EPFL

Support Vector Classification : sinusoidal interface

Antony Doukhan

June 12, 2020

We study the performance of Kernel method algorithms when the decision boundary is of sinusoidal shape. We first review the results of flat boundary and the dependence on the dimension in the absence of curse of dimensionality. Then we investigate on binary data classification (with intrinsic dimension $d_{\|} = 1$) the different regimes for the error function ϵ with respect to training sample size p, and how the parameters of the sine (amplitude and wavelength) influence it.

1 Support Vector Classification and related works

In supervised machine learning, a function mapping the input to an output is learnt from a finite collection of p training examples and a test error ϵ is computed from a set of test inputs/outputs. Usually, the error scales with a powerlaw decay $\epsilon \sim p^{-\beta}$ with β an exposant depending on the parameters of the model.
In the *Support Vector Classification* framework, the algorithm maximizes the margin between a decision boundary function and the points in the training set $\{\underline{x}^{\mu}\}_{\mu=1}^{p} \subset \mathbb{R}^d$. If the decision boundary can not separate linearly the data, a feature map $\underline{x}^{\mu} \mapsto \underline{\phi}(\underline{x}^{\mu})$ enables the data to live in a higher dimensional Hilbert space. The algorithm then exploits the scalar product of features which can be written as a positive-definite kernel $K(\underline{x}, \underline{x}') = \phi(\underline{x}) \cdot \phi(\underline{x}')$. Introducing the scale σ over which the kernel varies, $K(\underline{x}, \underline{x}')$ can be written as a radial function $K\left(\frac{||\underline{x}-\underline{x}'||}{\sigma}\right)$. The decision function can then be expressed according to this kernel and the label predicted $\hat{y}(\underline{x})$ is given by the sign of it :

$$f(\underline{x}) = \sum_{\mu=1}^{p} \alpha^{\mu} y^{\mu} K\left(\frac{||\underline{x} - \underline{x}^{\mu}||}{\sigma}\right) + b, \quad \hat{y}(\underline{x}) = \text{sign} f(\underline{x}) \tag{1.1}$$

The variables $\alpha^{\mu} \geq 0$ and bias b are fixed by solving the problem :

$$\max_{\{\alpha^\mu\}} \sum_{\mu=1}^{p} \alpha^\mu - \frac{1}{2} \sum_{\mu,\nu=1}^{p} \alpha^\mu \alpha^\nu y^\mu y^\nu K\left(\frac{||\underline{x} - \underline{x}^\mu||}{\sigma}\right) \tag{1.2}$$

subject to the constraints

$$\begin{aligned} \alpha^\mu > 0 \Longleftrightarrow y^\mu f(\underline{x}^\mu) = 1 \\ \sum_{\mu,\nu=1}^{p} \alpha^\mu y^\mu = 0 \\ \min_{\mu} |f(\underline{x}^\mu)| = 1 \end{aligned} \tag{1.3}$$

Vectors with $\alpha^\mu > 0$ are called support vectors and enter in the expansion of the decision function. Based on the work of J. Paccolat and S. Spigler [1], we can identify two regimes for the error function $\epsilon \sim p^{-\beta}$ that depend on the scale of the kernel σ and the nearest neighbor distance $\delta \sim p^{-1/d}$. In the case $\sigma \ll \delta$ the classification error is subject to the curse of dimensionality and $\beta = \mathcal{O}(1/d)$. For $\sigma \gg \delta$ however, we have $\beta = \mathcal{O}(1)$. In particular, if we consider a flat boundary decision interface, we can define a band Ω_Δ of thickness Δ around the interface where the support vectors are gathered. The main result from [1] that we use for our study are the following relations :

$$\boxed{\epsilon \sim \Delta \sim p^{-\beta}, \quad \text{with } \beta = \frac{d - 1 + \xi}{3d - 3 + \xi}} \tag{1.4}$$

where ξ corresponds to the exponent used to approximate (using Taylor expansion) the kernel.
Hence understanding where the support vectors are located around the interface allows us to predict the test error function.

2 Non linear interface : sinusoidal decision boundary

2.1 Framework

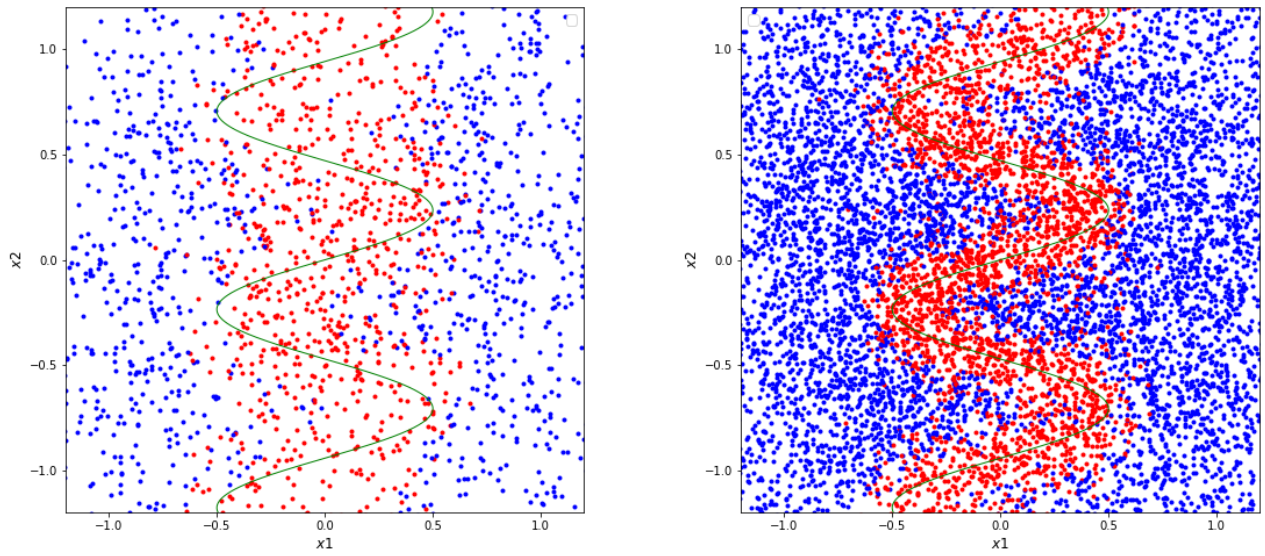
Our framework is defined by a set of training examples $\{\underline{x}^\mu\}_{\mu=1}^{p} \subset \mathbb{R}^d$. Each coordinate is sampled from a normal distribution of mean 0 and variance $\gamma^2 = 1$. Accordingly, the same distribution is used for the test set. To avoid the curse of dimensionality, we use a Laplace kernel ($\xi = 1$) with a scale parameter $\sigma = 100 \gg \delta \sim p^{-1/d}$. A 1-dimensional binary classification is made according to the x_1 coordinate. The decision function and the label read :

$$f(\underline{x}^\mu) = x_1^\mu - A \cdot \sin(x_2^\mu / r), \quad y^\mu = \text{sign} f(\underline{x}^\mu) \tag{2.1}$$

The interface is thus parametrized by the amplitude denoted A and the wavenumber r. In this section we are stating the dependence $\epsilon = F(A, r, p, d)$ and studying the limit cases. The test error is proportional to the probability of having the coordinate x_1 lying in the band of SVs. Ω_Δ :

$$\epsilon = \frac{1}{2}\mathbb{P}\{x_1 \in \Omega_\Delta\} = \frac{1}{2}\frac{1}{2\pi} \int_{\mathbb{R}} dx_2 e^{-x_2^2/2} \int_{A\sin(x_2/r)-\Delta}^{A\sin(x_2/r)+\Delta} dx_1 e^{-x_1^2/2} \tag{2.2}$$

We first highlight the existence of a critical training sample size value $p*$ that depends on the parameters d, A and r below which the algorithm is not able to properly discern the decision function. The support vectors are then gathered in a band of noise of width $2A$. Increasing the number of training examples eventually enables the learning procedure as shown on the Fig.1



(a) $p = 3000$ (b) $p = 14000$

Figure 1: Distribution of Support Vectors around the interface for $A = 0.5$, $r = 0.15$ and $d = 4$. The box size is slightly larger than the standard deviation of the distribution of points. In red are ploted the support vectors. For a small number of training examples, the algorithm is unable to expand the decision function correctly resulting in a band of noise. Increasing the number of examples eventually makes the learning possible.

2.1.1 Case $p \ll p*$: noise band

In the case $p \ll p*$ the error function can be rewritten as $1/2$ times the probability to be in the noise band and the error test is only dependent on the amplitude of the sinusoid :

$$\boxed{\epsilon = P(A) = \frac{1}{2\sqrt{2\pi}} \int_{-A}^{A} dx_1 e^{-x_1^2/2}} \tag{2.3}$$

Thus for a sufficiently small r, there exist a set of p sufficiently small for which the error is constant and independent of the number of training examples.

2.1.2 Case $p \gg p*$: recovery of the β exponent

For sufficiently large p, Spigler et al.[1] provide an expression for the bandwidth Δ. We suppose this value relatively small to approximate the error function in Eq.2.2 using the midpoint approximation :

$$\epsilon \approx \frac{2\Delta}{4\pi} \int_{\mathbb{R}} dx e^{-x^2/2} \cdot e^{-\frac{1}{2}A^2 \sin^2(x/r)} \propto \Delta \mathbb{E}\left[e^{-\frac{1}{2}A^2 \sin^2(X/r)}\right] \tag{2.4}$$

where the x_2 coordinate has been replaced by x to lighten the notation. A priori, we can not give an explicit behavior for Δ unless the boundary resembles a flat one, in which case $\Delta \sim p^{-\beta}$. Nevertheless, we can provide some analytical solutions for the test error in the limit values of A and r.

Case $r \gg \gamma$: Taylor expansion For large value of r, that we consider arbitrarily being greater than the standard deviation of the training examples distribution, we retrieve intuitively

the results of flat boundary, since the the training set is approximately crossed by 1 period of the sinus and this result is emphasized if A is small. We can solve the above integral expanding the expectation around $\mu_X = 0$ (see Appendix A) :

$$\boxed{\epsilon \sim \Delta \left(1 - \frac{A^2}{2r^2}\right) \sim p^{-\beta} \left(1 - \frac{A^2}{2r^2}\right)} \tag{2.5}$$

Case $A \gg \gamma$: Saddle-point approximation In the case A is arbitrarily greater than the standard deviation and (recall) $p \gg p*$, the maximum amplitude of the sine is outside the training set and thus there is no influence of the curvature on the error function, thus $\Delta \sim p^{-\beta}$. The error function in Eq.2.2 can be approximated using saddle point method :

$$\epsilon \approx \frac{r\Delta}{A\sqrt{2\pi}} \vartheta_3(0, e^{-(\pi r)^2/2}), \quad \vartheta_3(0, q) = \sum_{k=-\infty}^{\infty} q^{k^2} \tag{2.6}$$

The derivation of this result is presented in Appendix B. The complex behaviour obtained will not be tested in the experimental part of this paper.

Dependence on the curvature In the intermediate case where A, r are of the order of magnitude of γ, we can not find a closed form solution to the error function. Intuitively, as p grows to infinity, the error dependence on the training size is asymptotically the one of the flat boundary case that is $\epsilon \sim \Delta \sim p^{-\beta}$ for a fixed A, r. However, the main result is that the bandwidth Δ has to be rescaled by a factor proportional to the maximal curvature of the sinusoid :

$$\boxed{\epsilon \sim \kappa(A, r) \cdot p^{-\beta} \cdot I(r, A) = \frac{A}{r^2} p^{-\beta} \cdot I(r, A)} \tag{2.7}$$

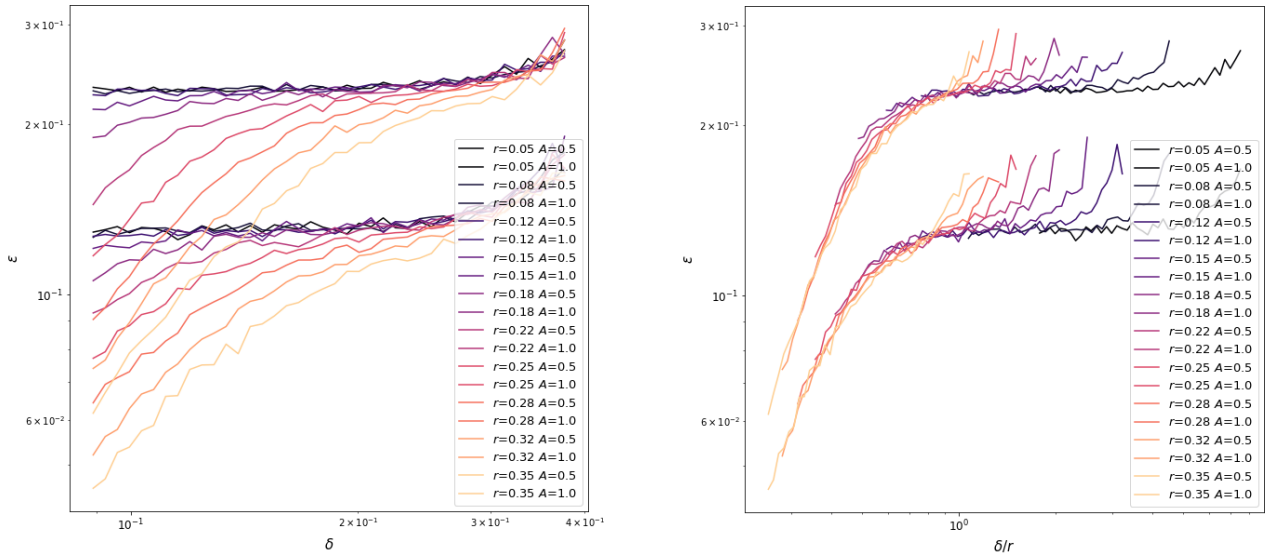
where $I(r, A)$ denotes the integral in Eq.2.4. Indeed, maximizing the margin of the decision function of sinusoidal shape at the peak of the wave, where the curvature is at its highest, is much more difficult than on a flat boundary, inducing an error.

3 Results

The experimentation is conducted using the `sklearn` package from python. Most of the measures were made in $d = 4$ vector space where the nearest neightbour distance scales as $\delta \sim p^{-1/4}$ and the exponent β is equal to $4/10$ according to Eq.1.4. Each slope represents the average value of the error over 20 independent experiments with a given amplitude and wavenumber parameters. The number of training examples varies between $p = 50$ and $p = 15438$. We often plot the error with respect to the nearest neighbour distance or the asymptotic Δ to emphasize on the predicted results.
Note that for the extreme cases $A \gg 1$, $r \gg 1$, we were not able to verify the theoretical results, as the decision function is too close to a flat boundary and the number of training data available was not enough to detect corrections.

3.1 Critical value of sample size $p*$

In this section we show numerically the existence of a critical value $p*$ under which the error function is constant, proportional to the amplitude of the sine. To identify the relation $p* = F(A, r, p, \delta)$ we rescale the x axis with respect to a function of d, A, r to identify a common value from which the behavior of each curve is the same.



(a) ϵ as function of $\delta = p^{-1/d}$

(b) ϵ as function of δ/r

Figure 2: Test error as function of nearest neighbor distance δ (supposedly taken as $p^{-1/d}$ the for A, r parameters of the magnitude of the standard deviation. The highest error values correspond to higher amplitude of the wave. By rescaling the x-axis by a the wavenumber, the curves collapse at a same point for a fixed amplitude.

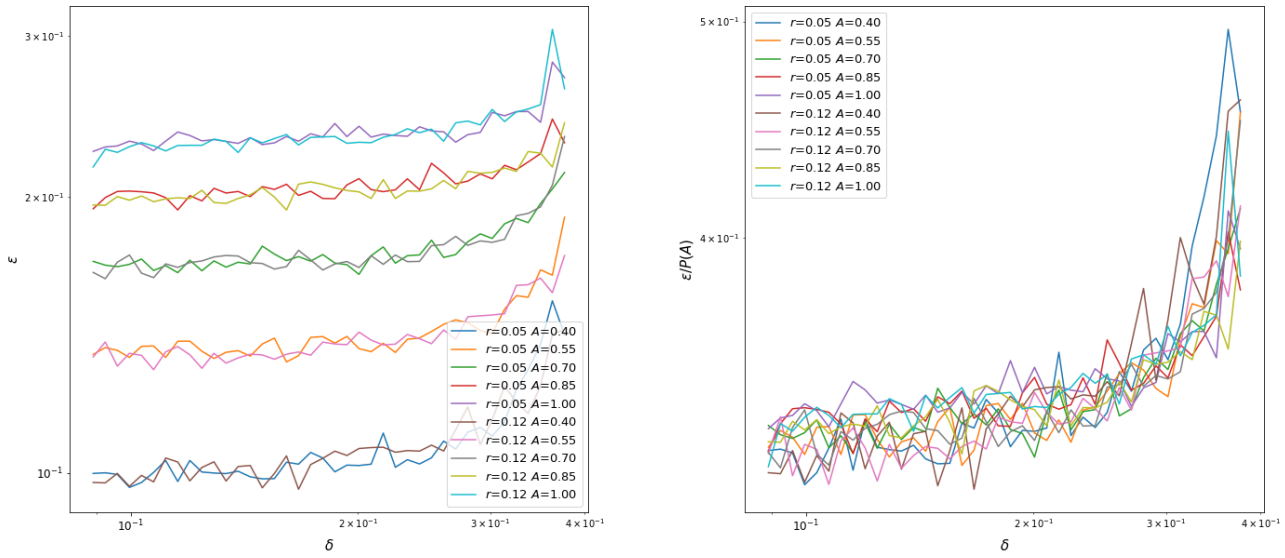
This first observation allows us to deduce experimentally a scaling dependence for the critical value $p*$ with respect to parameter r.

$$p* \sim g(A)r^{-d}$$

with $g(A)$ representing the A dependence left to be determine. Numerically we enlarge the set of amplitudes explored and for each amplitude we mark the point δ/r at which the curves collapse. Mainly due to noise , we were not able to retrieve the dependence in A. We shall however stress on the fact that the critical value $p*$ is highly more dependent on r rather then A. The main idea behind these results lies in the fact that if the nearest neighbour distance is much higher than the wavenumber of the sine, the algorithm is not able to draw the decision function properly.

3.2 Noise band case $p \ll p*$ and amplitude dependence

In this section we focus on the p values for which the algorithm is not able to discern the decision function. We calculate and probability of sampling our test data in the band of noise $P(A)$ using numerical integration method.



(a) ϵ as function of $\delta = p^{-1/d}$

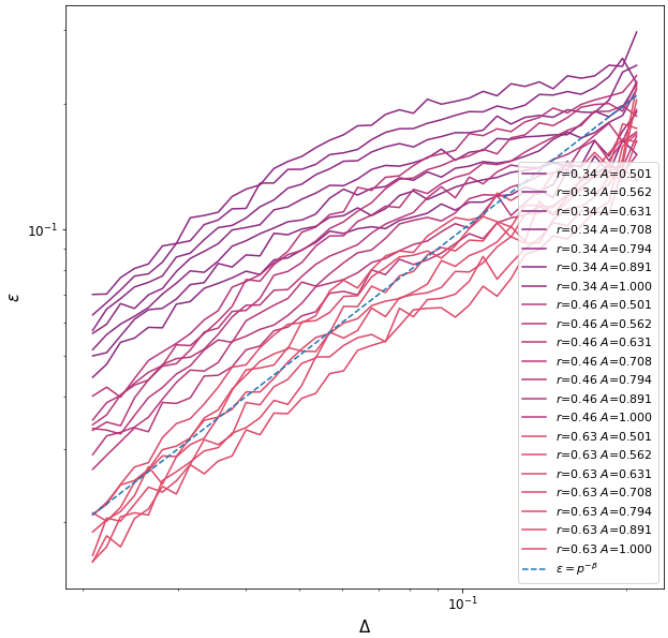
(b) $\epsilon/P(A)$ as function of $\delta = p^{-1/d}$

Figure 3: Test error for sufficiently small values of r to ensure $p \ll p*$. When rescaling by the function $P(A)$ defined in Eq.2.3, the test error curves collapse as predicted. The region of large δ is not relevant as it may suggest some other behaviour of the algorithm due to insufficient learning examples.

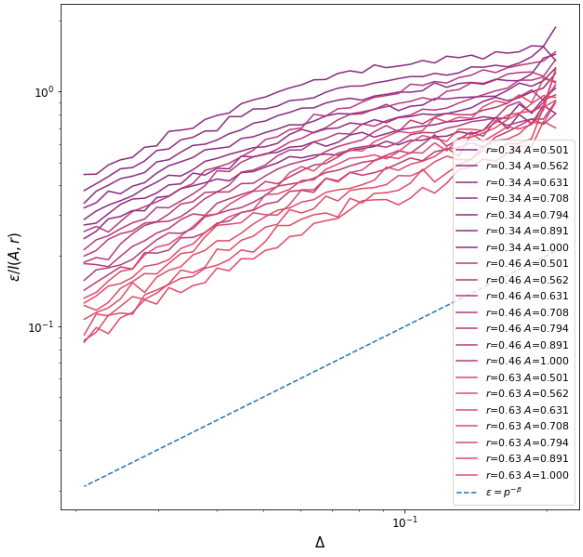
3.3 Case $p \gg p*$ and curvature dependence

We set ourselves in the context of section 2.1.2, where we work in a space of parameters relatively close to the standard deviation of the distribution. In the limit $p \longrightarrow \infty$, we wish to demonstrate the effect of curvature on the test error. We denote by $\kappa(A, r) := A/r^2$ the maximum curvature associated with the parameters and $I(A, r)$ the integral $\mathbb{E}\left[e^{-\frac{1}{2}A^2\sin^2(X/r)}\right]$ that is computed numerically using `scipy` package.

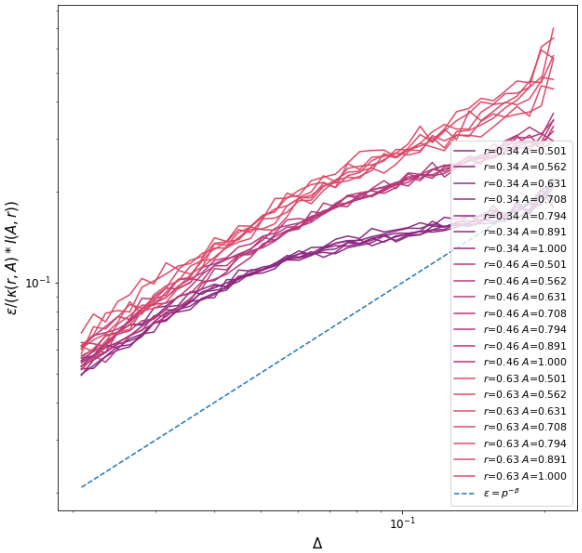
The results from Fig.4 suggest that a large number of training examples is required to detect the effects of curvature for given A, r. In the limit $p \longrightarrow \infty$ though, it should always be achievable.



(a) Original ϵ as function of $\Delta = p^{-\beta}$



(b) ϵ rescaled by the numerically computed integral $I(r, A)$.



(c) Twice rescaled ϵ with $\kappa(r, A) = A/r^2$

Figure 4: Test error as function of $p^{-\beta}$. Higher value for the same wavenumber correspond to higher amplitudes. By plotting the function $\epsilon = p^{-\beta}$, we see that for large p, the error behaves similarly to the flat interface case. In (b), we rescale the error by the integral function defined in Eq.2.4. This rescale seems to reduce the dependency in r, preserving the asymptotic behaviour $p^{-\beta}$. In (c), we rescale the error by the integral as well as the maximum curvature associated with A and r. We see that for large p the curves converge asymptotically to the same function being proportional to $p^{-\beta}$.

4 Conclusion and perspectives

In this paper we have elucidated the influence the topology of an interface has on the test error function when treating the case of a sine wave. Furthermore, we have provided the asymptotic formulas for the performance of SVC algorithm on such problems. In particular, there exist two regimes and a critical training set size $p*$ under which the kernel classification is unable to perform. This is mainly due to the nearest distance between training point being larger than the actual wavelength of the decision function.
On the other hand, we explored the limits of large training data size to recover the results from Spigler et al. to a certain extent. The dependence of the error with respect to the curvature of the function is what differentiates the most the two works.
This opens up some study around different topology of classifiers from which we can mention the Gaussian Process Classification.

References

[1] Jonas Paccolat, Stefano Spligler, Curse of Dimensionality in kernel methods, March 9, 2020 (in writing procedure paper).

[2] A. Engel, C. Van den Broeck, *Statistical Mechanics of Learning*, Cambridge University Press

[3] B. Schölkopf, A. J. Smola, *Learning with Kernels*, MIT Press

Appendix

Here we present derivations of results throughout the paper.

A Taylor expansion of the expectation for large r case

Result: Consider the error function being equal (up to an constant) to the following expected value :

$$\epsilon = \mathbb{E}\left[e^{-\frac{1}{2}A^2\sin^2(X/r)}\right] := \mathbb{E}\left[f(X)\right] \tag{A.1}$$

If r is sufficiently large, we can expand the function f around the mean value of X yielding :

$$\epsilon = \left(1 - \frac{A^2}{2r^2}\right) \tag{A.2}$$

Derivation:

$$\begin{aligned}\mathbb{E}\left[f(X)\right] &= \mathbb{E}\left[f(\mu_X + (X-\mu_X)\right] \\ &\approx \mathbb{E}[f(\mu_X) + f'(\mu_X)(X-\mu_X) + \frac{1}{2}f''(\mu_X)(X-\mu_X)^2] \\ &= f(\mu_X) + \frac{1}{2}f''(\mu_X)\sigma_X^2\end{aligned} \tag{A.3}$$

For a random variable $X \sim \mathcal{N}(0,1)$ and a function $f(X) = e^{-\frac{1}{2}A^2\sin^2(X/r)}$, deriving twice and evaluating at $\mu_X = 0$ provides the above result.

B Saddle-point approximation of the integral for large A case

Result For $A \gg 1$ we can make a saddle point-approximation yielding to the result :

$$\begin{aligned}\epsilon &= \frac{\Delta}{2\pi}\int_{\mathbb{R}} dx e^{-\frac{1}{2}(x^2 + A^2\sin^2(x/r))} \\ &\approx \frac{r\Delta}{A\sqrt{2\pi}}\vartheta_3(0, e^{-(\pi r)^2/2}), \quad \vartheta_3(0,q) = \sum_{k=-\infty}^{\infty} q^{k^2}\end{aligned} \tag{B.1}$$

Derivation Making use of the reference [2], the saddle-point approximation takes the form :

$$I = \int_{x_1}^{x_2} dx g(x) e^{A^2 f(x)} \approx g(x_0)e^{A^2 f(x_0)}\sqrt{\frac{2\pi}{A^2|f''(x_0)|}}, \quad A^2 \longrightarrow \infty \tag{B.2}$$

where x_0 stands for the point where f attains its maximum. First, we can rewrite our integral form of the error function by making the change of variable $x/r \mapsto x$. The $\sin^2$ function is periodic of period π thus we need to separate the integral in intervals $[(k-1/2)\pi, (k+1/2)\pi]$, $k \in \mathbb{Z}$. Finally the maximum of $-\sin^2(x)$ is attained in each interval at $x_k = k\pi$.

$$\begin{aligned}\epsilon &= \frac{r\Delta}{2\pi}\int_{\mathbb{R}} dx e^{(xr)^2/2} e^{A^2(-\sin^2(x)/2)} \\ &= \frac{r\Delta}{2\pi}\sum_{k=-\infty}^{\infty}\int_{k\pi-\pi/2}^{k\pi+\pi/2} dx e^{(xr)^2/2} e^{A^2(-\sin^2(x)/2)} \\ &\approx \frac{r\Delta}{A\sqrt{2\pi}}\sum_{k=-\infty}^{\infty} e^{-k^2(\pi r)^2/2} \equiv \frac{r\Delta}{A\sqrt{2\pi}}\vartheta_3(0, e^{-(\pi r)^2/2})\end{aligned} \tag{B.3}$$