

On equivalence between linear-chain conditional random fields and hidden Markov chains

Elie Azeraf^{1,2}^a, Emmanuel Monfrini²^b and Wojciech Pieczynski²^c

¹Watson Department, IBM GBS, avenue de l'Europe, Bois-Colombes, France

²SAMOVAR, CNRS, Telecom SudParis, Institut Polytechnique de Paris, Evry, France
 elie.azeraf@ibm.com, {emmanul.monfrini, wojciech.pieczynski}@telecom-sudparis.eu

Keywords: Linear-chain CRF, Hidden Markov chain, Bayesian segmentation, Natural language processing

Abstract: Practitioners successfully use hidden Markov chains (HMCs) in different problems for about sixty years. HMCs belong to the family of generative models and they are often compared to discriminative models, like conditional random fields (CRFs). Authors usually consider CRFs as quite different from HMCs, and CRFs are often presented as interesting alternatives to HMCs. In some areas, like natural language processing (NLP), discriminative models have completely supplanted generative models. However, some recent results show that both families of models are not so different, and both of them can lead to identical processing power. In this paper, we compare the simple linear-chain CRFs to the basic HMCs. We show that HMCs are identical to CRFs in that for each CRF we explicitly construct an HMC having the same posterior distribution. Therefore, HMCs and linear-chain CRFs are not different but just differently parametrized models.

1 INTRODUCTION


Let $Z_{1:N} = (Z_1, \dots, Z_N)$ be a stochastic sequence, with $Z_n = (X_n, Y_n)$. X_1, \dots, X_N take their values in a finite set Ω , while Y_1, \dots, Y_N take their values in a finite set Λ . Realizations of $X_{1:N} = (X_1, \dots, X_N)$ are hidden while realizations of $Y_{1:N} = (Y_1, \dots, Y_N)$ are observed, and the problem we deal with is to estimate $X_{1:N} = x_{1:N}$ from $Y_{1:N} = y_{1:N}$.


The simplest model allowing dealing with the problem is the well-known hidden Markov chain (HMC). In spite of their simplicity, HMCs are very robust and provide quite satisfactory results in many applications. We only cite the pioneering papers (Baum et al., 1970; Rabiner, 1989), and some books (Cappé et al., 2005; Koski, 2001), among great deal of publications. However, they can turn out to be too simple in complex cases and thus authors extended them in numerous directions. In particular, conditional random fields (Lafferty et al., 2001; Sutton and McCallum, 2006) are considered as interesting alternative to HMCs, especially in Natural Language Processing (NLP) area. They have also been used in different areas as diagnostic (Fang et al., 2018; Fang


et al., 2019), natural language processing (Jurafsky and Martin, 2009; Jurafsky and Martin, 2021), entity recognition (Song et al., 2019), or still relational learning (Sutton and McCallum, 2006). In general, authors consider CRFs as quite different from HMCs, and often prefers the former to the latter. In this paper, we show that CRFs and HMCs may be not so different. More precisely, we show that basic linear-chain CRFs are equivalent to HMCs.

Let us specify what “equivalence” in the paper’s title means. One can notice that HMCs and CRFs cannot be compared directly as they are of different nature. Assuming that a “model” is a distribution $p(x_{1:N}, y_{1:N})$, we may say that HMC is a model, while CRF is a family of models, in which all models have the same $p(x_{1:N}|y_{1:N})$, but can have any $p(y_{1:N})$. We will say that a CRF $p(x_{1:N}|y_{1:N})$ is equivalent to a HMC $q(x_{1:N}, y_{1:N})$ if and only if $p(x_{1:N}|y_{1:N}) = q(x_{1:N}|y_{1:N})$. To show that linear-chain CRFs are equivalent to HMCs it is thus sufficient to show that for each linear-chain CRF $p(x_{1:N}|y_{1:N})$, it is possible to find a HMC $q(x_{1:N}, y_{1:N})$ such that $p(x_{1:N}|y_{1:N}) = q(x_{1:N}|y_{1:N})$. This precisely is the contribution of the paper.

More generally, let us note that certain criticisms of the HMCs, put forward to justify the preference of the CRFs, currently appear to be not always entirely justified. For example, in monitoring prob-

^a <https://orcid.org/0000-0003-3595-0826>

^b <https://orcid.org/0000-0002-7648-2515>

^c <https://orcid.org/0000-0002-1371-2627>

lems, two independence conditions inherent to HMCs were put forward to justify this preference. However, these conditions are sufficient conditions for Bayesian processing, not necessary ones. Indeed, it is possible removing them considering Pairwise Markov Chains (PMCs), which extend HMCs and allow the same Bayesian processing (Pieczynski, 2003; Gorynin et al., 2018). Another example is related to NLP. HMCs are considered as generative models, and as such improper to NLP because of the fact $p(x_{1:N}|y_{1:N})$ are difficult to handle to (Jurafsky and Martin, 2021; Brants, 2000; McCallum et al., 2000). However, as recently shown in (Azeraf et al., 2020a), while defining Bayesian processing methods HMCs can also be used in discriminative way, without calling on $p(y_{1:N}|x_{1:N})$. The same is true in the case of other generative models like Naïve Bayes (Azeraf et al., 2020b).

2 LINEAR-CHAIN CRF AND HMC

2.1 Bayesian classifiers

In the Bayesian framework we consider, there is a loss function $L(x_{1:N}^*, x_{1:N})$, where $x_{1:N}$ is the true value and $x_{1:N}^*$ is the estimated one. Bayesian classifier $y_{1:N} \rightarrow \hat{x}_{1:N} = \hat{s}_B^L(y_{1:N})$ is optimal in that it minimizes the mean loss $\mathbb{E}[L(\hat{s}(Y), X)]$. It is defined with

$$\begin{aligned} \hat{x}_{1:N} &= \hat{s}_B^L(y_{1:N}) \\ &= \arg \inf_{x_{1:N}^*} \mathbb{E}[L(x_{1:N}^*, X_{1:N}) | Y_{1:N} = y_{1:N}], \end{aligned} \quad (1)$$

where $\mathbb{E}[L(x_{1:N}^*, X_{1:N}) | Y_{1:N}]$ denotes the conditional expectation. In this paper, we consider the Bayesian classifier \hat{s}_B^L corresponding to the loss function

$$L(x_{1:N}^*, x_{1:N}) = 1_{[x_1^* \neq x_1]} + \dots + 1_{[x_N^* \neq x_N]}, \quad (2)$$

which simply means that the loss is the number of wrongly classified data. Called "maximum posterior mode" (MPM), the related Bayesian classifier is defined with

$$\begin{aligned} [\hat{x}_{1:N} = (\hat{x}_1, \dots, \hat{x}_N) = \hat{s}_B^L(y_{1:N})] &\iff \\ [\forall n = 1, \dots, N, p(\hat{x}_n | y_{1:N}) = \sup_{x_n} (p(x_n | y_{1:N}))] \end{aligned} \quad (3)$$

Let us remark that Bayesian classifiers \hat{s}_B^L only depend on $p(x_{1:N}|y_{1:N})$, and are independent from $p(y_{1:N})$. In other words, for any distribution $q(y_{1:N})$, every other law of $(X_{1:N}, Y_{1:N})$ of the form $q(x_{1:N}|y_{1:N}) = p(x_{1:N}|y_{1:N})q(y_{1:N})$ gives the same Bayesian classifier \hat{s}_B^L . This shows that dividing classifiers into two categories "generative" and "discriminative" as usually done is somewhat misleading as they all are discriminative. Such a distinction is thus related to the way classifiers are defined, not to their intrinsic structure.

2.2 Equivalence between linear-chain CRF and a family of HMCs

We show in this section that for each linear-chain CRF one can find an equivalent HMC, with parameters specified from the considered CRF.

The following general Lemma will be useful in the sequence:

Lemma

Let $W_{1:N} = (W_1, \dots, W_N)$ be random sequence, taking its values in a finite set Ω . Then

- (i) $W_{1:N}$ is Markov chain iff there exist $N - 1$ functions $\varphi_1, \dots, \varphi_{N-1}$ from Ω^2 to \mathbb{R}^+ such that

$$p(w_1, \dots, w_N) \propto \varphi_1(w_1, w_2) \dots \varphi_{N-1}(w_{N-1}, w_N), \quad (4)$$

where " \propto " means "proportional to";

- (ii) for HMC defined with $\varphi_1, \dots, \varphi_{N-1}$ verifying (4), $p(w_1)$ and $p(w_{n+1}|w_n)$ are given with

$$p(w_1) = \frac{\beta_1(w_1)}{\sum_{w_1} \beta_1(w_1)}; \quad (5)$$

$$p(w_{n+1}|w_n) = \frac{\varphi_n(w_n, w_{n+1})\beta_{n+1}(w_{n+1})}{\beta_n(w_n)},$$

where $\beta_1(w_1), \dots, \beta_N(w_N)$ are defined with the following backward recursion:

$$\begin{aligned} \beta_N(w_N) &= 1, \\ \beta_n(w_n) &= \sum_{w_{n+1}} \varphi_n(w_n, w_{n+1})\beta_{n+1}(w_{n+1}) \end{aligned} \quad (6)$$

For the proof see (Lanchantin et al., 2011), Lemma 2.1, page 6.

We can state the following Proposition.

Proposition

Let $Z_{1:N} = (Z_1, \dots, Z_N)$ be stochastic sequence, with $Z_n = (X_n, Y_n)$. Each (X_n, Y_n) takes its values in $\Omega \times \Lambda$, with Ω and Λ finite. If $Z_{1:N}$ is a linear-chain conditional random field (CRF) with the distribution $p(x_{1:N}|y_{1:N})$ defined with

$$\begin{aligned} p(x_{1:N}|y_{1:N}) &= \\ &= \frac{1}{\kappa(y_{1:N})} \exp \left[\sum_{n=1}^{N-1} V_n(x_n, x_{n+1}) + \sum_{n=1}^N U_n(x_n, y_n) \right], \end{aligned} \quad (7)$$

where U_n and V_n are arbitrary "potential functions". Then (7) is the posterior distribution of the HMC

$$\begin{aligned} q(x_{1:N}, y_{1:N}) &= \\ &= q_1(x_1)q_1(y_1|x_1) \prod_{n=2}^N q_n(x_n|x_{n-1})q_n(y_n|x_n), \end{aligned} \quad (8)$$

defined as follows.

Let

$$\Psi_n(x_n) = \sum_{y_n} \exp(U(x_n, y_n)) \quad (9)$$

$$\Phi_1(x_1, x_2) = \exp(V_1(x_1, x_2))\Psi_1(x_1)\Psi_2(x_2); \quad (10)$$

and, for $n = 2, \dots, N-1$:

$$\Phi_n(x_n, x_{n+1}) = \exp(V_n(x_n, x_{n+1}))\Psi_{n+1}(x_{n+1}). \quad (11)$$

Besides, let

$$\begin{aligned} \beta_N(x_N) &= 1, \text{ and} \\ \beta_n(x_n) &= \sum_{x_{n+1}} \Phi_n(x_n, x_{n+1})\beta_{n+1}(x_{n+1}) \end{aligned} \quad (12)$$

for $n = N-1, \dots, 2$.

Then $q(x_{1:N}, y_{1:N})$ is given with

$$q(x_1) = \frac{\beta_1(x_1)}{\sum_{x_1} \beta_1(x_1)}; \quad (13)$$

$$q(x_{n+1}|x_n) = \frac{\Phi_n(x_n, x_{n+1})\beta_{n+1}(x_{n+1})}{\beta_n(x_n)} \quad (14)$$

$$q(y_n|x_n) = \frac{\exp(U(x_n, y_n))}{\Psi_n(x_n)}. \quad (15)$$

Proof

According to (9)-(15), the distribution (7) can be written:

$$p(x_{1:N}|y_{1:N}) = \frac{1}{\kappa(y_{1:N})} \prod_{n=1}^{N-1} \Phi_n(x_n, x_{n+1}) \prod_{n=1}^N q(y_n|x_n)$$

According to the Lemma, $\prod_{n=1}^{N-1} \Phi_n(x_n, x_{n+1})$ is a Markov chain defined by (13) and (14), with $\beta_n(x_n)$ defined (12), which ends the proof.

2.3 HMCs in natural language processing

Let us notice that the relationship between linear-chain CRFs and HMCs has been pointed out and discussed by some authors in the frame of natural language processing (NLP). For example, in (Sutton and McCallum, 2006) authors remark that in linear-chain CRFs it is possible to compute the posterior margins $p(x_n|y_{1:N})$ using the same forward-backward method as in HMCs. However, they keep on saying that CRFs are more general and better suited for applications in NLP. In particular, they consider that CRFs are able to model any kind of features while HMCs cannot. Similarly, in (Jurafsky and Martin, 2021), paragraph 8.5, authors recall that in general it's hard for generative

models like HMCs to add arbitrary features directly into the model in a clean way.

These arguments are no longer valid since the results presented in (Azeraf et al., 2020a). Indeed, according to the results the inability to take into account certain features is not due to the structure of HMCs, but is due to the way of calculating the *a posteriori* laws. More precisely, replacing the classic forward-backward computing by an "entropic" one allows HMCs to take into account the same features as discriminative linear-chain CRFs do. Similar kind of results concerning Naïve Bayes is specified in (Azeraf et al., 2021a).

Let us notice that HMCs defined with (8) are even more general than linear-chain CRFs defined with (7). Indeed, in the latter we have $p(x_{1:N}|y_{1:N}) > 0$, while in the former $q(x_{1:N}|y_{1:N}) \geq 0$. However, this is not a very serious advantage as one could extend (7) by removing the function exp and by considering $p(x_{1:N}|y_{1:N}) = \frac{1}{\kappa(y_{1:N})} \prod_{n=1}^{N-1} V_n(x_n, x_{n+1}) \prod_{n=1}^N U_n(x_n, y_n)$ with all $V_n(x_n, x_{n+1})$ and $U_n(x_n, y_n)$ positive or null.

3 CONCLUSION AND PERSPECTIVES

We discussed relationships between simple linear-chain CRFs and HMCs. We showed that for each linear-chain CRF, which is a family of models, one can find an HMCs giving the same posterior distribution. In addition, the related HMC's parameters can be computed from those of CRFs. In particular, joint to results in (Azeraf et al., 2020a), this shows that HMCs can be used in NLP with the same efficiency as CRFs do.

Let us mention some perspectives for further work. One recurrent argument in favour of CRFs with respect to HMCs is related to some independence properties assumed in HMCs and considered as binding. More precisely, in HMCs we have $p(y_n|x_{1:N}) = p(y_n|x_n)$ and $p(x_{n+1}|x_n, y_n) = p(x_{n+1}|x_n)$. These constraints can be removed by extending HMCs to pairwise Markov chains (PMCs) (Pieczynski, 2003; Gorynin et al., 2018; Azeraf et al., 2020b). More general than HMCs, PMCs allow strictly the same Bayesian processing. Furthermore, PMCs can be extended to triplet Markov chains (TMCs) (Boudaren et al., 2014; Gorynin et al., 2018), still allowing same Bayesian processing.

Extending HMCs considered in this paper to PMCs and TMCs should lead to extensions of recent hidden neural Markov chain (Azeraf et al., 2021b),

which is a first perspective for further works. Of course, there exist many CRFs much more sophisticated than the linear-chain CRF considered in the paper. Let us cite some recent papers (Siddiqi, 2021; Song et al., 2019; Quattoni et al., 2007; Kumar et al., 2003; Saa and Çetin, 2012), among others. Comparing different sophisticated CRFs to different PMCs and TMCs will undoubtedly be an interesting second perspective.

REFERENCES

- Azeraf, E., Monfrini, E., and Pieczynski, W. (2021a). Using the Naive Bayes as a Discriminative Model. In *Proceedings of the 13th International Conference on Machine Learning and Computing*, pages 106–110.
- Azeraf, E., Monfrini, E., Vignon, E., and Pieczynski, W. (2020a). Hidden Markov Chains, Entropic Forward-Backward, and Part-Of-Speech Tagging. *arXiv preprint arXiv:2005.10629*.
- Azeraf, E., Monfrini, E., Vignon, E., and Pieczynski, W. (2020b). Highly Fast Text Segmentation With Pairwise Markov Chains. In *Proceedings of the 6th IEEE Congress on Information Science and Technology*, pages 361–366.
- Azeraf, E., Monfrini, E., Vignon, E., and Pieczynski, W. (2021b). Introducing the Hidden Neural Markov Chain Framework. In *Proceedings of the 13th International Conference on Agents and Artificial Intelligence - Volume 2*, pages 1013–1020.
- Baum, L. E., Petrie, T., Soules, G., and Weiss, N. (1970). A Maximization Technique Occurring in the Statistical Analysis of Probabilistic Functions of Markov Chains. *The Annals of Mathematical Statistics*, 41(1):164–171.
- Boudaren, M. E. Y., Monfrini, E., Pieczynski, W., and Aissani, A. (2014). Phasic Triplet Markov Chains. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 36(11):2310–2316.
- Brants, T. (2000). TnT: a Statistical Part-of-Speech Tagger. In *Proceedings of the 6th Conference on Applied Natural Language Processing*, pages 224–231.
- Cappé, O., Moulines, E., and Ryden, T. (2005). *Inference in Hidden Markov Models*. Springer Series in Statistics, Berlin, Heidelberg.
- Fang, M., Kodamana, H., and Huang, B. (2019). Real-Time Mode Diagnosis for Processes with Multiple Operating Conditions using Switching Conditional Random Fields. *IEEE Transactions on Industrial Electronics*, 67(6):5060–5070.
- Fang, M., Kodamana, H., Huang, B., and Sammaknejad, N. (2018). A Novel Approach to Process Operating Mode Diagnosis using Conditional Random Fields in the Presence of Missing Data. *Computers & Chemical Engineering*, 111:149–163.
- Gorynin, I., Gangloff, H., Monfrini, E., and Pieczynski, W. (2018). Assessing the Segmentation Performance of Pairwise and Triplet Markov Models. *Signal Processing*, 145:183–192.
- Jurafsky, D. and Martin, J. H. (2009). *Speech and Language Processing: An Introduction to Natural Language Processing, Speech Recognition, and Computational Linguistics, 2nd Edition*. Prentice-Hall.
- Jurafsky, D. and Martin, J. H. (2021). *Speech and Language Processing, 3rd Edition Draft of December 2021*. Copyright ©. All rights reserved.
- Koski, T. (2001). *Hidden Markov Models for Bioinformatics*. Springer Netherlands.
- Kumar, S. et al. (2003). Discriminative Random Fields: A Discriminative Framework for Contextual Interaction in Classification. In *Proceedings of the 9th IEEE International Conference on Computer Vision*, pages 1150–1157.
- Lafferty, J. D., McCallum, A., and Pereira, F. C. (2001). Conditional Random Fields: Probabilistic Models for Segmenting and Labeling Sequence Data. In *Proceedings of the 18th International Conference on Machine Learning*, pages 282–289.
- Lanchantin, P., Lapuyade-Lahorgue, J., and Pieczynski, W. (2011). Unsupervised Segmentation of Randomly Switching Data Hidden with non-Gaussian Correlated Noise. *Signal Processing*, 91(2):163–175.
- McCallum, A., Freitag, D., and Pereira, F. C. N. (2000). Maximum Entropy Markov Models for Information Extraction and Segmentation. In *Proceedings of the 17th International Conference on Machine Learning*, page 591–598.
- Pieczynski, W. (2003). Pairwise Markov Chains. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 25(5):634–639.
- Quattoni, A., Wang, S., Morency, L.-P., Collins, M., and Darrell, T. (2007). Hidden Conditional Random Fields. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 29(10):1848–1852.
- Rabiner, L. R. (1989). A Tutorial on Hidden Markov Models and Selected Applications in Speech Recognition. *Proceedings of the IEEE*, 77(2):257–286.
- Saa, J. F. D. and Çetin, M. (2012). A Latent Discriminative Model-Based Approach for Classification of Imaginary Motor tasks from EEG Data. *Journal of Neural Engineering*, 9(2).
- Siddiqi, M. H. (2021). An Improved Gaussian Mixture Hidden Conditional Random Fields Model for Audio-Based Emotions Classification. *Egyptian Informatics Journal*, 22(1):45–51.
- Song, S., Zhang, N., and Huang, H. (2019). Named Entity Recognition Based on Conditional Random Fields. *Cluster Computing*, 22(3):5195–5206.
- Sutton, C. and McCallum, A. (2006). An Introduction to Conditional Random Fields for Relational Learning. *Introduction to Statistical Relational Learning*, 2:93–128.