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

Amrutha Varshini Ramesh AVRAMESH@CS.UBC.CA
University of British Columbia, Vancouver, Canada

Aaron Mishkin AMISHKIN@CS.STANFORD.EDU
Stanford Universitys

Mark Schmidt SCHMIDTM@CS.UBC.CA
University of British Columbia, Canada, CIFAR AI Chair (Amii)

[1]

Abstract

We consider minimizing a smooth function subject to an equality constraint. We analyze a greedy 2-coordinate update algorithm, and prove that greedy coordinate selection leads to faster convergence than random selection (under the Polyak-Łojasiewicz assumption). Our simple analysis exploits am equivalence between the greedy 2-coordinate update and equality-constrained steepest descent in the L1-norm. Unlike previous 2-coordinate analyses, our convergence rate is dimension independent.

1. Introduction

Coordinate descent (CD) is an iterative optimization algorithm where on each iteration we perform a gradient descent step on a single variable. CD methods are appealing because they have a convergence rate similar to gradient descent, but for some common objective functions the iterations have a much lower cost. Thus, there is substantial interest in using CD as an optimization algorithm for training machine learning models.

Coordinate descent with no constraints: Nesterov [8] considered CD with random choices of the coordinate to update, and proved explicit non-asymptotic linear convergence rates for strongly-convex functions with Lipschitz-continuous gradients. It was later shown that these linear convergence rates can be achieved under a generalization of strong convexity called the Polyak-Łojsiewicz condition [3]. Further, it was shown that greedy selection of the coordinate to update can lead to faster rates than randomized selection [9]. The faster greedy rates do not depend directly on the dimensionality of the problem, and rely on an equivalence between the greedy coordinate update and the steepest descent update on all coordinates in the L1-norm.

Coordinate descent with separable constraints: CD is commonly used for optimization with separable constraints, in the form of lower and/or upper bounds on each variable. Nesterov [8] showed that the unconstrained rates of randomized CD can be achieved under these separable constraints using a projected-gradient update of the coordinate. Richtárik and Takáč [11] generalize this result to include a non-smooth but separable term in the objective function, using a proximal-gradient update of the coordinate. These analyses justify using CD in various constrained and

1. This research was partially supported by the Canada CIFAR AI Chair Program, the Natural Sciences and Engineering Research Council of Canada (NSERC) Discovery Grants RGPIN-2022-03669

© A.V. Ramesh, A. Mishkin & M. Schmidt.

non-smooth settings, including ℓ_1-regularized least squares and SVMs where we regularize the bias term. Similar to the unconstrained case, Karimireddy et al. [4] show that several forms of greedy coordinate selection lead to faster convergence rates than random selection for problems with separable constraints and/or separable non-smooth terms.

Coordinate descent with an equality constraint: many problems in machine learning require us to satisfy an equality constraint. The most common example is that discrete probabilities must sum up to 1. Another common example is SVMs where we do not regularize the bias term. The (non-separable) equality constraint cannot be maintained if we only update one coordinate on each iteration, but it can be maintained if we update 2 variables on each iteration. Necoara et al. [5] analyze random selection of the two coordinates to update, while Fang et al. [2] discuss randomized selection with tighter rates. The popular LIBSVM package [1] uses a greedy 2-coordinate update for fitting SVMs without regularizing the bias. LIBSVM uses greedy coordinate selection since for the SVM problem greedy and random selection have similar iteration costs. But despite LIBSVM being perhaps the most widely-used CD method of all time, current analyses of greedy 2-coordinate updates [12] do not lead to faster rates than random selection.

Our contribution: we give a new analysis for a particular greedy 2-coordinate update for optimizing a smooth function an equality constraint. The analysis is based on an equivalence between the greedy update and equality-constrained steepest descent in the L1-norm. This leads to a dimension-independent analysis of greedy selection showing that it can converge substantially faster than random selection.

2. Notations and Definitions

In this work, we study the convergence of 2-GCD under GS-q block selection strategy ([12]) for problems with a single linear equality constraint. Specifically, we solve the following minimization problem:

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

We now introduce our notations and definitions. We call the indices of a parameter 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 linear operator $U_b : \mathbb{R}^{|b|} \to \mathbb{R}^n$ 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 gradient. That is, $\|\nabla f(x+\alpha) - \nabla f(x)\|_q \leq L_p \|\alpha\|_p$. This is a standard assumption for coordinate descent methods ([6–8, 12]). $||.||_p$ and $||.||_q$ are a pair of dual norms. 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 $i = \text{argmax}_i \nabla_i f(x), j = \text{argmin}_j \nabla_j f(x)$ and descends in the direction $d_b = -\frac{1}{L_2}\left(\nabla_b f(x) - \frac{1}{2}\langle \nabla_b f(x), 1\rangle 1\right)$.*
Proof sketch: *For the optimization problem (1), the GS-q rule selects the optimal block b, by solving the following minimization problem:*

$$b = \text{argmin}_b \min_{d_b | d_{b_1} + d_{b_2} = 0} \langle \nabla_b f(x), d_b \rangle + \frac{L_2}{2}||d_b||^2, \tag{2}$$

where d_b is the descent direction. To prove the Proposition 1, we first fix b and solve for d_b and then d_b is fixed to be the optimal d_b for each block, at which point it is easy to solve for the optimal b. 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 $d_{b^k} = -\frac{1}{L_2}\left(\nabla_b f(x^k) - \frac{1}{2}\langle \nabla_b f(x^k), 1\rangle 1\right)$.

3.1. Convergence Analysis

In this section, we prove that q-2-GCD converges linearly for functions satisfying the proximal-PL inequality [3]. 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 [10] 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 $f(x^{k+1}) = f(x^k) - \frac{1}{4L_2}\max_b(\nabla_i f(x^k) - \nabla_j f(x^k))^2$.*
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 [5, 7].

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 (14), with zeros for all elements in d except the two updated by the q-2-GCD and those coordinates follow (14).

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,

$$d \in \underset{d \in \mathbb{R}^n | d^T 1 = 0}{\text{argmin}} \nabla f(x)^T d + \frac{L_1}{2} ||d||_1^2. \tag{3}$$

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

$$\nabla_d \mathcal{L}(d, \lambda) = \nabla f(x) + \frac{L_1}{2} \partial ||d||_1^2 + \lambda 1 \text{ and } \nabla_\lambda \mathcal{L}(d, \lambda) = d^T 1.$$

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

$$\frac{2}{L_1}(-\nabla f(x) - \lambda 1) \in \partial ||d||_1^2 \quad \Rightarrow \quad \frac{1}{L_1}(-\nabla f(x) - \lambda 1) \in ||d||_1 sgn(d).$$

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,*

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

where, $\tau = \frac{L_1}{L_2}$ for a particular solution of the steepest descent problem.
Proof sketch: *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 [3] in the 1-norm. We have defined Prox-PL inequality in 1-norm and have shown its relationship with Prox-PL inequality in 2-norm is in appendix A.1.1.

Proposition 4 *The Lipschitz constants L_1 and L_2 of the block gradients with block size 2 of any function f measured in ℓ_1 norm and ℓ_2 norm is related as $L_2 = 2L_1$.*
The proof to this proposition follows from the result of Lemma 3.

3.1.3. CONVERGENCE RATE OF Q-2-GCD

Theorem 5 *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 (11) for some $\mu_1 > 0$. Then, the Algorithm q-2-GCD for the minimization problem in (1) obtains the following linear convergence rate:*

$$F(x^{k+1}) - F(x^*) \leq \left(1 - \frac{\mu_1}{L_2}\right)^k (F(x^0) - F^*). \tag{4}$$

Proof From the descent lemma of q-2-GCD,

$$F(x^{k+1}) \leq F(x^k) + \nabla f(x^k)^T (2\tau d_{SD}) + \frac{L_2}{2}||2\tau d_{SD}||^2 + g(x^{k+1}) - g(x^k).$$

Here, $f(x)$ is reformulated as $F(x) = f(x) + g(x)$ where, $g(x) = \begin{cases} 0 & \text{if } d^T 1 = 0 \\ \infty & \text{otherwise.} \end{cases}$.

From the relationship between ℓ_1 and ℓ_2 norms,

$$F(x^{k+1}) \leq F(x^k) + \nabla f(x^k)^T (2\tau d_{SD}) + \frac{L_2}{2}||2\tau d_{SD}||_1^2 + g(x^{k+1}) - g(x^k).$$

Change of notation: WLOG d_{SD} is written simply as d from this point. From the definition of GS-q rule and (4) we get,

$$\begin{aligned} F(x^{k+1}) &\leq F(x^k) + \min_d 2\nabla f(x^k)^T \tau d + 4\tau^2 L_1 ||d||_1^2 + g(x^{k+1}) - g(x^k) \\ &= F(x^k) + 2\tau \min_d \nabla f(x^k)^T d + 2L_1 \tau ||d||_1^2 + g(x^{k+1}) - g(x^k). \end{aligned} \tag{5}$$

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

$$F(x^{k+1}) \leq F(x^k) - \frac{1}{4L_1}\mathcal{D}_g(x, 2\tau L_1).$$

Therefore, from (Prox-PL Inequality in 1-norm) and (4) we get,

$$F(x^{k+1}) - F(x^k) \leq -\frac{\mu_1}{L_2}(F(x^k) - F^*) \quad \Rightarrow \quad F(x^{k+1}) - F(x^*) \leq \left(1 - \frac{\mu_1}{L_2}\right)(F(x^k) - F^*).$$

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

This shows the method's linear convergence to optimal F^*. Unlike the rates for random block selection [5, 7], this rate is independent of the problem dimension and significantly faster.

4. Conclusion

In this work we analyzed a greedy block-coordinate descent algorithm for a linearly-constrained objective which we call q-2-GCD. This has been a very successful method with various applications like SVM (LIBSVM). 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] 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.

[2] Qin Fang, Min Xu, and Yiming Ying. Faster convergence of a randomized coordinate descent method for linearly constrained optimization problems. *Analysis and Applications*, 16(05): 741–755, 2018.

[3] 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.

[4] 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.

[5] 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.

[6] 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.

[7] 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.

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

[9] 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.

[10] 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.

[11] Peter Richtárik and Martin Takáč. Iteration complexity of randomized block-coordinate descent methods for minimizing a composite function. *Mathematical Programming*, 144(1):1–38, 2014.

[12] 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 6 *(GS-q rule with one linear equality constraint) The GS-q rule chooses a block that maximizes progress while satisfying a single linear equality constraint, assuming a quadratic upper bound on the function f, where f is a smooth function with L-Lipschitz continuous gradient. That is,*

$$\text{GS-}q(x) = \arg\max_b |\mathcal{X}_b(x)|, \tag{6}$$

where,

$$\mathcal{X}_b(x) = \min_{d_b | \sum_i d_{b_i}=0} \left[\langle \nabla_b f(x), d_b \rangle + \frac{L}{2}\|d_b\|^2 \right].$$

Definition 7 *(Proximal Polyak- Lojasiewicz inequality [3])*
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 $\frac{\mu_2}{n} \leq \mu_1 \leq \mu_2$. 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,

$$\mathcal{D}_g(x, \alpha) = -2\alpha \min_y [\langle \nabla f(x), y - x\rangle + \frac{\alpha}{2}||y - x||_1^2 + g(y) - g(x)]. \tag{7}$$

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

$$x^{k+1} - x^k = d, \tag{8}$$

where $d = U_b(d_b)$. Applying (8) in (7), we get

$$\mathcal{D}_g(x, \alpha) = -2\alpha \min_d [\langle \nabla f(x), d\rangle + \frac{\alpha}{2}\|d\|_1^2 + g(y) - g(x)]. \tag{9}$$

We know that d has non-zero values in only 2 coordinates i and j. Using this (9) becomes,

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

$$= -2\alpha \min_d [\langle \nabla f(x), d\rangle + \frac{\alpha}{2}(d_i^2 + d_j^2 + 2|d_i||d_j|) + g(y) - g(x)]. \tag{11}$$

Following similar proof sketch, the Prox-PL inequality in 2-norm, introduced by [3], can be derived as,

$$\mathcal{D}_g(x,\alpha)_{[\text{2-norm}]} = -2\alpha \min_d [\langle \nabla f(x), d\rangle + \frac{\alpha}{2}(d_i^2 + d_j^2) + g(y) - g(x)]. \tag{12}$$

It is easy to see that,

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

Therefore,

$$\begin{aligned} -\mu_1(F(x) - F^*) &\geq -\mu_2(F(x) - F^*), \\ \mu_1(F(x) - F^*) &\leq \mu_2(F(x) - F^*). \end{aligned}$$

Therefore μ_1 is upper-bounded by μ_2. 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. Now, we show that $\mu_1 \geq \frac{\mu_2}{n}$. From the relationship between ℓ_1 and ℓ_2 norms, we know that $\|d\|_1 \leq \sqrt{n}\|d\|_2$. Now, (11) becomes

$$\begin{aligned} \mathcal{D}_g(x,\alpha) &\geq -2\alpha \min_d [\langle \nabla f(x), d\rangle + \frac{\alpha n}{2}\|d\|_2^2 + g(y) - g(x)] \\ &= \frac{\mathcal{D}_g(x,\alpha)_{[\text{2-norm}]}}{n}. \end{aligned}$$

Therefore,

$$\mu_1(F(x) - F^*) \geq \frac{\mu_2(F(x) - F^*)}{n}.$$

This implies, $\mu_1 \geq \frac{\mu_2}{n}$.

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:

$$b = \operatorname*{argmin}_b \min_{d_b | d_{b_1} + d_{b_2} = 0} \langle \nabla_b f(x), d_b\rangle + \frac{L_2}{2}||d_b||^2, \tag{13}$$

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

Solving for d_b. The Lagrangian of (13) is,

$$\mathcal{L}(d_b, \lambda) = \langle \nabla_b f(x), d_b\rangle + \frac{L_2}{2}||d_b||^2 + \lambda(d_{b_1} + d_{b_2}).$$

Taking the gradient with respect to d_b gives,

$$\nabla_{d_b}\mathcal{L}(d_b, \lambda) = \nabla_b f(x) + L_2 d_b + \lambda 1.$$

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

$$d_b = -\frac{1}{L_2}(\nabla_b f(x) + \lambda 1). \tag{14}$$

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

$$0 = -\frac{1}{L_2}\left(\nabla_1 f(x) + \lambda + \nabla_2 f(x) + \lambda\right),$$
$$\lambda = -\frac{1}{2}\langle \nabla_b f(x), 1\rangle.$$

Substituting in (14) we get,

$$d_b = -\frac{1}{L_2}\left(\nabla_b f(x) - \frac{1}{2}\langle \nabla_b f(x), 1\rangle 1\right). \tag{15}$$

That is,

$$\begin{bmatrix} d_1 \\ d_2 \end{bmatrix} = \frac{1}{2L_2}\left(\nabla_1 f(x) - \nabla_2 f(x)\right)\begin{bmatrix} -1 \\ 1 \end{bmatrix}.$$

Now, we plug in the optimal d_b from (15) in (13) and solve for b.

Solving for b. After substituting the optimal d_b, (13) becomes,

$$\begin{aligned}
&\operatorname*{argmin}_b -\frac{1}{L_2}\left\langle \nabla_b f(x), (\nabla_b f(x) - \frac{1}{2}\langle \nabla_b f(x), 1\rangle 1)\right\rangle + \frac{1}{2L_2}||(\nabla_b f(x) - \frac{1}{2}\langle \nabla_b f(x), 1\rangle 1)||^2 \\
&\equiv \operatorname*{argmin}_b -||\nabla_b f(x)||^2 + \frac{\langle \nabla_b f(x), 1\rangle}{2}\langle \nabla_b f(x), 1\rangle + \frac{1}{2}||\nabla_b f(x)||^2 - \frac{1}{2}\langle \nabla_b f(x), 1\rangle\langle \nabla_b f(x), 1\rangle + \\
&\qquad \frac{1}{8}(\langle \nabla_b f(x), 1\rangle)^2\langle 1, 1\rangle \\
&\equiv \operatorname*{argmin}_b -\frac{1}{2}||\nabla_b f(x)||^2 + \frac{1}{4}(\langle \nabla_b f(x), 1\rangle)^2 \\
&\equiv \operatorname*{argmin}_b -\nabla_1 f(x)^2 - \nabla_2 f(x)^2 + \frac{1}{2}(\nabla_1 f(x) + \nabla_2 f(x))^2 \\
&\equiv \operatorname*{argmax}_b \frac{1}{2}||\nabla_b f(x)||^2 - \nabla_1 f(x)\nabla_2 f(x) \\
&\equiv \operatorname*{argmax}_b \frac{1}{2}(\nabla_1 f(x) - \nabla_2 f(x))^2) \\
&\equiv \operatorname*{argmax}_b |\nabla_1 f(x) - \nabla_2 f(x)|.
\end{aligned} \tag{16}$$

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.

$$i = \operatorname*{argmax}_i \nabla_i f(x),\ j = \operatorname*{argmin}_j \nabla_j f(x). \tag{17}$$

A.3. Proof to Proposition 2

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

$$f(x^{k+1}) \leq f(x^k) + \langle \nabla_b f(x^k), d_b\rangle + \frac{L_2}{2}||d_b||^2. \tag{18}$$

Let $b = \{i, j\}$. We can now incorporate the linear equality constraints on d by substituting d_b from (14). Substituting d_b and further simplifying we get,

$$\begin{aligned} f(x^{k+1}) &\leq f(x^k) - \frac{1}{4L_2}(\nabla_i f(x^k) - \nabla_j f(x^k))^2 \\ &= f(x^k) - \frac{1}{4L_2}\max_b(\nabla_i f(x^k) - \nabla_j f(x^k))^2. \end{aligned}$$

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 $i = \text{argmax}_i \nabla_i f(x)$ and $j = \text{argmin}_j \nabla_j f(x)$. 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

$$-\frac{1}{L_1}\nabla_i f(x) + \lambda = -2\delta \tag{19}$$

$$-\frac{1}{L_1}\nabla_j f(x) + \lambda = 2\delta \tag{20}$$

$$-\frac{1}{L_1}\nabla_k f(x) + \lambda \in 2\delta[-1, 1]. \tag{21}$$

Solving for λ in (19) and substituting in (20) we get,

$$\lambda = \frac{1}{L_1}\nabla_i f(x) - 2\delta \tag{22}$$

$$\delta = \frac{1}{4L_1}(\nabla_i f(x) - \nabla_j f(x)). \tag{23}$$

Therefore we get

$$\begin{bmatrix} d_i \\ d_j \end{bmatrix} = \frac{1}{4L_1}\left(\nabla_i f(x) - \nabla_j f(x)\right)\begin{bmatrix} -1 \\ 1 \end{bmatrix}. \tag{24}$$

Recall from (15) the optimal d obtained by q-2-GCD. Before we derive an exact relationship between these two methods, we need to verify if (21) is satisfied as well. Substituting (22) in (21) gives:

$$-\frac{1}{L_1}\nabla_k f(x) + \frac{1}{L_1}\nabla_i f(x) - 2\delta \in 2\delta[-1, 1].$$

Now, substituting (23) in (21) gives

$$\begin{aligned} &-2\nabla_k f(x) + \nabla_i f(x) + \nabla_j f(x) \in (\nabla_i f(x) - \nabla_j f(x))[-1, 1], \\ &-2\nabla_k f(x) + \nabla_i f(x) + \nabla_j f(x) \leq |\nabla_i f(x) - \nabla_j f(x)|. \end{aligned}$$

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

$$\begin{aligned} -2(\theta\nabla_i f(x) + (1-\theta)\nabla_j f(x)) + \nabla_i f(x) + \nabla_j f(x) &\leq |\nabla_i f(x) - \nabla_j f(x)| \\ (1-2\theta)(\nabla_i f(x) - \nabla_j f(x)) &\leq |\nabla_i f(x) - \nabla_j f(x)|, \end{aligned}$$

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. Therefore,

$$U_b d_{b_{GCD}} = 2\tau d_{SD},$$

Where, $\tau = \frac{L_1}{L_2}$.