Coordinate Descent with Coupled Constraints

Aman Sinha

June 5, 2015

1 Introduction

For many big data applications, a relatively small parameter vector \( \theta \in \mathbb{R}^n \) is determined to fit a model to a very large dataset with \( N \) observations. We consider a different motivating problem in which both \( n \) and \( N \) are very large. In these situations, batch optimization techniques (that perform optimization utilizing all datapoints together) are not allowable; either these techniques are too costly for our computational memory and time budgets, or the model must be fit online as data is observed. More importantly, however, many stochastic techniques are also intractable. Stochastic descent methods iteratively perturb model parameters using only a single observation at each step. This assuages the dependence on \( N \), but most stochastic descent methods require working with the entire \( \theta \) when the corresponding elements are coupled via constraints. Since \( n \) is also very large, this is not allowable.

Coordinate descent methods work only with a single coordinate of \( \theta \) at each iteration. This is analogous to the idea of subsampling observations in a stochastic descent method. However, standard coordinate descent methods typically assume that the constraint sets for different coordinates are independent [Nes12, RT13, QRZ14]. This greatly limits the applicability of coordinate descent methods to practical problems, as many real-world situations have coupled constraints (such as \( \theta \) lying in the probability simplex or the \( l_p \) ball).

This project proposes to perform coordinate descent in the presence of coupled constraints. Specifically, we propose a simple algorithm that works with a group of \( k \) coordinates at each iteration; the minimum sufficient value of \( k \) depends on the constraint set. We prove convergence of the algorithm for polyhedral constraint sets and show experimental results over the probability simplex.

2 Problem Statement and Proposed Approach

We consider a generalized coordinate-descent framework that formally characterizes conditions under which coordinate descent can converge. Let \( X \) be a compact convex space with nonempty relative interior. Consider the following problem: minimize a differentiable convex function \( f(y) \) with \( L \)-Lipschitz gradient over \( X \),

\[
\min_{y \in X} f(y),
\]

with iterations that only perturb up to \( k \) coordinates of \( y \) at a time.
2.1 Approach

We denote the $k$ coordinates that we can change by a binary vector $s \in \mathbb{R}^n$ that is nonzero in at most $k$ coordinates, and we denote the set of all possible such vectors as $S$. We consider a method that iteratively minimizes a convex upper bound of the objective $f(y)$ using $k$ coordinates. Specifically, we have the following iterative update:

$$y^{i+1} = \arg\min_{z \in X_s(y^i)} \nabla f(y^i)^T(z - y^i) + \frac{L}{2} \|z - y^i\|^2,$$

where $X_s(y) := \{z : z \in X, \text{ diag}(1-s)(z-y) = 0\}$. This iterative update minimizes upper bounds on $f(y)$ since $f(y)$ has $L$-Lipschitz gradient; it is equivalent to a gradient descent update with a step size of $1/L$ projected onto $X_s(y^n)$. Because $y^n \in X_s(y^i)$, we know that $\nabla f(y^i)^T(y^{i+1} - y^i) + \frac{L}{2} \|y^{i+1} - y^i\|^2 \leq 0$ and $f(y^{i+1}) \leq f(y^i)$.

Note that we have not yet specified how the vector $s$ is chosen for each iteration. Whether we choose it deterministically or stochastically, what is most important is that, no matter how many iterations $n$ we have completed in the process, any of the $\binom{n}{k}$ possible vectors can occur once more using our deterministic or stochastic sampling approach. This is required, along with the right value of $k$, to guarantee convergence to the optimal value. To keep our approach general, we first prove convergence with these conditions presented as assumptions (and we will prove how to satisfy these assumptions for specific compact convex sets in Section 3).

2.2 Proof of Convergence

**Assumption 2.1** We have chosen $k$ so that for any vector $v \in \mathbb{R}^n$ and any $y^0 \in X$, there exists a path $y^0 \to y^1 \ldots \to y^m \to y^* = \arg\min_{y \in X} v^T y$, $v^T(y^i - y^{i+1}) < 0 \ \forall \ i$, and $y^{i+1} \in X_s(y^i) \ \forall \ i$.

**Assumption 2.2** The path given in Assumption 2.1 is achievable using a sampling approach that iteratively selects $s^i$ (although we allow multiple draws of $s^i$ in between moves from $y^i$ to $y^{i+1}$).

These assumptions guarantee that we can move towards the optimum if we are not already there, which is always true for $k = n$ (with $M = 0$) but is not necessarily true for smaller $k$. Note that Assumption 2.2 is easy to satisfy by either a deterministic sampling approach that cycles through all $\binom{n}{k}$ possibilities, or a stochastic approach that prescribes a nonzero probability to each of the $\binom{n}{k}$ possibilities. To be clear, these assumptions guarantee that we can make progress using some group of $k$ coordinates, but not necessarily any group of $k$ coordinates. Using this, we can give a condition for optimality.

**Proposition 2.3** We have $y^i \in \arg\min_{y \in X} f(y)$ if the following holds:

$$\min_{s^i \in S} \inf_{z \in X_s(y^i)} \nabla f(y^i)^T(z - y^i) + \frac{L}{2} \|z - y^i\|^2 \geq 0. \quad (3)$$

**Proof** Take $y^i \notin \arg\min_{y \in X} f(y)$. Then, using Assumptions 2.1 and 2.2, we can take a small enough step such that $\nabla f(y^i)^T(z - y^i) + \frac{L}{2} \|z - y^i\|^2 < 0$. The small step size is required to ensure the linear term $\nabla f(y^i)^T(z - y^i)$ dominates. \qed
Now we look at the following subset $X_\epsilon := \{ y : f(y) < \min_{y \in X} f(y) + \epsilon \}$. Intuitively, we can show convergence of (2) if we can show that the iterates $y^i$ must drop into $X_\epsilon$ for any value of $\epsilon$. We define $h_\epsilon(y) := \inf_{z \in X_\epsilon(y)} \nabla f(y)^T (z - y) + \frac{L}{2} \| z - y \|^2$, and $h(y) := \min_{s \in S} h_\epsilon(y)$. We need to show that $\sup_{y \notin X_\epsilon} h(y) < 0$.

**Assumption 2.4** For all $s \in S$, $h_\epsilon(y)$ is continuous over the restricted domain $X$.

Using this assumption, we can prove convergence.

**Proposition 2.5** The iterations (2) converge to any $\epsilon$-suboptimal neighborhood of $y^* \in \argmin_{y \in V} f(y)$ in finite time.

**Proof** Since all $h_\epsilon(y)$ are continuous, so is $h(y)$. Then $h(y)$ achieves its supremum over the compact space $X \setminus X_\epsilon$. We know that for every $y \notin X_\epsilon$, $h(y) < 0$ by Proposition 2.3, so $\sup_{y \notin X_\epsilon} h(y) < 0$. Since the iterations $y^i$ minimize an upper bound of $f(y^*)$, we cannot stay outside $X_\epsilon$ for infinitely many iterations, as we would then contradict the fact that the (finite) minimum exists only in $X_\epsilon$. Thus, we must fall into $X_\epsilon$ for any value of $\epsilon > 0$ in a finite number of iterations, so we have arbitrarily precise convergence in finite time.

Thus, we have a convergence result as long as we know that we can satisfy the aforementioned assumptions. In the next section, we prove Assumption 2.4 for all polyhedra and show that $k = 2$ satisfies Assumption 2.1 for the probability simplex.

### 3 Convergence Conditions for Specific Compact Convex Sets

#### 3.1 Assumption 2.4 for Polyhedra

We begin by showing that $h_\epsilon(y)$ is continuous when $X$ is a (compact) polyhedron, or $X = \{ x : Fx \preceq b \}$.\(^1\) In fact, this holds even when we consider moving not just $k$ coordinates but also in any linear subspace. Let $V \subset \mathbb{R}^n$ be a subspace of $\mathbb{R}^n$, meaning that $0 \in V$ and for any $v_1, v_2, \ldots, v_k \in V$, we have $\sum_{i=1}^k a_i v_i \in V$ for any $a_i \in \mathbb{R}$. Consider the set-valued mapping $A(y) = X \cap \{ y + V \}$. For $y \in X$, $A(y)$ is a nonempty compact convex set. Then, $A(y) = \{ x : Fx \preceq b, Gx = Gy \}$, where $G$ has rows that form an orthonormal basis for $V^\perp$.

The claim is that $p(y) := \inf_{x \in A(y)} \nabla f(y)^T (x - y) + \frac{L}{2} \| x - y \|^2$ is continuous in $y$.

**Proposition 3.1** The function $p(y) := \inf_{x \in A(y)} \nabla f(y)^T (x - y) + \frac{L}{2} \| x - y \|^2$ is continuous in $y$ over the restricted domain $X$.

**Proof** The primal problem can be written as

$$\begin{align*}
\text{minimize} & \quad \nabla f(y)^T (x - y) + \frac{L}{2} \| x - y \|^2 \\
\text{subject to} & \quad Fx \preceq b, Gx = Gy.
\end{align*}$$

\(^1\)Any equality constraints for the polyhedron can be rewritten as two inequality constraints.
Strong duality immediately obtains for this convex problem since all inequalities are affine and the problem is feasible (as \( y \) is a feasible point). Introducing the dual variables \( \lambda \) for \( \mu \) for the inequality and equality constraints respectively, we have the dual function:

\[
g(\lambda, \mu) = \inf_x \nabla f(y)^T(x - y) + \frac{1}{2} \|x - y\|^2 + \lambda^T(Fx - b) + \mu^T(Gx - Gy)
\]

\[
= -\frac{1}{2L}F^T\lambda + G^T\mu + \nabla f(y)^T + (Fy - b)^T\lambda.
\]

The dual problem is

- maximize \( g(\lambda, \mu) \)
- subject to \( \lambda \geq 0 \).

Taking the maximum over \( \mu \), we have the reduced problem:

- maximize \( -\frac{1}{2L}R(F^T\lambda + \nabla f(y))^2 + (Fy - b)^T\lambda \)
- subject to \( \lambda \geq 0 \),

where \( R = I - G^TG \). At optimality, we have (via the KKT conditions) [BV04]:

\[
x^\star = -\frac{1}{L}R(F^T\lambda^\star + \nabla f(y)) + y, \quad \mu^\star = -G(\nabla f(y) + F^T\lambda^\star)
\]

Now, strong duality holds and we know that \( p(y) \) is finite for \( y \in X \) since \( A(y) \) is compact and nonempty. Furthermore, we know that \( x^\star \) exists and is an element of \( A(y) \) since we have taken the infimum of a continuous function over a compact set. Thus \( x^\star \) is finite.

Now we show that we can choose \( \lambda^\star(y) \) such that \( \|\lambda^\star(y)\|_2 \) is finite as well. Note that the optimal objective must be at least the dual objective value at \( \lambda = 0 \), or \( -\frac{1}{2L}\|R\nabla f(y)\|^2_2 \), which implies that \( 0 \geq (Fy - b)^T\lambda^\star(y) \geq -\frac{1}{2L}\|R\nabla f(y)\|^2_2 \). Since \( \nabla f(y) \) is continuous, we know that \( \|R\nabla f(y)\|^2_2 \) is bounded above by some constant \( C \). Consider a sequence of points \( y^t \to y^\pm \), where \( y^t \in X \). If \( \lambda^\star(y^t) \to \infty \), then we must have \( (Fy^t - b)_i \to 0 \) as \( y^t \to y^\pm \) to satisfy the lower bound \( (Fy - b)^T\lambda^\star(y^\pm) \geq -C/2L \). Define the binary vector \( d_{\infty} \), which is nonzero at index \( i \) if the corresponding \( \lambda^\star_i(y^t) \to \infty \). Furthermore, define \( E := \text{diag} \: d_{\infty} \).

Then, since \( x^\star(y^\pm) \) is finite, we have (via the KKT condition):

\[
\limsup_{y^t \to y^\pm} \|RF^TEx^\star(y^t)\|_2 < \infty,
\]

so we can set \( Ex^\star(y^t) = \alpha(y^t) + \beta(y^t) \), where \( \|\alpha(y^t)\| \) is bounded and \( \beta(y^t) \in \ker(RF^T) \). Since \( Ex^\star(y^t) \) is unbounded above as \( t \to \infty \), we can choose some \( m \) along the path \( y^t \to y^\pm \) such that for \( t \geq m \), \( \beta(y^t) \geq 0 \). Then setting \( \beta(y^t) = 0 \) has no effect on the quadratic term in the objective, and it can only increase the linear term or leave it unchanged. Thus we either have a contradiction for the optimality of \( \lambda^\star(y^\pm) \) or we have found another finite optimal \( \lambda^\star(y^\pm) \). In either case, we have found that it is possible to have

\[
\limsup_{y^t \to y^\pm} \|\lambda^\star(y^t)\|_2 < \infty \quad \forall \: y^\pm \in X.
\]

Furthermore, we can achieve this result using a sequence of minimal-norm optimal dual variables \( \lambda^{**}(y) := \arg\min_{\lambda^\star(y)} \|\lambda^\star(y)\|_2 \). Now, if we have \( \sup_{y \in X} \|\lambda^{**}(y)\| = \infty \), there exists a sequence \( (y^p) \subset X \) such that \( \limsup_p \|\lambda^{**}(y^p)\| = \infty \). Then we can take a convergent
subsequence \( (y^q) \to y^1 \) satisfying \( \| \lambda^*(y^q) \| \to \infty \) since \( X \) is compact. This is a contradiction of the fact that this sequence satisfies (4). Then, we must have \( \sup_{y \in X} \| \lambda^*(y) \| < \infty \).

The above results show that the dual optimal variable can be taken to be bounded, so we can add a compactifying constraint \( \| \lambda \|_2 \leq K \) to the dual problem. In this way, we observe that \( p(y) \) is a supremum over \( \lambda \) in a compact set of a function continuous in \( (y, \lambda) \). Thus, \( p(y) \) is continuous over \( X \). ■

### 3.2 Assumption 2.1 for the Probability Simplex

Now we specialize to the probability simplex, \( \Delta := \{ x : x \geq 0, 1^T x = 1 \} \).

**Proposition 3.2** For the probability simplex, \( k = 2 \) is sufficient to satisfy Assumption 2.1.

**Proof** Denote a feasible pairwise perturbation from \( y \) as a perturbation from \( y \) that modifies only two coordinates and ends at a point \( z \in \Delta \). Satisfying the claim above is equivalent to proving the claim that \( y \) is optimal if and only if all feasible pairwise perturbation directions from \( y \) are ascent directions.

(\( \iff \)) If \( y \) is optimal, then we have the optimality condition \( v^T (z - y) \geq 0 \) for any \( z \in \Delta \). This includes all \( z \) that can be reached by a feasible pairwise (i.e. \( k = 2 \)) perturbation.

(\( \implies \)) Note that a pairwise perturbation direction is given by \( \delta_{ij} := (e_i - e_j)/\sqrt{2} \) where \( e_i \) is a standard basis vector and \( i \neq j \). For \( y \in \text{relint} \Delta, y > 0 \). Otherwise, some elements of \( y \) are 0 and for some \( \delta_{ij} \), we have either \( a_{ij} > 0 \) or \( a_{ij} < 0 \) yields an infeasible perturbation \( a_{ij} \delta_{ij} \). Nevertheless, we show that we can generate a perturbation towards any corner \( e_i \) using only feasible pairwise perturbations from any \( y \in \Delta \). Without loss of generality, suppose that the first \( m \geq 1 \) elements \( y_1, y_2, ..., y_m \) are nonzero followed by all \( y_{m+1}, ..., y_n = 0 \). Then, suppose we want to generate a perturbation in the direction of \( e_m \) with some scaling \( \alpha > 0 \).\(^2\)

This can be done in the following manner:

\[
\alpha \left( (1 - y_m) e_m - \sum_{i=1}^{m-1} y_i e_i \right) = -\alpha \left( y_1 \delta_{12} + (y_1 + y_2) \delta_{23} + ... + \left( \sum_{i=1}^{m-1} y_i \right) \delta_{m-1,m} \right),
\]

since \( \sum_{i=1}^{m-1} y_i = 1 - y_m \). For small enough \( \alpha \), all of the pairwise perturbations in the right hand side are feasible since \( y_i > 0 \) for \( 1 \leq i \leq m \). We can also generate a perturbation in the direction of \( e_{m+1} \) in the following manner:

\[
\alpha \left( e_{m+1} - \sum_{i=1}^{m} y_i e_i \right) = -\alpha \left( y_1 \delta_{12} + (y_1 + y_2) \delta_{23} + ... + \left( \sum_{i=1}^{m} y_i \right) \delta_{m,m+1} \right),
\]

since \( \sum_{i=1}^{m} y_i = 1 \). Again, for small enough \( \alpha \), all of the perturbations are feasible, so we have shown that we can generate perturbation directions towards any corner using only feasible pairwise perturbations. Now, since any \( z \in \Delta \) can be written as a convex combination of corners \( e_i \), we can generate a perturbation direction towards any \( z \in \Delta \) using feasible pairwise perturbations from \( y \in \Delta \). Thus, if all feasible pairwise perturbations are ascent directions, it must be the case that all feasible perturbation directions are ascent directions, so \( y \) is optimal. ■

\(^2\)Of course, if \( m = 1 \) and we are already at \( y = e_m \), then no action is necessary.
4 Experiments

We solve an LP over probability simplexes of various dimensions using uniformly random stochastic sampling of $k = 2$ coordinates. Specifically, we solve the problem

$$\begin{align*}
\text{minimize} \quad & v^T y \\
\text{subject to} \quad & y \succeq 0, \quad 1^T y = 1
\end{align*}$$

for $y \in \mathbb{R}^n$, where $n = 10, 50, 100, 250, 500, 1000$, $v_i$ drawn iid. from $\mathcal{N}(0, 0.25)$, and $y^0 = 1/n$. Results are shown Figure 1. Although we have not yet characterized theoretical guarantees for convergence rates, we see that we have convergence for this specific problem in $O(n^2)$ time using uniformly random stochastic sampling of coordinates. We also recognize that the rate of convergence will likely depend on the set $X$. Nevertheless, these results illustrate the benefits of our method, as each iteration is a simple two-dimensional minimization problem that has an analytical solution.

5 Conclusion

We have developed a framework for performing coordinate descent in the presence of coupled constraints. Namely, we have proved that one can achieve arbitrarily precise convergence in a finite number of iterations given various conditions that depend on the constraint set. We have proven these conditions for polyhedra and have demonstrated empirical convergence rates over the probability simplex. Generalizing these results to intersections of compact convex sets and determining theoretical guarantees for convergence rates (which will likely depend on $X$) are necessary to make our framework readily applicable to practical problems.
References


