Deep Posterior Sampling

Jonas Adler^{1, 2} Ozan Öktem¹

¹Department of Mathematics KTH - Royal Institute of Technology, Stockholm, Sweden

²Research and Physics Elekta, Stockholm, Sweden





Bayesian Inversion

Inverse problem (Statistical viewpoint)

Data $y \in Y$ is a single observation generated by Y-valued random variable **y** where

 $\mathbf{y} = \mathcal{A}(\mathbf{x}) + \mathbf{e}.$

Solution: A probability distribution on model parameter space X

$$\mathbb{P}(\mathsf{x} \mid \mathsf{y} = y)$$

This is a full characterization of the reconstruction, including uncertainty. We don't need to select estimators (e.g. task adapted becomes irrelevant).

Theoretical results:

- The posterior (almost) always exists
- The mapping

$$y \to \mathbb{P}(\mathsf{x} \mid \mathsf{y} = y)$$

is continuous.

• We can characterize convergence (Bernstein-von Mises)

• Bayes Law:

$$\mathbb{P}(x \mid y) = rac{\mathbb{P}(y \mid x)\mathbb{P}(x)}{\mathbb{P}(y)}$$

• We know the data likelihood

 $\mathbb{P}(y \mid x)$

• Only have to specify the prior

 $\mathbb{P}(x)$

• Standard approach: Gibbs priors

$$\mathbb{P}(x) = e^{-S(x)}$$

Jonas Adler jonasadler.com

Bayesian Inversion: Samples



Bayesian Inversion: Samples

Bayesian Inversion: Examples of natural images

Jonas Adler jonasadler.com

- Framework for solving inverse problems
- Strong regularizing properties
- Uncertainty quantification
- Basically parameter free
- Classical methods are relatively slow and require closed form prior

Deep posterior sampling

Generative models for uncertainty quantification in inverse problems

Bayesian Inversion: Hopes and dreams

What would we do if we had $\mathbb{P}(\mathbf{x} \mid \mathbf{y} = y)$?

• Variance

$$\mathbb{E}\Big[\big(\mathbf{x} - \mathbb{E}[\mathbf{x} \mid \mathbf{y} = y]\big)^2 \mid \mathbf{y} = y\Big]$$

• Covariance

$$\mathbb{E}\Big[\big(\mathbf{x}_1 - \mathbb{E}[\mathbf{x}_1 \mid \mathbf{y} = y]\big)\big(\mathbf{x}_2 - \mathbb{E}[\mathbf{x}_2 \mid \mathbf{y} = y]\big) \mid \mathbf{y} = y\Big]$$

• Bayesian hypothesis testing

$$\mathbb{P}(\mathbf{x} \in \Omega \mid \mathbf{y} = y) = \mathbb{E}\Big[\mathbbm{1}_{\Omega}(\mathbf{x}) \mid \mathbf{y} = y\Big]$$

Direct Estimation

• The quantities we're looking for have the form

$$\mathbb{E}\Big[\mathbf{w} \mid \mathbf{y} = y\Big]$$

• The quantities we're looking for have the form

$$\mathbb{E}\Big[\mathbf{w} \mid \mathbf{y} = y\Big]$$

Theorem (Conditional Mean)

Assume that Y is a measurable space, W a measurable Hilbert space, and \mathbf{y} and \mathbf{w} are Y- and W-valued random variables, respectively. Let

$$h^* = \operatorname*{arg\,min}_{h: Y \to W} \mathbb{E} \Big[\big\| h(\mathbf{y}) - \mathbf{w} \big\|_W^2 \Big].$$

Then $h^*(y) := \mathbb{E}[\mathbf{w} \mid \mathbf{y} = y].$

• The quantities we're looking for have the form

$$\mathbb{E}\Big[\mathbf{w} \mid \mathbf{y} = y\Big]$$

Theorem (Conditional Mean)

Assume that Y is a measurable space, W a measurable Hilbert space, and \mathbf{y} and \mathbf{w} are Y- and W-valued random variables, respectively. Let

$$h^* = \operatorname*{arg\,min}_{h: Y \to W} \mathbb{E} \Big[\big\| h(\mathbf{y}) - \mathbf{w} \big\|_W^2 \Big].$$

Then $h^*(y) := \mathbb{E}[\mathbf{w} \mid \mathbf{y} = y].$

• Deep learning can compute basically any estimator

• The quantities we're looking for have the form

$$\mathbb{E}\Big[\mathbf{w} \mid \mathbf{y} = y\Big]$$

Theorem (Conditional Mean)

Assume that Y is a measurable space, W a measurable Hilbert space, and \mathbf{y} and \mathbf{w} are Y- and W-valued random variables, respectively. Let

$$h^* = \operatorname*{arg\,min}_{h: Y \to W} \mathbb{E} \Big[\big\| h(\mathbf{y}) - \mathbf{w} \big\|_W^2 \Big].$$

Then $h^*(y) := \mathbb{E}[\mathbf{w} \mid \mathbf{y} = y].$

- Deep learning can compute basically any estimator
- But we need to train a network for each

Jonas Adler jonasadler.com

Deep Posterior Sampling: The main insight

• The quantities we're looking for have the form

$$\mathbb{E}\Big[\mathbf{w} \mid \mathbf{y} = y\Big]$$

- Mean (reconstruction): $\mathbf{w} = \mathbf{x}$
- Variance: $\mathbf{w} = (\mathbf{x} \mathbb{E}[\mathbf{x} \mid \mathbf{y} = y])^2$
- Hypothesis: $\mathbf{w} = \mathbb{1}_{\mathbf{x}_1 > \mathbf{x}_2}$
- Law of large numbers: Assume w_i I.I.D. from $\mathbf{w} \mid \mathbf{y} = y$, then a.s.

$$\lim_{N \to \infty} \frac{1}{N} \sum_{i=1}^{N} w_i \to \mathbb{E}\Big[\mathbf{w} \mid \mathbf{y} = y\Big]$$

• All we need is I.I.D. samples!

Jonas Adler jonasadler.com

Input: Training data (x_i) generated by (\mathbf{x}) .

Goal: Sample from unknown distribution $\mathbb{P}(x)$.

```
Input: Training data (x_i) generated by (\mathbf{x}).
```

Goal: Sample from unknown distribution $\mathbb{P}(x)$.

Approaches:

- Variational Auto-Encoders
- Plug and Play Generative Networks
- Pixel Recurrent Models
- Generative Adversarial Networks

Generative Advesarial Networks

- Main idea: train two networks, generator G and discriminator D
- Generator tries to generate "true" samples, discriminator tries to say "good/bad"

Generative Advesarial Networks

- Main idea: train two networks, generator G and discriminator D
- Generator tries to generate "true" samples, discriminator tries to say "good/bad"

Goal: Sample from unknown distribution $\mathbb{P}(x)$.

Approach: Learn how to sample from distribution by solving

 $\min_{\theta} \ \mathcal{W}(\mathsf{G}_{\theta}, \mathbb{P}(\mathbf{x}))$

Goal: Sample from unknown distribution $\mathbb{P}(x)$.

Approach: Learn how to sample from distribution by solving

 $\min_{\theta} \mathcal{W}(\mathsf{G}_{\theta}, \mathbb{P}(\mathbf{x}))$

- G_{θ} is a probability distribution on model parameters in X.
- \mathcal{W} is the Wasserstein 1-distance, measures how close G_{θ} is to the distribution.

Goal: Sample from unknown distribution $\mathbb{P}(x)$.

Approach: Learn how to sample from distribution by solving

 $\min_{\theta} \mathcal{W}(\mathsf{G}_{\theta}, \mathbb{P}(\mathsf{x}))$

Unfeasible: Not possible to evaluate \mathcal{W} ($\mathbb{P}(\mathbf{x})$ unknown).

Goal: Sample from unknown distribution $\mathbb{P}(x)$.

Approach: Learn how to sample from distribution by solving

$$\min_{\theta} \left\{ \max_{\mathsf{D} \in Lip(X)} \mathbb{E} \Big[\mathsf{D}(\mathbf{x}) - \mathsf{D}(\mathsf{G}_{\theta}) \Big] \right\}.$$

Unfeasible: Not possible to evaluate \mathcal{W} ($\mathbb{P}(\mathbf{x})$ unknown).

 \implies Re-write using the Kantorovich-Rubinstein dual characterization of \mathcal{W} .

Goal: Sample from unknown distribution $\mathbb{P}(x)$.

Approach: Learn how to sample from distribution by solving

$$\min_{\theta} \left\{ \max_{\mathsf{D} \in Lip(\mathsf{X})} \mathbb{E} \Big[\mathsf{D}(\mathsf{x}) - \mathsf{D}(\mathsf{G}_{\theta}) \Big] \right\}.$$

Unfeasible: Maximization over all Lipschitz operators

Goal: Sample from unknown distribution $\mathbb{P}(x)$.

Approach: Learn how to sample from distribution by solving

$$\min_{\theta} \left\{ \max_{\phi} \mathbb{E} \left[\mathsf{D}_{\phi}(\mathbf{x}) - \mathsf{D}_{\phi}(\mathsf{G}_{\theta}) \right] \right\}$$

Unfeasible: Maximization over *all* Lipschitz operators \implies Let discriminator be a NN.

Goal: Sample from unknown distribution $\mathbb{P}(x)$.

Approach: Learn how to sample from distribution by solving

$$\min_{\theta} \left\{ \max_{\phi} \mathbb{E} \left[\mathsf{D}_{\phi}(\mathbf{x}) - \mathsf{D}_{\phi}(\mathsf{G}_{\theta}) \right] \right\}$$

Unfeasible: How is G_{θ} random?

Goal: Sample from unknown distribution $\mathbb{P}(x)$.

Approach: Learn how to sample from distribution by solving

$$\min_{\theta} \left\{ \max_{\phi} \mathbb{E}_{\mathbf{x}, \mathbf{z}} \Big[\mathsf{D}_{\phi}(\mathbf{x}) - \mathsf{D}_{\phi}(\mathsf{G}_{\theta}(\mathbf{z})) \Big] \right\}.$$

Unfeasible: How is G_{θ} random?

 \implies Write as deterministic function of random input

Goal: Sample from unknown distribution $\mathbb{P}(x)$.

Approach: Learn how to sample from distribution by solving

$$\min_{\theta} \left\{ \max_{\phi} \mathbb{E}_{\mathbf{x},\mathbf{z}} \Big[\mathsf{D}_{\phi}(\mathbf{x}) - \mathsf{D}_{\phi}(\mathsf{G}_{\theta}(\mathbf{z})) \Big] \right\}.$$

Unfeasible: Expectation over samples

Goal: Sample from unknown distribution $\mathbb{P}(x)$.

Approach: Learn how to sample from distribution by solving

$$\min_{\theta} \left\{ \max_{\phi} \left[\frac{1}{N} \sum_{i=1}^{N} \mathsf{D}_{\phi}(x_i) - \mathbb{E}_{\mathbf{z}} \mathsf{D}_{\phi}(\mathsf{G}_{\theta}(\mathbf{z})) \right] \right\}.$$

Goal: Sample from unknown distribution $\mathbb{P}(x)$.

Approach: Learn how to sample from distribution by solving

$$\min_{\theta} \left\{ \max_{\phi} \left[\frac{1}{N} \sum_{i=1}^{N} \mathsf{D}_{\phi}(x_i) - \mathbb{E}_{\mathbf{z}} \mathsf{D}_{\phi}(\mathsf{G}_{\theta}(\mathbf{z})) \right] \right\}.$$

Approximation to Wasserstein distance useful for deep learning

Goal: Sample from unknown posterior $\mathbb{P}(x \mid y)$.

Approach: Learn how to sample from posterior by solving

$$\min_{\theta} \mathbb{E}_{\mathbf{y} \sim \mathbb{P}_{data}} \Big[\mathcal{W} \big(\mathsf{G}_{\theta}(\mathbf{y}), \mathbb{P}(\mathbf{x} \mid \mathbf{y}) \big) \Big].$$

Goal: Sample from unknown posterior $\mathbb{P}(x \mid y)$.

Approach: Learn how to sample from posterior by solving

$$\min_{\theta} \mathbb{E}_{\mathbf{y} \sim \mathbb{P}_{data}} \Big[\mathcal{W} \big(\mathsf{G}_{\theta}(\mathbf{y}), \mathbb{P}(\mathbf{x} \mid \mathbf{y}) \big) \Big].$$

Condition on data y, else same steps above (with some technical additions)

Goal: Sample from unknown posterior $\mathbb{P}(x \mid y)$.

Approach: Learn how to sample from posterior by solving

$$\min_{\theta} \left\{ \max_{\phi} \frac{1}{N} \sum_{i=1}^{N} \left[\mathsf{D}_{\phi}(x_i, \mathbf{y}_i) - \mathbb{E}_{\mathbf{z}} \left[\mathsf{D}_{\phi}(\mathsf{G}_{\theta}(\mathbf{z}, \mathbf{y}_i), \mathbf{y}_i) \right] \right] \right\}.$$

Condition on data y, else same steps above (with some technical additions)

Goal: Sample from unknown posterior $\mathbb{P}(x \mid y)$.

Approach: Learn how to sample from posterior by solving

$$\min_{\theta} \left\{ \max_{\phi} \frac{1}{N} \sum_{i=1}^{N} \left[\mathsf{D}_{\phi}(x_i, y_i) - \mathbb{E}_{\mathbf{z}} \left[\mathsf{D}_{\phi}(\mathsf{G}_{\theta}(\mathbf{z}, y_i), y_i) \right] \right] \right\}.$$

Formulation useful for deep learning

One of the images is the ground truth (phantom), can you figure out which one?

One of the images is the ground truth (phantom), can you figure out which one?

FBP

2

Conditional mean

One of the images is the ground truth (phantom), can you figure out which one?

FBP

Deep posterior sample

Phantom

Total variation

Conditional mean

Data

FBP

- Case: Patient with suspected metastasis to the liver.
- Data: Clinical helical 3D CT data, 2% of a normal dose (ultra low-dose).

Data

FBP

Posterior mean

- Case: Patient with suspected metastasis to the liver.
- Data: Clinical helical 3D CT data, 2% of a normal dose (ultra low-dose).

Data

FBP

Standard deviation

- Case: Patient with suspected metastasis to the liver.
- Data: Clinical helical 3D CT data, 2% of a normal dose (ultra low-dose).

Posterior mean with ROI

- Case: Patient with suspected metastasis to the liver.
- Data: Clinical helical 3D CT data, 2% of a normal dose (ultra low-dose).
- Liver lesion: $\bigtriangleup =$ difference in average contrast between ROI and liver.
- $\bullet\,$ Hypothesis test: Based on 1000 samples, the ROI contains a lesion at 95%

- Case: Patient with suspected metastasis to the liver.
- Data: Clinical helical 3D CT data, 2% of a normal dose (ultra low-dose).
- Liver lesion: $\bigtriangleup =$ difference in average contrast between ROI and liver.
- $\bullet\,$ Hypothesis test: Based on 1000 samples, the ROI contains a lesion at 95%

Normal dose image

Histogram of \triangle

Posterior mean with ROI

- Case: Patient with suspected metastasis to the liver.
- Data: Clinical helical 3D CT data, 2% of a normal dose (ultra low-dose).
- \bullet Liver lesion: $\bigtriangleup =$ difference in average contrast between ROI and liver.
- $\bullet\,$ Hypothesis test: Based on 1000 samples, the ROI contains a lesion at 95%

- Bayesian Inversion is an extremely powerful framework
- Historical problems with computational feasibility and unknown prior
- Deep Learning methods allow us to compute any estimator quickly and with the "true" prior

- Theory and methods for machine learning in image reconstruction.
- We've got the worlds first clinical photon counting spectral-CT data.
- Very nice position (great group, travel, salary)
- Pursued jointly with MedTechLabs and the Medical Imaging group at KTH led by Mats Danielsson.

Thank you for your attention!