Graphical Models: Modeling, Optimization, and Hilbert Space Embedding

Xinhua Zhang

A thesis submitted for the degree of
Doctor of Philosophy at
The Australian National University

March 2010
Declaration

Except where otherwise indicated, this thesis is my own original work.

Xinhua Zhang
March 2010

The following table gives the collaborators of each chapter, and the publications.

<table>
<thead>
<tr>
<th>Chapter</th>
<th>Collaborators</th>
<th>Publication</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Douglas Aberdeen and SVN Vishwanthan</td>
<td>ICML 2007, (Zhang et al., 2007)</td>
</tr>
<tr>
<td>2</td>
<td>Thore Graepel and Ralf Herbrich</td>
<td>AISTATS 2010, (Zhang et al., 2010)</td>
</tr>
<tr>
<td>3</td>
<td>Alex Smola, Le Song, and Arthur Gretton</td>
<td>NIPS 2008, (Zhang et al., 2009)</td>
</tr>
<tr>
<td>4</td>
<td>SVN Vishwanthan and Ankan Saha</td>
<td>Submitted to COLT 2010</td>
</tr>
</tbody>
</table>

Two other papers published but not included in this thesis are (Song, Zhang, Smola, Gretton, & Schölkopf, 2008b) and (Cheng, Vishwanathan, & Zhang, 2008).
To my parents for their love and support.
Acknowledgements

It is finally the time to summarize my four years’ PhD research in a single coherent document, and I would like to take this opportunity to thank all the people who helped me along the way.

First of all, I wish to express my utmost gratitude to S. V. N. Vishwanthan, known to all as Vishy, for his long time guidance in my research and life. Being energetic and enthusiastic with research, Vishy has given me most valuable hand-on guidance in both theoretical and practical problems. I still remember those long discussions into late night, both in Vishy’s office and at his home. After becoming a faculty at Purdue University, Vishy kindly offered me a visiting scholar position and I considerably benefited from this visit at Purdue.

My gratitude goes to Alex Smola, who showed me how to perform creative research and impressed me with his novel ideas. Despite the pain of following and digesting his thoughts, it is almost always fruitful to explore in the directions Alex suggested and this high level supervision greatly helped improve my ability to conduct independent research.

Thank you to Douglas Aberdeen for introducing reinforcement learning to me, and guiding me through the project of conditional random fields for reinforcement learning. It was really a pity that Doug left NICTA after the project grew into a good shape.

Many thanks to Microsoft Research Cambridge (MSRC), who offered me a 12-week internship from September 2008. My mentors Thore Graepel and Ralf Herbrich are especially creative, knowledgeable and friendly to help me in both theory and practice.

Thank you to Wray Buntine for teaching me nonparametric Bayesian, and chairing my supervisory panel after Alex and Vishy left NICTA.

My gratitude goes to the CSL/RSISE/CECS administration staff, especially Michelle Moravec, Di Kossatz, and Deb Pioch for creating a friendly and efficient environment for learning and research, keeping me on track, and helping me focus on research. Also thank you to Norma Lucas at Purdue University for all the administrative help.

I have also greatly benefited from discussions with other researchers, students and visitors of NICTA, of the Computer Sciences Lab of ANU, and of the Department of Statistics at Purdue University. The following list is absolutely not complete: Marconi Barbosa, Justin Bedo, Olivier Buffet, Tiberio Caetano, Dmitry Kamenetsky, Li Cheng, Nan Ding, Markus Hegland, Jiayuan Huang, Knut Hiiper, Marcus Hutter, Quoc Viet...
Le, Novi Quadrianto, Mark Reed, Scott Sanner, Nic Schraudolph, Hao Shen, Javen (Qinfeng) Shi, Le Song, Owen Thomas, Choon Hui Teo, Tao Wang, Chris Webers, and Jin Yu.

I gratefully acknowledge the scholarship from NICTA, fellowship from Microsoft Research Asia, Vice-Chancellor’s grant from ANU for attending conferences and visiting Purdue University, and travel grants from Pascal, Pascal 2, ICML, and NIPS.

Xinhua Zhang

March 2010
Abstract

Over the past two decades graphical models have been widely used as powerful tools for compactly representing distributions. On the other hand, kernel methods have been used extensively to come up with rich representations. This thesis aims to combine graphical models with kernels to produce compact models with rich representational abilities.

Graphical models are a powerful underlying formalism in machine learning. Their graph theoretic properties provide both an intuitive modular interface to model the interacting factors, and a data structure facilitating efficient learning and inference. The probabilistic nature ensures the global consistency of the whole framework, and allows convenient interface of models to data.

Kernel methods, on the other hand, provide an effective means of representing rich classes of features for general objects, and at the same time allow efficient search for the optimal model. Recently, kernels have been used to characterize distributions by embedding them into high dimensional feature space. Interestingly, graphical models again decompose this characterization and lead to novel and direct ways of comparing distributions based on samples.

Among the many uses of graphical models and kernels, this thesis is devoted to the following four areas:

**Conditional random fields for multi-agent reinforcement learning**  Conditional random fields (CRFs) are graphical models for modeling the probability of labels given the observations. They have traditionally been trained with using a set of observation and label pairs. Underlying all CRFs is the assumption that, conditioned on the training data, the label sequences of different training examples are independent and identically distributed (\( \text{iid} \)). We extended the use of CRFs to a class of temporal learning algorithms, namely policy gradient reinforcement learning (RL). Now the labels are no longer \( \text{iid} \). They are actions that update the environment and affect the next observation. From an RL point of view, CRFs provide a natural way to model joint actions in a decentralized Markov decision process. They define how agents can communicate with each other to choose the optimal joint action. We tested our framework on a synthetic network alignment problem, a distributed sensor network, and a road traffic control system. Using tree sampling by Hamze & de Freitas (2004) for inference, the RL methods employing CRFs clearly outperform those which do not
model the proper joint policy.

**Bayesian online multi-label classification** Gaussian density filtering (GDF) provides fast and effective inference for graphical models (Maybeck, 1982). Based on this natural online learner, we propose a Bayesian online multi-label classification (BOMC) framework which learns a probabilistic model of the linear classifier. The training labels are incorporated to update the posterior of the classifiers via a graphical model similar to TrueSkill (Herbrich et al., 2007), and inference is based on GDF with expectation propagation. Using samples from the posterior, we label the test data by maximizing the expected F-score. Our experiments on Reuters1-v2 dataset show that BOMC delivers significantly higher macro-averaged F-score than the state-of-the-art online maximum margin learners such as LaSVM (Bordes et al., 2005) and passive-aggressive online learning (Crammer et al., 2006). The online nature of BOMC also allows us to efficiently use a large amount of training data.

**Hilbert space embedment of distributions** Graphical models are also an essential tool in kernel measures of independence for non-iid data. Traditional information theory often requires density estimation, which makes it unideal for statistical estimation. Motivated by the fact that distributions often appear in machine learning via expectations, we can characterize the distance between distributions in terms of distances between means, especially means in reproducing kernel Hilbert spaces which are called kernel *embedment*. Under this framework, the undirected graphical models further allow us to factorize the kernel embedment onto cliques, which yields efficient measures of independence for non-iid data (Zhang et al., 2009). We show the effectiveness of this framework for ICA and sequence segmentation, and a number of further applications and research questions are identified.

**Optimization in maximum margin models for structured data** Maximum margin estimation for structured data, e.g. (Taskar et al., 2004), is an important task in machine learning where graphical models also play a key role. They are special cases of regularized risk minimization, for which bundle methods (BMRM, Teo et al., 2007) and the closely related SVMStruct (Tschantzaridis et al., 2005) are state-of-the-art general purpose solvers. Smola et al. (2007b) proved that BMRM requires $O(1/\epsilon)$ iterations to converge to an $\epsilon$ accurate solution, and we further show that this rate hits the lower bound. By utilizing the structure of the objective function, we devised an algorithm for the structured loss which converges to an $\epsilon$ accurate solution in $O(1/\sqrt{\epsilon})$ iterations. This algorithm originates from Nesterov’s optimal first order methods (Nesterov, 2003, 2005b).
# Contents

Acknowledgements v

Abstract vii

List of Symbols xix

1 Introduction 1
   1.1 Exponential families ............................ 3
   1.2 Graphical models and factorization .................. 6
      1.2.1 Markov random fields .......................... 6
   1.3 Conditional random fields .......................... 8
      1.3.1 Factorization of conditional distributions .......... 9
   1.4 Reproducing kernel Hilbert spaces for exponential families .... 10
      1.4.1 Positive semi-definite kernels .................. 11
      1.4.2 Reproducing kernel Hilbert spaces ................. 13
      1.4.3 Kernel exponential families and decomposition ....... 15
   1.5 Learning and inference ............................ 17
      1.5.1 Exact methods .................................. 19
      1.5.2 Message passing ............................... 20
      1.5.3 Sampling ...................................... 23
      1.5.4 Variational inference ........................... 23
   1.6 Regularized risk estimation and optimizations .............. 27
      1.6.1 Regularized risk minimization .................... 28
      1.6.2 Survey of existing optimization algorithms ......... 32
      1.6.3 Cutting plane ................................... 35
      1.6.4 Bundle methods for regularized risk minimization .... 37
   1.7 Outline ........................................... 39

2 Conditional Random Fields for Multi-agent Reinforcement Learning 41
   2.1 Conditional random fields and inference ................. 42
      2.1.1 Conditional exponential families .................. 43
      2.1.2 Inference and gradient computations ............... 43
      2.1.3 Tree MCMC sampler for CRFs ..................... 44
## Contents

2.2 Reinforcement learning ........................................... 46  
   2.2.1 Policy-gradient algorithms .................................. 47  
   2.2.2 Decentralized multi-agent RL ............................... 48  
2.3 Conditional random fields for RL .............................. 49  
2.4 Experimental results ........................................... 50  
   2.4.1 Grid alignment ............................................... 50  
   2.4.2 Sensor networks ........................................... 52  
   2.4.3 Traffic light control ....................................... 54  
2.5 Conclusions ..................................................... 55

3 Bayesian Online Learning for Multi-label and Multi-variate Measures 57  
   3.1 A Bayesian model for multi-label classification .......... 60  
      3.1.1 Multi-class classification .................................. 60  
      3.1.2 Multi-label classification ................................. 63  
   3.2 Online learning and inference ................................. 65  
      3.2.1 A Bayesian view of learning ............................... 65  
      3.2.2 Inference on the graph of a given training example with EP ..... 66  
      3.2.3 Dynamic models .......................................... 67  
   3.3 Generalization for multi-variate performance measure ... 68  
      3.3.1 Formulation of expected F-score ......................... 69  
      3.3.2 Algorithms for maximizing expected F-score .......... 72  
      3.3.3 Soundness of approximate Bayesian labeling criteria .... 75  
      3.3.4 Efficient calculation of empirical expected F-score ..... 76  
   3.4 Empirical evaluation ........................................... 79  
      3.4.1 Dataset .................................................. 79  
      3.4.2 Algorithms .............................................. 81  
      3.4.3 Performance measure ..................................... 84  
      3.4.4 Results .................................................. 84  
   3.5 Conclusion and future directions .............................. 90

4 Kernel Measures of Independence for non-iid Data 91  
   4.1 Preliminaries of RKHS embeddings of probabilities ...... 93  
      4.1.1 Distance between distributions ............................ 96  
      4.1.2 Hilbert-Schmidt independence criteria .................... 97  
      4.1.3 Applications of HSIC ..................................... 100  
   4.2 Embedding distributions with graphical models .......... 100  
      4.2.1 Factorization of mean operators .......................... 101  
      4.2.2 Factorization of RKHS for factorized kernels .......... 102  
      4.2.3 Injectivity of factored mean operators ................. 105
## Contents

4.2.4 Factorization of independence criteria .................................. 108
4.3 Estimates for special structures ............................................. 109
   4.3.1 Independent and identically distributed data .................... 109
   4.3.2 Sequence data ..................................................... 110
   4.3.3 TD-SEP as a special case ........................................ 111
   4.3.4 Grid structured data ............................................. 112
4.4 Experiments ............................................................... 112
   4.4.1 Independence test .................................................. 113
   4.4.2 Independent component analysis .................................. 115
   4.4.3 Time series clustering and segmentation ......................... 116
4.5 Conclusion ................................................................. 120

5 Lower Bounds for BMRM and Faster Rates for Training SVMs .......... 125
   5.1 Preliminaries ........................................................... 127
   5.2 Lower bounds .......................................................... 128
      5.2.1 Concepts and notations ......................................... 129
      5.2.2 Strong lower bounds ............................................ 131
      5.2.3 Weak lower bounds ............................................. 132
   5.3 A new algorithm with convergence rates $O(1/\sqrt{\epsilon})$ ........ 134
      5.3.1 Convergence rates ............................................. 136
      5.3.2 A linear time algorithm for simple QP ....................... 137
      5.3.3 Other versions of Nesterov’s algorithms ...................... 140
   5.4 Structured output space .............................................. 141
      5.4.1 Margin scaled maximum margin Markov network ............... 142
      5.4.2 Efficient projection onto factorized simplex .................. 144
   5.5 Experimental results .................................................. 147
   5.6 Discussion and conclusions ............................................ 149

A Fundamentals of Convex Analysis ........................................ 155
   A.1 Convex set and convex function ..................................... 155
   A.2 Fenchel conjugate .................................................... 157
   A.3 Convex analysis for the log partition function ..................... 160

B Message Update Formulae of Expectation Propagation .................. 167
   B.1 Preliminaries: canonical parametrization of multi-variate Gaussians .. 167
   B.2 EP updates for all factors in Figure 3.1 ........................... 168
   B.3 Message passing for the max factor .................................. 172
      B.3.1 Mathematical formulation of messages ........................ 172
      B.3.2 Moments of the maximum of multiple Gaussians ............... 174
## Contents

C Detailed Result of Empirical Optimal Threshold ........................................... 183

D Modeling Hierarchies in Labels ........................................................................... 189

E Statistical Estimation and Concentration of Measure ...................................... 191
   E.1 Desirable statistical properties of estimators ............................................. 191
   E.2 U-Statistics for iid observations ................................................................. 193
   E.3 Statistics for non-iid observations .............................................................. 195
   E.4 Mixing coefficients using Markov properties ............................................ 198
      E.4.1 Bounds on $\|\triangle_n\|_\infty$ using Markov property ....................... 199
   E.5 Concentration of measure with function space ......................................... 202
      E.5.1 Rademacher averages .................................................................. 203
      E.5.2 Extension to non-iid observations .................................................. 204

F Proof for Proposition 44 ...................................................................................... 205
   F.1 Finite linear combination ........................................................................ 205
   F.2 Limit of linear combination .................................................................... 206
   F.3 Proof of Lemma 92 ................................................................................ 207

G Incomplete Cholesky Decomposition for HSIC-Struct ...................................... 211

H Detailed proof for Theorem 57 ............................................................................. 213
   H.1 Example Dataset and Initial Few Steps .................................................. 213
   H.2 Asymptotic Rates .................................................................................. 215
   H.3 Proof of Theorem 93 ............................................................................. 216
List of Figures

1.1 Structure of introduction. ........................................... 3
1.2 2-D Grid CRF. $X_{ij}$ are observations and $Y_{ij}$ are latent states. .... 10
1.3 Example factor graph ................................................. 20
1.4 Simple break of injectivity ........................................... 25
1.5 Imaginary break of injectivity ....................................... 25
1.6 0-1 loss, hinge loss, logistic loss and exponential loss for binary classification. ........................................ 30
1.7 Slack rescaling and margin rescaling. ................................ 31

2.1 Four different partitions of a 5-by-6 CRF. Nodes in shaded and white regions are the two trees and the small black circles represent observations. 44
2.2 Abstract grid alignment domain. ................................... 50
2.3 Average reward over the last 1000 steps, and iterations to optimality. .. 51
2.4 Sensor network domain with 8 sensors, 2 targets, and 3 cells. ........... 52
2.5 Results over 100 runs of the Sensor Network scenario, varying the number of targets. ............................. 53
2.6 Results over 100 runs of the Sensor Network scenario, varying the number of cells. ............................. 53
2.7 Results over 50 runs of the road traffic offset scenario. The X-axis is intersections. The left Y-axis is the travel time. ............. 55
2.8 Convergence of NAC on the independent agents (upper curve) and on CRF model (lower curve). The number of controllable intersections is 14. 56

3.1 A factor graph for multi-class classification. Class 2 is the true label. . 61
3.2 A factor graph for multi-label classification via pairwise comparison. Here class 2 and 4 are the relevant labels. ..................... 63
3.3 A factor graph for multi-label classification using max and min factors. 64
3.4 A factor graph for multi-label classification via total ordering and a global bias. See text for explanation. ......................... 65
3.5 A dynamic graphical model with factors between the model of two adjacent examples. ............................................. 67
3.6 Example curves of $\tilde{\text{ExpFs}}(l(\theta))$ (blue) and $\text{F-score}(l(\theta), y^\ast)$ (red) v.s. $\theta$. 76
LIST OF FIGURES

3.7 Total number of features in the training set versus the number of training examples. .................................................. 80
3.8 Average number of non-zero features per example versus the number of training examples. .................................................. 80
3.9 F-score for the category group industries. ................................. 85
3.10 F-score for the category group regions. ................................... 86
3.11 F-score for the category group topics. ................................... 86
3.12 industries ............................................................................. 88
3.13 regions ............................................................................... 88
3.14 topics .................................................................................. 88
3.15 CPU time for training. .......................................................... 89
4.1 Graphical model for the XOR problem \( y_t = x_t \otimes x_{t-1} \). .................. 92
4.2 From left to right: (a) a graphical model representing \( iid \) observations, (b) a graphical model for first order sequential data, and (c) a graphical model for dependency on a two dimensional mesh. ................................. 109
4.3 Independence test for a Gaussian process. .................................. 115
4.4 Illustration of error calculation. Red lines denote the ground truth and blues line are the segmentation results. The error introduced for segment \( R_1 \) to \( R'_1 \) is \( a + b \), while that for segment \( R_2 \) to \( R'_2 \) is \( c + d \). The overall error in this example is then \( (a + b + c + d) / 4 \). ................................. 120
4.5 Segmentation results of BCI dataset produced by (a) structured HSIC, (b) spectral clustering and (c) HMM. In (c), we did specify 4 hidden states, but the Viterbi decoding showed only two states were used. ................. 121
4.6 Segmentation results of swimming dataset 1 produced by (a) structured HSIC, (b) spectral clustering and (c) HMM. .................................................. 121
4.7 Segmentation results of swimming dataset 2 produced by (a) structured HSIC, (b) spectral clustering and (c) HMM. .................................................. 122
4.8 Segmentation results of swimming dataset 3 produced by (a) structured HSIC, (b) spectral clustering and (c) HMM. .................................................. 122
5.1 \( h_i(\lambda) \) ................................................................................ 139
5.2 Primal function error versus number of iterations. .................................. 150
5.3 Primal function error versus time. .................................................. 151
5.4 Test accuracy versus number of iterations. ........................................ 152
5.5 Test accuracy versus time. .......................................................... 153
B.1 Recursively apply the max of two Gaussians in a chain structure to compute the max of \( n \) Gaussians. .................................................. 175
LIST OF FIGURES

B.2 Recursively apply the max of two Gaussians in a tree structure to compute the max of $n$ Gaussians. .................. 175

B.3 Relationship among symbols. Those without inbound arrows can be computed directly from $\theta$ and $\Sigma$. .................. 180

C.1 Example curves of $\tilde{\text{ExpFs}}(l(\theta))$ (blue) and $\text{F-score}(l(\theta), y^*)$ (red) v.s. $\theta$. 183
C.2 Example curves of $\tilde{\text{ExpFs}}(l(\theta))$ (blue) and $\text{F-score}(l(\theta), y^*)$ (red) v.s. $\theta$. 184
C.3 Example curves of $\tilde{\text{ExpFs}}(l(\theta))$ (blue) and $\text{F-score}(l(\theta), y^*)$ (red) v.s. $\theta$. 185
C.4 Example curves of $\tilde{\text{ExpFs}}(l(\theta))$ (blue) and $\text{F-score}(l(\theta), y^*)$ (red) v.s. $\theta$. 186
C.5 Example curves of $\tilde{\text{ExpFs}}(l(\theta))$ (blue) and $\text{F-score}(l(\theta), y^*)$ (red) v.s. $\theta$. 187
List of Tables

1.1 Example kernels on $\mathbb{R}^n$ ........................................... 12
1.2 Loss for multi-class classification ....................................... 29
1.3 Loss for binary classification ............................................. 29

2.1 Learning algorithm parameters. Lower $\beta$ and $\gamma$ give lower variance but higher bias in the gradient estimation. Too small $\alpha$ leads to slow learning, but too large $\alpha$ causes zigzag which also slows down the learning. $\alpha_{\text{node}}$ is N/A for Traffic because it does not use node features. ........ 51

3.1 Number of features under different sizes of idf set. ................. 80

4.1 The number of times HSIC and structured HSIC rejected the null hypothesis. ............................................................. 114
4.2 Median performance of ICA on music using HSIC, TDSEP, and structured HSIC. In the top row, the number $m$ of sources and $n$ of samples are given. In the second row, the number of time lags $\tau$ used by TDSEP and structured HSIC are given: thus the observation vectors $x_t, x_{t-1}, \ldots, x_{t-\tau}$ were compared. The remaining rows contain the median Amari divergence (multiplied by 100) for the three methods tested. The original HSIC method does not take into account time dependence ($\tau = 0$), and returns a single performance number. Results are in all cases averaged over 136 repetitions: for two sources, this represents all possible pairings, whereas for larger $m$ the sources are chosen at random without replacement. ...................................................... 117
4.3 Segmentation errors by various methods on the four studied time series. 120

5.1 Dataset statistics. $n$: #examples, $d$: #features, $s$: feature density. ... 147
5.2 $\lambda$ for datasets. ............................................................. 148

B.1 Correspondence Between Symbols in This Report and in (Afonja, 1972). 180
# List of Symbols

<table>
<thead>
<tr>
<th>Symbol</th>
<th>Description</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>$[n]$</td>
<td>Index set ${1, 2, \ldots, n}$</td>
<td>3</td>
</tr>
<tr>
<td>$\mathbb{R}_+$</td>
<td>Set of positive real numbers $(0, \infty)$</td>
<td>3</td>
</tr>
<tr>
<td>$\mathbb{R}_+$</td>
<td>Set of non-negative real numbers $[0, \infty)$</td>
<td>3</td>
</tr>
<tr>
<td>$\phi$</td>
<td>Vector of sufficient statistics</td>
<td>3</td>
</tr>
<tr>
<td>$\top$</td>
<td>Transpose of matrix or vector</td>
<td>3</td>
</tr>
<tr>
<td>$\mathcal{X}$</td>
<td>Sample space, or space of feature vectors</td>
<td>3</td>
</tr>
<tr>
<td>$\theta$</td>
<td>Natural parameter</td>
<td>3</td>
</tr>
<tr>
<td>$d$</td>
<td>Number of features</td>
<td>3</td>
</tr>
<tr>
<td>$\nu$</td>
<td>Base measure of exponential family of distributions</td>
<td>3</td>
</tr>
<tr>
<td>$\Theta$</td>
<td>The set of $\theta$ for which $g(\theta) &lt; +\infty$</td>
<td>4</td>
</tr>
<tr>
<td>$\mathcal{P}_\phi$</td>
<td>Exponential family generated by $\phi$</td>
<td>4</td>
</tr>
<tr>
<td>$g(\theta)$</td>
<td>Log partition function</td>
<td>3</td>
</tr>
<tr>
<td>$C^\infty$</td>
<td>Infinitely differentiable</td>
<td>4</td>
</tr>
<tr>
<td>$\text{MRF}_G$</td>
<td>Markov random field corresponding to graph $G$</td>
<td>6</td>
</tr>
<tr>
<td>$X \perp \perp Y \mid Z$</td>
<td>Random variable $X$ is independent of $Y$ given $Z$</td>
<td>6</td>
</tr>
<tr>
<td>$\mathcal{C}$</td>
<td>Set of maximal cliques</td>
<td>7</td>
</tr>
<tr>
<td>$k(x, y)$</td>
<td>Kernel function</td>
<td>11</td>
</tr>
<tr>
<td>$\mathbb{N}$</td>
<td>Set of natural numbers: ${1, 2, \ldots}$</td>
<td>12</td>
</tr>
<tr>
<td>$\mathbb{N}_0$</td>
<td>$\mathbb{N} \cup {0} = {0, 1, 2, \ldots}$</td>
<td>12</td>
</tr>
<tr>
<td>$\delta(\cdot)$</td>
<td>$\delta(\cdot) = 1$ if $\cdot$ is true, and 0 otherwise</td>
<td>12</td>
</tr>
<tr>
<td>$\mathcal{H}(p)$</td>
<td>Reproducing kernel Hilbert space (RKHS)</td>
<td>14</td>
</tr>
<tr>
<td>$\mathcal{H}$</td>
<td>$\mathcal{H}$</td>
<td>14</td>
</tr>
<tr>
<td>$\Rightarrow$</td>
<td>Uniform convergence of functions</td>
<td>14</td>
</tr>
<tr>
<td>$C(\mathcal{X})$</td>
<td>Space of continuous bounded functions $\mathcal{X}$</td>
<td>14</td>
</tr>
<tr>
<td>$\mathcal{P}_k$</td>
<td>Exponential family generated by kernel $k$</td>
<td>16</td>
</tr>
<tr>
<td>$n$</td>
<td>Number of data points</td>
<td>17</td>
</tr>
<tr>
<td>$\mathcal{Y}^m$</td>
<td>Cartesian product of $\mathcal{Y}$ for $m$ times</td>
<td>18</td>
</tr>
<tr>
<td>$\text{KL}(p, q)$</td>
<td>KL divergence from density $p$ to density $q$: $\int p(x) \frac{p(x)}{q(x)} dx$</td>
<td>21</td>
</tr>
<tr>
<td>$f^*$</td>
<td>Fenchel dual of function $f$</td>
<td>23</td>
</tr>
<tr>
<td>$\mathcal{M}_\phi$</td>
<td>Marginal polytope for sufficient statistics $\phi$</td>
<td>24</td>
</tr>
<tr>
<td>$\Lambda_\phi$</td>
<td>Mapping from $\theta \in \Theta$ to the mean $\mathbb{E}_\theta[\phi(x)]$</td>
<td>24</td>
</tr>
<tr>
<td>$\text{ri}$</td>
<td>Relative interior</td>
<td>24</td>
</tr>
<tr>
<td>Symbol</td>
<td>Description</td>
<td>Page</td>
</tr>
<tr>
<td>------------</td>
<td>--------------------------------------------------</td>
<td>------</td>
</tr>
<tr>
<td>$H(p)$</td>
<td>Differential entropy of density $p$</td>
<td>24</td>
</tr>
<tr>
<td>$R_{emp}(\mathbf{w})$</td>
<td>Empirical risk of weight vector $\mathbf{w}$</td>
<td>28</td>
</tr>
<tr>
<td>$\Omega(\mathbf{w})$</td>
<td>Regularizer</td>
<td>28</td>
</tr>
<tr>
<td>$|\mathbf{w}|_1$</td>
<td>$L_1$ norm, $|\mathbf{w}|_1 = \sum_i</td>
<td>w_i</td>
</tr>
<tr>
<td>$\epsilon_k$</td>
<td>Error at step $k$: $J(w_k) - J(w^*)$</td>
<td>32</td>
</tr>
<tr>
<td>$\text{qp-bmrm}$</td>
<td>BMRM with full inner solver</td>
<td>38</td>
</tr>
<tr>
<td>$\text{ls-bmrm}$</td>
<td>BMRM with line search inner solver</td>
<td>38</td>
</tr>
<tr>
<td>$R(\theta)$</td>
<td>Expected long term reward under policy $\theta$</td>
<td>46</td>
</tr>
<tr>
<td>$e_t$</td>
<td>Eligibility trace</td>
<td>47</td>
</tr>
<tr>
<td>$y^i$</td>
<td>Label (sequence) for example $i$</td>
<td>60</td>
</tr>
<tr>
<td>$\delta(x)$</td>
<td>Dirac function: $\delta(x) = 0$ for all $x \neq 0$ and $\int R \delta(x)dx = 1$</td>
<td>61</td>
</tr>
<tr>
<td>$\text{ExpFs}(l)$</td>
<td>Expected F-score of outputting label $l$</td>
<td>70</td>
</tr>
<tr>
<td>$\text{ExpFs}(l)$</td>
<td>Empirical estimate of $\text{ExpFs}(l)$ based on samples</td>
<td>74</td>
</tr>
<tr>
<td>$\mu[P]$</td>
<td>Mean element of distribution $P$</td>
<td>94</td>
</tr>
<tr>
<td>$\text{MMD}(P, Q)$</td>
<td>Maximum mean discrepancy between $P$ and $Q$</td>
<td>96</td>
</tr>
<tr>
<td>$R_n(\mathcal{H}, P)$</td>
<td>Rademacher average of $\mathcal{H}$ wrt $P$</td>
<td>96</td>
</tr>
<tr>
<td>$\text{Hsic}(X, Y)$</td>
<td>Hilbert-Schmidt independence criteria</td>
<td>97</td>
</tr>
<tr>
<td>$\text{HSIC}_b(Z^n)$</td>
<td>Biased estimator of HSIC</td>
<td>98</td>
</tr>
<tr>
<td>$\text{HSIC}_u(Z^n)$</td>
<td>Unbiased estimator of HSIC</td>
<td>99</td>
</tr>
<tr>
<td>$k_S(x_S, x_S')$</td>
<td>Kernel on the subset $S$</td>
<td>101</td>
</tr>
<tr>
<td>$\mathcal{H}_S$</td>
<td>Reproducing kernel Hilbert space on subset $S$</td>
<td>101</td>
</tr>
<tr>
<td>$\text{CE}$</td>
<td>Constant-exclusive kernel</td>
<td>107</td>
</tr>
<tr>
<td>$\text{CCE}$</td>
<td>Coordinate Constant-exclusive kernel</td>
<td>107</td>
</tr>
<tr>
<td>$k_G$</td>
<td>Ground kernel on the cliques of $X$</td>
<td>109</td>
</tr>
<tr>
<td>$[\cdot]_+$</td>
<td>Hinge function $[x]_+ := \max{0, x}$</td>
<td>126</td>
</tr>
<tr>
<td>$\mathbb{R}$</td>
<td>$\mathbb{R} \cup {+\infty}$</td>
<td>127</td>
</tr>
<tr>
<td>$L-l.c.g$</td>
<td>Lipschitz continuous gradient with modulus $L$</td>
<td>127</td>
</tr>
<tr>
<td>$\sigma-s.c.$</td>
<td>Strongly convex with modulus $\sigma$</td>
<td>127</td>
</tr>
<tr>
<td>$\partial f(\mathbf{w})$</td>
<td>The set of all subgradients of $f$ at $\mathbf{w}$</td>
<td>128</td>
</tr>
<tr>
<td>$\Delta(y, y^i)$</td>
<td>Discrepancy between $y$ and $y^i$</td>
<td>141</td>
</tr>
</tbody>
</table>