Appendix [Margins, Kernels and Non-linear Smoothed Perceptrons]

1. Unified Proof By Induction of Lemma 5, 8: $L_{\mu_k}(\alpha_k) \le -\frac{1}{2}\|p_k\|_G^2$

Let $d(p)$ be 1-strongly convex with respect to the #-norm, ie $d(q)-d(p)-\langle \nabla d(p), q-p\rangle \ge \frac{1}{2}\|q-p\|_\#^2$ for any $p,q \in \Delta_n$. Let the #-norm be lower bounded by the G-norm as $\|p\|_G^2 \le \lambda_\# \|p\|_\#^2$. For $d(p)=\sum_i p_i \log p_i + \log n$, # is the 1-norm, $\lambda_\# = 1$ and $p^* = \frac{\mathbf{1}_n}{n}$. For $d(p) = \frac{1}{2}\|q-p\|_2^2$, # is the 2-norm, $\lambda_\# = n$ and $p^* = q$. Choose $\mu_0 = 2\lambda_\#$.

Let the smoothed minimizer be defined by $p_\mu(\alpha) := \arg\min_{p\in\Delta_n} \langle G\alpha, p\rangle + \mu d(p)$, and $p^* := \arg\min_{p\in\Delta_n} d(p)$. The optimality condition of $p_\mu(\alpha)$ and p^* (the gradient is perpendicular to any feasible direction) is that for any $r \in \Delta_n$,

$$\begin{aligned} \langle G\alpha + \mu\nabla d(p_\mu(\alpha)), r-p\rangle &= 0 && (1)\\ \langle \nabla d(p^*), r-p\rangle &= 0 \Rightarrow d(p_0) \ge \tfrac{1}{2}\|p_0-p^*\|_\#^2. && (2) \end{aligned}$$

$$\begin{aligned} \text{For } k=0: \quad -\tfrac{1}{2}\|p_0\|_G^2 &= -\tfrac{1}{2}\|p_0-p^*\|_G^2 - \langle p^*, p_0-p^*\rangle_G - \tfrac{1}{2}\|p^*\|_G^2 && \text{writing } p_0 = (p_0-p^*)+p^*\\ &\ge -\tfrac{\lambda_\#}{2}\|p_0-p^*\|_\#^2 - \langle p^*, p_0\rangle_G + \tfrac{1}{2}\|p^*\|_G^2 && \text{using } \|p\|_G^2 \le \lambda_\#\|p\|_\#^2\\ &\ge -\mu_0 d(p_0) - \langle \alpha_0, p_0\rangle_G + \tfrac{1}{2}\|\alpha_0\|_G^2 && \text{adding } -\tfrac{\lambda_\#}{2}\|p_0-p^*\|_1^2\text{, using Eq. (2)}\\ &= L_{\mu_0}(\alpha_0). \end{aligned}$$

Assume it holds upto k. We drop index k, and write x_+ for x_{k+1}. Let $\hat{p} = (1-\theta)p + \theta p_\mu(\alpha)$ so $\alpha_+ = (1-\theta)\alpha + \theta\hat{p}$. (3)

$$\begin{aligned} L_{\mu_+}(\alpha_+) &= \tfrac{1}{2}\|\alpha_+\|_G^2 - \Big\langle \alpha_+, p_{\mu_+}(\alpha_+)\Big\rangle_G - \mu_+ d(p_{\mu_+}(\alpha_+))\\ &= \tfrac{1}{2}\Big\|(1-\theta)\alpha + \theta\hat{p}\Big\|_G^2 - \theta\Big\langle \hat{p}, p_{\mu_+}(\alpha_+)\Big\rangle_G - (1-\theta)\Big[\Big\langle \alpha, p_{\mu_+}(\alpha_+)\Big\rangle_G + \mu d(p_{\mu_+}(\alpha_+))\Big] \quad \text{using Eq. (3)}\\ &\le (1-\theta)\Big[\tfrac{1}{2}\|\alpha\|_G^2 - \Big\langle \alpha, p_{\mu_+}(\alpha_+)\Big\rangle_G - \mu d(p_{\mu_+}(\alpha_+))\Big]_1 + \theta\Big[-\tfrac{1}{2}\|\hat{p}\|_G^2 - \Big\langle \hat{p}, p_{\mu_+}(\alpha_+) - \hat{p}\Big\rangle_G\Big], \end{aligned}$$

where we used the convexity of $\|.\|_G^2$. Recall $p_+ = (1-\theta)p + \theta p_{\mu_+}(\alpha_+)$, so that $p_+ - \hat{p} = \theta(p_{\mu_+}(\alpha_+) - p_\mu(\alpha))$. (4)

$$\begin{aligned} \Big[\cdot\Big]_1 &= \Big[\tfrac{1}{2}\|\alpha\|_G^2 - \Big\langle \alpha, p_\mu(\alpha)\Big\rangle_G - \mu d(p_\mu(\alpha))\Big] - \Big\langle \alpha, p_{\mu_+}(\alpha_+) - p_\mu(\alpha)\Big\rangle_G - \mu\Big[d(p_{\mu_+}(\alpha_+)) - d(p_\mu(\alpha))\Big]\\ &= L_\mu(\alpha) - \mu\Big[d(p_{\mu_+}(\alpha_+)) - d(p_\mu(\alpha)) - \Big\langle \nabla d(p_\mu(\alpha)), p_{\mu_+}(\alpha_+) - p_\mu(\alpha)\Big\rangle\Big] && \text{using Eq. (1)}\\ &\le -\tfrac{1}{2}\|p\|_G^2 - \tfrac{\mu}{2}\|p_{\mu_+}(\alpha_+) - p_\mu(\alpha)\|_\#^2 && \text{using strong convexity of } d(p)\\ &\le -\tfrac{1}{2}\|\hat{p} + (p-\hat{p})\|_G^2 - \tfrac{\mu}{2\lambda_\#}\|p_{\mu_+}(\alpha_+) - p_\mu(\alpha)\|_G^2 && \text{using } \|p\|_G^2 \le \lambda_\#\|p\|_\#^2\\ &\le -\tfrac{1}{2}\|\hat{p}\|_G^2 - \Big\langle \hat{p}, p-\hat{p}\Big\rangle_G - \frac{\mu}{2\lambda_\#\theta^2}\|p_+ - \hat{p}\|_G^2 && \text{using Eq. (4) and dropping a } -\tfrac{1}{2}\|p-\hat{p}\|_G^2 \text{ term.} \end{aligned}$$

Using $(1-\theta)(p-\hat{p}) = -\theta(p_\mu(\alpha) - \hat{p})$ and substituting back,

$$\begin{aligned} L_{\mu_+}(\alpha_+) &\le (1-\theta)\Big[-\tfrac{1}{2}\|\hat{p}\|_G^2 + \tfrac{\theta}{1-\theta}\Big\langle \hat{p}, p_\mu(\alpha) - \hat{p}\Big\rangle_G - \frac{\mu}{2\lambda_\#\theta^2}\|p_+ - \hat{p}\|_G^2\Big] + \theta\Big[-\tfrac{1}{2}\|\hat{p}\|_G^2 - \Big\langle \hat{p}, p_{\mu_+}(\alpha_+) - \hat{p}\Big\rangle_G\Big]\\ &= -\tfrac{1}{2}\|\hat{p}\|_G^2 - \theta\Big\langle \hat{p}, p_{\mu_+}(\alpha_+) - p_\mu(\alpha)\Big\rangle_G - \frac{\mu(1-\theta)}{2\lambda_\#\theta^2}\|p_+ - \hat{p}\|_G^2\\ &\le -\tfrac{1}{2}\|\hat{p}\|_G^2 - \Big\langle \hat{p}, p_+ - \hat{p}\Big\rangle_G - \tfrac{1}{2}\|p_+ - \hat{p}\|_G^2 && \text{using Eq. (4) and } \tfrac{\theta^2}{1-\theta} = \tfrac{4}{(k+1)(k+3)} \le \tfrac{4}{(k+1)(k+2)} = \tfrac{\mu}{\lambda_\#}\\ &= -\tfrac{1}{2}\|p_+\|_G^2. \end{aligned}$$

This wraps up our unified proof for both settings.

2. Kaczmarz Perceptron (modified modified perceptron)

Start with $f_0 = \sum_i y_i \tilde{\phi}_{x_i}/n$ and $Z_0 = \|f_0\|_K$. So $Z_0 \geq \langle f_0, f^* \rangle_K \geq \rho_K$.

Halt when

$$|y_i f_k(x_i)| \leq \sigma Z_k \leq \sigma Z_0$$

for all mistakes.

Otherwise, for the worst mistake, update

$$f_{k+1} = f_k - y_i f_k(x_i)(y_i \tilde{\phi}_{x_i})$$

Now,

$$Z_{k+1} \geq \langle f_{k+1}, f^* \rangle = \langle f_k, f^* \rangle + \rho_K |y_i f_k(x_i)| \geq \langle f_k, f^* \rangle + \rho_K \sigma Z_0 = \rho_K(1 + \sigma \rho_K)$$

Also

$$Z_{k+1}^2 = Z_k^2 - |y_i f_k(x_i)|^2 \leq Z_k^2(1 - \sigma^2)$$

Assume we get

$$\rho_K(1 + \sigma \rho_K)^n \leq Z_0(1 - \sigma^2)^{n/2}$$

Hence

$$\rho_K e^{n\sigma\rho_K} \leq Z_0 e^{-\sigma^2 n/2}$$

Hence it halts before

$$n\sigma\rho_K \geq \log(Z_0/\sigma) - n\sigma^2/2$$