Normalizing Flows for Bayesian Experimental Design in Imaging Applications

Rafael Orozco¹, Abhinav Gahlot¹, Peng Chen¹, Mathias Louboutin¹* and Felix J. Herrmann¹

¹ now at Devito Codes

Seismic Laboratory for Imaging and Modeling (SLIM)
Georgia Institute of Technology

Released to public domain under Creative Commons license type BY (https://creativecommons.org/licenses/by/4.0)
Copyright (c) 2024, Rafael Orozco (Georgia Tech)
Gradient optimization of designs w.r.t a normalizing flow loss enables experimental design in problems with…
Gradient optimization of designs w.r.t a normalizing flow loss enables experimental design in problems with...

- large parameter designs (200,000 for medical imaging)
Gradient optimization of designs w.r.t a normalizing flow loss enables experimental design in problems with…

- large parameter designs (200,000 for medical imaging)
- non-linear, expensive forward operators (wave equation for CO2 monitoring).
Takeaways from presentation

1. Exact likelihood evaluation keeps normalizing flows relevant in this diffusion era.

SORA WHO????
Takeaways from presentation

1. Exact likelihood evaluation keeps normalizing flows relevant in this diffusion era.

2. Simulation based inference is a general framework for Bayesian inference and downstream tasks i.e. experimental design.
Bayesian experimental design

How should we collect data $y$ over observable $u$ to inform inference?

$$y = M(u)$$
Bayesian experimental design

How should we collect data $y$ over observable $u$ to inform inference?

$$y = M(u)$$

Bayesians have a powerful answer: “Collect the data that maximizes the information gained” - where information gain is quantified by Kullback-Leibler divergence:

$$\max_{M} D_{KL}(p(x|y) || p(x))$$
Bayesian experimental design

How should we collect data $\mathbf{y}$ over observable $\mathbf{u}$ to inform inference?

$$\mathbf{y} = \mathbf{M}(\mathbf{u})$$

Bayesians have a powerful answer: “Collect the data that maximizes the information gained” - where information gain is quantified by Kullback-Leibler divergence:

$$\max_{\mathbf{M}} D_{KL}(p(\mathbf{x} | \mathbf{y}) \mid | p(\mathbf{x}))$$

**Expected information gain (EIG) averages over all possible $\mathbf{y}$**

$$\max_{\mathbf{M}} \text{EIG}(\mathbf{M}) = \mathbb{E}_{p(\mathbf{y}|\mathbf{M})} \left[ D_{KL}(p(\mathbf{x} | \mathbf{y}) \mid | p(\mathbf{x})) \right]$$
Relation between EIG and posterior likelihood

Maximizing the expected information gain is equivalent to maximizing the expected posterior likelihood

\[
\max_M EIG(M) = \mathbb{E}_{p(y|M)} \left[ D_{KL}(p_\theta(x | y) || p(x)) \right]
\]

\[
= \mathbb{E}_{p(y|M)} \left[ \mathbb{E}_{p(x|y)} \left[ \log p_\theta(x | y) - \log p(x) \right] \right]
\]

\[
= \mathbb{E}_{p(y|M)} \left[ \mathbb{E}_{p(x|y)} [ \log p_\theta(x | y) ] \right]
\]

\[
= \mathbb{E}_{p(x,y|M)} [ \log p_\theta(x | y) ] \quad \text{same as neural posterior objective!}
\]
Relation between EIG and posterior likelihood

Maximizing the expected information gain is equivalent to maximizing the expected posterior likelihood

$$\max_{M} \quad EIG(M) = \mathbb{E}_{p(y|M)} \left[ D_{KL}(p_{\theta}(x | y) \ | \ | p(x)) \right]$$

$$= \mathbb{E}_{p(y|M)} \left[ \mathbb{E}_{p(x|y)} \left[ \log p_{\theta}(x | y) - \log p(x) \right] \right]$$

$$= \mathbb{E}_{p(y|M)} \left[ \mathbb{E}_{p(x|y)} \left[ \log p_{\theta}(x | y) \right] \right]$$

$$= \mathbb{E}_{p(x,y|M)} \left[ \log p_{\theta}(x | y) \right]$$

same as neural posterior objective!

Thus optimizing $M$ under posterior learning objective will increase its EIG
Normalizing flows for posteriors

They learn to sample posterior by maximizing the posterior likelihood under training examples

$$\max_{\theta} \mathbb{E}_{p(x,y)} \left[ \log p_\theta(x \mid y) \right]$$
Normalizing flows for posteriors

They learn to sample posterior by maximizing the posterior likelihood under training examples

\[
\max_{\theta} \mathbb{E}_{p(x,y)} \left[ \log p_\theta(x \mid y) \right]
\]

e.g. Normalizing flows are trained as such

\[
\hat{\theta} = \arg \max_{\theta} \frac{1}{N} \sum_{n=1}^{N} \left( -\|f_\theta(x^{(n)}; y^{(n)})\|_2^2 + \log \left| \det J_{f_\theta} \right| \right)
\]
Proposed method

Prepare posterior learning algorithm as typically:

- use prior samples and forward operator to make training pairs
  \[ \{ x^{(n)}, y^{(n)} \}_{i=1}^{N} \]

Instead of optimizing only network parameters:

\[
\hat{\theta} = \arg \max_{\theta} \frac{1}{N} \sum_{n=1}^{N} \left( -\|f_\theta(x^{(n)}; y^{(n)})\|_2^2 + \log \left| \det J_{f_\theta} \right| \right).
\]
Proposed method

Prepare posterior learning algorithm as typically:

- use prior samples and forward operator to make training pairs
\[ \{ \mathbf{x}^{(n)}, \mathbf{y}^{(n)} \}_{i=1}^{N} \]

Instead of optimizing only network parameters:

\[
\hat{\theta} = \arg \max_{\theta} \frac{1}{N} \sum_{n=1}^{N} \left( -\|f_{\theta}(\mathbf{x}^{(n)}; \mathbf{y}^{(n)})\|_2^2 + \log \left| \det J_{f_{\theta}} \right| \right).
\]

jointly optimize for design \( \mathbf{M} \) as well:

\[
\hat{\theta}, \hat{\mathbf{M}} = \arg \max_{\theta, \mathbf{M}} \frac{1}{N} \sum_{i=1}^{N} \left( -\|f_{\theta}(\mathbf{x}^{(n)}; \mathbf{M}(\mathbf{y}^{(n)}))\|_2^2 + \log \left| \det J_{f_{\theta}} \right| \right).
\]
Normalizing Flows
Normalizing flows

Likelihood-based generative models that:

‣ have exact likelihood evaluation

‣ scalable memory usage during training (more on this later)

‣ fast sampling
Normalizing flows

Learn distribution by mapping samples to simple distribution.

Mapping needs to be

• differentiable

• invertible

Normalizing flow during training

Learn distribution by mapping samples to Normal distribution.

\[ f_\theta(x) \]

Epoch 1

Epoch 2

Epoch 64

Normalizing flow during training

Learn distribution by mapping samples to Normal distribution.

\[ x \xrightarrow{\theta(x)} z_x \xrightarrow{\sim \mathcal{N}(0, I)} z \xrightarrow{f^{-1}_\theta(z)} \tilde{x} \]

Epoch 1

Epoch 2

Epoch 64

Application: medical imaging
Magnetic Resonance Imaging (MRI)

Established imaging modality for diagnosis in oncology, neurology and the musculoskeletal system.
Magnetic Resonance Imaging (MRI)

Established imaging modality for diagnosis in oncology, neurology and the musculoskeletal system.

Observation process involves magnetic field that captures the spatial frequency and phase of cross-section through patient tissue:

\[
\text{patient} \rightarrow \text{Magnetic pulses} \rightarrow \text{Fourier transform} \rightarrow \text{observations } \mathbb{C}^{m \times n}
\]
Accelerated Magnetic Resonance Imaging (MRI)

Process is lengthy (easily > 30 min), leads to low patient throughput, problems with patient comfort, artifacts from patient motion, and high exam costs.

Situation: accelerating MRI by subsampling data is important but…

all data

selected locations

accelerated data
Accelerated Magnetic Resonance Imaging (MRI)

Process is lengthy (easily > 30 min), leads to low patient throughput, problems with patient comfort, artifacts from patient motion, and high exam costs.

Situation: accelerating MRI by subsampling data is important but…

Problem: which data points should we measure for best image inference?

Solution: experimental design
Due to noise and subsampling the imaging is ill-posed thus best solved with Bayesian framework:

given observation $y$ (acquired w/ our experimental design) the goal is to sample the posterior:

$$\sim p(x | y = \text{image})$$
MRI experimental design with normalizing flows

FASTMRI pairs of high quality images $x^{(i)}$ and fully sampled k-space data $y^{(i)}$:

$\begin{align*}
x^{(1)} & \quad x^{(2)} & \quad x^{(1800)} \\
y^{(1)} & \quad y^{(2)} & \quad y^{(1800)}
\end{align*}$

MRI experimental design with normalizing flows

FASTMRI pairs of high quality images $\mathbf{x}^{(i)}$ and fully sampled k-space data $\mathbf{y}^{(i)}$:

Jointly train normalizing flow and subsampling pattern:

$$
\max_{\theta, \mathbf{M}} \frac{1}{N} \sum_{i=1}^{N} \left( -\frac{1}{2} \| f_{\theta}(\mathbf{x}^{(i)}; \mathbf{A}^\top \mathbf{M} \odot \mathbf{y}^{(i)}) \|_2^2 + \log \left| \text{det} \mathbf{J}_{f_\theta} \right| \right).
$$
MRI experimental design with normalizing flows

FASTMRI pairs of high quality images $\mathbf{x}^{(i)}$ and fully sampled k-space data $\mathbf{y}^{(i)}$:

Jointly train normalizing flow and subsampling pattern:

$$
\max_{\theta, M} \frac{1}{N} \sum_{i=1}^{N} \left( -\frac{1}{2} \| f_\theta(\mathbf{x}^{(i)} ; A^\top M \odot \mathbf{y}^{(i)}) \|^2_2 + \log |\det J_{f_\theta}| \right).
$$

Problem: how do you optimize binary mask?

---

MRI experimental design with normalizing flows

FASTMRI pairs of high quality images $\mathbf{x}^{(i)}$ and fully sampled k-space data $\mathbf{y}^{(i)}$:

Jointly train normalizing flow and subsampling pattern:

$$\max_{\theta, \mathbf{M}} \frac{1}{N} \sum_{i=1}^{N} \left( -\frac{1}{2} \| f_\theta(\mathbf{x}^{(i)}; \mathbf{A}^\top \mathbf{M} \odot \mathbf{y}^{(i)}) \|_2^2 + \log \left| \det J_{f_\theta} \right| \right).$$

Problem: how do you optimize binary mask?

Solution: reinterpret mask as a sampling density.
Instead of optimizing for binary mask $M \in \mathbb{Z}^{m \times n} : M_{i,j} = \{0,1\}$
Instead of optimizing for binary mask $M \in \mathbb{Z}^{m \times n}$: $M_{i,j} = \{0,1\}$

Instead of optimizing for sampling density $w \in \mathbb{R}^{m \times n}$


Sampling density for receiver placement

Instead of optimizing for binary mask $M \in \mathbb{Z}^{m \times n} : \quad M_{i,j} = \{0,1\}$

optimize for sampling density $w \in \mathbb{R}^{m \times n}$

\[ M(w) := 1_{w < u} \]

where $u \sim U(0,1)$.


MRI experimental design with normalizing flows

FASTMRI pairs of high quality images $\mathbf{x}^{(i)}$ and fully sampled k-space data $\mathbf{y}^{(i)}$:

Jointly train normalizing flow and sampling density:

$$
\hat{\theta}, \hat{w} = \arg\max_{\theta, w} \frac{1}{N} \sum_{i=1}^{N} \left( -\frac{1}{2} \| f_{\theta}(\mathbf{x}^{(i)}; \mathbf{A}^\top \mathbf{M}(w) \odot \mathbf{y}^{(i)}) \|_2^2 + \log \left| \det J_{f_{\theta}} \right| \right).
$$

MRI experimental design with normalizing flows

FastMRI pairs of high quality images $x^{(i)}$ and fully sampled k-space data $y^{(i)}$:

Jointly train normalizing flow and sampling density:

$$\hat{\theta}, \hat{w} = \arg\max_{\theta, w} \frac{1}{N} \sum_{i=1}^{N} \left( -\frac{1}{2} \| f_\theta(x^{(i)}; A^\top M(w) \odot y^{(i)}) \|_2^2 + \log \left| \det J_{f_\theta} \right| \right).$$

Binarize during training and enforce budget $s = 0.025$

$$M(w) := 1_{s \frac{w}{\mathbb{E}[w]} < u}$$

where $u \sim U(0,1)$.

Optimized experimental design

optimal density

optimal binary

baseline

full data
We conclude our optimized density is:

1. centered  -> prioritizes low frequencies
2. ellipsoid  -> prioritizes vertical elements in k-space
3. asymmetric -> learns to exploit Hermitian symmetry
Posterior sampling w/ optimal design

Baseline posterior samples:
Posterior sampling w/ optimal design

Baseline posterior samples:

Our posterior samples w/ optimal design:
Posterior sampling w/ optimal design

Baseline posterior samples:

Reference image

Our posterior samples w/ optimal design:
Posterior statistics

Fast sampling w/ normalizing flow to efficiently estimate statistical moments i.e. mean, standard deviation:

- Mean SSIM=0.57
- Standard deviation
- Error NMSE=0.105

Reference image
Posterior statistics

Fast sampling w/ normalizing flow to efficiently estimate statistical moments i.e. mean, standard deviation:

Mean SSIM=0.57
Standard deviation
Error NMSE=0.105

Reference image
Mean SSIM=0.68
Standard deviation
Error NMSE=0.022
Evaluation on leave-out test set

Posterior sampler generalizes to many observations thus can evaluate on many (100) test examples.

uncertainty is reduced
Evaluation on leave-out test set

Posterior sampler generalizes to many observations thus can evaluate on many (100) test examples.

uncertainty is reduced

error is reduced
Note on scalability

Normalizing flows give you crucial memory efficiency for free…

Note on scalability

Normalizing flows give you crucial memory efficiency for free… if you actually take advantage of it.

Application: monitoring carbon dioxide for mitigating climate change
Forecasts say it is not enough to reduce CO$_2$ emissions. We need to have negative CO$_2$ emissions i.e. take out CO$_2$ already in atmosphere...

but where do we store it?

Underground carbon dioxide storage

Demonstrated solution for large scale storage

- subsurface structures create natural barriers
- long term solution - CO₂ chemically seals into rock at geological time scales

but the plume is not stationary…

Carbon dioxide monitoring

$CO_2$ plume evolves over time due to injection and permeability effects
Carbon dioxide monitoring

CO₂ plume evolves over time due to injection and permeability effects thus monitoring plume is important to:

- prevent leakage
- avoid “seismic events”
- stay in licensed area.
Carbon dioxide monitoring

Two types of time-lapse CO$_2$ plume observations

- direct but local – borehole wells
- indirect but global – seismic
Optimal well locations

$\text{CO}_2$ project lasts years thus can drill more wells but:

- many location options
- expensive (1 million dollars - 100 million dollars)
Optimal well locations

CO$_2$ project lasts years thus can drill more wells but:

- many location options
- expensive (1 million dollars - 100 million dollars)

Operators deciding well locations should be informed by

- current knowledge of the CO$_2$ plumes (prior)
- physics simulations of plume forecasts (likelihood)
Optimal well locations

Optimize for probability *density* of well placement

- well budget agnostic
  - decide number of wells post-hoc
- easier optimization
  - stochastic sampling during training avoids local minima
Small module in full-stack digital twin

MS189
Uncertainty Quantification for Digital Twins - Part III of III

8:30 AM - 10:30 AM
Room: San Giusto - Hotel Savoia Excelsior Palace

For Part II, see MS169

A digital twin (DT) is a computational system that continuously and repeatedly assimilates observed otherwise guides decisions, using predictions from the updated model. Often DTs are employed from models to decisions. The resulting data assimilation and optimal control/decision problems must be tractable for large-scale complex systems. This minisymposium addresses mathematical, statistical analysis, and optimal experimental design subproblems, as well as the reduced order models and s

Organizer: Nicole Aretz
University of Texas at Austin, U.S.

Omar Ghattas
University of Texas at Austin, U.S.

Youssef M. Marzouk
Massachusetts Institute of Technology, U.S.

8:30-8:55 An Uncertainty-Aware Digital Twin for Geological Carbon Storage abstract
Felix Herrmann and Abhinav Gahlot, Georgia Institute of Technology, U.S.
CO2 storage project life cycle

Prior samples $p(x_i)$

Fluid flow simulations
CO2 storage project life cycle

Prior samples $p(x_t)$

Fluid flow simulations

Forecasted plumes $p(x_{t+1} | x_t)$
CO2 storage project life cycle

Prior samples \( p(x_t) \) → Fluid flow simulations → Forecasted plumes \( p(x_{t+1} | x_t) \) → Synthetic observations → Train inference network and well design using pairs \( p(x_{t+1}, y_{t+1}) \)
CO2 storage project life cycle

Prior samples $p(x_t)$  
Fluid flow simulations  
Forecasted plumes $p(x_{t+1} | x_t)$  
Synthetic observations  
Train inference network and well design using pairs $p(x_{t+1}, y_{t+1})$

Outputs: posterior sampler $p_\theta(x_{t+1} | y_{t+1})$ and optimal well density
CO2 storage project life cycle

Prior samples $p(x_t)$ → Fluid flow simulations → Forecasted plumes $p(x_{t+1} | x_t)$ → Synthetic observations → Train inference network and well design using pairs $p(x_{t+1}, y_{t+1})$

Outputs: posterior sampler $p_\theta(x_{t+1} | y_{t+1})$ and optimal well density

Drill well using optimal well density
CO2 storage project life cycle

Prior samples $p(x_t)$

Fluid flow simulations

Forecasted plumes $p(x_{t+1}|x_t)$

Synthetic observations

Collect field data $y_{t+1}^{obs}$ w/ optimal well

Outputs: posterior sampler $p_\theta(x_{t+1}|y_{t+1})$ and optimal well density

Drill well using optimal well density

Train inference network and well design using pairs $p(x_{t+1}, y_{t+1})$
CO2 storage project life cycle

1. **Prior samples** $p(x_t)$
   - Fluid flow simulations
   - Forecasted plumes $p(x_{t+1} | x_t)$
   - Synthetic observations
   - Collect field data $y_{t+1}^{obs}$ w/ optimal well
   - Inference from field data $p_\theta(x_{t+1} | y_{t+1}^{obs})$
   - Posterior inference
   - Field observation
   - Drill well using optimal well density

   *Outputs: posterior sampler $p_\theta(x_{t+1} | y_{t+1})$ and optimal well density*
CO2 storage project life cycle

Prior samples $p(x_t)$

Fluid flow simulations

Forecasted plumes $p(x_{t+1} | x_t)$

Collect field data $y_{t+1}^{obs}$ w/ optimal well

Posterior becomes prior and recurse

Inference from field data $p_{\theta}(x_{t+1} | y_{t+1}^{obs})$

Posterior inference

Synthetic observations

Outputs: posterior sampler $p_{\theta}(x_{t+1} | y_{t+1})$ and optimal well density

Train inference network and well design using pairs $p(x_{t+1}^{obs}, y_{t+1})$

Drill well using optimal well density

Field observation

Forecasted plumes $p(x_{t+1} | x_t)$
Monitor 1
Monitor 2

[Graphs and images related to well density and placement over distance.]
Monitor 4
Monitor 1

ground-truth CO₂

inference mean

inference error

inference variance
Monitor 2

ground-truth CO₂

inference mean

inference error

inference variance
Monitor 3

ground-truth CO₂

inference mean

inference error

inference variance
Improvement on baseline

Our algorithm places wells at optimal locations as measured by error.
Conclusions

Gradient optimization of designs w.r.t a normalizing flow loss enables experimental design in realistic problems:

MRI

optimal design    posterior samples

Underground CO2 monitoring

optimal design

posterior variance
Conclusions

Gradient optimization of designs w.r.t a normalizing flow loss enables experimental design in realistic problems:

and possible because normalizing flows have exact likelihood evaluation.
Conclusions

Gradient optimization of designs w.r.t a normalizing flow loss enables experimental design in realistic problems:

and possible because normalizing flows have exact likelihood evaluation.