Generative Adversarial Network - Derivations & Proofs

Paul F. Roysdon, Ph.D.

Contents

1 Mathematical Derivations & Proofs

1.1 Introduction

A Generative Adversarial Network (GAN) learns a generative model by solving a two-player game between a *generator* that maps latent noise to the data space and a *discriminator* that tries to distinguish real data from generated samples. From first principles, the discriminator is trained as a probabilistic classifier (maximum likelihood under a logistic model), while the generator is trained to *fool* the discriminator. We derive the canonical minimax objective, prove the optimal discriminator in closed form, show that the induced generator objective is (up to constants) the Jensen–Shannon (JS) divergence between the real and model distributions, and present the standard (non-saturating) generator objective used for stable gradients.

1.2 Data and Notation

We observe data $\mathcal{D} = \{\mathbf{x}_i\}_{i=1}^n$ with $\mathbf{x}_i \in \mathbb{R}^{d_x}$ drawn i.i.d. from an unknown distribution having density $p_{\text{data}}(\mathbf{x})$ (or law $\mathbb{P}_{\text{data}}$). Let $\mathbf{z} \in \mathbb{R}^r$ be a latent variable with a fixed prior $p_{\mathbf{z}}$ (e.g., $\mathcal{N}(\mathbf{0}, \mathbf{I}_r)$ or $\text{Unif}([-1,1]^r)$).

- **Generator** $G_{\boldsymbol{\theta}} : \mathbb{R}^r \to \mathbb{R}^{d_x}$ with parameters $\boldsymbol{\theta}$ maps $\mathbf{z}$ to a synthetic sample $\tilde{\mathbf{x}} = G_{\boldsymbol{\theta}}(\mathbf{z})$. The *model distribution* (push-forward) is

$$p_{g_{\boldsymbol{\theta}}} \triangleq (G_{\boldsymbol{\theta}})_{\#} p_{\mathbf{z}}, \qquad \text{i.e., } \tilde{\mathbf{x}} \sim p_{g_{\boldsymbol{\theta}}} \text{ when } \mathbf{z} \sim p_{\mathbf{z}}.$$

- **Discriminator** $D_{\boldsymbol{\phi}} : \mathbb{R}^{d_x} \to (0,1)$ with parameters $\boldsymbol{\phi}$ outputs the estimated probability that an input is real (from p_{data}) rather than generated.

1.3 Model Formulation: Minimax (Logistic) Game

The discriminator is trained as a logistic classifier on a 50–50 mixture of real and fake:

$$\underbrace{\tfrac{1}{2}\, p_{\text{data}}(\mathbf{x})}_{\text{real}} \quad \text{vs.} \quad \underbrace{\tfrac{1}{2}\, p_{g_{\boldsymbol{\theta}}}(\mathbf{x})}_{\text{fake}}.$$

The standard *value function* is

$$V(D_{\boldsymbol{\phi}}, G_{\boldsymbol{\theta}}) \;=\; \mathbb{E}_{\mathbf{x}\sim p_{\text{data}}}\big[\log D_{\boldsymbol{\phi}}(\mathbf{x})\big] \;+\; \mathbb{E}_{\mathbf{z}\sim p_{\mathbf{z}}}\big[\log\big(1 - D_{\boldsymbol{\phi}}(G_{\boldsymbol{\theta}}(\mathbf{z}))\big)\big]. \tag{1}$$

Training seeks the saddle point

$$\min_{\boldsymbol{\theta}} \max_{\boldsymbol{\phi}} \; V(D_{\boldsymbol{\phi}}, G_{\boldsymbol{\theta}}). \tag{2}$$

1.4 Derivation of the Optimal Discriminator and the JS Divergence Objective

Optimal discriminator for fixed generator. Fix $G_{\boldsymbol{\theta}}$; write $p_g \equiv p_{g_{\boldsymbol{\theta}}}$. Since Eqn. (1) decomposes pointwise in $\mathbf{x}$, at any $\mathbf{x}$ the discriminator maximizes

$$f(D) \;=\; p_{\text{data}}(\mathbf{x}) \log D \;+\; p_g(\mathbf{x}) \log(1 - D), \quad D \in (0, 1).$$

Claim. The maximizer is

$$D^\star(\mathbf{x}) \;=\; \frac{p_{\text{data}}(\mathbf{x})}{p_{\text{data}}(\mathbf{x}) + p_g(\mathbf{x})}. \tag{3}$$

Proof. $f'(D) = \frac{p_{\text{data}}}{D} - \frac{p_g}{1-D} = 0$ yields $D^\star = \frac{p_{\text{data}}}{p_{\text{data}}+p_g}$. Since $f''(D) = -\frac{p_{\text{data}}}{D^2} - \frac{p_g}{(1-D)^2} < 0$, f is strictly concave and the stationary point is unique and globally maximal. ■

Generator objective as JS divergence. *Proof.* Plugging Eqn. (3) into Eqn. (1):

$$\begin{aligned}
V(D^\star, G) &= \mathbb{E}_{\mathbf{x}\sim p_{\text{data}}}\left[\log \frac{p_{\text{data}}(\mathbf{x})}{p_{\text{data}}(\mathbf{x}) + p_g(\mathbf{x})}\right] + \mathbb{E}_{\mathbf{x}\sim p_g}\left[\log \frac{p_g(\mathbf{x})}{p_{\text{data}}(\mathbf{x}) + p_g(\mathbf{x})}\right] \\
&= \int p_{\text{data}} \log \frac{p_{\text{data}}}{\frac{1}{2}(p_{\text{data}} + p_g)}\, d\mathbf{x} + \int p_g \log \frac{p_g}{\frac{1}{2}(p_{\text{data}} + p_g)}\, d\mathbf{x} - 2\log 2 \\
&= \mathrm{KL}(p_{\text{data}} \,\|\, m) + \mathrm{KL}(p_g \,\|\, m) - 2\log 2, \quad m \triangleq \tfrac{1}{2}(p_{\text{data}} + p_g) \\
&= 2\,\mathrm{JS}(p_{\text{data}} \,\|\, p_g) - \log 4.
\end{aligned} \tag{4}$$

Thus, with an optimal discriminator, minimizing the minimax objective in G is equivalent (up to a constant shift) to minimizing $\mathrm{JS}(p_{\text{data}} \| p_g)$. In particular, the global optimum is attained at $p_g = p_{\text{data}}$, for which $D^\star \equiv \frac{1}{2}$ and $V(D^\star, G) = -\log 4$. ■

Since the Jensen-Shannon divergence is always non-negative and equals zero if and only if the two distributions are identical, minimizing the value function with respect to G (i.e., the generator) pushes p_g toward p_{data}.

1.5 Gradients, Saturation, and the Non-Saturating Generator Loss

Discriminator gradient (logistic MLE). For a mini-batch $\{\mathbf{x}^{(b)}\}_{b=1}^B$ of reals and $\{\tilde{\mathbf{x}}^{(b)}\}_{b=1}^B$ of fakes, the discriminator maximizes

$$\widehat{\mathcal{L}}_D(\boldsymbol{\phi}) = \frac{1}{B}\sum_{b=1}^{B} \log D_{\boldsymbol{\phi}}(\mathbf{x}^{(b)}) + \frac{1}{B}\sum_{b=1}^{B} \log\big(1 - D_{\boldsymbol{\phi}}(\tilde{\mathbf{x}}^{(b)})\big),$$

whose gradient is the standard logistic-regression gradient:

$$\nabla_{\boldsymbol{\phi}} \widehat{\mathcal{L}}_D = \frac{1}{B}\sum_{b=1}^{B} \frac{\nabla_{\boldsymbol{\phi}} D(\mathbf{x}^{(b)})}{D(\mathbf{x}^{(b)})} - \frac{1}{B}\sum_{b=1}^{B} \frac{\nabla_{\boldsymbol{\phi}} D(\tilde{\mathbf{x}}^{(b)})}{1 - D(\tilde{\mathbf{x}}^{(b)})}.$$

Generator gradient: minimax (saturating) form. The generator minimizes

$$\widehat{\mathcal{L}}_G^{\text{minimax}}(\boldsymbol{\theta}) = \frac{1}{B}\sum_{b=1}^{B} \log\big(1 - D_{\boldsymbol{\phi}}(G_{\boldsymbol{\theta}}(\mathbf{z}^{(b)}))\big),$$

with gradient (chain rule)

$$\nabla_{\boldsymbol{\theta}} \widehat{\mathcal{L}}_G^{\text{minimax}} = \frac{1}{B}\sum_{b=1}^{B} \frac{-\nabla_{\mathbf{x}} D(\mathbf{x})|_{\mathbf{x}=G(\mathbf{z}^{(b)})}}{1 - D(G(\mathbf{z}^{(b)}))} \cdot \nabla_{\boldsymbol{\theta}} G_{\boldsymbol{\theta}}(\mathbf{z}^{(b)}).$$

When the discriminator is strong and $D(G(\mathbf{z})) \approx 0$, $\log(1 - D)$ saturates and the gradient vanishes.

Non-saturating alternative. To avoid vanishing gradients early in training, one maximizes $\log D(G(\mathbf{z}))$ (equivalently minimizes $-\log D(G(\mathbf{z}))$):

$$\widehat{\mathcal{L}}_G^{\text{NS}}(\boldsymbol{\theta}) = -\frac{1}{B}\sum_{b=1}^{B} \log D_{\boldsymbol{\phi}}(G_{\boldsymbol{\theta}}(\mathbf{z}^{(b)})), \qquad \nabla_{\boldsymbol{\theta}} \widehat{\mathcal{L}}_G^{\text{NS}} = -\frac{1}{B}\sum_{b=1}^{B} \frac{\nabla_{\mathbf{x}} D(\mathbf{x})|_{\mathbf{x}=G(\mathbf{z}^{(b)})}}{D(G(\mathbf{z}^{(b)}))} \cdot \nabla_{\boldsymbol{\theta}} G_{\boldsymbol{\theta}}(\mathbf{z}^{(b)}). \tag{5}$$

This objective has the same fixed point ($p_g = p_{\text{data}}$) but provides stronger gradients when D is confident.

1.6 Existence of a Nash Equilibrium (Sketch)

From Eqn. (4), the game Eqn. (2) admits a global saddle point at $(D^\star, G^\star)$ with $p_{g_{\boldsymbol{\theta}^\star}} = p_{\text{data}}$ and $D^\star \equiv \frac{1}{2}$, where $V(D^\star, G^\star) = -\log 4$. This follows since $\text{JS}(p_{\text{data}} \| p_g) \geq 0$ with equality iff $p_g = p_{\text{data}}$, and $D^\star$ is the unique best response for any fixed G.

1.7 Algorithm (Stochastic Minimax Training)

1. **Input:** prior $p_{\mathbf{z}}$, batch size B, learning rates η_D, η_G, discriminator $D_{\boldsymbol{\phi}}$, generator $G_{\boldsymbol{\theta}}$.
2. **Repeat** for iterations $t = 1, 2, \ldots$:
   (a) **Discriminator step(s):** Sample $\{\mathbf{x}^{(b)}\}_{b=1}^{B} \sim \mathcal{D}$ and $\{\mathbf{z}^{(b)}\}_{b=1}^{B} \sim p_{\mathbf{z}}$. Form $\tilde{\mathbf{x}}^{(b)} = G_{\boldsymbol{\theta}}(\mathbf{z}^{(b)})$. Ascend
   $$\nabla_{\boldsymbol{\phi}} \left[\frac{1}{B}\sum_{b=1}^{B} \log D_{\boldsymbol{\phi}}(\mathbf{x}^{(b)}) + \frac{1}{B}\sum_{b=1}^{B} \log\big(1 - D_{\boldsymbol{\phi}}(\tilde{\mathbf{x}}^{(b)})\big) \right].$$
   Optionally take $k > 1$ discriminator steps per generator step.
   (b) **Generator step:** Sample $\{\mathbf{z}^{(b)}\}_{b=1}^{B} \sim p_{\mathbf{z}}$, set $\tilde{\mathbf{x}}^{(b)} = G_{\boldsymbol{\theta}}(\mathbf{z}^{(b)})$, and descend either
   $$\nabla_{\boldsymbol{\theta}} \ \frac{1}{B}\sum_{b=1}^{B} \log\big(1 - D_{\boldsymbol{\phi}}(\tilde{\mathbf{x}}^{(b)})\big) \quad \text{(minimax)},$$
   or the non-saturating gradient
   $$\nabla_{\boldsymbol{\theta}} \ \Big[-\frac{1}{B}\sum_{b=1}^{B} \log D_{\boldsymbol{\phi}}(\tilde{\mathbf{x}}^{(b)}) \Big].$$
3. **Output:** generator $G_{\boldsymbol{\theta}}$ whose push-forward distribution $p_{g_{\boldsymbol{\theta}}}$ approximates p_{data}.

1.8 Regularization and Practical Stabilizers (Brief)

While not required by the basic theory, the following improve optimization:

- **Label smoothing/noise:** replace $\log D(\mathbf{x})$ with $\log D(\mathbf{x}; y{=}0.9)$; add small noise to inputs/labels.
- **Spectral/gradient penalties:** control discriminator Lipschitzness to prevent sharp decision boundaries (e.g., spectral normalization, gradient penalty).
- **Hinge loss variant:** replace logistic terms by $\max(0, 1{-}D(\mathbf{x}))$ etc. (same equilibrium; different margin geometry).

1.9 Extensions (Pointers)

f-GAN. Replace the logistic loss with a generic f-divergence; the discriminator estimates a variational lower bound to $D_f(p_{\text{data}} \| p_g)$.

Wasserstein GAN (WGAN). Optimize the Earth Mover (Wasserstein-1) distance via the Kantorovich–Rubinstein dual:

$$\min_G \max_{D \in \mathcal{L}_1} \ \mathbb{E}_{\mathbf{x}\sim p_{\text{data}}}[D(\mathbf{x})] - \mathbb{E}_{\mathbf{z}\sim p_{\mathbf{z}}}[D(G(\mathbf{z}))],$$

with $\mathcal{L}_1$ the set of 1-Lipschitz critics (enforced by weight clipping, spectral norm, or gradient penalty).

1.10 Summary of Variables and Their Dimensions

- $\mathbf{x} \in \mathbb{R}^{d_x}$: data sample; $p_{\text{data}}(\mathbf{x})$ unknown.
- $\mathbf{z} \in \mathbb{R}^r$: latent noise; prior $p_{\mathbf{z}}$ (fixed).
- $G_{\boldsymbol{\theta}} : \mathbb{R}^r \to \mathbb{R}^{d_x}$: generator; parameters $\boldsymbol{\theta}$.
- $p_{g_{\boldsymbol{\theta}}}$: generator (push-forward) distribution over $\mathbb{R}^{d_x}$.
- $D_{\boldsymbol{\phi}} : \mathbb{R}^{d_x} \to (0,1)$: discriminator; parameters $\boldsymbol{\phi}$.
- $V(D,G)$: value function Eqn. (1) (scalar).
- $D^\star(\mathbf{x}) = \frac{p_{\text{data}}(\mathbf{x})}{p_{\text{data}}(\mathbf{x}) + p_g(\mathbf{x})}$: optimal discriminator for fixed G.
- $\mathrm{JS}(p\|q)$: Jensen–Shannon divergence; $V(D^\star, G) = 2\,\mathrm{JS}(p_{\text{data}}\|p_g) - \log 4$.

1.11 Summary

Starting from a logistic classification view of real vs. generated samples, the GAN objective Eqn. (1) defines a two-player minimax game. For fixed generator, the optimal discriminator is Eqn. (3); substituting yields the generator objective Eqn. (4), which minimizes the Jensen–Shannon divergence between p_{data} and p_g and achieves its global optimum at $p_g = p_{\text{data}}$ with $D^\star \equiv \frac{1}{2}$. Stochastic gradient updates implement ascent in the discriminator parameters (logistic MLE) and descent in the generator parameters. The non-saturating generator loss Eqn. (5) preserves the equilibrium while providing stronger gradients when the discriminator is confident. Extensions such as f-GAN and WGAN change the divergence/metric while retaining the same adversarial training paradigm.