A. Proofs for learning guarantees

A.1. Revenue formula

The simple expression of the expected revenue (2) can be obtained as follows:

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

A.2. Contraction lemma

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

Lemma 8. *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*

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

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:

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

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

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

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

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

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. □

A.3. Bounds on Rademacher complexity

Proposition 9. *For any hypothesis set H and 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

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

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 8, we have the inequality $\widehat{\mathfrak{R}}_S(l_{1H}) \leq \frac{1}{m} \mathbb{E}_{\boldsymbol{\sigma}}[\sup_{h \in H} \sum_{i=1}^m \sigma_i h(\mathbf{x}_i)] = \widehat{\mathfrak{R}}_S(H)$. □

Proposition 10. *Let $M = \sup_{\mathbf{b} \in \mathcal{B}} b^{(1)}$. Then, for any hypothesis set H with pseudo-dimension $d = Pdim(H)$ and 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:*

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

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

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

where for all $i \in [1, m]$, Ψ_i is the M-Lipschitz function $x \mapsto b_i^{(1)} x$. Thus, by Lemma 8 combined with Massart's lemma (see for example (Mohri et al., 2012)), we can write

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

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. □

A.4. Calibration

Theorem 2 (convex surrogates)**.** *There exists no non-constant function $L_c \colon \mathbb{R} \times \mathbb{R}_+ \to \mathbb{R}$ convex with respect to its first argument and satisfying the following conditions:*

- *for any $b_0 \in \mathbb{R}_+$, $\lim_{b \to b_0^-} L_c(b_0, b) = L_c(b_0, b_0)$.*
- *for any distribution D on $\mathbb{R}_+$, there exists a non-negative minimizer $r^* \in \operatorname{argmin}_r \mathbb{E}_{b \sim D}[\widetilde{L}(r, b)]$ such that $\min_r \mathbb{E}_{b \sim D} L_c(r, b) = \mathbb{E}_{b \sim D} L_c(r^*, b)$.*

Proof. For any loss L_c satisfying the assumptions, we can define a loss L'_c by $L'_c(r, b) = L_c(r, b) - L_c(b, b)$. L'_c then also satisfies the assumptions. Thus, without loss of generality, we can assume that $L_c(b, b) = 0$. Furthermore, since $\widetilde{L}(\cdot, b)$ is minimized at b we must have $L_c(r, b) \geq L_c(b, b) = 0$.

Notice that for any $b_1 \in \mathbb{R}+$, $b_1 < b_2 \in \mathbb{R}_+$ and $\mu \in [0, 1]$, the minimizer of $\mathbb{E}_\mu(\widetilde{L}(r, b)) = \mu \widetilde{L}(r, b_1) + (1 - \mu) \widetilde{L}(r, b_2)$ is either b_1 or b_2. In fact, by definition of $\widetilde{L}$, the solution is b_1 as long as $-b_1 \leq -(1 - \mu) b_2$, that is, when $\mu \geq \frac{b_2 - b_1}{b_2}$. Since the minimizing property of L_c should hold for every distribution we must have

$$\begin{aligned}&\mu L_c(b_1, b_1) + (1 - \mu) L_c(b_1, b_2) \\ &\qquad \leq \mu L_c(b_2, b_1) + (1 - \mu) L_c(b_2, b_2) \qquad (11)\end{aligned}$$

when $\mu \geq \frac{b_2 - b_1}{b_2}$ and the reverse inequality otherwise. This implies that (11) must hold as an equality when $\mu = \frac{b_2 - b_1}{b_2}$. This, combined with the equality $L_c(b, b) = 0$ valid for all b, yields

$$b_1 L_c(b_1, b_2) = (b_2 - b_1) L_c(b_2, b_1). \qquad (12)$$

Dividing by $b_2 - b_1$ and taking the limit $b_1 \to b_2$ result in

$$\lim_{b_1 \to b_2^-} b_1 \frac{L_c(b_1, b_2)}{b_2 - b_1} = \lim_{b_1 \to b_2^-} L_c(b_2, b_1). \qquad (13)$$

By convexity of L_c with respect to the first argument, we know that the left-hand side is well-defined and is equal to $-b_1 D_r^- L_c(b_2, b_2)$, where $D_r^- L_c$ denotes the left derivative of L_c with respect to the first coordinate. By assumption, the right-hand side is equal to $L_c(b_2, b_2) = 0$. Since $b_1 > 0$, this implies that $D_r^- L_c(b_2, b_2) = 0$.

Let $\mu < \frac{b_2 - b_1}{b_2}$. For this choice of μ, $\mathbb{E}_\mu(L_c(r, b))$ is minimized at b_2. This implies:

$$\mu D_r^- L_c(b_2, b_1) + (1 - \mu) D_r^- L_c(b_2, b_2) \leq 0. \qquad (14)$$

However, convexity implies that $D_r^- L_c(b_2, b_1) \geq D_r^- L_c(b_1, b_1) = 0$ for $b_2 \geq b_1$. Thus, inequality (14) can only be satisfied if $D_r^- L_c(b_2, b_1) = 0$.

Let $D_r^+ L_c$ denote the right derivative of L_c with respect to the first coordinate. The convexity of L_c implies that $D_r^- L_c(b_1, b_1) \leq D_r^+ L_c(b_1, b_1) \leq D_r^- L_c(b_2, b_1)$ for $b_2 > b_1$. Hence, $D_r^+ L_c(b_1, b_1) = 0$. If we let $\mu > \frac{b_2 - b_1}{b_2}$ then b_1 is a minimizer for $\mathbb{E}_\mu(L_c(r, b))$ and

$$\mu D_r^+(b_1, b_1) + (1 - \mu) D_r^+ L_c(b_1, b_2) \geq 0.$$

As before, since $b_1 < b_2$, $D_r^+(b_1, b_2) \leq D_r^+(b_2, b_2) = 0$ and we must have $D_r^+ L_c(b_1, b_2) = 0$ for this inequality to hold.

We have therefore proven that for every b, if $r \geq b$, then $D_r^- L_c(r, b) = 0$, whereas if $r \leq b$ then $D_r^+ L_c(r, b) = 0$. It is not hard to see that this implies $D_r L_c(r, b) = 0$ for all (r, b) and thus that $L_c(\cdot, b)$ must be a constant. In particular, since $L_c(b, b) = 0$, we have $L_c \equiv 0$. □

Lemma 4. *Let H be a closed, convex subset of a linear space of functions containing 0. Denote by h^*_γ the solution of $\min_{h\in H}\mathcal{L}_\gamma(h)$. If $\sup_{\mathbf{b}\in\mathcal{B}} b^{(1)} = M < \infty$, then*

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

Proof. Let $0<\lambda<1$, because $\lambda h^*_\gamma \in H$ by convexity and h^*_γ is a minimizer we must have:

$$\mathop{\mathbb{E}}_{\mathbf{x},\mathbf{b}}\Big[L_\gamma(h^*_\gamma(\mathbf{x}),\mathbf{b})\Big] \le \mathop{\mathbb{E}}_{\mathbf{x},\mathbf{b}}\Big[L_\gamma(\lambda h^*_\gamma(\mathbf{x}),\mathbf{b})\Big]. \qquad (15)$$

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. If on the other hand $h^*_\gamma(\mathbf{x})>0$, because $\lambda h^*_\gamma(\mathbf{x}) < h^*_\gamma(\mathbf{x})$ 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, because $L_\gamma \le 0$ and $L_\gamma(h^*_\gamma(\mathbf{x}),\mathbf{b}) = 0$ for $(\mathbf{x},\mathbf{b})\in I_4$ it is immediate that $L_\gamma(h^*_\gamma(\mathbf{x}),\mathbf{b}) \ge L_\gamma(\lambda h^*_\gamma(\mathbf{x}),\mathbf{b})$ for $(\mathbf{x},\mathbf{b})\in I_4$. The following inequality holds trivially:

$$\begin{aligned}&\mathop{\mathbb{E}}_{\mathbf{x},\mathbf{b}}\Big[L_\gamma(h^*_\gamma(\mathbf{x}),\mathbf{b})(\mathbb{1}_{I_1}(\mathbf{x})+\mathbb{1}_{I_4}(\mathbf{x}))\Big]\\ &\qquad\ge \mathop{\mathbb{E}}_{\mathbf{x},\mathbf{b}}\Big[L_\gamma(\lambda h^*_\gamma(\mathbf{x}),\mathbf{b})(\mathbb{1}_{I_1}(\mathbf{x})+\mathbb{1}_{I_4}(\mathbf{x}))\Big]. \qquad (16)\end{aligned}$$

Subtracting (16) from (15) we obtain

$$\begin{aligned}&\mathop{\mathbb{E}}_{\mathbf{x},\mathbf{b}}\Big[L_\gamma(h^*_\gamma(\mathbf{x}),\mathbf{b})(\mathbb{1}_{I_2}(\mathbf{x})+\mathbb{1}_{I_3}(\mathbf{x}))\Big]\\ &\qquad\le \mathop{\mathbb{E}}_{\mathbf{x},\mathbf{b}}\Big[L_\gamma(\lambda h^*_\gamma(\mathbf{x}),\mathbf{b})(\mathbb{1}_{I_2}(\mathbf{x})+\mathbb{1}_{I_3}(\mathbf{x}))\Big].\end{aligned}$$

By rearranging terms we can see this inequality is equivalent to

$$\begin{aligned}&\mathop{\mathbb{E}}_{\mathbf{x},\mathbf{b}}\Big[(L_\gamma(\lambda h^*_\gamma(\mathbf{x}),\mathbf{b}) - L_\gamma(h^*_\gamma(\mathbf{x}),\mathbf{b}))\mathbb{1}_{I_2}(\mathbf{x})\Big]\\ &\ge \mathop{\mathbb{E}}_{\mathbf{x},\mathbf{b}}\Big[(L_\gamma(h^*_\gamma(\mathbf{x}),\mathbf{b}) - L_\gamma(\lambda h^*_\gamma(\mathbf{x}),\mathbf{b}))\mathbb{1}_{I_3}(\mathbf{x})\Big] \qquad (17)\end{aligned}$$

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 $\lambda h^*_\gamma(\mathbf{x}) \le b^{(2)}$ then $L_\gamma(\lambda h^*_\gamma(\mathbf{x}),\mathbf{b}) = -b^{(2)} \le -\lambda h^*_\gamma(\mathbf{x})$. Thus

$$\begin{aligned}&\mathbb{E}(L_\gamma(\lambda h^*_\gamma(\mathbf{x}),\mathbf{b}) - L_\gamma(h^*_\gamma(\mathbf{x}),\mathbf{b}))\mathbb{1}_{I_2}(\mathbf{x}))\\ &\qquad\le (1-\lambda)\,\mathbb{E}(h^*_\gamma(\mathbf{x})\mathbb{1}_{I_2}(\mathbf{x})) \qquad (18)\end{aligned}$$

This gives an upper bound for the left-hand side of inequality (17). We now seek to derive a lower bound on the right-hand side. To do that, we analyze two different cases:

1. $\lambda h^*_\gamma(\mathbf{x}) \le b^{(1)}$;
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 $\lambda h^*_\gamma(\mathbf{x}) \le b^{(1)}$, then, by definition $L_\gamma(\lambda h^*_\gamma(\mathbf{x}),\mathbf{b}) = \min(-b^{(2)}, -\lambda h^*_\gamma(\mathbf{x})) \le -\lambda h^*_\gamma(\mathbf{x})$. Thus, we must have:

$$\begin{aligned}&L_\gamma(h^*_\gamma(\mathbf{x}),\mathbf{b}) - L_\gamma(\lambda h^*_\gamma(\mathbf{x}),\mathbf{b})\\ &\qquad > \lambda h^*_\gamma(\mathbf{x}) - b^{(1)} > (\lambda-1)b^{(1)} \ge (\lambda-1)M, \qquad (19)\end{aligned}$$

where we used the fact that $h^*_\gamma(\mathbf{x}) > b^{(1)}$ for the second inequality.

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 (17) as:

$$\begin{aligned}&\mathop{\mathbb{E}}_{\mathbf{x},\mathbf{b}}\Big[\Delta(\mathbf{x},\mathbf{b})\mathbb{1}_{I_3}(\mathbf{x})\Big] =\\ &\mathop{\mathbb{E}}_{\mathbf{x},\mathbf{b}}\Big[\Delta(\mathbf{x},\mathbf{b})\mathbb{1}_{I_3}(\mathbf{x})\mathbb{1}_{\{\lambda h^*_\gamma(\mathbf{x})>b^{(1)}\}}\Big]\\ &+\mathop{\mathbb{E}}_{\mathbf{x},\mathbf{b}}\Big[\Delta(\mathbf{x},\mathbf{b})\mathbb{1}_{I_3}(\mathbf{x})\mathbb{1}_{\{\lambda h^*_\gamma(\mathbf{x})\le b^{(1)}\}}\Big]\\ &\ge \frac{1-\lambda}{\gamma}\mathop{\mathbb{E}}_{\mathbf{x},\mathbf{b}}\Big[h^*_\gamma(\mathbf{x})\mathbb{1}_{I_3}(\mathbf{x})\mathbb{1}_{\{\lambda h^*_\gamma(\mathbf{x})>b^{(1)}\}}\Big]\\ &\qquad + (\lambda-1)M\,\mathbb{P}\Big[h^*_\gamma(\mathbf{x}) > b^{(1)} \ge \lambda h^*_\gamma(\mathbf{x})\Big], \qquad (20)\end{aligned}$$

where we have used (19) to bound the second summand. Combining inequalities (17), (18) and (20) and dividing by $(1-\lambda)$ we obtain the bound

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

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

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

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

A.5. Margin bounds

Theorem 5. *Fix $\gamma \in (0,1]$ and let S denotes 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:*

$$\mathcal{L}_\gamma(h) \le \widehat{\mathcal{L}}_\gamma(h) + \frac{2}{\gamma}\mathfrak{R}_m(H) + M\sqrt{\frac{\log\frac{1}{\delta}}{2m}}. \qquad (21)$$

Proof. Let $\mathcal{L}_{\gamma,H}$ 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 8) as in the proof of Proposition 9 gives $\mathfrak{R}_m(\mathcal{L}_{\gamma,H}) \le \frac{1}{\gamma}\mathfrak{R}_m(H)$. The application of a standard Rademacher complexity bound to the family of functions $\mathcal{L}_{\gamma,H}$ then shows that for any $\delta > 0$, with probability at least $1-\delta$, for any $h \in H$, the following holds:

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

□

We conclude this section by presenting a stronger form of consistency result. We will show that we can lower bound the generalization error of the best hypothesis in class $\mathcal{L}^* := \mathcal{L}(h^*)$ in terms of that of the empirical minimizer of L_γ, $\widehat{h}_\gamma := \operatorname{argmin}_{h\in H}\widehat{\mathcal{L}}_\gamma(h)$.

Theorem 11. *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:*

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

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

$$\mathcal{L}(\widehat{h}_\gamma) \le \widehat{\mathcal{L}}_S(\widehat{h}_\gamma) + 2\mathfrak{R}_m(H) +$$
$$2M\sqrt{\frac{2d\log\frac{em}{d}}{m}} + M\sqrt{\frac{\log\frac{2}{\delta}}{2m}}. \quad (22)$$

Furthermore, applying Lemma 4 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 $\widehat{h}_\gamma$. Finally, the same analysis as the one used in the proof of Theorem 5 shows that with probability $1-\delta/2$,

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

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

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

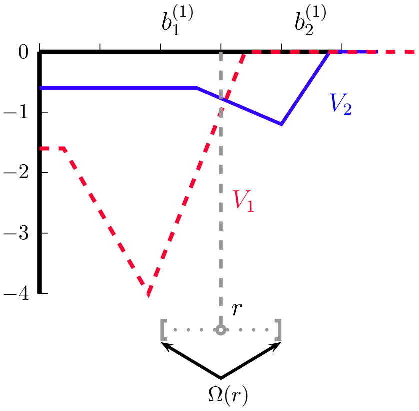

Figure 8. Illustration of the region $\Omega(r)$. The functions V_i are monotonic and concave when restricted to this region.

Combining this with (22) 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 and m (for instance $\gamma \gg 1/m^{1/4}$) it would guarantee the convergence of $\mathcal{L}(\widehat{h}_\gamma)$ to $\mathcal{L}^*$, a stronger form of consistency.

B. Combinatorial algorithm

B.1. Property of the solution

We will show that problem (8) admits a solution $r^* = b_i^{(1)}$ for some i. We will need the following definition.

Definition 12. *For any $r \in \mathbb{R}$, define the following subset of $\mathbb{R}$:*

$$\Omega(r) = \{\epsilon | r < b_i^{(1)} \leftrightarrow r+\epsilon \le b_i^{(1)}\ \forall i\}$$

We will drop the dependency on r when it is understood what value of r we are referring to.

Lemma 13. *Let $r \ne b_i^{(1)}$ for all i. If $\epsilon > 0$ is such that $[-\epsilon,\epsilon] \subset \Omega(r)$ then $F(r+\epsilon) < F(r)$ or $F(r-\epsilon) \le F(r)$.*

The condition that $r \ne b_i^{(1)}$ for all i implies that there exists ϵ small enough that satisfies $\epsilon \in \Omega(r)$.

Proof. Let $v_i = V_i(r,\mathbf{b}_i)$ and $v_i(\epsilon) = V_i(r+\epsilon,\mathbf{b}_i)$. For $\epsilon \in \Omega(r)$ define the sets $D(\epsilon) = \{i \mid v_i(\epsilon) \le v_i\}$ and $I(\epsilon) = \{i \mid v_i(\epsilon) > v_i\}$. If

$$\sum_{i\in D(\epsilon)} v_i + \sum_{i\in I(\epsilon)} v_i > \sum_{i\in D(\epsilon)} v_i(\epsilon) + \sum_{i\in I(\epsilon)} v_i(\epsilon),$$

then, by definition, we have $F(r) > F(r+\epsilon)$ and the result is proven. If this inequality is not satisfied, then, by grouping indices in $D(\epsilon)$ and $I(\epsilon)$ we must have

$$\sum_{i \in D(\epsilon)} v_i - v_i(\epsilon) \leq \sum_{i \in I(\epsilon)} v_i(\epsilon) - v_i \tag{23}$$

Notice that $v_i(\epsilon) \leq v_i$ if and only if $v_i(-\epsilon) \geq v_i$. Indeed, the function $V_i(r+\eta, \mathbf{b}_i)$ is monotone for $\eta \in [-\epsilon, \epsilon]$ as long as $[-\epsilon, \epsilon] \subset \Omega$ which is true by the choice of ϵ. This fact can easily be seen in Figure 8. Hence $D(\epsilon) = I(-\epsilon)$, similarly $I(\epsilon) = D(-\epsilon)$ Furthermore, because $V_i(r+\eta, \mathbf{b}_i)$ is also concave for $\eta \in [-\epsilon, \epsilon]$. We must have

$$\frac{1}{2}(v_i(-\epsilon) + v_i(\epsilon)) \leq v_i. \tag{24}$$

Using (24), we can obtain the following inequalities:

$$v_i(-\epsilon) - v_i \leq v_i - v_i(\epsilon) \qquad \text{for } i \in D(\epsilon) \tag{25}$$

$$v_i(\epsilon) - v_i \leq v_i - v_i(-\epsilon) \qquad \text{for } i \in I(\epsilon). \tag{26}$$

Combining inequalities (25), (23) and (26) we obtain

$$\sum_{i \in D(\epsilon)} v_i(-\epsilon) - v_i \leq \sum_{i \in I(\epsilon)} v_i - v_i(-\epsilon)$$
$$\Rightarrow \sum_{i \in I(-\epsilon)} v_i(-\epsilon) - v_i \leq \sum_{i \in D(-\epsilon)} v_i - v_i(-\epsilon).$$

By rearranging back the terms in the inequality we can easily see that $F(r-\epsilon) \leq F(r)$. □

Lemma 14. *Under the conditions of Lemma 13, if $F(r+\epsilon) \leq F(r)$ then $F(r+\lambda\epsilon) \leq F(r)$ for every λ that satisfies $\lambda\epsilon \in \Omega$ if and only if $\epsilon \in \Omega$.*

Proof. The proof follows the same ideas as those used in the previous lemma. By assumption, we can write

$$\sum_{D(\epsilon)} v_i - v_i(\epsilon) \geq \sum_{i \in I(\epsilon)} v_i(\epsilon) - v_i. \tag{27}$$

It is also clear that $I(\epsilon) = I(\lambda\epsilon)$ and $D(\epsilon) = D(\lambda\epsilon)$. Furthermore, the same concavity argument of Lemma 13 also yields:

$$v_i(\epsilon) \geq \frac{\lambda - 1}{\lambda} v_i + \frac{1}{\lambda} v_i(\lambda\epsilon),$$

which can be rewritten as

$$\frac{1}{\lambda}(v_i - v_i(\lambda\epsilon)) \geq v_i - v_i(\epsilon). \tag{28}$$

Applying inequality (28) in (27) we obtain

$$\frac{1}{\lambda} \sum_{D(\lambda\epsilon)} v_i - v_i(\lambda\epsilon) \geq \frac{1}{\lambda} \sum_{I(\lambda\epsilon)} v_i(\lambda\epsilon) - v_i.$$

Since $\lambda > 0$, we can multiply the inequality by λ to derive an inequality similar to (27) which implies that $F(r+\lambda\epsilon) \leq F(r)$. □

Proposition 7. *Problem (8) admits a solution r^* that satisfies $r^* = b_i^{(1)}$ for some $i \in [1, m]$.*

Proof. Let $r \neq b_i^{(1)}$ for every i. By Lemma 13, we can choose $\epsilon \neq 0$ small enough with $F(r+\epsilon) \leq F(r)$. Furthermore if $\lambda = \min_i \frac{|b_i^{(1)} - r|}{|\epsilon|}$ then λ satisfies the hypotheses of Lemma 14. Hence, $F(r) \geq F(r+\lambda\epsilon) = F(b_{i^*})$, where i^* is the minimizer of $\frac{|b_i^{(1)} - r|}{|\epsilon|}$. □

B.2. Algorithm

We now present a combinatorial algorithm to solve the optimization problem (8) 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 $V(\cdot, b_i)$ 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. 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 for the algorithm.

Proposition 15. *There exists an algorithm to solve the optimization problem (8) in $O(m \log m)$.*

Proof. The pseudocode for the desired algorithm is presented 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:

$$F(n_j) = c_j^{(1)} + c_j^{(2)} n_j + c_j^{(3)} n_j + c_j^{(4)}. \tag{29}$$

This holds trivially for n_1 since by construction $n_1 \leq b_i^{(2)}$ for all i and by definition then $F(n_1) = -\sum_{i=1}^m a_i^{(1)}$. Now, assume that (29) holds for j, we prove that then 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
        $c_j^{(2)} = c_j^{(1)} + a_i^{(2)}$;
        $c_j^{(3)} = c_j^{(3)} + a_i^{(3)}$;
        $c_j^{(4)} = c_j^{(1)} - a_i^{(4)}$;
    else
        $c_j^{(3)} = c_j^{(3)} - a_i^{(3)}$;
        $c_j^{(4)} = c_j^{(1)} + a_i^{(4)}$;
    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 $r \in [n_j, n_{j+1}]$. 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 $n_j = (1+\eta) b_i^{(1)}$.

Let us 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 done in linear time. Once all evaluations are done, finding the minimum can also be done in linear time. Thus, the overall time complexity of the algorithm is $O(m \log m)$. □