A. Omitted Proofs from Section 4

Proof of Proposition 4.2. Let $\mathcal{F}_{k-1}$ be the natural filtration up to iteration $k-1$. Observe that, as $\nabla_n f(\mathbf{x}_k) = \mathbf{0}$:

$$\mathbb{E}[\Delta_k | \mathcal{F}_{k-1}] = \nabla f(\mathbf{x}_k). \tag{A.1}$$

Since $\mathbf{x}_1$ is deterministic (fixed initial point) and the only random variable Δ_1 depends on is i_1, we have:

$$\begin{aligned}\mathbb{E}[a_1 \langle \Delta_1, \mathbf{x}_* - \mathbf{x}_1 \rangle] &= a_1 \langle \nabla f(\mathbf{x}_1), \mathbf{x}_* - \mathbf{x}_1 \rangle \\ &= \mathbb{E}[a_1 \langle \nabla f(\mathbf{x}_1), \mathbf{x}_* - \mathbf{x}_1 \rangle].\end{aligned} \tag{A.2}$$

Let $k > 1$. Observe that $a_j \langle \Delta_j, \mathbf{x}_* - \mathbf{x}_j \rangle$ is measurable with respect to $\mathcal{F}_{k-1}$ for $j \leq k-1$. By linearity of expectation, using (A.1):

$$\mathbb{E}[\sum_{j=1}^{k} a_j \langle \Delta_j, \mathbf{x}_* - \mathbf{x}_j \rangle | \mathcal{F}_{k-1}] = a_k \langle \nabla f(\mathbf{x}_k), \mathbf{x}_* - \mathbf{x}_k \rangle + \sum_{j=1}^{k-1} a_j \langle \Delta_j, \mathbf{x}_* - \mathbf{x}_j \rangle .$$

Taking expectations on both sides of the last equality gives a recursion on $\mathbb{E}[\sum_{j=1}^{k} a_j \langle \Delta_j, \mathbf{x}_* - \mathbf{x}_j \rangle]$, which, combined with (A.2), completes the proof. □

Proof of Lemma 4.5. As $A_{k-1}\Gamma_{k-1}$ is measurable with respect to the natural filtration $\mathcal{F}_{k-1}$, $\mathbb{E}[A_k\Gamma_k | \mathcal{F}_{k-1}] \leq A_{k-1}\Gamma_{k-1}$ is equivalent to $\mathbb{E}[A_k\Gamma_k - A_{k-1}\Gamma_{k-1} | \mathcal{F}_{k-1}] \leq 0$.

The change in the upper bound is:

$$A_k U_k - A_{k-1} U_{k-1} = A_k(f(\mathbf{y}_k) - f(\mathbf{x}_k)) + A_{k-1}(f(\mathbf{x}_k) - f(\mathbf{y}_{k-1})) + a_k f(\mathbf{x}_k).$$

By convexity, $f(\mathbf{x}_k) - f(\mathbf{y}_{k-1}) \leq \langle \nabla f(\mathbf{x}_k), \mathbf{x}_k - \mathbf{y}_{k-1} \rangle$. Further, as $\mathbf{y}_k = \mathbf{x}_k + I_N^{i_k} \frac{a_k}{p_{i_k} A_k}(\mathbf{v}_k - \mathbf{v}_{k-1})$, we have, by smoothness of $f(\cdot)$, that $f(\mathbf{y}_k) - f(\mathbf{x}_k) \leq \left\langle \nabla f(\mathbf{x}_k), I_N^{i_k} \frac{a_k}{p_{i_k} A_k}(\mathbf{v}_k - \mathbf{v}_{k-1}) \right\rangle + \frac{L_{i_k} {a_k}^2}{2{p_{i_k}}^2 {A_k}^2} \|\mathbf{v}_k^{i_k} - \mathbf{v}_{k-1}^{i_k}\|^2$. Hence:

$$\begin{aligned}&A_k U_k - A_{k-1} U_{k-1} \\ &\quad\leq a_k f(\mathbf{x}_k) + \left\langle \nabla f(\mathbf{x}_k), A_{k-1}(\mathbf{x}_k - \mathbf{y}_{k-1}) + I_N^{i_k} \frac{a_k}{p_{i_k}}(\mathbf{v}_k - \mathbf{v}_{k-1}) \right\rangle + \frac{L_{i_k} {a_k}^2}{2{p_{i_k}}^2 A_k} \|\mathbf{v}_k^{i_k} - \mathbf{v}_{k-1}^{i_k}\|^2.\end{aligned} \tag{A.3}$$

Let $m_k(\mathbf{u}) = \sum_{j=1}^{k} a_j \langle \Delta_j, \mathbf{u} - \mathbf{x}_j \rangle + \sum_{i=1}^{n} \frac{\sigma_i}{2} \|\mathbf{u}^i - \mathbf{x}_1^i\|^2$ denote the function under the minimum in the definition of Λ_k. Observe that $m_k(\mathbf{u}) = m_{k-1}(\mathbf{u}) + a_k \langle \Delta_k, \mathbf{u} - \mathbf{x}_k \rangle$ and $\mathbf{v}_k = \operatorname{argmin}_{\mathbf{u}} m_k(\mathbf{u})$. Then:

$$\begin{aligned}m_{k-1}(\mathbf{v}_k) =& m_{k-1}(\mathbf{v}_{k-1}) + \langle \nabla m_{k-1}(\mathbf{v}_{k-1}), \mathbf{v}_k - \mathbf{v}_{k-1} \rangle + \sum_{i=1}^{n-1} \frac{\sigma_i}{2} \|\mathbf{v}_k^i - \mathbf{v}_{k-1}^i\|^2 \\ =& m_{k-1}(\mathbf{v}_{k-1}) + \frac{\sigma_{i_k}}{2} \|\mathbf{v}_k^{i_k} - \mathbf{v}_{k-1}^{i_k}\|^2,\end{aligned}$$

as $\mathbf{v}_k$ and $\mathbf{v}_{k-1}$ only differ over the block i_k and $\mathbf{v}_{k-1} = \operatorname{argmin}_{\mathbf{u}} m_{k-1}(\mathbf{u})$ (and, thus, $\nabla m_{k-1}(\mathbf{v}_{k-1}) = \mathbf{0}$).

Hence, it follows that $m_k(\mathbf{v}_k) - m_{k-1}(\mathbf{v}_{k-1}) = a_k \langle \Delta_k, \mathbf{v}_k - \mathbf{x}_k \rangle + \frac{\sigma_{i_k}}{2} \|\mathbf{v}_k^{i_k} - \mathbf{v}_{k-1}^{i_k}\|^2$, and, thus:

$$A_k \Lambda_k - A_{k-1} \Lambda_{k-1} = a_k f(\mathbf{x}_k) + a_k \langle \Delta_k, \mathbf{v}_k - \mathbf{x}_k \rangle + \frac{\sigma_{i_k}}{2} \|\mathbf{v}_k^{i_k} - \mathbf{v}_{k-1}^{i_k}\|^2. \tag{A.4}$$

Combining (A.3) and (A.4):

$$\begin{aligned}A_k \Gamma_k - A_{k-1} \Gamma_{k-1} \leq& \left\langle \nabla f(\mathbf{x}_k), A_{k-1}(\mathbf{x}_k - \mathbf{y}_{k-1}) + I_N^{i_k} \frac{a_k}{p_{i_k}}(\mathbf{v}_k - \mathbf{v}_{k-1}) \right\rangle - a_k \langle \Delta_k, \mathbf{v}_k - \mathbf{x}_k \rangle \\ &+ \frac{L_{i_k} {a_k}^2}{2{p_{i_k}}^2 A_k} \|\mathbf{v}_k^{i_k} - \mathbf{v}_{k-1}^{i_k}\|^2 - \frac{\sigma_{i_k}}{2} \|\mathbf{v}_k^{i_k} - \mathbf{v}_{k-1}^{i_k}\|^2 \\ \leq& \left\langle \nabla f(\mathbf{x}_k), A_{k-1}(\mathbf{x}_k - \mathbf{y}_{k-1}) + I_N^{i_k} \frac{a_k}{p_{i_k}}(\mathbf{v}_k - \mathbf{v}_{k-1}) \right\rangle - a_k \langle \Delta_k, \mathbf{v}_k - \mathbf{x}_k \rangle ,\end{aligned}$$

as, by the initial assumptions, $\frac{{a_k}^2}{A_k} \leq \frac{p_{i_k}^2 \sigma_{i_k}}{L_{i_k}}$.

Finally, taking expectations on both sides, and as $\mathbf{x}_k, \mathbf{y}_{k-1}, \mathbf{v}_{k-1}$ are all measurable w.r.t. $\mathcal{F}_{k-1}$ and by the separability of the terms in the definition of $\mathbf{v}_k$:

$$\mathbb{E}[A_k\Gamma_k - A_{k-1}\Gamma_{k-1}|\mathcal{F}_{k-1}] \leq \langle \nabla f(\mathbf{x}_k), A_k\mathbf{x}_k - A_{k-1}\mathbf{y}_{k-1} - a_k\mathbf{v}_{k-1}\rangle = 0,$$

as, from (AAR-BCD), $\mathbf{x}_k = \frac{A_{k-1}}{A_k}\mathbf{y}_{k-1} + \frac{a_k}{A_k}\mathbf{v}_{k-1}$. □

B. Efficient Implementation of AAR-BCD Iterations

Using similar ideas as in (Fercoq & Richtárik, 2015; Lin et al., 2014; Lee & Sidford, 2013), here we discuss how to efficiently implement iterations of AAR-BCD, without requiring full-vector updates. First, due to the separability of the terms inside the minimum, between successive iterations $\mathbf{v}_k$ changes only over a single block. This is formalized in the following simple proposition.

Proposition B.1. *In each iteration $k \geq 1$, $\mathbf{v}_k^i = \mathbf{v}_{k-1}^i$, $\forall i \neq i_k$ and $\mathbf{v}_k^{i_k} = \mathbf{v}_{k-1}^{i_k} + \mathbf{w}^{i_k}$, where:*

$$\mathbf{w}^{i_k} = \operatorname*{argmin}_{\mathbf{u}^{i_k}}\{a_k \left\langle \Delta_k^{i_k}, \mathbf{u}\right\rangle + \frac{\sigma_{i_k}}{2}\|\mathbf{u}^{i_k} - \mathbf{v}_{k-1}^{i_k}\|^2\}.$$

Proof. Recall the definition of $\mathbf{v}_k$. We have:

$$\begin{aligned}
\mathbf{v}_k =& \operatorname*{argmin}_{\mathbf{u}}\Big\{\sum_{j=1}^{k}\langle\Delta_j, \mathbf{u}\rangle + \sum_{i=1}^{n-1}\frac{\sigma_i}{2}\|\mathbf{u}^i - \mathbf{x}_1^i\|^2\Big\}\\
=& \operatorname*{argmin}_{\mathbf{u}}\Big\{\sum_{j=1}^{k-1}\langle\Delta_j, \mathbf{u}\rangle + \langle\Delta_k, \mathbf{u}\rangle + \sum_{i=1}^{n-1}\frac{\sigma_i}{2}\|\mathbf{u}^i - \mathbf{x}_1^i\|^2\Big\}\\
=& \operatorname*{argmin}_{\mathbf{u}}\Big\{\sum_{j=1}^{k-1}\langle\Delta_j, \mathbf{u}\rangle + \left\langle\Delta_k^{i_k}, \mathbf{u}^{i_k}\right\rangle + \sum_{i=1}^{n-1}\frac{\sigma_i}{2}\|\mathbf{u}^i - \mathbf{x}_1^i\|^2\Big\}\\
=& \mathbf{v}_{k-1} + \operatorname*{argmin}_{\mathbf{u}^{i_k}}\Big\{\left\langle\Delta_k^{i_k}, \mathbf{u}^{i_k}\right\rangle + \frac{\sigma_{i_k}}{2}\|\mathbf{u}^{i_k} - \mathbf{v}_{k-1}^{i_k}\|^2\Big\},
\end{aligned}$$

where the third equality is by the definition of Δ_k ($\Delta_k^i = 0$ for $i \neq i_k$) and the last equality follows from block-separability of the terms under the min. □

Since $\mathbf{v}_k$ only changes over a single block, this will imply that the changes in $\mathbf{x}_k$ and $\mathbf{y}_k$ can be localized. In particular, let us observe the patterns in changes between successive iterations. We have that, $\forall i \neq n$:

$$\mathbf{x}_k^i = \frac{A_{k-1}}{A_k}\mathbf{y}_{k-1}^i + \frac{a_k}{A_k}\mathbf{v}_{k-1}^i = \frac{A_{k-1}}{A_k}\left(\mathbf{y}_{k-1}^i - \mathbf{v}_{k-1}^i\right) + \mathbf{v}_{k-1}^i \tag{B.1}$$

and

$$\begin{aligned}
\mathbf{y}_k^i &= \mathbf{x}_k^i + \frac{1}{p_i}\frac{a_k}{A_k}\left(\mathbf{v}_k^i - \mathbf{v}_{k-1}^i\right)\\
&= \frac{A_{k-1}}{A_k}\left(\mathbf{y}_{k-1}^i - \mathbf{v}_{k-1}^i\right) + \left(1 - \frac{1}{p_i}\frac{a_k}{A_k}\right)\left(\mathbf{v}_{k-1}^i - \mathbf{v}_k^i\right) + \mathbf{v}_k^i.
\end{aligned} \tag{B.2}$$

Due to Proposition B.1, $\mathbf{v}_k$ and $\mathbf{v}_{k-1}$ can be computed without full-vector operations (assuming the gradients can be computed without full-vector operations, which we will show later in this section). Hence, we need to show that it is possible to replace $\frac{A_{k-1}}{A_k}\left(\mathbf{y}_{k-1}^i - \mathbf{v}_{k-1}^i\right)$ with a quantity that can be computed without the full-vector operations. Observe that $\mathbf{y}_0 - \mathbf{v}_0 = 0$ (from the initialization of (AAR-BCD)) and that, from (B.2):

$$\mathbf{y}_k^i - \mathbf{v}_k^i = \frac{A_{k-1}}{A_k}\left(\mathbf{y}_{k-1}^i - \mathbf{v}_{k-1}^i\right) + \left(1 - \frac{1}{p_i}\frac{a_k}{A_k}\right)\left(\mathbf{v}_{k-1}^i - \mathbf{v}_k^i\right).$$

Dividing both sides by $\frac{{a_k}^2}{{A_k}^2}$ and assuming that $\frac{{a_k}^2}{A_k}$ is constant over iterations, we get:

$$\frac{{A_k}^2}{{a_k}^2}\left(\mathbf{y}_k^i - \mathbf{v}_k^i\right) = \frac{{A_{k-1}}^2}{{a_{k-1}}^2}\left(\mathbf{y}_{k-1}^i - \mathbf{v}_{k-1}^i\right) + \frac{{A_k}^2}{{a_k}^2}\left(1 - \frac{1}{p_i}\frac{a_k}{A_k}\right)\left(\mathbf{v}_{k-1}^i - \mathbf{v}_k^i\right). \tag{B.3}$$

Let N_n denote the size of the n^{th} block and define the $(N - N_n)$-length vector $\mathbf{u}_k$ by $\mathbf{u}_k^i = \frac{{A_k}^2}{{a_k}^2}\left(\mathbf{y}_k^i - \mathbf{v}_k^i\right)$, $\forall i \neq n$. Then (from (B.3)) $\mathbf{u}_k^i = \mathbf{u}_{k-1}^i + \frac{{A_k}^2}{{a_k}^2}\left(1 - \frac{1}{p_i}\frac{a_k}{A_k}\right)\left(\mathbf{v}_{k-1}^i - \mathbf{v}_k^i\right)$, and, hence, in iteration k, $\mathbf{u}_k$ changes only over block i_k. Combining with (B.1) and (B.2), we have the following lemma.

Lemma B.2. *Assume that $\frac{{a_k}^2}{A_k}$ is kept constant over the iterations of AAR-BCD. Let $\mathbf{u}_k$ be the $(N-N_n)$-dimensional vector defined recursively as $\mathbf{u}_0 = \mathbf{0}$, $\mathbf{u}_k^i = \mathbf{u}_{k-1}^i$ for $i \in \{1, ..., n-1\}$, $i \neq i_k$ and $\mathbf{u}_k^{i_k} = \mathbf{u}_{k-1}^{i_k} + \frac{{A_k}^2}{{a_k}^2}\left(1 - \frac{1}{p_i}\frac{a_k}{A_k}\right)\left(\mathbf{v}_{k-1}^i - \mathbf{v}_k^i\right)$. Then, $\forall i \in \{1, ..., n-1\}$: $\mathbf{x}_k^i = \frac{{a_k}^2}{{A_k}^2}\mathbf{u}_{k-1}^i + \mathbf{v}_{k-1}^i$ and $\mathbf{y}_k^i = \frac{{a_k}^2}{{A_k}^2}\mathbf{u}_{k-1}^i + \left(1 - \frac{1}{p_i}\frac{a_k}{A_k}\right)\left(\mathbf{v}_{k-1}^i - \mathbf{v}_k^i\right) + \mathbf{v}_k^i$.*

Note that we will never need to explicitly compute $\mathbf{x}_k, \mathbf{y}_k$, except for the last iteration K, which outputs $\mathbf{y}_K$. To formalize this claim, we need to show that we can compute the gradients $\nabla_i f(\mathbf{x}_k)$ without explicitly computing $\mathbf{x}_k$ and that we can efficiently perform the exact minimization over the n^{th} block. This will only be possible by assuming specific structure of the objective function, as is typical for accelerated block-coordinate descent methods (Fercoq & Richtárik, 2015; Lee & Sidford, 2013; Lin et al., 2014). In particular, we assume that for some $m \times N$ dimensional matrix $\mathbf{M}$:

$$f(\mathbf{x}) = \sum_{j=1}^{m} \phi_j(e_j^T \mathbf{M}\mathbf{x}) + \psi(\mathbf{x}), \tag{B.4}$$

where $\phi_j : \mathbb{R} \to \mathbb{R}$ and $\psi = \sum_{i=1}^{n} \psi_i : \mathbb{R}^N \to \mathbb{R}$ is block-separable.

Efficient Gradient Computations. Assume for now that $\mathbf{x}_k^n$ can be computed efficiently (we will address this at the end of this section). Let ind denote the set of indices of the coordinates from blocks $\{1, 2, ..., n-1\}$ and denote by $\mathbf{B}$ the matrix obtained by selecting the columns of $\mathbf{M}$ that are indexed by ind. Similarly, let ind_n denote the set of indices of the coordinates from block n and let $\mathbf{C}$ denote the submatrix of $\mathbf{M}$ obtained by selecting the columns of $\mathbf{M}$ that are indexed by ind_n. Denote $\mathbf{r}_{\mathbf{u}_k} = \mathbf{B}\mathbf{u}_k$, $\mathbf{r}_{\mathbf{v}_k} = \mathbf{B}[\mathbf{v}_k^1, \mathbf{v}_k^2, ..., \mathbf{v}_k^{n-1}]^T$, $\mathbf{r}_n = \mathbf{C}\mathbf{x}_k^n$. Let ind_{i_k} be the set of indices corresponding to the coordinates from block i_k. Then:

$$\nabla_{i_k} f(\mathbf{x}_k) = \sum_{j=1}^{m} (\mathbf{M}_{j, ind_{ik}})^T \phi_j' \left(\frac{{a_k}^2}{{A_k}^2}\mathbf{r}_{\mathbf{u}_{k-1}}^j + \mathbf{r}_{\mathbf{v}_{k-1}}^j + \mathbf{r}_n^j\right) + \nabla_{i_k}\psi(\mathbf{x}). \tag{B.5}$$

Hence, as long as we maintain $\mathbf{r}_{\mathbf{u}_k}, \mathbf{r}_{\mathbf{v}_k}$, and $\mathbf{r}_n$ (which do not require full-vector operations), we can efficiently compute the partial gradients $\nabla_{i_k} f(\mathbf{x}_k)$ without ever needing to perform any full-vector operations.

Efficient Exact Minimization. Suppose first that $\psi(\mathbf{x}) \equiv 0$. Then:

$$\mathbf{r}_n = \operatorname*{argmin}_{\mathbf{r} \in \mathbb{R}^m} \left\{ \sum_{j=1}^{m} \phi_j \left(\frac{{a_k}^2}{{A_k}^2}\mathbf{r}_{\mathbf{u}_{k-1}}^j + \mathbf{r}_{\mathbf{v}_{k-1}}^j + \mathbf{r}^j\right)\right\},$$

and $\mathbf{r}_n$ can be computed but solving m single-variable minimization problems, which can be done in closed form or with a very low complexity. Computing $\mathbf{r}_n$ is sufficient for defining all algorithm iterations, except for the last one (that outputs a solution). Hence, we only need to compute $\mathbf{x}_k^n$ once – in the last iteration.

More generally, $\mathbf{x}_k^n$ is determined by solving:

$$\mathbf{x}_k^n = \operatorname*{argmin}_{\mathbf{x} \in \mathbb{R}^{N_n}} \left\{ \sum_{j=1}^{m} \phi_j \left(\frac{{a_k}^2}{{A_k}^2}\mathbf{r}_{\mathbf{u}_{k-1}}^j + \mathbf{r}_{\mathbf{v}_{k-1}}^j + (\mathbf{C}\mathbf{x})^j\right) + \psi_n(\mathbf{x})\right\}.$$

When m and N_n are small, high-accuracy polynomial-time convex optimization algorithms are computationally inexpensive, and $\mathbf{x}_k^n$ can be computed efficiently.

In the special case of linear and ridge regression, $\mathbf{x}_k^n$ can be computed in closed form, with minor preprocessing. In particular, if $\mathbf{b}$ is the vector of labels, then the problem becomes:

$$\mathbf{x}_k^n = \operatorname*{argmin}_{\mathbf{x}\in\mathbb{R}^{N_n}} \left\{ \sum_{j=1}^{m} \left(\frac{{a_k}^2}{{A_k}^2}\mathbf{r}^j_{\mathbf{u}_{k-1}} + \mathbf{r}^j_{\mathbf{v}_{k-1}} + (\mathbf{C}\mathbf{x})^j - \mathbf{b}^j \right)^2 + \frac{\lambda}{2}\|\mathbf{x}\|_2^2 \right\},$$

where $\lambda = 0$ in the case of (simple) linear regression. Let $\mathbf{b}' = \mathbf{b} - \frac{{a_k}^2}{{A_k}^2}\mathbf{r}_{\mathbf{u}_{k-1}} - \mathbf{r}_{\mathbf{v}_{k-1}}$. Then:

$$\mathbf{x}_k^n = (\mathbf{C}^T\mathbf{C} + \lambda\mathbf{I})^\dagger(\mathbf{C}^T\mathbf{b}'),$$

where $(\cdot)^\dagger$ denotes the matrix pseudoinverse, and $\mathbf{I}$ is the identity matrix. Since $\mathbf{C}^T\mathbf{C} + \lambda\mathbf{I}$ does not change over iterations, $(\mathbf{C}^T\mathbf{C} + \lambda\mathbf{I})^\dagger$ can be computed only once at the initialization. Recall that $\mathbf{C}^T\mathbf{C} + \lambda\mathbf{I}$ is an $N_n \times N_n$ matrix, where N_n is the size of the n^{th} block, and thus inverting $\mathbf{C}^T\mathbf{C} + \lambda\mathbf{I}$ is computationally inexpensive as long as N_n is not too large. This reduces the overall per-iteration cost of the exact minimization to about the same cost as for performing gradient steps.