Supplementary Material

A Complete convergence analysis in the regularized case

Basic setup: We are minimizing a function f of the form $F + R$ where F is a convex differentiable function $F : \mathbb{R}^p \to \mathbb{R}$ that satisfies a second order upper bound

$$F(w + \Delta) \le F(w) + \nabla F(w)^T w + \frac{\beta}{2}\Delta^T A^T A \Delta$$

and $R : \mathbb{R}^p \to \mathbb{R}$ is convex (and possibly non-differentiable) and separable across coordinates:

$$R(w) = \sum_{j=1}^{p} r(w_j)$$

In our case $\mathbf{X}$ is the $n \times p$ design matrix. If columns of $\mathbf{X}$ are zero mean and unit variance normalized then entries in $\mathbf{X}^T\mathbf{X}$ measure the correlation between features. Also, $r(x) = \lambda|x|$.

Divide the p features into B blocks of p/B features each. The algorithm we analyze is block-greedy, a direct generalization of Shotgun ($B = p$ in the Shotgun case). In the regularized case, the block-greedy algorithm is

For P randomly chosen blocks in parallel **do**

- Within a block b, find $j = j_b \in b$ such that $|\eta_j|$ is maximum and update
$$w'_j \leftarrow w_j - \eta_j$$

Endfor

Here $|\eta_j|$ serves to quantify the guaranteed descent (based on second order upper bound) if feature j is updates and solves the one-dimensional problem

$$\eta_j = \operatorname*{argmin}_{\eta} \ \nabla_j F(w)\eta + \frac{\beta}{2}\eta^2 + r(w_j + \eta) - r(w_j) \ .$$

Note that if there is no regularization, then $\eta_j = -\nabla_j F(w)/\beta = g_j/\beta$ and this is the case we analyzed in the main body of the paper. In the general case, by first order optimality conditions for the above one-dimensional convex optimization problem, we have

$$g_j + \beta\eta_j + \nu_j = 0$$

where ν_j is a subgradient of r at $w_j + \eta_j$. That is, $\nu_j \in \partial r(w_j + \eta_j)$. This implies that

$$r(w_j + \eta_j) - r(w') \le \nu_j(w_j + \eta_j - w')$$

for any w'.

We first calculate the expected change in objective function following the Shotgun analysis. We will use w_b to denote w_{j_b} (similarly for ν_b, g_b)

$$\begin{aligned}\mathbb{E}\left[f(\mathbf{w}') - f(\mathbf{w})\right] = P\mathbb{E}_b &\left[\eta_b g_b + \frac{\beta}{2}(\eta_b)^2 + r(w_b + \eta_b) - r(w_b)\right] \\ &+ \frac{\beta}{2}P(P-1)\mathbb{E}_{b \ne b'}\left[\eta_b \cdot \eta_{b'} \cdot A_{j_b}^T A_{j_{b'}}\right]\end{aligned}$$

Define the $B \times B$ matrix M (depends on the current iteration) with entries $M_{b,b'} = A_{j_b}^T A_{j_b}$. Then, using $r(w_b + \eta_b) - r(w_b) \le \nu_b \eta_b$, we continue

$$\begin{aligned}&\le \frac{P}{B}\left[\eta^T g + \frac{\beta}{2}\eta^T\eta + \nu^T\eta\right] \\ &\quad + \frac{\beta P(P-1)}{2B(B-1)}\left[\eta^\top M\eta - \eta^T\eta\right]\end{aligned}$$

Above (with some abuse of notation), η, ν and g are B length vectors with components η_b, ν_b and g_b respectively.

Our generalization of Shotgun's ρ_{block} parameter is

$$\rho_{\text{block}} = \max_{M \in \mathcal{M}} \rho(M)$$

where $\mathcal{M}$ is the set of all $B \times B$ submatrices obtainable from $\mathbf{X}^T\mathbf{X}$ by selecting exactly one index from each of the B blocks.

So, we continue

$$\begin{aligned} &\leq \frac{P}{B}\left[\eta^T g + \frac{\beta}{2}\eta^T\eta - g^T\eta - \beta\eta^T\eta\right] \\ &\quad + \frac{\beta P(P-1)}{2B(B-1)}(\rho_{\text{block}} - 1)\eta^T\eta \end{aligned}$$

where we used $\nu = -g - \beta\eta$.

Simplifying we get

$$\mathbb{E}\left[f(\mathbf{w}') - f(\mathbf{w})\right] \leq \frac{P\beta}{2B}\left[-1 + \epsilon\right]\|\eta\|_2^2$$

where

$$\epsilon = \frac{(P-1)(\rho_{\text{block}} - 1)}{(B-1)}$$

should be less than 1.

Now note that

$$\|\eta\|_2^2 = \sum_b |\eta_{j_b}|^2 = \|\eta\|_{\infty,2}^2 \,.$$

where the "infinity-2" norm $\|\cdot\|_{\infty,2}$ of a p-vector is, by definition, as follows: take the ℓ_∞ norm within a block and take the ℓ_2 of the resulting values. Note that in the second step above, we moved from a B-length η to a p length η.

This gives us

$$\mathbb{E}\left[f(\mathbf{w}') - f(\mathbf{w})\right] \leq -\frac{(1-\epsilon)P\beta}{2B}\|\eta\|_{\infty,2}^2 \,. \tag{2}$$

From the results in Dhillon et al. [2011] we know that $f(\mathbf{w}) - f(\mathbf{w}^\star) \leq C\|\eta\|_\infty$ where the constant C depends on the function F (e.g. its smoothness and Lipschitz constants) and the maximum value $\|\mathbf{w} - \mathbf{w}^\star\|_1$ can take over the course of the algorithm. Because $\|\eta\|_\infty \leq \|\eta\|_{\infty,2}$, plugging this into (2), we get

$$\mathbb{E}\left[f(\mathbf{w}') - f(\mathbf{w})\right] \leq -\frac{(1-\epsilon)P\beta}{2BC}(f(\mathbf{w}) - f(\mathbf{w}^\star))^2 \,.$$

Defining the accuracy $\alpha_k = F(w_k) - F(w^\star)$, we translate the above into the recurrence

$$\mathbb{E}\left[\alpha_{k+1} - \alpha_k\right] \leq -\frac{(1-\epsilon)P\beta}{2BC}\mathbb{E}\left[\alpha_k^2\right]$$

and by Jensen's we have $(\mathbb{E}[\alpha_k])^2 \leq \mathbb{E}\left[\alpha_k^2\right]$ and therefore

$$\mathbb{E}\left[\alpha_{k+1}\right] - \mathbb{E}\left[\alpha_k\right] \leq -\frac{(1-\epsilon)P\beta}{2BC}(\mathbb{E}\left[\alpha_k\right])^2$$

which solves to (upto a universal constant factor)

$$\mathbb{E}\left[\alpha_k\right] \leq \frac{2BC}{(1-\epsilon)P\beta} \cdot \frac{1}{k} \,.$$