Generative Adversarial Networks

Yutao Chen

⧖ May 13 2026
⧗ May 13 2026

Contents

Generative adversarial networks (GANs) (Goodfellow et al., 2014) are generative models consisting of:

- A generator G that approximates the data distribution, and
- A discriminator D that learns to distinguish real / generated data.

Noise Contrastive Estimation

Generative adversarial networks are closely related to noise contrastive estimation (NCE) (Gutmann & Hyvärinen, 2010), a parameter estimation method for (unnormalized) probabilistic models.

Assume we have samples from an unknown data distribution $p_{\text{data}}(\boldsymbol{x})$, which we want to approximate with a probabilistic model $p_{\boldsymbol{\theta}}(\boldsymbol{x})$. Noise contrastive estimation learns $p_{\boldsymbol{\theta}}(\boldsymbol{x})$ by comparison:

> Given a noise distribution $q(\boldsymbol{x})$[1] with known density, we can learn $p_{\boldsymbol{\theta}}(\boldsymbol{x})$ by estimating the *density ratio*
>
> $$r(\boldsymbol{x}) = p_{\text{data}}(\boldsymbol{x})/q(\boldsymbol{x}).$$
>
> That is, if we know $r(\boldsymbol{x})$ and $q(\boldsymbol{x})$, we have $p_{\boldsymbol{\theta}}(\boldsymbol{x}) \approx r(\boldsymbol{x})q(\boldsymbol{x})$.

Furthermore, estimating the density ratio $r(\boldsymbol{x})$ can be casted as *binary classification*. Assuming we label data samples $\boldsymbol{x} \sim p_{\text{data}}(\boldsymbol{x})$ with $y = 1$ and noise $\boldsymbol{x} \sim q(\boldsymbol{x})$ with $y = 0$, we have

$$p_{\text{data}}(\boldsymbol{x}) = p(\boldsymbol{x}|y = 1), \qquad q(\boldsymbol{x}) = p(\boldsymbol{x}|y = 0).$$

[1] It must hold that $q(\boldsymbol{x})$ is non-zero everywhere $p_{\text{data}}(\boldsymbol{x})$ is non-zero.

By the Bayes theorem, we have $p(y|\boldsymbol{x}) = p(\boldsymbol{x}|y) \cdot p(y)/p(\boldsymbol{x})$. Therefore, the density ratio $r(\boldsymbol{x})$ can be equivalently written as

$$r(\boldsymbol{x}) = \frac{p_{\boldsymbol{\theta}}(\boldsymbol{x})}{q(\boldsymbol{x})} = \frac{p(\boldsymbol{x}|y=1)}{p(\boldsymbol{x}|y=0)} = \frac{D^*(\boldsymbol{x})}{1 - D^*(\boldsymbol{x})} \cdot \frac{1-\pi}{\pi},$$

where $D^*(\boldsymbol{x}) \coloneqq p(y=1|\boldsymbol{x})$ and $\pi \coloneqq p(y=1)$[2]. Note that $D^*(\boldsymbol{x})$ is the optimal binary classifier for $p_{\text{data}}(\boldsymbol{x})$ and $q(\boldsymbol{x})$, and

$$D^*(\boldsymbol{x}) = \frac{p_{\text{data}}(\boldsymbol{x})}{p_{\text{data}}(\boldsymbol{x}) + q(\boldsymbol{x})} = \frac{1}{1 + \exp(-\log r(\boldsymbol{x}))}.$$

Finally, we can find $D^*(\boldsymbol{x})$ by maximizing the Bernoulli log-likelihood for some parametric classifier $D_{\boldsymbol{\theta}}(\boldsymbol{x})$

$$\arg\max_{\boldsymbol{\theta}} \mathbb{E}_{p_{\text{data}}(\boldsymbol{x})}[\log D_{\boldsymbol{\theta}}(\boldsymbol{x})] + \mathbb{E}_{q(\boldsymbol{x})}[\log(1 - D_{\boldsymbol{\theta}}(\boldsymbol{x}))].$$

If we find $D_{\boldsymbol{\theta}}(\boldsymbol{x}) \approx D^*(\boldsymbol{x})$, we also find $r(\boldsymbol{x})$ and $p_{\boldsymbol{\theta}}(\boldsymbol{x}) \approx p_{\text{data}}(\boldsymbol{x})$.

Generative Adversarial Networks

NCE turns density estimation of $p_{\boldsymbol{\theta}}(\boldsymbol{x})$ into binary classification $D_{\boldsymbol{\theta}}(\boldsymbol{x})$. The noise distribution $q(\boldsymbol{x})$ is often chosen to be simplistic and tractable (e.g. Gaussians). In constrast, *generative adversarial networks* learn the noise distribution $q_{\boldsymbol{\phi}}(\boldsymbol{x})$ jointly with the binary classifier $D_{\boldsymbol{\theta}}(\boldsymbol{x})$:

$$\min_{\boldsymbol{\phi}} \max_{\boldsymbol{\theta}} \mathbb{E}_{p_{\text{data}}(\boldsymbol{x})}[\log D_{\boldsymbol{\theta}}(\boldsymbol{x})] + \mathbb{E}_{q_{\boldsymbol{\phi}}(\boldsymbol{x})}[\log(1 - D_{\boldsymbol{\theta}}(\boldsymbol{x}))]$$

With the above minimax optimization framework, a unique equilibrium exists at $q_{\boldsymbol{\phi}}(\boldsymbol{x}) = p_{\text{data}}(\boldsymbol{x})$ and $D_{\boldsymbol{\theta}}(\boldsymbol{x}) = \frac{1}{2}$ for all $\boldsymbol{x}$.

Proof. Note that for any fixed $\boldsymbol{\phi}$, the optimal $\boldsymbol{\theta}$ corresponds to

$$D_{\boldsymbol{\theta}}(\boldsymbol{x}) = \frac{p_{\text{data}}(\boldsymbol{x})}{p_{\text{data}}(\boldsymbol{x}) + q_{\boldsymbol{\phi}}(\boldsymbol{x})}.$$

Furthermore, for any optimal $\boldsymbol{\theta}$, the optimal $\boldsymbol{\phi}$ minimizes the following objective

[2] We assume without loss of generality $\pi = \frac{1}{2}$ hereafter.

$$
\begin{aligned}
& \mathbb{E}_{p_{\text{data}}(\boldsymbol{x})}\left[\log \frac{p_{\text{data}}(\boldsymbol{x})}{p_{\text{data}}(\boldsymbol{x})+q_{\boldsymbol{\phi}}(\boldsymbol{x})}\right]+\mathbb{E}_{q_{\boldsymbol{\phi}}(\boldsymbol{x})}\left[\log \frac{q_{\boldsymbol{\phi}}(\boldsymbol{x})}{p_{\text{data}}(\boldsymbol{x})+q_{\boldsymbol{\phi}}(\boldsymbol{x})}\right] \\
= & \quad \mathbb{D}_{\mathrm{KL}}\left(p_{\text{data}} \,\middle\|\, \frac{p_{\text{data}}+q_{\boldsymbol{\phi}}}{2}\right)+\mathbb{D}_{\mathrm{KL}}\left(q_{\boldsymbol{\phi}} \,\middle\|\, \frac{p_{\text{data}}+q_{\boldsymbol{\phi}}}{2}\right)-2 \log 2 \\
= & \quad 2 \text{ JSD }\left(p_{\text{data}}(\boldsymbol{x}) \,\|\, q_{\boldsymbol{\phi}}(\boldsymbol{x})\right)-\log 4,
\end{aligned}
$$

where $\mathrm{JSD}(p \parallel q)$ denotes the Jensen-Shannon divergence, minimized when $p=q$. Therefore, the optimal $\boldsymbol{\phi}$ for any optimal $\boldsymbol{\theta}$ is given by $p_{\text{data}}(\boldsymbol{x})=q_{\boldsymbol{\phi}}(\boldsymbol{x})$.

The noise distribution $q_{\boldsymbol{\phi}}(\boldsymbol{x})$ is often parametrized as a neural network $G_{\boldsymbol{\phi}}(\boldsymbol{z})$, where $\boldsymbol{z} \sim p(\boldsymbol{z})$ and $p(\boldsymbol{z})$ is a simple prior noise distribution.

We refer the readers to Murphy (2023) Section 26.2 for further analysis on connecting GAN, *proper scoring rule*, and statistical divergence.

References

Goodfellow, I. J., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., & Bengio, Y. (2014). Generative Adversarial Nets. In Z. Ghahramani, M. Welling, C. Cortes, N. Lawrence, & K. Weinberger (Eds.), *Advances in Neural Information Processing Systems: Vol. 27. Advances in Neural Information Processing Systems*.

Gutmann, M., & Hyvärinen, A. (2010). Noise-contrastive estimation: A new estimation principle for unnormalized statistical models. In Y. W. Teh & M. Titterington (Eds.), *Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics: Vol. 9. Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics*. https://proceedings.mlr.press/v9/gutmann10a.html

Murphy, K. P. (2023). *Probabilistic Machine Learning: Advanced Topics*. MIT Press. http://probml.github.io/book2