A Additional Proof Details

This section describes a functional boosting view of selecting features for generalized linear models of one-dimensional response. We then prove Lemma 3.3 and Lemma 3.4 for this more general setting. These more general results in turn extend Theorem 3.2 to generalized linear models.

A.1 Functional Boosting View of Feature Selection

We view each feature f as a function h_f that maps sample x to x_f. We define $f_S : \mathbb{R}^D \to \mathbb{R}$ to be the best linear predictor using features in S, i.e., $f_S(x) \triangleq w(S)^T x_S$. For each feature dimension $d \in D$, the coefficient of d is in $w(S)$ is $w(S)_d = f_S(e_d)$, where e_d is the d^{th} dimensional unit vector. So $\|w(S)\|_2^2 = \sum_{d=1}^D \|f_S(e_d)\|_2^2$. Given a generalized linear model with link function $\nabla\Phi$, the predictor is $E[y|x] = \nabla\Phi(w^T x)$ for some w and the calibrated loss is $r(w) = \sum_{i=1}^n (\Phi(w^T x_i) - y_i w^T x_i)$. Replacing $f_S(x_i) = w(S)^T x_i$, we have

$$r(w(S)) = \sum_{i=1}^n (\Phi(f_S(x_i)) - y_i f_S(x_i)). \quad (13)$$

Note that the risk function in Equation 1 can be rewritten as the following to resemble Equation 13:

$$R(S) = \mathcal{R}[f_S] = \frac{1}{n}\sum_{i=1}^n (\Phi(f_S(x_i)) - y_i^T f_S(x_i)) + \frac{\lambda}{2}\sum_{d=1}^D \|f_S(e_d)\|_2^2 + A, \quad (14)$$

where $\phi(x) = \frac{1}{2}x^2$ for linear predictions and constant $A = \frac{1}{2n}\sum_{i=1}^n y_i^2$. Next we define the inner product between two functions $f, h : \mathbb{R}^D \to \mathbb{R}$ over the training set to be:

$$\langle f, h\rangle \triangleq \frac{1}{n}\sum_{i=1}^n f(x_i)h(x_i) + \frac{\lambda}{2}\sum_{d=1}^D f(e_d)h(e_d). \quad (15)$$

With this definition of inner product, we can compute the derivative of $\mathcal{R}$:

$$\nabla\mathcal{R}[f] = \sum_{i=1}^n (\nabla\Phi(f(x_i)) - y_i)\delta_{x_i} + \sum_{d=1}^D f(e_d)\delta_{e_d}, \quad (16)$$

where $\nabla\phi(x) = x$ for linear predictions, and δ_x is an indicator function for x. Then the gradient of objective $F(S)$ w.r.t coefficient w_f of a feature dimension d can be written as:

$$b_d^S = -\frac{1}{n}\sum_{i=1}^n (\nabla\Phi_p(w(S)^T x^i) - y^i)x_d^i - \lambda w(S)_d \quad (17)$$

$$= -\langle \nabla\mathcal{R}[f_S], h_d\rangle. \quad (18)$$

In addition, the regularized covariance matrix of features C satisfies,

$$C_{ij} = \frac{1}{n}X_i^T X_j + \lambda I(i=j) = \langle h_i, h_j\rangle, \quad (19)$$

for all $i, j = 1, 2, ..., D$. So in this functional boosting view, Algorithm 1 greedily chooses group g that maximizes, with a slight abuse of notation of $\langle\ ,\ \rangle$, $\|\langle h_g, \nabla\mathcal{R}[f_S]\rangle\|_2^2 / c(g)$, i.e., the ratio between similarity of a feature group and the functional gradient, measured in sum of square of inner products, and the cost of the group

A.2 Proof of Lemma 3.3 and Lemma 3.4

The more general version of Lemma 3.3 and Lemma 3.4 assumes that the objective functional $\mathcal{R}$ is m-strongly smooth and M-strongly convex using our proposed inner product rule. M-strong convexity is a reasonable assumption, because the regularization term $\|w\|_2^2 = \sum_{d=1}^D \|f_S(e_d)\|_2^2$ ensures that all loss functional $\mathcal{R}$ with a convex Φ strongly convex. In the linear prediction case, both m and M equals 1.

The following two lemmas are the more general versions of Lemma 3.3 and Lemma 3.4.

Lemma A.1. *Let $\mathcal{R}$ be an m-strongly smooth functional with respect to our definition of inner products. Let S and G be some fixed sequences. Then*

$$F(S) - F(G) \le \frac{1}{2m}\langle b_{G\oplus S}^G, C_{G\oplus S}^{-1} b_{G\oplus S}^G\rangle$$

Proof. First we optimize over the weights in S.

$$\begin{aligned} &F(S) - F(G) \\ &= \mathcal{R}[f_G] - \mathcal{R}[f_S] = \mathcal{R}[f_G] - \mathcal{R}[\sum_{s\in S}\alpha_s^T h_s] \\ &\le \mathcal{R}[f_G] - \min_{w: w_i^T \in \mathbb{R}^{d_{s_i}}, s_i\in S} \mathcal{R}[\sum_{s_i\in S} w_{s_i}^T h_{s_i}] \end{aligned}$$

Adding dimensions in G will not increase the risk, we have:

$$\le \mathcal{R}[f_G] - \min_{w: w_i\in\mathbb{R}^{d_{s_i}}, s_i\in G\oplus S} \mathcal{R}[\sum_{s_i\in G\oplus S} w_{s_i} h_{s_i}]$$

Since $f_G = \sum_{g_i\in G}\alpha_i h_{g_i}$, we have:

$$\le \mathcal{R}[f_G] - \min_w \mathcal{R}[f_G + \sum_{s_i\in G\oplus S} w_i^T h_{s_i}]$$

Expanding using strong smoothness around f_G, we have:

$$\le \mathcal{R}[f_G] - \min_w(\mathcal{R}[f_G] + \langle\nabla\mathcal{R}[f_G], \sum_{s_i\in G\oplus S} w_i^T h_{s_i}\rangle$$

$$+\frac{m}{2}\|\sum_{s_i\in G\oplus S} w_i^T h_{s_i}\|_2^2)$$

$$=\max_w -\langle\nabla\mathcal{R}[f_G], \sum_{s_i\in G\oplus S} w_i^T h_{s_i}\rangle - \frac{m}{2}\|\sum_{s_i\in G\oplus S} w_i^T h_{s_i}\|_2^2$$

$$=\max_w \langle b^G_{G\oplus S}, w\rangle - \frac{m}{2}\langle w, C_{G\oplus S} w\rangle$$

Solving w directly we have:

$$F(S) - F(G) \leq \frac{1}{2m}\langle b^G_{G\oplus S}, C^{-1}_{G\oplus S} b^G_{G\oplus S}\rangle$$

□

Lemma A.2. *Let $\mathcal{R}$ be a M-strongly convex functional with respect to our definition of inner products. Then*

$$F(G_j) - F(G_{j-1}) \geq \frac{1}{2M(1+\lambda)}\langle b^{G_{j-1}}_{g_j}, b^{G_{j-1}}_{g_j}\rangle \quad (20)$$

Proof. After the greedy algorithm chooses some group g_j at step j, we form $f_{G_j} = \sum_{\alpha_i}\alpha_i^T h_{g_i}$, such that

$$\mathcal{R}[f_G] = \min_{\alpha_i\in\mathbb{R}^{d_{g_i}}} \mathcal{R}[\sum_{g_i\in G_j}\alpha_i^T h_{g_i}] \leq \min_{\beta\in\mathbb{R}^{d_{g_j}}} \mathcal{R}[f_{G_{j-1}} + \beta h_{g_j}]$$

Setting $\beta = \arg\min_{\beta\in\mathbb{R}^{d_{g_j}}} \mathcal{R}[f_{G_{j-1}} + \beta h_{g_j}]$, using the strongly convex condition at $f_{G_{j-1}}$, we have:

$$\begin{aligned}
&F(G_j) - F(G_{j-1}) \\
&= \mathcal{R}[f_{G_{j-1}}] - \mathcal{R}[f_{G_j}] \geq \mathcal{R}[f_{G_{j-1}}] - \mathcal{R}[f_{G_{j-1}} + \beta h_{g_j}] \\
&\geq \mathcal{R}[f_{G_{j-1}}] - (\mathcal{R}[f_{G_{j-1}}] + \langle\nabla\mathcal{R}[f_{G_{j-1}}], \beta h_{g_j}\rangle \\
&\quad + \frac{M}{2}\|\beta h_{g_j}\|_2^2) \\
&= -\langle\nabla\mathcal{R}[f_{G_{j-1}}], \beta h_{g_j}\rangle - \frac{M}{2}\|\beta h_{g_j}\|_2^2 \\
&= \langle b^{G_{j-1}}_{g_j}, \beta\rangle - \frac{M}{2}\langle\beta, C_{g_j}\beta\rangle \\
&\geq \frac{1}{2M}\langle b^{G_{j-1}}_{g_j}, C^{-1}_{g_j} b^{G_{j-1}}_{g_j}\rangle \\
&= \frac{1}{2M(1+\lambda)}\langle b^{G_{j-1}}_{g_j}, b^{G_{j-1}}_{g_j}\rangle
\end{aligned}$$

The last equality holds because each group is whitened, so that $C_{g_j} = (1+\lambda)I$. □

Note that the $(1+\lambda)$ constant is a result of group whitening, without which the constant can be as large as $(D_{g_j}+\lambda)$ for the worst case where all the D_{g_j} number of features are the same.

The proofs above for Lemma A.1 and A.2 are for one-dimensional output responses. They can be easily generalized to multi-dimensional responses by replacing 2-norms with Frobenius norms and vector inner-products with "Frobenius products", i.e., the sum of the products of all elements.

A.3 Proof of Main Theorem

Given Lemma A.1 and Lemma A.2, the proof of Lemma 3.1 holds with the same analysis with a more general constant $\gamma = \frac{m\lambda_{min}(C)}{M(1+\lambda)}$. The following prove our main theorem 3.2.

Proof. (of Theorem 3.2, given Lemma 3.1) Define $\Delta_j = F(S_{\langle K\rangle}) - F(G_{j-1})$. Then we have $\Delta_j - \Delta_{j+1} = F(G_j) - F(G_{j-1})$. By Lemma 3.1, we have:

$$\begin{aligned}
\Delta_j &= F(S_{\langle K\rangle}) - F(G_{j-1}) \\
&\leq \frac{K}{\gamma}[\frac{F(G_j) - F(G_{j-1})}{c(g_j)}] = \frac{K}{\gamma}[\frac{\Delta_j - \Delta_{j+1}}{c(g_j)}]
\end{aligned}$$

Rearranging we get $\Delta_{j+1} \leq \Delta_j(1 - \frac{\gamma c(g_j)}{K})$. Unroll we get:

$$\begin{aligned}
\Delta_{L+1} &\leq \Delta_1\prod_{j=1}^{L}(1 - \frac{\gamma c(g_j)}{K}) \leq \Delta_1(\frac{1}{L}\sum_{j=1}^{L}(1 - \frac{\gamma c(g_j)}{K}))^L \\
&= \Delta_1(1 - \frac{B\gamma}{LK})^L < \Delta_1 e^{-\gamma\frac{B}{K}}
\end{aligned}$$

By definition of Δ_1 and Δ_{L+1}, we have:

$$F(S_{\langle K\rangle}) - F(G_{\langle B\rangle}) < F(S_{\langle K\rangle})e^{-\gamma\frac{B}{K}}$$

The theorem follows and linear prediction is the special case that $m = M$. □