Conditional Sampling with Block-Triangular Transport Maps

Bamdad Hosseini
Department of Applied Mathematics
University of Washington

Kantorovich Initiative Seminar

October 21, 2021

Collaborators

"Conditional Sampling with Monotone GANs"

arXiv:2006.06755

- Nikola B Kovachki, Caltech.
- Ricardo Baptista, MIT.
- Youssef Marzouk, MIT.

Outline

Outline

Supervised learning and inverse problems

Deterministic supervised learning

- Input x and output y.
- Training data (x_j, y_j).
- Given new x^* predict $y^*(x^*)$.
- Assume input-output map $\Psi^\dagger$ so that $y = \Psi^\dagger(x)$.
- Typically, approximate $\Psi^\dagger$ with Ψ^* so that $y^* \approx \Psi^*(x^*)$.

Inverse problems

- Unknown parameter y and indirect measurement x so that $\mathcal{F}^\dagger(y) = x$.
- Given new measurement x^* find $y^* \approx \left(\mathcal{F}^\dagger\right)^{-1}(x^*)$.
- Typically, solve optimization problem to find minimal norm solution.

Supervised learning and inverse problems

Probabilistic supervised learning

- Training data (x_j, y_j).
- Joint measure $\nu(\mathrm{d}x, \mathrm{d}y)$ s.t. $(x_j, y_j) \sim \nu$.
- Given new x^* identify/sample $\nu(\mathrm{d}y|x^*)$.

Bayesian inverse problems

- Measurement and parameter pairs (x, y).
- Prior measure $\nu_0(\mathrm{d}y)$ on parameter.
- Joint measure $\nu(\mathrm{d}x, \mathrm{d}y) = \mathrm{Law}\{(x, y) : x = \mathcal{F}^\dagger(y),\ y \sim \nu_0\}$.
- Identify posterior $\nu(\mathrm{d}y|x^*)$.

[1] F Gressmann et al. “Probabilistic supervised learning”. In: *arXiv:1801.00753* (2019).

[2] Andrew M Stuart. “Inverse problems: a Bayesian perspective”. In: *Acta numerica* 19 (2010), pp. 451–559.

Generative modeling and measure transport

Generative modeling

- Given training data $z_i \sim \nu(\mathrm{d}z)$.
- Generate new samples z_j^* so that $z_j^* \sim \nu(\mathrm{d}z)$.
- Choose reference measure η and train map T so that $\mathsf{T}_\sharp \eta \approx \nu$.
- Draw $s_j \sim \eta$ and set $z_j^* = \mathsf{T}(s_j)$.

Measure transport

- Given measures η, ν find mapping T so that $\mathsf{T}_\sharp \eta = \nu$.
  - Optimal transport.
  - Image registration.
  - Triangular transformations.

[3] C. Villani. *Optimal transport: Old and new*. Springer, 2009.

[4] L. Younes. *Shapes and diffeomorphisms*. Springer, 2010.

[5] VI Bogachev, AV Kolesnikov, and KV Medvedev. "Triangular transformations of measures". In: *Sbornik: Mathematics* 196.3 (2005), p. 309.

Measure transport and conditioning

The missing link

Generative modeling

Measure transport

Conditional sampling

Supervised learning/inverse problems

Data-driven supervised learning/inversion with UQ.

[6] Gregory Ongie et al. "Deep learning techniques for inverse problems in imaging". In: *IEEE Journal on Selected Areas in Information Theory* 1.1 (2020), pp. 39–56.

Measure transport approach to conditioning

- Joint measure $\nu \in \mathbf{P}(\mathcal{X} \times \mathcal{Y})$.
- Reference measure $\eta_{\mathcal{Y}} \in \mathbf{P}(\mathcal{Y})$.
- **Goal:** Find transport map $\mathsf{S} : \mathcal{X} \times \mathcal{Y} \to \mathcal{Y}$ s.t.

$$\mathsf{S}(x^*, \cdot)_\sharp \eta_{\mathcal{Y}} = \nu(\cdot | x^*)$$

for $\nu_{\mathcal{X}}$ a.e. x^*.

Measure transport and conditioning

The Knothe-Rosenblatt map

- Concatenate variables $z = (x, y) \in \mathbb{R}^{m+n}$.
- Knothe-Rosenblatt (KR) map is a triangular example

$$T : \mathbb{R}^{m+n} \to \mathbb{R}^{m+n}, \qquad T(z) = \begin{bmatrix} T^1(z_1) \\ T^2(z_1, z_2) \\ \vdots \\ T^{m+n}(z_1, \ldots, z_{m+n}) \end{bmatrix}.$$

KR map in $\mathbb{R}^2$

- $\eta = h(z)dz, \nu = g(z)dz.$
- Define first marginals $h^1(z_1) = \int h(z_1, z_2)dz_2, \quad g^1(z_1) = \int f(z_1, z_2)dz_2.$
- $T^1(z_1)$ is OT map pushing h^1 to g^1.
- Disintegration/conditioning $h^2_{z_1}(z_2) := \frac{h(z_1,z_2)}{h^1(z_1)}, g^2_{z_1}(z_2) := \frac{g(z_1,z_2)}{g^1(z_1)}$.
- $T^2(z_1, z_2)$ is OT map pushing $h^2_{z_1}(z_2)$ to $g^2_{T^1(z_1)}(z_2)$.
- $T^2((T^1)^{-1}(z_1), z_2)$ is the desired conditioning map.

---

[7] F. Santambrogio. *Optimal transport for applied mathematicians*. Springer, 2015.
[8] N. Bonnotte. "From Knothe's rearrangement to Brenier's optimal transport map". In: *SIAM Journal on Mathematical Analysis* 45.1 (2013), pp. 64–87.

Monotone GANs

Overview of our approach

- Input and output spaces $\mathcal{X}, \mathcal{Y}$.
- Joint target measure: $\nu \in \mathbf{P}(\mathcal{X} \times \mathcal{Y})$.
- Reference: $\eta \in \mathbf{P}(\mathcal{X} \times \mathcal{Y})$ (mostly standard normal).
- Empirical measures with N samples ν_N, η_N.
- Block-triangular map:

$$\mathsf{T}(x, y) = \Big(\mathsf{K}(x), \mathsf{S}(\mathsf{K}(x), y)\Big), \qquad \mathsf{K} : \mathcal{X} \to \mathcal{X}, \quad \mathsf{S} : \mathcal{X} \times \mathcal{Y} \to \mathcal{Y}.$$

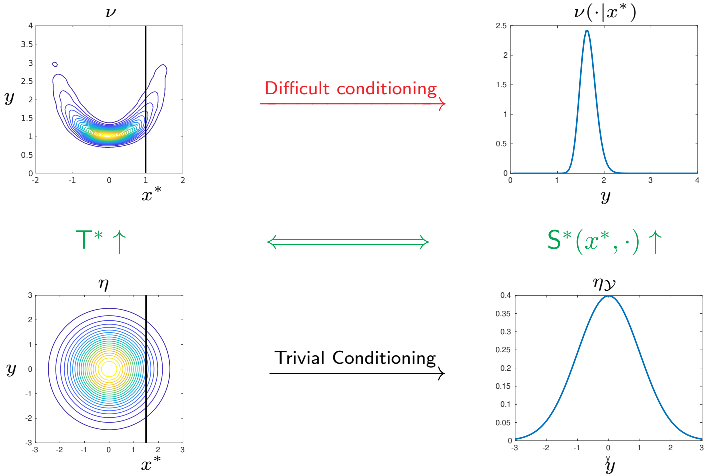
- Neural networks K, S.
- Optimization problem:

$$(\mathsf{K}^*, \mathsf{S}^*) = \underset{\mathsf{K},\mathsf{S}}{\arg\min}\, \text{Divergence}\ (\mathsf{T}_\sharp \eta_N || \nu_N) + \text{ Monotonicity penalty on T.}$$

- Result: $\mathsf{S}^*(x^*, \cdot)_\sharp \eta_{\mathcal{Y}} \approx \nu(\cdot|x^*)$.

Monotone GANs

Overview of our approach



Monotone GANs

Practical motivation

$$\mathsf{T}(x,y) = \Big(\mathsf{K}(x), \mathsf{S}(\mathsf{K}(x),y)\Big), \qquad \mathsf{T}^*_\sharp \eta = \nu.$$

- Probabilistic supervised learning and uncertainty quantification.
  - Compute statistics of $y|x^*$: mean, median, maximal probability points, variance, confidence intervals, error bars etc.
- Model agnostic/likelihood-free inference.
  - Only require knowledge of $\nu(dx,dy), \eta(dx,dy)$ and marginal $\eta_{\mathcal{Y}}(dy)$.
  - No explicit/functional model assumptions on the relationship between output y and input x.
- Conditional sampling with many new inputs.
  - Offline training of the map S^*.
  - Parallel acquisition of training data.
  - Sampling many conditionals.

$$\mathsf{S}^*(x^*_j, \cdot)_\sharp \eta_{\mathcal{Y}} \approx \nu(\cdot | x^*_j).$$

Related literature

GANs, Normalizing flows, VAEs

- Generative adversarial networks (GANs):
  - Goodfellow et al. (2014), "Generative adversarial nets".
  - Nowozin, Cseke, Tomioka (2016), "f-GAN: Training generative neural samplers using variational divergence minimization".
  - Arjovsky, Chintala, Bottou (2017), "Wasserstein GAN".
  - Creswell et al. (2018), "Generative adversarial networks: An overview".
  - Gui et al. (2020), "A review on generative adversarial networks: Algorithms, theory, and applications".
- Normalizing flows (NFs):
  - Tabak and Vanden-Eiden (2010), "Density estimation by dual ascent of the log-likelihood".
  - Moselhy and Marzouk (2012), "Bayesian Inference with Optimal Maps".
  - Rezende and Mohamed (2015), "Variational inference with normalizing flows".
  - Dinh et al, (2015), "NICE: Non-linear Independent Components Estimation"
  - Papamakarios et al. (2019), "Normalizing flows for probabilistic modeling and inference".
  - Kobyzev, Prince, Brubaker (2019), "Normalizing Flows: Introduction and Ideas".
- Variational auto encoders (VAEs):
  - Kingma and Welling (2014), "Auto-encoding variational Bayes".
  - Rezende, Mohamed, Wierstra (2014), "Stochastic back-propagation and approximate inference in deep generative models"
  - Kingma and Welling (2019), "An Introduction to Variational Autoencoders".

Related literature

Neural networks in inverse problems

- Inverse problems:
  - Vogel (2002), "Computational methods for inverse problems".
  - Kaipio and Somersalo (2006), "Statistical and computational inverse problems".
- Data-driven methods for inverse problems:
  - Adler and Öktem (2017), "Solving ill-posed inverse problems using iterative deep neural networks".
  - McCann, Jin and Unser (2017), "Convolutional neural networks for inverse problems in imaging: A review".
  - Lunz, Öktem and Schonlieb (2018), "Adversarial regularizers in inverse problems".
  - Gilton, Ongie and Willett (2019), "Learned patch-based regularization for inverse problems in imaging".
  - Arridge et al. (2019), "Solving inverse problems using data-driven models".
  - **Gottschling et al. (2020), "The troublesome kernel: why deep learning for inverse problems is typically unstable".

Related literature

Generative models and conditional sampling

- Mirza and Osindero (2014), "Conditional generative adversarial nets".
- Sohn, Yan and Lee (2015), "Learning structured output representation using deep conditional generative models"
- *Adler and Öktem (2018), "Deep Bayesian inversion".
- Belghazi et al. (2019), "Learning about an exponential amount of conditional distributions".
- *Whang, Lindgren and Dimakis (2020), "Approximate probabilistic inference with composed flows".
- [KBHM' 20], "Conditional sampling with monotone GANs".

We train a single map (network) that characterizes the conditional $\nu(\cdot|x^*)$ for any new input x^* + supported by theory.

Outline

Block-triangular maps

- Separable Banach spaces $\mathcal{X}, \mathcal{Y}$.
- Block-triangular map:

$$\mathsf{T} : \mathcal{X} \times \mathcal{Y} \to \mathcal{X} \times \mathcal{Y}, \qquad \mathsf{T}(x, y) := \Big(\mathsf{T}^1(x), \mathsf{T}^2(x, y)\Big).$$

- Jacobian of T is block-triangular when $\mathcal{X}, \mathcal{Y}$ are finite-dimensional Euclidean spaces.

Theorem [KBHM]

Suppose

- $\eta = \eta_{\mathcal{X}} \otimes \eta_{\mathcal{Y}} \in \mathbf{P}(\mathcal{X} \times \mathcal{Y})$.
- $\mathsf{T} : \mathcal{X} \times \mathcal{Y} \to \mathcal{X} \times \mathcal{Y}$ is a block-triangular map of the form

$$\mathsf{T}(x, y) = \Big(\mathsf{K}(x), \mathsf{S}(\mathsf{K}(x), y)\Big).$$

If $\mathsf{T}_\sharp \eta = \nu$ then

$$\mathsf{S}(x^*, \cdot)_\sharp \eta_{\mathcal{Y}} = \nu(\cdot | x^*),$$

for $\nu_{\mathcal{X}}$ a.e. x^*.

- Also true for $\mathsf{T} : \mathcal{X}' \times \mathcal{Y}' \to \mathcal{X} \times \mathcal{Y}$ where $\mathcal{X}', \mathcal{Y}'$ are also Banach spaces.

$$\mathsf{T}^* = \arg\min_{\mathsf{T}\in\mathcal{T}} D(\mathsf{T}_\sharp\eta||\nu)$$

- Statistical divergence $D : \mathbf{P}(\mathcal{X}\times\mathcal{Y})\times\mathbf{P}(\mathcal{X}\times\mathcal{Y}) \mapsto \mathbb{R}_+$ s.t. $D(\mu_1||\mu_2) = 0$ iff $\mu_1 = \mu_2$.
- Class of block-triangular maps

$$\mathcal{T} := \left\{\mathsf{T} : \mathcal{X}\times\mathcal{Y} \to \mathcal{X}\times\mathcal{Y} \,\middle|\, \mathsf{T}(x,y) = \Big(\mathsf{K}(x), \mathsf{S}(\mathsf{K}(x), y)\Big)\right\}.$$

Theoretical foundations

Existence of minimizers

$$\mathsf{T}^* = \arg\min_{\mathsf{T}\in\mathcal{T}} D(\mathsf{T}_\sharp \eta || \nu)$$

Theorem [KBHM]

Suppose

- $\eta = \eta_{\mathcal{X}} \otimes \eta_{\mathcal{Y}}$.
- $\eta_{\mathcal{X}}$ and $\eta_{\mathcal{Y}}$ have no atoms.

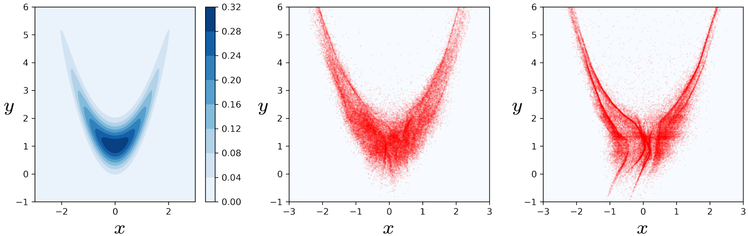
Then there exists at least one global minimizer T^* satisfying $\mathsf{T}^*_\sharp \eta = \nu$.

- Generalization of the KR map.
- All conditions stated for reference η.
- **Minimizers are not unique.**

[5] Bogachev, Kolesnikov, and Medvedev, "Triangular transformations of measures".

Theoretical foundations
Lack of uniqueness

$$\mathsf{T}^* = \arg\min_{\mathsf{T}\in\mathcal{T}} D(\mathsf{T}_\sharp\eta||\nu)$$



Target ν Good ordering Bad ordering

Conditional optimal transport

Theorem [Carlier, Chernozhukov and Galicon '16]

Suppose

- $\mathcal{X}, \mathcal{Y}$ are finite-dimensional Euclidean spaces.
- $\eta = \nu_{\mathcal{X}} \otimes \eta_{\mathcal{Y}}$ and $\eta_{\mathcal{Y}}$ has convex support.
- Conditionals $\nu(\cdot|x)$ have Lebesgue densities.
- $\int_{\mathcal{Y}} \int_{\mathcal{X}} \|y\|^2 \nu(\mathsf{d}x, \mathsf{d}y) + \int_{\mathcal{Y}} \int_{\mathcal{X}} \|y\|^2 \eta(\mathsf{d}x, \mathsf{d}y) < +\infty$.

Then there exists a map $\mathsf{S} = \nabla_y s(x, y)$ where $y \mapsto s(x, y)$ is convex for all $x \in \mathcal{X}$ and

$$\mathsf{S}(x^*, \cdot)_\sharp \eta_{\mathcal{Y}} = \nu(\cdot|x^*).$$

Furthermore, S is unique among such maps.

$$\begin{cases} \inf_{\mathsf{T}^{-1}} \int_{\mathcal{X} \times \mathcal{Y}} \|z - \mathsf{T}^{-1}(z)\|^2 \nu(\mathsf{d}z), \\ \text{subject to} \quad \mathsf{T}^{-1}_\sharp \nu = \eta \text{ and } \mathsf{T}^{-1}_\sharp \nu(\cdot|x) = \eta_{\mathcal{Y}}. \end{cases}$$

[9] G. Carlier, V. Chernozhukov, and A. Galichon. "Vector quantile regression: an optimal transport approach". In: *The Annals of Statistics* 44.3 (2016), pp. 1165–1192.

Theorem [KBHM]

Suppose

- $\mathcal{X}, \mathcal{Y}$ are finite-dimensional Eucledian spaces.
- $\eta = \eta_{\mathcal{X}} \otimes \eta_{\mathcal{Y}}$ and $\eta_{\mathcal{Y}}$ has convex support.
- Conditionals $\nu(\cdot|x)$ have Lebesgue densities.
- $\int_{\mathcal{Y}} \int_{\mathcal{X}} \|y\|^2 \nu(\mathrm{d}x, \mathrm{d}y) + \int_{\mathcal{Y}} \int_{\mathcal{X}} \|y\|^2 \eta(\mathrm{d}x, \mathrm{d}y) < +\infty.$

Consider the problem

$$
\begin{aligned}
\mathsf{T}^* &= \operatorname*{arg\,min}_{\mathsf{T} \in \mathcal{T}_B} D(\mathsf{T}_\sharp \eta || \nu), \\
\mathcal{T}_B &:= \Big\{ \mathsf{T}(x, y) = \big(\mathsf{K}(x), \mathsf{S}(\mathsf{K}(x), y\big) \\
&\qquad \text{such that } \mathsf{K} = \nabla_x k, \quad \mathsf{S} = \nabla_y s \\
&\qquad \text{and } x \mapsto k(x), y \mapsto s(x, y) \text{ are convex} \Big\}.
\end{aligned}
$$

Then T^* exists, is unique, and $\mathsf{S}^*(x^*, \cdot)_\sharp \eta_{\mathcal{Y}} = \nu(\cdot|x^*)$.

Monotone GANs

Building an algorithm based on theory

$$\mathsf{T}^* = \arg\min_{\mathsf{T}\in\widehat{\mathcal{T}}} D_{\mathsf{WGAN}}(\mathsf{T}_\sharp\eta||\nu),$$

- WGAN-GP discrepancy

$$D_{\mathsf{WGAN}}(\mathsf{T}_\sharp\eta||\nu) := \sup_{f\in\Gamma} \mathbb{E}_{z\sim\nu} f(z) - \mathbb{E}_{w\sim\eta} f(\mathsf{T}(w)) + \mathsf{Penalty}(f).$$

- Two choices for $\widehat{\mathcal{T}}$:

$$\mathcal{T}_B = \Big\{\mathsf{T}(x,y) = \big(\mathsf{K}(x), \mathsf{S}(\mathsf{K}(x),y\big) \\ \text{s.t. } \mathsf{K} = \nabla_x k, \quad \mathsf{S} = \nabla_y s, \quad k(\cdot), s(x,\cdot) \text{ are convex.}\Big\},$$

$$\mathcal{T}_M = \Big\{\mathsf{T}(x,y) = \big(\mathsf{K}(x), \mathsf{S}(\mathsf{K}(x),y\big) \\ \text{s.t. } \langle \mathsf{T}(w) - \mathsf{T}(w'), w - w'\rangle > 0 \quad \eta - \text{ a.e. } \Big\}.$$

[10] J. Birrell et al. "(f,Γ)-Divergences: Interpolating between f-Divergences and Integral Probability Metrics". In: *arXiv preprint:2011.05953* (2020).

Monotone GANs

Implementation with input-convex networks

$$\mathsf{T}^* = \arg\min_{\mathsf{T}} D_{\mathsf{WGAN}}(\mathsf{T}_\sharp \eta_N || \nu_N),$$

- Training data: $\{z_j = (x_j, y_j)\}_{j=1}^N \overset{\text{iid}}{\sim} \nu$.
- Reference samples (free): $\{w_k\}_{k=1}^{2J} \overset{\text{iid}}{\sim} \eta$.
- Empirical WGAN-GP approximation

$$D_{\mathsf{WGAN}}(\mathsf{T}_\sharp \eta_N || \nu_N) := \max_f \frac{1}{N} \sum_{j=1}^N f(z_j) - \frac{1}{J} \sum_{k=1}^J f(\mathsf{T}(w_k)) + \mathsf{Penalty}(f).$$

- Take $\mathsf{K}(x) = \nabla_x k(x)$ and $\mathsf{S}(x,y) = \nabla_y s(x,y)$.
- Parameterize k, s as input-convex neural nets.
- Parameterize f as neural net (discriminator).

Basic idea:

- Nonnegative sums of convex functions are convex.
- Composition of a convex and convex non-decreasing function is also convex.

[11] B. Amos, L. Xu, and JZ Kolter. "Input convex neural networks". In: *International Conference on Machine Learning*. 2017, pp. 146–155.

$$\mathsf{T}^* = \arg\min_{\mathsf{T}} \tilde{D}_{\mathsf{WGAN}}(\mathsf{T}_\sharp \eta || \nu),$$

- Similar setting as before.
- Parameterize $\mathsf{K}, \mathsf{S}, f$ as arbitrary neural networks.
- Impose average monotonicity by adding Lagrange multiplier

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

Input-convex neural nets versus average monotonicity

Input-convex neural nets:

- Pros:
  - Backed by theory (uniqueness).
  - Real valued neural network.
- Cons:
  - Requires specific architectures.
  - Training may be more involved.
  - Intrusive.

Average monotonicity:

- Pros:
  - Convenience.
  - Retrofit existing architectures.
  - Can track during training.
- Cons:
  - Sensitive dependence on Lagrange multiplier.
  - May be violated depending on problem.
  - Not supported theoretically (yet).

$$\mathsf{T}^* = \arg\min_{\mathsf{T}} D_{\mathsf{WGAN}}(\mathsf{T}_\sharp \eta_N || \nu_N),$$
$$\mathsf{T}^*(x, y) = \Big(\mathsf{K}^*(x), \mathsf{S}^*(\mathsf{K}^*(x), y)\Big)$$

- Compute T^* and extract the component S^*.
- New input $x^* \in \mathcal{X}$.
- Generate new samples from reference marginal $\tilde{y}_k \stackrel{\mathsf{iid}}{\sim} \eta_{\mathcal{Y}}$.
- Set $y_k = \mathsf{S}^*(x^*, \tilde{y}_k)$.
- Use the y_k to compute statistics of $\nu(\cdot|x^*)$.

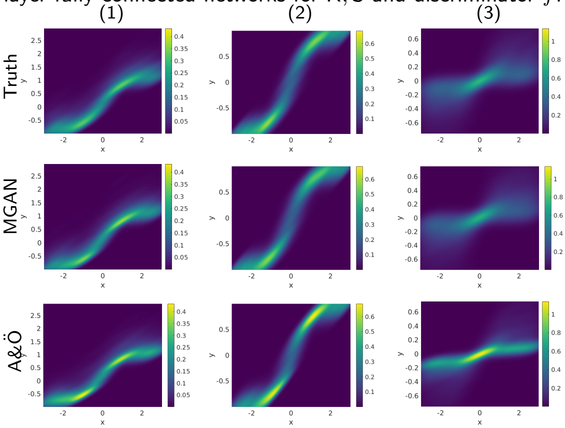
A 1D example

$$y = \tanh(x) + \gamma \qquad \gamma \sim \Gamma(1, 0.3), \tag{1}$$

$$y = \tanh(x + \gamma) \qquad \gamma \sim N(0, 0.05), \tag{2}$$

$$y = \gamma \tanh(x) \qquad \gamma \sim \Gamma(1, 0.3). \tag{3}$$

- $N = 5 \times 10^4$ training samples with input $x \sim U(-3, 3)$.
- Three layer fully connected networks for K, S and discriminator f.

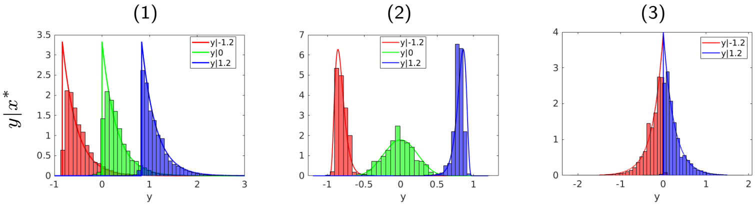


[12] J. Adler and O. Öktem. "Deep Bayesian inversion". In: *arXiv preprint:1811.05910* (2018).

A 1D example

$$y = \tanh(x) + \gamma \qquad \gamma \sim \Gamma(1, 0.3), \tag{4}$$
$$y = \tanh(x + \gamma) \qquad \gamma \sim N(0, 0.05), \tag{5}$$
$$y = \gamma \tanh(x) \qquad \gamma \sim \Gamma(1, 0.3). \tag{6}$$

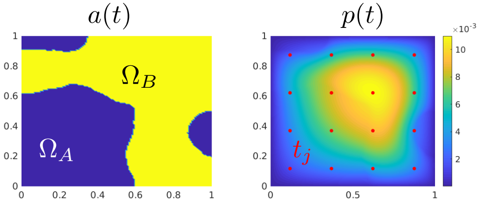
- $N = 5 \times 10^4$ training samples with input $x \sim U(-3, 3)$.
- Three layer fully connected networks for K, S and discriminator f.



Darcy flow

$$\begin{cases} -\nabla \cdot (a(t)\nabla p(t)) = 1, & t \in (0,1)^2, \\ p(t) = 0, & t \in \partial(0,1)^2. \end{cases}$$

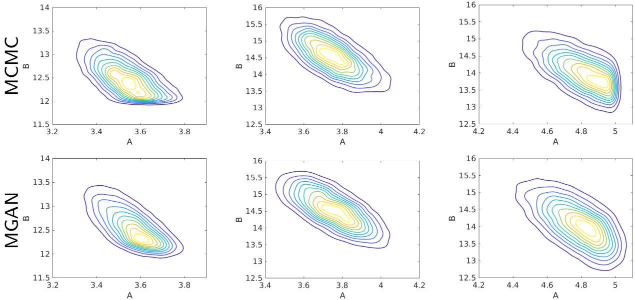
- Pressure field $p(t)$.
- Permeability field $a(t) = A\mathbf{1}_{\Omega_A}(t) + B\mathbf{1}_{\Omega_B}(t)$.
- **Goal:** Recover $y = (A, B)$ from measurements $x_j = p(t_j) + \gamma_j$ for $j = 1, \ldots, 16$ and $\gamma_j \sim N(0, 10^{-7})$.



Darcy flow

$$\begin{cases} -\nabla \cdot (a(t)\nabla p(t)) = 1, & t \in (0,1)^2, \\ p(t) = 0, & t \in \partial(0,1)^2. \end{cases}$$

- Prior $A \sim U(3,5), B \sim U(12,16)$.
- Discretize PDE using finite differences.
- Train MGAN using 10^5 training samples and three layer fully connected networks.
- Three data sets x_1^*, x_2^*, x_3^* generated from $y_1^* = (3.5, 13), y_2^* = (4, 14), y_3^* = (4.5, 15)$.
- Posterior KDEs of MGAN and MCMC (pCN), both with 3×10^4 samples.



- 100% monotone.

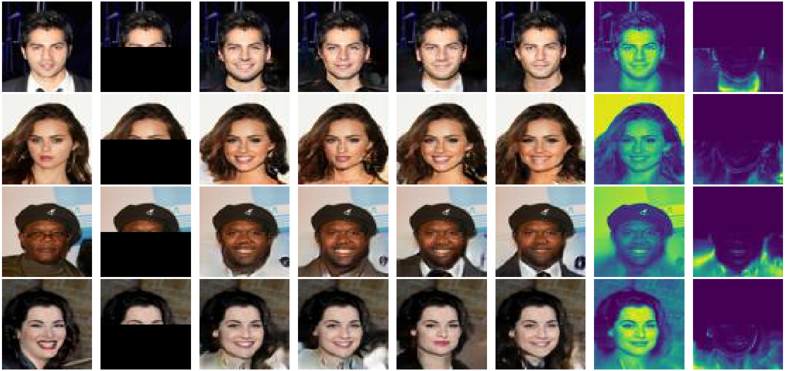
CelebA in-painting

- CelebA training set: 162,770 images of size $64 \times 64 \times 3$ (RGB).
- Observation x: top half of image.
- Parameter y: entire image.
- Given top half x^* from CelebA test set generate 1000 possible MGAN in-paintings $y|x^*$.
- Compute pixel-wise variances and sample mean.

y

x^*

CelebA in-painting



Summary

- A model agnostic method for conditional sampling.
  - Probabilistic supervised learning.
  - Bayesian inverse problems.
  - Likelihood-free inference.
- Measure transport view point towards conditioning.
- Straightforward implementation:
  - Retrofitting existing architectures with average monotonicity constraints.
  - Exact monotonicity constraints with input-convex neural nets.
- **When to use MGANs:**
  - Large training sets.
  - Inverse problems with unknown or black-box forward maps.
  - Intractable likelihood.
- **When not to use MGANs:**
  - Limited amounts of training data.
  - Expensive forward maps.
  - Highly concentrated target distributions.

Open questions and future directions

- Approximation theory: (KBHM + Sagiv)
  - Connecting map approximations to quality of samples.
  - Different approximation classes (polynomials, splines, RBFs, neural nets, etc).
  - Tail behavior of maps.
  - Consistency of minimizers in large data limits.
- Kernel formulation: (H + Owhadi)
  - Maximum mean discrepancy.
  - Parameterize T in RKHS (kernels, random features, etc).
  - Warped kernels and compositions of Gaussian processes.

Open questions and future directions

- Sample complexity.
- Choice of architecture:
  - GANs, ICNNs, Monotone nets, Neural ODEs, etc.
- Choice of cost function:
  - GAN, Wasserstein GAN, Kullback-Leibler divergence, Wasserstein metrics, etc.
- MGAN for inverse problems:
  - How does MGAN compare to MCMC?
  - Non-Gaussian priors and complicated posteriors.
  - Experimental design/active learning for MGANs?

Thank you

"Conditional Sampling with Monotone GANs"

arXiv:2006.06755

References

[1] F Gressmann et al. "Probabilistic supervised learning". In: *arXiv:1801.00753* (2019).

[2] Andrew M Stuart. "Inverse problems: a Bayesian perspective". In: *Acta numerica* 19 (2010), pp. 451–559.

[3] C. Villani. *Optimal transport: Old and new*. Springer, 2009.

[4] L. Younes. *Shapes and diffeomorphisms*. Springer, 2010.

[5] VI Bogachev, AV Kolesnikov, and KV Medvedev. "Triangular transformations of measures". In: *Sbornik: Mathematics* 196.3 (2005), p. 309.

[6] Gregory Ongie et al. "Deep learning techniques for inverse problems in imaging". In: *IEEE Journal on Selected Areas in Information Theory* 1.1 (2020), pp. 39–56.

[7] F. Santambrogio. *Optimal transport for applied mathematicians*. Springer, 2015.

[8] N. Bonnotte. "From Knothe's rearrangement to Brenier's optimal transport map". In: *SIAM Journal on Mathematical Analysis* 45.1 (2013), pp. 64–87.

[9] G. Carlier, V. Chernozhukov, and A. Galichon. "Vector quantile regression: an optimal transport approach". In: *The Annals of Statistics* 44.3 (2016), pp. 1165–1192.

[10] J. Birrell et al. "(f, Γ)-Divergences: Interpolating between f-Divergences and Integral Probability Metrics". In: *arXiv preprint:2011.05953* (2020).

[11] B. Amos, L. Xu, and JZ Kolter. "Input convex neural networks". In: *International Conference on Machine Learning*. 2017, pp. 146–155.

[12] J. Adler and O. Öktem. "Deep Bayesian inversion". In: *arXiv preprint:1811.05910* (2018).