Spectral Regularization: an Inductive Bias for Sequence Modeling

Kaiwen Hou KAIWEN.HOU@COLUMBIA.EDU
Tsinghua University

Guillaume Rabusseau GRABUS@IRO.UMONTREAL.CA
DIRO, Université de Montréal & Mila, CIFAR AI Chair

Abstract

Various forms of regularization in learning tasks strive for different notions of simplicity. This paper presents a spectral regularization technique, which attaches a unique inductive bias to sequence modeling based on an intuitive concept of simplicity defined in the Chomsky hierarchy. From fundamental connections between Hankel matrices and regular grammars, we propose to use the trace norm of the Hankel matrix, the tightest convex relaxation of its rank, as the spectral regularizer. To cope with the fact that the Hankel matrix is bi-infinite, we propose an unbiased stochastic estimator for its trace norm. Ultimately, we demonstrate experimental results on Tomita grammars, which exhibit the potential benefits of spectral regularization and validate the proposed stochastic estimator.

Keywords: Spectral Learning, Hankel Matrix, Sequence Modeling, Recurrent Neural Network

1. Introduction

The fundamental principle of *regularization* is at the heart of many machine learning algorithms and models. Informally speaking, regularization refers to the idea of adding a penalty term to the loss function optimized by a learning model in order to encourage learning *simple* functions. In particular, in regularized empirical risk minimization, the objective is to find the hypothesis h minimizing $\mathcal{L}(h, D) + \lambda\Omega(h)$ where $\mathcal{L}(h, D)$ denotes the empirical risk on a dataset D, $\Omega(h)$ is a penalty term that penalizes complex functions and λ is an hyper-parameter controlling the tradeoff between fitting the data and h being a "simple" hypothesis. Examples of regularization functions Ω include the ℓ_2 or ℓ_1 norm of the weight vector of a linear model, the degree of a polynomial model, the rank or the trace norm of the user-item matrix in a collaborative filtering task, etc.

In this work, we propose a novel regularization technique for sequential models. While there are many natural notions of simplicity for functions defined over vector spaces (e.g., sparsity, smoothness, etc.), defining a notion of simplicity suited for functions defined over sequences can be more tedious due to the discrete and sequential nature of the data arising in tasks such as language modelling. One such notion of simplicity naturally arises from the so-called Chomsky hierarchy, which categorizes functions over sequences into four different levels of complexity, the simplest of which is regular grammars (Chomsky, 1956). To encourage a sequential model such as an RNN (Recurrent Neural Network) to learn simple functions, i.e., functions that appear lower in Chomsky hierarchy, we introduce a novel inductive bias through *spectral regularization*.

In order to encourage learning such simple functions, we leverage a fundamental result relating the rank of the Hankel matrix of a function f to the minimal number of states of a weighted finite automaton computing f (Fliess, 1974; Carlyle and Paz, 1971). Functions computed by weighted finite automata corresponds to regular weighted languages, i.e., functions that are low in the Chomsky hierarchy. The idea behind spectral regularization is to encourage learning models whose Hankel matrix are approximately low rank. To do so, the spectral regularization is defined as the trace norm of the Hankel matrix. Using the trace norm instead of the rank of the Hankel matrix offers two advantages: (i) the trace norm is the tightest convex relaxation of the rank and is differentiable, allowing one to use automatic differentiation techniques to use the spectral regularization when training black box neural network sequence models, and (ii) the trace norm can be seen as a "soft" version of the rank, allowing learned models to only be *approximately* low rank, whereas a hard rank constraint would be too strong and forces the learned functions to be regular. The spectral regularization can thus incorporate a natural inductive bias towards regular functions in the training of any black box differentiable model.

A key technical challenge in implementing the proposed spectral regularization resides in the fact the Hankel matrix is a bi-infinite matrix whose trace norm cannot be explicitly computed. To address this issue we propose a Russian Roulette estimator to design a stochastic unbiased estimator of the Hankel matrix, whose trace norm is lower bounded (in expectation) by the trace norm of the Hankel matrix itself. We thus plug in the realizations of the Russian Roulette estimator in the minimization objective at each mini-batch in place of the actual trace norm of the Hankel matrix.

We provide a simple experimental study on Tomita grammars (Tomita, 1982) to illustrate the potential benefits of the spectral regularization.

2. Preliminaries

Let Σ be a finite nonempty set, also known as an alphabet. We denote by Σ^* the free monoid over Σ, where string concatenation is the binary operation and the empty string in the singleton set $\Sigma^0 := \{\epsilon\}$ serves as the unique unit element. Intuitively, Σ^* refers to the set of all finite sequences (or words) generated by Σ:

$$\Sigma^* := \bigcup_{n=0}^{\infty} \Sigma^n.$$

For two sequences $u, v \in \Sigma^*$, we use uv to denote the concatenation of u and v. The length of a sequence $w \in \Sigma^*$ is denoted as $|w|$. Finally, a grammar, or language, over Σ is a subset of Σ^*.

One of the simplest class of languages is the set of *regular languages*, which are languages that can be computed by deterministic finite automata. Regular languages forms the simplest class of languages in the so-called Chomsky hierarchy . In this work, we are interested in real-valued functions over Σ^*, sometimes called weighted languages. Such functions are of crucial interest for machine learning applications on sequence data such as language modeling. The Chomsky hierarchy easily extends to weighted languages using the weighted counterparts of the finite state machines used in the classical hierarchy. In particular, the simplest class of such functions is the set of regular functions (sometimes

called rational, or recognizable), which are functions that can be computed by weighted automata.

Definition 1 (Weighted Finite Automaton) *A weighted finite automaton (WFA) with n states is a tuple $\mathcal{A} = (\alpha, \{\boldsymbol{A}^\sigma\}_{\sigma\in\Sigma}, \omega)$, where $\alpha \in \mathbb{R}^n$ is the initial weight vector, $\omega \in \mathbb{R}^n$ the final weight vector, and $\boldsymbol{A}^\sigma \in \mathbb{R}^{n\times n}$ is the transition matrix for each symbol $\sigma \in \Sigma$. A WFA computes a function $f_\mathcal{A} : \Sigma^* \to \mathbb{R}$ that maps any sequence $\mathrm{u} = \mathrm{u}_1\mathrm{u}_2\cdots\mathrm{u}_k \in \Sigma^*$ to $f_\mathcal{A}(\mathrm{u}) = \alpha^T \boldsymbol{A}^\mathrm{u}\omega$, where $\boldsymbol{A}^\mathrm{u} := \boldsymbol{A}^{\mathrm{u}_1}\boldsymbol{A}^{\mathrm{u}_2}\cdots\boldsymbol{A}^{\mathrm{u}_k}$.*

It is worth briefly mentioning that any regular language is the support of a rational function (however, surprisingly, the converse is not true, see, e.g., Chapter 4, Section 6 in Droste et al. (2009)).

In this work, we will design a regularization scheme for sequential models that will favour functions that are close to the class of rational functions (i.e., low on Chomsky's hierarchy). In order to do so, we need a quantitative measure of the "rationality" of a function. We will see that the spectrum of the so-called Hankel matrix is a good candidate for this purpose.

Definition 2 (Hankel Matrix) *For a given function $f : \Sigma^* \to \mathbb{R}$, its Hankel matrix $\boldsymbol{H}_f \in \mathbb{R}^{\Sigma^*\times\Sigma^*}$ is the infinite matrix with entries $(\boldsymbol{H}_f)_{\mathrm{u},\mathrm{v}} = f(\mathrm{uv})$ for $\mathrm{u}, \mathrm{v} \in \Sigma^*$.*

The following classical theorem shows the fundamental (striking) relation between the Hankel matrix of a function and its "rationality".

Theorem 3 (Fliess (1974); Carlyle and Paz (1971)) *For any function $f : \Sigma^* \to \mathbb{R}$, $rank(\boldsymbol{H}_f)$ is equal to the minimal number of states of a WFA computing f. In particular, a function f is regular if and only if its Hankel matrix has finite rank.*

We will see in the next section how this result can be leveraged to design a regularization technique to favour simpler model during learning.

3. Spectral Regularization

In this section we propose the trace norm of the Hankel matrix as a natural spectral regularization for black box sequential models and show how to efficiently compute stochastic approximation of this regularization term for training through back-propagation.

3.1. Motivation and Definition

A naive idea to leverage Theorem 3 for regularization would be to enforce the Hankel matrix of the learned model to be low rank. However, this approach has two drawbacks. First, optimization under low rank constraints is known to be computationally hard. Second, such a constraint would be too strong: we want to incorporate an inductive bias towards simple functions in the learning process, but we do not want to actually enforce the learned function to be regular. In some sense, we want a softer version of the rank of the Hankel matrix which would also consider functions that can be well approximated by regular functions (i.e., functions whose Hankel matrix is approximately low rank) as simple.

Enforcing the trace norm (or nuclear norm) of the Hankel matrix to be small, instead of directly enforcing the rank to be small, will solve (to some extent) both of these issues. Indeed, the trace norm (which is the sum of the singular values) is the tightest convex relaxation of the matrix rank (Fazel et al., 2001) which naturally represents a soft version of the notion of rank. The trace norm of the Hankel matrix has actually been previously leveraged for this purpose in the context of learning (Balle et al., 2012). Using the trace norm of the Hankel matrix as a way to regularize models for sequence tagging was also previously explored in (Quattoni et al., 2014).

We formally introduce this regularization technique in the following definition.

Definition 4 (Spectral Regularization) *Let $f_\theta : \Sigma^* \to \mathbb{R}$ be the function computed by a model with parameters θ and let $\tilde{\mathcal{L}}(\theta)$ be the loss function associated with this model. Spectral regularization corresponds to the following minimization problem:*

$$\min_\theta \tilde{\mathcal{L}}(\theta) + \lambda \|\boldsymbol{H}_{f_\theta}\|_*, \tag{1}$$

where λ is the regularization coefficient, and the trace norm $\|\boldsymbol{H}_{f_\theta}\|_*$ is the spectral regularizer (or spectral loss).

Note that the previous definition does not make any assumptions on the class of models considered. One particular class of interest is the one of functions computed by recurrent neural networks, for which we would ideally want to, at the same time, benefit from their remarkable expressiveness while still steering the learning process towards functions that are, in some sense, low on the Chomsky hierarchy. In particular, when an RNN is used for sequential probabilistic modeling (i.e. trained to predict the probabilities of next symbol given a sequence), f_θ would denote the underlying probability distribution over Σ^*, i.e., $f_\theta(\mathrm{u}_1\mathrm{u}_2\cdots\mathrm{u}_k) = P(\mathrm{u}_1\mathrm{u}_2\cdots\mathrm{u}_k) = P(\mathrm{u}_1)P(\mathrm{u}_2 \mid \mathrm{u}_1)\cdots P(\mathrm{u}_k \mid \mathrm{u}_1\cdots\mathrm{u}_{k-1})$.

3.2. Russian Roulette Estimator

It is clear that the optimization problem in Eq. (1) can not be solved easily. To start with, the Hankel matrix is infinite! In order to tackle this optimization problem, we will make use of the so-called Russian Roulette estimator which allows one to stochastically approximate an infinite series with random realization of partial sums.

Definition 5 (Russian Roulette Estimator; Kahn (1955)) *Given a convergent series $S = \sum_{k=0}^{\infty} \alpha_k$, a Russian Roulette estimator of S is given by $\hat{S} := \sum_{i=0}^{\tau} \frac{\alpha_i}{P(\tau \geq i)}$, where $\tau \geq 0$ is a random variable with support over all nonnegative integers.*

Note that in this definition we do not require α_k's to be scalars. Instead, they could stand for vectors, matrices, tensors, or some abstract objects with well-defined component-wise addition.

Theorem 6 (Chen et al. (2019); Lemma 3; Lyne et al. (2015)) *If $P(\tau \geq n) > 0 \ \forall n > 0$ and the series S is absolutely convergent, then $\hat{S}$ given in Definition 5 is an unbiased estimator of S, i.e., $\mathbb{E}\hat{S} = S$.*

Although the Russian Roulette Estimator is unbiased under mild assumptions, its variance might be large or even unbounded with an ill-chosen random variable τ (McLeish, 2011; Beatson and Adams, 2019).

3.3. Stochastic Estimator for the Trace Norm of the Hankel Matrix

In order to leverage the Russian Roulette estimator for the trace norm of the Hankel matrix, we need to express the Hankel matrix as an infinite sum. We propose one way convenient way to do this in the following theorem.

Theorem 7 *Let $f : \Sigma^* \to \mathbb{R}$. For any $i \in \mathbb{N}$, let $\boldsymbol{H}_f^{(i)} \in \mathbb{R}^{\Sigma^* \times \Sigma^*}$ be defined by*

$$(\boldsymbol{H}_f^{(i)})_{\mathrm{u,v}} = \begin{cases} f(\mathrm{uv}) & \text{if } |\mathrm{uv}| = i \\ 0 & \text{otherwise} \end{cases}$$

for all $\mathrm{u}, \mathrm{v} \in \Sigma^*$. Then $\boldsymbol{H}_f = \sum_{i=0}^{\infty} \boldsymbol{H}_f^{(i)}$.

Although the $\boldsymbol{H}_f^{(i)}$'s defined above are infinite matrices, each of them only contains a finite number of nonzero elements. We can thus construct the Russian Roulette estimator of $\boldsymbol{H}_f$ as

$$\boldsymbol{H}_\tau = \sum_{i=0}^{\tau} \frac{\boldsymbol{H}_f^{(i)}}{P(\tau \geq i+1)} \tag{2}$$

where τ is a random variable taking its values in $\mathbb{N}$ such that $P(\tau \geq n) > 0$ for all n.

As mentioned previously, even though $\boldsymbol{H}_\tau$ still is an infinite matrix, it only has a finite number on non-zeros entries for any integer τ. Thus, informally, the trace norm of the infinite matrix $\boldsymbol{H}_\tau$ is equal to the trace norm of its smallest sub-block containing no columns or rows entirely filled with 0's, which is a finite sub-block whose trace norm can be computed in polynomial time. We now formalize this intuition. We start by showing that the Russian Roulette estimator of the Hankel matrix is unbiased.

Theorem 8 *Let $f : \Sigma^* \to \mathbb{R}$. The estimator $\boldsymbol{H}_\tau$ defined in Eq. (2) is an unbiased estimator of $\boldsymbol{H}_f$, i.e., $\mathbb{E}_\tau[\boldsymbol{H}_\tau] = \boldsymbol{H}_f$.*

Proof:
For some $\mathrm{u}, \mathrm{v} \in \Sigma^*$, we notice from Eq. (2) that

$$(\boldsymbol{H}_\tau)_{\mathrm{u,v}} = \begin{cases} \frac{(\boldsymbol{H}_f^{(i)})_{\mathrm{u,v}}}{P(\tau \geq i+1)} & \text{if } \tau \geq k+1 \\ 0 & \text{otherwise} \end{cases}$$

for some $k \in \mathbb{N}$, since the RHS of Eq. (2) contributes at most one term for an entry in LHS. Then

$$\mathbb{E}[(\boldsymbol{H}_\tau)_{\mathrm{u,v}}] = \mathbb{E}[\frac{(\boldsymbol{H}_f^{(i)})_{\mathrm{u,v}}}{P(\tau \geq i+1)} \mathbf{1}_{\tau \geq \mathrm{k}+1}] = \frac{(\boldsymbol{H}_f^{(i)})_{\mathrm{u,v}}}{P(\tau \geq k+1)} P(\tau \geq k+1) = (\boldsymbol{H}_f^{(i)})_{\mathrm{u,v}}.$$

Therefore, each entry of $\boldsymbol{H}_\tau$ is unbiased to estimate the corresponding entry in $\boldsymbol{H}_f$.

■

We showed that the infinite Hankel matrix of a function can be expressed as an infinite sum of matrices with a finite number of non-zero entries, allowing us to construct a Russian Roulette estimator of the Hankel matrix which can be computed efficiently. But we are interested in the trace norm of the Hankel matrix in the objective we wish to minimize in Eq. (1). It remains to show that the trace norm of the Russian Roulette estimator of the Hankel matrix is a good stochastic estimator of the trace norm of the Hankel matrix itself.

Theorem 9 *For any θ, we have that*

$$\tilde{\mathcal{L}}(\theta) + \lambda ||\boldsymbol{H}_{f_\theta}||_* \leq \tilde{\mathcal{L}}(\theta) + \lambda \mathbb{E}[||\boldsymbol{H}_\tau||_*], \tag{3}$$

where $\boldsymbol{H}_\tau$ is the Russian Roulette estimator defined in Eq. (2).

Proof:
It suffices to show that $||\cdot||_*$ is a convex operator, and then the claim follows by Jensen's inequality. The convexity of $||\cdot||_*$ follows naturally from the triangle inequality of the trace norm $||\mu \boldsymbol{H}_1 + (1-\mu)\boldsymbol{H}_2||_* \leq \mu ||\boldsymbol{H}_1||_* + (1-\mu)||\boldsymbol{H}_2||_*$ for any $\mu \in [0,1]$. Combining Jensen's inequality with Theorem 8 we have that $\mathbb{E}[||\boldsymbol{H}_\tau||_*] \geq ||\mathbb{E}[\boldsymbol{H}_\tau]||_* = ||\boldsymbol{H}_{f_\theta}||_*$.

■

The previous theorem shows that we can efficiently compute a stochastic approximation of the trace norm of the Hankel matrix, through which we can use the back-propagation algorithm to train any differentiable black-box model. In the next section, we implement this regularization technique to train RNNs on a synthetic language modeling task.

4. Experiments

We conduct experiments to validate that spectral regularization imposes an inductive bias for sequence modeling. In particular, we focus on synthetic data generated according to Tomita grammars #3 to #6 defined in Table 1, which is a benchmark study for grammatical inference (Tomita, 1982; Bengio and Frasconi, 1994). As shown in the table, all these grammars are some subsets of Σ^*, where the binary alphabet is $\Sigma := \{0,1\}$.

Tomita Grammars	Definitions
#3	not containing $1^{2n+1}0^{2m+1}$ as a substring
#4	not containing 000 as a substring
#5	containing even number of 01's and 10's
#6	(number of 0's − number of 1's) is a multiple of 3

Table 1: Definitions of Tomita grammars #3 to #6.

The training dataset for each grammar include synthetic sequences up to length 12 in that grammar. 20% of the training set is split out as the validation set. We also use a test dataset consisting of sequences of exact length 12 and disjoint with the training dataset.

We consider an RNN with one embedding layer of $|\Sigma|+2$ neurons (we use two additional symbols to mark the start and end of sequences) and one hidden layer of 50 neurons, which

has been just expressive enough for our training data. NLL (Negative Log-Likelihood) loss is used for both training and reporting performances of grammatical inference on the three test sets. The loss minimization is based on the *Adam* optimizer (Diederik et al., 2014) with an initial learning rate of 0.01 and a batch size of 32. Moreover, early stopping and a simple scheduler to reduce learning rate on detected plateaus of validation loss are adopted.

In our experiments, we compare the test NLL when training without spectral regularization versus that with spectral regularization for different size of training data sampled from the given training set. The latter chooses the hyperparameter λ according to validation NLL. In each mini-batch, we randomly draw $\tau \sim \text{Geometric}(0.2)$ (stopping probability is 0.2) to construct the Russian Roulette estimator of the Hankel matrix. To check the significance of the unbiased Russian Roulette estimator in spectral regularization, we also implement a naïve biased estimator of the Hankel matrix for comparison defined by $\hat{\boldsymbol{H}} = \sum_{i=0}^{10} \boldsymbol{H}_f^{(i)}$, which is a fixed-sized subblock of the Hankel matrix.

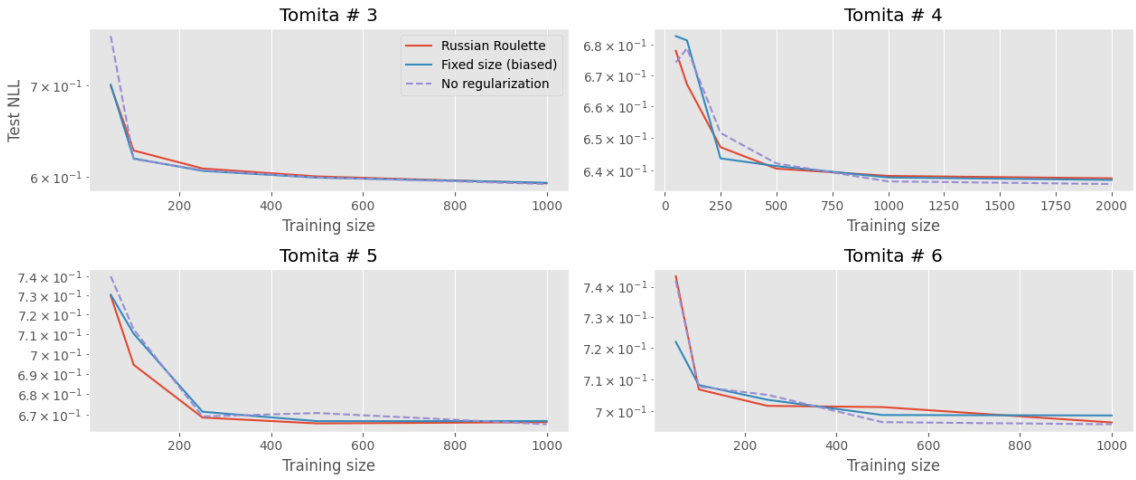

Figure 1: Test NLL against different training sizes for training with Russian Roulette estimator, with the fixed-sized (naïve biased) estimator, and without spectral regularization on Tomita grammars #3 to #6.

Results on Tomita grammars #3 to #6 are presented in Figure 1, where we see that spectral regularization marginally improves generalization for Tomita grammars #4, #5, and #6 on small training data sizes. Especially on Tomita grammar #5, the unbiased Russian Roulette estimator performs modestly better than the naïve biased one. However, there is no clear winner between the two estimators, biased or not, on other Tomita grammars. We hypothesize that this phenomenon might be due to the bias-variance tradeoff, i.e., the high variance of the unbiased Russian Roulette estimator makes the loss computation much coarser (Beatson and Adams, 2019), which will be further investigated in future work. We also consider whether more convincing results can be obtained on other tasks such as classification, or on other datasets, to be explored in upcoming studies.

5. Conclusion

This paper proposes spectral regularization according to an intuitive notion of simplicity arising from the Chomsky hierarchy, which serves as an extra inductive bias for any sequence modeling task and is formulated as an additional regularization term to be added to any loss function. Results on synthetic data of Tomita grammars show that spectral regularization indeed marginally helps encourage the model to learn approximately low-rank functions. Forthcoming research will also examine the effect of spectral regularization in other tasks and other datasets.

To estimate the trace norm of the bi-infinite Hankel matrix in the spectral regularizer, we construct an unbiased stochastic estimator to relax the loss minimization problem. However, the unbiased estimator does not exhibit significant advantages compared to a naïve biased one, which will be explored in further research.

Acknowledgments

We would like to acknowledge the support of the 2021 Globalink Research Internship Mitacs program (Project ID 24986).

References

Borja Balle, Ariadna Quattoni, and Xavier Carreras. Local loss optimization in operator models: a new insight into spectral learning. In *Proceedings of the 29th International Coference on International Conference on Machine Learning*, pages 1819–1826, 2012.

Alex Beatson and Ryan P. Adams. Efficient optimization of loops and limits with randomized telescoping sums. In *Proceedings of the 36th International Conference on Machine Learning, ICML 2019, 9-15 June 2019, Long Beach, California, USA*, pages 534–543, 2019.

Yoshua Bengio and Paolo Frasconi. An em approach to learning sequential behavior. *Advances in Neural Information Processing Systems*, 7, 1994.

Jack W. Carlyle and Azaria Paz. Realizations by stochastic finite automata. *Journal of Computer and System Sciences*, 5(1):26–40, 1971.

Ricky T. Q. Chen, Jens Behrmann, David K Duvenaud, and Joern-Henrik Jacobsen. Residual flows for invertible generative modeling. In *Advances in Neural Information Processing Systems*, volume 32, 2019.

N. Chomsky. Three models for the description of language. *IRE Transactions on Information Theory*, 2(3):113–124, 1956.

Kingma Diederik, Ba Jimmy, et al. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, pages 273–297, 2014.

Manfred Droste, Werner Kuich, and Heiko Vogler. *Handbook of weighted automata*. Springer Science & Business Media, 2009.

Maryam Fazel, Haitham Hindi, and Stephen P Boyd. A rank minimization heuristic with application to minimum order system approximation. In *Proceedings of the 2001 American Control Conference.(Cat. No. 01CH37148)*, volume 6, pages 4734–4739. IEEE, 2001.

Michel Fliess. Matrices de hankel. *J. Math. Pures Appl*, 53(9):197–222, 1974.

Herman Kahn. *Use of Different Monte Carlo Sampling Techniques*. RAND Corporation, Santa Monica, CA, 1955.

Anne-Marie Lyne, Mark Girolami, Yves Atchadé, Heiko Strathmann, and Daniel Simpson. On russian roulette estimates for bayesian inference with doubly-intractable likelihoods. *Statistical science*, 30(4):443–467, 2015.

Don McLeish. A general method for debiasing a monte carlo estimator. *Monte Carlo Methods Appl.*, 17(4):301–315, 2011.

Ariadna Quattoni, Borja Balle, Xavier Carreras, and Amir Globerson. Spectral regularization for max-margin sequence tagging. In *International Conference on Machine Learning*, pages 1710–1718. PMLR, 2014.

Masaru Tomita. Dynamic construction of finite-state automata from examples using hill-climbing. In *Proceedings of the Fourth Annual Conference of the Cognitive Science Society*, pages 105–108, 1982.