



## یادگیری عمیق

نیم سال دوم ۰۴-۰۵  
مدرس: مهدیه سلیمانی

ددلاین تمرین : ۱۸ اردیبهشت

تمرین اول

- برای ارسال هر تمرین تا ساعت ۲۳:۵۹ روز ددلاین فرصت دارید. مهلت تاخیر (مجاز و غیر مجاز) برای این تمرین ۷ روز است (یعنی حداکثر تاریخ ارسال تمرین ۲۵ اردیبهشت است)
- در هر کدام از سوالات، اگر از منابع خارجی استفاده کرده اید باید آن را ذکر کنید. در صورت همفکری با افراد دیگر هم باید نام ایشان را در سوال مورد نظر ذکر نمایید.
- پاسخ تمرین باید ماحصل دانسته های خود شما باشد. در صورت رعایت این موضوع، استفاده از ابزارهای هوش مصنوعی با ذکر نحوه و مصداق استفاده بلامانع است.
- پاسخ ارسالی واضح و خوانا باشد. در غیر این صورت ممکن است منجر به از دست دادن نمره شود.
- از تحویل تمرین نظری به شکل تایپ شده (فایل Word-Latex یا هر نوع تایپ شده دیگری) بپرهیزید و حتما پاسخ دست نویس تحویل دهید.
- پاسخ ارسالی باید توسط خود شما نوشته شده باشد. به اسکرین شات از منابع یا پاسخ افراد دیگر نمره ای تعلق نمی گیرد.
- در صورتی که بخشی از سوالها را جای دیگری آپلود کرده و لینک آن را قرار داده باشید، حتما باید تاریخ آپلود مشخص و قابل اتکا باشد.
- محل بارگذاری سوالات نظری و عملی در هر تمرین مجزا خواهد بود. به منظور بارگذاری بایستی تمرین تئوری در یک فایل pdf با نام `HW1_[First-Name]_[Last-Name]_[Student-Id].pdf` بارگذاری شوند.
- در صورت وجود هرگونه ابهام یا مشکل، در کوثرای درس آن مشکل را بیان کنید و از پیغام دادن مستقیم به دستیاران آموزشی خودداری کنید.
- طراحان این تمرین: رضا قربانی- عرفان صبحائی- مزدک تیموریان- محمد محمدبیگی- مهشید دهقانی- حسین شریفی- امیرحسین کوچکیان- حسین شاه آبادی

بخش نظری (۱۰۰+۱۰ نمره)

پرسش ۱. مرور ریاضیات برای یادگیری عمیق (۱۵+۵ نمره)

## بخش اول

فرض کنید

$$f: \mathbb{R}^n \rightarrow \mathbb{R}^m$$

و

$$g: \mathbb{R}^m \rightarrow \mathbb{R}^k$$

دو تابع مشتق پذیر باشند و تابع ترکیبی زیر تعریف شود:

$$h(x) = g(f(x))$$

آ) ماتریس ژاکوبین تابع  $h(x)$  را بر حسب ماتریس‌های ژاکوبین  $J_f(x)$  و  $J_g(x)$  بدست آورید.

ب) فرض کنید

$$f(x) = \text{ReLU}(Ax + b), \quad g(z) = Cz$$

که در آن

$$A \in \mathbb{R}^{m \times n}, \quad b \in \mathbb{R}^m, \quad C \in \mathbb{R}^{k \times m}$$

و تابع ReLU به صورت زیر تعریف می‌شود:

$$\text{ReLU}(u_i) = \max(0, u_i)$$

۱. ترکیبی  $h(x) = g(f(x))$  را بنویسید.
۲. ماتریس ژاکوبین  $J_h(x)$  را محاسبه کنید.
۳. مشتق تابع ReLU را با استفاده از یک ماتریس قطری بیان کنید.

ج) فرض کنید تابع  $f$  برابر با تابع Softmax باشد. برای یک بردار

$$z \in \mathbb{R}^m$$

تابع Softmax به صورت زیر تعریف می‌شود:

$$\text{Softmax}(z)_i = \frac{e^{z_i}}{\sum_{j=1}^m e^{z_j}}$$

اکنون فرض کنید

$$f(x) = \text{Softmax}(Ax + b), \quad g(z) = Cz$$

۱. تابع ترکیبی  $h(x) = g(f(x))$  را بنویسید.
۲. ژاکوبین تابع Softmax یعنی

$$J_{\text{Softmax}}(z)$$

را محاسبه کنید و نشان دهید که مؤلفه‌های آن به صورت زیر هستند:

$$\frac{\partial s_i}{\partial z_j} = s_i(\delta_{ij} - s_j)$$

- که در آن  $s = \text{Softmax}(z)$  و  $\delta_{ij}$  دلتای کرونکر است.
۳. با استفاده از قاعده زنجیره‌ای، ماتریس ژاکوبین

$$J_h(x)$$

را برای این حالت محاسبه کنید.

## بخش دوم

فرض کنید

$$x \in \mathbb{R}^n, \quad y \in \mathbb{R}^m, \quad W \in \mathbb{R}^{m \times n}, \quad b \in \mathbb{R}^m$$

و تابع زیر تعریف شده باشد:

$$z = Wx + b$$

$$L(W, b, x) = \frac{1}{2} \|z - y\|^2$$

(آ) ژاکوبین بردار  $z$  نسبت به بردار  $x$  را محاسبه کنید:

$$\frac{\partial z}{\partial x}$$

و اندازه این ماتریس را مشخص کنید.

(ب) گرادیان تابع ضرر را نسبت به بردار بایاس  $b$  محاسبه کنید:

$$\nabla_b L$$

(ج) گرادیان تابع ضرر را نسبت به ورودی  $x$  محاسبه کنید:

$$\nabla_x L$$

(د) گرادیان تابع ضرر را نسبت به ماتریس وزن  $W$  محاسبه کنید:

$$\nabla_W L$$

## پرسش ۲. پس انتشار خطا (۱۵ نمره)

فرض کنید یک شبکه عصبی دولایه به صورت زیر داریم:

$$\begin{aligned} z_1 &= W_1 x^{(i)} + b_1 \\ a_1 &= \tanh(z_1) \\ z_2 &= W_2 a_1 + b_2 \\ \hat{y}^{(i)} &= z_2 \\ L^{(i)} &= \frac{1}{2} (\hat{y}^{(i)} - y^{(i)})^2 \\ J &= \frac{1}{m} \sum_{i=1}^m L^{(i)} \end{aligned}$$

در این جا،  $x^{(i)}$  نشان دهنده یک نمونه ورودی منفرد با ابعاد  $D_x \times 1$  است. همچنین  $y^{(i)}$  برچسب خروجی (هدف) متناظر با این نمونه بوده و یک اسکالر است. تعداد کل نمونه‌های داده برابر با  $m$  است. فرض می‌کنیم لایه مخفی دارای  $D_{a1}$  نود باشد؛ بنابراین  $z_1$  دارای ابعاد  $D_{a1} \times 1$  خواهد بود.

**نکته:** در بخش مربوط به پیاده‌سازی برداری (vectorized)، منظور این است که کل داده‌ها، یعنی هر  $m$  نمونه، به صورت همزمان به شبکه داده می‌شوند.

۱. ابعاد ماتریس‌ها و بردارهای  $W_1, b_1, W_2, b_2$  را مشخص کنید. اگر شبکه را روی چندین نمونه به صورت برداری (vectorized) پیاده‌سازی کنیم، ابعاد وزن‌ها و بایاس‌ها چه تغییری می‌کنند؟ در حالت برداری، ابعاد  $X$  و  $Y$  چگونه خواهند بود؟

۲.  $\frac{\partial J}{\partial \hat{y}^{(i)}}$  را محاسبه کنید و آن را با  $\delta_1^{(i)}$  نشان دهید.

۳.  $\frac{\partial \hat{y}^{(i)}}{\partial z_2}$  را محاسبه کنید و آن را با  $\delta_2^{(i)}$  نشان دهید.

۴.  $\frac{\partial z_2}{\partial a_1}$  را محاسبه کنید و آن را با  $\delta_3^{(i)}$  نشان دهید.

۵.  $\frac{\partial a_1}{\partial z_1}$  را محاسبه کنید و آن را با  $\delta_4^{(i)}$  نشان دهید.

۶.  $\frac{\partial z_1}{\partial W_1}$  را محاسبه کنید و آن را با  $\delta_5^{(i)}$  نشان دهید.

۷.  $\frac{\partial J}{\partial W_1}$  را محاسبه کنید. می‌توانید از نتایج قسمت‌های قبل استفاده کنید. **راهنما:** به ابعاد ماتریس‌ها دقت کنید.

## پرسش ۳. روش‌های بهینه‌سازی مرتبه اول (۲۵ نمره)

در این مسئله، تکامل الگوریتم‌های بهینه‌سازی مرتبه اول را بررسی می‌کنیم. هدف، تحلیل رفتار این الگوریتم‌ها برای کمینه‌سازی تابع محدب زیر است:

$$f(x, y) = 50x^2 + y^2$$

که در آن  $x, y \in \mathbb{R}$  است. این تابع دارای انحنا بسیار زیاد در راستای محور  $x$  و انحنا کم در راستای محور  $y$  است، بنابراین یک مثال مناسب برای بررسی چالش‌های بهینه‌سازی محسوب می‌شود.

## • الگوریتم Gradient Descent

الگوریتم گرادیان کاهشی<sup>۱</sup> پارامترها را در جهت بیشترین کاهش تابع هدف به روزرسانی می‌کند:

$$x_{k+1} = x_k - \eta \nabla f(x_k)$$

که در آن  $\eta > 0$  نرخ یادگیری<sup>۲</sup> است.

(الف) گرادیان تابع داده شده  $\nabla f(x, y)$  را محاسبه کنید.

(ب) چرا استفاده از یک نرخ یادگیری ثابت  $\eta$  برای کمینه‌سازی این تابع، با توجه به تفاوت انحنای آن در راستای محوره‌های  $x$  و  $y$ ، می‌تواند مشکل‌ساز باشد؟

(ج) با فرض  $\eta = 0.1$  و نقطه اولیه  $(x_0, y_0) = (1, 1)$ ، دو تکرار اول الگوریتم گرادیان کاهشی را محاسبه کنید. آیا این نتایج نشان‌دهنده مشکلات ذکر شده در بخش (ب) هستند؟

## • الگوریتم Adaptive Learning Rates (AdaGrad)

الگوریتم AdaGrad نرخ یادگیری هر پارامتر را بر اساس تاریخچه گرادیان‌های آن تنظیم می‌کند:

$$\eta_i^{(k)} = \frac{\eta}{\sqrt{\sum_{j=1}^k (g_i^{(j)})^2 + \epsilon}}$$

$$x_{i,k+1} = x_{i,k} - \eta_i^{(k)} g_i^{(k)}$$

که در آن  $g_i^{(j)}$  مؤلفه  $i$ -ام گرادیان در گام  $j$  است و  $\epsilon > 0$  برای جلوگیری از تقسیم بر صفر است.

(الف) چرا تنظیم نرخ یادگیری به صورت جداگانه توسط AdaGrad برای هر پارامتر در تابع  $f(x, y)$  مفید است؟

(ب) با  $\eta = 0.1$  و  $\epsilon = 10^{-8}$ ، نرخ‌های یادگیری مؤثر  $\eta_x^{(1)}$  و  $\eta_y^{(1)}$  را بعد از اولین تکرار محاسبه کنید.

(ج) نرخ یادگیری در طول زمان چه تغییری می‌کند و چه اثری دارد؟

## • الگوریتم RMSProp

الگوریتم RMSProp نسخه بهبودیافته AdaGrad است که به جای جمع کل گرادیان‌ها، از میانگین متحرک نمایی<sup>۳</sup> آن‌ها استفاده می‌کند:

$$s_{k+1} = \gamma s_k + (1 - \gamma) g_k^2$$

$$x_{k+1} = x_k - \frac{\eta}{\sqrt{s_{k+1} + \epsilon}} g_k$$

که در آن  $s_k$  میانگین متحرک نمایی مربع گرادیان‌ها و  $\gamma \in [0, 1)$  ضریب کاهش است.

(الف) استفاده از میانگین متحرک نمایی مربع گرادیان‌ها چه تفاوتی در مقایسه با AdaGrad ایجاد می‌کند؟ یک مزیت واضح RMSProp نسبت به AdaGrad را بیان کنید.

(ب) نقش پارامتر  $\gamma$  در RMSProp را با نقش  $\beta$  در Momentum مقایسه کنید. هر کدام چه چیزی را کنترل می‌کنند؟

<sup>۱</sup> Gradient Descent

<sup>۲</sup> Learning Rate

<sup>۳</sup> Exponential Moving Average (EMA)

## • الگوریتم Adam

Adam ترکیبی از Momentum و AdaGrad است:

$$\begin{aligned} m_{k+1} &= \beta_1 m_k + (1 - \beta_1) g_k \\ v_{k+1} &= \beta_2 v_k + (1 - \beta_2) g_k^2 \\ \hat{m}_{k+1} &= \frac{m_{k+1}}{1 - \beta_1^k}, \quad \hat{v}_{k+1} = \frac{v_{k+1}}{1 - \beta_2^k} \\ x_{k+1} &= x_k - \frac{\eta}{\sqrt{\hat{v}_{k+1} + \epsilon}} \hat{m}_{k+1} \end{aligned}$$

(الف) بر اساس روابط بالا، مشخص کنید که کدام یک از متغیرهای  $m_k$  و  $v_k$  مربوط به Momentum و کدام مربوط به AdaGrad می‌باشند.

(ب) چرا در الگوریتم Adam نیاز به تصحیح بایاس<sup>۴</sup> برای  $m_k$  و  $v_k$  وجود دارد؟ اگر این تصحیح انجام نشود، در چند تکرار ابتدایی چه مشکلی رخ می‌دهد؟

(ج) برای تابع  $f(x, y) = 50x^2 + y^2$  نتیجه‌گیری کنید که چرا انتظار می‌رود Adam عملکرد بهتری نسبت به گرادیان کاهشی ساده داشته باشد؟

## پرسش ۴. بهینه‌سازی مرتبه دوم (۱۵+۵ نمره)

در این پرسش، قصد داریم تابع ضرر  $E$  (با فرض داشتن شرایط لازم) را کمینه سازیم. فرض کنید  $w^*$  یک نقطه بهینه است به طوری که داریم  $\nabla E(w^*) = 0$  و ماتریس هسیان  $\nabla^2 E(w^*)$  مثبت معین است. با فرض بودن در همسایگی به اندازه کافی کوچک  $w^*$ ، طبق روش نیوتون داریم:

$$w^{(k+1)} = w^{(k)} + p^{(k)}$$

که در آن  $p^{(k)} := -\nabla^2 E(w^{(k)})^{-1} \cdot \nabla E(w^{(k)})$  است.

### (الف)

نشان دهید تساوی زیر برقرار است:

$$\|\nabla E(w^{(k+1)})\| = \|\nabla E(w^{(k+1)}) - \nabla E(w^{(k)}) - \nabla^2 E(w^{(k)})p^{(k)}\|. \quad (۱)$$

### (ب)

طبق قضیه تیلور، برای هر تابعی مانند  $f: \mathbb{R}^d \rightarrow \mathbb{R}$  که دو مرتبه مشتق پذیر باشد و مشتق آن پیوسته باشد و به ازای هر  $p \in \mathbb{R}^d$  و مقدار  $t \in (0, 1)$  داریم:

$$\nabla f(x + p) = \nabla f(x) + \int_0^1 \nabla^2 f(x + tp) p \, dt. \quad (۲)$$

با استفاده از روابط (۱) و (۲) و خواص عملگرهای نرم و انتگرال، نامساوی زیر را اثبات کنید.

$$\|\nabla E(w^{(k+1)})\| \leq \int_0^1 \|\nabla^2 E(w^{(k)} + tp^{(k)}) - \nabla^2 E(w^{(k)})\| \cdot \|p^{(k)}\| \, dt \quad (۳)$$

Bias Correction<sup>۴</sup>

(ج)

فرض می‌کنیم  $\nabla^2 E(\cdot)$  یک تابع  $L$ -لیپشیتز است؛ یعنی به ازای هر  $u$  و  $v$  داریم:

$$\|\nabla^2 E(u) - \nabla^2 E(v)\| \leq L \cdot \|u - v\|.$$

با استفاده از رابطه (۳) و فرض فوق، نشان دهید:

$$\|\nabla E(w^{(k+1)})\| \leq \frac{1}{2} L \|\nabla^2 E(w^{(k)})^{-1}\|^2 \cdot \|\nabla E(w^{(k)})\|^2$$

(د)

با توجه به معکوس‌پذیری  $\nabla^2 E(w^*)$ ، وجود دارد شعاع  $r > 0$  طوری که داریم  $\|\nabla^2 E(w^{(k)})^{-1}\| \leq 2 \|\nabla^2 E(w^*)^{-1}\|$  به ازای تمام  $w^{(k)}$  هایی که حداکثر فاصله  $r$  دارند یعنی  $\|w^{(k)} - w^*\| \leq r$ . با استفاده از این خاصیت و بخش‌های قبل، نتیجه گرفت:

$$\|\nabla E(w^{(k+1)})\| \leq 2L \|\nabla^2 E(w^*)^{-1}\|^2 \cdot \|\nabla E(w^{(k)})\|^2 \quad (۴)$$

با توجه به آنکه رابطه (۴) بیانگر تغییرات نرم‌گرادیان در هر گام نسبت به گام قبلی است، چه تفسیری می‌توان در مورد نقطه هم‌گرایی و نرخ هم‌گرایی برای نرم‌گرادیان‌ها ارائه کرد؟

## پرسش ۵. خواص پایه‌ای توابع لیپشیتس (۱۵ نمره)

در این بخش، چند خاصیت بنیادی توابع لیپشیتس را که به عنوان ابزارهای ریاضی برای تحلیل شبکه‌های عصبی به کار می‌روند، اثبات و بررسی می‌کنیم. ابتدا به تعریف لیپشیتس می‌پردازیم. اگر برای تابعی نظیر  $f: \mathbb{R}^n \rightarrow \mathbb{R}$  داشته باشیم:

$$\|f(x_1) - f(x_2)\| \leq L \|x_1 - x_2\|, \quad \forall x_1, x_2 \in \mathbb{R}^n$$

می‌گوییم این تابع  $L$ -Lipschitz است. برای درک بهتر این مفهوم  $\delta$  را در نظر داشته باشید. اکنون می‌توان ادعا کرد که برای هر  $x_2$  و به ازای هر  $\delta$  دلخواه داریم:

$$\|f(x_2 + \delta) - f(x_2)\| \leq L \|\delta\|$$

یعنی اگر به ازای هر مقدار ورودی نظیر  $x$ ، آن را به اندازه  $\delta$  تغییر دهیم، میزان تغییرات در خروجی تابع کمتر از  $L \|\delta\|$  خواهیم داشت و به بیانی یک کران بالا برای میزان تغییرات تعریف کردیم. می‌دانیم که در هر لایه از یک شبکه عصبی داریم:

$$x^{(\ell+1)} = h^{(\ell)}(W^{(\ell)} x^{(\ell)})$$

که در اینجا  $h(\cdot)$  یک تابع فعال‌سازی<sup>۵</sup> دلخواه است. حال به پرسش‌های زیر پاسخ دهید.

۱. **جمع توابع لیپشیتس:** ثابت کنید اگر  $f: \mathbb{R}^n \rightarrow \mathbb{R}^m$  یک تابع  $L_f$ -لیپشیتس و  $g: \mathbb{R}^n \rightarrow \mathbb{R}^m$  یک تابع  $L_g$ -لیپشیتس باشد، آنگاه تابع مجموع  $h(x) = f(x) + g(x)$  نیز لیپشیتس است. کران بالای ثابت لیپشیتس آن  $(L_h)$  چیست؟

<sup>۵</sup>Activation Function

## ۲. ترکیب توابع لپشیتس:

(آ) اثبات خاصیت ترکیب: ثابت کنید اگر  $g : \mathbb{R}^n \rightarrow \mathbb{R}^m$  یک تابع  $L_g$ -لپشیتس و  $f : \mathbb{R}^m \rightarrow \mathbb{R}^k$  یک تابع  $L_f$ -لپشیتس باشد، آنگاه تابع مرکب  $h(x) = f(g(x))$  نیز یک تابع لپشیتس است. کران بالای ثابت لپشیتس آن ( $L_h$ ) چیست؟

(ب) کران ثابت لپشیتس یک لایه: یک لایه شبکه عصبی به صورت  $f_k(x) = \sigma(W_k x)$  را در نظر بگیرید که در آن  $\sigma$  یک تابع فعال سازی  $L_\sigma$ -لپشیتس و  $W_k$  ماتریس وزن است. با استفاده از خاصیت ترکیب توابع که در بند (الف) اثبات کردید، نشان دهید که ثابت لپشیتس این لایه ( $L_k$ ) توسط  $L_k \leq L_\sigma \cdot \|W_k\|_2$  کران دار می شود. (یادآوری: ثابت لپشیتس تبدیل خطی  $x \mapsto Wx$  برابر با نورم طیفی ماتریس، یعنی  $\|W\|_2$  است).

## ۳. ارتباط با مشتق:

برای یک تابع مشتق پذیر  $f : \mathbb{R} \rightarrow \mathbb{R}$ ، ثابت کنید که  $f$  یک تابع  $L$ -لپشیتس است اگر و تنها اگر برای تمام مقادیر  $x$  داشته باشیم  $|f'(x)| \leq L$ .

## پرسش ۶. (۱۵ نمره) Regularization

ماتریس طراحی  $X$  (که سطر  $i$ -ام آن نقطه نمونه  $x_i^T$  است) و یک بردار  $n$ -تایی از برجسبها  $y = [y_1, \dots, y_n]^T$  به شما داده شده است. برای سادگی، فرض کنید  $X$  سفید (whitened) شده است، به طوری که  $X^T X = nI$  باشد. جمله بایاس یا بعد اضافی اضافه نکنید؛ برای ورودی صفر، خروجی همیشه صفر است. فرض کنید  $x_{*i}$  ستون  $i$ -ام ماتریس  $X$  را نشان می دهد.

۱. نشان دهید که تابع هزینه برای کمترین مربعات با تنظیم گر  $\ell_1$  (Lasso)، یعنی:

$$J_1(w) \triangleq \|Xw - y\|^2 + \lambda \|w\|_1 \quad (\lambda > 0 \text{ در آن})$$

می تواند به صورت زیر بازنویسی شود:

$$J_1(w) = \|y\|^2 + \sum_{i=1}^d f(x_{*i}, w_i)$$

که در آن  $f(\cdot, \cdot)$  یک تابع مناسب است که آرگومان اول آن یک بردار و آرگومان دوم آن یک اسکالر است.

۲. با استفاده از پاسخ خود به بخش (۱)، شرایط لازم و کافی برای مولفه  $i$ -ام بهینه ساز  $w^*$  تابع  $J_1(\cdot)$  را استخراج کنید به طوری که هر یک از سه خاصیت زیر را برآورده کند:

$$w_i^* > 0, \quad w_i^* = 0, \quad w_i^* < 0.$$

۳. برای بهینه ساز  $w^\#$  تابع هزینه کمترین مربعات با تنظیم گر  $\ell_2$  (Ridge)، یعنی:

$$J_2(w) \triangleq \|Xw - y\|^2 + \lambda \|w\|_2^2 \quad (\lambda > 0 \text{ در آن})$$

یک شرط لازم و کافی برای  $w_i^\# = 0$  استخراج کنید، که در آن مولفه  $i$ -ام بردار  $w^\#$  است.

۴. یک بردار را تُنک (Sparse) می نامیم اگر بیشتر مولفه های آن صفر باشند. بر اساس پاسخ های خود به بخش های (۲) و (۳)، کدام یک از بردارهای  $w^*$  و  $w^\#$  به احتمال زیاد تُنک تر است؟ چرا؟