OPT2022: 14th Annual Workshop on Optimization for Machine Learning

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 University

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

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 a Polyak-Łojasiewicz assumption). Our simple analysis exploits an equivalence between the greedy 2-coordinate update and equality-constrained steepest descent in the 1-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 [9] 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 [10]. The faster greedy rates do not depend directly on the dimensionality of the problem, and are a consequence of an equivalence between the greedy coordinate update and the steepest descent update on all coordinates in the 1-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 [9] 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. AM is supported by NSF GRF Grant No. DGE-1656518 and NSERC PGS D, Grant No. PGSD3-547242-2020.

non-smooth settings, including least squares regularized by the 1-norm and support vector machines 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 with an unregularized 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. [6] analyze random selection of the two coordinates to update, while Fang et al. [2] discuss randomized selection with tighter rates. The 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 [13] 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 1-norm. This leads to a dimension-independent analysis of greedy selection showing that it can converge substantially faster than random selection.

2. Optimization with an Equality, Greedy 2-Coordinate Updates, and Proof Outline

We consider the problem of minimizing a twice-differentiable function f subject to a simple linear equality constraint,

$$\min_{x \in \mathbb{R}^n} f(x), \quad \text{subject to } \sum_{i=1}^{n} x_i = \gamma, \tag{1}$$

where n is the number of variables and γ is a constant. On iteration k the 2-coordinate descent method chooses a coordinate i_k and a coordinate j_k and updates these two coordinates using

$$x_{i_k}^{k+1} = x_{i_k}^k + \delta^k, \quad x_{j_k}^{k+1} = x_{j_k}^k - \delta^k,$$

for a scalar δ^k (the other coordinates are unchanged). We write this update for all coordinates as

$$x^{k+1} = x^k + d^k, \tag{2}$$

where $d_i^k = \delta^k$, $d_j^k = -\delta^k$, and $d_m^k = 0$ for $m \neq i$ and $m \neq j$. If the iterate x^k satisfies the constraint, then this update maintains the constraint.

To choose the coordinate to update, the greedy rule chooses the coordinates to maximize the difference in their partial derivatives,

$$i_k \in \arg\max_i \nabla_i f(x^k), \quad j_j \in \arg\min_j \nabla_j f(x^k). \tag{3}$$

At the solution of the problem we must have partial derivatives being equal, and intuitively this greedy choice updates the coordinates that are furthest above/below the average partial derivative. We

can also derive this choice as the minimizer among a set of 2-coordinate quadratic approximations to the function

$$\arg\min_{i,j} \left\{ \min_{d_{ij} | d_i + d_j = 0} f(x^k) + \nabla_{ij} f(x^k)^T d_{ij} + \frac{1}{2\alpha} \|d_{ij}\|^2 \right\}, \tag{4}$$

for a step size α — see Appendix A.1. This is a special case of the Gauss-Southwell-q (GS-q) rule of Tseng and Yun [13].

We assume that the gradient of f is Lipschitz continuous, and our analysis will depend on a quantity we call L_2. The quantity L_2 bounds the change in the 2-norm of the gradient with respect to any two coordinates i and j under a two-coordinate update of any x of the form in (2).

$$\|\nabla_{ij} f(x + d) - \nabla_{ij} f(x)\| \leq L_2 \|d\|. \tag{5}$$

Note that L_2 is less than or equal to the Lipschitz constant of the gradient of f. Our analysis will focus on the case of $\delta^k = -\frac{1}{2L}(\nabla_{i_k} f(x^k) - \nabla_{j_k} f(x^k))$, resulting in an update to the coordinates of

$$\begin{aligned} x_{i_k}^{k+1} &= x_{i_k}^k - \frac{1}{2L_2}(\nabla_{i_k} f(x^k) - \nabla_{j_k} f(x^w)), \\ x_{j_k}^{k+1} &= x_{j_k}^k - \frac{1}{2L_2}(\nabla_{j_k} f(x^k) - \nabla_{i_k} f(x^w)). \end{aligned} \tag{6}$$

However, we note that our analysis also applies if we choose δ^k to maximally decrease f.

Our analysis relies on the following properties related to the 1-norm:

1. For vectors d^k of the form given above, we have $\|d^k\|_1^2 = 2\|d^k\|_2^2$,

   $$\|d^k\|_1^2 = (|\delta^k| + |-\delta^k|)^2 = (\delta^k)^2 + (\delta^k)^2 + 2|\delta^k| \cdot |\delta^k| = 4(\delta^k)^2 = 2((\delta^k)^2 + (-\delta^k)^2) = 2\|d^k\|_2^2.$$

2. If a twice-differentiable function's gradient satisfies the 2-coordinate Lipschitz continuity assumption (5) with constant L_2, then the full gradient is Lipschitz continuous in the 1-norm with constant $L_1 = L_2/2$ (see Appendix B).

3. Applying the 2-coordinate update (6) is an instance of applying steepest descent over all coordinates in the 1-norm.

4. For a function satisfying the proximal-PL inequality, we can measure the proximal-PL inequality in the 1-norm.

The next section outlines the latter components, and we use these to give a simple proof for the greedy 2-coordinate algorithm in Section 4.

3. Connections to the 1-Norm

In this section we outline the connection between the greedy update and steepest descent in the 1-norm, and then we discuss the proximal-PL condition.

3.1. Steepest Descent in the 1-Norm

Now we show that steepest descent in the 1-norm can also lead to sparse update directions. In particular, we show that steepest descent in the 1-norm always admits at least one solution which updates only two coordinates. Crucially, this implies that the best two-coordinate update makes as much progress in the 1-norm as any full-coordinate update.

Steepest descent in the 1-norm, subject to the equality constraint, takes steps in the direction d that minimizes the following model of the objective:

$$d \in \underset{d \in \mathbb{R}^n | d^T 1 = 0}{\arg\min} \nabla f(x)^T d + \frac{1}{2\alpha}||d||_1^2, \tag{7}$$

where α is the step size. This is a convex optimization problem for which strong duality holds. Introducing a dual variable $\lambda \in \mathbb{R}$, we obtain the Lagrangian

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

The subdifferential with respect to d and λ yields necessary and sufficient optimality conditions for a steepest descent direction,

$$\nabla_d \mathcal{L}(d, \lambda) = \nabla f(x) + \frac{1}{2\alpha} g + \lambda 1 = 0 \qquad \text{(for some subgradient } g \in \partial||d||_1^2\text{)}$$
$$\nabla_\lambda \mathcal{L}(d, \lambda) = d^T 1 = 0.$$

The second condition is simply feasibility of d, while from the first we obtain,

$$2\alpha(-\nabla f(x) - \lambda 1) \in \partial||d||_1^2 \quad \Longrightarrow \quad \alpha(-\nabla f(x) - \lambda 1) \in ||d||_1 \text{sgn}(d), \tag{8}$$

where element m of $\text{sgn}(d)$ is 1 if d_m is positive, -1 if d_m is negative, and can be any value in $[-1, 1]$ if d_m is 0. The following lemma shows that these conditions are always satisfied by a two-coordinate update.

Lemma 1 *Let $\alpha > 0$. Then at least one steepest descent direction with respect to the 1-norm has exactly two non-zero coordinates. That is,*

$$\min_{d \in \mathbb{R}^n | d^T 1 = 0} \nabla f(x)^T d + \frac{1}{2\alpha}||d||_1^2 = \min_{i,j} \left\{ \min_{d_{ij} \in \mathbb{R}^2 | d_i + d_j = 0} \nabla_{ij} f(x)^T d_{ij} + \frac{1}{2\alpha}||d_{ij}||_1^2 \right\}. \tag{9}$$

See Appendix A.2 for the proof. Lemma 1 allows us to relate the progress of a block-coordinate update on just two coordinates to the progress made by a full-coordinate steepest descent step. This will be a key step in our analysis of the GS-q method. However, first we introduce the proximal-PL inequality in the 1-norm, which is our main technique for lower bounding the sub-optimality of an iterate in terms of the squared 1-norm..

3.2. Proximal-PL Inequality in the 1-Norm

The proximal-PL condition was introduced to allow simpler proofs for various constrained and non-smooth optimization problems [3]. The proximal-PL condition is normally defined based on the 2-norm and non-smooth functions. Below, we define a variant where distances are measured in the 1-norm and we include the equality constraint explicitly.

Definition 2 *A function f, that is L_1-Lipschitz with respect to the 1-norm and has a summmation constraint on its parameters, satisfies the proximal-PL condition in the 1-norm if for a positive constants μ_1 we have*

$$\frac{1}{2}\mathcal{D}(x, L_1) \geq \mu_1(f(x) - f^*), \tag{10}$$

for all x satisfying the equality constraint, where f^* is the constrained optimum function value and where

$$\mathcal{D}(x, \alpha) = -2\alpha \min_{\{y \mid y^T 1 = 0\}} \left[\langle \nabla f(x), y - x\rangle + \frac{\alpha}{2}||y - x||_1^2\right]. \tag{11}$$

It follows from the equivalence between norms that summation-constrained functions satisfying the proximal-PL condition in the 2-norm will also satisfy the above proximal-PL condition in the 1-norm. In particular, if μ_2 is the proximal-PL constant in the 2-norm, then we have $\frac{\mu_2}{n} \leq \mu_1 \leq \mu_2$ (see Appendix C). Functions satisfying these conditions include any strongly-convex function f, as well as relaxations of this such as functions of the form $f = g(Ax)$ for a strongly-convex g and a matrix A (in this case f may not be strongly-convex) [3].

4. Convergence Result

We now combine the previously-stated results to give a convergence rate for the greedy 2-coordinate method.

Theorem 3 *Let f be a twice-differentiable function whose gradient is 2-coordinate-wise Lipschitz (4) and restricted to the set where $x^T 1 = \gamma$. If this function satisfies the proximal-PL inequality in the 1-norm (10) for some positive μ_1, then the 2-coordinate update (6) with the greedy GS-q rule (3) satisfies:*

$$f(x^k) - f(x^*) \leq \left(1 - \frac{2\mu_1}{L_2}\right)^k (f(x^0) - f^*). \tag{12}$$

Proof *Starting from the descent lemma applied to the function restricted to the coordinates i_k and j_k that we update, we have*

$$\begin{aligned}
f(x^{k+1}) &\leq f(x^k) + \nabla f(x^k)^T d + \frac{L_2}{2}\|d\|^2 \\
&= f(x^k) + \min_{i,j}\left\{\min_{d_{ij}\in\mathbb{R}^2 | d_i + d_j = 0} \nabla_{ij} f(x^k)^T d_{ij} + \frac{L_2}{2}\|d_{ij}\|^2\right\} && \textit{(GS-q rule)} \\
&= f(x^k) + \min_{i,j}\left\{\min_{d_{ij}\in\mathbb{R}^2 | d_i + d_j = 0} \nabla_{ij} f(x^k)^T d_{ij} + \frac{L_2}{4}\|d_{ij}\|_1^2\right\} && (\|d\|_1^2 = 2\|d\|^2) \\
&= f(x^k) + \min_{i,j}\left\{\min_{d_{ij}\in\mathbb{R}^2 | d_i + d_j = 0} \nabla_{ij} f(x^k)^T d_{ij} + \frac{L_1}{2}\|d_{ij}\|_1^2\right\} && (L_1 = L_2/2) \\
&= f(x^k) + \min_{d | d^T 1 = 0}\left\{\nabla f(x^k)^T d + \frac{L_1}{2}\|d\|_1^2\right\} && \textit{(Lemma 1)}
\end{aligned}$$

Now subtracting f^* from both sides and using the definition of $\mathcal{D}$ we get

$$\begin{aligned} f(x^{k+1}) - f(x^*) &\leq f(x^k) - f(x^*) - \frac{1}{2L_1}\mathcal{D}(x^k, L_1) \\ &= f(x^k) - f(x^*) - \frac{\mu_1}{L_1}(f(x^k) - f^*) \qquad \textit{(proximal PL)} \\ &= f(x^k) - f(x^*) - \frac{2\mu_1}{L_2}(f(x^k) - f^*) \\ &= \left(1 - \frac{2\mu_1}{L_2}\right)(f(x^k) - f^*) \end{aligned}$$

Applying the inequality recursively completes the proof. ■

5. Comparison to Randomized Selection

If we sample the two coordinates i_k and j_k from a uniform distribution, then it is known that the 2-coordinate descent method satisfies [12]

$$\mathbb{E}[f(x^k)] - f(x^*) \leq \left(1 - \frac{\mu_2}{n^2 L_2}\right)^k (f(x^0) - f^*). \tag{13}$$

A similar result for a more-general problem class was shown by Necoara and Patrascu [7]. This is substantially slower the rate we show for the greedy 2-coordinate descent method. This rate is slower even in the extreme case where μ_1 is similar to μ/n, due to the presence of the n^2 term.

Necoara and Patrascu [7] also show that faster rates than (13) under the additional assumption that we know coordinate-wise Lipschitz constant values (and give faster non-uniform sampling strategies). It is also possible to derive faster rates for problems were f is separable [2, 6, 8], but this restricts the applicability of the result. Finally, unlike the convergence rates shown for random coordinate selection, we note that the linear convergence rate shown in this work for the greedy 2-coordinate method avoids requiring a direct dependence on the problem dimension.

6. Extension to Bound Constraints

In this work we only considered a single equality constraint. But in machine learning where we have a single equality constraint, we typically also have bound constraints on the variables. This includes our motivating problems of optimizing over the probability simplex, or optimizing SVMs with an unregularized bias. The first versions of LIBSVM used a variation on the Gauss-Southwell-q for the case of bound constraints an a linear equality. Unfortunately, our proof technique does not directly apply if we add bound constraints. This is because with bound constraints and an equality constraint the steepest descent direction in the 1-norm may update more than two coordinates. We are exploring whether ideas like those of Karimireddy et al. [4] will allow us to use this simple analysis this more-general setting.

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] Sai Praneeth Reddy Karimireddy, Sebastian Stich, and Martin Jaggi. Adaptive balancing of gradient and update computation times using global geometry and approximate subproblems. In *International Conference on Artificial Intelligence and Statistics*, pages 1204–1213. PMLR, 2018.

[6] 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.

[7] 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.

[8] 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.

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

[10] 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.

[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] Jennifer She and Mark Schmidt. Linear convergence and support vector identification of sequential minimal optimization. In *10th NIPS Workshop on Optimization for Machine Learning*, volume 5, page 50, 2017.

[13] 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. Greedy GS-q Rule for Summation Constraints

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

$$b = \arg\min_{b} \min_{d_b | d_{b_1} + d_{b_2} = 0} \langle \nabla_b f(x), d_b \rangle + \frac{1}{2\alpha} ||d_b||^2, \tag{14}$$

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

Solving for d_b. The Lagrangian of (14) is,

$$\mathcal{L}(d_b, \lambda) = \langle \nabla_b f(x), d_b \rangle + \frac{1}{2\alpha} ||d_b||^2 + \lambda(d_{b_1} + d_{b_2}).$$

Taking the gradient with respect to d_b gives,

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

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

$$d_b = -\alpha(\nabla_b f(x) + \lambda 1). \tag{15}$$

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

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

Substituting in (15) we get,

$$d_b = -\alpha \left(\nabla_b f(x) - \frac{1}{2} \langle \nabla_b f(x), 1 \rangle 1 \right). \tag{16}$$

That is,

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

Now, we plug in the optimal d_b from (16) in (14) and solve for b.

Solving for b. After substituting the optimal d_b, (14) becomes,

$$\begin{aligned}
&\arg\min_b -\alpha \left\langle \nabla_b f(x), (\nabla_b f(x) - \frac{1}{2}\langle \nabla_b f(x), 1\rangle 1) \right\rangle + \frac{\alpha}{2}||(\nabla_b f(x) - \frac{1}{2}\langle \nabla_b f(x), 1\rangle 1)||^2 \\
&\equiv \arg\min_b -||\nabla_b f(x)||^2 + \frac{\langle \nabla_b f(x), 1\rangle}{2}\langle \nabla_b f(x), 1\rangle + \frac{1}{2}||\nabla_b f(x)||^2 - \frac{1}{2}\langle \nabla_b f(x), 1\rangle\langle \nabla_b f(x), 1\rangle + \\
&\qquad \frac{1}{8}(\langle \nabla_b f(x), 1\rangle)^2\langle 1, 1\rangle \\
&\equiv \arg\min_b -\frac{1}{2}||\nabla_b f(x)||^2 + \frac{1}{4}(\langle \nabla_b f(x), 1\rangle)^2 \\
&\equiv \arg\min_b -\nabla_1 f(x)^2 - \nabla_2 f(x)^2 + \frac{1}{2}(\nabla_1 f(x) + \nabla_2 f(x))^2 \\
&\equiv \arg\max_b \frac{1}{2}||\nabla_b f(x)||^2 - \nabla_1 f(x)\nabla_2 f(x) \\
&\equiv \arg\max_b \frac{1}{2}(\nabla_1 f(x) - \nabla_2 f(x))^2) \\
&\equiv \arg\max_b |\nabla_1 f(x) - \nabla_2 f(x)|. \qquad (17)
\end{aligned}$$

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.

$$i = \arg\max_i \nabla_i f(x),\ j = \arg\min_j \nabla_j f(x). \qquad (18)$$

A.2. Proof to Lemma 1

The proof follows by constructing a solution to the steepest descent problem in Eq. 7 which only has two non-zero entries. Let $i = \arg\max_i \nabla_i f(x)$ and $j = \arg\min_j \nabla_j f(x)$. Our proposal solution is d such that $d_i = -\delta, d_j = \delta$ for some $\delta \in \mathbb{R}$ and $d_{k,k\neq i,j} = 0$. Clearly d satisfies the sum-to-zero constraint required for feasibility.

Now we check that (8) is satisfied by every coordinate of d. The definition of the 1-norm implies $||d||_1 = 2\delta$, while $sgn(d_i) = -1, sgn(d_j) = 1$ and $sgn(d_k) \in [-1, 1]$ follow by construction. Thus, for d to be a steepest descent direction we must have the following:

$$-\alpha\nabla_i f(x) + \lambda = -2\delta \qquad (19)$$
$$-\alpha\nabla_j f(x) + \lambda = 2\delta \qquad (20)$$
$$-\alpha\nabla_k f(x) + \lambda \in 2\delta[-1, 1]. \qquad (21)$$

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

$$\lambda = \alpha\nabla_i f(x) - 2\delta \qquad (22)$$
$$\delta = \frac{\alpha}{4}(\nabla_i f(x) - \nabla_j f(x)). \qquad (23)$$

Therefore we get

$$\begin{bmatrix} d_i \\ d_j \end{bmatrix} = \frac{\alpha}{4}(\nabla_i f(x) - \nabla_j f(x)) \begin{bmatrix} -1 \\ 1 \end{bmatrix}. \qquad (24)$$

It remains only to show that (21) is satisfied by d. Using the value of λ in the constraint yields,

$$-\alpha\nabla_k f(x) + \alpha\nabla_i f(x) - 2\delta \in 2\delta[-1,1].$$

Now, substituting the value for δ 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 have shown that a two-coordinate update d satisfies the sufficient conditions to be a steepest descent direction in the 1-norm. Substituting d back into the expression for steepest descent gives

$$\begin{aligned}\min_{d\in\mathbb{R}^n | d^T 1 = 0} \nabla f(x)^T d + \frac{1}{2\alpha}||d||_1^2 &= \nabla_{ij} f(x)^T d_{ij} + \frac{1}{2\alpha}||d_{ij}||_1^2\\ &\geq \min_{i,j}\left\{\min_{d_{i,j}\in\mathbb{R}^2 | d_i + d_j = 0} \nabla_{ij} f(x)^T d_{ij} + \frac{1}{2\alpha}||d_{ij}||_1^2\right\}.\end{aligned}$$

Since the reverse inequality follows trivially, we deduce that

$$\min_{d\in\mathbb{R}^n | d^T 1 = 0} \nabla f(x)^T d + \frac{1}{2\alpha}||d||_1^2 = \min_{i,j}\left\{\min_{d_{i,j}\in\mathbb{R}^2 | d_i + d_j = 0} \nabla_{ij} f(x)^T d_{ij} + \frac{1}{2\alpha}||d_{ij}||_1^2\right\},$$

as claimed.

Appendix B. Relating Lipschitz Constants

Proposition 4 *Suppose f is twice differentiable and*

$$\max_d \left\{d^\top \nabla^2 f(x) d : \langle d, 1\rangle = 0, \textit{supp}(d) = 2, \|d\|_1 \leq 1\right\} \leq L_1. \tag{25}$$

for any x such that $\langle x, 1\rangle = a$. Then f satisfies the following inequality:

$$f(x+d) \leq f(x) + \langle \nabla f(x), d\rangle + \frac{L_1}{2}\|d\|_1^2, \tag{26}$$

for any such x and any d such that $\langle d, 1\rangle = 0$.

Proof Consider the optimization problem

$$\max_d \left\{d^\top \nabla^2 f(x) d : \langle d, 1\rangle = 0, \|d\|_1 \leq 1\right\}. \tag{27}$$

This is a convex maximization problem over a polyhedral constraint set; standard results from convex optimization imply that at least one maximizer of (27) occurs at an extreme point of the constraint set

$$\mathcal{D} = \{d : \langle d, a\rangle = 0, \|d\|_1 \leq 1\}.$$

We now show that all extreme points of $\mathcal{D}$ contain exactly two non-zero entries.

Let d_e be any extreme point of $\mathcal{D}$ and suppose by way of contradiction that d_e has at least three non-zero entries. Denote these entries as d_1, d_2, d_3. Since at least one entry of d_e must negative and one must be positive, we may assume without loss of generality that $d_1, d_2 > 0$ and $d_3 < 0$.

Let $\epsilon > 0$ and define $d'_e = d_e + e_1\epsilon - e_2\epsilon$. For ϵ sufficiently small ϵ it holds that $d_1 + \epsilon > 0$ and $d_2 - \epsilon > 0$ so that

$$(d_1 + \epsilon) + (d_2 - \epsilon) + d_3 = d_1 + d_2 + d_3$$

and

$$\begin{aligned} |d_1 + \epsilon| + |d_2 - \epsilon| + |d_3| &= (d_1 + \epsilon) + (d_2 - \epsilon) + d_3 \\ &= d_1 + d_2 + d_3. \end{aligned}$$

Thus, $d'_e \in \mathcal{D}$. Repeating this argument for $d''_e = d_e - e_1\epsilon + e_2\epsilon$, we obtain $d''_e \in \mathcal{D}$ and

$$d_e = \frac{1}{2}d'_e + \frac{1}{2}d''_e,$$

which contradicts that d_e is an extreme point of $\mathcal{D}$. Thus, every extreme point of $\mathcal{D}$ has at exactly two non-zero entries.

Returning to our argument, we see that (27) is maximized at d_e, where $\text{supp}(d_e) = 2$. Thus,

$$\begin{aligned} \max_d \left\{ d^\top \nabla^2 f(x) d : \langle d, 1\rangle = 0, \|d\|_1 \leq 1 \right\} &= \max_d \left\{ d^\top \nabla^2 f(x) d : \langle d, 1\rangle = 0, \text{supp}(d) = 2, \|d\|_1 \leq 1 \right\} \\ &\leq L_1. \end{aligned}$$

The result is now easily obtained using a Taylor expansion and the intermediate value theorem. ■

Proposition 5 *The constant L_1 in (25) is exactly equal to $\frac{L_2}{2}$.*

Proof Let $d \in \mathbb{R}^n$ such that $\text{supp}(d) = 2$ and $\langle d, 1\rangle = 0$. WLOG, suppose that the two non-zero entries of d are d_1 and d_2. Observe that $\langle d, 1\rangle = 0$ implies $d_1 = -d_2$. Using these facts, we obtain

$$\begin{aligned} \|d\|_1^2 &= d_1^2 + d_2^2 + 2|d_1||d_2| \\ \|d\|_1^2 &= d_1^2 + d_2^2 + |d_1||d_1| + |d_2||d_2| \\ \|d\|_1^2 &= 2\|d\|_2^2. \end{aligned}$$

Thus, $\|d\|_1 = \sqrt{2}\|d\|_2$ and

$$\begin{aligned} &\max_d \left\{ d^\top \nabla^2 f(x) d : \langle d, 1\rangle = 0, \text{supp}(d) = 2, \|d\|_2 \leq 1 \right\} \\ &\qquad = 2 \max_d \left\{ d^\top \nabla^2 f(x) d : \langle d, 1\rangle = 0, \text{supp}(d) = 2, \|d\|_1 \leq 1 \right\} \\ &\qquad \leq 2L_1. \end{aligned}$$

which implies $L_2 \leq 2L_1$. Similarly, we have

$$\begin{aligned}
&\max_d \left\{ d^\top \nabla^2 f(x) d : \langle d, 1 \rangle = 0, \text{supp}(d) = 2, \|d\|_1 \leq 1 \right\} \\
&\qquad = \frac{1}{2} \max_d \left\{ d^\top \nabla^2 f(x) d : \langle d, 1 \rangle = 0, \text{supp}(d) = 2, \|d\|_2 \leq 1 \right\} \\
&\qquad \leq \frac{L_2}{2}.
\end{aligned}$$

This completes the proof. ■

Appendix C. Relationship Between Proximal-PL Constants

Lemma 6 *Suppose that $F(x) = f(x) + g(x)$ satisfies the proximal-PL inequality in the ℓ_2-norm with constants L_2, μ_2. Then F also satisfies the proximal-PL inequality in the ℓ_1-norm with constants L_1 and $\mu_1 \in [\mu_2/n, \mu_2]$.*

Proof Proximal-PL inequality in the ℓ_2-norm implies

$$\begin{aligned}
F(x) - F(x^*) &\leq -\frac{L_2}{\mu_2} \min_y \left\{ \langle \nabla f(x), y - x \rangle + \frac{L_2}{2} \|y - x\|_2^2 + g(y) - g(x) \right\} \\
&\leq -\frac{L_2}{\mu_2} \min_y \left\{ \langle \nabla f(x), y - x \rangle + \frac{L_2}{2n} \|y - x\|_1^2 + g(y) - g(x) \right\} \\
&\leq -\frac{L_2 L_1 n}{L_2 \mu_2} \min_y \left\{ \langle \nabla f(x), y - x \rangle + \frac{L_1}{2} \|y - x\|_1^2 + g(y) - g(x) \right\} \\
&= -\frac{L_1 n}{\mu_2} \min_y \left\{ \langle \nabla f(x), y - x \rangle + \frac{L_1}{2} \|y - x\|_1^2 + g(y) - g(x) \right\},
\end{aligned}$$

where the last inequality follows from Karimireddy et al. [5][Lemma 9] with the choice of $\beta = \frac{L_2}{L_1 n}$, $h(y) = \langle \nabla f(x), y - x \rangle + g(y) - g(x)$, and $V(y) = \sqrt{L_2/2n} \|y - x\|_1$. Note that $\beta \in (0, 1]$ since $L_2 n \geq L_1$ and $h(x) = V(x) = 0$ so that the conditions of the lemma are satisfied. We conclude that proximal-PL inequality holds with $\mu_1 \geq \mu_2/n$.

We establish the reverse direction similarly; starting from proximal-PL in the ℓ_1-norm,

$$\begin{aligned}
F(x) - F(x^*) &\leq -\frac{L_1}{\mu_1} \min_y \left\{ \langle \nabla f(x), y - x \rangle + \frac{L_1}{2} \|y - x\|_1^2 + g(y) - g(x) \right\} \\
&\leq -\frac{L_1}{\mu_1} \min_y \left\{ \langle \nabla f(x), y - x \rangle + \frac{L_1}{2} \|y - x\|_2^2 + g(y) - g(x) \right\} \\
&\leq -\frac{L_1 L_2}{L_1 \mu_1} \min_y \left\{ \langle \nabla f(x), y - x \rangle + \frac{L_2}{2} \|y - x\|_2^2 + g(y) - g(x) \right\} \\
&= -\frac{L_2}{\mu_1} \min_y \left\{ \langle \nabla f(x), y - x \rangle + \frac{L_2}{2} \|y - x\|_2^2 + g(y) - g(x) \right\},
\end{aligned}$$

where now we have used the same lemma with $V(y) = \sqrt{L_1/2} \|y - x\|_2$ and $\beta = \frac{L_1}{L_2}$, noting that $\beta \in (0, 1]$ since $L_1 \leq L_2$. This shows that $\mu_2 \geq \mu_1$, which completes the proof. ■