy \mapsto \max$$

Note: EIG reduces to **D-optimal** for linear inverse problems

A. Alexanderian, P. Gloor, and O. Ghattas, *On Bayesian A- and D-optimal experimental designs in infinite dimensions*, Bayesian Analysis, 2016.

A. Alexanderian, N. Petra, G. Stadler, and O. Ghattas, *A fast and scalable method for A-optimal design of experiments for infinite-dimensional Bayesian nonlinear inverse problems*, SISC, 2016

Expected Information Gain (EIG)

- EIG is the **expected value of the KL divergence** from prior to posterior over all possible realizations of the data

$$\begin{aligned} EIG &:= \mathtt{E}_{\mathcal{Y}}\left[D_{KL}(\mu_{\text{post}}(\cdot|y)\|\mu_{\text{pr}})\right] \\ &= \int_{\mathcal{Y}} D_{KL}(\mu_{\text{post}}(\cdot|y)\|\mu_{\text{pr}})\,\pi(y)\,dy \\ &= \int_{\mathcal{M}}\int_{\mathcal{Y}} D_{KL}(\mu_{\text{post}}(\cdot|y)\|\mu_{\text{pr}})\pi_{\text{like}}(y|m)dy\,d\mu_{\text{pr}}(m) \end{aligned}$$

where

$$D_{KL}(\mu_{\text{post}}(\cdot|y)\|\mu_{\text{pr}}) := \int_{\mathcal{M}} \ln\left(\frac{d\mu_{\text{post}}(m|y)}{d\mu_{\text{pr}}(m)}\right) d\mu_{\text{post}}(m|y)$$

- **Sample average approximation** (SAA) gives

$$EIG \approx \frac{1}{N_s}\sum_{i=1}^{N_s} D_{KL}(\mu_{\text{post}}(\cdot|y_i)\|\mu_{\text{pr}})$$

where

$$y_i = \mathcal{B}u(m_i) + \varepsilon_i \text{ with } m_i \sim \mu_{\text{pr}}, \varepsilon_i \sim \mu_{\text{n}} \text{ i.i.d.}$$

Evaluation of EIG

- **Double Loop Monte Carlo** (DLMC):
  - Numerical underflow issues for *concentrated* posteriors
  - Extremely expensive to compute
- **Monte Carlo with Laplace Approximation** (MCLA):
  - Laplace approx introduces bias, which decays with the number of observations
  - It exploits closed form expression for KL distance between Gaussian distributions
  - Efficient evaluation of the function & its gradient
- **Double Loop Monte Carlo with Importance Sampling** (DLMC-IS):
  - Bias can be made arbitrarily small by increasing number of inner samples
  - Uses Laplace approx as biasing distribution to evaluate evidence
  - Extremely expensive to compute

J. Beck, B. Mansour Dia, L. F. Espath, Q. Long, and R. Tempone, *Fast Bayesian experimental design: Laplace-based importance sampling for the expected information gain*, CMAME, 2018

Laplace Approximation of the Posterior

- Computing the KL divergence is **intractable for large-scale problems**
- **Laplace approximation is a fast and scalable approach** for:
  - Approximates posterior as Gaussian with **mean given by maximum a posteriori** (MAP) point, and **covariance by inverse of Hessian of negative log posterior**
  - Combined **with low-rank approximation of Hessian of negative log likelihood**, renders posterior estimate for large-scale inverse problems tractable
  - Permits **tractable computation of KL divergence** (since between two Gaussians)
- Our approach exploits **three approximations**:
  - **Sample average approximation** (SAA) of expectation over the data
  - **Laplace approximation** to the true posterior
  - **Spectrum truncation & randomized eigensolver** to efficiently represent Hessian of negative log likelihood

U. Villa, O. Ghattas. *Scalable Laplace approximation-based optimal experimental design for infinite dimensional Bayesian inverse problems*, in preparation, 2019

Bayesian Inverse Problem

- **Forward problem**

$$r(u, m) = 0$$

where:

  - u is state
  - $m \sim \mathcal{N}(m_{\mathrm{pr}}, \mathcal{C}_{\mathrm{pr}})$ is parameter to be inferred (random field)

- **Observation process** ($\mathcal{B}$ is observation operator)

$$y = \mathcal{B}u(m) + \varepsilon, \quad \varepsilon \sim \mathcal{N}(0, \Gamma_n)$$

- **Bayes theorem**

$$d\mu_{\mathrm{post}}(m|y) \propto \pi_{\mathrm{like}}(y|m)\, d\mu_{\mathrm{pr}}(m)$$

where

$$\pi_{\mathrm{like}}(y|m) \propto \exp\left(-\frac{1}{2}\|\mathcal{B}u(m) - y\|^2_{\Gamma_n^{-1}}\right)$$

Laplace Approximation

- For nonlinear inverse problems, the **posterior is non-Gaussian**
- **Laplace approximation** of posterior reads

$$m \sim \nu := \mathcal{N}\left(m_{\mathrm{MAP}}, \mathcal{C}_{\mathrm{post}}\right)$$

where m_{MAP} is the **maximum a posteriori point** (minimizes negative log posterior)

$$m_{\mathrm{MAP}} := \underset{m \in \mathcal{M}}{\operatorname{argmin}} \frac{1}{2}\|\mathcal{B}u(m) - y\|^2_{\Gamma_n^{-1}} + \frac{1}{2}\|m - m_{\mathrm{pr}}\|^2_{\mathcal{C}_{\mathrm{pr}}^{-1}}$$

and C_{post} is the inverse of the **(Gauss Newton) Hessian** at m_{MAP}

$$\mathcal{C}_{\mathrm{post}} = \left(\mathcal{H}_m + \mathcal{C}_{\mathrm{pr}}^{-1}\right)^{-1}$$

- **Closed form expression** for KL distance between Gaussian distributions

$$D_{KL}(\nu\|\mu_{\mathrm{pr}}) = \frac{1}{2}\Big[\operatorname{logdet}\left(\mathcal{C}_{\mathrm{pr}}^{\frac{1}{2}}\mathcal{C}_{\mathrm{post}}^{-1}\mathcal{C}_{\mathrm{pr}}^{\frac{1}{2}}\right) - \operatorname{trace}\left(\mathcal{C}_{\mathrm{post}}^{\frac{1}{2}}(\mathcal{C}_{\mathrm{post}}^{-1} - \mathcal{C}_{\mathrm{pr}}^{-1})\mathcal{C}_{\mathrm{post}}^{\frac{1}{2}}\right) + \|m_{\mathrm{MAP}} - m_{\mathrm{pr}}\|^2_{\mathcal{C}_{\mathrm{pr}}^{-1}}\Big]$$

Laplace Approximation

- For nonlinear inverse problems, the **posterior is non-Gaussian**
- **Laplace approximation** of posterior reads

$$m \sim \nu := \mathcal{N}\left(m_{\mathrm{MAP}}, \mathcal{C}_{\mathrm{post}}\right)$$

where m_{MAP} is the **maximum a posteriori point** (maximizes negative log posterior)

$$m_{\mathrm{MAP}} := \underset{m \in \mathcal{M}}{\operatorname{argmin}} \frac{1}{2}\|\mathcal{B}u(m) - y\|^2_{\Gamma_n^{-1}} + \frac{1}{2}\|m - m_{\mathrm{pr}}\|^2_{\mathcal{C}_{\mathrm{pr}}^{-1}}$$

and C_{post} is the inverse of the **(Gauss Newton) Hessian** at m_{MAP}

$$\mathcal{C}_{\mathrm{post}} = \left(\mathcal{H}_m + \mathcal{C}_{\mathrm{pr}}^{-1}\right)^{-1}$$

- **Closed form expression** for KL distance between Gaussian distributions

$$D_{KL}(\nu\|\mu_{\mathrm{pr}}) = \frac{1}{2}\Big[\operatorname{logdet}\left(\mathcal{I} + \widetilde{\mathcal{H}}_m\right) - \operatorname{trace}\left(\mathcal{C}_{\mathrm{post}}^{\frac{1}{2}}\mathcal{H}_m\mathcal{C}_{\mathrm{post}}^{\frac{1}{2}}\right) + \|m_{\mathrm{MAP}} - m_{\mathrm{pr}}\|^2_{\mathcal{C}_{\mathrm{pr}}^{-1}}\Big]$$

where $\widetilde{\mathcal{H}}_m = \mathcal{C}_{\mathrm{pr}}^{\frac{1}{2}}\mathcal{H}_m\mathcal{C}_{\mathrm{pr}}^{\frac{1}{2}}$ is **prior-preconditioned data misfit Hessian**

Finding the MAP point scalably

$$m_{\mathrm{MAP}} := \underset{m \in \mathcal{M}}{\operatorname{argmin}} \frac{1}{2}\|\mathcal{B}u(m) - y\|^2_{\Gamma_n^{-1}} + \frac{1}{2}\|m - m_{\mathrm{pr}}\|^2_{\mathcal{C}_{\mathrm{pr}}^{-1}}$$

- Equivalent to solving a deterministic inverse problem:
  - Minimize a regularized data misfit
- Can be done at cost (measured in forw/adj PDE solves) that is **independent of the parameter & data dimensions**; depends only on **"information dimension"** (how many modes in parameter space are informed by the data)
- But must be done properly:
  - Hessian-free inexact Newton-conjugate gradient solver (terminate CG when negative curvature direction encountered & avoid oversolving)
  - Hessian action in CG directions formed by solving a pair of second-order linearized forward/adjoint PDEs
  - Precondition by prior covariance to yield compact perturbation of identity → CG converges in dimension-independent steps
  - Globalize via continuation and trust region or line search

Example: Scalability of MAP Point Computation for Ice Sheet Inverse Problem

#sdof	#pdof	#N	#CG	avgCG	#Stokes
95,796	10,371	42	2718	65	7031
233,834	25,295	39	2342	60	6440
848,850	91,787	39	2577	66	6856
3,372,707	364,649	39	2211	57	6193
22,570,303	1,456,225	40	1923	48	5376

- **sdof** : number of degrees of freedom for the state variables
- **pdof** : number of degrees of freedom for the inversion parameter field
- **N** : number of Newton iterations for the inverse problem
- **totCG** : total number of CG iterations for the inverse problem
- **avgCG** : average number of CG iterations per Newton iteration
- **linfwd** : total number of linear(ized) state solves
- **Stokes** : total number of linear(ized) Stokes solves
- refinements obtained by decreasing max area of an element by a factor of 4
- convergence = reduction of gradient by factor of 10^5
- cost (measured by # of forward solves) is independent of parameter dimension and data dimension (CG performance consequence of preconditioned Hessian operator of form compact perturbation of identity)

T. Isaac, N. Petra, G. Stadler, O. Ghattas, *Scalable and efficient algorithms for the propagation of uncertainty from data through inference to prediction for large-scale problems, with application to flow of the Antarctic ice sheet*, Journal of Computational Physics, 2015.

Spectrum Truncation for D_{KL} Computation

- Computing logdet and trace of **large implicitly defined operators** is **intractable using naïve algorithms**
- Exploit fact that prior-preconditioned Hessian misfit $\widetilde{\mathcal{H}}_m$ is **compact**:
  - Prior covariance operator $\mathcal{C}_{\text{pr}}$ is compact
  - Hessian of data misfit $\mathcal{H}_m$ is usually compact (observables are sensitive to a limited number of modes in parameter space)
  - Make **low rank approximation** of $\widetilde{\mathcal{H}}_m$
- Let $\lambda_j \; (j = 1, \ldots, r)$ be the *r* dominant eigenvalues of $\widetilde{\mathcal{H}}_m$, then

$$\text{logdet}\left(\mathcal{I} + \widetilde{\mathcal{H}}_m\right) \approx \sum_{j=1}^{r} \ln(1+\lambda_j), \;\; \text{trace}\left(\mathcal{C}_{\text{post}}^{\frac{1}{2}} \mathcal{H}_m \mathcal{C}_{\text{post}}^{\frac{1}{2}}\right) \approx \sum_{j=1}^{r} \frac{\lambda_j}{1+\lambda_j}$$

- **Randomized eigensolvers** can extract accurate approximation of dominant eigenvalues of $\widetilde{\mathcal{H}}_m$ at a **cost** (measured in PDE solves)
  - that **is independent of the parameter and data dimensions**
  - and depends only on the numerical rank *r*
- Randomized eigensolver requires only **Hessian actions**; can be formed **operator-free** at cost of one incremental forward/adjoint PDE solve

Randomized Eigensolver

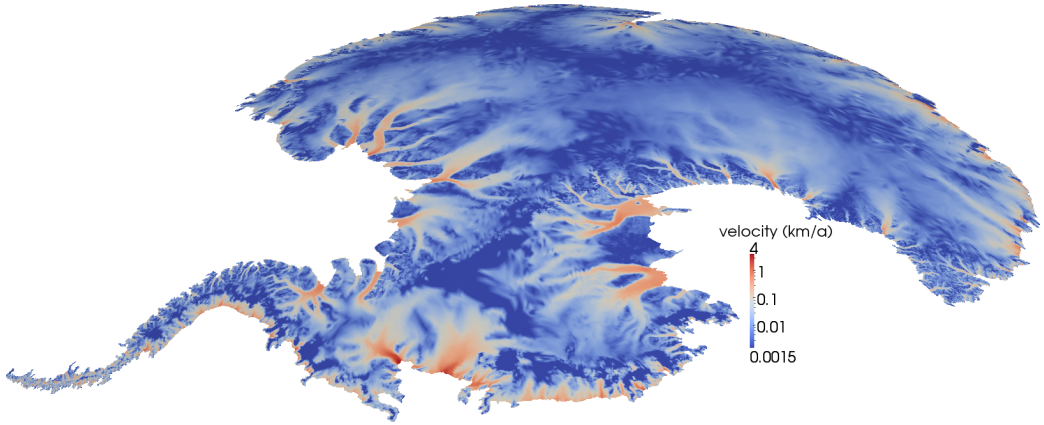
Double pass algorithm:

1. Generate i.i.d. Gaussian matrix $\boldsymbol{\Omega} \in \mathbb{R}^{n \times r}$
2. Form $\mathbf{Y} = \widetilde{\mathbf{H}}\boldsymbol{\Omega}$,
3. Compute $\mathbf{Q} =$ orthonormal basis for $\mathbf{Y}$
4. Compute $\mathbf{T} = \mathbf{Q}^T \widetilde{\mathbf{H}} \mathbf{Q}$
5. Decompose $\mathbf{T} = \mathbf{Z} \boldsymbol{\Lambda} \mathbf{Z}^T$.
6. Low rank approximation $\widetilde{\mathbf{H}} = \mathbf{V} \boldsymbol{\Lambda} \mathbf{V}^T$, where $\mathbf{V} \in \mathbb{R}^{n \times r} := \mathbf{Q}\mathbf{Z}$

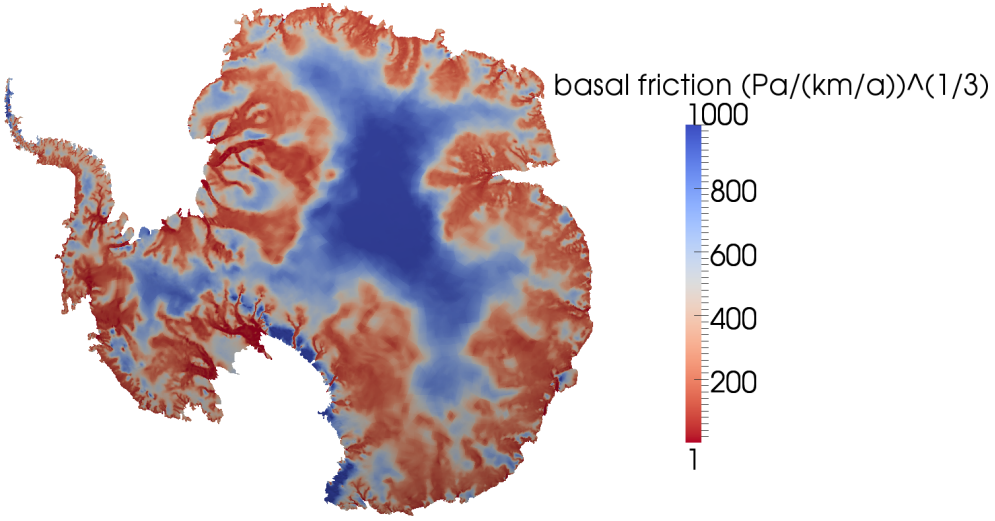
Main cost: 2r incremental forward solves, 2r incremental adjoint solves, 4r prior square root applications

- N. Halko, P. G. Martinsson, and J. A. Tropp. *Finding structure with randomness: Probabilistic algorithms for constructing approximate matrix decompositions*. SIAM Review, 2011.
- A. K. Saibaba, A. Alexanderian, and I. C. F. Ipsen, *Randomized matrix-free trace and log-determinant estimators*, Numerische Mathematik, 2017.

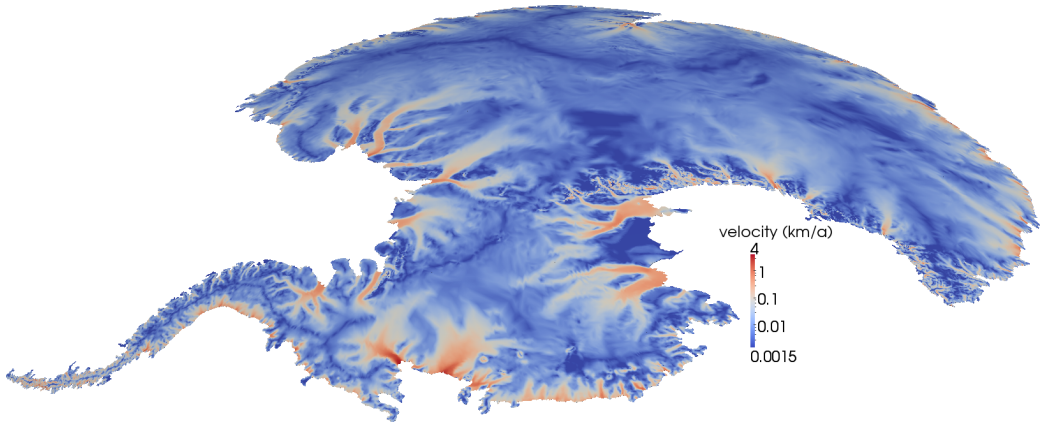
Example: Inversion for basal friction field in Antarctica



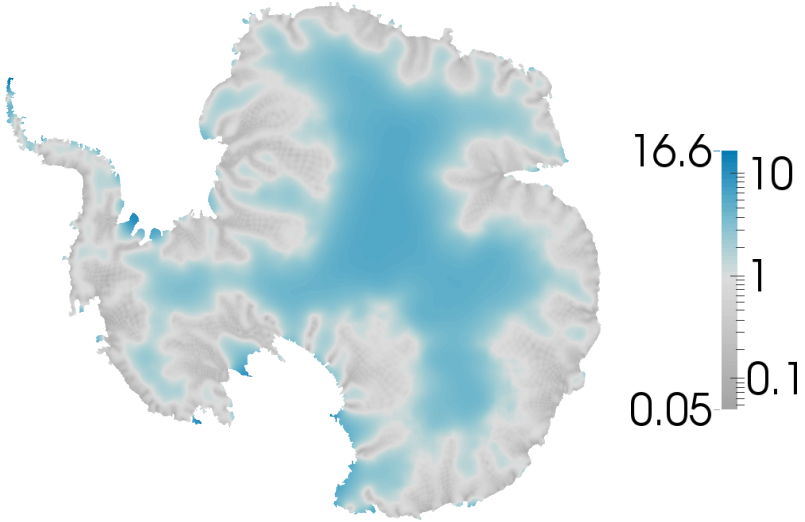
InSAR-based ice surface velocity observations



Inferred mean of basal friction field



Reconstructed ice surface velocity field (based on inferred mean of basal friction field)



Inferred uncertainty in basal friction field (standard deviation of Laplace approx of posterior of log basal friction)

Joint work with Tobin Isaac (Georgia Tech), Noemi Petra (UC-Merced), Georg Stadler (NYU)

Nonlinear Stokes Ice Sheet Flow

Balance of linear momentum, mass, and energy

$$-\nabla \cdot \overbrace{[2\eta(\boldsymbol{u},\theta)\,\dot{\boldsymbol{\varepsilon}}_{\boldsymbol{u}} - \boldsymbol{I}p]}^{\boldsymbol{\sigma}_{\boldsymbol{u}}} = \rho \boldsymbol{g} \qquad [\dot{\boldsymbol{\varepsilon}}_{\boldsymbol{u}} = \tfrac{1}{2}(\nabla \boldsymbol{u} + \nabla \boldsymbol{u}^T)]$$

$$\nabla \cdot \boldsymbol{u} = 0$$

$$\rho c \boldsymbol{u} \cdot \nabla \theta - \nabla \cdot (K \nabla \theta) = \eta(\boldsymbol{u},\theta)\ \mathrm{tr}(\dot{\boldsymbol{\varepsilon}}_{\boldsymbol{u}}^2)$$

Constitutive relation

$$\eta(\boldsymbol{u},\theta) = \frac{1}{2}\left\{A_0 \exp\left(-\frac{Q}{R\theta}\right)\right\}^{-\frac{1}{n}} \left(\tfrac{1}{2}\mathrm{tr}(\dot{\boldsymbol{\varepsilon}}_{\boldsymbol{u}}^2)\right)^{\frac{1-n}{2n}}$$

Boundary conditions

unknown parameter fields

$$\boldsymbol{\sigma}_{\boldsymbol{u}}\boldsymbol{n} = \boldsymbol{0} \qquad \theta = \theta_s \quad \text{on } \Gamma_t$$

$$\boldsymbol{u}\cdot\boldsymbol{n} = \frac{M}{\rho L_i} \qquad \boldsymbol{T} := \boldsymbol{I} - \boldsymbol{n}\otimes\boldsymbol{n},\ \boldsymbol{T}\boldsymbol{\sigma}_{\boldsymbol{u}}\boldsymbol{n} + \beta \boldsymbol{T}\boldsymbol{u} = \boldsymbol{0} \quad \text{on } \Gamma_b$$

$$K\nabla\theta\cdot\boldsymbol{n} = G - M + \beta \boldsymbol{T}\boldsymbol{u}\cdot\boldsymbol{T}\boldsymbol{u}, \qquad \theta \le \theta_m,\ M \ge 0,\ M(\theta-\theta_m) = 0 \quad \text{on } \Gamma_b$$

Forward & Adjoint Stokes Equations

u and p satisfy the *forward (nonlinear) Stokes equations*

$$\begin{aligned} \nabla \cdot \boldsymbol{u} &= 0 && \text{in } \Omega \\ -\nabla \cdot [\eta(\boldsymbol{u})(\nabla \boldsymbol{u} + \nabla \boldsymbol{u}^T) - \boldsymbol{I}p] &= \rho \boldsymbol{g} && \text{in } \Omega \\ \boldsymbol{\sigma}_{\boldsymbol{u}} \boldsymbol{n} &= \boldsymbol{0} && \text{on } \Gamma_t \\ \boldsymbol{u} \cdot \boldsymbol{n} = 0, \quad (\boldsymbol{\sigma}_{\boldsymbol{u}} \boldsymbol{n})_\Gamma + \exp(\beta) \boldsymbol{u}_\Gamma &= \boldsymbol{0} && \text{on } \Gamma_b \end{aligned}$$

v and q satisfy the *adjoint Stokes equations*

$$\begin{aligned} \nabla \cdot \boldsymbol{v} &= 0 && \text{in } \Omega \\ -\nabla \cdot \boldsymbol{\sigma}_{\boldsymbol{v}} &= \boldsymbol{0} && \text{in } \Omega \\ \boldsymbol{\sigma}_{\boldsymbol{v}} \boldsymbol{n} &= -\mathcal{B}^*(\mathcal{B}\boldsymbol{u} - \boldsymbol{d}_{\text{obs}}) && \text{on } \Gamma_t \\ \boldsymbol{v} \cdot \boldsymbol{n} = 0, \quad (\boldsymbol{\sigma}_{\boldsymbol{v}} \boldsymbol{n})_\Gamma + \exp(\beta) \boldsymbol{v}_\Gamma &= \boldsymbol{0} && \text{on } \Gamma_b \end{aligned}$$

where the adjoint stress $\boldsymbol{\sigma}_{\boldsymbol{v}}$ is

$$\boldsymbol{\sigma}_{\boldsymbol{v}} := 2\eta(\boldsymbol{u}) \left(\mathsf{I} + \frac{1-n}{n} \frac{\dot{\boldsymbol{\varepsilon}}_{\boldsymbol{u}} \otimes \dot{\boldsymbol{\varepsilon}}_{\boldsymbol{u}}}{\dot{\boldsymbol{\varepsilon}}_{\boldsymbol{u}} \cdot \dot{\boldsymbol{\varepsilon}}_{\boldsymbol{u}}} \right) \dot{\boldsymbol{\varepsilon}}_{\boldsymbol{v}} - \boldsymbol{I} q$$

Incremental Forward Stokes Equation

Action of Hessian operator in direction $\hat{\beta}$ evaluated at β

$$\mathcal{H}(\beta)\hat{\beta} := \exp(\beta)\left(\hat{\beta}\,\boldsymbol{u}_\Gamma \cdot \boldsymbol{v}_\Gamma + \hat{\boldsymbol{u}}_\Gamma \cdot \boldsymbol{v}_\Gamma + \boldsymbol{u}_\Gamma \cdot \hat{\boldsymbol{v}}_\Gamma\right) + \alpha\Delta_\Gamma\hat{\beta}$$

where $\hat{u}$ and $\hat{p}$ satisfy the *incremental forward equations*

$$\begin{aligned}
\nabla \cdot \hat{\boldsymbol{u}} &= 0 && \text{in } \Omega \\
-\nabla \cdot \boldsymbol{\sigma}_{\hat{\boldsymbol{u}}} &= \boldsymbol{0} && \text{in } \Omega \\
\boldsymbol{\sigma}_{\hat{\boldsymbol{u}}}\boldsymbol{n} &= \boldsymbol{0} && \text{on } \Gamma_t \\
\hat{\boldsymbol{u}} \cdot \boldsymbol{n} = 0, \quad (\boldsymbol{\sigma}_{\hat{\boldsymbol{u}}}\boldsymbol{n})_\Gamma + \exp(\beta)\hat{\boldsymbol{u}}_\Gamma &= -\hat{\beta}\exp(\beta)\boldsymbol{u}_\Gamma && \text{on } \Gamma_b
\end{aligned}$$

with $\boldsymbol{\sigma}_{\hat{\boldsymbol{u}}} := 2\eta(\boldsymbol{u})\left(\mathsf{I} + \frac{1-n}{n}\frac{\dot{\boldsymbol{\varepsilon}}_{\boldsymbol{u}} \otimes \dot{\boldsymbol{\varepsilon}}_{\boldsymbol{u}}}{\dot{\boldsymbol{\varepsilon}}_{\boldsymbol{u}} \cdot \dot{\boldsymbol{\varepsilon}}_{\boldsymbol{u}}}\right)\dot{\boldsymbol{\varepsilon}}_{\hat{\boldsymbol{u}}} - \boldsymbol{I}\hat{p}$

N. Petra, H. Zhu, G. Stadler, T.J.R. Hughes, O. Ghattas, *A scalable adjoint-based inexact Newton method for inversion of basal sliding and rheology parameters in a nonlinear Stokes ice sheet model*, Journal of Glaciology, 58(211):889903, 2012.

Incremental Adjoint Stokes Equation

Action of Hessian operator in direction $\hat{\beta}$ evaluated at β

$$\mathcal{H}(\beta)\hat{\beta} := \exp(\beta)\left(\hat{\beta}\,\boldsymbol{u}_\Gamma \cdot \boldsymbol{v}_\Gamma + \hat{\boldsymbol{u}}_\Gamma \cdot \boldsymbol{v}_\Gamma + \boldsymbol{u}_\Gamma \cdot \hat{\boldsymbol{v}}_\Gamma\right) + \alpha\Delta_\Gamma\hat{\beta}$$

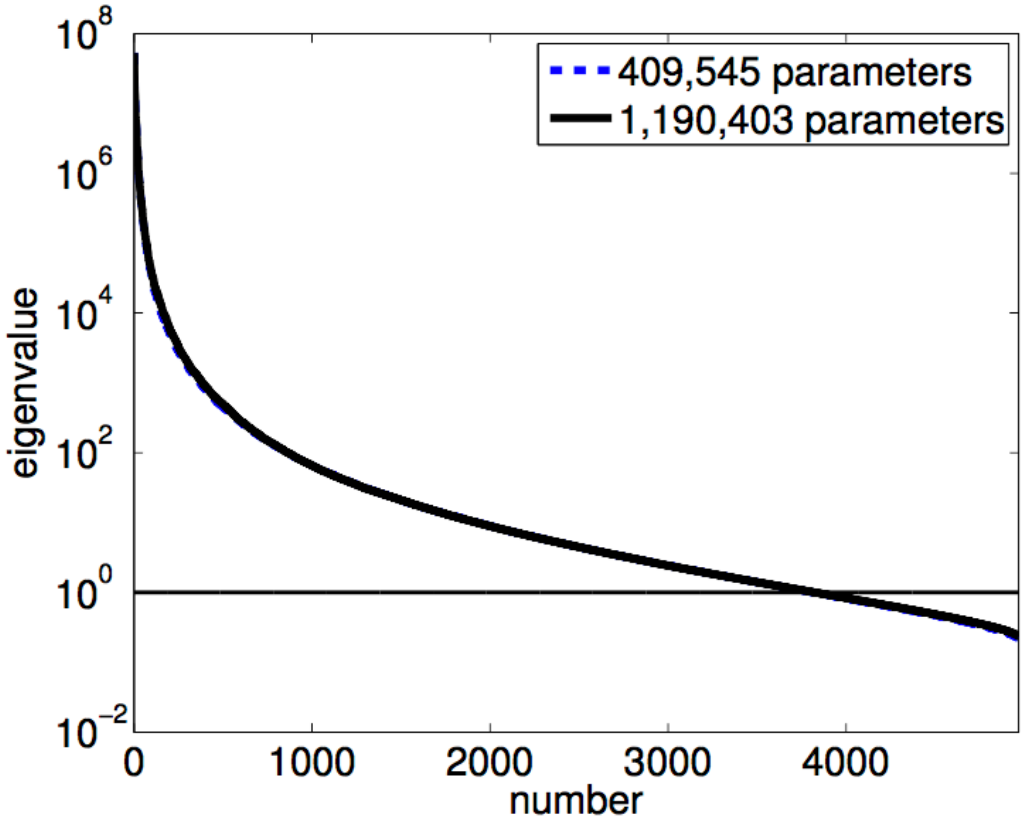
where $\tilde{v}$, $\tilde{q}$ satisfy the *incremental adjoint equations*

$$\begin{aligned}
\boldsymbol{\nabla}\cdot\hat{\boldsymbol{v}} &= 0 && \text{in } \Omega\\
-\boldsymbol{\nabla}\cdot\boldsymbol{\sigma}_{\hat{\boldsymbol{v}}} &= -\boldsymbol{\nabla}\cdot\boldsymbol{\tau}_{\hat{\boldsymbol{u}}} && \text{in } \Omega\\
\boldsymbol{\sigma}_{\hat{\boldsymbol{v}}}\boldsymbol{n} &= -\mathcal{B}^*\mathcal{B}\hat{\boldsymbol{u}} - \boldsymbol{\tau}_{\hat{\boldsymbol{u}}}\boldsymbol{n} && \text{on } \Gamma_t\\
\hat{\boldsymbol{v}}\cdot\boldsymbol{n} = 0,\quad (\boldsymbol{\sigma}_{\hat{\boldsymbol{v}}}\boldsymbol{n})_\Gamma + \exp(\beta)\hat{\boldsymbol{v}}_\Gamma &= -(\boldsymbol{\tau}_{\hat{\boldsymbol{u}}}\boldsymbol{n})_\Gamma && \text{on } \Gamma_b
\end{aligned}$$

with $\boldsymbol{\sigma}_{\hat{\boldsymbol{v}}} := 2\eta(\boldsymbol{u})\left(\mathsf{I} + \frac{1-n}{n}\frac{\dot{\boldsymbol{\varepsilon}}_{\boldsymbol{u}}\otimes\dot{\boldsymbol{\varepsilon}}_{\boldsymbol{u}}}{\dot{\boldsymbol{\varepsilon}}_{\boldsymbol{u}}\cdot\dot{\boldsymbol{\varepsilon}}_{\boldsymbol{u}}}\right)\dot{\boldsymbol{\varepsilon}}_{\hat{\boldsymbol{v}}} - \boldsymbol{I}\hat{q}$, and $\boldsymbol{\tau}_{\hat{\boldsymbol{u}}} = 2\eta(\boldsymbol{u})\Psi\dot{\boldsymbol{\varepsilon}}_{\hat{\boldsymbol{u}}}$, where

$$\Psi = (1 + \frac{1-n}{n}\dot{\boldsymbol{\varepsilon}}_{\boldsymbol{u}}\cdot\dot{\boldsymbol{\varepsilon}}_{\boldsymbol{u}})\mathsf{I} + \frac{1-n}{n}\left[\frac{\dot{\boldsymbol{\varepsilon}}_{\boldsymbol{u}}\otimes\dot{\boldsymbol{\varepsilon}}_{\boldsymbol{u}}}{\dot{\boldsymbol{\varepsilon}}_{\boldsymbol{u}}\cdot\dot{\boldsymbol{\varepsilon}}_{\boldsymbol{u}}} + 2\frac{\dot{\boldsymbol{\varepsilon}}_{\boldsymbol{u}}\otimes\dot{\boldsymbol{\varepsilon}}_{\boldsymbol{v}}}{\dot{\boldsymbol{\varepsilon}}_{\boldsymbol{u}}\cdot\dot{\boldsymbol{\varepsilon}}_{\boldsymbol{u}}} + \frac{1-3n}{n}\frac{\dot{\boldsymbol{\varepsilon}}_{\boldsymbol{u}}\otimes\dot{\boldsymbol{\varepsilon}}_{\boldsymbol{u}}}{(\dot{\boldsymbol{\varepsilon}}_{\boldsymbol{u}}\cdot\dot{\boldsymbol{\varepsilon}}_{\boldsymbol{u}})^2}\right]$$

Spectrum of the prior-preconditioned data misfit Hessian



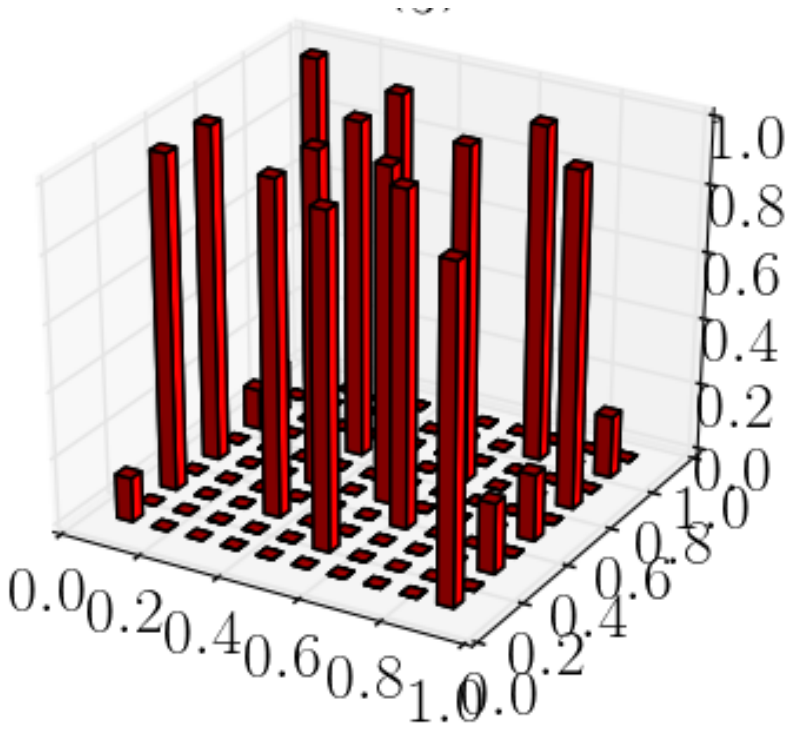
- Spectrum of $\widetilde{\mathcal{H}}_m = \mathcal{C}_{\mathrm{pr}}^{\frac{1}{2}} \mathcal{H}_m \mathcal{C}_{\mathrm{pr}}^{\frac{1}{2}}$ for Antarctica inverse problem with 410K and 1.19M basal sliding parameters (observed to decay like i^{-3})
- ~5000 dominant modes, **indep. of parameter & data dimensions**
- intrinsic problem dimension depends on **information content of data**

Design Space & Weighted Likelihood

$$\text{design} := \left\{ \begin{array}{c} \mathbf{x}_1, \ldots, \mathbf{x}_{n_d} \\ w_1, \ldots, w_{n_d} \end{array} \right\}$$

- $\mathbf{x}_i$: candidate sensor locations
- w_i: weights
- *Ideally:* $w_i \in \{0, 1\}$
  Combinatorial complexity
- *Relax:* $0 \leq w_i \leq 1$
- *Sparsity-inducing penalization:*

$$\mathcal{P}_\epsilon(w) \sim \left\{ \begin{array}{ll} \|w\|_1 & \epsilon = 1 \\ \|w\|_0 & \epsilon \to 0 \end{array} \right.$$



Weighted likelihood function

$$\pi_{\text{like}}(y|m; w) \propto \exp\left(-\frac{1}{2}\|\mathcal{B}u(m) - y\|^2_{\Gamma_W^{-1}}\right), \quad \Gamma_W = \text{diag}\left(w^{-\frac{1}{2}}\right) \Gamma_n \, \text{diag}\left(w^{-\frac{1}{2}}\right)$$

A. Alexanderian, N. Petra, G. Stadler, and O. Ghattas, *A-optimal design of experiments for infinite dimensional Bayesian linear inverse problems with regularized l_0-sparsification*, SISC, 2014

OED Problem with EIG-MCLA

EIG-MCLA Sparsity-inducing penalization

$$\min_{w\in\mathcal{W}} -\frac{1}{2\,N_s}\sum_{i=1}^{N_s}\left[\sum_{j=1}^{r_i}\left(\ln(1+\lambda_{ij})-\frac{\lambda_{ij}}{1+\lambda_{ij}}\right)+\|m_i-m_{\rm pr}\|^2_{\mathcal{C}_{\rm pr}^{-1}}\right]+\beta\mathcal{P}(w)$$

Subject to:

- **Inverse problem**

$$\begin{cases} m_i := \underset{m\in\mathcal{M}}{\operatorname{argmin}} \frac{1}{2}\|\mathcal{B}\,u_i-y_i\|^2_{\Gamma_W^{-1}}+\frac{1}{2}\|m-m_{\rm pr}\|^2_{\mathcal{C}_{\rm pr}^{-1}} \\ \text{subject to:} \\ \quad \langle r(u_i,m),\tilde{p}\rangle = 0 \quad \forall\tilde{p}\in\mathcal{U} \end{cases} \qquad i=1,\dots,N_s$$

with $y_i=\mathcal{B}u(m_i)+\varepsilon_i$ with $m_i\sim\mu_{\rm pr},\varepsilon_i\sim\mu_{\rm n}$ i.i.d.

- **Eigenproblem**

$$\begin{cases} \langle\tilde{a}_{ij},\mathcal{H}_m(m_i,w)\,a_{ij}\rangle=\lambda_{ij}\langle\tilde{a}_{ij},\mathcal{C}_{\rm pr}^{-1}a_{ij}\rangle \quad \tilde{a}_{ij}\in\mathcal{U} \\ \|a_{ij}\|^2_{\mathcal{C}_{\rm pr}^{-1}}=1 \end{cases} \qquad j=1,\dots,r_i;\,i=1,\dots,N_s$$

OED Problem with EIG-MCLA

$$\min_{w\in\mathcal{W}} -\frac{1}{2\,N_s}\sum_{i=1}^{N_s}\left[\sum_{j=1}^{r_i}\left(\ln(1+\lambda_{ij})-\frac{\lambda_{ij}}{1+\lambda_{ij}}\right)+\|m_i-m_{\rm pr}\|^2_{\mathcal{C}_{\rm pr}^{-1}}\right]+\beta\mathcal{P}(w)$$

Subject to:

- **Inverse problem**

$$\begin{cases}\langle r(u_i,m_i),\tilde{p}_i\rangle=0 & \forall\tilde{p}_i\in\mathcal{U}\\ \langle\partial_u r|_\star[\tilde{u}_i],p_i\rangle=-\langle\mathcal{B}\tilde{u}_i,\Gamma_W^{-1}(\mathcal{B}\,u_i-y_i)\rangle & \forall\tilde{u}_i\in\mathcal{U}\\ \langle\tilde{m}_i,\mathcal{C}_{\rm pr}^{-1}(m_i-m_{\rm pr})\rangle=-\langle\partial_m r|_\star[\tilde{m}_i],p_i\rangle & \forall\tilde{m}_i\in\mathcal{M}\end{cases}\qquad i=1,\dots,N_s$$

- **Eigenproblem**

$$\begin{cases}\langle\partial_u r|_\star[v_{ij}],\tilde{q}_{ij}\rangle=-\langle\partial_m r|_\star[a_{ij}],\tilde{q}_{ij}\rangle & \forall\tilde{q}_{ij}\in\mathcal{U}\\ \langle\partial_u r|_\star[\tilde{v}_{ij}],q_{ij}\rangle=-\langle\mathcal{B}\tilde{v}_{ij},\Gamma_W^{-1}\mathcal{B}\rangle v_{ij}\rangle & \forall\tilde{v}_{ij}\in\mathcal{U}\\ \langle\partial_m r|_\star[\tilde{m}_{ij}],q_{ij}\rangle=\lambda_{ij}\langle\tilde{m}_{ij},\mathcal{C}_{\rm pr}^{-1}a_{ij}\rangle & \forall\tilde{m}_{ij}\in\mathcal{M}\\ \langle a_{ij},\mathcal{C}_{\rm pr}^{-1}a_{ij}\rangle\tilde{\lambda}_{ij}=\tilde{\lambda}_{ij} & \forall\tilde{\lambda}_{ij}\in\mathbb{R}\\ i=1,\dots,N_s,\,j=1,\dots,r_i.\end{cases}$$

Cost Per Optimization Iteration (Measured in Forward/Adjoint PDE Solves)

- **Objective functional**:
  - **MAP Point**: Solve an inverse problem using **inexact Newton-CG** algorithm for each SAA sample.
  - **Eigenvalues of Hessian at MAP point**: Compute the low-rank factorization of the prior-preconditioned Hessian of the data misfit using **randomized eigensolver** for each SAA sample
- **Gradient (needed for quasi-Newton optimization)**:
  - **Adjoint of Hessian eigenproblem:** Adjoint eigenvalues are ~ free
  - **Adjoint of MAP point necessary conditions:** Solve a linear system involving **the Hessian of the inverse problem** for each SAA sample; spectral decomposition of Hessian is available, so cost is ~ free

Newton method, variational adjoint-based methods, & randomized eigensolvers allow for **efficient computation of objective functional and its gradient at a cost** (measured in number of PDE solves) **that is asymptotically independent of parameter dimension and the data/sensor dimension**.

1D Elliptic PDE Coefficient Inversion

- **Prior distribution:** $m \sim \mathcal{N}(0, \mathcal{C}_{\text{pr}})$

$$\mathcal{C}_{\text{pr}} = \sigma^2 \left(\kappa^{-2}\mathcal{I} - \Delta\right)^{-1}$$

- **Forward model**

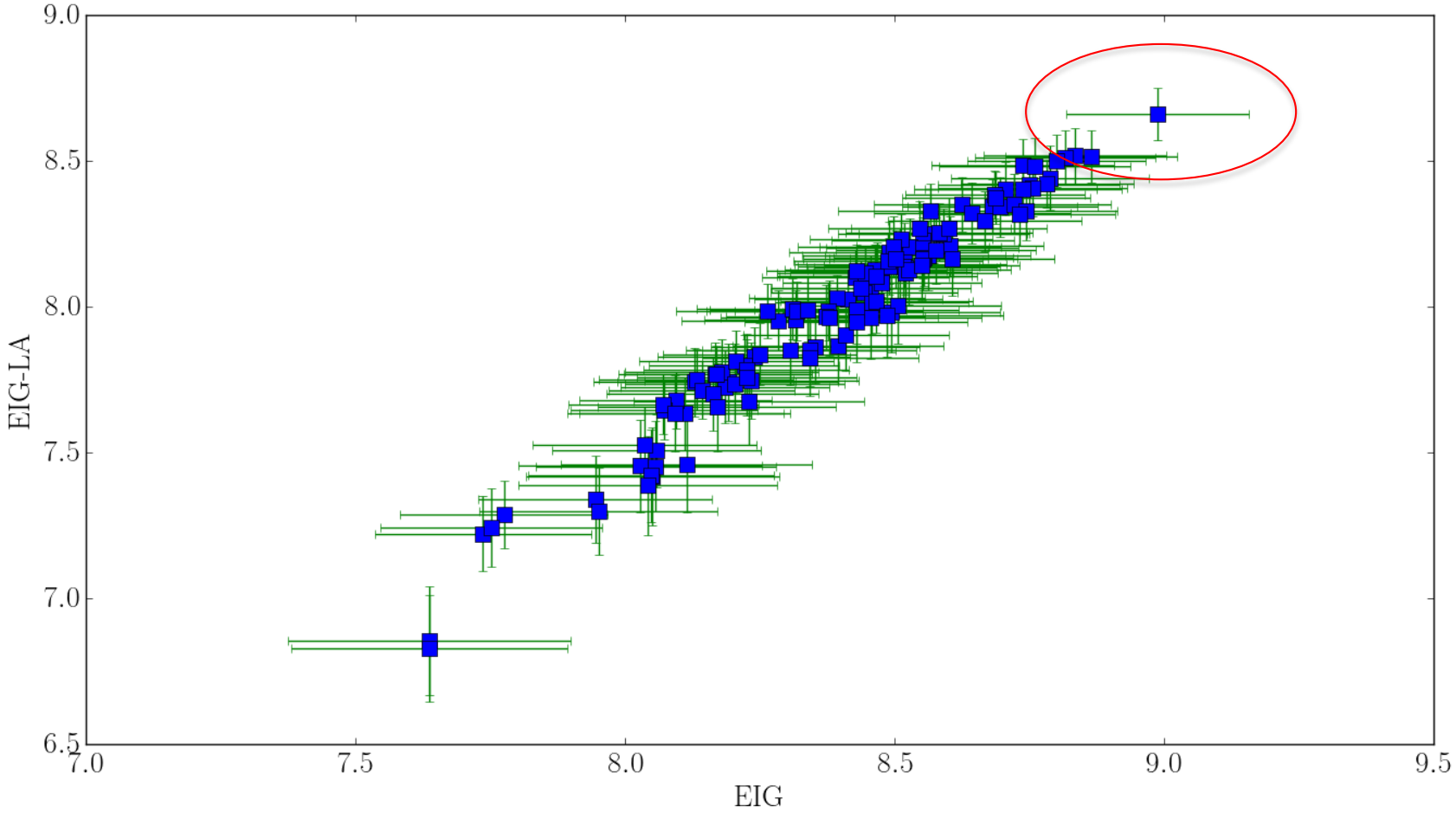
$$\begin{cases} -\frac{d}{dx}\left(e^m \frac{du}{dx}\right) = 0, & x \in (0,1) \\ u(0) = 0,\ u(1) = 1 \end{cases}$$

- **Design space**
  9 equispaced candidate sensor locations
  Find 4 optimal locations

- **Small problem:** Only 126 different designs to compare using brute force approach

U. Villa, N. Petra, and O. Ghattas, *hIPPYlib: An Extensible Software Framework for Large-Scale Deterministic and Linearized Bayesian Inverse Problems*
http://hippylib.github.io

Accuracy of Laplace Approximation to EIG



Blue squares: All 126 possible designs (9 choose 4)

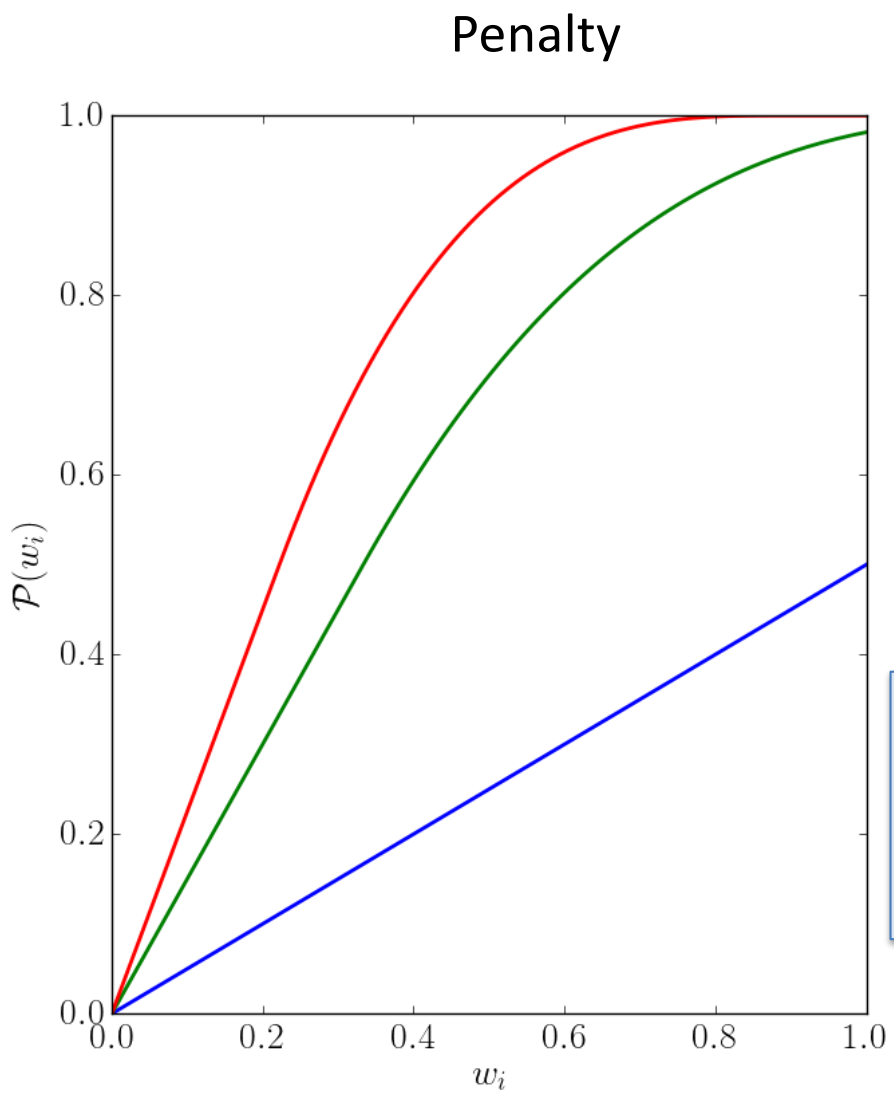
EIG: Double loop Monte Carlo (w/IS) with 1000 inner + 1000 outer samples

EIG-LA: Monte Carlo-Laplace Approximation with 1000 outer samples

Error bars: 3 standard deviations (99.7% CI)

Correlation coefficient: 0.98

Sparsifying Penalty



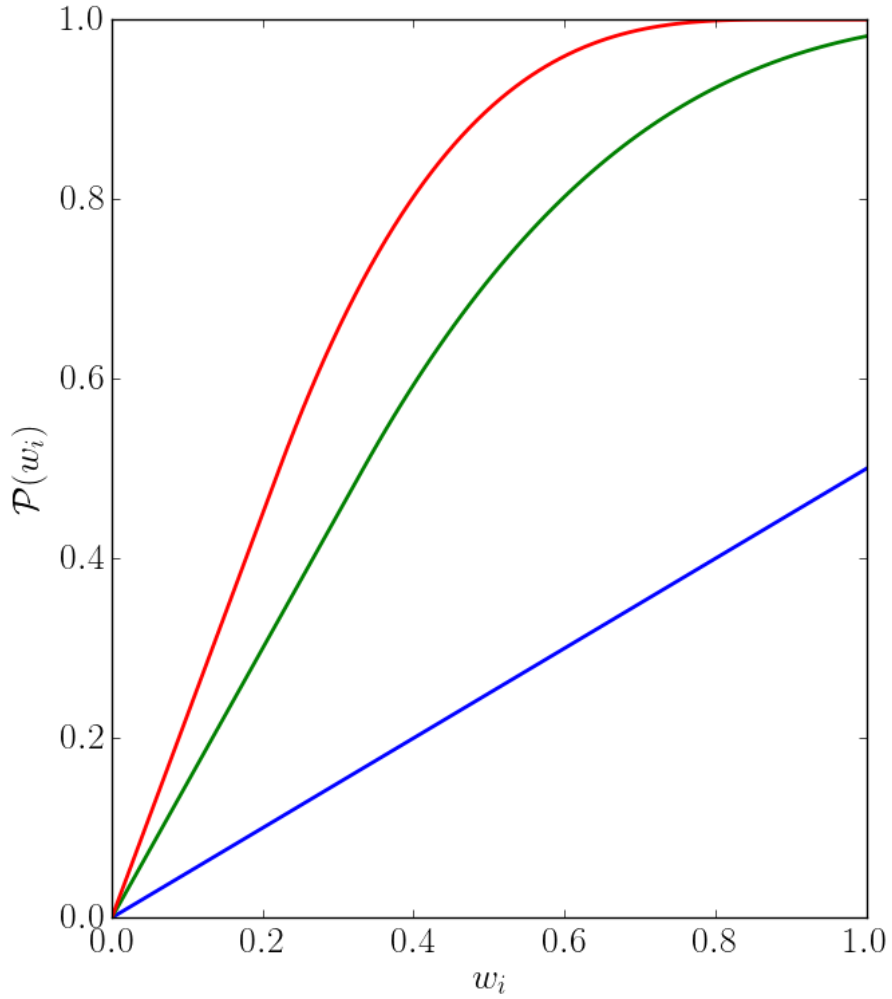
$$\mathcal{P}_\varepsilon(w_i) = \begin{cases} \frac{w_i}{\varepsilon}, & 0 \le w_i \le \frac{1}{2}\varepsilon \\ p_\varepsilon(w_i), & \frac{1}{2}\varepsilon \le w_i \le 2\varepsilon \\ 1, & 2\varepsilon \le w_i \le 1 \end{cases}$$

$$p_\varepsilon : \text{cubic polynomial s.t. } \mathcal{P}_\varepsilon \in C^1$$

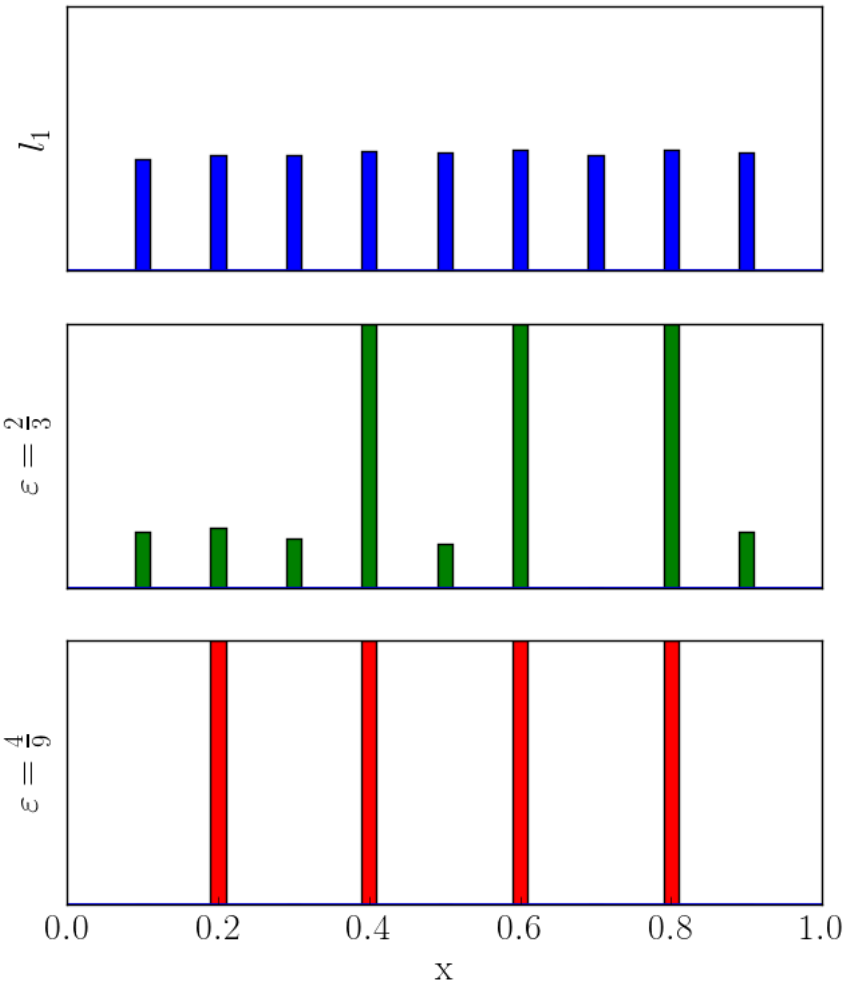
A. Alexanderian, N. Petra, G. Stadler, O. Ghattas, *A-optimal design of experiments for infinite-dimensional Bayesian linear inverse problems with regularized l_0-sparsification,* SISC, 2014

Sparsifying Penalty

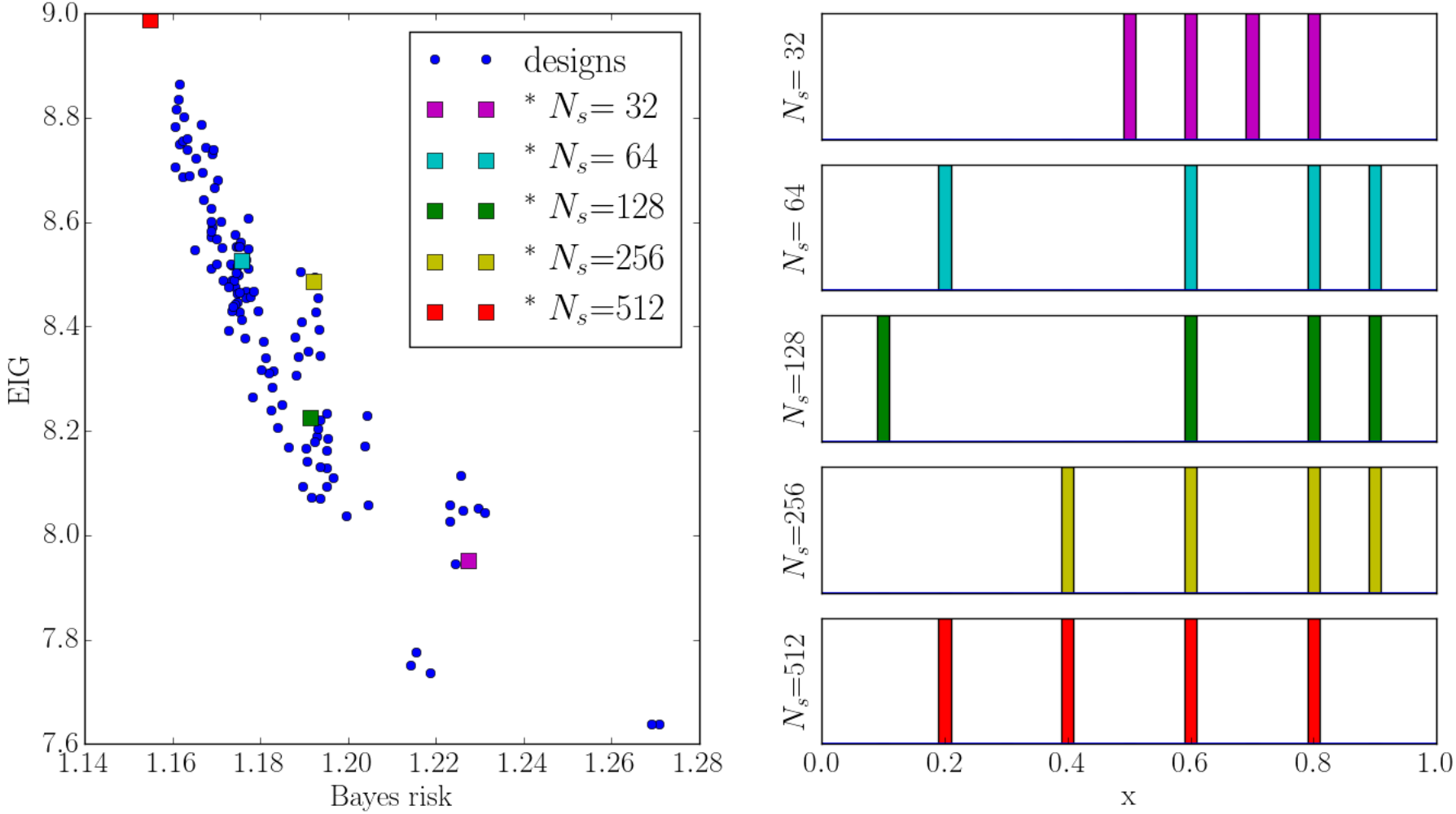
Penalty



Optimal designs (N_s = 512)



Effect of SAA Sample Size



Based on all 126 possible sensor locations

L-BFGS optimizer with bound constraints

2D Elliptic PDE Coefficient Inversion

- **Prior distribution:**

$$m \sim \mathcal{N}(0, \mathcal{C}_{\text{pr}})$$

$$\mathcal{C}_{\text{pr}} = \sigma^2 \left[\mathcal{I} - \text{div}(\Theta\nabla)\right]^{-2}$$

- **Forward model:**

$$\begin{cases} -\text{div}\left(e^m \nabla u\right) = 0 & \text{in } \Omega \\ u = u_d & \text{on } \Gamma_D \subset \partial\Omega \\ e^m \nabla u \cdot \mathbf{n} = 0 & \text{on } \Gamma_N \subset \partial\Omega \end{cases}$$

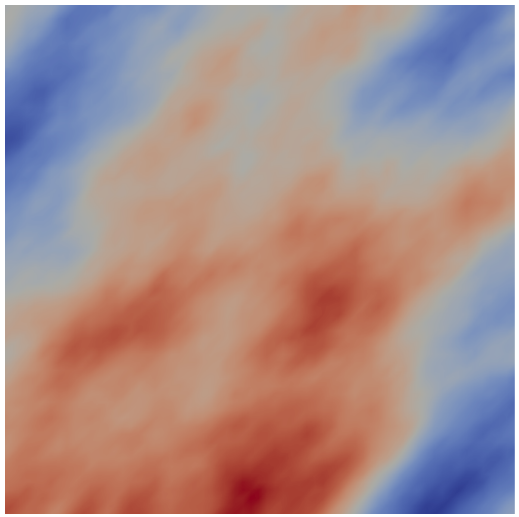
- **Design space:**
  10-by-10 grid of candidate sensor locations
  Select 15 locations

- **Brute force approach is not feasible:**

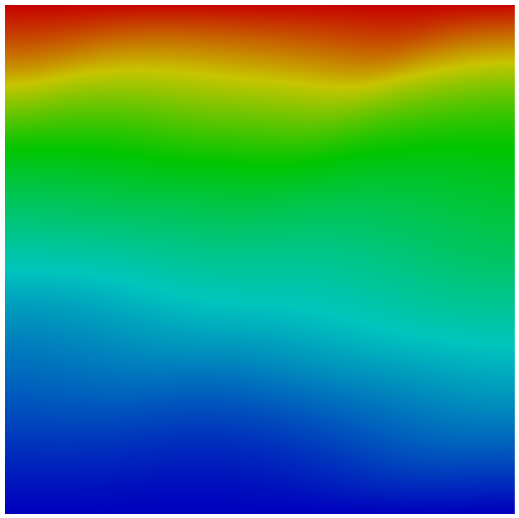
$$\begin{pmatrix} 100 \\ 15 \end{pmatrix} \approx 10^{17} \text{ combinations}$$

- Solve **continuous relaxation** using our optimization method

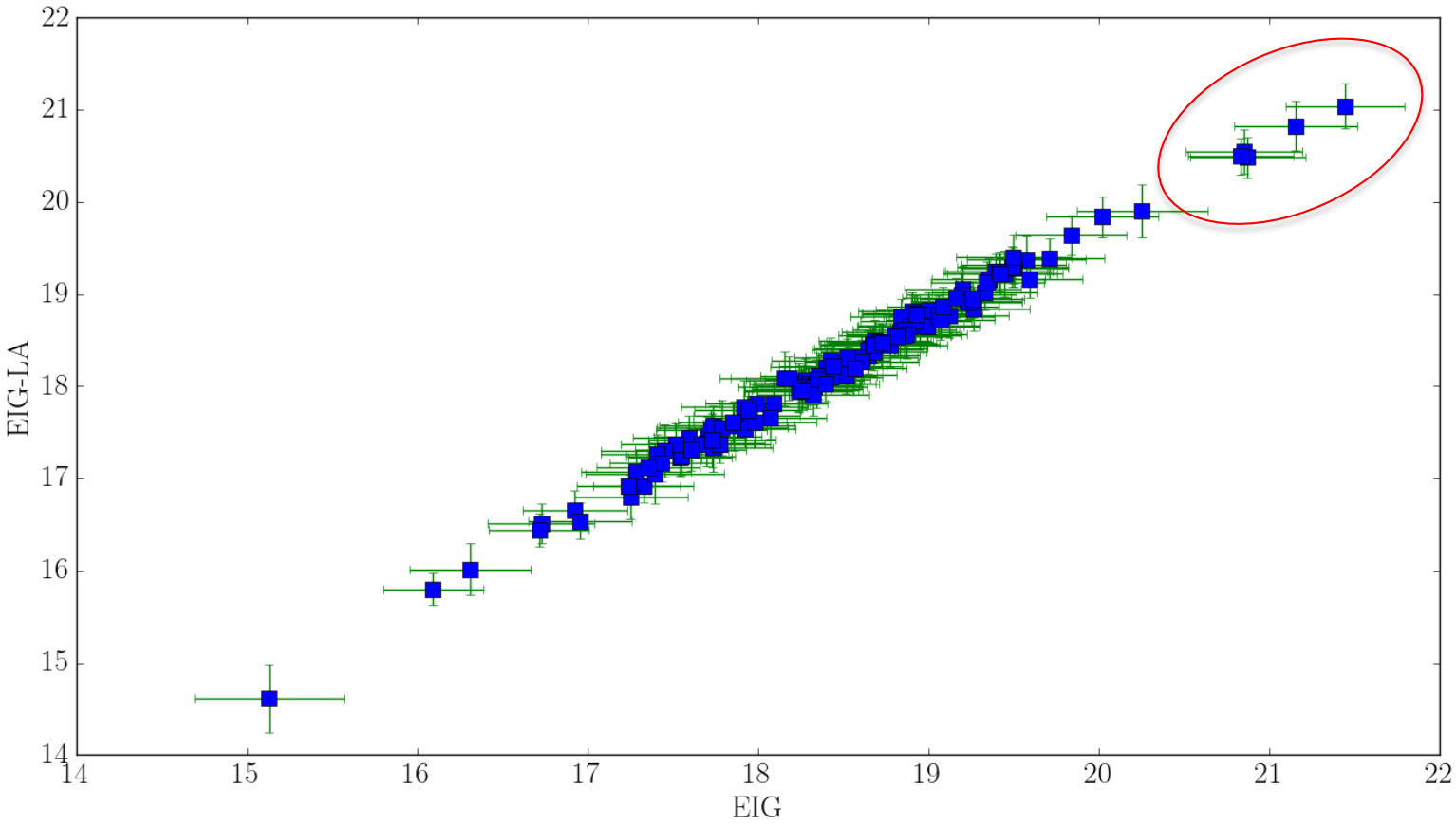
m: random realization



u(m): forward solution



Accuracy of Laplace Approximation



Blue squares: 123 random designs tested (from 10^{17} combinations) + 5 optimal in red

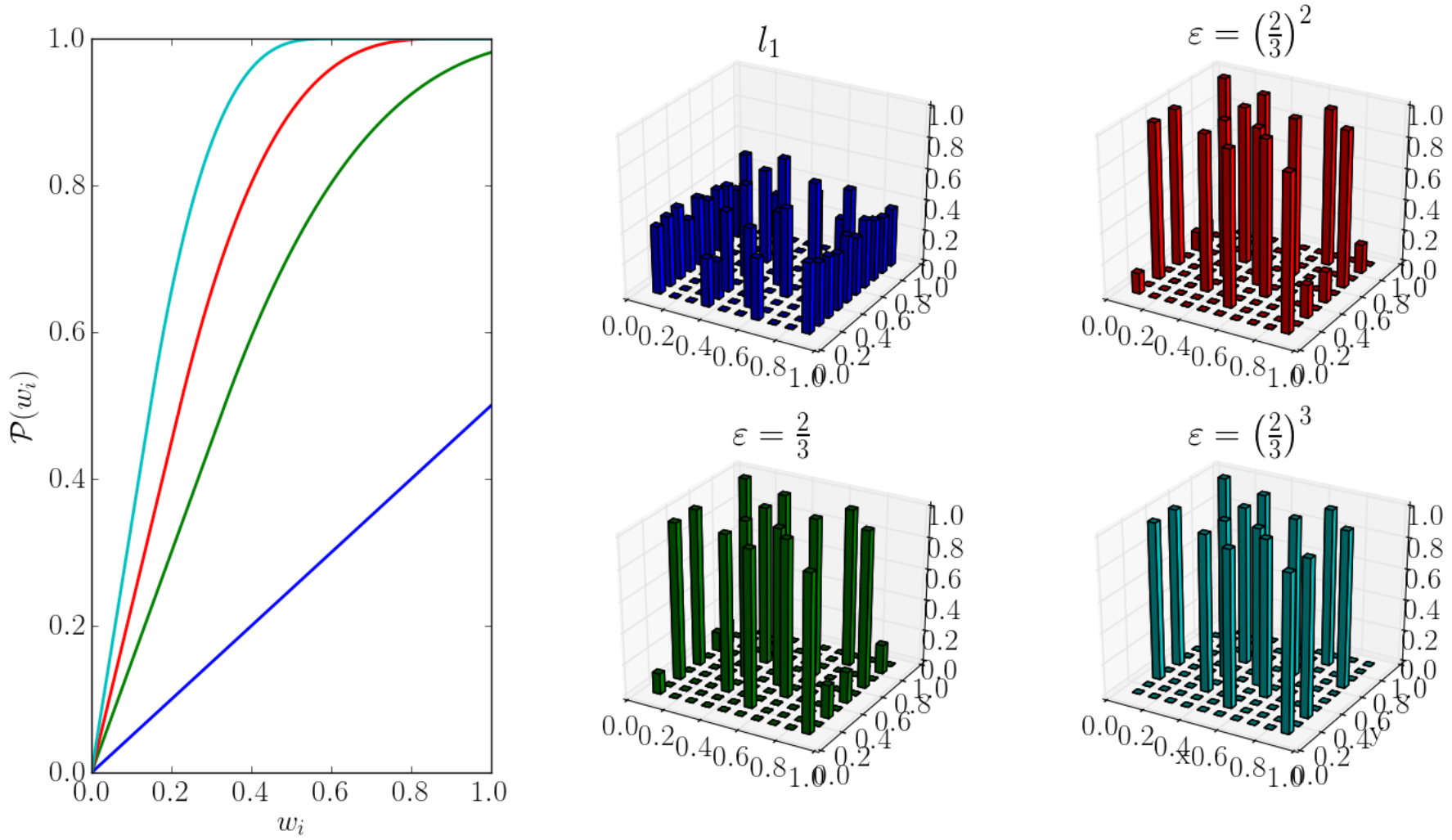
EIG: Double loop Monte Carlo (w/IS) with 1000 inner + 1000 outer samples

EIG-LA: Monte Carlo-Laplace Approximation with 1000 outer samples

Error bars: 3 standard deviations (99.7% CI)

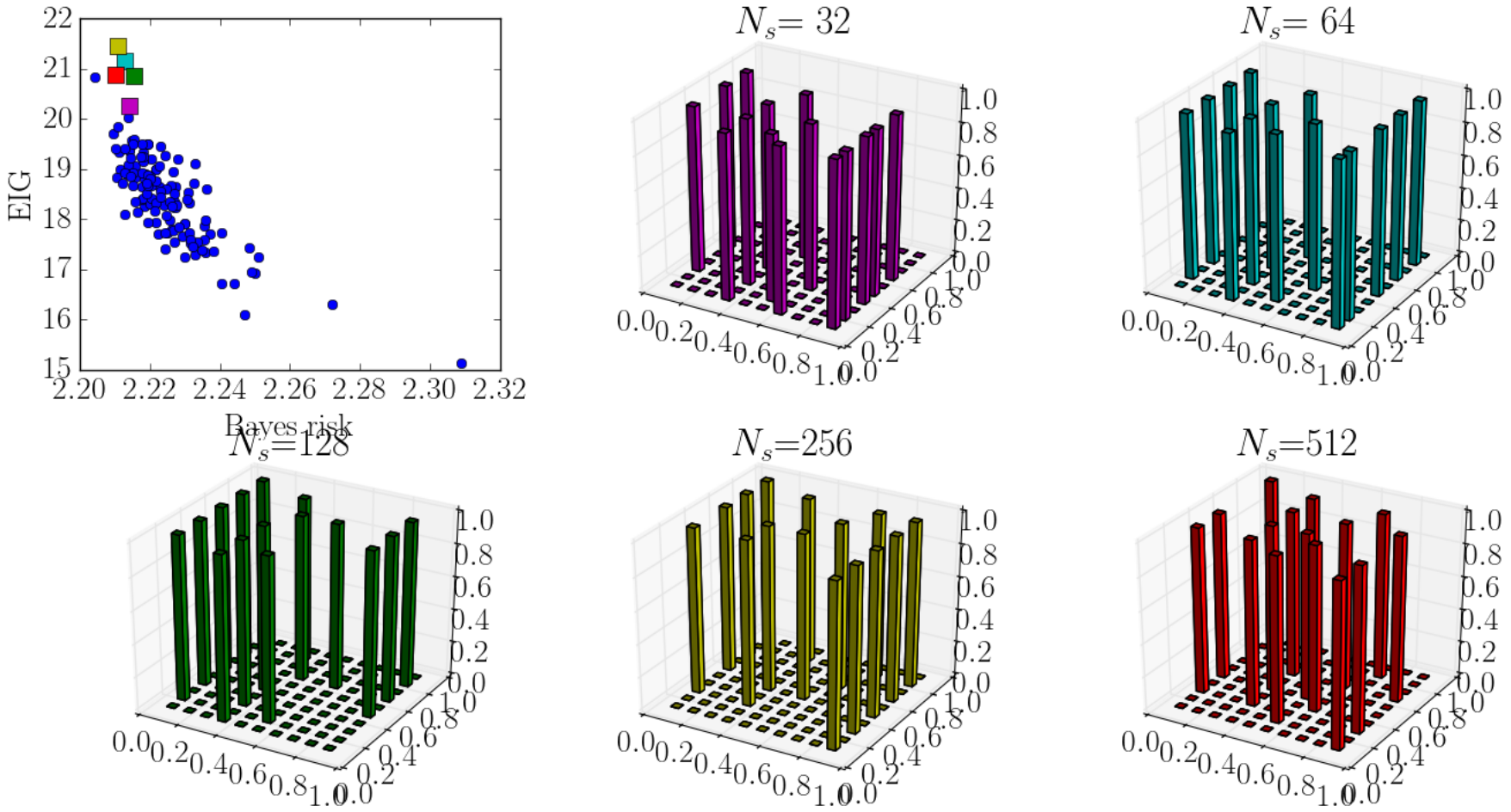
Correlation coefficient: 0.99

Sparsifying Penalty (continuation toward l_0)



N_s = 512. L-BFGS optimizer with bound constraints

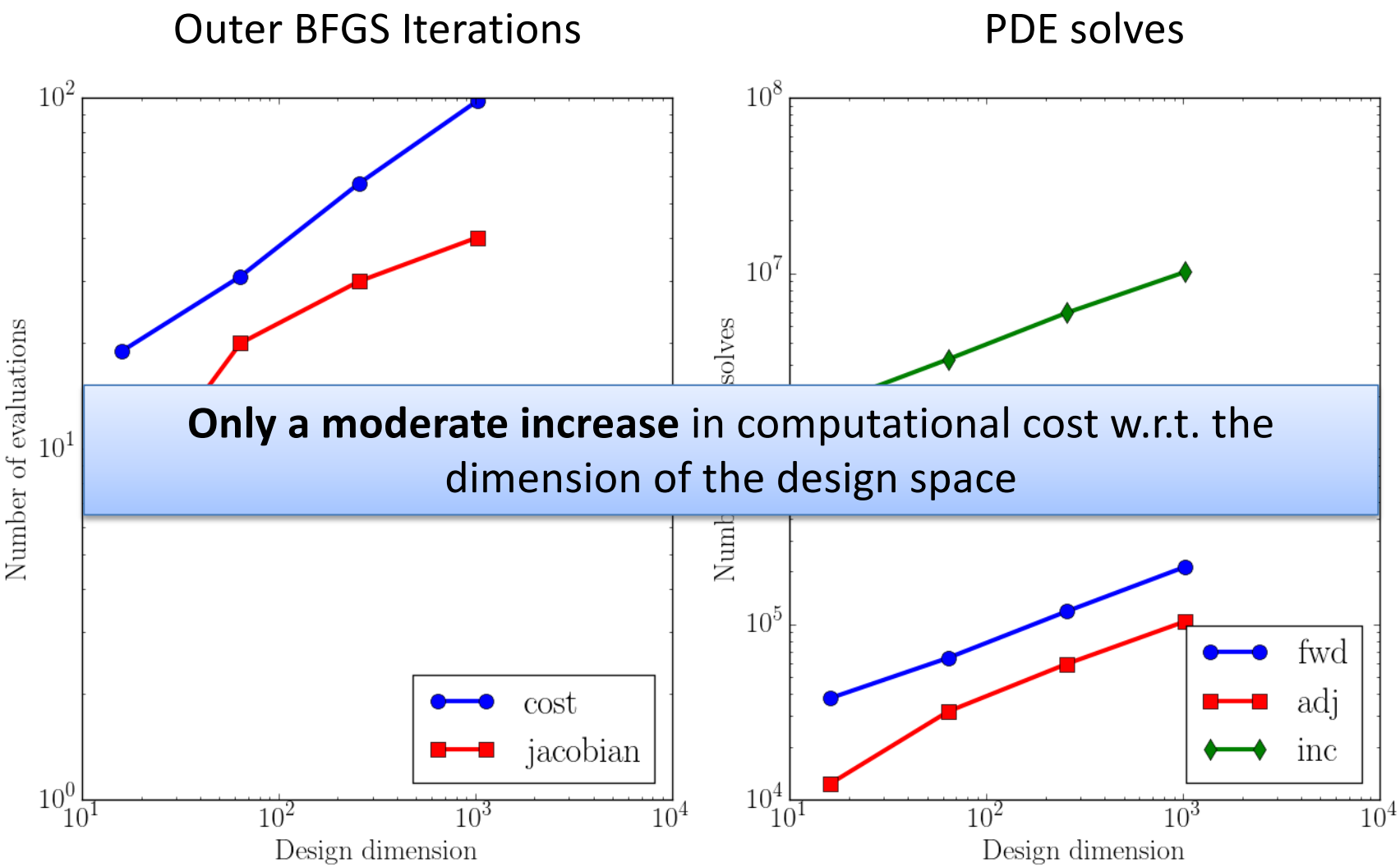
Effect of SAA Sample Size



100-choose-15 sensor locations

L-BFGS optimizer with bound constraints

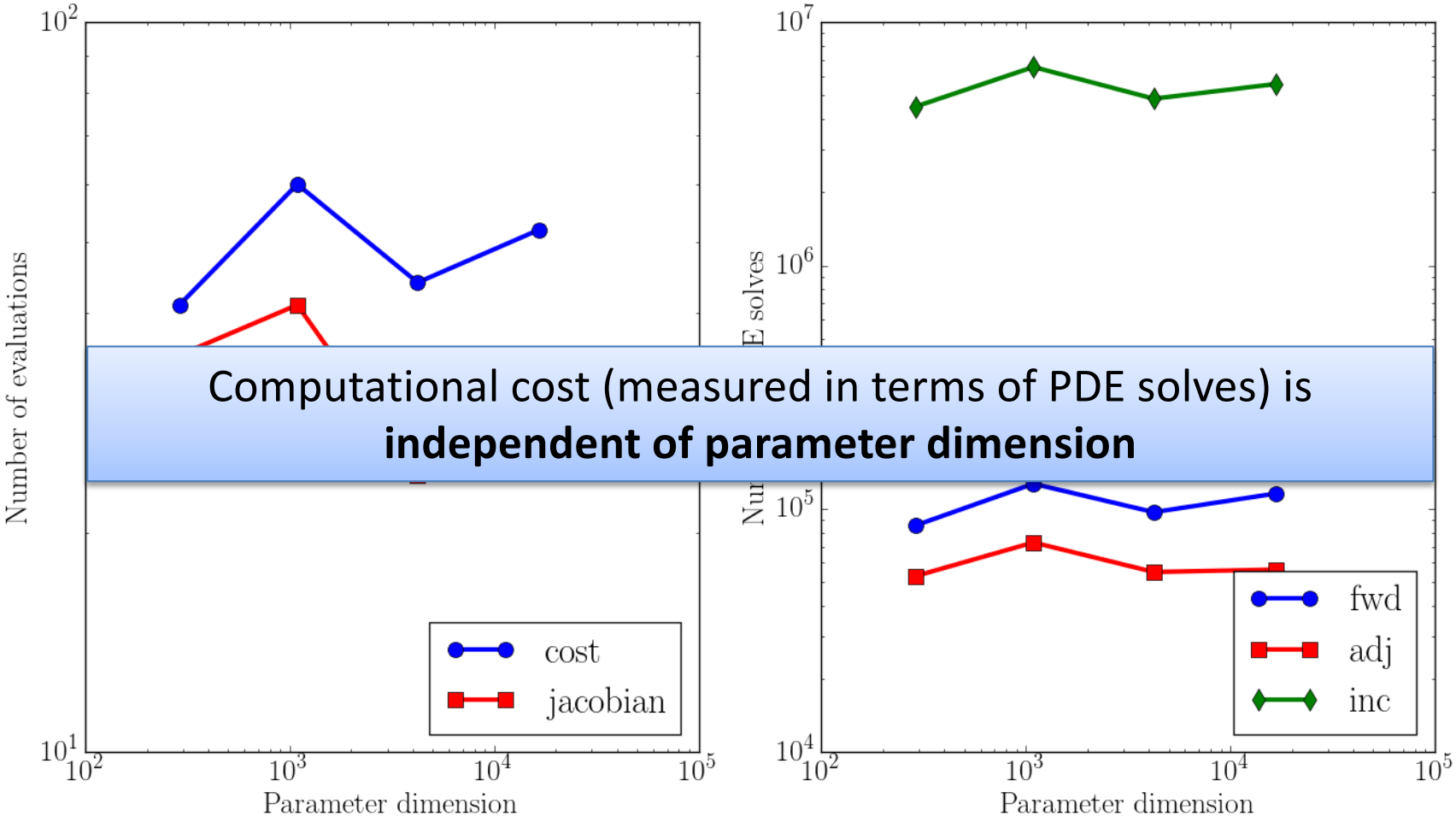
Computational Cost: Design-Dim



Computational Cost: Param-Dim

Outer BFGS Iterations

PDE solves



Computational cost (measured in terms of PDE solves) is **independent of parameter dimension**

Number of degrees of freedom in parameter space

Conclusions & Ongoing Work

- Repeated evaluation of EIG using **double Loop Monte Carlo (w/IS) is prohibitive** for large-scale problems
- **MC-Laplace approximation** allows for **efficient & accurate** estimation of EIG:
  - **SAA over** possible realization of **the data**
  - **Closed form expression for KL distance** between Gaussian involves "only" log-determinant & trace of (preconditioned) posterior covariance
  - **Randomized eigensolvers** permit efficient evaluation of these invariants
- Gradient computation using **variational adjoint methods**
- **Cost** measured in forward PDE solves is **independent of parameter space dimension**, and slowly increases with experimental design space dimension
- Ongoing/future work:
  - Employ Laplace approximation as control variate for expectation over data
  - For Helmholtz, rank of Hessian grows with increasing frequency; pursue compression schemes that do not depend on global low rank (H-matrix, product-convolution)
  - Pursue Newton methods for OED problem to achieve sensor-dimension independence