SAND2021-5029PE

Enabling and interpreting hyper-differential sensitivity analysis for Bayesian inverse problems

Joseph Hart†
with Bart van Bloemen Waanders†

† Sandia National Laboratories[1]
Center for Computing Research
Optimization and UQ Group

RISE Group Meeting

[1] 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. SANDXXXX

1. Bayesian inverse problems
2. Hyper-differential sensitivity analysis (HDSA)
3. The Bayesian interpretation of HDSA
4. Enabling HDSA for ill-posed problems
5. Numerical Results

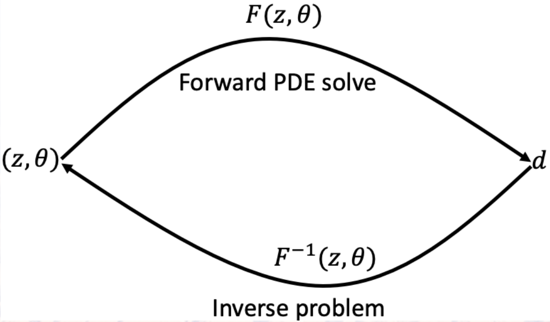
Inverse problems

Find parameters (z, θ) such that

$$F(z,\theta) = \mathcal{Q}(u(z,\theta)) \approx \boldsymbol{d}$$

where

- $\boldsymbol{d}$ are sparse and noisy observations of a state variable $u(z,\theta)$.
- $u(z,\theta)$ is the solution of a PDE and $\mathcal{Q}$ is the observation operator.



The joint Bayesian formulation

- Assume a prior distribution for $(\boldsymbol{z}, \boldsymbol{\theta}) \sim \mathcal{N}((\boldsymbol{z}_{\text{prior}}, \boldsymbol{\theta}_{\text{prior}}), \Gamma_{\text{prior}})$.
- Assume that $\boldsymbol{d} = F(\boldsymbol{z}^\star, \boldsymbol{\theta}^\star) + \epsilon$ where $\epsilon \sim \mathcal{N}(0, \Gamma_{\text{noise}})$.
- The joint posterior probability density function (PDF) is

$$\pi_{\text{post}}(\boldsymbol{z}, \boldsymbol{\theta}) \propto \pi_{\text{like}}(\boldsymbol{d}|\boldsymbol{z}, \boldsymbol{\theta})\pi_{\text{prior}}(\boldsymbol{z}, \boldsymbol{\theta})$$

  where

- π_{like} is the likelihood function,
- π_{prior} is the prior PDF.

Analyzing properties of $\pi_{\text{post}}(\boldsymbol{z}, \boldsymbol{\theta})$ provide a wealth of information, but may be computationally intractable.

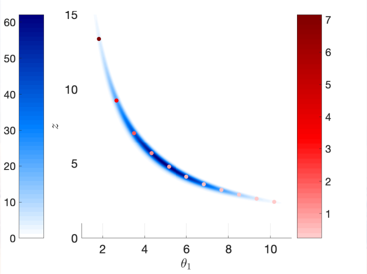
The conditional Bayesian formulation

- The posterior probability density function (PDF) for $\boldsymbol{z}$ given $\boldsymbol{\theta} = \boldsymbol{\theta}_{\text{prior}}$ is

$$\pi_{\text{post}}(\boldsymbol{z}|\boldsymbol{\theta}_{\text{prior}}) \propto \pi_{\text{like}}(\boldsymbol{d}|\boldsymbol{z}, \boldsymbol{\theta}_{\text{prior}})\pi_{\text{prior}}(\boldsymbol{z}, \boldsymbol{\theta}_{\text{prior}}).$$

- Fixing $\boldsymbol{\theta} = \boldsymbol{\theta}_{\text{prior}}$ to its prior mean simplifies the analysis.
- How does $\boldsymbol{\theta}_{\text{prior}}$ influence the posterior distribution for $\boldsymbol{z}$?

$\pi_{\text{post}}(z, \theta_1|\theta_2)$



Leveraging PDE-constrained optimization

- The maximum a posteriori probability (MAP) point(s) for $\pi_{\text{post}}(\boldsymbol{z}|\boldsymbol{\theta}_{\text{prior}})$ are local minima of

$$\min_{\boldsymbol{z} \in \mathbb{R}^m} J(\boldsymbol{z}; \boldsymbol{\theta}_{\text{prior}}) := M(\boldsymbol{z}, \boldsymbol{\theta}_{\text{prior}}) + R(\boldsymbol{z}, \boldsymbol{\theta}_{\text{prior}})$$

where $M(\boldsymbol{z}, \boldsymbol{\theta}_{\text{prior}})$ and $R(\boldsymbol{z}, \boldsymbol{\theta}_{\text{prior}})$ are the negative log likelihood and prior.

Leveraging PDE-constrained optimization

- The maximum a posteriori probability (MAP) point(s) for $\pi_{\text{post}}(\boldsymbol{z}|\boldsymbol{\theta}_{\text{prior}})$ are local minima of

$$\min_{\boldsymbol{z}\in\mathbb{R}^m} J(\boldsymbol{z};\boldsymbol{\theta}_{\text{prior}}) := M(\boldsymbol{z},\boldsymbol{\theta}_{\text{prior}}) + R(\boldsymbol{z},\boldsymbol{\theta}_{\text{prior}})$$

where $M(\boldsymbol{z},\boldsymbol{\theta}_{\text{prior}})$ and $R(\boldsymbol{z},\boldsymbol{\theta}_{\text{prior}})$ are the negative log likelihood and prior.

Benefit of PDE-constrained optimization

- Leverages computationally scalable, matrix free, and parallel algorithms.

Limitation of PDE-constrained optimization

- Only provides a limited characterization of the posterior distribution.

Leveraging PDE-constrained optimization

- The maximum a posteriori probability (MAP) point(s) for $\pi_{\text{post}}(\boldsymbol{z}|\boldsymbol{\theta}_{\text{prior}})$ are local minima of

$$\min_{\boldsymbol{z}\in\mathbb{R}^m} J(\boldsymbol{z};\boldsymbol{\theta}_{\text{prior}}) := M(\boldsymbol{z},\boldsymbol{\theta}_{\text{prior}}) + R(\boldsymbol{z},\boldsymbol{\theta}_{\text{prior}})$$

  where $M(\boldsymbol{z},\boldsymbol{\theta}_{\text{prior}})$ and $R(\boldsymbol{z},\boldsymbol{\theta}_{\text{prior}})$ are the negative log likelihood and prior.

Benefit of PDE-constrained optimization

- Leverages computationally scalable, matrix free, and parallel algorithms.

Limitation of PDE-constrained optimization

- Only provides a limited characterization of the posterior distribution.

This work focuses using hyper-differential sensitivity analysis (HDSA) to analyze the influence of $\boldsymbol{\theta}_{\text{prior}}$ on the MAP point for $\boldsymbol{z}|\boldsymbol{\theta}_{\text{prior}}$.

Post optimality sensitivity analysis

$$\min_{\boldsymbol{z} \in \mathbb{R}^m} J(\boldsymbol{z}; \boldsymbol{\theta}_{\text{prior}}) := M(\boldsymbol{z}, \boldsymbol{\theta}_{\text{prior}}) + R(\boldsymbol{z}, \boldsymbol{\theta}_{\text{prior}})$$

- Let $\boldsymbol{z}^\star$ denote a local minimum when $\boldsymbol{\theta} = \boldsymbol{\theta}_{prior}$ is fixed,

$$\nabla_{\boldsymbol{z}} J(\boldsymbol{z}^\star, \boldsymbol{\theta}_{\text{prior}}) = 0 \qquad \text{and} \qquad \nabla_{\boldsymbol{z},\boldsymbol{z}} J(\boldsymbol{z}^\star, \boldsymbol{\theta}_{\text{prior}}) \succ 0$$

Post optimality sensitivity analysis

$$\min_{\boldsymbol{z}\in\mathbb{R}^m} J(\boldsymbol{z};\boldsymbol{\theta}_{\text{prior}}) := M(\boldsymbol{z},\boldsymbol{\theta}_{\text{prior}}) + R(\boldsymbol{z},\boldsymbol{\theta}_{\text{prior}})$$

- Let $\boldsymbol{z}^\star$ denote a local minimum when $\boldsymbol{\theta} = \boldsymbol{\theta}_{prior}$ is fixed,

$$\nabla_{\boldsymbol{z}} J(\boldsymbol{z}^\star,\boldsymbol{\theta}_{\text{prior}}) = 0 \qquad \text{and} \qquad \nabla_{\boldsymbol{z},\boldsymbol{z}} J(\boldsymbol{z}^\star,\boldsymbol{\theta}_{\text{prior}}) \succ 0$$

- The implicit function theorem gives

$$\mathcal{G} : \mathcal{N}(\boldsymbol{\theta}_{\text{prior}}) \to \mathcal{N}(\boldsymbol{z}^\star),$$

defined on neighborhoods of $\boldsymbol{\theta}_{\text{prior}}$ and $\boldsymbol{z}^\star$, such that

$$\nabla_{\boldsymbol{z}} J(\mathcal{G}(\boldsymbol{\theta}),\boldsymbol{\theta}) = 0 \qquad \forall \boldsymbol{\theta} \in \mathcal{N}(\boldsymbol{\theta}_{\text{prior}})$$

Post optimality sensitivity analysis

$$\mathcal{G} : \mathcal{N}(\boldsymbol{\theta}_{\text{prior}}) \to \mathcal{N}(\boldsymbol{z}^{\star})$$

- $\mathcal{G}$ associates parameters $\boldsymbol{\theta}$ with the corresponding MAP points for $\boldsymbol{z}$ given $\boldsymbol{\theta}$
- Further, $\mathcal{G}$ is differentiable at $\boldsymbol{\theta}_{\text{prior}}$ and its Jacobian is given by

$$\mathcal{G}'(\boldsymbol{\theta}_{\text{prior}}) = -\mathcal{H}^{-1}\mathcal{B} \in \mathbb{R}^{m \times n}$$

- $\mathcal{H} = \nabla_{\boldsymbol{z},\boldsymbol{z}} J$ is the Hessian of J and,
- $\mathcal{B} = \nabla_{\boldsymbol{z},\boldsymbol{\theta}} J$ is the Jacobian of $\nabla_{\boldsymbol{z}} J$ with respect to $\boldsymbol{\theta}$

Post optimality sensitivity analysis

$$\mathcal{G} : \mathcal{N}(\boldsymbol{\theta}_{\text{prior}}) \to \mathcal{N}(\boldsymbol{z}^\star)$$

- $\mathcal{G}$ associates parameters $\boldsymbol{\theta}$ with the corresponding MAP points for $\boldsymbol{z}$ given $\boldsymbol{\theta}$
- Further, $\mathcal{G}$ is differentiable at $\boldsymbol{\theta}_{\text{prior}}$ and its Jacobian is given by

$$\mathcal{G}'(\boldsymbol{\theta}_{\text{prior}}) = -\mathcal{H}^{-1}\mathcal{B} \in \mathbb{R}^{m \times n}$$

- $\mathcal{H} = \nabla_{\boldsymbol{z},\boldsymbol{z}} J$ is the Hessian of J and,
- $\mathcal{B} = \nabla_{\boldsymbol{z},\boldsymbol{\theta}} J$ is the Jacobian of $\nabla_{\boldsymbol{z}} J$ with respect to $\boldsymbol{\theta}$

$\mathcal{G}'(\boldsymbol{\theta}_{\text{prior}})$ is like a Newton step to update the MAP point given a perturbation of $\boldsymbol{\theta}$.

Hyper-differential sensitivity analysis

Compute properties of

$$\mathcal{G}'(\boldsymbol{\theta}_{\text{prior}}) = -\mathcal{H}^{-1}\mathcal{B} \in \mathbb{R}^{m \times n},$$

a large dense matrix which is only accessible through matrix vector products.

Hyper-differential sensitivity analysis

Compute properties of

$$\mathcal{G}'(\boldsymbol{\theta}_{\text{prior}}) = -\mathcal{H}^{-1}\mathcal{B} \in \mathbb{R}^{m \times n},$$

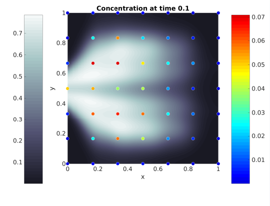
a large dense matrix which is only accessible through matrix vector products.

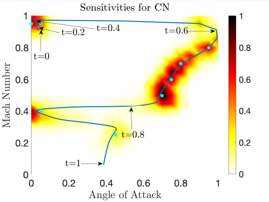
- Compute matrix-vector products with $\mathcal{H}$ and $\mathcal{B}$ using
  - adjoint-based derivative computations,
  - matrix free linear algebra.
- Approximation $\mathcal{G}'(\boldsymbol{\theta}_{\text{prior}})$ via a randomized generalized SVD
  - efficiently expose low rank structure,
  - embarrassingly parallel,
  - appropriate inner products to facilitate the interpretation.

Hyper-differential sensitivity analysis: Previous work

Has been used in the context of:

- Parameter uncertainty in PDE-constrained control
- Nuisance parameter uncertainty in deterministic inverse problems
- Data sensitivity to augment optimal experimental design
- Feedback controller robustness
- Model form error in PDE-constrained optimization





Has been applied in:

- Aerospace vehicle trajectory planning
- Thermal fluid system control
- Subsurface source inversion
- Ice sheet bedrock topography inversion

Pulling it all together

- State-of-the-art tools enables efficient analysis for the sensitivity of $\boldsymbol{z}$'s MAP point with respect to perturbations of $\boldsymbol{\theta}_{\text{prior}}$.
- Scalability in the parameter dimension is achieved through adjoint calculations and low rank approximation.

There are two questions we must address to facilitate this analysis:

Pulling it all together

- State-of-the-art tools enables efficient analysis for the sensitivity of $\boldsymbol{z}$'s MAP point with respect to perturbations of $\boldsymbol{\theta}_{\text{prior}}$.
- Scalability in the parameter dimension is achieved through adjoint calculations and low rank approximation.

There are two questions we must address to facilitate this analysis:

1. What is the Bayesian interpretation of $\mathcal{G}'(\boldsymbol{\theta}_{\text{prior}})$?

Pulling it all together

- State-of-the-art tools enables efficient analysis for the sensitivity of $\boldsymbol{z}$'s MAP point with respect to perturbations of $\boldsymbol{\theta}_{\text{prior}}$.
- Scalability in the parameter dimension is achieved through adjoint calculations and low rank approximation.

There are two questions we must address to facilitate this analysis:

1. What is the Bayesian interpretation of $\mathcal{G}'(\boldsymbol{\theta}_{\text{prior}})$?
2. Can I still compute/interpret $\mathcal{G}'(\boldsymbol{\theta}_{\text{prior}})$ for ill-posed problems where:
   - An ill-conditioned Hessian introduces theoretical and computational challenges,
   - The optimizer may struggle to solve the MAP point estimation problem to optimality (satisfaction of the first order optimality condition)?

The linear case

Theorem

If $F(\boldsymbol{z},\boldsymbol{\theta}) = A\boldsymbol{z} + B\boldsymbol{\theta}$ *then the posterior is Gaussian with covariance*

$$\Sigma_{post} = \begin{pmatrix} \Sigma_{\boldsymbol{z},\boldsymbol{z}} & \Sigma_{\boldsymbol{z},\boldsymbol{\theta}} \\ \Sigma_{\boldsymbol{\theta},\boldsymbol{z}} & \Sigma_{\boldsymbol{\theta},\boldsymbol{\theta}} \end{pmatrix}$$

and the post-optimality sensitivity is given by

$$\mathcal{G}'(\boldsymbol{\theta}_{prior}) = \Sigma_{\boldsymbol{z},\boldsymbol{\theta}}\Sigma_{\boldsymbol{\theta},\boldsymbol{\theta}}^{-1}.$$

- The post-optimality sensitivity is a correlation between $\boldsymbol{z}$ and $\boldsymbol{\theta}$.

The linear case

Theorem

If $F(\boldsymbol{z},\boldsymbol{\theta}) = A\boldsymbol{z} + B\boldsymbol{\theta}$ *then the posterior is Gaussian with covariance*

$$\Sigma_{post} = \begin{pmatrix} \Sigma_{\boldsymbol{z},\boldsymbol{z}} & \Sigma_{\boldsymbol{z},\boldsymbol{\theta}} \\ \Sigma_{\boldsymbol{\theta},\boldsymbol{z}} & \Sigma_{\boldsymbol{\theta},\boldsymbol{\theta}} \end{pmatrix}$$

and the post-optimality sensitivity is given by

$$\mathcal{G}'(\boldsymbol{\theta}_{prior}) = \Sigma_{\boldsymbol{z},\boldsymbol{\theta}}\Sigma_{\boldsymbol{\theta},\boldsymbol{\theta}}^{-1}.$$

- The post-optimality sensitivity is a correlation between $\boldsymbol{z}$ and $\boldsymbol{\theta}$.
- Connection between optimization/analysis and Bayesian statistics.

The linear case

Theorem

If $F(\boldsymbol{z},\boldsymbol{\theta}) = A\boldsymbol{z} + B\boldsymbol{\theta}$ *then the posterior is Gaussian with covariance*

$$\Sigma_{post} = \begin{pmatrix} \Sigma_{\boldsymbol{z},\boldsymbol{z}} & \Sigma_{\boldsymbol{z},\boldsymbol{\theta}} \\ \Sigma_{\boldsymbol{\theta},\boldsymbol{z}} & \Sigma_{\boldsymbol{\theta},\boldsymbol{\theta}} \end{pmatrix}$$

and the post-optimality sensitivity is given by

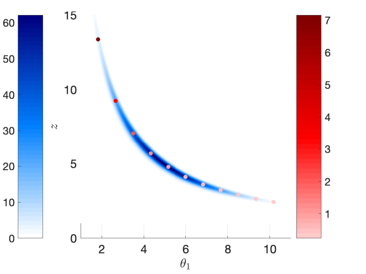
$$\mathcal{G}'(\boldsymbol{\theta}_{prior}) = \Sigma_{\boldsymbol{z},\boldsymbol{\theta}}\Sigma_{\boldsymbol{\theta},\boldsymbol{\theta}}^{-1}.$$

- The post-optimality sensitivity is a correlation between $\boldsymbol{z}$ and $\boldsymbol{\theta}$.
- Connection between optimization/analysis and Bayesian statistics.
- Local correlation for nonlinear inverse problems (Laplace approximation).

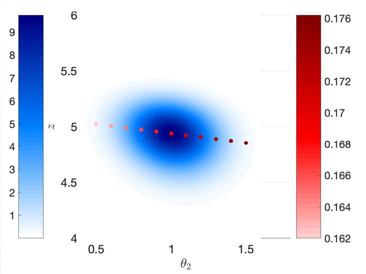
An illustrative example

$$F(z, \theta_1, \theta_2) = e^{\frac{1}{10} z \theta_1} + \theta_2$$

$\pi_{\text{post}}(z, \theta_1 | \theta_2)$



$\pi_{\text{post}}(z, \theta_2 | \theta_1)$



- High sensitivity (left panel) corresponds to stronger correlations
- Approximately Gaussian distribution (right panel) has nearly constant sensitivity (local correlations = global correlations for Gaussians)

Handling ill-conditioning

$$\mathcal{G}'(\boldsymbol{\theta}_{\text{prior}}) = -\mathcal{H}^{-1}\mathcal{B}$$

- For ill-posed inverse problems, $\mathcal{H}$ may be ill-conditioned.
- Can yield high sensitivity as a result of lacking information/data.
- Analyzing $\mathcal{G}'(\boldsymbol{\theta}_{\text{prior}})$ will be numerically troublesome and the resulting sensitivities may be dominated by what the data does not tell you.

[2] T. Cui, J. Martin, Y. M. Marzouk, A. Solonen, and A. Spantini, Likelihood-informed dimension reduction for nonlinear inverse problems, Inverse Problems, 30 (2014), pp. 1– 28.

Handling ill-conditioning

$$\mathcal{G}'(\boldsymbol{\theta}_{\text{prior}}) = -\mathcal{H}^{-1}\mathcal{B}$$

- For ill-posed inverse problems, $\mathcal{H}$ may be ill-conditioned.
- Can yield high sensitivity as a result of lacking information/data.
- Analyzing $\mathcal{G}'(\boldsymbol{\theta}_{\text{prior}})$ will be numerically troublesome and the resulting sensitivities may be dominated by what the data does not tell you.

Proposed Approach: Compute sensitivities

$$\mathcal{P}\mathcal{G}'(\boldsymbol{\theta}_{\text{prior}}) = -\mathcal{P}\mathcal{H}^{-1}\mathcal{B}$$

where $\mathcal{P}$ projects onto the likelihood informed subspace [2].

[2] T. Cui, J. Martin, Y. M. Marzouk, A. Solonen, and A. Spantini, Likelihood-informed dimension reduction for nonlinear inverse problems, Inverse Problems, 30 (2014), pp. 1– 28.

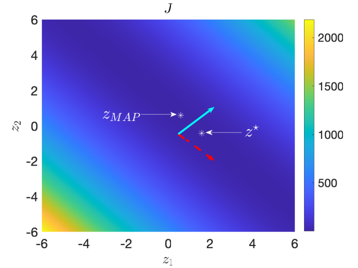
Likelihood informed subspaces (LIS)

$$\mathcal{H}_M \boldsymbol{v}_j = \lambda_j \mathcal{H}_R \boldsymbol{v}_j \qquad \mathcal{H} = \mathcal{H}_M + \mathcal{H}_R$$

The eigenvalue

$$\lambda_j = \frac{\boldsymbol{v}_j^T \mathcal{H}_M \boldsymbol{v}_j}{\boldsymbol{v}_j^T \mathcal{H}_R \boldsymbol{v}_j}$$

measures the ratio of the likelihood and prior in the direction of $\boldsymbol{v}_j$.



$$\mathcal{P}\mathcal{G}'(\boldsymbol{\theta}_{\text{prior}}) = -\mathcal{P}\mathcal{H}^{-1}\mathcal{B}$$

Project the sensitivities onto the span of $\{\boldsymbol{v}_1, \boldsymbol{v}_2, \ldots, \boldsymbol{v}_r\}$, the subspace which is most informed by the likelihood.

Computing LIS sensitivities

Theorem

The LIS sensitivity is given by

$$S(\overline{\boldsymbol{\theta}}) = \left|\left|\mathcal{P}\mathcal{H}^{-1}\mathcal{B}\overline{\boldsymbol{\theta}}\right|\right|_{W_z} = \sqrt{\sum_{k=1}^{r}\sum_{j=1}^{r}\left(\frac{\boldsymbol{v}_j^T\mathcal{B}\overline{\boldsymbol{\theta}}}{1+\lambda_j}\right)\left(\frac{\boldsymbol{v}_k^T\mathcal{B}\overline{\boldsymbol{\theta}}}{1+\lambda_k}\right)\boldsymbol{v}_k^T W_Z \boldsymbol{v}_j}.$$

- Need the leading eigenpairs

$$\mathcal{H}_M \boldsymbol{v}_j = \lambda_j \mathcal{H}_R \boldsymbol{v}_j$$

  rather than $\mathcal{H}^{-1}$.
- Sensitivities inherit the likelihood to prior ratio interpretation.
- W_Z measures the inner products in the original function space.

Failure to satisfy the first order optimality condition?

It may not be practical to solve the MAP point estimation problem

$$\min_{\boldsymbol{z} \in \mathbb{R}^m} J(\boldsymbol{z}; \boldsymbol{\theta}_{\text{prior}}) := M(\boldsymbol{z}, \boldsymbol{\theta}_{\text{prior}}) + R(\boldsymbol{z}, \boldsymbol{\theta}_{\text{prior}})$$

to optimality if ill-conditioning yields slow convergence.

- Early iterations refine features which are well informed by the data.
- Ill-conditioning may yield slow convergence in the uniformed subspace.

Failure to satisfy the first order optimality condition?

It may not be practical to solve the MAP point estimation problem

$$\min_{\boldsymbol{z} \in \mathbb{R}^m} J(\boldsymbol{z}; \boldsymbol{\theta}_{\text{prior}}) := M(\boldsymbol{z}, \boldsymbol{\theta}_{\text{prior}}) + R(\boldsymbol{z}, \boldsymbol{\theta}_{\text{prior}})$$

to optimality if ill-conditioning yields slow convergence.

- Early iterations refine features which are well informed by the data.
- Ill-conditioning may yield slow convergence in the uniformed subspace.

Question: HDSA assume satisfaction of the optimality criteria. What can I do when converging the optimization is impractical/unnecessary?

Failure to satisfy the first order optimality condition?

Idea: Compute sensitivities of a nearby problem which is solved to optimality.

Failure to satisfy the first order optimality condition?

Idea: Compute sensitivities of a nearby problem which is solved to optimality.

- Assume that $\boldsymbol{z}^\star$ is an approximation of the MAP point but

$$\nabla_{\boldsymbol{z}} J(\boldsymbol{z}^\star, \boldsymbol{\theta}_{\text{prior}}) \neq 0.$$

- Find a minimum norm perturbation $\tilde{R}$ so that $\nabla_{\boldsymbol{z}} J(\boldsymbol{z}^\star; \boldsymbol{\theta}_{\text{prior}}) + \nabla_{\boldsymbol{z}} \tilde{R}(\boldsymbol{z}^\star) = 0$,

$$\min_{\tilde{R} \in Q} ||\tilde{R}||_{L^1(\mu)}$$

$$\text{s.t. } \nabla_{\boldsymbol{z}} \tilde{R}(\boldsymbol{z}^\star) = -\nabla_{\boldsymbol{z}} J(\boldsymbol{z}^\star; \boldsymbol{\theta}_{\text{prior}})$$

Failure to satisfy the first order optimality condition?

Idea: Compute sensitivities of a nearby problem which is solved to optimality.

- Assume that $\boldsymbol{z}^\star$ is an approximation of the MAP point but

$$\nabla_{\boldsymbol{z}} J(\boldsymbol{z}^\star, \boldsymbol{\theta}_{\text{prior}}) \neq 0.$$

- Find a minimum norm perturbation $\tilde{R}$ so that $\nabla_{\boldsymbol{z}} J(\boldsymbol{z}^\star; \boldsymbol{\theta}_{\text{prior}}) + \nabla_{\boldsymbol{z}} \tilde{R}(\boldsymbol{z}^\star) = 0$,

$$\min_{\tilde{R} \in Q} ||\tilde{R}||_{L^1(\mu)}$$

$$\text{s.t. } \nabla_{\boldsymbol{z}} \tilde{R}(\boldsymbol{z}^\star) = -\nabla_{\boldsymbol{z}} J(\boldsymbol{z}^\star; \boldsymbol{\theta}_{\text{prior}})$$

where

- $Q = \{\tilde{R} : \mathbb{R}^m \to \mathbb{R} | \tilde{R} \geq 0,\ \tilde{R} \text{ is quadratic, } \tilde{R} \text{ is convex}\}$
- $L^1(\mu)$ is defined by a Gaussian measure μ with mean $\boldsymbol{z}^\star$ and covariance $\alpha^2 I$
- α is a user defined length scale parameter (will revisit it later)

First order a posteriori update

Can solve

$$\min_{\tilde{R} \in Q} ||\tilde{R}||_{L^1(\mu)}$$
$$\text{s.t. } \nabla_{\boldsymbol{z}} \tilde{R}(\boldsymbol{z}^\star) = -\nabla_{\boldsymbol{z}} J(\boldsymbol{z}^\star; \boldsymbol{\theta}_{\text{prior}})$$

First order a posteriori update

Can solve

$$\min_{\tilde{R} \in Q} ||\tilde{R}||_{L^1(\mu)}$$

$$\text{s.t. } \nabla_{\boldsymbol{z}} \tilde{R}(\boldsymbol{z}^\star) = -\nabla_{\boldsymbol{z}} J(\boldsymbol{z}^\star; \boldsymbol{\theta}_{\text{prior}})$$

in closed form (with judicious algebraic manipulations) to find the update

$$\tilde{R}(\boldsymbol{z}) = \frac{\alpha}{2} ||\boldsymbol{g}||_2 - (\boldsymbol{z} - \boldsymbol{z}^\star)^T \boldsymbol{g} + \frac{1}{2} (\boldsymbol{z} - \boldsymbol{z}^\star)^T \frac{1}{\alpha ||\boldsymbol{g}||_2} \boldsymbol{g}\boldsymbol{g}^T (\boldsymbol{z} - \boldsymbol{z}^\star),$$

where $\boldsymbol{g} = \nabla_{\boldsymbol{z}} J(\boldsymbol{z}^\star; \boldsymbol{\theta}_{\text{prior}})$.

First order a posteriori update

Can solve

$$\min_{\tilde{R} \in Q} ||\tilde{R}||_{L^1(\mu)}$$
$$\text{s.t. } \nabla_{\boldsymbol{z}} \tilde{R}(\boldsymbol{z}^\star) = -\nabla_{\boldsymbol{z}} J(\boldsymbol{z}^\star; \boldsymbol{\theta}_{\text{prior}})$$

in closed form (with judicious algebraic manipulations) to find the update

$$\tilde{R}(\boldsymbol{z}) = \frac{\alpha}{2}||\boldsymbol{g}||_2 - (\boldsymbol{z} - \boldsymbol{z}^\star)^T \boldsymbol{g} + \frac{1}{2}(\boldsymbol{z} - \boldsymbol{z}^\star)^T \frac{1}{\alpha ||\boldsymbol{g}||_2} \boldsymbol{g}\boldsymbol{g}^T (\boldsymbol{z} - \boldsymbol{z}^\star),$$

where $\boldsymbol{g} = \nabla_{\boldsymbol{z}} J(\boldsymbol{z}^\star; \boldsymbol{\theta}_{\text{prior}})$.

- $\tilde{R}$ is a "nice" function.
- The length scale parameter α dictates the curvature.
- Computational cost is negligible since a closed form expression is available.

The perturbed MAP point problem

$$\min_{\boldsymbol{z}\in\mathbb{R}^m} J(\boldsymbol{z};\boldsymbol{\theta}_{\text{prior}}) + \tilde{R}(\boldsymbol{z})$$

- $\boldsymbol{z}^\star$ satisfies the first order optimality condition.
- Post-optimality sensitivities are well defined for this perturbed problem.

The perturbed MAP point problem

$$\min_{\boldsymbol{z} \in \mathbb{R}^m} J(\boldsymbol{z}; \boldsymbol{\theta}_{\text{prior}}) + \tilde{R}(\boldsymbol{z})$$

- $\boldsymbol{z}^\star$ satisfies the first order optimality condition.
- Post-optimality sensitivities are well defined for this perturbed problem.

Some important questions:

- What is the Bayesian interpretation of $\tilde{R}$?
- How does the perturbation influence the sensitivities?
- How do the sensitivities depend on α (the length scale parameter)?

A perturbed Gaussian prior

- What is the Bayesian interpretation of $\tilde{R}$?
- How does the perturbation influence the sensitivities?
- How do the sensitivities depend on α (the length scale parameter)?

Theorem

The perturbed inverse problem has a Gaussian prior with mean

$$\tilde{\boldsymbol{z}}_{prior} = \boldsymbol{z}_{prior} + \frac{\alpha - (\boldsymbol{z}^\star - \boldsymbol{z}_{prior})^T \boldsymbol{s}}{\alpha - \boldsymbol{v}^T \boldsymbol{s}} \boldsymbol{v}$$

and covariance

$$\tilde{\Gamma}_{prior} = \Gamma_{prior} - \frac{1}{||\boldsymbol{g}||_2} \frac{1}{\alpha - \boldsymbol{v}^T \boldsymbol{s}} \boldsymbol{v}\boldsymbol{v}^T$$

where

$$\boldsymbol{g} = \nabla_{\boldsymbol{z}} J(\boldsymbol{z}^\star; \boldsymbol{\theta}_{prior}), \qquad \boldsymbol{s} = -\frac{\boldsymbol{g}}{||\boldsymbol{g}||_2}, \qquad \text{and} \qquad \boldsymbol{v} = \Gamma_{prior} \boldsymbol{g}.$$

The perturbation shifts the mean and reduces uncertainty in the direction $\boldsymbol{v}$.

Difference in sensitivities

- What is the Bayesian interpretation of $\tilde{R}$?
- How does the perturbation influence the sensitivities?
- How do the sensitivities depend on α (the length scale parameter)?

Theorem

The quantities

$$S(\overline{\boldsymbol{\theta}}) = \left|\left|\mathcal{P}\mathcal{H}^{-1}\mathcal{B}\overline{\boldsymbol{\theta}}\right|\right|_{W_z} \qquad \text{and} \qquad \tilde{S}(\overline{\boldsymbol{\theta}}) = \left|\left|\mathcal{P}\tilde{\mathcal{H}}^{-1}\tilde{\mathcal{B}}\overline{\boldsymbol{\theta}}\right|\right|_{W_z}$$

satisfy

$$\frac{|\tilde{S}(\overline{\boldsymbol{\theta}}) - S(\overline{\boldsymbol{\theta}})|}{\left|\left|\mathcal{H}^{-1}\mathcal{B}\overline{\boldsymbol{\theta}}\right|\right|_{W_z}} \leq \frac{||\mathcal{P}\boldsymbol{n}||_{W_z}}{\boldsymbol{s}^T\boldsymbol{n} + \alpha},$$

where

$$\boldsymbol{g} = \nabla_{\boldsymbol{z}} J(\boldsymbol{z}^\star; \boldsymbol{\theta}_{prior}), \qquad \boldsymbol{s} = -\frac{\boldsymbol{g}}{||\boldsymbol{g}||_2}, \qquad \text{and} \qquad \boldsymbol{n} = -\mathcal{H}^{-1}\boldsymbol{g}.$$

Robustness with respect to α

- What is the Bayesian interpretation of $\tilde{R}$?
- How does the perturbation influence the sensitivities?
- How do the sensitivities depend on α (the length scale parameter)?

Theorem

Letting $\tilde{S}_\alpha(\overline{\boldsymbol{\theta}})$ be the sensitivity as a function of α,

$$\frac{|\tilde{S}_{\alpha+\alpha\beta}(\overline{\boldsymbol{\theta}}) - \tilde{S}_\alpha(\overline{\boldsymbol{\theta}})|}{||\mathcal{H}^{-1}\mathcal{B}\boldsymbol{e}_i||_{W_Z}} < |\beta| \cdot \frac{||\mathcal{P}\boldsymbol{n}||_{W_Z}}{\boldsymbol{s}^T\boldsymbol{n} + \alpha(1+\beta)} \qquad \text{for } -1 < \beta < 1$$

where

$$\boldsymbol{g} = \nabla_{\boldsymbol{z}} J(\boldsymbol{z}^\star; \boldsymbol{\theta}_{prior}), \qquad \boldsymbol{s} = -\frac{\boldsymbol{g}}{||\boldsymbol{g}||_2}, \qquad \text{and} \qquad \boldsymbol{n} = -\mathcal{H}^{-1}\boldsymbol{g}.$$

First order a posteriori update

- What is the Bayesian interpretation of $\tilde{R}$?
- How does the perturbation influence the sensitivities?
- How do the sensitivities depend on α (the length scale parameter)?

First order a posteriori update

- What is the Bayesian interpretation of $\tilde{R}$?
- How does the perturbation influence the sensitivities?
- How do the sensitivities depend on α (the length scale parameter)?

- $\tilde{R}$ shifts the prior mean and reduces the variance in the direction $\boldsymbol{v} = \Gamma_{\text{prior}}\boldsymbol{g}$.
- The change in the sensitivity indices and their robustness with respect to α are bounded by quantities proportional to $||\mathcal{P}\boldsymbol{n}||$.

Take away message: If the optimizer has converged in the likelihood informed subspace, then HDSA is a robust, interpretable, scalable, and efficient (RISE) way to assess correlations in the joint Bayesian posterior distribution.

An illustrative example

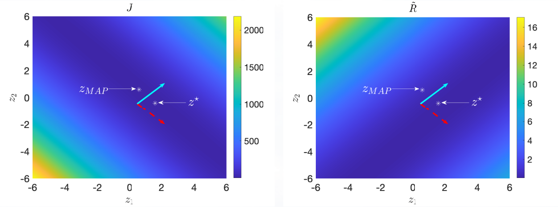


Figure: The solid cyan arrow indicates the likelihood informed subspace and the broken red arrow indicates the uninformed subspace.

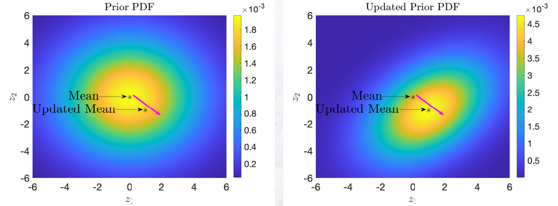


Figure: The magenta arrow indicates the direction of $\boldsymbol{v} = \Gamma_{\text{prior}} \boldsymbol{g}$.

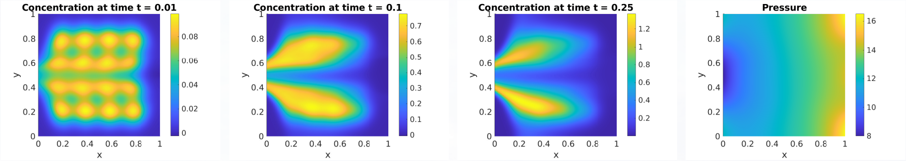
The two questions

1. What is the Bayesian interpretation of $\mathcal{G}'(\boldsymbol{\theta}_{\text{prior}})$?

   The local correlation between $\boldsymbol{z}$ and $\boldsymbol{\theta}$ in the Bayesian posterior.

2. Can I still compute/interpret $\mathcal{G}'(\boldsymbol{\theta}_{\text{prior}})$ for ill-posed problems where:
   - An ill-conditioned Hessian introduces theoretical and computational challenges,
     Project on likelihood informed subspaces.
   - The optimizer may struggle to solve the MAP point estimation problem to optimality (satisfaction of the first optimality condition)?
     A posteriori update.

Subsurface permeability inversion



$$-\nabla \cdot (e^{\kappa} \nabla p) = 0 \qquad \text{in } \Omega$$

$$c_t - \nabla \cdot \Big(\epsilon(\boldsymbol{\theta}) \nabla c \Big) + \nabla \cdot \big(-e^{\kappa} \nabla p c \big) = g(\boldsymbol{\theta}) \qquad \text{in } [0, T] \times \Omega$$

$$p = p_1(\boldsymbol{\theta}) \qquad \text{on } \Gamma_1$$

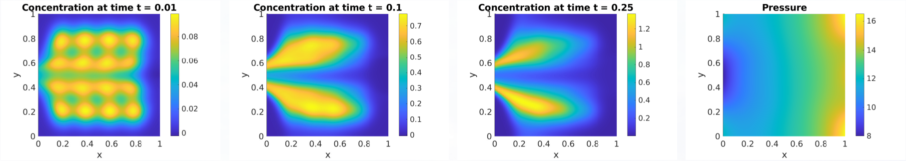
$$p = p_2(\boldsymbol{\theta}) \qquad \text{on } \Gamma_3$$

$$e^{\kappa} \nabla p \cdot n = 0 \qquad \text{on } \Gamma_0 \cup \Gamma_2$$

$$\nabla c \cdot n = 0 \qquad \text{on } [0, T] \times \{\Gamma_0 \cup \Gamma_1 \cup \Gamma_2 \cup \Gamma_3\}$$

$$c(0, x) = 0 \qquad \text{in } \Omega$$

Subsurface permeability inversion



$$-\nabla \cdot (e^{\kappa} \nabla p) = 0 \qquad \text{in } \Omega$$

$$c_t - \nabla \cdot \Big(\epsilon(\boldsymbol{\theta}) \nabla c \Big) + \nabla \cdot \big(-e^{\kappa} \nabla p c \big) = g(\boldsymbol{\theta}) \qquad \text{in } [0, T] \times \Omega$$

$$p = p_1(\boldsymbol{\theta}) \qquad \text{on } \Gamma_1$$

$$p = p_2(\boldsymbol{\theta}) \qquad \text{on } \Gamma_3$$

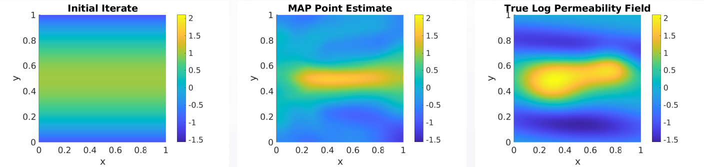
$$e^{\kappa} \nabla p \cdot n = 0 \qquad \text{on } \Gamma_0 \cup \Gamma_2$$

$$\nabla c \cdot n = 0 \qquad \text{on } [0, T] \times \{\Gamma_0 \cup \Gamma_1 \cup \Gamma_2 \cup \Gamma_3\}$$

$$c(0, x) = 0 \qquad \text{in } \Omega$$

Subsurface permeability inversion

$$\min_{\kappa} \sum_{i=1}^{n_c} w_c(\mathcal{Q}_c^i c(\kappa) - d_c^i)^2 + \sum_{i=1}^{n_p} w_p(\mathcal{Q}_p^i p(\kappa) - d_p^i)^2 + \gamma_1 ||\nabla \kappa||^2 + \gamma_2 ||\kappa||^2$$

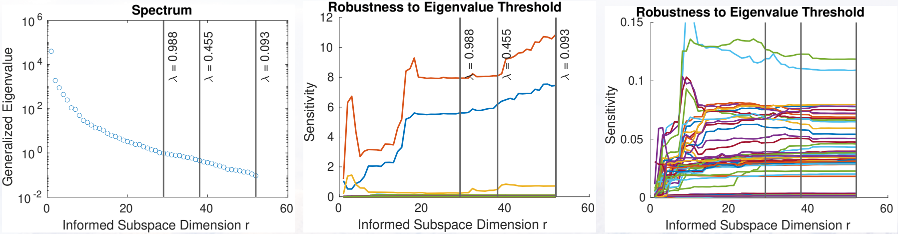


Iteration	Objective	Gradient Norm	Step Size
0	17.2	1.33	N/A
4	9.59	.697	15.6
10	3.29	.676	2.38
41	.897	.113	2.02
65	.578	.331	.115
75	.571	.102	.109
125	.529	.081	.034

Significant computational efforts gives little improvement.

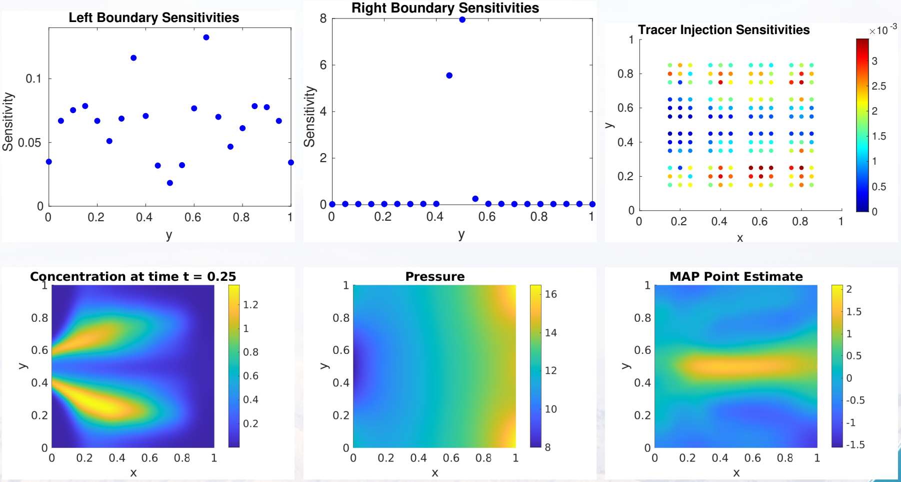
Computing the likelihood informed subspace

$$S_i = ||\mathcal{P}\mathcal{H}^{-1}\mathcal{B}\boldsymbol{e}_i|| = \sqrt{\sum_{k=1}^{r}\sum_{j=1}^{r}\left(\frac{\boldsymbol{v}_j^T\mathcal{B}\boldsymbol{e}_i}{1+\lambda_j}\right)\left(\frac{\boldsymbol{v}_k^T\mathcal{B}\boldsymbol{e}_i}{1+\lambda_k}\right)\boldsymbol{v}_k^T W_Z \boldsymbol{v}_j}$$



Automatically check for robustness with respect to rank truncation choice.

Sensitivities



Summary

- Established the Bayesian interpretation of post-optimality sensitivity analysis.
- Addressed ill-conditioning by projecting on likelihood informed subspaces.
- Theoretically justified HDSA when optimization fails to converge.
- Provided strong error bounds establishing the robust of the analysis.
- HDSA gives a robust, interpretable, scalable, and efficient (RISE) way to assess correlations in the joint Bayesian posterior distribution.

Joseph Hart, Sandia National Laboratories
joshart@sandia.gov