A Curiosity Regarding Steganographic Capacity of Pathologically Nonstationary Sources

Andrew D. Ker

Oxford University Computing Laboratory, Parks Road, Oxford OX1 3QD, England.

ABSTRACT

Square root laws state that the capacity of an imperfect stegosystem – where the embedding does not preserve the cover distribution exactly – grows with the square root of cover size. Such laws have been demonstrated empirically and proved mathematically for a variety of situations, but not for nonstationary covers. Our aim here is to examine a highly simplified nonstationary source, which can have pathological and unpredictable behaviour. Intuition suggests that, when the cover source distribution is not perfectly known in advance, it should be impossible to distinguish covers and stego objects because the detector can never learn enough information about the varying cover source. However we show a strange phenomenon, whereby it is possible to distinguish stego and cover objects as long as the cover source is stationary for two pixels at a time, and then the capacity follows *neither* a square root law *nor* a linear law.

Keywords: Steganographic Capacity, Square Root Law, Steganography, Steganalysis, Information Theory

1. INTRODUCTION

Imperfect steganography, where the embedding does not preserve all the statistics of the cover, is very different from perfect steganography. Although there are theoretical constructions for the latter, the former applies to all known practical steganographic embedding in digital media, text, and other non-artificial sources, because the cover source distributions will never be perfectly known[1] and so cannot be preserved exactly. And the capacity of the two scenarios differs fundamentally: perfect steganography typically allows embedding of payload linear in the size of the cover (at up to the entropy rate of the cover source[2]), whereas the "Square Root Law" applies to imperfect embedding.

Square root laws have been demonstrated empirically,[3] and proved for a variety of situations, including various models of cover and imperfect embedding,[4–6] but the case of nonstationary covers has never been considered. Our aim here is to examine a highly simplified nonstationary source: a binary stream where the probability of the two symbols can vary arbitrarily at every step. Even given access to another cover source with synchronized probabilities, it should be impossible to distinguish covers and stego objects because the detector can never learn enough information about a constantly varying source. This turns out to be true as long as the embedding preserves the first-order statistics of the cover source (a much weaker condition than perfect security). However we show a strange phenomenon, whereby it *is* possible to distinguish stego and cover objects, given a synchronized cover oracle, as long as the cover source is stationary for two symbols at a time, and then imperfect embedding capacity follows an intermediate law, $O(n^{3/4})$ where n is the cover size.

We stress that these results apply to a completely artificial model and we do not claim that it reflects the practice of steganography in digital media. Nonetheless, it demonstrates that asymptotically perfect detection is possible for nonstationary sources (given a synchronized cover oracle). Moreover, the dichotomy between linear capacity laws (perfect steganography) and square root laws (imperfect steganography) is not necessarily correct. It also highlights the importance for first-order security, even when perfect security is impossible.

We will briefly review the classic square root law in Sect. 2 then turn to constantly varying sources in Sect. 3, where we prove capacity laws for the cases of perfect and imperfect knowledge. The results are briefly discussed in Sect. 4.

Further author information:
A. D. Ker: E-mail: adk@comlab.ox.ac.uk, Telephone: +44 1865 283530

Throughout the paper, vectors will be denoted boldface ($\boldsymbol{X}$, $\boldsymbol{\mu}$, $\mathbf{0}$ the zero vector) and matrices as uppercase Greek (Σ, I the identity). The notation $\boldsymbol{X} \sim \mathbf{N}(\boldsymbol{\mu}, \Sigma)$ indicates that the random vector $\boldsymbol{X}$ has the multivariate normal distribution with mean $\boldsymbol{\mu}$ and covariance matrix Σ, including the case when Σ is singular. The notation $\boldsymbol{S} \sim \mathbf{M}(n, \boldsymbol{\phi})$ indicates that $\boldsymbol{S}$ has the multinomial distribution, where n samples are allocated into categories with respective probabilities $\boldsymbol{\phi}$. $X \sim \mathrm{U}[0, 1]$ indicates a uniform random scalar. $\mathrm{E}[-]$ denotes expectation, and $\mathrm{Var}[-]$ the variance (covariance) matrix of a random scalar (vector). $\mathbb{I}_A$ is the indicator random variable for the event A. The Kullback-Leibler (KL) divergence between two distributions will be denoted $D_{\mathrm{KL}}(X \,\|\, Y)$, where X and Y are random variables or vectors with the distributions concerned.

2. SQUARE ROOT LAWS

In the terminology of statistics, a detector is a hypothesis test. Typically there is a scenario which depends on a number of parameters, including a problem size n, all of which are known except for a single unknown (things are more complex when additional parameters are unknown, as we shall discuss later). The detector is to decide between (usually) two values for the unknown parameter. In the case of steganalysis, the parameter n is usually the size of the cover object, and the unknown parameter is γ, the rate of embedded payload for which the two cases are zero or some known, positive, alternative*. Square root laws concern the accuracy of the detectors, asymptotically as $n \to \infty$, when γ bears some asymptotic relationship to n.

To shorten the statement of our results, we introduce some convenient terminology for situations which appear in the literature on square root laws:

(i) We say that there is *asymptotically perfect detection* if there exists family of detectors (parameterised by n) such that, for sufficiently large n, the false positive and false negative error rates become arbitrarily small. In statistics this concept is referred to as *consistency*.

(ii) We say that there is *asymptotic perfect security* if, for every possible detector and any given bound, for sufficiently large n the detector's error rates exceed the bound.

These embody the two cases of an embedder's ever-increasing, or ever-decreasing, risk which appeared in the very first publication on the capacity of imperfect stegosystems.[7] In this paper, we will also prove some results which say that there is *no asymptotically perfect detection*: this is an intermediate situation in which we can say that the embedder's risk of perfect detection does not tend to one, but neither does it guarantee that the risk tends to zero.

We now re-state the simplest possible square root law (even simpler than that found in Ref. 4, from which the proof can be adapted). It applies to "hiding" in independent and identically distributed (i.i.d.) random bit streams, where the hiding involves (pseudorandomly) replacing cover bits with stego bits which have a different distribution. In practice this need not be literal replacement: it also covers the case where the cover bits are modified in any way which does not preserve their distribution.

Theorem 1. *Suppose that a cover consists of n bits, independent and identically distributed, taking value 1 with fixed probability $p \neq 0, 1$. Each cover bit may be replaced by a stego bit, and the stego bits are independent of everything else, taking value 1 with probability $q \neq p$. Suppose that each location is used as a stego bit with probability γ, independent of everything else. As $n \to \infty$,*

(i) if $\gamma\sqrt{n} \to \infty$ then there is asymptotically perfect detection;

(ii) if $\gamma\sqrt{n} \to 0$ then there is asymptotic perfect security.

Note that we ruled out the case of deterministic covers ($p = 0$ or 1) and *perfect embedding* ($p = q$). It is of key importance that the embedding be *imperfect*, i.e. it does not preserve the cover distribution exactly. We, and others, have argued[1,6] that it is a practical impossibility for embedding to preserve completely the distribution of realistic covers, though later in this paper we will consider a weaker form of distribution-preservation which is achievable.

*In fact, it will not make any difference to our results whether the alternative is known or not, because all the tests we construct are uniform in γ.

The embedding probability γ is related to the nominal payload size which caused the embedding changes: in the absence of adaptive source coding, the payload size m would be proportional to γn. Hence $m = O(\sqrt{n})$ is the critical rate: asymptotically perfect detection is possible above this rate, asymptotic perfect security holds below it. Thus a *square root law*.

The model of covers is highly simplified, but the same result can be shown for covers with arbitrary numbers of pixel colours,[4] where the covers form a Markov chain,[5] for payload which is of fixed size (rather than each location being used independently at random),[4] and equivalent results for embedding using adaptive source coding.[8] Note that in Theorem 1 the detector must know the value of p, which is to say that they must have perfect knowledge of the cover source. An interesting alternative is that they learn about the cover source from a cover oracle, and a suitably modified square root law is proved in Ref. 6. Our aim in this paper is to consider whether a related result could possibly hold for nonstationary sources.

3. THE SIMPLEST PATHOLOGICALLY NONSTATIONARY SOURCE

We consider a very simple nonstationary cover source, which represents a severe challenge. We continue to suppose that the cover consists of independent bits, which we denote $(X_1, X_2, \ldots)$, but suppose that the probability of a 1 bit can vary *at every step*. That is, $\Pr[X_i = 1] = p_i$, where p_i is some arbitrary sequence in $(0, 1)$. This represents a very difficult cover to perform steganalysis on, as its future behaviour is completely unpredictable from past behaviour. We also assume that stego bits can vary in probability, again changing arbitrarily often, so that if pixel i is used as a stego location then $\Pr[X_i = 1] = q_i$. Again, (q_i) may be arbitrary. Since are considering imperfect steganography, we assume that $q_i \neq p_i$ for all i, though it is sufficient for the inequality to hold for any positive proportion of indices.

We need to rule out $p_i = 0, 1$ so that the cover is not deterministic at any point, and we also need to rule out the fiddly situations $p_i \to 0$, $p_i \to 1$, or $p_i - q_i \to 0$ as $i \to \infty$; this last would mean that the distribution of stego bits converges to that of cover bits, which would certainly disturb any asymptotic results. To exclude these, and similar, cases we will make the following broad assumption:

$$(*) \quad \exists \epsilon > 0 \text{ such that, for all } i,\ \epsilon \leq p_i \leq 1 - \epsilon, \text{ and } |q_i - p_i| \geq \epsilon.$$

It is possible to perform a more delicate analysis, giving a slight weakening of $(*)$ without changing the conclusions of our results, but it complicates the exposition and we are content with the assumption as stated.

3.1 Perfect Knowledge

First consider the case analogous to Theorem 1, when the detector has perfect knowledge of the distribution of cover and stego bits. With nonstationary sources, this means knowledge of all p_i and q_i. Then the nonstationarity presents no particular difficulty:

Theorem 2. *Suppose that a cover consists of n independent bits, and bit i takes value 1 with probability p_i. If replaced by a stego bit, bit i will take value 1 with probability q_i instead. Each location is used as a stego bit with probability γ, independent of everything else. Also, assume $(*)$. Then, as $n \to \infty$,*

(i) if $\gamma\sqrt{n} \to \infty$ then there is asymptotically perfect detection;

(ii) if $\gamma\sqrt{n} \to 0$ then there is asymptotic perfect security.

Proof. The proof is very similar to that of Theorem 1, *mutatis mutandis*. For convenience, let us write $r_i = q_i - p_i$.

For (i) we construct an asymptotically perfect detector as follows. Let $(X_1, X_2, \ldots)$ be the observed binary stream and set

$$T = \sum_{i=1}^{n} (X_i - p_i) r_i. \tag{1}$$

Each bit is a random mixture, in the ratio $1-\gamma:\gamma$, between the streams with probability (p_i) and (q_i), and everything is independent of everything else, so $P(X_i = 1) = p_i + \gamma r_i$. Hence

$$\mathrm{E}[T] = \gamma \sum_{i=1}^{n} r_i^2 \geq \gamma c n, \quad \mathrm{Var}[T] = \sum_{i=1}^{n} r_i^2 (p_i + \gamma r_i)(1 - p_i - \gamma r_i) \leq dn, \tag{2}$$

where c and d are positive constants. The inequalities are justified by $(*)$, which ensures that all sums are linear in n.

Of course, $\mathrm{E}[T] = 0$ if $\gamma = 0$. So we define a detector to give a positive detection if $T > k\sqrt{n}$, where k is a positive constant. The probability of a false positive is

$$\Pr[T > k\sqrt{n}] \leq \frac{\mathrm{Var}[T]}{\big(k\sqrt{n} - \mathrm{E}[T]\big)^2} \leq \frac{dn}{(k\sqrt{n})^2} = \frac{d}{k^2},$$

by Chebychev's inequality, and this can be made arbitrarily small by large enough choice of k. By the same argument, the probability of a false negative is

$$\Pr[T \leq k\sqrt{n}] \leq \frac{\mathrm{Var}[T]}{\big(\mathrm{E}[T] - k\sqrt{n}\big)^2} \leq \frac{dn}{(\gamma c n - k\sqrt{n})^2} \to 0$$

as long as $\gamma\sqrt{n} \to \infty$, regardless of k. Thus we have constructed a detector with arbitrarily small error rates for sufficiently large n.

For (ii), consider the KL divergence between two individual bits X and X', which take 1 with probabilities p and $p + \gamma r$, respectively:

$$\begin{aligned} D_{\mathrm{KL}}(X \,\|\, X') &= -p \log\big(1 + \gamma \tfrac{r}{p}\big) - (1-p) \log\big(1 - \gamma \tfrac{r}{1-p}\big) \\ &\leq \gamma^2 \frac{r^2}{p(1-p)}, \end{aligned}$$

the inequality holding at least as long as γ is sufficiently small that the arguments to both logarithms are at least $\frac{1}{2}$, whereby we can apply $\log(1+x) \geq x - x^2$. Therefore, because all bits are independent, the KL divergence between a sequence of cover bits $\boldsymbol{X}$ and a sequence of stego bits $\boldsymbol{X'}$ satisfies

$$D_{\mathrm{KL}}(\boldsymbol{X} \,\|\, \boldsymbol{X'}) = \sum_{i=1}^{n} D_{\mathrm{KL}}(X_i \,\|\, X_i') \leq \gamma^2 \sum_{i=1}^{n} \frac{r_i^2}{p_i(1-p_i)} \to 0$$

if $\gamma\sqrt{n} \to 0$, again using $(*)$ to ensure that the sum is linear in n.

This ensures that the distributions of $\boldsymbol{X}$ and $\boldsymbol{X'}$ converge for sufficiently large n, and any detector attempting to distinguish them has arbitrarily high error rates: a standard argument using the data processing theorem, for example as in Ref. 7. ■

So the situation is similar to the classic square root law but with one significant difference: the asymptotically perfect detector (1) uses knowledge of all q_i as well as all p_i, whereas in the classic case it is not required to know q as well as p. In the nonstationary case, it is necessary at least to know whether $q_i > p_i$ or $p_i > q_i$, so that a "1" in place i represents a little piece of evidence in favour of, or against, a positive detection.

3.2 Imperfect Knowledge and Parallel Streams

Now consider the case, analogous to that in Ref. 6, where the detector does not have knowledge of the cover source, i.e. is ignorant of (p_i), but must learn about the cover from an oracle. The oracle is useless if its bit probabilities change independently of the embedder's source, so let us assume that the cover oracle is synchronized with that of the embedder. This means that the detector sees *two* parallel streams of bits, one a guaranteed cover (X_i), with arbitrarily varying probabilities $p_i = \Pr[X_i = 1]$, and one a potential mixture of cover and stego,

(Y_i) with $\Pr[Y_i = 1] = p_i + \gamma(q_i - p_i)$ for an arbitrary sequence q_i. Either $\gamma = 0$ (no steganography) or $\gamma > 0$ (steganography present). We may assume that γ the embedding rate, if not zero, is known to the detector, but in fact it will turn out that the detector will not use such knowledge. The detector is supposed to be ignorant of the sequence (p_i), and we will also assume that they are ignorant of (q_i) too, although this is inessential.

In this situation it is not always possible fully to distinguish cover and stego streams, even if the embedding rate does not diminish at all:

Theorem 3. *Given the observations*

$$(X_i), \text{ with } \Pr[X_i = 1] = p_i, \text{ and } (Y_i), \text{ with } \Pr[Y_i = 1] = p_i + \gamma(q_i - p_i),$$

assuming $(*)$*, there does not necessarily exist a detector that is ignorant of* (p_i) *and can distinguish* $\gamma = 0$ *and* $\gamma > 0$ *with asymptotically perfect detection, even if* γ *does not diminish.*

This is true because, given no knowledge of the p_i or q_i, the detector is forced to rely on first-order statistics, and there exist sources for which the first-order statistics are identical even when the cover and stego distributions are not.

Proof. It is useful to write indicator random variables for the four possible occurrences at each position in the parallel streams (X_i) and (Y_i):

$$\begin{aligned} Z_i^0 &= \mathbb{I}_{(X_i,Y_i)=(0,0)} \\ Z_i^1 &= \mathbb{I}_{(X_i,Y_i)=(0,1)} \\ Z_i^2 &= \mathbb{I}_{(X_i,Y_i)=(1,0)} \\ Z_i^3 &= \mathbb{I}_{(X_i,Y_i)=(1,1)} \end{aligned}$$

and the probabilities $\pi_i^j = \Pr[Z_i^j = 1]$, if $p_i' = p_i + \gamma(q_i - p_i)$, and as usual writing $r_i = q_i - p_i$,

$$\begin{aligned} \pi_i^0 &= (1-p_i)(1-p_i') &&= (1-p_i)^2 &&- (1-p_i)\gamma r_i \\ \pi_i^1 &= (1-p_i)p_i' &&= p_i(1-p_i) &&+ (1-p_i)\gamma r_i \\ \pi_i^2 &= p_i(1-p_i') &&= p_i(1-p_i) &&- p_i\gamma r_i \\ \pi_i^3 &= p_i p_i' &&= p_i^2 &&+ p_i\gamma r_i. \end{aligned} \tag{3}$$

We must consider what it means for a detector to be *ignorant of* (p_i) (and perhaps also of (q_i)). As discussed in Refs. 6 and 9, it can be difficult to impose a lack of knowledge. For any sequence (p_i) an asymptotically perfect detector *does* exist: it is the one from the previous section, which happens to have the correct sequence (p_i) hardwired into it. In Ref. 6 we solved the problem by imposing *unbiasedness* on the detector. Here, we have a more attractive option: a detector ignorant of (p_i) is certainly ignorant of any permutation applied to (p_i), so its behaviour should be the invariant under all permutations to the observations (X_i, Y_i). This is the statistical property of *invariance* (see, for example, chapter 6 of Ref. 10) and in our case it forces a detector to make a decision solely from the *number* of occurrences of the different cases $(X_i = 0, Y_i = 0)$, $(X_i = 0, Y_i = 1)$, etc., rather than the positions in which they occur. This is intuitive, as well as statistically rigorous.

So an ignorant detector is required to decide whether $\gamma = 0$ or $\gamma > 0$ based on the 4-dimensional vector $\boldsymbol{S} = (S^0, S^1, S^2, S^3)$, where $S^j = \sum_i Z_i^j$. This amounts to saying that only the first-order statistics of the observations can be considered, and points the way to an example of cover and stego distributions which cannot be perfectly separated.

To prove our result, it suffices to find *one* example of sequences (p_i) and (q_i), satisfying $(*)$, for which no asymptotically perfect detector distinguishes $\gamma = 0$ and $\gamma > 0$. We use

$$p_i = \tfrac{1}{2}, \quad q_i = \tfrac{1}{2} + (-1)^i\epsilon, \tag{4}$$

where ϵ is any positive number strictly less than $\frac{1}{2}$. In which case, (3) gives $\pi_i^0 = \pi_i^2 = \frac{1}{4} - (-1)^i\frac{\epsilon}{2}\gamma$ and $\pi_i^1 = \pi_i^3 = \frac{1}{4} + (-1)^i\frac{\epsilon}{2}\gamma$.

We may assume that the number of observations is even, say $2n$. We decompose $\boldsymbol{S}$ into the parts arising from odd and even positions:

$$S^j = T_1^j + T_2^j, \quad T_1^j = \sum_{i=0}^{n-1} Z_{2i+1}^j, \quad T_2^j = \sum_{i=1}^{n} Z_{2i}^j,$$

and note that the corresponding vectors $\boldsymbol{T_1}$ and $\boldsymbol{T_2}$ are both examples of multinomial distributions with four possible outcomes

$$\begin{aligned} \boldsymbol{T_1} &\sim \mathbf{M}(n, \boldsymbol{\phi}), \quad \boldsymbol{T_2} \sim \mathbf{M}(n, \boldsymbol{\psi}), \quad \text{where} \\ \boldsymbol{\phi} &= (\tfrac{1}{4} + \tfrac{\epsilon}{2}\gamma, \tfrac{1}{4} - \tfrac{\epsilon}{2}\gamma, \tfrac{1}{4} + \tfrac{\epsilon}{2}\gamma, \tfrac{1}{4} - \tfrac{\epsilon}{2}\gamma), \\ \boldsymbol{\psi} &= (\tfrac{1}{4} - \tfrac{\epsilon}{2}\gamma, \tfrac{1}{4} + \tfrac{\epsilon}{2}\gamma, \tfrac{1}{4} - \tfrac{\epsilon}{2}\gamma, \tfrac{1}{4} + \tfrac{\epsilon}{2}\gamma). \end{aligned}$$

The rest of the proof is conceptually simple, but technically difficult. The idea is to use the convergence of the multinomial to the multivariate normal distribution (the multivariate central limit theorem, see e.g. 2.18 of Ref. 11) which states that $\boldsymbol{T_1} \mathrel{\dot\sim} \mathbf{N}(n\boldsymbol{\phi}, n\Sigma_\phi)$ with Σ_ϕ some covariance matrix depending on ϕ. Similarly, $\boldsymbol{T_2} \mathrel{\dot\sim} \mathbf{N}(n\boldsymbol{\psi}, n\Sigma_\psi)$, hence $\boldsymbol{S} \mathrel{\dot\sim} \mathbf{N}\big(n(\boldsymbol{\phi}+\boldsymbol{\psi}), n(\Sigma_\phi + \Sigma_\psi)\big)$. Since $\boldsymbol{\phi}+\boldsymbol{\psi} = (\frac{1}{4}, \frac{1}{4}, \frac{1}{4}, \frac{1}{4})$ is a fixed mean, we would appeal to Lemma 6, in the Appendix, to deduce that no asymptotically perfect detector exists. However, "$\mathrel{\dot\sim}$" here is not a mode of convergence which applies to probabilities of false positive and negative. Furthermore, Σ_ϕ and Σ_ψ are singular.

We now address these technical issues. The reader may safely skip the rest of the proof.

In order to relate a detector for a limiting distribution with its performance on approximants, we need a strong version of convergence. Recall that the *total variation* between random variables or vectors X and Y (strictly speaking, between their distributions), is given by

$$D_{\mathrm{TV}}(X \,\|\, Y) = \sup_A \big|\Pr[X \in A] - \Pr[Y \in A]\big|$$

where the supremum is over measurable sets. We say that X_n *converges in total variation* to Y if $D_{\mathrm{TV}}(X_n \,\|\, Y) \to 0$ as $n \to \infty$. See section 2.9 of Ref. 11 for more on total variation.

Let $\boldsymbol{T_1}$ be as above and set $\boldsymbol{V_1} \sim \mathbf{N}(n\boldsymbol{\phi}, n\Sigma_\phi)$, where $n\Sigma_\phi$ is the covariance matrix of $\boldsymbol{T_1}$ (we need not compute it, though we note that it will be singular because the components of $\boldsymbol{T_1}$ are constrained to add to n). Although $\boldsymbol{T_1}$ converges to $\boldsymbol{V_1}$ in distribution, it does not do so in total variation, because $\boldsymbol{T_1}$ is constrained to an integer lattice while $\boldsymbol{V_1}$ inhabits a 3-dimensional hyperplane in $\mathbb{R}^4$. The solution is to add small random perturbations to the components of $\boldsymbol{T_1}$, making it into a continuous distribution on the same hyperplane while not completely destroying its connection to the original, unperturbed, variable: this idea is studied in detail in Ref. 12, from which we extract the key result. Note that the probabilities in $\boldsymbol{\phi}$ are bounded away from 0 (uniformly in γ), hence their ratios are bounded, and this allows the results of Ref. 12 to be applied.

Define another random vector $\boldsymbol{U_1} = (U^1 + U^2, U^1 - U^2, -U^1 + U^3, -U^1 - U^3)$ where each $U^j \sim \mathrm{U}[-\frac{1}{2}, \frac{1}{2}]$. This is two steps of the recursive construction in Ref. 12, designed to preserve the sum of components, and giving $D_{\mathrm{TV}}(\boldsymbol{T_1} + \boldsymbol{U_1} \,\|\, \boldsymbol{V_1}) \to 0$ as $n \to \infty$.

We repeat the construction by adding $\boldsymbol{U_2} = (U^4 + U^5, U^4 - U^5, -U^4 + U^6, -U^4 - U^6)$ to $\boldsymbol{T_2}$; if $\boldsymbol{V_2} \sim \mathbf{N}(n\boldsymbol{\psi}, n\Sigma_\psi)$, where $n\Sigma_\psi$ is the covariance matrix of $\boldsymbol{T_2}$, we deduce $D_{\mathrm{TV}}(\boldsymbol{T_2} + \boldsymbol{U_2} \,\|\, \boldsymbol{V_2}) \to 0$. Finally, set

$$\boldsymbol{V} \sim \mathbf{N}\big(n(\boldsymbol{\phi}+\boldsymbol{\psi}), n(\Sigma_\phi + \Sigma_\psi)\big)$$

and apply VIII.10.14 of Ref. 13, deducing that

$$D_{\mathrm{TV}}(\boldsymbol{S} + \boldsymbol{U_1} + \boldsymbol{U_2} \,\|\, \boldsymbol{V}) \to 0. \tag{5}$$

To apply Lemma 6 we need a multivariate normal distribution with nondegenerate covariance matrix: not true of $\boldsymbol{V}$. Thankfully, we can simply eliminate the final component of $\boldsymbol{V}$ without losing any information, since its components are constrained to sum to n. Let us write $\boldsymbol{V'}$ for the 3-dimensional vector thus created. Then

$$\boldsymbol{V'} \sim \mathbf{N}\big(n\boldsymbol{\mu}, n\Sigma(\gamma)\big), \tag{6}$$

with $\boldsymbol{\mu} = (\frac{1}{4}, \frac{1}{4}, \frac{1}{4})$ independent of γ, so we can apply Lemma 6 to deduce that there is no asymptotically perfect detector for $\gamma = 0$ against $\gamma > 0$ based on $\boldsymbol{V}$.

This implies that there cannot be an asymptotically perfect detector based on $\boldsymbol{S} + \boldsymbol{U_1} + \boldsymbol{U_2}$ either, because for any region $R \subseteq \mathbb{R}^4$ where a positive or negative decision is given, $\Pr[\boldsymbol{S} + \boldsymbol{U_1} + \boldsymbol{U_2} \in R] - \Pr[\boldsymbol{V} \in R] \to 0$, by (5). Of course $\boldsymbol{S} + \boldsymbol{U_1} + \boldsymbol{U_2}$ is a version of $\boldsymbol{S}$ "corrupted" by some noise. Happily, we know that any decision on $\boldsymbol{S}$ need consider only integer arguments, and there is a strictly positive probability (independent of n) that the integer part of $\boldsymbol{U_1} + \boldsymbol{U_2}$ will be $\mathbf{0}$, so there is a strictly positive probability that a decision based on $\boldsymbol{S}$ will be identical to one based on $\boldsymbol{S} + \boldsymbol{U_1} + \boldsymbol{U_2}$, which in turn must have probability of error not tending to zero. ■

We have *not* proved asymptotic perfect security in this case, and indeed it does not hold: the mean of $\boldsymbol{S}$ does not depend on γ, but the covariance matrix does, and one can construct detectors based on this difference. Their performance is better than random, but not asymptotically perfect. Note that, for Theorem 3, it is necessary that the mean of (6) does not depend on γ. That follows because the overall effect of the embedding, in the example (4), is *first-order secure* in the sense that the long-run proportion of 0s and 1s is identical in cover and stego bit streams. It is clear that the same result would hold for any sequences (p_i) and (q_i) with the same first-order statistics. But if the sequences (p_i) and (q_i) were such that more 1s would be expected in stego streams than cover streams, or vice versa, a detector could be constructed similar to that in Theorem 1, and under mild conditions a square root law would hold. In Sect. 4 we will discuss the plausibility of first-order security, given implausibility of perfect embedding.

The reader might think that this is the whole story: one cannot perform steganalysis in pathologically nonstationary sources, even when given a reference source with synchronized probabilities, because there is no time to learn about the source before it changes. However, the situation is different if a very small modification is made.

Suppose it is known that (p_i) and (q_i) remain stationary for two bits each time, i.e. $p_{2i} = p_{2i+1}$ and $q_{2i} = q_{2i+1}$ for all i. The detector still does not have time for more than a cursory estimation of p_i, before it changes. But even this amount of stationarity unlocks a new detector, with the following performance.

Theorem 4. *Given the observations*

$$(X_i),\ \textit{with } \Pr[X_i = 1] = p_i,\ \textit{and } (Y_i),\ \textit{with } \Pr[Y_i = 1] = p_i + \gamma(q_i - p_i),$$

where $p_{2i} = p_{2i+1}$ *and* $q_{2i} = q_{2i+1}$ *for all* i*, and* $(*)$*, as* $n \to \infty$*,*

(i) if $\gamma n^{1/4} \to \infty$ *then an asymptotically perfect detector exists;*

(ii) if $\gamma n^{1/4} \to 0$ *then there does not necessarily exist a detector, ignorant of* (p_i)*, which can distinguish* $\gamma = 0$ *and* $\gamma > 0$ *with asymptotically perfect detection.*

Proof. We may assume that the cover size is even (if not, for (i) disregard the last observation and for (ii) we may permit a detector an additional observation). The proof now follows the same argument as that of Theorem 3. Considering now pairs of pixels in both cover and stego stream, define indicators for the 16 possibilities

$$\begin{aligned} Z_i^0 &= \mathbb{I}_{(X_{2i},X_{2i+1},Y_{2i},Y_{2i+1})=(0,0,0,0)} \\ Z_i^1 &= \mathbb{I}_{(X_{2i},X_{2i+1},Y_{2i},Y_{2i+1})=(0,0,0,1)} \\ Z_i^2 &= \mathbb{I}_{(X_{2i},X_{2i+1},Y_{2i},Y_{2i+1})=(0,0,1,0)} \\ &\vdots \qquad\qquad\qquad \vdots \\ Z_i^{15} &= \mathbb{I}_{(X_{2i},X_{2i+1},Y_{2i},Y_{2i+1})=(1,1,1,1)} \end{aligned}$$

and the probabilities $\pi_i^j = \Pr[Z_i^j = 1]$, with $p_i' = p_i + \gamma(q_i - p_i) = p_i + \gamma r_i$,

$$\begin{aligned}
\pi_i^0 &= (1-p_i)^2(1-p_i')^2 &&= (1-p_i)^4 &&- 2(1-p_i)^3\gamma r_i &&+ (1-p_i)^2\gamma^2 r_i^2 \\
\pi_i^1 = \pi_i^2 &= (1-p_i)^2 p_i'(1-p_i') &&= p_i(1-p_i)^3 + (1-p_i)^2(1-2p_i)\gamma r_i &&- (1-p_i)^2\gamma^2 r_i^2 \\
\pi_i^3 &= (1-p_i)^2 p_i'^2 &&= p_i^2(1-p_i)^2 + 2p_i(1-p_i)^2\gamma r_i &&+ (1-p_i)^2\gamma^2 r_i^2 \\
\pi_i^4 = \pi_i^8 &= p_i(1-p_i)(1-p_i')^2 &&= p_i(1-p_i)^3 - 2p_i(1-p_i)^2\gamma r_i &&+ p_i(1-p_i)\gamma^2 r_i^2 \\
\pi_i^5 = \pi_i^6 = \pi_i^9 = \pi_i^{10} &= p_i(1-p_i)p_i'(1-p_i') &&= p_i^2(1-p_i)^2 + p_i(1-p_i)(1-2p_i)\gamma r_i &&- p_i(1-p_i)\gamma^2 r_i^2 \\
\pi_i^7 = \pi_i^{11} &= p_i(1-p_i)p_i'^2 &&= p_i^3(1-p_i) + 2p_i^2(1-p_i)\gamma r_i &&+ p_i(1-p_i)\gamma^2 r_i^2 \\
\pi_i^{12} &= p_i^2(1-p_i')^2 &&= p_i^2(1-p_i)^2 - 2p_i^2(1-p_i)\gamma r_i &&+ p_i^2\gamma^2 r_i^2 \\
\pi_i^{13} = \pi_i^{14} &= p_i^2 p_i'(1-p_i') &&= p_i^3(1-p_i) + p_i^2(1-2p_i)\gamma r_i &&- p_i^2\gamma^2 r_i^2 \\
\pi_i^{15} &= p_i^2 p_i'^2 &&= p_i^4 + 2p_i^3\gamma r_i &&+ p_i^2\gamma^2 r_i^2.
\end{aligned} \tag{7}$$

We know that detectors ignorant of (p_i) and (q_i) must be based solely on the vector $\boldsymbol{S} = (S^0, \ldots, S^{15})$ with $S^j = \sum_i Z_i^j$.

For (i) we construct an asymptotically perfect detector: the key is to combine the probabilities to ensure that only terms quadratic in γr_i appear: this means that $\gamma > 0$ always gives a signature of positive sign, whether r_i is positive or negative. A statistic with this property is

$$T = S^3 - S^5 - S^{10} + S^{12}.$$

(In other words, count the number of times we observe $X_{2i} = X_{2i+1} \neq Y_{2i} = Y_{2i+1}$ and subtract the number of occurrences of $X_{2i} = Y_{2i} \neq X_{2i+1} = Y_{2i+1}$.)

Write $T = \sum T_i$ where $T_i = S_i^3 - S_i^5 - S_i^{10} + S_i^{12}$. It is easy to check that $\mathrm{E}[T_i] = \pi_i^3 - \pi_i^5 - \pi_i^{10} + \pi_i^{12} = \gamma^2 r_i^2$, and $|T_i| \leq 1$ gives $\mathrm{Var}[T_i] \leq 1$, so we have

$$\mathrm{E}[T] = \gamma^2 \sum r_i^2 \geq \gamma^2 cn, \quad \mathrm{Var}[T] \leq n.$$

Now the situation is identical to that in (2), but with γ^2 replacing γ. So for identical reasons to those in Theorem 2, part (i), a detector based on $T > k\sqrt{n}$ is asymptotically perfect as long as $\gamma n^{1/4} \to \infty$.

The proof of (ii) is related to that of Theorem 3. We use the same example

$$p_i = \tfrac{1}{2}, \quad q_i = \tfrac{1}{2} + (-1)^i\epsilon,$$

for which (7) gives

$$\begin{aligned}
\pi_i^0 = \pi_i^4 = \pi_i^8 = \pi_i^{12} &= \tfrac{1}{16} - (-1)^i\tfrac{\epsilon}{4}\gamma + \tfrac{\epsilon^2}{4}\gamma^2 \\
\pi_i^1 = \pi_i^2 = \pi_i^5 = \pi_i^6 = \pi_i^9 = \pi_i^{10} = \pi_i^{13} = \pi_i^{14} &= \tfrac{1}{16} - \tfrac{\epsilon^2}{4}\gamma^2 \\
\pi_i^3 = \pi_i^7 = \pi_i^{11} = \pi_i^{15} &= \tfrac{1}{16} + (-1)^i\tfrac{\epsilon}{4}\gamma + \tfrac{\epsilon^2}{4}\gamma^2.
\end{aligned} \tag{8}$$

Again, we isolate the contribution from odd and even positions,

$$S^j = T_1^j + T_2^j, \quad T_1^j = \sum_{i=0}^{n-1} Z_{2i+1}^j, \quad T_2^j = \sum_{i=1}^{n} Z_{2i}^j,$$

where

$$\boldsymbol{T_1} \sim \mathbf{M}(n, \boldsymbol{\phi}), \quad \boldsymbol{T_2} \sim \mathbf{M}(n, \boldsymbol{\psi})$$

and $\boldsymbol{\phi}$ and $\boldsymbol{\psi}$ are 16-dimensional vectors from (8), with i odd and even respectively.

Define

$$\boldsymbol{V} \sim \mathbf{N}\big(n(\boldsymbol{\phi} + \boldsymbol{\psi}), n(\Sigma_\phi + \Sigma_\psi)\big).$$

By applying the result from Ref. 12 again, which applies to multinomials of arbitrary dimension, we can deduce that $D_{\mathrm{TV}}(\boldsymbol{S}+\boldsymbol{U}\,\|\,\boldsymbol{V}) \to 0$ as $n \to \infty$, where $\boldsymbol{U}$ is some combination of bounded, uniform, random numbers (a more complex combination than in Theorem 3, but this affects nothing).

As before, we can drop the final component, leaving a 15-dimensional random vector which has nonsingular covariance matrix. This time, the key is to note that, by (8), $\boldsymbol{\phi}+\boldsymbol{\psi} = \boldsymbol{\mu}+\gamma^2\boldsymbol{\nu}$, where $\boldsymbol{\mu}$ and $\boldsymbol{\nu}$ do not depend on γ: there is no first-order term. Hence Lemma 7 applies, which gives no asymptotically perfect detector for $\gamma = 0$ against $\gamma > 0$ (in fact it gives asymptotically no detection power at all), and the same total variation argument transfers the conclusion to $\boldsymbol{S}$. ■

In this model of cover and stego streams, γ does have to diminish as n increases, but not as fast as either the stationary case (Theorem 1) or the perfect knowledge nonstationary case (Theorem 2). The rate must be $\gamma = O(n^{-1/4})$, which would mean (in the absence of source coding) a payload size of order $n^{3/4}$. One might expect that the law changes, moving towards a payload of order $n^{1/2}$, if the cover source has to be stationary for more than two bits at a time, but this is not the case. Exactly *the same* result holds in more restrictive cover models, where p_i and q_i are allowed to vary only once every k steps:

Theorem 5. *For any $k \geq 2$, if $p_{ki} = p_{ki+1} = \cdots = p_{ki+k-1}$ and $q_{ki} = q_{ki+1} = \cdots = q_{ki+k-1}$ then we have the same conclusion as in Theorem 4.*

Sketch Proof. We can construct the asymptotically perfect detector, if $\gamma n^{1/4} \to \infty$, in exactly the same way: use only the first 2 out of each group of k bits and ignore the rest. For (ii), we re-use the example

$$p_i = \tfrac{1}{2}, \quad q_i = \tfrac{1}{2} + (-1)^i\epsilon,$$

if k is even, and a simple modification such as

$$p_i = \tfrac{1}{2}, \quad q_{ki} = q_{ki+1} = \tfrac{1}{2} - \tfrac{1}{2}\epsilon, \quad q_{ki+j} = \tfrac{1}{2} + (-1)^j\epsilon \text{ for } 1 < j < k,$$

for which the probabilities $\sum_i \pi_i^j$ have no first-order dependence on γ. The random vector $\boldsymbol{S}$ has dimensionality 4^k, but the results of Ref. 12 still apply and the multivariate normal approximation has all the same properties as in Theorem 4. ■

4. DISCUSSION

We have analysed the asymptotic capacity of some artificial stego systems where the cover source is nonstationary. We do not claim that the model is realistic for steganography in practical covers: our choice of cover model was deliberately as unconstrained as possible, because if steganalysis is possible even in constantly varying sources, then it should be possible anywhere.

We have shown that a square root law remains valid in constantly varying sources, if the characteristics of the source are known exactly to the detector (Theorem 2). If they are not, we must assume that the detector learns something about the covers from another source, and in our model this only makes sense if their cover "oracle" has bit probabilities synchronized with that of the source used by the embedder (though we have not attempted to describe a realistic scenario in which this might arise). Then, we have shown, asymptotically perfect detection becomes impossible (Theorem 3). Note that this is not to say that there is asymptotic perfect security, and indeed there is not, but that the risk of detection does not grow without bound as n increases, even if payload is linear in the cover size.

However, it only requires the cover source to remain stationary for two bits at a time for the result to change (Theorem 4). Then the embedder must diminish their embedding rate γ at order $n^{-1/4}$. Assuming no adaptive source coding, this means that the capacity law is $n^{3/4}$: a curious result indeed. It indicates that there *is* a situation to fill the gap between perfect embedding (the linear capacity law) and imperfect embedding in stationary sources (the square root law). We highlight a difference between the statements of Theorems 2 and 4: in the former case, for sufficiently-fast diminishing γ, there is asymptotic perfect security; in the latter, there is no asymptotic perfect detection. Unlike the case of Theorem 3, we believe that this can be improved:

Conjecture. Under the conditions of Theorem 4, as $n \to \infty$,

(i) if $\gamma n^{1/4} \to \infty$ then an asymptotically perfect detector exists;

(ii) if $\gamma n^{1/4} \to 0$ then there is asymptotic perfect security.

This should hold because the asymptotic distribution of $\boldsymbol{S}$ and $\boldsymbol{V}$, in the proof of Theorem 4, is so close. The randomization needed to show convergence in total variation is of bounded magnitude, and hence asymptotically negligible as $n \to \infty$, and according to Lemma 7 there is asymptotic perfect security against an observation of $\boldsymbol{V}$. Perhaps the conjecture can be proved with only minor modifications to our current techniques.

Note that the detector would need to know exactly which pairs of bits, in the cover, had the same distribution: it must not be desynchronized. Again, we have not attempted to find a realistic situation in which this would occur. According to Theorem 5, any source which is stationary for $k \geq 2$ bits at a time, assuming that the detector knows which bits have the same distribution, gives the same result. This would give a $n^{3/4}$ capacity law for any k except $k = 1$, which has a linear law. It is also counterintuitive that, as $k \to \infty$, the cover source tends to stationarity, and yet a square root law has already been proved for exactly this situation in Ref. 6. Apparently the limiting behaviour as $k \to \infty$ does not match the behaviour at any finite k. This is not as paradoxical as it appears because such "behaviour" is itself a limit as $n \to \infty$. Were we to examine the performance of detectors at any finite n we would naturally see a strong dependence on k.

Finally, we note that the "secure" rate of $\gamma = O(n^{-1/4})$ only occurs if the embedding is *first-order secure*, i.e. that the first-order statistics of a cover stream match those of a stego stream. Given that we have ruled out perfect embedding, which preserves all statistics, as impracticable, how realistic is it to consider embedding which manages to preserve first-order statistics? (In stationary sources with independent components, first-order security is equivalent to perfect security.)

We believe that it is a highly plausible situation. It is not necessary to know the first-order statistics to preserve them because, for example, every change of pixel colour x to y can be counterbalanced with one of y to x: steganographic embedding with this property has been around for many years (e.g. Outguess[14]). Indeed it is the fact that permutation preserves first-order statistics which drives the perfect steganography construction of Ref. 2. The impracticability of perfect steganography arises from the need to preserve *all* statistics of the cover source, including the joint distributions of all tuples of bits/pixels/coefficients. So, in the continuum between completely imperfect embedding which preserves nothing, and perfect embedding which preserves everything, at least some possibilities are practicable.

It is much less plausible that the cover symbols should be uncorrelated, and indeed the failure of steganographic embedding which preserves first-order statistics (e.g. detectors for Outguess[15, 16]) effectively exploit a failure to preserve higher-order statistics of the cover source. An obvious extension of this work is to attempt to extend it to Markov chain cover models, in analogy with Ref. 5. The analysis is likely to be challenging.

ACKNOWLEDGMENTS

The author is a Royal Society University Research Fellow. Andrew Carter provided some useful ideas on multinomial approximations to the multivariate normal.

REFERENCES

[1] Böhme, R., *Improved Statistical Steganalysis using Models of Heterogeneous Cover Signals*, PhD thesis, Technische Universität Dresden (2008).

[2] Ryabko, B. and Ryabko, D., "Asymptotically optimal perfect steganographic systems," *Problems of Information Transmission* **45**(2), 184–190 (2009).

[3] Ker, A., Pevný, T., Kodovský, J., and Fridrich, J., "The square root law of steganographic capacity," in [*Proc. 10th ACM Workshop on Multimedia and Security*], 107–116 (2008).

[4] Ker, A., "The Square Root Law requires a linear key," in [*Proc. 11th ACM Workshop on Multimedia and Security*], 85–92 (2009).

[5] Filler, T., Ker, A., and Fridrich, J., "The square root law of steganographic capacity for Markov covers," in [*Media Forensics and Security XI*], *Proc. SPIE* **7254**, 0801–0811 (2009).

[6] Ker, A., "The square root law in stegosystems with imperfect information," in [*Proc. 12th Information Hiding Workshop*], *Springer LNCS* **6387**, 145–160 (2010).
[7] Ker, A., "A capacity result for batch steganography," *IEEE Signal Processing Letters* **14**(8), 525–528 (2007).
[8] Ker, A., "The Square Root Law does not require a linear key," in [*Proc. 12th ACM Workshop on Multimedia and Security*], 213–223 (2010).
[9] Ker, A., "The Uniform Prior and Zero Information: A Technical Note," Oxford University Computing Laboratory Research Report CS-RR-10-06 (2010).
[10] Lehmann, E. and Romano, J., [*Testing Statistical Hypotheses*], Springer, third ed. (2005).
[11] van der Vaart, A., [*Asymptotic Statistics*], Cambridge University Press (2000).
[12] Carter, A., "Deficiency distance between multinomial and multivariate normal experiments," *Annals of Statistics* **30**(3), 708–730 (2002).
[13] Feller, W., [*An Introduction to Probability Theory and Its Applications, Volume II*], Wiley (1966).
[14] Provos, N., "Defending against statistical steganalysis," in [*Proc. 10th USENIX Security Symposium*], (2001).
[15] Pevný, T. and Fridrich, J., "Merging Markov and DCT features for multi-class JPEG steganalysis," in [*Security, Steganography and Watermarking of Multimedia Contents IX*], *Proc. SPIE* **6505**, 0301–0314 (2007).
[16] Shi, Y., Chen, C., and Chen, W., "A Markov process based approach to effective attacking JPEG steganography," in [*Proc. 8th Information Hiding Workshop*], *Springer LNCS* **4437**, 249–264 (2006).

APPENDIX A. ASYMPTOTICALLY PERFECT DETECTION AND SECURITY FOR MULTIVARIATE NORMAL OBSERVATIONS

We prove some lemmas which are used in the body of the paper. Recall that the density function of a multivariate normal random variable $X \sim \mathbf{N}(\boldsymbol{\mu}, \Sigma)$ is

$$|2\pi\Sigma|^{-1/2} \exp\bigl(-\tfrac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu})^T\Sigma^{-1}(\boldsymbol{x}-\boldsymbol{\mu})\bigr),$$

as long as Σ is (symmetric) positive definite. When Σ is singular, the random vector's domain is constrained to a plane and the density does not exist.

The first lemma applies to perturbations in the covariance matrix, but not the mean, of a multivariate normal observation.

Lemma 6. *Let $\boldsymbol{\mu} \in \mathbb{R}^p$ be known, and let $\{\Sigma(\delta) \mid \delta \in [0,d]\}$ be a known family of symmetric, positive definite, $p \times p$ matrices. Let $\delta \in [0,d]$ be unknown. From an observation*

$$\boldsymbol{X}_\delta \sim \mathbf{N}\bigl(n\boldsymbol{\mu}, n\Sigma(\delta)\bigr)$$

there is no asymptotically perfect detector for the cases $\delta = 0$ against $\delta = \delta_1 \in (0,d]$.

Proof. Assume that $\Sigma(\delta_1) \neq \Sigma(0)$, otherwise the result is immediate. Since this detection is a simple hypothesis test, the Neyman-Pearson Lemma gives us the optimal decision function: a negative decision (that $\delta = 0$) if

$$\frac{|2\pi n\Sigma(0)|^{-1/2} \exp\bigl(-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu})^T\bigl(n\Sigma(0)\bigr)^{-1}(\boldsymbol{x}-\boldsymbol{\mu})\bigr)}{|2\pi n\Sigma(\delta_1)|^{-1/2} \exp\bigl(-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu})^T\bigl(n\Sigma(\delta_1)\bigr)^{-1}(\boldsymbol{x}-\boldsymbol{\mu})\bigr)}$$

is greater than some threshold. This minimizes the false negative rate for any given false positive rate. The region is equivalent to

$$A_c = \{\, \boldsymbol{y} \mid \boldsymbol{y}^T\bigl(\Sigma(0)^{-1} - \Sigma(\delta_1)^{-1}\bigr)\boldsymbol{y} < c\}$$

where $\boldsymbol{y}$ is an observation of

$$\boldsymbol{Y}_\delta = \frac{\boldsymbol{X}_\delta - \boldsymbol{\mu}}{\sqrt{n}}$$

and c is a constant determining the true and false negative rates. Note that the distribution

$$\boldsymbol{Y}_\delta \sim \mathbf{N}\big(\mathbf{0}, \Sigma(\delta)\big),$$

and A_c, are independent of n, and that A_c is a family of sets monotone increasing in c. Furthermore, the density of $\boldsymbol{Y}_\delta$ is strictly positive on $\mathbb{R}^p$. Therefore a sequence of detectors, one for each n, can only have the false negative rate tending to zero if $c \to \infty$, which forces the true negative rate to tend to zero also. Hence no asymptotically perfect detector exists. ■

The second lemma applies to locally-square perturbations in the mean of a multivariate normal observation.

Lemma 7. *Let $\boldsymbol{\mu}, \boldsymbol{\nu} \in \mathbb{R}^p$ be known, and let $\{\Sigma(\delta) \,|\, \delta \in [0, d]\}$ be a known family of symmetric, positive definite, $p \times p$ matrices continuous in δ. Let $\delta \in [0, d]$ be unknown. From an observation*

$$\boldsymbol{X}_\delta \sim \mathbf{N}\big(n\big(\boldsymbol{\mu} + \delta^2\boldsymbol{\nu} + O(\delta^3)\big), n\Sigma(\delta)\big)$$

there is asymptotically perfect security against detectors for the cases $\delta = 0$ against $\delta = \delta_1 \in (0, d]$, as long as $n\delta_1^4 \to 0$.

Proof. We will show that the KL divergence $D_{\mathrm{KL}}(\boldsymbol{X}_\delta \,\|\, \boldsymbol{X}_0)$ (note that this is the other way around from the usual square root law proofs) tends to zero if $n\delta^4 \to 0$.

It is known that the KL divergence between p-dimensional distributions $\mathbf{N}(\boldsymbol{\mu_0}, \Sigma_0)$ and $\mathbf{N}(\boldsymbol{\mu_1}, \Sigma_1)$ is

$$\frac{1}{2}\Big[\log\tfrac{|\Sigma_1|}{|\Sigma_0|} + \mathrm{tr}(\Sigma_1^{-1}\Sigma_0) - p + (\boldsymbol{\mu_1} - \boldsymbol{\mu_0})^T\Sigma_1^{-1}(\boldsymbol{\mu_1} - \boldsymbol{\mu_0})\Big]$$

so

$$D_{\mathrm{KL}}(\boldsymbol{X}_\delta \,\|\, \boldsymbol{X}_0) = \frac{1}{2}\Big[\log\Big(\tfrac{|n\Sigma(0)|}{|n\Sigma(\delta)|}\Big) + \mathrm{tr}\big(\Sigma(0)^{-1}\Sigma(\delta)\big) - p + \big(n\delta^2\boldsymbol{\nu} + O(n\delta^3)\big)^T\big(n\Sigma(0)\big)^{-1}\big(n\delta^2\boldsymbol{\nu} + O(n\delta^3)\big)\Big].$$

The first term tends to zero by continuity of Σ, determinant, and logarithm; similarly, the second term to $\mathrm{tr(I)} = p$ by continuity of Σ and trace. That leaves a leading term of $O(n\delta^4)$, and the usual data processing theorem argument completes the result. ■