Fast Convergence of Greedy 2-Coordinate Updates for Optimizing with an Equality Constraint

author names withheld

Under Review for OPT 2022

Abstract

In this work, we study the Block Coordinate Descent (BCD) algorithm with greedy block selection rule for minimizing functions with one linear equality constraint. We show a faster linear convergence rate for BCD with block size 2 (2-BCD) on functions satisfying the Polyak-Lojasiewicz inequality. Our analysis exploits similarity between the solutions of 2-BCD and equality-constrained steepest descent (SD) in the L1-norm. This yields a simple proof.

1. Introduction

Block coordinate descent (BCD) is an iterative algorithm, where in each iteration, the objective function is improved in a chosen block coordinate.More precisely, BCD optimizes a differentiable function $f(\mathbf{x}), \mathbf{x} \in \mathbb{R}^d$, by using iterations of the form

$$\mathbf{x}_{b^k}^{k+1} = \arg\min_{\mathbf{x}_{b^k}} f(\mathbf{x}^k),$$

only at the sub-vector $\mathbf{x}_{b^k}$ chosen by a block selection rule. All other coordinates stay the same. In this work, we study the convergence of BCD for problems with a single linear equality constraint. Specifically, we want to solve the following minimization problem:

$$\min_{\mathbf{x}} f(\mathbf{x}), \quad \text{subject to } \sum_{i=1}^{n} \mathbf{x}_i = c. \tag{1}$$

This problem commonly appears in many machine learning applications and (with additional bound constraints) in the dual-SVM formulation [2, 14].

Recently, BCD has become one of the popular methods in solving large-scale machine learning applications due to its simplicity, easier adaptability to distributed setting and good performance among others. For instance, BCD has been commonly used to efficiently solve some important problems such as Support Vector Machines (SVM) [2, 5] and L1-regularized least squares (LASSO) [4]. Specifically SVMs have been trained using the Sequential minimal optimization (SMO) method [14] and other greedy variants of BCD in the dual problem. For example, one of the most widely used SVM implementations, LIBSVM package [2], updates 2 coordinates greedily in every iteration.

Despite the success and widespread application of BCD, it is known that analyzing convergence of BCD for non-convex problems without any assumptions is challenging. Several works in the past have studied the convergence properties by under the assumptions of either strict convexity or pseudo-convexity of the objective [1, 15]. Others have made assumptions on the structure of f and

© .

proved tighter convergence. [15] analyzes a non-differentiable function assuming separability on the non-differentiable part of the function. Few works in the past have studied the convergence properties of BCD for problem under linear equality constraints as in (1) under assumptions such as convex separability of the objective [8, 10].

None of the previous works on equality-constrained optimization show a theoretical advantage to using greedy rules, despite their success in SVM training. On the other hand, several recent works have explored the effectiveness of greedy rules for unconstrained BCD [6, 7, 12, 13]. The authors of [12] analyses one such greedy selection rule called Gauss-Southwell (GS) which chooses the block with largest gradient norm , $b^k \in \text{argmax}_{b \in \mathcal{B}} ||\nabla_b f(\mathbf{x}^k)||$, where $\mathcal{B}$ is the set of all blocks, $\nabla_b f(\mathbf{x}^k)$ is the gradient of the sub-vector $\mathbf{x}_{b^k}$. They show that GS rule has a significantly faster convergence rate than random coordinate selection in many cases. In the same work, they proved the linear convergence of a strongly convex function using one variant of GS rule, called GS-q rule, which chooses a block that minimizes the function value. The key step behind attaining this convergence is by performing analysis in the L_1 norm utilizing the well studied relationship between L_1 and L_2 norms. [6] derives analogous result as [12], but relaxes the strong convexity assumption by considering f that satisfies Polyak- Lojasiewicz inequality (PL-inequality). [7] has extended this analysis for composite problems for GS-s, GS-r, and GS-q variants of GS rule. It is noteworthy that all these works have studied greedy coordinate descent (GCD) where the block size is 1.

Prior literature on general greedy BCD is sparse [3, 13]. Further, the convergence properties of greedy BCD for a variety of problem settings and structures, especially under equality constraints as in (1) are not well studied. In this work, we aim to address this gap by adapting the techniques used to analyze GCD for block size 1 and for unconstrained problems, to obtain tighter convergence of greedy BCD with block size 2 for minimization problems with linear equality constraints.

In this work, we are primarily focused on analysing BCD with block size 2 *using the GS-q block selection rule with linear equality constraints*. Specifically, we have derived convergence of 2-BCD with GS-q rule with a linear equality constraint under the Prox-PL-inequality [6], which was proposed as a generalized version of PL-inequality. This inequality is weaker than strong convexity of f, and is satisfied by several important models, including SVMs. The convergence rate we present in this work is faster than existing rates, is the first rate to show an advantage of the type of greedy coordinate selection used in practice, and unlike previous rates is dimension independent.

2. Notations and Definitions

We now introduce our notation and definitions. We call the indices of parameter a vector $x \in \mathbb{R}^n$, *coordinates*. The set of coordinates is $\mathcal{C} = \{1, 2, ..., n\}$. Our algorithm will operate on *coordinate blocks* $x_b \subseteq x$, where $b \subseteq \mathcal{C}$. We use $\mathcal{B}$ to indicate the set of all coordinate blocks.

The notation $d_b \in \mathbb{R}^{|b|}$ denotes the elements d corresponding to the coordinate block b and $\nabla_b f(x) \in \mathbb{R}^{|b|}$ denotes the gradient of the block. The function $U_b : \mathbb{R}^{|b|} \to \mathbb{R}^n$ is the mapping,

$$[U_b(d_b)] = \begin{cases} [d_b]_j, & \text{if } j \in b \\ 0, & \text{otherwise.} \end{cases}$$

A basic property we assume is that f has Lipschitz-continuous gradients. This is a standard assumption for coordinate descent methods [9–11, 16]. The definitions for GS-q rule and Proximal Polyak-Lojasiewicz inequality are given in the Appendix A.1.

3. Algorithm and Theoretical Results

Let x^0 be a feasible initial iterate and $f(x)$ be the function to optimize. Let $d_b \in \mathbb{R}^{|b|}$ be the descent direction. We analyze a block coordinate descent algorithm that chooses a coordinate block b^k according to the GS-q rule assuming block size $|b| = 2$ (similar to the early versions of libSVM). Throughout the paper, we call it **q-2-GCD**. Without loss of generality, we will assume that $b = \{1, 2\}$.

Proposition 1 *For the optimization problem (1), the GS-q rule selects the optimal block $b = [i, j]$, where $i = \text{argmax}_i \nabla_i f(x), j = \text{argmin}_j \nabla_j f(x)$ by descending in the direction $d_b = -\frac{1}{L}\left(\nabla_b f(x) - \frac{1}{2}\langle \nabla_b f(x), 1\rangle 1\right)$.*

Proof *For the optimization problem (1), the GS-q rule selects the optimal block b, by solving the following minimization problem:*

$$b = \underset{b}{\text{argmin}} \min_{d_b | d_{b_1} + d_{b_2} = 0} \langle \nabla_b f(x), d_b \rangle + \frac{L}{2}||d_b||^2, \tag{2}$$

where d_b is the descent direction. To prove the Proposition 1, we first fix b and solve for d_b and vice versa. The full proof to Proposition 1 is produced in the appendix A.2. ■

The algorithm q-2-GCD can be summarized as follows:

Algorithm 1 Greedy Block Coordinate Descent with GS-q Rule for block size 2 (q-2-GCD)

1: At iteration k, choose a block $b^k = \{i, j\}$ that solves: $b^k = \text{argmax}_{i,j} |\nabla_i f(x^k) - \nabla_j f(x^k)|$.
2: Set $x^{k+1} = x^k + U_{b^k}(d_{b^k})$, where $d_{b^k} = -\frac{1}{L}\left(\nabla_b f(x^k) - \frac{1}{2}\langle \nabla_b f(x^k), 1\rangle 1\right)$.

3.1. Convergence Analysis

In this section, we prove that q-2-GCD converges linearly for functions satisfying the proximal-PL inequality [6]. The key step here is the observation that the descent direction d_b of q-2-GCD is (almost) recovered by computing the equality-constrained steepest descent direction in the $1-$norm.

The challenge in analyzing q-2-GCD stems from the presence of equality constraints in (1). Without constraints, and assuming PL-inequality of $f(x)$, previous works [13] showed that linear convergence can be derived by using GS-q rule in the progress bound and applying the parameter update, where the progress is measured by a norm. Now, let us derive the progress bound of q-2-GCD and see if we could do the same.

Proposition 2 *The progress of q-2-GCD for the problem in 1 can be measured as $f(x^{k+1}) = f(x^k) - \frac{1}{4L}\max_b(\nabla_i f(x^k) - \nabla_j f(x^k))^2$.*

Proof *The proof to Proposition (2) is in appendix A.3.* ■

Here, the progress depends on the Lipschitz constant and maximum difference between the partial derivatives and there is no straight-forward way to measure this progress in terms of a norm. This makes it harder to derive linear convergence using existing methods. Whereas, when $b = 1$, the update of the descent direction d_b has no coupling of block coordinate gradients, making it easier to represent the progress as a norm [8, 10].

The core idea of our analysis is deriving the convergence of steepest descent in 1-norm as a proxy for q-2-GCD, thereby bypassing the complication of having block coordinates in the progress bound. The idea stems from the observation that the descent direction d of the steepest descent in 1-norm algorithm is almost equivalent to d_b of q-2-GCD as in (13), with zeros for all elements in d except the two updated by the q-2-GCD and those coordinates follows (13).

3.1.1. STEEPEST SUBSPACE DESCENT IN 1-NORM

The steepest descent method finds the descent direction that minimizes the function value in every iteration. That is,

$$d = \underset{d \in \mathbb{R}^n | d^T 1 = 0}{\text{argmin}} \nabla f(x)^T d + \frac{L}{2}||d||_1^2. \tag{3}$$

Taking gradient with respect to d and λ of the Lagrangian of (3),

$$\nabla_d \mathcal{L}(d, \lambda) = \nabla f(x) + \frac{L}{2}\partial||d||_1^2 + \lambda 1 \text{ and } \nabla_\lambda \mathcal{L}(d, \lambda) = d^T 1.$$

At a solution we need $d^T 1 = 0$. Now from $\nabla_d \mathcal{L}(d, \lambda)$ we get,

$$\frac{2}{L}(-\nabla f(x) - \lambda 1) \in \partial||d||_1^2 \quad \Rightarrow \quad \frac{2}{L}(-\nabla f(x) - \lambda 1) \in ||d||_1 sgn(d).$$

This means that any optimal d must satisfy the above condition. We will use this as a test to show the similarity between the solutions recovered by steepest descent and q-2-GCD and this will be the basis of our convergence analysis.

3.1.2. SIMILARITY BETWEEN STEEPEST DESCENT AND Q-2-GCD

In this section we will argue that the optimal solution found by q-2-GCD can also be recovered by steepest descent in 1-norm.

Lemma 3 *Consider the problem* (1)*. The optimal direction* d *solved by the steepest descent in* $1-$ *norm algorithm recovers the direction* d_b *solved by q-2-GCD with zeroes in all its coordinates except the* 2 *coordinates chosen and updated by q-2-GCD. Precisely,*

$$U_b(d_b)_{GCD} = 2d_{SD},$$

for a particular solution of the steepest descent problem.

Proof *We begin by proposing a* d *and show that this* d *respects the steepest descent optimality condition, derived in* (4)*. Now, one possible solution for* d *would be where* d *is a vector of all zeros except* 2 *coordinates* d_i *and* d_j*, where* $i = \text{argmax}_i \nabla_i f(x)$ *and* $j = \text{argmin}_j \nabla_j f(x)$*, the coordinates chosen by q-2-GCD. We prove that this* d *can be a solution by showing that* (4) *is satisfied by every coordinate of* d*. The full proof to this lemma is in appendix A.4.* ■

Therefore, we can now analyze the convergence of q-2-GCD by directly analyzing steepest descent in 1-norm. We will express our results using Proximal-PL inequality [6] in the 1-norm. The definition of the Prox-PL inequality in 1-norm and its relationship with Prox-PL inequality in 2-norm is in appendix A.1.1.

3.1.3. CONVERGENCE RATE OF Q-2-GCD

Theorem 4 *Assume a function $F(x) = f(x) + g(x)$, where $g(x) = 0$ if $x^T 1 = c$ and ∞ otherwise, with Lipschitz continuous gradient, satisfies Prox-PL inequality in 1-norm (9) for some $\mu_1 > 0$. Then, the Algorithm q-2-GCD for the minimization problem in (1) obtains the following linear convergence rate:*

$$F(x^{k+1}) - F(x^*) \leq \left(1 - \frac{\mu_1}{4L}\right)^k (F(x^0) - F^*). \tag{4}$$

Proof From the relationship between steepest descent and q-2-GCD as shown in the previous sections, the descent lemma of q-2-GCD can be written as

$$F(x^{k+1}) \leq F(x^k) + \min_{d|d^T 1=0} \nabla F(x^k)^T (2d_{SD}) + \frac{L}{2}||2d_{SD}||_1^2 + g(y) - g(x^k).$$

Here, $f(x)$ is reformulated as $F(x) = f(x) + g(x)$ where, $g(x) = \begin{cases} 0 & \text{if } d^T 1 = 0 \\ \infty & \text{otherwise.} \end{cases}$.
Change of notation: WLOG d_{SD} is written simply as d from this point.

$$\begin{aligned} F(x^{k+1}) &\leq F(x^k) + \min_{d|d^T 1=0} 2\nabla f(x^k)^T d + 2L||d||_1^2 + g(y) - g(x^k) \\ &= F(x^k) + 2 \min_{d|d^T 1=0} \nabla f(x^k)^T d + L||d||_1^2 + g(y) - g(x^k). \end{aligned} \tag{5}$$

Applying (Prox-PL Inequality in 1-norm) in (5),

$$F(x^{k+1}) \leq F(x^k) - \frac{1}{2L}\mathcal{D}_g(x, 2L).$$

Therefore,

$$F(x^{k+1}) - F(x^k) \leq -\frac{\mu_1}{4L}(F(x^k) - F^*), \quad \Rightarrow \quad F(x^{k+1}) - F(x^*) \leq \left(1 - \frac{\mu_1}{4L}\right)(F(x^k) - F^*).$$

Subtracting F^* on both sides and applying the inequality recursively completes the proof. ■

This shows the method's linear convergence to optimal F^* and this rate is independent of problem dimension.

4. Conclusion

In this work we analyzed a greedy block-coordinate descent algorithm for linearly-constrained objectives which we call q-2-GCD. This has been a very successful method with various applications and multiple works have analyzed this method. We provide tighter convergence rate than the existing works. The key idea behind our analysis was the heuristic notion that greedy block-selection rules can only improve in performance as the set of coordinate blocks is expanded. We showed the convergence of q-2-GCD on functions satisfying proximal PL inequality thereby furthering our understanding of greedy methods.

References

[1] Amir Beck and Luba Tetruashvili. On the convergence of block coordinate descent type methods. *SIAM journal on Optimization*, 23(4):2037–2060, 2013.

[2] Chih-Chung Chang and Chih-Jen Lin. Libsvm: a library for support vector machines. *ACM transactions on intelligent systems and technology (TIST)*, 2(3):1–27, 2011.

[3] Dominik Csiba and Peter Richtárik. Global convergence of arbitrary-block gradient methods for generalized polyak-{\L} ojasiewicz functions. *arXiv preprint arXiv:1709.03014*, 2017.

[4] Wenjiang J Fu. Penalized regressions: the bridge versus the lasso. *Journal of computational and graphical statistics*, 7(3):397–416, 1998.

[5] Thorsten Joachims. 11 making large-scale support vector machine learning practical. In *Advances in kernel methods: support vector learning*, page 169. MIT press, 1999.

[6] Hamed Karimi, Julie Nutini, and Mark Schmidt. Linear convergence of gradient and proximal-gradient methods under the polyak-łojasiewicz condition. In *Joint European conference on machine learning and knowledge discovery in databases*, pages 795–811. Springer, 2016.

[7] Sai Praneeth Karimireddy, Anastasia Koloskova, Sebastian U Stich, and Martin Jaggi. Efficient greedy coordinate descent for composite problems. In *The 22nd International Conference on Artificial Intelligence and Statistics*, pages 2887–2896. PMLR, 2019.

[8] I Necoara, Y Nesterov, and F Glineur. A random coordinate descent method on large optimization problems with linear constraints. *Technical Report, University Politehnica Bucharest*, 2011.

[9] Ion Necoara and Andrei Patrascu. A random coordinate descent algorithm for optimization problems with composite objective function and linear coupled constraints. *Computational Optimization and Applications*, 57(2):307–337, 2014.

[10] Ion Necoara, Yurii Nesterov, and François Glineur. Random block coordinate descent methods for linearly constrained optimization over networks. *Journal of Optimization Theory and Applications*, 173(1):227–254, 2017.

[11] Yu Nesterov. Efficiency of coordinate descent methods on huge-scale optimization problems. *SIAM Journal on Optimization*, 22(2):341–362, 2012.

[12] Julie Nutini, Mark Schmidt, Issam Laradji, Michael Friedlander, and Hoyt Koepke. Coordinate descent converges faster with the gauss-southwell rule than random selection. In *International Conference on Machine Learning*, pages 1632–1641. PMLR, 2015.

[13] Julie Nutini, Issam Laradji, and Mark Schmidt. Let's make block coordinate descent converge faster: Faster greedy rules, message-passing, active-set complexity, and superlinear convergence. *Journal of Machine Learning Research*, 23(131):1–74, 2022.

[14] John Platt. Sequential minimal optimization: A fast algorithm for training support vector machines. Technical report, Microsoft, April 1998.

[15] Paul Tseng. Convergence of a block coordinate descent method for nondifferentiable minimization. *Journal of optimization theory and applications*, 109(3):475–494, 2001.

[16] Paul Tseng and Sangwoon Yun. Block-coordinate gradient descent method for linearly constrained nonsmooth separable optimization. *Journal of optimization theory and applications*, 140(3):513–535, 2009.

Appendix A. Additional proofs

A.1. Definitions

Definition 5 *(GS-q rule) The GS-q rule [16] chooses the block that minimizes the function value $f(x)$ the most. For our problem in (1), with linear equality constraints solved with 2-BCD, the GS-q selects a block b as*

$$b = \underset{b}{\operatorname{argmin}} \min_{d_b | d_{b_1} + d_{b_2} = 0} \langle \nabla_b f(x), d_b \rangle + \frac{L}{2} ||d_b||^2,$$

where f is the function to be solved, $d_b \in \mathbb{R}^2$ is the descent direction.

Definition 6 *(Proximal Polyak- Lojasiewicz inequality [6])*
A composite function F of the form $F(X) = f(x) + g(x)$ satisfies Prox-PL inequality with constant μ if

$$\frac{1}{2}\mathcal{D}_g(x, L) \geq \mu(F(x) - F^*),$$

where,

$$\mathcal{D}_g(x, \alpha) \equiv -2\alpha \min_y [\langle \nabla f(x), y - x \rangle + \frac{\alpha}{2}||y - x||^2 + g(y) - g(x)].$$

A.1.1. PROX-PL INEQUALITY IN 1-NORM

In this section, we define Prox-PL inequality in 1-norm. We compare the constants μ_1 and μ_2 corresponding to Prox-PL inequality in 1-norm and 2-norm and show that $\mu_1 \leq \mu_2$. Then, we analyze q-2-GCD assuming the function satisfies Prox-PL inequality in 1-norm and we derive convergence that is tighter and independent of dimension.

A composite function F of form $F(X) = f(x) + g(x)$ satisfies Prox-PL inequality with constant μ_1 when,

$$\frac{1}{2}\mathcal{D}_g(x, L) \geq \mu_1(F(x) - F^*) \qquad \text{(Prox-PL Inequality in 1-norm)}$$

where,

$$\mathcal{D}_g(x, \alpha) = -2\alpha \min_y [\langle \nabla f(x), y - x \rangle + \frac{\alpha}{2}||y - x||_1^2 + g(y) - g(x)]. \tag{6}$$

For the problem as in (1), the update rule of 2-BCD is given by

$$x^{k+1} - x^k = d, \tag{7}$$

where $d = U_b(d_b)$. Applying (7) in (6), we get

$$\mathcal{D}_g(x, \alpha) = -2\alpha \min_y \left[\langle \nabla f(x), d \rangle + \frac{\alpha}{2}||d||_1^2 + g(y) - g(x) \right]. \tag{8}$$

We know that d has non-zero values in only 2 coordinates i and j. Using this (8) becomes,

$$\begin{aligned} \mathcal{D}_g(x, \alpha) &= -2\alpha \min_y [\langle \nabla f(x), d \rangle + \frac{\alpha}{2}(|d_i| + |d_j|)^2 + g(y) - g(x)] \\ &= -2\alpha \min_y [\langle \nabla f(x), d \rangle + \frac{\alpha}{2}(d_i^2 + d_j^2 + 2|d_i||d_j|) + g(y) - g(x)]. \end{aligned} \tag{9}$$

Following similar proof sketch, the Prox-PL inequality in 2-norm, introduced by [6], can be derived as,

$$\mathcal{D}_g(x,\alpha)_{[2-norm]} = -2\alpha \min_y [\langle \nabla f(x), d\rangle + \frac{\alpha}{2}(d_i^2 + d_j^2) + g(y) - g(x)]. \tag{10}$$

It is easy to see that,

$$2\alpha \min_y \left[\langle \nabla f(x), d\rangle + \frac{\alpha}{2}(d_i^2 + d_j^2 + 2|d_i||d_j|) + g(y) - g(x)\right] \geq$$
$$2\alpha \min_y \left[\langle \nabla f(x), d\rangle + \frac{\alpha}{2}(d_i^2 + d_j^2) + g(y) - g(x)\right].$$

Therefore,

$$-\mu_1(F(x) - F^*) \geq -\mu_2(F(x) - F^*). \tag{11}$$

This shows that a bound obtained by assuming Prox-PL in 1-norm can be least as tight as assuming Prox-PL in 2-norm.

A.2. Proof to Proposition (1)

For the optimization problem (1), the GS-q rule selects the optimal block b, by solving the following minimization problem:

$$b = \operatorname*{argmin}_b \min_{d_b | d_{b_1} + d_{b_2} = 0} \langle \nabla_b f(x), d_b\rangle + \frac{L}{2}||d_b||^2, \tag{12}$$

where d_b is the descent direction. First let us fix b and solve for d_b.

Solving for d_b. The Lagrangian of (12) is,

$$\mathcal{L}(d_b, \lambda) = \langle \nabla_b f(x), d_b\rangle + \frac{L}{2}||d_b||^2 + \lambda(d_{b_1} + d_{b_2}).$$

Taking the gradient with respect to d_b gives,

$$\nabla_{d_b}\mathcal{L}(d_b, \lambda) = \nabla_b f(x) + Ld_b + \lambda 1.$$

Setting the gradient equal to 0 and solving for d_b gives,

$$d_b = -\frac{1}{L}(\nabla_b f(x) + \lambda 1). \tag{13}$$

From our constraint, $d_{b_1} + d_{b_2} = 0$, we get

$$0 = -\frac{1}{L}\left(\nabla_1 f(x) + \lambda + \nabla_2 f(x) + \lambda\right),$$
$$\lambda = -\frac{1}{2}\langle \nabla_b f(x), 1\rangle.$$

Substituting in (13) we get,

$$d_b = -\frac{1}{L}\left(\nabla_b f(x) - \frac{1}{2}\langle \nabla_b f(x), 1\rangle 1\right). \tag{14}$$

That is,

$$\begin{bmatrix} d_1 \\ d_2 \end{bmatrix} = \frac{1}{2L}\left(\nabla_1 f(x) - \nabla_2 f(x)\right)\begin{bmatrix} -1 \\ 1 \end{bmatrix}.$$

Now, we plug in the optimal d_b from (14) in (12) and solve for b.

Solving for b. After substituting the optimal d_b, (12) becomes,

$$\begin{aligned}
&\operatorname*{argmin}_b -\frac{1}{L}\left\langle \nabla_b f(x), (\nabla_b f(x) - \frac{1}{2}\langle \nabla_b f(x), 1\rangle 1)\right\rangle + \frac{1}{2L}||(\nabla_b f(x) - \frac{1}{2}\langle \nabla_b f(x), 1\rangle 1)||^2 \\
&\equiv \operatorname*{argmin}_b -||\nabla_b f(x)||^2 + \frac{\langle \nabla_b f(x), 1\rangle}{2}\langle \nabla_b f(x), 1\rangle + \frac{1}{2}||\nabla_b f(x)||^2 - \frac{1}{2}\langle \nabla_b f(x), 1\rangle\langle \nabla_b f(x), 1\rangle + \\
&\qquad \frac{1}{8}(\langle \nabla_b f(x), 1\rangle)^2\langle 1, 1\rangle \\
&\equiv \operatorname*{argmin}_b -\frac{1}{2}||\nabla_b f(x)||^2 + \frac{1}{4}(\langle \nabla_b f(x), 1\rangle)^2 \\
&\equiv \operatorname*{argmin}_b -\nabla_1 f(x)^2 - \nabla_2 f(x)^2 + \frac{1}{2}(\nabla_1 f(x) + \nabla_2 f(x))^2 \\
&\equiv \operatorname*{argmax}_b \frac{1}{2}||\nabla_b f(x)||^2 - \nabla_1 f(x)\nabla_2 f(x) \\
&\equiv \operatorname*{argmax}_b \frac{1}{2}(\nabla_1 f(x) - \nabla_2 f(x))^2) \\
&\equiv \operatorname*{argmax}_b |\nabla_1 f(x) - \nabla_2 f(x)|. \qquad (15)
\end{aligned}$$

Therefore, *q-2-GCD* chooses two coordinates $b = \{i, j\}$ that are farthest apart. That is, coordinates with maximum and minimum gradient values of f.

$$i = \operatorname*{argmax}_i \nabla_i f(x),\ j = \operatorname*{argmin}_j \nabla_j f(x). \qquad (16)$$

A.3. Proof to Proposition 2

The block wise descent lemma for q-2-GCD can be written as,

$$f(x^{k+1}) \leq f(x^k) + \langle \nabla_b f(x^k), d_b\rangle + \frac{L}{2}||d_b||^2. \qquad (17)$$

Let $b = \{i, j\}$. We can now incorporate the linear equality constraints on d by substituting d_b from (13). Substituting d_b and further simplifying we get,

$$\begin{aligned}
f(x^{k+1}) &\leq f(x^k) - \frac{1}{4L}(\nabla_i f(x^k) - \nabla_j f(x^k))^2 \\
&= f(x^k) - \frac{1}{4L}\max_b(\nabla_i f(x^k) - \nabla_j f(x^k))^2.
\end{aligned}$$

A.4. Proof to Lemma 3

One possible solution for d would be where d is a vector of all zeros except 2 coordinates d_i and d_j. Let $i = \operatorname{argmax}_i \nabla_i f(x)$ and $j = \operatorname{argmin}_j \nabla_j f(x)$. Let us set $d_i = -\delta, d_j = \delta$ for some value $\delta > 0$ and $d_{k,k\neq i,j} = 0$, so that the linear equality constraint is satisfied by our construction. We then have $||d||_1 = 2\delta$, $sgn(d_i) = -1, sgn(d_j) = 1$ and $sgn(d_k) \in [-1, 1]$. We now want to see if (4) is

satisfied by every coordinate of d, that is, in order for d to be a solution, we would need

$$-\frac{1}{L}\nabla_i f(x) + \lambda = -2\delta \tag{18}$$
$$-\frac{1}{L}\nabla_j f(x) + \lambda = 2\delta \tag{19}$$
$$-\frac{1}{L}\nabla_k f(x) + \lambda \in 2\delta[-1, 1]. \tag{20}$$

Solving for λ in (18) and substituting in (19) we get,

$$\lambda = \frac{1}{L}\nabla_i f(x) - 2\delta \tag{21}$$
$$\delta = \frac{1}{4L}(\nabla_i f(x) - \nabla_j f(x)). \tag{22}$$

Therefore we get

$$\begin{bmatrix} d_i \\ d_j \end{bmatrix} = \frac{1}{4L}\left(\nabla_i f(x) - \nabla_j f(x)\right)\begin{bmatrix} -1 \\ 1 \end{bmatrix}. \tag{23}$$

Recall from (14) the optimal d obtained by q-2-GCD. Equation (23) shows that steepest descent indeed recovers the solution of q-2-GCD, but dividing by an extra factor of 2. Before we derive an exact relationship between these two methods, we need to verify if (20) is satisfied as well. Substituting (21) in (20) gives:

$$-\frac{1}{L}\nabla_k f(x) + \frac{1}{L}\nabla_i f(x) - 2\delta \in 2\delta[-1, 1].$$

Now, substituting (22) in (20) gives

$$-2\nabla_k f(x) + \nabla_i f(x) + \nabla_j f(x) \in (\nabla_i f(x) - \nabla_j f(x))[-1, 1],$$
$$-2\nabla_k f(x) + \nabla_i f(x) + \nabla_j f(x) \leq |\nabla_i f(x) - \nabla_j f(x)|.$$

As $\nabla_k f(x)$ is between $\nabla_i f(x)$ and $\nabla_j f(x)$, we can write it as $\theta\nabla_i f(x) + (1-\theta)\nabla_j f(x)$.

$$-2(\theta\nabla_i f(x) + (1-\theta)\nabla_j f(x)) + \nabla_i f(x) + \nabla_j f(x) \leq |\nabla_i f(x) - \nabla_j f(x)|$$
$$(1-2\theta)(\nabla_i f(x) - \nabla_j f(x)) \leq |\nabla_i f(x) - \nabla_j f(x)|,$$

which holds because $(1-2\theta) \in [-1, 1]$.

We now showed that steepest descent in 1-norm recovers the direction d that is equivalent to using q-2-GCD rule with zeros in all the coordinates of d except the two coordinates precisely chosen and updated by the GS-q rule.

$$d_{GCD} = 2d_{SD}.$$