Coordinate Descent with Coupled Constraints

Aman Sinha
EE364b: Convex Optimization II Class Project

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 large. Thus, both batch optimization techniques and many stochastic techniques that require working with the entire \( \theta \) vector (e.g., mirror descent methods) are too inefficient.

Coordinate descent methods work only with a single coordinate of \( \theta \) at each iteration. However, standard coordinate descent methods typically assume that the constraint sets for different coordinates are independent. We 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 \( \theta \) is determined to fit a model to a very large dataset with \( N \) observations.

**Problem Setup and Approach**

Consider a nonempty compact convex space \( X \subset \mathbb{R}^n \) and a differentiable convex function \( f(y) \) with \( L \)-Lipschitz gradient. Solve the problem:

\[
\begin{align*}
\text{minimize}_{y \in X} f(y) \\
\end{align*}
\]

by moving only \( k \) coordinates at a time.

**Approach:** minimize a convex upper bound of \( f(y) \):

\[
y^{(i+1)} = \min_{s} \nabla f(y^{(i)})^T (z - y^{(i)}) + \frac{L}{2} ||z - y^{(i)}||^2
\]

- \( s \) is a binary vector determining the \( k \) coordinates to change
- \( x_{i,y} := \{ z \in X, \text{diag}(1-s)(z-y) = 0 \} \)
- Equivalent to gradient step with size \( 1/L \) projected onto \( x_{i,y} \)
- \( \nabla f(y^{(i)})^T (y^{(i+1)} - y^{(i)}) + \frac{L}{2} ||y^{(i+1)} - y^{(i)}||^2 \leq 0 \) and \( f(y^{(i+1)}) \leq f(y^{(i)}) \)

**Convergence Analysis (High-Level)**

**Assumptions:**

- Choose \( k \) such that we can always make strict progress towards the optimum if we aren’t already there (e.g., \( f(y^{(i+1)}) < f(y^{(i)}) \))
- We sample \( s \) such that at any iteration \( m \), we perform the necessary move to make strict progress (possibly with multiple draws of \( s \)).

**Key idea:**

Given above assumptions, suppose \( \min_{y \in x_{i,y}} h_s(y) < 0 \), where \( x_{i,y} := \{ y : f(y) < \min_{y \in x_{i,y}} f(y) + \epsilon \} \). Then, the iterations converge to an \( \epsilon \)-suboptimal neighborhood of an optimal \( y^* \in \arg\min_{y \in X} f(y) \) in finite time.

**Convergence Analysis (Details)**

\( h_s(y) \) is continuous over \( X \) when \( X \) is a polyhedron.

**Key ideas:**

The primal problem is

\[
\begin{align*}
\text{minimize} & \quad \nabla f(y)^T (x - y) + \frac{L}{2} ||x - y||^2 \\
\text{subject to} & \quad F y \succeq b, G y = c y
\end{align*}
\]

and strong duality holds. The (reduced) dual problem is

\[
\begin{align*}
\text{maximize} & \quad \frac{1}{2} L \nabla f(y)^T \lambda + \nabla f(y)^T x \\
\text{subject to} & \quad L \lambda \succeq 0
\end{align*}
\]

and the KKT condition is \( x^* = -1/2 L \nabla f(y)^T \lambda + \nabla f(y) + y, y \succeq 0 \). Since \( X \) is compact, \( x^* \) is finite. Then, we can choose \( \lambda^* \) such that \( ||\lambda^*||_2 \leq M \) (see paper for full proof of this result). Thus, we can add a compactifying constraint \( ||\lambda||_2 \leq M \) to the dual problem. So \( h_s(y) \) is a supremum over \( \lambda \) in a compact set of a function continuous in \( (y, \lambda) \). Thus, \( h_s(y) \) is continuous over \( X \).

We can set \( k = 2 \) for the probability simplex \( \Delta \)

**Key ideas:**

Denote a feasible pairwise perturbation as a perturbation from \( y \) that modifies only two coordinates and ends at a point \( z \in \Delta \). Need to show that \( y \) is optimal if and only if if feasible pairwise perturbations are ascent directions.

\( \iff \) Optimality condition \( v^T (x - y) \succeq 0 \) for any \( z \in \Delta \).

\( \implies \) Consider two cases. In the first case, \( y \notin \text{relint} \Delta \). In this case, \( v^T (x - y) \) is finite. Then, we can choose \( \lambda^* \) such that \( ||\lambda||_2 \leq M \) (see paper for full proof of this result). Thus, we can add a compactifying constraint \( ||\lambda||_2 \leq M \) to the dual problem. So \( h_s(y) \) is a supremum over \( \lambda \) in a compact set of a function continuous in \( (y, \lambda) \). Thus, \( h_s(y) \) is continuous over \( X \).

**Experiments**

We solve an LP over probability simplexes of various dimensions using uniformly random stochastic sampling of \( k = 2 \) coordinates. Specifically, we solve

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

for \( y \in \mathbb{R}^n \), where \( n \in \{10, 50, 100, 250, 500, 1000\} \), and \( v_i \) are drawn iid. from \( N(0, 0.25) \), and \( y^0 = 1/n \).

**Conclusion**

We have developed a framework for doing coordinate descent in the presence of coupled constraints. We proved high-level convergence given various assumptions, and then proved these assumptions for special cases of compact convex sets. Experimental results are given for solving an LP over the probability simplex. Future work will attempt to generalize convergence results to other compact convex sets and characterize convergence rates.

**Acknowledgements**

Many thanks to Prof. Duchi and the course staff for their feedback.