Toward Practical Reinforcement Learning Algorithms: Classification Based Policy Iteration and Model-Based Learning

by

Bernardo Ávila Pires

A thesis submitted in partial fulfillment of the requirements for the degree of

Doctor of Philosophy

in

Statistical Machine Learning

Department of Computing Science

University of Alberta

Abstract

In this dissertation, we advance the theoretical understanding of two families of Reinforcement Learning (RL) methods: Classification-based policy iteration (CBPI) and model-based reinforcement learning (MBRL) with factored semi-linear models.

In contrast to generalized policy iteration, CBPI does not rely on value-function estimates (value estimates for all states and actions). Instead, CBPI uses a classifier to learn how to discriminate value-maximizing actions based on value estimates for a set of observed states and actions. This creates the potential for learning effective policies in settings where estimating value functions is challenging, but where good value estimates can be obtained and good actions can be distinguished from sub-optimal ones.

Previous theoretical work on CBPI has required classifiers that are computed by solving a combinatorial problem, which we can expect to be computationally hard (with minimization of the infamous 0-1 loss as a special case). In contrast, classifiers that are computed by solving convex minimization problems (which can be done efficiently) enjoy limited or no performance guarantees, namely, bounds on the cost-sensitive generalization of the misclassification probability, the so-called "true risk". Therefore, we only have instances of CBPI that enjoy theoretical guarantees but cannot be used in practice, or vice-versa. We present a theoretical analysis of CBPI that fills this gap, by providing theoretical guarantees for instances of CBPI

that can be used in practice (by virtue of the classification methods used).

Our analysis extends an existing theoretical analysis of CBPI, and incorporates performance guarantees for classification methods that minimize a cost-sensitive multiclass convex surrogate loss that generalizes the hinge loss. The hinge loss has been widely used in the design of classification methods, including the popular support vector machines (SVMs). As part of our analysis, we also present results for cost-sensitive multiclass classification: Novel expected surrogate loss (surrogate-risk) bounds, as well as tools for converting surrogate risk bounds into true risk bounds. This is done with the help of novel calibration functions that are specific to cost-sensitive multiclass classification losses. Moreover, our analysis of CBPI can be easily adjusted to accommodate for different classification methods, provided that the corresponding surrogate risk bounds and calibration functions are available.

We also present policy error bounds for MBRL methods that use factored semi-linear models. The factored semi-linear model framework generalizes the factored linear model framework and many existing MBRL methods. Notably, factored semi-linear models generalize a recent trend of MBRL methods that depart from learning "traditional" MDP models in order to achieve flexibility, computational efficiency, data efficiency and good empirical performance. As such, factored semi-linear models are both flexible and geared toward efficient policy computation, with instances that have been shown to be promising in practice.

The policy error bounds that we present improve the previously existing bounds by relaxing conditions, refining the bounds themselves, and increasing the scope of models that they apply to—namely, factored semi-linear models. These bounds allow us to understand the policy error in norms

other than the overly-harsh supremum norm. For example, our $L^p(\mu)$ norm results allow us to see that policy error bounds for MBRL methods with factored semi-linear models are less sensitive to covariate-shift than policy error bounds for competing methods, such as approximate linear programming or approximate dynamic programming methods. This robustness suggests that MBRL methods with factored semi-linear models have much potential to be a valid alternative to popular non-model-based RL methods.

Preface

The introduction to reinforcement learning in Chapter 1 presents concepts and ideas from Puterman (1994); Bertsekas and Tsitsiklis (1996); Sutton and Barto (1998); Bertsekas (2007); Szepesvári (2010); Bertsekas (2016). The rest of this dissertation is grounded on joint work with the author's supervisor, Csaba Szepesvári.

Chapter 2 is original to this dissertation, and the original results therein have not been previously published. The results are closest to recent work of the author, Csaba Szepesvári and Mohammad Ghavamzadeh (Ávila Pires et al., 2013; Ávila Pires and Szepesvári, 2016a).

Chapter 3 is also original to this dissertation, and the original results therein have not been previously published. The results build on those of Chapter 2, and most closely on the work of Lazaric et al. (2016).

Chapter 4 is an updated version of a conference paper authored by the author and Csaba Szepesvári (Ávila Pires and Szepesvári, 2016b). While the text and comments are largely taken from the conference paper, in Chapter 4 I fully develop an extension of factored linear models that was only briefly discussed in the related work section (Section 6) of Ávila Pires and Szepesvári (2016b): In the updated results, we are no longer required to have a join-homomorphism assumption (cf. Theorems 11 and 12 of Ávila Pires and Szepesvári, 2016b), and I generalize the factored linear model framework to what we call factored semi-linear models.

To my lovely wife, Mahsa
To God

Acknowledgments

I have had a great time at the University of Alberta. My time there has been not just two graduate programs, MSc and PhD, but life: My experiences at the University are part of who I am, and have dramatically shaped my foreseeable future.

At the University of Alberta, during my PhD, I met my sweet, dear wife Mahsa—without her support and love, I would have achieved nothing. She is the joy of my days.

Although I have been physically distant from my beloved parents, Fernando and Lélia, they have incessantly and warmly supported me, loving me without reservation. All my achievements stems from their dedication and love as parents and friends.

For these seven years that comprised the MSc and PhD, I have worked with Prof. Csaba Szepesvári, who teaches and works with absolute patience, diligence and passionate devotion. Csaba is a fantastic supervisor, researcher, mentor, colleague and friend, from whom I have learned immensely, and to whom I am extremely grateful.

I must also extend my gratitude to all who have supported me throughout these many years. I give especial thanks to

- ◇ all my friends and family, in Canada, Brazil and elsewhere, in particular Levi, Estácio, Letícia (and their baby whom we all love, Helena), Gabriel, Thaís, and Arthur;
- ◇ the Basilian Fathers, in particular Frs. Glenn, Dave, Terry and Don, as well as all the community and friends from St. Joseph's College, where I have grown in faith, and found hope and support in the difficult times;
- ◇ my Sensei, Sean Bowen, and my fellow karateka;
- ◇ the University of Alberta faculty, in particular Prof. Dale Schuurmans;
- ◇ the staff at the Department of Computer Science, the International Centre, and the Counseling and Clinical Services, in particular Jason Murray;

Thanks be also, and above all, to God, from Whom flow all the generosity and love that I have received from all around me, and from Whom comes all that is good, and all the gifts that I have been given. May this work and my work serve to further His love in this world.

***

This work was supported by Alberta Innovates Technology Futures and NSERC. My special thanks to Professors Csaba Szepesvári, Dale Schuurmans, András György, Ivan Mizera and Ambuj Tewari for their valuable assistance and feedback in the preparation of this work.

Contents

List of Figures

List of Symbols

List of Abbreviations

RL	Reinforcement Learning	1
MDP	Markov Decision Process	1
DP	Dynamic Programming	1
w.r.t.	with respect to	2
w.p.	with probability	3
s.t.	such that	3
CBPI	Classification-Based Policy Iteration	8
MBRL	Model-Based Reinforcement Learning	8
DPL	Direct Policy Learning	8
GPI	Generalized Policy Iteration	8
ERM	Empirical Risk Minimization	12
SVM	Support Vector Machine	17

Chapter 1

Introduction

Consider an agent that interacts with the environment by taking an action and then observing some information about the environment, along with a numerical reward signal. *Reinforcement learning* (RL) is the problem of constructing agents that interact with an unknown environment in the described fashion and are capable of maximizing the amount of reward that they collect on the long run.

Reinforcement Learning is well-suited for designing goal-oriented agents with the potential to succeed in a variety of tasks without requiring too much (though often benefiting from) prior knowledge about each specific task. The field of RL has been subject of much research and the years have seen many successful applications. From the theoretical point of view, RL is related to a series of challenging and interesting problems, *e.g.*, controlling systems with complex dynamics, dealing with delayed feedback structure, exploration, covariate shift, and even supervised learning (*e.g.*, regression and classification).

In this chapter, we will give a minimalistic introduction to the reinforcement learning problem in its most studied setting, where the environment is treated as a Markov Decision Process (MDP). We also introduce Dynamic Programming (DP), the standard approach to "solve" MDPs. There is a vast body of literature on various reinforcement learning methods that we do not discuss in this dissertation, as well as variants of the basic MDP framework. We refer the reader to the works of Puterman (1994); Bertsekas

and Tsitsiklis (1996); Sutton and Barto (1998); Bertsekas (2007); Szepesvári (2010); Buşoniu et al. (2010b); Powell (2011); Wiering and van Otterlo (2012); Bertsekas (2016), who give a comprehensive treatment of Reinforcement Learning.

We will close this introductory chapter with an outline of this dissertation and an overview of the scientific contributions presented in each of the subsequent chapters.

1.1 Reinforcement Learning

In this section we define the reinforcement learning problem in the MDP framework, and introduce dynamic programming. This section is derived from the works of Puterman (1994), Bertsekas and Tsitsiklis (1996), Sutton and Barto (1998), Szepesvári (2010) and Bertsekas (2010).

1.1.1 Problem Definition

We will treat the environment as a Markov Decision Process (MDP). An MDP is a discrete-time sequential decision framework where at each time-step t the agent is assumed to observe[1] a random state[2] X_t taking values in a *state space* $\mathcal{X}$, takes a (possibly-random) action A_t taking values in an *action space* $\mathcal{A}$, observers a random state $X_{t+1} \in \mathcal{X}$ and a random reward

[1] It may not always be the case that the agent is able to fully observe the state of the environment, or that the environment is even Markovian. The framework of partially-observable MDPs (POMDPs) (Spaan, 2012) allows us to consider scenarios where agents do not fully observe the state of a finite MDP, and to account for these partial observations. Hutter (2014) shows that if the environment (be it partially-observable, non-Markov, adversarial, etc.) enjoys some regularities then we do not lose much by assuming that it is an MDP and then using an optimal policy for this MDP to act in the environment.

[2] In this dissertation, we omit technical details related to measurability, as they are well-understood. We will assume that sets and functions are measurable with respect to (w.r.t.) the underlying measure spaces. Moreover, predicates of the form [for any $\mathcal{Z}' \subset \mathcal{Z}$] should be read as [for any measurable $\mathcal{Z}' \subset \mathcal{Z}$]. We will mention measurability explicitly in the few situations where this technicality must be treated with care. To shorten notation, for a random variable Z taking values in some $\mathcal{Z}$ and some random variable W, we will treat $\mathbb{P}(Z|W)$ as a $\sigma(W)$-measurable distribution p such that $p(\mathcal{Z}') = \mathbb{P}(Z \in \mathcal{Z}'|W)$ for all measurable $\mathcal{Z}' \subset \mathcal{Z}$. This notation only clashes with standard notation when Z takes values in $\{0,1\}$ (almost-surely), which will not happen in this dissertation. As usual, conditioning on W should be read as conditioning on the sigma-algebra generated by W.

$R_{t+1} \in \mathbb{R}$, takes another action $A_{t+1} \in \mathcal{A}$, etc. The random variables (X_{t+1}, R_{t+1}) $(t \geq 0)$ are assumed to be *Markov*, that is, they are conditionally independent of the "past" $(X_0, A_0, R_1, X_1, \ldots, X_{t-1}, A_{t-1})$ given the "present" (X_t, A_t):

$$\mathbb{P}(\, X_{t+1}, R_{t+1} |\, X_0, A_0, R_1, X_1, \ldots, X_t, A_t) = \mathbb{P}(\, X_{t+1}, R_{t+1} |\, X_t, A_t), \quad (1.1.1)$$

where $X_0 \sim \alpha$ is a random initial state and α is the *initial state distribution* of the MDP. We will refer to the sequence $X_0, A_0, R_1, X_1, \ldots, X_t$ for $t \geq 0$ as a *trajectory*.

In the MDP framework, agents are merely strategies for choosing $(A_0, A_1, \ldots)$. A strategy for choosing actions is called a *policy*, which maps trajectories to distributions over the set of actions $\mathcal{A}$, that is

$$A_t \sim \pi(X_0, A_0, R_1, X_1, \ldots, X_t). \quad (1.1.2)$$

Optionally, we may restrict the actions that policies may choose at each state. In this straightforward extension of the MDP framework, there is a function $a : \mathcal{X} \to 2^{\mathcal{A}}$ and policies are restricted to satisfy $A_t \in a(X_t)$ with probability (w.p.) one. In this text, in order to simplify notation, we will not use this extension, but the results do not depend on the simplifying assumption that all actions are allowed at all states.

We say that a policy π is *stationary* if the actions A_t depend only on the respective X_t, that is, if for all $t \geq 0$

$$\mathbb{P}(\, A_t |\, X_0, A_0, R_1, X_1, \ldots, X_t) = \mathbb{P}(\, A_t |\, X_t),$$

in which case we can write $A_t \sim \pi(X_t)$. We say that π is *deterministic* if $\pi(X_t)$ is degenerate, *i.e.*, there exists $f : \mathcal{X} \to \mathcal{A}$ such that (s.t.) $A_t = f(X_t)$ w.p. one. In this case we will abuse notation and write $A_t = \pi(X_t)$.

As mentioned before, the goal in the reinforcement learning problem is to compute policies that are capable of maximizing the amount of reward that they collect on the long run. The long-run aspect is important, as normally we are not interested in policies that maximize only the expected

immediate reward $\mathbb{E}(R_1)$, or $\mathbb{E}(R_t)$ in step t. Rather, we want a policy π that chooses $(A_0, A_1, \ldots)$ so as to maximize the *expected discounted total reward*, or *expected return*[3]

$$\mathbb{E}\left(\sum_{t=1}^{\infty} \gamma^{t-1} R_t\right), \tag{1.1.3}$$

where $\gamma \in [0, 1)$ is a so-called discount factor. The closer γ is to one, the greater the weight given to rewards in the distant future. We will assume that the rewards are s.t. (1.1.3) is well-defined and finite, regardless of π and α. As an example, it would be sufficient take uniformly bounded rewards: $\sup_t |R_t| \leq r_{\max}$ w.p. one, for some constant $r_{\max} < \infty$.

It is possible to impose additional assumptions so that (1.1.3) also holds with $\gamma = 1$ (see Bertsekas and Tsitsiklis, 1996, Section 2.1, pp. 12–14 and Szepesvári, 2010, Section 2.2, p. 11), leading to an *undiscounted* objective. When $\gamma = 1$, the MDP is assumed to be episodic, in the following sense: An MDP is *episodic* if it has a set of *terminal states* $\mathcal{T} \subset \mathcal{X}$ satisfying $\mathbb{P}(X_{t+1} \in \mathcal{T}, R_{t+1} = 0 | X_t \in \mathcal{T}) = 1$. In other words, terminal states are absorbing and incur zero return for any agent acting in the MDP. An additional restricting and simplifying assumption often made in literature is that a terminal state is eventually always reached, regardless of how the actions are chosen by the policy. Formally, for some t large enough we have $X_t \in \mathcal{T}$ w.p. one.

It is also possible to have undiscounted ($\gamma = 1$) MDPs with a different goal, such as maximizing the *average reward* (Bertsekas and Tsitsiklis, 1996, Section 2.1, p. 15):

$$\limsup_{t \to \infty} \mathbb{E}\left(\frac{1}{t+1} \sum_{s=0}^{t} R_{s+1}\right). \tag{1.1.4}$$

Throughout this text, we will assume that maximizing the expected discounted/undiscounted *total* reward is the objective of the agent.

From now on, we will assume that the action space is $\mathcal{A}$ is finite, and that rewards R_{t+1} are given by a reward function $r : \mathcal{X} \times \mathcal{A} \to \mathbb{R}$ so that

[3] We define $\gamma^0 \doteq 1$ if $\gamma = 0$.

$r(X_t, A_t) = \mathbb{E}(R_{t+1} | X_t, A_t)$ holds w.p. one. The assumption that rewards are deterministic is a simplifying assumption that can be removed without much effort. The MDP can then be written as a tuple $(\mathcal{X}, \mathcal{A}, \mathcal{P}, r, \alpha, \gamma)$, where the only component that remains to be introduced is $\mathcal{P}$, the *transition probability kernel* of the MDP, defined such that

$$X_{t+1} \sim \mathcal{P}(X_t, A_t)$$

w.p. one. Formally, $\mathcal{P}$ maps $\mathcal{X} \times \mathcal{A}$ to probability measures over $\mathcal{X}$ and satisfies that, for $\mathcal{U} \subset \mathcal{X}$, $(x, a) \mapsto \mathcal{P}(x, a)(\mathcal{U})$ is measurable. That such a kernel exists follows from (1.1.1) under mild assumptions on $\mathcal{X}$.

1.1.2 Dynamic Programming

Our goal is to construct a policy π so as to maximize the expected return given in (1.1.3). Whether and under which circumstances we can succeed will depend on a variety of factors, including the type and the amount of information available to us. In the simplest setting, the MDP is known, in which case we can choose π so as to maximize the *value* of each state x, *i.e.*, the total expected discounted reward of π given that the initial state X_0 is x:

$$V^\pi(x) \doteq \mathbb{E}\left(\sum_{t=1}^{\infty} \gamma^{t-1} R_t \,\middle|\, X_0 = x \right),$$

with (X_t, A_t) satisfying (1.1.1) and (1.1.2). V^π is called the *value function* of π. The largest possible value function is the *optimal value function*, defined by

$$V^*(x) \doteq \sup_\pi V^\pi(x). \tag{1.1.5}$$

A policy whose value function is optimal is called an *optimal policy*, and will be denoted by π^*. Note that while the optimal value function is unique, there may be multiple optimal policies (in which case π^* will denote an arbitrary optimal policy). Letting $\Pi \doteq \mathcal{A}^\mathcal{X}$ be the set of stationary deterministic policies, it is known that if $\gamma < 1$ then $V^*(x) = \sup_{\pi \in \Pi} V^\pi(x)$ and an optimal stationary deterministic policy $\pi^* \in \Pi$ exists (Szepesvári,

2010, Section 2.3, p. 14). When $\gamma = 1$, similar statements can be made under additional assumptions on the reward function (see Bertsekas, 2010, Chapter 7: Corollary 2.2 and Proposition 3).

To find π^*, we can first solve the *Bellman optimality equation* (Szepesvári, 2010, Section 2.2, p. 15) for V^*. This equation states that V^* must satisfy, for all $x \in \mathcal{X}$

$$V^*(x) = \max_a \mathbb{E}\left(R_{t+1} + \gamma V^*(X_{t+1}) \middle| X_t = x, A_t = a \right). \tag{1.1.6}$$

Once V^* is found, for each x we can simply take $\pi^*(x)$ to be a maximizing action in (1.1.6). Another powerful fact is that any $V : \mathcal{X} \to \mathbb{R}$ that satisfies the fixed point equation (1.1.6) must be equal to V^*, *i.e.*, (1.1.6) has a unique fixed point. Therefore, we have a clear approach to find π^*: Find the fixed point of (1.1.6) and take, for each x, $\pi^*(x)$ as a maximizing action in (1.1.6). This approach is known as *dynamic programming*.

We can re-write (1.1.6) compactly using operator notation, which we now introduce (reproduced here from Ávila Pires and Szepesvári, 2016b, Section 2). Recall that "*a Banach space is a normed* [vector] *space that is also a complete metric space under the metric induced by its norm*" (Aliprantis and Border, 2007, Definition 6.1, p. 228). We let $(\mathcal{V}, \|\cdot\|_{\mathcal{V}})$ be a Banach space of real-valued functions over $\mathcal{X}$, equipped with a given norm, and $(\mathcal{V}^{\mathcal{A}}, \|\cdot\|_{\mathcal{V}^{\mathcal{A}}})$ be a Banach space of functions mapping $\mathcal{A}$ to $\mathcal{V}$. We assume $\mathcal{V}$ contains the value functions of all deterministic stationary policies (all $\pi \in \Pi$), $\mathcal{V}^{\mathcal{A}}$ contains the action-value functions[4] of all $\pi \in \Pi$. Of course, since $\mathcal{A}$ is finite, $\mathcal{V}^{\mathcal{A}}$ can also be identified with the set $\left\{(x,a) \mapsto (V(a))(x) : V \in \mathcal{V}^{\mathcal{A}}\right\}$, which is a set of $\mathcal{X} \times \mathcal{A} \to \mathbb{R}$ functions. For $V \in \mathcal{V}^{\mathcal{A}}$ and $a \in \mathcal{A}$, we will use V^a as an alternate notation to $V(a) = x \mapsto V(x,a)$. Conveniently, $V^a \in \mathcal{V}$. With

[4] The *action-value function* of a policy π is defined by

$$V^\pi(x,a) \doteq \mathbb{E}\left(\sum_{t=1}^{\infty} \gamma^{t-1} R_t \middle| X_0 = x, A_0 = a \right),$$

for $x \in \mathcal{X}$ and $a \in \mathcal{A}$, with $A_t \sim \pi(X_0, A_0, R_1, X_1, \ldots, X_t)$ $(t \geq 1)$. Typically, action-value functions are denoted by the symbol Q^π, but we use V^π for both (state-) value functions and action-value functions, and the distinction will be clear from the context.

a slight abuse of notation, we take $\mathcal{P} \doteq (\mathcal{P}^a)_{a\in\mathcal{A}}$ where $\mathcal{P}^a$ is the $\mathcal{V} \to \mathcal{V}$ right linear operator defined by $(\mathcal{P}^a V)(x) \doteq \mathbb{E}(V(X_{t+1})|\, X_t = x, A_t = a)$ with $X_{t+1} \sim \mathcal{P}(X_t, A_t)$ (we assume that $V \in \mathcal{V}$ implies integrability, so that the integrals are well defined). We also treat $\mathcal{P}$ as a $\mathcal{V} \to \mathcal{V}^{\mathcal{A}}$ linear operator defined by $(\mathcal{P}V)^a \doteq \mathcal{P}(V^a)$ for $a \in \mathcal{A}$ and $V \in \mathcal{V}$. We also assume that the reward function r is an element of $\mathcal{V}^{\mathcal{A}}$, so $r^a(x)$ will be used to denote $r(x, a)$.

The *Bellman return operator* $T_{\mathcal{P}} : \mathcal{V} \to \mathcal{V}^{\mathcal{A}}$, is defined by $T_{\mathcal{P}}V \doteq r + \gamma\mathcal{P}V$. and the so-called *maximum selection* operator $M : \mathcal{V}^{\mathcal{A}} \to \mathcal{V}$ is defined by $(MV)(x) \doteq \max_a V^a(x)$. We can therefore write (1.1.6) in compact form:

$$V^* = MT_{\mathcal{P}}V^* \tag{1.1.7}$$

and we can also define the *greedy operator* $G : \mathcal{V}^{\mathcal{A}} \to \Pi$, which selects the maximizing actions chosen by M:

$$GV(x) \doteq \operatorname*{argmax}_a V^a(x)$$

for all $x \in \mathcal{X}$, with ties broken arbitrarily. As mentioned, under some conditions (*e.g.*, $\gamma < 1$) GTV^* is an optimal policy.

The performance of a policy π will be given by the *policy error*

$$\|V^* - V^\pi\|_{\mathcal{V}},$$

and different choices of the norm $\|\cdot\|_{\mathcal{V}}$ can be made. For example, a common choice is the *supremum norm* $\|V\|_\infty \doteq \sup_{x\in\mathcal{X}}|V(x)|$. In Sections 3.2.3 and 4.4 we will look at policy error bounds with the supremum norm, as well as other choices of $\|\cdot\|_{\mathcal{V}}$.

1.2 Contributions and Dissertation Outline

It is already known that DP allows us find an optimal stationary deterministic policy in an MDP, *i.e.*, it essentially solves the MDP. It can be carried out, for example, via linear programming (de Farias and Van Roy, 2003),

policy iteration, or value iteration (Szepesvári, 2010, Section 2.4, pp. 16–17). However, DP requires that we know r and $\mathcal{P}$, and, even if we do, DP is intractable if the cardinality of $\mathcal{X} \times \mathcal{A}$ is too large. As remarked by Szepesvári (2010), DP is intractable in all but the simplest MDPs, and RL methods (insofar as an optimal policy is being sought) one way or another are *approximately* doing DP in settings where there is only "indirect access" to $\mathcal{P}$ and maybe r, and where $\mathcal{X} \times \mathcal{A}$ may be prohibitively large.

In this dissertation, we present advances in the theoretical understanding of Classification-Based Policy Iteration (CBPI, Farahmand et al., 2014; Lazaric et al., 2016) and Model-Based Reinforcement Learning (MBRL) with factored semi-linear models. Our emphasis is on analyzing *practical* methods, that is, reinforcement learning methods that compute policies efficiently. Evidently, we want practical algorithms to produce effective policies, and in this text we focus on provable effectiveness, namely, theoretical guarantees in the form of policy error bounds.

CBPI falls under the category of so-called Direct Policy Learning (DPL) methods, which also includes, *e.g.*, policy gradient (Sutton et al., 1999), conservative policy iteration (Kakade and Langford, 2002), and classification-based methods for learning non-stationary policies (Langford and Zadrozny, 2003; Bagnell et al., 2003; Langford and Zadrozny, 2005). The notable trait of DPL methods is that they do not necessarily rely on estimating value-functions, but only on value estimates at certain state-action pairs. In contrast, many popular RL methods, *e.g.*, SARSA, Q-learning, (Sutton and Barto, 1998), approximate policy iteration (Scherrer, 2014) and fitted Q-iteration (Antos et al., 2008a), are instances of Generalized Policy Iteration (GPI, Sutton and Barto, 1998, Section 4.6, p. 106), and, as such, rely (at least procedurally) on estimating value functions. While it is not clear whether GPI inherently depends on "accurate" value-function estimates, the existing performance guarantees (policy error bounds) for these methods do degrade when value-function estimates are poor (see, for example Bertsekas, 2012, Proposition 3.1).

It is important to emphasize that estimating value functions is different from estimating values at a given set of state-action pairs—which DPL methods do resort to. Accurate value function estimates should yield good (greedy) policies, however, it is plausible to expect that such estimates not be necessary for constructing effective policies. Therefore, one can anticipate that there will be cases where accurate value function estimates cannot be constructed and GPI may fail, but where DPL can succeed in constructing effective policies. As a motivating example, in SZ-Tetris (Burgiel, 1997) a simple set of features proposed by Bertsekas and Ioffe (1996) is likely not expressive enough for representing value functions (parametrically), as evidenced by poor performance of GPI methods (Bertsekas and Ioffe, 1996; Szita and Szepesvári, 2010). However, the features are still expressive enough for good policies to be represented, as we can see from the performance of local-search methods (Szita and Lörincz, 2006). Because DPL methods may avoid estimating value functions, they hold the promise to leverage the representation in order to learn good policies directly, using value estimates only.

CBPI, at its core, makes use of a classification method, and therefore has the potential to succeed where classifiers can be effective in using value estimates to discriminate better actions (actions with higher value) from sub-optimal ones, and where the representation allows generalizing this "effective action discrimination" across the state space. Indeed, the performance of a CBPI method is inherently tied to the performance of the classification algorithm used, and in order to present policy error bounds for CBPI, we need performance bounds classification.

The two existing analysis of CBPI (Farahmand et al., 2014; Lazaric et al., 2016) rely on classification methods that solve a combinatorial problem, which can often be computationally hard. In contrast, classification methods that require solving convex minimization problems (which can often be done efficiently) enjoy limited or no performance guarantees. Therefore, we only have instances of CBPI that enjoy theoretical guarantees but cannot be

used in practice, or vice-versa. The theoretical analysis of CBPI presented in this dissertation fills this gap, by providing theoretical guarantees for instances of CBPI that can be used in practice (by virtue of the classification methods used).

Our first analysis, therefore is not of reinforcement learning methods, but of cost-sensitive multiclass classification algorithms. In Chapter 2, we look at the popular approach of empirical risk minimization applied to surrogate convex losses. This approach leads to what we informally call *practical* classification methods, that is, methods that are efficient and can be used in practice. We present novel surrogate risk bounds for these "empirical surrogate risk minimizers" and a particular family of cost-sensitive multiclass classification losses. We also present specific results for converting these surrogate risk bounds into true risk bounds. Our surrogate risk bounds allow us to understand, as a special case, the performance of a cost-sensitive generalization of the hinge loss, which has been used in the design of many classification algorithms, including the popular support vector machine (SVM, Hastie et al., 2009, Chapter 12, p. 417).

In Chapter 3, we then present our analysis of CBPI. The central result of this analysis is a policy error bound derived from the work of Lazaric et al. (2016), and which incorporates the results developed in Chapter 2.

In Chapter 4, we present policy error bounds for MBRL methods with factored semi-linear models. MBRL methods have often been regarded as inefficient or not scalable. However, they have recently been the subject of a number of works that have focused on efficient methods with a great potential to be competitive with other ("model-free") methods (Ormoneit and Sen, 2002; Barreto et al., 2011; Grünewälder et al., 2012; Kveton and Theocharous, 2012; Lever et al., 2016). We present an updated version of the work of Ávila Pires and Szepesvári (2016b), with policy error bounds for MBRL methods that use what we call factored semi-linear models. Factored-*linear*-model methods generalize many previously proposed model-based reinforcement learning methods, including the efficient MBRL methods just

mentioned. Although factored semi-linear models are a mild generalization of factored linear models, we are able to increase the scope of the policy error bounds of Ávila Pires and Szepesvári (2016b), which, differently from ours, did not apply to all factored-*linear*-model methods. Moreover, the factored semi-linear model framework is flexible and allows the design of a number of "practical" (*i.e.*, computationally efficient) MBRL methods.

The policy error bounds presented in Chapter 4, besides having an increased scope (*i.e.*, factored semi-linear models), also improve on the previously existing bounds by relaxing conditions and refining the bounds themselves. The bounds allow us to understand the policy error in norms other than the overly-harsh supremum norm, which has, nevertheless, been a common choice for policy error bounds in the RL literature (and, to the best of our knowledge, the only such choice for policy error bounds concerning MBRL methods).

We conclude this dissertation with Chapter 5, where we have a final discussion of CBPI and MBRL.

Chapter 2

Cost-sensitive Multiclass Classification

In this chapter, we present some supervised-learning results, for the problem of cost-sensitive multiclass classification (henceforth simply called classification). We first introduce the classification problem (Section 2.1), as well as Empirical Risk Minimization (ERM) with surrogate losses as a means to solve the classification problem, and the concept of calibration functions (Section 2.2). Calibration functions allow us to convert surrogate risk bounds into true risk bounds for minimizers of the surrogate risk.

The first set of results that we present (Section 2.3) concerns bounds on the true risk, *i.e.*, the original risk in the classification problem, called the cost-sensitive error. The cost-sensitive error generalizes the misclassification probability of cost-insensitive multiclass classification. More specifically, we present calibration functions for a family of novel cost-sensitive surrogate loss based on the work of Zhang (2004) for multiclass classification.

The second set of results (Section 2.4) are surrogate bounds for classifiers obtained by a specific classification method based on a cost-sensitive surrogate loss proposed by Ávila Pires et al. (2013) as a generalization of a cost-insensitive multiclass classification loss proposed by Lee et al. (2004). Surprisingly, to the best of our knowledge, these are the first surrogate risk bounds in the context of cost-sensitive multiclass classification, and the first "calibration-compatible" bounds in the context of (cost-sensitive and

cost-insensitive) multiclass classification. By "calibration-compatible" we mean that we are able to use calibration functions to convert these surrogate risk bounds into true risk bounds.

We conclude this chapter (Section 2.5) with a discussion about extensions of our result, most notably extensions that obtain so-called fast-rates by incorporating the Mammen-Tsybakov noise condition (see Boucheron et al., 2005).

The results in Sections 2.3 and 2.4 set the foundations for extending the analysis of Lazaric et al. (2016) to what we call practical instances of CBPI, that is, instances with classification methods that employ surrogate empirical risk minimization with a convex loss. Interestingly, the extensions discussed in Section 2.5 are quite pertinent to the analysis of CBPI presented by Farahmand et al. (2014), suggesting that their analysis could be extended to practical instances of CBPI, based on our analysis in Chapter 3 and fast-rate extensions of our results in this chapter.

2.1 Problem Definition

In classification (in the formulation used by Ávila Pires et al., 2013) we observe the jointly distributed random variables (X, C) taking values in $\mathcal{X} \times \mathbb{R}^{|\mathcal{Y}|}$ and distributed according to an unknown p. The set $\mathcal{X}$ is a measurable set[1] and the set $\mathcal{Y}$ is a finite set of labels[2], here assumed to be (without loss of generality) $\{1, \ldots, |\mathcal{Y}|\}$. Differently from multiclass classification, or more conventional views of cost-sensitive multiclass classification, in our framework there is no notion of a (random) "correct label" Y. The goal is to find a *classifier* $g : \mathcal{X} \to \mathcal{Y}$ that minimizes the *risk* or *classification cost*:

$$R(g) \doteq \mathbb{E}\left(C_{g(X)}\right),$$

[1] Note that for now we are outside of the reinforcement learning setting outlined in Chapter 1, but in Chapter 3 the set $\mathcal{X}$ will be both the "input space" in classification, and the state space of the MDP.

[2] We use "class" and "label" interchangeably, and the distinction from "class" in the sense of a collection or set (*e.g.*, a hypothesis class) will be clear from context.

or, alternatively, the *excess risk* w.r.t. a "baseline" set of classifiers $\mathcal{G} \subset \mathcal{Y}^{\mathcal{X}}$

$$R(g) - \inf_{g' \in \mathcal{G}} R(g'),$$

whenever $\left|\inf_{g' \in \mathcal{G}} R(g')\right| < \infty$.

Cost-sensitive classification with random costs has been studied by Zadrozny and Elkan (2001); Brefeld et al. (2003); Zadrozny et al. (2003). If the costs are fixed we recover "traditional" cost-sensitive classification (Shalev-Shwartz and Ben-David, 2014, Section 17.2.2, p. 194). In cost-insensitive classification, one starts from the pair (X, Y) of jointly-distributed $\mathcal{X} \times \mathcal{Y}$-valued random variables, where Y is the random "true label" mentioned earlier. Cost-insensitive classification can be seen as a special case of cost-sensitive classification where the costs satisfy $C_k = \mathbb{I}\{Y = k\}$ ($k \in \mathcal{Y}$) w.p. one, where $\mathbb{I}\{\cdot\}$ is the indicator function. In binary classification we simply have $|\mathcal{Y}| = 2$.

In the classification learning problem, we observe a sample[3]

$$S \doteq ((X_1, C_1), \ldots, (X_n, C_n))$$

and our goal is to construct a classifier G depending on S that minimizes, with high probability, the *conditional risk given the sample*:

$$R(G, S) \doteq \mathbb{E}\Big(C_{G(X)} \Big| S\Big),$$

where $((X_1, C_1), \ldots, (X_n, C_n), (X, C)) \sim p^{n+1}$.

Additional notation. We let $\Delta_{|\mathcal{Y}|}$ be the $|\mathcal{Y}|$-dimensional simplex. We abuse notation and let $\Delta_{\mathbb{R}^{|\mathcal{Y}|}}$ denote the set of all probability measures p over $\mathbb{R}^{|\mathcal{Y}|}$ s.t. $|\mathbb{E}(C_k)|$ is well-defined and finite for all $k \in \mathcal{Y}$ when $C \sim p$. We call these distributions *cost distributions*. We will call cost distributions for which the costs are non-negative w.p. one *non-negative-cost distributions*.

[3] Note that a subscripted C will denote both a coordinate of C (C_k) and a cost-vector observed in the sample S (C_i), and these two uses will be easy to distinguish from context. The k-th coordinate of the i-th cost-vector in S will be denoted by $C_{i,k}$.

We let $p_{C|X}$ denote the conditional distribution of C given X, $\mathbf{1}$ and $\mathbf{0}$ denote, respectively, all-ones and all-zeros vectors of appropriate dimension, $\mathbb{N}$ denote the natural numbers, $[n] \doteq \{1, \ldots, n\}$ for $n \in \mathbb{N}$, $[0] \doteq \emptyset$, $a \wedge b \doteq \min(a, b)$ and $a \vee b \doteq \max(a, b)$. For simplicity, we assume that argmin and argmax are singletons (which can be enforced by breaking ties in some arbitrary, fixed way). We also assume that objectives with an argmin do have a minimizer, but it is easy to re-do our derivations with infima and approximate minimizers.

2.2 Empirical Risk Minimization, Surrogate Risk, and Risk Bounds

ERM (Steinwart and Christmann, 2008, p. 8; Shalev-Shwartz and Ben-David, 2014, p. 15), a typical approach to solve classification, prescribes that we take a minimizer $\widehat{G}$ of the *empirical risk*

$$\widehat{R}(g, S) \doteq \frac{1}{n} \sum_{i=1}^{n} C_{i,g(X)}$$

over the set of classifiers $\mathcal{G}$, that is

$$\widehat{G} \doteq \operatorname*{argmin}_{g \in \mathcal{G}} \widehat{R}(g, S).$$

ERM has been widely studied (see, for example, Koltchinskii, 2011; Vapnik, 2013) and enjoys performance guarantees in the form high-probability bounds on $R(\widehat{G}, S)$ (see, *e.g.* Vapnik, 2013, Chapter 3).

The empirical risk is not convex, however: Calculating $\widehat{G}$ is a combinatorial problem and can be computationally hard for typical choices of $\mathcal{G}$ (Höffgen et al., 1995; Steinwart and Christmann, 2008, p. 59 and p. 62). One such typical choice is $\bigcup_{B=0}^{\infty} \mathcal{G}_{\phi,B}$, where

$$\mathcal{G}_{\phi,B} \doteq \left\{ x \mapsto \operatorname*{argmax}_k \langle \phi(x, k), w \rangle : w \in \mathbb{R}^d \text{ s.t. } \|w\| \leq B \right\},$$

is the set of bounded linear classifiers w.r.t. a *feature extractor* $\phi : \mathcal{X} \times \mathcal{Y} \to \mathbb{R}^d$, with $B > 0$ and some norm $\|\cdot\|$.

Following the approach of empirical surrogate-risk minimization (see Steinwart and Christmann, 2008, Chapter 3), we introduce a convex cost-sensitive surrogate loss $L : \mathcal{S} \times \mathbb{R}^{|\mathcal{Y}|} \to \mathbb{R}$, where $\mathcal{S} \subset \mathbb{R}^{|\mathcal{Y}|}$ is a non-empty *set of scores* compatible with the loss. While a number of losses for cost-insensitive multiclass classification have been proposed (see Mason et al., 2000; Crammer and Singer, 2003; Lee et al., 2004; Rifkin and Klautau, 2004; Zhang, 2004; Zou et al., 2006; Gneiting and Raftery, 2007; Liu, 2007; Nock and Nielsen, 2009; Reid and Williamson, 2010; Mroueh et al., 2012; Beijbom et al., 2014; Shi et al., 2015), surrogate losses for cost-sensitive multiclass losses are fewer in number (see, for example Tsochantaridis et al., 2005; Guruprasad and Agarwal, 2012; Ávila Pires et al., 2013; Ramaswamy et al., 2013). As we will see in Section 2.3.3, however, some general principles can be used to generalize cost-insensitive losses to the cost-sensitive case.

As an example, L can be the cost-sensitive generalization (proposed by Ávila Pires et al., 2013) of the loss of Lee et al. (2004)[4], :

$$L^{\mathrm{LLW}}(s, c) \doteq \sum_{k=1}^{|\mathcal{Y}|} c_k \varphi(s_k) \tag{2.2.1}$$

where $\varphi : \mathbb{R} \to \mathbb{R}$ is a convex[5] *score transformation function*, for example, the hinge transformation function $\varphi^{\mathrm{hinge}}(t) \doteq (1 + t)_+$. The score set for L^{LLW} is the set of *sum-to-zero* scores $\mathcal{S}_0 \doteq \left\{ s \in \mathbb{R}^{|\mathcal{Y}|} : \mathbf{1}^\top s = 0 \right\}$. Typical choices of score sets for other losses include itself $\mathbb{R}^{|\mathcal{Y}|}$ (Zhang, 2004), $\mathbb{R}^{|\mathcal{Y}|-1}$ (Mroueh et al., 2012) and $\Delta_{|\mathcal{Y}|}$ (Reid and Williamson, 2010).

We wish to find a *score function* $h : \mathcal{X} \to \mathcal{S}$ that minimizes the *surrogate risk*

$$R_L^{\mathrm{surr}}(h) \doteq \mathbb{E}(L(h(X), C)),$$

or, alternatively, the *excess surrogate risk* s.t. to a baseline set of score functions

[4] The loss $L^{\mathrm{LLW,CI}} : \mathcal{S} \times \mathcal{Y} \to \mathbb{R}$ proposed by Lee et al. (2004) is defined as $L^{\mathrm{LLW,CI}}(s, y) \doteq \sum_{k \neq y} \varphi(s_k)$ and has score set $\mathcal{S}_0$, defined below.

[5] The reader may notice that convexity of φ alone is not enough to ensure that L^{LLW} is convex (in the first argument), because the costs can be negative. For L^{LLW}, assuming that the costs are non-negative w.p. one addresses the convexity issue, and, as we will see in Section 2.3.4, will create only mild limitations.

$\mathcal{H} \subset (\mathbb{R}^{|\mathcal{Y}|})^{\mathcal{X}}$

$$R_L^{\text{surr}}(g) - \inf_{h' \in \mathcal{H}} R_L^{\text{surr}}(h'),$$

whenever $\left|\inf_{h' \in \mathcal{H}} R_L^{\text{surr}}(h')\right| < \infty$. Where the choice of L is clear from the context, we will drop the subscript from the surrogate risks. We can use empirical surrogate-risk minimization, so we take a minimizer $\widehat{H}$ of the *empirical surrogate risk*

$$\widehat{R}^{\text{surr}}(h, S) \doteq \frac{1}{n} \sum_{i=1}^{n} L(h(X_i), C_i)$$

over the set of score functions $\mathcal{H}$, that is

$$\widehat{H} \doteq \operatorname*{argmin}_{h \in \mathcal{H}} \widehat{R}^{\text{surr}}(h, S). \tag{2.2.2}$$

Then L, $\mathcal{S}$ and $\mathcal{H}$ are chosen so that $\widehat{H}$ can be computed efficiently. A number of popular classification algorithms can be seen to perform empirical surrogate-risk minimization—to mention a few, SVMs, ridge regression, the Lasso, logistic regression, and AdaBoost (see Hastie et al., 2009, Table 21.1, and Sections 3.4, 4.4.1 and 10.4).

Even though we attempt to minimize the surrogate risk using ERM, our goal is still to minimize the *true risk*, that is, the classification cost. In order to convert scores into class predictions, we use the *maximum selector* $f : \mathbb{R}^{|\mathcal{Y}|} \to \mathcal{Y}$ defined by

$$f(s) \doteq \operatorname*{argmax}_k s_k.$$

We can also hope to obtain high-probability bounds on the surrogate $R^{\text{surr}}(\widehat{H}, S)$, but how can we obtain high-probability bounds on the *true* risk $R(f \circ \widehat{H}, S)$? One way to do so is to use *calibration functions*, which allow us to convert surrogate risk bounds into true risk bounds. Definition 2.2.1 defines calibration functions as introduced by Steinwart (2007), and Theorem 2.2.2 (due to Bartlett et al., 2006 for the binary case and to Steinwart, 2007 for the general case) shows how they can be used to obtain true risk bounds.

Definition 2.2.1 (Calibration function, Definition 2.7 of Steinwart, 2007)**.** *A function $\delta : (0, \infty) \times \Delta_{\mathbb{R}^{|\mathcal{Y}|}} \to [0, \infty]$ is a calibration function for a surrogate loss L with score set $\mathcal{S}$ if $\delta(\varepsilon, p) > 0$ for all $\varepsilon > 0$ and every cost distribution p, and if, for all $s \in \mathcal{S}$ and $\varepsilon > 0$, the inequality*

$$\mathbb{E}(L(s, C)) - \inf_{s' \in \mathcal{S}} \mathbb{E}(L(s, C)) < \delta(\varepsilon, p)$$

implies that

$$\mathbb{E}\left(C_{f(S)}\right) - \inf_{k \in \mathcal{Y}} \mathbb{E}(C_k) < \varepsilon,$$

when $C \sim p$.

Theorem 2.2.2 (Theorem 2.8 of Steinwart, 2007)**.** *Given a surrogate loss L with score set $\mathcal{S}$, assume that L and $\mathcal{S}$ have a calibration function δ. Assume also that $|\mathbb{E}(C_k | X)|$ exists and is finite w.p. one for all $k \in \mathcal{Y}$ and that $\mathbb{E}(\inf_{s \in \mathcal{S}} L(s, C) | X)$ is measurable. Then, for any $\delta \in (0, 1)$ and $\varepsilon > 0$, the following holds: If*

$$R^{\text{surr}}(\widehat{H}, S) - \inf_h R^{\text{surr}}(h) < \mathbb{E}\left(\delta(\varepsilon, p_{C|X})\right)$$

holds with probability at least $1 - \delta$, then

$$R(f \circ \widehat{H}, S) - \inf_g R(g) < \varepsilon$$

also holds with probability at least $1 - \delta$.

Typically, the calibration function will be easy to invert; for example, for L^{LLW} with φ^{hinge} we can take $\delta(\varepsilon, p) = \varepsilon$ for every non-negative-cost distribution[6] p (Ávila Pires et al., 2013, Table 1). As another example, for L^{LLW} with $\varphi^{\text{squared}}(t) \doteq (1 + t)^2$, if there is a constant c s.t. for every non-negative-cost distribution p s.t. $C \sim p$ we have $\mathbb{E}(\min_k C_k + \max_k C_k) \leq c$, then we can take $\delta(\varepsilon, p) = \frac{\varepsilon^2}{c}$ (Ávila Pires et al., 2013, Table 1). We can then apply Theorem 2.2.2 by starting from a guarantee that

$$R^{\text{surr}}(\widehat{H}, S) - \inf_h R^{\text{surr}}(h) < t$$

[6] In Section 2.3.4 we discuss how to obtain guarantees similar to those of Theorem 2.2.2 for cost distributions when the calibration function is only defined for to non-negative-cost distributions.

for some $t > 0$ with probability at least $1 - \delta$, to obtain that

$$R(f \circ \widehat{H}, S) - \inf_g R(g) < \delta^{-1}(t)$$

with probability at least $1 - \delta$, where $\delta^{-1}(t) \doteq \inf\{\varepsilon : \inf_{p \in \mathcal{P}} \delta(\varepsilon, p) \geq t\}$ for some $\mathcal{P}$ s.t. $p_{C|X} \in \mathcal{P}$ w.p. one. Note that without constraints on $\mathcal{P}$ (for example, the assumption on c above) we may have $\inf_{p \in \mathcal{P}} \delta(\varepsilon, p) = 0$, in which case the true risk bound is vacuous.

It is important to emphasize that the strategy for using Theorem 2.2.2 is inherently non-parametric, that is, it involves bounds where we "compete" against all score functions and classifiers, not just baselines. On the other hand, surrogate risk guarantees often have the form

$$R^{\text{surr}}(\widehat{H}, S) - \inf_{h \in \mathcal{H}} R^{\text{surr}}(h) < t$$

with high-probability, in which case Theorem 2.2.2 gives us

$$R(f \circ \widehat{H}, S) - \inf_g R(g) < \delta^{-1}(t + A^{\text{surr}}(\mathcal{H})),$$

where $A^{\text{surr}}(\mathcal{H})$ is the *surrogate approximation error* of $\mathcal{H}$:

$$A^{\text{surr}}(\mathcal{H}) \doteq \inf_{h \in \mathcal{H}} R^{\text{surr}}(h) - \inf_{h'} R^{\text{surr}}(h),$$

where the second infimum is taken over all (measurable) score functions. In such a non-parametric setting, one should trade off t and $\mathcal{H}$ in such a way that $\delta^{-1}(t + A^{\text{surr}}(\mathcal{H}))$ is minimized (see Steinwart and Christmann, 2008, p. 8).

In Section 3.2, we will apply Theorem 2.2.2 following the steps outlined above in order to obtain true risk bounds for classifiers computed by a classification methods used in CBPI. In the remainder of this chapter, we will take a closer look at existing and novel calibration functions for specific surrogate losses, as well as novel surrogate risk bounds for empirical surrogate-risk minimization with L^{LLW} as the surrogate loss.

2.3 Calibration Functions

In this section we present different cost-sensitive multiclass classification surrogate loss functions and their respective calibration functions.

We start by instantiating the framework of Steinwart (2007) for the specific case of classification in Section 2.3.1, as done by Ávila Pires et al. (2013). In Section 2.3.2, we present a calibration function for L^{LLW} derived by Ávila Pires et al. (2013).

In Section 2.3.3, we present L^{Zhang}, a novel cost-sensitive generalization of a multiclass loss studied by Zhang (2004) (which, in turn generalizes the multiclass logistic regression loss and a number of so-called "decoupled" losses). We reuse results of Ávila Pires and Szepesvári (2016a) for the cost-insensitive decoupled losses proposed by Zhang (2004) to derive calibration functions for decoupled versions of L^{Zhang}.

In order to be convex (and for the calibration functions to hold), the losses L^{LLW} and L^{Zhang} both require the costs to be non-negative. In Section 2.3.4, we show that this limitation can be partially overcome: We can obtain true risk bounds with real-valued costs by shifting these costs to be non-negative and then using L^{LLW} or L^{Zhang}.

Finally, for the sake of completeness, we also show how to construct cost-sensitive losses from cost-insensitive losses that is similarly to the strategy employed by Zadrozny et al., 2003 to solve cost-sensitive binary classification using cost-insensitive binary classification algorithms. We also show how to obtain calibration functions for these cost-sensitive losses, from calibration functions of the cost-insensitive losses. While this approach is relatively straightforward, we show that it can produce losses with poor scaling, as well as calibration functions that introduce undesirable factors into the true risk bounds.

2.3.1 General Calibration Functions

Steinwart (2007) defined a function $\delta_{\max} : [0, \infty) \times \Delta_{\mathbb{R}^{|\mathcal{Y}|}} \to [0, \infty)$ that

depends on the given surrogate loss and constitutes a key notion for calibration functions. $\delta_{\max}$ is special because no calibration function for the given surrogate loss is larger than $\delta_{\max}$ (Steinwart, 2007). Hence, if $\delta_{\max}$ is a calibration function, it is called the *maximum calibration function*. Conveniently, any positive lower bound to the maximum calibration function is also a calibration function, which is a useful fact for understanding and calculating calibration functions for specific surrogate losses.

In order to define $\delta_{\max}$, we must define three useful concepts (see Definition 2.3.1): The set of scores in $\mathcal{S}$ whose maximum coordinate is j ($\mathcal{M}(\mathcal{S}, j)$), the set of scores that give ε-sub-optimal class predictions ($\mathcal{T}(\mathcal{S}, \varepsilon, c)$), and the set of ε-sub-optimal maximum-probability indices ($\mathcal{J}(\varepsilon, c)$). A score $s \in \mathcal{S}$ is *ε-sub-optimal* for a given $c \in \mathbb{R}^{|\mathcal{Y}|}$ if $c_{f(s)} - \min_k c_k \geq \varepsilon$. Otherwise, s is *ε-optimal*.

Definition 2.3.1. *Given a set of scores $\mathcal{S} \subset \mathbb{R}^{|\mathcal{Y}|}$ let, for $\varepsilon \geq 0$ and $c \in \mathbb{R}^{|\mathcal{Y}|}$*

$$\mathcal{M}(\mathcal{S}, j) \doteq \left\{ s \in \mathcal{S} : s_j = \max_k s_k \right\}, \quad \mathcal{T}(\mathcal{S}, \varepsilon, c) \doteq \bigcup_{j : c_j - \min_k c_k \geq \varepsilon} \mathcal{M}(\mathcal{S}, \varepsilon, j),$$

$$\mathcal{J}(\varepsilon, c) \doteq \operatorname*{argmax}_j \left\{ c_j : c_j - \min_k c_k \geq \varepsilon \right\}.$$

We will override notation and use R^{surr} to denote the *pointwise surrogate risk* $R^{\text{surr}} : \mathcal{S} \times \Delta_{|\mathcal{Y}|} \to \mathbb{R}$ for a surrogate loss $L : \mathcal{S} \times \mathcal{Y} \to \mathbb{R}$ with $\mathcal{S} \subset \mathbb{R}^{|\mathcal{Y}|}$, defined by

$$R_L^{\text{surr}}(s, p) \doteq \mathbb{E}(L(s, C)),$$

where $C \sim p$ (cf. Definition 2.2.1). As with the surrogate risk, where the choice of L is clear from context, we will drop the subscript from the pointwise surrogate risk.

In Definition 2.3.2, we present $\delta_{\max}(\varepsilon, p)$, which is the difference between the smallest surrogate risk of any ε-suboptimal score and the optimal surrogate risk, when $C \sim p$. If any score has surrogate risk closer to the optimal surrogate risk than $\delta_{\max}(\varepsilon, p)$, the score must be ε-optimal w.r.t. $\mathbb{E}(C)$. Confronting this fact with Definition 2.2.1, we see that if $\delta_{\max}$ is

positive for all $\varepsilon > 0$, then it is a calibration function. It is, however, a calibration function only for the pointwise surrogate risk defined in terms of cost distributions p.

Definition 2.3.2. *Given a surrogate loss L with score set $\mathcal{S}$, let*

$$\delta_{\max}(\varepsilon, p) \doteq \inf_{s \in \mathcal{T}(\mathcal{S}, \varepsilon, \mathbb{E}(C))} R^{\text{surr}}(s, p) - \inf_{s \in \mathcal{S}} R^{\text{surr}}(s, p).$$

If $\delta_{\max}(\varepsilon) > 0$ for all $\varepsilon > 0$, then it is called the maximum calibration function.

In the binary case, we can expect $\delta_{\max}$ to be easy to calculate, but this may not be the case in the multiclass case. Thus, one strategy to obtain calibration functions for multiclass losses is to lower-bound $\delta_{\max}$ by a function that resembles $\delta_{\max}$ in the binary case. We will use the function δ_{binary} given in Definition 2.3.3 to do so. With some assumptions on the surrogate loss, it is possible to show that $\delta_{\max}$ and δ_{binary} coincide in the binary case for non-negative-cost distributions (see Ávila Pires and Szepesvári, 2016a, Proposition 15).

Definition 2.3.3. *Given a surrogate loss L with score set $\mathcal{S}$, let*

$$\delta_{\text{binary}}(\varepsilon, p) \doteq \inf_{\substack{s \in \mathcal{S}: \\ s_1 = s_2}} R^{\text{surr}}(s, p') - \inf_{s \in \mathcal{S}} R^{\text{surr}}(s, p'),$$

where in the right-hand side $\mathcal{S} \subset \mathbb{R}^2$ and p' is the distribution of the random variable $(C_{j_0}, C_{j_\varepsilon})$, with $j_\varepsilon \in \mathcal{J}(\varepsilon, \mathbb{E}(C))$ and $j_0 \in \mathcal{J}(0, \mathbb{E}(C))$.

2.3.2 Calibration Functions for L^{LLW}

Ávila Pires et al. (2013) proposed L^{LLW}—defined in (2.2.1)—as a cost-sensitive generalization of the loss of Lee et al. (2004), and lower-bounded $\delta_{\max}$ by δ_{binary}, as shown in Theorem 2.3.4. Ávila Pires et al. (2013, Table 1) also calculated δ_{binary} for different choices of φ. Theorem 2.3.4 only applies to non-negative-cost distributions, but, as mentioned earlier, this limitation will be addressed in Section 2.3.4.

Theorem 2.3.4 (Theorem 2.2, Ávila Pires et al., 2013). *Consider $L = L^{\mathrm{LLW}}$ with φ convex and $\mathcal{S} = \mathcal{S}_0$. For every non-negative-cost distribution p, if*

$$\inf_{s \in \mathcal{S}} R^{\mathrm{surr}}(s, p) > -\infty,$$

then for all $\varepsilon > 0$ we have

$$\delta_{\max}(\varepsilon, p) \geq \delta_{\mathrm{binary}}(\varepsilon, p'),$$

where p' is the distribution of the random variable $(C_{j_0}, C_{j_\varepsilon})$, with $j_\varepsilon \in \mathcal{J}(\varepsilon, \mathbb{E}(C))$ and $j_0 \in \mathcal{J}(0, \mathbb{E}(C))$ (breaking ties arbitrarily). Moreover, the above holds with equality when $|\mathcal{Y}| = 2$.

2.3.3 Calibration Functions for L^{Zhang}

Zhang (2004) investigated, among other surrogate losses, the loss

$$L^{\mathrm{Zhang,CI}}(s, y) \doteq \psi(s_y) + F\left(\sum_{k=1}^{|\mathcal{Y}|} \varphi(s_k)\right),$$

with $\mathcal{S} = \mathbb{R}^{|\mathcal{Y}|}$, ψ non-decreasing, and with ψ, F and φ chosen so that $L^{\mathrm{Zhang,CI}}$ is convex. As pointed out by Zhang (2004), $L^{\mathrm{Zhang,CI}}$ generalizes, for example, the multiclass logistic regression loss (obtained by taking $\psi(t) = t$, $F(t) = \ln t$ and $\varphi(t) = \varphi^{\exp}(t) \doteq e^t$). Apart from the difference in score sets, we can also recover L^{LLW} in the cost-insensitive setting from $L^{\mathrm{Zhang,CI}}$ (by taking $\psi(t) = -\varphi(t)$ and $F(t) = t$). Moreover, if $F(t) = t$, then the loss is said to be *decoupled*, in the sense that the surrogate risk can be written as a summation over $k \in \mathcal{Y}$ with unconstrained scores.

Our generalization of $L^{\mathrm{Zhang,CI}}$ to the cost-sensitive case, called L^{Zhang}, is given in Definition 2.3.5. We use the surrogate risk as the basis for generalizing the loss: In the cost-insensitive case, with $Y \sim p$ and $p \in \Delta_{\mathcal{Y}}$, we have

$$\mathbb{E}\left(L^{\mathrm{Zhang,CI}}(s, Y)\right) = \sum_{k=1}^{|\mathcal{Y}|} p_k \psi(s_k) + F\left(\sum_{k=1}^{|\mathcal{Y}|} \varphi(s_k)\right),$$

and we would like that if $s \in \mathcal{S}$ minimizes $s' \mapsto R^{\mathrm{surr}}_{L^{\mathrm{Zhang}}}(s', p)$, then $p_{f(s)} = \max_k p_k$ holds. It is natural, therefore, that in L^{Zhang} we have a notion of

"gain" multiplying each $\psi(s_k)$ term. Using a negative cost $-c_k$ as gain is the first idea that comes to mind, but we instead use $\max_{k'} c_{k'} - c_k$. Because this notion of "gain" is non-negative, we will able to re-use some results by Ávila Pires and Szepesvári (2016a) concerning $L^{\text{Zhang,CI}}$ when computing calibration functions for L^{Zhang}. The term multiplying $F\left(\sum_{k=1}^{|\mathcal{Y}|} \varphi(s_k)\right)$ is chosen so that it dominates all the "gains" (as 1 dominates all p_k) and helps ensure that $\inf_{s\in\mathcal{S}} L^{\text{Zhang}}(s, c) > -\infty$ for all $c \in \mathbb{R}^{|\mathcal{Y}|}$. This lower-boundedness is important so that $\delta_{\max}$ is well-defined. Other options for dominating terms are available: For example, $\sum_{k=1}^{|\mathcal{Y}|}(\max_{k'} c_{k'} - c_k)$, which is explored in Section 2.3.5.

Definition 2.3.5. *The loss* $L^{\text{Zhang}} : \mathcal{S} \times \mathcal{Y} \times \mathbb{R}^{|\mathcal{Y}|} \to \mathbb{R}$ *is defined as*

$$L^{\text{Zhang}}(s, c) \doteq \sum_{k=1}^{|\mathcal{Y}|} (\max_{k'} c_{k'} - c_k)\psi(s_k) + (\max_{k'} c_{k'})\, F\left(\sum_{k=1}^{|\mathcal{Y}|} \varphi(s_k)\right),$$

and has score set $\mathcal{S} = \mathbb{R}^{|\mathcal{Y}|}$.

We can see that L^{Zhang} indeed generalizes $L^{\text{Zhang,CI}}$, by taking $c_k = \mathbb{I}\{k \neq y\}$, where y is the correct class.

In Theorem 2.3.6 we lower-bound $\delta_{\max}$ by δ_{binary} for L^{Zhang} with ψ non-decreasing, $F(t) = t$, and φ convex, and score set $\mathcal{S} = \mathbb{R}^{|\mathcal{Y}|}$. Both the form and the proof techniques used are similar for Theorems 2.3.4 and 2.3.6. In particular, Theorem 2.3.6 also only applies to non-negative-cost distributions.

Theorem 2.3.6. *Consider* $L = L^{\text{Zhang}}$ *convex with* ψ *non-decreasing and* $F(t) = t$. *For every non-negative-cost distribution* p, *if*

$$\inf_{s\in\mathcal{S}} R^{\text{surr}}(s, p) > -\infty,$$

then for all $\varepsilon > 0$ *we have*

$$\delta_{\max}(\varepsilon, p) \geq \delta_{\text{binary}}(\varepsilon, p'),$$

where p' *is the distribution of the random variable* $(C_{j_0}, C_{j_\varepsilon})$, *with* $j_\varepsilon \in \mathcal{J}(\varepsilon, -\mathbb{E}(C))$ *and* $j_0 \in \mathcal{J}(0, -\mathbb{E}(C))$ *(breaking ties arbitrarily). Moreover, the above holds with equality when* $|\mathcal{Y}| = 2$.

Proof. See Appendix A.1.1, page 100. □

2.3.4 Relaxing the Assumption of Non-negative Costs

As discussed earlier, Theorems 2.3.4 and 2.3.6 require the costs to be non-negative w.p. one. Proposition 2.3.7 shows us how to convert surrogate risk bounds into true risk bounds using calibration functions that requires non-negative-cost distributions. More specifically, Proposition 2.3.7 instructs to shift the costs to make them non-negative, and then use the surrogate loss whose calibration function requires non-negative costs.

Proposition 2.3.7. *Consider a cost-sensitive surrogate loss L with score set $\mathcal{S}$. Assume that there is a calibration function δ for $L : \mathcal{S} \times \mathbb{R}^{|\mathcal{Y}|} \to \mathbb{R}$ and any non-negative-cost distribution p. Then for all $s \in \mathcal{S}$, $\varepsilon > 0$ and any cost distribution p,*

$$R^{\text{surr}}(s, p') - \inf_{s' \in \mathcal{S}} R^{\text{surr}}(s', p') < \delta(\varepsilon, p')$$

implies that

$$R(s, p) - \inf_{s' \in \mathcal{S}} R(s', p) < \varepsilon,$$

where p' is the distribution of the random variable $C - \mathbf{1}(\min_k C_k)$, and $C \sim p$.

Proof. See Appendix A.1.1, page 101. □

So far, we have reported a calibration function for L^{LLW}, established one for L^{Zhang}, and shown how to sidestep the requirement of these calibration functions that the cost distribution be a non-negative-cost distribution. We will conclude this section with a presentation on a different and simple strategy to construct cost-sensitive surrogate losses and obtain calibration functions for them.

2.3.5 Calibration Functions by Reduction to Cost-Insensitive Classification

An alternative way to obtain calibration functions for cost-sensitive classification is to use losses that effectively reduce cost-sensitive classification

to cost-insensitive classification. This strategy was used by Zadrozny et al. (2003) to solve cost-sensitive classification problems using cost-insensitive classification algorithms. This reduction allows us to easily use calibration functions for the cost-insensitive loss to construct calibration functions for the cost-sensitive loss.

To perform the reduction, we are given a cost-insensitive surrogate loss $L : \mathcal{S} \times \mathcal{Y} \to \mathbb{R}$ and we define the loss $L^{\text{Red}} : \mathcal{S} \times \mathbb{R}^{|\mathcal{Y}|} \to \mathbb{R}$

$$L^{\text{Red}}_{L,u}(s, c) \doteq \sum_{k=1}^{|\mathcal{Y}|} (u(c) - c_k) L(s, k), \tag{2.3.1}$$

where $u : \mathbb{R}^{|\mathcal{Y}|} \to \mathbb{R}$ is an *coordinate-wise upper-bound function* satisfying $u(c) \geq \max_k c_k$ for all $c \in \mathbb{R}^{|\mathcal{Y}|}$. For convenience, we will use the shorthand $\overline{c} \doteq \sum_{k=1}^{|\mathcal{Y}|} (u(c) - c_k)$. We can see that taking $L = L^{\text{LLW,CI}}$ (resp. $L = L^{\text{Zhang,CI}}$) and constructing $L^{\text{Red}}_{L,u}$ does *not* give us the same surrogate risk as L^{LLW} (resp. L^{Zhang}). Indeed, taking $L = L^{\text{Zhang,CI}}$ gives us

$$\begin{aligned} R^{\text{surr}}_{L^{\text{Red}}_{L,u}}(s, p) = &\sum_{k=1}^{|\mathcal{Y}|} (\max_{k'} c'_k - c_k) \psi(s_k) \\ &+ \left(\sum_{k=1}^{|\mathcal{Y}|} (\max_{k'} c'_k - c_k) \right) F\left(\sum_{k=1}^{|\mathcal{Y}|} \varphi(s_k) \right). \end{aligned}$$

In addition, the scale of $L^{\text{Red}}_{L,u}(s, c)$ can be undesirably sensitive to large $|\mathcal{Y}|$ and large costs. For example, with $L^{\text{Red}}_{L,u}$ obtained from $L = L^{\text{LLW,CI}}$ we have

$$L^{\text{Red}}_{u}(\mathbf{0}, c) = (|\mathcal{Y}| - 1)\overline{c}\varphi(0).$$

In contrast,

$$L^{\text{LLW}}(\mathbf{0}, c) = \overline{c}\varphi(0).$$

Poor scaling is undesirable for risk bounds, since the scale will usually appear as a constant factor (for example, cf. the scale of L^{LLW} and Lemma 2.4.8 in Section 2.4).

As an advantage of using a "reduction loss", it is straightforward to construct a calibration function for $L^{\text{Red}}_{L,u}$ from a calibration function for L, as shown in Theorem 2.3.8. For example, Theorem 2.3.8 with $L = L^{\text{LLW,CI}}$ and

φ^{hinge} gives us $\delta(\varepsilon, p) = \varepsilon$. However, using Theorem 2.3.8 for other losses may introduce additional constant factors during the bound conversion: For example, if $\delta(\varepsilon, q) = \frac{\varepsilon^2}{2}$ (for example, for $L^{\text{LLW,CI}}$ with φ^{squared}, as seen in Table 1 of Ávila Pires et al., 2013), we get an extra factor of $\sqrt{\bar{c}}$ in the true risk bound.

Theorem 2.3.8. *Let L be a cost-insensitive surrogate loss and $L^{\text{Red}}_{L,u}$ be the corresponding cost-sensitive loss for reduction, as given in* (2.3.1)*. If δ is a calibration function for L, then for any cost-distribution p, the function $(\varepsilon, p) \mapsto \bar{c}\delta\left(\frac{\varepsilon}{\bar{c}}, q\right)$ is a calibration function for $L^{\text{Red}}_{L,u}$ and the cost-distribution p, where*

$$\bar{c} = |\mathcal{Y}|\mathbb{E}(u(C)) - \sum_{k=1}^{|\mathcal{Y}|} \mathbb{E}(C_k),$$

$C \sim p$ and $q \in \Delta_{|\mathcal{Y}|}$ is given by $q_k = \frac{1}{\bar{c}}(\mathbb{E}(u(C)) - \mathbb{E}(C_k))$ $(k \in \mathcal{Y})$.

Proof. See Appendix A.1.1, page 102. □

2.4 Surrogate Risk Bounds

In this section, we present surrogate risk bounds for empirical risk minimizers of the cost-sensitive multiclass classification surrogate loss L^{LLW} (scaled by $\frac{1}{|\mathcal{Y}|}$). As we will discuss in Section 2.4.3, there are no surrogate risk bounds that apply to cost-sensitive multiclass classification methods.

This section's results are part of our effort toward an analysis of CBPI with practical classification methods: Besides having the means to convert surrogate risk bounds into true risk bounds (namely, calibration functions), it is imperative that we establish surrogate risk bounds for some cost-sensitive classification methods. Therefore, the setup of the classification learning problem in this section will be slightly more general than the one presented in Section 2.1, and substantially similar to the one considered in Section 3.2.

To obtain the risk bounds, we will use a usual strategy for similar bounds in cost-insensitive classification (and, more generally, the concentration of

some empirical processes, Pollard, 1984): Bound the deviation between the empirical surrogate risk of an individual score function and its surrogate risk, take a union-bound over an appropriately chosen covering of the set of all score functions, and bound the size of the covering. A similar strategy involves a Rademacher complexity appearing explicitly (following Bartlett and Mendelson, 2002, Proof of Theorem 5, p. 228). A Rademacher complexity does appear in the proofs of Lemmas 2.4.9 and 2.4.12, and we further upper-bound it terms of covering numbers of the hypothesis class, with the hope that such upper-bounding will be instructive, and will address some of the challenges related to bounding surrogate risks in the context of cost-sensitive multiclass classification.

The novel surrogate risk bounds for cost-sensitive classification presented here are given in terms of ∞/∞-norm covering numbers of the set of score functions restricted to the sample (Lemma 2.4.9), and in terms of Frobenius-norm covering numbers (Lemma 2.4.12). We also bound these covering numbers for well-known classes of linear score functions, and instantiate the respective surrogate risk bounds. We start by introducing the cost-sensitive classification setting for our bounds (Section 2.4.1), then we present our results (Section 2.4.2), and last we discuss related work (Section 2.4.3). Extensions of the results in this section will be discussed in Section 2.5.

2.4.1 A Variant of the Classification Learning Problem

We will introduce a variant of the classification learning problem with $m \in \mathbb{N}$ cost observations per instance X_i. This variant is strongly related to the setting considered by Lazaric et al. (2016) for their analysis of CBPI, and will allow us to apply the surrogate risk bounds developed in this section in order to obtain our performance guarantees for CBPI in Section 3.2. With an abuse of notation, we will override S and $\widehat{H}$ defined in Section 2.1.

Fix a surrogate loss $L : \mathcal{S} \times \mathbb{R}^{|\mathcal{Y}|} \to \mathbb{R}$ where $\mathcal{S} \subset \mathbb{R}^{|\mathcal{Y}|}$ is its set of scores. We are given a sample

$$S \doteq ((X_1, C'_{1,1}, \ldots, C'_{1,m}), \ldots, (X_n, C'_{n,1}, \ldots, C'_{n,m})), \tag{2.4.1}$$

with each $X_i \in \mathcal{X}$ and $C'_{i,j} \in \mathbb{R}^{|\mathcal{Y}|}$; the statistical nature of S will be detailed later in Assumption 2.4.2.

We want to calculate $T_{L,\mathcal{H}}(n,\delta)$ for a $\mathcal{H} \subset (\mathbb{R}^{|\mathcal{Y}|})^{\mathcal{X}}$ so that for all positive δ small enough we have, with probability at least $1-\delta$,

$$R_L^{\text{surr}}(\widehat{H},S) - \inf_{h\in\mathcal{H}} R_L^{\text{surr}}(h) \leq T_{L,\mathcal{H}}(n,\delta),$$

where[7]

$$\widehat{H} \doteq \operatorname*{argmin}_{h\in\mathcal{H}} \frac{1}{nm}\sum_{i=1}^{n}\sum_{j=1}^{m} L(h(X_i),C'_{i,j}) \tag{2.4.2}$$

(cf. $\widehat{H}$ in (2.2.2), Section 2.1).

Our first result, Proposition 2.4.1, can be easily obtained by looking at the proof of Theorem 5 of Bartlett and Mendelson (2002). Proposition 2.4.1 gives us an excess risk bound provided that we can: i) bound the deviation between the empirical surrogate risk of any individual score function in $\mathcal{H}$ and its surrogate risk (the second term in the right-hand side of the bound in Proposition 2.4.1); and ii) bound the same deviation for all score functions in $\mathcal{H}$ simultaneously (the third term in right-hand side of the bound in Proposition 2.4.1).

Proposition 2.4.1 (Adapted from the proof Theorem 5 of Bartlett and Mendelson, 2002)**.** *For any surrogate loss $L : \mathcal{S}\times\mathbb{R}^{|\mathcal{Y}|}\to\mathbb{R}$ and $h\in\mathcal{H}$,*

$$\begin{aligned} R_L^{\text{surr}}(\widehat{H},S) \leq\ & R_L^{\text{surr}}(h,S) \\ & + \left|\widehat{R}_L^{\text{surr}}(h,S) - R_L^{\text{surr}}(h,S)\right| + \sup_{h'\in\mathcal{H}}\left|\widehat{R}_L^{\text{surr}}(h',S) - R_L^{\text{surr}}(h',S)\right| \end{aligned}$$

holds with probability one.

The rest of this section is devoted to bounding the two terms in the second line of the inequality in Proposition 2.4.1. There is a large number of settings we can study, simply by making different choices for

1. the statistical nature of the sample S,

[7] In our results, we will ensure that $\inf_{h\in\mathcal{H}} R_L^{\text{surr}}(h) > -\infty$, so that the excess surrogate risk is well-defined.

2. the surrogate loss L that we minimize, and its components (*e.g.*, φ for L^{LLW}, or ψ, F and φ for L^{Zhang}),

3. the set of score functions $\mathcal{H}$.

We will present risk bounds in the following setting:

1. The sample S satisfies Assumption 2.4.2 [8].

   Assumption 2.4.2 (Statistical properties of the sample)**.** *The random variables $X_1, \ldots, X_n$ are i.i.d. For all $i \in [n]$, $j \in [m]$, $k \in \mathcal{Y}$*

   1. *$(C'_{i,1,k}, \ldots, C'_{i,m,k})$ are conditionally independent given X_i*
   2. *$(C'_{1,j,k}, \ldots, C'_{n,j,k})$ are independent*
   3. *$(X_i, C'_{i,j})$ and (X, C) are i.i.d.*

   Furthermore, there exist non-negative constants $C_{\max}$ and $C'_{\max}$ s.t., with probability one, $\max_k C_k \in [0, C_{\max}]$ *and* $\max_k \mathbb{E}(C_k | X) \in [0, C'_{\max}]$.

2. We use the *scaled* surrogate loss $(s, c) \mapsto \frac{1}{|\mathcal{Y}|} L^{\mathrm{LLW}}(s, c)$ (with an abuse of notation, we will call this loss $\frac{1}{|\mathcal{Y}|} L^{\mathrm{LLW}}$), with score transformation function $\varphi : \mathbb{R} \to \mathbb{R}$ satisfying Assumptions 2.4.3 and 2.4.4. Assumption 2.4.4 ensures that φ is Lipschitz in a specific interval (which later will be the range of the scores). Lipschitzness in an interval (as opposed to Lipschitzness in $\mathbb{R}$) is used so that our results apply to more choices of φ, including $\varphi^{\exp}$, for example. For the remainder of Section 2.4, the surrogate risks and empirical surrogate risks will all be defined w.r.t. $\frac{1}{|\mathcal{Y}|} L^{\mathrm{LLW}}$, unless subscripted to denote otherwise.

   Assumption 2.4.3 (φ is non-negative-finite-valued)**.** *We have* $\inf_t \varphi(t) \geq 0$ *and* $\sup_{t' \leq t} \varphi(t') < \infty$ *for all* $t \in \mathbb{R}$.

[8] We point out that $(C'_{i,j,1}, \ldots, C'_{i,j,|\mathcal{Y}|})$ (for any $i \in [n]$ and $j \in [m]$) need not be independent.

Assumption 2.4.4 (φ is Lipschitz in an interval)**.** *The function* $\mathrm{Lip}_\varphi : [0,\infty) \to [0,\infty]$ *upper-bounds the* Lipschitz constant of φ in the interval $[-T,T]$ *for all* $T \in \mathbb{R}$*:*

$$\mathrm{Lip}_\varphi(T) \geq \sup_{t,t' \in [-T,T]: t \neq t'} \frac{|\varphi(t) - \varphi(t')|}{|t - t'|}.$$

3. Our risk bounds given in terms of covering numbers will hold for any $\mathcal{H} \subset (\mathbb{R}^{|\mathcal{Y}|})^{\mathcal{X}}$ with bounded scores (see Assumption 2.4.5), but we will only present covering-number bounds for *linear score functions* (with bounded weights, w.r.t. a feature extractor) from

$$\mathcal{H}_{\phi,B} \doteq \left\{(x,k) \mapsto \langle \varphi(x,k), w \rangle : w \in (\mathbb{R}^d, \|\cdot\|) \text{ s.t. } \|w\| \leq B\right\}$$

for a norm $\|\cdot\|$ on $\mathbb{R}^d$ with dual $\|\cdot\|_*$, $B \geq 0$ and a feature extractor $\phi : \mathcal{X} \times \mathcal{Y} \to (\mathbb{R}^d, \|\cdot\|_*)$ satisfying Assumption 2.4.6. In particular, if ϕ satisfies Assumption 2.4.5 and φ is non-decreasing, Assumption 2.4.5 is satisfied with $\varphi_{\max,\mathcal{H}_{B,\phi}} = \varphi(BB_*)$. If φ is finite-valued, Assumption 2.4.5 is an assumption on $\mathcal{H}$ only, but at our convenience we use a constant that bounds the transformed scores.

Assumption 2.4.5 (Scores are bounded)**.** *There exists a constant* $\varphi_{\max,\mathcal{H}}$ *s.t.*

$$\sup_{h \in \mathcal{H}, x \in \mathcal{X}, k \in \mathcal{Y}} \varphi(h(x)_k) \leq \varphi_{\max,\mathcal{H}}.$$

Assumption 2.4.6 (Feature vectors have bounded norm)**.** *There exists* $B_* \geq 0$ *s.t.* $\sup_{x \in \mathcal{X}, k \in \mathcal{Y}} \|\phi(x,k)\|_* \leq B_*$.

The structure we chose for the costs will allow us to analyze CBPI in the setting studied by Lazaric et al. (2016) (detailed in Section 3.2). As a special case, with $m = 1$ we recover our original classification learning problem from Section 2.1 (although taking $m = 1$ in our bounds ahead will give slightly worse constant factors than if we use $m = 1$ in the analysis from the outset). The surrogate loss L^{LLW} is amenable to our analysis due to its simple structure. Moreover, Theorem 2.3.4 gives us calibration functions

for this loss if we impose some additional assumptions on $\mathcal{H}$ to ensure that scores always sum to zero. As we discuss in Section 2.5, it would be interesting to generalize our bounds beyond our specific choice of surrogate loss.

To conclude these preliminaries, we present some additional definitions. For $p \geq 1$ the *p-norm* over $\mathbb{R}^d$ is defined by

$$\|v\|_p^p \doteq \sum_{l=1}^{d} |v_l|^p,$$

and $\|v\|_\infty \doteq \max_l |v_l|$ for $v \in \mathbb{R}^d$. Definition 2.4.7 introduces coverings in vector spaces equipped with a semi-norm.

Definition 2.4.7. *Given a vector space $\mathcal{F}$ and a semi-norm $\|\cdot\|_{\mathcal{F}}$ over $\mathcal{F}$, we say that a set $\mathcal{C} \subset \mathcal{F}' \subset \mathcal{F}$ is an* ε-covering *of $\mathcal{F}'$ in $\|\cdot\|_{\mathcal{F}}$ if*

$$\sup_{f' \in \mathcal{F}'} \inf_{f \in \mathcal{C}} \left\| f - f' \right\|_{\mathcal{F}} < \varepsilon.$$

The $\mathcal{F}$-norm-ε-covering number *of $\mathcal{F}'$ is defined for all $\varepsilon > 0$ as*

$$N_{\|\cdot\|_{\mathcal{F}}}(\varepsilon, \mathcal{F}') \doteq \inf\{|\mathcal{C}| : \mathcal{C} \text{ is an } \varepsilon\text{-covering of } \mathcal{F}' \text{ in } \|\cdot\|_{\mathcal{F}}\}.$$

In this section, we are interested in $\|\cdot\|$-covering numbers of the restriction of score functions to n-sized samples, denoted by $N_{\|\cdot\|} : \mathbb{R}_+ \times \mathbb{N} \times (\mathbb{R}^{|\mathcal{Y}|})^{\mathcal{X}} \to \mathbb{N} \cup \{\infty\}$ (the $\|\cdot\|$ is a norm over $\mathbb{R}^{n \times |\mathcal{Y}|}$) and defined by

$$N_{\|\cdot\|}(\varepsilon, n, \mathcal{H}) \doteq \sup_{(x_1,\ldots,x_n) \in \mathcal{X}^n} N_{\|\cdot\|}(\varepsilon, \{(i,k) \mapsto h(x_i)_k : h \in \mathcal{H}\}). \tag{2.4.3}$$

We also define $\varphi \circ \mathcal{H} \doteq \{\varphi \circ h : h \in \mathcal{H}\}$. We will only consider two norms over $\mathbb{R}^{n \times |\mathcal{Y}|}$, the max-norm and the Frobenius norm, defined respectively by

$$\|v\|_{\infty/\infty} \doteq \max_{i \in [n], k \in \mathcal{Y}} |v_{i,k}|,$$

$$\|v\|_F \doteq \sqrt{\sum_{i=1}^{n} \sum_{k=1}^{|\mathcal{Y}|} (v_{i,k})^2},$$

for $v \in \mathbb{R}^{n \times |\mathcal{Y}|}$, and we will write $N_{\infty/\infty} \doteq N_{\|v\|_{\infty/\infty}}$ as well as $N_F \doteq N_{\|v\|_F}$.

2.4.2 Risk Bounds

Our first result, Lemma 2.4.8, gives an upper-bound on how fast the empirical surrogate risk concentrates around the surrogate risk, with high probability. We need to account for the fact that for each $i \in [n]$ the cost estimates $C'_{i,1}, \ldots, C'_{i,m}$ are not independent, albeit conditionally independent given X_i. To do so, we follow the strategy used by Lazaric et al. (2016, Lemmas 3 and 4): We first study the concentration of the empirical risk around its expectation conditioned on $X_1, \ldots, X_n$, then we study the concentration of this quantity around the surrogate risk. However, differently from Lazaric et al. (2016), who upper-bound the costs uniformly (using only $C_{\max}$), we note that this two-step concentration analysis allows us to have $C'_{\max}$ appear multiplying the "slower" $n^{-\frac{1}{2}}$ term, with the potentially bigger $C_{\max}$ multiplying the "faster" $(nm)^{-\frac{1}{2}}$ term. Occasionally, we may be able to choose m, in which case we can make a choice that minimizes the upper-bound of Lemma 2.4.8 up to constant factors: $m = \left(\frac{C_{\max}}{C'_{\max}}\right)^2$ (cf. $m = 1$ suggested by Lazaric et al., 2016, Theorem 5).

Lemma 2.4.8. *For any $\mathcal{H} \subset \mathcal{H}$ satisfying Assumption 2.4.5, φ satisfying Assumptions 2.4.3 and 2.4.4, $\delta \in (0,1)$ and $h \in \mathcal{H}$, if S satisfies Assumption 2.4.2 then we have with probability at least $1 - \delta$ that*

$$\left| R^{\text{surr}}(h) - \widehat{R}^{\text{surr}}(h, S) \right| \leq \varphi_{\max,\mathcal{H}} \sqrt{\frac{1}{2n} \ln \frac{4}{\delta}} \left(C'_{\max} + C_{\max} \sqrt{\frac{1}{m}} \right).$$

We can follow the steps of Pollard (1984, Symmetrization, pp. 14–15, and Theorem 24, pp. 25–26) and use Lemma 2.4.8 to bound the concentration of the empirical risk for all score functions in $\mathcal{H}$ (simultaneously). Differently from Pollard (1984), we must account for dependencies in the sample and make sure that the bound is given in terms of a covering number of the set $\{(i,k) \mapsto \varphi(h(X_i)_k) : h \in \mathcal{H}\}$, rather than $\left\{(i,j,k) \mapsto C'_{i,j,k} \varphi(h(X_i)_k) : h \in \mathcal{H}\right\}$. Covering number bounds for the latter set may be significantly looser than covering number bounds for the former set.

Lemma 2.4.9 gives a bound on the concentration of empirical risks of all score functions in $\mathcal{H}$. We can see that the dependencies on n and m are inherited from Lemma 2.4.8 (up to constant factors), and that the covering number $N_{\infty/\infty}\left(2\frac{\varphi_{\max,\mathcal{H}}}{nm}, n, \mathcal{H}\right)$ appears. As we will see in Lemma 2.4.10 and Theorem 2.4.11, for $\mathcal{H}_{\phi,B}$, our upper-bound on the logarithm of this covering number scales linearly with d (the number of features) and logarithmically with m and n.

Lemma 2.4.9. *For any $\mathcal{H} \subset \mathcal{H}$ satisfying Assumption 2.4.5, φ satisfying Assumptions 2.4.3 and 2.4.4, and $\delta \in (0,1)$, if S satisfies Assumption 2.4.2 then we have with probability at least $1-\delta$ that*

$$\begin{aligned}&\sup_{h\in\mathcal{H}}\left|\widehat{R}^{\text{surr}}(h,S) - R^{\text{surr}}(h)\right| \\ &\quad\leq \varphi_{\max,\mathcal{H}}\sqrt{\frac{32}{n}\ln\frac{16N}{\delta}}\left(2C'_{\max} + C_{\max}\sqrt{\frac{1}{m}}\right),\end{aligned}$$

where

$$N = N_{\infty/\infty}\left(2\frac{\varphi_{\max,\mathcal{H}}}{nm}, n, \varphi\circ\mathcal{H}\right).$$

Lemma 2.4.10 gives us a covering number bound for $\mathcal{H}_{\phi,B}$. The result relies on well-known bounds on the maximum size of a minimum-covering of the d-dimensional ball of radius B. Indeed, we can transform such a covering into a covering of $\{(i,k) \mapsto \varphi(h(X_i)_k)) : h \in \mathcal{H}\}$, by Lipschitzness of $w \mapsto \langle\phi(x,k), w\rangle$ and of φ in the interval $[-BB_*, BB_*]$.

Lemma 2.4.10. *With ϕ satisfying Assumption 2.4.6 and φ satisfying Assumptions 2.4.3 and 2.4.4, for any $n \in \mathbb{N}$, we have that*

$$\ln N_{\infty/\infty}(\varepsilon, n, \varphi\circ\mathcal{H}_{\phi,B}) \leq d\ln\left(1 + \frac{2BB_*}{\varepsilon}\text{Lip}_\varphi(BB_*)\right)$$

We can plug Lemma 2.4.10 into Lemma 2.4.9 to get risk bounds for empirical surrogate risk minimizers (of L^{LLW}) in $\mathcal{H}_{\phi,B}$, as seen in Theorem 2.4.11. The scaling of the risk bound in Theorem 2.4.11 is the "usual" in terms of n: proportional to $\sqrt{\frac{\ln n}{n}}$, where the $\ln n$ factor comes from the covering number bound (as we will mention in Section 2.5, the dependency

may be improved if refine the proof of Lemma 2.4.9). On the other hand, in terms of m the bound in Theorem 2.4.11 scales with $\sqrt{\ln m}$, since ultimately the term $2C'_{\max}$ dominates $C_{\max}\sqrt{\frac{1}{m}}$. As a sanity check, we see that the risk bound scales with the square-root of the number of features, which upper-bounds the graph dimension of $\mathcal{H}_{\phi,B}$ (combine Daniely et al., 2013, Theorem 3.5 and Anthony and Bartlett, 2009, Theorem 3.5, p. 37). In this sense, the right-hand side is somewhat similar to what we would obtain for empirical true-risk minimization over linear classifiers. (Of course, Theorem 2.4.11 gives us a *surrogate-risk* bound for empirical surrogate risk minimizers in $\mathcal{H}_{\phi,B}$.)

Theorem 2.4.11. *With φ satisfying Assumptions 2.4.3 and 2.4.4, ϕ satisfying Assumption 2.4.6, for any $\delta \in (0,1)$ if S satisfies Assumption 2.4.2 then we have with probability at least $1-\delta$ that*

$$\begin{aligned} &R^{\text{surr}}(\widehat{H}_{\phi,B}, S) - \inf_{h \in \mathcal{H}_{\phi,B}} R^{\text{surr}}(h, S) \\ &\quad \leq 2\varphi(BB_*)\sqrt{\frac{32d}{n}\ln\frac{20c}{\delta}}\left(2C'_{\max} + C_{\max}\sqrt{\frac{1}{m}}\right) \end{aligned}$$

where

$$c = \left(1 + \frac{BB_*\text{Lip}_\varphi(BB_*)}{\varphi(BB_*)}nm\right).$$

In a non-parametric setting, we see that Theorem 2.4.11 allows us to have B "grow" polynomially with n without affecting the risk bound by more than a constant factor. This is convenient when the feature extractor is expressive enough for the approximation error, the surrogate risk of the best score function in $\mathcal{H}_{\phi,B}$, to decrease significantly as $B \to \infty$, in which case one could choose B to trade off the risk bound and the approximation error.

Because we are controlling the magnitude of the weights underlying the score functions (via the constant B), we would expect risk bounds with a better scaling in terms of d, as, for example, controlling the 1-norm does (Bartlett et al., 2012, Theorem 1.2), rather than, for example, $\sqrt{d}$ as in

Theorem 2.4.11. In this sense, can we improve over Theorem 2.4.11? Indeed we can, if we use different covering numbers.

Lemma 2.4.12 is an analogue of Lemma 2.4.9 that uses Frobenius-norm covering numbers instead of ∞/∞-norm covering numbers. The bound in Lemma 2.4.12 is quite similar to the one in Lemma 2.4.9, except for the covering number and the resolution of the covering (for an ε-covering, we informally refer to $\frac{1}{\varepsilon}$ as its resolution). In Lemma 2.4.9, we need a covering whose resolution scales with nm, whereas in Lemma 2.4.12 the resolution of the covering scales with $\sqrt{\frac{m}{|\mathcal{Y}|}}$. This scaling may be an artifact of the proof, and the factor of $\left(\sqrt{\frac{1}{m}} \vee \frac{C'_{\max}}{C_{\max}}\right)$ in (2.4.4) may be avoidable, but we have not been able to remove it. On its own, Lemma 2.4.12 does not necessarily improve over Lemma 2.4.9. However, for $\mathcal{H}_{\phi,B}$ we can use a result by Zhang (2002) to show an improvement over Theorem 2.4.11.

Lemma 2.4.12. *For any $\mathcal{H} \subset \mathcal{H}$ satisfying Assumption 2.4.5, φ satisfying Assumptions 2.4.3 and 2.4.4, and $\delta \in (0,1)$, if S satisfies Assumption 2.4.2 then we have with probability at least $1-\delta$ that*

$$\begin{aligned}&\sup_{h\in\mathcal{H}}\left|\widehat{R}^{\text{surr}}(h,S) - R^{\text{surr}}(h)\right| \\ &\quad\leq \varphi_{\max,\mathcal{H}}\sqrt{\frac{32}{n}\ln\frac{16N}{\delta}}\left(2C'_{\max} + C_{\max}\sqrt{\frac{1}{m}}\right),\end{aligned}$$

where

$$N = N_F\left(\varphi_{\max,\mathcal{H}}\sqrt{|\mathcal{Y}|}\left(\sqrt{\frac{1}{m}} \vee \frac{C'_{\max}}{C_{\max}}\right), n, \varphi\circ\mathcal{H}\right). \tag{2.4.4}$$

Lemma 2.4.13 is the analogue of Lemma 2.4.10 for Frobenius-norm covering numbers. The result itself is a simple consequence of covering numbers bounds due to Zhang (2002, Theorem 3 and Corollary 3). We can see that the bound on the logarithm of the covering number scales quadratically with resolution of the covering, but only logarithmically with the number of features—or, if the norm underlying the definition of $\mathcal{H}_{\phi,B}$ is a Frobenius norm, the bound does not depend at all on the number of

features (but scales logarithmically with the sample size and the number of classes).

Lemma 2.4.13. *If ϕ satisfies Assumption 2.4.6, φ satisfies Assumptions 2.4.3 and 2.4.4, and the norm underlying the definition of $\mathcal{H}_{\phi,B}$ is $\|\cdot\|_p$ for $p \geq 1$, then for any $n \in \mathbb{N}$ we have that*

$$\ln N_F(\varepsilon, n, \varphi \circ \mathcal{H}_{\phi,B}) \leq \left\lceil \left(\frac{\mathrm{Lip}_\varphi(BB_*)BB_*}{\varepsilon} \right)^2 \right\rceil \log_2(2d+1).$$

If, additionally, $p = 2$, then we also have

$$\ln N_F(\varepsilon, n, \varphi \circ \mathcal{H}_{\phi,B}) \leq \left\lceil \left(\frac{\mathrm{Lip}_\varphi(BB_*)BB_*}{\varepsilon} \right)^2 \right\rceil \log_2(n|\mathcal{Y}|+1).$$

To conclude this section, we present Theorem 2.4.14, which is obtained by combining Lemmas 2.4.12 and 2.4.13. We see that the bound in Theorem 2.4.14 is the maximum of two bounds, and the condition (2.4.5) determines (up to constant factors) which one of the two is worse. Indeed, if Equation (2.4.5) holds we have $\varepsilon_1 \vee \varepsilon_2 = \varepsilon_1$, otherwise we have $\varepsilon_1 \vee \varepsilon_2 \leq 6\varepsilon_2$. The bound in Theorem 2.4.14 does not scale with m, since for m large enough the constant $2C'_{\max}$ dominates in ε_1, and ε_2 is constant w.r.t. m. In contrast, the bound Theorem 2.4.11 scales with $\sqrt{\ln m}$. Moreover, the scaling of the bound in Theorem 2.4.14 in terms of d greatly improves over Theorem 2.4.11, with a stronger dependence on the constants BB_* (and the Lipschitz constant of φ). In a non-parametric setting, in contrast to Theorem 2.4.11, where we could scale B polynomially with n without affecting the bound significantly, here we can scale d polynomially with n without significant effects on the bound, for B constant. The most interesting aspect of Theorem 2.4.14 is that the bound does not scale with the number of classes—ultimately, as $|\mathcal{Y}|$ increases (and everything else is fixed), we will have $\varepsilon_1 \geq \varepsilon_2$. How is the the cost-sensitive classification problem in our setting not statistically harder for $|\mathcal{Y}|$ very large? It is likely that even the best hypothesis in $\mathcal{H}_{\varphi,B}$ cannot do well in challenging instances with too

many classes, since the hypotheses in $\mathcal{H}_{\varphi,B}$ have limited expressiveness (constrained by d and B), so competing against the best hypothesis in $\mathcal{H}_{\varphi,B}$ should not become more challenging with very large $|\mathcal{Y}|$.

Theorem 2.4.14. *With φ satisfying Assumptions 2.4.3 and 2.4.4, ϕ satisfying Assumption 2.4.6, for any $\delta \in (0,1)$ if S satisfies Assumption 2.4.2 then we have with probability at least $1-\delta$ that*

$$R^{\text{surr}}(\widehat{H}, S) - \inf_{h \in \mathcal{H}_{\phi,B}} R^{\text{surr}}(h, S) \leq \varepsilon_1 \vee \varepsilon_2,$$

where

$$\varepsilon_1 = 2\varphi(BB_*)\sqrt{\frac{32}{n}\ln\frac{20(2d+1)}{\delta}}\left(2C'_{\max} + C_{\max}\sqrt{\frac{1}{m}}\right),$$

$$\varepsilon_2 = 12\text{Lip}_\varphi(BB_*)BB_*C_{\max}\sqrt{\frac{32}{n|\mathcal{Y}|}\ln\frac{20(2d+1)}{\delta}}.$$

Moreover, if

$$\sqrt{\frac{1}{m}} \vee \frac{C'_{\max}}{C_{\max}} \geq \frac{\text{Lip}_\varphi(BB_*)BB_*}{\varphi(BB_*)}\sqrt{\frac{1}{|\mathcal{Y}|}}. \tag{2.4.5}$$

then $\varepsilon_1 \geq \varepsilon_2$, otherwise $\varepsilon_1 \leq 6\varepsilon_2$.

In the common setting where we have one d'-dimensional weight vector for each class ("class-independent weights"), each constrained to have norm at most some B', the bound in Theorem 2.4.14 scales with $\sqrt{\ln|\mathcal{Y}|}$. To see this, consider the setting where $\phi(x,k) = \phi'(x) \otimes e_k$ where $\phi' : \mathcal{X} \to \mathbb{R}^{d'}$ is a feature extractor over $\mathcal{X}$ only, $\otimes$ denotes the Kronecker product, and e_k is the k-th $|\mathcal{Y}|$-dimensional elementary vector. If the norm underlying $\mathcal{H}_{\phi,B}$ is the Frobenius norm, we can constrain the 2-norm of the weight vectors for each class individually, to be at most B', which gives us the hypothesis class $\mathcal{H}_{\phi,\sqrt{|\mathcal{Y}|}B'}$. In this case, we still have $\sup_{h\in\mathcal{H},x\in\mathcal{X},k\in\mathcal{Y}}|h(x)_k| \leq B'B_*$, but $d = |\mathcal{Y}|d'$. Therefore, ε_1 in Theorem 2.4.14 scales with $\sqrt{\ln|\mathcal{Y}|}$ and ε_2 becomes

$$\begin{aligned} &R^{\text{surr}}(\widehat{H}, S) - \inf_{h \in \mathcal{H}_{\phi,B}} R^{\text{surr}}(h, S) \\ &\quad\leq 12\text{Lip}_\varphi(B'B_*)B'B_*C_{\max}\sqrt{\frac{32}{n}\ln\frac{20(2d'|\mathcal{Y}|+1)}{\delta}}. \end{aligned}$$

2.4.3 Discussion and Related Work

Comparison to other surrogate risk bounds. To the best of our knowledge, there are no surrogate risk bounds in cost-sensitive multiclass classification. In binary classification, Vapnik (2013, p. 76) upper-bounded the performance of the empirical true risk minimizer in the context of cost-sensitive classification. The surrogate risk bound presented by Zadrozny et al. (2003, Theorem 2.2) is about converting bounds on the cost-insensitive true risk into bounds on the cost-sensitive true risk, whereas Scott (2011) have explored bound conversion in cost-sensitive binary classification using calibration functions.

Koltchinskii and Panchenko (2002), Mohri et al. (2012, Theorem 8.1, p. 187), Kuznetsov et al. (2014), Lei et al. (2015) and Maximov and Reshetova (2015) present surrogate bounds in the context of cost-insensitive multiclass classification. The surrogate loss used is the *margin loss*, defined by $L_{\varphi,z}(s,y) \doteq \varphi(z + \max_{y' \neq y} s_{y'} - s_y)$ for some φ s.t. $\varphi(t) \geq \mathbb{I}\{t \geq 0\}$ for all $t \in \mathbb{R}$, and $z \geq 0$. There is some intersection of all those results and ours, because in all cases the main challenge is bounding

$$\sup_{h \in \mathcal{H}} \left| \widehat{R}^{\text{surr}}(\widehat{H}, S) - R^{\text{surr}}(\widehat{H}, S) \right|$$

with high probability (see Proposition 2.4.1). However, the surrogate loss we consider, $\frac{1}{|\mathcal{Y}|} L^{\text{LLW}}$ in the cost-insensitive case, is substantially different from the margin loss $L_{\varphi,z}$, so that we cannot meaningfully compare bounds if the goal is minimizing the surrogate risk. Nevertheless, we could consider comparing true risk bounds obtained from these surrogate risk bounds.

True risk bounds. While no $|\mathcal{Y}|$ appears in the surrogate risk bounds given by Theorem 2.4.14, that is not the case for the true risk bounds we can get from our surrogate risk bounds given by Theorem 2.4.11 or Theorem 2.4.14 (see also Theorem 3.2.4). By reduction to Theorem 2.3.4, if $\delta(\varepsilon)$ is a calibration function for L^{LLW}, then $\frac{1}{|\mathcal{Y}|}\delta(\varepsilon)$ is a calibration function for $\frac{1}{|\mathcal{Y}|} L^{\text{LLW}}$ (see also Definition 2.2.1), so based on the current results we

need to scale surrogate risk bounds by $|\mathcal{Y}|$ before converting them into true risk bounds. As we will discuss in Section 2.5, the appearance of this undesirable factor of $|\mathcal{Y}|$ is a limitation of the calibration results.

Koltchinskii and Panchenko (2002), Mohri et al. (2012, Theorem 8.1, p. 187), Kuznetsov et al. (2014), Lei et al. (2015) and Maximov and Reshetova (2015) obtain true risk bounds for their surrogate risk bounds by using the fact that for every score function h we have (w.p. one)

$$\mathbb{I}\{f(h(X)) \neq Y\} \leq \inf_{z \geq 0} L_{\varphi,z}(h(X), Y). \tag{2.4.6}$$

It is possible to show (see Ávila Pires and Szepesvári, 2016a, Section 2) that (2.4.6) can be hopelessly loose even if $f \circ h$ for some $h \in \mathcal{H}$ is optimal, and that we should use calibration functions to perform bound conversion instead. However, the margin loss is not amenable to a calibration argument: For example, if φ is non-decreasing, we can show that $L_{\varphi,z}$ can be transformed into an instance of the loss of Crammer and Singer (2003), for which no calibration function exists[9] (see Zhang, 2004).

2.5 Conclusion

In this chapter, we presented a novel cost-sensitive multiclass classification loss, and a calibration function for it. This novel loss, L^{Zhang}, generalizes the loss proposed by Zhang (2004), which in turn generalizes, among other losses the multiclass logistic regression loss. We also described a simple process to obtain cost-sensitive losses and calibration functions from cost-insensitive losses and their calibration functions. Moreover, we have also presented novel surrogate risk bounds for the cost-sensitive multiclass

[9] Nevertheless, margin losses are popular choices in empirical studies (see, *e.g.* Bakir et al., 2007), with evidence of good performance in low-noise settings (Doğan et al., 2016). Moreover, the loss of Crammer and Singer (2003) in the cost-insensitive setting is known to have a calibration function w.r.t. distributions of Y where some one label has probability at least $\frac{1}{2}$ (Zhang, 2004, p. 1233). It may be the case that under the Mammen-Tsybakov Noise condition (Mammen and Tsybakov, 1999; Boucheron et al., 2005; Bartlett et al., 2006) one can obtain calibration functions for the loss of Crammer and Singer (2003) (and other margin losses), but to the best of our knowledge this line of work has not yet been pursued.

classification. These bounds apply to the loss L^{LLW}. There is a number of interesting refinements and extensions that we can consider to our bounds.

The calibration function in Theorem 2.3.6 applies to a decoupled variant of the loss proposed by Zhang (2004). Ávila Pires and Szepesvári (2016a) presents calibration functions to the cost-insensitive logistic regression loss as well, so it would be interesting to have results for the coupled formulation, at the very least for a cost-sensitive generalization of the logistic regression loss.

Because we presented a calibration function for a decoupled L^{Zhang} (with ψ non-decreasing and $F(t) = t$), we could think of surrogate risk bounds for it as well, besides just L^{LLW}. These results would be a standard exercise of re-doing the analysis in Section 2.4 with a loss that, from the point of view of the analysis, is similar to L^{LLW}.

A more interesting endeavor would be to re-do the analyses in Sections 2.3 and 2.4 using the Mammen-Tsybakov Noise condition (Mammen and Tsybakov, 1999; Boucheron et al., 2005; Bartlett et al., 2006). The Mammen-Tsybakov noise condition, in binary classification, condition interpolates between a noiseless-label (realizable) scenario and a scenario with no assumptions on the marginal distributions of Y given X. The rates of the true-risk upper-bounds obtained for empirical true-risk minimization interpolate between the usual $\sqrt{\frac{1}{n}}$-rates (in the no-assumption scenario) and fast, $\frac{1}{n}$-rates in the noiseless scenario (Boucheron et al., 2005). Bartlett et al. (2006) presented similar results (interpolating rates for true-risk upper-bounds) for empirical surrogate-risk minimizers in the case of cost-insensitive binary classification with strictly convex margin-based losses (that is, L^{LLW} with φ strictly convex and $|\mathcal{Y}| = 2$). Ávila Pires and Szepesvári (2016a) extended the calibration results of Bartlett et al. (2006) to the cost-insensitive multiclass case, but did not present surrogate risk bounds as Bartlett et al. (2006) did. Farahmand et al. (2014) generalized the Mammen-Tsybakov noise condition to the cost-sensitive multiclass classification case, and presented true-risk bounds for empirical true-risk minimizers.

Therefore, what remains to be done is to: i) generalize the fast-rate calibration results of Bartlett et al. (2006); Ávila Pires and Szepesvári (2016a) to the cost-sensitive case; ii) re-do the analysis in Section 2.4 with Bernstein's inequality (Steinwart and Christmann, 2008, Theorem 6.12, p. 213) instead of Hoeffding's inequality (see Boucheron et al., 2005, Section 5.2); and iii) bound $\mathrm{Var}(L(h(X), C) - L(h'(X), C))$ for any $h, h' \in \mathcal{H}$. The third step has been done for the cost-insensitive binary case by Bartlett et al. (2006, see Theorems 4 and 5), where it has been noted that, in order to obtain fast-rates, strictly convex surrogate losses are recommended.

While the extensions of our results to L^{Zhang} do not seem very interesting, the fast-rate extensions using the Mammen-Tsybakov noise condition with strongly convex L^{Zhang} are much more appealing. For example, we could take $\psi(t) = -t$, and $\varphi = \varphi^{\exp}$, which give us losses related to logistic regression (in particular, as mentioned before, with $F(t) = \ln t$ in the cost-insensitive case we get the logistic regression loss).

We are also able to improve Lemmas 2.4.9 and 2.4.12. In the proofs of these results, after a symmetrization step, we have used a union bound over a covering of $\{(i,k) \mapsto \varphi(h(X_i)_k : h \in \mathcal{H}\}$ to obtain the final result. Using this union-bound is sub-optimal, and the bounds of Lemmas 2.4.9 and 2.4.12 can be improved (with better dependencies on covering numbers) if we using chaining instead (see Koltchinskii, 2011, Chapter 3 and Steinwart and Christmann, 2008, Chapter 7).

The techniques used for obtaining calibration functions also have room for improvement. As claimed in Section 2.4.3 (and as we will see in Section 3.2), when converting surrogate risk bounds using Theorem 2.3.4 (and also Theorem 2.3.6), a factor of $|\mathcal{Y}|$ is introduced to the true risk bounds. This is a flaw of results that lower-bound $\delta_{\max}$ by δ_{binary}. We speculate that a bound of the form $\delta_{\max}(\varepsilon, p) \geq \delta_{\mathrm{binary}}(\varepsilon, p')$ (cf. Theorems 2.3.4 and 2.3.6) is not possible (even up to constant factors) for $\frac{1}{|\mathcal{Y}|} L^{\mathrm{LLW}}$ or $\frac{1}{|\mathcal{Y}|} L^{\mathrm{Zhang}}$, but at the same time we may be able to improve Theorems 2.3.4 and 2.3.6.

While the focus of this chapter and, in fact, this dissertation, is theo-

retical, there is interesting empirical research where surrogate losses are compared. For example, Doğan et al. (2016) experimentally compare different cost-insensitive multiclass classification surrogate losses, where most loss choices (including $L^{\text{LLW,CI}}$) provide competitive results on the datasets considered, except in seemingly low-noise scenario, where the loss proposed by Crammer and Singer (2003) is preferable in both predictive and computational performance. To the best of our knowledge, empirical comparisons in the cost-sensitive multiclass case have not been pursued, but, in our opinion, they would be interesting. The cost-sensitive setting (with random costs) can be quite different from the cost-insensitive setting. On the one hand, misclassification (w.r.t. the label with lowest expected cost) is not an issue if another low-cost label is chosen. On the other hand, noise on the costs can affect classifiers in ways that are not understood (as observed by Pires and Szepesvári, 2015, in the context of CBPI).

Chapter 3

Classification-Based Policy Iteration

In this chapter, we present an extended analysis of Classification-Based Policy Iteration. As remarked in Section 1.2, CBPI falls under the category of so-called direct policy learning (DPL) methods, which also includes policy gradient (Sutton et al., 1999), conservative policy iteration (Kakade and Langford, 2002), and classification-based methods for learning non-stationary policies (Langford and Zadrozny, 2003; Bagnell et al., 2003; Langford and Zadrozny, 2005). The common feature of these methods is that they attempt to learn a policy without estimating value functions. This creates the potential for DPL methods to perform well in scenarios where generalized policy iteration (GPI) may fail, *e.g.*, when the representation is expressive enough for representing near-optimal policies, but not near-optimal value functions. One such case is the (SZ-)Tetris scenario outlined in Section 1.2.

At the core of CBPI there is a classification method, whose performance (the true risk) is determinant to the performance (the policy error) of the policies computed by CBPI. The two existing analysis of CBPI (Farahmand et al., 2014; Lazaric et al., 2016) apply to instances of CBPI that rely on an "impractical" classification method (recall that by "practical" we mean classification algorithms that can be executed efficiently and therefore used in practice). The method used is empirical true-risk minimization, where the true risk is the cost-sensitive classification error. Because the empirical

true risk is not convex, empirical true-risk minimization may require solving a combinatorial problem that is often computationally hard (see Section 2.2). Therefore, we have performance guarantees for CBPI methods that we cannot use in practice, and we can define instances of CBPI that we can use in practice, but which enjoy no performance guarantees. What we do in this chapter is to extend the analysis of Lazaric et al. (2016), so that we have performance guarantees for instances of CBPI that use a practical classification method.

The classification algorithm that we consider in this chapter corresponds to empirical surrogate-risk minimization with the surrogate loss L^{LLW} and φ^{hinge} over the set of linear score functions $\mathcal{H}_{\phi,B}$. Not surprisingly, this is the classification method for which we have presented surrogate risk bounds in Section 2.4. Moreover, we can use calibration functions from Theorem 2.3.4 to convert these surrogate risk bounds into true risk bounds. These true risk bounds can then be plugged into a result by Lazaric et al. (2016) that gives us a bound on the policy error of the policy constructed by CBPI, as a function of the performance of the classification method used.

The outline of this chapter is as follows. We start by introducing CBPI through a novel, unified view of the method, in Section 3.1. This unified view of CBPI encompasses the different variants of the method studied in the literature, and allows us to really understand what is, in our opinion, the essential structure of CBPI, and also to identify the components that can be easily changed. In Section 3.2, we present the analysis of CBPI *per se*, and in Section 3.3 we discuss future work.

While we do not develop an extension of the analysis of CBPI by Farahmand et al. (2014), the extensions discussed in Section 2.5 are much relevant to their analysis of CBPI, and we discuss these connections in Section 3.3.

3.1 A Unified View of CBPI

The classification-based policy iteration algorithm is given in Algorithms 3.1.1 and 3.1.2. Each iteration in CBPI requires three (blackbox) components to

Algorithm 3.1.1 Classification-based policy iteration

input: An initial policy π_0, a number of iterations K, a per-iteration number of observed states n
output: A policy π_K
  procedure CBPI(K, π_0, n)
    for $k \in [K]$ **do**
      Set the state provider States
      Set the cost estimator CostEstimate
      Set the cost-sensitive multiclass classification method Classifier
      $(x_1, \ldots, x_n) \leftarrow \textsc{States}(n)$
      $\pi_k \leftarrow \textsc{Iteration}((x_1, \ldots, x_n), \textsc{CostEstimate}, \textsc{Classifier})$
    end for
  end procedure

be set: a state provider States, a cost estimator CostEstimate, and a classification method Classifier (the components are named after what they yield). Then CBPI constructs a policy using the Iteration procedure, which is the defining trait of CBPI in comparison to other policy iteration methods. Algorithm 3.1.2 gives a detailed description of Iteration. Different variants

Algorithm 3.1.2 Iteration routine for CBPI.

input: States $(x_1, \ldots, x_n) \in \mathcal{X}^n$, a cost estimator CostEstimate, a cost-sensitive classification method Classifier
output: A policy
  procedure Iteration($(x_1, \ldots, x_n)$, CostEstimate, Classifier)
    for all $i \in [n]$, $a \in \mathcal{A}$ **do**
      $C_{i,a} \leftarrow \textsc{CostEstimate}(x_i, a)$
    end for
    return $\textsc{Classifier}(x_1, \ldots, x_n, C_1, \ldots, C_n)$ $\quad \triangleright\ \forall i, x_i \in \mathcal{X},\ C_i \in \mathbb{R}^{|\mathcal{A}|}$
  end procedure

of CBPI have been analyzed and evaluated empirically (Bagnell et al., 2003; Fern et al., 2003; Lagoudakis and Parr, 2003a; Lazaric et al., 2010; Gabillon et al., 2011, 2013; Farahmand et al., 2014; Lazaric et al., 2016), and all these

methods are described by Algorithm 3.1.1 and Algorithm 3.1.2, with the differences lying in which specific choices are made for the three components, States, CostEstimate and Classifier.

Indeed, at each iteration these CBPI instances can be seen to execute the following steps: i) for a given set of states and actions, a blackbox CostEstimate produces an estimate of the cost associated with taking each action at each given state; and ii) a Classifier produces an approximately greedy policy from the costs estimates. Between iterations, as seen in Algorithm 3.1.1, States, CostEstimate, Classifier and the states $(x_1, \ldots, x_n)$ may change. For example, the CostEstimate blackbox may work as follows. Monte Carlo value estimates are generated by performing rollouts with a behavior policy π (which can change between iterations) and then these value estimates are used to construct the cost estimates. For a given value estimate $V \in \mathbb{R}^{|\mathcal{A}|}$ for a state x, at least two different cost-estimates can be obtained: The negative value estimates $-V$ and the estimated *disadvantages* of the actions $(\max_a V_a)\mathbf{1} - V$. The second option may be used if the costs are required by the Classifier to be non-negative (as will be the case in Section 3.2).

In practice, we may want to use Algorithm 3.1.2 with state-dependent action spaces, which is typically an extension that can be easily incorporated into common classification methods. We could further generalize Algorithm 3.1.2 to allow a sequence of state-action pairs $((x_1, a_1), \ldots, (x_n, a_n)) \in (\mathcal{X} \times \mathcal{A})^n$ to be given as an argument. This generalization would require Classifier to be able to generalize both across states and actions, which is to the best of our knowledge unexplored in cost-sensitive classification, albeit conceivable, based the structured prediction[1] literature (see, *e.g.*, Pérez-Cruz et al., 2007; Bakir et al., 2007).

[1] Structured prediction can be seen as cost-insensitive classification with an extremely large number of classes where classes may be related under some notion of similarity, and where minimizing typical classification losses cannot be done because the cost of computing minimizers and evaluating these losses typically scales linearly with the number of classes, as we can see from Table 3 in Ávila Pires and Szepesvári (2016a).

The nature of the value estimates underlying the cost estimates can vary greatly. For example, Lazaric et al. (2010), Lazaric et al. (2016) and Bagnell et al. (2003) use the average of Monte Carlo rollouts as the value estimator. Farahmand et al. (2014), Gabillon et al. (2011) and Gabillon et al. (2013), in contrast, use hybrid estimates that combine value-function estimates and value estimates. Moreover, it is possible to have CostEstimate reuse observations (similarly to Lagoudakis and Parr, 2003b) over different iterations, or to combine policies from previous iterations into a rollout policy (Bagnell et al., 2003).

The state provider States is critical to the performance and analysis of CBPI. Bagnell et al. (2003); Farahmand et al. (2014); Lazaric et al. (2016) assume that States allows us to observe i.i.d. states from a given distribution. Evidently, the nature of the distribution will affect the policy error guarantees that we can obtain (see the concentrability coefficients in Theorem 3.2.7). For example, a distribution that is closer to a stationary distribution of the optimal policy can be expected to lead to better classifiers (that is, better policies). Relaxations of the i.i.d. assumption, *e.g.* fast mixing, have been explored in the context of other policy iteration methods (*e.g.* Antos et al., 2008b), but not in CBPI, to the best of our knowledge.

As for Classifier, we can see it as an operator $G' : \mathbb{R}^{n|\mathcal{A}|} \to \Pi$, in which case the k-th iteration of CBPI produces

$$\pi_k = G'v$$

where $v_i^a \approx (T_{\mathcal{P}} V^{\pi'_k})^a(x_i)$ for a given sequence $(x_1, \ldots, x_n) \in \mathcal{X}^n$, and π'_k is a policy underlying the cost estimates produced by CostEstimate (in some instances we may have $\pi'_k = \pi_{k-1}$). For classifiers based on score functions, we can further write

$$\pi_k = GHv$$

where $H : \mathbb{R}^{n|\mathcal{A}|} \to \mathcal{V}^{\mathcal{A}}$ gives us score functions over the whole state-action space, and G is the greedy operator. In contrast, GPI outputs

$$\pi_k = GV$$

at iteration k, where $V \approx T_{\mathcal{P}} V^{\pi_{k-1}}$ and π_0 is some initial policy. The fact that Hv is a scores function (and not a value function estimate, such as V) means that we can be much more flexible about its construction. Indeed, it suffices for $GHv \approx \pi^*$ for us to perform well, even if MHv is not close to $MT_{\mathcal{P}} V^*$ at all. On the other hand, GPI relies on having $V \approx T_{\mathcal{P}} V^*$, as we can see from policy error bounds, which are given in terms of $\|V - V^*\|$ (Bertsekas, 2012, Proposition 3.1).

The Classifier routine has seen different instantiations in the literature. Fern et al. (2003) use rule-based classifiers. Lagoudakis and Parr (2003a) take (with ties broken arbitrarily)

$$\operatorname*{argmin}_{\pi \in \Pi'} \frac{1}{n} \sum_{i=1}^{n} \mathbb{I}\left\{ \pi(x_i) \neq \operatorname*{argmin}_k C_{i,k} \right\},$$

which Li et al. (2007) show not to be a sensible objective for Classifier because it is not cost-sensitive (thus underweighting errors that cause large losses in terms of the return, and overweighting errors that do not significantly affect the return). Lazaric et al. (2010); Farahmand et al. (2014); Lazaric et al. (2016) propose taking

$$\operatorname*{argmin}_{\pi \in \Pi'} \frac{1}{n} \sum_{i=1}^{n} C_{i,\pi(x_i)}, \tag{3.1.1}$$

with $\Pi' \subset \Pi$ given. Lazaric et al. (2010); Farahmand et al. (2014) use disadvantages as the costs, whereas Lazaric et al. (2016) use negative values. Objective (3.1.1) is the empirical classification cost (see Section 2.2), which is non-convex and is usually not efficiently minimizable. This non-convexity issue can be remedied by replacing the costs $(C_1, \ldots, C_n)$ with a convex cost-sensitive surrogate loss, as we have done in Section 2.2 Replacing the costs with a convex loss in (3.1.1) solves a practical problem, but creates a gap between instances of CBPI that can be used and instances that enjoy the statistical guarantees presented by Farahmand et al. (2014); Lazaric et al. (2016). Bridging this gap is the theme of the next section, and to do so we will use the classification results established in Chapter 2.

3.2 An Extended Analysis of CBPI

With the cost-sensitive classification surrogate risk bounds from Section 2.4 and the ability to convert surrogate risk bounds into true risk bounds, thanks to Section 2.3, we are able to extend the analysis of Lazaric et al. (2016) so that it applies to practical instances of CBPI, *i.e.*, instances that rely on empirical surrogate-risk minimization. Once we have a true risk bound for the classifier, we can simply plug it in the bound of Theorem 7 of Lazaric et al. (2016) (presented as Theorem 3.2.7 here), to get error propagation results.

We will start this section by introducing the setting considered by Lazaric et al. (2016) in a "single-iteration" manner, in Section 3.2.1. This will give us single-iteration descriptions of States and CostEstimate, and simplify the statement of our true risk bounds for Classifier, which are given in Section 3.2.2.

At the end of this section, in Section 3.2.3, we will describe how States and CostEstimate change at each iteration of CBPI and report Theorem 3.2.7, the policy error bound shown by Lazaric et al. (2016) into which we can plug our true risk bounds.

3.2.1 Preliminaries

Consider states $(X_1, \dots, X_n) \in \mathcal{X}^n$ and define, for $j \in [m]$, $a \in \mathcal{A}$ and $t \in [h-1]$

$$\begin{aligned} &X_{i,j,a,0} \doteq X_i, && A_{i,j,a,0} \doteq a, \\ &X_{i,j,a,t} \sim \mathcal{P}(X_{i,j,a,t-1}, A_{i,j,a,t-1}), && A_{i,j,a,t} \sim \pi(X_{i,j,a,t}) \\ &T_{i,j,a} \doteq X_{i,j,a,0}, A_{i,j,a,0}, \dots, X_{i,j,a,h-1}, A_{i,j,a,h-1}. \end{aligned} \tag{3.2.1}$$

where π is a given evaluation policy and $h > 0$ is a horizon. The sample $(X_1, \dots, X_n)$ and the trajectories will be assumed to satisfy Assumption 3.2.1. For convenience, we let (X, T) be jointly-distributed random variables that share the common distribution and are independent of $(X_i, T_{i,j})$ ($i \in [n], j \in [m]$).

Assumption 3.2.1. *The* $(X_1, \ldots, X_n)$ *arei.i.d. For all* $i \in [n]$, $j \in [m]$, $a \in \mathcal{A}$,

1. $(T_{i,1,a}, \ldots, T_{i,m,a})$ *are conditionally independent given* X_i,
2. $(T_{1,j,a}, \ldots, T_{n,j,a})$ *are independent,*
3. $(T_{i,j,1}, \ldots, T_{i,j,|\mathcal{A}|})$ *need not be independent.*

One can generate trajectories that satisfy Assumption 3.2.1 by using Monte Carlo rollouts with common random numbers (Schruben, 2010) shared among actions, but independent for each $i \in [n]$ and $j \in [m]$. We define the *return along a trajectory*

$$\mathrm{Ret}(x_0, a_0, \ldots, x_{h-1}, a_{h-1}) \doteq \sum_{t=0}^{h-1} \gamma^t r(x_t, a_t),$$

which provides a biased estimate of the value of the state-action pair (X_i, a). Rather than assuming that this bias is bounded, as done by Lazaric et al. (2016), we will require that the bias of the differences between the returns along two trajectories (for the same state but different actions) be bounded, as seen in Assumption 3.2.2. If the rewards are uniformly bounded in absolute value by $r_{\max}$, then Assumption 3.2.2 is satisfied with $B_{\pi,h} = 4\frac{\gamma^h}{1-\gamma} r_{\max}$. However, Assumption 3.2.2 can be satisfied with smaller values, depending on the statistical nature of the trajectories. For example, informally, if the rollout policy π is likely to recover from mistakes in a long trajectory (and yield close returns for two long trajectories starting from the same state but different actions), we can expect the bias in Assumption 3.2.2 to be much smaller than $4\frac{\gamma^h}{1-\gamma} r_{\max}$, especially for large γ. The truncation bias bounded by Assumption 3.2.2 can help control the variance of the returns, which has may negatively affect the performance of classifiers and CBPI (Pires and Szepesvári, 2015).

Assumption 3.2.2. *There exists a non-negative constant* $B_{\pi,h}$ *s.t.*

$$\mathbb{E}\left(\max_{a,a'} (V^\pi(X, a) - V^\pi(X, a')) - (\mathrm{Ret}(T_a) - \mathrm{Ret}(T_{a'})) \right) \leq B_{\pi,h}.$$

In the instance of CBPI analyzed by Lazaric et al. (2016), the Classifier procedure is required minimize the empirical classification cost, which (as discussed in Section 2.2) can be computationally hard. Therefore, we will resort to empirical surrogate-risk minimization with L^{LLW}, so that we can benefit from the guarantees from Chapter 2. We want to set up the classification problem so that we can satisfy Assumption 2.4.2 (Section 2.4). To that end, we the state space $\mathcal{X}$ as the input space, take $\mathcal{Y} = \mathcal{A}$, and define

$$C'_{i,j,a} \doteq \max_{a'} \operatorname{Ret}(T_{i,j,a'}) - \operatorname{Ret}(T_{i,j,a}) \tag{3.2.2}$$

and $C \doteq \max_{a'} \operatorname{Ret}(T_{a'}) - \operatorname{Ret}(T_a)$. (We need the costs to be non-negative, so we cannot take $C'_{i,j,a} = -\operatorname{Ret}(T_{i,j,a})$ unless the returns are non-positive with probability one.)

We also choose φ in L^{LLW} to be φ^{hinge}, which has calibration function $\delta(\varepsilon) = \varepsilon$ (Ávila Pires et al., 2013, Table 1). Choosing φ^{hinge} allows us to satisfy Assumption 2.4.3, Assumption 2.4.4 with $\mathrm{Lip}_{\varphi^{\mathrm{hinge}}}(t) = 1$ for all $t \geq 0$, and Assumption 2.4.5 with $\varphi_{\max,\mathcal{H}_{\phi,B}} = 1 + BB^*$. In order to use our calibration-function results (see Section 2.3.2) to convert surrogate risk bounds into true risk bounds, we must ensure that all scores output by score functions in $\mathcal{H}_{B,\phi}$ sum to zero, so we will impose Assumption 3.2.3 on ϕ.

Assumption 3.2.3. *For all $x \in \mathcal{X}$, $\sum_{a=1}^{|\mathcal{A}|} \phi(x,a) = \mathbf{0}_d$.*

The policy output in an iteration of CBPI will be the output of the Classifier procedure outlined in Algorithm 3.2.3, that is, the classifier outputs $\hat{\pi} \doteq f \circ \widehat{H}$, where $\widehat{H}$ is as given in (2.4.2) (page 29), with $L = L^{\mathrm{LLW}}$, $\varphi = \varphi^{\mathrm{hinge}}$ and $\mathcal{H} = \mathcal{H}_{\phi,B}$. The sample given to the Classifier procedure will be as given in (2.4.1) (page 28):

$$S \doteq ((X_1, C'_{1,1}, \ldots, C'_{1,m}), \ldots, (X_n, C'_{n,1}, \ldots, C'_{n,m})).$$

We can then use Theorems 2.4.11 and 2.4.14 to obtain true risk bounds for the classifier $\hat{\pi}$.

Algorithm 3.2.3 CLASSIFIER procedure analyzed in this section.

input: A sample $((x_1, c_1), \dots, (x_n, c_n)) \in (\mathcal{X} \times \mathbb{R}^{|\mathcal{A}|})^n$, a set of score functions $\mathcal{H} \subset (\mathbb{R}^{|\mathcal{A}|})^{\mathcal{X}}$

output: A classifier $f \circ h$

procedure CLASSIFIER($(x_1, c_1), \dots, (x_n, c_n), \mathcal{H}$)

return

$$h \in \operatorname*{argmin}_{h' \in \mathcal{H}} \frac{1}{n|\mathcal{A}|} \sum_{i=1}^{n} \sum_{a=1}^{|\mathcal{A}|} c_{i,a} \varphi(h(x_i)_a) \tag{3.2.3}$$

end procedure

3.2.2 True Risk Bounds

Our first result is Theorem 3.2.4, which is obtained by combining Theorems 2.3.4 and 2.4.11. Differently from the surrogate risk bounds, which had mild (or constant) scaling with $|\mathcal{A}|$, the scaling of the true risk bounds in Theorem 3.2.4 with $|\mathcal{A}|$ is linear. This scaling comes from a limitation of the calibration analysis, as discussed in Section 2.5. The factor of $|\mathcal{A}|$ also scales the approximation error, which in Theorem 3.2.4 is given in terms of L^{LLW}, not $\frac{1}{|\mathcal{A}|} L^{\mathrm{LLW}}$. A choice of m that minimizes the bound in Theorem 3.2.4 (up to constant factors) is $m = \left(\frac{C_{\max}}{C'_{\max}}\right)^2$. In contrast, Lazaric et al. (2016) recommend (based on their bounds) that one choose $m = 1$. The choice informed by Theorem 3.2.4 is somewhat in line with some of the conclusions of Pires and Szepesvári (2015), which stated that variance in the cost estimates could negatively affect the performance of CLASSIFIER and CBPI. Although we did not express the bounds in terms of variances, we carried out our analysis so that we could understand whether there could be any benefit in using more than one rollout ($m > 1$) for the cost estimates. Asymptotically, there is no benefit, but m can be used to eliminate potentially large constant factors from the bound, since $C_{\max}$ could be much larger than $C'_{\max}$. The bias of the truncated rollouts, $B_{\pi,h}$, is also different from the one used by Lazaric et al. (2016). Our choice emphasizes that truncation of the trajectories can help, since we can trade-off $B_{\pi,h}$, $C_{\max}$, and $C'_{\max}$. In the illustrative scenario where π is likely to recover early from the forceful choice of A_0 in the

rollouts, we would benefit from truncating the trajectories with a small h, which would not produce a large bias $B_{\pi,h}$, but would have the potential to significantly decrease $C_{\max}$ and $C'_{\max}$. Theorem 3.2.4 indicates that proper rollout truncation may benefit performance.

Theorem 3.2.4. *Consider a sample S satisfying Assumptions 3.2.1 and 3.2.2 , where costs defined as in* (3.2.2) *and trajectories as in* (3.2.1)*). Consider also ϕ satisfying Assumptions 3.2.3 and 2.4.6. For any $\delta \in (0,1)$ we have with probability at least $1-\delta$ that*

$$\begin{aligned}
&\sup_{\pi' \in \Pi} \mathbb{E}\big(V^{\pi}(X, \pi'(X))\big) - \mathbb{E}\big(V^{\pi}(X, \hat{\pi}(X))\big|\, S\big) \\
&\leq \inf_{h \in \mathcal{H}} R^{\text{surr}}_{L^{\text{LLW}}}(h) - \inf_{h \in (\mathbb{R}^{|\mathcal{A}|})^{\mathcal{X}}} R^{\text{surr}}_{L^{\text{LLW}}}(h) + B_{\pi,h} \\
&\quad + 2|\mathcal{A}|(1 + BB_*)\sqrt{\frac{32d}{n} \ln \frac{20(1+nm)}{\delta}} \left(2C'_{\max} + C_{\max}\sqrt{\frac{1}{m}}\right)
\end{aligned}$$

We are also able to use our results based on Frobenius-norm coverings to get Theorem 3.2.5. As expected from how Theorem 2.4.11 compares to Theorem 2.4.14, we see a logarithmic dependence on the number of features the bound of Theorem 3.2.5, as opposed to the square-root dependence seen in Theorem 3.2.4. As in Theorem 3.2.4, the bound in Theorem 3.2.5 scales with $|\mathcal{A}|$, also as a result of the calibration function used. More importantly, Theorem 3.2.5 informs us to the same choice of m as Theorem 3.2.4 to minimize the bound in Theorem 3.2.5 (up to constant factors), that is $m = \left(\frac{C_{\max}}{C'_{\max}}\right)^2$. However, the impact of this choice is not as favorable as in Theorem 3.2.4, because if the ratio $\frac{C'_{\max}}{C_{\max}}$ is too small in the sense of (3.2.4) (which would mean a larger benefit in using the suggested m), the bound in Theorem 3.2.5 will be given by ε_2.

Theorem 3.2.5. *Consider a sample S satisfying Assumptions 3.2.1 and 3.2.2 , where costs defined as in* (3.2.2) *and trajectories as in* (3.2.1)*). Consider also ϕ satisfying Assumptions 3.2.3 and 2.4.6. For any $\delta \in (0,1)$ we have with probability*

at least $1-\delta$ *that*

$$\sup_{\pi' \in \Pi} \mathbb{E}\big(V^{\pi}(X, \pi'(X))\big) - \mathbb{E}\big(V^{\pi}(X, \hat{\pi}(X))\big|\, S\big)$$
$$\leq \inf_{h \in \mathcal{H}} R^{\text{surr}}_{L^{\text{LLW}}}(h) - \inf_{h \in (\mathbb{R}^{|\mathcal{A}|})^{\mathcal{X}}} R^{\text{surr}}_{L^{\text{LLW}}}(h) + (\varepsilon_1 \vee \varepsilon_2) + B_{\pi,h}$$

where

$$\varepsilon_1 = 2|\mathcal{A}|(1+BB_*)\sqrt{\frac{32}{n}\ln\frac{20(2d+1)}{\delta}}\left(2C'_{\max} + C_{\max}\sqrt{\frac{1}{m}}\right),$$

$$\varepsilon_2 = 12|\mathcal{A}|BB_*C_{\max}\sqrt{\frac{32}{n|\mathcal{A}|}\ln\frac{20(2d+1)}{\delta}}.$$

Moreover, if

$$\sqrt{\frac{1}{m}} \vee \frac{C'_{\max}}{C_{\max}} \geq \frac{BB_*}{1+BB_*}\sqrt{\frac{1}{|\mathcal{A}|}}. \tag{3.2.4}$$

then $\varepsilon_1 \geq \varepsilon_2$*, otherwise* $\varepsilon_1 \leq 6\varepsilon_2$*.*

To conclude our analysis, we will present Theorem 7 of Lazaric et al. (2016), which gives us policy error bounds for CBPI, and with which we can immediately combine Theorems 3.2.4 and 3.2.5.

3.2.3 Policy Error Bounds

So far, we have described the setup for a single iteration of CBPI. For the policy iteration with $K \in \mathbb{N}$ iterations, we assume that we are given a policy $\pi_0 \in \Pi$, a sequence of measures $\rho_0, \ldots, \rho_{K-1}$ over $\mathcal{X}$, and the Classifier method. In Definition 3.2.6 we introduce the samples used by Classifier at each iteration.

Definition 3.2.6. *For each* $k \in [K]$*, let* $\pi_k \doteq \mathrm{A}(S_{k-1})$ *and where each* S_k *is an independent sample satisfying*

$$S_k = ((X_1, C'_{1,1}, \ldots, C'_{1,m}), \ldots, (X_n, C'_{n,1}, \ldots, C'_{n,m}))$$

with $(X_1, \ldots, X_n) \sim \rho^n_{k-1}$*, the costs* $C'_{i,j}$ *defined as in* (3.2.2)*, and the trajectories as in* (3.2.1)*, with* $\pi = \pi_{k-1}$*.*

We also define, for $V \in \mathcal{V}$,

$$\|V\|_{1,\mu} \doteq \int_{\mathcal{X}} |V(x)| d\mu(x),$$

and

$$\|V\|_{\infty} \doteq \sup_{x \in \mathcal{X}} |V(x)|,$$

under the respective assumptions that value functions are bounded in norm.

Now that we have outlined the setting studied by Lazaric et al. (2016), we are able to report their policy error bounds as a function of the true risk of the classifiers used: Theorem 3.2.7 gives us policy error bounds for the policy π_K obtained after K iterations of CBPI, using Classifier at each iteration. We can then immediately plug Theorems 3.2.4 and 3.2.5 as (3.2.7) into Theorem 3.2.7 to obtain policy error bounds for our practical classifiers. We see that the concentrability coefficients, the quality of the initial policy π_0 and γ affect our decision of when to stop the iteration, but if we have a fixed "data budget" we need to look at the bound differently: It may seem as though the bounds scale logarithmically on the number of iterations, but the fact that the samples $S_1, \ldots, S_K$ are independent introduces a factor of $\sqrt{K}$ to the bounds, since we would observe $\frac{n}{\sqrt{K}}$ states at each iteration.

Theorem 3.2.7 (Adapted from Theorem 7, Lazaric et al., 2016)**.** *Given $K \in \mathbb{N}$; measures μ, ρ and $\rho_1 = \ldots = \rho_K = \rho$ over $\mathcal{X}$; a cost-sensitive classification algorithm* Classifier*; and $\pi_0 \in \Pi$, assume that* Classifier *runs algorithm* A *at each iteration with inputs $S_1, \ldots, S_k$ defined as in Definition 3.2.6. Assume that for any $\delta \in (0,1)$ and $k \in [K]$ we have with probability at least $1 - \delta$*

$$\sup_{\pi' \in \Pi} \mathbb{E}\left(V^{\pi_{k-1}}(X, \pi'(X))\right) - \mathbb{E}\left(V^{\pi_{k-1}}(X, \hat{\pi}(X)) \,|\, S_{k-1}\right) \leq \varepsilon_{k,\delta}$$

for some $\varepsilon_{k,\delta}$, where $X \sim \rho_{k-1} = \rho$.

If the concentrability coefficient $C_{\mu,\rho}$ satisfies Assumption 1 of Lazaric et al. (2016), then for any $\delta \in (0,1)$ we have with probability at least $1 - \delta$

$$\|V^* - V^{\pi_K}\|_{1,\mu} \leq \gamma^K \|V^* - V^{\pi_0}\|_{1,\mu} + \frac{C_{\mu,\rho}}{(1-\gamma)^2} \max_{k \in [K]} \varepsilon_{k, \frac{\delta}{K}},$$

Alternatively, if the concentrability coefficient $C_{\infty,\rho}$ satisfies Assumption 2 of Lazaric et al. (2016), then for any $\delta \in (0,1)$ we have with probability at least $1-\delta$

$$\|V^* - V^{\pi_K}\|_\infty \leq \gamma^K \|V^* - V^{\pi_0}\|_\infty + \frac{C_{\infty,\rho}}{(1-\gamma)^2} \max_{k \in [K]} \varepsilon_{k,\frac{\delta}{K}}.$$

3.3 Conclusion

In this chapter, we have used the surrogate and true risk bounds presented in Chapter 2 in order to extend the analysis of Lazaric et al. (2016) so that it applies to classification methods that can be used in practice, in particular, the empirical risk minimizer of L^{LLW} with φ^{hinge} (a convex surrogate loss) over $\mathcal{H}_{\phi,B}$ (a space of linear classifiers), which can be seen to correspond to a cost-sensitive multiclass SVM.

In Section 2.5, we have touched on fast rates based on the Mammen-Tsybakov noise condition and strongly convex surrogate losses. Bounds of this nature are quite relevant to CBPI. Farahmand et al. (2014) have used the Mammen-Tsybakov noise condition to obtain fast rates for the empirical risk minimizer of classification cost. As remarked by them, large gaps between optimal and sub-optimal actions should be detected by the classifier, and taking sub-optimal actions is not really an issue where gaps are small. This intuition translates as true risk bounds with faster rates, due to the Mammen-Tsybakov noise condition.

Following the discussion in Section 2.5, if the Mammen-Tsybakov noise condition can be leveraged to obtain cost-sensitive classification surrogate and true risk bounds with faster rates, then it would be interesting to see them applied in order to extend the analysis of Farahmand et al. (2014).

Also as mentioned in Section 2.5, refining the analysis of Section 2.4 to incorporate the Mammen-Tsybakov noise condition would require replacing Hoeffding's inequality with Bernstein's inequality. This would expose the cost variances and help analyze some of the phenomena observed by Pires and Szepesvári (2015)—in particular, that large cost variance could render classification methods ineffective. An interesting development, then,

would be to seek a better understanding of the truncation bias $B_{\pi,h}$ (see Assumption 3.2.2), in order to make a more informed bias-variance tradeoff.

One can also consider relaxing the i.i.d. assumption underlying the data provided to the classifier, for example, introducing fast mixing (Antos et al., 2008b) in $(X_1, \ldots, X_n)$. Alternatively, fast mixing could be assumed for the trajectories $T_{i,a}$, which could be a start for better understanding the truncation bias $B_{\pi,h}$. In particular, can fast-mixing trajectories make a case in favor of having small h (trajectories truncated early) without incurring a large return bias?

Directions pertaining to empirical evaluations of CBPI can also be pursued. It remains to be understood the practical benefits of a particular choice of surrogate loss in the context of CBPI. As mentioned in Section 2.5, there are empirical comparisons between surrogate losses in the context of cost-insensitive classification, but a comparison in the cost-sensitive case is lacking. In particular, there is very little understanding (theoretical or empirical) of the cost-sensitive case with random costs. Pires and Szepesvári (2015) have shown a simple MDP where large variance in the cost estimates lead to poor classifiers (classifiers with large true risk) and, as a consequence, policies that do not perform well. Therefore, it would be interesting to understand how to construct low-variance cost estimates that yield effective classifiers. One way to construct such estimates is the aforementioned truncation of Monte Carlo rollouts, another are the hybrid cost estimates (Gabillon et al., 2013; Farahmand et al., 2014).

Chapter 4

Model-Based Reinforcement Learning with Factored Semi-Linear Models

An alternative to approximately solving an MDP by policy iteration (*e.g.* GPI, CPI, CBPI) is to learn a *model* of the MDP dynamics $\mathcal{P}$, and somehow use the model instead of $\mathcal{P}$ to perform dynamic programming. This is called *model-based reinforcement learning* (MBRL).

According to Sutton and Barto (1998) (Section 9.1, p. 227), a model should approximate $\mathcal{P}(x, a)$ for every $(x, a) \in \mathcal{X} \times \mathcal{A}$. While this requirement is sufficient for dealing with the statistical challenges of DP, is often difficult to satisfy and it does not address the potential intractability of DP. The recent years have witnessed a renewed interest in MBRL, with the emergence of approaches that did not necessarily try to approximate $\mathcal{P}(x, a)$ for every $(x, a) \in \mathcal{X} \times \mathcal{A}$, and eventually led to a quite flexible concept of what a model should be.

Barreto et al. (2011); Kveton and Theocharous (2012) and Precup et al. (2012), building on the seminal work of Ormoneit and Sen (2002), studied various approaches to stochastic factorizations of the transition probability kernel, while Grünewälder et al. (2012) proposed to use RKHS embeddings to approximate the transition kernel, with further enhancements proposed recently by Lever et al. (2016). A key common feature of these otherwise distant-looking works is that once the model is set up, it leads to a policy in

a computationally efficient way (i.e., in polynomial time and space in the size of the model). Having realized that this is not a mere coincidence, Yao et al. (2014) introduced the concept of factored linear models, which keeps the advantageous computational properties, while generalizing all previous works. While efficient computation is a necessity, efficient learning and good performance of the policy are equally important. In this chapter we focus on the second of these criteria, namely the performance of the policy derived from the model, more specifically, the policy error as a function of model errors. The argument for omitting the learning part for the time being is that one should better understand first what errors need to be controlled because this will influence the choice of the learning objective and hence the algorithms (we also note in passing that, in the above-mentioned examples, the statistical analysis of the model learning algorithms is well understood by now).

We are not the first to consider the performance of the policy (the policy error) as a function of the model errors. Most of the previously mentioned works also present policy error bounds of this nature. However, all these works derive bounds that express model errors in a supremum norm. While the supremum norm is a convenient choice when working with MDPs (which give the theoretical foundations in these works), an observation that goes back to at least Whitt (1978) is that the supremum norm is also known to be a rather unforgiving metric: In learning settings, when data comes from a large cardinality set, and the data may have an uneven distribution, while the objects of interest lack appropriate smoothness, or other helpful structural properties, we expect errors measured in the supremum norm to decrease rather slowly. Furthermore, most learning algorithms aim to reduce some weighted norms, hence deriving bounds for the supremum norm is neither natural, nor desirable. Can existing bounds of the policy error from the MBRL literature be extended to other norms? In the analogue context of approximate dynamic programming methods, Munos (2003) pioneered a technique to allow the use of weighted L^p-norms to bound the

policy error, while in the context of approximate linear programming (ALP), de Farias and Van Roy (2003) proposed a different technique to allow the use of weighted supremum norms, both leading to substantial further work (Buşoniu et al., 2010a, 2012). While the use of weighted norms is a major advance, these bounds do not come without any caveats. In particular, in ALP, the bounds rely on the similarity of the so-called constraint sampling distribution to the stationary distribution μ^* of the optimal policy, while in ADP they rely on the similarity of the data sampling distribution and the start-state distribution, leading to hard to control error terms. Can this be avoided by model-based approaches?

In this chapter, we present bounds on the policy error of policies derived from factored semi-linear models in MBRL, following the work of Ávila Pires and Szepesvári (2016b), but mildly generalizing the factored linear model framework. The policy error is bounded in supremum, weighted supremum and weighted L^p norms (Theorems 4.4.3, 4.4.5 and 4.4.7). The results hold under some conditions: the left factor of the approximate factorization of the transition kernel must satisfy a mild boundedness condition (Assumption 4.3.7), while the right is not constrained. Ávila Pires and Szepesvári (2016b) assume that the right factor is a join-homomorphism[1]. We introduce, however, a third component of the factored semi-linear models, which is restricted to be a collection of linear operators (which is, itself an operator). In the work of Ávila Pires and Szepesvári (2016b), each operator in this collection coincided with the (linear) right factor of the model. The last condition is that the product of the third-component operator and the left factor satisfy a norm constraint. This last condition is not mild as the others, but it i) generalizes the conditions used to derive previous policy error bounds; and ii) can be easier to enforce as it constrains the norm of a low-dimensional operator, unlike the analogue constraints in previous works. In addition, relinquishing the join-homomorphism condition used

[1] An operator J from a semi-lattice $(\mathcal{U}, \vee)$ into the semi-lattice $(\mathcal{U}', \vee)$ is a *join-homomorphism* if $J(U \vee U') = (JU) \vee (JU')$ for any $U, U' \in \mathcal{U}$ (see Ávila Pires and Szepesvári, 2016b, Assumption 2)).

by Ávila Pires and Szepesvári (2016b) allows us to generalize essentially all previous work on MBRL that uses the model to compute a policy using DP.

We recover results for unfactored semi-linear models that satisfy a contraction assumption, including previously proposed supremum norm bounds. In addition to being able to recover previous results, we also provide a new type of analysis, which has interesting implications. The new analysis shows that MBRL can in fact escape the sensitivities in ALP and ADP (cf. Theorem 4.4.7, term ε_1), answering the above major question on the positive. In fact, the new bound also shows the potential for better scaling with the discount factor, which is another surprising result. We attribute this success to the systematic use of the language of Banach lattices, which forced us to discover amongst other things a definition of *mixed* norms for action-value functions which is general, yet makes the so-called value selection operators non-expansions (cf. Proposition 4.3.2). For the skeptics who believe that MBRL is "hard" because the derived policy cannot be good before the model approximates "reality" uniformly everywhere, we point out that already the first ever bound derived for policy error in MBRL (due to Whitt, 1978) shows that the model has to be accurate only in an extremely localized way. Our bounds also share this characteristic of previous bounds.

Our analysis builds on techniques borrowed from approximate policy iteration (API) and approximate linear programming (ALP), and provide new insights to existing results for ALP (Proposition 4.3.5). However, the MBRL setup we consider is nevertheless different from API and ALP, so the connections in our proofs are not a mere translation of API or ALP results to MBRL, as we will explain in Section 4.5, which is also attested by the novel features of our bounds.

In Section 4.2, we will introduce factored semi-linear models. After this, we state our assumptions in Section 4.3, present our main results in Section 4.4, and close with placing our work in the context of existing work, and providing an outlook for future work in Section 4.5. The novel proofs can be found in Appendix A.3, whereas the proofs of accessory results from

Ávila Pires and Szepesvári (2016b) are not presented here.

4.1 Preliminaries

In this section, we build on the definitions given in Section 1.1.2. There, we have introduced the Banach spaces $(\mathcal{V}, \|\cdot\|_{\mathcal{V}})$ and $(\mathcal{V}^{\mathcal{A}}, \|\cdot\|_{\mathcal{V}^{\mathcal{A}}})$, which have been assumed to contain the value functions and action value functions, respectively, of all deterministic stationary policies. In this chapter, we will choose $\|\cdot\|_{\mathcal{V}}$ to be supremum, weighted supremum, or $L^p(\mu)$ norms. The choice of $\|\cdot\|_{\mathcal{V}^{\mathcal{A}}}$ will in general depend on that of $\|\cdot\|_{\mathcal{V}}$, but this will be made clear in the actual context. As mentioned in Chapter 1, $\mathcal{V}^{\mathcal{A}}$ can also be identified with the set of real-valued functions with domain $\mathcal{X} \times \mathcal{A}$ (since $\mathcal{A}$ is finite). Recall that we are using V^a as an alternate notation to $V(a)$, and that we denote by $\mathcal{P}^a$ the $\mathcal{V} \to \mathcal{V}$ right linear operator defined by $(\mathcal{P}^a V)(x) \doteq \mathbb{E}(V(X_{t+1}) | X_t = x, A_t = a)$ (we have assumed that $V \in \mathcal{V}$ implies integrability, so the integrals are well defined). We also view $\mathcal{P}^a$ as a left linear operator, acting over the space of probability measures defined over $\mathcal{X}$: $\mathcal{P}^a : \mathcal{M}_1(\mathcal{X}) \to \mathcal{M}_1(\mathcal{X})$, $(\mu\mathcal{P}^a)(\mathcal{X}') = \int d\mu(x) d\mathcal{P}^a(\mathcal{X}'|x)$, $\mu \in \mathcal{M}_1(\mathcal{X})$, $\mathcal{X}' \subset \mathcal{X}$. In what follows, whenever a norm is uniquely identifiable from its argument, we will drop the index of the norm denoting the underlying space.

We extend *Bellman return operator* defined in Section 1.1.2 to accept any linear operator $\mathcal{J} : \mathcal{V} \to \mathcal{V}^{\mathcal{A}}$ by defining: $T_{\mathcal{J}} V \doteq r + \mathcal{J} V$ $(V \in \mathcal{V})$. Then $T_{\mathcal{P}}$ is the Bellman return operator originally defined in Section 1.1.2. Recall that the *maximum selection* operator $M : \mathcal{V}^{\mathcal{A}} \to \mathcal{V}$ is defined by $(MV)(x) \doteq \max_a V^a(x)$.

Then $MT_{\mathcal{P}}$ corresponds to the *Bellman optimality operator* and the optimal value function is known to satisfy $V^* = MT_{\mathcal{P}} V^*$ (Puterman, 1994, Section 6.2). a non-linear fixed-point equation, which is known as the *Bellman optimality equation*. The *greedy operator* $G : \mathcal{V}^{\mathcal{A}} \to \Pi$, which selects the maximizing actions chosen by M, is defined by $GV(x) \doteq \operatorname{argmax}_a V^a(x)$

($x \in \mathcal{X}$, with ties broken arbitrarily). Recall that GTV^* is an optimal policy (Puterman, 1994, Section 6.2.4).

As we have postulated in the introduction, RL methods are, one way or another, trying to perform DP efficiently. In this chapter, we will do so with *online planning*. In the online planning problem, we wish to compute, at any given state x, an action that a near-optimal policy would take. The attribute "online" signifies that one is allowed some amount of calculation for each state. By collecting all actions at all states, a planning method defines a policy $\hat{\pi}$. Apart from computation, planning methods are compared by how good the policy they return is, *i.e.*, by the policy error of $\hat{\pi}$. One approach to efficient online planning is to use an abstract model which i) contains relevant information about the MDP, ii) can be efficiently constructed, and iii) allows $\hat{\pi}(x)$ to be computed efficiently at any state x. The online planning we are interested in uses a special type of abstract models, called *factored semi-linear models*.

4.2 Factored Semi-Linear Models

In this section, we define factored semi-linear models, the core of our MBRL approach. We also show examples of MBRL approaches that use factored linear models. We are not aware of methods that rely on non-linear factored semi-linear models, so the generalization from factored linear models, which were introduced by Yao et al. (2014), and factored semi-linear models is mild and technical.

In a *factored linear model* we approximate the MDP's stochastic kernel $\mathcal{P}$ as the product of two operators, $\mathcal{Q}\mathcal{R}$. Similarly to factored linear models, Yao et al. (2014), we have $\mathcal{Q} \doteq (\mathcal{Q}^a)_{a\in\mathcal{A}}$, with each $\mathcal{Q}^a : \mathcal{W} \to \mathcal{V}^{\mathcal{A}}$ linear, and also a collection of linear operators $\mathcal{R}' \doteq (\mathcal{R}'^a)_{a\in\mathcal{A}}$ and each $\mathcal{R}'^a : \mathcal{W} \to \mathcal{V}$ linear. The operator $\mathcal{R} : \mathcal{W} \to \mathcal{V}$ (with $\mathcal{W}$ defined below), differently from the one proposed by Yao et al. (2014), need not be linear or equal to each $\mathcal{R}'^a$. The term "semi-linear" refers to the fact that each $(\mathcal{Q})^a$ is linear, but

not $\mathcal{R}$. While only $\mathcal{Q}$ and $\mathcal{R}$ form the approximation of $\mathcal{P}$, we will see that the operator $\mathcal{R}'$ is also essential to the factored semi-linear model approach we use, and to the construction of $\hat{\pi}$.

The space $\mathcal{W} = (\mathcal{W}, \|\cdot\|_{\mathcal{W}})$ is a Banach space of functions with (measurable) domain $\mathcal{I}$, and $\mathcal{W}^{\mathcal{A}}$ is a Banach space of $\mathcal{A} \to \mathcal{W}$ (cf. $\mathcal{V}$, $\mathcal{V}^{\mathcal{A}}$, $\mathcal{W}$ and $\mathcal{V}^{\mathcal{A}}$). We will refer to $\mathcal{W}$ and $\mathcal{W}^{\mathcal{A}}$ as the compressed spaces, and, occasionally, the spaces $\mathcal{V}$ and $\mathcal{V}^{\mathcal{A}}$ will be called uncompressed. These names come from the fact that often we will want to choose $\mathcal{I}$ to be "small". In fact, for computational reasons one should choose $\mathcal{I}$ to be finite, in which case $\mathcal{W}$ will be a finite-dimensional Euclidean space. We also allow infinite $\mathcal{I}$, so that we can then use $\mathcal{I} = \mathcal{X}$ and compare the tightness of our results to existing results that consider unfactored linear models.

In this work, for simplicity, we assume that the reward function r remains the same in the factored linear model (the extension of our results to the case when the reward function is also approximated is routine). Formally, we will call a tuple of the form $\langle \mathcal{X}, \mathcal{A}, \mathcal{Q}, \mathcal{R}, \mathcal{R}', r \rangle$ a factored semi-linear model, where $\mathcal{Q}$, $\mathcal{R}$ and $\mathcal{R}'$ are as above. While a factored linear model defines a *pseudo-MDP* (Yao et al., 2014), a factored semi-linear model defines a generalization of pseudo-MDPs where the transition dynamics are non-linear.

We must define some additional operators in order to describe how we use factored linear models to derive policies. We define the shorthands $T_{\mathcal{R}'\mathcal{Q}} \doteq \mathcal{R}' T_{\mathcal{Q}} = \mathcal{R}' r + \gamma \mathcal{R}' \mathcal{Q}$ (the equality holds by linearity of $\mathcal{R}'$) and $T_{\mathcal{Q}\mathcal{R}} \doteq T_{\mathcal{Q}} \mathcal{R} = r + \gamma \mathcal{Q}\mathcal{R}$ (by definition of the Bellman return operator). Finally, $M' : \mathcal{W}^{\mathcal{A}} \to \mathcal{W}$, the compressed counterpart of the maximum selection operator M, is defined by $(M'w)(i) = \max_{a \in \mathcal{A}} w^a(i)$ $(i \in \mathcal{I})$, and the compressed counterpart of the greedy operator is G', mapping elements of $\mathcal{W}^{\mathcal{A}}$ to policies over $\mathcal{I}$. (*i.e.*, $M'^{G'w} w = M'w$ for $w \in \mathcal{W}^{\mathcal{A}}$). The relationship between these operators is shown on Figure 4.1.

The *factored semi-linear model approach to reinforcement learning* is as follows:

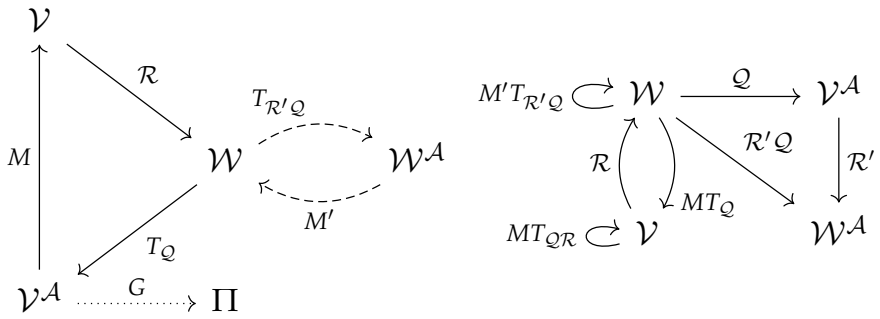

Figure 4.1: Commutative diagrams showing the operators and the spaces that they act on.

Given the factored linear model $\langle \mathcal{X}, \mathcal{A}, \mathcal{Q}, \mathcal{R}, \mathcal{R}', r \rangle$, we take the policy

$$\hat{\pi} \doteq G T_{\mathcal{Q}} u^*, \tag{4.2.1}$$

where

$$u^* = M' T_{\mathcal{R}'\mathcal{Q}} u^* \,. \tag{4.2.2}$$

that is, the policy $\hat{\pi}$ does a *Bellman lookahead* with $T_{\mathcal{Q}}$ from $u^* \in \mathcal{W}$, a function that satisfies a fixed-point equation. Note that even when $\mathcal{X}$ is very large, or infinite, $\mathcal{W}$ can be finite dimensional, in which case a good approximation to u^* can often be found in a computationally efficient manner, for example by iterating $u_{k+1} = M' T_{\mathcal{R}'\mathcal{Q}} u_k$, which can be seen as a form of value iteration (Yao et al., 2014). The dashed lines on the left subfigure on Figure 4.1 show that this computation can be done over the compressed spaces $\mathcal{W}$ and $\mathcal{W}^{\mathcal{A}}$. The diagram also shows that once u^* is found, $T_{\mathcal{Q}}$ extends this function to $\mathcal{V}^{\mathcal{A}}$, from where using the greedy operator G one obtains a policy. Note that in the applications the policy itself does not need to be explicitly represented, but the actions that the policy takes in a particular state $x \in \mathcal{X}$ can be computed "on demand" given u^* and the Bellman return operator $T_{\mathcal{Q}}$. (The right-hand side figure shows some more useful relationships between the operators involved.) We will say that this approach is *viable* when u^* is well-defined. In our bounds, we will make use of an additional policy, $\hat{\pi}' : \mathcal{I} \to \mathcal{A}$, defined by

$$\hat{\pi}' \doteq G' T_{\mathcal{R}'\mathcal{Q}} u^*, . \tag{4.2.3}$$

This is a policy over $\mathcal{I}$ derived from u^* and its use is merely technical.

Factored semi-linear models (presently, also factored linear models) allow one to analyze modeling errors in seemingly distant model-based planning methods in a *unified manner*. This will be illustrated soon by describing how models proposed in numerous previous works can be written in a factored form (this was also shortly mentioned by Yao et al., 2014). Before describing these previous models, we need some more definitions, to be able to describe the differences and similarities between them. In particular, the models will differ in terms of whether $\mathcal{R}$ is stochastic, or more specifically $\mathcal{R}$ is also a point-evaluator. Recall that the operator $\mathcal{R}$ is stochastic if $\inf_{V \geq 0} \inf_x (\mathcal{R}V)(x) \geq 0$ and $\mathcal{R}\mathbf{1}_{\mathcal{V}} = \mathbf{1}_{\mathcal{W}}$ where $\mathbf{1}_{\mathcal{V}}(x) = 1$ for all $x \in \mathcal{X}$ and $(\mathbf{1}_{\mathcal{W}})_i = 1$ for all $i \in \mathcal{I}$. Here, we started to use w_i instead of $w(i)$ to reduce clutter. Also, we say that $\mathcal{R}$ is a *point-evaluator* if $\mathcal{I}$ indexes elements of $\mathcal{X}$ and $(\mathcal{R}V)_i = V(x_i)$ for all $i \in \mathcal{I}, V \in \mathcal{V}$. Note that point evaluators are stochastic. Choosing $\mathcal{I} = \mathcal{X}$ allows us to choose $\mathcal{R}$ to be the identity, which becomes a point evaluator when choosing $x_i = i$, $i \in \mathcal{I}$.

When $\mathcal{R}$ is a point selector and $\mathcal{R}'^a = \mathcal{R}$ for all $a \in \mathcal{A}$, a short direct calculation shows that $\mathcal{R}M = M'\mathcal{R}'$, which means that on Figure 4.1 the solid cycle and the dashed cycle starting from $\mathcal{W}$ are equivalent and we can interweave solid and dashed lines. For example, starting from $\mathcal{V}$: $MT_{\mathcal{Q}}M'T_{\mathcal{R}'\mathcal{Q}}\mathcal{R} = (MT_{\mathcal{Q}\mathcal{R}})^2$. The equivalence $M'T_{\mathcal{R}'\mathcal{Q}} = \mathcal{R}MT_{\mathcal{Q}}$ gives that $U^* \doteq MT_{\mathcal{Q}}u^*$ is a fixed point of $MT_{\mathcal{Q}\mathcal{R}}$, and that the identity $u^* = \mathcal{R}U^*$ also holds (see Theorem 4.4.1). It also follows that if $M'T_{\mathcal{R}'\mathcal{Q}}$ is a contraction (though $MT_{\mathcal{Q}\mathcal{R}}$ may not be), the factored semi-linear model approach (now a factored linear model approach due to the choice of $\mathcal{R}$) is viable. To the best of our knowledge, Ávila Pires and Szepesvári (2016b) were the first to make this observation. In all previous works, viability was achieved by assuming that $\mathcal{Q}$ and $\mathcal{R}$ are both stochastic, or that $\mathcal{R}$ is a point evaluator and $\mathcal{Q}\mathcal{R}$ is a non-expansion in supremum norm. (In both cases, both $MT_{\mathcal{Q}\mathcal{R}'}$ and $M'T_{\mathcal{R}'\mathcal{Q}}$ are contractions in supremum norm, so u^* is well-defined and the factored linear model approach is viable.)

With this, we are ready to present different instances of the factored linear model approach:

Example 4.2.1 (Kernel-based reinforcement learning). In kernel-based reinforcement learning (KBRL), introduced by Ormoneit and Sen (2002), $\mathcal{I}$ is indexing elements of $\mathcal{X}$, and $\mathcal{Q}$ is a stochastic operator constructed from kernel functions at elements of $S \doteq \{x_i : i \in \mathcal{I}\}$. Moreover,

(a) S is an i.i.d. sample from $\mathcal{X} \doteq \mathbb{R}^d$ and $\mathcal{R}$ is a point evaluator (Ormoneit and Sen, 2002); or

(b) S is a set of *reference states* and $\mathcal{R}$ is stochastic (Barreto et al., 2011; Kveton and Theocharous, 2012; Precup et al., 2012).

KBRL is viable because $\mathcal{Q}$ and $\mathcal{R}$ are stochastic, so $\mathcal{R}'\mathcal{Q}$ is also stochastic.

Example 4.2.2 (Pseudo-MDPs). Pseudo-MDPs (Yao et al., 2014) are factored linear models with a point evaluator $\mathcal{R}$. In pseudo-MDPs, $\mathcal{Q}$ is no longer stochastic, but $\mathcal{Q}\mathcal{R}$ is assumed to be a non-expansion in supremum norm (Grünewälder et al., 2012; Yao et al., 2014; Lever et al., 2016). It can be shown that under this assumption both $MT_{\mathcal{Q}\mathcal{R}}$ and $M'T_{\mathcal{R}'\mathcal{Q}}$ are contractions. In the approach of these authors, one should take $\tilde{\pi} \doteq GT_{\mathcal{Q}\mathcal{R}}U^*$, where U^* is the fixed point of $MT_{\mathcal{Q}\mathcal{R}}$. Our formulation still applies, though, because we can show that $u^* = \mathcal{R}U^*$ is the fixed point of $M'T_{\mathcal{R}'\mathcal{Q}}$ (see Theorem 4.4.1), so that $\tilde{\pi} = GT_{\mathcal{Q}\mathcal{R}}U^* = GT_{\mathcal{Q}}u^* = \hat{\pi}$. Here, $\mathcal{Q}$ is essentially learned using a penalized least-squared approach.

Example 4.2.3 (State aggregation). State aggregation (Whitt, 1978; Bertsekas, 2011) in MBRL generalizes KBRL. Here, too, $\mathcal{I}$ is an index set over $\mathcal{X}$, and $\{x_i : i \in \mathcal{I}\}$ is the set of reference states. In hard aggregation, $\mathcal{R}$ is a point evaluator, while in soft aggregation (Singh et al., 1995) it is stochastic.

Example 4.2.4 (MDP homomorphisms). MDP homomorphisms (Ravindran, 2004; Sorg and Singh, 2009) can be used for transfer learning in reinforcement learning. Here, $\mathcal{I}$ is not identified with an index set over $\mathcal{X}$. If $\mathcal{R}$ is a point-evaluator, we recover MDP homomorphisms *per se* (Ravindran, 2004),

and the more general case of $\mathcal{R}$ stochastic yields soft MRP homomorphisms (Sorg and Singh, 2009).

Example 4.2.5 (Unfactored linear models). It is possible to recover unfactored linear models as a special case of factored linear models by taking $\mathcal{W} = \mathcal{V}$, and $\mathcal{R}$ to be the identity mapping. For the approach to be viable, it is sufficient for $\mathcal{Q}$ to be stochastic, which is often assumed with unfactored linear models.

4.3 Assumptions

The purpose of this section is to state and discuss the assumptions that will be used in our subsequent results.

Our first assumption states that the operators $M : \mathcal{V}^{\mathcal{A}} \to \mathcal{V}$, $M' : \mathcal{W}^{\mathcal{A}} \to \mathcal{W}$, and the related policy based value selector operators $M^{\pi} : \mathcal{V}^{\mathcal{A}} \to \mathcal{V}$ and $M'^{\pi'} : \mathcal{W}^{\mathcal{A}} \to \mathcal{W}$ to be defined soon are non-expansions. The operator M^{π} is defined by $(M^{\pi}V)(x) \doteq V^{\pi(x)}(x)$ ($x \in \mathcal{X}$, $\pi \in \Pi$), while $(M'^{\pi'}w)_i \doteq w_i^{\pi'(i)}$ ($i \in \mathcal{I}$, $\pi' : \mathcal{I} \to \mathcal{A}$). Now, recall that an operator $J : \mathcal{E} \to \mathcal{F}$ mapping between Banach spaces $\mathcal{E} = (\mathcal{E}, \|\cdot\|_{\mathcal{E}})$, $\mathcal{F} = (\mathcal{F}, \|\cdot\|_{\mathcal{F}})$ is called a non-expansion when its Lipschitz constant does not exceed one. The Lipschitz constant of J is defined by

$$\mathrm{Lip}(J) \doteq \sup_{e,e' \in \mathcal{E} : e \neq e'} \frac{\|Je - Je'\|}{\|e - e'\|},$$

where we follow the convention that the identity of the norm is derived from what space the argument belongs to. Note the dependence of Lip on the norms of $\mathcal{E}$ and $\mathcal{F}$, which we suppressed. The definition implies that for any e, e', $\|Je - Je'\| \leq \mathrm{Lip}(J)\|e - e'\|$. Useful properties of Lip include that it is submultiplicative ($\mathrm{Lip}(JJ') \leq \mathrm{Lip}(J)\,\mathrm{Lip}(J')$), it is invariant to constant shifts of operators ($\mathrm{Lip}(J + e) = \mathrm{Lip}(J)$, where $J + e$ is defined by $(J + e)e' = e + Je'$) and when J is a linear operator, $\mathrm{Lip}(J) = \|J\|$, the induced operator norm of J, which is defined by

$$\|J\| \doteq \sup_{e \in \mathcal{E}, e \neq 0} \frac{\|Je\|}{\|e\|}.$$

Again, the induced norm depends on the norms that the operator acts between, but we suppress this dependence.

Let us now formally state the aforementioned assumption:

Assumption 4.3.1 (Non-expanding selectors)**.** *We have* $\mathrm{Lip}(M) \leq 1$, *and* $\mathrm{Lip}(M') \leq 1$. *For any* $\pi_1 \in \Pi$, $\pi_2 : \mathcal{I} \to \mathcal{A}$, *we have* $\mathrm{Lip}(M^{\pi_1}) \leq 1$ *and* $\mathrm{Lip}(M'^{\pi_2}) \leq 1$.

Note that this assumption constrains what norms can be selected for the spaces $\mathcal{V}^{\mathcal{A}}$, $\mathcal{V}$, $\mathcal{W}^{\mathcal{A}}$ and $\mathcal{W}$. Assumption 4.3.1 will be helpful to establish that various operators involving M are Lipschitz with a factor strictly below one, *i.e.*, that they are *contractions*. For example, to establish that $MT_{\mathcal{P}}$ is a contraction, one can use $\mathrm{Lip}(MT_{\mathcal{P}}) \leq \mathrm{Lip}(M)\,\mathrm{Lip}(T_{\mathcal{P}}) \leq \gamma\,\mathrm{Lip}(\mathcal{P}) = \gamma\|\mathcal{P}\|$, reducing the question to showing $\gamma\|\mathcal{P}\| < 1$. Similar arguments work the other operators that will involve M', M^{π}, or $M'^{\pi'}$.

As it was alluded to earlier, we will use a number of different norms. However, in all cases we choose the norm for $\mathcal{V}^{\mathcal{A}}$ ($\mathcal{W}^{\mathcal{A}}$) based on the norm of $\mathcal{V}$ (respectively, the norm of $\mathcal{W}$) to be a mixed max-norm: In particular, for $\mathcal{U}$ being either $\mathcal{V}$ or $\mathcal{W}$, the norm of $\mathcal{U}^{\mathcal{A}}$ will be defined as $\|U\|_{\mathcal{U}^{\mathcal{A}}} = \|M_{|\cdot|}U\|_{\mathcal{U}}$ where $M_{|\cdot|} : \mathcal{U}^{\mathcal{A}} \to \mathcal{U}$ is defined by $(M_{|\cdot|}U)(\cdot) = \max_a |U^a(\cdot)|$. We call the resulting norm the *mixed max-norm* w.r.t. the norm of $\mathcal{U}$.

The next proposition shows that this choice of the mixed norm makes Assumption 4.3.1 hold whenever the underlying spaces are so-called Banach lattices (Meyer-Nieber, 1991). Recall that a lattice is a non-empty set $\mathcal{U}$ with a partial ordering $\leq$ such that every pair $f, g \in \mathcal{U}$ has a supremum (or least upper bound), denoted by $f \vee g$, and an infimum (greatest lower bound), denoted by $f \wedge g$. Spaces of real-valued functions are lattices with the componentwise ordering, our default choice in what follows when it comes to $\mathcal{V}$ and $\mathcal{W}$. Operator $\vee$ is also called a *join*, a terminology we will adopt. A vector lattice $\mathcal{U}$ is a lattice that is also a vector space. In a vector lattice, for $f \in \mathcal{U}$, $f_+ = f \vee 0$, $f_- = (-f) \vee 0$ and $|f| = f_+ + f_-$ (these generalize the usual definitions of positive part, negative part and absolute value). A

Banach lattice $\mathcal{U}$ is a normed vector lattice where $\mathcal{U}$ is also a Banach space and the norm satisfies that for any $f, g \in \mathcal{V}$, $|f| \leq |g| \implies \|f\| \leq \|g\|$. Note that $(\mathcal{V}, \vee)$ is a semi-lattice (a lattice with only a join). With this we are ready to restate and prove the said statement:

Proposition 4.3.2. *Assume that $\mathcal{V}$ and $\mathcal{W}$ are Banach lattices. Then Assumption 4.3.1 is satisfied.*

Proof. See Appendix A.3, page 111. □

Let us now define the norms we will use in this paper. The weighted supremum norm of a function $f : \mathcal{Z} \to \mathbb{R}$ with respect to weight $w : \mathcal{Z} \to \mathbb{R}_+$ is defined as $\|f\|_{\infty,w} = \sup_{z \in \mathcal{Z}} |f(z)|/w(z)$. When $w = \mathbf{1}$ (*i.e.*, $w(z) = 1$ for all $z \in \mathcal{Z}$), we drop w from the index and use $\|f\|_\infty$. For $p \geq 1$, the $L^p(\mu)$-norm of f is defined as $\|f\|_{\mu,p}^p \doteq \int_{\mathcal{Z}} |f(z)|^p d\mu(z)$. By slightly abusing notation, the mixed norm of space $\mathcal{U}^{\mathcal{A}}$ derived from $\|\cdot\|_{\infty,w}$, or $\|\cdot\|_{p,\mu}$ will be denoted identically (*i.e.*, for $V \in \mathcal{V}^{\mathcal{A}}$, $\|V\|_{\infty,w}$ is a mixed norm defined using $M_{|\cdot|}$). Since these norms make their underlying spaces a Banach lattice, we immediately get the following corollary to Proposition 4.3.2:

Corollary 4.3.3. *Assume that the norms over $\mathcal{V}$ and $\mathcal{W}$ are supremum norms, weighted supremum norms, or $L^p(\mu)$ and $L^p(\rho)$ norms, and equip the spaces $\mathcal{V}^{\mathcal{A}}$ and $\mathcal{W}^{\mathcal{A}}$ with the respective mixed norms. Then Assumption 4.3.1 is satisfied.*

Proof. See Corollary 2, Ávila Pires and Szepesvári (2016b). □

Our subsequent assumptions will ensure that certain operators are contractions in appropriate norms. We start with the simplest of these assumptions:

Assumption 4.3.4. *The following holds for $\mathcal{Q}$ and $\mathcal{R}'$: $\|\mathcal{R}'\mathcal{Q}\| \leq 1$.*

Note that $\mathcal{R}'\mathcal{Q}$ is a $(\mathcal{W}, \|\cdot\|_{\mathcal{W}}) \to (\mathcal{W}^{\mathcal{A}}, \|\cdot\|_{\mathcal{W}^{\mathcal{A}}})$ operator and the norm used in Assumption 4.3.4 is the respective operator norm. As mentioned earlier, whenever Assumption 4.3.1 holds (which is the case for the norms under which we bound the policy error, see Corollary 4.3.3), we have that

$\mathrm{Lip}(M'T_{\mathcal{R}'\mathcal{Q}}) \leq \gamma\|\mathcal{R}'\mathcal{Q}\|$, and then Assumption 4.3.4 implies that $M'T_{\mathcal{R}'\mathcal{Q}}$ is a γ-contraction (again, for the respective operator norm). That $\mathcal{R}'\mathcal{Q}$ is a map between the compressed spaces $\mathcal{W}$ and $\mathcal{W}^{\mathcal{A}}$ is significant: When $\mathcal{W}$ is a finite dimensional space, Assumption 4.3.4 can be enforced during a learning procedure as done, e.g., by Yao et al. (2014). In fact, Yao et al. (2014) argue by means of some examples that enforcing this constraint as opposed to enforcing $\|\mathcal{Q}\mathcal{R}\| \leq 1$ (which may be difficult to enforce as it constrains the norm of an operator between potentially infinite dimensional spaces) can lead to better results in some learning settings.

When the norms are specifically chosen to be weighted supremum norms, the previous assumption can be replaced by a weaker one, to be stated next. To state this assumption, we need to introduce the concept of Lyapunov functions, building on a more specialized definition due to de Farias and Van Roy (2003). As de Farias and Van Roy (2003) showed by means of an example, using weighted supremum norms can greatly reduce the error bounds. Intuitively, one achieves this by assigning large weights to unimportant states, *i.e.*, to states that are infrequently visited by any policy. Indeed, one should not expect much data, or a good behavior at such states, but since they are not visited often, the errors made at such states can be safely discounted.

Given $\mathcal{Z} = (\mathcal{Z}, \|\cdot\|_{\infty,w})$, with $w : \mathcal{Z} \to \mathbb{R}_+$, and an operator $J : \mathcal{Z} \to \mathcal{Z}$, first let us define

$$\beta_{w,J} = \gamma \sup_{f:|f|=w} \|Jf\|_{\infty,w}\,.$$

Then, we say that the function w is γ-*Lyapunov* with respect to operator J if $\beta_{w,J} < 1$. We also extend the definition for operators of the form $K : \mathcal{Z} \to \mathcal{Z}^{\mathcal{A}}$, i.e., when $K = (K^a)_{a\in\mathcal{A}}$. In this case, we say that w is γ-Lyapunov w.r.t. K if it is γ-Lyapunov w.r.t. each operator K^a for any $a \in \mathcal{A}$. If J satisfies $Jf \leq J|f|$ for all $f \in \mathcal{Z}$ (*e.g.*, if J is a stochastic operator), then the definition of $\beta_{w,J}$ simplifies to $\gamma\|Jw\|_{\infty,w}$, coinciding with the definition of de Farias and Van Roy (2003).

Lyapunov functions enable us to ensure that $MT_{\mathcal{P}}$, $M^{\pi}T_{\mathcal{P}}$ ($\pi \in \Pi$) and

$M'T_{\mathcal{R}'\mathcal{Q}}$ are contractions in the corresponding weighted supremum norms. For this, notice that the following hold:

Proposition 4.3.5. *Given $(\mathcal{U}, \|\cdot\|_{\infty,\nu})$ with $\nu : \mathcal{U} \to \mathbb{R}_+$, and $J : \mathcal{U} \to \mathcal{U}^{\mathcal{A}}$, if each J^a is a linear operator, then $\gamma \operatorname{Lip}(J) = \beta_{\nu,J}$.*

Now, if ν is γ-Lyapunov w.r.t. the probability kernel $\mathcal{P}$, then we immediately get from Corollary 4.3.3 and Proposition 4.3.5 that $MT_{\mathcal{P}}$ and $M^{\pi}T_{\mathcal{P}}$ (for any $\pi \in \Pi$) are $\beta_{\nu,\mathcal{P}}$-contractions in ν-weighted supremum norm. Similarly, if η is γ-Lyapunov w.r.t. $\mathcal{R}'\mathcal{Q}$, then $M'T_{\mathcal{R}'\mathcal{Q}}$ is a $\beta_{\eta,\mathcal{R}'\mathcal{Q}}$-contraction in η-weighted supremum norm.

With this, we can state the assumption that we will use to relax Assumption 4.3.4 when the norms used the respective function spaces are weighted supremum norms. In what follows we fix two functions, $\nu : \mathcal{V} \to \mathbb{R}_+$ and $\eta : \mathcal{W} \to \mathbb{R}_+$, which will act as weighting functions.

Assumption 4.3.6 (Lyapunov weights)**.** *The following hold for $\mathcal{Q}$, $\mathcal{R}'$, ν, and η:*

(i) ν is γ-Lyapunov w.r.t. $\mathcal{P}$;

(ii) η is γ-Lyapunov w.r.t. $\mathcal{R}'\mathcal{Q}$.

Note that choosing the weight function ν to be the constant one function, Assumption 4.3.6(i) is automatically satisfied, while choosing η to be the constant one function, Assumption 4.3.6(ii) is equivalent to Assumption 4.3.4 when the norm used there is the supremum norm.

Some (but not all) of our bounds will have a dependency on $\operatorname{Lip}(T_{\mathcal{Q}}) = \gamma\|\mathcal{Q}\|$. Therefore, we will also make Assumption 4.3.7 to avoid vacuous bounds.

Assumption 4.3.7. *We have that $B \doteq \|\mathcal{Q}\| < \infty$.*

Note that this assumption is mild: Learning procedures would more often than not guarantee finiteness of the objects they return. In fact, by appropriate normalization, even $\|Q\| \leq 1$ can be arranged (if necessary), as done for example by Grünewälder et al. (2012).

4.4 Results

In this section we present our main results. We start with a viability result (explaining why our minimal assumptions are sufficient for the existence of the policy whose performance we are interested in), followed by a short review of previous bounds on the policy error. These previous bounds provide the context for our new results, which we present afterwards. After each result we discuss their relative merits and present their proofs. Differently from Ávila Pires and Szepesvári (2016b), we do not not assume that $\mathcal{R}$ is a join-homomorphism, or that $\mathcal{R}'^a = \mathcal{R}$ for each $a \in \mathcal{A}$.

4.4.1 A Viability Result

Theorem 4.4.1 formalizes that u^* is well-defined (the MBRL approach with factored semi-linear models is viable) under Assumption 4.3.4 or Assumption 4.3.6 (ii), provided that the norm over $\mathcal{W}^{\mathcal{A}}$ is a mixed max-norm w.r.t. the norm over $\mathcal{W}$. Theorem 4.4.1 shows that $M'T_{\mathcal{R}'\mathcal{Q}}$ is a contraction (in $\|\cdot\|_{\mathcal{W}}$) and we can compute u^* by value iteration. Therefore, as remarked in Section 4.1, if the compressed space $\mathcal{W}$ is finite dimensional, we are able to evaluate $M'T_{\mathcal{R}'\mathcal{Q}}$ and thus also approximate u^* efficiently (up to the desired accuracy). Evaluating $\hat{\pi}(x)$ can be done by computing $(T_{\mathcal{Q}}u^*)(x)$ for each x as needed. As we have relinquished the join-homomorphism assumption, we can no longer say that $MT_{\mathcal{Q}}$ has a fixed point (which holds if $\mathcal{R}$ is a join-homomorphism and $\mathcal{R}'^a = \mathcal{R}$ for each $a \in \mathcal{A}$, cf. Theorem 4.4.1 Ávila Pires and Szepesvári, 2016b, Theorem 5).

The fixed point u^*, the contraction $M'T_{\mathcal{Q}\mathcal{R}}$, and the quantity[2] $U^* \doteq MT_{\mathcal{Q}}u^*$, will play pivotal roles in our bounds.

Theorem 4.4.1. *Assume that the norm over $\mathcal{W}^{\mathcal{A}}$ is the mixed max-norm w.r.t. the norm over $\mathcal{W}$, and let Assumption 4.3.4 or Assumption 4.3.6 (ii) hold. Then $M'T_{\mathcal{R}'\mathcal{Q}}$ is a contraction w.r.t. the norm underlying $\mathcal{W}$, $M'T_{\mathcal{R}'\mathcal{Q}}$ has a unique*

[2] If we define $\mathcal{R}$ to satisfy $\mathcal{R}U^* = u^*$, we get back that U^* is a fixed point of $MT_{\mathcal{Q}\mathcal{R}}$, although that does not seem to have any practical significance in our bounds.

fixed point u^*, and the iteration $u_{k+1} = M' T_{\mathcal{R}'\mathcal{Q}} u_k$ converges geometrically to u^*, for any $u_0 \in \mathcal{W}$.

Proof. See the first part of the proof of Theorem 5 of Ávila Pires and Szepesvári (2016b). In essence, this result is a simple application of Banach's fixed point theorem. □

As pointed out in Section 4.1, to the best of our knowledge, all previous works either assumed or imposed a contraction property on $MT_{\mathcal{QR}}$. In fact, with the exception of Yao et al. (2014), all previous works required $\mathcal{QR}$ to be stochastic, so Theorem 4.4.1 is a notable relaxation of viability requirements.

4.4.2 Previous Results on the Policy Error

The typical MBRL performance bound is a supremum-norm bound on the policy error of $\tilde{\pi} \doteq GT_{\widetilde{\mathcal{P}}}\widetilde{V}$, where $\widetilde{\mathcal{P}}$ is stochastic and $\widetilde{V}$ is the fixed point of $MT_{\widetilde{\mathcal{P}}}$.

Theorem 4.4.2 (Baseline bound on MBRL policy error)**.** *Consider some transition probability kernel $\widetilde{\mathcal{P}}$ for the state and action spaces $\mathcal{X}$ and $\mathcal{A}$. Let $\widetilde{V}$ be the fixed point of $MT_{\widetilde{\mathcal{P}}}$, and $\tilde{\pi} = GT_{\widetilde{\mathcal{P}}}\widetilde{V}$. Then*

$$\left\|V^* - V^{\tilde{\pi}}\right\|_\infty \leq \frac{2\gamma}{1-\gamma}\left\|(\mathcal{P} - \widetilde{\mathcal{P}})\widetilde{V}\right\|_\infty.$$

This result is essentially contained in the works of Whitt (1978, Corollary to Theorem 3.1), Singh and Yee (1994, Corollary 2)[3], Bertsekas (2012, Proposition 3.1), and Grünewälder et al. (2011, Lemma 1.1).

An important implication of this result, which we feel is often overlooked, is that the approximation $\widetilde{\mathcal{P}}$ to $\mathcal{P}$ does not have to be precise everywhere (at all functions $V \in \mathcal{V}$), but only at $\widetilde{V}$, the fixed point of the approximate model—a self-fulfilling prophecy, prone to failure? To understand why this works, consider the case when $\widetilde{\mathcal{P}}\widetilde{V}$ perfectly matches $\mathcal{P}\widetilde{V}$, *i.e.*, when the bound on the right-hand side is zero. In this case $\widetilde{V} = MT_{\widetilde{\mathcal{P}}}\widetilde{V} = MT_{\mathcal{P}}\widetilde{V}$,

[3] Singh and Yee (1994) correctly bound $\left\|V^* - V^{\tilde{\pi}}\right\|_\infty$, but their statement of Corollary 2 suggests that they are bounding a different quantity.

which implies that $\widetilde{V} = V^*$ and, $\tilde{\pi} = GT_{\widetilde{\mathcal{P}}}\widetilde{V} = GT_{\mathcal{P}}V^*$ is optimal. *The moral is that models do not have to be precise everywhere; if $\mathcal{P}\tilde{V}$ can be estimated, the above inequality can be used to derive a posteriori bounds on the policy error and even form the basis of improving the model.* This can be viewed as a major, unexpected win for model-based RL.

Ormoneit and Sen (2002); Barreto et al. (2011); Barreto and Fragoso (2011); Precup et al. (2012); Barreto et al. (2014b,a) bound $\|V^* - \widetilde{V}\|_\infty$ rather than the policy error. We emphasize (see Ávila Pires and Szepesvári, 2016b, Appendix D) that $\|V^* - \widetilde{V}\|_\infty$ is not the correct quantity to bound in order to understand the quality of $\hat{\pi}$, and that the policy error should be bounded. As remarked by Ávila Pires and Szepesvári (2016b, Appendix D), in ADP it is sufficient to bound the deviation between the optimal value function and the value estimate that generates the policy, in order to understand the policy error in supremum norm.

4.4.3 Bounds on the Policy Error in Factored Semi-linear Models

Our first novel result is a supremum-norm bound for policy error when we use factored semi-linear models: Theorem 4.4.3. Because we can recover results for unfactored linear models by taking $\mathcal{R}$ and $\mathcal{R}'$ to be the identity mapping over $\mathcal{X}$, we can use Theorem 4.4.3 to get a bound that is tighter than Theorem 4.4.2. Strictly speaking, taking $\mathcal{Q}$ stochastic, $\mathcal{R}$ and each $\mathcal{R}'^a$ for $a \in \mathcal{A}$ as the identity mapping, and upper-bounding the right-hand side of Theorem 4.4.3 by $2\varepsilon_2$ gives us Theorem 4.4.2. Ávila Pires and Szepesvári (2016b, Proposition 16) show that Theorem 4.4.3 if $\mathcal{R}$ is a join-homomorphism and $\mathcal{R}'^a = \mathcal{R}$ for each $a \in \mathcal{A}$. is tight.

Theorem 4.4.3 (Supremum-norm bound with linear $\mathcal{R}'$)**.** *Given linear $\mathcal{R}'$, let $\hat{\pi}$ be the policy derived from the factored semi-linear model defined using* (4.2.1) *and* (4.2.2)*. If Assumptions 4.3.4 and 4.3.7 hold, then*

$$\left\| V^* - V^{\hat{\pi}} \right\|_\infty \leq \varepsilon(V^*) + \varepsilon(V^{\hat{\pi}}),$$

where

$$
\begin{aligned}
\varepsilon(V) &= \min(\varepsilon_1(V), \varepsilon_2)\ , \\
\varepsilon_1(V) &= \inf_{\mathcal{R}:\mathcal{V}\to\mathcal{W}} \Big(\gamma \|\mathcal{P}V - \mathcal{Q}\mathcal{R}V\|_\infty + \frac{B\gamma^2}{1-\gamma} \big\|\mathcal{R}'(\mathcal{P}V - \mathcal{Q}\mathcal{R}V)\big\|_\infty \\
&\qquad\qquad + \varepsilon_3(V, \mathcal{R}) \Big)\ , \\
\varepsilon_2 &= \frac{\gamma}{1-\gamma} \big\|\mathcal{P}U^* - \mathcal{Q}u^*\big\|_\infty\ , \\
\varepsilon_3(V^*, \mathcal{R}) &= \frac{B\gamma}{1-\gamma} \big\|\mathcal{R}MT_{\mathcal{P}}V^* - M'\mathcal{R}'T_{\mathcal{P}}V^*\big\|_\infty\ , \\
\varepsilon_3(V^{\hat{\pi}}, \mathcal{R}) &= \frac{B\gamma}{1-\gamma} \big\|\mathcal{R}M^{\hat{\pi}}T_{\mathcal{P}}V^{\hat{\pi}} - M'^{\hat{\pi}'}\mathcal{R}'T_{\mathcal{P}}V^{\hat{\pi}}\big\|_\infty\ .
\end{aligned}
$$

Proof. See Appendix A.3, page 114. □

From Theorem 4.4.3, we see that it is enough if the model is good in the sense of minimizing ε_2, which depends on how the model interacts u^*. If $\mathcal{Q}u^*$ is close $\mathcal{P}U^* = \mathcal{P}MT_{\mathcal{Q}}u^*$ our model will be good. We can say that the term with ε_2 may lead to *a posteriori* bounds, while the ε_1 terms are better treated as *a priori* bounds, due to the presence of V^* and $V^{\hat{\pi}}$, objects in the true MDP.

A striking feature of Theorem 4.4.3 is the $\varepsilon_1(V)$ term. It means that if B is not too big, and if the error of the model at V^* and $V^{\hat{\pi}}$ *in the compressed space* $\mathcal{W}^{\mathcal{A}}$ and the term $\varepsilon_3(V, \mathcal{R})$ are small, then the term that depends on $\frac{1}{1-\gamma}$ is small. Moreover, we can expect the error in the compressed space to be easier to control than $\|\mathcal{P}V - \mathcal{Q}\mathcal{R}V\|_\infty$, depending on the choice of $\mathcal{R}$.

The term $\varepsilon_3(V, \mathcal{R})$ (cf. Theorem 4.4.3 and Theorem 8 of Ávila Pires and Szepesvári, 2016b) is the price that we pay for not being able to use the identity $\mathcal{R}M = M'\mathcal{R}'$. If the identity does hold, and if we define linear $\mathcal{R}$ so that $\mathcal{R}U^* = u^*$, we immediately Theorem 8.

Perhaps more interestingly, one will notice that $\mathcal{R}$ is only used in the bound, since $\hat{\pi} = GT_{\mathcal{Q}}u^*$ and $u^* = M'T_{\mathcal{R}'\mathcal{Q}}u^*$, so we can make an oracle choice of $\mathcal{R}$ that minimizes $\varepsilon_1(V)$ Moreover, a different choice of $\mathcal{R}$ can be made for each individual bound, so different choices can be made for $\varepsilon_1(V^{\hat{\pi}})$

and $\varepsilon_1(V^*)$. In particular, because $\mathcal{R}$ may be non-linear, we may choose it so that $\mathcal{R}MT_{\mathcal{P}}V^* = M'\mathcal{R}'T_{\mathcal{P}}V^*$, which gives us (using that $V^* = MT_{\mathcal{P}}V^*$)

$$\varepsilon_1(V^*) \leq \gamma \left\| \mathcal{P}V^* - \mathcal{Q}M'\mathcal{R}'T_{\mathcal{P}}V^* \right\|_\infty + \frac{B\gamma^2}{1-\gamma} \left\| \mathcal{R}'(\mathcal{P}V^* - \mathcal{Q}M'\mathcal{R}'T_{\mathcal{P}}V^*) \right\|_\infty. \tag{4.4.1}$$

We are also able to compare Theorem 4.4.3 with Proposition 1 of Barreto et al. (2014a), who study a factored linear model approach with stochastic $\mathcal{R}'$, finite-dimensional $\mathcal{V}$ and $\mathcal{W}$, and $\left\|V^*\right\|_\infty \leq \frac{C}{1-\gamma}$ for some C. In their setting, the reward vector is not known, so $\mathcal{R}'r$ is replaced by an $r' \in \mathcal{W}^{\mathcal{A}}$. Proposition 1 of Barreto et al. (2014a) gives us

$$\begin{aligned} \left\| V^* - M\mathcal{Q}M'(r' + \mathcal{Q}u^*) \right\|_\infty \leq \frac{1}{1-\gamma} \left\| r - \mathcal{Q}r' \right\|_\infty \\ + \frac{\gamma C}{(1-\gamma)^2} \left\| \mathcal{P} - \mathcal{Q}\mathcal{R}' \right\|_\infty \\ + \frac{C}{(1-\gamma)^2} \max_{a,i} \min_j (1 - \mathcal{Q}^a_{i,j}). \end{aligned} \tag{4.4.2}$$

We can eliminate the first term in (4.4.2) by using an uncompressed guess $r'' \in \mathcal{V}^{\mathcal{A}}$ for r and bounding $\left\| V^* - M(r'' + \mathcal{Q}u^*) \right\|_\infty$. Moreover, the factor of $\frac{C}{1-\gamma}$ can be avoided (see the discussion of Theorem 4.4.2 in Section 4.4). If $\mathcal{R}'$ is stochastic and $B \geq 1$, in Theorem 4.4.3 we can upper-bound

$$\varepsilon_1(V^*) \leq \inf_{\mathcal{R}:\mathcal{V}\to\mathcal{W}} \left(\frac{B\gamma}{1-\gamma} \|\mathcal{P}V^* - \mathcal{Q}\mathcal{R}V^*\|_\infty + \varepsilon_3(V^*, \mathcal{R}) \right) \tag{4.4.3}$$

since $\mathrm{Lip}(\mathcal{R}') \leq 1$. We may choose $\mathcal{R}$ in (4.4.3) so that, for example, $\varepsilon_3(V^*, \mathcal{R}) = 0$, although this choice may adversely affect the first term in the right-hand side of (4.4.3). In contrast, the term in (4.4.2) greatly restricts the choices of $\mathcal{Q}$. Having $\max_{a,i} \min_j (1 - \mathcal{Q}^a_{i,j}) = 0$ in the setting of Barreto et al. (2014a) but being able to chose $\mathcal{R}'$ to be stochastic is comparable to, in our setting, allowing $\mathcal{Q}$ to be stochastic, requiring $\mathcal{R}$ to be a point selector, and having $(\mathcal{R}')^a = \mathcal{R}$ for all $a \in \mathcal{A}$.

The proof of Theorem 4.4.3 uses the triangle inequality

$$\|V^* - V^{\hat{\pi}}\| \leq \|V^* - U^*\| + \|U^* - V^{\hat{\pi}}\|, \tag{4.4.4}$$

combined with Lemma 4.4.4 stated next. Though technical, Lemma 4.4.4 is the cornerstone of our policy error bounds, just as Lemma 9 of Ávila Pires

and Szepesvári (2016b) is the cornerstone of theirs. Indeed, once we generalize Lemma 9 to factored semi-linear models (thus obtaining Lemma 4.4.4), the policy error bounds in this text follow with little extra effort: We simply need to redo the proofs of each policy error bound by Ávila Pires and Szepesvári (2016b)) with Lemma 4.4.4 instead of Lemma 9, which is a straightforward task.

Lemma 4.4.4. *Given linear $\mathcal{R}'$, assume that $\gamma \operatorname{Lip}(\mathcal{R}'\mathcal{Q}) \leq \alpha < 1$ and that Assumptions 4.3.1 and 4.3.7 hold. The following holds with $(V, N, N', N'') = (V^*, M, M', I)$ and with $(V, N, N', N'') = \left(V^{\hat{\pi}}, M^{\hat{\pi}}, M'^{\hat{\pi}'}, M^{\hat{\pi}}\right)$:*

$$\|V - U^*\| \leq \gamma \|\mathcal{P}V - \mathcal{Q}u^*\|, \tag{4.4.5}$$

and for any $\mathcal{R} : \mathcal{V} \to \mathcal{W}$ (that may depend on V):

$$\begin{aligned} \|V - U^*\| \leq \gamma \|\mathcal{P}V - \mathcal{Q}\mathcal{R}V\| + \frac{B\gamma^2}{1-\alpha} \left\|\mathcal{R}'(\mathcal{P}V - \mathcal{Q}\mathcal{R}V)\right\| \\ + \frac{B\gamma}{1-\alpha} \left\|\mathcal{R}NT_{\mathcal{P}}V - N'\mathcal{R}'T_{\mathcal{P}}V\right\| \end{aligned} \tag{4.4.6}$$

Additionally, if $\gamma \operatorname{Lip}(N''\mathcal{P}) \leq \beta < 1$, we also have that

$$\|V - U^*\| \leq \frac{\gamma}{1-\beta} \|\mathcal{P}U^* - \mathcal{Q}u^*\|. \tag{4.4.7}$$

Proof. See Appendix A.3, page 111. □

Compared to Lemma 9, we also added an extra inequality to Lemma 4.4.4: (4.4.5). This inequality is a simple observation that may be useful as a criterion to minimize—we see that the ADMM objective used by Yao et al. (2014) is related to minimizing the right-hand side of (4.4.5) subject to ensuring that Assumption 4.3.4 is satisfied.

Lemma 4.4.4 (4.4.6) can be interpreted as the bound we get by doing a Bellman lookahead with $MT_{\mathcal{Q}}$, followed by application of the well-known bound for an α-contraction T with fixed point $\widetilde{V}$ (Bertsekas, 2007):

$$\left\|V - \widetilde{V}\right\| \leq \frac{1}{1-\alpha} \|V - TV\| \tag{4.4.8}$$

(with $T = M'T_{\mathcal{R}'\mathcal{Q}}$ in the case of Lemma 4.4.4). Similarly, taking $T = MT_{\mathcal{P}}$ ($T = M^{\hat{\pi}}T_{\mathcal{P}}$) in (4.4.8) combined with $\gamma \operatorname{Lip}(\mathcal{P}) \leq \beta < 1$ ($\gamma \operatorname{Lip}(M^{\hat{\pi}}\mathcal{P}) \leq$

$\beta < 1$), allows us to see that $MT_{\mathcal{P}}$ ($M^{\hat{\pi}}T_{\mathcal{P}}$) is a β-contraction, so (4.4.8) gives us Lemma 4.4.4 (4.4.7) for V^* ($V^{\hat{\pi}}$). Lemma 4.4.4 (4.4.6) is also interesting in the special case of unfactored linear models (when $\mathcal{R}$ is the identity mapping) with $\mathcal{Q}$ as a non-expansion (*e.g.*, $\mathcal{Q}$ stochastic): Because $B = 1$ and $\alpha = \gamma$, the bound becomes

$$\|V - U^*\| \leq \frac{\gamma}{1-\gamma}\|(\mathcal{P} - \mathcal{Q}\mathcal{R})V\|,$$

and in this case no looseness was introduced by doing a Bellman lookahead and then applying (4.4.8), relative to applying (4.4.8) directly. This will allow us to recover results for unfactored linear models from the bounds we derive from Lemma 4.4.4.

Many of the remarks about Theorem 4.4.3 pertaining to $\mathcal{R}$ and $\varepsilon_3(V, \mathcal{R})$ are in fact a consequence of the bounds in Lemma 4.4.4. As this is our cornerstone lemma, we will be able to make similar remarks for the other bounds as well. For example, we can see that $\varepsilon_3(V, \mathcal{R})$ from Theorem 4.4.3 appears in (4.4.6) as the price we pay for not having the identity $\mathcal{R}M = M'\mathcal{R}'$. Moreover, in Lemma 4.4.4 we are able to make oracle choices of $\mathcal{R}$ (different choices for V^* and $V^{\hat{\pi}}$), and we can generalize (4.4.1) by choosing $\mathcal{R}$ so that it gives $\varepsilon_3(V, \mathcal{R}) = 0$. Turning back to Theorem 4.4.3, we can use it to crudely upper-bound the policy error in $L^p(\mu)$ norm, but the bound we obtain this way is not very interesting. This is because supremum norm bounds, though easy to prove, can be too harsh: V^* and $V^{\hat{\pi}}$ can be close in other meaningful norms, while not being close in supremum norm, in which case the right-hand side of the bound in Theorem 4.4.3 can be large even if the left-hand side is small (see Ávila Pires and Szepesvári, 2016b, Proposition 17).

De Farias and Van Roy (2003) show that the harshness of the supremum norm can be mitigated by considering the policy error in weighted supremum norm. Intuitively, the error in states that are unlikely to be visited by π^* should be underweighted, as we discussed earlier. Thus, one alternative to supremum norm bounds is to use a generalization of Theorem 4.4.3 for

the weighted supremum norm:

Theorem 4.4.5 (Weighted supremum norm bound with linear $\mathcal{R}'$)**.** *Given linear $\mathcal{R}'$, let $\hat{\pi}$ be the policy derived from the factored linear model defined using* (4.2.1) *and* (4.2.2)*. If Assumptions 4.3.6 and 4.3.7 hold, then*

$$\left\| V^* - V^{\hat{\pi}} \right\|_{\infty,\nu} \leq \varepsilon(V^*) + \varepsilon(V^{\hat{\pi}}),$$

where

$$\begin{aligned}
&\varepsilon(V) = \min(\varepsilon_1(V), \varepsilon_2)\ , \\
&\varepsilon_1(V) = \inf_{\mathcal{R}:\mathcal{V}\to\mathcal{W}} \Bigg(\gamma \|\mathcal{P}V - \mathcal{Q}\mathcal{R}V\|_{\infty,\nu} + \frac{B\gamma^2}{1-\beta_{\eta,\mathcal{R}'\mathcal{Q}}} \left\|\mathcal{R}'(\mathcal{P}V - \mathcal{Q}\mathcal{R}V)\right\|_{\infty,\eta} \\
&\qquad\qquad + \varepsilon_3(V,\mathcal{R}) \Bigg)\ , \\
&\varepsilon_2 = \frac{\gamma}{1-\beta_{\nu,\mathcal{P}}} \left\|\mathcal{P}U^* - \mathcal{Q}u^*\right\|_{\infty,\nu}\ , \\
&\varepsilon_3(V^*,\mathcal{R}) = \frac{B\gamma}{1-\beta_{\eta,\mathcal{R}'\mathcal{Q}}} \left\|\mathcal{R}MT_{\mathcal{P}}V^* - M'\mathcal{R}'T_{\mathcal{P}}V^*\right\|_{\infty,\eta}\ , \\
&\varepsilon_3(V^{\hat{\pi}},\mathcal{R}) = \frac{B\gamma}{1-\beta_{\eta,\mathcal{R}'\mathcal{Q}}} \left\|\mathcal{R}M^{\hat{\pi}}T_{\mathcal{P}}V^{\hat{\pi}} - M'^{\hat{\pi}'}\mathcal{R}'T_{\mathcal{P}}V^{\hat{\pi}}\right\|_{\infty,\eta}\ .
\end{aligned}$$

Proof. See Appendix A.3, page 113. □

Under Assumption 4.3.4 and Assumption 4.3.6 (i), Theorem 4.4.5 holds with $\beta_{\eta,\mathcal{R}'\mathcal{Q}} = \gamma$. The comments about $\varepsilon_1(V)$, ε_2 and $\varepsilon_3(V,\mathcal{R})$ in Theorems 4.4.2 and 4.4.3 are also valid for Theorem 4.4.5, but the dependencies are, evidently, expressed in different norms. Moreover, by taking $\nu = x \mapsto 1$ and $\eta = i \mapsto 1$, and by realizing that ν is γ-Lyapunov w.r.t. $\mathcal{P}$ and, under Assumption 4.3.4, η is γ-Lyapunov w.r.t. $\mathcal{R}'\mathcal{Q}$, we recover Theorem 4.4.3 from Theorem 4.4.5. Previously, weighted-supremum norm bounds were derived for ALP. However, the weakness of these bounds is that they are sensitive to the measure-change between the "ideal constraint sampling distribution" (which depends on unknown quantities whose knowledge basically implies the knowledge of the optimal policy) and the actual one used in the algorithm (de Farias and Van Roy, 2003).

Normally, we are interested in the policy error w.r.t. an initial state distribution, or a stationary distribution of a policy (*e.g.*, a stationary distribution of π^*), and we can naturally consider the policy error in $L^1(\mu)$ norm, where μ is a measure over $\mathcal{X}$ that we are interested in. We can get an immediate bound for the more general $L^p(\mu)$ norm (for any $p \geq 1$) of the policy error, using Theorem 4.4.5.

Theorem 4.4.6 (Weighted supremum norm bound for the policy error in $L^p(\mu)$ norm, with linear $\mathcal{R}'$)**.** *Given linear $\mathcal{R}'$, let $\hat{\pi}$ be the policy derived from the factored linear model defined using* (4.2.1) *and* (4.2.2)*. If Assumptions 4.3.6 and 4.3.7 holds for the weighted supremum norm over $\mathcal{V}^{\mathcal{A}}$ and $\mathcal{W}^{\mathcal{A}}$, then*

$$\left\| V^* - V^{\hat{\pi}} \right\|_{\mu,p} \leq \|\nu\|_{\mu,p} \Big(\varepsilon(V^*) + \varepsilon(V^{\hat{\pi}}) \Big),$$

where ε, ε_1, ε_2 and ε_3 are as in Theorem 4.4.5.

Proof. See Appendix A.3, page 114. □

However, we can also bound the policy error in $L^p(\mu)$ "directly", *i.e.*, in terms of model errors in $L^p(\mu)$ norm, as Theorem 4.4.7, to be stated next, shows. In order to state Theorem 4.4.7, we need to use a *concentrability coefficient* $C_{\hat{\pi},\mathcal{P},\mu,\xi}$ (although part of our bound will be free of this coefficient). Consider a measure ξ over $\mathcal{X}$, and the operator $I - \gamma M^{\hat{\pi}} \mathcal{P} : (\mathcal{V}, \|\cdot\|_{\xi,p}) \to (\mathcal{V}, \|\cdot\|_{\mu,p})$. If $I - \gamma M^{\hat{\pi}} \mathcal{P}$ has no inverse (as an operator acting between the above two spaces), define $C_{\gamma,\hat{\pi},\mathcal{P},\mu,\xi} \doteq \infty$, otherwise let the concentrability coefficient be

$$C_{\gamma,\hat{\pi},\mathcal{P},\mu,\xi} \doteq (1-\gamma) \operatorname{Lip}((I - \gamma M^{\hat{\pi}} \mathcal{P})^{-1}) = (1-\gamma) \left\| (I - \gamma M^{\hat{\pi}} \mathcal{P})^{-1} \right\|. \tag{4.4.9}$$

(Note that here both $\operatorname{Lip}(\cdot)$ and $\|\cdot\|$ hide a dependence on ξ, π and p.) As opposed to previous uses of concentrability coefficients (Munos, 2003; Farahmand et al., 2010), our coefficient depends only on the policy computed, which makes it more suitable for the estimation of our bound. In case the $C_{\gamma,\hat{\pi},\mathcal{P},\mu,\xi}$ is not very large, we can get meaningful bounds from

Theorem 4.4.7 from ε_2, but even if $C_{\gamma,\hat{\pi},\mathcal{P},\mu,\xi} = \infty$ and ε_2 is vacuous, we can still get a priori bounds with a dependence on $\varepsilon_1(V^{\hat{\pi}})$, in addition to the dependence on $\varepsilon_1(V^*)$. The $\varepsilon_1(V)$ term can be analyzed similarly to its analogues in Theorems 4.4.3 and 4.4.5, modulo the norm differences. We are flexible about the choice of $\|\cdot\|_{\mathcal{W}}$ (which nonetheless affects Assumptions 4.3.4 and 4.3.7). One may think of choosing $\|\cdot\|_{\mathcal{W}} = \|\cdot\|_{\rho,p}$ for some ρ, however with this norm choice, Assumption 4.3.4 becomes restrictive. When it comes to satisfying Assumption 4.3.4, a weighted supremum norm is reasonable, as discussed earlier, so we choose this norm as the norm over the compressed space $\mathcal{W}$ in Theorem 4.4.7. We emphasize that $\varepsilon_1(V)$ is independent of the concentrability coefficient. Further, as remarked beforehand, its dependence on the discount factor can be quite mild (if the second term in the definition of ε_1 is small).

Theorem 4.4.7 ($L^p(\mu)$ norm bound with linear $\mathcal{R}'$)**.** *Given linear $\mathcal{R}'$, let $\hat{\pi}$ be the policy derived from the factored linear model defined using* (4.2.1) *and* (4.2.2)*. Choose the norms so that $\|\cdot\|_{\mathcal{V}} = \|\cdot\|_{\mu,p}$ and $\|\cdot\|_{\mathcal{W}} = \|\cdot\|_{\infty,\eta}$. If Assumptions 4.3.4 and 4.3.7 hold, then*

$$\left\|V^* - V^{\hat{\pi}}\right\|_{\mu,p} \leq \varepsilon_1(V^*) + \min\Big(\varepsilon_1(V^{\hat{\pi}}), \varepsilon_2\Big),$$

where

$$\begin{aligned}
\varepsilon_1(V) &= \inf_{\mathcal{R}:\mathcal{V}\to\mathcal{W}} \Bigg(\gamma\|\mathcal{P}V - \mathcal{Q}\mathcal{R}V\|_{\mu,p} + \frac{B\gamma^2}{1-\gamma}\big\|\mathcal{R}'(\mathcal{P}V - \mathcal{Q}\mathcal{R}V)\big\|_{\infty,\eta} \\
&\qquad\qquad + \varepsilon_3(V,\mathcal{R})\Bigg)\,, \\
\varepsilon_2 &= C_{\gamma,\hat{\pi},\mathcal{P},\mu,\xi}\,\frac{\gamma}{1-\gamma}\|\mathcal{P}U^* - \mathcal{Q}u^*\|_{\xi,p}\,, \\
\varepsilon_3(V^*,\mathcal{R}) &= \frac{B\gamma}{1-\gamma}\big\|\mathcal{R}M T_{\mathcal{P}} V^* - M'\mathcal{R}' T_{\mathcal{P}} V^*\big\|_{\infty,\eta}\,, \\
\varepsilon_3(V^{\hat{\pi}},\mathcal{R}) &= \frac{B\gamma}{1-\gamma}\big\|\mathcal{R}M^{\hat{\pi}} T_{\mathcal{P}} V^{\hat{\pi}} - M'^{\hat{\pi}'}\mathcal{R}' T_{\mathcal{P}} V^{\hat{\pi}}\big\|_{\infty,\eta}\,.
\end{aligned}$$

where $C_{\gamma,\hat{\pi},\mathcal{P},\mu,\xi}$ is defined in (4.4.9)*.*

Proof. See Appendix A.3, page 113. □

4.5 Conclusion

Our results in this chapter generalize the results of Ávila Pires and Szepesvári (2016b), from which this chapter is largely derived. The results of Ávila Pires and Szepesvári (2016b), in turn, unify, strengthen and extend previous works The unifying framework of factored linear models was introduced by Yao et al. (2014), and in this chapter we have mildly generalized it to a factored semi-linear model framework. This generalization is not significant in terms of coverage, as factored linear models already cover all previous work that we are aware of and that is also covered by factored semi-linear models. Nevertheless, having $\mathcal{R}$ non-linear adds flexibility to the framework.

Our focus has been the derivation of policy error bounds, and we have put aside issues of designing and analyzing algorithms to learn models. We believe that the when developing theories for reinforcement learning one should start by figuring out what quantities control the policy error of a given methods. Then, one is in a better position to design learning algorithms which then control the said quantities (this is distantly reminiscent to choosing surrogate losses in supervised learning).

Previous work that derives policy error bounds goes back to at least Whitt (1978). In fact, looking at the literature we see that the results of Whitt (1978) have been independently re-derived in part or as a whole multiple times (often confounded with the issue of statistical questions), *e.g.*, in the works mentioned in Section 4.2. Compared to the work of Whitt (1978), main advances in deriving policy error bounds have been the introduction of norms other than the supremum norm, though this happened in different contexts (*e.g.* de Farias and Van Roy, 2003; Munos, 2003), and breaking down the bound of Whitt (1978) to more specialized models (*e.g.* Ormoneit and Sen, 2002; Ravindran, 2004; Barreto et al., 2011; Sorg and Singh, 2009).

One of the main novelties of Ávila Pires and Szepesvári (2016b) presented in this chapter is that previous techniques are imported to model-based RL to obtain policy error bounds in norms other than (unweighted) supre-

mum norms. In particular, to derive policy error bounds that use weighted supremum norms, we are building on the work of de Farias and Van Roy (2003), and we bring Lyapunov analysis from the approximate linear programming (ALP) methodology to model-based RL. At the same time, to derive policy error bounds that use weighted L^p-norms we import ideas from Munos (2003), who analyzed approximate dynamic programming (ADP) algorithms. During this process we streamlined the definitions from these works by sticking to the language of operator algebras (specifically, Banach lattices). The use of this language has two main benefits: It allowed us to present shorter and rather direct proofs, while it also shed light on the algebraic and geometric assumptions that were key in the proofs. We believe that our operator algebra approach could also improve previous results in either ALP or ADP. An interesting avenue for further work is to investigate the minimum set of assumptions under which our calculations remain valid: At present it appears that we use very little of the rich structure of the function spaces involved. We speculate that the results can also be proven in certain max-plus (a.k.a. tropical) algebras, leading to results that may hold, *e.g.*, for various versions of sequential games.

Another major novel aspect of the present work is that we tightened previous bounds. In particular, our bounds come in two forms: One form (the "ε_1" term) tells us how model errors should be controlled in the compressed space (with the compression depending on the choice of $\mathcal{R}$) and traded off with the error incurred by us not being able to use the identity $\mathcal{R}M = M'\mathcal{R}'$. The other form (the "$\varepsilon_2$" term) tells us that it is enough if the compressed fixed point applied to the model operator ($\mathcal{Q}u^*$) approximates the true model operator at U^* ($\mathcal{P}U^*$), with $U^* = MT_{\mathcal{Q}}u^*$ being a quantity derived from the model.

While we shorten and improve previous results, we also managed to relax the key condition of previous works that required that the Bellman operator acting on uncompressed value functions and underlying the model needs to be a contraction. While we are still relying on contraction-type

arguments, the contraction arguments are used with the compressed space, as previously suggested (but not analyzed) by Yao et al. (2014). We feel that it is more natural to require that the Bellman operator for the compressed space be a contraction than to require the same for the respective operator acting on the uncompressed space. Indeed, our bounds show that this second assumption is entirely superfluous (cf. the "ε_2" terms).

In this chapter, we have also relaxed the assumption used by Ávila Pires and Szepesvári (2016b) that $\mathcal{R}$ is a join-homomorphism and that $\mathcal{R}'^a = \mathcal{R}$ for each $a \in \mathcal{A}$. This allows us to extend the scope of our results to MBRL methods that use factored linear models, but where $\mathcal{R}$ and $\mathcal{R}'$ can be any linear operator (and different from each other). This family of methods includes, for example, state-aggregation (soft or not) or stochastic factorization Van Roy (2006); Barreto et al. (2011), where $\mathcal{R}$ and $\mathcal{R}'$ are linear (and stochastic) but *not* join-homomorphisms.

While Ávila Pires and Szepesvári (2016b) sketched a generalization of their results to linear $\mathcal{R}$, we have realized that $\mathcal{R}$ need not be linear or known (so oracle choices can be made to optimize the bound). This relaxation does not affect the error term additional error terms introduced in Theorem 12 of Ávila Pires and Szepesvári (2016b) (cf. Theorem 4.4.7).

Our results (as a consequence of the cornerstone result, Lemma 4.4.4, and similarly to Lemma 9 and the other results of Ávila Pires and Szepesvári, 2016b) show a curious scaling as a function of $1/(1-\gamma)$. In fact, the astute reader may recall that policy error bounds typically scale with $1/(1-\gamma)^2$. A little thinking reveals that our result may be subject to the same scaling: Just like in Theorem 4.4.2, where $\widetilde{V}$ hides $1/(1-\gamma)$, in the above bounds the value functions themselves bring in another $1/(1-\gamma)$, too. Is the scaling with $1/(1-\gamma)^2$ necessary? The answer is no: Theorem 4.1 of Van Roy (2006) shows that in some version of state-aggregation the policy error can scale with $1/(1-\gamma)$ only (as a side-note, the only result so far with this property). Thus, it may be worthwhile to look at the differences between Theorem 4.1 and the above result. First, recall that in his Theorem 4.1 Van Roy (2006)

bounds the error of the policy $\tilde{\pi}$ that is greedy with respect to the fixed point $\tilde{U}^*$ of $MT_{\mathcal{QR}}$, where $\mathcal{R} = \mathcal{R}_{\tilde{\pi}}$ is chosen to depend on the policy (for some policy π, $\mathcal{R}_\pi$ is a weighted Euclidean projection to the compressed space induced by the aggregation, where the weights depend on the stationary distribution of π). Formally, the policy is defined by $\tilde{\pi} = GT_{\mathcal{QR}_{\tilde{\pi}}}\tilde{U}^*$ where $\tilde{U}^* = MT_{\mathcal{QR}_{\tilde{\pi}}}\tilde{U}^*$. Similar to our results (and differently from the ones in Ávila Pires and Szepesvári, 2016b), $U^* = MT_{\mathcal{Q}}u^*$ is not necessarily the fixed point of $MT_{\mathcal{QR}}$. In contrast, however, our result is proven for general $\mathcal{R}$. At this time it is not clear whether with a specific choice of $\mathcal{R}$ (like $\mathcal{R}_{\hat{\pi}}$) the terms involved in the definition of ϵ_1 would cancel the additional $1/(1-\gamma)$ factor. For what it is worth, we note that for the "counterexample" that Van Roy (2006) presents, when $\mathcal{R} = \mathcal{R}_{\hat{\pi}}$, ϵ_1 scales with $1/(1-\gamma)$ only (as opposed to scaling with $1/(1-\gamma)^2$), showing that our bound has the ability to exploit the benefits of a "good" choice of $\mathcal{R}$. However, it remains to be seen whether this, or some other systematic way of choosing $\mathcal{R}$, or at the very least some choice of $\mathcal{R}$ will always cancel the extra $1/(1-\gamma)$ factor.

Empirical studies can also clarify which specific instances of MBRL with factored semi-linear models (perhaps, with factored *linear* models) are preferable. We can mention successful applications in the context of kernel-based reinforcement learning (Grünewälder et al., 2012; Yao et al., 2014; Lever et al., 2016), where, however, scaling with the amount of data available for constructing the model is still needs to be improved. Lever et al. (2016) attack this scalability issue using compression techniques, and there are a number of other works that provide potential tools for addressing the scalability issue (see Le et al., 2013; Hsieh et al., 2014, and references therein).

To summarize, this chapter advances our understanding of model errors on policy error in reinforcement learning. We improve previous bounds by using a versatile set of norms and introduce new bounds which has the potential of better scaling with the discount factor, while at the same time we extend the range of the models by relaxing previous assumptions.

These generalized results yield bounds for (to the best of our knowledge) all previous MBRL works that rely on factored linear models. By effectively using the language of Banach lattices, the proofs the results by Ávila Pires and Szepesvári (2016b) are shorter, while at the same time hold the promise of being generalizable beyond MDPs. We believe that our approach may lead to advances in the analysis and design of alternate approaches to reinforcement learning, namely both in approximate linear programming and approximate dynamic programming.

Chapter 5

Conclusion

In this dissertation, we have analyzed CBPI with practical classifiers and MBRL with factored semi-linear models. Both analyses have been an effort toward the development of practical RL algorithms, *i.e.*, algorithms that can be executed with a small (polynomial) amount of computation, while yielding effective policies in a number of scenarios.

We have presented results in the context of classification. Although these results are also of independent interest, our aim was to establish supporting results for policy error bounds referring to CBPI. We have also presented policy error bounds for MBRL methods that use factored semi-linear models, an abstract framework that generalizes many MBRL methods, including promising approaches that have been subject of recent interested in the community.

With these results, we have strengthened the theoretical foundations of CBPI and MBRL methods, and with them we hope to encourage experimental studies both to validate our findings, and also reveal where the analyses can be refined.

A number of experimental studies related to our theoretical work can be carried out, concerning CBPI and MBRL with factored semi-linear models. For example, empirical comparisons of different surrogate losses in the contexts of cost-sensitive classification and of CBPI are lacking. One can also investigate and compare, based on empirical evidence, specific instances of MBRL methods with factored semi-linear (or linear) models.

We can also mention a interesting extensions of our reinforcement learning results to different settings. For example, the average reward setting and MDPs with infinitely many actions. Extensions to infinite $|\mathcal{A}|$ seem more plausible for the MBRL results, because classification methods are still often quite dependent on finiteness of the number of actions, both statistically, as we can see from the upper-bounds, and computationally, as we can see from the structure of losses often considered and the need to evaluate the maximum selector for score functions. The structured prediction literature has made progress in the direction of handling a large number of classes, but, to the best of of knowledge and in our opinion, there is still quite a jump from finite to infinite $\mathcal{A}$ ($\mathcal{Y}$).

A very interesting (and challenging!) direction for future work is to understand the covariate-shift issues that affect the policy error bounds that we have presented. In CBPI, we may be able to improve the policy error bounds that we have used (Theorem 3.2.7). Clearly, the behavior policy used to generate data at the beginning of each iteration has a determining effect on the policy error bounds. MBRL with factored semi-linear models, on the other hand, are more robust to covariate-shift effects. Is this a peculiarity of factored semi-linear models, or can we obtain similar results for, say, CBPI?

Bibliography

Aliprantis, C. D. and Border, K. C. (2007). *Infinite Dimensional Analysis: A Hitchhiker's Guide*. Springer.

Anthony, M. and Bartlett, P. (2009). *Neural Network Learning: Theoretical Foundations*. Cambridge University Press.

Antos, A., Csaba Szepesvári, and Munos, R. (2008a). Fitted Q-iteration in continuous action-space MDPs. In Platt, J. C., Koller, D., Singer, Y., and Roweis, S. T., editors, *Advances in Neural Information Processing Systems 20*, pages 9–16. Curran Associates, Inc.

Antos, A., Szepesvári, C., and Munos, Munos, R. (2008b). Learning near-optimal policies with Bellman-residual minimization based fitted policy iteration and a single sample path. *Machine Learning*, 71(1):89–129.

Ávila Pires, B. and Szepesvári, C. (2016a). Multiclass classification calibration functions. *CoRR*, arXiv:0902.0885.

Ávila Pires, B. and Szepesvári, C. (2016b). Policy error bounds for model-based reinforcement learning with factored linear models. In Feldman, V. and Rakhlin, A., editors, *Proceedings of the 29th Annual Conference on Learning Theory*, volume 49, pages 1–31.

Ávila Pires, B., Szepesvári, C., and Ghavamzadeh, M. (2013). Cost-sensitive multiclass classification risk bounds. In Dasgupta, S. and McAllester, D., editors, *Proceedings of The 30th International Conference on Machine Learning*, volume 28 of *JMLR: Workshop & Conference Proceedings (ICML'13)*, pages 1391–1399.

Bagnell, J. A., Kakade, S., Ng, A. Y., and Schneider, J. G. (2003). Policy search by dynamic programming. In Thrun, S., Saul, L. K., and Schölkopf, B., editors, *Advances in Neural Information Processing Systems*, volume 16, pages 831–838, Cambridge, MA. MIT Press.

Bakir, G., Hofmann, T., Schölkopf, B., Smola, A. J., Taskar, B., and Vishwantathan, S. V. N. (2007). *Predicting Structured Data*. Advances in neural information processing systems. MIT Press.

Barreto, A., Precup, D., and Pineau, J. (2014a). Practical kernel-based reinforcement learning. *arXiv preprint arXiv:1407.5358*.

Barreto, A. M. S. and Fragoso, M. D. (2011). Computing the stationary distribution of a finite Markov chain through stochastic factorization. *SIAM Journal on Matrix Analysis and Applications*, 32(4):1513–1523.

Barreto, A. M. S., Pineau, J., and Precup, D. (2014b). Policy iteration based on stochastic factorization. *Journal of Artificial Intelligence Research*, 50:763–803.

Barreto, A. S. M., Precup, D., and Pineau, J. (2011). Reinforcement learning using kernel-based stochastic factorization. In Shawe-Taylor, J., Zemel, R. S., Bartlett, P. L., Pereira, F., and Weinberger, K. Q., editors, *Advances in Neural Information Processing Systems 24*, pages 720–728. Curran Associates, Inc.

Bartlett, P. L., Jordan, M. I., and McAuliffe, J. D. (2006). Convexity, classification, and risk bounds. *Journal of the American Statistical Association*, 101(473):138–156.

Bartlett, P. L. and Mendelson, S. (2002). Rademacher and Gaussian complexities: Risk bounds and structural results. *Journal of Machine Learning Research*, 3:463–482.

Bartlett, P. L., Mendelson, S., and Neeman, J. (2012). ℓ_1-regularized linear regression: persistence and oracle inequalities. *Probability theory and related fields*, 154(1-2):193–224.

Beijbom, O., Saberian, M., Kriegman, D., and Vasconcelos, N. (2014). Guess-averse loss functions for cost-sensitive multiclass boosting. In Xing, E. P. and Jebara, T., editors, *Proceedings of The 31st International Conference on Machine Learning*, volume 32 of *JMLR: Workshop & Conference Proceedings (ICML'14)*, pages 586–594.

Bertsekas, D. P. (2007). *Dynamic Programming and Optimal Control*, volume 2 of *Athena Scientific optimization and computation series*. Athena Scientific, 3rd edition.

Bertsekas, D. P. (2010). *Dynamic programming and optimal control*, volume 2. Athena Scientific, 3rd edition.

Bertsekas, D. P. (2011). Approximate policy iteration: A survey and some new methods. *Journal of Control Theory and Applications*, 9(3):310–335.

Bertsekas, D. P. (2012). Weighted sup-norm contractions in dynamic programming: A review and some new applications. *Dept. Elect. Eng. Comput. Sci., Massachusetts Inst. Technol., Cambridge, MA, USA, Tech. Rep. LIDS-P-2884*.

Bertsekas, D. P. (2016). *Noncontractive Total Cost Problems*, volume 2 of *Athena Scientific optimization and computation series*, chapter 4, Noncontractive Total Cost Problems. Athena Scientific. Updated version `http://web.mit.edu/dimitrib/www/DP2_Chapter%204_UPDATED.pdf`.

Bertsekas, D. P. and Ioffe, S. (1996). Temporal differences-based policy iteration and applications in neuro-dynamic programming. Technical Report LIDS-P-2349, Laboratory for Information and Decision Systems Report, MIT, Cambridge, MA.

Bertsekas, D. P. and Tsitsiklis, J. N. (1996). *Neuro-dynamic Programming*. Anthropological Field Studies. Athena Scientific.

Boucheron, S., Bousquet, O., and Lugosi, G. (2005). Theory of classification: A survey of some recent advances. *ESAIM: Probability and Statistics*, 9:323–375.

Brefeld, U., Geibel, P., and Wysotzki, F. (2003). Support vector machines with example dependent costs. In Lavrač, N., Gamberger, D., Blockeel, H., and Todorovski, L., editors, *Machine Learning: ECML 2003: Proceedings of the 14th European Conference on Machine Learning*, pages 23–34, Berlin, Heidelberg. Springer Berlin Heidelberg.

Burgiel, H. (1997). How to lose at tetris. *Mathematical Gazette*, 81:194–200.

Buşoniu, L., Babuška, R., Schutter, B. D., and Ernst, D. (2010a). *Reinforcement Learning and Dynamic Programming Using Function Approximators*. CRC Press, Inc., Boca Raton, FL, USA, 1st edition.

Buşoniu, L., Babuška, R., Schutter, D., and Ernst, D. (2010b). *Reinforcement Learning and Dynamic Programming Using Function Approximators*. Automation and Control Engineering. CRC Press.

Buşoniu, L., Lazaric, A., Ghavamzadeh, M., Munos, R., Babuška, R., and De Schutter, B. (2012). *Least-squares methods for policy iteration*, chapter 3, pages 75–109. Springer.

Crammer, K. and Singer, Y. (2003). Ultraconservative online algorithms for multiclass problems. *Journal of Machine Learning Research*, 3:951–991.

Daniely, A., Sabato, S., Ben-David, S., and Shalev-Shwartz, S. (2013). Multiclass learnability and the ERM principle. *CoRR*, arXiv:1308.2893.

de Farias, D. P. and Van Roy, B. (2003). The linear programming approach to approximate dynamic programming. *Operations Research*, 51(6):850–865.

Doğan, Ü., Glasmachers, T., and Igel, C. (2016). A unified view on multiclass support vector classification. *Journal of Machine Learning Research*, 17(45):1–32.

Farahmand, A.-M., Precup, D., Barreto, A., and Ghavamzadeh, M. (2014). Classification-based approximate policy iteration: Experiments and extended discussions. *arXiv preprint arXiv:1407.0449*.

Farahmand, A.-M., Szepesvári, C., and Munos, R. (2010). Error propagation for approximate policy and value iteration. In Lafferty, J. D., Williams, C. K. I., Shawe-Taylor, J., Zemel, R. S., and Culotta, A., editors, *Advances in Neural Information Processing Systems*, volume 23, pages 568–576. Curran Associates, Inc.

Fern, A., Yoon, S. W., and Givan, R. (2003). Approximate policy iteration with a policy language bias. In Thrun, S., Saul, L. K., and Schölkopf, B., editors, *Advances in Neural Information Processing Systems*, volume 16, pages 847–854, Cambridge, MA. MIT Press.

Gabillon, V., Ghavamzadeh, M., and Scherrer, B. (2013). Approximate dynamic programming finally performs well in the game of Tetris. In Burges, C. J. C., Bottou, L., Welling, M., Ghahramani, Z., and Weinberger, K. Q., editors, *Advances in Neural Information Processing Systems*, volume 26, pages 1754–1762. Curran Associates, Inc.

Gabillon, V., Lazaric, A., Ghavamzadeh, M., and Scherrer, B. (2011). Classification-based policy iteration with a critic. In Getoor, L. and Scheffer, T., editors, *Proceedings of the 28th International Conference on Machine Learning*, ICML'11, pages 1049–1056, New York, NY, USA. ACM.

Gneiting, T. and Raftery, A. E. (2007). Strictly proper scoring rules, prediction, and estimation. *Journal of the American Statistical Association*, 102(477):359–378.

Grünewälder, S., Baldassarre, L., Pontil, M., Gretton, A., and Lever, G. (2011). Modeling transition dynamics in MDPs with RKHS embeddings of conditional distributions. *CoRR*, arXiv:1112.4722.

Grünewälder, S., Lever, G., Baldassarre, L., Pontil, M., and Gretton, A. (2012). Modelling transition dynamics in MDPs with RKHS embeddings. In Langford, J. and Pineau, J., editors, *Proceedings of the 29th International Conference on Machine Learning*, ICML'12, pages 535–542, New York, NY, USA. Omnipress.

Guruprasad, H. and Agarwal, S. (2012). Classification calibration dimension for general multiclass losses. In Bartlett, P. L., Pereira, F., Burges, C. J. C., Bottou, L., and Weinberger, K. Q., editors, *Advances in Neural Information Processing Systems*, volume 25, pages 2087–2095. Curran Associates, Inc.

Hastie, T., Tibshirani, R., and Friedman, J. (2009). *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. Springer Series in Statistics. Springer, 2nd edition.

Höffgen, K.-U., Simon, H.-U., and Horn, K. S. V. (1995). Robust trainability of single neurons. *Journal of Computer and System Sciences*, 50:114–125.

Hsieh, C.-J., Si, S., and Dhillon, I. S. (2014). Fast prediction for large-scale kernel machines. In Ghahramani, Z., Welling, M., Cortes, C., Lawrence, N. D., and Weinberger, K. Q., editors, *Advances in Neural Information Processing Systems*, volume 27, pages 3689–3697. Curran Associates, Inc.

Hutter, M. (2014). *Extreme State Aggregation beyond MDPs*, volume 8776 of *Lecture Notes in Computer Science*, pages 185–199. Springer International Publishing, Cham.

Kakade, S. and Langford, J. (2002). Approximately optimal approximate reinforcement learning. In Sammut, C. and Hoffmann, A. G., editors, *Proceedings of the Nineteenth International Conference on Machine Learning*, ICML'02, pages 267–274, San Francisco, CA, USA. Morgan Kaufmann Publishers Inc.

Koltchinskii, V. (2011). *Oracle Inequalities in Empirical Risk Minimization and Sparse Recovery Problems: École d'Été de Probabilités de Saint-Flour XXXVIII-2008*. Lecture Notes in Mathematics. Springer Berlin Heidelberg.

Koltchinskii, V. and Panchenko, D. (2002). Empirical margin distributions and bounding the generalization error of combined classifiers. *Annals of Statistics*, pages 1–50.

Kuznetsov, V., Mohri, M., and Syed, U. (2014). Multi-class deep boosting. In *Advances in Neural Information Processing Systems*, pages 2501–2509.

Kveton, B. and Theocharous, G. (2012). Kernel-based reinforcement learning on representative states. In *Proceedings of the Twenty-Sixth AAAI Conference on Artificial Intelligence*, AAAI'12, pages 977–983. AAAI Press.

Lagoudakis, M. G. and Parr, R. (2003a). Approximate policy iteration using large-margin classifiers. *IJCAI'03 Proceedings of the 18th international joint conference on Artificial intelligence*, pages 1432–1434.

Lagoudakis, M. G. and Parr, R. (2003b). Least-squares policy iteration. *Journal of Machine Learning Research*, 4:1107–1149.

Langford, J. and Zadrozny, B. (2003). Reducing T-step reinforcement learning to classification. In *Proceedings of the Machine Learning Reductions Workshop*.

Langford, J. and Zadrozny, B. (2005). Relating reinforcement learning performance to classification performance. In *Proceedings of the 22nd International Conference on Machine Learning*, ICML'05, pages 473–480, New York, NY, USA. ACM.

Lazaric, A., Ghavamzadeh, M., and Munos, R. (2010). Analysis of a classification-based policy iteration algorithm. In Fürnkranz, J. and Joachims, T., editors, *Proceedings of the 27th International Conference on Machine Learning*, ICML'10, pages 607–614, Haifa, Israel. Omnipress.

Lazaric, A., Ghavamzadeh, M., and Munos, R. (2016). Analysis of classification-based policy iteration algorithms. *Journal of Machine Learning Research*, 17(19):1–30.

Le, Q., Sarlós, T., and Smola, A. (2013). Fastfood–approximating kernel expansions in loglinear time. In Dasgupta, S. and McAllester, D., editors, *Proceedings of The 30th International Conference on Machine Learning*, volume 28 of *JMLR: Workshop & Conference Proceedings (ICML'13)*, page 244–252.

Lee, Y., Lin, Y., and Wahba, G. (2004). Multicategory support vector machines: Theory and application to the classification of microarray data and satellite radiance data. *Journal of the American Statistical Association*, 99(465):67–81.

Lei, Y., Dogan, U., Binder, A., and Kloft, M. (2015). Multi-class SVMs: From tighter data-dependent generalization bounds to novel algorithms. In Cortes, C., Lawrence, N. D., Lee, D. D., Sugiyama, M., and Garnett, R., editors, *Advances in Neural Information Processing Systems*, volume 28, pages 2035–2043. Curran Associates, Inc.

Lever, G., Shawe-Taylor, J., Stafford, R., and Szepesvári, C. (2016). Compressed conditional mean embeddings for model-based reinforcement learning. In *Proceedings of the Thirtieth AAAI Conference on Artificial Intelligence*, AAAI'16. AAAI Press.

Li, L., Bulitko, V., and Greiner, R. (2007). Focus of attention in reinforcement learning. *Journal of Universal Computer Science*, 13(9):1246–1269.

Liu, Y. (2007). Fisher consistency of multicategory support vector machines. *Proceedings of the Eleventh International Conference on Artificial Intelligence and Statistics*, 2:289–296.

Mammen, E. and Tsybakov, A. B. (1999). Smooth discrimination analysis. *The Annals of Statistics*, 27(6):1808–1829.

Mason, L., Baxter, J., Bartlett, P. L., and Frean, M. R. (2000). Boosting algorithms as gradient descent. In Solla, S. A., Leen, T. K., and Müller, K., editors, *Advances in Neural Information Processing Systems*, volume 12, pages 512–518, Cambridge, MA. MIT Press.

Maximov, Y. and Reshetova, D. (2015). Tight risk bounds for multi-class margin classifiers. *arXiv preprint arXiv:1507.03040*.

Meyer-Nieber, P. (1991). *Banach Lattices*. Springer.

Mohri, M., Rostamizadeh, A., and Talwalkar, A. (2012). *Foundations of Machine Learning*. MIT Press.

Mroueh, Y., Poggio, T., Rosasco, L., and Slotine, J.-J. (2012). Multiclass learning with simplex coding. In Bartlett, P. L., Pereira, F., Burges, C. J. C., Bottou, L., and Weinberger, K. Q., editors, *Advances in Neural Information Processing Systems*, volume 25, pages 2798–2806. Curran Associates, Inc.

Munos, R. (2003). Error bounds for approximate policy iteration. In Fawcett, T. and Mishra, N., editors, *Machine Learning, Proceedings of the Twentieth International Conference*, ICML'03, pages 560–567. AAAI Press.

Nock, R. and Nielsen, F. (2009). Bregman divergences and surrogates for learning. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 31(11):2048–2059.

Ormoneit, D. and Sen, Ś. (2002). Kernel-based reinforcement learning. *Machine Learning*, 49(2–3):161–178.

Pérez-Cruz, F., Ghahramani, Z., and Pontil, M. (2007). *Predicting Structured Data*, chapter 12, Kernel Conditional Graphical Models. In Bakir et al. (2007). Available at `http://mlg.eng.cam.ac.uk/zoubin/papers/CGM.pdf`.

Pires, B. A. and Szepesvári, C. (2015). Pathological effects of variance on classification-based policy iteration. In *Workshops at the Twenty-Ninth AAAI Conference on Artificial Intelligence*.

Pollard, D. (1984). *Convergence of Stochastic Processes*. Springer Series in Statistics. Springer-Verlag New York.

Powell, W. (2011). *Approximate Dynamic Programming: Solving the Curses of Dimensionality*. Wiley Series in Probability and Statistics. Wiley.

Precup, D., Pineau, J., and Barreto, A. S. (2012). On-line reinforcement learning using incremental kernel-based stochastic factorization. In *Advances in Neural Information Processing Systems 25*, pages 1484–1492.

Puterman, M. L. (1994). *Markov Decision Processes — Discrete Stochastic Dynamic Programming*. John Wiley & Sons, Inc.

Ramaswamy, H. G., Agarwal, S., and Tewari, A. (2013). Convex calibrated surrogates for low-rank loss matrices with applications to subset ranking losses. In Burges, C. J. C., Bottou, L., Welling, M., Ghahramani, Z., and Weinberger, K. Q., editors, *Advances in Neural Information Processing Systems*, volume 26, pages 1475–1483. Curran Associates, Inc.

Ravindran, B. (2004). *An algebraic approach to abstraction in reinforcement learning*. PhD thesis, University of Massachusetts Amherst.

Reid, M. D. and Williamson, R. C. (2010). Composite binary losses. *Journal of Machine Learning Research*, 11:2387–2422.

Rifkin, R. and Klautau, A. (2004). In defense of one-vs-all classification. *Journal of Machine Learning Research*, 5:101–141.

Scherrer, B. (2014). Approximate policy iteration schemes: A comparison. In Xing, E. P. and Jebara, T., editors, *Proceedings of The 31st International Conference on Machine Learning*, volume 32 of *JMLR: Workshop & Conference Proceedings (ICML'14)*, pages 1314—-1322.

Schruben, L. W. (2010). Common random numbers. In *Wiley Encyclopedia of Operations Research and Management Science*. John Wiley & Sons, Inc. Retrieved from `http://dx.doi.org/10.1002/9780470400531.eorms0166`.

Scott, C. (2011). Surrogate losses and regret bounds for cost-sensitive classification with example-dependent costs. In Getoor, L. and Scheffer, T., editors, *Proceedings of the 28th International Conference on Machine Learning*, ICML'11, pages 153–160, New York, NY, USA. ACM.

Shalev-Shwartz, S. and Ben-David, S. (2014). *Understanding Machine Learning: From Theory to Algorithms*. Understanding Machine Learning: From Theory to Algorithms. Cambridge University Press.

Shi, Q., Reid, M. D., Caetano, T., Van den Hengel, A., and Wang, Z. (2015). A hybrid loss for multiclass and structured prediction. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 37(1):2–12.

Singh, S., Jaakkola, T., and Jordan, M. (1995). Reinforcement learning with soft state aggregation. In *NIPS-7*, pages 361–368.

Singh, S. P. and Yee, R. C. (1994). An upper bound on the loss from approximate optimal-value functions. *Machine Learning*, 16(3):227–233.

Sorg, J. and Singh, S. (2009). Transfer via soft homomorphisms. In *Proceedings of The 8th International Conference on Autonomous Agents and Multiagent Systems – Volume 2*, pages 741–748. International Foundation for Autonomous Agents and Multiagent Systems.

Spaan, M. T. J. (2012). *Partially Observable Markov Decision Processes*, chapter 12, pages 387–414. Springer.

Steinwart, I. (2007). How to compare different loss functions and their risks. *Constructive Approximation*, 26(2):225–287.

Steinwart, I. and Christmann, A. (2008). *Support vector machines*. Springer.

Sutton, R. S. and Barto, A. G. (1998). *Reinforcement Learning: An Introduction (Adaptive Computation and Machine Learning)*. The MIT Press.

Sutton, R. S., McAllester, D. A., Singh, S. P., and Mansour, Y. (1999). Policy gradient methods for reinforcement learning with function approximation. In Solla, S. A., Leen, T. K., and Müller, K., editors, *Advances in Neural Information Processing Systems*, volume 12, Cambridge, MA. MIT Press.

Szepesvári, C. (2010). *Algorithms for Reinforcement Learning*. Morgan & Claypool.

Szita, I. and Lörincz, A. (2006). Learning Tetris using the noisy cross-entropy method. *Neural computation*, 18(12):2936–2941.

Szita, I. and Szepesvári, C. (2010). SZ-Tetris as a benchmark for studying key problems of reinforcement learning. In *Proceedings of the ICML 2010 Workshop on Machine Learning and Games*.

Tsochantaridis, I., Joachims, T., Hofmann, T., and Altun, Y. (2005). Large margin methods for structured and interdependent output variables. *Journal of Machine Learning Research*, 6:1453–1484.

Van Roy, B. (2006). Performance loss bounds for approximate value iteration with state aggregation. *Mathematics of Operations Research*, 31(2):234–244.

Vapnik, V. (2013). *The Nature of Statistical Learning Theory*. Springer New York.

Whitt, W. (1978). Approximations of dynamic programs, I. *Mathematics of Operations Research*, 3(3):231–243.

Wiering, M. and van Otterlo, M., editors (2012). *Reinforcement Learning: State-of-the-Art*. Springer.

Yao, H., Szepesvári, C., Pires, B. A., and Zhang, X. (2014). Pseudo-MDPs and factored linear action models. In *IEEE Symposium on Adaptive Dynamic Programming and Reinforcement Learning (ADPRL), 2014*, pages 1–9. IEEE.

Zadrozny, B. and Elkan, C. (2001). Learning and making decisions when costs and probabilities are both unknown. In *Proceedings of the seventh ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 204–213. ACM.

Zadrozny, B., Langford, J., and Abe, N. (2003). Cost-sensitive learning by cost-proportionate example weighting. In *Third IEEE International Conference on Data Mining, ICDM 2003*, pages 435–442.

Zhang, T. (2002). Covering number bounds of certain regularized linear function classes. *Journal of Machine Learning Research*, 2(Mar):527–550.

Zhang, T. (2004). Statistical analysis of some multi-category large margin classification methods. *Journal of Machine Learning Research*, 5:1225–1251.

Zou, H., Zhu, J., and Hastie, T. (2006). The margin vector, admissible loss and multiclass margin-based classifiers. Technical report, Statistics Departement, Stanford University.

Appendix A

Proofs

A.1 Chapter 2 Proofs

A.1.1 Section 2.3 Proofs

To prove Theorem 2.3.6, we will use the following result by Ávila Pires and Szepesvári (2016a).

Lemma A.1.1 (Adapted from Lemma 19 of Ávila Pires and Szepesvári, 2016a)**.** *Consider $L^{\text{Zhang,CI}}$ convex with ψ non-decreasing and $\mathcal{S} = \mathbb{R}^{|\mathcal{Y}|}$. Then for all $p \in \Delta_{|\mathcal{Y}|}$, with $Y \sim p$, and every $\varepsilon > 0$, we have that*

$$\inf_{s \in \mathcal{T}(\mathcal{S}, \varepsilon, -p)} \mathbb{E}\left(L^{\text{Zhang,CI}}(s, Y)\right) = \inf_{s \in \mathcal{M}(\mathcal{S}, j_\varepsilon) \cap \mathcal{M}(\mathcal{S}, j_0)} \mathbb{E}\left(L^{\text{Zhang,CI}}(s, Y)\right).$$

Proof of Theorem 2.3.6, page 24. The this proof reuses Lemma A.1.1 and then mirrors the proof of Lemma 20 of Ávila Pires and Szepesvári (2016a).

Fix the distribution p of C, any $\varepsilon > 0$, and pick $j_\varepsilon \in \mathcal{J}(\varepsilon, \mathbb{E}(C))$ and $j_0 \in \mathcal{J}(0, \mathbb{E}(C))$ (breaking ties arbitrarily). Let $\rho_k \doteq \mathbb{E}(\max_{k'} c_{k'} - c_k)$ $(k \in \mathcal{Y})$, $\overline{\rho} \doteq \mathbf{1}_{|\mathcal{Y}|}^\top \rho$, $q_k \doteq \overline{\rho}^{-1} \rho_k$ $(k \in \mathcal{Y})$, $\varepsilon' \doteq \overline{\rho}^{-1} \varepsilon$, $j'_{\varepsilon'} \in \mathcal{J}(\varepsilon', -q)$ and $j'_0 \in \mathcal{J}(0, -q)$ (again, breaking ties arbitrarily). We have that $q \in \Delta_{|\mathcal{Y}|}$ and

$$\begin{aligned} R^{\text{surr}}(s, p) &= \sum_{k=1}^{|\mathcal{Y}|} \rho_k \psi(s_k) + \mathbb{E}\left(\max_k c_k\right) \sum_{k=1}^{|\mathcal{Y}|} \varphi(s_k) \\ &= \overline{\rho}\, R^{\text{surr}}(s, q), \end{aligned}$$

where R^{surr} is the surrogate risk w.r.t. $L^{\text{Zhang,CI}}$ (and the same choices of ψ,

φ, but with $F(t) = \overline{\rho}^{-1}\mathbb{E}(\max_k c_k)\, t$). Since for any $j \in |\mathcal{Y}|$,

$$\max_k -c_k + c_j = (\max_k \max_{k'} c_{k'} - c_k) - (\max_{k'} c_{k'} - c_j) = \overline{\rho}\left(\max_k q_k - q_j\right),$$

we get that $\mathcal{T}(\mathcal{S}, \varepsilon, \mathbb{E}(C)) = \mathcal{T}(\mathcal{S}, \overline{\rho}^{-1}\varepsilon, -q)$ and $\mathcal{J}(t, \mathbb{E}(C)) = \mathcal{J}(\overline{\rho}^{-1}t, -q)$ for every $t \geq 0$.

We can now apply Lemma A.1.1, by recalling that ψ is non-decreasing by assumption:

$$\begin{aligned}
\inf_{s \in \mathcal{T}(\mathcal{S}, \varepsilon, \mathbb{E}(C))} R^{\text{surr}}(s, p) &= \inf_{s \in \mathcal{T}(\mathcal{S}, \overline{\rho}^{-1}\varepsilon, -q)} R^{\text{surr}}(s, p) \\
&= \overline{\rho} \cdot \inf_{s \in \mathcal{T}(\mathcal{S}, \overline{\rho}^{-1}\varepsilon, -q)} R^{\text{surr}}(s, q) \\
&= \overline{\rho} \cdot \inf_{s \in \mathcal{M}(\mathcal{S}, j'_{\varepsilon'}) \cap \mathcal{M}(\mathcal{S}, j'_0)} R^{\text{surr}}(s, q) \\
&= \inf_{s \in \mathcal{M}(\mathcal{S}, j'_{\varepsilon'}) \cap \mathcal{M}(\mathcal{S}, j'_0)} R^{\text{surr}}(s, p) \\
&= \inf_{s \in \mathcal{M}(\mathcal{S}, j_\varepsilon) \cap \mathcal{M}(\mathcal{S}, j_0)} R^{\text{surr}}(s, p).
\end{aligned}$$

Letting p' be the distribution of the random variable $(C_{j_0}, C_{j_\varepsilon})$, we get that

$$\begin{aligned}
\delta_{\max}(\varepsilon, p) &= \inf_{s \in \mathcal{M}(\mathcal{S}, j_\varepsilon) \cap \mathcal{M}(\mathcal{S}, j_0)} \sup_{s' \in \mathcal{S}} R^{\text{surr}}(s, p) - R^{\text{surr}}(s', p) \\
&\geq \inf_{s \in \mathcal{M}(\mathcal{S}, j_\varepsilon) \cap \mathcal{M}(\mathcal{S}, j_0)} \sup_{\substack{s' \in \mathcal{S}: \\ s'_k = s_k, k \notin \{j_\varepsilon, j_0\}}} R^{\text{surr}}(s, p) - R^{\text{surr}}(s', p) \\
&= \inf_{\substack{s \in \mathcal{S}: \\ s_1 = s_2}} \sup_{s' \in \mathcal{S}} R^{\text{surr}}(s, p') - R^{\text{surr}}(s', p') \\
&= \delta_{\text{binary}}(\varepsilon, p).
\end{aligned}$$

The result for $|\mathcal{Y}| = 2$ is follows from Definitions 2.3.2 and 2.3.3 combined with the result of Lemma A.1.1, since in this case $\mathcal{M}(\mathcal{S}, j_\varepsilon) \cap \mathcal{M}(\mathcal{S}, j_0) = \{s \in \mathcal{S} : s_1 = s_2\}$. □

Proof of Proposition 2.3.7, page 25. Then the result follows by combining Definition 2.2.1 and the fact that

$$\begin{aligned}
R(s, p') - \inf_{s' \in \mathcal{S}} R(s', p') &= \mathbb{E}\left(C_{f(s)} - \min_k C_k\right) - \min_{k'} \mathbb{E}\left(C_{k'} - \min_k C_k\right) \\
&= \mathbb{E}\left(C_{f(s)}\right) - \min_{k'} \mathbb{E}(C_{k'}) \\
&= R(s, p) - \inf_{s' \in \mathcal{S}} R(s', p).
\end{aligned}$$

□

Proof of Theorem 2.3.8, page 27. We have that

$$R^{\text{surr}}_{L_u^{\text{Red}}}(s,p) = \overline{c} R^{\text{surr}}_L(s,q).$$

For all $s \in \mathcal{S}$ and every non-negative-cost distribution p and $\varepsilon > 0$, if

$$R^{\text{surr}}_L(s,q) - \inf_{s' \in \mathcal{S}} R^{\text{surr}}_L(s,q) < \delta\Big(\frac{\varepsilon}{\overline{c}}, q\Big) \tag{A.1.1}$$

then

$$\max_k q_k - q_{f(s)} < \frac{\varepsilon}{\overline{c}},$$

by the assumption that L has calibration function δ. Since

$$\begin{aligned}\overline{c}\Big(\max_k q_k - q_{f(s)}\Big) &= \max_k \mathbb{E}(u(C)) - \mathbb{E}(C_k) - \mathbb{E}(u(C)) + \mathbb{E}\Big(C_{f(s)}\Big) \\ &= \mathbb{E}\Big(C_{f(s)}\Big) - \min_k \mathbb{E}(C_k),\end{aligned}$$

it follows that $(\varepsilon, p) \mapsto \overline{c}\delta\big(\frac{\varepsilon}{\overline{c}}, q\big)$ is a calibration function for L^{Red} and the cost-distribution p. □

A.1.2 Section 2.4 Proofs

Proof of Proposition 2.4.1, page 29. By simple algebra,

$$\begin{aligned}R^{\text{surr}}(\widehat{H}, S) &\le \widehat{R}^{\text{surr}}(\widehat{H}, S) + \sup_{h \in \mathcal{H}} \Big|\widehat{R}^{\text{surr}}(h,S) - R^{\text{surr}}(h,S)\Big| \\ &= \inf_{h \in \mathcal{H}} \widehat{R}^{\text{surr}}(h,S) + \sup_{h \in \mathcal{H}} \Big|\widehat{R}^{\text{surr}}(h,S) - R^{\text{surr}}(h,S)\Big| \\ &\le \inf_{h \in \mathcal{H}} R^{\text{surr}}(h,S) + \Big|\widehat{R}^{\text{surr}}(h,S) - R^{\text{surr}}(h,S)\Big| \\ &\quad + \sup_{h \in \mathcal{H}} \Big|\widehat{R}^{\text{surr}}(h,S) - R^{\text{surr}}(h,S)\Big|.\end{aligned}$$

□

Proof of Lemma 2.4.8, page 33. Fix $\delta \in (0,1)$. We have that

$$\widehat{R}^{\text{surr}}(h,S) = \frac{1}{nm|\mathcal{Y}|} \sum_{i=1}^{n} \sum_{j=1}^{m} \sum_{i=1}^{|\mathcal{Y}|} C'_{i,j,k} \varphi(h(X_i)_k)$$

and also

$$\mathbb{E}\left(\widehat{R}^{\text{surr}}(h,S)\middle|\, X_1,\ldots,X_n\right) = \frac{1}{n|\mathcal{Y}|}\sum_{i=1}^{n}\sum_{k=1}^{|\mathcal{Y}|}\mathbb{E}(\,C_{i,k}\varphi(h(X_i)_k)|\,X_i),$$

for any $j \in [m]$.

We will first study the concentration of the empirical risk of a single $h \in \mathcal{H}$ "with respect to" the cost estimates, that is, conditioned on $(X_1,\ldots,X_n)$ (in which case all $C_{i,j}$ are independent). Then we will analyze the analogue, not conditioned on $(X_1,\ldots,X_n)$.

By Hoeffding's inequality (Steinwart and Christmann, 2008, Theorem 6.10, p. 211), for any $h \in \mathcal{H}$

$$\begin{aligned}&\left|\widehat{R}^{\text{surr}}(h,S) - \mathbb{E}\left(\widehat{R}^{\text{surr}}(h,S)\middle|\, X_1,\ldots,X_n\right)\right| \\ &\leq C_{\max}\varphi_{\max,\mathcal{H}}\sqrt{\frac{1}{2nm}\ln\frac{4}{\delta}}\end{aligned}$$

with probability at least $1-\frac{\delta}{2}$, since for each $i \in [n]$ and $j \in [m]$ we have, with probability one,

$$0 \leq \frac{1}{|\mathcal{Y}|}\sum_{k=1}^{|\mathcal{Y}|} C'_{i,j,k}\varphi(h(X_i)_k) \leq C_{\max}\varphi_{\max,\mathcal{H}}. \tag{A.1.2}$$

Applying Hoeffding's inequality once more to remove the conditioning on $(X_1,\ldots,X_n)$ gives us

$$\begin{aligned}&\left|\mathbb{E}\left(\widehat{R}^{\text{surr}}(h,S)\middle|\, X_1,\ldots,X_n\right) - \mathbb{E}\left(\widehat{R}^{\text{surr}}(h,S)\right)\right| \\ &\leq C'_{\max}\varphi_{\max,\mathcal{H}}\sqrt{\frac{1}{2n}\ln\frac{4}{\delta}},\end{aligned}$$

with probability at least $1-\frac{\delta}{2}$, where now

$$0 \leq \frac{1}{|\mathcal{Y}|}\sum_{k=1}^{|\mathcal{Y}|}\mathbb{E}\left(C'_{i,j,k}\varphi(h(X_i)_k)\middle|\, X_i\right) \leq C'_{\max}\varphi_{\max,\mathcal{H}}. \tag{A.1.3}$$

The result follows by taking a union-bound over both uses of Hoeffding's inequality, and by realizing that, by Assumption 2.4.2, $\mathbb{E}\left(\widehat{R}^{\text{surr}}(h,S)\right) = R^{\text{surr}}(h)$ for any $h \in \mathcal{H}$. □

Proof of Lemma 2.4.9, page 34. Fix $\delta \in (0,1)$, as well as $\alpha > 0$ and

$$0 < \alpha' \leq \alpha \tag{A.1.4}$$

to be specified later.

Let

$$\begin{aligned}
S'' &\doteq ((X_1, C'_{1,1}, \ldots, C'_{1,m}), \ldots, (X_{2n}, C'_{2n,1}, \ldots, C'_{2n,m})), \\
S &\doteq ((X_1, C'_{1,1}, \ldots, C'_{1,m}), \ldots, (X_n, C'_{n,1}, \ldots, C'_{n,m})), \\
S' &\doteq ((X_{n+1}, C'_{n+1,1}, \ldots, C'_{n+1,m}), \ldots, (X_{2n}, C'_{2n,1}, \ldots, C'_{2n,m})).
\end{aligned}$$

We will apply symmetrization in this proof), and let $(\sigma_1, \ldots, \sigma_n)$ be i.i.d. *Rademacher* ($\{-1,1\}$-valued uniform) random variables. Define the vectors $\ell(h,S) \doteq \mathrm{vec}((i,j,k) \mapsto \varphi(h(X_i,k)))$ and $c(S) \doteq \mathrm{vec}((i,j,k) \mapsto \frac{1}{nm|\mathcal{Y}|}\sigma_i C'_{i,j,k})$, and note that $\|c(S)\|_1 \leq C_{\max}$ with probability one.

Let $\mathcal{C}$ be a minimum $\frac{\alpha'}{C_{\max}}$-covering of $\{\ell(h,S) : h \in \mathcal{H}\}$ in ∞-norm. We first claim that $|\mathcal{C}| \leq N_{\infty/\infty}\left(\frac{\alpha'}{C_{\max}}, n, \mathcal{H}\right)$. We have that[1] $\ell(h,S)_{i,j,k} = \ell(h,S)_{i,j',k}$ for every $h \in \mathcal{H}$, $i \in [n]$, $j, j' \in [m]$ and $k \in \mathcal{Y}$, so without loss of generality we drop the second index in $\ell(h,S)$, and then it is easy to see from (2.4.3) that $|\mathcal{C}| \leq N_{\infty/\infty}\left(\frac{\alpha'}{C_{\max}}, n, \mathcal{H}\right)$.

[1] We abuse notation and denote let $\ell(h,S)_{(i,j,k)}$ denote the appropriate coordinate of $\ell(h,S)$ with the "flattened" index corresponding to (i,j,k).

By symmetrization (Pollard, 1984, pp. 14–15), we get that for some $h' \in \mathcal{C}$

$$
\begin{aligned}
&\mathbb{P}\left(\sup_{h\in\mathcal{H}}\left|\widehat{R}^{\mathrm{surr}}(h,S) - R^{\mathrm{surr}}(h)\right| > 8\alpha\right) \\
&\quad\leq 2\mathbb{P}\left(\sup_{h\in\mathcal{H}}\left|\widehat{R}^{\mathrm{surr}}(h,S) - \widehat{R}^{\mathrm{surr}}(h,S')\right| > 4\alpha\right) \\
&\quad= 2\mathbb{P}\left(\sup_{h\in\mathcal{H}}\left|\langle c(S), \ell(h,S)\rangle - \langle c(S'), \ell(h,S')\rangle\right| > 4\alpha\right) \\
&\quad\leq 4\mathbb{P}\left(\sup_{h\in\mathcal{H}}\left|\langle c(S), \ell(h,S)\rangle\right| > 2\alpha\right) \\
&\quad\leq 4\mathbb{P}\left(\left|\langle c(S), \ell(h',S)\rangle\right| + \sup_{h\in\mathcal{H}}\left|\langle c(S), \ell(h,S) - \ell(h',S)\rangle\right| > 2\alpha\right) \\
&\quad\leq 4\mathbb{P}\left(\left|\langle c(S), \ell(h',S)\rangle\right| + \|c(S)\|_1 \sup_{h\in\mathcal{H}}\left\|\ell(h,S) - \ell(h',S)\right\|_\infty > 2\alpha\right) \\
&\quad\leq 4\mathbb{P}\left(\left|\langle c(S), \ell(h',S)\rangle\right| > \alpha\right) + 4\mathbb{P}\left(\|c(S)\|_1 \sup_{h\in\mathcal{H}}\left\|\ell(h,S) - \ell(h',S)\right\|_\infty > \alpha\right) \\
&\quad= 4\mathbb{P}\left(\left|\langle c(S), \ell(h',S)\rangle\right| > \alpha\right).
\end{aligned}
$$

In order to perform the symmetrization, we must use (in the first line above) Lemma 2.4.8 with $\delta = \frac{1}{2}$, which imposes a restriction on α:

$$
4\alpha \geq \varphi_{\max,\mathcal{H}}\left(C'_{\max}\sqrt{\frac{1}{2n}\ln 8} + C_{\max}\sqrt{\frac{1}{2nm}\ln 8}\right). \tag{A.1.5}
$$

Moreover, in the symmetrization we can only "swap" the observations $(X_i, C'_{i,1}, \ldots, C'_{i,m})$ with their counterparts in S', but not each individual $C'_{i,j}$ and their counterparts in S', hence the Rademacher variables $(\sigma_1, \ldots, \sigma_n)$ appear.

Similar to Pollard (1984, Theorem 24, pp. 25–26), we use the union bound to get that

$$
\mathbb{P}\left(\sup_{h'\in\mathcal{C}}\left|\langle c(S), \ell(h',S)\rangle\right| > \alpha\right) \leq \sum_{h'\in\mathcal{C}}\mathbb{P}\left(\left|\langle c(S), \ell(h',S)\rangle\right| > \alpha\right),
$$

and then we bound each summand on the right-hand side individually[2].

[2] We have to be mindful that the supremum over $h' \in \mathcal{C}$ can be shown to be measurable,

By realizing that for every $h' \in \mathcal{C}$ we have

$$\mathbb{E}\left(\left\langle c(S), \ell(h', S)\right\rangle \middle| S\right) = 0,$$

and adapting the proof of Lemma 2.4.8 with minor modifications, we can see that for all $h' \in \mathcal{C}$ we have, with probability at least $1 - \frac{\delta}{4N(\alpha')}$,

$$\begin{aligned} &\left|\left\langle c(S), \ell(h', S)\right\rangle\right| \\ &\quad \leq 2\varphi_{\max,\mathcal{H}} C'_{\max} \sqrt{\frac{1}{2n} \ln \frac{16N(\alpha')}{\delta}} \\ &\qquad + \varphi_{\max,\mathcal{H}} C_{\max} \sqrt{\frac{1}{2nm} \ln \frac{16N(\alpha')}{\delta}}. \end{aligned} \tag{A.1.6}$$

In contrast to the result of Lemma 2.4.8, a factor of 2 appears multiplying the first term in the right-hand side of (A.1.6), because the range of $\sigma_i \mathbb{E}\left(C_{i,j,k} \middle| X_i, \sigma_i\right)$ is $[-C'_{\max}, C'_{\max}]$. The factor of two does not appear on the second term of the right-hand side of (A.1.6) because for all $i \in [n]$, the range of $\left|\sigma_i C_{i,j,k}\right|$ is $[0, C_{\max}]$ with probability one. Take

$$\alpha' = 2\frac{C_{\max}\varphi_{\max,\mathcal{H}}}{nm},$$

and

$$\alpha = 8\varphi_{\max,\mathcal{H}}\left(2C'_{\max}\sqrt{\frac{1}{2n}\ln\frac{16N(\alpha')}{\delta}} + C_{\max}\sqrt{\frac{1}{2nm}\ln\frac{16N(\alpha')}{\delta}}\right).$$

We can see that both (A.1.4) and (A.1.5) are satisfied, and the result follows by simple algebra. □

Proof of Lemma 2.4.10, page 34. Fix any $(x_1, \ldots, x_n) \in \mathcal{X}^n$ and define the shorthand $\mathcal{V} \doteq \left\{(i,k) \mapsto h(x_i)_k : h \in \mathcal{H}_{\phi,B}\right\}$. For any $v, v' \in \mathcal{V}$, we have that

$$\begin{aligned} \left\|\varphi(v) - \varphi(v')\right\|_{\infty/\infty} &= \max_{i,k} \left|\varphi(v_{i,k}) - \varphi(v'_{i,k})\right| \\ &\leq \mathrm{Lip}_\varphi(BB_*) \max_{i,k} \left|v_{i,k} - v'_{i,k}\right| \\ &= \mathrm{Lip}_\varphi(BB_*) \left\|v - v'\right\|_{\infty/\infty}. \end{aligned}$$

because we can assume that $\mathcal{C}$ is finite (otherwise $N_\infty\left(\frac{\alpha'}{C_{\max}}, n, \mathcal{H}\right) = \infty$ and the lemma holds vacuously). Then the supremum is just a maximum, which is measurable as long as $\mathcal{H}$ is a set of measurable score functions, which is assumed.

Above, we have used that $\sup_{h\in\mathcal{H},x\in\mathcal{X},k\in\mathcal{Y}}|h(x)_k| \leq BB_*$ (by Hölder's inequality). Moreover, each $v \in \mathcal{V}$ corresponds to at least one $w \in \mathbb{R}^d$ s.t. $\|w\| \leq B$ and $v_{i,k} = \langle \phi(x_i, k), w\rangle$ for all $i \in [n]$ and $k \in \mathcal{Y}$, so for any $v, v' \in \mathcal{V}$ we have $w, w' \in \mathbb{R}^d$ s.t.

$$\left\|v - v'\right\|_{\infty/\infty} = \max_{i,k}\left|\left\langle \phi(x_i, k), w - w'\right\rangle\right| \leq B_*\|w - w'\|.$$

Therefore, the ε-covering number of the d-dimensional unit ball in $\|\cdot\|$ gives us an $\frac{\varepsilon}{\mathrm{Lip}_\varphi(BB_*)BB_*}$-covering number of $\varphi \circ \mathcal{V}$ in ∞/∞-norm. It is well-known that there exists an ε-covering of the d-dimensional unit ball in $\|\cdot\|$ with size at most $(1 + \frac{2}{\varepsilon})^d$, which gives the result. □

Proof of Theorem 2.4.11, page 35. This result follows by combining Proposition 2.4.1 and Lemmas 2.4.8 to 2.4.10. We take Lemma 2.4.8 to hold with probability at least $1 - \frac{1}{5}\delta$ and Lemma 2.4.9 with probability at least $1 - \frac{4}{5}\delta$, and we upper-bound the left-hand side of Lemma 2.4.8 by the left-hand side of Lemma 2.4.9.

We point out that Assumptions 2.4.2 to 2.4.4 and 2.4.6 are assumed to hold and that Assumption 2.4.5 is satisfied with $\varphi_{\max,\mathcal{H}_{\phi,B}} = BB_*$, so the assumptions of Lemmas 2.4.8 and 2.4.9 are satisfied. □

Sketch of proof, Lemma 2.4.12, page 36. Lemma 2.4.12 follows from some (careful) modifications to the proof of Lemma 2.4.9, and different choices of α, α' and, of course, the covering $\mathcal{C}$. Let $\mathcal{C}$ be a $\frac{\sqrt{nm|\mathcal{Y}|}}{C_{\max}}\alpha'$-covering of $\{\ell(h, S) : h \in \mathcal{H}\}$ in 2-norm.

In the symmetrization step, we use the Cauchy-Schwarz inequality

(instead of Hölder's inequality with $\|\cdot\|_1$ and $\|\cdot\|_\infty$) to see that

$$
\begin{aligned}
&\mathbb{P}\left(\sup_{h\in\mathcal{H}}\left|\widehat{R}^{\mathrm{surr}}(h,S)-R^{\mathrm{surr}}(h)\right|>8\alpha\right)\\
&\quad\le 4\mathbb{P}\left(\left|\left\langle c(S),\ell(h',S)\right\rangle\right|+\sup_{h\in\mathcal{H}}\left|\left\langle c(S),\ell(h,S)-\ell(h',S)\right\rangle\right|>2\alpha\right)\\
&\quad\le 4\mathbb{P}\left(\left|\left\langle c(S),\ell(h',S)\right\rangle\right|+\|c(S)\|_2\sup_{h\in\mathcal{H}}\left\|\ell(h,S)-\ell(h',S)\right\|_2>2\alpha\right)\\
&\quad\le 4\mathbb{P}\left(\left|\left\langle c(S),\ell(h',S)\right\rangle\right|>\alpha\right)+4\mathbb{P}\left(\|c(S)\|_2\sup_{h\in\mathcal{H}}\left\|\ell(h,S)-\ell(h',S)\right\|_2>\alpha\right)\\
&\quad\le 4\mathbb{P}\left(\left|\left\langle c(S),\ell(h',S)\right\rangle\right|>\alpha\right)+4\mathbb{P}\left(\alpha'>\alpha\right)\\
&\quad\le 4\mathbb{P}\left(\left|\left\langle c(S),\ell(h',S)\right\rangle\right|>\alpha\right),
\end{aligned}
$$

where we have used that

$$\|c(S)\|_2\le\frac{C_{\max}}{\sqrt{nm|\mathcal{Y}|}}$$

and that $\alpha'\le\alpha$. We observe that $j\mapsto\ell(h,S)(i,j,k)$ is constant for all $i\in[n]$ and $k\in\mathcal{Y}$, so an $\frac{\varepsilon}{\sqrt{m}}$-covering of $\{(i,k)\mapsto\varphi(h(X_i)_k:h\in\mathcal{H}\}$ in Frobenius norm immediately gives us an ε-covering of $\{\ell(h,S):h\in\mathcal{H}\}$ in 2-norm, and vice-versa.

Choosing α as in the proof of Lemma 2.4.9 and

$$\alpha'=\varphi_{\max,\mathcal{H}}\sqrt{\frac{1}{n}}\left(\frac{C_{\max}}{\sqrt{m}}\vee C'_{\max}\right)$$

ensures that $\alpha'\le\alpha$, and gives the result. □

Proof of Lemma 2.4.13, page 37. Fix any $(x_1,\dots,x_n)\in\mathcal{X}^n$ and define the shorthand $\mathcal{V}\doteq\left\{(i,k)\mapsto h(x_i)_k:h\in\mathcal{H}_{\phi,B}\right\}$. For any $v,v'\in\mathcal{V}$, we have that

$$\left\|\varphi(v)-\varphi(v')\right\|_F=\mathrm{Lip}_\varphi(BB_*)\left\|v-v'\right\|_F,$$

since $\sup_{h\in\mathcal{H},x\in\mathcal{X},k\in\mathcal{Y}}|h(x)_k|\le BB_*$ (by Hölder's inequality). Thus,

$$N_F(\varepsilon,n,\varphi\circ\mathcal{H}_{\phi,B})\le N_2\left(\frac{\varepsilon}{\mathrm{Lip}_\varphi(BB_*)},\{\mathrm{vec}(v):v\in\mathcal{V}\}\right).$$

To obtain the first bound in Lemma 2.4.13, we use Theorem 2 of Zhang (2002), which gives us covering number bounds for $\mathcal{V}$:

$$\log_2 N_2(\varepsilon, \{\mathrm{vec}(v) : v \in \mathcal{V}\}) \leq \left\lceil \frac{B^2 B_*^2}{\varepsilon^2} \right\rceil \log_2(2d+1).$$

For the second result, we use Corollary 3 of Zhang (2002), which tells us that if $p = 2$ then

$$\log_2 N_2(\varepsilon, \{\mathrm{vec}(v) : v \in \mathcal{V}\}) \leq \left\lceil \frac{B^2 B_*^2}{\varepsilon^2} \right\rceil \log_2(2n|\mathcal{Y}|+1),$$

since $\mathrm{vec}(v)$ is $n|\mathcal{Y}|$-dimensional for $v \in \mathcal{V}$. □

A.2 Chapter 3 Proofs

Proof of Theorem 2.4.14, page 38. This result follows by combining Proposition 2.4.1 and Lemmas 2.4.8, 2.4.12 and 2.4.13. Assumptions 2.4.2 to 2.4.4 and 2.4.6 are assumed to hold and Assumption 2.4.5 is satisfied with $\varphi_{\max,\mathcal{H}_{\phi,B}} = BB_*$, so the assumptions of Lemmas 2.4.8 and 2.4.12 are satisfied. With simple algebra, we can verify the the claim that if (2.4.5) then $\varepsilon_1 \geq \varepsilon_2$, otherwise $\varepsilon_1 \leq 6\varepsilon_2$.

We take Lemma 2.4.8 to hold with probability at least $1 - \frac{1}{5}\delta$ and Lemma 2.4.12 with probability at least $1 - \frac{4}{5}\delta$, and we upper-bound the left-hand side of Lemma 2.4.8 by the left-hand side of Lemma 2.4.12. If (2.4.5) holds, we also use

$$\frac{C'_{\max}}{\sqrt{n}} \left(\sqrt{\frac{1}{m}} \vee \frac{C'_{\max}}{C_{\max}} \right)^{-1} \leq \frac{C_{\max}}{\sqrt{n}},$$

and

$$\frac{C_{\max}}{\sqrt{nm}} \left(\sqrt{\frac{1}{m}} \vee \frac{C'_{\max}}{C_{\max}} \right)^{-1} \leq \frac{C_{\max}}{\sqrt{n}}.$$

Then, if (2.4.5) holds,

$$R^{\mathrm{surr}}(\widehat{H}, S) - \inf_{h \in \mathcal{H}_{\phi,B}} R^{\mathrm{surr}}(h, S), \leq \varepsilon_1 = \varepsilon_1 \vee \varepsilon_2,$$

otherwise,

$$R^{\mathrm{surr}}(\widehat{H}, S) - \inf_{h \in \mathcal{H}_{\phi,B}} R^{\mathrm{surr}}(h, S) \leq \varepsilon_2.$$

The result follows by upper-bounding $\varepsilon_2 \leq \varepsilon_1 \vee \varepsilon_2$, which is not too loose: If Equation (2.4.5) fails to hold, then $\varepsilon_1 \vee \varepsilon_2 \leq 6\varepsilon_2$. □

Proof of Theorem 3.2.4, page 54. Let h be as in (3.2.3), which gives us $h = \widehat{H}$ (see (2.4.2) with $L = \frac{1}{|\mathcal{A}|} L^{\mathrm{LLW,CS}}$ and $\mathcal{H} = \mathcal{H}_{\phi,B}$) and $\hat{\pi} = f \circ \widehat{H}$. We define the shorthands

$$\varepsilon_1 \doteq 2|\mathcal{A}|(1 + BB_*)\sqrt{\frac{32d}{n} \ln \frac{20(1+nm)}{\delta}} \left(2C'_{\max} + C_{\max}\sqrt{\frac{1}{m}} \right)$$

$$\varepsilon_2 \doteq \inf_{h \in \mathcal{H}_{\phi,B}} R^{\mathrm{surr}}_{\frac{1}{|\mathcal{A}|}L^{\mathrm{LLW}}}(h) - \inf_{h \in (\mathbb{R}^{|\mathcal{A}|})^{\mathcal{X}}} R^{\mathrm{surr}}_{\frac{1}{|\mathcal{A}|}L^{\mathrm{LLW}}}(h).$$

The true risk of $\pi' \in \Pi$ in our setting corresponds to, for any $i \in [n]$, $j \in [m]$,

$$R(\pi') = \mathbb{E}\left(\frac{1}{m} \sum_{j'=1}^{m} C'_{i,j',\pi(X_i)} \right) = \mathbb{E}\left(C_{\pi(X)} \right).$$

Moreover, the true risk of $\hat{\pi}$ conditioned on the sample S is

$$R(\hat{\pi}, S) \doteq \mathbb{E}\left(C_{\hat{\pi}(X)} \,\middle|\, S \right).$$

The sample S satisfies Assumption 2.4.2, φ^{hinge} satisfies Assumption 2.4.3 as well as Assumption 2.4.4 with $\mathrm{Lip}_{\varphi}(T) = 1$ for all $T \in \mathbb{R}$, ϕ has been assumed to satisfy Assumption 2.4.6 and $\mathcal{H}_{\phi,B}$ satisfies Assumption 2.4.5 with $\varphi_{\max,\mathcal{H}_{\phi,B}} = 1 + BB_*$. Thus, we can apply Theorem 2.4.11, so that

$$R^{\mathrm{surr}}_{\frac{1}{|\mathcal{A}|}L^{\mathrm{LLW}}}(\widehat{H}) - \inf_{h \in \mathcal{H}_{\phi,B}} R^{\mathrm{surr}}_{\frac{1}{|\mathcal{A}|}L^{\mathrm{LLW}}}(h) \leq \varepsilon_1$$

with probability at least $1 - \delta$, where we have also used that $c \leq (1 + nm)$. Then

$$R^{\mathrm{surr}}_{L^{\mathrm{LLW}}}(\widehat{H}) - \inf_{h \in (\mathbb{R}^{|\mathcal{A}|})^{\mathcal{X}}} R^{\mathrm{surr}}_{L^{\mathrm{LLW}}}(h) \leq |\mathcal{A}|(\varepsilon_1 + \varepsilon_2).$$

Note that above the surrogate risks are w.r.t. the unscaled loss $L^{\mathrm{LLW,CS}}$, which has calibration function $\delta(\varepsilon) = \varepsilon$, as shown by Ávila Pires et al. (2013, Table 1) provided that $\{h(x) : h \in \mathcal{H}_{\phi,B}\} \subset \mathcal{S}_0$ for all $x \in \mathcal{X}$ (that is, all scores sum to zero), which is ensured by Assumption 3.2.3.

Therefore, by Theorem 2.2.2, we have

$$R(\widehat{H}, S) - \inf_{h \in (\mathbb{R}^{|\mathcal{A}|})^{\mathcal{X}}} R(h) \leq |\mathcal{A}|(\varepsilon_1 + \varepsilon_2)$$

with probability at least $1-\delta$. Let $\pi'' = GV^{\pi}$ (the greedy policy w.r.t. the value-function of π). For all $i \in [n]$ and $j \in [m]$, we have

$$\begin{aligned} R(\widehat{H}, S) - \inf_{\pi' \in \Pi} R(\pi') &\geq \mathbb{E}\left(C_{\hat{\pi}(X)} - C_{\pi''(X)} \middle| S \right) \\ &= \mathbb{E}\left(\mathrm{Ret}(T_{\pi''(X)}) - \mathrm{Ret}(T_{\hat{\pi}(X)}) \middle| S \right) \\ &\geq \mathbb{E}\big(V^{\pi}(X, \pi''(X))\big) - \mathbb{E}\big(V^{\pi}(X, \hat{\pi}(X)) \big| S\big) - B_{\pi,h} \\ &= \sup_{\pi' \in \Pi} \mathbb{E}\big(V^{\pi}(X, \pi'(X))\big) - \mathbb{E}\big(V^{\pi}(X, \hat{\pi}(X)) \big| S\big) - B_{\pi,h}. \end{aligned}$$

Above, we have used Assumption 3.2.2.

Combining the derivations above, we get that for all $\pi \in \Pi$

$$\mathbb{E}(V^{\pi}(X, h(X))) - \mathbb{E}\big(V^{\pi}(X, \hat{\pi}H(X)) \big| S\big) \leq |\mathcal{A}|(\varepsilon_1 + \varepsilon_2) + B_{\pi,h},$$

which concludes the proof. □

Sketch of the proof of Theorem 3.2.5, page 54. The proof of Theorem 3.2.5 follows the same arguments as Theorem 3.2.4, but we use Theorem 2.4.14 instead of Theorem 2.4.11. □

A.3 Chapter 4 Proofs

Proof of Proposition 4.3.2, page 71. To see why this holds, take for example M. Then for any $U, V \in \mathcal{V}^{\mathcal{A}}$, $MU - MV \leq M_{|\cdot|}(U - V)$ ($\leq$ denotes the componentwise ordering) and by swapping the order of U, V, we also get $|MU - MV| \leq M_{|\cdot|}(U - V)$. Now, since for any $f, g \in \mathcal{V}$, $|f| \leq |g|$ implies $\|f\| \leq \|g\|$, we get $\|MU - MV\| \leq \|M_{|\cdot|}(U - V)\| = \|U - V\|_{\mathcal{V}^{\mathcal{A}}}$. For M^{π}, since it is a linear operator, $\mathrm{Lip}(M^{\pi}) = \|M^{\pi}\|$, and for any $V \in \mathcal{V}^{\mathcal{A}}$, $\left|M^{\pi}V^{\mathcal{A}}\right| \leq M\left|V^{\mathcal{A}}\right|$, so $\mathrm{Lip}(M^{\pi}) \leq \mathrm{Lip}(M) \leq 1$. The statement is proven for the other operators analogously. □

Proof of Lemma 4.4.4, page 79. We start by recalling the definition of U^*, $U^* \doteq MT_{\mathcal{Q}}u^*$, and the identity $u^* = N'T_{\mathcal{R}'\mathcal{Q}}u^*$, which holds because $\mathrm{Lip}(N'T_{\mathcal{R}'\mathcal{Q}}) < 1$ as a consequence of our assumptions. We also have

$V = NT_{\mathcal{P}}V$. Here, however, the identities $\mathcal{R}M = M'\mathcal{R}'$, $U^* = MT_{\mathcal{Q}\mathcal{R}}U^*$ and $\mathcal{R}U^* = u^*$ do not necessarily hold.

Inequality (4.4.5) is a fairly simple observation:

$$\begin{aligned}\|V - U^*\| &= \|NT_{\mathcal{P}}V - NT_{\mathcal{Q}}u^*\| \\ &\leq \gamma \operatorname{Lip}(N)\|\mathcal{P}V - \mathcal{Q}u^*\| \\ &\leq \gamma \operatorname{Lip}(N)\|\mathcal{P}V - \mathcal{Q}u^*\|.\end{aligned}$$

In order to prove (4.4.6), it suffices to use a variant of the proof of Lemma 15 of Ávila Pires and Szepesvári (2016b), as follows:

$$\begin{aligned}\|V - U^*\| &= \|NT_{\mathcal{P}}V - NT_{\mathcal{Q}}u^*\| \\ &\leq \|NT_{\mathcal{P}}V - NT_{\mathcal{Q}\mathcal{R}}V\| + \|NT_{\mathcal{Q}\mathcal{R}}V - NT_{\mathcal{Q}}u^*\| \\ &\leq \gamma \operatorname{Lip}(N)\|\mathcal{P}V - \mathcal{Q}\mathcal{R}V\| + \operatorname{Lip}(NT_{\mathcal{Q}})\|\mathcal{R}V - u^*\| \\ &\leq \gamma\|\mathcal{P}V - \mathcal{Q}\mathcal{R}V\| + B\gamma\,\|\mathcal{R}V - u^*\|.\end{aligned}$$

Lemma 14 of Ávila Pires and Szepesvári (2016b) with $T = N'T_{\mathcal{R}'\mathcal{Q}}$ and $a = b = \alpha < 1$ gives us

$$\begin{aligned}\|\mathcal{R}V - u^*\| &= \inf_{k\geq 1} \frac{1}{1-\alpha^k}\left\|\mathcal{R}V - (N'T_{\mathcal{R}'\mathcal{Q}})^k\mathcal{R}V\right\| \\ &\leq \frac{1}{1-\alpha}\left\|\mathcal{R}NT_{\mathcal{P}}V - N'\mathcal{R}'T_{\mathcal{Q}\mathcal{R}}V\right\| \\ &\leq \frac{1}{1-\alpha}\Big(\left\|\mathcal{R}NT_{\mathcal{P}}V - N'\mathcal{R}'T_{\mathcal{P}}V\right\| \\ &\qquad\qquad + \left\|N'\mathcal{R}'T_{\mathcal{P}}V - N'\mathcal{R}'T_{\mathcal{Q}\mathcal{R}}V\right\|\Big) \\ &\leq \frac{1}{1-\alpha}\left(\left\|\mathcal{R}NT_{\mathcal{P}}V - N'\mathcal{R}'T_{\mathcal{P}}V\right\| + \gamma\left\|\mathcal{R}'(\mathcal{P}V - \mathcal{Q}\mathcal{R}V)\right\|\right).\end{aligned}$$

Note that linearity of $\mathcal{R}'$ is used so that $\mathcal{R}'T_{\mathcal{P}}V - \mathcal{R}'T_{\mathcal{Q}\mathcal{R}}V = \mathcal{R}'(\mathcal{P}V - \mathcal{Q}\mathcal{R}V)$. Hence, (4.4.6) holds.

To obtain (4.4.7), we first point out that $\operatorname{Lip}(MT_{\mathcal{P}}) \leq \gamma \operatorname{Lip}(\mathcal{P})$ and that, since $M^{\hat{\pi}}$ is linear, $\operatorname{Lip}(M^{\hat{\pi}}T_{\mathcal{P}}) = \gamma \operatorname{Lip}(M^{\hat{\pi}}\mathcal{P})$. We cannot use that $U^* = MT_{\mathcal{Q}\mathcal{R}}U^*$, so we just use Lemma 14 of Ávila Pires and Szepesvári

(2016b) with $T = NT_{\mathcal{P}}$ and $a = b = \beta < 1$ instead:

$$\begin{aligned}\|V - U^*\| &= \inf_{k\geq 1} \frac{1}{1-\beta^k}\left\|U^* - NT_{\mathcal{P}}U^*\right\| \\ &\leq \frac{1}{1-\beta}\|NT_{\mathcal{Q}}u^* - NT_{\mathcal{P}}U^*\| \\ &\leq \frac{\gamma}{1-\beta}\|\mathcal{Q}u^* - \mathcal{P}U^*\|.\end{aligned}$$

□

Proof of Theorem 4.4.5, page 81. We start with the triangle inequality in (4.4.4). To obtain $\varepsilon_1(V)$ we use Lemma 4.4.4-(4.4.6) with $\alpha = \beta_{\eta,\mathcal{R}'\mathcal{Q}}$. The conditions of Lemma 4.4.4-(4.4.6) are fulfilled by Corollary 4.3.3 and Assumption 4.3.7, and because η is γ-Lyapunov w.r.t. $\mathcal{R}'\mathcal{Q}$ (via Assumption 4.3.6-(ii)).

Lemma 4.4.4-(4.4.6) gives ε_2 after we realize that $\operatorname{Lip}(MT_{\mathcal{P}}) \leq \gamma \operatorname{Lip}(\mathcal{P}) = \gamma\beta_\nu < 1$ and that $\operatorname{Lip}(M^{\hat{\pi}}T_{\mathcal{P}}) \leq \gamma \operatorname{Lip}(\mathcal{P}) = \gamma\beta_\nu < 1$, since ν is γ-Lyapunov w.r.t. $\mathcal{P}$ by Assumption 4.3.6-(i). □

Proof of Theorem 4.4.7, page 83. The first step is to use (4.4.4). Then we see that Corollary 4.3.3 ensures that Assumption 4.3.1 is satisfied, and Assumption 4.3.4 guarantees that $\|\mathcal{R}'\mathcal{Q}\| \leq 1$. Thus, Lemma 4.4.4-(4.4.6) with $\alpha = \gamma$ gives us $\|U^* - V\|_{\mu,p} \leq \varepsilon_1(V)$ for $V \in \left\{V^*, V^{\hat{\pi}}\right\}$.

To bound $\left\|U^* - V^{\hat{\pi}}\right\|_{\mu,p} \leq \varepsilon_2(V^{\hat{\pi}})$ we proceed exactly as in the proof of Theorem 4.4.7. If $(I - \gamma M^{\hat{\pi}}\mathcal{P})$ is not invertible, then $C_{\gamma,\hat{\pi},\mathcal{P},\mu,\xi} = \infty$ and the result holds vacuously, so assume otherwise. Since $V^{\hat{\pi}} = M^{\hat{\pi}}T_{\mathcal{P}}V^{\hat{\pi}}$,

$$(I - \gamma M^{\hat{\pi}}\mathcal{P})V^{\hat{\pi}} = M^{\hat{\pi}}r.$$

Moreover,

$$U^* - \gamma M^{\hat{\pi}}\mathcal{P}U^* - M^{\hat{\pi}}r = U^* - M^{\hat{\pi}}T_{\mathcal{P}}U^*.$$

Hence,

$$\begin{aligned}
\left\|U^* - V^{\hat{\pi}}\right\|_{\mu,p} &= \left\|(I - \gamma M^{\hat{\pi}}\mathcal{P})^{-1}(I - \gamma M^{\hat{\pi}}\mathcal{P})(U^* - V^{\hat{\pi}})\right\|_{\mu,p} \\
&\leq \mathrm{Lip}((I - \gamma M^{\hat{\pi}}\mathcal{P})^{-1}) \left\|(I - \gamma M^{\hat{\pi}}\mathcal{P})(U^* - V^{\hat{\pi}})\right\|_{\xi,p} \\
&= C_{\gamma,\hat{\pi},\mathcal{P},\mu,\xi} \frac{1}{1-\gamma}\left\|U^* - M^{\hat{\pi}}T_{\mathcal{P}}U^*\right\|_{\xi,p} \\
&= C_{\gamma,\hat{\pi},\mathcal{P},\mu,\xi} \frac{1}{1-\gamma}\left\|M^{\hat{\pi}}T_{\mathcal{Q}}u^* - M^{\hat{\pi}}T_{\mathcal{P}}U^*\right\|_{\xi,p} \\
&\leq C_{\gamma,\hat{\pi},\mathcal{P},\mu,\xi}\ \mathrm{Lip}(M^{\hat{\pi}})\frac{\gamma}{1-\gamma}\|\mathcal{P}U^* - \mathcal{Q}u^*\|_{\xi,p}
\end{aligned}$$

To conclude, we use that $\mathrm{Lip}(M^{\hat{\pi}}) \leq 1$ by Corollary 4.3.3. □

Proof of Theorem 4.4.6, page 82. Since

$$\|V\|_\mu \leq (\mu(\nu^p))^{\frac{1}{p}}\||V|^p\|_{\infty,\nu^p} = \|\nu\|_{\mu,p}\,\|V\|_{\infty,\nu},$$

we can apply Theorem 4.4.5 to obtain the result. □

Proof of Theorem 4.4.3, page 76. We start by verifying the assumptions of Lemma 4.4.4, so that we can bound the terms on the right-hand side of (4.4.4) with the help of this lemma. Lemma 4.4.4 needs: Assumption 4.3.1, Assumption 4.3.7, $\gamma\,\mathrm{Lip}(\mathcal{R}'\mathcal{Q}) < 1$, $\gamma\,\mathrm{Lip}(\mathcal{P}) < 1$ and $\gamma\,\mathrm{Lip}(M^{\hat{\pi}}\mathcal{P}) < 1$. Assumption 4.3.1 holds by Corollary 4.3.3, whose assumptions are satisfied because Theorem 4.4.3 uses supremum norms. Assumption 4.3.7 holds by assumption. Next, Assumption 4.3.4 implies that $\gamma\,\mathrm{Lip}(\mathcal{R}'\mathcal{Q}) \leq \gamma < 1$. Because $\mathrm{Lip}(\mathcal{P}) = 1$ in supremum norm, we get $\gamma\,\mathrm{Lip}(\mathcal{P}) \leq \gamma < 1$. Finally, $\mathrm{Lip}(\mathcal{P}) = 1$ and Assumption 4.3.1 imply together that $\mathrm{Lip}(M^{\hat{\pi}}\mathcal{P}) \leq \gamma < 1$. The result is obtained by using Lemma 4.4.4 (with $\alpha = \beta = \gamma$) to bound the terms on the right-hand side of (4.4.4). □