A. Proof of Theorem 1

We first prove a lemma of the gradient estimation quality which samples from the entire subspace:

Lemma 1. *For a boundary point x, suppose that $S(x)$ has L-Lipschitz gradients in a neighborhood of x, and that $\mathbf{u}_1, \ldots, \mathbf{u}_B$ are sampled from the unit ball in $\mathbb{R}^m$ and orthogonal to each other. Then the expected cosine similarity between $\widetilde{\nabla S}$ and ∇S can be bounded by:*

$$\left(2\left(1-(\frac{L\delta}{2||\nabla S||_2})^2\right)^{\frac{m-1}{2}}-1\right)c_m\sqrt{\frac{B}{m}} \tag{14}$$

$$\leq \mathbb{E}\left[\cos(\widetilde{\nabla S}, \nabla S)\right] \tag{15}$$

$$\leq c_m\sqrt{\frac{B}{m}} \tag{16}$$

where c_m is a constant related with m and can be bounded by $c_m \in (2/\pi, 1)$. In particular, we have:

$$\lim_{\delta\to 0}\mathbb{E}\left[\cos(\widetilde{\nabla S}, \nabla S)\right] = c_m\sqrt{\frac{B}{m}}. \tag{17}$$

Proof. Let $\mathbf{u}_1, \ldots, \mathbf{u}_B$ be the random orthonormal vectors sampled from $\mathbb{R}^m$. We expand the vectors to an orthonormal basis in $\mathbb{R}^m$: $\mathbf{q}_1 = \mathbf{u}_1, \ldots, \mathbf{q}_B = \mathbf{u}_B, \mathbf{q}_{B+1}, \ldots, \mathbf{q}_m$. Hence, the gradient direction can be written as:

$$\frac{\nabla S}{||\nabla S||_2} = \sum_{i=1}^{m} a_i \mathbf{q}_i \tag{18}$$

where $a_i = \langle \frac{\nabla S}{||\nabla S||_2}, \mathbf{q}_i\rangle$ and its distribution is equivalent to the distribution of one coordinate of an $(m-1)$-sphere. Then each a_i follows the probability distribution function:

$$p_a(x) = \frac{(1-x^2)^{\frac{m-3}{2}}}{\mathcal{B}(\frac{m-1}{2}, \frac{1}{2})},\ x \in (-1, 1) \tag{19}$$

where $\mathcal{B}$ is the beta function. According to the conclusion in the proof of Theorem 1 in [9], if we let $w = \frac{L\delta}{2||\nabla S||_2}$, then it always holds true that $\phi(\mathbf{x}+\delta\mathbf{u_i}) = 1$ when $a_i > w$, -1 when $a_i < -w$ regardless of u_i and the decision boundary shape. Hence, we can rewrite ϕ_i in term of a_i:

$$\phi_i = \phi(\mathbf{x}+\delta\mathbf{u_i}) = \begin{cases} 1, & \text{if } a_i \in [w, 1) \\ -1, & \text{if } a_i \in (-1, -w] \\ \text{undetermined}, & \text{otherwise} \end{cases} \tag{20}$$

Therefore, the estimated gradient can be rewritten as:

$$\widetilde{\nabla S} = \frac{1}{B}\sum_{i=1}^{B}\phi_i \mathbf{u}_i \tag{21}$$

Combining Eqn. 18 and 21, we can calculate the cosine similarity:

$$\mathbb{E}\left[\cos(\widetilde{\nabla S}, \nabla S)\right] = \mathop{\mathbb{E}}_{a_1,\ldots,a_B}\frac{\sum_{i=1}^{B} a_i\phi_i}{\sqrt{B}} \tag{22}$$

$$= \sqrt{B}\cdot\mathop{\mathbb{E}}_{a_1}\left[a_1\phi_1\right] \tag{23}$$

In the best case, ϕ_1 has the same sign with a_1 everywhere on $(-1, 1)$; in the worst case, ϕ_1 has different sign with a_1 on $(-w, w)$. In addition, $p_a(x)$ is symmetric on $(-1, 1)$. Therefore, the expectation is bounded by:

$$2\int_{w}^{1} p_a(x)\cdot xdx - 2\int_{0}^{w} p_a(x)\cdot xdx \tag{24}$$

$$\leq \mathop{\mathbb{E}}_{a_1}\left[a_1\phi_1\right] \tag{25}$$

$$\leq 2\int_{0}^{1} p_a(x)\cdot xdx \tag{26}$$

By calculating the integration, we have:

$$\left(2\left(1-w^2\right)^{\frac{m-1}{2}}-1\right)\cdot\frac{2\sqrt{B}}{\mathcal{B}(\frac{m-1}{2}, \frac{1}{2})\cdot(m-1)} \tag{27}$$

$$\leq \mathbb{E}\left[\cos(\widetilde{\nabla S}, \nabla S)\right] \tag{28}$$

$$\leq \frac{2\sqrt{B}}{\mathcal{B}(\frac{m-1}{2}, \frac{1}{2})\cdot(m-1)} \tag{29}$$

The only problem is to calculate $\mathcal{B}(\frac{m-1}{2}, \frac{1}{2})\cdot(m-1)$. It is easy to prove by scaling that $\mathcal{B}(\frac{m-1}{2}, \frac{1}{2})\cdot(m-1) \in (2\sqrt{m}, \pi\sqrt{m})$. Hence we can get the conclusion in the theorem. ☐

Having Lemma 1, Theorem 1 follows by noticing that $\mathbb{E}\left[\cos(\widetilde{\nabla S}, \nabla S)\right] = \rho\mathbb{E}\left[\cos(\widetilde{\nabla S}, \text{proj}_{\text{span}(W)}(\nabla S))\right]$.