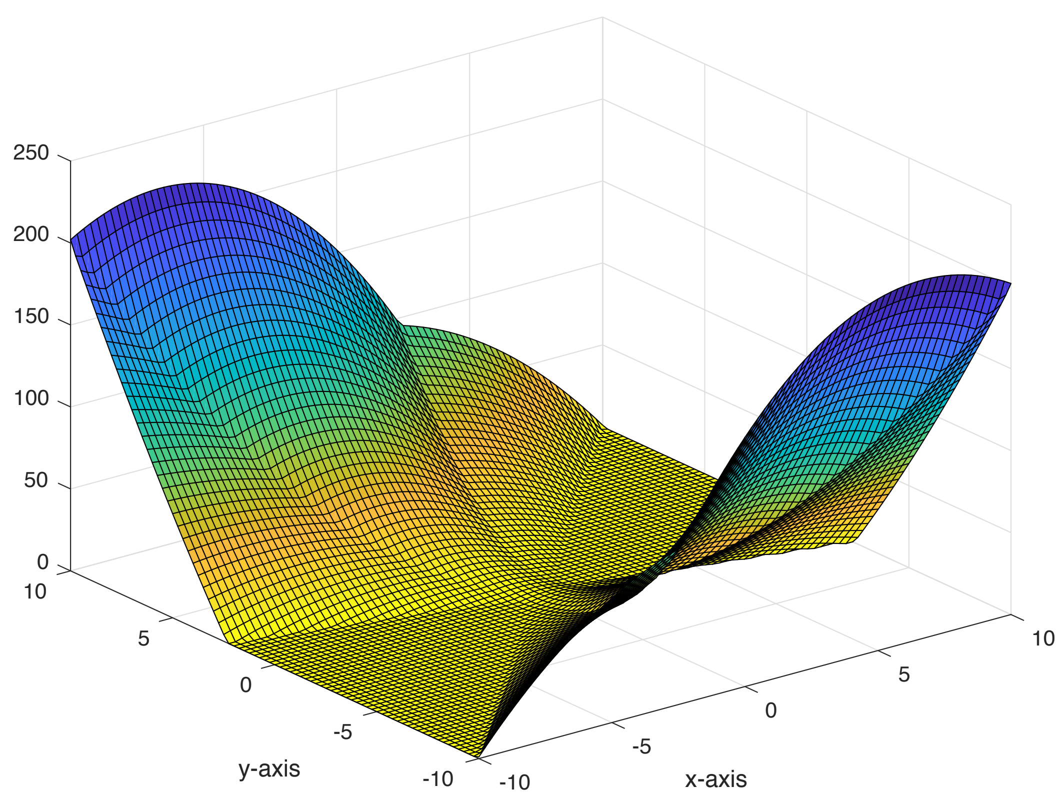
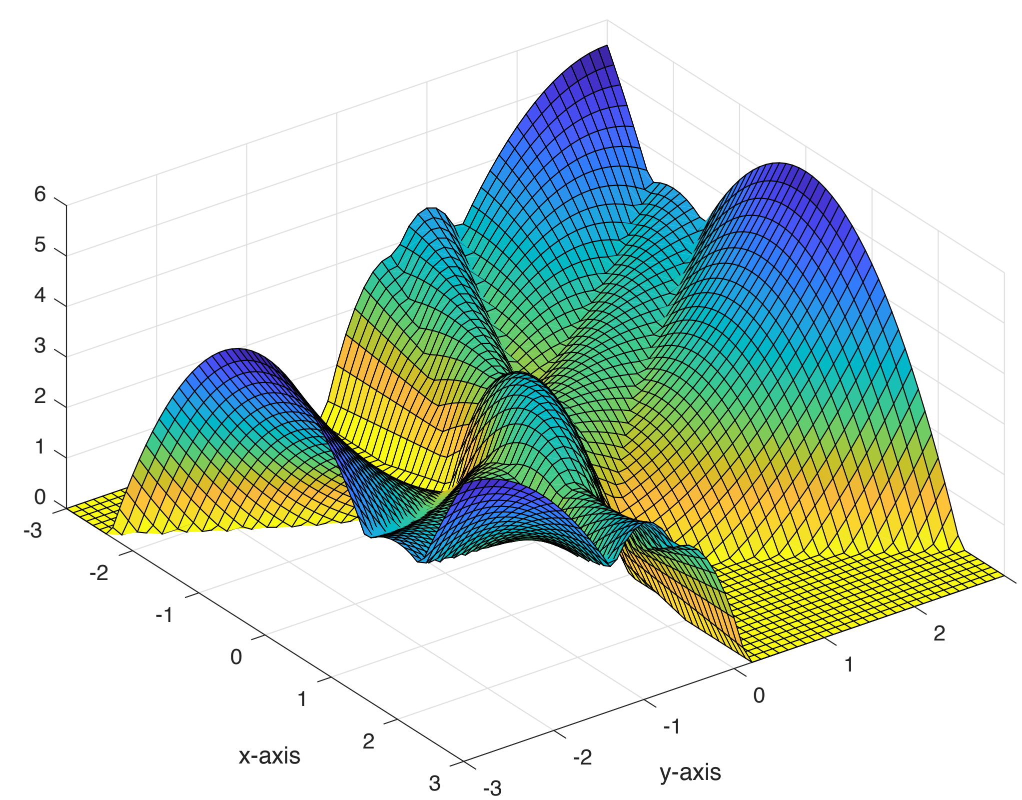
The Landscape of Non-convex Quadratic Feasibility

Amanda Bower, Lalit Jain, Laura Balzano





Quadratic Feasibility

Let $P_1, \ldots, P_m \in \mathbb{R}^{n \times n}$. The homogenous quadratic feasibility problem is defined as

$$\begin{aligned} \text{find} \quad & x \in \mathbb{R}^n \\ \text{subject to} \quad & x^T P_i x > 0, \quad \forall\, i = 1, \ldots, m\,. \end{aligned} \tag{1}$$

Quadratic Feasibility

Let $P_1, \ldots, P_m \in \mathbb{R}^{n \times n}$. The homogenous quadratic feasibility problem is defined as

$$\begin{aligned} \text{find} \quad & x \in \mathbb{R}^n \\ \text{subject to} \quad & x^T P_i x > 0, \quad \forall\, i = 1, \ldots, m\ . \end{aligned} \tag{1}$$

A solution to (1) is called a feasible point.

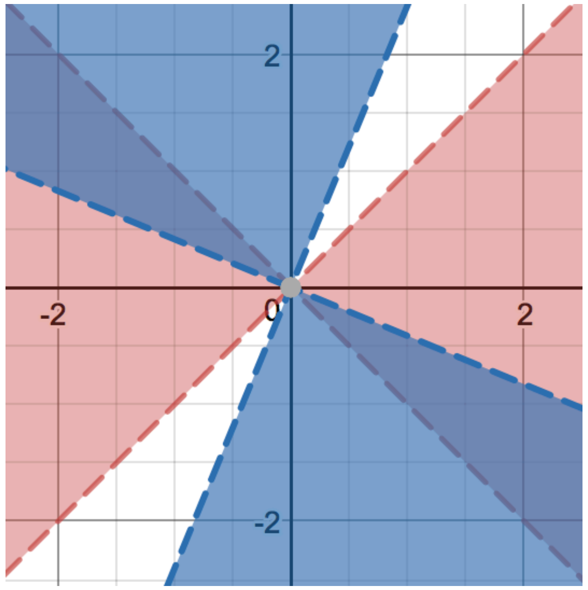
Quadratic Feasibility

Let $P_1, \ldots, P_m \in \mathbb{R}^{n \times n}$. The homogenous quadratic feasibility problem is defined as

$$\begin{aligned} \text{find} \quad & x \in \mathbb{R}^n \\ \text{subject to} \quad & x^T P_i x > 0, \quad \forall\, i = 1, \ldots, m\ . \end{aligned} \tag{1}$$

A solution to (1) is called a feasible point.

Example: Feasible region of $x^2 - y^2 > 0$ and $\frac{1}{2}x^2 + \frac{1}{2}y^2 - xy > 0$.



Quadratic Feasibility

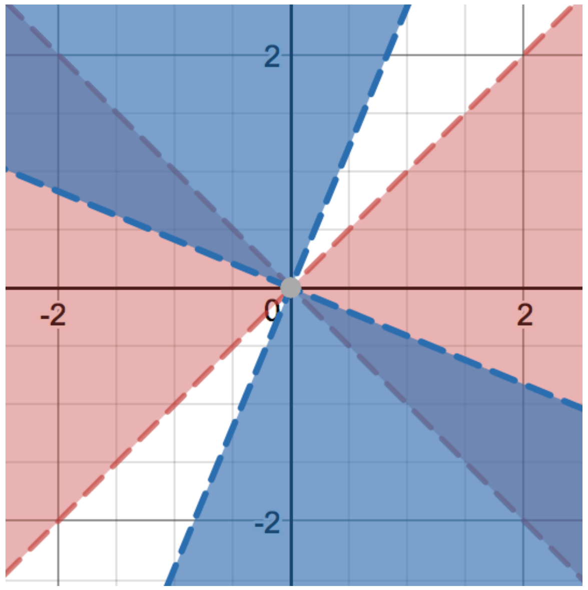
Let $P_1, \ldots, P_m \in \mathbb{R}^{n \times n}$. The homogenous quadratic feasibility problem is defined as

$$\begin{aligned} \text{find} \quad & x \in \mathbb{R}^n \\ \text{subject to} \quad & x^T P_i x > 0, \quad \forall\, i = 1, \ldots, m\ . \end{aligned} \tag{1}$$

A solution to (1) is called a feasible point.

Example: Feasible region of $x^2 - y^2 > 0$ and $\frac{1}{2}x^2 + \frac{1}{2}y^2 - xy > 0$.

$$\begin{pmatrix} x \\ y \end{pmatrix}^T \begin{pmatrix} 1 & 0 \\ 0 & -1 \end{pmatrix} \begin{pmatrix} x \\ y \end{pmatrix} > 0$$



Quadratic Feasibility: Motivation

Ordinal Embedding (aka non-metric multidimensional scaling):

Let $D(\cdot, \cdot)$ be a distance function.

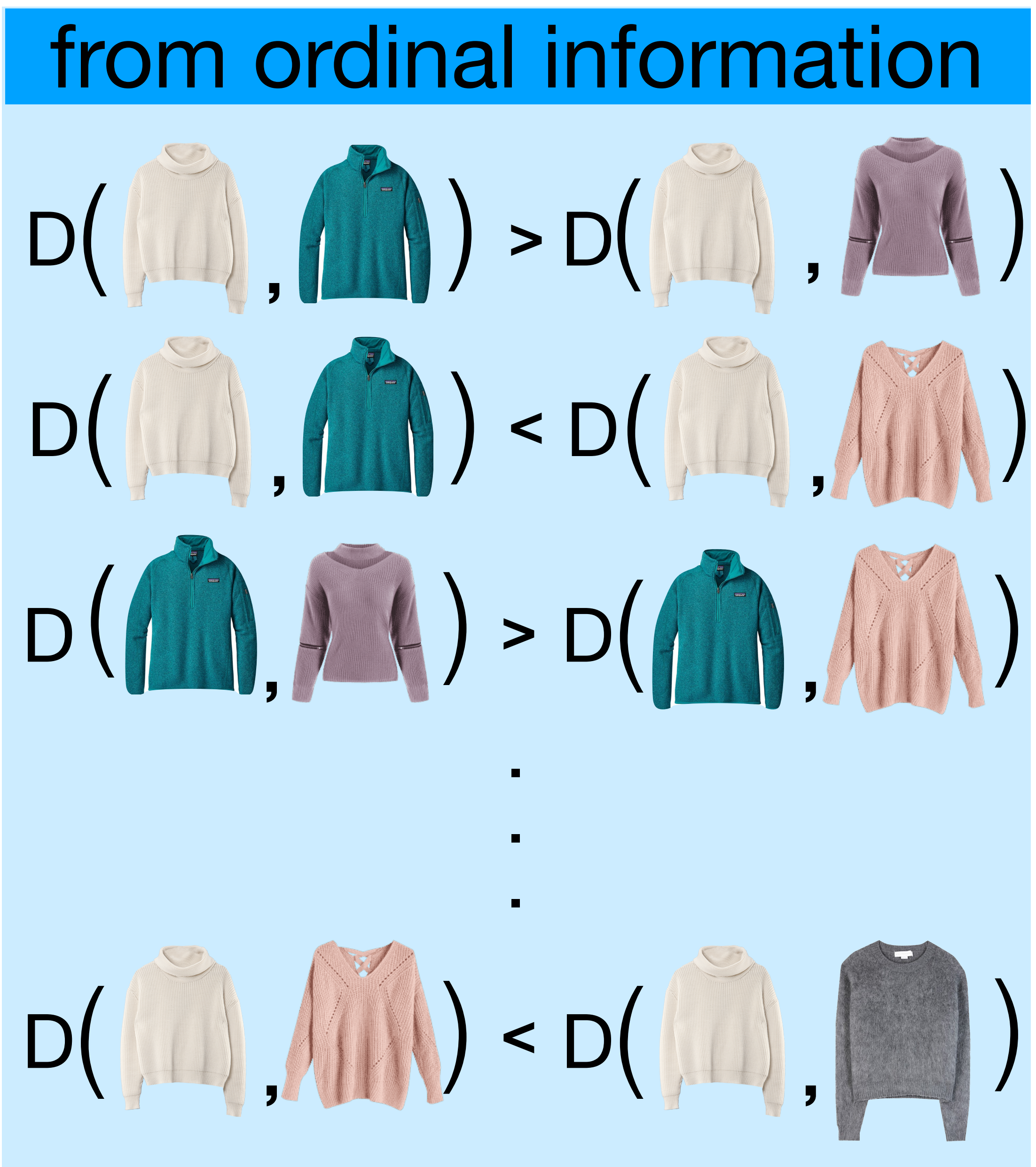
from ordinal information

D( , ) $\overset{?}{\gtrless}$ D( , )

Quadratic Feasibility: Motivation

Ordinal Embedding (aka non-metric multidimensional scaling):

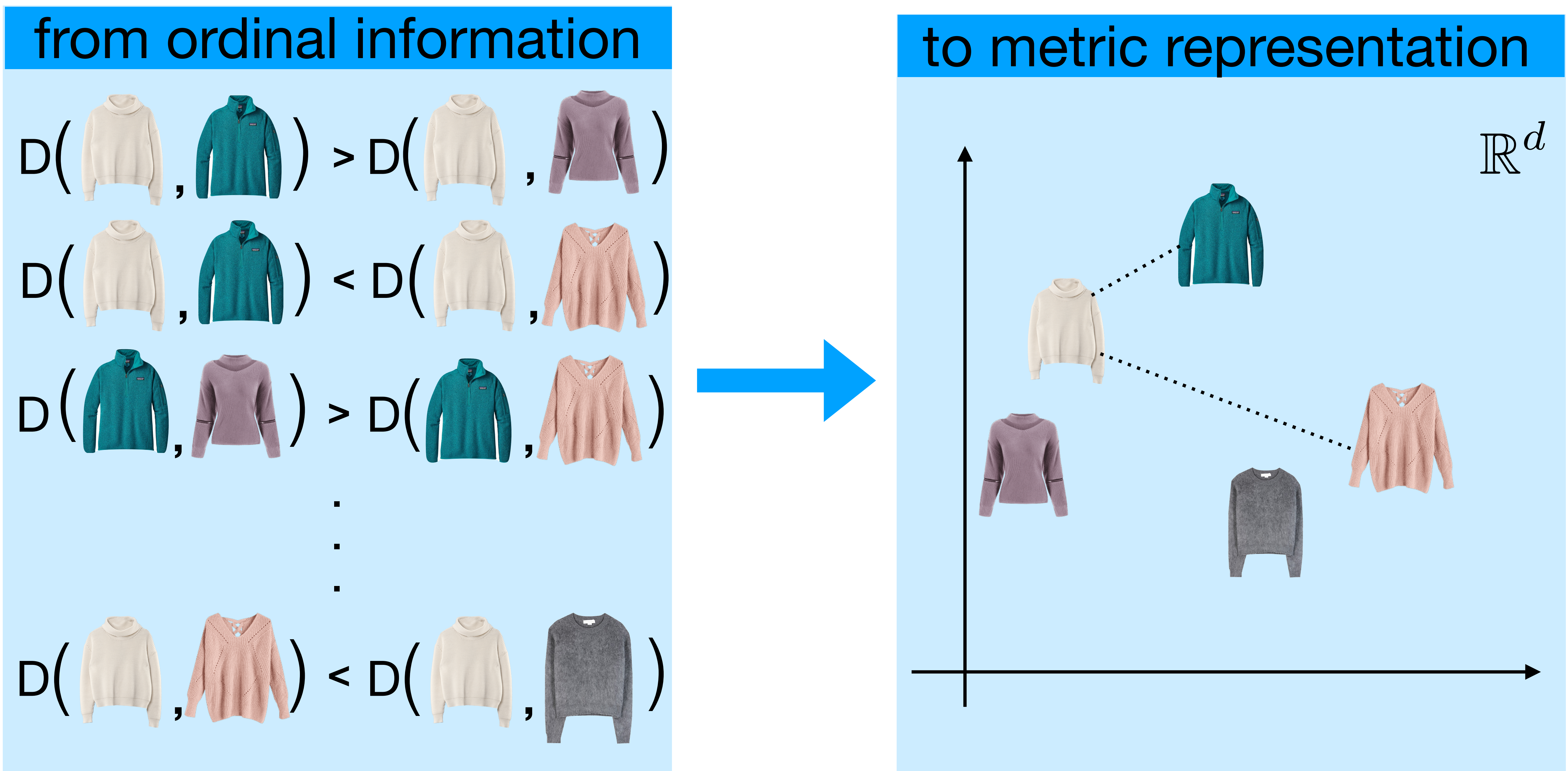
Let $D(\cdot, \cdot)$ be a distance function.



Quadratic Feasibility: Motivation

Ordinal Embedding (aka non-metric multidimensional scaling):

Let $D(\cdot, \cdot)$ be a distance function.



Quadratic Feasibility: Motivation

Ordinal embedding: Find $\{x_i \in \mathbb{R}^d\}$ such that

$$\|x_i - x_k\|_2^2 < \|x_i - x_j\|_2^2$$
$$\implies \langle x_i - x_k, x_i - x_k\rangle < \langle x_i - x_j, x_i - x_j\rangle$$
$$\implies 0 < x^T P_{ijk} x,$$

where $x = \begin{pmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{pmatrix} \in \mathbb{R}^{nd}$ and $P_{ijk} \in \mathbb{R}^{nd \times nd}$.

Quadratic Feasibility: Motivation

"w is closer to y than z":

Example in $\mathbb{R}^2$:

$$0 < \|w - z\|_2^2 - \|w - y\|_2^2$$

$$\Rightarrow 0 < \langle y, y \rangle + 2\langle w, z - y \rangle - \langle z, z \rangle$$

$$\Rightarrow 0 < x^T P_{wyz} x$$

$$x = \begin{pmatrix} w_1 \\ w_2 \\ y_1 \\ y_2 \\ z_1 \\ z_2 \end{pmatrix} \qquad P_{wyz} = \begin{pmatrix} 0 & 0 & -1 & 0 & 1 & 0 \\ 0 & 0 & 0 & -1 & 0 & 1 \\ -1 & 0 & 1 & 0 & 0 & 0 \\ 0 & -1 & 0 & 1 & 0 & 0 \\ 1 & 0 & 0 & 0 & -1 & 0 \\ 0 & 1 & 0 & 0 & 0 & -1 \end{pmatrix}$$

Quadratic Feasibility: Motivation

"w is closer to y than z":

Example in $\mathbb{R}^2$:

$$0 < \|w - z\|_2^2 - \|w - y\|_2^2$$

$$\Rightarrow 0 < \langle y, y \rangle + 2\langle w, z - y \rangle - \langle z, z \rangle$$

$$\Rightarrow 0 < x^T P_{wyz} x$$

$$x = \begin{pmatrix} w_1 \\ w_2 \\ y_1 \\ y_2 \\ z_1 \\ z_2 \end{pmatrix} \qquad P_{wyz} = \begin{pmatrix} 0 & 0 & -1 & 0 & 1 & 0 \\ 0 & 0 & 0 & -1 & 0 & 1 \\ -1 & 0 & 1 & 0 & 0 & 0 \\ 0 & -1 & 0 & 1 & 0 & 0 \\ 1 & 0 & 0 & 0 & -1 & 0 \\ 0 & 1 & 0 & 0 & 0 & -1 \end{pmatrix}$$

trace(P_{wyz})

Quadratic Feasibility: Motivation

"w is closer to y than z":

Example in $\mathbb{R}^2$:

$$0 < \|w - z\|_2^2 - \|w - y\|_2^2$$

$$\Rightarrow 0 < \langle y, y \rangle + 2\langle w, z - y \rangle - \langle z, z \rangle$$

$$\Rightarrow 0 < x^T P_{wyz} x$$

$$x = \begin{pmatrix} w_1 \\ w_2 \\ y_1 \\ y_2 \\ z_1 \\ z_2 \end{pmatrix} \qquad P_{wyz} = \begin{pmatrix} 0 & 0 & -1 & 0 & 1 & 0 \\ 0 & 0 & 0 & -1 & 0 & 1 \\ -1 & 0 & 1 & 0 & 0 & 0 \\ 0 & -1 & 0 & 1 & 0 & 0 \\ 1 & 0 & 0 & 0 & -1 & 0 \\ 0 & 1 & 0 & 0 & 0 & -1 \end{pmatrix}$$

trace(P_{wyz})

indefinite
(has positive & negative eigenvalues)

The Optimization Problem

$$\begin{aligned} \text{find} \quad & x \in \mathbb{R}^n \\ \text{subject to} \quad & x^T P_i x > 0, \quad i = 1, \ldots, m\ . \end{aligned} \tag{1}$$

The Optimization Problem

$$\begin{aligned} \text{find} \quad & x \in \mathbb{R}^n \\ \text{subject to} \quad & x^T P_i x > 0, \quad i = 1, \ldots, m \,. \end{aligned} \tag{1}$$

We formulate (1) as a non-convex, unconstrained optimization problem with the hinge loss:

$$\underset{x \in \mathbb{R}^n}{\text{minimize}} \sum_{i=1}^{m} \max\{0, 1 - x^T P_i x\}. \tag{2}$$

The Optimization Problem

$$\begin{aligned} \text{find} \quad & x \in \mathbb{R}^n \\ \text{subject to} \quad & x^T P_i x > 0, \quad i = 1, \ldots, m \;. \end{aligned} \qquad (1)$$

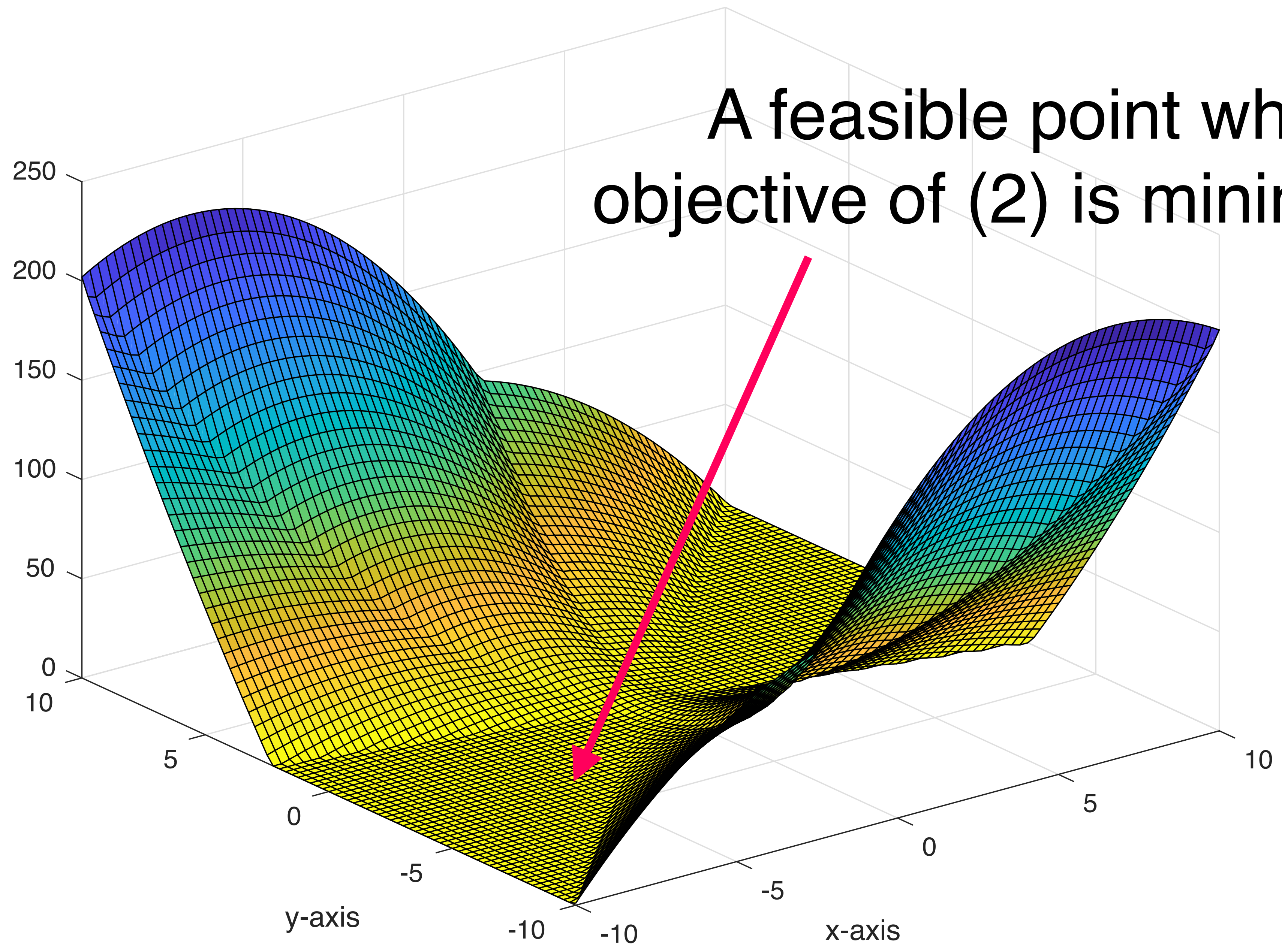
We formulate (1) as a non-convex, unconstrained optimization problem with the hinge loss:

$$\underset{x \in \mathbb{R}^n}{\text{minimize}} \sum_{i=1}^{m} \max\{0, 1 - x^T P_i x\}. \qquad (2)$$

When a feasible point of (1) exists:

global minimizers of (2) $\Leftrightarrow$ feasible points of (1).

The Optimization Problem



A feasible point where the objective of (2) is minimized (=0).

Goal

$$\underset{x \in \mathbb{R}^n}{\text{minimize}} \sum_{i=1}^{m} \text{max}\{0, 1 - x^T P_i x\} \qquad (2)$$

Goal

$$\underset{x\in\mathbb{R}^n}{\text{minimize}} \sum_{i=1}^{m} \text{max}\{0, 1 - x^T P_i x\} \qquad (2)$$

Solving the optimization problem (2):

Standard optimization tools such as stochastic gradient descent have a chance at solving (2).

Goal

$$\underset{x \in \mathbb{R}^n}{\text{minimize}} \sum_{i=1}^{m} \max\{0, 1 - x^T P_i x\} \qquad (2)$$

Solving the optimization problem (2):

Standard optimization tools such as stochastic gradient descent have a chance at solving (2).

However, assuming feasibility, success of recovering a feasible point crucially depends on every local minimizer of (2) being a global minimizer. To this end, **the goal is to classify the non-global minimizers of (2)**.

Related Work

Examples of non-convex problems where all local minima are global minima under suitable assumptions:

Phase retrieval [Sun, Qu, Wright 2016]; **Neural networks** [Kawaguchi 2016; Haeffele and Vidal 2017; Ge, Lee, Ma 2017]; **Matrix completion** [Ma 2016]; **Burer-Montiero Factorization for Semidefinite Programs** [Boumal, Voroninski, and Bandeira 2016]

Related Work

Work that proposes optimization problems for finding an ordinal embedding:

- Kruskal, 1964;
- Agarwal et al., 2007
- Terada & Von Luxberg, 2014
- Jain, Jamieson, & Nowak, 2016.

Related Work

Other work that proposes (2) or studies (1):

- Konar and Sidiropoulos. *Fast feasibility pursuit for non-convex QCQPS via first-order methods,* 2017.
- Boyd and Park. *General Heuristics for Nonconvex Quadratically Constrained Quadratic Programming,* 2017.
- Luo et al. *Semidefinite Relaxation of Quadratic Optimization Problems,* 2010.

Related Work

Other work that proposes (2) or studies (1):

- Konar and Sidiropoulos. *Fast feasibility pursuit for non-convex QCQPS via first-order methods,* 2017.
- Boyd and Park. *General Heuristics for Nonconvex Quadratically Constrained Quadratic Programming,* 2017.
- Luo et al. *Semidefinite Relaxation of Quadratic Optimization Problems,* 2010.

However, none of these works theoretically studies the landscape of (2).

Our Results

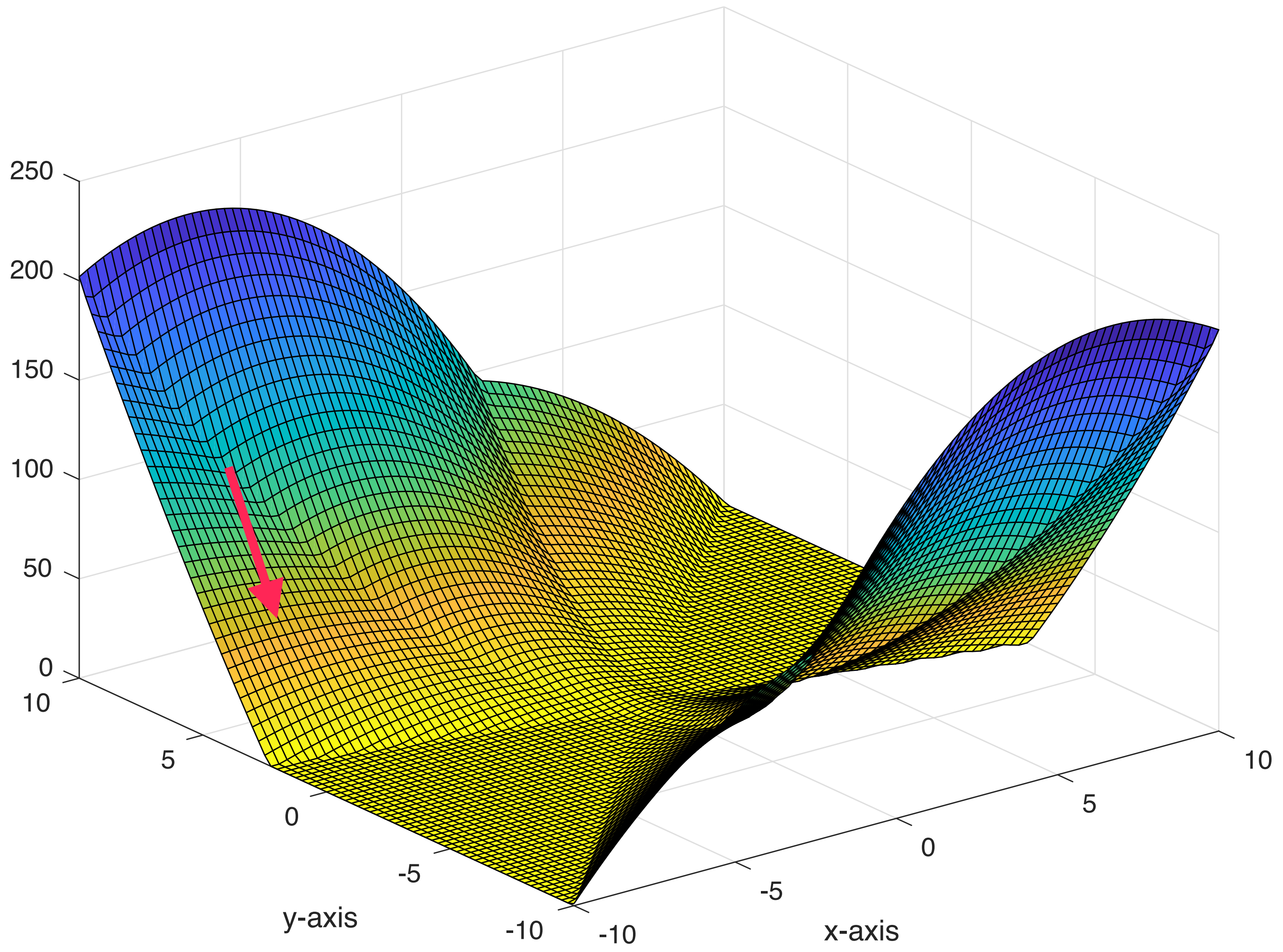
Two dimensions:

Theorem: Assume $P_1, \ldots, P_m \in \mathbb{R}^{2\times 2}$ are trace zero and symmetric such that a feasible point exists. Furthermore, assume no three of the curves $x^T P_i x = 1$ intersect at a point. Every local minimizer of (2) is a global minimizer.

$$\underset{x\in\mathbb{R}^n}{\text{minimize}} \sum_{i=1}^{m} \max\{0, 1 - x^T P_i x\} \quad (2)$$

Our Results

Proof idea: For any point that is not a global minimizer, we exhibit a descent direction.

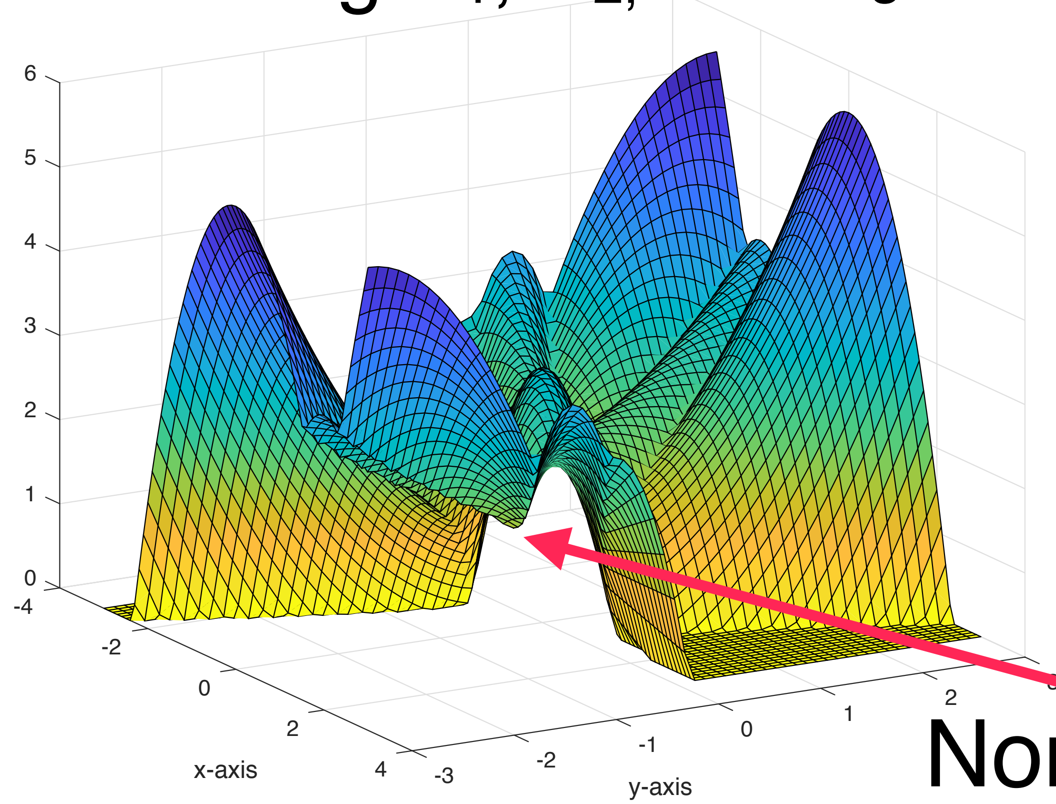


Importance of Assumptions

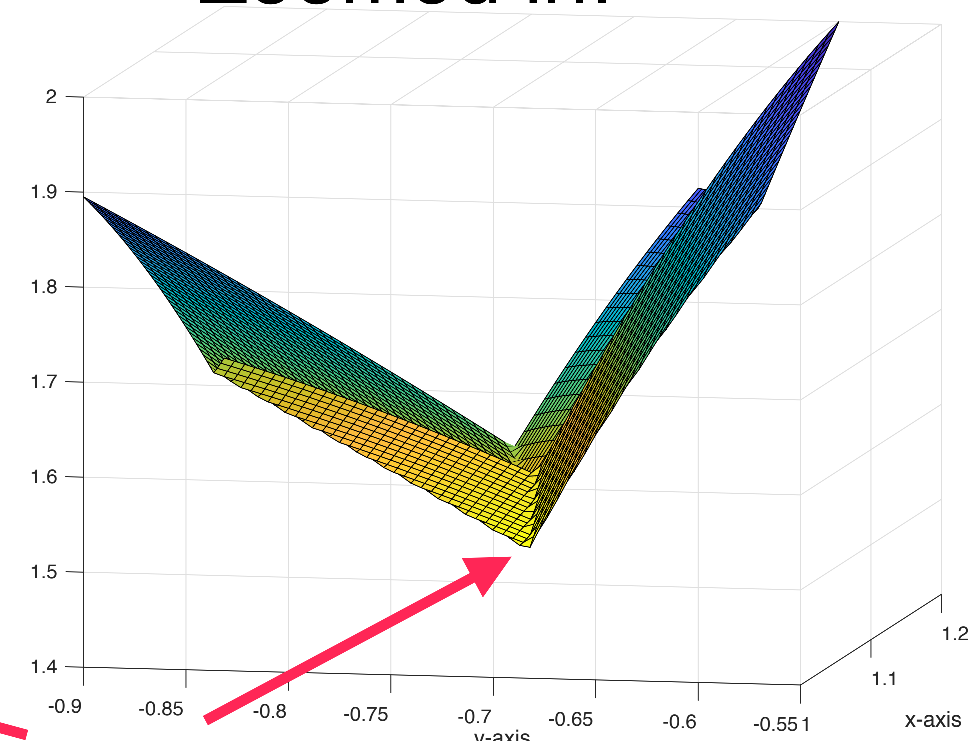
Consider $P_1 = \begin{pmatrix} 1 & 0 \\ 0 & -.5 \end{pmatrix}$, $P_2 = \begin{pmatrix} .5 & 1 \\ 1 & 1 \end{pmatrix}$, $P_3 = \begin{pmatrix} 0 & 1 \\ 1 & 5 \end{pmatrix}$.

- Indefinite, but not trace zero.
- $[1, 1]^T$ is a feasible point.
- $[1.1, -.7]$ is approximately a non-global minimizer.

Objective of (2) using P_1, P_2, and P_3:



Zoomed in:



Non-global, local minimizer

Our Results

Theorem: Let $\{P_i \in \mathbb{R}^{n \times n}\}$ be a set of real, symmetric trace 0 matrices. Assume the P_i share a feasible point. If $x \in \mathbb{R}^n$ is a non-global minimizer of (2), x must satisfy the following two equations:

P1) $\sum_{\{i : x^T P_i x < 1\}} x^T P_i x < 0$

P2) $\sum_{\{i : x^T P_i x < 1\}} x^T P_i x + \sum_{\{i : x^T P_i x = 1\}} x^T P_i x \geq 0.$

In particular, $\{i : x^T P_i x = 1\} \neq \varnothing$.

Take-away: Non-global minimizers arise when at least one constraint is equal to one.

$$\underset{x \in \mathbb{R}^n}{\text{minimize}} \sum_{i=1}^{m} \text{max}\{0, 1 - x^T P_i x\} \quad (2)$$

Experiments

We used mini-batch stochastic gradient descent to solve the optimization problem in all experiments:

step size

$$x^{(i)} = \underline{x^{(i-1)}} - \underline{\eta_i} \underline{\sum_{\{j \in \underline{T_i} : x^{(i-1)^T} P_j x^{(i-1)} < 1\}} -2P_j x^{(i-1)}}.$$

current estimate

mini-batch

sub-gradient

We call an experiment successful if stochastic gradient descent converged to a feasible point.

Experiments

Step size:

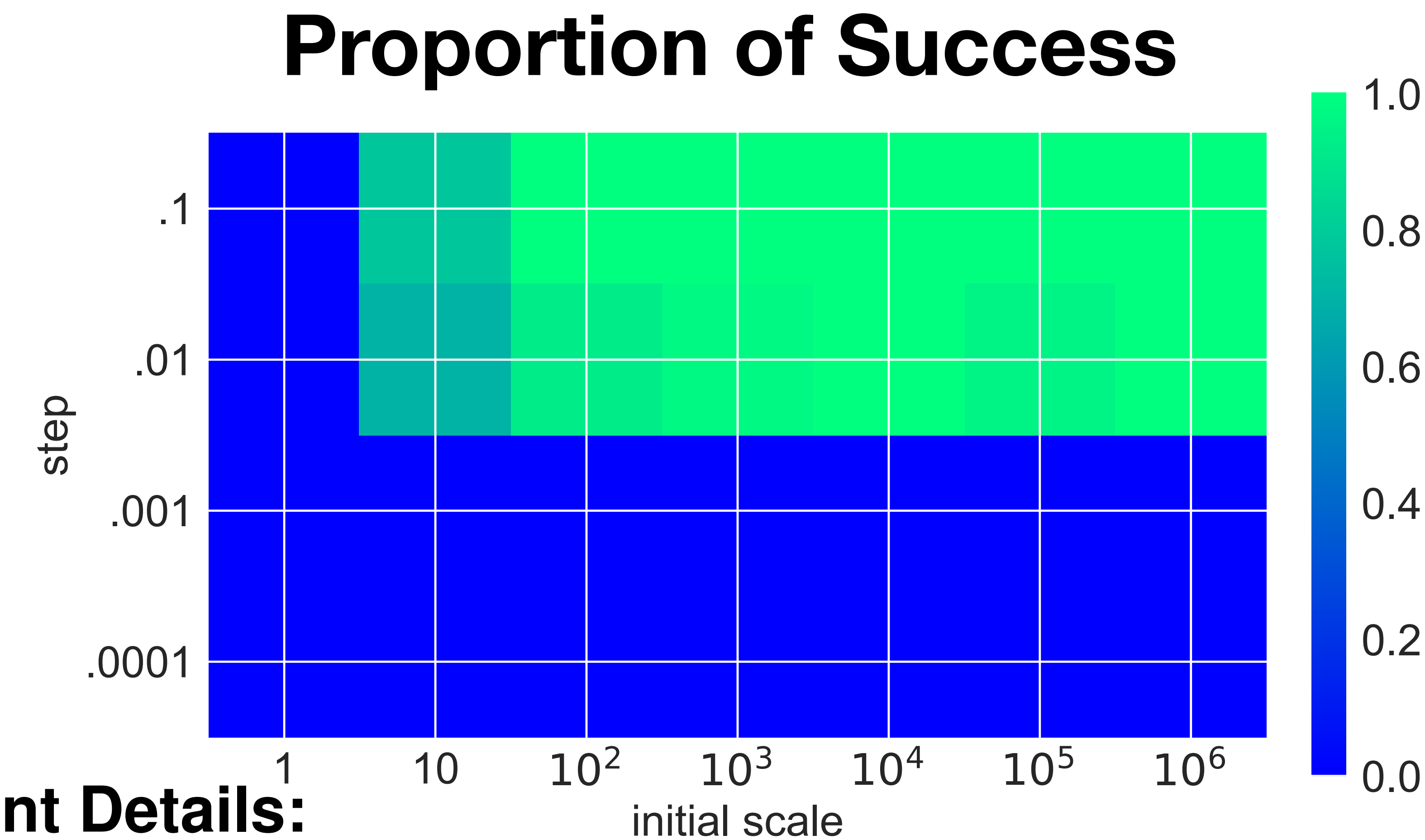
$$\eta_i = \text{(initial step)} * .5^{(i/\text{num of quadratic constraints})},$$

where we vary the initial step.

Initialization:

We pick a random point for initialization but we vary the norm, which we call “initial scale.”

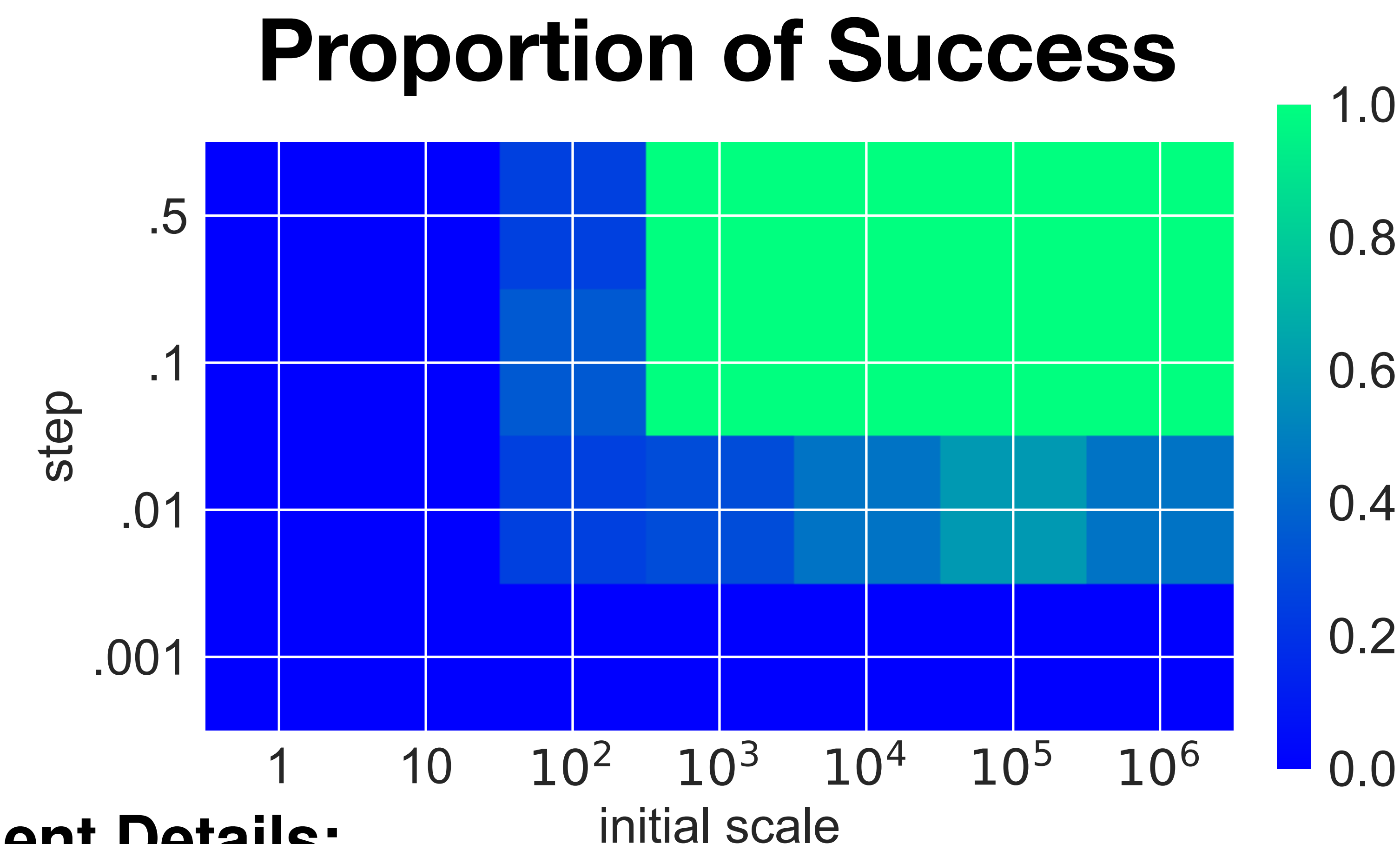
Experiments: Random Constraints



Experiment Details:

- 2000 trace zero symmetric $\mathbb{R}^{20\times 20}$ matrices with entries were drawn from $\mathcal{N}(0,1)$ with a feasible point.
- Stochastic gradient descent capped at 4000 epochs with mini-batch sizes of 300.
- 50 experiments were ran per initial step and initial scale.

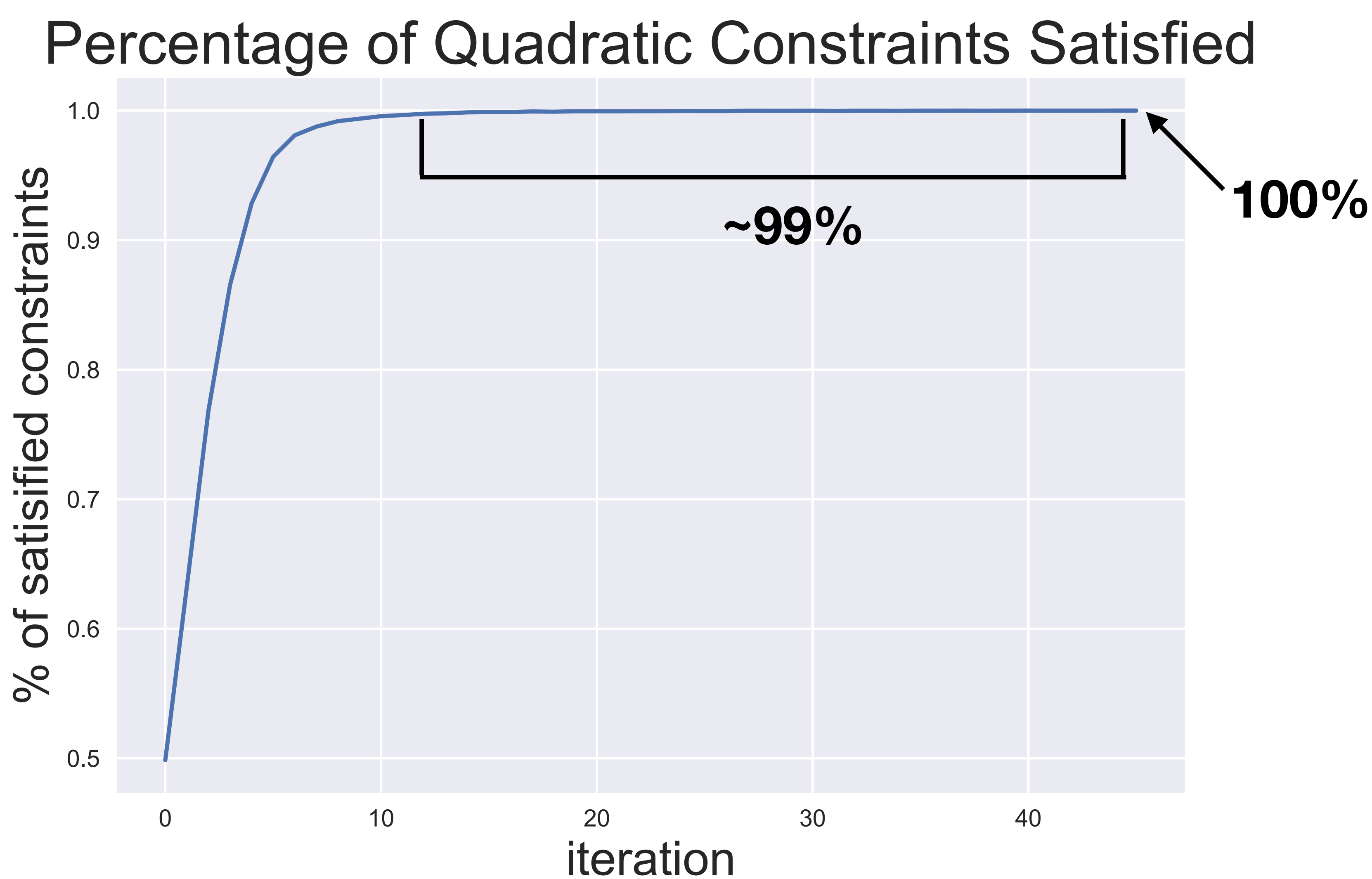
Experiments: Ordinal Embedding



Experiment Details:

- All $O(50^3)$ triplet constraints were collected from 50 points in $\mathbb{R}^2$ whose coordinates were drawn from $\mathcal{N}(0,1)$.
- Stochastic gradient descent capped at 8000 epochs with mini-batch sizes of 1000.
- 20 experiments were ran per initial step and initial scale.

Experiments: Ordinal Embedding



one ordinal embedding experiment where initial step = .5 and initial scale = 10000

Open Questions

1. Classify landscape in higher dimensions.

Open Questions

1. Classify landscape in higher dimensions.
2. Guided by the landscape, prove convergence rates for stochastic gradient descent.

Open Questions

1. Classify landscape in higher dimensions.
2. Guided by the landscape, prove convergence rates for stochastic gradient descent.
3. Understand why large norm initialization works.