Faster convex optimization
Simulated annealing & Interior point

Elad Hazan

PRINCETON UNIVERSITY

Joint work with Jacob Abernethy – U MICH

Convex optimization

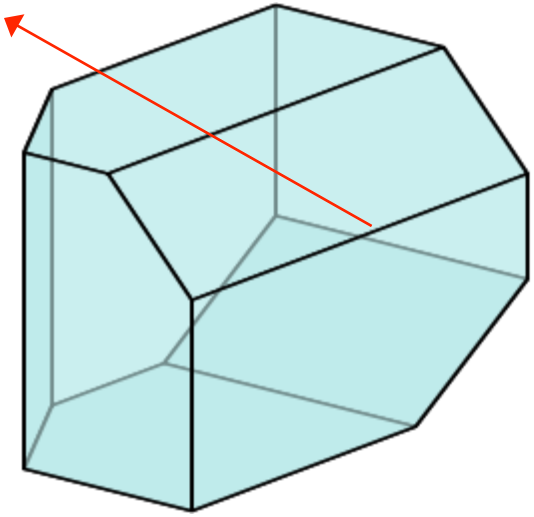
fundamental problem of optimization:

minimize a convex (linear) function over a convex set

$$\min_{x \in \mathcal{K}} f(x)$$

$$\Downarrow$$

$$\min_{x \in \mathcal{K} \cap \{f(x) \leq t\}} t$$



Convex optimization

A few examples

1. ERM/stochastic minimization for machine learning

2. Semi-definite programming for block model, 3D-reconstruction

3. Bayesian inference relaxations.

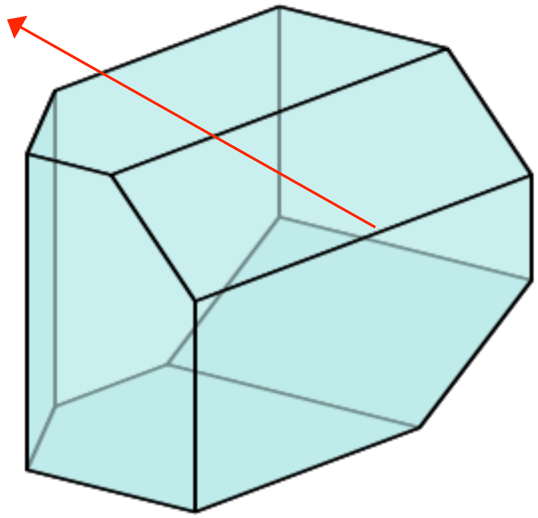
4. Matrix completion problems, sparse reconstruction, nuclear norm minimization, metric learning….

Convex optimization

fundamental problem of optimization:

minimize a convex (linear) function over a convex set

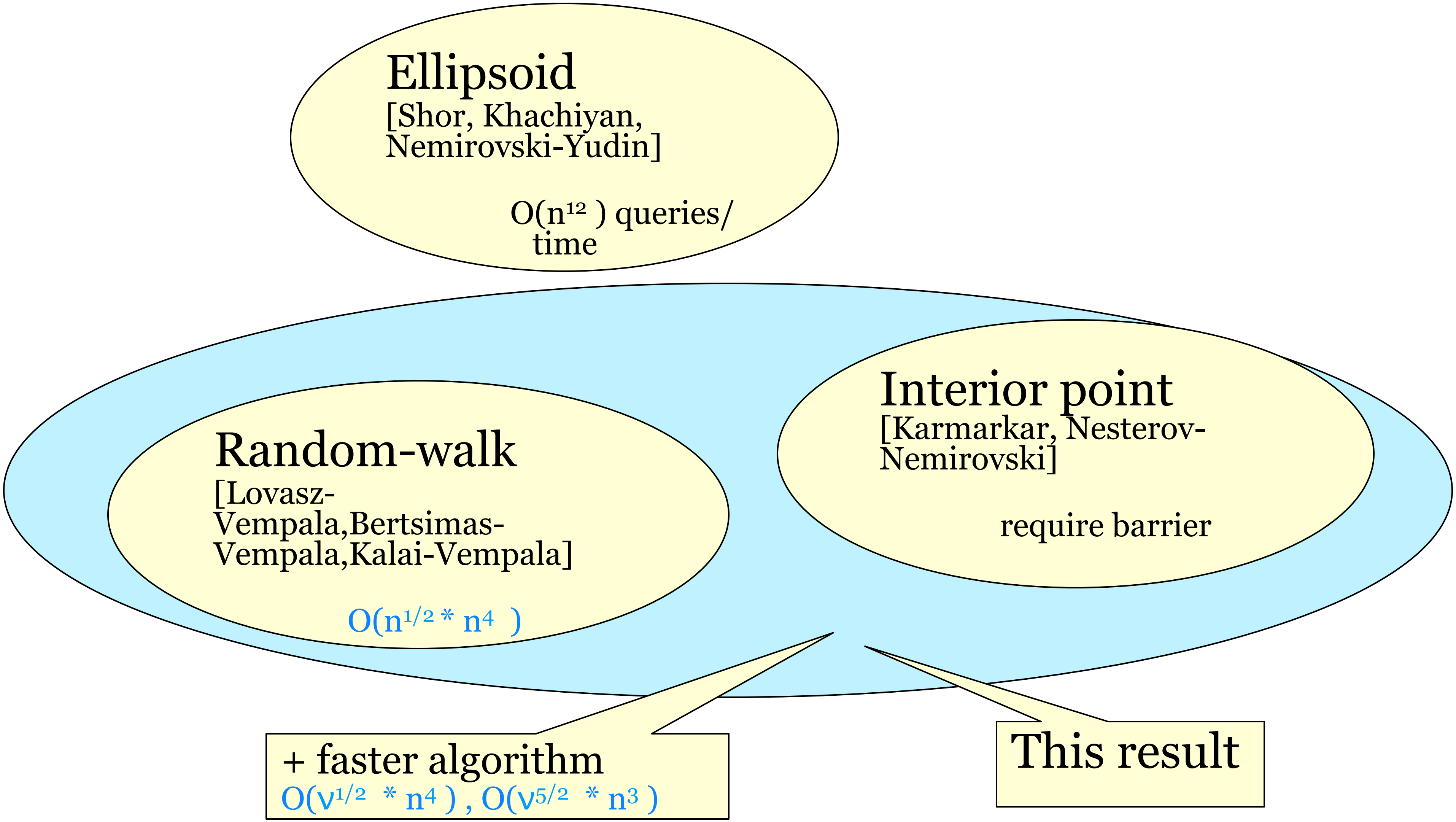
$$\min_{x \in \mathcal{K}} c^\top x$$



Convex set given by:

1. linear constraints (LP)
2. Semi-definite constraints
3. Separation oracle
4. Membership oracle

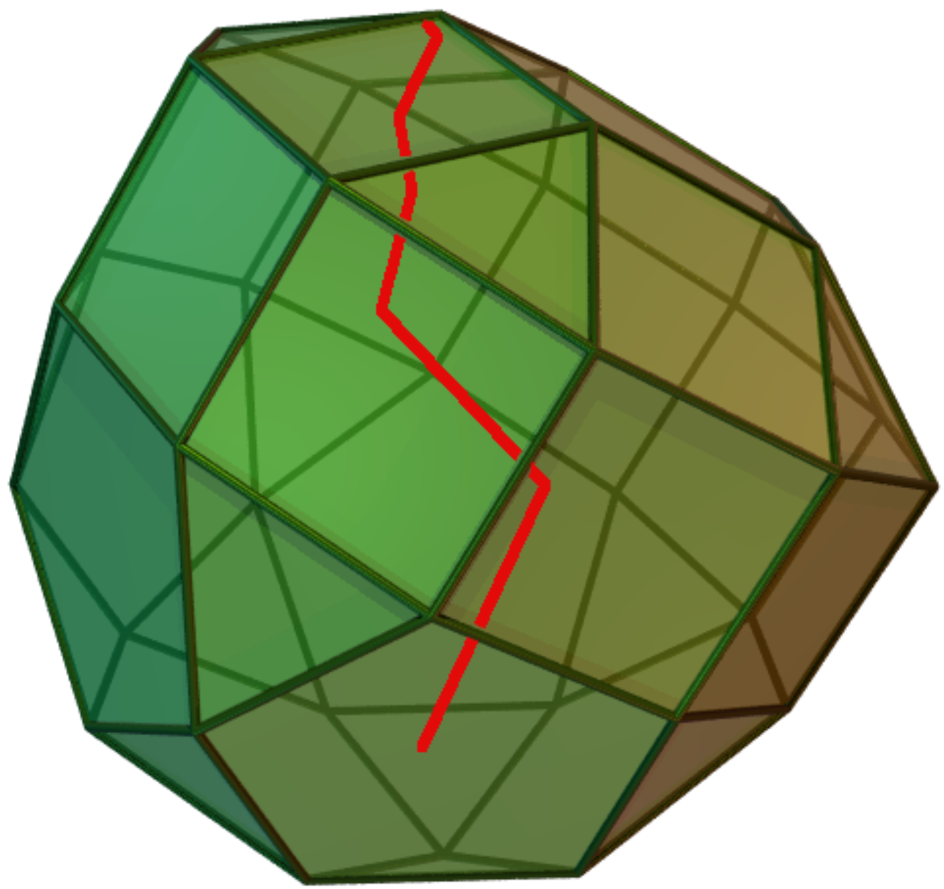
Polynomial time convex opt



Agenda

1. Mini tutorial on IPM
2. Mini tutorial on SA
3. The equivalence of SA and IPM
4. How to get faster convex opt

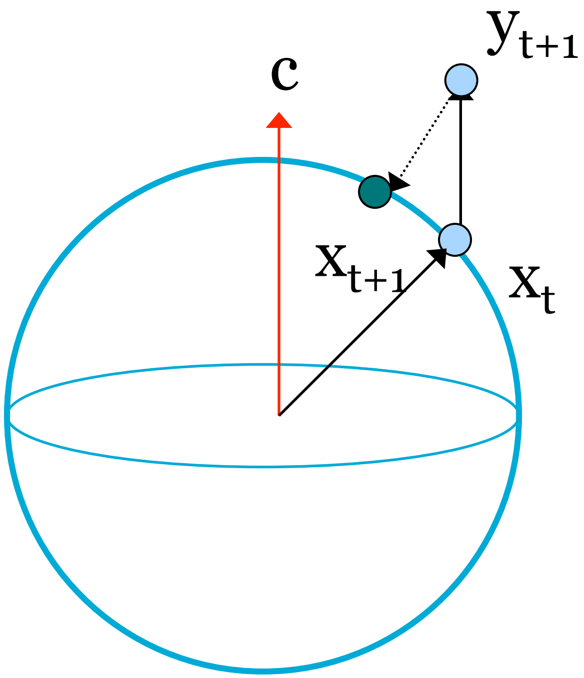
Interior point methods: mini-tutorial



Gradient descent

move in the direction of the

steepest decrease (-gradient)



$$y_{t+1} = x_t - \eta \nabla f(x_t)$$

$$x_{y+1} = \text{project}_{\mathcal{K}}[y_{t+1}]$$

Projection –

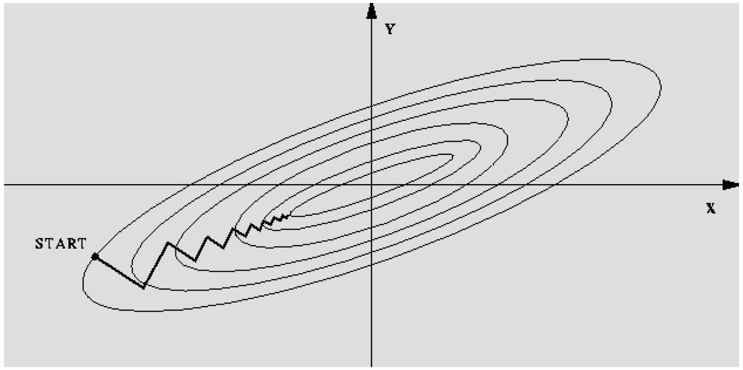
Can be as hard as the original problem!

$$\min \|x - y\|^2$$

$$x \in \mathcal{K}$$

steepest decrease direction

– no information on curvature!



Newton's method ("smart gradient"):

$$y_{t+1} = x_t - \eta[\nabla^2 f(x_t)]^{-1} \nabla f(x_t)$$
$$x_{y+1} = \text{project}_{\mathcal{K}}[y_{t+1}]$$

For quadratic functions: solution in 1 step

Interior point methods

Avoid projections → remain in the interior always

Add curvature → add a "super-smooth" barrier function

$\min c^T x$
$A_1 \; x \; - b_1 \leq 0$
$\ldots$
$A_m \, x \; - b_m \leq 0$
$x \sim R^n$

$\min \; c^T x - \; \sum_i \log(b_i - A_i \, x)$
$x \sim R^n$

R(x)

Barrier function

Self-concordant barrier

Allow polynomial-time convex optimization [Nesterov, Nemirovski 1994]. Properties:

Self-concordance parameter

1. as x-> ϑK, R(x) → ∞

2.
$$\nabla^3 R(x)[h,h,h] \leq 2(\nabla^2 R(x)[h,h])^{3/2}$$
$$\nabla R(x)[h] \leq \sqrt{\nu \nabla^2 R(x)[h,h]}$$

Property 1: remain in the interior
Properties 2: ensure that Newton's method can exploit curvature

Linear programming:

$$Ax \leq b \;\Rightarrow\; R(x) = \sum_i \log(A_i x - b_i)$$

Interior point methods

But now:

Objective is skewed – barrier distorts

$$\min_{x \in \mathcal{K}} c^\top x \quad \Rightarrow \quad \min_{x \in \mathcal{R}^d} \left\{ c^\top x + R(x) \right\}$$

Interior point methods

→

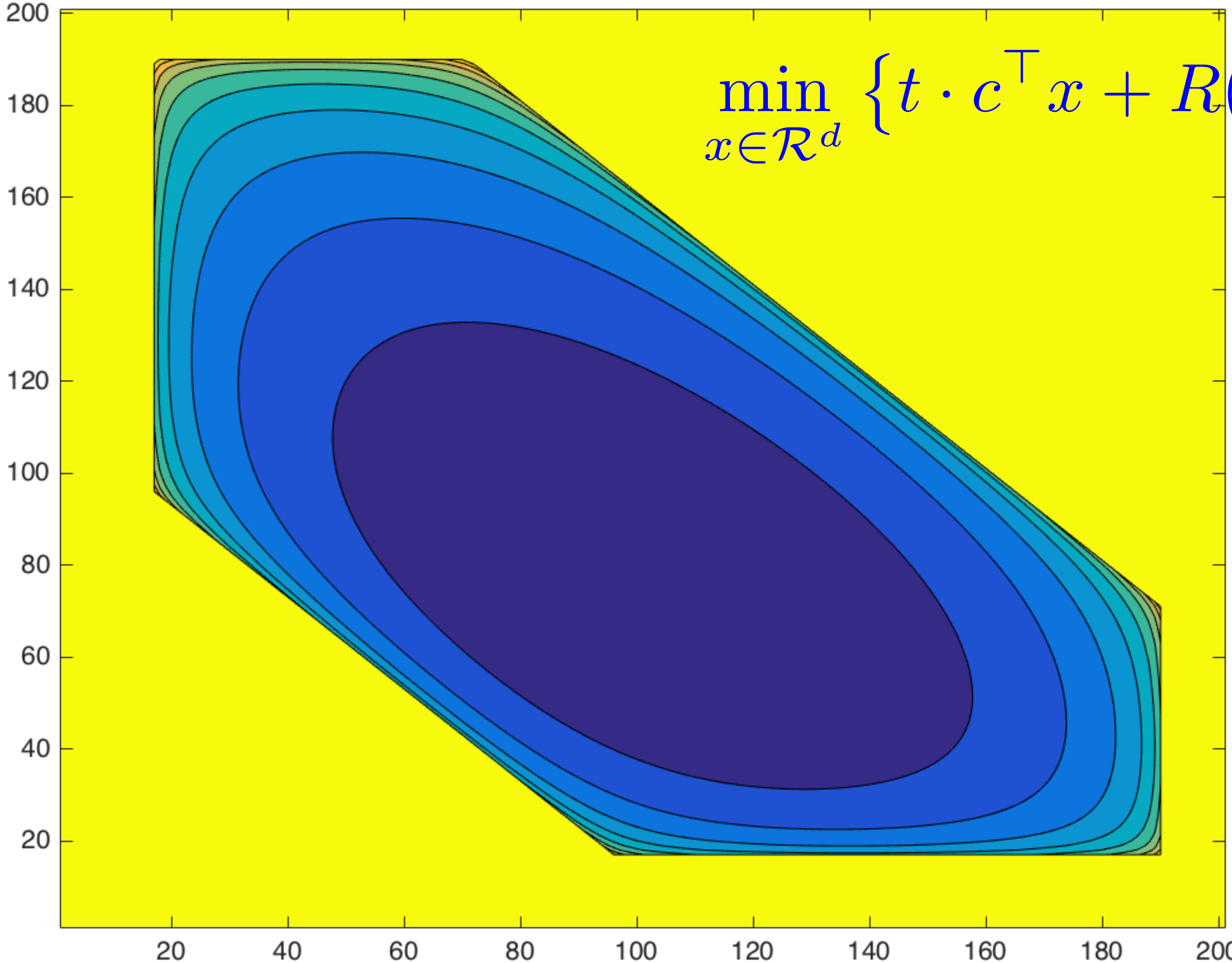
Add & change barrier scale

$$\min_{x \in \mathcal{K}} c^\top x \quad \Rightarrow \quad \min_{x \in \mathcal{R}^d} \left\{ t \cdot c^\top x + R(x) \right\}$$

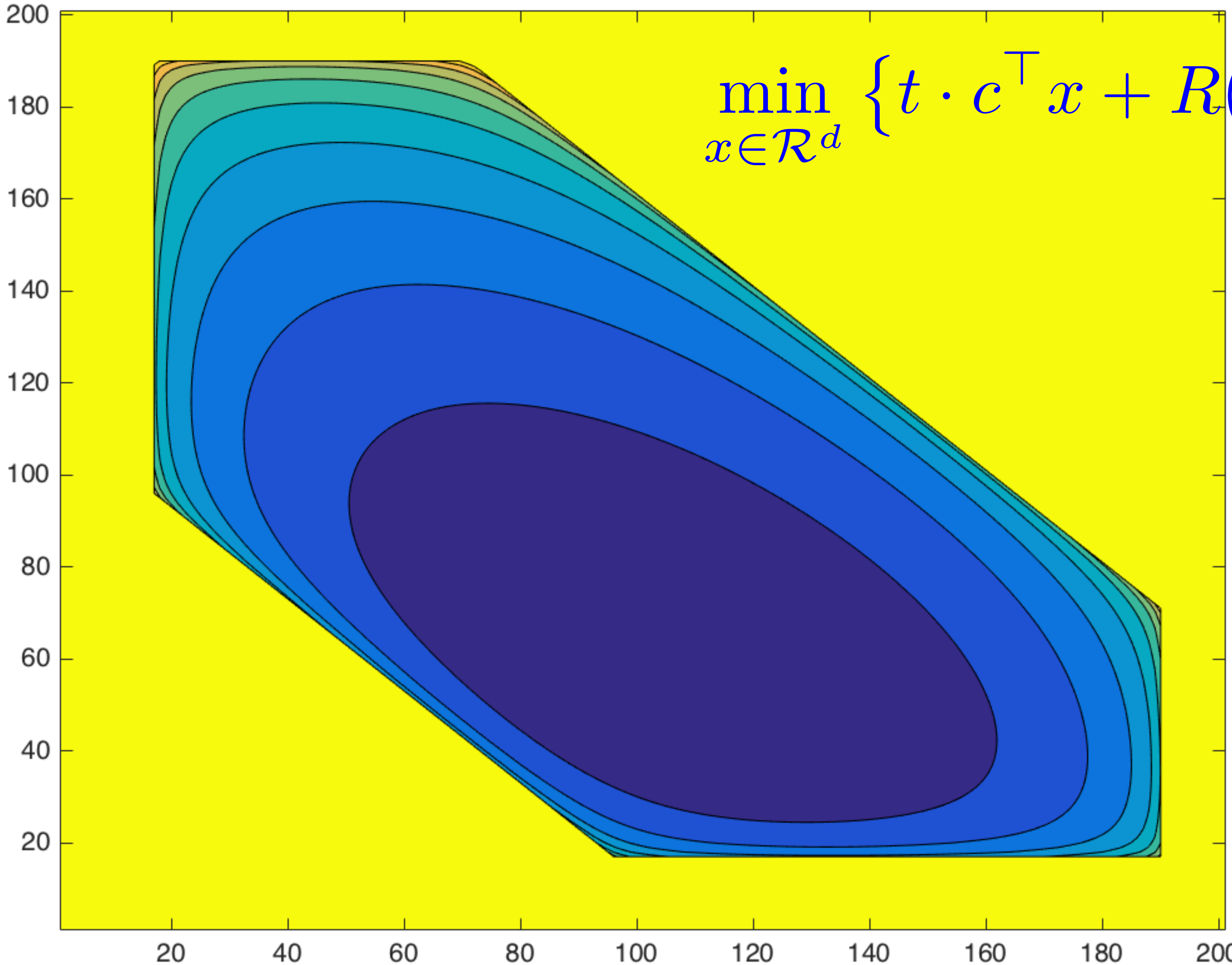
$$t :\sim 0 \Rightarrow \infty$$

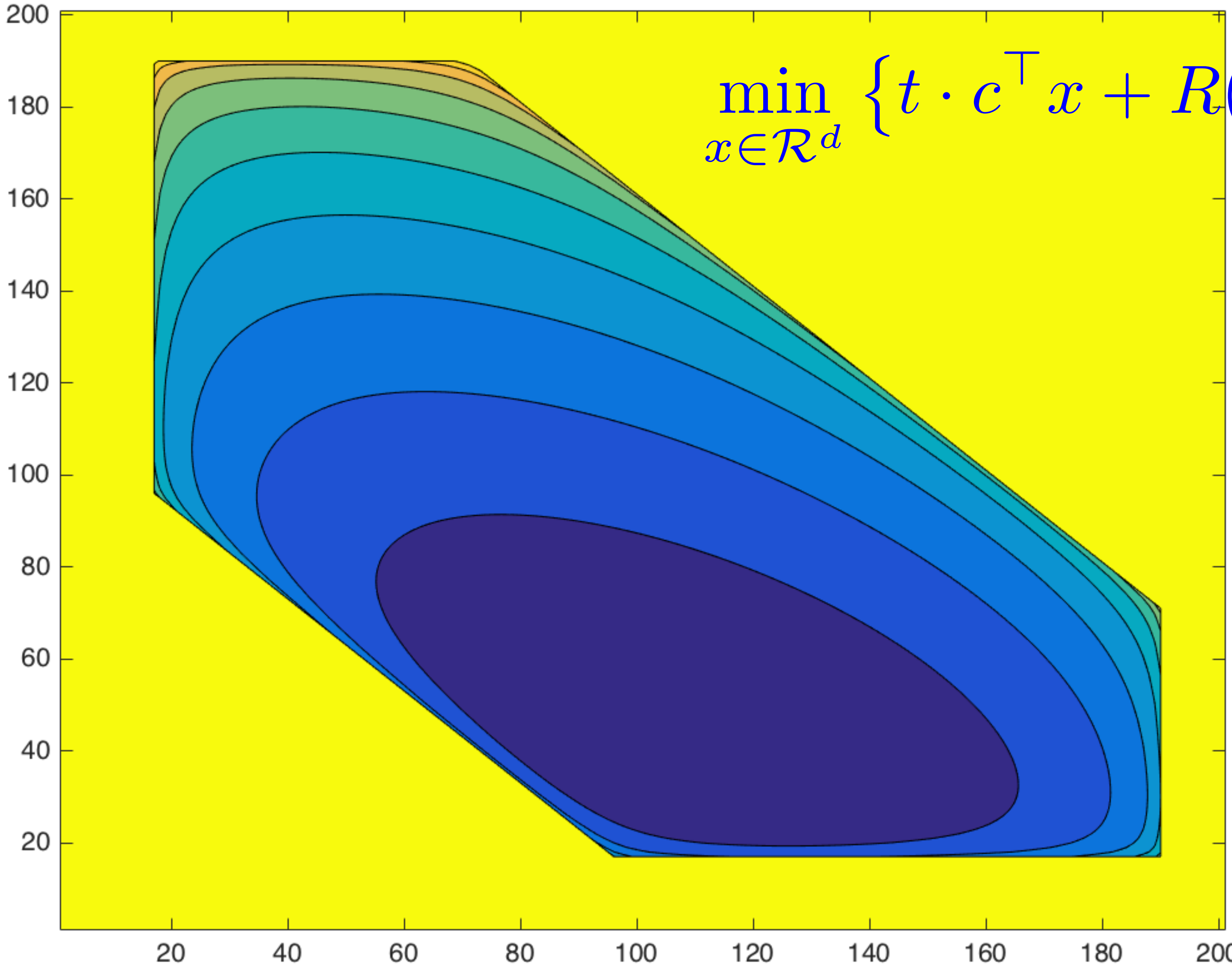
$$t_{k+1} = t_k \left(1 + \frac{1}{\sqrt{\nu}}\right)$$

$$\min_{x \in \mathcal{R}^d} \left\{ t \cdot c^\top x + R(x) \right\}$$

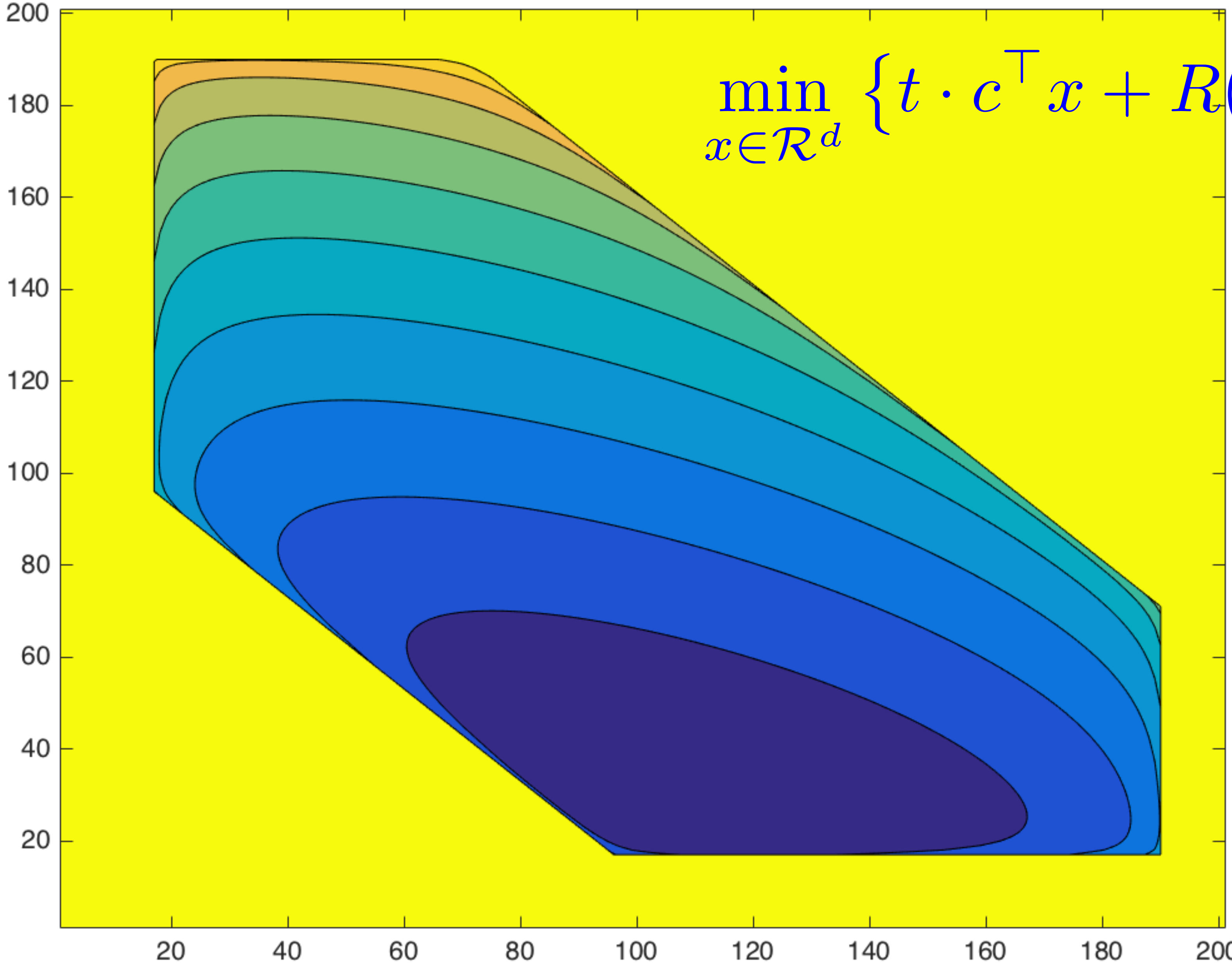


$$\min_{x \in \mathcal{R}^d} \left\{ t \cdot c^\top x + R(x) \right\}$$

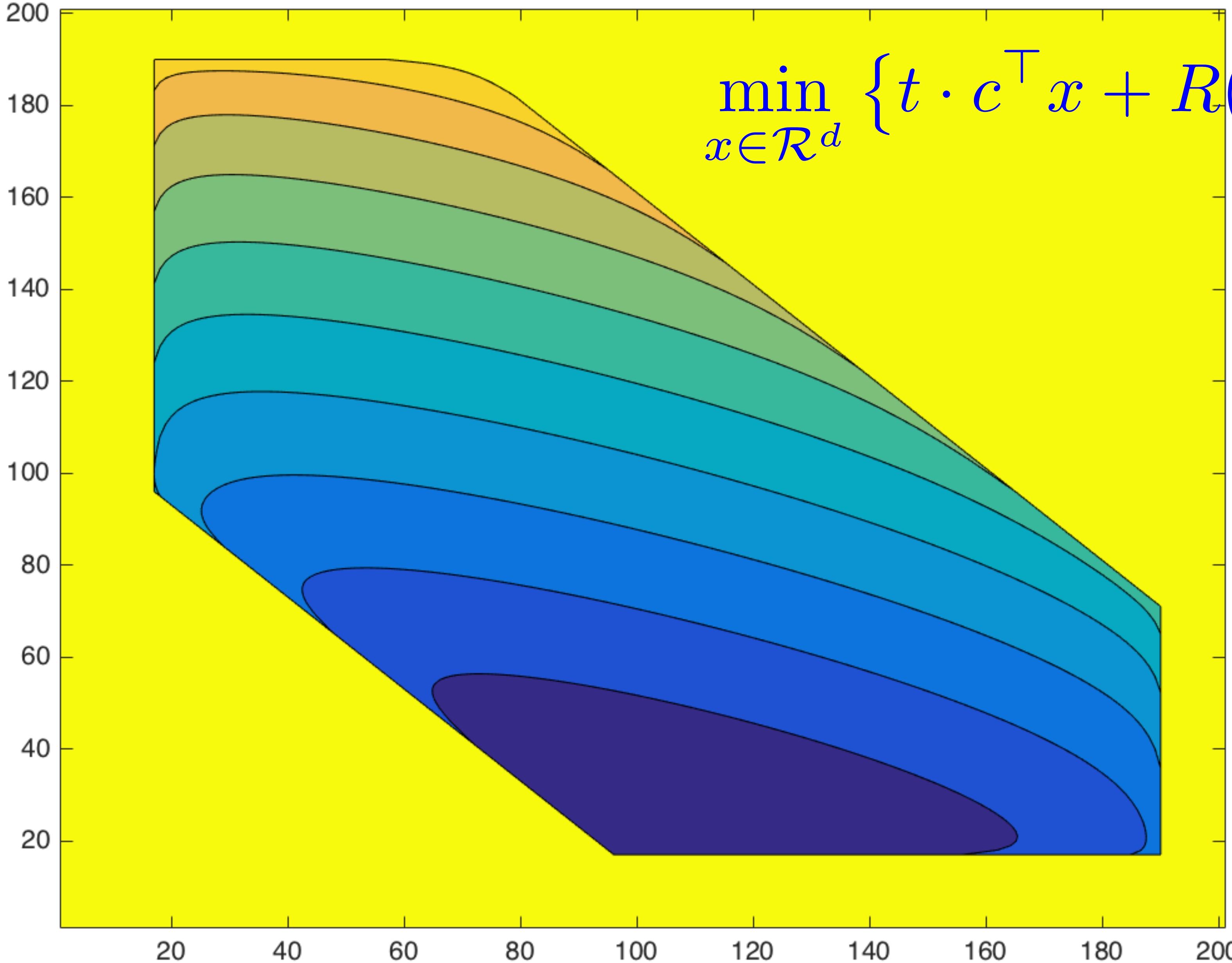




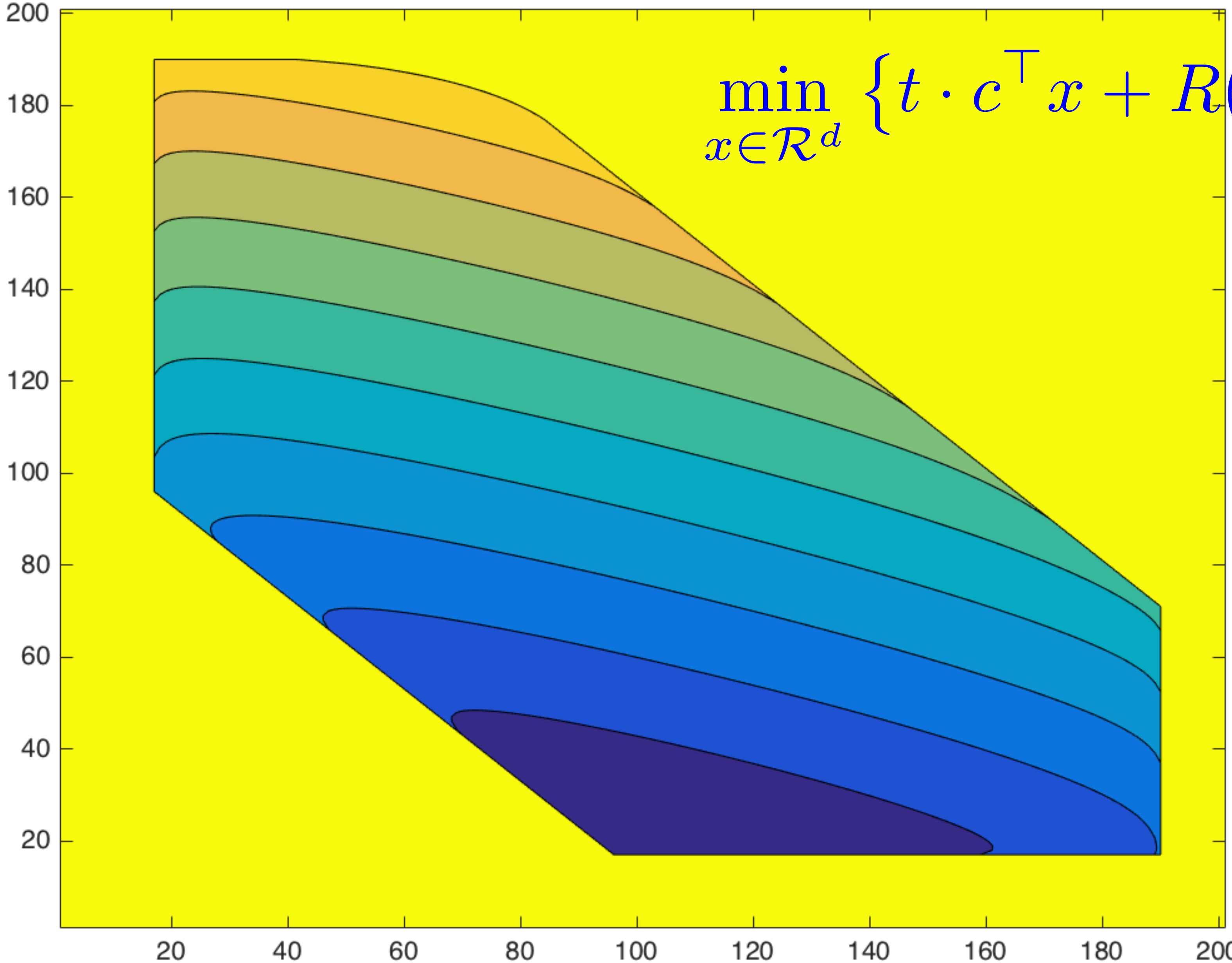
$$\min_{x \in \mathcal{R}^d} \left\{ t \cdot c^\top x + R(x) \right\}$$



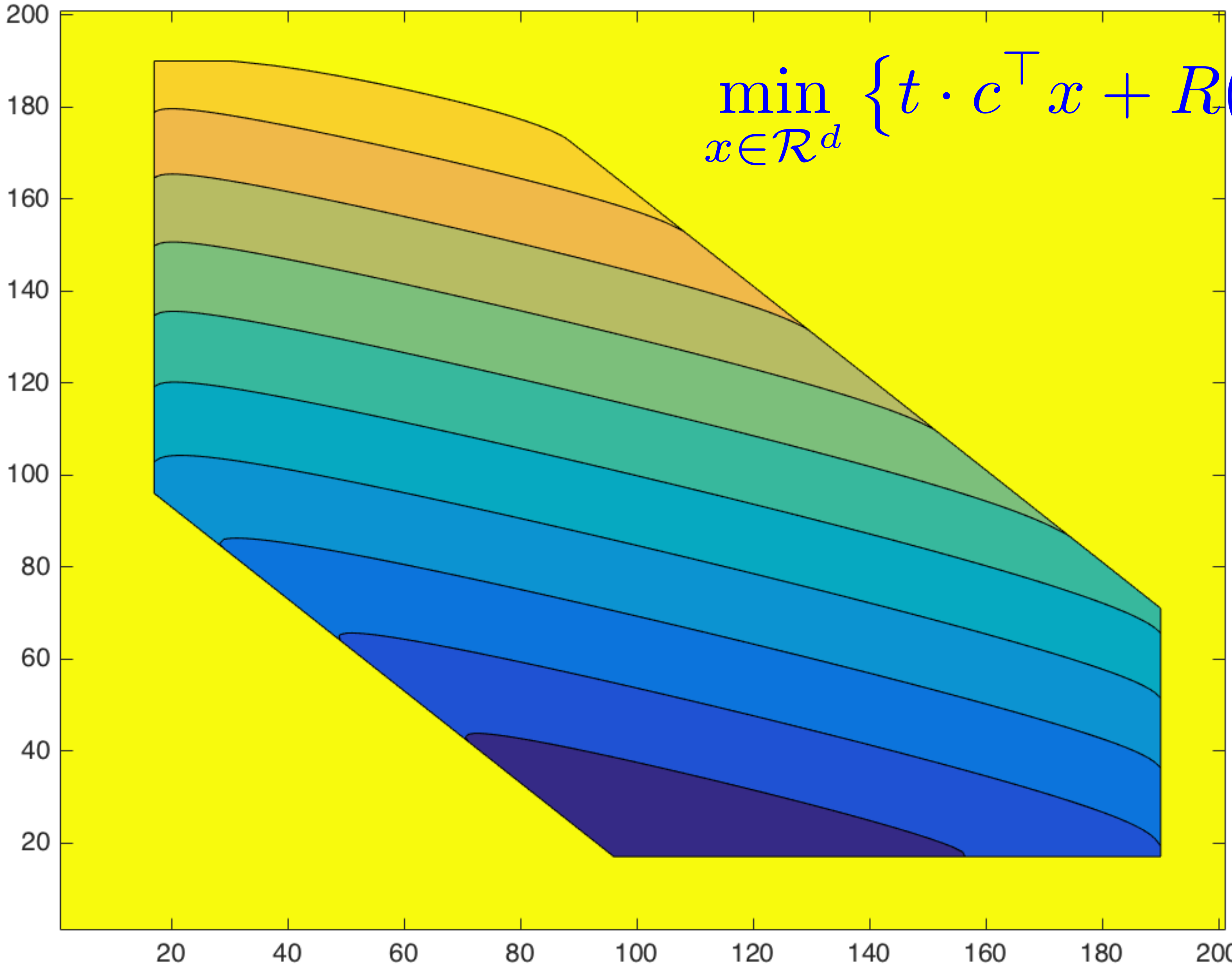
$$\min_{x \in \mathcal{R}^d} \left\{ t \cdot c^\top x + R(x) \right\}$$



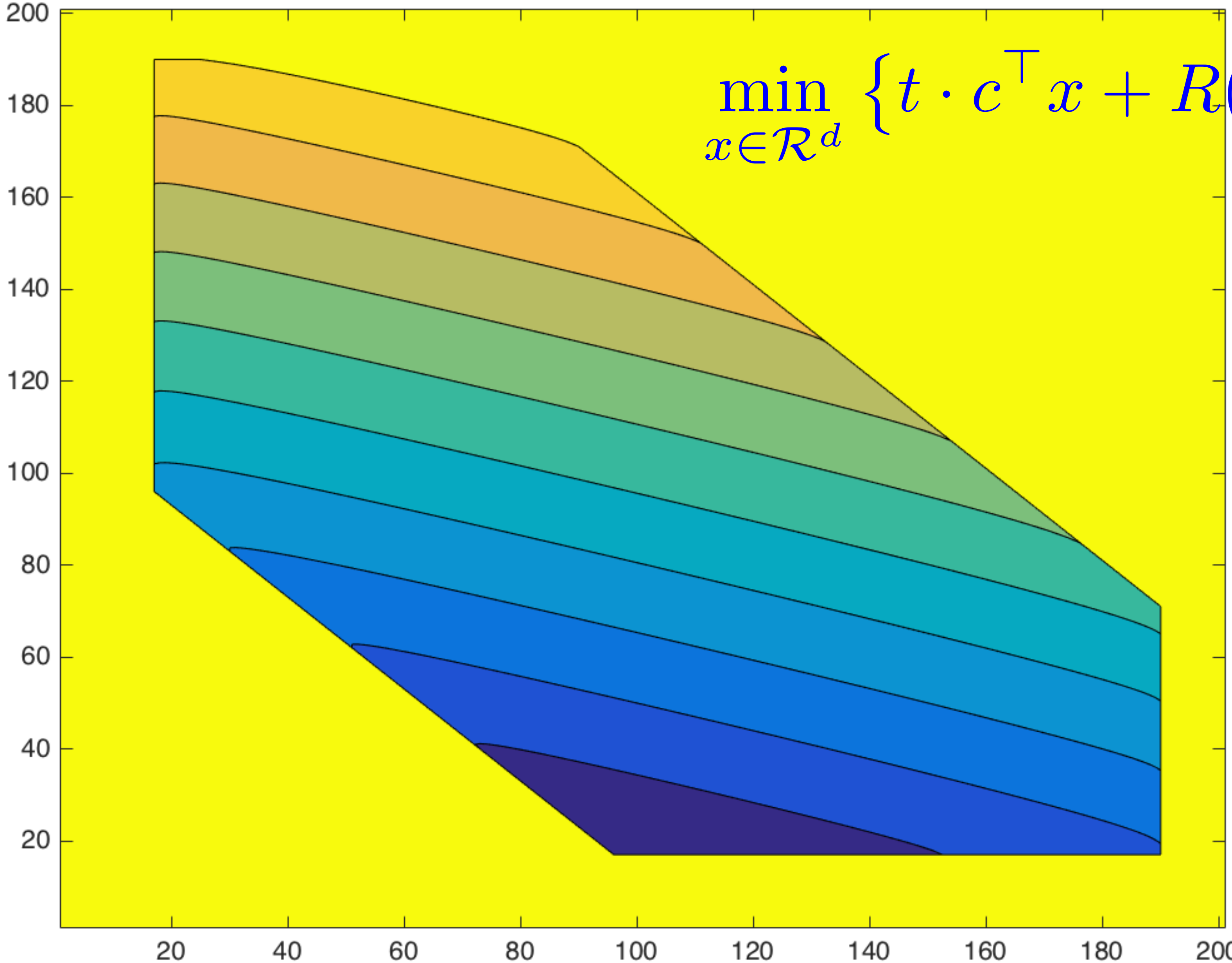
$$\min_{x \in \mathcal{R}^d} \left\{ t \cdot c^\top x + R(x) \right\}$$



$$\min_{x \in \mathcal{R}^d} \left\{ t \cdot c^\top x + R(x) \right\}$$



$$\min_{x \in \mathcal{R}^d} \left\{ t \cdot c^\top x + R(x) \right\}$$



$$\min_{x \in \mathcal{R}^d} \left\{ t \cdot c^\top x + R(x) \right\}$$

Path following method

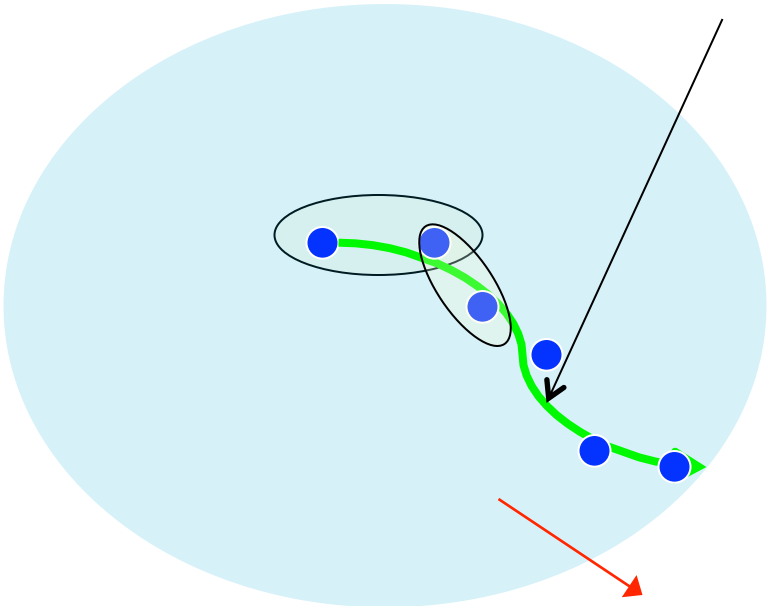
Changing the parameter t from 0 to ∞

$$\min_{x \in \mathcal{R}^d} \left\{ t \cdot c^\top x + R(x) \right\}$$

Iteratively:

1. Update t
2. Optimize new objective (inside the yellow ellipse)

$$\beta(t) = \arg\min_{x \in \mathcal{R}^n} \left\{ t \cdot c^\top x + R(x) \right\}$$



Inside the yellow ellipse: self concordant functions

R - self concordant for convex set K, at each x, hessian of R at x defines local norm:

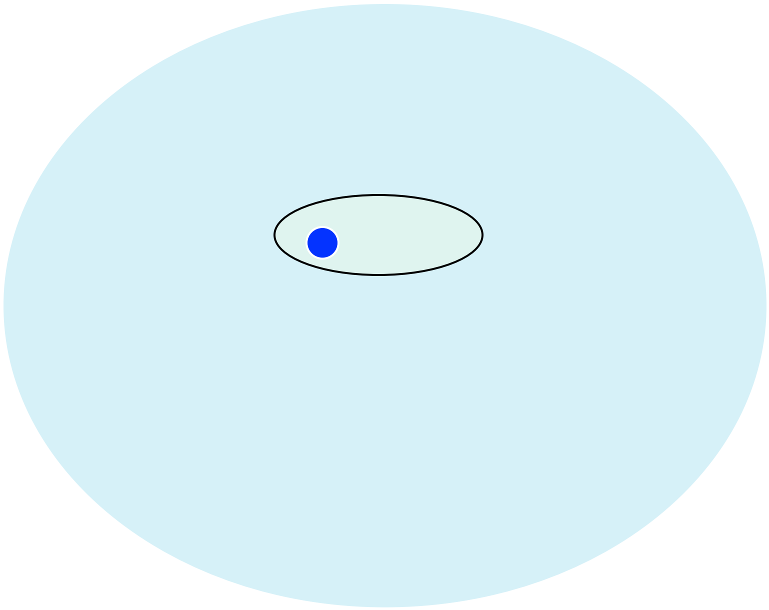
$$\|y\|_x = y^\top \nabla^2 R(x) y \geq 0$$

The Dikin ellipsoid

$$D_1(x) = \{y \text{ such that } \|y - x\|_x \leq 1\}$$

Inside Dikin ellipsoid: function is
strongly convex and smooth
with respect to the local norm

One newton step suffices!



Path following method – complexity

$$\min_{x \in \mathcal{R}^d} \left\{ t \cdot c^\top x + R(x) \right\}$$

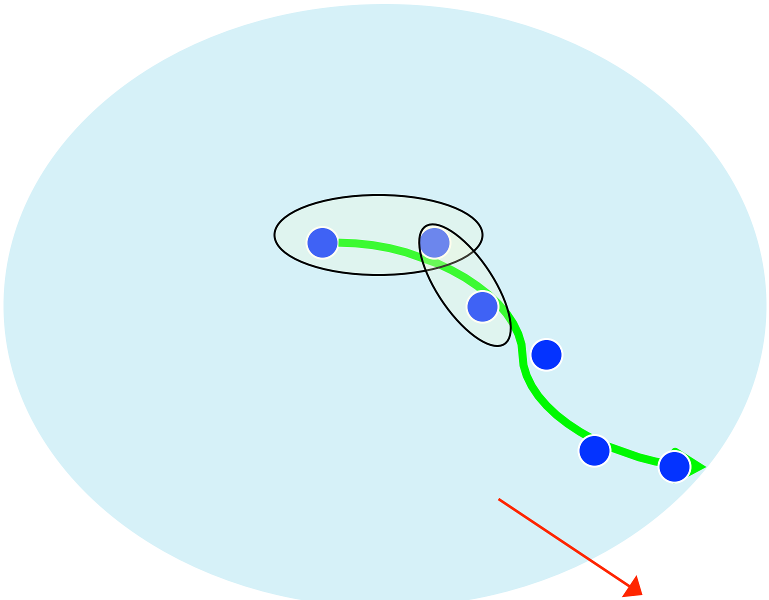
Self-concordance parameter ~ isoperimetric constant of K

1. Geometric update of t → # of iterations <= $\nu^{1/2}$
2. Each iteration: mirror descent (Newton), matrix inversion

REQUIRE EFFICIENT BARRIER!!

Long standing question:

efficient universal barrier?



Interior point: summary

$$\min_{x \in \mathcal{R}^d} \left\{ t \cdot c^\top x + R(x) \right\}$$

Problems with gradient descent: projections, cannot exploit curvature

Moved to Newton's method + barrier + changed scaling → interior algorithm,
provably converging in poly time

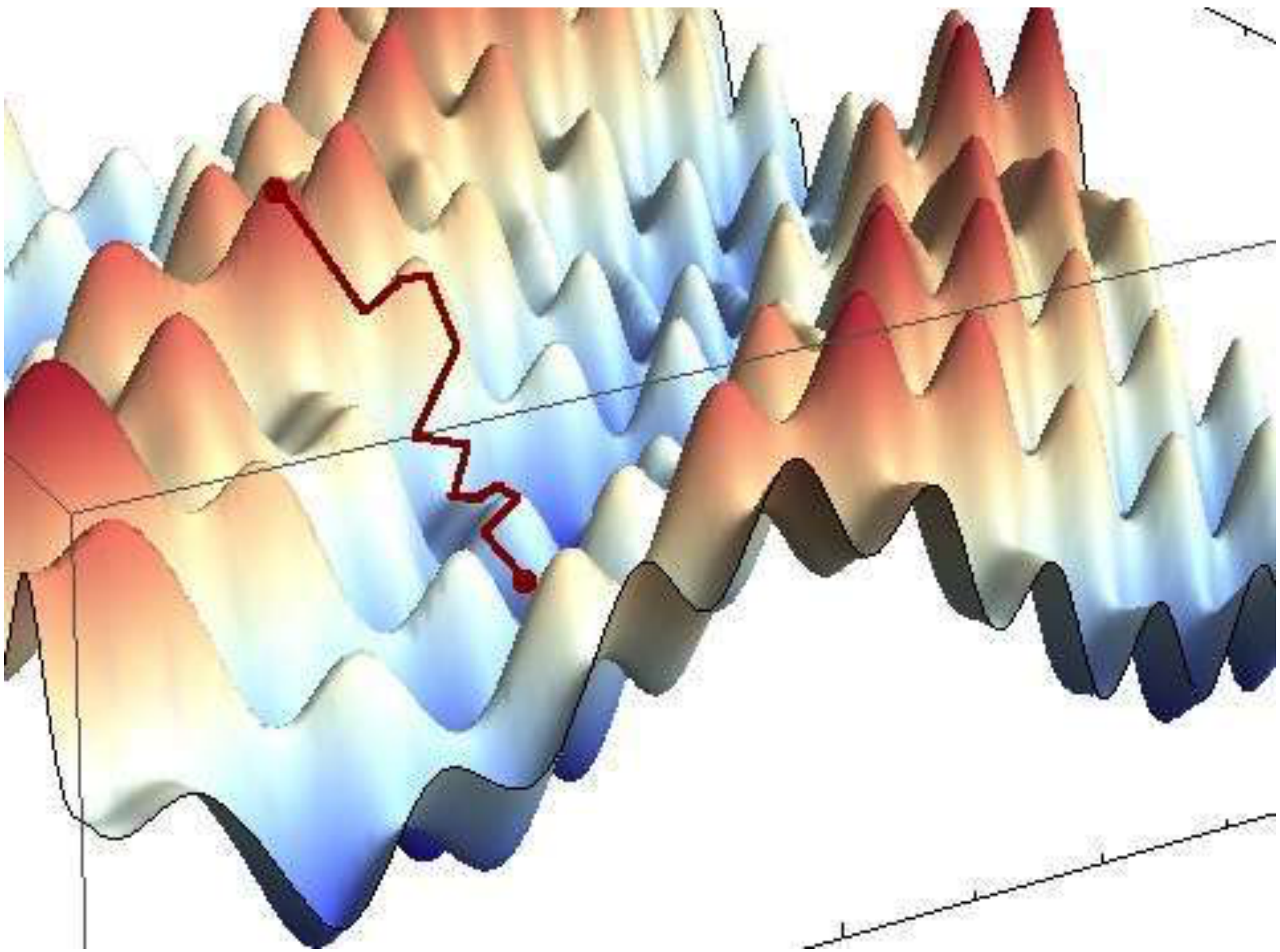
BUT: REQUIRE EFFICIENT BARRIER!!

Long standing open question: efficient universal barrier?

Agenda

1. Mini tutorial on IPM
2. Mini tutorial on SA
3. The equivalence of SA and IPM
4. How to get faster convex opt

Simulated annealing: mini-tutorial



Simulated annealing

Common heuristic for non-convex optimization:

Boltzman distribution over a set K: (w.r.t. function f or direction c)

$$P_{t,f}(x) \equiv \frac{e^{-\frac{f(x)}{t}}}{\int_{y \in \mathcal{K}} e^{-\frac{f(y)}{t}} dy}$$

t = ∞: uniform over K

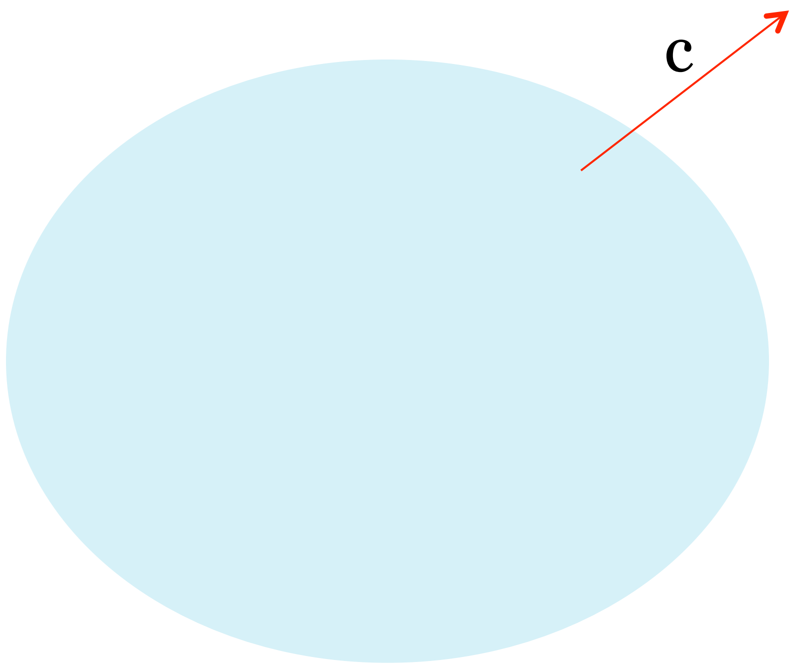
t → 0: approach min f(x) over K

Simulated annealing

Common heuristic for non-convex optimization:

Boltzman distribution over a set K: (w.r.t. function f or direction c)

$$P_{t,c}(x) \equiv \frac{e^{-\frac{c^\top x}{t}}}{\int_{y \in K} e^{-\frac{c^\top y}{t}} dy}$$



t = ∞: uniform over K

t → 0: approach min $c^T x$ over K

Simulated annealing - intuition

$$P_{t,c}(x) \equiv \frac{e^{-\frac{c^\top x}{t}}}{\int_{y \in K} e^{-\frac{c^\top y}{t}} dy}$$

Initially: sampling uniformly at random

When temperature is very low → sample from minimum = goal

If successive distributions are "close" – can use "warm start" to sample efficiently from P_{t+1} given an efficient method for sampling from P_t

1. What is a warm start?
2. How to sample from P_t? (there are many methods…)

Hit-and-Run

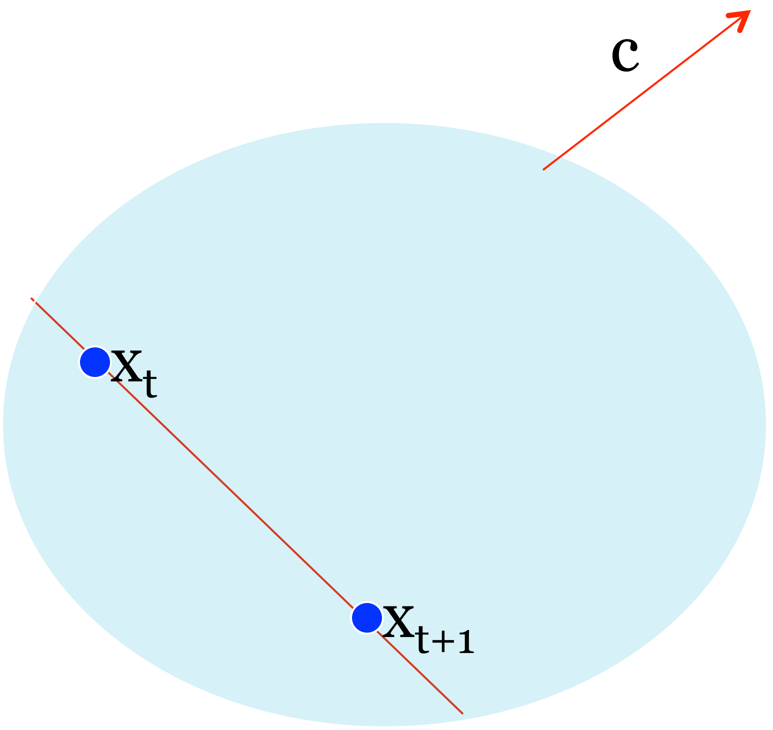
Iteratively:

$$P_{t,c}(x) \equiv \frac{e^{-\frac{c^\top x}{t}}}{\int_{y \in K} e^{-\frac{c^\top y}{t}} dy}$$

1. Sample line from distribution

$$u \sim N(X_t, C_t)$$

2. Consider interval = restriction to K
3. Sample from induced distribution P_t on interval – this is X_{t+1}

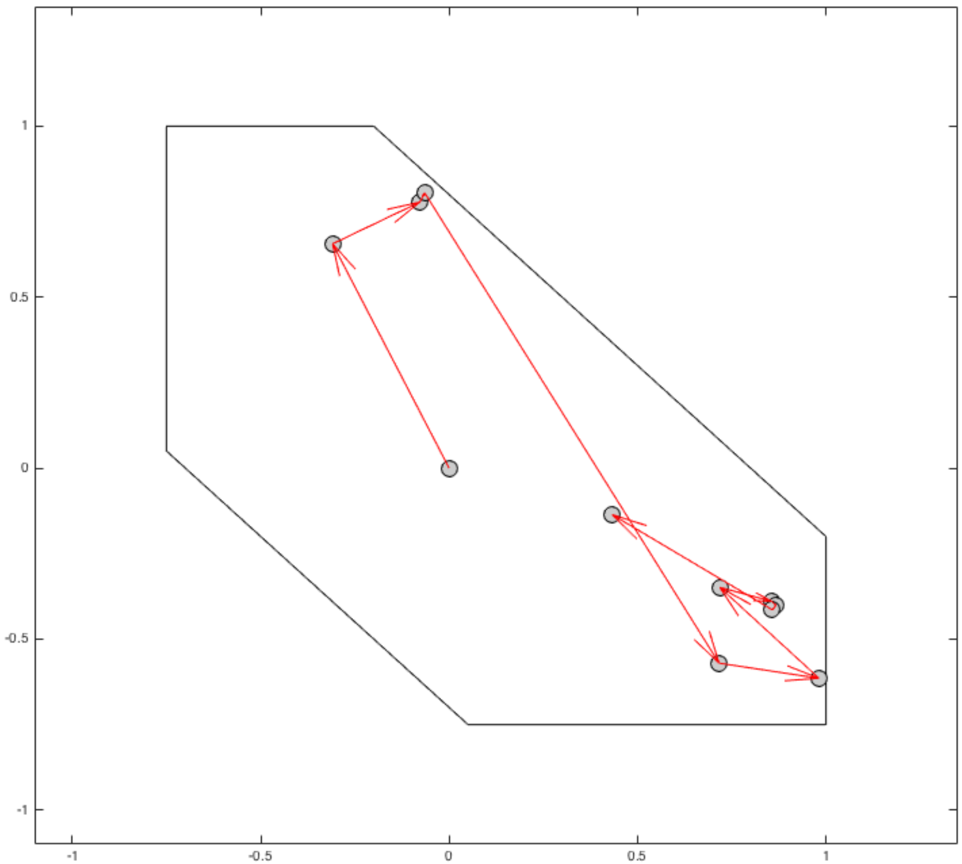


Theorem: HNR has stationary dist. P_t

How does K enter the random walk?

Notice– only membership oracle needed for K!

hit & run



Simulated annealing w. Hit-and-Run

First polynomial-time algorithm [Kalai, Vempala '06]:

1. Sample from $P_{t,c}(x) \equiv \frac{e^{-\frac{c^\top x}{t}}}{\int_{y \in K} e^{-\frac{c^\top y}{t}} dy}$

   using Hit-and-Run

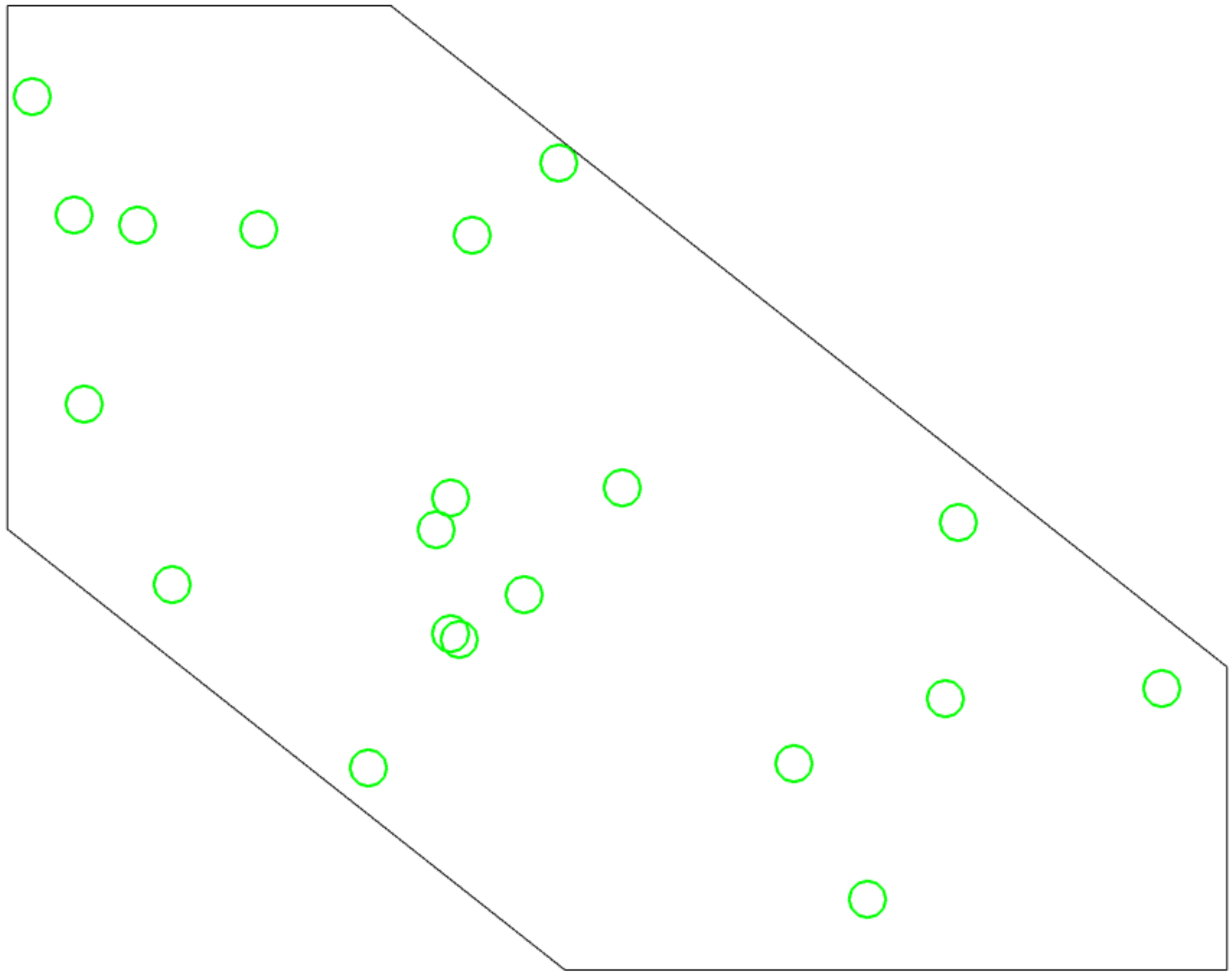
2. Successive distributions are close enough if

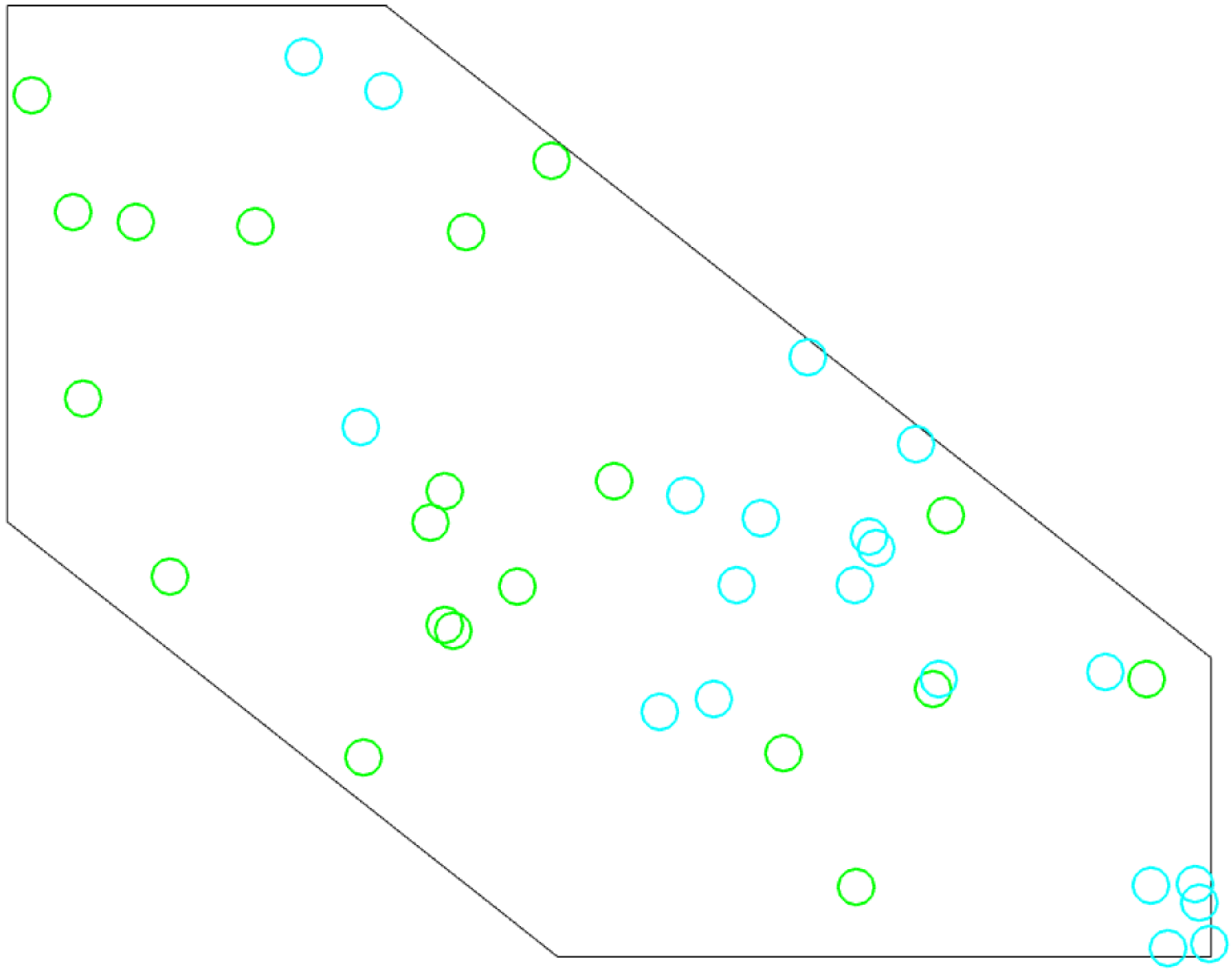
$$KL(P_{t_k}, P_{t_{k+1}}) \leq \frac{1}{2} \quad \Leftrightarrow \quad \|\mathrm{cov}(P_{t_k}) - \mathrm{cov}(P_{t_{k+1}})\| \leq \frac{1}{2}$$

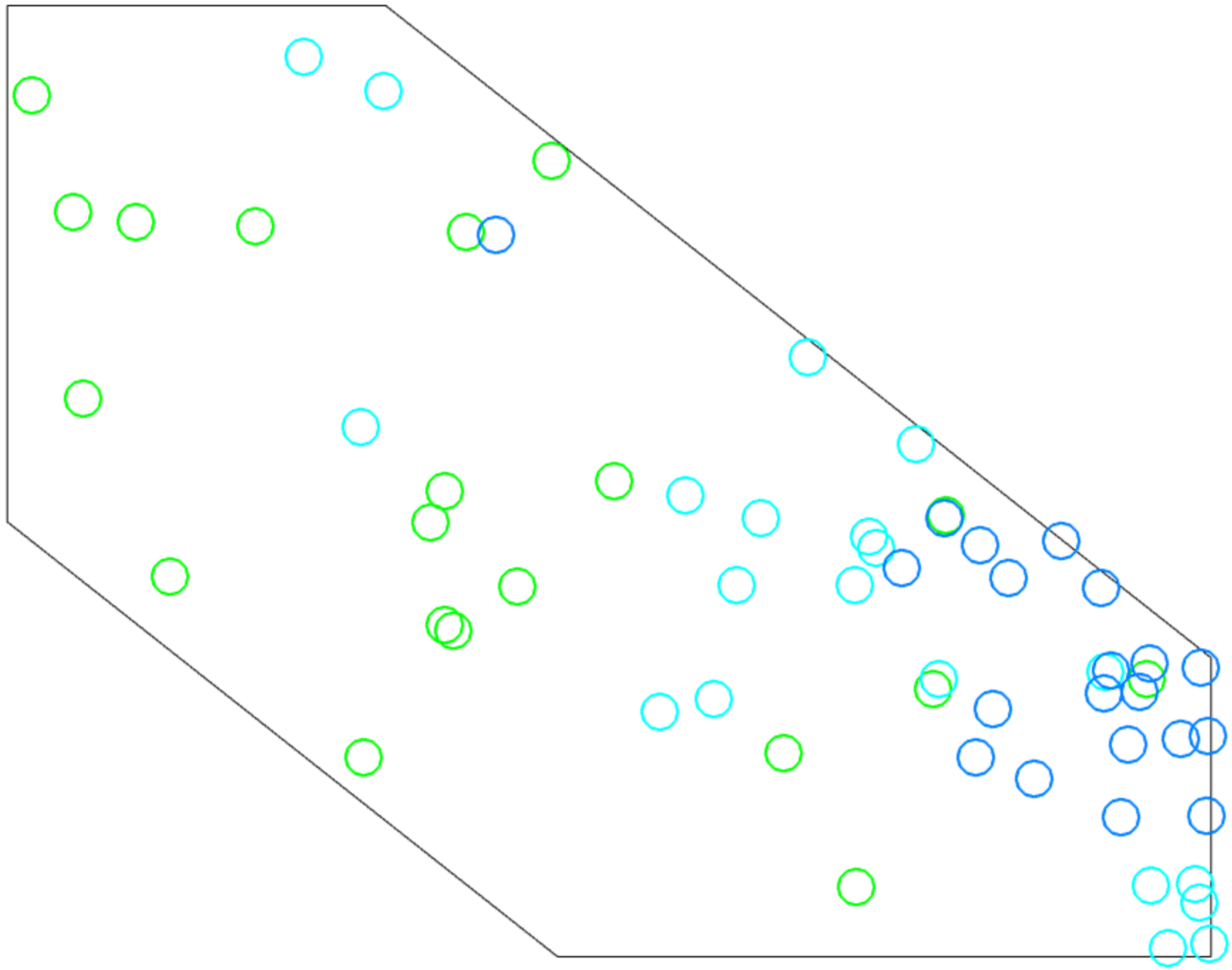
3. SA with HNR, temperature schedule of $t_{k+1} = t_k(1 - \frac{1}{\sqrt{n}})$

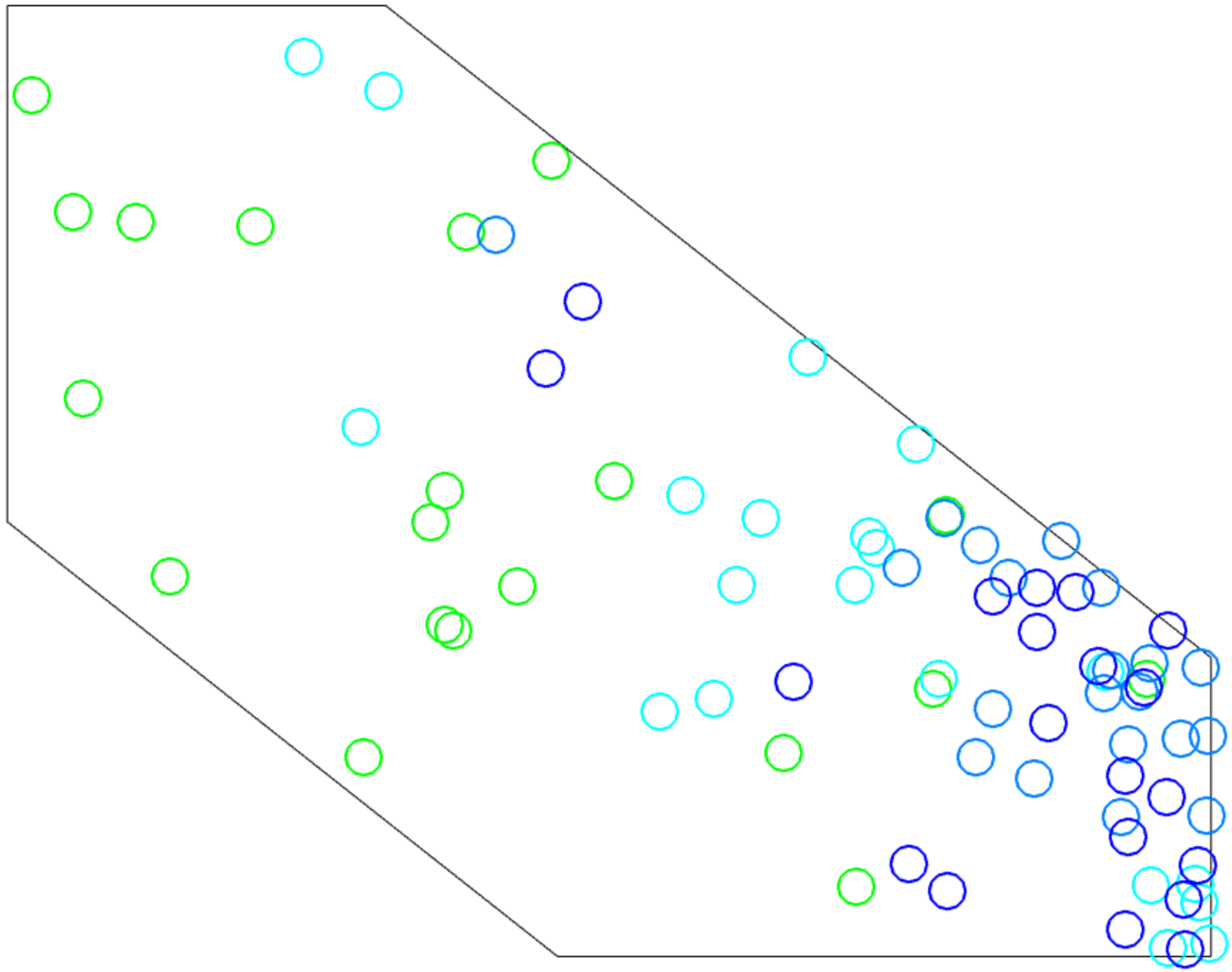
Their main theorem: algorithm returns approximate solution in $O(\sqrt{n} \log \frac{1}{\epsilon})$ iterations, and overall time

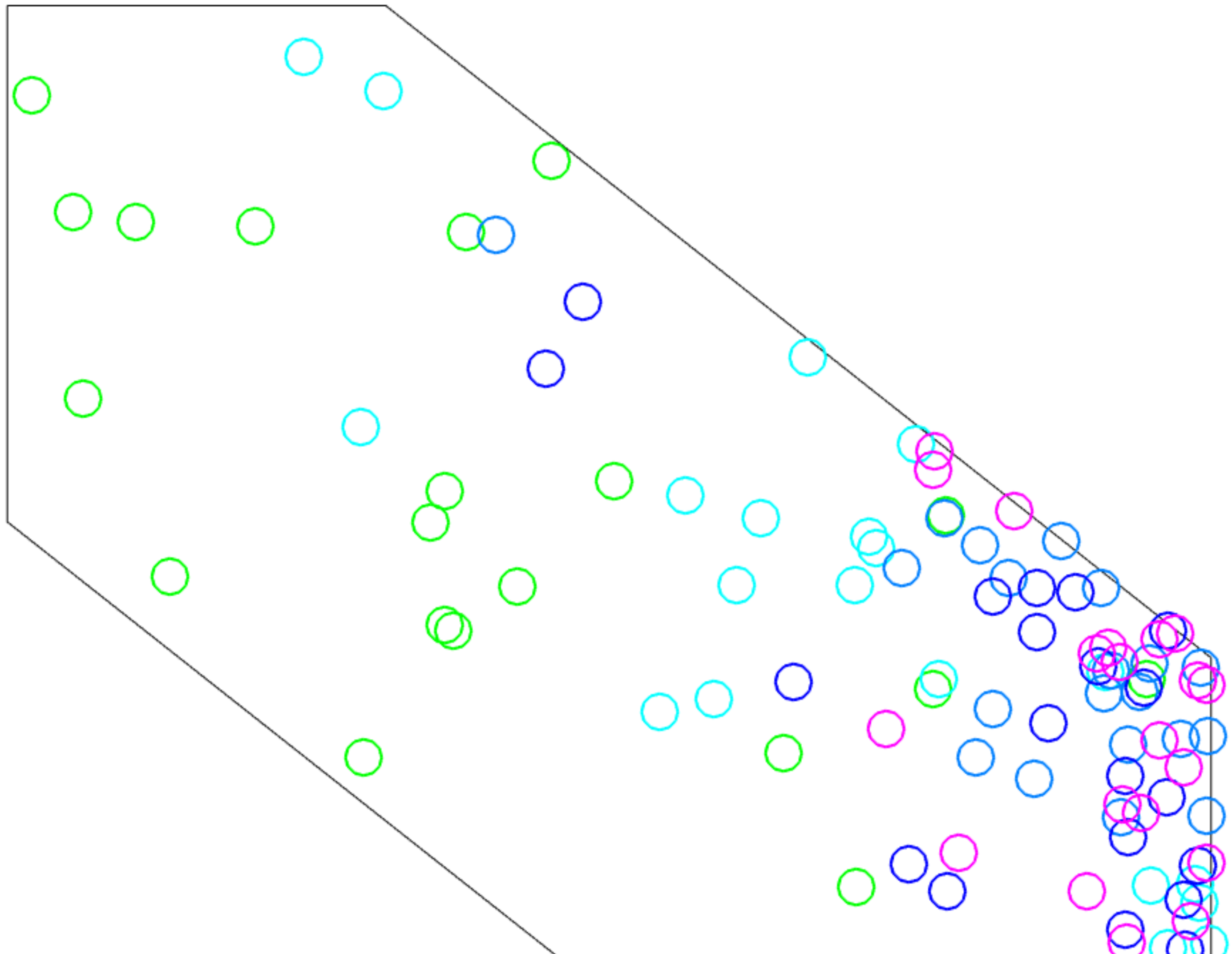
$$O(\sqrt{n} \log \frac{1}{\epsilon} \times n \times n^3) = \tilde{O}(n^{4.5})$$

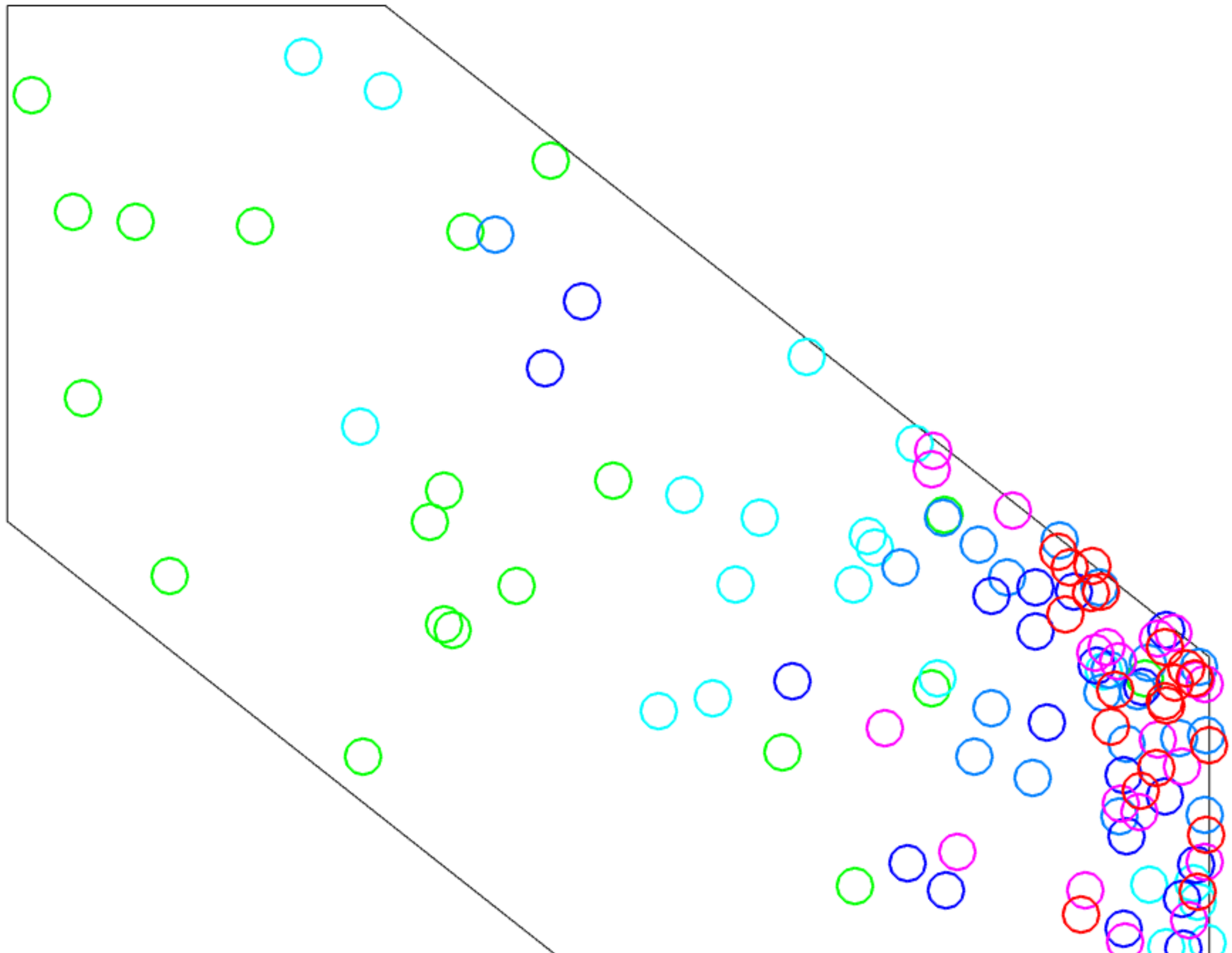


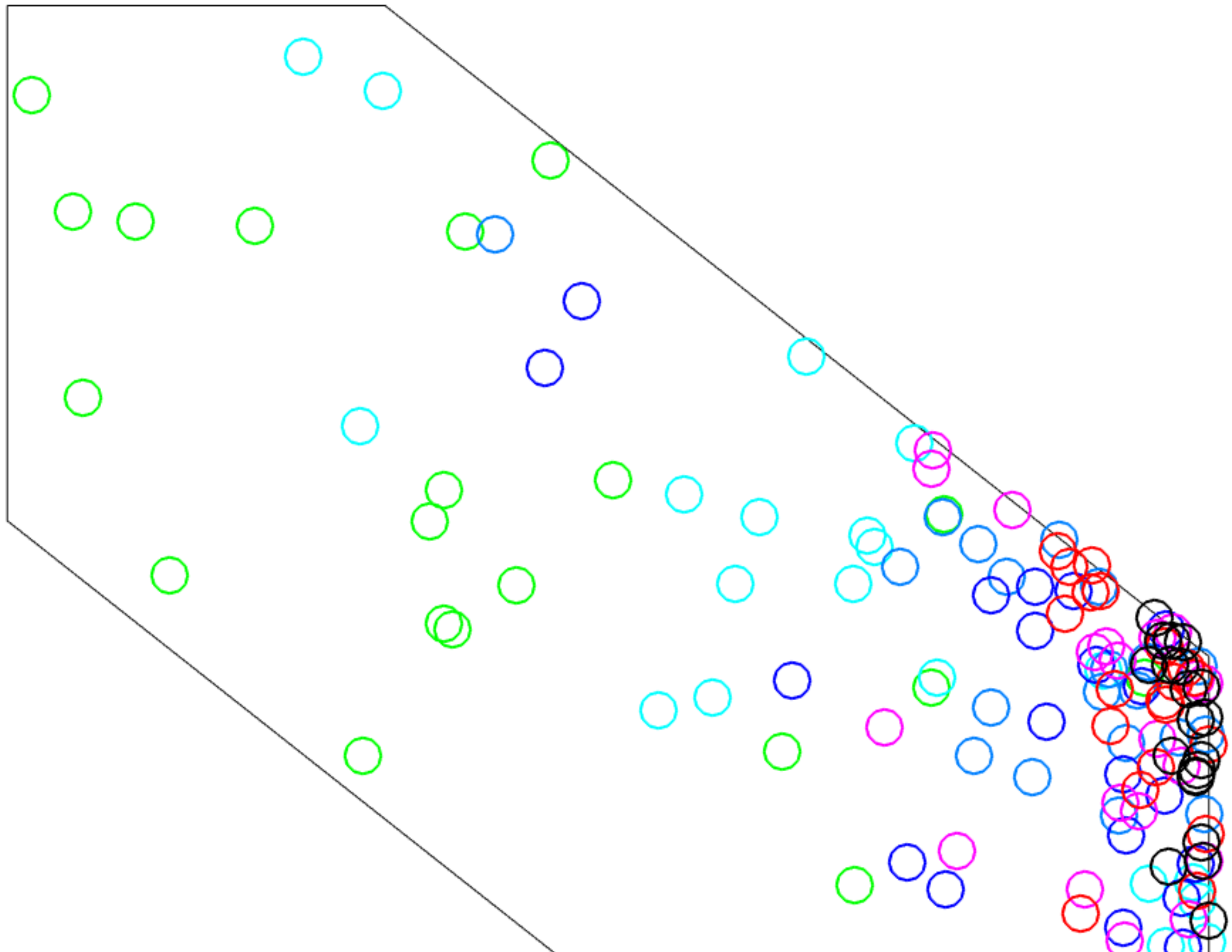








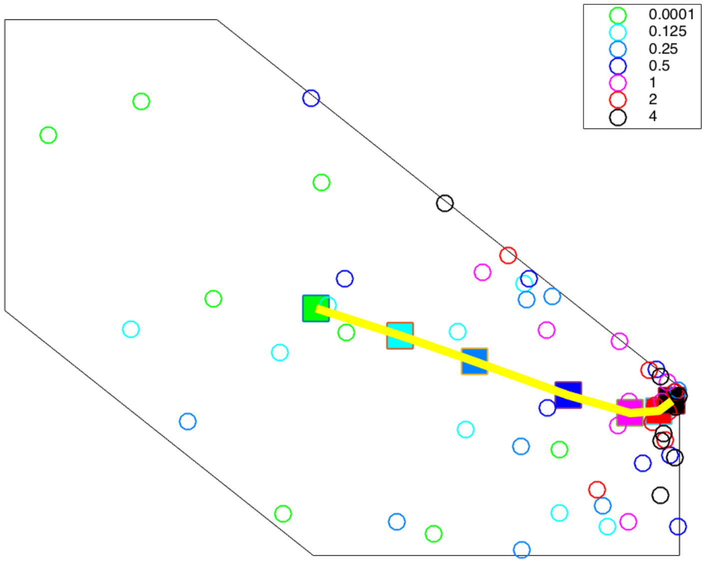




New: heat path

Curve of mean of Boltzman distribution, parameterized by temperature

$$\mu(t) = E_{x \sim P_{t,c}(x)}[x] \ , \ P_{t,c}(x) = \frac{e^{-c^\top x/t}}{\int_{y \in K} e^{-c^\top y/t} dy}$$



Two different convex optimization methods

Not really different

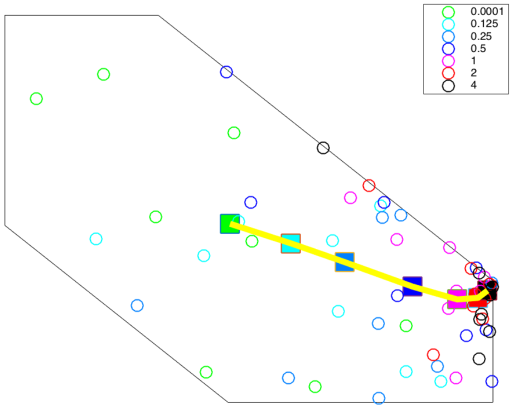
Simulated Annealing via Hit-and-Run

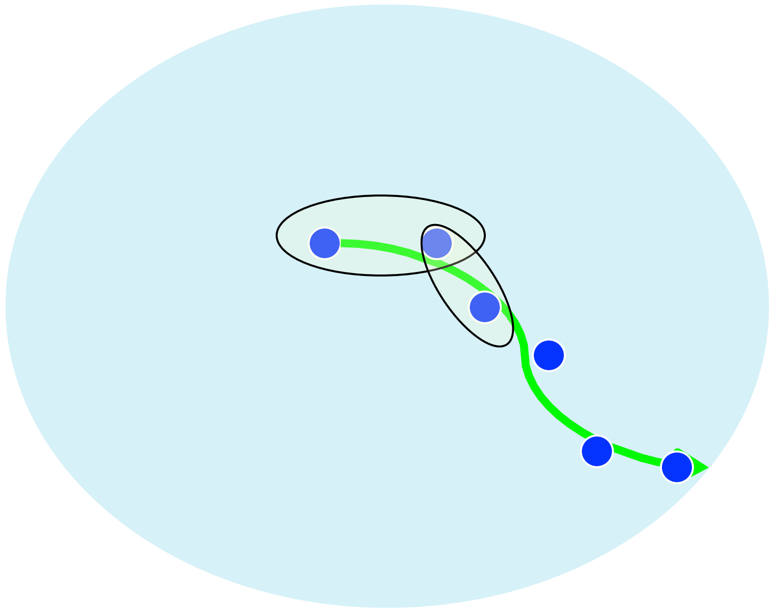
Interior Point Methods via Path Following

Our key result: there exists a barrier R(x) for any convex set such that CentralPath is **identically** the HeatPath

$$\mu(t) = \underset{K \ni x \sim e^{-\frac{c^\top x}{t}}}{E}[x]$$

$$\beta(t) = \arg\min_{x \in \mathcal{R}^n} \{t \cdot c^\top x + R(x)\}$$





What is this special function?

the entropic barrier:

$$A(c) = \log \int_{x \in K} e^{-c^\top x} dx$$

= log partition function for the exponential family

$$\nabla A(c) = -E_{x \sim P_c}[x] \ , \ \nabla^2 A(c) = E_{x \sim P_c}[(x - E[x])(x - E[x])^\top]$$

entropic barrier for K:

$$A^*(x) = \sup_c \{c^\top x - A(c)\}$$

1. Guller '96 + Nesterov/Nemirovski '94

   $\nu = O(n)$
   PSD cone - $\nu = O(n^{1/2})$

2. Bubeck-Eldan '15:
   $\nu = n + o(n)$

Convergence/running time analysis

Method	Interior point methods	Simulated annealing
Inside each temperature	Fast convergence of Newton's method	Fast convergence of Hit-and-Run to stationary distribution
Change temperature	After Newton converged	stationary distribution, estimate covariance
Condition	Newton decrement << 1	Distance between consecutive dist.

Why is this interesting?

- Unifies two distinct literatures
- One less algorithm to teach/learn in your class!
- Using IPM ideas we get a faster algorithm for convex optimization

$$\tilde{O}(\sqrt{n}) \Rightarrow \tilde{O}(\sqrt{\nu})$$

- For semi-definite programming:

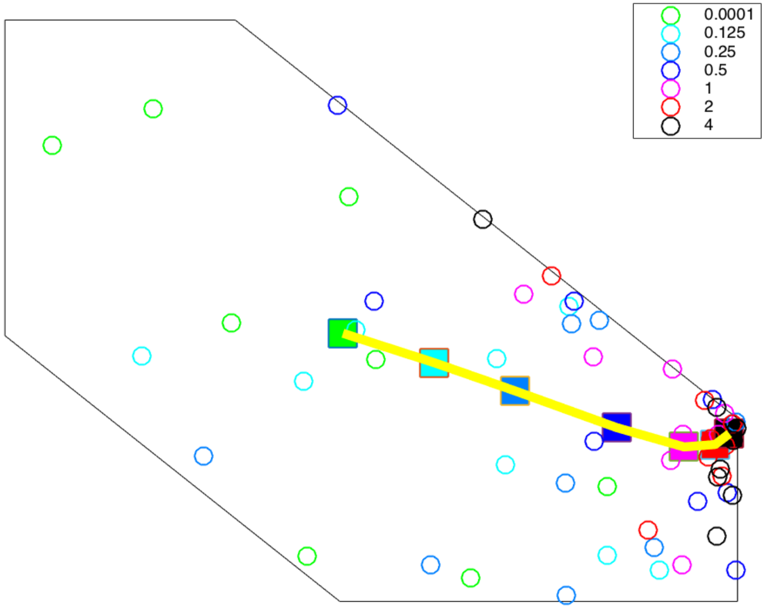
$$\nu = O(\sqrt{n})$$

- Randomized efficient interior-point path-following algorithm for any convex set! (long-standing open problem in optimization)

- Time for a Demo?

- Time for a proof sketch?

- Fin…



When can we increase the temperature?

Theorem [Kalai-Vempala '06]:
Temperature schedule suffices to satisfy: ($c_k = t_k * c$)

$$\|P_{c_k} - P_{c_{k+1}}\|_{TV2} = \max\left\{ \left\| \frac{P_{c_k}}{P_{c_{k+1}}} \right\|_2, \left\| \frac{P_{c_{k+1}}}{P_{c_k}} \right\|_2 \right\} \leq O(1)$$

For hit-N-run-based simulated annealing to work.

Our main lemma: for the above, we can have :

$$\frac{t_{k+1}}{t_k} = 1 + \frac{O(1)}{\sqrt{\nu}}$$

Proof: $$\frac{t_{k+1}}{t_k} = 1 + \frac{O(1)}{\sqrt{\nu}}$$

Part 1:
duality of Bregman divergence, equivalence to Kullback-Leibler for exponential families:

$$KL(P_{c_k}, P_{c_{k+1}}) = D_A(c_k, c_{k+1}) = D_{A^*}(x(c_k), x(c_{k+1}))$$

(reminder, Bregman divergence w.r.t. A ~ local norm)

$$D_A(x, y) \equiv A(x) - A(y) - \nabla A(y)^\top (x - y) \approx \|x - y\|^2_{A(x)}$$

$$A(\theta) = \log \int_{x \in K} e^{-\theta^\top x} dx \qquad x(c) = E_{x \sim P_c}[x] = -\nabla A(c)$$

Proof: $$\frac{t_{k+1}}{t_k} = 1 + \frac{O(1)}{\sqrt{\nu}}$$

Part 2:
by definition and calculation:

$$\log\left\|\frac{P_{c_{k+1}}}{P_{c_k}}\right\| = D_A(c_{k+1}, c_k) + D_A(c_k, c_{k+1})$$

Proof: $\frac{t_{k+1}}{t_k} = 1 + \frac{O(1)}{\sqrt{\nu}}$

Part 3 – using IPM:

Bregman divergence between local means bounded inside the Dikin ellipsoid by O(1).

$$D_A(c_{k+1}, c_k) \sim \|c_k - c_{k+1}\|^2_{A(c_k)}$$

$$\sim \|x(c_k) - x(c_{k+1})\|^{*\,2}_{A(c_k)}$$

$$= \|x_k - x_{k+1}\|^2_{A^*(x_k)}$$

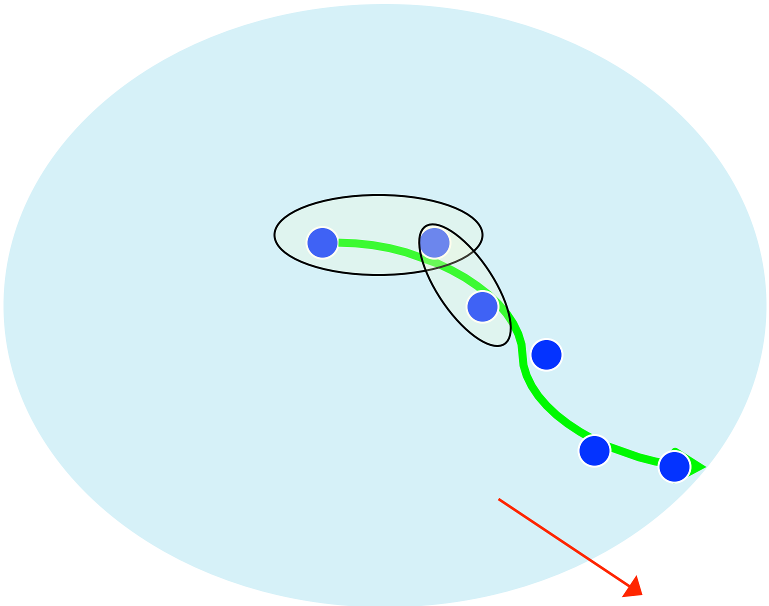
$$= O(1)$$

Putting it together

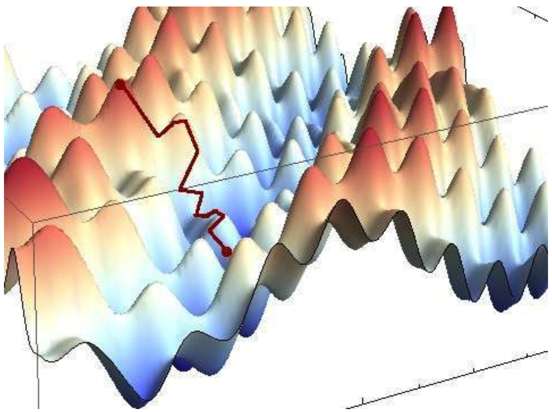
1. Nemirovski: # of Dikin ellipsoids on the path <= $\nu^{1/2}$
2. This bounds the total # of temperature updates

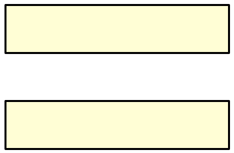
Complexity:

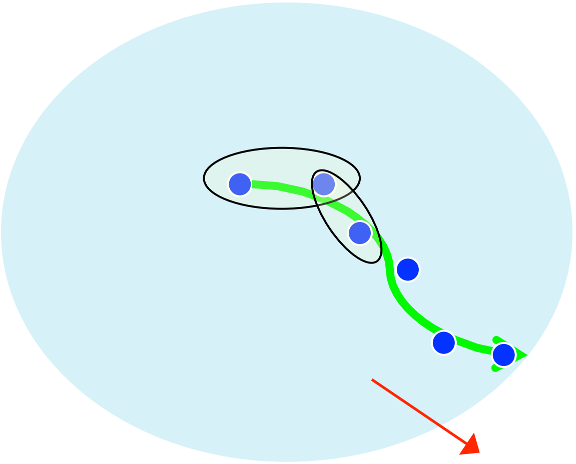
1. Each iteration requires Hit-And-Run * N times
   (for mean & covariance)



Conclusion

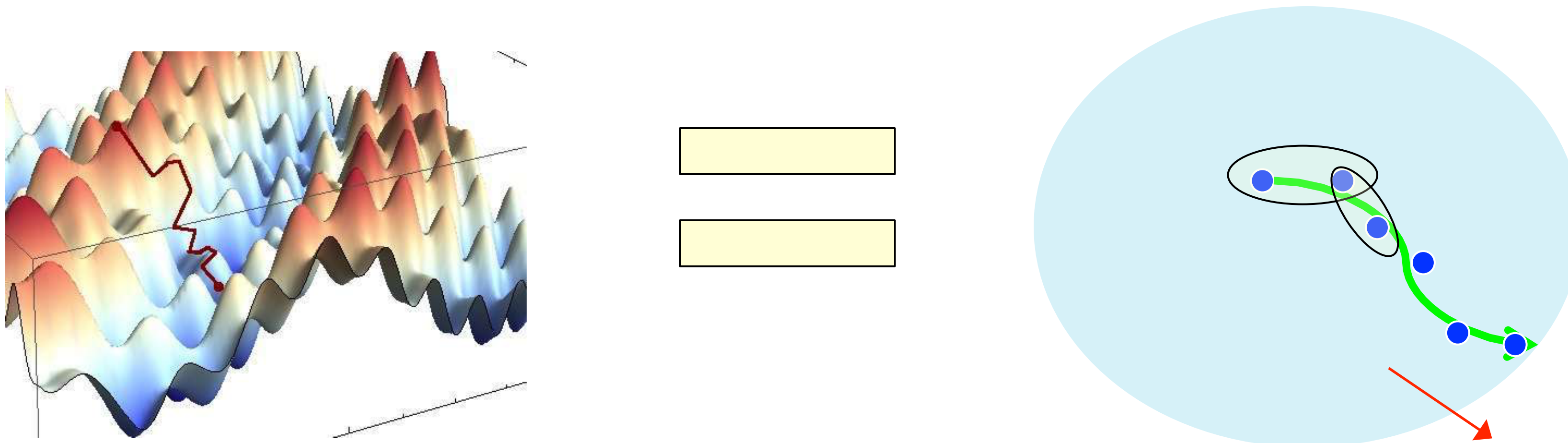






1. Faster convex optimization ➔ $\nu^{1/2}$ iterations vs. $n^{1/2}$, faster SDP
   each iteration $n^3\nu^2$ vs n^4
2. Efficient randomized IPM for any convex body (open Q in optimization)
3. Defined the Heat path, showed equivalence to Central Path

Where do we go from here?



1. Heat path for non-convex optimization
2. Regret minimization – geometric connection
3. Gradient descent analogue?