Score-Regularized GAN for Simulating Private Equity Contribution Patterns

Author: Param Sejpal, M.S. Data Science (pns2129@columbia.edu)

COLUMBIA | ENGINEERING
The Fu Foundation School of Engineering and Applied Science

Abstract

We propose a Score-regularized Generative Adversarial Network (GAN) framework for simulating Private Equity (PE) contribution cash flows. This novel approach combines GANs with a pre-trained trend classifier, addressing issues like data scarcity and realistic cash flow shape preservation. By utilizing Brownian bridge augmentation and a trend-aware scoring mechanism, we generate high-quality synthetic data that closely mirrors real PE cash flow patterns, enhancing forecasting and risk modeling.

Introduction

Private Equity (PE) contribution data typically follow a unique pattern of **rapid early growth** followed by a **plateau**, making it difficult for **traditional GANs** to generate realistic synthetic sequences. To address this, we propose a **Score-regularized GAN framework** that incorporates **financial domain knowledge**. By using a **trend-aware classifier** to evaluate the quality of generated data, our model ensures not only **statistical similarity** but also accurate **economic shape fidelity**, producing more realistic **PE cash flow simulations**. At the heart of our approach is the **trend-aware score classifier**, which guides the **GAN** in generating realistic sequences. This classifier analyzes the shape of each generated curve using **changepoint detection** and **slope analysis** to compute a **trend score**. The score reflects how well the sequence matches the expected PE pattern - a **sharp early growth** followed by **stabilization**. Only **high-scoring sequences** are used to refine the generator, ensuring that the model learns to produce **economically plausible** and **high-quality synthetic data**.

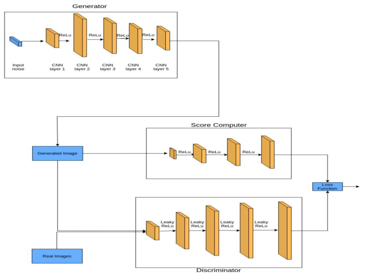

Fig 1: Score-regularized GAN Framework Overview

Dataset

The dataset consists of the following key components:
1. Identifiers: Fund metadata (names, vintages, classifications).
2. Cash Flows: Raw data for individual contributions.
3. Benchmark Data: Aggregated quarterly data used as ground truth to evaluate model outputs.
4. Synthetic Data: Generated using Brownian bridges and an iTransformer-GAN to augment the training data for score computer
5. Filtered Data: High-quality synthetic data that pass the trend score filtering process.

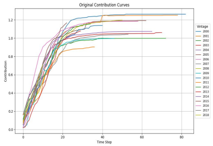

Fig 2: Contribution data

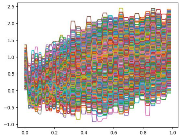

Fig 3: Synthetic Data

Score Computer

The following steps outline the process for preparing and training the Score Machine, a classifier later used to penalize the GAN:

1. Preprocessing & Standardization – Sequences are padded to a fixed length; undersized sequences are excluded.
2. Augmentation – Brownian bridge resampling generates synthetic PE-like time series to expand the dataset.
3. Trend Score Calculation – Sequences are split at a changepoint; slope differences (pre vs. post) adjusted by R^2 define a trend score.
4. Labeling & Data Preparation – Real + Brownian sequences labeled good (1); GAN sequences labeled good (1) if trend score ≥ median threshold, else bad (0).
5. Classifier Training – A 3-layer feedforward neural net is trained with BCEWithLogitsLoss and Adam
6. Use in GAN Training – The frozen Score Machine is integrated as a regularizer; generator outputs classified as bad incur a penalty, steering GAN toward trend-consistent "good" sequences.

iTransformer-GAN

The following steps describe the process by which the iTransformer GAN is trained and regularized by a pre-trained Score Machine:

1. Data selection & standardization – Vintage contribution series are selected and all sequences are standardized to a fixed length for batching.
2. Augmentation – Brownian-bridge resampling is used to expand the dataset with smooth, plausible time-series variants.
3. Model architecture & normalization – Transformer-based generator and discriminator are constructed and stabilization/normalization layers are applied for reliable sequence modeling.
4. Autocorrelation-aware comparison – Autocorrelation weights are computed to emphasize temporal structure when real and generated sequences are compared.
5. Wasserstein objective + gradient penalty – Training is performed with a Wasserstein-style loss and a gradient penalty is applied to stabilize the critic.
6. Adversarial training loop – Discriminator and generator updates are alternated in the GAN loop to improve sample realism.
7. Trend regularization via Score Machine – The frozen Score Machine is integrated as a regularizer: its BCE penalty is added to the generator loss so outputs classified as "bad" are penalized.
8. Smoothness regularizers – Total-variation and Fourier penalties are added to discourage abrupt jumps and promote investment-like smoothness.

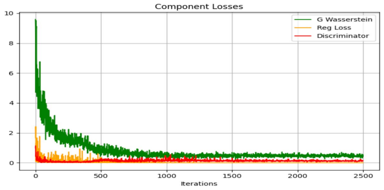

Figure 4: GAN Training Losses

Results & Conclusion

The Score-regularized iTransformer GAN produced realistic private equity (PE) contribution J-curves that broadly matched real temporal patterns. Incorporating a trend-aware Score Machine significantly improved generation quality, making the framework a robust tool for scenario analysis and risk modeling in PE.

Visual Results:

Side-by-side panels and mean J-curves with confidence bands showed that GAN-generated series closely aligned with real (Original + Brownian augmented) data, producing smooth, investment-like patterns.

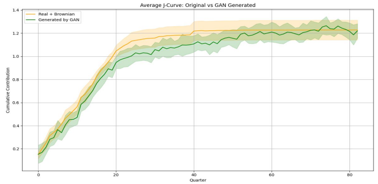

Figure 5: Original Data vs GAN Outputs

Quantitative Results:

1. Wasserstein distance (0.059): Generated distributions closely resemble real data.
2. Row fidelity (0.668): Temporal correlations are well preserved, showing strong capture of sequence dynamics.
3. Privacy (DCR median = 0.672): Synthetic records maintain safe distance from real data; no overlaps detected.
4. Utility (0.958): Models trained on synthetic data generalized strongly to real data, confirming practical usability.
5. Synthesis score (1.000): No exact matches occurred (0/200), ensuring novelty of generated records.

Future Work:

1. Hard negative sampling – introduce near-realistic but trend-violating sequences to strengthen discriminator learning.
2. Dataset expansion – include larger datasets and additional financial features for improved generalization.
3. Multi-faceted modeling – extend beyond contributions to also model distributions and NAV, enabling a fuller PE portfolio view.

Under the Mentorship of Professor Ali Hirsa and Miao Wang
AI Applications in Finance 2025 (IEOR E4743)