Conditional Sampling with Monotone GANs

Nikola B. Kovachki

Computing and Mathematical Sciences
California Institute of Technology

Second Symposium on Machine Learning and Dynamical Systems
September 28th, 2020

Conditional Sampling with Monotone GANs

arXiv:2006.06755

Nikola B. Kovachki
Computing and Mathematical Sciences
California Institute of Technology
Pasadena, CA 91125
nkovachki@caltech.edu

Ricardo Baptista
Aeronautics and Astronautics
Massachusetts Institute of Technology
Cambridge, MA 02139
rsb@mit.edu

Bamdad Hosseini
Computing and Mathematical Sciences
California Institute of Technology
Pasadena, CA 91125
bamdadh@caltech.edu

Youssef Marzouk
Aeronautics and Astronautics
Massachusetts Institute of Technology
Cambridge, MA 02139
ymarz@mit.edu

Table of Contents

Caltech

Problem Setting

Setting

Target measure: $\nu(dx, dy)$ supported on $\mathbb{R}^{n+m}$

Data: $\{x_n, y_n\}_{n=1}^{N} \overset{i.i.d.}{\sim} \nu.$

Goal: sample $\nu(dy|x)$ for any $x \in \operatorname{supp} \nu(dx)$

Approach

Reference measure : $\eta(dy)$ supported on $\mathbb{R}^m$

Approximate: $S(x, \cdot) : \mathbb{R}^m \to \mathbb{R}^m$ s.t.

$$S(x, \cdot)_\sharp \eta(dy) = \nu(dy|x)$$

Key Idea

Construct block-triangular $T : \mathbb{R}^{n+m} \to \mathbb{R}^{n+m}$ s.t. $T_\sharp \eta(dx, dy) = \nu$ and extract S.

Applications

Probabilistic Supervised Learning and UQ

Model: $y = G(x)$, with G stochastic mapping

Statsitics of $y|x$: mean, median, maximal probability points, variance, confidence intervals, error bars, etc.

Does not require knowledge of G, only data.

Bayesian Inverse Problem

Model: $x = G(y)$, with G stochastic mapping

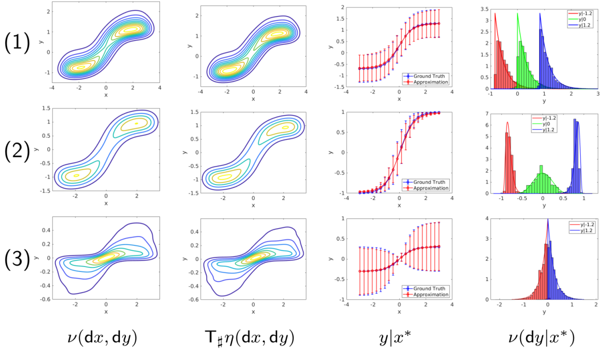
Characterize posterior $y|x$.

Does not require a know prior on parameters, only data.

1D Example

Different noise models

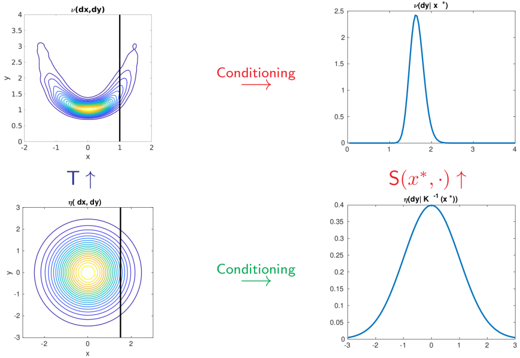
$$y = \tanh(x) + \gamma, \qquad \gamma \sim \Gamma(1, 0.3) \qquad (1)$$
$$y = \tanh(x + \gamma), \qquad \gamma \sim N(0, 0.05) \qquad (2)$$
$$y = \gamma \tanh(x), \qquad \gamma \sim \Gamma(1, 0.3) \qquad (3)$$



Block-triangular map

$$\text{Joint: } T(x,y) = \begin{bmatrix} K(x) \\ S(K(x),y) \end{bmatrix}, \quad K : \mathbb{R}^n \to \mathbb{R}^n, S : \mathbb{R}^{n+m} \to \mathbb{R}^m$$

$$\text{Conditioning: } S(x,y)$$



Block-triangular measure transport

Properties

- Block-triangular mappings exist under mild conditions (e.g. ν and η are absolutely continuous without atoms).
- Block-triangular maps can be constructed explicitly (e.g. the Knothe-Rosenblatt rearrangement).

Optimization

$$\begin{aligned}
\text{Objective:}\quad & T = \operatorname*{arg\,min}_{T \in \mathcal{T}} D(T_\sharp \eta(dx, dy) || \nu(dx, dy)) \\
\text{Statistical Divergence:}\quad & D : \mathcal{P}(\mathbb{R}^{n+m}) \times \mathcal{P}(\mathbb{R}^{n+m}) \to \mathbb{R}_+ \\
& \text{s.t. } D(\mu, \nu) = 0 \text{ iff } \mu = \nu \\
\text{Approximation Space:}\quad & \mathcal{T} = \{\text{block-triangular, continuous, monotone maps}\}
\end{aligned}$$

Block-triangular measure transport

$$T^* = \arg\min_{T \in \mathcal{T}} D(T_\sharp \eta(dx, dy) || \nu(dx, dy)) \qquad (1)$$

Theorem [KBHM 20]

Let T^* be given by (1) and suppose it has the form

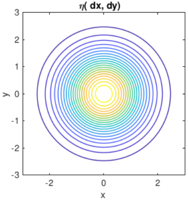
$$T^*(x, y) = \begin{bmatrix} K^*(x) \\ S^*(K^*(x), y) \end{bmatrix}.$$

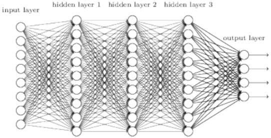
If T^* is surjective then for any $x \in \operatorname{supp} \nu(dx)$, we have that

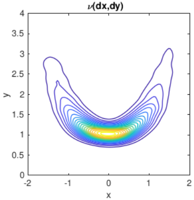
$$S^*(x, \cdot)_\sharp \eta(dy) = \nu(dy|x).$$

- By Browder-Minty theorem: continuity + monotonicity + coercivity $\implies$ surjectivity.

GANs


η(dx, dy)
3
2
1
0
-1
-2
-3
-2
0
2
x
y


input layer
hidden layer 1
hidden layer 2
hidden layer 3
output layer


ν(dx,dy)
4
3.5
3
2.5
2
1.5
1
0.5
0
-2
-1
0
1
2
x
y

The GAN "divergence"

Objective: $T^* = \arg\min_{T \in \mathcal{T}} D_{\text{GAN}}(T_\sharp \eta || \nu)$

Divergence: $D_{\text{GAN}}(T_\sharp \eta || \nu) = \sup_{f \in C(\mathbb{R}^{n+m};[0,1])} \mathbb{E}_{z \sim \nu} \log f(z) + \mathbb{E}_{w \sim \eta} \log(1 - f(T(w)))$

Approximation Space

Choose $\mathcal{T}$ as the set of continuous, strictly monotone, block-triangular mappings:

$$\begin{cases} T(x,y) = \begin{bmatrix} K(x) \\ S(K(x),y) \end{bmatrix} \in C(\mathbb{R}^{n+m};\mathbb{R}^{n+m}) \\ \langle T(w) - T(w'), w - w' \rangle > 0, \quad \eta - \text{a.e.} \end{cases}$$

Discretize

Data: $\{z_j = (x_j, y_j)\}_{j=1}^N \overset{i.i.d.}{\sim} \nu$

Reference samples: $\{w_k\}_{k=1}^{2J} \overset{i.i.d.}{\sim} \eta$

Neural Networks: K, S, f

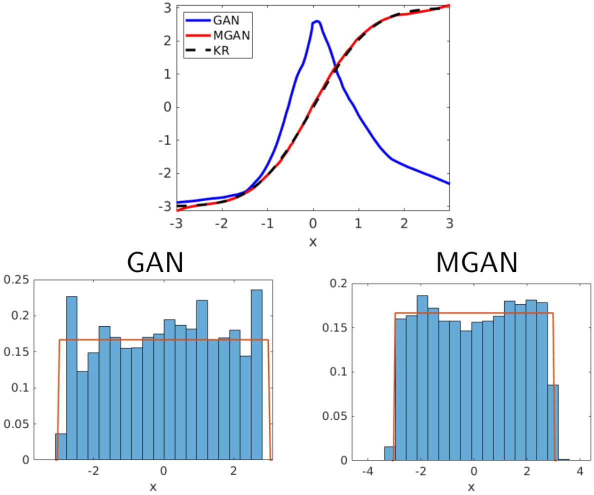
Divergence: $$\min_f \frac{1}{N}\sum_{j=1}^N \log f(z_j) + \frac{1}{2J}\sum_{k=1}^{2J} \log(1 - f(T(w_k)))$$

Average Monotonicity: $$\frac{1}{J}\sum_{k=1}^{J} \langle T(w_k) - T(w_{k+J}), w_k - w_{k+J} \rangle > 0$$

Caltech

GANs: Monotonicity

- $\nu = U(-3,3)$, $\eta = N(0,1)$



GANs: Sampling and UQ

$$T^* = \arg\min_{T \in \mathcal{T}} D_{\mathrm{GAN}}(T_\sharp \eta || \nu)$$

- Extract $S^*(x, y)$ from $T^*(x, y)$.
- New input $x^\dagger \in \mathbb{R}^n$.
- Generate new samples from reference marginal $\tilde{y}_k \overset{i.i.d.}{\sim} \eta(dy)$.
- Set $y_k = S(x^\dagger, \tilde{y}_k)$ then $y_k \sim \nu(dy|x^\dagger)$.
- Use y_k to compute statistics of $\nu(dy|x^\dagger)$.

Example: Blood Oxygen Demand Model

Model

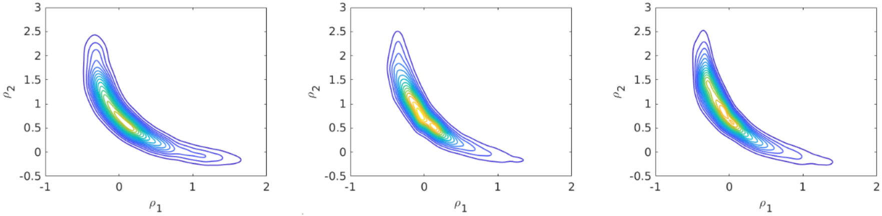
$$\begin{aligned} \text{Model:} \quad & B(t) = A(1 - e^{-Bt}) + \gamma \\ \text{Prior:} \quad & A \sim U(0.4, 1, 2), B \sim U(0.01, 0.31) \\ \text{Noise:} \quad & \gamma \sim N(0, 10^{-3}) \end{aligned}$$

Data

$$\begin{aligned} \text{Without prior:} \quad & x = (B(1), B(2), \dots, B(5)), y = (A, B) \\ \text{With prior:} \quad & x = (B(1), B(2), \dots, B(5)), \\ & y = \left(\rho_1 = \sqrt{2}\mathrm{erf}^{-1}\left(\frac{A - 0.4}{0.4} - 1\right), \rho_2 = \sqrt{2}\mathrm{erf}^{-1}\left(\frac{A - 0.01}{0.15} - 1\right)\right) \\ & \text{s.t. } (\rho_1, \rho_2) \sim N(0, I_2) \end{aligned}$$

Example: Blood Oxygen Demand Model

Map type	N	Mean		Variance		Skewness		Kurtosis	
		ρ_1	ρ_2	ρ_1	ρ_2	ρ_1	ρ_2	ρ_1	ρ_2
MCMC "truth"		0.075	0.875	0.190	0.397	1.935	0.681	8.537	3.437
[30]	5000	0.090	**0.908**	0.180	0.490	2.968	**0.707**	29.589	16.303
	50000	0.034	0.902	0.206	0.457	1.628	0.872	7.568	3.876
MGAN (without prior)	5000	**0.075**	0.762	**0.193**	0.331	1.079	0.451	3.980	**3.048**
	50000	**0.071**	0.843	**0.196**	0.358	**2.003**	**0.538**	7.961	3.246
MGAN (with prior)	5000	0.084	0.805	0.278	**0.402**	**1.947**	0.805	**8.122**	4.363
	50000	0.038	**0.882**	0.205	**0.397**	2.132	0.490	**8.9747**	**3.408**



(a) MCMC (b) Without prior (c) With prior

Example: Darcy Flow

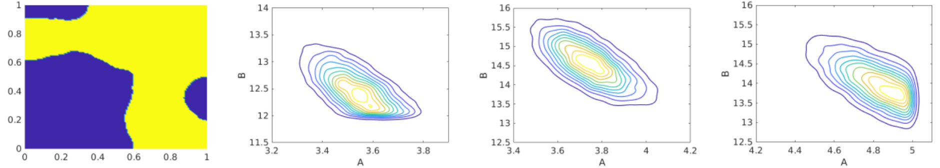
Model

$$\begin{aligned} \text{Model:} \quad & -\nabla \cdot (a \nabla p) = f \\ & a(s) = A\mathbb{1}_{\Omega_A} + B\mathbb{1}_{\Omega_B} \\ \text{Prior:} \quad & A \sim U(3,5), B \sim U(12,16) \\ \text{Noise (additive):} \quad & \gamma \sim N(0, 10^{-7} I_{16}) \end{aligned}$$

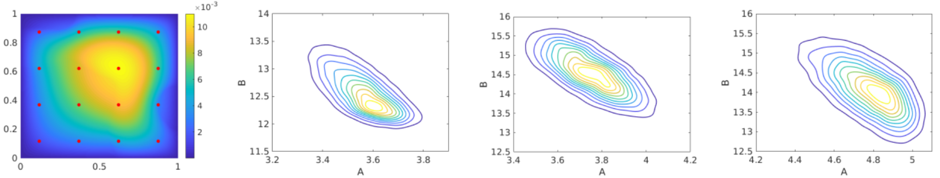
Data

$$\text{Without prior:} \quad x = (p(1), p(2), \ldots, p(16)), y = (A, B)$$

Example: Darcy Flow



(a) Ω_A, Ω_B (b) MCMC $\boldsymbol{y}|\boldsymbol{x}_1^*$ (c) MCMC $\boldsymbol{y}|\boldsymbol{x}_2^*$ (d) MCMC $\boldsymbol{y}|\boldsymbol{x}_3^*$



(e) $\boldsymbol{x}_2^*$ (f) $\mathsf{F}_\sharp(\boldsymbol{x}_1^*, \cdot)\eta(\boldsymbol{y})$ (g) $\mathsf{F}_\sharp(\boldsymbol{x}_2^*, \cdot)\eta(\boldsymbol{y})$ (h) $\mathsf{F}_\sharp(\boldsymbol{x}_3^*, \cdot)\eta(\boldsymbol{y})$

Example: MNIST

Caltech

Conditional Sampling

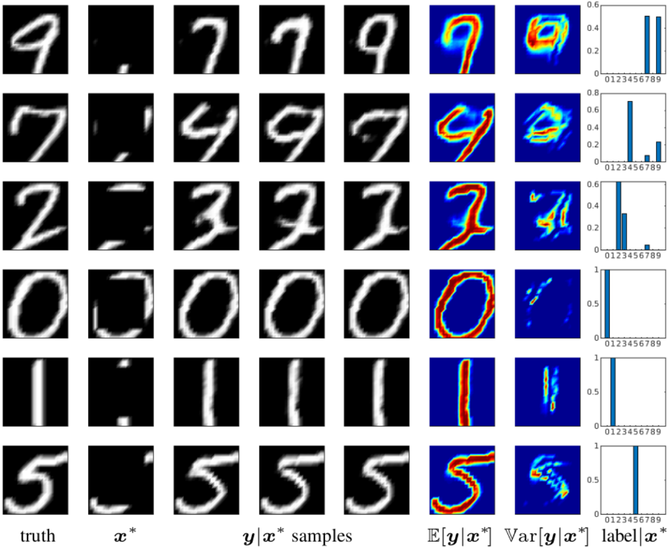
N.B.K.

Problem Setting

Block-triangular measure transport

GANs

Numerics



Conclusion

- A model agnostic method for probabilistic supervised learning.
- Applications to Bayesian inverse problems.
- Using measure transport for conditioning.
- A straightforward extension of GAN framework.
- Future extensions:
  - Generalize to infinite dimensions.
  - Analysis: approximations of T and the pushforward $T_\sharp \eta$, properties of the minimizer, sample sizes, choice of divergence, etc.