Learn distributed GAN with Temporary Discriminators - Appendix

1 Training algorithm of TDGAN

The training algorithm and communication details between the central generator and distributed discriminators of TDGAN are shown in Algorithm 1.

2 Loss function of TDGAN

$$\begin{aligned}
V_t(G_t, D_t^{1:K_t}) &= \min_{G_t} \; Digesting \;\; Loss + \lambda \cdot Reminding \;\; Loss \\
Digesting \;\; Loss &\triangleq \max_{D_t^{1:K_t}} \sum_{k=1}^{K_t} \pi_t^k \mathbb{E}_{y \sim g_t^k(y)} \{ \mathbb{E}_{x \sim p(x|y)} [\log D_t^k(x, y)] \\
&\quad + \mathbb{E}_{u \sim unif(0,1)} [\log(1 - D_t^k(G_t(u, y), y))] \} \\
Reminding \;\; Loss &\triangleq \mathbb{E}_{y \sim s_{t-1}(y)} \mathbb{E}_{u \sim unif(0,1)} [\| G_t(u, y) - G_{t-1}(u, y) \|^2]
\end{aligned} \tag{1}$$

3 Missing Proof in Analysis Section

Lemma 1 (Reminding Loss enforces consistency). *Suppose G_t has enough model capacity, the optimal G_t for loss function:*

$$\min_{G_t} \mathbb{E}_{y \sim s_{t-1}(y)} \mathbb{E}_{u \sim unif(0,1)} [\| G_t(u, y) - G_{t-1}(u, y) \|^2]$$

given G_{t-1} is $G_t(u, y) = G_{t-1}$ for all u and $y \in \Omega(s_{t-1}(y))$.

Proof:

$$\begin{aligned}
&\min_{G_t} \mathbb{E}_{y \sim s_{t-1}(y)} \mathbb{E}_{u \sim unif(0,1)} [\| G_t(u, y) - G_{t-1}(u, y) \|^2] \\
= &\min_{G_t} \int_y s_{t-1}(y) \int_u \| G_t(u, y) - G_{t-1}(u, y) \|^2 du dy \\
\geq &\int_y s_{t-1}(y) \int_u \min_{G_t} \| G_t(u, y) - G_{t-1}(u, y) \|^2 du dy
\end{aligned}$$

When $G_t(u, y) = G_{t-1}(u, y)$ for all u, y the inequality becomes equality. $\square$

Algorithm 1 Training algorithm of TDGAN at time step t.

Initialized G_t with G_{t-1} if $t > 1$.
for number of total training iterations **do**
  // Update online Discriminators
  for each online node $k \in [K_t]$ **do**
    – Sample minibatch of of m variables $\{y_1^k, ..., y_m^k\}$ from $g_t^k(y)$.
    – Send the minibatch from D_t^k to G_t.
    – Generate m fake data from G_t, $\{\hat{x}_1^k, ..., \hat{x}_m^k\} \sim q_t(\hat{x}|y)$.
    – Send the fake data from G_t to D_t^k.
    – Update the discriminator D_t^k by ascending its stochastic gradient:

$$\nabla_{\theta_{D_t^k}} \frac{1}{m} \sum_{i=1}^{m} \left[\log D_t^k(x_i^k) + \log(1 - D_t^k(\hat{x}_i^k)) \right].$$

  end for
  // Compute the gradients of G_t using the digesting loss
  for each online node $k \in [K_t]$ **do**
    – Sample minibatch of m variables $\{y_1^k, ..., y_m^k\}$ from $g_t^k(y)$.
    – Send the minibatch from D_t^k to G_t.
    – Generate m fake data from G_t, $\{\hat{x}_1^k, ..., \hat{x}_m^k\} \sim q_t(\hat{x}|y)$.
    – Send the fake data from G_t to D_t^k.
    – Collect error from D_t^k for G_t.
  end for
  – Compute gradients on the digesting loss:

$$\nabla_{\theta_{G_t}} \frac{1}{m} \sum_{k=1}^{K_t} \pi_t^k \sum_{i=1}^{m} \log(1 - D_t^k(\hat{x}_i^k)).$$

  // Compute the gradients using the reminding loss if $t > 1$
  if $t > 1$ **then**
    – Sample minibatch of n variables $\{y_1, ..., y_n\}$ from $s_{t-1}(y)$. (We approximate $s_{t-1}(y)$ by storing the empirical distribution in central server)
    – Generate n copies of u from $unif(0,1)$: $\{u_1, ..., u_n\}$.
    – Compute gradients on the reminding loss:

$$\nabla_{\theta_{G_t}} \frac{1}{n} \sum_{i=1}^{n} \|G_t(u_i, y_i) - G_{t-1}(u_i, y_i)\|^2.$$

  end if
  Update G_t using gradients from both losses.
end for

Lemma 2 (Digesting Loss Learns correct distribution). *Suppose discriminator $D_t^k, k \in [K_t]$ always behave optimally and let $q_t(x|y)$ be the distribution of $G_t(u,y)$, the the optimal $G_t(u,y)$ for digesting loss:*

$$\min_{G_t} \max_{D_t^{1:K_t}} \sum_{k=1}^{K_t} \pi_t^k \mathbb{E}_{y \sim g_t^k(y)} \{ \mathbb{E}_{x \sim p(x|y)} [\log D_t^k(x, y)] + \mathbb{E}_{u \sim unif(0,1)} [\log(1 - D_t^k(G_t(u, y), y))] \}$$

is $q_t(x|y) = p(x|y)$ *for all* $y \in \Omega(g_t(y))$.

Proof:

Similar to [1], we first analyze the behavior of optimal discriminators w.r.t a fixed generator.

$$\max_{D_t^{1:K_t}} Loss(D_t) = \max_{D_t^{1:K_t}} \sum_{k=1}^{K_t} \pi_t^k \int_y g_t^k(y) \int_x p(x|y) log D_t^k(y,x)$$
$$+ q_t(y|x) log(1 - D_t^k(y,x)) dxdy$$
$$\leq \sum_{k=1}^{K_t} \pi_t^k \int_y g_t^k(y) \int_x \max_{D_t} \{p(x|y) log D_t^k(x,y) + q(y|x) log(1 - D_t^k(x,y))\} dxdy$$

by setting $D_t^k(y,x) = \frac{p(x|y)}{p(x|y)+q_t(x|y)}$ for all $y \in \Omega(g_t^k(y))$ we can make the inequality hold with equality. Given a consistent optimal discriminator in each step of optimization process, the loss function of generator becomes:

$$Loss(G_t) = \sum_{k=1}^{K_t} \pi_t^k \mathbb{E}_{y \sim g_t^k(y)} \{\mathbb{E}_{x \sim p(x|y)[log D_t^k(x,y)]} + \mathbb{E}_{\hat{x} \sim q(x|y)}[log(1 - D_t^k(x,y))]\}$$
$$\Longleftrightarrow$$
$$Loss(q_t, \gamma) = \sum_{k=1}^{K_t} \pi_t^k \int_y g_t(y) \int_x p(x|y) log \frac{p(x|y)}{p(x|y) + q_t(x|y)}$$
$$+ q_t(x|y) log \frac{q_t(x|y)}{p(x|y) + q_t(x|y)} dx + \int_x q_t(x|y) - 1 dxdy$$

where γ is Lagrangian Multiplier for constraint $\int_x q_t(x|y)dx = 1$. We have:

$$Loss(q_t, \gamma) \underset{*}{\geq} \int_y g_t(y) \int_x \min_{q_t} p(x|y) log \frac{p(x|y)}{p(x|y) + q_t(x|y)}$$
$$+ q_t(x|y) log \frac{q_t(x|y)}{p(x|y) + q_t(x|y)} + \gamma q_t(x|y) - \gamma \; dxdy$$

Minimizing $p(x|y) log \frac{p(x|y)}{p(x|y)+q_t(x|y)} + q_t(x|y) log \frac{q_t(x|y)}{p(x|y)+q_t(x|y)} + \gamma q_t(x|y)$ requires $\frac{p(x|y)}{p(x|y)+q_t(x|y)}$ to be constant for all possible value of x and y. Such constraint enforces $p(x|y) = q_t(x|y)$ and $\gamma = -\log 2$, which makes inequality $*$ holds with equality. $\square$

Above two lemmas describes the behavior of digesting loss and reminding loss separately. In next theorem, we show that the design of loss can work cooperatively when mixtured thus the overall loss function leads to a correct global distribution.

Theorem 1. *Suppose the generator has enough model capacity to obtain $q_1(x|y) = p(x|y)$ for all $y \in g_1(y)$ and the loss $V_\tau(G_\tau, D_\tau)$ defined in Equation 1 is optimized optimally for each $\tau \in [t]$, then $q_t(x|y) = p(x|y)$ for all $y \in \Omega_t$.*

Proof:
We will rely on induction for proof of the statement. The statement is true for $t = 1$ according to our assumption and the fact that $g_1(y) = s_1(y)$. Assuming $q_{t-1}(x|y) = p(x|y)$ for all $y \in \Omega_{t-1}$, we will show $q_t(x|y) = p(x|y)$ for all $y \in \Omega_t$. Formally:

$$
\begin{aligned}
&V_t(G_t, D_t) = \min_{G_t} \max_{D_t^{1:K_t}} \mathbb{E}_{y \sim g_t(y)} \{\mathbb{E}_{x \sim p(x|y)} [\log D_t^k(x, y)] \\
&+ \mathbb{E}_{u \sim unif(0,1))} [\log(1 - D_t^k(G_t(u, y), y))]\} \\
&+ \lambda \min_{G_t} \mathbb{E}_{y \sim s_{t-1}(y)} \mathbb{E}_{u \sim unif(0,1)} [\|G_t(u, y) - G_{t-1}(u, y)\|^2] \\
&= \min_{q_t} \int\limits_{y \in y \in \Omega(g_t(y))} \sum_{k=1}^{K_t} \pi_t^k g_t^k(y) \int\limits_x p(x|y) log \frac{p(x|y)}{p(x|y) + q_t(x|y)} \\
&+ q_t(x|y) log \frac{q_t(x|y)}{p(x|y) + q_t(x|y)} dxdy \\
&+ \min_{G_t} \lambda \int\limits_{y \in \Omega_{t-1}} s_{t-1}(y) \int_u \|G_t(u, y) - G_{t-1}(u, y)\|^2 dudy \\
&\underset{*}{\geq} \int\limits_{y \in y \in \Omega(g_t(y))} \sum_{k=1}^{K_t} \pi_t^k g_t^k(y) \int\limits_x \min_{q_t} p(x|y) log \frac{p(x|y)}{p(x|y) + q_t(x|y)} \\
&+ q_t(x|y) log \frac{q_t(x|y)}{p(x|y) + q_t(x|y)} dxdy \\
&+ \lambda \int\limits_{y \in \Omega_{t-1}} s_{t-1}(y) \int_u \min_{G_t} \|G_t(u, y) - G_{t-1}(u, y)\|^2 dudy
\end{aligned}
$$

Next we show the inequality $*$ attains equality if $q_t(x|y) = p(x|y)$ for all $y \in \Omega_t$. First we note that for $y \in \Omega(g_t(y)) \cap \Omega_{t-1}$ the digesting loss and reminding loss shares the same optimal solution. Note $G_t(u, y) = G_{t-1}(u, y)$ is equivalent to $q_t(x|y) = q_{t-1}(x|y)$ since G_t and G_{t-1} shares the same random seed u. We have for $y \in \Omega(g_t(y)) \cap \Omega_{t-1}$, $q_t(x|y) = q_{t-1}(x|y) = p(x|y)$ due to the inductive assumption for reminding loss. The optimality of $q_t(x|y) = p(x|y)$ is due to Lemma 2.

Next we have $q_t(x|y) = p(x|y)$ for $y \in \Delta\Omega_t$. For $y \in \Omega(g_t(y)) - \Omega(s_{t-1}(y))$, $q_t(x|y) = p(x|y)$ according to Lemma 2. For $y \in \Omega_{t-1} - \Omega(g_t(y))$, $G_t(u, y) = G_{t-1}(u, y)$ according to Lemma 1. □

References

1. Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., Bengio, Y.: Generative adversarial nets. In: Advances in neural information processing systems. pp. 2672–2680 (2014)