

Contents

Abstract (English/Français)	i
Acknowledgements	v
List of figures	xi
List of tables	xiii
List of algorithms	xv
	1
1 Introduction	1
1.1 Motivations	1
1.1.1 Learning from small data with large number of classes	1
1.1.2 Learning to embed sequences	2
1.1.3 Efficient estimation of an empirical expectation	3
1.2 Summary of contributions	3
1.3 Thesis outline	4
1.4 Notation	4
2 Background on Machine Learning	5
2.1 Introduction	6
2.2 Empirical Risk Minimization (ERM) principle	7
2.2.1 Bias-variance trade-off	8
2.3 Perceptron learning algorithm	9
2.3.1 Remarks	11
2.4 Stochastic gradient descent (SGD)	11
2.5 Representer theorem and non-linear learning algorithms	12
2.6 Kernel learning and neural networks	15
2.6.1 Random feature map for Gaussian RBF kernel	16
2.6.2 Neural networks	17
2.7 Discussion	19
3 Weighted Approximate Rank Component Analysis	21

3.1	Introduction	22
3.2	Related work	23
3.3	Weighted Approximate Rank Component Analysis (WARCA)	26
3.3.1	Problem formulation	26
3.3.2	Approximate OrthoNormal (AON) regularizer	28
3.3.3	Max-margin reformulation	28
3.3.4	WARCA in kernel space	30
3.4	Experiments	32
3.4.1	Data-sets and baselines	32
3.4.2	Technical details	33
3.4.3	Comparison against state-of-the-art	37
3.4.4	Analysis of the AON regularizer	37
3.4.5	Analysis of the training time	39
3.5	Discussion	39
4	Kronecker Recurrent Units	41
4.1	Introduction	42
4.2	Recurrent neural network formalism	44
4.2.1	Over parametrization and computational efficiency	45
4.2.2	Poor conditioning implies gradients explode or vanish	45
4.2.3	Why complex field?	45
4.3	Kronecker recurrent units (KRU)	46
4.3.1	Soft unitary constraint	47
4.4	Experiments	47
4.4.1	Copy memory problem	47
4.4.2	Adding problem	49
4.4.3	Pixel by pixel MNIST	51
4.4.4	Character level language modelling on Penn TreeBank (PTB)	52
4.4.5	Polyphonic music modeling	53
4.4.6	Frameworkise phoneme classification on TIMIT	54
4.4.7	Influence of soft unitary constraints	55
4.5	Discussion	56
5	Importance Sampling Tree	59
5.1	Introduction	60
5.2	Related work	61
5.3	Weighted averages in machine learning	63
5.3.1	Importance sampling for Monte-carlo simulations	64
5.4	Importance Sampling Tree (IST)	64
5.4.1	Adaptive sampling	65
5.5	Experiments and results	66
5.5.1	Multi-layer Neural Network on a 2D synthetic data-sets	66
5.5.2	Deep Convolution Network on CIFAR10	68

5.6	Discussion	70
6	Conclusions	73
6.1	Summary	73
6.2	Future directions	74
A	Chapter 2 Appendix	77
A.1	Perceptron convergence proof	77
B	Chapter 3 Appendix	79
B.1	Metric learning	79
B.1.1	Principal component analysis	80
B.1.2	Fisher discriminant analysis	80
B.1.3	Information-theoretic metric learning (Davis et al., 2007)	81
B.1.4	KISS metric learning (Köstinger et al., 2012)	82
B.1.5	Siamese neural network (Chopra et al., 2005)	82
B.1.6	Chopping (Fleuret and Blanchard, 2005)	83
B.2	Person re-identification	83
B.2.1	Performance measures for person re-identification	84
B.3	Feature space visualization for WARCA	84
B.4	Maximizing AUC with a Mahalanobis metric	87
B.4.1	Kernelization	88
B.4.2	Optimization	89
B.4.3	Experiments	91
C	Chapter 4 Appendix	95
C.1	Analysis of vanishing and exploding gradients in RNN	95
C.2	Long short-term memory (LSTM) (Hochreiter and Schmidhuber, 1997)	96
C.3	Unitary evolution RNN (Arjovsky et al., 2016)	97
C.4	Full capacity unitary RNN (Wisdom et al., 2016)	97
C.5	Orthogonal RNN (Mhammedi et al., 2016)	98
C.6	Properties of Kronecker matrix (Van Loan, 2000)	98
C.7	Product between a dense matrix and a Kronecker matrix	99
C.8	Gradient computation in a Kronecker layer	102
Bibliography		105
Curriculum Vitae		116