IMPORTANT PROPERTIES OF NORMALIZED KL DIVERGENCE UNDER THE HMC MODEL

TOMÁŠ FILLER
(TOMAS.FILLER@BINGHAMTON.EDU)
(HTTP://DDE.BINGHAMTON.EDU/FILLER/)

Abstract. Normalized Kullback-Leibler (KL) divergence between cover and stego objects was shown to be important quantity for proofs around Square Root Law. In these proofs, we need to find Taylor expansion of this function w.r.t. change-rate β around $\beta = 0$ or be able to find an upper bound for some derivatives of this function. We can expect that this can be done since normalized KL divergence should be arbitrarily smooth in β. In this report, we show that every derivative of this function is uniformly upper-bounded and Lipschitz-continuous w.r.t. β and thus Taylor expansion of any order can be done.

1. Report in a nutshell

If you are reading this report just because you need its main result, you can only read this section. If cover image is modeled as Markov Chain and embedding operation is mutually independent (LSB, ± 1) and done with change-rate $\beta \in [0, \beta_{MAX}]$ (we call this HMC model), then we define normalized KL divergence between n-element cover distribution $P^{(n)}$ and n-element stego distribution $Q_\beta^{(n)}$ as

$$\frac{1}{n} d_n(\beta) = \frac{1}{n} D_{KL}\big(P^{(n)} || Q_\beta^{(n)}\big). \tag{1.1}$$

Main result of this report, formally stated in Theorem 3, tells us that every derivative of $d_n(\beta)/n$ w.r.t. β (and function $d_n(\beta)/n$ itself) is uniformly bounded and Lipschitz-continuous (or simply continuous) on $[0, \beta_0]$. These properties are independent of $n \geq 1$.

2. Introduction

Normalized KL divergence between n-element cover and stego distributions, defined by equation (1.1), was shown to be a key quantity for studying and proving Square Root Law (SRL). Although it is not hard to believe, that this function and its derivatives are well behaved functions (bounded and continuous), the proofs are somewhat lengthy and not inspirational and thus they are presented in this report, separate from other and more important results.

In the rest of this section, we describe the notation used in this report. Section 3 gives a summary of the assumptions we are using and derives some of their basic consequences. Main result of this report is formulated in Section 4 and proved in Section 5. All necessary results from the theory of hidden Markov chains are presented in Appendix A.

In addition to the notation developed before, we use the following symbols. We [notation]

use $\mathcal{P}_\epsilon(\mathcal{X})$ to denote set of probability distributions on set $\mathcal{X} = \{1, \ldots, N\}$ lower-bounded by ϵ, i.e., $p = (p_1, \ldots, p_N)^T \in \mathcal{P}_\epsilon(\mathcal{X}) \Rightarrow p_i \geq \epsilon$. We define $\mathbb{B}(y) = (b_{i,j}(y))$ as diagonal matrix with $b_{ii}(y) = b_{i,y}$ and vectors $b(y) = (b_{1,y}, \ldots, b_{N,y})^T$, $e = (1, \ldots, 1)^T$, e_i as i-th standard basis vector. Sometimes we write $\mathbb{B}_\beta(y)$ and $b_\beta(y)$ to stress the dependency on parameter β. We write ∂f as a shorthand for $\frac{\partial}{\partial \beta} f$. For vector x and matrix $\mathbb{M}$, we denote $\|x\|_1$the L_1 norm, $\|x\|_1 = \sum_i |x_i|$,$\|x\|$ the L_2 norm, $\|x\| = (\sum_i x_i^2)^{1/2}$, and $\|\mathbb{M}\|$ the 2-norm of matrix $\mathbb{M}$, i.e. $\|\mathbb{M}\| = \sup_{x \neq 0} \frac{\|\mathbb{M}x\|}{\|x\|} = \sup_{\|x\|=1} \|\mathbb{M}x\|$.

[3, S. 2.2,p. 14–15]

3. Basic assumptions and their consequences

We use the following assumption for deriving all the results.

Assumption 1. [HMC model] *Let $\{X_n\}$ be Markov Chain defined over set $\mathcal{X} = \{1, \ldots, N\}$ with probability transition matrix $\mathbb{A} = (a_{ij})$ and initial distribution on $Pr(X_1) = \pi$, where $\pi = (\pi_1, \ldots, \pi_N)$ is stationary distribution (left eigenvector of 1) of $\mathbb{A}$. Let $\{Y_n\}$ be non-deterministic function of $\{X_n\}$ defined by matrix $\mathbb{B} = (b_{i,j})$, where $Pr(Y_n = j | X_n = i) = b_{i,j}$. We assume $\mathbb{B}$ in the form $\mathbb{B}_\beta = \mathbb{I} + \beta\mathbb{C}$, where $\mathbb{C} = (c_{i,j})$ with $c_{i,j} \geq 0$ for $i \neq j$ and $\sum_j c_{i,j} = 0$ for all $i \in \mathcal{X}$. Parameter β is in range $[0, \beta_{MAX}]$, where β_{MAX} is given by embedding method. Finally, we assume that elements of $\mathbb{A}$ are lower-bounded by δ, $a_{i,j} \geq \delta$.*

Direct consequences of the above assumption are summarized in this corollary.

Corollary 2. *By the Perron-Frobenius theorem $\|\mathbb{A}^T\| = 1$. By $a_{i,j} \geq \delta > 0$, MC $\{X_n\}$ is irreducible and $\pi_i \geq \delta$ (see [1, p. 173, eq. 2.1]), $\pi \in \mathcal{P}_\delta(\mathcal{X})$. If $p \in \mathcal{P}_\delta(\mathcal{X})$, then $b^T(y)p \geq \delta \sum_i b_{i,y} \geq \delta(1 + \beta c_{y,y}) \geq \delta(1 + \beta_1 \min_y c_{y,y}) = \delta_1 > 0$ for $\beta \in [0, \beta_1]$, where $1 + \beta_1 \min_y c_{y,y} > 0$. We will need the following bounds, $\|b_\beta(y)\| \leq S_0$, $\|\partial b_\beta(y)\| = \|\mathbb{C}_{\bullet,y}\| \leq S_1$. By the assumption, $S_0 < \infty$ and $S_1 < \infty$.*

[definition of δ_1, β_1]

[definition of S_0, S_1]

4. Important properties of normalized KL divergence

Under the mentioned assumption, we state the main result of this report.

Theorem 3. *Every derivative of normalized KL divergence $\frac{1}{n} d_n(\beta) = D_{KL}\big(P^{(n)} || Q_\beta^{(n)}\big)$ between n-sample distributions of $\{X_n\}$ ($P^{(n)}$) and $\{Y_n\}$ ($Q_\beta^{(n)}$) embedded with change-rate β is uniformly bounded,*

$$\forall k \geq 0, \exists C_k < \infty, \forall n, \forall \beta \in [0, \beta_0], \Big| \frac{1}{n} \frac{\partial^k d_n(\beta)}{\partial \beta^k} \Big| < C_k, \tag{4.1}$$

and is Lipschitz-continuous (shortly Lipschitz) w.r.t. parameter β, i.e.,

$$\forall k \geq 0, \exists L_k < \infty, \forall n, \forall \beta, \beta' \in [0, \beta_0], \frac{1}{n} \Big| \frac{\partial^k d_n(\beta)}{\partial \beta^k} - \frac{\partial^k d_n(\beta')}{\partial \beta^k} \Big| < L_k |\beta - \beta'|. \tag{4.2}$$

Constant $\beta_0 > 0$ is given in the proof.

As a result of the above theorem, we have the fact, that $\frac{1}{n} d_n(\beta)$ and its derivatives are continuous functions of β.

5. PROOF OF THEOREM 3

We highly recommend the reader to read the Appendix A prior to this section. We use the concept of prediction filter from Appendix A to calculate normalized KL divergence and its derivatives.

First, approximate prediction filter is closed to $\mathcal{P}_\delta(\mathcal{X})$, i.e. if $p \in \mathcal{P}_\delta(\mathcal{X})$, then $f_\beta(y,p) \in \mathcal{P}_\delta(\mathcal{X})$. This holds, because equation (A.2) can be seen as a convex combination of rows of matrix $\mathbb{A}$ which are in $\mathcal{P}_\delta(\mathcal{X})$. Therefore, if $p^{(1)} = \pi$, then $p^{(n)} \in \mathcal{P}_\delta(\mathcal{X})$. From this we obtain the proof of (4.1) for $k=0$, because by using $\log P(X_1^n) \le 0$ and $\sum_{y_1^n} P(X_1^n = y_1^n) = 1$ it is sufficient to bound normalized log-likelihood $|l_n(\beta, y_1^n)| \le C_0$. This can be done, because $p_\beta^{(n)} \in \mathcal{P}_\delta(\mathcal{X})$ and $b_\beta^T(y) p_\beta^{(n)} \ge \delta_1$ for $\beta \in [0, \beta_1]$ and by (A.3) $C_0 = -\log \delta_1$.

[proof for $k=0$]

To prove (4.2) for $k=0$, it is sufficient to prove Lipschitz property for function $\log(b_\beta^T(y_i) p_\beta^{(i)})$. By Mean Value Theorem (MVT) used on function $\beta \to \log(v(\beta)^T z)$, for some vectors v and z, $|\log(v(\beta)^T z)/(v(\beta')^T z)| \le \max |\frac{(\partial v(\tilde\beta)^T) z}{v(\tilde\beta)^T z}| |\beta - \beta'|$ and thus

$$
\begin{aligned}
|\log(b_\beta^T(y_i) p_\beta^{(i)}) - \log(b_{\beta'}^T(y_i) p_{\beta'}^{(i)})| &\le \left| \log \frac{b_\beta^T(y_i) p_\beta^{(i)}}{b_{\beta'}^T(y_i) p_\beta^{(i)}} \right| + \left| \log \frac{b_{\beta'}^T(y_i) p_\beta^{(i)}}{b_{\beta'}^T(y_i) p_{\beta'}^{(i)}} \right| \\
&\le \frac{S_1}{\delta_1} |\beta - \beta'| + \frac{Lip(f) S_0}{\delta_1} |\beta - \beta'|.
\end{aligned}
$$

We use the fact that Lipschitz property of $p_\beta^{(i)}$ w.r.t. β (see Lemma 6), i.e. $\|f_\beta[y_1^n, p] - f_{\beta'}[y_1^n, p]\| \le Lip(f)|\beta - \beta'|$ implies $\|\partial f_\beta[y_1^n, p]\| \le Lip(f)$. This completes the proof of Theorem 3 for $k=0$.

The following lemma will be useful for proving Lipschitz property of some class of functions.

Lemma 4. *Let g_1, g_2 be real Lipschitz functions, then the following holds: (A) function $g_1 \pm g_2$ is Lipschitz; (B) if $|g_1|$, $|g_2|$ are upper bounded, then function $g_1 \cdot g_2$ is Lipschitz; (C) if $|g_1|$, $|g_2|$ are bounded from above and below, respectively and if $1/g_2$ is differentiable, then function $\frac{g_1}{g_2}$ is Lipschitz; (D) if g_1' and g_2' are Lipschitz and $|g_1'|$, $|g_2'|$ bounded, then $(g_1 \cdot g_2)'$ and $(g_1/g_2)'$ are Lipschitz.*

By (B), $(g_1)^k$ is Lipschitz for some fixed $k > 0$.

Proof. Let $|g_i(x) - g_i(x')| \le G_i |x - x'|$ for $i \in \{1,2\}$. (A) $|(g_1 \pm g_2)(x) - (g_1 \pm g_2)(x')| \le (G_1 + G_2)|x - x'|$. Let $G_1^- \le |g_i(x)| \le G_1^+$ for all possible x. (B) $|(g_1 \cdot g_2)(x) - (g_1 \cdot g_2)(x')| \le |(g_1(x)||g_2(x) - g_2(x')| + |g_2(x')||g_1(x) - g_1(x')| \le (G_1^+ G_2 + G_2^+ G_1)|x - x'|$. (C) $|\frac{g_1(x)}{g_2(x)} - \frac{g_1(x')}{g_2(x')}| \le \frac{1}{|g_2(x)|}|g_1(x) - g_1(x')| + |g_1(x')||\frac{1}{g_2(x)} - \frac{1}{g_2(x')}|$. By the MVT for function $1/g_2$, $\frac{1}{g_2(x)} - \frac{1}{g_2(x')} = \frac{-g_2'(\tilde x)}{(g_2(\tilde x))^2}(x - x')$. From the Lipschitz property of g_2, we obtain $|g_2'(\tilde x)| \le G_2$ and hence $|\frac{g_1(x)}{g_2(x)} - \frac{g_1(x')}{g_2(x')}| \le \left(\frac{G_1}{G_2^-} + \frac{G_1^+}{(G_2^-)^2} G_2\right)|x - x'|$. Case (D) holds, because $(g_1 \cdot g_2)' = g_1' \cdot g_2 + g_1 \cdot g_2'$ is Lipschitz by using (A),(B). Same holds for $(g_1/g_2)'$. □

$|g_2(x) - g_2(x')| \le G_2|x - x'|$ implies $|\frac{g_2(x) - g_2(x')}{x - x'}| \le G_2$

[proof for $k > 0$]

Now we show that if we have Lipschitz property and upper bound for derivatives of prediction filter up to order k, then we can prove (4.1) and (4.2) for k, i.e. we need $\|\partial^j p_\beta^{(i)} - \partial^j p_{\beta'}^{(i)}\| \le Lip(\partial^j f)|\beta - \beta'|$ and $\|\partial^j p_\beta^{(i)}\| \le P_j < \infty$ for $j \le k$. Result for $k=0$ has been established already. To prove (4.1) and (4.2) for $k > 0$, it is

sufficient to study the derivatives of normalized log-likelihood. First derivative of $l_n(\beta)$ w.r.t. β can be written as

$$\frac{\partial}{\partial\beta} l_n(\beta) = \frac{1}{n}\sum_{i=1}^{n} \frac{(\partial b^T(y_i))p^{(i)} + b^T(y_i)(\partial p^{(i)})}{b^T(y_i)p^{(i)}}. \tag{5.1}$$

Derivatives of $l_n(\beta)$ of order k can be expressed as an average of terms of the form $g_1/(b^T(y_i)p^{(i)})^{2^{k-1}}$, where g_1 is linear combination of dot-products of the following vectors $b^T(y_i)$, $\partial b^T(y_i)$, $p^{(i)}, \partial p^{(i)}, \ldots, \partial^k p^{(i)}$. By Lemma 4, we need upper bound on L_2 norm and Lipschitz property of these vectors to prove boundedness and Lipschitz property for this type of functions, because $|b^T(y_i)p^{(i)}| \geq \delta_1 > 0$ for $\beta \in [0, \beta_1]$. Vectors $b^T(y_i)$ and $\partial b^T(y_i)$ are bounded and Lipschitz in L_2 norm by Assumption 1 and thus we need to prove the same for $\partial^k p^{(i)}$. Uniform upper bound and Lipschitz property of $\partial^k p^{(i)}$ in L_2 norm are stated and proved in Lemma 7. Finally, we set $\beta_0 = \beta_3$, because $0 < \beta_3 \leq \beta_2 \leq \beta_1$ and thus all bounds are valid for $\beta \in [0, \beta_3]$.

Appendix A. Basics of Hidden Markov Chains

In this section, we give a summary of some important results about Hidden Markov Chains (HMCs). They are mainly obtained from the work of Mevel and Finesso [5] and are considered classics.

We view HMC as stochastic process $\{X_n, Y_n\}$, with $\{X_n\}$ being Markov Chain (MC), $X_n \in \mathcal{X} = \{1, \ldots, N\}$, and $\{Y_n\}$ non-deterministic function of $\{X_n\}$, however only $\{Y_n\}$ is observable.

$p^{(n)}$ is column vector.

For some fixed output $y_1^{n-1} \in \mathcal{X}^{n-1}$, we define vector $p^{(n)} = (p_1^{(n)}, \ldots, p_N^{(n)})^T$, called *prediction filter,* as $p_i^{(n)} = P(X_n = i | Y_1^{n-1} = y_1^{n-1})$. Sometimes we use $p_\beta^{(n)}$ to stress the dependency on β. Filter $p^{(n+1)}$ can be recursively calculated from $p^{(n)}$ and given observation y_n by using so called forward Baum equation as

This can be easily proved and is considered as a classical description of HMC. For details see [2, p. 1538].

$$p^{(n+1)} = \frac{\mathbb{A}^T \mathbb{B}(y_n) p^{(n)}}{b^T(y_n) p^{(n)}}. \tag{A.1}$$

Similarly as in [5], we define *approximate prediction filter* as

$$f_\beta(y, p) \triangleq \mathbb{A}^T \frac{\mathbb{B}(y)p}{e^T \mathbb{B}(y)p} = \mathbb{A}^T \frac{p_y e_y + \beta \mathbb{C}(y)p}{p_y + \beta e^T \mathbb{C}(y)p} = \mathbb{A}^T \frac{e_y + \frac{\beta}{p_y}\mathbb{C}(y)p}{1 + \frac{\beta}{p_y} e^T \mathbb{C}(y)p}, \tag{A.2}$$

where $\mathbb{C}(y) = diag(\mathbb{C}_{\bullet,y})$. Important case of this expression is $\beta = 0$, then $f_0(y, p) = (\mathbb{A}_{y,\bullet})^T$ regardless of p. This reflects the fact that there is no uncertainty about x if we observe y, because the case $\beta = 0$ represents MC.

For given observation sequence y_1^n, we define the *normalized log-likelihood function* as $l_n(\beta, y_1^n) = \frac{1}{n} \log Q_\beta(Y_1^n = y_1^n)$. This can be written in terms of prediction filter $p_\beta^{(i)}$ as

$$l_n(\beta, y_1^n) = \frac{1}{n}\sum_{i=1}^{n} \log\left(b_\beta^T(y_i) p_\beta^{(i)}\right), \tag{A.3}$$

because $\log Q_\beta(Y_1^n = y_1^n) = \log \prod_{i=1}^n b^T(y_i) Q_\beta(X_i | Y_1^{i-1} = y_1^{i-1})$.

Under the assumption that the initial distribution on MC $\{X_n\}$ is chosen to be stationary distribution π, then $p^{(1)} = \pi$. If $\beta = 0$, then from (A.1) we have $p^{(n)} = \pi$.

One of the key property of the approximate prediction filter is expressed in the following lemma. It states that for all $\beta \in [0, \beta_2]$ the approximate prediction filter satisfies contraction property in p and in β, independently of the choice of $y \in \mathcal{X}$.

Lemma 5. *Approximate prediction filter $f_\beta(y,p)$ satisfies the following contraction properties*

$$\|f_\beta(y,p) - f_\beta(y,q)\| \leq \lambda_1 \|p-q\| \tag{A.4}$$

$$\|f_\beta(y,p) - f_{\beta'}(y,p)\| \leq \lambda_2 |\beta - \beta'| \tag{A.5}$$

for all values of $\beta, \beta' \in [0,\beta_2]$, $p, q \in \mathcal{P}_\delta(\mathcal{X})$ and output $y \in \mathcal{X}$, where constants $\lambda_1 < 1$ and $\lambda_2 < \infty$ depend only on δ and matrix $\mathbb{C}$.

Proof. We use the vector form of the mean value theorem (V-MVT)[4] to derive both inequalities

$$\|f_\beta(y,p) - f_\beta(y,q)\| \leq \sup_{t\in[0,1]} \left\| \frac{\partial f(y, p+t(q-p))}{\partial p} \right\| \|p-q\|,$$

where $\mathbb{J}(\tilde{p}) = (j_{i,l}) \triangleq \frac{\partial f(y,\tilde{p})}{\partial p}$ is Jacobian matrix of function $f_\beta(y,p)$ w.r.t. p calculated at point $\tilde{p} = p + t(q-p)$. We consider the following bound for matrix 2-norm (see [3, 2.2-15, p. 15]) $\|\mathbb{M}\| \leq \sqrt{N} \max_j \sum_i |m_{i,j}| = \sqrt{N} \max_j \|\mathbb{M}_{\bullet,j}\|_1$ and calculate the j-th column of the Jacobian matrix $\mathbb{J}$ by differentiating (A.2). If $j \neq y$, then

Differentiate the last but one expression in (A.2).

$$\mathbb{J}(\tilde{p})_{\bullet,j} = \mathbb{A}^T \left(\frac{\beta \mathbb{C}(y) e_i}{\tilde{p}_y + \beta e^T \mathbb{C}(y) \tilde{p}} - \frac{\beta(\tilde{p}_y e_y + \beta \mathbb{C}(y)\tilde{p}) c_{i,y}}{(\tilde{p}_y + \beta e^T \mathbb{C}(y)\tilde{p})^2} \right) = \mathbb{A}^T \beta \mathbb{M}_j(\beta, y, p),$$

if $j = y$, then

Differentiate the last expression in (A.2).

$$\begin{aligned} \mathbb{J}(\tilde{p})_{\bullet,y} &= \mathbb{A}^T \beta \left(\frac{\mathbb{C}(y)}{1 + \beta e^T \mathbb{C}(y)\tilde{p}/\tilde{p}_y} - \frac{e^T \mathbb{C}(y)(e_y + \beta \mathbb{C}(y)\tilde{p}/\tilde{p}_y)}{(1+\beta e^T \mathbb{C}(y)\tilde{p}/\tilde{p}_y)^2} \right) \left(\frac{\tilde{p}_y e_y - p}{(\tilde{p}_y)^2} \right) \\ &= \mathbb{A}^T \beta \mathbb{M}_y(\beta, y, p), \end{aligned}$$

where $\mathbb{C}(y) = diag(\mathbb{C}_{\bullet,y})$. We know that $\|\mathbb{A}\| = 1$. By the Assumption 1 and by $e^T \mathbb{B}(y) p \geq \delta_1$ for $\beta \in [0, \beta_1]$, we can find $C < \infty$, such that $\|\mathbb{M}_j(\beta, y, p)\|_1 \leq C$ for all $j \in \mathcal{X}$ and thus we set $\lambda_1 = \sqrt{N} C \beta_2$, where β_2 satisfies $\beta_2 < (\sqrt{N}C)^{-1}$. If $\beta_2 > \beta_1$, then we set $\beta_2 = \beta_1$. Constant $\lambda_1 < 1$ does not depend on the choice of y, $\beta \in [0, \beta_2]$ and $p \in \mathcal{P}_\delta(\mathcal{X})$.

In order to prove the second statement, we find an upper bound for $\|\partial f(\tilde{\beta})/\partial\beta\|$, $\tilde{\beta} \in [\beta, \beta']$ by using V-MVT. Partial derivative of (A.2) w.r.t. β can be written as

$$\frac{\partial f_{\tilde{\beta}}(y,p)}{\partial \beta} = \mathbb{A}^T \left(\frac{\mathbb{C}(y)p}{e^T \mathbb{B}(y) p} - \frac{(p_y e_y + \tilde{\beta}\mathbb{C}(y)p) e^T \mathbb{C}(y) p}{(e^T \mathbb{B}(y) p)^2} \right).$$

Since $\beta, \beta' \in [0, \beta_2] \subset [0, \beta_1]$, $\tilde{\beta} \in [0, \beta_1]$ and thus $e^T \mathbb{B}(y) p \geq \delta_1$. By $\|\mathbb{A}\| = 1$, we can prove that $\|\partial f(\tilde{\beta})/\partial\beta\|$ is finite and can be bounded by λ_2. □

By using the above lemma, we can prove Lipschitz property of approximate prediction filter w.r.t. parameter β.

Lemma 6. *The functions $\beta \to f_\beta(y_1^n, p)$, such as $f_\beta(y_1^n, p) \triangleq f_\beta(y_n, f_\beta(y_1^{n-1}, p))$ are Lipschitz on $\mathcal{P}_\delta(\mathcal{X})$ w.r.t. $\beta \in [0, \beta_2]$, i.e., if $\beta, \beta' \in [0, \beta_2]$ then*

$$\omega(n) \triangleq \sup_{p \in \mathcal{P}_\delta(\mathcal{X})} \|f_\beta(y_1^n, p) - f_{\beta'}(y_1^n, p)\| \le Lip(f)|\beta - \beta'|.$$

Constant $Lip(f)$ does not depend on the choice of $y_1^n \in \mathcal{X}^n$.

Proof. We prove $\omega(n) \le \big(\lambda_2 + \lambda_2 \sum_{i=1}^{n-1} \lambda_1^i\big)|\beta - \beta'|$ for $\beta \in [0, \beta_2]$ by induction on n. By using (A.5), we have

$$\big\|f_\beta(y, p) - f_{\beta'}(y, p)\big\| \le \lambda_2 |\beta - \beta'|.$$

For $n > 1$ we have

$$\begin{aligned}\|f_\beta(y_1^n, p) - f_{\beta'}(y_1^n, p)\| \le \|f_\beta(y_n, f_\beta(y_1^{n-1}, p)) - f_{\beta'}(y_n, f_\beta(y_1^{n-1}, p))\| + \\ + \|f_{\beta'}(y_n, f_\beta(y_1^{n-1}, p)) - f_{\beta'}(y_n, f_{\beta'}(y_1^{n-1}, p))\|.\end{aligned}$$

By definition of prediction filter (A.2), $f_\beta(y_1, \bullet) : \mathcal{P}_\delta(\mathcal{X}) \to \mathcal{P}_\delta(\mathcal{X})$, because (A.2) can be seen as convex combination of rows of $\mathbb{A}$. By Lemma 5 , $\|f_\beta(y_n, f_\beta(y_1^{n-1}, p)) - f_{\beta'}(y_n, f_\beta(y_1^{n-1}, p))\| \le \lambda_2 |\beta - \beta'|$. By (A.4) and by the induction hypothesis, we can bound the second term as

$$\begin{aligned}\|f_{\beta'}(y_n, f_\beta(y_1^{n-1}, p)) - f_{\beta'}(y_n, f_{\beta'}(y_1^{n-1}, p))\| &\le \lambda_1 \|f_\beta(y_1^{n-1}, p) - f_{\beta'}(y_1^{n-1}, p)\| \\ &\le \lambda_1 \big(\lambda_2 + \lambda_2 \sum_{i=1}^{n-2} \lambda_1^i\big)|\beta - \beta'|\end{aligned}$$

and thus $\omega(n) \le \big(\lambda_2 + \lambda_2 \sum_{i=1}^{n-1} \lambda_1^i\big)|\beta - \beta'|$. By Lemma 5, $\lambda_1 < 1$ for $\beta \in [0, \beta_2]$ and thus the whole bound is convergent and $Lip(f) = \lim_{n\to\infty} \lambda_2 + \lambda_2 \sum_{i=1}^{n-1} \lambda_1^i = \lambda_2 + \frac{\lambda_1}{1-\lambda_1}$. □

Boundedness and Lipschitz property of $\|\partial^k p^{(i)}\|$ are stated and proved below.

Lemma 7. *The functions $\beta \to \partial^l f_\beta(y_1^n, p)$ are bounded and Lipschitz on $\mathcal{P}_\delta(\mathcal{X})$ w.r.t. $\beta \in [0, \beta_2]$, i.e., if $\beta, \beta' \in [0, \beta_3]$ then*

$$\sup_{p \in \mathcal{P}_\delta(\mathcal{X})} \|\partial^l f_\beta(y_1^n, p)\| \le P_l, \tag{A.6}$$

$$\sup_{p \in \mathcal{P}_\delta(\mathcal{X})} \|\partial^l f_\beta(y_1^n, p) - \partial^l f_{\beta'}(y_1^n, p)\| \le Lip(\partial^l f)|\beta - \beta'|. \tag{A.7}$$

Constants $Lip(\partial^l f)$ and P_l does not depend on the choice of $y_1^n \in \mathcal{X}^n$.

Proof. We prove (A.6) and (A.7) for $l = 1$ and show how to generalize this approach for higher derivatives. First derivative of prediction filter can be written as

$$\partial p^{(n+1)} = \partial f_\beta(y_1^n, p) = \mathbb{A}^T \partial \frac{\mathbb{B}(y_n) p^{(n)}}{b^T(y_n) p^{(n)}} = \mathbb{A}^T \mathbb{F} \partial p^{(n)} + \mathbb{A}^T \mathbb{G} p^{(n)}, \tag{A.8}$$

where

$$\mathbb{F} = \frac{\mathbb{B}(y_n)}{b^T(y_n) p^{(n)}} \left(\mathbb{I} - \frac{p^{(n)} b^T(y_n)}{b^T(y_n) p^{(n)}}\right) \quad \mathbb{G} = \frac{\partial \mathbb{B}(y_n)}{b^T(y_n) p^{(n)}} - \frac{\mathbb{B}(y_n) p^{(n)} \partial b^T(y_n)}{(b^T(y_n) p^{(n)})^2}. \tag{A.9}$$

In the rest of this proof, we will need $\|\mathbb{A}^T\mathbb{F}\| < 1$ for $\beta \in [0, \beta_2]$ which we prove now. If $\mathbb{C}(y) = diag(\mathbb{C}_{\bullet,y})$, then by $b^T(y)p \geq \delta_1$ and $\|\mathbb{A}\| = 1$

$$\|\mathbb{A}^T\mathbb{F}\| \leq \delta_1^{-1}\left\|\mathbb{A}^T\Big(e_y - \frac{p_y e_y + \beta\mathbb{C}(y)p}{p_y + \beta e^T\mathbb{C}(y)p}\Big)e_y^T + \beta\mathbb{A}^T\Big(\mathbb{C}(y) - \frac{p_y e_y + \beta\mathbb{C}(y)p}{p_y + \beta e^T\mathbb{C}(y)p}\mathbb{C}_{\bullet,y}^T\Big)\right\|$$
$$\leq \delta_1^{-1}\big\|f_0(y,p) - f_\beta(y,p)\big\| + \beta\big\|\mathbb{C}(y) - f_\beta(y,p)\mathbb{C}_{\bullet,y}^T\big\| \leq \beta\delta_1^{-1}(\lambda_2 + 2S_1),$$

where $p = p^{(n)}$, $y = y_n$ and thus we can find $0 < \beta_3 \leq \beta_2$ such that $\|\mathbb{A}^T\mathbb{F}_\beta\| \leq \beta_3\delta_1^{-1}(\lambda_2 + 2S_1) = \lambda_3 < 1$ for $\beta \in [0, \beta_3]$. We call this "contraction property" of $\mathbb{A}^T\mathbb{F}$. [definition of β_3]

By Assumption 1, $\|\mathbb{G}\|$ is upper bounded. By this and by contraction property of $\mathbb{A}^T\mathbb{F}$, $\|\partial p^{(n+1)}\| \leq \|\mathbb{A}^T\mathbb{F}\|\|\partial p^{(n)}\| + \|\mathbb{A}^T\|\|\mathbb{G}\|\|p^{(n)}\|$ is recurrent expression for an upper bound on $\|\partial p^{(n+1)}\|$. This upper bound converges to finite number P_1, because $\partial p^{(1)} = 0$ – initial distribution does not depend on β, it is equal to π. This bound does not depend on $p \in \mathcal{P}_\delta(\mathcal{X})$, $y \in \mathcal{X}$. $\|p^{(n)}\| \leq 1$ Same proof as in Lemma 6.

By Lemma 4, $\mathbb{F}$ and $\mathbb{G}$ are Lipschitz in 2-norm w.r.t. β, because they were obtained by combination of Lipschitz and bounded terms, remember $b^T(y)p \geq \delta_1$ and $\|\partial\mathbb{B}(y)\|$ if finite. Now we can prove (A.7), because by (A.8) Hint: add and subtract $\mathbb{A}^T\mathbb{F}_{\beta'}\partial p_\beta^{(n)}$.

$$\|\partial p_\beta^{(n+1)} - \partial p_{\beta'}^{(n+1)})\| \leq \|\mathbb{A}^T\mathbb{F}_{\beta'}\|\|\partial p_\beta^{(n)} - \partial p_{\beta'}^{(n)}\| + \|\mathbb{A}^T\|\|\mathbb{F}_\beta - \mathbb{F}_{\beta'}\|\|\partial p_\beta^{(n)}\| +$$
$$+ \|\mathbb{A}^T\|\|\mathbb{G}_\beta p_\beta^{(n)} - \mathbb{G}_{\beta'}p_{\beta'}^{(n)}\| \leq \lambda_3\|\partial p_\beta^{(n)} - \partial p_{\beta'}^{(n)}\| + (Lip(\mathbb{F}) + Lip(\mathbb{G}p))|\beta - \beta'|,$$

where we used Lipschitz property of $\mathbb{G}_\beta p_\beta^{(n)}$ w.r.t. β (use Lemma 4) and Lipschitz property of $\mathbb{F}$. Again, this recurrent bound converges to finite limit $Lip(\partial f)$, because of contraction property of $\mathbb{A}^T\mathbb{F}_{\beta'}$ and $\|\partial p_\beta^{(1)} - \partial p_{\beta'}^{(1)}\| = 0|\beta - \beta'|$.

By a closer look at higher derivatives of (A.2), we can realize that

$$\partial^l p^{(n+1)} = \mathbb{A}^T\mathbb{F}\partial^l p^{(n)} + R_{n,\beta}(y_n, p^{(n)}, \partial p^{(n)}, \ldots, \partial^{l-1}p^{(n)}),$$

where $R_{n,\beta}(\cdots)$ is Lipschitz w.r.t. derivatives of $p^{(n)}$ up to order $l-1$. Same observation was mentioned and used by Mevel and Finesso in [5, p. 1127]. This observation is possible, since $\mathbb{B}(y)$, $\partial\mathbb{B}(y)$ are bounded and Lipschitz and $\partial^l\mathbb{B}(y) = 0$ for $l \geq 2$. By this recursion, induction hypothesis ((A.6) and (A.7) holds up to $l-1$) and the fact that $\mathbb{A}^T\mathbb{F}$ is contracting, we can find finite upper bound P_l for $\|\partial^l p^{(n)}\|$. Same approach applies to (A.7). □

The following lemmas are related to the problem of exponential forgetting of the derivatives of the prediction filter. From Lemma 5, we know that prediction filter is forgetting its initial condition with exponential rate. By this result, sequence of realizations of prediction filters can be seen as nearly mutually independent and thus classical laws such as Central Limit Theorem (CLT) and Law of Large Numbers (LLN) can be proved. In the next, we will show that derivatives of prediction filter have similar property and thus as a result, we can prove the CLT for first derivative and LLN for second derivative of log-likelihood function. This is because from (A.3) the log-likelihood can be written as a sum of terms of prediction filters and its derivatives. The CLT for first derivative allows us to prove the LAN (local asymptotic normality) of log-likelihood ratio test statistics.

First we show some simple properties of matrices $\mathbb{F}$ and $\mathbb{G}$ from (A.8) which will be necessary.

Lemma 8. *Let matrix $\mathbb{F}(p)$ and $\mathbb{G}(p)$ be defined as in (A.9) for fixed prediction filter p, then these matrices are continuous and bounded in L_2 norm w.r.t. p for all $\beta \in [0, \beta_3]$, i.e. for $p, p' \in \mathcal{P}_\delta$*

$$\|\mathbb{F}(p) - \mathbb{F}(p')\| \le C_f \|p - p'\|, \qquad \|\mathbb{F}_p\| \le D_f < 1,$$
$$\|\mathbb{G}(p) - \mathbb{G}(p')\| \le C_g \|p - p'\|, \qquad \|\mathbb{G}_p\| \le D_g,$$

for some finite constants C_f, C_g, D_f, and D_g.

Proof. Boundedness of both matrices was proved and mentioned in previous lemmas (use $\|\mathbb{A}\| \le 1$), thus we prove the continuity only by using V-MVT [4]. By the same approach as in Lemma 5, it is sufficient to be interested in an upper bound on $\|\mathbb{J}(\tilde{p})_{\bullet,j}\|_1$, where $\mathbb{J}(\tilde{p}) = (j_{(i,l),k}) \triangleq (\partial \mathbb{F}_{il}(\tilde{p})/\partial p_k)$ is Jacobian matrix of size $N^2 \times N$ calculated at point $\tilde{p}$ on line between p and p'. We start with matrix $\mathbb{F}$ and calculate the k-th column of the Jacobian matrix as

$$\mathbb{J}(\tilde{p})_{\bullet,k} = -\frac{\mathbb{B}(y) b^T(y) e_k}{(b^T(y)p)^2} - \mathbb{B}(y) \frac{e_k b^T(y)(b^T(y)p)^2 - 2pb^T(y)b^T(y)pb^T(y)e_k}{(b^T(y)p)^4}.$$

By Assumption 1 and Corollary 2, the above matrix (think of it as big vector) is bounded in L_1 norm by some constant C_f. The same steps can be done to show the upper bound in the case of matrix $\mathbb{G}$. □

Now we can prove the fact that sequence $\big((p^{(n)}, \partial p^{(n)})\big)_{n=1}^{\infty}$ and possible extensions to higher order derivatives are exponentially forgetting their initial values $(p^{(1)}, \partial p^{(1)})$.

Lemma 9. *Function $(f, \partial f)_\beta(y_1^n, p, \partial p)$ defined as*

$$\big(f, \partial f\big)_\beta(y_1^n, p, \partial p) \triangleq \big(f_\beta(y_1^n, p), \partial f_\beta(y_1^n, p, \partial p)\big)$$

is forgetting its initial values $p \in \mathcal{P}_\delta(\mathcal{X})$ and $\partial p \in R^N$ on $\beta \in [0, \beta_3]$ with exponential rate, i.e., if $p, \hat{p} \in \mathcal{P}_\delta(\mathcal{X})$ and $\partial p, \partial \hat{p} \in R^N$ then

$$\|(f, \partial f)_\beta(y_1^n, p, \partial p) - (f, \partial f)_\beta(y_1^n, \hat{p}, \partial \hat{p})\| \le C\rho^n \|p - \hat{p}\| + \rho^n \|\partial p - \partial \hat{p}\|,$$

where $\rho < 1$ and C are constants independent of $y_1^n \in \mathcal{X}^n$ and choice of $\beta \in [0, \beta_3]$.

Proof. For fixed $y_1^n \in \mathcal{X}^n$ and $\beta \in [0, \beta_3]$, define

$$p^{(n+1)} = \begin{cases} p & \text{if } n = 0 \\ f_\beta(y_1^n, p) & \text{otherwise} \end{cases} \qquad \hat{p}^{(n+1)} = \begin{cases} \hat{p} & \text{if } n = 0 \\ f_\beta(y_1^n, \hat{p}) & \text{otherwise} \end{cases}$$
$$\partial p^{(n+1)} = \begin{cases} \partial p & \text{if } n = 0 \\ \partial f_\beta(y_1^n, p, \partial p) & \text{otherwise} \end{cases} \quad \partial \hat{p}^{(n+1)} = \begin{cases} \partial \hat{p} & \text{if } n = 0 \\ \partial f_\beta(y_1^n, \hat{p}, \partial \hat{p}) & \text{otherwise} \end{cases}$$

and sequences Δ_n and δ_n as $\Delta_n = \|\partial p^{(n)} - \partial \hat{p}^{(n)}\|$ and $\delta_n = \|p^{(n)} - \hat{p}^{(n)}\|$. By expanding $\partial p^{(n+1)}$ as in (A.8), we can find an upper bound on Δ_{n+1}

$$\begin{aligned} \Delta_{n+1} &\le \|\mathbb{A}^T\| \|\mathbb{F}(p^{(n)}) \partial p^{(n)} - \mathbb{F}(\hat{p}^{(n)}) \partial \hat{p}^{(n)} + \mathbb{G}(p^{(n)}) p^{(n)} - \mathbb{G}(\hat{p}^{(n)}) \hat{p}^{(n)}\| \\ &\le \|\mathbb{F}(p^{(n)})\| \Delta_n + \|\mathbb{F}(p^{(n)}) - \mathbb{F}(\hat{p}^{(n)})\| \|\partial \hat{p}^{(n)}\| + \\ &\quad + \|\mathbb{G}(p^{(n)})\| \delta_n + \|\mathbb{G}(p^{(n)}) - \mathbb{G}(\hat{p}^{(n)})\| \|\hat{p}^{(n)}\| \\ &\le D_f \Delta_n + P_1 C_f \delta_n + D_g \delta_n + C_g \delta_n = D_f \Delta_n + C_1 \delta_n, \end{aligned}$$

Matrix $\mathbb{F}$ depends on p, therefore $\mathbb{F}(p)$.

where we used continuity and boundedness proved in Lemma 8, the fact that $\|\partial p\|$ is bounded (see Lemma 7) and $C_1 = P_1 C_f + D_g + C_g$. By recursion, we obtain

$$\Delta_{n+1} \leq D_f \Delta_n + C_1 \delta_n \leq \cdots \leq D_f^n \Delta_1 + C_1 \sum_{i=0}^{n-1} D_f^i \delta_{n-i}.$$

From (A.4), we have $\delta_{n+1} \leq \lambda_1 \|p^{(n)} - \hat{p}^{(n)}\| \leq \lambda_1^n \delta_1$ and thus we can get rid of the sum

$$\begin{aligned}
\Delta_{n+1} &\leq D_f^n \Delta_1 + C_1 \Big(\sum_{i=0}^{n-1} (D_f/\lambda_1)^i \Big) \lambda_1^{n-1} \delta_1 \\
&\leq D_f^n \Delta_1 + C_1 \Big(\sum_{i=0}^{n-1} (\hat{D}_f/\lambda_1)^i \Big) \lambda_1^{n-1} \delta_1 \\
&\leq \hat{D}_f^n \Delta_1 + C_1 \frac{\hat{D}_f^n - \lambda_1^n}{\hat{D}_f - \lambda_1} \delta_1 \leq \hat{D}_f^n \Delta_1 + C_2 \hat{D}_f^n \delta_1,
\end{aligned}$$

We choose $\hat{D}_f$ in order to avoid case $\lambda_1 = D_f$. If $\lambda_1 \geq D_f$, then we choose $\lambda_1 < \hat{D}_f < 1$ and $\hat{D}_f = D_f$ otherwise.

where $C_2 = \frac{C_1}{\hat{D}_f - \lambda_1}$. Finally, we have

$$\begin{aligned}
&\|(p^{(n+1)}, \partial p^{(n+1)}) - (\hat{p}^{(n+1)}, \partial \hat{p}^{(n+1)})\| = \sqrt{\delta_{n+1}^2 + \Delta_{n+1}^2} \leq \delta_{n+1} + \Delta_{n+1} \\
&\quad \leq (\lambda_1^n + C_2 \hat{D}_f^n) \delta_1 + \hat{D}_f^n \Delta_1 \leq 2 C_2 \hat{D}_f^n \delta_1 + \hat{D}_f^n \Delta_1.
\end{aligned}$$

□

From the proof, we can see that exponential forgetting of ∂p is a consequence of exponential forgetting of p, continuity of matrices $\mathbb{F}$ and $\mathbb{G}$ and contraction of matrix $\mathbb{A}^T \mathbb{F}$ (forgetting previous ∂p). When we consider (A.8) and its higher order derivatives w.r.t. β, the same result (exponential forgetting) can be proved for vectors of higher order derivatives of the prediction filter $(p, \partial p, \ldots, \partial^l p)$. We formulate this in the next corollary which is presented without the proof, because all the asumptions (continuity, boundedness and contraction) of respective matrices are satisfied (this was discussed earlier in this report) and thus same approach can be used in the proof.

Corollary 10. *Function* $(f, \partial f, \ldots, \partial^l f)_\beta(y_1^n, p, \partial p, \ldots, \partial^l p)$ *defined as*

$$\big(f, \ldots, \partial^l f\big)_\beta(y_1^n, p, \ldots, \partial^l p) \triangleq \big(f_\beta(y_1^n, p), \partial f_\beta(y_1^n, p, \partial p), \ldots, \partial^l f_\beta(y_1^n, p, \partial p, \ldots \partial^l p)\big)$$

is forgetting its initial values $p \in \mathcal{P}_\delta(\mathcal{X})$ *and* $\partial p, \ldots, \partial^l p \in R^N$ *on* $\beta \in [0, \beta_3]$ *with exponential rate, i.e., if* $p, \hat{p} \in \mathcal{P}_\delta(\mathcal{X})$ *and* $\partial p, \partial \hat{p}, \ldots, \partial^l p, \partial^l \hat{p} \in R^N$ *then*

$$\begin{aligned}
&\|(f, \ldots, \partial^l f)_\beta(y_1^n, p, \ldots, \partial^l p) - (f, \ldots, \partial^l f)_\beta(y_1^n, \hat{p}, \ldots, \partial^l \hat{p})\| \\
&\quad \leq C_1 \rho^n \|p - \hat{p}\| + C_2 \rho^n \|\partial p - \partial \hat{p}\| + \cdots + C_{l+1} \rho^n \|\partial^l p - \partial^l \hat{p}\|
\end{aligned}$$

where $\rho < 1$ *and* C_i *are constants independent of* $y_1^n \in \mathcal{X}^n$ *and choice of* $\beta \in [0, \beta_3]$.

References

[1] J. L. Doob. *Stochastic processes*. Wiley, New York, 1st edition, 1953.

[2] Y. Ephraim and N. Merhav. Hidden Markov processes. *Information Theory, IEEE Transactions on*, 48(6):1518–1569, June 2002.

[3] G. Golub and C. Van Loan. *Matrix Computations*. The John Hopkins University Press, 1st edition, 1983.

[4] W. Hall and M. Newell. The mean value theorem for vector valued functions: a simple proof. *Mathematics Magazine*, 52(3):157–158, May 1979.

[5] L. Mevel and L. Finesso. Asymptotical statistics of misspecified hidden Markov models. *Automatic Control, IEEE Transactions on*, 49(7):1123–1132, July 2004.