Subgradient Regularization: A Descent-Oriented Subgradient Method for Nonsmooth Optimization

Hanyang Li

IEOR, UC Berkeley

Joint work with Prof. Ying Cui

ICCOPT 2025

Descent directions in nonsmooth optimization

The **steepest descent direction** at x

$$g_x = -\operatorname*{argmin}_{v \in \partial f(x)} \|v\|$$

- If f is smooth, $g_x = -\nabla f(x)$

Descent directions in nonsmooth optimization

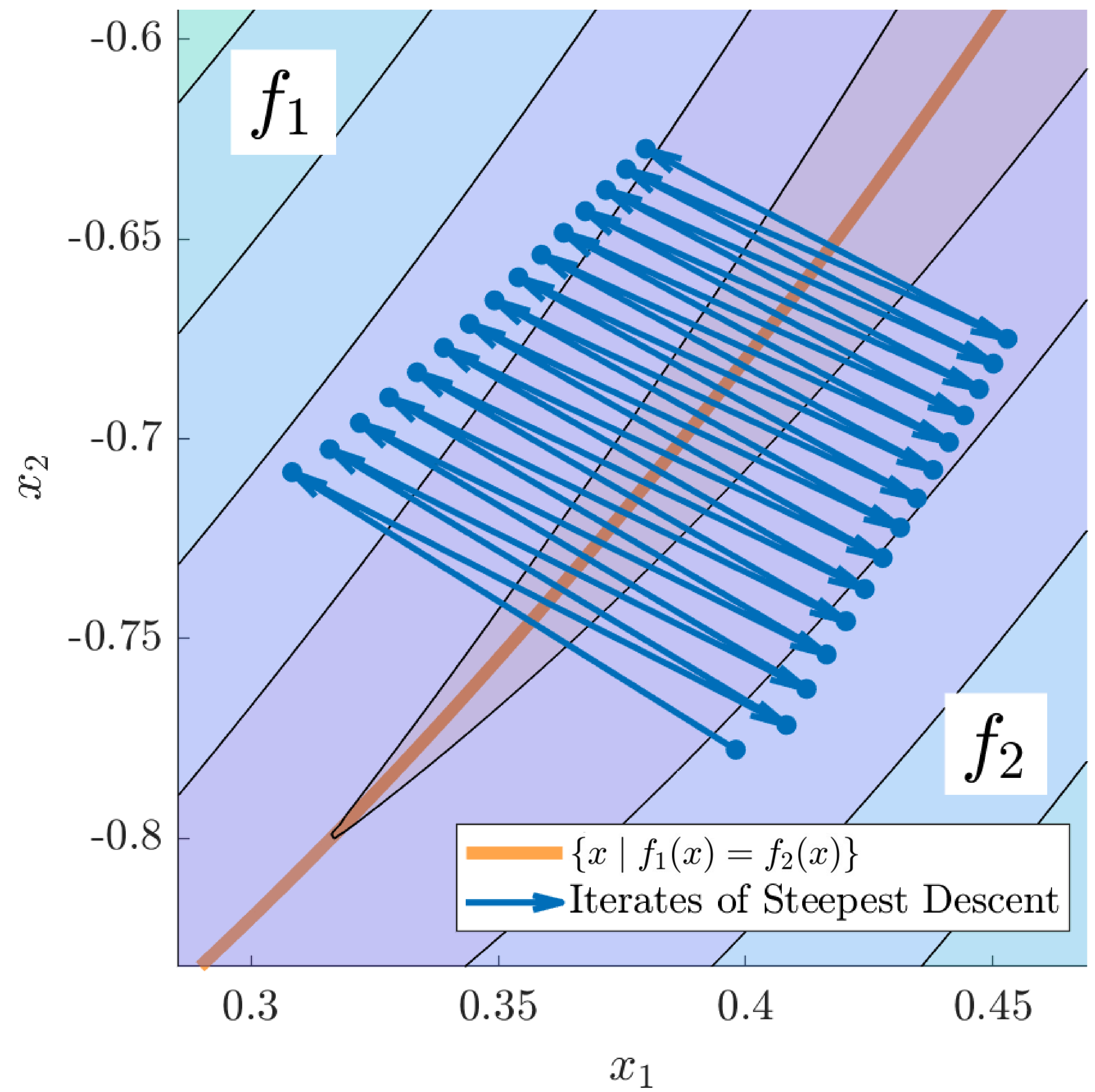
The **steepest descent direction** at x

$$g_x = -\operatorname*{argmin}_{v \in \partial f(x)} \|v\|$$

- If f is smooth, $g_x = -\nabla f(x)$
- Generally, g_x is discontinuous in x

— zigzag phenomenon

— may converge to non-stationary points



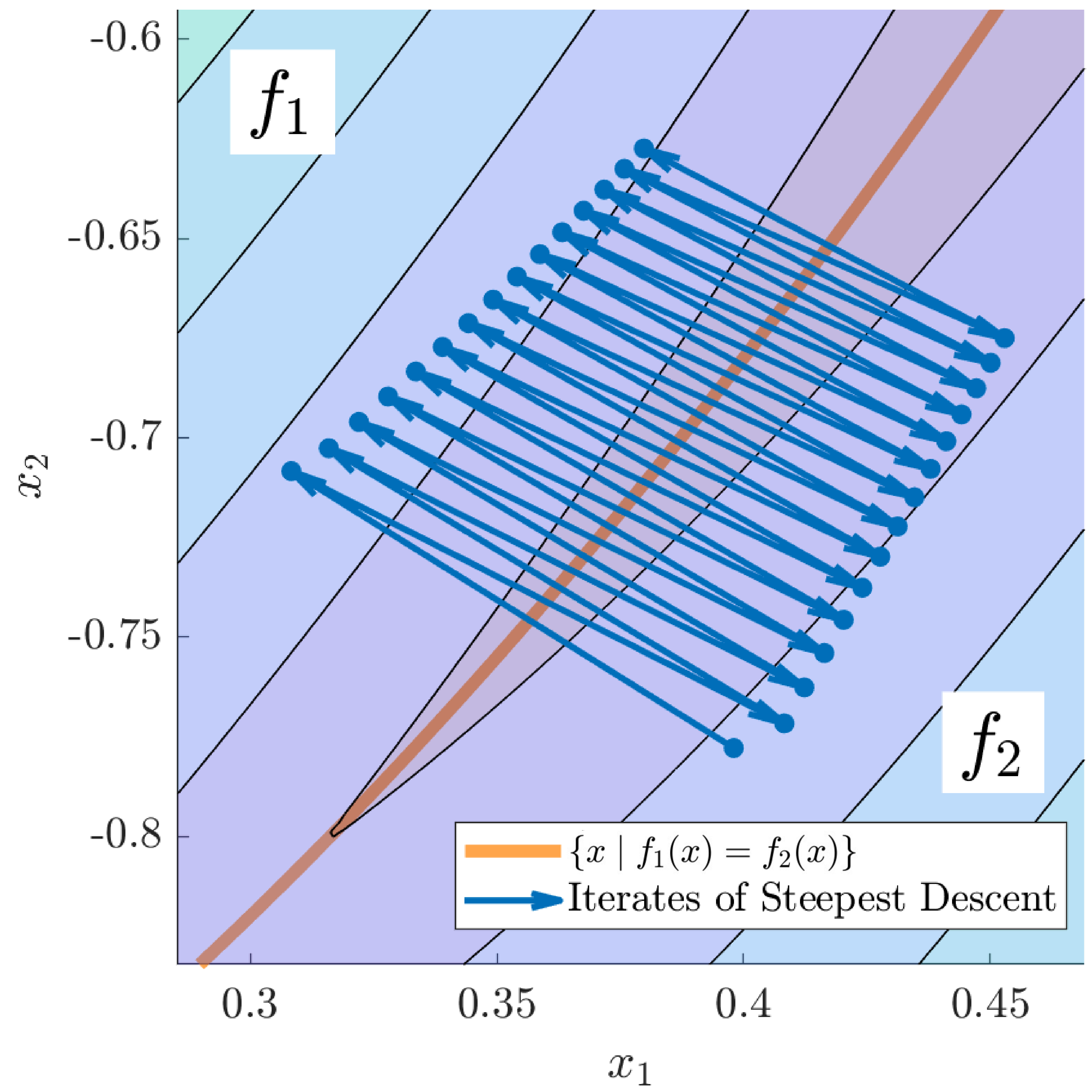
$f(x) = \max\{f_1(x), f_2(x)\}$

Descent directions in nonsmooth optimization

The **steepest descent direction** at x

$$g_x = -\operatorname*{argmin}_{v \in \partial f(x)} \|v\|$$

- If f is smooth, $g_x = -\nabla f(x)$
- Generally, g_x is discontinuous in x
  - — zigzag phenomenon
  - — may converge to non-stationary points
- **Improvement:** $g_x \xrightarrow{\text{regularization}} ??$ (stable in x)



$$f(x) = \max\{f_1(x), f_2(x)\}$$

Two types of descent algorithms

1. Goldstein-type methods

Idea: ϵ-neighborhood of x^k stabilizes the direction

Goldstein ϵ-subdifferential $\partial_{\epsilon}^{G} f(x) = \mathrm{conv}\left\{ \bigcup_{\|z-x\|\leq\epsilon} \partial f(z) \right\}$

Two types of descent algorithms

1. Goldstein-type methods

Idea: ϵ-neighborhood of x^k stabilizes the direction

Goldstein ϵ-subdifferential $\partial_\epsilon^G f(x) = \operatorname{conv}\left\{ \bigcup_{\|z-x\|\le\epsilon} \partial f(z) \right\}$

$$x^{k+1} = x^k - \epsilon \frac{g_k}{\|g_k\|} \quad \text{with} \quad g_k = \operatorname*{argmin}_{v\in\partial_\epsilon^G f(x^k)} \|v\|$$

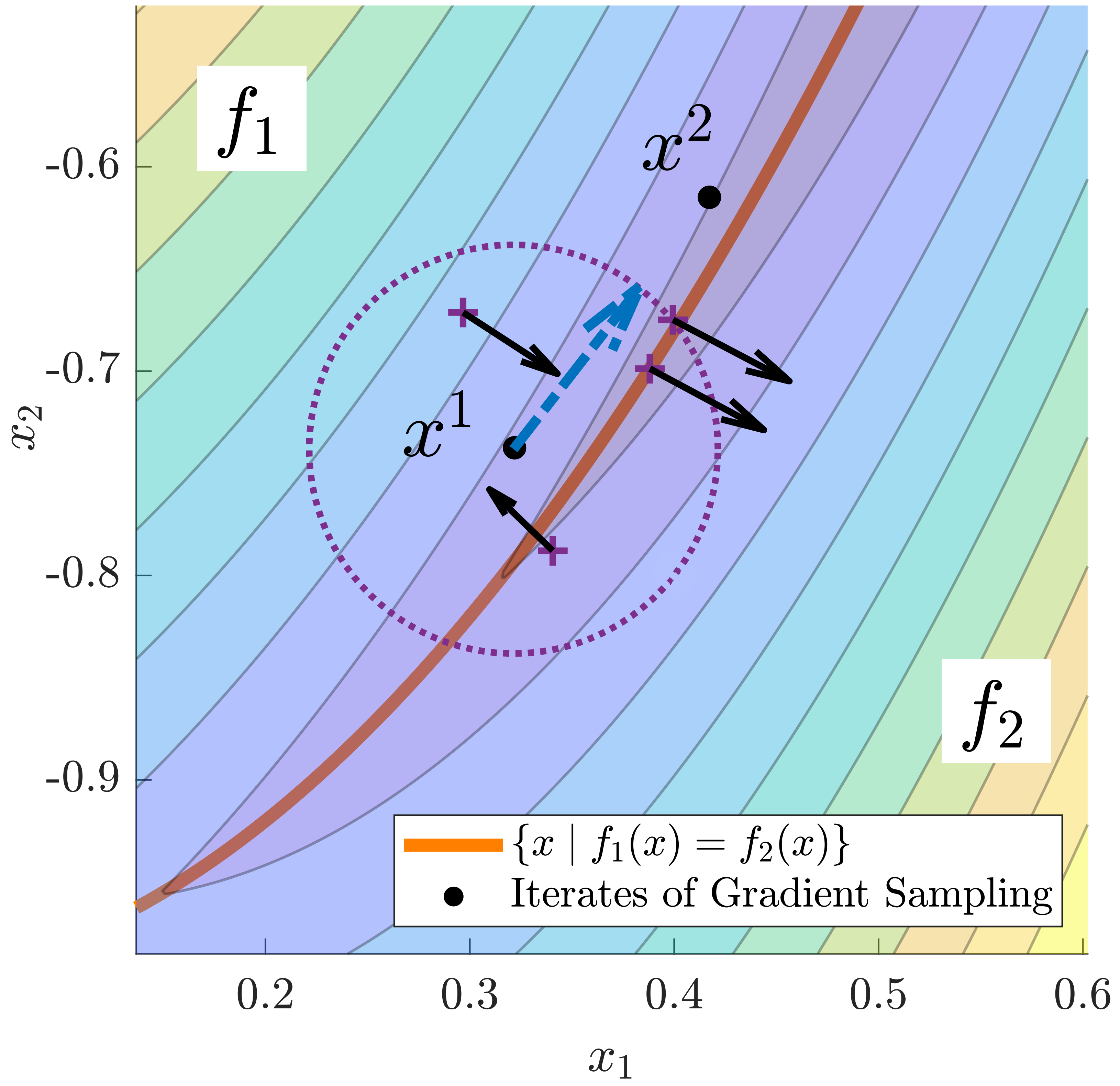
Two types of descent algorithms

1. Goldstein-type methods

Idea: ϵ-neighborhood of x^k stabilizes the direction

Goldstein ϵ-subdifferential $\partial_{\epsilon}^{G} f(x) = \operatorname{conv}\left\{ \bigcup_{\|z-x\| \leq \epsilon} \partial f(z) \right\}$

$$x^{k+1} = x^k - \epsilon \frac{g_k}{\|g_k\|} \quad \text{with} \quad g_k = \underset{v \in \partial_{\epsilon}^{G} f(x^k)}{\operatorname{argmin}} \|v\|$$



Practical issue: computation of g_k

$\xrightarrow{\text{approx.}}$ *Gradient Sampling [Burke, Lewis, Overton '05], INGD [Zhang, Lin, Jegelka, Sra, Jadbabaie '20], NTD [Davis, Jiang '23], …*

Two types of descent algorithms

2. Bundle-type methods

"Bundle": subgradients & function values over past iterations

$$\begin{array}{l} \left\{ v_1 \in \partial f(x^1), v_2 \in \partial f(x^2), \cdots, v_k \in \partial f(x^k) \right\} \\ \left\{ \quad f(x^1), \qquad\quad f(x^2), \quad \cdots, \quad f(x^k) \quad \right\} \end{array}$$

Two types of descent algorithms

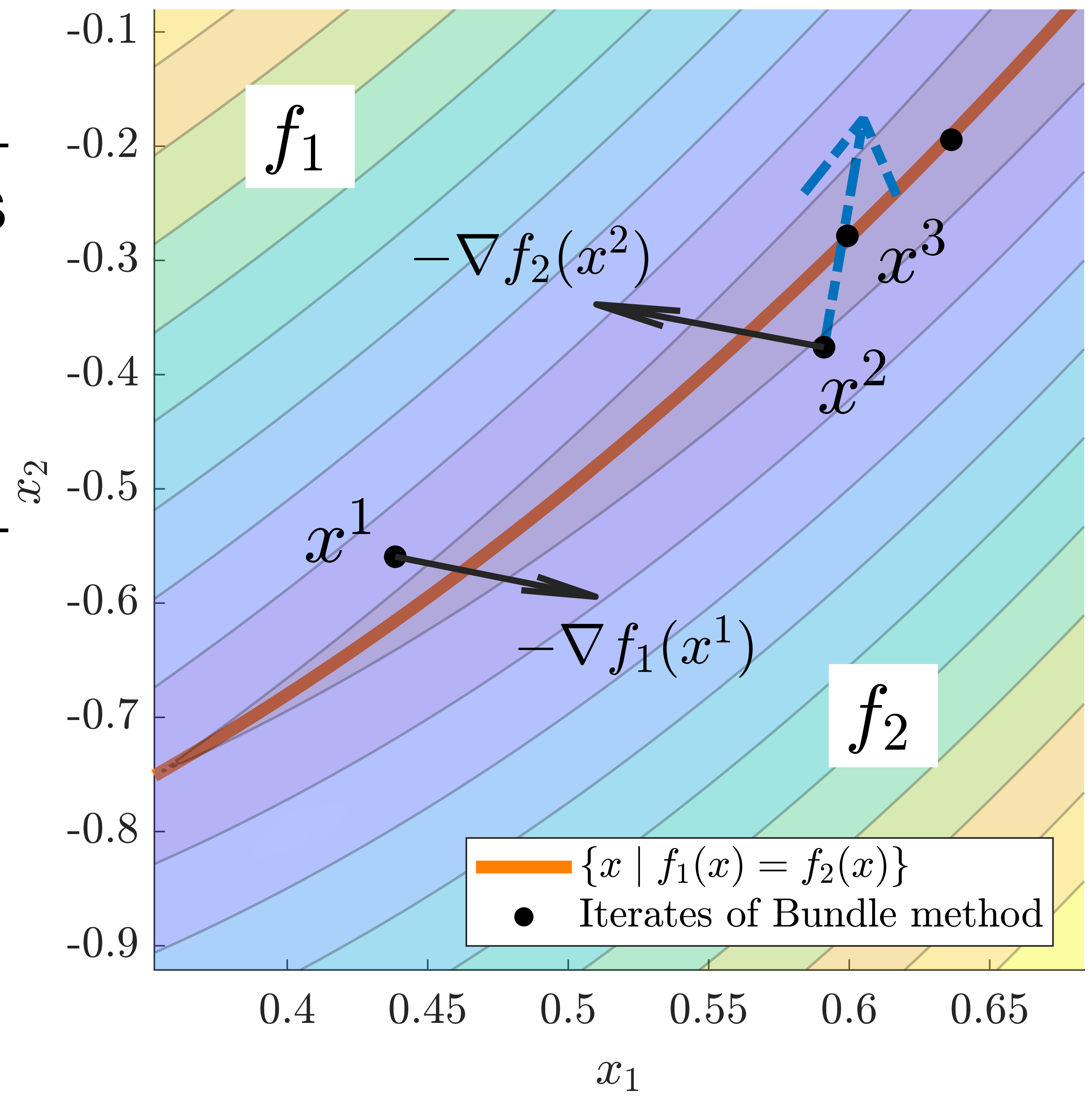
2. Bundle-type methods

"Bundle": subgradients & function values over past iterations

$$\left\{ v_1 \in \partial f(x^1), v_2 \in \partial f(x^2), \cdots, v_k \in \partial f(x^k) \right\}$$
$$\left\{ \quad f(x^1), \qquad f(x^2), \quad \cdots, \quad f(x^k) \quad \right\}$$

$$x^{k+1} = x^k - \alpha_k \, g_k$$

- g_k is a convex combination of $\{v_1, v_2, \cdots, v_k\}$
- $f(x^i)$ closer to $f(x^k)$ → larger weights for v_i



Two types of descent algorithms

2. Bundle-type methods

“Bundle”: subgradients & function values over past iterations

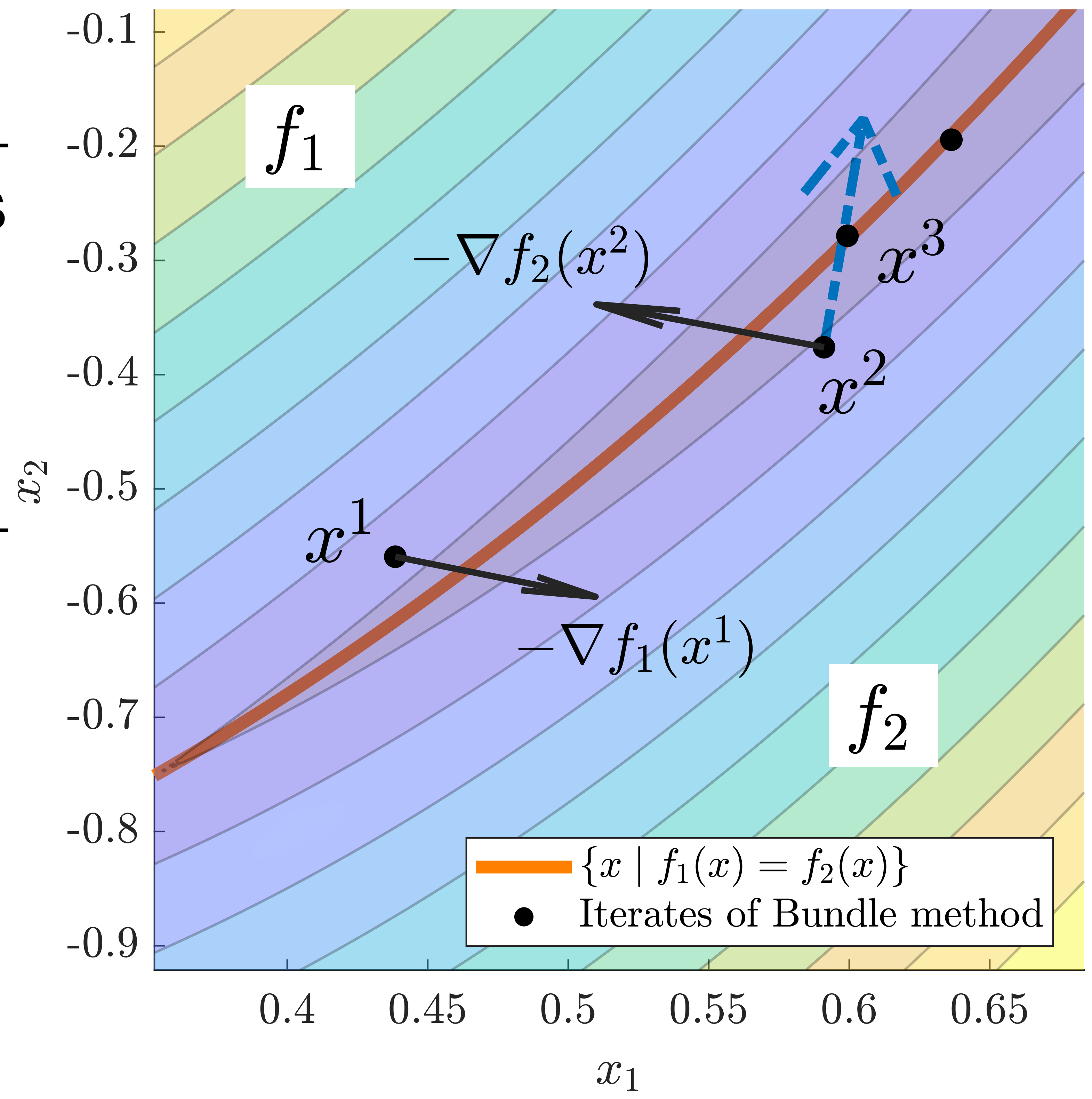
$$\left\{ v_1 \in \partial f(x^1), v_2 \in \partial f(x^2), \cdots, v_k \in \partial f(x^k) \right\}$$
$$\left\{ \quad f(x^1), \qquad f(x^2), \qquad \cdots, \qquad f(x^k) \quad \right\}$$

$$x^{k+1} = x^k - \alpha_k\, g_k$$

- g_k is a convex combination of $\{v_1, v_2, \cdots, v_k\}$
- $f(x^i)$ closer to $f(x^k)$ → larger weights for v_i

Idea: ~~ϵ-neighborhood of x^k~~ stabilizes the direction

ϵ-neighborhood of $f(x^k)$



Perspectives via “enlarged subdifferential”

$$x^{k+1} = x^k - \alpha_k \cdot g_k \quad \text{with} \quad g_k = \operatorname*{argmin}_{v \in S_k} \|v\|$$

Methods	Convex set S_k
Steepest descent	$\partial f(x^k)$ ✗
Goldstein-type	$\partial^G_{\epsilon_k} f(x^k) = \text{conv}\left\{ \bigcup_{\|z-x^k\| \le \epsilon_k} \partial f(z) \right\}$
Bundle-type (for convex f)	$\partial_{\epsilon_k} f(x^k) = \left\{ v \mid f(z) \ge f(x^k) + v^\top (z - x^k) - \epsilon_k, \forall z \right\}$

Key message:

To get a stable descent direction,

select & combine (sub)gradients in some “neighborhood”!

Key message:

To get a stable descent direction,

select & combine (sub)gradients in some "neighborhood"!

Questions:

- What is the general principle?
- What if more structures are known?

Part 1: A unifying principle for constructing stable descent directions

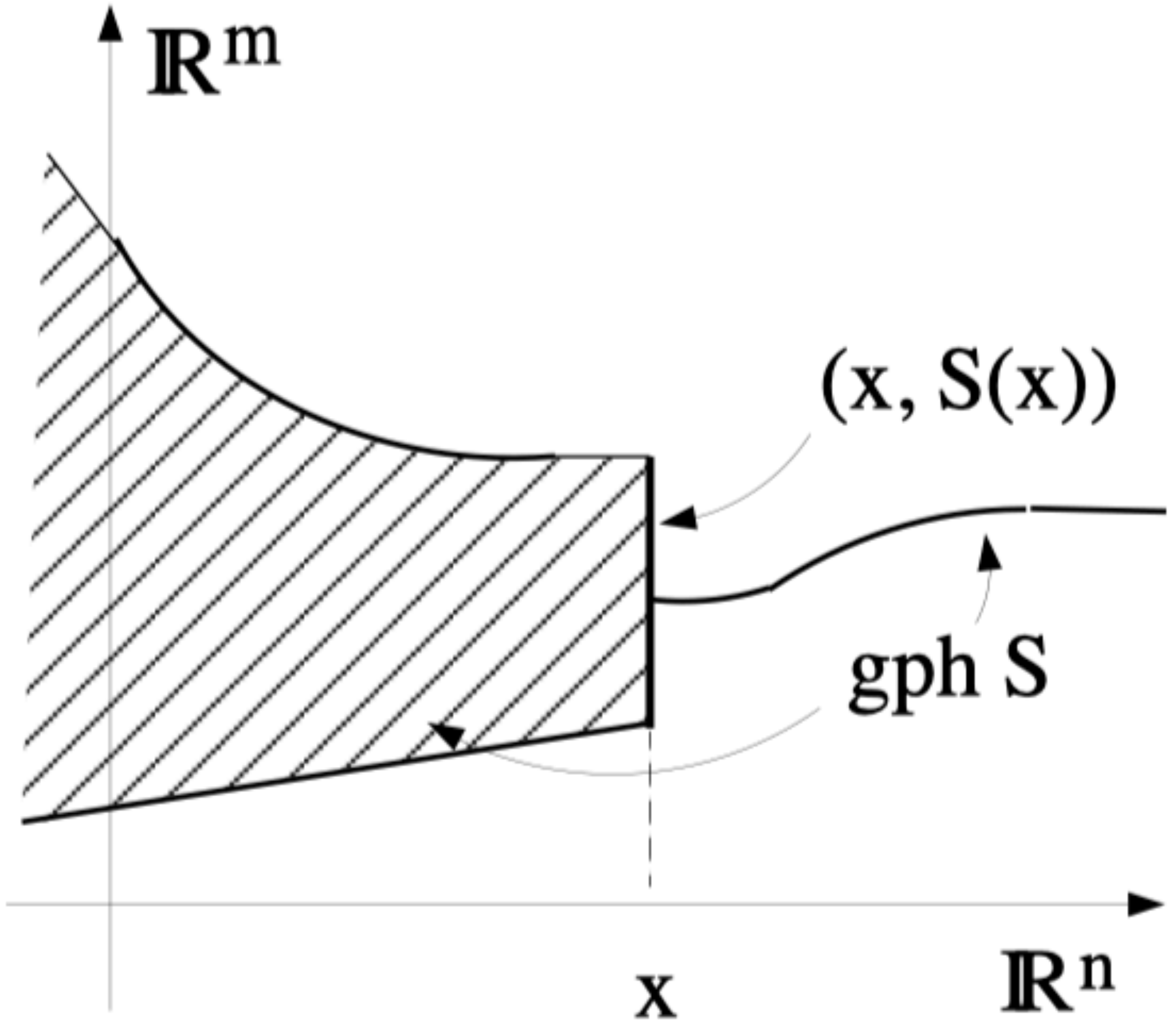
Preparation: set-valued analysis

For a set-valued mapping $S : \mathbb{R}^n \rightrightarrows \mathbb{R}^m$

- **Outer & Inner limits:**

$$\limsup_{x \to \bar{x}} S(x) = \bigcup_{x^k \to \bar{x}} \left\{ \text{accumulation points of } \{S(x^k)\}_{k \in \mathbb{N}} \right\}$$

$$\liminf_{x \to \bar{x}} S(x) = \bigcap_{x^k \to \bar{x}} \left\{ \text{limit points of } \{S(x^k)\}_{k \in \mathbb{N}} \right\}$$



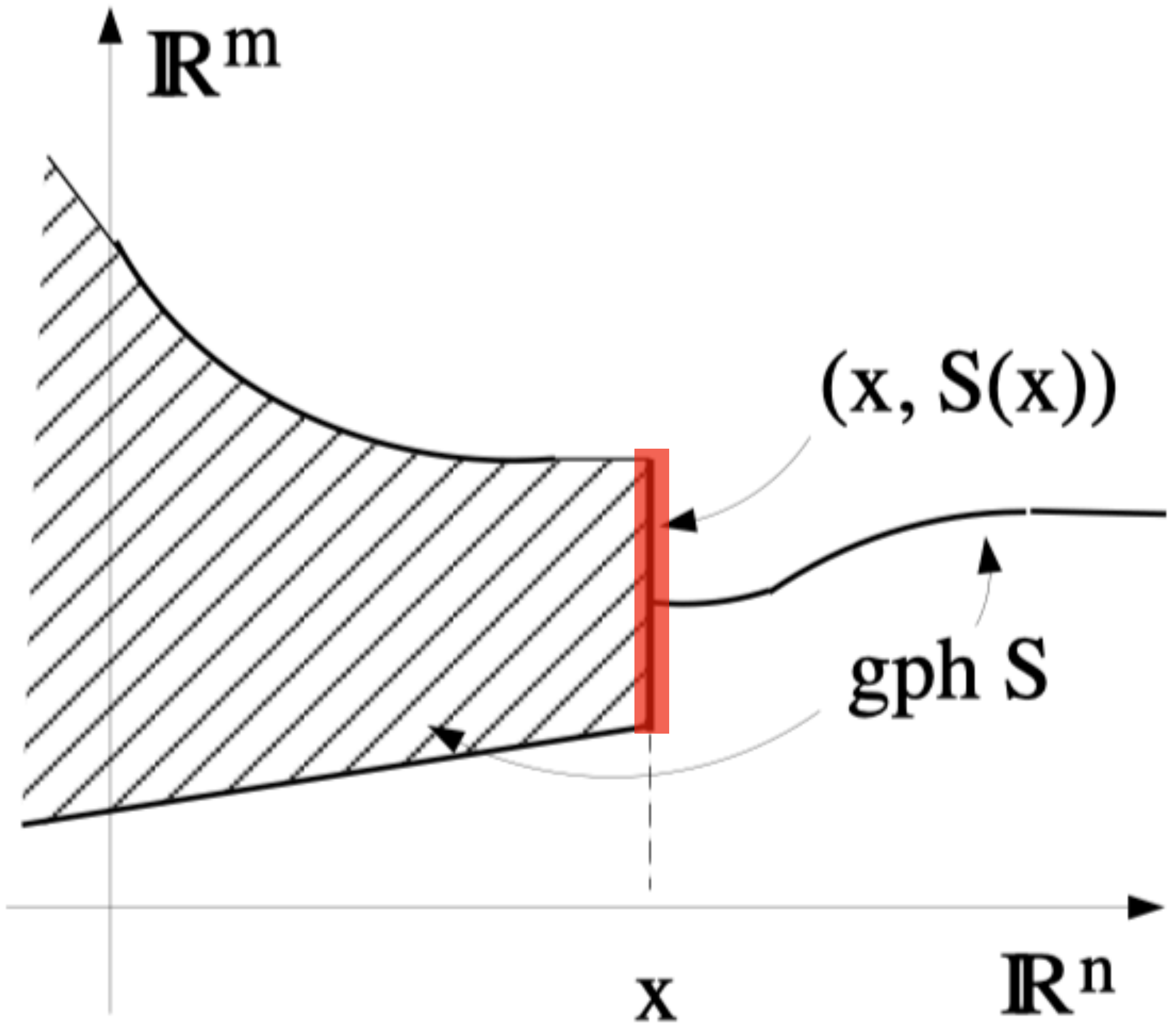
Preparation: set-valued analysis

For a set-valued mapping $S : \mathbb{R}^n \rightrightarrows \mathbb{R}^m$

- **Outer & Inner limits:**

$$\limsup_{x\to\bar{x}} S(x) = \bigcup_{x^k\to\bar{x}} \left\{\text{accumulation points of } \{S(x^k)\}_{k\in\mathbb{N}}\right\}$$

$$\liminf_{x\to\bar{x}} S(x) = \bigcap_{x^k\to\bar{x}} \left\{\text{limit points of } \{S(x^k)\}_{k\in\mathbb{N}}\right\}$$



Preparation: set-valued analysis

For a set-valued mapping $S : \mathbb{R}^n \rightrightarrows \mathbb{R}^m$

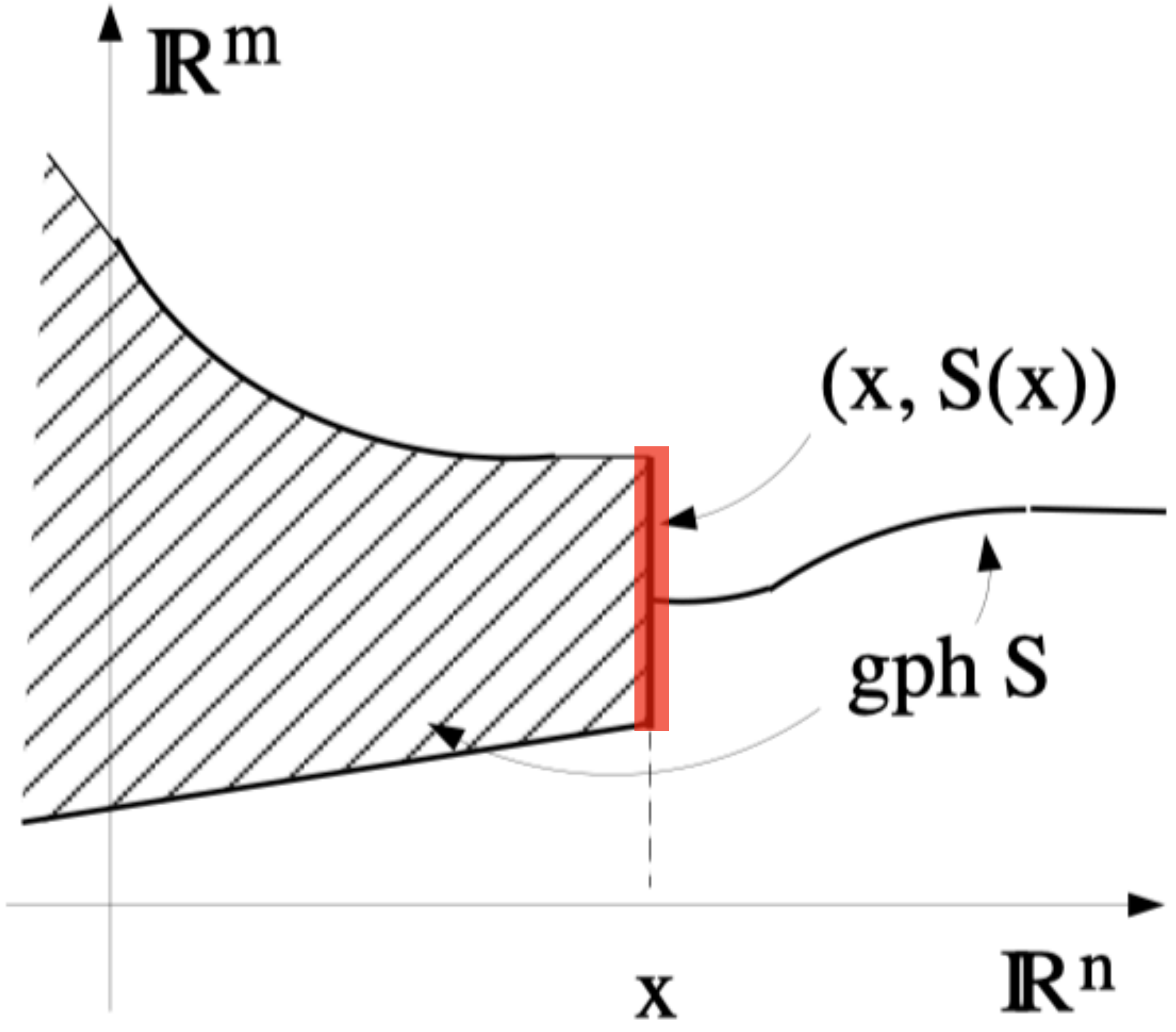
- **Outer & Inner limits:**

$$\limsup_{x\to\bar{x}} S(x) = \bigcup_{x^k\to\bar{x}} \left\{\text{accumulation points of } \{S(x^k)\}_{k\in\mathbb{N}}\right\}$$

$$\liminf_{x\to\bar{x}} S(x) = \bigcap_{x^k\to\bar{x}} \left\{\text{limit points of } \{S(x^k)\}_{k\in\mathbb{N}}\right\}$$

- **Outer semi-continuous** (osc) if $\limsup_{x\to\bar{x}} S(x) = S(\bar{x})$

- **Continuous** if $\limsup_{x\to\bar{x}} S(x) = \liminf_{x\to\bar{x}} S(\bar{x})$



$S(\cdot)$ *is osc, not continuous*

Preparation: set-valued analysis

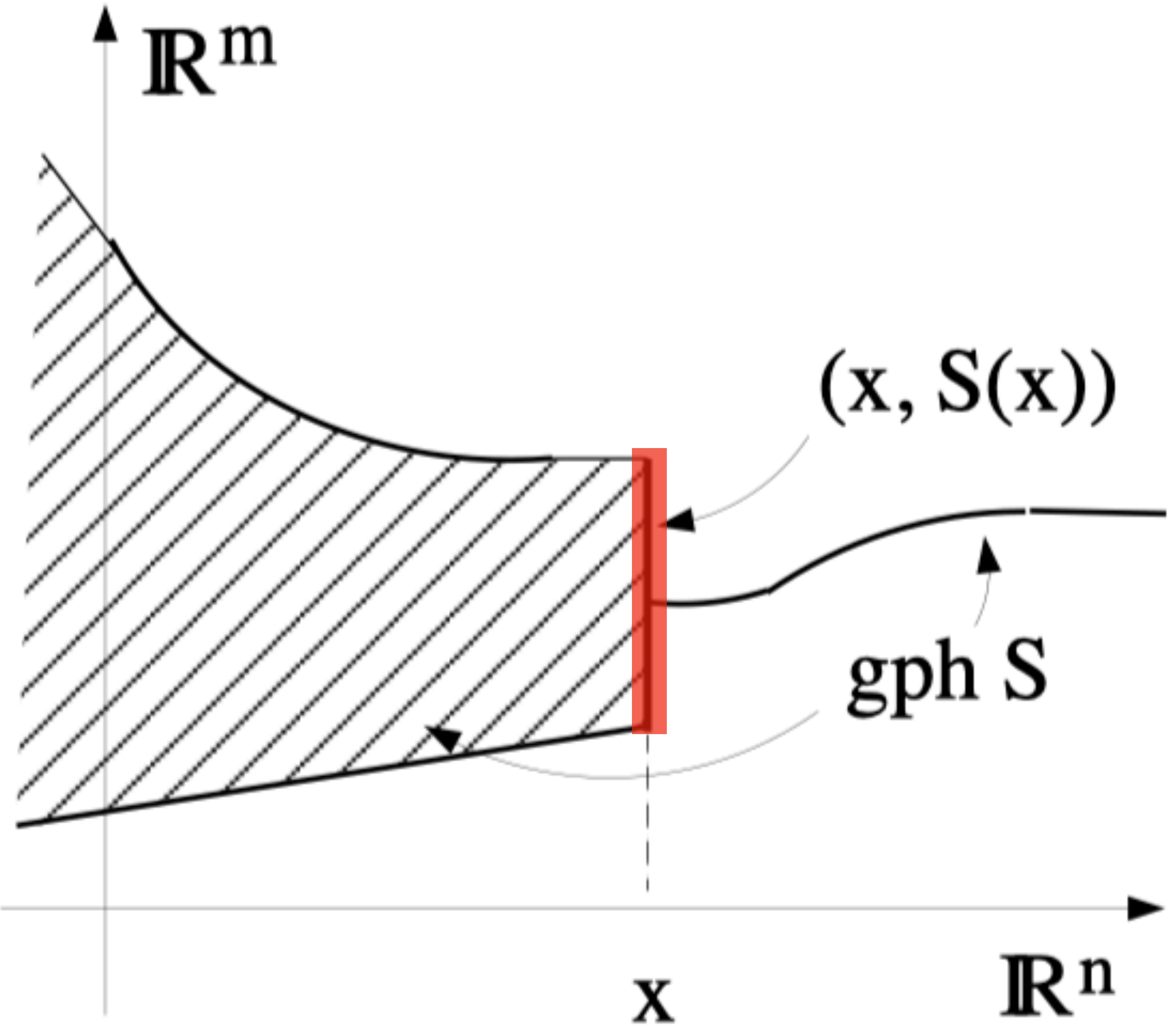
For a set-valued mapping $S : \mathbb{R}^n \rightrightarrows \mathbb{R}^m$

- **Outer & Inner limits:**

$$\limsup_{x \to \bar{x}} S(x) = \bigcup_{x^k \to \bar{x}} \left\{ \text{accumulation points of } \{S(x^k)\}_{k \in \mathbb{N}} \right\}$$

$$\liminf_{x \to \bar{x}} S(x) = \bigcap_{x^k \to \bar{x}} \left\{ \text{limit points of } \{S(x^k)\}_{k \in \mathbb{N}} \right\}$$

- **Outer semi-continuous** (osc) if $\limsup_{x \to \bar{x}} S(x) = S(\bar{x})$
- **Continuous** if $\limsup_{x \to \bar{x}} S(x) = \liminf_{x \to \bar{x}} S(\bar{x})$



$S(\cdot)$ is osc, not continuous

Facts: $\partial f(\cdot)$ is osc;

$\partial_\epsilon f(\cdot)$ is continuous for every fixed $\epsilon > 0$ when f is convex

Descent-oriented subdifferential

A map $G : \mathbb{R}^n \times (0,\infty) \rightrightarrows \mathbb{R}^m$ is a descent-oriented subdifferential for f if

(G1) Outer limit jointly in (x, ϵ) stays in the Clarke subdifferential:

$$\limsup_{\epsilon \downarrow 0,\, x \to \bar{x}} G(x, \epsilon) \subset \partial f(\bar{x})$$

(G2) Separate limit yields the minimal norm subgradient:

$$\lim_{\epsilon \downarrow 0} \left(\limsup_{x \to \bar{x}} G(x, \epsilon) \right) = \operatorname{argmin}\{ \|v\| \mid v \in \partial f(\bar{x}) \}$$

Descent-oriented subdifferential

A map $G : \mathbb{R}^n \times (0,\infty) \rightrightarrows \mathbb{R}^m$ is a descent-oriented subdifferential for f if

(G1) Outer limit jointly in (x, ϵ) stays in the Clarke subdifferential:

$$\limsup_{\epsilon \downarrow 0,\, x \to \bar{x}} G(x, \epsilon) \subset \partial f(\bar{x})$$

(G2) Separate limit yields the minimal norm subgradient:

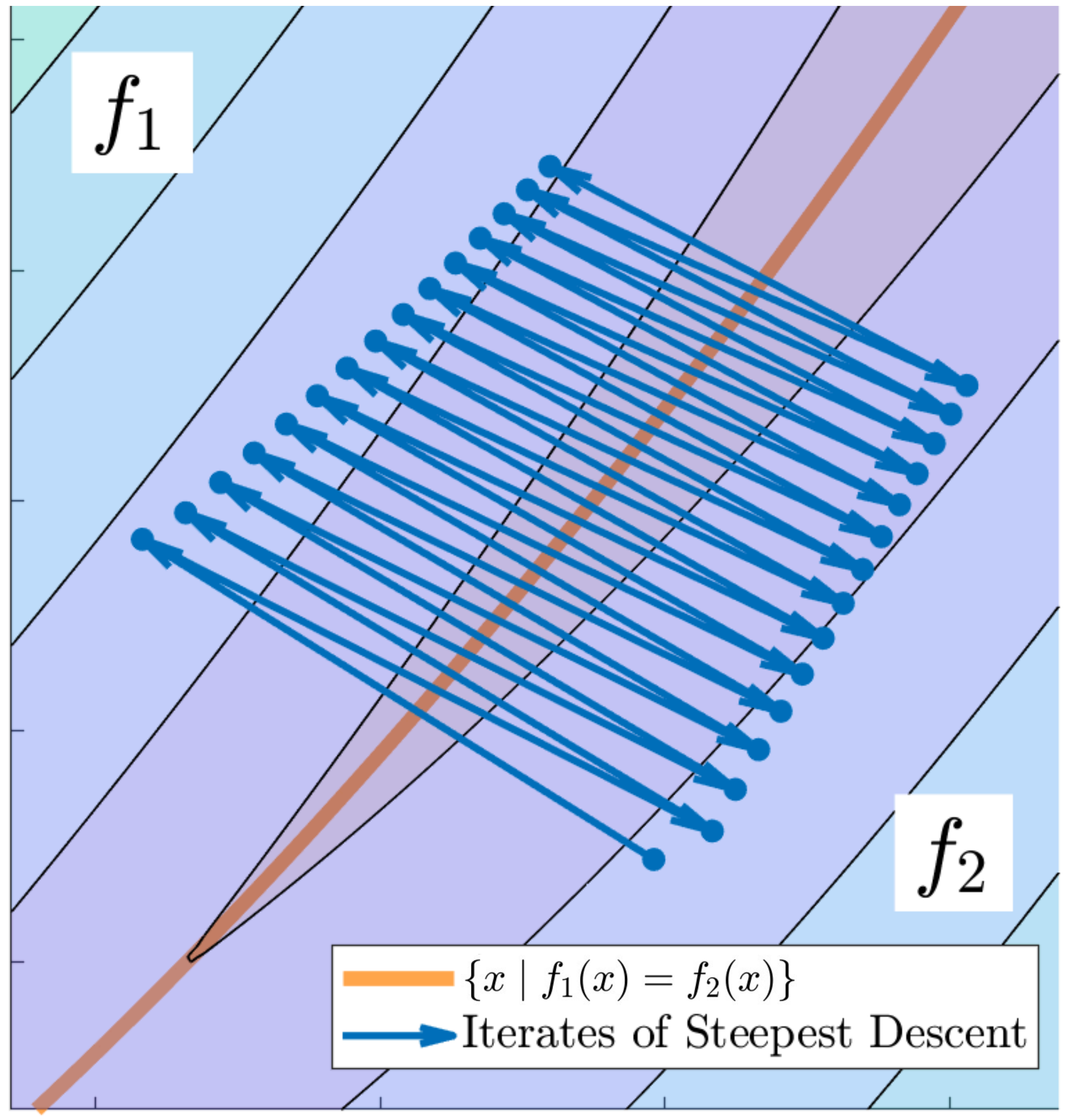
$$\lim_{\epsilon \downarrow 0} \left(\limsup_{x \to \bar{x}} G(x, \epsilon) \right) = \operatorname{argmin}\{\|v\| \mid v \in \partial f(\bar{x})\}$$

- Sufficient conditions for (G2):

$$\underbrace{\limsup_{x \to \bar{x}} G(x, \epsilon) = G(\bar{x}, \epsilon),\ \ \forall \epsilon > 0}_{G(\,\cdot\,, \epsilon) \text{ is osc}} \quad \text{and} \quad \lim_{\epsilon \downarrow 0} G(\bar{x}, \epsilon) = \operatorname*{argmin}_{v \in \partial f(\bar{x})} \|v\|$$

Descent-oriented subdifferential

A map $G : \mathbb{R}^n \times (0,\infty) \rightrightarrows \mathbb{R}^m$ is a descent-oriented subdifferential for f if



(G1) Outer limit jointly in (x, ϵ) stays in the Clarke subdifferential:

$$\limsup_{\epsilon \downarrow 0,\, x \to \bar{x}} G(x, \epsilon) \subset \partial f(\bar{x})$$

(G2) Separate limit yields the minimal norm subgradient:

$$\lim_{\epsilon \downarrow 0} \left(\limsup_{x \to \bar{x}} G(x, \epsilon) \right) = \operatorname{argmin}\{\|v\| \mid v \in \partial f(\bar{x})\}$$

- The minimal norm subgradient map

$$G : (x, \epsilon) \mapsto \operatorname{argmin}\{\|v\| \mid v \in \partial f(x)\}$$

violates (G2)!

Examples of descent-oriented subdifferential

A map $G : \mathbb{R}^n \times (0,\infty) \rightrightarrows \mathbb{R}^m$ is a descent-oriented subdifferential for f if

(G1) $\limsup_{\epsilon \downarrow 0,\, x \to \bar{x}} G(x, \epsilon) \subset \partial f(\bar{x})$

(G2) $\lim_{\epsilon \downarrow 0} \left(\limsup_{x \to \bar{x}} G(x, \epsilon) \right) = \operatorname{argmin}\{ \|v\| \mid v \in \partial f(\bar{x}) \}$

- **Goldstein direction**:

$$G : (x, \epsilon) \mapsto \operatorname{argmin}\left\{ \|v\| \mid v \in \partial_\epsilon^G f(x) \right\}$$

- **Bundle direction** (when f is convex):

$$G : (x, \epsilon) \mapsto \operatorname{argmin}\left\{ \|v\| \mid v \in \partial_\epsilon f(x) \right\}$$

- Gradient of Moreau envelope (when f is weakly convex):

$$G : (x, \epsilon) \mapsto \nabla e_\epsilon f(x) \quad \text{with} \quad e_\epsilon f(x) := \inf_z \left\{ f(z) + (2\epsilon)^{-1} \|z - x\|^2 \right\}$$

Existence of descent directions

A map $G : \mathbb{R}^n \times (0,\infty) \rightrightarrows \mathbb{R}^m$ is a descent-oriented subdifferential for f if

(G1) $\limsup_{\epsilon \downarrow 0,\, x \to \bar{x}} \; G(x,\epsilon) \subset \partial f(\bar{x})$

(G2) $\lim_{\epsilon \downarrow 0} \left(\limsup_{x \to \bar{x}} G(x,\epsilon) \right) = \operatorname{argmin}\{ \|v\| \mid v \in \partial f(\bar{x}) \}$

Proposition: For nonstationary point x and constant $\alpha \in (0,1)$,

$$f(\bar{x} - \eta g) \le f(\bar{x}) - \alpha \eta \|g\|^2, \qquad \forall g \in \limsup_{x \to \bar{x}} G(x,\epsilon),$$

holds for sufficiently small ϵ and η.

Algorithm

A descent-oriented subgradient method

```
Given a descent-oriented subdifferential $G : \mathbb{R}^n \times (0,\infty) \rightrightarrows \mathbb{R}^m$
for $k = 0,1,\cdots$
    for $i = 0,1,\cdots$
        Generate a direction $g^{k,i} \in G(x^k, \epsilon_{k,0}\, 2^{-i})$
        if $\exists \eta_k \in \left\{\epsilon_{k,0}, \cdots, \epsilon_{k,0}\, 2^{-i}\right\}$ with $f\left(x^k - \eta_k g^{k,i}\right) \leq f(x^k) - \alpha \eta_k \|g^{k,i}\|^2$    } line-search
            Update $x^{k+1} = x^k - \eta_k g^{k,i}$ and break
    if $\|g^{k,i}\| \leq \nu_k$
        Update $\epsilon_{k+1,0} = \epsilon_{k,0}/2$ and $\nu_{k+1} = \nu_k/2$
    else set $\epsilon_{k+1,0} = \epsilon_{k,0}$ and $\nu_{k+1} = \nu_k$
```

Algorithm

A descent-oriented subgradient method

```
Given a descent-oriented subdifferential $G : \mathbb{R}^n \times (0,\infty) \rightrightarrows \mathbb{R}^m$
for $k = 0,1,\cdots$
    for $i = 0,1,\cdots$
        Generate a direction $g^{k,i} \in G(x^k, \epsilon_{k,0}\, 2^{-i})$
        if $\exists \eta_k \in \left\{\epsilon_{k,0}, \cdots, \epsilon_{k,0}\, 2^{-i}\right\}$ with $f\left(x^k - \eta_k g^{k,i}\right) \leq f(x^k) - \alpha\eta_k \|g^{k,i}\|^2$   } line-search
            Update $x^{k+1} = x^k - \eta_k g^{k,i}$ and break                                                            }
    if $\|g^{k,i}\| \leq \nu_k$
        Update $\epsilon_{k+1,0} = \epsilon_{k,0}/2$ and $\nu_{k+1} = \nu_k/2$
    else set $\epsilon_{k+1,0} = \epsilon_{k,0}$ and $\nu_{k+1} = \nu_k$
```

- The inner-loop terminates for sufficiently large i ($\exists$ descent directions at x^k)

Algorithm

A descent-oriented subgradient method

```
Given a descent-oriented subdifferential $G : \mathbb{R}^n \times (0,\infty) \rightrightarrows \mathbb{R}^m$
for $k = 0,1,\cdots$
    for $i = 0,1,\cdots$
        Generate a direction $g^{k,i} \in G(x^k, \epsilon_{k,0}\, 2^{-i})$
        if $\exists \eta_k \in \left\{\epsilon_{k,0}, \cdots, \epsilon_{k,0}\, 2^{-i}\right\}$ with $f\left(x^k - \eta_k g^{k,i}\right) \le f(x^k) - \alpha\eta_k \|g^{k,i}\|^2$   } line-search
            Update $x^{k+1} = x^k - \eta_k g^{k,i}$ and break
    if $\|g^{k,i}\| \le \nu_k$
        Update $\epsilon_{k+1,0} = \epsilon_{k,0}/2$ and $\nu_{k+1} = \nu_k/2$
    else set $\epsilon_{k+1,0} = \epsilon_{k,0}$ and $\nu_{k+1} = \nu_k$
```

Algorithm

A descent-oriented subgradient method

```
Given a descent-oriented subdifferential $G : \mathbb{R}^n \times (0,\infty) \rightrightarrows \mathbb{R}^m$
for $k = 0,1,\cdots$
    for $i = 0,1,\cdots$
        Generate a direction $g^{k,i} \in G(x^k, \epsilon_{k,0} 2^{-i})$
        if $\exists \eta_k \in \left\{\epsilon_{k,0}, \cdots, \epsilon_{k,0} 2^{-i}\right\}$ with $f\left(x^k - \eta_k g^{k,i}\right) \le f(x^k) - \alpha\eta_k \|g^{k,i}\|^2$  } line-search
            Update $x^{k+1} = x^k - \eta_k g^{k,i}$ and break
    if $\|g^{k,i}\| \le \nu_k$
        Update $\epsilon_{k+1,0} = \epsilon_{k,0}/2$ and $\nu_{k+1} = \nu_k/2$
    else set $\epsilon_{k+1,0} = \epsilon_{k,0}$ and $\nu_{k+1} = \nu_k$
```

Theorem: Any accumulation point $\bar{x}$ of $\{x^k\}$ is a stationary point, i.e., $0 \in \partial f(\bar{x})$.

Idea: If x^k close to a non-stationary point $\bar{x}$ $\Rightarrow$ $G(x^k, \epsilon)$ close to $G(\bar{x}, \epsilon)$ [for a fixed $\epsilon > 0$]

$\Rightarrow$ x^k escapes $\bar{x}$ for sufficiently small ϵ

A general principle: Descent-oriented subdifferential G

- *Examples: Goldstein & Bundle directions*

A framework of descent algorithms using $G(x, \epsilon)$

A general principle: Descent-oriented subdifferential G

- *Examples: Goldstein & Bundle directions*

A framework of descent algorithms using $G(x, \epsilon)$

Question 2: What if more structures are known? e.g., $f(x) = \max\{\underbrace{f_1(x), f_2(x)}_{\text{smooth}}\}$

Part 2: Efficient construction for nonsmooth marginal functions

Goal:

- $G(\,\cdot\,,\epsilon)$ is osc
- $\lim_{\epsilon\downarrow 0} G(\bar{x},\epsilon) = \operatorname*{argmin}_{v\in\partial f(\bar{x})} \|v\|$

A toy example

For a piecewise smooth function $f(x) = \max\{f_1(x), f_2(x)\}$,

$$\partial f(x) = \left\{ \bar{y}_1 \nabla f_1(x) + \bar{y}_2 \nabla f_2(x) \;\middle|\; \bar{y} \in \operatorname*{argmax}_{y\in\Delta^2} \left[y_1 f_1(x) + y_2 f_2(x) \right] \right\},$$

- $\Delta^2 = \{y \geq 0 \mid y_1 + y_2 = 1\}$

Goal:

- $G(\cdot, \epsilon)$ is osc
- $\lim_{\epsilon \downarrow 0} G(\bar{x}, \epsilon) = \operatorname{argmin}_{v \in \partial f(\bar{x})} \|v\|$

A toy example

For a piecewise smooth function $f(x) = \max\{f_1(x), f_2(x)\}$,

$$\partial f(x) = \left\{ \bar{y}_1 \nabla f_1(x) + \bar{y}_2 \nabla f_2(x) \;\middle|\; \bar{y} \in \operatorname*{argmax}_{y \in \Delta^2} \left[y_1 f_1(x) + y_2 f_2(x) \right] \right\},$$

- $\Delta^2 = \{y \geq 0 \mid y_1 + y_2 = 1\}$

For any $\epsilon > 0$, define

$$G(x, \epsilon) = \left\{ \bar{y}_1^\epsilon \nabla f_1(x) + \bar{y}_2^\epsilon \nabla f_2(x) \;\middle|\; \bar{y}^\epsilon \in \operatorname*{argmax}_{y \in \Delta^2} \left[y_1 f_1(x) + y_2 f_2(x) - \frac{\epsilon}{2} \boxed{\underset{\text{subgradient regularization}}{\|y_1 \nabla f_1(x) + y_2 \nabla f_2(x)\|^2}} \right] \right\}$$

Goal:

- $G(\cdot, \epsilon)$ is osc
- $\lim_{\epsilon \downarrow 0} G(\bar{x}, \epsilon) = \operatorname{argmin}_{v \in \partial f(\bar{x})} \|v\|$

A toy example

For a piecewise smooth function $f(x) = \max\{f_1(x), f_2(x)\}$,

$$\partial f(x) = \left\{ \bar{y}_1 \nabla f_1(x) + \bar{y}_2 \nabla f_2(x) \;\middle|\; \bar{y} \in \operatorname*{argmax}_{y \in \Delta^2} \left[y_1 f_1(x) + y_2 f_2(x) \right] \right\},$$

- $\Delta^2 = \{y \geq 0 \mid y_1 + y_2 = 1\}$

For any $\epsilon > 0$, define

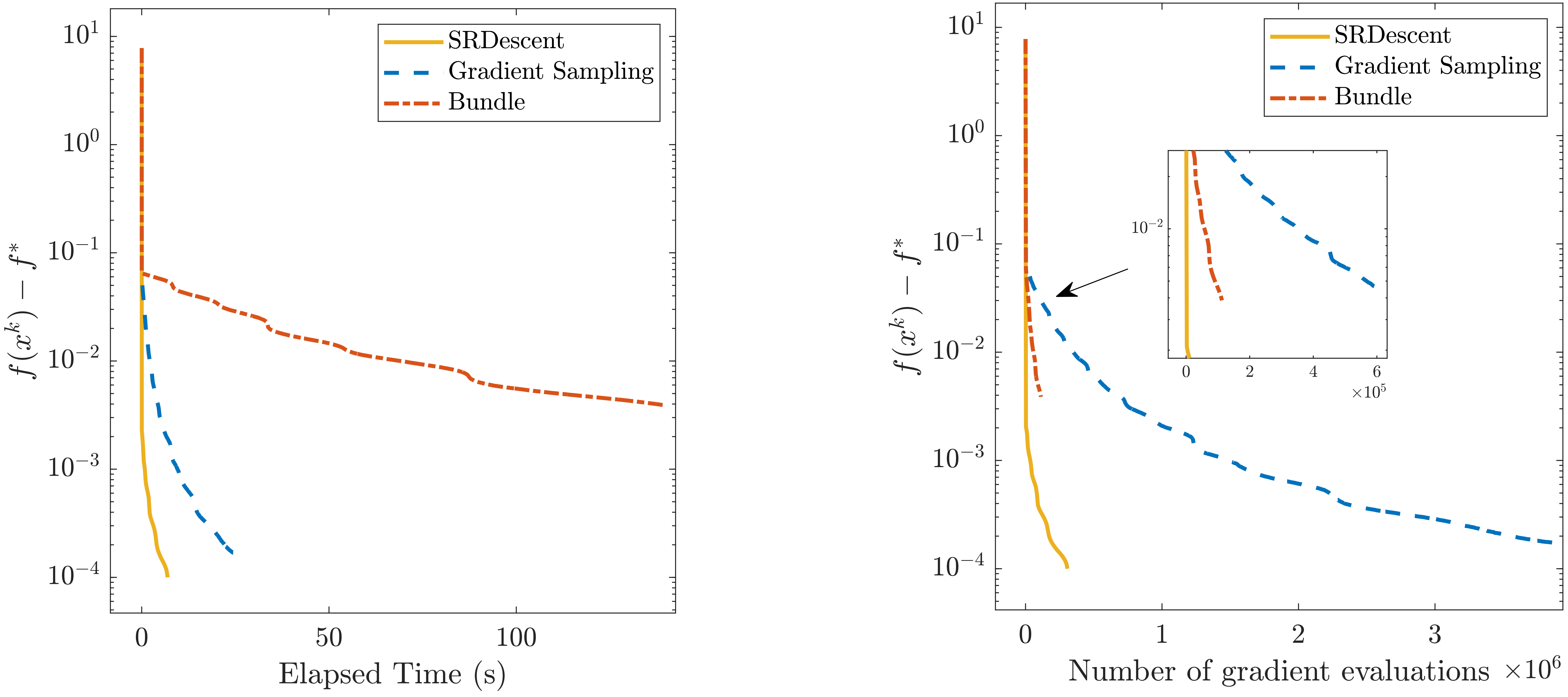
$$G(x, \epsilon) = \left\{ \bar{y}_1^\epsilon \nabla f_1(x) + \bar{y}_2^\epsilon \nabla f_2(x) \;\middle|\; \bar{y}^\epsilon \in \operatorname*{argmax}_{y \in \Delta^2} \left[y_1 f_1(x) + y_2 f_2(x) - \frac{\epsilon}{2} \underset{\text{subgradient regularization}}{\boxed{\|y_1 \nabla f_1(x) + y_2 \nabla f_2(x)\|}}^2 \right] \right\}$$

Fact: G is a descent-oriented subdifferential

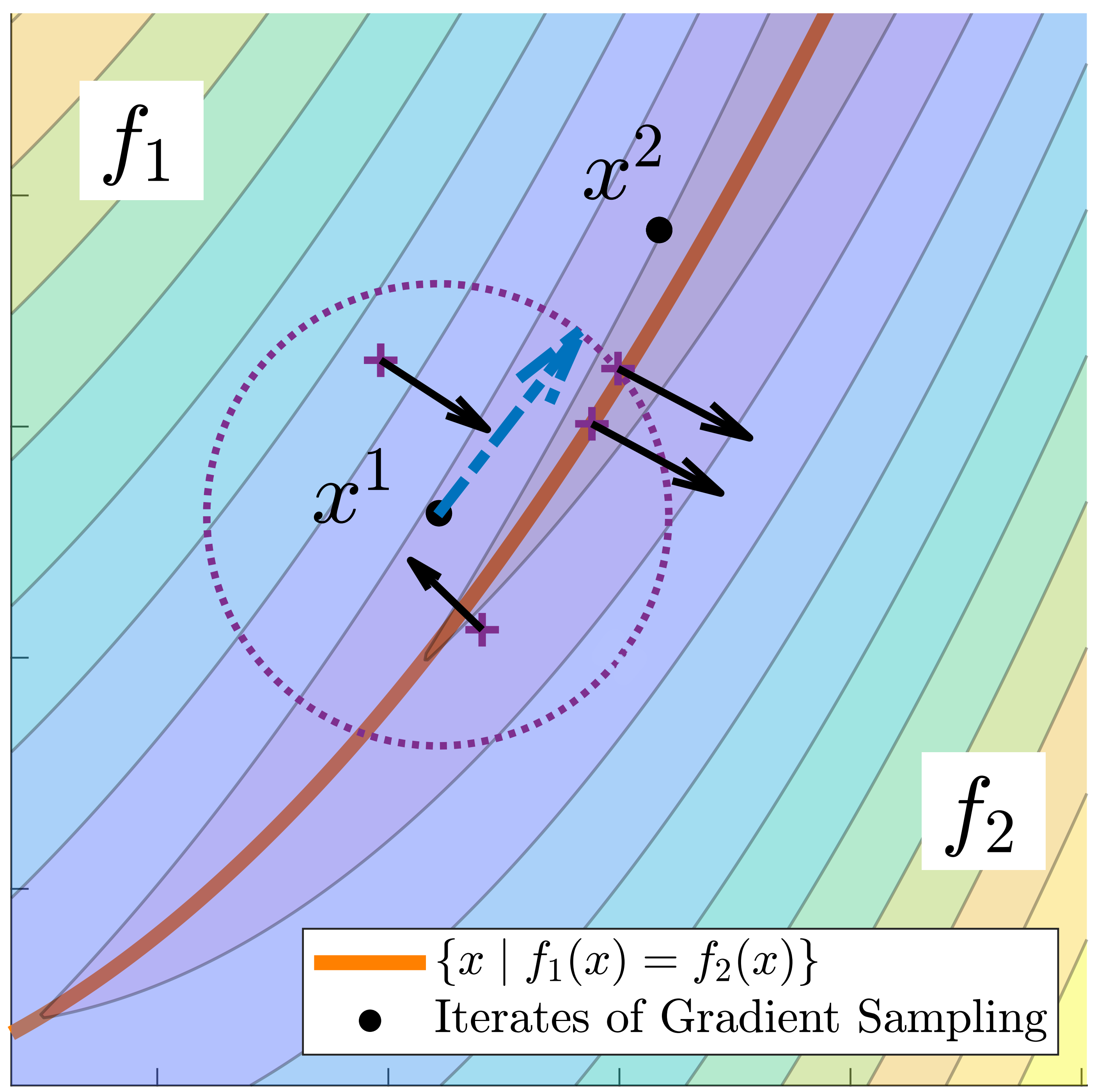
Comparison with Goldstein & Bundle

A nonconvex piecewise smooth function $f(x) = \frac{1}{4}(x_1 - 1)^2 + \sum_{i=1}^{6} \left| x_{i+1} - 2(x_i)^2 + 1 \right|$ with $f^* = 0$

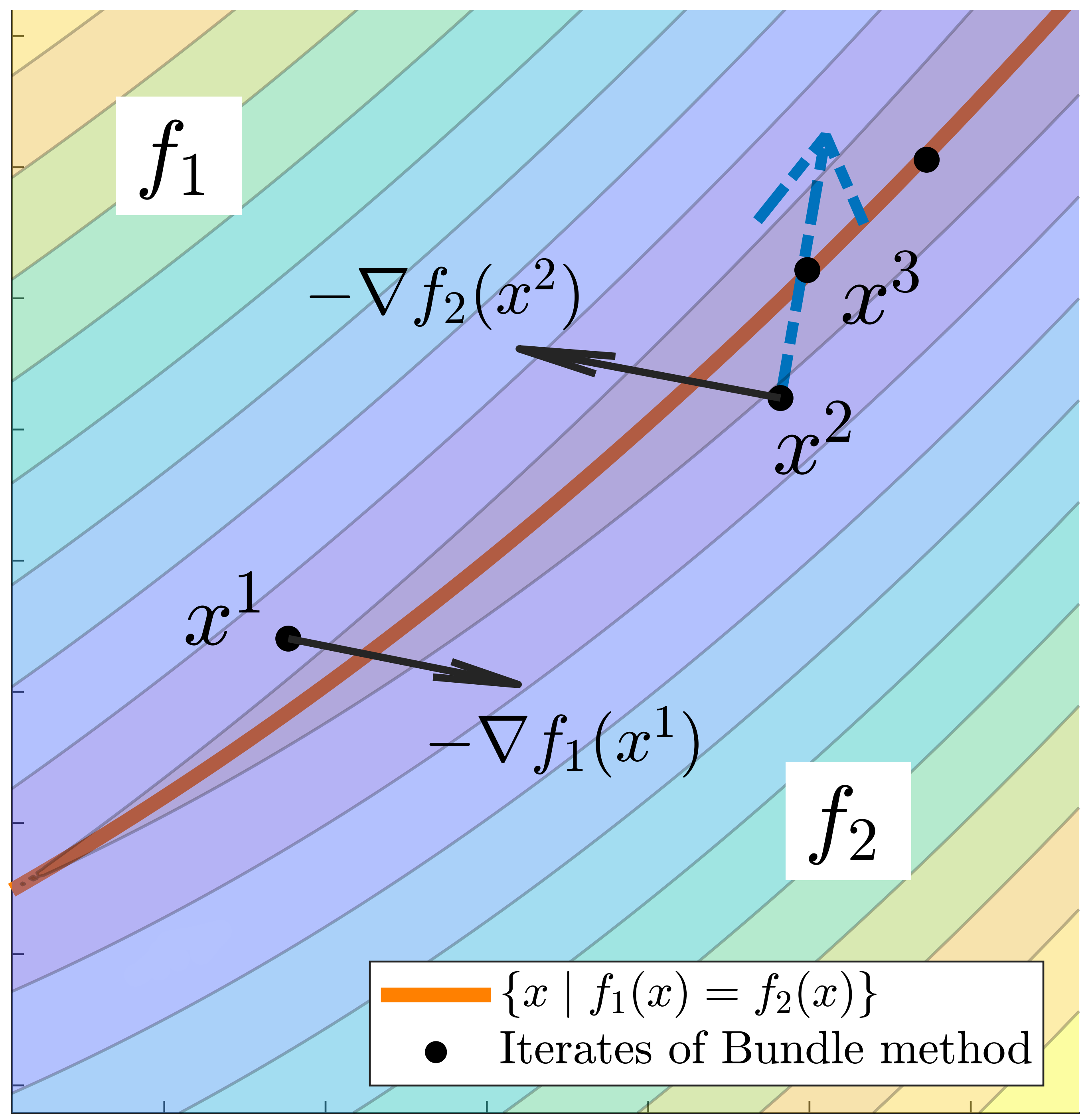
- **SRDescent**: the descent-oriented subgradient method + $G(x, \epsilon)$ via subgradient regularization



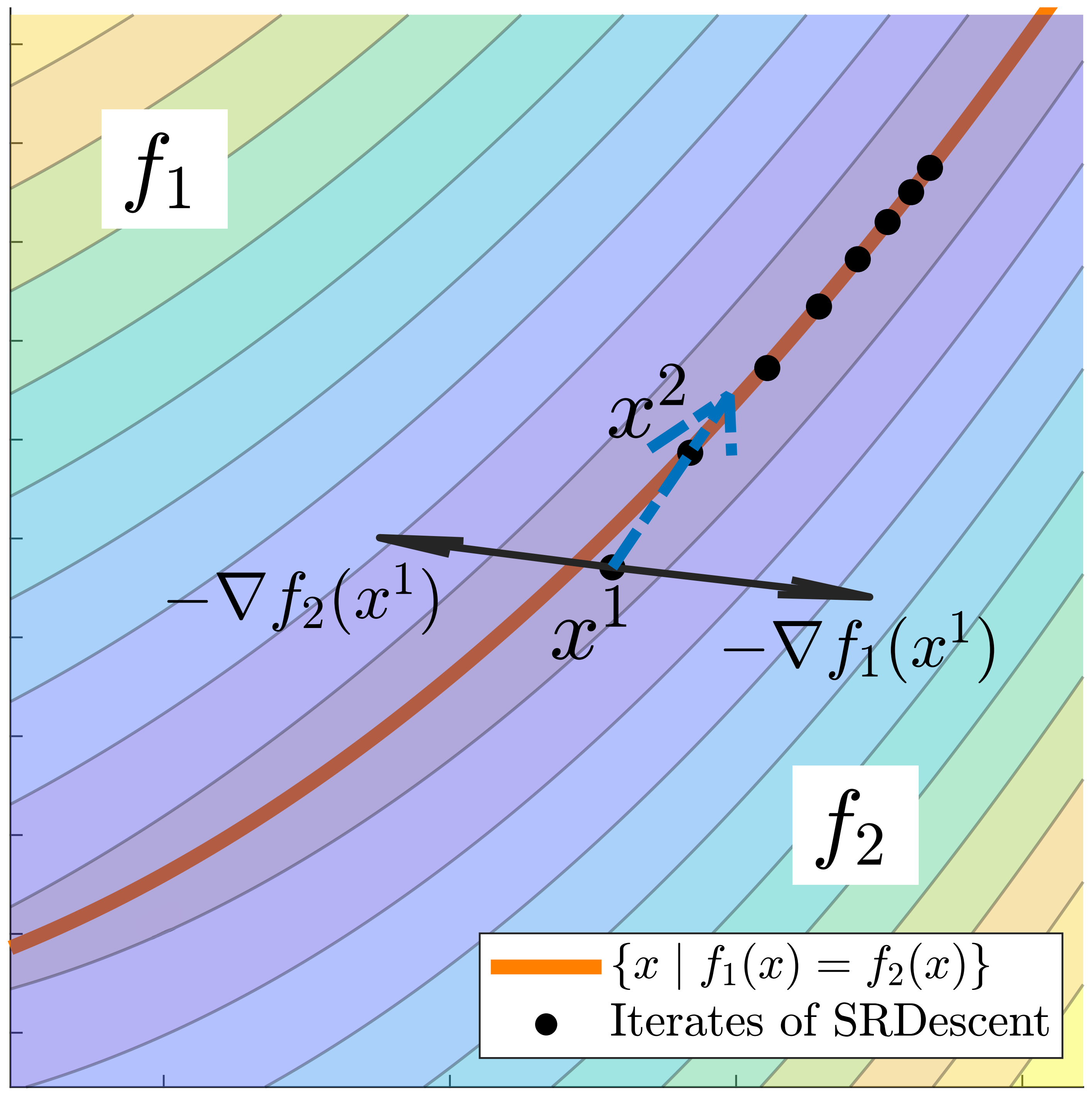
Comparison with Goldstein & Bundle



Gradient Sampling



Bundle method



Subgradient Regularization

combining (sub)gradients at nearby points

Comparison with the prox-linear method

A piecewise linear approximation of $f(x) = \max\{f_1(x), f_2(x)\}$ at x^k:

$$f(x; x^k) = \max_{i=1,2} \left\{ f_i(x^k) + \nabla f_i(x^k)^\top (x - x^k) \right\}.$$

The prox-linear update:

$$x^{k+1} = \operatorname*{argmin}_x \left\{ f(x; x^k) + \frac{1}{2\epsilon} \|x - x^k\|^2 \right\},$$

Comparison with the prox-linear method

A piecewise linear approximation of $f(x) = \max\{f_1(x), f_2(x)\}$ at x^k:

$$f(x; x^k) = \max_{i=1,2} \left\{ f_i(x^k) + \nabla f_i(x^k)^\top (x - x^k) \right\}.$$

The prox-linear update:

$$x^{k+1} = \operatorname*{argmin}_x \left\{ f(x; x^k) + \frac{1}{2\epsilon} \|x - x^k\|^2 \right\},$$

⇒ A minimax formulation:

$$x^{k+1} = \operatorname*{argmin}_x \left\{ \max_{y \in \Delta^2} \sum_{i=1}^{2} y_i \big(f_i(x^k) + \nabla f_i(x^k)^\top (x - x^k) \big) + \frac{1}{2\epsilon} \|x - x^k\|^2 \right\}$$

Comparison with the prox-linear method

A piecewise linear approximation of $f(x) = \max\{f_1(x), f_2(x)\}$ at x^k:

$$f(x; x^k) = \max_{i=1,2} \left\{ f_i(x^k) + \nabla f_i(x^k)^\top (x - x^k) \right\}.$$

The prox-linear update:

$$x^{k+1} = \operatorname*{argmin}_x \left\{ f(x; x^k) + \frac{1}{2\epsilon} \|x - x^k\|^2 \right\},$$

⇒ A minimax formulation:

$$\begin{aligned} x^{k+1} &= \operatorname*{argmin}_x \left\{ \max_{y \in \Delta^2} \sum_{i=1}^{2} y_i \big(f_i(x^k) + \nabla f_i(x^k)^\top (x - x^k) \big) + \frac{1}{2\epsilon} \|x - x^k\|^2 \right\} \\ &= x^k - \epsilon \big[\bar{y}_1 \nabla f_1(x^k) + \bar{y}_2 \nabla f_2(x^k) \big], \end{aligned}$$

where $\bar{y} \in \operatorname*{argmax}_{y \in \Delta^2} \left\{ y_1 f_1(x) + y_2 f_2(x) - \frac{\epsilon}{2} \underset{\text{subgradient regularization}}{\boxed{\|y_1 \nabla f_1(x) + y_2 \nabla f_2(x)\|^2}} \right\}$

Comparison with the prox-linear method

A piecewise linear approximation of $f(x) = \max\{f_1(x), f_2(x)\}$ at x^k:

$$f(x; x^k) = \max_{i=1,2} \left\{ f_i(x^k) + \nabla f_i(x^k)^\top (x - x^k) \right\}.$$

The prox-linear update:

$$x^{k+1} = \operatorname*{argmin}_x \left\{ f(x; x^k) + \frac{1}{2\epsilon} \|x - x^k\|^2 \right\},$$

Observation: For $f(x) = \max\{f_1(x), f_2(x)\}$,

$G(x, \epsilon)$ via subgradient regularization $\iff$ the prox-linear update with stepsize ϵ

- A dual interpretation of the prox-linear method
- can be extended to composite function (convex) ∘ (smooth) by conjugate duality

Subgradien regularization beyond composite structure

For the marginal function:

$$f(x) = \max_{y \in Y} \ \varphi(x, y)$$

- Y is convex and compact, φ is C^1 and concave in y

Subgradien regularization beyond composite structure

For the marginal function:

$$f(x) = \max_{y \in Y} \ \varphi(x, y)$$

- Y is convex and compact, φ is C^1 and concave in y

$$G(x, \epsilon) = \bigcup \left\{ \nabla_x \varphi(x, \bar{y}) \;\middle|\; \bar{y} \in \operatorname*{argmax}_{y \in Y} \left\{ \varphi(x, y) - \frac{\epsilon}{2} \underset{\text{subgradient regularization}}{\boxed{\|\nabla_x \varphi(x, y)\|}}^2 \right\} \right\}$$

Subgradien regularization beyond composite structure

For the marginal function:

$$f(x) = \max_{y \in Y} \ \varphi(x, y)$$

- Y is convex and compact, φ is C^1 and concave in y

$$G(x, \epsilon) = \bigcup \left\{ \nabla_x \varphi(x, \bar{y}) \ \middle| \ \bar{y} \in \operatorname*{argmax}_{y \in Y} \left\{ \varphi(x, y) - \frac{\epsilon}{2} \boxed{\underset{\text{subgradient regularization}}{\| \nabla_x \varphi(x, y) \|}}^2 \right\} \right\}$$

Extension:

$$f(x) = \left[\max_{y} \ \varphi_0(x, y) \ \ \text{subject to } \varphi_j(x, y) \le 0, j = 1, \cdots, r \right]$$

- Characterize $\partial f(x)$, and apply subgradient regularization

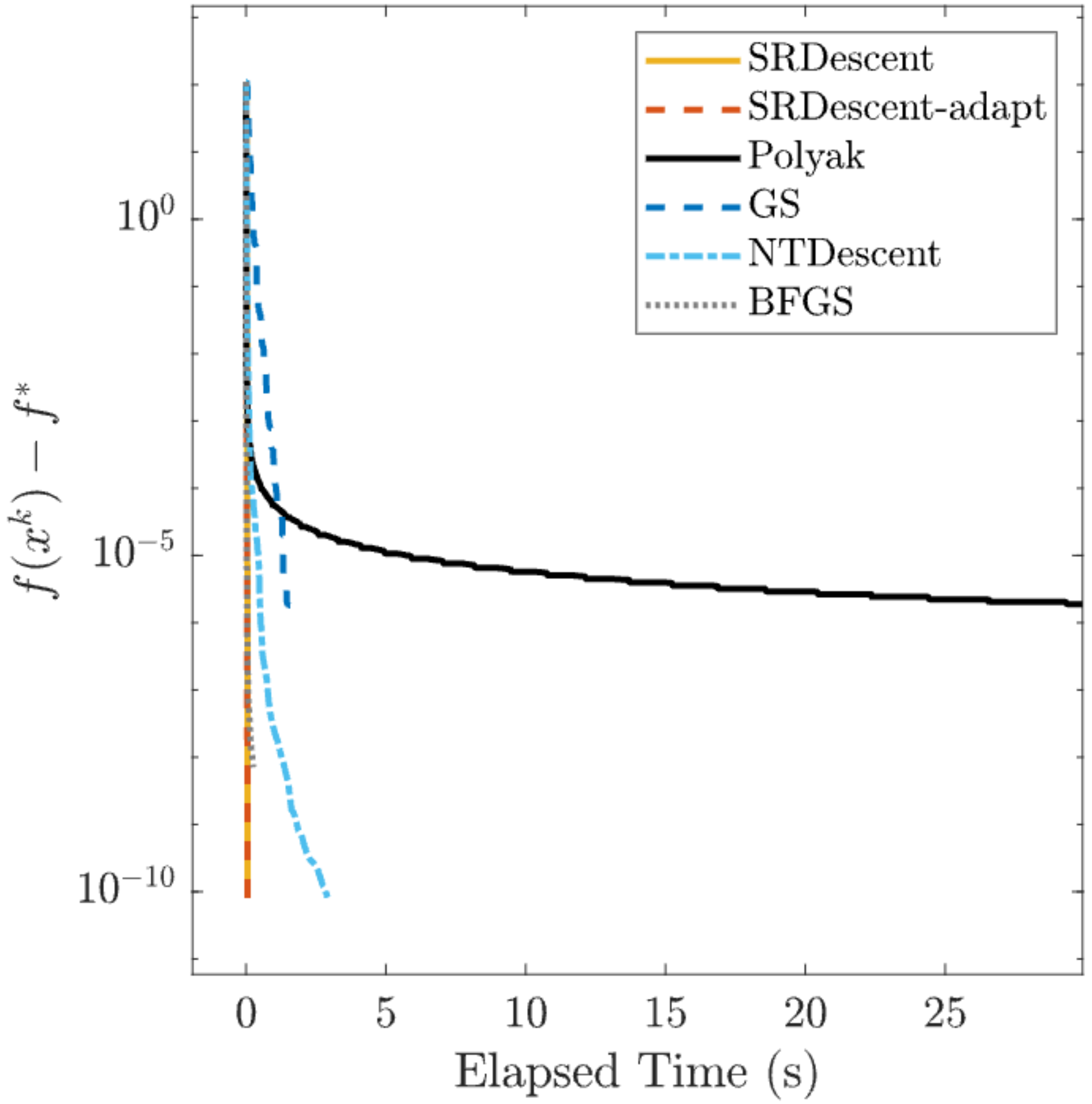
Takeaways

1. A unifying principle for stable descent directions
    - $G(\underset{\substack{\nearrow \\ \text{continuity}}}{x}, \underset{\substack{\nwarrow \\ \text{minimal norm subgradient}}}{\epsilon})$ — ‘gradient’ in nonsmooth optimization
2. Efficient construction of descent directions
    - Subgradient Regularization for marginal functions
    - For (convex) $\circ$ (smooth), Subgradient Regularization $\iff$ Prox-linear update

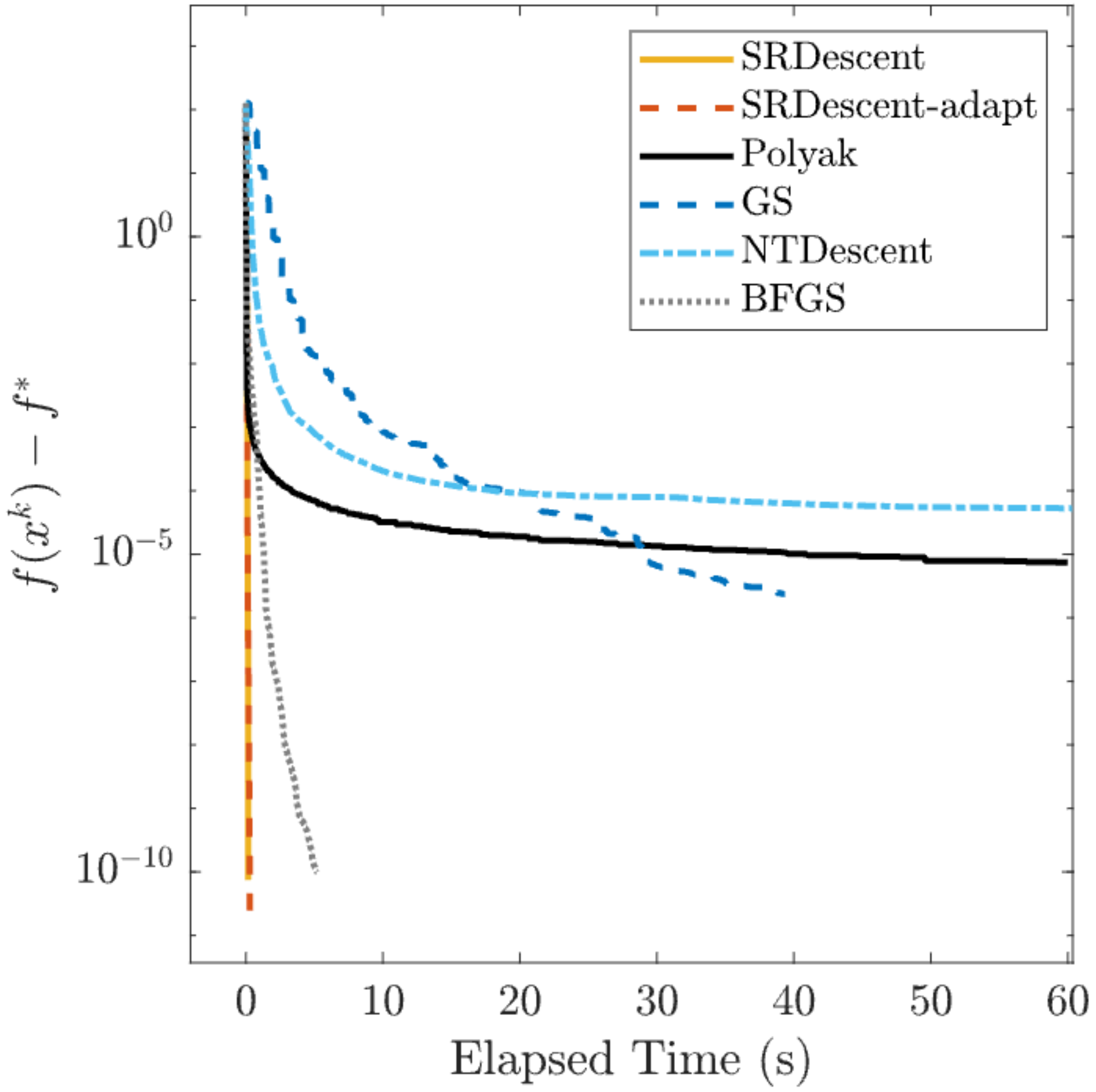
Thank you!

Example: max-of-smooth

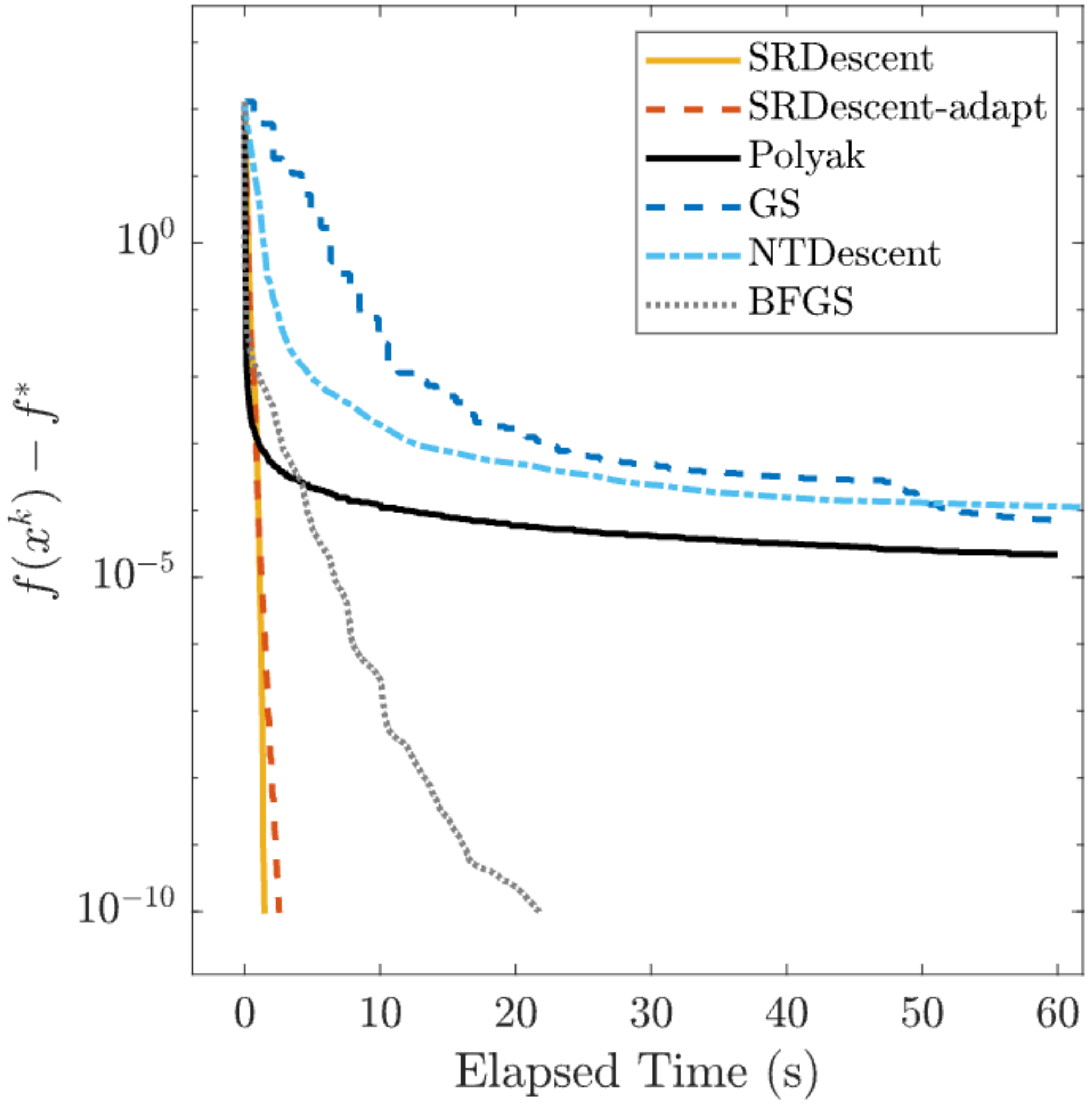
$$f(x) = \max_{1 \le i \le m} \left(g_i^\top x + \frac{1}{2} x^\top H_i x \right)$$



Dim. = 200, #pieces = 10

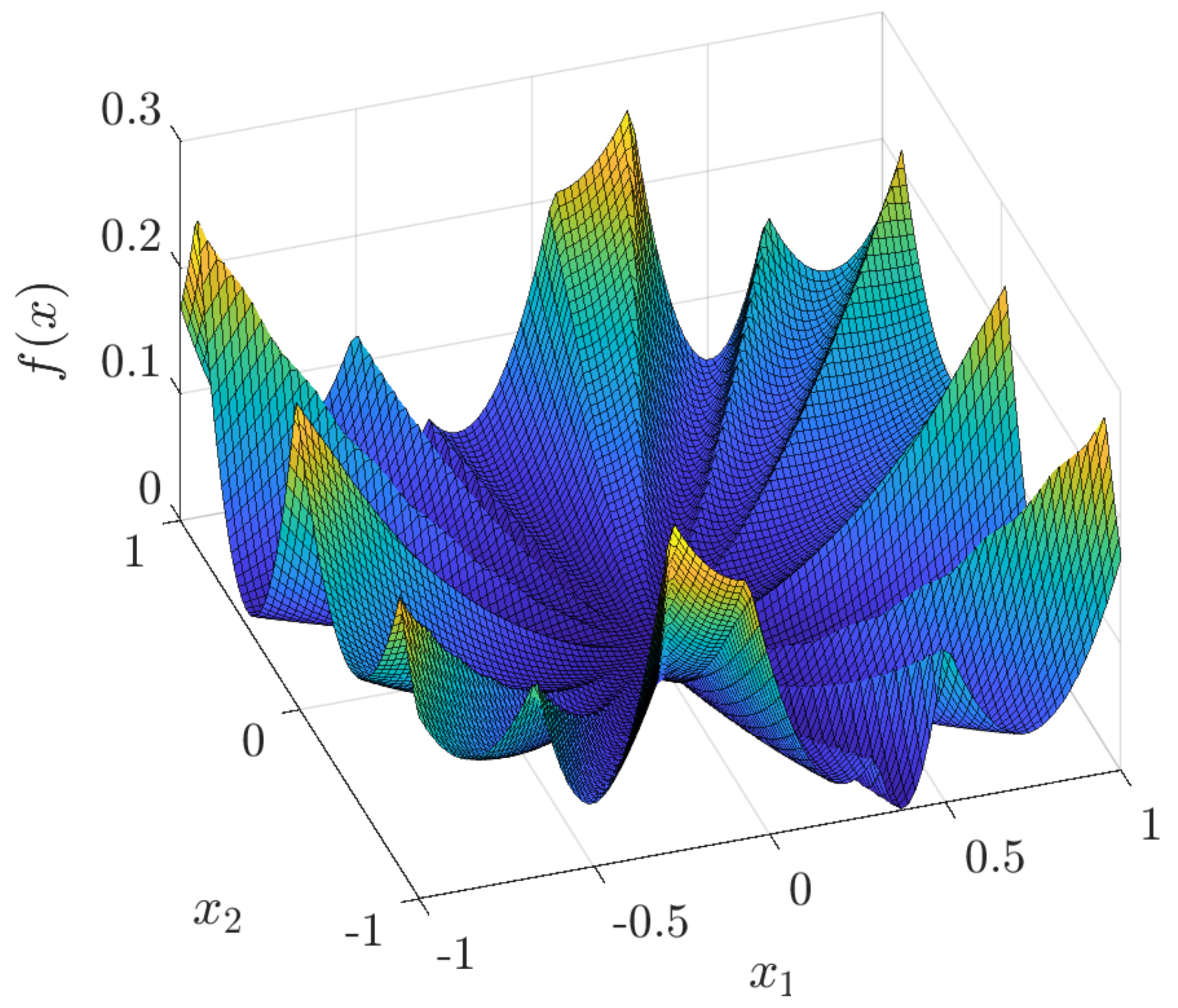


Dim. = 200, #pieces = 100

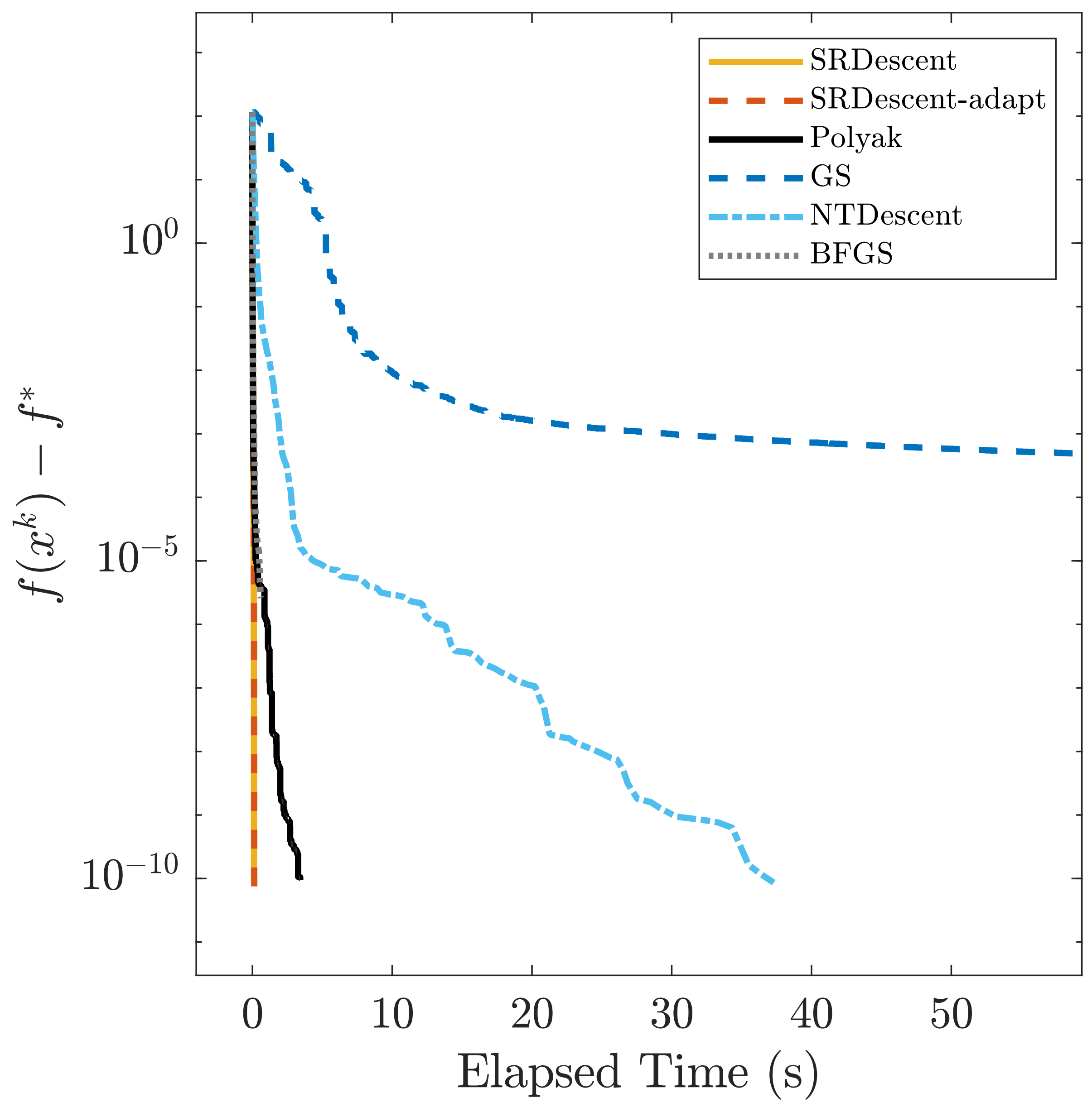


Dim. = 200, #pieces = 200

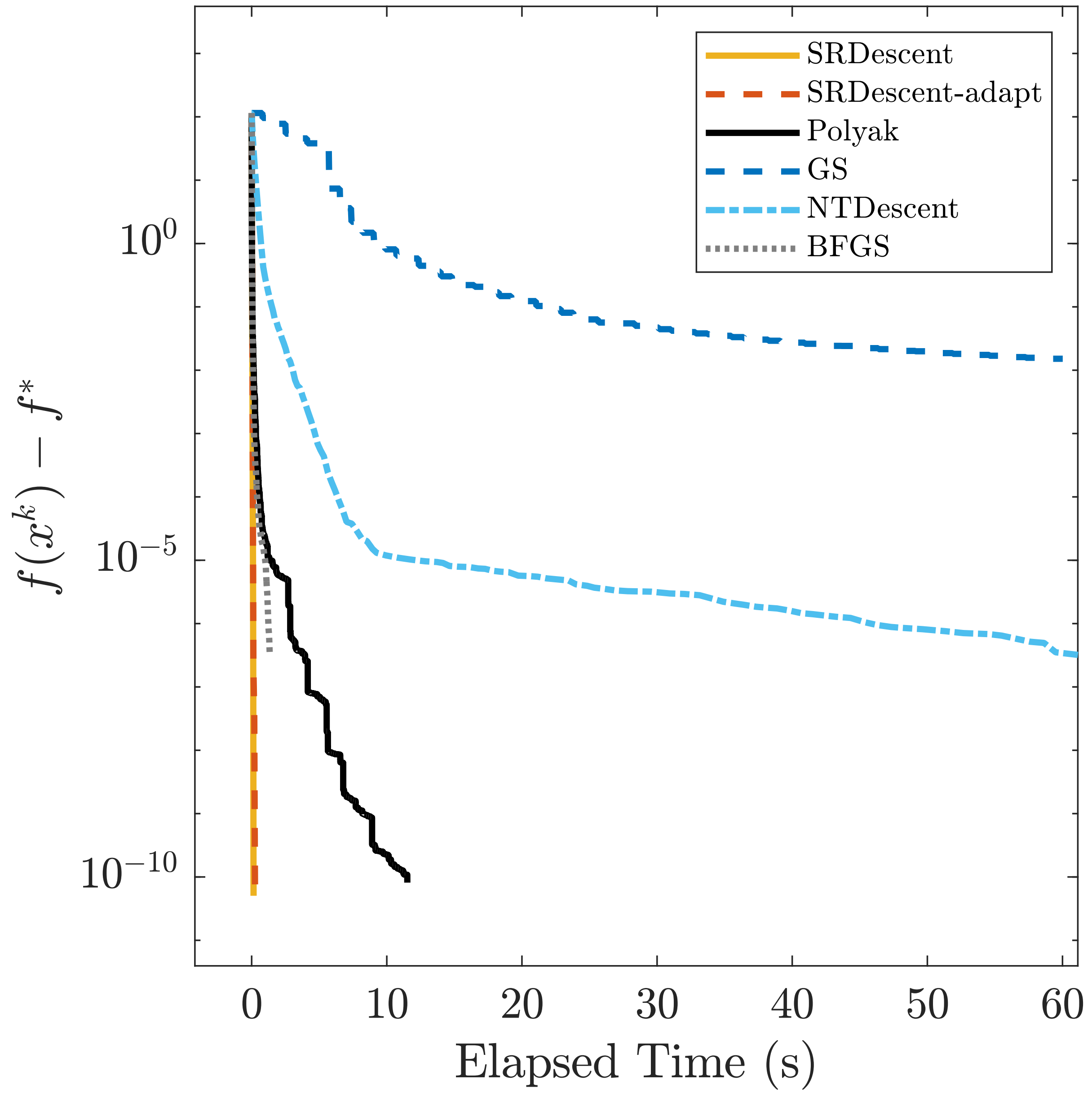
Example: min-of-smooth



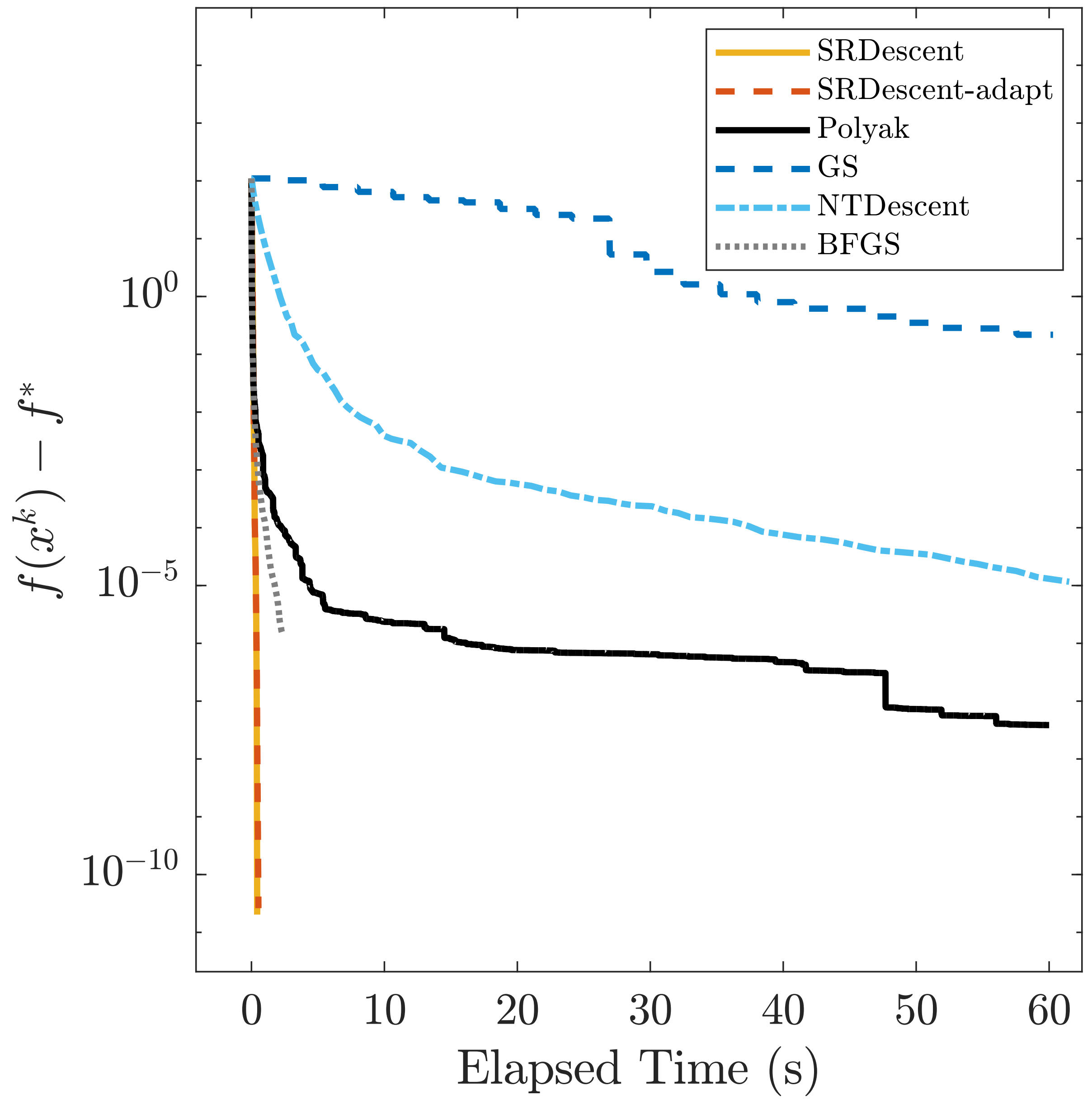
$$f(x) = \min_{1 \le i \le m} \frac{1}{2} \|A_i x - b_i\|^2$$



Dim. = 300, #pieces = 10

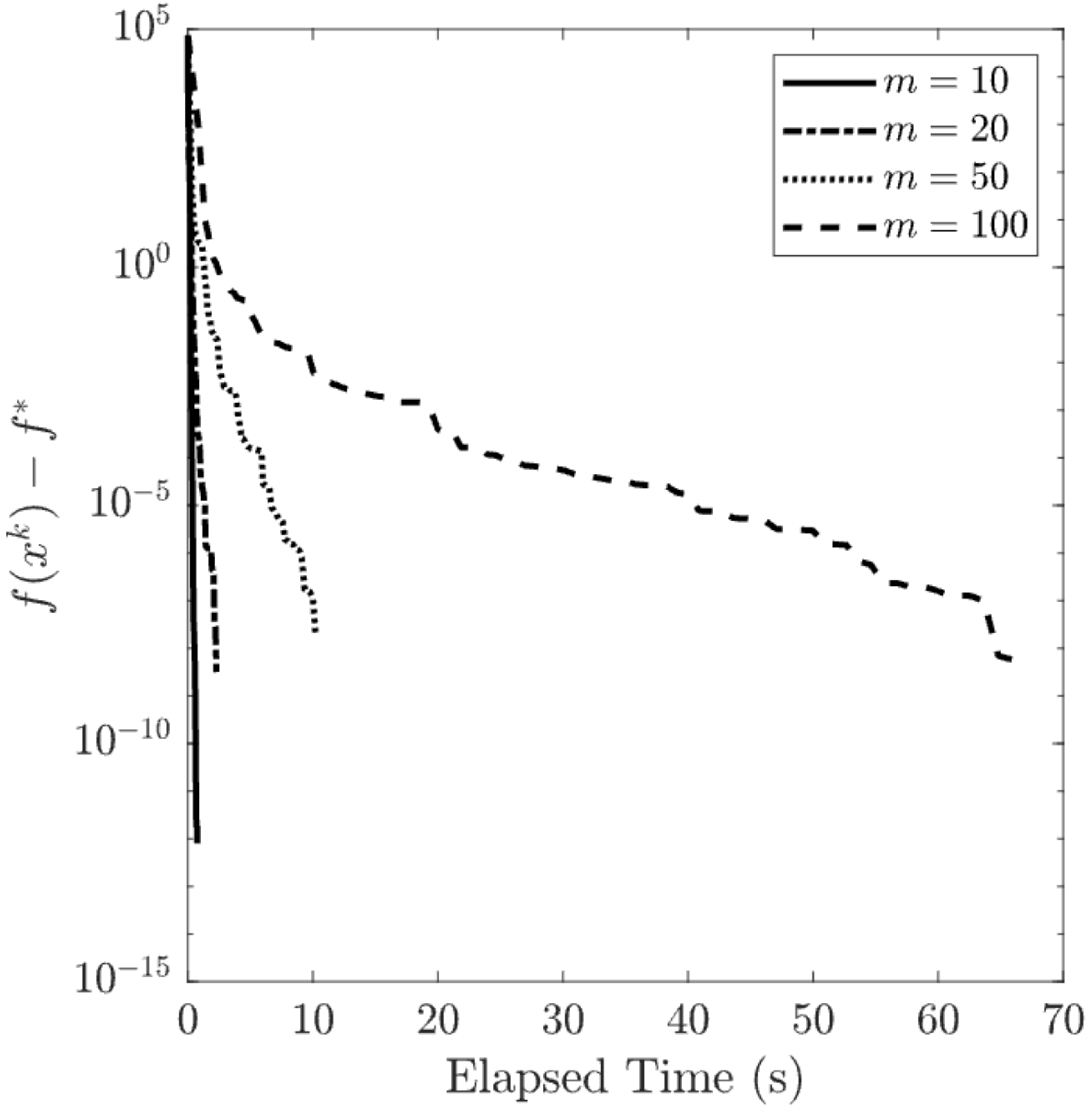


Dim. = 300, #pieces = 50



Dim. = 300, #pieces = 100

Example: general marginal functions



Dim. of x = 300

$$f(x) = \min_{y \in \mathbb{R}^m} \left\{ (c + Dx)^\top y + \frac{1}{2} y^\top Q y + \|x\|^4 \right\}$$

subject to $b - Ax - \mathbf{1} \le Wy \le b - Ax\,.$