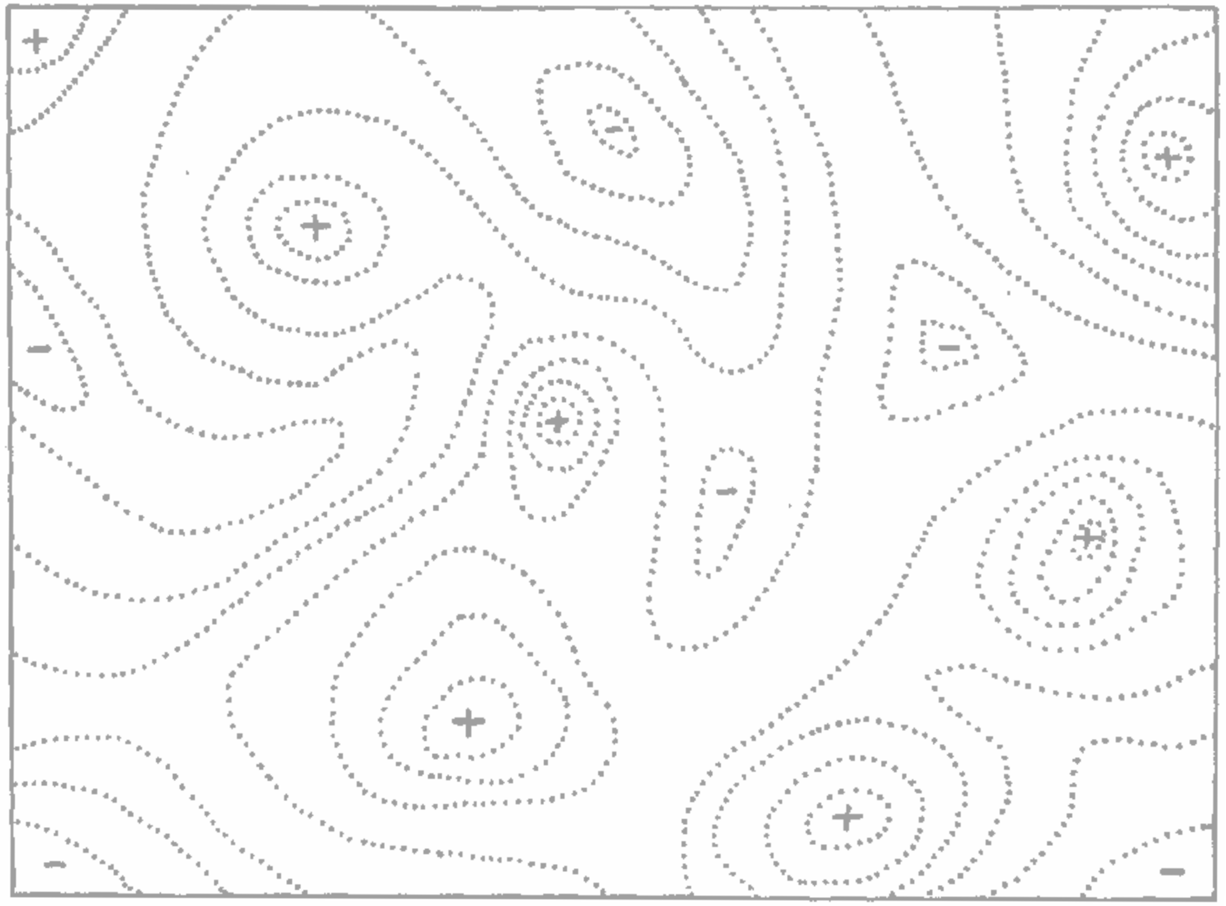
OPTIMIZATION LANDSCAPES



CHRISTIAN L. MÜLLER
CENTER FOR COMPUTATIONAL MATHEMATICS, FLATIRON INSTITUTE, NEW YORK
INSTITUTE FOR STATISTICS, LUDWIG-MAXIMILIANS-UNIVERSITÄT &
INSTITUTE OF COMPUTATIONAL BIOLOGY, HELMHOLTZ ZENTRUM, MUNICH

FWAM Flatiron Conference 10/30/2019

CONGRATULATIONS

SIMONS FOUNDATION 25
1994–2019

CONGRATULATIONS

FLATIRON INSTITUTE
Center for Computational Mathematics

5 YEARS OF SIMONS CENTER FOR DATA ANALYSIS (SCDA)/ FLATIRON INSTITUTE

CONGRATULATIONS

FLATIRON INSTITUTE
Center for Computational Mathematics

SIMONS FOUNDATION
1994–2019
25

5 YEARS OF SIMONS CENTER FOR DATA ANALYSIS (SCDA)/ FLATIRON INSTITUTE

OPTIMIZATION

FLATIRON INSTITUTE
Center for Computational Mathematics

OPTIMIZATION

op·ti·mi·za·tion

/ˌäptəməˈzāSHən,ˌäptəˌmīˈzāSHən/

noun

noun: **optimization**; plural noun: **optimizations**; noun: **optimisation**; plural noun: **optimisations**

1. the action of making the best or most effective use of a situation or resource.

google dictionary

OPTIMIZATION

op·ti·mi·za·tion

/ˌäptəməˈzāSHən,ˌäptəˌmīˈzāSHən/

noun

noun: **optimization**; plural noun: **optimizations**; noun: **optimisation**; plural noun: **optimisations**

1. the action of making the best or most effective use of a situation or resource.

google dictionary

Mathematical optimization

Discipline

Description

Mathematical optimization or mathematical programming is the selection of a best element from some set of available alternatives. Wikipedia

wikipedia

OPTIMIZATION

Mathematical optimization (alternatively spelled *optimisation*) or **mathematical programming** is the selection of a best element (with regard to some criterion) from some set of available alternatives.[1]

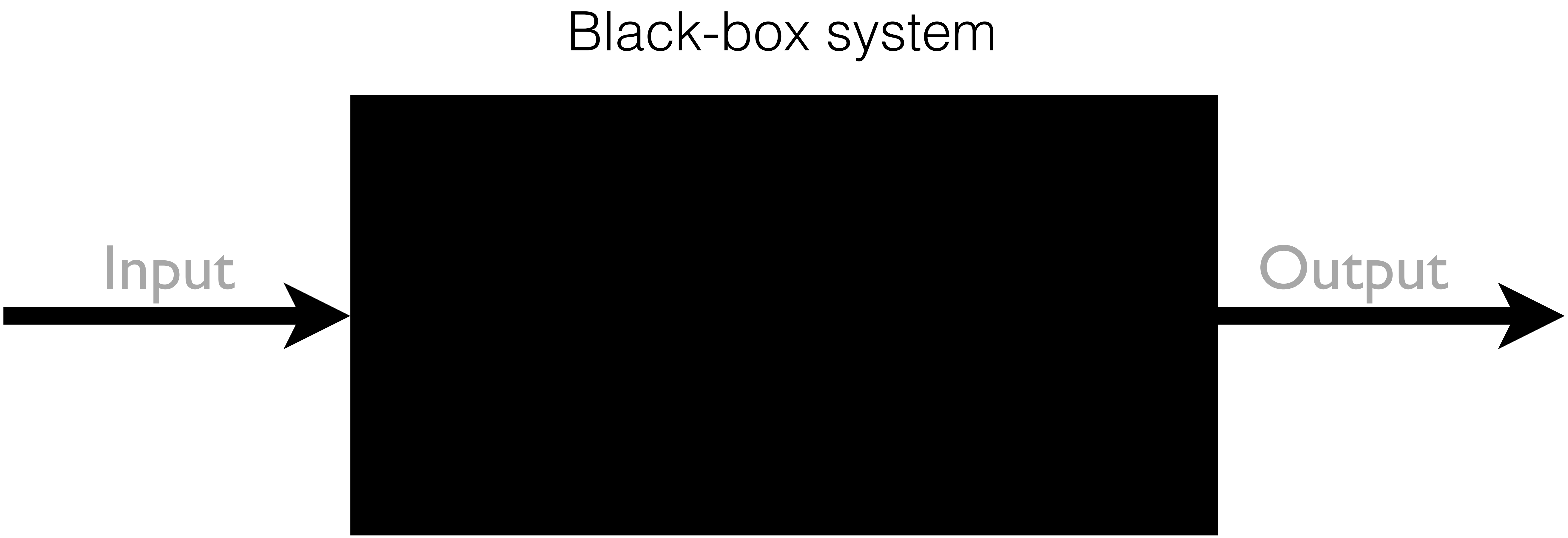
Optimization problems of sorts arise in all quantitative disciplines from computer science and engineering to operations research and economics, and the development of solution methods has been of interest in mathematics for centuries.[2]

wikipedia

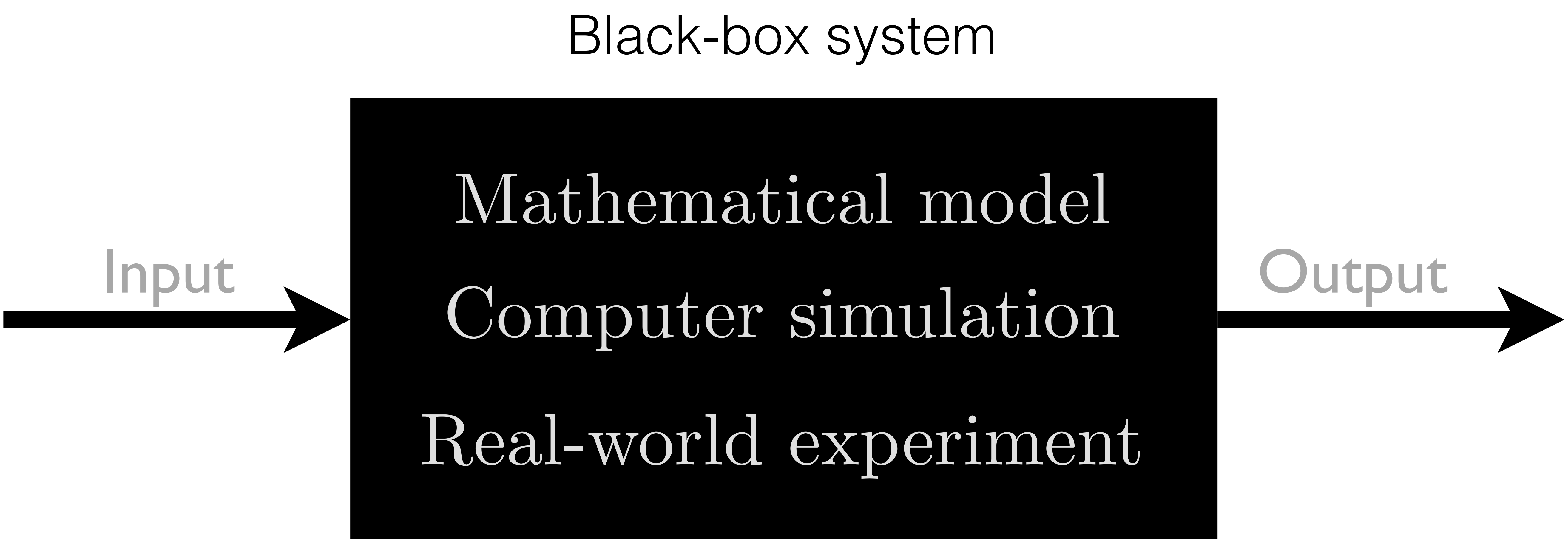
1. "The Nature of Mathematical Programming Archived 2014-03-05 at the Wayback Machine," *Mathematical Programming Glossary*, INFORMS Computing Society.
2. ^ Du, D. Z.; Pardalos, P. M.; Wu, W. (2008). "History of Optimization". In Floudas, C.; Pardalos, P. (eds.). *Encyclopedia of Optimization*. Boston: Springer. pp. 1538–1542.

OPTIMIZING A BLACK-BOX

FLATIRON INSTITUTE
Center for Computational Mathematics



OPTIMIZING A BLACK-BOX



BLACK-BOX OPTIMIZATION

Input

$\mathbf{x} \in \mathbb{R}^n$

Black-box system

Output

$f(\mathbf{x}) \in \mathbb{R}$

BLACK-BOX OPTIMIZATION

Input

$\mathbf{x} \in \mathbb{R}^n$

Black-box system

Output

$f(\mathbf{x}) \in \mathbb{R}$

x_1

x_2

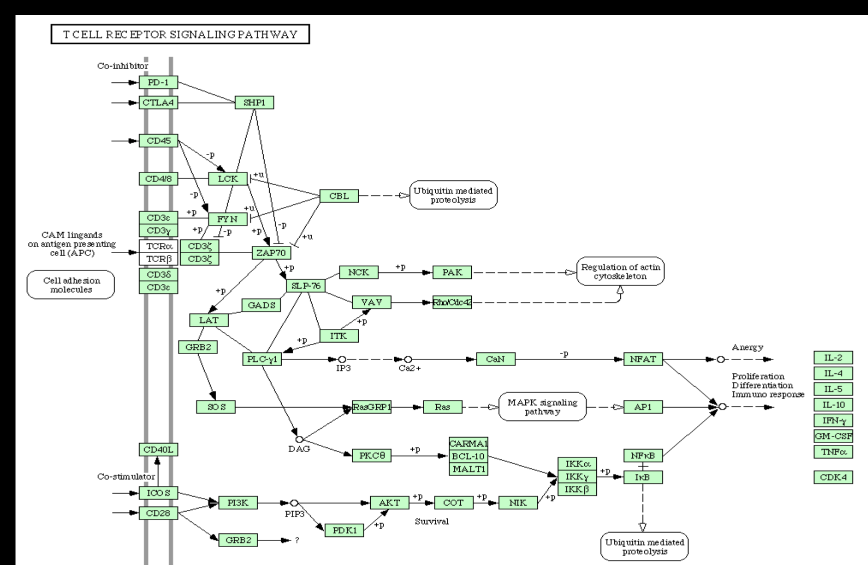
...

x_i

...

x_n

Systems biology models



BLACK-BOX OPTIMIZATION

FLATIRON INSTITUTE
Center for Computational Mathematics

Input

$\mathbf{x} \in \mathbb{R}^n$

Black-box system

Output

$f(\mathbf{x}) \in \mathbb{R}$

x_1

x_2

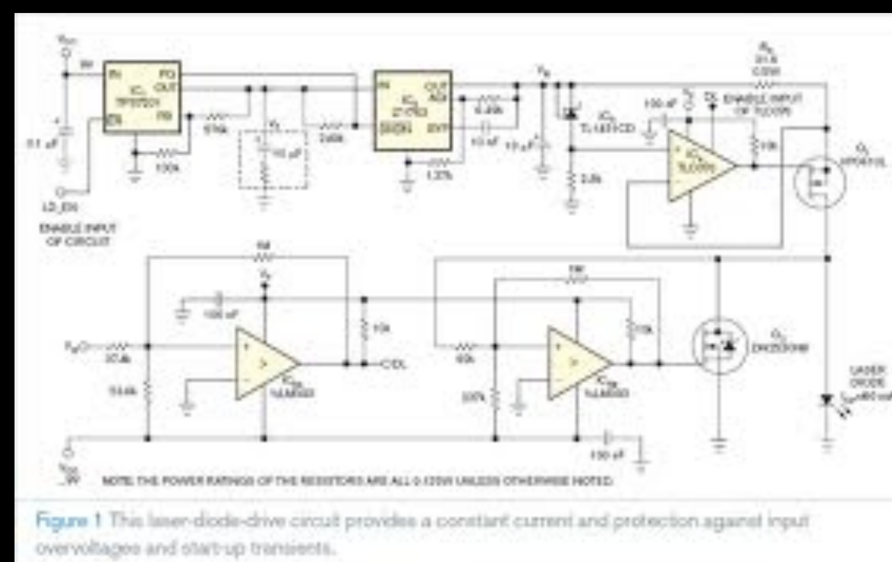
...

x_i

...

x_n

Analog circuit design

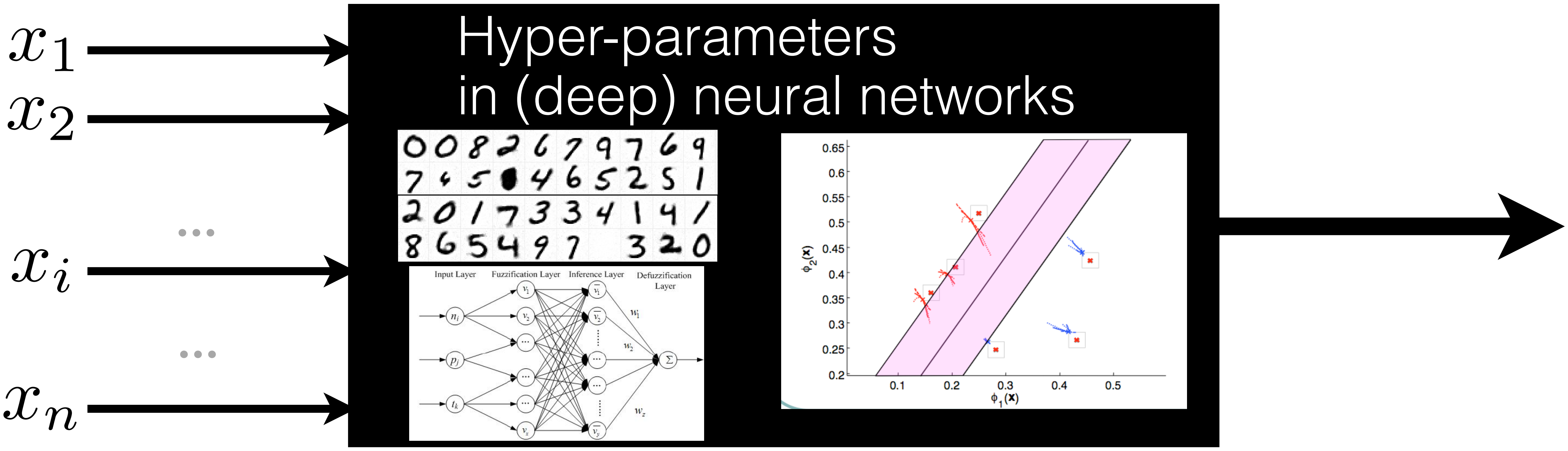


BLACK-BOX OPTIMIZATION

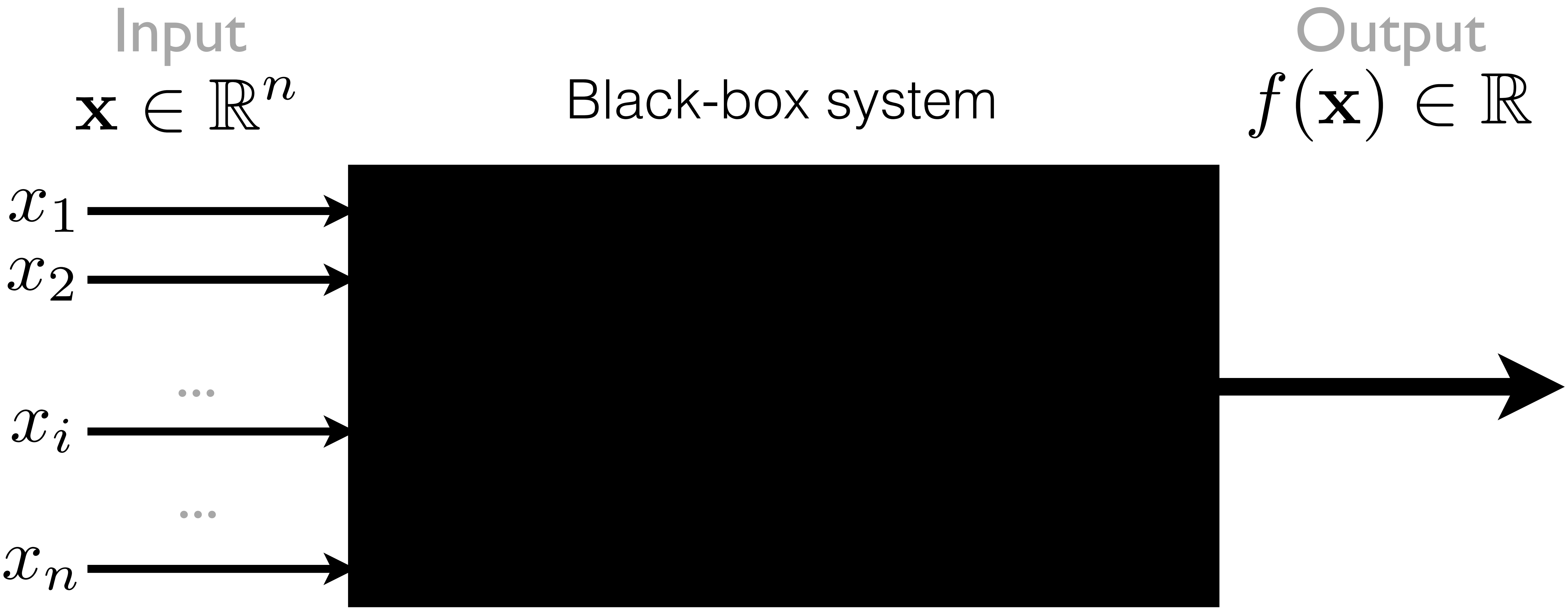
Input
$\mathbf{x} \in \mathbb{R}^n$

Black-box system

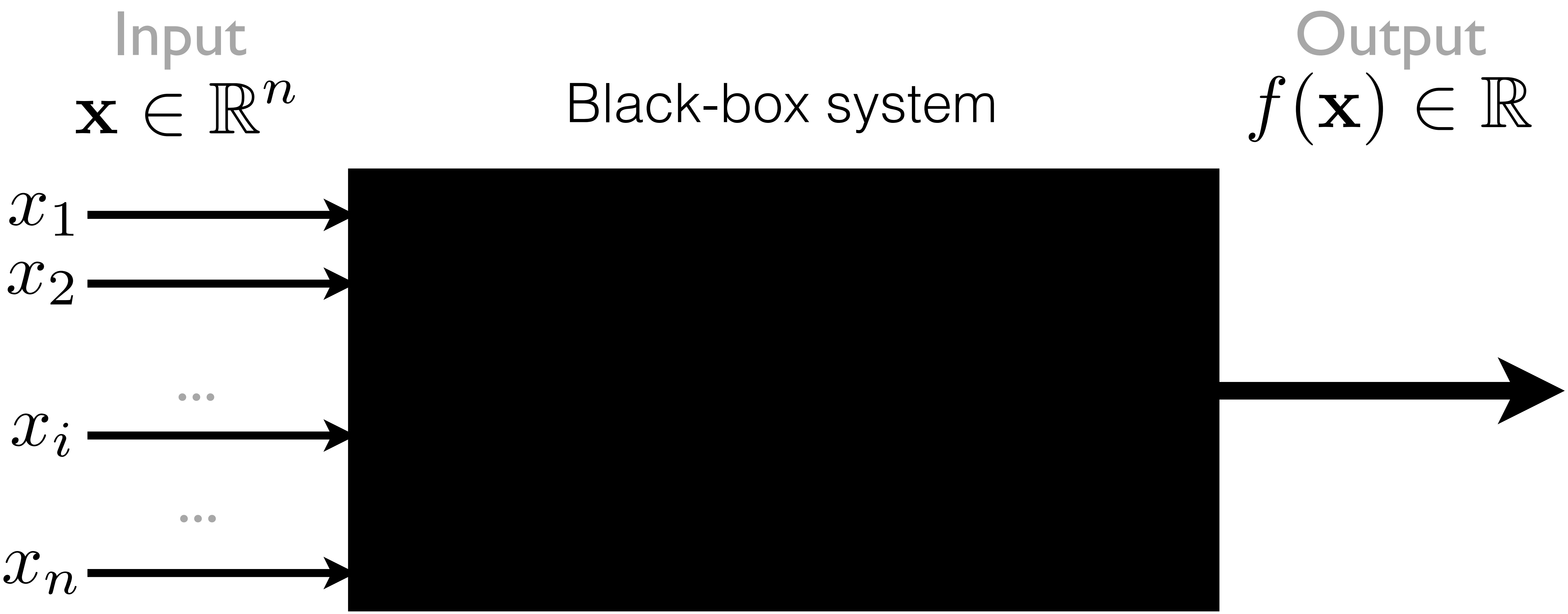
Output
$f(\mathbf{x}) \in \mathbb{R}$



BLACK-BOX OPTIMIZATION



BLACK-BOX OPTIMIZATION



- Variables
- Parameters
- Configuration
- Factors

BLACK-BOX OPTIMIZATION

Input

$\mathbf{x} \in \mathbb{R}^n$

Black-box system

Output

$f(\mathbf{x}) \in \mathbb{R}$

- Variables
- Parameters
- Configuration
- Factors

- Cost
- Criterion
- Objective
- Energy
- Fitness

BLACK-BOX OPTIMIZATION

FLATIRON INSTITUTE
Center for Computational Mathematics

Input

$\mathbf{x} \in \mathbb{R}^n$

Black-box system

Output

$f(\mathbf{x}) \in \mathbb{R}$

- Variables
- Parameters
- Configuration
- Factors

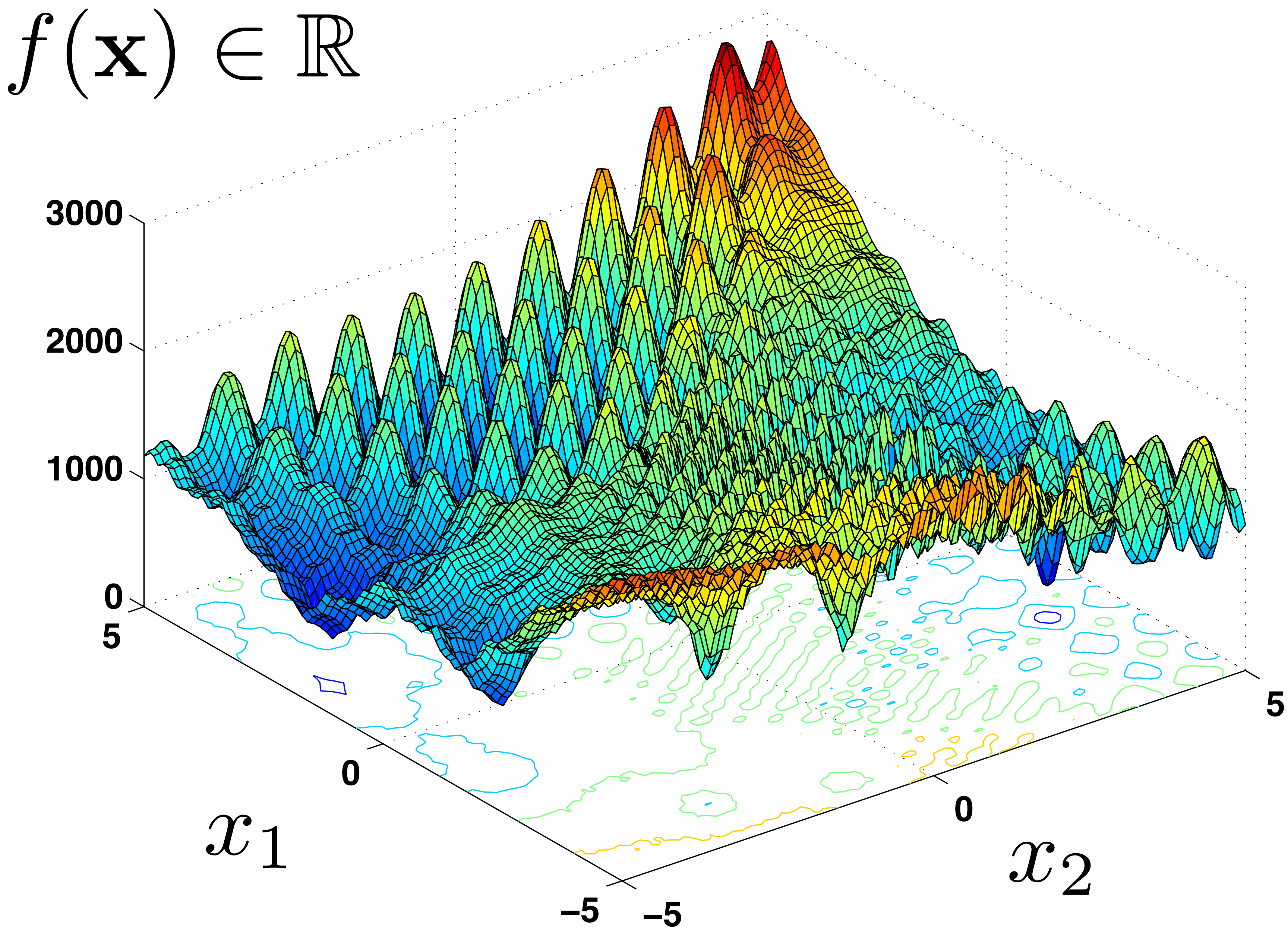
Find x* such that f(x*)≤f(x) ∀x

- Cost
- Criterion
- Objective
- Energy
- Fitness

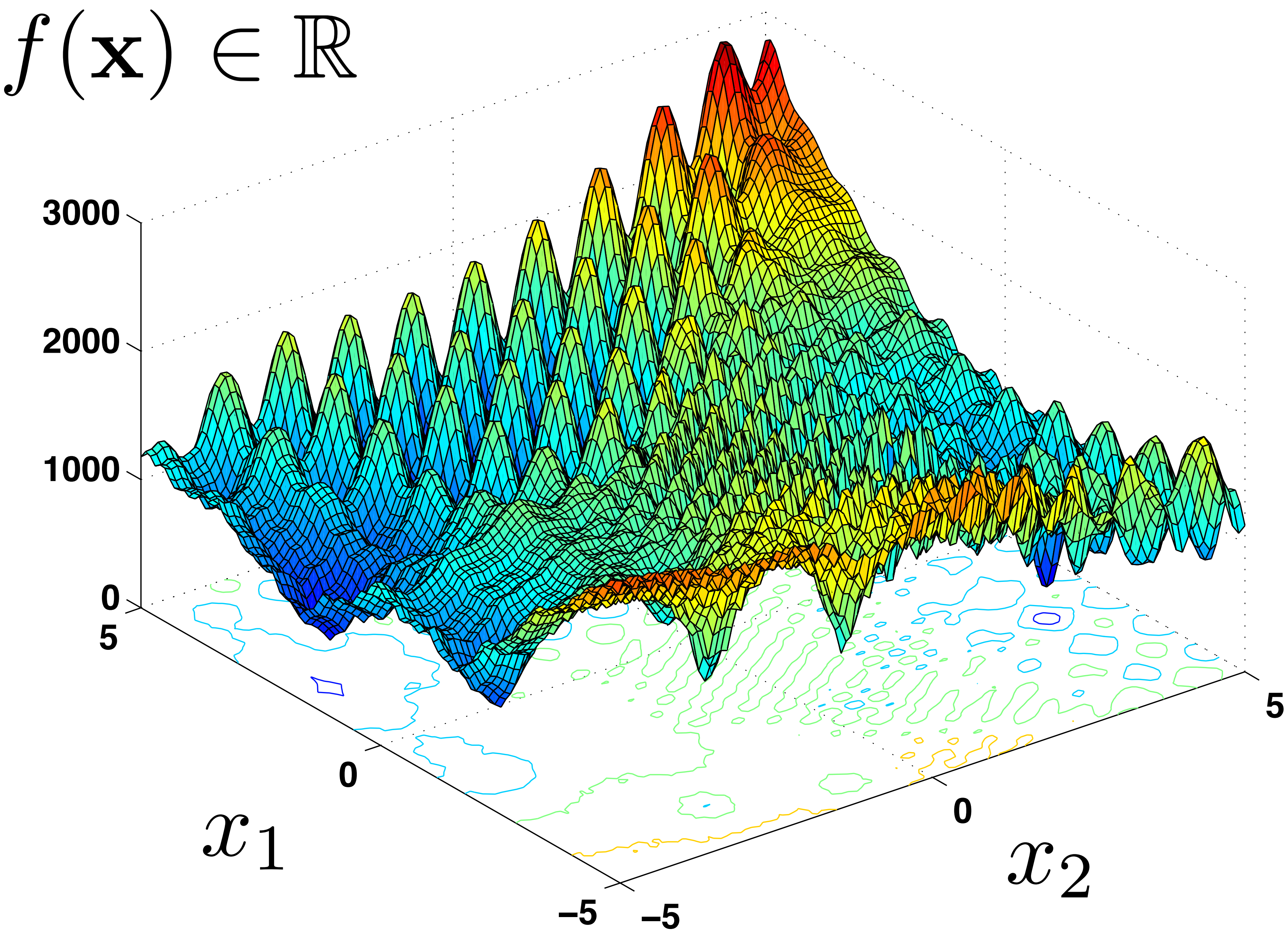
OPTIMIZATION LANDSCAPES

OPTIMIZATION LANDSCAPES

FLATIRON INSTITUTE
Center for Computational Mathematics

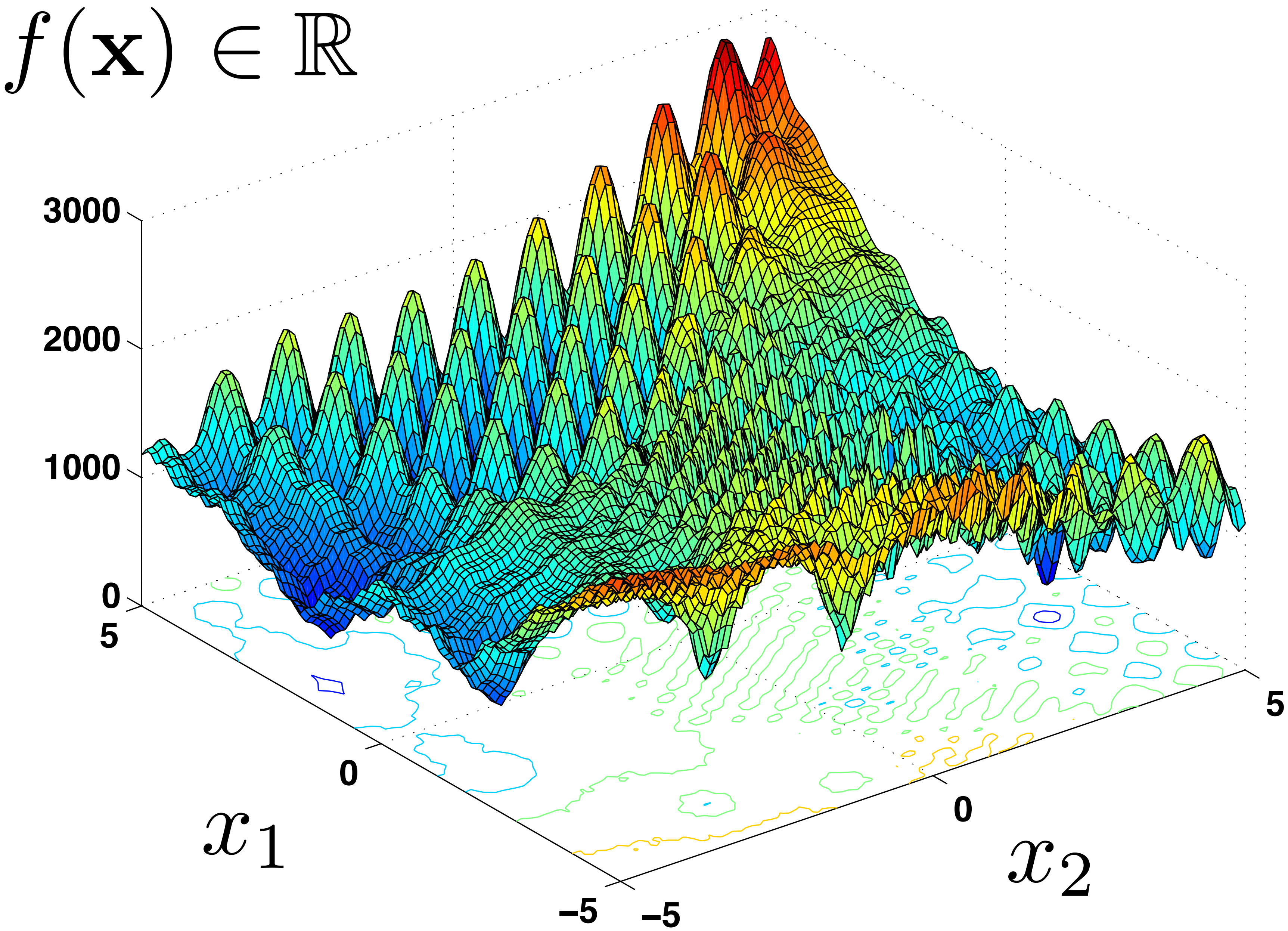


BLACK-BOX LANDSCAPES $\mathcal{L}_{\mathrm{B}}$



BLACK-BOX LANDSCAPES $\mathcal{L}_{\mathrm{B}}$

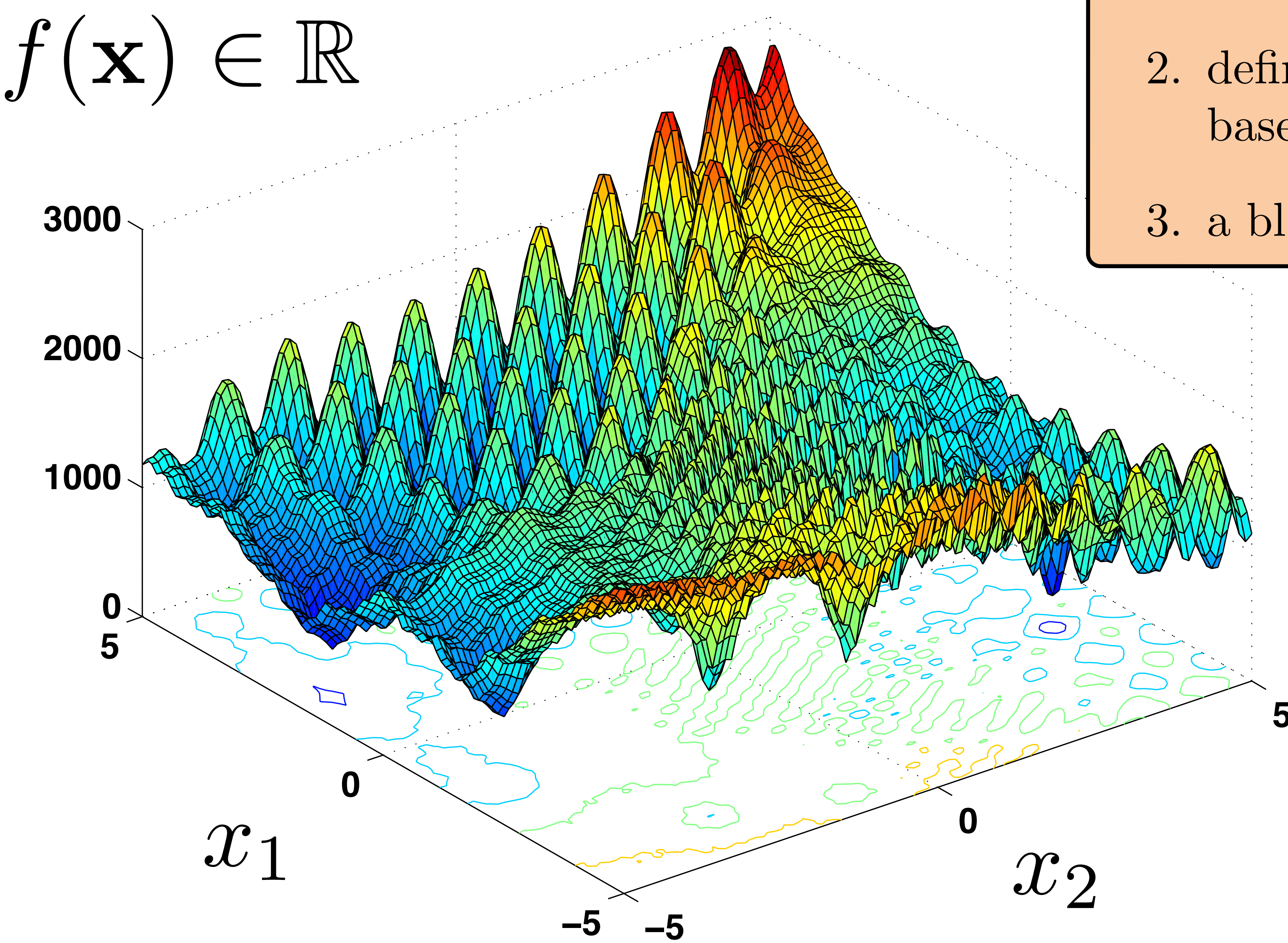
FLATIRON INSTITUTE
Center for Computational Mathematics



BLACK-BOX LANDSCAPES $\mathcal{L}_{\mathrm{B}}$

$\mathcal{L}_{\mathrm{B}}$ is the triple $(\mathcal{X}, d_{\mathrm{X}}, f)$ consisting of

1. $\mathcal{X} = [\mathbf{l}, \mathbf{u}] \subset \mathbb{R}^n$ with $\mathbf{l}, \mathbf{u} \in \mathbb{R}^n$.
2. definition of neighborhood/similarity based on a distance d_{X}.
3. a black-box function f.

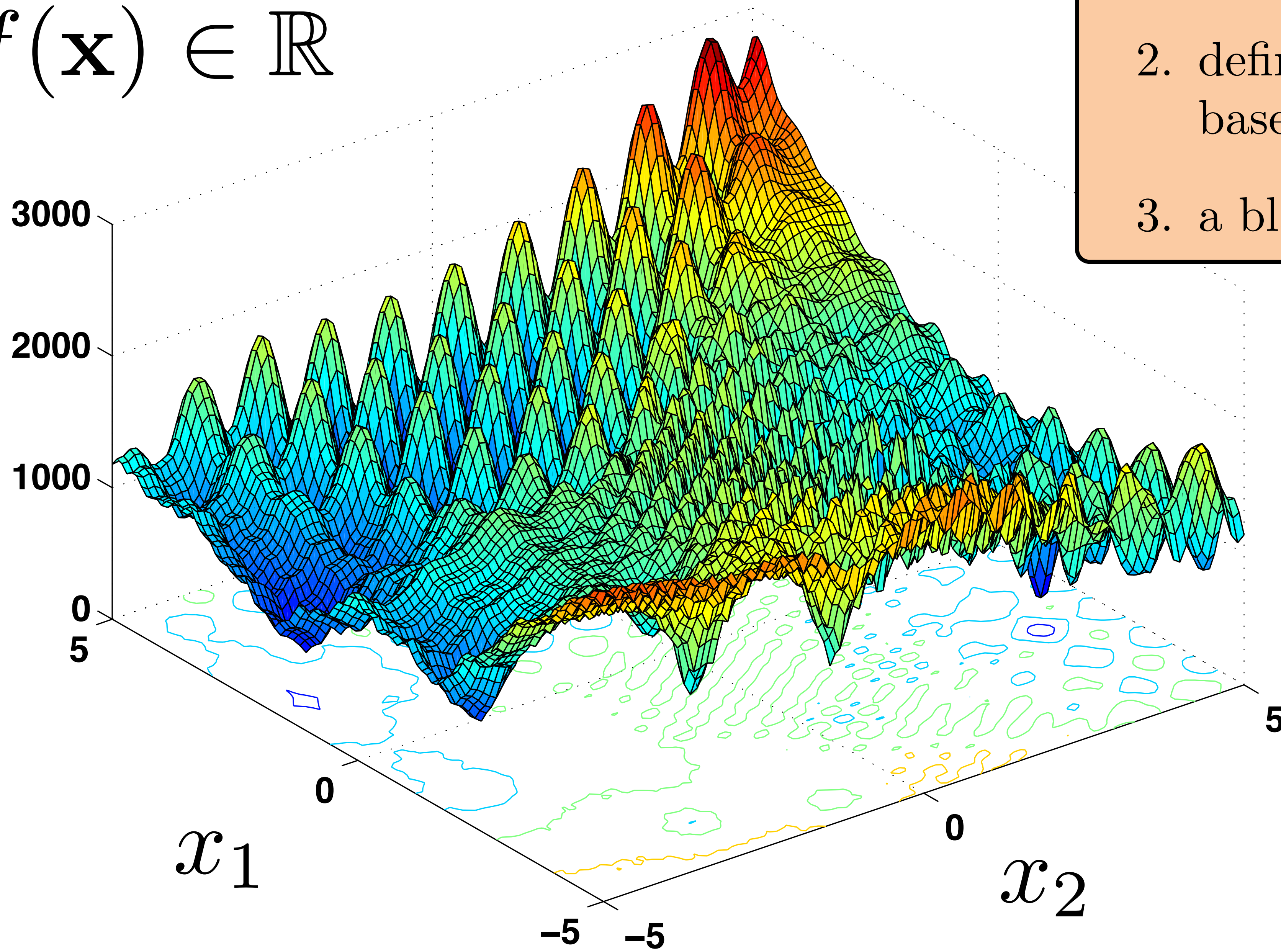


BLACK-BOX LANDSCAPES $\mathcal{L}_{\mathrm{B}}$

$\mathcal{L}_{\mathrm{B}}$ is the triple $(\mathcal{X}, d_{\mathrm{X}}, f)$ consisting of

1. $\mathcal{X} = [\mathbf{l}, \mathbf{u}] \subset \mathbb{R}^n$ with $\mathbf{l}, \mathbf{u} \in \mathbb{R}^n$.
2. definition of neighborhood/similarity based on a distance d_{X}.
3. a black-box function f.

$f(\mathbf{x}) \in \mathbb{R}$



Topographic description:

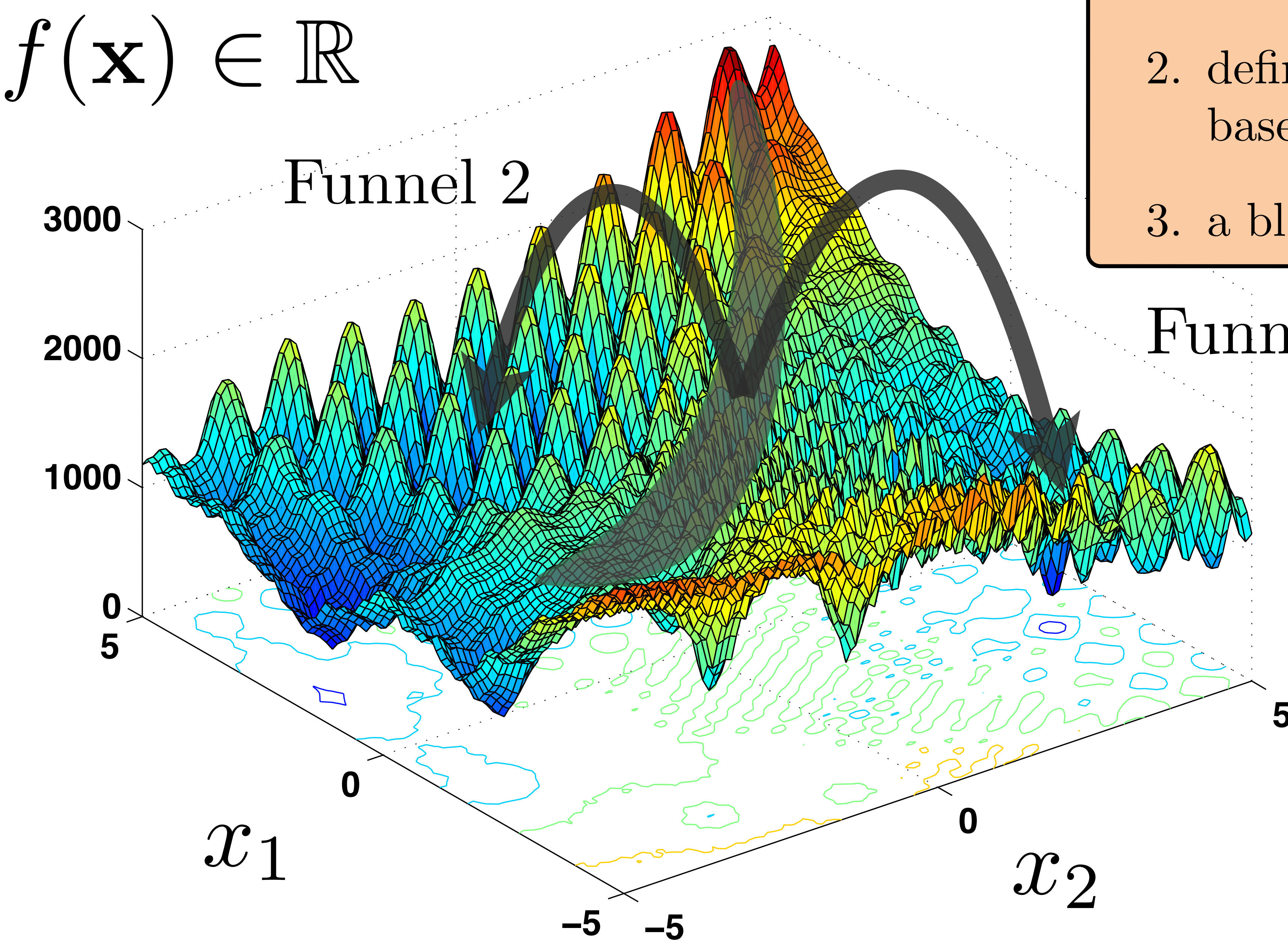
- Peaks and valleys
- Plateaus and basins
- Ridges and funnels

BLACK-BOX LANDSCAPES $\mathcal{L}_{\mathrm{B}}$

FLATIRON INSTITUTE
Center for Computational Mathematics

$\mathcal{L}_{\mathrm{B}}$ is the triple $(\mathcal{X}, d_{\mathrm{X}}, f)$ consisting of

1. $\mathcal{X} = [\mathbf{l}, \mathbf{u}] \subset \mathbb{R}^n$ with $\mathbf{l}, \mathbf{u} \in \mathbb{R}^n$.
2. definition of neighborhood/similarity based on a distance d_{X}.
3. a black-box function f.



Topographic description:

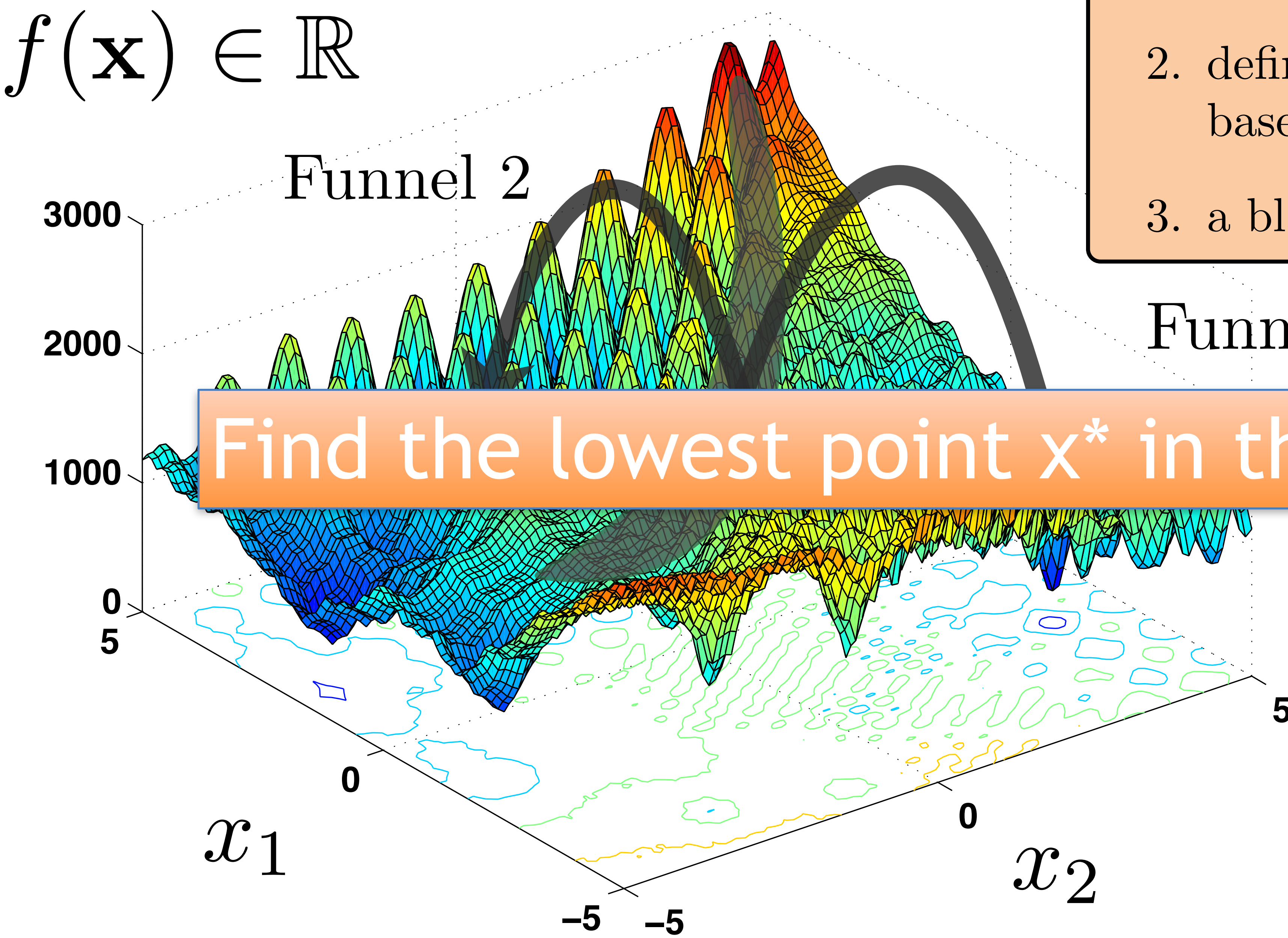
- Peaks and valleys
- Plateaus and basins
- Ridges and funnels

BLACK-BOX LANDSCAPES $\mathcal{L}_\mathrm{B}$

FLATIRON INSTITUTE
Center for Computational Mathematics

$\mathcal{L}_\mathrm{B}$ is the triple $(\mathcal{X}, d_\mathrm{X}, f)$ consisting of

1. $\mathcal{X} = [\mathbf{l}, \mathbf{u}] \subset \mathbb{R}^n$ with $\mathbf{l}, \mathbf{u} \in \mathbb{R}^n$.
2. definition of neighborhood/similarity based on a distance d_X.
3. a black-box function f.

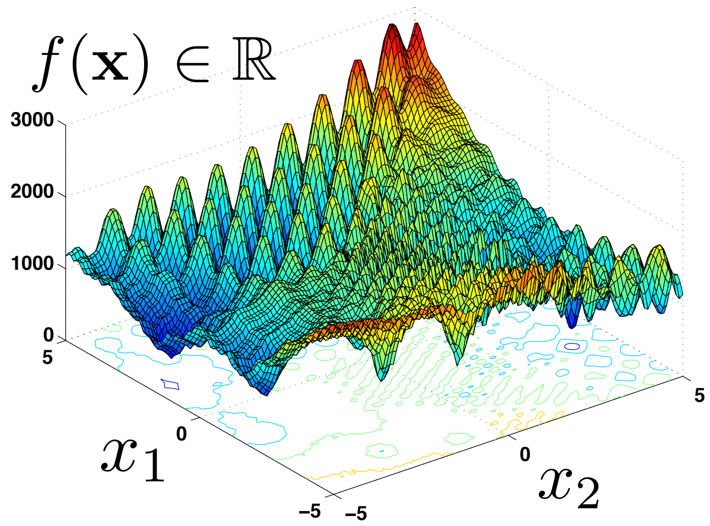


Find the lowest point x* in the landscape!

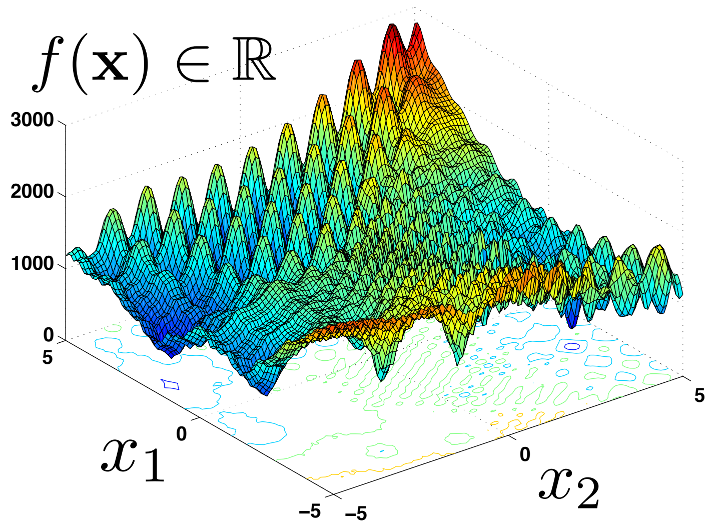
Topographic description:

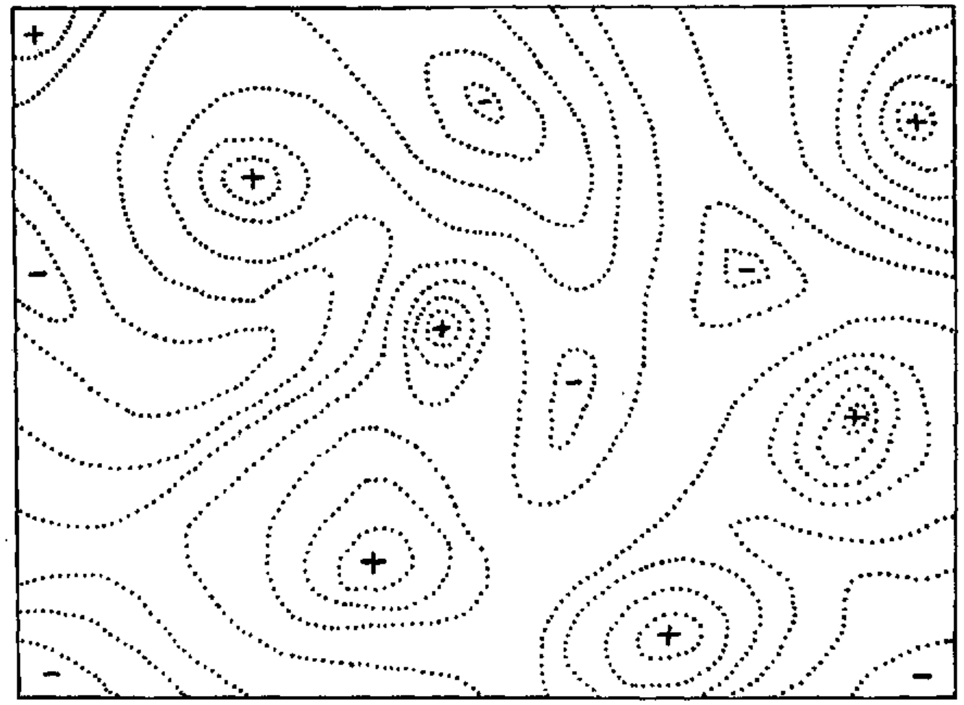
- Peaks and valleys
- Plateaus and basins
- Ridges and funnels

LANDSCAPES IN SCIENCE



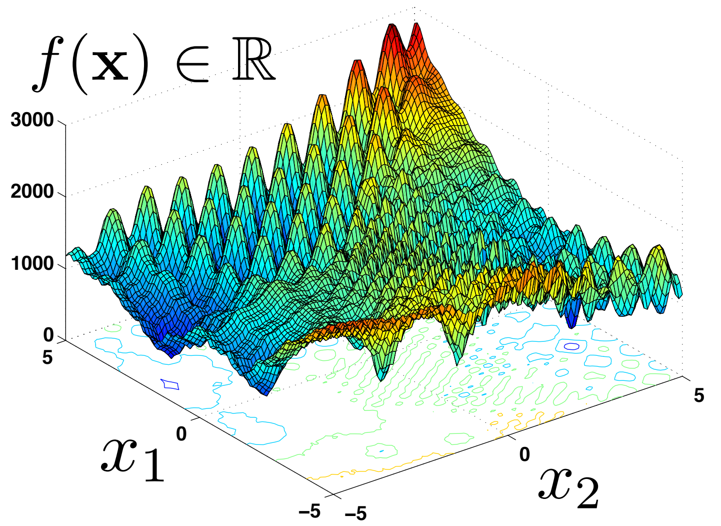
LANDSCAPES IN SCIENCE

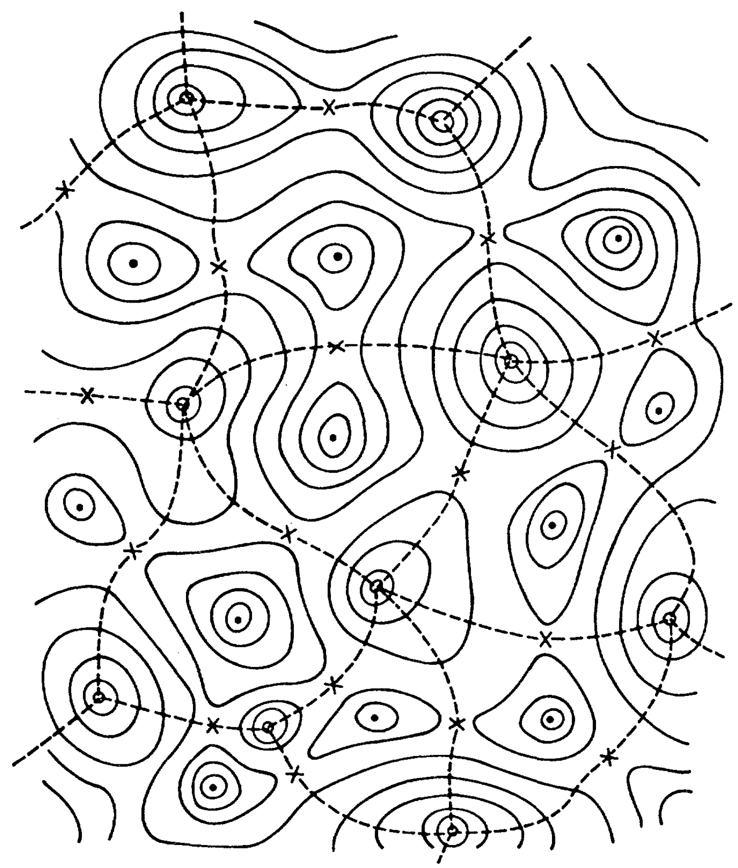




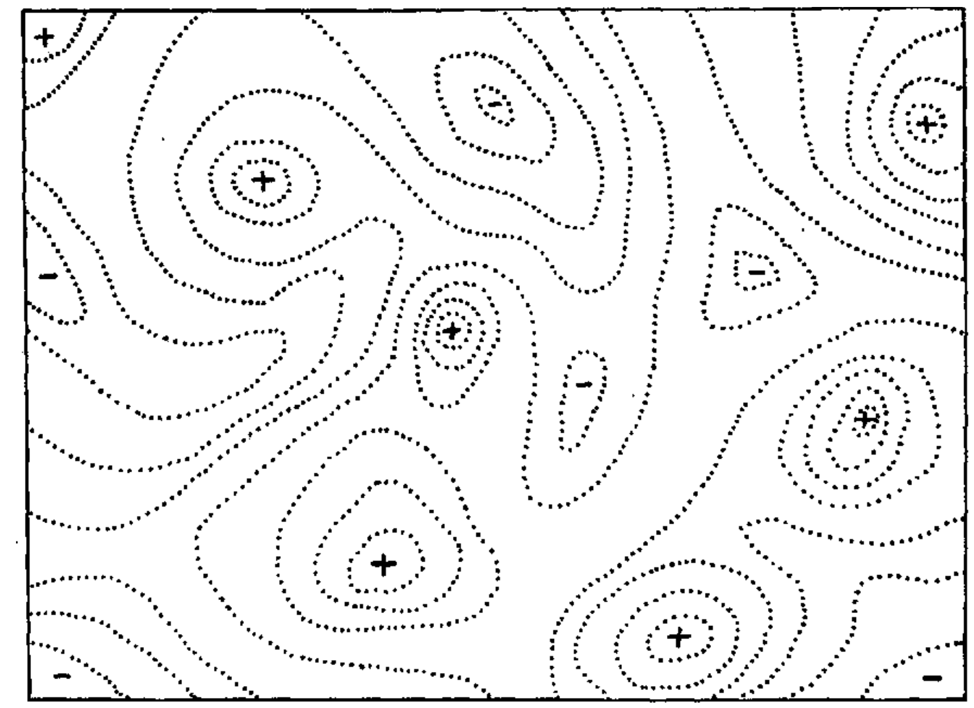
Fitness landscape

LANDSCAPES IN SCIENCE



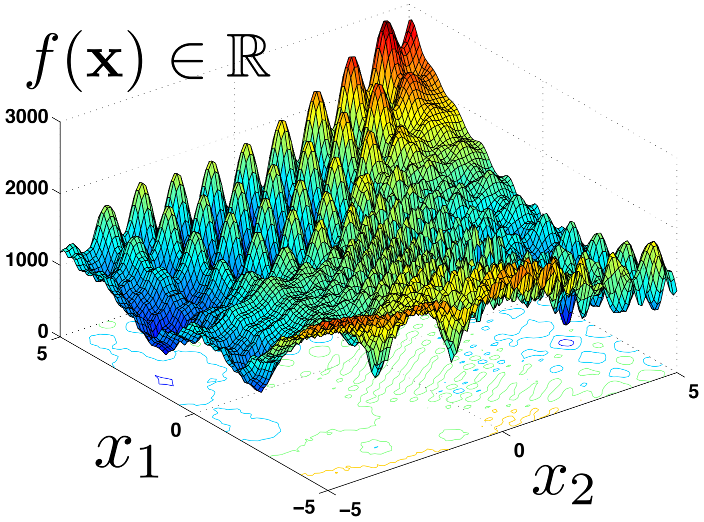


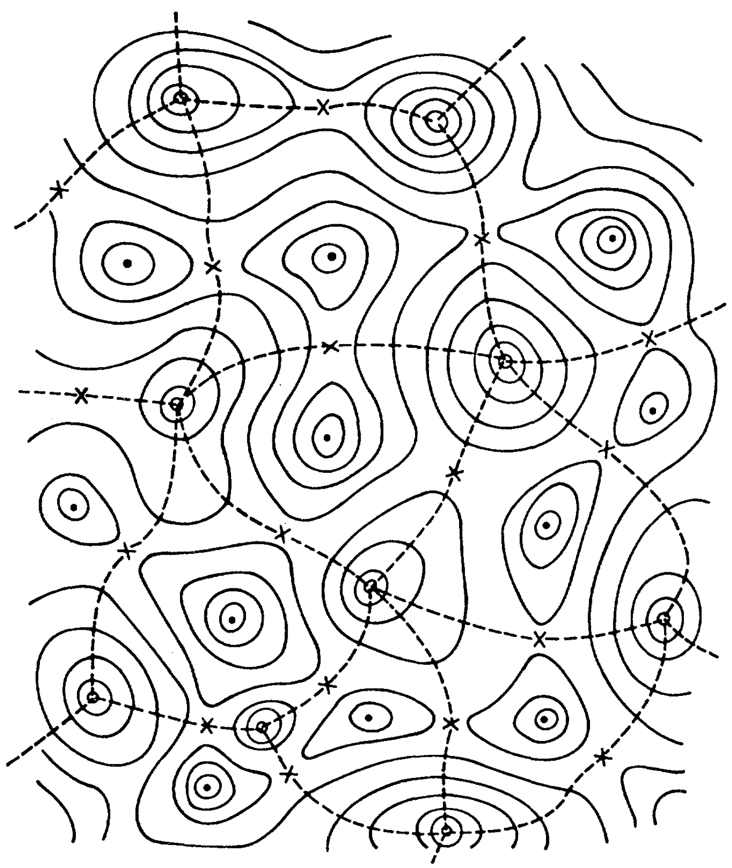
Potential energy landscape



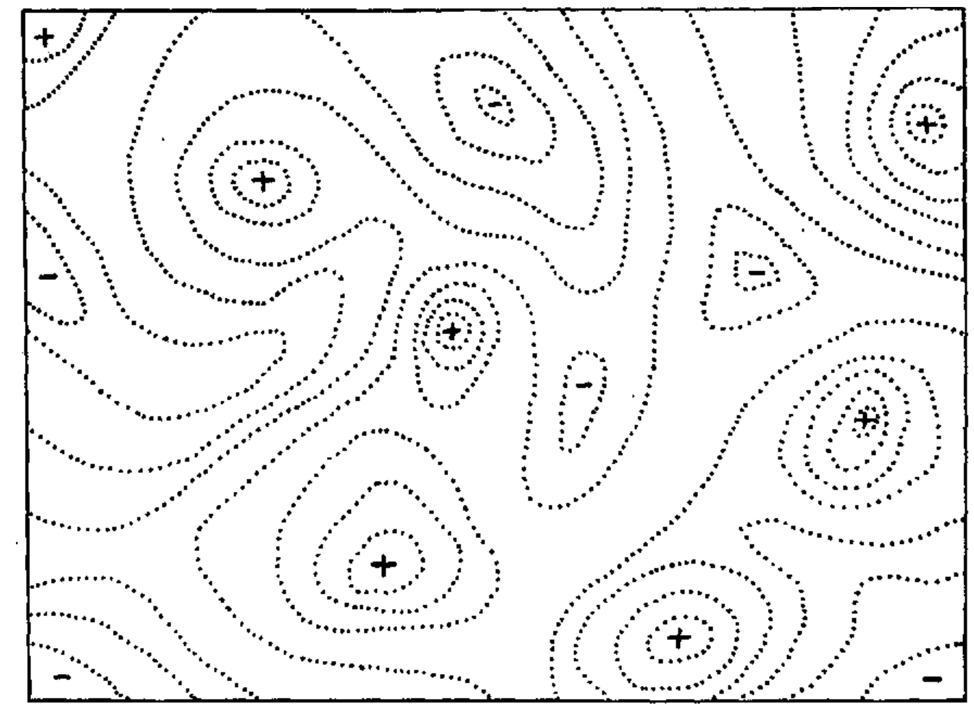
Fitness landscape

LANDSCAPES IN SCIENCE

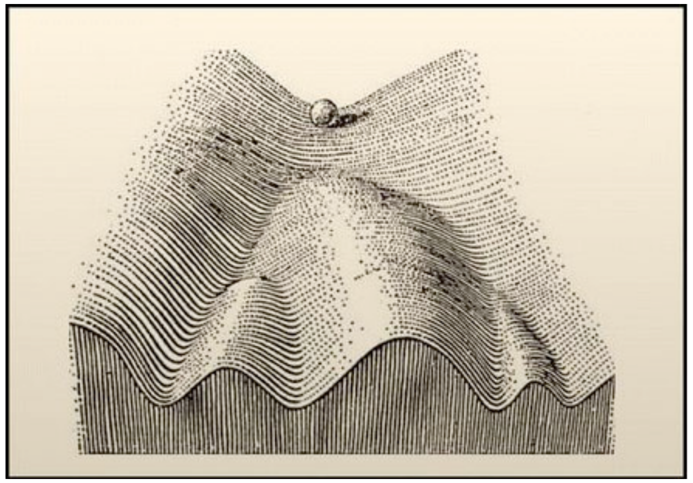




Potential energy landscape

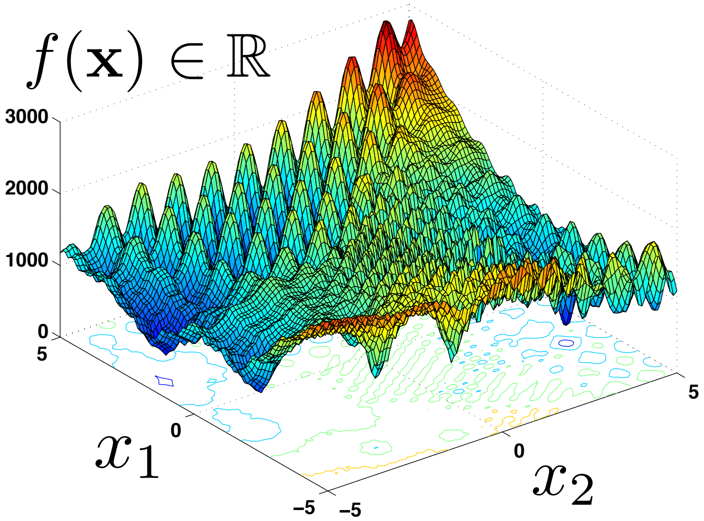


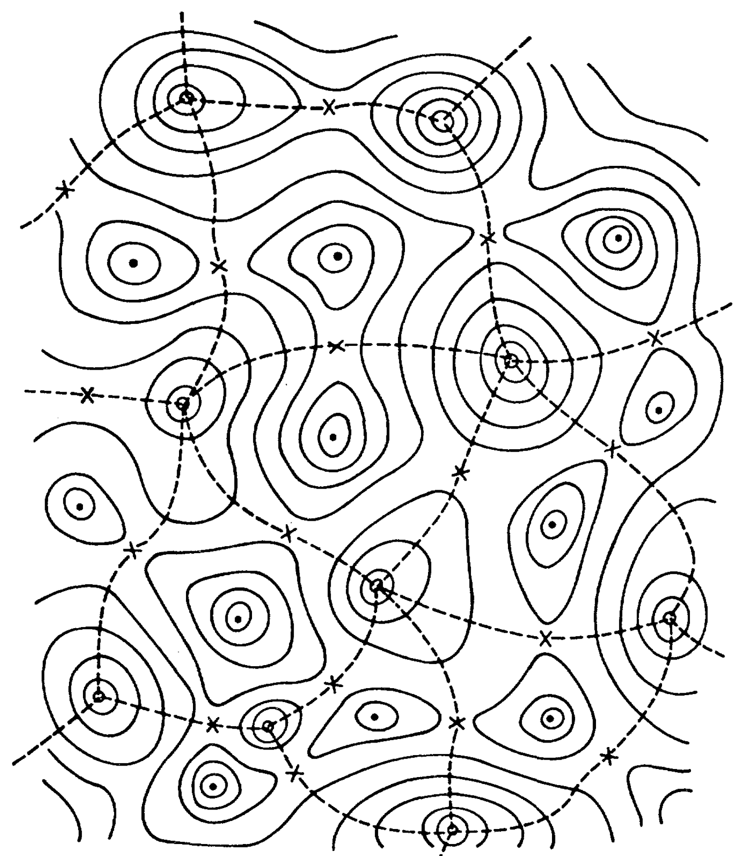
Fitness landscape



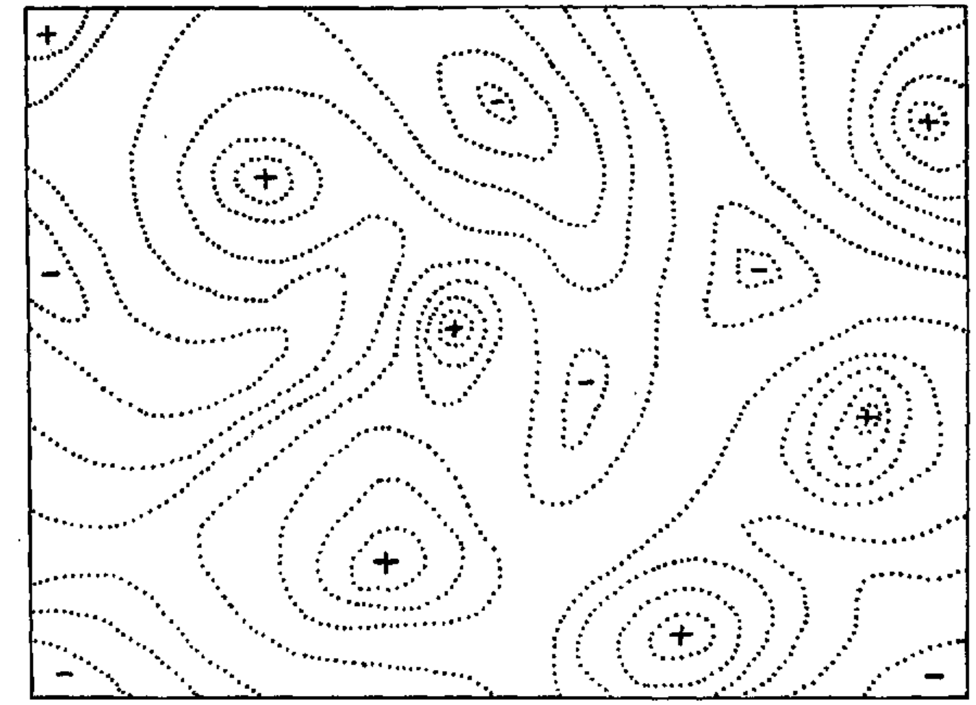
Epigenetic landscape

LANDSCAPES IN SCIENCE

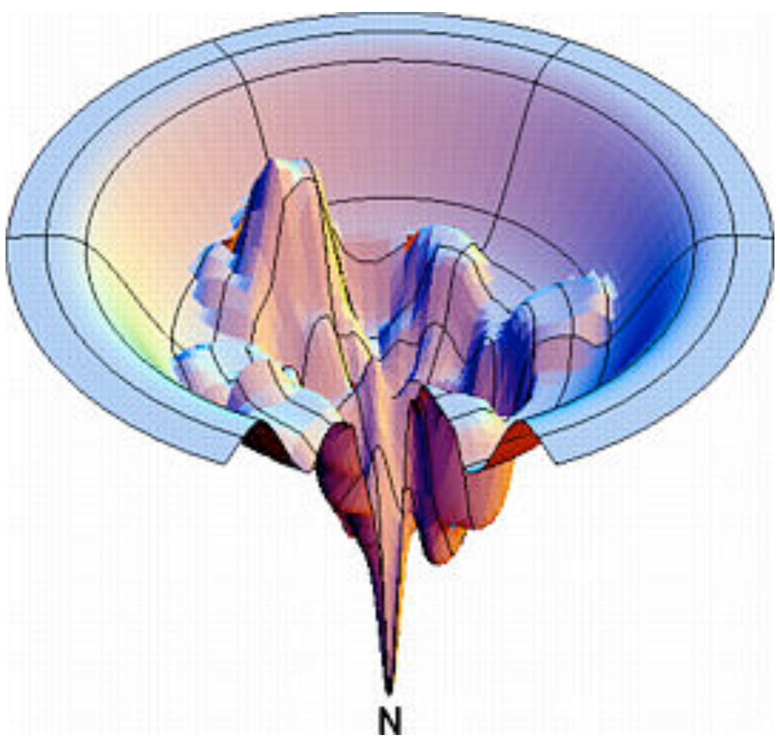




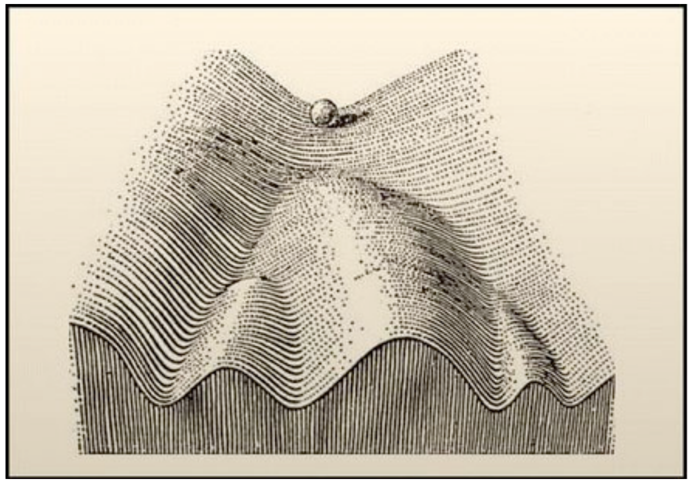
Potential energy landscape



Fitness landscape



Folding funnel



Epigenetic landscape

LANDSCAPES ARE METAPHORS

“The price of metaphor is eternal vigilance."

Norbert Wiener

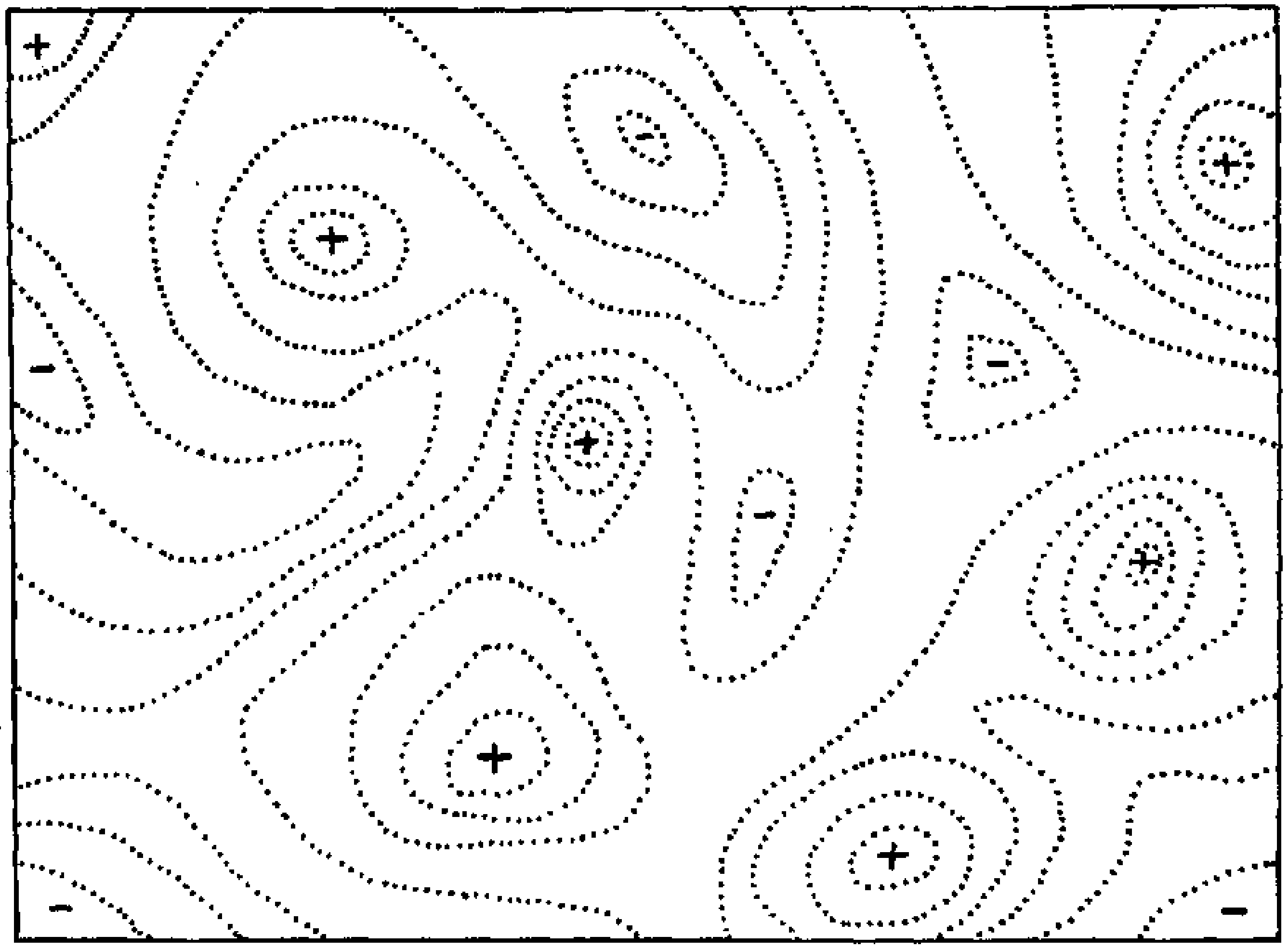
La condition humaine, René Magritte

FITNESS LANDSCAPES

FLATIRON INSTITUTE
Center for Computational Mathematics

Wright, S., “The Roles of Mutation, Inbreeding, Crossbreeding, and Selection in Evolution,”

Proceedings of the Sixth International Congress on Genetics, 1932.

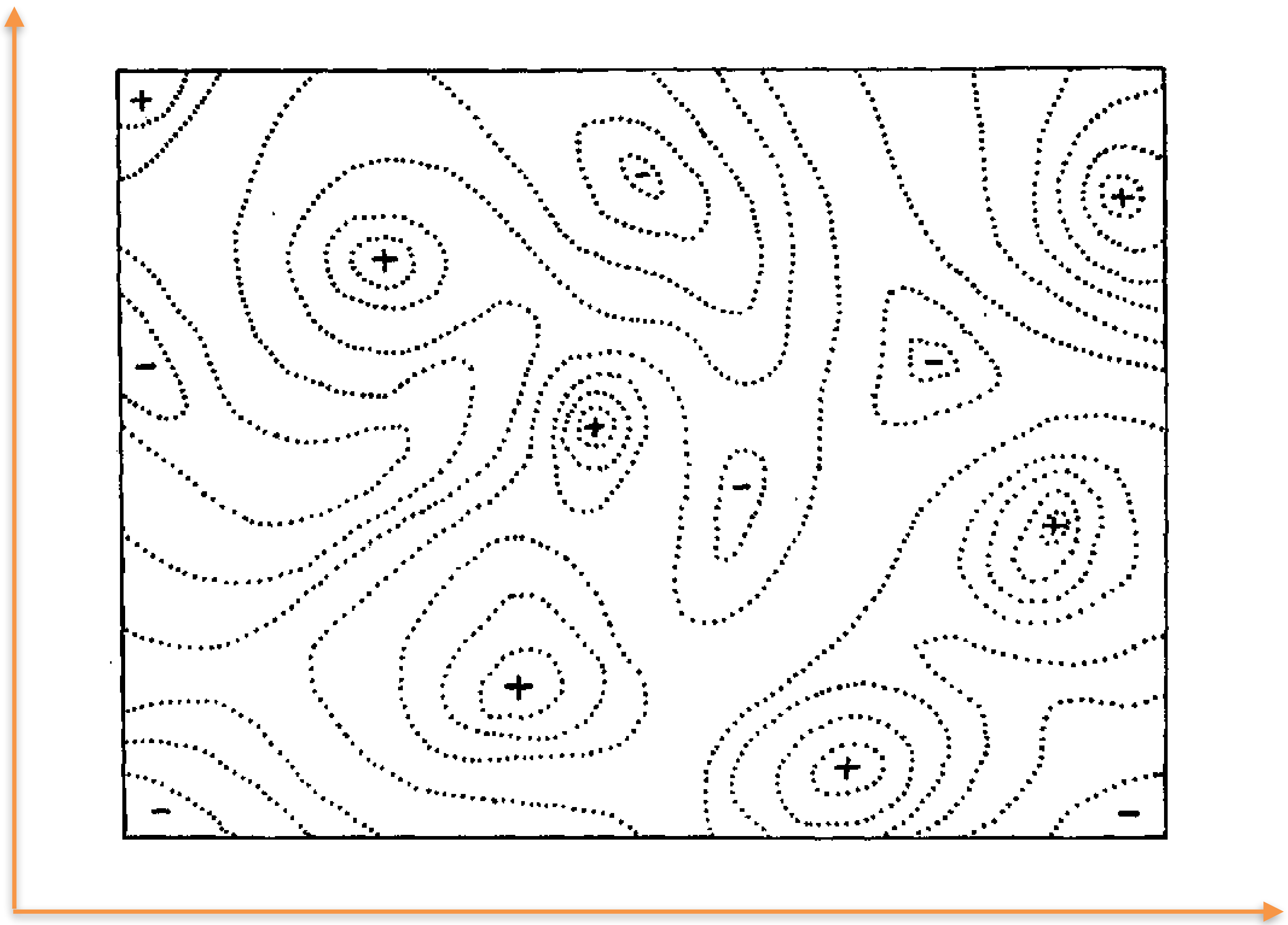


FITNESS LANDSCAPES

FLATIRON INSTITUTE
Center for Computational Mathematics

Wright, S., "The Roles of Mutation, Inbreeding, Crossbreeding, and Selection in Evolution," Proceedings of the Sixth International Congress on Genetics, 1932.

gene 1/
trait 1/…



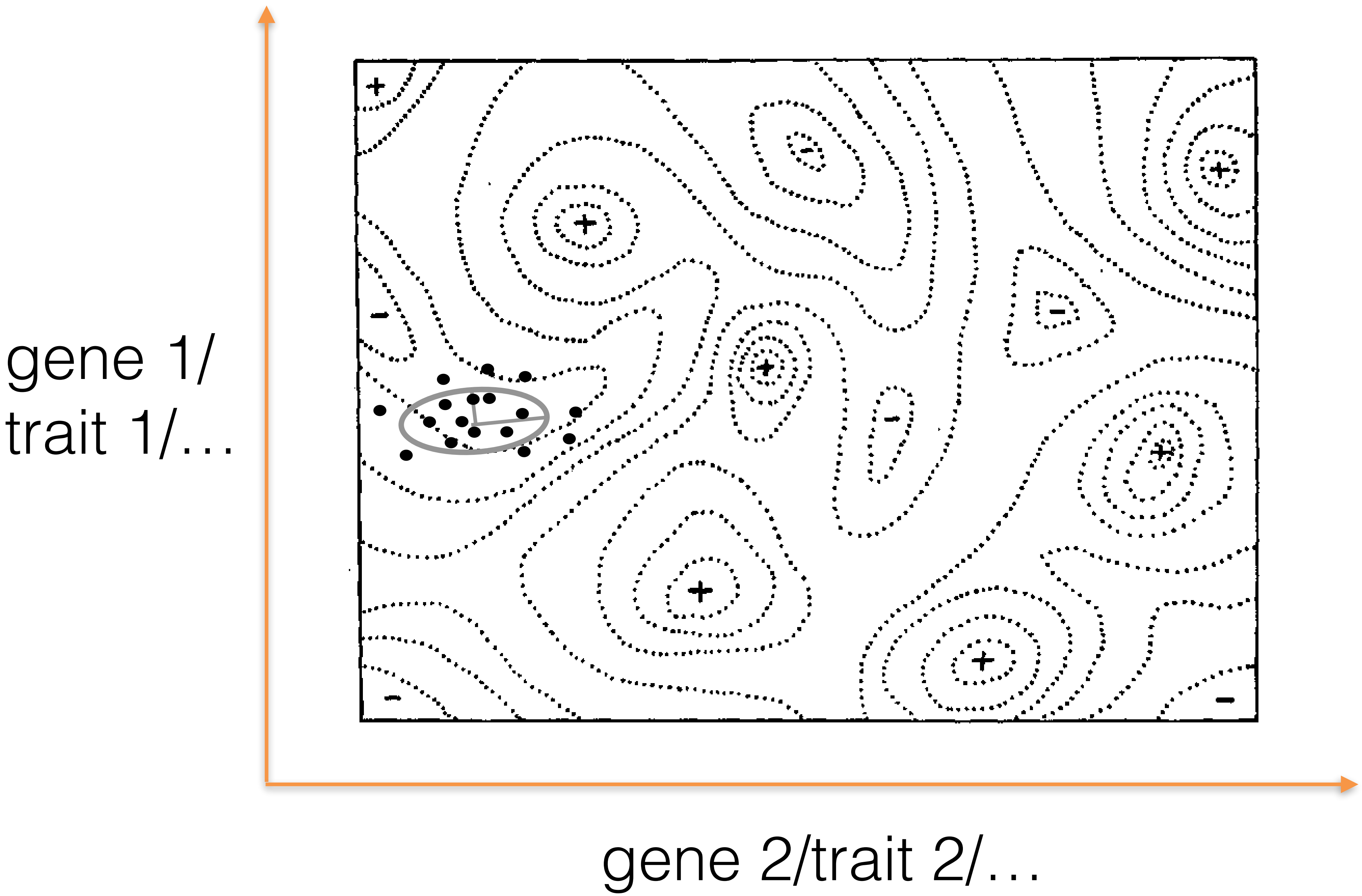
gene 2/trait 2/…

FITNESS LANDSCAPES

FLATIRON INSTITUTE
Center for Computational Mathematics

Wright, S., "The Roles of Mutation, Inbreeding, Crossbreeding, and Selection in Evolution,"

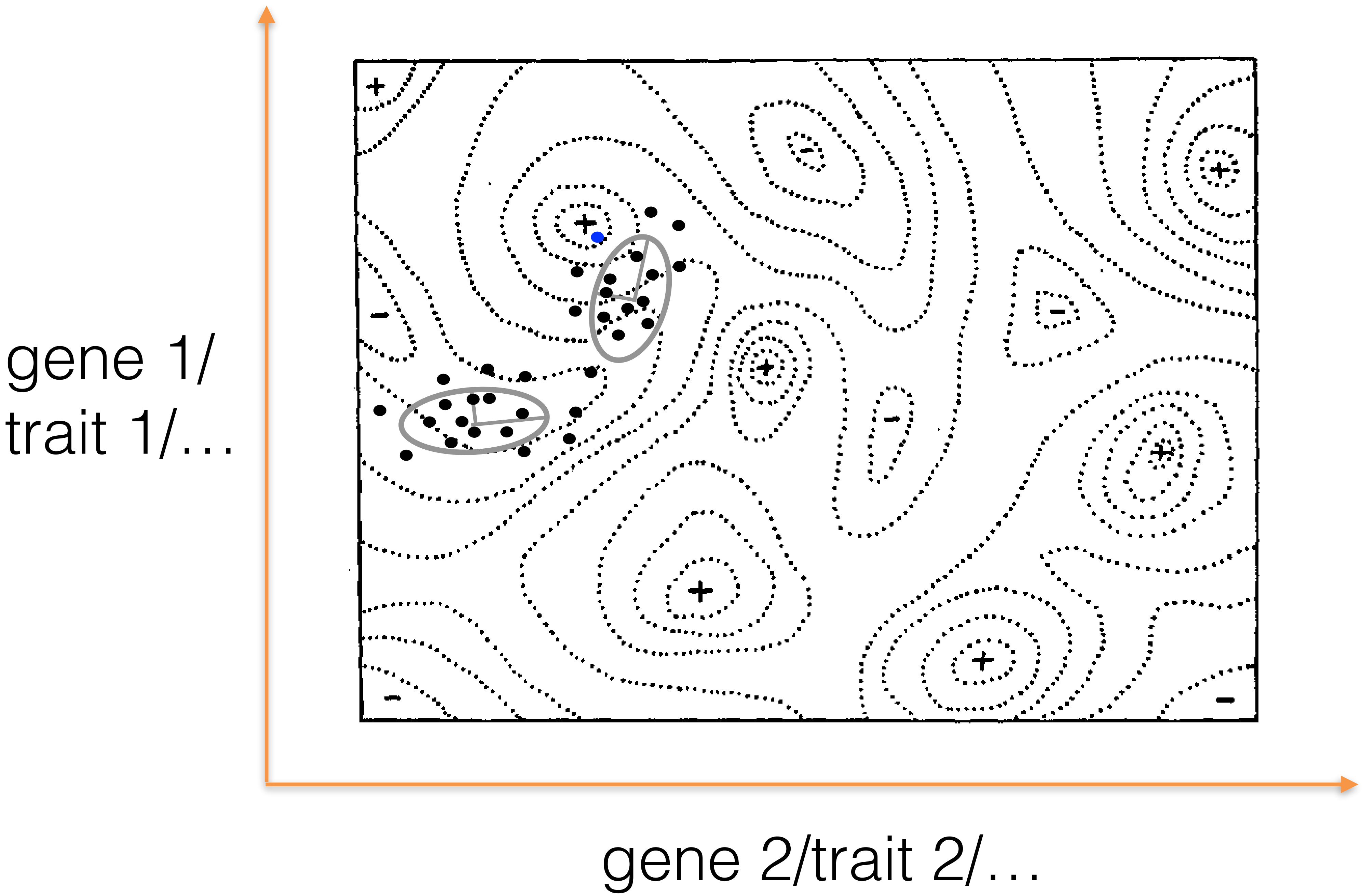
Proceedings of the Sixth International Congress on Genetics, 1932.



FITNESS LANDSCAPES

FLATIRON INSTITUTE
Center for Computational Mathematics

Wright, S., "The Roles of Mutation, Inbreeding, Crossbreeding, and Selection in Evolution," Proceedings of the Sixth International Congress on Genetics, 1932.

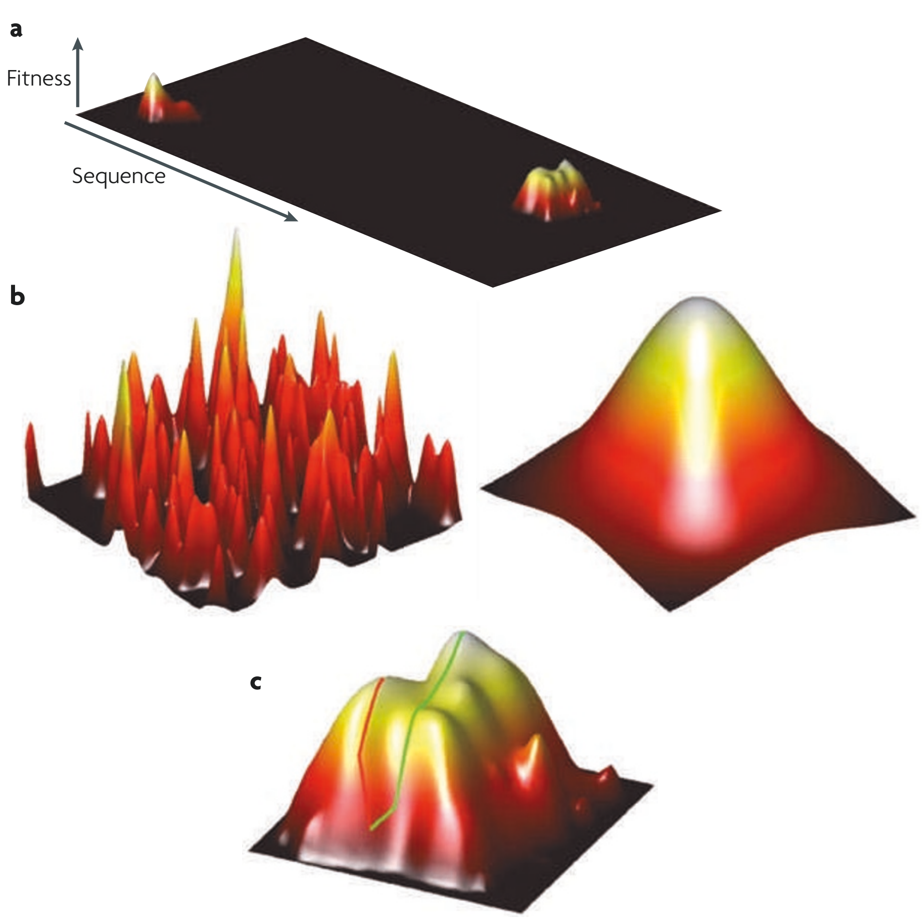


FITNESS LANDSCAPES

Exploring protein fitness landscapes by directed evolution

Philip A. Romero and Frances H. Arnold

Abstract | Directed evolution circumvents our profound ignorance of how a protein's sequence encodes its function by using iterative rounds of random mutation and artificial selection to discover new and useful proteins. Proteins can be tuned to adapt to new functions or environments by simple adaptive walks involving small numbers of mutations. Directed evolution studies have shown how rapidly some proteins can evolve under strong selection pressures and, because the entire 'fossil record' of evolutionary intermediates is available for detailed study, they have provided new insight into the relationship between sequence and function. Directed evolution has also shown how mutations that are functionally neutral can set the stage for further adaptation.



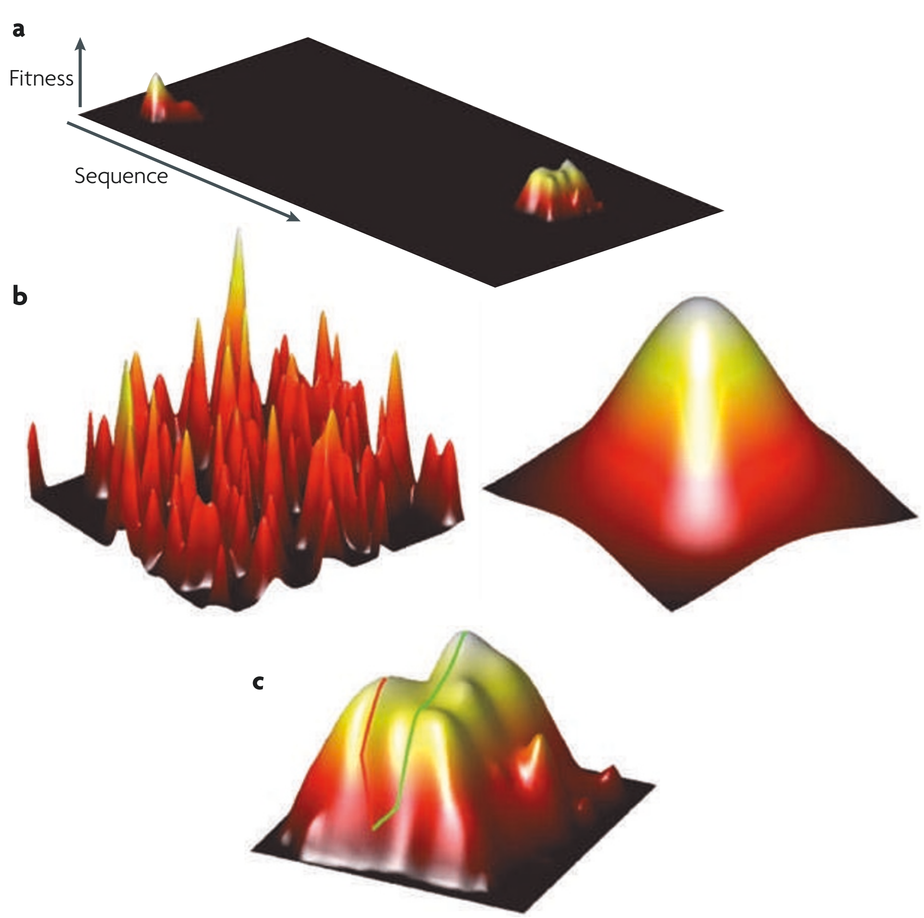
FITNESS LANDSCAPES

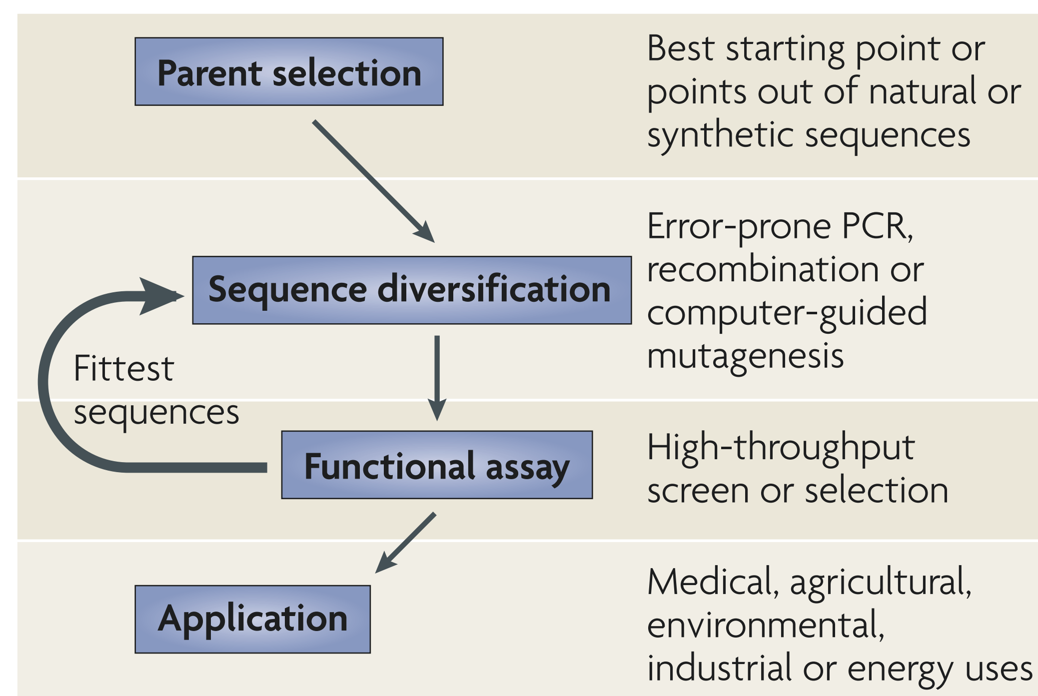
FLATIRON INSTITUTE
Center for Computational Mathematics

Exploring protein fitness landscapes by directed evolution

Philip A. Romero and Frances H. Arnold

Abstract | Directed evolution circumvents our profound ignorance of how a protein's sequence encodes its function by using iterative rounds of random mutation and artificial selection to discover new and useful proteins. Proteins can be tuned to adapt to new functions or environments by simple adaptive walks involving small numbers of mutations. Directed evolution studies have shown how rapidly some proteins can evolve under strong selection pressures and, because the entire 'fossil record' of evolutionary intermediates is available for detailed study, they have provided new insight into the relationship between sequence and function. Directed evolution has also shown how mutations that are functionally neutral can set the stage for further adaptation.





FITNESS LANDSCAPES

FLATIRON INSTITUTE
Center for Computational Mathematics

PNAS PLUS

Navigating the protein fitness landscape with Gaussian processes

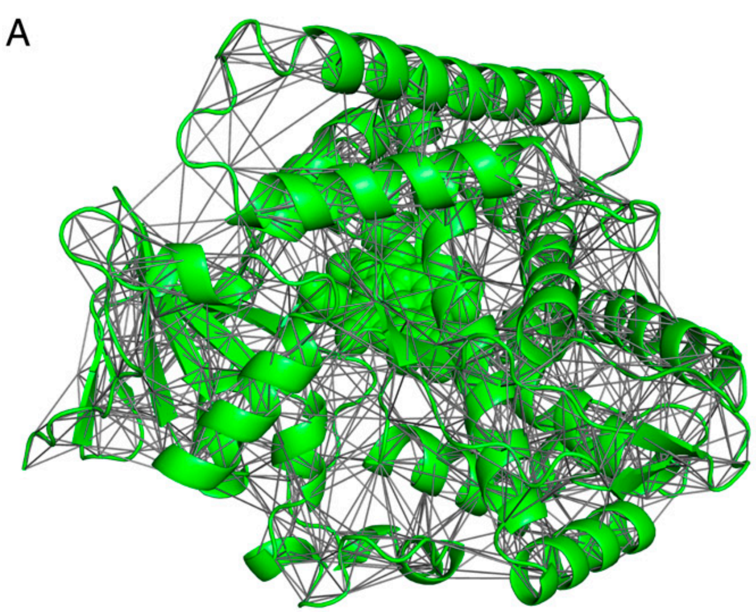
Philip A. Romero[a], Andreas Krause[b], and Frances H. Arnold[a,1]

[a]Division of Chemistry and Chemical Engineering, California Institute of Technology, Pasadena, CA 91125; and [b]Department of Computer Science, Swiss Federal Institute of Technology, 8092 Zurich, Switzerland

Edited by Michael Levitt, Stanford University School of Medicine, Stanford, CA, and approved November 28, 2012 (received for review September 9, 2012)

Enzyme to be optimized

A



FITNESS LANDSCAPES

PNAS PLUS

Navigating the protein fitness landscape with Gaussian processes

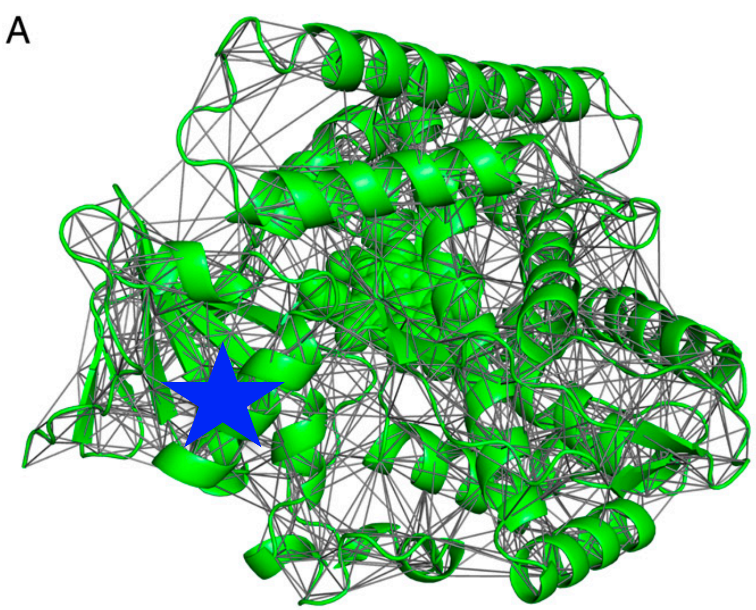
Philip A. Romero[a], Andreas Krause[b], and Frances H. Arnold[a,1]

[a]Division of Chemistry and Chemical Engineering, California Institute of Technology, Pasadena, CA 91125; and [b]Department of Computer Science, Swiss Federal Institute of Technology, 8092 Zurich, Switzerland

Edited by Michael Levitt, Stanford University School of Medicine, Stanford, CA, and approved November 28, 2012 (received for review September 9, 2012)

Enzyme to be optimized

A



FITNESS LANDSCAPES

FLATIRON INSTITUTE
Center for Computational Mathematics

PNAS PLUS

Navigating the protein fitness landscape with Gaussian processes

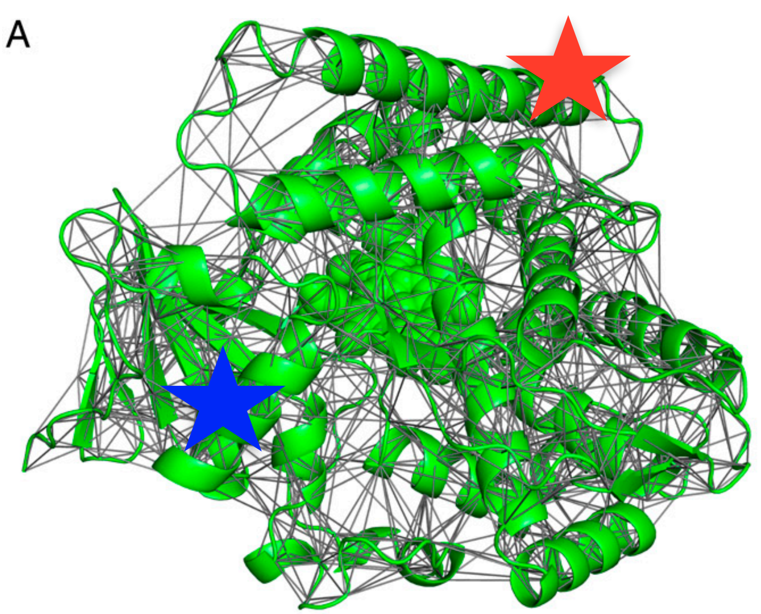
Philip A. Romero[a], Andreas Krause[b], and Frances H. Arnold[a,1]

[a]Division of Chemistry and Chemical Engineering, California Institute of Technology, Pasadena, CA 91125; and [b]Department of Computer Science, Swiss Federal Institute of Technology, 8092 Zurich, Switzerland

Edited by Michael Levitt, Stanford University School of Medicine, Stanford, CA, and approved November 28, 2012 (received for review September 9, 2012)

Enzyme to be optimized

A



FITNESS LANDSCAPES

PNAS PLUS

Navigating the protein fitness landscape with Gaussian processes

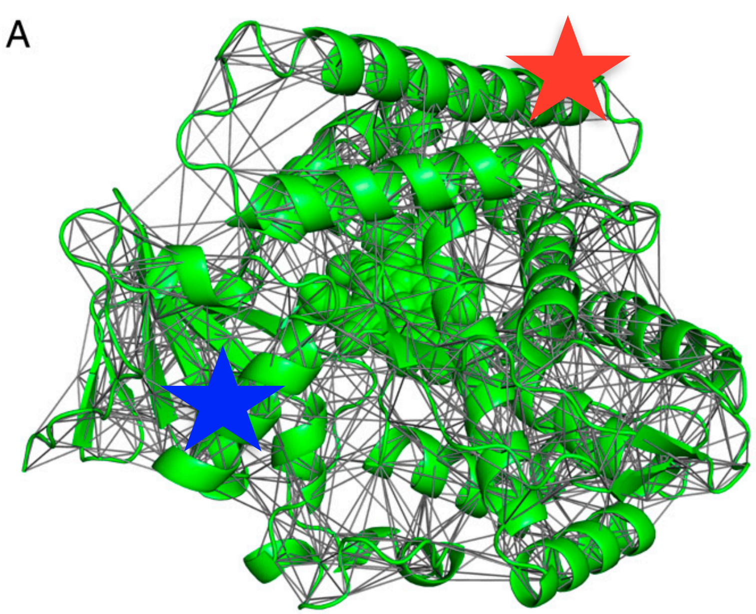
Philip A. Romero[a], Andreas Krause[b], and Frances H. Arnold[a,1]

[a]Division of Chemistry and Chemical Engineering, California Institute of Technology, Pasadena, CA 91125; and [b]Department of Computer Science, Swiss Federal Institute of Technology, 8092 Zurich, Switzerland

Edited by Michael Levitt, Stanford University School of Medicine, Stanford, CA, and approved November 28, 2012 (received for review September 9, 2012)

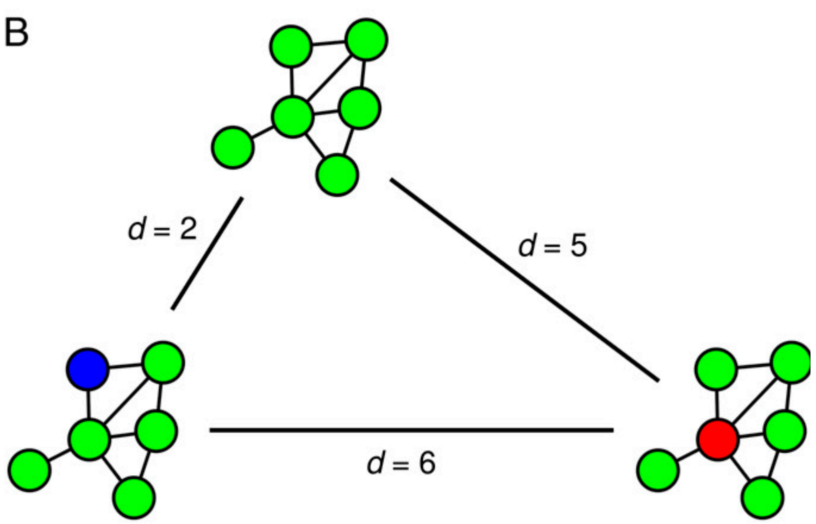
Enzyme to be optimized

A



Network representation and distance definition

B



FITNESS LANDSCAPES

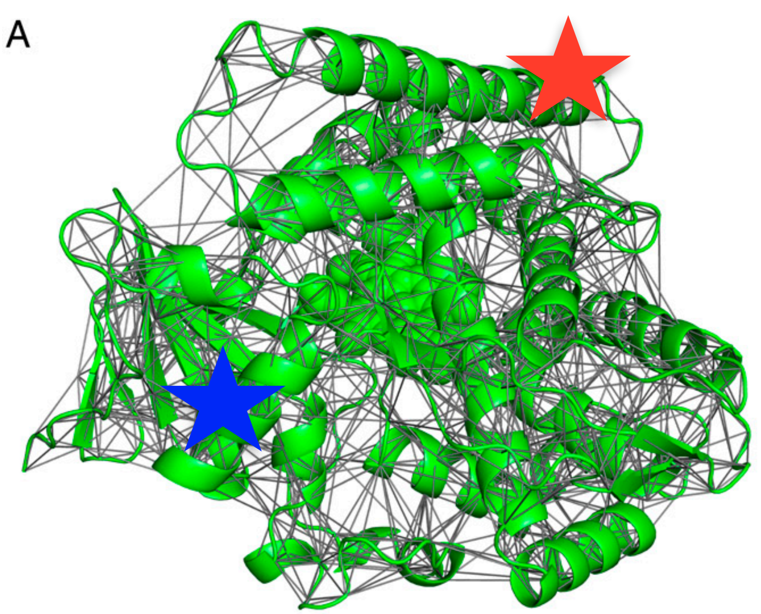
Navigating the protein fitness landscape with Gaussian processes

Philip A. Romero[a], Andreas Krause[b], and Frances H. Arnold[a,1]

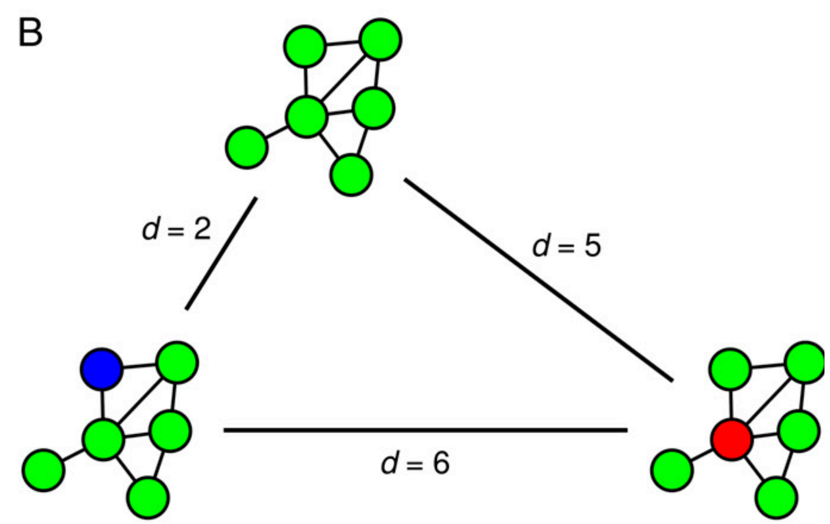
[a]Division of Chemistry and Chemical Engineering, California Institute of Technology, Pasadena, CA 91125; and [b]Department of Computer Science, Swiss Federal Institute of Technology, 8092 Zurich, Switzerland

Edited by Michael Levitt, Stanford University School of Medicine, Stanford, CA, and approved November 28, 2012 (received for review September 9, 2012)

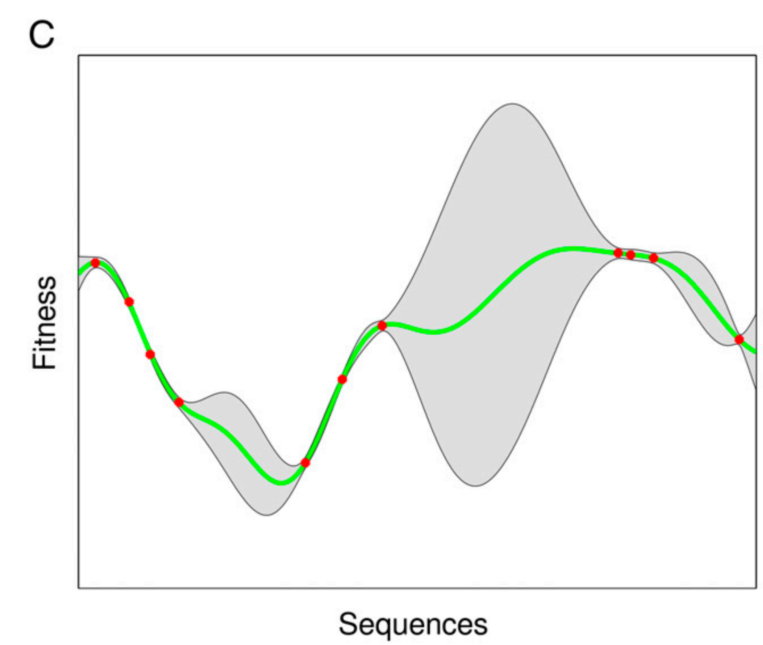
Enzyme to be optimized



Network representation and distance definition



Modeling of measured fitness as GP



FITNESS LANDSCAPES

FLATIRON INSTITUTE
Center for Computational Mathematics

PNAS PLUS

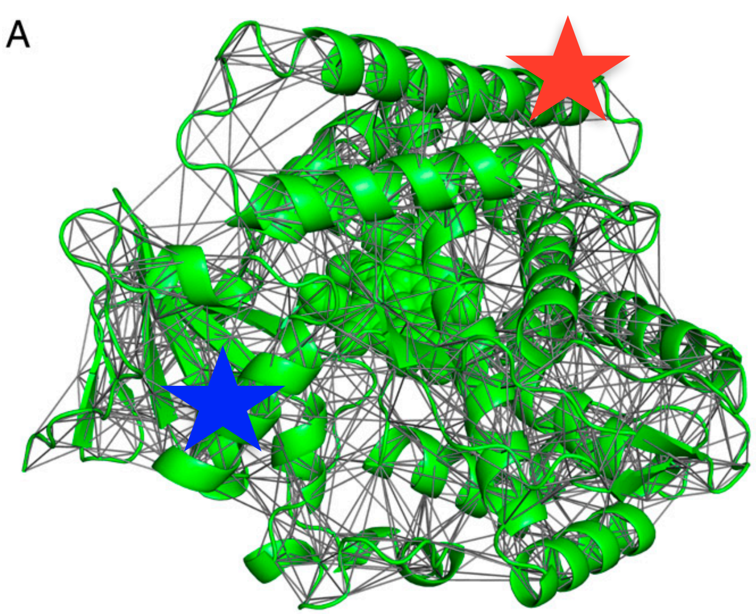
Navigating the protein fitness landscape with Gaussian processes

Philip A. Romero[a], Andreas Krause[b], and Frances H. Arnold[a,1]

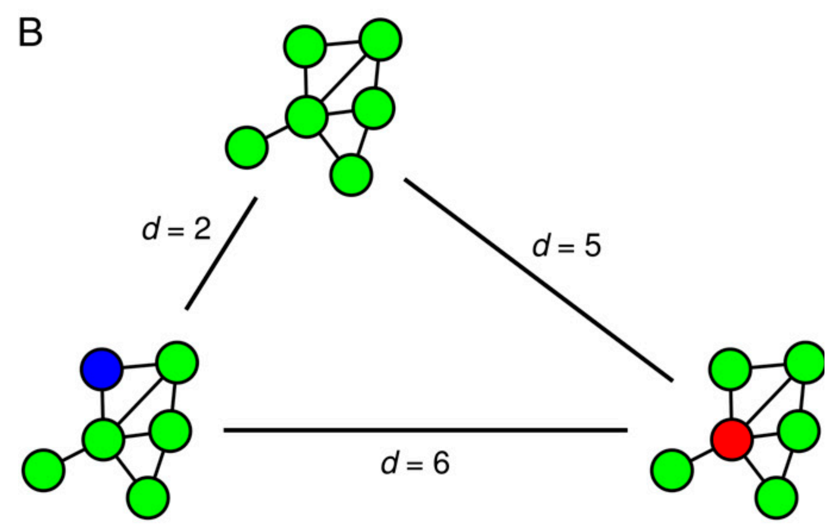
[a]Division of Chemistry and Chemical Engineering, California Institute of Technology, Pasadena, CA 91125; and [b]Department of Computer Science, Swiss Federal Institute of Technology, 8092 Zurich, Switzerland

Edited by Michael Levitt, Stanford University School of Medicine, Stanford, CA, and approved November 28, 2012 (received for review September 9, 2012)

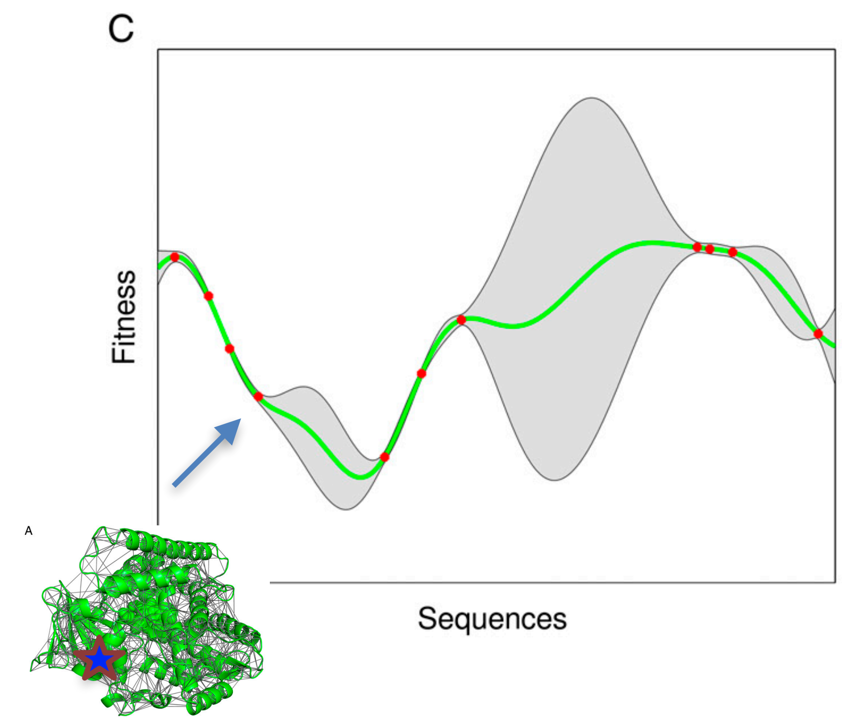
Enzyme to be optimized



Network representation and distance definition



Modeling of measured fitness as GP



FITNESS LANDSCAPES

PNAS PLUS

Navigating the protein fitness landscape with Gaussian processes

Philip A. Romero[a], Andreas Krause[b], and Frances H. Arnold[a,1]

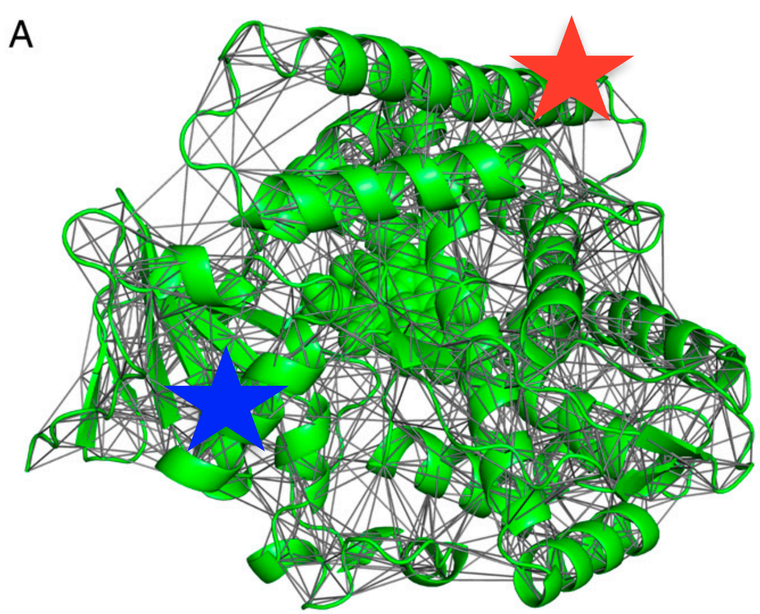
[a]Division of Chemistry and Chemical Engineering, California Institute of Technology, Pasadena, CA 91125; and [b]Department of Computer Science, Swiss Federal Institute of Technology, 8092 Zurich, Switzerland

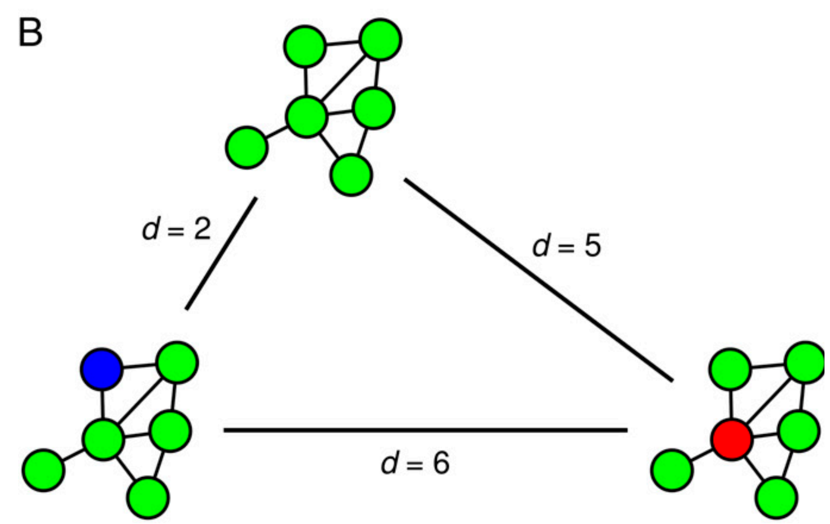
Edited by Michael Levitt, Stanford University School of Medicine, Stanford, CA, and approved November 28, 2012 (received for review September 9, 2012)

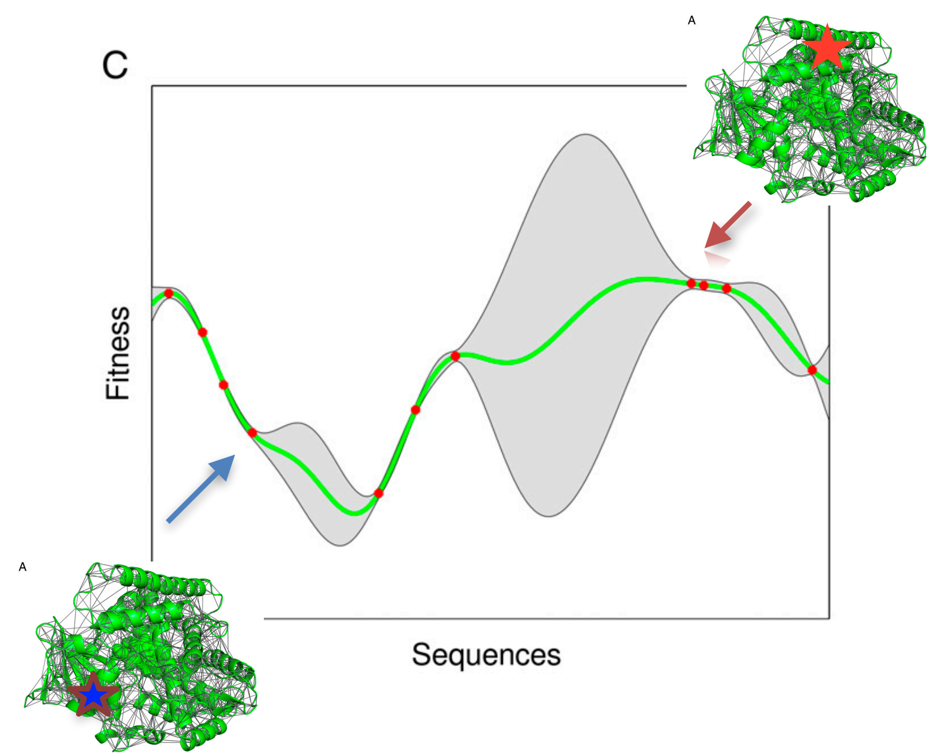
Enzyme to be optimized

Network representation and distance definition

Modeling of measured fitness as GP







FITNESS LANDSCAPES AND OPTIMIZATION

- Evolution can be seen as optimization process over a fitness landscapes.
- The optimization process is based on a **population** of individuals.
- Key operations are **mutation** and **selection**.

FITNESS LANDSCAPES AND OPTIMIZATION

FLATIRON INSTITUTE
Center for Computational Mathematics

- Evolution can be seen as optimization process over a fitness landscapes.
- The optimization process is based on a **population** of individuals.
- Key operations are **mutation** and **selection**.

The entire field of *evolutionary computation*, a subfield of continuous optimization, is based on this idea (>100k publications).

Keywords: Genetic algorithms, genetic programs, Evolution Strategies

ENERGY LANDSCAPES -

LONDON-EYRING-POLANYI POTENTIAL ENERGY SURFACE

Eyring, H, Polanyi, M., “Über einfache Gasreaktionen,”

Zeitschrift für Physikalische Chemie B, Band 12, S. 279–311,1931

ENERGY LANDSCAPES -

LONDON-EYRING-POLANYI POTENTIAL ENERGY SURFACE

Flatiron Institute
Center for Computational Mathematics

Eyring, H, Polanyi, M., “Über einfache Gasreaktionen,”

Zeitschrift für Physikalische Chemie B, Band 12, S. 279–311,1931

$H + H_2 \Leftrightarrow H_2 + H$ reaction for a collinear collision geometry

ENERGY LANDSCAPES -

LONDON-EYRING-POLANYI POTENTIAL ENERGY SURFACE

Eyring, H, Polanyi, M., “Über einfache Gasreaktionen,”

Zeitschrift für Physikalische Chemie B, Band 12, S. 279–311,1931

H + H2 ⇔ H2 + H reaction for a collinear collision geometry

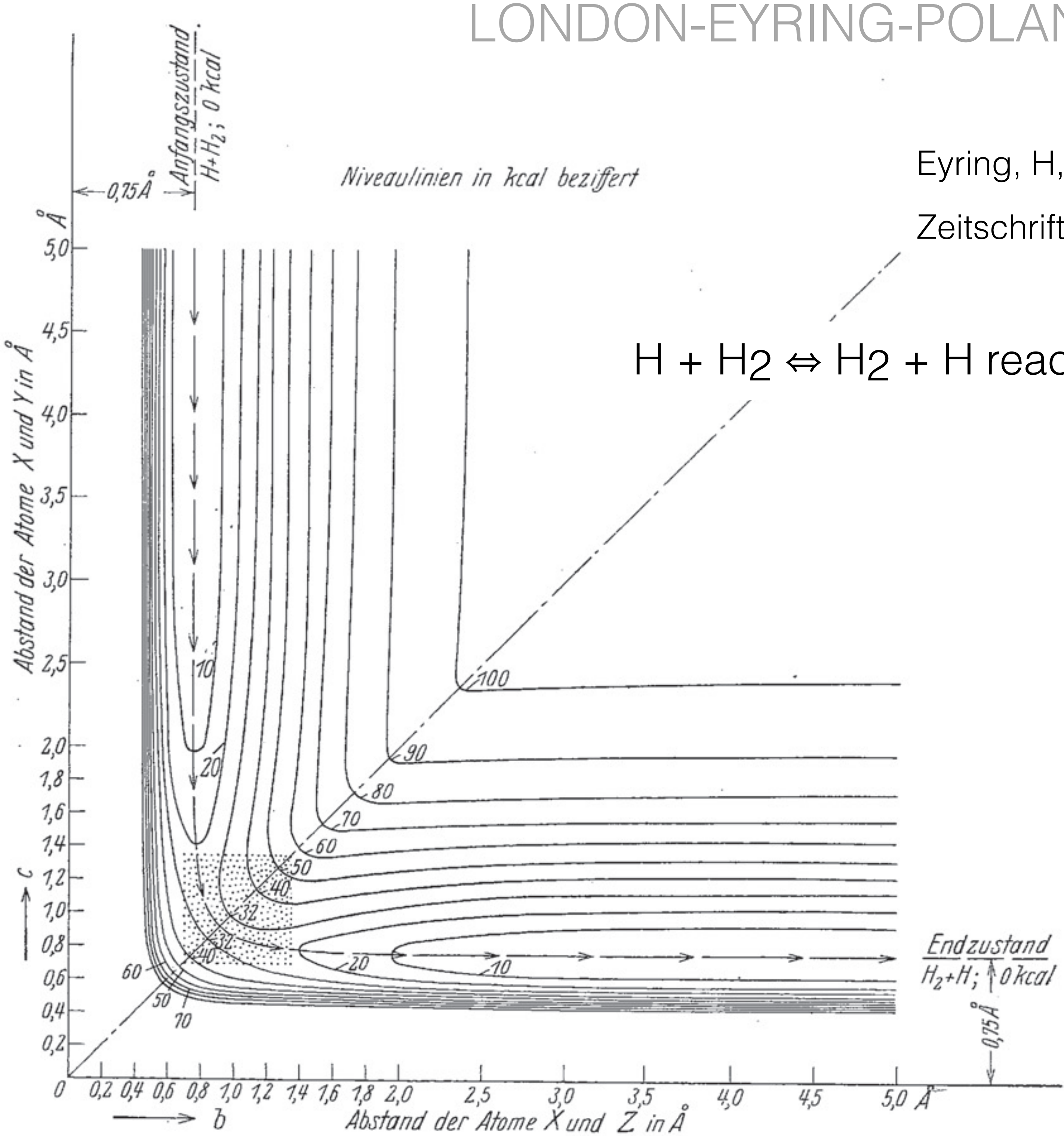


Fig. 5. Resonanzenergie von 3 geradlinig angeordneten H-Atomen als Funktion der Abstände („Resonanzgebirge“).
aus der optischen Energiekurve von H_2 (Fig. 4) unter Vernachlässigung des Coulombschen Anteils berechnet.

ENERGY LANDSCAPES -

LONDON-EYRING-POLANYI POTENTIAL ENERGY SURFACE

Eyring, H, Polanyi, M., “Über einfache Gasreaktionen,”

Zeitschrift für Physikalische Chemie B, Band 12, S. 279–311,1931

H + H2 ⇔ H2 + H reaction for a collinear collision geometry

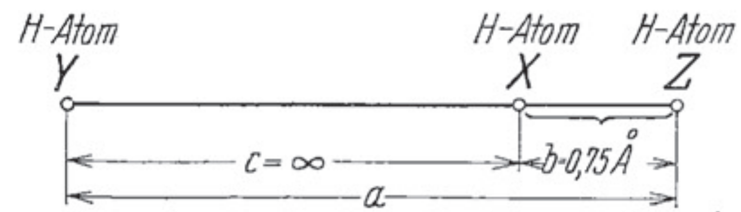


Fig. 5a. Ausgangszustand der in Fig. 5 dargestellten Umsetzung $H + H_2 \rightarrow H_2 + H$.

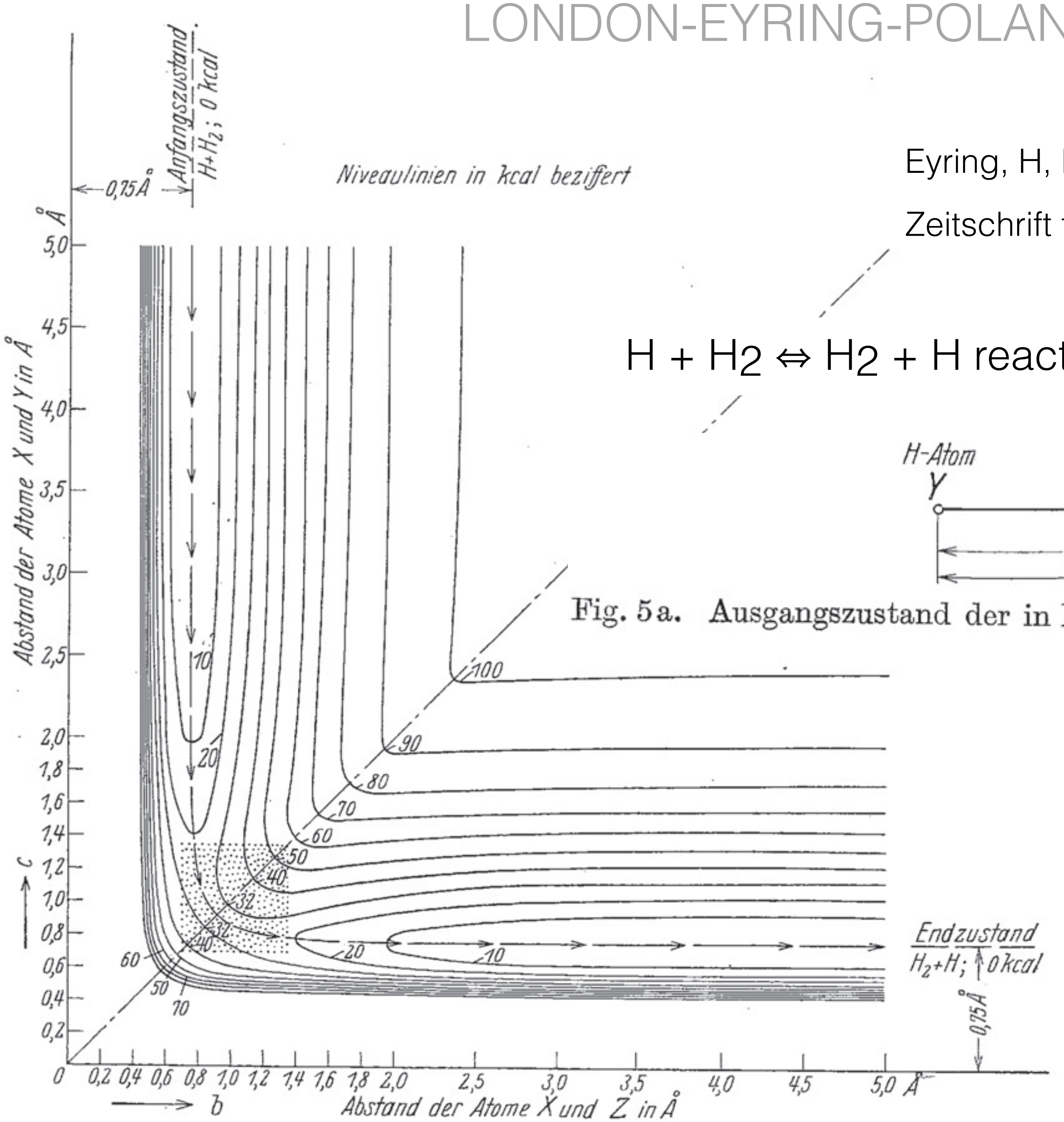


Fig. 5. Resonanzenergie von 3 geradlinig angeordneten H-Atomen als Funktion der Abstände („Resonanzgebirge“).
aus der optischen Energiekurve von H_2 (Fig. 4) unter Vernachlässigung des Coulombschen Anteils berechnet.

ENERGY LANDSCAPES -

LONDON-EYRING-POLANYI POTENTIAL ENERGY SURFACE

FLATIRON INSTITUTE
Center for Computational Mathematics

Eyring, H, Polanyi, M., “Über einfache Gasreaktionen,”

Zeitschrift für Physikalische Chemie B, Band 12, S. 279–311,1931

H + H2 ⇔ H2 + H reaction for a collinear collision geometry

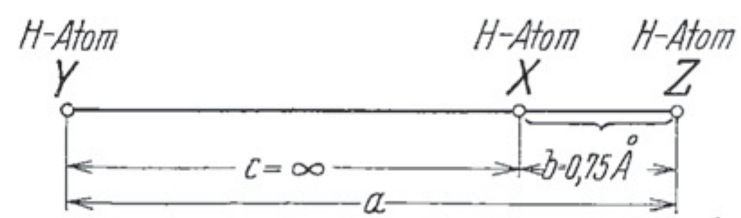


Fig. 5a. Ausgangszustand der in Fig. 5 dargestellten Umsetzung $H + H_2 \rightarrow H_2 + H$.

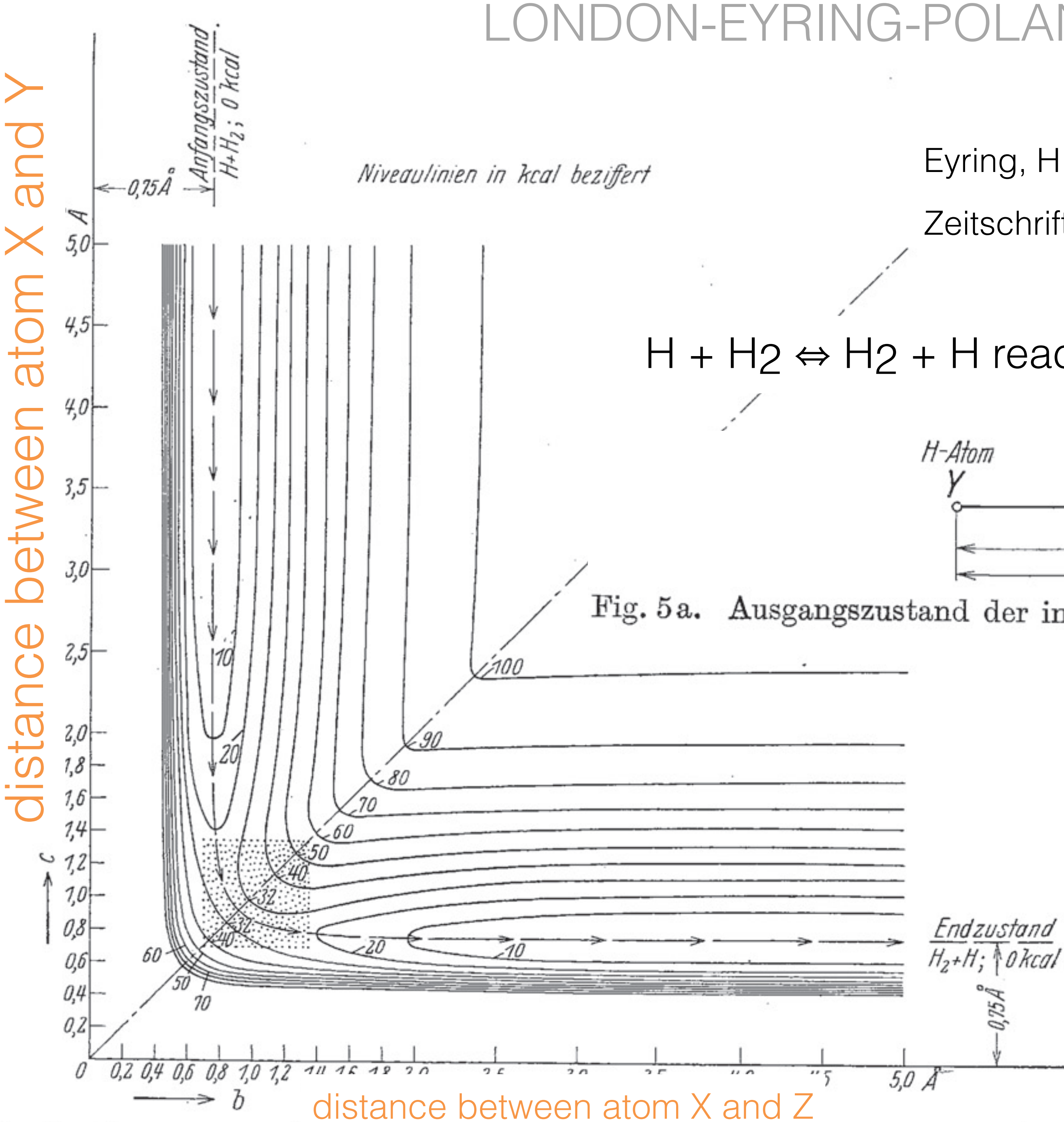


Fig. 5. Resonanzenergie von 3 geradlinig angeordneten H-Atomen als Funktion der Abstände („Resonanzgebirge“).
aus der optischen Energiekurve von H_2 (Fig. 4) unter Vernachlässigung des Coulombschen Anteils berechnet.

ENERGY LANDSCAPES -

LONDON-EYRING-POLANYI POTENTIAL ENERGY SURFACE

FLATIRON INSTITUTE
Center for Computational Mathematics

Eyring, H, Polanyi, M., "Über einfache Gasreaktionen,"

Zeitschrift für Physikalische Chemie B, Band 12, S. 279–311,1931

H + H2 ⇔ H2 + H reaction for a collinear collision geometry

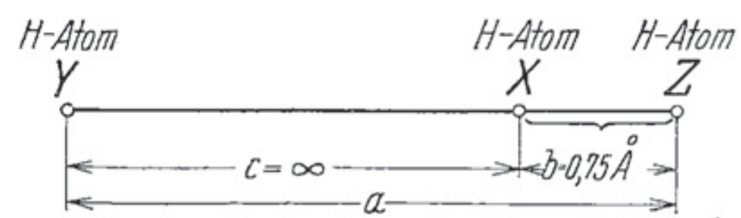


Fig. 5a. Ausgangszustand der in Fig. 5 dargestellten Umsetzung $H + H_2 \rightarrow H_2 + H$.

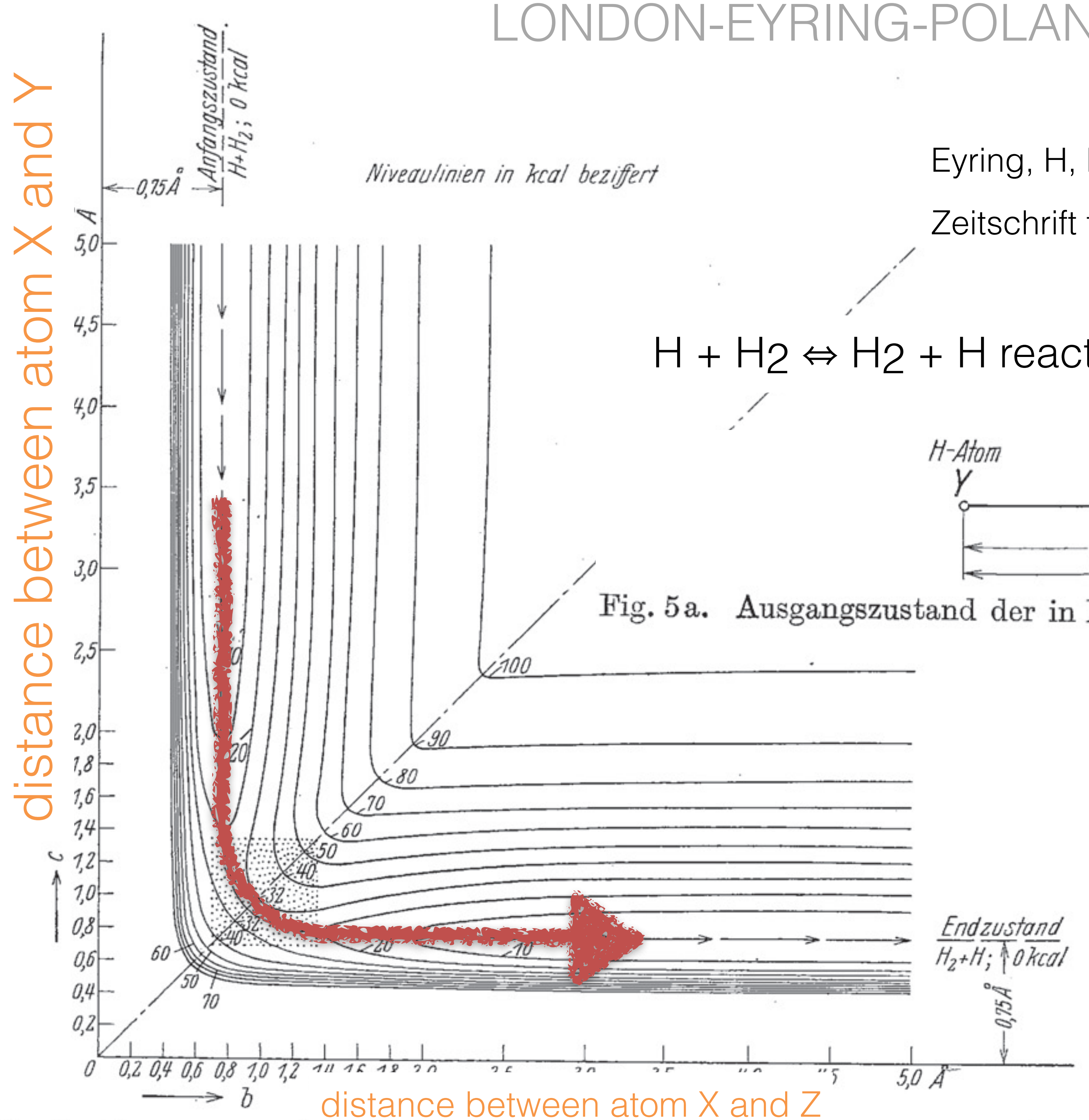


Fig. 5. Resonanzenergie von 3 geradlinig angeordneten H-Atomen als Funktion der Abstände („Resonanzgebirge").
aus der optischen Energiekurve von H_2 (Fig. 4) unter Vernachlässigung des Coulombschen Anteils berechnet.

FLATIRON INSTITUTE
Center for Computational Mathematics

ENERGY LANDSCAPES -

LONDON-EYRING-POLANYI POTENTIAL ENERGY SURFACE

Eyring, H, Polanyi, M., “Über einfache Gasreaktionen,”

Zeitschrift für Physikalische Chemie B, Band 12, S. 279–311,1931

$H + H_2 \Leftrightarrow H_2 + H$ reaction for a collinear collision geometry

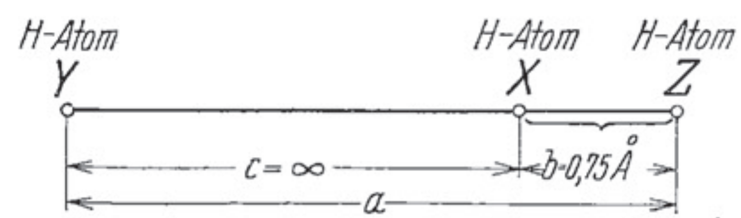


Fig. 5a. Ausgangszustand der in Fig. 5 dargestellten Umsetzung $H + H_2 \rightarrow H_2 + H$.

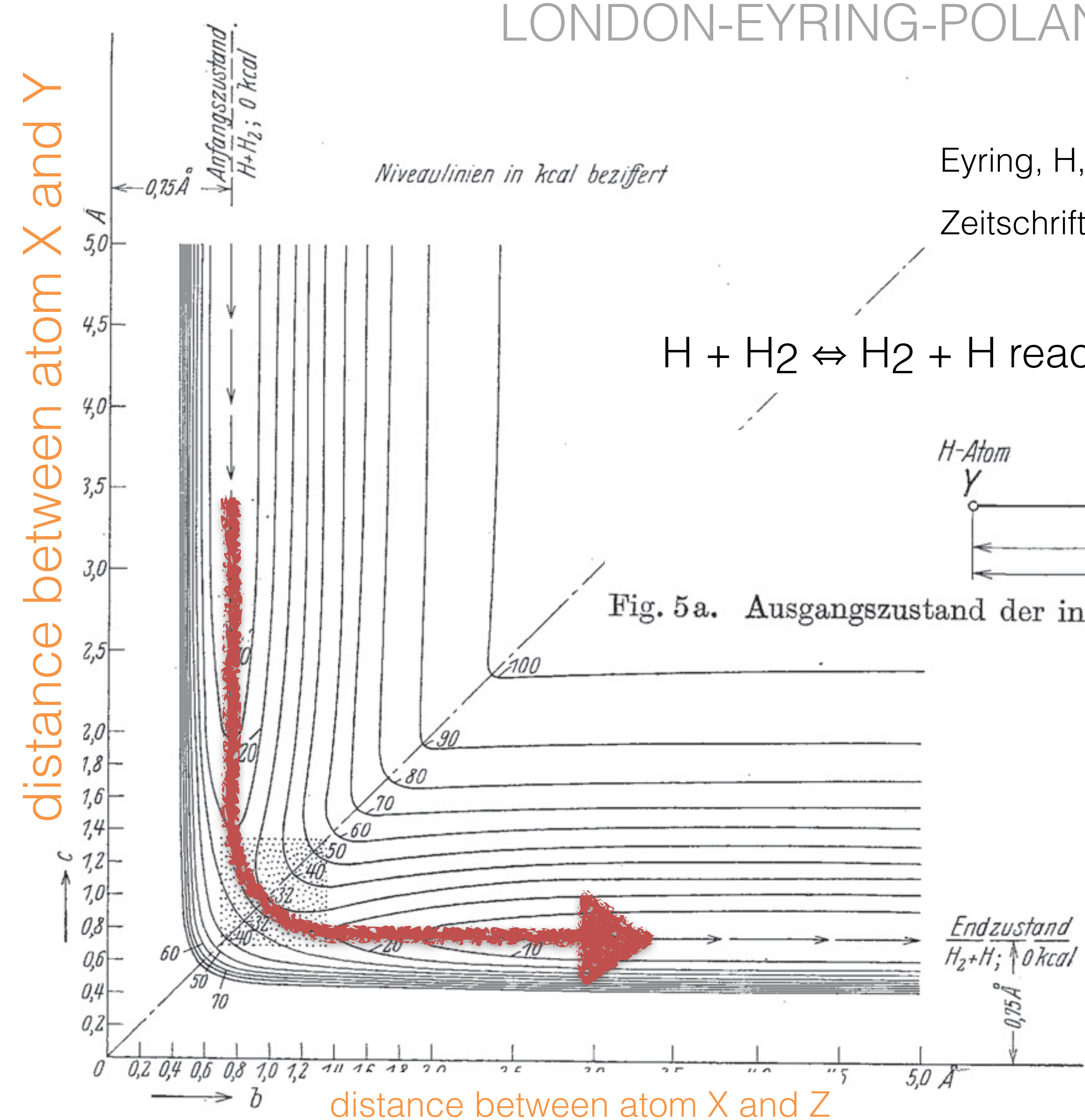


Fig. 5. Resonanzenergie von 3 geradlinig angeordneten H-Atomen als Funktion der Abstände („Resonanzgebirge“).
aus der optischen Energiekurve von H_2 (Fig. 4) unter Vernachlässigung des Coulombschen Anteils berechnet.

ENERGY LANDSCAPES -

LONDON-EYRING-POLANYI POTENTIAL ENERGY SURFACE

Eyring, H, Polanyi, M., “Über einfache Gasreaktionen,”

Zeitschrift für Physikalische Chemie B, Band 12, S. 279–311,1931

$H + H_2 \Leftrightarrow H_2 + H$ reaction for a collinear collision geometry

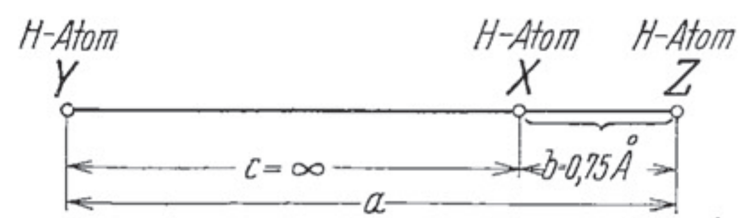


Fig. 5a. Ausgangszustand der in Fig. 5 dargestellten Umsetzung $H + H_2 \rightarrow H_2 + H$.

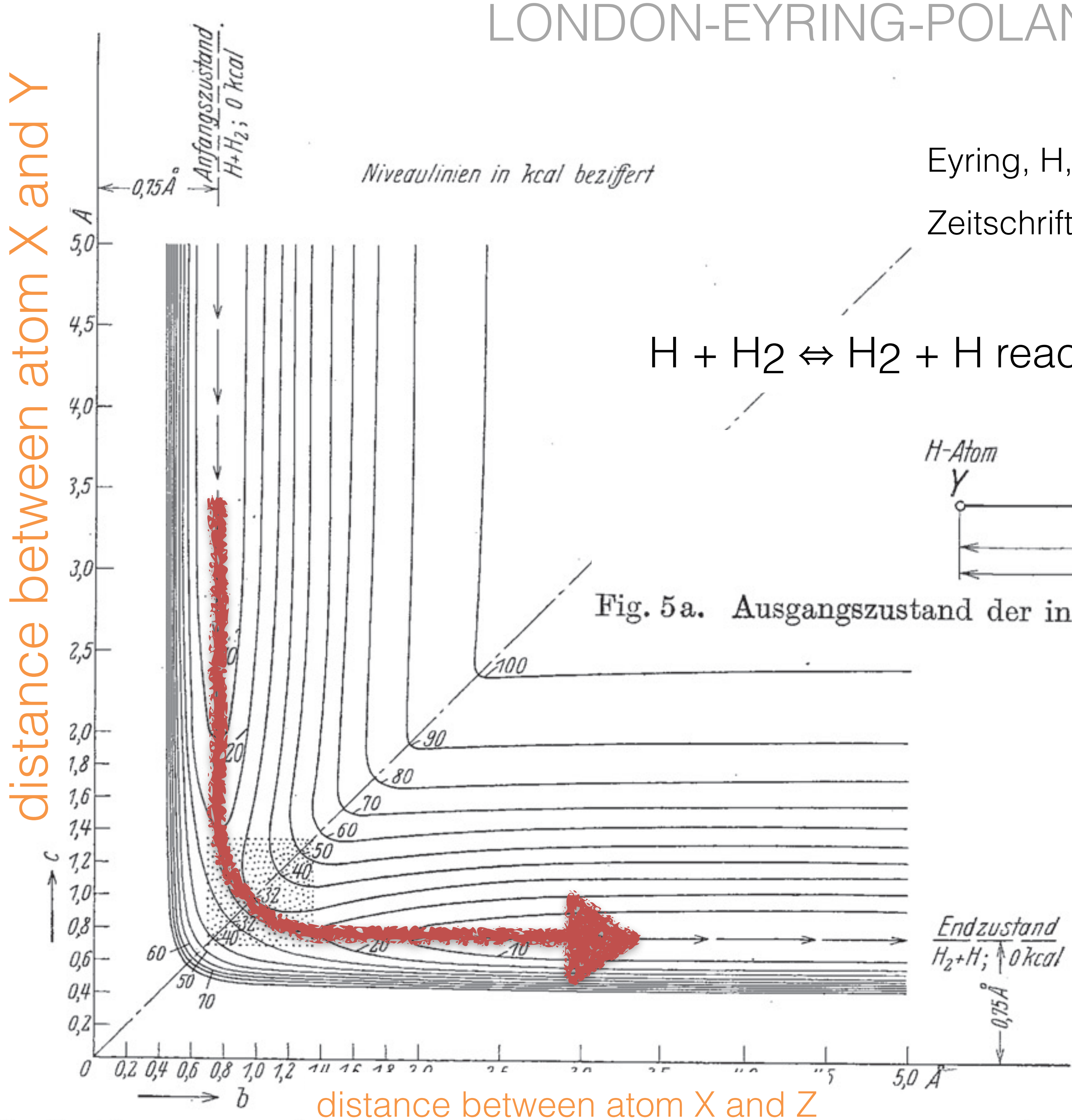


Fig. 5. Resonanzenergie von 3 geradlinig angeordneten H-Atomen als Funktion der Abstände („Resonanzgebirge“). aus der optischen Energiekurve von H_2 (Fig. 4) unter Vernachlässigung des Coulombschen Anteils berechnet.

Resonance energy as a function of distances (“resonance mountain”)

ENERGY LANDSCAPES

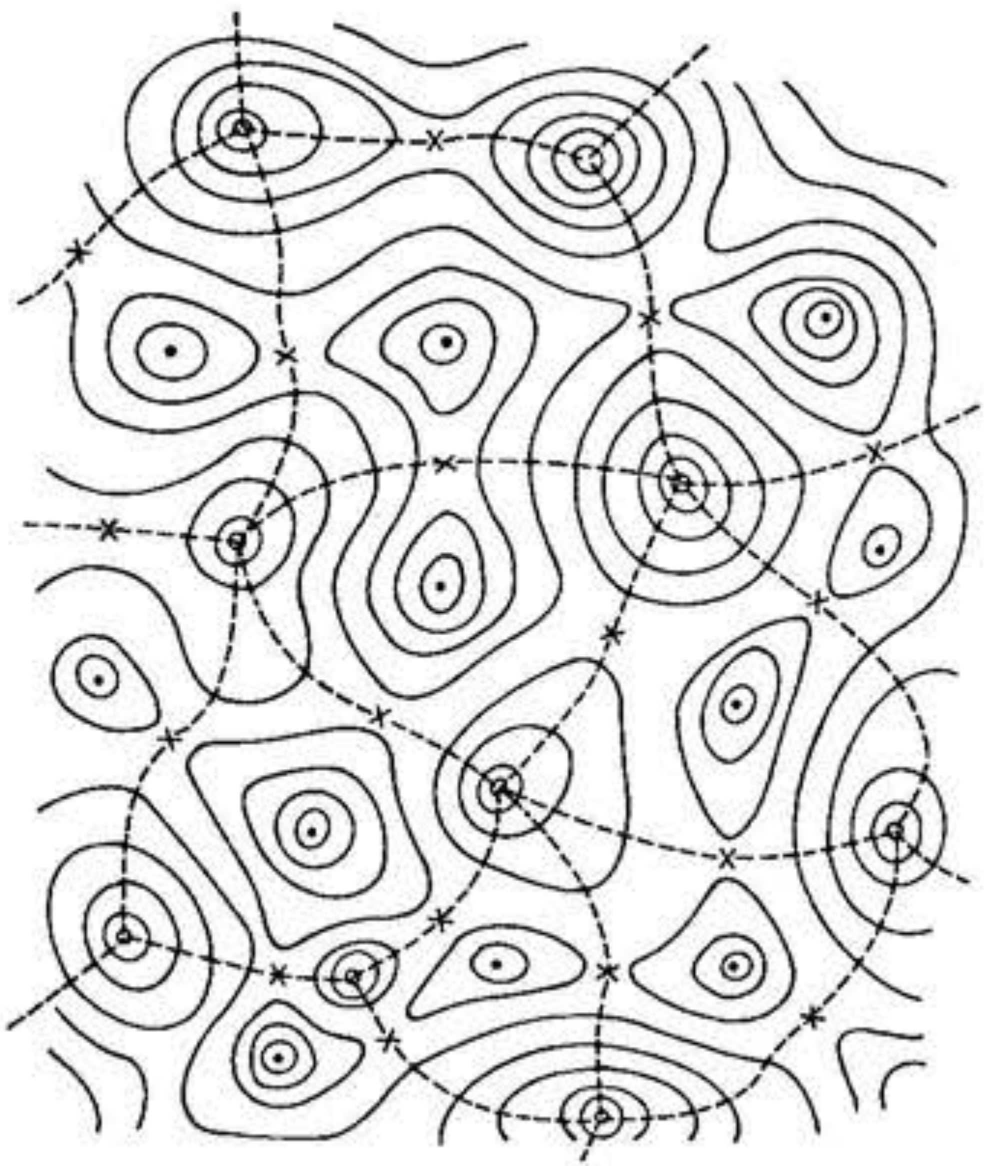


Fig. 1. Schematic representation of the potential energy surface for an *N*-atom system. Minima are shown as filled circles and saddle points as crosses. Potential energy is constant along the continuous curves. Regions belonging to different minima are indicated by dashed curves.

7 September 1984, Volume 225, Number 4666 **SCIENCE**

Packing Structures and Transitions in Liquids and Solids

Frank H. Stillinger and Thomas A. Weber

ENERGY LANDSCAPES

FLATIRON INSTITUTE
Center for Computational Mathematics

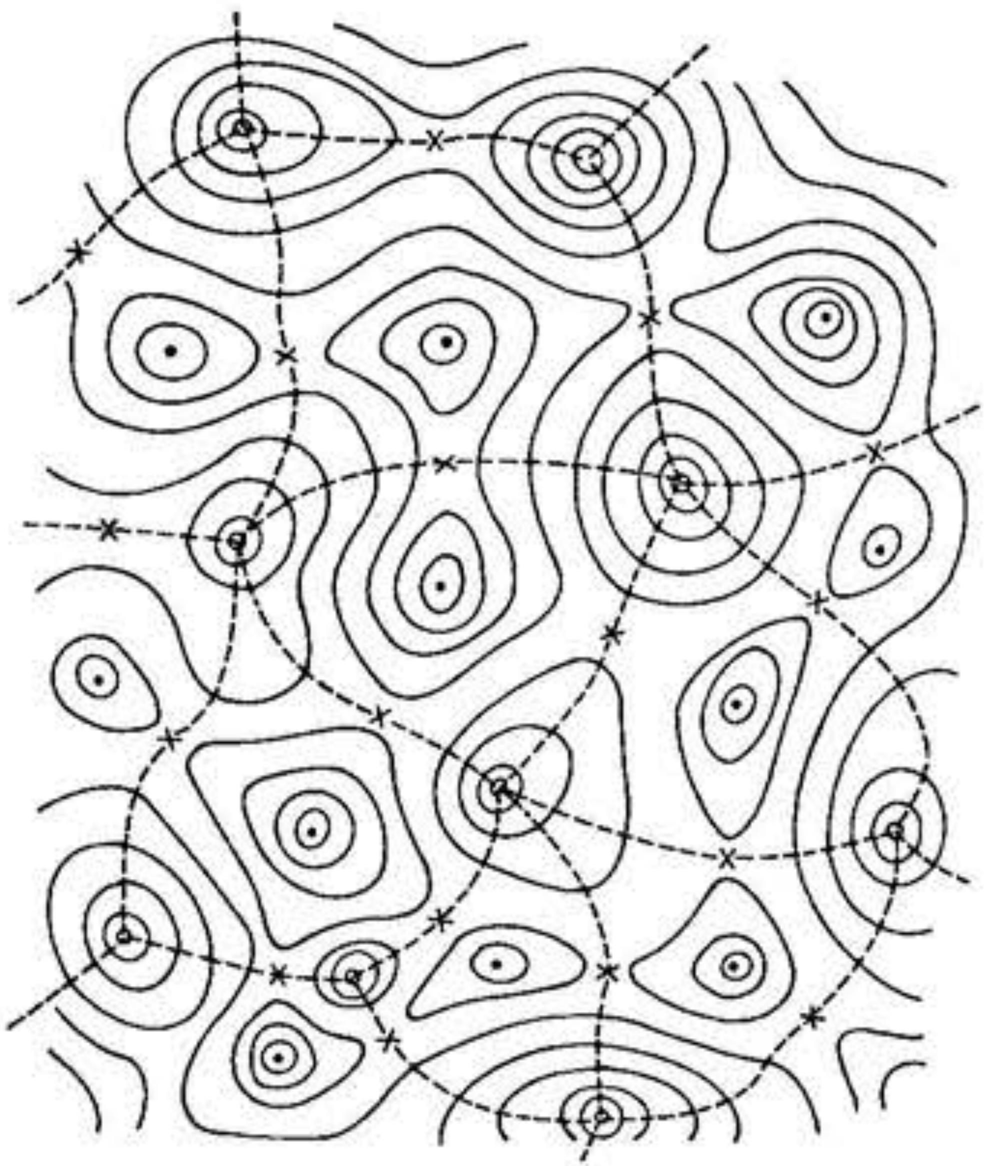


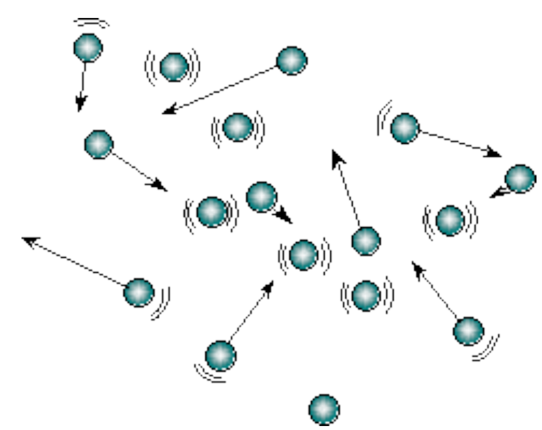
Fig. 1. Schematic representation of the potential energy surface for an *N*-atom system. Minima are shown as filled circles and saddle points as crosses. Potential energy is constant along the continuous curves. Regions belonging to different minima are indicated by dashed curves.

7 September 1984, Volume 225, Number 4666 **SCIENCE**

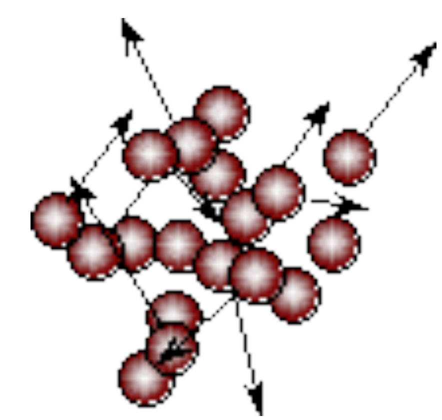
Packing Structures and Transitions in Liquids and Solids

Frank H. Stillinger and Thomas A. Weber

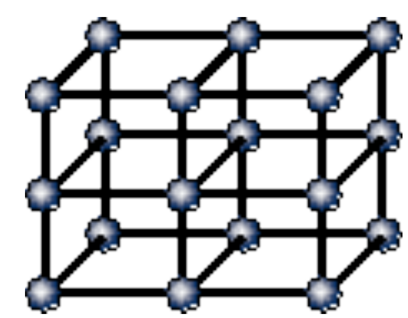
Gas



Liquid



Solid



https://www.learnthermo.com/T1-tutorial/ch03/lesson-A/pg01.php

ENERGY LANDSCAPES AND OPTIMIZATION

The transition process from gas to liquid to solid can be seen as optimization process

13 May 1983, Volume 220, Number 4598

SCIENCE

Optimization by Simulated Annealing

S. Kirkpatrick, C. D. Gelatt, Jr., M. P. Vecchi

ENERGY LANDSCAPES AND OPTIMIZATION

13 May 1983, Volume 220, Number 4598 SCIENCE

The transition process from gas to liquid to solid can be seen as optimization process

Optimization by Simulated Annealing

S. Kirkpatrick, C. D. Gelatt, Jr., M. P. Vecchi

Ingredients:

- A procedure to explore local configurations
- An **temperature**-dependent acceptance criterion for new configurations
- An **temperature** annealing schedule

ENERGY LANDSCAPES AND OPTIMIZATION

13 May 1983, Volume 220, Number 4598 **SCIENCE**

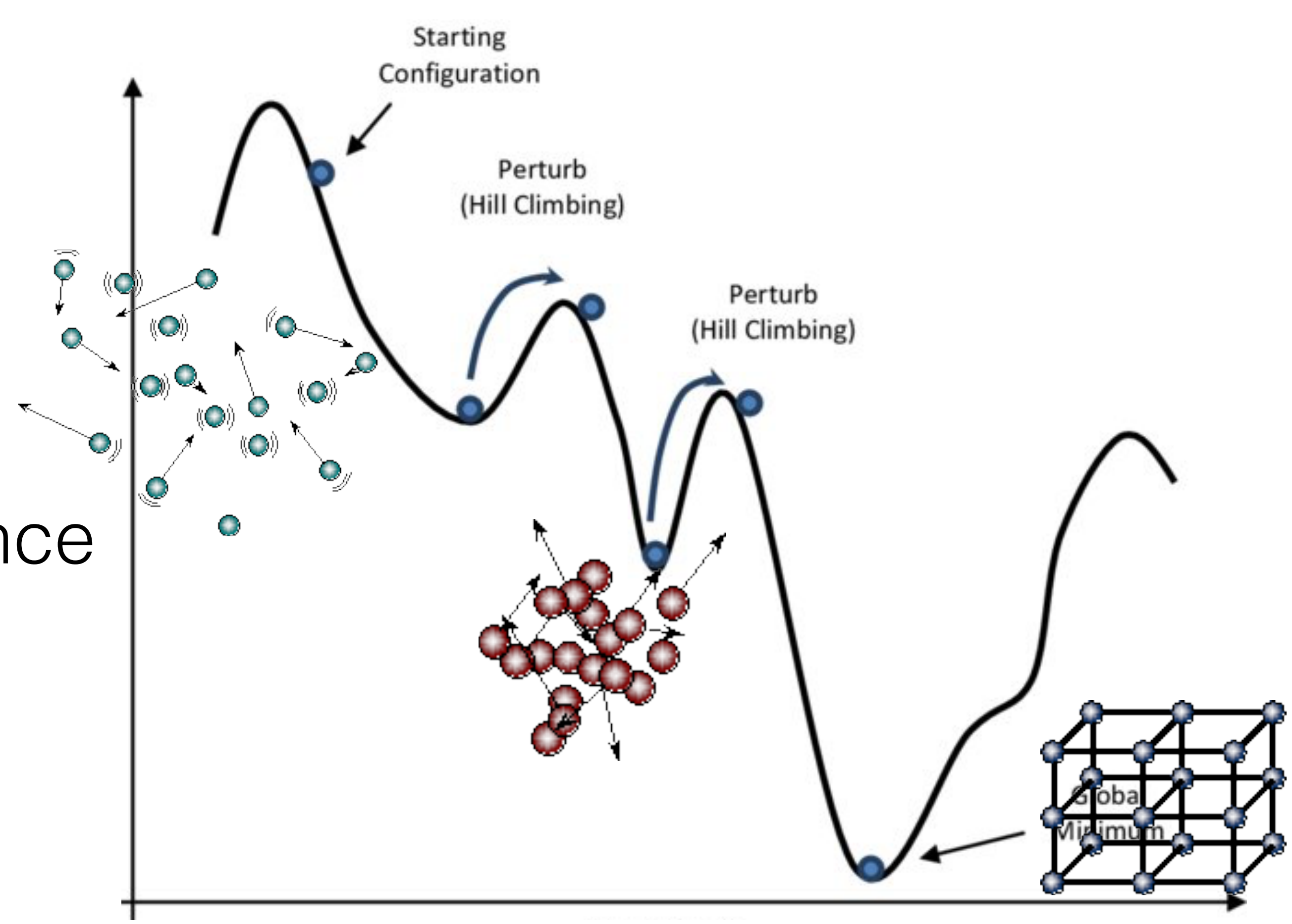
Optimization by Simulated Annealing

S. Kirkpatrick, C. D. Gelatt, Jr., M. P. Vecchi

The transition process from gas to liquid to solid can be seen as optimization process

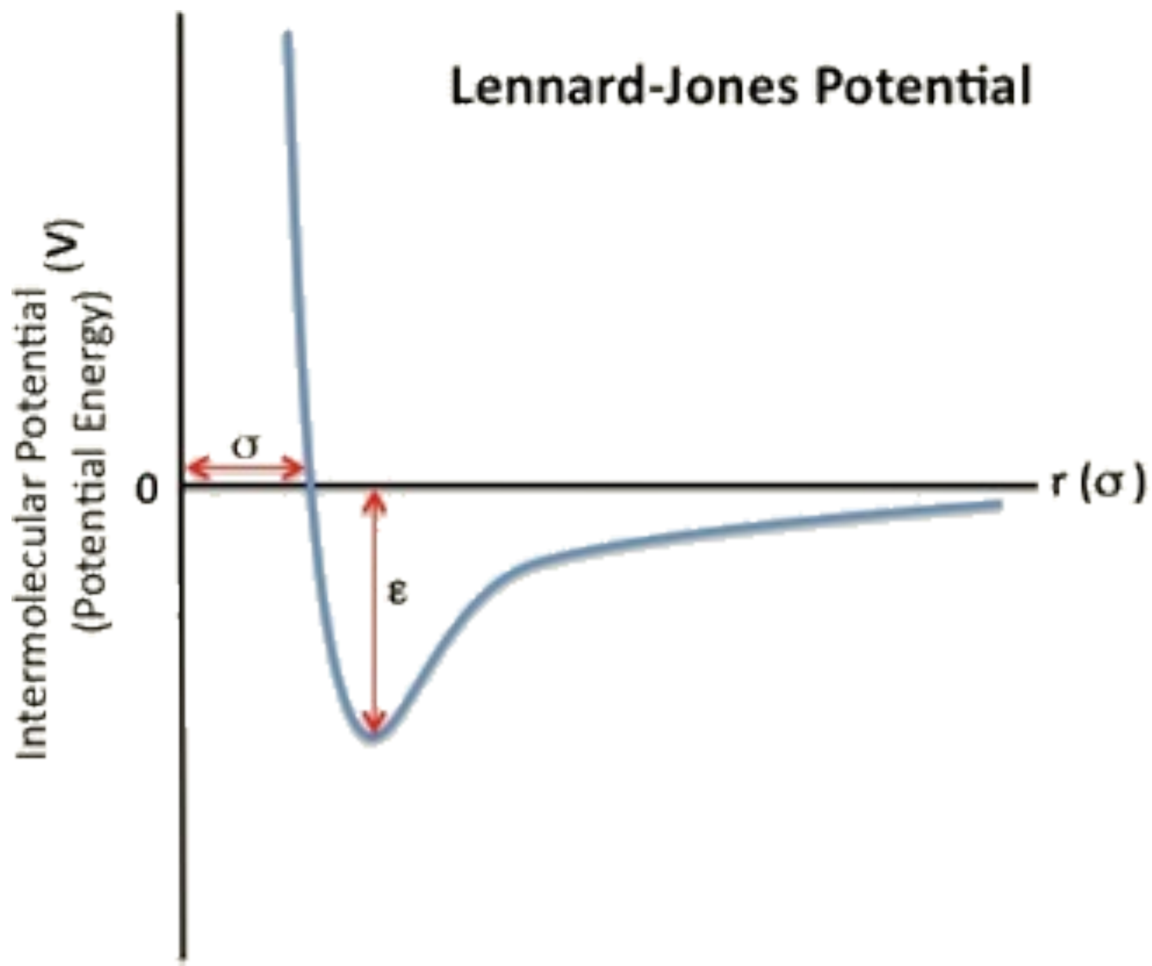
Ingredients:

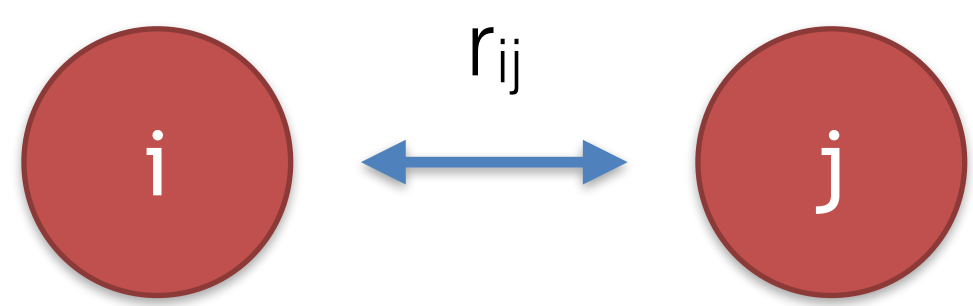
- A procedure to explore local configurations
- An **temperature**-dependent acceptance criterion for new configurations
- An **temperature** annealing schedule



ENERGY LANDSCAPES - LENNARD-JONES CLUSTERS

- Lennard-Jones potential as pair potential between noble gas atoms
- What is the best (lowest potential energy) configuration at temperature T = 0?
- How does the energy landscape look like for N number of atoms?





$$E = 4\epsilon \sum_{i<j} \left[\left(\frac{\sigma}{r_{ij}} \right)^{12} - \left(\frac{\sigma}{r_{ij}} \right)^{6} \right]$$

ENERGY LANDSCAPES - BASIN HOPPING

FLATIRON INSTITUTE
Center for Computational Mathematics

J. Phys. Chem. A **1997,** *101,* 5111−5116 **5111**

Global Optimization by Basin-Hopping and the Lowest Energy Structures of Lennard-Jones Clusters Containing up to 110 Atoms

David J. Wales*
University Chemical Laboratories, Lensfield Road, Cambridge CB2 1EW, U.K.

Jonathan P. K. Doye
FOM Institute for Atomic and Molecular Physics, Kruislaan 407, 1098 SJ Amsterdam, The Netherlands

Received: March 19, 1997; In Final Form: April 29, 1997⊗

ENERGY LANDSCAPES - BASIN HOPPING

FLATIRON INSTITUTE
Center for Computational Mathematics

Ingredients:

- A procedure to explore local configurations as best as possible (e.g., a gradient descent)
- Simulated annealing

J. Phys. Chem. A **1997,** *101,* 5111−5116 **5111**

Global Optimization by Basin-Hopping and the Lowest Energy Structures of Lennard-Jones Clusters Containing up to 110 Atoms

David J. Wales*
University Chemical Laboratories, Lensfield Road, Cambridge CB2 1EW, U.K.

Jonathan P. K. Doye
FOM Institute for Atomic and Molecular Physics, Kruislaan 407, 1098 SJ Amsterdam, The Netherlands

Received: March 19, 1997; In Final Form: April 29, 1997⊗

ENERGY LANDSCAPES - BASIN HOPPING

J. Phys. Chem. A **1997,** *101,* 5111−5116 **5111**

Global Optimization by Basin-Hopping and the Lowest Energy Structures of Lennard-Jones Clusters Containing up to 110 Atoms

David J. Wales*
University Chemical Laboratories, Lensfield Road, Cambridge CB2 1EW, U.K.

Jonathan P. K. Doye
FOM Institute for Atomic and Molecular Physics, Kruislaan 407, 1098 SJ Amsterdam, The Netherlands

Received: March 19, 1997; In Final Form: April 29, 1997[⊗]

Ingredients:

- A procedure to explore local configurations as best as possible (e.g., a gradient descent)
- Simulated annealing

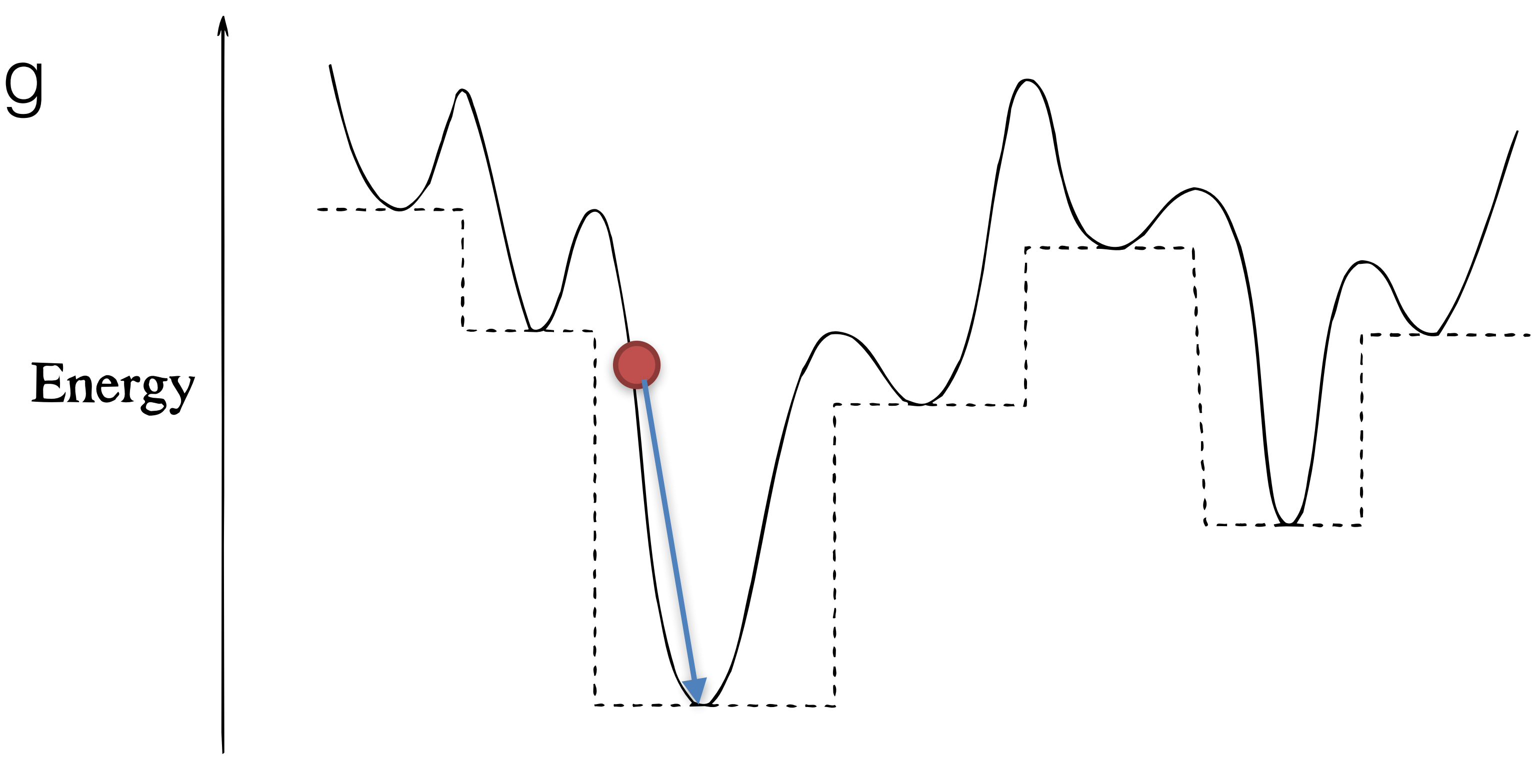
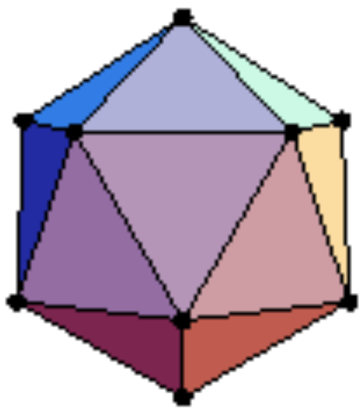


Figure 2. A schematic diagram illustrating the effects of our energy transformation for a one-dimensional example. The solid line is the energy of the original surface and the dashed line is the transformed energy $\tilde{E}$.

ENERGY LANDSCAPES - LJ CLUSTER MINIMA

FLATIRON INSTITUTE
Center for Computational Mathematics

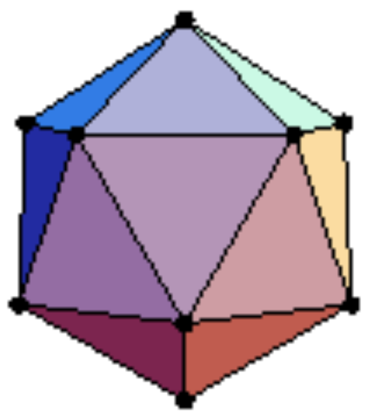
LJ 13



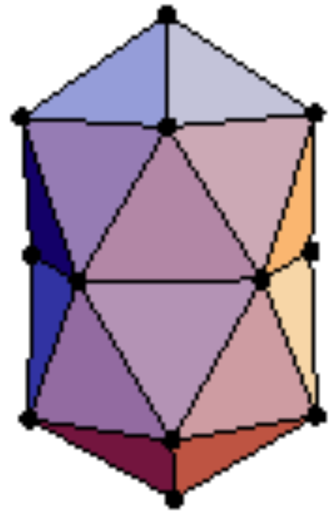
ENERGY LANDSCAPES - LJ CLUSTER MINIMA

FLATIRON INSTITUTE
Center for Computational Mathematics

LJ 13

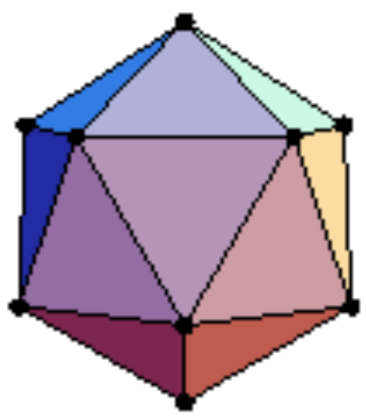


LJ 19

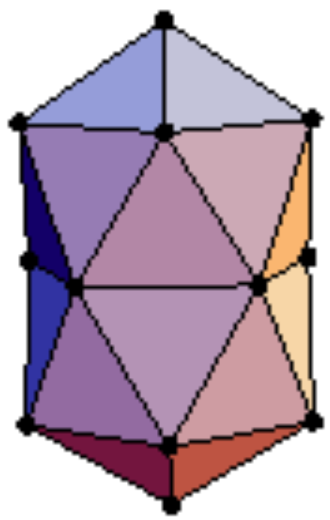


ENERGY LANDSCAPES - LJ CLUSTER MINIMA

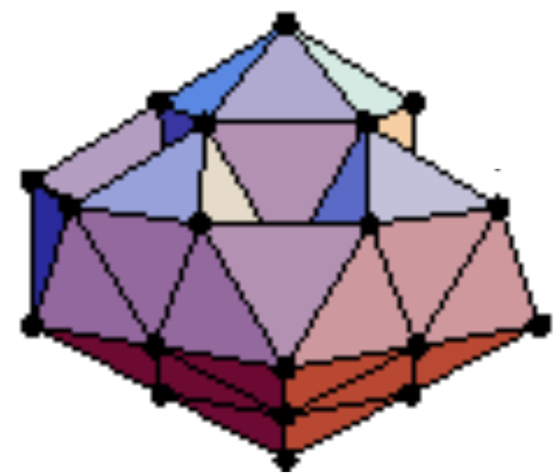
LJ 13



LJ 19

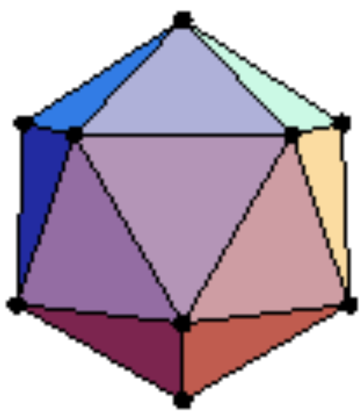


LJ 31

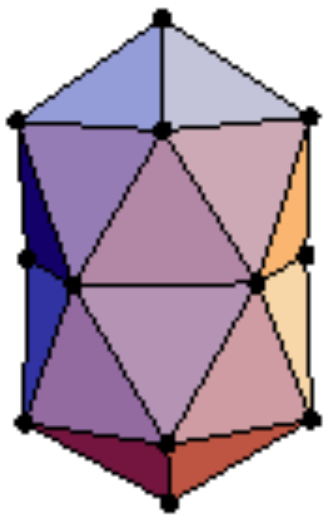


ENERGY LANDSCAPES - LJ CLUSTER MINIMA

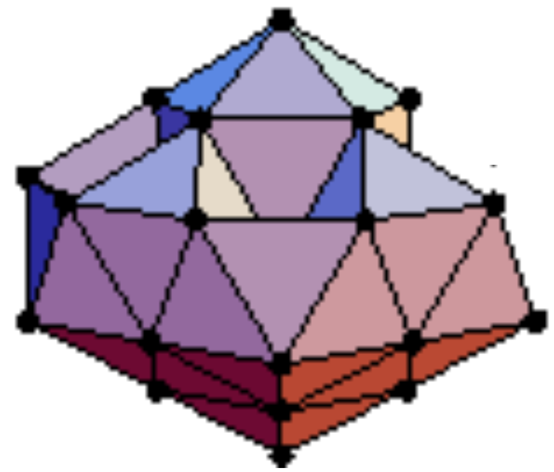
LJ 13



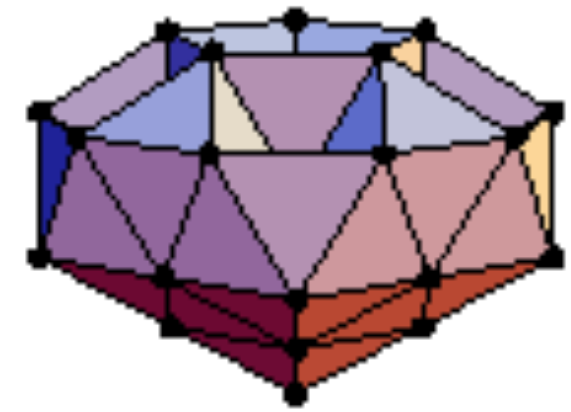
LJ 19



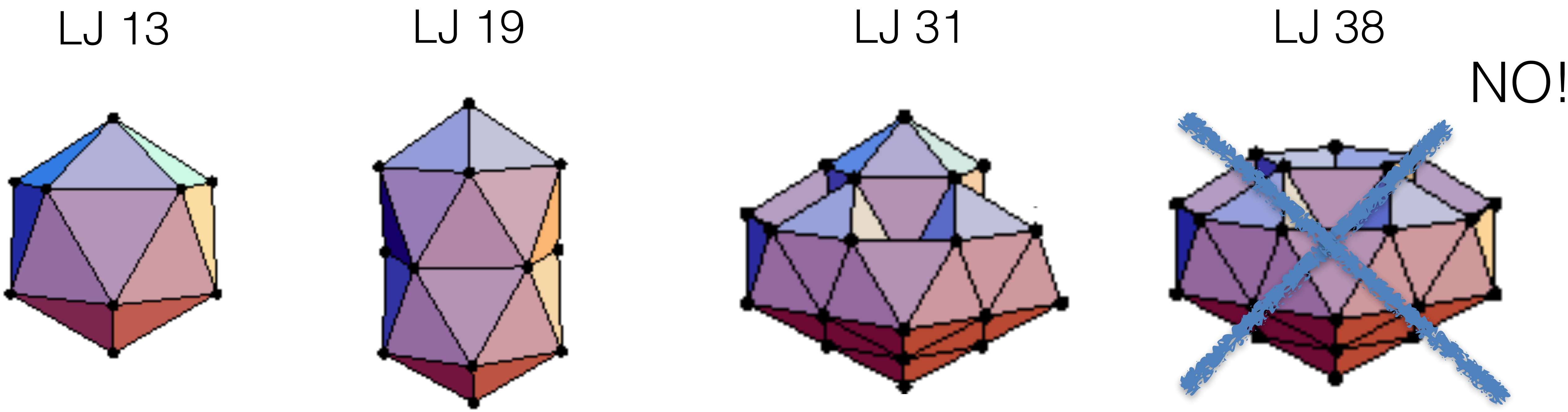
LJ 31



LJ 38

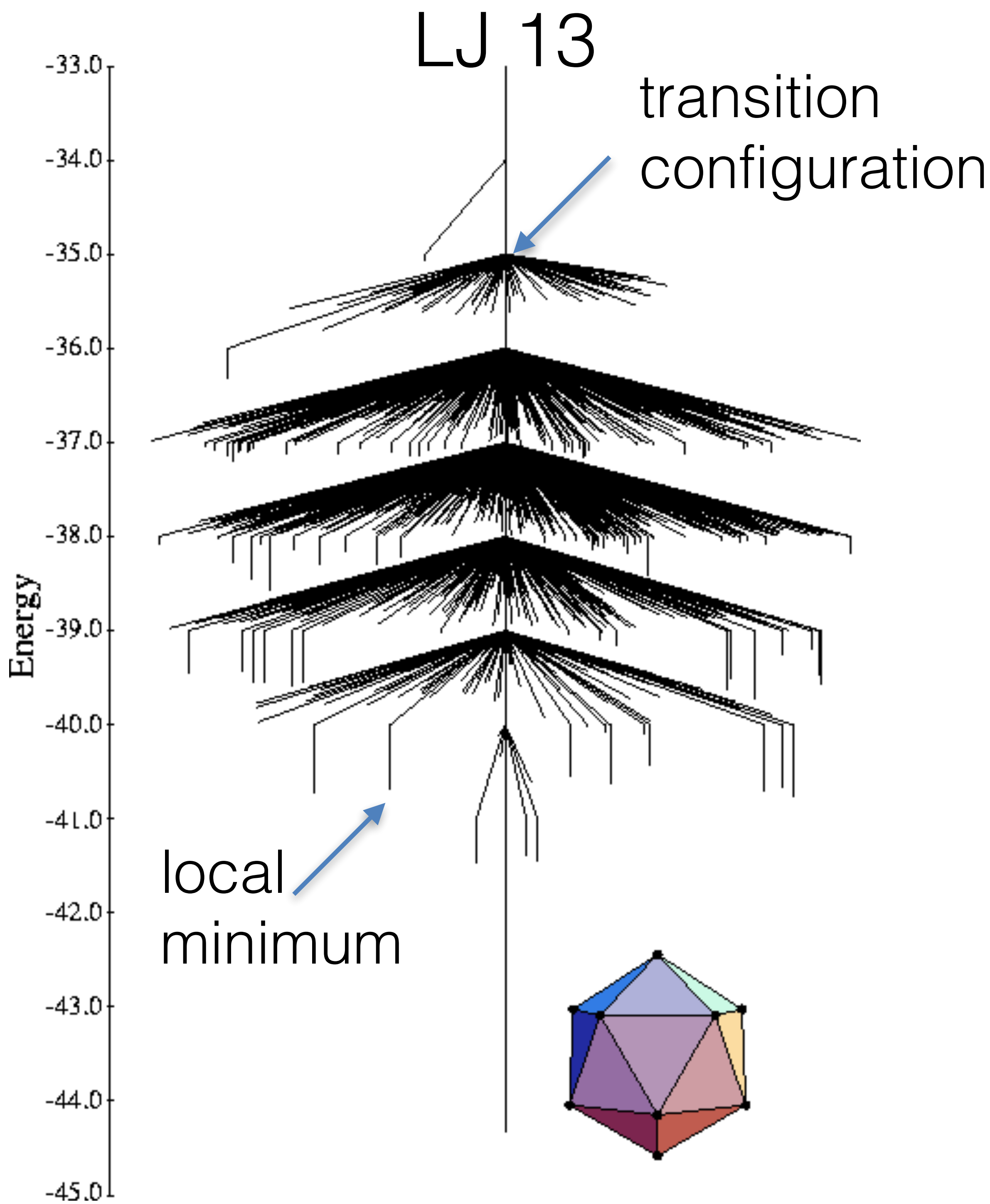


ENERGY LANDSCAPES - LJ CLUSTER MINIMA



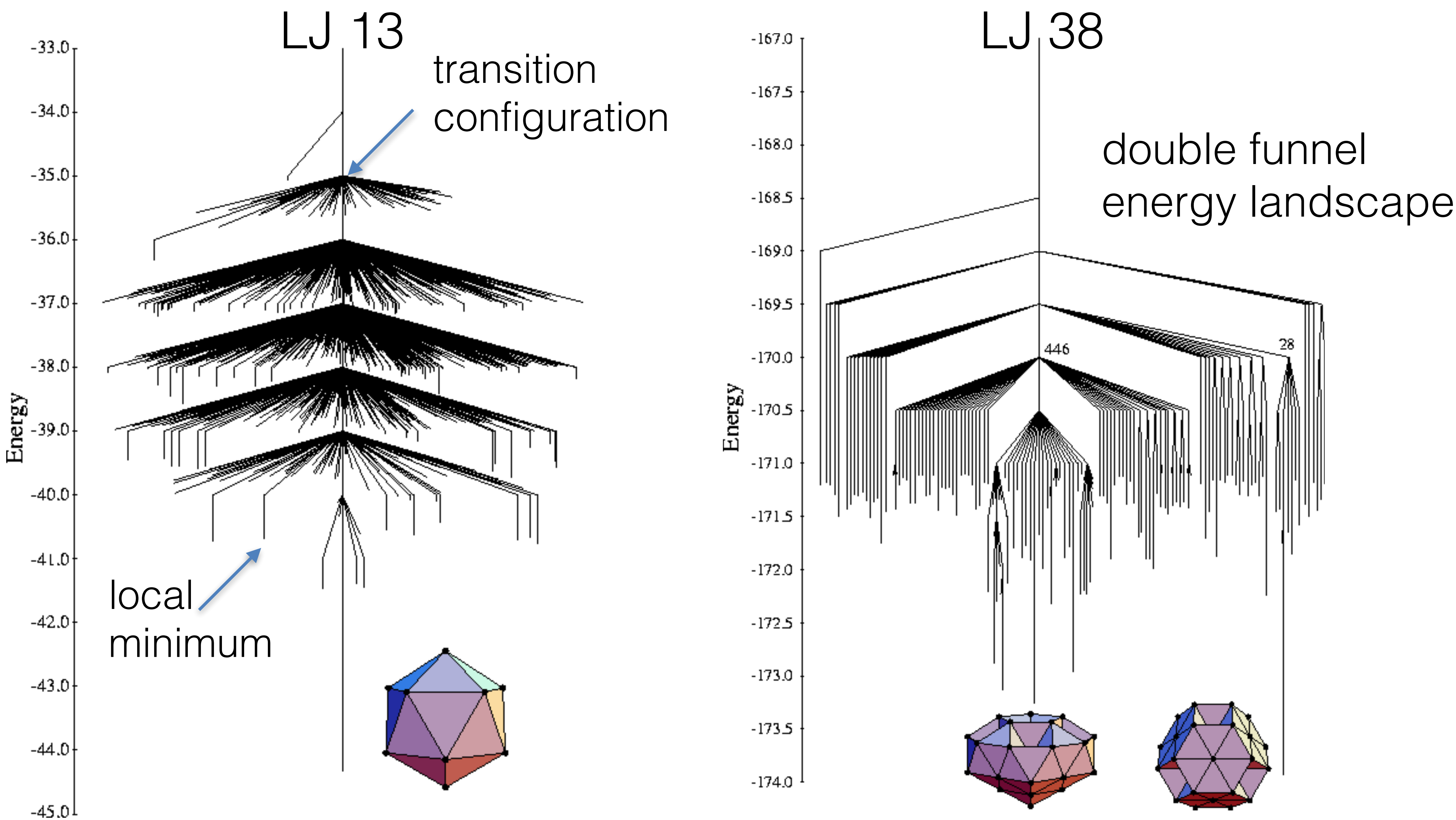
This face-centered cubic octahedron (fcc) structure is the global minimum.

ENERGY LANDSCAPES AND DISCONNECTIVITY GRAPHS



ENERGY LANDSCAPES AND DISCONNECTIVITY GRAPHS

FLATIRON INSTITUTE
Center for Computational Mathematics



ENERGY LANDSCAPES AND THE SIMONS FOUNDATION

FLATIRON INSTITUTE
Center for Computational Mathematics

SIMONS COLLABORATION ON CRACKING THE GLASS PROBLEM

Home | Our Team | Affiliates | Collaborators | Alumni | Tutorials | Events | Jobs | Publications | News

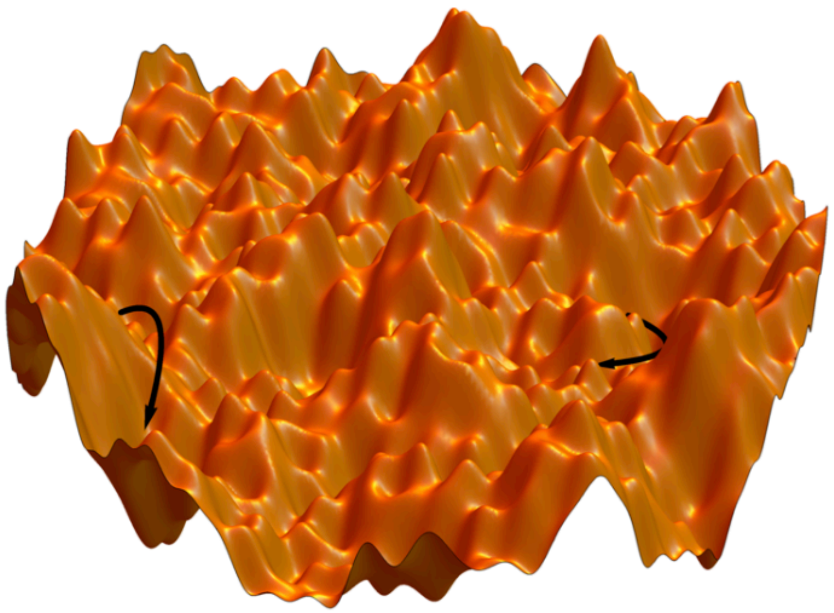
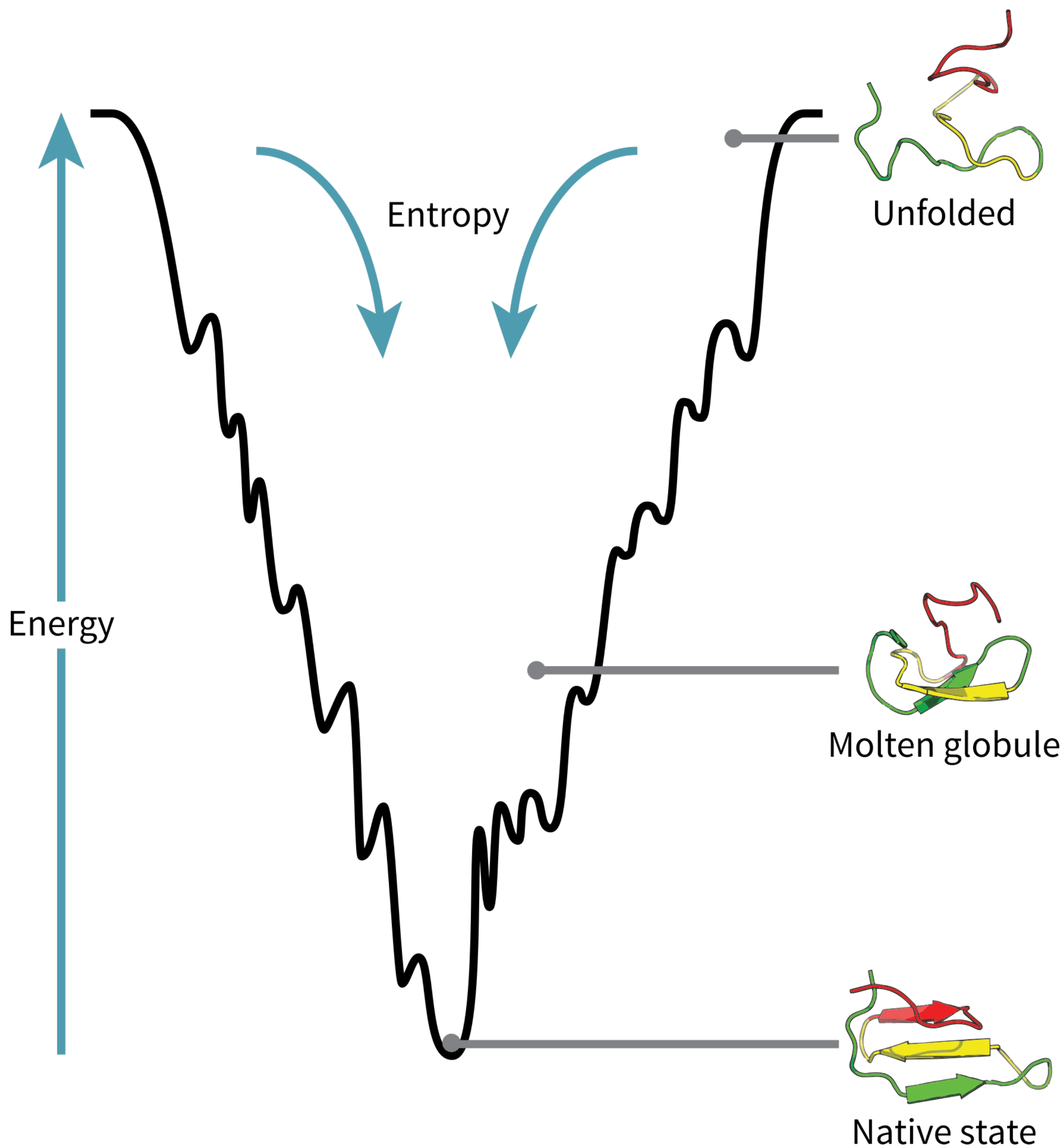


Figure (credit: Chiara Cammarota): A schematic rugged energy landscape with a multitude of energy minima, maxima, and saddles. Arrows denote some of the possible relaxation pathways.

ENERGY LANDSCAPES AND PROTEIN FOLDING



Science 13 Dec 1991:
Vol. 254, Issue 5038, pp. 1598-1603
DOI: 10.1126/science.1749933

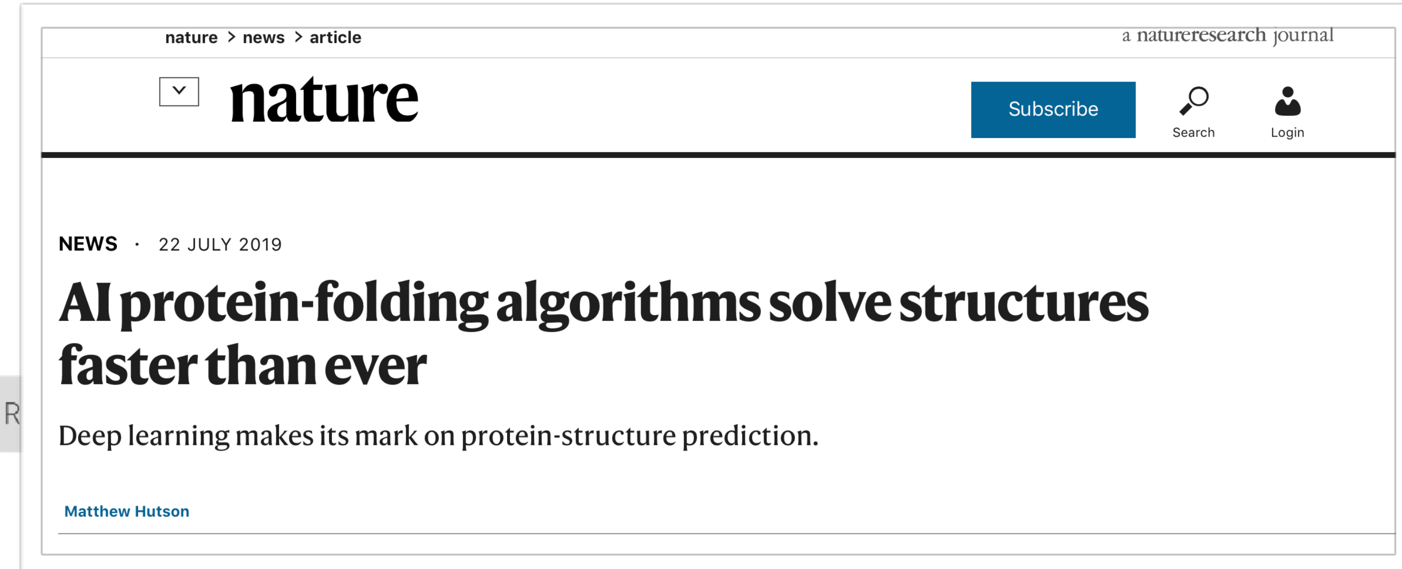
Articles

The Energy Landscapes and Motions of Proteins

HANS FRAUENFELDER, STEPHEN G. SLIGAR, PETER G. WOLYNES

ENERGY LANDSCAPES AND DEEP MIND'S ALPHA-FOLD

FLATIRON INSTITUTE
Center for Computational Mathematics

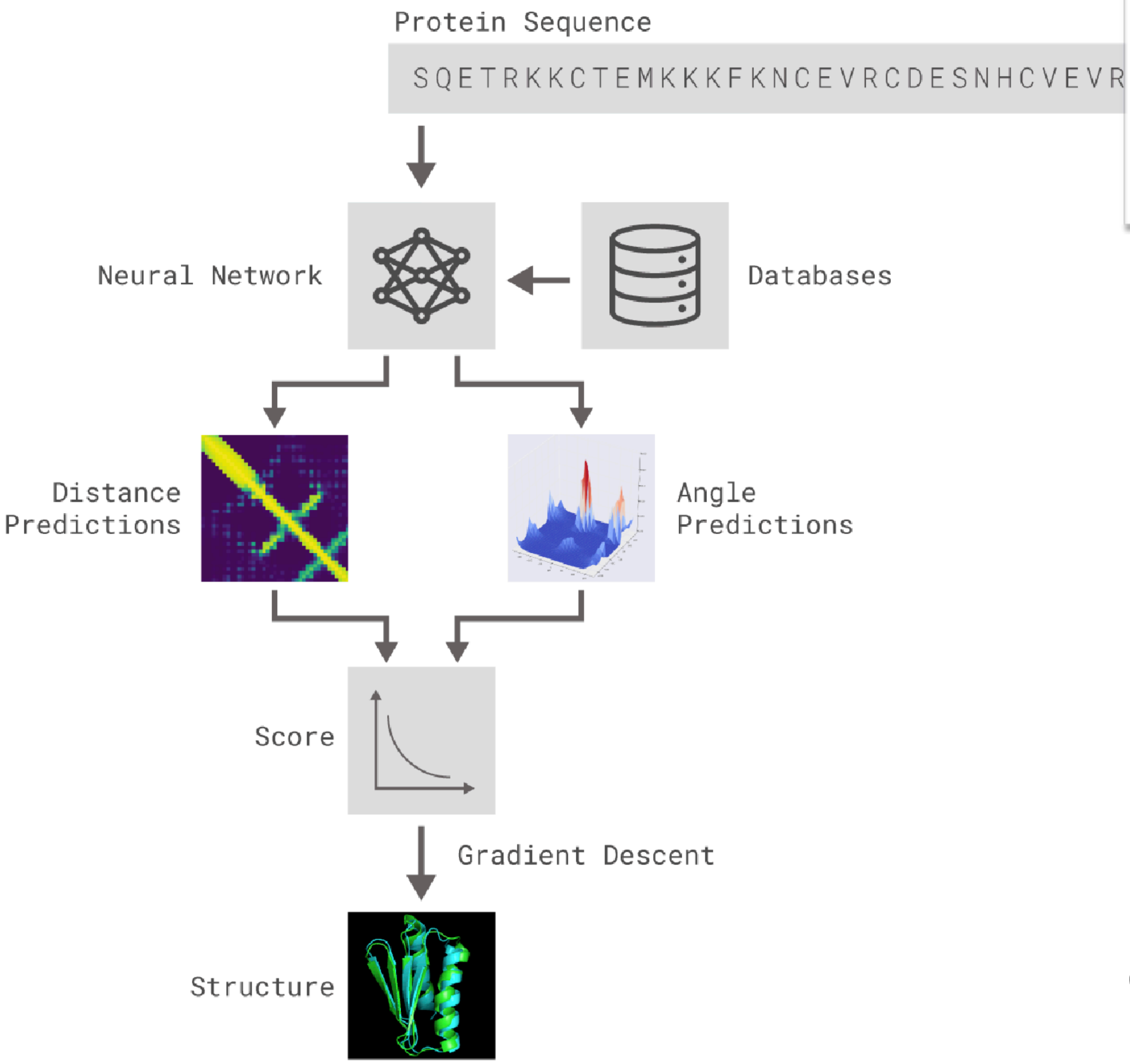

nature > news > article
a natureresearch journal
nature
Subscribe
Search
Login

NEWS · 22 JULY 2019

AI protein-folding algorithms solve structures faster than ever

Deep learning makes its mark on protein-structure prediction.

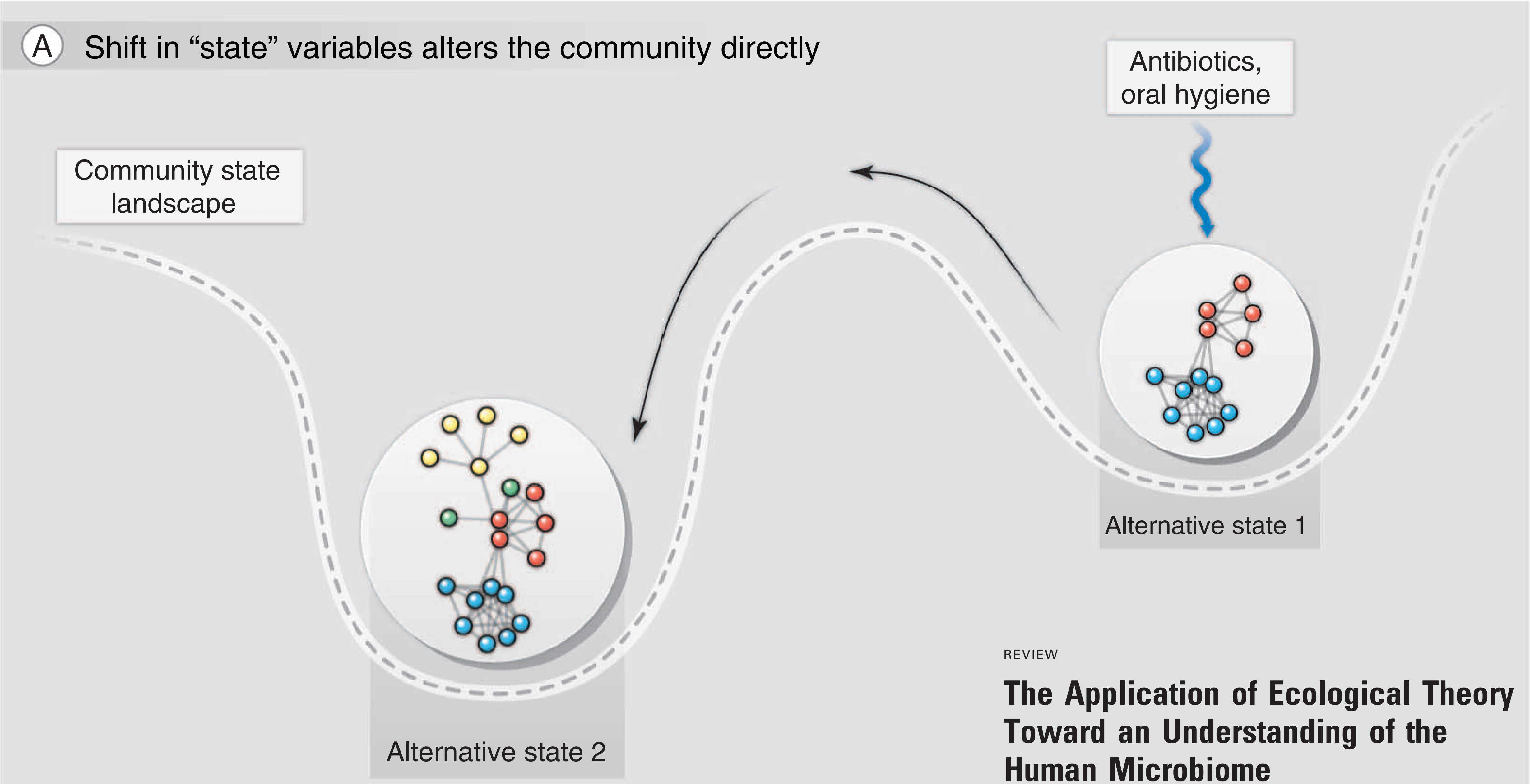
Matthew Hutson



Build a single-funnel energy landscape approximation

Gradient descent on landscape

COMMUNITY STATE LANDSCAPES AND ECOSYSTEMS



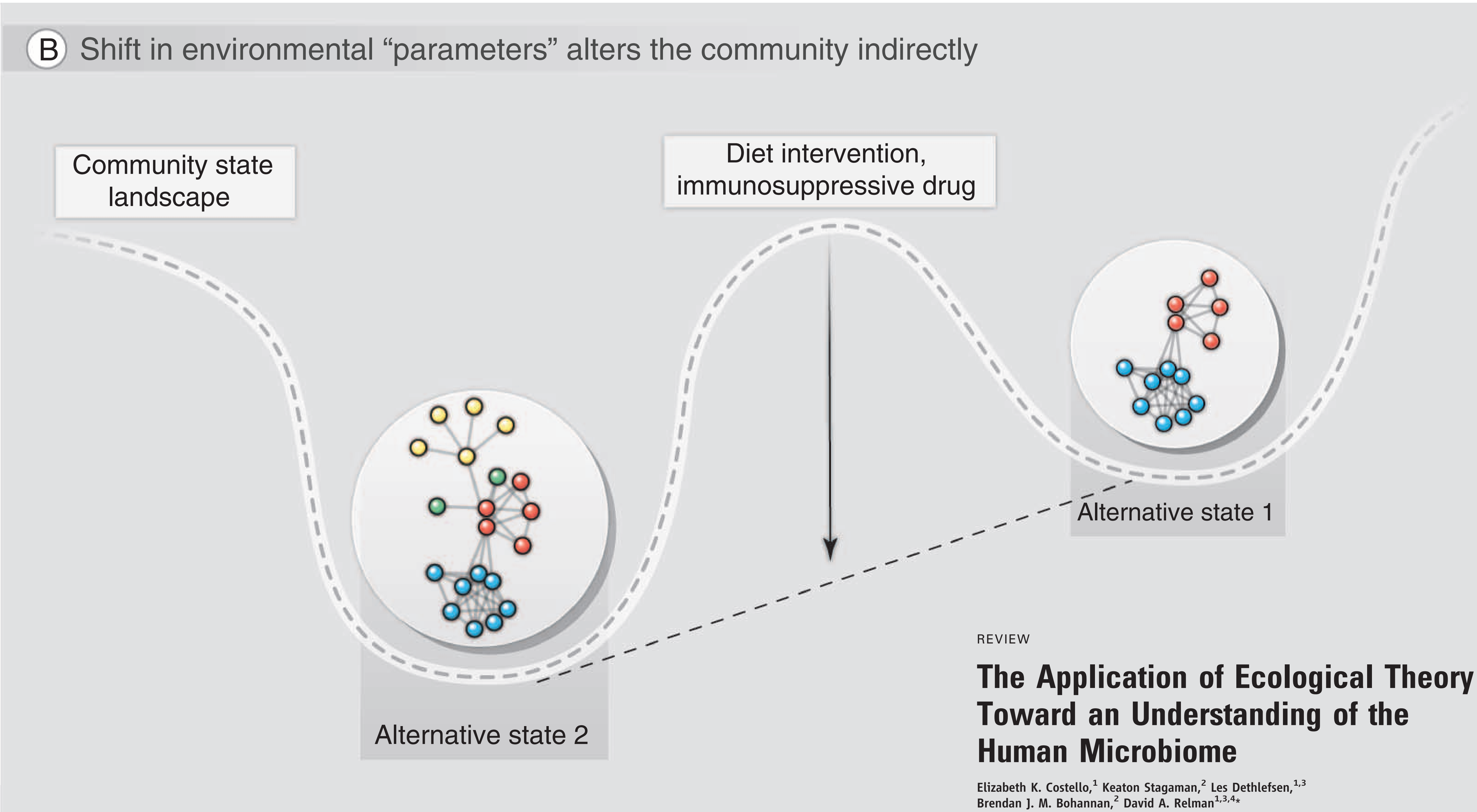
REVIEW

The Application of Ecological Theory Toward an Understanding of the Human Microbiome

Elizabeth K. Costello,[1] Keaton Stagaman,[2] Les Dethlefsen,[1,3] Brendan J. M. Bohannan,[2] David A. Relman[1,3,4]*

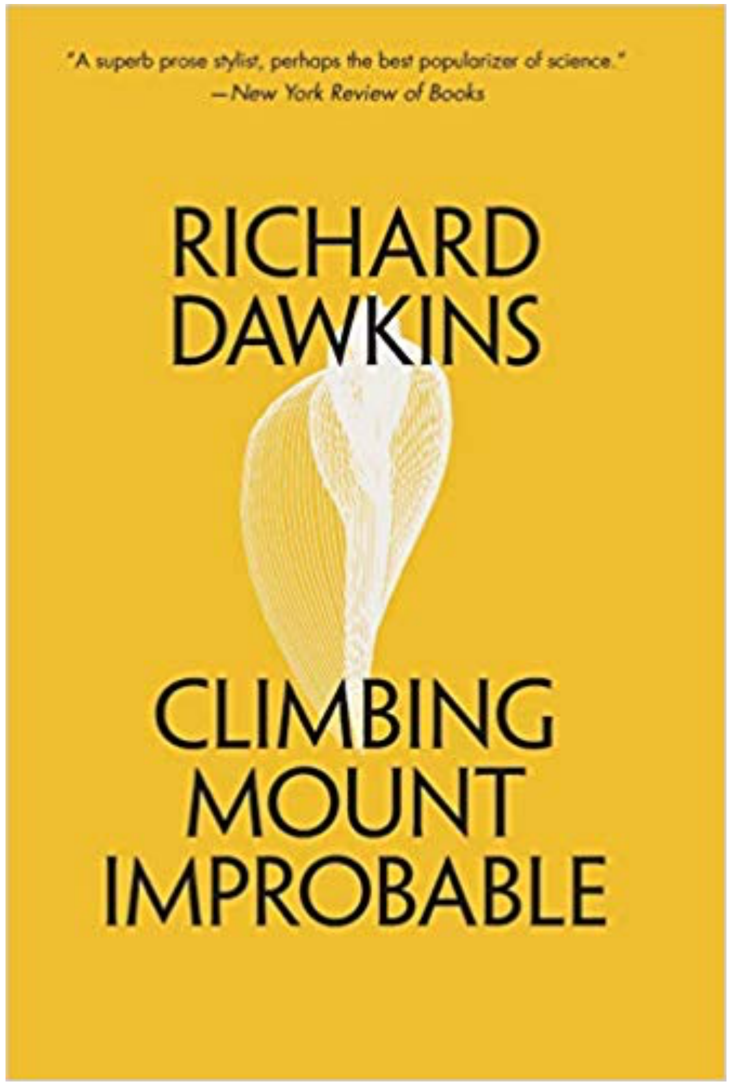
COMMUNITY STATE LANDSCAPES AND ECOSYSTEMS

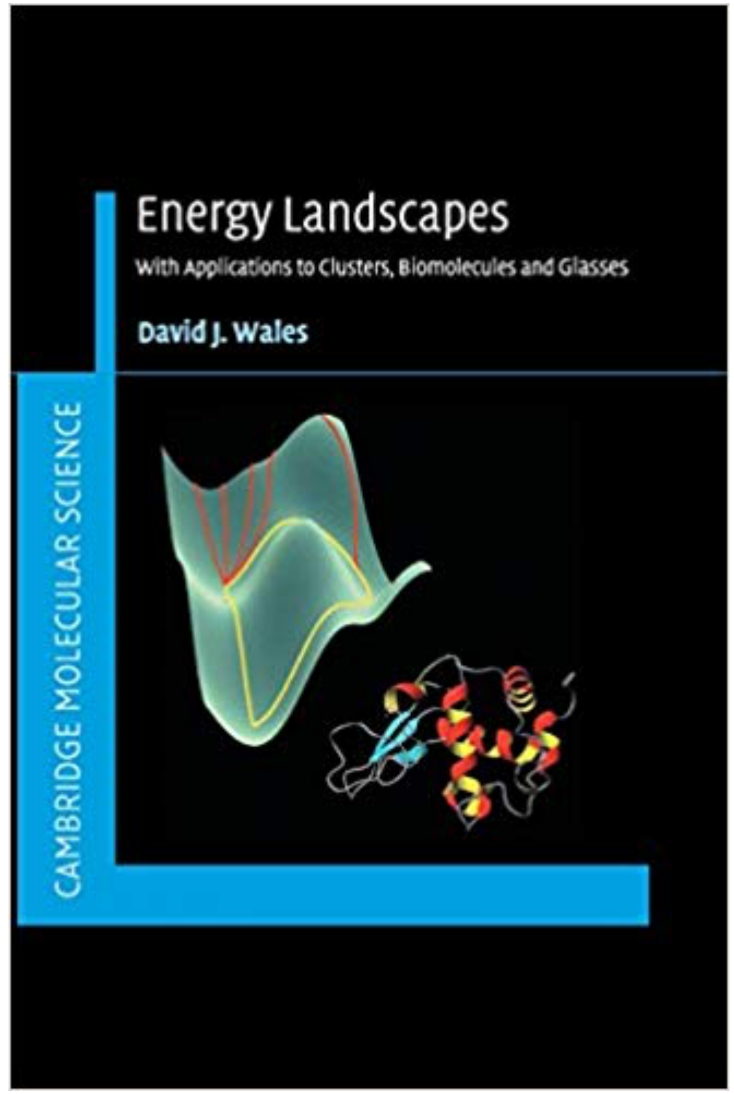
FLATIRON INSTITUTE
Center for Computational Mathematics

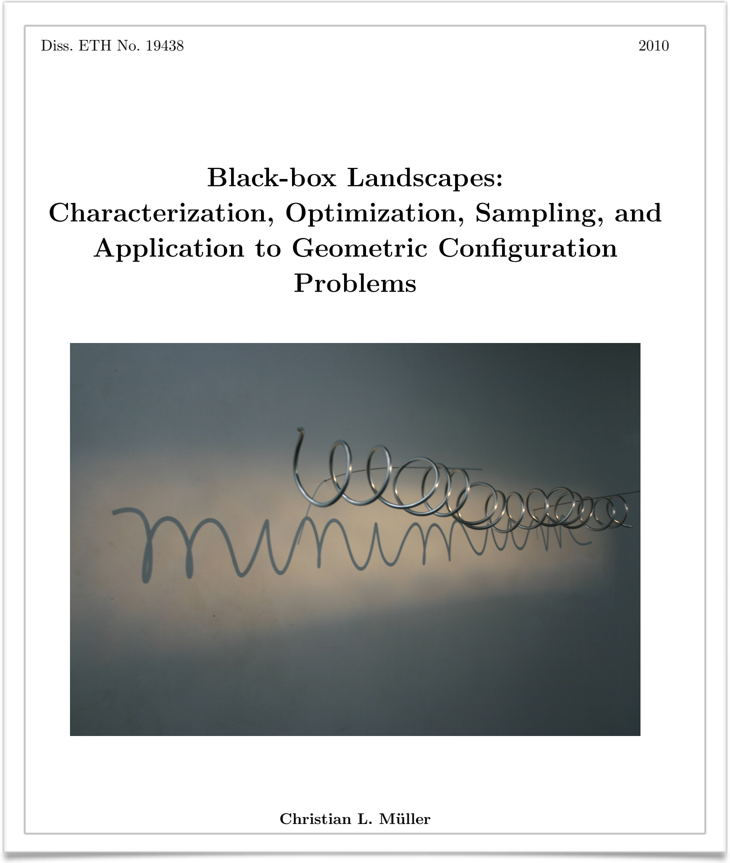


WANT TO KNOW MORE?

FLATIRON INSTITUTE
Center for Computational Mathematics

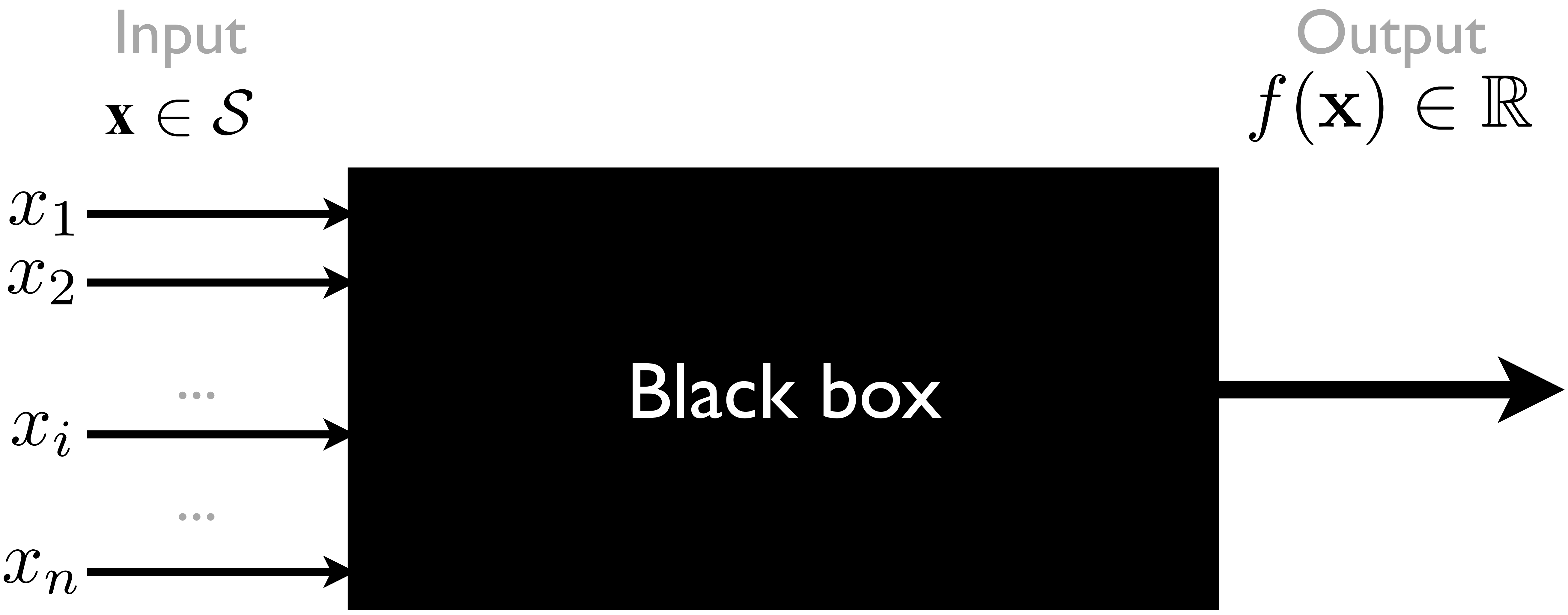






FROM LANDSCAPES TO MATHEMATICAL OPTIMIZATION

FLATIRON INSTITUTE
Center for Computational Mathematics



FROM LANDSCAPES TO MATHEMATICAL OPTIMIZATION

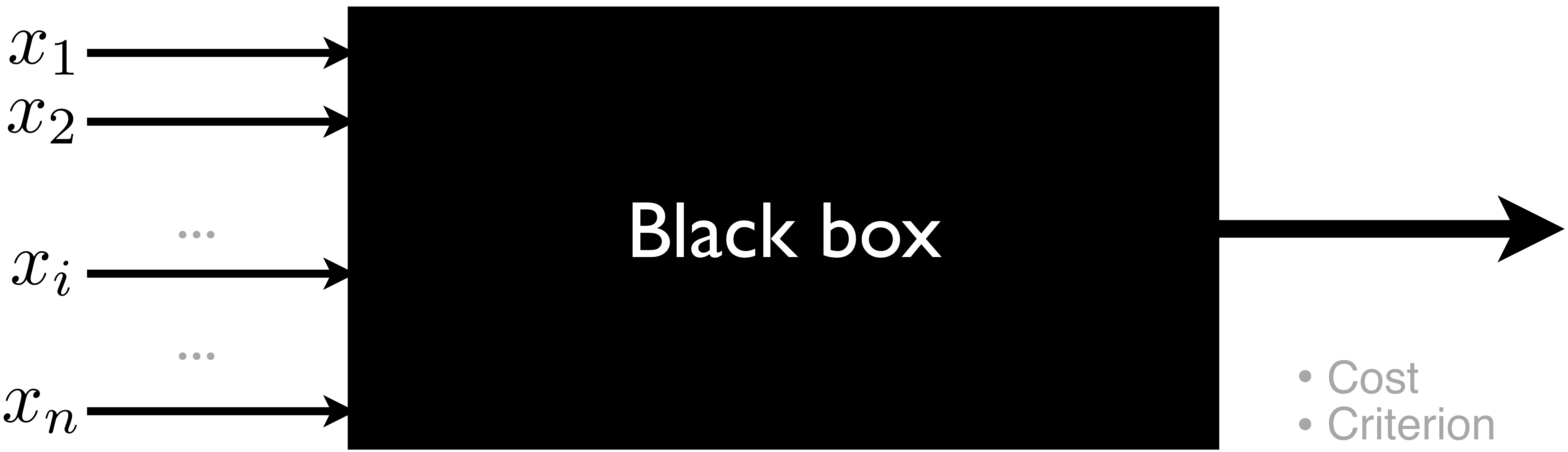
Center for Computational Mathematics

Input

$\mathbf{x} \in \mathcal{S}$

Output

$f(\mathbf{x}) \in \mathbb{R}$



- Variables
- Parameters
- Configuration
- Factors

- Cost
- Criterion
- Objective
- Energy
- Fitness

OPENING UP THE BLACK-BOX: CONTINUOUS OPTIMIZATION PROBLEM

The *standard form* of a continuous optimization problem is[1]

$$\begin{aligned} &\underset{x}{\operatorname{minimize}} && f(x) \\ &\operatorname{subject\ to} && g_i(x) \leq 0, \quad i = 1, \dots, m \\ &&& h_j(x) = 0, \quad j = 1, \dots, p \end{aligned}$$

where

- $f : \mathbb{R}^n \to \mathbb{R}$ is the **objective function** to be minimized over the n-variable vector x,
- $g_i(x) \leq 0$ are called **inequality constraints**
- $h_j(x) = 0$ are called **equality constraints**, and
- $m \geq 0$ and $p \geq 0$.

If $m = p = 0$, the problem is an unconstrained optimization problem. By convention, the standard form defines a **minimization problem**. A **maximization problem** can be treated by negating the objective function.

wikipedia

OPENING UP THE BLACK BOX

FLATIRON INSTITUTE
Center for Computational Mathematics

- What do you know about $\mathbf{x} \in \mathcal{S}$?
- What is the dimensionality of the problem?
- Does the function $f(\mathbf{x})$ have special properties? What are good properties?
- Can you evaluate gradients or higher-order information of the function?

OPENING UP THE BLACK BOX

FLATIRON INSTITUTE
Center for Computational Mathematics

- What do you know about $\mathbf{x} \in \mathcal{S}$?
- What is the dimensionality of the problem?
- Does the function $f(\mathbf{x})$ have special properties? What are good properties?
- Can you evaluate gradients or higher-order information of the function?
- How much does it cost (in computation time/experimental time) to evaluate the function? How often can you evaluate it?
- Is the function value deterministic? Is it stochastic?
- How accurate does the solution need to be?
- …

PURE RANDOM SEARCH

FLATIRON INSTITUTE
Center for Computational Mathematics

Rastrigin, L.A. (1963). "The convergence of the random search method in the extremal control of a many parameter system". *Automation and Remote Control*. **24** (10): 1337–1342.

PURE RANDOM SEARCH

FLATIRON INSTITUTE
Center for Computational Mathematics

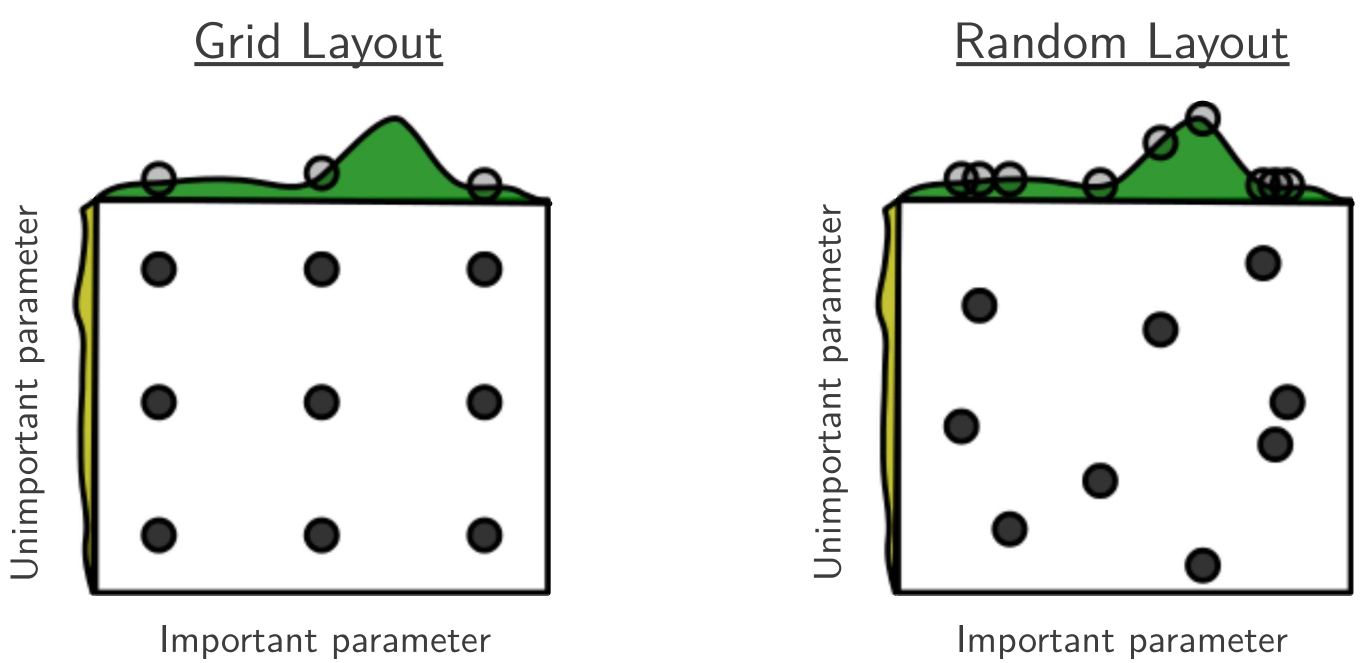
Rastrigin, L.A. (1963). "The convergence of the random search method in the extremal control of a many parameter system". *Automation and Remote Control*. **24** (10): 1337–1342.

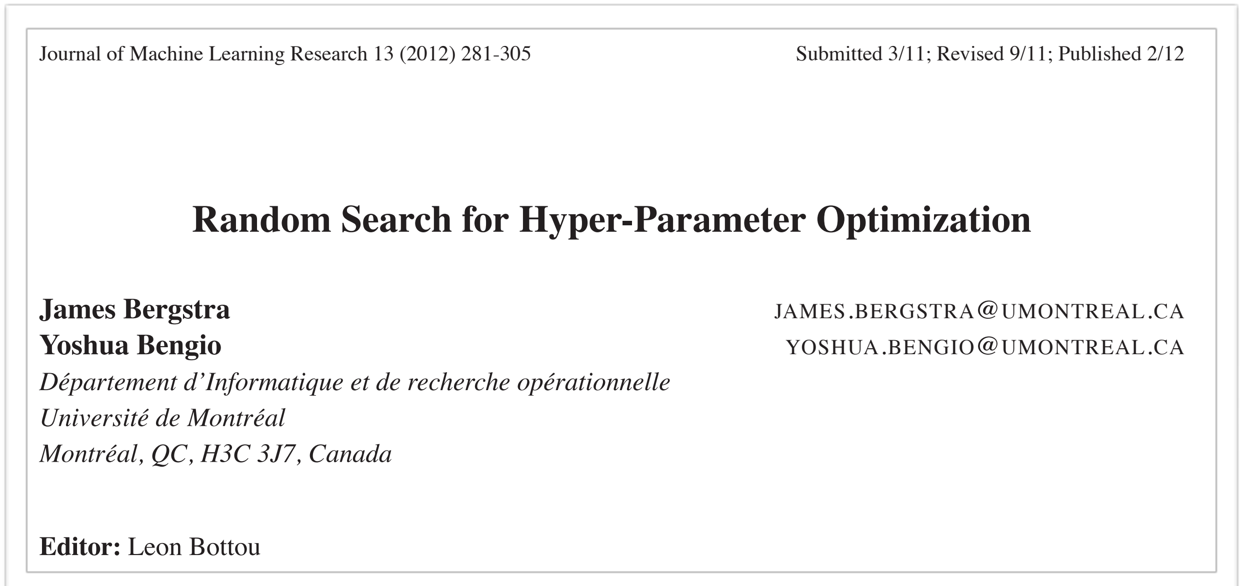
- Use it when you know very little about the function and the function is **costly**
- Useful when your input domain is simple, e.g., a hyper-cube
- Only requires function evaluations, no other information needed
- Better coverage than grid search

PURE RANDOM SEARCH

Rastrigin, L.A. (1963). "The convergence of the random search method in the extremal control of a many parameter system". *Automation and Remote Control*. **24** (10): 1337–1342.

- Use it when you know very little about the function and the function is **costly**
- Useful when your input domain is simple, e.g., a hyper-cube
- Only requires function evaluations, no other information needed
- Better coverage than grid search




Journal of Machine Learning Research 13 (2012) 281-305 Submitted 3/11; Revised 9/11; Published 2/12

Random Search for Hyper-Parameter Optimization

James Bergstra JAMES.BERGSTRA@UMONTREAL.CA
Yoshua Bengio YOSHUA.BENGIO@UMONTREAL.CA
Département d'Informatique et de recherche opérationnelle
Université de Montréal
Montréal, QC, H3C 3J7, Canada

Editor: Leon Bottou

cited 3k times since 2012

QUASI-RANDOM SEARCH

Sobol,I.M. (1967), "Distribution of points in a cube and approximate evaluation of integrals". *Zh. Vych. Mat. Mat. Fiz.* **7**: 784–802 (in Russian); *U.S.S.R Comput. Maths. Math. Phys.* **7**: 86–112 (in English).

QUASI-RANDOM SEARCH

Sobol,I.M. (1967), "Distribution of points in a cube and approximate evaluation of integrals". *Zh. Vych. Mat. Mat. Fiz.* **7**: 784–802 (in Russian); *U.S.S.R Comput. Maths. Math. Phys.* **7**: 86–112 (in English).

- Use quasi-random points rather than random ones to cover the space
- Better space-filling properties
- Works well for up to n=50 dimensions
- (Scrambled) Sobol sequences are good

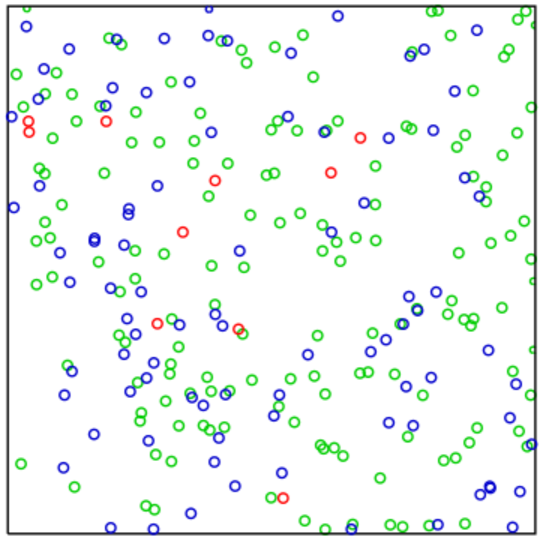
QUASI-RANDOM SEARCH

FLATIRON INSTITUTE
Center for Computational Mathematics

Sobol,I.M. (1967), "Distribution of points in a cube and approximate evaluation of integrals". *Zh. Vych. Mat. Mat. Fiz.* **7**: 784–802 (in Russian); *U.S.S.R Comput. Maths. Math. Phys.* **7**: 86–112 (in English).

- Use quasi-random points rather than random ones to cover the space
- Better space-filling properties
- Works well for up to n=50 dimensions
- (Scrambled) Sobol sequences are good

Pseudo-random points



QUASI-RANDOM SEARCH

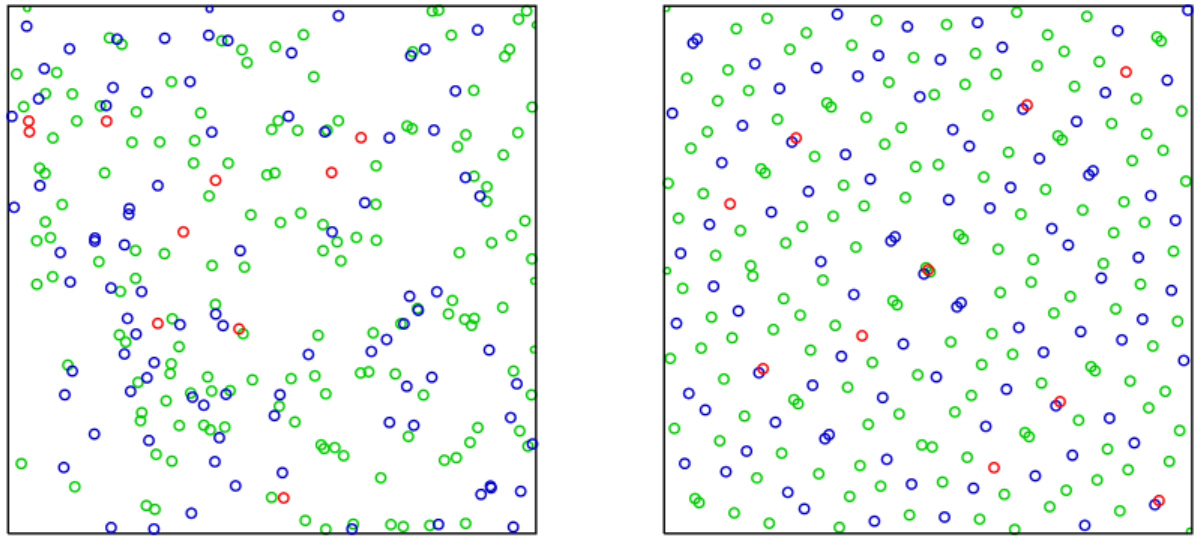
FLATIRON INSTITUTE
Center for Computational Mathematics

Sobol,I.M. (1967), "Distribution of points in a cube and approximate evaluation of integrals". *Zh. Vych. Mat. Mat. Fiz.* **7**: 784–802 (in Russian); *U.S.S.R Comput. Maths. Math. Phys.* **7**: 86–112 (in English).

- Use quasi-random points rather than random ones to cover the space
- Better space-filling properties
- Works well for up to n=50 dimensions
- (Scrambled) Sobol sequences are good

Pseudo-random points

Quasi-random points



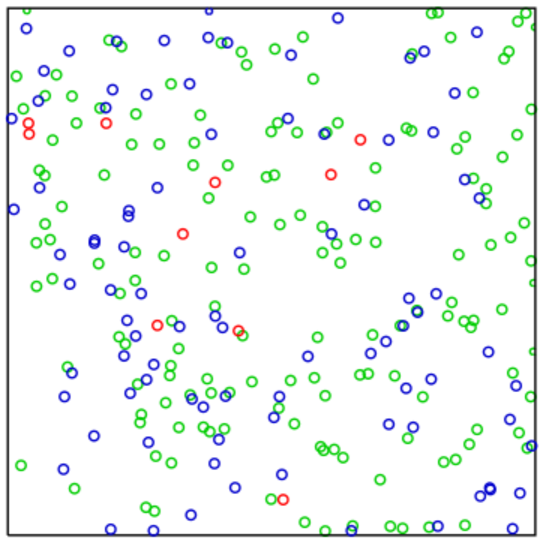
QUASI-RANDOM SEARCH

FLATIRON INSTITUTE
Center for Computational Mathematics

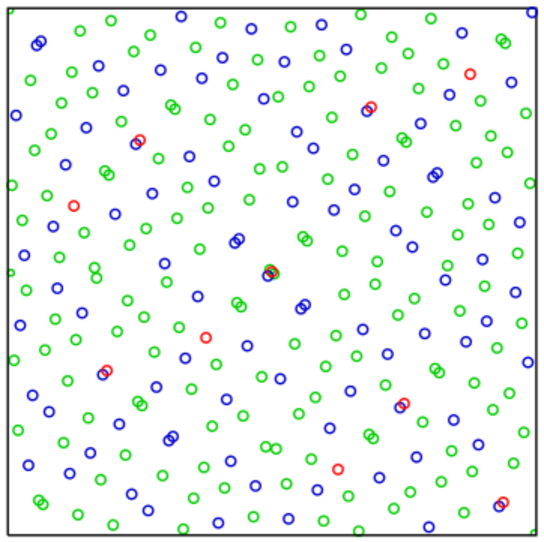
Sobol,I.M. (1967), "Distribution of points in a cube and approximate evaluation of integrals". *Zh. Vych. Mat. Mat. Fiz.* **7**: 784–802 (in Russian); *U.S.S.R Comput. Maths. Math. Phys.* **7**: 86–112 (in English).

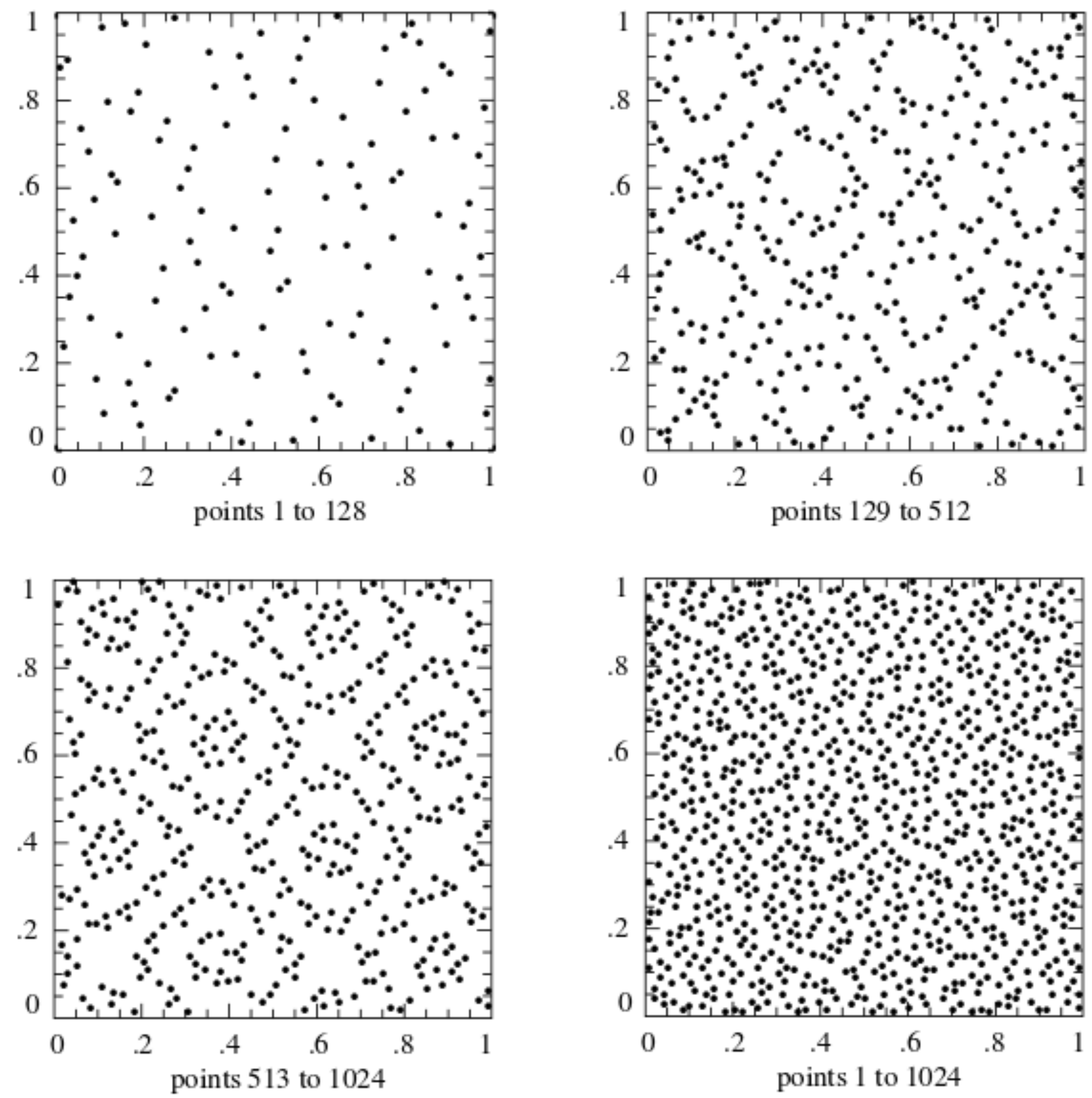
- Use quasi-random points rather than random ones to cover the space
- Better space-filling properties
- Works well for up to n=50 dimensions
- (Scrambled) Sobol sequences are good

Pseudo-random points



Quasi-random points





DERIVATIVE-FREE OPTIMIZATION AND EVOLUTION STRATEGIES

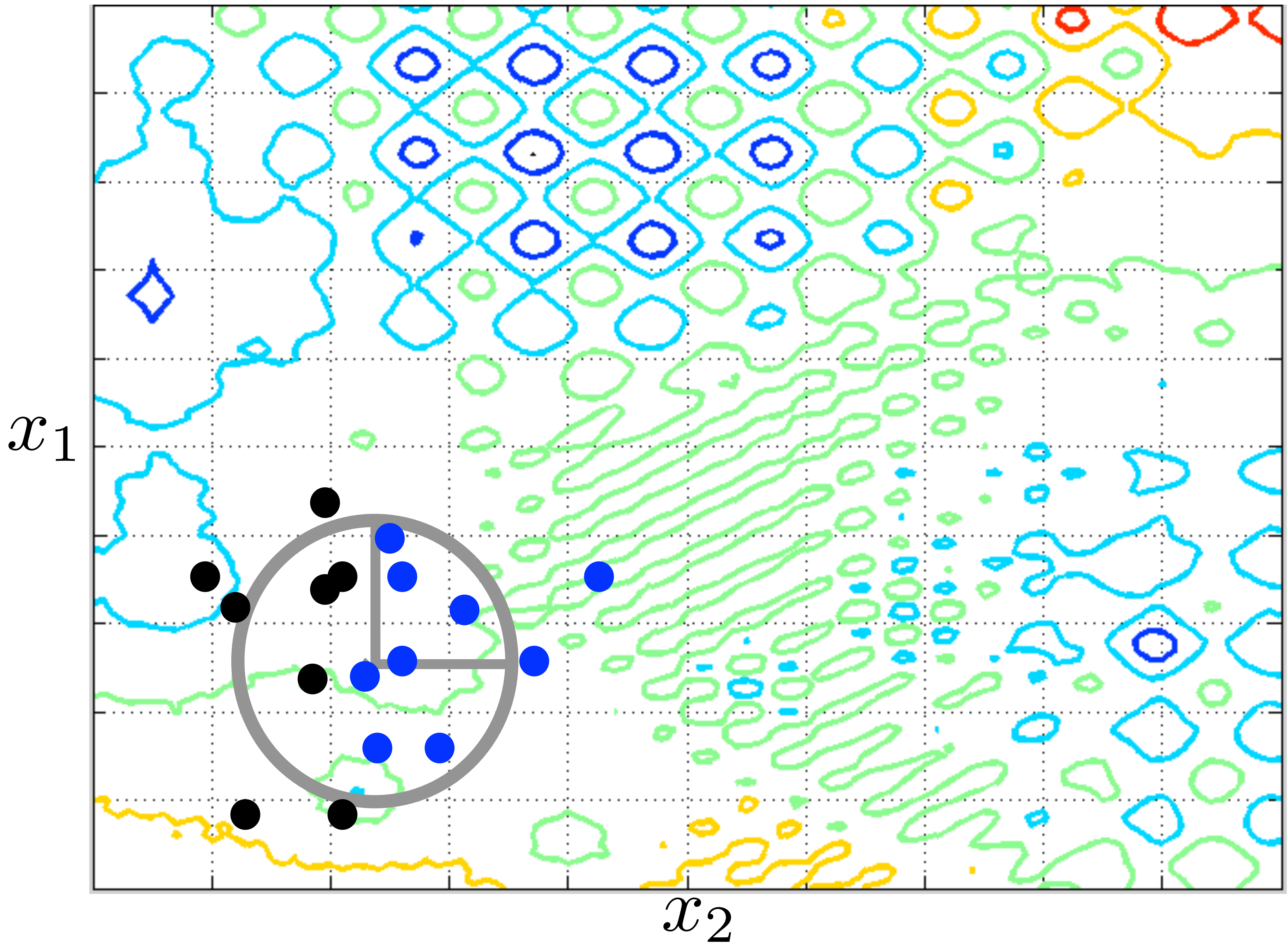
- Use it when you know very little about the function and the function is **not costly**, i.e., you can evaluate $O(n^2)$ points
- Input domain is simple, e.g. a hyper-cube, not too high-dimensional
- Typically used in **simulation-based optimization** where only function evaluations are available
- Popular method: Nelder-Mead Simplex method (not recommended), Pattern search, Covariance Matrix Adaptation ES

CMA-ES resources

http://www.cmap.polytechnique.fr/~nikolaus.hansen/

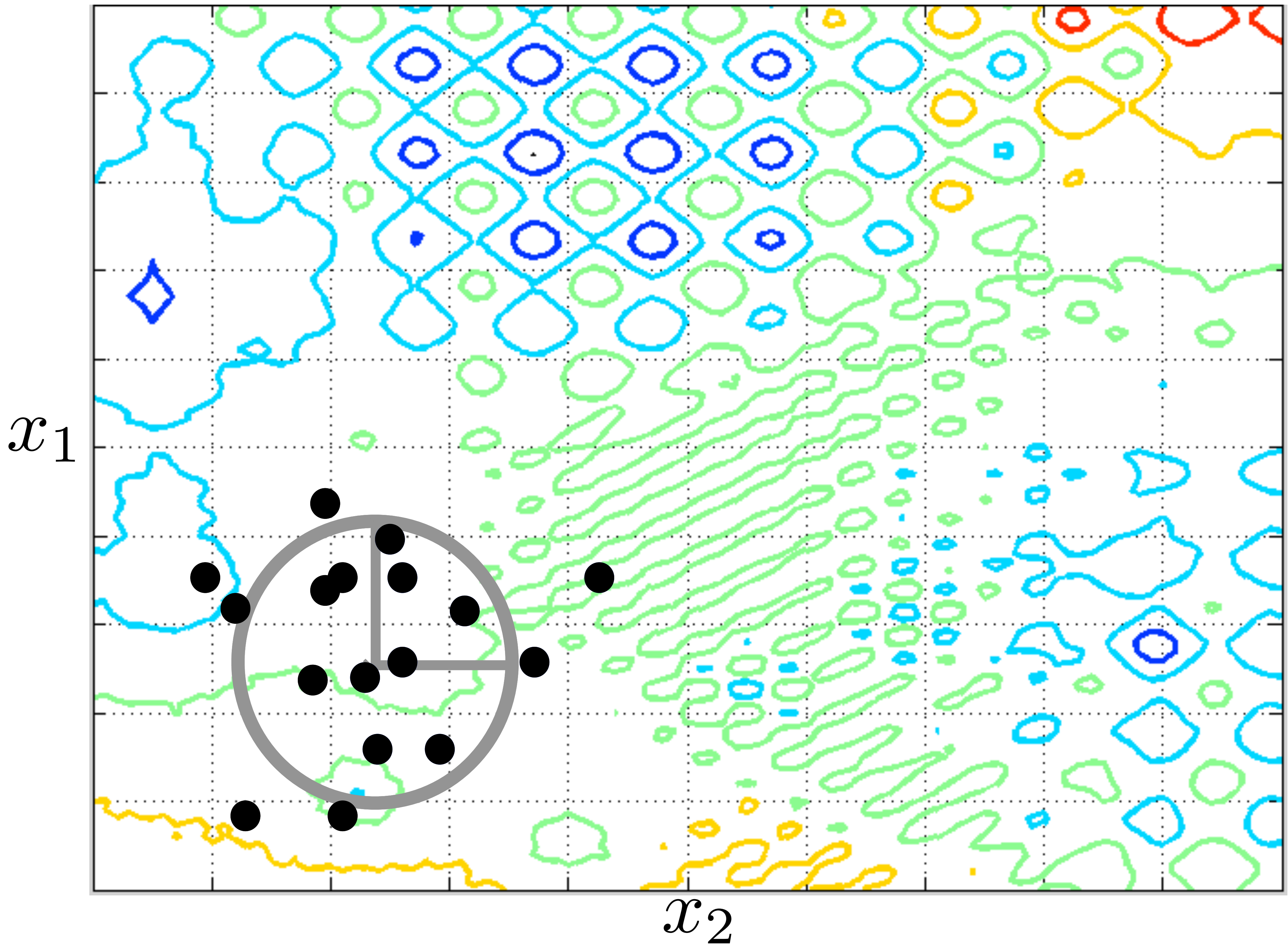
GRADIENT-FREE OPTIMIZATION WITH CMA-ES

FLATIRON INSTITUTE
Center for Computational Mathematics

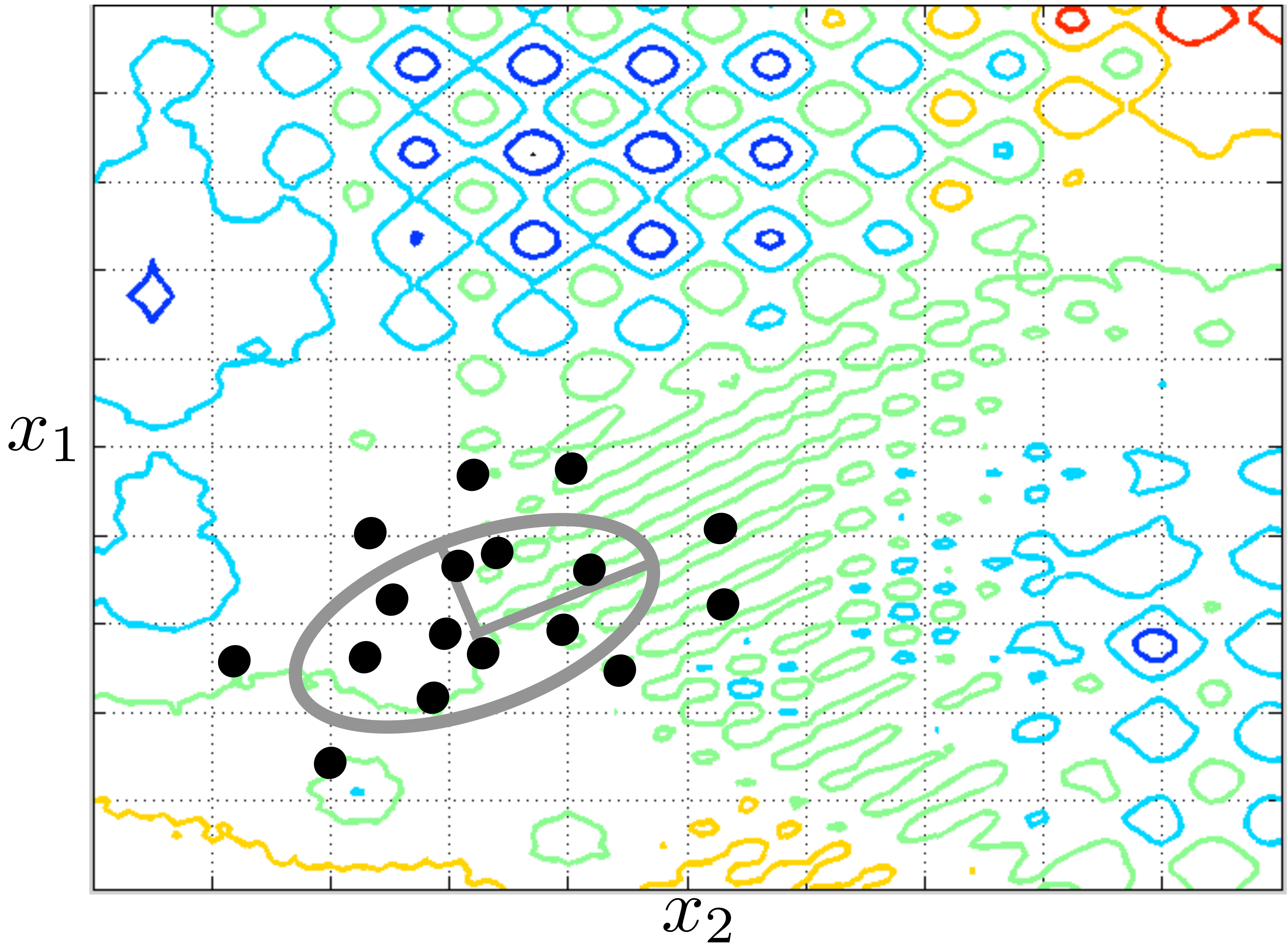


GRADIENT-FREE OPTIMIZATION WITH CMA-ES

FLATIRON INSTITUTE
Center for Computational Mathematics

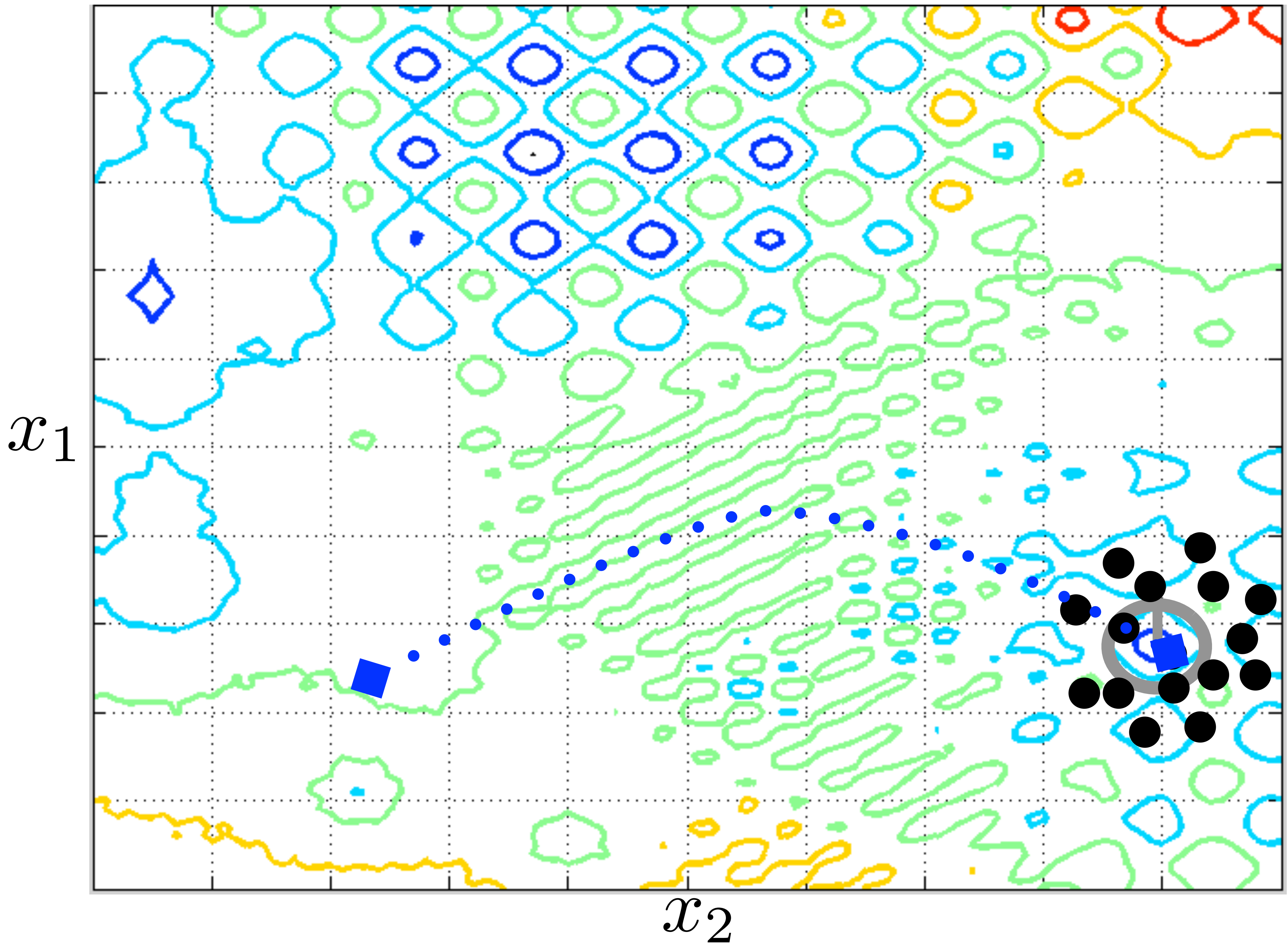


GRADIENT-FREE OPTIMIZATION WITH CMA-ES



GRADIENT-FREE OPTIMIZATION WITH CMA-ES

FLATIRON INSTITUTE
Center for Computational Mathematics



The ($\mu/\mu_w,\lambda$)-CMA-ES in mathematical terms

Sampling

$$\mathbf{x}_k^{(g+1)} \sim \mathbf{m}^{(g)} + \sigma^{(g)} \mathcal{N}\left(\mathbf{0}, \mathbf{C}^{(g)}\right) \quad \text{for } k = 1, \ldots, \lambda.$$

GRADIENT-FREE OPTIMIZATION WITH CMA-ES

FLATIRON INSTITUTE
Center for Computational Mathematics

The ($\mu/\mu_w,\lambda$)-CMA-ES in mathematical terms

Sampling

$$\mathbf{x}_k^{(g+1)} \sim \mathbf{m}^{(g)} + \sigma^{(g)} \mathcal{N}\left(\mathbf{0}, \mathbf{C}^{(g)}\right) \qquad \text{for } k = 1, \ldots, \lambda.$$

Evaluation

Calculate fitness of all λ individuals and sort them

GRADIENT-FREE OPTIMIZATION WITH CMA-ES

The ($\mu/\mu_w,\lambda$)-CMA-ES in mathematical terms

Sampling

$$\mathbf{x}_k^{(g+1)} \sim \mathbf{m}^{(g)} + \sigma^{(g)} \mathcal{N}\left(\mathbf{0}, \mathbf{C}^{(g)}\right) \quad \text{for } k = 1, \ldots, \lambda.$$

Evaluation

Calculate fitness of all λ individuals and sort them

Selection

$$\mathbf{m}^{(g+1)} = \sum_{i=1}^{\mu} w_i \mathbf{x}_{i:\lambda}^{(g+1)} \qquad \sum_{i=1}^{\mu} w_i = 1, \quad w_1 \geq w_2 \geq \ldots \geq w_\mu > 0$$

FLATIRON INSTITUTE
Center for Computational Mathematics

The ($\mu/\mu_w,\lambda$)-CMA-ES in mathematical terms

Sampling

$$\mathbf{x}_k^{(g+1)} \sim \mathbf{m}^{(g)} + \sigma^{(g)} \mathcal{N}\left(\mathbf{0}, \mathbf{C}^{(g)}\right) \qquad \text{for } k = 1, \ldots, \lambda.$$

Evaluation

Calculate fitness of all λ individuals and sort them

Selection

$$\mathbf{m}^{(g+1)} = \sum_{i=1}^{\mu} w_i \mathbf{x}_{i:\lambda}^{(g+1)} \qquad \sum_{i=1}^{\mu} w_i = 1, \quad w_1 \geq w_2 \geq \ldots \geq w_\mu > 0$$

Recombination Adaptation

$$\mathbf{C}^{(g+1)} = (1 - c_{\text{cov}})\mathbf{C}^{(g)} + \frac{c_{\text{cov}}}{\mu_{\text{cov}}} \underbrace{\mathbf{p}_c^{(g+1)} \mathbf{p}_c^{(g+1)^T}}_{\text{rank-one-update}} + c_{\text{cov}} \left(1 - \frac{1}{\mu_{\text{cov}}}\right) \times \underbrace{\sum_{i=1}^{\mu} w_i \mathbf{y}_{i:\lambda}^{(g+1)} \left(\mathbf{y}_{i:\lambda}^{(g+1)}\right)^T}_{\text{rank-}\mu\text{-update}},$$

GRADIENT-FREE OPTIMIZATION WITH CMA-ES

FLATIRON INSTITUTE
Center for Computational Mathematics

The ($\mu/\mu_w,\lambda$)-CMA-ES in mathematical terms

Sampling

$$\mathbf{x}_k^{(g+1)} \sim \mathbf{m}^{(g)} + \sigma^{(g)} \mathcal{N}\left(\mathbf{0}, \mathbf{C}^{(g)}\right) \qquad \text{for } k = 1, \ldots, \lambda.$$

Evaluation

Calculate fitness of all λ individuals and sort them

Selection

$$\mathbf{m}^{(g+1)} = \sum_{i=1}^{\mu} w_i \mathbf{x}_{i:\lambda}^{(g+1)} \qquad \sum_{i=1}^{\mu} w_i = 1, \quad w_1 \geq w_2 \geq \ldots \geq w_\mu > 0$$

Recombination
Adaptation

$$\mathbf{C}^{(g+1)} = (1 - c_{\text{cov}})\mathbf{C}^{(g)} + \frac{c_{\text{cov}}}{\mu_{\text{cov}}} \underbrace{\mathbf{p}_c^{(g+1)} \mathbf{p}_c^{(g+1)^T}}_{\text{rank-one-update}} + c_{\text{cov}} \left(1 - \frac{1}{\mu_{\text{cov}}}\right) \times \underbrace{\sum_{i=1}^{\mu} w_i \mathbf{y}_{i:\lambda}^{(g+1)} \left(\mathbf{y}_{i:\lambda}^{(g+1)}\right)^T}_{\text{rank-}\mu\text{-update}},$$

$$\sigma^{(g+1)} = \sigma^{(g)} \exp\left(\frac{c_\sigma}{d_\sigma}\left(\frac{||\mathbf{p}_\sigma^{(g+1)}||}{E||\mathcal{N}(\mathbf{0}, \mathbf{I})||} - 1\right)\right).$$

A NOTE ON DESIGN PRINCIPLES FOR OPTIMIZATION HEURISTICS

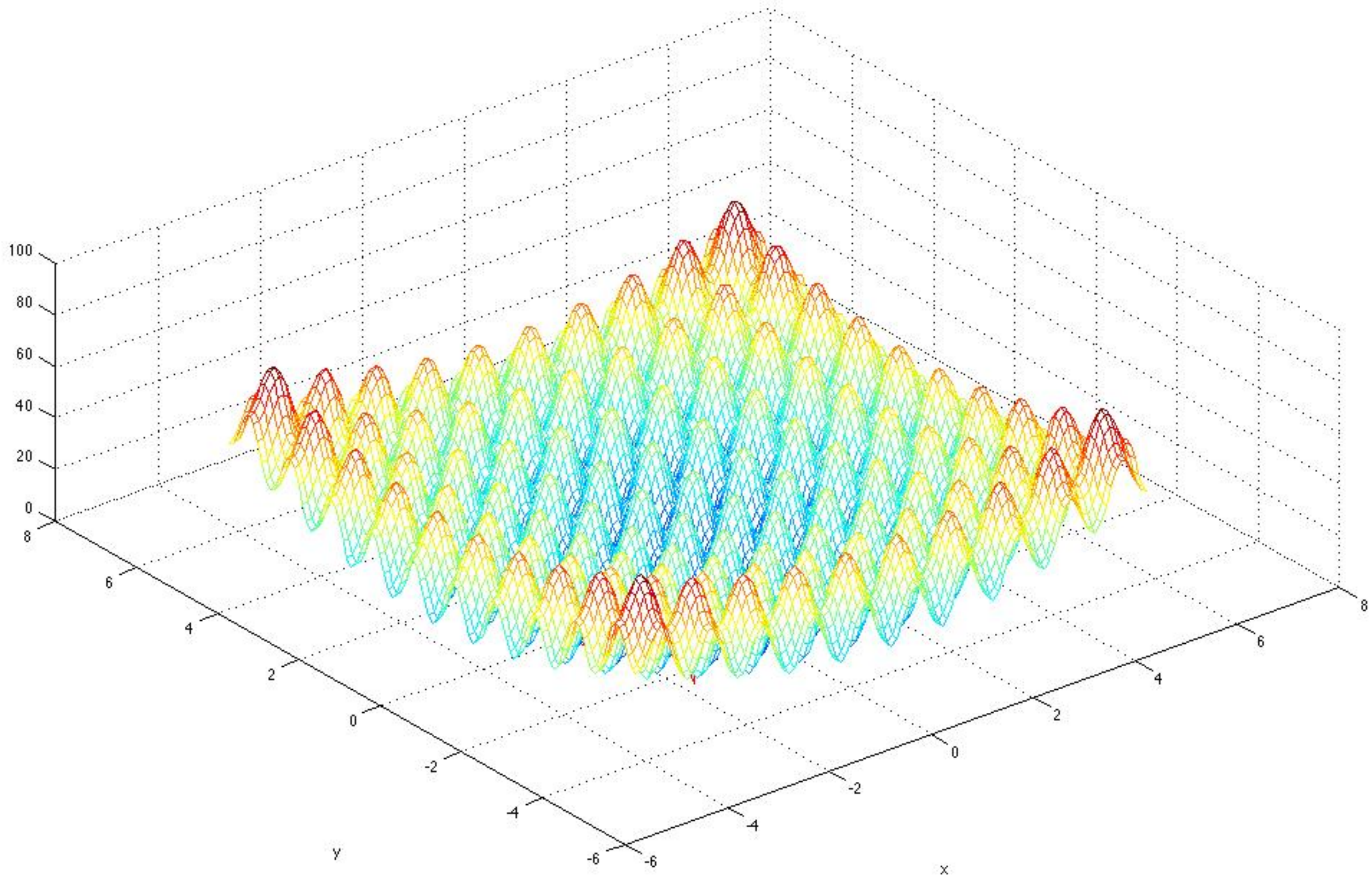
FLATIRON INSTITUTE
Center for Computational Mathematics

- Use invariance (symmetry) principles as much as possible
- CMA-ES is invariant to affine transformations of the domain
- CMA-ES is invariant to monotone transformations of the objective function

CMA-ES ON RASTRIGIN FUNCTION

Rastrigin's Function

$$f(\vec{x}) = 10 \times n + \sum_{i=1}^{n} \left(x_i^2 - 10 \times \cos(2\pi x_i) \right)$$

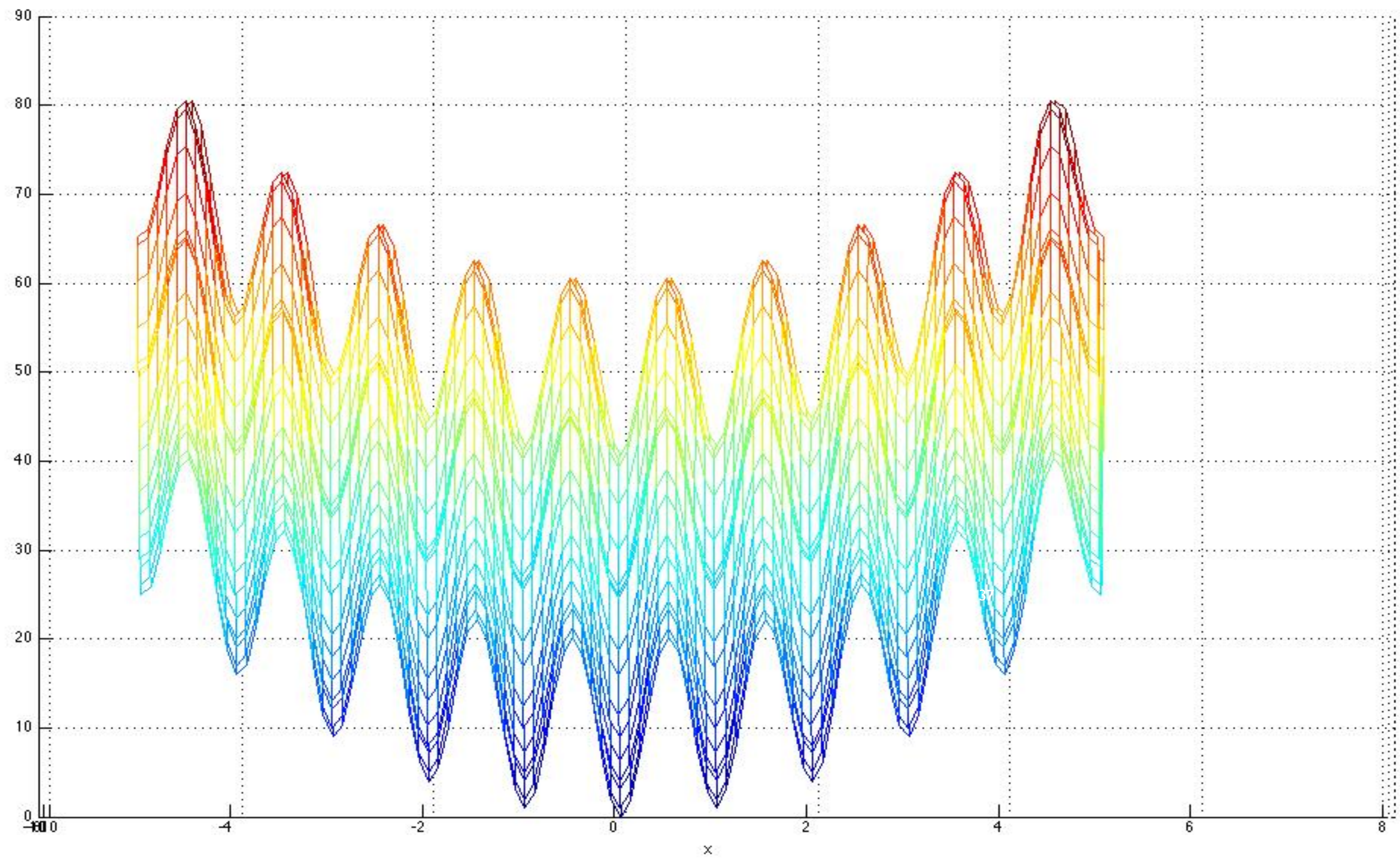


CMA-ES ON RASTRIGIN FUNCTION

FLATIRON INSTITUTE
Center for Computational Mathematics

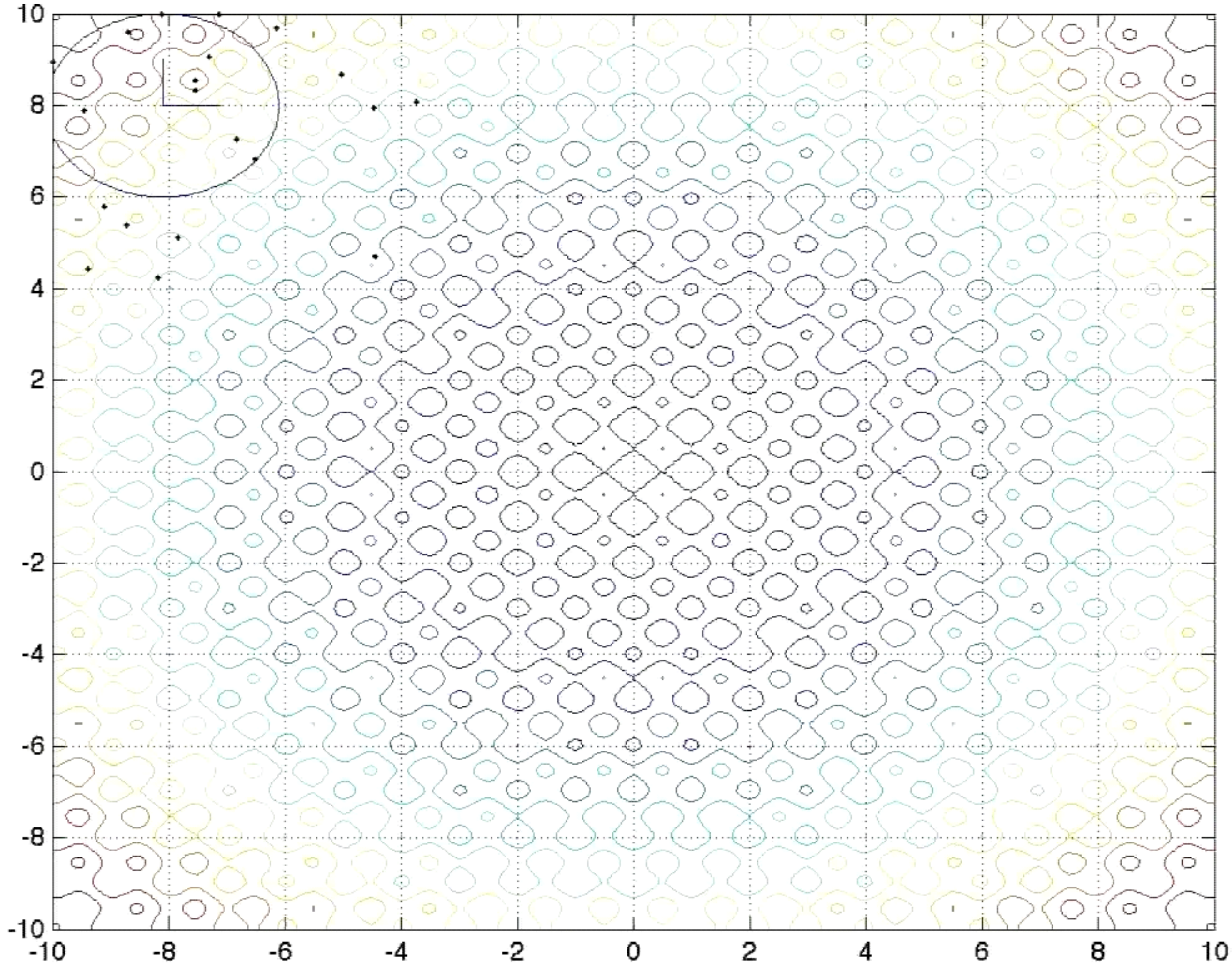
Rastrigin's Function

$$f(\vec{x}) = 10 \times n + \sum_{i=1}^{n} \left(x_i^2 - 10 \times \cos(2\pi x_i) \right)$$

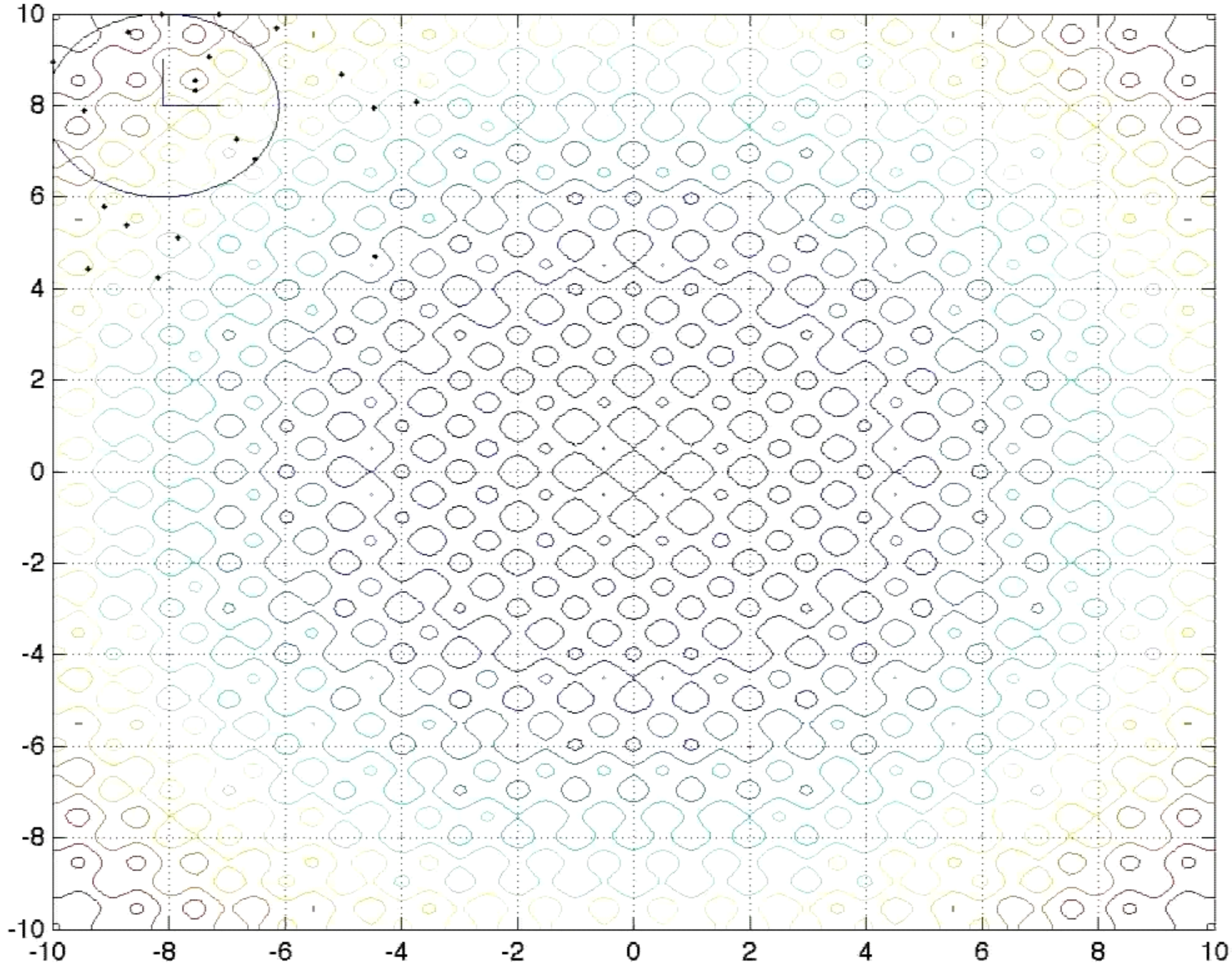


CMA-ES ON RASTRIGIN FUNCTION

FLATIRON INSTITUTE
Center for Computational Mathematics

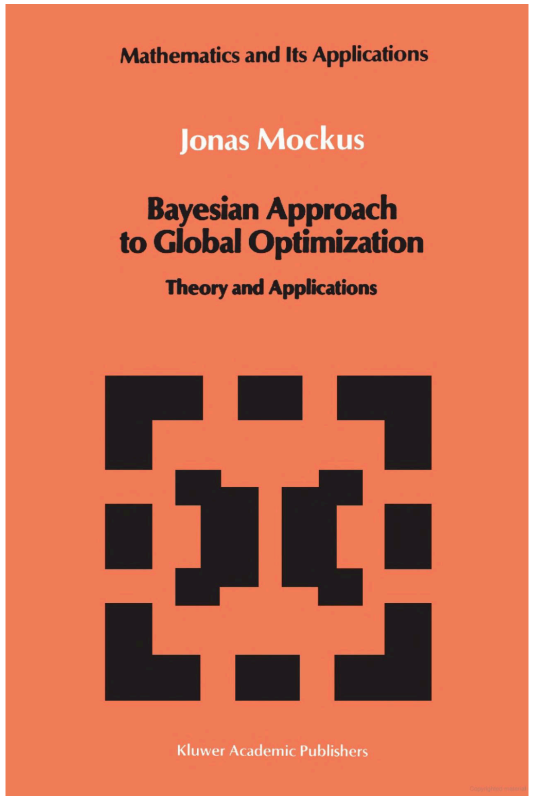


CMA-ES ON RASTRIGIN FUNCTION



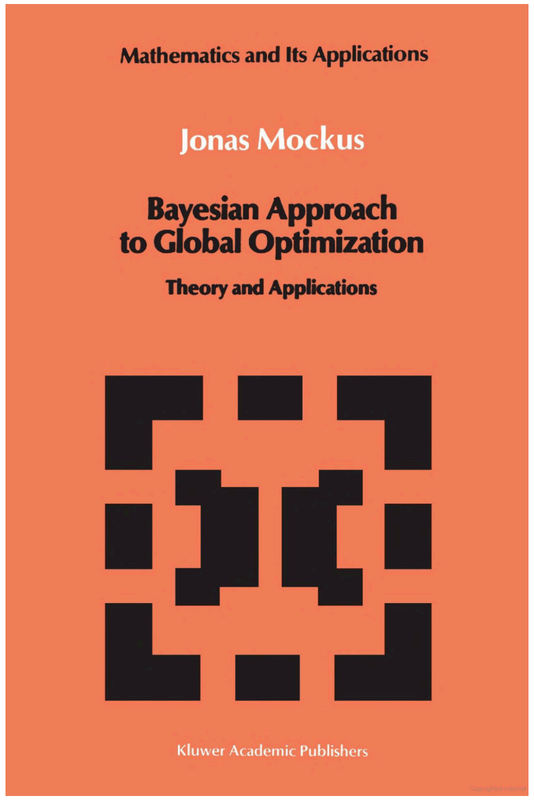
BAYESIAN OPTIMIZATION

- Bayesian optimization is a type of sequential design scheme
- An acquisition function guides the generation of a new function evaluation that balances exploration and exploitation
- Builds a surrogate model of the function (often with Gaussian Processes) (see Directed Evolution example)
- Use it when you know very little about the function and the function is **costly** and low-dimensional
- Input domain is simple, e.g. a hyper-cube



BAYESIAN OPTIMIZATION

- Bayesian optimization is a type of sequential design scheme
- An acquisition function guides the generation of a new function evaluation that balances exploration and exploitation
- Builds a surrogate model of the function (often with Gaussian Processes) (see Directed Evolution example)
- Use it when you know very little about the function and the function is **costly** and low-dimensional
- Input domain is simple, e.g. a hyper-cube



Practical Bayesian Optimization of Machine Learning Algorithms

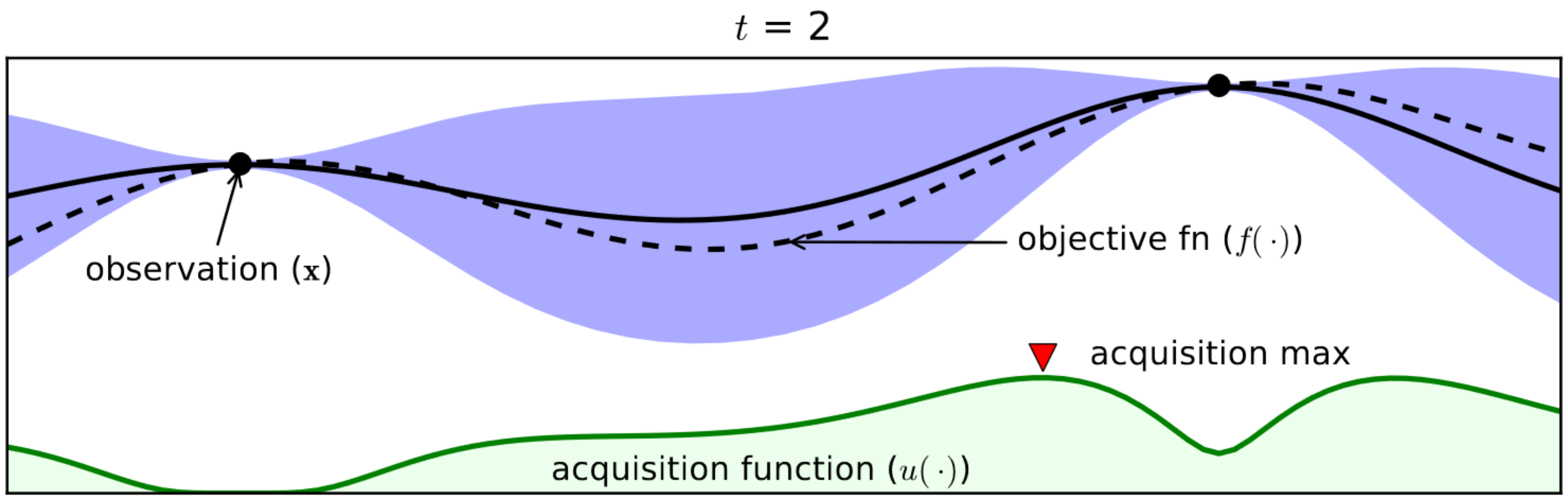
Jasper Snoek
Department of Computer Science
University of Toronto
jasper@cs.toronto.edu

Hugo Larochelle
Department of Computer Science
University of Sherbrooke
hugo.larochelle@usherbrooke.edu

Ryan P. Adams
School of Engineering and Applied Sciences
Harvard University
rpa@seas.harvard.edu

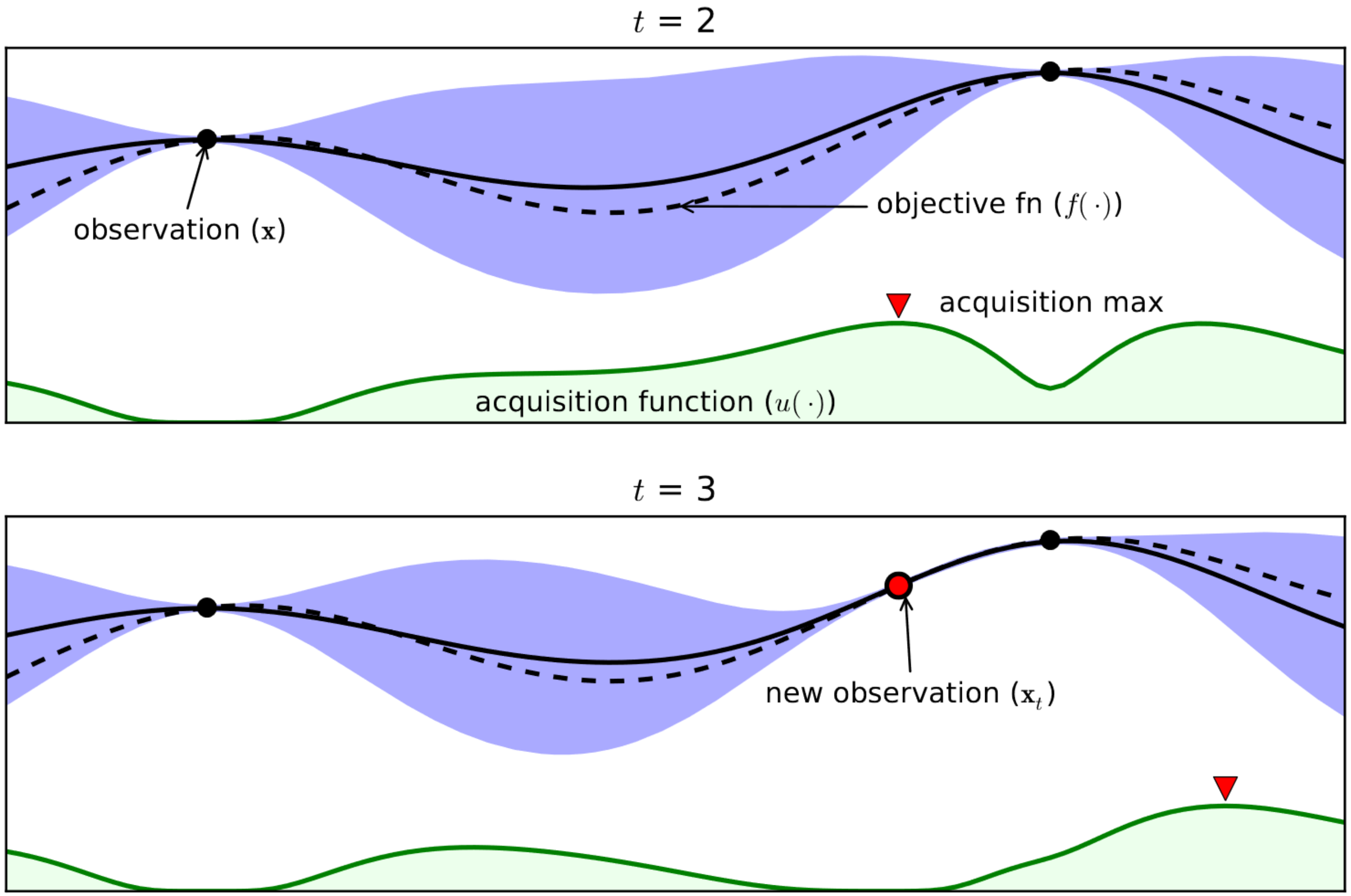
BAYESIAN OPTIMIZATION

FLATIRON INSTITUTE
Center for Computational Mathematics

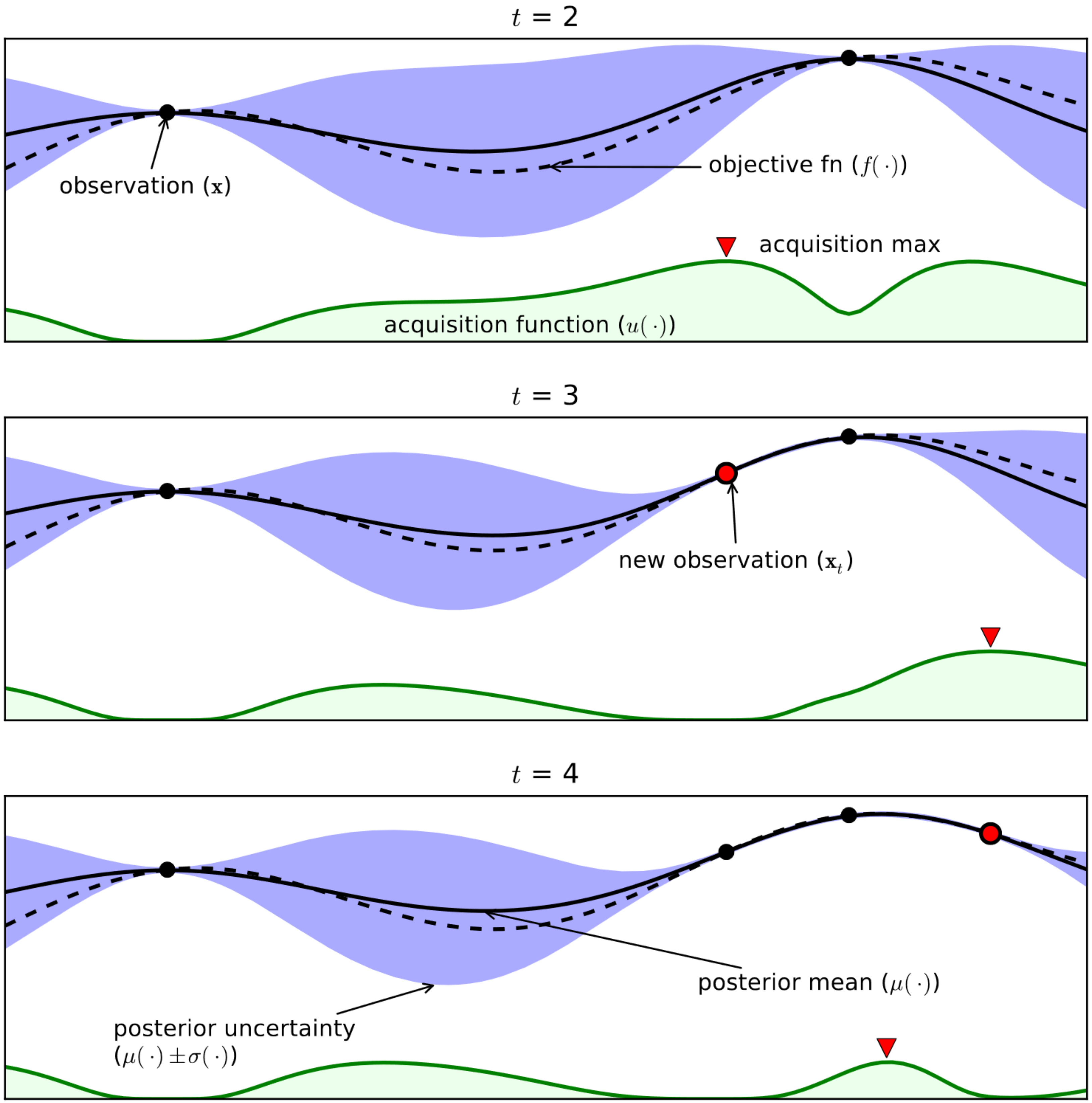


BAYESIAN OPTIMIZATION

FLATIRON INSTITUTE
Center for Computational Mathematics

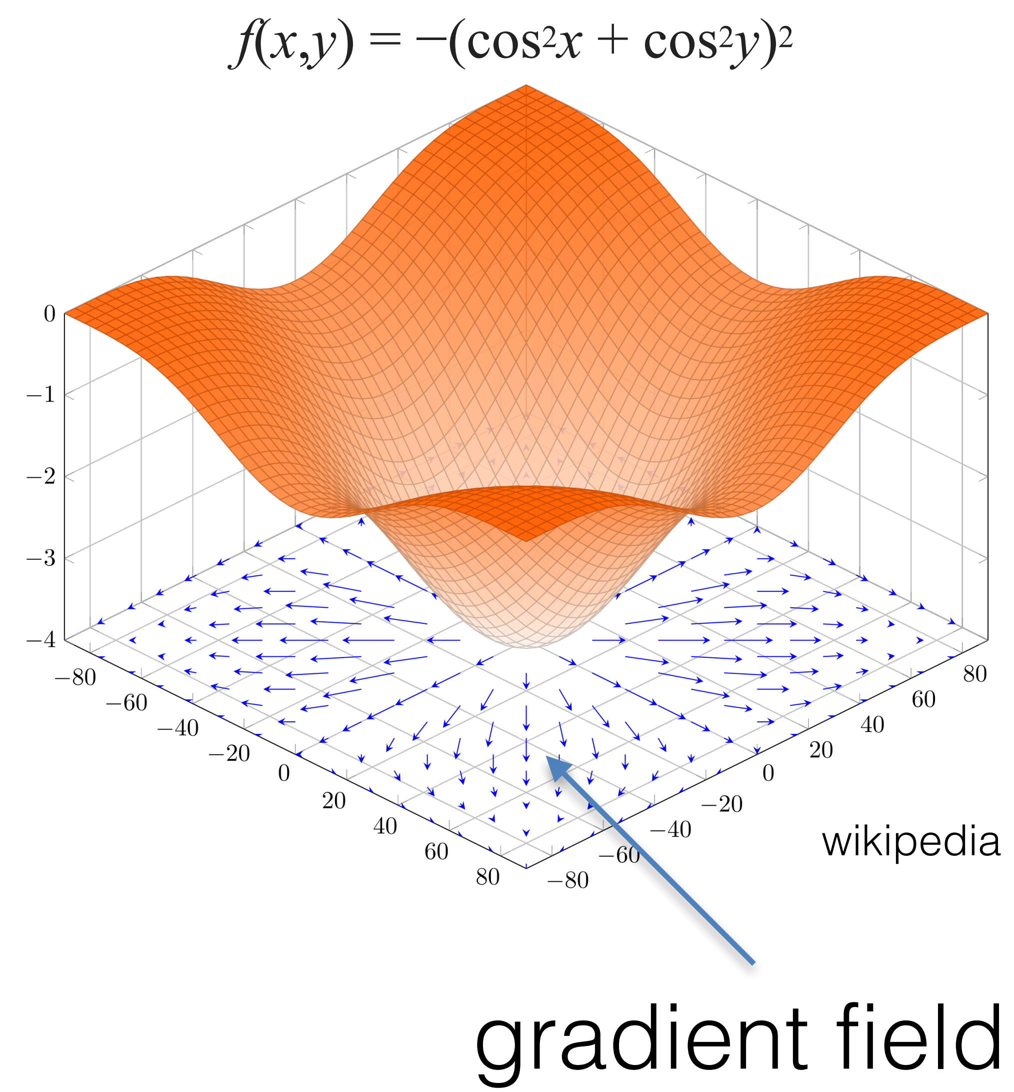


BAYESIAN OPTIMIZATION



GRADIENT-BASED OPTIMIZATION

FLATIRON INSTITUTE
Center for Computational Mathematics

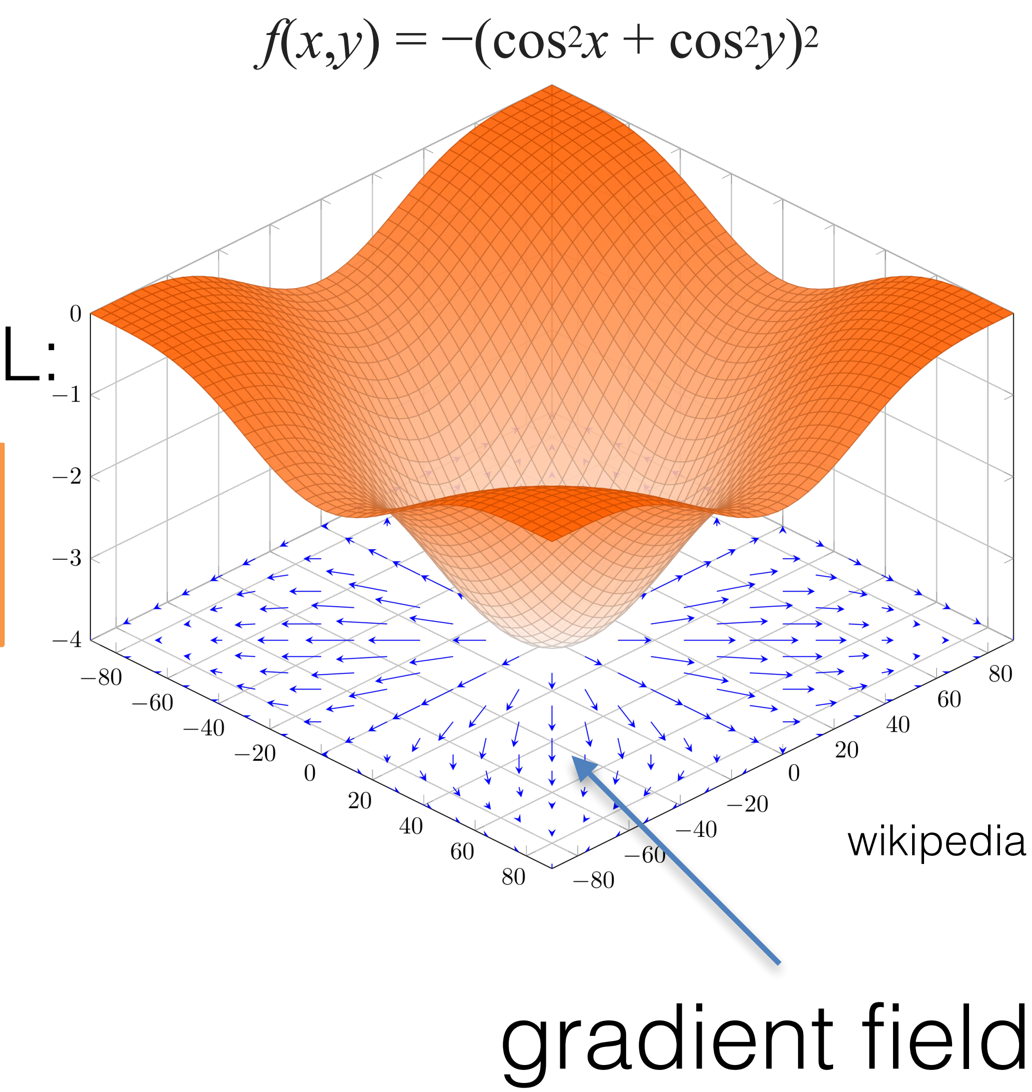


GRADIENT-BASED OPTIMIZATION

FLATIRON INSTITUTE
Center for Computational Mathematics

- The gradient of the function f is available
- The function can be high-dimensional
- The function is smooth with Lipschitz constant L:

$$f(\mathbf{y}) \leq f(\mathbf{x}) + \nabla f(\mathbf{x})^\top (\mathbf{y} - \mathbf{x}) + \frac{L}{2}\|\mathbf{x} - \mathbf{y}\|^2, \quad \forall \mathbf{x}, \mathbf{y} \in \mathbb{R}^d.$$



GRADIENT-BASED OPTIMIZATION

FLATIRON INSTITUTE
Center for Computational Mathematics

- The gradient of the function f is available
- The function can be high-dimensional
- The function is smooth with Lipschitz constant L:

$$f(\mathbf{y}) \leq f(\mathbf{x}) + \nabla f(\mathbf{x})^\top(\mathbf{y}-\mathbf{x}) + \frac{L}{2}\|\mathbf{x}-\mathbf{y}\|^2, \quad \forall \mathbf{x},\mathbf{y} \in \mathbb{R}^d.$$

- Gradient descent:

Goal: Find $\mathbf{x} \in \mathbb{R}^d$ such that

$$f(\mathbf{x}) - f(\mathbf{x}^\star) \leq \varepsilon.$$

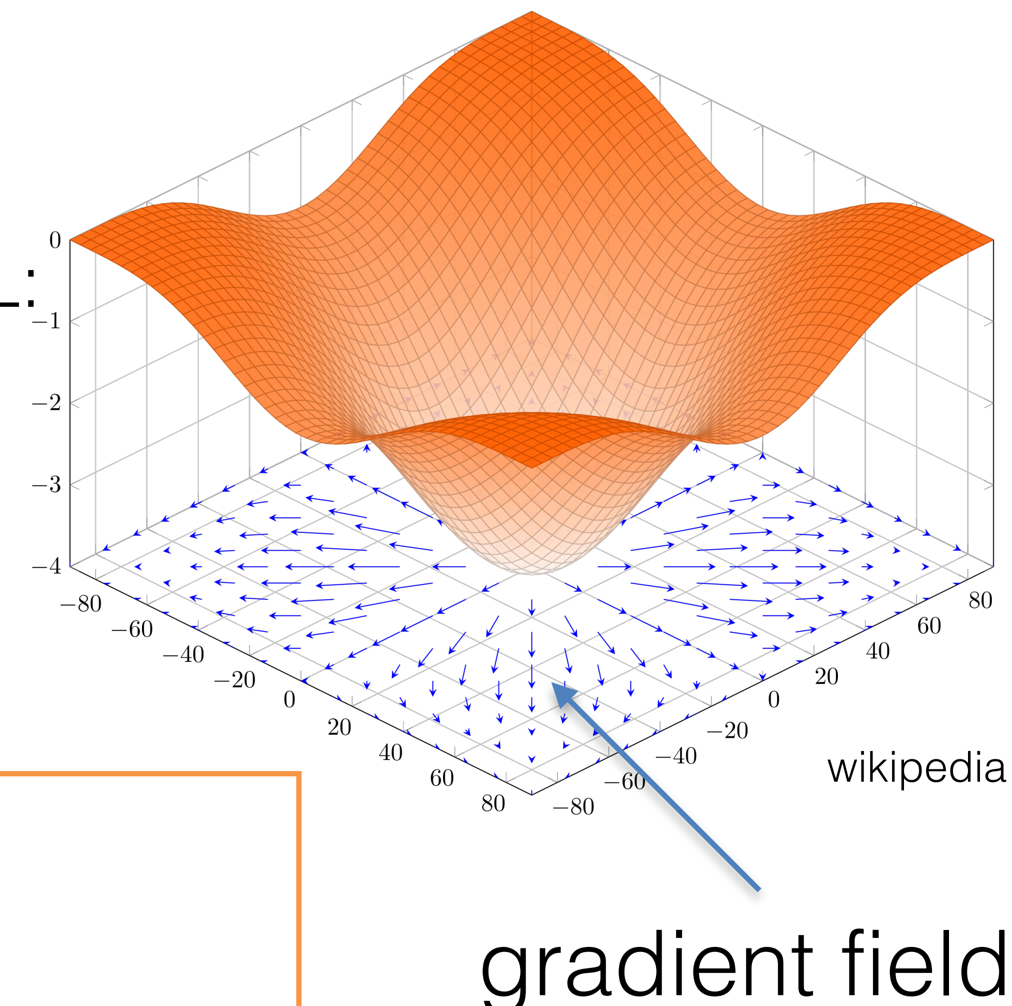
Note that there can be several minima $\mathbf{x}_1^\star \neq \mathbf{x}_2^\star$ with $f(\mathbf{x}_1^\star) = f(\mathbf{x}_2^\star)$.

Iterative Algorithm:

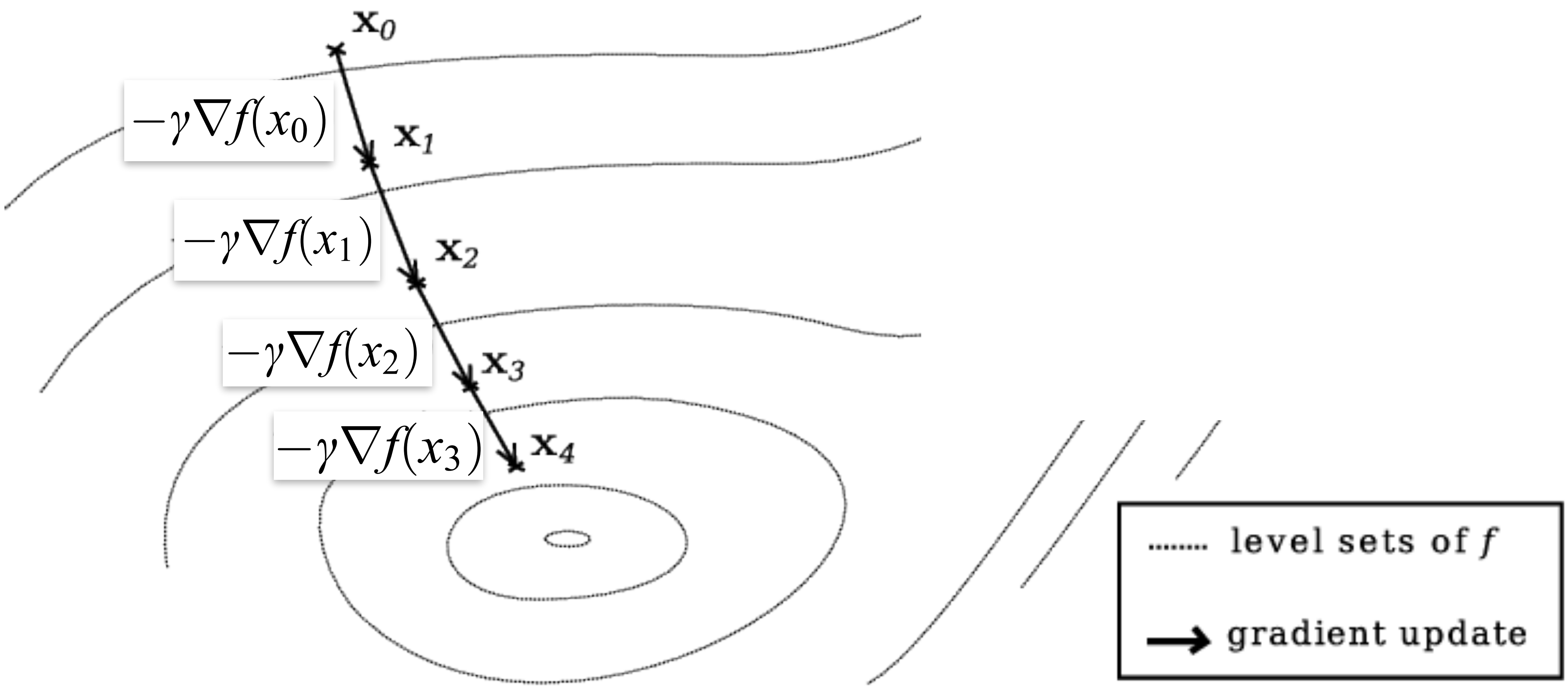
$$\mathbf{x}_{t+1} := \mathbf{x}_t - \gamma \nabla f(\mathbf{x}_t),$$

for **timesteps** $t = 0, 1, \ldots,$ and **stepsize** $\gamma \geq 0$.

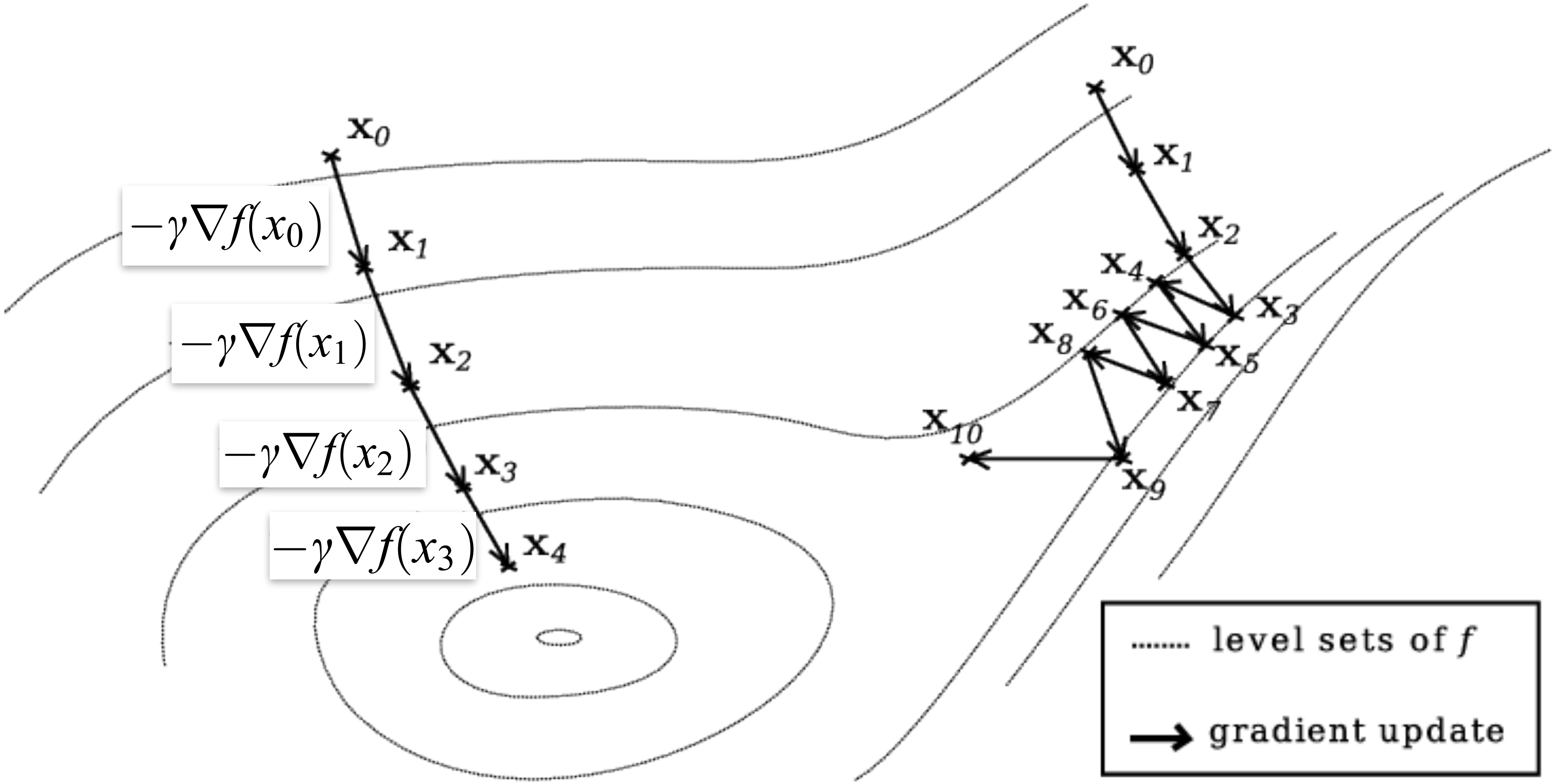
$f(x,y) = -(\cos^2 x + \cos^2 y)^2$



GRADIENT-BASED OPTIMIZATION



GRADIENT-BASED OPTIMIZATION



GRADIENT DESCENT RULES THE WORLD!!!

FLATIRON INSTITUTE
Center for Computational Mathematics

GRADIENT DESCENT RULES THE WORLD!!!

FLATIRON INSTITUTE
Center for Computational Mathematics

- When the function is VERY high-dimensional, only stochastic gradients are computable (see Elad's talk)
- Adaptive gradient descent (ADAGRAD) or Nesterov acceleration is a standard workhorse in large-scale optimization in (online) machine learning
- Stochastic, batch, mini-batch gradient descent (with adaptive step sizes), such as ADAM, is the standard optimizer for Deep NN

GRADIENT DESCENT RULES THE WORLD!!!

- When the function is VERY high-dimensional, only stochastic gradients are computable (see Elad's talk)
- Adaptive gradient descent (ADAGRAD) or Nesterov acceleration is a standard workhorse in large-scale optimization in (online) machine learning
- Stochastic, batch, mini-batch gradient descent (with adaptive step sizes), such as ADAM, is the standard optimizer for Deep NN

Journal of Machine Learning Research 12 (2011) 2121-2159 Submitted 3/10; Revised 3/11; Published 7/11

Adaptive Subgradient Methods for Online Learning and Stochastic Optimization*

John Duchi JDUCHI@CS.BERKELEY.EDU
Computer Science Division
University of California, Berkeley
Berkeley, CA 94720 USA

Elad Hazan EHAZAN@IE.TECHNION.AC.IL
Technion - Israel Institute of Technology
Technion City
Haifa, 32000, Israel

Yoram Singer SINGER@GOOGLE.COM
Google
1600 Amphitheatre Parkway
Mountain View, CA 94043 USA

Published as a conference paper at ICLR 2015

ADAM: A METHOD FOR STOCHASTIC OPTIMIZATION

Diederik P. Kingma* University of Amsterdam, OpenAI dpkingma@openai.com
Jimmy Lei Ba* University of Toronto jimmy@psi.utoronto.ca

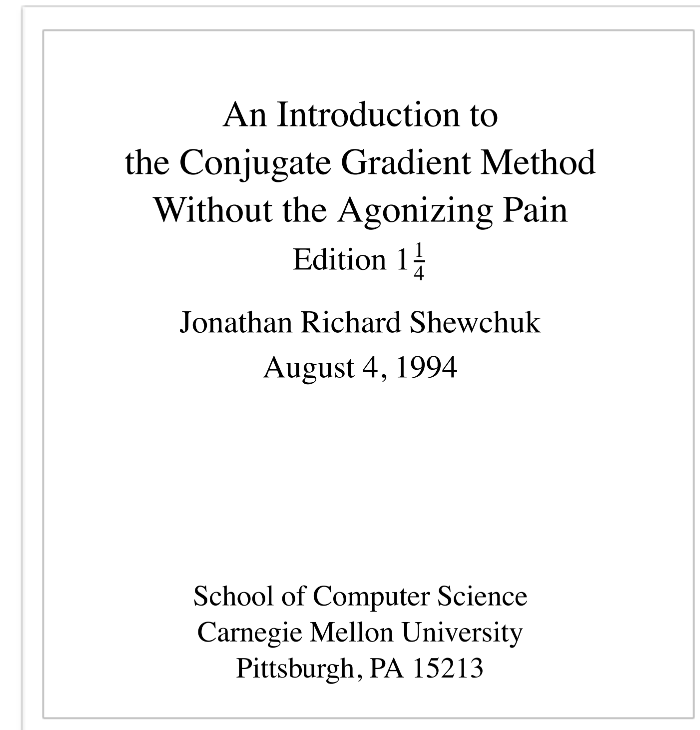
ABSTRACT

We introduce *Adam*, an algorithm for first-order gradient-based optimization of stochastic objective functions, based on adaptive estimates of lower-order moments. The method is straightforward to implement, is computationally efficient, has little memory requirements, is invariant to diagonal rescaling of the gradients, and is well suited for problems that are large in terms of data and/or parameters. The method is also appropriate for non-stationary objectives and problems with very noisy and/or sparse gradients. The hyper-parameters have intuitive interpretations and typically require little tuning. Some connections to related algorithms, on which *Adam* was inspired, are discussed. We also analyze the theoretical convergence properties of the algorithm and provide a regret bound on the convergence rate that is comparable to the best known results under the online convex optimization framework. Empirical results demonstrate that Adam works well in practice and compares favorably to other stochastic optimization methods. Finally, we discuss *AdaMax*, a variant of *Adam* based on the infinity norm.

GRADIENT-BASED OPTIMIZATION

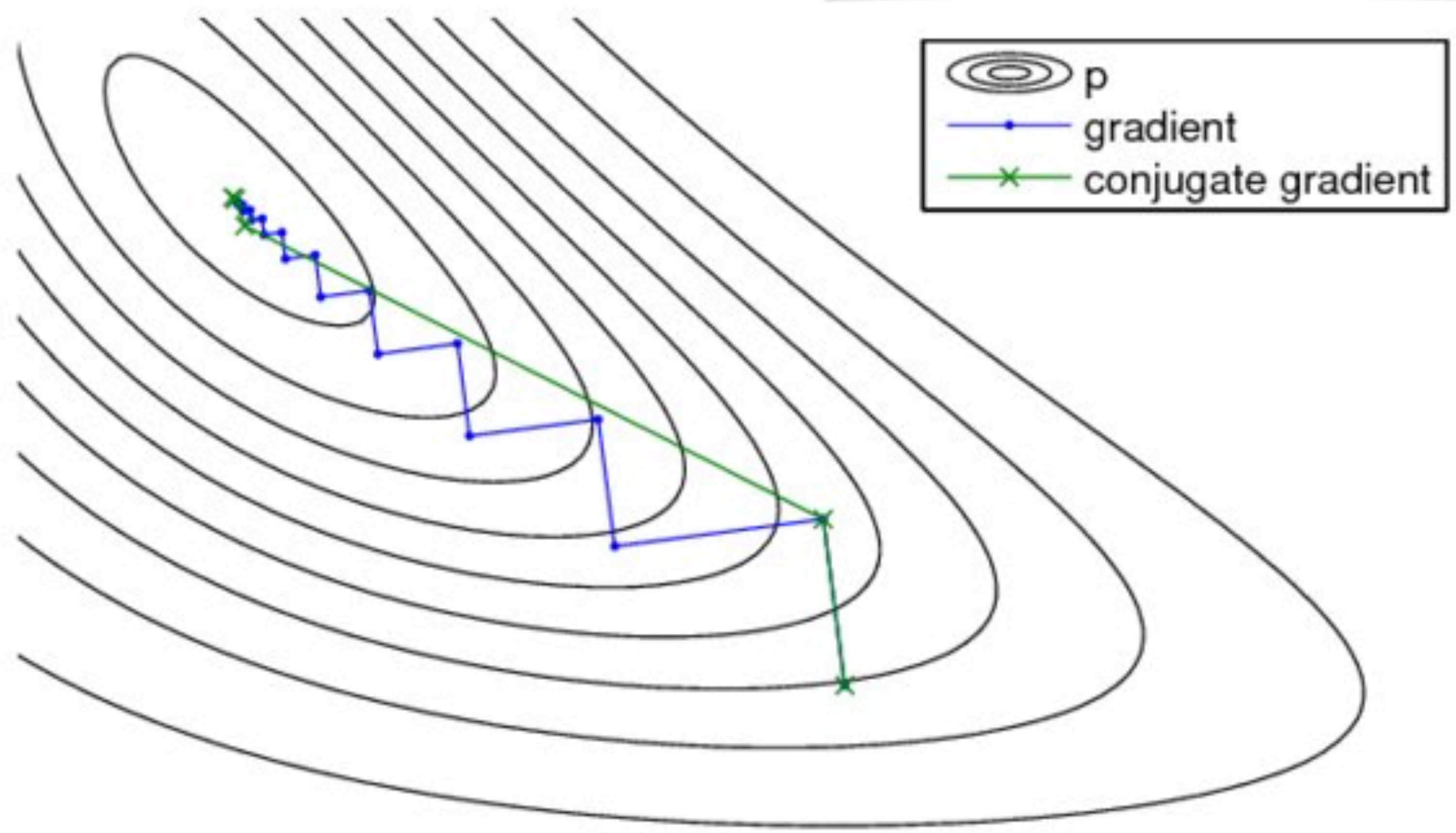
FLATIRON INSTITUTE
Center for Computational Mathematics

- Extension: **Nonlinear conjugate** gradient descent
- Use consecutive gradient directions to generate better search directions (conjugate directions)
- Use line search along the new search directions
- Keywords: Fletcher-Reeves, Polak–Ribière


An Introduction to
the Conjugate Gradient Method
Without the Agonizing Pain
Edition $1\frac{1}{4}$

Jonathan Richard Shewchuk
August 4, 1994

School of Computer Science
Carnegie Mellon University
Pittsburgh, PA 15213



SECOND-ORDER OPTIMIZATION

FLATIRON INSTITUTE
Center for Computational Mathematics

- The gradient and the **Hessian** of the function f is available, i.e. local curvature information
- The function is moderately high-dimensional
- The function is smooth with Lipschitz constant L

SECOND-ORDER OPTIMIZATION

FLATIRON INSTITUTE
Center for Computational Mathematics

- The gradient and the **Hessian** of the function f is available, i.e. local curvature information
- The function is moderately high-dimensional
- The function is smooth with Lipschitz constant L
- Gradient descent:

General update scheme:

$$\mathbf{x}_{t+1} = \mathbf{x}_t - H(\mathbf{x}_t)\nabla f(\mathbf{x}_t),$$

where $H(\mathbf{x}) \in \mathbb{R}^{d\times d}$ is some matrix.

Newton's method: $H = \nabla^2 f(\mathbf{x}_t)^{-1}$.

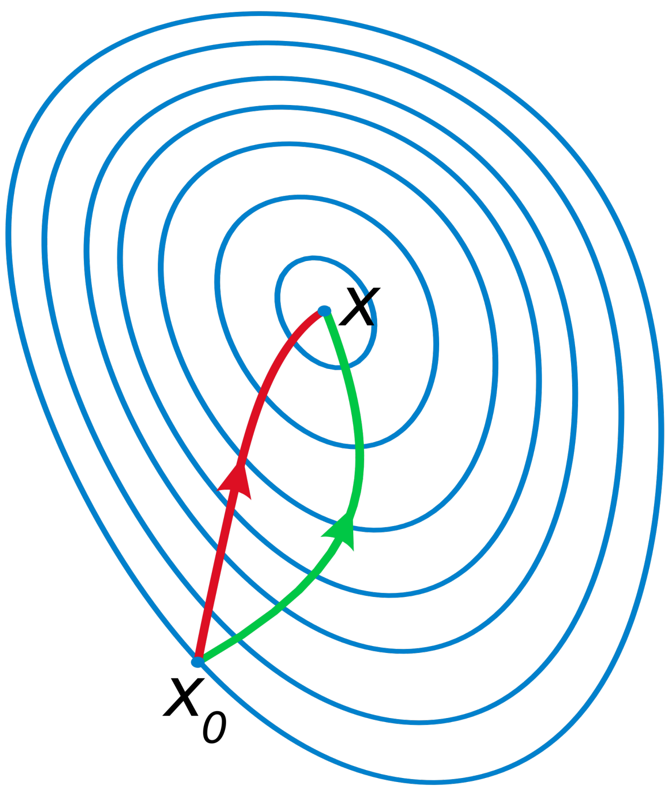
Gradient descent: $H = \gamma I$.

Newton's method: "adaptive gradient descent", adaptation is w.r.t. the local geometry of the function at $\mathbf{x}_t$.

SECOND-ORDER OPTIMIZATION

FLATIRON INSTITUTE
Center for Computational Mathematics

- The gradient and the **Hessian** of the function f is available, i.e. local curvature information
- The function is moderately high-dimensional
- The function is smooth with Lipschitz constant L
- Gradient descent:



General update scheme:

$$\mathbf{x}_{t+1} = \mathbf{x}_t - H(\mathbf{x}_t)\nabla f(\mathbf{x}_t),$$

where $H(\mathbf{x}) \in \mathbb{R}^{d\times d}$ is some matrix.

Newton's method: $H = \nabla^2 f(\mathbf{x}_t)^{-1}$.

Gradient descent: $H = \gamma I$.

Newton's method: "adaptive gradient descent", adaptation is w.r.t. the local geometry of the function at $\mathbf{x}_t$.

SECOND-ORDER OPTIMIZATION AND APPROXIMATIONS

FLATIRON INSTITUTE
Center for Computational Mathematics

- Second-order very useful when the dimension is not too high; otherwise storage of the Hessian becomes prohibitive ($O(n^2)$)
- When the function has many saddle-points, Newton's method needs to be modified
- Variable-metric methods provide an efficient alternative, e.g., BFGS (Broyden, Fletcher, Goldfarb, Shanno) and L-BFGS

SIAM J. OPTIMIZATION
Vol. 1, No. 1, pp. 1–17, February 1991

VARIABLE METRIC METHOD FOR MINIMIZATION*

WILLIAM C. DAVIDON†

Abstract. This is a method for determining numerically local minima of differentiable functions of several variables. In the process of locating each minimum, a matrix which characterizes the behavior of the function about the minimum is determined. For a region in which the function depends quadratically on the variables, no more than N iterations are required, where N is the number of variables. By suitable choice of starting values, and without modification of the procedure, linear constraints can be imposed upon the variables.

Key words. variable metric algorithms, quasi-Newton, optimization

AMS(MOS) subject classifications. primary, 65K10; secondary, 49D37, 65K05, 90C30

PDE-CONSTRAINT OPTIMIZATION

Complicated!

$$
\begin{aligned}
\underset{x}{\text{minimize}} \quad & f(x) \\
\text{subject to} \quad & g_i(x) \le 0, \quad i = 1, \ldots, m \\
& h_j(x) = 0, \quad j = 1, \ldots, p
\end{aligned}
$$

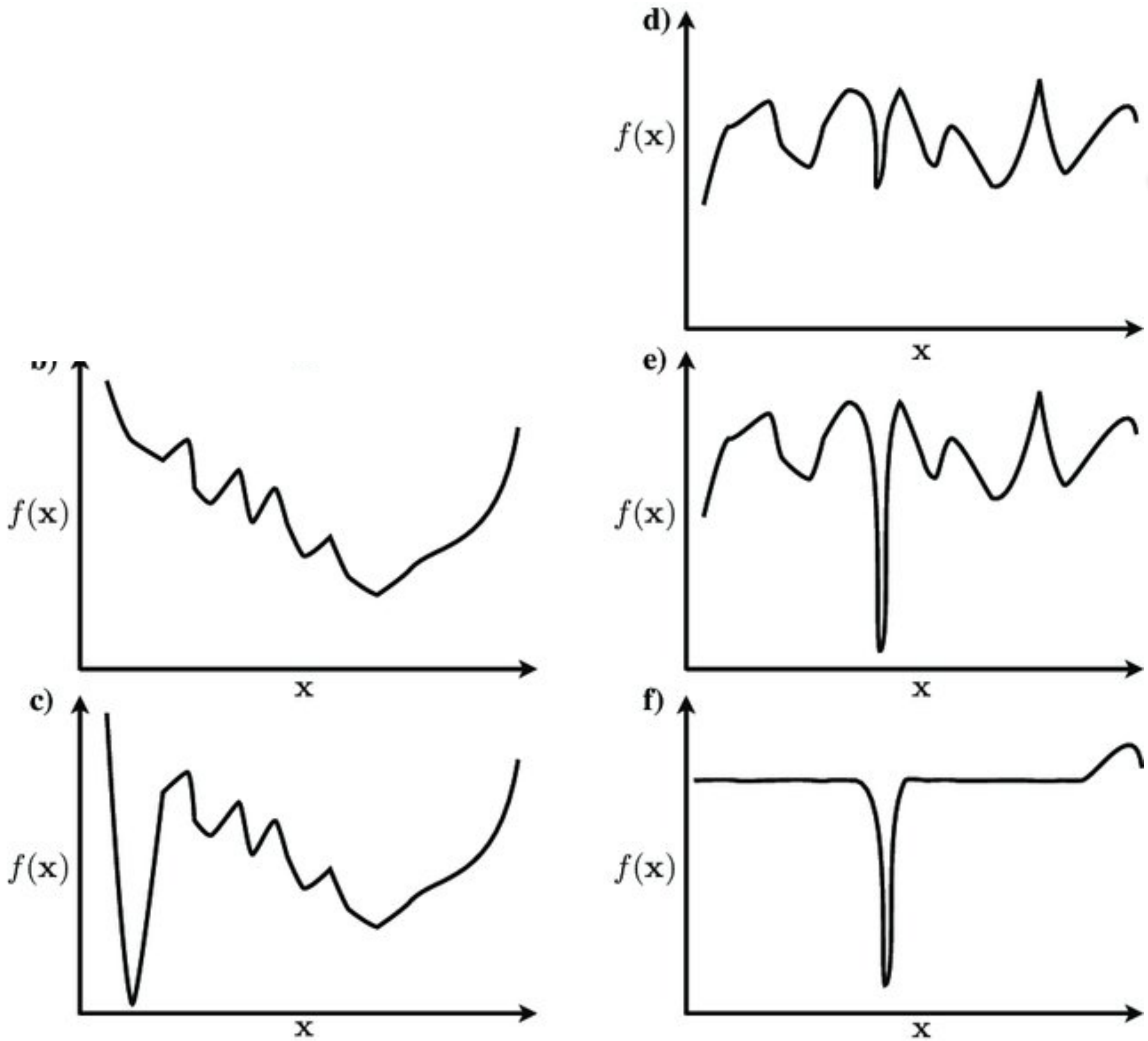
Solution of a (parameterized) partial differential equation!

- Arises in many optimal control problems
- Extremely costly is moderately high-dimensional
- Certain tricks allow efficient optimization (see Leslie's talk)

WHAT ARE GOOD FUNCTIONS?

Hard
but doable?!

Deceiving

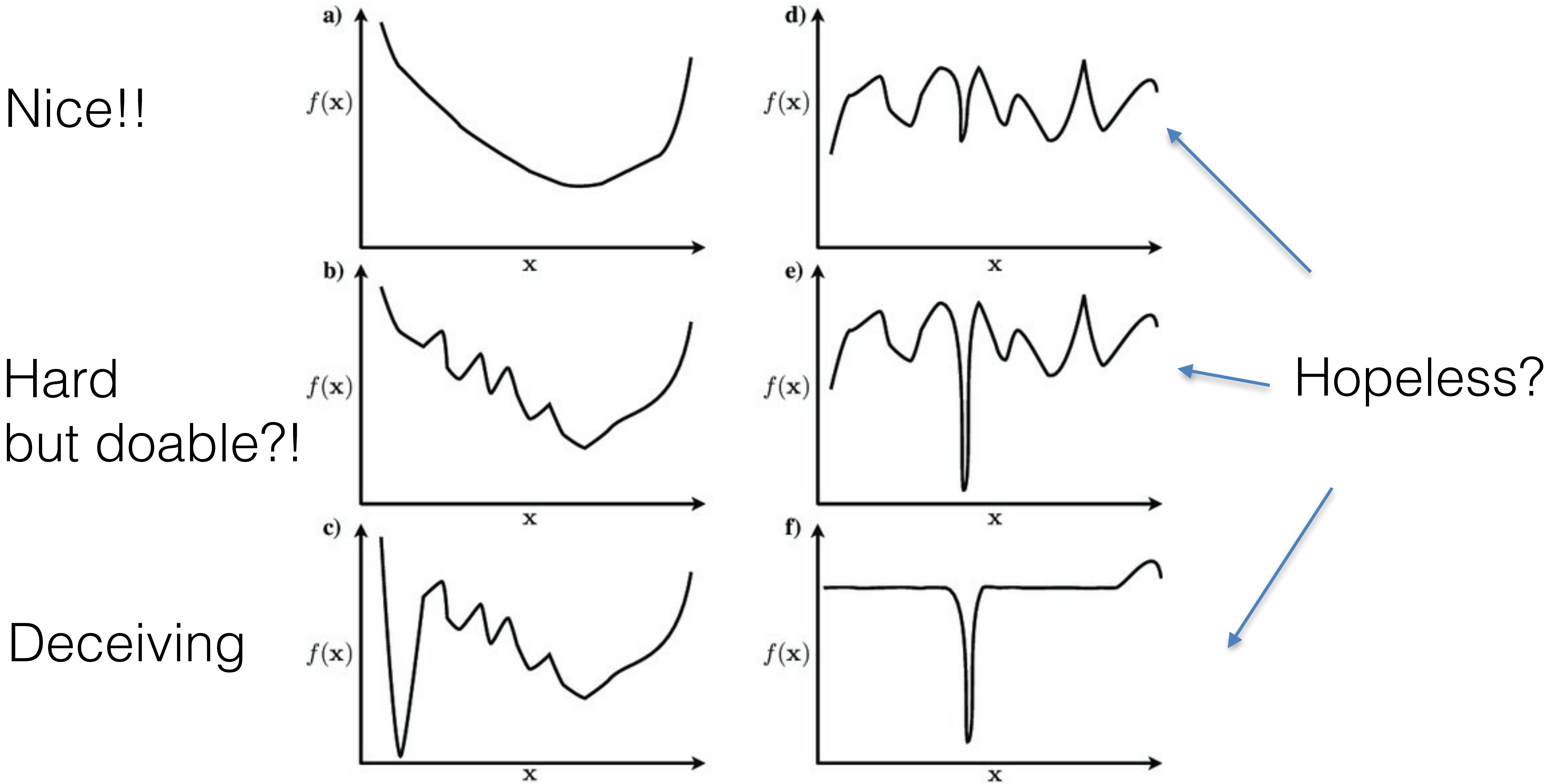


Hopeless?

Stochastic Methods for Single Objective Global Optimization, Christian L. Müller, in: Computational Intelligence in Aerospace Sciences - Fundamental Concepts and Methods (2015) https://doi.org/10.2514/5.9781624102714.0063.0112

WHAT ARE GOOD FUNCTIONS?

FLATIRON INSTITUTE
Center for Computational Mathematics



Stochastic Methods for Single Objective Global Optimization, Christian L. Müller, in: Computational Intelligence in Aerospace Sciences - Fundamental Concepts and Methods (2015) https://doi.org/10.2514/5.9781624102714.0063.0112

WHAT ARE GOOD FUNCTIONS?

FLATIRON INSTITUTE
Center for Computational Mathematics

WHAT ARE GOOD FUNCTIONS?

FLATIRON INSTITUTE
Center for Computational Mathematics

CONVEX FUNCTIONS!

WHAT ARE GOOD FUNCTIONS?

FLATIRON INSTITUTE
Center for Computational Mathematics

CONVEX FUNCTIONS!

"…in fact, the great watershed in optimization isn't between linearity and nonlinearity, but convexity and nonconvexity."
- R. Tyrrell Rockafellar, in SIAM Review, 1993

WHAT ARE GOOD FUNCTIONS?

FLATIRON INSTITUTE
Center for Computational Mathematics

CONVEX FUNCTIONS!

“…in fact, the great watershed in optimization isn't between linearity and nonlinearity, but convexity and nonconvexity.”
- R. Tyrrell Rockafellar, in SIAM Review, 1993

“if it’s not convex, it’s not science”
- attributed to Emmanuel Candes, undated

CONVEX OPTIMIZATION

A convex optimization problem is said to be in the *standard form* if it is written as

$$\begin{aligned} \underset{\mathbf{x}}{\text{minimize}} \quad & f(\mathbf{x}) \\ \text{subject to} \quad & g_i(\mathbf{x}) \leq 0, \quad i = 1, \ldots, m \\ & h_i(\mathbf{x}) = 0, \quad i = 1, \ldots, p, \end{aligned}$$

where $x \in \mathbb{R}^n$ is the optimization variable, the functions $f, g_1, \ldots, g_m$ are convex, and the functions $h_1, \ldots, h_p$ are affine.

CONVEX OPTIMIZATION

FLATIRON INSTITUTE
Center for Computational Mathematics

A convex optimization problem is said to be in the *standard form* if it is written as

$$\begin{aligned} \underset{\mathbf{x}}{\text{minimize}} \quad & f(\mathbf{x}) \\ \text{subject to} \quad & g_i(\mathbf{x}) \leq 0, \quad i = 1, \ldots, m \\ & h_i(\mathbf{x}) = 0, \quad i = 1, \ldots, p, \end{aligned}$$

where $x \in \mathbb{R}^n$ is the optimization variable, the functions $f, g_1, \ldots, g_m$ are convex, and the functions $h_1, \ldots, h_p$ are affine.

Let X be a convex set in a real vector space and let $f : X \to \mathbb{R}$ be a function.

- f is called **convex** if:

$$\forall x_1, x_2 \in X, \forall t \in [0, 1] : \qquad f(tx_1 + (1-t)x_2) \leq tf(x_1) + (1-t)f(x_2)$$

- f is called **strictly convex** if:

$$\forall x_1 \neq x_2 \in X, \forall t \in (0, 1) : \qquad f(tx_1 + (1-t)x_2) < tf(x_1) + (1-t)f(x_2)$$

CONVEX OPTIMIZATION

A convex optimization problem is said to be in the *standard form* if it is written as

$$\begin{aligned} \underset{\mathbf{x}}{\text{minimize}} \quad & f(\mathbf{x}) \\ \text{subject to} \quad & g_i(\mathbf{x}) \leq 0, \quad i = 1, \dots, m \\ & h_i(\mathbf{x}) = 0, \quad i = 1, \dots, p, \end{aligned}$$

where $x \in \mathbb{R}^n$ is the optimization variable, the functions $f, g_1, \dots, g_m$ are convex, and the functions $h_1, \dots, h_p$ are affine.

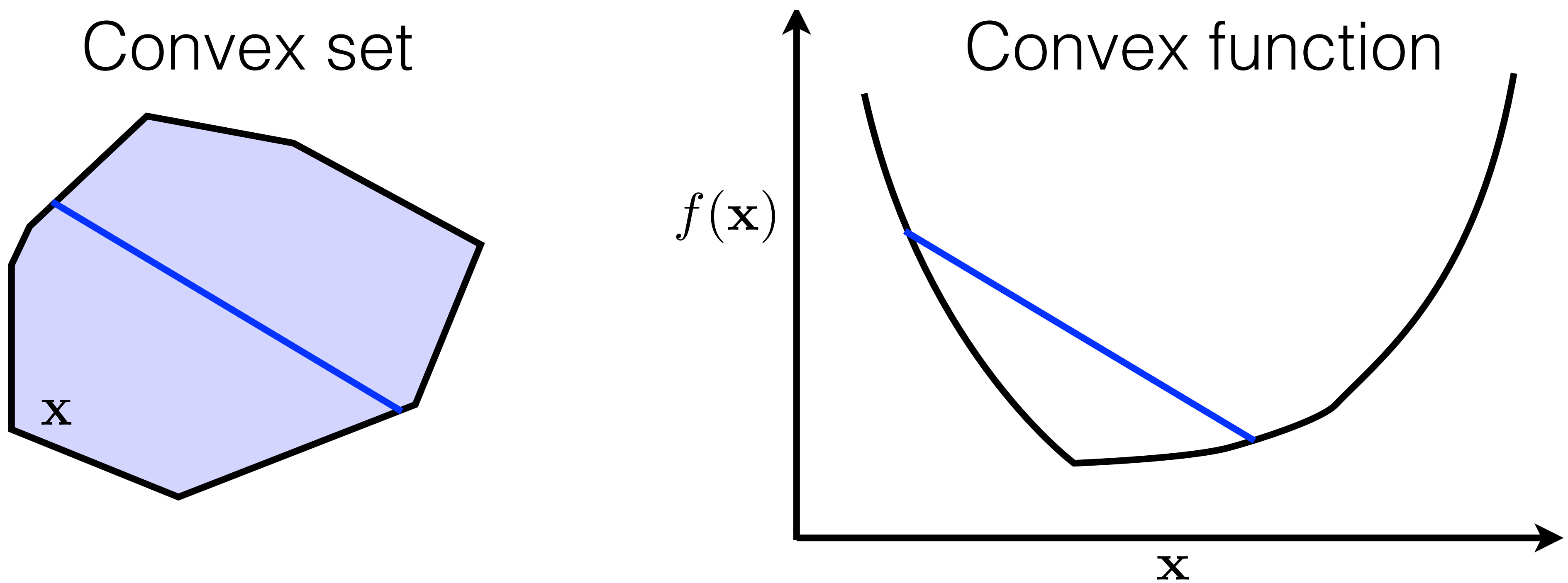
Let X be a convex set in a real vector space and let $f : X \to \mathbb{R}$ be a function.

- f is called **convex** if:

$$\forall x_1, x_2 \in X, \forall t \in [0,1] : \qquad f(tx_1 + (1-t)x_2) \leq tf(x_1) + (1-t)f(x_2)$$

- f is called **strictly convex** if:

$$\forall x_1 \neq x_2 \in X, \forall t \in (0,1) : \qquad f(tx_1 + (1-t)x_2) < tf(x_1) + (1-t)f(x_2)$$



CONVEX OPTIMIZATION

FLATIRON INSTITUTE
Center for Computational Mathematics

A convex optimization problem is said to be in the *standard form* if it is written as

$$\begin{aligned} \underset{\mathbf{x}}{\text{minimize}} \quad & f(\mathbf{x}) \\ \text{subject to} \quad & g_i(\mathbf{x}) \leq 0, \quad i = 1, \ldots, m \\ & h_i(\mathbf{x}) = 0, \quad i = 1, \ldots, p, \end{aligned}$$

where $x \in \mathbb{R}^n$ is the optimization variable, the functions $f, g_1, \ldots, g_m$ are convex, and the functions $h_1, \ldots, h_p$ are affine.

Let X be a convex set in a real vector space and let $f : X \to \mathbb{R}$ be a function.

- f is called **convex** if:

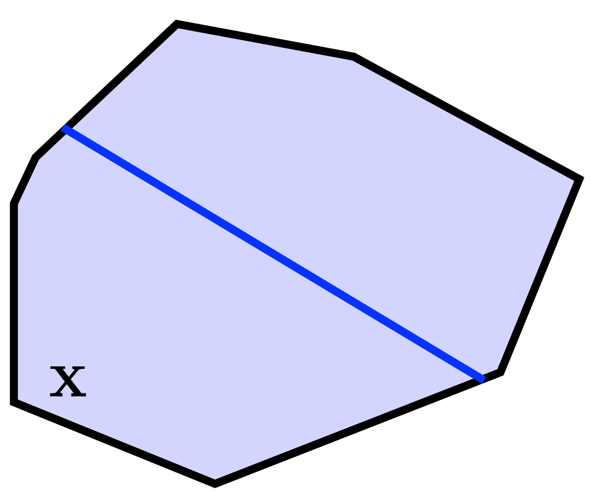
  $\forall x$ …

- f is called **strictly convex** if:

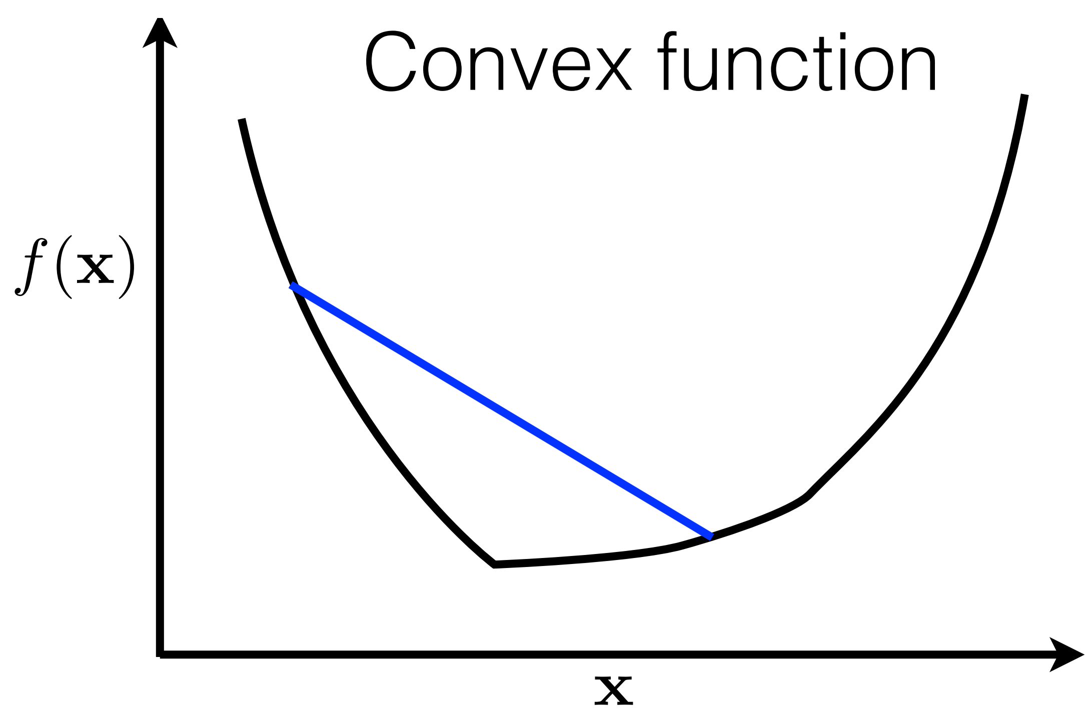
  $\forall x_1 \neq x_2 \in X, \forall t \in (0,1) : \qquad f(tx_1 + (1-t)x_2) < tf(x_1) + (1-t)f(x_2)$

Every local minimum is a global minimum!

Convex set



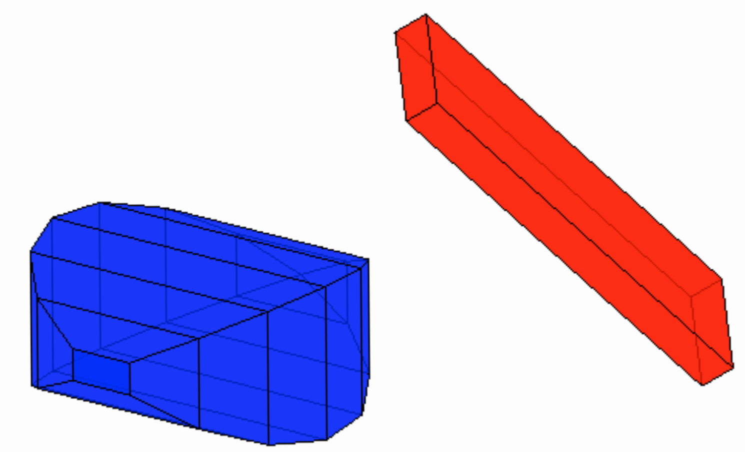
Convex function



CONVEX OPTIMIZATION WITH CONVEX CONSTRAINTS

FLATIRON INSTITUTE
Center for Computational Mathematics

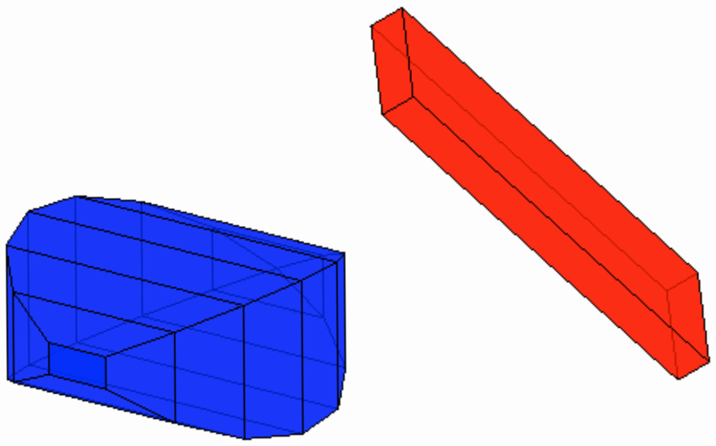
$$\min_{\mathbf{x} \in \mathbb{R}^n} \quad f(\mathbf{x})$$
$$\text{s.t.} \quad \mathbf{A}\mathbf{x} \leq \mathbf{b}\,.$$



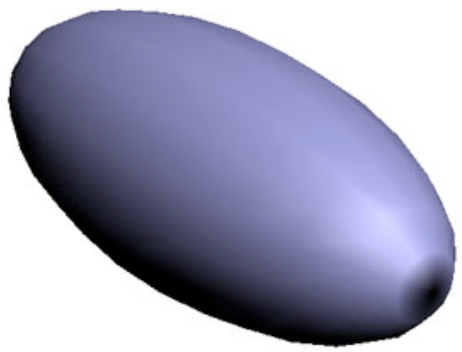
CONVEX OPTIMIZATION WITH CONVEX CONSTRAINTS

$$\min_{\mathbf{x}\in\mathbb{R}^n} \quad f(\mathbf{x})$$

$$\text{s.t.} \quad \mathbf{A}\mathbf{x} \leq \mathbf{b}\,.$$



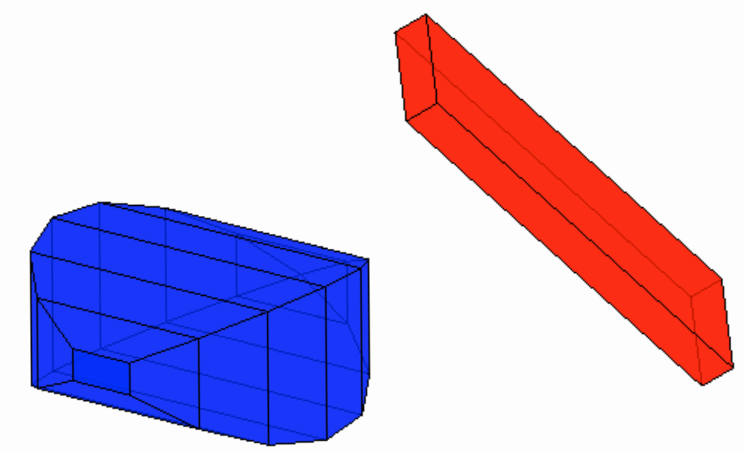
$$\text{s.t.} \quad \mathbf{x}^T\mathbf{A}\mathbf{x} \leq 1\,.$$



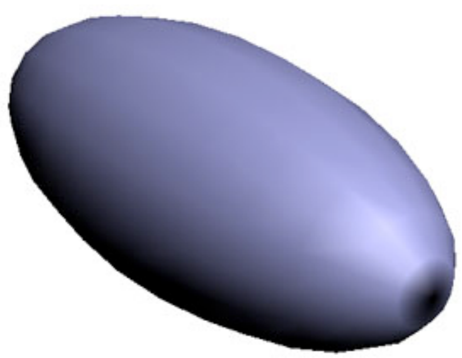
CONVEX OPTIMIZATION WITH CONVEX CONSTRAINTS

$$\min_{\mathbf{x}\in\mathbb{R}^n} \quad f(\mathbf{x})$$

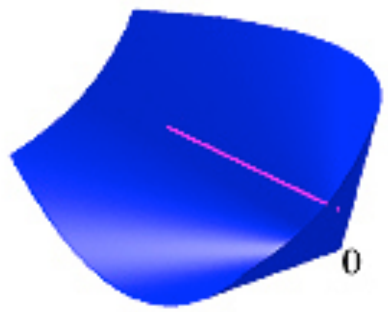
$$\text{s.t.} \quad \mathbf{A}\mathbf{x} \leq \mathbf{b}\,.$$



$$\text{s.t.} \quad \mathbf{x}^T\mathbf{A}\mathbf{x} \leq 1\,.$$

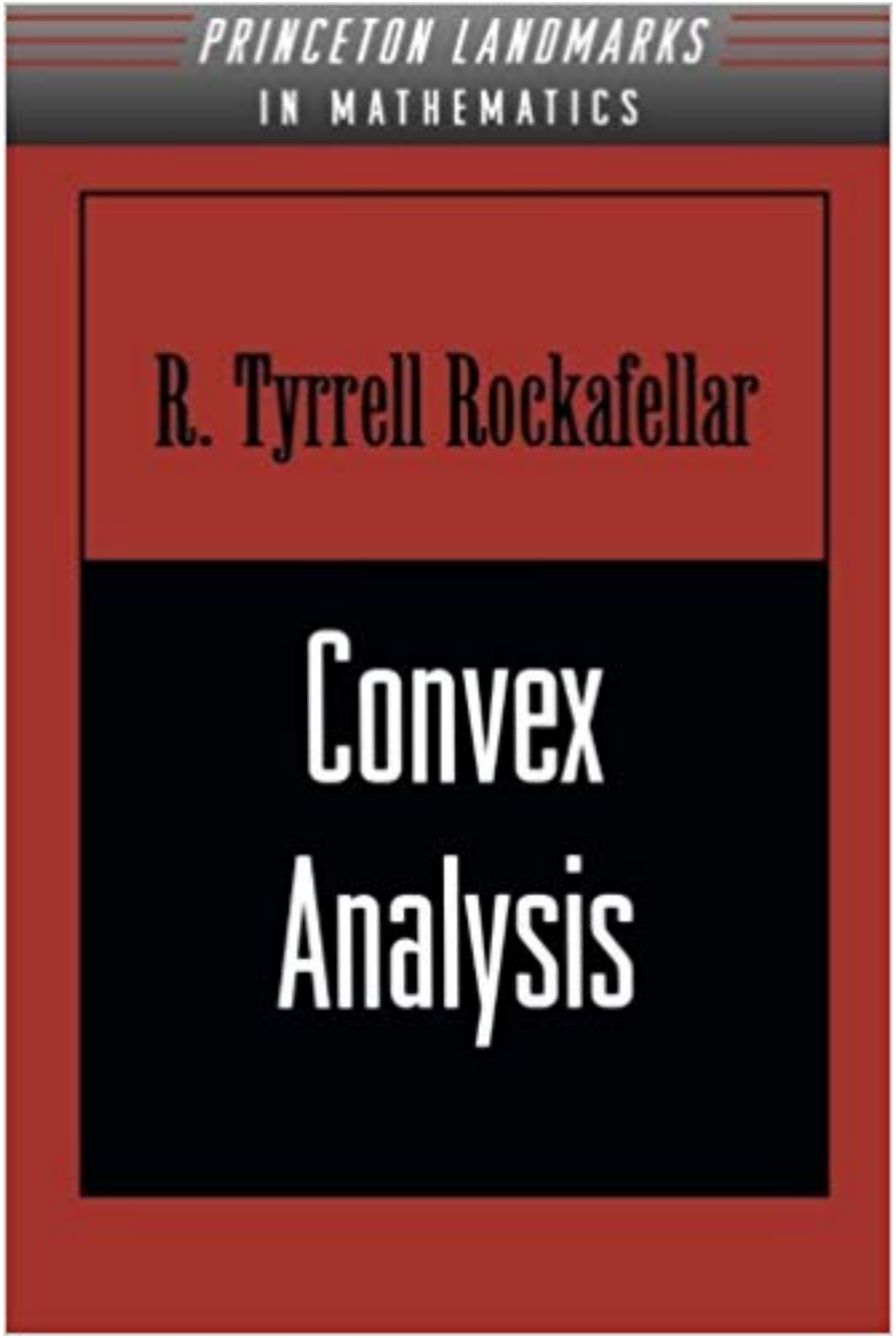


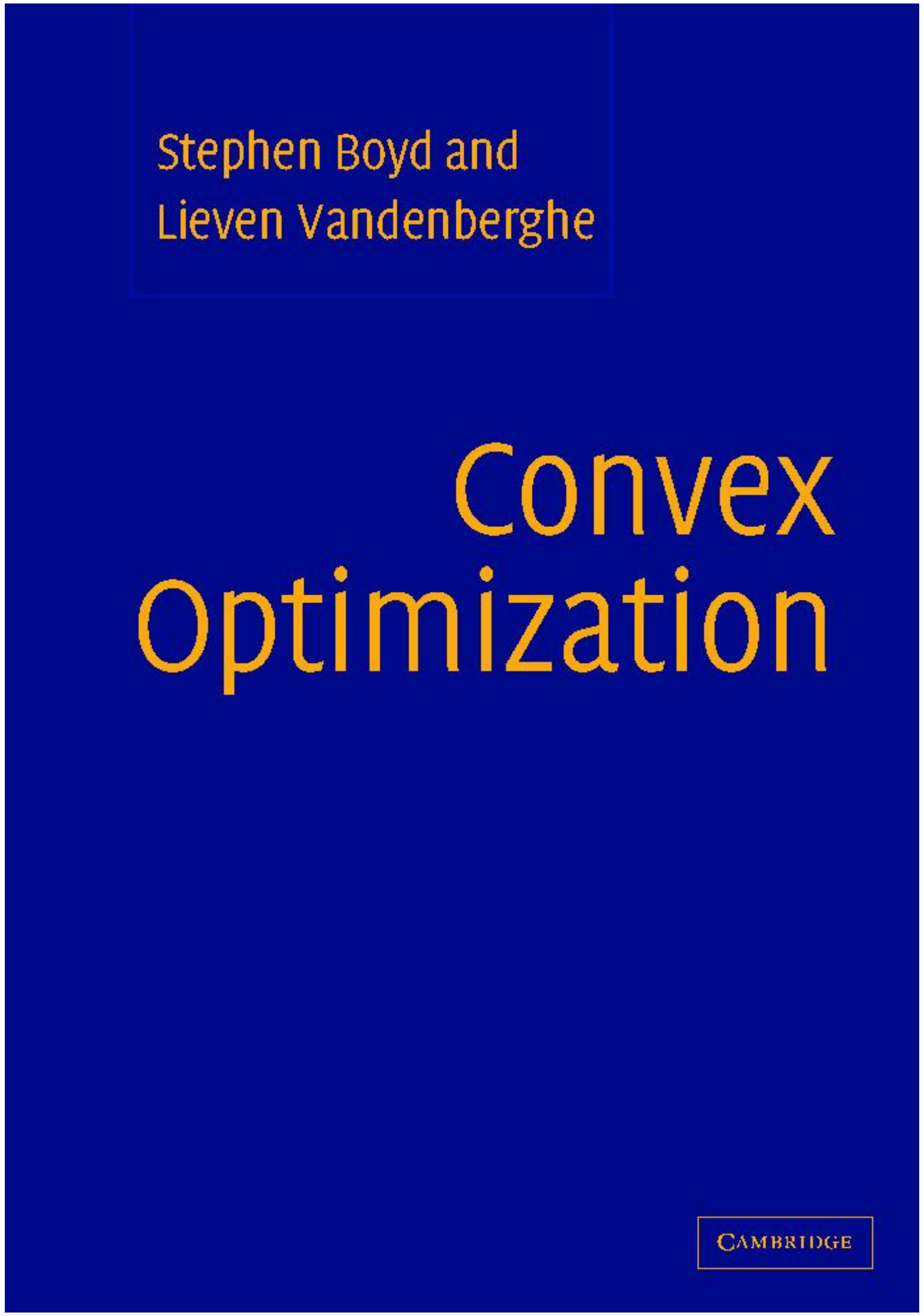
$$\text{s.t.} \quad \mathbf{A}_0 + x_1\mathbf{A}_1 + \ldots + x_n\mathbf{A}_n \preceq 0\,.$$



CONVEX ANALYSIS/MODELING

FLATIRON INSTITUTE
Center for Computational Mathematics

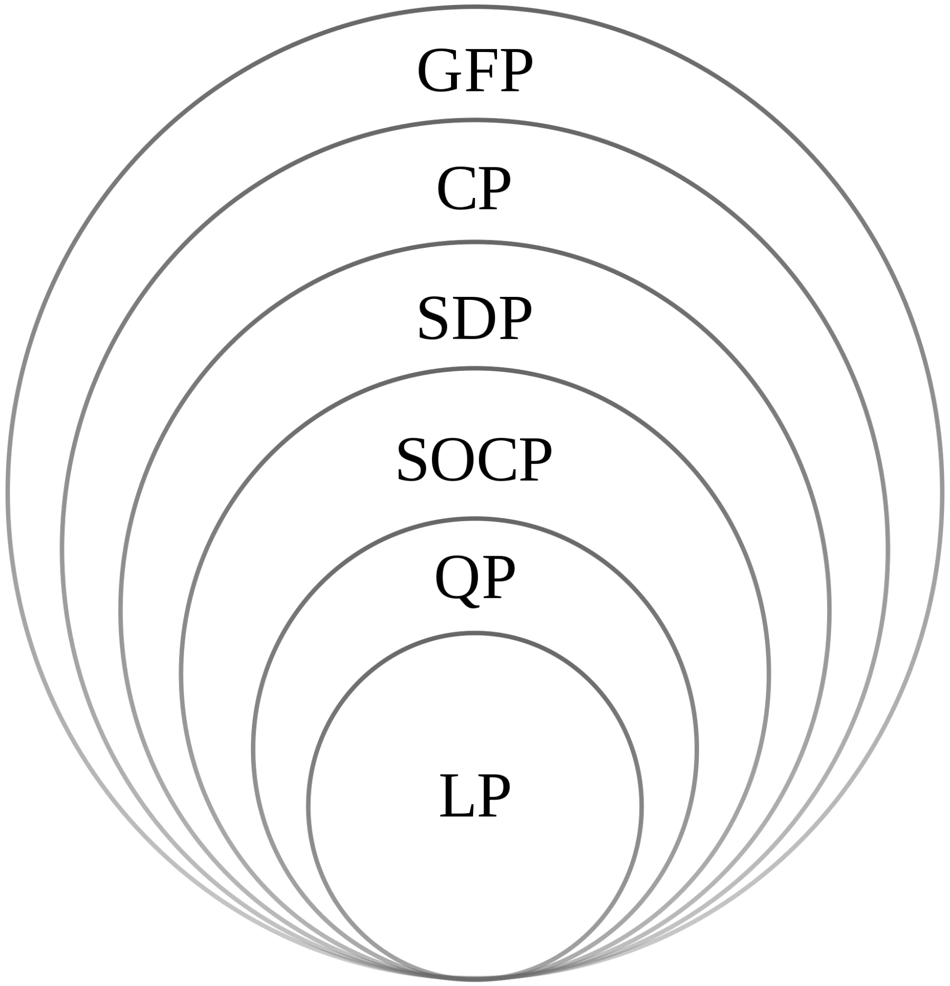




THE HIERARCHY OF CONVEX PROGRAMS

FLATIRON INSTITUTE
Center for Computational Mathematics

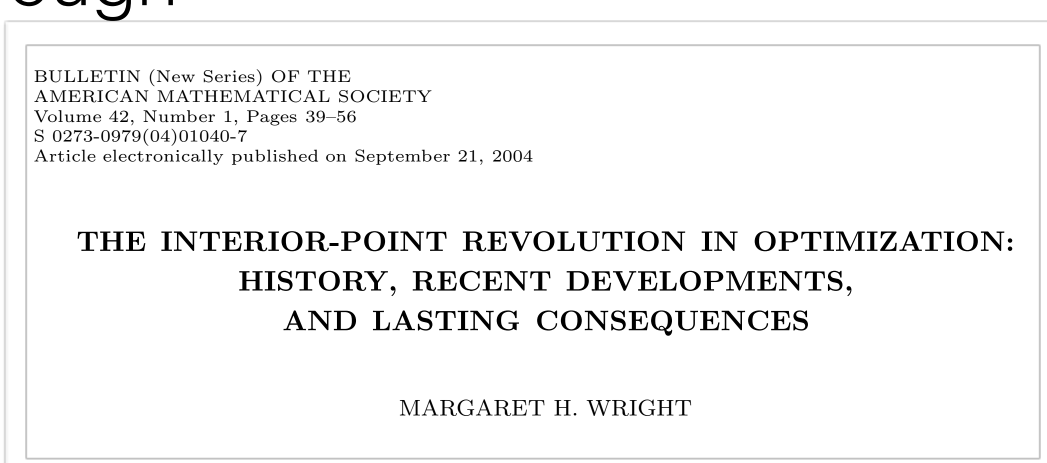
- Each category has a standard form and associated **generic solvers**
- Many engineering problems can be formulated as one of these problems and efficiently solved with theoretical guarantees
- Convergence guarantees and rates can be proven under certain conditions
- Interior-point methods as fundamental breakthrough



LP: linear program
QP: quadratic program
SOCP second-order cone program
SDP: semidefinite program
CP: cone program
GFP: graph form program

THE HIERARCHY OF CONVEX PROGRAMS

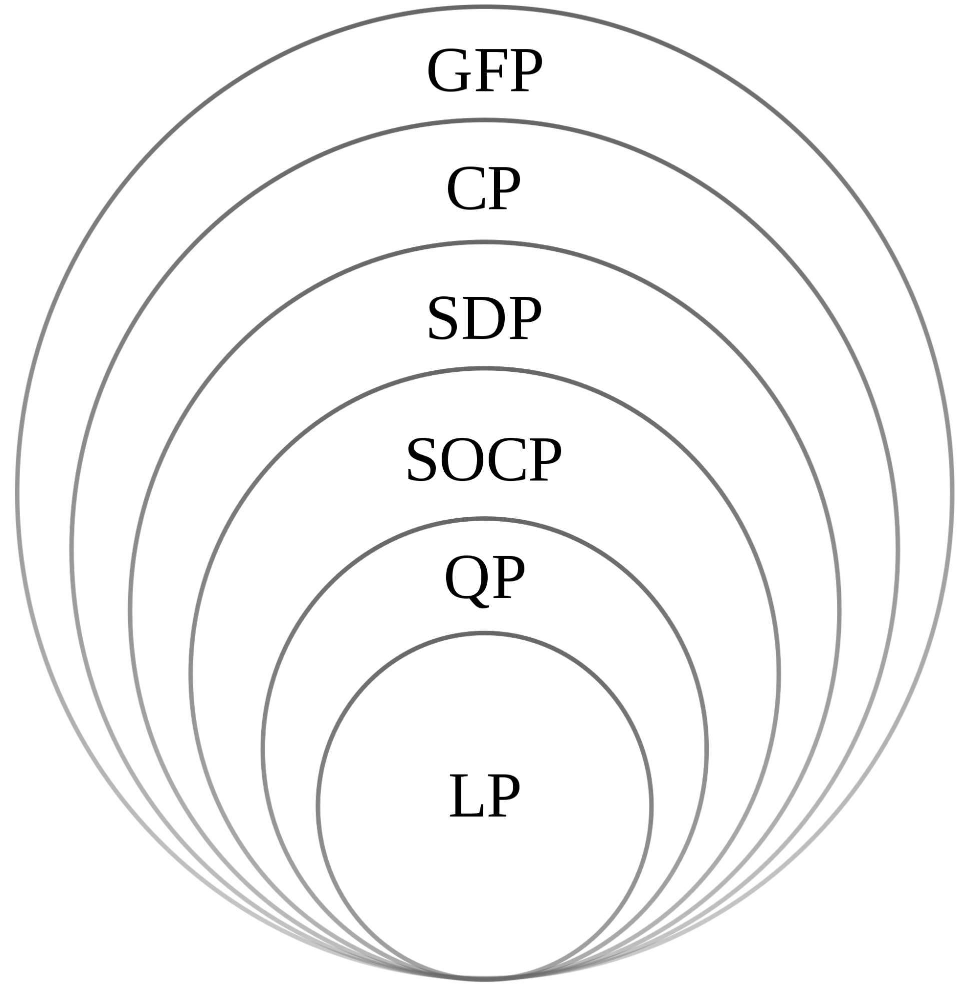
- Each category has a standard form and associated **generic solvers**
- Many engineering problems can be formulated as one of these problems and efficiently solved with theoretical guarantees
- Convergence guarantees and rates can be proven under certain conditions
- Interior-point methods as fundamental breakthrough



BULLETIN (New Series) OF THE
AMERICAN MATHEMATICAL SOCIETY
Volume 42, Number 1, Pages 39–56
S 0273-0979(04)01040-7
Article electronically published on September 21, 2004

THE INTERIOR-POINT REVOLUTION IN OPTIMIZATION: HISTORY, RECENT DEVELOPMENTS, AND LASTING CONSEQUENCES

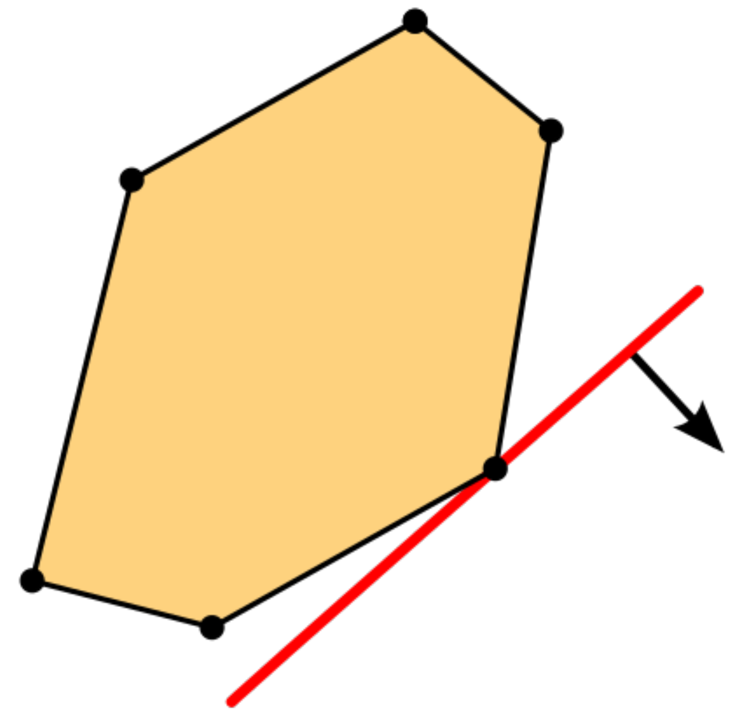
MARGARET H. WRIGHT



LP: linear program
QP: quadratic program
SOCP second-order cone program
SDP: semidefinite program
CP: cone program
GFP: graph form program

LINEAR PROGRAMS

$$\begin{array}{ll} \text{Maximize} & \mathbf{c}^{\mathrm{T}}\mathbf{x} \\ \text{subject to} & A\mathbf{x} \leq \mathbf{b} \\ \text{and} & \mathbf{x} \geq \mathbf{0} \end{array}$$



- Dantzig's simplex algorithm popular in practice (but exponential runtime in worst case)
- Khachiyan's ellipsoidal algorithm and Karmarkar's projective algorithm give polynomial-time guarantees

PROPERTIES OF CONVEX FUNCTIONS AND OPTIMIZATION

FLATIRON INSTITUTE
Center for Computational Mathematics

- Choice, run time, and applicability of different methods depend on the **specific properties** of the convex functions and the constraints
- Keywords: Strongly convex, smooth, non-smooth, constrained, unconstrained,…
- Optimal convergence rates (in function value and iterates) can be proven for many algorithms for specific classes of convex function

RECENT NICE EXAMPLE

FLATIRON INSTITUTE
Center for Computational Mathematics

- Gradient descent with adaptive step size

Revisiting the Polyak step size

Elad Hazan *† Sham M. Kakade *‡

	convex	β-smooth	α-strongly convex	(α, β)-well conditioned
error	$\frac{1}{\sqrt{T}}$	$\frac{\beta}{T}$	$\frac{1}{\alpha T}$	$e^{-\frac{\beta}{\alpha}T}$
step size	$\frac{1}{\sqrt{T}}$	$\frac{1}{\beta}$	$\frac{1}{\alpha T}$	$\frac{1}{\beta}$

RECENT NICE EXAMPLE

Flatiron Institute — Center for Computational Mathematics

- Gradient descent with adaptive step size

Revisiting the Polyak step size

Elad Hazan [*†]   Sham M. Kakade [*‡]

	convex	β-smooth	α-strongly convex	(α, β)-well conditioned
error	$\frac{1}{\sqrt{T}}$	$\frac{\beta}{T}$	$\frac{1}{\alpha T}$	$e^{-\frac{\beta}{\alpha}T}$
step size	$\frac{1}{\sqrt{T}}$	$\frac{1}{\beta}$	$\frac{1}{\alpha T}$	$\frac{1}{\beta}$

Algorithm 1 GD with the Polyak stepsize

1: Input: time horizon T, x_0
2: **for** $t = 0, \ldots, T-1$ **do**
3: Set $\eta_t = \frac{h_t}{\|\nabla_t\|^2}$
4: $\mathbf{x}_{t+1} = \mathbf{x}_t - \eta_t \nabla_t$
5: **end for**
6: Return $\bar{\mathbf{x}} = \operatorname{argmin}_{\mathbf{x}_t} \{f(\mathbf{x}_t)\}$

PROXIMAL ALGORITHMS FOR NON-SMOOTH CONVEX OPTIMIZATION

FLATIRON INSTITUTE
Center for Computational Mathematics

- Many high-dimensional statistics problems are non-smooth convex problems (e.g., Lasso, structured sparsity, …)
- Proximity operator as fundamental building block
- Efficient schemes and exact convergence guarantees

Chapter 10
Proximal Splitting Methods in Signal Processing

Patrick L. Combettes and Jean-Christophe Pesquet

Foundations and Trends® in Optimization
Vol. 1, No. 3 (2013) 123–231

DOI: xxx

now
the essence of knowledge

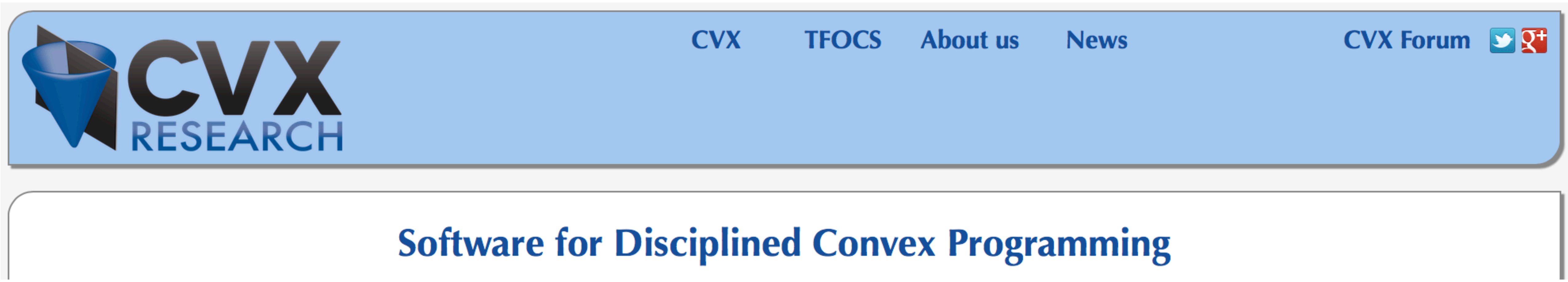
Proximal Algorithms

Neal Parikh
Department of Computer Science
Stanford University
npparikh@cs.stanford.edu

Stephen Boyd
Department of Electrical Engineering
Stanford University
boyd@stanford.edu

NEXT TALK

OPTIMIZATION SOFTWARE

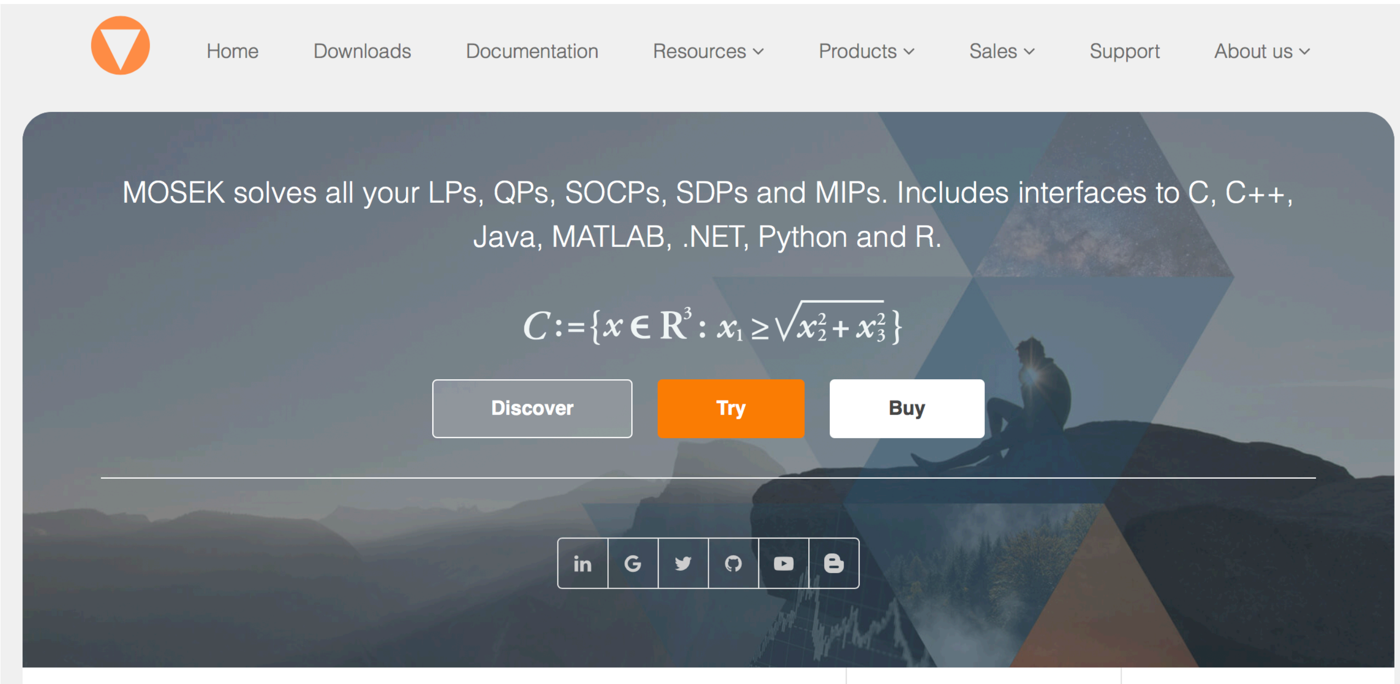


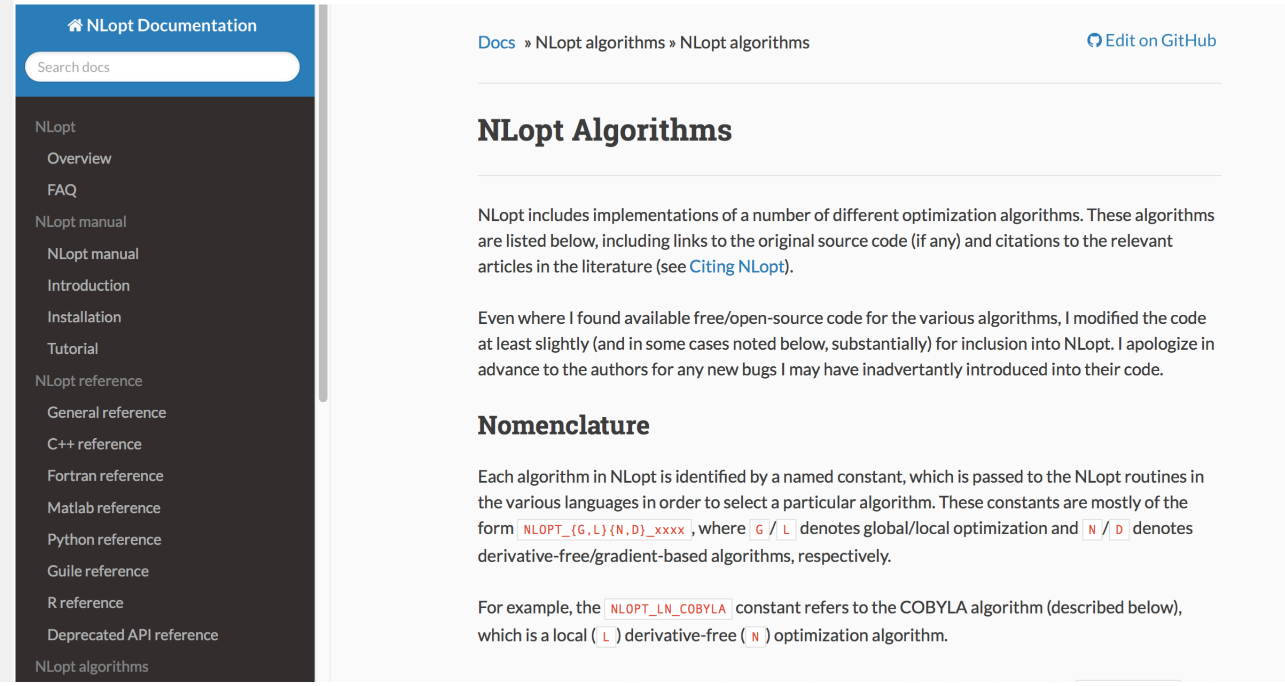
$$\begin{array}{ll} \text{minimize} & \|Ax - b\|_2 \\ \text{subject to} & Cx = d \\ & \|x\|_\infty \leq e \end{array}$$

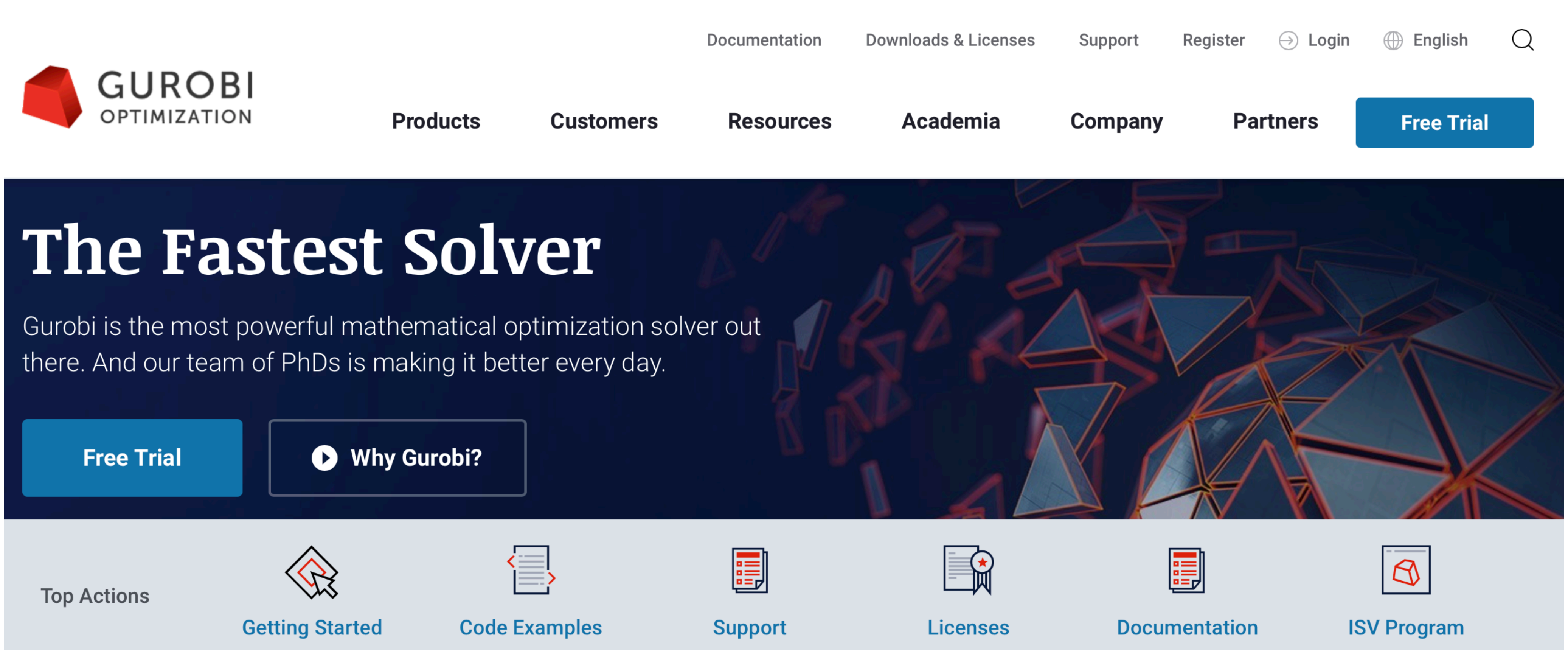
```
m = 20; n = 10; p = 4;
A = randn(m,n); b = randn(m,1);
C = randn(p,n); d = randn(p,1); e = rand;
cvx_begin
    variable x(n)
    minimize( norm( A * x - b, 2 ) )
    subject to
        C * x == d
        norm( x, Inf ) <= e
cvx_end
```

OPTIMIZATION SOFTWARE

FLATIRON INSTITUTE
Center for Computational Mathematics







OUTLOOK - DNN LANDSCAPES

FLATIRON INSTITUTE
Center for Computational Mathematics

Towards Understanding Generalization of Deep Learning: Perspective of Loss Landscapes

Lei Wu
School of Mathematics, Peking University
Beijing, China
leiwu@pku.edu.cn

Zhanxing Zhu
Beijing Institute of Big Data Research (BIBDR)
Center for Data Science, Peking University
Beijing, China
zhanxing.zhu@pku.edu.cn

Weinan E
Beijing Institute of Big Data Research (BIBDR)
Center for Data Science and BICMR, Peking University, Beijing, China
Department of Mathematics and PACM, Princeton University, Princeton, NJ, USA
weinan@math.princeton.edu

OUTLOOK - DNN LANDSCAPES

Towards Understanding Generalization of Deep Learning: Perspective of Loss Landscapes

Lei Wu
School of Mathematics, Peking University
Beijing, China
leiwu@pku.edu.cn

Zhanxing Zhu
Beijing Institute of Big Data Research (BIBDR)
Center for Data Science, Peking University
Beijing, China
zhanxing.zhu@pku.edu.cn

Weinan E
Beijing Institute of Big Data Research (BIBDR)
Center for Data Science and BICMR, Peking University, Beijing, China
Department of Mathematics and PACM, Princeton University, Princeton, NJ, USA
weinan@math.princeton.edu

Large Scale Structure of Neural Network Loss Landscapes

Stanislav Fort*
Google Research
Zurich, Switzerland
stanislavfort@google.com

Stanislaw Jastrzebski†
New York University
New York, United States

Abstract

There are many surprising and perhaps counter-intuitive properties of optimization of deep neural networks. We propose and experimentally verify a unified phenomenological model of the loss landscape that incorporates many of them. High dimensionality plays a key role in our model. Our core idea is to model the loss landscape as a set of high dimensional *wedges* that together form a large-scale, inter-connected structure and towards which optimization is drawn. We first show that hyperparameter choices such as learning rate, network width and L_2 regularization, affect the path optimizer takes through the landscape in a similar ways, influencing the large scale curvature of the regions the optimizer explores. Finally, we predict and demonstrate new counter-intuitive properties of the loss-landscape. We show an existence of low loss subspaces connecting a set (not only a pair) of solutions, and verify it experimentally. Finally, we analyze recently popular ensembling techniques for deep networks in the light of our model.

OUTLOOK - DNN LANDSCAPES

FLATIRON INSTITUTE
Center for Computational Mathematics

Towards Understanding Generalization of Deep Learning: Perspective of Loss Landscapes

Lei Wu
School of Mathematics, Peking University
Beijing, China
leiwu@pku.edu.cn

Zhanxing Zhu
Beijing Institute of Big Data Research (BIBDR)
Center for Data Science, Peking University
Beijing, China
zhanxing.zhu@pku.edu.cn

Weinan E
Beijing Institute of Big Data Research (BIBDR)
Center for Data Science and BICMR, Peking University, Beijing, China
Department of Mathematics and PACM, Princeton University, Princeton, NJ, USA
weinan@math.princeton.edu

Visualizing the Loss Landscape of Neural Nets

Hao Li[1], Zheng Xu[1], Gavin Taylor[2], Christoph Studer[3], Tom Goldstein[1]
[1]University of Maryland, College Park [2]United States Naval Academy [3]Cornell University
{haoli,xuzh,tomg}@cs.umd.edu, taylor@usna.edu, studer@cornell.edu

Abstract

Neural network training relies on our ability to find "good" minimizers of highly non-convex loss functions. It is well-known that certain network architecture designs (e.g., skip connections) produce loss functions that train easier, and well-chosen training parameters (batch size, learning rate, optimizer) produce minimizers that generalize better. However, the reasons for these differences, and their effect on the underlying loss landscape, are not well understood. In this paper, we explore the structure of neural loss functions, and the effect of loss landscapes on generalization, using a range of visualization methods. First, we introduce a simple "filter normalization" method that helps us visualize loss function curvature and make meaningful side-by-side comparisons between loss functions. Then, using a variety of visualizations, we explore how network architecture affects the loss landscape, and how training parameters affect the shape of minimizers.

Large Scale Structure of Neural Network Loss Landscapes

Stanislav Fort*
Google Research
Zurich, Switzerland
stanislavfort@google.com

Stanislaw Jastrzebski†
New York University
New York, United States

Abstract

There are many surprising and perhaps counter-intuitive properties of optimization of deep neural networks. We propose and experimentally verify a unified phenomenological model of the loss landscape that incorporates many of them. High dimensionality plays a key role in our model. Our core idea is to model the loss landscape as a set of high dimensional *wedges* that together form a large-scale, inter-connected structure and towards which optimization is drawn. We first show that hyperparameter choices such as learning rate, network width and L_2 regularization, affect the path optimizer takes through the landscape in a similar ways, influencing the large scale curvature of the regions the optimizer explores. Finally, we predict and demonstrate new counter-intuitive properties of the loss-landscape. We show an existence of low loss subspaces connecting a set (not only a pair) of solutions, and verify it experimentally. Finally, we analyze recently popular ensembling techniques for deep networks in the light of our model.

OUTLOOK - DNN LANDSCAPES

FLATIRON INSTITUTE
Center for Computational Mathematics

Towards Understanding Generalization of Deep Learning: Perspective of Loss Landscapes

Lei Wu
School of Mathematics, Peking University
Beijing, China
leiwu@pku.edu.cn

Zhanxing Zhu
Beijing Institute of Big Data Research (BIBDR)
Center for Data Science, Peking University
Beijing, China
zhanxing.zhu@pku.edu.cn

Weinan E
Beijing Institute of Big Data Research (BIBDR)
Center for Data Science and BICMR, Peking University, Beijing, China
Department of Mathematics and PACM, Princeton University, Princeton, NJ, USA
weinan@math.princeton.edu

Visualizing the Loss Landscape of Neural Nets

Hao Li[1], Zheng Xu[1], Gavin Taylor[2], Christoph Studer[3], Tom Goldstein[1]
[1]University of Maryland, College Park [2]United States Naval Academy [3]Cornell University
{haoli,xuzh,tomg}@cs.umd.edu, taylor@usna.edu, studer@cornell.edu

Abstract

Neural network training relies on our ability to find "good" minimizers of highly non-convex loss functions. It is well-known that certain network architecture designs (e.g., skip connections) produce loss functions that train easier, and well-chosen training parameters (batch size, learning rate, optimizer) produce minimizers that generalize better. However, the reasons for these differences, and their effect on the underlying loss landscape, are not well understood. In this paper, we explore the structure of neural loss functions, and the effect of loss landscapes on generalization, using a range of visualization methods. First, we introduce a simple "filter normalization" method that helps us visualize loss function curvature and make meaningful side-by-side comparisons between loss functions. Then, using a variety of visualizations, we explore how network architecture affects the loss landscape, and how training parameters affect the shape of minimizers.

Large Scale Structure of Neural Network Loss Landscapes

Stanislav Fort*
Google Research
Zurich, Switzerland
stanislavfort@google.com

Stanislaw Jastrzebski†
New York University
New York, United States

Abstract

There are many surprising and perhaps counter-intuitive properties of optimization of deep neural networks. We propose and experimentally verify a unified phenomenological model of the loss landscape that incorporates many of them. High dimensionality plays a key role in our model. Our core idea is to model the loss landscape as a set of high dimensional *wedges* that together form a large-scale, inter-connected structure and towards which optimization is drawn. We first show that hyperparameter choices such as learning rate, network width and L_2 regularization, affect the path optimizer takes through the landscape in a similar ways, influencing the large scale curvature of the regions the optimizer explores. Finally, we predict and demonstrate new counter-intuitive properties of the loss-landscape. We show an existence of low loss subspaces connecting a set (not only a pair) of solutions, and verify it experimentally. Finally, we analyze recently popular ensembling techniques for deep networks in the light of our model.

Journal of Machine Learning Research 20 (2019) 1-34 Submitted 9/18; Revised 7/19; Published 8/19

Spurious Valleys in One-hidden-layer Neural Network Optimization Landscapes

Luca Venturi VENTURI@CIMS.NYU.EDU
Courant Institute of Mathematical Sciences, 251 Mercer Street, New York, NY 10012

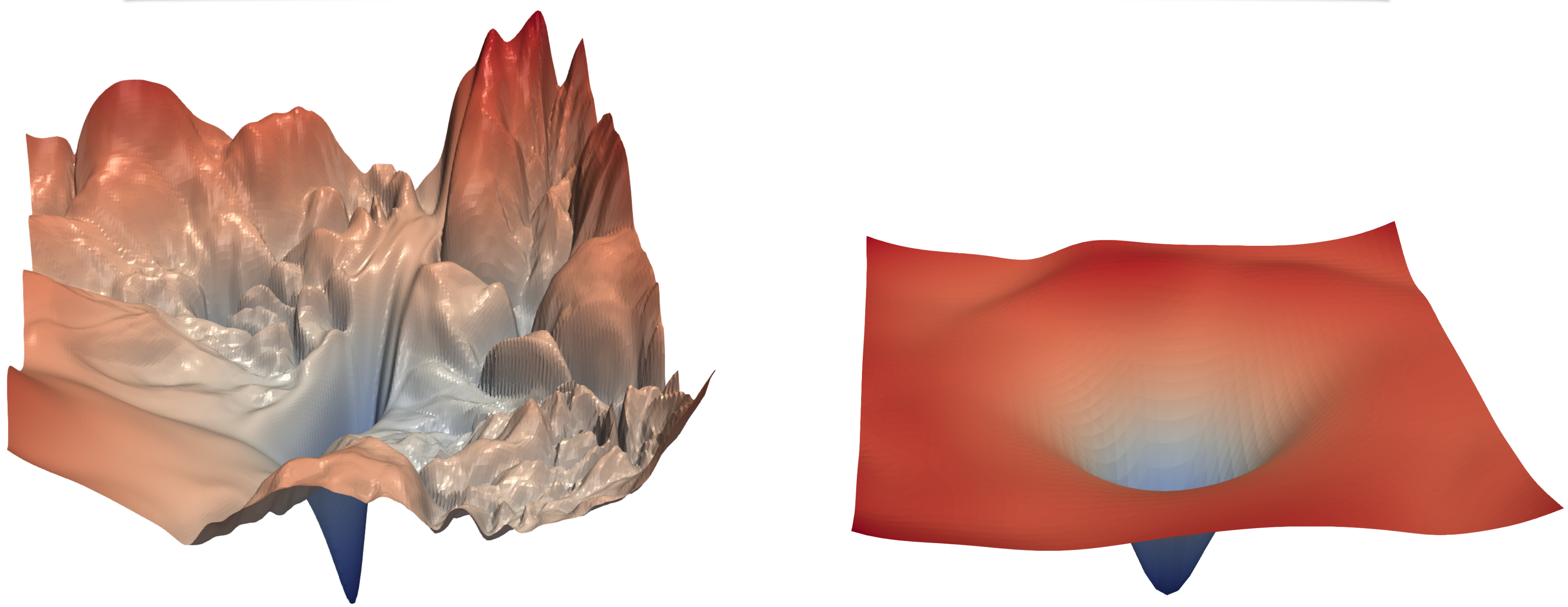
Afonso S. Bandeira BANDEIRA@CIMS.NYU.EDU
Joan Bruna BRUNA@CIMS.NYU.EDU
Courant Institute of Mathematical Sciences, 251 Mercer Street, New York, NY 10012
Center for Data Science, 60 5th Avenue, New York, NY 10011

Editor: Animashree Anandkumar

OUTLOOK

Visualizing the Loss Landscape of Neural Nets

Hao Li[1], Zheng Xu[1], Gavin Taylor[2], Christoph Studer[3], Tom Goldstein[1]
[1]University of Maryland, College Park [2]United States Naval Academy [3]Cornell University
{haoli,xuzh,tomg}@cs.umd.edu, taylor@usna.edu, studer@cornell.edu

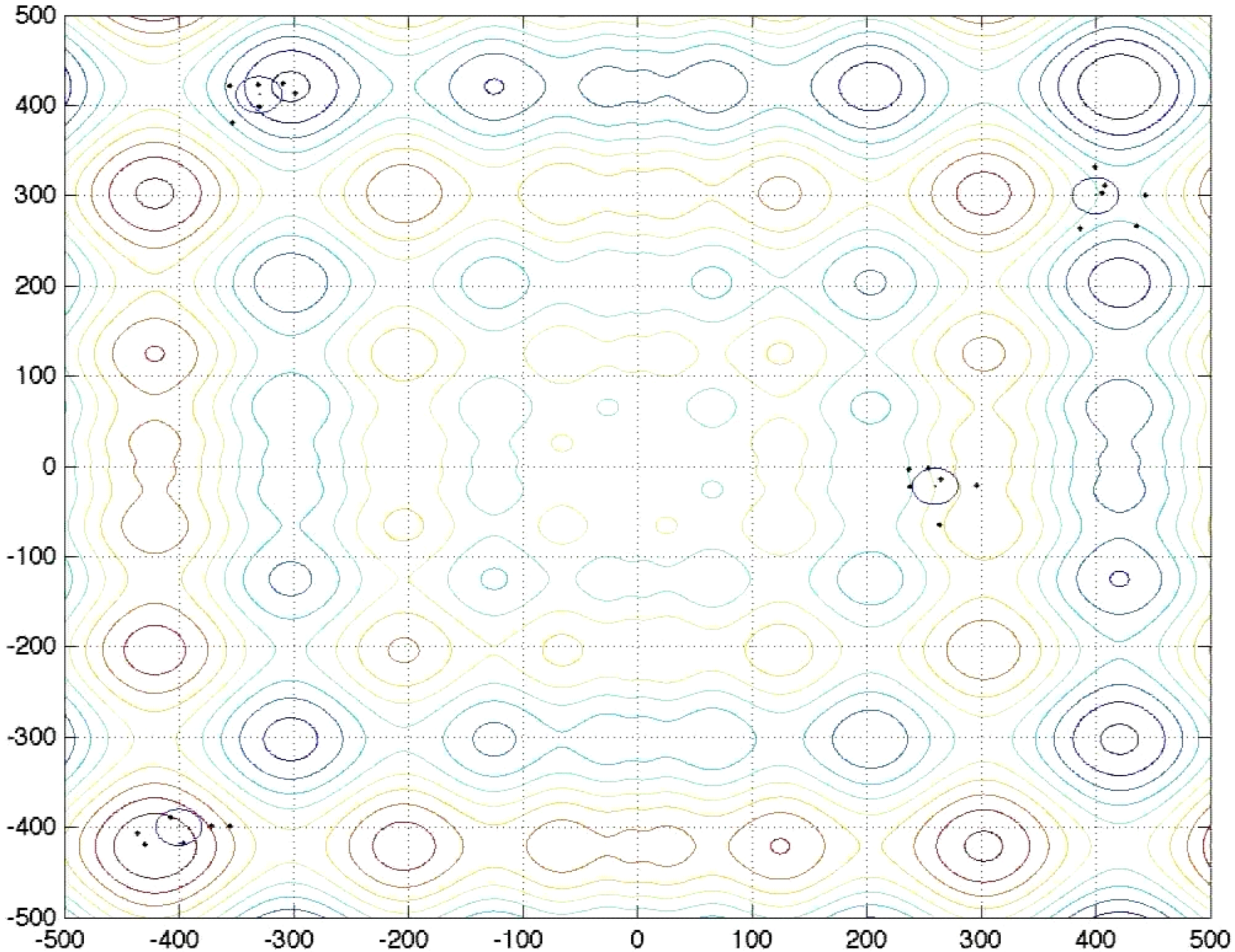


(a) without skip connections (b) with skip connections

Figure 1: The loss surfaces of ResNet-56 with/without skip connections. The proposed filter normalization scheme is used to enable comparisons of sharpness/flatness between the two figures.

Thank you for your time! Questions?

FLATIRON INSTITUTE
Center for Computational Mathematics

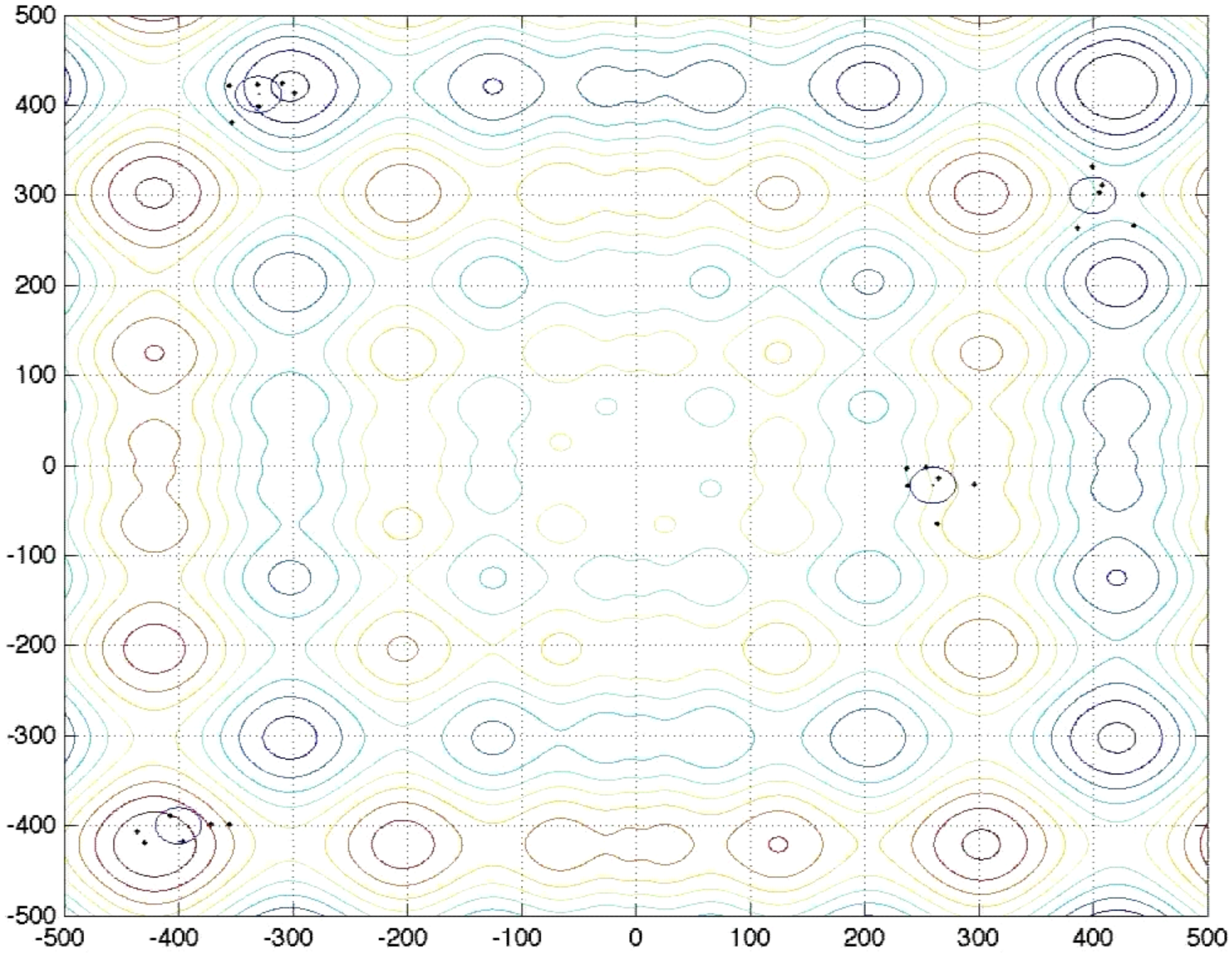


@microbionaut

cmueller@flatironinstitute.org

Thank you for your time! Questions?

FLATIRON INSTITUTE
Center for Computational Mathematics



@microbionaut

cmueller@flatironinstitute.org