8 Coordinate descent methods

8.1 Introduction

Coordinate-based algorithms solve optimization problems by advancing along coordinate directions towards a solution. Namely, they iteratively (and approximately) minimize the objective along one or a handful of coordinates at a time. These methods have a long history, being one of the first algorithms proposed for solving optimization problems computationally. Before the rise of modern day huge-scale applications, these methods were regarded as too simple/simplistic and examples of failure as below consolidated a view and search in the optimization community for better, more sophisticated methods, based on using (full) derivatives. They continued to be used in *derivative-free* optimization though, where access to derivatives is impossible/too expensive[14] However, these methods were brought to the forefront of research with the need to solve ever-increasingly large scale problems in terms of the number of variables/parameters (sparse optimization in signal processing, supervised learning and more, as seen at the start of this course), for the past twenty years or so (though some prominent researchers such as Paul Tseng (and others) foresaw the upcoming need for such methods and analysed them already in the 1990s).

We return here to the general form of the unconstrained optimization problem, encountered in Section 4 (equation (8)), namely,

$$\min_{x \in \mathbb{R}^n} f(x) \tag{66}$$

where $f : \mathbb{R}^n \to \mathbb{R}$ is continuously differentiable; extensions to the regularized (also called composite) optimization formulation in Section 5 (equation (15)) are also available/popular and may be briefly described later on, time permitting. (Note that (66) includes, of course, the case when f is a sum of functions that we studied in the last chapter- but no special exploitation of such a structure is used here.)

Coordinate-based methods proceed by successively fixing most components of the variables and (approximately) minimizing the objective along the remaining (small number of) variables; for example, in the first iteration, we would be minimizing say, only in the first two coordinates $f(x_1, x_2, 0, 0, ...0)$. Thus at each iteration, we are only addressing/dealing with a small-dimensional optimization problem/iterate calculation - so each iteration is (far) less computationally expensive compared to the case when we would be calculating a change to each entry in the iterate x^k. To clarify, at x^k, the general approach approximately solves

$$\min_{t \in \mathbb{R}^b} f(x^k + U_{\mathcal{B}} t) = f(x_1^k, \dots x_{i_1}^k + t_1, \dots, x_{i_b}^k + t_b, x_{b+1}^k, \dots, x_n^k),$$

[14] The format of these methods is a bit different in that context, and we refer the reader to so-called *direct-search* methods for a closer look.

where $t = (t_1, \ldots, t_b)^T$, $b = |\mathcal{B}|$, and $U_{\mathcal{B}} = [e_{i_1}\ e_{i_2}\ \ldots e_{i_b}]$ is an $n \times b$ matrix with columns e_{i_j}, the i_jth coordinate vector in $\mathbb{R}^n$; $b \ll n$. The matrix $U_{\mathcal{B}}$ changes at the next iteration.

More simply, Coordinate Descent (CD) variants can be viewed as variants of steepest descent that only move along the negative gradient components for a subset of the variables at a time. Namely, at iteration $k \geq 0$, given the (current) iterate x^k, the CD algorithm constructs an update x^{k+1} to x^k as follows,

$$x^{k+1} = x^k - \alpha_k g^k, \tag{67}$$

where $\alpha_k > 0$ is as in earlier lectures, the step-length (can be constant, or varying, adaptive and often pre-defined at the start of the algorithm); and $g^k := \nabla_{\mathcal{B}_k} f(x^k)$, for coordinates' block $\mathcal{B}_k \subset \{1, 2, \ldots, n\}$, has the gradient components $\dfrac{\partial f}{\partial x_i}(x^k)$ for $i \in \mathcal{B}_k$ and is set to zero on the remaining entries,

$$g_i^k = \begin{cases} \dfrac{\partial f}{\partial x_i}(x^k) & i \in \mathcal{B}_k \\ 0, & i \in \{1, \ldots, n\} \setminus \mathcal{B}_k. \end{cases} \tag{68}$$

If $|\mathcal{B}_k| = n$, then (68) is just gradient descent. If $|\mathcal{B}_k| = 1$, then we are only moving along one coordinate, and (68) becomes

$$g^k = \frac{\partial f}{\partial x_{i_k}}(x^k) e_{i_k} = \left(0, \ldots, \frac{\partial f}{\partial x_{i_k}}(x^k), \ldots, 0\right)^T, \tag{69}$$

where $\mathcal{B}_k = \{i_k\} \subset \{1, \ldots, n\}$ and e_{i_k} is the i_kth coordinate vector.

A summary of the CD algorithm is given next.

Coordinate descent (CD) method

Algorithm 4 (CD). *Given $x^0 \in \mathbb{R}^n$, for $k = 0, 1, 2, \ldots$, repeat:*

```
select $\mathcal{B}_k \subset \{1, \ldots, n\}$
calculate $g^k = \nabla_{\mathcal{B}_k} f(x^k)$ according to (68)
calculate $x^{k+1} = x^k - \alpha_k g^k$
```

The choice of block of coordinates can be as follows:

- *randomized*: choose $\mathcal{B}_k$ i.i.d. $\sim U(\{1, \ldots, n\}) \rightarrow (|\mathcal{B}_k| \geq 1$ leads to a randomized CD algorithm); other randomized choices possible.
- *cyclic*: $|\mathcal{B}_k| = 1$; $\mathcal{B}_0 = 1$ and $\mathcal{B}_{k+1} = [\mathcal{B}_k \text{mod} n] + 1$ (cycles through coordinate directions in order; leads to a deterministic algorithm); other choices are possible such as reversing the order in which coordinates are traversed.

- *Gauss-Southwell*: $|\mathcal{B}_k| = 1$; $\mathcal{B}_k = \arg\max_{1\leq i\leq n} \left|\frac{\partial f}{\partial x_i}(x^k)\right|$ (choose the components corresponding to the largest componentwise decrease in the gradient; leads to a deterministic algorithm) [See Problem Sheet 4]

Terminating CD algorithms is similarly problematic to SGD ones; evaluating when progress is stagnating and monitoring successive block gradient values are common heuristics.

There are some positive differences to stochastic gradient methods. Firstly, g^k here is a descent direction (whether it is random or not), whenever it is nonzero[15] $\nabla f(x^k)^T(-g^k) = -\sum_{i\in\mathcal{B}_k}\left(\frac{\partial f}{\partial x_i}(x^k)\right)^2 < 0$, due to the definition of g^k and the Euclidean inner product. Secondly, due to this, we are able to measure/ensure decrease in f more easily, typically; there is monotonic decrease in f, rather than stochastic. Also, note that the variance of G^k here shrinks to zero as k increases since every component of $\nabla f(x^*)$ is zero at a solution; this may not be the case for SGD, since the gradient of some f_i in the sum of functions may not be zero at a stationary point of f. Thus as we will see, there is no limiting 'noise' level for CD as there is for SGD.

Illustration. Figure 8.1 illustrates the iterates (numbered) of randomized coordinate descent method on a two-dimensional scaled quadratic; blocks of size one are chosen uniformly at random (with replacement).

It is not uncommon though, to notice oscillations/failure or very slow convergence/stagnation when experimenting with both the deterministic and randomized variants of block methods.

8.2 Global convergence of randomized CD methods

Assumptions needed for convergence Until now, we have used the following assumption of sufficient smoothness of f, namely, that the gradient is Lipschitz continuous with constant L, so that for all x and y in $\mathbb{R}^n$,

$$\|\nabla f(x+d) - \nabla f(x)\| \leq L\|d\|. \qquad (70)$$

Here we use a more refined version of the L-smoothness of f, namely we assume the existence of (individual) *component Lipschitz constants* L_i, for each $i \in \{1,\ldots,n\}$, such that

$$\left|\frac{\partial f}{\partial x_i}(x+te_i) - \frac{\partial f}{\partial x_i}(x)\right| \leq L_i|t|, \qquad (71)$$

for all $x \in \mathbb{R}^n$ and all $t \in \mathbb{R}$. We define the *componentwise Lipschitz constant* $L_{\max}$ as

$$L_{\max} := \max_{i\in\{1,\ldots,n\}} L_i. \qquad (72)$$

[15] But note that g^k could be zero even if we are not at a stationary point.

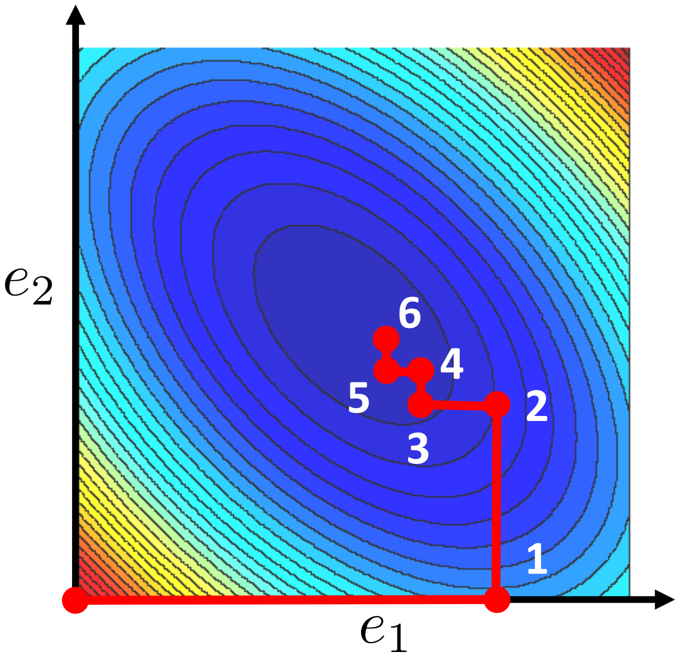

Figure 15: Randomized CD method applied to a scaled quadratic function [@Peter Richtarik, Kaust].

L and $L_{\max}$ are connected as follows (see Problem Sheet 4)

$$L_{\max} \leq L \leq nL_{\max}. \tag{73}$$

A useful property follows, an overestimation property along coordinate directions, which is similar to Lemma 5 and to Proposition 2(iv) and equation (11) (Lectures 1–8). Note that as in SGD, the random choice of (block) coordinates introduces randomness at each iteration. Thus again, we talk about random iterates X^k and a random estimator G^k of $\nabla f(X^k)$. Again we will have conditional expectation depending on the choice of block/random coordinate, and also total expectation depending on the past history (which here will again reduce to the 'history' of the previous iterate).

Lemma 7. *[An overestimation property for coordinate directions] Assume f satisfies (71) and (72). Then for each $i \in \{1, \ldots, n\}$, $x \in \mathbb{R}^n$ and $\alpha \geq 0$, we have*

$$f(x - \alpha g) \leq f(x) - \alpha \left(1 - \frac{L_{\max}}{2}\alpha\right) \left(\frac{\partial f}{\partial x_i}(x)\right)^2, \tag{74}$$

where $g = \nabla_i f(x) = \frac{\partial f}{\partial x_i}(x) e_i$. Apply randomized CD method to (66) with $|\mathcal{B}_k| = 1$

and $\mathcal{B}_k$ chosen uniformly at random (with replacement) and $\alpha^k > 0$. Then

$$\mathbb{E}_{\mathcal{B}_k}\left[f(X^{k+1})\right] \leq f(X^k) - \frac{\alpha^k}{n}\left(1 - \frac{L_{\max}\alpha^k}{2}\right)\|\nabla f(X^k)\|^2, \tag{75}$$

where $\mathbb{E}_{\mathcal{B}_k}$ denotes conditional expectation with respect to the random variable $\mathcal{B}_k$.

Proof. Given that we now want to make use of the coordinate Lipschitz constants L_i, we need to prove a slightly different variant of equation (11) (Lectures 1–8). Use the Taylor expansion $f(x+d) = f(x) + \nabla f(x)^T d + \int_0^1 d^T[\nabla f(x+td) - \nabla f(x)]dt$ with $d := -\alpha g$ to deduce

$$\begin{aligned} f(x-\alpha g) &= f(x) + \nabla f(x)^T(-\alpha g) + \textstyle\int_0^1 [\nabla f(x - t\alpha g) - \nabla f(x)]^T(-\alpha g)dt, \\ &\leq f(x) - \alpha g^T \nabla f(x) + \alpha\|g\| \textstyle\int_0^1 \|\nabla f(x - t\alpha g) - \nabla f(x)\|dt, \end{aligned}$$

where in the first inequality, we used the Cauchy-Schwarz inequality. Now note that by definition of g, we have $g^T\nabla f(x) = \left[\frac{\partial f}{\partial x_i}(x)\right] e_i^T \nabla f(x) = \left[\frac{\partial f}{\partial x_i}(x)\right]^2$. Also,

$$\begin{aligned} \|\nabla f(x - t\alpha g) - \nabla f(x)\| &= \left\|\nabla f\left(x - t\alpha\tfrac{\partial f}{\partial x_i}(x)e_i\right) - \nabla f(x)\right\| \\ &\leq L_i\left|t\alpha\tfrac{\partial f}{\partial x_i}(x)\right| \\ &\leq L_{\max} t\alpha\left|\tfrac{\partial f}{\partial x_i}(x)\right|. \end{aligned}$$

where we used (71). Thus we deduce

$$f(x-\alpha g) \leq f(x) - \alpha\left[\frac{\partial f}{\partial x_i}(x)\right]^2 + L_{\max}\alpha^2\left[\frac{\partial f}{\partial x_i}(x)\right]^2 \int_0^1 t dt$$

which gives (74).

To prove (75), we let $x := X^k$, $g := G^k = \nabla_{\mathcal{B}_k} f(X^k) = \frac{\partial f}{\partial x_{\mathcal{B}_k}}(X^k)e_{\mathcal{B}_k}$ and $\alpha := \alpha^k$ in (74) and apply expectation with respect to $\mathcal{B}_k$ on both sides of the ensuing (74),

$$\mathbb{E}_{\mathcal{B}_k}\left[f(X^{k+1})\right] \leq f(X^k) - \alpha^k\left(1 - \frac{L_{\max}}{2}\alpha^k\right)\mathbb{E}_{\mathcal{B}_k}\left[\left(\frac{\partial f}{\partial x_{\mathcal{B}_k}}(X^k)\right)^2\right],$$

where we also used that $X^{k+1} = X^k - \alpha^k G^k$. The definition of $\mathbb{E}_{\mathcal{B}_k}$ gives us

$$\begin{aligned} \mathbb{E}_{\mathcal{B}_k}\left[\left(\tfrac{\partial f}{\partial x_{\mathcal{B}_k}}(X^k)\right)^2\right] &= \textstyle\sum_{i=1}^n \mathbb{E}\left[\left(\tfrac{\partial f}{\partial x_{\mathcal{B}_k}}(X^k)\right)^2 |\mathcal{B}_k = i\right]\mathbb{P}(\mathcal{B}_k = i) \\ &= \textstyle\sum_{i=1}^n \left(\tfrac{\partial f}{\partial x_i}(X^k)\right)^2 \cdot \tfrac{1}{n} = \tfrac{1}{n}\|\nabla f(X^k)\|^2. \end{aligned}$$

Now (75) follows. □

Global convergence of randomized CD: general case We have the following result for the general case when using randomized CD with one coordinate at a time.

Theorem 14. *[Randomized CD method with fixed stepsize: general case] Consider problem (66), with f satisfying (71) and (72) and assume f is bounded below by f_{low} over $\mathbb{R}^n$. Apply randomized CD method to (66) starting at $x^0 \in \mathbb{R}^n$, with $|\mathcal{B}_k| = 1$ and $\mathcal{B}_k$ chosen uniformly at random (with replacement) from $\{1, \ldots, n\}$, independently of the other iterations, and $\alpha_k = \frac{1}{L_{\max}}$. Then for $k \geq 1$,*

$$\min_{0 \leq i \leq k-1} \mathbb{E}[\|\nabla f(X^i)\|^2] \leq \frac{2nL_{\max}(f(x^0) - f_{\text{low}})}{k}, \tag{76}$$

and so the randomized CD method takes at most $k \leq 2nL_{\max}(f(x^0) - f_{\text{low}})\frac{1}{\epsilon}$ iterations to generate $\mathbb{E}[\|\nabla f(X^{k-1})\|^2] \leq \epsilon$.

Remarks

- Theorem 14 implies that and so $\liminf_{k\to\infty} \mathbb{E}[\|\nabla f(X^k)\|^2] = 0$. With more work, one can show that $\lim_{k\to\infty} \mathbb{E}[\|\nabla f(X^k)\|] = 0$, ensuring global convergence in expectation of randomized CD; see Problem Sheet 4.

- Compare this result to the gradient descent method and its general convergence rate; see Theorem 1 (Lectures 1–8) (both have sublinear rate of convergence). Comparing the constants involved: In GD with stepsize $\alpha^k := 1/L$, we see (by squaring the result in Theorem 1 (Lectures 1–8)) that for GD, $\min_{0\leq i\leq k-1} \|\nabla f(x^i)\|^2 \leq \frac{2L(f(x^0)-f_{\text{low}})}{k}$. Thus, besides the distinction of expected gradient versus true gradient being guaranteed to be made small, the remaining difference is that the randomized CD bound is a multiple of $nL_{\max}$, while the GD one is a multiple of L. Recalling relation (73), we see that in the worst case, these bounds coincide. However, it is not uncommon for L and $L_{\max}$ to be similar in magnitude, and so then, the CD bound is worse by a factor of n (which could be large). Thus, as expected, there may be a penalty to pay for using incomplete problem information in the algorith.

- Other stepsize choices are possible and the result in the above theorem continues to hold: for example, a potentially longer stepsize would be to set α^k to $1/L_{\mathcal{B}_k}$, or set to a global minimizer of f along g^k.

Proof. (Theorem 14) Lemma 7, namely (75), holds for each $k \geq 0$ with $\alpha^k = 1/L_{\max}$, and so we have

$$\mathbb{E}_{\mathcal{B}_k}\left[f(X^{k+1})\right] \leq f(X^k) - \frac{1}{2nL_{\max}}\|\nabla f(X^k)\|^2, \quad k \geq 0. \tag{77}$$

Passing to total expectation in (77), namely, taking expectation $\mathbb{E}$ with respect to the past, namely, $\mathcal{B}_0, \ldots, \mathcal{B}_{k-1}$ on both sides of the above, we note that we have a memoryless property so current iterate only depends on previous block, so at the kth iteration, $\mathbb{E} = \mathbb{E}_k := \mathbb{E}(\cdot|\mathcal{B}_0, \ldots, \mathcal{B}_k) = \mathbb{E}_{\mathcal{B}_k}$,, while at iteration $k-1$, $\mathbb{E} = \mathbb{E}_{k-1} = \mathbb{E}_{\mathcal{B}_{k-1}}$ and so on. We obtain

$$\mathbb{E}_k\left[f(X^{k+1})\right] \le \mathbb{E}_{k-1}[f(X^k)] - \frac{1}{2nL_{\max}}\mathbb{E}_{k-1}[\|\nabla f(X^k)\|^2], \quad k \ge 0, \tag{78}$$

which re-arranges to give a lower bound on the expected decrease in f from one iteration to the next,

$$\mathbb{E}_{i-1}[f(X^i)] - \mathbb{E}_i\left[f(X^{i+1})\right] \ge \frac{1}{2nL_{\max}}\mathbb{E}_{i-1}[\|\nabla f(X^i)\|^2], \quad i \ge 0, \tag{79}$$

where by convention $\mathbb{E}_{-1}[f(X^0)] = f(x^0)$ since x^0 is deterministic; similarly for $\nabla f(x^0)$; and where we re-indexed by i instead of k. Following similar approaches to earlier proofs for GD and SGD, we now sum up (79) from $i = 0$ to $i = k$, (for any $k \ge 0$) to deduce that on the left hand side of the summed inequalities we have a telescoping sum and consecutive terms cancel leading to

$$f(x^0) - \mathbb{E}_k\left[f(X^{k+1})\right] \ge \frac{1}{2nL_{\max}}\sum_{i=0}^{k}\mathbb{E}_{i-1}[\|\nabla f(X^i)\|^2]$$

and using $f(X^{k+1}) \ge f_{\text{low}}$ and $\mathbb{E}_{i-1}[\|\nabla f(X^i)\|^2] \ge \min_{0\le i\le k}\mathbb{E}_{i-1}[\|\nabla f(X^i)\|^2]$, we deduce

$$f(x^0) - f_{\text{low}} \ge \frac{1}{2nL_{\max}}(k+1)\min_{0\le i\le k}\mathbb{E}_{i-1}[\|\nabla f(X^i)\|^2],$$

which gives (76) with $k+1$ instead of k, and where we let $\mathbb{E} = \mathbb{E}_{i-1}$. □

Again, as we have seen for earlier methods, the randomized CD performance improves when f is (strongly) convex.

Global convergence of randomized CD: convex and strongly convex cases
We have the following result in the case when f is convex and when using randomized CD with one coordinate at a time.

Theorem 15. *[Randomized CD method with fixed stepsize: convex case] Consider problem (66), with f satisfying (71) and (72) and $f(x) \ge f(x^*)$ for all $x \in \mathbb{R}^n$ and for some $x^* \in \mathbb{R}^n$. Assume also that f is a convex function (Definition 2, Lectures 1–8) and that $\|x - x^*\| \le D$ for all x with $f(x) \le f(x^0)$. Apply randomized CD method to (66) starting at $x^0 \in \mathbb{R}^n$, with $|\mathcal{B}_k| = 1$ and $\mathcal{B}_k$ chosen uniformly at random (with replacement) from $\{1, \ldots, n\}$, independently of the other iterations, and $\alpha_k = \frac{1}{L_{\max}}$. Then for $k \ge 0$,*

$$\mathbb{E}[f(X^k)] - f(x^*) \le \frac{2nL_{\max}D^2}{k}, \tag{80}$$

and so the randomized CD method takes at most $k \le 2nL_{\max}D^2\frac{1}{\epsilon}$ iterations to generate $\mathbb{E}[f(X^k)] - f(x^*) \le \epsilon$.

Remarks. Note that we obtain similar guarantees as for GD method in the same case, with similar distinctions as in the general case. In particular, we obtain sublinear rate $\mathcal{O}(1/k)$ for driving the objective gap (not just the gradient) to being sufficiently small. The theorem also implies that $\mathbb{E}[f(X^k)] - f(x^*) \to 0$ as $k \to \infty$.

The proof of the above theorem is very similar to the case of GD and SGD applied to the same class of convex functions.

Proof. (Theorem 15) The more general conditions of Theorem 14 are satisfied here, and so its proof is valid in this case. We now refine that proof starting at equation (78), in which we take away $f(x^*)$ on each side to obtain that $\Delta_k := \mathbb{E}[f(X^k)] - f(x^*)$ satisfies

$$\Delta_{k+1} \leq \Delta_k - \frac{1}{2nL_{\max}} \mathbb{E}\left[\|\nabla f(X^k)\|^2\right]. \tag{81}$$

As seen in the proof of Theorem 2 (or for Problem Sheet 3), f being convex implies: $0 \leq f(X^k) - f(x^*) \leq \nabla f(X^k)^T(X^k - x^*) \leq \|\nabla f(X^k)\| \cdot \|(X^k - x^*)\| \leq D\|\nabla f(X^k)\|$, where the latter inequality follows from the fact that $f(x^k) \leq f(x^0)$ for all k and so X^k is in the $f(x^0)$ level set of f. Squaring this and passing to total expectation, we deduce $\mathbb{E}[(f(X^k) - f(x^*))^2] \leq D^2\,\mathbb{E}[\|\nabla f(X^k)\|^2]$; the LHS of the latter satisfies $\mathbb{E}[(f(X^k) - f(x^*))^2] \geq (E[(f(X^k)] - f(x^*))^2$ since $\mathrm{var}(f(X^k) - f(x^*)) \geq 0$. Thus $\Delta_k^2 \leq D^2\,\mathbb{E}[\|\nabla f(X^k)\|^2]$. The latter implies

$$\Delta_{k+1} \leq \Delta_k - \frac{1}{2nL_{\max}D^2}\Delta_k^2, \quad k \geq 0,$$

or equivalently,

$$\Delta_k - \Delta_{k+1} \geq \frac{1}{2nL_{\max}D^2}\Delta_k^2, \quad k \geq 0. \tag{82}$$

To resolve the above recurrence, consider the difference of reciprocals

$$\frac{1}{\Delta_{k+1}} - \frac{1}{\Delta_k} = \frac{\Delta_k - \Delta_{k+1}}{\Delta_k\Delta_{k+1}} \geq \frac{\Delta_k - \Delta_{k+1}}{\Delta_k^2},$$

where we used that $\Delta_{k+1} \leq \Delta_k$. This and (82) imply

$$\frac{1}{\Delta_{k+1}} - \frac{1}{\Delta_k} \geq \frac{1}{2nL_{\max}D^2}, \quad k \geq 0. \tag{83}$$

Summing up the inequality (83) for $k = 0$ to $k = i$ (for any $i \geq 0$), we note that the terms on the left hand side are consecutive and so they cancel, apart from the first and last term, while the term on the right-hand side is constant with respect to the summation index. Thus we deduce, for any $i \geq 0$,

$$\frac{1}{\Delta_{i+1}} - \frac{1}{\Delta_0} \geq \frac{i+1}{2nL_{\max}D^2},$$

which gives (84) since $-\frac{1}{\Delta_0} \leq 0$ and if we let $k = i + 1$. □

We have the following result in the case when f is strongly convex and when using randomized CD with one coordinate at a time.

Theorem 16. *[Randomized CD method with fixed stepsize: strongly convex case] Consider problem (66), with f satisfying (71) and (72). Assume also that f is a γ-strongly convex function (Definition 2, Lectures 1–8). Apply randomized CD method to (66) starting at $x^0 \in \mathbb{R}^n$, with $|\mathcal{B}_k| = 1$ and $\mathcal{B}_k$ chosen uniformly at random (with replacement) from $\{1, \ldots, n\}$, independently of the other iterations, and $\alpha_k = \frac{1}{L_{\max}}$. Then for $k \geq 0$,*

$$\mathbb{E}[f(X^k)] - f(x^*) \leq \left(1 - \frac{\gamma}{nL_{\max}}\right)^k (f(x^0) - f(x^*)), \tag{84}$$

and so the randomized CD method takes at most $k \leq \mathcal{O}(|\log \epsilon|)$ iterations to generate $\mathbb{E}[f(X^k)] - f(x^*) \leq \epsilon$.

Remarks. Note that we obtain similar guarantees as for GD method in the same case, with similar distinctions as in the general case. In particular, we obtain linear rate of convergence for driving the objective gap to being sufficiently small. The theorem also implies that $\mathbb{E}[f(X^k)] - f(x^*) \to 0$ as $k \to \infty$, linearly.

The proof of the above theorem is very similar to the case of GD and SGD applied to the same class of strongly convex functions.

Proof. (Theorem 16) The more general conditions of Theorem 14 are satisfied here, and so its proof is valid in this case. We now refine that proof starting at equation (78), in which we take away $f(x^*)$ on each side to obtain that $\Delta_k := \mathbb{E}[f(X^k)] - f(x^*)$ satisfies

$$\Delta_{k+1} \leq \Delta_k - \frac{1}{2nL_{\max}} \mathbb{E}\left[\|\nabla f(X^k)\|^2\right]. \tag{85}$$

(We note that (85) is the same as (81) in Theorem 15 but it will be distinct from that proof from now on, as in the case of strongly convex functions we are able to improve on the bound on $\mathbb{E}\left[\|\nabla f(X^k)\|^2\right]$.) A consequence of the strong convexity property of f is that $f(x) - f(x^*) \leq \frac{1}{2\gamma}\|\nabla f(x)\|^2$ for all x. Thus, letting $x := X^k$, and passing to total expectation, $\Delta_k = \mathbb{E}[f(X^k)] - f(x^*) \leq \frac{1}{2\gamma}\mathbb{E}[\|\nabla f(X^k)\|^2]$. Now we substitute this into (85) to deduce $\Delta_{k+1} \leq \Delta_k - \frac{2\gamma}{2nL_{\max}}\Delta_k$, $k \geq 0$, and so

$$\Delta_{k+1} \leq \left(1 - \frac{\gamma}{nL_{\max}}\right)\Delta_k, \quad k \geq 0. \tag{86}$$

Since $\gamma \leq L \leq nL_{\max}$, the convergence factor in (86) is in $(0, 1]$, and so (84) follows inductively. □

Block methods If $|\mathcal{B}_k| > 1$, we have two possibilities : we partition $\{1, 2, \ldots, n\}$ into blocks (of equal size or importance) and pick one block at random (with replacement). Alternatively, as already mentioned, we choose a size and pick each entry in the block independently at random from $\{1, 2, \ldots, n\}$ (without replacement). (But at the next iteration it is with replacement to allow any component to be picked again). Blocks of adaptive size can also be used.

8.3 Deterministic CD methods

Recall the cyclic and Gauss-Southwell rule for choosing the coordinates to traverse in CD methods. CD methods with Gauss-Southwell rule will be addressed in Problem Sheet 4, while here we discuss briefly some results - or lack of - when using cyclic CD methods.

Cyclic CD methods: general case Examples of failure exist for cyclic CD methods.

Cyclic CD methods: an example of failure In a dedicated paper, MJD Powell ('On search directions for minimization algorithms', *Mathematical Programming*, 1973) constructed an example function in three dimensions on which cyclic CD method fails to converge to a stationary point. The function is

$$f(x_1, x_2, x_3) = -(x_1x_2 + x_2x_3 + x_1x_3) + \sum_{i=1}^{3}(|x_i| - 1)_+^2,$$

where $a_+ = max\{a, 0\}$. This function is nonconvex, continuously differentiable and its minimizers are at the corners $(1, 1, 1)^T$ and $(-1, -1, -1)^T$ of the unit cube; see Figure 8.3. We apply the cyclic CD method to minimizing this f, that calculates the linesearch $\alpha^k \in \mathbb{R}$ as the global minimizer along each coordinate (i.e., the best possible/ideal step as it gives largest decrease in the objective along the search direction). Note that then, cyclic CD is just a cyclic coordinate search method (the gradient component multiple is irrelevant). If the starting point is chosen close to one of the other – non-optimal – vertices of the unit cube, then the cyclic CD method cycles around these non-optimal vertices, between their neighbourhoods. (This example is special and small perturbations of this example function leads to convergent behaviour of cyclic CD. Also, a randomized CD method would typically converge as well.)

It is not uncommon to notice oscillations/failure or very slow convergence or stagnation when experimenting with both the deterministic and randomized variants of block methods.

Cyclic CD methods: convergence results Under some stronger assumptions (than for R-CD and the above) on general objectives, one can show convergence of cyclic coordinate search methods.

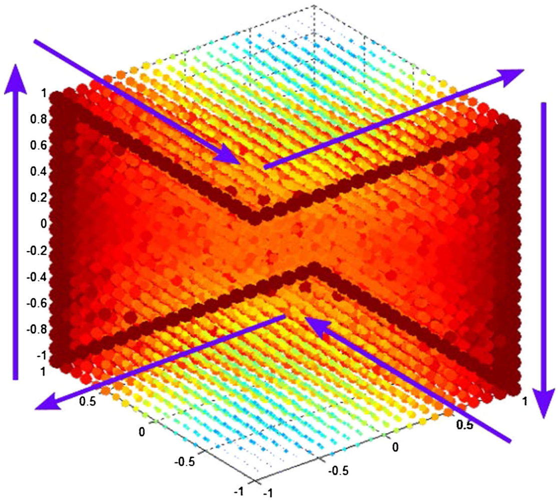

Figure 16: Cyclic CD method applied to Powell's objective: nonconvergent behaviour [3].

Theorem 17 (Cyclic CD for general objectives)**.** *Let f be continuously differentiable and bounded below on $\mathbb{R}^n$*[16]*. Suppose that for each i, the minimum*

$$\min_{t\in\mathbb{R}} f(x_1, x_2, \ldots, x_{i-1}, t, x_{i+1}, \ldots, x_n)$$

is uniquely attained. Apply a variant of cyclic CD to (66) *that globally minimises f along each coordinate direction at a time, starting with e_1, and keeping the others fixed at the current iterate. Then every limit point of the sequence of iterates x^k where $k = n, 2n, 3n, \ldots$, is a stationary point of f.*

This result and the ones that follow can be extended to blocks of variables as long as the blocks form a partition of $\{1, \ldots, n\}$ and we cycle through the partition blocks in order. In particular, we could have two blocks of variables, in which case we have alternating minimization algorithms.

The next theorem follows the more familiar framework of the cyclic CD along negative gradient components.

Theorem 18 (Cyclic CD for convex functions with fixed stepsize)**.** *Consider problem* (66)*, with f satisfying* (71) *and* (72) *and $f(x) \geq f(x^*)$ for all $x \in \mathbb{R}^n$ and for some $x^* \in \mathbb{R}^n$. Assume also that f is a convex function (Definition 2, Lectures 1–8) and that $\|x - x^*\| \leq D$ for all x with $f(x) \leq f(x^0)$. Apply the cyclic CD method (Algorithm 4 with cyclic rule on page 52) to* (66) *starting at $x^0 \in \mathbb{R}^n$, with*

[16] It could be a subset as well, which would make the restrictive assumptions of this result more acceptable.

$\alpha_k = \frac{1}{L_{\max}}$. *Then for* $k \in \{n, 2n, 3n, \ldots\}$,

$$f(x^k) - f(x^*) \leq \frac{4nL_{\max}D^2\left(1 + n\frac{L^2}{L_{\max}^2}\right)}{k+8}. \tag{87}$$

If we compare this result to Theorem 15, since $L \geq L_{\max}$, the bound for cyclic CD has a factor of n^2 compared to a factor of n for randomized CD; the rates are still sublinear, but this result shows deterministic convergence while the earlier one was in expectation.

Similarly, we have a result for strongly convex functions.

Theorem 19 (Cyclic CD for strongly convex functions with fixed stepsize)**.** *Consider problem* (66)*, with* f *satisfying* (71) *and* (72)*. Assume also that* f *is a* γ*-strongly convex function (Definition 2, Lectures 1–8). Apply the cyclic CD method (Algorithm 4 with cyclic rule on page 52) to* (66) *starting at* $x^0 \in \mathbb{R}^n$*, with* $\alpha_k = \frac{1}{L_{\max}}$*. Then for* $k \in \{n, 2n, 3n, \ldots\}$,

$$f(x^k) - f(x^*) \leq \left(1 - \frac{\gamma}{2L_{\max}(1 + nL^2/L_{\max}^2)}\right)^{k/n} (f(x^0) - f(x^*)). \tag{88}$$

Comparing this result to Theorem 16, again we see there is an extra n factor here as $(1-a)^{1/n} \approx (1-a/n)$ and so the denominator in the convergence factor here has n^2 factor.

We can think of the ratio $L/L_{\max}$ as being close to 1 when the function can be decoupled along coordinates, and being large(r) otherwise; indeed for the former case it is when we expect CD methods to work well (whether randomized or deterministic).

8.4 Perspectives

CD methods are popular candidates for parallelization, an active area of research. Accelerated variants of CD exist, as well as combinations with proximal methods for composite or regularized optimization problems. They can be applied to many problems including the iterative solution of linear systems.

Other variants move beyond block coordinates and consider random subspace methods. Namely, instead of choosing a subspace aligned with the coordinate directions at each iteration, they choose a random subspace (generated for example by the column space of a Gaussian random matrix, a matrix with iid entries from the standard normal distribution). As long as the subspace gradient can be related to the full gradient – which it usually possible, probabilistically – such methods have almost sure convergence (not just in expectation).