Supplementary Materials

Contents

A Proofs for "TWO MOTIVATING EXAMPLES" **2**

A.1 Proofs for Sec 3.1 2

B Definitions for "THEORETICAL CHARACTERIZATION OF THE SHRINKING PHENOMENON" **5**

B.1 Definitions 5

C Proofs for "THEORETICAL CHARACTERIZATION OF THE SHRINKING PHENOMENON" **6**

C.1 Proofs for Sec 4.1 6

C.2 Proofs for Sec 4.2 9

D Additional Analysis **13**

D.1 Shrinking effect for unidimensional data 13

D.2 Bounded decision region behaviors 13

D.3 Semi-bounded decision region certified radius behaviors w.r.t data dimensions 14

A Proofs for "TWO MOTIVATING EXAMPLES"

A.1 Proofs for Sec 3.1

Lemma 1. $\Phi[x] + \Phi[\frac{1}{x}] \geq 1.5$ *with equality holds iff* $x \in \{0, \infty\}$.

Proof. Let $f(x) = \Phi[x] + \Phi[\frac{1}{x}]$. We observe that $f(x) = f(1/x)$ by definition. So, it is sufficient to show that for x in the interval $(1, \infty)$, $f(x) \geq 1.5$ with equality at $x \to \infty$. We prove this by showing that in the interval $(1, \infty)$, $f(x)$ is strictly decreasing and $\lim_{x\to\infty} f(x) = \lim_{x\to\infty} \Phi(x) + \Phi(1/x) = \Phi(\infty) + \Phi(0) = 1 + 0.5 = 1.5$. To show f(x) is strictly decreasing we proceed by taking the derivative wrt x,

$$\frac{d}{dx} f(x) = \frac{e^{-\frac{x^2}{2}}}{\sqrt{2\pi}} - \frac{e^{-\frac{1}{2x^2}}}{x^2\sqrt{2\pi}}$$

we show that for the interval $(1, \infty)$ this derivative is less than 0. So, we need to show that

$$\begin{aligned}
&\frac{e^{-\frac{x^2}{2}}}{\sqrt{2\pi}} - \frac{e^{-\frac{1}{2x^2}}}{x^2\sqrt{2\pi}} < 0 \\
\Leftrightarrow \quad & x^2 e^{-\frac{x^2}{2}} < e^{-\frac{1}{2x^2}} \\
\Leftrightarrow \quad & \log\left(x^2\right) + \frac{1}{2x^2} < x^2 \\
& \text{Let } t = \log\left(x^2\right), x > 1 \to t > 0 \\
\Leftrightarrow \quad & 2t < e^t - e^{-t}
\end{aligned}$$

This holds for $t > 0$ as we have that at $t = 0$. $2 \cdot 0 = 0 = e^0 - e^{-0}$ and $2t$ increases at a rate of 2 while $e^t - e^{-t}$ increases at a rate of $e^t + e^{-t} > 2 \cdot \sqrt{e^t \cdot e^{-t}} = 2$ as $t > 1 \to e^t \neq e^{-t}$. Finally for $x = 1$, we calculate $f(x) \approx 1.6829 > 1.5$. □

Theorem 1. *Consider a classifier* f_{train,σ_t} *given as the naive-Bayes classifier obtained by training on the dataset* $\mathcal{X}$ *with data augmentation of variance* σ_t*. Let the class-wise accuracy of* f_{train,σ_t} *using the randomized smoothing prediction rule be given as* $Acc_1(\sigma_t), Acc_2(\sigma_t)$*. Then we define the bias* ($\Delta(\sigma_t)$) *to be the gap between class-wise accuracies* ($\Delta(\sigma_t) = |Acc_1(\sigma_t) - Acc_2(\sigma_t)|$)*. For* $k > \frac{1}{2\epsilon} - 1$*, class I decision region grows in size at a rate of* $O(\sigma_t^2)$ *and thus the bias is large for large* σ_t*.*

Proof. In order to determine the accuracies we start by looking at the decision regions given by the two classifiers. We show that the decision region of class 1 increases with increasing σ effectively increasing the bias by increasing the class 1 accuracy while decreasing the class 2 accuracy.

From the structure of the dataset it is easy to show that the naive Bayes classifier yield decision regions:
class 1 : $[-(\frac{a}{2} + c_0(\sigma)), \frac{ka}{2} + d_0(\sigma)]$
class 2 : $[-\infty, -(\frac{a}{2} + c_0(\sigma))] \cup [\frac{ka}{2} + d_0(\sigma), +\infty]$
The likelihood ratio function $r_\sigma(x) = \frac{p(x \in \text{ class 2})}{p(x \in \text{ class 1})} = (1 - 2\epsilon)e^{\frac{-a(2x+a)}{2\sigma^2}} + 2\epsilon e^{\frac{(2x-ka)ka}{2\sigma^2}}$. This is a convex function is x resulting in the previous form of decision regions. Thus, we get the following decision regions after smoothing, class 1 $[-(\frac{a}{2} + c_1(\sigma)), \frac{ka}{2} + d_1(\sigma)]$ and the rest being class 2.

In this case we show that for $c_0(\sigma)$ grows at $\Theta(\sigma^2)$ with increasing σ by establishing a lower bound and upper bound which both grow at the rate of $O(\sigma^2)$.
For the lower bound consider the function $r_\sigma^u(x) = (1 - 2\epsilon)e^{\frac{ax}{\sigma^2}} + 2\epsilon e^{\frac{-kax}{\sigma^2}} > r_\sigma(-(\frac{a}{2} + x))$. If for any $c_l(\sigma)$ we have $r_\sigma^u(c_l(\sigma)) = 1$, then $r_\sigma(-(\frac{a}{2} + c_l(\sigma))) < 1$. Thus, we see that using the convexity argument from before $c_0(\sigma) > c_l(\sigma)$. But it is easy to see that if $c_l(1)$ is a solution of the equation $r_1^u(x) = 1$ at $\sigma = 1$, then $\sigma^2 c_l(1)$ is a solution for $r_\sigma^u(x) = 1$.
As r_1^u is a continuous function with $r_1^u(0) = 1$ and $\lim_{x\to\infty} r_1^u(x) \to \infty$, it is sufficient to show that $\frac{d}{dx} r_1^u(0) = a(1 - 2\epsilon(k+1)) < 0$ (follows from the case condition) to show that $r_1^u(x) = 1$ has a positive real solution and consequently $c_0(\sigma) > \sigma^2 c_l(1) = O(\sigma^2)$. From the likelihood function, we can also clearly see that $r_\sigma(-(\frac{a}{2} + x)) > (1 - 2\epsilon)e^{\frac{ax}{\sigma^2}}$. Using this we can establish that $c_0(\sigma) < \frac{\sigma^2 - \log(1-2\epsilon)}{a}$ making $c_0(\sigma) = \Theta(\sigma^2)$.

As $d_0(\sigma) \geq 0$, we have that for all $\sigma \in (0, \infty)$ the size of the interval $[-(\frac{a}{2} + c_0(\sigma)), \frac{ka}{2} + d_0(\sigma)]$ is bigger that $C\sigma^2 + C$ for some positive constant C. Thus, we have that at $x = -(\frac{a}{2} + c_0(\sigma) - \frac{1}{C})$ the probability $x \in$ Class I after smoothing is given as $\Phi(\frac{C\sigma^2}{\sigma_t}) - \Phi(\frac{-1/C}{\sigma_t})$. By Lemma 1, we get that $\Phi(\frac{C\sigma^2}{\sigma_t}) - \Phi(\frac{-1}{C\sigma_t}) > \Phi(\sigma_t C) - \Phi(\frac{-1}{C\sigma_t}) = \Phi(\sigma_t C) - (1 - \Phi(\frac{1}{C\sigma_t})) = \Phi(\sigma_t C) + \Phi(\frac{1}{C\sigma_t}) - 1 > 0.5$. Thus, we have $c_1(\sigma) > c_0(\sigma) - \frac{1}{C}$. Combining this with the fact that clearly $c_0(\sigma) > c_1(\sigma)$, we have $c_1(\sigma) \in (c_0(\sigma) - \frac{1}{C}, c_0(\sigma))$ and similarly, we also have $d_1(\sigma) \in (d_0(\sigma) - \frac{1}{C}, d_0(\sigma))$. This also gives us $c_1(\sigma) = \Theta(\sigma^2) = \Theta(\sigma_t^2)$.

Consider the function $f_x(\sigma) = r_\sigma(x)$. By differentiating this function wrt σ we see that it has only one extremum point. Using the fact that $\lim_{\sigma \to \infty} f_x(\sigma) = 1$ we have that if for any x, $f_x(\sigma) = 1$ then we see that there the extremum point lies between σ and ∞. If for any $\sigma' > \sigma$, $f_x(\sigma') = 1$, then there would be a two extremum points one between σ, σ' and another between σ', ∞. Using this along with the continuity of f_x we get that either $f_x(\sigma') < 1 \forall \sigma' > \sigma$ or $f_x(\sigma') > 1 \forall \sigma' > \sigma$. We can further use the fact that $f_x(0) \to \infty$ to see that f_x is decreasing at σ making $f_x(\sigma') < 1 \forall \sigma' > \sigma$. Thus, we see that $d_0(\sigma), c_0(\sigma)$ are increasing functions of σ. Combining this with the previous result shows that the decision region of class I after smoothing increases at $O(\sigma_t^2)$.

For the bias we see that as $\sigma_t \to \infty$, class I at least occupies the region $(-\infty, \frac{ka}{2}]$ while class II occupies at most the region $(\frac{ka}{2}, \infty)$. As a result the bias is lower bounded by $(1 - \Phi(\frac{-ka}{2\sigma_o})) - \epsilon(1 - \Phi(\frac{-ka}{2\sigma_o})) = (1 - \epsilon)(1 - \Phi(\frac{-ka}{2\sigma_o}))$ which is very high. □

Theorem 2. *Consider a classifier f_{train,σ_t} given as the naive-Bayes classifer obtained by training on the dataset $\mathcal{X}'$ with data augmentation of variance σ_t. The bias of the classifier f_{train,σ_t} using the randomized smoothing prediction rule is $1 - \epsilon$, if $k > \frac{e^2}{\epsilon} - 1$ and $\sigma_t \geq a\sqrt{\frac{k(k+1)}{2ln(2\epsilon(k+1)) - \frac{2k}{k+2}}}$.*

Proof. At $x = -a$, we see that if the decision region for class 1 is $[-(a + c), \frac{ka}{2} + d]$, then the probability after smoothing is

$$\begin{aligned}
g(-a, 1) &= \int_{x' \in \mathbb{R}^d} d(-a, x')\psi(x', 1)dx' \\
&= \int_{-(a+c)}^{\frac{ka}{2}+d} d(-a, x')dx' \\
&= \int_{-\infty}^{\frac{ka}{2}+d} d(-a, x')dx' - \int_{-\infty}^{-(a+c)} d(-a, x')dx' \\
&= \Phi(\frac{\frac{ka}{2} + d + a}{\sigma}) - \Phi(\frac{-c}{\sigma}) \\
&\geq \Phi(\frac{\frac{k+2}{2}a}{\sigma}) - \Phi(\frac{-c}{\sigma}) \quad (\text{if } d \geq 0) \\
&\geq \Phi(\frac{\frac{k+2}{2}a}{\sigma}) - \Phi(-\frac{\sigma}{\frac{k+2}{2}a}) \quad (\text{if } c \geq \frac{2\sigma^2}{(k+2)a}) \\
&> 0.5. \quad (\text{by Lemma 1})
\end{aligned}$$

That's said, the bias will be atleast $1 - \epsilon$ if $d \geq 0$ and $c \geq \frac{2\sigma^2}{(k+2)a}$ are true. We now check for $d \geq 0$: for $x \in [0, \frac{ka}{2}]$,

$$\begin{aligned}
\psi(x, 1) &= \int_{x' \in \mathbb{R}^d} d(x, x')\rho(x', 1)dx' && = d(x, 0)\rho(0, 1) \\
&= \frac{1}{\sqrt{2\pi\sigma^2}}[\frac{1}{2}e^{-\frac{x^2}{2\sigma^2}}] && = \frac{1}{\sqrt{2\pi\sigma^2}}[(\frac{1}{2} - \epsilon)e^{-\frac{x^2}{2\sigma^2}} + \epsilon e^{-\frac{x^2}{2\sigma^2}}] \\
&> \frac{1}{\sqrt{2\pi\sigma^2}}[(\frac{1}{2} - \epsilon)e^{-\frac{(x+a)^2}{2\sigma^2}} + \epsilon e^{-\frac{(ka-x)^2}{2\sigma^2}}] && = d(x, -a)\rho(-a, 2) + d(x, ka)\rho(ka, 2) \\
&= \int_{x' \in \mathbb{R}^d} d(x, x')\rho(x', 2)dx' = \psi(x, 2),
\end{aligned}$$

implying $x \in [0, \frac{ka}{2}]$ belongs to class 1 for the naive bayes classifier. Therefore the decision region for class 1

extends at least to $\frac{ka}{2}+d$ with $d\geq 0$. Next, we check for $c\geq\frac{2\sigma^2}{(k+2)a}$: at $x=-a-\frac{2\sigma^2}{(k+2)a}$, the probability is

$$\psi(-a-\frac{2\sigma^2}{(k+2)a},1)=\int_{x'\in\mathbb{R}^d}d(-\frac{2\sigma+a}{(k+2)\frac{a}{\sigma}},x')\rho(x',1)dx' \quad =\frac{1}{\sqrt{2\pi\sigma^2}}[\frac{1}{2}e^{-\frac{x^2}{2\sigma^2}}]\big|_{x=-a-\frac{2\sigma^2}{(k+2)a}}$$
$$\psi(-a-\frac{2\sigma^2}{(k+2)a},2)=\int_{x'\in\mathbb{R}^d}d(-\frac{2\sigma+a}{(k+2)\frac{a}{\sigma}},x')\rho(x',2)dx' \quad =\frac{1}{\sqrt{2\pi\sigma^2}}[(\frac{1}{2}-\epsilon)e^{-\frac{(x+a)^2}{2\sigma^2}}+\epsilon e^{-\frac{(ka-x)^2}{2\sigma^2}}]\big|_{x=-a-\frac{2\sigma^2}{(k+2)a}}.$$

Therefore we see that $\psi(-a-\frac{2\sigma^2}{(k+2)a},1)>\psi(-a-\frac{2\sigma^2}{(k+2)a},2)$ if

$$\begin{aligned}
&(1-2\epsilon)e^{(\frac{a}{\sigma})^2\frac{1}{2}+\frac{2}{k+2}}+2\epsilon e^{-\frac{k(k+2)}{2}(\frac{a}{\sigma})^2-\frac{2k}{k+2}}<1\\
\Leftrightarrow\quad&(1-2\epsilon)[e^{(\frac{a}{\sigma})^2\frac{1}{2}+\frac{2}{k+2}}-1]<2\epsilon[1-e^{-\frac{k(k+2)}{2}(\frac{a}{\sigma})^2-\frac{2k}{k+2}}]\\
\Leftrightarrow\quad&\frac{1}{2\epsilon}-1<\frac{1-e^{-\frac{k(k+2)}{2}(\frac{a}{\sigma})^2-\frac{2k}{k+2}}}{e^{(\frac{a}{\sigma})^2\frac{1}{2}+\frac{2}{k+2}}-1}\\
\Leftrightarrow\quad&\frac{1}{2\epsilon}<\frac{e^{(\frac{a}{\sigma})^2\frac{1}{2}+\frac{2}{k+2}}-e^{-\frac{k(k+2)}{2}(\frac{a}{\sigma})^2-\frac{2k}{k+2}}}{e^{(\frac{a}{\sigma})^2\frac{1}{2}+\frac{2}{k+2}}-1}\\
\Leftrightarrow\quad&\frac{1}{2\epsilon}<\frac{\tau l-\tau^{-k(k+2)}l^{-k}}{\tau l-1}\quad(\text{let }\tau=e^{(\frac{a}{\sigma})^2\frac{1}{2}},l=e^{\frac{2}{k+2}})\\
\Leftrightarrow\quad&\frac{1}{2\epsilon}<\tau^{-k(k+2)}l^{-k}\frac{\tau^{(k+1)^2}l^{k+1}-1}{\tau l-1}\\
\Leftrightarrow\quad&\frac{1}{2\epsilon}<\tau^{-k(k+2)}l^{-k}\frac{\tau^{(k+1)^2}l^{k+1}-1}{\tau^{k+1}l-1}\frac{\tau^{k+1}l-1}{\tau l-1}\\
\Leftrightarrow\quad&\frac{1}{2\epsilon}<\tau^{-k(k+1)}l^{-k}(\Sigma_{i=0}^{k}(\tau^{k+1}l)^i)\frac{\tau^{k+1}l-1}{\tau l-1}\tau^{-k}\\
\Leftrightarrow\quad&\frac{1}{2\epsilon}<(\Sigma_{i=0}^{k}(\tau^{k+1}l)^{-i})\frac{\tau l-\tau^{-k}}{\tau l-1}\\
\Leftarrow\quad&\frac{1}{2\epsilon}\leq\Sigma_{i=0}^{k}(\tau^{k+1}l)^{-i}\leq(k+1)(\tau^{k+1}l)^{-k}\\
\Leftrightarrow\quad&0<ln(\tau)\leq\frac{ln(2\epsilon(k+1))-kln(l)}{k(k+1)}=\frac{ln(2\epsilon(k+1))-\frac{2k}{k+2}}{k(k+1)}\\
\Leftarrow\quad&(\frac{a}{\sigma})^2\frac{1}{2}\leq\frac{ln(2\epsilon(k+1))-\frac{2k}{k+2}}{k(k+1)},\ k>\frac{e^2}{\epsilon}-1\\
\Leftrightarrow\quad&\sigma\geq a\sqrt{\frac{k(k+1)}{2ln(2\epsilon(k+1))-\frac{2k}{k+2}}},\ k>\frac{e^2}{\epsilon}-1.
\end{aligned}$$

These conclude our proof. □

B Definitions for "THEORETICAL CHARACTERIZATION OF THE SHRINKING PHENOMENON"

B.1 Definitions

Definition 1 (Smoothed). *If we use f to denote an original neural network function with outputs in the simplex $\Delta^c = \{z \in \mathbb{R}^c \mid \sum_{i=1}^c z_i = 1, 0 \leq z_i \leq 1, \forall i\}$, then its smoothed counterpart defined on d-dimensional inputs $x \in \mathbb{R}^d$ is defined by*

$$f_{smooth}(x) = \int_{x' \in \mathbb{R}^d} f(x')p(x')dx',$$

where $p(x')$ is the probability density function of the filter.

Definition 2 (Gaussian smoothing). *If $p(x')$ is the probability density function of a normally-distributed random variable with an expected value x and standard deviation σ, then we call f_{smooth} a Gaussian-smoothed function and denote it by f_σ.*

Definition 3 (Bounded Decision Regions). *If the decision region (disconnected or connected) of class 1 data is a bounded set in the Euclidean space (can be bounded by a ball of finite radius), then we call these decision regions bounded decision regions.*

Definition 4 (Shrinking of Bounded Decision Regions). *A bounded decision region is distinguished as shrinked after applying smoothing filters if the radius R_σ of $S_{\mathcal{D}_\sigma}$ is rigorously smaller than the radius R of $S_{\mathcal{D}}$, i.e. $R_\sigma < R$, where $S_{\mathcal{D}}$ and $S_{\mathcal{D}_\sigma}$ are the smallest balls containing the original decision region and the smoothed decision region, respectively.*

Definition 5 (Unbounded Decision Regions). *If for any ball there exists at least one point in the decision regions that reside outside the ball, then we call these decision regions unbounded decision regions.*

Definition 6 (Semi-bounded Decision Regions). *For an unbounded decision region, if there exists any half-space $\mathcal{H}$ (decided by a hyperplane) that contains the unbounded decision region, then we call it semi-bounded decision region. We say a semi-bounded decision region is bounded in v-direction if there $\exists k \in \mathbb{R}/\infty$ such that for $\forall x \in \mathcal{D}$, $v^T x < k$.*

Definition 7 (Shrinking of Semi-bounded Decision Regions). *A semi-bounded decision region bounded in v-direction is distinguished as shrinked along the direction after applying smoothing filters if the upper bound of projections of the decision region onto direction v shrinks, i.e. $\Upsilon^v_{\mathcal{D}_\sigma} < \Upsilon^v_{\mathcal{D}}$, where $\Upsilon^v_{\mathcal{D}} = \max_{x \in \mathcal{D}} v^T x, \Upsilon^v_{\mathcal{D}_\sigma} = \max_{x \in \mathcal{D}_\sigma} v^T x$.*

Definition 8 (θ, v-Bounding Cone for a Decision Region). *A θ, v cone is defined as a right circular cone $\mathcal{C}$ with axis along $-v$ and aperture 2θ. Then we define the θ, v-bounding cone $\mathcal{C}^{\mathcal{D}}_{\theta,v}$ for $\mathcal{D}$ as the θ, v cone that has the smallest projection on v and contains $\mathcal{D}$, i.e., $\mathcal{C}^{\mathcal{D}}_{\theta,v} = \arg\min_{\mathcal{D} \subseteq \mathcal{C}_{\theta,v}} \Upsilon^v_{\mathcal{C}_{\theta,v}}$.*

C Proofs for "THEORETICAL CHARACTERIZATION OF THE SHRINKING PHENOMENON"

C.1 Proofs for Sec 4.1

Lemma 2. *For any two original decision regions A, B, if we have that $A \subset B$, then we also have that $A_\sigma \subset B_\sigma$, where A_σ and B_σ are the decision regions of the Gaussian-smoothed functions.*

Proof. Recalling that decision regions A_σ and B_σ satisfy $D_\sigma = \{x \in \mathbb{R}^d | f_\sigma^D(x)_1 \geq \frac{1}{c}\}$ for $D = A, B$. Therefore for $\forall x \in A_\sigma$, we have $f_\sigma^A(x) \geq \frac{1}{c}$. And

$$\begin{aligned} f_\sigma^B(x)_1 &= \int_{x' \in \mathbb{R}^d} f^B(x')_1 p(x') dx' = \int_{x' \in \mathbb{R}^d} \mathbb{1}_{x' \in B} p(x') dx' \\ &= \int_{x' \in B} p(x') dx' > \int_{x' \in A} p(x') dx' \\ &= \int_{x' \in \mathbb{R}^d} \mathbb{1}_{x' \in A} p(x') dx' = \int_{x' \in \mathbb{R}^d} f^A(x')_1 p(x') dx' \\ &= f_\sigma^A(x)_1 \geq \frac{1}{c}, \end{aligned}$$

implying $x \in B_\sigma$. That said, we have that if $x \in A_\sigma$, then $x \in B_\sigma$, making $A_\sigma \subseteq B_\sigma$. □

Corollary 1. *The smallest ball $S_{\mathcal{D}_\sigma}$ containing the smoothed decision region is contained within the smoothed version of $S_\mathcal{D}$, i.e. $S_{\mathcal{D}_\sigma} \subseteq (S_\mathcal{D})_\sigma$.*

Proof. As we have $\mathcal{D} \subseteq S_\mathcal{D}$, from Lemma 2 we get $\mathcal{D}_\sigma \subseteq (S_\mathcal{D})_\sigma$. Then by isotropy we have that $(S_\mathcal{D})_\sigma$ is also a ball centered at the same point as $S_\mathcal{D}$. As $S_{\mathcal{D}_\sigma}$ is the smallest ball containing $\mathcal{D}_\sigma$, we have that $S_{\mathcal{D}_\sigma} \subseteq (S_\mathcal{D})_\sigma$. □

We also need another important definition for the coming theorem, the regularized Gamma function:

Definition 9 (Regularized Gamma Function)**.** *The lower regularized gamma functions $Q(s, x)$ is defined by*

$$Q(s, x) = \frac{\int_0^x t^{s-1} e^{-t} dt}{\int_0^\infty t^{s-1} e^{-t} dt}.$$

Moreover, it is well-known that

$$Q\left(\frac{d}{2}, \frac{R^2}{2\sigma^2}\right) = \int_{x' \in \mathbb{R}^d, \|x'\|_2 \leq R} (2\pi\sigma^2)^{-\frac{d}{2}} e^{\frac{x'^T x'}{2\sigma^2}} dx'.$$

We also give a short proof of this in the proof of Theorem 4. For the number of dimensions d, we summarize the lemma based on regularized Gamma functions below.

Lemma 3. *For $\forall d, c \in \mathbb{N}^+$, $Q(\frac{d}{2}, \frac{d}{2c}) < \frac{1}{c}$ holds.*

Proof. To prove $Q(\frac{d}{2}, \frac{d}{2c}) < \frac{1}{c})$, by definition 9, we aim at proving $\int_0^\infty t^{\frac{d}{2}-1} e^{-t} dt > c \cdot \int_0^{\frac{d}{2c}} t^{\frac{d}{2}-1} e^{-t} dt$ $(\forall d \in \mathbb{N}^+)$. For c = 1, this is clearly true as $t^{x-1} e^{-t} \geq 0$ is true for $t \geq 0$. Then we show it also holds for $c \geq 2$.

Let $g(t) = t^{x-1} e^{-t}$, we have $g'(t) = t^{x-2} e^{-t}(x - 1 - t)$. Therefore $g(t)$ is increasing when $t \leq x - 1$ and decreasing when $t > x - 1$. Thus, giving us two equations

$$\begin{aligned} &\int_{\frac{x}{c}}^x t^{x-1} e^{-t} dt > \min\{x^{x-1} e^{-x}, (\frac{x}{c})^{x-1} e^{-\frac{x}{c}}\} \frac{(c-1)x}{c} \\ &\frac{x}{c} (\frac{x}{c})^{x-1} e^{-\frac{x}{c}} > \int_0^{\frac{x}{c}} t^{x-1} e^{-t} dt \end{aligned}$$

So, we see that for any x, c if we have $x^{x-1}e^{-x} \geq (\frac{x}{c})^{x-1}e^{-\frac{x}{c}}$ then $\int_{\frac{x}{c}}^{x} t^{x-1}e^{-t}dt > (c-1)\cdot\int_0^{\frac{x}{c}} t^{x-1}e^{-t}dt \Leftrightarrow \int_0^x t^{x-1}e^{-t}dt > c\cdot\int_0^{\frac{x}{c}} t^{x-1}e^{-t}dt$. Using $t^{x-1}e^{-t} \geq 0, \forall x \int_0^\infty t^{x-1}e^{-t}dt \geq \int_0^x t^{x-1}e^{-t}dt$. So, we have $\int_0^\infty t^{x-1}e^{-t}dt > c\cdot\int_0^{\frac{x}{c}} t^{x-1}e^{-t}dt$ as needed. So, for any x, c it is sufficient to show

$$x^{x-1}e^{-x} \geq (\frac{x}{c})^{x-1}e^{-\frac{x}{c}}$$

in order to prove $\int_0^\infty t^{x-1}e^{-t}dt > c\cdot\int_0^{\frac{x}{c}} t^{x-1}e^{-t}dt$. The inequality can be re-written as $(x-1)\log(c) > \frac{c-1}{c}x$ or $(1-\frac{1}{x}) > (1-\frac{1}{c})\frac{1}{\log(c)}$. We observe that $(1-\frac{1}{c})\frac{1}{\log(c)}$ is a decreasing function of c for $c \geq 1$ and $(1-\frac{1}{x})$ is an increasing function of x.
For $x \geq 4, c \geq 2$, we see $(1-\frac{1}{x}) \geq 1-\frac{1}{4} = 0.75 > (1-\frac{1}{2})\frac{1}{\log(2)} \geq (1-\frac{1}{c})\frac{1}{\log(c)}$.
For $x \geq \frac{3}{2}, c \geq 20$, we have $(1-\frac{1}{x}) \geq 1-\frac{2}{3} > (1-\frac{1}{20})\frac{1}{\log(20)} \geq (1-\frac{1}{c})\frac{1}{\log(c)}$.
For $\frac{3}{2} \leq x < 4 \rightarrow 3 \leq d < 8$ and $2 \leq c < 20$, we numerically verify the values of $Q(\frac{d}{2}, \frac{d}{2c})$ to see the inequality is satisfied.
Thus, for $d \geq 3, c \geq 2$ we have the inequality.

For $d = 2$, we have $Q(\frac{d}{2}, \frac{d}{2c}) = Q(1, \frac{1}{c})$. This has a closed form solution $Q(1, x) = 1 - e^{-x}$. So, we need to show that for $c \geq 2$ $1 - e^{-\frac{1}{c}} < \frac{1}{c}$ or $e^{\frac{1}{c}} < \frac{c}{c-1}$ or $\frac{1}{c} < \log\Big(1 + \frac{1}{c-1}\Big)$. But we know that for $x > -1, x \neq 0$, $\log(1+x) > \frac{x}{x+1}$, so $\log\Big(1 + \frac{1}{c-1}\Big) > \frac{\frac{1}{c-1}}{1+\frac{1}{c-1}} = \frac{1}{c}$ which concludes the proof for $d = 2, c \geq 2$. □

Theorem 3. *A bounded decision region shrinks after applying Gaussian smoothing filters with large σ, i.e. if $\sigma > \frac{R\sqrt{c}}{\sqrt{2(d-1)}}$, then $R_\sigma < R$, where R and R_σ are the radii of $S_\mathcal{D}$ and $S_{\mathcal{D}_\sigma}$, the smallest balls bounding the original decision region and the smoothed decision region, respectively.*

Proof. Considering the ball $S_\mathcal{D}$, we see that from Corollary 1, $\mathcal{D}_\sigma \subseteq (S_\mathcal{D})_\sigma$. Thus, we see that by the definition of radius $R_{\mathcal{D}_\sigma} \leq R_{(S_\mathcal{D})_\sigma}$. It is sufficient to show that for large σ, $R_{(S_\mathcal{D})_\sigma} < R_{S_\mathcal{D}}$. Then we observe that due to the isotropic nature of Gaussian smoothing, $(S_\mathcal{D})_\sigma$ is also a sphere concentric to $S_\mathcal{D}$. So, it is sufficient to show that for a point x at distance $R_{S_\mathcal{D}}$ from the center x_0 of the sphere, $f_\sigma(x)_1 < \frac{1}{c}$.
Without loss of generality consider $\mathcal{D}$ to be the origin-centered sphere of radius R and $x = [0, \ldots, 0, R]^T$. It is sufficient to show for large σ $f_\sigma(x)_1 < \frac{1}{c}$. By definition 2, we have

$$\begin{aligned} f_\sigma(x)_1 &= \int_{x' \in \mathbb{R}^d} f(x')_1 p(x')dx' \\ &= \int_{\|x'\|_2 \leq R} (2\pi)^{-\frac{d}{2}}|\Sigma|^{-\frac{1}{2}} e^{-\frac{1}{2}(x'-x)^T\Sigma^{-1}(x'-x)}dx' \\ &= \int_{\|x'\|_2 \leq R} (2\pi\sigma^2)^{-\frac{d}{2}} e^{-\frac{(x'-x)^T(x'-x)}{2\sigma^2}}dx'. \end{aligned} \tag{1}$$

Then substituting the value of x, we get the equation.

$$\begin{aligned} f_\sigma(x)_1 &= \int_{\|x'\|_2 \leq R} (2\pi\sigma^2)^{-\frac{d}{2}} e^{\frac{\sum_{i=1}^{d-1} x_i'^2 + (x_d' - R)^2}{2\sigma^2}}dx' \\ &= \int_{-R}^{R}\int_{\sum_{k=1}^{d-1} {x_k'}^2 \leq R^2 - {x_d'}^2} (2\pi\sigma^2)^{-\frac{d}{2}} e^{-\frac{\sum_{k=1}^{d-1}(x_k' - x_k)^2}{2\sigma^2}}dx_1' \ldots dx_{d-1}' e^{-\frac{(x_d' - x_d)^2}{2\sigma^2}}dx_d' \\ &< \int_{-R}^{R}\int_{\sum_{k=1}^{d-1} {x_k'}^2 \leq R^2} (2\pi\sigma^2)^{-\frac{d}{2}} e^{-\frac{\sum_{k=1}^{d-1}(x_k' - x_k)^2}{2\sigma^2}}dx_1' \ldots dx_{d-1}' e^{-\frac{(x_d' - x_d)^2}{2\sigma^2}}dx_d' \\ &= \Big(\int_{-R}^{R} (2\pi\sigma^2)^{-\frac{1}{2}} e^{-\frac{(x_d' - x_d)^2}{2\sigma^2}}dx_d'\Big)\Big(\int_{\sum_{k=1}^{d-1} {x_k'}^2 \leq R^2} (2\pi\sigma^2)^{-\frac{d-1}{2}} e^{-\frac{\sum_{k=1}^{d-1}(x_k' - x_k)^2}{2\sigma^2}}dx_1' \ldots dx_{d-1}'\Big) \\ &= (\Phi(\frac{2R}{\sigma}) - \Phi(0)) \cdot Q(\frac{d-1}{2}, \frac{R^2}{2\sigma^2}) \end{aligned}$$

$$< \frac{1}{2} \cdot Q(\frac{d-1}{2}, \frac{R^2}{2\sigma^2}).$$

Using Lemma 3 we get that for $d \geq 3$, if $\frac{R^2}{2\sigma^2} \leq \frac{d-1}{c}$, then we have $\frac{1}{2} \cdot Q(\frac{d-1}{2}, \frac{R^2}{2\sigma^2}) < \frac{1}{c}$. Now, $\frac{R^2}{2\sigma^2} < \frac{d-1}{c}$ gives

$$\sigma > \frac{R\sqrt{c}}{\sqrt{2(d-1)}}.$$

□

For class 1 data x, the point at the origin has the highest probability to be classified as class 1, *i.e.* $f_\sigma(x)_1 = \int_{x' \in \mathbb{R}^d} f(x')p(x')dx' = \int_{\|x'\|_2 \leq R} (2\pi)^{-\frac{d}{2}} |\Sigma|^{-\frac{1}{2}} e^{-\frac{1}{2}(x'-x)^T \Sigma^{-1}(x'-x)} dx' \leq f_\sigma(0)_1$.

Lemma 4. *Assume the decision region of class 1 data is $\{x \in \mathbb{R}^d \mid \|x\|_2 \leq R\}$, the point at the origin has the highest probability to be classified as class 1 by the gaussian-smoothed classifier f_σ, i.e. $f_\sigma(x)_1 \leq f_\sigma(0)_1$.*

Proof. We do the proof by mathematical induction and begin by giving $d = 1$ case. For $\forall R > 0$ and $d = 1$, Equation (1) reduces to

$$\begin{aligned}
f_\sigma(x)_1 &= \int_{-R}^{R} (2\pi\sigma^2)^{-\frac{1}{2}} e^{-\frac{(x'-x)^2}{2\sigma^2}} dx' \\
&\overset{a=x'-x}{=\!=\!=} \int_{-R-x}^{R-x} (2\pi\sigma^2)^{-\frac{1}{2}} e^{-\frac{a^2}{2\sigma^2}} da \\
f'_\sigma(x)_1 &= -(2\pi\sigma^2)^{-\frac{1}{2}} e^{-\frac{(R-x)^2}{2\sigma^2}} - (-1)(2\pi\sigma^2)^{-\frac{1}{2}} e^{-\frac{(-R-x)^2}{2\sigma^2}}
\end{aligned}$$

and $f'_\sigma(x)_1$ equals to zero only when $x = 0$. Now suppose the conclusion holds for $d-1$ dimensional case, then when $x \in \mathbb{R}^d$ we scale $f_\sigma(x)_1$ by $(2\pi\sigma^2)^{\frac{d}{2}}$ and obtain

$$\begin{aligned}
&\int_{\|x'\|_2 \leq R} e^{-\frac{(x'-x)^T(x'-x)}{2\sigma^2}} dx' \\
&= \int_{\sum_{k=1}^d {x'_k}^2 \leq R^2} e^{-\frac{\sum_{k=1}^d (x'_k - x_k)^2}{2\sigma^2}} dx' \\
&= \int_{-R}^{R} \int_{\sum_{k=1}^{d-1} {x'_k}^2 \leq R^2 - {x'_d}^2} e^{-\frac{\sum_{k=1}^{d-1}(x'_k - x_k)^2}{2\sigma^2}} dx'_1 \dots dx'_{d-1} e^{-\frac{(x'_d - x_d)^2}{2\sigma^2}} dx'_d \\
&\leq \int_{-R}^{R} \int_{\sum_{k=1}^{d-1} {x'_k}^2 \leq R^2 - {x'_d}^2} e^{-\frac{\sum_{k=1}^{d-1} {x'_k}^2}{2\sigma^2}} dx'_1 \dots dx'_{d-1} e^{-\frac{(x'_d - x_d)^2}{2\sigma^2}} dx'_d \\
&= \int_{\sum_{k=1}^d {x'_k}^2 \leq R^2} e^{-\frac{\sum_{k=1}^{d-1} {x'_k}^2}{2\sigma^2}} e^{-\frac{(x'_d - x_d)^2}{2\sigma^2}} dx' \\
&= \int_{\sum_{k=1}^{d-1} {x'_k}^2 \leq R^2} \int_{{x'_d}^2 \leq R^2 - \sum_{k=1}^{d-1} {x'_k}^2} e^{-\frac{(x'_d - x_d)^2}{2\sigma^2}} dx'_d e^{-\frac{\sum_{k=1}^{d-1} {x'_k}^2}{2\sigma^2}} dx'_1 \dots dx'_{d-1} \\
&\leq \int_{\sum_{k=1}^{d-1} {x'_k}^2 \leq R^2} \int_{{x'_d}^2 \leq R^2 - \sum_{k=1}^{d-1} {x'_k}^2} e^{-\frac{{x'_d}^2}{2\sigma^2}} dx'_d e^{-\frac{\sum_{k=1}^{d-1} {x'_k}^2}{2\sigma^2}} dx'_1 \dots dx'_{d-1} \\
&= \int_{\sum_{k=1}^d {x'_k}^2 \leq R^2} e^{-\frac{\sum_{k=1}^d {x'_k}^2}{2\sigma^2}} dx',
\end{aligned}$$

where the first inequality comes from the assumption that the conclusion holds for $d-1$ dimensional case with equality if and only if $x_1 = \dots x_{d-1} = 0$, and the second inequality comes from an one dimensional observation with equality precisely when $x_d = 0$. This concludes our proof. □

Since the value of $f_\sigma(0)_1$ depends on the radius R of the decision region, the dimension d, and the smoothing factor σ, we denote $f_\sigma(0)_1$ by $q(R, d, \sigma)$, *i.e.* $q(R, d, \sigma) := f_\sigma(0)_1$.

Theorem 4 (Vanishing Rate in the Ball-like Decision Region Case)**.** *The decision region of class 1 data vanishes at smoothing factor* $\sigma_{van} > \frac{R\sqrt{c}}{\sqrt{d}}$.

Proof. Noticing that the surface area of a d-dimensional ball of radius r is proportional to r^{d-1}, we can therefore write out the probability of the point at the origin be classified as class 1 as

$$\begin{aligned}
q(R,d,\sigma) &= \frac{\int_0^R r^{d-1}(\frac{1}{2\pi\sigma^2})^{\frac{d}{2}} e^{-\frac{r^2}{2\sigma^2}} dr}{\int_0^\infty r^{d-1}(\frac{1}{2\pi\sigma^2})^{\frac{d}{2}} e^{-\frac{r^2}{2\sigma^2}} dr} \\
&= \frac{\int_0^R r^{d-1} e^{-\frac{r^2}{2\sigma^2}} dr}{\int_0^\infty r^{d-1} e^{-\frac{r^2}{2\sigma^2}} dr} \\
&\xlongequal{t=\frac{r^2}{2\sigma^2}} \frac{\int_0^{\frac{R^2}{2\sigma^2}} (2\sigma^2 t)^{\frac{d-1}{2}} e^{-t}\sigma^2 (2\sigma^2 t)^{-\frac{1}{2}} dt}{\int_0^\infty (2\sigma^2 t)^{\frac{d-1}{2}} e^{-t}\sigma^2 (2\sigma^2 t)^{-\frac{1}{2}} dt} \\
&= \frac{\int_0^{\frac{R^2}{2\sigma^2}} t^{\frac{d}{2}-1} e^{-t} dt}{\int_0^\infty t^{\frac{d}{2}-1} e^{-t} dt} \\
&= Q(\frac{d}{2}, \frac{R^2}{2\sigma^2}).
\end{aligned}$$

Now let $\sigma = \sqrt{\frac{c}{d}}R$ yields $q(R,d,\sqrt{\frac{c}{d}}R) = Q(\frac{d}{2}, \frac{d}{2c})$. By Lemma 3, we then have $Q(\frac{d}{2}, \frac{d}{2c}) < \frac{1}{c}$, implying the decision region of class 1 data has already vanished and making $\sigma = \sqrt{\frac{c}{d}}R$ an upper bound of the vanishing smoothing factor. □

C.2 Proofs for Sec 4.2

Corollary 2. *As* $\mathcal{D} \subseteq \mathcal{C}^{\mathcal{D}}_{\theta,v}$*, using Lemma 2, we have that the smoothed decision region is contained within the smoothed version of* $\mathcal{C}^{\mathcal{D}}_{\theta,v}$*, i.e.* $\mathcal{D}_\sigma \subseteq (\mathcal{C}^{\mathcal{D}}_{\theta,v})_\sigma$. □

Lemma 5. *If the decision region of class 1 data is* $\mathcal{D} = \{x \in \mathbb{R}^d \mid v^T x + \|v\|\|x\| cos(\theta) \leq 0\}$*, where* $v = [0,\ldots,0,1]^T \in \mathbb{R}^d$ *and* $2\theta \in (-\pi,\pi)$*, then after smoothing among the set of points* S_a *with the same projection on* v *the point on the axis has the highest probability of being in class 1. For* $S_a = \{x \mid v^T x = a\}$*, we have* $argsup_{x\in S_a} f_\sigma(x)_1 = a \cdot v$*. Moreover if* $a_1 > a_2$*, then* $f_\sigma(a_1 \cdot v)_1 < f_\sigma(a_2 \cdot v)_1$.

Proof. For the first part of the proof consider the set of points $S_a = \{x \mid v^T x = a\}$. For any point x is S_a, we see that

$$\begin{aligned}
f_\sigma(x)_1 &= \int_{x'\in\mathbb{R}^d} f(x')_1 p(x') dx' \\
&= \int_{x'_d + \|x'\| cos(\theta) \leq 0} (2\pi\sigma^2)^{-\frac{d}{2}} e^{-\frac{(x'-x)^T(x'-x)}{2\sigma^2}} dx' \\
&= (2\pi\sigma^2)^{-\frac{d}{2}} \int_{-\infty}^0 \int_{\sum_{k=1}^{d-1} {x'_k}^2 \leq tan^2(\theta) {x'_d}^2} e^{-\frac{\sum_{k=1}^{d-1}(x'_k - x_k)^2}{2\sigma^2}} dx'_1 \ldots dx'_{d-1} e^{-\frac{(x'_d - a)^2}{2\sigma^2}} dx'_d \\
&\leq (2\pi\sigma^2)^{-\frac{d}{2}} \int_{-\infty}^0 \int_{\sum_{k=1}^{d-1} {x'_k}^2 \leq tan^2(\theta) {x'_d}^2} e^{-\frac{\sum_{k=1}^{d-1} {x'_k}^2}{2\sigma^2}} dx'_1 \ldots dx'_{d-1} e^{-\frac{(x'_d - a)^2}{2\sigma^2}} dx'_d \\
&= f_\sigma(av)_1.
\end{aligned}$$

where the inequality comes from Theorem 4 with equality iff $x_1 = \ldots x_{d-1} = 0$, *i.e.* $x = [0,\ldots,0,a] \in \mathcal{V}$. Now for the second part of the proof, let $x_1 = a_1 v$, $x_2 = a_2 v$ such that $a_1 > a_2$. Then

$$\begin{aligned}
f_\sigma(x_1)_1 = (2\pi\sigma^2)^{-\frac{d}{2}} \int_{-\infty}^{-a_1} \int_{\sum_{k=1}^{d-1} {x'_k}^2 \leq tan^2(\theta)(x'_d + a_1)^2} & e^{-\frac{\sum_{k=1}^{d-1} {x'_k}^2}{2\sigma^2}} dx'_1 \ldots dx'_{d-1} e^{-\frac{{x'_d}^2}{2\sigma^2}} dx'_d \\
\text{As } a_1 + x'_d \leq 0, (a_1 + x'_d)^2 &< (a_2 + x'_d)^2
\end{aligned}$$

$$
\begin{aligned}
&< (2\pi\sigma^2)^{-\frac{d}{2}} \int_{-\infty}^{-a_1} \int_{\sum_{k=1}^{d-1} {x'_k}^2 \leq tan^2(\theta)(x'_d+a_2)^2} e^{-\frac{\sum_{k=1}^{d-1} {x'_k}^2}{2\sigma^2}} dx'_1 \dots dx'_{d-1} e^{-\frac{x'^2_d}{2\sigma^2}} dx'_d \\
&< (2\pi\sigma^2)^{-\frac{d}{2}} \int_{-\infty}^{-a_2} \int_{\sum_{k=1}^{d-1} {x'_k}^2 \leq tan^2(\theta)(x'_d+a_2)^2} e^{-\frac{\sum_{k=1}^{d-1} {x'_k}^2}{2\sigma^2}} dx'_1 \dots dx'_{d-1} e^{-\frac{x'^2_d}{2\sigma^2}} dx'_d \\
&= f_\sigma(x_2)_1.
\end{aligned}
$$

□

Lemma 6. $\forall a > 0, k \geq 1,\ \frac{\Phi(-a)}{\Phi(-ka)} \geq e^{\frac{(k^2-1)a^2}{2}}$.

Proof. Consider the function $h(x) = \frac{\sqrt{2\pi}\Phi(-x)}{e^{-x^2/2}}$ and we will show in the following that it is strictly decreasing for $x > 0$. Alternatively, we take the derivative *w.r.t.* x,

$$\frac{d}{dx}h(x) = \frac{\sqrt{2\pi}x\Phi(-x)}{e^{-x^2/2}} - 1,$$

and show that it is negative for $x > 0$. Since $e^{-x^2/2} > 0$, it is sufficient to show that $\sqrt{2\pi}x\Phi(-x) - e^{-x^2/2} < 0$. Combining that 1) $\sqrt{2\pi}x\Phi(-x) - e^{-x^2/2}$ is increasing as

$$
\begin{aligned}
\frac{d}{dx}\left(x\Phi(-x) - \frac{e^{-x^2/2}}{\sqrt{2\pi}}\right) &= \Phi(-x) - \frac{xe^{-x^2/2}}{\sqrt{2\pi}} - \frac{-xe^{-x^2/2}}{\sqrt{2\pi}} \\
&= \Phi(-x) > 0
\end{aligned}
$$

and 2) $\sqrt{2\pi}x\Phi(-x) - e^{-x^2/2} \to 0$ when $x \to \infty$, we have that $\sqrt{2\pi}x\Phi(-x) - e^{-x^2/2} < 0$. As $h(x)$ is strictly decreasing we have that for any $a > 0$ and $k > 1$, $ka > a$. Thus,

$$\frac{\sqrt{2\pi}\Phi(-a)}{e^{-a^2/2}} > \frac{\sqrt{2\pi}\Phi(-ka)}{e^{-(ka)^2/2}}.$$

Rearranging the terms gives the inequality. □

Theorem 5. *A semi-bounded decision region that has a narrow bounding cone shrinks along v-direction after applying Gaussian smoothing filters with high σ, i.e. if the region admits a bounding cone $\mathcal{C}^{\mathcal{D}}_{\theta,v}$ with* $\tan(\theta) < \sqrt{\frac{(d-1)}{2c\log(c-1)}}$*, then for* $\sigma > (\Upsilon^v_{\mathcal{C}^{\mathcal{D}}_{\theta,v}} - \Upsilon^v_{\mathcal{D}})\tan(\theta)\sqrt{\frac{c}{d-1}} \cdot \frac{2(d-1)}{(d-1)-2\tan^2(\theta)c\log(c-1)}$, $\Upsilon^v_{\mathcal{D}_\sigma} < \Upsilon^v_{\mathcal{D}}$.

Proof. In this derivation we assume without loss of generality, $v = [0, \dots, 0, 1]^T \in \mathbb{R}^d$ (It is always possible to orient the axis to make this happen). From Corollary 2, we can see that $\mathcal{D}_\sigma \subseteq (\mathcal{C}^{\mathcal{D}}_{\theta,v})_\sigma$ which gives us $\Upsilon^v_{\mathcal{D}_\sigma} = \max_{x\in\mathcal{D}_\sigma} v^T x \leq \max_{x\in(\mathcal{C}^{\mathcal{D}}_{\theta,v})_\sigma} v^T x = \Upsilon^v_{(\mathcal{C}^{\mathcal{D}}_{\theta,v})_\sigma}$. Then to show that $\Upsilon^v_{\mathcal{D}_\sigma} < \Upsilon^v_{\mathcal{D}}$ it is sufficient to show that $\Upsilon^v_{(\mathcal{C}^{\mathcal{D}}_{\theta,v})_\sigma} < \Upsilon^v_{\mathcal{D}}$.
We observe that we only need to check the point x on the axis of the cone at distance $\Upsilon^v_{\mathcal{C}^{\mathcal{D}}_{\theta,v}} - \Upsilon^v_{\mathcal{D}}$ from the tip x_0 of the cone, *i.e.*, $x = x_0 - (\Upsilon^v_{\mathcal{C}^{\mathcal{D}}_{\theta,v}} - \Upsilon^v_{\mathcal{D}})v$. If x is not classified as Class 1 then by Lemma 5, we have that

$$
\begin{aligned}
\Upsilon^v_{(\mathcal{C}^{\mathcal{D}}_{\theta,v})_\sigma} < v^T x &= v^T(x_0 - (\Upsilon^v_{\mathcal{C}^{\mathcal{D}}_{\theta,v}} - \Upsilon^v_{\mathcal{D}})v) \\
&= v^T x_0 - (\Upsilon^v_{\mathcal{C}^{\mathcal{D}}_{\theta,v}} - \Upsilon^v_{\mathcal{D}})v^T v \\
&= \Upsilon^v_{\mathcal{C}^{\mathcal{D}}_{\theta,v}} - (\Upsilon^v_{\mathcal{C}^{\mathcal{D}}_{\theta,v}} - \Upsilon^v_{\mathcal{D}}) = \Upsilon^v_{\mathcal{D}}
\end{aligned}
$$

From the above argument and the definition of the decision boundary we see that if $f_\sigma(x)_1 < \frac{1}{c}$, then $\Upsilon^v_{\mathcal{D}_\sigma} < \Upsilon^v_{\mathcal{D}}$. Without loss of generality we let x_0 be the origin. By definition 2, we have

$$f_\sigma(x)_1 = \int_{x'\in\mathbb{R}^d} f(x')_1 p(x') dx'$$

$$= \int_{x'_d + \|x'\| cos(\theta) \leq 0} (2\pi\sigma^2)^{-\frac{d}{2}} e^{-\frac{(x'-x)^T(x'-x)}{2\sigma^2}} dx'$$

$$= (2\pi\sigma^2)^{-\frac{d}{2}} \int_{-\infty}^{0} \int_{\Sigma_{k=1}^{d-1} {x'_k}^2 \leq tan^2(\theta) {x'_d}^2} e^{-\frac{\Sigma_{k=1}^{d-1}(x'_k - x_k)^2}{2\sigma^2}} dx'_1 \dots dx'_{d-1} e^{-\frac{(x'_d - x_d)^2}{2\sigma^2}} dx'_d$$

$$= (2\pi\sigma^2)^{-\frac{d}{2}} \int_{-\infty}^{0} \int_{\Sigma_{k=1}^{d-1} {x'_k}^2 \leq tan^2(\theta) {x'_d}^2} e^{-\frac{\Sigma_{k=1}^{d-1} {x'_k}^2}{2\sigma^2}} dx'_1 \dots dx'_{d-1} e^{-\frac{(x'_d - x_d)^2}{2\sigma^2}} dx'_d$$

$$= (2\pi\sigma^2)^{-\frac{1}{2}} \int_{-\infty}^{0} q(|x'_d \tan(\theta)|, d-1, \sigma) e^{-\frac{(x'_d - x_d)^2}{2\sigma^2}} dx'_d$$

$$= (2\pi\sigma^2)^{-\frac{1}{2}} \int_{-\infty}^{0} Q(\frac{d-1}{2}, \frac{tan^2(\theta) {x'_d}^2}{2\sigma^2}) e^{-\frac{(x'_d - x_d)^2}{2\sigma^2}} dx'_d$$

Substitute $X_d = \frac{x_d}{\sigma}, X'_d = \frac{x'_d}{\sigma}$

$$= (2\pi)^{-\frac{1}{2}} \int_{-\infty}^{0} Q(\frac{d-1}{2}, \frac{tan^2(\theta) {X'_d}^2}{2}) e^{-\frac{(X'_d - X_d)^2}{2}} dX'_d$$

Let $M \leq \sqrt{\frac{d-1}{c \tan^2(\theta)}}, k = \frac{M}{X_d}$

$$= (2\pi)^{-\frac{1}{2}} \int_{M}^{0} Q(\frac{d-1}{2}, \frac{tan^2(\theta) {X'_d}^2}{2}) e^{-\frac{(X'_d - X_d)^2}{2}} dX'_d$$

$$+ (2\pi)^{-\frac{1}{2}} \int_{-\infty}^{M} Q(\frac{d-1}{2}, \frac{tan^2(\theta) {X'_d}^2}{2}) e^{-\frac{(X'_d - X_d)^2}{2}} dX'_d$$

$$< (\Phi(-X_d) - \Phi(M - X_d)) Q(\frac{d-1}{2}, \frac{\tan^2(\theta) M^2}{2}) + \Phi(M - X_d)$$

$$< \frac{\Phi(-X_d) - \Phi(M - X_d)}{c} + \Phi(M - X_d)$$

$$= \frac{1}{c} + \frac{(c-1)\Phi((k-1)X_d) - \Phi(X_d)}{c}.$$

Then we see that using Lemma 6, we see that we see that if $e^{\frac{X_d^2((k-1)^2-1)}{2}} \geq c-1$ then $(c-1)\Phi((k-1)X_d) \leq \Phi(X_d)$. So, we need to satisfy the inequalities for some k:

$$\sqrt{\frac{2\log(c-1)}{(k-1)^2 - 1}} \leq -X_d \leq \sqrt{\frac{d-1}{k^2 c \tan^2(\theta)}}.$$

This is only possible if for some k, we have $\sqrt{\frac{2\log(c-1)}{(k-1)^2-1}} \leq \sqrt{\frac{d-1}{k^2 c \tan^2(\theta)}}$ or $\tan(\theta) \leq \sqrt{\frac{d-1}{2c\log(c-1)}} \cdot \sqrt{1 - \frac{2}{k}}$. So, we need that

$$\tan(\theta) < \sqrt{\frac{d-1}{2c\log(c-1)}}.$$

Then we see that giving the cone is narrow enough, we have the required shrinking if we have X_d satisfies the inequalities for some k. So, we see that if we have $-X_d = \sqrt{\frac{d-1}{k^2 c \tan^2(\theta)}}$ for some k such that $\tan(\theta) \leq \sqrt{\frac{d-1}{2c\log(c-1)}} \cdot \sqrt{1 - \frac{2}{k}}$ is satisfied. So, we need that $\frac{-x_d}{\sigma} = \sqrt{\frac{d-1}{k^2 c \tan^2(\theta)}}$ for some suitable k. Thus we need $\sigma = -x_d \tan(\theta) \sqrt{\frac{c}{d-1}} k$ for some suitable k. Including the constraint on k and substituting the value for x_d, we get that shrinking always happens for

$$\sigma \geq (\Upsilon^v_{\mathcal{C}^{\mathcal{D}}_{\theta,v}} - \Upsilon^v_{\mathcal{D}}) \tan(\theta) \sqrt{\frac{c}{d-1}} \cdot \frac{2(d-1)}{(d-1) - 2\tan^2(\theta) c \log(c-1)}.$$

□

Theorem 6. *The shrinkage of class 1 decision region is proportional to the smoothing factor, i.e.* $\Upsilon^v_{\mathcal{D}} - \Upsilon^v_{\mathcal{D}_\sigma} \propto \sigma$.

Proof. In this case we assume a cone-like decision region which can be represented as $\mathcal{D} = \{x \in \mathbb{R}^d \mid v^T x + \|v\|\|x\|cos(\theta) \leq 0\}$ with $v = [0, \ldots, 0, 1]^T$ without loss of generality. By Lemma 5, we see that in order to get bounds on $\Upsilon^v_{\mathcal{D}_\sigma}$, we only need to analyze the value of $f_\sigma(x)_1$ for points x along the axis of the cone. Then we see that for a general point $x = av$ along the axis of the cone, using the same ideas as in proof of Theorem 5, we have

$$\begin{aligned}
f_\sigma(x)_1 &= \int_{x' \in \mathbb{R}^d} f(x')_1 p(x') dx' \\
&= (2\pi\sigma^2)^{-\frac{d}{2}} \int_{-\infty}^{0} \int_{\sum_{k=1}^{d-1} {x'_k}^2 \leq tan^2(\theta) {x'_d}^2} e^{-\frac{\sum_{k=1}^{d-1} {x'_k}^2}{2\sigma^2}} dx'_1 \ldots dx'_{d-1} e^{-\frac{(x'_d - a)^2}{2\sigma^2}} dx'_d \\
&= (2\pi\sigma^2)^{-\frac{1}{2}} \int_{-\infty}^{0} Q(\frac{d-1}{2}, \frac{tan^2(\theta) {x'_d}^2}{2\sigma^2}) e^{-\frac{(x'_d - a)^2}{2\sigma^2}} dx'_d \\
&\text{Substitute } A = \frac{a}{\sigma}, x'_d = \frac{x'_d}{\sigma} \\
&= (2\pi)^{-\frac{1}{2}} \int_{-\infty}^{0} Q(\frac{d-1}{2}, \frac{tan^2(\theta) {x'_d}^2}{2}) e^{-\frac{(x'_d - A)^2}{2}} dx'_d \\
&= f_1(Av)_1 = f_1(\frac{1}{\sigma} x)_1.
\end{aligned}$$

Using the equation above we see that for smoothing by a general σ,

$$\begin{aligned}
\Upsilon_{\mathcal{D}_\sigma} &= \sup_{x | f_\sigma(x) \geq \frac{1}{c}} v^T x = \sup_{x | f_1(\frac{1}{\sigma} x) \geq \frac{1}{c}} v^T x = \sup_{x' | f_1(x') \geq \frac{1}{c}} v^T (\sigma x') \\
&= \sigma \sup_{x' | f_1(x') \geq \frac{1}{c}} v^T x' = \sigma \Upsilon_{\mathcal{D}_1}.
\end{aligned}$$

In this case we have $\Upsilon_{\mathcal{D}} = 0$ by construction, so $\Upsilon_{\mathcal{D}} - \Upsilon_{\mathcal{D}_\sigma} = 0 - \sigma \Upsilon_{\mathcal{D}_1} = \sigma \cdot (-\Upsilon_{\mathcal{D}_1}) \propto \sigma$. □

With the above Theorem 6, we can fix the smoothing factor to $\sigma = 1$ and further obtain a lower bound of the shrinking rate *w.r.t* c, θ, and d:

Theorem 7. *The shrinking rate of class 1 decision region is at least* $\sqrt{\frac{d-1}{c \tan^2(\theta)}} \cdot \frac{(d-1) - 2\tan^2(\theta) c \log(c-1)}{2(d-1)}$, *i.e.* $\frac{\Upsilon^v_{\mathcal{D}_\sigma} - \Upsilon^v_{\mathcal{D}_{\sigma+\delta}}}{\delta} > \sqrt{\frac{d-1}{c \tan^2(\theta)}} \cdot \frac{(d-1) - 2\tan^2(\theta) c \log(c-1)}{2(d-1)}$.

Proof. As in Theorem 6, we assume a cone at origin along $v = [0, \ldots, 0, 1]^T$ given by $\mathcal{D} = \{x \in \mathbb{R}^d \mid v^T x + \|v\|\|x\|cos(\theta) \leq 0\}$. Following the same proof idea as Theorem 6, we see that the rate is given by the value $-\Upsilon_{\mathcal{D}_1}$. So, we try to get a bound on the value of $-\Upsilon_{\mathcal{D}_1}$. To establish a lower bound we show that for the point $x = av$, $f_1(x)_1 < \frac{1}{c}$. Then by Lemma 5 we have $\Upsilon_{\mathcal{D}_1} < a$ or $-\Upsilon_{\mathcal{D}_1} > -a$.
Using the same procedure as in the proof of Theorem 5, we get that if x satisfies the two inequalities

$$\sqrt{\frac{2\log(c-1)}{(k-1)^2 - 1}} \leq -v^T x \leq \sqrt{\frac{d-1}{k^2 c \tan^2(\theta)}}$$

for suitable real k, then we have $f_1(x)_1 < \frac{1}{c}$. So, we need $v^T x = -\sqrt{\frac{d-1}{k^2 c \tan^2(\theta)}}$ for some k such that $\sqrt{\frac{2\log(c-1)}{(k-1)^2 - 1}} \leq x \leq \sqrt{\frac{d-1}{k^2 c \tan^2(\theta)}}$. The constraint on k can be re-written as $k \geq \frac{2(d-1)}{(d-1) - 2\tan^2(\theta) c \log(c-1)}$. Taking k to be lower bound, we get that for

$$-a = -v^T x = \sqrt{\frac{d-1}{c \tan^2(\theta)}} \cdot \frac{(d-1) - 2\tan^2(\theta) c \log(c-1)}{2(d-1)}$$

$f_1(x)_1 \leq \frac{1}{c}$. So, we get that the rate is $-\Upsilon_{\mathcal{D}_1} \geq -a \geq \sqrt{\frac{d-1}{c \tan^2(\theta)}} \cdot \frac{(d-1) - 2\tan^2(\theta) c \log(c-1)}{2(d-1)}$. □

D Additional Analysis

D.1 Shrinking effect for unidimensional data

Bounded decision region. Without loss of generality, let the decision region be interval $\mathcal{D} = [-R, R]$. By the symmetric nature of Gaussian smoothing, we see that $\mathcal{D}_\sigma$ is also an interval of the form $[-a, a]$. We claim that for large σ, $a < R$ and for even larger σ, $\mathcal{D}_\sigma$ disappears. Formally, we do the analysis as follows.

For the shrinking, we check the value of $f_\sigma(R)_1$. By definition 2, we see that $f_\sigma(R)_1 = \Phi(\frac{2R}{\sigma}) - \Phi(0)$ and if

$$\sigma > \frac{2R}{\Phi^{-1}(\frac{1}{2} + \frac{1}{c})},$$

$f_\sigma(R) < \frac{1}{c}$ is true. Thus, the bounded decision region of unidimensional data shrinks with smoothing factor $\sigma > \frac{2R}{\Phi^{-1}(\frac{1}{c}+\frac{1}{2})}$.

For the vanishing rate, we check the value of $f_\sigma(x)_1$ at $x = 0$. Now since $f_\sigma(0)_1 = \Phi(\frac{R}{\sigma}) - \Phi(-\frac{R}{\sigma})$, we have that if

$$\sigma > \frac{R}{\Phi^{-1}(\frac{1}{2} + \frac{1}{2c})},$$

$f_\sigma(0)_1 < \frac{1}{c}$ is true, *i.e.*, $\mathcal{D}_\sigma$ vanishes.

Semi-bounded decision region. In a unidimensional case, our definition of semi-bounded regions degenerates into an interval I of the form $[a, \infty)$. In this case, Theorem 7 gives a trivial bound of 0 for the shrinkage of the decision region, suggesting that no shrinking happens. However, we emphasize that in practice, shrinking might still happen and more detailed analysis is left for future work.

D.2 Bounded decision region behaviors

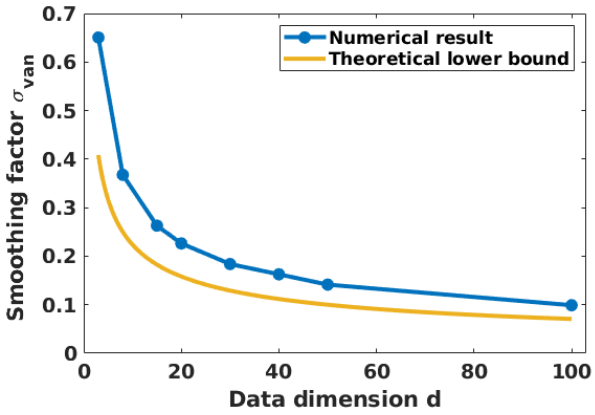

Figure S1: The vanishing smoothing factor σ_{van} with an increasing input-space dimension in the exemplary adversarial ball.

The vanishing smoothing factors σ_{van} with different data dimensions implied by Figure 2 of the main text together with the theoretical lower bound found in Theorem 4 is given as Figure S1.

Figure S2 shows the certified radius behavior as a function of the distance of points from the origin (y-axis) and the smoothing factor σ (x-axis) for dimension $d = 30$. The contour lines in Figure S2 mark the certified radius of points under Gaussian smoothing. It is notable that points closer to the origin generally have larger certified radii and the certified radius of the point at the origin (y-axis $y = 0$) drops to zero at vanishing smoothing factor $\sigma_{\text{van}} = 0.184$ as specified in Figure S1. Specifically, one can readily verify that the certified radii of points closer to the origin increase with the growing smoothing factor σ but begin to decrease at certain point, which is coherent with our observations through Figure 3(a) of the main text. Conducting similar experiments for different dimensions completes the maximum certified radius vs. data dimension relationship as shown in Figure S3.

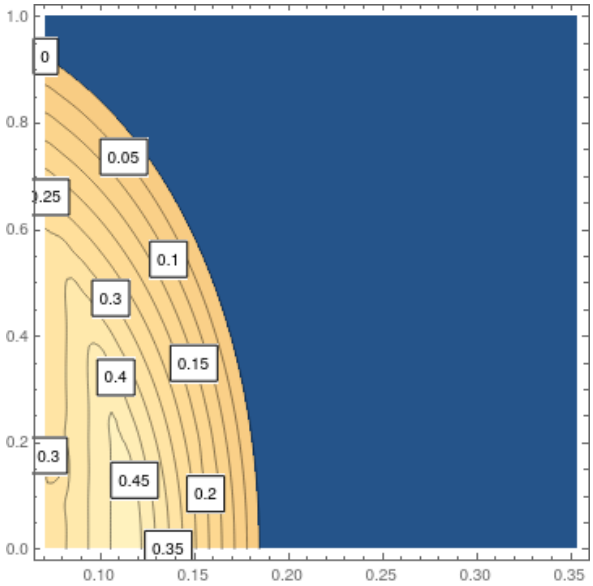

Figure S2: The certified radius of smoothed classifiers with an increasing input-space dimension when $d = 30$.

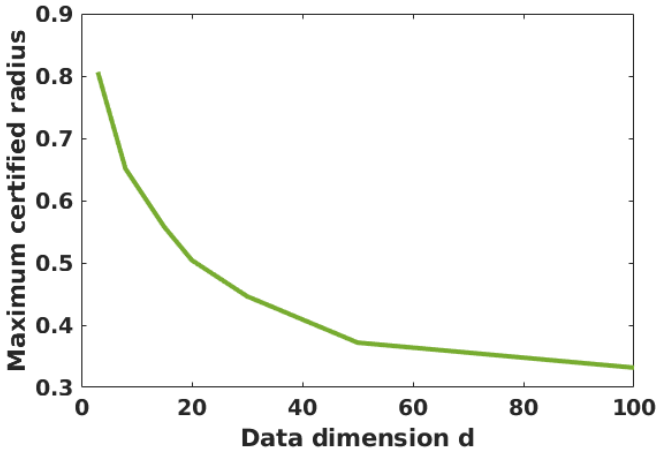

Figure S3: The maximum certified radius with an increasing input-space dimension in the exemplary case.

D.3 Semi-bounded decision region certified radius behaviors w.r.t data dimensions

In Figure S4, we show the unscaled certified radius r as a function of an increasing smoothing factor σ for different input data dimension d with fixed narrowness $\theta = 45°$. One can then see similar trend as told in Figure 3(a) of the main text in the bounded decision region case, the maximum certified radius (the peak) also decreases with the increasing dimension.

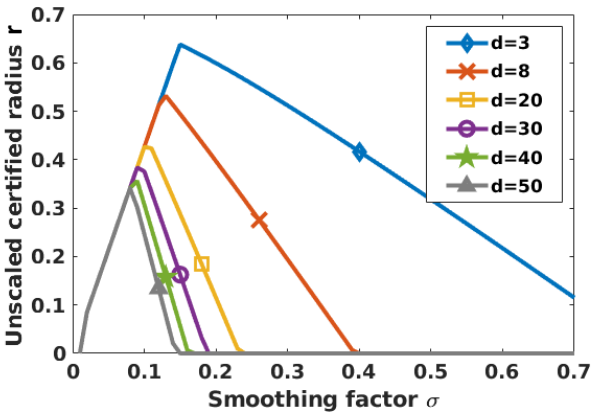

Figure S4: The unscaled certified radius r of a point on the axis v for different input data dimension d.