Global Arguments in Non-Convex Optimization: Finding Stationary Points in Low Dimensions

Dan Mikulincer

Joint work with Sébastien Bubeck

The Problem

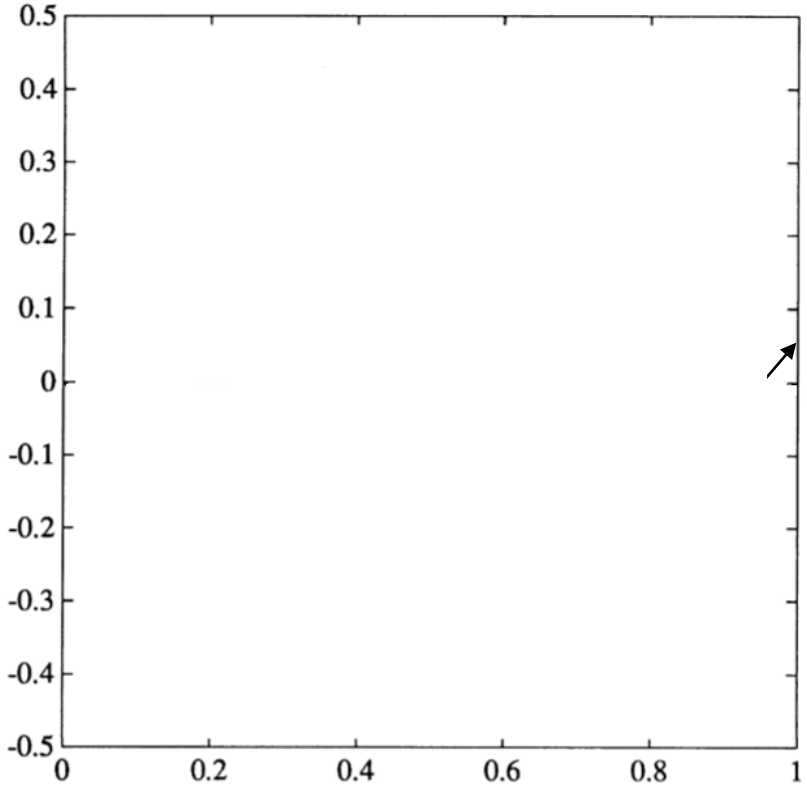
- Given $f:[0,1]^d \to \mathbb{R}$, differentiable (non convex) and $\varepsilon > 0$, find
$$x \in [0,1]^d, \|\nabla f(x)\| \leq \varepsilon.$$

- First-order oracle model:
  - Algorithm may query a point for the function and gradient values.

- The goal is to find an ε-stationary point, with a minimal number of queries.

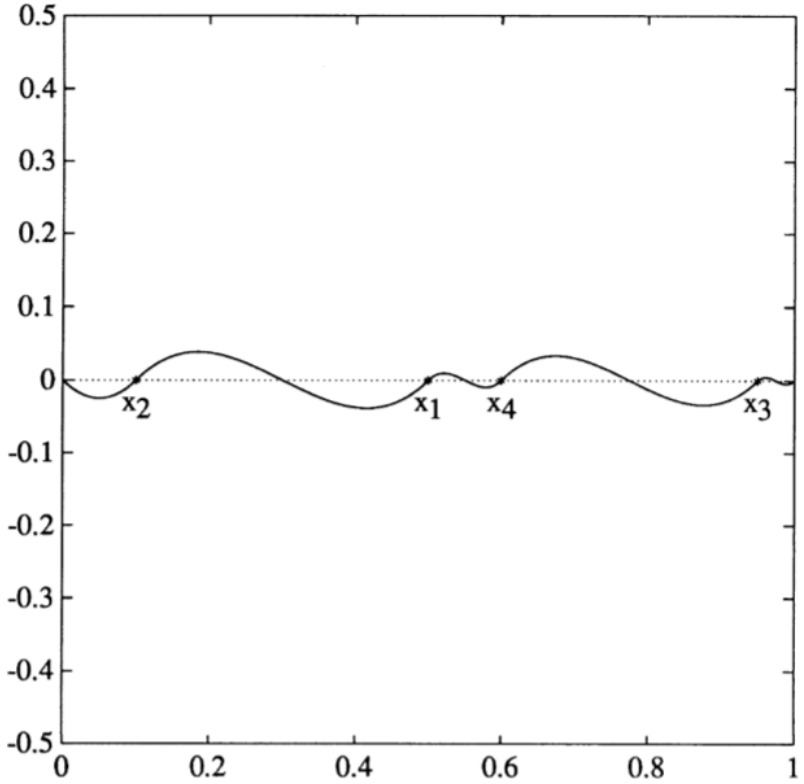
The Problem

- Without further restriction the problem is intractable.



The Problem

- Without further restriction the problem is intractable.



The Problem

- In the example, the second derivative might be very large.

- We impose the following (gradient-) Lipschitz condition on f

$$\|f(x) - f(y)\|, \|\nabla f(x) - \nabla f(y)\| \leq \|x - y\|.$$

Gradient Descent

- Choose x_0 arbitrarily from $[0,1]^d$.
- Update rule: $x_{i+1} = x_i - \varepsilon \frac{\nabla f(x_i)}{\|\nabla f(x_i)\|}$
- If at any point $\|\nabla f(x_i)\| \leq \varepsilon$, terminate.
- Claim: the algorithm terminates after $O\left(\frac{1}{\varepsilon^2}\right)$ queries.

Gradient Descent

- Set $x_* = \text{argmin } f(x)$, and note that $f(x_0) - f(x_*) = O(1)$.

- Using the smoothness of the function, a second order approximation shows:

$$f(x_i) - f(x_{i+1}) \geq \|\nabla f(x_i)\|\varepsilon - \frac{\varepsilon^2}{2} \geq \frac{\varepsilon^2}{2}.$$

- So, the algorithm terminates after $O\left(\frac{1}{\varepsilon^2}\right)$ gradient steps.

Optimality of Gradient Descent

- In high dimensions, gradient descent is optimal

- Carmon, Duchi, Hinder, Sidford, 17':

> For every $\varepsilon > 0$, there exists $d_0 \in \mathbb{N}$, such that for any $d > d_0$, and algorithm A there exists a 'smooth' function $f: [0,1]^d \to \mathbb{R}$ such that A requires $\Omega\left(\frac{1}{\varepsilon^2}\right)$ queries in order to find an ε-stationary point.

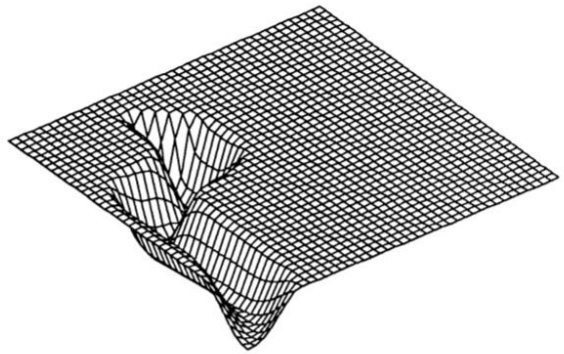
Lower Dimensions

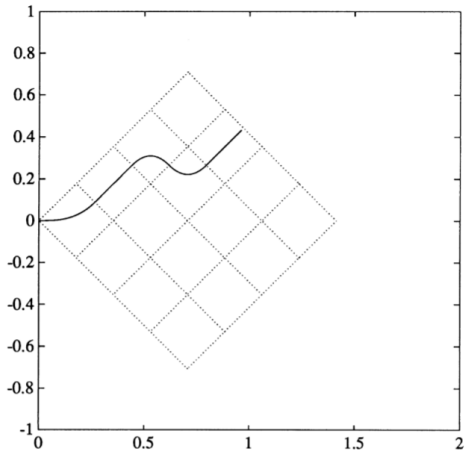
- The situation is more mysterious in low dimensions.

- In 91′ Vavasis gave a $\Omega\left(\frac{1}{\sqrt{\varepsilon}}\right)$ lower bound.
- The bound applies for deterministic algorithms in dimension 2.

Dimension 2

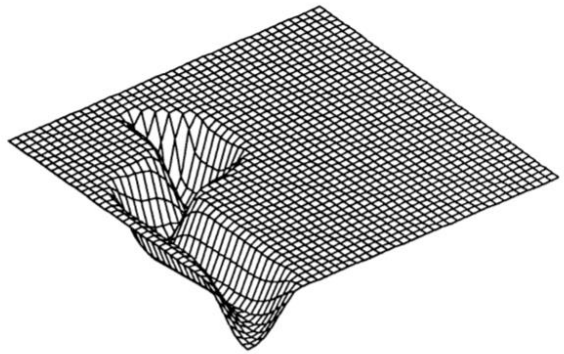
- The idea is to reduce the problem to a discrete setting.
- Specifically, finding ε-stationary point in $[0,1]^2$, is reduced to finding a local minima on the discrete $\left[\frac{1}{\sqrt{\varepsilon}}\right] \times \left[\frac{1}{\sqrt{\varepsilon}}\right]$ grid.

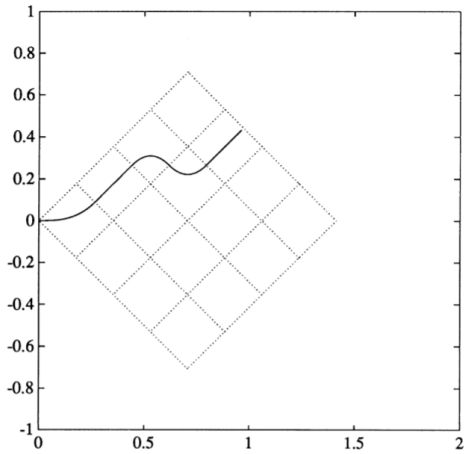




Dimension 2 – Random Algorithms

- Sun and Yao (06′) gave a $\Omega^*(n)$ lower bound for finding local minima in $[n] \times [n]$, which applies to random algorithms.
- This translates to a corresponding $\Omega^*\left(\frac{1}{\sqrt{\varepsilon}}\right)$ bound, for ε-stationary points.





Going Up in the Dimension Scale

- Zhang (04′) gives a $\Omega\left(n^{\frac{d}{2}}\right)$ lower bound for local minima on $[n]^d$.

- Vavasis' construction implies a lower bound of $\Omega\left(\varepsilon^{-\frac{d}{d+2}}\right)$, for finding ε-stationary points in $[0,1]^d$.

Improving Gradient Descent

- One possible improvement to gradient descent would be to choose the initial point in a smarter way.

- If $f(x_0) - f(x_*) = \Delta$, the analysis shows that gradient descent will find an ε-stationary in $O\left(\frac{\Delta}{\varepsilon^2}\right)$ queries.

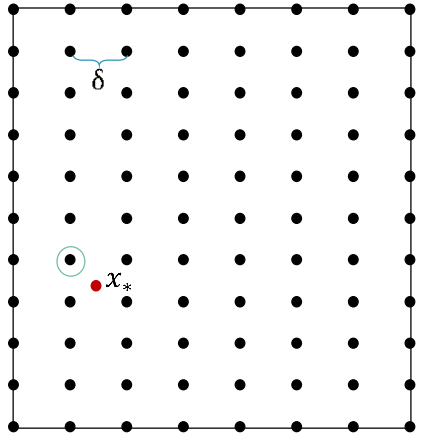
A Global Argument

- Vavasis proposed the following way to choose x_0.

- For some $\delta > 0$, take a grid $\mathrm{N} \subset [0,1]^d$ with points spaced δ apart.
- Put $x_0 = \operatorname{argmin}\{f(x) | x \in N\}$.

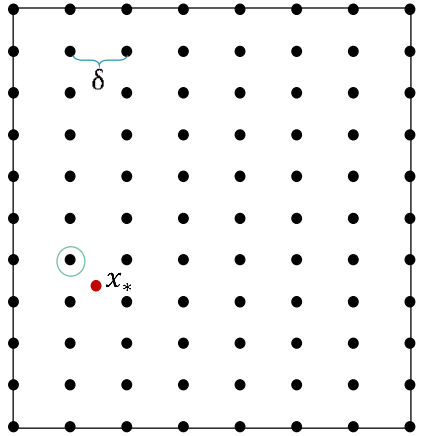
- Since $\nabla \mathrm{f}(x_*) = 0$

$$f(x_0) - f(x_*) \leq \frac{\delta^2 \cdot d}{2}.$$



A Global Argument

- The main idea is that we are able to observe a value which is close to the value of the optimum by querying a small box around it.



A Global Argument - Analysis

- The grid has $O\left(\frac{1}{\delta^d}\right)$ points. So, the total number of queries is:

$$O\left(\frac{1}{\delta^d}\right) + O\left(\frac{d\delta^2}{\varepsilon^2}\right).$$

- Optimizing over δ, we get

$$O\left(\left(\frac{d}{\varepsilon^2}\right)^{\frac{d}{d+2}}\right).$$

- In particular, in dimension 2, this is $O\left(\frac{1}{\varepsilon}\right)$.

Further Improvement

- In $[0,1]^2$, we present a better, optimal, algorithm.

- Our algorithm consists of two main ideas:
  - A local improvement to gradient descent.
  - A global argument.

Key Observation

- The main observation is that the function is relatively flat, in directions which are perpendicular to the gradient.

- Formally,

  If $x \in [0,1]^2$ and $v \in S^2$, is such that $\langle v, \nabla f(x) \rangle = 0$. Then

$$|f(x + t \cdot v) - f(x)| \leq \frac{t^2}{2}.$$

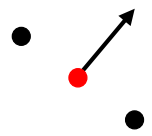
Key Observation

- So, given a point $x \in [0,1]^2$.

- We are able to detect another point $y \in [0,1]^2$ such that

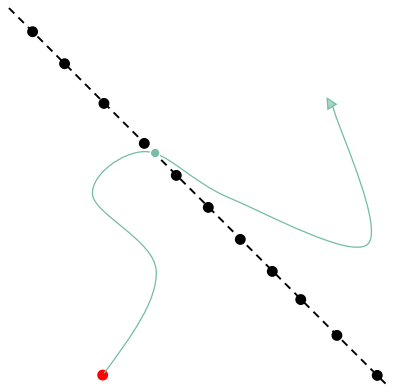
$$|f(x) - f(y)| \leq \frac{\delta^2}{2},$$

provided that $|y - x| \leq \delta$ and that y lies in the direction orthogonal to $\nabla f(x)$.



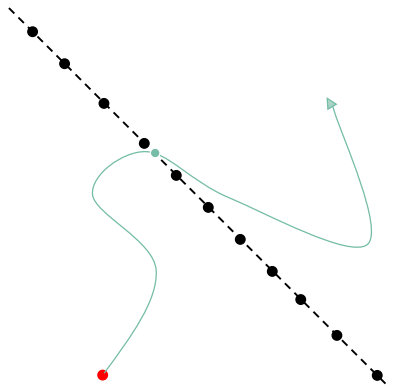
Key Observation

- Let $L \subset [0,1]^2$ be a line segment and $w \in [0,1]^2$.
- Suppose that we know that the gradient flow originating at w crosses L.
- We can detect the value at the exit point by querying a δ-net of the L.



Key Observation

- If w is 'far' from L, and there is no ε-stationary point on the gradient flow leading to L, we will see a decrease in function value on the net.



The ‘Net Lemma’

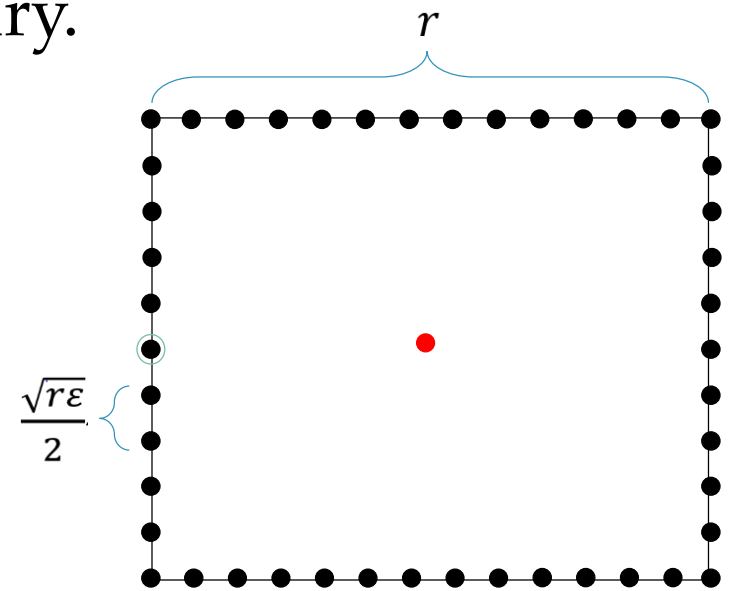
Lemma

Let $L \subset [0,1]^2$ be a line segment, and let $N \subset L$ be a δ-net of L.
Fix $w \in [0,1]^2$, and suppose that the gradient flow originating at w crosses L and that there is no ε-stationary on the flow, between w and L. Then,

$$f(w) \geq \min\{f(x) | f(x) \in N\} - 2\delta^2 + \varepsilon \cdot \mathrm{dist}(w, L).$$

A Local Improvement Procedure

- Fix a scale $r > 0$ and a point $x \in [0,1]^2$.
- Put a square with perimeter $4r$ around x.
- Query an $\frac{\sqrt{r\varepsilon}}{2}$-net, N of the box's boundary.
- Update $x_{new} = \text{argmin}\,\{f(x)|x \in N\}$.



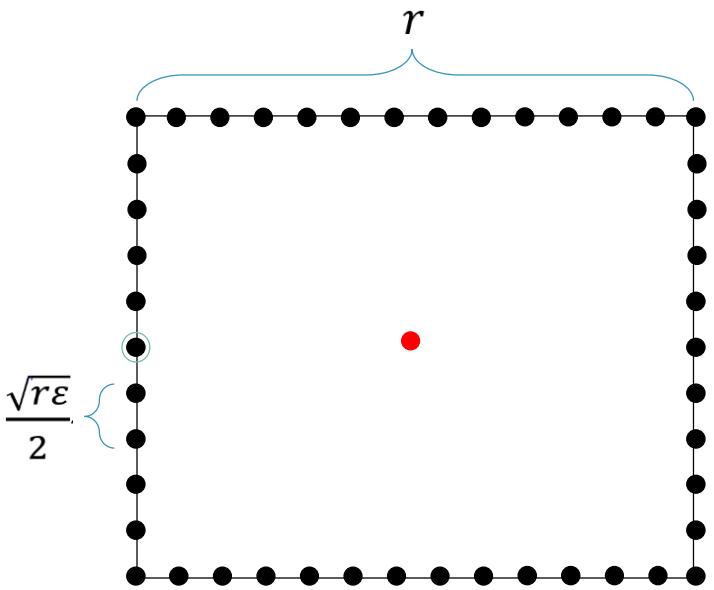
A Local Improvement Procedure

- Suppose that the square does not contain an ε-stationary point.
- The 'net lemma' tells us that

$$f(x) - f(x_{new}) \geq \frac{r\varepsilon}{2} - \frac{r\varepsilon}{4} \geq \frac{r\varepsilon}{4}.$$

- So, using $O\left(\frac{r}{\sqrt{r\varepsilon}}\right) = O\left(\sqrt{\frac{r}{\varepsilon}}\right)$ queries

we achieved an improvement of $\frac{r\varepsilon}{4}$.



A Local Improvement Procedure

- Compare to gradient descent, which would get an improvement of $O\left(\sqrt{\frac{r}{\varepsilon}}\varepsilon^2\right) = O\left(\sqrt{r}\,\varepsilon^{\frac{3}{2}}\right)$,using $O\left(\sqrt{\frac{r}{\varepsilon}}\right)$ queries.

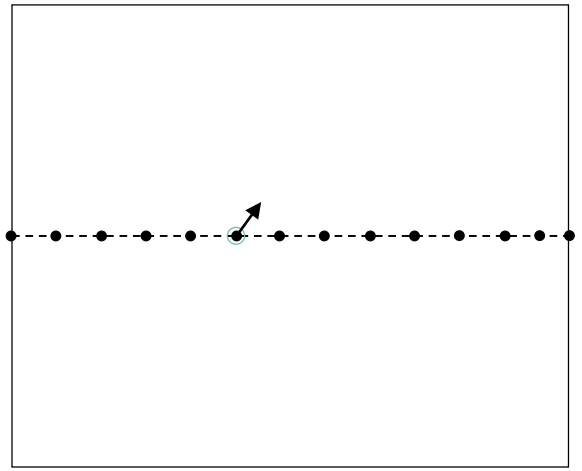
- As long as $r \geq \varepsilon$. We are doing something better gradient descent!

- Remark that if we know the box contains an ε-stationary point, we may find it using $O\left(\frac{r^2}{\varepsilon^2}\right)$ queries.

A Global Argument

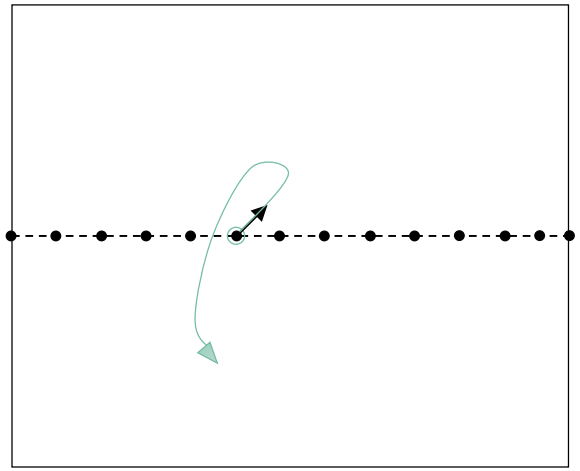
- Idea: cut the space into two pieces and trap a stationary point in one.

- This could be achieved by looking at a net of the meridian.



A Global Argument

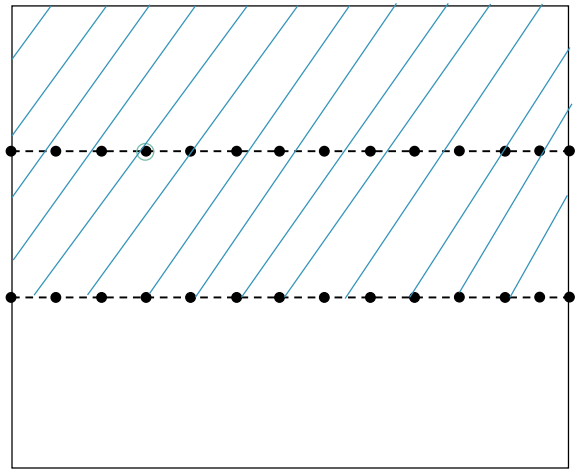
- Problem: the gradient flow could escape the ‘holes’ in the net.



Parallel Plates Solution

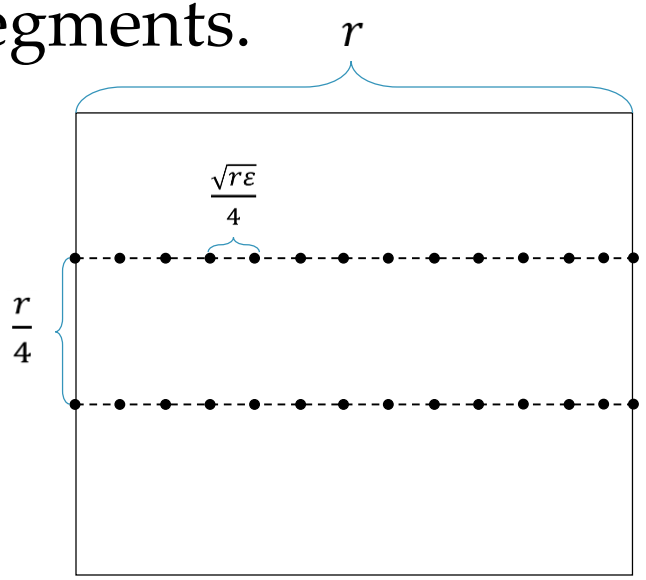
- Instead we cut the domain using two parallel lines.

- We choose the side which contains the best point on the net.



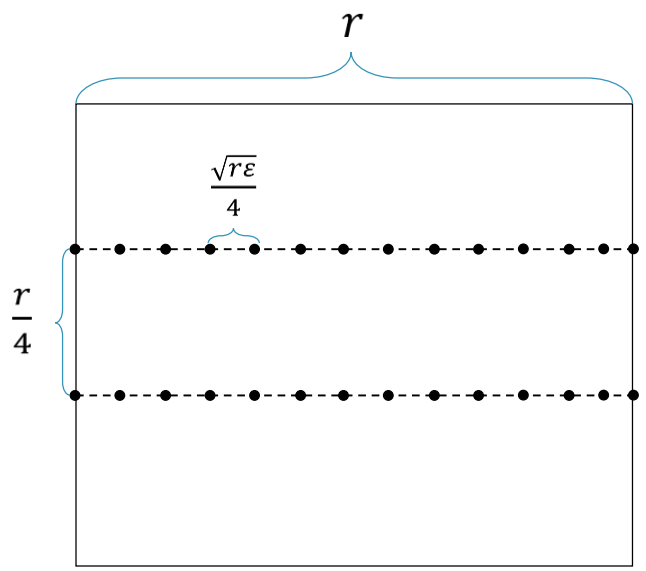
Parallel Plates Solution

- Suppose current domain has side-length r.
- We can take a $\frac{\sqrt{r\varepsilon}}{4}$-net of the line segments.
- As well as a larger distance between the segments.



Parallel Plates Solution

- The ‘net lemma’ tells us that as long as the distance between the segments is large enough, there has to exist an ε-stationary point in the side we chose.

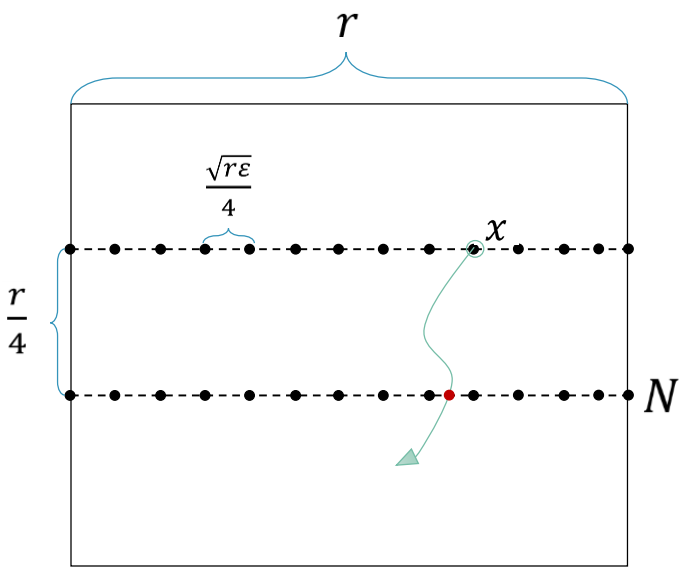


Parallel Plates Solution

- Suppose that the gradient flow originating at the best point escapes the new domain.
- If there is no ε-stationary point, by the lemma

$$f(x) \geq \min\{f(y) | y \in N\} - \frac{r\varepsilon}{8} + \frac{r\varepsilon}{4}.$$

- This contradicts the minimality of x.



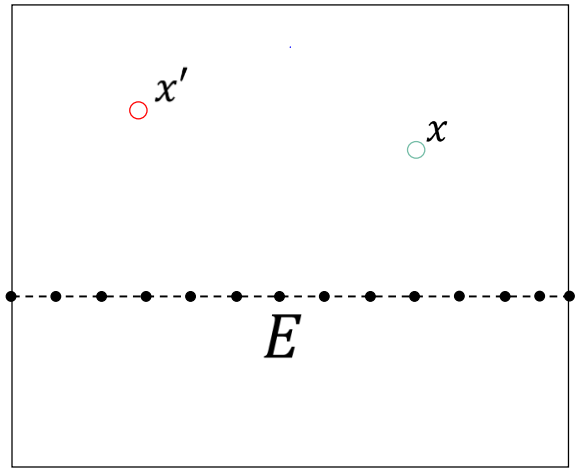
Parallel Plates Solution

- If E is the new edge of the domain.

- We also know that if x' is any other point in the domain with

$f(x') \leq f(x)$ and $\text{dist}(x', E) \geq \text{dist}(x, E)$

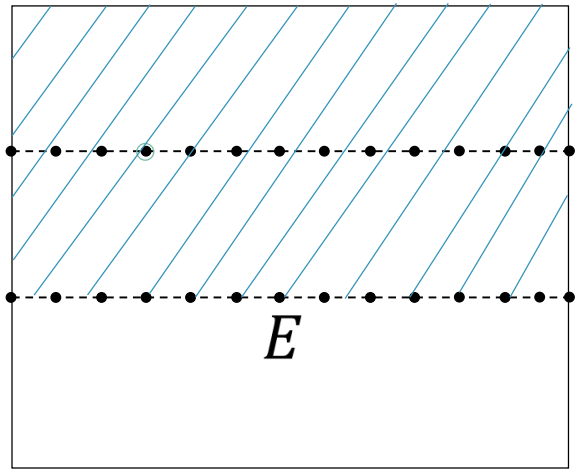
then the gradient flow originating at x' must contain an ε-stationary point.



Parallel Plates Solution

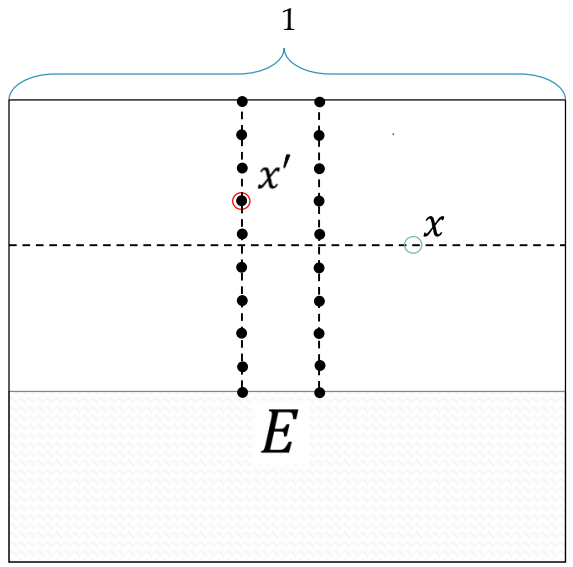
- The number of queries is $O\left(\frac{r}{\sqrt{r\varepsilon}}\right) = O\left(\sqrt{\frac{r}{\varepsilon}}\right)$, and we've removed a constant fraction of the domain.

- Hopefully, by iterating the construction we could find an ε-stationary point with $O\left(\frac{1}{\sqrt{\varepsilon}}\right)$ queries.



Further steps

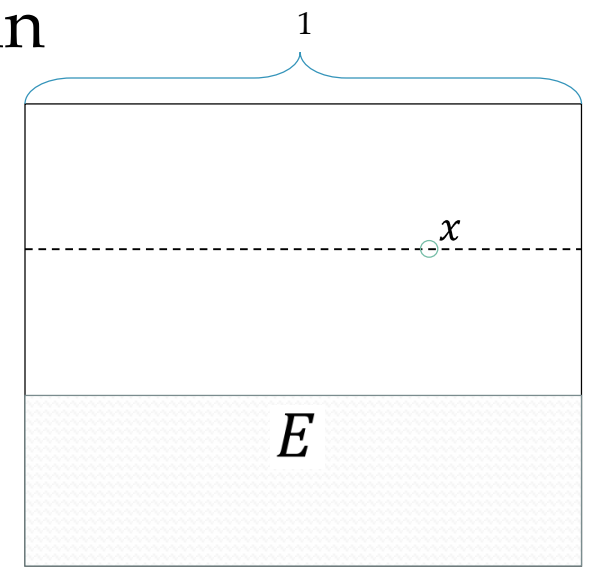
- When iterating, let x' be the 'best point' of the new net.
- We will only update to x', if $f(x') \leq f(x)$.
- Otherwise, we'll keep the side at which x lies.



A Problem

- A the beginning $r = 1$ and we know the domain contains a stationary point.

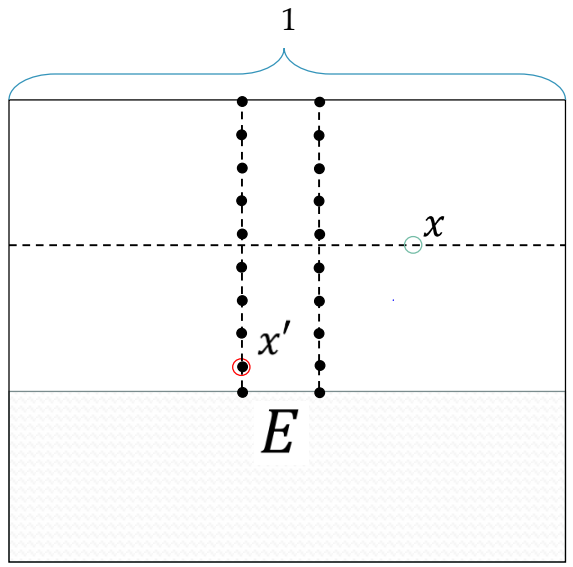
- After cutting, we only know the new domain contains an an ε-stationary point.



A Problem

- It could be that the new point we choose is too close to the new edge.

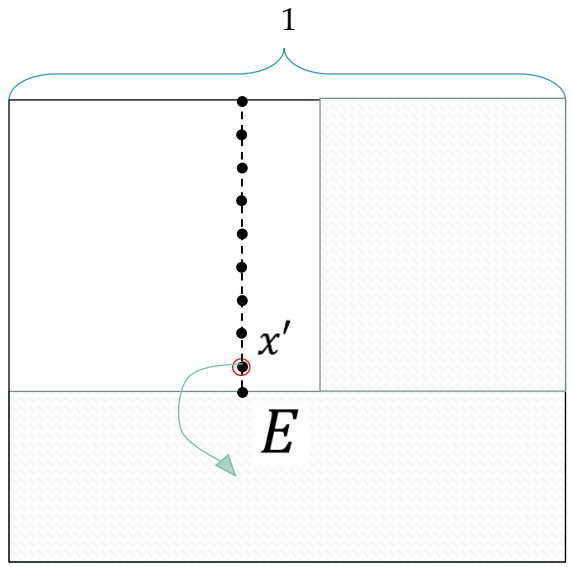
- We can make no guarantees in this case.



A Problem

- It could be that the new point we choose is too close to the new edge.

- We can make no guarantees in this case.



A Possible Solution

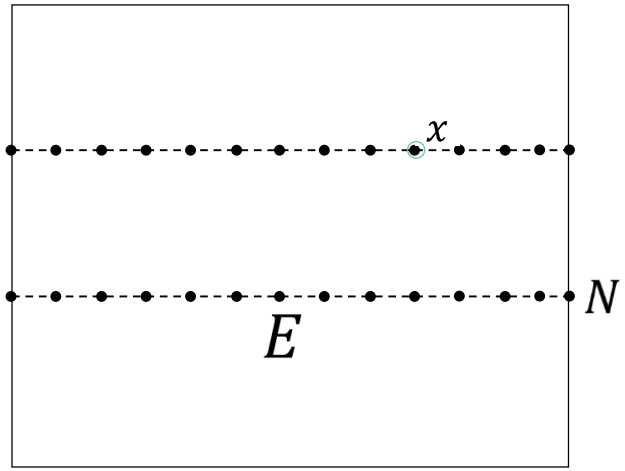
- At the end of the first step, we knew that

$$f(x) \leq \min\{f(y) | y \in N\},$$

where N is the queried net.

- Suppose we knew instead that

$$f(x) \leq \min\{f(y) | y \in N\} - \varepsilon \cdot \mathrm{dist}(x, E)$$

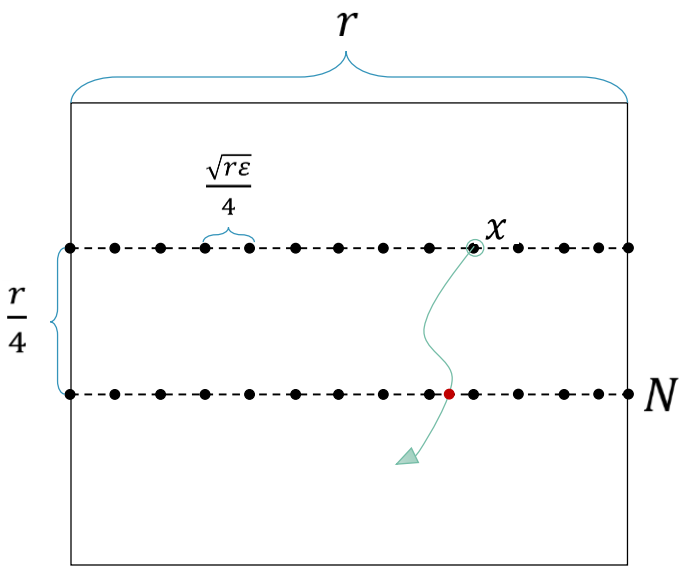


A Possible Solution

- In this case, the same calculation as before would show that:

  There is a **stationary** point, on the gradient flow originating at x. Otherwise

$$f(x) > \min\{\mathrm{f(y)}|y \in N\} - \frac{r\varepsilon}{8}.$$



- Which contradicts

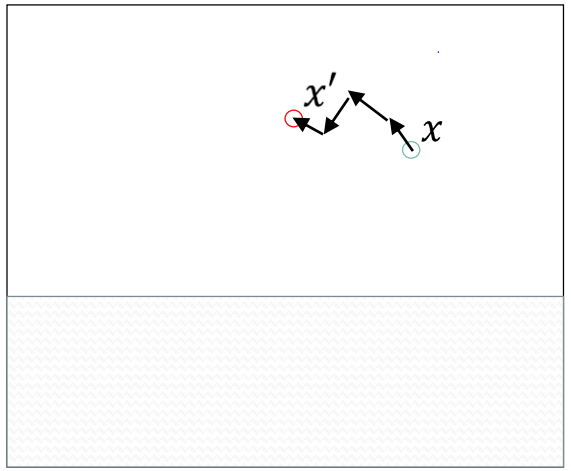
$$f(x) \leq \min\{\mathrm{f(y)}|\mathrm{y} \in N\} - \varepsilon \cdot \mathrm{dist}(x, E)$$
$$= \min\{\mathrm{f(y)}|\mathrm{y} \in N\} - \frac{r\varepsilon}{4}$$

Parallel Plates Solution

- It would be enough to find a point x' in the domain such that

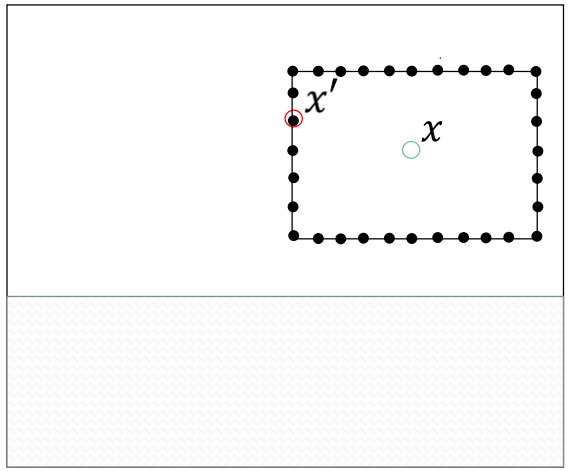
$$f(x') \leq f(x) - \varepsilon \cdot \text{dist}(x, E) \leq \min\{f(y) | y \in N\} - \frac{r\varepsilon}{4}.$$

- One possibility is to run gradient descent, starting from x.



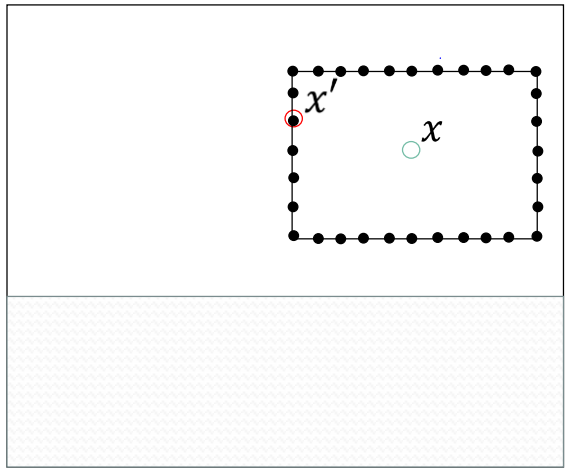
Global to Local

- A better, faster possibility, is to run the local improvement procedure, in order to find a point with a lower function value.



Global to Local

- Working out all the details gives an algorithm which finds an ε- stationary point with $O^*\left(\frac{1}{\sqrt{\varepsilon}}\right)$ queries.



Open Questions

- In dimension 3, each computation will consist of querying a $\sqrt{r\varepsilon}$-net of a square section of length r.
- This would require $O\left(\frac{r}{\varepsilon}\right)$ queries and the entire algorithm would use $O^*\left(\frac{1}{\varepsilon}\right)$ queries.
- This is better than Vavasis' $O\left(\frac{1}{\varepsilon^{\frac{6}{5}}}\right)$ queries algorithm.
- But misses the known lower bound of $O\left(\frac{1}{\varepsilon^{\frac{3}{5}}}\right)$.

Open Questions

- In higher dimensions the algorithm proposed by Vavasis is better.
- The number of queries required by the algorithm is always $\gg 1/\varepsilon$.
- In any fixed dimension the lower bound is $\ll 1/\varepsilon$.

Open Questions

- Is there an algorithm which finds an ε-stationary point with $O\left(\frac{f(d)}{\varepsilon}\right)$ queries?

- Or, can we find a fixed dimension in which any algorithm will require $\Omega\left(\frac{1}{\varepsilon}\right)$ queries?