Addressing the lack of reliability in machine learning

A novel probabilistic viewpoint on self-supervised data representations

Michael Arbel

CRCN and ISFP recruitment campaign at Inria
Grenoble Rhône-Alpes

23 may 2022

Academic background

2021-present
Starting Research Position at Thoth team
Inria Grenoble Rhône-Alpes, France

2016-2021
Ph.D. student at the Gatsby Unit
University College London, United Kingdom

2015-2016
Computer vision engineer
Prophesee, Paris, France

2014-2015
Student at Ecole Normale Supérieure de Cachan
Master M2 (MVA), Cachan, France

2011-2015
Student at Ecole Polytechnique
INRIA Grenoble Rhône-Alpes

Ph.D. contributions

Learning structure from data with generative models

Learning structure from data with generative models

Explicit Generative Models

$$p_\theta(X) = \frac{1}{Z_\theta} e^{-E_\theta(x)}$$

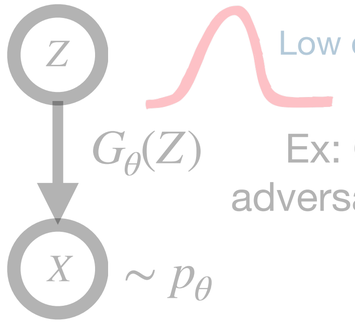
Ex: Energy-based models

✓Capture multimodality
Hard to sample from

$$\approx p_\theta$$

$$\min_\theta \mathscr{L}(p_\theta)$$

Implicit Generative Models



✓Capture low dimensionality
Hard to optimize

Learning structure from data with generative models

Which model to use?

$$\approx p_\theta$$

$$\min_\theta \mathcal{L}(p_\theta)$$

Explicit Generative Models

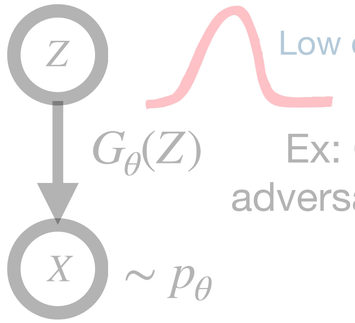
$$p_\theta(X) = \frac{1}{Z_\theta} e^{-E_\theta(x)}$$

Ex: Energy-based models

✓Capture multimodality

Hard to sample from

Implicit Generative Models



✓Capture low dimensionality

Hard to optimize

Learning structure from data with generative models

Which model to use?

$$\approx p_\theta$$

$$\min_\theta \mathcal{L}(p_\theta)$$

Explicit Generative Models

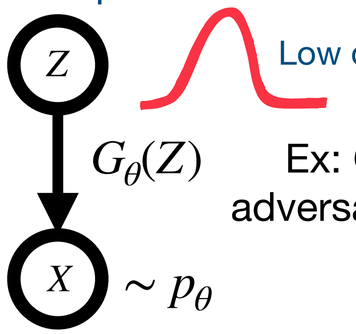
$$p_\theta(X) = \frac{1}{Z_\theta} e^{-E_\theta(x)}$$

Ex: Energy-based models

✓Capture multimodality

Hard to sample from

Implicit Generative Models



Ex: Generative adversarial networks

✓Capture low dimensionality

Hard to optimize

Learning structure from data with generative models

Which model to use?

$$\approx p_\theta$$

$$\min_\theta \mathcal{L}(p_\theta)$$

Explicit Generative Models

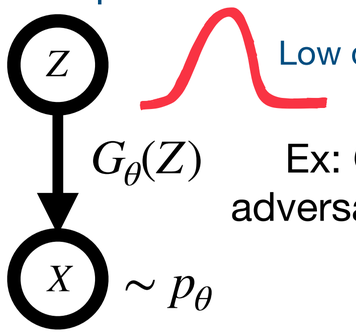
$$p_\theta(X) = \frac{1}{Z_\theta} e^{-E_\theta(x)}$$

Ex: Energy-based models

✓Capture multimodality

Hard to sample from

Implicit Generative Models



✓Capture low dimensionality

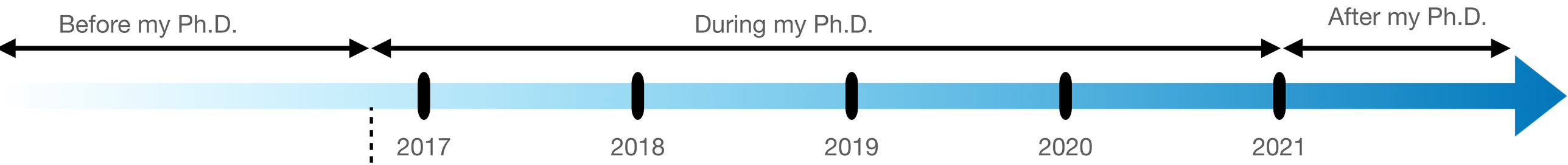
Hard to optimize

Contributions in sampling

Contributions in generative modeling

Contributions in optimization

Major contributions



Kernel Methods

Generative Modeling

Distributional optimization and Monte Carlo sampling

Information geometry and Optimization

Reinforcement learning

Industrial experience

Event-based cameras

- Pedestrian tracking systems
- Autocalibration algorithms

Learning structure from data with generative models

Explicit Generative Models

$$p_\theta(X) = \frac{1}{Z_\theta} e^{-E_\theta(x)}$$

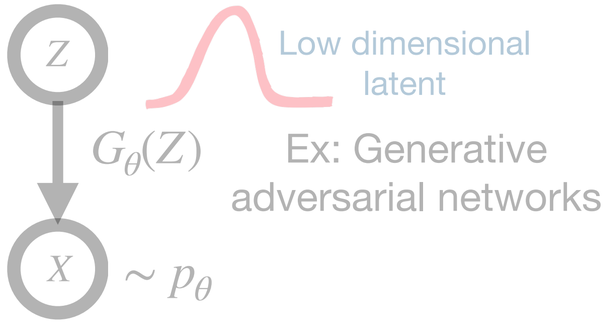
Ex: Energy-based models

✓Capture multimodality
Hard to sample from

$$\approx p_\theta$$

$$\min_\theta \mathscr{L}(p_\theta)$$

Implicit Generative Models



✓Capture low dimensionality
Hard to optimize

Contributions in sampling

Contributions in generative modeling

Contributions in optimization

Major contributions in generative modelling

Before my Ph.D.

Unstable algorithms for learning generative models

During my Ph.D.

Towards **stable and principled** algorithms for learning generative models

Explicit Generative Models

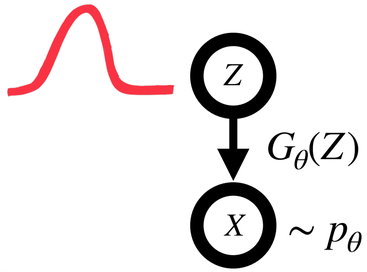
$$p_\theta(X) = \frac{1}{Z_\theta} e^{-E_\theta(x)}$$

✓Capture multimodality
Hard to sample from

$$\approx p_\theta$$

$$\min_\theta \mathscr{L}(p_\theta)$$

Implicit Generative Models



✓Capture low dimensionality
Hard to optimize

Generalized Energy-Based models

✓Capture multimodality + low dimensionality
✓A unified theory for learning generative models
✓Practical methods to ensure stable optimization

Contributions: A framework unifying both implicit and explicit models

5 publications at top-tier international conferences (NeurIPS, ICLR, AISTATS), more than 600 citations

Learning structure from data with generative models

Explicit Generative Models

$$p_\theta(X) = \frac{1}{Z_\theta} e^{-E_\theta(x)}$$

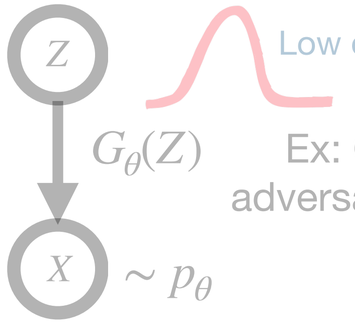
Ex: Energy-based models

✓Capture multimodality
Hard to sample from

$$\approx p_\theta$$

$$\min_\theta \mathscr{L}(p_\theta)$$

Implicit Generative Models



✓Capture low dimensionality
Hard to optimize

Contributions in sampling

Contributions in generative modeling

Contributions in optimization

Contributions in high dimensional sampling

Before my Ph.D.

Limited reliability of model-based sampling

During my Ph.D.

Efficient model-based samplers with theoretical guarantees

High dimensional sampling*

A fundamental problem

$$X \sim \pi$$

Challenges

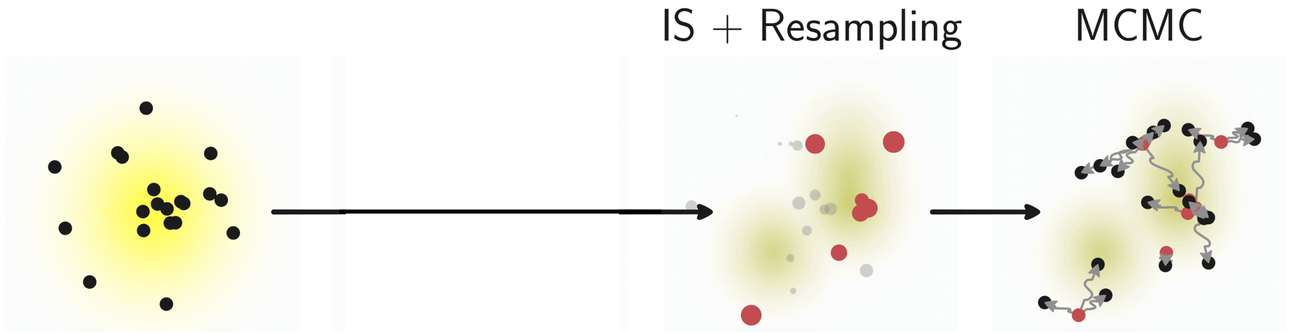
- Complex multi-modal distributions
- Curse of dimensions

Scientific context

Sequential Monte-Carlo (SMC) methods

✓ Reliable + Strong theoretical guarantees

Agnostic to the target density → High cost in general



Contributions: Principled and efficient model-based SMC samplers

1 publication at top-tier international conference (ICML 2021 Long Oral, top 3% submissions)

Equation of State Calculations by Fast Computing Machines, Metropolis et. al., 1953

Contributions in high dimensional sampling

Before my Ph.D.

Limited reliability of model-based sampling

During my Ph.D.

Efficient model-based samplers with theoretical guarantees

High dimensional sampling*

A fundamental problem

$$X \sim \pi$$

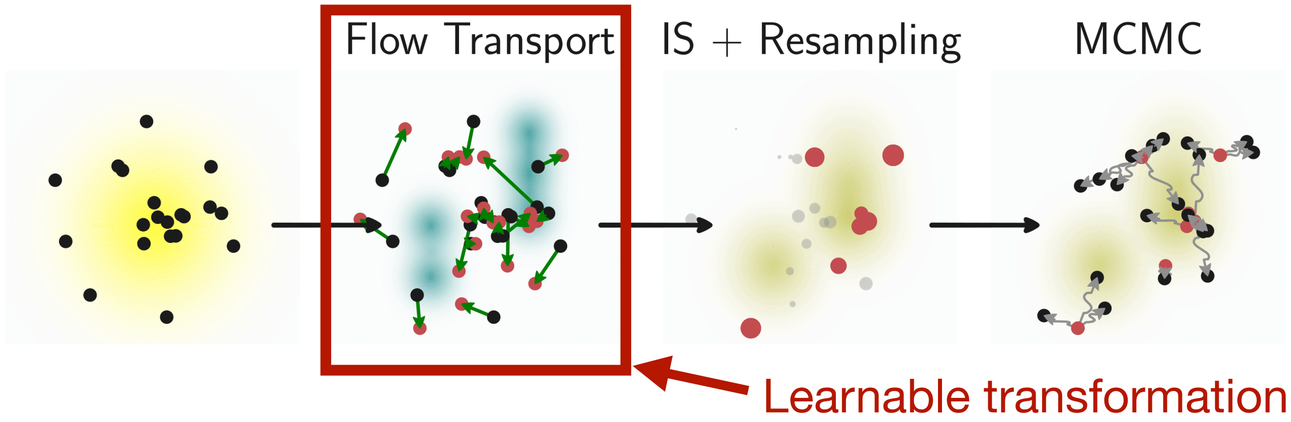
Key idea

Exploiting structure in the target density through modeling for more efficient sampling

↓

SMC with a learnable transformation

- ✓Fast sampling exploiting shared structure
- ✓Provably improves over than SMC
- ✓**Strong** approximation guarantees



Contributions: Principled and efficient model-based SMC samplers

1 publication at top-tier international conference (ICML 2021 Long Oral, top 3% submissions)

Equation of State Calculations by Fast Computing Machines, Metropolis et. al., 1953

Learning structure from data with generative models

Explicit Generative Models

$$p_\theta(X) = \frac{1}{Z_\theta} e^{-E_\theta(x)}$$

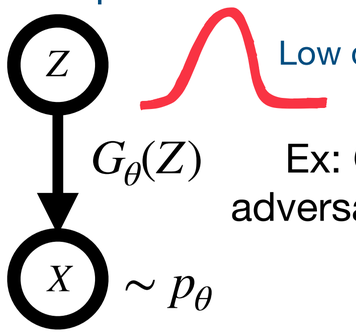
Ex: Energy-based models

✓Capture multimodality
Hard to sample from

$$\approx p_\theta$$

$$\min_\theta \mathscr{L}(p_\theta)$$

Implicit Generative Models



✓Capture low dimensionality
Hard to optimize

Contributions in sampling

Contributions in generative modeling

Contributions in optimization

Starting Research Position at Inria Thoth

Bilevel Optimization for Machine Learning

Contributions in bilevel optimization for machine learning

Bilevel optimization

$$\min_{x \in \mathbb{R}^p} \mathscr{L}(x) := f(x, y^\star(x))$$
$$\text{s.t. } y^\star(x) = \arg\min_{y \in \mathbb{R}^d} g(x, y)$$

Applications in machine learning

- Learning hyper-parameters of a learning algorithm
  - Data-augmentation
  - Regularization
  - Architecture search

Challenges

- Only approximations to $y^\star(x)$ are available
- Stochastic setting: additional errors due to noise

Scientific context

Gap between theory and practice

Theory: - Costly convergent algorithms
Practice: - Cheaper algorithms
- Not necessarily convergent

Contributions: First optimal convergence results for bilevel problems

1 publication at top-tier international conference (ICLR, 2022)

*Stabilité Asymptotique Pour des Problèmes de Perturbations Singulières, Habets, 1974

Contributions in bilevel optimization for machine learning

Bilevel optimization

$$\min_{x \in \mathbb{R}^p} \mathscr{L}(x) := f(x, y^\star(x))$$
$$\text{s.t. } y^\star(x) = \arg\min_{y \in \mathbb{R}^d} g(x, y)$$

Applications in machine learning

- Learning hyper-parameters of a learning algorithm
  - Data-augmentation
  - Regularization
  - Architecture search

Key idea

Introducing tools from dynamical systems:
*Singularly perturbed systems analysis**

A tight analysis of simple and efficient algorithms

✓ The first analysis providing optimal convergence rates
✓ Significant reduction of the computational cost

Contributions: First optimal convergence results for bilevel problems

1 publication at top-tier international conference (ICLR, 2022)

*Stabilité Asymptotique Pour des Problèmes de Perturbations Singulières, Habets, 1974

Addressing the lack of reliability in Machine learning

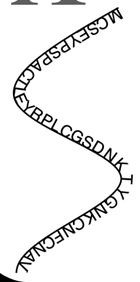
A novel probabilistic viewpoint on self-supervised data representations

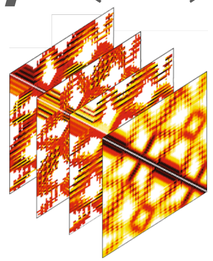
Key Observation

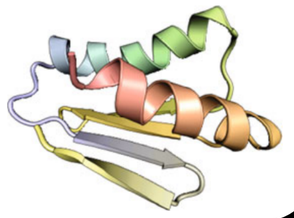
Prediction task

Data → Representation → Target

$X \rightarrow \phi(X) \rightarrow Y$





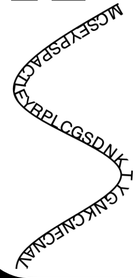


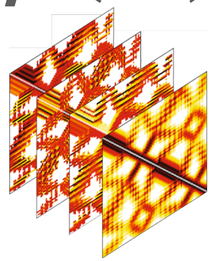
Key Observation

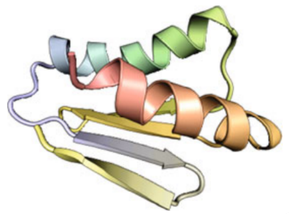
Prediction task

Data → Representation → Target

$X \rightarrow \phi(X) \rightarrow Y$







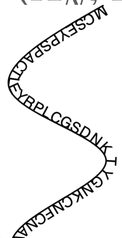
Supervised Representation Learning

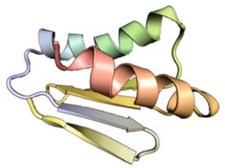
Labelled Data

(X_1, Y_1)

$\vdots$

(X_N, Y_N)

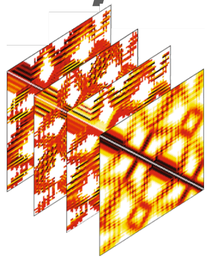




Machine Learning Algorithm

Learned Representation

ϕ

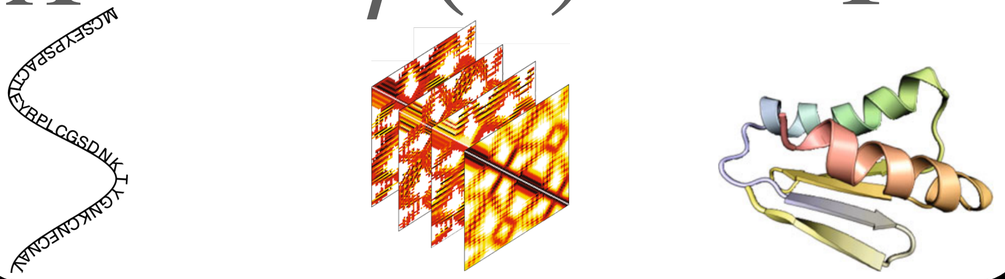


Key Observation

Prediction task

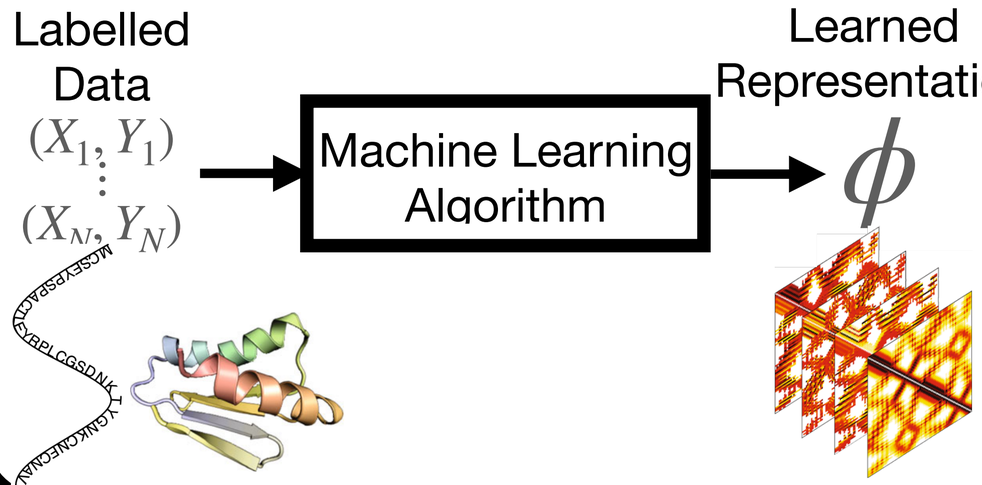
Data → Representation → Target

$X \rightarrow \phi(X) \rightarrow Y$

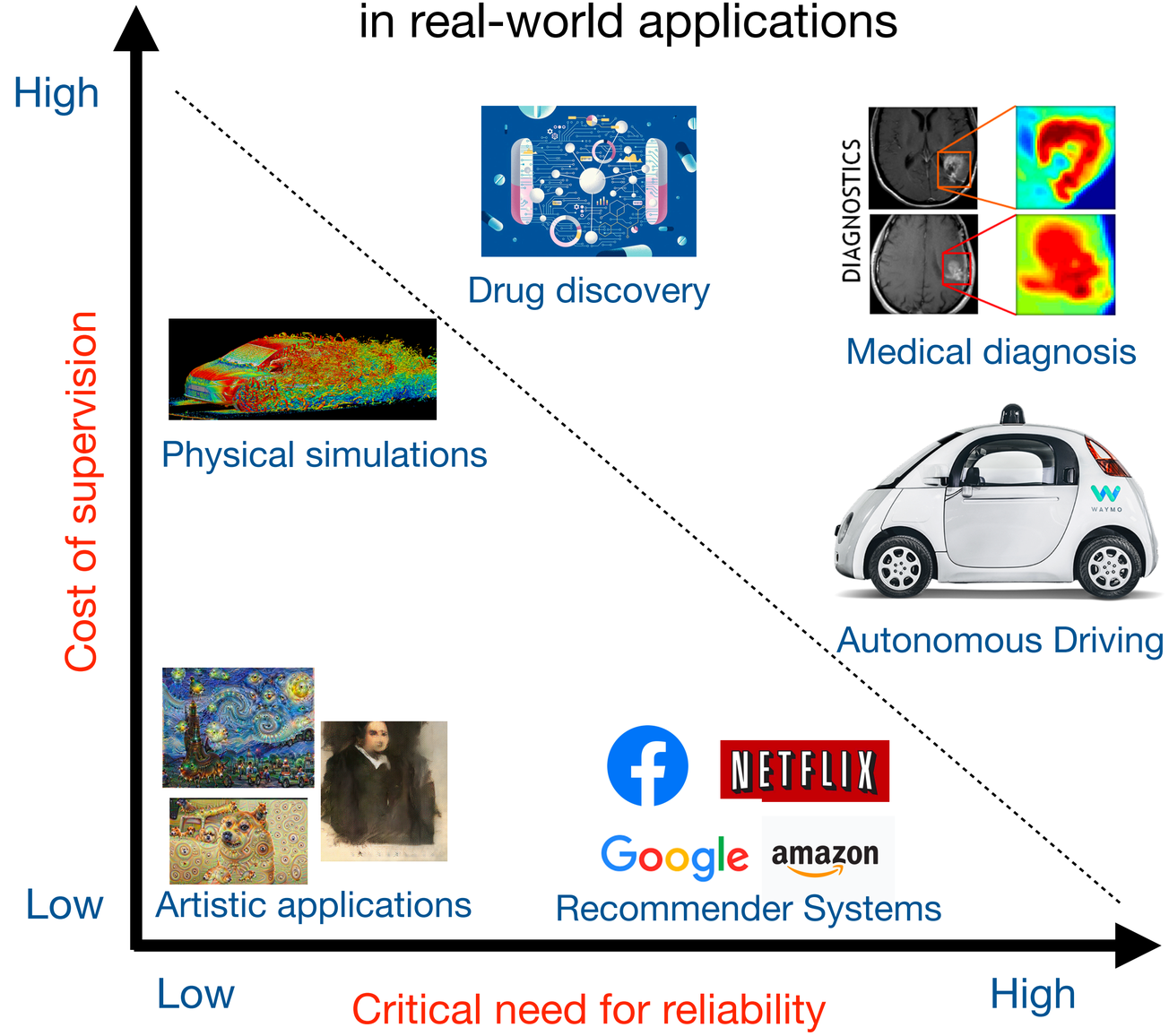


Supervised Representation Learning

Labelled Data $(X_1, Y_1) \ldots (X_N, Y_N)$ → Machine Learning Algorithm → Learned Representation ϕ

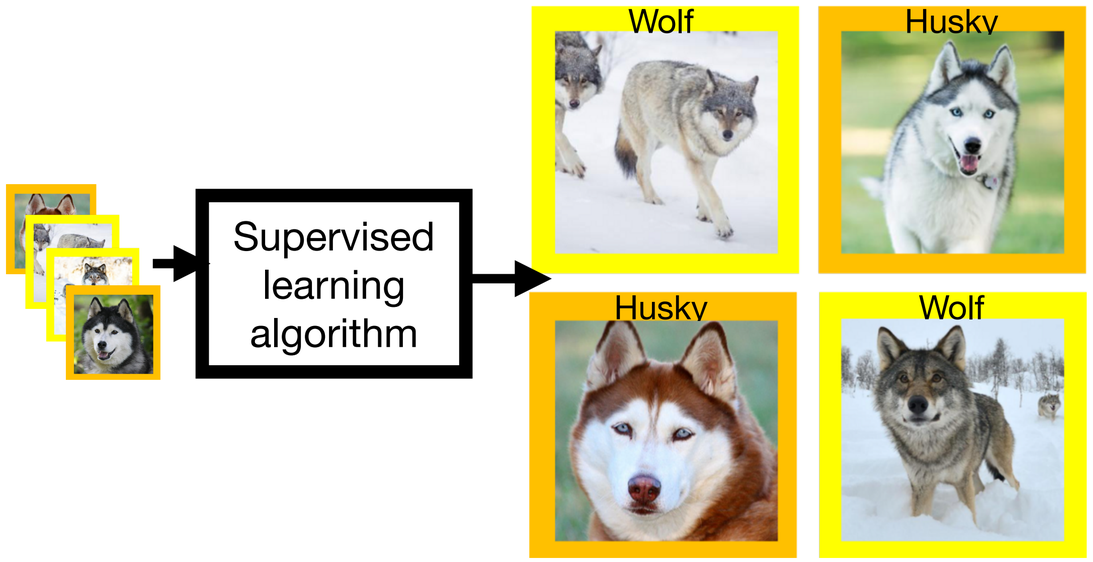


Deployment of Machine Learning methods in real-world applications



Some sources of unreliability in Machine Learning

Lack of robustness to distributional shift



Training dataset

Wolf images → Snowy background
Husky images → Grassy background

Some sources of unreliability in Machine Learning

Lack of robustness to distributional shift

Supervised learning algorithm

Husky classified as Wolf

Training dataset

Wolf images → Snowy background

Husky images → Grassy background

Systems learn spurious correlations

Do not generalize to out of distribution shifts

Need to develop robust representations

Some sources of unreliability in Machine Learning

Lack of robustness to distributional shift

Supervised learning algorithm →

Husky classified as Wolf

Training dataset

Wolf images → Snowy background

Husky images → Grassy background

Systems learn spurious correlations

Do not generalize to out of distribution shifts

↓

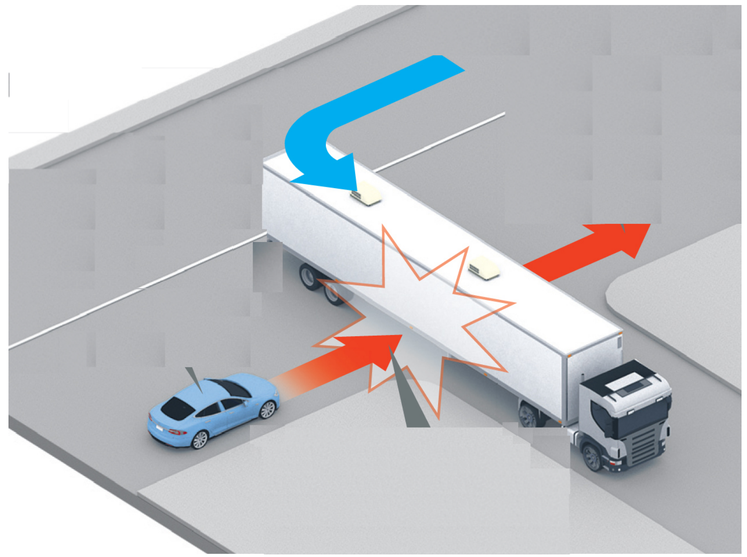
Need to develop robust representations

Lack of uncertainty quantification

Autopilot confuses a truck with a bright sky

↓

Car continues driving at full speed



Overly-confident systems

Do not account for alternative outcomes

↓

Need to account for uncertainty in the predictions

Representation learning

$X \longrightarrow \phi(X) \longrightarrow Y$ Toxicity

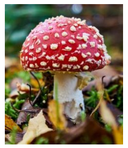
Self-supervised representation learning

Robustness to spurious correlations*
by encoding knowledge relevant for prediction

Prior knowledge

Relevant for the task: Color, Texture, Shape,…

Unlabelled data



X_1
$\vdots$
X_N

Prior work

- *DeepCluster*
- *SwAV*
- *DINO*

Learned Representation

ϕ

Does not account for uncertainty

*An Empirical Study on Robustness to Spurious Correlations using Pre-trained Language Models, Tu et. al., 2020

Representation learning

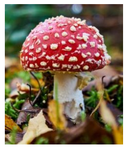
$X \rightarrow \phi(X) \rightarrow Y$ Toxicity

Self-supervised representation learning

Robustness to spurious correlations*
by encoding knowledge relevant for prediction

Prior knowledge
Relevant for the task: Color, Texture, Shape,…

Unlabelled data



X_1
$\vdots$
X_N

Prior work
- *DeepCluster*
- *SwAV*
- *DINO*

Learned Representation

ϕ

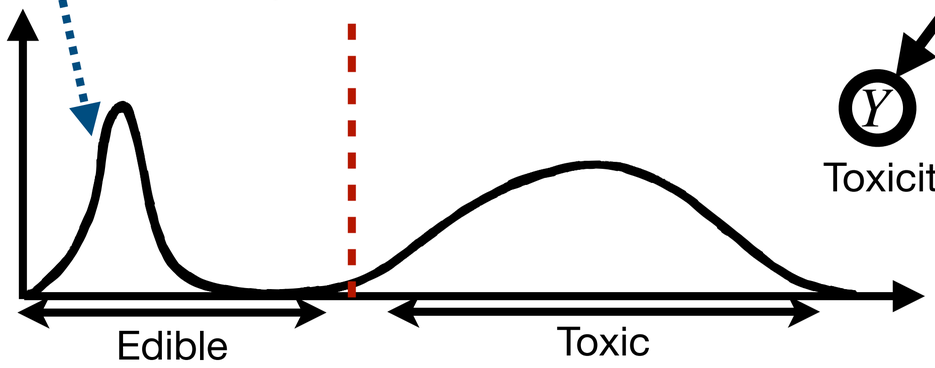
Does not account for uncertainty

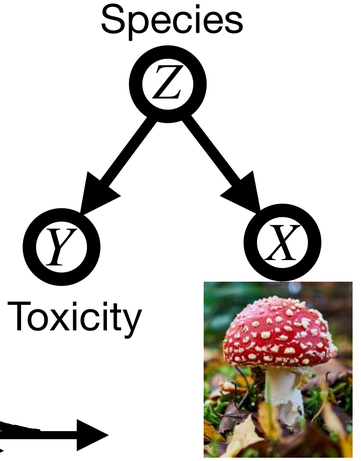
Probabilistic unsupervised representation learning

Accounting for uncertainty in the predictions

Probabilistic representation

$$p(Y|X) = \int p(Y|Z)p(Z|X)dZ$$





*An Empirical Study on Robustness to Spurious Correlations using Pre-trained Language Models, Tu et. al., 2020

Representation learning

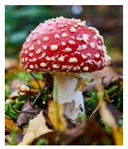
$X \rightarrow \phi(X) \rightarrow Y$ Toxicity

Self-supervised representation learning

Robustness to spurious correlations
by encoding knowledge relevant for prediction

Prior knowledge
Relevant for the task: Color, Texture, Shape,…

Unlabelled data



X_1
$\vdots$
X_N

Prior work
- *DeepCluster*
- *SwAV*
- *DINO*

Learned Representation

ϕ

Does not account for uncertainty

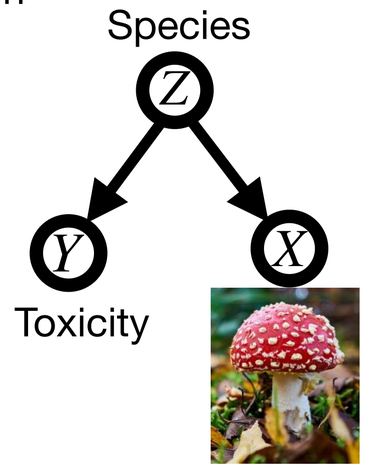
Probabilistic unsupervised representation learning

Accounting for uncertainty in the predictions

Probabilistic representation

$$p(Y|X) = \int p(Y|Z)p(Z|X)dZ$$

Prior work
- *Variational Auto-encoders*
- *Deep Bayesian networks*
- *Deep Gaussian processes*
- *Non-linear ICA*



Unsupervised → may not be relevant for prediction*

*Nonlinear ICA using auxiliary variables and generalized contrastive learning, Hyvarinen et al., 2019

Representation learning

Self-Supervised representation Learning

Encodes knowledge relevant for prediction

Does not account for uncertainty

Prior work
- *DeepCluster*
- *SwAV*
- *DINO*

Probabilistic unsupervised representation learning

Accounts for uncertainty

Not always relevant for prediction

Prior work
- *Variational Auto-encoders*
- *Deep Bayesian networks*
- *Deep Gaussian processes*
- *Non-linear ICA*

Representation learning

Self-Supervised representation Learning

Encodes knowledge relevant for prediction

Does not account for uncertainty

Prior work
- *DeepCluster*
- *SwAV*
- *DINO*

Probabilistic unsupervised representation learning

Accounts for uncertainty

Not always relevant for prediction

Prior work
- *Variational Auto-encoders*
- *Deep Bayesian networks*
- *Deep Gaussian processes*
- *Non-linear ICA*

My vision

Learning probabilistic representations that reflect the data generating mechanism

Probabilistic self-supervised representation learning

Axes 1 & 2

Theory and method

A new probabilistic self-supervised learning framework

✓Accounts for uncertainty

✓Robust to distributional shift

↓

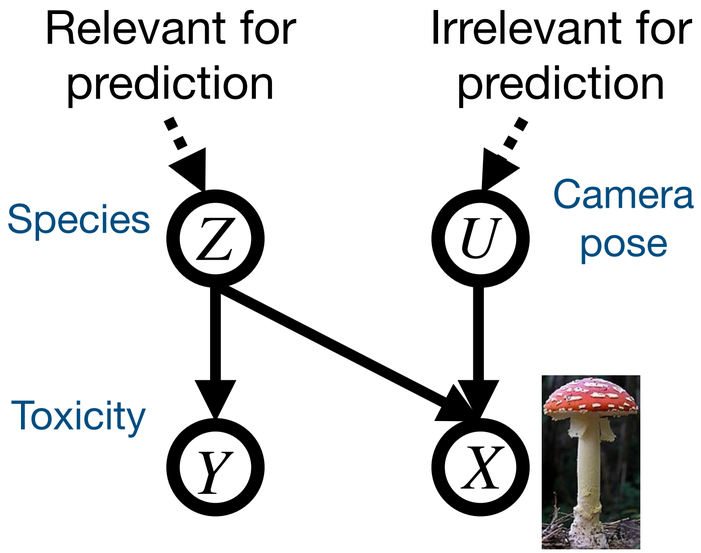
Efficient and reliable machine learning

Axes 3 & 4

Tools

Efficient and reliable sampling and optimization

Axis 1: A new probabilistic self-supervised learning framework



Probabilistic representation

$$p(Y|X) \propto \int p(Y|Z)p(Z|X)dZ$$

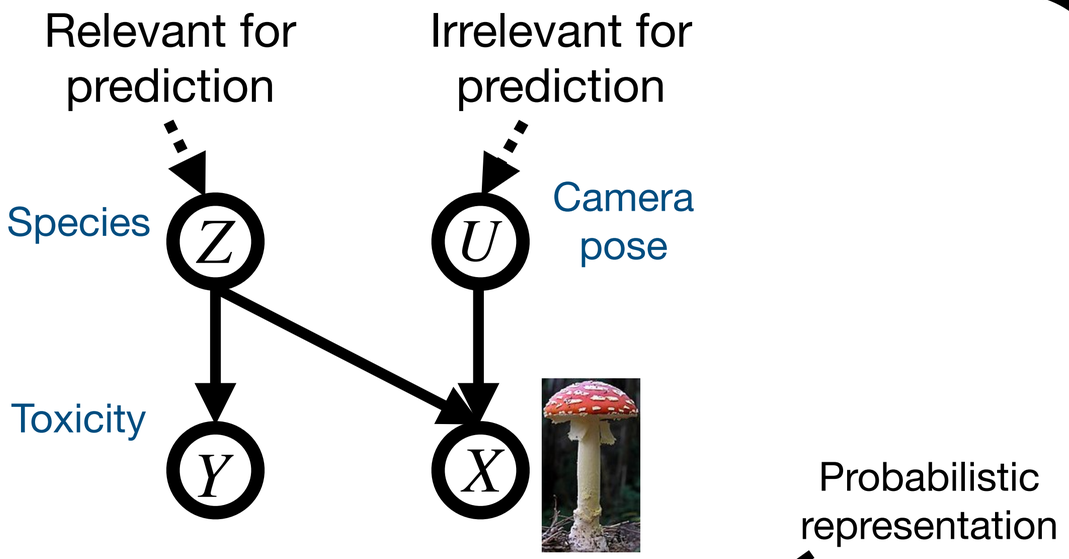
Challenges

- Robustness to distributional shift
- Identifying the relevant latent variable

Impact

- Better robustness to distribution shift
- Modeling uncertainty
- Provides mechanisms for basic reasoning

Axis 1: A new probabilistic self-supervised learning framework



Probabilistic representation

$$p(Y|X) \propto \int p(Y|Z)p(Z|X)dZ$$

Challenges

- Robustness to distributional shift
- Identifying the relevant latent variable

Impact

- Better robustness to distribution shift
- Modeling uncertainty
- Provides mechanisms for basic reasoning

Self-supervised learning of $p(Z|X)$

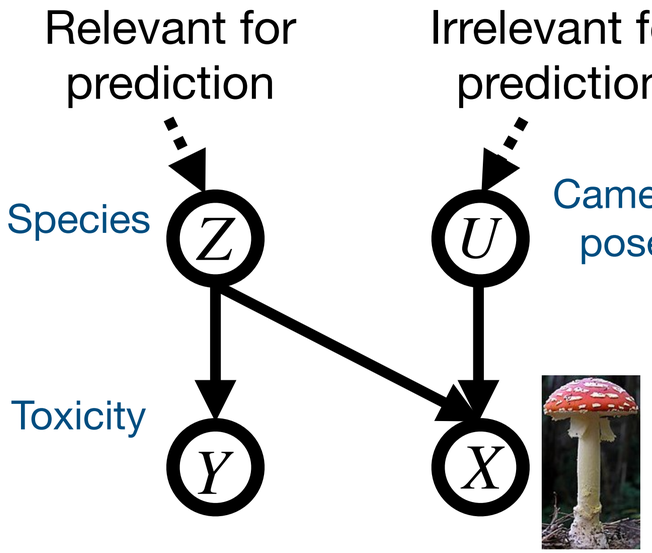
Models robust to distributional shift

Preserving the structure of the generating mechanism*

$$p(Z|X) \approx p_\theta(X|Z)p_\theta(Z)$$

* On causal and anticausal learning, Schölkopf et al. , 2012

Axis 1: A new probabilistic self-supervised learning framework



Probabilistic representation

$$p(Y|X) \propto \int p(Y|Z)p(Z|X)dZ$$

Challenges

- ❖ Robustness to distributional shift
- ❖ Identifying the relevant latent variable

Impact

- ❖ Better robustness to distribution shift
- ❖ Modeling uncertainty
- ❖ Provides mechanisms for basic reasoning

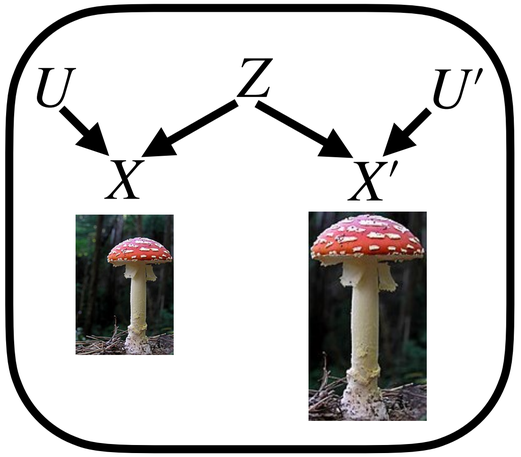
Self-supervised learning of $p(Z|X)$

Models robust to distributional shift

Preserving the structure of the generating mechanism*

$$p(Z|X) \approx p_\theta(X|Z)p_\theta(Z)$$

Identifying the relevant latent variable

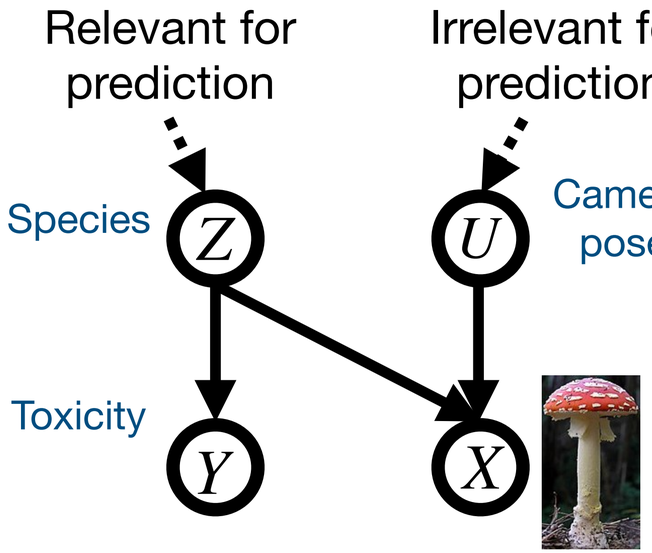


Examples:

- Data augmentation
- Ordering in time series
- Interventions on causal variables

* On causal and anticausal learning, Schölkopf et al. , 2012

Axis 1: A new probabilistic self-supervised learning framework



Probabilistic representation

$$p(Y|X) \propto \int p(Y|Z)p(Z|X)dZ$$

Challenges

- ❖ Robustness to distributional shift
- ❖ Identifying the relevant latent variable

Impact

- ❖ Better robustness to distribution shift
- ❖ Modeling uncertainty
- ❖ Provides mechanisms for basic reasoning

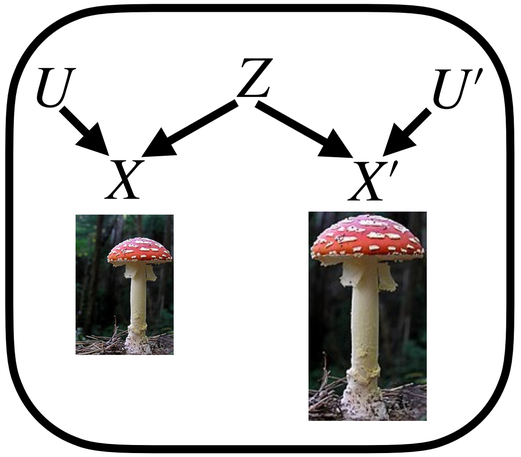
Self-supervised learning of $p(Z|X)$

Models robust to distributional shift

Preserving the structure of the generating mechanism*

$$p(Z|X) \approx p_\theta(X|Z)p_\theta(Z)$$

Identifying the relevant latent variable



Examples:

- Data augmentation
- Ordering in time series
- Interventions on causal variables

Beyond maximum likelihood:

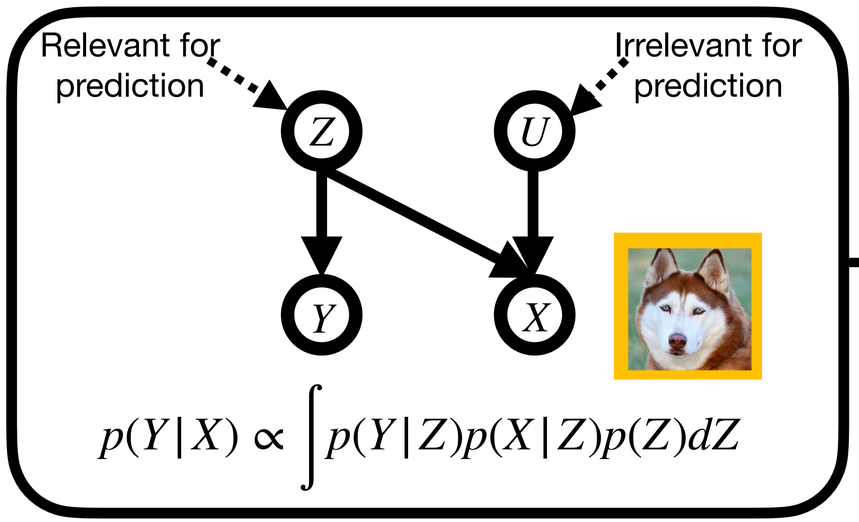
Learning to associate correlated pairs (X, X')

$$p(X, X') \approx \int p_\theta(X|Z)p_\theta(X'|Z)p_\theta(Z)dZ$$

* On causal and anticausal learning, Schölkopf et al. , 2012

Axis 2: A learning theory for probabilistic self-supervised learning (PSSL)

Data generating mechanism



Data

Learning algorithm

Probabilistic Self-supervised model

$\hat{p}$

Generalization theory

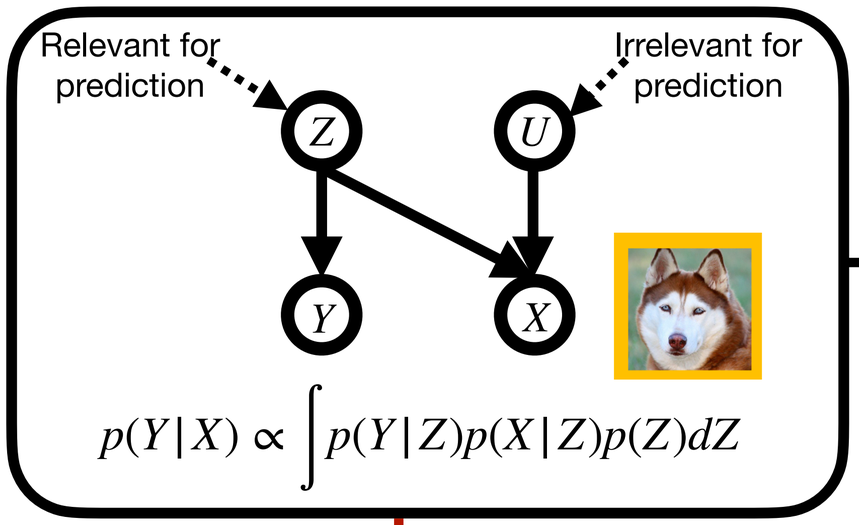
Prediction error

$|\mathscr{R}(\hat{p}) - \mathscr{R}(p)| \xrightarrow[\text{data size} \to \infty]{} 0$

Generalization error

Axis 2: A learning theory for probabilistic self-supervised learning (PSSL)

Data generating mechanism



Data

Learning algorithm

Probabilistic Self-supervised model

$\hat{p}$

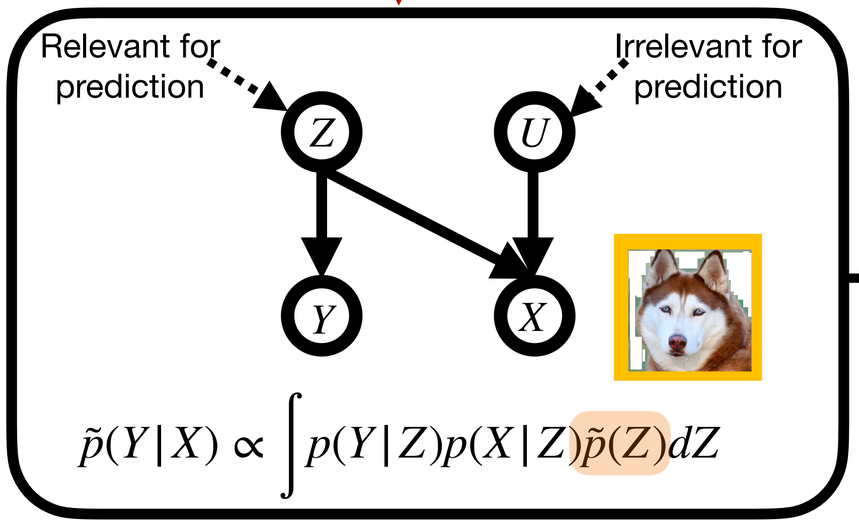
Generalization theory

Prediction error

$|\underbrace{\mathscr{R}(\hat{p}) - \mathscr{R}(p)}_{\text{Generalization error}}| \xrightarrow[\text{data size} \to \infty]{} 0$

Distribution shift

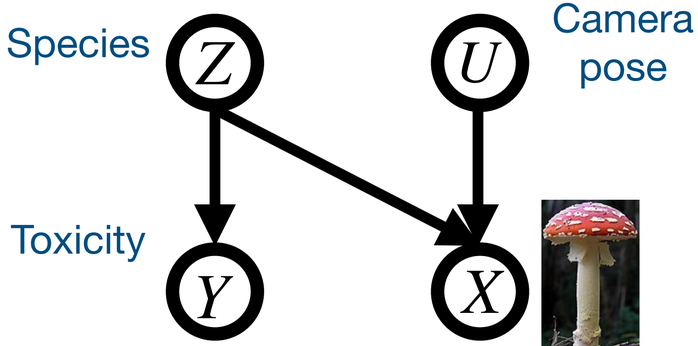
Adaptation of $\hat{p}$



Perturbed Data

Learning algorithm

Axis 3: Efficient and reliable sampling from distributions with a shared structure



$$p(Y|X) = \int p(Y|Z)p(Z|X)dZ$$

Sampling from multiple target densities

$$Z \sim p(Z|X)$$

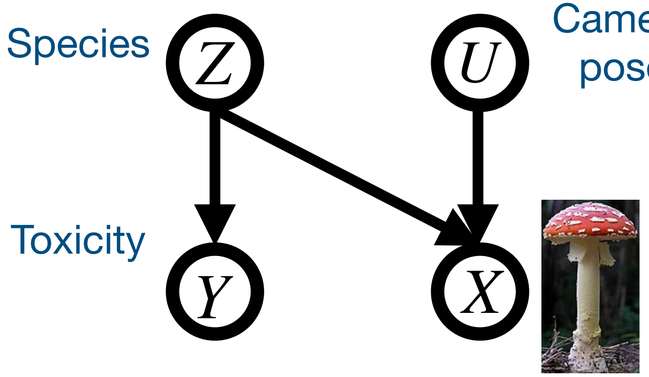
Each data point X defines its own target

Popular approach: *Amortized Inference*

- Fast approximate sampling
- Sample quality can degrade drastically*
- **Unreliable:** no approximation guarantees

*Inference suboptimality in variational autoencoders, Cremer et. al., 2018

Axis 3: Efficient and reliable sampling from distributions with a shared structure



$$p(Y|X) = \int p(Y|Z)p(Z|X)dZ$$

Sampling from multiple target densities

$$Z \sim p(Z|X)$$

Each data point X defines its own target

Popular approach: *Amortized Inference*

- Fast approximate sampling
- Sample quality can degrade drastically*
- **Unreliable:** no approximation guarantees

Model-based samplers for fast and reliable inference

Accounting for **shared** structure in $p(Z|X)$

MCMC methods

Reliability

Approximation guaranteed

Models of $p(Z|X)$

Shared structure

- Smoothness
- Symmetries

Model-based MCMC methods

Reliability + Efficiency

- Fast sampling exploiting shared structure
- Asymptotically exact samples
- **Strong** approximation guarantees

*Inference suboptimality in variational autoencoders, Cremer et. al., 2018

Axis 4: Non-convex optimization of intractable objectives

Bilevel Optimization

$$\min_{\theta} \mathscr{L}(\theta) := f(\theta, y^\star(\theta))$$
$$\text{s.t. } y^\star(\theta) = \arg\min_{y \in \mathbb{R}^d} g(\theta, y)$$

Optimization

Correspondence*

Probabilistic self-supervised learning

Needs samples from q_θ

$$\min_{\theta} \mathscr{L}(\theta) := \ell(q_\theta)$$
$$\text{s.t. } q_\theta = \arg\min_{q} KL(q \,||\, p_\theta)$$

Langevin dynamics, MCMC

Key idea

Amortizing the cost of previous computations

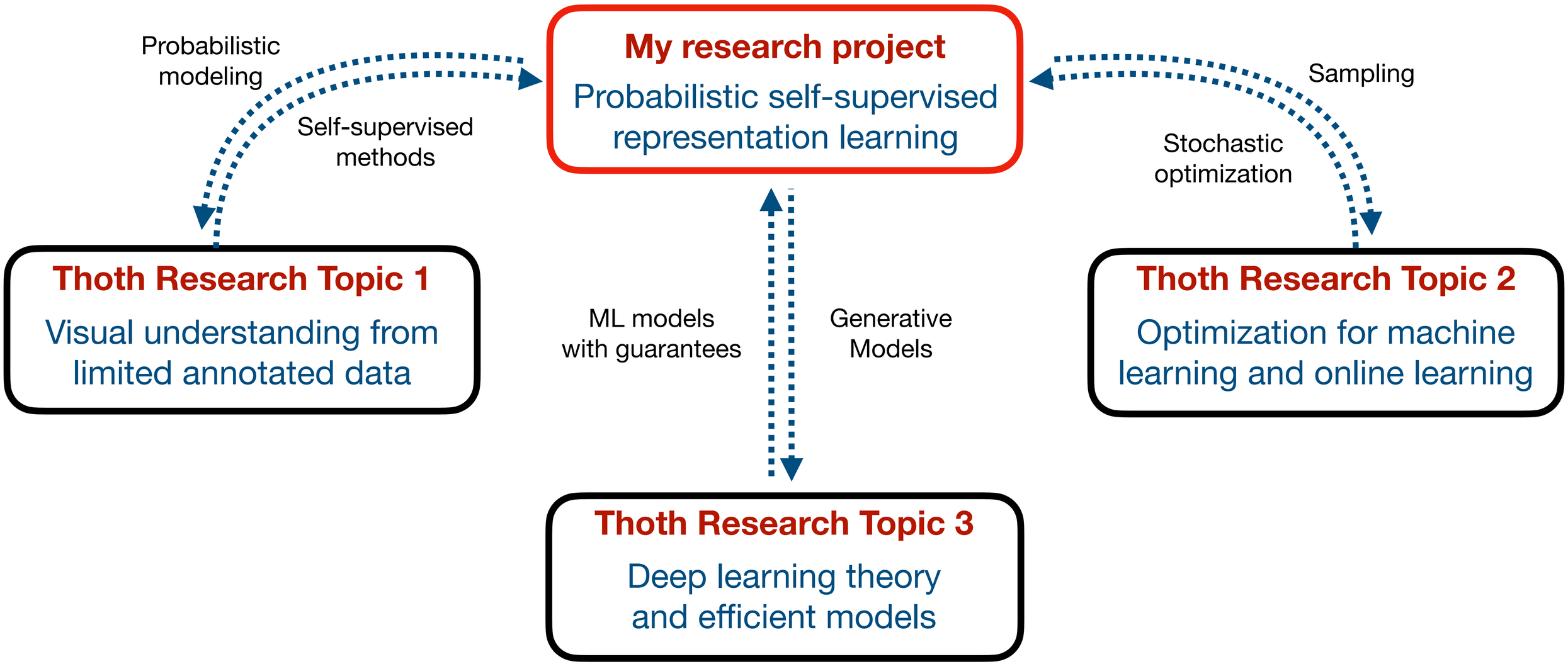
✓ Computationally efficient algorithms
✓ Strong convergence result

↓

Reliable and efficient optimization methods

*The variational formulation of the Fokker--Planck equation, Jordan et. al., 1998

Integration within Thoth team



Integration at Grenoble Rhône-Alpes ecosystem

Potential collaborations with Inria teams:

Statify (Bayesian models)

Polaris (Game theory)

Dantes (Lyon) (Frugal learning)

Potential collaborations with LJK

DAO: Optimization and Learning for Data Sciences

Potential industrial collaborations in Grenoble

Naver Labs (ongoing)

Criteo

Scientific work and collaborations

- Experience in generative modeling, sampling, and statistics
- 17 papers at top-tier international conferences in machine learning
- 17 seminars and invited talks (USA, UK, Germany, France)
- Academic and industrial collaborations

Multidisciplinary research project

Addressing the lack of reliability in Machine learning

A novel probabilistic viewpoint on self-supervised data representations

Axes 1 & 2

A new probabilistic self-supervised learning framework

Theory and method

Axes 3 & 4

Efficient and reliable sampling and optimization

Scientific work and collaborations

❖ 17 papers at top-tier international conferences in machine learning
- [Moskovitz et al. 2022]: Best Paper Award Finalist (Top 4 papers)
- [Arbel et al. 2021]: Long Oral Presentation (Top 3%)
- [Arbel et al. 2020]: Spotlight Presentation (Top 6%)

❖ 17 seminars and invited talks (USA, UK, Germany, France, Switzerland)
- Workshop on Machine Learning Augmented Sampling for the Molecular Sciences, (EPFL), May 2022
- Les journées MAS, (SMAI), August 2021
- Deep Learning Theory Kickoff Meeting, (MPI), March 2019

❖ Academic and industrial collaborations
- Optimal transport geometry: W. Li (South Carolina), G. Montufar (MPI)
- Reinforcement learning: A. Pacchiano (Microsoft)
- Computer Vision: T. Birdal (Stanford), U. Simsekli (Inria)
- High dimensional Sampling: A. Matthews (DeepMind), A. Doucet (Oxford)

❖ Service to the community
- **Reviewing**: International journals and conferences
- **Organizing seminars**: Ellis seminar series, THOTH seminar series
- **Supervision**: 3 Ph.D. students in co-supervision
- **Teaching**: 50% of a Master course at ENS Paris-Saclay (MVA)

Major contributions in generative modelling

Before my Ph.D.

Unstable algorithms for learning generative models

During my Ph.D.

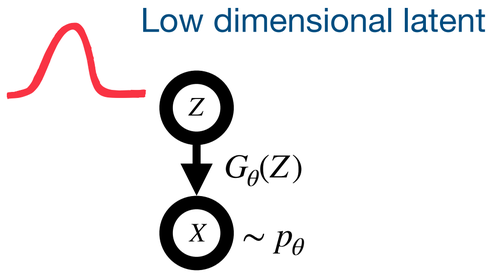
Towards **stable and principled** algorithms for learning generative models

Explicit Generative Models

$$p_\theta(X) = \frac{1}{Z_\theta} e^{-E_\theta(x)}$$

$$\min_\theta \mathscr{L}(p_\theta)$$

Implicit Generative Models



✓No need for a density for the model

Often leads to unstable algorithms

Scientific context: Variational objectives

$$\mathscr{D}(p_{data}, p_\theta) = \sup_{\phi \in \mathscr{F}} \mathbb{E}[\phi(X_{data})] - \mathbb{E}[\phi(X_{model})]$$

Contributions: A framework unifying both implicit and explicit models

5 publication at top-tier international conference (NeurIPS, ICLR, AISTATS), more than 600 citations

Contributions in high dimensional sampling

Before my Ph.D.

Limited reliability of model-based sampling

During my Ph.D.

Efficient model-based samplers with theoretical guarantees

High dimensional sampling*

A fundamental problem

$$X \sim \pi$$

Challenges

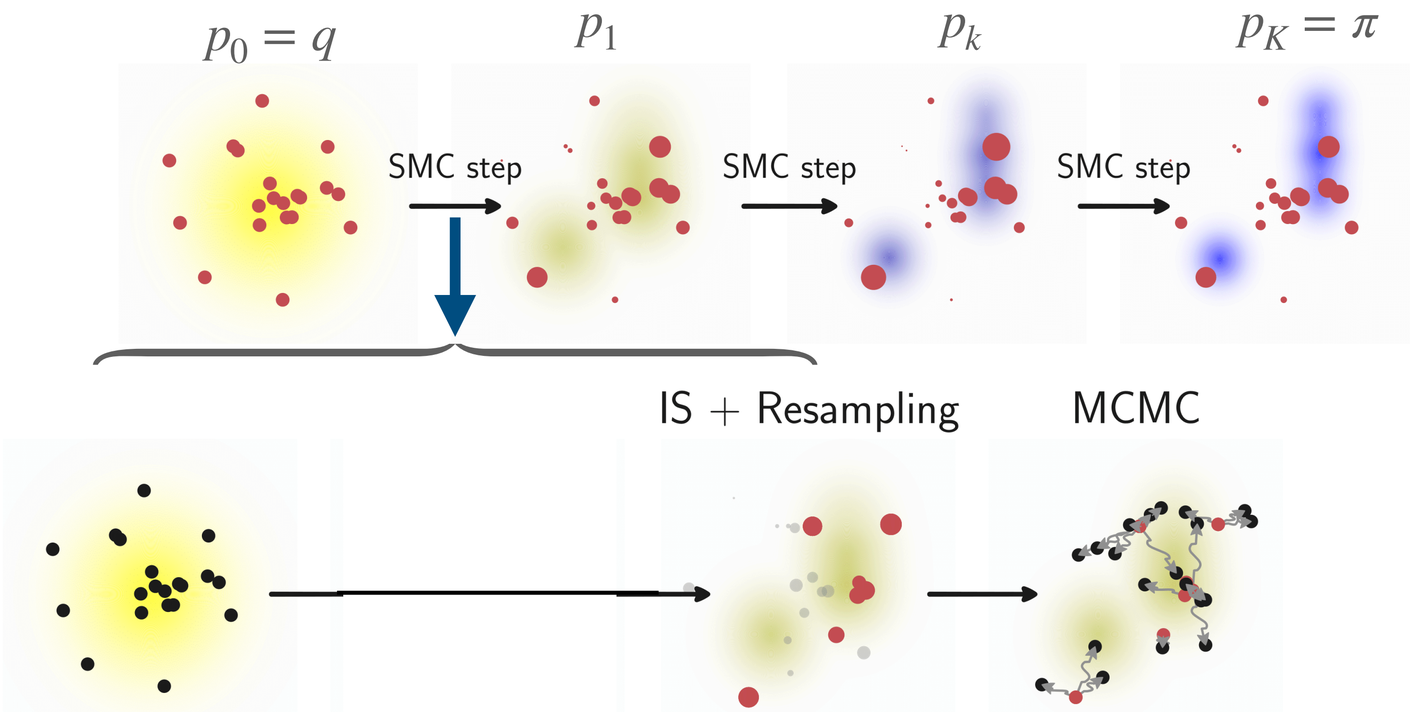
- Complex multi-modal distributions
- Curse of dimensions

Scientific context

Sequential Monte-Carlo (SMC) methods

✓ Reliable + Strong theoretical guarantees

Agnostic to the target density → High cost in general



Contributions: Principled and efficient model-based SMC samplers that capture structure in the target density

1 publication at top-tier international conference (ICML 2021 Long Oral, top 3% submissions)

Equation of State Calculations by Fast Computing Machines, Metropolis et. al., 1953

Contributions in high dimensional sampling

Before my Ph.D.

Limited reliability of model-based sampling

During my Ph.D.

Efficient model-based samplers with theoretical guarantees

High dimensional sampling*

A fundamental problem

$$X \sim \pi$$

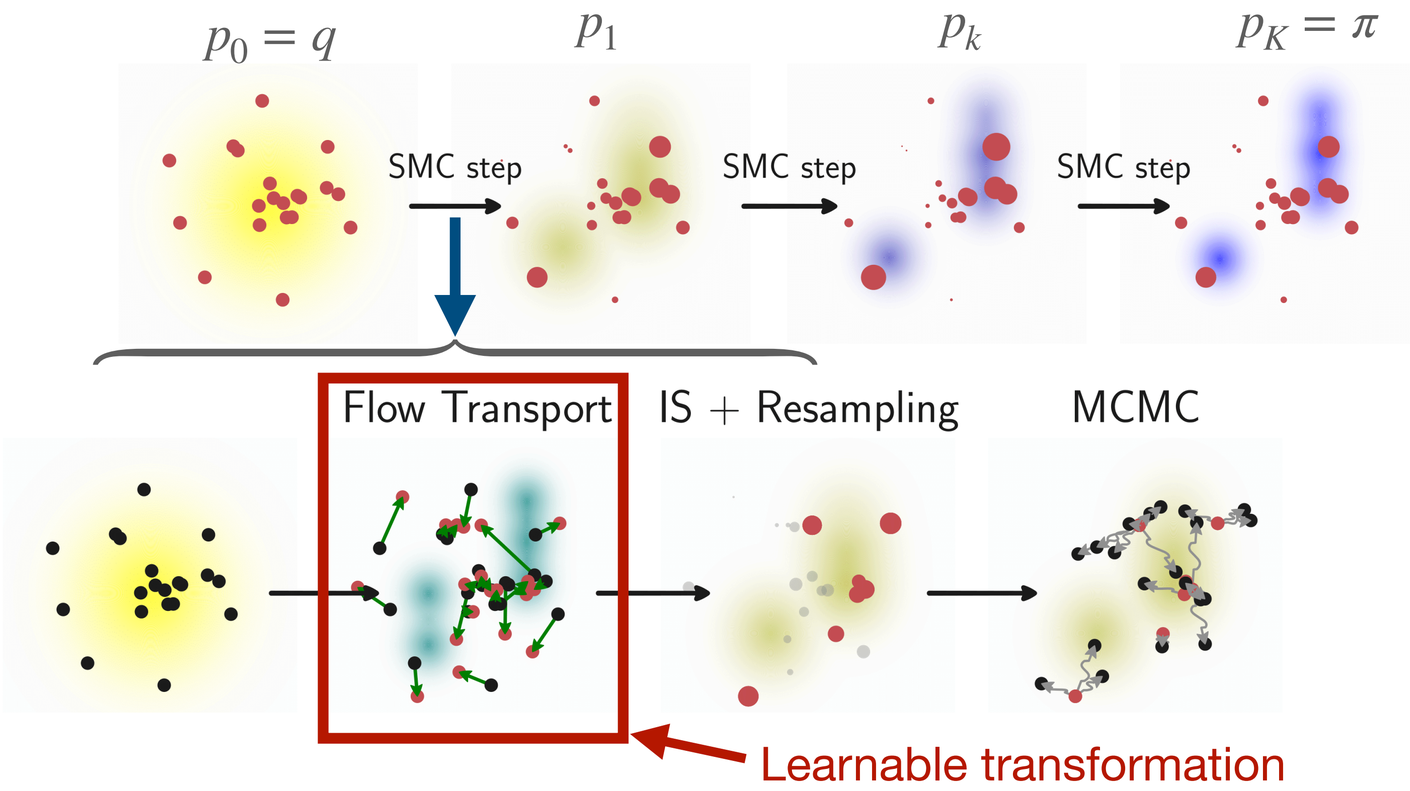
Key idea

Exploiting structure in the target density through modeling for more efficient sampling

↓

SMC with a learnable transformation

✓Fast sampling exploiting shared structure
✓Provably improves over than SMC
✓**Strong** approximation guarantees



Contributions: Principled and efficient model-based SMC samplers that capture structure in the target density

1 publication at top-tier international conference (ICML 2021 Long Oral, top 3% submissions)

Equation of State Calculations by Fast Computing Machines, Metropolis et. al., 1953

Contributions in large-scale optimization for generative models

Before my Ph.D.

Slow convergence of gradient methods for ill-conditioned generative models

During my Ph.D.

Fast and **scalable** optimizers that are **robust** to ill-conditioned models

Optimizing a generative model

$$\min_{\theta} \mathscr{L}(p_\theta)$$

Dependence on θ only through p_θ

Challenges

Hard to optimize when models are ill-conditioned

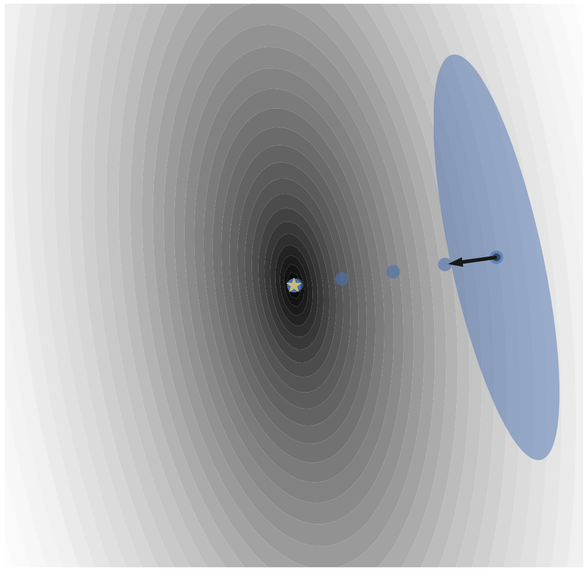
Scientific context

Fisher natural gradient

- ✓Robust to ill-conditionned models
- Not defined for Implicit generative models
- Expensive to compute for large models

Fisher natural gradient descent

$$\theta_{k+1} = \theta_k - \gamma G(\theta_k)^{-1} \nabla \mathscr{L}(p_{\theta_k})$$



Contributions: Robust, fast and scalable optimizers based on optimal transport geometry

2 publications at top-tier international conferences (ICLR 2020 Spotlight, top 6% submissions, ICLR 2021)

Contributions in large-scale optimization for generative models

Before my Ph.D.

Slow convergence of gradient methods for ill-conditioned generative models

During my Ph.D.

Fast and **scalable** optimizers that are **robust** to ill-conditioned models

Optimizing a generative model

$$\min_{\theta} \mathscr{L}(p_\theta)$$

Dependence on θ only through p_θ

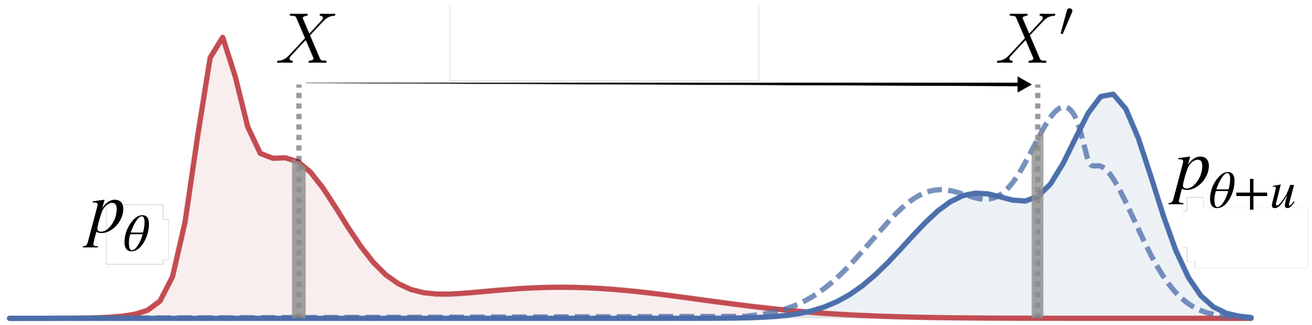
Key idea

Wasserstein Natural gradient

✓Robust to ill-conditionned models
✓Well-defined implicit generative models
✓Scalable and efficient estimation using kernel methods
✓Statistical guarantees on the estimation

Wasserstein natural gradient descent

$$\theta_{k+1} = \theta_k - \gamma G(\theta_k)^{-1} \nabla \mathscr{L}(p_{\theta_k})$$



$$\frac{1}{2} u^\top G(\theta) u \approx W^2(p_\theta, p_{\theta+u})$$

Contributions: Robust, fast and scalable optimizers based on optimal transport geometry

2 publications at top-tier international conferences (ICLR 2020 Spotlight, top 6% submissions, ICLR 2021)