Testable Sanitization via Boundary Geometry

Manolis Zampetakis

Yale University

Yang Cai

Katerina Mamali

Katerina Sotiraki

Zihan Yu

STOC 2026 - Workshop on Testing in the Modern World

Motivating Example – Loan Application

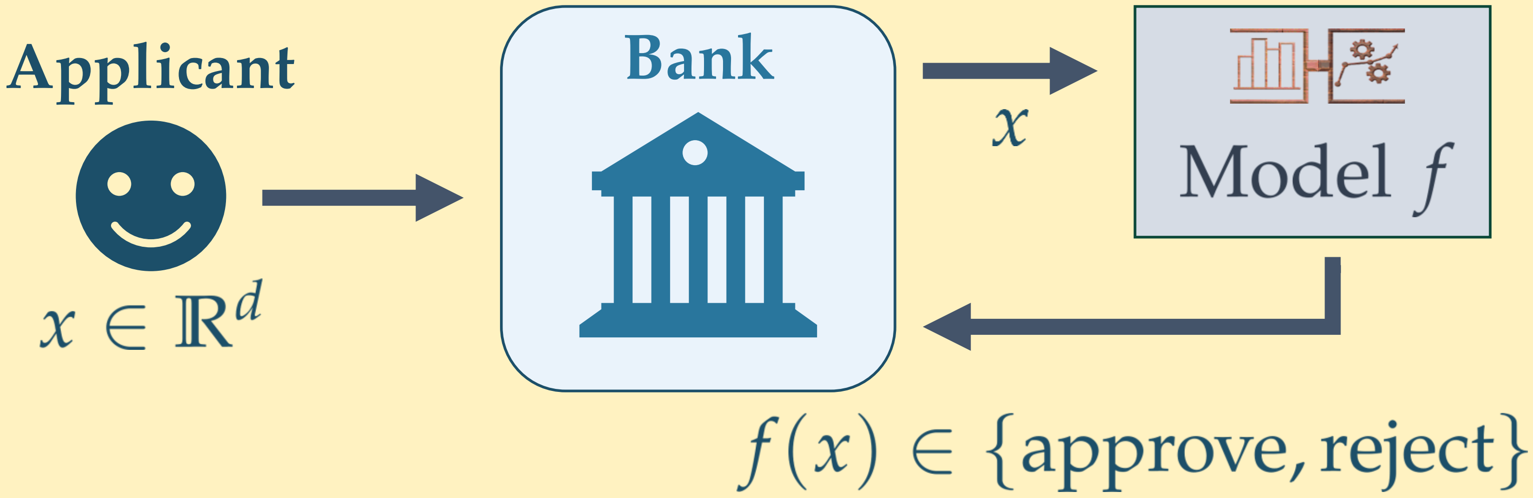
Bank
wants a classifier to for loan approvals

Model f

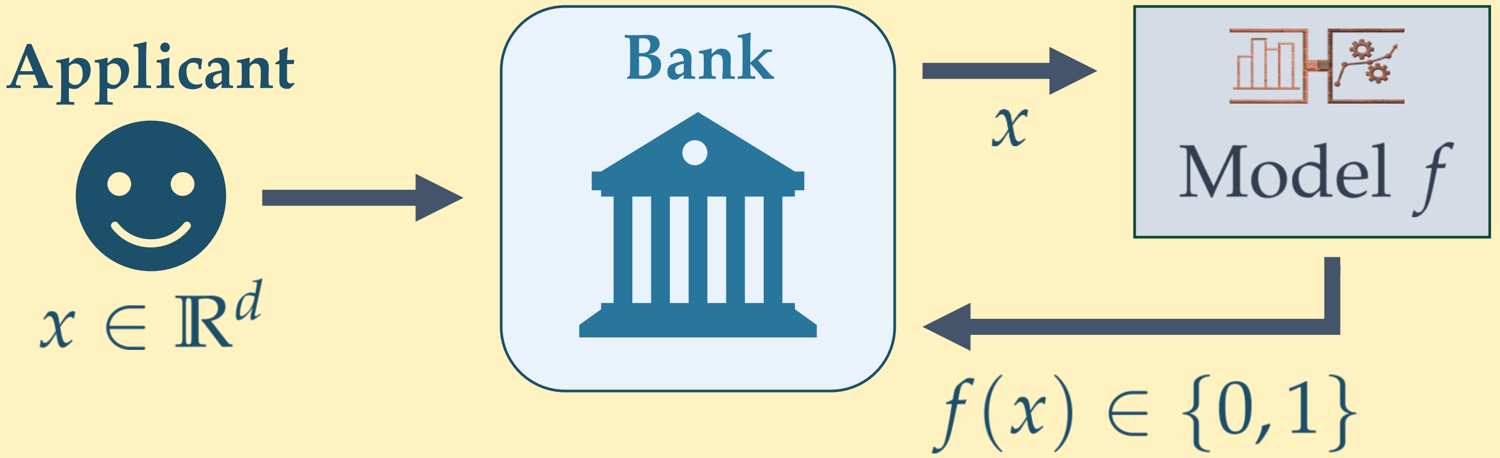
ML Company
trains classifier with good performance

The bank now has one question: is this model safe to deploy?

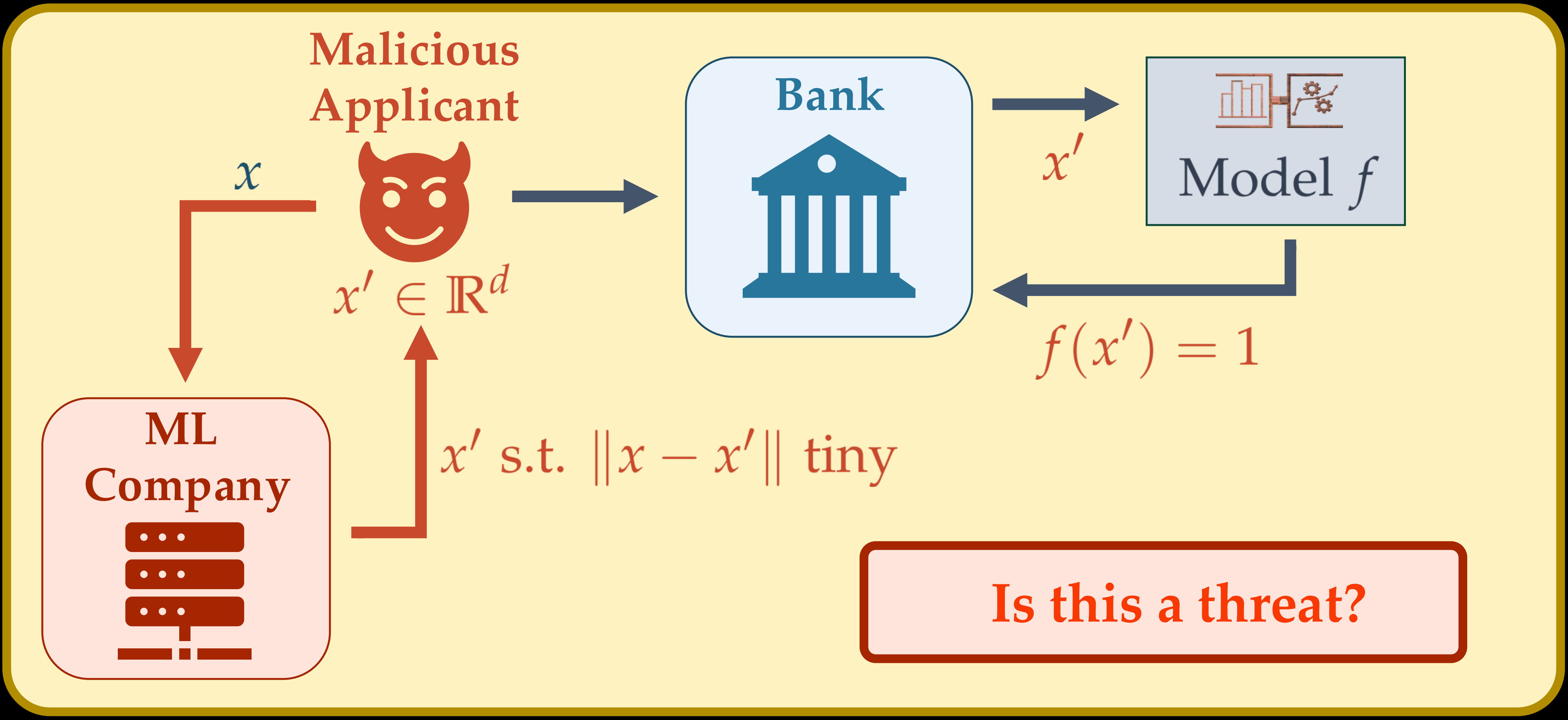
Motivating Example – Loan Application



Motivating Example – Loan Application



Motivating Example – Loan Application



Threat #1: Backdoors [Goldwasser Kim Vainkutanathan Zamir '22]

ML Company

Model f

Model $\tilde{f}$

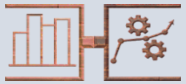
Indistinguishable!

for **all** x exists x'

s.t. $\|x - x'\|$ tiny and $f(x') = 1$

Threat #2: Adversarial Examples

[Szegedy Zaremba Sutskever Bruna Erhan Goodfellow Fergu ’13]

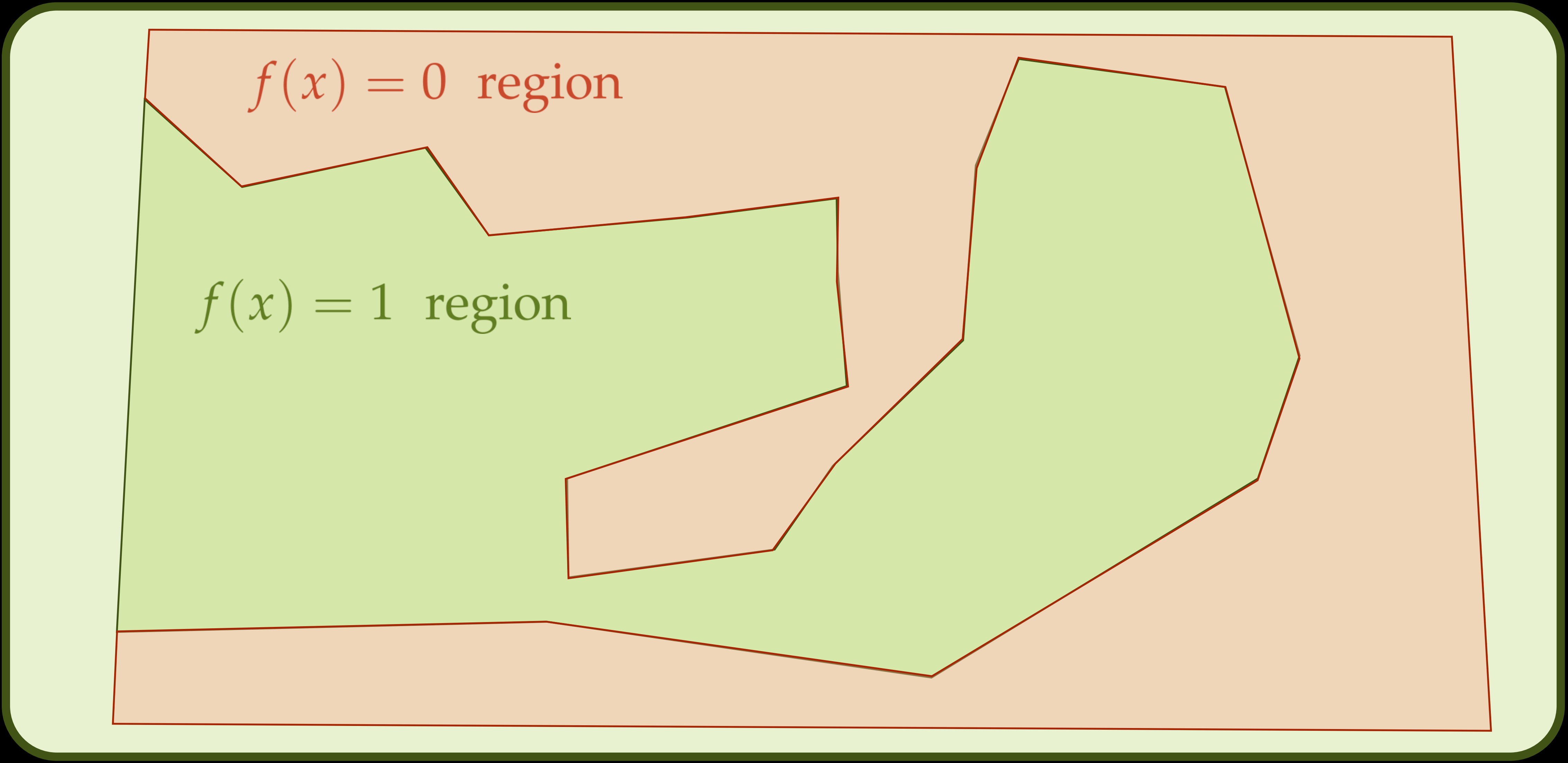


Model f

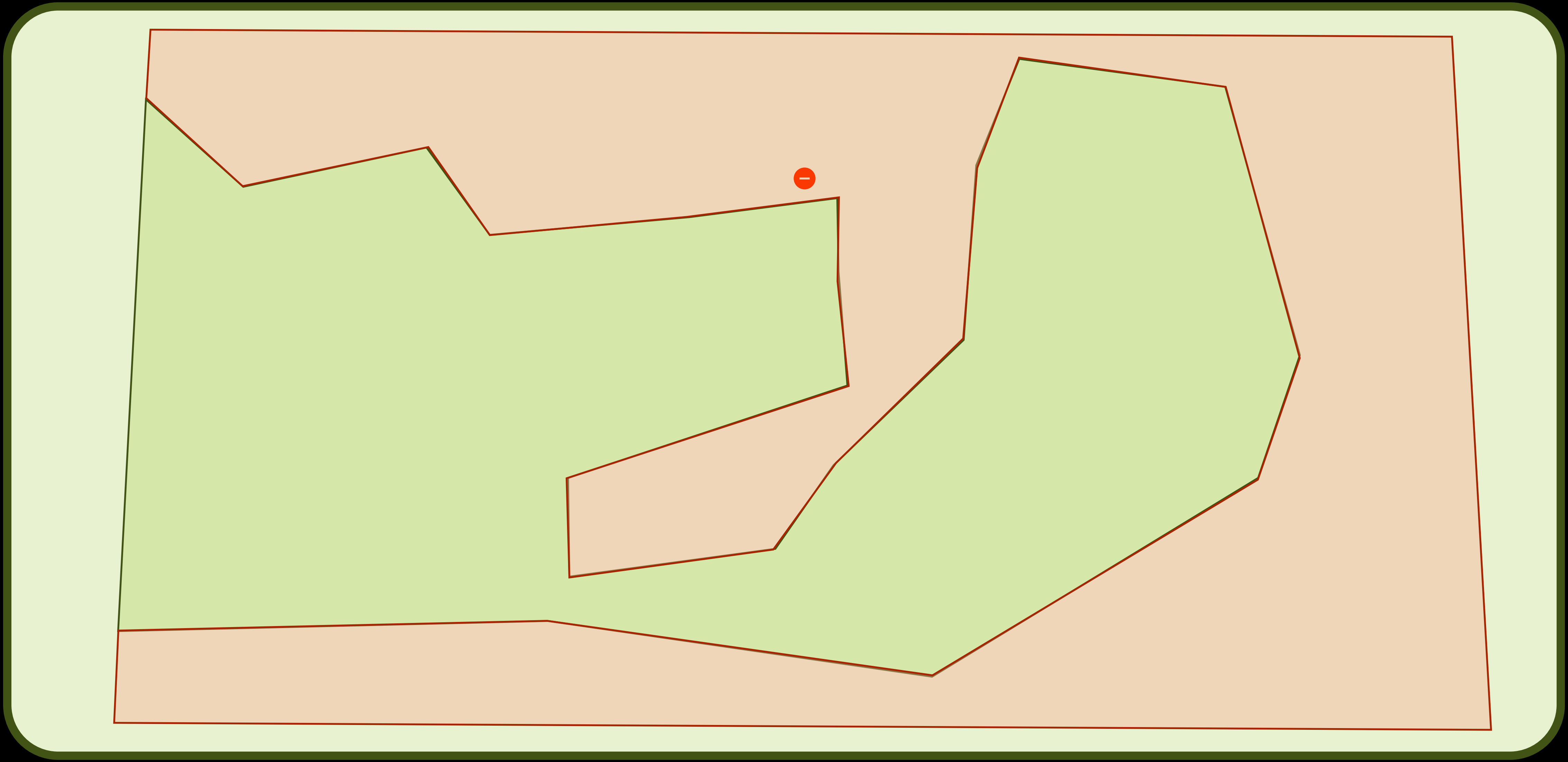
for most x exists x'

s.t. $\|x - x'\|$ tiny and $f(x') = 1$

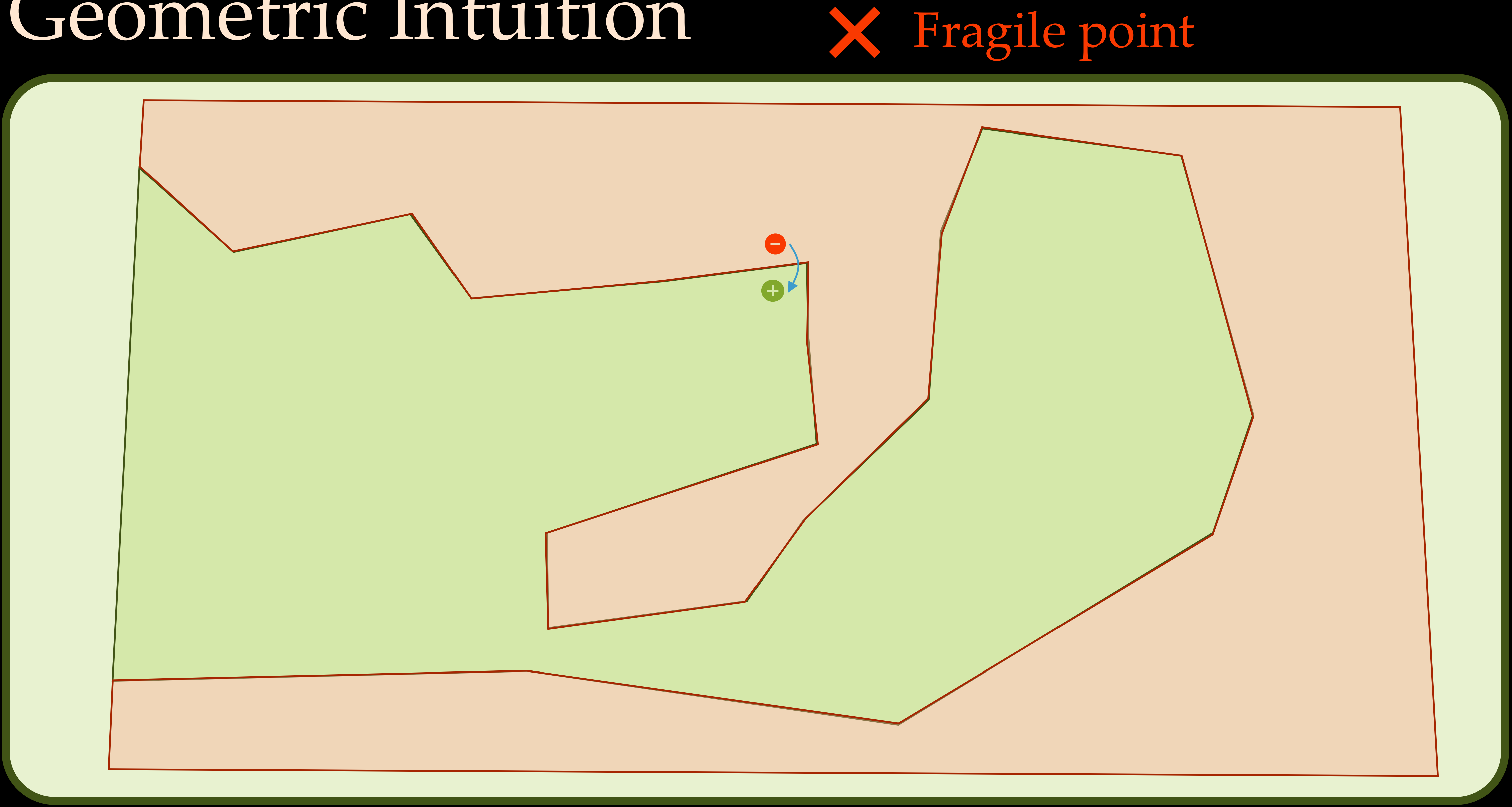
Geometric Intuition



Geometric Intuition

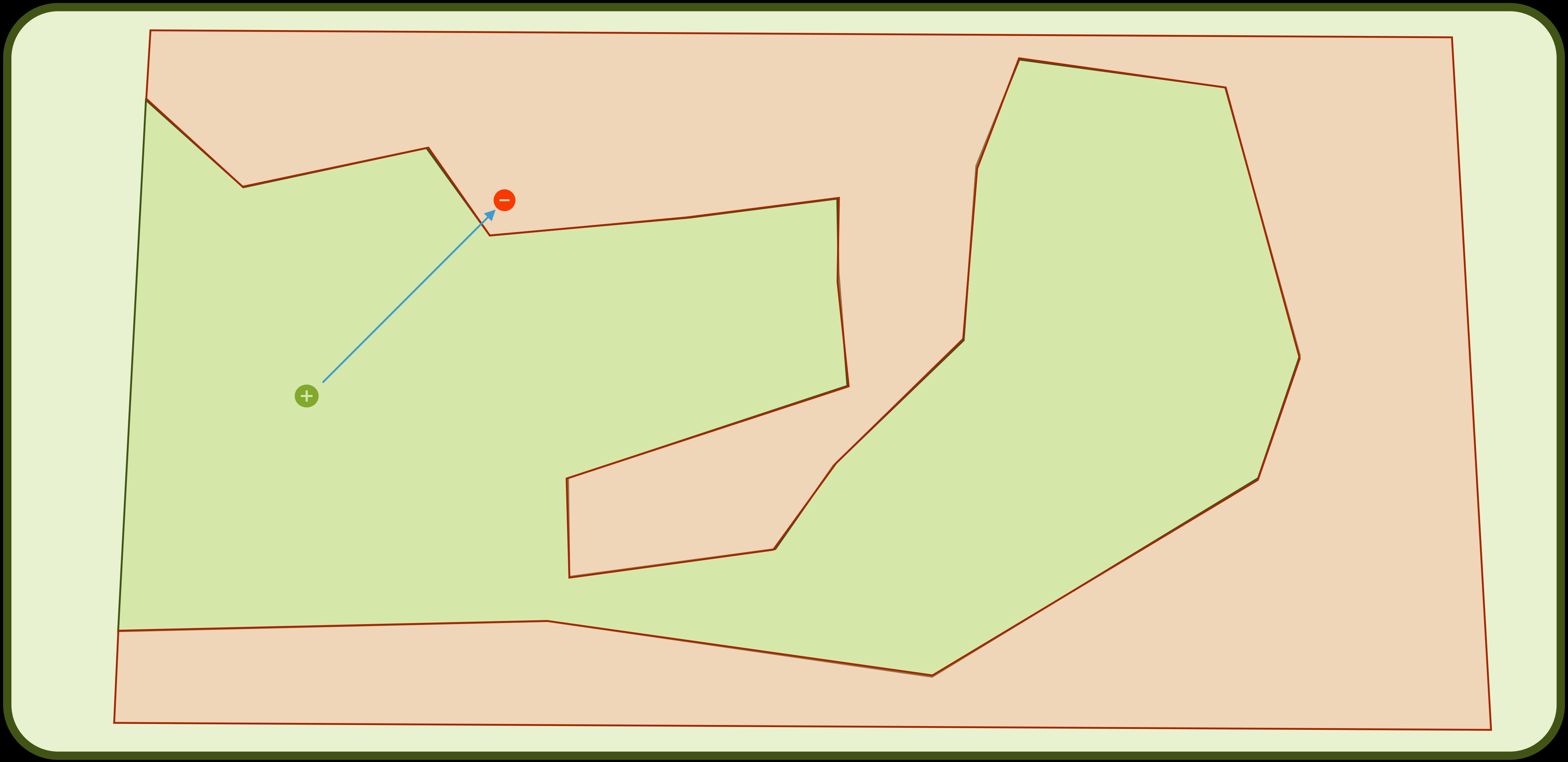


Geometric Intuition



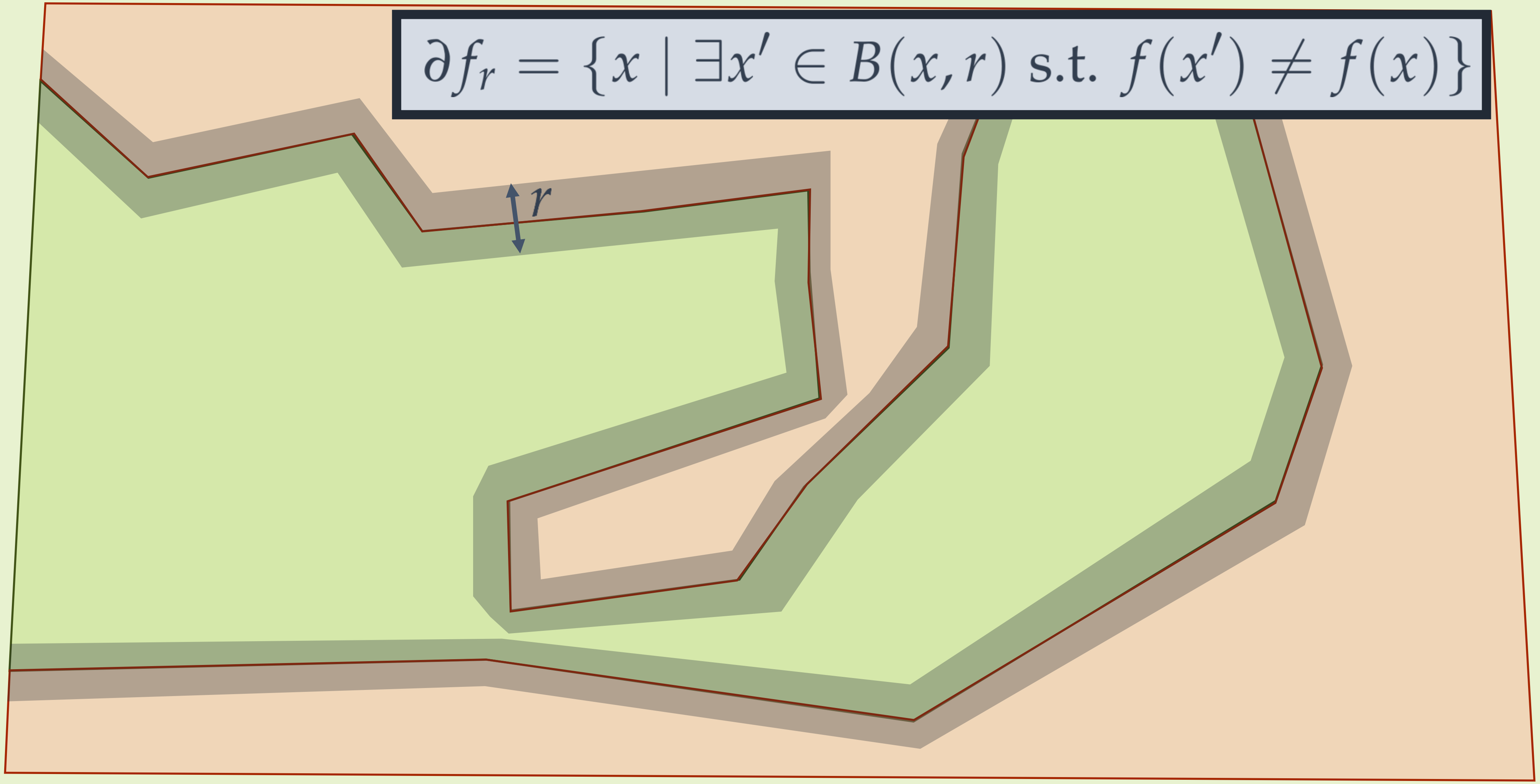
Geometric Intuition

✓ Robust point

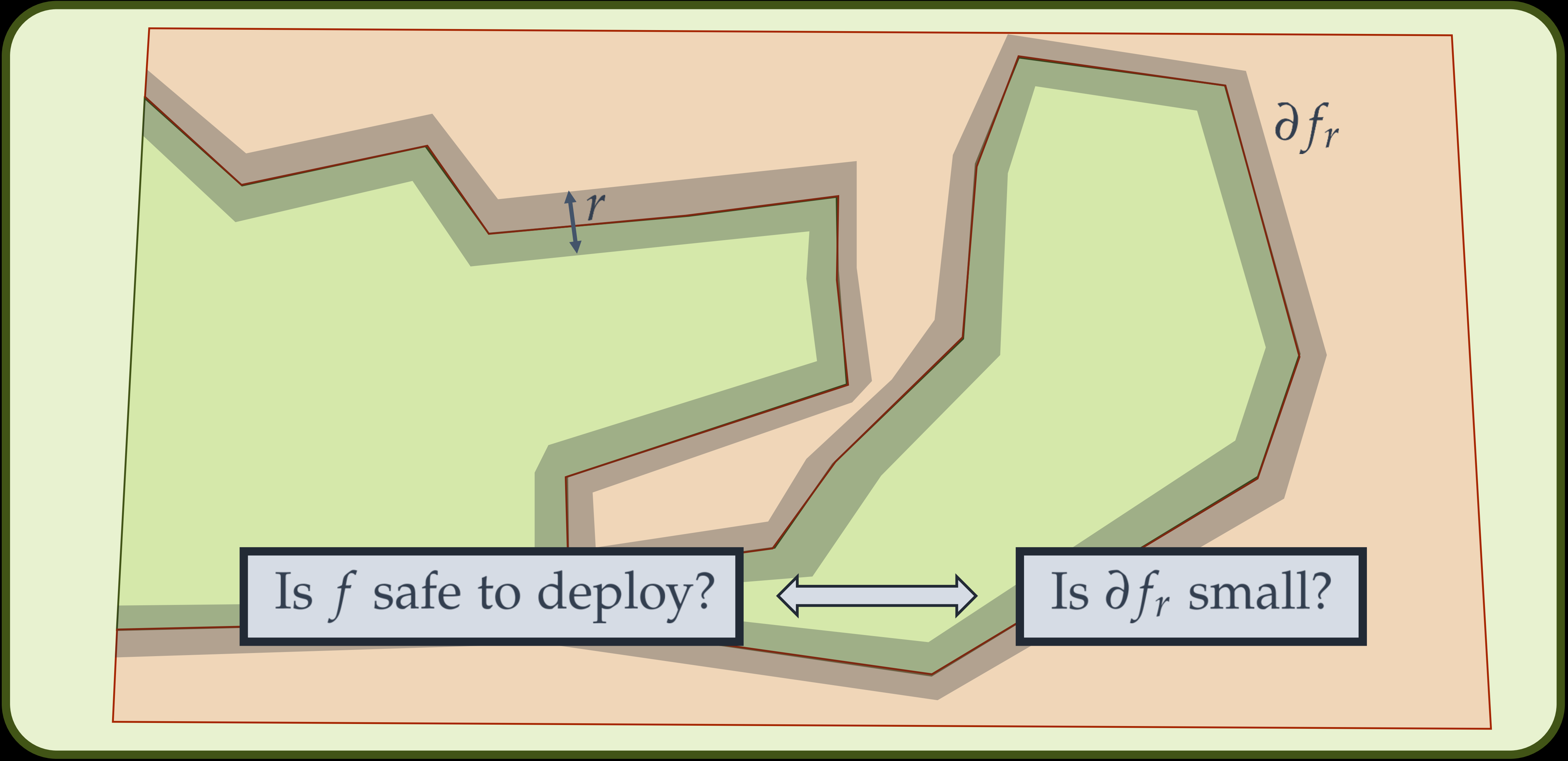


Geometric Intuition

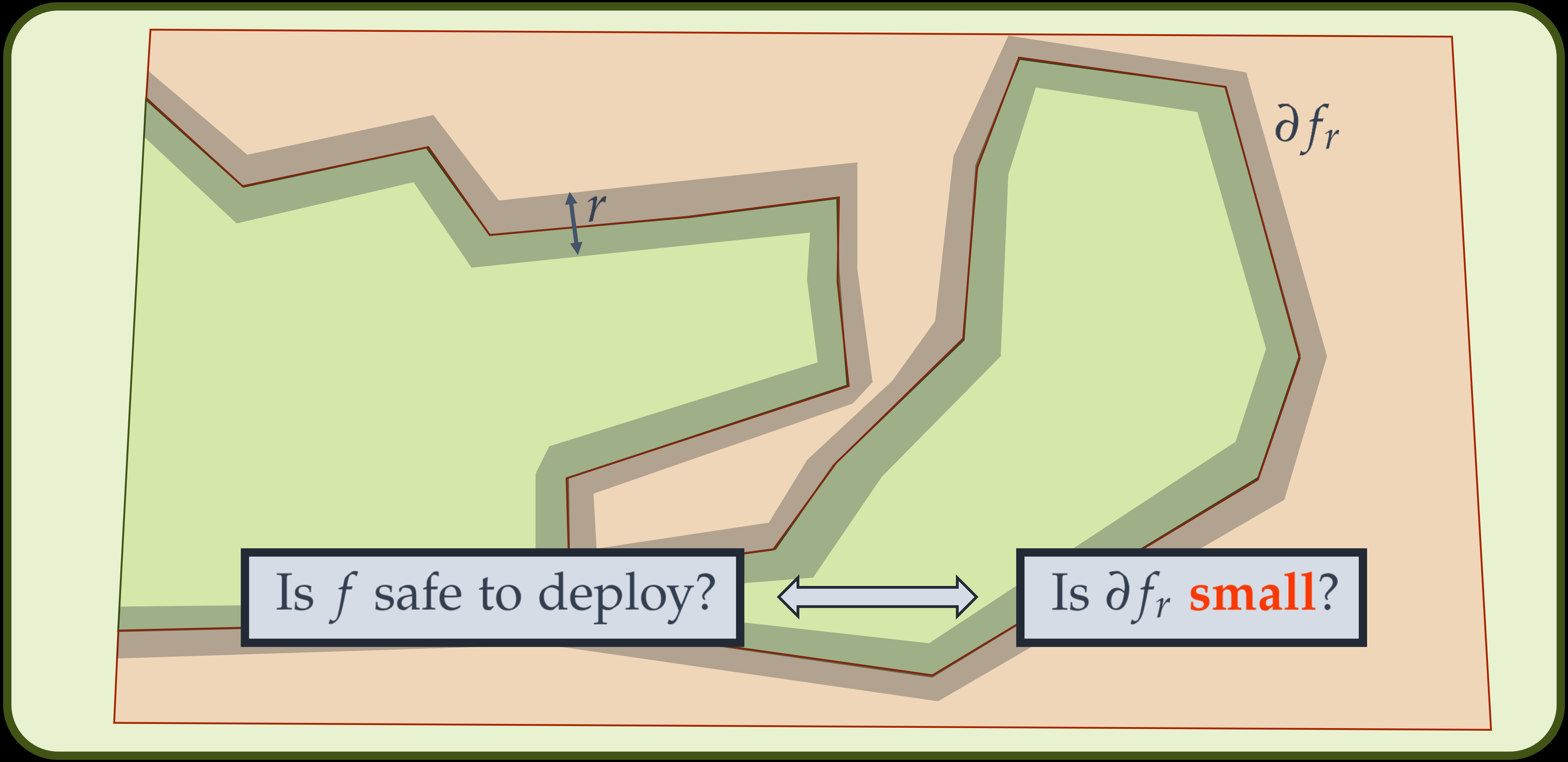
$$\partial f_r = \{x \mid \exists x' \in B(x,r) \text{ s.t. } f(x') \neq f(x)\}$$



Geometric Intuition

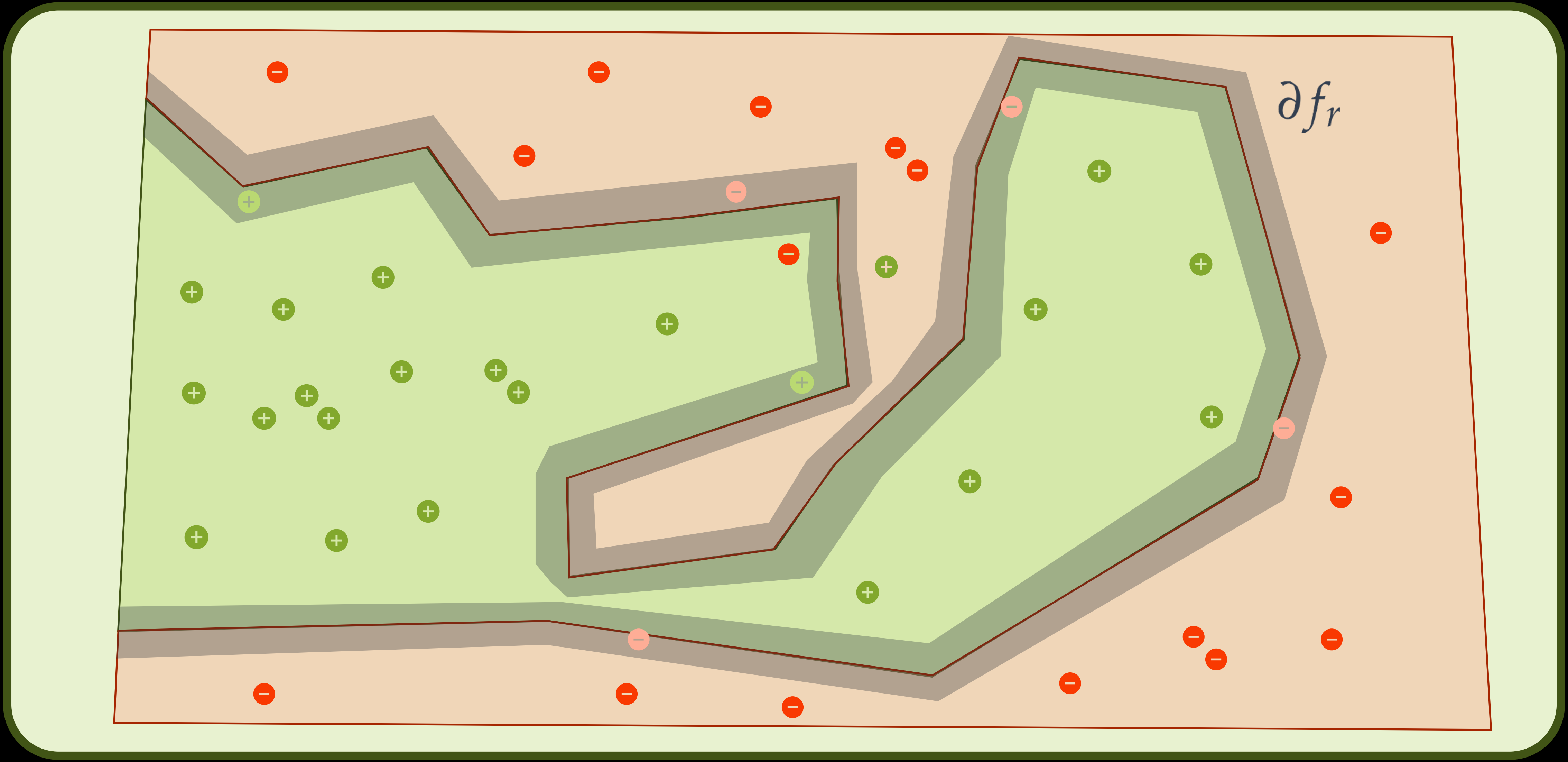


Geometric Intuition



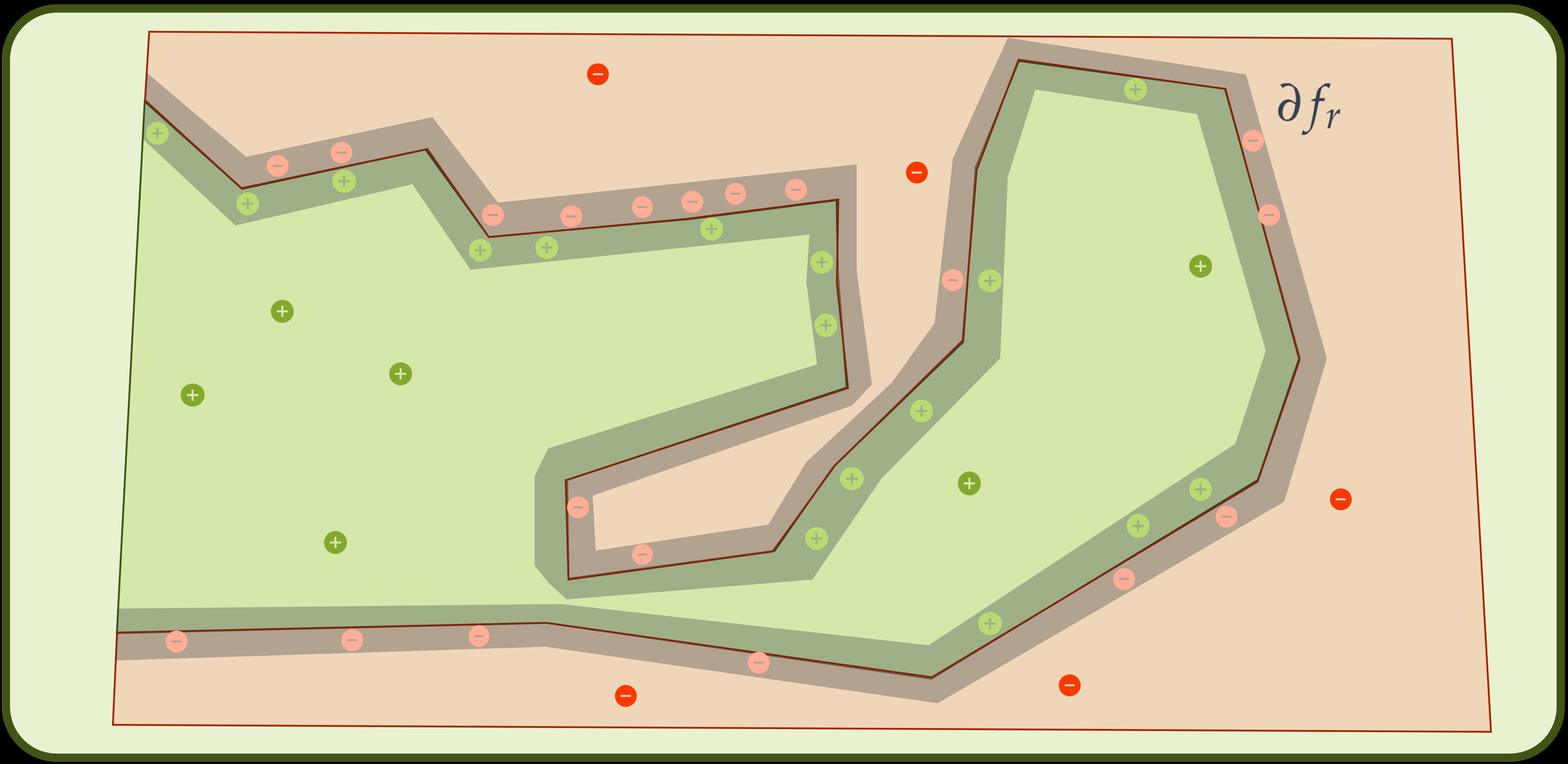
Geometric Intuition

✓ Good case



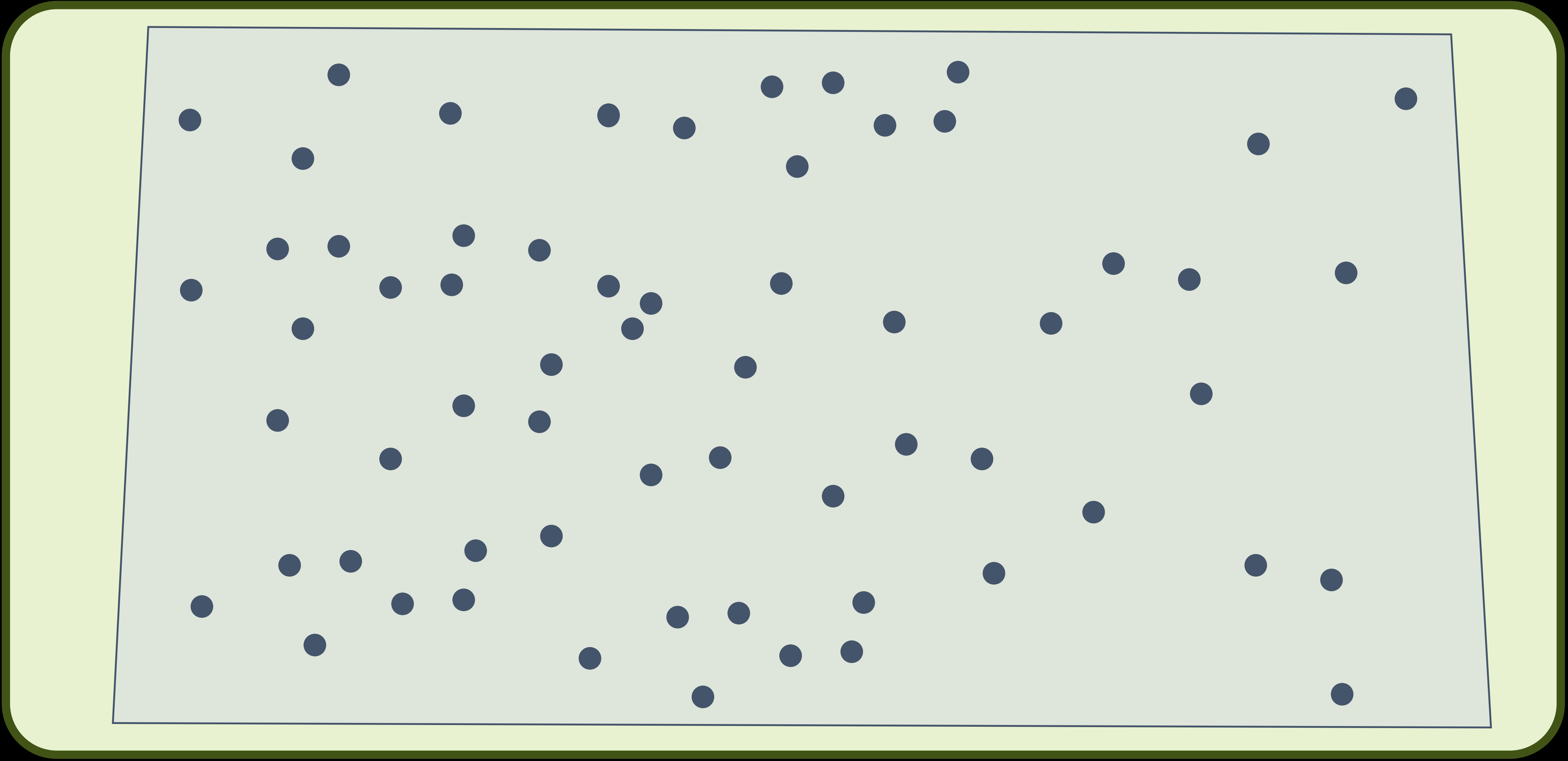
Geometric Intuition

✕ Bad case



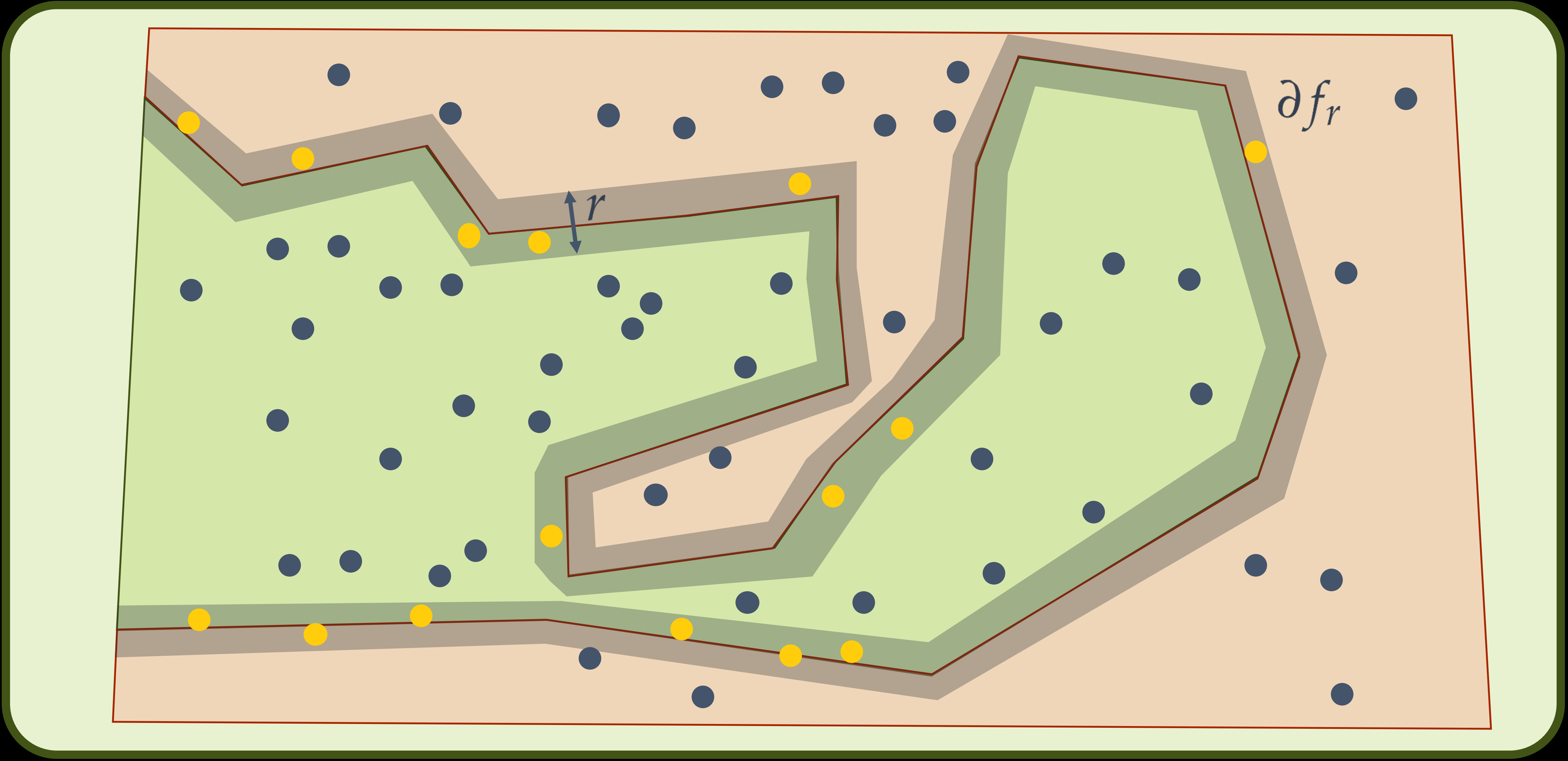
Geometric Intuition

Distribution $\mathcal{D}$ over $\mathbb{R}^d$.

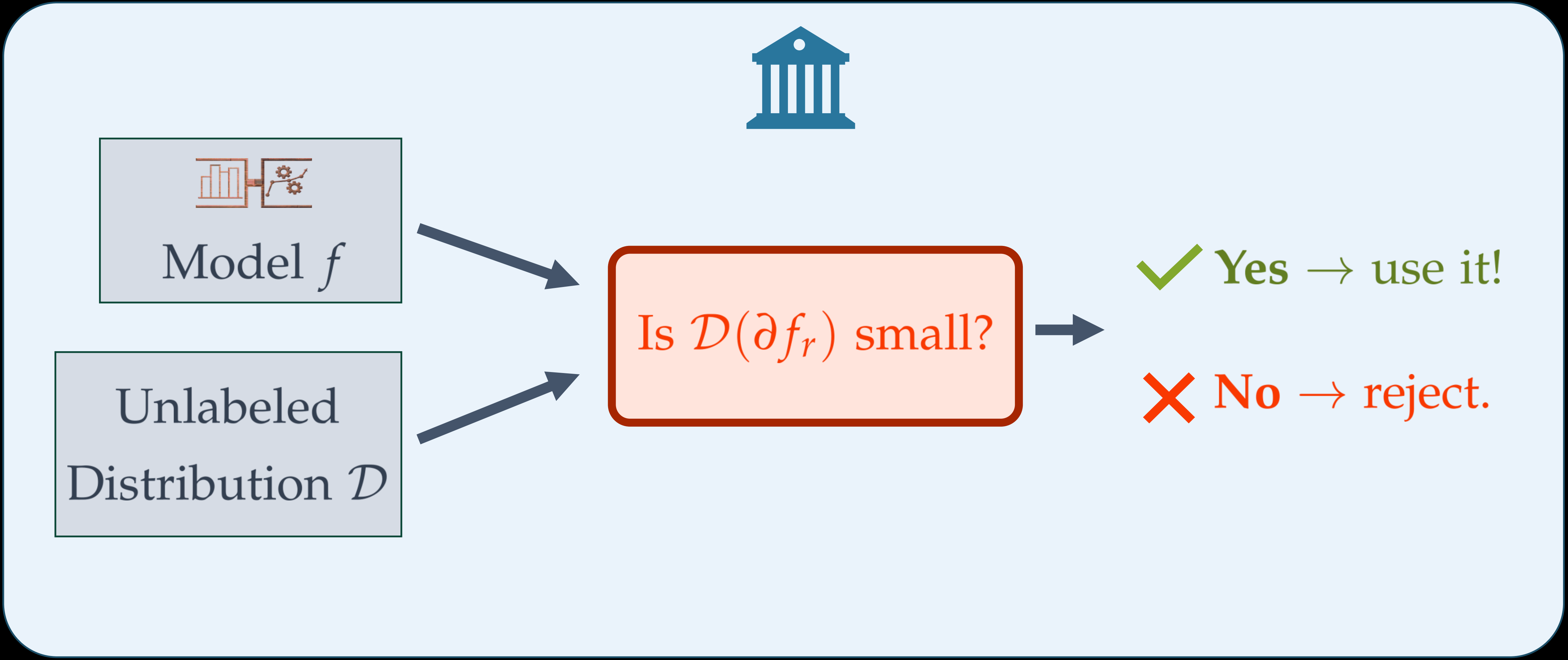


Geometric Intuition

Is $\mathcal{D}(\partial f_r)$ small?



Task #1: Test Size of Boundary



Task #1: Test Size of Boundary

Corollary [Goldwasser Kim Vainkutanathan Zamir '22]

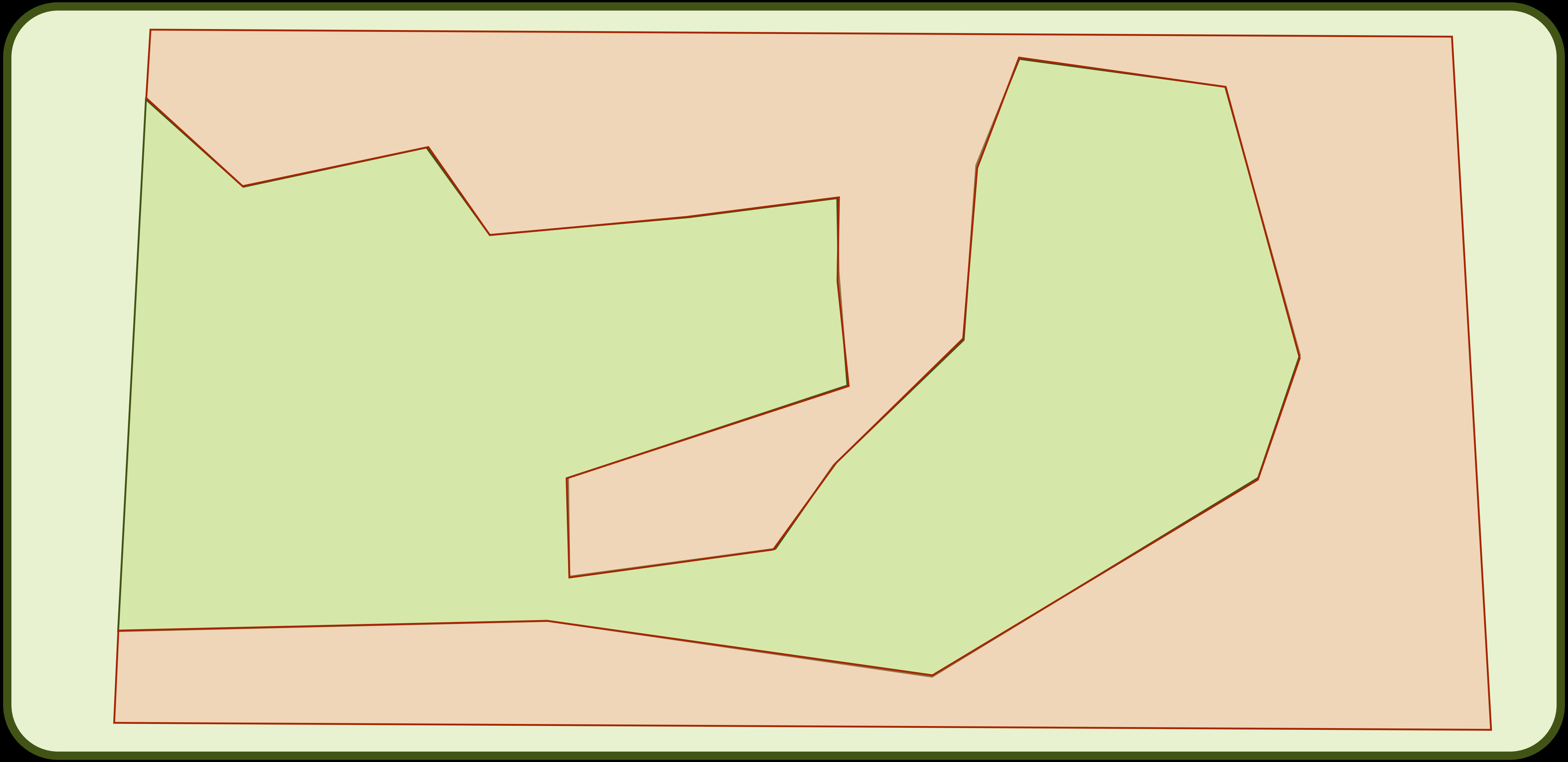
Testing the size of the boundary is an intractable problem.

Proof idea.

$\tilde{f}$ can be constructed using tools from cryptography to be computationally indistinguishable from f.

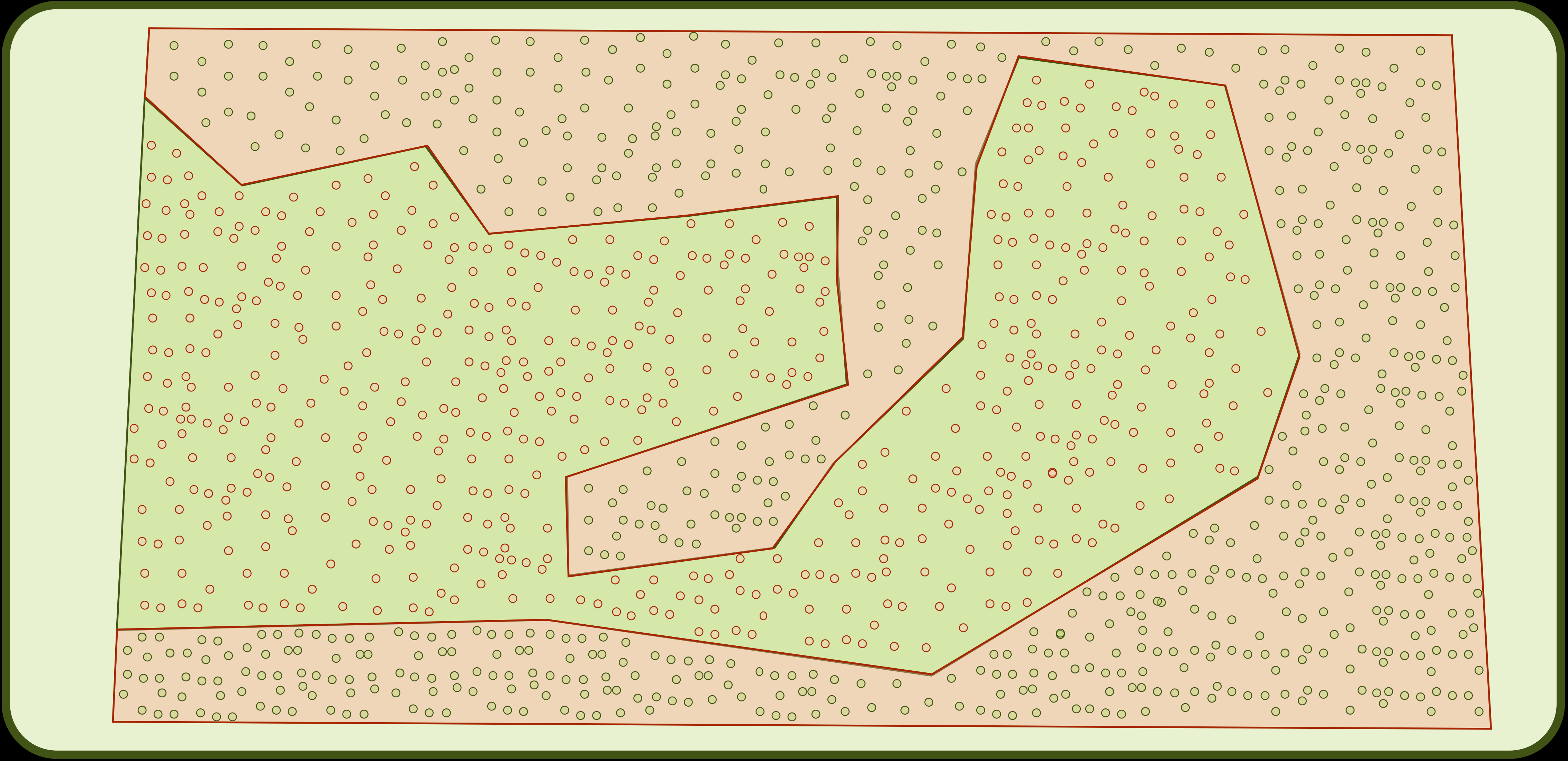
Geometric Intuition

Model f

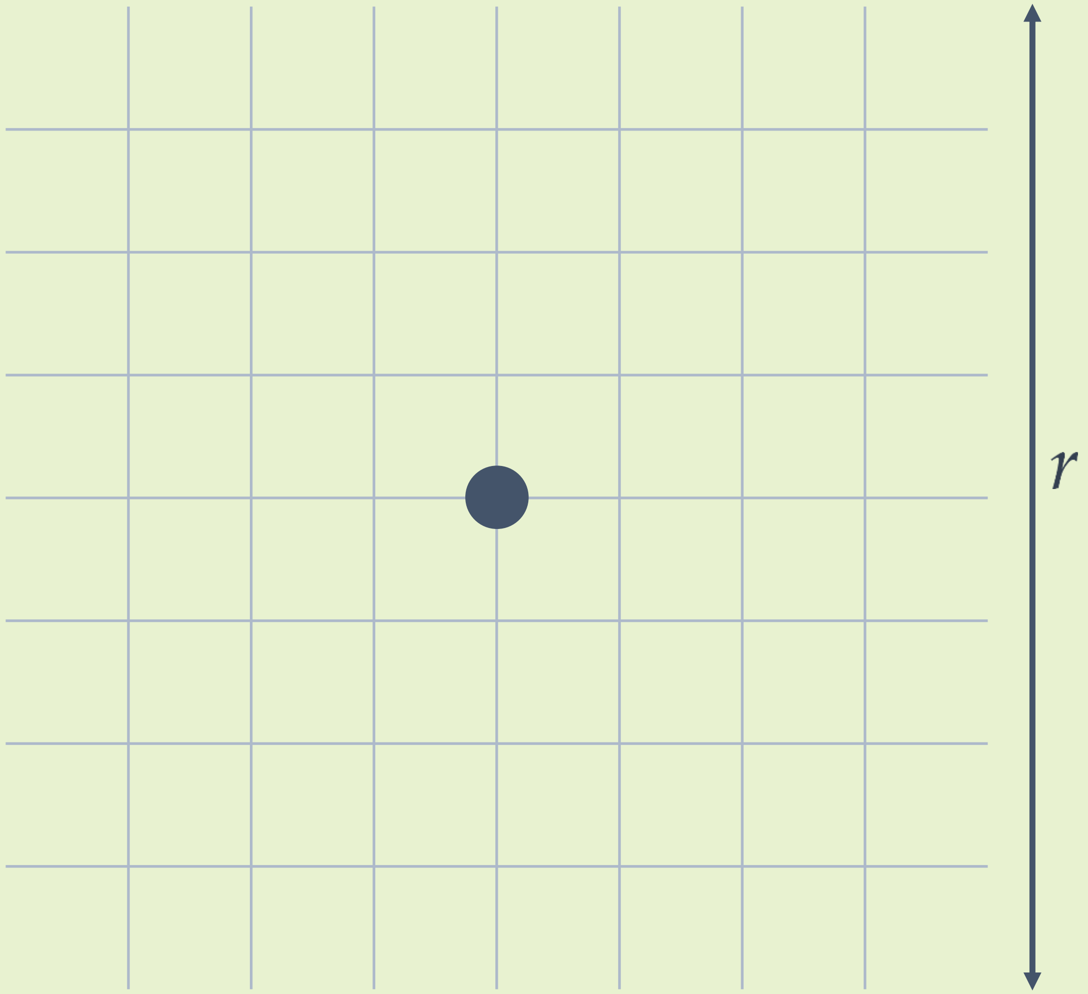


Geometric Intuition

Model $\tilde{f}$

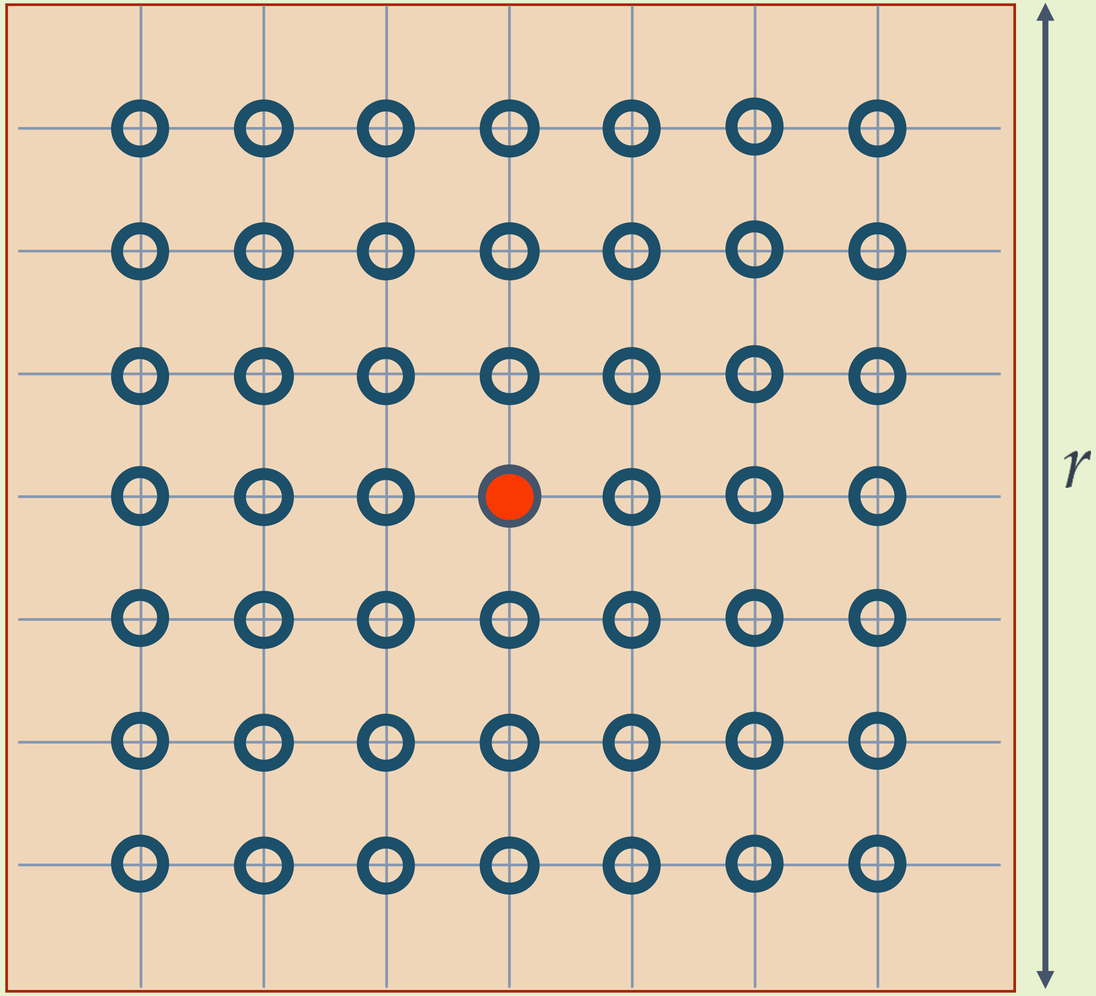


Hardness of Testing the Size of the Boundary



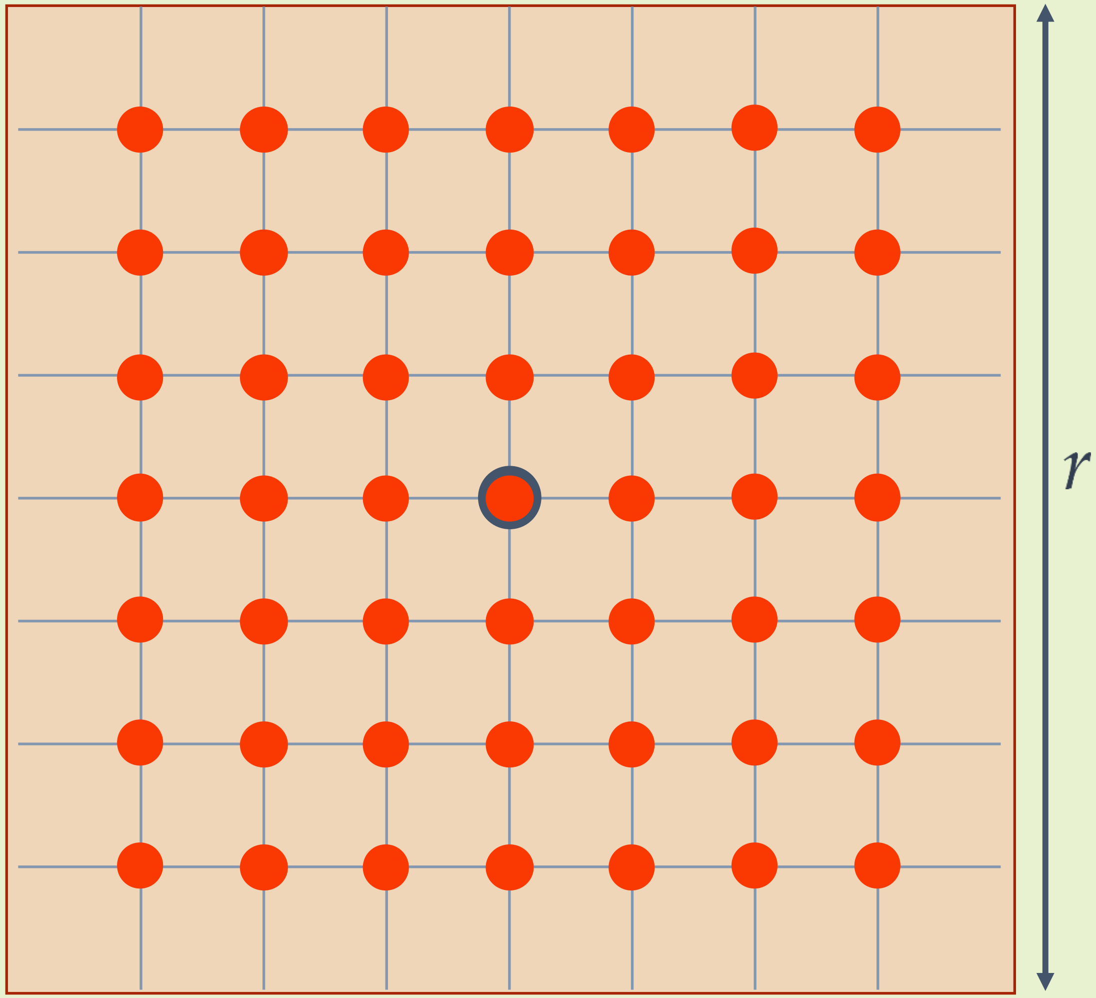
- $\mathcal{D}$: point mass at 0.

Hardness of Testing the Size of the Boundary



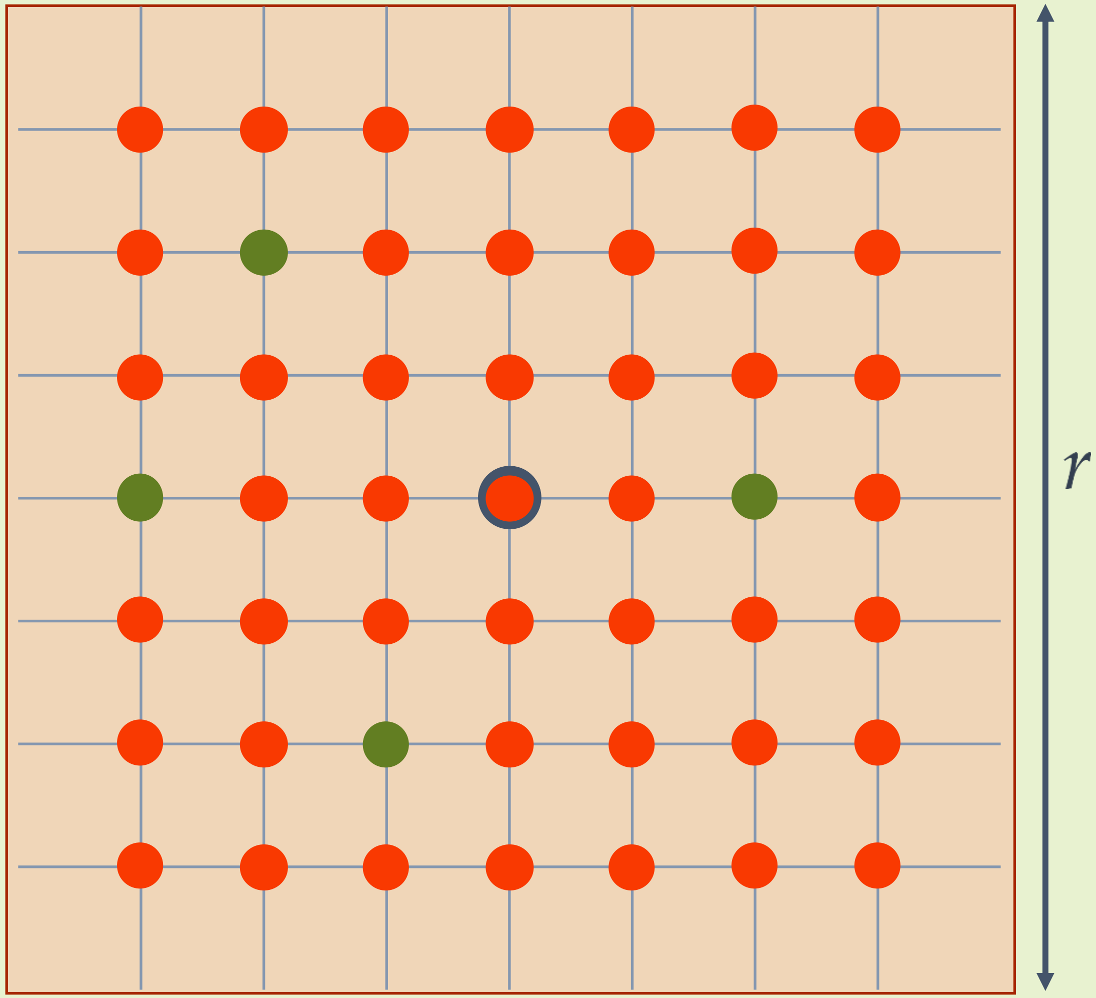
- $\mathcal{D}$: point mass at 0.
- $f(x) = \mathrm{C}(x)$ where C is a circuit for x on the grid and 0 otherwise.

Hardness of Testing the Size of the Boundary



- $\mathcal{D}$: point mass at 0.
- $f(x) = C(x)$ where C is a circuit for x on the grid and 0 otherwise.
- $\mathcal{D}(\partial_r f) = 0$ if $C(x) = 0$ for all x.

Hardness of Testing the Size of the Boundary



- $\mathcal{D}$: point mass at 0.
- $f(x) = C(x)$ where C is a circuit for x on the grid and 0 otherwise.
- $\mathcal{D}(\partial_r f) = 0$ if $C(x) = 0$ for all x.
- $\mathcal{D}(\partial_r f) = 1$, otherwise.

Task #1: Test Size of Boundary

Observation 1

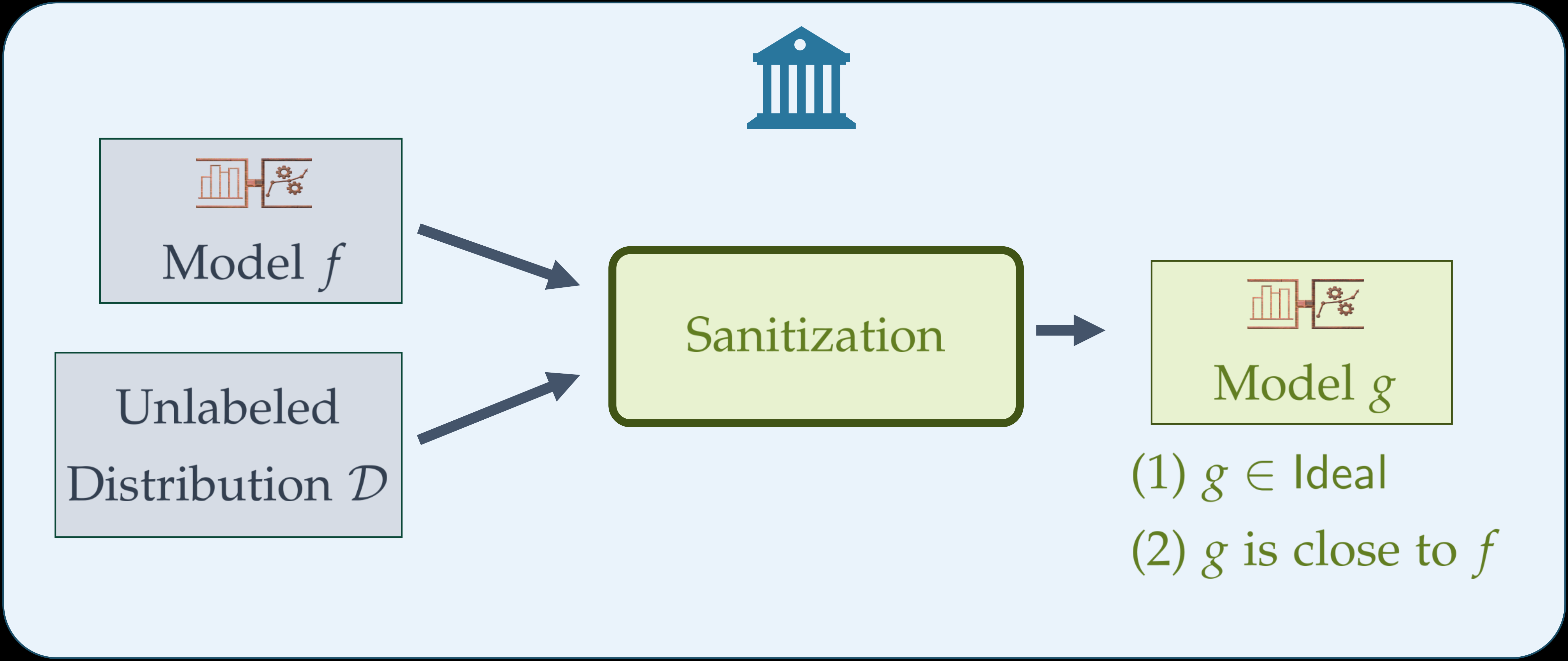
Testing the size of the boundary is coNP-complete.

Observation 2

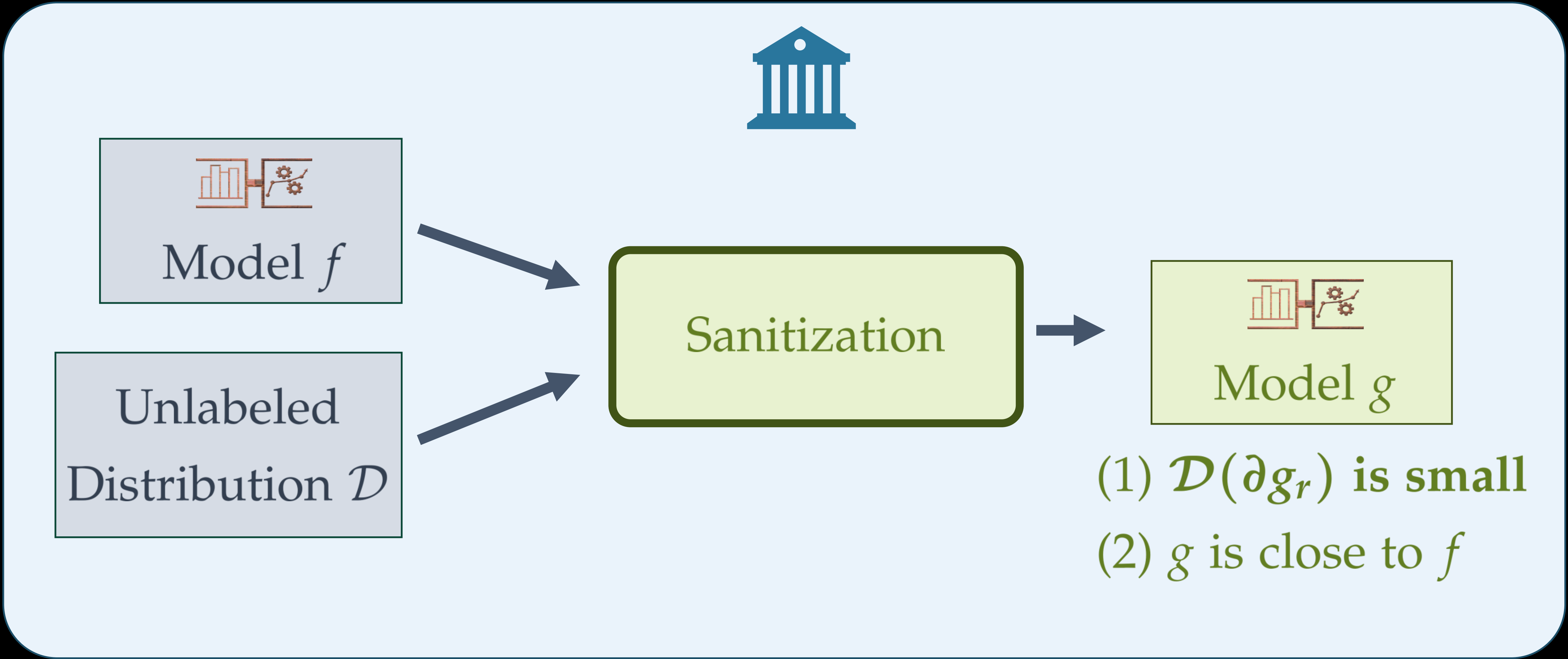
If we only have oracle access to f, testing the size of the boundary requires exponential in d number of queries.

Task #2: Sanitization

[Goldwasser Shafer Vafa Vainkutanathan ‘25]



Task #2: Sanitization



Task #2: Sanitization

Lemma [Cai Mamali Sotiraki Yu Zampetakis]

Sanitization is not always possible.

Proof. There are models f that are far from any model g with small $\mathcal{D}(\partial g_r)$.

Testable Sanitization

Inspired by Testable Learning by [Rubinfeld Vasilyan ‘23]

Testable Sanitization

Definition

g is (r, η)-robust if $\mathcal{D}(\partial g_r) \leq \eta$.

Testable Sanitization

Definition

g is (r, η)-robust if $\mathcal{D}(\partial g_r) \leq \eta$.

Let $\mathcal{F}(r, \eta)$ the set of (r, η)-robust models.

Testable Sanitization

Definition

g is (r, η)-robust if $\mathcal{D}(\partial g_r) \leq \eta$.

Let $\mathcal{F}(r, \eta)$ the set of (r, η)-robust models.

Definition

f is ε-close to g if $\mathbb{P}_{\mathcal{D}}(f(x) \neq g(x)) \leq \varepsilon$.

Testable Sanitization

$$(r, \quad \eta, \quad \varepsilon)$$

Testable Sanitization

$$(r, \quad \eta, \quad \varepsilon)$$

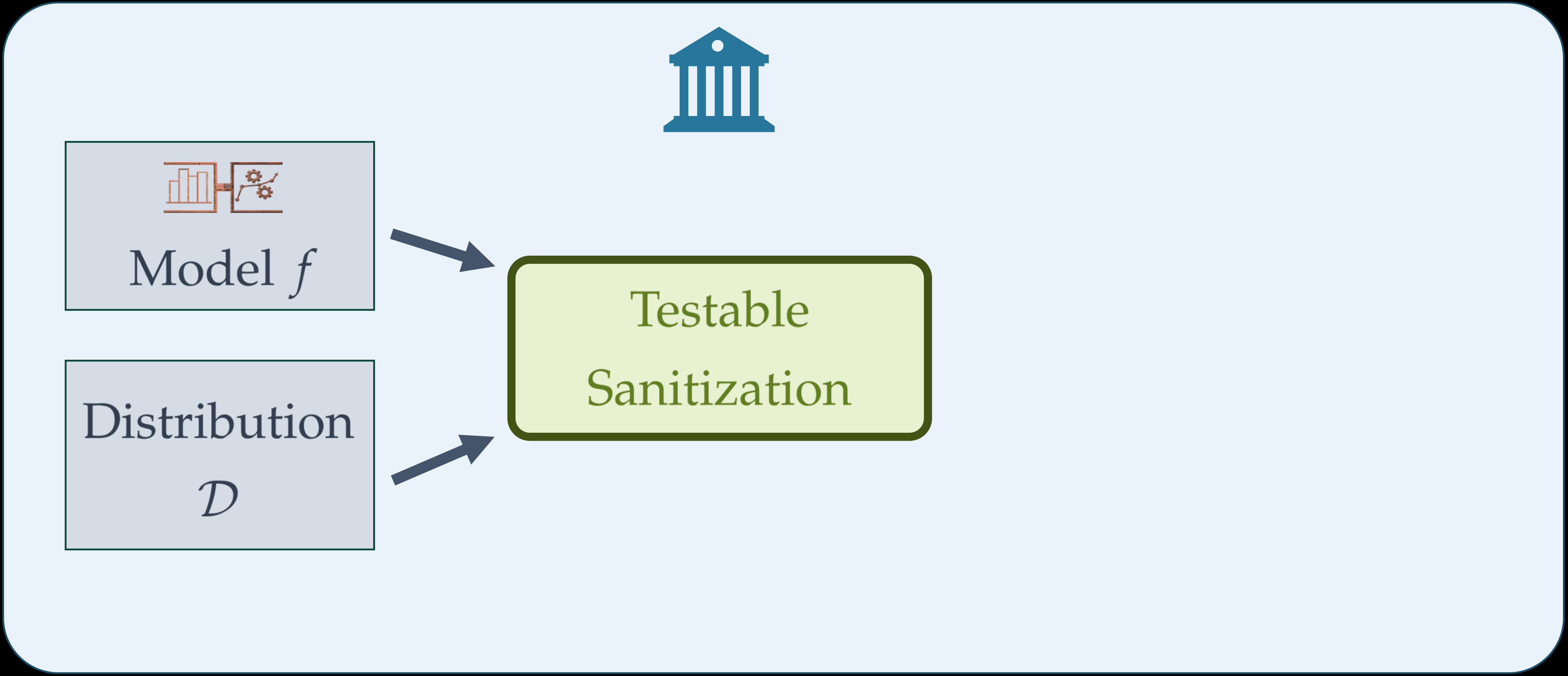
Adversarial loss

Testable Sanitization

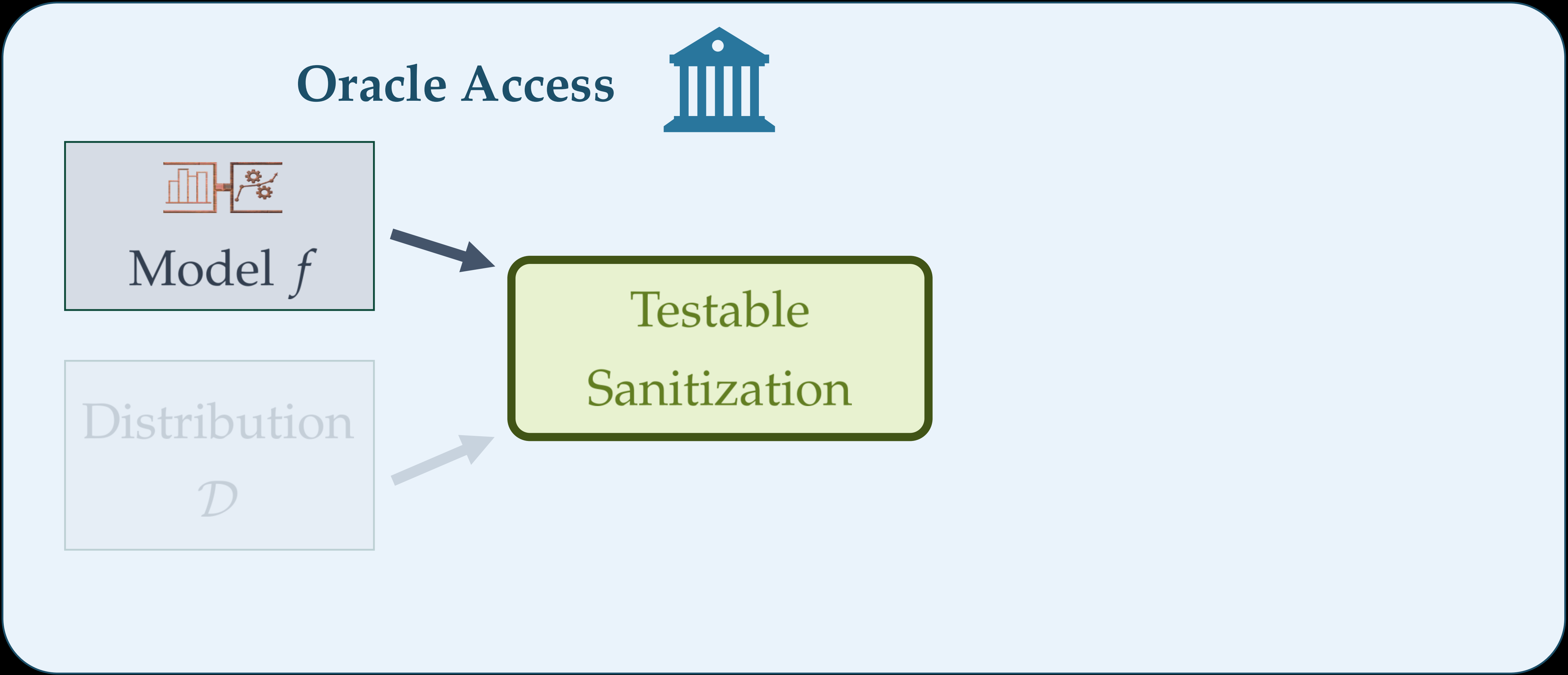
$$(r, \quad \eta, \quad \varepsilon)$$

Honest loss

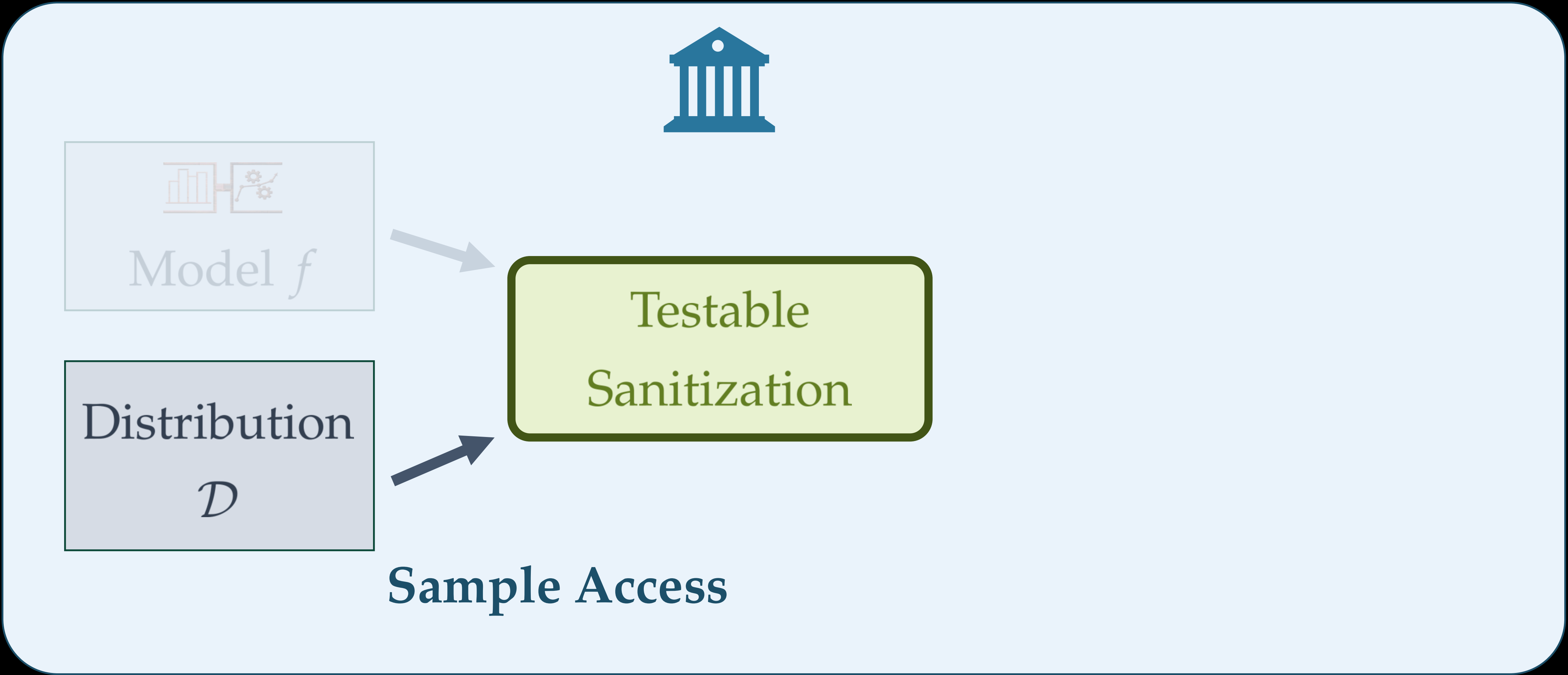
Testable Sanitization



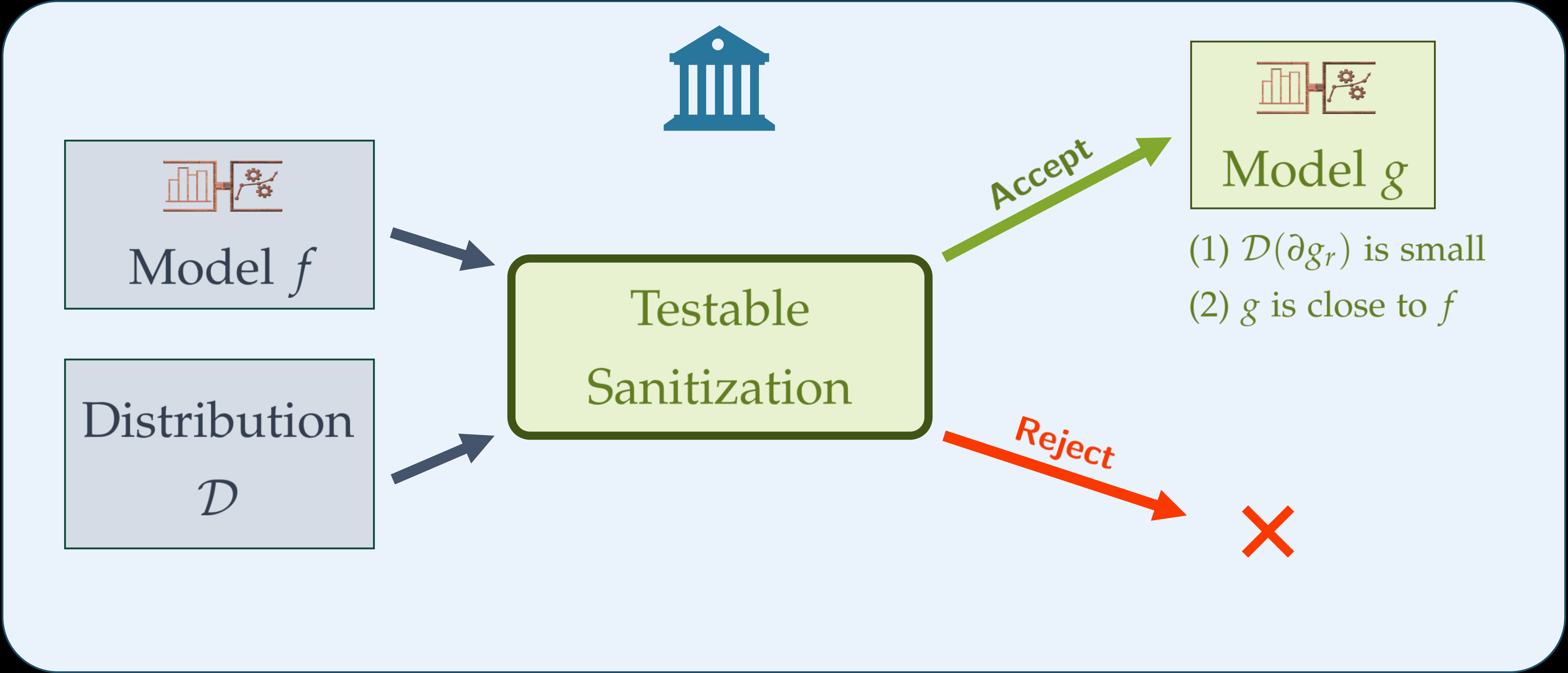
Testable Sanitization



Testable Sanitization

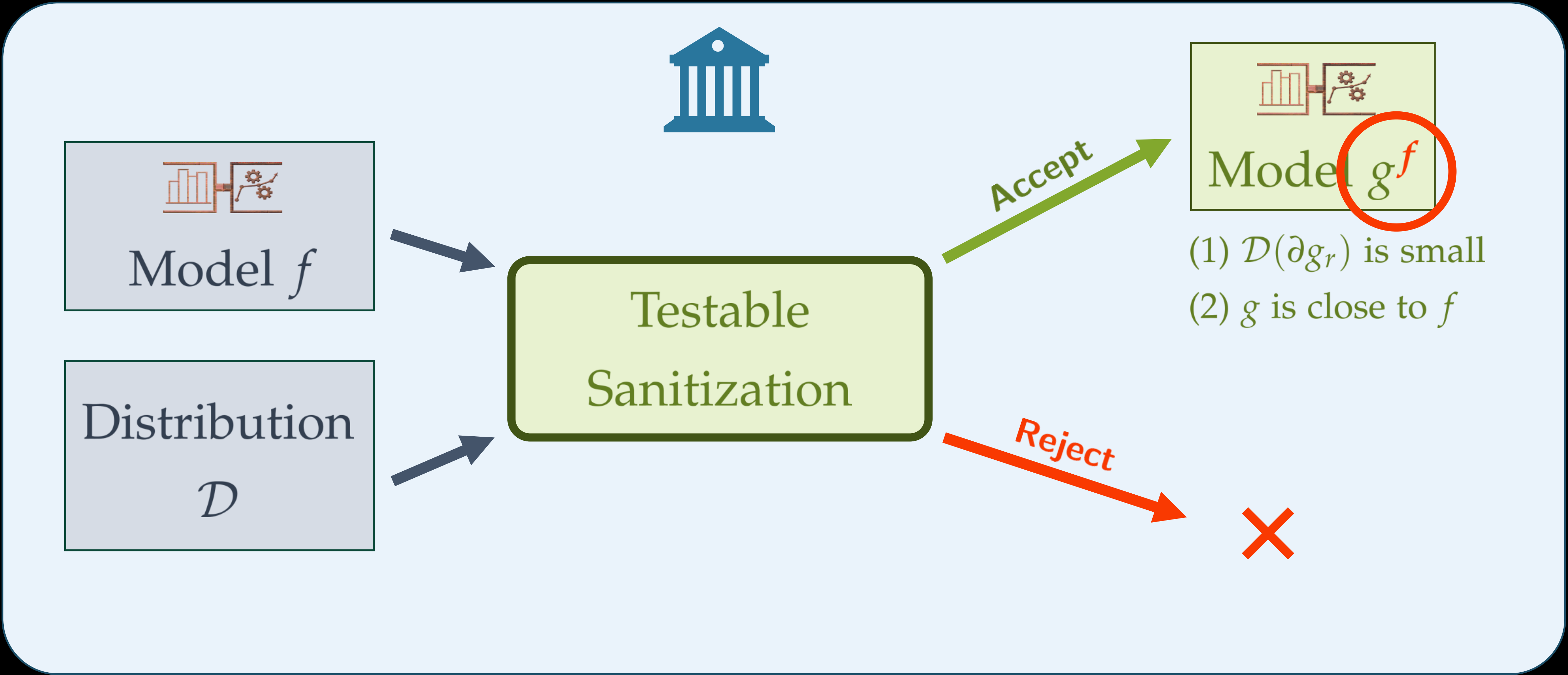


Testable Sanitization



Testable Sanitization

g has oracle access to f



Testable Sanitization

Definition

An algorithm A with access to $f, \mathcal{D}$ is (r, η, ε)-*testable sanitization* for a class $\mathcal{F}$ if it satisfies:

1. **Soundness:** **Accept** $\implies$ g is (r, η)-robust and ε-close to f.
2. **Completeness:** $f \in \mathcal{F}$ $\implies$ A outputs **Accept**.

Testable Sanitization

Definition

An algorithm A with access to $f, \mathcal{D}$ is (r, η, ε)-*testable sanitization* for a class $\mathcal{F}$ if it satisfies:

Does not depend on $\mathcal{F}$!

1. **Soundness: Accept** $\implies$ g is (r, η)-robust and ε-close to f.
2. **Completeness:** $f \in \mathcal{F}$ $\implies$ A outputs **Accept**.

Testable Sanitization

Definition

An algorithm A with access to $f, \mathcal{D}$ is (r, η, ε)-*testable sanitization* for a class $\mathcal{F}$ if it satisfies:

1. **Soundness:** **Accept** $\implies$ g is (r, η)-robust and ε-close to f.
2. **Completeness:** $f \in \mathcal{F} \implies A$ outputs **Accept**.

Testable Sanitization

Definition

An algorithm A with access to $f, \mathcal{D}$ is (r, η, ε)-*testable sanitization* for a class $\mathcal{F}$ if it satisfies:

1. **Soundness:** **Accept** $\implies$ g is (r, η)-robust and ε-close to f.
2. **Completeness:** $f \in \mathcal{F} \implies A$ outputs **Accept**.

Main Results

Theorem 1 [Cai Mamali Sotiraki Yu Zampetakis]
There is a (r, η, η)-testable sanitization algorithm for the class $\mathcal{F}(\sqrt{d} \cdot r, \eta)$ in time and sample complexity $\mathrm{poly}(d, 1/\eta)$.

1. **Soundness:** **Accept** $\implies$ g is (r, η)-robust and η-close to f.
2. **Completeness:** f is $(\sqrt{d} \cdot r, \eta)$-robust $\implies$ A outputs **Accept**.

Main Results

Theorem 1 [Cai Mamali Sotiraki Yu Zampetakis]
There is a (r, η, η)-testable sanitization algorithm for the class $\mathcal{F}(\sqrt{d} \cdot r, \eta)$ in time and sample complexity $\mathrm{poly}(d, 1/\eta)$.

1. No assumptions on $\mathcal{D}$ or on the expressivity of $\mathcal{F}$!
2. Sample complexity: $\tilde{O}(d/\eta^2)$.
3. Time complexity: $\tilde{O}(d^2/\eta^4)$.

- $\varepsilon = \eta$,

Main Results

Theorem 1 [Cai Mamali Sotiraki Yu Zampetakis]
There is a $(r, \eta, \boldsymbol{\eta})$-testable sanitization algorithm for the class $\mathcal{F}(\sqrt{d} \cdot r, \eta)$ in time and sample complexity $\mathrm{poly}(d, 1/\eta)$.

1. **Soundness:** **Accept** $\implies$ g is (r, η)-robust and $\boldsymbol{\eta}$-close to f.
2. **Completeness:** f is $(\sqrt{d} \cdot r, \eta)$-robust $\implies$ A outputs **Accept**.

Main Results

Theorem 2 [Cai Mamali Sotiraki Yu Zampetakis]

1. Any $(r, \eta, \eta/4)$-testable sanitization algorithm for the class $\mathcal{F}(2^{d/2} \cdot r, \eta)$ requires exponential in d samples.

2. Any $(r, 1/4, 1/4)$-testable sanitization algorithm for the class $\mathcal{F}(\tilde{o}(\sqrt{d}) \cdot r, 0)$ requires exponential in d time.

3. Even 0-tolerant $(r, 1/4, 1/4)$-testable sanitization for the class $\mathcal{F}(2^{o(d)} \cdot r, 0)$ requires exponential in d time.

Main Results

- $\varepsilon = \eta$,
- completeness for $\mathcal{F}(\sqrt{d} \cdot r, \eta)$,

Theorem 1 [Cai Mamali Sotiraki Yu Zampetakis]
There is a (r, η, η)-testable sanitization algorithm for the class $\mathcal{F}(\sqrt{d} \cdot r, \eta)$ in time and sample complexity $\mathrm{poly}(d, 1/\eta)$.

1. **Soundness:** **Accept** $\implies$ g is (r, η)-robust and η-close to f.

2. **Completeness:** f is $(\sqrt{d} \cdot r, \eta)$-robust $\implies$ A outputs **Accept**.

Main Results

Theorem 2 [Cai Mamali Sotiraki Yu Zampetakis]

1. Any $(r, \eta, \eta/4)$-testable sanitization algorithm for the class $\mathcal{F}(2^{d/2} \cdot r, \eta)$ requires exponential in d samples.

2. Any $(r, 1/4, 1/4)$-testable sanitization algorithm for the class $\mathcal{F}(\tilde{o}(\sqrt{d}) \cdot r, 0)$ requires exponential in d time.

3. Even 0-tolerant $(r, 1/4, 1/4)$-testable sanitization for the class $\mathcal{F}(2^{o(d)} \cdot r, 0)$ requires exponential in d time.

Main Results

- $\varepsilon = \eta$,
- completeness for $\mathcal{F}(\sqrt{d} \cdot r, \eta)$,
- non-tolerant.

Theorem 1 [Cai Mamali Sotiraki Yu Zampetakis]
There is a (r, η, η)-testable sanitization algorithm for the class $\mathcal{F}(\sqrt{d} \cdot r, \eta)$ in time and sample complexity $\mathrm{poly}(d, 1/\eta)$.

1. **Soundness:** **Accept** $\implies$ g is (r, η)-robust and η-close to f.
2. **Completeness:** f is $(\sqrt{d} \cdot r, \eta)$-robust $\implies$ A outputs **Accept**.

Main Results

Theorem 2 [Cai Mamali Sotiraki Yu Zampetakis]

1. Any $(r, \eta, \eta/4)$-testable sanitization algorithm for the class $\mathcal{F}(2^{d/2} \cdot r, \eta)$ requires exponential in d samples.

2. Any $(r, 1/4, 1/4)$-testable sanitization algorithm for the class $\mathcal{F}(\tilde{o}(\sqrt{d}) \cdot r, 0)$ requires exponential in d time.

3. Even 0-tolerant $(r, 1/4, 1/4)$-testable sanitization for the class $\mathcal{F}(2^{o(d)} \cdot r, 0)$ requires exponential in d time.

Main Results

Theorem 2 [Cai Mamali Sotiraki Yu Zampetakis]

1. Any $(r, \eta, \eta/4)$-testable sanitization algorithm for the class $\mathcal{F}(2^{d/2} \cdot r, \eta)$ requires exponential in d samples.
2. Any $(r, 1/4, 1/4)$-testable sanitization algorithm for the class $\mathcal{F}(\tilde{o}(\sqrt{d}) \cdot r, 0)$ requires exponential in d time.
3. Even 0-tolerant $(r, 1/4, 1/4)$-testable sanitization for the class $\mathcal{F}(2^{o(d)} \cdot r, 0)$ requires exponential in d time.

We have almost matching upper bounds.

Main Results

Theorem 1 [Cai Mamali Sotiraki Yu Zampetakis]

There is a (r, η, η)-testable sanitization algorithm for the class $\mathcal{F}(\sqrt{d} \cdot r, \eta)$ in time and sample complexity $\mathrm{poly}(d, 1/\eta)$.

Main Results

Theorem 1 [Cai Mamali Sotiraki Yu Zampetakis]

There is an algorithm (r, η, η)-testable sanitization for the class $\mathcal{F}(\sqrt{d} \cdot r, \eta)$ in time and sample complexity $\mathrm{poly}(d, 1/\eta)$.

Can we get (r, η, ε)-testable sanitization for $\varepsilon \ll \eta$?

Boundaries without Singular Points

Boundaries without Singular Points

Definition (Informal)

$f \in \mathcal{F}(r, \eta, \ell, L)$ if $f(x) = \mathbf{1}\{h(x) \geq 0\}$ with:

1. $f \in \mathcal{F}(r, \eta)$,
2. $\|\nabla^2 h(x)\| \leq L$,
3. $\|\nabla h(x)\| \geq \ell$ for all x s.t. $h(x) = 0$.

Boundaries without Singular Points

Definition (Informal)

$f \in \mathcal{F}(r, \eta, \ell, L)$ if $f(x) = \mathbf{1}\{h(x) \geq 0\}$ with:

1. $f \in \mathcal{F}(r, \eta)$,
2. $\|\nabla^2 h(x)\| \leq L$, **Smoothness**
3. $\|\nabla h(x)\| \geq \ell$ for all x s.t. $h(x) = 0$.

Boundaries without Singular Points

Definition (Informal)

$f \in \mathcal{F}(r, \eta, \ell, L)$ if $f(x) = \mathbf{1}\{h(x) \geq 0\}$ with:

1. $f \in \mathcal{F}(r, \eta)$,
2. $\|\nabla^2 h(x)\| \leq L$,
3. $\|\nabla h(x)\| \geq \ell$ for all x s.t. $h(x) = 0$.

Regularity

Boundaries without Singular Points

Definition (Informal)

$f \in \mathcal{F}(r, \eta, \ell, L)$ if $f(x) = \mathbf{1}\{h(x) \geq 0\}$ with:

1. $f \in \mathcal{F}(r, \eta)$,
2. $\|\nabla^2 h(x)\| \leq L$,
3. $\|\nabla h(x)\| \geq \ell$ for all x s.t. $h(x) = 0$.

Regularity

$x \in \mathbb{R}^d$

$$\left.\begin{aligned} \nabla h(x) &= 0 \\ h(\underbrace{x}_{d \text{ variables}}) &= 0 \end{aligned}\right\} (d+1) \text{ equations}$$

In "general position" this system should be unsatisfiable.

Regularity

Lemma (Informal) [Cai Mamali Sotiraki Yu Zampetakis]
For any $h : \mathbb{R}^d \to \mathbb{R}$, let $\tilde{h}(x) = h(x) + \langle a, x \rangle + b$, where a, b are random. Then $\tilde{h}$ satisfies regularity with high-probability.

Similar result if h is a neural network with randomness in the weights of last layer.

Main Results

Theorem 3 (Informal) [Cai Mamali Sotiraki Yu Zampetakis]
If $r \leq \ell/(dL)$, there is a (r, η, ε)-testable sanitization algorithm for the class $\mathcal{F}(\sqrt{d} \cdot r, \eta, \ell, L)$ in time and sample complexity $\mathrm{poly}(d, 1/\varepsilon)$.

Main Results

Theorem 4 (Very Informal) [Cai Mamali Sotiraki Yu Zampetakis]

If we can sample $f_1, \ldots, f_k$ such that:

- on average f_i has error ε, and
- a constant fraction of f_i's satisfy smoothness & robustness

then we can find a (r, η)-robust g with ε error in time and sample complexity $\mathrm{poly}(d, 1/\varepsilon)$.

Sanitization Algorithms

1: Efficient – Single Model

Theorem 1 [Cai Mamali Sotiraki Yu Zampetakis]
There is a (r, η, η)-testable sanitization algorithm for the class $\mathcal{F}(\sqrt{d} \cdot r, \eta)$ in time and sample complexity $\mathrm{poly}(d, 1/\eta)$.

Theorem 3 (Informal) [Cai Mamali Sotiraki Yu Zampetakis]
If $r \leq \ell/(dL)$, there is a (r, η, ε)-testable sanitization algorithm for the class $\mathcal{F}(\sqrt{d} \cdot r, \eta, \ell, L)$ in time and sample complexity $\mathrm{poly}(d, 1/\varepsilon)$.

2: Inefficient – Single Model

Theorem 2 [Cai Mamali Sotiraki Yu Zampetakis]

1. Any $(r, \eta, \eta/4)$-testable sanitization algorithm for the class $\mathcal{F}(2^{d/2} \cdot r, \eta)$ requires exponential in d samples.
2. Any $(r, 1/4, 1/4)$-testable sanitization algorithm for the class $\mathcal{F}(\tilde{o}(\sqrt{d}) \cdot r, 0)$ requires exponential in d time.
3. Even 0-tolerant $(r, 1/4, 1/4)$-testable sanitization for the class $\mathcal{F}(2^{o(d)} \cdot r, 0)$ requires exponential in d time.

We have almost matching upper bounds.

3: Efficient – Multi Model

Theorem 4 (Very Informal) [Cai Mamali Sotiraki Yu Zampetakis]
If we can sample $f_1, \ldots, f_k$ such that:

- on average f_i has error ε, and
- a constant fraction of f_i's satisfy smoothness & robustness

then we can find a (r, η)-robust g with ε error in time and sample complexity $\mathrm{poly}(d, 1/\varepsilon)$.

Sanitization Algorithms

1: Efficient – Single Model

Theorem 1 [Cai Mamali Sotiraki Yu Zampetakis]
There is a (r, η, η)-testable sanitization algorithm for the class $\mathcal{F}(\sqrt{d} \cdot r, \eta)$ in time and sample complexity $\mathrm{poly}(d, 1/\eta)$.

Theorem 3 (Informal) [Cai Mamali Sotiraki Yu Zampetakis]
If $r \leq \ell/(dL)$, there is a (r, η, ε)-testable sanitization algorithm for the class $\mathcal{F}(\sqrt{d} \cdot r, \eta, \ell, L)$ in time and sample complexity $\mathrm{poly}(d, 1/\varepsilon)$.

2: Inefficient – Single Model

Theorem 2 [Cai Mamali Sotiraki Yu Zampetakis]

1. Any $(r, \eta, \eta/4)$-testable sanitization algorithm for the class $\mathcal{F}(2^{d/2} \cdot r, \eta)$ requires exponential in d samples.
2. Any $(r, 1/4, 1/4)$-testable sanitization algorithm for the class $\mathcal{F}(\tilde{o}(\sqrt{d}) \cdot r, 0)$ requires exponential in d time.
3. Even 0-tolerant $(r, 1/4, 1/4)$-testable sanitization for the class $\mathcal{F}(2^{o(d)} \cdot r, 0)$ requires exponential in d time.

We have almost matching upper bounds.

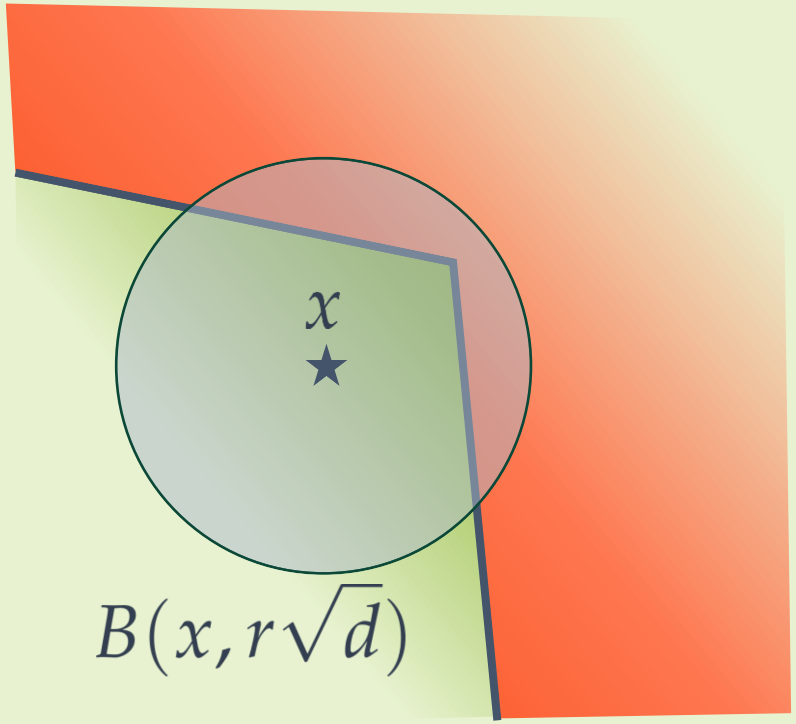
3: Efficient – Multi Model

Theorem 4 (Very Informal) [Cai Mamali Sotiraki Yu Zampetakis]
If we can sample $f_1, \ldots, f_k$ such that:

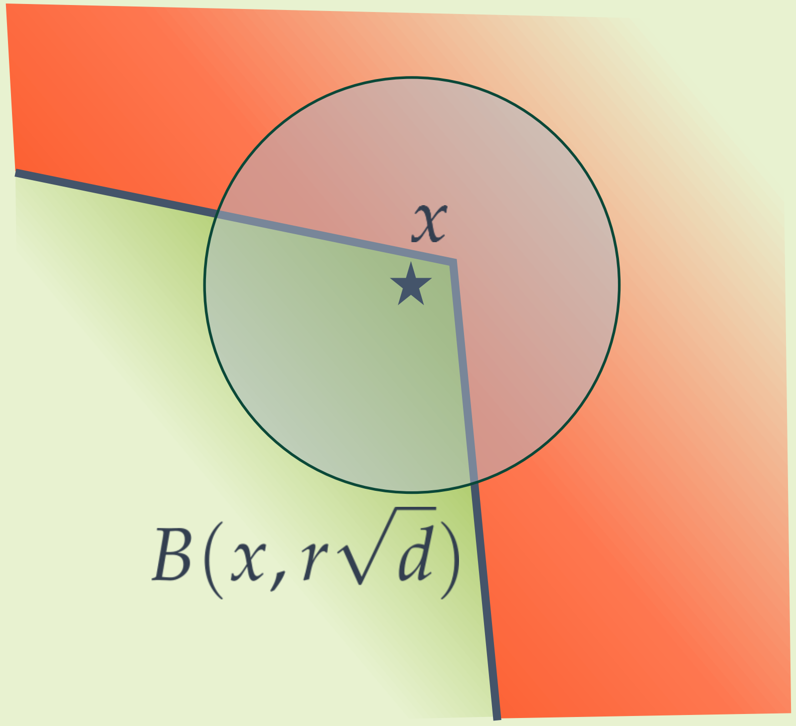
- on average f_i has error ε, and
- a constant fraction of f_i's satisfy smoothness & robustness

then we can find a (r, η)-robust g with ε error in time and sample complexity $\mathrm{poly}(d, 1/\varepsilon)$.

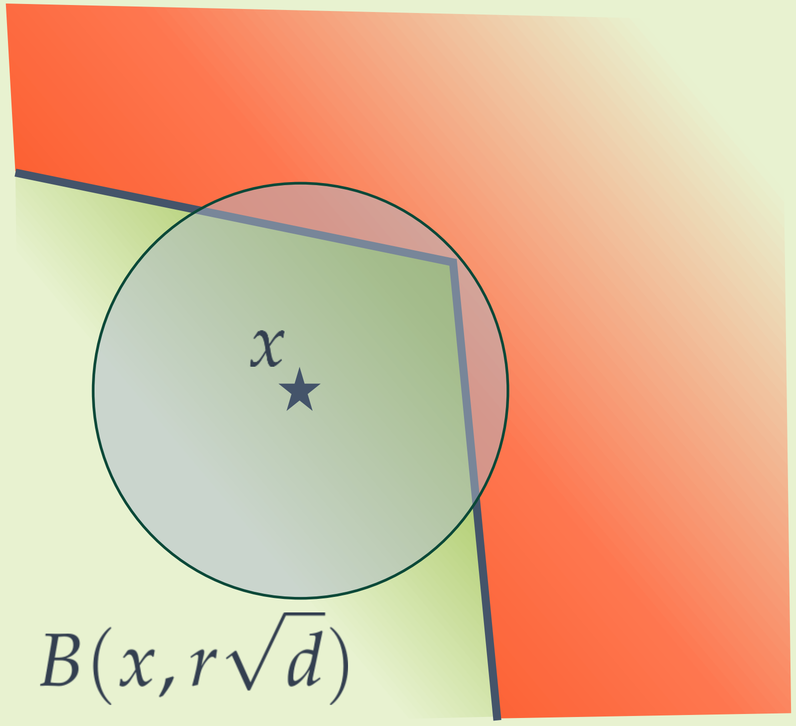
Sanitization Algorithm (Informal)



Sanitization Algorithm (Informal)



Sanitization Algorithm (Informal)



Sanitization. inspired by randomized smoothing and [Lange Vasilyan '25]

1. If $f(x) =$ strict majority in $B(x, r\sqrt{d})$,
   $g(x) = f(x)$
2. Else,
   $g(x) = 1 - f(x)$

Testing.

Is f ε-close to g?

And, does f have small noise sensitivity?

Main Steps of Proof

1. Prove that the output g of sanitization has small $\mathcal{D}(\partial g_r)$.

2. Prove if $f \in \mathcal{F}$ then g is close to f and f has small noise sensitivity.

Summary and Future Directions

1. We define testable sanitization for binary classifiers in terms of boundary conditions.
2. We build testable sanitization algorithms under no assumptions on the input distribution and for general classes of classifiers.

Open Questions:

- Testable sanitization for properties beyond robustness?
- Is (r, η, ε)-testable sanitization possible for a class larger than $\mathcal{F}(r, \eta, \ell, L)$?

Thank you! ☺