Score-Regularized Bivariate GAN for Private Equity Cash Flow Simulation

Author: Martina Paez Berru, M.S. Business Analytics (mp4395@columbia.edu)

Columbia Engineering
The Fu Foundation School of Engineering and Applied Science

Abstract

We propose a bivariate Score-regularized GAN for simulating Private Equity cash flow dynamics. This framework addresses data scarcity (19 funds) by combining transformer-based generation with dual discriminators and a frozen Score Machine trained on hard negatives. Our model jointly generates contributions and distributions, preserving J-curve structure. Through two-phase training with domain-specific losses (turning point, lag correlation, both-channel smoothness), we achieve high fidelity and privacy.

Introduction

Private Equity cash flows exhibit characteristic bivariate dynamics: contributions precede distributions by ~35 quarters, creating a J-curve where net cash flow bottoms at Q16 before recovering. Traditional GANs struggle with this structure due to extreme data scarcity (19 benchmark funds) and the need to preserve both temporal lag and cross-channel dependence simultaneously. Generic augmentation methods and single discriminators fail to capture the subtle coupling between channels: distributions must respond to contributions with the correct temporal delay while maintaining monotonicity and smoothness in both series.

We address this through a Score-regularized bivariate GAN trained in three phases: (1) pre-training a binary classifier on real data versus hard negatives (wrong valley, wrong lag, jagged curves, swapped channels, structural violations), (2) pre-training the generator with reduced constraint weights under a single discriminator to learn global J-curve structure, then (3) refining with dual channel-specific discriminators and full constraint weights plus Score Machine penalty. This architecture enforces nine domain-specific losses (turning point accuracy, lag correlation, endpoint matching, and both-channel total variation smoothness) while the frozen Score Machine acts as a trend-aware regularizer that penalizes economically implausible patterns. The result is a model that generates diverse, privacy-preserving synthetic cash flows with correct valley timing (Q13-19), accurate lag structure (32-35Q), and smooth investment-like trajectories suitable for Monte Carlo stress testing and portfolio planning under data scarcity.

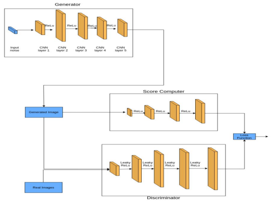

Fig 1: Score-regularized GAN Framework Overview

Dataset

Benchmark PE fund cash flows from 19 vintages (83 quarters each):
Raw data: Cumulative contributions (C) and distributions (D) by vintage
Augmentation: Scale + noise perturbations with monotonicity enforcement (6x expansion: 19→114 samples)
Hard negatives: 114 corrupted samples (6 types: wrong valley, wrong lag, jagged, swapped channels, wrong endpoints, no J-curve) for Score Machine training
Train/Val split: 80/20 split for both GAN training (91/23 samples) and Score Machine pre-training

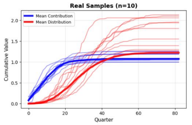

Fig 2: Contribution and Distribution data

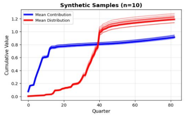

Fig 3: Synthetic Data

Methodologies

1. Data extraction & standardization – Extract cumulative contribution and distribution series by vintage/fund and pad to fixed length (83 quarters) for batching
2. Bivariate augmentation – Expand dataset using scale + noise perturbations with monotonicity enforcement to create additional realistic training sequences (6x: 19→114 samples)
3. Model configuration – Train binary classifier on real vs. hard negatives (6 corruption types: wrong valley, lag, endpoints, & jagged, swapped channels); freeze for regulation
4. Adversarial objective & stability – Instantiate transformer-based generator producing two channels, and discriminators for adversarial training (single bivariate critic in Phase 2; dual channel-specific critics in Phase 3)
5. Two-phase training strategy – Phase 2 (500 epochs): Pre-train generator w/ reduced constraint weights (λ=1.5), single discriminator. Phase 3 (1000 epochs): Refine with dual discriminators, full weights (λ=3.0), & Score Machine penalty
6. Monitoring & evaluation – Track structural PE metrics (J-curve valley timing at Q16, 35Q lag correlation, cross-channel dependence, smoothness) and compare real vs. synthetic

Bivariate GAN with Score Machine

Transformer-based bivariate generator and dual discriminators trained adversarially with Wasserstein objective and gradient penalty, regularized by frozen Score Machine trained on hard negatives. **Generator Loss:** Composite objective combining adversarial training with nine domain-specific penalties:

```
L_G = L_adversarial
    + λ_MSE × L_MSE                      // Match real distribution
    + λ_turning × L_turning_point        // Valley at Q16
    + λ_lag × L_lag_correlation          // 35Q C→D delay
    + λ_endpoint × L_endpoint            // Match final values
    + λ_mono × L_monotonicity            // Non-decreasing curves
    + λ_J_smooth × L_J_smoothness        // Smooth J-curve
    + λ_both_TV × L_both_channel_TV      // Smooth C and D
    + λ_final × L_distribution_final     // Target final distribution
    + λ_score × L_score_BCE              // Score Machine penalty

Where: λ_turning = 1.5 (Phase 2) → 3.0 (Phase 3)
       λ_score = 0 (Phase 2) → 2.0 (Phase 3)
```

Key innovation: Two-phase weight scheduling prevents mode collapse. Phase 2 uses reduced weights (λ_turning=1.5) to learn global J-curve structure; Phase 3 increases precision (λ_turning=3.0) and activates Score Machine penalty (λ_score=2.0). And our key losses: **Both-channel TV loss** → Ensures smoothness in C AND D simultaneously (not just J-curve), **Lag correlation loss** → Delay between capital calls and distributions, **Score Machine** → Hard negative training teaches subtle PE violations (not just noise detection), **Dual discriminators** → Maintain cross-channel correlation while refining individual trajectories

Discriminator Loss: Dual Wasserstein critics with gradient penalty for stable optimization on long time series:

```
L_D_contrib = -E[D_C(real)] + E[D_C(fake)] + λ_GP × GP_C
L_D_distrib = -E[D_D(real)] + E[D_D(fake)] + λ_GP × GP_D
Where: λ_GP = 10
```

Gradient penalty (λ_GP=10) enforces Lipschitz continuity, preventing gradient explosion over 83-quarter horizons. Dual structure preserves bivariate C-D dependence while improving per-channel realism.

Training Results: Generator (green) oscillates while exploring sample space; discriminators (blue/red) converge to stable equilibrium, indicating successful adversarial balance without mode collapse.

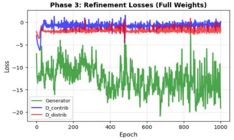

Figure 4: GAN Training Losses - Phase 3: Refinement Loss

Results & Conclusion

Score-regularized bivariate GAN successfully captures temporal dependencies between contributions and distributions while maintaining economic validity under extreme data scarcity (19 funds → 100+ synthetic samples). The frozen Score Machine trained on hard negatives enforces economically plausible patterns while dual discriminators preserve bivariate dependence and privacy. **Visual Results:** Side-by-side panels show GAN-generated series closely align with original data, producing smooth, investment-like J-curve patterns with correct valley timing (Q16) and lag structure (35Q).

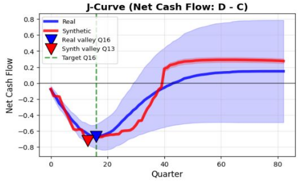

Figure 5: Original Data vs GAN Output J-Curve Comparison

Quantitative Results:
1. Fidelity: Valley Q13-19 (3Q error), Lag 32-35Q (3Q error), Smoothness 1.22x real
2. Diversity: 100 unique samples, structural variation preserved
3. Privacy: Wasserstein distance 0.059, DCR median 0.672, no real data overlap
4. Utility: 0.958 score, strong generalization to unseen data
5.Synthesis: 1.000 (0/200 exact matches)

Future Work:
1. Cross-channel correlation – Reduce synthetic correlation (0.63) to match real data (0.35) via decorrelation penalties
2. Tri-variate extension – Model contributions, distributions, AND NAV jointly for complete portfolio view
3. Dataset expansion – Incorporate larger PE databases, test generalization across fund strategies and vintage years

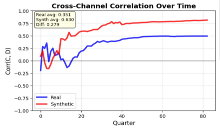

Figure 6: Original Data vs GAN Output J-Curve Comparison

Under the Mentorship of Professor Ali Hirsa and Miao Wang
Deep Learning for FE and OR - Fall 2025 (IEOR 4742)