Problem 1

We have shown that probability mass of the spherical Gaussian with variance 1 in each coordinate has almost zero mass, since the volume of such a ball is negligible and the probability density is bounded above by $1/(2\pi)^{d/2}$. Now we show that for $\sigma = O(1/d)$ the probability mass is non-negligible.

Proof (a) We note that the probability density function is

$$p(x) = \frac{1}{(2\pi/\sqrt{d})^{d/2}} \exp\left(-\frac{|x|^2}{2/d}\right)$$

Then the probability mass function $P(x \in B)$ is

$$\int_B p(x)dx = \int_{S^d}\int_{r=0}^{1} p(x)r^{d-1}drd\Omega = \int_{S^d}\int_{r=0}^{1} \frac{1}{(2\pi/d)^{d/2}} \exp\left(-\frac{r^2}{2/d}\right) r^{d-1}drd\Omega$$

Since the variables Ω, r, and d do not intersect, we have

$$\frac{1}{(2\pi/d)^{d/2}} A(d) \int_0^1 \exp\left(-\frac{r^2}{2/d}\right) r^{d-1}dr$$

We have previously shown that

$$A(d) = \frac{\pi^{d/2}}{\frac{1}{2}\Gamma\left(\frac{d}{2}\right)}$$

and by substituting $x = \frac{r^2}{2/d}$ we have

$$\int_0^1 \exp\left(-\frac{r^2}{2/d}\right) r^{d-1}dr = 2^{d/2-1}d^{-d/2} \int_0^{d/2} \exp(-x)x^{d/2-1}dx$$

Combining the above, we have that the probability mass is

$$\int_0^{d/2} \frac{\exp(-x)x^{d/2-1}}{\Gamma(\frac{d}{2})}dx$$

i.e., the CDF of $\Gamma(d/2, 1)$ evaluated at $d/2$, its expectation. Note that this is greater than the CDF of $\Gamma(d/2, 1)$ evaluated at its median, since its median is strictly less than its mean [1]. The CDF of $\Gamma(d/2, 1)$ evaluated at its median is $1/2$. Thus, we are done.

Proof (b) Alternatively, we can easily show that the probability that $R = X_1^2 + X_2^2 + \cdots + X_n^2 < 1$ is bounded below for $\sigma = 1/\sqrt{2d}$. Equivalently, we can show that $R^2 > 1$. Observe that $E(X_i^2) = 1/2d$ and $Var(X_i^2) = 3/4d^2$. Then by the law of large numbers

$$P\left(\left|R^2 - d \cdot E(X_i^2)\right| \geq d\epsilon\right) = P\left(\left|R^2 - \frac{1}{2}\right| \geq d\epsilon\right) \leq \frac{Var(X_i^2)}{d\epsilon^2} = \frac{3}{4d^3\epsilon^2}$$

Choose $\epsilon = 1/2d$. Then we have that

$$P(R^2 > 1) \leq \frac{3}{4d}$$

Thus, $P(R^2 < 1) \geq 1 - \frac{3}{4d}$ so we are done.

Problem 2

We try to amend Proof (a) by considering the rate of change of the probability mass, and specifically showing that as $d \to \infty$, the rate of change is 0 beyond the unit ball, i.e., $R = 1$. We recall the prior definition of the probability mass M of the R-radius ball

$$M = \frac{2\int_0^R \exp\left(-\frac{r^2}{2/d}\right) r^{d-1}dr}{(2/d)^{d/2}\Gamma\left(\frac{d}{2}\right)}$$

Let $a = d/2$. The derivative is

$$\frac{dM}{dR} = \frac{2\exp\left(-aR^2\right)R^{2a-1}a^a}{\Gamma\left(a\right)}$$

For $R \geq 1$, the derivative is smallest at $R = 1$, and can be further bounded above using Stirling's approximation

$$\frac{dM}{dR} \leq \frac{2\left(a/e\right)^a}{a!} \leq \frac{2\left(a/e\right)^a}{\sqrt{2\pi a}(a/e)^a}$$

This converges to 0 as $a \to \infty$ so we are done.

Problem 3

When training a linear classifier to data x of large dimension d, we can use a projection P of shape $d \times d'$, according to the JL-Lemma to reduce the dimensionality of the data to $d' << d$ while preserving relative distances. This leads us to fit a classifier W of shape $k \times d'$, giving us scores of the form

$$WP^Tx$$

We could also train a linear classifier U on the original data, which would have shape $k \times d$ and give us

$$U^Tx$$

In the first case, we are essentially fitting a linear classifier W^TP to the data which has the same shape as U but significantly lower rank. We note that this increases the *approximation error*, the best error achievable by the classifier, but may decrease the *estimation error*, indicating more generalizability. We also note that P will lead to distortion of the data, which may cause *distortion* of the data, which manifests itself in the approximation error.

The suggestion was to look into how to balance the introduction of *distortion* with the *generalizability* of the linear classifier. To do so, I will try to get a better understanding of the bias-complexity tradeoff and experiment with some high-dimensional datasets. For the latter, I want to get an idea of the point at which the error begins to increase as d' decreases. I could use cardinality bounds to induce a finite hypothesis class.

Primarily, I want to formalize this tradeoff. I assume this means: can I give some relationship between d', d and the (change in) the error of the model. One starting point would be that the JL-Lemma gives a probabilistic bound on the distortion of the dataset for d' larger than a function of d. I would want to look into error bounds for linear classifiers.

Problem 3 Idea Sketch

We consider the case of binary classification. We know that the linear classifier is PAC-learnable, and thus with a sufficient number of examples m, we can show that with high probability $1 - \delta$, the error is less than ϵ.

We know that the JL-Lemma approximately preserves pairwise distances. Then, it also approximately preserves angular distances. Notably, it preserves angular distances from all the points x to w. Thus, it approximately preserves the classifications.

Consider any w obtained from training on X. To ensure that the distortions do not lead to misclassifications, I think it would be sufficient to scale w with respect to the smallest value of $|wx|$ that is nonzero. For those that are zero, we were unable to predict before, hence the distortions I think would be tolerable.

One thing you could show in this way is that with the same sample size m, you could have a similar bound on the error with the projection with high probability. You could decrease the m required because of the reduced dimensions, and you could have a similar upper bound on the error with low fewer samples.

References

[1] Jeesen Chen and Herman Rubin. Bounds for the difference between median and mean of gamma and poisson distributions. *Statistics & probability letters*, 4(6):281–283, 1986. `http://aiweb.techfak.uni-bielefeld.de/content/bworld-robot-control-software/`.