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 $u_1, \ldots, u_B$ are sampled from the unit ball in $\mathbb{R}^m$ and orthogonal to each other. Then the expected cosine similarity $B$ where $a$ is the beta function. According to the conclusion in [9], if we let $L\phi_1(x + \delta u_i) = 1$ when $a_i > w$, $-1$ when $a_i < -w$ regardless of $w$, and the decision boundary shape. Hence, we can rewrite $\phi_1$ in term of $a_i$:

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

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

\[
\mathbb{E} \left[ \cos(\nabla S, \nabla S) \right] = \mathbb{E} \frac{\sum_{i=1}^{B} a_i \phi_i}{\sqrt{B}} \quad (22)
\]

\[
= \sqrt{B} \cdot \mathbb{E} \left[ a_1 \phi_1 \right] \quad (23)
\]

In the best case, $\phi_1$ has the same sign with $a_1$ everywhere on $(-1, 1)$; in the worst case, $\phi_1$ 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:

\[
2 \int_{-1}^{1} p_a(x) \cdot x dx - 2 \int_{-w}^{w} p_a(x) \cdot x dx \quad (24)
\]

\[
\leq \mathbb{E} a_1 \phi_1 \quad (25)
\]

\[
\leq 2 \int_{0}^{1} p_a(x) \cdot x dx \quad (26)
\]

By calculating the integration, we have:

\[
\left( 2 \left( 1 - \frac{m-1}{2} x^2 \right) - 1 \right) \frac{2 \sqrt{B}}{B(m-1, \frac{1}{2}) \cdot (m-1)} \quad (27)
\]

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

\[
\leq 2 \sqrt{B} \quad (29)
\]

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

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

Hence, the gradient direction can be written as:

\[
\nabla S = \frac{1}{B} \sum_{i=1}^{B} \phi_i u_i \quad (21)
\]