Randomized Kaczmarz Algorithms: Exact MSE Analysis and Optimal Sampling Probabilities

Yue M. Lu

Signals, Information, and Networks Group (SING)
Harvard University

Joint work with Ameya Agaskar and Chuang Wang

April 17, 2015

Support: The Department of Air Force under Contract #FA8721T05TCT0002 and NSF CCF-1319140

Large Linear Systems

$$\underbrace{\boldsymbol{y}}_{\text{observation}} = \underbrace{\boldsymbol{A}}_{\text{system matrix}} \times \boldsymbol{x} + \underbrace{\boldsymbol{z}}_{\text{noise}}$$

$$\begin{bmatrix} y_1 \\ y_2 \\ \cdot \\ \cdot \\ \cdot \\ y_m \end{bmatrix} = \begin{bmatrix} -\boldsymbol{a_1}^T- \\ -\boldsymbol{a_2}^T- \\ \cdot \\ \cdot \\ \cdot \\ -\boldsymbol{a_m}^T- \end{bmatrix} \times \begin{bmatrix} x_1 \\ \cdot \\ \cdot \\ \cdot \\ x_n \end{bmatrix} + \begin{bmatrix} z_1 \\ z_2 \\ \cdot \\ \cdot \\ \cdot \\ z_m \end{bmatrix}$$

observation

system matrix

noise

Large Linear Systems

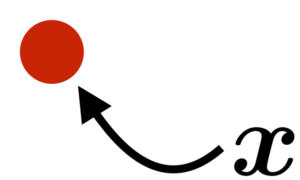
$$\begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_m \end{bmatrix} = \begin{bmatrix} -\boldsymbol{a_1}^T- \\ -\boldsymbol{a_2}^T- \\ \vdots \\ -\boldsymbol{a_m}^T- \end{bmatrix} \times \begin{bmatrix} x_1 \\ \vdots \\ x_n \end{bmatrix} + \begin{bmatrix} z_1 \\ z_2 \\ \vdots \\ z_m \end{bmatrix}$$

New considerations:

- Massive sizes
- Streaming data
- Distributed storage
- Parallel computing platform

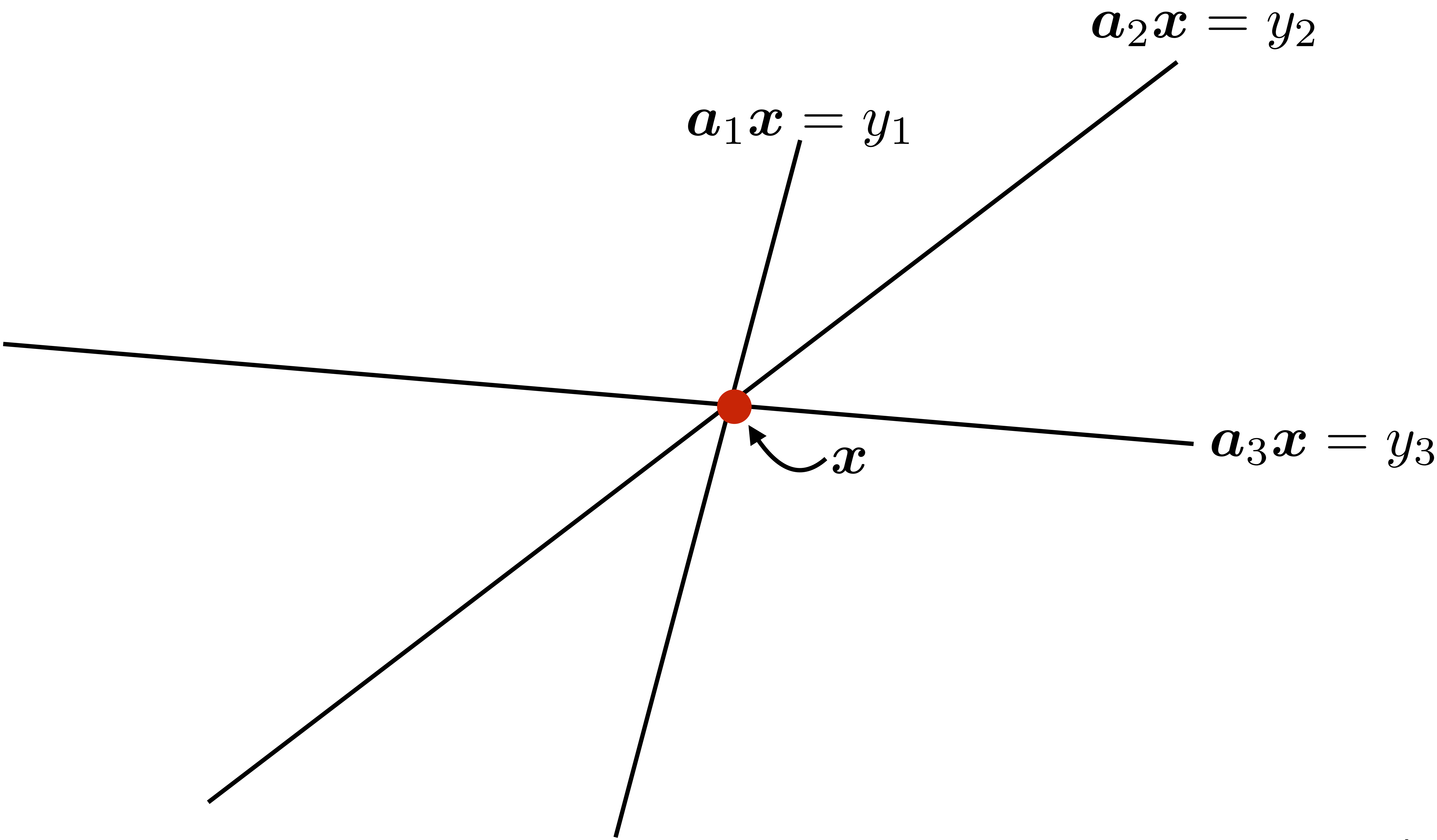
$\Longrightarrow$ *Iterative greedy methods*

- Noiseless case: $\boldsymbol{y} = \boldsymbol{A}\boldsymbol{x}$ encodes a system of m equations.



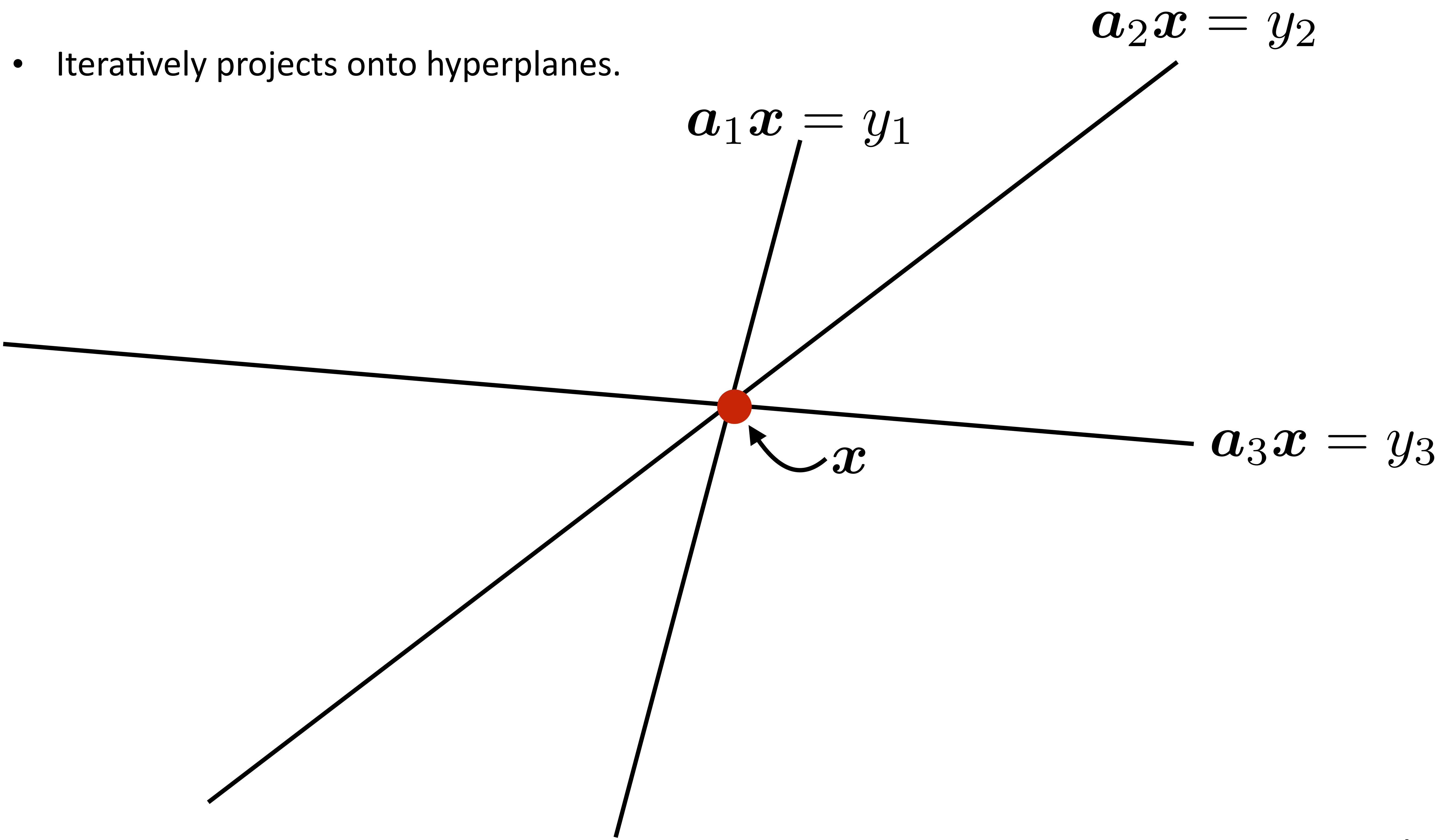
Kaczmarz Algorithm — Iterative Projections

- Noiseless case: $\boldsymbol{y} = \boldsymbol{A}\boldsymbol{x}$ encodes a system of m equations.



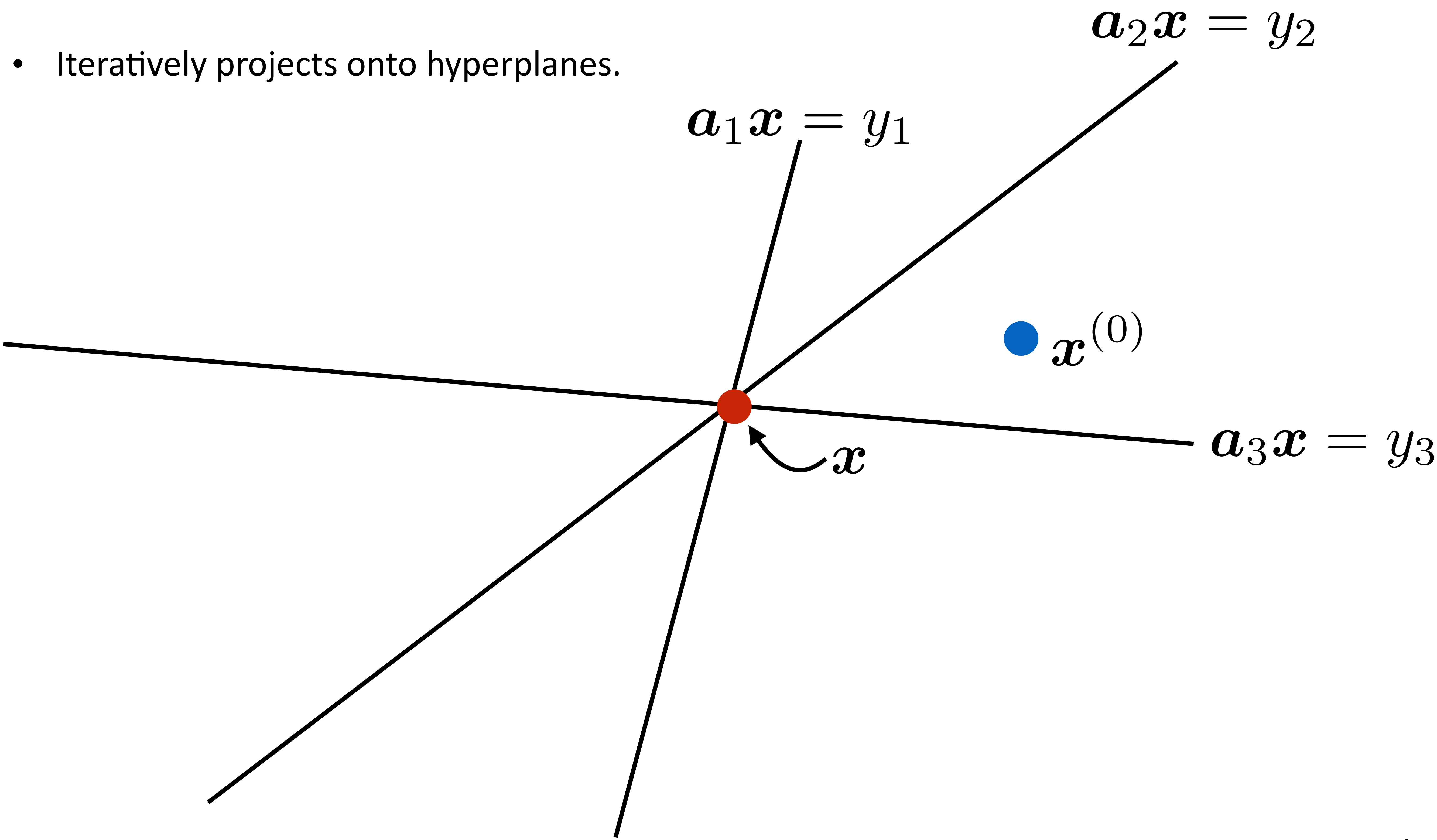
Kaczmarz Algorithm — Iterative Projections

- Noiseless case: $\boldsymbol{y} = \boldsymbol{A}\boldsymbol{x}$ encodes a system of m equations.
- Iteratively projects onto hyperplanes.



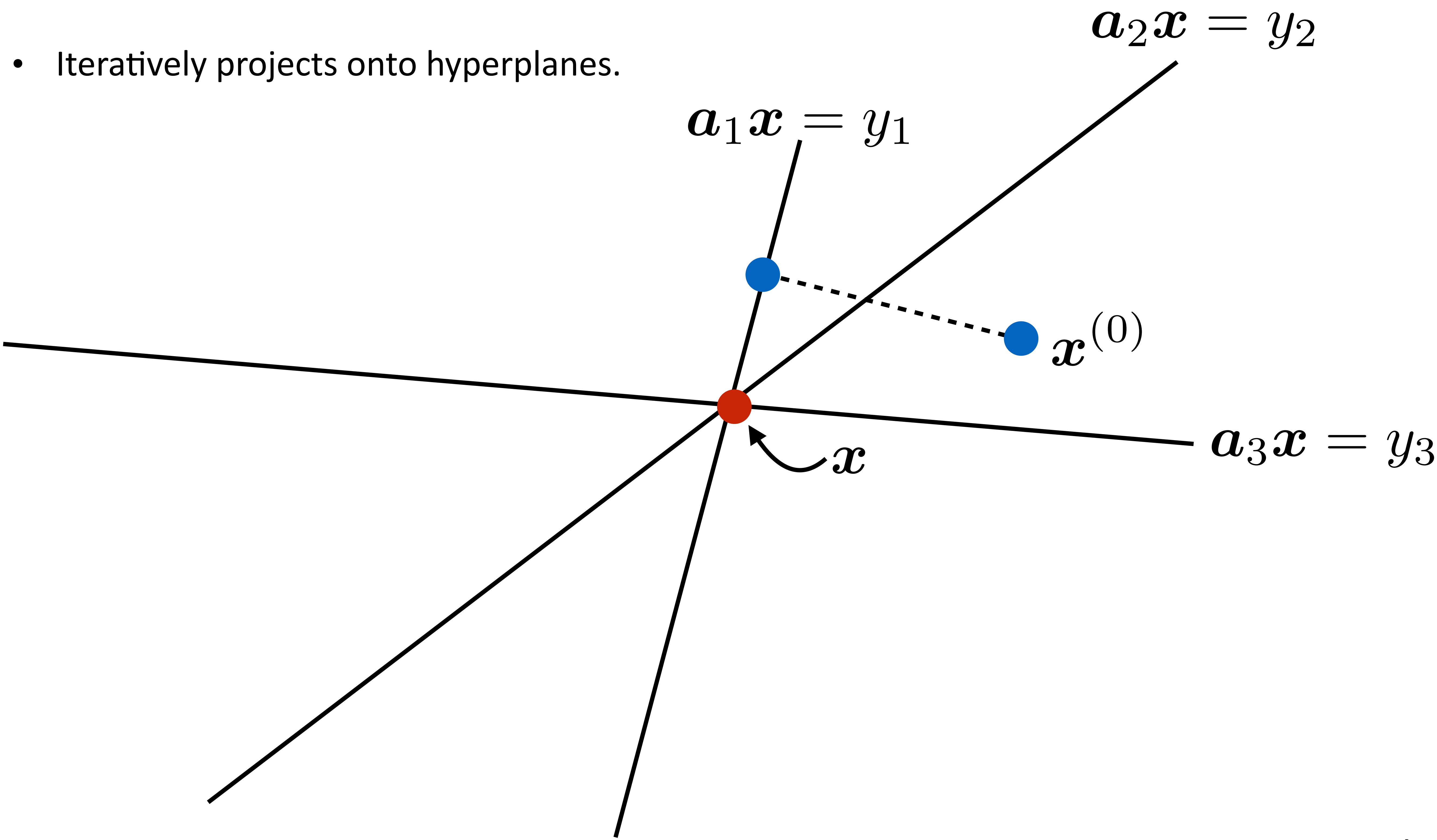
Kaczmarz Algorithm — Iterative Projections

- Noiseless case: $\boldsymbol{y} = \boldsymbol{A}\boldsymbol{x}$ encodes a system of m equations.
- Iteratively projects onto hyperplanes.



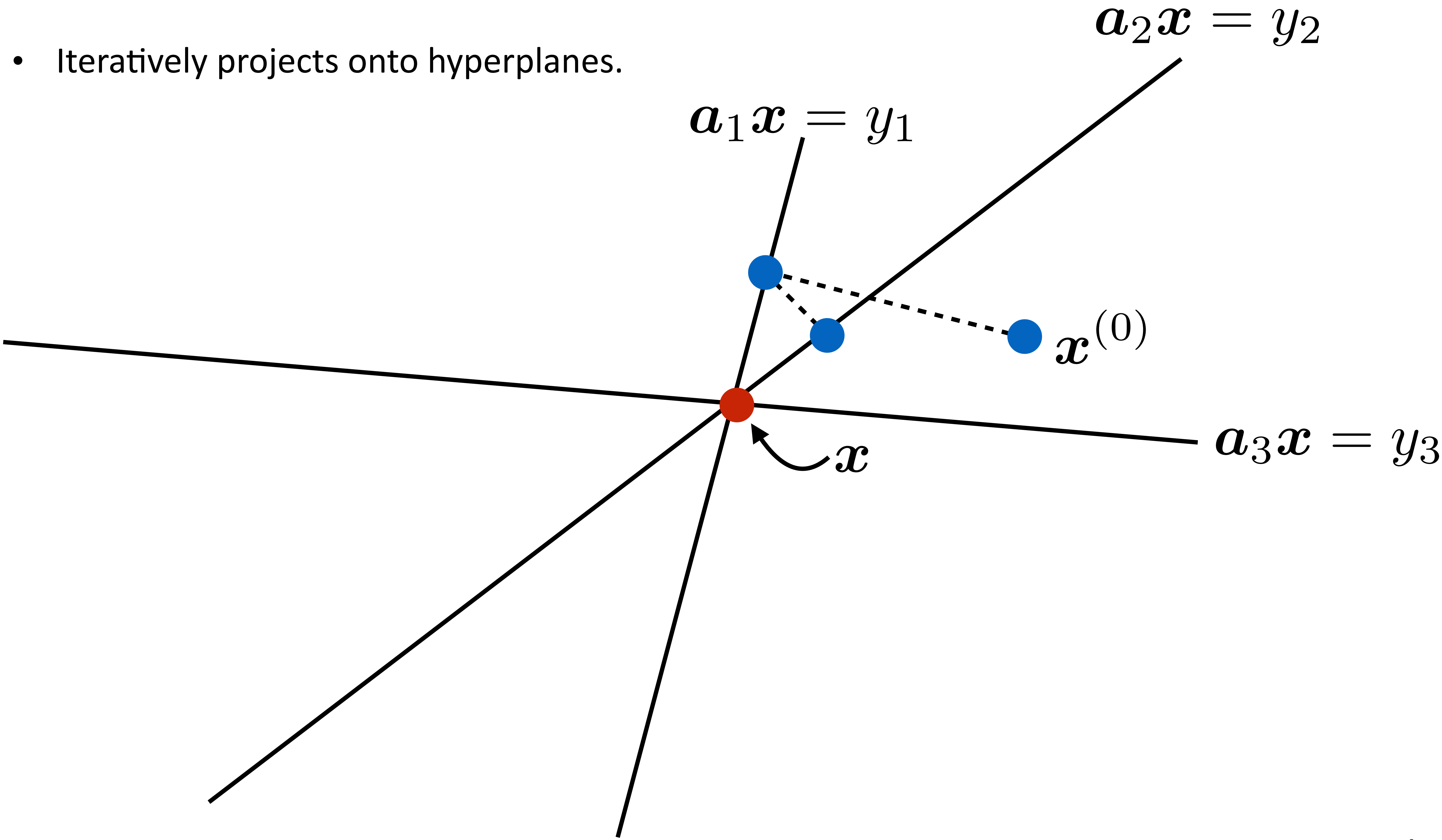
Kaczmarz Algorithm — Iterative Projections

- Noiseless case: $\boldsymbol{y} = \boldsymbol{A}\boldsymbol{x}$ encodes a system of m equations.
- Iteratively projects onto hyperplanes.



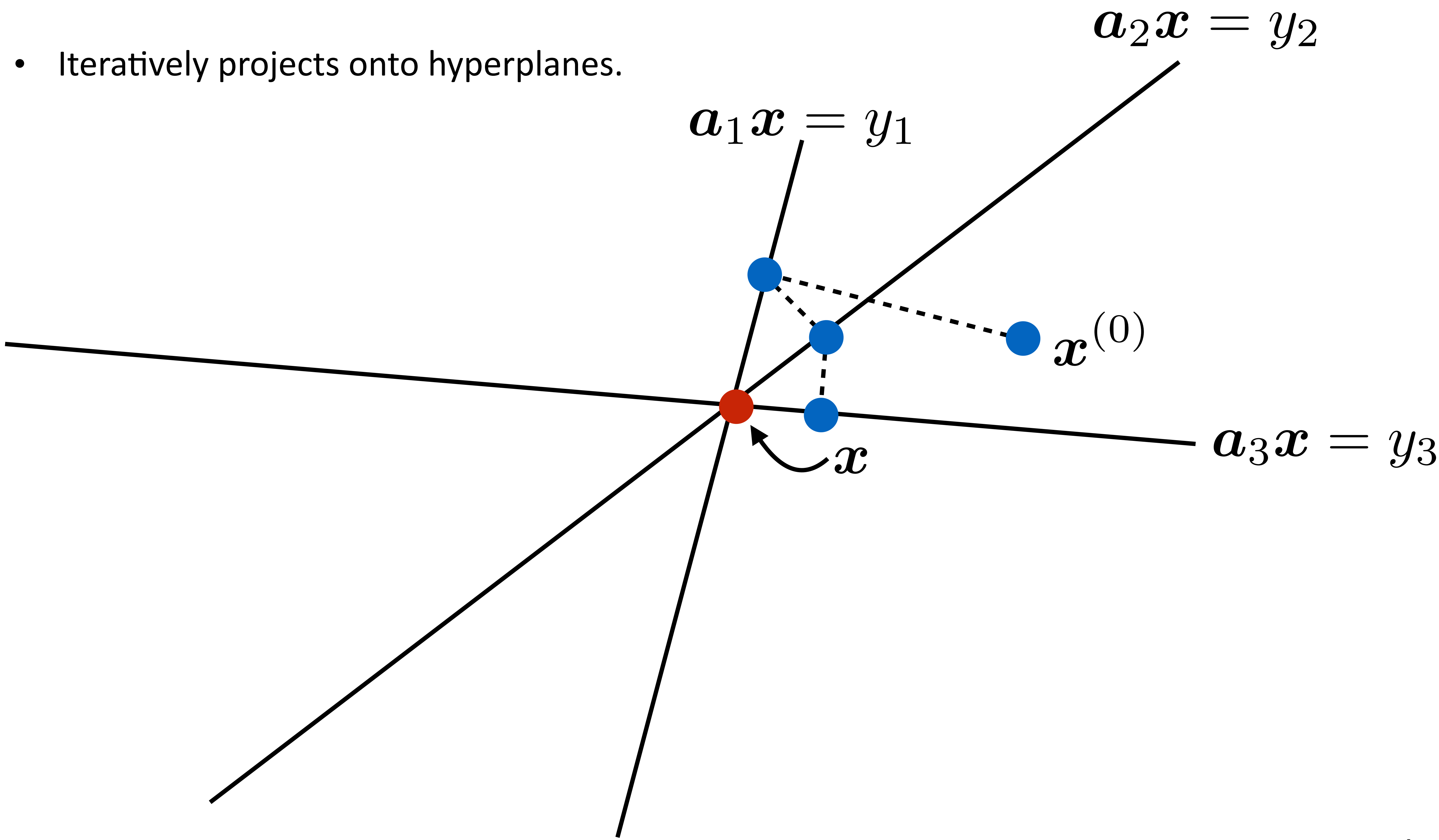
Kaczmarz Algorithm — Iterative Projections

- Noiseless case: $\boldsymbol{y} = \boldsymbol{A}\boldsymbol{x}$ encodes a system of m equations.
- Iteratively projects onto hyperplanes.



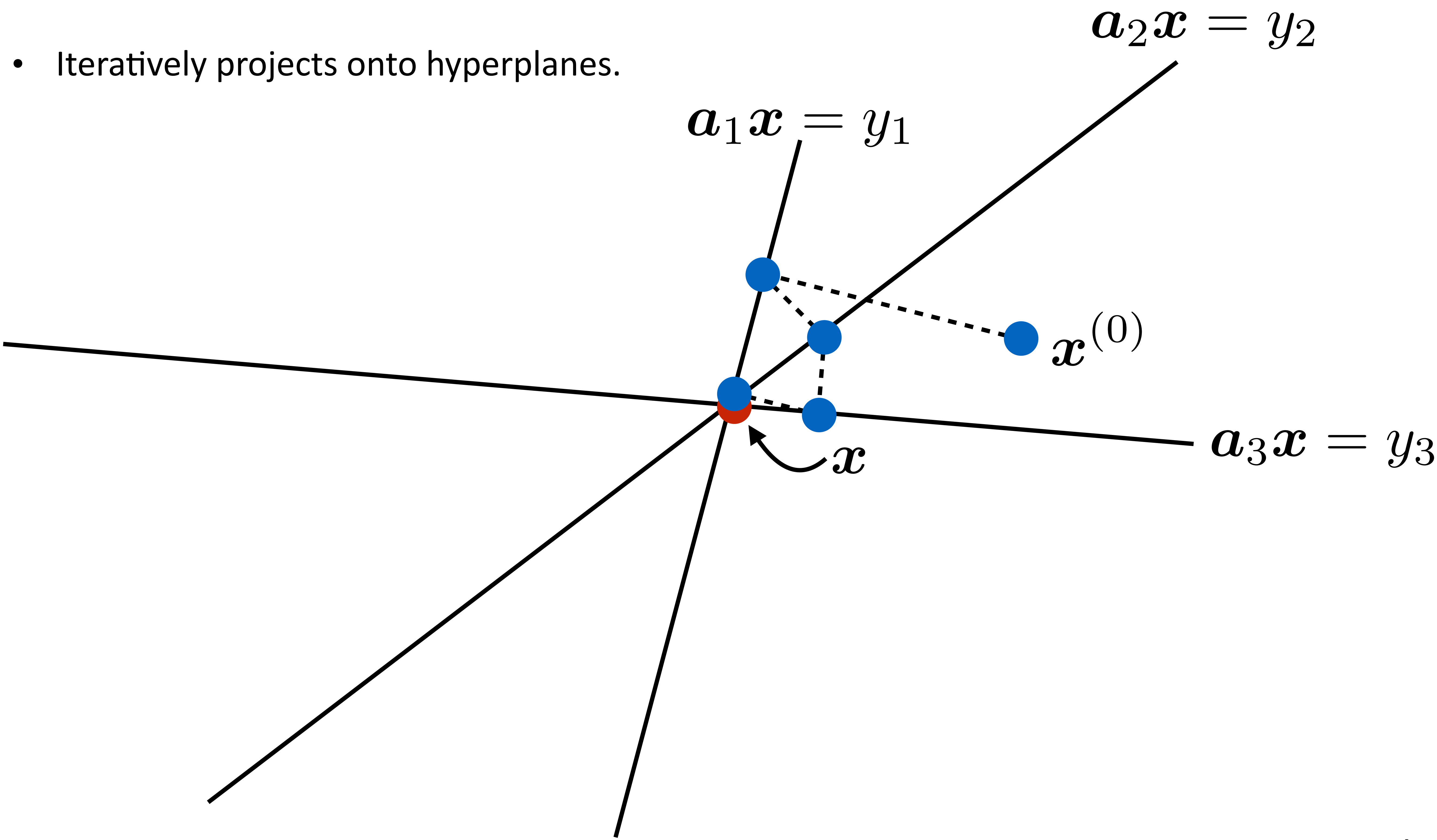
Kaczmarz Algorithm — Iterative Projections

- Noiseless case: $\boldsymbol{y} = \boldsymbol{A}\boldsymbol{x}$ encodes a system of m equations.
- Iteratively projects onto hyperplanes.



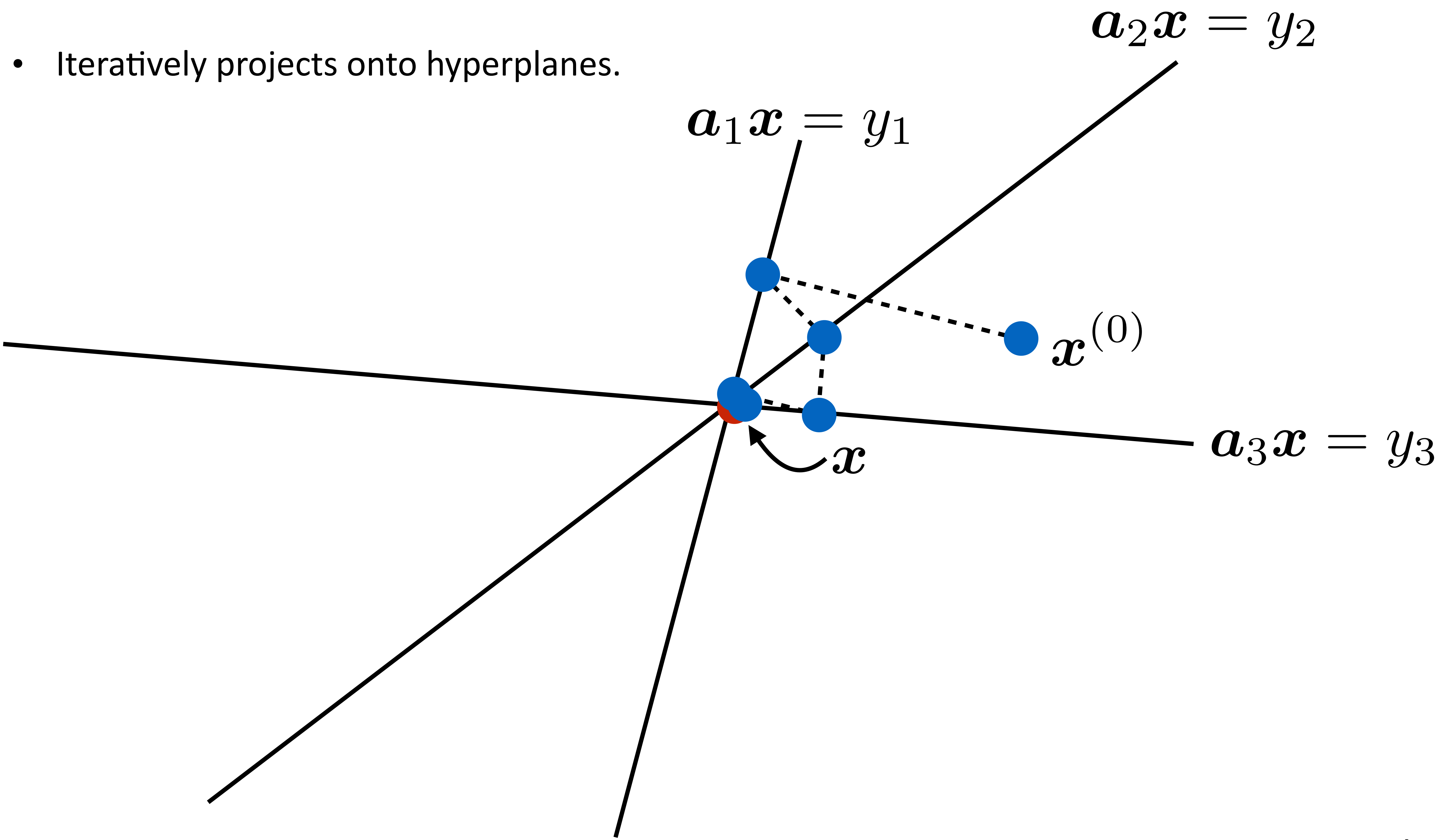
Kaczmarz Algorithm — Iterative Projections

- Noiseless case: $\boldsymbol{y} = \boldsymbol{A}\boldsymbol{x}$ encodes a system of m equations.
- Iteratively projects onto hyperplanes.



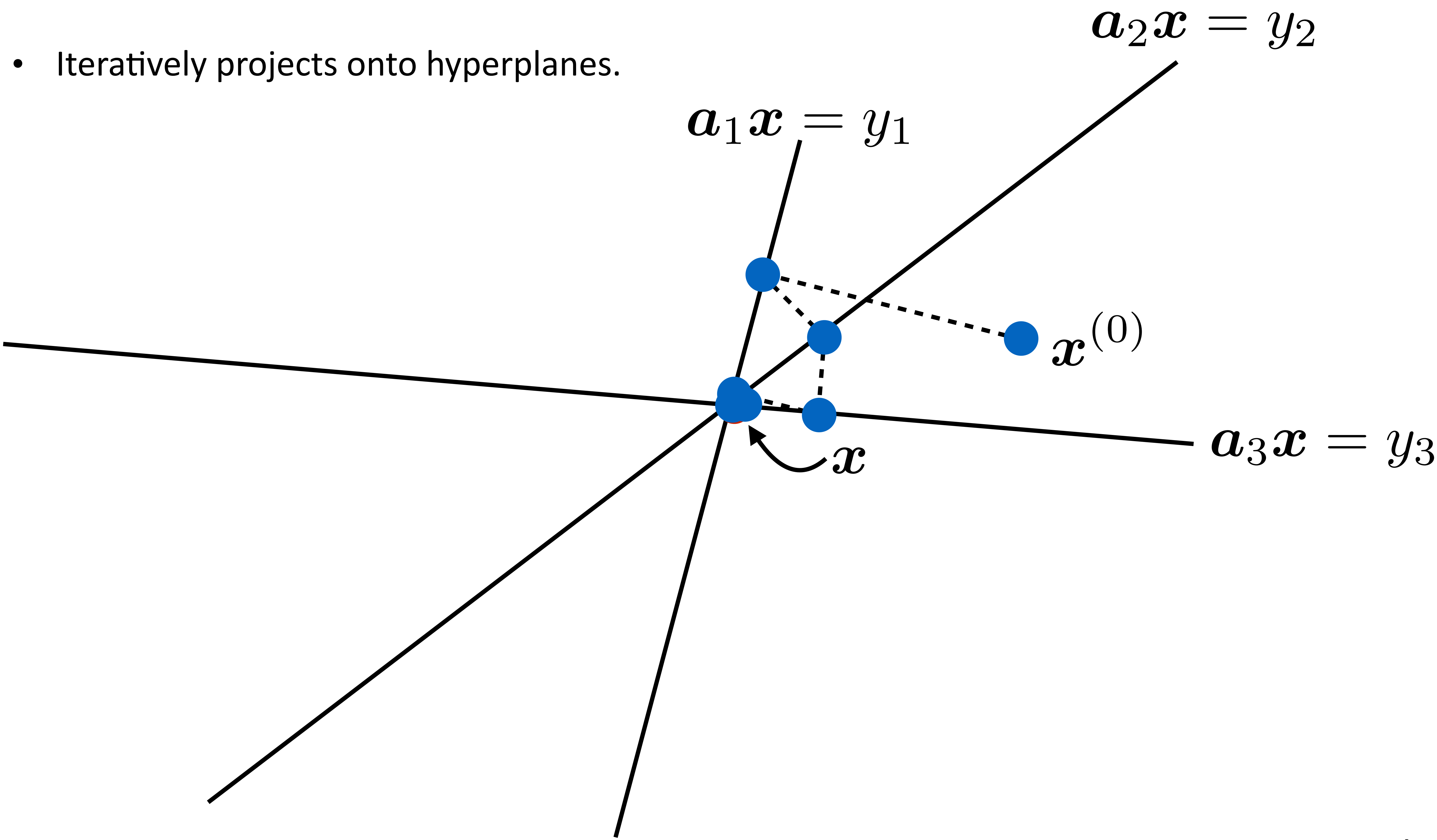
Kaczmarz Algorithm — Iterative Projections

- Noiseless case: $\boldsymbol{y} = \boldsymbol{A}\boldsymbol{x}$ encodes a system of m equations.
- Iteratively projects onto hyperplanes.



Kaczmarz Algorithm — Iterative Projections

- Noiseless case: $\boldsymbol{y} = \boldsymbol{A}\boldsymbol{x}$ encodes a system of m equations.
- Iteratively projects onto hyperplanes.



Kaczmarz Algorithm

- Iterative algorithm introduced by S. Kaczmarz (1937)
- Also known as algebraic reconstruction technique (ART)
- Special case of projection onto convex sets (POCS)

Pseudocode

Initialize arbitrary $\boldsymbol{x}^{(0)}$.

For $k = 1$ to N_{iter}:

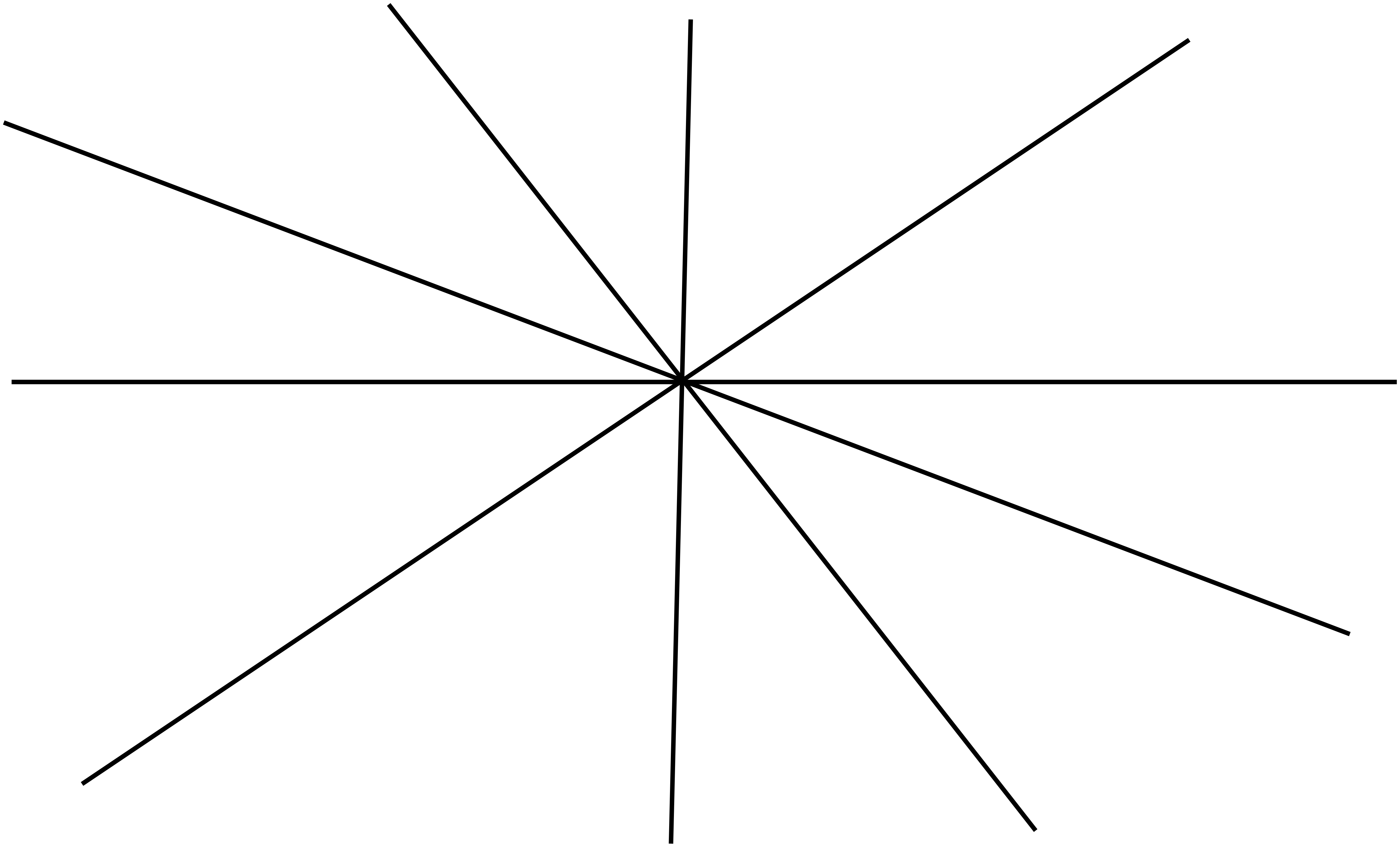
$$r \leftarrow (k \bmod m) + 1$$ *select the next row*

$$\boldsymbol{x}^{(k)} \leftarrow \boldsymbol{x}^{(k-1)} + \frac{y_r - \boldsymbol{a}_r^T \boldsymbol{x}^{(k-1)}}{||\boldsymbol{a}_r||^2} \boldsymbol{a}_r$$ *projection*

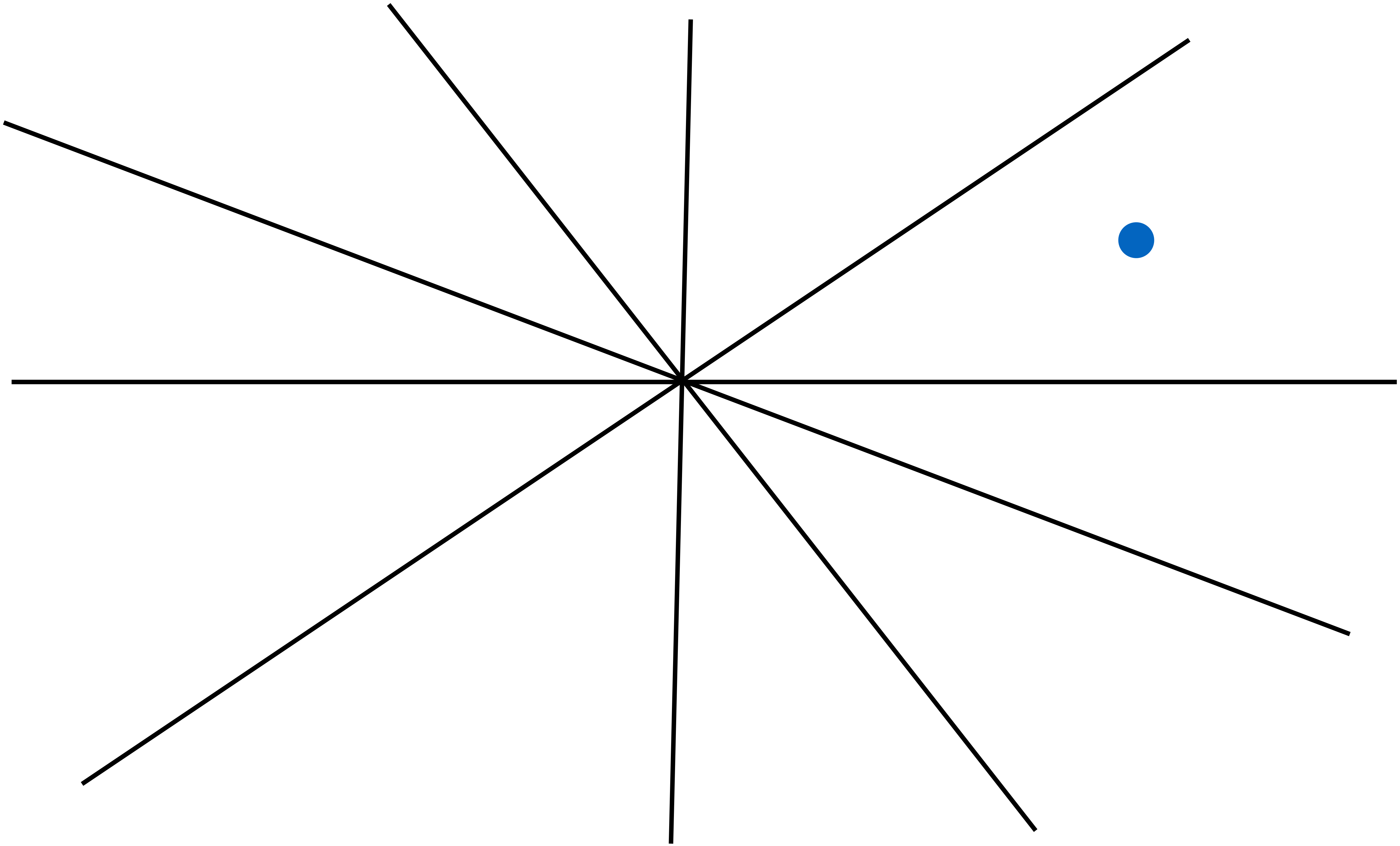
$$\widehat{\boldsymbol{x}} \leftarrow \boldsymbol{x}^{(N_{\mathrm{iter}})}$$

- Can be extended to find least squares estimate from noisy measurements (Zouzias & Freris 2013)

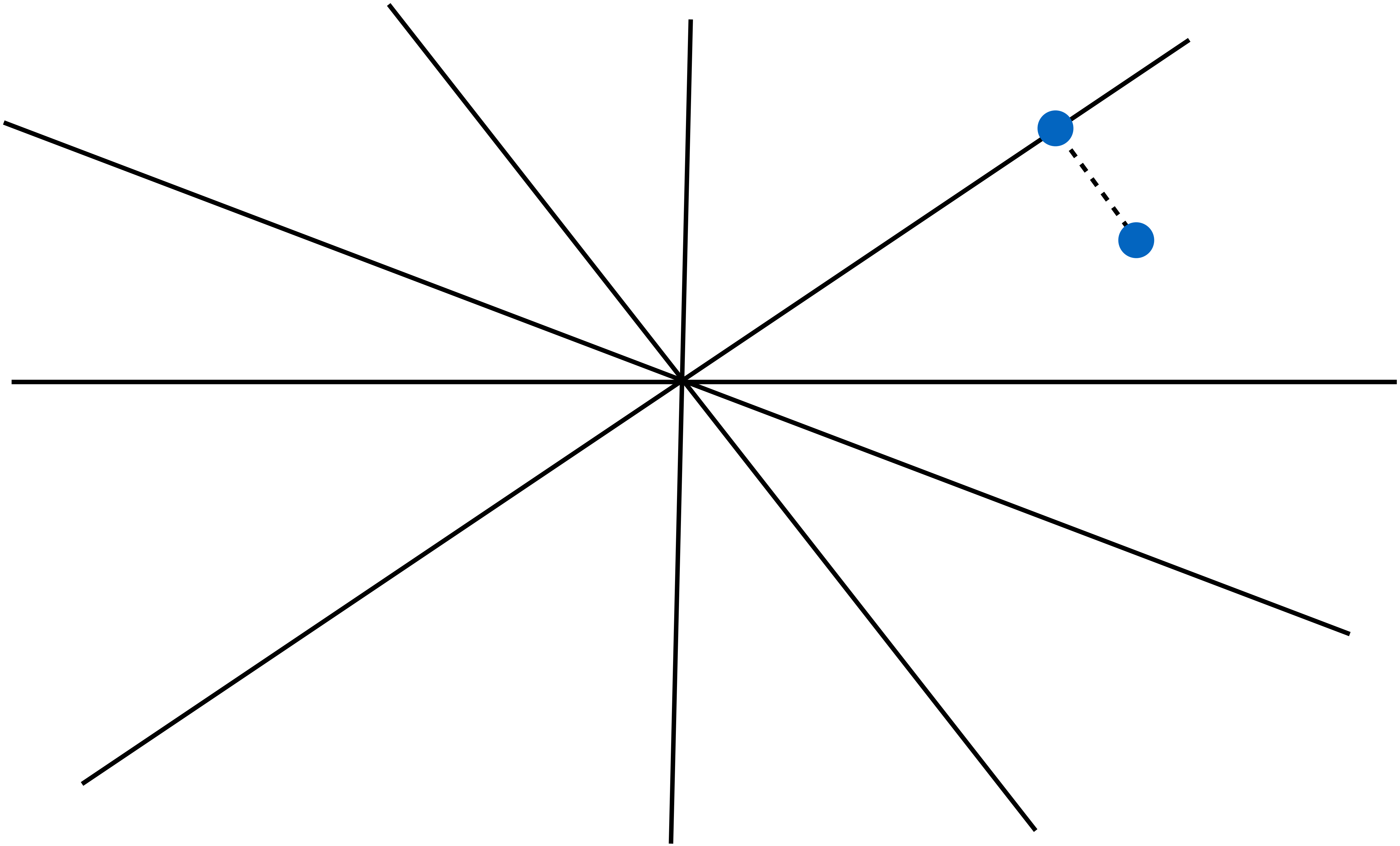
Order matters!

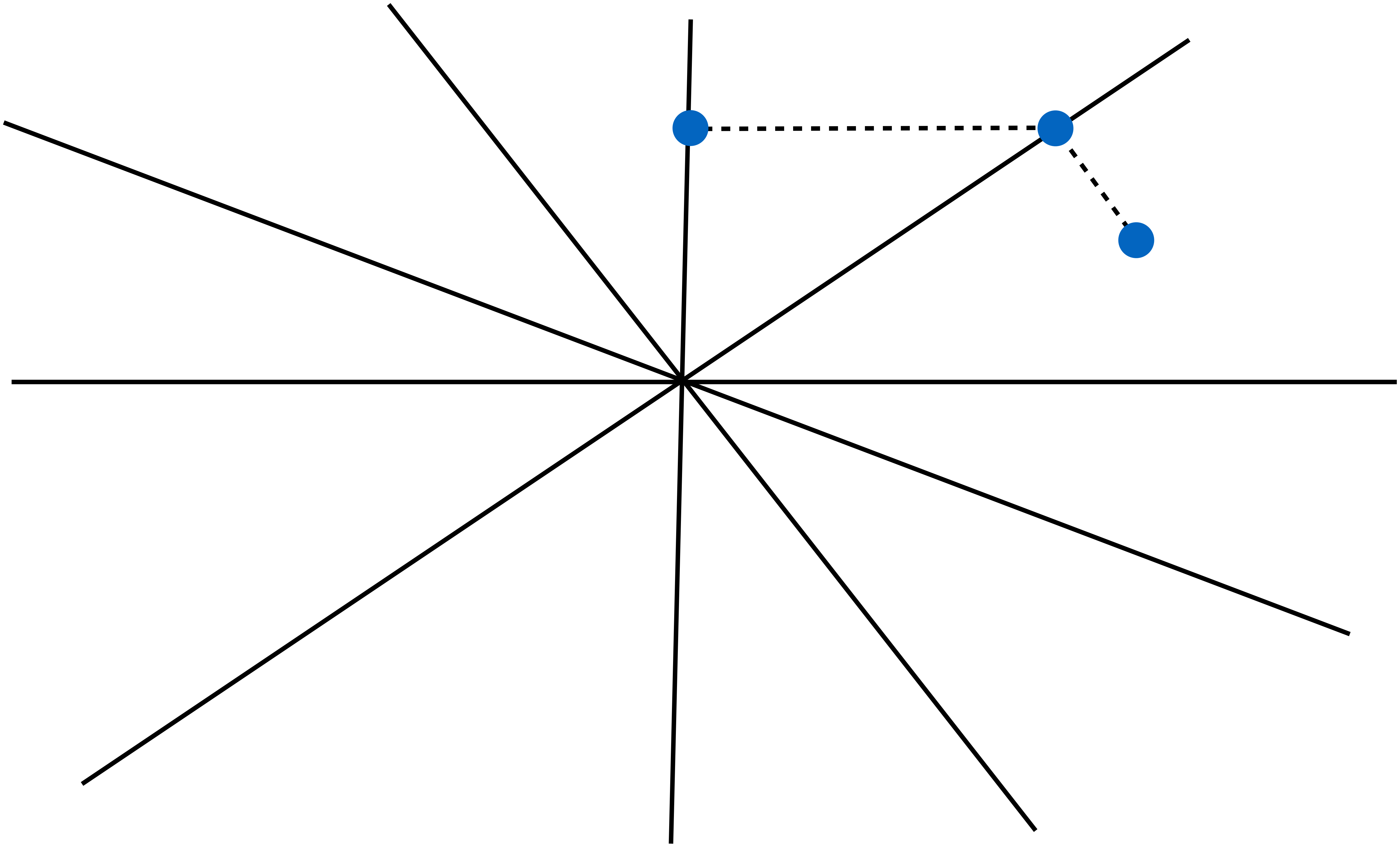


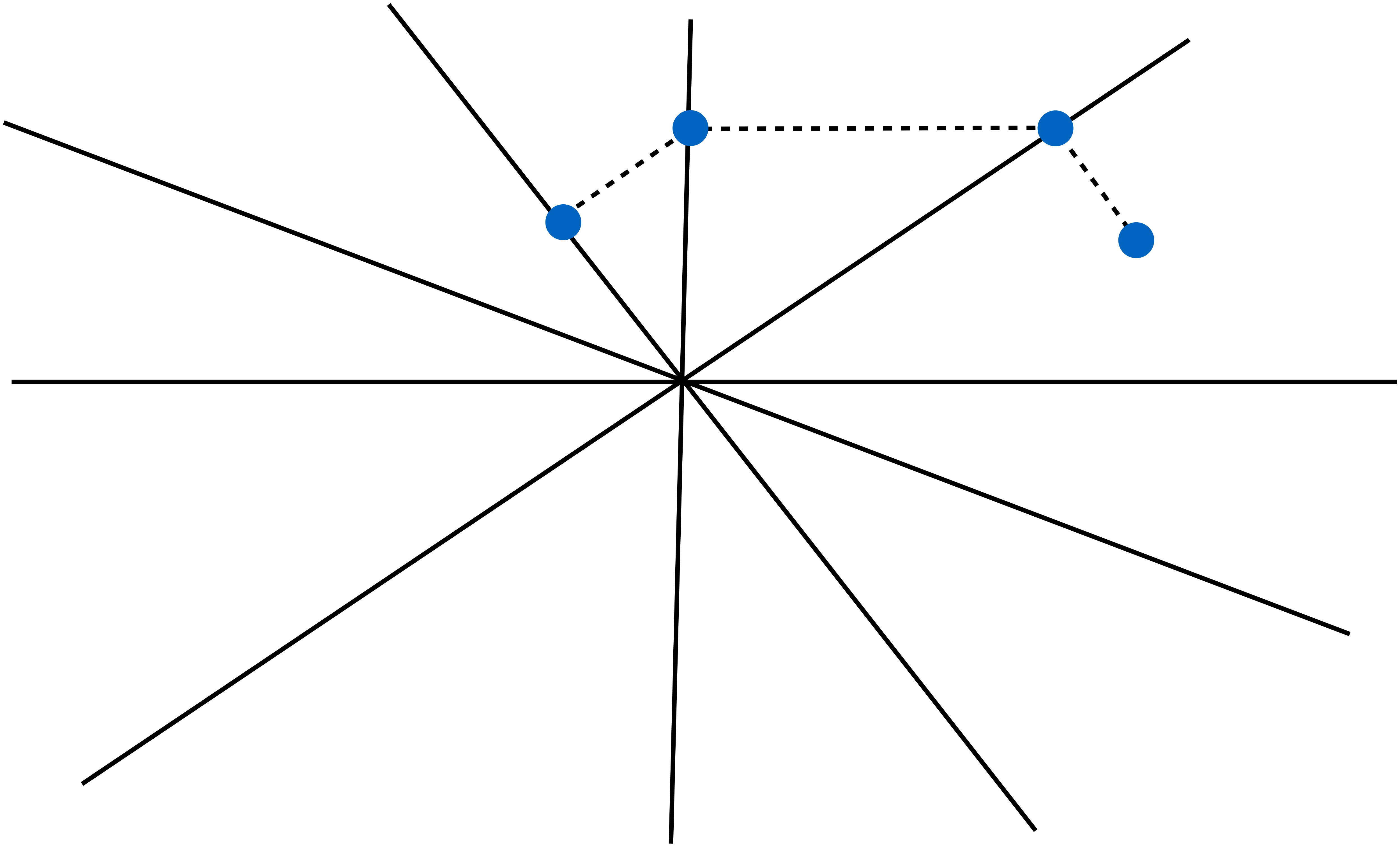
Order matters!

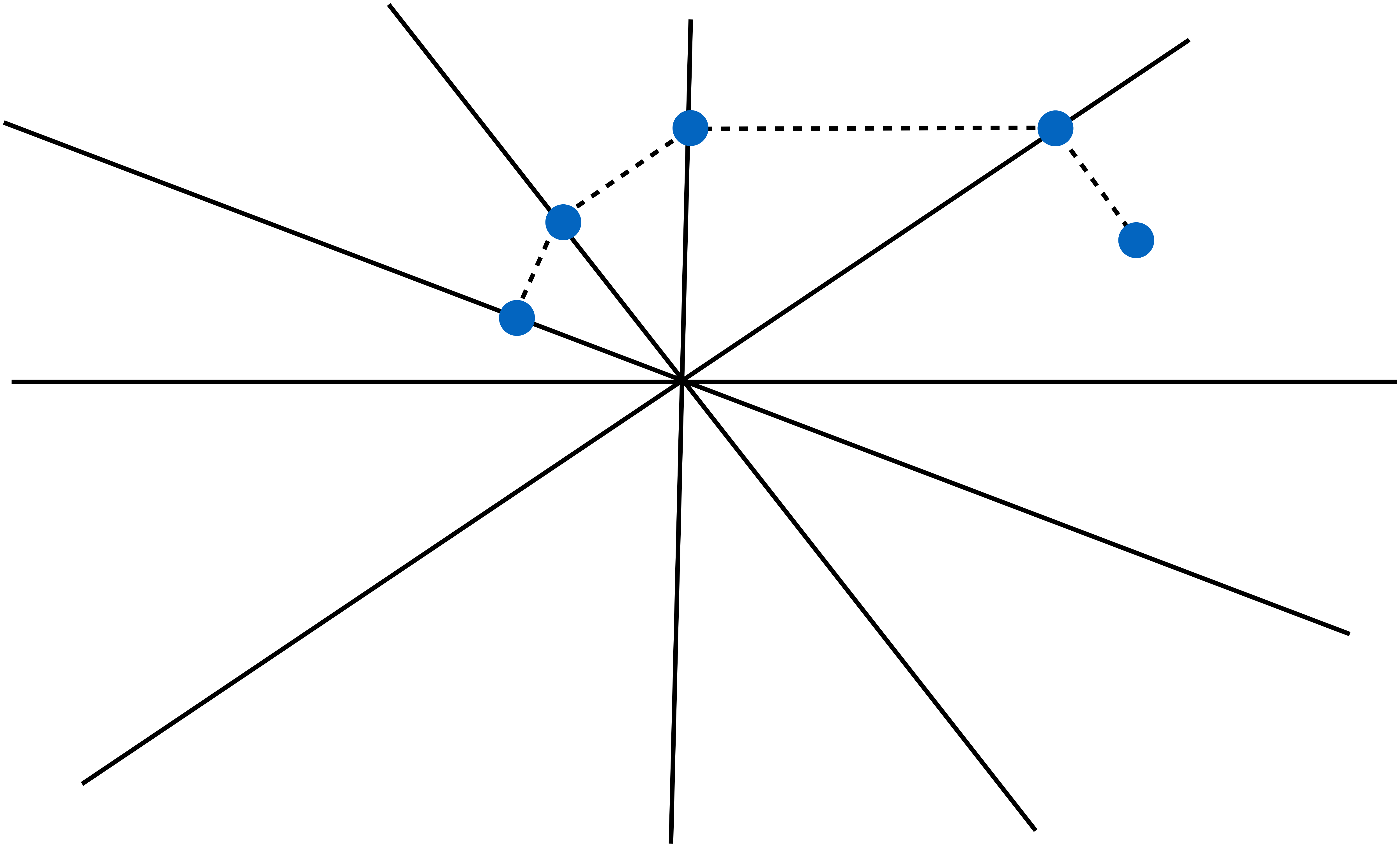


Order matters!

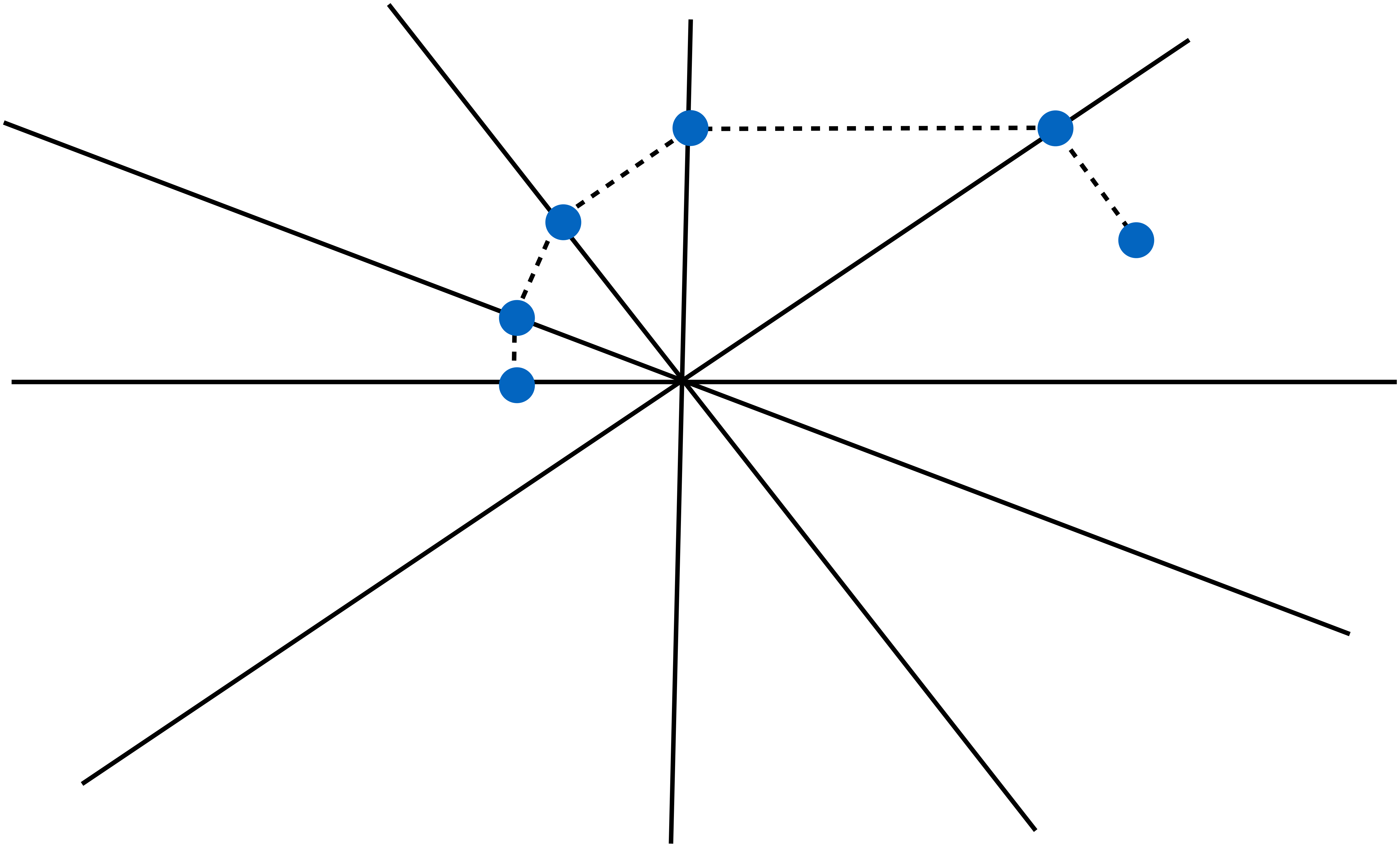




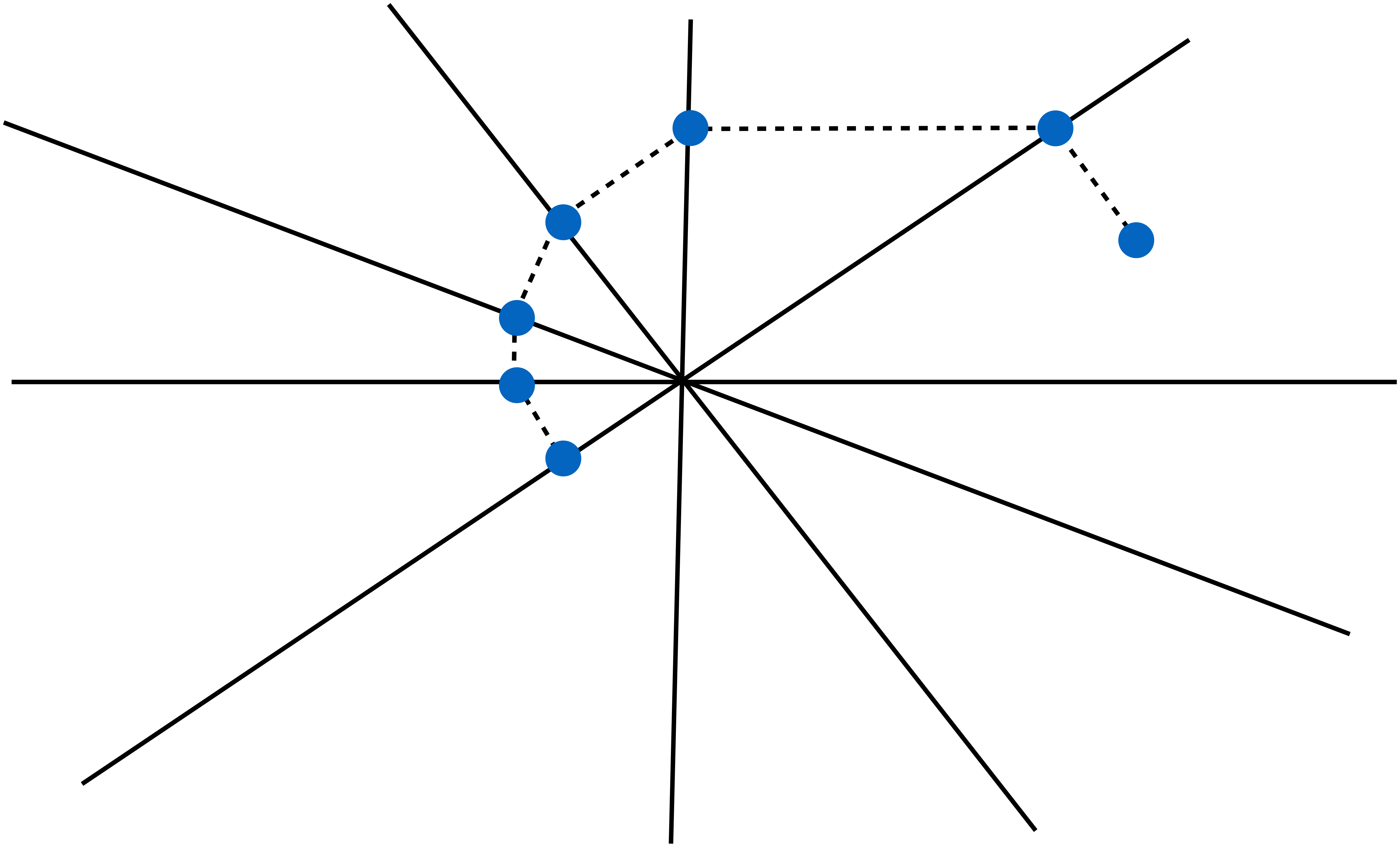




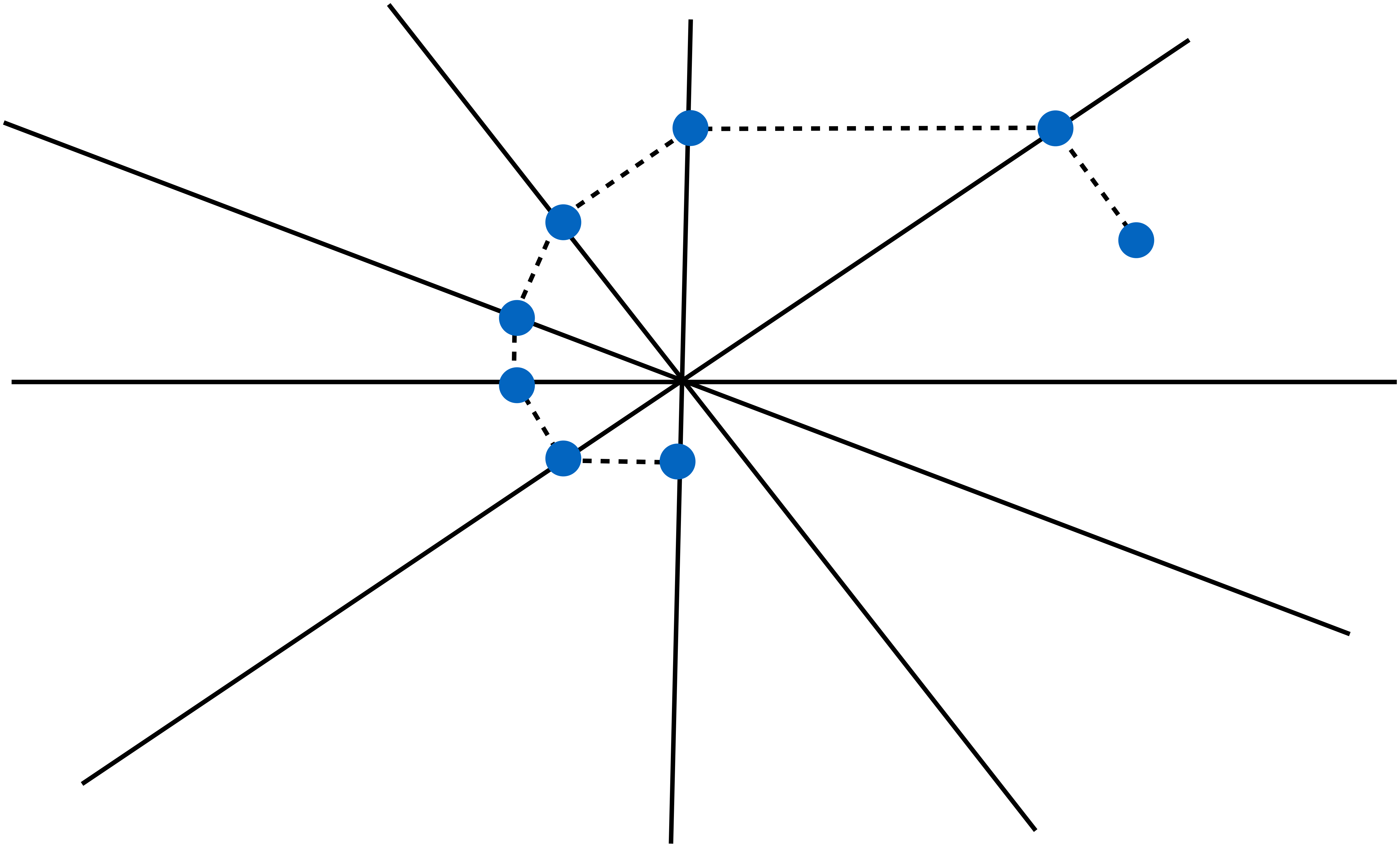
Order matters!



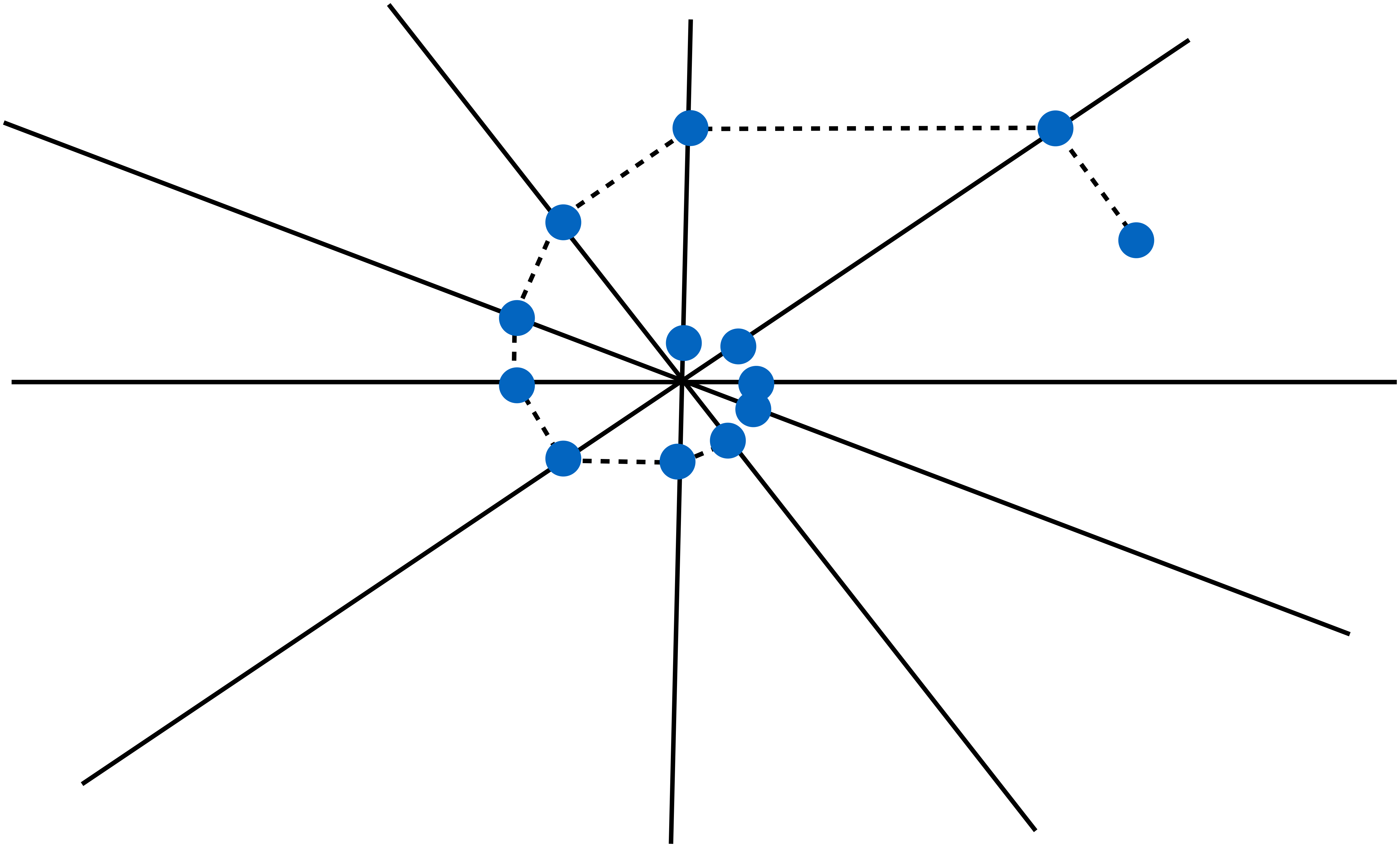
Order matters!



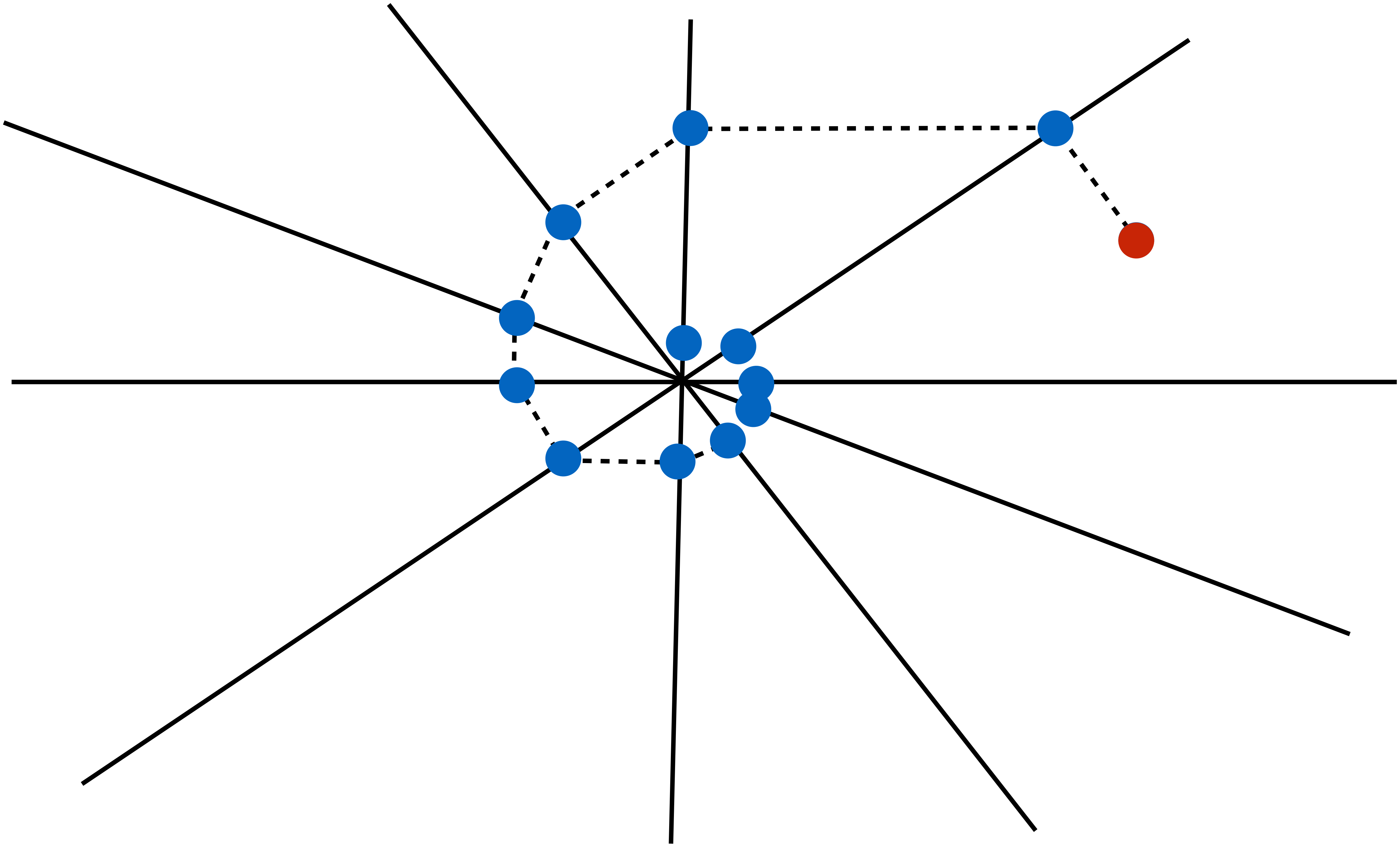
Order matters!



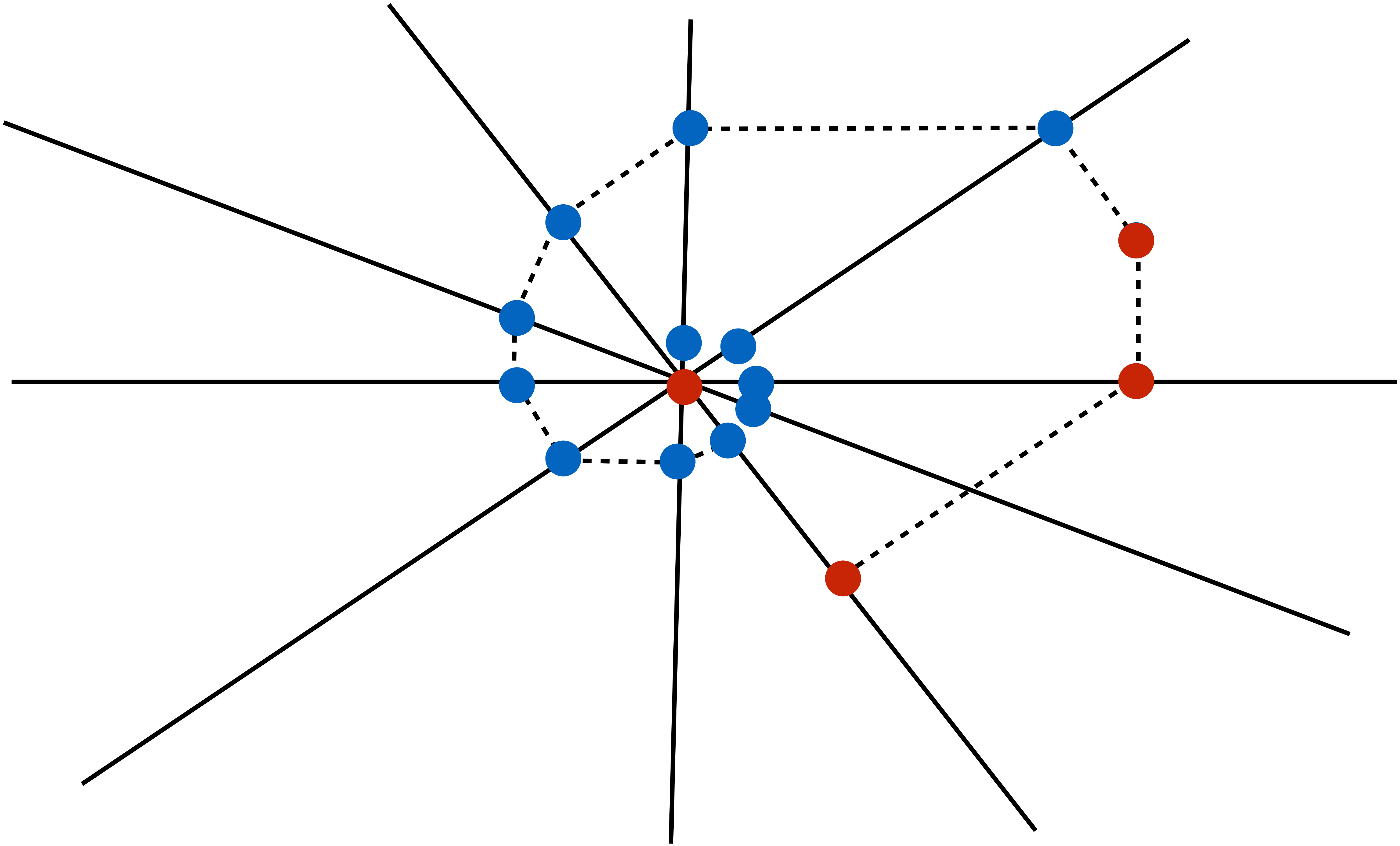
Order matters!



Order matters!



Order matters!




IF YOU USE THE WRONG ROW ORDER
INSTRUCTO
YOU'RE GONNA HAVE A BAD TIME
quickmeme.com

Solution: Randomized Kaczmarz

- Strohmer and Vershynin (2009) proposed randomizing the order:

Choose row $\boldsymbol{a}_i$ with probability proportional to $||\boldsymbol{a}_i||^2$.

- Guarantees exponential convergence:

$$\frac{\mathbb{E}||\boldsymbol{x}^{(N)} - \boldsymbol{x}||^2}{||\boldsymbol{x}^{(0)} - \boldsymbol{x}||^2} \leq (1 - \underbrace{\kappa(A)}_{||A||_F ||A^{-1}||_2}{}^{-2})^N$$

- Works for arbitrary probabilities by preconditioning, so we assume row i chosen with probability p_i.

Solution: Randomized Kaczmarz

- Strohmer and Vershynin (2009) proposed randomizing the order:

Choose row $\boldsymbol{a}_i$ with probability proportional to $||\boldsymbol{a}_i||^2$.

- Guarantees exponential convergence:

$$\frac{\mathbb{E}||\boldsymbol{x}^{(N)} - \boldsymbol{x}||^2}{||\boldsymbol{x}^{(0)} - \boldsymbol{x}||^2} \leq (1 - \underbrace{\kappa(A)}_{||A||_F||A^{-1}||_2}{}^{-2})^N$$

- Works for arbitrary probabilities by preconditioning, so we assume row i chosen with probability p_i.

Related work

- Error bounds for ***inconsistent systems*** (Needell 2012)
- ***Almost-sure*** convergence (Chen & Powell 2012)
- Extension to find ***least-square solution*** in noise (Zouzias & Freris 2013)
- ***Block*** Kaczmarz (Needell & Tropp 2014)

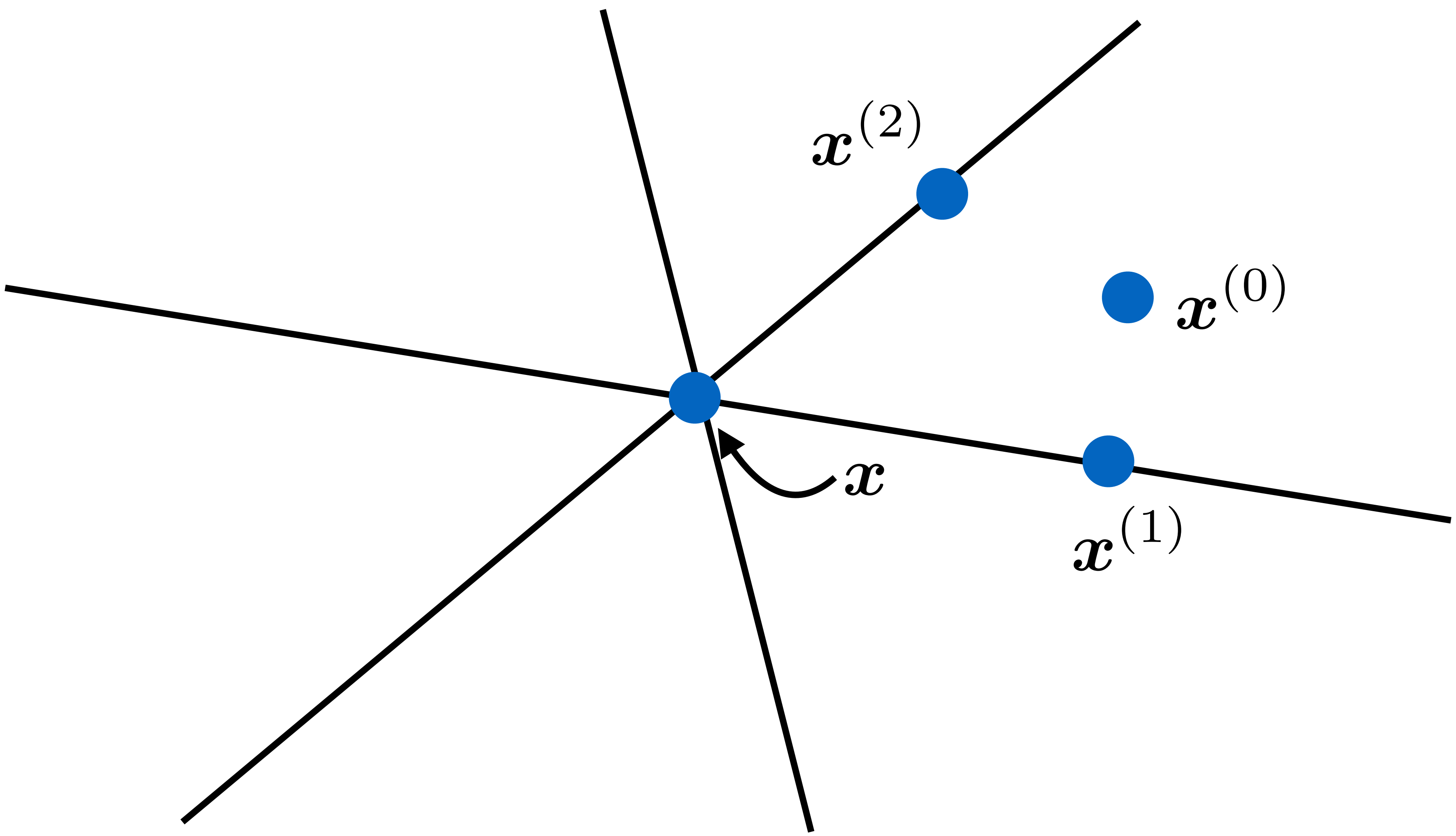
This talk

1. Exact MSE formula and decay rate

2. Optimization of row selection probabilities

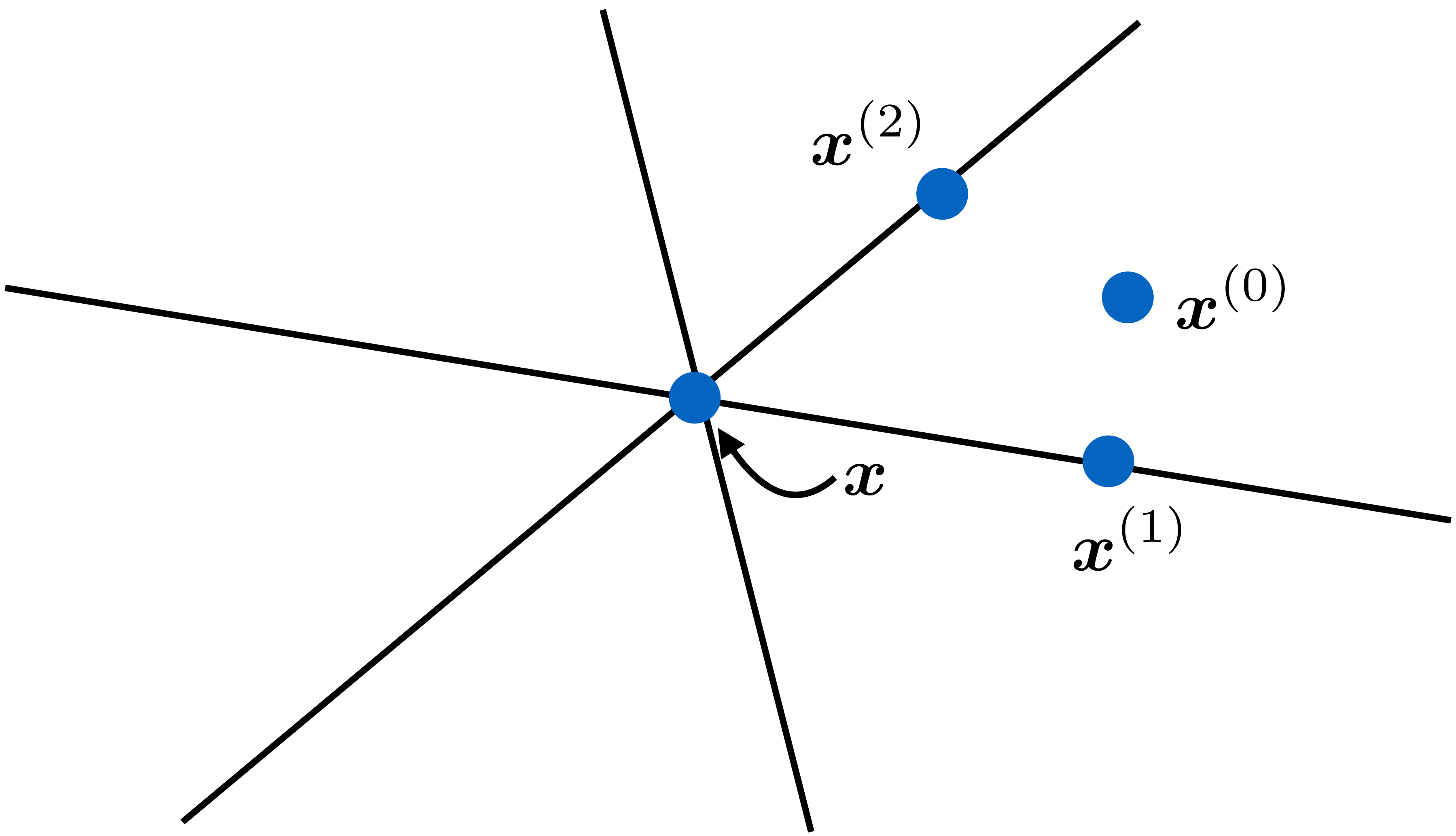
3. “Quenched error exponent”

$$\boldsymbol{x}^{(k)} = \boldsymbol{x}^{(k-1)} + \frac{y_r - \boldsymbol{a}_r^T \boldsymbol{x}^{(k-1)}}{||\boldsymbol{a}_r||^2} \boldsymbol{a}_r$$



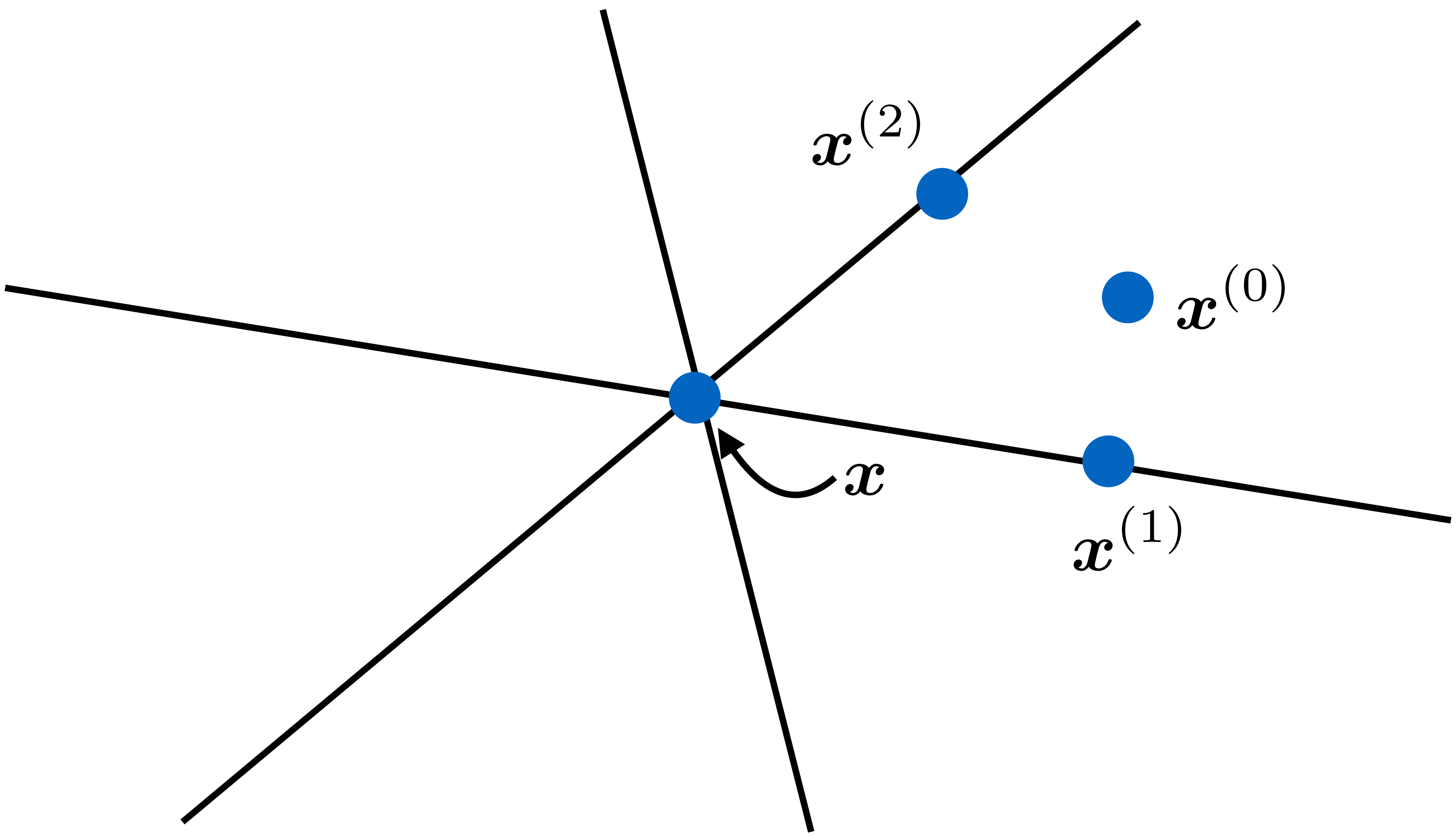
Exact MSE Formula

$$\boldsymbol{x}^{(k)} - \boldsymbol{x} = \boldsymbol{x}^{(k-1)} - \boldsymbol{x} + \frac{y_r - \boldsymbol{a}_r^T \boldsymbol{x}^{(k-1)}}{||\boldsymbol{a}_r||^2}\boldsymbol{a}_r$$



Exact MSE Formula

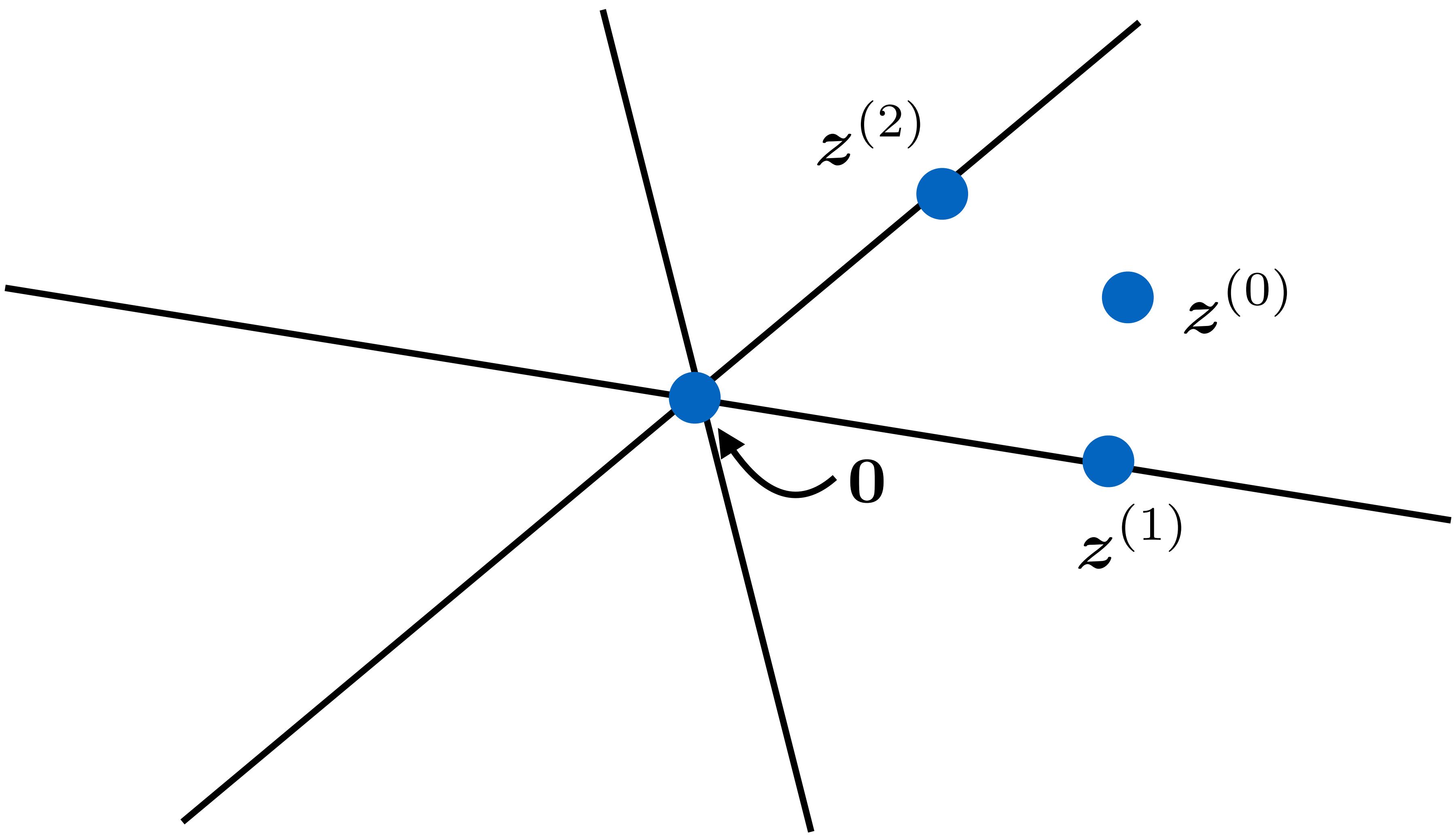
$$(\boldsymbol{x}^{(k)} - \boldsymbol{x}) = (\boldsymbol{x}^{(k-1)} - \boldsymbol{x}) - \frac{\boldsymbol{a}_r^T(\boldsymbol{x}^{(k-1)} - \boldsymbol{x})}{||\boldsymbol{a}_r||^2}\boldsymbol{a}_r$$



Exact MSE Formula

$$(\boldsymbol{x}^{(k)} - \boldsymbol{x}) = (\boldsymbol{x}^{(k-1)} - \boldsymbol{x}) - \frac{\boldsymbol{a}_r^T(\boldsymbol{x}^{(k-1)} - \boldsymbol{x})}{||\boldsymbol{a}_r||^2}\boldsymbol{a}_r$$

$\boldsymbol{z}^{(k)} = Q_k \boldsymbol{z}^{(k-1)}$ where $\boldsymbol{Q}_k = \left(I - \frac{\boldsymbol{a}_r \boldsymbol{a}_r^T}{||\boldsymbol{a}_r||^2}\right)$

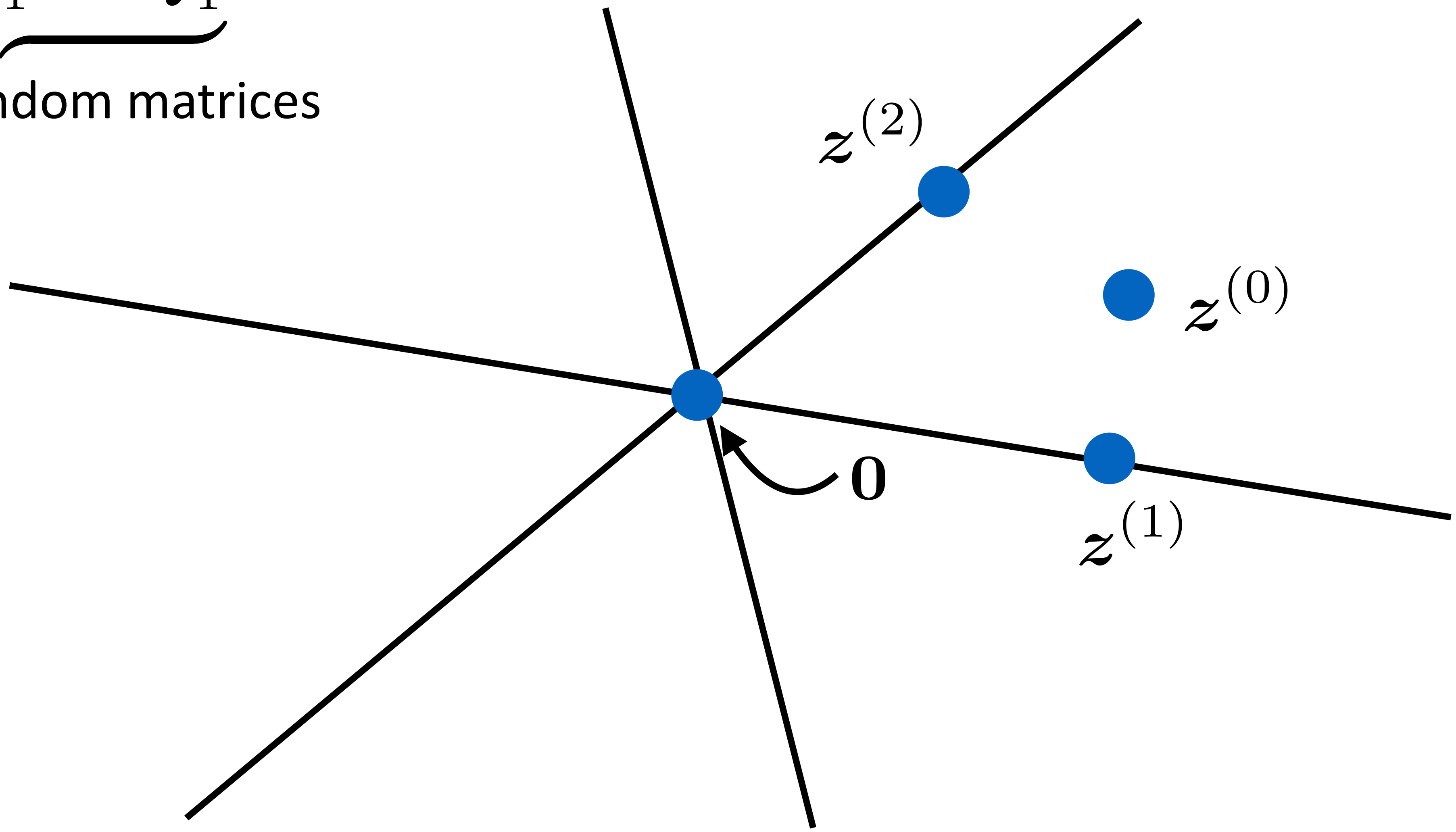


Exact MSE Formula

$$(\boldsymbol{x}^{(k)} - \boldsymbol{x}) = (\boldsymbol{x}^{(k-1)} - \boldsymbol{x}) - \frac{\boldsymbol{a}_r^T(\boldsymbol{x}^{(k-1)} - \boldsymbol{x})}{||\boldsymbol{a}_r||^2}\boldsymbol{a}_r$$

$$\boldsymbol{z}^{(k)} = Q_k \boldsymbol{z}^{(k-1)} \quad \text{where} \quad \boldsymbol{Q}_k = \left(I - \frac{\boldsymbol{a}_r \boldsymbol{a}_r^T}{||\boldsymbol{a}_r||^2} \right)$$

$$\boldsymbol{z}^{(k)} = \underbrace{\boldsymbol{Q}_k \boldsymbol{Q}_{k-1} \cdots \boldsymbol{Q}_1}_{\text{product of random matrices}} \boldsymbol{z}^{(0)}$$



Exact MSE Formula

Proposition (A.-Wang-Lu 2014)

$$\mathrm{MSE}_N = \mathrm{vec}\, \boldsymbol{I}^T (\mathbb{E}\boldsymbol{Q} \otimes \boldsymbol{Q})^N \,\mathrm{vec}(\boldsymbol{z}^{(0)} \boldsymbol{z}^{(0)T})$$

where $\mathbb{E}\boldsymbol{Q} \otimes \boldsymbol{Q} = \sum_i p_i \left(\boldsymbol{I} - \frac{\boldsymbol{a}_i \boldsymbol{a}_i^T}{||\boldsymbol{a}_i||^2} \right)^{\otimes 2}$

vec — vectorization operator; stack columns of matrix into vector.

$\otimes$ — matrix Kronecker product.

Proof sketch

$$\underbrace{\mathbb{E}||\boldsymbol{z}^{(N)}||^2}_{\text{MSE}} = \mathbb{E}||\boldsymbol{Q}_N \boldsymbol{Q}_{N-1} \cdots \boldsymbol{Q}_1 \boldsymbol{z}^{(0)}||^2$$

$$= \mathbb{E}\boldsymbol{z}^{(0)T} \boldsymbol{Q}_1 \boldsymbol{Q}_2 \cdots \boldsymbol{Q}_N \boldsymbol{Q}_N \cdots \boldsymbol{Q}_2 \boldsymbol{Q}_1 \boldsymbol{z}^{(0)}$$

$$= \mathbb{E}\,\mathrm{trace}(\boldsymbol{Q}_1 \boldsymbol{Q}_2 \cdots \boldsymbol{Q}_N \boldsymbol{Q}_N \cdots \boldsymbol{Q}_2 \boldsymbol{Q}_1 \boldsymbol{z}^{(0)} \boldsymbol{z}^{(0)T})$$

$$= \mathbb{E}\,\mathrm{vec}(\boldsymbol{Q}_1 \boldsymbol{Q}_2 \cdots \boldsymbol{Q}_N \boldsymbol{Q}_N \cdots \boldsymbol{Q}_2 \boldsymbol{Q}_1)^T \,\mathrm{vec}(\boldsymbol{z}^{(0)} \boldsymbol{z}^{(0)T})$$

Proof sketch

Matrix identities

$$\mathrm{vec}(ABC) = (C^T \otimes A)\,\mathrm{vec}(B)$$

$$\mathrm{trace}(AB) = \mathrm{vec}(A)^T\,\mathrm{vec}(B)$$

$$\mathrm{trace}\,AB = \mathrm{trace}\,BA$$

$$= \mathbb{E}\,\mathrm{vec}(\boldsymbol{Q}_1\boldsymbol{Q}_2\cdots\boldsymbol{Q}_N\boldsymbol{Q}_N\cdots\boldsymbol{Q}_2\boldsymbol{Q}_1)^T\,\mathrm{vec}(\boldsymbol{z}^{(0)}\boldsymbol{z}^{(0)T})$$

Exact MSE formula

$$
\begin{aligned}
&\mathrm{MSE}_N \\
&= \mathbb{E}\{\mathrm{vec}(\boldsymbol{Q}_1\boldsymbol{Q}_2\cdots\boldsymbol{Q}_N\boldsymbol{Q}_N\cdots\boldsymbol{Q}_2\boldsymbol{Q}_1)\}^T\,\mathrm{vec}(\boldsymbol{z}^{(0)}\boldsymbol{z}^{(0)T}) \\
&= \mathbb{E}\{(\boldsymbol{Q}_1\otimes\boldsymbol{Q}_1)\,\mathrm{vec}(\boldsymbol{Q}_2\cdots\boldsymbol{Q}_N\boldsymbol{Q}_N\cdots\boldsymbol{Q}_2)\}^T\,\mathrm{vec}(\boldsymbol{z}^{(0)}\boldsymbol{z}^{(0)T})
\end{aligned}
$$

$\blacksquare$

Exact MSE formula

$$
\begin{aligned}
&\mathrm{MSE}_N \\
&= \mathbb{E}\{\mathrm{vec}(\boldsymbol{Q}_1\boldsymbol{Q}_2\cdots\boldsymbol{Q}_N\boldsymbol{Q}_N\cdots\boldsymbol{Q}_2\boldsymbol{Q}_1)\}^T\,\mathrm{vec}(\boldsymbol{z}^{(0)}\boldsymbol{z}^{(0)T}) \\
&= \{(\mathbb{E}\boldsymbol{Q}\otimes\boldsymbol{Q})^N\,\mathrm{vec}(\boldsymbol{I})\}^T\,\mathrm{vec}(\boldsymbol{z}^{(0)}\boldsymbol{z}^{(0)T})
\end{aligned}
$$

$\blacksquare$

Exact MSE formula

$$
\begin{aligned}
&\mathrm{MSE}_N \\
&= \mathbb{E}\{\mathrm{vec}(\boldsymbol{Q}_1\boldsymbol{Q}_2\cdots\boldsymbol{Q}_N\boldsymbol{Q}_N\cdots\boldsymbol{Q}_2\boldsymbol{Q}_1)\}^T\,\mathrm{vec}(\boldsymbol{z}^{(0)}\boldsymbol{z}^{(0)T}) \\
&= \{(\mathbb{E}\boldsymbol{Q}\otimes\boldsymbol{Q})^N\,\mathrm{vec}(\boldsymbol{I})\}^T\,\mathrm{vec}(\boldsymbol{z}^{(0)}\boldsymbol{z}^{(0)T}) \\
&= \mathrm{vec}\,\boldsymbol{I}^T(\mathbb{E}\boldsymbol{Q}\otimes\boldsymbol{Q})^N\,\mathrm{vec}(\boldsymbol{z}^{(0)}\boldsymbol{z}^{(0)T})\ \blacksquare
\end{aligned}
$$

MSE decays exponentially: $\mathbb{E}||\boldsymbol{z}^{(N)}||^2 = \exp(-\gamma_a N + o(N))$

Error exponent

$$\gamma_a \stackrel{\text{def}}{=} \lim_{N\to\infty} -\frac{1}{N} \log \mathbb{E}||\boldsymbol{z}^{(N)}||^2$$

We can compute the error exponent:

$$\gamma_a = -\log \lambda_{\max} \left(\sum_i p_i \left(\boldsymbol{I} - \frac{\boldsymbol{a}_i \boldsymbol{a}_i^T}{||\boldsymbol{a}_i||^2} \right)^{\otimes 2} \right)$$

- Can be computed in $O(mn^2)$ time
- Must be positive; exponential convergence confirmed

1. Exact MSE formula and decay rate

2. **Optimization of row selection probabilities**

3. “Quenched error exponent”

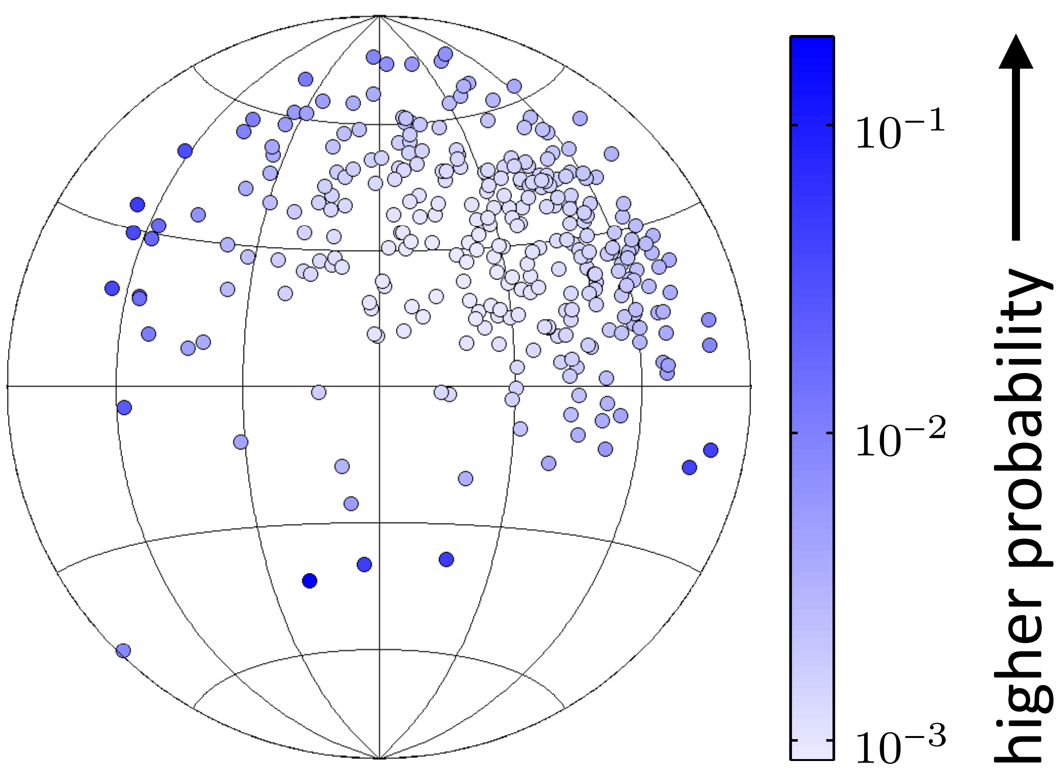
Optimal row selection

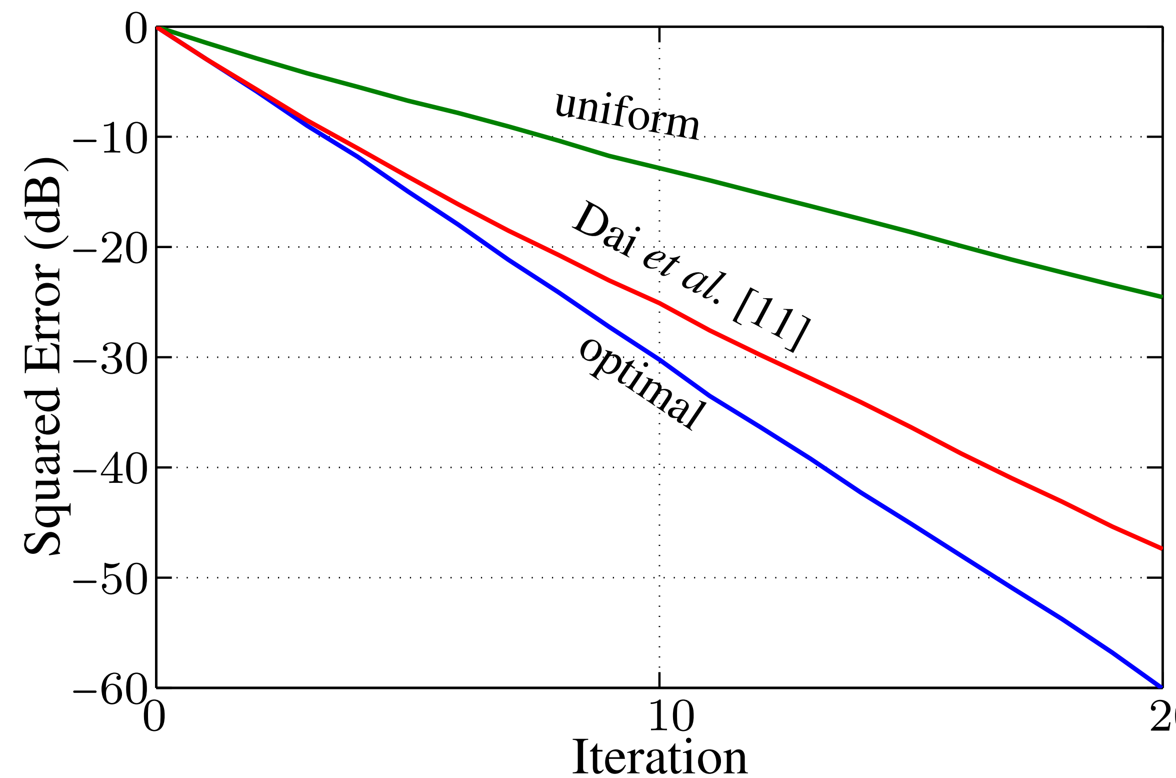
Convex optimization problem: minimize error exponent.

$$(p_1, \ldots, p_m) = \operatorname{argmin}_{\boldsymbol{p}} \lambda_{\max} \left(\sum_i p_i \left(\boldsymbol{I} - \frac{\boldsymbol{a}_i \boldsymbol{a}_i^T}{||\boldsymbol{a}_i||^2} \right)^{\otimes 2} \right)$$

semi-definite programming

n = 3 lets us easily visualize the optimal probabilities





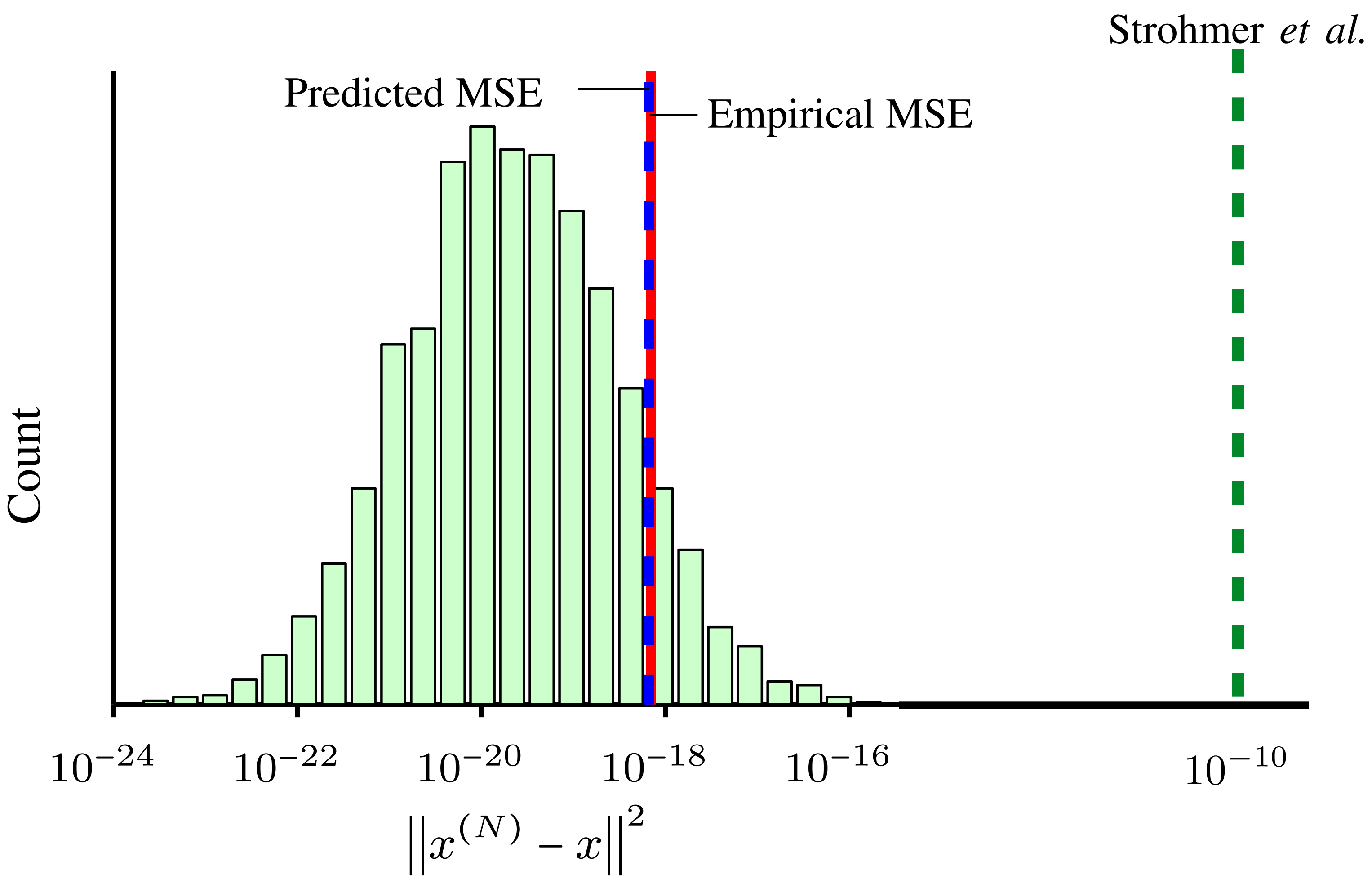
Intuition: explorers of sparsely-populated regions chosen with higher probability

This talk

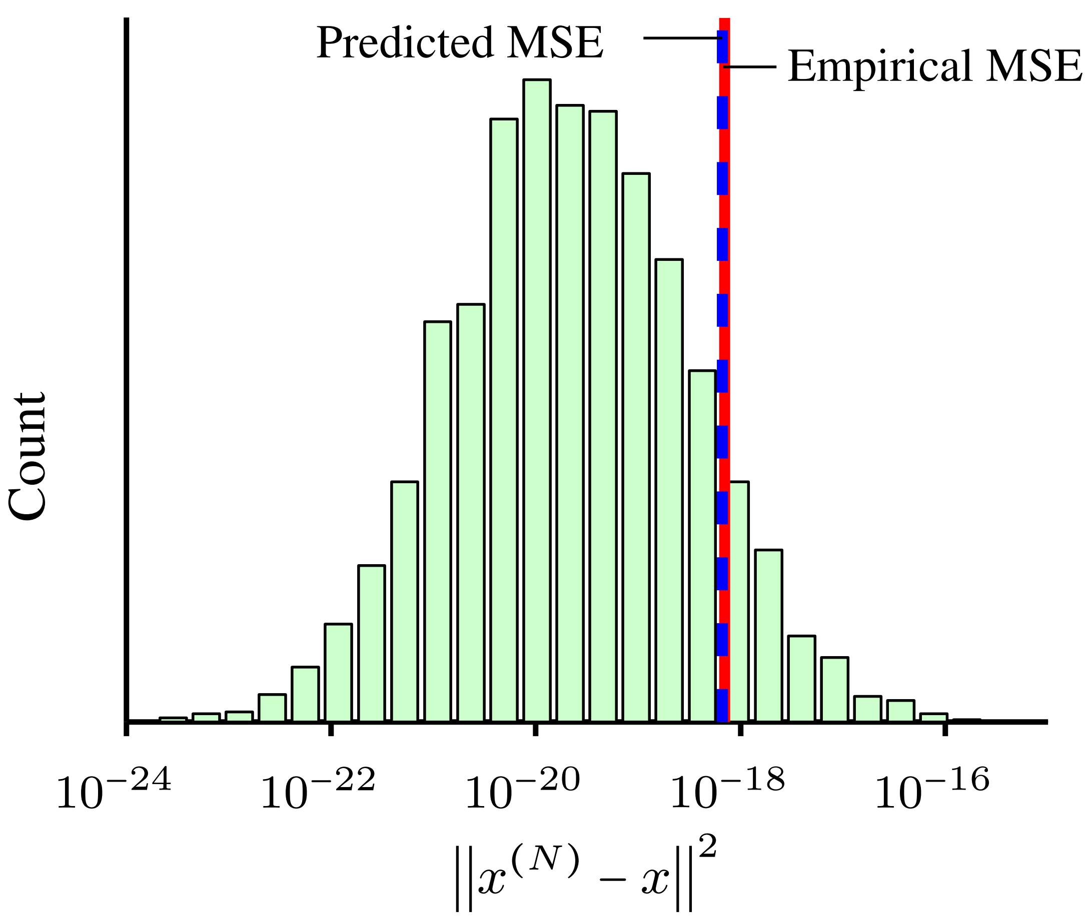
1. Exact MSE formula and decay rate

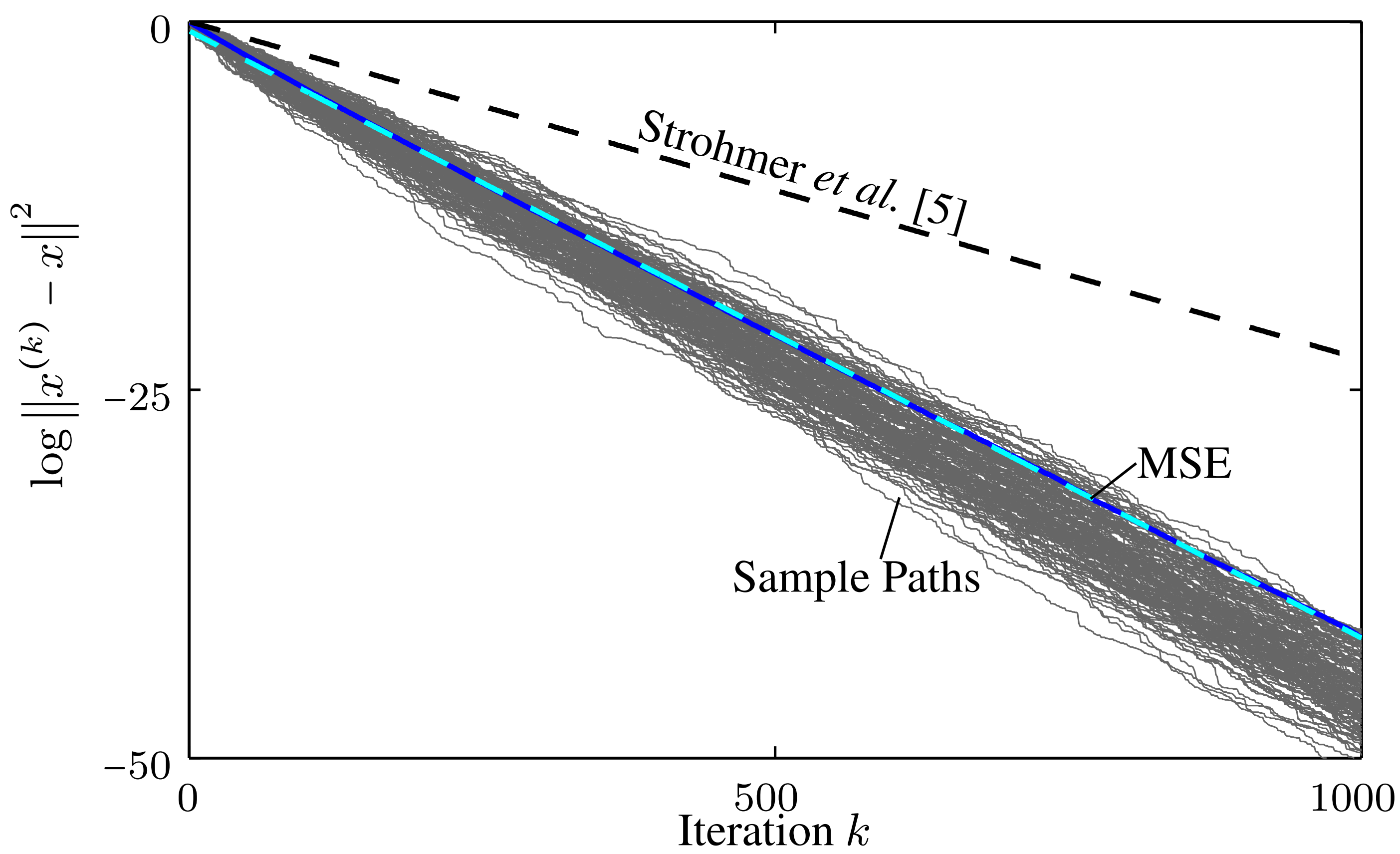
2. Optimization of row selection probabilities

3. **“Quenched error exponent”**



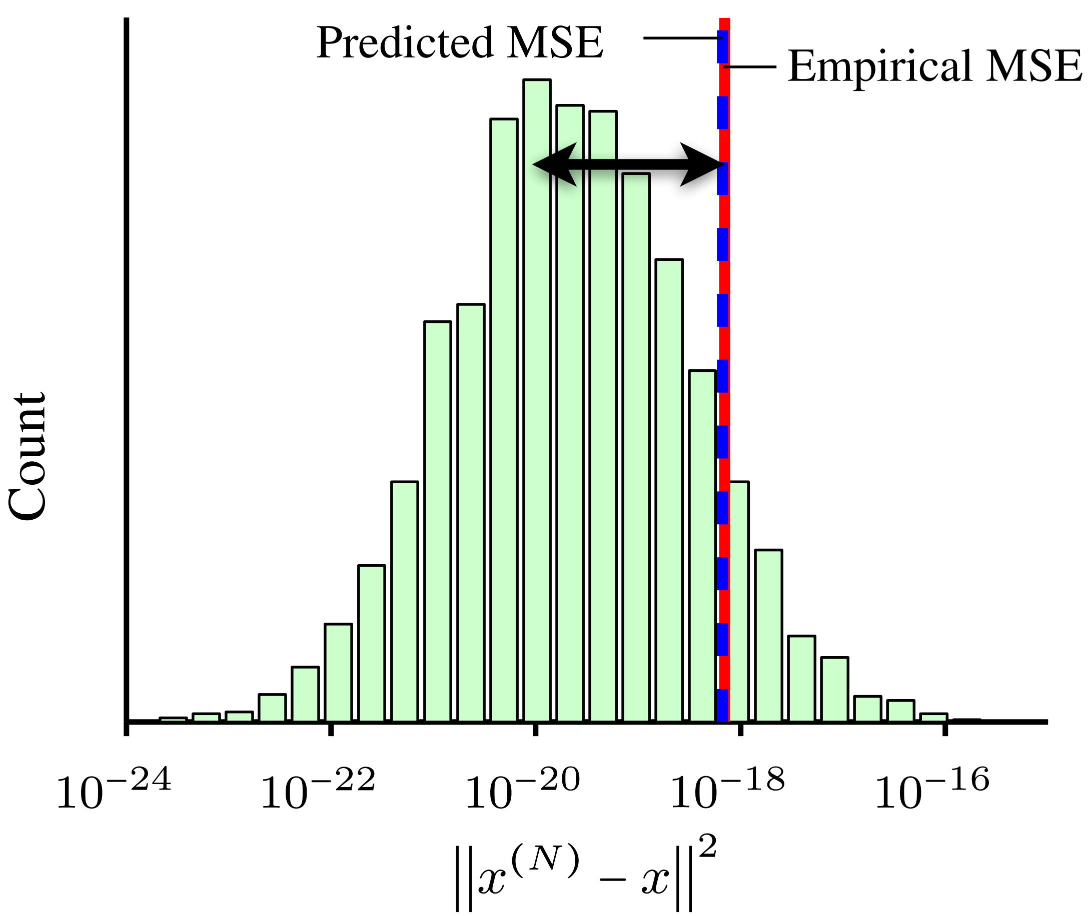
150 x 20 matrix w/ Gaussian entries.

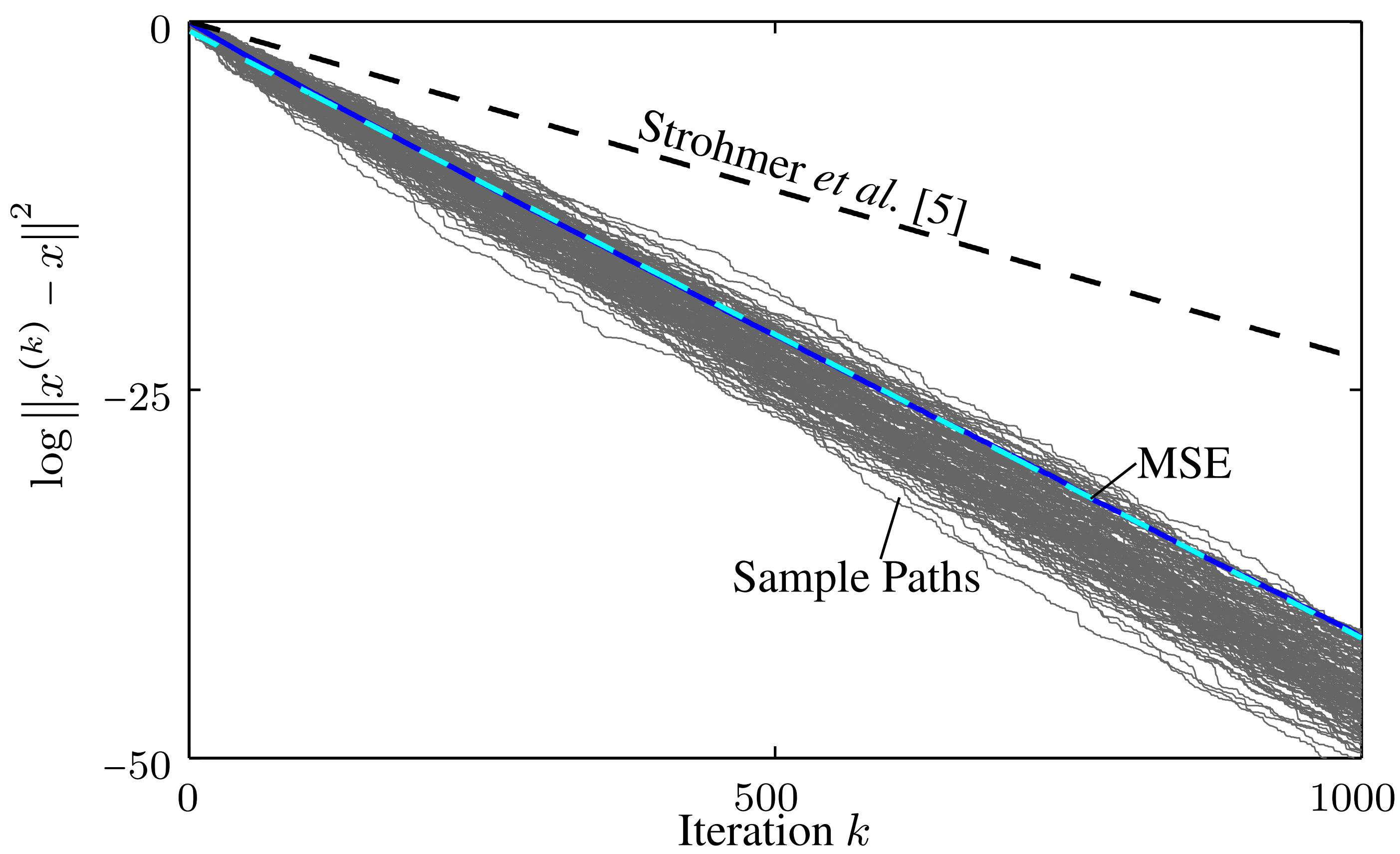




150 x 20 matrix w/ Gaussian entries.

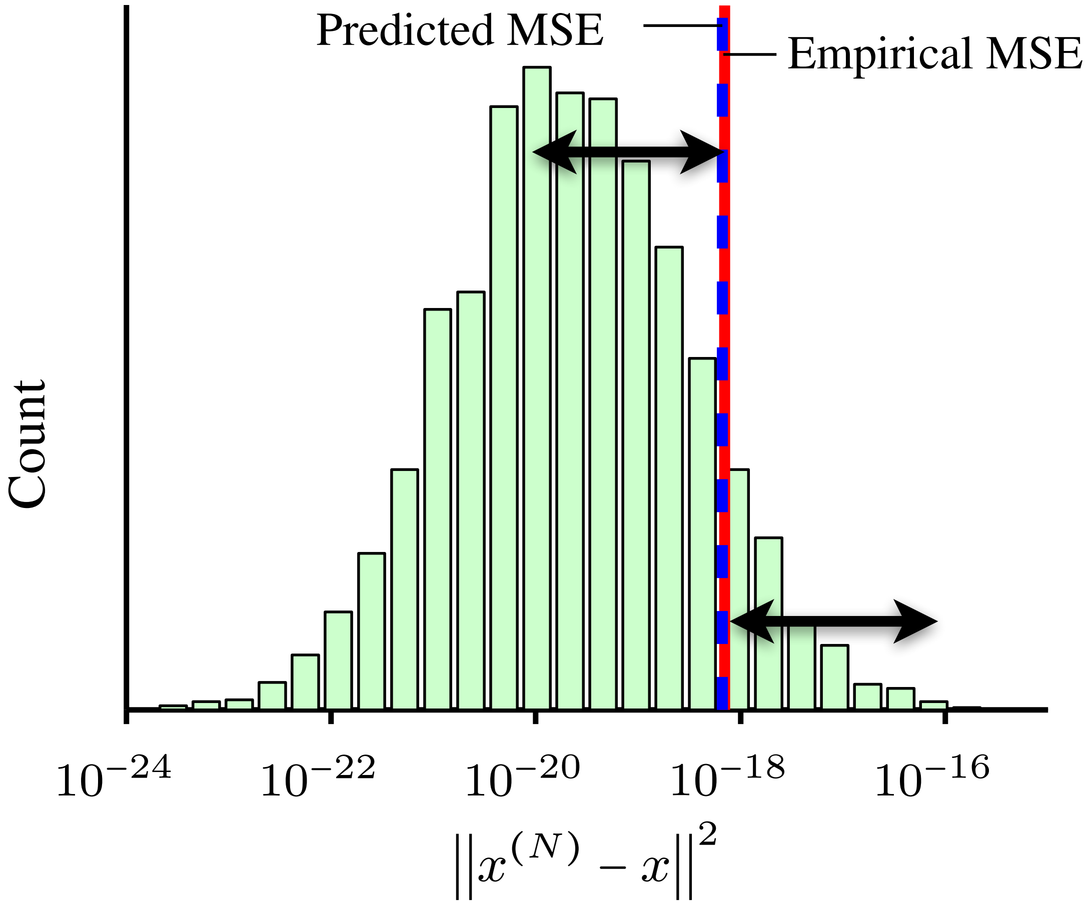
Simulations

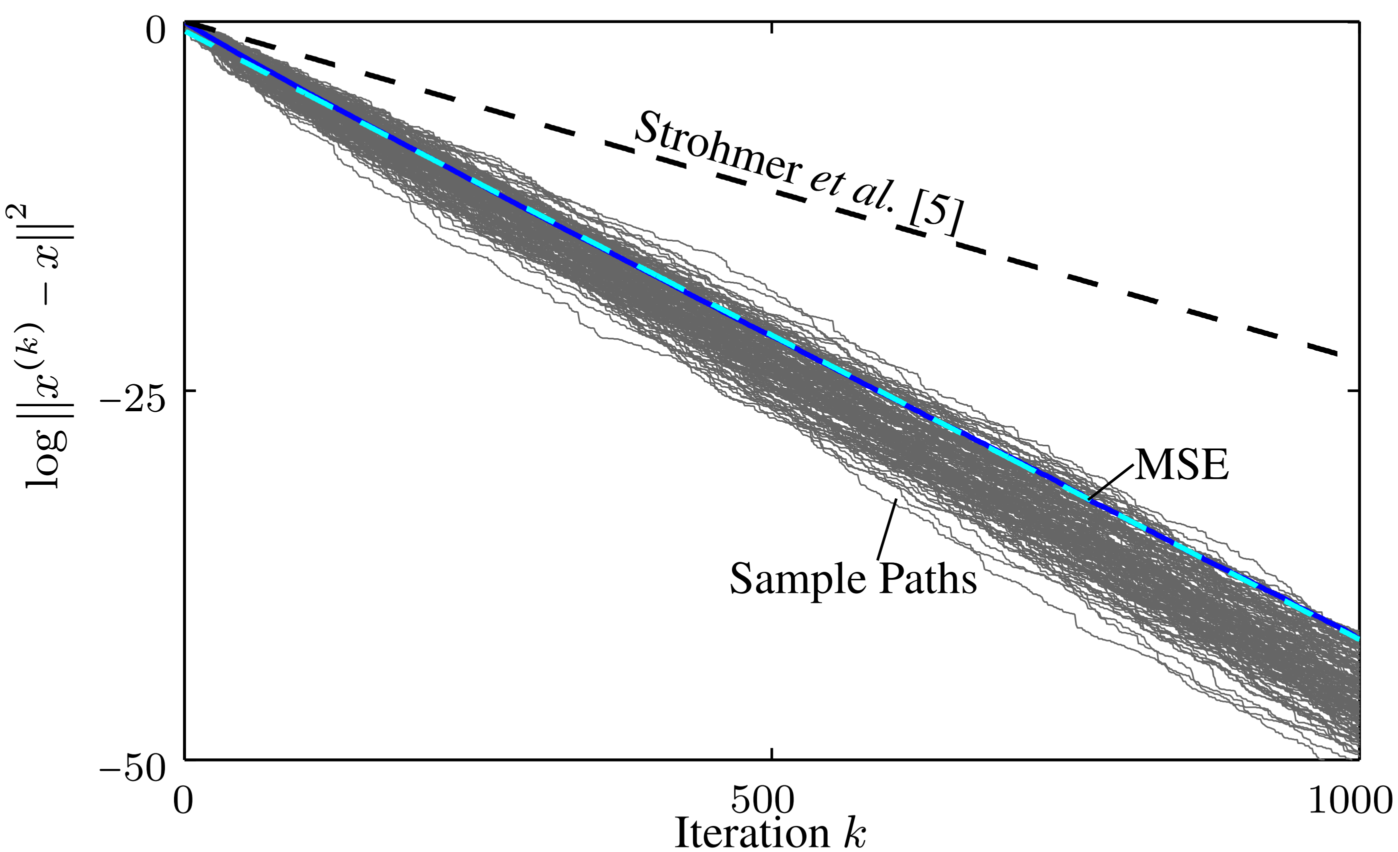




150 x 20 matrix w/ Gaussian entries.

Simulations





150 x 20 matrix w/ Gaussian entries.


IF YOU WANT TO MEASURE
TYPICAL PERFORMANCE
DON'T
USE the MSE!
quickmeme.com

Quenched error exponent

Average performance:

Annealed error exponent

$$\gamma_a \stackrel{\text{def}}{=} \lim_{N\to\infty} -\frac{1}{N} \log \mathbb{E}||\boldsymbol{z}^{(N)}||^2$$

Typical performance:

Quenched error exponent

$$\gamma_q \stackrel{\text{def}}{=} \lim_{N\to\infty} -\frac{1}{N} \mathbb{E} \log ||\boldsymbol{z}^{(N)}||^2$$

- Much more difficult to analyze.
- Known to physicists as the top ***Lyapunov exponent***.
- They use heuristics to solve.

Physicists have their own intuition for this trick, but we can get the same result by assuming the error is log-normal:

Assume

$$\log ||\boldsymbol{z}^{(N)}||^2 \sim \mathcal{N}(N\mu, N\sigma^2).$$

Then

$$\gamma_q = -\mu$$

$$\mathbb{E}||\boldsymbol{z}^{(N)}||^2 = \exp(N[\mu + \frac{1}{2}\sigma^2])$$

$$\mathbb{E}||\boldsymbol{z}^{(N)}||^4 = \exp(N[2\mu + 2\sigma^2])$$

Naive replica method $\log Z = \lim_{n \to 0} \frac{Z^n - 1}{n}$

Solve

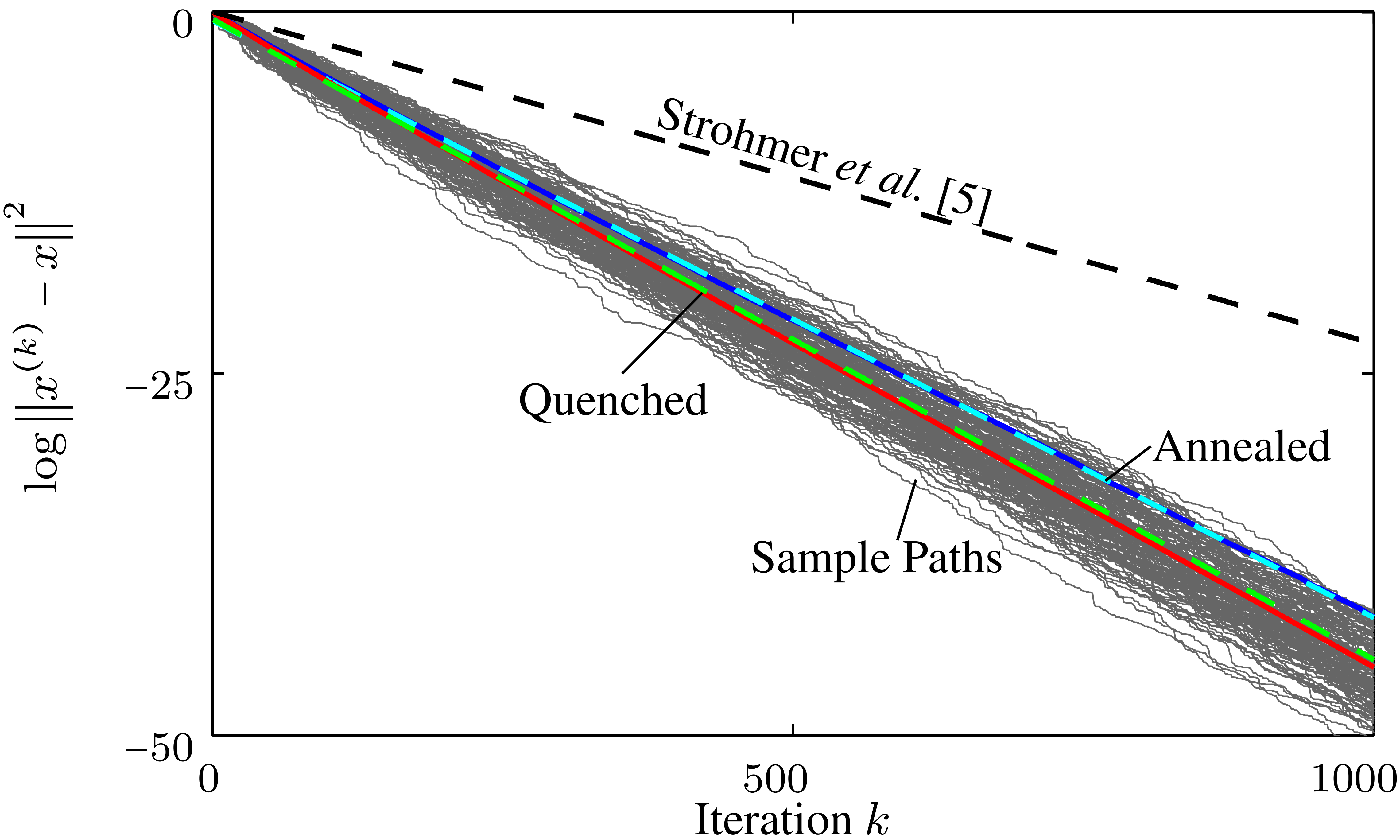
$$\mu = \frac{1}{N}\left[2\log\mathbb{E}||\boldsymbol{z}^{(N)}||^2 - \frac{1}{2}\log\mathbb{E}||\boldsymbol{z}^{(N)}||^4\right]$$

$$\gamma_q \approx 2\gamma_a - \frac{1}{2}\gamma_a^{(2)}$$

where $\gamma_a^{(2)} = -\log\lambda_{\max}\left(\sum_i p_i\left(\boldsymbol{I} - \frac{\boldsymbol{a}_i\boldsymbol{a}_i^T}{||\boldsymbol{a}_i||^2}\right)^{\otimes 4}\right)$

Quenched error exponent

$$\gamma_q \stackrel{\text{def}}{=} \lim_{N\to\infty} -\frac{1}{N}\mathbb{E}\log||\boldsymbol{z}^{(N)}||^2$$



0
−25
−50
0
500
1000
Iteration k
$\log \|x^{(k)} - x\|^2$
Strohmer *et al.* [5]
Quenched
Annealed
Sample Paths

Summary

- ***Exact MSE formula for randomized Kaczmarz algorithms (and its generalizations)***

  Lifting!

- ***Annealed and quenched error exponents give decay rate***

  Average vs. typical performance

- ***Finding optimal row selection probabilities***

  Convex optimization