نام مدرس: دکتر محمدرضا محمدی **دستیاران آموزشی مرتبط:** آرش فرزانه نژاد، آیسا میاهی‌نیا، آیدا خالقی، امیرحسین حسینی جبلی **مهلت تحویل (بدون کسر نمره):** ۱۰ آذر ماه	**تمرین سری سوّم** درس یادگیری عمیق	دانشگاه علم و صنعت ایران

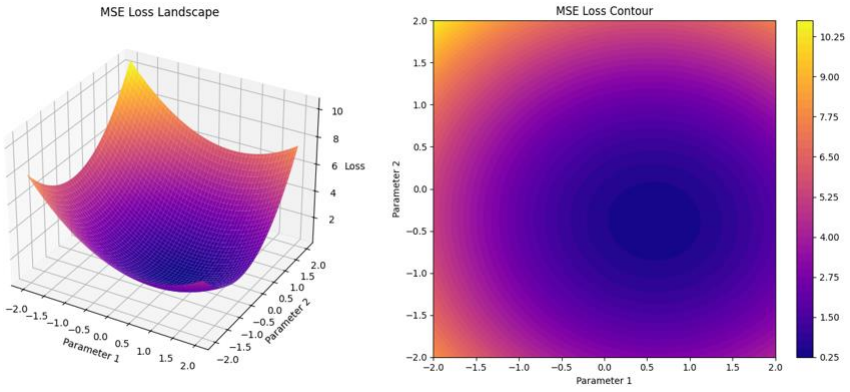
سوالات تئوری (۱۰+۵۰ نمره)

سوال ۱ (۱۵ نمره)

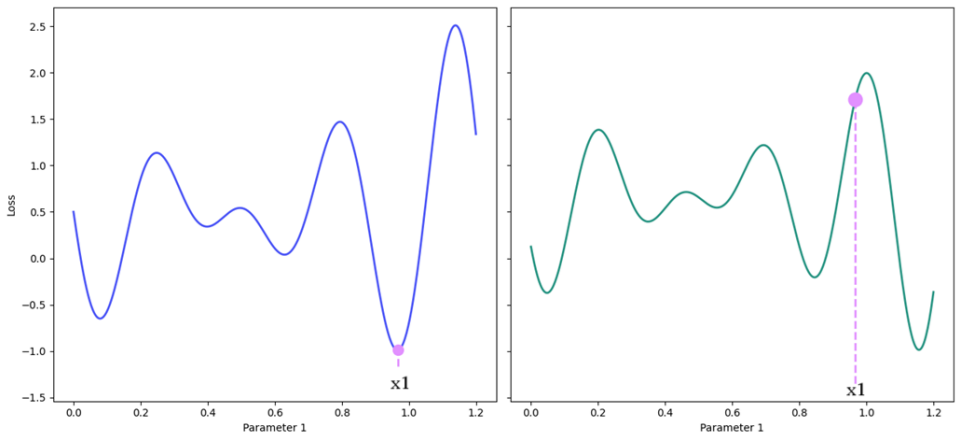
در این سوال به بررسی Loss Landscape میپردازیم. جستجو و مطالعه‌ای در مورد آن انجام دهید.

الف) به طور کامل شرح دهید که Loss Landscape چیست، بُعدهای آن و هر نقطه در آن نماینده‌ی چیست و ما در فرایند یادگیری مدل چطوری آن را تحلیل میکنیم؟

ب) هر دو تصویر زیر مرتبط با MSE Loss یک مساله‌ی یکسان (با ۲ پارامتر) هستند. ارتباط این دو شکل چیست و هر کدام را چطور تفسیر میکنید؟ در شکل سمت راست دایره‌هایی که مشخص هستند چه معنایی دارند و نقاطی که روی یک دایره مشترک قرار میگیرند چه ویژگی یکسانی دارند؟

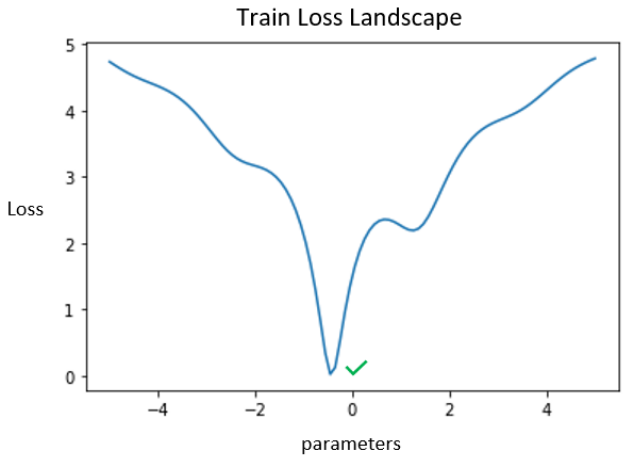


ج) ۲ شکل زیر هر دو Training Loss Landscape هایی در یک مساله‌ی یادگیری ماشین یکسان (با ۱ پارامتر) هستند. چطور یکسان نبودن دو شکل را توجیه میکنید؟



د) آنطور که در شکل بالا مشخص است، نقطه‌ی x_1 ای که یافت شده، در شکل سمت چپ نقطه‌ی بهینه میباشد اما همان نقطه در شکل سمت راست مقدار Loss متفاوت و زیادی دارد. این را چطور تفسیر میکنید؟ و آیا میتوان گفت optimizer یکسانی که استفاده شده به طور کلی ضعیف است؟

ه) فرض کنید در مساله‌ای optimizer ما توانسته به هنگام train به نقطه‌ی بهینه‌ی سراسری (مجموعه پارامترهایی که به کمترین Loss منجر میشود) برسد.



با این حال مدل روی داده‌ی تست عملکرد خوبی ندارد. علت چیست؟ مگر به نقطه‌ی بهینه‌ی سراسری دست پیدا نکرده ایم؟ با توجه به مطالبی که تا کنون آموخته‌اید چه راه‌حل‌هایی برای این مشکل پیشنهاد میکنید؟

سوال ۲ (۱۲ نمره)

فرض کنید یک شبکه عصبی با ۲ ورودی و ۱ خروجی داریم. رابطه بین ورودی و خروجی به شکل زیر است (پارامترهای w_1, w_2, w_3, b قابل آموزش‌اند):

$$y = w_1 x_1^2 + w_2 x_2^2 + w_3 x_1 x_2 + b$$

اگر داده‌های آموزشی مطابق جدول زیر باشند و از نقطه‌ی اولیه: $b = +1$ ، $w_3 = -1$ ، $w_2 = -1$ ، $w_1 = +1$ شروع کنیم، نتیجه حاصل را برای یک دوره آموزش (epoch = 1) با بهینه‌ساز SGD + Momentum و SGD محاسبه کنید. فرض کنید از تابع میانگین مربعات خطا (MSE) استفاده می‌شود. لطفاً مراحل محاسبات را نشان دهید و سپس هر دو روش را مقایسه کنید (دو نمونه اول را اولین Batch و دو نمونه دیگری را به عنوان دومین Batch در نظر بگیرید).

$$learning\_rate = 0.1 \quad beta = 0.9 \quad batch\_size = 2$$

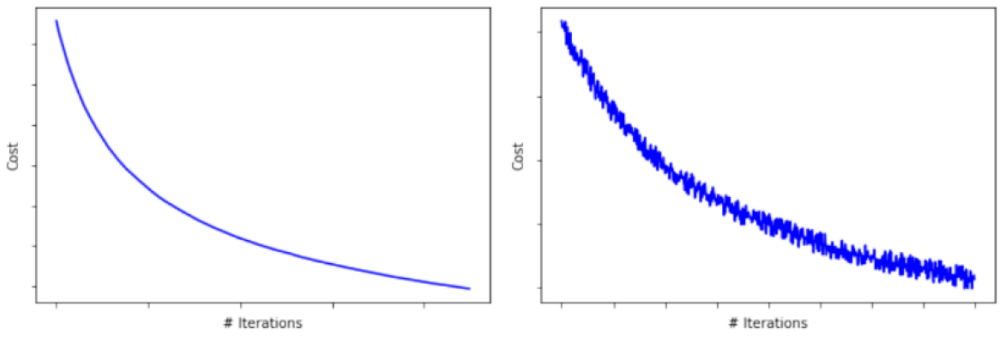
$$MSE(y, \hat{y}) = \left(\frac{1}{n}\right) \sum_{i=1}^{n} (y_i - \hat{y}_i)^2$$

x_1	x_2	y
۱	-۱	۱۰
۲	۰	۱۳
۰	۲	۱۱
-۱	۱	۴

سوال ۳ (۸ نمره)

به پرسش‌های زیر درباره بهینه‌سازها پاسخ دهید.

- استفاده از نرخ یادگیری بسیار بالا چه مشکلاتی ایجاد می‌کند؟ چگونه می‌توان این مشکلات را تشخیص داد؟
- استفاده از نرخ یادگیری بسیار پایین چه مشکلاتی ایجاد می‌کند؟ چگونه می‌توان این مشکلات را تشخیص داد؟
- «نقطه زینی» چیست؟ دو الگوریتم SGD و Adam را در برخورد با این نقاط مقایسه کنید و مزایا/معایب هرکدام را بنویسید.
- شکل‌های زیر کاهش خطا را بر حسب تعداد تکرارها نشان می‌دهند هنگامی که از دو الگوریتم بهینه‌سازی متفاوت برای آموزش استفاده شده است. مشخص کنید کدام نمودار مربوط به Batch Gradient Descent و کدام مربوط به Mini-Batch Gradient Descent است و دلیل بیاورید (منظور از Cost در لیبل نمودارها همان خطا هستش).



سوال ۴ (۱۰ نمره ـ اختیاری)

نشان دهید در گرادیان نزولی با به‌روزرسانی به شکل:

$$w^{t+1} = w^t - \eta \nabla J(w^t) \quad \eta > 0$$

برای η کوچک داریم:

$$J(w^{t+1}) \leq J(w^t)$$

(راهنمایی : معادل این است که عبارت روبه‌رو برقرار باشد: $(f(x + \Delta x) \leq f(x),\ \Delta x = -\eta \nabla f(x)$

سوال ۵ (۱۵ نمره)

در سال ۲۰۱۵، بهینه‌ساز Adam در مقاله‌ای از Diederik P. Kingma (از OpenAI) معرفی شد و امروزه همچنان یکی از پراستفاده‌ترین بهینه‌سازها در یادگیری عمیق میباشد.

ابتدا مقاله‌ی آن را بررسی کنید و سپس به سوالات مربوطه پاسخ دهید:

ADAM: A METHOD FOR STOCHASTIC OPTIMIZATION

1. ایده اصلی:

- توضیح دهید که Adam قصد حل چه مشکلی را دارد. چرا روش‌های سنتی گرادیان کاهشی تصادفی (SGD) همیشه برای شبکه‌های عصبی کارآمد نیستند؟
- توضیح دهید که دو تخمین مومان در Adam (m_t و v_t) به چه معنا هستند و چرا در نظر گرفتن هر دو کمک‌کننده است؟
- قانون به‌روزرسانی پارامترهای Adam را که در مقاله بررسی شده (الگوریتم ۱) بنویسید و هر عبارت را به طور خلاصه توضیح دهید.

2. اصلاح بایاس و پایداری:

مقاله عبارت‌های اصلاح بایاس یعنی $\hat{m}_t$ و $\hat{v}_t$ را معرفی می‌کند.

- چرا m_t و v_t نیاز به اصلاح بایاس دارند؟
- چرا این اصلاح ضروری است و اگر در ابتدای آموزش از آن‌ها استفاده نشود چه اتفاقی میفتد؟
- همگرایی یک مدل که با Adam بدون تصحیح بایاس بهینه‌سازی شود چطور با SGD قابل مقایسه است؟

3. مقایسه با سایر بهینه‌سازها:

- به صورت خلاصه Adam را با موارد زیر مقایسه کنید (همچنین برای هر کدام توضیح دهید که Adam چه ایده‌ای از آن روش را مورد استفاده قرار داده و چه مواردی را در آن‌ها بهبود داده است):

a) SGD + Momentum

b) AdaGrad

c) RMSDrop

- طبق بخش ۶ مقاله (نتایج) برای هر یک از موارد زیر، یک مثال ذکر کرده و توضیح دهید:
  a) Adam عملکرد بهتری از SGD داشته است.
  b) عملکرد Adam مشابه و یا کمی بهتر از SGD بوده است.
  c) AdaGrad نسبت به Adam عملکرد ضعیف‌تری داشته.

4. چکیده:

- موارد زیر در مورد الگوریتم Adam را با نوشتن به صورت چکیده‌ای مورد بررسی قرار دهید:
  - نوآوری و دستاورد (contribution)
  - علت اثرگذاری آن در یادگیری عمیق
  - نقاط قوت (تئوری و عملی)
  - محدودیت‌ها و شرایطی که Adam در آن ایده‌آل نیست
  - آیا استفاده از آن همچنان منطقی و کاربردیست؟

نکات تکمیلی:

سوالات عملی را می‌توانید در نوتبوک های تهیه شده در پوشه Practical مشاهده کنید.

سوال موجود در پوشه اختیاری ۱۰ امتیاز و سوالات موجود در پوشه دیگر هرکدام ۲۵ امتیاز دارند.

دانشجویان محترم حتماً فایل قوانین را مطالعه کرده و در انجام و ارسال تمارین رعایت بفرمایید.

موفق و سربلند باشید.