למידה עמוקה 046211
גיליון יבש 1

גל גרנות	315681593
ניר טבת	208731422

שאלה 1

נתחיל עם הביטוי משמאל ונשתמש ביצוג של הפונקציה כאינטגרל:

$$f(w_1) - f(w_2) - \nabla f(w_2)^T(w_1 - w_2)$$

$$= \int_0^1 \nabla f\big(w_2 + t(w_1 - w_2)\big)^T (w_1 - w_2)dt - \nabla f(w_2)^T(w_1 - w_2)$$

הביטוי השני קבוע ביחס לאינטגרל, ולכן:

$$= \int_0^1 \left(\nabla f\big(w_2 + t(w_1 - w_2)\big)^T (w_1 - w_2) - \nabla f(w_2)^T(w_1 - w_2)\right) dt$$

$$= \int_0^1 (\nabla f\big(w_2 + t(w_1 - w_2) - \nabla f(w_2)\big)^T (w_1 - w_2)dt$$

$$\leq \left| \int_0^1 (\nabla f\big(w_2 + t(w_1 - w_2) - \nabla f(w_2)\big)^T (w_1 - w_2)dt \right|$$

$$\leq \int_0^1 |(\nabla f\big(w_2 + t(w_1 - w_2) - \nabla f(w_2)\big)^T (w_1 - w_2)|\ dt$$

כאשר המעבר האחרון מתבצע לפי אי שוויון המשולש האינטגרלי. נשתמש בחלקות β על הגורם הראשון במכפלה באינטגרנד:

$$\leq \int_0^1 \beta|w_2 + t(w_1 - w_2) - w_2||w_1 - w_2|dt = |w_1 - w_2| \int_0^1 \beta t|w_1 - w_2|dt$$

$$= \beta|w_1 - w_2|^2 \int_0^1 tdt = \frac{\beta}{2}|w_1 - w_2|^2$$

כאשר השתמשנו במשפט קושי-שוורץ ביחס למכפלה פנימית אינטגרלית בין הפונקציות המתאימות.

שאלה 2

סעיף א':

הפונקציה f היא תבנית בילינארית עם מטריצה positive-definite, ולכן המינימום שלה ביחס ל-w הוא וקטור האפס. נמצא את הפרמטרים A, η עבורם האלגוריתם מתכנס בצעד אחד:

$$w(t+1) - w(t) = -\eta A \nabla f\big(w(t)\big) = -\eta A \nabla \frac{1}{2} w^T H w = -\eta A H w$$

עבור $A = H^{-1}$ (הקיימת מכיוון ש-H היא positive definite) נקבל:

$$w(t+1) = w(t) - \eta H^{-1} H w(t) = w(t) - \eta w(t) = (1-\eta) w(t) = 0 \Longrightarrow \eta = 1$$

כלומר, נבחר $A = H^{-1}$ ו-$\eta = 1$ לקבלת התכנסות בצעד בודד. נשים לב שעבור $H_{d\times d}$, חישוב H^{-1} הוא אלגוריתם בסיבוכיות $\theta(d^{2.373})$, ולכן אינו פרקטי ב-d גדול.

מה הקירוב המתאים??

סעיף ב':

נשתמש בהגדרות ה-GD וברמז:

$$w(t+1) = w(t) - \eta A \nabla f\big(w(t)\big)$$

$$\frac{\partial f}{\partial t} = \frac{\partial f}{\partial w}\frac{\partial w}{\partial t} = \nabla f\big(w(t)\big)^T \dot{w}(t) = \nabla f\big(w(t)\big)^T \Big(-A \nabla f\big(w(t)\big)\Big) =$$

$$= -\nabla f\big(w(t)\big)^T A \nabla f\big(w(t)\big)$$

נשתמש באי השוויון ברמז עם $v = \nabla f\big(w(t)\big) \in \mathbb{R}^d$ ונקבל כי מתקיים:

$$0 \le \lambda_{min} \big\|\nabla f\big(w(t)\big)\big\| \le -\frac{\partial f}{\partial t} = \nabla f\big(w(t)\big)^T A \nabla f\big(w(t)\big) \le \lambda_{max} \big\|\nabla f\big(w(t)\big)\big\|$$

כלומר, הגרדיאנט של f אי-חיובי ולכן f לא עולה. כמו כן היא לא מתבדרת לפי הנתון, ולכן מתכנסת לנקודה קריטית.

סעיף ג':

מאותם שיקולים נקבל:

$$f\big(w(t+1)\big) \le f\big(w(t)\big) + \big(w(t+1) - w(t)\big)^T \nabla f\big(w(t)\big) + \frac{\beta}{2}\|w(t+1) - w(t)\|^2$$

$$= -\eta\left(A\nabla f\big(w(t)\big)\right)^T \nabla f\big(w(t)\big) + \frac{\beta}{2}\left\|-\eta A f\big(w(t)\big)\right\|^2$$

$$= f\big(w(t)\big) - \eta A\left\|\nabla f\big(w(t)\big)\right\|^2 + \frac{\beta\eta^2}{2}\left\|\nabla f\big(w(t)\big)\right\|^2$$

$$= f\big(w(t)\big) - \eta\left(A - \frac{\beta\eta}{2}I\right)\left\|\nabla f\big(w(t)\big)\right\|^2$$

נסמן $A - \frac{\beta\eta}{2}I = C$ ונקבל:

$$f\big(w(t+1)\big) - f\big(w(t)\big) \le -C\left\|\nabla f\big(w(t)\big)\right\|^2 \Longrightarrow C\left\|\nabla f\big(w(t)\big)\right\|^2 \le f\big(w(t)\big) - f\big(w(t+1)\big)$$

התכונה מתקיימת לכל t. לכן, נבחר $T \in \mathbb{N}$ כלשהו ונקבל:

$$C\sum_{t=0}^{T}\left\|\nabla f\big(w(t)\big)\right\|^2 \le \sum_{t=0}^{T}\Big(f\big(w(t)\big) - f\big(w(t+1)\big)\Big) = f\big(w(0)\big) - f\big(w(T+1)\big)$$

$$\le f\big(w(0)\big) - \min_w f(w)$$

כאשר המעבר לפני האחרון מכיוון שהטור המתקבל באגף ימין הוא טלסקופי. ניקח $T \to \infty$ ונקבל:

$$0 \le C\sum_{t=0}^{\infty}\left\|\nabla f\big(w(t)\big)\right\|^2 \le f\big(w(0)\big) - \min_w f(w) < \infty$$

טור פונקציות מהצורה $C\sum_{t=0}^{\infty}\left\|\nabla f\big(w(t)\big)\right\|^2$ לא מתכנס אם האיבר הכללי שלו לא שואף ל-0, ולכן נסיק:

$$\lim_{t\to\infty}\left\|\nabla f\big(w(t)\big)\right\|^2 = 0 = \lim_{t\to\infty} \mathrm{C}\nabla f\big(w(t)\big) = 0$$

כלומר, ב-$t \to \infty$ האלגוריתם מתכנס לנקודה סטציונרית כנדרש.

שאלה 3

סעיף א':

ראשית מכלל השרשרת:

$$\frac{\partial L\big(w(t-1)\big)}{\partial \eta_{t-1}} = \frac{\partial L}{\partial w(t-1)} \cdot \frac{\partial w(t-1)}{\partial \eta_{t-1}} = (**)$$

עבור האיבר השמאלי:

$$\frac{\partial L}{\partial w(t-1)} = \nabla L\big(w(t-1)\big)$$

ועבור האיבר הימני:

$$w(t-1) = w(t-2) - \eta_{t-1}\nabla L\big(w(t-2)\big)$$

$$\frac{\partial w(t-1)}{\partial \eta_{t-1}} = -\nabla L(w(t-2))$$

וסך הכל קיבלנו:

$$(**) = \nabla L\big(w(t-1)\big) \cdot \Big(-\nabla L\big(w(t-2)\big)\Big) = -\nabla L\big(w(t-1)\big)^T \nabla L\big(w(t-2)\big)$$

סעיף ב':

באופן דומה לסעיף א', נשתמש בכלל השרשרת:

$$\frac{\partial \mathrm{L}\big(\mathrm{w}(\mathrm{t}-1)\big)}{\partial \alpha_{\mathrm{t}-1}} = \frac{\partial \mathrm{L}\big(\mathrm{w}(\mathrm{t}-1)\big)}{\partial \eta_{\mathrm{t}-1}} \cdot \frac{\partial \eta_{\mathrm{t}-1}}{\partial \alpha_{\mathrm{t}-1}} = (**)$$

האיבר השמאלי הוא כמובן התוצאה מסעיף א'. נחשב את האיבר הימני:

$$\eta_{t-1} = \eta_{t-2} - \alpha_{t-1} \cdot \frac{\partial L\big(w(t-2)\big)}{\partial \eta_{t-2}}$$

$$\frac{\partial \eta_{\mathrm{t}-1}}{\partial \alpha_{\mathrm{t}-1}} = -\frac{\partial L\big(w(t-2)\big)}{\partial \eta_{t-2}} = \nabla L\big(w(t-2)\big)^T \nabla L\big(w(t-3)\big)$$

כאשר המעבר האחרון הוא הזזה של תוצאת סעיף א' ב-1 והכפלה במינוס.

סך הכל קיבלנו:

$$(**) = \Big(-\nabla L\big(w(t-1)\big)^T \nabla L\big(w(t-2)\big)\Big) \cdot \Big(\nabla L\big(w(t-2)\big)^T \nabla L\big(w(t-3)\big)\Big)$$

סעיף ג':

נבצע כעת את כלל השרשרת הבא:

$$\frac{\partial L\big(w(t-1)\big)}{\partial \eta_{t-2}} = \frac{\partial L}{\partial w(t-1)} \cdot \frac{\partial w(t-1)}{\partial \eta_{t-1}} = \frac{\partial L}{\partial w(t-1)} \cdot \frac{\partial w(t-1)}{\partial w(t-2)} \cdot \frac{\partial w(t-2)}{\partial \eta_{t-2}}$$

כעת האיבר הראשון הוא:

$$\frac{\partial L}{\partial w(t-1)} = \nabla L\big(w(t-1)\big)$$

האיבר הימני, זהה לאיבר הימני מסעיף א' בהזזה של 1:

$$\frac{\partial w(t-1)}{\partial \eta_{t-1}} = -\nabla L\big(w(t-2)\big) \rightarrow \frac{\partial w(t-2)}{\partial \eta_{t-2}} = -\nabla L\big(w(t-3)\big)$$

נותר למצוא רק את האיבר האמצעי, מתוך הקשר:

$$w(t-1) = w(t-2) - \eta_{t-1} \nabla L\big(w(t-2)\big)$$

$$\frac{\partial w(t-1)}{\partial w(t-2)} = I - \eta_{t-1} \nabla^2 L\big(w(t-2)\big)$$

סך הכל קיבלנו:

$$(**) = \Big(\nabla L\big(w(t-1)\big)\Big) \cdot \Big(I - \eta_{t-1} \nabla^2 L\big(w(t-2)\big)\Big) \cdot \Big(-\nabla L\big(w(t-3)\big)\Big)$$

סעיף ד':

סעיף זה הוא הכללה של סעיף ג' למספר צעדים גדול, נחשב באותה דרך:

$$\frac{\partial L\big(w(t-1)\big)}{\partial \eta_{t-\tau}} = \frac{\partial L}{\partial w(t-1)} \cdot \frac{\partial w(t-1)}{\partial \eta_{t-\tau}}$$
$$= \frac{\partial L}{\partial w(t-1)} \cdot \frac{\partial w(t-1)}{\partial w(t-2)} \cdot \frac{\partial w(t-2)}{\partial w(t-3)} \cdot \frac{\partial w(t-3)}{\partial w(t-4)} \cdots \frac{\partial w(t-\tau)}{\partial \eta_{t-\tau}}$$

ניעזר בסעיף ג' ונאמר שעבור איברים כלליים:

$$\frac{\partial w(t-k)}{\partial w(t-k-1)} = I - \eta_{t-k}\nabla^2 L\big(w(t-k-1)\big)$$
$$\frac{\partial w(t-\tau)}{\partial \eta_{t-\tau}} = -\nabla L\big(w(t-\tau-1)\big)$$

נכפול בכל האיברים ונקבל:

$$(**) = \Big(\nabla L\big(w(t-1)\big)\Big) \cdot \left(\prod_{k=1}^{\tau-1} \Big(I - \eta_{t-k}\nabla^2 L\big(w(t-k-1)\big)\Big)\right) \cdot \Big(-\nabla L\big(w(t-\tau-1)\big)\Big)$$

סעיף ה':

עבור הגישה הראשונה שבה מחשבים כל מחזור פרמטר אחד:

- לכל עדכון יש סיבוכיות חישוב נמוכה שכן אנחנו מעדכנים רק פרמטר אחד
- קל יותר למימוש
- מתאים יותר לפונקציות שכדי להתכנס בהן גודל הצעד צריך להשתנות משמעותית בין סיבוב לסיבוב.

עבור הגישה השנייה שבה מחשבים כל מחזור T פרמטרים:

- לכל עדכון יש סיבוכיות חישוב גדולה שכן אנחנו מעדכנים T פרמטרים בו זמנית
- מסובך יותר למימוש
- מתאים יותר לפונקציות בהן גודל הצעד צריך להתחשב בשינויים רחבים יותר בפונקציה ולא רק בצעד האחרון. עבור T גדול מדי, ייתכן שיהיה קשה להתכנס כי גודל הצעד לא ישתנה במשך T מחזורים.

שאלה 4

נתונה הפונקציה:

$$f(x) = e^{e^x+e^{2x}} + \sin(e^x + e^{2x})$$

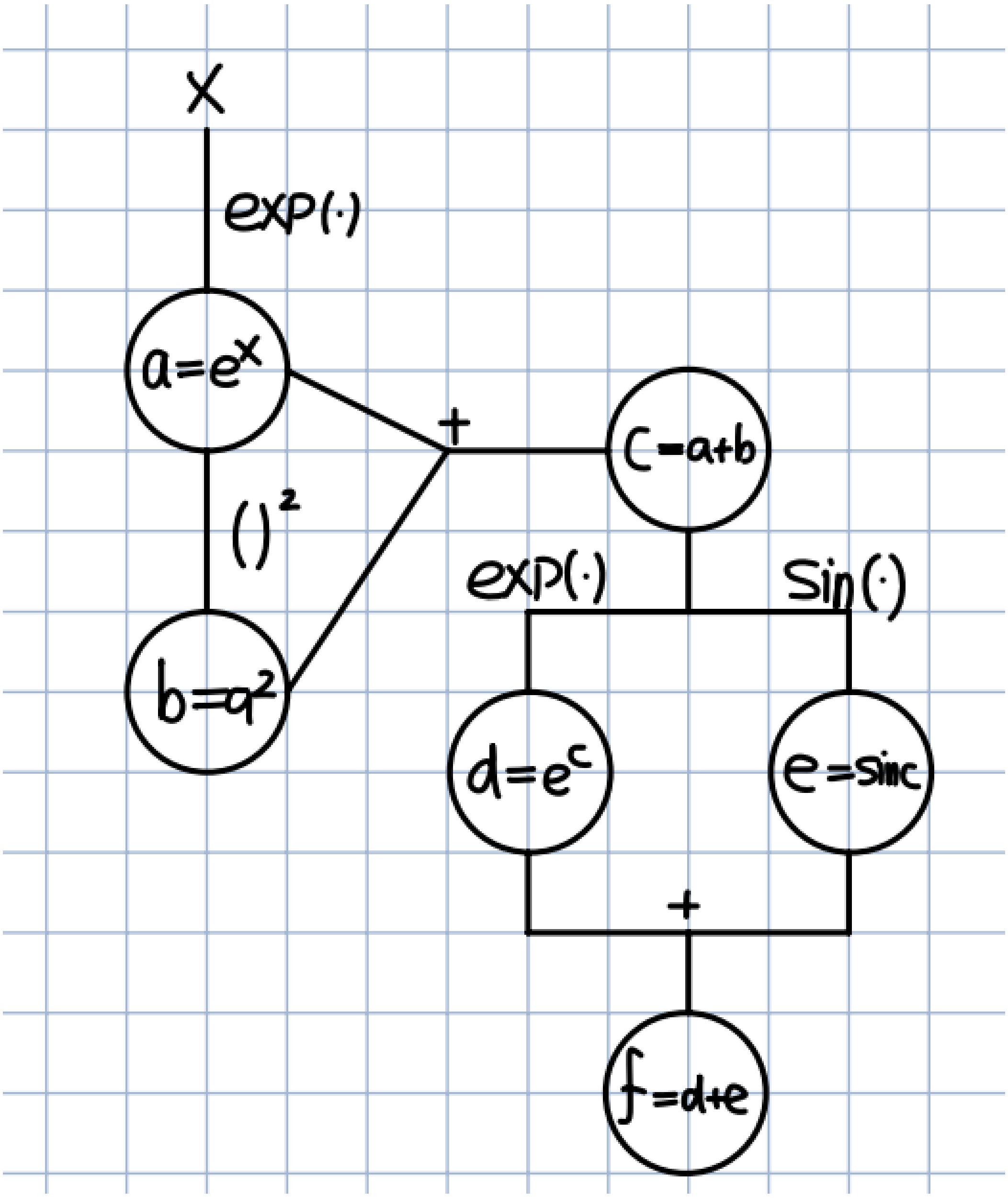
סעיף א':

נחשב ישירות את הנגזרת:

$$\frac{df}{dx} = (e^x + 2e^{2x})e^{e^x+e^{2x}} + (e^x + 2e^{2x})\cos(e^x + e^{2x})$$

$$= (e^x + e^{2x})\left(e^{e^x+e^{2x}} + \cos(e^x + e^{2x})\right)$$

סעיף ב':



סעיף ג':

מכלל השרשרת נקבל:

$$f = f(d, e) = d + e$$

$$e = e(c) = \sin c$$

$$d = d(c) = e^c$$

$$c = c(a, b) = a + b$$

$$b = b(a) = a^2$$

$$a = a(x) = e^x$$

ולכן:

$$\frac{df}{dx} = \frac{df}{dd}\frac{dd}{dc}\left(\frac{dc}{da}\frac{da}{dx} + \frac{dc}{db}\frac{db}{da}\frac{da}{dx}\right) + \frac{df}{de}\frac{de}{dc}\left(\frac{dc}{da}\frac{da}{dx} + \frac{dc}{db}\frac{db}{da}\frac{da}{dx}\right)$$

$$= \left(\frac{df}{dd}\frac{dd}{dc} + \frac{df}{de}\frac{de}{dc}\right)\left(\frac{dc}{da}\frac{da}{dx} + \frac{dc}{db}\frac{db}{da}\frac{da}{dx}\right)$$

$$= (1 \cdot e^c + 1 \cdot \cos c)(1 \cdot e^x + 1 \cdot 2a \cdot e^x) = \left(e^{a+b} + \cos(a+b)\right)(e^x + 2e^x e^x)$$

$$\left(e^{e^x + a^2} + \cos(e^x + a^2)\right)(e^x + 2e^{2x}) = \left(e^{e^x + e^{2x}}\right)(e^x + 2e^{2x})$$

כלומר, התוצאה המתקבלת זהה לתוצאה המתקבלת לפי החישוב הישיר. פרקטית, גם את המשתנה $a + b$ חישבנו יותר מפעם אחת ויכולנו לשמור בצד לחישוב יעיל יותר.