Supplementary Material for the Article "Accelerating Greedy Coordinate Descent Methods"

1 Proofs in Section 2

Lemma 2.2

$$f(x^{k+1}) \leq f(y^k) + \langle \nabla f(y^k), s^{k+1} - y^k \rangle + \tfrac{n}{2}\|s^{k+1} - y^k\|_L^2 \ .$$

Proof:

$$\begin{aligned} f(x^{k+1}) &\leq f(y^k) - \tfrac{1}{2L_{j_k^1}} \left(\nabla_{j_k^1} f(y^k)\right)^2 \\ &\leq f(y^k) - \tfrac{1}{2n}\|\nabla f(y^k)\|_{L^{-1}}^2 \qquad (1) \\ &= f(y^k) + \langle \nabla f(y^k), s^{k+1} - y^k \rangle + \tfrac{n}{2}\|s^{k+1} - y^k\|_L^2 \ , \end{aligned}$$

where the first inequality of (1) derives from the smoothness of $f(\cdot)$, and is a simple instance of equation (8) of the paper using $x = y^k$, $i = j_k^1$, and $h = -\frac{1}{L_{j_k^1}} \nabla_{j_k^1} f(y^k)$. The second inequality of (1) follows from the definition of j_k^1 which yields:

$$n\left[\tfrac{1}{L_{j_k^1}} \left(\nabla_{j_k^1} f(y^k)\right)^2\right] \geq \sum_{i=1}^{n} \tfrac{1}{L_i} \left(\nabla_i f(y^k)\right)^2 = \left\|\nabla f(y^k)\right\|_{L^{-1}}^2 \ .$$

The last equality of (1) follows by using the definition of s^{k+1} and rearranging terms. □

Lemma 2.3

$$f(x^{k+1}) \leq (1-\theta_k) f(x^k) + \theta_k f(x^*) + \tfrac{n\theta_k^2}{2}\|x^* - z^k\|_L^2 - \tfrac{n\theta_k^2}{2}\|x^* - t^{k+1}\|_L^2 \ . \qquad (2)$$

Proof: Recall that we can rewrite t^{k+1} as

$$t^{k+1} = \arg\min_z \ \langle \nabla f(y^k), z - z^k \rangle + \tfrac{n\theta_k}{2}\|z - z^k\|_L^2 \ . \qquad (3)$$

Following from Lemma 2.2, we have

$$\begin{aligned} f(x^{k+1}) &\leq f(y^k) + \langle \nabla f(y^k), s^{k+1} - y^k \rangle + \tfrac{n}{2}\|s^{k+1} - y^k\|_L^2 \\ &= f(y^k) + \theta_k \left(\langle \nabla f(y^k), t^{k+1} - z^k \rangle + \tfrac{n\theta_k}{2}\|t^{k+1} - z^k\|_L^2 \right) \\ &= f(y^k) + \theta_k \left(\langle \nabla f(y^k), x^* - z^k \rangle + \tfrac{n\theta_k}{2}\|x^* - z^k\|_L^2 - \tfrac{n\theta_k}{2}\|x^* - t^{k+1}\|_L^2 \right) \\ &= (1-\theta_k)\left(f(y^k) + \langle \nabla f(y^k), x^k - y^k \rangle \right) + \theta_k \left(f(y^k) + \langle \nabla f(y^k), x^* - y^k \rangle \right) \\ &\quad + \tfrac{n\theta_k^2}{2}\|x^* - z^k\|_L^2 - \tfrac{n\theta_k^2}{2}\|x^* - t^{k+1}\|_L^2 \\ &\leq (1-\theta_k) f(x^k) + \theta_k f(x^*) + \tfrac{n\theta_k^2}{2}\|x^* - z^k\|_L^2 - \tfrac{n\theta_k^2}{2}\|x^* - t^{k+1}\|_L^2 \end{aligned} \tag{4}$$

where the first equality of (4) utilizes $s^{k+1} - y^k = \theta_k(t^{k+1} - z^k)$. The second equality of (4) follows as an application of the Three-Point-Property (Lemma 2.1) together with (3), where we set $\phi(x) = \langle \nabla f(y^k), x - z^k \rangle$ and $h(x) = \frac{n\theta_k}{2}\|x\|_L^2$ (whereby $D_h(x,v) = \frac{n\theta_k}{2}\|x - v\|_L^2$). The third equality of (4) is derived from $y^k = (1-\theta_k)x^k + \theta_k z^k$ and rearranging the terms. And the last inequality of (4) is an application of the gradient inequality at y^k applied to x^k and also to x^*. □

Lemma 2.4

$$\tfrac{n}{2}\|x^* - z^k\|_L^2 - \tfrac{n}{2}\|x^* - t^{k+1}\|_L^2 = \tfrac{n^2}{2}\|x^* - z^k\|_L^2 - \tfrac{n^2}{2}E_{j_k^2}\left[\|x^* - z^{k+1}\|_L^2\right] .$$

Proof:

$$\begin{aligned} \tfrac{n}{2}\|x^* - z^k\|_L^2 - \tfrac{n}{2}\|x^* - t^{k+1}\|_L^2 &= \tfrac{n}{2}\left\langle t^{k+1} - z^k, 2x^* - 2z^k \right\rangle_L - \tfrac{n}{2}\left\|t^{k+1} - z^k\right\|_L^2 \\ &= \tfrac{n^2}{2}E_{j_k^2}\left[\left\langle z^{k+1} - z^k, 2x^* - 2z^k \right\rangle_L - \left\|z^{k+1} - z^k\right\|_L^2\right] \\ &= \tfrac{n^2}{2}\|x^* - z^k\|_L^2 - \tfrac{n^2}{2}E_{j_k^2}\left[\|x^* - z^{k+1}\|_L^2\right], \end{aligned} \tag{5}$$

where the first and third equations above are straightforward arithmetic rearrangements, and the second equation follows from the two easy-to-verify identities $t^{k+1} - z^k = nE_{j_k^2}\left[z^{k+1} - z^k\right]$ and $\left\|t^{k+1} - z^k\right\|_L^2 = nE_{j_k^2}\left[\left\|z^{k+1} - z^k\right\|_L^2\right]$. □

2 Proofs in Section 3

Theorem 3.1. *Consider the Accelerated Semi-Greedy Coordinate Descent method for the strongly convex case (Algorithm 2 with rule (5) in the main paper for ASCD). If $f(\cdot)$ is coordinate-wise L-smooth and μ-strongly convex with respect to $\|\cdot\|_L$, it holds for all $k \geq 1$ that:*

$$E_{\xi_k}\left[f(x^k) - f^* + \tfrac{n^2}{2}(a^2+b)\|z^k - x^*\|_L^2\right] \leq \left(1 - \tfrac{\sqrt{\mu}}{n+\sqrt{\mu}}\right)^k \left(f(x^0) - f^* + \tfrac{n^2}{2}(a^2+b)\|x^0 - x^*\|_L^2\right). \tag{6}$$

□

In order to prove Theorem 3.1, we first prove the following three lemmas:

Lemma 3.1.

$$a^2\|x^*-z^k\|_L^2+b\|x^*-y^k\|_L^2=(a^2+b)\|x^*-u^k\|_L^2+\tfrac{a^2b}{a^2+b}\|y^k-z^k\|_L^2\ .$$

Proof:

$$\begin{aligned}
&(a^2+b)\|x^*-u^k\|_L^2+\tfrac{a^2b}{a^2+b}\|y^k-z^k\|_L^2\\
=\ &(a^2+b)\left(\|x^*\|_L^2-2\langle x^*,u^k\rangle_L+\|u^k\|_L^2\right)+\tfrac{a^2b}{a^2+b}\|y^k-z^k\|_L^2\\
=\ &(a^2+b)\|x^*\|_L^2-2\langle x^*,a^2z^k+by^k\rangle_L+\tfrac{1}{a^2+b}\|a^2z^k+by^k\|_L^2+\tfrac{a^2b}{a^2+b}\|y^k-z^k\|_L^2 \qquad (7)\\
=\ &(a^2+b)\|x^*\|_L^2-2\langle x^*,a^2z^k+by^k\rangle_L+a^2\|z^k\|_L^2+b\|y^k\|_L^2\\
=\ &a^2\|x^*-z^k\|_L^2+b\|x^*-y^k\|_L^2\ ,
\end{aligned}$$

where the second equality utilizes $u^k=\frac{a^2}{a^2+b}z^k+\frac{b}{a^2+b}y^k$ and the other equalities are just mathematical manipulations. □

Lemma 3.2. Define $t^{k+1}:=u^k-\frac{a}{a^2+b}\frac{1}{n}\mathbf{L}^{-1}\nabla f(y^k)$, then

$$\|x^*-t^{k+1}\|_L^2-\|x^*-u^k\|_L^2=nE_{j_k^2}\left[\|x^*-z^{k+1}\|_L^2-\|x^*-u^k\|_L^2\right]$$

Proof:

$$\begin{aligned}
\|x^*-t^{k+1}\|_L^2-\|x^*-u^k\|_L^2 &= 2\langle x^*-u^k,u^k-t^{k+1}\rangle_L+\|u^k-t^{k+1}\|_L^2\\
&= 2nE_{j_k^2}\left[\langle x^*-u^k,u^k-z^{k+1}\rangle_L+\|u^k-z^{k+1}\|_L^2\right] \qquad (8)\\
&= nE_{j_k^2}\left[\|x^*-z^{k+1}\|_L^2-\|x^*-u^k\|_L^2\right],
\end{aligned}$$

where the second equality is from the relationship of $t^{k+1}=u^k-\frac{a}{a^2+b}\frac{1}{n}\mathbf{L}^{-1}\nabla f(y^k)$ and $z^{k+1}=u^k-\frac{a}{a^2+b}\frac{1}{nL_{j_k^2}}\nabla_{j_k^2}f(y^k)e_{j_k^2}$, and the first and third equations are just rearrangement. □

Lemma 3.3.

$$a^2\le(1-a)(a^2+b)\ .$$

Proof: Remember that $b=\frac{\mu a}{n^2}$ and $a>0$, thus the above inequality is equivalent to

$$a\le(1-a)(a+\tfrac{\mu}{n^2})\ .$$

Substituting $a=\frac{\sqrt{\mu}}{n+\sqrt{\mu}}$, the above inequality becomes

$$\tfrac{\sqrt{\mu}}{n}\le\tfrac{\sqrt{\mu}}{n+\sqrt{\mu}}+\tfrac{\mu}{n^2}\ .$$

We furnish the proof by noting $\frac{\sqrt{\mu}}{n} - \frac{\sqrt{\mu}}{n+\sqrt{\mu}} = \frac{\mu}{n(n+\sqrt{\mu})} \leq \frac{\mu}{n^2}$. □

Proof of Theorem 3.1: Recall that $t^{k+1} = u^k - \frac{a}{a^2+b}\frac{1}{n}\mathbf{L}^{-1}\nabla f(y^k)$, then it is easy to check that

$$t^{k+1} = \arg\min_z \ a\langle \nabla f(y^k), z - z^k\rangle + \frac{n}{2}a^2\|z - z^k\|_L^2 + \frac{n}{2}b\|z - y^k\|_L^2$$

by writing the optimality conditions of the right-hand side.

We have

$$\begin{aligned}
& f(x^{k+1}) - f(y^k) \\
\leq\ & \langle \nabla f(y^k), x^{k+1} - y^k\rangle + \tfrac{1}{2}\|x^{k+1} - y^k\|_L^2 \\
=\ & -\tfrac{1}{2L_{j_k^1}}\left(\nabla_{j_k^1} f(y^k)\right)^2 \\
\leq\ & -\tfrac{1}{2n}\|\nabla f(y^k)\|_{L^{-1}}^2 \\
\leq\ & a\langle \nabla f(y^k), t^{k+1} - z^k\rangle + \tfrac{n}{2}a^2\|t^{k+1} - z^k\|_L^2 \\
=\ & a\langle \nabla f(y^k), x^* - z^k\rangle + \tfrac{n}{2}a^2\|x^* - z^k\|_L^2 - \tfrac{n}{2}a^2\|x^* - t^{k+1}\|_L^2 + \tfrac{n}{2}b\|x^* - y^k\|_L^2 \\
& -\tfrac{n}{2}b\|y^k - t^{k+1}\|_L^2 - \tfrac{n}{2}b\|t^{k+1} - x^*\|_L^2 \\
\leq\ & a\langle \nabla f(y^k), x^* - z^k\rangle + \tfrac{n}{2}a^2\|x^* - z^k\|_L^2 - \tfrac{n}{2}a^2\|x^* - t^{k+1}\|_L^2 + \tfrac{n}{2}b\|x^* - y^k\|_L^2 - \tfrac{n}{2}b\|t^{k+1} - x^*\|_L^2 \\
=\ & a\langle \nabla f(y^k), x^* - z^k\rangle + \tfrac{n}{2}(a^2+b)\left(\|x^* - u^k\|_L^2 - \|x^* - t^{k+1}\|_L^2\right) + \tfrac{n}{2}\tfrac{a^2 b}{a^2+b}\|y^k - z^k\|_L^2 \\
=\ & a\langle \nabla f(y^k), x^* - z^k\rangle + \tfrac{n^2}{2}(a^2+b)E_{j_k^2}\left[\|x^* - u^k\|_L^2 - \|x^* - z^{k+1}\|_L^2\right] + \tfrac{n}{2}\tfrac{a^2 b}{a^2+b}\|y^k - z^k\|_L^2 \\
\leq\ & a\langle \nabla f(y^k), x^* - z^k\rangle + \tfrac{n^2}{2}(a^2+b)E_{j_k^2}\left[\|x^* - u^k\|_L^2 - \|x^* - z^{k+1}\|_L^2\right] + \tfrac{n^2}{2}\tfrac{a^2 b}{a^2+b}\|y^k - z^k\|_L^2 \\
=\ & a\langle \nabla f(y^k), x^* - z^k\rangle + \tfrac{n^2}{2}\left((a^2+b)\|x^* - u^k\|_L^2 + \tfrac{a^2 b}{a^2+b}\|y^k - z^k\|_L^2\right) - \tfrac{n^2}{2}(a^2+b)\mathbb{E}_{j_k^2}\left[\|x^* - z^{k+1}\|_L^2\right] \\
=\ & a\langle \nabla f(y^k), x^* - z^k\rangle + \tfrac{n^2}{2}\left(a^2\|x^* - z^k\|_L^2 + b\|x^* - y^k\|_L^2\right) - \tfrac{n^2}{2}(a^2+b)E_{j_k^2}\left[\|x^* - z^{k+1}\|_L^2\right],
\end{aligned} \tag{9}$$

where the first inequality is due to coordinate-wise smoothness, the first equality utilizes $x^{k+1} = y^k - \frac{1}{L_{j_k^1}}\nabla_{j_k^1} f(y^k) e_{j_k^1}$, the second inequality follows from the fact that j_k^1 is the greedy coordinate of $\nabla f(y^k)$ in the $\|\cdot\|_{L^{-1}}$ norm, the third inequality follows from the basic inequality $\|v\|_L^2 + \|w\|_{L^{-1}}^2 \geq 2\langle v, w\rangle$ for all v, w, the second equality is from Three Point Property by noticing

$$t^{k+1} = \arg\min_z \ a\langle \nabla f(y^k), z - z^k\rangle + \frac{n}{2}a^2\|z - z^k\|_L^2 + \frac{n}{2}b\|z - y^k\|_L^2 \ ,$$

the third equality follows from Lemma 3.1, and the fourth and sixth equalities each utilize Lemma 3.2.

On the other hand, by strong convexity we have

$$\begin{aligned} f(y^k) - f(x^*) &\leq \langle \nabla f(y^k), y^k - x^* \rangle - \tfrac{1}{2}\mu \|y^k - x^*\|_L^2 \\ &= \langle \nabla f(y^k), y^k - z^k \rangle + \langle \nabla f(y^k), z^k - x^* \rangle - \tfrac{1}{2}\mu \|y^k - x^*\|_L^2 \\ &= \tfrac{1-a}{a} \langle \nabla f(y^k), x^k - y^k \rangle + \langle \nabla f(y^k), z^k - x^* \rangle - \tfrac{1}{2}\mu \|y^k - x^*\|_L^2 \\ &\leq \tfrac{1-a}{a} \left(f(x^k) - f(y^k) \right) + \langle \nabla f(y^k), z^k - x^* \rangle - \tfrac{1}{2}\mu \|y^k - x^*\|_L^2, \end{aligned} \tag{10}$$

where the second equality uses the fact that $y^k = (1-a)x^k + az^k$ and the last inequality is from the gradient inequality.

By rearranging (10), we obtain

$$f(y^k) - f(x^*) \leq (1-a) \left(f(x^k) - f(x^*) \right) + a \langle \nabla f(y^k), z^k - x^* \rangle - \frac{1}{2}\mu a \|y^k - x^*\|_L^2 \, . \tag{11}$$

Notice that $b = \frac{\mu a}{n^2}$ and $a^2 \leq (1-a)(a^2+b)$ following from Lemma 3.3. Thus summing up (9) and (11) leads to

$$\begin{aligned} & E_{j_k^2} \left[f(x^{k+1}) - f(x^*) + \tfrac{n^2}{2}(a^2+b) \|z^{k+1} - x^*\|_L^2 \right] \\ \leq \; & (1-a) \left(f(x^k) - f(x^*) \right) + \tfrac{n^2}{2} a^2 \|z^k - x^*\|_L^2 \\ \leq \; & (1-a) \left(f(x^k) - f(x^*) + \tfrac{n^2}{2}(a^2+b) \|z^k - x^*\|_L^2 \right) \, , \end{aligned} \tag{12}$$

which furnishes the proof using a telescoping series. □

3 More Material on the Numerical Experiments

3.1 Implementation Detail

To be consistent with the notation in statistics and machine learning we use p to denote the dimension of the variables in the optimization problems describing linear and logistic regression. Then the per-iteration computation cost of AGCD and ASCD is dominated by three computations: (i) p-dimensional vector operations (such as in computing y^k using x^k and z^k), (ii) computation of the gradient $\nabla f(\cdot)$, and (iii) computation of the maximum (weighted) magnitude coordinate of the gradient $\nabla f(\cdot)$. [3] proposed an efficient way to avoid (i) by changing variables. Distinct from the dual approaches discussed in [3], [4], and [1], here we only consider the primal problem in the regime $n > p$, and therefore the cost of (ii) dominates the cost of (i) in these cases. For this reason in our numerical experiments we use the simple implementation of ARCD proposed by Nesterov [6] and which we adopt for AGCD and ASCD as well. We note that both the randomized methods and the greedy methods can take advantage of the efficient calculations proposed in [3] as well.

For the linear regression experiments we focused on synthetic problem instances with different condition numbers κ of the matrix $X^T X$ and where X is dense. In this case the cost for computation (i) is $O(p)$. And by taking advantage of the coordinate update structure, we can implement (ii) in $O(p)$ operations by pre-computing and storing $X^T X$ in memory, see [6] and [4] for further details. The cost of (iii) is simply $O(p)$.

The data (X, y) for the linear regression problems is generated as follows. For a given number of samples n and problem dimension p (in the experiments we used $n = 200$ and $p = 100$), we generate a standard Gaussian random matrix $\bar{X} \in \mathbb{R}^{p \times n}$ with each entry drawn $\sim N(0, 1)$. In order to generate the design matrix X with fixed condition number κ, we first decompose of $\bar{X}$ as $\bar{X} = U^T \bar{D} V$. Then we rescale the diagonal matrix $\bar{D}$ of singular values linearly to D such that the smallest singular value of D is $\frac{1}{\sqrt{\kappa}}$ and the largest singular value of D is 1. We then compute the final design matrix $X = U^T D V$ and therefore the condition number of $X^T X$ becomes κ. We generate the response vector y using the linear model $y \sim N(X\beta^*, \sigma^2)$, with true model β^* chosen randomly by a Gaussian distribution as well. For the cases with finite κ, we are able to compute the strong convexity parameter μ exactly because the objective function is quadratic, and we use that μ to implement our Algorithm Framework for strongly convex problems (Algorithm Framework 2). When $\kappa = \infty$, we instead use the smallest positive eigenvalue of $X^T X$ to compute μ.

For the logistic regression experiments, the cost of (ii) at each iteration of AGCD and ASCD can be much larger than $O(p)$ because there is no easy way to update the full gradient $\nabla f(\cdot)$. For these problems we have

$$\nabla f(\beta) = -\frac{1}{n} X^T w(\beta) \tag{13}$$

where X is the sample matrix with x_i composing the i-th row, and $w(\beta)_i := \frac{1}{1+\exp(y_i \beta^T x_i)}$. Notice that calculating $w(\beta)$ can be done using a rank 1-update with cost $O(n)$. But calculating the matrix-vector product $X^T w(\beta)$ will cost $O(np)$, which dominates the cost of (i) and/or (iii). However, in the case when X is a sparse matrix with density ρ, the cost can be decreased to $O(\rho np)$.

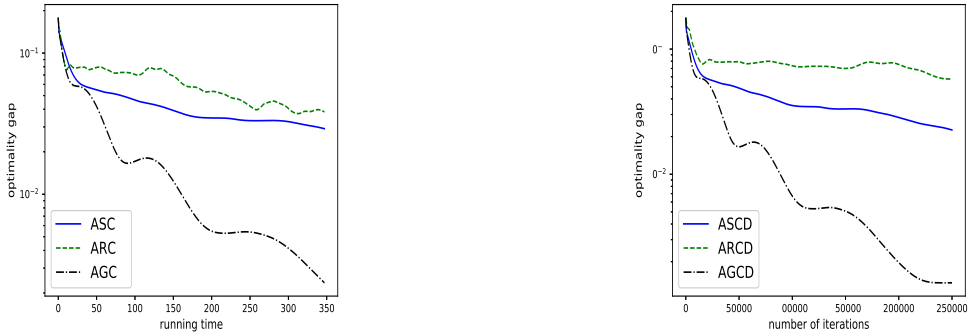

Figure 1: Plots showing the optimality gap versus run-time (in seconds) on the left and versus the number of iterations on the right, for the logistic regression instance madelon with $\bar{\mu} = 10^{-7}$, solved by ASCD, ARCD and AGCD.

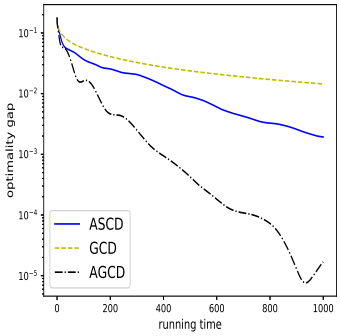

Figure 2: Plots showing the optimality gap versus run-time (in seconds) for the logistic regression instance madelon with $\bar{\mu} = 10^{-6}$, solved by ASCD, GCD and AGCD.

3.2 Comparing the Algorithms using Running Time and the Number of Iterations

Figure 1 shows the optimality gap versus running time (seconds) in the left plot and and versus the number of iterations in the right plot, logistic regression problem using the dataset a1a in LIBSVM [2], with $\bar{\mu} = 10^{-7}$. Here we see that AGCD and ASCD are vastly superior to ARCD in terms of the number of iterations, but not nearly as much in terms of running time, because one iteration of AGCD or ASCD can be more expensive than an iteration of ARCD.

3.3 Comparing Accelerated Method with Non-Accelerated Method

Figure 2 shows the optimality gap versus running time (seconds) of GCD, ASCD and AGCD for the logistic regression problem instance madelon in LIBSVM [2], with $\bar{\mu} = 10^{-6}$. Here we see that ASCD and AGCD are superior to non-accelerated GCD.

3.4 Numerical Results for Logistic Regression with Other Datasets

We present numerical results for logistic regression problems for several other datasets in LIBSVM solved by ASCD, ARCD and AGCD in Figure 3. Here we see that AGCD always has superior performance as compared to ASCD and ARCD, and ASCD outperforms ARCD in most of the instances.

4 Regarding Connections with a Concurrent Paper [5]

In a concurrent paper [5], the authors develop computational theory for matching pursuit algorithms, which can be viewed as a generalized version of greedy coordinate descent where the directions do not need to form an orthogonal basis. The paper also develops an accelerated version of the matching pursuit algorithms, which turns out to be equivalent to the algorithm ASCD discussed here in the special case where the chosen directions are orthogonal. Although the focus in [5] and in our paper are different – [5] is more focused on (accelerated) greedy direction updates along a certain linear subspace whereas our focus is on when and how one can accelerate greedy coordinate updates – both of the works share a similar spirit and similar approaches in developing accelerated methods. Moreover, both works use a decoupling of the coordinate update for the $\{x^k\}$ sequence (with a greedy rule) and the $\{z^k\}$ sequence (with a randomized rule). In fact, [5] is consistent with the argument in our paper as to why one cannot accelerate greedy coordinate descent in general.

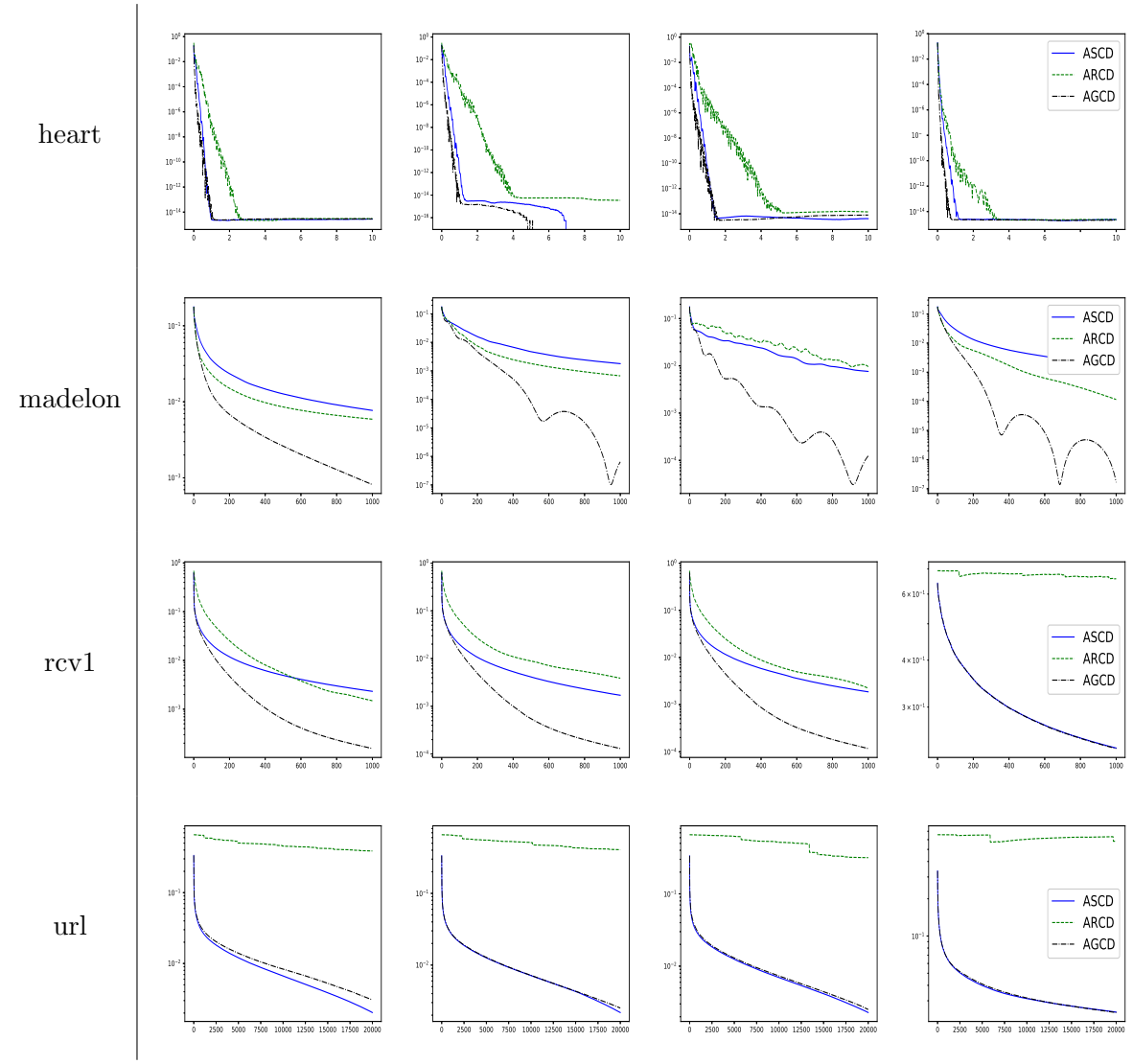

Figure 3: Plots showing the optimality gap versus run-time (in seconds) for some other logistic regression instances in LIBSVM, solved by ASCD, ARCD and AGCD.

References

[1] Zeyuan Allen-Zhu, Zheng Qu, Peter Richtarik, and Yang Yuan, *Even faster accelerated coordinate descent using non-uniform sampling*, International Conference on Machine Learning, 2016.

[2] Chih-Chung Chang and Chih-Jen Lin, *Libsvm: a library for support vector machines*, ACM transactions on intelligent systems and technology (TIST) **2** (2011), no. 3, 27.

[3] Yin Tat Lee and Aaron Sidford, *Efficient accelerated coordinate descent methods and faster algorithms for solving linear systems*, Proceedings of the 2013 IEEE 54th Annual Symposium on Foundations of Computer Science (Washington, DC, USA), FOCS '13, IEEE Computer Society, 2013, pp. 147–156.

[4] Qihang Lin, Zhaosong Lu, and Lin Xiao, *An accelerated randomized proximal coordinate gradient method and its application to regularized empirical risk minimization*, SIAM Journal on Optimization **25** (2015), no. 4, 2244–2273.

[5] Francesco Locatello, Anant Raj, Sai Praneeth Reddy, Gunnar Rätsch, Bernhard Schölkopf, Sebastian U Stich, and Martin Jaggi, *Revisiting first-order convex optimization over linear spaces*, arXiv preprint arXiv:1803.09539 (2018).

[6] Yu Nesterov, *Efficiency of coordinate descent methods on huge-scale optimization problems*, SIAM Journal on Optimization **22** (2012), no. 2, 341–362.