دانشگاه صنعتی شریف

دانشکده برق

ارائه درس آنالیز احتمالاتی در ابعاد بالا

عنوان:

Convergence Analysis of GAN training procedures

نگارش:

رضا پیشکو - ۹۸۱۰۰۳۶۷

کسری خوشجو - ۹۸۱۰۰۴۲۹

استاد درس:

دکتر یاسایی

تیر ۱۴۰۲

۱- مقدمه

GANها مدل هایی هستند که برای یادگیری توزیع های پیچیده استفاده می‌شوند. به خصوص در تصاویر، GANها به این منظور استفاده می‌شوند که بتوانند با آموزش داده شدن روی یک مجموعه داده، داده های مشابه تولید کنند. با این وجود آموزش دادن GANها می‌تواند سخت باشد و در عمل معمولا دیده می‌شود که در GANهایی که بر پایه گرادیان هستند، همگرایی رخ نمی‌دهد.

با وجود پیشرفت های کاربردی GANها، تئوری آموزش دادن آنها هنوز به طور کامل فهمیده نشده است. Mescheder et al.(2017) نشان دادند که همگرایی محلی آموزش GANها را می‌توان با بررسی مقدار ویژه های ژاکوبین بردار گرادیان مربوطه، آنالیز کرد. اگر ژاکوبین در نقطه تعادل فقط شامل مقادیر ویژه با قسمت حقیقی منفی باشد، آموزش GAN برای طول قدم ها کوچک به صورت محلی همگرا می‌شود.

۲- Generative Adversarial Network

GANها را می‌توان به صورت یک بازی min-max که دو بازیکن دارد، دید که یک طرف شبکه discriminative هست که آن را با $D_\psi(x)$ نشان می‌دهیم و یک طرف بازی هم شبکه generative است که آن را با $G_\theta(z)$ نشان می‌دهیم. در این بازی هدف discriminator این است که بتواند داده های واقعی را از داده‌هایی که generator تولید کرده تمایز دهد و هدف generator هم این است که بتواند داده‌هایی تولید کند که discriminator نتواند آنها را از داده های اصلی تشخیص دهد.

تابع هدفی که در GANها برای آموزش دادن مدل مورد استفاده قرار می‌گیرد به فرم زیر است :

$$\mathcal{L}(\theta, \psi) = \mathbb{E}_{P_Z}\left[f(D_\psi(G_\theta(Z)))\right] + \mathbb{E}_{X\sim P_D}\left[f(-D_\psi(X))\right]$$

که در آن f هر تابع دلخواه حقیقی مقدار است.

مسئله‌ای که در GANها مورد بررسی قرار می‌گیرد، مسئله min-max روی تابع هدف بالا است یعنی :

$$\min_\theta \max_\psi \mathcal{L}(\theta, \psi)$$

که به آن zero-sum game هم می‌گویند به این دلیل که هر چقدر یکی از بین این دو شبکه سود کند به همان اندازه آن یکی شبکه ضرر کرده است.

هدف ما در مسائل GAN پیدا کردن نقطه تعادل است؛ به این معنی که پارامتر (θ^*, ψ^*) را به گونه‌ای پیدا کنیم که هیچ کدام از discriminator و generator نتوانند نتیجه بهتری بگیرند و با عوض کردن پارامتر مربوط به خودشان ضرر کنند. در GANها از یک نقطه ابتدایی شروع می‌کنیم و با آپدیت کردن آنها سعی می‌کنیم به نقطه تعادل برسیم. در واقع می‌توان این فرایند آپدیت کردن را به این صورت در نظر بگیریم که یک تابع F روی پارامتر های θ و ψ اعمال می شود و به پارامتر های جدید می‌رسیم. یعنی داریم :

$$(\theta, \psi)^{(k)} = F(\theta^{(k-1)}, \psi^{(k-1)})$$

معمولا روش هایی که برای آپدیت کردن پارامتر ها استفاده می‌شود یکی از روش های زیر است :

- simultaneous gradient descent : $F_h(\theta, \psi) = (\theta, \psi) + hv(\theta, \psi)$ where $v(\theta, \psi) = \begin{bmatrix} -\nabla_\theta \mathcal{L} \\ \nabla_\psi \mathcal{L} \end{bmatrix}$
- alternating gradient descent : $F_h = F_{2,h} \circ F_{1,h}$
- continuous(ODE solver) : $\frac{d(\theta,\psi)}{dt} = v(\theta, \psi)$

که در مورد اول (simultaneous gradient descent) بردار v بردار گرادیان است و تابع F_h همزمان پارامترهای θ و ψ را آپدیت می‌کند. در مورد دوم (alternating gradient descent) تابع $F_{1,h}$ پارامتر discriminator و تابع $F_{2,h}$ پایامتر generator را به صورت یکی در میان آپدیت می‌کنند. مورد سوم هم که در واقع روی خم انتگرال حرکت می‌کند، در عمل ممکن نیست و فقط برای بررسی استفاده می‌شود.

در مقاله اولیه GAN [1](Goodfellow et al, 2014) در مورد یکتایی نقطه تعادل این قضیه وجود دارد که اگر هر دوی generator(G_θ) و discriminator(D_ψ) بتوانند هر تابع دلخواه را تخمین بزنند، نقطه تعادل یکتا است و جایی اتفاق میفتد که generator توزیع واقعی داده را بسازد و discriminator هم همه جا روی توزیع داده برابر با مقدار صفر باشد. در واقع جایی اتفاق می افتد که داشته باشیم $G_\theta(Z) \sim P_D$

با این وجود همگرایی به نقطه تعادل لزوما صورت نمی‌گیرد و قضیه زیر در مقاله [2](Mescheder et al, 2017) در مورد رخ دادن همگرایی وجود دارد.

قضیه ۱ (شرط کافی برای همگرایی محلی[۲]). *فرض کنید تابع تکرار F یک تابع پیوسته مشتق پذیر باشد که در نقطه $\bar{x}$ دارای دو شرط زیر است :*

- *نقطه ثابت باشد یعنی داشته باشیم :* $F(\bar{x}) = \bar{x}$
- *دارای ماتریس ژاکوبین با مقادیر ویژه کمتر از ۱ باشد یعنی داشته باشیم :* $|\lambda_i(F'(\bar{x}))| < 1$

در این صورت یک همسایگی باز U از $\bar{x}$ وجود دارد که اگر داخل آن همسایگی باشیم تکرار های $F^{(k)}$ به $\bar{x}$ همگرا می‌شوند و سرعت این همگرایی حداقل خطی است. در واقع خطای $\|F^{(k)}(x_0) - \bar{x}\|$ برای $k \to \infty$ از $O(|\lambda_{max}(\bar{x})|^k)$ خواهد بود که منظور از λ_{max} بزرگ ترین مقدار ویژه ماتریس ژاکوبین در نقطه $\bar{x}$ است.

با توجه به این قضیه معمولا تابع F را به صورت $x + hv$ در نظر می‌گیرند که پیدا کردن نقطه ثابت معادل با پیدا کردن ریشه v باشد. در این صورت برای ماتریس ژاکوبین F خواهیم داشت :

$$F'(\theta, \psi) = I + hv'(\theta, \psi)$$

البته از آنجایی که F و v لزوما متقارن نیستند مقادیر ویژه آنها می‌تواند مختلط هم باشد.

برای بررسی ساده تر شرط بالا، مقاله [2](Mescheder et al, 2017)، یک شرط ساده تر ارائه داده است که به صورت زیر است.

گزاره ۱ (شرط کافی برای همگرایی[۲]). *اگر قسمت حقیقی مقادیر ویژه‌ی ژاکوبین v در نقطه‌ی ثابت $\bar{x}$ مقادیر منفی داشته باشند، در این صورت $h > 0$ به اندازه کافی کوچک وجود دارد که مقادیر ویژه‌ی $I + hv'(\bar{x})$ داخل گوی با شعاع واحد باشند.*

۳- همگرایی روش های موجود برای GANها

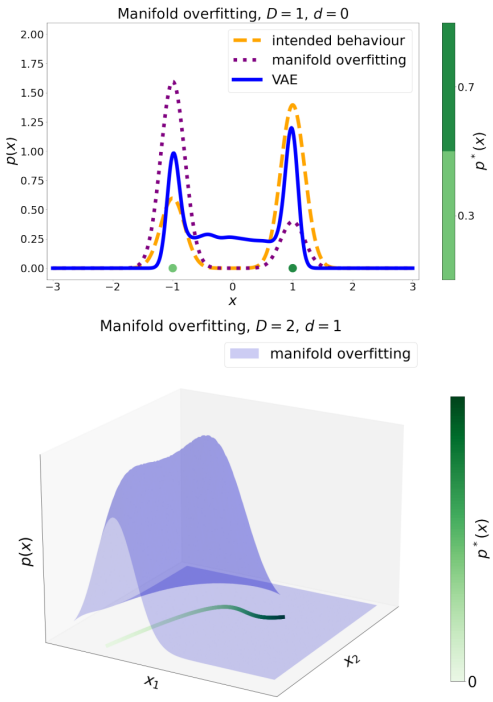
در ابتدای این بخش به بررسی دو مثال در مورد مشکلات f-divergenceها می‌پردازیم.

مثال ۱ (یادگیری خطوط موازی[۳]). *فرض کنید که $Z \sim U(0,1)$ و $\mathbb{P}_0$ توزیع $(0, Z)$ و $\mathbb{P}_\theta$ هم توزیع (θ, Z) باشد. در این صورت خواهیم داشت :*

- $W_1(\mathbb{P}_0, \mathbb{P}_\theta) = |\theta|$
- $\mathrm{JS}(\mathbb{P}_0, \mathbb{P}_\theta) = \log 2 \cdot \mathbf{1}_{\{\theta \neq 0\}}$
- $D_{KL}(\mathbb{P}_0, \mathbb{P}_\theta) = \mathbb{I}_{\{0\}}(\theta)$

مشاهده می‌شود تحت Wasserstein divergence وقتی که $\theta \to 0$، آنگاه دنباله $\mathbb{P}_\theta$ هم به سمت $\mathbb{P}_0$ می‌رود ولی تحت KL و Jensen Shannon همگرایی اتفاق نمیفتد. این مثال یک حالتی بود که با استفاده از Wasserstein divergence میتوانیم توزیع احتمال را روی یک منیفلد با بعد کمتر به دست آوریم ولی با بقیه فاصله ها و divergenceها نمی‌توانیم.

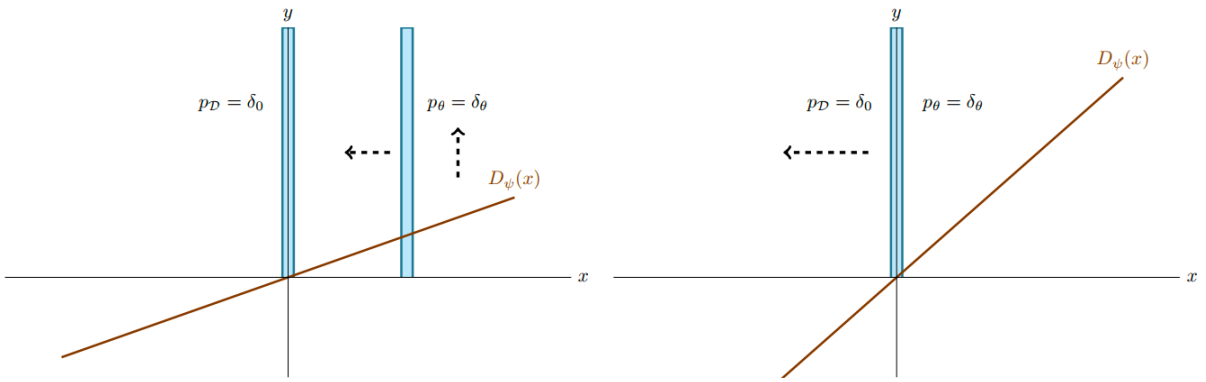
مثال ۲. *(Manifold overfitiing)*



این مثال هم به طور مشابه نشان دهنده مشکل f-divergenceها هست. همانطور که در تصاویر نشان داده شده است خط های آبی نشان دهنده توزیع اصلی هستند که با استفاده از f-divergence ها فقط توانسته‌است maniflodای که توزیع روی آن است را پیدا کند ولی توزیع را به درستی پیدا نکرده است.

حال برای بررسی همگرا بودن روش های آموزش GAN یک مساله آزمایشی یادگیری توزیع مطرح می‌کنیم:

مساله Dirac-GAN[۴]: توزیع اصلی داده ها دارای تابع چگالی $P_D = \delta_0$ روی مقادیر حقیقی است، توزیع مدل مولد $P_\theta = \delta_\theta$روی مقادیر حقیقی با پارامتر یادگیری θ است. مدل discriminator نیز تابع $D_\psi(x) = \psi \cdot x$ است.



(a) $t = t_0$ (b) $t = t_1$

در اینصورت تابع هدفی که قبل تر تعریف کردیم برای مدل Dirac-GAN به صورت زیر خواهد بود:

$$\mathcal{L}(\theta, \psi) = f(\psi \cdot \theta) + f(0) \quad ; \quad f(x) = -\log\left(1 + \exp\left(-x\right)\right)$$

که نقطه تعادل نش آن در $\theta = \psi = 0$ قرار دارد. مقادیر ویژه$v'(\theta, \psi)$ برابر با $\pm f'(0)i$ هستند و بنابراین شرط کافی برای همگرایی نداریم. در ادامه مقادیر ویژه ماتریس ژاکوبی برای روش های بهینه سازی تکراری مختلف را بررسی می‌کنیم:

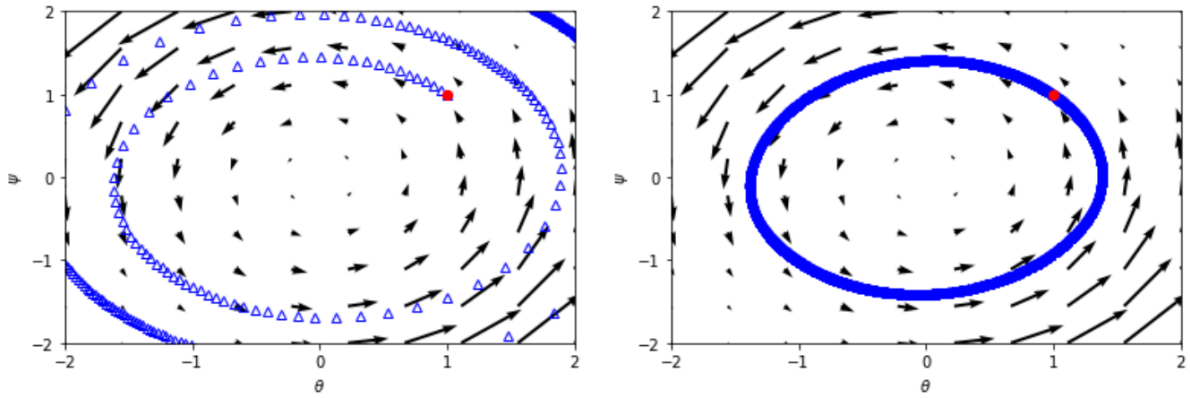
لم ۱. *در مساله Dirac-GAN در هیچ همسایگی مشتبی از x خم های انتگرال میدان $v(\theta, \psi)$ به نقطه تعادل همگرا نمی‌شوند. در حقیقت هر خم انتگرال شامل نقاطی به فرم $\theta(t)^2 + \psi(t)^2 = const.\,; \forall t$ است.*

لم ۲. *در روش گرادیان کاهشی همزمان، در نقطه تعادل مقادیر ویژه ژاکوبی تابع $F_h(\theta, \psi)$ به فرم $\lambda_{1,2} = 1 \pm hf'(0)i$ هستند که نرم آنها از یک بیشتر است و درنتیجه الگوریتم تکراری حول نقطه تعادل ناپایدار است.*

لم ۳. *در روش کاهش گرادیان یکی در میان،اگر n_g بار مقدار θ و بعد n_d بار مقدار ψ در هر قدم به روز رسانی شوند مقادیر ویژه ژاکوبی F در نقطه تعادل به فرم زیر هستند:*

$$\lambda_{1,2} = 1 - \frac{\alpha^2}{2} \pm \sqrt{\left(1 - \frac{\alpha^2}{2}\right)^2 - 1} \quad ; \quad \alpha = \sqrt{n_g n_d} h f'(0)$$

که یعنی داخل دایره واحد نیستند.



(a) SimGD (b) AltGD

در آزمایش های عددی هم می توانیم ببینیم که همگرایی در این روش ها رخ نمی دهد. مساله ای که پیش می آید این است که ریشه مشکل عدم همگرایی چیست. در ادامه قضیه ای را مطرح می کنیم که مارا در حل کردن این مشکل راهنمایی می کند.

قضیه ۲ (همگرایی برای توزیع های پیوسته[۵]). *با فرضیات معقولی در مورد مساله، اگر توزیع مدل generative و توزیع اصلی داده ها در همه نقاط فضای داده ها پیوسته باشند ، روش گرادیان کاهشی به نقطه تعادل همگرا می‌شود.*

بنابراین عدم پیوستگی و صادق بودن فرضیه manifold احتمالا علت بوجود آمدن مشکل عدم همگرایی است. با این مشاهده در ادامه به ارائه روش هایی می پردازیم که می توانند در مساله Dirac-GAN و احتمالا مسائل دنیای واقعی به نقطه تعادل همگرا شوند.

۴- طراحی روش هایی که به نقطه تعادل همگرا می شوند

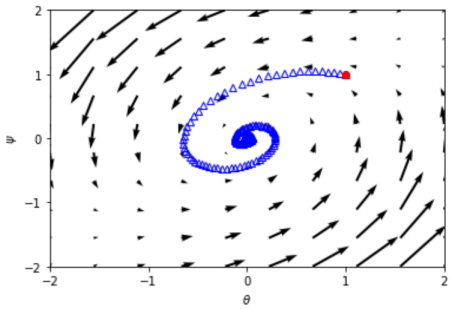
۴-۱- اضافه کردن نویز گاوسی

با اضافه کردن نویز گاوسی به مثال ها، هر دو توزیع در تمام نقاط پیوسته می شوند و بنابراین طبق قضیه ای که قبل تر دیدیم انتظار می رود همگرایی رخ دهد.

لم ۴. *با اضافه کردن نویز گاوسی به مثال ها در مساله Dirac-GAN مقادیر ویژه $v(\theta,\psi)$ در نقطه تعادل به فرم زیر هستند.*

$$\lambda_{1,2} = f''(0)\sigma^2 \pm \sqrt{f''(0)^2\sigma^4 - f'(0)^2}$$

بنابراین انتظار داریم وقتی $f''(0) < 0$ همگرایی رخ دهد. آزمایش عددی این روش در تصویر زیر قابل مشاهده است.



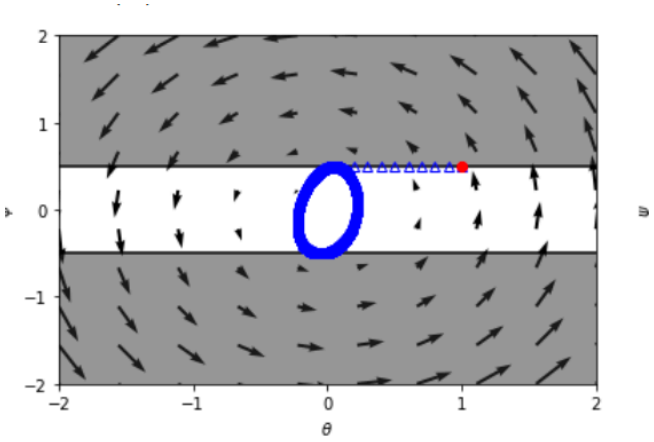
(f) Instance noise

۴-۲- فاصله Wasserstein

همان طور که در مثال ۱ دیدیم تنها فاصله ای که وقتی دو توزیع به هم نزدیک می شدند به صفر همگرا می شد فاصله Wasserstein بود. در مدل های Wasserstein-GAN تابع هدف از فرم وردشی فاصله به صورت زیر بدست می‌آید.

$$\mathcal{L}(\theta, \psi) = \mathbb{E}_{P_Z}\left[D_\psi(G_\theta(Z))\right] - \mathbb{E}_{X\sim P_D}\left[D_\psi(X)\right]; \quad D_\psi \text{ is 1-Lipschitz}$$

اگر این روش را روی Dirac-GAN اعمال کنیم همگرایی رخ نمی دهد و آزمایشی عددی آنرا در شکل زیر می توانیم ببینیم:



(c) WGAN ($n_d = 5$)

اگر تابع هدف را به صورت زیر تغییر می‌دهیم می‌توان نشان داد که الگوریتم همگرا می شود.[۶]

- ابتدا مساله خطی زیر را برای batch داده های این قدم حل کنیم:

$$\begin{cases} \max\limits_{H_i, H_j} & \frac{1}{m}\sum H_j - \frac{1}{n}\sum H_i \\ \text{to subject} & |H_i - H_j| \le \frac{1}{2}\|x_i - y_j\|^2 \end{cases}$$

که درحقیقت فاصله Wasserstein بین دو توزیع تجربی را بدست می‌آورد.

- حال مساله انتقال بهینه زیر را حل می‌کنیم:

$$\sigma(j) = \underset{i}{\text{argmin}} \ \frac{1}{2}\|x_i - y_j\|^2 + H_i^* - H_j^*$$

- حال تابع هدف را به این شکل بهینه سازی می کنیم:

$$\min_{\psi} \frac{1}{2}\left(\frac{1}{m}\sum D_\psi(y_j) - \frac{1}{n}H_j^*\right)^2 + \frac{1}{2}\cdot\frac{1}{n}\sum\left(D_\psi(x_i) - H_i^*\right)^2$$
$$+ \underbrace{\frac{\lambda}{2}\mathbb{E}_{X\sim G_\theta(Z)}\left[\left(\|\nabla_x D_\psi(X)\| - \|y_{\sigma^*(X)} - X\|\right)^2\right]}_{\text{regulrization gradient}}$$
$$\min_{\theta} \mathcal{L}(\theta) = -\frac{1}{n}\sum D_\psi(G_\theta(Z_i))$$

مشتق گیری از توابعی که در این روش به دنبال بهینه سازیشان هستیم و اعمال گرادیان کاهشی منجر به رسیدن به ماتریس ژاکوبی زیر می‌شود:

$$-v'(\theta^*, \psi^*) = \begin{bmatrix} M_{DD} + M_R & M_{GD} \\ 0 & 0 \end{bmatrix},$$

$$M_R = \lambda \cdot \mathbb{E}_{P_Z}\left[\nabla_{\theta,\psi} D_{\psi^*}(G_\theta(Z))\nabla_{\theta,\psi} D_{\psi^*}(G_\theta(Z))^T\right]$$
$$M_{DD} = \mathbb{E}_{y\sim P_D}\left[\nabla_\psi D_{\psi^*}(y)\right]\mathbb{E}_{y\sim P_D}\left[\nabla_\psi D_{\psi^*}(y)\right]^T$$
$$+ \mathbb{E}_{y\sim P_D}\left[\nabla_\psi D_{\psi^*}(y)\nabla_\psi D_{\psi^*}(y)^T\right]$$
$$M_{GD} = -\mathbb{E}_{Z\sim P_Z}\left[\nabla^2 D_{\psi^*}(G_\theta(Z))\nabla_\theta G_{\theta^*}(Z)^T\right]$$

بنابراین به وضوح $v'(\theta, \psi)$منفی معین است و طبق قضیه ای که قبلا مطرح کرده بودیم این روش حتما با شروع از نقطه ای در همسایگی نقطه تعادل همگرا می‌شود.

مراجع

[1] G. et al., "Generative adversarial networks," *NIPS*, 2014.
[2] M. et al., "The numerics of gans," *NIPS*, 2017.
[3] A. et al., "Wasserstein gan," *PMLR*, 2014.
[4] M. et al., "Which training methods for gans do actually converge?," *ICML*, 2018.
[5] N. et al., "Gradient descent gan optimization is locally stable.," *NIPS*, 2017.
[6] L. et al., "Wasserstein gan with quadratic transport cost," *CVF*, 2019.