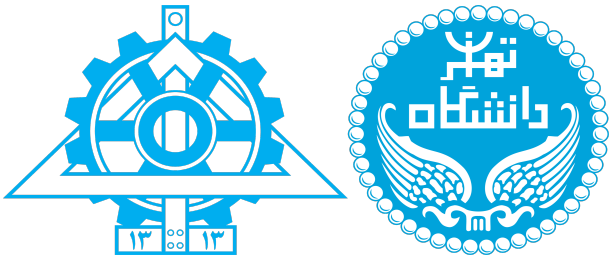


دانشگاه تهران

پردیس دانشکده‌های فنی

دانشکده مهندسی برق و کامپیوتر

تمرین شماره: ۳

مدل‌های مولد عمیق

نام و نام‌خانوادگی: آرتین توسلی

شماره دانشجویی: ۸۱۰۱۰۲۵۴۳

نیم‌سال اول

سال تحصیلی ۴۰۵-۴۰۴

فهرست مطالب

۱ **Energy Based Models** ۴

۱.۱ تئوری ۴

۱.۱.۱ Conditional EBMs ۴

۲.۱.۱ Rejection Sampling ۶

۳.۱.۱ Contrastive Divergence ۹

۲.۱ عملی ۱۰

۱.۲.۱ بارگذاری دیتاست و آماده‌سازی داده ۱۰

۲.۲.۱ پیاده‌سازی معماری و آموزش مدل ۱۲

۳.۲.۱ تولید تصویر از نویز خالص ۱۴

۴.۲.۱ نویززدایی تصویر نویزی ۱۶

۲ **Score Based Models** ۲۱

۱.۲ تئوری ۲۱

۱.۱.۲ استقلال از تابع پارتیشن ۲۱

۲.۱.۲ چالش‌های Matching Score و روش DSM ۲۲

۳.۱.۲ چالش تکیه گاه های مجزا ۲۵

۴.۱.۲ چالش‌های داده‌های واقعی و راهکار NCSN ۲۷

۲.۲ عملی ۲۸

۱.۲.۲ پیاده‌سازی مدل پایه ۲۸

۲.۲.۲ پیاده‌سازی مدل شرطی ۳۱

فهرست تصاویر

۱.۱ نمونه تصاویر تصادفی از مجموعه داده آموزش و آزمون ۱۱
۲.۱ نمودارهای میانگین انرژی داده‌های واقعی، مصنوعی و تابع هزینه کل در طول ۱۰ اپاک . ۱۳
۳.۱ روند بهبود کیفیت نمونه‌های تولیدی (CD-۶۰) از اپاک ۰ تا ۹ . . . . . ۱۳
۴.۱ مقایسه تصاویر آموزشی اصلی (ردیف بالا) و خروجی نمونه‌برداری Langevin (ردیف پایین) ۱۴
۵.۱ مقایسه تصاویر تولید شده از نویز خالص (سمت چپ) و تصاویر واقعی مجموعه آموزش (سمت راست) ۱۵
۶.۱ نتایج نویززدایی در ۳۰ گام برای سطوح نویز 0.2, 0.4, 0.6 (ردیف بالا: واقعی، وسط: نویزی، پایین: بازسازی شده) ۱۷
۷.۱ اثر "تغییر هویت" تصاویر به عدد ۱ در تعداد گام‌های بالا (۱۰۰ گام) ۱۹

۱.۲ نمودار کاهش تابع هزینه؛ پایداری در آموزش مدل SBM مشهود است. ۲۹
۲.۲ گرید تصاویر تولید شده؛ وضوح لبه‌ها و تنوع اعداد تاییدکننده عملکرد صحیح مدل پایه است. ۳۰
۳.۲ روند تبدیل شدن نویز خالص به عدد نهایی طی مراحل کاهش نویز ۳۰
۴.۲ نمودار کاهش تابع هزینه مدل شرطی. ۳۲
۵.۲ گرید اعداد تولید شده به صورت شرطی؛ هر ردیف مربوط به یک کلاس خاص است که نشان‌دهنده کنترل کامل بر فرآیند تولید است. ۳۲

فهرست جداول

۱.۲ مقایسه ویژگی‌های کلیدی مدل پایه و مدل شرطی ۳۳

سوال ۱

Energy Based Models

۱.۱ تئوری

۱.۱.۱ Conditional EBMs

استفاده از مدل‌های مبتنی بر انرژی شرطی (Conditional EBMs)

به جای اینکه مدل تنها توزیع $p(x)$ را یاد بگیرد، توزیع شرطی $p(x|y)$ را مدل‌سازی می‌کند که در آن y بردار ویژگی‌ها (مانند مرد و جوان) است.

تابع انرژی به صورت $E_\theta(x, y)$ تعریف می‌شود. در این حالت توزیع احتمال برابر است با:

$$p_\theta(x|y) = \frac{e^{-E_\theta(x,y)}}{Z(\theta, y)} \tag{۱.۱}$$

که در آن $Z(\theta, y) = \int e^{-E_\theta(x,y)} dx$ تابع پارتیشن وابسته به کلاس است. **این روش آموزش را راحتر میکند بخاطر:**

۱. **یادگیری ویژگی‌های مشترک (Shared Features):** تمام تصاویر چهره (زن، مرد، پیر، جوان) دارای ساختارهای مشترکی هستند (محل چشم‌ها، ساختار پوست و ...). وقتی یک مدل شرطی روی

کل دیتاست آموزش می‌بیند، این ویژگی‌های پایه را با قدرت و داده‌های بسیار بیشتری یاد می‌گیرد. اگر داده‌ها را جدا کنیم، هر مدل تنها بخش کوچکی از داده‌ها را می‌بیند.

۲. **هدایت گرادیان (Gradient Guidance):** در زمان تولید تصویر (Sampling)، ما متغیر y را روی ویژگی مطلوب (مثلاً $y = \{\text{Male} \text{ ، } \text{Young}\}$) فیکس می‌کنیم. سپس با استفاده از الگوریتم Langevin Dynamics، نمونه‌ها را دقیقاً به سمت مینیمم انرژی مربوط به آن کلاس هل می‌دهیم:

$$x_{t+1} = x_t - \frac{\epsilon}{2}\nabla_x E_\theta(x_t, y_{target}) + \sqrt{\epsilon}\cdot\omega_t \qquad (۲.۱)$$

یک ایده دیگری که میشه داد این است که از خاصیت composition مدل های انرژی استفاده کنیم. آموزش دو مدل جداگانه (یکی برای مرد $P_m(x)$ و یکی برای جوان $P_y(x)$) و سپس ضرب کردن آن‌ها،

$$P_{new}(x) \propto P_m(x)\cdot P_y(x) \qquad (۳.۱)$$

با جایگذاری فرم انرژی $P(x) \propto e^{-E(x)}$ خواهیم داشت:

$$P_{new}(x) \propto e^{-E_m(x)}\cdot e^{-E_y(x)} = e^{-(E_m(x)+E_y(x))} \qquad (۴.۱)$$

این یعنی انرژی کل برابر با مجموع انرژی‌ها خواهد بود: $E_{total}(x) = E_m(x) + E_y(x)$.

گرچه در این حالت دیتاست پیدا کردن راحتر است یعنی گر ما N ویژگی داشته باشیم (جوان، عینکی، خندان، مرد، ...)، برای آموزش یک مدل شرطی کامل نیاز داریم که تمام ترکیب‌های ممکن (2^N حالت) در دیتاست وجود داشته باشند. اما با روش شما، فقط N مدل جداگانه آموزش می‌دهیم و بعداً آن‌ها را با هم جمع می‌کنیم. و یعنی data efficiency بیشتری داریم اما آموزش این مدل سخت میباشد بخاطر اینکه اولا برای صحت اینکار باید این دو ویژگی مستقل فرض بشوند و همینطور فرض کنید دو مدل جداگانه آموزش داده‌اید:مدل جنسیت: $E_{male}(x)$ مقادیری بین 0 تا 1000 تولید می‌کند.مدل سن: $E_{young}(x)$ مقادیری بین 0 تا 10 تولید می‌کند.وقتی این‌ها را جمع می‌کنید:

$$E_{total}(x) = E_{male}(x) + E_{young}(x)$$

تاثیر مدل "جنسیت" ۱۰۰ برابر بیشتر از "سن" است. در نتیجه تصویر تولید شده قطعاً "مرد" خواهد بود، اما

اصلاً "جوان" نمی‌شود (چون گرادیان مدل دوم در برابر اولی گم می‌شود).در روش ،Conditional چون مدل همزمان روی هر دو ویژگی آموزش می‌بیند، شبکه عصبی خودش به صورت خودکار وزن‌ها را تنظیم می‌کند تا هر دو ویژگی تاثیر متناسبی داشته باشند. در روش جداگانه، شما باید دستی ضریب λ را تنظیم کنید $(E_m + \lambda E_y)$ که کار دشواری است.

۲.۱.۱ Rejection Sampling

روش Rejection Sampling تکنیکی برای تولید نمونه از یک توزیع هدف پیچیده $P(x)$ است که نمونه‌برداری مستقیم از آن دشوار است، اما می‌توانیم مقدار چگالی آن را (معمولاً تا حد یک ثابت نرمال‌سازی) محاسبه کنیم (مثلاً $P(x) = \tilde{p}(x)/Z$). برای انجام این کار، از یک توزیع پیشنهادی (Proposal Distribution) ساده‌تر مانند $Q(x)$ استفاده می‌کنیم که نمونه‌برداری از آن آسان است (مانند توزیع گوسی). شرط اصلی این الگوریتم وجود یک ثابت M است به طوری که توزیع پیشنهادی اسکیل‌شده، همیشه بالاتر از توزیع هدف قرار گیرد:

$$\forall x : \quad M \cdot q(x) \geq \tilde{p}(x) \tag{۵.۱}$$

مراحل الگوریتم:

۱. یک نمونه x از توزیع پیشنهادی $Q(x)$ تولید می‌کنیم.

۲. یک عدد تصادفی u از توزیع یکنواخت در بازه $[0, 1]$ تولید می‌کنیم.

۳. شرط پذیرش را بررسی می‌کنیم:

$$u \leq \frac{\tilde{p}(x)}{M \cdot q(x)} \tag{۶.۱}$$

۴. اگر شرط برقرار بود، x را به عنوان نمونه‌ای از $P(x)$ می‌پذیریم (Accept). در غیر این صورت، آن را دور ریخته (Reject) و دوباره به مرحله اول برمی‌گردیم.

حال Optimal Acceptance Rate را برای دیتاست MNIST محاسبه میکنیم:

- توزیع هدف: $P(x) = \mathcal{N}(\mu, \sigma_p^2 I)$
- توزیع پیشنهادی: $Q(x) = \mathcal{N}(\mu, \sigma_q^2 I)$

- رابطه انحراف معیارها: $\sigma_q = 1.01\sigma_p$
- ابعاد داده (MNIST): تصاویر 28×28 هستند، بنابراین بُعد فضا $D = 784$ است.

محاسبه ثابت M: همانطورکه در فرمول بالا دیدیم، باید نامساوی برای هر x ای صدق کند پس M را حداقل مقدار بیشترین نسبت p به q تعریف میکنیم یعنی:

$$M = \max_x \frac{p(x)}{q(x)} \tag{۷.۱}$$

نسبت چگالی دو توزیع گوسی با میانگین برابر به صورت زیر است:

$$\frac{p(x)}{q(x)} = \frac{(2\pi\sigma_p^2)^{-D/2} \exp\left(-\frac{\|x-\mu\|^2}{2\sigma_p^2}\right)}{(2\pi\sigma_q^2)^{-D/2} \exp\left(-\frac{\|x-\mu\|^2}{2\sigma_q^2}\right)} \tag{۸.۱}$$

با ساده‌سازی عبارت، به رابطه زیر می‌رسیم:

$$\frac{p(x)}{q(x)} = \left(\frac{\sigma_q}{\sigma_p}\right)^D \exp\left(-\frac{\|x-\mu\|^2}{2}\left(\frac{1}{\sigma_p^2} - \frac{1}{\sigma_q^2}\right)\right) \tag{۹.۱}$$

از آنجا که $\sigma_q > \sigma_p$، عبارت داخل پرانتزِ توان ($\frac{1}{\sigma_p^2} - \frac{1}{\sigma_q^2}$) مثبت است. بنابراین کل عبارت نمایی همواره کوچکتر یا مساوی ۱ است ($\exp(-\text{positive} \times \text{distance})$). ماکسیمم این کسر زمانی رخ می‌دهد که توان نمایی صفر شود، یعنی در $x = \mu$:

$$M = \left(\frac{\sigma_q}{\sigma_p}\right)^D \cdot e^0 = (1.01)^{784} \tag{۱۰.۱}$$

محاسبه نرخ پذیرش (Acceptance Rate):

اثبات فرمول زیر در آخر این بخش آورده شده است.

$$\text{Acceptance Rate} = \frac{1}{M} = \left(\frac{1}{1.01}\right)^{784} \tag{۱۱.۱}$$

حال مقدار عددی آن را محاسبه می‌کنیم:

$$\text{Acceptance Rate} \approx (0.990099)^{784} \approx 4.08 \times 10^{-4} \tag{۱۲.۱}$$

این نرخ بسیار کم میباشد و این نتیجه نشان دهنده Curse of Dimensionality است. ناحیه‌ی همپوشانی موثر دو توزیع بسیار ناچیز است. این یعنی برای تولید هر ۱ نمونه موفق، باید به طور میانگین حدود ۲۵۰۰ بار نمونه‌برداری کنیم که بسیار ناکارآمد است.

اثبات نرخ پذیرش درRejection Sampling:

$$P(\text{Accept}) = \int P(\text{Accept}|x) \cdot q(x)\, dx \tag{۱۳.۱}$$

(دقت کنید که چون x از $q(x)$ نمونه‌برداری می‌شود، وزن احتمالی آن $q(x)$ است).

حال، احتمال شرطی $P(\text{Accept}|x)$ چیست؟ چون u به صورت یکنواخت در بازه $[0,1]$ توزیع شده است، احتمال اینکه u کوچکتر یا مساوی یک مقدار A (که $0 \leq A \leq 1$) باشد، دقیقاً برابر با خود A است. در اینجا $A = \frac{p(x)}{Mq(x)}$ است. بنابراین:

$$P(\text{Accept}|x) = \frac{p(x)}{Mq(x)} \tag{۱۴.۱}$$

این مقدار را در رابطه انتگرالی اول جایگذاری می‌کنیم:

$$P(\text{Accept}) = \int \left(\frac{p(x)}{Mq(x)} \right) \cdot q(x)\, dx \tag{۱۵.۱}$$

عبارت $q(x)$ از صورت و مخرج کسر ساده می‌شود:

$$P(\text{Accept}) = \int \frac{p(x)}{M}\, dx \tag{۱۶.۱}$$

$$P(\text{Accept}) = \frac{1}{M} \int p(x)\, dx \tag{۱۷.۱}$$

می‌دانیم که $p(x)$ یک تابع چگالی احتمال (PDF) معتبر است، بنابراین انتگرال آن روی کل فضا برابر با ۱ است ($\int p(x)dx = 1$):

$$P(\text{Accept}) = \frac{1}{M} \cdot 1 = \frac{1}{M} \tag{۱۸.۱}$$

۳.۱.۱ Contrastive Divergence

(الف) تحلیل تغییرات انرژی در فرآیند آموزش:

رابطه گرادیان لگاریتم درست‌نمایی (Log-Likelihood) به صورت زیر داده شده است:

$$\nabla_\theta \log p_\theta(x_{train}) = \nabla_\theta f_\theta(x_{train}) - \mathbb{E}_{x_{sample} \sim p_\theta}[\nabla_\theta f_\theta(x_{sample})] \tag{۱۹.۱}$$

از آنجا که در روش Maximum Likelihood هدف ما بیشینه کردن تابع درست‌نمایی است، باید در جهت مثبت گرادیان حرکت کنیم. با توجه به علامت‌های مثبت و منفی در معادله بالا، فرآیند آموزش سعی دارد تغییرات زیر را اعمال کند:

۱. برای داده‌های آموزشی (x_{train}): جمله اول ($\nabla_\theta f_\theta(x_{train})$) دارای علامت مثبت است. آموزش سعی می‌کند مقدار $f(x)$ را برای داده‌های واقعی افزایش دهد. (pullup)

۲. برای نمونه‌های تولیدی مدل (x_{sample}): جمله دوم ($-\mathbb{E}[\nabla_\theta f_\theta(x_{sample})]$) دارای علامت منفی است. آموزش سعی می‌کند مقدار $f(x)$ را برای نمونه‌هایی که خود مدل تولید کرده است (و احتمالا هنوز واقعی نیستند) کاهش دهد. (pulldown)

از آنجا که انرژی معکوس $f(x)$ است، مدل سعی دارد:

- انرژی داده‌های واقعی (x_{train}) را کاهش دهد (به سمت دره‌های انرژی هل دهد).
- انرژی نمونه‌های تولیدی فعلی (x_{sample}) را افزایش دهد (آن‌ها را بالا ببرد تا احتمال وقوعشان کم شود، مگر اینکه با داده‌های واقعی منطبق شوند).

(ب) چالش محاسباتی و راهکار CD:

چالش محاسباتی ترم دوم: ترم دوم شامل محاسبه امید ریاضی (Expectation) روی نمونه‌های

تولید شده از خود مدل ($x_{sample} \sim p_\theta$) است:

$$\mathbb{E}_{x\sim p_\theta}[\nabla f(x)] = \int p_\theta(x)\nabla f(x)\,dx \tag{۲۰.۱}$$

برای محاسبه دقیق این ترم، نیاز داریم از توزیع فعلی مدل $p_\theta(x)$ نمونه‌برداری کنیم. از آنجا که ثابت نرمال‌سازی $Z(\theta)$ را نداریم، باید از روش‌های MCMC (مانند Langevin Dynamics) استفاده کنیم. مشکل اصلی این است که برای رسیدن به نمونه‌های دقیق از توزیع مدل، زنجیره MCMC باید به تعادل برسد که نظریهً نیازمند تعداد گام‌های بسیار زیاد (بی‌نهایت) است. انجام این کار در هر گام از آموزش (Iteration) بسیار زمان‌بر و عملاً غیرممکن است.

راهکار Contrastive Divergence:

این روش یک تقریب برای حل این مشکل است. این روش به جای اجرای زنجیره MCMC تا تعادل کامل، تغییرات زیر را اعمال می‌کند:

۱. مقداردهی اولیه: به جای شروع از نویز تصادفی، زنجیره MCMC را از داده‌های آموزشی (x_{train}) شروع می‌کند.

۲. محدود کردن گام‌ها: زنجیره تنها به تعداد محدودی گام (k گام، معمولاً حتی $k = 1$) اجرا می‌شود.

ایده این است که اگرچه نمونه‌های حاصل دقیقاً از p_θ نیستند، اما جهت تقریبی گرادیان را به درستی نشان می‌دهند (مدل سعی می‌کند از فرار کردن داده‌های واقعی به سمت نواحی با انرژی پایین‌تر جلوگیری کند) و محاسبات بسیار سریع‌تر انجام می‌شود.

۲.۱ عملی

۱.۲.۱ بارگذاری دیتاست و آماده‌سازی داده

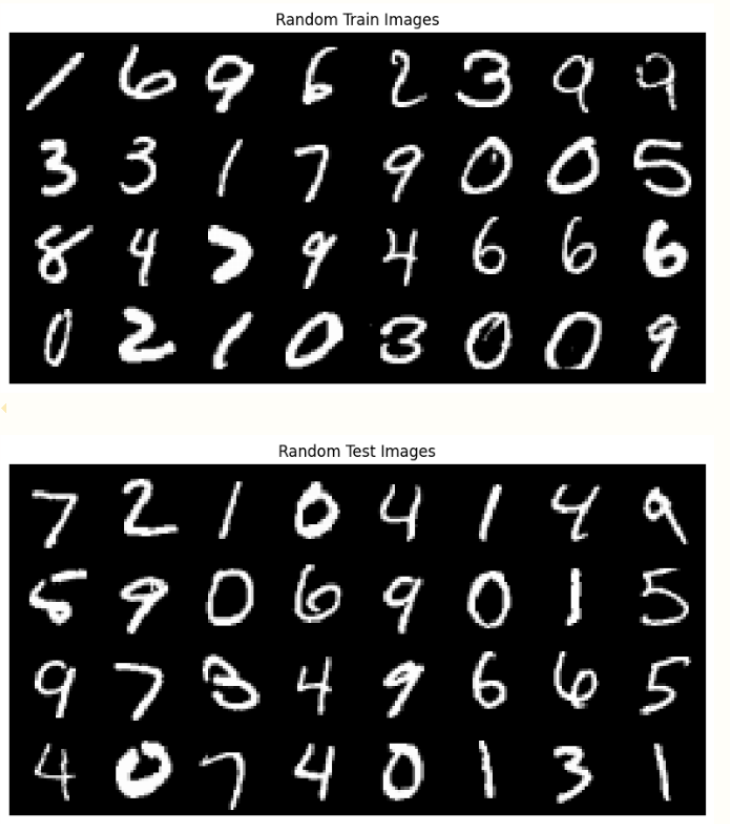
در این بخش، مجموعه داده MNIST را با استفاده از کتابخانه torchvision بارگذاری کردیم. این دیتاست شامل تصاویر دست‌نویس ارقام ۰ تا ۹ است.

۱. پیش‌پردازش داده‌ها: با استفاده از تبدیل transforms.ToTensor، تصاویر به تانسورهایی با ابعاد $(1, 28, 28)$ تبدیل شدند. این تبدیل به صورت خودکار مقادیر پیکسل‌ها را از بازه $[0, 255]$ به بازه $[0, 1]$ نرمال‌سازی می‌کند.

۲. ایجاد DataLoader: برای داده‌های آموزش (Train) و آزمون (Test)، DataLoaderهای جداگانه با اندازه دسته (Batch Size) برابر با ۶۴ تعریف شد. داده‌های آموزشی در هر Epoch بر زده (Shuffle) می‌شوند تا فرآیند آموزش تعمیم‌پذیری بهتری داشته باشد.

۳. نمایش داده‌ها: تابعی برای نمایش یک دسته از تصاویر به صورت شبکه‌ای (Grid) پیاده‌سازی شد تا صحت بارگذاری داده‌ها تایید شود.

در ادامه، نمونه‌ای از تصاویر بارگذاری شده از مجموعه آموزش و آزمون نمایش داده شده است:



شکل ۱.۱: نمونه تصاویر تصادفی از مجموعه داده آموزش و آزمون

۲.۲.۱ پیاده‌سازی معماری و آموزش مدل

در این زیربخش، یک شبکه عصبی کانولوشنی مطابق با مشخصات جدول ۱ پیاده‌سازی شد. این مدل با دریافت تصویر ورودی، یک مقدار عددی به عنوان سطح انرژی اختصاص می‌دهد. برای آموزش، از الگوریتم Contrastive Divergence (CD) و نمونه‌برداری Langevin استفاده شده است.

ساختار شبکه شامل لایه‌های Conv2d با فعال‌ساز LeakyReLU، لایه AdaptiveAvgPool2d و در نهایت یک لایه خطی برای تولید خروجی تک‌بعدی است. تابع هزینه استفاده شده برای آموزش به صورت زیر تعریف شده است:

$$\mathcal{L}_{data} = \frac{1}{B}\sum_{i=1}^{B} E_{real}^{(i)} - \frac{1}{B}\sum_{i=1}^{B} E_{fake}^{(i)}$$

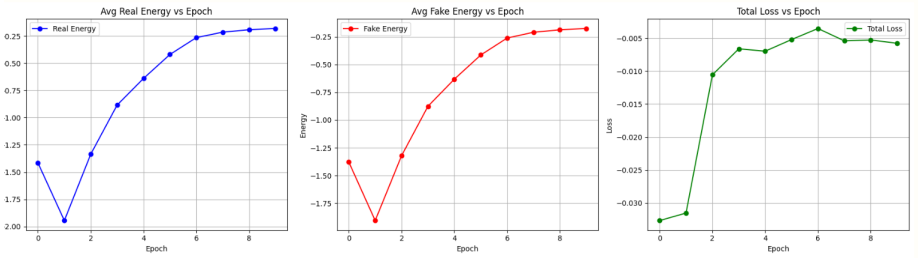
$$\mathcal{L}_{reg} = \lambda \left(\frac{1}{B}\sum_{i=1}^{B} (E_{real}^{(i)})^2 + \frac{1}{B}\sum_{i=1}^{B} (E_{fake}^{(i)})^2 \right)$$

$$\mathcal{L} = \mathcal{L}_{data} + \mathcal{L}_{reg}$$

که در آن تفاوت میانگین انرژی داده‌های واقعی و مصنوعی را محاسبه کرده و برای جلوگیری از بزرگ شدن بیش از حد مقادیر انرژی (Regularization) به کار می‌رود.

تحلیل روند آموزش (نمودارها)

با توجه به نتایج حاصل از آموزش در شکل ۲.۱، روند همگرایی مدل قابل تحلیل است:

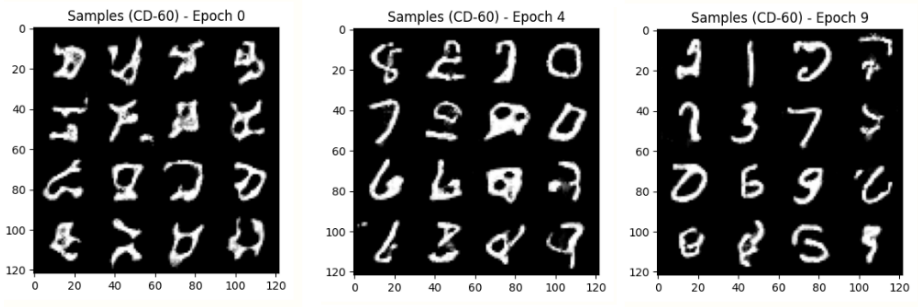


شکل ۲.۱: نمودارهای میانگین انرژی داده‌های واقعی، مصنوعی و تابع هزینه کل در طول ۱۰ اپاک

- **روند صعودی انرژی‌ها:** برخلاف انتظار اولیه مبنی بر کاهش مطلق انرژی واقعی، مشاهده می‌شود که هر دو نمودار انرژی واقعی (آبی) و مصنوعی (قرمز) پس از اپاک ۱ روند صعودی داشته و از مقادیر منفی (حدود) به سمت صفر (حدود) حرکت کرده‌اند. این رفتار به دلیل غلبه ترم منظم‌ساز است که سعی دارد توان دوم انرژی‌ها را کمینه کرده و آن‌ها را به سمت صفر سوق دهد تا از ناپایداری عددی جلوگیری کند.
- **تابع هزینه کل** روند صعودی نمودار خطا و تثبیت آن در نزدیکی صفر، نشان‌دهنده رسیدن به نقطه تعادل بین ترم داده (که انرژی‌ها را از هم دور می‌کند) و ترم منظم‌ساز (که انرژی‌ها را به سمت صفر می‌کشد) می‌باشد.

بررسی کیفی تصاویر تولید شده در طول آموزش

تغییرات خروجی مدل در اپاک‌های مختلف در شکل ۳.۱ نمایش داده شده است:



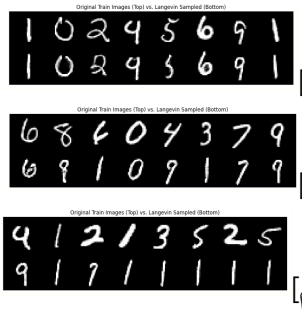
شکل ۳.۱: روند بهبود کیفیت نمونه‌های تولیدی (CD-۶۰) از اپاک ۰ تا ۹

در اپاک ابتدایی، تصاویر صرفاً نویزهای ساختاریافته هستند. با پیشرفت آموزش در اپاک ۴ و ۹، مشاهده

می‌شود که مدل ویژگی‌های هندسی ارقام MNIST مانند خطوط و حلقه‌ها را یاد گرفته و تصاویری شبیه به اعداد تولید می‌کند. (در اپاگ آخر اعداد ۰ و ۱ و ۳ و ۵ و ۶ و ۷ و ۹ قابل تشخیص هستند)

نمونه‌برداری Langevin از تصاویر آموزشی

در این آزمایش، تصاویر واقعی به عنوان نقطه شروع به تابع Langevin Sampling داده شدند. نتایج با تعداد گام‌های مختلف در شکل ۴.۱ مقایسه شده است:



شکل ۴.۱: مقایسه تصاویر آموزشی اصلی (ردیف بالا) و خروجی نمونه‌برداری Langevin (ردیف پایین)

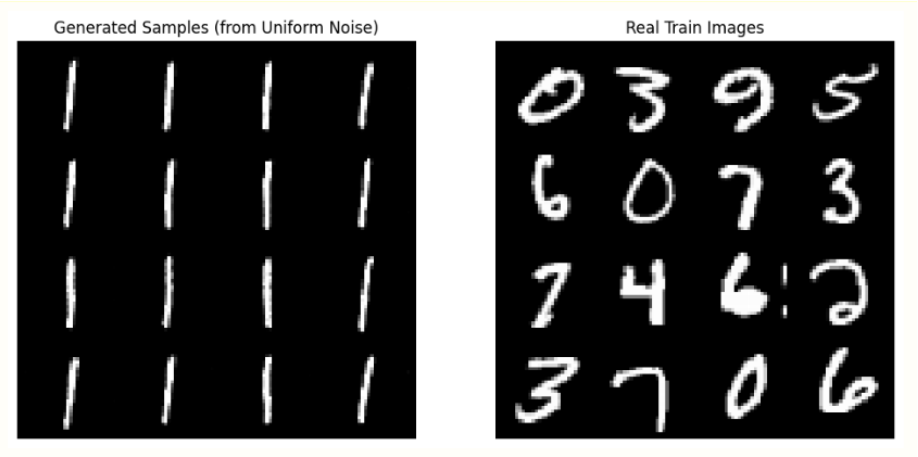
مشاهده می‌شود که با افزایش تعداد گام‌های Langevin از ۲۰ به ۱۰۰، تصاویر خروجی تمایل پیدا می‌کنند تا به "مدهای" (Modes) اصلی توزیع آموخته شده توسط مدل نزدیک شوند. در برخی موارد (مانند ۱۰۰ گام)، تصویر ممکن است تغییر شکل داده و به عدد دیگری که در فضای انرژی مدل دارای کمینه محلی است تبدیل شود (مثلاً تغییر شکل جزئی در عدد ۹). در تعداد گام های زیاد مشاهده میکنیم که تعداد ۱ ها زیاد میشود (این یعنی مدل به عدد ۱ انرژی بسیار کمی تخصیص داده است (احتمالا بخاطر ساختار بسیار ساده آن) و تعداد گام ها وقتی زیاد میشود به این سمت میرود)

۳.۲.۱ تولید تصویر از نویز خالص

در این مرحله، فرآیند تولید تصویر با شروع از نویز تصادفی در بازه و طی کردن ۶۰ گام نمونه‌برداری Langevin انجام شد تا توانایی مدل در تولید داده‌های جدید از فضای نویز بررسی شود. هدف این آزمایش، سنجش میزان موفقیت مدل در شکل‌دهی به نویز و هدایت آن به سمت مدهای با انرژی پایین (ارقام دست‌نویس) است.

تحلیل کیفی و مقایسه با داده‌های واقعی

نتایج حاصل از تولید ۱۶ نمونه تصادفی در کنار ۱۶ تصویر از داده‌های آموزشی در شکل ۵.۱ نمایش داده شده است:



شکل ۵.۱: مقایسه تصاویر تولید شده از نویز خالص (سمت چپ) و تصاویر واقعی مجموعه آموزش (سمت راست)

- **کیفیت تصاویر تولیدی:** تصاویر تولید شده دارای ساختار واضحی هستند و نویز پس‌زمینه به خوبی حذف شده است، که نشان‌دهنده یادگیری نسبی گرادیان‌های تابع انرژی توسط مدل است.
- **عدم تنوع (Mode Collapse):** همان‌طور که در تصویر مشخص است، تمامی ۱۶ نمونه تولید شده شباهت بسیار زیادی به عدد "۱" دارند. این در حالی است که داده‌های واقعی (سمت راست) شامل تنوع بالایی از ارقام مختلف (۰، ۳، ۹، ۵ و غیره) هستند.

علت اصلی عدم مطلوبیت نتایج

اگرچه مدل توانسته است نویز را به یک رقم تبدیل کند، اما در بازسازی توزیع کامل داده‌ها شکست خورده است. علل اصلی این پدیده عبارتند از:

۱. **مشکل اختلاط در MCMC (MCMC Mixing Problem):** نمونه‌برداری Langevin نوعی زنجیره مارکوف است. در مدل‌های EBM، فضای انرژی معمولاً دارای "چاله‌های" عمیق (مدهای محلی) است که توسط سدهای پتانسیل بلند از هم جدا شده‌اند. عبور از نویز خالص و رسیدن به مدهای مختلف (اعداد ۰ تا ۹) دشوار است و زنجیره معمولاً در اولین مدی که به آن نزدیک می‌شود (در اینجا

عدد ۱ که ساختار ساده‌تری دارد) گرفتار می‌شود.

۲. **تعداد اپاک‌های محدود و چالش آموزش:** آموزش مدل‌های مبتنی بر انرژی با استفاده از Contrastive Divergence دشوار است. مدل ممکن است فضای انرژی را به گونه‌ای شکل داده باشد که مدهای مربوط به عدد "۱" دارای حوزه جذب بسیار وسیع‌تری نسبت به سایر ارقام باشند.

۳. **وابستگی به نقطه شروع:** بر خلاف بخش قبل که شروع از تصاویر واقعی بود، شروع از نویز خالص نیازمند گرادیان‌های بسیار دقیق در نواحی دور از داده است. در صورتی که مدل این نواحی را به خوبی یاد نگرفته باشد، نمونه‌ها به سمت ساده‌ترین مد ممکن هدایت می‌شوند.

به طور خلاصه، مدل در یادگیری "ساختار محلی" اعداد موفق بوده اما در بازنمایی "توزیع سراسری" و حفظ تنوع ارقام با چالش جدی روبرو است.

در این زیربخش، توانایی مدل در بازیابی تصاویر اصلی از نسخه‌های نویزی آن‌ها مورد بررسی قرار می‌گیرد. این آزمایش نه تنها قدرت نویززدایی مدل را نشان می‌دهد، بلکه بینش عمیق‌تری نسبت به ساختار "منظره انرژی" (Energy Landscape) که در بخش قبل تحلیل شد، ارائه می‌دهد.

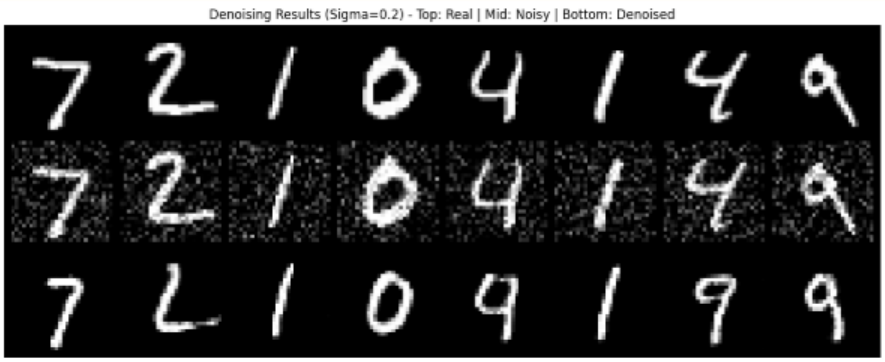
۴.۲.۱ نویززدایی تصویر نویزی

در این مرحله، دسته‌ای از تصاویر مجموعه تست انتخاب شده و با سطوح مختلف نویز گاوسی ترکیب شدند. سپس از فرآیند نمونه‌برداری Langevin برای هدایت این تصاویر نویزی به سمت مناطق با انرژی کمتر (که نشان‌دهنده ارقام معتبر هستند) استفاده شد.

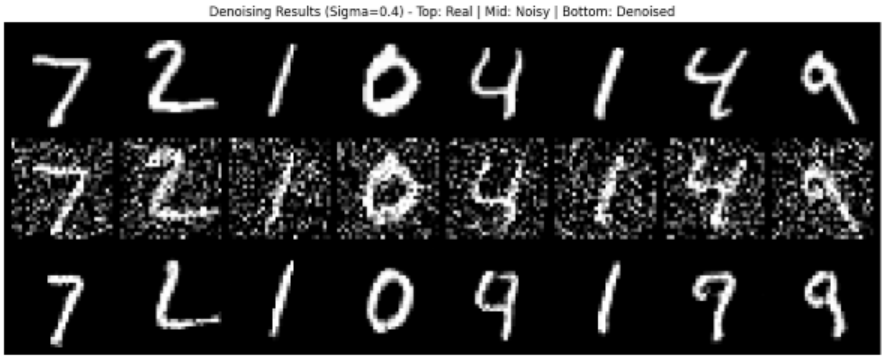
نمایش نتایج نویززدایی

با بررسی نتایج در تعداد گام‌های مختلف، مشخص گردید که ۳۰ گام نمونه‌برداری نقطه تعادل مناسبی میان بازسازی ساختار و حفظ هویت تصویر است. نتایج این بخش در شکل ۶.۱ نمایش داده شده است:

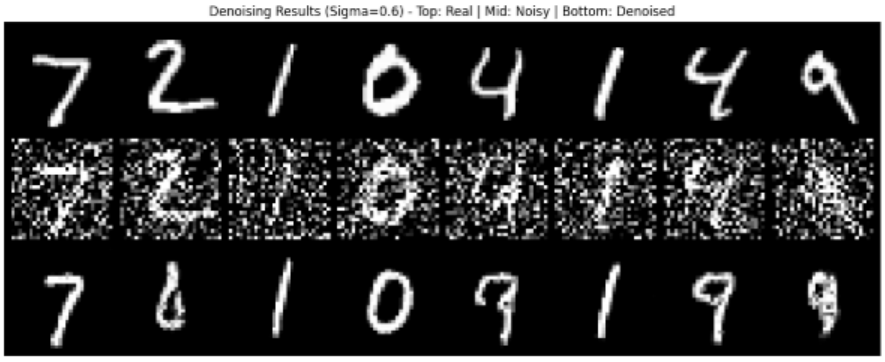
Results for Noise Level (Sigma): 0.2



Results for Noise Level (Sigma): 0.4



Results for Noise Level (Sigma): 0.6



شکل ۶.۱: نتایج نویززدایی در ۳۰ گام برای سطوح نویز 0.2, 0.4, 0.6 (ردیف بالا: واقعی، وسط: نویزی، پایین: بازسازی شده)

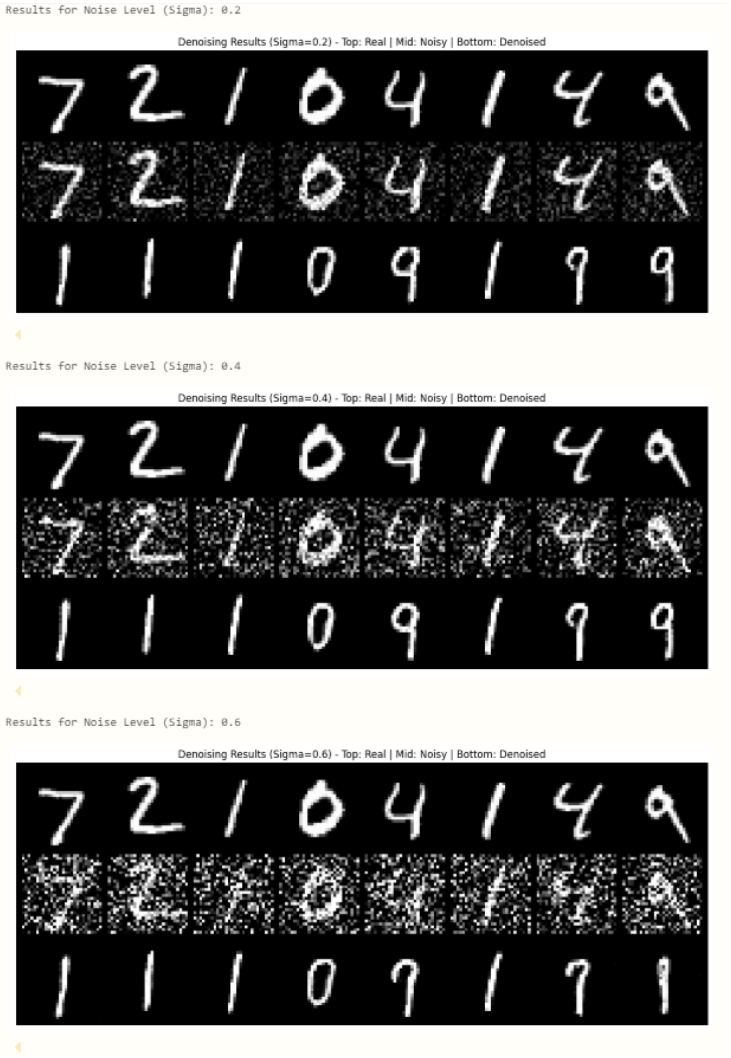
تحلیل عملکرد بر اساس سطح نویز

مطابق با خروجی‌های مشاهده شده، عملکرد مدل را می‌توان به صورت زیر تحلیل کرد:

- **نویز کم:** مدل در این سطح بسیار موفق عمل کرده است. تصاویر بازسازی شده تقریباً با تصاویر اصلی تطابق دارند و لبه‌ها به خوبی تیز شده‌اند.
- **نویز متوسط:** مدل همچنان موفق به حذف نویز است، اما در برخی ارقام (مانند عدد ۲) شاهد تغییرات جزئی در ضخامت یا فرم هستیم. با این حال، هویت عدد حفظ شده است.
- **نویز زیاد:** در این سطح، ساختار اصلی تصویر تا حد زیادی زیر نویز پنهان شده است. اگرچه نویز حذف شده و خروجی شبیه به یک عدد است، اما در برخی موارد هویت عدد تغییر می‌کند.

ارتباط با تحلیل بخش قبل (تأیید فرضیه چاله انرژی)

نتایج نویززدایی، به ویژه در ۱۰۰ گام (شکل ۷.۱)، به طرز جالبی تحلیل بخش سوم (تولید از نویز خالص) را تأیید می‌کند:



شکل ۷.۱: اثر ”تغییر هویت” تصاویر به عدد ۱ در تعداد گام‌های بالا (۱۰۰ گام)

- **غلبه مد عدد ۱:** در تصاویر ۱۰۰ گام مشاهده می‌شود که حتی در نویز کم، مدل تمایل دارد ارقامی مثل ۷ یا ۴ را به عدد ۱ تبدیل کند. این نشان می‌دهد که در منظره انرژی آموخته شده، ”چاله انرژی” مربوط به عدد ۱ بسیار عمیق‌تر و وسیع‌تر از سایر اعداد است.
- **شکست در حفظ هویت در گام‌های بالا:** زمانی که تعداد گام‌های Langevin زیاد می‌شود، تصویر از ”ساختار محلی” خود خارج شده و به سمت ”کمینه مطلق” انرژی حرکت می‌کند که در این مدل، عدد ۱ است.

- **نتیجه‌گیری:** نویززدایی در ۳۰ گام بهترین نتیجه را دارد، زیرا زمان کافی برای حذف نویز را فراهم می‌کند اما هنوز اجازه نداده است که تصویر کاملاً در چاله انرژی عدد ۱ سقوط کند. این موضوع توضیح می‌دهد که چرا در بخش قبل، شروع از نویز خالص (که هیچ ساختار اولیه‌ای ندارد) مستقیماً به تولید عدد ۱ ختم می‌شد.

سوال ۲

Score Based Models

۱.۲ تئوری

۱.۱.۲ استقلال از تابع پارتیشن

تابع امتیاز برای یک توزیع احتمال $p(x)$، به صورت گرادیان لگاریتم تابع چگالی احتمال نسبت به متغیر ورودی x تعریف می‌شود:

$$\mathbf{s}_\theta(x) \coloneqq \nabla_x \log p_\theta(x)$$

فرض کنید یک مدل مبتنی بر انرژی (EBM) با پارامترهای θ به صورت زیر تعریف شده است:

$$p_\theta(x) = \frac{e^{-E_\theta(x)}}{Z(\theta)}$$

که در آن $E_\theta(x)$ تابع انرژی و $Z(\theta)$ تابع پارتیشن (ثابت نرمال‌سازی) است. برای به دست آوردن تابع امتیاز این مدل، ابتدا از طرفین رابطه لگاریتم طبیعی می‌گیریم:

$$\log p_\theta(x) = \log \left(\frac{e^{-E_\theta(x)}}{Z(\theta)} \right)$$

با استفاده از خواص لگاریتم، عبارت فوق به صورت زیر باز می‌شود:

$$\log p_\theta(x) = \log(e^{-E_\theta(x)}) - \log Z(\theta)$$

$$\log p_\theta(x) = -E_\theta(x) - \log Z(\theta)$$

حال عملگر گرادیان ∇_x را بر روی عبارت حاصل اعمال می‌کنیم:

$$\nabla_x \log p_\theta(x) = \nabla_x \left[-E_\theta(x) - \log Z(\theta)\right]$$

با توجه به خطی بودن عملگر گرادیان، داریم:

$$\nabla_x \log p_\theta(x) = -\nabla_x E_\theta(x) - \nabla_x \log Z(\theta)$$

از آنجایی که تابع پارتیشن $Z(\theta) = \int e^{-E_\theta(x)} dx$ تنها تابعی از پارامترهای مدل (θ) است و به متغیر ورودی x بستگی ندارد، مشتق آن نسبت به x برابر صفر خواهد بود:

$$\nabla_x \log Z(\theta) = 0$$

در نتیجه، تابع امتیاز به صورت زیر حاصل می‌شود که نشان‌دهنده استقلال کامل آن از تابع پارتیشن است:

$$\mathbf{s}_\theta(x) = -\nabla_x E_\theta(x)$$

مزیت این ویژگی در آموزش مدل:

محاسبه مستقیم تابع پارتیشن در ابعاد بالا بسیار دشوار است (محاسبات از اوردر نمایی میباشد)؛ این رویکرد نیاز به محاسبه یا تخمین آن را کاملاً حذف می‌کند. در این مدل، به جای تخمین چگالی، مدل بر یادگیری میدان برداری تمرکز میکند که چارچوبی انعطاف پذیر تر برای تولید داده میباشد. تا وقتی که شبکه عصبی مشتق پذیر باشد میتوان برای آن تابع امتیاز تعریف کرد و دیگر نگران فرم بسته توزیع احتمال نباشیم.

۲.۱.۲ چالش‌های Matching Score و روش DSM

۱. **علت دشواری محاسبه ترم $tr(\nabla_x \mathbf{s}_\theta(x))$ در ابعاد بالا:** ترم $tr(\nabla_x \mathbf{s}_\theta(x))$ نشان‌دهنده تراک (اثر) ماتریس ژاکوبین تابع امتیاز نسبت به ورودی است. محاسبه این ترم برای داده‌های با ابعاد بالا (مانند تصاویر)

به دلایل زیر بسیار پرهزینه است:

- **تعداد عملیات پس‌انتشار (Backpropagation):** برای محاسبه دقیق تراک ماتریس ژاکوبین در فضایی با بعد d، نیاز به d بار انجام عملیات پس‌انتشار است. در تصاویر که تعداد پیکسل‌ها (d) بسیار زیاد است، این کار از نظر زمان محاسباتی غیرممکن می‌شود.
- **پیچیدگی زمانی و حافظه:** محاسبه مشتقات مرتبه دوم (هسین) یا مولفه‌های قطری ژاکوبین در شبکه‌های عصبی عمیق، به حافظه و توان پردازشی بسیار بالایی نیاز دارد که باعث می‌شود آموزش مدل روی دیتاست‌های واقعی مقیاس‌پذیر نباشد.

۲. ایده اصلی روش Denoising Score Matching (DSM): ایده اصلی روش DSM این است که به جای تلاش برای تخمین مستقیم تابع امتیاز توزیع داده‌های اصلی (p_{data})، ابتدا به داده‌ها یک نویز (معمولاً گاوسی) اضافه کرده و سپس تابع امتیاز توزیع نویزی حاصل (q_σ) را یاد بگیریم. در واقع مدل آموزش می‌بیند تا جهت حرکت از یک نمونه نویزی $\tilde{x}$ به سمت داده سالم x را تخمین بزند (فرآیند نویززدایی). این کار نیاز به محاسبه ترم دشوار Trace را کاملاً حذف می‌کند.

۳. منطق برابری هدف DSM با Score Matching اصلی:

هدف اصلی در Score Matching کمینه کردن فاصله بین مدل امتیاز و تابع امتیاز واقعی توزیع نویزی است:

$$\mathcal{L}_{SM} = \mathbb{E}_{q_\sigma(\tilde{x})}[\|\mathbf{s}_\theta(\tilde{x}) - \nabla_{\tilde{x}} \log q_\sigma(\tilde{x})\|^2]$$

اثبات می‌شود که تحت شرایط ملایم، این تابع هدف با تابع هدف زیر (که همان DSM است) معادل است:

$$\mathcal{L}_{DSM} = \mathbb{E}_{p_{data}(x)}\mathbb{E}_{q_\sigma(\tilde{x}|x)}[\|\mathbf{s}_\theta(\tilde{x}) - \nabla_{\tilde{x}} \log q_\sigma(\tilde{x}|x)\|^2]$$

محاسبه‌پذیری گرادیان شرطی: اگر از نویز گاوسی $q_\sigma(\tilde{x}|x) = \mathcal{N}(\tilde{x}; x, \sigma^2 I)$ استفاده کنیم، تابع امتیاز شرطی به سادگی به دست می‌آید:

$$\nabla_{\tilde{x}} \log q_\sigma(\tilde{x}|x) = -\frac{\tilde{x} - x}{\sigma^2}$$

این ترم بر خلاف تابع امتیاز کل، به راحتی قابل محاسبه است.

ارتباط با توزیع کلی: طبق قواعد آماری و انتگرال‌گیری جزء به جزء، ثابت می‌شود که امید ریاضی گرادیان لگاریتم توزیع شرطی ($\nabla_{\tilde{x}} \log q_\sigma(\tilde{x}|x)$) با گرادیان لگاریتم توزیع حاشیه‌ای نویزی ($\nabla_{\tilde{x}} \log q_\sigma(\tilde{x})$) برابر است. (اثبات این نیز در پایین آورده شده است.) بنابراین، وقتی مدل را آموزش می‌دهیم تا بردار نویز را در سطح نمونه‌ها تخمین بزند، به طور غیرمستقیم در حال یادگیری تابع امتیاز توزیع نویزی شده هستیم. این روش به دلیل حذف مشتقات مرتبه دوم، بسیار سریع‌تر و پایدارتر از روش اصلی است.

اثبات هم‌ارزی Score Matching و DSM

$$q_\sigma(\tilde{x}) = \int q_\sigma(\tilde{x}|x) p_{data}(x) dx$$

تابع هدف اصلی بر روی توزیع نویزی به صورت زیر است:

$$\mathcal{L}_{SM}(\theta) = \mathbb{E}_{q_\sigma(\tilde{x})} \left[\frac{1}{2} \|\mathbf{s}_\theta(\tilde{x})\|^2 - \mathbf{s}_\theta(\tilde{x}) \cdot \nabla_{\tilde{x}} \log q_\sigma(\tilde{x}) \right] + C$$

که در آن C ترمی ثابت و مستقل از θ است.

تابع هدف DSM به صورت زیر تعریف می‌شود:

$$\mathcal{L}_{DSM}(\theta) = \mathbb{E}_{p_{data}(x)} \mathbb{E}_{q_\sigma(\tilde{x}|x)} \left[\frac{1}{2} \|\mathbf{s}_\theta(\tilde{x})\|^2 - \mathbf{s}_\theta(\tilde{x}) \cdot \nabla_{\tilde{x}} \log q_\sigma(\tilde{x}|x) \right] + C'$$

$$\mathbb{E}_{q_\sigma(\tilde{x})}[\mathbf{s}_\theta(\tilde{x}) \cdot \nabla_{\tilde{x}} \log q_\sigma(\tilde{x})] = \int q_\sigma(\tilde{x}) \mathbf{s}_\theta(\tilde{x}) \cdot \frac{\nabla_{\tilde{x}} q_\sigma(\tilde{x})}{q_\sigma(\tilde{x})} d\tilde{x} = \int \mathbf{s}_\theta(\tilde{x}) \cdot \nabla_{\tilde{x}} q_\sigma(\tilde{x}) d\tilde{x}$$

با جایگذاری تعریف $q_\sigma(\tilde{x})$ از گام اول:

$$\int \mathbf{s}_\theta(\tilde{x}) \cdot \nabla_{\tilde{x}} \left(\int q_\sigma(\tilde{x}|x) p_{data}(x) dx \right) d\tilde{x}$$

با جابجا کردن عملگر گرادیان و انتگرال (تحت شرایط پایداری):

$$\int \mathbf{s}_\theta(\tilde{x}) \cdot \left(\int p_{data}(x) \nabla_{\tilde{x}} q_\sigma(\tilde{x}|x) dx \right) d\tilde{x}$$

با بازنویسی عبارت داخل پرانتز بر حسب امتیاز شرطی ($\nabla_{\tilde{x}} q_\sigma = q_\sigma \nabla_{\tilde{x}} \log q_\sigma$):

$$\iint p_{data}(x) q_\sigma(\tilde{x}|x) [\mathbf{s}_\theta(\tilde{x}) \cdot \nabla_{\tilde{x}} \log q_\sigma(\tilde{x}|x)] dx d\tilde{x}$$

که این دقیقاً برابر است با:

$$\mathbb{E}_{p_{data}(x)} \mathbb{E}_{q_\sigma(\tilde{x}|x)} [\mathbf{s}_\theta(\tilde{x}) \cdot \nabla_{\tilde{x}} \log q_\sigma(\tilde{x}|x)]$$

از آنجایی که ترم اول هر دو تابع هدف ($\frac{1}{2}\|\mathbf{s}_\theta\|^2$) و ترم‌های ضرب داخلی آن‌ها با هم برابر هستند، مینیمم کردن $\mathcal{L}_{DSM}$ نسبت به θ دقیقاً معادل با مینیمم کردن $\mathcal{L}_{SM}$ روی توزیع نویزی است. تفاوت اصلی در این است که در DSM ما نیازی به دانستن $q_\sigma(\tilde{x})$ (که مجهول است) نداریم و فقط از نویز شرطی $q_\sigma(\tilde{x}|x)$ (که خودمان اضافه کرده‌ایم و امتیاز آن معلوم است) استفاده می‌کنیم.

۳.۱.۲ چالش تکیه گاه های مجزا

الف) اثبات عدم وابستگی تابع امتیاز به ضریب مخلوط (π): فرض کنید توزیع داده‌ها ترکیبی از دو توزیع گوسی با وزن‌های متفاوت باشد:

$$p_{data}(x) = \pi p_1(x) + (1 - \pi) p_2(x)$$

طبق فرض سوال، تکیه‌گاه‌های این دو توزیع (A و B) کاملاً از هم جدا (Disjoint) هستند. این بدان معناست که:

- اگر $x \in A$ باشد، آنگاه $p_1(x) > 0$ و $p_2(x) = 0$ است.
- اگر $x \in B$ باشد، آنگاه $p_2(x) > 0$ و $p_1(x) = 0$ است.

حال تابع امتیاز واقعی ($\nabla_x \log p_{data}(x)$) را برای ناحیه A محاسبه می‌کنیم:

$$\nabla_x \log p_{data}(x) = \nabla_x \log(\pi p_1(x) + (1-\pi)p_2(x))$$

با جایگذاری $p_2(x) = 0$ برای $x \in A$:

$$\nabla_x \log p_{data}(x) = \nabla_x \log(\pi p_1(x))$$

با استفاده از ویژگی لگاریتم حاصل‌ضرب ($\log(ab) = \log a + \log b$):

$$\nabla_x \log p_{data}(x) = \nabla_x (\log \pi + \log p_1(x))$$

از آنجایی که عملگر گرادیان نسبت به x اعمال می‌شود و π یک مقدار ثابت (مستقل از x) است، داریم $\nabla_x \log \pi = 0$. در نتیجه:

$$\nabla_x \log p_{data}(x) = \nabla_x \log p_1(x)$$

به طور مشابه برای ناحیه B حاصل می‌شود:

$$\nabla_x \log p_{data}(x) = \nabla_x \log p_2(x)$$

این روابط ثابت می‌کنند که در هر ناحیه، تابع امتیاز تنها به شکل توزیع محلی بستگی دارد و هیچ وابستگی به ضرایب مخلوط (π و $1-\pi$) ندارد.

ب) علت مشکل در نمونه‌برداری Langevin Dynamics: الگوریتم نمونه‌برداری Langevin Dynamics برای به‌روزرسانی نمونه‌ها تنها از تابع امتیاز استفاده می‌کند:

- **فقدان اطلاعات جرم کلی:** همان‌طور که در بخش قبل اثبات شد، تابع امتیاز در هر نقطه فقط حاوی اطلاعات محلی درباره جهت افزایش چگالی است و هیچ اطلاعی از وزن کلی (جرم احتمالی) آن ناحیه نسبت به نواحی دیگر ندارد.
- **مشکل تکیه‌گاه‌های جدا از هم:** وقتی تکیه‌گاه‌ها کاملاً جدا باشند، در فضای بین A و B چگالی احتمال صفر است. در این حالت تابع امتیاز (گرادیان) وجود ندارد تا زنجیره را از یک ناحیه به ناحیه دیگر هدایت کند.

- **عدم رعایت نسبت وزن‌ها:** به دلیل اینکه نرخ انتقال بین دو ناحیه عملاً صفر است، زنجیره مارکوف در هر ناحیه‌ای که شروع شده باشد محبوس می‌شود. در نتیجه، نسبت نمونه‌های تولیدی از هر عدد به جای اینکه تابع π باشد، تابع توزیع اولیه نمونه‌ها خواهد بود و مدل نمی‌تواند نسبت‌های صحیح وزن‌ها را رعایت کند.

۴.۱.۲ چالش‌های داده‌های واقعی و راهکار NCSN

الف) مشکلات اعمال مستقیم Score Matching روی داده‌های بدون نویز: اعمال مستقیم روش‌های تطبیق امتیاز روی داده‌های واقعی معمولاً با دو چالش اساسی روبرو می‌شود:

- **Manifold Hypothesis:** داده‌های واقعی (مانند تصاویر) معمولاً در زیرفضاهایی با ابعاد بسیار پایین‌تر نسبت به فضای اصلی ورودی قرار دارند. در نتیجه، تابع امتیاز در اکثر نقاط فضای اصلی که داده‌ای در آن‌ها وجود ندارد، تعریف نشده است یا تخمین آن بسیار ناپایدار است. این موضوع باعث می‌شود مدل در نواحی دور از داده‌ها (Low-density regions) عملکرد درستی نداشته باشد.
- **مشکل داده‌های کم‌تراکم:** در مناطقی که چگالی داده‌ها بسیار کم است، تخمین گرادیان لگاریتم چگالی با خطای بسیار زیادی همراه است. این خطا باعث می‌شود در هنگام نمونه‌برداری، زنجیره Langevin به جای حرکت به سمت مودهای اصلی داده، به سمت نواحی بی‌استفاده فضای ورودی منحرف شود.

ب) حل مشکلات با NCSN و Annealed Langevin Dynamics: مقاله NCSN برای حل مشکلات فوق، از دو استراتژی کلیدی استفاده می‌کند:

- **Multi-scale Noise Perturbation:** به جای اضافه کردن یک سطح ثابت از نویز، داده‌ها با سطوح مختلفی از نویز $\sigma_1 > \sigma_2 > \cdots > \sigma_L$ مشوش می‌شوند. نویزهای بزرگتر باعث می‌شوند که تمام فضای ورودی پوشش داده شده و تابع امتیاز در همه جا به خوبی تعریف شود.
- **الگوریتم Annealed Langevin Dynamics:** در این روش، نمونه‌برداری ابتدا از بزرگترین سطح نویز (σ_1) شروع می‌شود که در آن تابع امتیاز بسیار روان و ساده است. سپس به تدریج سطح نویز کاهش می‌یابد تا به کوچکترین مقدار (σ_L) برسد. این فرآیند به مدل اجازه می‌دهد ابتدا ساختار کلی داده را پیدا کرده و سپس جزئیات دقیق را بازسازی کند.

ج) نقش بزرگترین و کوچکترین سطح نویز در کیفیت نمونه‌برداری:

- **بزرگترین نویز (σ_{max}):** نقش اصلی آن غلبه بر مشکل تکیه‌گاه‌های جدا از هم است. نویز بزرگ

باعث می‌شود شکاف‌های بین مودهای مختلف داده پر شود و زنجیره نمونه‌برداری بتواند آزادانه بین مودها حرکت کرده و توزیع کلی را به درستی یاد بگیرد.

- **کوچکترین نویز** (σ_{min}): این سطح از نویز باید به قدری کوچک باشد که تصویر نهایی تولید شده، تفاوت ناچیزی با داده‌های واقعی بدون نویز داشته باشد. در واقع σ_{min} مسئول حفظ کیفیت بصری، وضوح لبه‌ها و جزئیات دقیق در تصویر نهایی است.

۲.۲ عملی

۱.۲.۲ پیاده‌سازی مدل پایه

در این مرحله، هدف آموزش یک مدل برای یادگیری توزیع داده‌های MNIST با استفاده از تکنیک Annealed Langevin Dynamics است.

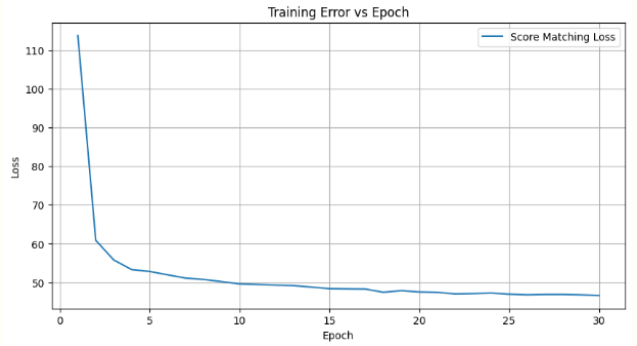
برای غلبه بر مشکل داده‌های کم‌تراکم در فضای بالا-بعد، از چندین سطح نویز (σ) استفاده می‌کنیم. مدل طراحی شده (NCSN) یک شبکه عصبی است که علاوه بر تصویر نویزی، مقدار σ را نیز به عنوان ورودی دریافت می‌کند.

مراحل کلیدی در این پیاده‌سازی عبارتند از:

- **دنباله نویز:** یک دنباله هندسی از σها از مقدار بزرگ ($\sigma_{max} = 30$) تا مقدار کوچک ($\sigma_{min} = 0.01$) تعریف شده است. نویز بزرگ باعث می‌شود مدل ساختار کلی داده را یاد بگیرد و نویز کوچک به یادگیری جزئیات کمک می‌کند.
- **معماری شبکه:** از یک ساختار U-Net بهبود یافته استفاده شده است که دارای لایه‌های *Adaptive ResBlock* است. این لایه‌ها اجازه می‌دهند اطلاعات مربوط به سطح نویز به طور موثر در تمام بخش‌های شبکه تزریق شود.
- **تابع هزینه:** از تابع هزینه *Denoising Score Matching* استفاده شده است. هدف شبکه کمینه کردن تفاوت بین خروجی مدل و نویز اضافه شده (با مقیاس‌بندی مناسب) است:

$$\mathcal{L} = \frac{1}{L}\sum_{i=1}^{L} \sigma_i^2 \mathbb{E}_{p_{data}(\mathbf{x})} \mathbb{E}_{p_{\sigma_i}(\tilde{\mathbf{x}}|\mathbf{x})} [\|\mathbf{s}_\theta(\tilde{\mathbf{x}}, \sigma_i) + \frac{\tilde{\mathbf{x}} - \mathbf{x}}{\sigma_i^2}\|_2^2]$$

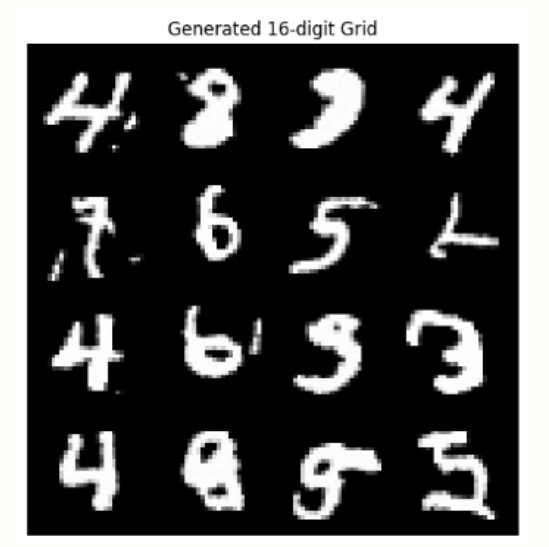
پایداری آموزش و نمودار تابع هزینه: نمودار لوس (شکل ۱.۲) روند یادگیری شبکه را در طول ۳۰ اپاک نشان می‌دهد. نزولی بودن نمودار نشان‌دهنده این است که شبکه U-Net به خوبی توانسته نگاشت بین تصاویر نویزی و بردارهای امتیاز (Score Vectors) را یاد بگیرد. نوسانات جزئی در نمودار به دلیل ماهیت تصادفی انتخاب سطح نویز در هر Batch است؛ اما روند کلی حاکی از آن است که شبکه در تخمین نویز برای هر دو حالت نویز شدید و نویز ضعیف به مهارت رسیده است.



شکل ۱.۲: نمودار کاهش تابع هزینه؛ پایداری در آموزش مدل SBM مشهود است.

تصاویر تولید شده: خروجی نهایی حاصل از الگوریتم Annealed Langevin Dynamics در شکل ۲.۲ نمایش داده شده است.

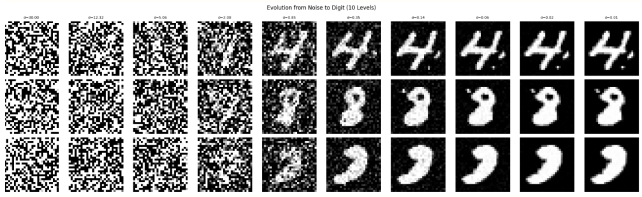
تصاویر تولید شده دارای وضوح بسیار بالا و ساختار کاملاً معنادار هستند. عدم وجود لکه‌های نویز در پس‌زمینه نشان‌دهنده دقت بالای شبکه در سطوح نویز پایین است. همچنین تنوع ارقام تولید شده نشان می‌دهد که مدل دچار مشکل Mode Collapse نشده و توانسته است چگالی احتمالی تمامی اعداد (۰ تا ۹) را به خوبی مدل‌سازی کند.



شکل ۲.۲: گرید تصاویر تولید شده؛ وضوح لبه‌ها و تنوع اعداد تاییدکننده عملکرد صحیح مدل پایه است.

دنباله نویززدایی: نتایج روند تبدیل شدن نویز خالص به عکسای با معنا در شکل ۳.۲ آورده شده است.

- **نویز بالا (σ از 30.00 تا 5.06):** تصویر از توزیع اولیه نویز است و هیچ شباهتی به عدد ندارد.
- **مقادیر میانی (σ از 2.08 تا 0.85):** این مرحله بحرانی است که در آن *ساختارکلی* عدد شکل می‌گیرد. مدل محدوده کلی داده را تشخیص می‌دهد.
- **نویز پایین (σ از 0.35 تا 0.01):** مدل عملیات *نویززدایی* و بهبود جزئیات را انجام می‌دهد تا تصویر نهایی شفاف و دقیق شود.



شکل ۳.۲: روند تبدیل شدن نویز خالص به عدد نهایی طی مراحل کاهش نویز

۲.۲.۲ پیاده‌سازی مدل شرطی

در این مرحله، هدف ارتقای مدل به یک ساختار شرطی ($p(\mathbf{x}|y)$) است تا بتوانیم فرآیند تولید تصویر را هدایت کنیم؛ برای مثال از مدل بخواهیم صرفاً عدد خاصی (مانند رقم ۵) را تولید کند این رویکرد به جای یادگیری توزیع کلی، بر یادگیری توزیع به شرط لیبل تمرکز دارد که قابلیت کنترل مدل را به شدت افزایش می‌دهد.

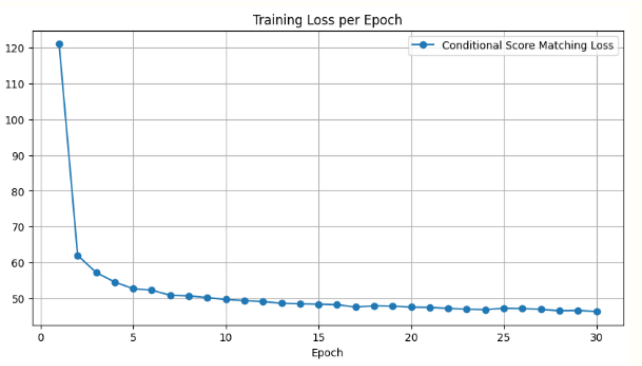
تغییرات کلیدی نسبت به مدل پایه عبارتند از:

- **لایه امبدینگ:** یک لایه nn.Embedding برای ۱۰ کلاس ارقام (۰ تا ۹) به مدل اضافه شده است تا لیبل‌های گسسته به فضایی پیوسته و قابل یادگیری توسط شبکه تبدیل شوند.
- **ترکیب اطلاعات:** در ورودی بلوک‌های *Adaptive ResBlock*، بردار امبدینگ نویز و بردار امبدینگ کلاس با یکدیگر ترکیب می‌شوند تا شبکه در هر مرحله از بازسازی، هم از سطح نویز فعلی و هم از نوع عدد درخواستی آگاه باشد.
- **مکانیزم FiLM:** برای تزریق موثر اطلاعات شرطی، از لایه‌های خطی برای تبدیل امبدینگ به پارامترهای *Scale* و *Shift* استفاده شده است که خروجی نرمال‌سازی را در بلوک‌های شبکه تغییر می‌دهند.

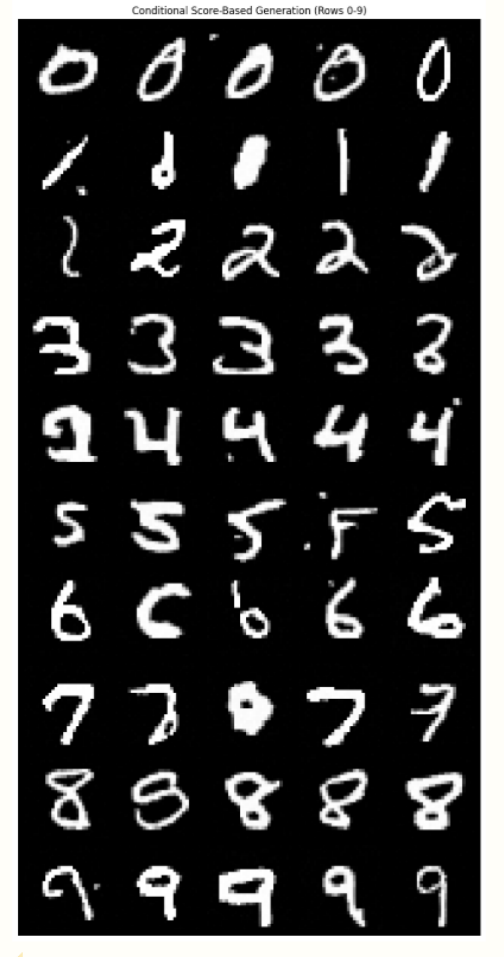
تحلیل روند آموزش و نمودار خطا: نمودار مربوط به مدل شرطی در شکل ۴.۲ رسم شده است. مشاهده می‌شود که خطا با سرعت بسیار بالایی در اپاک‌های ابتدایی کاهش یافته (از حدود ۱۲۰ به زیر ۶۰) و سپس به یک پایداری مطلوب در حدود مقدار ۴۶ رسیده است. این ثبات نشان می‌دهد که اضافه کردن شرط کلاس نه تنها باعث پیچیدگی مخرب در آموزش نشده، بلکه به دلیل جهت‌دهی به فضای جستجو، به مدل کمک کرده است تا سریع‌تر به میدان گرادیان هر کلاس مسلط شود.

تصاویر تولید شده: خروجی نهایی مدل در قالب یک گرید که در آن هر ردیف مختص یک عدد خاص است (از ردیف بالا عدد ۰ تا ردیف پایین عدد ۹)، در شکل ۵.۲ نمایش داده شده است.

دقت بالای مدل در رعایت شرط ورودی بسیار قابل توجه است؛ در تمامی ردیف‌ها، اعداد دقیقاً مطابق با لیبل درخواستی تولید شده‌اند که نشان‌دهنده عملکرد صحیح امبدینگ کلاس و لایه‌های *FiLM* در هدایت الگوریتم Langevin است. علاوه بر دقت، تنوع استایل‌های نوشتاری در هر ردیف نشان می‌دهد که مدل هنوز قدرت تولیدی خود را حفظ کرده و برای هر کلاس، کل فضای احتمالی آن رقم را پوشش می‌دهد.



شکل ۴.۲: نمودار کاهش تابع هزینه مدل شرطی.



شکل ۵.۲: گرید اعداد تولید شده به صورت شرطی؛ هر ردیف مربوط به یک کلاس خاص است که نشان‌دهنده کنترل کامل بر فرآیند تولید است.

مقایسه مدل پایه و مدل شرطی در این بخش به بررسی و مقایسه عملکرد مدل پایه (یادگیری توزیع کلی $p(\mathbf{x})$) و مدل شرطی (یادگیری توزیع مشروط $p(\mathbf{x}|y)$) پرداخته می‌شود. این مقایسه بر اساس سه محور ساختار، پایداری آموزش و قابلیت تولید انجام می‌گیرد.

۱. **تفاوت‌های ساختاری و مفهومی:** مدل پایه تنها بر یادگیری میدان گرادیان چگالی احتمال کل داده‌ها تمرکز داشت. در مقابل، مدل شرطی با بهره‌گیری از لایه $nn.Embedding$ و مکانیزم $FiLM$، توانایی یادگیری ویژگی‌های منحصربه‌فرد هر کلاس را پیدا کرده است. این تغییر ساختاری باعث شده است که مدل از یک مولد تصادفی به یک ابزار تولید هدفمند تبدیل شود

۲. **تحلیل روند همگرایی:** با مقایسه نمودار خطای مدل پایه (شکل ۱.۲) و مدل شرطی (شکل ۴.۲)، نکات زیر استخراج می‌شود:

- **مقدار نهایی خطا:** هر دو مدل پس از ۳۰ اپاک در محدوده خطای ۴۶ تثبیت شده‌اند. این موضوع نشان می‌دهد که گنجاندن اطلاعات شرطی، پیچیدگی مخربی به فرآیند بهینه‌سازی اضافه نکرده است.
- **سرعت همگرایی:** مدل شرطی به دلیل داشتن جهت‌دهی مشخص برای هر دسته از داده‌ها، در اپاک‌های اولیه افت خطای تندتری را نسبت به مدل پایه تجربه کرده است. (در مدل پایه در ۵ اپاک اول از ۱۱۳ به ۵۲ افت کرده است اما در مدل شرطی از ۱۲۱ به ۵۲)

۳. **ارزیابی خروجی‌های بصری:** تفاوت اصلی در تصاویر تولید شده مشهود است:

- **مدل پایه:** ارقامی با کیفیت بالا اما به صورت تصادفی تولید می‌کند که در آن توزیع ارقام در گرید خروجی قابل پیش‌بینی نیست.
- **مدل شرطی:** علاوه بر حفظ کیفیت و شفافیت لبه‌ها، تفکیک کامل کلاس‌ها را در هر ردیف نمایش می‌دهد. تنوع ارقام (*Diversity*) در هر ردیف نشان‌دهنده این است که شرطی‌سازی باعث محدود شدن مدل به چند نمونه خاص (Mode Collapse) نشده است.

جدول ۱.۲: مقایسه ویژگی‌های کلیدی مدل پایه و مدل شرطی

ویژگی	مدل پایه	مدل شرطی
توزیع هدف	$p(\mathbf{x})$	$p(\mathbf{x}\|y)$
کنترل بر خروجی	ندارد (تصادفی)	دارد (بر اساس لیبل)
مکانیزم تزریق اطلاعات	فقط نویز (σ)	نویز (σ) + کلاس (y)
پایداری در آموزش	بسیار بالا	بسیار بالا

تحلیل نهایی نشان می‌دهد که مدل $NCSN$ به خوبی پتانسیل ارتقا به مدل‌های شرطی را دارد. مدل شرطی پیاده‌سازی شده، ضمن حفظ تمامی مزایای مدل پایه از جمله وضوح بالا و پایداری در آموزش، محدودیت اصلی آن یعنی عدم کنترل بر نوع رقم تولیدی را کاملاً مرتفع ساخته است.