DS603: Robust Machine Learning **Autumn 2025**

Lecture 14 — Oct 10

Lecturer: Arjun Bhagoji *Scribe: Pranav Bhagwat*

14.1 Theorem Statement

If $g_{\vec{x}}(\vec{x}_t)$ has gradients that are L-Lipschitz (L-smooth) then

$$\cos\angle(\mathbb{E}[\tilde{\nabla} g_{\vec{x}}(\vec{x}_t,b)],\nabla g_{\vec{x}}(\vec{x}_t)) \geqslant 1-\frac{9L^2b^2d^2}{8||\nabla g_{\vec{x}}(\vec{x}_t)||_2^2},$$

and as $b \to 0$, the angle tends to 0, if $\vec{x}_t \in \mathit{boundary}(g_{\vec{x}})$,

Key question

Why is the estimate of the gradient defined solely using the decision information reasonable?

Proof idea

The main idea is to show that the probability of the event where the sign of the function $g(\,.)$ is *not* a good indicator of the projection of its gradient along a random direction is small.
This proof is taken from **HopSkipJumpAttack paper** [1].

14.2 Proof of Theorem

We use a gradient estimate based on sign decisions along random directions. For an iterate x_t and radius b let us define

$$\widehat{\nabla} g_{\vec{x}}(\vec{x}_t,b) := \frac{1}{k}\sum_{k=1}^{k} \psi_{\vec{x}_t}(\vec{x}_t+b\vec{u}_k)\,\vec{u}_k, \tag{1}$$

To bound the expectation of this gradient with respect to the random unit direction $\vec{u}_k \sim S^{d-1}$, let us consider a single unit vector $\vec{u}$ (with $||\vec{u}||=1$).
Why?: The estimator is an average over independent random directions, so to bound the expectation it is enough to analyze one randomly drawn unit direction.

Taylor expansion of the perturbed function

By Taylor's theorem,

$$g(\vec{x}+b\vec{u}) = b\nabla g(\vec{x}_t)^\top\vec{u}+\frac{b^2}{2}\vec{u}^\top\nabla^2 g(\vec{x}_t)\,\vec{u},\ \texttt{(since}\ \ g(\vec{x})=0\texttt{)} \tag{2}$$

As the gradient of g is L–Lipschitz, then the remainder term satisfies

$$\left|\tfrac{b^2}{2}\vec{u}^\top \nabla^2 g(\vec{x}_t)\,\vec{u}\right| \leqslant \tfrac{1}{2}Lb^2.$$

Thus the linear term dominates the perturbation whenever $|\nabla g(\vec{x}_t)^\top \vec{u}| > \frac{1}{2}Lb$. Consequently,

$$\begin{aligned}
\text{If } \nabla g(\vec{x})^\top \vec{u} > \tfrac{1}{2}Lb &\implies g(\vec{x}+b\vec{u}) > 0 \quad \text{and} \quad \psi(\vec{x}+b\vec{u}) = +1,\\
\text{If } \nabla g(\vec{x})^\top \vec{u} < -\tfrac{1}{2}Lb &\implies g(\vec{x}+b\vec{u}) < 0 \quad \text{and} \quad \psi(\vec{x}+b\vec{u}) = -1.
\end{aligned}$$

Therefore the decision $\psi(g(\vec{x}+b\vec{u}))$ correctly reflects the sign of the directional derivative in extreme case, but not in centre case when $|\nabla g(\vec{x})^\top u| \leqslant \frac{1}{2}Lb$. We call this the *ambiguous* set of directions.

Let's define the following events for a fixed iterate $\vec{x}$ and vector $\vec{u}$:

$$\begin{aligned}
E_1 &:= \{\,\nabla g(\vec{x}_t)^\top \vec{u} > \tfrac{1}{2}Lb\,\},\\
E_2 &:= \{\,|\nabla g(\vec{x})^\top \vec{u}| \leqslant \tfrac{1}{2}Lb\,\},\\
E_3 &:= \{\,\nabla g(\vec{x})^\top \vec{u} < -\tfrac{1}{2}Lb\,\}.
\end{aligned}$$

On E_1 and E_3 the sign of $g(\vec{x}+b\vec{u})$ matches the sign of $\nabla g(\vec{x})^\top u$. The only problematic directions belong to E_2.

Orthonormal-basis decomposition

Let us fix an orthonormal basis to simplify things.
Fix $\vec{v}_1 = \frac{\nabla g(\vec{x}_t)}{\|\nabla g(\vec{x}_t)\|_2}$, and choose $\{\vec{v}_i\}_{i=1}^d$ to be appropriately orthonormal. Then

$$\vec{u} = \sum_{i=1}^{d} r_i v_i, \qquad \text{where } \vec{r} \sim S^{d-1}$$

Why?: As v is unit vector and has orthonormal components, when it is multiplied with any $\vec{r}$ then it can generate any $\vec{u}$ In this basis we have

$$\nabla g(\vec{x})^\top \vec{u} = \vec{r}_1 \nabla g(\vec{x})^\top \vec{v}_1 = \vec{r}_1 \|\nabla g(\vec{x}_t)\|_2\,.$$

Recall, we want to lower bound $cos\angle([E[\widehat{\nabla} g_{\vec{x}}(\vec{x}_t,b)],\nabla g_{\vec{x}_t}(\vec{x}_t)])$,
Instead of that we could upper-bound

$$\|E[\widehat{\nabla} g_{\vec{x}_t}(\vec{x}_t,b)] - \nabla g_{\vec{x}_t}(\vec{x}_t)\|_2$$

Why?: By using the identity

$$\cos\angle(a,b) = \frac{\|a\|^2 + \|b\|^2 - \|a-b\|^2}{2\,\|a\|\,\|b\|}$$

Upper bounding $\|a-b\|$ will give a lower bound on cos and will also be algebraically simpler.

Consider

$$E[|r_1|\vec{v}_1] = \vec{v}_1 E[r_1] = \frac{\nabla g(\vec{x}_t)}{\|\nabla g(\vec{x}_t)\|_2} E[|r_1|]$$

$$\Rightarrow \frac{\nabla g(\vec{x}_t)}{\|\nabla g(\vec{x}_t)\|_2} = \frac{E[|r_1|v_1]}{E[|r_1|]}$$

So, if we can bound the following by some ρ':

$$\|E[\psi_{\vec{x}}\,(\vec{x}_t + b\vec{u})\,\vec{u}] - E[|\vec{r}_1|\vec{v}_1]\|_2 \leqslant \rho'$$

$$\implies \|E[\psi_{\vec{x}}(\vec{x}_t + b\vec{u})\,\vec{u}] - r\nabla g(\vec{x}_t)\|_2 \leqslant \rho'(\mathbf{r} \text{ is random}) \qquad 1$$

and recalling, $||\vec{a} - r\vec{b}||^2 = \|\vec{a}\|^2 + r^2\|\vec{b}\|^2 - 2r||\vec{a}||_2||\vec{b}||_2 \cos\angle(\vec{a},\vec{b})$

As in our case,

$$||\vec{a} - r\vec{b}||_2^2 \leqslant (\rho')^2$$

so

$$\cos\angle(\vec{a},\vec{b}) \geqslant \frac{\|\vec{a}\|^2 + r^2\|\vec{b}\|^2 - (\rho')^2}{2r\|\vec{a}\|\|\vec{b}\|}$$

In our case, $r = E[|\vec{r}_1|]$, and $||\vec{b}|| = 1$,

$$\Rightarrow \cos\angle(\vec{a},\vec{b}) \geqslant \frac{||\vec{a}||^2 + r^2 - (\rho')^2}{2r||\vec{a}||} \geqslant 1 - \tfrac{1}{2}(\tfrac{\rho'}{r})^2 \quad (\text{when } \|\vec{a}\| \geqslant r) \qquad 2$$

Therefore above equation (2) will hold if

$$\|E[\psi_{\vec{x}}(\vec{x}_t + b\vec{u})\,\vec{u}]\|_2 \geqslant E[|r_1|]$$

Now, we need to determine what ρ' is, and verify that above equation (2) holds.
As defined earlier,

$$E_1 := \{\,\nabla g(\vec{x}_t)^\top \vec{u} > \tfrac{1}{2}Lb\,\},$$
$$E_2 := \{\,|\nabla g(\vec{x}_t)^\top \vec{u}| \leqslant \tfrac{1}{2}Lb\,\},$$
$$E_3 := \{\,\nabla g(\vec{x}_t)^\top \vec{u} < -\tfrac{1}{2}Lb\,\}$$

In the basis that we have, $E_1 \iff \nabla g(\vec{x}) \sum r_i v_i > w \iff r_1\|\nabla g(\vec{x})\|_2 > w$
Note also that $E_1 \implies \psi_{\vec{x}}(\vec{x}_t + b\vec{u}) = 1$.

14.2.1 Bounding the expectation

Now consider

$$\mathbb{E}[\psi_{\vec{x}}(\vec{x}_t + b\vec{u})\,\vec{u}] = \rho\,\mathbb{E}[\,\psi_{\vec{x}} \mid E_2\,] + \frac{1-\rho}{2}\,\mathbb{E}[\psi_{\vec{x}} \mid E_1\,] + \frac{1-\rho}{2}\,\mathbb{E}[\psi_{\vec{x}} \mid E_3\,]$$

as,

$$\mathbb{E}[\psi_{\vec{x}}\,\vec{u} \mid E_1\,] = \mathbb{E}[\Sigma_i\; r_i\vec{v}_i \mid E_1] = \mathbb{E}[r_1\;\vec{v}_1|E_1]$$

$$\therefore \mathbb{E}[\psi_{\vec{x}}\vec{u}] = \rho\left(\mathbb{E}[\psi_{\vec{x}}|E_2] - \frac{1}{2}\mathbb{E}[r_1\vec{v}_1|E_1] - \frac{1}{2}\mathbb{E}[-r_1\vec{v}_1|E_3]\right) + \frac{1}{2}\mathbb{E}[r_1\vec{v}_1|E_1] + \frac{1}{2}\mathbb{E}[-r_1\vec{v}_1|E_3]$$

Now consider

$$\mathbb{E}[|r_1|\vec{v}_1] = \frac{1}{2}\mathbb{E}[r_1\vec{v}_1|r_1 > 0] + \frac{1}{2}\mathbb{E}[-r_1\vec{v}_1|r_1 < 0]$$

$$\mathbb{E}[\psi_{\vec{x}}\vec{u}] = \rho\left(\mathbb{E}[\psi_{\vec{x}}|E_2] - \frac{1}{2}\mathbb{E}[r_1\vec{v}_1|E_1] - \frac{1}{2}\mathbb{E}[-r_1\vec{v}_1|E_3]\right) + \frac{1}{2}\mathbb{E}[r_1\vec{v}_1|E_1] + \frac{1}{2}\mathbb{E}[-r_1\vec{v}_1|E_3]$$

Using the triangle inequality and the fact that norms of each expectation are bounded from above by 1, we obtain

$$\left\|\mathbb{E}[\psi_{\vec{x}}\vec{u}] - \mathbb{E}[|r_1|\vec{v}_1]\right\|_2 \leqslant \rho + 2\rho = 3\rho$$

therefore $\rho' = 3\rho$, where

$$\rho = \mathbb{P}\left[|\nabla g(\vec{x}_t)^\top u| < \frac{1}{2}Lb\right] = \mathbb{P}\left[r_1^2 \leqslant \frac{(\frac{1}{2}Lb)^2}{||\nabla g(\vec{x}_t)||_2^2}\right]$$

Note: Each element of a unit random vector defined over the unit sphere follows a beta distribution, i.e.,

$$r_i^2 \sim \text{Beta}(p,q) = \frac{1}{\text{beta}(p,q)}x^{p-1}(1-x)^{q-1}$$

with $p = \frac{1}{2}, q = \frac{d-1}{2}$.

About Beta distribution

The Beta distribution is a continuous probability distribution defined on the interval $x \in [0,1]$. It is parameterized by two positive parameters p and q, which control the shape of the distribution.

$$\begin{aligned}\beta(x) &= \frac{\Gamma(p+q)}{\Gamma(p)\Gamma(q)}x^{p-1}(1-x)^{q-1} \\ &= \frac{1}{beta(p,q)}x^{p-1}(1-x)^{q-1}, \qquad 0 \leqslant x \leqslant 1,\end{aligned}$$

$$beta(p,q) \;=\; \int_0^1 x^{p-1}(1-x)^{q-1}\,dx$$

Here, $\Gamma(\cdot)$ denotes the Gamma function, which generalizes the factorial: $\Gamma(n) = (n-1)!$ for any positive integer n. The beta function serves as a normalization constant ensuring that the total probability integrates to 1.

14.2.2 Continuing the proof

For small t, $\mathbb{P}[X \leqslant t]$ for a beta distribution, $Beta(\frac{1}{2}, \frac{d-1}{2})$ can be upper bounded as

$$\frac{\sqrt{t}}{\text{beta}(p,q)}$$

What other assumption do we need? Why is this true?
We need to assume that t is small, formally $0 < t << 1$. We also assume that $(q-1) > 0$ or equivalently, on putting $q = \frac{d-1}{2}$, we assumme that $d > 3$ which is true for high dimensions.
Under these assumptions, the term of Beta distribution, $(1-x)^{q-1}$ can be approximated as 1, and remaining terms give the above bound.

therefore

$$\rho \leqslant \frac{2(\frac{1}{2}Lb)}{\text{beta}(\frac{1}{2}, \frac{d-1}{2})||\nabla g(\vec{x}_t)||_2}$$

Plugging everything back into equation (2), we get

$$\cos\angle(\mathbb{E}[\psi_{\vec{x}}(\vec{x}_t + b\vec{u})\,\vec{u}], \nabla g(\vec{x}_t)) \geqslant 1 - \frac{1}{2}\left(\frac{p'}{r}\right)^2$$

$$\begin{aligned} \text{RHS: } 1 - \frac{1}{2}\left(\frac{\rho'}{r}\right)^2 &= 1 - \frac{1}{2}\left(\frac{3\rho}{\mathbb{E}[|r_1|]}\right)^2 \\ &= 1 - \frac{1}{2}\frac{9\rho^2}{(\mathbb{E}[|r_1|])^2} \\ &\geqslant 1 - \frac{1}{2}\frac{9L^2b^2}{\text{beta}\left(\frac{1}{2}, \frac{d-1}{2}\right)^2 ||\nabla g(\vec{x}_t)||_2^2\,\mathbb{E}[|r_1|]^2} \\ &\geqslant 1 - \frac{9L^2b^2d^2}{8||\nabla g(\vec{x}_t)||_2^2} \end{aligned}$$

Hence Proved

Bibliography

[1] Jianbo Chen, Michael I. Jordan, and Martin J. Wainwright. Hopskipjumpattack: A query-efficient decision-based attack, 2020.