Domain-Constrained Diffusion Models to Synthesize Tabular Data: A Case Study in Power Systems

Milad Hoseinpour, Vladimir Dvorkin

Department of Electrical and Computer Engineering, University of Michigan

Motivation

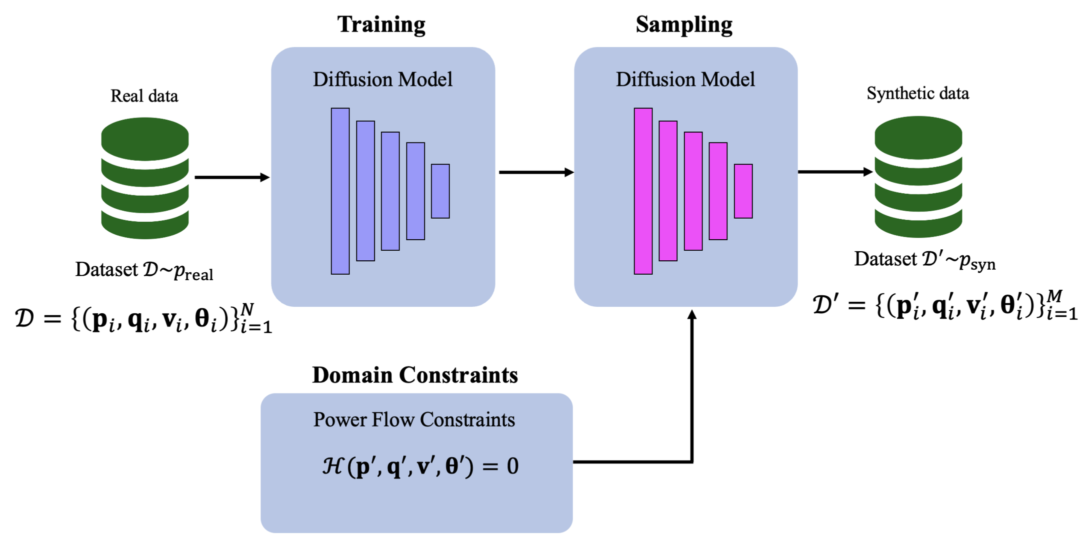
- Access to real-world data is often limited due to privacy, security, and legal barriers, hindering the training of Machine Learning (ML) models across domains.

A synthetic dataset is artificially generated data that enjoys the statistical properties of real-world data without containing any actual records.

- High-quality synthetic data must go beyond statistics by adhering to domain-specific constraints that ensure real-world feasibility.

Problem Setup

Goal: Given a dataset including real power flow data points, we aim to synthesize (1) statistically representative and (2) high fidelity power flow data points:



A high-level view of the problem setup.

Diffusion Models

- **Forward** diffusion process gradually adds noise to input:

$$\mathbf{x}_t = \sqrt{\bar{\alpha}_t}\mathbf{x}_0 + \sqrt{1-\bar{\alpha}_t}\epsilon_t, , \quad \epsilon_t \sim \mathcal{N}(0, \mathbb{I}), t \in (0, T].$$

- **Reverse** diffusion process learns to generate data by denoising:

$$\mathbf{x}_{t-1} = \mu_\theta(\mathbf{x}_t, t) + \sigma_t \epsilon_t, \quad \epsilon_t \sim \mathcal{N}(0, \mathbb{I}), t \in (T, 0].$$

- **Training:** The loss function to train the denoiser neural network:

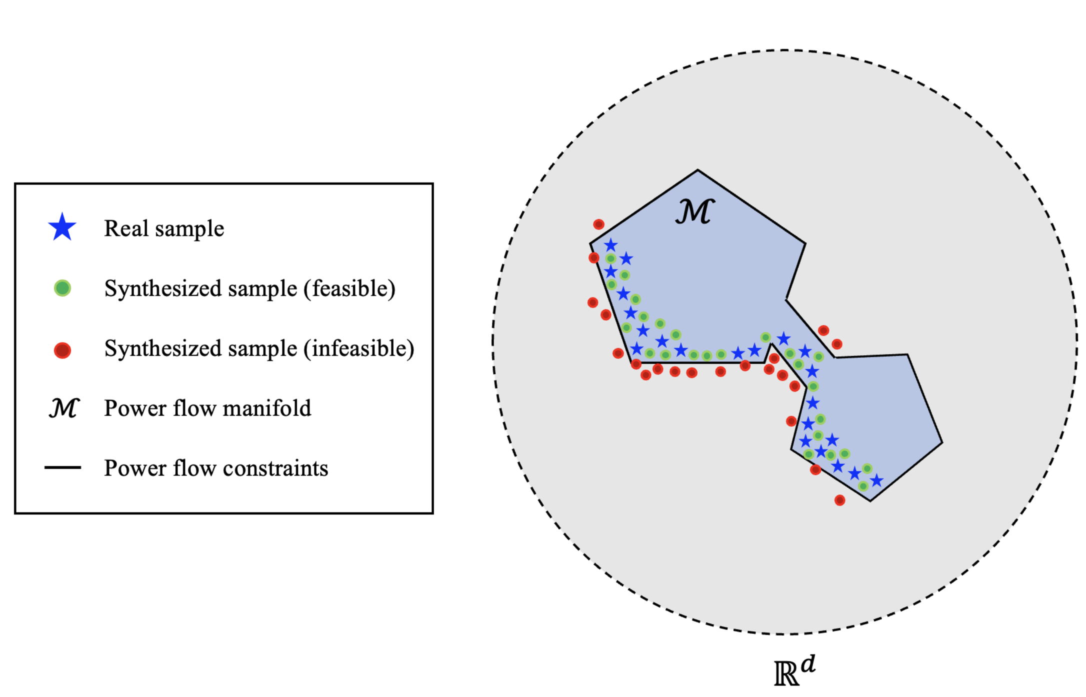
$$\mathcal{L}_{\text{diff}} = \mathbb{E}_{\mathbf{x}_0, \epsilon, t}\left[\|\epsilon - \epsilon_\theta(\mathbf{x}_t, t)\|^2\right].$$

- **Sampling:**

$$\mathbf{x}_{t-1} = \frac{\sqrt{\alpha_t}(1-\bar{\alpha}_{t-1})}{1-\bar{\alpha}_t}\mathbf{x}_t + \frac{\sqrt{\bar{\alpha}_{t-1}}\beta_t}{1-\bar{\alpha}_t}\hat{\mathbf{x}}_0 + \sigma_t z, \quad z \sim \mathcal{N}(0, \mathbb{I}), t \in [T, 0).$$

Diffusion Guidance based on Power Flow Constraints

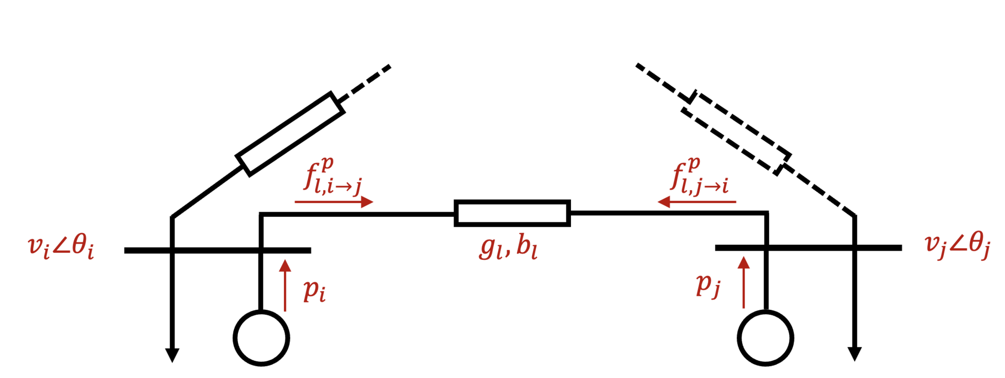
- In practice, a diffusion model may generate power flow data points that are **infeasible** due to learning and sampling errors.



- Active and reactive power balance constraints:

$$p_b - \sum_{l \in \mathcal{L}: i=b} f^p_{l, i\to j} - \sum_{l \in \mathcal{L}: j=b} f^p_{l, j\to i} = 0, \quad \forall b \in \mathcal{B}$$

$$q_b - \sum_{l \in \mathcal{L}: i=b} f^q_{l, i\to j} - \sum_{l \in \mathcal{L}: j=b} f^q_{l, j\to i} = 0, \quad \forall b \in \mathcal{B}$$



How can we enforce power flow constraints in generated samples?

- Our goal is to minimize the data consistency loss $R_{\mathcal{H}}(\mathbf{x})$ on the clean data manifold $\mathcal{M}$:

$$\min_{\mathbf{x} \in \mathcal{M}} R_{\mathcal{H}}(\mathbf{x}),$$

where $\mathcal{H}(\cdot)$ encodes the equality constraints and

$$R_{\mathcal{H}}(\mathbf{x}) = \|\mathcal{H}(\mathbf{x})\|_2^2.$$

- We take one step of Riemannian gradient descent on $\mathcal{M}$:

$$\hat{\mathbf{x}}'_{0|t} = \hat{\mathbf{x}}_{0|t} - \tau_t \text{ grad } R_{\mathcal{H}}(\hat{\mathbf{x}}_{0|t}),$$

where

$$\text{grad } R_{\mathcal{H}}(\hat{\mathbf{x}}_{0|t}) = \mathcal{P}_{T_{\hat{\mathbf{x}}_{0|t}}\mathcal{M}}\left(\nabla_{\mathbf{x}_t} R_{\mathcal{H}}(\hat{\mathbf{x}}_{0|t})\right).$$

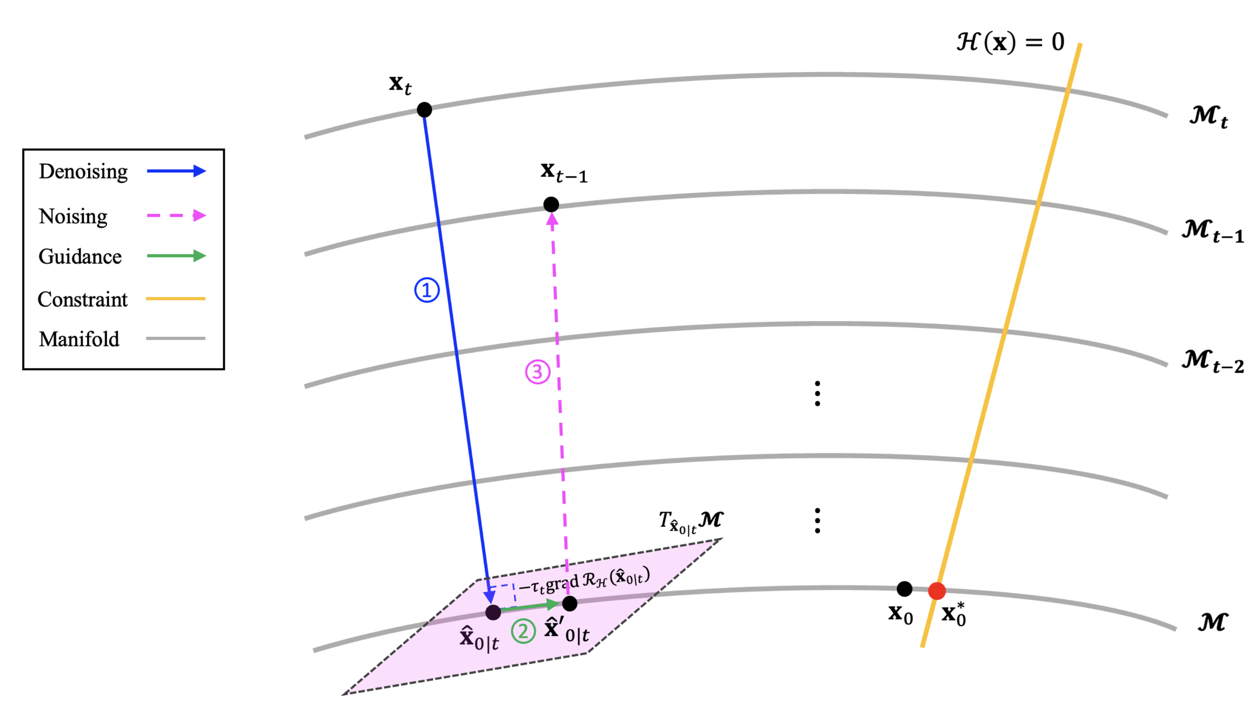
- Under affine subspace assumption of clean data manifold $\mathcal{M}$, we can prove:

$$\mathcal{P}_{T_{\hat{\mathbf{x}}_{0|t}}\mathcal{M}}\left(\nabla_{\mathbf{x}_t} R_{\mathcal{H}}(\hat{\mathbf{x}}_{0|t})\right) \approx \nabla_{\mathbf{x}_t} R_{\mathcal{H}}(\hat{\mathbf{x}}_{0|t}).$$

$$\hat{\mathbf{x}}'_{0|t} = \hat{\mathbf{x}}_{0|t} - \lambda_t \nabla_{\mathbf{x}_t} R_{\mathcal{H}}(\hat{\mathbf{x}}_{0|t}).$$

Sampling steps can be characterized as transitions from $\mathcal{M}_i$ to $\mathcal{M}_{i-1}$:

- (1) we do a denoising step based on $\mathbf{x}_t$ and estimate the clean data $\hat{\mathbf{x}}_0$,
- (2) add the gradient guidance term,
- (3) add noise w.r.t. the corresponding noise schedule and obtain $\mathbf{x}_{t-1}$.

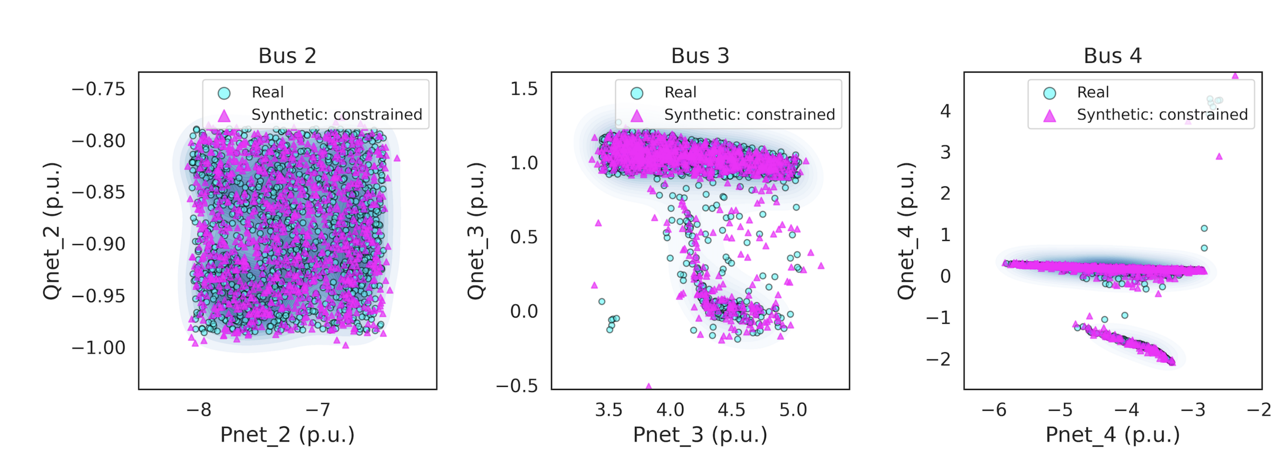


Geometry of sampling with guidance.

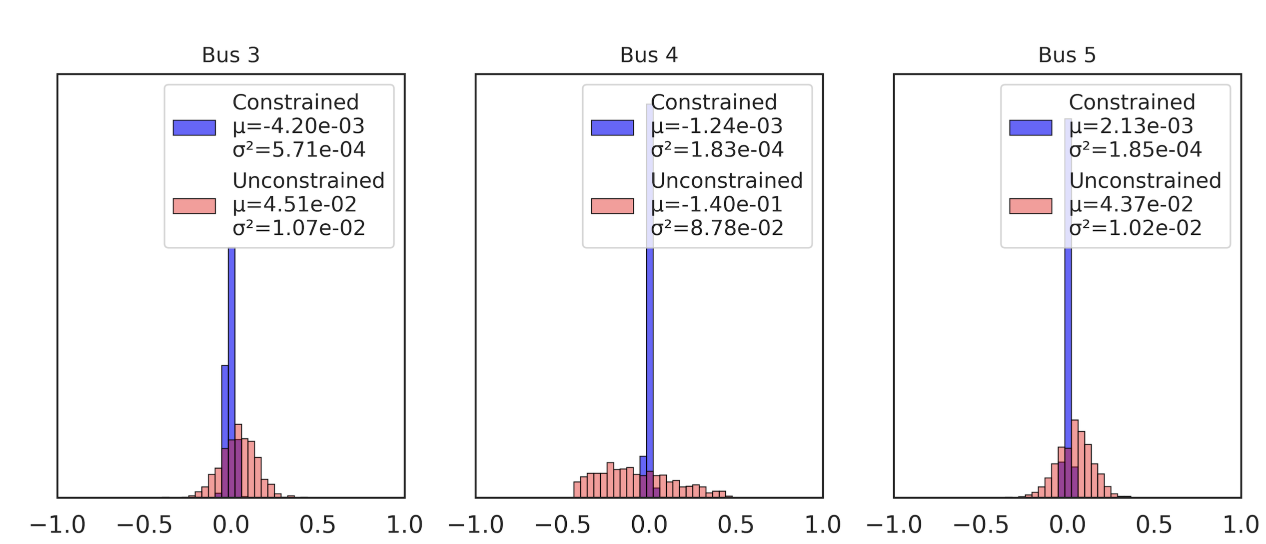
Results

Test System: PJM 5-BUS System

- Distribution Matching: joint distribution



- Histograms of violation magnitude for active power balance constraints



Conclusion

- Synthesized power flow data points effectively capture the pattern, domain, and modes of underlying distributions of the real data.
- The proposed gradient guidance approach successfully enforces power flow constraints during sampling, ensuring the feasibility of the generated data.

ICML 2025 Workshop on DataWorld: Unifying Data Curation Frameworks Across Domains Email: miladh@umich.edu