Cutoff Averaging: Technical Appendix

A Proof of Thm. 1: A Regret Bound for Margin-Based Perceptron

Proof. Throughout this proof, ℓ denotes the hinge loss. We define $\Delta_i = \|\mathbf{u} - \mathbf{w}_{i-1}\|^2 - \|\mathbf{u} - \mathbf{w}_i\|^2$ and prove the theorem by proving upper and lower bounds on $\sum_{i=1}^m \Delta_i$. Beginning with the upper bound, we notice that $\sum_{i=1}^m \Delta_i$ is a telescopic sum that collapses to

$$\sum_{i=1}^{m} \Delta_i \;=\; \|\mathbf{u} - \mathbf{w}_0\|^2 - \|\mathbf{u} - \mathbf{w}_m\|^2 \;\; .$$

Neglecting $\|\mathbf{u} - \mathbf{w}_m\|^2$ and using the facts that $\mathbf{w}_0 = (0, \ldots, 0)$ and that $\|\mathbf{u}\| \leq 1$, we obtain the upper bound

$$\sum_{i=1}^{m} \Delta_i \;\leq\; 1 \;\; . \tag{5}$$

Moving on to the lower bound, we focus on rounds where $\ell(\mathbf{w}_{i-1}; (\mathbf{w}_i, y_i)) > 0$. We rewrite Δ_i as $\alpha_i + \beta_i$, where

$$\alpha_i \;=\; \|\mathbf{u} - \mathbf{w}_{i-1}\|^2 - \|\mathbf{u} - \mathbf{w}'_{i-1}\|^2 \quad \text{and} \quad \beta_i \;=\; \|\mathbf{u} - \mathbf{w}'_{i-1}\|^2 - \|\mathbf{u} - \mathbf{w}_i\|^2 \;\; .$$

Setting $\eta = 1/(\sqrt{m}R)$, we can rewrite α_i as

$$\begin{aligned} \alpha_i &= \|\mathbf{u} - \mathbf{w}_{i-1}\|^2 - \|\mathbf{u} - \mathbf{w}_{i-1} - \eta y_i \mathbf{x}_i\|^2 \\ &= 2\eta \left(y_i \langle \mathbf{u}, \mathbf{x}_i \rangle - y_i \langle \mathbf{w}_{i-1}, \mathbf{x}_i \rangle \right) \;-\; \eta^2 \|\mathbf{x}_i\|^2 \;\; , \end{aligned} \tag{6}$$

where the first inequality follows from the definition of $\mathbf{w}'_{i-1}$ and the second equality is straightforward linear algebra. Next, we combine the term in Eq. (6) with three additional facts: (1) by assumption $\|\mathbf{x}\| \leq R$, (2) by the assumption that $\ell(\mathbf{w}_{i-1}; (\mathbf{x}_i, y_i)) > 0$ and using the definition of the hinge loss, we have $\ell(\mathbf{w}_{i-1}; (\mathbf{x}_i, y_i)) = 1 - y_i \langle \mathbf{w}_{i-1}, \mathbf{x}_i \rangle$, and (3) by the definition of the hinge loss $\ell(\mathbf{u}; (\mathbf{x}_i, y_i)) \geq 1 - y_i \langle \mathbf{u}, \mathbf{x}_i \rangle$. We obtain the lower bound

$$\alpha_i \;\geq\; 2\eta \Big(-\ell\big(\mathbf{u}; (\mathbf{x}_i, y_i)\big) + \ell\big(\mathbf{w}_{i-1}; (\mathbf{x}_i, y_i)\big) \Big) \;-\; \eta^2 R^2 \;\; .$$

Next we prove that β_i is always non-negative. If $1 \leq \frac{1}{\|\mathbf{w}'_{i-1}\|}$ then this claim is an immediate consequence of the definition of $\mathbf{w}_i$. Otherwise, it holds that $\|\mathbf{w}'_{i-1}\| \geq 1$, $\mathbf{w}_i = \mathbf{w}'_{i-1}/\|\mathbf{w}'_{i-1}\|$, and we have that

$$\begin{aligned} \beta_i &= \|\mathbf{u} - \mathbf{w}'_{i-1}\|^2 - \left\| \mathbf{u} - \frac{\mathbf{w}'_{i-1}}{\|\mathbf{w}'_{i-1}\|} \right\|^2 \\ &= -2 \left(1 - \frac{1}{\|\mathbf{w}'_{i-1}\|} \right) \langle \mathbf{u}, \mathbf{w}'_{i-1} \rangle \;+\; \|\mathbf{w}'_{i-1}\|^2 \;-\; 1 \;\; . \end{aligned} \tag{7}$$

Using the Cauchy-Schwartz inequality and the assumption that $\|\mathbf{u}\| \leq 1$, we lower bound the term

$$-2 \left(1 - \frac{1}{\|\mathbf{w}'_{i-1}\|} \right) \langle \mathbf{u}, \mathbf{w}'_{i-1} \rangle$$

with $-2\|\mathbf{w}'_{i-1}\| + 2$. Plugging this lower bound into Eq. (7) gives

$$\beta_i \;\geq\; 1 - 2\|\mathbf{w}'_{i-1}\| + \|\mathbf{w}'_{i-1}\|^2 \;=\; \left(1 - \|\mathbf{w}'_{i-1}\|\right)^2 \;\; .$$

We have proven that β_i is non-negative and we conclude that

$$\Delta_i \;\geq\; 2\eta \Big(-\ell\big(\mathbf{u}; (\mathbf{x}_i, y_i)\big) + \ell\big(\mathbf{w}_{i-1}; (\mathbf{x}_i, y_i)\big) \Big) \;-\; \eta^2 R^2 \;\; . \tag{8}$$

Note that the above holds trivially whenever $\ell(\mathbf{w}_{i-1}; (\mathbf{x}_i, y_i)) = 0$, and therefore the above holds for all i. Summing Δ_i over all i, we get

$$\sum_{i=1}^{m} \Delta_i \;\geq\; -2\eta \sum_{1=1}^{m} \ell\big(\mathbf{u}; (\mathbf{x}_i, y_i)\big) + 2\eta \sum_{1=1}^{m} \ell\big(\mathbf{w}_{i-1}; (\mathbf{x}_i, y_i)\big) \;-\; m\eta^2 R^2 \;\; .$$

Comparing the above to the upper bound in Eq. (5) and rearranging terms gives the bound

$$\frac{1}{m} \sum_{1=1}^{m} \ell\big(\mathbf{w}_{i-1}; (\mathbf{x}_i, y_i)\big) \;\leq\; \frac{1}{m} \sum_{1=1}^{m} \ell\big(\mathbf{u}; (\mathbf{x}_i, y_i)\big) \;+\; \frac{1}{2m\eta} + \frac{\eta R^2}{2} \;\; .$$

Recalling that $\eta = 1/(\sqrt{m}R)$ proves the bound. □

B Proof of Lemma 1: An Adaptation of Freedman's Bound

The following is a detailed proof of Lemma 1. We show that the lemma is a direct corollary from Freedman's tail bound for martingales [6]. This proof is adapted from the work of Cesa-Bianchi and Gentile in [3, Proposition 2] with two exceptions: First we use the full power of Freedman's theorem and prove a Kolmogorov-type maximal inequality, namely, an inequality that holds uniformly for any prefix of the random variable sequence. Second, we build on Freeman's original bound, as it appears in [6], rather than the slightly different version used in [3].

One of the straightforward techniques used in our proof is the *square root trick*. There is really nothing tricky about this elementary technique: it involves finding the positive root of a second-degree polynomial, in order to satisfy a quadratic constraint. The term "square root trick" has been coined elsewhere and we stick with this name.

Lemma 2. *Let b and c be positive numbers. Then,*

(1) $x^2 - bx - c > 0$ *and* $x \geq 0 \quad \iff \quad x > \frac{b+\sqrt{b^2+4c}}{2}$

(2) $x^2 - bx - c < 0$ *and* $x \geq 0 \quad \iff \quad 0 \leq x < \frac{b+\sqrt{b^2+4c}}{2}$

Proof. The left-hand side of (1) above is a second degree polynomial in x with a positive leading term, one negative root N and one positive root P. Therefore, it is positive in the region $(-\infty, N) \cup (P, \infty)$. Intersecting this constraint with $x \geq 0$, gives $x > P$. Equivalently, the left-hand side of (2) is negative between N and P. Intersecting this constraint with $x \geq 0$ results in the constraint $0 \leq x < P$. In both cases, the value of P can be calculated using the quadratic formula. □

For completeness, we give Freedman's original theorem:

Theorem 3 (Freedman, [6])**.** *Let $(A_i)_{i=0}^m$ be a martingale with respect to $(Z_i)_{i=1}^m$. Let $B_i = A_i - A_{i-1}$ be the corresponding sequence of martingale differences and let $D_i = \mathrm{Var}[B_i|(Z_j)_{j=1}^{i-1}]$ be the corresponding sequence of conditional variances. Assume $|B_i| \leq 1$ for all i. For any positive numbers a and b,*

$$\Pr\left(\exists t \quad \sum_{i=1}^{t} B_i \geq a \;, \quad \sum_{i=1}^{t} D_i \leq b\right) \;\leq\; \exp\left(-\frac{a^2}{2(a+b)}\right) \;.$$

We are now ready to prove Lemma 1.

Proof of Lemma 1. Define, for all $i \in \{1, \ldots, m\}$

$$B_i \;=\; \frac{U_i - L_i}{C} \quad \text{and} \quad V_i \;=\; \mathrm{Var}\left[B_i \middle| (Z_j)_{j=1}^{i-1}\right] \;.$$

Note that $(B_i)_{i=1}^m$ is a sequence of martingale differences with respect to $(Z_i)_{i=1}^m$, and that $|B_i| \leq 1$ for all i. For brevity, define $\alpha = \ln(\frac{m}{\delta})$. We begin by examining the probability

$$\Pr\left(\exists t \quad \sum_{i=1}^{t} B_i \;\geq\; \alpha + \sqrt{\alpha^2 + 2\alpha\left(1 + \textstyle\sum_{i=1}^{t} V_i\right)}\right) \;.$$

Since $|B_i| \leq 1$, it holds that $\sum_{i=1}^m V_i \leq m$. Therefore, we can upper-bound the above by

$$\sum_{s=1}^{m} \Pr\left(\exists t \quad \sum_{i=1}^{t} B_i \;\geq\; \alpha + \sqrt{\alpha^2 + 2\alpha s}\;, \quad \sum_{i=1}^{t} V_i \leq s\right) \;.$$

Each summand above satisfies the requirements of Freedman's bound, Thm. 3. Applying the theorem for each summand gives the upper bound

$$\sum_{s=1}^{m} \exp\left(-\frac{\left(\alpha + \sqrt{\alpha^2 + 2\alpha s}\right)^2}{2\left(\alpha + \sqrt{\alpha^2 + 2\alpha s} + s\right)}\right) \;=\; \sum_{s=1}^{m} \exp(-\alpha) \;=\; \delta \;.$$

Overall, we have proven that, with probability at least $1-\delta$, it holds that

$$\forall\, t \quad \sum_{i=1}^{t} B_i \;<\; \alpha + \sqrt{\alpha^2 + 2\alpha(1+\textstyle\sum_{i=1}^{t} V_i)}\ , \tag{9}$$

Given a concrete value of $(Z_j)_{j=1}^{i-1}$, U_i is just a constant and does not effect the variance. Therefore,

$$V_i \;=\; \frac{\mathrm{Var}\big[L_i \big| (Z_j)_{j=1}^{i-1}\big]}{C^2} \;\leq\; \frac{\mathbb{E}\big[L_i^2 \big| (Z_j)_{j=1}^{i-1}\big]}{C^2} \;\leq\; \frac{\mathbb{E}\big[L_i \big| (Z_j)_{j=1}^{i-1}\big]}{C} \;=\; \frac{U_i}{C}\ .$$

where the first inequality follows from the definition of variance, the second inequality follows from the fact that $L_i \in [0, C]$, and the last equality uses the definition of U_i. Plugging this bound into Eq. (9), we have

$$\forall\, t \quad \sum_{i=1}^{t} B_i \;<\; \alpha + \sqrt{\alpha^2 + 2\alpha\left(1 + \tfrac{1}{C}\textstyle\sum_{i=1}^{t} U_i\right)}\ .$$

Using the definition of B_i and the fact that $\sqrt{a+b} \leq \sqrt{a} + \sqrt{b}$, we have

$$\forall\, t \quad \tfrac{1}{C}\textstyle\sum_{i=1}^{t} U_i - \tfrac{1}{C}\sum_{i=1}^{t} L_i \;<\; 2\alpha + \sqrt{2\alpha\left(1 + \tfrac{1}{C}\sum_{i=1}^{t} U_i\right)}\ .$$

Focus for a moment on one value of t. Substituting $\gamma = \sqrt{1 + \frac{1}{C}\sum_{i=1}^{t} U_1}$ and $\lambda = \frac{1}{C}\sum_{i=1}^{t} L_i$, the above can be rewritten as the following quadratic constraint on γ

$$\gamma^2 - \sqrt{2\alpha}\gamma - (2\alpha + \lambda + 1) \;<\; 0\ .$$

Using the square-root trick, outlined in Lemma 2, the above is equivalent to

$$\gamma \;<\; \frac{\sqrt{2\alpha} + \sqrt{10\alpha + 4\lambda + 4}}{2}\ .$$

Taking the square of both sides above, we get

$$\gamma^2 \;<\; 3\alpha + \lambda + 1 + \sqrt{5\alpha^2 + 2\alpha\lambda + 2\alpha}\ .$$

Once again using the inequality $\sqrt{a+b+c} \leq \sqrt{a} + \sqrt{b} + \sqrt{c}$, we get

$$\gamma^2 \;<\; \lambda + (3+\sqrt{5})\alpha + \sqrt{2\alpha\lambda} + \sqrt{2\alpha}\ .$$

Finally, assuming $m \geq 4$ we have that $\alpha > \sqrt{\alpha}$ and therefore

$$\gamma^2 \;<\; \lambda + (3+\sqrt{5}+\sqrt{2})\alpha + \sqrt{2\alpha\lambda}\ .$$

Plugging in the definitions of γ and λ and using $3 + \sqrt{5} + \sqrt{2} < 7$ concludes the proof. □