A Proofs

A.1 Proof for Theorem 1

A.1.1 Proof for (I) and (II)

First, observe that the constraint in Equation (I) can be equivalently replaced by an inequality constraint \( f_{\pi_0}(x_0) \geq g_{\pi_0}(x_0) \). Therefore, the Lagrangian multiplier can be restricted to be \( \lambda \geq 0 \). We have

\[
L_{\pi_0}(F, B) = \min_{\delta \in B} \min_{f \in F} \max_{\lambda \geq 0} E_{\pi}[f(x_0 + z)] + \lambda \left( f_{\pi_0}^*(x_0) - E_{\pi_0}[f(x_0 + z)] \right)
\]

\[
= \max_{\lambda \geq 0} \min_{f \in F} \min_{\delta \in B} E_{\pi}[f(x_0 + z)] + \lambda \left( f_{\pi_0}^*(x_0) - E_{\pi_0}[f(x_0 + z)] \right)
\]

\[
= \max_{\lambda \geq 0} \min_{f \in F} \min_{\delta \in B} \left( \lambda f_{\pi_0}^*(x_0) + \min_{\delta \in B} E_{\pi}[f(x_0 + z)] - \lambda E_{\pi_0}[f(x_0 + z)] \right)
\]

\[
= \max_{\lambda \geq 0} \min_{f \in F} \min_{\delta \in B} \{ \lambda f_{\pi_0}^*(x_0) - D_F(\lambda \pi_0 || \pi_0) \}.
\]

II follows a straightforward calculation.

A.1.2 Proof for (III), the strong duality

We first introduce the following lemma, which is a straightforward generalization of the strong Lagrange duality to functional optimization case.

**Lemma 1.** Given some \( \delta^* \), we have

\[
\max_{\lambda \in \mathbb{R}} \min_{f \in F_{[0,1]}} E_{\pi}[f(x_0 + z)] + \lambda \left( f_{\pi_0}^*(x_0) - E_{\pi_0}[f(x_0 + z)] \right) = \min_{f \in F_{[0,1]}} \max_{\lambda \in \mathbb{R}} E_{\pi}[f(x_0 + z)] + \lambda \left( f_{\pi_0}^*(x_0) - E_{\pi_0}[f(x_0 + z)] \right).
\]

The proof of Lemma II is standard. However, for completeness, we include it here.

**Proof.** Without loss of generality, we assume \( f_{\pi_0}^*(x_0) \in (0, 1) \), otherwise the feasible set is trivial. Let \( \alpha^* \) be the value of the optimal solution of the primal problem. We define \( f_{\pi_0}^*(x_0) - E_{\pi_0}[f(x_0 + z)] = h[f] \) and \( g[f] = E_{\pi_0}[f(x_0 + z)] \). We define the following two sets:

\[
A = \{(v, t) \in \mathbb{R} \times \mathbb{R} : \exists f \in F_{[0,1]}, h[f] = v, g[f] \leq t\}
\]

\[
B = \{(0, s) \in \mathbb{R} \times \mathbb{R} : s < \alpha^*\}.
\]

Notice that both sets \( A \) and \( B \) are convex. This is obvious for \( B \). For any \( (v_1, t_1) \in A \) and \( (v_2, t_2) \in A \), we define \( f_1 \in F_{[0,1]} \) such that \( h[f_1] = v_1, g[f_1] \leq t_1 \) (and similarly we define \( f_2 \)). Notice that for any \( \gamma \in [0, 1] \), we have

\[
\gamma f_1 + (1 - \gamma) f_2 \in F_{[0,1]}
\]

\[
\gamma h[f_1] + (1 - \gamma) h[f_2] = \gamma v_1 + (1 - \gamma) v_2
\]

\[
\gamma g[f_1] + (1 - \gamma) g[f_2] \leq \gamma t_1 + (1 - \gamma) t_2,
\]

which implies that \( \gamma (v_1, t_1) + (1 - \gamma) (v_2, t_2) \in A \) and thus \( A \) is convex. Also notice that by definition, \( A \cap B = \emptyset \). Using separating hyperplane theorem, there exists a point \( (q_1, q_2) \neq (0, 0) \) and a value \( \alpha \) such that for any \( (v, t) \in A \), \( q_1 v + q_2 t \geq \alpha \) and for any \( (0, s) \in B \), \( q_2 s \leq \alpha \). Notice that we must have \( q_2 \geq 0 \), otherwise, for sufficient \( s \), we will have \( q_2 s \geq \alpha \). We thus have, for any \( f \in F_{[0,1]} \), we have

\[
q_1 h[f] + q_2 g[f] \geq \alpha^* \geq q_2 \alpha^*.
\]

If \( q_2 > 0 \), we have

\[
\max_{\lambda \in \mathbb{R}} \min_{f \in F_{[0,1]}} g[f] + \lambda h[f] \geq \min_{f \in F_{[0,1]}} g[f] + \frac{q_1}{q_2} h[f] \geq \alpha^*,
\]

which gives the strong duality. If \( q_2 = 0 \), we have for any \( f \in F_{[0,1]} \), \( q_1 h[f] \geq 0 \) and by the separating hyperplane theorem, \( q_1 \neq 0 \). However, this case is impossible: If \( q_1 > 0 \), choosing
Given our confidence lower bound

\[ q_1 h[f] = q_1 \left( f_{\pi_0}^*(x_0) - 1 \right) < 0; \]  
\[[q_1 < 0, \text{by choosing } f \equiv 0, \text{we have } q_1 h[f] = q_1 \left( f_{\pi_0}^*(x_0) - 0 \right) < 0. \]

Both cases give contradiction. \( \Box \)

Based on Lemma \( \text{III} \), we have the proof of the strong duality as follows.

Notice that by Lagrange multiplier method, our primal problem can be rewritten as follows:

\[
\min_{\delta \in \mathcal{B}} \min_{f \in \mathcal{P}[0,1]} \max_{\lambda \in \mathbb{R}} \mathbb{E}_{\pi_\delta} \left[ f(x_0 + z) + \lambda \left( f_{\pi_0}^2(x_0) - \mathbb{E}_{\pi_0} \left[ f(x_0 + z) \right] \right) \right],
\]

and the dual problem is

\[
\max_{\lambda \geq 0} \min_{\delta \in \mathcal{B}} \min_{f \in \mathcal{P}[0,1]} \mathbb{E}_{\pi_\delta} \left[ f(x_0 + z) + \lambda \left( f_{\pi_0}^2(x_0) - \mathbb{E}_{\pi_0} \left[ f(x_0 + z) \right] \right) \right] = \max_{\lambda \geq 0} \min_{\delta \in \mathcal{B}} \min_{f \in \mathcal{P}[0,1]} \mathbb{E}_{\pi_\delta} \left[ f(x_0 + z) + \lambda \left( f_{\pi_0}^2(x_0) - \mathbb{E}_{\pi_0} \left[ f(x_0 + z) \right] \right) \right],
\]

By the assumption that for any \( \lambda \geq 0 \), we have

\[
\max_{\lambda \geq 0} \min_{\delta \in \mathcal{B}} \min_{f \in \mathcal{P}[0,1]} \mathbb{E}_{\pi_\delta} \left[ f(x_0 + z) + \lambda \left( f_{\pi_0}^2(x_0) - \mathbb{E}_{\pi_0} \left[ f(x_0 + z) \right] \right) \right],
\]

for some \( \delta^* \in \mathcal{B} \). We have

\[
\max_{\lambda \geq 0} \min_{\delta \in \mathcal{B}} \min_{f \in \mathcal{P}[0,1]} \mathbb{E}_{\pi_\delta} \left[ f(x_0 + z) + \lambda \left( f_{\pi_0}^2(x_0) - \mathbb{E}_{\pi_0} \left[ f(x_0 + z) \right] \right) \right]
\]

\[
= \max_{\lambda \geq 0} \min_{f \in \mathcal{P}[0,1]} \mathbb{E}_{\pi_{\delta^*}} \left[ f(x_0 + z) + \lambda \left( f_{\pi_0}^2(x_0) - \mathbb{E}_{\pi_0} \left[ f(x_0 + z) \right] \right) \right],
\]

\[
\geq \min_{\delta \in \mathcal{B}} \max_{\lambda \in \mathbb{R}} \mathbb{E}_{\pi_\delta} \left[ f(x_0 + z) + \lambda \left( f_{\pi_0}^2(x_0) - \mathbb{E}_{\pi_0} \left[ f(x_0 + z) \right] \right) \right],
\]

where the second equality (*) is by Lemma \( \text{III} \).

A.2 Proof for Corollary \( \text{III} \)

Proof. Given our confidence lower bound

\[
\max_{\lambda \geq 0} \min_{||\delta||_1 \leq r} \left\{ \lambda \pi_0 - \int (\lambda \pi_0(z) - \pi_\delta(z))_+ \, dz \right\},
\]

One can show that the worst case for \( \delta \) is obtained when \( \delta^* = (r, 0, \cdots, 0) \) (see following subsection), thus the bound is

\[
\max_{\lambda \geq 0} \left\{ \lambda \pi_0 - \int \frac{1}{2b} \exp \left( -\frac{|z_1|}{b} \right) \left[ \lambda - \exp \left( \frac{|z_1| - |z_1 + r|}{b} \right) \right] \, dz_1 \right\}.
\]

Denote \( a \) to be the solution of

\[
\lambda = \exp \left( \frac{|a| - |a + r|}{b} \right),
\]

then obviously we have

\[
a = \begin{cases} -\infty, & b \log \lambda \geq r \\ -\frac{1}{2} (b \log \lambda + r), & -r < b \log \lambda < r \\ +\infty, & b \log \lambda \leq -r \end{cases}
\]

So the bound above is

\[
\lambda \int_{z_1 > a} \frac{1}{2b} \exp \left( -\frac{|z_1|}{b} \right) \, dz_1 - \int_{z_1 > a} \frac{1}{2b} \exp \left( -\frac{|z_1 + r|}{b} \right) \, dz_1.
\]
i) $b \log \lambda \geq r \iff \lambda \geq \exp \left( \frac{r}{b} \right)$
the bound is
$$
\max_{\lambda \geq e^{r/b}} \{ \lambda p_0 - (\lambda - 1) \} = 1 - \exp \left( \frac{r}{b} \right) (1 - p_0).
$$

ii) $-r < b \log \lambda < r \iff \exp \left( \frac{-r}{b} \right) < \lambda < \exp \left( \frac{r}{b} \right)$
the bound is
$$
\max_{\lambda} \left\{ \lambda (p_0 - 1) + \frac{\lambda}{2} \exp \left( \frac{b \log \lambda + r}{2b} \right) + \frac{1}{2} \exp \left( \frac{b \log \lambda - r}{2b} \right) \right\}
= \max_{\lambda} \left\{ \lambda (p_0 - 1) + \frac{\lambda}{2} \exp \left( \frac{b \log \lambda + r}{2b} \right) + \frac{1}{2} \exp \left( \frac{b \log \lambda - r}{2b} \right) \right\}
= \left\{ \frac{1}{2} \exp \left( - \log [2(1 - p_0)] - \frac{r}{b} \right). \right\}
$$
the extremum is achieved when $\hat{\lambda} = \exp \left( \frac{-r}{b} \right)$, Notice that $\hat{\lambda}$ does not necessarily locate in $(e^{-r/b}, e^{r/b})$, so the actual bound is always equal or less than $\frac{1}{2} \exp \left( - \log [2(1 - p_0)] - \frac{r}{b} \right)$.

iii) $b \log \lambda \leq -r \iff \lambda \leq \exp \left( \frac{-r}{b} \right)$
the bound is
$$
\max_{\lambda \leq \exp \left( \frac{-r}{b} \right)} \lambda \cdot p_0 = p_0 \exp \left( \frac{-r}{b} \right).
$$

Since $\hat{\lambda} > e^{r/b} \iff p_0 > 1 - \frac{1}{2} \exp \left( \frac{-r}{b} \right)$, notice that the lower bound is a concave function w.r.t. $\lambda$, making the final lower bound become
$$
\begin{cases}
1 - \exp \left( \frac{r}{b} \right) (1 - p_0), & \text{when } p_0 > 1 - \frac{1}{2} \exp \left( \frac{-r}{b} \right) \\
\frac{1}{2} \exp \left( - \log [2(1 - p_0)] - \frac{r}{b} \right), & \text{otherwise}
\end{cases}
$$

**Remark** Actually, we have $1 - \exp \left( \frac{r}{b} \right) (1 - p_0) \leq \frac{1}{2} \exp \left( - \log [2(1 - p_0)] - \frac{r}{b} \right)$ all the time. Another interesting thing is that both the bound can lead to the same radius bound:
$$
1 - \exp \left( \frac{r}{b} \right) (1 - p_0) \geq \frac{1}{2} \iff r < - b \log [2(1 - p_0)]
$$
$$
\frac{1}{2} \exp \left( - \log [2(1 - p_0)] - \frac{r}{b} \right) > \frac{1}{2} \iff r < - b \log [2(1 - p_0)]
$$

### A.3 Proof for Corollary 2

**Proof.** With strong duality, our confidence lower bound is
$$
\min_{\|\delta\|_2 \leq r} \max_{\lambda \geq 0} \left\{ \lambda p_0 - \int (\lambda \pi_0(z) - \pi_\delta(z))_+ dz \right\},
$$
define $C_\lambda = \{ z : \lambda \pi_0(z) \geq \pi_\delta(z) \} = \{ z : \delta^T z \leq \frac{\|\delta\|_2^2}{2r} + \sigma^2 \ln \lambda \}$ and $\Phi(\cdot)$ to be the cdf of standard gaussian distribution, then
$$
\int (\lambda \pi_0(z) - \pi_\delta(z))_+ dz
= \int_{C_\lambda} (\lambda \pi_0(z) - \pi_\delta(z)) dz
= \lambda \cdot \mathbb{P} (N(z; 0, \sigma^2 I) \in C_\lambda) - \mathbb{P} (N(z; \delta, \sigma^2 I) \in C_\lambda)
= \lambda \cdot \Phi \left( \frac{\|\delta\|_2}{2\sigma} + \frac{\sigma \ln \lambda}{\|\delta\|_2} \right) - \Phi \left( \frac{-\|\delta\|_2}{2\sigma} + \frac{\sigma \ln \lambda}{\|\delta\|_2} \right).
$$
Define
\[ F(\delta, \lambda) := \lambda p_0 - \int (\lambda \pi_0(z) - \pi_\delta(z))_+ dz = \lambda p_0 - \lambda \Phi \left( \frac{\| \delta \|_2}{2\sigma} + \frac{\sigma \ln \lambda}{\| \delta \|_2} \right) + \Phi \left( -\frac{\| \delta \|_2}{2\sigma} + \frac{\sigma \ln \lambda}{\| \delta \|_2} \right). \]
For all \( \delta \), \( F \) is a concave function w.r.t. \( \lambda \), as \( F \) is actually a summation of many concave piece wise linear function. See [55] for more discussions of properties of concave functions.

Define \( \lambda_\delta = \exp \left( \frac{2\sigma \| \delta \|_2 \Phi^{-1}(p_0) - \| \delta \|_2^2}{2\sigma^2} \right) \), simple calculation can show \( \frac{\partial F(\delta, \lambda)}{\partial \lambda} \bigg|_{\lambda = \lambda_\delta} = 0 \), which means
\[ \min_{\| \delta \|_2 \leq r} \max_{\lambda \geq 0} F(\delta, \lambda) = \min_{\| \delta \|_2 \leq r} \left\{ 0 + \Phi \left( \frac{\| \delta \|_2}{2\sigma} + \frac{\sigma \ln \lambda_\delta}{\| \delta \|_2} \right) \right\} \]
\[ = \min_{\| \delta \|_2 \leq r} \Phi \left( \Phi^{-1}(p_0) - \frac{\| \delta \|_2}{\sigma} \right) \]
This tells us
\[ \min_{\| \delta \|_2 \leq r} \max_{\lambda \geq 0} F(\delta, \lambda) > 1/2 \iff \Phi \left( \Phi^{-1}(p_0) - \frac{r}{\sigma} \right) > 1/2 \iff r < \sigma \cdot \Phi^{-1}(p_0), \]
I.e. the certification radius is \( \sigma \cdot \Phi^{-1}(p_0) \). This is exactly the core theoretical contribution of [2]. This bound has a straightforward expansion for multi-class classification situations, we refer interesting readers to Appendix A.

A.4 Proof For Theorem 2 and 3

A.4.1 Proof for \( \ell_2 \) and \( \ell_\infty \) cases

Here we consider a more general smooth distribution \( \pi_0(z) \propto \| z \|_{\infty}^{-k_1} \| z \|_2^{-k_2} \exp \left( -\frac{\| z \|_2^2}{2\sigma^2} \right) \), for some \( k_1, k_2 \geq 0 \) and \( \sigma > 0 \). We first gives the following key theorem shows that \( D_{\mathbb{F}_{[0,1]} \lambda \pi_0 \| \pi_\delta} \) increases as \( |\delta_i| \) becomes larger for every dimension \( i \).

**Theorem 4.** Suppose \( \pi_0(z) \propto \| z \|_{\infty}^{-k_1} \| z \|_2^{-k_2} \exp \left( -\frac{\| z \|_2^2}{2\sigma^2} \right) \), for some \( k_1, k_2 \geq 0 \) and \( \sigma > 0 \), for any \( \lambda \geq 0 \) we have
\[ \text{sgn}(\delta_i) \frac{\partial}{\partial \delta_i} D_{\mathbb{F}_{[0,1]} \lambda \pi_0 \| \pi_\delta} \geq 0, \]
for any \( i \in \{1, 2, \ldots, d\} \).

Theorem 2 and 3 directly follows the above theorem. Notice that in Theorem 2 as our distribution is spherical symmetry, it is equivalent to set \( \mathcal{B} = \{ \delta : \delta = [a, 0, \ldots, 0]^T, a \leq r \} \) by rotating the axis.

**Proof.** Given \( \lambda, k_1 \) and \( k_2 \), we define \( \phi_1(s) = s^{-k_1}, \phi_2(s) = s^{-k_2}e^{-\frac{s^2}{2\sigma^2}} \). Notice that \( \phi_1 \) and \( \phi_2 \) are monotone decreasing for non-negative \( s \). By the symmetry, without loss of generality, we assume \( \delta = [\delta_1, \ldots, \delta_d]^T \) for \( \delta_i \geq 0, i \in [d] \). Notice that
\[ \frac{\partial}{\partial \delta_i} \| x_0 - \delta \|_{\infty} = \mathbb{I}\{\| x_0 - \delta \|_{\infty} = |x_i - \delta_i|\} \frac{\partial}{\partial \delta_i} \sqrt{(x_i - \delta_i)^2} = \mathbb{I}\{\| x_0 - \delta \|_{\infty} = |x_i - \delta_i|\} \frac{(x_i - \delta_i)}{\| x_0 - \delta \|_{\infty}}. \]
And also
\[ \frac{\partial}{\partial \delta_i} ||x_0 - \mu||_2 = \frac{\partial}{\partial \delta_i} \sqrt{\sum_i (x_i - \mu_i)^2} = \frac{-(x_i - \mu_i)}{||x_0 - \mu||_2}. \]

We thus have
\[ \frac{\partial}{\partial \delta_i} \int (\lambda \pi_0(x_0) - \pi_\delta(x_0))_+ dx_0 = -\int I \{ \lambda \pi_0(x_0) \geq \pi_\delta(x_0) \} \frac{\partial}{\partial \delta_i} \pi_\delta(x_0) dx_0 \\
= \int I \{ \lambda \pi_0(x_0) \geq \pi_\delta(x_0) \} F_1 (||x_0 - \delta||_\infty, ||x_0 - \delta||_2) dx_0 \\
= \int I \{ \lambda \pi_0(x_0) \geq \pi_\delta(x_0), x_1 > \delta_1 \} F_1 (||x_0 - \delta||_\infty, ||x_0 - \delta||_2) dx_0 \\
+ \int I \{ \lambda \pi_0(x_0) \geq \pi_\delta(x_0), x_1 < \delta_1 \} F_1 (||x_0 - \delta||_\infty, ||x_0 - \delta||_2) dx_0, \]

where we define
\[ F_1 (||x_0 - \delta||_\infty, ||x_0 - \delta||_2) = \phi'_1 (||x_0 - \delta||_\infty) \phi_2 (||x_0 - \delta||_2) I\{ ||x_0 - \delta||_\infty = |x_1 - \delta_1| \} \frac{(x_1 - \delta_1)}{||x_0 - \delta||_\infty} \\
+ \phi_1 (||x_0 - \delta||_\infty) \phi'_2 (||x_0 - \delta||_2) \frac{(x_1 - \delta_1)}{||x_0 - \delta||_2}. \]

Notice that as \( \phi'_1 \leq 0 \) and \( \phi'_2 \leq 0 \) and we have
\[ \int I \{ \lambda \pi_0(x_0) \geq \pi_\delta(x_0), x_1 > \delta_1 \} F_1 (||x_0 - \delta||_\infty, ||x_0 - \delta||_2) dx_0 \leq 0 \]
\[ \int I \{ \lambda \pi_0(x_0) \geq \pi_\delta(x_0), x_1 < \delta_1 \} F_1 (||x_0 - \delta||_\infty, ||x_0 - \delta||_2) dx_0 \geq 0. \]

Our target is to prove that \( \frac{\partial}{\partial \delta_i} \int (\lambda \pi_0(x_0) - \pi_\delta(x_0))_+ dx_0 \geq 0. \) Now define the set
\[ H_1 = \{ x_0 : \lambda \pi_0(x_0) \geq \pi_\delta(x_0), x_1 > \delta_1 \} \]
\[ H_2 = \{ [2\delta_1 - x_1, x_2, ..., x_d]^T : x_0 = [x_1, ..., x_d]^T \in H_1 \}. \]

Here the set \( H_2 \) is defined as a image of a bijection
\[ proj(x_0) = [2\delta_1 - x_1, x_2, ..., x_d]^T = \bar{x}_0, \]
that is constrained on the set \( H_1. \) Notice that under our definition,
\[ \int I \{ \lambda \pi_0(x_0) \geq \pi_\delta(x_0), x_1 > \delta_1 \} F_1 (||x_0 - \delta||_\infty, ||x_0 - \delta||_2) dx_0 \\
= \int_{H_1} F_1 (||x_0 - \delta||_\infty, ||x_0 - \delta||_2) dx_0. \]

Now we prove that
\[ \int I \{ \lambda \pi_0(x_0) \geq \pi_\delta(x_0), x_1 < \delta_1 \} F_1 (||x_0 - \delta||_\infty, ||x_0 - \delta||_2) dx_0 \]
\[ \geq \int_{H_2} F_1 (||x_0 - \delta||_\infty, ||x_0 - \delta||_2) dx_0 \]
\[ \geq \left[ \int_{H_1} F_1 (||x_0 - \delta||_\infty, ||x_0 - \delta||_2) dx_0 \right]. \]
Property of the projection  Before we prove the (1) and (2), we give the following property of the
defined projection function. For any \( \hat{x}_0 = \text{proj}(x_0) \), \( x_0 \in H_1 \), we have
\[
\| x_0 - \delta \|_\infty = \| \hat{x}_0 - \delta \|_\infty
\]
\[
\| x_0 - \delta \|_2 = \| \hat{x}_0 - \delta \|_2
\]
\[
\| x_0 \|_2 \geq \| \hat{x}_0 \|_2
\]
\[
\| x_0 \|_\infty \geq \| \hat{x}_0 \|_\infty.
\]
This is because
\[
\hat{x}_i = x_i, \; i \in [d] - \{1\}
\]
\[
\hat{x}_1 = 2\delta_1 - x_1,
\]
and by the fact that \( x_1 \geq \delta_1 \geq 0 \), we have \( |\hat{x}_1| \leq |x_1| \) and \( |\hat{x}_1 - \delta_1| \leq |x_1 - \delta_1| \).

Proof of Equality (2)  By the fact that \( \text{proj} \) is bijective constrained on the set \( H_1 \) and the property
of \( \text{proj} \), we have
\[
\int_{H_2} F_1 (\| x_0 - \delta \|_\infty, \| \hat{x}_0 - \delta \|_2) \, dx_0
\]
\[
= \int_{H_2} \phi'_1 (\| x_0 - \delta \|_\infty) \phi_2 (\| \hat{x}_0 - \delta \|_2) \mathbb{I} (\| x_0 - \delta \|_\infty = \| \hat{x}_0 - \delta \|_\infty = |\hat{x}_1 - \delta_1|) \frac{(\hat{x}_1 - \delta_1)}{\| \hat{x}_0 - \delta \|_2} \, dx_0
\]
\[
+ \int_{H_1} \phi_1 (\| x_0 - \delta \|_\infty) \phi'_2 (\| x_0 - \delta \|_2) \frac{(\delta_1 - x_1)}{\| \hat{x}_0 - \delta \|_2} \, dx_0 \quad (\ast)
\]
\[
= \int_{H_1} \phi_1 (\| x_0 - \delta \|_\infty) \phi'_2 (\| x_0 - \delta \|_2) \frac{(\delta_1 - x_1)}{\| \hat{x}_0 - \delta \|_2} \, dx_0
\]
\[
= \int_{H_1} F_1 (\| x_0 - \delta \|_\infty, \| \hat{x}_0 - \delta \|_2) \, dx_0,
\]
where \( \ast \) is by change of variable \( \hat{x}_0 = \text{proj}(x_0) \) and \( J \) is the Jacobian matrix \( J =
\begin{bmatrix}
-1 & 0 & \cdots & 0 \\
0 & 1 & \cdots & 0 \\
\vdots & \vdots & \ddots & \vdots \\
0 & 0 & \cdots & 1
\end{bmatrix}
\)
and here we have the fact that \( \hat{x}_1 - \delta_1 = (2\delta_1 - x_1) - \delta_1 = -(x_1 - \delta_1) \).

Proof of Inequality (1)  This can be done by verifying that \( H_2 \subseteq \{ x_0 : \lambda \pi_0(x_0) \geq \pi_\delta(x_0), x_1 < \delta \} \). By the property of the projection, for any \( x_0 \in H_1 \), let \( x_0 = \text{proj}(x_0) \), then \( \lambda \pi_0(x_0) \geq \lambda \pi_0(x_0) \geq \lambda \pi_\delta(x_0) = \pi_\delta(x_0) \) (by the fact that \( t \phi_1 \) and \( \phi_2 \) are monotone decreasing). It implies that for any \( x_0 \in H_2 \), we have \( \lambda \pi_0(x_0) \geq \pi_\delta(x_0) \) and thus \( H_2 \subseteq \{ x_0 : \lambda \pi_0(x_0) \geq \pi_\delta(x_0), x_1 < \delta \} \).

Final statement  By the above result, we have
\[
\frac{\partial}{\partial \delta_1} \int (\lambda \pi_0(x_0) - \pi_\delta(x_0)) \, dx_0 \geq 0,
\]
and the same result holds for any \( \frac{\partial}{\partial \delta_i} \int (\lambda \pi_0(x_0) - \pi_\delta(x_0)) \, dx_0, \; i \in [d] \), which implies our result.

A.4.2  Proof for \( \ell_1 \) case
Slightly different for former cases, apart from proving \( \frac{\partial}{\partial \delta_i} \mathbb{D}_{\mathcal{F}_{[0,1]}} (\lambda \pi_0 \parallel \pi_\delta) \geq 0 \) for \( \forall \delta_i \geq 0 \), we also need to demonstrate

Theorem 5.  Suppose \( \pi_0(x_0) \propto \| x_0 \|^{-k} \exp \left( -\frac{\| x_0 \|_1}{b} \right) \), then for \( \delta = (r, d - r, \delta_3, \delta_4, \ldots) \) and
\( \delta = (0, d, \delta_3, \delta_4, \ldots) \), \( 0 < r < d \), we have
\[
\mathbb{D}_{\mathcal{F}_{[0,1]}} (\lambda \pi_0 \parallel \pi_\delta) \geq \mathbb{D}_{\mathcal{F}_{[0,1]}} (\lambda \pi_0 \parallel \pi_\delta)
\]

17
Proof. We turn to show that
\[ \frac{\partial}{\partial r} \mathcal{D}_{\mathcal{F}_{[0,1]}}(\lambda \pi_0 \parallel \pi_\delta) \leq 0, \]
for \( \delta = (r, d - r, \delta_3, \delta_4, \cdots) \) and \( r < d/2 \). We define \( \phi(s) = s^{-k} \exp(-\frac{\pi}{s}). \) With
\[ \frac{\partial}{\partial \delta_i} \| x_0 - \delta \|_1 = \frac{\partial}{\partial \delta_i} |x_i - \delta_i| = -\text{sgn}(x_i - \delta_i) = \frac{\delta_i - x_i}{|x_i - \delta_i|}, \]
We have
\[ \frac{\partial}{\partial r} \mathcal{D}_{\mathcal{F}_{[0,1]}}(\lambda \pi_0 \parallel \pi_\delta) = -\int I \{\lambda \pi_0(x_0) \geq \pi_\delta(x_0)\} \frac{\partial}{\partial r} \pi_\delta(x_0) dx_0 \]
\[ = \int \| \lambda \pi_0(x_0) \geq \pi_\delta(x_0) \} F(x_0) dx_0, \]
where
\[ F(x_0) = -\frac{\partial}{\partial r} \phi(\|x_0 - \delta\|_1) = -\phi'(\|x_0 - \delta\|_1) \frac{\partial}{\partial r} \|x_0 - \delta\|_1 \]
\[ = \phi'(\|x_0 - \delta\|_1) \frac{\partial}{\partial r} (|x_1 - r| + |x_2 - d + r|) \]
\[ = \phi'(\|x_0 - \delta\|_1) \cdot (\text{sgn}(x_1 - r) + \text{sgn}(d - x_2 - r)). \]
Thus the original derivative becomes
\[ = \int I \{\lambda \pi_0(x_0) \geq \pi_\delta(x_0), x_1 > r, x_2 < d - r \} F(x_0) dx_0 \]
\[ + \int \| \lambda \pi_0(x_0) \geq \pi_\delta(x_0), x_1 > r, x_2 > d - r \} F(x_0) dx_0 \]
\[ + \int \| \lambda \pi_0(x_0) \geq \pi_\delta(x_0), x_1 < r, x_2 > d - r \} F(x_0) dx_0 \]
\[ + \int \| \lambda \pi_0(x_0) \geq \pi_\delta(x_0), x_1 < r, x_2 < d - r \} F(x_0) dx_0 \]
\[ = 2 \int \| \lambda \pi_0(x_0) \geq \pi_\delta(x_0), x_1 > r, x_2 < d - r \} \phi'(\|x_0 - \delta\|_1) dx_0 \]
\[ - 2 \int \| \lambda \pi_0(x_0) \geq \pi_\delta(x_0), x_1 < r, x_2 > d - r \} \phi'(\|x_0 - \delta\|_1) dx_0 \]
We only need to show that
\[ \int I \{\lambda \pi_0(x_0) \geq \pi_\delta(x_0), x_1 > r, x_2 < d - r \} \phi'(\|x_0 - \delta\|_1) dx_0 \geq \]
\[ \int \| \lambda \pi_0(x_0) \geq \pi_\delta(x_0), x_1 < r, x_2 > d - r \} \phi'(\|x_0 - \delta\|_1) dx_0. \]
Notice that \( r < d/2 \), therefore this can be proved with a similar projection \( x_0 \mapsto \tilde{x}_0: \)
\[ (x_1, x_2, x_3, x_4, \cdots) \mapsto (2r - x_1, 2d - 2r - x_2, x_3, x_4, \cdots) \]
and the similar deduction as previous theorem.

\[ \square \]

A.5 Theoretical Demonstration about the Ineffetivity of Equation (12)

Theorem 6. Consider the adversarial attacks on the \( \ell_\infty \) ball \( B_{\ell_\infty, r} = \{ \delta : \| \delta \|_\infty \leq r \} \). Suppose we use the smoothing distribution \( \pi_0 \) in Equation (12) and choose the parameters \((k, \sigma)\) such that
1) \( \|z\|_\infty \) is stochastic bounded when \( z \sim \pi_0 \), in that for any \( \epsilon > 0 \), there exists a finite \( M > 0 \) such that \( P_{\pi_0}(|z| > M) \leq \epsilon \);

2) the mode of \( \|z\|_\infty \) under \( \pi_0 \) equals \( Cr \), where \( C \) is some fixed positive constant, then for any \( \epsilon \in (0, 1) \) and sufficiently large dimension \( d \), there exists a constant \( t > 1 \), such that,

\[
\max_{\delta \in B_{\|z\|_\infty}} \left\{ D_{\mathcal{F}_{[0,1]}}(\lambda \pi_0 \| \pi_\delta) \right\} \geq (1 - \epsilon) \left( \lambda - \mathcal{O}(t^{-d}) \right).
\]

This shows that, in very high dimensions, the maximum distance term is arbitrarily close to \( \lambda \) which is the maximum possible value of \( D_{\mathcal{F}_{[0,1]}}(\lambda \pi_0 \| \pi_\delta) \) (see Theorem 6). In particular, this implies that in high dimensional scenario, once \( f_{\pi_0}^*(x_0) \leq (1 - \epsilon) \) for some small \( \epsilon \), we have \( \mathcal{L}_{\pi_0}(\mathcal{F}_{[0,1]}; B_{\|z\|_\infty}) = \mathcal{O}(t^{-d}) \) and thus fail to certify.

**Remark** The condition 1) and 2) in Theorem 6 are used to ensure that the magnitude of the random perturbations generated by \( \pi_0 \) is within a reasonable range such that the value of \( f_{\pi_0}^*(x_0) \) is not too small, in order to have a high accuracy in the trade-off in Equation (5). Note that the natural images are often contained in cube \([0, 1]^d\). If \( \|z\|_\infty \) is too large to exceed the region of natural images, the accuracy will be obviously rather poor. Note that if we use variants of Gaussian distribution, we only need \( \|z\|_2/\sqrt{d} \) to be not too large. Theorem 6 says that once \( \|z\|_\infty \) is in a reasonably small scale, the maximum distance term must be unreasonably large in high dimensions, yielding a vacuous lower bound.

**Proof.** First notice that the distribution of \( z \) can be factorized by the following hierarchical scheme:

\[
a \sim \pi_R(a) \propto a^{d-1-k}e^{-\frac{a^2}{2\sigma^2}}1\{a \geq 0\} \\
s \sim \text{Unif}_{[d]}(-1, 1) \\
z \leftarrow \frac{s}{\|s\|_\infty}a.
\]

Without loss of generality, we assume \( \delta^* = [r, \ldots, r]^T \). (see Theorem 4)

\[
D_{\mathcal{F}_{[0,1]}}(\lambda \pi_0 \| \pi_\delta^*) = E_{z \sim \pi_0} \left( \lambda - \frac{\pi_\delta}{\pi_0}(z) \right)_{+}.
\]

Notice that as the distribution is symmetry,

\[
P_{\pi_0}(\|z + \delta^*\|_\infty = a + r \mid \|z\|_\infty = a) = \frac{1}{2}.
\]

Define \( |z|^{(j)} \) is the \( j \)-th order statistics of \( |z_j| \), \( j = 1, \ldots, d \) conditioning on \( \|z\|_\infty = a \). By the factorization above and some algebra, we have, for any \( \epsilon \in (0, 1) \),

\[
P\left( \frac{|z|^{(j-1)}}{|z|^{(d)}} > (1 - \epsilon) \mid \|z\|_\infty = a \right) \geq 1 - (1 - \epsilon)^{d-1}.
\]

And \( \frac{|z|^{(d-1)}}{|z|^{(d)}} \perp |z|^{(d)} \). Now we estimate \( D_{\mathcal{F}_{[0,1]}}(\lambda \pi_0 \| \pi_\delta^*) \).

\[
E_{z \sim \pi_0} \left( \lambda - \frac{\pi_\delta}{\pi_0}(z) \right)_{+} = E_a E_{z \sim \pi_0} \left( \lambda - \frac{\pi_\delta}{\pi_0}(z) \right)_{+} \mid \|z\|_\infty = a
\]

\[
= \frac{1}{2} E_a E_{z \sim \pi_0} \left( \lambda - \frac{\pi_\delta}{\pi_0}(z) \right)_{+} \mid \|z\|_\infty = a, \|z + \delta^*\|_\infty = a + r
\]

\[
+ \frac{1}{2} E_a E_{z \sim \pi_0} \left( \lambda - \frac{\pi_\delta}{\pi_0}(z) \right)_{+} \mid \|z\|_\infty = a, \|z + \delta^*\|_\infty \neq a + r.
\]
Conditioning on \(\|z\|_\infty = a, \|z + \delta^*\|_\infty = a + r\), we have
\[
\frac{\pi_{\delta}}{\pi_0}(z) = \left(\frac{1}{1 + \frac{r}{a}}\right)^k e^{-\frac{1}{2\sigma^2} (2ra + r^2)} = \left(\frac{1}{1 + \frac{r}{a}}\right)^k e^{-\frac{a+1-k}{2c^2} (2z+1)}.
\]
Here the second equality is because we choose \(\text{mode}(\|z\|_\infty) = Cr\), which implies that \(\sqrt{a - 1 - K\sigma} = Cr\). And thus we have
\[
\mathbb{E}_a \mathbb{E}_{z \sim \pi_0} \left[ \left( \lambda - \frac{\pi_{\delta}}{\pi_0}(z) \right) \mathbb{1}_{\|z\|_\infty = a, \|z + \delta^*\|_\infty = a + r} \right] = \int \left( \lambda - \left(1 + \frac{r}{a}\right)^{-k} e^{-\frac{a+1-k}{2c^2} (2\lambda+1)} \right) \pi(a) da
\]
for some \(t > 1\). Here the last equality is by the assumption that \(\|z\|_\infty = O_p(1)\).

Next we bound the second term \(\mathbb{E}_a \mathbb{E}_{z \sim \pi_0} \left[ \left( \lambda - \frac{\pi_{\delta}}{\pi_0}(z) \right) \mathbb{1}_{\|z\|_\infty = a, \|z + \delta^*\|_\infty \neq a + r} \right] \). By the property of uniform distribution, we have
\[
P\left( \frac{|z|}{|z|} > (1 - \epsilon) \mid \|z\|_\infty = a, \|z + \delta^*\|_\infty \neq a + r \right) = 1 - (1 - \epsilon)^{d-1}.
\]
And thus, for any \(\epsilon \in (0, 1)\),
\[
P\left( \|z + \delta^*\|_\infty \geq ((1 - \epsilon)a + r)^2 \mid \|z\|_\infty = a, \|z + \delta^*\|_\infty \neq a + r \right) \geq \frac{1}{2} \left( 1 - (1 - \epsilon)^{d-1} \right).
\]
It implies that
\[
\mathbb{E}_{z \sim \pi_0} \left[ \left( \lambda - \frac{\pi_{\delta}}{\pi_0}(z) \right) \mathbb{1}_{\|z\|_\infty = a, \|z + \delta^*\|_\infty = a + r} \right] \geq \frac{1}{2} (1 - (1 - \epsilon)^{d-1}) \left( \lambda - \left(1 + \frac{r}{a}\right)^{-k} e^{-\frac{a+1-k}{2c^2} (2\lambda+1)} \right) + \frac{1}{2} (1 - (1 - \epsilon)^{d-1}) \left( \lambda - \left(1 + \frac{r}{a}\right)^{-k} e^{-\frac{d-k}{2c^2} \left( \epsilon(1-\epsilon)^2/a^2 + 2(1-\epsilon)^2 r^2 \right)} \right).
\]
For any \(\epsilon' \in (0, 1)\), by choosing \(\epsilon = \frac{\log(2/\epsilon')}{d-1}\), for large enough \(d\), we have
\[
\mathbb{E}_{z \sim \pi_0} \left[ \left( \lambda - \frac{\pi_{\delta}}{\pi_0}(z) \right) \mathbb{1}_{\|z\|_\infty = a, \|z + \delta^*\|_\infty = a + r} \right] \geq \frac{1}{2} (1 - (1 - \epsilon)^{d-1}) \left( \lambda - \left(1 + \frac{r}{a}\right)^{-k} e^{-\frac{d-k}{2c^2} (2(1-\epsilon)a/r + 1)} e^{\frac{d\log(2/\epsilon')}{c^2r^2}} \right) + \frac{1}{2} (1 - (1 - \epsilon)^{d-1}) \left( \lambda - \left(1 - \frac{\log(2/\epsilon')}{d-1}\right)^{-k} e^{-\frac{d-k}{2c^2} (2(1-\epsilon)a/r + 1)} e^{\frac{d\log(2/\epsilon')}{c^2r^2}} \right).
\]
Thus we have

$$\frac{1}{2} \mathbb{E}_a \mathbb{E}_{z \sim \pi_0} \left[ \left( \lambda - \frac{\pi_0}{\pi_0}(z) \right)_+ | \| z \|_\infty = a, \| z + \delta^* \|_\infty \neq a + r \right]$$

$$= \frac{1}{2} \left( 1 - \epsilon' \right) \left( \lambda - O(t^{-d}) \right).$$

Combine the bounds, for large $d$, we have

$$\mathbb{D}_{\mathcal{F}_{[0,1]}} (\lambda \pi_0 \parallel \pi_{\delta^*}) = \left( 1 - \epsilon' \right) \left( \lambda - O(t^{-d}) \right).$$

\[\square\]

\section*{B More about Experiments}

\subsection*{B.1 Practical Algorithm}

In this section, we give our algorithm for certification. Our target is to give a high probability bound for the solution of

$$\mathcal{L}_{\pi_0}(\mathcal{F}_{[0,1]} \mathcal{B}_{\ell_\infty}, \epsilon) = \max_{\lambda \geq 0} \left\{ \lambda f^\sharp_{\pi_0} - \mathbb{D}_{\mathcal{F}_{[0,1]}} (\lambda \pi_0 \parallel \pi_{\delta^*}) \right\}$$

given some classifier $f^\sharp$. Following \cite{9}, the given classifier here has a binary output $\{0, 1\}$. Computing the above quantity requires us to evaluate both $f^\sharp_{\pi_0}$ and $\mathbb{D}_{\mathcal{F}_{[0,1]}} (\lambda \pi_0 \parallel \pi_{\delta^*})$. A lower bound $\hat{p}_0$ of the former term is obtained through binominal test as \cite{9} do, while the second term can be estimated with arbitrary accuracy using Monte Carlo samples. We perform grid search to optimize $\lambda$ and given $\lambda$, we draw $N$ i.i.d. samples from the proposed smoothing distribution $\pi_0$ to estimate $\lambda f^\sharp_{\pi_0} - \mathbb{D}_{\mathcal{F}_{[0,1]}} (\lambda \pi_0 \parallel \pi_{\delta^*})$. This can be achieved by the following importance sampling manner:

$$\lambda f^\sharp_{\pi_0} - \mathbb{D}_{\mathcal{F}_{[0,1]}} (\lambda \pi_0 \parallel \pi_{\delta^*})$$

$$\geq \lambda \hat{p}_0 - \int \left( \lambda - \frac{\pi_{\delta^*}}{\pi_0}(z) \right)_+ \pi_0(z) dz$$

$$\geq \lambda \hat{p}_0 - \frac{1}{N} \sum_{i=1}^{N} \left( \lambda - \frac{\pi_{\delta^*}}{\pi_0}(z_i) \right)_+ - \epsilon.$$

And we use reject sampling to obtain samples from $\pi_0$. Notice that, we restrict the search space of $\lambda$ to a finite compact set so the importance samples is bounded. Since the Monte Carlo estimation is not exact with an error $\epsilon$, we give a high probability concentration lower bound of the estimator. Algorithm \[\square\] summarized our algorithm.
Algorithm 1 Certification algorithm

**Input:** input image $x_0$; original classifier: $f^\sharp$; smoothing distribution $\pi_0$; radius $r$; search interval $[\lambda_{\text{start}}, \lambda_{\text{end}}]$ of $\lambda$; search precision $h$ for optimizing $\lambda$; number of samples $N_1$ for testing $p_0$; pre-defined error threshold $\epsilon$; significant level $\alpha$;

compute search space for $\lambda$: $\Lambda = \text{range}(\lambda_{\text{start}}, \lambda_{\text{end}}, h)$

compute $N_2$: number of Monte Carlo estimation given $\epsilon, \alpha$ and $\Lambda$

compute optimal disturb: $\delta$ depends on specific setting for $\lambda$ in $\Lambda$

for $\lambda \in \Lambda$
    sample $z_1, \ldots, z_{N_1} \sim \pi_0$
    compute $n_1 = \frac{1}{N_1} \sum_{i=1}^{N_1} f^\sharp(x_0 + z_i)$
    compute $\hat{p}_0 = \text{LowerConfBound}(n_1, N_1, 1 - \alpha)$
    sample $z_1, \ldots, z_{N_2} \sim \pi_0$
    compute $\hat{D}_{F[0,1]}(\lambda \pi_0 \| \pi_\delta) = \frac{1}{N_2} \sum_{i=1}^{N_2} \left( \lambda - \frac{\pi_\delta(z_i)}{\pi_0(z_i)} \right)_+$
    compute confidence lower bound $b_\lambda = \lambda \hat{p}_0 - \hat{D}_{F[0,1]}(\lambda \pi_0 \| \pi_\delta) - \epsilon$
end
if $\max_{\lambda \in \Lambda} b_\lambda \geq 1/2$ then
    $x_0$ can be certified
else
    $x_0$ cannot be certified
end

The LowerConfBound function performs a binomial test as described in [9]. The $\epsilon$ in Algorithm 1 is given by concentration inequality.

**Theorem 7.** Let $h(z_1, \ldots, z_N) = \frac{1}{N} \sum_{i=1}^{N} \left( \lambda - \frac{\pi_\delta(z_i)}{\pi_0(z_i)} \right)_+$, we yield

$$
\Pr\{|h(z_1, \ldots, z_N) - \int (\lambda \pi_0(z) - \pi_\delta(z))_+ dz| \geq \varepsilon\} \leq \exp \left(-\frac{2N\varepsilon^2}{\lambda^2}\right).
$$

**Proof.** Given McDiarmid’s Inequality, which says

$$
\sup_{x_1, x_2, \ldots, x_n, \hat{x}_i} |h(x_1, x_2, \ldots, x_n) - h(x_1, x_2, \ldots, x_{i-1}, \hat{x}_i, x_{i+1}, \ldots, x_n)| \leq c_i \quad \text{for} \quad 1 \leq i \leq n,
$$

we have $c_i = \frac{\lambda}{N}$, and then obtain

$$
\Pr\{|h(z_1, \ldots, z_N) - \int (\lambda \pi_0(z) - \pi_\delta(z))_+ dz| \geq \varepsilon\} \leq \exp \left(-\frac{2N\varepsilon^2}{\lambda^2}\right).
$$

The above theorem tells us that, once $\epsilon, \lambda, N$ is given, we can yield a bound with high-probability $1 - \alpha$. One can also get $N$ when $\epsilon, \lambda, \alpha$ is provided. Note that this is the same as the Hoeffding bound mentioned in Section 4.2 as McDiarmid bound is a generalization of Hoeffding bound.

However, in practice we can use a small trick as below to certify with much less computation:
Algorithm 2 Practical certification algorithm

**Input:** input image $x_0$; original classifier: $f^*$; smoothing distribution $\pi_0$; radius $r$; search interval for $\lambda$: $[\lambda_{\text{start}}, \lambda_{\text{end}}]$; search precision $h$ for optimizing $\lambda$; number of Monte Carlo for first estimation: $N_0^0, N_1^0$; number of samples $N_1$ for a second test of $p_0$; pre-defined error threshold $\epsilon$; significant level $\alpha$; optimal perturbation $\delta$ ($\delta = [r, 0, \ldots, 0]^T$ for $\ell_2$ attacking and $\delta = [r, \ldots, r]^T$ for $\ell_{\infty}$ attacking).

for $\lambda$ in $\Lambda$ do
  sample $z_1, \ldots, z_{N_1^0} \sim \pi_0$
  compute $n_1^0 = \frac{1}{N_1^0} \sum_{i=1}^{N_1^0} f^*(x_0 + z_i)$
  compute $\hat{p}_0 = \text{LowerConfBound}(n_1^0, N_1^0, 1 - \alpha)$
  sample $z_1, \ldots, z_{N_2^0} \sim \pi_0$
  compute $\bar{D}_{f_{[0,1]}}(\lambda \pi_0 \parallel \pi_\delta) = \frac{1}{N_2^0} \sum_{i=1}^{N_2^0} \left( \lambda - \frac{\pi_\delta}{\pi_0}(z_i) \right)_+$
  compute confidence lower bound $b_\lambda = \lambda \hat{p}_0 - \bar{D}_{f_{[0,1]}}(\lambda \pi_0 \parallel \pi_\delta)$

end

compute $\lambda^* = \arg \max_{\lambda \in \Lambda} b_\lambda$

compute $N_2$: number of Monte Carlo estimation given $\epsilon, \alpha$ and $\lambda$
  sample $z_1, \ldots, z_{N_2} \sim \pi_0$
  compute $n_1 = \frac{1}{N_1} \sum_{i=1}^{N_1} f^*(x_0 + z_i)$
  compute $\hat{p}_0 = \text{LowerConfBound}(n_1, N_1, 1 - \alpha)$
  sample $z_1, \ldots, z_{N_2} \sim \pi_0$
  compute $\bar{D}_{f_{[0,1]}}(\lambda \pi_0 \parallel \pi_\delta) = \frac{1}{N_2} \sum_{i=1}^{N_2} \left( \lambda - \frac{\pi_\delta}{\pi_0}(z_i) \right)_+$

compute $b = \lambda^* \hat{p}_0 - \bar{D}_{f_{[0,1]}}(\lambda \pi_0 \parallel \pi_\delta) - \epsilon$

if $b \geq 1/2$ then
  $x_0$ can be certified
else
  $x_0$ cannot be certified
end

Algorithm 2 allow one to begin with small $N_0^0, N_2^0$ to obtain the first estimation and choose a $\lambda^*$. Then a rigorous lower bound can be achieved with $\lambda^*$ with enough (i.e. $N_1, N_2$) Monte Carlo samples.

### B.2 Experiment Settings

The details of our method are shown in the supplementary material. Since our method requires Monte Carlo approximation, we draw $0.1M$ samples from $\pi_0$ and construct $\alpha = 99.9\%$ confidence lower bounds of that in Equation (2). The optimization on $\lambda$ is solved using grid search. For $\ell_2$ attacks, we set $k = 500$ for CIFAR-10 and $k = 50000$ for ImageNet in our non-Gaussian smoothing distribution Equation (11). If the used model was trained with a Gaussian perturbation noise of $\mathcal{N}(0, \sigma_0^2)$, then the $\sigma$ parameter of our smoothing distribution is set to be $\sqrt{(d - 1)/(d - 1 - k)}\sigma_0$, such that the expectation of the norm $\|z\|_2$ under our non-Gaussian distribution Equation (11) matches with the norm of $\mathcal{N}(0, \sigma_0^2)$. For $\ell_1$ situation, we keep the same rule for hyperparameter selection as $\ell_2$ case, in order to make the norm of proposed distribution has the same mean with original distribution. For $\ell_{\infty}$ situation, we set $k = 250$ and $\sigma$ also equals to $\sqrt{(d - 1)/(d - 1 - k)}\sigma_0$ for the mixed norm smoothing distribution Equation (13) just for consistency. More ablation study about $k$ is deferred to Appendix B.3.

### B.3 Ablation Study

On CIFAR10, we also do ablation study to show the influence of different $k$ for the $\ell_2$ certification case as shown in Table 3.
Table 4: Certified top-1 accuracy of the best classifiers on cifar10 at various $\ell_2$ radius. We use the same model as [9] and do not train any new models.

<table>
<thead>
<tr>
<th>$\ell_2$ Radius</th>
<th>0.25</th>
<th>0.50</th>
<th>0.75</th>
<th>1.00</th>
<th>1.25</th>
<th>1.50</th>
<th>1.75</th>
<th>2.00</th>
<th>2.25</th>
</tr>
</thead>
<tbody>
<tr>
<td>Baseline (%)</td>
<td>60</td>
<td>43</td>
<td>34</td>
<td>23</td>
<td>17</td>
<td>14</td>
<td>12</td>
<td>10</td>
<td>8</td>
</tr>
<tr>
<td>$k = 100$ (%)</td>
<td>60</td>
<td>43</td>
<td>34</td>
<td>23</td>
<td>18</td>
<td>15</td>
<td>12</td>
<td>10</td>
<td>8</td>
</tr>
<tr>
<td>$k = 200$ (%)</td>
<td>60</td>
<td>44</td>
<td>36</td>
<td>24</td>
<td>18</td>
<td>15</td>
<td>13</td>
<td>10</td>
<td>8</td>
</tr>
<tr>
<td>$k = 500$ (%)</td>
<td>61</td>
<td>46</td>
<td>37</td>
<td>25</td>
<td>19</td>
<td>16</td>
<td>14</td>
<td>11</td>
<td>9</td>
</tr>
<tr>
<td>$k = 1000$ (%)</td>
<td>59</td>
<td>44</td>
<td>36</td>
<td>25</td>
<td>19</td>
<td>16</td>
<td>14</td>
<td>11</td>
<td>9</td>
</tr>
<tr>
<td>$k = 2000$ (%)</td>
<td>56</td>
<td>41</td>
<td>35</td>
<td>24</td>
<td>19</td>
<td>16</td>
<td>15</td>
<td>12</td>
<td>9</td>
</tr>
</tbody>
</table>

C Illumination about Bilateral Condition

The results in the main context is obtained under binary classification setting. Here we show it has a natural generalization to multi-class classification setting. Suppose the given classifier $f^2$ classifies an input $x_0$ correctly to class $A$, i.e.,

$$f^2_A(x_0) > \max_{B \neq A} f^2_B(x_0) \quad (14)$$

where $f^2_B(x_0)$ denotes the prediction confidence of any class $B$ different from ground truth label $A$. Notice that $f^2_A(x_0) + \sum_{B \neq A} f^2_B(x_0) = 1$, so the necessary and sufficient condition for correct binary classification $f^2_A(x_0) > 1/2$ becomes a sufficient condition for multi-class prediction.

Similarly, the necessary and sufficient condition for correct classification of the smoothed classifier is

$$\min_{f \in F} \left\{ \mathbb{E}_{z \sim p_0} [f_A(x_0 + \delta + z)] \mid \mathbb{E}_{\pi_0} [f_A(x_0)] = f^2_A(x_0) \right\} >$$

$$\max_{f \in F} \left\{ \mathbb{E}_{z \sim p_0} [f_B(x_0 + \delta + z)] \mid \mathbb{E}_{\pi_0} [f_B(x_0)] = f^2_B(x_0) \right\}$$

for $\forall B \neq A$ and any perturbation $\delta \in \mathcal{B}$. Writing out their Langragian forms makes things clear:

$$\max_{\lambda} \lambda f^2_{\pi_0,A}(x_0) - \mathbb{D}_{\mathcal{F}_{[0,1]}}(\lambda p_0 || p_\delta) > \min_{\lambda} \max_{B \neq A} \lambda f^2_{\pi_0,B}(x_0) + \mathbb{D}_{\mathcal{F}_{[0,1]}}(\pi_\delta || \lambda p_0)$$

Thus the overall necessary and sufficient condition is

$$\min_{\delta \in \mathcal{B}} \left\{ \max_{\lambda} \left( \lambda f^2_{\pi_0,A}(x_0) - \mathbb{D}_{\mathcal{F}_{[0,1]}}(\lambda p_0 || p_\delta) \right) - \max_{\lambda} \min_{B \neq A} \left( \lambda f^2_{\pi_0,B}(x_0) + \mathbb{D}_{\mathcal{F}_{[0,1]}}(\pi_\delta || \lambda p_0) \right) \right\} > 0$$

Optimizing this bilateral object will theoretically give a better certification result than our method in main context, especially when the number of classes is large. But we do not use this bilateral formulation as reasons stated below.

When both $p_0$ and $p_\delta$ are gaussian, which is [9]'s setting, this condition is equivalent to:

$$\min_{\delta \in \mathcal{B}} \left\{ \Phi \left( \Phi^{-1}(f^2_{\pi_0,A}(x_0)) - \frac{||\delta||_2}{\sigma} \right) - \max_{B \neq A} \Phi \left( \Phi^{-1}(f^2_{\pi_0,B}(x_0)) + \frac{||\delta||_2}{\sigma} \right) \right\} > 0$$

$$\Leftrightarrow \quad \Phi^{-1}(f^2_{\pi_0,A}(x_0)) - \frac{r}{\sigma} > \Phi^{-1}(f^2_{\pi_0,B}(x_0)) + \frac{r}{\sigma}, \quad \forall B \neq A$$

$$\Leftrightarrow \quad r < \frac{\sigma}{2} \left( \Phi^{-1}(f^2_{\pi_0,A}(x_0)) - \Phi^{-1}(f^2_{\pi_0,B}(x_0)) \right), \forall B \neq A$$

with a similar derivation process like Appendix [3]. This is exactly the same bound in the (restated) theorem 1 of [3].

[9] use $1 - p_A$ as a naive estimate of the upper bound of $f^2_{\pi_0,B}(x_0)$, where $p_A$ is a lower bound of $f^2_{\pi_0,A}(x_0)$. This leads the confidence bound decay to the bound one can get in binary case, i.e., $r \leq \sigma \Phi^{-1}(f^2_{\pi_0,A}(x_0))$.

As the two important baselines [9, 10] do not take the bilateral form, we also do not use this form in experiments for fairness.

---

3In fact, the theoretical part of [12] share some similar discussion with this section.