Journal of Machine Learning Research 3 (2015) XX-XX Submitted 11/14; Published 03/15

Learning Algorithms for Second-Price Auctions with Reserve

Mehryar Mohri
Courant Institute of Mathematical Sciences
251 Mercer Street
New York, NY

Andrés Muñoz Medina
Courant Institute of Mathematical Sciences
251 Mercer Street
New York, NY

Editor: Gabor Lugosi

Abstract

Second-price auctions with reserve play a critical role in the revenue of modern search engine and popular online sites since the revenue of these companies often directly depends on the outcome of such auctions. The choice of the reserve price is the main mechanism through which the auction revenue can be influenced in these electronic markets. We cast the problem of selecting the reserve price to optimize revenue as a learning problem and present a full theoretical analysis dealing with the complex properties of the corresponding loss function. We further give novel algorithms for solving this problem and report the results of several experiments in both synthetic and real-world data demonstrating their effectiveness.

Keywords: Learning Theory, Auctions, Revenue Optimization.

1. Introduction

Over the past few years, advertisement has gradually moved away from the traditional printed promotion to the more tailored and directed online publicity. The advantages of online advertisement are clear: since most modern search engine and popular online site companies such as Microsoft, Facebook, Google, eBay, or Amazon, may collect information about the users' behavior, advertisers can better target the population sector their brand is intended for.

More recently, a new method for selling advertisements has gained momentum. Unlike the standard contracts between publishers and advertisers where some amount of impressions is required to be fulfilled by the publisher, an Ad Exchange works in a way similar to a financial exchange where advertisers bid and compete between each other for an ad slot. The winner then pays the publisher and his ad is displayed.

The design of such auctions and their properties are crucial since they generate a large fraction of the revenue of popular online sites. These questions have motivated extensive research on the topic of auctioning in the last decade or so, particularly in the theoretical computer science and economic theory communities. Much of this work has focused on the analysis of mechanism design, either to prove some useful property of an existing auctioning mechanism, to analyze its computational efficiency, or to search for an optimal revenue maximization truthful mechanism (see Muthukrishnan (2009) for a good discussion of key research problems related to Ad Exchange and references to a fast growing literature therein).

One particularly important problem is that of determining an auction mechanism that achieves optimal revenue (Muthukrishnan, 2009). In the ideal scenario where the valuation of the bidders is drawn i.i.d. from a given distribution, this is known to be achievable (see for example (Myerson, 1981)). But, even good approximations of such distributions are not known in practice. Game theoretical approaches to the design of auctions have given a series of interesting results including (Riley and Samuelson, 1981; Milgrom and Weber, 1982; Myerson, 1981; Nisan et al., 2007), all of them based on some assumptions about the distribution of the bidders, e.g., the monotone hazard rate assumption.

The results of these publications have set the basis for most Ad Exchanges in practice: the mechanism widely adopted for selling ad slots is that of a *Vickrey auction* (Vickrey, 1961) or *second-price auction with reserve price* (Easley and Kleinberg, 2010). In such auctions, the winning bidder (if any) pays the maximum of the second-place bid and the reserve price. The reserve price can be set by the publisher or automatically by the exchange. The popularity of these auctions relies on the fact that they are incentive-compatible, i.e., bidders bid exactly what they are willing to pay. It is clear that the revenue of the publisher depends greatly on how the reserve price is set: if set too low, the winner of the auction might end up paying only a small amount, even if his bid was high; on the other hand, if it is set too high, then bidders may not bid higher than the reserve price and the ad slot will not be sold.

We propose a learning approach to the problem of determining the reserve price to optimize revenue in such auctions. The general idea is to leverage the information gained from past auctions to predict a beneficial reserve price. Since every transaction on an Exchange is logged, it is natural to seek to exploit that data. This could be used to estimate the probability distribution of the bidders, which can then be used indirectly to come up with the optimal reserve price (Myerson, 1981; Ostrovsky and Schwarz, 2011). Instead, we will seek a discriminative method making use of the loss function related to the problem and taking advantage of existing user features.

Learning methods have been used in the past for the related problems of designing incentive-compatible auction mechanisms (Balcan et al., 2008; Blum et al., 2004), for algorithmic bidding (Langford et al., 2010; Amin et al., 2012), and even for predicting bid landscapes (Cui et al., 2011). Another closely related problem for which machine learning solutions have been proposed is that of revenue optimization for sponsored search ads and click-through rate predictions (Zhu et al., 2009; He et al., 2013; Devanur and Kakade, 2009). But, to our knowledge, no prior work has used historical data in combination with user features for the sole purpose of revenue optimization in this context. In fact, the only publications we are aware of that are directly related to our objective are (Ostrovsky and Schwarz, 2011) and (Cesa-Bianchi et al., 2013), which considers a more general case than (Ostrovsky and Schwarz, 2011).

The scenario studied by Cesa-Bianchi et al. (2013) is that of censored information, which motivates their use of a regret minimization algorithm to optimize the revenue of the seller. Our analysis assumes instead access to full information. We argue that this is a more realistic scenario since most companies do in fact have access to the full historical data. The learning scenario we consider is also more general since it includes the use of features, as is standard in supervised learning. Since user information is communicated to advertisers and bids are made based on that information, it is only natural to include user features in the formulation of the learning problem. A special case of our analysis coincides with the no-feature scenario considered by Cesa-Bianchi et al. (2013), assuming full information. But, our results further extend those of this paper even in that scenario. In particular, we present an $O(m \log m)$ algorithm for solving a key optimization problem used as a subroutine by these authors, for which they do not seem to give an algorithm. We also do not assume that buyers' bids are sampled i.i.d. from a common distribution. Instead, we only assume that the full outcome of each auction is independent and identically distributed. This subtle distinction makes our scenario closer to reality as it is unlikely for all bidders to follow the same underlying value distribution. Moreover, even though our scenario does not take into account a possible strategic behavior of bidders between rounds, it allows for bidders to be correlated, which is common in practice.

This paper is organized as follows: in Section 2, we describe the setup and give a formal description of the learning problem. We discuss the relations between the scenario we consider and previous work on learning in auctions in Section 3. In particular, we show that, unlike previous work, our problem can be cast as that of minimizing the expected value of a loss function, which is a standard learning problem. Unlike most work in this field, however, the loss function naturally associated to this problem does not admit favorable properties such as convexity or Lipschitz continuity. In fact the loss function is discontinuous. Therefore, the theoretical and algorithmic analysis of this problem raises several non-trivial technical issues. Nevertheless, we use a decomposition of the loss to derive generalization bounds for this problem (see Section 4). These bounds suggest the use of structural risk minimization to determine a learning solution benefiting from strong guarantees. This, however, poses a new challenge: solving a highly non-convex optimization problem. Similar algorithmic problems have been of course previously encountered in the learning literature, most notably when seeking to minimize a regularized empirical 0-1 loss in binary classification. A standard method in machine learning for dealing with such issues consists of resorting to a convex surrogate loss (such as the hinge loss commonly used in linear classification). However, we show in Section 4.2 that no convex loss function is calibrated for the natural loss function for this problem. That is, minimizing a convex surrogate could in fact be detrimental to learning. This fact is further empirically verified in Section 6.

The impossibility results of Section 4.2 prompt us to search for surrogate loss functions with weaker regularity properties such as Lipschitz continuity. We describe a loss function with precisely that property which we further show to be consistent with the original loss. We also provide finite sample learning guarantees for that loss function, which suggest minimizing its empirical value while controlling the complexity of the hypothesis set. This leads to an optimization problem which, albeit non-convex, admits a favorable decomposition as a difference of two convex functions (DC-programming). Thus, we suggest using the DC-programming algorithm (DCA) introduced by Tao and An (1998) to solve our optimization problem. This algorithm admits favorable convergence guarantees to a *local minimum*. To further improve upon DCA, we propose a combinatorial algorithm to cycle through different local minima with the guarantee of reducing the objective function at every iteration. Finally, in Section 6, we show that our algorithm outperforms several different baselines in various synthetic and real-world revenue optimization tasks.

2. Setup

We start with the description of the problem and our formulation of the learning setup. We study second-price auctions with reserve, the type of auctions adopted in many Ad Exchanges. In such auctions, the bidders submit their bids simultaneously and the winner, if any, pays the maximum of the value of the second-place bid and a reserve price r set by the seller. This type of auctions benefits from the same truthfulness property as second-price auctions (or Vickrey auctions) Vickrey (1961): truthful bidding can be shown to be a dominant strategy in such auctions. The choice of the reserve price r is the only mechanism through which the seller can influence the auction revenue. Its choice is thus critical: if set too low, the amount paid by the winner could be too small; if set too high, the ad slot could be lost. How can we select the reserve price to optimize revenue?

We consider the problem of learning to set the reserve prices to optimize revenue in second-price auctions with reserve. The outcome of an auction can be encoded by the highest and second-highest bids which we denote by a vector $\mathbf{b} = (b^{(1)}, b^{(2)}) \in \mathcal{B} \subset \mathbb{R}_+^2$. We will assume that there exists an upper bound $M \in (0, +\infty)$ for the bids: $\sup_{b \in \mathcal{B}} b^{(1)} = M$. For a given reserve price r and bid pair $\mathbf{b}$, by definition, the revenue of an auction is given by

$$\text{Revenue}(r, \mathbf{b}) = b^{(2)} \mathbb{1}_{r < b^{(2)}} + r \mathbb{1}_{b^{(2)} \leq r \leq b^{(1)}}.$$

We consider the general scenario where a feature vector $\mathbf{x} \in \mathcal{X} \subset \mathbb{R}^N$ is associated with each auction. In the auction theory literature, this feature vector is commonly referred to as *public information*. In the context of online advertisement, this could be for example information about the user's location, gender or age. The learning problem can thus be formulated as that of selecting out of a hypothesis set H of functions mapping $\mathcal{X}$ to $\mathbb{R}$ a hypothesis h with high expected revenue

$$\mathop{\mathbb{E}}_{(\mathbf{x},\mathbf{b}) \sim D} [\text{Revenue}(h(\mathbf{x}), \mathbf{b})], \tag{1}$$

where D is an unknown distribution according to which pairs $(\mathbf{x}, \mathbf{b})$ are drawn. Instead of the revenue, we will consider a loss function L defined for all $(r, \mathbf{b})$ by $L(r, \mathbf{b}) = -\text{Revenue}(r, \mathbf{b})$, and will equivalently seek a hypothesis h with small expected loss $\mathcal{L}(h) := \mathbb{E}_{(\mathbf{x},\mathbf{b}) \sim D}[L(h(\mathbf{x}), \mathbf{b})]$. As in standard supervised learning scenarios, we assume access to a training sample $S = ((\mathbf{x}_1, \mathbf{b}_1), \ldots, (\mathbf{x}_m, \mathbf{b}_m))$ of size $m \geq 1$ drawn i.i.d. according to D. We will denote by $\widehat{\mathcal{L}}_S(h)$ the empirical loss $\widehat{\mathcal{L}}_S(h) = \frac{1}{m} \sum_{i=1}^m L(h(\mathbf{x}_i, \mathbf{b}_i)$. Notice that we only assume that the auction outcomes are i.i.d. and not that bidders are independent of each other with the same underlying bid distribution, as in some previous work (Cesa-Bianchi et al., 2013; Ostrovsky and Schwarz, 2011). In the next sections, we will present a detailed study of this learning problem, starting with a review of the related literature.

3. Previous work

Here, we briefly discuss some previous work related to the study of auctions from a learning standpoint. One of the earliest contributions in this literature is that of Blum et al. (2004) where the authors studied a posted-price auction mechanism where a seller offers some good at a certain price and where the buyer decides to either accept that price or reject it. It is not hard to see that this type of auctions is equivalent to second-price auctions with reserve with a single buyer. The authors consider a scenario of repeated interactions with different buyers where the goal is to design an incentive-compatible method of setting prices that is competitive with the best fixed-priced strategy in hindsight. A fixed-price strategy is one that simply offers the same price to all buyers. Using a variant of the EXP3 algorithm of Auer et al. (2002), the authors designed a pricing algorithm achieving a $(1+\epsilon)$-approximation to the best fixed-price strategy. This same scenario was also studied by Kleinberg and Leighton (2003) who gave an online algorithm whose regret after T rounds is in $O(T^{2/3})$.

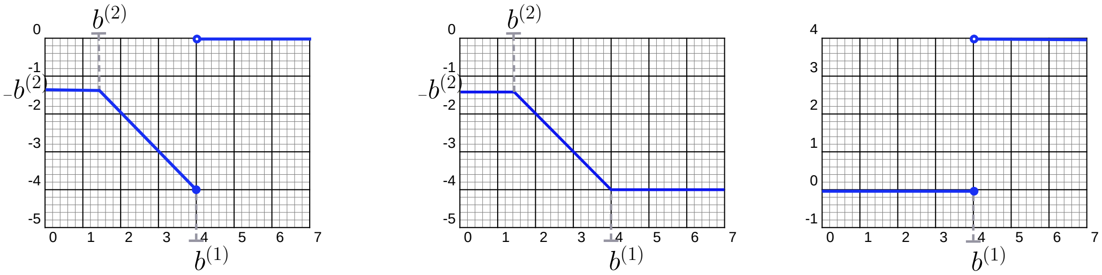
A step further in the design of optimal pricing strategies was proposed by Balcan et al. (2008). One of the problems considered by the authors was that of setting prices for n buyers in a posted-price auction as a function of their public information. Unlike the on-line scenario of Blum et al. (2004), Balcan et al. (2008) considered a batch scenario where all buyers are known in advance. However, the comparison class considered was no longer that of simple fixed-price strategies but functions mapping public information to prices. This makes the problem more challenging and in fact closer to the scenario we consider. The authors showed that finding a $(1+\epsilon)$-optimal truthful mechanism is equivalent to finding an algorithm to optimize the empirical risk associated to the loss function we consider (in the case $b^{(2)} \equiv 0$). There are multiple connections between this work and our results. In particular, the authors pointed out that the discontinuity and asymmetry of the loss function presented several challenges to their analysis. We will see that, in fact, the same problems appear in the derivation of our learning guarantees. But, we will present an algorithm for minimizing the empirical risk which was a crucial element missing in their results.

A different line of work by Cui et al. (2011) focused on predicting the highest bid of a second-price auction. To estimate the distribution of the highest bid, the authors partitioned the space of advertisers based on their campaign objectives and estimated the distribution for each partition. Within each partition, the distribution of the highest bid was modeled as a mixture of log-normal distributions where the means and standard deviations of the mixtures were estimated as a function of the data features. While it may seem natural to seek to predict the highest bid, we show that this is not necessary and that in fact accurate predictions of the highest bid do not necessarily translate into algorithms achieving large revenue (see Section 6).

As already mentioned, the closest previous work to ours is that of Cesa-Bianchi et al. (2013), who studied the problem of directly optimizing the revenue under a partial information setting where the learner can only observe the value of the second-highest bid, if it is higher than the reserve price. In particular, the highest bid remains unknown to the learner. This is a natural scenario for auctions such as those of eBay where only the price at which an object is sold is reported. To do so, the authors expressed the expected revenue in terms of the quantity $q(t) = \mathbb{P}[b^{(2)} > t]$. This can be done as follows:

$$\begin{aligned}
\mathop{\mathbb{E}}_{\mathbf{b}}[\mathrm{Revenue}(r, \mathbf{b})] &= \mathop{\mathbb{E}}_{b^{(2)}}[b^{(2)} \mathbb{1}_{r<b^{(2)}}] + r\,\mathbb{P}[b^{(2)} \leq r \leq b^{(1)}] \qquad (2)\\
&= \int_0^{+\infty} \mathbb{P}[b^{(2)} \mathbb{1}_{r<b^{(2)}} > t]\,dt + r\,\mathbb{P}[b^{(2)} \leq r \leq b^{(1)}]\\
&= \int_0^{r} \mathbb{P}[r < b^{(2)}]\,dt + \int_r^{\infty} \mathbb{P}[b^{(2)} > t]dt + r\,\mathbb{P}[b^{(2)} \leq r \leq b^{(1)}]\\
&= \int_r^{\infty} \mathbb{P}[b^{(2)} > t]\,dt + r(\mathbb{P}[b^{(2)} > r] + 1 - \mathbb{P}[b^{(2)} > r] - \mathbb{P}[b^{(1)} < r])\\
&= \int_r^{\infty} \mathbb{P}[b^{(2)} > t]\,dt + r\,\mathbb{P}[b^{(1)} \geq r].
\end{aligned}$$

The main observation of Cesa-Bianchi et al. (2013) was that the quantity $q(t)$ can be estimated from the observed outcomes of previous auctions. Furthermore, if the buyers' bids are i.i.d., then, one can express $\mathbb{P}[b^{(1)} \geq r]$ as a function of the estimated value of $q(r)$. This implies that the right-hand side of (2) can be accurately estimated and therefore an optimal reserve price can be selected. Their algorithm makes calls to a procedure that maximizes the empirical revenue. The authors, however, did not describe an algorithm for that maximization. A by-product of our work is an efficient algorithm for that procedure. The guarantees of Cesa-Bianchi et al. (2013) are similar to those presented in the next section in the special case of learning without features. However, our derivation is different since we consider a batch scenario while Cesa-Bianchi et al. (2013) treated an online setup for which they presented regret guarantees.



(a) (b)

Figure 1: (a) Plot of the loss function $r \mapsto L(r, \mathbf{b})$ for fixed values of $b^{(1)}$ and $b^{(2)}$; (b) Functions l_1 on the left and l_2 on the right.

4. Learning Guarantees

The problem we consider is an instance of the well known family of supervised learning problems. However, the loss function L does not admit any of the properties such as convexity or Lipschitz continuity often assumed in the analysis of the generalization error, as shown by Figure 1(a). Furthermore, L is discontinuous and, unlike the 0-1 loss function whose discontinuity point is independent of the label, its discontinuity depends on the outcome $\mathbf{b}$ of the auction. Thus, the problem of learning with the loss function L requires a new analysis.

4.1 Generalization bound

To analyze the complexity of the family of functions L_H mapping $\mathcal{X} \times \mathcal{B}$ to $\mathbb{R}$ defined by

$$L_H = \{(\mathbf{x}, \mathbf{b}) \mapsto L(h(\mathbf{x}), \mathbf{b}) \colon h \in H\},$$

we decompose L as a sum of two loss functions l_1 and l_2 with more favorable properties than L. We have $L = l_1 + l_2$ with l_1 and l_2 defined for all $(r, \mathbf{b}) \in \mathbb{R} \times \mathcal{B}$ by

$$l_1(r, \mathbf{b}) = -b^{(2)} \mathbb{1}_{r < b^{(2)}} - r \mathbb{1}_{b^{(2)} \leq r \leq b^{(1)}} - b^{(1)} \mathbb{1}_{r > b^{(1)}}$$
$$l_2(r, \mathbf{b}) = b^{(1)} \mathbb{1}_{r > b^{(1)}}.$$

These functions are shown in Figure 1(b). Note that, for a fixed $\mathbf{b}$, the function $r \mapsto l_1(r, \mathbf{b})$ is 1-Lipschitz since the slope of the lines defining the function is at most 1. We will consider the corresponding families of loss functions: $l_{1H} = \{(\mathbf{x}, \mathbf{b}) \mapsto l_1(h(\mathbf{x}), \mathbf{b}) \colon h \in H\}$ and $l_{2H} = \{(\mathbf{x}, \mathbf{b}) \mapsto l_2(h(\mathbf{x}), \mathbf{b}) \colon h \in H\}$ and use the notion of pseudo-dimension as well as those of empirical and average Rademacher complexity to measure their complexities. The pseudo-dimension is a standard complexity measure (Pollard, 1984) extending the notion of VC-dimension to real-valued functions (see also Mohri et al. (2012)). For a family of functions G and finite sample $S = (z_1, \ldots, z_m)$ of size m, the empirical Rademacher complexity is defined by $\widehat{\mathfrak{R}}_S(G) = \mathbb{E}_{\boldsymbol{\sigma}} \left[\sup_{g \in G} \frac{1}{m} \sum_{i=1}^m \sigma_i g(z_i)\right]$, where $\boldsymbol{\sigma} = (\sigma_1, \ldots, \sigma_m)^\top$, with σ_is independent uniform random variables taking values in $\{-1, +1\}$. The Rademacher complexity of G is defined as $\mathfrak{R}_m(G) = \mathbb{E}_{S \sim D^m}[\widehat{\mathfrak{R}}_S(G)]$.

To bound the complexity of L_H, we will first bound the complexity of the family of loss functions l_{1H} and l_{2H}. Since l_1 is 1-Lipschitz, the complexity of the class l_{1H} can be readily bounded by that of H, as shown by the following proposition.

Proposition 1 *For any sample $S = ((\mathbf{x}_1, \mathbf{b}_1), \ldots, (\mathbf{x}_m, \mathbf{b}_m))$, the empirical Rademacher complexity of l_{1H} can be bounded as follows:*

$$\widehat{\mathfrak{R}}_S(l_{1H}) \leq \widehat{\mathfrak{R}}_S(H).$$

Proof By definition of the empirical Rademacher complexity, we can write

$$\widehat{\mathfrak{R}}_S(l_{1H}) = \frac{1}{m} \mathop{\mathbb{E}}_{\boldsymbol{\sigma}} \left[\sup_{h \in H} \sum_{i=1}^m \sigma_i l_1(h(\mathbf{x}_i), \mathbf{b}_i)\right] = \frac{1}{m} \mathop{\mathbb{E}}_{\boldsymbol{\sigma}} \left[\sup_{h \in H} \sum_{i=1}^m \sigma_i (\psi_i \circ h)(\mathbf{x}_i)\right],$$

where, for all $i \in [1, m]$, ψ_i is the function defined by $\psi_i \colon r \mapsto l_1(r, \mathbf{b}_i)$. For any $i \in [1, m]$, ψ_i is 1-Lipschitz, thus, by the contraction lemma of Appendix A (Lemma 14), the following inequality holds:

$$\widehat{\mathfrak{R}}_S(l_{1H}) \leq \frac{1}{m} \mathop{\mathbb{E}}_{\boldsymbol{\sigma}} \left[\sup_{h \in H} \sum_{i=1}^m \sigma_i h(\mathbf{x}_i) \right] = \widehat{\mathfrak{R}}_S(H),$$

which completes the proof. ■

As shown by the following proposition, the complexity of l_{2H} can be bounded in terms of the pseudo-dimension of H.

Proposition 2 *Let $d = Pdim(H)$ denote the pseudo-dimension of H, then, for any sample $S = ((\mathbf{x}_1, \mathbf{b}_1), \ldots, (\mathbf{x}_m, \mathbf{b}_m))$, the empirical Rademacher complexity of l_{2H} can be bounded as follows:*

$$\widehat{\mathfrak{R}}_S(l_{2H}) \leq M \sqrt{\frac{2d \log \frac{em}{d}}{m}}.$$

Proof By definition of the empirical Rademacher complexity, we can write

$$\widehat{\mathfrak{R}}_S(l_{2H}) = \frac{1}{m} \mathop{\mathbb{E}}_{\boldsymbol{\sigma}} \left[\sup_{h \in H} \sum_{i=1}^m \sigma_i b_i^{(1)} \mathbb{1}_{h(\mathbf{x}_i) > b_i^{(1)}} \right] = \frac{1}{m} \mathop{\mathbb{E}}_{\boldsymbol{\sigma}} \left[\sup_{h \in H} \sum_{i=1}^m \sigma_i \Psi_i(\mathbb{1}_{h(\mathbf{x}_i) > b_i^{(1)}}) \right],$$

where, for all $i \in [1, m]$, Ψ_i is the M-Lipschitz function $x \mapsto b_i^{(1)} x$. Thus, by Lemma 14 combined with Massart's lemma (see for example Mohri et al. (2012)), we can write

$$\widehat{\mathfrak{R}}_S(l_{2H}) \leq \frac{M}{m} \mathop{\mathbb{E}}_{\boldsymbol{\sigma}} \left[\sup_{h \in H} \sum_{i=1}^m \sigma_i \mathbb{1}_{h(\mathbf{x}_i) > b_i^{(1)}} \right] \leq M \sqrt{\frac{2d' \log \frac{em}{d'}}{m}},$$

where $d' = \text{VCdim}(\{(\mathbf{x}, \mathbf{b}) \mapsto \mathbb{1}_{h(\mathbf{x}) - b^{(1)} > 0} \colon (\mathbf{x}, \mathbf{b}) \in \mathcal{X} \times \mathcal{B}\})$. Since the second bid component $b^{(2)}$ plays no role in this definition, d' coincides with $\text{VCdim}(\{(\mathbf{x}, b^{(1)}) \mapsto \mathbb{1}_{h(\mathbf{x}) - b^{(1)} > 0} \colon (\mathbf{x}, b^{(1)}) \in \mathcal{X} \times \mathcal{B}_1\})$, where $\mathcal{B}_1$ is the projection of $\mathcal{B} \subseteq \mathbb{R}^2$ onto its first component, and is upper-bounded by $\text{VCdim}(\{(\mathbf{x}, t) \mapsto \mathbb{1}_{h(\mathbf{x}) - t > 0} \colon (\mathbf{x}, t) \in \mathcal{X} \times \mathbb{R}\})$, that is, the pseudo-dimension of H. ■

Propositions 1 and 2 can be used to derive the following generalization bound for the learning problem we consider.

Theorem 3 *For any $\delta > 0$, with probability at least $1 - \delta$ over the draw of an i.i.d. sample S of size m, the following inequality holds for all $h \in H$:*

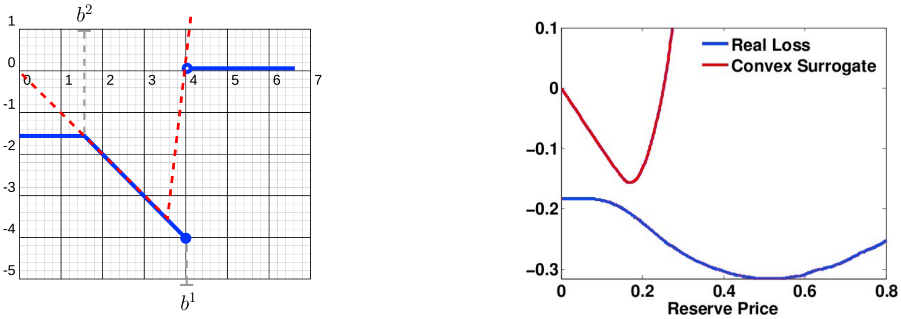
$$\mathcal{L}(h) \leq \widehat{\mathcal{L}}_S(h) + 2\mathfrak{R}_m(H) + 2M \sqrt{\frac{2d \log \frac{em}{d}}{m}} + M \sqrt{\frac{\log \frac{1}{\delta}}{2m}}.$$

Proof By a standard property of the Rademacher complexity, since $L = l_1 + l_2$, the following inequality holds: $\mathfrak{R}_m(L_H) \leq \mathfrak{R}_m(l_{1H}) + \mathfrak{R}_m(l_{2H})$. Thus, in view of Propositions 1 and 2, the Rademacher complexity of L_H can be bounded via

$$\mathfrak{R}_m(L_H) \leq \mathfrak{R}_m(H) + M \sqrt{\frac{2d \log \frac{em}{d}}{m}}.$$

The result then follows by the application of a standard Rademacher complexity bound (Koltchinskii and Panchenko, 2002). ■

This learning bound invites us to consider an algorithm seeking $h \in H$ to minimize the empirical loss $\widehat{\mathcal{L}}_S(h)$, while controlling the complexity (Rademacher complexity and pseudo-dimension) of the hypothesis set H. In the following section, we discuss the computational problem of minimizing the empirical loss and suggest the use of a surrogate loss leading to a more tractable problem.



(a) (b)

Figure 2: (a) Piecewise linear convex surrogate loss L_p. (b) Comparison of the sum of real losses $\sum_{i=1}^m L(\cdot, \mathbf{b}_i)$ for $m = 500$ with the sum of convex surrogate losses. Note that the minimizers are significantly different.

4.2 Surrogate Loss

As pointed out in Section 4, the loss function L does not admit most properties of traditional loss functions used in machine learning: for any fixed $\mathbf{b}$, $L(\cdot, \mathbf{b})$ is not differentiable (at two points), it is not convex nor Lipschitz, and in fact it is discontinuous. For any fixed $\mathbf{b}$, $L(\cdot, \mathbf{b})$ is quasi-convex,[1] a property that is often desirable since there exist several solutions for quasi-convex optimization problems. However, in general, a sum of quasi-convex functions, such as the sum $\sum_{i=1}^m L(\cdot, \mathbf{b}_i)$ appearing in the definition of the empirical loss, is not quasi-convex and a fortiori not convex.[2] In fact, in general, such a sum may admit exponentially many local minima. This leads us to seek a surrogate loss function with more favorable optimization properties.

A standard method in machine learning consists of replacing the loss function L with a convex upper bound (Bartlett et al., 2006). A natural candidate in our case is the piecewise linear function L_p shown in Figure 2(a). While this is a convex loss function, and thus convenient for optimization, it is not calibrated. That is, it is possible for $r_p \in \operatorname{argmin} \mathbb{E}_{\mathbf{b}}[L_p(r, \mathbf{b})]$ to have a large expected true loss. Therefore, it does not provide us with a useful surrogate. The calibration problem is illustrated by Figure 2(b) in dimension one, where the true objective function to be minimized $\sum_{i=1}^m L(r, \mathbf{b}_i)$ is compared with the sum of the surrogate losses. The next theorem shows that this problem in fact affects *any* non-constant convex surrogate. It is expressed in terms of the loss $\widetilde{L} \colon \mathbb{R} \times \mathbb{R}_+ \to \mathbb{R}$ defined by $\widetilde{L}(r, b) = -r \mathbb{1}_{r \leq b}$, which coincides with L when the second bid is 0.

Definition 4 *We say that a function* $L_c \colon [0, M] \times [0, M] \to \mathbb{R}$ *is* consistent with $\widetilde{L}$ *if, for any distribution D, there exists a minimizer* $r^* \in \operatorname{argmin}_r \mathbb{E}_{b \sim D}[L_c(r, b)]$ *such that* $r^* \in \operatorname{argmin}_r \mathbb{E}_{b \sim D}[\widetilde{L}(r, b)]$.

Definition 5 *We say that a sequence of functions* $(L_n)_{n \in \mathbb{N}}$ *mapping* $[0, M] \times [0, M]$ *to* $\mathbb{R}$ *is* weakly consistent with $\widetilde{L}$ *if there exists a sequence* $(r_n)_{n \in \mathbb{N}}$ *in* $\mathbb{R}$ *with* $r_n \in \operatorname{argmin}_r \mathbb{E}_{b \sim D}[L_n(r, b)]$ *for all* $n \in \mathbb{N}$ *such that* $\lim_{n \to +\infty} r_n = r^*$ *with* $r^* \in \operatorname{argmin} \mathbb{E}_{b \sim D}[\widetilde{L}(r, b)]$.

Proposition 6 (Convex surrogates) *Let* $L_c \colon [0, M] \times [0, M] \to \mathbb{R}$ *be a bounded function, convex with respect to its first argument. If L_c is consistent with $\widetilde{L}$, then $L_c(\cdot, b)$ is constant for any* $b \in [0, M]$.

Proof The idea behind the proof is the following: for any two bids $b_1 < b_2$, there exists a distribution D with support $\{b_1, b_2\}$ such that $\mathbb{E}_{b \sim D}[\widetilde{L}(r, b)]$ is minimized at both $r = b_1$ and $r = b_2$. We show this implies that $\mathbb{E}_{b \sim D}[L_c(r, b)]$ must attain a minimum at both points too. By convexity of L_c it follows that $\mathbb{E}_{b \sim D}[L_c(r, b)]$ must be constant on the interval $[b_1, b_2]$. The main part of the proof will be showing that this implies that the function $L_c(\cdot, b_1)$ must also be constant on the interval $[b_1, b_2]$. Finally, since the value of b_2 was chosen arbitrarily, it will follow that $L_c(\cdot, b_1)$ is constant.

[1] A function $f : \mathbb{R} \to \mathbb{R}$ is said to be *quasi-convex* if for any $\alpha \in \mathbb{R}$ the sub-level set $\{x \colon f(x) \leq \alpha\}$ is convex.

[2] It is known that, under some separability condition, if a finite sum of quasi-convex functions on an open convex set is quasi-convex, then all but perhaps one of them is convex (Debreu and Koopmans, 1982).

Let $0 < b_1 < b_2 < M$ and, for any $\mu \in [0, 1]$, let D_μ denote the probability distribution with support included in $\{b_1, b_2\}$ defined by $D_\mu(b_1) = \mu$ and let $\mathbb{E}_\mu$ denote the expectation with respect to this distribution. A straightforward calculation shows that the unique minimizer of $\mathbb{E}_\mu[\widetilde{L}(r, b)]$ is given by b_2 if $\mu > \frac{b_2 - b_1}{b_2}$ and by b_1 if $\mu < \frac{b_2 - b_1}{b_2}$. Therefore, if $F_\mu(r) = \mathbb{E}_\mu[L_c(r, b)]$, it must be the case that b_2 is a minimizer of F_μ for $\mu > \frac{b_2 - b_1}{b_2}$ and b_1 is a minimizer of F_μ for $\mu < \frac{b_2 - b_1}{b_2}$.

For a convex function $f \colon \mathbb{R} \to \mathbb{R}$, we denote by f^- its left-derivative and by f^+ its right-derivative, which are guaranteed to exist. We will also denote here, for any $b \in \mathbb{R}$, by $g^-(\cdot, b)$ and $g^+(\cdot, b)$ the left- and right-derivatives of the function $g(\cdot, b)$ and by $g'(\cdot, b)$ its derivative, when it exists. Recall that for a convex function f, if x_0 is a minimizer, then $f^-(x_0) \leq 0 \leq f^+(x_0)$. In view of that and the minimizing properties of b_1 and b_2, the following inequalities hold:

$$0 \geq F_\mu^-(b_2) = \mu L_c^-(b_2, b_1) + (1 - \mu) L_c^-(b_2, b_2) \qquad \text{for } \mu > \frac{b_2 - b_1}{b_2}, \tag{3}$$

$$0 \leq F_\mu^+(b_1) \leq F_\mu^-(b_2) \qquad \text{for } \mu < \frac{b_2 - b_1}{b_2}, \tag{4}$$

where the second inequality in (4) holds by convexity of F_μ and the fact that $b_1 < b_2$. By setting $\mu = \frac{b_2 - b_1}{b_2}$, it follows from inequalities (3) and (4) that $F_\mu^-(b_2) = 0$ and $F_\mu^+(b_1) = 0$. By convexity of F_μ, it follows that F_μ is constant on the interval (b_1, b_2). We now show this may only happen if $L_c(\cdot, b_1)$ is also constant. By rearranging terms in (3) and plugging in the expression of μ, we obtain the equivalent condition

$$(b_2 - b_1) L_c^-(b_2, b_1) = -b_1 L_c^-(b_2, b_2).$$

Since L_c is a bounded function, it follows that $L_c^-(b_2, b_1)$ is bounded for any $b_1, b_2 \in (0, M)$, therefore as $b_1 \to b_2$ we must have $b_2 L_c^-(b_2, b_2) = 0$, which implies $L_c^-(b_2, b_2) = 0$ for all $b_2 > 0$. In view of this, inequality (3) may only be satisfied if $L_c^-(b_2, b_1) \leq 0$. However, the convexity of L_c implies $L_c^-(b_2, b_1) \geq L_c^-(b_1, b_1) = 0$. Therefore, $L_c^-(b_2, b_1) = 0$ must hold for all $b_2 > b_1 > 0$. Similarly, by definition of F_μ, the first inequality in (4) implies

$$\mu L_c^+(b_1, b_1) + (1 - \mu) L_c^+(b_1, b_2) \geq 0. \tag{5}$$

Nevertheless, for any $b_2 > b_1$ we have $0 = L_c^-(b_1, b_1) \leq L_c^+(b_1, b_1) \leq L_c^-(b_2, b_1) = 0$. Consequently, $L_c^+(b_1, b_1) = 0$ for all $b_1 > 0$. Furthermore, $L_c^+(b_1, b_2) \leq L_c^+(b_2, b_2) = 0$. Therefore, for inequality (5) to be satisfied, we must have $L_c^+(b_1, b_2) = 0$ for all $b_1 < b_2$.

Thus far, we have shown that for any $b > 0$, if $r \geq b$, then $L_c^-(r, b) = 0$, while $L_c^+(r, b) = 0$ for $r \leq b$. A simple convexity argument shows that $L_c(\cdot, b)$ is then differentiable and $L_c'(r, b) = 0$ for all $r \in (0, M)$, which in turn implies that $L_c(\cdot, b)$ is a constant function. ■

The result of the previous proposition can be considerably strengthened, as shown by the following theorem. As in the proof of the previous proposition, to simplify the notation, for any $b \in \mathbb{R}$, we will denote by $g'(\cdot, b)$ the derivative of a differentiable function $g(\cdot, b)$.

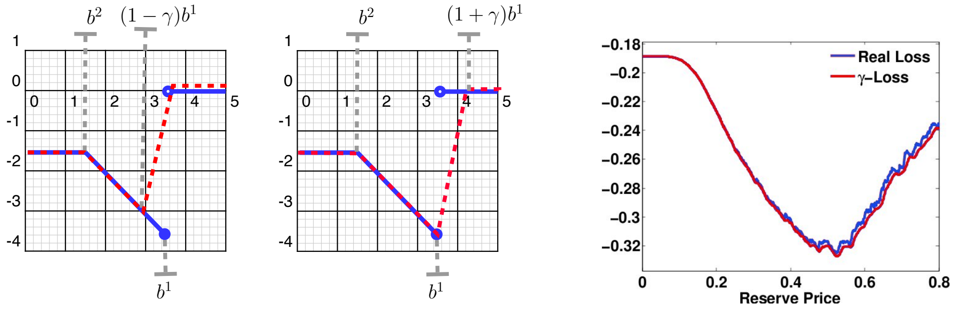
Theorem 7 *Let $(L_n)_{n \in \mathbb{N}}$ denote a sequence of functions mapping $[0, M] \times [0, M]$ to $\mathbb{R}$ that are convex and differentiable with respect to their first argument and satisfy the following conditions:*

- $\sup_{b \in [0,M], n \in \mathbb{N}} \max(|L_n'(0, b)|, |L_n'(M, b)| = K < \infty$;
- *$(L_n)_n$ is weakly consistent with $\widetilde{L}$;*
- *$L_n(0, b) = 0$ for all $n \in \mathbb{N}$ and for all b.*

If the sequence $(L_n)_n$ converges pointwise to a function L_c, then $L_n(\cdot, b)$ converges uniformly to $L_c(\cdot, b) \equiv 0$.

We defer the proof of this theorem to Appendix B and present here only a sketch of the proof. We first show that the convexity of the functions L_n implies that the convergence to L_c must be uniform and that L_c is convex with respect to its first argument. This fact and the weak consistency of the sequence L_n will then imply that L_c is consistent with $\widetilde{L}$ and therefore must be constant by Proposition 6.

The theorem just presented shows that even a weakly consistent sequence of convex losses is uniformly close to a constant function and therefore not helpful to tackle the learning task we consider. This suggests searching for surrogate losses that admit weaker regularity assumptions such as Lipschitz continuity.



(a) (b)

Figure 3: (a) Comparison of the true loss L with surrogate loss $\overline{L}_\gamma$ on the left and surrogate loss L_γ on the right, for $\gamma = 0.1$. (b) Comparison of $\sum_{i=1}^{500} L(r, \mathbf{b}_i)$ and $\sum_{i=1}^{500} L_\gamma(r, \mathbf{b}_i)$

Perhaps, the most natural surrogate loss function is then $\overline{L}_\gamma$, an upper bound on L defined for all $\gamma > 0$ by:

$$\overline{L}_\gamma(r, \mathbf{b}) = -b^{(2)}\mathbb{1}_{r \le b^{(2)}} - r\mathbb{1}_{b^{(2)} < r \le \left((1-\gamma)b^{(1)}\right) \vee b^{(2)}} + \left(\frac{1-\gamma}{\gamma} \vee \frac{b^{(2)}}{b^{(1)} - b^{(2)}}\right)(r - b^{(1)})\mathbb{1}_{\left((1-\gamma)b^{(1)}\right) \vee b^{(2)} < r \le b^{(1)}},$$

where $c \vee d = \max(c, d)$. The plot of this function is shown in Figure 3(a). The $\max$ terms ensure that the function is well defined if $(1-\gamma)b^{(1)} < b^{(2)}$. However, this turns out to be also a poor choice as $\overline{L}_\gamma$ is a loose upper bound on L in the most critical region, that is around the minimum of the loss L. Thus, instead, we will consider, for any $\gamma > 0$, the loss function L_γ defined as follows:

$$L_\gamma(r, \mathbf{b}) = -b^{(2)}\mathbb{1}_{r \le b^{(2)}} - r\mathbb{1}_{b^{(2)} < r \le b^{(1)}} + \frac{1}{\gamma}(r - (1+\gamma)b^{(1)})\mathbb{1}_{b^{(1)} < r \le (1+\gamma)b^{(1)}}, \tag{6}$$

whose plot is shown in Figure 3(a).[3] A comparison between the sum of L-losses and the sum of L_γ-losses is shown in Figure 3(b). Observe that the fit is considerably better than that of the piecewise linear convex surrogate loss shown in Figure 2(b). A possible concern associated with the loss function L_γ is that it is a lower bound for L. One might think then that minimizing it would not lead to an informative solution. However, we argue that this problem arises significantly with upper bounding losses such as the convex surrogate, which we showed not to lead to a useful minimizer, or $\overline{L}_\gamma$, which is a poor approximation of L near its minimum. By matching the original loss L in the region of interest, around the minimal value, the loss function L_γ leads to more informative solutions for this problem. In fact, we show that that the expected loss $\mathcal{L}_\gamma(h) \colon = \mathbb{E}_{\mathbf{x},\mathbf{b}}[L_\gamma(h)]$ admits a minimizer close to the minimizer of $\mathcal{L}(h)$. Since $L_\gamma \to L$ as $\gamma \to 0$, this result may seem trivial. However, this convergence is not uniform and therefore calibration is not guaranteed.

Theorem 8 *Let H be a closed, convex subset of a linear space of functions containing 0. Then, the following inequality holds for all $\gamma \ge 0$:*

$$\mathcal{L}(h^*_\gamma) - \mathcal{L}_\gamma(h^*_\gamma) \le \gamma M.$$

Notice that, since $L \ge L_\gamma$ for all $\gamma \ge 0$, the theorem implies that $\lim_{\gamma\to 0} \mathcal{L}(h^*_\gamma) = \mathcal{L}(h^*)$. Indeed, let h^* denote the best-in-class hypothesis for the loss function L. Then, the following straightforward inequalities hold:

$$\begin{aligned}\mathcal{L}(h^*) &\le \mathcal{L}(h^*_\gamma) \\ &\le \mathcal{L}_\gamma(h^*_\gamma) + \gamma M \\ &\le \mathcal{L}_\gamma(h^*) + \gamma M \le \mathcal{L}(h^*) + \gamma M.\end{aligned}$$

[3] Technically, the theoretical and algorithmic results we present for L_γ could be developed in a somewhat similar way for $\overline{L}_\gamma$.

By letting $\gamma \to 0$ we see that $\mathcal{L}(h^*_\gamma) \to \mathcal{L}(h^*)$. This is a remarkable result as it not only provides a convergence guarantee but it also gives us an explicit rate of convergence. We will later exploit this fact to come up with an optimal choice for γ.

The proof of Theorem 8 is based on the following partitioning of $\mathcal{X} \times \mathcal{B}$ in four regions where L_γ is defined as an affine function:

$$I_1 = \{(\mathbf{x}, \mathbf{b}) | h^*_\gamma(\mathbf{x}) \leq b^{(2)}\} \qquad I_2 = \{(\mathbf{x}, \mathbf{b}) | h^*_\gamma(\mathbf{x}) \in (b^{(2)}, b^{(1)}]\}$$
$$I_3 = \{(\mathbf{x}, \mathbf{b}) | h^*_\gamma(\mathbf{x}) \in (b^{(1)}, (1+\gamma) b^{(1)}]\} \qquad I_4 = \{(\mathbf{x}, \mathbf{b}) | h^*_\gamma(\mathbf{x}) > (1+\gamma) b^{(1)}\},$$

Notice that L_γ and L differ only on I_3. Therefore, we only need to bound the measure of this set which can be done as in Lemma 15 (see Appendix C).

Proof [Theorem 8]. We can express the difference as

$$\begin{aligned}
\mathop{\mathbb{E}}_{\mathbf{x},\mathbf{b}} \Big[L(h^*_\gamma(\mathbf{x}), \mathbf{b}) - L_\gamma(h^*_\gamma(\mathbf{x}), \mathbf{b})\Big] &= \sum_{k=1}^{4} \mathop{\mathbb{E}}_{\mathbf{x},\mathbf{b}} \Big[(L(h^*_\gamma(\mathbf{x}), \mathbf{b}) - L_\gamma(h^*_\gamma(\mathbf{x}), \mathbf{b}))\mathbb{1}_{I_k}(\mathbf{x}, \mathbf{b})\Big] \\
&= \mathop{\mathbb{E}}_{\mathbf{x},\mathbf{b}} \Big[(L(h^*_\gamma(\mathbf{x}), \mathbf{b}) - L_\gamma(h^*_\gamma(\mathbf{x}), \mathbf{b}))\mathbb{1}_{I_3}(\mathbf{x}, \mathbf{b})\Big] \\
&= \mathop{\mathbb{E}}_{\mathbf{x},\mathbf{b}} \Big[\frac{1}{\gamma}((1+\gamma) b^{(1)} - h^*_\gamma(\mathbf{x}))\mathbb{1}_{I_3}(\mathbf{x}, \mathbf{b})\Big]. \qquad (7)
\end{aligned}$$

Furthermore, for $(\mathbf{x}, \mathbf{b}) \in I_3$, we know that $b^{(1)} < h^*_\gamma(\mathbf{x})$. Thus, we can bound (7) by $\mathbb{E}_{\mathbf{x},\mathbf{b}}[h^*_\gamma(\mathbf{x})\mathbb{1}_{I_3}(\mathbf{x}, \mathbf{b})]$, which, by Lemma 15 in Appendix C, is upper bounded by $\gamma\, \mathbb{E}_{\mathbf{x},\mathbf{b}} \Big[h^*_\gamma(\mathbf{x})\mathbb{1}_{I_2}(\mathbf{x}, \mathbf{b})\Big]$. Thus, the following inequalities hold:

$$\mathop{\mathbb{E}}_{\mathbf{x},\mathbf{b}} \Big[L(h^*_\gamma(\mathbf{x}), \mathbf{b})\Big] - \mathop{\mathbb{E}}_{\mathbf{x},\mathbf{b}} \Big[L_\gamma(h^*_\gamma(\mathbf{x}), \mathbf{b})\Big] \leq \gamma \mathop{\mathbb{E}}_{\mathbf{x},\mathbf{b}} \Big[h^*_\gamma(\mathbf{x})\mathbb{1}_{I_2}(\mathbf{x}, \mathbf{b})\Big] \leq \gamma \mathop{\mathbb{E}}_{\mathbf{x},\mathbf{b}} \Big[b^{(1)}\mathbb{1}_{I_2}(\mathbf{x}, \mathbf{b})\Big] \leq \gamma M,$$

using the fact that $h^*_\gamma(\mathbf{x}) \leq b^{(1)}$ for $(\mathbf{x}, \mathbf{b}) \in I_2$. ■

The $1/\gamma$-Lipschitzness of L_γ can be used to prove the following generalization bound.

Theorem 9 *Fix $\gamma \in (0, 1]$ and let S denote a sample of size m. Then, for any $\delta > 0$, with probability at least $1 - \delta$ over the choice of the sample S, for all $h \in H$, the following holds:*

$$\mathcal{L}_\gamma(h) \leq \widehat{\mathcal{L}}_\gamma(h) + \frac{2}{\gamma}\mathfrak{R}_m(H) + M\sqrt{\frac{\log \frac{1}{\delta}}{2m}}. \qquad (8)$$

Proof Let $L_{\gamma,H}$ denote the family of functions $\{(\mathbf{x}, \mathbf{b}) \to L_\gamma(h(\mathbf{x}), b) \colon h \in H\}$. The loss function L_γ is $\frac{1}{\gamma}$-Lipschitz since the slope of the lines defining it is at most $\frac{1}{\gamma}$. Thus, using the contraction lemma (Lemma 14) as in the proof of Proposition 1, gives $\mathfrak{R}_m(L_{\gamma,H}) \leq \frac{1}{\gamma}\mathfrak{R}_m(H)$. The application of a standard Rademacher complexity bound to the family of functions $L_{\gamma,H}$ then shows that for any $\delta > 0$, with probability at least $1 - \delta$, for any $h \in H$, the following holds:

$$\mathcal{L}_\gamma(h) \leq \widehat{\mathcal{L}}_\gamma(h) + \frac{2}{\gamma}\mathfrak{R}_m(H) + M\sqrt{\frac{\log \frac{1}{\delta}}{2m}}.$$

■

We conclude this section by showing that L_γ admits a stronger form of consistency. More precisely, we prove that the generalization error of the best-in-class hypothesis $\mathcal{L}^* := \mathcal{L}(h^*)$ can be lower bounded in terms of that of the empirical minimizer of L_γ, $\widehat{h}_\gamma\colon = \operatorname{argmin}_{h\in H} \widehat{\mathcal{L}}_\gamma(h)$.

Theorem 10 *Let $M = \sup_{b\in\mathcal{B}} b^{(1)}$ and let H be a hypothesis set with pseudo-dimension $d = Pdim(H)$. Then, for any $\delta > 0$ and a fixed value of $\gamma > 0$, with probability at least $1 - \delta$ over the choice of a sample S of size m, the following inequality holds:*

$$\mathcal{L}^* \leq \mathcal{L}(\widehat{h}_\gamma) \leq \mathcal{L}^* + \frac{2\gamma + 2}{\gamma}\mathfrak{R}_m(H) + \gamma M + 2M\sqrt{\frac{2d \log \frac{\epsilon m}{d}}{m}} + 2M\sqrt{\frac{\log \frac{2}{\delta}}{2m}}.$$

Proof By Theorem 3, with probability at least $1 - \delta/2$, the following holds:

$$\mathcal{L}(\widehat{h}_\gamma) \leq \widehat{\mathcal{L}}_S(\widehat{h}_\gamma) + 2\mathfrak{R}_m(H) + 2M\sqrt{\frac{2d\log\frac{\epsilon m}{d}}{m}} + M\sqrt{\frac{\log\frac{2}{\delta}}{2m}}. \tag{9}$$

Applying Lemma 15 with the empirical distribution induced by the sample, we can bound $\widehat{\mathcal{L}}_S(\widehat{h}_\gamma)$ by $\widehat{\mathcal{L}}_\gamma(\widehat{h}_\gamma) + \gamma M$. The first term of the previous expression is less than $\widehat{\mathcal{L}}_\gamma(h^*_\gamma)$ by definition of $\widehat{h}_\gamma$. Moreover, the same analysis used in the proof of Theorem 9 shows that with probability $1 - \delta/2$,

$$\widehat{\mathcal{L}}_\gamma(h^*_\gamma) \leq \mathcal{L}_\gamma(h^*_\gamma) + \frac{2}{\gamma}\mathfrak{R}_m(H) + M\sqrt{\frac{\log\frac{2}{\delta}}{2m}}.$$

Finally, by definition of h^*_γ and using the fact that L is an upper bound on L_γ, we can write $\mathcal{L}_\gamma(h^*_\gamma) \leq \mathcal{L}_\gamma(h^*) \leq \mathcal{L}(h^*)$. Thus,

$$\widehat{\mathcal{L}}_S(\widehat{h}_\gamma) \leq \mathcal{L}(h^*) + \frac{2}{\gamma}\mathfrak{R}_m(H) + M\sqrt{\frac{\log\frac{2}{\delta}}{2m}} + \gamma M.$$

Replacing this inequality in (9) and applying the union bound yields the result. ■

This bound can be extended to hold uniformly over all γ at the price of a term in $O\Big(\frac{\sqrt{\log\log_2\frac{1}{\gamma}}}{\sqrt{m}}\Big)$. Thus, for appropriate choices of γ as a function of m (for instance $\gamma = 1/m^{1/4}$) we can guarantee the convergence of $\mathcal{L}(\widehat{h}_\gamma)$ to $\mathcal{L}^*$, a stronger form of consistency (See Appendix C).

These results are reminiscent of the standard margin bounds with γ playing the role of a margin. The situation here is however somewhat different. Our learning bounds suggest, for a fixed $\gamma \in (0, 1]$, to seek a hypothesis h minimizing the empirical loss $\widehat{\mathcal{L}}_\gamma(h)$ while controlling a complexity term upper bounding $\mathfrak{R}_m(H)$, which in the case of a family of linear hypotheses could be $\|h\|^2_K$ for some PSD kernel K. Since the bound can hold uniformly for all γ, we can use it to select γ out of a finite set of possible grid search values. Alternatively, γ can be set via cross-validation. In the next section, we present algorithms for solving this regularized empirical risk minimization problem.

5. Algorithms

In this section, we show how to minimize the empirical risk under two regimes: first we analyze the no-feature scenario considered in Cesa-Bianchi et al. (2013) and then we present an algorithm to solve the more general feature-based revenue optimization problem.

5.1 No-Feature Case

We now present a general algorithm to optimize sums of functions similar to L_γ or L in the one-dimensional case.

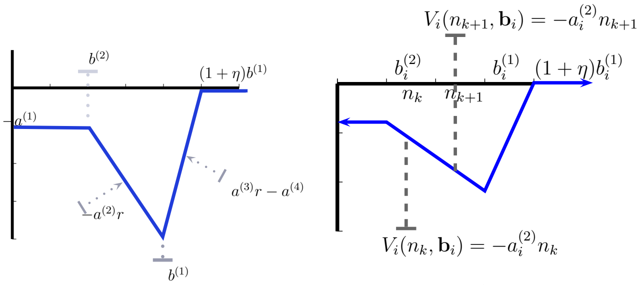
Definition 11 *We will say that function $V : \mathbb{R} \times \mathcal{B} \to \mathbb{R}$ is a* v-function *if it admits the following form:*

$$V(r, \mathbf{b}) = -a^{(1)}\mathbb{1}_{r \leq b^{(2)}} - a^{(2)}r\mathbb{1}_{b^{(2)} < r \leq b^{(1)}} + (a^{(3)}r - a^{(4)})\mathbb{1}_{b^{(1)} < r < (1+\eta)b^{(1)}},$$

with $a^{(1)} > 0$ and $\eta > 0$ constants and $a^{(2)}, a^{(3)}, a^{(4)}$ defined by $a^{(1)} = \eta a^{(3)} b^{(2)}$, $a^{(2)} = \eta a^{(3)}$, and $a^{(4)} = a^{(3)}(1+\eta)b^{(1)}$.

Figure 4(a) illustrates this family of loss functions. A v-function is a generalization of L_γ and L. Indeed, any v-function V satisfies $V(r, \mathbf{b}) \leq 0$ and attains its minimum at $b^{(1)}$. Finally, as can be seen straightforwardly from Figure 3, L_γ is a v-function for any $\gamma > 0$. We consider the following general problem of minimizing a sum of v-functions:

$$\min_{r \geq 0} F(r) := \sum_{i=1}^{m} V_i(r, \mathbf{b}_i). \tag{10}$$



(a)

(b)

Figure 4: (a) Prototypical v-function. (b) Illustration of the fact that the definition of $V_i(r, \mathbf{b}_i)$ does not change on an interval $[n_k, n_{k+1}]$.

Observe that this is not a trivial problem since, for any fixed $\mathbf{b}_i$, $V_i(\cdot, \mathbf{b}_i)$ is non-convex and that, in general, a sum of m such functions may admit many local minima. Of course, we can seek a solution that is ϵ-close to the optimal reserve via a grid search over points $r_i = i\epsilon$. However, the guarantees for that algorithm would depend on the continuity of the function. In particular, this algorithm might fail for the loss L. Instead, we exploit the particular structure of a v-function to exactly minimize F. The following proposition, which is proven in Appendix D, shows that the minimum is attained at one of the highest bids, which matches the intuition. Notice that for the loss function L this is immediate since if r is not a highest bid, one can raise the reserve price without increasing any of the component losses.

Proposition 12 *Problem (10) admits a solution r^* that satisfies $r^* = b_i^{(1)}$ for some $i \in [1, m]$.*

Problem (10) can thus be reduced to examining the value of the function for the m arguments $b_i^{(1)}$, $i \in [1, m]$. This yields a straightforward method for solving the optimization which consists of computing $F(b_i^{(1)})$ for all i and taking the minimum. But, since the computation of each $F(b_i^{(1)})$ takes $O(m)$, the overall computational cost is in $O(m^2)$, which can be prohibitive for even moderately large values of m.

Instead, we present a combinatorial algorithm to solve the optimization problem (10) in $O(m \log m)$. Let $\mathcal{N} = \bigcup_i \{b_i^{(1)}, b_i^{(2)}, (1+\eta)b_i^{(1)}\}$ denote the set of all *boundary points* associated with the functions $V(\cdot, \mathbf{b}_i)$. The algorithm proceeds as follows: first, sort the set $\mathcal{N}$ to obtain the ordered sequence $(n_1, \ldots, n_{3m})$, which can be achieved in $O(m \log m)$ using a comparison-based sorting algorithm. Next, evaluate $F(n_1)$ and compute $F(n_{k+1})$ from $F(n_k)$ for all k.

The main idea of the algorithm is the following: since the definition of $V_i(\cdot, b_i)$ can only change at boundary points (see also Figure 4(b)), computing $F(n_{k+1})$ from $F(n_k)$ can be achieved in constant time. Indeed, since between n_k and n_{k+1} there are only two boundary points, we can compute $V(n_{k+1}, \mathbf{b}_i)$ from $V(n_k, \mathbf{b}_i)$ by calculating V for only two values of $\mathbf{b}_i$, which can be done in constant time. We now give a more detailed description and proof of correctness of our algorithm.

Proposition 13 *There exists an algorithm to solve the optimization problem (10) in $O(m \log m)$.*

Proof The pseudocode of the algorithm is given in Algorithm 1, where $a_i^{(1)}, ..., a_i^{(4)}$ denote the parameters defining the functions $V_i(r, \mathbf{b}_i)$. We will prove that, after running Algorithm 1, we can compute $F(n_j)$ in constant time using:

$$F(n_j) = c_j^{(1)} + c_j^{(2)} n_j + c_j^{(3)} n_j + c_j^{(4)}. \tag{11}$$

This holds trivially for n_1 since by definition $n_1 \leq b_i^{(2)}$ for all i and therefore $F(n_1) = -\sum_{i=1}^m a_i^{(1)}$. Now, assume that (11) holds for j, we prove that it must also hold for $j+1$. Suppose $n_j = b_i^{(2)}$ for some i (the cases $n_j = b_i^{(1)}$ and $n_j = (1+\eta)b_i^{(1)}$ can be handled in the same way). Then $V_i(n_j, \mathbf{b}_i) = -a_i^{(1)}$ and we can write

$$\sum_{k \neq i} V_k(n_j, \mathbf{b}_k) = F(n_j) - V(n_j, \mathbf{b}_i) = (c_j^{(1)} + c_j^{(2)} n_j + c_j^{(3)} n_j + c_j^{(4)}) + a_i^{(1)}.$$

Algorithm 1 Sorting

```
$\mathcal{N} := \bigcup_{i=1}^m \{b_i^{(1)}, b_i^{(2)}, (1+\eta)b_i^{(1)}\}$;
$n_1, ..., n_{3m}) = \mathbf{Sort}(\mathcal{N})$;
Set $\mathbf{c}_i := (c_i^{(1)}, c_i^{(2)}, c_i^{(3)}, c_i^{(4)}) = 0$ for $i = 1, ..., 3m$;
Set $c_1^{(1)} = -\sum_{i=1}^m a_i^{(1)}$;
for $j = 2, ..., 3m$ do
  Set $\mathbf{c}_j = \mathbf{c}_{j-1}$;
  if $n_{j-1} = b_i^{(2)}$ for some $i$ then
    $c_j^{(1)} = c_j^{(1)} + a_i^{(1)}$;
    $c_j^{(2)} = c_j^{(2)} - a_i^{(2)}$;
  else if $n_{j-1} = b_i^{(1)}$ for some $i$ then
    $c_j^{(2)} = c_j^{(2)} + a_i^{(2)}$;
    $c_j^{(3)} = c_j^{(3)} + a_i^{(3)}$;
    $c_j^{(4)} = c_j^{(4)} - a_i^{(4)}$;
  else
    $c_j^{(3)} = c_j^{(3)} - a_i^{(3)}$;
    $c_j^{(4)} = c_j^{(4)} + a_i^{(4)}$;
  end if
end for
```

Thus, by construction we would have:

$$\begin{aligned} c_{j+1}^{(1)} + c_{j+1}^{(2)} n_{j+1} + c_{j+1}^{(3)} n_{j+1} + c_{j+1}^{(4)} &= c_j^{(1)} + a_i^{(1)} + (c_j^{(2)} - a_i^{(2)}) n_{j+1} + c_j^{(3)} n_{j+1} + c_j^{(4)} \\ &= (c_j^{(1)} + c_j^{(2)} n_{j+1} + c_j^{(3)} n_{j+1} + c_j^{(4)}) + a_i^{(1)} - a_i^{(2)} n_{j+1} \\ &= \sum_{k \neq i} V_k(n_{j+1}, \mathbf{b}_k) - a_i^{(2)} n_{j+1}, \end{aligned}$$

where the last equality holds since the definition of $V_k(r, \mathbf{b}_k)$ does not change for $r \in [n_j, n_{j+1}]$ and $k \neq i$. Finally, since n_j was a boundary point, the definition of $V_i(r, \mathbf{b}_i)$ must change from $-a_i^{(1)}$ to $-a_i^{(2)} r$, thus the last equation is indeed equal to $F(n_{j+1})$. A similar argument can be given if $n_j = b_i^{(1)}$ or $n_j = (1+\eta) b_i^{(1)}$.

We proceed to analyze the complexity of the algorithm: sorting the set $\mathcal{N}$ can be performed in $O(m \log m)$ and each iteration takes only constant time. Thus, the evaluation of all points can be achieved in linear time and, clearly, the minimum can then also be obtained in linear time. Therefore, the overall time complexity of the algorithm is in $O(m \log m)$. ■

The algorithm just proposed can be straightforwardly extended to solve the minimization of F over a set of r-values bounded by Λ, that is $\{r \colon 0 \leq r \leq \Lambda\}$. Indeed, we need only compute $F(b_i^{(1)})$ for $i \in [1, m]$ such that $b_i^{(1)} < \Lambda$ and of course also $F(\Lambda)$, thus the computational complexity in this regularized case remains in $O(m \log m)$.

5.2 General Case

We now present our main algorithm for revenue optimization in the presence of features. This problem presents new challenges characteristic of non-convex optimization problems in higher dimensions. Therefore, our proposed algorithm can only guarantee convergence to a local minimum. Nevertheless, we provide a simple method for cycling through these local minima with the guarantee of reducing the objective function at each time.

We consider the case of a hypothesis set $H \subset \mathbb{R}^N$ of linear functions $\mathbf{x} \mapsto \mathbf{w} \cdot \mathbf{x}$ with bounded norm, $\|\mathbf{w}\| \leq \Lambda$, for some $\Lambda \geq 0$. This can be immediately generalized to non-linear hypotheses by using a positive definite kernel.

The results of Theorem 9 suggest seeking, for a fixed $\gamma \geq 0$, the vector $\mathbf{w}$ solution to the following optimization problem: $\min_{\|\mathbf{w}\| \leq \Lambda} \sum_{i=1}^{m} L_\gamma(\mathbf{w} \cdot \mathbf{x}_i, \mathbf{b}_i)$. Replacing the original loss L with L_γ helped us remove the discontinuity of the loss. But, we still face an optimization problem based on a sum of non-convex functions. This problem can be formulated as a DC-programming (difference of convex functions programming) problem which is a well studied problem in non-convex optimization. Indeed, L_γ can be decomposed as follows for all $(r, \mathbf{b}) \in \mathbb{R} \times \mathcal{B}$: $L_\gamma(r, \mathbf{b}) = u(r, \mathbf{b}) - v(r, \mathbf{b})$, with the convex functions u and v defined by

$$u(r, \mathbf{b}) = -r \mathbb{1}_{r < b^{(1)}} + \tfrac{r - (1+\gamma) b^{(1)}}{\gamma} \mathbb{1}_{r \geq b^{(1)}}$$
$$v(r, \mathbf{b}) = (-r + b^{(2)}) \mathbb{1}_{r < b^{(2)}} + \tfrac{r - (1+\gamma) b^{(1)}}{\gamma} \mathbb{1}_{r > (1+\gamma) b^{(1)}}.$$

Using the decomposition $L_\gamma = u - v$, our optimization problem can be formulated as follows:

$$\min_{\mathbf{w} \in \mathbb{R}^N} U(\mathbf{w}) - V(\mathbf{w}) \qquad \text{subject to } \|\mathbf{w}\| \leq \Lambda, \tag{12}$$

where $U(\mathbf{w}) = \sum_{i=1}^{m} u(\mathbf{w} \cdot \mathbf{x}_i, \mathbf{b}_i)$ and $V(\mathbf{w}) = \sum_{i=1}^{m} v(\mathbf{w} \cdot \mathbf{x}_i, \mathbf{b}_i)$, which shows that it can be formulated as a DC-programming problem. The global minimum of the optimization problem (12) can be found using a cutting plane method (Horst and Thoai, 1999), but that method only converges in the limit and does not admit known algorithmic convergence guarantees.[4] There exists also a branch-and-bound algorithm with exponential convergence for DC-programming (Horst and Thoai, 1999) for finding the global minimum. Nevertheless, in (Tao and An, 1997), it is pointed out that such combinatorial algorithms fail to solve real-world DC-programs in high dimensions. In fact, our implementation of this algorithm shows that the convergence of the algorithm in practice is extremely slow for even moderately high-dimensional problems. Another attractive solution for finding the global solution of a DC-programming problem over a polyhedral convex set is the combinatorial solution of Tuy (1964). However, this method requires explicitly specifying the slope and offsets for the piecewise linear function corresponding to a sum of L_γ losses and incurs an exponential cost in time and space.

An alternative consists of using the DC algorithm (DCA), a primal-dual sub-differential method of Dinh Tao and Hoai An Tao and An (1998), (see also Tao and An (1997) for a good survey). This algorithm is applicable when u and v are proper lower semi-continuous convex functions as in our case. When v is differentiable, the DC algorithm coincides with the CCCP algorithm of Yuille and Rangarajan (2003), which has been used in several contexts in machine learning and analyzed by Sriperumbudur and Lanckriet (2012).

The general proof of convergence of the DC algorithm was given by Tao and An (1998). In some special cases, the DC algorithm can be used to find the global minimum of the problem as in the trust region problem (Tao and An, 1998), but, in general, the DC algorithm or its special case CCCP are only guaranteed to converge to a critical point (Tao and An, 1998; Sriperumbudur and Lanckriet, 2012). Nevertheless, the number of iterations of the DC algorithm is relatively small. Its convergence has been shown to be in fact linear for DC-programming problems such as ours (Yen et al., 2012). The algorithm we are proposing goes one step further than that of Tao and An (1998): we use DCA to find a local minimum but then restart our algorithm with a new seed that is guaranteed to reduce the objective function. Unfortunately, we are not in the same regime as in the trust region problem of Tao and An (1998) where the number of local minima is linear in the size of the input. Indeed, here, the number of local minima can be exponential in the number of dimensions of the feature space and it is not clear to us how the combinatorial structure of the problem could help us rule out some local minima faster and make the optimization more tractable.

In the following, we describe more in detail the solution we propose for solving the DC-programming problem (12). The functions v and V are not differentiable in our context but they admit a sub-gradient at all points. We will denote by $\delta V(\mathbf{w})$ an arbitrary element of the sub-gradient $\partial V(\mathbf{w})$, which coincides with $\nabla V(\mathbf{w})$ at points $\mathbf{w}$ where V is differentiable. The DC algorithm then coincides with CCCP, modulo the replacement of the gradient of V by $\delta V(\mathbf{w})$. It consists of starting with a weight vector $\mathbf{w}_0 \leq \Lambda$ and of iteratively solving a sequence of convex optimization problems obtained by replacing V with its linear approximation giving $\mathbf{w}_t$ as a function of $\mathbf{w}_{t-1}$, for $t = 1, \ldots, T$: $\mathbf{w}_t \in \operatorname{argmin}_{\|\mathbf{w}\| \leq \Lambda} U(\mathbf{w}) - \delta V(\mathbf{w}_{t-1}) \cdot$

[4] Some claims of Horst and Thoai (1999), e.g., Proposition 4.4 used in support of the cutting plane algorithm, are incorrect (Tuy, 2002).

```
DC Algorithm
$\mathbf{w} \leftarrow \mathbf{w}_0$            ▷ initialization
while $\mathbf{v} \neq \mathbf{w}$ do
    $\mathbf{v} \leftarrow \mathrm{DCA}(\mathbf{w})$    ▷ DC algorithm
    $\mathbf{u} \leftarrow \frac{\mathbf{v}}{\|\mathbf{v}\|}$
    $\eta^* \leftarrow \min_{0 \leq \eta \leq \Lambda} \sum_{\mathbf{u} \cdot \mathbf{x}_i > 0} L_\gamma(\eta \mathbf{u} \cdot \mathbf{x}_i, \mathbf{b}_i)$
    $\mathbf{w} \leftarrow \eta^* \mathbf{v}$
end while
```

Figure 5: Pseudocode of our DC-programming algorithm.

$\mathbf{w}$. This problem can be rewritten in our context as the following:

$$\min_{\|\mathbf{w}\|^2 \leq \Lambda^2, \mathbf{s}} \sum_{i=1}^m s_i - \delta V(\mathbf{w}_{t-1}) \cdot \mathbf{w} \tag{13}$$
$$\text{subject to } (s_i \geq -\mathbf{w} \cdot \mathbf{x}_i) \wedge \Big[s_i \geq \frac{1}{\gamma} \big(\mathbf{w} \cdot \mathbf{x}_i - (1+\gamma) b_i^{(1)} \big) \Big].$$

The problem is equivalent to a QP (quadratic-programming). Indeed, by convex duality, there exists a $\lambda > 0$ such that the above problem is equivalent to

$$\min_{w \in R^N} \lambda \|w\|^2 + \sum_{i=1}^m s_i - \delta V(\mathbf{w}_{t-1}) \cdot \mathbf{w}$$
$$\text{subject to } (s_i \geq -\mathbf{w} \cdot \mathbf{x}_i) \wedge \Big[s_i \geq \frac{1}{\gamma} \big(\mathbf{w} \cdot \mathbf{x}_i - (1+\gamma) b_i^{(1)} \big) \Big]$$

which is a simple QP that can be tackled by one of many off-the-shelf QP solvers. Of course, the value of λ as a function of Λ does not admit a simple expression. Instead, we select λ through validation which is then equivalent to choosing the optimal value of Λ through validation.

We now address the problem of the DC algorithm converging to a local minimum. A common practice is to restart the DC algorithm at a new random point. Instead, we propose an algorithm that iterates along different local minima, with the guarantee of reducing the function at every change of local minimum. The algorithm is simple and is based on the observation that the function L_γ is positive homogeneous. Indeed, for any $\eta > 0$ and $(r, \mathbf{b})$,

$$\begin{aligned} L_\gamma(\eta r, \eta \mathbf{b}) &= -\eta b^{(2)} 1_{\eta r < \eta b^{(2)}} - \eta r 1_{\eta b^{(2)} \leq \eta r \leq \eta b^{(1)}} + \frac{\eta r - (1+\gamma) \eta b^{(1)}}{\gamma} 1_{\eta b^{(1)} < \eta r < \eta (1+\gamma) b^{(1)}} \\ &= \eta L_\gamma(r, \mathbf{b}). \end{aligned}$$

Minimizing the objective function of (12) in a fixed direction $\mathbf{u}$, $\|\mathbf{u}\| = 1$, can be reformulated as follows: $\min_{0 \leq \eta \leq \Lambda} \sum_{i=1}^m L_\gamma(\eta \mathbf{u} \cdot \mathbf{x}_i, \mathbf{b}_i)$. Since for $\mathbf{u} \cdot \mathbf{x}_i \leq 0$ the function $\eta \mapsto L_\gamma(\eta \mathbf{u} \cdot \mathbf{x}_i, \mathbf{b}_i)$ is constant and equal to $-b_i^{(2)}$, the problem is equivalent to solving

$$\min_{0 \leq \eta \leq \Lambda} \sum_{\mathbf{u} \cdot \mathbf{x}_i > 0} L_\gamma(\eta \mathbf{u} \cdot \mathbf{x}_i, \mathbf{b}_i).$$

Furthermore, since L_γ is positive homogeneous, for all $i \in [1, m]$ with $\mathbf{u} \cdot \mathbf{x}_i > 0$, $L_\gamma(\eta \mathbf{u} \cdot \mathbf{x}_i, \mathbf{b}_i) = (\mathbf{u} \cdot \mathbf{x}_i) L_\gamma(\eta, \mathbf{b}_i / (\mathbf{u} \cdot \mathbf{x}_i))$. But $\eta \mapsto (\mathbf{u} \cdot \mathbf{x}_i) L_\gamma(\eta, \mathbf{b}_i / (\mathbf{u} \cdot \mathbf{x}_i))$ is a v-function and thus the problem can efficiently be optimized using the combinatorial algorithm for the no-feature case (Section 5.1). This leads to the optimization algorithm described in Figure 5. The last step of each iteration of our algorithm can be viewed as a *line search* and this is in fact the step that reduces the objective function the most in practice. This is because we are then precisely minimizing the objective function even though this is for some fixed direction. Since in general this line search does not find a local minimum (we are likely to decrease the objective value in other directions that are not the one in which the line search was performed) running DCA helps us find a better direction for the next iteration of the line search.

6. Experiments

In this section, we report the results of several experiments with synthetic and real-world data demonstrating the benefits of our algorithm. Since the use of features for reserve price optimization has not been previously studied in the literature, we are not aware of any baseline for comparison with our algorithm. Therefore, its performance is measured against three natural strategies that we now describe.

As mentioned before, a standard solution for solving this problem would be the use of a convex surrogate loss. In view of that, we compare against the solution of the regularized empirical risk minimization of the convex surrogate loss L_α shown in Figure 2(a) parametrized by $\alpha \in [0, 1]$ and defined by

$$L_\alpha(r, \mathbf{b}) = \begin{cases} -r & \text{if } r < b^{(1)} + \alpha(b^{(2)} - b^{(1)}) \\ \left(\frac{(1-\alpha)b^{(1)} + \alpha b^{(2)}}{\alpha(b^{(1)} - b^{(2)})}\right)(r - b^{(1)}) & \text{otherwise.} \end{cases}$$

A second alternative consists of using ridge regression to estimate the first bid and of using its prediction as the reserve price. A third algorithm consists of minimizing the loss while ignoring the feature vectors $\mathbf{x}_i$, i.e., solving the problem $\min_{r \leq \Lambda} \sum_{i=1}^n L(r, \mathbf{b}_i)$. It is worth mentioning that this third approach is very similar to what advertisement exchanges currently use to suggest reserve prices to publishers. By using the empirical version of equation (2), we see that this algorithm is equivalent to finding the empirical distribution of bids and optimizing the expected revenue with respect to this empirical distribution as in (Ostrovsky and Schwarz, 2011) and (Cesa-Bianchi et al., 2013).

6.1 Artificial Data Sets

We generated 4 different synthetic data sets with different correlation levels between features and bids. For all our experiments, the feature vectors $\mathbf{x} \in \mathbb{R}^{21}$ were generated in as follows: $\tilde{\mathbf{x}} \in \mathbb{R}^{20}$ was sampled from a standard Gaussian distribution and $\mathbf{x} = (\tilde{\mathbf{x}}, 1)$ was created by adding an offset feature. We now describe the bid generating process for each of the experiments as a function of the feature vector $\mathbf{x}$. For our first three experiments, shown in Figure 6(a)-(c), the highest bid and second highest bid were set to $\max\left(\left|\sum_{i=1}^{21} x_i\right| + \epsilon_1, \left|\sum_{i=1}^{21} \frac{x_i}{2}\right| + \epsilon_2\right)_+$ and $\min\left(\left|\sum_{i=1}^{21} x_i\right| + \epsilon_1, \left|\sum_{i=1}^{21} \frac{x_i}{2}\right| + \epsilon_2\right)_+$ respectively, where ϵ_i is a Gaussian random variable with mean 0. The standard deviation of the Gaussian noise was varied over the set $\{0, 0.25, 0.5\}$.

For our last artificial experiment, we used a generative model motivated by previous empirical observations (Ostrovsky and Schwarz, 2011; Lahaie and Pennock, 2007): bids were generated by sampling two values from a log-normal distribution with means $\mathbf{x} \cdot \mathbf{w}$ and $\frac{\mathbf{x} \cdot \mathbf{w}}{2}$ and standard deviation 0.5, with $\mathbf{w}$ a random vector sampled from a standard Gaussian distribution.

For all our experiments, the parameters λ, γ and α were selected respectively from the sets $\{2^i | i \in [-5, 5]\}$, $\{0.1, 0.01, 0.001\}$, and $\{0.1, 0.2, \ldots, 0.9\}$ via validation over a set consisting of the same number of examples as the training set. Our algorithm was initialized using the best solution of the convex surrogate optimization problem. The test set consisted of $5{,}000$ examples drawn from the same distribution as the training set. Each experiment was repeated 10 times and the mean revenue of each algorithm is shown in Figure 6. The plots are normalized in such a way that the revenue obtained by setting no reserve price is equal to 0 and the maximum possible revenue (which can be obtained by setting the reserve price equal to the highest bid) is equal to 1. The performance of the ridge regression algorithm is not included in Figure 6(d) as it was too inferior to be comparable with the performance of the other algorithms.

By inspecting the results in Figure 6(a), we see that, even in the simplest noiseless scenario, our algorithm outperforms all other techniques. The reader could argue that these results are, in fact, not surprising since the bids were generated by a locally linear function of the feature vectors, thereby ensuring the success of our algorithm. Nevertheless, one would expect this to be the case too for algorithms that leverage the use of features such as the convex surrogate and ridge regression. But one can see that this is in fact not true even for low levels of noise. It is also worth noticing that the use of ridge regression is actually worse than setting the reserve price to 0. This fact can be easily understood by noticing that the square loss used in regression is symmetric. Therefore, we can expect several reserve prices to be above the highest bid, making the revenue of these auctions equal to zero. Another notable feature is that as the noise level increases, the performance of feature-based algorithms decreases. This is true for any learning algorithm: if the features are not relevant to the prediction task, the performance of the algorithm will suffer. However, for the convex surrogate algorithm, a more critical issue occurs: the performance of this algorithm actually decreases as the sample size increases,

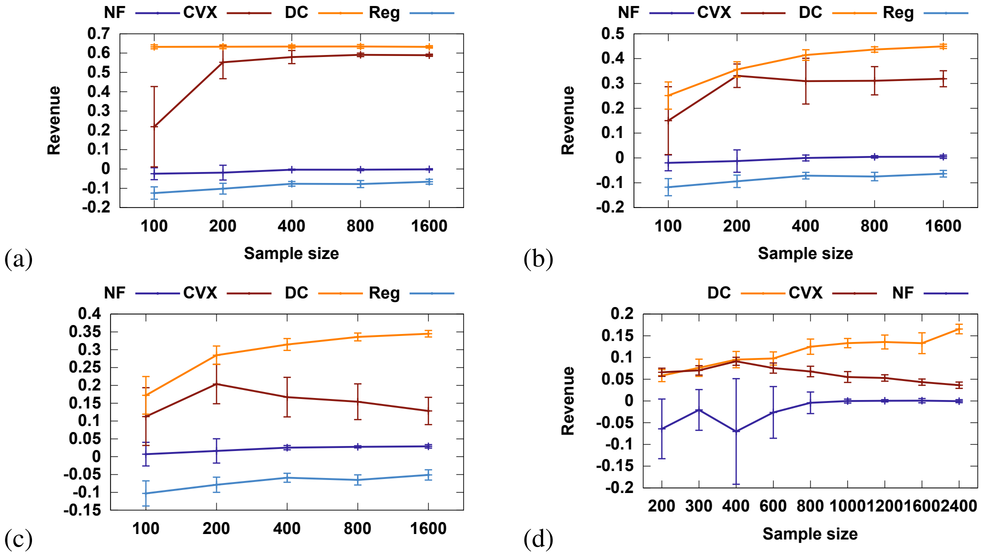


Figure 6: Plots of expected revenue against sample size for different algorithms: DC algorithm (DC), convex surrogate (CVX), ridge regression (Reg) and the algorithm that uses no feature to set reserve prices (NF). For (a)-(c) bids are generated with different noise standard deviation (a) 0, (b) 0.25, (c) 0.5. The bids in (d) were generated using a generative model.

which shows that in general learning with a convex surrogate is not possible. This is an empirical verification of the inconsistency result provided in Section 4.2. This lack of calibration can also be seen in Figure 6(d), where in fact the performance of this algorithm approaches the use of no reserve price.

In order to better understand the reason behind the performance discrepancy between feature-based algorithms, we analyze the reserve prices offered by each algorithm. In Figure 7 we see that the convex surrogate algorithm tends to offer lower reserve prices. This should be intuitively clear as high reserve prices are over-penalized by the chosen convex surrogate as shown in Figure 2(b). On the other hand, reserve prices suggested by the regression algorithm seem to be concentrated and symmetric around their mean. Therefore we can infer that about 50% of the reserve prices offered will be higher than the highest bid thereby yielding zero revenue. Finally, our algorithm seems to generally offer higher prices. This suggests that the increase in revenue comes from auctions where the highest bid is large but the second bid is small. This bidding phenomenon is in fact commonly observed in practice (Amin et al., 2013).

6.2 Real-world Data Sets

Due to proprietary data and confidentiality reasons, we cannot present empirical results for AdExchange data. However, we were able to procure an eBay data set consisting of approximately 70,000 second-price auctions of collector sport cards. The full data set can be accessed using the following URL: http://cims.nyu.edu/~munoz/data. Some other sources of auction data are accessible (e.g., http://modelingonlineauctions.com/datasets), but features are not available for those datasets. To the best of our knowledge, with the exception of the one used here, there is no publicly available data set for online auctions including features that could be readily used with our algorithm. The features used here include information about the seller such as positive feedback percent, seller rating and seller country; as well as information about the card such as whether the player is in the sport's Hall of Fame. The final dimension of the feature vectors is 78. The values of these features are both continuous and categorical. For our experiments we also included an extra offset feature.

Since the highest bid is not reported by eBay, our algorithm cannot be straightforwardly used on this data set. In order to generate highest bids, we calculated the mean price of each object (each card was generally

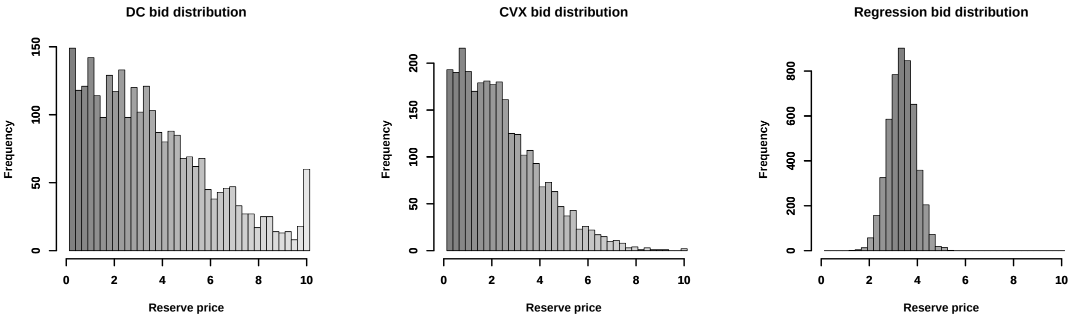


Figure 7: Distribution of reserve prices for each algorithm. The algorithms were trained on 800 samples using noisy bids with standard deviation 0.5.

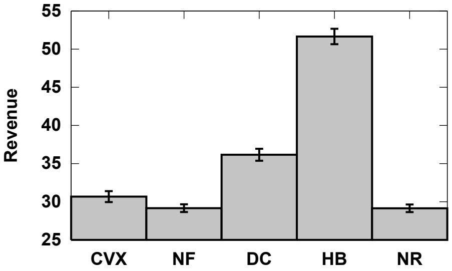


Figure 8: Results of the eBay data set. Comparison of our algorithm (DC) against a convex surrogate (CVX), using no features (NF), setting no reserve (NR) and setting reserve price to highest bid (HB).

sold more than once) and set the highest bid to be the maximum between this average and the second highest bid.

Figure 8 shows the revenue gained using different algorithms including our DC algorithm, using a convex surrogate, or the algorithm that ignores features. It also shows the results obtained by using no reserve price (NR) and the highest possible revenue obtained by setting the reserve price to the highest bid (HB). We randomly sampled $2{,}000$ examples for training, $2{,}000$ examples for validation and $2{,}000$ examples for testing. This experiment was repeated 10 times. Figure 8(b) shows the mean revenue for each algorithm and their standard deviations. The results of this experiment show that the use of features is crucial for revenue optimization. Indeed, setting an optimal reserve price for all objects seems to achieve the same revenue as no reserve price. Instead, our algorithm achieves a 22% increase on the revenue obtained by not setting a reserve price whereas the non-calibrated convex surrogate algorithm only obtains a 3% revenue improvement. Furthermore, our algorithm is able to obtain as much as 70% of the achievable revenue with knowledge of the highest bid.

7. Conclusion

We presented a comprehensive theoretical and algorithmic analysis of the learning problem of revenue optimization in second-price auctions with reserve. The specific properties of the loss function for this problem required a new analysis and new learning guarantees. The algorithmic solutions we presented are practically applicable to revenue optimization problems for this type of auctions in most realistic settings. Our experimental results further demonstrate their effectiveness. Much of the analysis and algorithms presented, in particular our study of calibration questions, can also be of interest in other learning problems. In particular, they are relevant to the study of learning problems arising in the study of generalized second-Price auctions (Mohri and Medina, 2015).

Acknowledgments

We thank Afshin Rostamizadeh and Umar Syed for several discussions about the topic of this work and ICML and JMLR reviewers for useful comments. We also warmly thank Jay Grossman for providing us access to the eBay data set used in this paper. This work was partly funded by the NSF award IIS-1117591.

References

Kareem Amin, Michael Kearns, Peter Key, and Anton Schwaighofer. Budget optimization for sponsored search: Censored learning in MDPs. In *UAI*, pages 54–63, 2012.

Kareem Amin, Afshin Rostamizadeh, and Umar Syed. Learning prices for repeated auctions with strategic buyers. In *Proceedings of NIPS*, pages 1169–1177, 2013.

Peter Auer, Nicolò Cesa-Bianchi, Yoav Freund, and Robert E. Schapire. The nonstochastic multiarmed bandit problem. *SIAM J. Comput.*, 32(1):48–77, 2002.

Maria-Florina Balcan, Avrim Blum, Jason D. Hartline, and Yishay Mansour. Reducing mechanism design to algorithm design via machine learning. *J. Comput. Syst. Sci.*, 74(8):1245–1270, 2008.

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

Avrim Blum, Vijay Kumar, Atri Rudra, and Felix Wu. Online learning in online auctions. *Theor. Comput. Sci.*, 324(2-3): 137–146, 2004.

Nicolò Cesa-Bianchi, Claudio Gentile, and Yishay Mansour. Regret minimization for reserve prices in second-price auctions. In *SODA*, pages 1190–1204, 2013.

Ying Cui, Ruofei Zhang, Wei Li, and Jianchang Mao. Bid landscape forecasting in online ad exchange marketplace. In *Proceedings of KDD*, pages 265–273, 2011.

Gerard Debreu and Tjalling C. Koopmans. Additively decomposed quasiconvex functions. *Mathematical Programming*, 24, 1982.

Nikhil R. Devanur and Sham M. Kakade. The price of truthfulness for pay-per-click auctions. In *Proceedings of EC*, pages 99–106, 2009.

David A. Easley and Jon M. Kleinberg. *Networks, Crowds, and Markets - Reasoning About a Highly Connected World*. Cambridge University Press, 2010.

Di He, Wei Chen, Liwei Wang, and Tie-Yan Liu. Online learning for auction mechanism in bandit setting. *Decision Support Systems*, 56:379–386, 2013.

R Horst and Nguyen V Thoai. DC programming: overview. *Journal of Optimization Theory and Applications*, 103(1): 1–43, 1999.

Robert D. Kleinberg and Frank Thomson Leighton. The value of knowing a demand curve: Bounds on regret for online posted-price auctions. In *Proceedings of FOCS*, pages 594–605, 2003.

V. Koltchinskii and D. Panchenko. Empirical margin distributions and bounding the generalization error of combined classifiers. *Ann. Statist.*, 30(1):1–50, 2002.

Sébastien Lahaie and David M. Pennock. Revenue analysis of a family of ranking rules for keyword auctions. In *Proceedings of EC*, pages 50–56, 2007.

John Langford, Lihong Li, Yevgeniy Vorobeychik, and Jennifer Wortman. Maintaining equilibria during exploration in sponsored search auctions. *Algorithmica*, 58(4):990–1021, 2010.

Michel Ledoux and Michel Talagrand. *Probability in Banach spaces*. Classics in Mathematics. Springer-Verlag, Berlin, 2011. Isoperimetry and processes, Reprint of the 1991 edition.

P.R. Milgrom and R.J. Weber. A theory of auctions and competitive bidding. *Econometrica: Journal of the Econometric Society*, pages 1089–1122, 1982.

Mehryar Mohri and Andrés Muñoz Medina. Non-parametric revenue optimization for generalized second-price auctions. In *Proceedings of UAI*, Amsterdam, The Netherlands, 2015.

Mehryar Mohri, Afshin Rostamizadeh, and Ameet Talwalkar. *Foundations of machine learning*. MIT Press, Cambridge, MA, 2012.

S Muthukrishnan. Ad exchanges: Research issues. *Internet and network economics*, pages 1–12, 2009.

R.B. Myerson. Optimal auction design. *Mathematics of operations research*, 6(1):58–73, 1981.

Noam Nisan, Tim Roughgarden, Éva Tardos, and Vijay V. Vazirani, editors. *Algorithmic game theory*. Cambridge University Press, Cambridge, 2007.

Michael Ostrovsky and Michael Schwarz. Reserve prices in internet advertising auctions: a field experiment. In *Proceedings of EC*, pages 59–60, 2011.

David Pollard. *Convergence of stochastic processes*. Springer Series in Statistics. Springer-Verlag, New York, 1984.

J.G. Riley and W.F. Samuelson. Optimal auctions. *The American Economic Review*, pages 381–392, 1981.

Bharath K. Sriperumbudur and Gert R. G. Lanckriet. A proof of convergence of the concave-convex procedure using Zangwill's theory. *Neural Computation*, 24(6):1391–1407, 2012.

Pham Dinh Tao and Le Thi Hoai An. Convex analysis approach to DC programming: theory, algorithms and applications. *Acta Mathematica Vietnamica*, 22(1):289–355, 1997.

Pham Dinh Tao and Le Thi Hoai An. A DC optimization algorithm for solving the trust-region subproblem. *SIAM Journal on Optimization*, 8(2):476–505, 1998.

Hoang Tuy. Concave programming under linear constraints. *Translated Soviet Mathematics*, 5:1437–1440, 1964.

Hoang Tuy. Counter-examples to some results on D.C. optimization. Technical report, Institute of Mathematics, Hanoi, Vietnam, 2002.

William Vickrey. Counterspeculation, auctions, and competitive sealed tenders. *The Journal of finance*, 16(1):8–37, 1961.

Ian E.H. Yen, Nanyun Peng, Po-Wei Wang, and Shou-De Lin. On convergence rate of concave-convex procedure. In *Proceedings of the NIPS 2012 Optimization Workshop*, 2012.

Alan L. Yuille and Anand Rangarajan. The concave-convex procedure. *Neural Computation*, 15(4):915–936, 2003.

Yunzhang Zhu, Gang Wang, Junli Yang, Dakan Wang, Jun Yan, Jian Hu, and Zheng Chen. Optimizing search engine revenue in sponsored search. In *Proceedings of ACM-SIGIR*, pages 588–595, 2009.

Appendix A. Contraction Lemma

The following is a version of Talagrand's contraction lemma Ledoux and Talagrand (2011). Since our definition of Rademacher complexity does not use absolute values, we give an explicit proof below.

Lemma 14 *Let H be a hypothesis set of functions mapping $\mathcal{X}$ to $\mathbb{R}$ and $\Psi_1, \ldots, \Psi_m$, μ-Lipschitz functions for some $\mu > 0$. Then, for any sample S of m points $x_1, \ldots, x_m \in \mathcal{X}$, the following inequality holds*

$$\frac{1}{m} \mathop{\mathbb{E}}_{\boldsymbol{\sigma}} \left[\sup_{h \in H} \sum_{i=1}^{m} \sigma_i (\Psi_i \circ h)(x_i) \right] \leq \frac{\mu}{m} \mathop{\mathbb{E}}_{\boldsymbol{\sigma}} \left[\sup_{h \in H} \sum_{i=1}^{m} \sigma_i h(x_i) \right] = \mu \, \widehat{\mathfrak{R}}_S(H).$$

Proof The proof is similar to the case where the functions Ψ_i are all equal. Fix a sample $S = (x_1, \ldots, x_m)$. Then, we can rewrite the empirical Rademacher complexity as follows:

$$\frac{1}{m} \mathop{\mathbb{E}}_{\boldsymbol{\sigma}} \Big[\sup_{h \in H} \sum_{i=1}^{m} \sigma_i (\Psi_i \circ h)(x_i) \Big] = \frac{1}{m} \mathop{\mathbb{E}}_{\sigma_1, \ldots, \sigma_{m-1}} \Big[\mathop{\mathbb{E}}_{\sigma_m} \Big[\sup_{h \in H} u_{m-1}(h) + \sigma_m (\Psi_m \circ h)(x_m) \Big] \Big],$$

where $u_{m-1}(h) = \sum_{i=1}^{m-1} \sigma_i (\Psi_i \circ h)(x_i)$. Assume that the suprema can be attained and let $h_1, h_2 \in H$ be the hypotheses satisfying

$$\begin{aligned} u_{m-1}(h_1) + \Psi_m(h_1(x_m)) &= \sup_{h \in H} u_{m-1}(h) + \Psi_m(h(x_m)) \\ u_{m-1}(h_2) - \Psi_m(h_2(x_m)) &= \sup_{h \in H} u_{m-1}(h) - \Psi_m(h(x_m)). \end{aligned}$$

When the suprema are not reached, a similar argument to what follows can be given by considering instead hypotheses that are ϵ-close to the suprema for any $\epsilon > 0$.

By definition of expectation, since σ_m uniform distributed over $\{-1, +1\}$, we can write

$$\begin{aligned} \mathop{\mathbb{E}}_{\sigma_m} \Big[\sup_{h \in H} u_{m-1}(h) + \sigma_m (\Psi_m \circ h)(x_m) \Big] &= \frac{1}{2} \sup_{h \in H} u_{m-1}(h) + (\Psi_m \circ h)(x_m) + \frac{1}{2} \sup_{h \in H} u_{m-1}(h) - (\Psi_m \circ h)(x_m) \\ &= \frac{1}{2} [u_{m-1}(h_1) + (\Psi_m \circ h_1)(x_m)] + \frac{1}{2} [u_{m-1}(h_2) - (\Psi_m \circ h_2)(x_m)]. \end{aligned}$$

Let $s = \operatorname{sgn}(h_1(x_m) - h_2(x_m))$. Then, the previous equality implies

$$\begin{aligned} \mathop{\mathbb{E}}_{\sigma_m} \Big[\sup_{h \in H} u_{m-1}(h) + \sigma_m (\Psi_m \circ h)(x_m) \Big] &\leq \frac{1}{2} [u_{m-1}(h_1) + u_{m-1}(h_2) + s\mu(h_1(x_m) - h_2(x_m))] \\ &= \frac{1}{2} [u_{m-1}(h_1) + s\mu h_1(x_m)] + \frac{1}{2} [u_{m-1}(h_2) - s\mu h_2(x_m)] \\ &\leq \frac{1}{2} \sup_{h \in H} [u_{m-1}(h) + s\mu h(x_m)] + \frac{1}{2} \sup_{h \in H} [u_{m-1}(h) - s\mu h(x_m)] \\ &= \mathop{\mathbb{E}}_{\sigma_m} \Big[\sup_{h \in H} u_{m-1}(h) + \sigma_m \mu h(x_m) \Big], \end{aligned}$$

where we used the $\mu-$Lipschitzness of Ψ_m in the first equality and the definition of expectation over σ_m for the last equality. Proceeding in the same way for all other σ_i's ($i \neq m$) proves the lemma. ■

Appendix B. Proof of Theorem 7

Proof We first show that the functions L_n are uniformly bounded for any b:

$$\begin{aligned} |L_n(r, b)| = \Big| \int_0^r L_n'(r, b) dr \Big| &\leq \int_0^M \max\Big(\Big| L_n'(0, b) \Big|, \Big| L_n'(M, b) \Big| \Big) dr \\ &\leq \int_0^M K dr = MK, \end{aligned}$$

where the first inequality holds since, by convexity, the derivative of L_n with respect to r is an increasing function.

Next, we show that the sequence $(L_n)_{n\in\mathbb{N}}$ is also equicontinuous. It will follow then by the theorem of Arzela-Ascoli that the sequence $L_n(\cdot, b)$ converges uniformly to $L_c(\cdot, b)$. Let $r_1, r_2 \in [0, M]$, for any $b \in [0, M]$ we have

$$\begin{aligned}|L_n(r_1, b) - L_n(r_2, b)| &\leq \sup_{r\in[0,M]} \left|L_n'(r, b)\right| |r_1 - r_2| \\ &= \max\left(\left|L_n'(0, b)\right|, \left|L_n'(M, b))\right|\right) |r_1 - r_2| \\ &\leq K|r_1 - r_2|,\end{aligned}$$

where, again, the convexity of L_n was used for the first equality. Let $F_n(r) = \mathbb{E}_{b\sim D}[L_n(r, b)]$ and $F(r) = \mathbb{E}_{b\sim D}[L_c(r, b)]$. F_n is a convex function as the expectation of a convex function. By the theorem of Arzela-Ascoli, the sequence $(F_n)_n$ admits a uniformly convergent subsequence. Furthermore, by the dominated convergence theorem, we have $(F_n(r))_n$ converges pointwise to $F(r)$. Therefore, the uniform limit of F_n must be F. This implies that

$$\min_{r\in[0,M]} F(r) = \lim_{n\to+\infty} \min_{r\in[0,M]} F_n(r) = \lim_{n\to+\infty} F_n(r_n) = F(r^*),$$

where the first and third equalities follow from the uniform convergence of F_n to F. The last equation implies that L_c is consistent with $\widetilde{L}$. Furthermore, the function $L_c(\cdot, b)$ is convex since it is the uniform limit of convex functions. It then follows by Proposition 6 that $L_c(\cdot, b) \equiv L_c(0, b) = 0$. ■

Appendix C. Consistency of L_γ

Lemma 15 *Let H be a closed, convex subset of a linear space of functions containing 0 and let $h_\gamma^* = \operatorname{argmin}_{h\in H} \mathcal{L}_\gamma(h)$. Then, the following inequality holds:*

$$\mathop{\mathbb{E}}_{\mathbf{x},\mathbf{b}}\left[h_\gamma^*(\mathbf{x})\mathbb{1}_{I_2}(\mathbf{x}, \mathbf{b})\right] \geq \frac{1}{\gamma} \mathop{\mathbb{E}}_{\mathbf{x},\mathbf{b}}\left[h_\gamma^*(\mathbf{x})\mathbb{1}_{I_3}(\mathbf{x}, \mathbf{b})\right].$$

Proof Let $0 < \lambda < 1$. Since H is a convex set, it follows that $\lambda h_\gamma^* \in H$. Furthermore, by the definition of h_γ^*, we must have:

$$\mathop{\mathbb{E}}_{\mathbf{x},\mathbf{b}}\left[L_\gamma(h_\gamma^*(\mathbf{x}), \mathbf{b})\right] \leq \mathop{\mathbb{E}}_{\mathbf{x},\mathbf{b}}\left[L_\gamma(\lambda h_\gamma^*(\mathbf{x}), \mathbf{b})\right]. \tag{14}$$

If $h_\gamma^*(\mathbf{x}) < 0$, then $L_\gamma(h_\gamma^*(\mathbf{x}), \mathbf{b}) = L_\gamma(\lambda h_\gamma^*(\mathbf{x})) = -b^{(2)}$ by definition of L_γ. If on the other hand $h_\gamma^*(\mathbf{x}) > 0$, since $\lambda h_\gamma^*(\mathbf{x}) < h_\gamma^*(\mathbf{x})$, we must have that for $(\mathbf{x}, \mathbf{b}) \in I_1$ $L_\gamma(h_\gamma^*(\mathbf{x}), \mathbf{b}) = L_\gamma(\lambda h_\gamma^*(\mathbf{x}), \mathbf{b}) = -b^{(2)}$ too. Moreover, from the fact that $L_\gamma \leq 0$ and $L_\gamma(h_\gamma^*(\mathbf{x}), \mathbf{b}) = 0$ for $(\mathbf{x}, \mathbf{b}) \in I_4$ it follows that $L_\gamma(h_\gamma^*(\mathbf{x}), \mathbf{b}) \geq L_\gamma(\lambda h_\gamma^*(\mathbf{x}), \mathbf{b})$ for $(\mathbf{x}, \mathbf{b}) \in I_4$, and therefore the following inequality trivially holds:

$$\mathop{\mathbb{E}}_{\mathbf{x},\mathbf{b}}\left[L_\gamma(h_\gamma^*(\mathbf{x}), \mathbf{b})(\mathbb{1}_{I_1}(\mathbf{x}, \mathbf{b}) + \mathbb{1}_{I_4}(\mathbf{x}, \mathbf{b}))\right] \geq \mathop{\mathbb{E}}_{\mathbf{x},\mathbf{b}}\left[L_\gamma(\lambda h_\gamma^*(\mathbf{x}), \mathbf{b})(\mathbb{1}_{I_1}(\mathbf{x}, \mathbf{b}) + \mathbb{1}_{I_4}(\mathbf{x}, \mathbf{b}))\right]. \tag{15}$$

Subtracting (15) from (14) we obtain

$$\mathop{\mathbb{E}}_{\mathbf{x},\mathbf{b}}\left[L_\gamma(h_\gamma^*(\mathbf{x}), \mathbf{b})(\mathbb{1}_{I_2}(\mathbf{x}, \mathbf{b}) + \mathbb{1}_{I_3}(\mathbf{x}, \mathbf{b}))\right] \leq \mathop{\mathbb{E}}_{\mathbf{x},\mathbf{b}}\left[L_\gamma(\lambda h_\gamma^*(\mathbf{x}), \mathbf{b})(\mathbb{1}_{I_2}(\mathbf{x}, \mathbf{b}) + \mathbb{1}_{I_3}(\mathbf{x}, \mathbf{b}))\right].$$

Rearranging terms shows that this inequality is equivalent to

$$\mathop{\mathbb{E}}_{\mathbf{x},\mathbf{b}}\left[(L_\gamma(\lambda h_\gamma^*(\mathbf{x}), \mathbf{b}) - L_\gamma(h_\gamma^*(\mathbf{x}), \mathbf{b}))\mathbb{1}_{I_2}(\mathbf{x}, \mathbf{b})\right] \geq \mathop{\mathbb{E}}_{\mathbf{x},\mathbf{b}}\left[(L_\gamma(h_\gamma^*(\mathbf{x}), \mathbf{b}) - L_\gamma(\lambda h_\gamma^*(\mathbf{x}), \mathbf{b}))\mathbb{1}_{I_3}(\mathbf{x}, \mathbf{b})\right] \tag{16}$$

Notice that if $(\mathbf{x}, \mathbf{b}) \in I_2$, then $L_\gamma(h_\gamma^*(\mathbf{x}), \mathbf{b}) = -h_\gamma^*(\mathbf{x})$. If $\lambda h_\gamma^*(\mathbf{x}) > b^{(2)}$ too then $L_\gamma(\lambda h_\gamma^*(\mathbf{x}), \mathbf{b}) = -\lambda h_\gamma^*(\mathbf{x})$. On the other hand if $\lambda h_\gamma^*(\mathbf{x}) \leq b^{(2)}$ then $L_\gamma(\lambda h_\gamma^*(\mathbf{x}), \mathbf{b}) = -b^{(2)} \leq -\lambda h_\gamma^*(\mathbf{x})$. Thus

$$\mathbb{E}(L_\gamma(\lambda h_\gamma^*(\mathbf{x}), \mathbf{b}) - L_\gamma(h_\gamma^*(\mathbf{x}), \mathbf{b}))\mathbb{1}_{I_2}(\mathbf{x}, \mathbf{b})) \leq (1 - \lambda)\,\mathbb{E}(h_\gamma^*(\mathbf{x})\mathbb{1}_{I_2}(\mathbf{x}, \mathbf{b})) \tag{17}$$

This gives an upper bound for the left-hand side of inequality (16). We now seek to derive a lower bound on the right-hand side. To do so, we analyze two different cases:

1. $\lambda h^*_\gamma(\mathbf{x}) \leq b^{(1)}$;
2. $\lambda h^*_\gamma(\mathbf{x}) > b^{(1)}$.

In the first case, we know that $L_\gamma(h^*_\gamma(\mathbf{x}), \mathbf{b}) = \frac{1}{\gamma}(h^*_\gamma(\mathbf{x}) - (1+\gamma)b^{(1)}) > -b^{(1)}$ (since $h^*_\gamma(\mathbf{x}) > b^{(1)}$ for $(\mathbf{x}, \mathbf{b}) \in I_3$). Furthermore, if $\lambda h^*_\gamma(\mathbf{x}) \leq b^{(1)}$, then, by definition $L_\gamma(\lambda h^*_\gamma(\mathbf{x}), \mathbf{b}) = \min(-b^{(2)}, -\lambda h^*_\gamma(\mathbf{x})) \leq -\lambda h^*_\gamma(\mathbf{x})$. Thus, we must have:

$$L_\gamma(h^*_\gamma(\mathbf{x}), \mathbf{b}) - L_\gamma(\lambda h^*_\gamma(\mathbf{x}), \mathbf{b}) > \lambda h^*_\gamma(\mathbf{x}) - b^{(1)} > (\lambda - 1)b^{(1)} \geq (\lambda - 1)M, \tag{18}$$

where we used the fact that $h^*_\gamma(\mathbf{x}) > b^{(1)}$ for the second inequality and the last inequality holds since $\lambda - 1 < 0$.

We analyze the second case now. If $\lambda h^*_\gamma(\mathbf{x}) > b^{(1)}$, then for $(\mathbf{x}, \mathbf{b}) \in I_3$ we have $L_\gamma(h^*_\gamma(\mathbf{x}), \mathbf{b}) - L_\gamma(\lambda h^*_\gamma(\mathbf{x}), \mathbf{b}) = \frac{1}{\gamma}(1-\lambda)h^*_\gamma(\mathbf{x})$. Thus, letting $\Delta(\mathbf{x}, \mathbf{b}) = L_\gamma(h^*_\gamma(\mathbf{x}), \mathbf{b}) - L_\gamma(\lambda h^*_\gamma(\mathbf{x}), \mathbf{b})$, we can lower bound the right-hand side of (16) as:

$$\begin{aligned}\mathop{\mathbb{E}}_{\mathbf{x},\mathbf{b}}\Big[\Delta(\mathbf{x}, \mathbf{b})\mathbb{1}_{I_3}(\mathbf{x}, \mathbf{b})\Big] &= \mathop{\mathbb{E}}_{\mathbf{x},\mathbf{b}}\Big[\Delta(\mathbf{x}, \mathbf{b})\mathbb{1}_{I_3}(\mathbf{x}, \mathbf{b})\mathbb{1}_{\{\lambda h^*_\gamma(\mathbf{x}) > b^{(1)}\}}\Big] + \mathop{\mathbb{E}}_{\mathbf{x},\mathbf{b}}\Big[\Delta(\mathbf{x}, \mathbf{b})\mathbb{1}_{I_3}(\mathbf{x}, \mathbf{b})\mathbb{1}_{\{\lambda h^*_\gamma(\mathbf{x}) \leq b^{(1)}\}}\Big] \\ &\geq \frac{1-\lambda}{\gamma}\mathop{\mathbb{E}}_{\mathbf{x},\mathbf{b}}\Big[h^*_\gamma(\mathbf{x})\mathbb{1}_{I_3}(\mathbf{x}, \mathbf{b})\mathbb{1}_{\{\lambda h^*_\gamma(\mathbf{x}) > b^{(1)}\}}\Big] + (\lambda - 1)M\,\mathbb{P}\Big[h^*_\gamma(\mathbf{x}) > b^{(1)} \geq \lambda h^*_\gamma(\mathbf{x})\Big],\end{aligned} \tag{19}$$

where we have used (18) to bound the second summand. Combining inequalities (16), (17) and (19) and dividing by $(1 - \lambda)$ we obtain the bound

$$\mathop{\mathbb{E}}_{\mathbf{x},\mathbf{b}}\Big[h^*_\gamma(\mathbf{x})\mathbb{1}_{I_2}(\mathbf{x}, \mathbf{b})\Big] \geq \frac{1}{\gamma}\mathop{\mathbb{E}}_{\mathbf{x},\mathbf{b}}\Big[h^*_\gamma(\mathbf{x})\mathbb{1}_{I_3}(\mathbf{x}, \mathbf{b})\mathbb{1}_{\{\lambda h^*_\gamma(\mathbf{x}) > b^{(1)}\}}\Big] - M\,\mathbb{P}\Big[h^*_\gamma(\mathbf{x}) > b^{(1)} \geq \lambda h^*_\gamma(\mathbf{x})\Big].$$

Finally, taking the limit $\lambda \to 1$, we obtain

$$\mathop{\mathbb{E}}_{\mathbf{x},\mathbf{b}}\Big[h^*_\gamma(\mathbf{x})\mathbb{1}_{I_2}(\mathbf{x}, \mathbf{b})\Big] \geq \frac{1}{\gamma}\mathop{\mathbb{E}}_{\mathbf{x},\mathbf{b}}\Big[h^*_\gamma(\mathbf{x})\mathbb{1}_{I_3}(\mathbf{x}, \mathbf{b})\Big].$$

Taking the limit inside the expectation is justified by the bounded convergence theorem and $\mathbb{P}[h^*_\gamma(\mathbf{x}) > b^{(1)} \geq \lambda h^*_\gamma(\mathbf{x})] \to 0$ holds by the continuity of probability measures. ■

Proposition 16 *For any $\delta > 0$, with probability at least $1 - \delta$ over the choice of a sample S of size m, the following holds for all $\gamma \in (0, 1]$ and $h \in H$:*

$$\mathcal{L}_\gamma(h) \leq \widehat{\mathcal{L}}_\gamma(h) + \frac{2}{\gamma}\mathfrak{R}_m(H) + M\left[\sqrt{\frac{\log\log_2\frac{1}{\gamma}}{m}} + \sqrt{\frac{\log\frac{1}{\delta}}{2m}}\right].$$

Proof Consider two sequences $(\gamma_k)_{k\geq1}$ and $(\epsilon_k)_{k\geq1}$, with $\epsilon_k \in (0, 1)$. By theorem 9, for any fixed $k \geq 1$,

$$\mathbb{P}\left[\mathcal{L}_{\gamma_k}(h) - \widehat{\mathcal{L}}_{\gamma_k}(h) > \frac{2}{\gamma_k}\mathfrak{R}_m(H) + M\epsilon_k\right] \leq \exp(-2m\epsilon_k^2).$$

Choose $\epsilon_k = \epsilon + \sqrt{\frac{\log k}{m}}$, then, by the union bound,

$$\begin{aligned}\mathbb{P}\left[\exists k \colon \mathcal{L}_{\gamma_k}(h) - \widehat{\mathcal{L}}_{\gamma_k}(h) > \frac{1}{\gamma_k}\mathfrak{R}_m(H) + M\epsilon_k\right] &\leq \sum_{k\geq1}\exp\big[-2m(\epsilon + \sqrt{(\log k)/m})^2\big] \\ &\leq \big(\sum_{k\geq1}1/k^2\big)\exp(-2m\epsilon^2) \\ &= \frac{\pi^2}{6}\exp(-2m\epsilon^2) \leq 2\exp(-2m\epsilon^2).\end{aligned}$$

For any $\gamma \in (0,1]$, there exists $k \geq 1$ such that $\gamma \in (\gamma_k, \gamma_{k-1})$ with $\gamma_k = 1/2^k$. For such a k, $\frac{1}{\gamma_{k-1}} \leq \frac{1}{\gamma}$, $\gamma_{k-1} \leq \frac{\gamma}{2}$, and $\sqrt{\log(k-1)} = \sqrt{\log\log_2(1/\gamma_{k-1})} \leq \sqrt{\log\log_2(1/\gamma)}$. Since for any $h \in H$, $\mathcal{L}_{\gamma_{k-1}}(h) \leq \mathcal{L}_\gamma(h)$, we can write

$$\mathbb{P}\left[\mathcal{L}(h) - \widehat{\mathcal{L}}_\gamma(h) > \frac{2}{\gamma}\mathfrak{R}_m(H) + M\Big(K(\gamma) + \epsilon\Big)\right] \leq \exp(-2m\epsilon^2),$$

where $K(\gamma) = \sqrt{\frac{\log\log_2 \frac{1}{\gamma}}{m}}$. This concludes the proof. ■

Corollary 17 *Let H be a hypothesis set with pseudo-dimension $d = Pdim(H)$. Then, for any $\delta > 0$ and any $\gamma > 0$, with probability at least $1 - \delta$ over the choice of a sample S of size m, the following inequality holds:*

$$\mathcal{L}(\widehat{h}_\gamma) \leq \mathcal{L}^* + \frac{2\gamma + 2}{\gamma}\mathfrak{R}_m(H) + \gamma M + M\left[2\sqrt{\frac{2d\log\frac{\epsilon m}{d}}{m}} + 2\sqrt{\frac{\log\frac{2}{\delta}}{2m}} + \sqrt{\frac{\log\log_2\frac{1}{\gamma}}{m}}\right].$$

The proof follows the same steps as Theorem 10 and uses the results of Proposition 16. Notice that, by setting $\gamma = \frac{1}{m^{1/4}}$, we can guarantee the convergence of $\mathcal{L}(\widehat{h}_\gamma)$ to $\mathcal{L}^*$. Indeed, with this choice, the bound can be expressed as follows:

$$\mathcal{L}(\widehat{h}_\gamma) \leq \mathcal{L}^* + (2 + m^{1/4})\mathfrak{R}_m(H) + \frac{1}{m^{1/4}}M + M\left[2\sqrt{\frac{2d\log\frac{\epsilon m}{d}}{m}} + 2\sqrt{\frac{\log\frac{2}{\delta}}{2m}} + \sqrt{\frac{\log\log_2 m^{1/4}}{m}}\right].$$

Furthermore, when H has finite pseudo-dimension, it is known that $\mathfrak{R}_m(H)$ is in $O\big(\frac{1}{m^{1/2}}\big)$. Thus, this shows that $\mathcal{L}(\widehat{h}_\gamma) = \mathcal{L}^* + O\big(\frac{1}{m^{1/4}}\big)$.

Appendix D. Proof of Proposition 12

Proof From the definition of v-function, it is immediate that V_i is differentiable everywhere except at the three points $n_i^{(1)} = b_i^{(2)}, n_i^{(2)} = b_i^{(1)}$ and $n_i^{(3)} = (1+\eta)b_i^{(1)}$. Let r^* be a minimizer of F. If $r^* \neq n_i^{(j)}$ for every $j \in \{1, 2, 3\}$ and $i \in \{1, \ldots, m\}$, then F must be differentiable at r^* and $F'(r^*) = 0$. Now, let $n^* = \max\{n_i^{(j)} | n_i^{(j)} < r^*\}$. Since F is a linear function over the interval $(n^*, r^*]$, we must have $F'(r) = F'(r^*) = 0$ for every $r \in (n^*, r^*]$. Thus, F reduces to a constant over this interval and continuity of F implies that $F(n^*) = F(r^*)$.

We conclude the proof by showing that n^* is equal to $b_i^{(1)}$ for some i. Suppose this is not the case and let U be an open interval around n^* satisfying $b_i^{(1)} \notin U$ for all i. It is not hard to verify that V_i is a concave function over every interval not containing $b_i^{(1)}$. In particular V_i is concave over U for any i and, as a sum of concave functions, F is concave too over the interval U. Moreover, by definition, n^* minimizes F restricted to U. This implies that F is constant over U as a non-constant concave function cannot reach its minimum over an open set. Finally, let $b^* = \operatorname{argmin}_i |n^* - b_i^{(1)}|$. Since U was an arbitrary open interval, it follows that there exists r arbitrarily close to b^* such that $F(r) = F(n^*)$. By the continuity of F, we must then have $F(b^*) = F(n^*)$. ■