سوال ۱ بخش (الف)

فرض کنیم $x' \in S_\phi(x)$ یک نمونه مشابه دلخواه باشد. و $z' = f_\theta(x')$ بازنمایی ویژگی (feature representation) آن باشد. طبق تعریف ε داریم:

$$\|z - z'\|_\infty = \|z - f_\theta(x')\|_\infty \leq \varepsilon$$

که این یعنی $z' \in B_\infty(z, \varepsilon)$

z' عضو ℓ_∞ ball باشعاع ε به مرکزیت z است.

همچنین پیش بینی مدل برای x' برابر است با:

$$M(x') = \arg\max_i \{h_\psi^{(i)}(z')\}$$

همچنین ما می دانیم که $M(x) = y$ که یعنی:

$$y = \arg\max_i \{h_\psi^{(i)}(z)\}$$

از شرط (۱) که در فرض سوال داده شده است برای تمامی $y' \neq y$ و تمامی $z' \in B_\infty(z, \varepsilon)$ داریم:

$$h_\psi^{(y')}(z') - h_\psi^{(y)}(z') < 0$$

$$h_\psi^{(y')}(z') < h_\psi^{(y)}(z')$$

حال چون $z' = f_\theta(x') \in B_\infty(z, \varepsilon)$ داریم:

$$h_\psi^{(y')}(f_\theta(x')) < h_\psi^{(y)}(f_\theta(x')) \quad \text{for all} \quad y' \neq y$$

بنابراین چون امتیاز کلاس y اکیداً بزرگتر از هر کلاس دیگری در نقطهٔ x' است داریم:

$$M(x') = \arg\max_i \{h_\psi^{(i)}(f_\theta(x'))\} = y = M(x)$$

بنابراین حکم ثابت می شود چون برای هر ورودی $x' \in S_\phi(x)$ داریم $M(x') = M(x)$.

بخش (ب) نتیجهٔ اثبات شده در بخش (الف) انصاف فردی را تضمین می کند چرا که ورودی های مشابه، خروجی های مشابه دارند. (similar inputs receive similar (identical) outputs) اگر دو نمونه x' و x توسط تابع شباهت ϕ (یعنی $x' \in S_\phi(x)$) مشابه در نظر گرفته شوند، مدل تضمین می کند که آنها طبقه بندی یکسانی دارند. این بدان معناست که مدل با افرادی که طبق معیار انصاف ما مشابه هستند، به طور یکسان رفتار می کند و از تبعیض براساس تغییرات کوچک

در رویژگی ها جلوگیری می کند. این یک تضمین قوی برای انصاف است زیرا تضمین می کند که اختلالات یا تفاوت های جزئی بین افراد مشابه منجر به رفتار متفاوت توسط دسته بند نمی شود که لازمهٔ انصاف فردی است. در واقع اگر هر بردار ویژگی درون ℓ_∞-ball حول $z=f_\theta(x)$ با شعاع ε امتیاز مدل را برای کلاس اصلی y کاملاً بزرگتر از تمام امتیازات کلاس های دیگر نگه دارد، آنگاه هر ورودی x' که مشابه x باشد (پس بردار ویژگی آن درون آن ball قرار دارد) مانند x دسته بندی می شود. یعنی اگر دسته بند حاشیه ای برای کلاس صحیح داشته باشد که کل همسایگی ورودی های مشابه را حفظ کند، ورودی های مشابه را نمی توان به کلاس دیگری منتقل کرد.

بخش (ج) نقطهٔ ضعف اصلی در فضاهای ویژگی با ابعاد بالا مشکل نفرین ابعاد یا همان curse of dimensionality است: ناحیهٔ verification یا همان ناحیهٔ تایید به شکل نمایی یا exponentially بزرگ می‌شود. گوی ℓ_∞ یا همان $B_\infty(z,\varepsilon)$ در یک فضای k بعدی شامل حجمی متناسب با $(2\varepsilon)^k$ است. با افزایش ابعاد k، این ناحیه به صورت نمایی بزرگ می‌شود و فضای ویژگی بسیار بیشتری از ویژگی‌های نمونه‌های مشابه را پوشش می‌دهد. بنابراین ارضای شرط (۱) بسیار سخت‌تر می‌شود.

پیامدها: ۱. تضمین بیش از حد محافظه‌کارانه: این شرط باید برای همه نقاط در $B_\infty(z,\varepsilon)$ از جمله بسیاری از نقاطی که با هیچ نمونهٔ مشابه واقعی مطابقت ندارند برقرار باشد. این امر برآورده کردن تضمین را بسیار دشوارتر می‌کند.

۲. عدم قطعیت محاسباتی: تایید شرط (۱) نیاز به بررسی یک ناحیهٔ exponentially بزرگ را دارد که تایید را از نظر محاسباتی گران یا غیرممکن می‌کند.

۳. کاهش کاربردپذیری: با افزایش ابعاد، مدل‌های کمتری چنین شرط دقیق را برآورده می‌کنند و کاربرد عملی این تضمین انصاف را محدود می‌کند.

۴. کاهش دقت و انعطاف‌پذیری مدل: برای حفظ نابرابری (۱) در $B_\infty(z,\varepsilon)$ دسته‌بند باید بسیار robust باشد که طبق robustness tradeoff accuracy دقت را کاهش می‌دهد به‌خصوص در ابعاد بالا. حاشیه‌های بسیار زیاد بین کلاس‌ها انعطاف‌پذیری مدل را کم می‌کند و مرزهای تصمیم‌گیری محافظه‌کارانه ممکن است دقت را کاهش دهند. و در عمل نیز با افزایش k یافتن دسته‌بندی که شرط (۱) را ارضا کند و در عین حال عملکرد خوبی داشته باشد به طور فزاینده‌ای دشوار یا حتی غیرممکن است.

۵. تمرکز در جهت‌های متضاد: در ابعاد بالا، جهت‌های زیادی وجود دارند که تغییرات کوچک در آنها می‌تواند دسته را تغییر دهد (پدیده concentration of measure) حاشیه‌ای که در ابعاد پایین ممکن است امن باشد در ابعاد بالا ممکن است ناکافی باشد.

همچنین توجه کنید که این گزاره به استخراج‌کننده ویژگی f_θ متکی است. اگر f_θ لیپشیتز نباشد یا در برخی مختصات تغییرپذیری محلی بالایی داشته باشد، آنگاه ε محاسبه شده از شباهت ورودی ممکن است معیار ضعیفی از تغییرپذیری واقعی در فضای ویژگی باشد.

سوال ۲ بخش اول (الف) داریم: $\boxed{d_t := x_t - x^*}$ و $\xi_t = \eta_t \|x_t - x^*\|_2^2$

$$\boxed{\varepsilon_t = \eta_t \|d_t\|_2^2} \quad , \quad \tau_t = \frac{\varepsilon_t}{\|\nabla S_{x^*}(x_t)\|_2} \rightarrow \boxed{\tau_t = \frac{\eta_t \|d_t\|_2^2}{\|\nabla S_{x^*}(x_t)\|_2}}$$

از معادلهٔ (۵) داریم: $x_{t+1} = \alpha_t x^* + (1-\alpha_t)\left[x_t + \tau_t \nabla S_{x^*}(x_t)\right]$ بنابراین:

$$d_{t+1} = x_{t+1} - x^* = \alpha_t x^* + (1-\alpha_t)\left[x_t + \tau_t \nabla S_{x^*}(x_t)\right] - x^*$$

$$d_{t+1} = \alpha_t x^* - x^* + (1-\alpha_t) x_t + (1-\alpha_t)\tau_t \nabla S_{x^*}(x_t)$$

$$= -(1-\alpha_t) x^* + (1-\alpha_t) x_t + (1-\alpha_t)\tau_t \nabla S_{x^*}(x_t)$$

$$= (1-\alpha_t)\Big[\underbrace{x_t - x^*}_{d_t} + \tau_t \nabla S_{x^*}(x_t)\Big] = (1-\alpha_t)\big[d_t + \tau_t \nabla S_{x^*}(x_t)\big]$$

توجه کنید $\eta_t = t^{-q}$ زمان بندی اندازه گام انتخاب شده را مشخص می کند و این اتحادات جبری را تغییر نمی دهد.

$\Downarrow \|\cdot\|_2^2$

حال از طرفین معادله نرم ۲ به توان ۲ می گیریم:

$$\|d_{t+1}\|_2^2 = \|\underbrace{(1-\alpha_t)}_{\text{scaler}}[d_t + \tau_t \nabla S_{x^*}(x_t)]\|_2^2$$

$$\boxed{\|d_{t+1}\|_2^2 = (1-\alpha_t)^2 \|\tau_t \nabla S_{x^*}(x_t) + d_t\|_2^2}$$

توجه کنید scaler از عملگر $\|\cdot\|$ به شکل قدر مطلق خارج می شود و چون توان ۲ داریم می توان علامت قدر مطلق را حذف کرد.

بخش (ب): برای اثبات عبارت $\|\tau_t \nabla S_{x^*}(x_t) + d_t\|_2^2$ را بسط می‌دهیم:

$$\|\underbrace{\tau_t}_{\text{Scaler}} \nabla S_{x^*}(x_t) + d_t\|_2^2 = \tau_t^2 \|\nabla S_{x^*}(x_t)\|_2^2 + 2\tau_t \langle d_t, \nabla S_{x^*}(x_t)\rangle + \|d_t\|_2^2 \quad (*)$$

توجه کنید که $\boxed{\tau_t = \frac{\varepsilon_t}{\|\nabla S_{x^*}(x_t)\|_2} = \frac{\eta_t \|d_t\|_2}{\|\nabla S_{x^*}(x_t)\|_2}}$ $(**)$ چرا که $\varepsilon_t = \eta_t \|x_t - x^*\|_2 = \eta_t \|d_t\|_2$

همچنین

$$r_t = \frac{\langle x_t - x^*, \nabla S_{x^*}(x_t)\rangle}{\|x_t - x^*\|_2 \|\nabla S_{x^*}(x_t)\|_2} = \frac{\langle d_t, \nabla S_{x^*}(x_t)\rangle}{\|d_t\|_2 \|\nabla S_{x^*}(x_t)\|_2}$$

$$\Longrightarrow \boxed{\langle d_t, \nabla S_{x^*}(x_t)\rangle = r_t \|d_t\|_2 \|\nabla S_{x^*}(x_t)\|_2} \quad (***)$$

حال $(**)$ و $(***)$ را در $(*)$ جایگذاری می‌کنیم:

$$\|\tau_t \nabla S_{x^*}(x_t) + d_t\|_2^2 = \frac{\eta_t^2 \|d_t\|_2^2}{\|\nabla S_{x^*}(x_t)\|_2^2} \|\nabla S_{x^*}(x_t)\|_2^2 + 2\frac{\eta_t \|d_t\|_2}{\|\nabla S_{x^*}(x_t)\|_2} \cdot r_t \|d_t\|_2 \|\nabla S_{x^*}(x_t)\|_2 + \|d_t\|_2^2$$

$$\|\tau_t \nabla S_{x^*}(x_t) + d_t\|_2^2 = \eta_t^2 \|d_t\|_2^2 + 2\eta_t r_t \|d_t\|_2^2 + \|d_t\|_2^2$$

$$\boxed{\|\tau_t \nabla S_{x^*}(x_t) + d_t\|_2^2 = \|d_t\|_2^2 (\eta_t^2 + 2\eta_t r_t + 1)}$$

حال آن را در داخل رابطه (۶) قرار می‌دهیم:

$$\|d_{t+1}\|_2^2 = (1-\alpha_t)^2 \|\tau_t \nabla S_{x^*}(x_t) + d_t\|_2^2 = (1-\alpha_t)^2 \|d_t\|_2^2 (\eta_t^2 + 2\eta_t r_t + 1)$$

$\xrightarrow{\text{تقسیم طرفین بر } \underbrace{\|d_t\|_2^2 > 0}_{\|x_t - x^*\|_2^2 > 0}}$

$$\boxed{\frac{\|d_{t+1}\|_2^2}{\|d_t\|_2^2} = (1-\alpha_t)^2 (\eta_t^2 + 2\eta_t r_t + 1)}$$

حال از رابطهٔ (۷) استفاده می‌کنیم:

$$\downarrow \quad (1-\alpha_t)^2 \le \left(\frac{r_t + \eta_t K_t}{r_t + \eta_t (1+K_t)}\right)^2$$

$$\frac{\|d_{t+1}\|_2^2}{\|d_t\|_2^2} \le \left(\frac{r_t + \eta_t K_t}{r_t + \eta_t (1+K_t)}\right)^2 (\eta_t^2 + 2\eta_t r_t + 1) =: \theta_t$$

که دقیقاً رابطه (۸) است.

بخش ج) ۱. توجه کنید که $\eta_t := t^{-q}$ برای $q \in (\frac{1}{2}, 1)$. حال با توجه به اینکه $\eta_t > r_t$ بیاید فاکتور اول θ_t را آنالیز کنیم. فرض کنید $a = r_t + \eta_t \kappa_t$ و $b = r_t + \eta_t(1+\kappa_t)$:

$$\frac{a}{b} = \frac{r_t + \eta_t \kappa_t}{r_t + \eta_t(1+\kappa_t)} = \frac{r_t + \eta_t \kappa_t}{r_t + \eta_t + \eta_t \kappa_t} < 1 \Rightarrow \left(\frac{a}{b}\right)^2 < 1 \longrightarrow$$

فاکتور اول کوچکتر از ۱ است.

به بیان دقیقتر برای $\eta_t > r_t$ یا همان $\frac{\eta_t}{r_t} > 1$ داریم:

$$\frac{a}{b} = \frac{r_t(1 + \frac{\eta_t}{r_t}\kappa_t)}{r_t(1 + \frac{\eta_t}{r_t}(1+\kappa_t))} = \frac{1 + \lambda_t \kappa_t}{1 + \lambda_t(1+\kappa_t)} \quad \text{where } \lambda_t = \frac{\eta_t}{r_t} > 1$$

چون κ_t محدود است ($\kappa_t \le K$ برای یک ثابت K) داریم:

$$\frac{a}{b} \le \frac{1 + \lambda_t K}{1 + \lambda_t(1+K)} = \frac{1 + \lambda_t K}{1 + \lambda_t + \lambda_t K} < 1 \longrightarrow \text{فاکتور اول} = \left(\frac{a}{b}\right)^2 < 1$$

چون $\lambda_t > 1$ و $K > 0$ پس صورت کوچکتر از مخرج است و در نتیجه $\frac{a}{b} < 1$ است.

حال بیایید فاکتور دوم θ_t یعنی $\eta_t^2 + 2\eta_t r_t + 1$ را آنالیز کنیم. توجه کنید چون $t \to \infty$ پس $\eta_t \to 0$ پس $r_t \to 0$ در نتیجه فاکتور دوم صفر می‌شود.

$$\left.\begin{array}{l} \eta_t := t^{-q} \\ q \in (\frac{1}{2}, 1) \end{array}\right\} \Rightarrow \lim_{t\to\infty} \eta_t = 0 \longrightarrow \lim_{t\to\infty} \eta_t^2 + 2\eta_t r_t + 1 = 1$$

همچنین چون $r_t < \eta_t$ پس فاکتور دوم صفر می‌شود.

پس داریم:

$$\theta_t = \underbrace{\left(\frac{r_t + \eta_t \kappa_t}{r_t + \eta_t(1+\kappa_t)}\right)^2}_{\text{عددی کوچکتر از یک است (فاکتور اول)}} \underbrace{(\eta_t^2 + 2\eta_t r_t + 1)}_{\text{وقتی } t\to\infty \text{ به } 1 \text{ میل می‌کند. (فاکتور دوم)}}$$

بنابراین چون فاکتور اول اکیداً کوچکتر از 1 هست و فاکتور دوم به 1 میل می‌کند وقتی $t \to \infty$ وجود دارد یک $c_1 > 0$ به طوری که

$$\boxed{\theta_t < 1 - c_1}$$

به طور دقیق‌تر از آنجایی که $\frac{a}{b} < 1$ با حاشیه‌ای که به $\eta_t - r_t$ بستگی دارد و κ_t محدود است ($0 \le \kappa_t \le K$) پس $\left(\frac{a}{b}\right)^2 < 1 - \varepsilon$ برای یک $\varepsilon > 0$ ثابت c_1 را که $\theta_t < 1 - c_1$ می‌دهد.

۲. از آنجایی که $S_{x^*}(x_t) > 0$ برای همه iteration های الگوریتم (الگوریتم feasibility را حفظ می‌کند) و فواصل $\|d_t\|_2$ محدود باقی می‌ماند (چرا که ما در حال بهینه‌سازی روی یک دامنه فشرده $[0,1]^d$ هستیم)، حاصل ضرب $\prod_{t=1}^{\infty} \theta_t$ نمی‌تواند به صفر همگرا شود زیرا این به معنای $\|d_t\|_2 \to 0$ است یعنی $x_t \to x^*$ که $S_{x^*}(x_t) > 0$ را با پیوستگی نقض می‌کند.

ادامه بخش د) ۱. که در آن تقریباً $c_1 \approx 2$ و c_2 برای جملات مرتبه بالاتر در نظر گرفته شده که با محاسبات جبری به دست می‌آید.

۱	۴	۷
۲	۵	جمع نمرات
۳	۶	

دانشگاه صنعتی شریف

شماره دانشجویی:

نام و نام خانوادگی:

شماره صندلی: نام استاد:

شماره سوال: ۳. با برهان خلف فرض کنید تعداد نامتناهی اندیس وجود دارد که $\eta_t > r_t$ را برآورده می‌کند و در آنها صدق می‌کند. آنگاه:

$$\prod_t \theta_t = \prod_{t\in T}\theta_t \cdot \prod_{t\notin T}\theta_t \quad (t\in T \text{ برای } \theta_t < 1-c_1)$$

$$\prod_{t=1}^{\infty}\theta_t \leqslant \prod_{t\in T}(1-c_1)\cdot\prod_{t\notin T}\theta_t = 0\cdot\prod_{t\notin T}\theta_t = 0 \quad \text{※}$$

با رابطهٔ (۹)

که در آن T مجموعه نامتناهی از اندیس‌های با $\eta_t > r_t$ است. از آنجایی که $1 > (1-c_1)$ و ما یک حاصل ضرب نامتناهی در نظر می‌گیریم پس $\prod_{t\in T}(1-c_1) = 0$ که با معادلهٔ (۹) در تناقض است. ※ بنابراین فرض خلف باطل است و فقط تعداد نامتناهی از اندیس‌ها $\eta_t > r_t$ را برآورده می‌کنند.

بخش د) ۱. بسط دادن θ_t با $\lambda_t = \frac{\eta_t}{r_t} \leqslant 1$:

$$\theta_t = \underbrace{\left(\frac{r_t+\eta_t K_t}{r_t+\eta_t(1+K_t)}\right)^2}_{\text{فاکتور ۱}}\underbrace{(\eta_t^2+2\eta_t K_t+1)}_{\text{فاکتور ۲}}$$

حال از r_t در فاکتور اول فاکتور می‌گیریم:

$$\frac{r_t+\eta_t K_t}{r_t+\eta_t(1+K_t)} = \frac{r_t(1+\lambda_t K_t)}{r_t(1+\lambda_t(1+K_t))} = \frac{1+\lambda_t K_t}{1+\lambda_t(1+K_t)}$$

برای مقادیر کوچک $\lambda_t \leqslant 1$ با استفاده از بسط تیلور (Taylor $(1+x)^{-1} \approx 1-x+x^2$) خواهیم داشت:

$$\frac{1+\lambda_t K_t}{1+\lambda_t(1+K_t)} = (1+\lambda_t K_t)(1-\lambda_t(1+K_t)+O(\lambda_t^2)) = 1+\lambda_t K_t-\lambda_t(1+K_t)+O(\lambda_t^2) = 1-\lambda_t+O(\lambda_t^2)$$

حال به مربع گرفتن داریم:

$$\left(\frac{1+\lambda_t K_t}{1+\lambda_t(1+K_t)}\right)^2 = (1-\lambda_t+O(\lambda_t^2))^2 = 1-2\lambda_t+O(\lambda_t^2)$$

همچنین برای فاکتور دوم داریم:

$$\eta_t^2+2\eta_t r_t+1 = 1+2\eta_t r_t+\eta_t^2 = 1+2\lambda_t r_t^2+\lambda_t^2 r_t^2$$

حال آنها را ترکیب می‌کنیم:

$$\theta_t = (1-2\lambda_t+O(\lambda_t^2))(1+2\lambda_t r_t^2+\lambda_t^2 r_t^2) = 1+2\lambda_t r_t^2-2\lambda_t+O(\lambda_t^2)$$

$$\theta_t = 1-2\lambda_t(1-r_t^2)+O(\lambda_t^2) \xrightarrow[\text{پس } \lambda_t^2 = \frac{\eta_t^2}{r_t^2}]{\lambda_t = \frac{\eta_t}{r_t} \text{ چون}} \boxed{\theta_t \leqslant 1-c_1\lambda_t(1-r_t^2)+c_2\eta_t^2}$$

ادامه توضیح بالای صفحه

بخش د) ۲. رابطه (۹) $\prod_{t=1}^{\infty}\theta_t>0 \xrightarrow{\log} \sum_{t=1}^{\infty}\log(\theta_t)>-\infty$

حال بیایید از نامساوی $\log(1+x)\geq x$ برای $x\geq -1$ استفاده کنیم:

$$\sum_{t=1}^{\infty}\log(\theta_t)\geq\sum_{t=1}^{\infty}(\theta_t-1)\geq\sum_{t=1}^{\infty}\left(1-c_1\lambda_t(1-r_t^2)+c_2\eta_t^2-1\right)$$

$1+x=\theta_t$ و $\theta_t-1\geq -1$

استفاده از رابطه (۱۰)

$$\Rightarrow \sum_{t=1}^{\infty}\log(\theta_t)\geq\sum_{t=1}^{\infty}-c_1\lambda_t(1-r_t^2)+c_2\eta_t^2$$

$$\underbrace{\sum_{t=1}^{\infty}\log(\theta_t)}\geq-\sum_{t=1}^{\infty}\left(c_1\lambda_t(1-r_t)^2-c_2\eta_t^2\right)$$

$\sum_{t=1}^{\infty}\log(\theta_t)>-\infty$ عبارت سمت چپ محدود است.

$$-\sum_{t=1}^{\infty}\log(\theta_t)<\infty \quad , \quad -\sum_{t=1}^{\infty}\log(\theta_t)\leq\sum_{t=1}^{\infty}\left(c_1\lambda_t(1-r_t)^2-c_2\eta_t^2\right)$$

$$\Downarrow$$

$$\sum_{t=1}^{\infty}c_1\lambda_t(1-r_t)^2-c_2\eta_t^2<\infty$$

جایگذاری $\lambda_t=\frac{\eta_t}{r_t}$

$$\sum_{t=1}^{\infty}c_1\eta_t\frac{(1-r_t)^2}{r_t}-c_2\eta_t^2<\infty$$

اثبات شرط حکم!

توجه: در صورت سوال در مخرج از r_t^2 استفاده شده که چون r_t محدود است تفاوتی ندارد.

بخش ه) چون $\eta_t = t^{-q}$ with $q \in (\frac{1}{2}, 1)$ و داریم

$\sum_{t=1}^{\infty} \eta_t = \sum t^{-q} = \infty$ (سری هارمونیک واگرا می‌شود.)

ولی از رابطه (۱۱) اگر $\sum_{t=1}^{\infty} \left(c_1 \eta_1 \frac{1-r_t^2}{r_t^2} - c_2 \eta_t^2 \right) > \infty$

چون $\sum \eta_t^2 = \sum t^{-2q} < \infty$ که برای $q > \frac{1}{2}$ همگرا می‌شود و ما هم داریم

$2q \in (1, 2)$ چون $q \in (\frac{1}{2}, 1)$ بنابراین نیاز داریم که

$$\sum_{t=1}^{\infty} \frac{\eta_t (1-r_t^2)}{r_t^2} < \infty \qquad \eta_t = t^{-q} = O(t^{-q})$$

که به این منجر می‌شود که:

$$\frac{(1-r_t^2)}{r_t^2} = O\left(\frac{1}{\eta_t}\right) = O(t^q)$$

حال چون $r_t \to 1$ پس $r_t^2 \to 1$ همچنین می‌توان صورت را به این شکل

نوشت: $1 - r_t^2 = (1-r_t)\underbrace{(1+r_t)}_{2} \approx 2(1-r_t)$

بنابراین باید $1 - r_t = O(t^{q-1})$ باشد.

$$1 - r_t = O(t^{-(1+q+\varepsilon)}) = O(t^{q-1}) \quad \text{for some } \varepsilon > 0$$

چون ε کوچک است.

$$1 - r_t = O(t^{q-1})$$

۱	۴	۷
۲	۵	جمع نمرات
۳	۶	

دانشگاه صنعتی شریف

شماره دانشجویی:

نام و نام خانوادگی:

شماره صندلی: نام استاد:

شماره سوال: بخش دوم (تخمین گرادیان) بخش الف)

به علت داشتن خاصیت Lipschitz برای ∇S_{x^*} داریم:

$$\boxed{\|\nabla S_{x^*}(x') - \nabla S_{x^*}(x_t)\|_2 \leq L\|x' - x_t\|_2}$$

چون x' نقطه‌ای روی خط مابین x_t و $x_t + \delta u$ است داریم:

$$\|x' - x_t\|_2 \leq \|\delta u\|_2 = \delta \underbrace{\|u\|_2}_{1} = \delta$$

δu = حداکثر فاصله

توجه: چون بردار u روی کره واحد است پس $\|u\|_2 = 1$

چون $\delta \in (0,1)$ یک scalar است از عملگر $\|\cdot\|_2$ بیرون می‌آید.

$$\boxed{\|x' - x_t\|_2 \leq \delta}$$

همچنین مشتق دوم در جهت بردار u برابر است با:

$$\frac{d}{dh}\left(u^T \nabla S_{x^*}(x' + hu)\right)\Big|_{h=0} = u^T \nabla^2 S_{x^*}(x')\, u$$

و بنابراین طبق نامساوی Cauchy-Schwarz و خاصیت Lipschitz داریم:

$$\left|\underbrace{u^T}_{\text{قسمت اول}} \underbrace{\nabla^2 S_{x^*}(x') u}_{\text{قسمت دوم}}\right| \leq \|u\|_2 \, \|\nabla^2 S_{x^*}(x') u\|_2$$

بخش دوم (تخمین گرادیان) بخش الف) طبق تعریف مشتق دوم در جهت u داریم:

$$u^T \nabla^2 S_{x^*}(x') u = \frac{d}{dh}\left(u^T \nabla S_{x^*}(x' + hu)\right)\Big|_{h=0}$$

چون ∇S_{x^*} یک تابع Lipschitz با ضریب L است داریم:

$$\|\nabla S_{x^*}(x' + hu) - \nabla S_{x^*}(x')\|_2 \leq L\|x' + hu - x'\|_2 = L\|hu\|_2 = Lh\underbrace{\|u\|_2}_{1} = Lh$$

$\|u\|_2 = 1$ چرا که روی کره واحد است. $h > 0$ فرض کنیم

$$\Longrightarrow \boxed{\|\nabla S_{x^*}(x' + hu) - \nabla S_{x^*}(x')\|_2 \leq Lh}$$

حال طبق نامساوی Cauchy-Schwarz داریم:

$$\left|\underbrace{u^T}_{\text{قسمت اول}} \underbrace{\left(\nabla S_{x^*}(x' + hu) - \nabla S_{x^*}(x')\right)}_{\text{قسمت دوم}}\right| \leq \underbrace{\|u\|_2}_{\|u\|_2 = 1} \cdot \underbrace{\|\nabla S_{x^*}(x' + hu) - \nabla S_{x^*}(x')\|_2}_{\text{بالا نشان دادیم که } \leq Lh} \leq Lh$$

توجه کنید اگر $h<0$ هم باشد تفاوتی ندارد و فقط $|h|=-h$ و به همین شکل و با $h\to 0$ به همین صورت اثبات می‌شود.

بنابراین با تقسیم طرفین بر $h>0$ داریم: (توجه کنید که $|h|=h$ و فرقی ندارد که بر h یا $|h|$ تقسیم کنیم.)

$$\left|\frac{u^T(\nabla S_{x^*}(x'+hu)-\nabla S_{x^*}(x'))}{h}\right|\leq L$$

حال با گرفتن حد در $h\to 0$ داریم:

$$\lim_{h\to 0}\left|\frac{u^T(\nabla S_{x^*}(x'+hu)-\nabla S_{x^*}(x'))}{h}\right|\leq L \longrightarrow |u^T\nabla S_{x^*}(x')u|\leq L$$

حال چون $\delta\in(0,1)$ پس $\frac{1}{2}\delta^2>0$ و با ضرب طرفین در $\frac{1}{2}\delta^2>0$ داریم:

$$\frac{1}{2}\delta^2|u^T\nabla S_{x^*}(x')u|\leq\frac{1}{2}\delta^2 L \longrightarrow \boxed{\left|\frac{1}{2}\delta^2 u^T\nabla S_{x^*}(x')u\right|\leq\frac{1}{2}\delta^2 L}$$

بخش ب) ۱. فرض سوال: $\boxed{\nabla S_{x^*}(x_t)^T u > w = \frac{1}{2}L\delta}$

طبق رابطه ۵ داریم:

$$S_{x^*}(x_t+\delta u)=\underbrace{S_{x^*}(x_t)}_{0}+\delta\nabla S_{x^*}(x_t)^T u+\frac{1}{2}\delta^2 u^T\nabla^2 S_{x^*}(x')u$$

$$\longrightarrow S_{x^*}(x_t+\delta u)=\delta\underbrace{\nabla S_{x^*}(x_t)^T u}_{>w}+\frac{1}{2}\delta^2 u^T\nabla^2 S_{x^*}(x')u$$

از فرض سوال استفاده می‌کنیم.

$$S_{x^*}(x_t+\delta u)>\delta w+\frac{1}{2}\delta^2 u^T\nabla^2 S_{x^*}(x')u$$

طبق رابطه (۱۹) با بازگرین قدر مطلق داریم:

$$\boxed{-\frac{1}{2}L\delta^2\leq\frac{1}{2}\delta^2 u^T\nabla^2 S_{x^*}(x')u\leq\frac{1}{2}L\delta^2}$$

$$S_{x^*}(x_t+\delta u)>\delta w-\frac{1}{2}L\delta^2$$

$$\delta w-\frac{1}{2}L\delta^2=\delta w-\delta\left(\frac{1}{2}L\delta\right)=\delta w-\delta w=0$$

$$\boxed{S_{x^*}(x_t+\delta u)>0}$$

۲. فرض سوال: $\boxed{\nabla S_{x^*}(x_t)^T u < -w = -\frac{1}{2}L\delta}$

طبق رابطه (۱۵) داریم:

$$S_{x^*}(x_t+\delta u)=\delta\underbrace{\nabla S_{x^*}(x_t)^T u}_{<-w}+\frac{1}{2}\delta^2 u^T\nabla^2 S_{x^*}(x')u$$

از فرض سوال استفاده می‌کنیم.

با بازگرین رابطه (۱۹) داریم

$$-\frac{1}{2}L\delta^2\leq\boxed{\frac{1}{2}\delta^2 u^T\nabla^2 S_{x^*}(x')u\leq\frac{1}{2}L\delta^2}$$

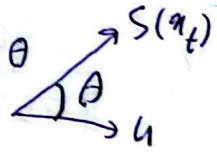
$$S_{x^*}(x_t+\delta u)<-\delta w+\frac{1}{2}\delta^2 u^T\nabla^2 S_{x^*}(x')u$$

$$S_{x^*}(x_t+\delta u)<\underbrace{-\delta w+\frac{1}{2}L\delta^2}_{0}=-\delta\left(\frac{1}{2}L\delta\right)+\frac{1}{2}L\delta^2=-\frac{1}{2}L\delta^2+\frac{1}{2}L\delta^2=0$$

بنابراین داریم: $S_{x^*}(x_t+\delta u) < 0$

بخش ج) توجه: ضرب داخلی شباهت دو بردار را می‌سنجد. اگر مثبت باشد بردارها align و در یک جهت هستند. اگر منفی باشد بردارها anti aligned و در خلاف هم هستند و اگر صفر باشد بردارها به هم عمود هستند.

$\nabla S(x_t)^T u = \|S(x_t)^T\|_2 \underbrace{\|u\|_2}_{1} \cos(\theta) = \|S(x_t)^T\|_2 \cos(\theta)$ θ زاویه بین دو بردار است.



$\nabla S(x_t)^T u > 0 \rightarrow \cos(\theta) > 0 \rightarrow 0 < \theta < 90^\circ \rightarrow$ u و $S(x_t)$ هم جهت اند

$\nabla S(x_t)^T u < 0 \rightarrow \cos(\theta) < 0 \rightarrow 90^\circ < \theta < 180^\circ \rightarrow$ $S(x_t)$, u خلاف هم اند

$\nabla S(x_t)^T u = 0 \rightarrow \cos(\theta) = 0 \rightarrow \theta = 90^\circ \rightarrow$ u و $S(x_t)$ برهم عمودند.

حال به بررسی هر رویداد می‌پردازیم:

- $E_1 = \{\nabla S(x_t)^T u > w\}$ ← جهت تصادفی u به شرت با گرادیان همسو (aligned) است چرا که ضرب داخلی از آستانه w بیشتر است. این بدان معنا است که حرکت در جهت u باعث افزایش S_{x^*} می‌شود که برای خاصیت تخاصمی مطلوب است. همچنین $S>0$ و در نتیجه این u منجر به $\Phi(x_t+\delta u) = +1$ می‌شود. درواقع u در تخمین گرادیان به شکل مثبت مشارکت می‌کند.

- $E_2 = \{-w < |\nabla S(x_t)^T u| < w\}$ ← جهت تصادفی u تقریباً متعامد بر گرادیان است چون ضرب داخلی نزدیک صفر است. ($|\nabla S(x_t)^T u| < w$) همچنین علامت $S_{x^*}(x_t+\delta u)$ به دلیل اثرات مرتبه دوم نامشخص است و ابهام دارد. این جهت u اطلاعات کمی در مورد جهت گرادیان ارائه می‌دهد و چون عبارت مرتبه دوم ممکن است غالب باشد این علامت غیر قابل اعتماد است و به نوعی نویز (noise) است.

- $E_3 = \{|\nabla S(x_t)^T u| < -w\}$ ← جهت تصادفی u به شرت با گرادیان همسو نیست چون ضرب داخلی از آستانه $-w$ کوچکتر است و بردارها خلاف جهت هم هستند. حرکت در جهت u باعث کاهش S_{x^*} می‌شود که برای خاصیت تخاصمی نامطلوب است. همچنین $S<0$ و در نتیجه این جهت تصادفی u منجر به $\Phi(x_t+\delta u) = -1$ می‌شود. درواقع u در تخمین گرادیان به شکل منفی مشارکت می‌کند.

بخش د) فرض کنیم $\{V_i\}_{i=1}^{d}$ یک پایه متعامد با $V_1 = \frac{\nabla S(x_t)}{\|\nabla S(x_t)\|_2}$ باشد. هر بردار تصادفی u که از یک توزیع یکنواخت (uniform) بر روی کره واحد می‌آید می‌تواند به شکل $u = \sum_{i=1}^{d} \beta_i V_i$ نوشته شود که $\sum_{i=1}^{d} \beta_i^2 = 1$ و $\beta_i = \langle V_i, u \rangle$ چون $\{V_i\}_{i=1}^{d}$ یک پایه متعامد (orthogonal basis) است.

همچنین داریم:

$$V_1 = \frac{\nabla S(x_t)}{\|\nabla S(x_t)\|_2} \longrightarrow \boxed{\nabla S(x_t) = V_1 \cdot \|\nabla S(x_t)\|_2}$$

پس داریم:

$$\nabla S(x_t)^T u = (\underbrace{\|\nabla S(x_t)\|_2}_{\text{scalar}} \cdot V_1)^T u = \|\nabla S(x_t)\|_2 \cdot \underbrace{V_1^T u}_{} = \|\nabla S(x_t)\|_2 \beta_1$$

چون $\{V_i\}_{i=1}^{d}$ یک پایه متعامد است.

$$i \neq j \rightarrow V_i \perp V_j \rightarrow \langle V_i, V_j \rangle = 0$$

$$\rightarrow \boxed{\nabla S(x_t)^T u = \|\nabla S(x_t)\|_2 \beta_1}$$

حال برای E_1 داریم:

$$E_1: \nabla S(x_t)^T u > w \Rightarrow \|\nabla S(x_t)\|_2 \beta_1 > w \Rightarrow \beta_1 > \underbrace{\frac{w}{\|\nabla S(x_t)\|_2}}_{} > 0$$

چون صورت و مخرج اعداد نامنفی هستند.

$$\boxed{\beta_1 > 0 \quad \text{with} \quad \text{sign}(S(x_t + \delta u)) = +1}$$

برای E_3 داریم:

$$E_3: \nabla S(x_t)^T u < -w \Rightarrow \|\nabla S(x_t)\|_2 \beta_1 < -w \Rightarrow \beta_1 < \underbrace{\frac{-w}{\|\nabla S(x_t)\|_2}}_{} < 0$$

چون $w \geq 0$ و $\|\cdot\|_2 \geq 0$ و فرض می‌کنیم گرادیان صفر مطلق نیست.

$$\boxed{\beta_1 < 0 \quad \text{with} \quad \text{sign}(S(x_t + \delta u)) = -1}$$

حال با توجه به تقارن توزیع یکنواخت (symmetry of uniform distribution) روی کره، توزیع β_1 در حدود صفر متقارن است.

$\mathbb{E}[\beta_1 | E_1]$ امید ریاضی شرطی β_1 با فرض $\beta_1 > \frac{w}{\|\nabla S(x_t)\|_2}$ است.

$\mathbb{E}[\beta_1 | E_3]$ امید ریاضی شرطی β_1 با فرض $\beta_1 < \frac{-w}{\|\nabla S(x_t)\|_2}$ است.

بنابراین طبق symmetry یا همان تقارن توزیع یکنواخت روی کره داریم:

$$\boxed{\mathbb{E}[\beta_1 | E_1] = -\mathbb{E}[\beta_1 | E_3] > 0}$$

۲. برای $i \geq 2$ بردار V_i عمود بر بردار $V_1 = \frac{\nabla S(x_t)}{\|\nabla S(x_t)\|_2}$ یعنی $V_i \perp V_1$, for $i \geq 2$

درواقع $0 = \langle V_1, V_i \rangle$, for $i \geq 2$.

رویدادهای E_1 و E_3 فقط با شرط روی B_1 (مؤلفه در امتداد V_1) تعریف می‌شوند چرا که V_1 بردار پایه در جهت $\nabla S(x_t)$ است و بقیه بردارهای V_i طبق تعریف پایه متعامد $\{V_i\}_{i=1}^{d}$ بر V_1 عمودند. ($V_i \perp V_1$, for $i \geq 2$). ازآنجایی که توزیع یکنواخت روی کره از نظر چرخشی ثابت است (rotationally invariant)، شرط گذاشتن روی B_1 بر توزیع مؤلفه‌های B_i برای $i \geq 2$ تأثیری ندارد که منجر به این می‌شود که در حدود صفر متقارن (symmetric around zero) باقی می‌مانند. توجه کنید که رویدادهای E_1 و E_3 فقط روی B_1 شرطی گذارند چرا که فقط V_1 در جهت گرادیان است.

Conditioning event depends only on B_1

درواقع توزیع حاشیه‌ای B_i اطراف صفر متقارن است و مستقل از تغییرات علامت در B_1 است.
پس طبق تقارن، میانگین شرطی B_i برای هر رویداد متقارن با علامت B_1 برابر با صفر می‌شود.
بنابراین:

$$\boxed{\mathbb{E}[B_i \mid E_1] = \mathbb{E}[B_i \mid E_3] = 0 \quad \text{for all } i \geq 2}$$

خلاصه: ازآنجا که توزیع joint به شکل rotationally invariant است و رویدادهای E_1 و E_3 فقط B_1 بستگی دارند توزیع شرطی هر B_i که $i \geq 2$ حول صفر متقارن است و بنابراین میانگین شرطی صفر است.

بخش هـ) با استفاده از قانون امید کل داریم:

$$\mathbb{E}[\Phi_{x^*}(x+\delta u)] = P[E_1]\,\mathbb{E}[\Phi_{x^*}(x+\delta u) \mid E_1] + P[E_2]\,\mathbb{E}[\Phi_{x^*}(x+\delta u) \mid E_2] + P[E_3]\,\mathbb{E}[\Phi_{x^*}(x+\delta u) \mid E_3]$$

توجه کنید که: در رویداد E_1 : $E_1 \Longrightarrow S_{x^*}(x_t + \delta u) > 0 \longrightarrow \Phi_{x^*}(x+\delta u) = +1$

در رویداد E_3 : $E_3 \Longrightarrow S_{x^*}(x_t + \delta u) < 0 \longrightarrow \Phi_{x^*}(x+\delta u) = -1$

در رویداد E_2 علامت مبهم است و مشخص نیست.

فرض کنید $P[E_2] = p$ پس:

$$\mathbb{E}[\Phi_{x^*}(x+\delta u)] = P[E_1] \times 1 + p \cdot \mathbb{E}[\Phi_{x^*}(x+\delta u) \mid E_2] + P[E_3] \times (-1)$$

$$\mathbb{E}[\Phi_{x^*}(x+\delta u)] = p \cdot \underbrace{\mathbb{E}[\Phi_{x^*}(x+\delta u)]}_{A} + \underbrace{(P[E_1] - P[E_3])}_{B} = pA + B$$

$$\boxed{A = \mathbb{E}[\Phi_{x^*}(x+\delta u)], \quad B = P[E_1] - P[E_3]}$$

همچنین توجه کنید که طبق تقارن (symmetry) داریم که

$$\left.\begin{array}{l} P[E_1]+P[E_2]+P[E_3]=1 \\ P[E_1]=P[E_3] \\ P[E_2]=p \end{array}\right\} \Rightarrow P[E_1]=P[E_3]=\frac{1-p}{2} \Rightarrow P[E_1]-P[E_3]=0 \Downarrow \boxed{B=0}$$

$$\underbrace{pA}_{\mathbb{E}[\phi_{x^*}(x+\delta u)|E_2]}+\underbrace{B}_{0}=\mathbb{E}[\phi_{x^*}(x+\delta u)]=p\cdot\mathbb{E}[\phi_{x^*}(x+\delta u)|E_2]$$

حال سعی می‌کنیم با تتابع بخش د) آن را ساده‌تر کنیم.

امید ریاضی کامل جمله تخمین گر می‌شود:

$$\mathbb{E}[\phi_{x^*}(x+\delta u)u]=P[E_1]\underbrace{\mathbb{E}[u|E_1]}_{\text{طبق د } \mathbb{E}[\beta_1|E_1]v_1}+p\,\mathbb{E}[\phi_{x^*}(x+\delta u)u|E_2]+P[E_3]\underbrace{\mathbb{E}[u|E_3]}_{\text{طبق د } \mathbb{E}[\beta_1|E_3]v_1}$$

$$\cancel{\mathbb{E}[\phi_{x^*}(x+\delta u)u]=\underbrace{\left(\frac{1-p}{2}\right)}_{P[E_1]=P[E_3]=\frac{1-p}{2}}(\mathbb{E}[\beta_1|E_1]+\mathbb{E}[\beta_1|E_3])}$$

$$\mathbb{E}[\beta_1|E_3]=-\mathbb{E}[\beta_1|E_1]$$

$$\mathbb{E}[\phi_{x^*}(x+\delta u)u]=\underbrace{(P[E_1]-P[E_3])}_{\text{می‌توان با تقارن این عبارت را صفر کرد.}}\mathbb{E}[\beta_1|E_1]v_1+p\,\mathbb{E}[\phi_{x^*}(x+\delta u)u|E_2]$$

بخش و) ۱. ابتدا $\|a-b\|^2$ را بسط می‌دهیم:

$$\|a-b\|_2^2=\|a\|_2^2+\|b\|_2^2-2a^Tb$$

$$2a^Tb=\|a\|_2^2+\|b\|_2^2-\|a-b\|_2^2 \longrightarrow a^Tb=\frac{1}{2}(\underbrace{\|a\|_2^2+\|b\|_2^2}_{\geq 2\|a\|_2\|b\|_2}-\|a-b\|_2^2)$$

حال با استفاده از نامساوی AM-GM (Arithmetic Mean-Geometric mean) داریم:

$$a^Tb=\frac{1}{2}(\|a\|_2^2+\|b\|_2^2-\|a-b\|_2^2)\geq\frac{1}{2}(2\|a\|_2\|b\|_2-\|a-b\|_2^2)$$

$\|a\|_2^2+\|b\|_2^2\geq 2\|a\|_2\|b\|_2$

اثبات با $\|a-b\|_2^2\geq 0$

$$\rightarrow \boxed{a^Tb\geq\|a\|_2\|b\|_2-\frac{1}{2}\|a-b\|_2^2}$$

$$\cos\angle(a,b)=\frac{a^Tb}{\|a\|_2\|b\|_2}\geq\frac{\|a\|_2\|b\|_2-\frac{1}{2}\|a-b\|_2^2}{\|a\|_2\|b\|_2}=\frac{\|a\|_2\|b\|_2}{\|a\|_2\|b\|_2}-\frac{\frac{1}{2}\|a-b\|_2^2}{\|a\|_2\|b\|_2}=1-\frac{\|a-b\|_2^2}{2\|a\|_2\|b\|_2}$$

حال طبق فرض سوال a و b در فضایی مشخص هستند بنابراین فرض می‌کنیم که $\|a\|_2\approx\|b\|_2$ (یا به طور دقیق‌تر اگر $a=\mathbb{E}[\phi u]$ و $b=\mathbb{E}[|\beta_1|v_1]$ که هردو norm های مشابهی دارند) آنگاه:

$$\cos\angle(a,b)\geq 1-\frac{\|a-b\|_2^2}{2\|b\|_2\|b\|_2}=1-\frac{\|a-b\|_2^2}{2\|b\|_2^2}\Longrightarrow \boxed{\cos\angle(a,b)\geq 1-\frac{\|a-b\|_2^2}{2\|b\|_2^2}}$$

۲. فرض کنید:

طبق تعریف $V_1 = \frac{\nabla S(x_t)}{\|\nabla S(x_t)\|_2}$

$$b = \mathbb{E}[|\beta_1| v_1] = \underbrace{\mathbb{E}[|\beta_1|]}_{\text{positive scalar}} \frac{\nabla S(x_t)}{\|\nabla S(x_t)\|_2} \quad , \quad a = \mathbb{E}[\phi_{x^*}(x_t+\delta u)u]$$

$$\hat{\nabla} S(x_t,\delta) = \frac{1}{B}\sum_{b=1}^{B} \phi_{x^*}(x_t+\delta u_b)u_b$$

expectation

$$\|\cdot\|_2 \rightarrow \|b\|_2 = \mathbb{E}[|\beta_1|] \underbrace{\frac{\|\nabla S(x_t)\|_2}{\|\nabla S(x_t)\|_2}}_{1} = \mathbb{E}[|\beta_1|]$$

$$\mathbb{E}[\hat{\nabla} S(x_t,\delta)] = \mathbb{E}[\phi_{x^*}(x_t+\delta u)u] = a$$

با توجه به رابطه (۲۶) داریم: $(\|a-b\|_2 \le 3\rho)$

توان ۲

$$\|a-b\|_2^2 \le 9\rho^2$$

$$\times(-1) \quad \boxed{-\|a-b\|_2^2 \ge -9\rho^2} \quad (*)$$

حال از رابطه (۲۵) استفاده می‌کنیم:

$$\cos\angle(\mathbb{E}[\hat{\nabla} S(x_t,\delta)], \nabla S(x_t)) = \cos\angle(a, \nabla S(x_t))$$

حال توجه کنید که تابع cos میان دو بردار scale invariant است چرا که فقط به زاویه میان آنها بستگی دارد. پس

$$\cos\angle(a, \nabla S(x_t)) = \cos\angle(a, \nabla S(x_t) \times \underbrace{\frac{\mathbb{E}[|\beta_1|]}{\|\nabla S(x_t)\|_2}}_{\text{scalar عدد ثابت}}) = \cos\angle(a,b)$$

$$\cos\angle(a,b) \ge 1 - \frac{\|a-b\|_2^2}{2\|b\|_2^2} \overset{(*)}{\ge} 1 - \frac{9\rho^2}{2\|b\|_2^2} = 1 - \frac{9\rho^2}{2(\mathbb{E}[|\beta_1|])^2}$$

بنابراین داریم:

$$\cos\angle(\mathbb{E}[\hat{\nabla} S(x_t,\delta)], \nabla S(x_t)) \ge 1 - \frac{9\rho^2}{2(\mathbb{E}[|\beta_1|])^2}$$

بخش ز) توجه کنید در مرحله قبل اثبات کردیم که $\cos\angle(\mathbb{E}[\hat{\nabla} S(x_t,\delta)], \nabla S(x_t)) \ge 1 - \frac{9\rho^2}{2(\mathbb{E}[|\beta_1|])^2}$

و کافی است بر ρ و $\mathbb{E}[|\beta_1|]$ روابطی پیدا کنیم.

$$\rho = \mathbb{P}(E_2) = \mathbb{P}(|\nabla S(x_t)^T u| < w), \quad w = \frac{L\delta}{2}$$

برای بردار تصادفی u روی کره واحد در $\mathbb{R}^d$ داریم:

$$\beta_1^2 = \langle v_1, u\rangle^2 \sim \text{Beta}\left(\frac{1}{2}, \frac{d-1}{2}\right)$$

با توجه به راهنمایی داده شده در صورت سوال داریم:

$$\mathbb{E}[|\beta_1|] = \frac{2}{d-1} \xrightarrow{\text{بنابراین}} \mathbb{E}[|\beta_1|] = \left(\frac{2}{d-1}\right)^2$$

حال باید کرانی برای ρ پیدا کنیم.

$$\cos\angle(\mathbb{E}[\underbrace{\hat{\nabla} S(x_t,\delta)}_{a}], \underbrace{\nabla S(x_t)}_{b}) \ge 1 - \frac{9\rho^2}{2\left(\frac{2}{d-1}\right)^2} = 1 - \frac{9\rho^2(d-1)^2}{8}$$

کران بالا برای p:

$$p = \mathbb{P}(|\nabla S(x_t)^T u| < w) \quad , \quad \nabla S(x_t)^T u = \|\nabla S(x_t)\|_2 B_1 \Rightarrow$$

$$p = \mathbb{P}\left(|B_1| < \frac{w}{\|\nabla S(x_t)\|_2}\right) \xrightarrow{w = \frac{L\delta}{2}} p = \mathbb{P}\left(|B_1| < \frac{L\delta}{2\|\nabla S(x_t)\|_2}\right)$$

حال چون $|B_1|$ دارای توزیع متقارن مشتق شده از توزیع Beta است، یک حد بالای ساده (concentration bounds) به شکل زیر است:

$$p \le \frac{L\delta}{\|\nabla S(x_t)\|_2}$$

(ثابت های دقیق بعداً جذب می شوند؛ این حرکاتی است.)

حال باید حد بالای p^2 را در رابطه A جایگزین کنیم.

$$p^2 \le \frac{L^2\delta^2}{\|\nabla S(x_t)\|_2^2} \quad (A) \longrightarrow \cos\angle(a,b) \ge 1 - \frac{9(d-1)^2}{\Lambda} \cdot \frac{L^2\delta^2}{\|\nabla S(x_t)\|_2^2}$$

که دقیقاً همان خواسته مسئله است.

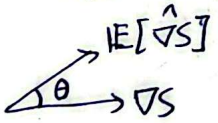
$$\cos\angle\left(\mathbb{E}[\hat{\nabla} S(x_t,\delta)], \nabla S(x_t)\right) \ge 1 - \frac{9L^2\delta^2(d-1)^2}{\Lambda\|\nabla S(x_t)\|_2^2}$$

۲. بیایید دوباره به طرف راست نامساوی دقت کنیم:

$$1 \leftarrow 1 - 0 \leftarrow 1 - \underbrace{\frac{9L^2\delta^2(d-1)^2}{\Lambda\|\nabla S(x_t)\|_2^2}}_{0}$$

وقتی $\delta \to 0$: عبارت خطای تخمین برحسب δ درجه دو است. بنابراین خیلی سریع به صفر میل می کند.

پس:



$$\lim_{\delta \to 0} \cos\angle\left(\mathbb{E}[\hat{\nabla} S(x_t,\delta)], \nabla S(x_t)\right) = 1$$

وقتی $\cos\theta = 1$ یعنی زاویه بین ∇S و $\mathbb{E}[\hat{\nabla} S]$ به صفر میل می کند و در بردار هم سو و هم جهت می شوند.

$$\mathbb{E}[\hat{\nabla} S(x_t,\delta)] \longrightarrow \nabla S(x_t) \longleftarrow \theta \to 0$$

تفسیر: وقتی $\delta \to 0$ لذا $\theta \to 0$ می کند. یعنی تخمین گر گرادیان جعبه سیاه مورد انتظار کاملاً با گرادیان واقعی هم سو و هم جهت می شود. با اینکه ما هرگز احتمالات یا گرادیان ها را مشاهده نمی کنیم و فقط برچسب های پیش بینی شده را مشاهده می کنیم، تخمین گر جهت صحیح گرادیان را بازیابی می کند.

δ کوچکتر ← ابهام کمتر در جهت ها ← همسویی بهتر

این موضوع توجیه می کند که چرا حملات جعبه سیاه فقط با برچسب هنوز می توانند همگرا شوند.

شهود: با کوچک شدن δ، probe های گروی تصادفی almost exactly در امتداد گرادیان واقعی از مرز تصمیم گیری عبور می کنند. بنابراین میانگین جهت های sign ها به گرادیان واقعی همگرا می شود. پس با کوچک کردن δ به اندازه کافی حملات black box عملکرد خوبی خواهد داشت.