# Duality Connections Between Variants of the Kaczmarz and Motzkin Methods and Variants of Randomized Coordinate Descent

**Jamie Haddock**[*]
Graduate Group in Applied Mathematics
Department of Mathematics
University of California, Davis
Davis, CA 95616
jhaddock@math.ucdavis.edu

## Abstract

The Kaczmarz method for solving a system of equations is well known to be equivalent (under primal-dual variable transformation) to coordinate descent applied to a dual problem. Here we examine the dual equivalency of a coordinate descent type method to Motzkin's relaxation method for solving a system of equations and to a hybrid of Motzkin's method and the randomized Kaczmarz method. Finally, we discuss what can be said about equivalencies for these methods applied to the linear feasibility problem.

Projection type methods are popular in numerical linear algebra as well as numerical optimization. In particular, the randomized Kaczmarz method for solving systems of equations is widely used in computer tomography and signal/image processing. Meanwhile, Motzkin's relaxation method was developed to solve linear programming problems in the 1950's but has fallen out of vogue due to its computational inefficiency. However, there is an entire class of methods that can be seen to interpolate between these two classical methods.

All of these methods can be phrased in the same way (with each method differing in step 2):
**Method 1.** *Suppose $A \in \mathbb{R}^{m\times n}$, $b \in \mathbb{R}^m$ and the associated polyhedron $P := \{x \in \mathbb{R}^n : Ax \le b\}$ is nonempty. Let $x_0 \in \mathbb{R}^n$ be given. Fix $0 < \lambda \le 2$. We iteratively construct approximations to a solution lying in $P$ in the following way:*

1. *If $x_k$ is feasible, stop.*
2. *Choose $i_k$ to be the index of a row constraint of $A$, $i_k \in [m]$.*
3. *Define $x_k := x_{k-1} - \lambda \frac{(a_{i_k}^T x_{k-1} - b_{i_k})^+}{||a_{i_k}||^2} a_{i_k}$.*
4. *Repeat.*

In the randomized Kaczmarz method $i_k$ is chosen randomly with probability proportional to the norm of the associated row of $A$, $a_{i_k}$ [1]. In Motzkin's relaxation method, $i_k$ is chosen as the index of the constraint

[*]math.ucdavis.edu/˜jhaddock

which is violated maximally: $i_k := \underset{i \in [m]}{\text{argmax}} a_i^T x - b_i$ [3]. Each of these methods can be seen as a member of a generalized class of methods which we refer to as the Sampling-Kaczmarz Motzkin methods:
**Method 2** (SKMM)**.** *Suppose $A \in \mathbb{R}^{m \times n}$ ($A$ normalized so that $||a_i|| = 1$), $b \in \mathbb{R}^m$ and the associated polyhedron $P := \{x \in \mathbb{R}^n : Ax \leq b\}$ is nonempty. Let $x_0 \in \mathbb{R}^n$ be given. We iteratively construct approximations to a solution lying in $P$ in the following way:*

1. *If $x_k$ is feasible, stop.*
2. *Choose $\tau_k$ to be a sample of size $\beta$ constraints chosen uniformly at random from among the rows of $A$.*
3. *From among these $\beta$ rows, choose $i_k := \underset{i \in \tau_k}{\text{argmax}} a_i^T x_{k-1} - b_i$.*
4. *Define $x_k := x_{k-1} - \lambda \frac{(a_{i_k} x_{k-1} - b_{i_k})^+}{||a_{i_k}||^2} a_{i_k}$.*
5. *Repeat.*

Note that the randomized Kaczmarz method is SKMM with $\beta = 1$ while Motzkin's method is SKMM with $\beta = m$.

**Duality Connections for $Ax = b$**

It is well known that the randomized Kaczmarz method for solving a system of linear equations can be shown to construct iterates which are equivalent to those iterates constructed by the randomized coordinate descent method applied the the dual problem [2]. If $A \in \mathbb{R}^{m \times n}$ with $m \geq n$ and $A$ full rank then we can pose solving the system of linear equations as the primal problem

$$\begin{aligned} &\min \frac{1}{2} x^T x \\ &\text{s.t. } Ax = b \end{aligned} \qquad (P)$$

This problem has Lagrangian, $L(x, y) = \frac{1}{2} x^T x + y^T (b - Ax)$ and then $\nabla_x L(x, y) = x - A^T y$ so the minimum of $L(x, y)$ with respect to $x$ is achieved when $x = A^T y$. Thus, the dual objective is $g(y) = \frac{1}{2} ||A^T y||^2 + y^T b - ||A^T y||^2 = b^T y - \frac{1}{2} ||A^T y||^2$ so the dual problem is

$$\min \frac{1}{2} ||A^T y||^2 - b^T y \qquad (D)$$

Now if we apply randomized coordinate descent to (D) with arbitrary $y_0$ and coordinate $i$ chosen with probability proportional to $\frac{||a_i||^2}{||A||_F^2}$ and step size $\frac{\lambda}{||a_i||^2}$ then we obtain the update

$$y_{k+1} = y_k - \lambda \frac{(a_i^T A^T y_k - b_i)}{||a_i||^2} e_i$$

Now, we use the dual-primal variable relationship $x = A^T y$ and see that

$$\begin{aligned} x_{k+1} &= A^T y_{k+1} \\ &= A^T y_k - \lambda \frac{(a_i^T A^T y_k - b_i)}{||a_i||^2} A^T e_i \\ &= x_k - \lambda \frac{(a_i^T x_k - b_i)}{||a_i||^2} a_i \end{aligned}$$

which are exactly the Kaczmarz iterates produced on the system $Ax = b$.

Now, in the same way we see that Motzkin's method is equivalent to coordinate descent where the index selected is the index of the maximum entry in the gradient vector. In particular we can say that any SKMM method (with a particular choice of $\beta$ and $\lambda$) is equivalent, via the primal-dual variable relationship to the coordinate descent method variant:

**Method 3.** *Suppose $A \in \mathbb{R}^{m\times n}$, $b \in \mathbb{R}^m$ with $m \geq n$ and $A$ full rank. Let $y_0 \in \mathbb{R}^m$ be given. Fix $0 < \lambda \leq 2$. We iteratively construct approximations to the solution of* $\min f(y) := \frac{1}{2}||A^T y||^2 - b^T y$ *in the following way:*

1. *If $\nabla f(y_k) = AA^T y_k - b = 0$, stop.*
2. *Choose $\tau_k$ to be a sample of size $\beta$ entries chosen uniformly at random from among the entries of $\nabla f(y_k)$.*
3. *From among these $\beta$ entries, choose $i_k := \underset{i\in\tau_k}{\text{argmax}}\ a_i^T A^T y_{k-1} - b_i$, the largest entry of this sample of the gradient vector.*
4. *Define $y_k := y_{k-1} - \lambda \frac{a_{i_k} A y_{k-1} - b_{i_k}}{||a_{i_k}||^2} e_{i_k}$.*
5. *Repeat.*

Thus, when the solution of the system of linear equations is unique, SKMM methods can be seen as simple variants of the coordinate descent method applied to *a* dual problem corresponding to one of the infinite choices of primal problems we could have defined (since there is a unique feasible point).

**Duality Connections for $Ax \leq b$**

Given our previous analysis, one could hope that we will see the same equivalencies for these methods applied to systems of linear inequalities. However, the main issue with this attempt is that the solution to the system of linear inequalities is not necessarily unique and thus, these methods are not guaranteed to solve any primal problem corresponding to a choice of objective function,

$$\min f(x) \text{ s. t. } Ax \leq b,$$

unless the objective function is a constant function (like 0).

However, if we define the variable relationship $x = A^T y$, then we can still find a relationship between the SKMM methods and something that resembles a variant of coordinate descent (although it is not clear what problem it should be applied to). If we use the above mentioned relationship, $x = A^T y$, then we see that the SKMM method is equivalent to:

**Method 4.** *Suppose $A \in \mathbb{R}^{m\times n}$ ($A$ normalized so that $||a_i|| = 1$), $b \in \mathbb{R}^m$ with $P := \{x|Ax \leq b\}$ nonempty . Let $y_0 \in \mathbb{R}^m$ be given. Fix $0 < \lambda \leq 2$. We iteratively construct:*

1. *If $AA^T y_k - b \leq 0$, stop.*
2. *Choose $\tau_k$ to be a sample of size $\beta$ entries chosen uniformly at random from among the entries of $AA^T y_k - b$.*
3. *From among these $\beta$ entries, choose $i_k := \underset{i\in\tau_k}{\text{argmax}}\ a_i^T A^T y_{k-1} - b_i$, the largest entry of this sample of the gradient vector.*
4. *Define $y_k := y_{k-1} - \lambda \frac{(a_{i_k} A y_{k-1} - b_{i_k})^+}{||a_{i_k}||^2} e_{i_k}$.*
5. *Repeat.*

As mentioned before, however, if this is to be considered a coordinate descent method applied to the gradient of some objective function, it is not clear that this method solves that problem.

However, notice that the SKMM methods will approximate a feasible solution (but not optimal solution) [2] for the problem

$$\min \frac{1}{2}x^T x \text{ s. t. } Ax \leq b \qquad (P)$$

Now, note that this problem is equivalent to

$$\min \frac{1}{2}x^T x + \delta_{\leq b}(Ax)$$

so we have that the dual problem is

$$\max -f^*(A^T y) - g^*(-y)$$

where $f(x) = \frac{1}{2}||x||^2$ and $g(x) = \delta_{\leq b}(x)$. Now, the convex conjugate of $f(x)$ is $f^*(y) = f(y)$ and the convex conjugate of $g(x)$ is $g^*(y) = b^T y + \delta_{\geq 0}(y)$, so the dual problem is

$$\min \frac{1}{2}||A^T y||^2 - b^T y \text{ s.t. } y \leq 0 \qquad (D).$$

Now, note that provided $y_0 \leq 0$, Method 4 will maintain feasibility of the point and will decrease the gradient of this problem, provided it never increases any entry of $y$. Thus, the method will continue with all $y$ iterates feasible until all entries of the gradient are at most 0.

Thus, in the case of linear equations with a unique solution, when the SKMM methods solved the primal problem, they could be seen as equivalent to coordinate descent methods which solved the primal problem. However, in the case of linear inequalities, since the SKMM methods are only guaranteed to converge to a feasible solution of the primal problem (thus they are not a method for solving the problem described), we do not get neat equivalency with coordinate descent methods applied to the dual problem. The equivalent method acts to decrease the entries of the gradient (residual) until they are non-positive

**Conclusion**

The SKMM methods and in particular, the randomized Kaczmarz method and Motzkin's method applied to linear systems of equations with a unique solution can be easily seen to be equivalent to variants of the coordinate descent method applied to the dual problem, precisely because we could choose how to describe the primal problem (given a unique feasible solution). However, in the case of inequalities, this equivalency is nowhere near as neat since the SKMM methods are not designed to find the minimum norm solution to a system of inequalities. They are only able to solve the linear feasibility problem and thus, the method viewed in the variable $y$ (related to $x$ given the relationship $x = A^T y$) does not have a neat geometric interpretation.

**Acknowledgments**

This paper discusses unpublished work due to collaboration with Jesús De Loera and Deanna Needell.

**References**

## References

[1] S. Kaczmarz. Angenaherte auflosung von systemen linearer gleichungen. *Bull.Internat.Acad.Polon.Sci.Lettres A*, pages 335–357, 1937.

[2] D. Leventhal and A. S. Lewis. Randomized methods for linear constraints: convergence rates and conditioning. *Math.Oper.Res.*, 35(3):641–654, 2010. 65F10 (15A39 65K05 90C25); 2724068 (2012a:65083); Raimundo J. B. de Sampaio.

[3] T. S. Motzkin and I. J. Schoenberg. The relaxation method for linear inequalities. *Canadian J. Math.*, 6:393–404, 1954.