A. Proof of Theorem 1

Theorem 1. *Assume $p > 1$. Fix $\rho > 0$. Then, for any $\delta > 0$, with probability at least $1 - \delta$ over the choice of a sample S of size m drawn i.i.d. according to $\mathcal{D}$, the following inequality holds for all $f = \sum_{t=1}^T \alpha_t h_t$:*

$$R(f) \leq \widehat{R}_{S,\rho}(f) + \frac{4}{\rho} \sum_{t=1}^T \alpha_t \mathfrak{R}_m(H_{k_t}) + \frac{2}{\rho}\sqrt{\frac{\log p}{m}} + \sqrt{\left\lceil \frac{4}{\rho^2} \log\left[\frac{\rho^2 m}{\log p}\right]\right\rceil \frac{\log p}{m} + \frac{\log \frac{2}{\delta}}{2m}}.$$

Thus, $R(f) \leq \widehat{R}_{S,\rho}(f) + \frac{4}{\rho}\sum_{t=1}^T \alpha_t \mathfrak{R}_m(H_{k_t}) + C(m,p)$ with $C(m,p) = O\left(\sqrt{\frac{\log p}{\rho^2 m} \log\left[\frac{\rho^2 m}{\log p}\right]}\right)$.

Proof. For a fixed $\mathbf{h} = (h_1, \ldots, h_T)$, any $\boldsymbol{\alpha} \in \Delta$ defines a distribution over $\{h_1, \ldots, h_T\}$. Sampling from $\{h_1, \ldots, h_T\}$ according to $\boldsymbol{\alpha}$ and averaging leads to functions g of the form $g = \frac{1}{n}\sum_{i=1}^T n_t h_t$ for some $\mathbf{n} = (n_1, \ldots, n_T)$, with $\sum_{t=1}^T n_t = n$, and $h_t \in H_{k_t}$.

For any $\mathbf{N} = (N_1, \ldots, N_p)$ with $|\mathbf{N}| = n$, we consider the family of functions

$$G_{\mathcal{F},\mathbf{N}} = \left\{\frac{1}{n}\sum_{k=1}^p \sum_{j=1}^{N_k} h_{k,j} \,|\, \forall (k,j) \in [p] \times [N_k], h_{k,j} \in H_k\right\},$$

and the union of all such families $G_{\mathcal{F},n} = \bigcup_{|\mathbf{N}|=n} G_{\mathcal{F},\mathbf{N}}$. Fix $\rho > 0$. For a fixed $\mathbf{N}$, the Rademacher complexity of $G_{\mathcal{F},\mathbf{N}}$ can be bounded as follows for any $m \geq 1$: $\mathfrak{R}_m(G_{\mathcal{F},\mathbf{N}}) \leq \frac{1}{n}\sum_{k=1}^p N_k \mathfrak{R}_m(H_k)$. Thus, the following standard margin-based Rademacher complexity bound holds (Koltchinskii & Panchenko, 2002). For any $\delta > 0$, with probability at least $1 - \delta$, for all $g \in G_{\mathcal{F},\mathbf{N}}$,

$$R_\rho(g) - \widehat{R}_{S,\rho}(g) \leq \frac{2}{\rho}\frac{1}{n}\sum_{k=1}^p N_k \mathfrak{R}_m(H_k) + \sqrt{\frac{\log \frac{1}{\delta}}{2m}}.$$

Since there are at most p^n possible p-tuples $\mathbf{N}$ with $|\mathbf{N}| = n$, by the union bound, for any $\delta > 0$, with probability at least $1 - \delta$, for all $g \in G_{\mathcal{F},n}$, we can write

$$R_\rho(g) - \widehat{R}_{S,\rho}(g) \leq \frac{2}{\rho}\frac{1}{n}\sum_{k=1}^p N_k \mathfrak{R}_m(H_k) + \sqrt{\frac{\log \frac{p^n}{\delta}}{2m}}.$$

Thus, with probability at least $1 - \delta$, for all functions $g = \frac{1}{n}\sum_{i=1}^T n_t h_t$ with $h_t \in H_{k_t}$, the following inequality holds

$$R_\rho(g) - \widehat{R}_{S,\rho}(g) \leq \frac{2}{\rho}\frac{1}{n}\sum_{k=1}^p \sum_{t: k_t = k} n_t \mathfrak{R}_m(H_{k_t}) + \sqrt{\frac{\log \frac{p^n}{\delta}}{2m}}.$$

Taking the expectation with respect to $\boldsymbol{\alpha}$ and using $\mathrm{E}_{\boldsymbol{\alpha}}[n_t/n] = \alpha_t$, we obtain that for any $\delta > 0$, with probability at least $1 - \delta$, for all $\mathbf{h}$, we can write

$$\mathop{\mathrm{E}}_{\boldsymbol{\alpha}}[R_\rho(g) - \widehat{R}_{S,\rho}(g)] \leq \frac{2}{\rho}\sum_{t=1}^T \alpha_t \mathfrak{R}_m(H_{k_t}) + \sqrt{\frac{\log \frac{p^n}{\delta}}{2m}}.$$

Fix $n \geq 1$. Then, for any $\delta_n > 0$, with probability at least $1 - \delta_n$,

$$\mathop{\mathrm{E}}_{\boldsymbol{\alpha}}[R_{\rho/2}(g) - \widehat{R}_{S,\rho/2}(g)] \leq \frac{4}{\rho}\sum_{t=1}^T \alpha_t \mathfrak{R}_m(H_{k_t}) + \sqrt{\frac{\log \frac{p^n}{\delta_n}}{2m}}.$$

Choose $\delta_n = \frac{\delta}{2p^{n-1}}$ for some $\delta > 0$, then for $p \geq 2$, $\sum_{n \geq 1} \delta_n = \frac{\delta}{2(1-1/p)} \leq \delta$. Thus, for any $\delta > 0$ and any $n \geq 1$, with probability at least $1 - \delta$, the following holds for all $\mathbf{h}$:

$$\mathop{\mathrm{E}}_{\boldsymbol{\alpha}}[R_{\rho/2}(g) - \widehat{R}_{S,\rho/2}(g)] \leq \frac{4}{\rho}\sum_{t=1}^T \alpha_t \mathfrak{R}_m(H_{k_t}) + \sqrt{\frac{\log \frac{2p^{2n-1}}{\delta}}{2m}}. \quad (10)$$

Now, for any $f = \sum_{t=1}^T \alpha_t h_t \in \mathcal{F}$ and any $g = \frac{1}{n}\sum_{i=1}^T n_t h_t$, we can upper bound $R(f) = \Pr_{(x,y)\sim\mathcal{D}}[yf(x) \leq 0]$, the generalization error of f, as follows:

$$\begin{aligned} R(f) &= \Pr_{(x,y)\sim\mathcal{D}}[yf(x) - yg(x) + yg(x) \leq 0] \\ &\leq \Pr[yf(x) - yg(x) < -\rho/2] + \Pr[yg(x) \leq \rho/2] \\ &= \Pr[yf(x) - yg(x) < -\rho/2] + R_{\rho/2}(g). \end{aligned}$$

We can also write

$$\begin{aligned} \widehat{R}_{\rho/2}(g) &= \widehat{R}_{S,\rho/2}(g - f + f) \\ &\leq \widehat{\Pr}[yg(x) - yf(x) < -\rho/2] + \widehat{R}_{S,\rho}(f). \end{aligned}$$

Combining these inequalities yields

$$\begin{aligned} &\Pr_{(x,y)\sim\mathcal{D}}[yf(x) \leq 0] - \widehat{R}_{S,\rho}(f) \\ &\leq \Pr[yf(x) - yg(x) < -\rho/2] \\ &+ \widehat{\Pr}[yg(x) - yf(x) < -\rho/2] + R_{\rho/2}(g) - \widehat{R}_{S,\rho/2}(g). \end{aligned}$$

Taking the expectation with respect to $\boldsymbol{\alpha}$ yields

$$\begin{aligned} R(f) - \widehat{R}_{S,\rho}(f) &\leq \mathop{\mathrm{E}}_{x\sim\mathcal{D},\boldsymbol{\alpha}}[1_{yf(x)-yg(x)<-\rho/2}] + \\ &\mathop{\mathrm{E}}_{x\sim\mathcal{D},\boldsymbol{\alpha}}[1_{yg(x)-yf(x)<-\rho/2}] + \mathop{\mathrm{E}}_{\boldsymbol{\alpha}}[R_{\rho/2}(g) - \widehat{R}_{S,\rho/2}(g)]. \end{aligned}$$

Since $f = \mathrm{E}_{\boldsymbol{\alpha}}[g]$, by Hoeffding's inequality, for any x,

$$\mathop{\mathrm{E}}_{\boldsymbol{\alpha}}[1_{yf(x)-yg(x)<-\rho/2}] = \Pr_{\boldsymbol{\alpha}}[yf(x) - yg(x) < -\rho/2] \leq e^{-\frac{n\rho^2}{8}}$$

$$\mathop{\mathrm{E}}_{\boldsymbol{\alpha}}[1_{yg(x)-yf(x)<-\rho/2}] = \Pr_{\boldsymbol{\alpha}}[yg(x) - yf(x) < -\rho/2] \leq e^{-\frac{n\rho^2}{8}}.$$

Thus, for any fixed $f \in \mathcal{F}$, we can write

$$R(f) - \widehat{R}_{S,\rho}(f) \leq 2e^{-n\rho^2/8} + \operatorname*{E}_{\boldsymbol{\alpha}}[R_{\rho/2}(g) - \widehat{R}_{S,\rho/2}(g)].$$

Thus, the following inequality holds:

$$\sup_{f \in \mathcal{F}} R(f) - \widehat{R}_{S,\rho}(f) \leq 2e^{-n\rho^2/8} + \sup_{\mathbf{h}} \operatorname*{E}_{\boldsymbol{\alpha}}[R_{\rho/2}(g) - \widehat{R}_{S,\rho/2}(g)].$$

Therefore, in view of (10), for any $\delta > 0$ and any $n \geq 1$, with probability at least $1 - \delta$, the following holds for all $f \in \mathcal{F}$:

$$\begin{aligned} & R(f) - \widehat{R}_{S,\rho}(f) \\ & \leq \frac{4}{\rho} \sum_{t=1}^{T} \alpha_t \mathfrak{R}_m(H_{k_t}) + 2e^{-n\rho^2/8} + \sqrt{\frac{\log \frac{2p^{2n-1}}{\delta}}{2m}} \\ & = \frac{4}{\rho} \sum_{t=1}^{T} \alpha_t \mathfrak{R}_m(H_{k_t}) + 2e^{-n\rho^2/8} + \sqrt{\frac{(2n-1)\log p + \log \frac{2}{\delta}}{2m}}. \end{aligned}$$

To select n, we seek to minimize

$$f \colon n \mapsto 2e^{-n\rho^2/8} + \sqrt{\frac{n \log p}{m}} = 2e^{-nu} + \sqrt{nv},$$

with $u = \rho^2/8$ and $v = (\log p)/m$. f is differentiable and for all n, $f'(n) = -2ue^{-nu} + \frac{\sqrt{v}}{2\sqrt{n}}$. The minimum of f is thus for n such that

$$\begin{aligned} f'(n) = 0 \Leftrightarrow 2ue^{-nu} = \frac{\sqrt{v}}{2\sqrt{n}} & \Leftrightarrow -2une^{-2un} = -\frac{v}{8u} \\ & \Leftrightarrow n = \frac{-1}{2u} W_{-1}\Big(\frac{-v}{8u}\Big), \end{aligned}$$

where W_{-1} is the second branch of the Lambert function (inverse of $x \mapsto xe^x$. It is not hard to verify that the following inequalities hold for all $x \in (0, 1/e]$:

$$-\log(x) \leq -W_{-1}(-x) \leq 2\log(x).$$

Bounding $-W_{-1}$ using the lower bound leads to the following choice for n:

$$n = \left\lceil \frac{-1}{2u} \log\Big(\frac{v}{8u}\Big) \right\rceil = \left\lceil \frac{4}{\rho^2} \log\Big(\frac{\rho^2 m}{\log p}\Big) \right\rceil.$$

Plugging in this value of n yields the following bound:

$$\begin{aligned} R(f) - \widehat{R}_{S,\rho}(f) \leq \frac{4}{\rho} \sum_{t=1}^{T} \alpha_t \mathfrak{R}_m(H_{k_t}) + \frac{2}{\rho} \sqrt{\frac{\log p}{m}} \\ + \sqrt{\left\lceil \frac{4}{\rho^2} \log \Big[\frac{\rho^2 m}{\log p}\Big] \right\rceil \frac{\log p}{m} + \frac{\log \frac{2}{\delta}}{2m}}, \end{aligned}$$

which concludes the proof. □

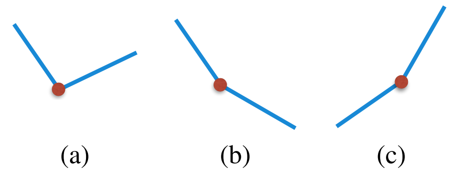

Figure 5. Illustration of the directional derivatives in the three cases of definition (11).

B. Coordinate descent

B.1. Maximum descent coordinate

For a differentiable convex function, the definition of coordinate descent along the direction with maximal descent is standard: the direction selected is the one maximizing the absolute value of the directional derivative. Here, we clarify the definition of the maximal descent strategy for a non-differentiable convex function.

For any function $Q \colon \mathbb{R}^N \to \mathbb{R}$, we denote by $Q'_+(\boldsymbol{\alpha}, \mathbf{e})$ the right directional derivative of Q at $\boldsymbol{\alpha} \in \mathbb{R}^N$ and by $Q'_-(\boldsymbol{\alpha}, \mathbf{e})$ its left directional derivative at $\boldsymbol{\alpha} \in \mathbb{R}^N$ along the direction $\mathbf{e} \in \mathbb{R}^N$, $\|\mathbf{e}\| = 1$, when they exist:

$$\begin{aligned} Q'_+(\boldsymbol{\alpha}, \mathbf{e}) &= \lim_{\eta \to 0^+} \frac{Q(\boldsymbol{\alpha} + \eta \mathbf{e}) - Q(\boldsymbol{\alpha})}{\eta} \\ Q'_-(\boldsymbol{\alpha}, \mathbf{e}) &= \lim_{\eta \to 0^-} \frac{Q(\boldsymbol{\alpha} + \eta \mathbf{e}) - Q(\boldsymbol{\alpha})}{\eta}. \end{aligned}$$

For the remaining of this section, we will assume that Q is a convex function. It is known that in that case these quantities always exist and that $Q'_-(\boldsymbol{\alpha}, \mathbf{e}) \leq Q'_+(\boldsymbol{\alpha}, \mathbf{e})$ for all $\boldsymbol{\alpha}$ and $\mathbf{e}$. The left and right directional derivatives coincide with the directional derivative $Q'(\boldsymbol{\alpha}, \mathbf{e})$ of Q along the direction $\mathbf{e}$ when Q is differentiable at $\boldsymbol{\alpha}$ along the direction $\mathbf{e}$: $Q'(\boldsymbol{\alpha}, \mathbf{e}) = Q'_+(\boldsymbol{\alpha}, \mathbf{e}) = Q'_-(\boldsymbol{\alpha}, \mathbf{e})$.

For any $j \in [1, N]$, let $\mathbf{e}_j$ denote the jth unit vector in $\mathbb{R}^N$. For any $\boldsymbol{\alpha} \in \mathbb{R}^N$ and $j \in [1, N]$, we define the *descent gradient* $\delta Q(\boldsymbol{\alpha}, \mathbf{e}_j)$ *of* Q *along the direction* $\mathbf{e}_j$ as follows:

$$\delta Q(\boldsymbol{\alpha}, \mathbf{e}_j) = \begin{cases} 0 & \text{if } Q'_-(\boldsymbol{\alpha}, \mathbf{e}_j) \leq 0 \leq Q'_+(\boldsymbol{\alpha}, \mathbf{e}_j) \\ Q'_+(\boldsymbol{\alpha}, \mathbf{e}_j) & \text{if } Q'_-(\boldsymbol{\alpha}, \mathbf{e}_j) \leq Q'_+(\boldsymbol{\alpha}, \mathbf{e}_j) \leq 0 \\ Q'_-(\boldsymbol{\alpha}, \mathbf{e}_j) & \text{if } 0 \leq Q'_-(\boldsymbol{\alpha}, \mathbf{e}_j) \leq Q'_+(\boldsymbol{\alpha}, \mathbf{e}_j). \end{cases} \tag{11}$$

$\delta Q(\boldsymbol{\alpha}, \mathbf{e}_j)$ is the element of the subgradient along $\mathbf{e}_j$ that is the closest to 0. Figure 5 illustrates the three cases in that definition. Note that when Q is differentiable along $\mathbf{e}_j$, then $Q'_+(\boldsymbol{\alpha}, \mathbf{e}_j) = Q'_-(\boldsymbol{\alpha}, \mathbf{e}_j)$ and $\delta Q(\boldsymbol{\alpha}, \mathbf{e}_j) = Q'(\boldsymbol{\alpha}, \mathbf{e}_j)$. The maximum descent coordinate can then be defined by

$$k = \operatorname*{argmax}_{j \in [1, N]} |\delta Q(\boldsymbol{\alpha}, \mathbf{e}_j)| \tag{12}$$

This coincides with the standard definition when Q is convex and differentiable.

B.2. Direction

In view of (12), at each iteration $t \geq 1$, the direction $\mathbf{e}_k$ selected by coordinate descent with maximum descent is $k = \operatorname{argmax}_{j\in[1,N]} |\delta Q(\boldsymbol{\alpha}_{t-1}, \mathbf{e}_j)|$. To determine k, we compute $\delta Q(\boldsymbol{\alpha}_{t-1}, \mathbf{e}_j)$ for all $j \in [1, N]$ by distinguishing two cases: $\alpha_{t-1,j} \neq 0$ and $\alpha_{t-1,j} = 0$.

Assume first that $\alpha_{t-1,j} \neq 0$ and let s denote the sign of $\alpha_{t-1,j}$. For η sufficiently small, $\alpha_{t-1,j} + \eta$ has the sign of $\alpha_{t-1,j}$, that is s and

$$F(\boldsymbol{\alpha}_{t-1} + \eta\mathbf{e}_j) = \frac{1}{m}\sum_{i=1}^{m} \Phi\Big(1 - y_i f_{t-1}(x_i) - \eta y_i h_j(x_i)\Big) + \sum_{p\neq j} \Lambda_j |\alpha_{t-1,p}| + s\Lambda_j(\alpha_{t-1,j} + \eta).$$

Thus, when $\alpha_{t-1,j} \neq 0$, F admits a directional derivative along $\mathbf{e}_j$ given by

$$\begin{aligned} F'(\boldsymbol{\alpha}_{t-1}, \mathbf{e}_j) &= -\frac{1}{m}\sum_{i=1}^{m} y_i h_j(x_i)\Phi'\big(1 - y_i f_{t-1}(x_i)\big) + s\Lambda_j \\ &= -\frac{1}{m}\sum_{i=1}^{m} y_i h_j(x_i)\mathcal{D}_t(i)S_t + s\Lambda_j \\ &= (2\epsilon_{t,j} - 1)\frac{S_t}{m} + s\Lambda_j, \end{aligned}$$

and $\delta F(\boldsymbol{\alpha}_{t-1}, \mathbf{e}_j) = (2\epsilon_{t,j} - 1)\frac{S_t}{m} + \operatorname{sgn}(\alpha_{t-1,j})\Lambda_j$. When $\alpha_{t-1,j} = 0$, we find similarly that

$$\begin{aligned} F'_+(\boldsymbol{\alpha}_{t-1}, \mathbf{e}_j) &= (2\epsilon_{t,j} - 1)\frac{S_t}{m} + \Lambda_j \\ F'_-(\boldsymbol{\alpha}_{t-1}, \mathbf{e}_j) &= (2\epsilon_{t,j} - 1)\frac{S_t}{m} - \Lambda_j. \end{aligned}$$

The condition $(F'_-(\boldsymbol{\alpha}, \mathbf{e}_j) \leq 0 \leq F'_+(\boldsymbol{\alpha}, \mathbf{e}_j))$ is equivalent to

$$\Big(-\Lambda_j \leq (2\epsilon_{t,j} - 1)\frac{S_t}{m} \leq \Lambda_j \Big) \Leftrightarrow \left|\epsilon_{t,j} - \frac{1}{2}\right| \leq \frac{\Lambda_j m}{2S_t}.$$

Thus, in summary, we can write, for all $j \in [1, N]$,

$$\delta F(\boldsymbol{\alpha}_{t-1}, \mathbf{e}_j) = \begin{cases} (2\epsilon_{t,j} - 1)\frac{S_t}{m} + \operatorname{sgn}(\alpha_{t-1,j})\Lambda_j & \text{if } (\alpha_{t-1,j} \neq 0) \\ 0 & \text{else if } \left|\epsilon_{t,j} - \frac{1}{2}\right| \leq \frac{\Lambda_j m}{2S_t} \\ (2\epsilon_{t,j} - 1)\frac{S_t}{m} + \Lambda_j & \text{else if } \epsilon_{t,j} - \frac{1}{2} \leq -\frac{\Lambda_j m}{2S_t} \\ (2\epsilon_{t,j} - 1)\frac{S_t}{m} - \Lambda_j & \text{otherwise.} \end{cases}$$

This can be simplified by unifying the last two cases and observing that the sign of $(\epsilon_{t,j} - \frac{1}{2})$ suffices to distinguish between the last two cases:

$$\delta F(\boldsymbol{\alpha}_{t-1}, \mathbf{e}_j) = \begin{cases} (2\epsilon_{t,j} - 1)\frac{S_t}{m} + \operatorname{sgn}(\alpha_{t-1,j})\Lambda_j & \text{if } (\alpha_{t-1,j} \neq 0) \\ 0 & \text{else if } \left|\epsilon_{t,j} - \frac{1}{2}\right| \leq \frac{\Lambda_j m}{2S_t} \\ (2\epsilon_{t,j} - 1)\frac{S_t}{m} - \operatorname{sgn}(\epsilon_{t,j} - \frac{1}{2})\Lambda_j & \text{otherwise.} \end{cases}$$

B.3. Step

Given the direction $\mathbf{e}_k$, the optimal step value η is given by $\operatorname{argmin}_\eta F(\boldsymbol{\alpha}_{t-1} + \eta\,\mathbf{e}_k)$. In the most general case, η can be found via a line search or other numerical methods. In some special cases, we can derive a closed-form solution for the step by minimizing an upper bound on $F(\boldsymbol{\alpha}_{t-1} + \eta\,\mathbf{e}_k)$. For convenience, in what follows, we use the shorthand ϵ_t for $\epsilon_{t,k}$.

Since $y_i h_k(x_i) = \frac{1+y_i h_k(x_i)}{2} \cdot (1) + \frac{1-y_i h_k(x_i)}{2} \cdot (-1)$, by the convexity of $u \mapsto \Phi(1 - \eta u)$, the following holds for all $\eta \in \mathbb{R}$:

$$\begin{aligned} &\Phi\Big(1 - y_i f_{t-1}(x_i) - \eta y_i h_k(x_i)\Big) \\ &\quad\leq \frac{1 + y_i h_k(x_i)}{2}\Phi\Big(1 - y_i f_{t-1}(x_i)) - \eta\Big) \\ &\quad\quad + \frac{1 - y_i h_k(x_i)}{2}\Phi\Big(1 - y_i f_{t-1}(x_i)) + \eta\Big). \end{aligned} \tag{13}$$

Thus, we can write

$$\begin{aligned} &F(\boldsymbol{\alpha}_{t-1} + \eta\mathbf{e}_k) - \sum_{j\neq k} \Lambda_j |\alpha_{t-1,j}| \\ &\leq \frac{1}{m}\sum_{i=1}^{m} \frac{1 + y_i h_k(x_i)}{2}\Phi\Big(1 - y_i f_{t-1}(x_i)) - \eta\Big) \\ &\quad + \frac{1}{m}\sum_{i=1}^{m} \frac{1 - y_i h_k(x_i)}{2}\Phi\Big(1 - y_i f_{t-1}(x_i)) + \eta\Big) \\ &\quad + \Lambda_k |\alpha_{t-1,k} + \eta|. \end{aligned}$$

Let $J(\eta)$ denote that upper bound. We can select η to minimize $J(\eta)$. J is convex and admits a subdifferential at all points. Thus, η^* is a minimizer of $J(\eta)$ iff $0 \in \partial J(\eta^*)$, where $\partial J(\eta^*)$ denotes the subdifferential of J at η^*.

B.4. Exponential loss

In the case $\Phi = \exp$, $J(\eta)$ can be expressed as follows

$$\begin{aligned} J(\eta) &= \frac{1}{m}\sum_{i=1}^{m} \frac{1 + y_i h_k(x_i)}{2} e^{1 - y_i f_{t-1}(x_i)} e^{-\eta} \\ &\quad + \frac{1}{m}\sum_{i=1}^{m} \frac{1 - y_i h_k(x_i)}{2} e^{1 - y_i f_{t-1}(x_i)} e^{\eta} \\ &\quad + \Lambda_k |\alpha_{t-1,k} + \eta|, \end{aligned}$$

and $e^{1 - y_i f_{t-1}(x_i)} = \Phi'(1 - y_i f_{t-1}(x_i)) = S_t \mathcal{D}_t(i)$. Thus, J can be rewritten as follows:[2]

$$J(\eta) = (1 - \epsilon_t)\frac{S_t}{m} e^{-\eta} + \epsilon_t \frac{S_t}{m} e^{\eta} + \Lambda_k |\alpha_{t-1,k} + \eta|,$$

[2]Note that when the functions in H take values in $\{-1, +1\}$, (13) is in fact an equality and $J(\eta)$ coincides with $F(\boldsymbol{\alpha}_{t-1} + \eta\mathbf{e}_t) - \sum_{j\neq k} \Lambda_j |\alpha_{t-1,j}|$.

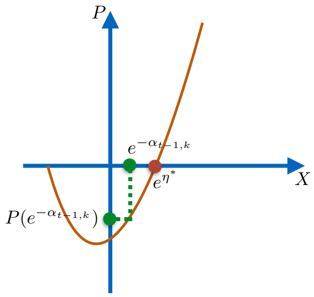

Figure 6. Plot of the polynomial function P.

where we used the shorthand $\epsilon_t = \epsilon_{t,k}$ where k is the index of the direction $\mathbf{e}_k$ selected. If $\alpha_{t-1,k} + \eta^* = 0$, then the subdifferential of $|\alpha_{t-1,k} + \eta|$ at η^* is the set $\{\nu\colon \nu \in [-1, +1]\}$. Thus, $\partial J(\eta^*)$ contains 0 iff there exists $\nu \in [-1, +1]$ such that

$$\begin{aligned} &-(1-\epsilon_t)\frac{S_t}{m}e^{-\eta^*} + \epsilon_t \frac{S_t}{m}e^{\eta^*} + \Lambda_k \nu = 0 \\ \Leftrightarrow &-(1-\epsilon_t)e^{\alpha_{t-1,k}} + \epsilon_t e^{-\alpha_{t-1,k}} + \frac{\Lambda_k m}{S_t}\nu = 0. \end{aligned}$$

This is equivalent to the condition

$$\left|(1-\epsilon_t)e^{\alpha_{t-1,k}} - \epsilon_t e^{-\alpha_{t-1,k}}\right| \leq \frac{\Lambda_k m}{S_t}. \tag{14}$$

If $\alpha_{t-1,k} + \eta^* > 0$, then the subdifferential of $|\alpha_{t-1,k} + \eta|$ at η^* is reduced to $\{1\}$ and $\partial J(\eta^*)$ contains 0 iff

$$\begin{aligned} &-(1-\epsilon_t)e^{-\eta^*} + \epsilon_t e^{\eta^*} + \frac{\Lambda_k m}{S_t} = 0 \\ \Leftrightarrow\ &\epsilon_t e^{2\eta^*} + \frac{\Lambda_k m}{S_t}e^{\eta^*} - (1-\epsilon_t) = 0. \end{aligned} \tag{15}$$

Solving the resulting second-degree equation in e^{η^*} gives

$$e^{\eta^*} = -\frac{\Lambda_k m}{2\epsilon_t S_t} + \sqrt{\left(\frac{\Lambda_k m}{2\epsilon_t S_t}\right)^2 + \frac{1-\epsilon_t}{\epsilon_t}},$$

that is

$$\eta^* = \log\left[-\frac{\Lambda_k m}{2\epsilon_t S_t} + \sqrt{\left(\frac{\Lambda_k m}{2\epsilon_t S_t}\right)^2 + \frac{1-\epsilon_t}{\epsilon_t}}\right].$$

Let P be the second-degree polynomial of (15) whose solution is e^{η^*}. P is convex, has one negative solution, one positive solution, and the positive solution is e^{η^*}. Since $e^{-\alpha_{t-1,k}}$ is positive, the condition $\alpha_{t-1,k} + \eta^* > 0$ or $-\alpha_{t-1,k} < \eta^*$ is then equivalent to $P(e^{-\alpha_{t-1,k}}) < 0$ (see Figure 6), that is

$$\begin{aligned} &\epsilon_t e^{-2\alpha_{t-1,k}} + \frac{\Lambda_k m}{S_t}e^{-\alpha_{t-1,k}} - (1-\epsilon_t) < 0 \\ \Leftrightarrow &(1-\epsilon_t)e^{\alpha_{t-1,k}} - \epsilon_t e^{-\alpha_{t-1,k}} > \frac{\Lambda_k m}{S_t}. \end{aligned} \tag{16}$$

Note that $\eta^* \leq \eta^0$, where $\eta^0 = \log\left[\sqrt{\frac{1-\epsilon_t}{\epsilon_t}}\right]$ is the step size used is AdaBoost.

The case $\alpha_{t-1,k} + \eta^* < 0$ can be treated similarly. It is equivalent to the condition

$$(1-\epsilon_t)e^{\alpha_{t-1,k}} - \epsilon_t e^{-\alpha_{t-1,k}} < -\frac{\Lambda_k m}{S_t}, \tag{17}$$

and leads to the step size

$$\eta^* = \log\left[\frac{\Lambda_k m}{2\epsilon_t S_t} + \sqrt{\left(\frac{\Lambda_k m}{2\epsilon_t S_t}\right)^2 + \frac{1-\epsilon_t}{\epsilon_t}}\right].$$

B.5. Logistic loss

In the case of logistic loss, for any $u \in \mathbb{R}$, $\Phi(-u) = \log_2(1+e^{-u})$ and $\Phi'(-u) = \frac{1}{\log 2}\frac{1}{(1+e^u)}$. To determine the step size, we use the following general upper bound:

$$\begin{aligned} \Phi(-u-v) - \Phi(-u) &= \log_2\left[\frac{1+e^{-u-v}}{1+e^{-u}}\right] \\ &= \log_2\left[\frac{1+e^{-u}+e^{-u-v}-e^{-u}}{1+e^{-u}}\right] \\ &= \log_2\left[1 + \frac{e^{-v}-1}{1+e^u}\right] \\ &\leq \frac{e^{-v}-1}{(\log 2)(1+e^u)} \\ &= \Phi'(-u)(e^{-v}-1). \end{aligned}$$

Thus, we can write

$$\begin{aligned} &F(\boldsymbol{\alpha}_{t-1} + \eta \mathbf{e}_t) - F(\boldsymbol{\alpha}_{t-1}) \\ &\quad\leq \frac{1}{m}\sum_{i=1}^m \Phi'(1 - y_i f_{t-1}(x_i))(e^{-\eta y_i h_k(x_i)} - 1) \\ &\qquad + \Lambda_k(|\alpha_{t-1,k} + \eta| - |\alpha_{t-1,k}|) \\ &\quad= \frac{1}{m}\sum_{i=1}^m \mathcal{D}_t(i) S_t (e^{-\eta y_i h_k(x_i)} - 1) \\ &\qquad + \Lambda_k(|\alpha_{t-1,k} + \eta| - |\alpha_{t-1,k}|). \end{aligned}$$

To determine η, we can minimize this upper bound, or equivalently the following

$$\frac{1}{m}\sum_{i=1}^m \mathcal{D}_t(i) S_t\, e^{-\eta y_i h_k(x_i)} + \Lambda_k |\alpha_{t-1,k} + \eta|.$$

This expression is syntactically the same as the one considered in the case of the exponential loss with only the distribution weights $\mathcal{D}_t(i)$ and S_t being different. Indeed, in the case of the exponential loss ($\Phi = \exp$), we can write

$$F(\boldsymbol{\alpha}_{t-1} + \eta \mathbf{e}_k) - \sum_{j \neq k} \Lambda_j |\alpha_{t-1,j}|$$
$$= \frac{1}{m} \sum_{i=1}^{m} \Phi(1 - y_i f_{t-1}(x_i) - \eta y_i h_k(x_i)) + \Lambda_k |\alpha_{t-1,k} + \eta|,$$
$$= \frac{1}{m} \sum_{i=1}^{m} \Phi(1 - y_i f_{t-1}(x_i))\, e^{-\eta y_i h_k(x_i)} + \Lambda_k |\alpha_{t-1,k} + \eta|,$$
$$= \frac{1}{m} \sum_{i=1}^{m} \Phi'(1 - y_i f_{t-1}(x_i))\, e^{-\eta y_i h_k(x_i)} + \Lambda_k |\alpha_{t-1,k} + \eta|,$$
$$= \frac{1}{m} \sum_{i=1}^{m} D_t(i) S_t\, e^{-\eta y_i h_k(x_i)} + \Lambda_k |\alpha_{t-1,k} + \eta|.$$

Thus, we obtain immediately the same expressions for the step size in the case of the logistic loss with the same three cases, but with $S_t = \sum_{i=1}^m \frac{1}{1+e^{y_i f_{t-1}(x_i)}}$ and $\mathcal{D}_t(i) = \frac{1}{S_t} \frac{1}{1+e^{y_i f_{t-1}(x_i)}}$.

C. Alternative DeepBoost$_\gamma$ algorithm

We also devised and implemented an alternative algorithm, DeepBoost$_\gamma$, which is inspired by the learning bound of Theorem 1 but does not seek to minimize it. The algorithm admits a parameter $\gamma > 0$ representing the edge value demanded at each boosting round. This is the amount by which we require the error ϵ_t of the base hypothesis h_t selected at round t to be better than $\frac{1}{2}$: $\epsilon_t - \frac{1}{2} > \gamma$. We assume given p distinct hypothesis sets with increasing degrees of complexity $H_1, \ldots, H_p$. DeepBoost$_\gamma$ proceeds as if we were running AdaBoost using only as base hypothesis set H_1. But, at each round, if the edge achieved by the best hypothesis found in H_1 is not sufficient, that is if it is not larger than the demanded edge γ, then it selects instead the hypothesis in H_2 with the smallest error on the sample weighted by D_t. If the edge of that hypothesis is also not sufficient, it proceeds with the next hypothesis set and so forth. If the edge is insufficient even with the best hypothesis in H_p, then it just uses the best hypothesis found in $H = \bigcup_{k=1}^p H_k$. The edge parameter γ is determined via cross-validation.

DeepBoost$_\gamma$ is inspired by the bound of Theorem 1 since it seeks to use as much as possible hypotheses from H_1 or lower complexity families and only when necessary functions from more complex families. Since it tends to choose rarely hypotheses from more complex H_ks, the complexity term of the bound of Theorem 1 remains close to the one using only H_1. On the other hand, DeepBoost$_\gamma$ can achieve a smaller empirical margin loss (first term of the bound) by selecting, when needed, more powerful hypotheses than those accessible using H_1 alone.

We carried out some early experiments on several datasets with DeepBoost$_\gamma$ using boosting stumps, in which the performance of the algorithm was found to be superior to that of AdaBoost. A more extensive study of the theoretical and empirical properties of this algorithm are left to the future.

Table 4. Dataset statistics. `german` refers more specifically to the `german (numeric)` dataset.

	`breastcancer`	`ionosphere`	`german`
Examples	699	351	1000
Attributes	9	34	24

	`diabetes`	`ocr17`	`ocr49`
Examples	768	2000	2000
Attributes	8	196	196

	`ocr17-mnist`	`ocr49-mnist`
Examples	15170	13782
Attributes	400	400

D. Additional empirical information

D.1. Dataset sizes and attributes

The size and the number of attributes for the datasets used in our experiments are indicated in Table 4.