Extra content for obsessed people

Here's a quick reference for the GAN variants we'll explore:

GAN Variant	Key Innovation	Main Benefit
LSGAN	Least squares loss	Better gradients, less saturation
RWGAN	Relaxed Wasserstein framework	Balance between WGAN variants
McGAN	Mean/covariance matching	Statistical feature alignment
GMMN	Maximum mean discrepancy	No discriminator needed
MMD GAN	Adversarial kernels for MMD	Improved GMMN performance
Cramer GAN	Cramer distance	Unbiased sample gradients
Fisher GAN	Chi-square distance	Training stability + efficiency
EBGAN	Autoencoder discriminator	Reconstruction-based losses
BEGAN	Boundary equilibrium	WGAN + EBGAN hybrid
MAGAN	Adaptive margin	Dynamic loss boundaries

Why Objective Functions Matter

The objective function is the mathematical heart of any GAN – it defines how we measure the "distance" between our generated distribution and the real data distribution. This choice profoundly impacts:

- Training stability: Some objectives lead to more stable convergence
- Sample quality: Different losses emphasize different aspects of realism
- Mode collapse: The tendency to generate limited variety
- Computational efficiency: Some objectives are faster to compute

The original GAN uses Jensen-Shannon Divergence (JSD), but researchers have discovered many alternatives that address specific limitations. Let's explore this evolution.

LSGAN: The Power of Least Squares

Least Squares GAN takes a different approach: replace the logarithmic loss with L2 (least squares) loss.

Motivation: Beyond Binary Classification

Traditional GANs use log loss, which focuses primarily on correct classification:

- Real sample correctly classified → minimal penalty
- Fake sample correctly classified → minimal penalty
- Distance from decision boundary ignored

L2 Loss: Distance Matters

LSGAN uses L2 loss, which penalizes proportionally to distance:

Discriminator Minimization (D):

$$\min_{\mathbb{D}} V_{\text{LSGAN}}(\mathbb{D}) = \frac{1}{2}\mathbb{E}_{x \sim p_{\text{data}}(x)}[(\mathbb{D}(x) - b)^2] + \frac{1}{2}\mathbb{E}_{z \sim p_z(z)}[(\mathbb{D}(G(z)) - a)^2]$$

Generator Minimization (G):

$$\min_{G} V_{\text{LSGAN}}(G) = \frac{1}{2}\mathbb{E}_{z \sim p_z(z)}[(\mathbb{D}(G(z)) - c)^2]$$

Where typically: $a = 0$ (fake label), $b = c = 1$ (real label).

Benefits of L2 Loss:

Log Loss	L2 Loss
Binary focus	Distance-aware
Can saturate	Informative gradients
Sharp decision boundary	Smooth decision regions

Relaxed Wasserstein GAN (RWGAN)

Relaxed WGAN bridges the gap between WGAN and WGAN-GP, proposing a **general framework** for designing GAN objectives.

Key Innovations:

- **Asymmetric weight clamping:** Instead of symmetric clamping (original WGAN) or gradient penalties (WGAN-GP), RWGAN uses an asymmetric approach that provides better balance.
- **Relaxed Wasserstein divergences:** A generalized framework that extends the Wasserstein distance, enabling systematic design of new GAN variants while maintaining theoretical guarantees.

Benefits

- Better convergence properties than standard WGAN
- Framework for designing new loss functions and GAN architectures
- Competitive performance with state-of-the-art methods

Key insight: RWGAN parameterized with KL divergence shows excellent performance while maintaining the theoretical foundations that make Wasserstein GANs attractive.

Statistical Distance Approaches

Several GAN variants focus on minimizing specific statistical distances between distributions.

McGAN: Mean and Covariance Matching

McGAN belongs to the Integral Probability Metric (IPM) family, using **statistical moments** as the distance measure.

Approach: Match first and second-order statistics:

- **Mean matching:** Align distribution centers
- **Covariance matching:** Align distribution shapes

Limitation:

Relies on weight clipping like original WGAN.

GMMN: Maximum Mean Discrepancy

Generative Moment Matching Networks eliminates the discriminator entirely, directly minimizing **Maximum Mean Discrepancy (MMD)**.

MMD Intuition:

Compare distributions by their means in a high-dimensional feature space:

$$ \text{MMD}^2(X, Y) = \Vert \mathbb{E}[\phi(x)] - \mathbb{E}[\phi(y)] \Vert^2 $$

Benefits:

- Simple, discriminator-free training
- Theoretical guarantees
- Can incorporate autoencoders for better MMD estimation

Drawbacks:

- Computationally expensive
- Often weaker empirical results

MMD GAN: Learning Better Kernels

MMD GAN improves GMMN by **learning optimal kernels** adversarially rather than using fixed Gaussian kernels.

Innovation:

Combine **GAN** adversarial training with the **MMD objective** for the best of both worlds.

Different Distance Metrics

Cramer GAN: Addressing Sample Bias

Cramer GAN identifies a critical issue with WGAN: **biased sample gradients**.

The Problem:

WGAN's Wasserstein distance lacks three important properties:

1. **Sum invariance** (satisfied)

2. **Scale sensitivity** (satisfied)
3. **Unbiased sample gradients** (not satisfied)

The Solution:

Use the **Cramer distance**, which satisfies all three properties:

$$ d_C^2(\mu, \nu) = \int \Vert \mathbb{E}_{x \sim \mu}[\Vert X - x \Vert^2] - \mathbb{E}_{y \sim \nu}[\Vert Y - x \Vert^2] \Vert^2 d \pi(x) $$

Benefit:

More reliable gradients lead to better training dynamics.

Fisher GAN: Chi-Square Distance

Fisher GAN uses a **data-dependent constraint** on the critic's second-order moments (variance).

Key Innovation: The constraint naturally bounds the critic without manual techniques:

- No weight clipping needed
- No gradient penalties required
- Constraint emerges from the objective itself

Distance: Approximates the Chi-square distance as critic capacity increases:

$$ \chi^2(P, Q) = \int \frac{(P(x) - Q(x))^2}{Q(x)}\ dx $$

The Fisher GAN essentially measures the **Mahalanobis distance**, which accounts for correlated variables relative to the distribution's centroid. This ensures the generator and critic remain bounded, and as the critic's capacity increases, it estimates the Chi-square distance.

Benefits:

- Efficient computation
- Training stability
- Unconstrained critic capacity

Beyond Traditional GANs: Alternative Approaches

The following variants explore fundamentally different architectures and training paradigms.

EBGAN: Energy-Based Discrimination

Energy-Based GAN replaces the discriminator with an autoencoder.

Key insight: Use reconstruction error as the discrimination signal:

- Good data → Low reconstruction error
- Poor data → High reconstruction error

Architecture:

1. Train autoencoder on real data
2. Generator creates samples
3. Poor generated samples have high reconstruction loss
4. This loss drives generator improvement

Benefits:

- Fast and stable training
- Robust to hyperparameter changes
- No need to balance discriminator/generator

BEGAN: Boundary Equilibrium

BEGAN combines EBGAN's autoencoder approach with WGAN-style loss functions.

Innovation

- Dynamic equilibrium parameter k_t that balances:
  - Real data reconstruction quality
  - Generated data reconstruction quality

Equilibrium equation

The Discriminator loss function ($\mathbb{L}_{\mathbb{D}}$) is given by:

$$\mathbb{L}_{\mathbb{D}} = \mathbb{L}(x) - k_t\mathbb{L}(G(z))$$

Where the update for the equilibrium parameter k_t is:

$$k_{t+1} = k_t + \lambda(\gamma\mathbb{L}(x) - \mathbb{L}(G(z)))$$

MAGAN: Adaptive Margins

MAGAN improves EBGAN by making the margin in the hinge loss adaptive over time.

Concept: Start with a large margin, gradually reduce it as training progresses:

- Early training: Focus on major differences
- Later training: Fine-tune subtle details

Result: Better sample quality and training stability.