11.1. Generative Adversarial Networks

François Fleuret
https://fleuret.org/dlc/
Dec 20, 2020
A popular approach to learn high-dimension densities are the **Generative Adversarial Networks** proposed by Goodfellow et al. (2014), where two networks are trained jointly:

- A **discriminator** $D$ to classify samples as “real” or “fake”,
- a **generator** $G$ to map a [simple] fixed distribution to samples that fool $D$. 

The approach is adversarial since the two networks have antagonistic objectives.
A popular approach to learn high-dimension densities are the **Generative Adversarial Networks** proposed by Goodfellow et al. (2014), where two networks are trained jointly:

- A **discriminator** $D$ to classify samples as “real” or “fake”,
- a **generator** $G$ to map a [simple] fixed distribution to samples that fool $D$.

![Diagram showing the process of a sample being classified by $D$ as "real".]

The approach is adversarial since the two networks have antagonistic objectives.
A popular approach to learn high-dimension densities are the **Generative Adversarial Networks** proposed by Goodfellow et al. (2014), where two networks are trained jointly:

- A **discriminator** $D$ to classify samples as “real” or “fake”,
- a **generator** $G$ to map a [simple] fixed distribution to samples that fool $D$.

![Diagram of Generative Adversarial Networks](diagram.png)
A popular approach to learn high-dimension densities are the **Generative Adversarial Networks** proposed by Goodfellow et al. (2014), where two networks are trained jointly:

- A **discriminator** $D$ to classify samples as “real” or “fake”,
- a **generator** $G$ to map a [simple] fixed distribution to samples that fool $D$.

The approach is **adversarial** since the two networks have antagonistic objectives.
Let $\mathcal{X}$ be the signal space, and $D$ the latent space dimension.

- The generator $G : \mathbb{R}^D \rightarrow \mathcal{X}$ is trained so that ideally if it gets a random normal-distributed $Z$ as input, it produces a sample following the data distribution as output.
Let $\mathcal{X}$ be the signal space, and $D$ the latent space dimension.

- The generator
  \[ G : \mathbb{R}^D \to \mathcal{X} \]
  is trained so that [ideally] if it gets a random normal-distributed $Z$ as input, it produces a sample following the data distribution as output.

- The discriminator
  \[ D : \mathcal{X} \to [0, 1] \]
  is trained so that if it gets a sample as input, it predicts if it comes from $G$ or from the real data.
If $G$ is fixed, to train $D$ given a set of “real points”

$$x_n \sim \mu, \quad n = 1, \ldots, N,$$

we can generate

$$z_n \sim \mathcal{N}(0, I), \quad n = 1, \ldots, N,$$

build a two-class data-set

$$\mathcal{D} = \left\{ (x_1, 1), \ldots, (x_N, 1), (G(z_1), 0), \ldots, (G(z_N), 0) \right\},$$

where $\mu$ is the true data distribution, and $\mu_G$ is the distribution of $G(Z)$ with $Z \sim \mathcal{N}(0, I)$. 

And minimize the binary cross-entropy

$$\mathcal{L}(D) = -\frac{1}{2}N \sum_{n=1}^{N} \log D(x_n) + \sum_{n=1}^{N} \log(1 - D(G(z_n), 0)).$$

This is equivalent to

$$\mathcal{L}(D) = -\frac{1}{2} \left( \hat{E}_{X \sim \mu}[\log D(X)] + \hat{E}_{X \sim \mu_G}[\log(1 - D(X))] \right).$$
If $G$ is fixed, to train $D$ given a set of “real points”

$$x_n \sim \mu, \; n = 1, \ldots, N,$$

we can generate

$$z_n \sim \mathcal{N}(0, I), \; n = 1, \ldots, N,$$

build a two-class data-set

$$\mathcal{D} = \left\{ (x_1, 1), \ldots, (x_N, 1), (G(z_1), 0), \ldots, (G(z_N), 0) \right\},$$

where $\mu$ is the true data distribution, and $\mu_G$ is the distribution of $G(Z)$ with $Z \sim \mathcal{N}(0, I)$

And minimize the binary cross-entropy

$$\mathcal{L}(D) = -\frac{1}{2N} \left( \sum_{n=1}^{N} \log D(x_n) + \sum_{n=1}^{N} \log(1 - D(G(z_n))) \right)$$

$$= -\frac{1}{2} \left( \hat{E}_{X \sim \mu} \left[ \log D(X) \right] + \hat{E}_{X \sim \mu_G} \left[ \log(1 - D(X)) \right] \right).$$
The situation is slightly more complicated since we also want to optimize $G$ to maximize $D$'s loss.
The situation is slightly more complicated since we also want to optimize $G$ to maximize $D$'s loss.

Goodfellow et al. (2014) provide an analysis of the resulting equilibrium of that strategy.
Let’s define the loss of $G$

$$\mathcal{L}_G(D, G) = \mathbb{E}_{X \sim \mu} \left[ \log D(X) \right] + \mathbb{E}_{X \sim \mu_G} \left[ \log(1 - D(X)) \right]$$

which is high if $D$ is doing a good job (low cross entropy), and low if $G$ fools $D$. 
Let’s define the loss of $G$

$$\mathcal{L}_G(D, G) = \mathbb{E}_{X \sim \mu} \left[ \log D(X) \right] + \mathbb{E}_{X \sim \mu_G} \left[ \log(1 - D(X)) \right]$$

which is high if $D$ is doing a good job (low cross entropy), and low if $G$ fools $D$.

Our ultimate goal is a $G^*$ that fools any $D$, so

$$G^* = \arg\min_G \max_D \mathcal{L}_G(D, G).$$
If we define the optimal discriminator for a given generator

$$D_G^* = \arg\max_D \mathcal{L}_G(D, G),$$

our objective becomes

$$G^* = \arg\min_G \mathcal{L}_G(D_G^*, G),$$

that is:

**Find a $G$ whose loss against its best adversary $D_G^*$ is low.**
We have

\[ L_G(D, G) = \mathbb{E}_{X \sim \mu} \left[ \log D(X) \right] + \mathbb{E}_{X \sim \mu_G} \left[ \log(1 - D(X)) \right] \]

\[ = \int x \mu(x) \log D(x) + \mu_G(x) \log(1 - D(x)) dx. \]
We have

\[
\mathcal{L}_G(D, G) = \mathbb{E}_{X \sim \mu} \left[ \log D(X) \right] + \mathbb{E}_{X \sim \mu_G} \left[ \log(1 - D(X)) \right]
\]

\[
= \int \mu(x) \log D(x) + \mu_G(x) \log(1 - D(x)) \, dx.
\]

Since

\[
\arg\max_d \mu(x) \log d + \mu_G(x) \log(1 - d) = \frac{\mu(x)}{\mu(x) + \mu_G(x)},
\]

and

\[
D^*_G = \arg\max_D \mathcal{L}_G(D, G),
\]

if there is no regularization on \( D \), we get

\[
\forall x, \quad D^*_G(x) = \frac{\mu(x)}{\mu(x) + \mu_G(x)}.
\]
So, since

$$\forall x, \ D^*_G(x) = \frac{\mu(x)}{\mu(x) + \mu_G(x)}.$$ 

we get

$$\mathcal{L}_G(D^*_G, G) = \mathbb{E}_{X \sim \mu} \left[ \log D^*_G(X) \right] + \mathbb{E}_{X \sim \mu_G} \left[ \log(1 - D^*_G(X)) \right]$$

$$= \mathbb{E}_{X \sim \mu} \left[ \log \frac{\mu(X)}{\mu(X) + \mu_G(X)} \right] + \mathbb{E}_{X \sim \mu_G} \left[ \log \frac{\mu_G(X)}{\mu(X) + \mu_G(X)} \right]$$

$$= \mathcal{D}_{KL} \left( \mu \left\| \frac{\mu + \mu_G}{2} \right\| \right) + \mathcal{D}_{KL} \left( \mu_G \left\| \frac{\mu + \mu_G}{2} \right\| \right) - \log 4$$
So, since

$$\forall x, \; D_G^*(x) = \frac{\mu(x)}{\mu(x) + \mu_G(x)}.$$ 

we get

$$\mathcal{L}_G(D_G^*, G) = \mathbb{E}_{X \sim \mu} \left[ \log D_G^*(X) \right] + \mathbb{E}_{X \sim \mu_G} \left[ \log(1 - D_G^*(X)) \right]$$

$$= \mathbb{E}_{X \sim \mu} \left[ \log \frac{\mu(X)}{\mu(X) + \mu_G(X)} \right] + \mathbb{E}_{X \sim \mu_G} \left[ \log \frac{\mu_G(X)}{\mu(X) + \mu_G(X)} \right]$$

$$= \mathbb{D}_{KL} \left( \mu \mid \mid \frac{\mu + \mu_G}{2} \right) + \mathbb{D}_{KL} \left( \mu_G \mid \mid \frac{\mu + \mu_G}{2} \right) - \log 4$$

$$= 2 \mathbb{D}_{JS} (\mu, \mu_G) - \log 4$$

where $\mathbb{D}_{JS}$ is the Jensen-Shannon Divergence, a standard similarity measure between distributions.
To recap: if there is no capacity limitation for $D$, and if we define

$$
\mathcal{L}_G(D, G) = \mathbb{E}_{X \sim \mu} \left[ \log D(X) \right] + \mathbb{E}_{X \sim \mu_G} \left[ \log(1 - D(X)) \right],
$$

computing

$$
G^* = \arg \min_G \max_D \mathcal{L}_G(D, G)
$$

clearly amounts to compute

$$
G^* = \arg \min_G \mathbb{D}_{JS}(\mu, \mu_G),
$$

where $\mathbb{D}_{JS}$ is a reasonable similarity measure between distributions.
To recap: if there is no capacity limitation for $D$, and if we define

$$
\mathcal{L}_G(D, G) = \mathbb{E}_{X \sim \mu} \left[ \log D(X) \right] + \mathbb{E}_{X \sim \mu G} \left[ \log(1 - D(X)) \right],
$$

computing

$$
G^{\ast} = \arg\min_G \max_D \mathcal{L}_G(D, G)
$$

amounts to compute

$$
G^{\ast} = \arg\min_G D_{JS}(\mu, \mu_G),
$$

where $D_{JS}$ is a reasonable similarity measure between distributions.

Although this derivation provides a nice formal framework, in practice $D$ is not “fully” optimized to [come close to] $D_G^{\ast}$ when optimizing $G$.

In our minimal example, we alternate gradient steps to improve $G$ and $D$. 

For our toy example, we take $D = 8$, and $\mathcal{X} = \mathbb{R}^2$.

$z_{\text{dim}} = 8$
$\text{nb}_{\text{hidden}} = 100$

```python
model_G = nn.Sequential(nn.Linear(z_dim, nb_hidden),
                        nn.ReLU(),
                        nn.Linear(nb_hidden, 2))

model_D = nn.Sequential(nn.Linear(2, nb_hidden),
                        nn.ReLU(),
                        nn.Linear(nb_hidden, 1),
                        nn.Sigmoid())
```
batch_size, lr = 10, 1e-3

optimizer_G = optim.Adam(model_G.parameters(), lr = lr)
optimizer_D = optim.Adam(model_D.parameters(), lr = lr)

for e in range(nb_epochs):
    for t, real_batch in enumerate(real_samples.split(batch_size)):
        z = real_batch.new(real_batch.size(0), z_dim).normal_()
        fake_batch = model_G(z)

        D_scores_on_real = model_D(real_batch)
        D_scores_on_fake = model_D(fake_batch)

        if t%2 == 0:
            loss = (1 - D_scores_on_fake).log().mean()
            optimizer_G.zero_grad()
            loss.backward()
            optimizer_G.step()
        else:
            loss = - (1 - D_scores_on_fake).log().mean() - D_scores_on_real.log().mean()
            optimizer_D.zero_grad()
            loss.backward()
            optimizer_D.step()
$D = 2$

The image shows a 2D scatter plot with two clusters labeled 'Real' and 'Fake'. The plot appears to be generated by a generative adversarial network (GAN) model.
$D = 8$
$D = 32$

The diagram shows a scatter plot with two categories: Real and Fake. The points are distributed across a 2D space, with the x-axis ranging from -6 to 6 and the y-axis from -4 to 4. The points are clustered together, indicating a distribution of data points.
$D = 2$
$D = 8$
$D = 32$

![Graph showing real and fake data distribution with $D = 32$.](image)

- **Real** data is represented by blue dots.
- **Fake** data is represented by red dots.
\( D = 2 \)

- **Real**
- **Fake**
$D = 8$
$D = 32$
\( D = 2 \)
\[ D = 8 \]

![Graph showing real and fake data distribution](image)
\[ D = 32 \]
$D = 2$

![Graph showing real and fake data points with $D = 2$.](image)
\( D = 8 \)

![Graph showing real and fake data points in a 2D space with axes labeled from -6 to 6, and values from -4 to 4. The data points are differentiated by color: blue for real and red for fake.](image-url)
$D = 32$

Graph showing the distribution of real and fake data points in a 2D space.
In more realistic settings, the fake samples may be initially so “unrealistic” that the response of \( D \) saturates. That causes the loss for \( G \)

\[
\hat{G}_{X \sim \mu_G} \left[ \log(1 - D(X)) \right]
\]

to be far in the exponential tail of \( D \)’s sigmoid, and have zero gradient since \( \log(1 + \epsilon) \approx \epsilon \) does not correct it in any way.
In more realistic settings, the fake samples may be initially so “unrealistic” that the response of $D$ saturates. That causes the loss for $G$

$$\hat{\mathbb{E}}_{X \sim \mu_G} \left[ \log(1 - D(X)) \right]$$

to be far in the exponential tail of $D$’s sigmoid, and have zero gradient since $\log(1 + \epsilon) \simeq \epsilon$ does not correct it in any way.

Goodfellow et al. suggest to replace this term with a non-saturating cost

$$-\hat{\mathbb{E}}_{X \sim \mu_G} \left[ \log(D(X)) \right]$$

so that the log fixes $D$’s exponential behavior. The resulting optimization problem has the same optima as the original one.
In more realistic settings, the fake samples may be initially so “unrealistic” that the response of \( D \) saturates. That causes the loss for \( G \)

\[
\hat{\mathbb{E}}_{X \sim \mu_G} \left[ \log(1 - D(X)) \right]
\]

to be far in the exponential tail of \( D \)’s sigmoid, and have zero gradient since \( \log(1 + \epsilon) \approx \epsilon \) does not correct it in any way.

Goodfellow et al. suggest to replace this term with a non-saturating cost

\[
-\hat{\mathbb{E}}_{X \sim \mu_G} \left[ \log(D(X)) \right]
\]

so that the \( \log \) fixes \( D \)’s exponential behavior. The resulting optimization problem has the same optima as the original one.

⚠️ The loss for \( D \) remains unchanged.
Table 1: Parzen window-based log-likelihood estimates. The reported numbers on MNIST are the mean log-likelihood of samples on test set, with the standard error of the mean computed across examples. On TFD, we computed the standard error across folds of the dataset, with a different $\sigma$ chosen using the validation set of each fold. On TFD, $\sigma$ was cross validated on each fold and mean log-likelihood on each fold were computed. For MNIST we compare against other models of the real-valued (rather than binary) version of dataset.

Results are reported in Table 1. This method of estimating the likelihood has somewhat high variance and does not perform well in high dimensional spaces but it is the best method available to our knowledge. Advances in generative models that can sample but not estimate likelihood directly motivate further research into how to evaluate such models.

In Figures 2 and 3 we show samples drawn from the generator net after training. While we make no claim that these samples are better than samples generated by existing methods, we believe that these samples are at least competitive with the better generative models in the literature and highlight the potential of the adversarial framework.

Figure 2: Visualization of samples from the model. Rightmost column shows the nearest training example of the neighboring sample, in order to demonstrate that the model has not memorized the training set. Samples are fair random draws, not cherry-picked. Unlike most other visualizations of deep generative models, these images show actual samples from the model distributions, not conditional means given samples of hidden units. Moreover, these samples are uncorrelated because the sampling process does not depend on Markov chain mixing. a) MNIST b) TFD c) CIFAR-10 (fully connected model) d) CIFAR-10 (convolutional discriminator and “deconvolutional” generator)

(Goodfellow et al., 2014)
Deep Convolutional GAN
“We also encountered difficulties attempting to scale GANs using CNN architectures commonly used in the supervised literature. However, after extensive model exploration we identified a family of architectures that resulted in stable training across a range of datasets and allowed for training higher resolution and deeper generative models.”

(Radford et al., 2015)
Radford et al. converged to the following rules:

- Replace pooling layers with strided convolutions in $D$ and strided transposed convolutions in $G$,
- use batchnorm in both $D$ and $G$,
- remove fully connected hidden layers,
- use ReLU in $G$ except for the output, which uses Tanh,
- use LeakyReLU activation in $D$ for all layers.
Figure 1: DCGAN generator used for LSUN scene modeling. A 100 dimensional uniform distribution $Z$ is projected to a small spatial extent convolutional representation with many feature maps. A series of four fractionally-strided convolutions (in some recent papers, these are wrongly called deconvolutions) then convert this high level representation into a $64 \times 64$ pixel image. Notably, no fully connected or pooling layers are used.

(Radford et al., 2015)
Figure 1: DCGAN generator used for LSUN scene modeling. A 100 dimensional uniform distribution \( Z \) is projected to a small spatial extent convolutional representation with many feature maps. A series of four fractionally-strided convolutions (in some recent papers, these are wrongly called deconvolutions) then convert this high level representation into a \( 64 \times 64 \) pixel image. Notably, no fully connected or pooling layers are used.

(Radford et al., 2015)

We can have a look at the reference implementation provided in

https://github.com/pytorch/examples.git
# default nz = 100, ngf = 64

class Generator(nn.Module):
    def __init__(self, ngpu):
        super().__init__()
        self.ngpu = ngpu
        self.main = nn.Sequential(
            # input is Z, going into a convolution
            nn.ConvTranspose2d(nz, ngf * 8, 4, 1, 0, bias=False),
            nn.BatchNorm2d(ngf * 8),
            nn.ReLU(True),  # state size. (ngf*8) x 4 x 4
            nn.ConvTranspose2d(ngf * 8, ngf * 4, 4, 2, 1, bias=False),
            nn.BatchNorm2d(ngf * 4),
            nn.ReLU(True),  # state size. (ngf*4) x 8 x 8
            nn.ConvTranspose2d(ngf * 4, ngf * 2, 4, 2, 1, bias=False),
            nn.BatchNorm2d(ngf * 2),
            nn.ReLU(True),  # state size. (ngf*2) x 16 x 16
            nn.ConvTranspose2d(ngf * 2, ngf, 4, 2, 1, bias=False),
            nn.BatchNorm2d(ngf),
            nn.ReLU(True),  # state size. (ngf) x 32 x 32
            nn.ConvTranspose2d(nz, nc, 4, 2, 1, bias=False),
            nn.Tanh()
            # state size. (nc) x 64 x 64
        )
# default nz = 100, ndf = 64

class Discriminator(nn.Module):
    def __init__(self, ngpu):
        super().__init__()
        self.ngpu = ngpu
        self.main = nn.Sequential(
            # input is (nc) x 64 x 64
            nn.Conv2d(nc, ndf, 4, 2, 1, bias=False),
            nn.LeakyReLU(0.2, inplace=True),
            # state size. (ndf) x 32 x 32
            nn.Conv2d(ndf, ndf * 2, 4, 2, 1, bias=False),
            nn.BatchNorm2d(ndf * 2),
            nn.LeakyReLU(0.2, inplace=True),
            # state size. (ndf*2) x 16 x 16
            nn.Conv2d(ndf * 2, ndf * 4, 4, 2, 1, bias=False),
            nn.BatchNorm2d(ndf * 4),
            nn.LeakyReLU(0.2, inplace=True),
            # state size. (ndf*4) x 8 x 8
            nn.Conv2d(ndf * 4, ndf * 8, 4, 2, 1, bias=False),
            nn.BatchNorm2d(ndf * 8),
            nn.LeakyReLU(0.2, inplace=True),
            # state size. (ndf*8) x 4 x 4
            nn.Conv2d(ndf * 8, 1, 4, 1, 0, bias=False),
            nn.Sigmoid()
        )
# custom weights initialization called on netG and netD

def weights_init(m):
    classname = m.__class__.__name__
    if classname.find('Conv') != -1:
        m.weight.data.normal_(0.0, 0.02)
    elif classname.find('BatchNorm') != -1:
        m.weight.data.normal_(1.0, 0.02)
    m.bias.data.fill_(0)

criterion = nn.BCELoss()
fixed_noise = torch.randn(opt.batchSize, nz, 1, 1, device=device)
real_label = 1
fake_label = 0

# setup optimizer
optimizerD = optim.Adam(netD.parameters(), lr=opt.lr, betas=(opt.beta1, 0.999))
optimizerG = optim.Adam(netG.parameters(), lr=opt.lr, betas=(opt.beta1, 0.999))
def weights_init(m):
    classname = m.__class__.__name__
    if classname.find('Conv') != -1:
        m.weight.data.normal_(0.0, 0.02)
    elif classname.find('BatchNorm') != -1:
        m.weight.data.normal_(1.0, 0.02)
    m.bias.data.fill_(0)

criterion = nn.BCELoss()

fixed_noise = torch.randn(opt.batchSize, nz, 1, 1, device=device)
real_label = 1
fake_label = 0

# setup optimizer
optimizerD = optim.Adam(netD.parameters(), lr=opt.lr, betas=(opt.beta1, 0.999))
optimizerG = optim.Adam(netG.parameters(), lr=opt.lr, betas=(opt.beta1, 0.999))
# (1) Update D network: maximize log(D(x)) + log(1 - D(G(z)))

# train with real
netD.zero_grad()
real_cpu = data[0].to(device)
batch_size = real_cpu.size(0)
label = torch.full((batch_size,), real_label, device=device)
output = netD(real_cpu)
errD_real = criterion(output, label)
errD_real.backward()
D_x = output.mean().item()

# train with fake
noise = torch.randn(batch_size, nz, 1, 1, device=device)
fake = netG(noise)
label.fill_(fake_label)
output = netD(fake.detach())
errD_fake = criterion(output, label)
errD_fake.backward()
D_G_z1 = output.mean().item()
errD = errD_real + errD_fake
optimizerD.step()
# (2) Update G network: maximize log(D(G(z)))

```python
netG.zero_grad()
label.fill_(real_label)  # fake labels are real for generator cost
output = netD(fake)
errG = criterion(output, label)
errG.backward()
D_G_z2 = output.mean().item()
optimizerG.step()
```

Note that this update implements the \(-\log(D(G(z)))\) trick.
Real images from LSUN's “bedroom” class.
Fake images after 1 epoch (3M images)
Fake images after 2 epochs
Fake images after 5 epochs
Fake images after 10 epochs
Fake images after 20 epochs
Training a standard GAN often results in two pathological behaviors:

- Oscillations without convergence. Contrary to standard loss minimization, we have no guarantee here that it will actually decrease.

- The infamous “mode collapse”, when $G$ models very well a small sub-population, concentrating on a few modes.
Training a standard GAN often results in two pathological behaviors:

- Oscillations without convergence. Contrary to standard loss minimization, we have no guarantee here that it will actually decrease.

- The infamous “mode collapse”, when $G$ models very well a small sub-population, concentrating on a few modes.

Additionally, performance is hard to assess. Two standard measures are the Inception Score (Salimans et al., 2016) and the Fréchet Inception Distance (Heusel et al., 2017), but assessment is often a “beauty contest”.

(Brock et al., 2018)
(Brock et al., 2018)
(Karras et al., 2018)
The end


