Adaptive Symmetrization of the KL Divergence

Omri Ben-Dov[1], Luiz F.O. Chamon[2]

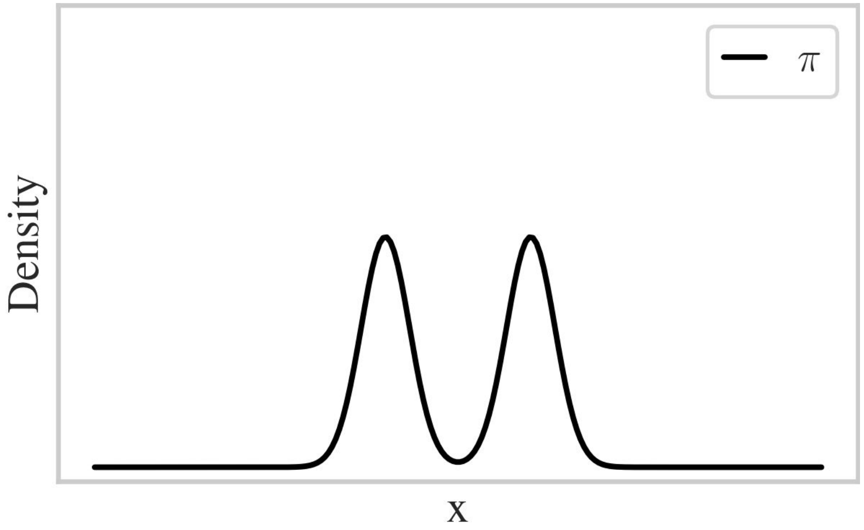
[1]Max Planck Institute for Intelligent Systems, Tübingen AI Center [2]Department of Applied Mathematics, École Polytechnique, Paris, France

It’s all about probabilities

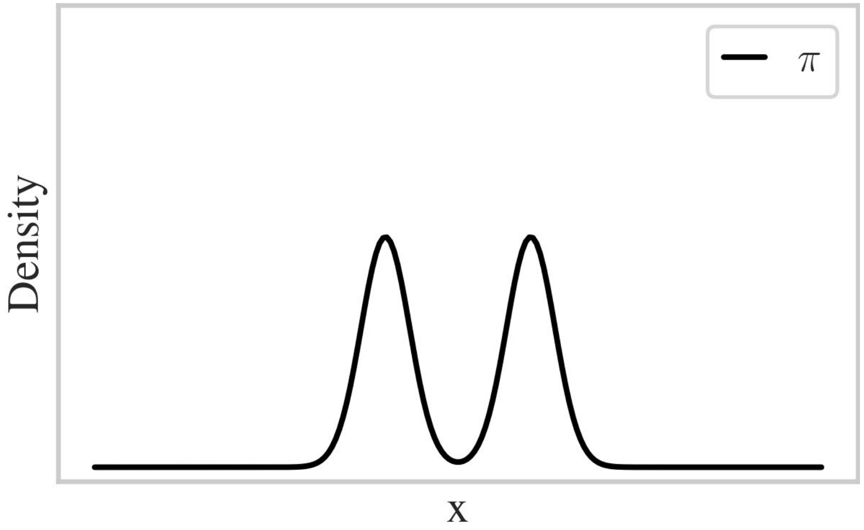
- Learn the underlying distribution π
- We can see only a finite set of observation x_i

- Parameterized distribution p_θ
- Find θ that minimizes a distance to π, based on x_i

Consequence of Distance Choice



Consequence of Distance Choice

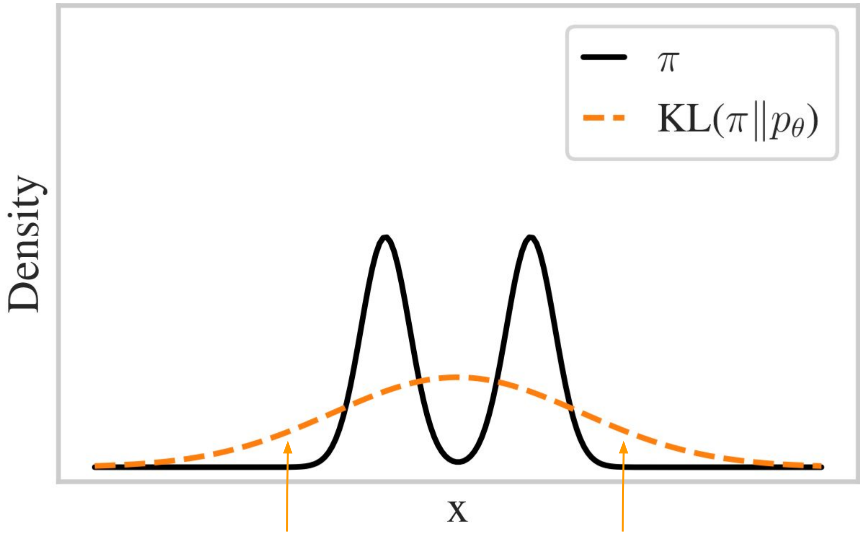


Forward KL:

$$\mathrm{KL}\left(\pi \| p_\theta\right) = \mathbb{E}_{x\sim\pi}\left[\log \pi\left(x\right) - \log p_\theta\left(x\right)\right]$$

- Tractable (= NLL, cross entropy)

Consequence of Distance Choice



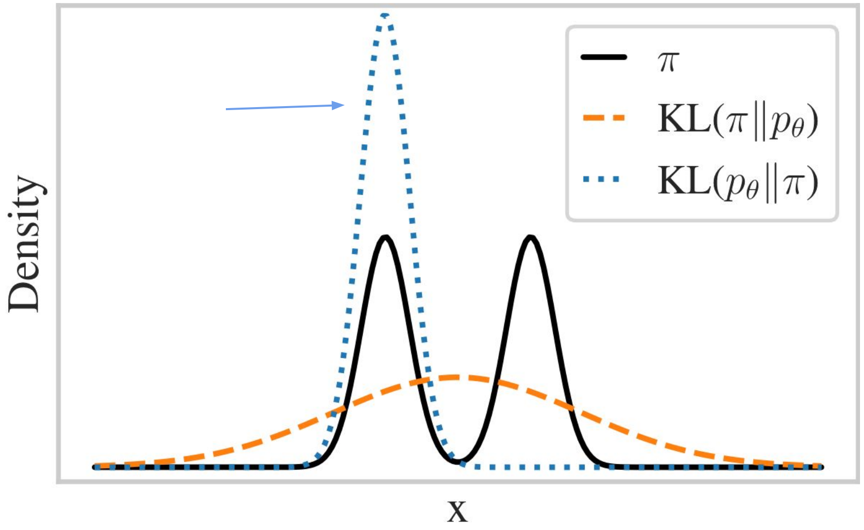
High values in low density regions

Forward KL:

$$\mathrm{KL}\left(\pi \| p_\theta\right) = \mathbb{E}_{x\sim\pi}\left[\log \pi\left(x\right) - \log p_\theta\left(x\right)\right]$$

- Tractable (= NLL, cross entropy)
- Mode-covering
- Asymmetric

Consequence of Distance Choice



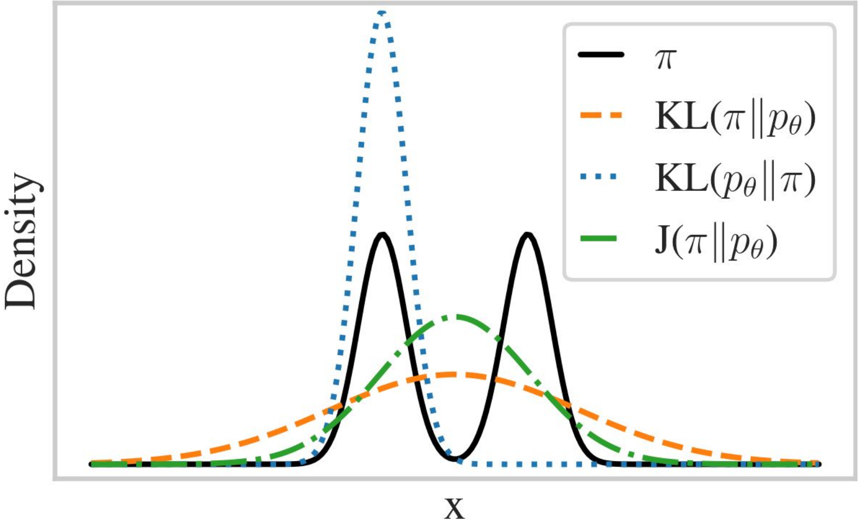
$$\mathrm{KL}\left(\pi \| p_\theta\right) = \mathbb{E}_{x\sim\pi}\left[\log \pi\left(x\right) - \log p_\theta\left(x\right)\right]$$

Reverse KL:

$$\mathrm{KL}\left(p_\theta \| \pi\right) = \mathbb{E}_{x\sim p_\theta}\left[\log p_\theta\left(x\right) - \boxed{\log \pi\left(x\right)}\right]$$

- Asymmetric
- Mode-seeking
- Intractable

Consequence of Distance Choice



$$\mathrm{KL}\left(\pi \| p_\theta\right) = \mathbb{E}_{x \sim \pi}\left[\log \pi\left(x\right) - \log p_\theta\left(x\right)\right]$$

$$\mathrm{KL}\left(p_\theta \| \pi\right) = \mathbb{E}_{x \sim p_\theta}\left[\log p_\theta\left(x\right) - \log \pi\left(x\right)\right]$$

Jeffreys divergence:

$$\mathrm{J}\left(\pi \| p_\theta\right) = \mathrm{KL}\left(\pi \| p_\theta\right) + \mathrm{KL}\left(p_\theta \| \pi\right)$$

- Symmetric
- Balances both behaviors
- Still intractable

Generative adversarial networks (GANs)

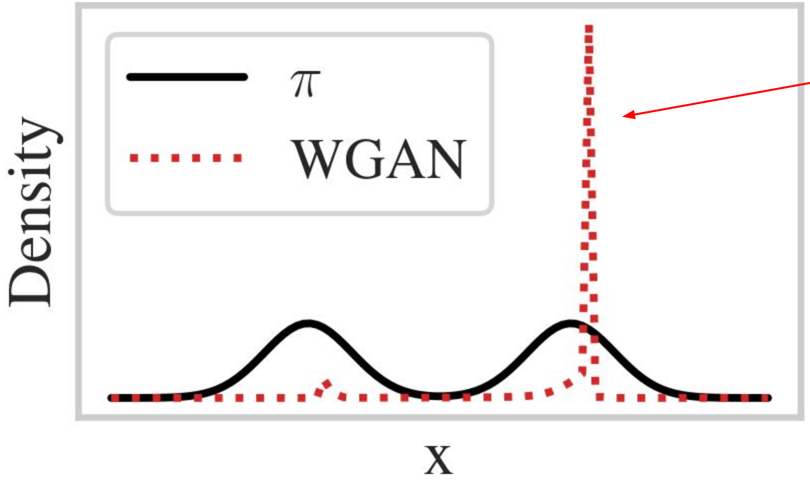
$$\min_{\theta} D_f\left(\pi \parallel p_\theta\right) = \min_{\theta} \max_{\psi} \left(\mathbb{E}_{x\sim\pi}\left[g_\psi\left(x\right)\right] - \mathbb{E}_{x\sim p_\theta}\left[f^*\left(g_\psi\left(x\right)\right)\right]\right)$$

- Symmetric
- Tractable
- Stable?

Generative adversarial networks (GANs)

Tractable but unstable

$$\min_{\theta} D_f\left(\pi \parallel p_\theta\right) = \min_{\theta}\max_{\psi}\left(\mathbb{E}_{x\sim\pi}\left[g_\psi\left(x\right)\right] - \mathbb{E}_{x\sim p_\theta}\left[f^*\left(g_\psi\left(x\right)\right)\right]\right)$$

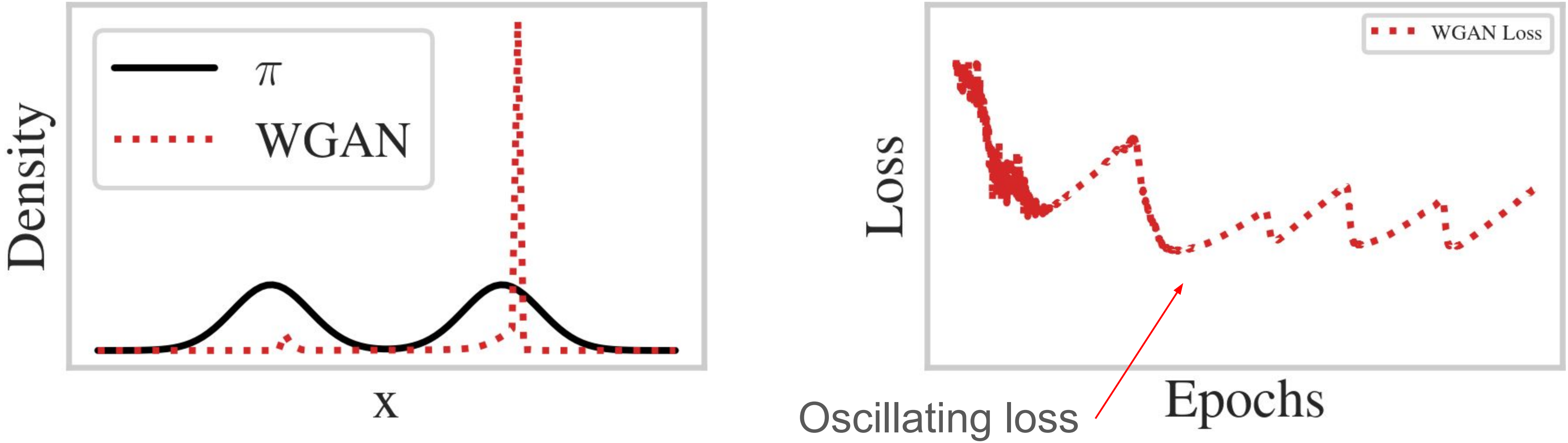


Prone to mode collapse

Generative adversarial networks (GANs)

Tractable but unstable

$$\min_{\theta} D_f\left(\pi \parallel p_\theta\right) = \min_{\theta} \max_{\psi} \left(\mathbb{E}_{x\sim\pi}\left[g_\psi\left(x\right)\right] - \mathbb{E}_{x\sim p_\theta}\left[f^*\left(g_\psi\left(x\right)\right)\right]\right)$$



Approximating the reverse KL

$$\mathrm{J}\left(\pi \| p_\theta\right) = \mathrm{KL}\left(\pi \| p_\theta\right) + \boxed{\mathrm{KL}\left(p_\theta \| \pi\right)}$$

How to make this tractable?

Approximating the reverse KL

$$\mathrm{J}\left(\pi \| p_\theta\right) = \mathrm{KL}\left(\pi \| p_\theta\right) + \mathrm{KL}\left(p_\theta \| \pi\right)$$

- GANs use an extra model to estimate the divergence
- Instead, can we use it as a proxy of the true distribution? $q_\psi \approx \pi$

$$\min_{\theta,\psi} \quad \mathrm{KL}\left(\pi \| p_\theta\right) + \mathrm{KL}\left(p_\theta \| q_\psi\right)$$

- We need to force q_ψ to be *close enough*

Approximating the reverse KL

$J(\pi \| p_\theta) = \mathrm{KL}(\pi \| p_\theta) + \mathrm{KL}(p_\theta \| \pi)$

- GANs use an extra model to estimate the divergence
- In...
- We need to force q_ψ to be close enough

Constrained optimization:

Forward KL — Reverse KL (appx.)

$$\begin{aligned} \underset{\theta,\psi \in \mathbb{R}^k}{\text{minimize}} \quad & \underbrace{D_{\mathrm{KL}}(\pi \parallel p_\theta)}_{\text{Forward KL}} + \underbrace{D_{\mathrm{KL}}(p_\theta \parallel q_\psi)}_{\text{Reverse KL (appx.)}} \\ \text{subject to} \quad & \underbrace{D_{\mathrm{KL}}(\pi \parallel q_\psi) \leq \epsilon}_{\text{Proxy quality}} \end{aligned}$$

Proxy quality

Critical Decisions

$$\begin{aligned} \underset{\theta, \psi \in \mathbb{R}^k}{\text{minimize}} \quad & D_{\mathrm{KL}}(\pi \parallel p_\theta) + D_{\mathrm{KL}}(p_\theta \parallel q_\psi) \\ \text{subject to} \quad & D_{\mathrm{KL}}(\pi \parallel q_\psi) \leq \epsilon \end{aligned}$$

1. What epsilon to use? Is it feasible?

2. Should the KLs have the same weight as the forward? Will it focus on mode-seeking (reverse) or mode-covering (forward)?

Critical Decisions

Hounie, I., Ribeiro, A., & Chamon, L. F. (2023). Resilient constrained learning. *Advances in Neural Information Processing Systems*, *36*, 71767-71798.

arXiv

Resilient Constrained Learning:

optimize the constraint specifications

$$\begin{aligned}\underset{\substack{\theta,\psi\in\mathbb{R}^k\\ \epsilon_{\mathrm{fw}},\epsilon_{\mathrm{rv}},\epsilon_{\mathrm{prx}}\geq 0}}{\text{minimize}} \quad & \epsilon_{\mathrm{fw}}^2+\epsilon_{\mathrm{rv}}^2+\epsilon_{\mathrm{prx}}^2\\ \text{subject to} \quad & D_{\mathrm{KL}}(\pi \,\|\, p_\theta)\leq \epsilon_{\mathrm{fw}}, \quad D_{\mathrm{KL}}(p_\theta \,\|\, q_\psi)\leq \epsilon_{\mathrm{rv}}\\ & D_{\mathrm{KL}}(\pi \,\|\, q_\psi)\leq \epsilon_{\mathrm{prx}}\end{aligned}$$

Adaptive: Shift focus to the difficult constraints

1. What
2. Shoul ... on
   mode

Unconstrained Dual Problem

Solving constrained optimization usually requires heavy computations.

We can just solve the unconstrained dual problem

$$D^\star = \max_{\boldsymbol{\lambda} \geq 0} \min_{\theta, \psi \in \mathbb{R}^k, \boldsymbol{\epsilon} \geq 0} \mathcal{L}(\theta, \psi, \boldsymbol{\epsilon}, \boldsymbol{\lambda})$$

With the intuitive Lagrangian

$$\mathcal{L}(\theta, \psi, \boldsymbol{\epsilon}, \boldsymbol{\lambda}) = \epsilon_{\text{fw}}^2 + \epsilon_{\text{rv}}^2 + \epsilon_{\text{prx}}^2 + \lambda_{\text{fw}} \left[-\frac{1}{N} \sum_{i=1}^{N} \log p_\theta(x_i) - \epsilon_{\text{fw}} \right] + \lambda_{\text{rv}} \left[D_{\text{KL}}(p_\theta \parallel q_\psi) - \epsilon_{\text{rv}} \right] + \lambda_{\text{prx}} \left[-\frac{1}{N} \sum_{i=1}^{N} \log q_\psi(x_i) - \epsilon_{\text{prx}} \right] + \lambda_h \left[h(p_\theta, q_\psi) - c \right]$$

Model Choice

arXiv

$$p_\theta$$

Main model: Normalizing flow

- Exact probability
- Limited architecture
- Quick sampling

$$q_\psi$$

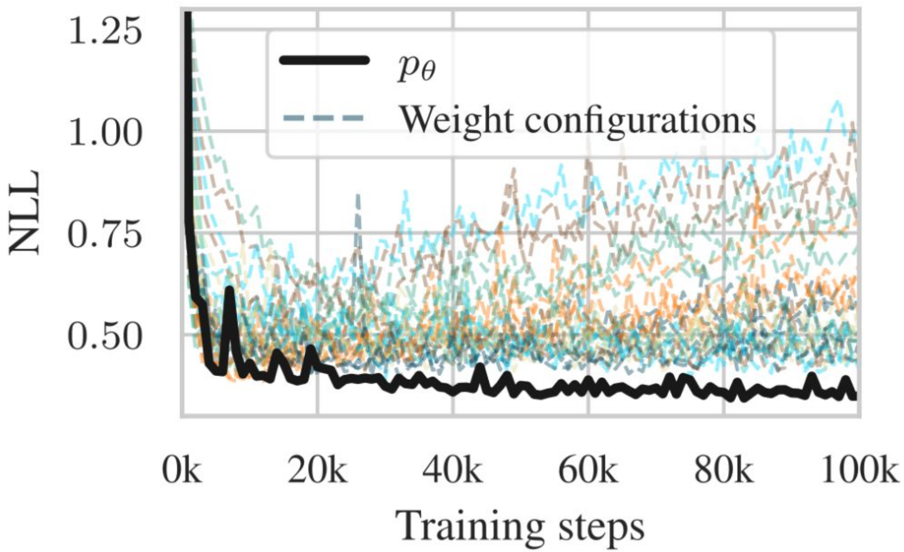
Proxy: Energy-based model

- Unnormalized distributions
- Architecture freedom
- Slow sampling

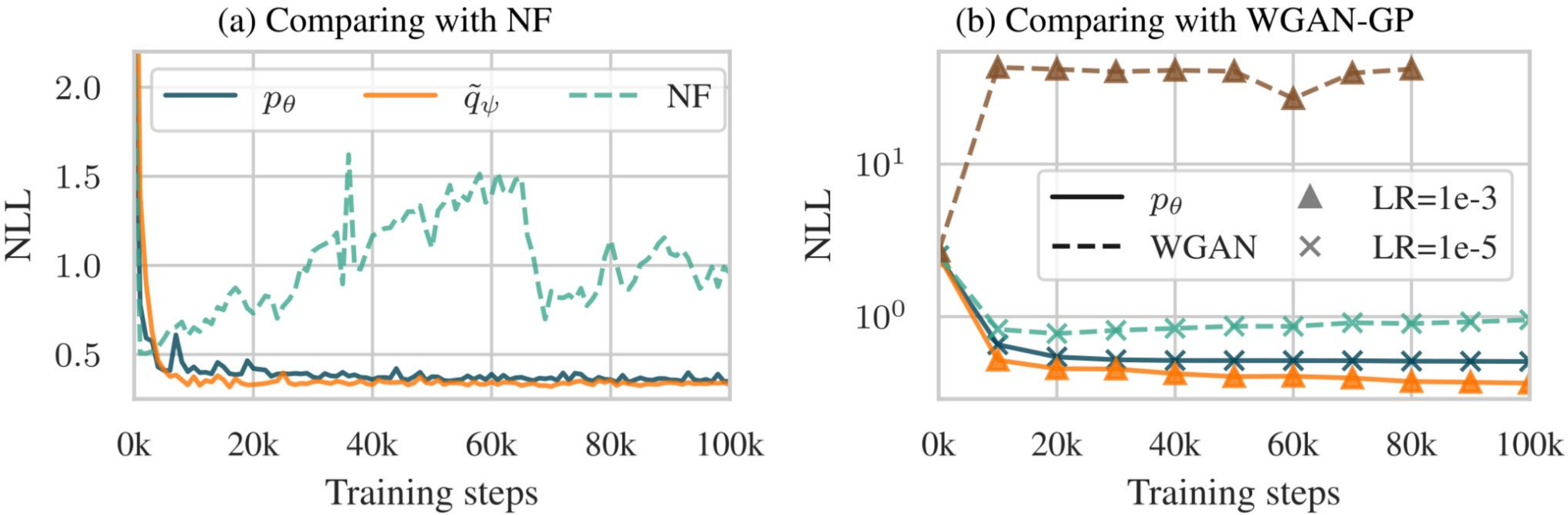
Ablation of adaptivity

Fixed weights objective:

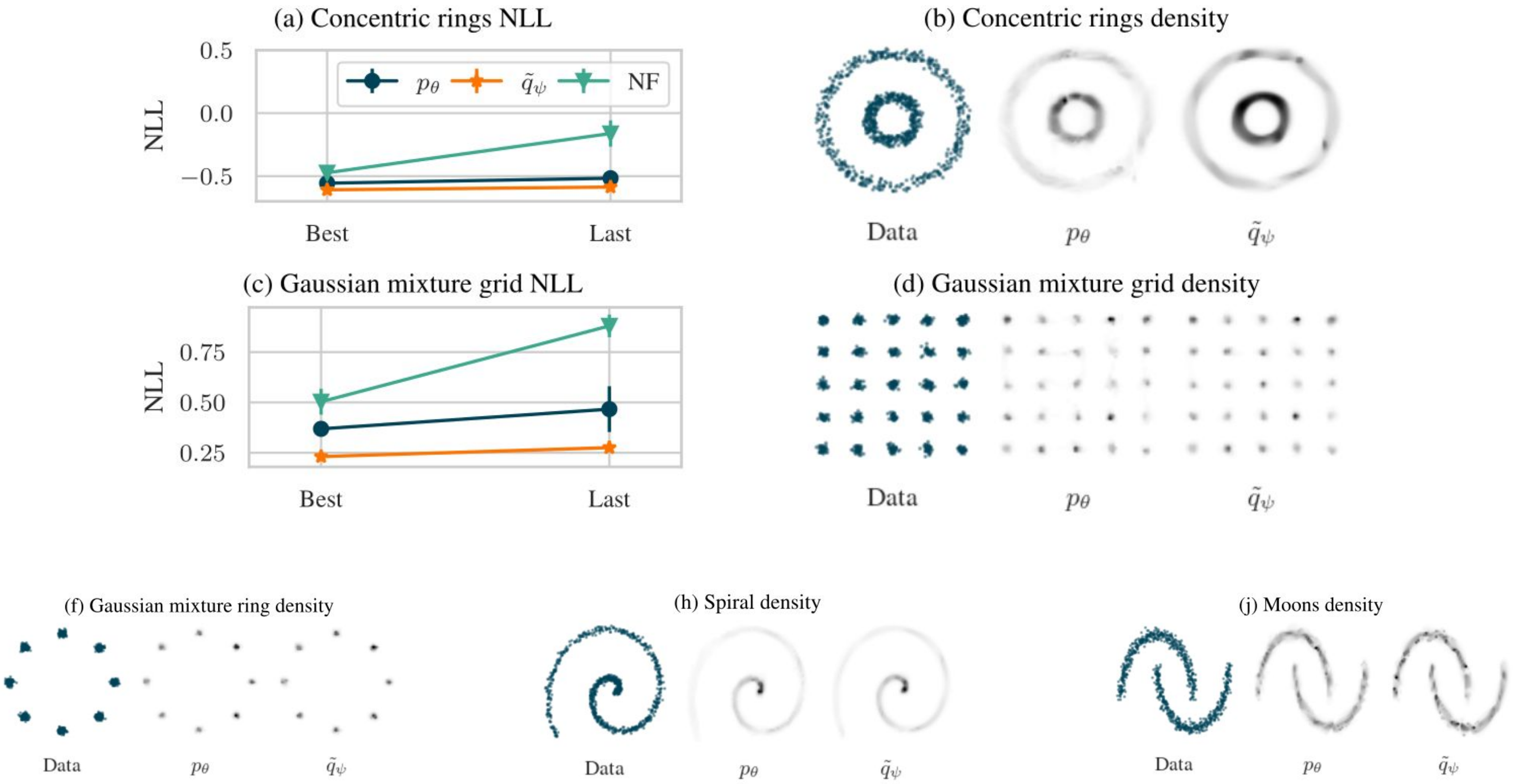
$$\underset{\theta,\psi\in\mathbb{R}^k}{\text{minimize}}\ w_{\text{fw}} D_{\text{KL}}(\pi \parallel p_\theta) + w_{\text{rv}} D_{\text{KL}}(p_\theta \parallel q_\psi) + w_{\text{prx}} D_{\text{KL}}(\pi \parallel q_\psi)$$



Improvement over KL and WGAN

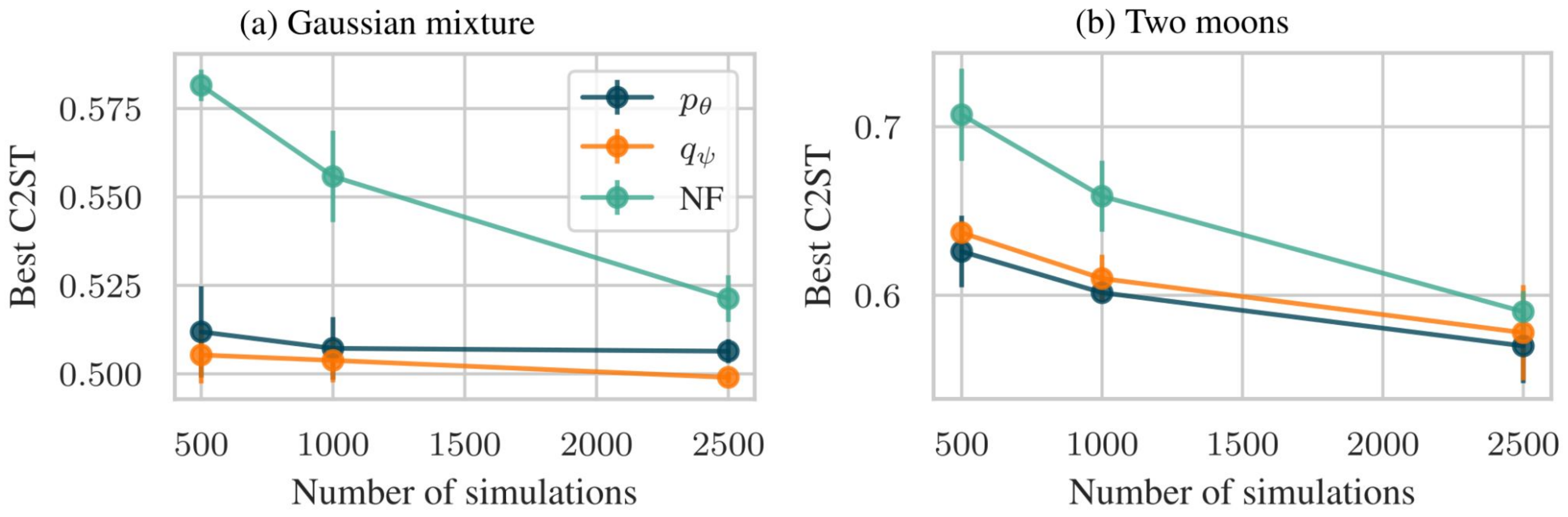


Density estimation



Simulation-based inference (SBI)

Learning conditional distributions



C2ST: classifier 2-sample test, closer to 0.5 is better

Conclusion

arXiv

1. Goal: Tractable and stable optimization of a symmetric divergence.
2. Approximate the reverse KL with a proxy model.
3. Resilient constrained learning makes the constraints adaptive.
4. Solve the unconstrained dual problem.
5. More accurate than KL, more stable than GANs and needs less training data.

More details and experiments in the preprint https://arxiv.org/abs/2511.11159