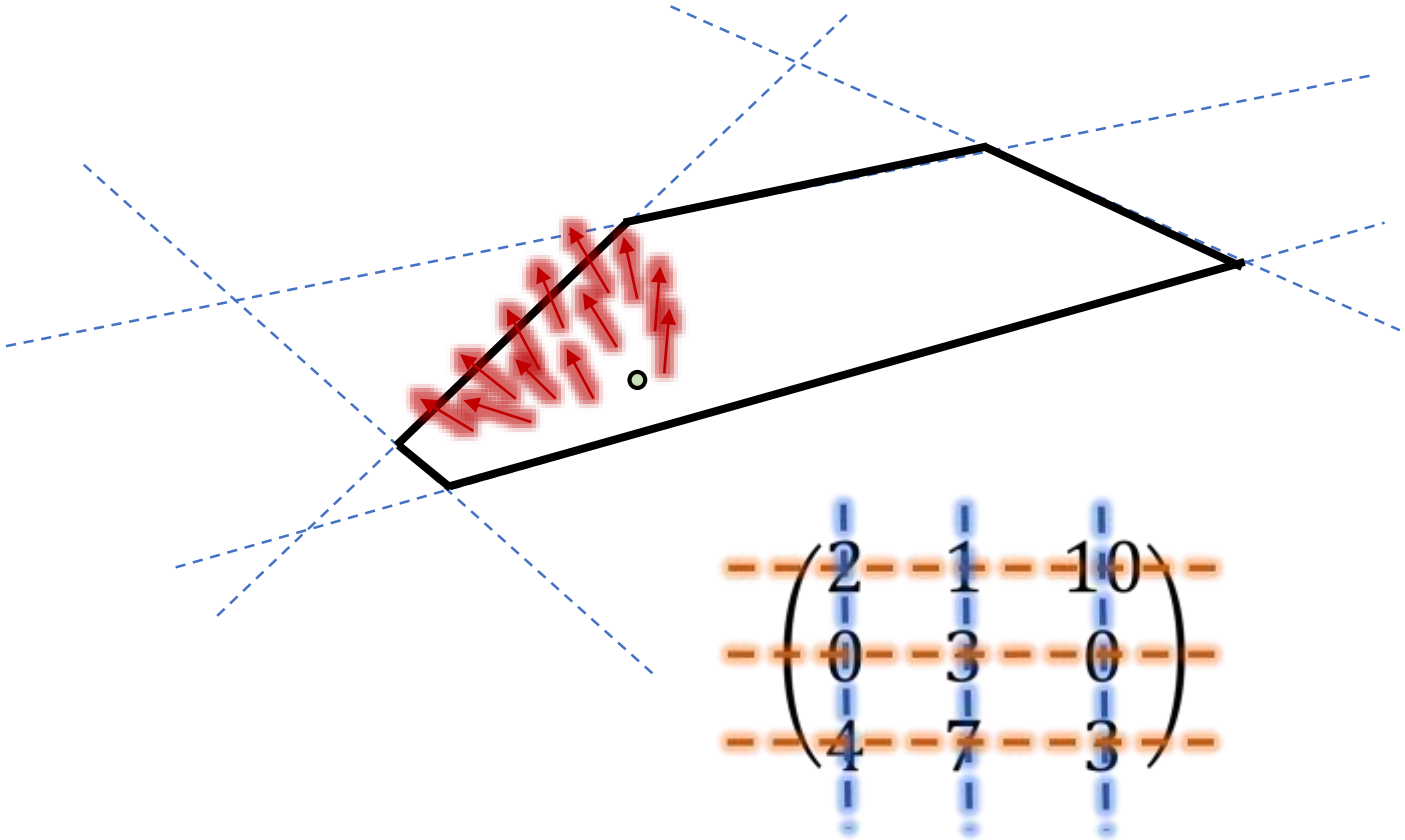
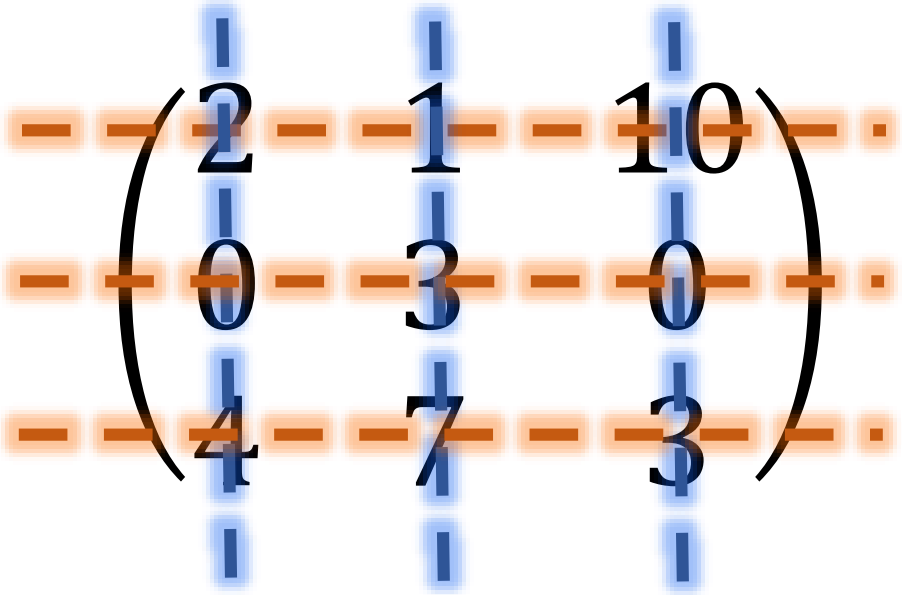
regularized gradient flows on convex bodies

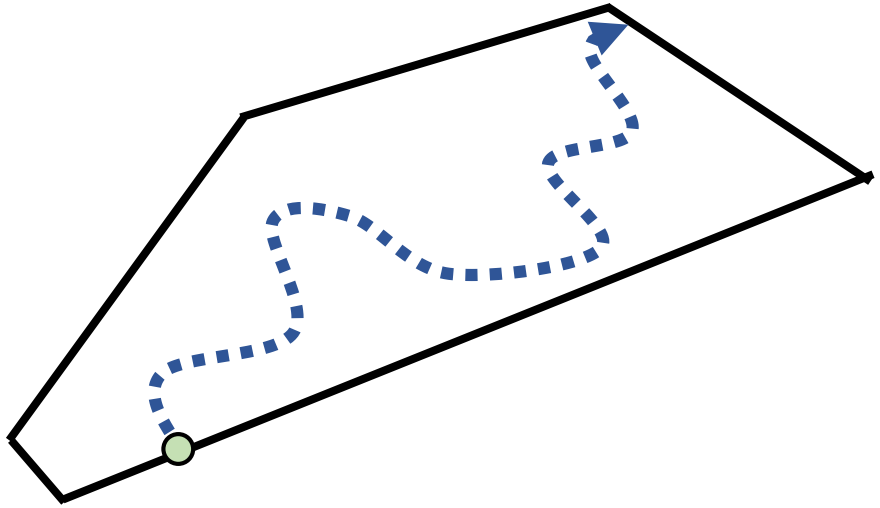
James R. Lee
University of Washington



alternating minimization and competing constraints

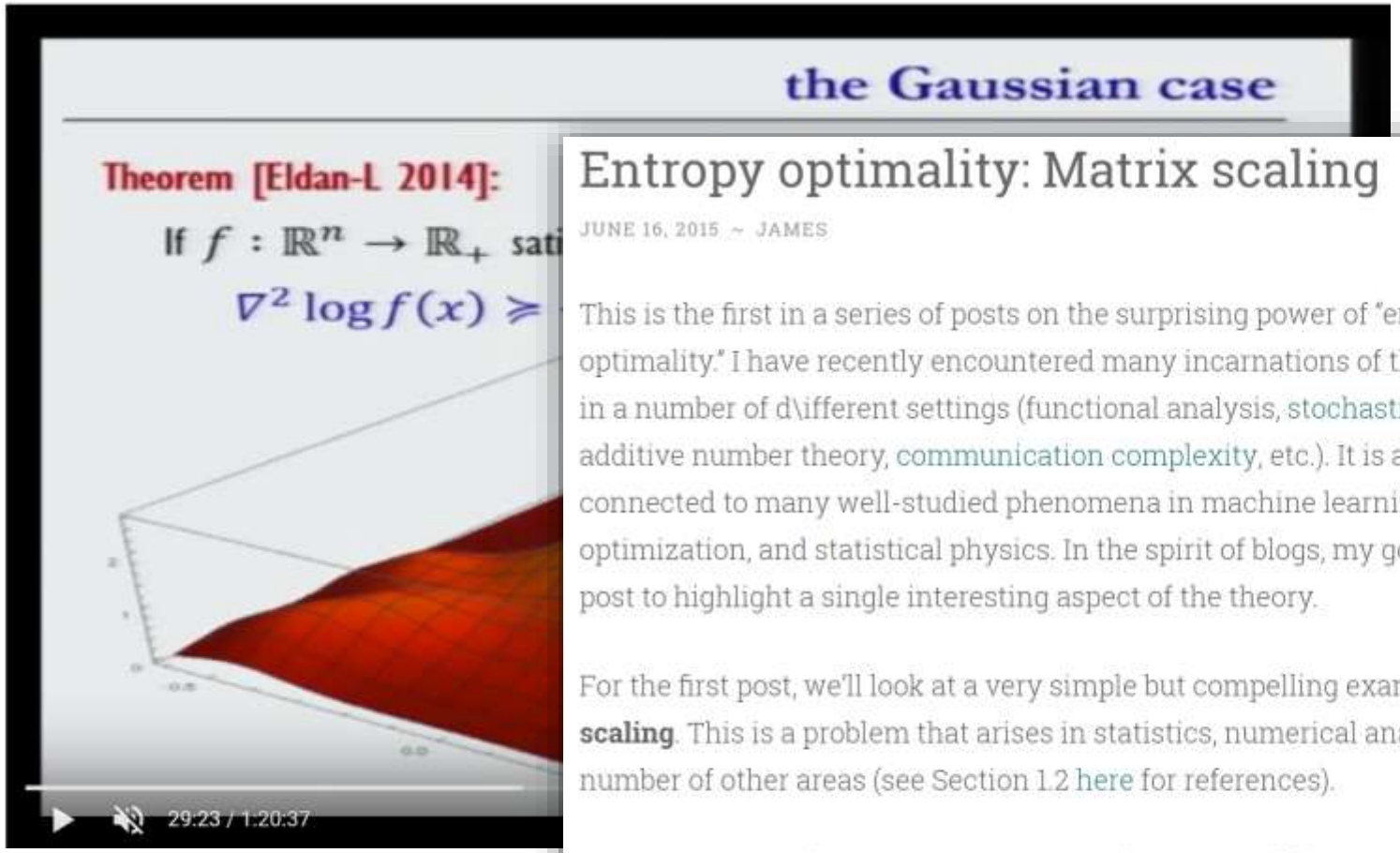


Potential function (e.g., the capacity) drives progress forward.



The dynamics can tell us interesting things about the endpoint.

a small world



Entropy optimality: Matrix scaling

JUNE 16, 2015 ~ JAMES

This is the first in a series of posts on the surprising power of "entropy optimality." I have recently encountered many incarnations of this principle in a number of d\ifferent settings (functional analysis, stochastic processes, additive number theory, communication complexity, etc.). It is also connected to many well-studied phenomena in machine learning, optimization, and statistical physics. In the spirit of blogs, my goal is for each post to highlight a single interesting aspect of the theory.

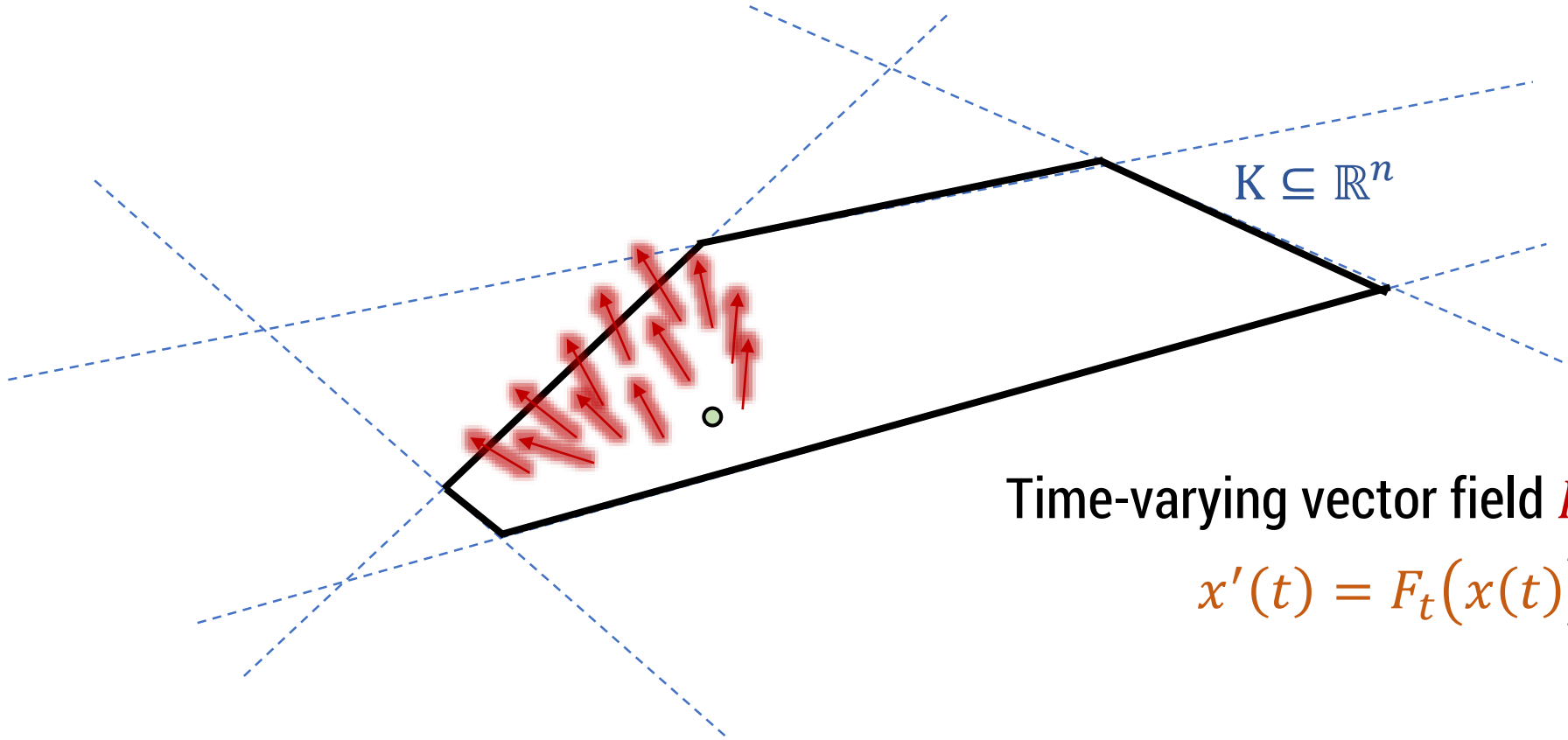
For the first post, we'll look at a very simple but compelling example: **matrix scaling**. This is a problem that arises in statistics, numerical analysis, and a number of other areas (see Section 1.2 here for references).

Suppose we are given an $n \times n$ target matrix $T = (t_{ij})$ with nonnegative entries. We also have an $n \times n$ input matrix $X = (x_{ij})$ and our goal is to multiply the rows and columns of X by positive weights so that the resulting matrix has the same row and column sums as T.

Entropy optimality:

- Matrix scaling
- Chang's Lemma
- A potential function
- Bloom's Chang's Lemma
- Forster's isotropy
- Primer: Lifts of polytopes
- Lifts of polytopes
- Non-negative rank
- Quantum lifts
- Analytic PSD rank
- HIM Lecture notes
- Optimality on path spaces
- Follmer drift and log-Sobolev

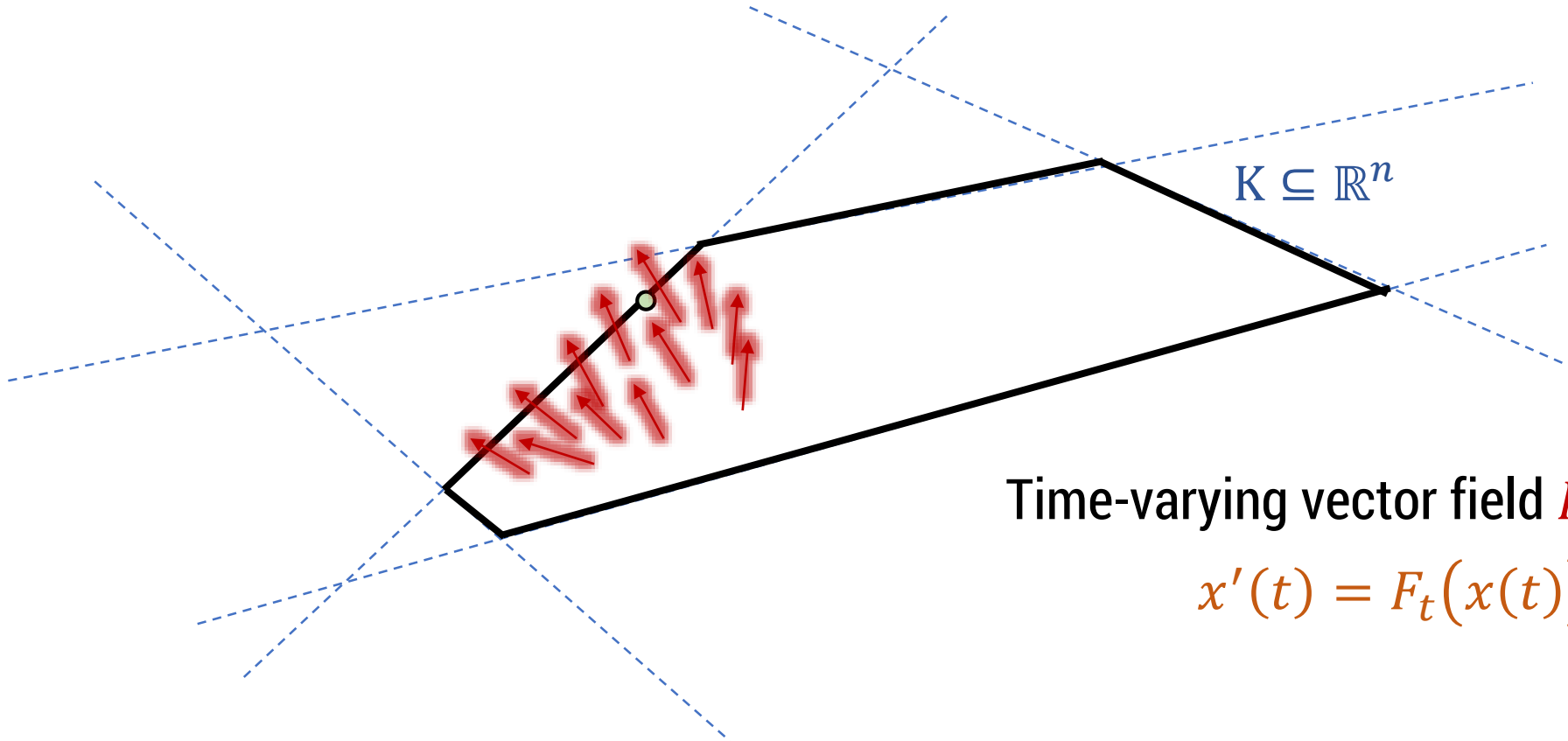
navigating a convex body online



Time-varying vector field $F_t : \mathrm{K} \to \mathbb{R}^n$

$$x'(t) = F_t\big(x(t)\big)$$

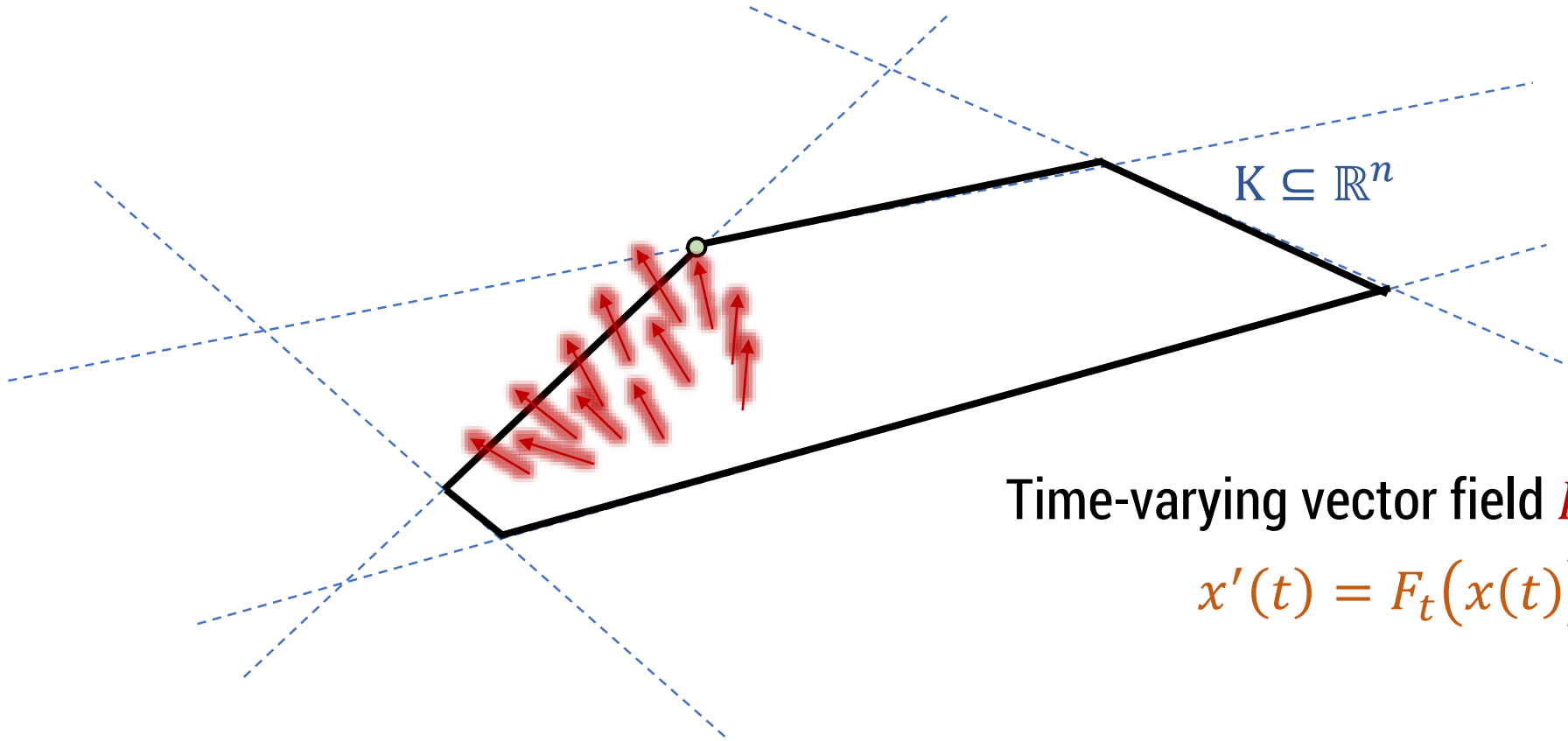
navigating a convex body online



Time-varying vector field $F_t : \mathrm{K} \to \mathbb{R}^n$

$$x'(t) = F_t\big(x(t)\big)$$

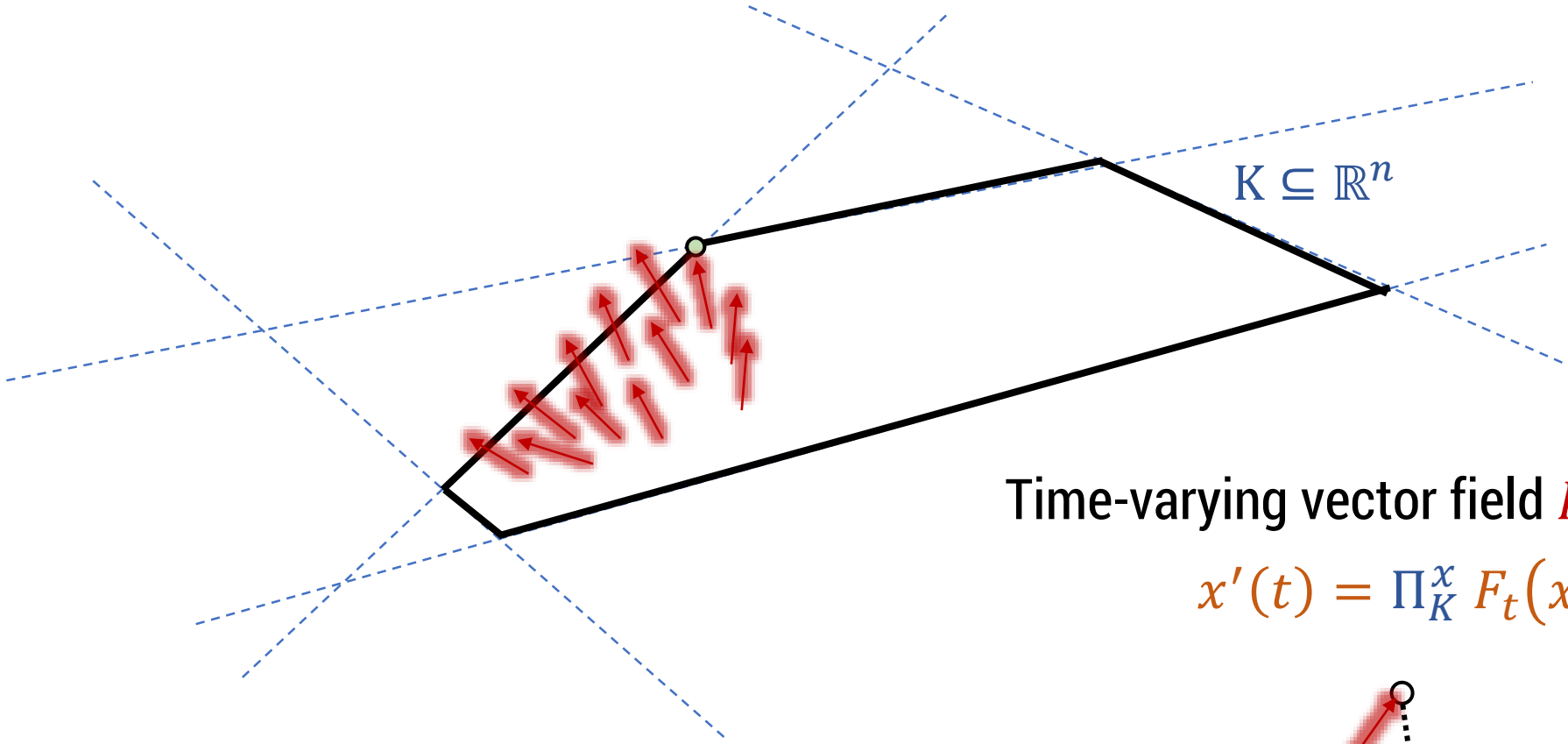
navigating a convex body online



Time-varying vector field $F_t : K \to \mathbb{R}^n$

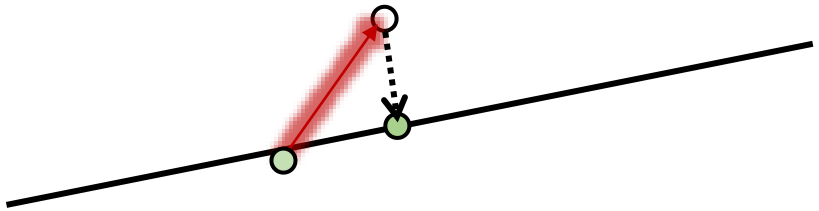
$$x'(t) = F_t\big(x(t)\big)$$

navigating a convex body online



Time-varying vector field $F_t : K \to \mathbb{R}^n$

$$x'(t) = \Pi_K^x F_t\big(x(t)\big)$$

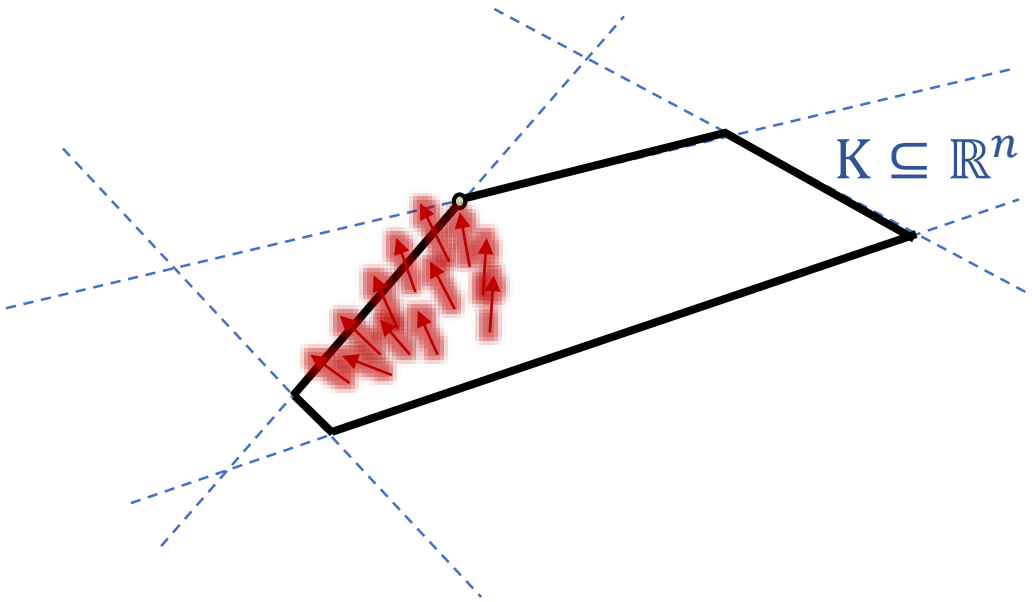


navigating a convex body online

Hessian structure:

If $\Phi : \mathrm{K} \to \mathbb{R}$ is strictly convex, it gives a local inner product at $x \in \mathrm{K}$:

$$\langle u, v \rangle_x = \langle u, \nabla^2 \Phi(x) v \rangle$$



Resulting algorithm is called…

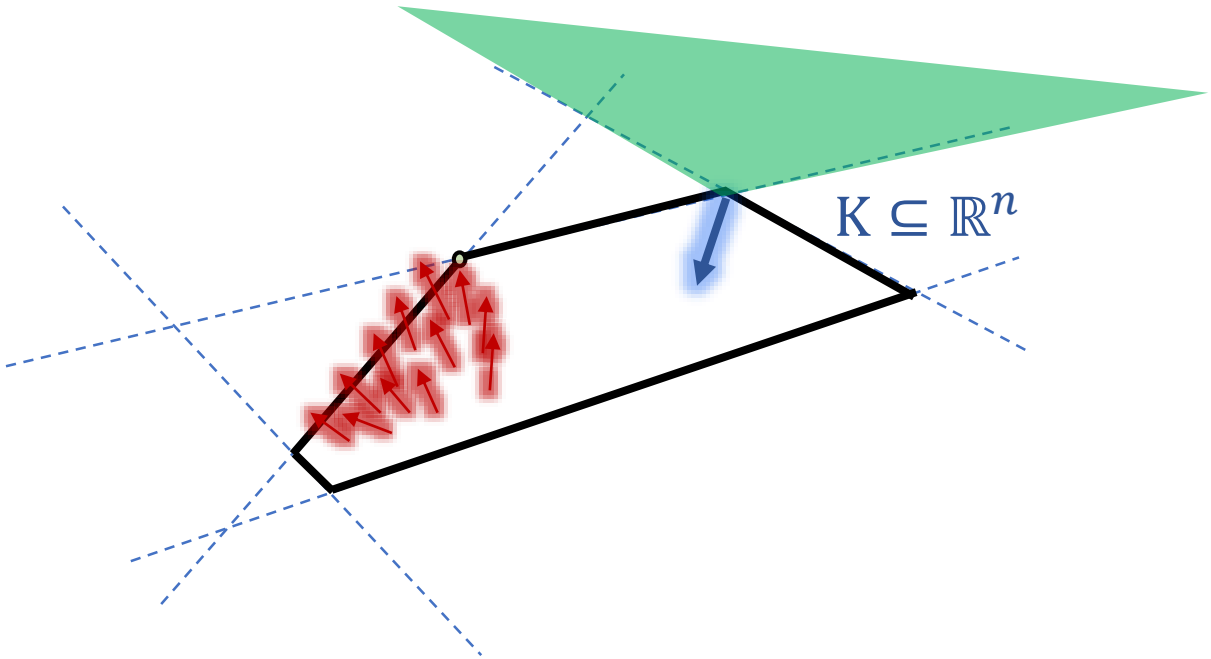
- Mirror descent
- Interior point method
- “Natural gradient” in information geometry

navigating a convex body online

Hessian structure:

If $\Phi : \mathrm{K} \to \mathbb{R}$ is strictly convex, it gives a local inner product at $x \in \mathrm{K}$:

$$\langle u, v \rangle_x = \langle u, \nabla^2 \Phi(x) v \rangle$$



Dynamics:

$$\nabla^2 \Phi\big(x(t)\big)\, x'(t) = F_t\big(x(t)\big) - \lambda(t)$$

$$x(0) = x_0 \in \mathrm{K}$$

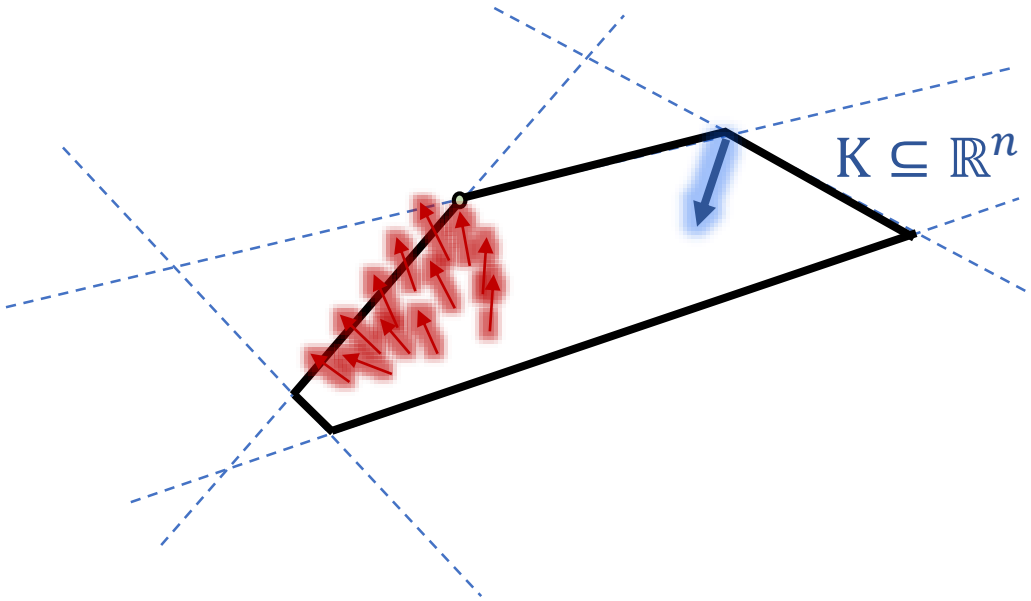
$$\lambda(t) \in \mathrm{N}_{\mathrm{K}}(x(t))$$

navigating a convex body online

Hessian structure:

If $\Phi : \mathrm{K} \to \mathbb{R}$ is strictly convex, it gives a local inner product at $x \in \mathrm{K}$:

$$\langle u, v \rangle_x = \langle u, \nabla^2 \Phi(x) v \rangle$$



Dynamics: [Bubeck-Cohen-L-Lee-Madry 2017]

$$\nabla^2 \Phi\big(x(t)\big)\, x'(t) = F_t\big(x(t)\big) - \lambda(t)$$

$$x(0) = x_0 \in \mathrm{K}$$

$$\lambda(t) \in \mathrm{N}_{\mathrm{K}}(x(t))$$

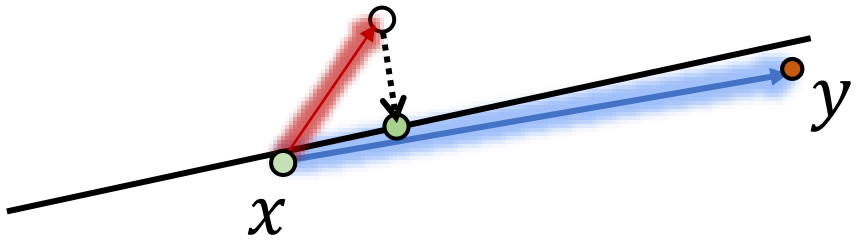
Lyapunov function:

$$D_\Phi(y; x) :=$$

$$\Phi(y) - \Phi(x) - \langle \nabla \Phi(x), y - x \rangle$$

For any $y \in \mathrm{K}$:

$$\partial_t D_\Phi\big(y; x(t)\big) \le -\big\langle F_t\big(x(t)\big), y - x(t) \big\rangle$$



Matrix scaling [Sinkhorn-Knopp 1967, Franklin-Lorenz 1989]

$$\mathrm{K} = \left\{ X \in \mathbb{R}_{+}^{n \times n} : \sum_{ij} X_{ij} = 1 \right\} \qquad X(0) = \text{input matrix}$$

$$F_t(X) \in \left\{ \pm \begin{pmatrix} 0 & 0 & 0 & 0 \\ 1 & 1 & 1 & 1 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{pmatrix}, \ldots, \pm \begin{pmatrix} 0 & 0 & 1 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 1 & 0 \end{pmatrix} \right\}$$

$$\Phi(\mathrm{X}) = \sum_{ij} X_{ij} \log X_{ij} \qquad \nabla^2 \Phi\big(x(t)\big)\, x'(t) = F_t\big(x(t)\big) - \lambda(t)$$

Matrix scaling [Sinkhorn-Knopp 1967, Franklin-Lorenz 1989]

$$\mathrm{K} = \left\{ X \in \mathbb{R}_+^{n\times n} : \sum_{ij} X_{ij} = 1 \right\}$$

$X(0)$ = input matrix

$$F_t(X) \in \left\{ \pm \begin{pmatrix} 0 & 0 & 0 & 0 \\ 1 & 1 & 1 & 1 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{pmatrix}, \ldots, \pm \begin{pmatrix} 0 & 0 & 1 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 1 & 0 \end{pmatrix} \right\}$$

Converges to target marginals $\Leftrightarrow$
There exists a left-right scaling $Y \in \mathrm{K}$ of $X(0)$ such that

$$\Phi(\mathrm{X}) = \sum_{ij} X_{ij} \log X_{ij}$$

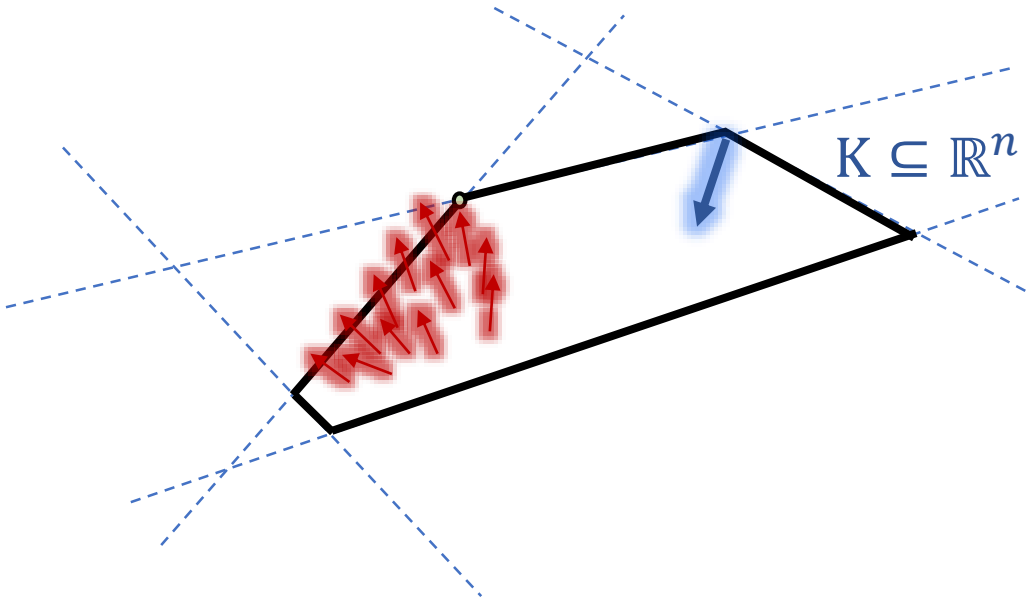
$$D_\Phi\big(Y; X(0)\big) = \sum_{ij} Y_{ij} \log \frac{Y_{ij}}{X_{ij}(0)} < \infty$$

navigating a convex body online

Hessian structure:

If $\Phi : \mathrm{K} \to \mathbb{R}$ is strictly convex, it gives a local inner product at $x \in \mathrm{K}$:

$$\langle u, v \rangle_x = \langle u, \nabla^2 \Phi(x) v \rangle$$



Dynamics: [Bubeck-Cohen-L-Lee-Madry 2017]

$$\nabla^2 \Phi\big(x(t)\big)\, x'(t) = F_t\big(x(t)\big) - \lambda(t)$$

$$x(0) = x_0 \in \mathrm{K}$$

$$\lambda(t) \in \mathrm{N}_{\mathrm{K}}(x(t))$$

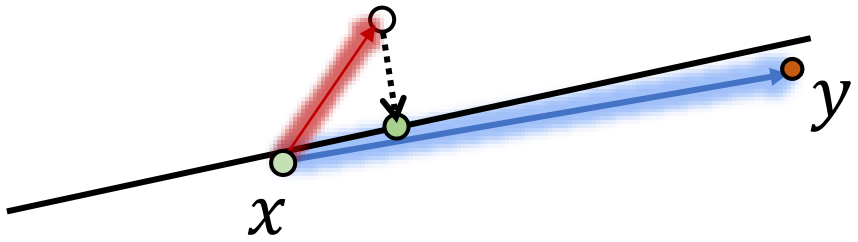
Lyapunov function:

$$D_\Phi(y; x) :=$$

$$\Phi(y) - \Phi(x) - \langle \nabla \Phi(x), y - x \rangle$$

For any $y \in \mathrm{K}$:

$$\partial_t D_\Phi\big(y; x(t)\big) \le -\big\langle F_t\big(x(t)\big), y - x(t) \big\rangle$$



Quantitative Roth's theorem for 3-term progressions in dense subsets of $\mathbb{Z}$

Theorem [Roth, …, Bourgain, Sanders, Bloom 2015]:

If $A \subseteq \{1,2,\dots,N\}$ contains no non-trivial 3-term arithmetic progressions, then

$$|A| \leq c\frac{(\log\log N)^4}{\log N}N$$

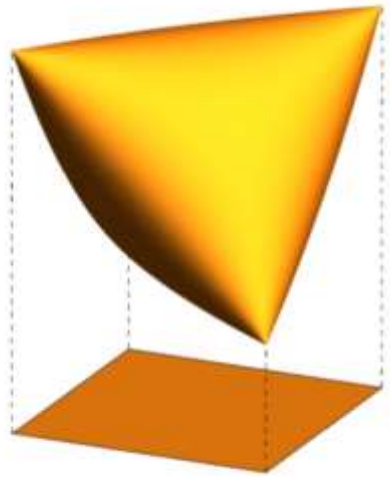
Bloom's main Fourier-analytic tool falls out of this setup [L 2016]:

$$\mathrm{K} = \{\, f : G \to \mathbb{R}_+ \mid \textstyle\sum_{x\in G} f(x) = 1 \,\} \qquad F_t(f) \in \{\, u_g : g \in G^* \}$$

$$f(0) = \frac{\mathbf{1}_G}{|G|} \qquad Y = \frac{\mathbf{1}_A}{|A|} \qquad \Phi(f) = \sum_{x\in G} f(x)\log f(x)$$

Spectrahedral lifts of the cut polytope [L-Raghavendra-Steurer 2015]

Denote $\mathrm{CUT}_n := \mathrm{conv}(x \otimes x \in \mathbb{R}^{n\times n} : \mathrm{x} \in \{0,1\}^n)$



Theorem: If Z is a spectrahedron that linearly projects to CUT_n, then

$$\dim(Z) \geq c^{n^{2/13}}, \qquad c > 1$$

$$\mathrm{K} = \left\{ D : \{-1,1\}^n \to \mathcal{S}_+^r \;\middle|\; \mathbb{E}_x \operatorname{tr}\big(D(x)\big) = 1 \right\}$$

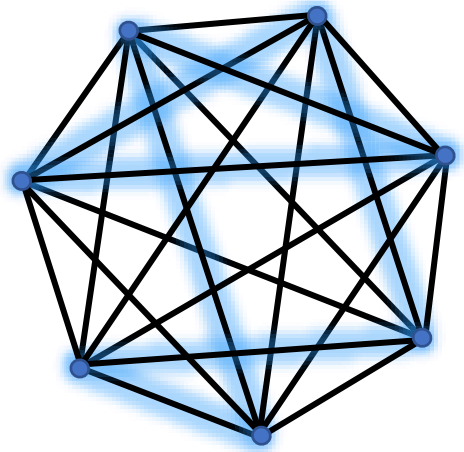
$D(0) =$ maximally (QC) mixed state

$$\Phi(D) = \mathbb{E}_x \operatorname{tr}(D(x) \log D(x))$$

$$F_t(D) \in \left\{ P(x)^2 : \deg\big(P(x)\big) \leq d \right\}$$

Twice-Ramanujan sparsifiers

[Batson, Spielman, Srivastava 2012; Allen-Zhu, Liao, Orecchia 2015]



Roughly: Every graph G can be spectrally approximated by a graph supported on a subset of $O(n)$ edges of G.

$$\mathrm{K} = \{ X \in \mathcal{S}_+^n : \mathrm{tr}(X) = 1 \}$$

$$\Phi(X) = -\mathrm{tr}\left(\sqrt{X}\right)$$

$$F_t(X) \in \{ x_e x_e^T : e \in E(G) \}$$

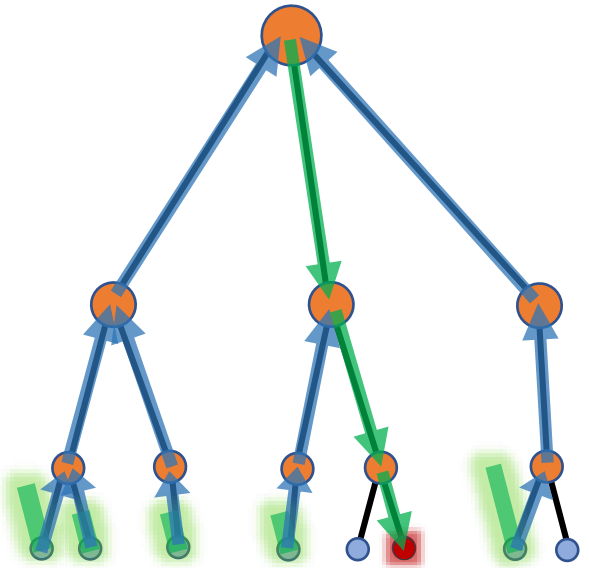
Solution of the (weak) randomized k-server conjecture

[Bubeck-Cohen-L-Lee-Madry 2017, L 2018]

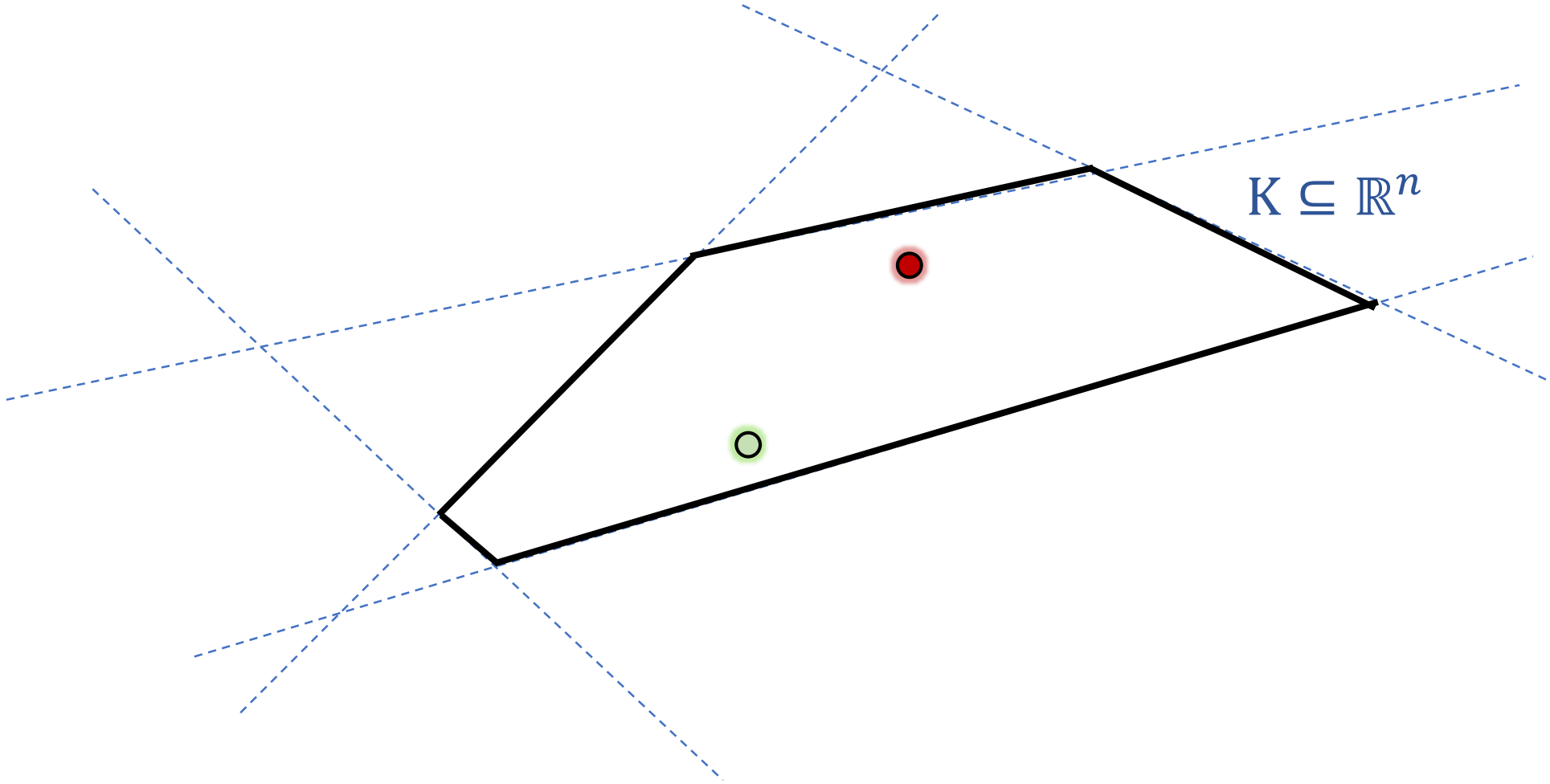
$\mathrm{K} =$ allocation polytope

$$\Phi(x) = \sum_{v \in V} w_v \sum_{i \geq 1} \left(x_{v,i} + \delta\right) \log \left(x_{v,i} + \delta\right), \ \delta \asymp \frac{1}{k}$$

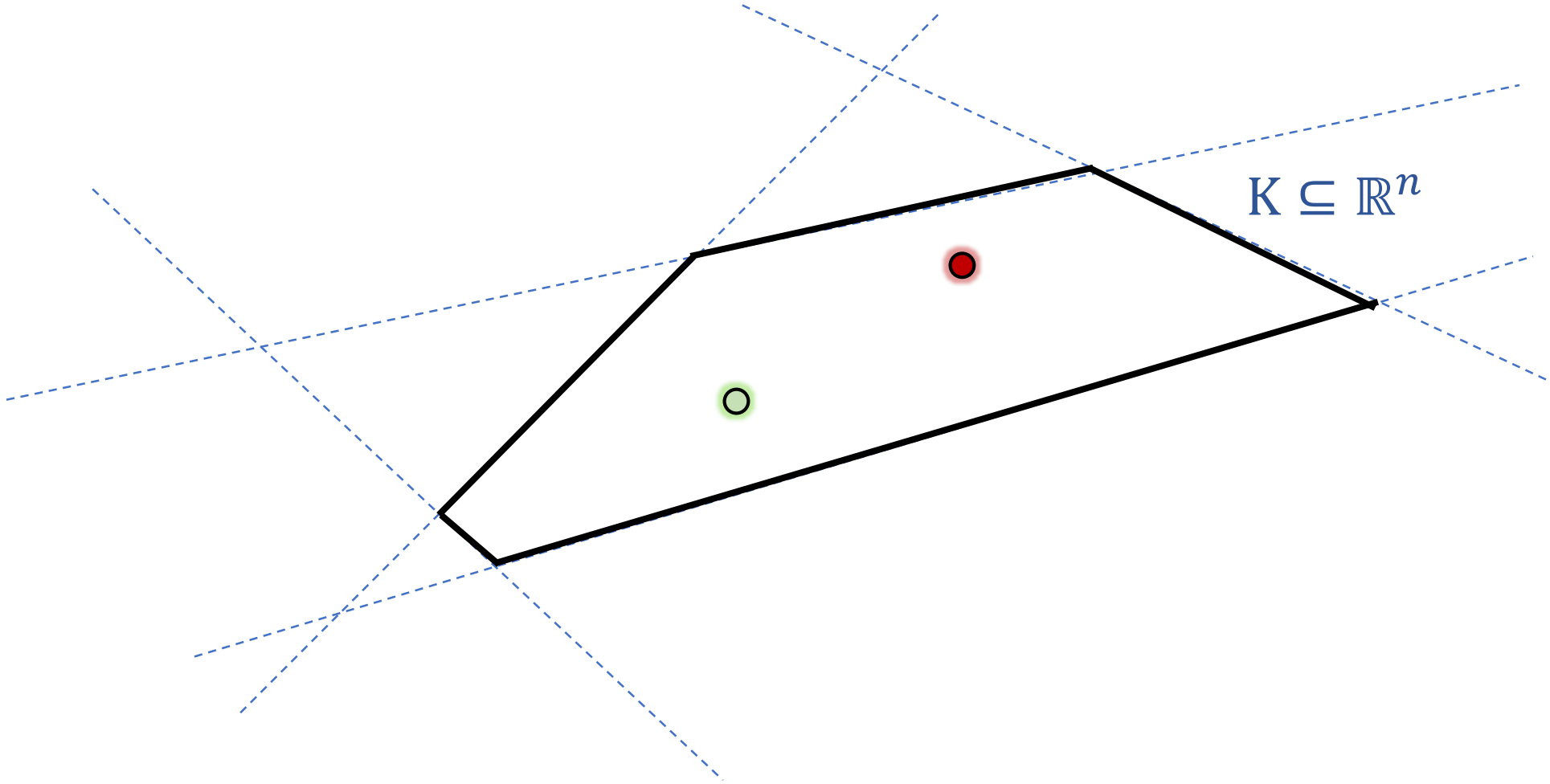
(multiscale shifted metric entropy)



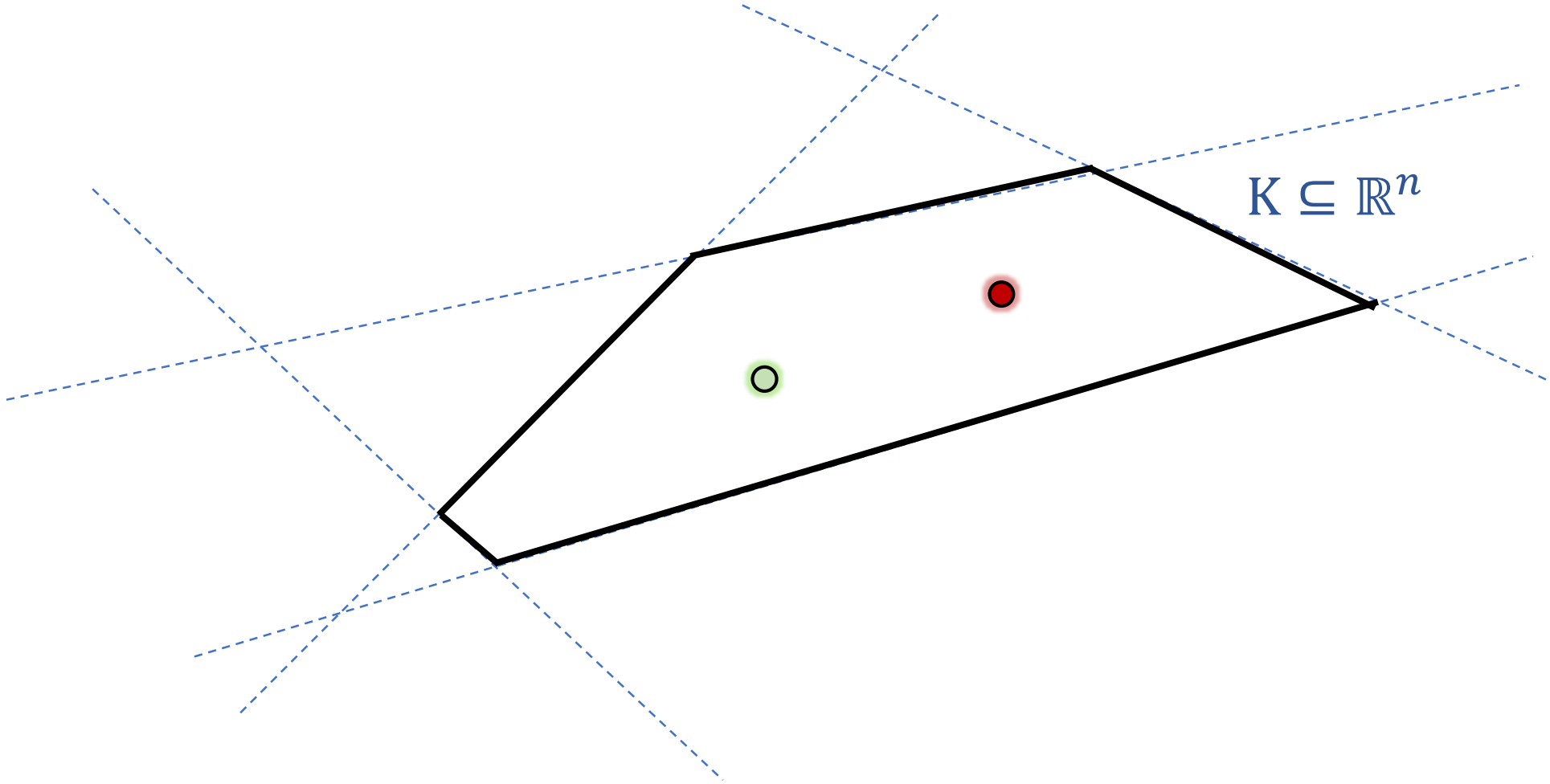
chasing a moving target (and bit complexity?)



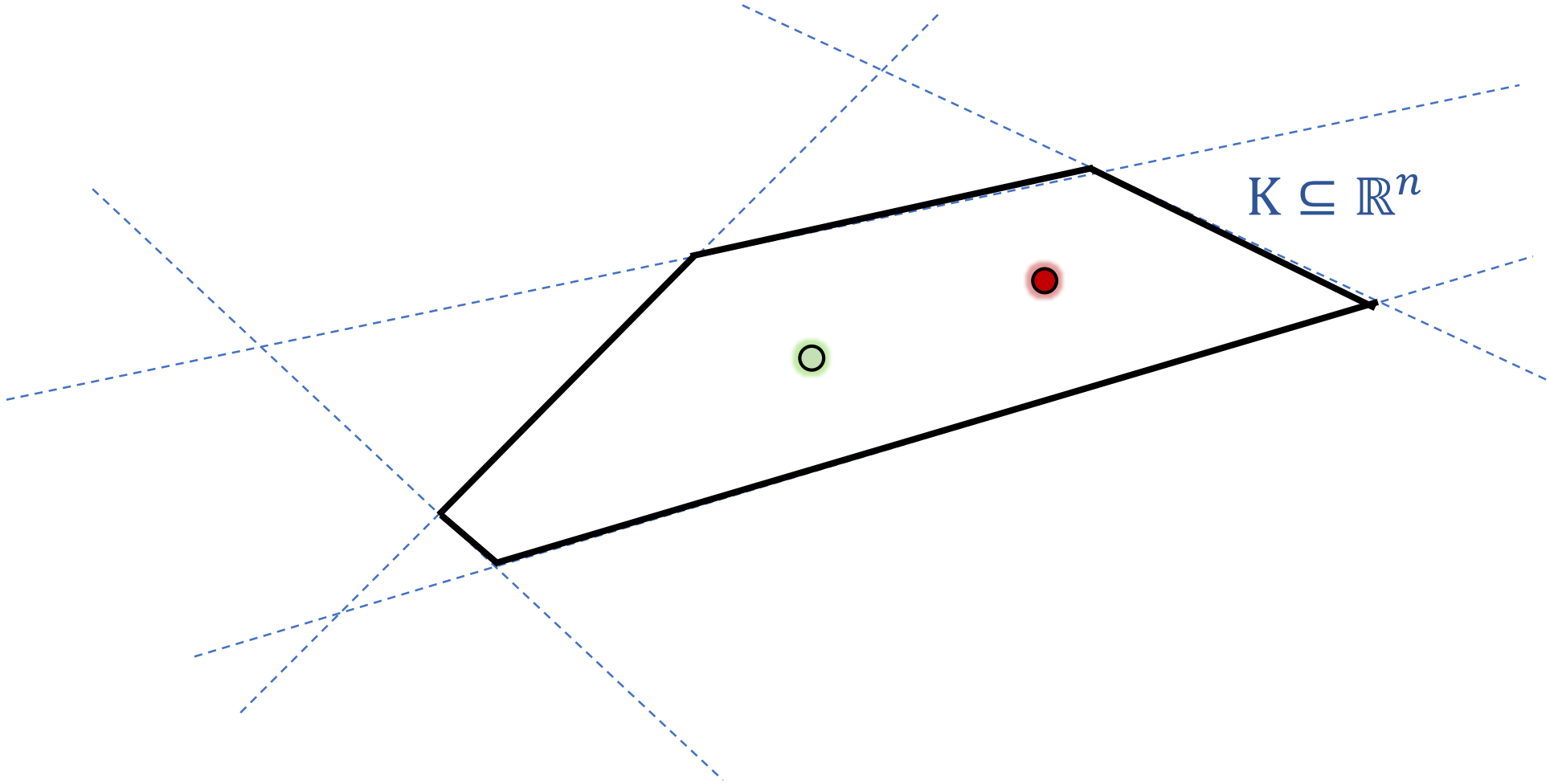
chasing a moving target (and bit complexity?)



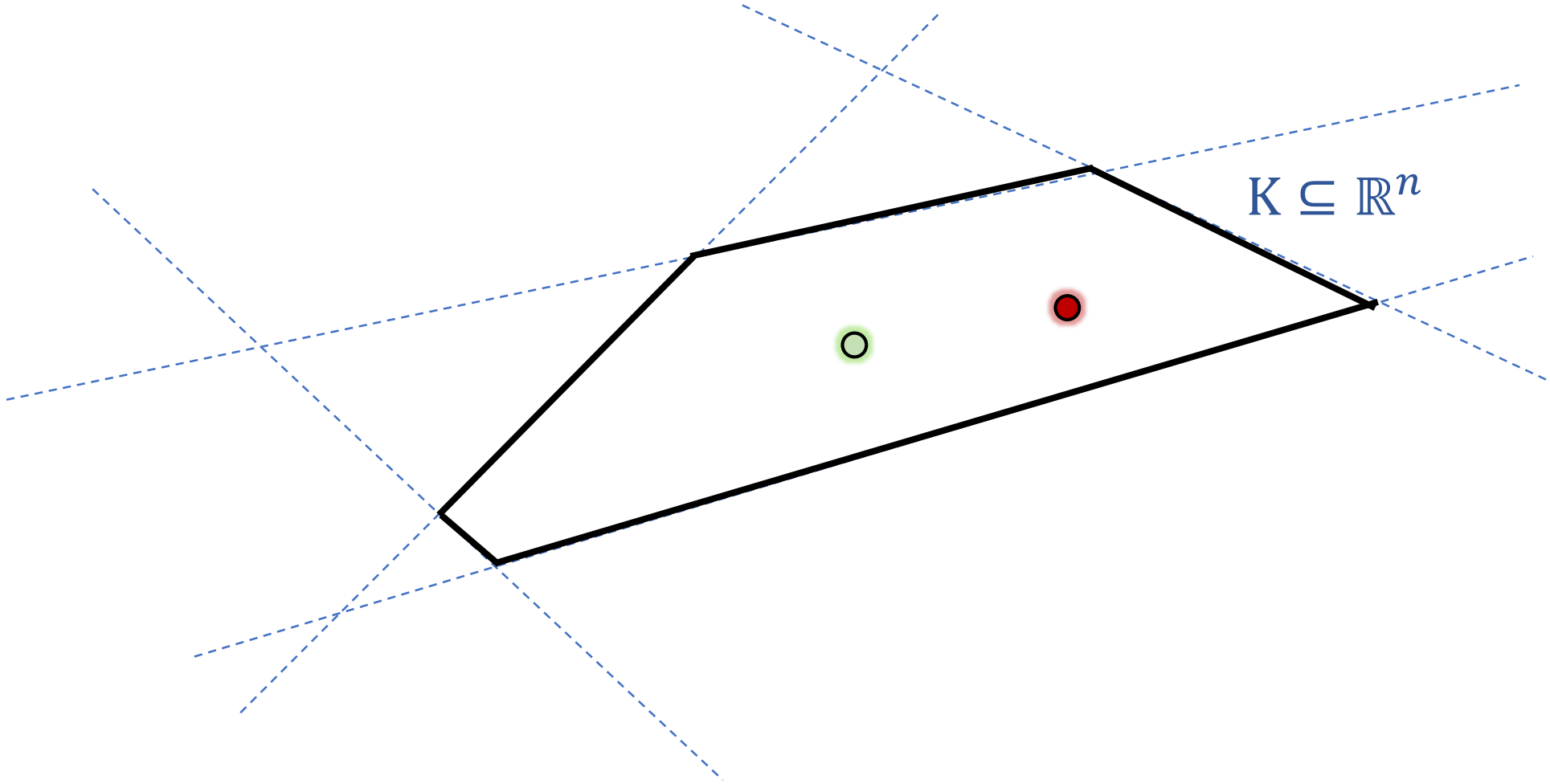
chasing a moving target (and bit complexity?)



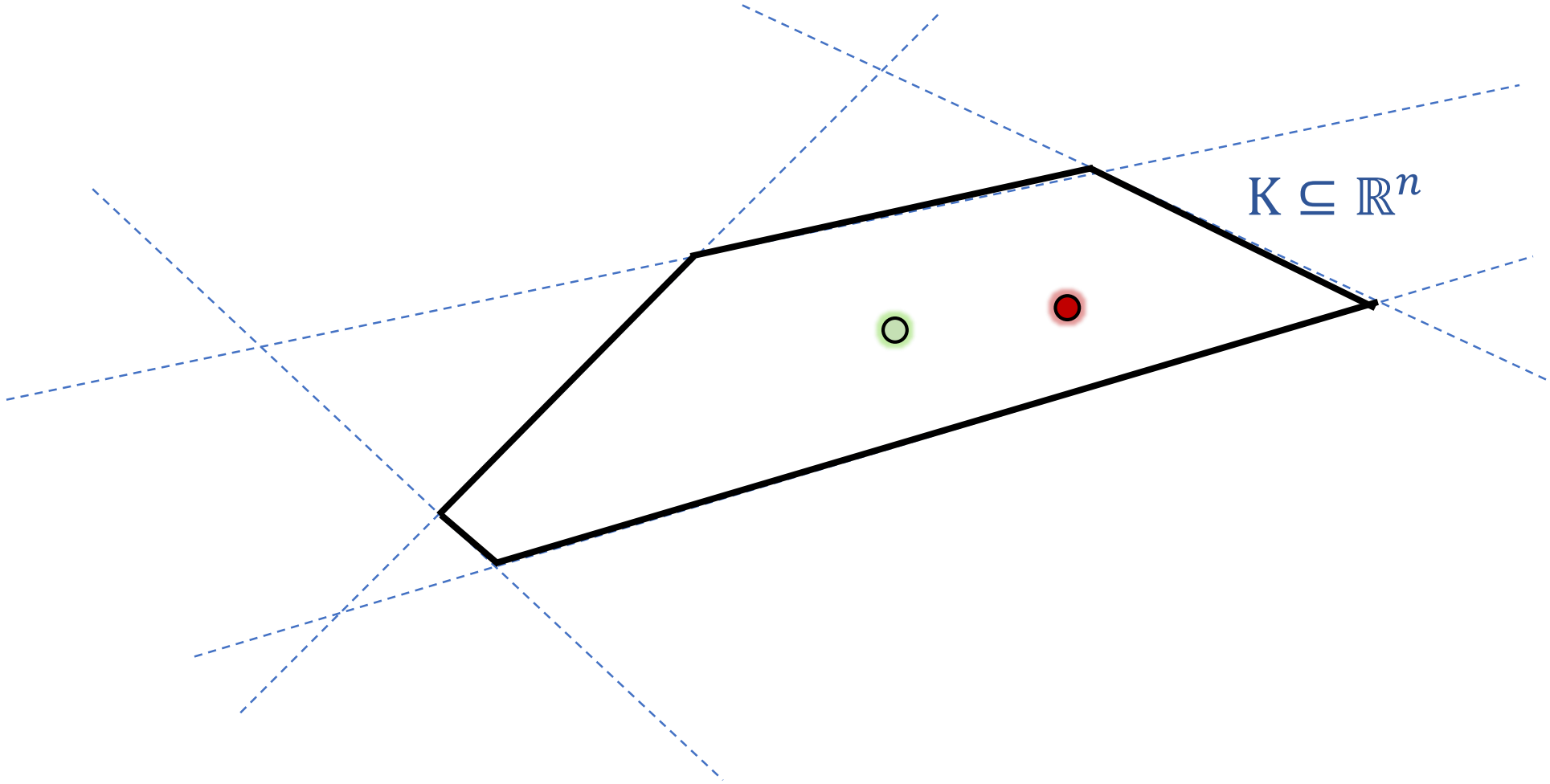
chasing a moving target (and bit complexity?)



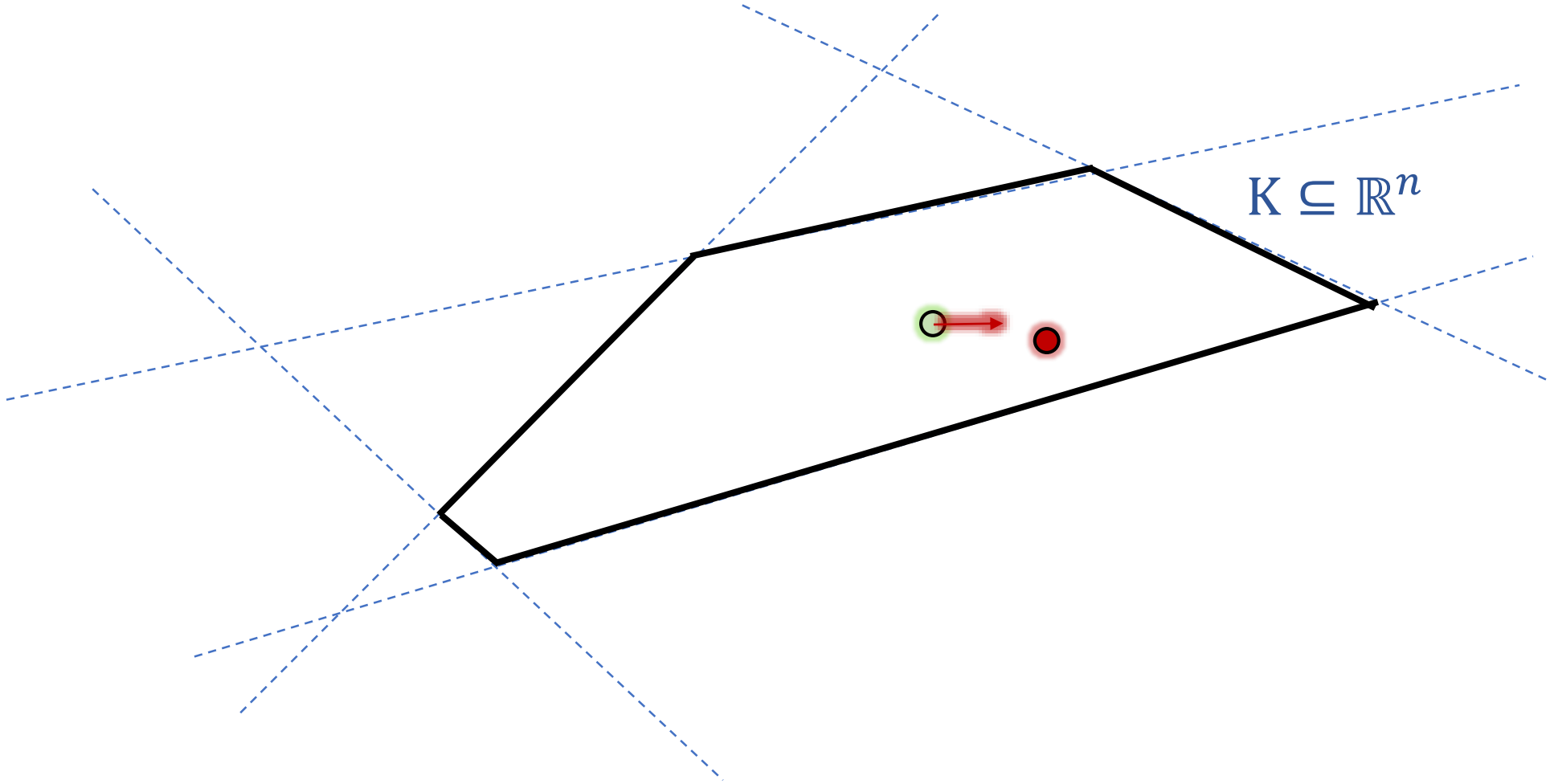
chasing a moving target (and bit complexity?)



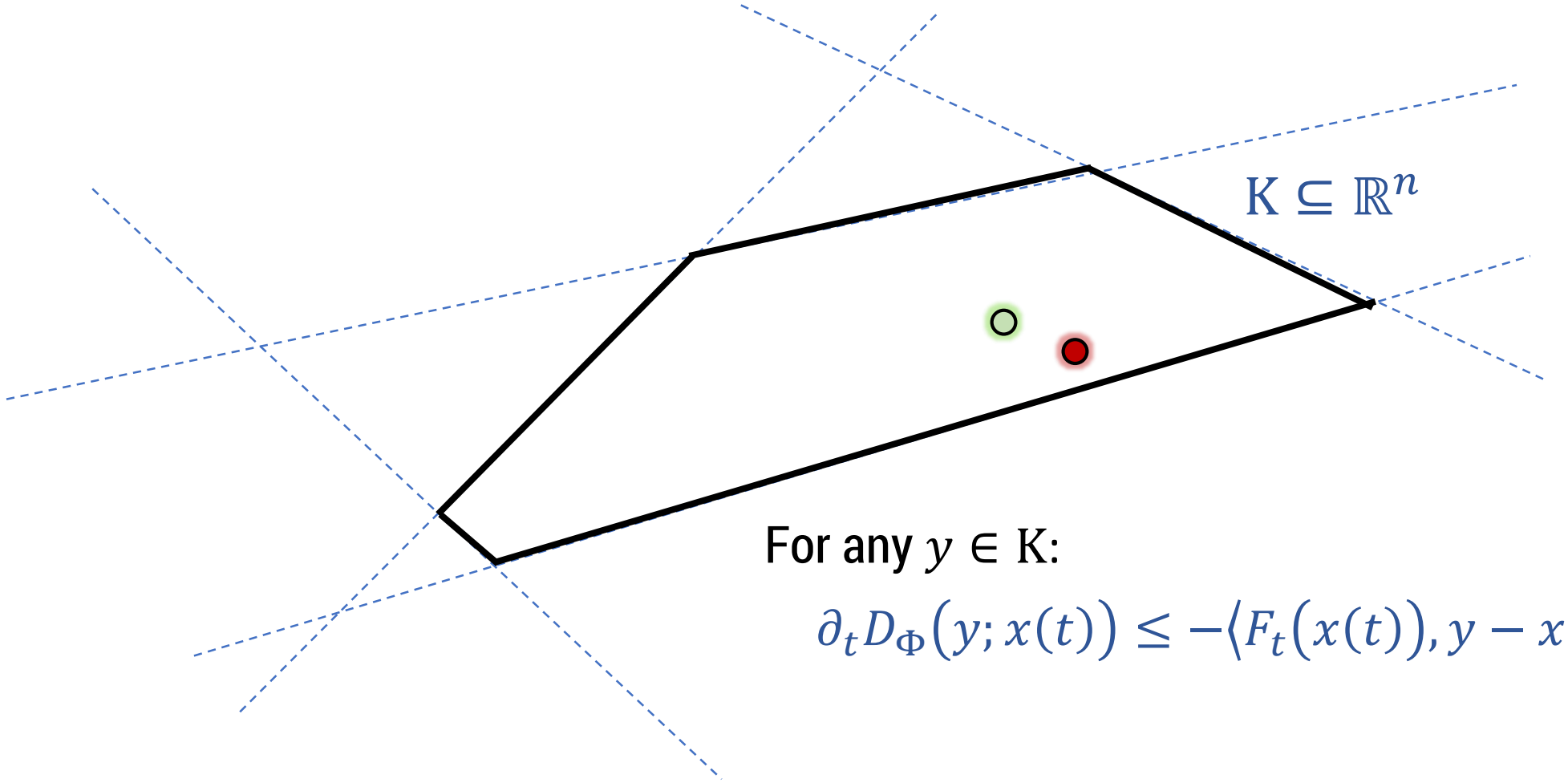
chasing a moving target (and bit complexity?)



chasing a moving target (and bit complexity?)



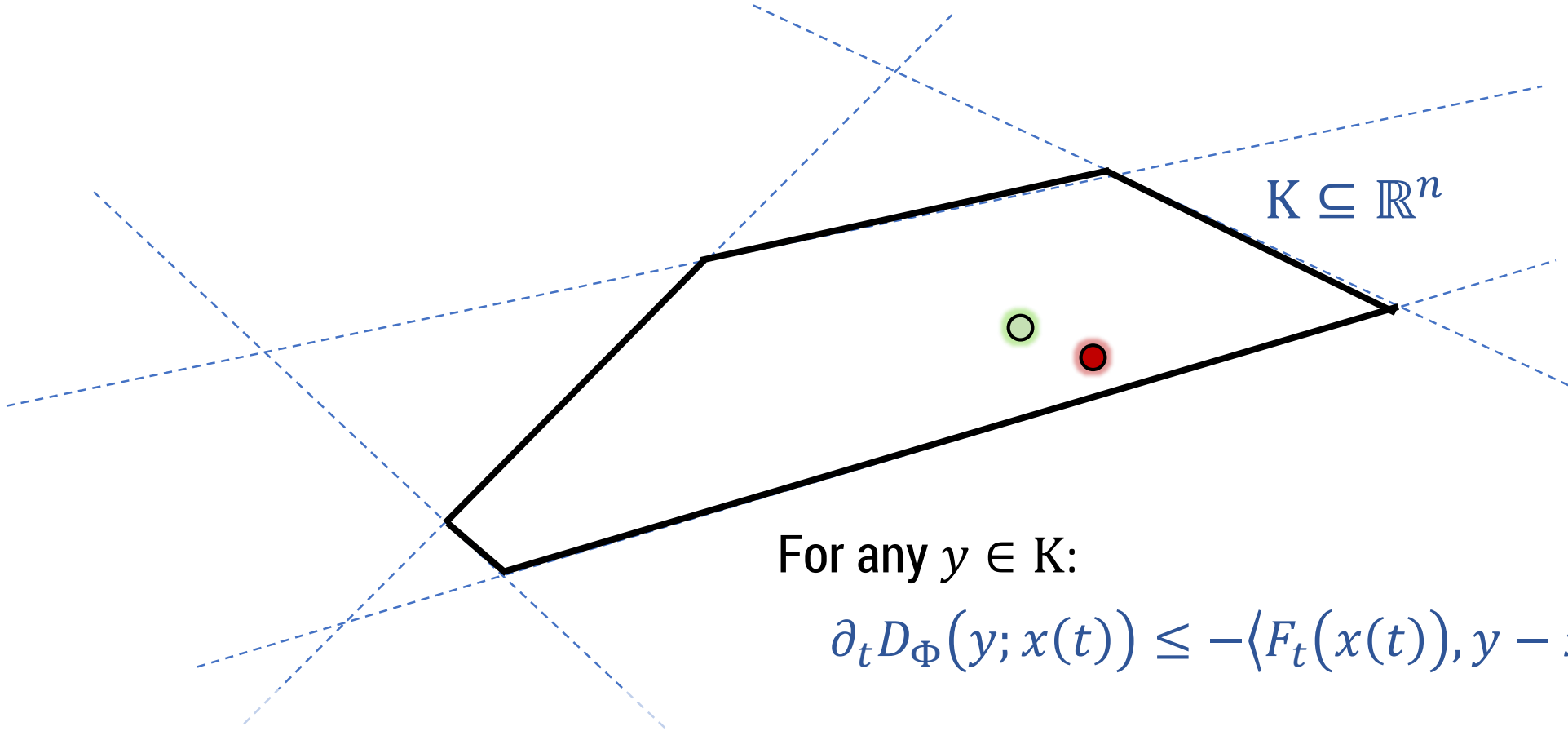
chasing a moving target (and bit complexity?)



For any $y \in \mathrm{K}$:

$$\partial_t D_\Phi\big(y; x(t)\big) \leq -\big\langle F_t\big(x(t)\big), y - x(t)\big\rangle$$

chasing a moving target (and bit complexity?)



For any $y \in \mathrm{K}$:

$$\partial_t D_\Phi\big(y; x(t)\big) \leq -\big\langle F_t\big(x(t)\big), y - x(t)\big\rangle$$

To chase a moving target, need smoothness of $y \mapsto D_\Phi(y; x)$!

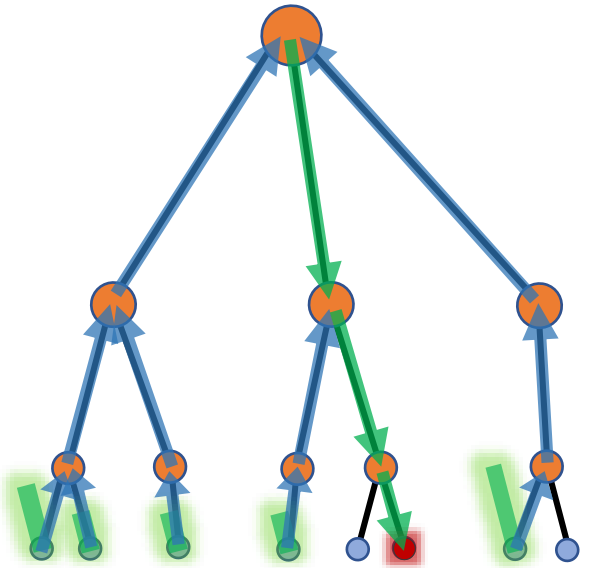
Solution of the (weak) randomized k-server conjecture

[Bubeck-Cohen-L-Lee-Madry 2017, L 2018]

$$\mathrm{K} = \text{ allocation polytope}$$

$$\Phi(x) = \sum_{v \in V} w_v \sum_{i \geq 1} \left(x_{v,i} + \delta\right) \log \left(x_{v,i} + \delta\right), \; \delta \asymp \frac{1}{k}$$

(multiscale shifted metric entropy)



geometric BL data (Gaussian version)

Consider subspaces $E_1, E_2, \ldots, E_m \subseteq \mathbb{R}^n$. Let P_i denote the orthogonal projection of E onto E_i, and suppose that $c_1, c_2, \ldots, c_m > 0$ are numbers so that the **frame condition** holds:

$$\sum_{i=1}^{m} c_i P_i = \mathrm{id}_E$$

Let Z denote a standard Gaussian on $\mathbb{R}^n$, and Z_i a standard Gaussian on E_i. Then for any random vector $X \in \mathbb{R}^n$:

$$D(X \mid Z) \geq \sum_{i=1}^{m} c_i \, D(\, P_i X \mid Z_i \,)$$

[Ball / Barthe / Carlen & Cordero-Erausquin]

entropy optimal drifts

Consider a probability μ on $\mathbb{R}^n$, γ_n = standard Gaussian measure

Let $\{ B_t : t \in [0,1] \}$ be a Brownian motion with $B_0 = 0$ (so $B_1 \sim \gamma_n$)

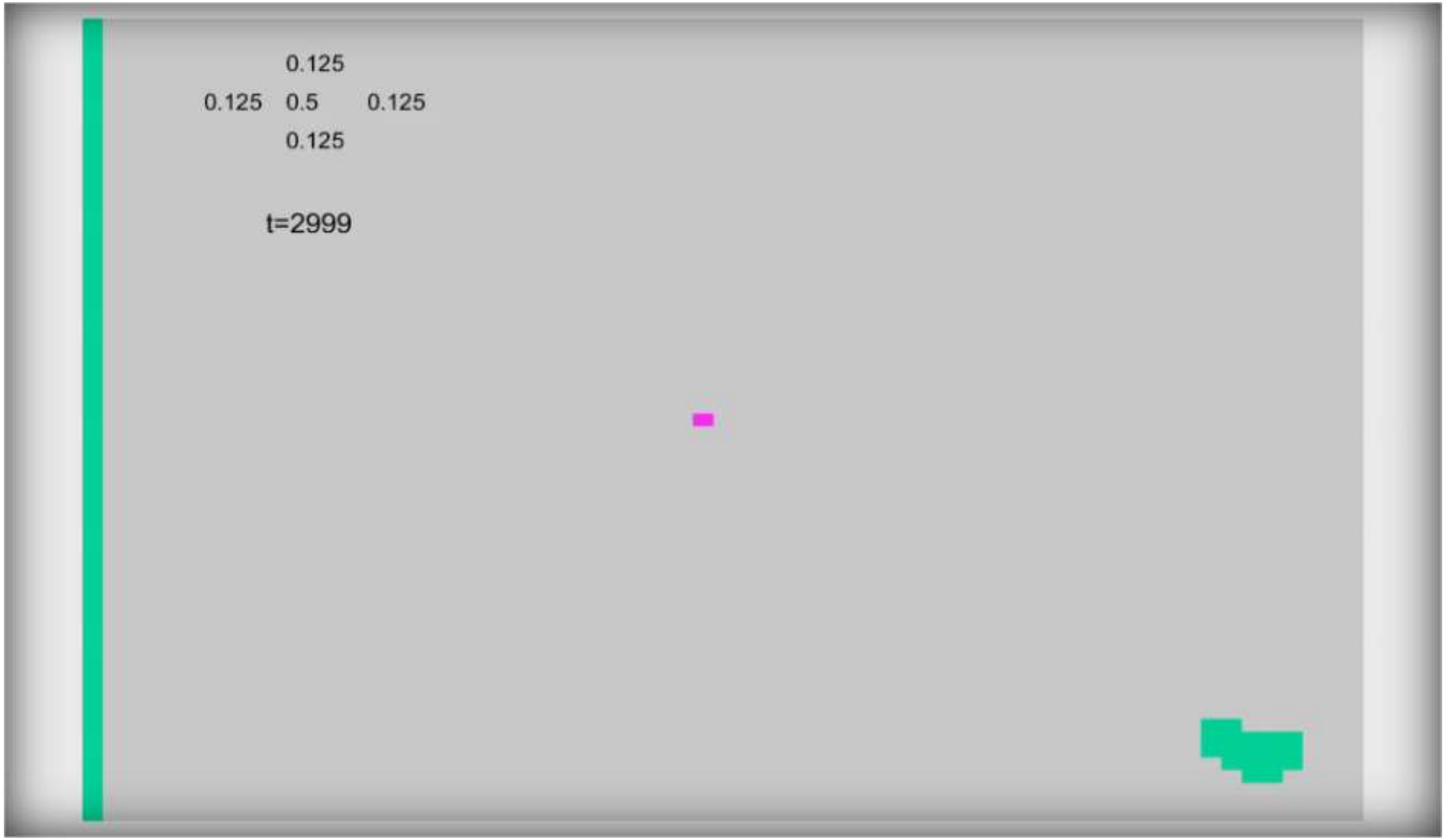
Consider the family $\{v_t : t \in [0,1] \}$ of all adapted drifts so that if $X_0 = 0$ and

$$dX_t = dB_t + v_t \, dt$$

then X_1 has law μ.

[Follmer-Borell]: $$D(\mu \mid \gamma_n) = \frac{1}{2} \min_{\{v_t\}} \int_0^1 \mathbb{E} \, \|v_t\|^2 \, dt$$

entropy optimal drifts



geometric BL data (Gaussian version)

Subspaces $E_1 \dots, E_m \subseteq \mathbb{R}^n$

$$\sum_{i=1}^{m} c_i P_i = \mathrm{id}_E$$

Let Z denote a standard Gaussian on $\mathbb{R}^n$, and Z_i a standard Gaussian on E_i. For any random vector $X \in \mathbb{R}^n$:

$$D(X \mid Z) \geq \sum_{i=1}^{m} c_i \, D(P_i X \mid Z_i)$$

[Lehec 2010]

Proof:

Let $\{v_t\}$ be the energy-optimal drift so that $dX_t = dB_t + v_t \, dt$ has $X_1 \sim X$

By Follmer-Borel:

$$D(X \mid Z) = \frac{1}{2} \int_0^1 \mathbb{E} \, \|v_t\|^2 \, dt$$

$$= \frac{1}{2} \int_0^1 \sum_{i=1}^{m} c_i \, \mathbb{E} \, \|P_i v_t\|^2 \, dt$$

$$\geq \sum_{i=1}^{m} c_i \, D(P_i X \mid Z_i)$$

how do operator scaling (or non-geometric BL) fit?

Rank one operator scaling (Barthe, Forster):

Suppose $x_1, x_2, \ldots, x_n \in \mathbb{R}^k$ are unit vectors so that every subset of k vectors is linearly independent. Then there is a linear mapping $A : \mathbb{R}^k \to \mathbb{R}^k$ such that

$$\sum_{i=1}^{n} \frac{(Ax_i)(Ax_i)^T}{\|Ax_i\|^2} = \frac{n}{k} \mathrm{Id}_k$$

Define the determinantal measure

$$D_S = \det\left(\sum_{i \in S} x_i x_i^T\right) / \det\left(\sum_{i=1}^{n} x_i x_i^T\right)$$

(measure on $S \subseteq [n], |S| = k$)

Minimize: $\sum_{|S|=k} p_S \log \frac{p_S}{D_S}$

over probability measures $\{p_S : |S| = k\}$ satisfying

$$\sum_{S : i \in S} p_S = \frac{k}{n} \quad \forall i = 1, \ldots, n$$