دانشگاه صنعتی امیرکبیر

(پلی‌تکنیک تهران)

دانشکده علوم کامپیوتر

پایان‌نامه کارشناسی‌ارشد

گزارش پروژه درس داده کاوی محاسباتی

پروژه ۴

نگارش

محمدصادق قلی زاده

استاد راهنما

دکتر مهدی قطعی


بسم الله الرحمن الرحيم

صفحه فرم ارزیابی و تصویب پایان نامه- فرم تأیید اعضاء کمیته دفاع

در این صفحه فرم دفاع یا تایید و تصویب پایان نامه موسوم به فرم کمیته دفاع- موجود در پرونده آموزشی- را قرار دهید.

نکات مهم:

- نگارش پایان نامه/رساله باید به زبان فارسی و بر اساس آخرین نسخه دستورالعمل و راهنمای تدوین پایان نامه های دانشگاه صنعتی امیرکبیر باشد.(دستورالعمل و راهنمای حاضر)
- رنگ جلد پایان نامه/رساله چاپی کارشناسی، کارشناسی ارشد و دکترا باید به ترتیب مشکی، طوسی و سفید رنگ باشد.
- چاپ و صحافی پایان نامه/رساله بصورت پشت و رو(دورو) بلامانع است و انجام آن توصیه می شود.

دانشگاه صنعتی امیرکبیر
(پلی‌تکنیک تهران)

به نام خدا

تعهدنامه اصالت اثر

تاریخ:

اینجانب **محمدصادق قلی زاده** متعهد می‌شوم که مطالب مندرج در این پایان‌نامه حاصل کار پژوهشی اینجانب تحت نظارت و راهنمایی اساتید دانشگاه صنعتی امیرکبیر بوده و به دستاوردهای دیگران که در این پژوهش از آنها استفاده شده است مطابق مقررات و روال متعارف ارجاع و در فهرست منابع و مآخذ ذکر گردیده است. این پایان‌نامه قبلاً برای احراز هیچ مدرک هم‌سطح یا بالاتر ارائه نگردیده است.

در صورت اثبات تخلف در هر زمان، مدرک تحصیلی صادر شده توسط دانشگاه از درجه اعتبار ساقط بوده و دانشگاه حق پیگیری قانونی خواهد داشت.

کلیه نتایج و حقوق حاصل از این پایان‌نامه متعلق به دانشگاه صنعتی امیرکبیر می‌باشد. هرگونه استفاده از نتایج علمی و عملی، واگذاری اطلاعات به دیگران یا چاپ و تکثیر، نسخه‌برداری، ترجمه و اقتباس از این پایان نامه بدون موافقت کتبی دانشگاه صنعتی امیرکبیر ممنوع است. نقل مطالب با ذکر مآخذ بلامانع است.

محمدصادق قلی زاده

امضا

نویسنده پایان نامه، در صورت تمایل میتواند برای سپاسگزاری پایان نامه خود را به شخص یا اشخاص و یا ارگان خاصی تقدیم نماید.

سپاس‌گزاری

نویسنده پایان‌نامه می‌تواند مراتب امتنان خود را نسبت به استاد راهنما و استاد مشاور و یا دیگر افرادی که طی انجام پایان‌نامه به نحوی او را یاری و یا با او همکاری نموده‌اند ابراز دارد.

محمدصادق قلی‌زاده

چکیده

این گزارش محدودیت‌های هندسی و محاسباتی الگوریتم‌های بهینه‌سازی در یادگیری ماشین را بررسی می‌کند؛ از سطوح درجه‌دوم ساده گرفته تا شبکه‌های عصبی عمیق. از طریق یک تحلیل چندمرحله‌ای، ما به بررسی موازنه بنیادین میان کارایی گام‌ها (نرخ همگرایی) و مقیاس‌پذیری (هزینه حافظه و محاسبات) می‌پردازیم. در مرحله نخست، اثر بدشرطی [۱] بر گرادیان نزولی به‌صورت بصری نشان داده می‌شود و رفتار ناکارآمد «زیگ‌زاگی» آن در دره‌هایی با خمیدگی بالا برجسته می‌گردد. این رفتار در مقابل روش نیوتن قرار می‌گیرد که با استفاده از اطلاعات مرتبه دوم (ماتریس Hessian)، خمیدگی را نرمال‌سازی کرده و به همگرایی مستقیم دست می‌یابد. در مرحله دوم، یک «رژیم نیوتنی» برای مسائل با بُعد کم شناسایی می‌شود (با استفاده از مجموعه‌داده Breast Cancer)، که در آن روش‌های شبه‌نیوتنی مانند L-BFGS و گرادیان مزدوج [۲] از نظر زمان واقعی اجرا (wall-clock time) به‌طور قابل‌توجهی از گرادیان نزولی تصادفی [۳] پیشی می‌گیرند. در مرحله سوم، تله مقیاس‌پذیری ذاتی در یادگیری عمیق تحلیل می‌شود. محاسبات نظری روی یک شبکه عمیق Fashion-MNIST نشان می‌دهد که ذخیره ماتریس هسین حتی برای یک معماری نسبتاً ساده به حافظه‌ای بسیار عظیم (در حدود ۴۰ گیگابایت) نیاز دارد، که روش‌های خالص مرتبه دوم را عملاً غیرقابل استفاده می‌سازد. در نتیجه، ضرورت استفاده از روش‌های مرتبه اول تطبیقی مانند Adam برای فضاهای با بُعد بالا تأیید می‌شود. در نهایت، نشان داده می‌شود که عمودسازی داده‌ها (Data Orthogonalization) یک جایگزین هندسی برای بهینه‌سازهای پیچیده فراهم می‌کند. با اعمال تجزیه QR بر ویژگی‌های هم‌بسته، عدد شرطی مسئله به یک کاهش می‌یابد و عملاً به یک بهینه‌ساز ساده مانند SGD اجازه داده می‌شود تا سرعت همگرایی مشابه روش نیوتن را تجربه کند. در پایان نتیجه‌گیری می‌شود که اگرچه روش‌های مرتبه دوم از نظر تئوری همگرایی برتری دارند، یادگیری عمیق مدرن بر یک سازش متکی است: استفاده از تقریب‌های مرتبه اول همراه با بهبود هندسه داده‌ها برای پیمایش

[۱] (ill-conditioning)

[۲] (Conjugate Gradient)

[۳] (SGD یا Stochastic Gradient Descent)

مؤثر فضاهای با بُعد بالا. GitHub

واژه‌های کلیدی:

روش نیوتن، گرادیان کاهشی، هندسه بهینه سازی

فهرست مطالب

عنوان صفحه

فهرست اشکال

شکل صفحه

فهرست جداول

جدول صفحه

فهرست نمادها

نماد	مفهوم
$\mathbb{R}^n$	فضای اقلیدسی با بعد n
$\mathbb{S}^n$	کره یکه n بعدی
M^m	خمینه m-بعدی M
$\mathfrak{X}(M)$	جبر میدان‌های برداری هموار روی M
$\mathfrak{X}^1(M)$	مجموعه میدان‌های برداری هموار یکه روی (M,g)
$\Omega^p(M)$	مجموعه p-فرمی‌های روی خمینه M
Q	اپراتور ریچی
$\mathcal{R}$	تانسور انحنای ریمان
ric	تانسور ریچی
L	مشتق لی
Φ	۲-فرم اساسی خمینه تماسی
∇	التصاق لوی-چویتای
Δ	لاپلاسین ناهموار
∇^*	عملگر خودالحاق صوری القا شده از التصاق لوی-چویتای
g_s	متر ساساکی
∇	التصاق لوی-چویتای وابسته به متر ساساکی
Δ	عملگر لاپلاس-بلترامی روی p-فرم‌ها

فصل ۱

مقدمه

۱-۱ مقدمه

بهینه‌سازی در قلب یادگیری ماشین قرار دارد. چه در حال برازش یک رگرسیون خطی ساده باشیم و چه در حال آموزش یک شبکه عصبی عمیق، هدف همواره یکسان است: پیمایش یک چشم‌انداز خطا با بُعد بالا برای یافتن پارامترهایی که خطا را کمینه می‌کنند. با وجود ثبات هدف، روش‌های دستیابی به آن بسته به مقیاس و هندسه مسئله، تفاوت‌های چشمگیری دارند.

این گزارش سازوکارهای بهینه‌سازی را از منظر خمیدگی و مقیاس‌پذیری بررسی می‌کند. به‌طور مشخص، بررسی می‌کنیم که چرا راه‌حل‌های «کتاب درسی» ریاضی ،مانند روش نیوتن که با استفاده از مشتقات مرتبه دوم گام ایده‌آل را محاسبه می‌کند، در یادگیری عمیق مدرن اغلب کنار گذاشته می‌شوند و جای خود را به تقریب‌های ساده‌تر مرتبه اول مانند گرادیان نزولی تصادفی [۱] می‌دهند.

۱-۱-۱ بیان مسئله

چالش اصلی در بهینه‌سازی، موازنه میان کیفیت گام و هزینه محاسباتی است.

روش‌های مرتبه اول روش‌های مرتبه اول (مانند گرادیان نزولی) تنها به گرادیان (شیب) متکی هستند. هزینه محاسباتی هر تکرار آن‌ها پایین است، اما نسبت به خمیدگی سطح «نابینا» بوده و اغلب در دره‌های بدشرط، دچار حرکت زیگ‌زاگی و ناکارآمد می‌شوند.

[۱] (Stochastic Gradient Descent یا SGD)

روش‌های مرتبه دوم روش‌های مرتبه دوم (مانند روش نیوتن) از ماتریس هسین [۲] برای نرمال‌سازی هندسه سطح استفاده می‌کنند. این روش‌ها گام‌هایی بسیار کارآمد و مستقیم برمی‌دارند، اما هزینه محاسباتی آن‌ها به‌صورت درجه دوم یا سوم با تعداد پارامترها افزایش می‌یابد.

۱-۱-۲ اهداف

هدف این تمرین، تحلیل تجربی این موازنه در سه رژیم متمایز است:

- **تحلیل نظری:** تجسم رفتار هندسی بهینه‌سازها بر روی سطوح درجه دوم بدشرط برای درک «مسئله دره».
- **رژیم نیوتنی:** نمایش برتری روش‌های مرتبه دوم (مانند L-BFGS و گرادیان مزدوج [۳]) در شبکه‌های عصبی کلاسیک و کوچک‌مقیاس که محاسبه هسین در آن‌ها امکان‌پذیر است.
- **رژیم یادگیری عمیق:** کمی‌سازی «تله مقیاس‌پذیری» که در آن نیازهای حافظه‌ای ماتریس هسین، استفاده از روش‌های خالص نیوتنی را برای شبکه‌های عمیق غیرممکن می‌سازد.
- **شرطی‌سازی هندسی:** بررسی این موضوع که چگونه پیش‌پردازش داده‌ها (عمودسازی از طریق تجزیه QR) می‌تواند هندسه مسئله را بهبود دهد و به روش‌های ساده مرتبه اول اجازه دهد به سرعت همگرایی روش‌های مرتبه دوم دست یابند.

۱-۱-۳ ساختار گزارش

ادامه این گزارش به‌صورت زیر سازمان‌دهی شده است:

- بخش اول، تحلیل بصری گرادیان نزولی در مقایسه با روش نیوتن را بر روی یک تابع درجه دوم مصنوعی ارائه می‌دهد.
- بخش دوم، این روش‌ها را بر روی یک مسئله دسته‌بندی دودویی (مجموعه‌داده Breast Cancer) به‌کار می‌گیرد تا همگرایی آن‌ها را در یک محیط عملی و کم‌بُعد مقایسه کند.

[۲] (Hessian)
[۳] (Conjugate Gradient)

- بخش سوم، محدودیت‌های سخت‌افزاری آموزش شبکه‌های عصبی عمیق (Fashion-MNIST) را محاسبه می‌کند و نشان می‌دهد چرا هوش مصنوعی مدرن به Adam و SGD متکی است.
- بخش چهارم، تأثیر عمودبودگی (تجزیه QR) بر همگرایی را نشان می‌دهد و ثابت می‌کند که بهبود هندسه داده‌ها می‌تواند جایگزینی برای الگوریتم‌های پیچیده بهینه‌سازی باشد.

فصل ۲

مروری بر ادبیات

۲-۱ تعاریف مفاهیم پایه

۲-۱-۱ روش نیوتن

روش نیوتن [۱] یک روش بهینه‌سازی مرتبه دوم است که از اطلاعات انحنای سطح خطا (ماتریس هسین، یعنی مشتقات مرتبه دوم) استفاده می‌کند [۱، ۲]. برخلاف گرادیان نزولی که تنها جهت شیب را در نظر می‌گیرد، روش نیوتن شکل هندسی سطح خطا را نیز به‌طور کامل در نظر می‌گیرد.

قانون به‌روزرسانی این روش به‌صورت زیر بیان می‌شود:

$$\theta_{\text{new}} = \theta_{\text{old}} - H^{-1}\nabla J(\theta)$$

که در آن H ماتریس هسین تابع هزینه است. روش نیوتن دارای همگرایی درجه دوم [۲] بوده و در صورت نزدیک بودن مقدار اولیه به جواب بهینه، در تعداد گام‌های بسیار کم همگرا می‌شود [۱].

با این حال، محاسبه و ذخیره‌سازی ماتریس هسین و همچنین معکوس‌سازی آن برای مسائل با تعداد پارامترهای زیاد (مانند شبکه‌های عصبی عمیق) از نظر محاسباتی و حافظه‌ای بسیار پرهزینه یا حتی غیرممکن است، که کاربرد عملی این روش را به مسائل با بُعد پایین محدود می‌کند [۲].

[۱] (Newton's Method)

[۲] (Quadratic Convergence)

۲-۱-۲ گرادیان مزدوج

گرادیان مزدوج [۳] روشی تکرارشونده برای حل مسائل بهینه‌سازی و دستگاه‌های خطی بزرگ‌مقیاس است که با هدف جلوگیری از حرکت‌های زیگ‌زاگی در گرادیان نزولی طراحی شده است [۱]. در این روش، جهت‌های جستجو به‌گونه‌ای انتخاب می‌شوند که نسبت به جهت‌های قبلی مزدوج (Conjugate) باشند.

ویژگی مهم روش CG آن است که نیازی به محاسبه یا ذخیره‌سازی صریح ماتریس هسین ندارد، اما همچنان می‌تواند از اطلاعات هندسی مسئله بهره ببرد. در مسائل محدب با ساختار درجه دوم، این روش در حداکثر d گام (که d تعداد پارامترهاست) به جواب دقیق می‌رسد [۳].

به همین دلیل، گرادیان مزدوج برای مسائل بزرگ‌مقیاس، به‌ویژه در بهینه‌سازی عددی و یادگیری ماشین، گزینه‌ای بسیار کارآمدتر از روش نیوتن محسوب می‌شود [۱].

۲-۱-۳ تجزیه QR و تعامد

هر ماتریس ویژگی X را می‌توان با استفاده از تجزیه QR به‌صورت زیر نوشت:

$$X = QR$$

که در آن:

- Q ماتریسی با ستون‌های متعامد [۴] است،
- R یک ماتریس بالامثلثی است [۴].

در حالتی که ویژگی‌های ورودی یک مدل دارای هم‌خطی یا هم‌بستگی بالا [۵] باشند، سطح خطای تابع هزینه به‌شکل دره‌هایی باریک و کشیده ظاهر می‌شود. در این شرایط، کانتورهای خطا بیضوی و کشیده بوده و روش‌های گرادیانی با نوسان و همگرایی کند مواجه می‌شوند [۲].

با ضرب ماتریس ویژگی‌ها در Q یا استفاده از نمایش متعامد داده‌ها، ویژگی‌ها عمودسازی شده و کانتورهای خطا به شکل‌هایی نزدیک به دایره تبدیل می‌شوند. این تغییر هندسی باعث بهبود قابل توجه در نرخ همگرایی گرادیان نزولی می‌شود، بدون آنکه نیازی به استفاده از بهینه‌سازهای مرتبه دوم پیچیده باشد [۴].

---

[۳] (CG یا Conjugate Gradient)

[۴] (Orthogonal)

[۵] (Collinear)

۲-۱-۴ عدد وضعیت

عدد وضعیت [۶] معیاری برای سنجش میزان بدشرطی یک مسئله بهینه‌سازی است و به‌صورت نسبت بزرگ‌ترین مقدار ویژه به کوچک‌ترین مقدار ویژه ماتریس هسین تعریف می‌شود [۵]:

$$\kappa = \frac{\lambda_{\max}}{\lambda_{\min}}$$

هرچه مقدار κ بزرگ‌تر باشد، سطح خطای تابع هزینه کشیده‌تر و دره‌ای‌تر بوده و همگرایی روش‌های مبتنی بر گرادیان به‌طور قابل توجهی کندتر خواهد شد. در چنین شرایطی، گرادیان نزولی عملکرد ضعیفی از خود نشان می‌دهد [۲].

در مقابل، روش نیوتن نسبت به عدد وضعیت ناوردا [۷] است، زیرا با استفاده از معکوس ماتریس هسین، خمیدگی سطح خطا را نرمال‌سازی کرده و حرکت به سمت کمینه را به‌طور مستقیم انجام می‌دهد [۱].

---

۶ (Condition Number)

۷ (Invariant)

فصل ۳

تحلیل نتایج

۳-۱ تحلیل ریاضی (سطوح بدشرط)

۳-۱-۱ هدف

هدف اصلی این بخش، مشاهده و تحلیل بصری تفاوت رفتار الگوریتم‌های بهینه‌سازی مرتبه اول (گرادیان نزولی) و مرتبه دوم (روش نیوتن) است. تمرکز ما به‌طور خاص بر یک مسئله بدشرط [۱] قرار دارد: یک تابع هزینه درجه دوم با ساختار دره‌ای، که در آن میزان خمیدگی در ابعاد مختلف به‌طور قابل‌توجهی متفاوت است.

۳-۱-۲ پیاده‌سازی

ما یک تابع هزینه درجه دوم مصنوعی به‌صورت زیر تعریف کردیم:

$$J(v) = \frac{۱}{۲} v^T H v$$

که در آن ماتریس هسین قطری به‌شکل زیر است:

$$H = \mathrm{diag}(۱, ۵۰)$$

[۱] (ill-conditioned)

این انتخاب عدد وضعیت

$$\kappa = ۵۰$$

را ایجاد می‌کند که منجر به کانتورهای بیضوی بسیار کشیده می‌شود.

الگوریتم‌های بهینه‌سازی از ابتدا در زبان Python پیاده‌سازی شدند. قواعد به‌روزرسانی اصلی به‌صورت زیر تعریف شده‌اند:

```
Update Descent Gradient #
derivative) (first gradient the on only Relies #
learning_rate): ,gradient ,gradient_descent_step(theta def
    gradient * learning_rate - theta return

Update Method 'sNewton #
curvature for correct to Hessian inverse the Uses #
Hessian): ,gradient ,newtons_method_step(theta def
    np.linalg.inv(Hessian) = H_inv
    gradient @ H_inv - theta return
```

۳-۱-۳ نتایج تجربی

هر دو الگوریتم از نقطه شروع

$$\theta_{\circ} = \begin{bmatrix} ۴/۰ \\ ۱/۰ \end{bmatrix}$$

آغاز شدند. مسیر گام‌های بهینه‌سازی بر روی خطوط تراز (کانتورهای) تابع هزینه ترسیم شد.

شکل ۱: مقایسه مسیرها در شکل ۱ (بالا)، خط قرمز مسیر گرادیان نزولی و خط آبیِ چین‌دار مسیر روش نیوتن را نشان می‌دهد.

مشاهدات

- **گرادیان نزولی (قرمز):** مسیر حرکت نوسان شدید («زیگ‌زاگ») را نشان می‌دهد. الگوریتم به‌سرعت عرض دره باریک (در راستای محور y با شیب تند) را طی می‌کند، اما در امتداد دره (در راستای محور x با شیب کم) پیشرفت بسیار کندی دارد.
- **روش نیوتن (آبی):** مسیر یک خط مستقیم از نقطه شروع تا کمینه سراسری

$$[۰, ۰]$$

است و تنها در یک گام همگرا می‌شود.

۳-۱-۴ تحلیل

تفاوت عملکرد این دو روش با نحوه برخورد آن‌ها با خمیدگی سطح توضیح داده می‌شود.

پدیده زیگ‌زاگ (گرادیان نزولی) گرادیان نزولی در جهت بیشترین کاهش حرکت می‌کند:

$$-\nabla J$$

در یک دره بدشرط، گرادیان تحت سلطه جهتی با بیشترین خمیدگی (مقدار ویژه $\lambda = ۵۰$) قرار دارد. با این حال، کمینه واقعی عمدتاً در راستای جهتی با کمترین خمیدگی ($\lambda = ۱$) واقع شده است.

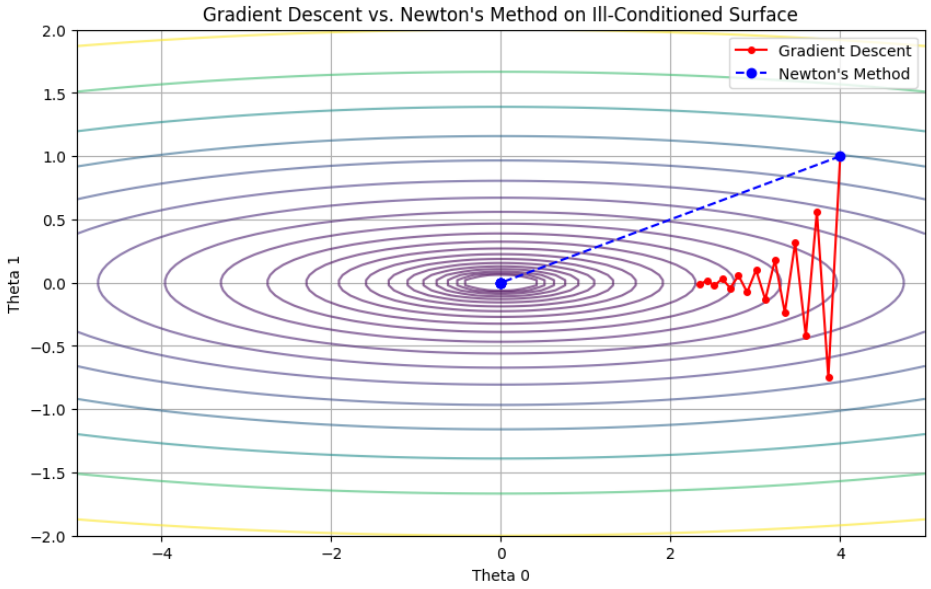
این ناهماهنگی باعث می‌شود بردار گرادیان تقریباً عمود بر مسیر بهینه اشاره کند. در نتیجه، الگوریتم از کف دره عبور کرده، جهت خود را معکوس می‌کند و این چرخه تکرار می‌شود؛ رفتاری که به نوسان مشاهده‌شده منجر می‌گردد.

اصلاح هندسی (روش نیوتن) روش نیوتن با استفاده از معکوس ماتریس هسین ($H^{-۱}$)، هندسه سطح خطا را به‌طور مؤثر بازمقیاس می‌کند:

$$\theta_{\text{new}} = \theta - H^{-۱}\nabla J$$

با ضرب در $H^{-۱}$، الگوریتم مؤلفه گرادیان در جهت پرشیب را بر ۵۰ و در جهت کم‌شیب را بر ۱

تقسیم می‌کند. این پیش‌شرطی‌سازی [۲] یک چشم‌انداز مؤثر همسان‌گرد (تقریباً دایره‌ای) ایجاد می‌کند که در آن گرادیان مستقیماً به سمت کمینه اشاره می‌کند. در نتیجه، بر روی این سطح درجه دوم، همگرایی آنی حاصل می‌شود.



شکل ۳-۱: مقایسه نتایج نیوتن با گرادیان کاهشی

۳-۲ شبکه عصبی کلاسیک (فضای نیوتنی)

۳-۲-۱ هدف

هدف این بخش شناسایی «رژیم نیوتنی» است؛ یعنی مقیاس خاصی از مسئله که در آن روش‌های بهینه‌سازی مرتبه دوم (که از اطلاعات خمیدگی استفاده می‌کنند) نه‌تنها قابل اجرا هستند، بلکه به‌طور چشمگیری از روش‌های استاندارد مرتبه اول برتری دارند. برای آزمون این فرضیه، یک مسئله دسته‌بندی دودویی با استفاده از یک شبکه عصبی کوچک بررسی می‌شود.

۳-۲-۲ روش‌شناسی

مجموعه‌داده از مجموعه‌داده Breast Cancer Wisconsin (Diagnostic) استفاده شد؛ یک مسئله کلاسیک دسته‌بندی دودویی با ۳۰ ویژگی حقیقی.

[۲] (preconditioning)

معماری مدل یک پرسپترون چندلایه کم‌عمق [۳] با ساختار زیر ساخته شد:

- **لایه ورودی:** ۳۱ واحد (۳۰ ویژگی به‌علاوه ۱ بایاس)
- **لایه پنهان:** ۵ نورون با تابع فعال‌سازی سیگموید
- **لایه خروجی:** ۱ نورون با تابع فعال‌سازی سیگموید

تعداد کل پارامترهای مدل برابر با ۱۶۰ است که به‌مراتب کمتر از محدودیت تکلیف (کمتر از ۵۰۰ پارامتر) می‌باشد.

بهینه‌سازهای مقایسه‌شده برای تضمین یک مقایسه کاملاً منصفانه بر اساس زمان اجرا، یک حلقه آموزش سفارشی در زبان Python پیاده‌سازی شد که زمان واقعی اجرا [۴] را در برابر مقدار تابع خطا ثبت می‌کند، برای روش‌های زیر:

- **SGD:** گرادیان نزولی تصادفی استاندارد (روش مرتبه اول)
- **L-BFGS:** روش شبه‌نیوتنی محدودحافظه Broyden–Fletcher–Goldfarb–Shanno
- **گرادیان مزدوج** [۵]: روشی تکرارشونده که با استفاده از جهت‌های مزدوج، همگرایی را تسریع می‌کند

۳-۲-۳ نتایج تجربی

شکل ۲: رقابت بهینه‌سازها (خطا در برابر زمان) شکل ۲ (در بالا) مقدار Log Loss را بر حسب زمان واقعی اجرا (بر حسب ثانیه) نمایش می‌دهد.

مشاهدات

- **SGD (خط آبی):** نرخ همگرایی آهسته و تقریباً خطی را نشان می‌دهد. پس از حدود ۰٫۳۵ ثانیه، مقدار خطا در حدود 10^{-1} متوقف شده و الگوریتم در کاهش بیشتر خطا با مشکل مواجه می‌شود.

[۳] (MLP یا Multilayer Perceptron)
[۴] (wall-clock time)
[۵] (CG یا Conjugate Gradient)

- **L-BFGS (خط نارنجی):** افتی بسیار چشمگیر و تقریباً «عمودی» دارد. در کمتر از ۰٫۵ ثانیه، مقدار خطا به‌شدت کاهش یافته و تقریباً بلافاصله به یک راه‌حل نزدیک به بهینه همگرا می‌شود.
- **گرادیان مزدوج (خط سبز):** این روش نیز به‌طور قابل‌توجهی از SGD بهتر عمل می‌کند و در حدود ۰٫۲ ثانیه به خطایی بسیار کوچک (کمتر از 10^{-4}) می‌رسد.

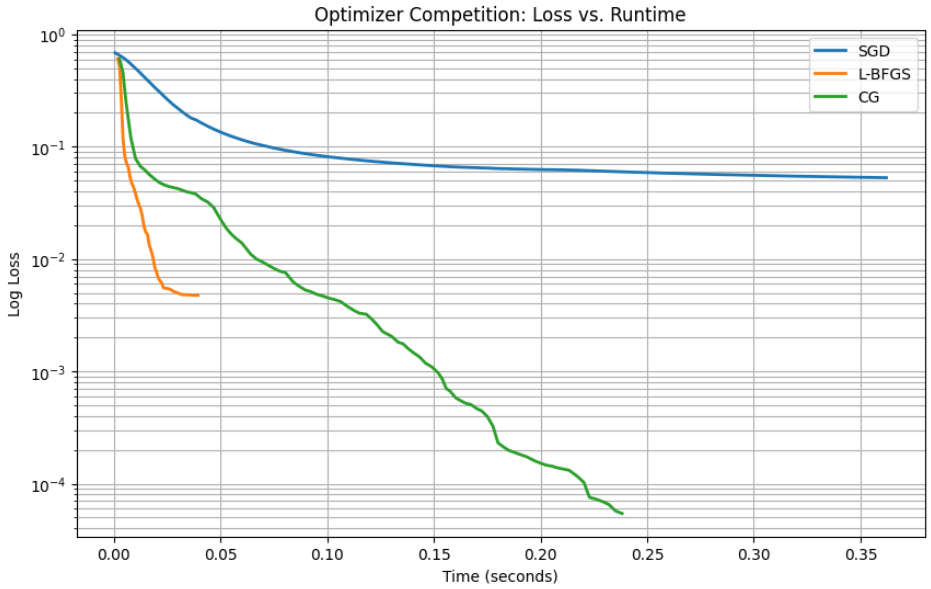
۳-۲-۴ تحلیل

قدرت اطلاعات خمیدگی افزایش سرعت چشمگیر در L-BFGS و CG ناشی از استفاده آن‌ها از اطلاعات خمیدگی (مشتقات مرتبه دوم) است. در حالی که SGD به‌صورت کورکورانه در جهت بیشترین شیب حرکت می‌کند، L-BFGS یک تقریب از ماتریس هسین می‌سازد. این امر به الگوریتم اجازه می‌دهد:

- **گام‌های بهینه بردارد:** به‌جای تکیه بر یک نرخ یادگیری ثابت، به‌طور تطبیقی تعیین می‌کند که در هر جهت چه مقدار حرکت لازم است.
- **دره‌ها را به‌درستی مدیریت کند:** با اصلاح اثر بدشرطی، از رفتار زیگ‌زاگی که موجب کندی SGD می‌شود جلوگیری می‌کند.

چرا «فضای نیوتنی»؟ این آزمایش مرزهای «فضای نیوتنی» را مشخص می‌کند. از آنجا که مدل تنها ۱۶۰ پارامتر دارد، هزینه محاسباتی تقریب معکوس هسین در مقایسه با مزیت برداشتن گام‌های هوشمندانه، ناچیز است. در این رژیم کم‌بُعد، «هوشمندی گام» به‌مراتب مهم‌تر از «ارزان بودن هر تکرار» در SGD است.

جمع‌بندی بخش دوم برای مسائل کوچک‌مقیاس (مجموعه‌داده‌های کوچک و شبکه‌های کم‌عمق)، روش‌های مرتبه دوم به‌طور قاطع برتر هستند. این روش‌ها نیازی به تنظیم ابرپارامترها (مانند نرخ یادگیری) ندارند و چندین مرتبه بزرگی سریع‌تر از SGD همگرا می‌شوند.



شکل ۳-۲: رقابت روش های بهینه سازی

۳-۳ یادگیری عمیق و تله مقیاس‌پذیری

۳-۳-۱ هدف

هدف این بخش، کمی‌سازی موانع محاسباتی مرتبط با بهینه‌سازی مرتبه دوم در یادگیری عمیق است. ما نشان می‌دهیم که اگرچه روش نیوتن از نظر هندسی برتر است (همان‌گونه که در بخش‌های ۱ و ۲ مشاهده شد)، اما برای مدل‌های با بُعد بالا از نظر محاسباتی غیرقابل اجراست.

۳-۳-۲ معماری مدل

ما یک شبکه عصبی عمیق [۶] را برای دسته‌بندی تصاویر مجموعه‌داده Fashion-MNIST تعریف کردیم.

- **ورودی:** ۷۸۴ ویژگی (۲۸ × ۲۸ پیکسل)
- **معماری:** سه لایه پنهان با ۱۰۰ نورون در هر لایه (تابع فعال‌سازی ReLU)، به‌دنبال آن یک لایه خروجی با ۱۰ کلاس

تعداد کل پارامترهای مدل برابر با $N \approx ۱۰۰,۰۰۰$ است که به‌صورت زیر محاسبه می‌شود:

[۶] (DNN یا Deep Neural Network)

$$۷۸۴ \times ۱۰۰ + ۱۰۰ \times ۱۰۰ + ۱۰۰ \times ۱۰۰ + ۱۰۰ \times ۱۰ \approx ۹۹,۷۱۰$$

که با در نظر گرفتن بایاس‌ها، تقریباً برابر با ۱۰۰,۰۰۰ پارامتر است.

۳-۳-۳ محاسبه حافظه هسین

برای اجرای «خالص» روش نیوتن، باید ماتریس هسین (H) را محاسبه و وارون کنیم؛ ماتریسی که شامل مشتقات مرتبه دوم تابع خطا نسبت به هر جفت از پارامترهاست.

ابعاد ماتریس

$$H \in \mathbb{R}^{N \times N} = ۱۰۰,۰۰۰ \times ۱۰۰,۰۰۰$$

تعداد کل درایه‌ها

$$N^۲ = ۱۰,۰۰۰,۰۰۰,۰۰۰ \quad \text{(ده میلیارد درایه)}$$

حافظه موردنیاز با فرض استفاده از اعداد ممیز شناور ۳۲ بیتی (۴ بایت برای هر درایه):

$$۱۰^{۱۰} \times ۴ = ۴۰ \ \square\square$$

نتیجه‌گیری ذخیره تنها یک ماتریس هسین برای این شبکه عمیق نسبتاً کوچک، به حدود ۴۰ گیگابایت حافظه RAM نیاز دارد. این مقدار از ظرفیت حافظه اغلب GPUهای مصرفی (معمولاً بین ۸ تا ۲۴ گیگابایت) فراتر است. افزون بر این، وارون‌سازی این ماتریس به عملیاتی با مرتبه

$$\mathcal{O}(N^۳)$$

نیاز دارد که برای $N = ۱۰۰,۰۰۰$ عملاً غیرممکن است.

۴-۳-۳ جایگزین‌ها: بهینه‌سازهای مرتبه اول

از آنجا که محاسبه دقیق اطلاعات خمیدگی بسیار پرهزینه است، ناچار به استفاده از روش‌های مرتبه اول هستیم. در این مطالعه، دو جایگزین رایج مقایسه شدند:

- **SGD به‌همراه مومنتوم:** از میانگین متحرک گرادیان‌ها برای شبیه‌سازی مفهوم سرعت استفاده می‌کند.
- **Adam:** از نرخ‌های یادگیری تطبیقی برای هر پارامتر بهره می‌برد و به‌طور ضمنی، خمیدگی قطری سطح خطا را تقریب می‌زند.

شکل ۳: مقایسه همگرایی (Adam در برابر SGD) شکل ۳ (در بالا) مقدار خطای Cross-Entropy را بر حسب تکرارهای آموزش نمایش می‌دهد.

مشاهدات

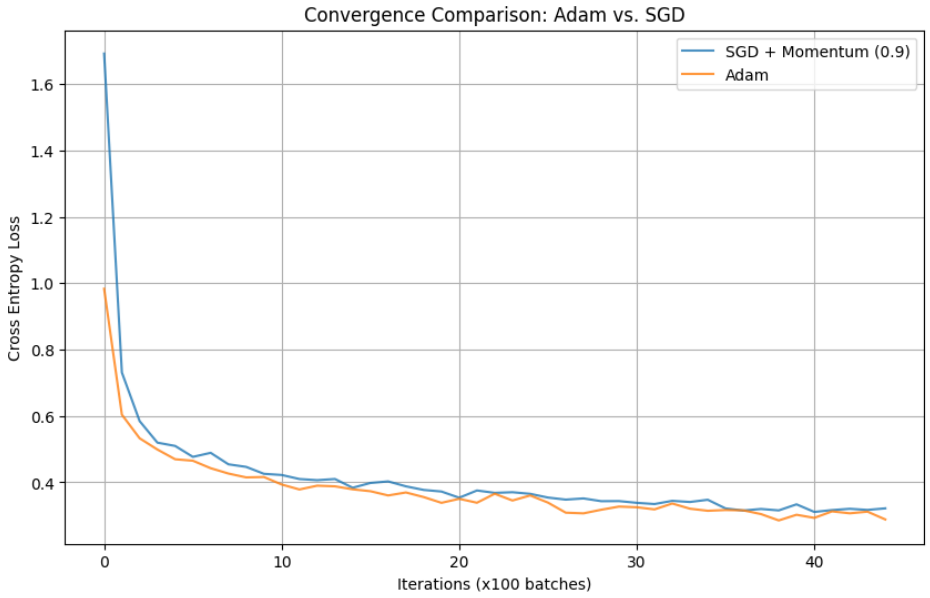
- **Adam (خط نارنجی):** افت اولیه بسیار سریعی را نشان می‌دهد و تنها در چند تکرار نخست، مقدار خطا را به کمتر از ۰٪۶ می‌رساند. ماهیت تطبیقی این روش اجازه می‌دهد حتی بدون تنظیم دقیق نرخ یادگیری سراسری، پیشرفت سریعی حاصل شود.
- **SGD به‌همراه مومنتوم (خط آبی):** در مراحل اولیه همگرایی کندتری دارد. اگرچه در نهایت به سطح عملکردی مشابه می‌رسد، اما فاقد سرعت اولیه «تهاجمی» روش Adam است.

۳-۳-۵ تحلیل

یادگیری عمیق در رژیمی قرار دارد که در آن تعداد پارامترها N آن‌قدر بزرگ است که استفاده از روش نیوتن غیرممکن می‌شود. در نتیجه، با یک «تله مقیاس‌پذیری» مواجه هستیم:

- **مدل‌های کوچک (بخش ۲):** می‌توان هزینه حافظه‌ای $\mathcal{O}(N^2)$ را پذیرفت، بنابراین از روش‌های نیوتنی یا شبه‌نیوتنی برای همگرایی تقریباً آنی استفاده می‌شود.
- **مدل‌های عمیق (این بخش):** تحمل هزینه $\mathcal{O}(N^2)$ ممکن نیست. در نتیجه، «گام کامل و ایده‌آل» با هزاران «گام ارزان» جایگزین می‌شود و از روش‌هایی مانند Adam بهره گرفته می‌شود.

الگوریتم Adam این شکاف را با برآورد قطری ماتریس هسین (از طریق ممان‌های مرتبه دوم گرادیان) پر می‌کند، بدون آنکه نیازی به ذخیره کل ماتریس هسین داشته باشد.



شکل ۳-۳: مقایسه همگرایی

۳-۴ عمودبودگی و تجزیه QR (شرطی‌سازی)

۳-۴-۱ هدف

بخش پایانی به بررسی یک جایگزین هندسی برای الگوریتم‌های پیچیده بهینه‌سازی می‌پردازد. به‌جای ارتقای بهینه‌ساز (برای مثال از SGD به روش نیوتن)، هدف ما بهبود هندسه خود مسئله است. فرضیه اصلی این است که با حذف هم‌بستگی میان ویژگی‌ها از طریق تجزیه QR، می‌توان به یک بهینه‌ساز ساده مانند گرادیان نزولی اجازه داد تا به سرعت‌های همگرایی قابل مقایسه با روش‌های مرتبه دوم دست یابد.

۳-۴-۲ روش‌شناسی

تولید داده یک مجموعه‌داده رگرسیونی مصنوعی با $N = ۱۰۰۰$ نمونه و دو ویژگی به‌شدت هم‌بسته تولید شد تا یک «دره بدشرط» شبیه‌سازی شود:

$$x_۱ \sim \mathcal{U}(۰, ۱)$$

$$x_۲ = x_۱ + \varepsilon, \qquad \varepsilon \sim \mathcal{N}(۰, ۰/۰۵)$$

عمودسازی تجزیه QR بر روی ماتریس ویژگی‌ها X اعمال شد:

$$X = QR$$

که در آن، Q یک پایه متعامد برای فضای ویژگی‌ها را نمایش می‌دهد.

گام مقیاس‌بندی حیاتی تجزیه QR استاندارد، بردارهای واحد (با نرم برابر با ۱) تولید می‌کند. برای تضمین یک مقایسه منصفانه با داده خام X (که واریانس بالاتری دارد)، ماتریس Q در $\sqrt{N}$ ضرب شد تا «انرژی کل» ویژگی‌ها با داده اولیه هم‌تراز شود. در ادامه، این داده با نماد $Q_{\square\square\square\square\square\square\square}$ مشخص می‌شود.

۳-۴-۳ نتایج تجربی

تحلیل هم‌بستگی

در گام نخست، ماتریس کوواریانس ویژگی‌های خام بررسی شد.

شکل ۴: ماتریس کوواریانس شکل ۴ (بالا) نقشه حرارتی ماتریس کوواریانس را نشان می‌دهد. خروجی، هم‌بستگی بسیار بالایی را تأیید می‌کند:

```
[08468374.0 08534424.0]]
[[08646821.0 08468374.0]]
```

درایه‌های خارج از قطر (حدود ۰٫۰۸۴۷) تقریباً با واریانس‌های روی قطر (حدود ۰٫۰۸۵۳) برابرند، که نشان می‌دهد x_1 و x_2 تقریباً هم‌خط هستند. این وضعیت دقیقاً همان هندسه «دره باریک» را ایجاد می‌کند که گرادیان نزولی را به دام می‌اندازد.

مقایسه همگرایی

یک مدل رگرسیون خطی با استفاده از گرادیان نزولی استاندارد، هم روی داده‌های خام X و هم روی داده‌های عمودشده ($Q_{\square\square\square\square\square\square\square}$) آموزش داده شد.

شکل ۵: تأثیر عمودبودگی بر همگرایی شکل ۵ (بالا) خطای میانگین مربعات (Mean Squared Error یا MSE) را بر حسب تعداد تکرارها مقایسه می‌کند.

مشاهدات

- **داده خام (خط قرمز چین‌دار):** خطا به‌تدریج کاهش می‌یابد. هم‌بستگی بالا، بهینه‌ساز را محدود کرده و آن را مجبور به برداشتن گام‌های کوچک و ناکارآمد می‌کند تا از نوسان عرضی در دره جلوگیری شود.
- **داده عمودشده (خط آبی پیوسته):** خطا تقریباً به‌صورت عمودی کاهش می‌یابد و در حدود ۵ تکرار به کف عددی می‌رسد.

۳-۴-۴ تحلیل

این آزمایش نشان می‌دهد که شرطی‌سازی داده‌ها از نظر ریاضی معادل پیش‌شرطی‌سازی بهینه‌ساز است.

تبدیل هندسی کانتورهای تابع هزینه برای داده خام X بیضی‌های کشیده هستند (عدد وضعیت $\kappa \gg 1$). با تبدیل داده‌ها به Q، عملاً ماتریس هسین در $(X^TX)^{-1}$ ضرب شده و به ماتریس همانی می‌رسد:

$$Q^TQ = I$$

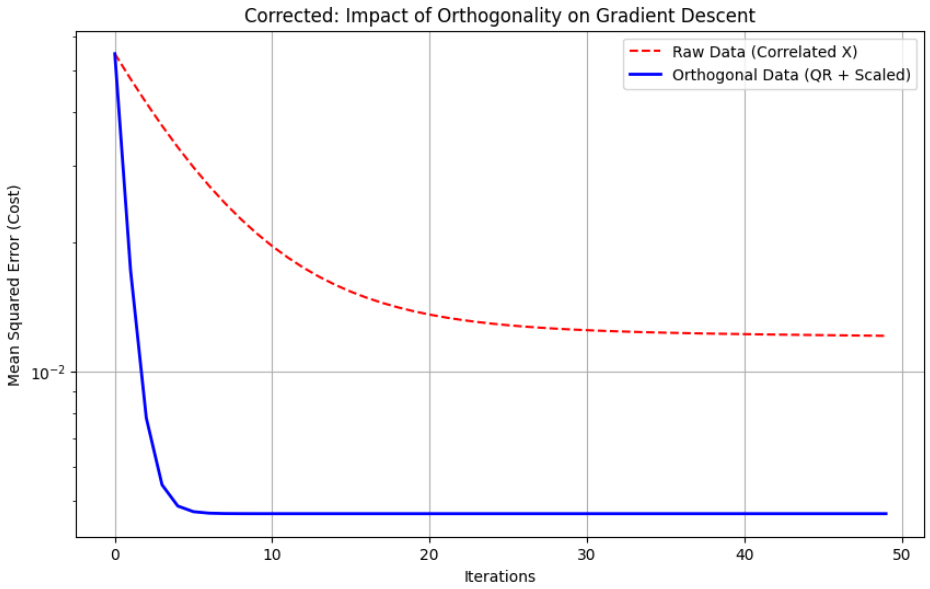
در نتیجه، عدد وضعیت برابر با

$$\kappa(I) = 1$$

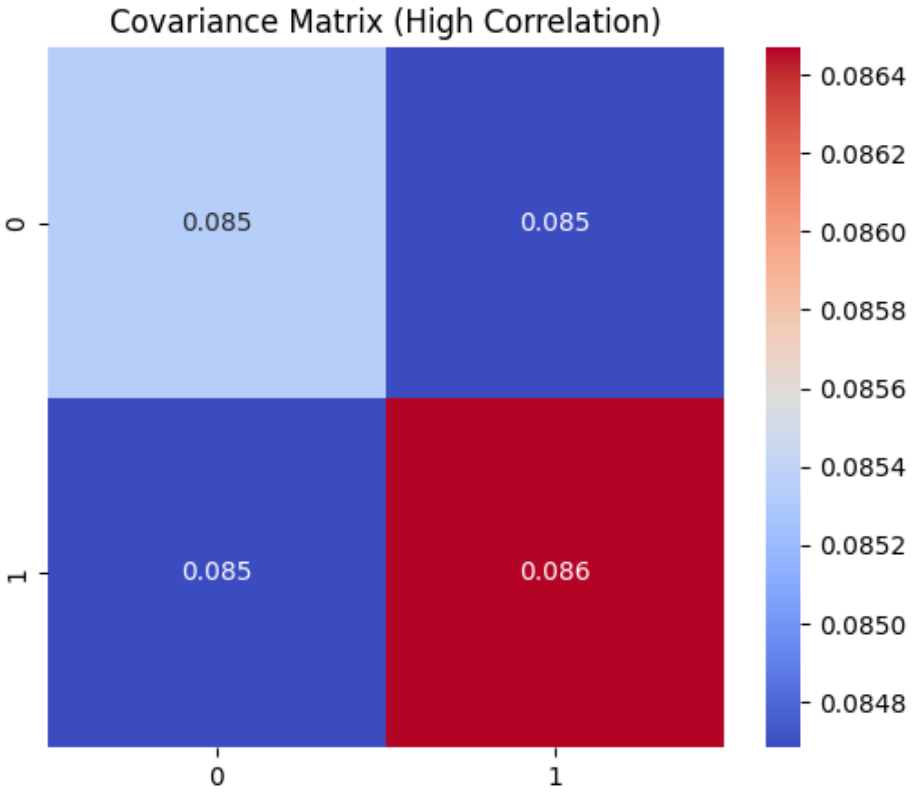
می‌شود. از دید هندسی، سطح خطا از یک دره باریک به یک ابرکره کامل تبدیل می‌گردد.

هم‌ارزی با روش نیوتن بر روی یک سطح کروی (عدد وضعیت برابر با ۱)، بردار گرادیان دقیقاً به سمت کمینه سراسری اشاره می‌کند. بنابراین، گرادیان نزولی بر روی داده‌های عمودشده، رفتاری کاملاً مشابه روش نیوتن بر روی داده‌های خام دارد:

- **روش نیوتن:** جهت گام را برای جبران هندسه سطح تغییر می‌دهد.
- **عمودسازی / QR:** هندسه مسئله را تغییر می‌دهد تا با جهت گام سازگار شود.



شکل ۳-۴: نسخه اصلاح شده ی تاثیر برروی گرادیان کاهشی



شکل ۳-۵: ماتریس کواریانس

فصل ۴

جمع‌بندی و نتیجه‌گیری و پیشنهادات

۴-۱ نتیجه‌گیری

این تکلیف به بررسی تنش بنیادین میان دقت ریاضی و امکان‌پذیری محاسباتی در بهینه‌سازی پرداخت. از طریق مجموعه‌ای از آزمایش‌ها از سطوح مصنوعی ساده تا شبکه‌های عصبی عمیق نشان دادیم که «سرعت» یادگیری صرفاً تابع الگوریتم مورد استفاده نیست، بلکه حاصل برهم‌کنش پیچیده‌ای میان منطق بهینه‌ساز و هندسه داده‌هاست.

۴-۱-۱ خلاصه یافته‌ها

هندسه شکست (بخش ۱) ما به‌صورت بصری تأیید کردیم که مسائل بدشرط که با دره‌های باریک مشخص می‌شوند پاشنه آشیل گرادیان نزولی هستند. در حالی که روش نیوتن می‌تواند با نرمال‌سازی خمیدگی به همگرایی آنی برسد، گرادیان نزولی استاندارد بخش قابل‌توجهی از انرژی محاسباتی خود را صرف نوسان (حرکت زیگ‌زاگی) در امتداد شیب‌های تند می‌کند.

جایگاه نیوتنی (بخش ۲) ما یک «رژیم نیوتنی» مشخص برای مسائل کم‌بُعد (برای مثال، MLPهای کوچک با کمتر از ۵۰۰ پارامتر) شناسایی کردیم. در این فضا، روش‌های مرتبه دوم مانند L-BFGS و گرادیان مزدوج [1] به‌مراتب برتر از SGD هستند و بدون نیاز به تنظیم ابرپارامترها، به خطایی نزدیک به صفر دست می‌یابند.

[1] (Conjugate Gradient)

دیوار مقیاس‌پذیری (بخش ۳) ما نشان دادیم که روش نیوتن «خالص» از نظر ریاضی ایده‌آل است، اما برای یادگیری عمیق از نظر محاسباتی غیرممکن می‌باشد. هزینه حافظه‌ای

$$\mathcal{O}(N^2)$$

برای ذخیره ماتریس هسین (در حدود ۴۰ گیگابایت برای یک شبکه سه‌لایه نسبتاً ساده) ناگزیر ما را به عقب‌نشینی به سمت روش‌های مرتبه اول سوق می‌دهد. در این میان، مشاهده کردیم که روش‌های تطبیقی مانند Adam یک سازش ضروری هستند که مزایای خمیدگی را بدون سربار حافظه‌ای شبیه‌سازی می‌کنند.

هندسه به‌عنوان راه‌حل (بخش ۴) در نهایت، نشان دادیم که همیشه برای حل بدشرطی به بهینه‌ساز پیچیده‌تر نیاز نیست. با عمودسازی داده‌های ورودی (از طریق تجزیه QR)، هندسه خود مسئله تغییر داده شد و به ساده‌ترین بهینه‌ساز (SGD) اجازه داده شد تا به سرعت همگرایی بهینه‌سازهای پیچیده دست یابد.

۴-۱-۲ جمع‌بندی نهایی

سیر تکامل بهینه‌سازی در یادگیری ماشین، روایت مواجهه با «تله مقیاس‌پذیری» است. اگرچه روش‌های مرتبه دوم از نظر تئوری ایده‌آل هستند، یادگیری عمیق مدرن در رژیمی عمل می‌کند که در آن محاسبه دقیق خمیدگی عملاً غیرقابل انجام است.

ازاین‌رو، موفقیت‌های پیشرفته امروزی بر یک راهبرد دوگانه تکیه دارند:

- **تقریب الگوریتمی:** استفاده از بهینه‌سازهایی مانند Adam که خمیدگی را (به‌صورت قطری) تخمین می‌زنند، نه اینکه آن را به‌طور دقیق محاسبه کنند.
- **پیش‌شرطی‌سازی هندسی:** به‌کارگیری تکنیک‌های معماری (مانند Batch Normalization یا عمودسازی نشان‌داده‌شده در بخش ۴) برای حفظ هندسه‌ای کروی در داده‌ها، به‌گونه‌ای که نیاز به گام‌های پیچیده بهینه‌سازی به حداقل برسد.

در پایان، نتیجه کلیدی این است که **هندسه بهتر داده‌ها از نظر محاسباتی معادل یک بهینه‌ساز بهتر است**.

منابع و مراجع

[1] Nocedal, Jorge and Wright, Stephen J. Numerical Optimization. Springer, New York, 2 ed. , 2006.

[2] Boyd, Stephen and Vandenberghe, Lieven. Convex Optimization. Cambridge University Press, Cambridge, 2004.

[3] Shewchuk, Jonathan Richard. An introduction to the conjugate gradient method without the agonizing pain. Technical Report, 1994.

[4] Golub, Gene H. and Van Loan, Charles F. Matrix Computations. Johns Hopkins University Press, Baltimore, 4 ed. , 2013.

[5] Higham, Nicholas J. Accuracy and Stability of Numerical Algorithms. SIAM, Philadelphia, 2 ed. , 2002.

Abstract

This report investigates the geometric and computational limitations of optimization algorithms in machine learning, ranging from simple quadratic surfaces to deep neural networks. Through a multi-stage analysis, we explore the fundamental trade-off between step efficiency (convergence rate) and scalability (memory and compute cost).First, we visually demonstrate the impact of ill-conditioning on Gradient Descent, highlighting its inefficient ”zig-zagging” behavior in high-curvature valleys, contrasted with Newton’s Method, which utilizes second-order information (the Hessian) to normalize curvature and achieve direct convergence. Second, we identify a ”Newtonian Regime” for small-dimensional problems (Breast Cancer dataset), where Quasi-Newton methods like L-BFGS and Conjugate Gradient significantly outperform Stochastic Gradient Descent (SGD) in wall-clock time.Third, we analyze the Scalability Trap inherent in Deep Learning. Theoretical calculations on a Fashion-MNIST Deep Network reveal that storing the Hessian matrix for even a modest architecture requires prohibitively large memory ($\sim$40 GB), rendering pure second-order methods infeasible. Consequently, we validate the necessity of first-order adaptive methods like Adam for high-dimensional spaces.Finally, we demonstrate that Data Orthogonalization offers a geometric alternative to complex optimizers. By applying QR Decomposition to correlated features, we reduce the problem’s condition number to unity, effectively allowing a simple SGD optimizer to mimic the convergence speed of Newton’s method. We conclude that while second-order methods offer superior theoretical convergence, modern deep learning relies on a compromise: utilizing first-order approximations combined with improved data geometry to navigate high-dimensional landscapes.

Key Words:

SGD, Newton Method, Hessian, Convergence rate

Amirkabir University of Technology
(Tehran Polytechnic)

Department of Comptuer Science

M. Sc. Thesis

Fourth CDM Project

By

Mohammad Sadegh Gholizadeh

Supervisor

Dr. Mahdi Ghatei