A Sampling Kaczmarz-Motzkin Algorithm for Linear Feasibility

Jamie Haddock

Graduate Group in Applied Mathematics,
Department of Mathematics,
University of California, Davis

ICCOPT
August 11, 2016

Joint work with Jesus De Loera and Deanna Needell

Linear Feasibility Problem

We are interested in solving the *linear feasibility problem* (LF):

Linear Feasibility Problem

We are interested in solving the *linear feasibility problem* (LF):

Find x such that $Ax \leq b$ or conclude one does not exist.

Linear Feasibility Problem

We are interested in solving the *linear feasibility problem* (LF):

Find x such that $Ax \leq b$ or conclude one does not exist.

We consider large-scale problems in which $A \in \mathbb{R}^{m \times n}$, $m >> n$.

Linear Feasibility Problem

We are interested in solving the *linear feasibility problem* (LF):

Find x such that $Ax \leq b$ or conclude one does not exist.

We consider large-scale problems in which $A \in \mathbb{R}^{m \times n}$, $m >> n$.

These problems arise in machine learning classification, *support-vector machines* (Boser, Guyon, Vapnik 1992), (Cortes, Vapnik 1995).

Projection Methods

If $P := \{x \in \mathbb{R}^n : Ax \leq b\}$ is nonempty, these methods construct an approximation to an element of P:

1. Motzkin's Relaxation Method(s)
2. Randomized Kaczmarz Method
3. Sampling Kaczmarz-Motzkin Method (SKM)

Motzkin's Relaxation Method(s)

Given $x_0 \in \mathbb{R}^n$, fix $0 < \lambda \leq 2$ and iteratively construct approximations to P:

1. If x_k is feasible, stop.
2. Choose $i_k \in [m]$ as $i_k := \operatorname*{argmax}_{i \in [m]} a_i^T x_{k-1} - b_i$.
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.

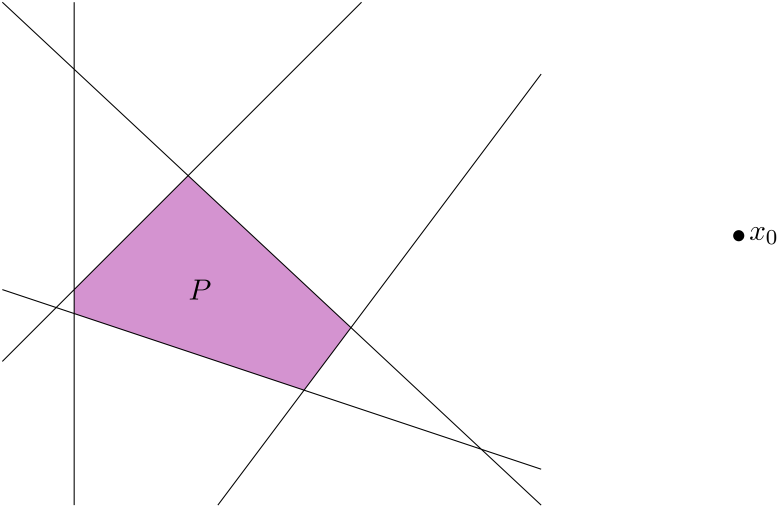
Motzkin's Relaxation Method(s)

Given $x_0 \in \mathbb{R}^n$, fix $0 < \lambda \leq 2$ and iteratively construct approximations to P:

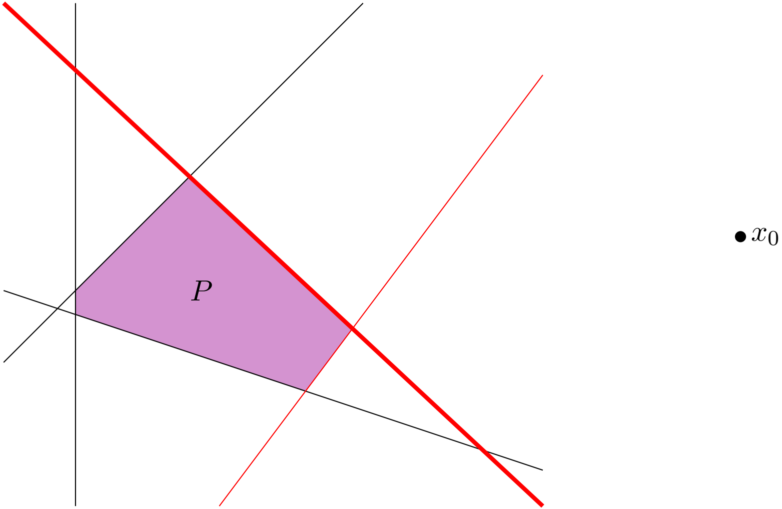
1. If x_k is feasible, stop.
2. Choose $i_k \in [m]$ as $i_k := \underset{i \in [m]}{\operatorname{argmax}}\, a_i^T x_{k-1} - b_i$.
3. Define $x_k := x_{k-1} - \lambda \frac{a_{i_k}^T x_{k-1} - b_{i_k}}{||a_{i_k}||^2} a_{i_k}$.
4. Repeat.

λ is the *projection (or relaxation) parameter*

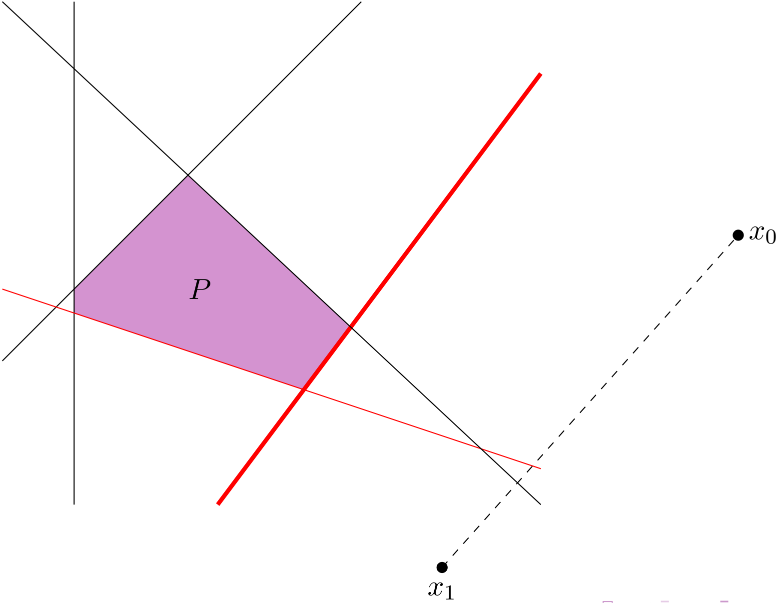
Motzkin's Method



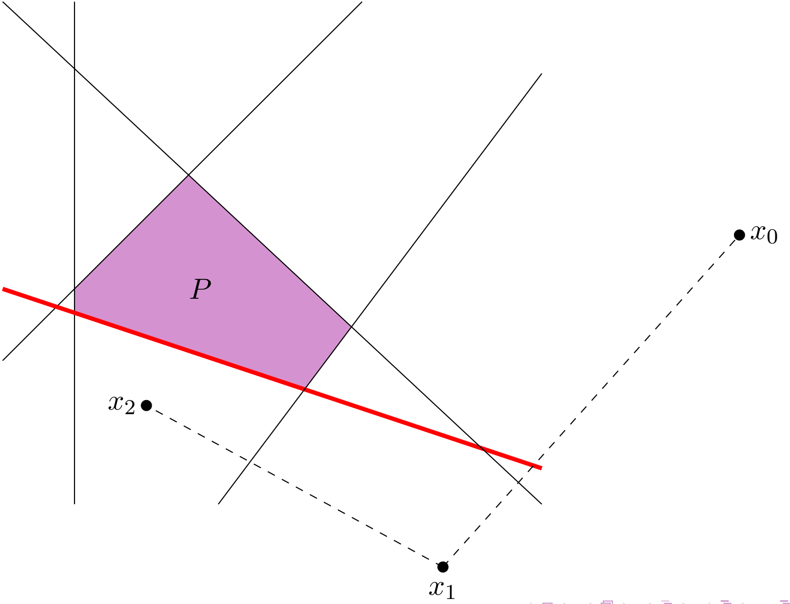
Motzkin's Method



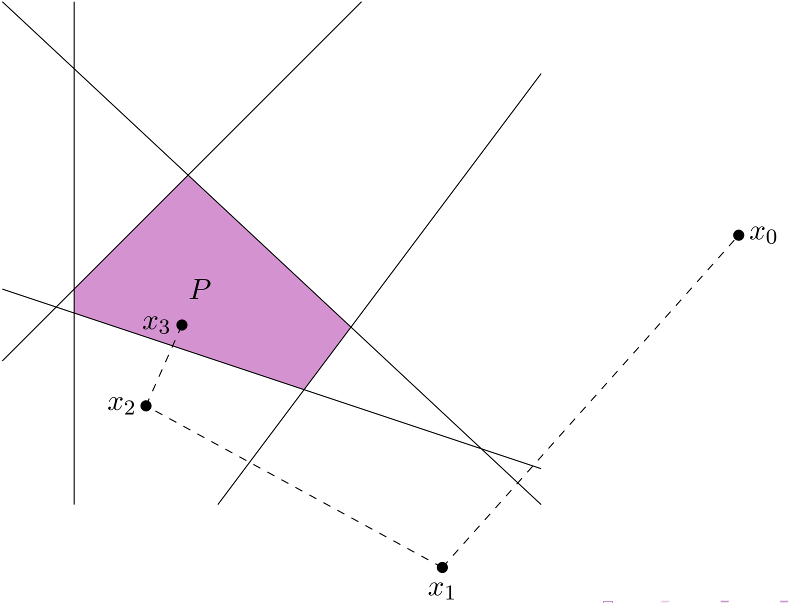
Motzkin's Method



Motzkin's Method



Motzkin's Method

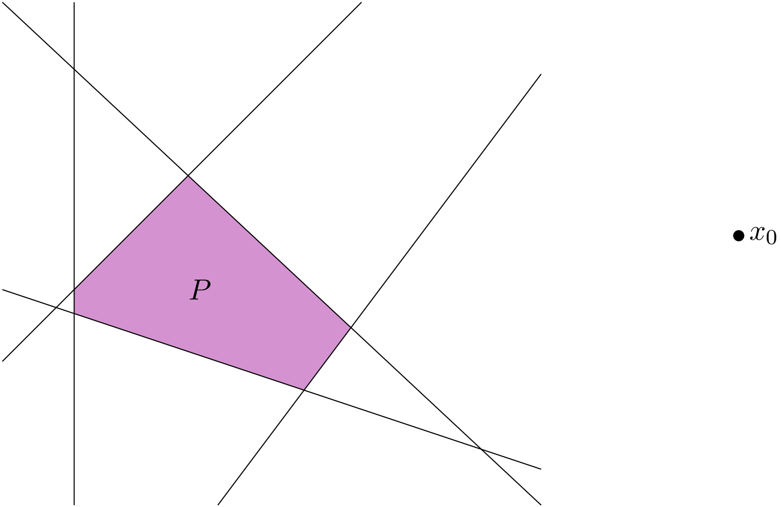


Randomized Kaczmarz Method

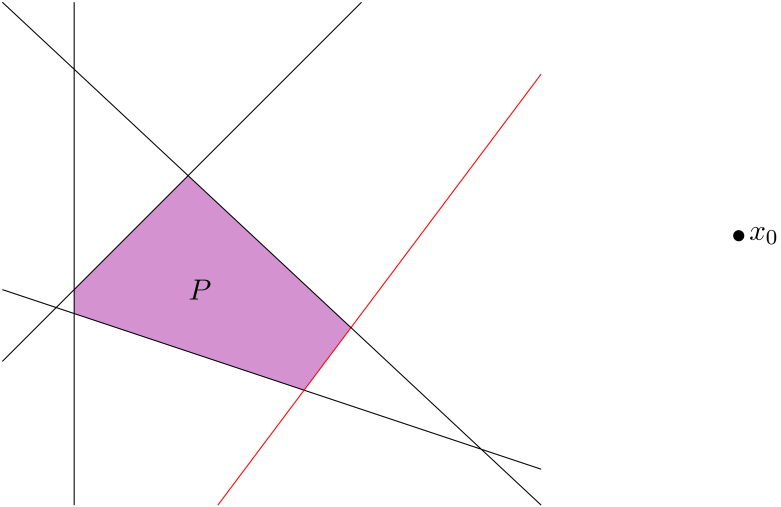
Given $x_0 \in \mathbb{R}^n$, iteratively construct approximations to P:

1. If x_k is feasible, stop.
2. Choose $i_k \in [m]$ with probability $\frac{||a_{i_k}||^2}{||A||_F^2}$.
3. Define $x_k := x_{k-1} - \frac{(a_{i_k}^T x_{k-1} - b_{i_k})^+}{||a_{i_k}||^2} a_{i_k}$.
4. Repeat.

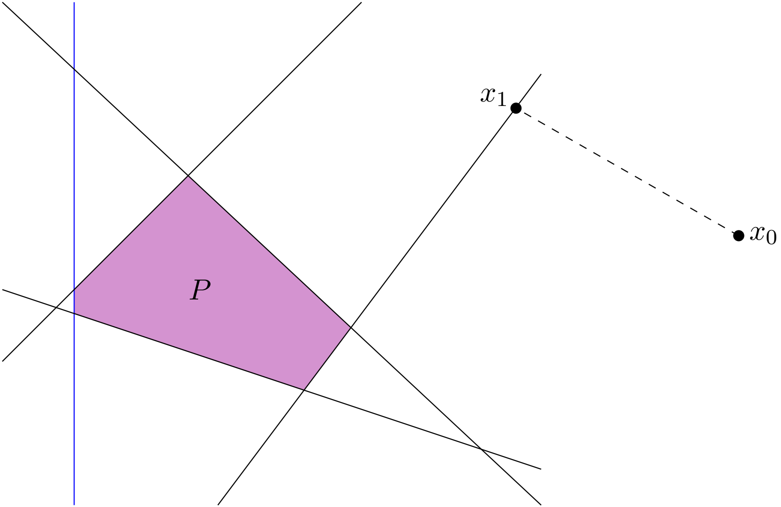
Kaczmarz Method



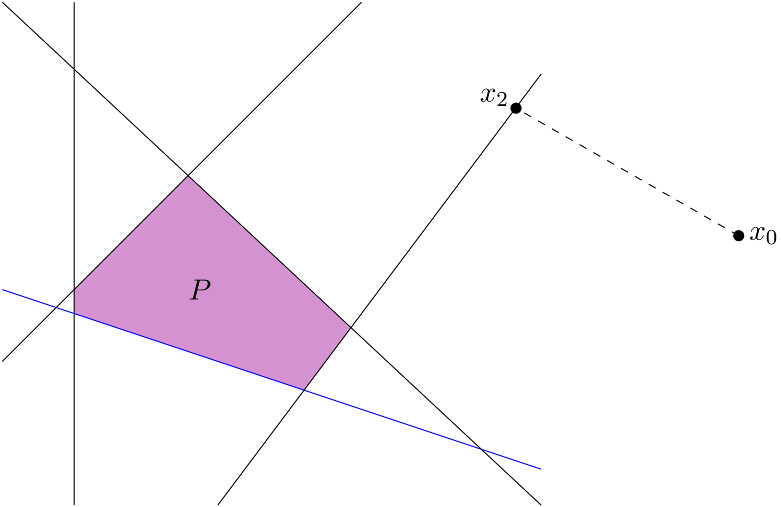
Kaczmarz Method



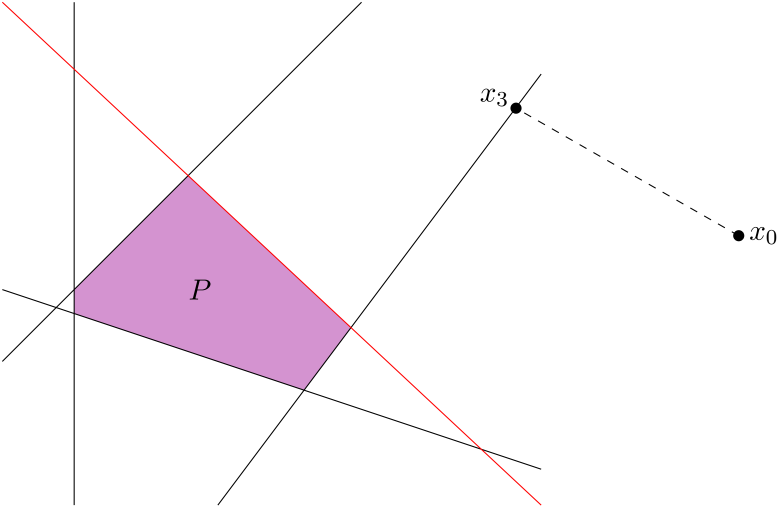
Kaczmarz Method



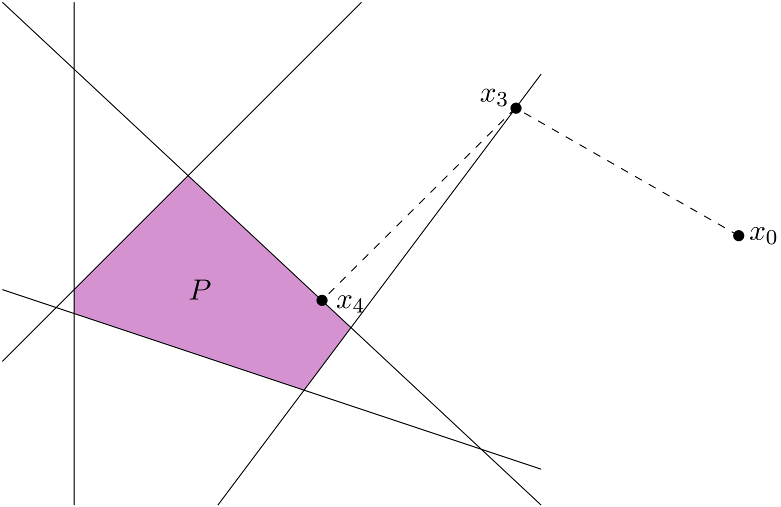
Kaczmarz Method



Kaczmarz Method



Kaczmarz Method

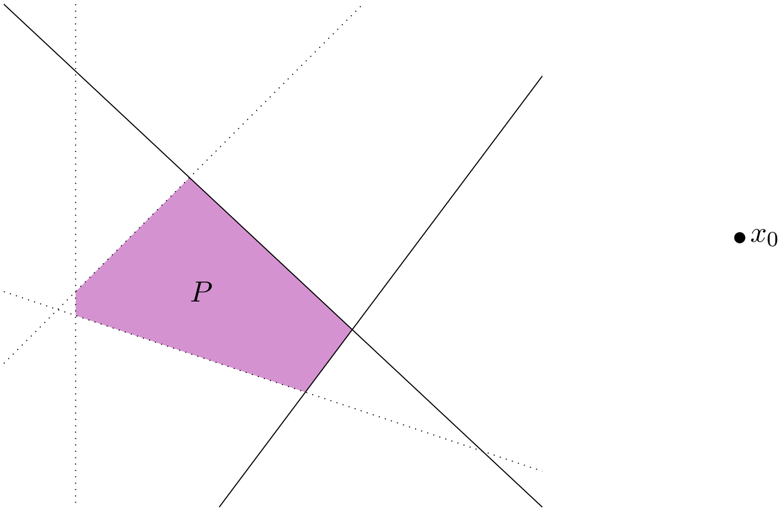


A Hybrid Method (SKM)

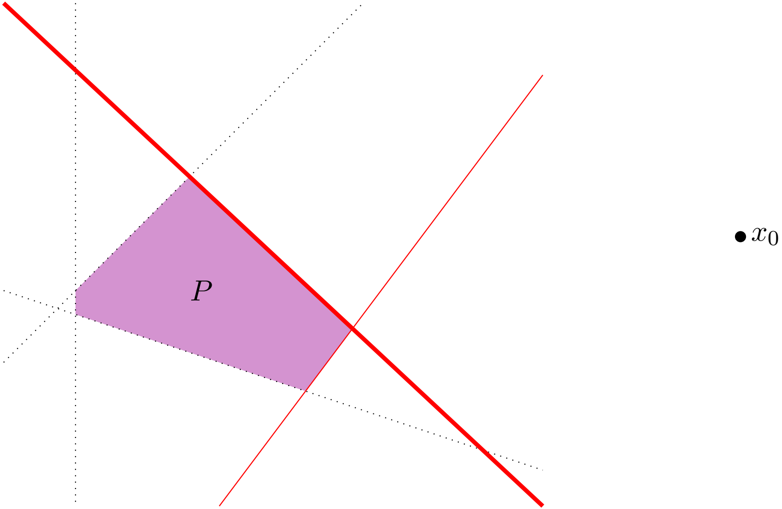
Given $x_0 \in \mathbb{R}^n$, fix $0 < \lambda \leq 2$ and iteratively construct approximations to P in the following way:

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

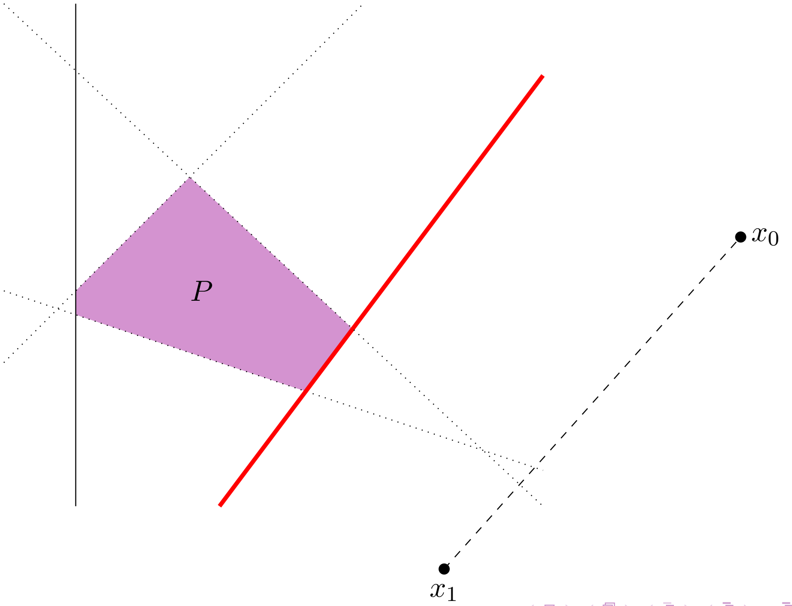
A Hybrid Method



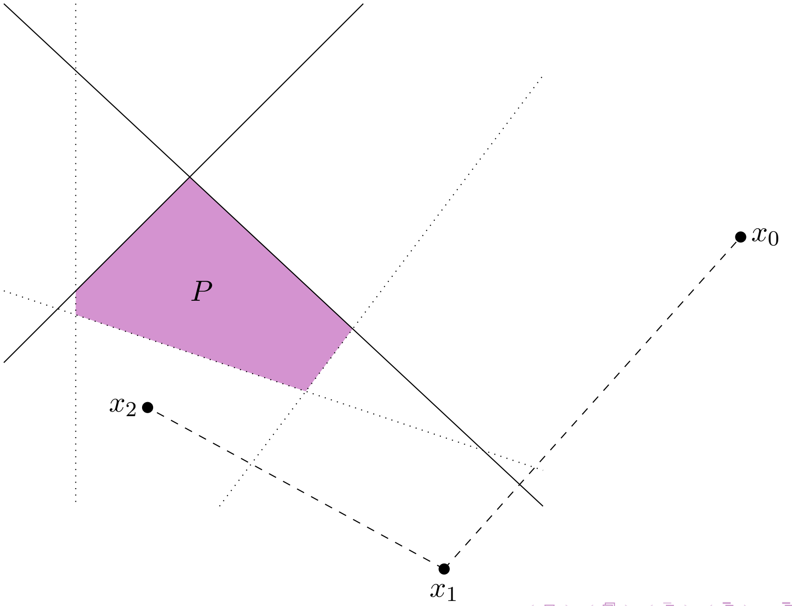
A Hybrid Method



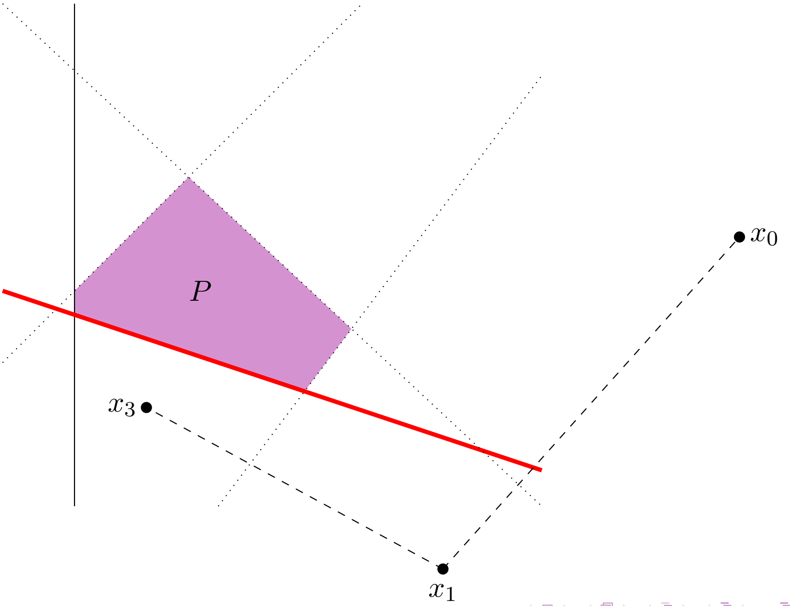
A Hybrid Method



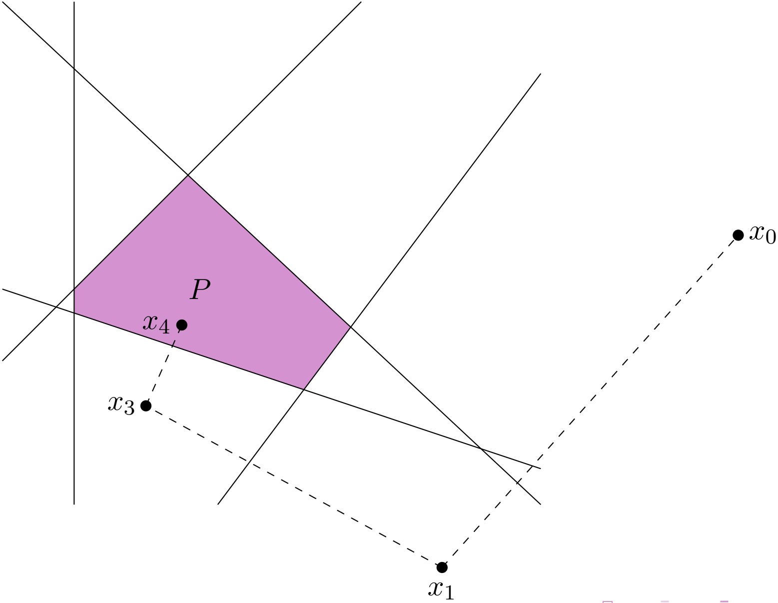
A Hybrid Method



A Hybrid Method



A Hybrid Method



SKM Method Convergence Rate

Theorem (De Loera, H., Needell)

If the feasible region (for row-normalized A) is nonempty, then the SKM methods with samples of size β converge at least linearly in expectation: If s_{k-1} is the number of constraints satisfied by x_{k-1} and $V_{k-1} = \max\{m - s_{k-1}, m - \beta + 1\}$ then

$$\mathbb{E}[d(x_k, P)^2] \leq \left(1 - \frac{2\lambda - \lambda^2}{V_{k-1} L_2^2}\right) d(x_{k-1}, P)^2$$

$$\leq \left(1 - \frac{2\lambda - \lambda^2}{m L_2^2}\right)^k d(x_0, P)^2.$$

SKM Method Convergence Rate

Theorem (De Loera, H., Needell)

If the feasible region (for row-normalized A) is nonempty, then the SKM methods with samples of size β converge at least linearly in expectation: If s_{k-1} is the number of constraints satisfied by x_{k-1} and $V_{k-1} = \max\{m - s_{k-1}, m - \beta + 1\}$ then

$$\mathbb{E}[d(x_k, P)^2] \leq \left(1 - \frac{2\lambda - \lambda^2}{V_{k-1} L_2^2}\right) d(x_{k-1}, P)^2$$

$$\leq \left(1 - \frac{2\lambda - \lambda^2}{m L_2^2}\right)^k d(x_0, P)^2.$$

The *Hoffman constant*, L_2 is an error bound defined as the minimum constant that satisfies

$$d(x, P) \leq L_2 ||(Ax - b)^+||_2.$$

Improved Rate

Theorem (De Loera, H., Needell)

If the feasible region, $P = \{x | Ax \leq b\}$ is nondegenerate (generic) and nonempty (for normalized A), then an SKM method with samples of size $\beta \leq m - n$ is guaranteed an increased convergence rate after some K:

$$\mathbb{E}[d(x_k, P)^2] \leq \left(1 - \frac{2\lambda - \lambda^2}{mL_2^2}\right)^K \left(1 - \frac{2\lambda - \lambda^2}{(m - \beta + 1)L_2^2}\right)^{k-K} d(x_0, P)^2.$$

Finiteness of Motzkin's Method

Theorem (Goffin 1980, Telgen 1982)

Suppose A, b are rational matrices with binary encoding length σ, and that we run a relaxation method on the normalized system $\tilde{A}x \leq \tilde{b}$ with $x_0 = 0$. Then either the relaxation method detects feasibility of the system within $k = \left\lceil \frac{2^{4\sigma}}{n\lambda(2-\lambda)} \right\rceil$ iterations or the system is infeasible.

Finiteness of Motzkin's Method

Theorem (Goffin 1980, Telgen 1982)

Suppose A, b are rational matrices with binary encoding length σ, and that we run a relaxation method on the normalized system $\tilde{A}x \leq \tilde{b}$ with $x_0 = 0$. Then either the relaxation method detects feasibility of the system within $k = \left\lceil \frac{2^{4\sigma}}{n\lambda(2-\lambda)} \right\rceil$ iterations or the system is infeasible.

The *binary encoding length* of the problem is

$$\sigma = \sum_{i=1}^{m}\sum_{j=1}^{n} \log(|a_{ij}| + 1) + \sum_{i=1}^{m} \log(|b_i| + 1) + \log(nm) + 2.$$

Certificates of Feasibility

Define the *maximum violation* in the point x to be

$$\theta(x) := \max\{0, \max_{i \in [m]} a_i^T x - b_i\}.$$

Certificates of Feasibility

Define the *maximum violation* in the point x to be

$$\theta(x) := \max\{0, \max_{i \in [m]} a_i^T x - b_i\}.$$

Lemma

If the rational system $Ax \leq b$ (with binary encoding length σ) is infeasible, then for all $x \in \mathbb{R}^n$, the maximum violation satisfies $\theta(x) \geq 2^{1-\sigma}$.

Certificates of Feasibility

Define the *maximum violation* in the point x to be

$$\theta(x) := \max\{0, \max_{i \in [m]} a_i^T x - b_i\}.$$

Lemma

If the rational system $Ax \leq b$ (with binary encoding length σ) is infeasible, then for all $x \in \mathbb{R}^n$, the maximum violation satisfies $\theta(x) \geq 2^{1-\sigma}$.

Thus, to detect feasibility of the rational system $Ax \leq b$, we need only find a point, x_k with $\theta(x_k) < 2 * 2^{-\sigma}$; such a point will be called a *certificate of feasibility*.

Expected Finiteness of SKM methods

Theorem (De Loera, H., Needell)

Suppose A, b are rational matrices with binary encoding length σ, and that we run an SKM method on the normalized system $\tilde{A}x \leq \tilde{b}$ with $x_0 = 0$. Suppose the number of iterations k satisfies

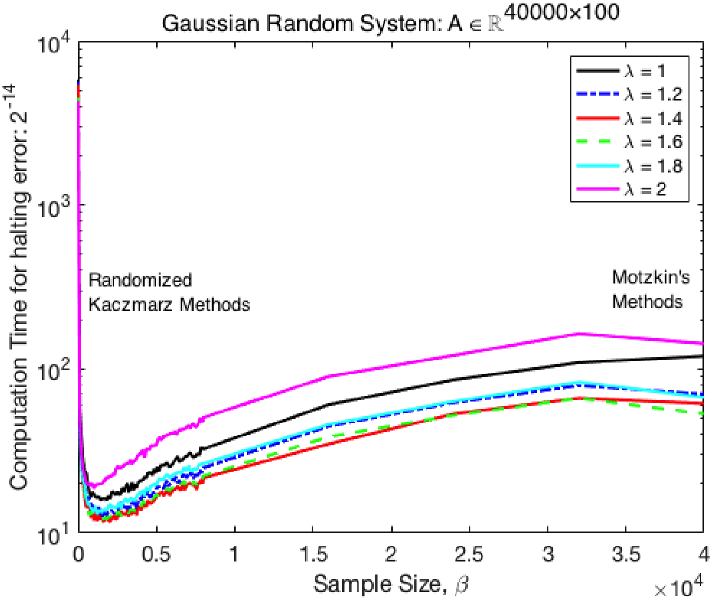
$$k > \frac{4\sigma - 4 - \log n + 2\log\left(\max_{j\in[m]}||a_j||\right)}{\log\left(\frac{mL_2^2}{mL_2^2 - 2\lambda + \lambda^2}\right)}.$$

If the system $Ax \leq b$ is feasible, the probability that the iterate x_k is not a certificate of feasibility is at most

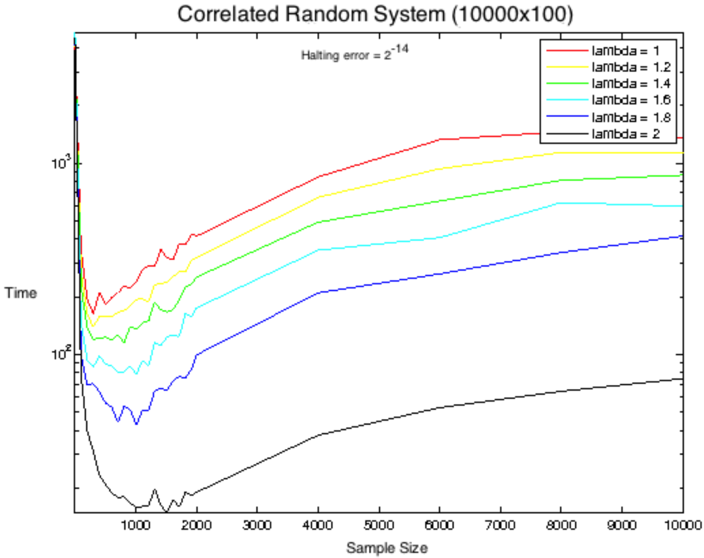
$$\frac{\max||a_j||\, 2^{2\sigma-2}}{n^{1/2}}\left(1 - \frac{2\lambda - \lambda^2}{mL_2^2}\right)^{k/2},$$

which decreases with k.

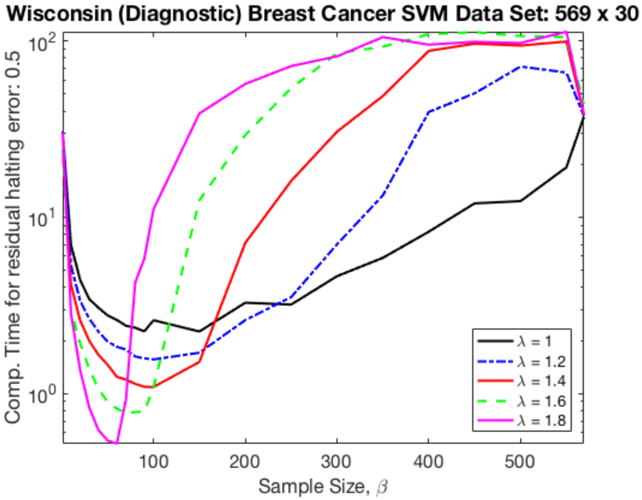
Experimental Results



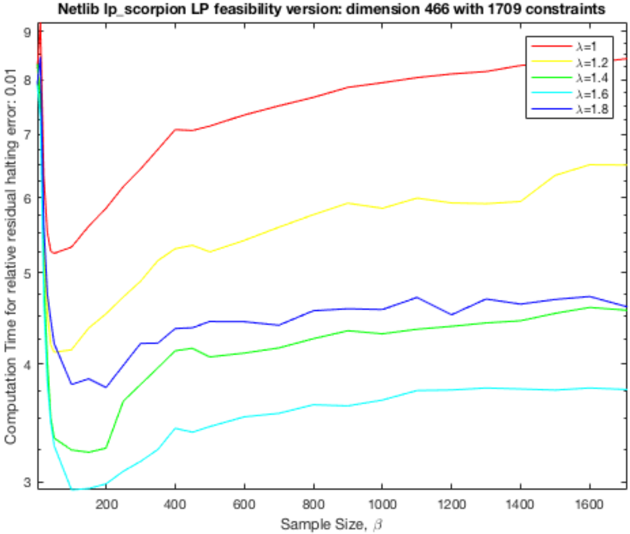
Experimental Results



Experimental Results



Experimental Results



Acknowledgements

Thanks to you for attending!

Are there any questions?

References I

- Kaczmarz, S. (1937).

  Angenaherte auflosung von systemen linearer gleichungen.

  Bull.Internat.Acad.Polon.Sci.Lettres A, pages 335–357.

- Leventhal, D. and Lewis, A. S. (2010).

  Randomized methods for linear constraints: convergence rates and conditioning.

  Math.Oper.Res., 35(3):641–654.

  65F10 (15A39 65K05 90C25); 2724068 (2012a:65083); Raimundo J. B. de Sampaio.

- Motzkin, T. S. and Schoenberg, I. J. (1954).

  The relaxation method for linear inequalities.

  Canadian J. Math., 6:393–404.

- Needell, D. (2010).

  Randomized kaczmarz solver for noisy linear systems.

  BIT, 50(2):395–403.

References II

- Needell, D., Sbrero, N., and Ward, R. (2013).

  Stochastic gradient descent and the randomized kaczmarz algorithm.

  submitted.

- Needell, D. and Tropp, J. A. (2013).

  Paved with good intentions: Analysis of a randomized block kaczmarz method.

  Linear Algebra Appl.

- Schrijver, A. (1986).

  Theory of linear and integer programming.

  Wiley-Interscience Series in Discrete Mathematics. John Wiley & Sons, Ltd., Chichester.

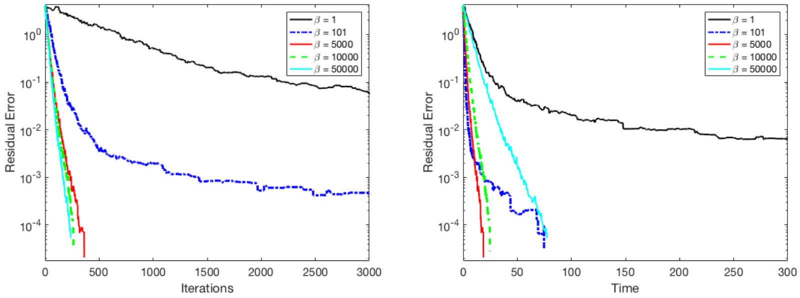
  A Wiley-Interscience Publication.

- Strohmer, T. and Vershynin, R. (2009).

  A randomized kaczmarz algorithm with exponential convergence.

  J. Fourier Anal. Appl., 15:262–278.

Iterations vs. Time



SKM on Gaussian random system, $A \in \mathbb{R}^{50000 \times 100}$

Heuristics for β Selection

In an iteration, the expected improvement is

$$d(x_j, P)^2 - d(x_{j+1}, P)^2 = \mathbb{E}\Big[||(A_{\tau_j} x_j - b_{\tau_j})^+||_\infty^2\Big].$$

Heuristics for β Selection

In an iteration, the expected improvement is

$$d(x_j, P)^2 - d(x_{j+1}, P)^2 = \mathbb{E}\Big[||(A_{\tau_j} x_j - b_{\tau_j})^+||_\infty^2\Big].$$

The worst case will be when the $m - s$ non-zero entries of the residual all are the same, assume they are 1.

Heuristics for β Selection

In an iteration, the expected improvement is

$$d(x_j, P)^2 - d(x_{j+1}, P)^2 = \mathbb{E}\Big[||(A_{\tau_j} x_j - b_{\tau_j})^+||_\infty^2\Big].$$

The worst case will be when the $m - s$ non-zero entries of the residual all are the same, assume they are 1.

We consider this case and model the computation in a fixed iteration as the overhead cost, C, and a factor $cn\beta$ for checking the feasibility of β constraints.

Heuristics for β Selection

Note that

$$\mathbb{E}\left[||(A_{\tau_j}x_j - b_{\tau_j})^+||_\infty^2\right] = \begin{cases} 1 - \frac{\binom{s}{\beta}}{\binom{m}{\beta}} \approx 1 - \left(\frac{s}{m}\right)^\beta & \text{if } \beta \leq s \\ 1 & \text{if } \beta > s \end{cases}$$

Heuristics for β Selection

Note that

$$\mathbb{E}\left[||(A_{\tau_j}x_j - b_{\tau_j})^+||_\infty^2\right] = \begin{cases} 1 - \frac{\binom{s}{\beta}}{\binom{m}{\beta}} \approx 1 - \left(\frac{s}{m}\right)^\beta & \text{if } \beta \le s \\ 1 & \text{if } \beta > s \end{cases}$$

Thus, we look for β that maximizes the improvement per unit of computation time:

$$\text{gain}(\beta) := \frac{\mathbb{E}\left[||(A_{\tau_j}x_j - b_{\tau_j})^+||_\infty^2\right]}{C + cn\beta} \approx \frac{1 - \left(\frac{s}{m}\right)^\beta}{C + cn\beta}.$$

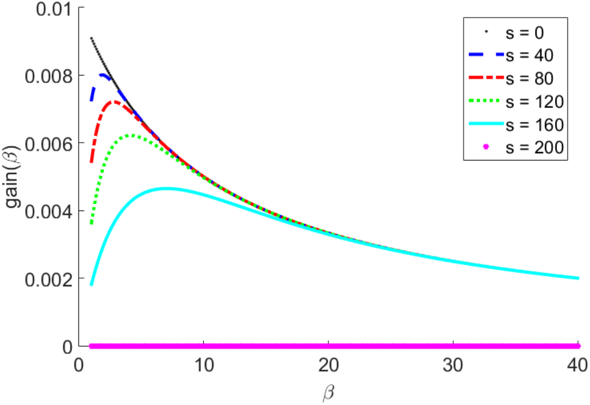


Figure : The quantity gain(β) as a function of β for various numbers of satisfied constraints s. Here we set $m = 200$, $n = 10$, $c = 1$ and $C = 100$. Optimal values of β maximize the gain function.