Submodular Optimization: From Discrete to Continuous and Back

Hamed Hassani

Amin Karbasi

Penn

Yale

Slides + references: http://iid.yale.edu/icml/icml-20.md/

Submodularity: A General View

- Recall the two equivalent definitions of discrete submodularity:

Yale

Penn

Submodularity: A General View

- Recall the two equivalent definitions of discrete submodularity:

Submodularity: A General View

- Recall the two equivalent definitions of discrete submodularity:

(1) Submodularity $A, B \longrightarrow f(A \cup B) + f(A \cap B) \leq f(A) + f(B)$

Submodularity: A General View

- Recall the two equivalent definitions of discrete submodularity:

(1) Submodularity A, B ➡ $f(A \cup B) + f(A \cap B) \leq f(A) + f(B)$

(2) Diminishing Returns: $A \subseteq B$ ➡ $f(A \cup \{s\}) - f(A) \geq f(B \cup \{s\}) - f(B)$

Submodularity: A General View

- Recall the two equivalent definitions of discrete submodularity:

(1) Submodularity $A, B \longrightarrow f(A \cup B) + f(A \cap B) \leq f(A) + f(B)$

(2) Diminishing Returns: $A \subseteq B \longrightarrow f(A \cup \{s\}) - f(A) \geq f(B \cup \{s\}) - f(B)$

- Natural extension: Submodularity on $\mathbb{R}_+^n$

$\mathbf{x}, \mathbf{y} \in \mathbb{R}^n$:

- $\mathbf{x} \wedge \mathbf{y} = \min(\mathbf{x}, \mathbf{y})$ (coordinate-wise)
- $\mathbf{x} \vee \mathbf{y} = \max(\mathbf{x}, \mathbf{y})$ (coordinate-wise)

Submodularity: A General View

- Recall the two equivalent definitions of discrete submodularity:

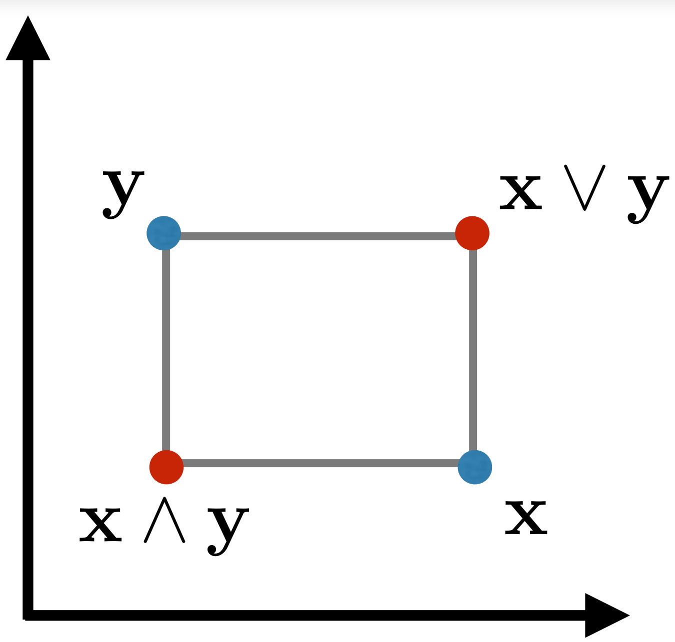
(1) Submodularity $A, B \longrightarrow f(A \cup B) + f(A \cap B) \leq f(A) + f(B)$

(2) Diminishing Returns: $A \subseteq B \longrightarrow f(A \cup \{s\}) - f(A) \geq f(B \cup \{s\}) - f(B)$

- Natural extension: Submodularity on $\mathbb{R}_+^n$

$\mathbf{x}, \mathbf{y} \in \mathbb{R}^n$:

- $\mathbf{x} \wedge \mathbf{y} = \min(\mathbf{x}, \mathbf{y})$ (coordinate-wise)
- $\mathbf{x} \vee \mathbf{y} = \max(\mathbf{x}, \mathbf{y})$ (coordinate-wise)



Submodularity: A General View

- Recall the two equivalent definitions of discrete submodularity:

(1) Submodularity $A, B \longrightarrow f(A \cup B) + f(A \cap B) \leq f(A) + f(B)$

(2) Diminishing Returns: $A \subseteq B \longrightarrow f(A \cup \{s\}) - f(A) \geq f(B \cup \{s\}) - f(B)$

- Natural extension: Submodularity on $\mathbb{R}_+^n$

$\mathbf{x}, \mathbf{y} \in \mathbb{R}^n$:

- $\mathbf{x} \wedge \mathbf{y} = \min(\mathbf{x}, \mathbf{y})$ (coordinate-wise)
- $\mathbf{x} \vee \mathbf{y} = \max(\mathbf{x}, \mathbf{y})$ (coordinate-wise)

$F : \mathcal{X} \subseteq \mathbb{R}^n \to \mathbb{R}$

Submodularity: A General View

- Recall the two equivalent definitions of discrete submodularity:

(1) Submodularity $A, B \longrightarrow f(A \cup B) + f(A \cap B) \leq f(A) + f(B)$

(2) Diminishing Returns: $A \subseteq B \longrightarrow f(A \cup \{s\}) - f(A) \geq f(B \cup \{s\}) - f(B)$

- Natural extension: Submodularity on $\mathbb{R}_+^n$

$\mathbf{x}, \mathbf{y} \in \mathbb{R}^n$:

- $\mathbf{x} \wedge \mathbf{y} = \min(\mathbf{x}, \mathbf{y})$ (coordinate-wise)
- $\mathbf{x} \vee \mathbf{y} = \max(\mathbf{x}, \mathbf{y})$ (coordinate-wise)

$F : \mathcal{X} \subseteq \mathbb{R}^n \to \mathbb{R}$

Assumptions:

(1) $\mathcal{X}$ is a box: $\mathcal{X} = \prod_{i=1}^{n} [0, u_i]$

(2) $F(\mathbf{0}) = 0$

Submodularity: A General View

- Recall the two equivalent definitions of discrete submodularity:

(1) Submodularity $A, B \longrightarrow f(A \cup B) + f(A \cap B) \leq f(A) + f(B)$

(2) Diminishing Returns: $A \subseteq B \longrightarrow f(A \cup \{s\}) - f(A) \geq f(B \cup \{s\}) - f(B)$

- Natural extension: Submodularity on $\mathbb{R}_+^n$

$\mathbf{x}, \mathbf{y} \in \mathbb{R}^n$:

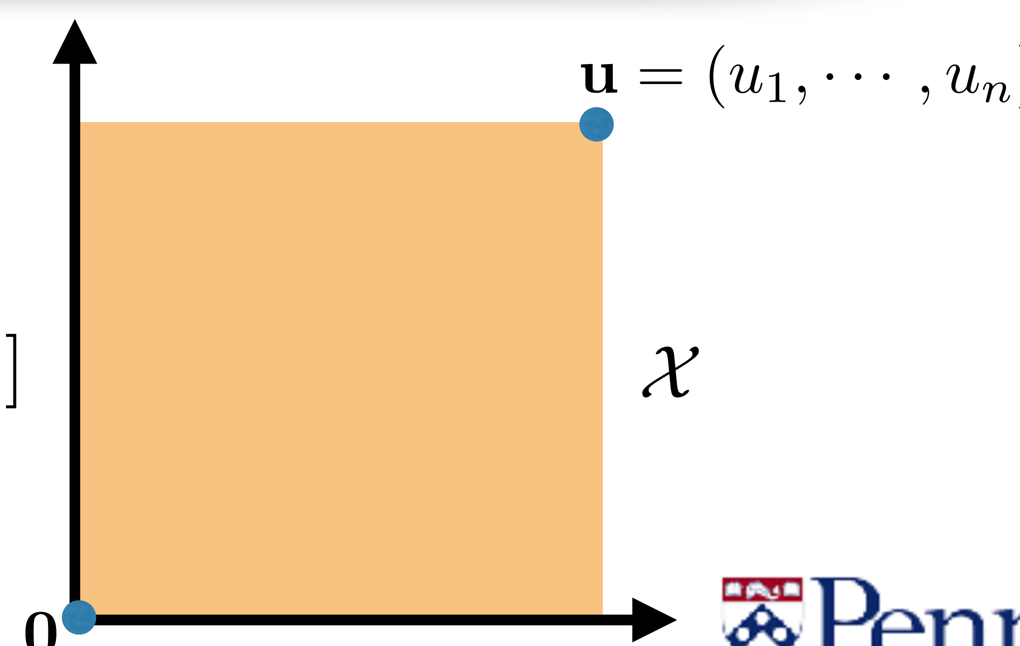
- $\mathbf{x} \wedge \mathbf{y} = \min(\mathbf{x}, \mathbf{y})$ (coordinate-wise)
- $\mathbf{x} \vee \mathbf{y} = \max(\mathbf{x}, \mathbf{y})$ (coordinate-wise)

$F : \mathcal{X} \subseteq \mathbb{R}^n \to \mathbb{R}$

Assumptions:

(1) $\mathcal{X}$ is a box: $\mathcal{X} = \prod_{i=1}^{n} [0, u_i]$

(2) $F(\mathbf{0}) = 0$



Submodularity: A General View

- Recall the two equivalent definitions of discrete submodularity:

(1) Submodularity $A, B \longrightarrow f(A \cup B) + f(A \cap B) \leq f(A) + f(B)$

(2) Diminishing Returns: $A \subseteq B \longrightarrow f(A \cup \{s\}) - f(A) \geq f(B \cup \{s\}) - f(B)$

- Natural extension: Submodularity on $\mathbb{R}_+^n$

$\mathbf{x}, \mathbf{y} \in \mathbb{R}^n$:

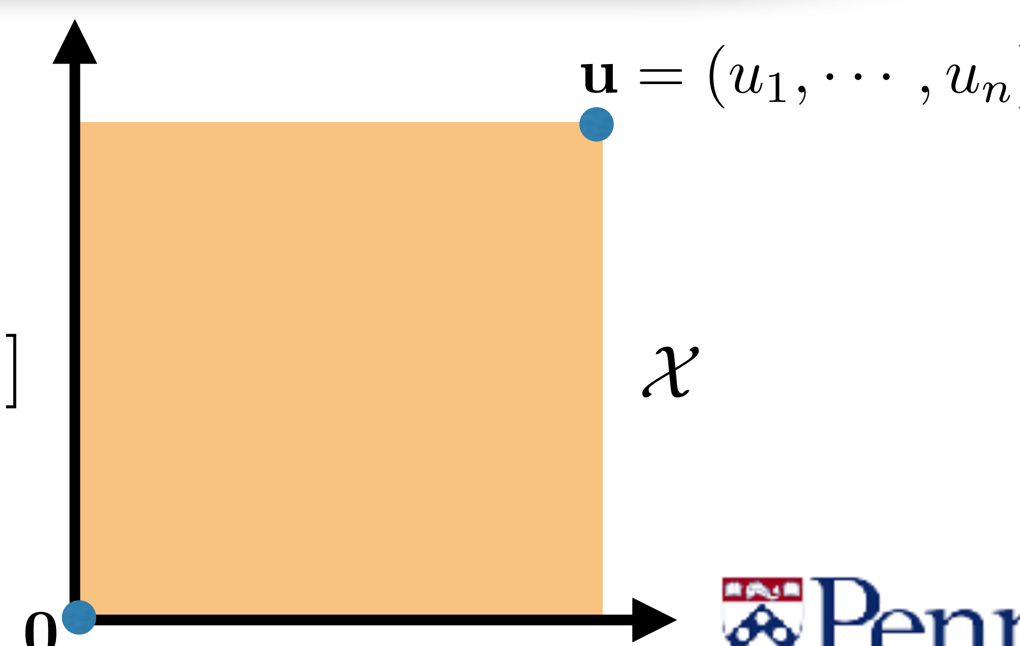
- $\mathbf{x} \wedge \mathbf{y} = \min(\mathbf{x}, \mathbf{y})$ (coordinate-wise)
- $\mathbf{x} \vee \mathbf{y} = \max(\mathbf{x}, \mathbf{y})$ (coordinate-wise)

$F : \mathcal{X} \subseteq \mathbb{R}^n \to \mathbb{R}$

Assumptions:

(1) $\mathcal{X}$ is a box: $\mathcal{X} = \prod_{i=1}^{n} [0, u_i]$

(2) $F(\mathbf{0}) = 0$



Continuous Submodular Functions

- A function $F : \mathcal{X} \subseteq \mathbb{R}^n \to \mathbb{R}$ is called **Continuous Submodular** if

$$\forall\ \mathbf{x}, \mathbf{y} \in \mathcal{X} \quad \longrightarrow \quad F(\mathbf{x} \vee \mathbf{y}) + F(\mathbf{x} \wedge \mathbf{y}) \leq F(\mathbf{x}) + F(\mathbf{y})$$

"Minimizing a submodular function on a lattice", Topkis, 1978

"Maximizing real-valued submodular functions: Primal and dual heuristics for location problems", Wolsey, 1982

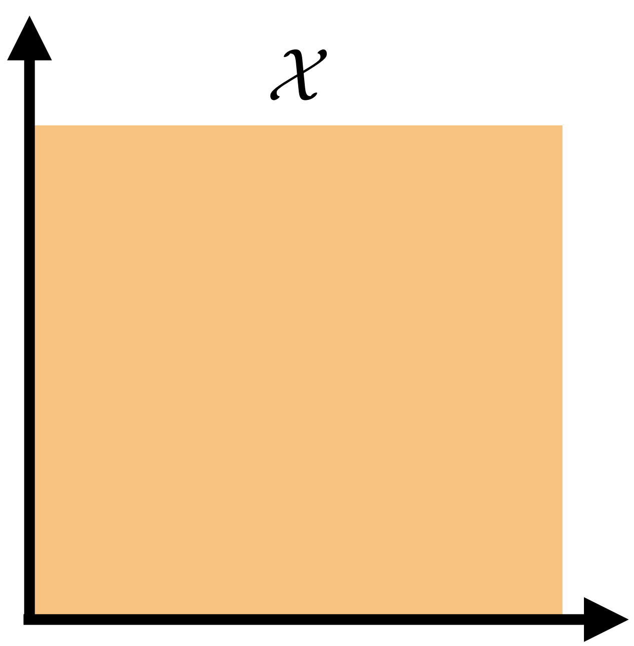
"Submodular Functions and Optimization", Fujishige, 2005

"Submodular functions: from discrete to continuous domains", Bach, 2015

Continuous Submodular Functions

- A function $F : \mathcal{X} \subseteq \mathbb{R}^n \to \mathbb{R}$ is called **Continuous Submodular** if

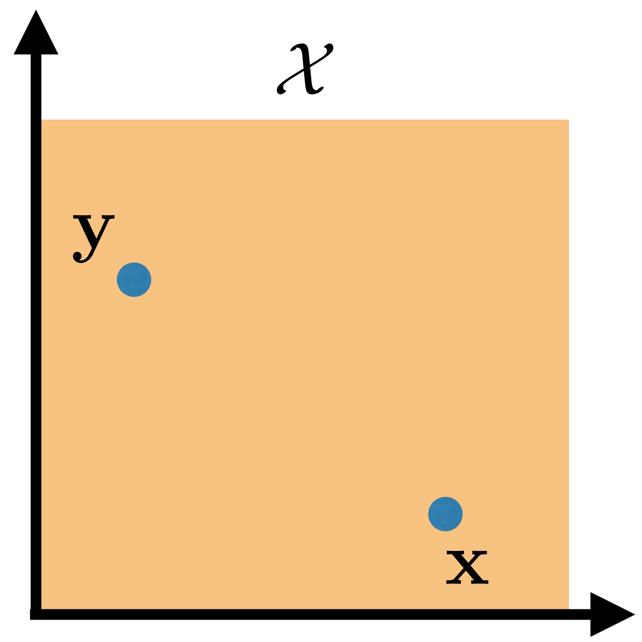
$$\forall\ \mathbf{x}, \mathbf{y} \in \mathcal{X} \quad \Rightarrow \quad F(\mathbf{x} \vee \mathbf{y}) + F(\mathbf{x} \wedge \mathbf{y}) \leq F(\mathbf{x}) + F(\mathbf{y})$$



Continuous Submodular Functions

- A function $F : \mathcal{X} \subseteq \mathbb{R}^n \to \mathbb{R}$ is called **Continuous Submodular** if

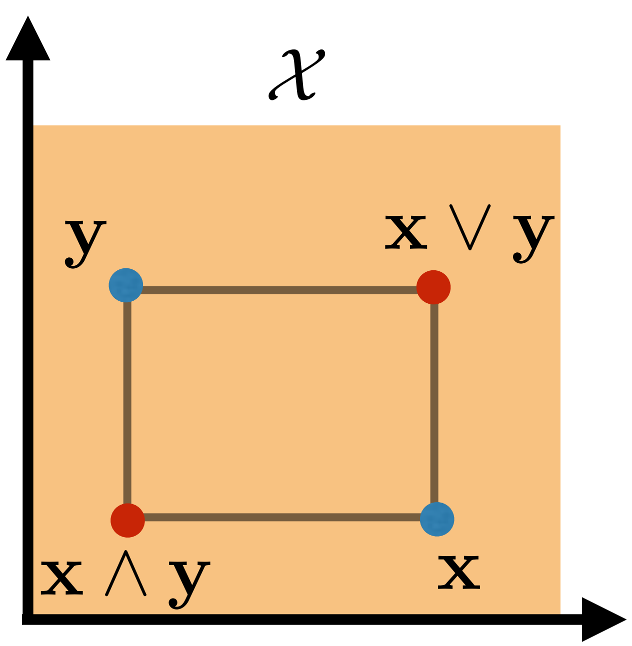
$$\forall\, \mathbf{x}, \mathbf{y} \in \mathcal{X} \quad \Rightarrow \quad F(\mathbf{x} \vee \mathbf{y}) + F(\mathbf{x} \wedge \mathbf{y}) \leq F(\mathbf{x}) + F(\mathbf{y})$$



Continuous Submodular Functions

- A function $F : \mathcal{X} \subseteq \mathbb{R}^n \rightarrow \mathbb{R}$ is called **Continuous Submodular** if

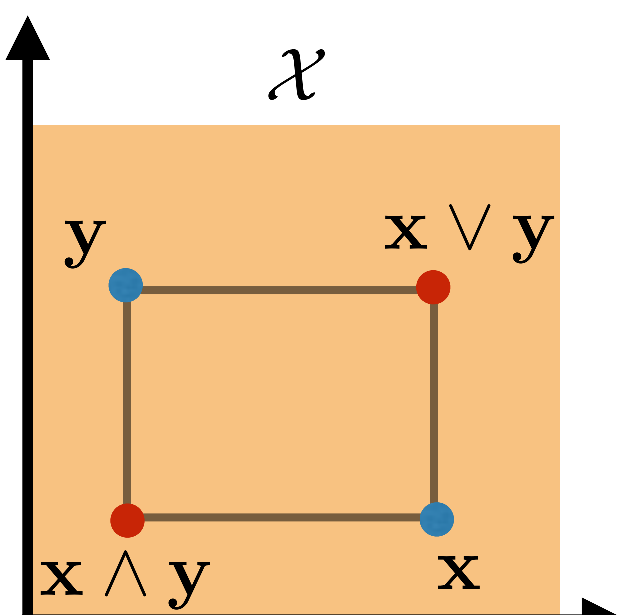
$$\forall\, \mathbf{x}, \mathbf{y} \in \mathcal{X} \quad \Longrightarrow \quad F(\mathbf{x} \vee \mathbf{y}) + F(\mathbf{x} \wedge \mathbf{y}) \leq F(\mathbf{x}) + F(\mathbf{y})$$

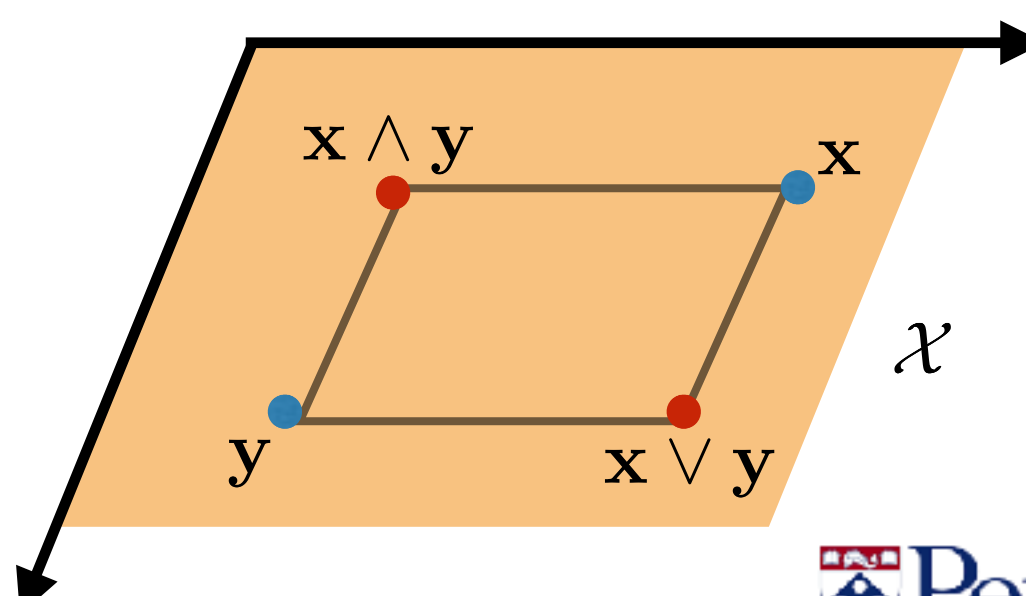


Continuous Submodular Functions

- A function $F : \mathcal{X} \subseteq \mathbb{R}^n \rightarrow \mathbb{R}$ is called **Continuous Submodular** if

$$\forall\, \mathbf{x}, \mathbf{y} \in \mathcal{X} \quad \longrightarrow \quad F(\mathbf{x} \vee \mathbf{y}) + F(\mathbf{x} \wedge \mathbf{y}) \leq F(\mathbf{x}) + F(\mathbf{y})$$

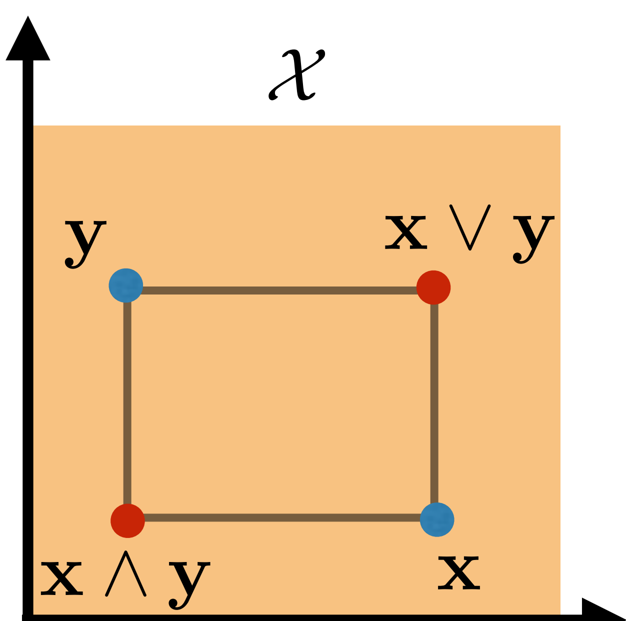


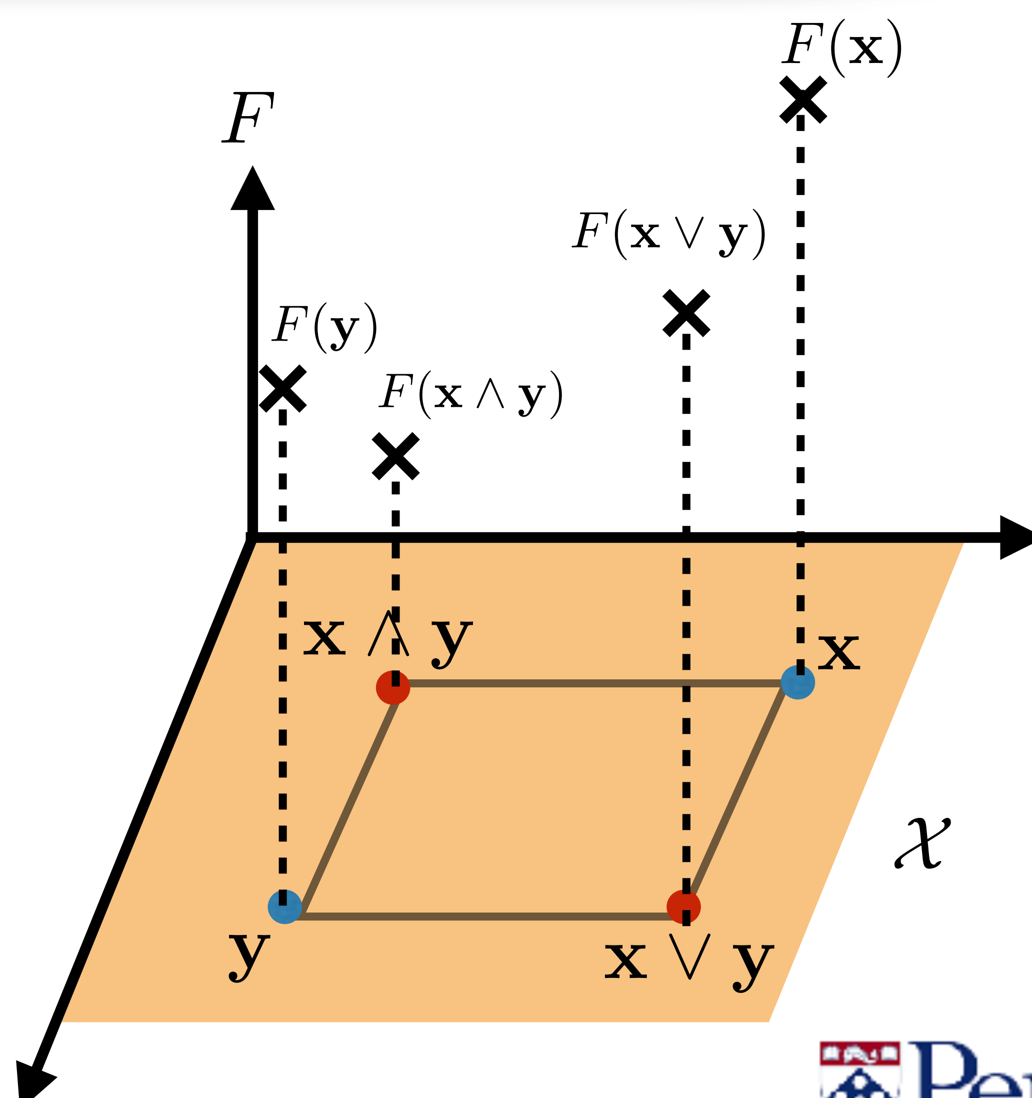


Continuous Submodular Functions

- A function $F : \mathcal{X} \subseteq \mathbb{R}^n \to \mathbb{R}$ is called **Continuous Submodular** if

$$\forall\, \mathbf{x}, \mathbf{y} \in \mathcal{X} \quad \longrightarrow \quad F(\mathbf{x} \vee \mathbf{y}) + F(\mathbf{x} \wedge \mathbf{y}) \leq F(\mathbf{x}) + F(\mathbf{y})$$





Continuous Submodular Functions

- A function $F : \mathcal{X} \subseteq \mathbb{R}^n \to \mathbb{R}$ is called **Continuous Submodular** if

$$\forall\, \mathbf{x}, \mathbf{y} \in \mathcal{X} \quad \Rightarrow \quad F(\mathbf{x} \vee \mathbf{y}) + F(\mathbf{x} \wedge \mathbf{y}) \leq F(\mathbf{x}) + F(\mathbf{y})$$

Continuous Submodular Functions

- A function $F : \mathcal{X} \subseteq \mathbb{R}^n \to \mathbb{R}$ is called **Continuous Submodular** if

$$\forall\ \mathbf{x}, \mathbf{y} \in \mathcal{X} \quad \longrightarrow \quad F(\mathbf{x} \vee \mathbf{y}) + F(\mathbf{x} \wedge \mathbf{y}) \leq F(\mathbf{x}) + F(\mathbf{y})$$

- When twice **differentiable**, continuous submodularity is equivalent to

$$\forall i \neq j, \forall \mathbf{x} \in \mathcal{X}, \quad \frac{\partial^2 F(\mathbf{x})}{\partial x_i \partial x_j} \leq 0$$

"Minimizing a submodular function on a lattice", Topkis, 1978

"Submodular functions: from discrete to continuous domains", Bach, 2015

Continuous Submodular Functions

- A function $F : \mathcal{X} \subseteq \mathbb{R}^n \to \mathbb{R}$ is called **Continuous Submodular** if

$$\forall\ \mathbf{x}, \mathbf{y} \in \mathcal{X} \quad \longrightarrow \quad F(\mathbf{x} \vee \mathbf{y}) + F(\mathbf{x} \wedge \mathbf{y}) \leq F(\mathbf{x}) + F(\mathbf{y})$$

- When twice **differentiable**, continuous submodularity is equivalent to

$$\forall i \neq j, \forall \mathbf{x} \in \mathcal{X}, \quad \frac{\partial^2 F(\mathbf{x})}{\partial x_i \partial x_j} \leq 0 \quad \longrightarrow \quad \nabla^2 F(\mathbf{x}) = \begin{pmatrix} * & \leq 0 & \leq 0 \\ \leq 0 & * & \leq 0 \\ \leq 0 & \leq 0 & * \end{pmatrix}$$

Hessian

"Minimizing a submodular function on a lattice", Topkis, 1978

"Submodular functions: from discrete to continuous domains", Bach, 2015

Yale

Penn

Continuous Submodular Functions

- A function $F: \mathcal{X} \subseteq \mathbb{R}^n \to \mathbb{R}$ is called **Continuous Submodular** if

$$\forall\ \mathbf{x}, \mathbf{y} \in \mathcal{X} \longrightarrow F(\mathbf{x} \vee \mathbf{y}) + F(\mathbf{x} \wedge \mathbf{y}) \leq F(\mathbf{x}) + F(\mathbf{y})$$

- When twice **differentiable**, continuous submodularity is equivalent to

$$\forall i \neq j, \forall \mathbf{x} \in \mathcal{X}, \quad \frac{\partial^2 F(\mathbf{x})}{\partial x_i \partial x_j} \leq 0 \longrightarrow \nabla^2 F(\mathbf{x}) = \begin{pmatrix} * & \leq 0 & \leq 0 \\ \leq 0 & * & \leq 0 \\ \leq 0 & \leq 0 & * \end{pmatrix}$$

Hessian

- BUT, does it mean Diminishing Returns (DR)?

DR-Submodular Functions

(1) Submodularity A, B → $f(A \cup B) + f(A \cap B) \leq f(A) + f(B)$

(2) Diminishing Returns: $A \subseteq B$ → $f(A \cup \{s\}) - f(A) \geq f(B \cup \{s\}) - f(B)$

DR-Submodular Functions

(1) Submodularity $A, B \longrightarrow f(A \cup B) + f(A \cap B) \leq f(A) + f(B)$

(2) Diminishing Returns: $A \subseteq B \longrightarrow f(A \cup \{s\}) - f(A) \geq f(B \cup \{s\}) - f(B)$

- Natural extension: Diminishing returns on $\mathbb{R}^n$

$$\forall \mathbf{x} \leq \mathbf{y} \longrightarrow F(\mathbf{x} + \alpha\, \mathbf{e}_i) - F(\mathbf{x}) \geq F(\mathbf{y} + \alpha\, \mathbf{e}_i) - F(\mathbf{y}) \qquad \alpha > 0$$

$\mathbf{e}_i = i\text{-th standard vector}$

"Guaranteed non-convex optimization: Submodular maximization over continuous domains", Bian, Mirzasoleiman, Buhmann, Krause, 2017

DR-Submodular Functions

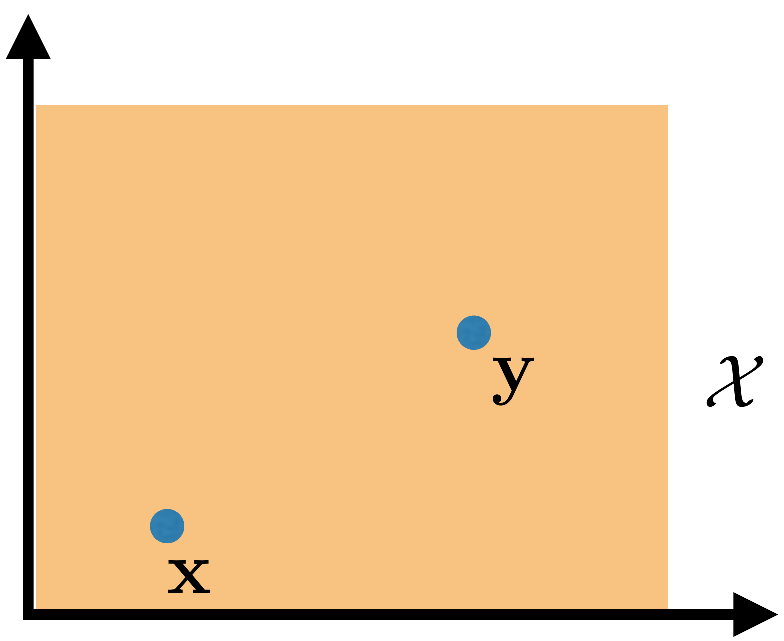
(1) Submodularity $A, B \longrightarrow f(A \cup B) + f(A \cap B) \le f(A) + f(B)$

(2) Diminishing Returns: $A \subseteq B \longrightarrow f(A \cup \{s\}) - f(A) \ge f(B \cup \{s\}) - f(B)$

- Natural extension: Diminishing returns on $\mathbb{R}^n$

$$\forall \mathbf{x} \le \mathbf{y} \longrightarrow F(\mathbf{x} + \alpha\, \mathbf{e}_i) - F(\mathbf{x}) \ge F(\mathbf{y} + \alpha\, \mathbf{e}_i) - F(\mathbf{y}) \qquad \alpha > 0$$

$\mathbf{e}_i = i$-th standard vector



DR-Submodular Functions

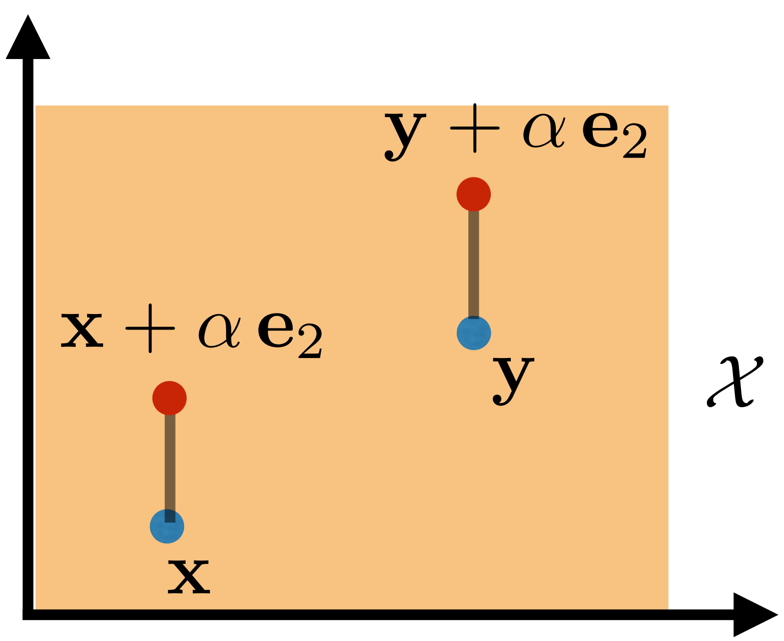
(1) Submodularity $A, B \longrightarrow f(A \cup B) + f(A \cap B) \leq f(A) + f(B)$

(2) Diminishing Returns: $A \subseteq B \longrightarrow f(A \cup \{s\}) - f(A) \geq f(B \cup \{s\}) - f(B)$

- Natural extension: Diminishing returns on $\mathbb{R}^n$

$$\forall \mathbf{x} \leq \mathbf{y} \longrightarrow F(\mathbf{x} + \alpha\, \mathbf{e}_i) - F(\mathbf{x}) \geq F(\mathbf{y} + \alpha\, \mathbf{e}_i) - F(\mathbf{y}) \qquad \alpha > 0$$

$\mathbf{e}_i = i$-th standard vector



DR-Submodular Functions

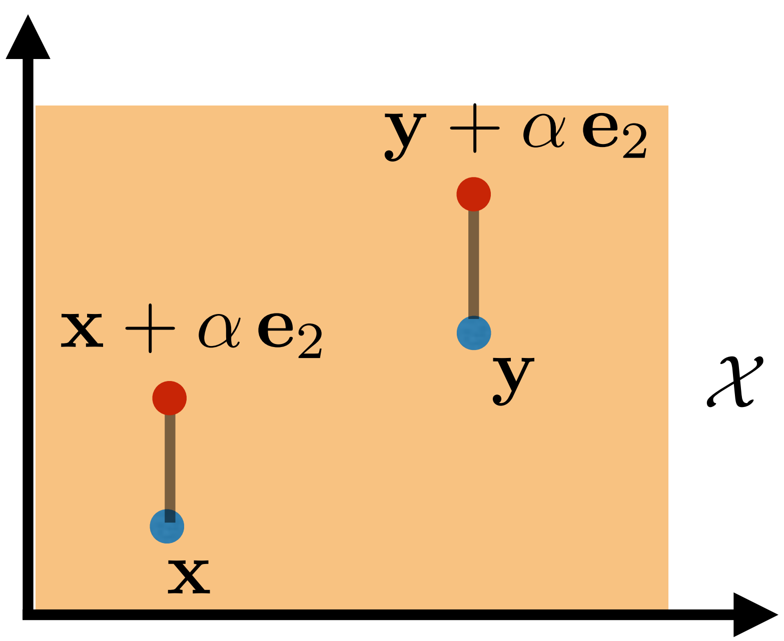
(1) Submodularity $A, B \longrightarrow f(A \cup B) + f(A \cap B) \leq f(A) + f(B)$

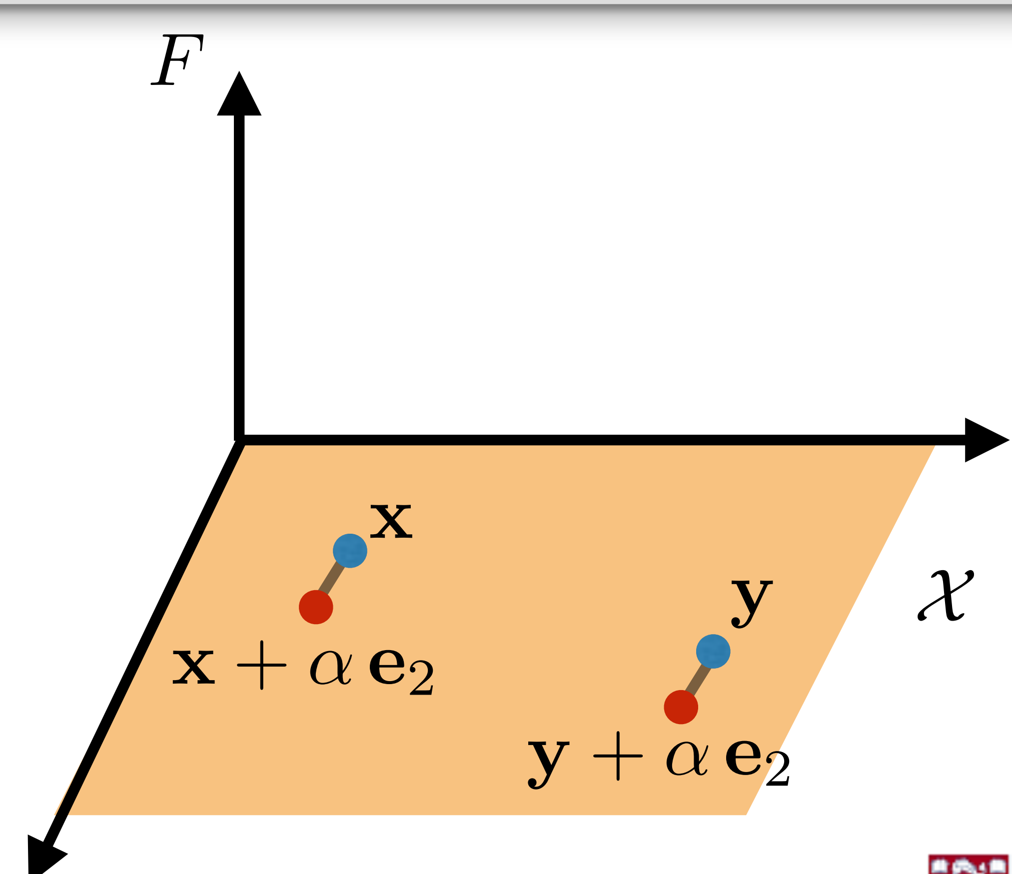
(2) Diminishing Returns: $A \subseteq B \longrightarrow f(A \cup \{s\}) - f(A) \geq f(B \cup \{s\}) - f(B)$

- Natural extension: Diminishing returns on $\mathbb{R}^n$

$$\forall \mathbf{x} \leq \mathbf{y} \longrightarrow F(\mathbf{x} + \alpha\,\mathbf{e}_i) - F(\mathbf{x}) \geq F(\mathbf{y} + \alpha\,\mathbf{e}_i) - F(\mathbf{y}) \qquad \alpha > 0$$

$\mathbf{e}_i = i$-th standard vector





Yale

Penn

DR-Submodular Functions

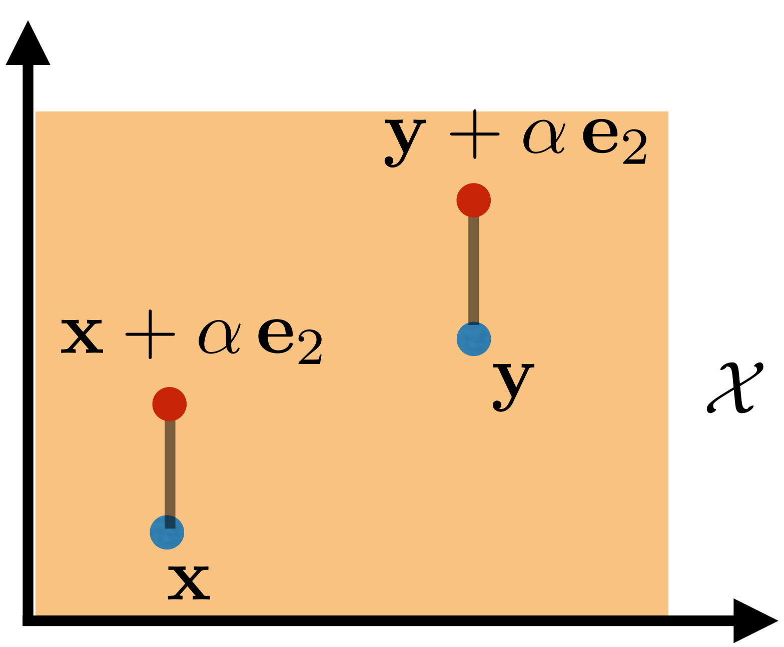
(1) Submodularity $A, B \longrightarrow f(A \cup B) + f(A \cap B) \leq f(A) + f(B)$

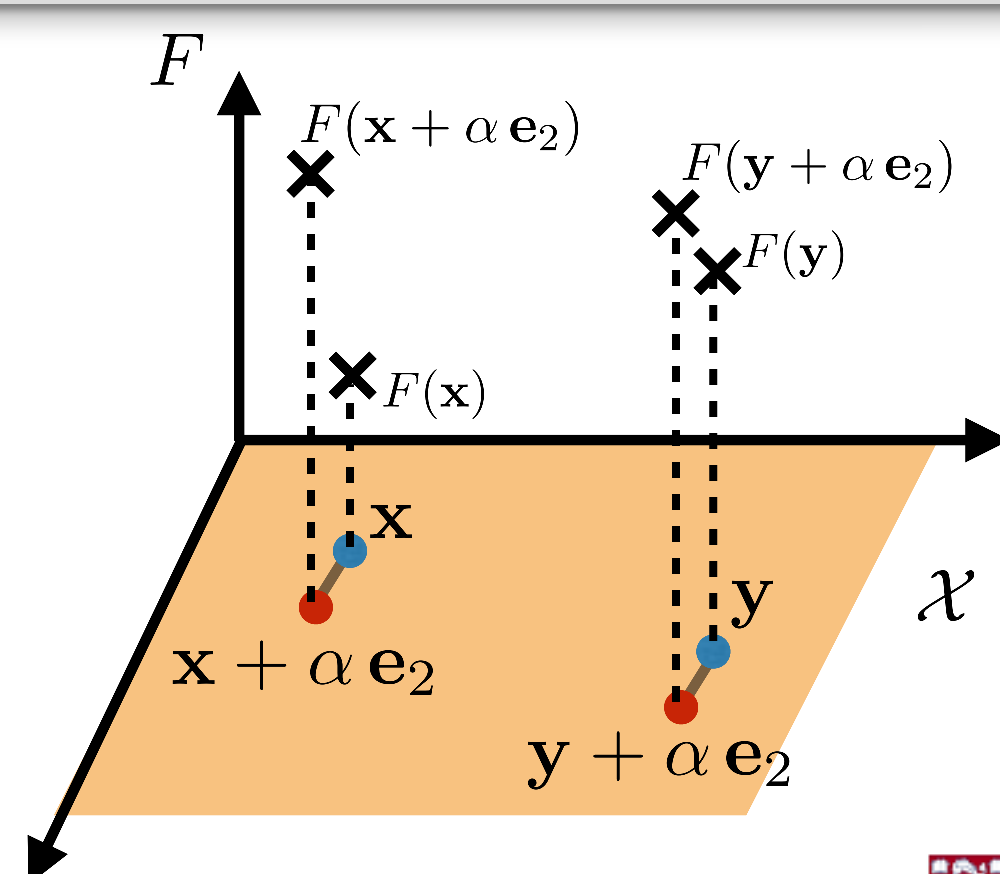
(2) Diminishing Returns: $A \subseteq B \longrightarrow f(A \cup \{s\}) - f(A) \geq f(B \cup \{s\}) - f(B)$

- Natural extension: Diminishing returns on $\mathbb{R}^n$

$$\forall \mathbf{x} \leq \mathbf{y} \longrightarrow F(\mathbf{x} + \alpha\, \mathbf{e}_i) - F(\mathbf{x}) \geq F(\mathbf{y} + \alpha\, \mathbf{e}_i) - F(\mathbf{y}) \qquad \alpha > 0$$

$\mathbf{e}_i = i\text{-th standard vector}$





DR-Submodular Functions

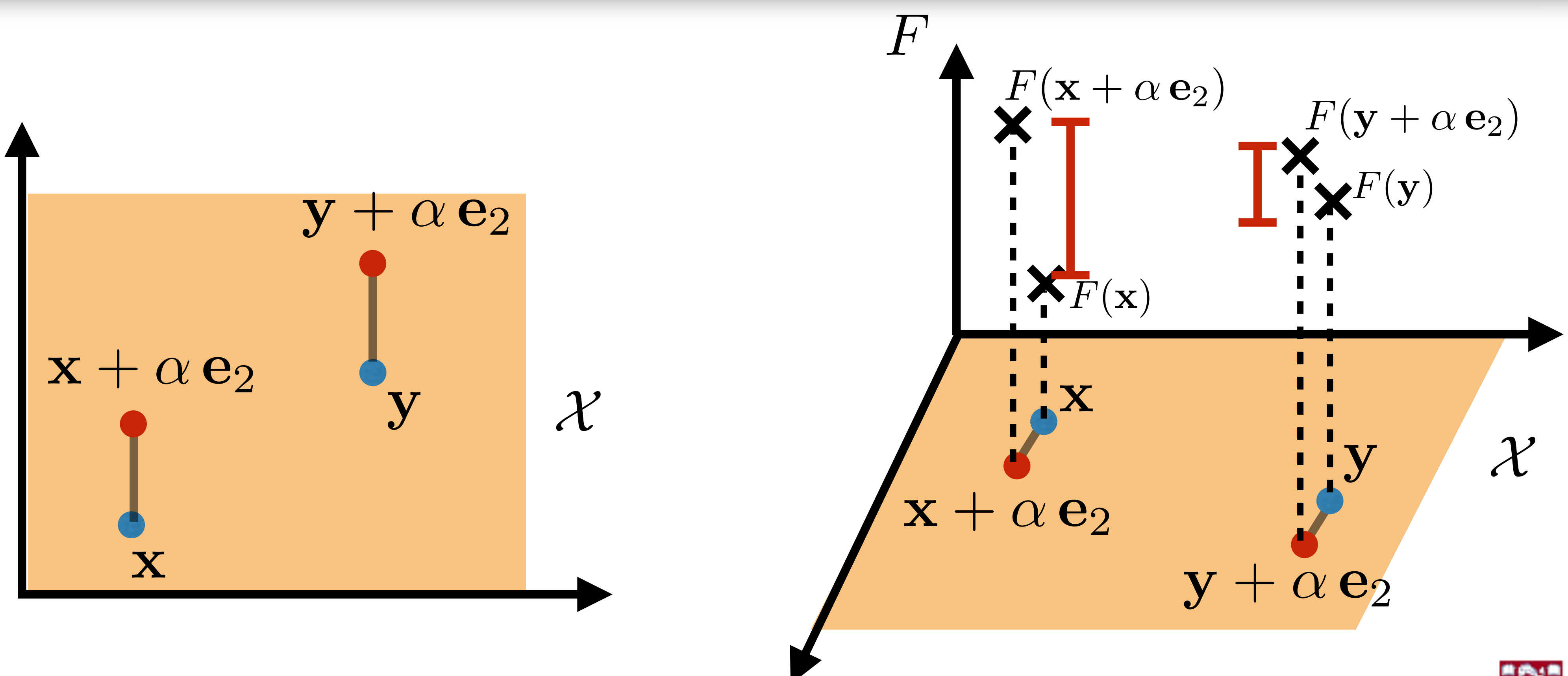
(1) Submodularity $A, B \longrightarrow f(A \cup B) + f(A \cap B) \leq f(A) + f(B)$

(2) Diminishing Returns: $A \subseteq B \longrightarrow f(A \cup \{s\}) - f(A) \geq f(B \cup \{s\}) - f(B)$

- Natural extension: Diminishing returns on $\mathbb{R}^n$

$$\forall \mathbf{x} \leq \mathbf{y} \longrightarrow F(\mathbf{x} + \alpha\, \mathbf{e}_i) - F(\mathbf{x}) \geq F(\mathbf{y} + \alpha\, \mathbf{e}_i) - F(\mathbf{y}) \qquad \alpha > 0$$

$\mathbf{e}_i = i$-th standard vector



Yale Penn

DR-Submodular Functions

(1) Submodularity $A, B \longrightarrow f(A \cup B) + f(A \cap B) \leq f(A) + f(B)$

(2) Diminishing Returns: $A \subseteq B \longrightarrow f(A \cup \{s\}) - f(A) \geq f(B \cup \{s\}) - f(B)$

- Natural extension: Diminishing returns on $\mathbb{R}^n$

$$\forall \mathbf{x} \leq \mathbf{y} \longrightarrow F(\mathbf{x} + \alpha\, \mathbf{e}_i) - F(\mathbf{x}) \geq F(\mathbf{y} + \alpha\, \mathbf{e}_i) - F(\mathbf{y}) \qquad \alpha > 0$$

$\mathbf{e}_i = i\text{-th standard vector}$

DR-Submodular Functions

(1) Submodularity $A, B \longrightarrow f(A \cup B) + f(A \cap B) \leq f(A) + f(B)$

(2) Diminishing Returns: $A \subseteq B \longrightarrow f(A \cup \{s\}) - f(A) \geq f(B \cup \{s\}) - f(B)$

- Natural extension: Diminishing returns on $\mathbb{R}^n$

$$\forall \mathbf{x} \leq \mathbf{y} \longrightarrow F(\mathbf{x} + \alpha\, \mathbf{e}_i) - F(\mathbf{x}) \geq F(\mathbf{y} + \alpha\, \mathbf{e}_i) - F(\mathbf{y}) \qquad \alpha > 0$$

$\mathbf{e}_i = i\text{-th standard vector}$

F is DR-submodular $\longleftrightarrow$ F is concave along positive directions

DR-Submodular Functions

(1) Submodularity $A, B \longrightarrow f(A \cup B) + f(A \cap B) \leq f(A) + f(B)$

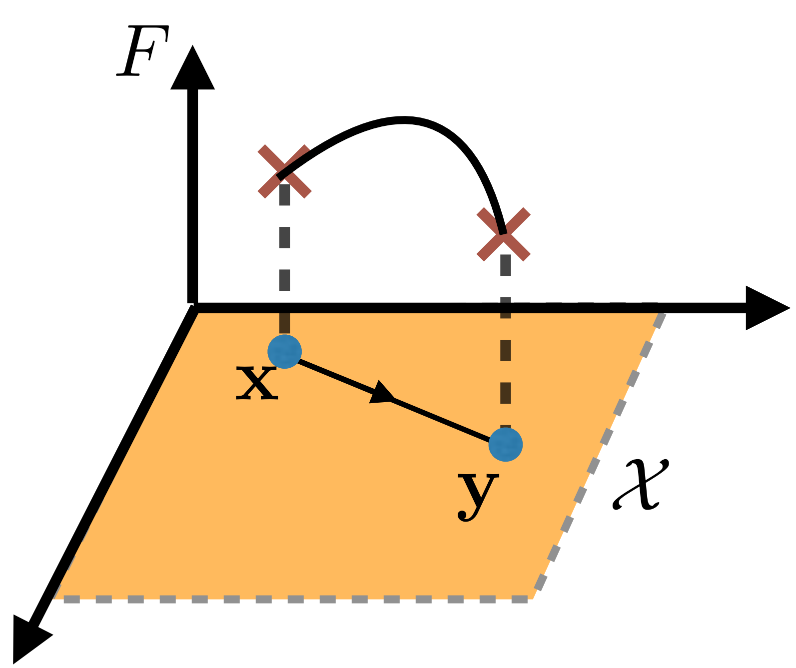
(2) Diminishing Returns: $A \subseteq B \longrightarrow f(A \cup \{s\}) - f(A) \geq f(B \cup \{s\}) - f(B)$

- Natural extension: Diminishing returns on $\mathbb{R}^n$

$$\forall \mathbf{x} \leq \mathbf{y} \longrightarrow F(\mathbf{x} + \alpha\, \mathbf{e}_i) - F(\mathbf{x}) \geq F(\mathbf{y} + \alpha\, \mathbf{e}_i) - F(\mathbf{y}) \qquad \alpha > 0$$

$\mathbf{e}_i = i\text{-th standard vector}$

F is DR-submodular $\longleftrightarrow$ F is concave along positive directions



DR-Submodular Functions

(1) Submodularity $A, B \longrightarrow f(A \cup B) + f(A \cap B) \leq f(A) + f(B)$

(2) Diminishing Returns: $A \subseteq B \longrightarrow f(A \cup \{s\}) - f(A) \geq f(B \cup \{s\}) - f(B)$

- Natural extension: Diminishing returns on $\mathbb{R}^n$

$$\forall \mathbf{x} \leq \mathbf{y} \longrightarrow F(\mathbf{x} + \alpha\, \mathbf{e}_i) - F(\mathbf{x}) \geq F(\mathbf{y} + \alpha\, \mathbf{e}_i) - F(\mathbf{y}) \qquad \alpha > 0$$

$\mathbf{e}_i = i\text{-th standard vector}$

DR-Submodular Functions

(1) Submodularity $A, B \longrightarrow f(A \cup B) + f(A \cap B) \leq f(A) + f(B)$

(2) Diminishing Returns: $A \subseteq B \longrightarrow f(A \cup \{s\}) - f(A) \geq f(B \cup \{s\}) - f(B)$

- Natural extension: Diminishing returns on $\mathbb{R}^n$

$$\forall \mathbf{x} \leq \mathbf{y} \longrightarrow F(\mathbf{x} + \alpha\, \mathbf{e}_i) - F(\mathbf{x}) \geq F(\mathbf{y} + \alpha\, \mathbf{e}_i) - F(\mathbf{y}) \qquad \alpha > 0$$

$\mathbf{e}_i = i\text{-th standard vector}$

- When twice **differentiable**, DR-Submodularity is equivalent to

$$\forall i, j,\ \forall \mathbf{x} \in \mathcal{X}, \quad \frac{\partial^2 F(\mathbf{x})}{\partial x_i \partial x_j} \leq 0$$

DR-Submodular Functions

(1) Submodularity $A, B \longrightarrow f(A \cup B) + f(A \cap B) \leq f(A) + f(B)$

(2) Diminishing Returns: $A \subseteq B \longrightarrow f(A \cup \{s\}) - f(A) \geq f(B \cup \{s\}) - f(B)$

- Natural extension: Diminishing returns on $\mathbb{R}^n$

$$\forall \mathbf{x} \leq \mathbf{y} \longrightarrow F(\mathbf{x} + \alpha\, \mathbf{e}_i) - F(\mathbf{x}) \geq F(\mathbf{y} + \alpha\, \mathbf{e}_i) - F(\mathbf{y}) \qquad \alpha > 0$$

$\mathbf{e}_i = i\text{-th standard vector}$

- When twice **differentiable**, DR-Submodularity is equivalent to

$$\forall i, j,\ \forall \mathbf{x} \in \mathcal{X}, \quad \frac{\partial^2 F(\mathbf{x})}{\partial x_i \partial x_j} \leq 0 \longrightarrow \nabla^2 F(\mathbf{x}) = \begin{pmatrix} \leq 0 & \leq 0 & \leq 0 \\ \leq 0 & \leq 0 & \leq 0 \\ \leq 0 & \leq 0 & \leq 0 \end{pmatrix}$$

Hessian

"Guaranteed non-convex optimization: Submodular maximization over continuous domains", Bian, Mirzasoleiman, Buhmann, Krause, 2017

Submodularity: A General View

Summary

cont. submodular

$$\begin{pmatrix} * & \leq 0 & \leq 0 \\ \leq 0 & * & \leq 0 \\ \leq 0 & \leq 0 & * \end{pmatrix}$$

DR-submodular

$$\begin{pmatrix} \leq 0 & \leq 0 & \leq 0 \\ \leq 0 & \leq 0 & \leq 0 \\ \leq 0 & \leq 0 & \leq 0 \end{pmatrix}$$

Submodularity: A General View

Summary

cont. submodular

$$\begin{pmatrix} * & \leq 0 & \leq 0 \\ \leq 0 & * & \leq 0 \\ \leq 0 & \leq 0 & * \end{pmatrix}$$

DR-submodular

$$\begin{pmatrix} \leq 0 & \leq 0 & \leq 0 \\ \leq 0 & \leq 0 & \leq 0 \\ \leq 0 & \leq 0 & \leq 0 \end{pmatrix}$$

Submodularity: A General View

Summary

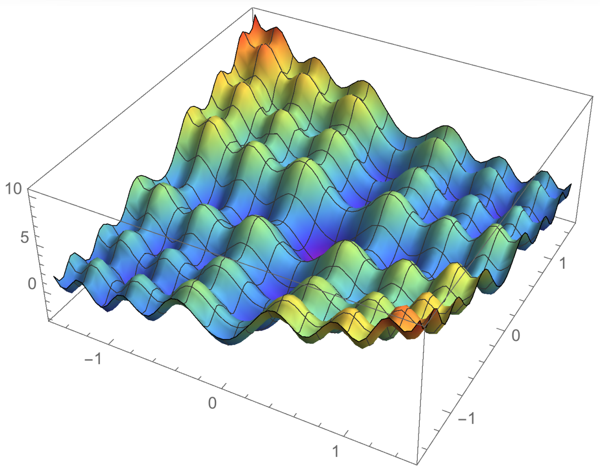
cont. submodular

$$\begin{pmatrix} * & \leq 0 & \leq 0 \\ \leq 0 & * & \leq 0 \\ \leq 0 & \leq 0 & * \end{pmatrix}$$

DR-submodular

$$\begin{pmatrix} \leq 0 & \leq 0 & \leq 0 \\ \leq 0 & \leq 0 & \leq 0 \\ \leq 0 & \leq 0 & \leq 0 \end{pmatrix}$$

non-convex

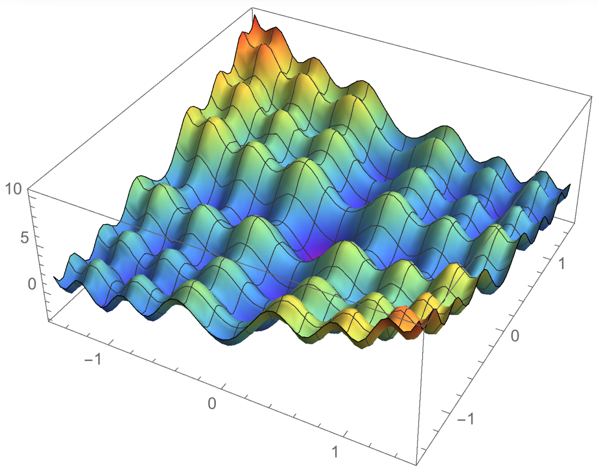


Submodularity: A General View

Summary

cont. submodular	DR-submodular	non-convex

$$\begin{pmatrix} * & \leq 0 & \leq 0 \\ \leq 0 & * & \leq 0 \\ \leq 0 & \leq 0 & * \end{pmatrix} \qquad \begin{pmatrix} \leq 0 & \leq 0 & \leq 0 \\ \leq 0 & \leq 0 & \leq 0 \\ \leq 0 & \leq 0 & \leq 0 \end{pmatrix}$$



- F is called monotone if for any $\mathbf{x} \leq \mathbf{y}$ we have $F(\mathbf{x}) \leq F(\mathbf{y})$

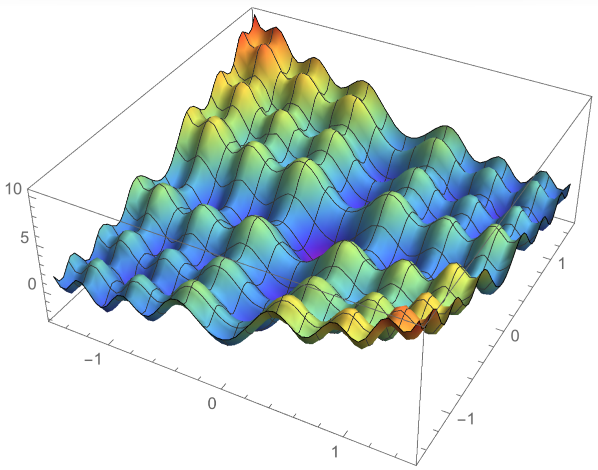
Submodularity: A General View

Summary

cont. submodular	DR-submodular	non-convex

$$\begin{pmatrix} * & \leq 0 & \leq 0 \\ \leq 0 & * & \leq 0 \\ \leq 0 & \leq 0 & * \end{pmatrix}$$

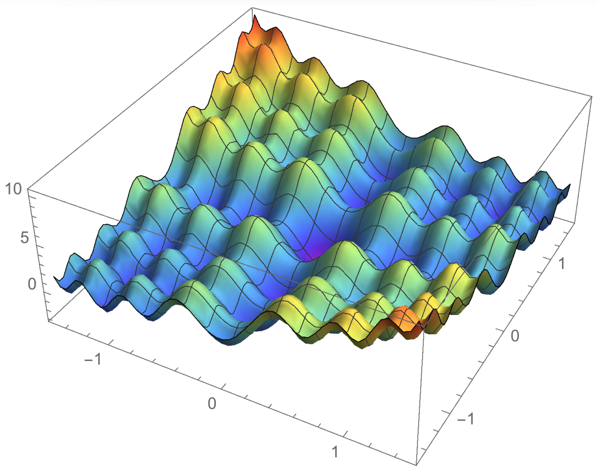
$$\begin{pmatrix} \leq 0 & \leq 0 & \leq 0 \\ \leq 0 & \leq 0 & \leq 0 \\ \leq 0 & \leq 0 & \leq 0 \end{pmatrix}$$



Examples and Applications

Submodularity: A General View

Summary

cont. submodular	DR-submodular	non-convex
$$\begin{pmatrix} * & \leq 0 & \leq 0 \\ \leq 0 & * & \leq 0 \\ \leq 0 & \leq 0 & * \end{pmatrix}$$	$$\begin{pmatrix} \leq 0 & \leq 0 & \leq 0 \\ \leq 0 & \leq 0 & \leq 0 \\ \leq 0 & \leq 0 & \leq 0 \end{pmatrix}$$	

Examples and Applications

$$F(\mathbf{x}) = \mathbf{x}^T \mathbf{A}\mathbf{x} + \mathbf{B}\mathbf{x}$$

Yale Penn

Submodularity: A General View

Summary

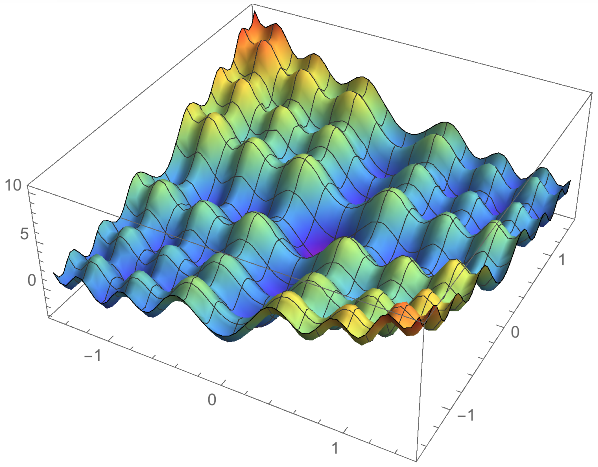
cont. submodular

$$\begin{pmatrix} * & \leq 0 & \leq 0 \\ \leq 0 & * & \leq 0 \\ \leq 0 & \leq 0 & * \end{pmatrix}$$

DR-submodular

$$\begin{pmatrix} \leq 0 & \leq 0 & \leq 0 \\ \leq 0 & \leq 0 & \leq 0 \\ \leq 0 & \leq 0 & \leq 0 \end{pmatrix}$$

non-convex



Examples and Applications

Calinescu et al., 2011

$$F(\mathbf{x}) = \mathbf{x}^T \mathbf{A}\mathbf{x} + \mathbf{B}\mathbf{x}$$

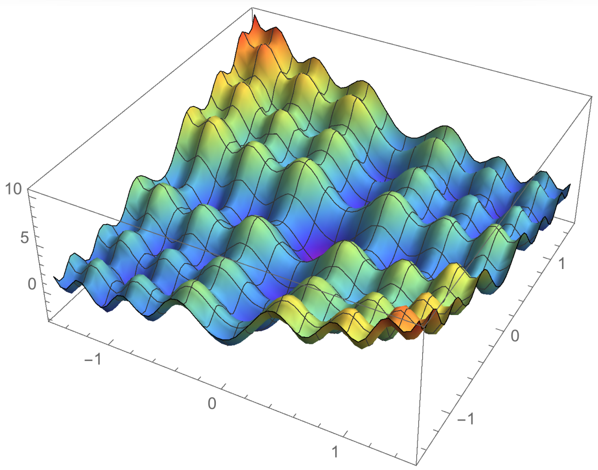
Multi-linear extension of submodular set functions

Yale

Penn

Submodularity: A General View

Summary

cont. submodular	DR-submodular	non-convex
$\begin{pmatrix} {\color{red}*} & \leq 0 & \leq 0 \\ \leq 0 & {\color{red}*} & \leq 0 \\ \leq 0 & \leq 0 & {\color{red}*} \end{pmatrix}$	$\begin{pmatrix} {\color{red}\leq 0} & \leq 0 & \leq 0 \\ \leq 0 & {\color{red}\leq 0} & \leq 0 \\ \leq 0 & \leq 0 & {\color{red}\leq 0} \end{pmatrix}$	

Examples and Applications

$F(\mathbf{x}) = \mathbf{x}^T \mathbf{A} \mathbf{x} + \mathbf{B} \mathbf{x}$

Calinescu et al., 2011

Multi-linear extension of submodular set functions

Bian et al, 2020

Influence maximization

Kulesza, Taskar, 2012

MAP Inference of DPPs

Hartline et al, 2014 Soma et al, 2017

Revenue Maximization

Soma et al, '14 Hotano et al, '15 Stalib, Jegelka, '17 Eghbali et al., '16

Resource Allocation

Bian et al, 2020

coverage, summarization, etc

Bian et al, 2018

Mean-Field Inference

Yale Penn

Submodular Maximization

- We consider the following optimization problem:

$$\text{maximize } F(\mathbf{x})$$
$$\text{subject to: } \mathbf{x} \in \mathcal{K}$$

Submodular Maximization

- We consider the following optimization problem:

$$\text{maximize } F(\mathbf{x})$$
$$\text{subject to: } \mathbf{x} \in \mathcal{K}$$

Submodular Maximization

- We consider the following optimization problem:

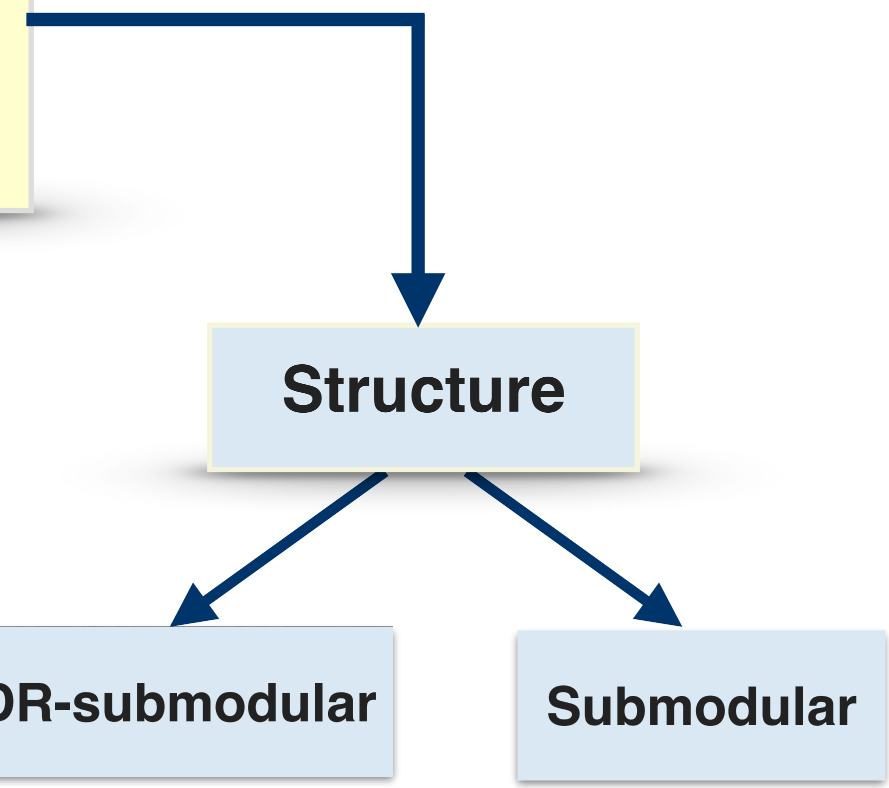
$$\begin{aligned}&\text{maximize } F(\mathbf{x})\\ &\text{subject to: } \mathbf{x} \in \mathcal{K}\end{aligned}$$

Structure

Submodular Maximization

- We consider the following optimization problem:

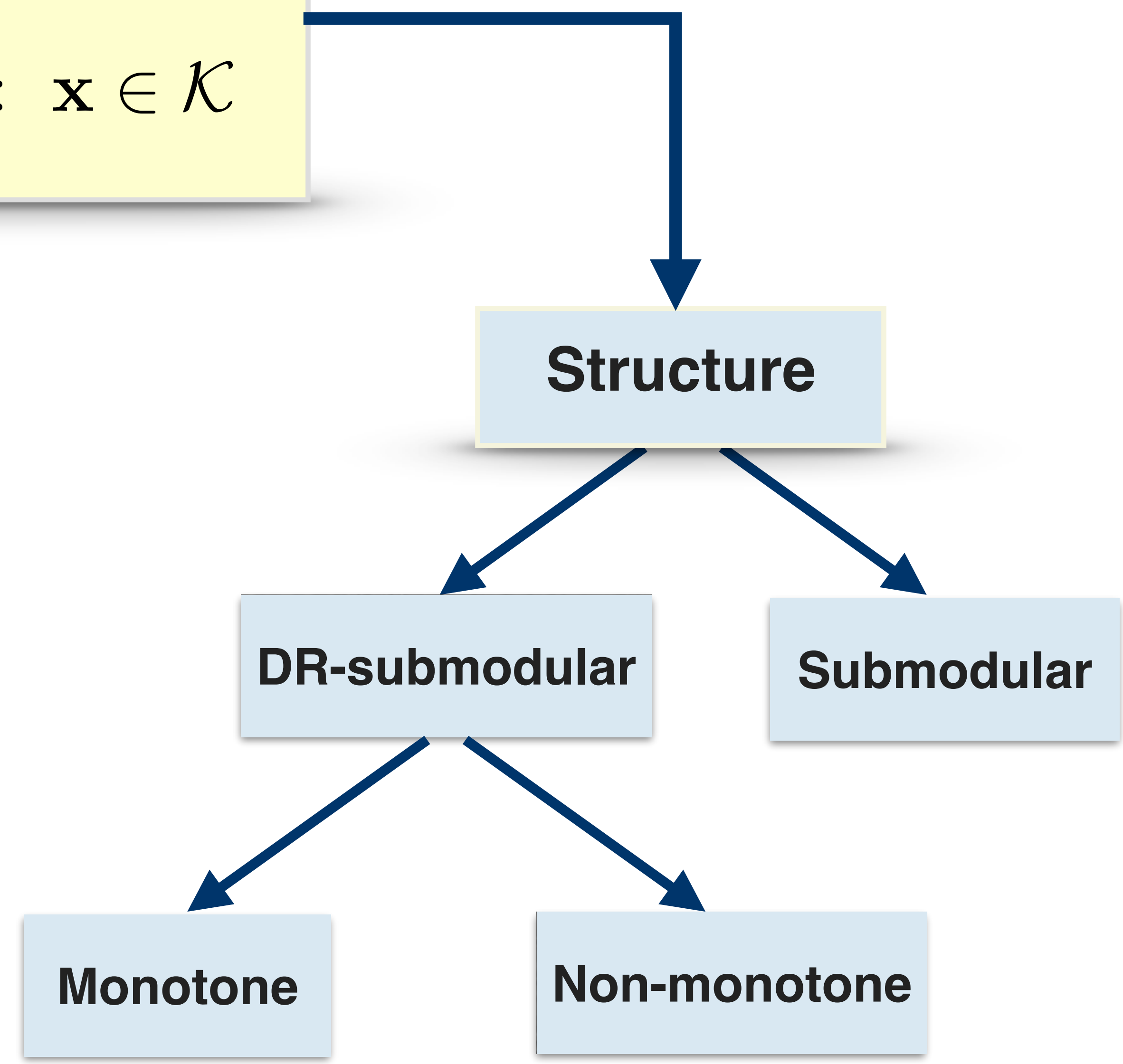
$$\begin{aligned} &\text{maximize } F(\mathbf{x}) \\ &\text{subject to: } \mathbf{x} \in \mathcal{K} \end{aligned}$$



Submodular Maximization

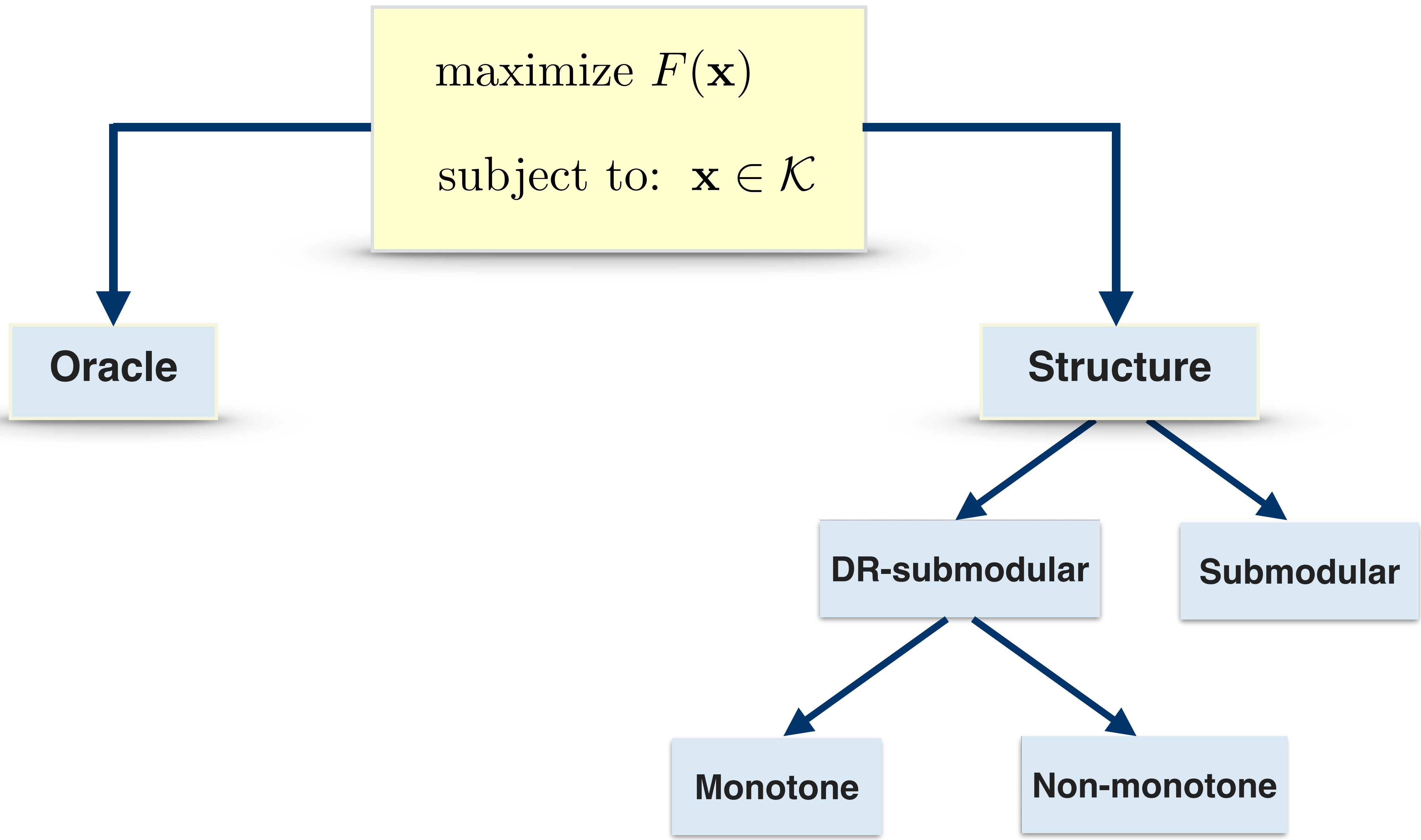
- We consider the following optimization problem:

$$\text{maximize } F(\mathbf{x})$$
$$\text{subject to: } \mathbf{x} \in \mathcal{K}$$



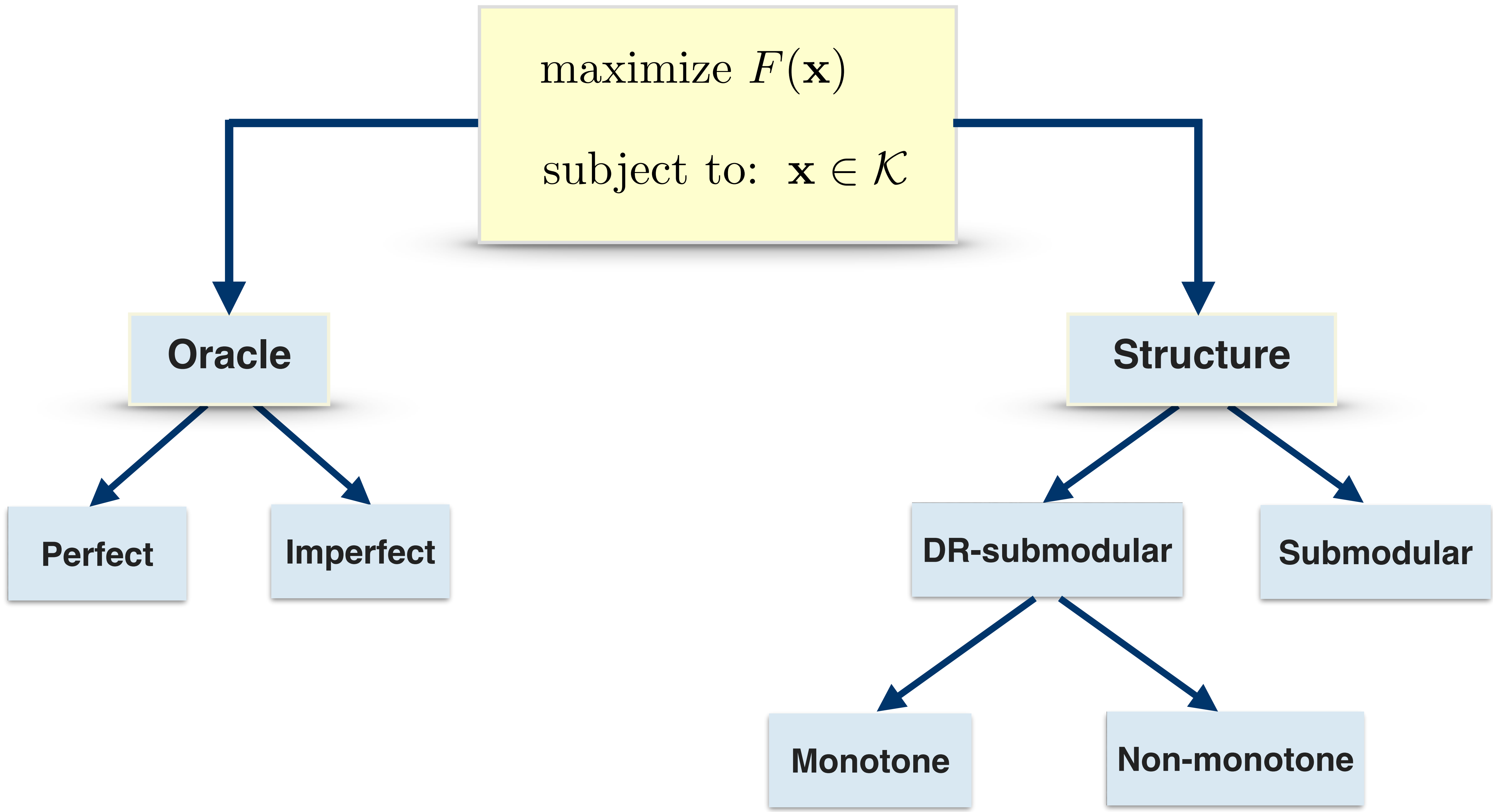
Submodular Maximization

- We consider the following optimization problem:



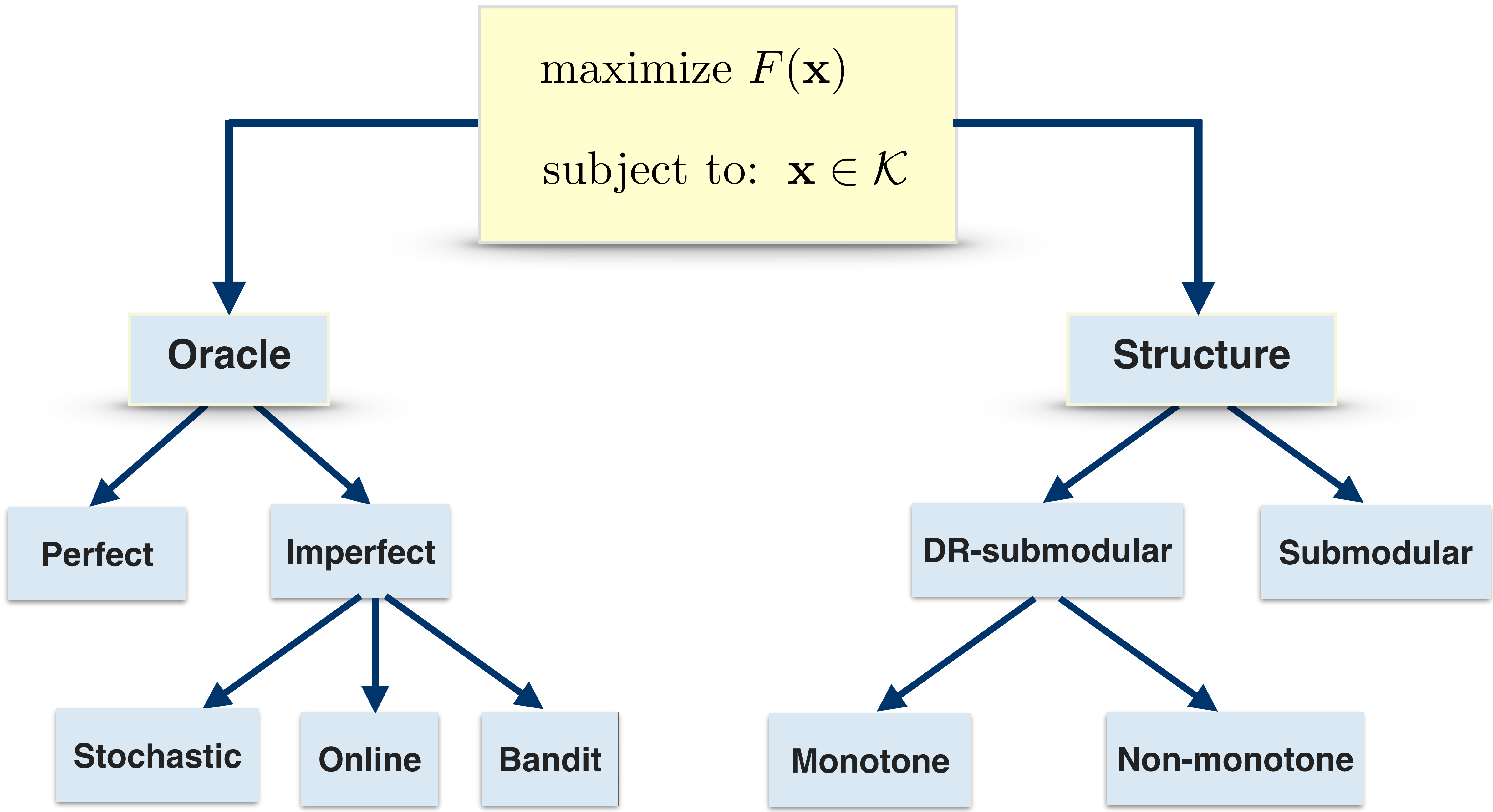
Submodular Maximization

- We consider the following optimization problem:



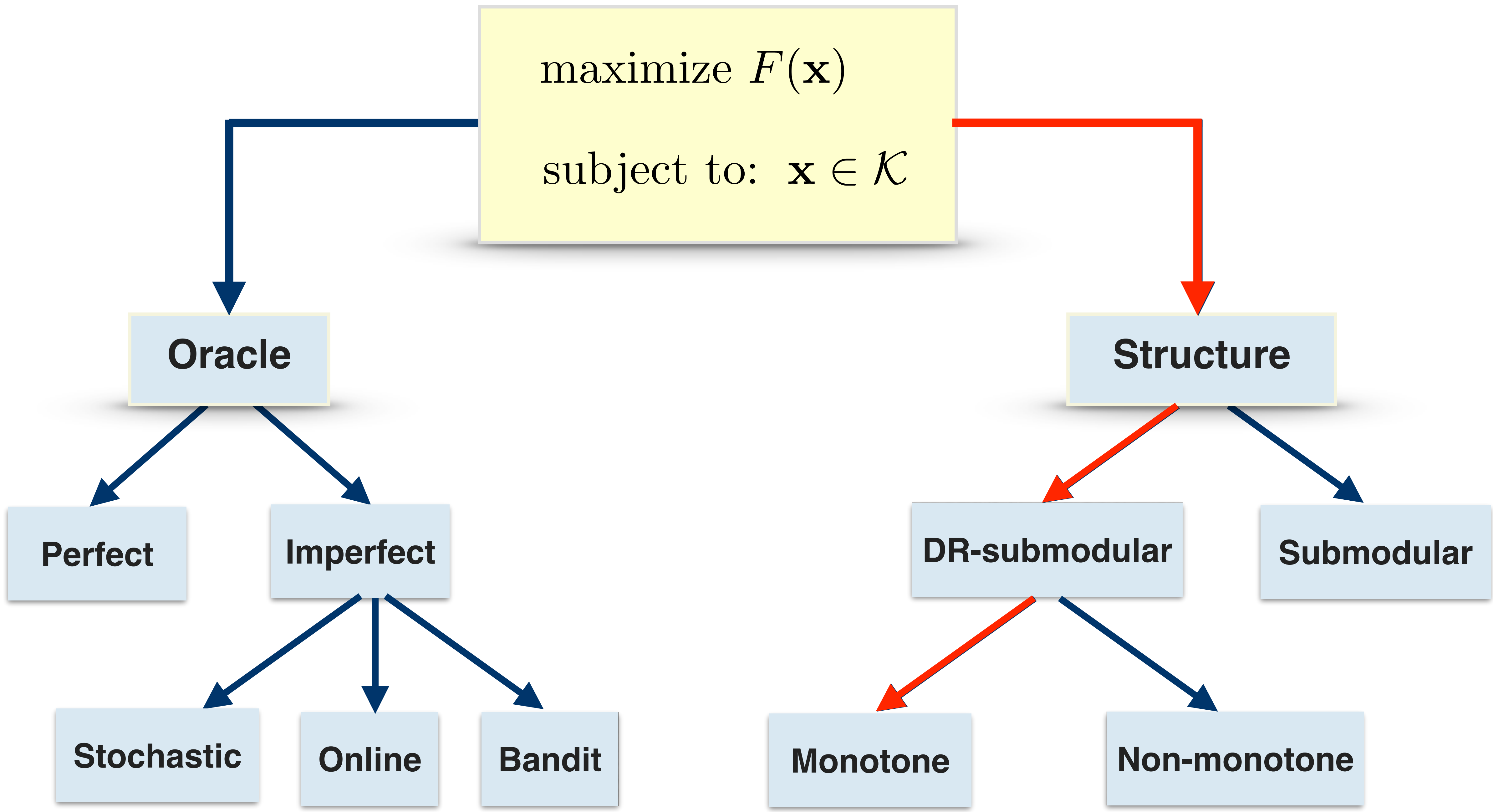
Submodular Maximization

- We consider the following optimization problem:



Submodular Maximization

- We consider the following optimization problem:

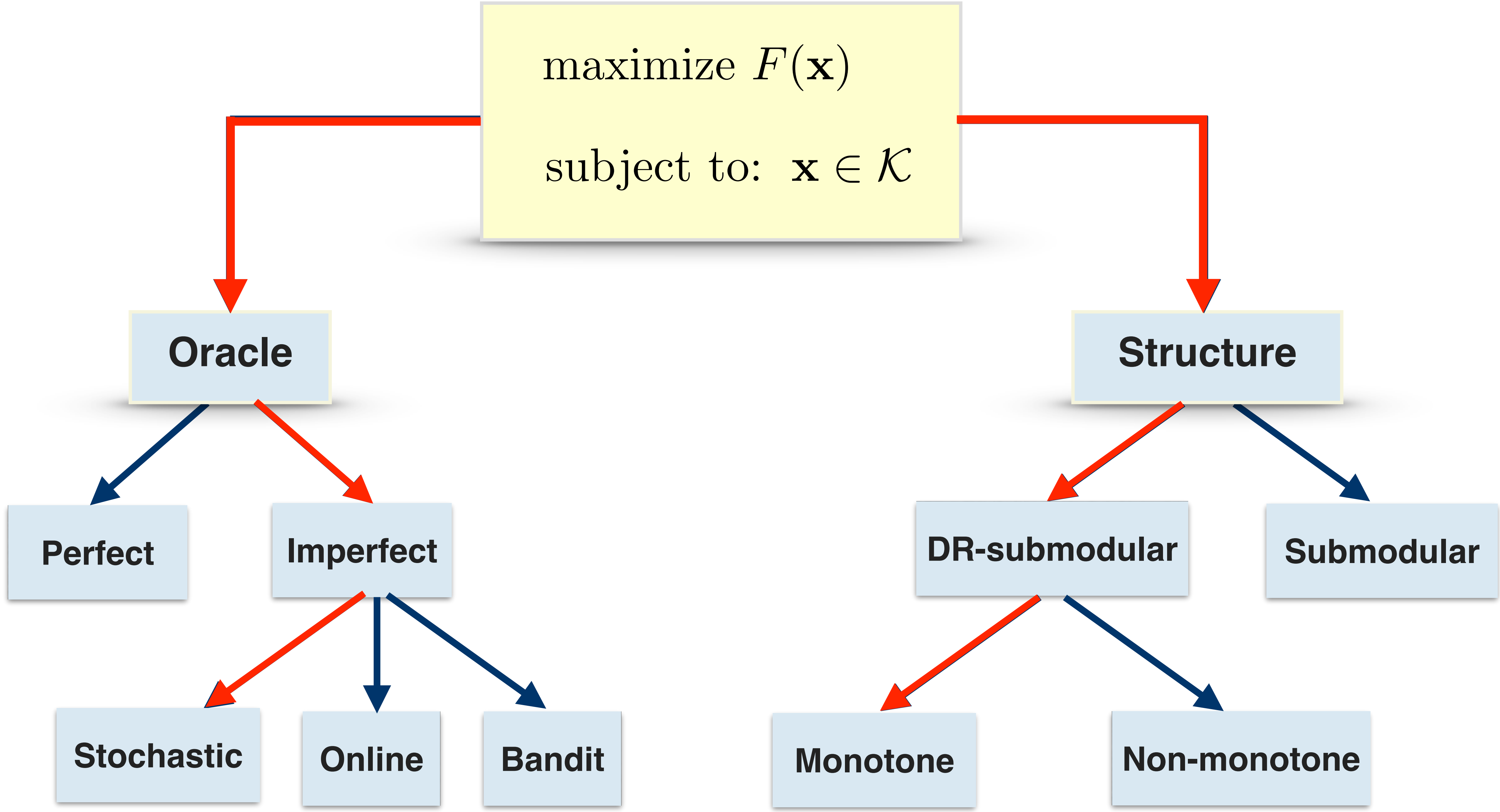


Submodular Maximization

- We consider the following optimization problem:

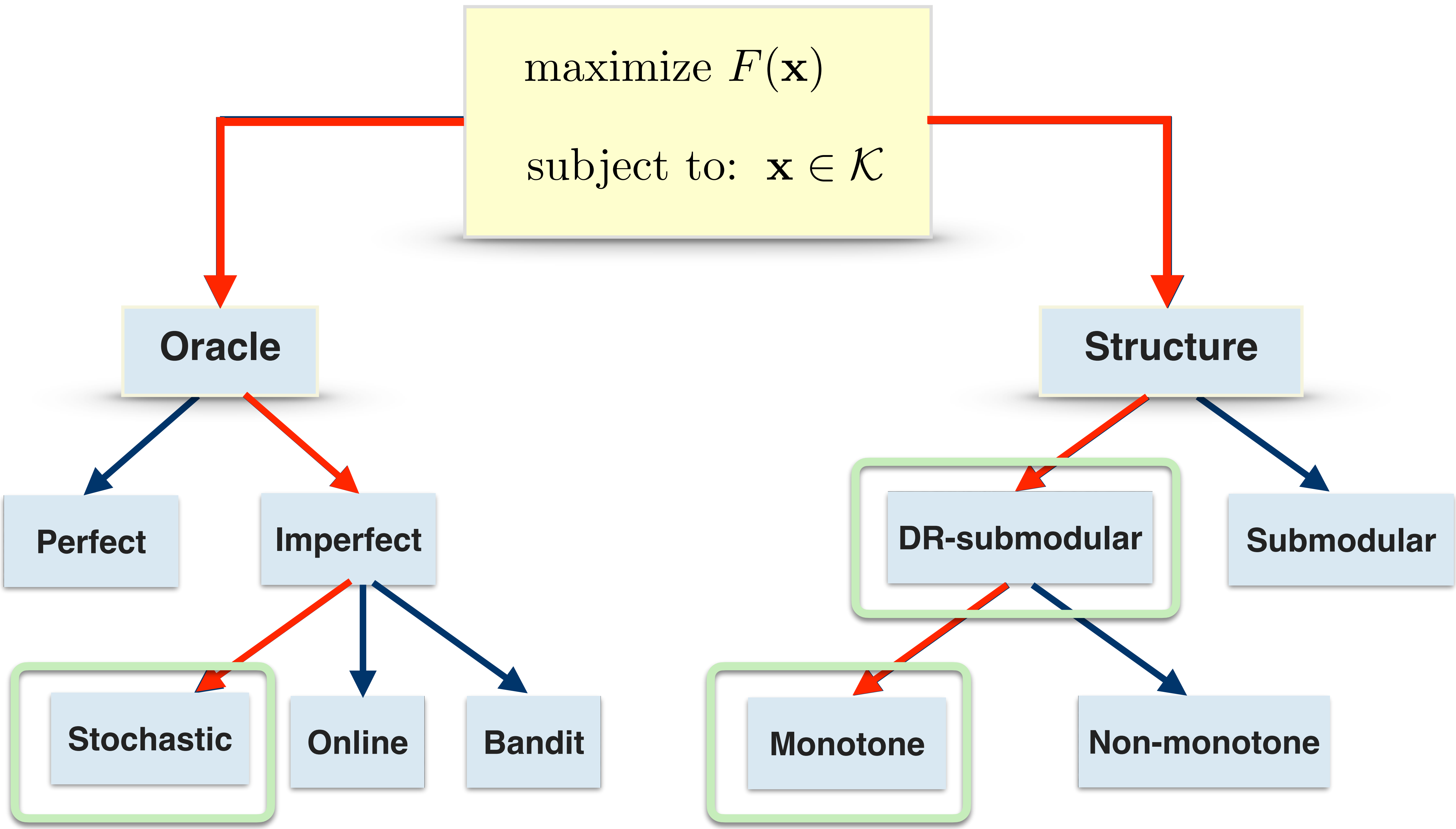
$$\text{maximize } F(\mathbf{x})$$

$$\text{subject to: } \mathbf{x} \in \mathcal{K}$$



Submodular Maximization

- We consider the following optimization problem:



(Stochastic) Submodular Maximization

$$\max_{\mathbf{x} \in \mathcal{K}} F(\mathbf{x})$$

(Stochastic) Submodular Maximization

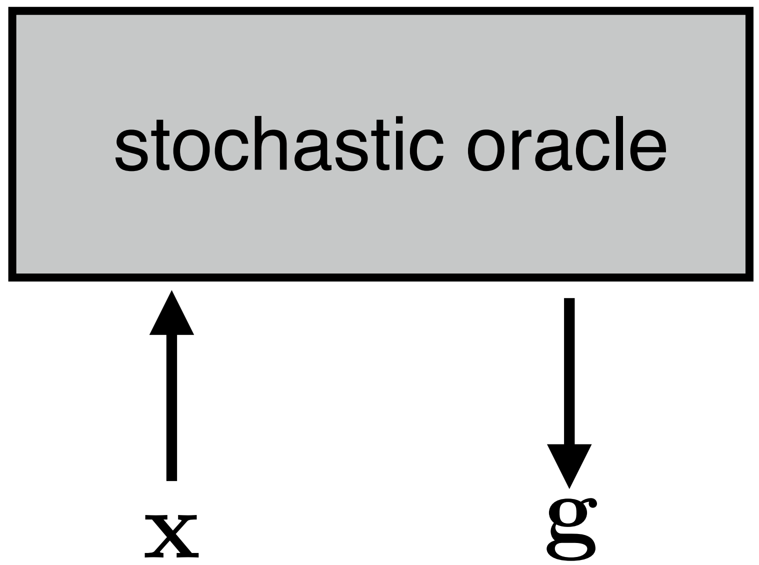
$$\max_{\mathbf{x}\in\mathcal{K}} F(\mathbf{x})$$

Yale

Penn

(Stochastic) Submodular Maximization

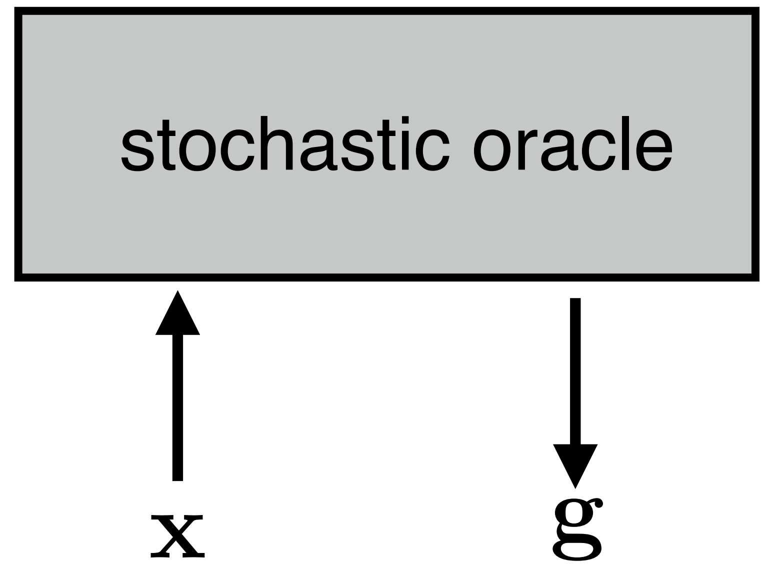
$$\max_{\mathbf{x} \in \mathcal{K}} F(\mathbf{x})$$



“Gradient Methods for Submodular Maximization”, Hassani, Soltanolkotabi, Karbasi, 2017

(Stochastic) Submodular Maximization

$$\max_{\mathbf{x} \in \mathcal{K}} F(\mathbf{x})$$

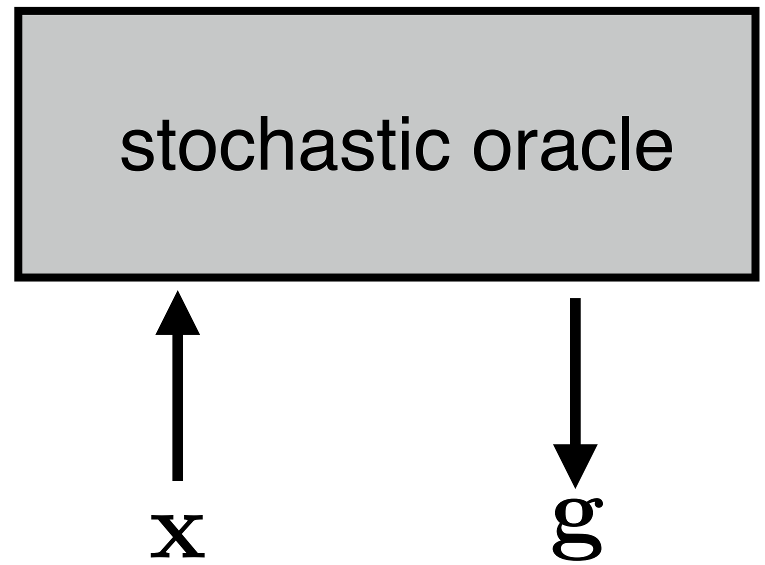


$$\mathbb{E}[\mathbf{g} \mid \mathbf{x}] = \nabla F(\mathbf{x})$$

"Gradient Methods for Submodular Maximization", Hassani, Soltanolkotabi, Karbasi, 2017

(Stochastic) Submodular Maximization

$$\max_{\mathbf{x}\in\mathcal{K}} F(\mathbf{x})$$



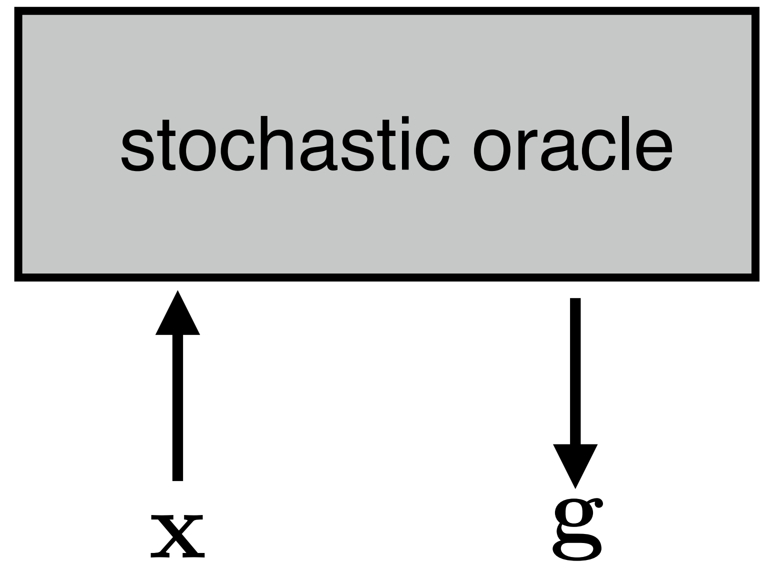
$\mathbf{x}$

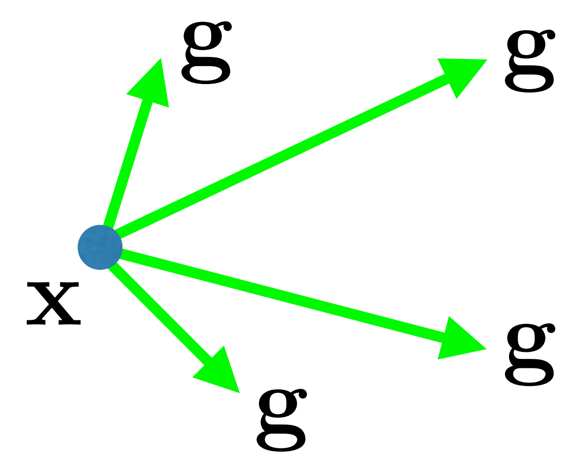
$$\mathbb{E}[\mathbf{g} \mid \mathbf{x}] = \nabla F(\mathbf{x})$$

"Gradient Methods for Submodular Maximization", Hassani, Soltanolkotabi, Karbasi, 2017

(Stochastic) Submodular Maximization

$$\max_{\mathbf{x} \in \mathcal{K}} F(\mathbf{x})$$



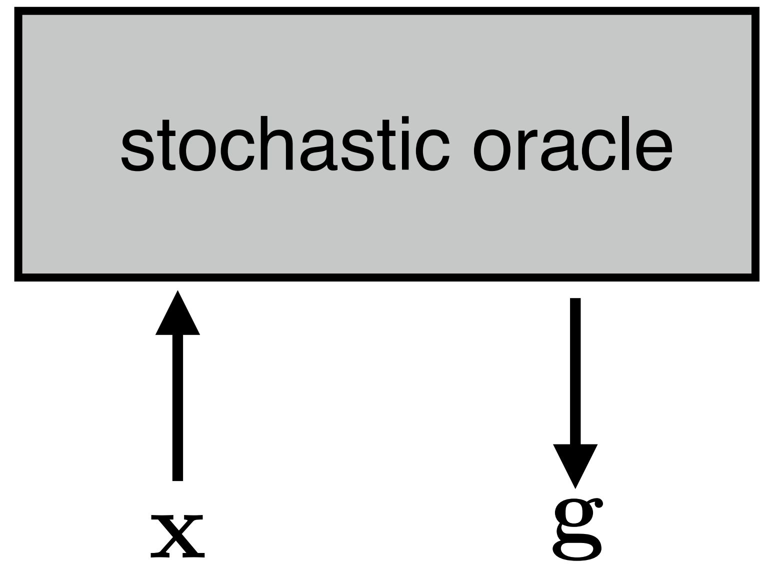


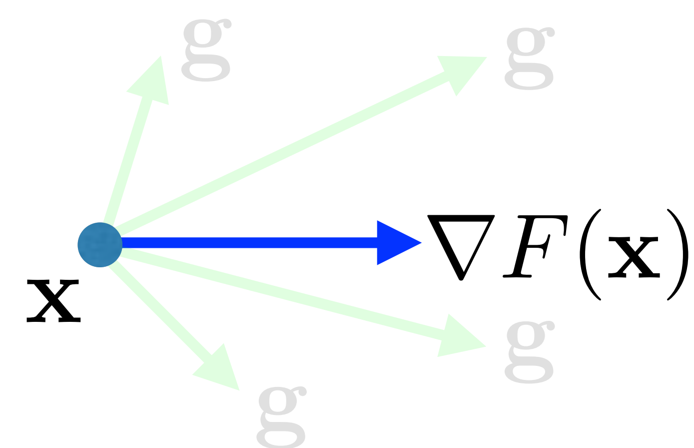
$$\mathbb{E}[\mathbf{g} \mid \mathbf{x}] = \nabla F(\mathbf{x})$$

"Gradient Methods for Submodular Maximization", Hassani, Soltanolkotabi, Karbasi, 2017

(Stochastic) Submodular Maximization

$$\max_{\mathbf{x} \in \mathcal{K}} F(\mathbf{x})$$



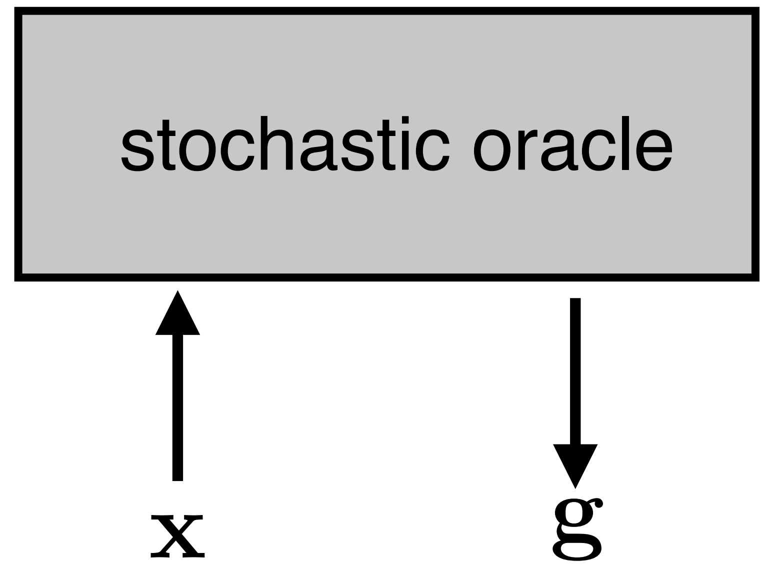


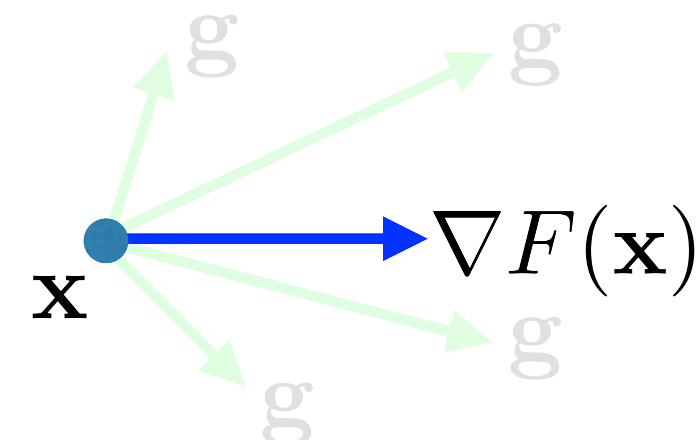
$$\mathbb{E}[\mathbf{g} \mid \mathbf{x}] = \nabla F(\mathbf{x})$$

"Gradient Methods for Submodular Maximization", Hassani, Soltanolkotabi, Karbasi, 2017

(Stochastic) Submodular Maximization

$$\max_{\mathbf{x}\in\mathcal{K}} F(\mathbf{x})$$



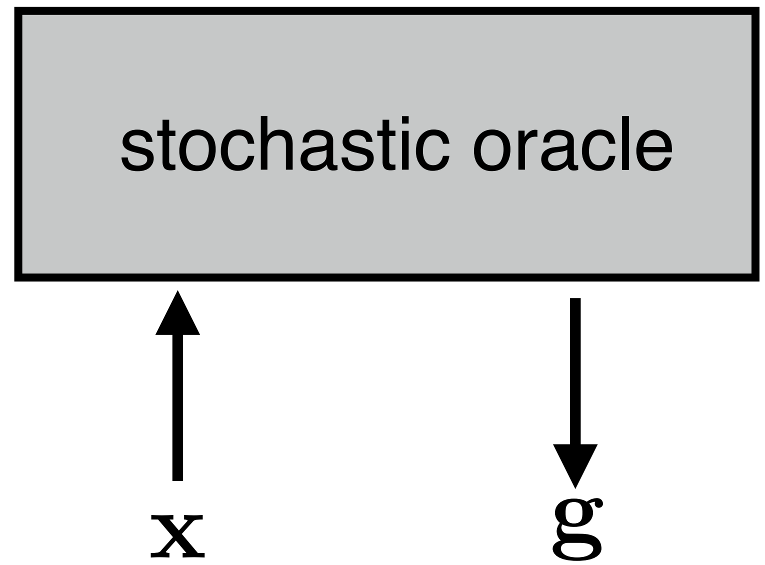


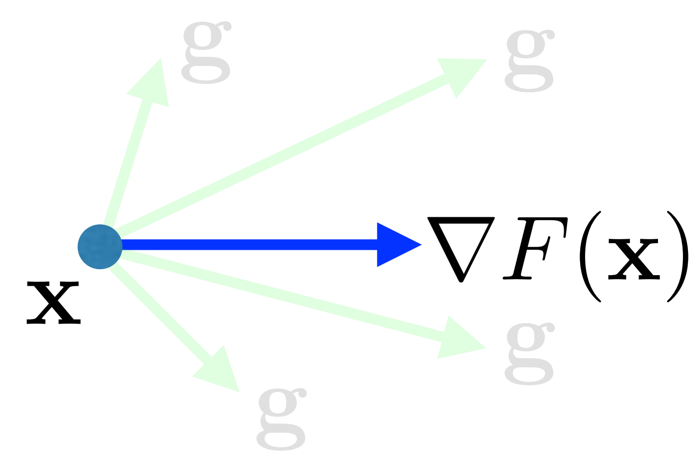
$$\mathbb{E}[\mathbf{g} \mid \mathbf{x}] = \nabla F(\mathbf{x})$$

- F is monotone and DR-submodular
- $\mathcal{K}$ is a convex and bounded set

(Stochastic) Submodular Maximization

$$\max_{\mathbf{x}\in\mathcal{K}} F(\mathbf{x})$$





$$\mathbb{E}[\mathbf{g} \mid \mathbf{x}] = \nabla F(\mathbf{x})$$

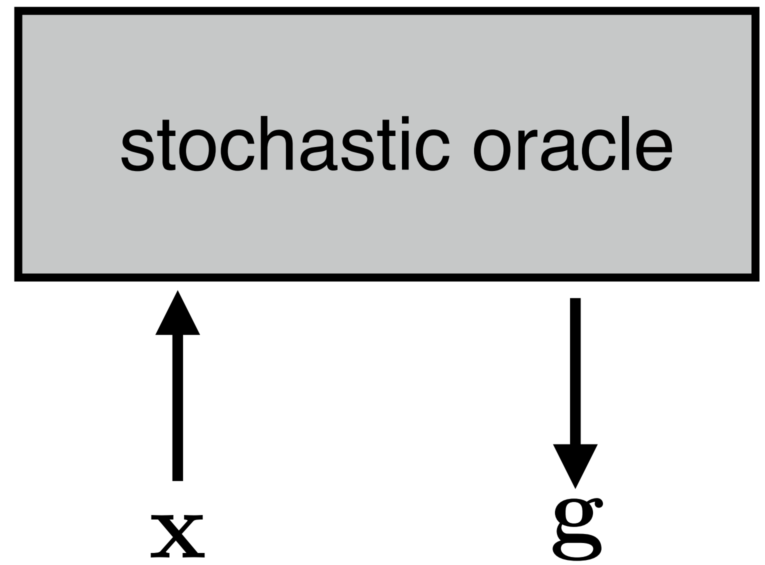
- F is monotone and DR-submodular
- $\mathcal{K}$ is a convex and bounded set

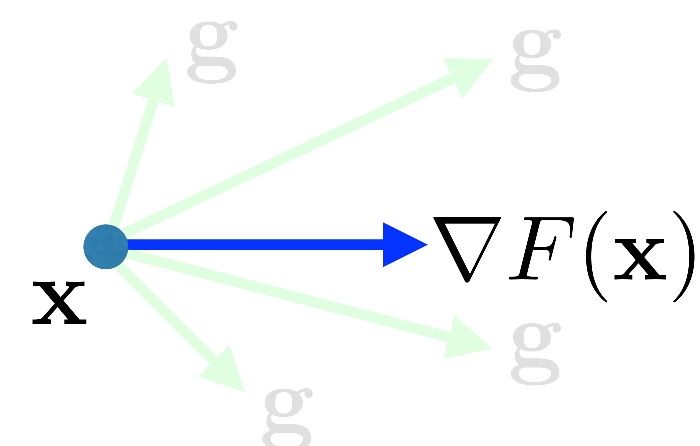
Devise efficient algorithms to solve the problem above.

Yale Penn

(Stochastic) Submodular Maximization

$$\max_{\mathbf{x} \in \mathcal{K}} F(\mathbf{x})$$





$$\mathbb{E}[\mathbf{g} \mid \mathbf{x}] = \nabla F(\mathbf{x})$$

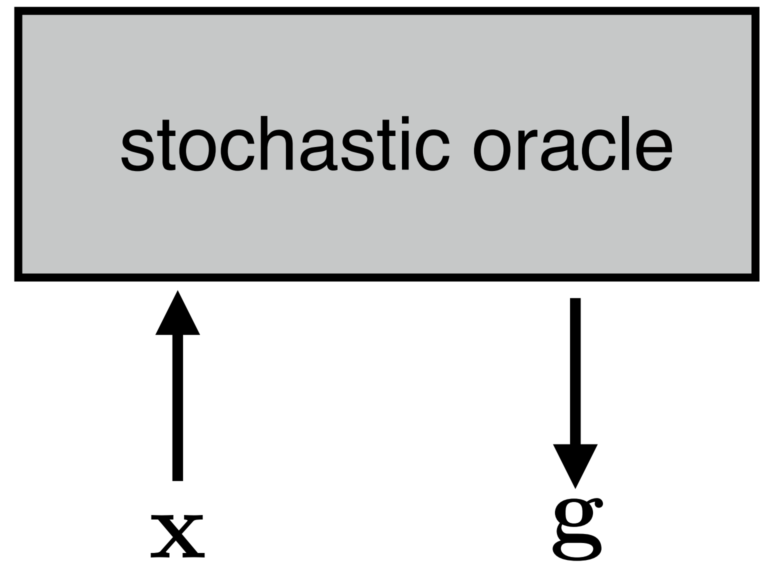
- F is monotone and DR-submodular
- $\mathcal{K}$ is a convex and bounded set

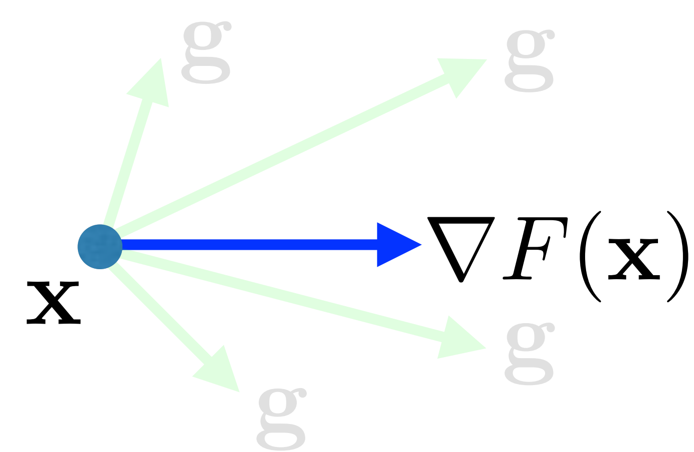
Devise efficient algorithms to solve the problem above.

Yale Penn

(Stochastic) Submodular Maximization

$$\max_{\mathbf{x} \in \mathcal{K}} F(\mathbf{x})$$





$$\mathbb{E}[\mathbf{g} \mid \mathbf{x}] = \nabla F(\mathbf{x})$$

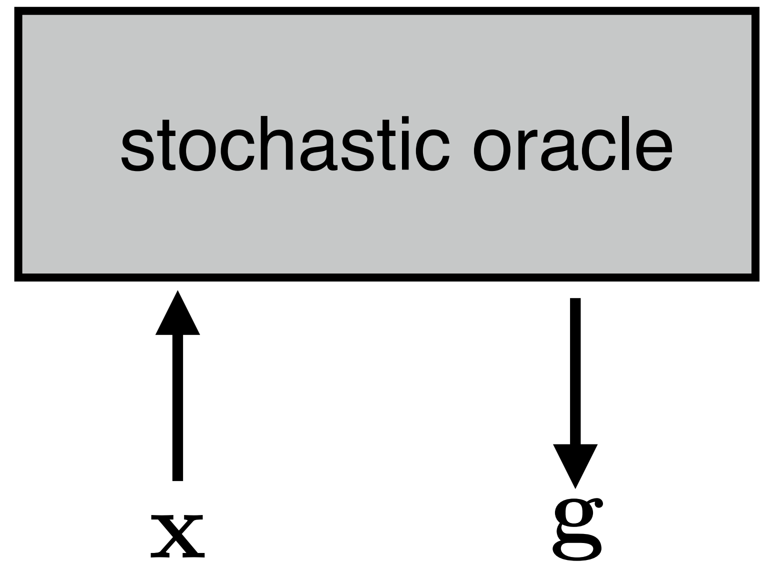
- F is monotone and DR-submodular
- $\mathcal{K}$ is a convex and bounded set

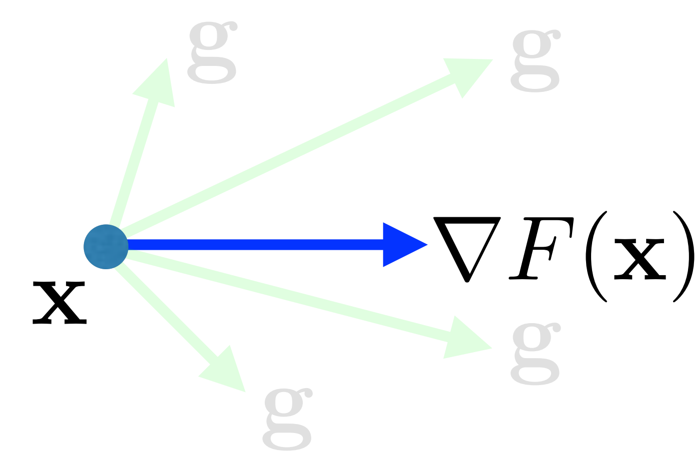
Devise <u>efficient</u> algorithms to <u>solve</u> the problem above.

- *computational complexity?*
- *sample complexity?*

(Stochastic) Submodular Maximization

$$\max_{\mathbf{x} \in \mathcal{K}} F(\mathbf{x})$$





$$\mathbb{E}[\mathbf{g} \mid \mathbf{x}] = \nabla F(\mathbf{x})$$

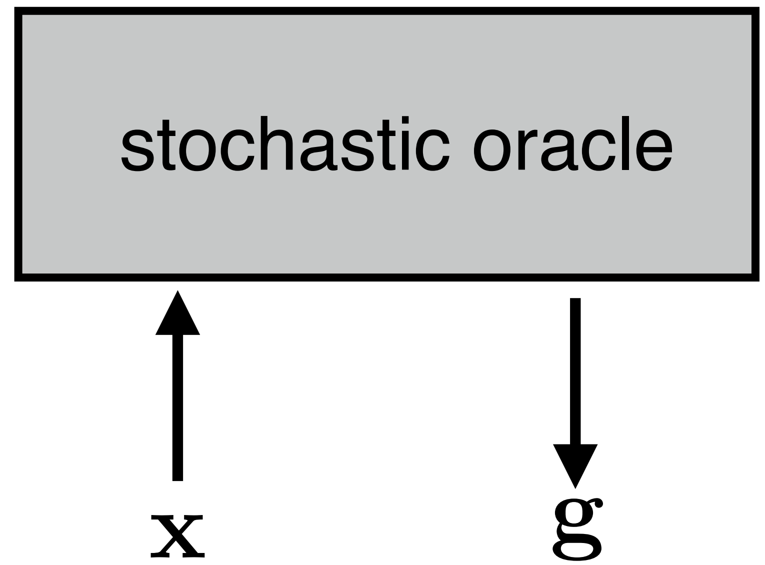
- F is monotone and DR-submodular
- $\mathcal{K}$ is a convex and bounded set

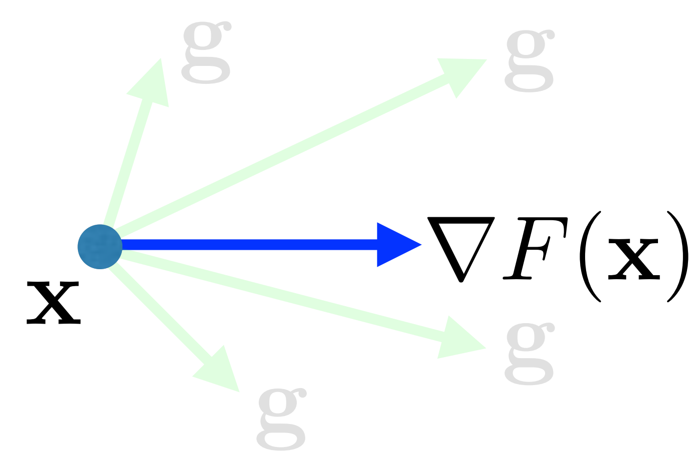
Devise efficient algorithms to solve the problem above.

- *computational complexity?* - *quality of the solution?*
- *sample complexity?*

(Stochastic) Submodular Maximization

$$\mathrm{OPT} \triangleq \max_{\mathbf{x}\in\mathcal{K}} F(\mathbf{x})$$





$$\mathbb{E}[\mathbf{g} \mid \mathbf{x}] = \nabla F(\mathbf{x})$$

- F is monotone and DR-submodular
- $\mathcal{K}$ is a convex and bounded set

Devise efficient algorithms to solve the problem above.

- computational complexity? *- quality of the solution?*

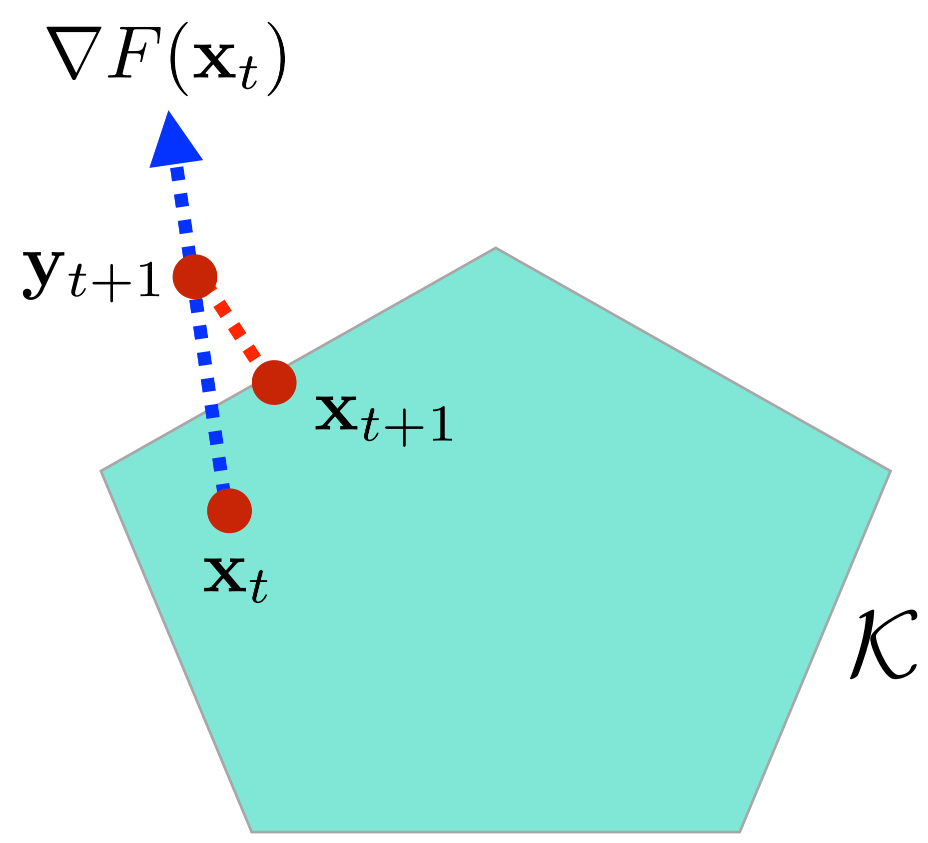
- sample complexity?

Projected Gradient Methods

- Repeat for T iterations:

$$\mathbf{y}_{t+1} = \mathbf{x}_t + \eta_t \nabla F(\mathbf{x}_t)$$

$$\mathbf{x}_{t+1} = \mathrm{Proj}_{\mathcal{K}}(\mathbf{y}_{t+1})$$



"Gradient Methods for Submodular Maximization", Hassani, Soltanolkotabi, Karbasi, 2017

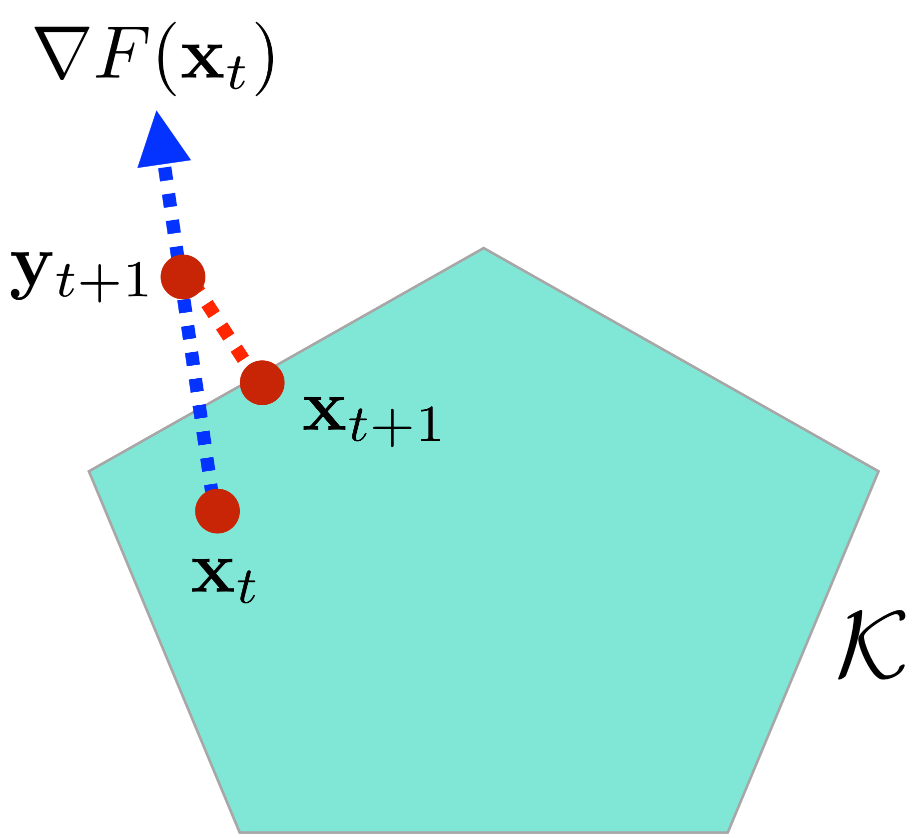
Yale Penn

Projected Gradient Methods

- Repeat for T iterations:

$$\mathbf{y}_{t+1} = \mathbf{x}_t + \eta_t \nabla F(\mathbf{x}_t)$$

$$\mathbf{x}_{t+1} = \mathrm{Proj}_{\mathcal{K}}(\mathbf{y}_{t+1})$$



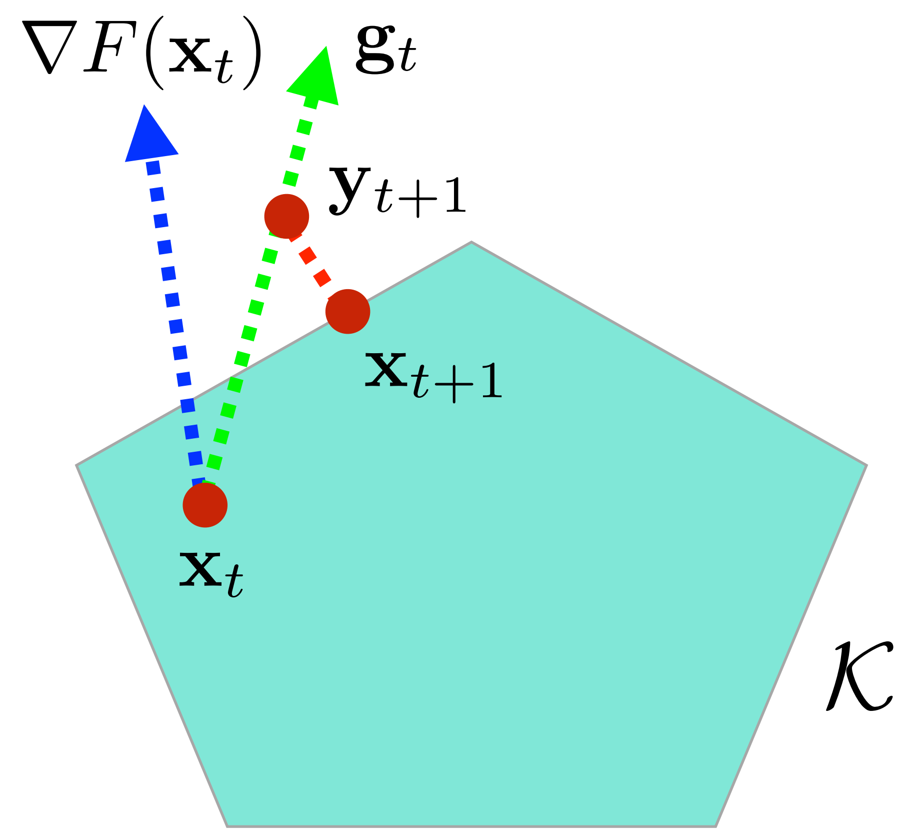
"Gradient Methods for Submodular Maximization", Hassani, Soltanolkotabi, Karbasi, 2017

Projected Gradient Methods

- Repeat for T iterations:

$$\mathbf{y}_{t+1} = \mathbf{x}_t + \eta_t \mathbf{g}_t$$

$$\mathbf{x}_{t+1} = \mathrm{Proj}_{\mathcal{K}}(\mathbf{y}_{t+1})$$



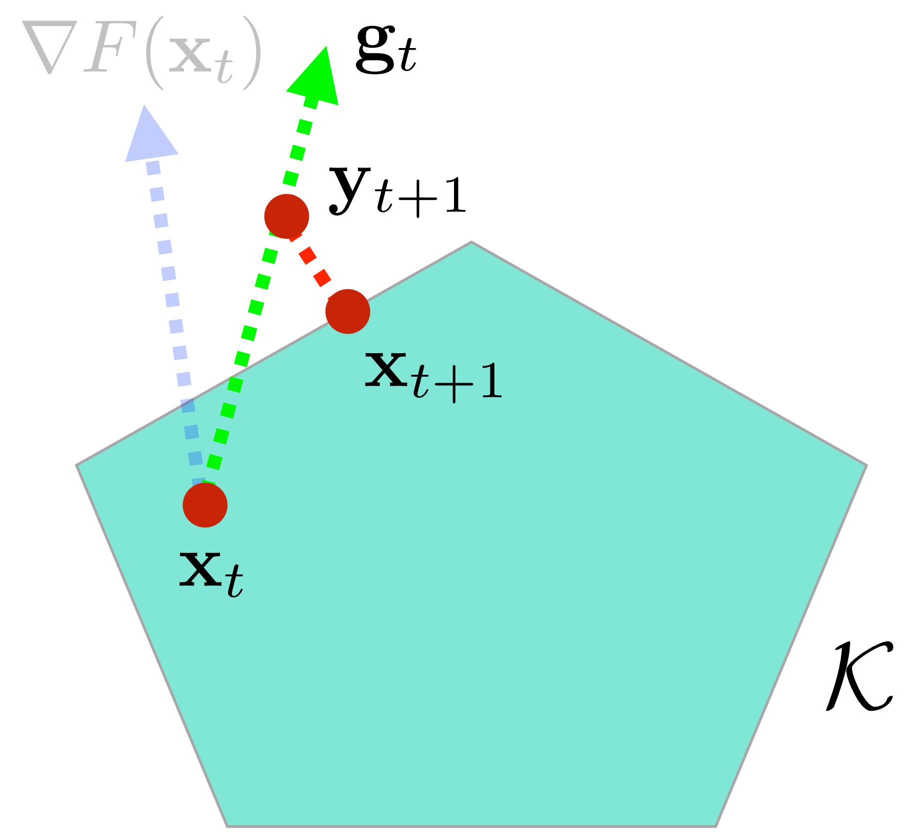
“Gradient Methods for Submodular Maximization”, Hassani, Soltanolkotabi, Karbasi, 2017

Projected Gradient Methods

- Repeat for T iterations:

$$\mathbf{y}_{t+1} = \mathbf{x}_t + \eta_t \mathbf{g}_t$$

$$\mathbf{x}_{t+1} = \mathrm{Proj}_{\mathcal{K}}(\mathbf{y}_{t+1})$$



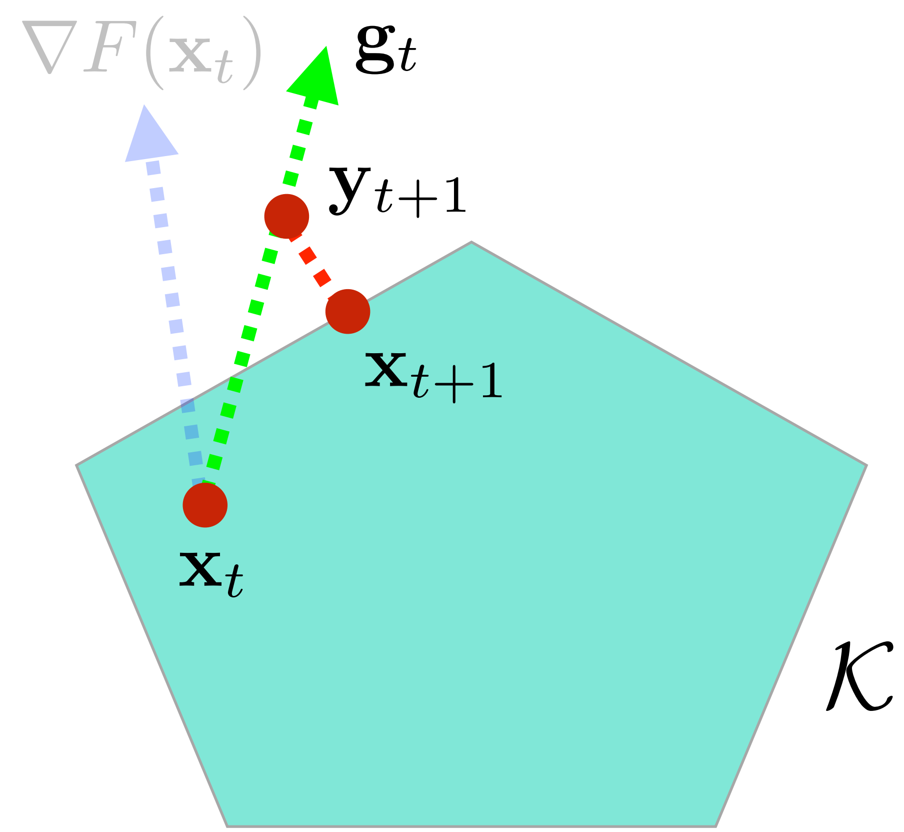
"Gradient Methods for Submodular Maximization", Hassani, Soltanolkotabi, Karbasi, 2017

Projected Gradient Methods

- Repeat for T iterations:

$$\mathbf{y}_{t+1} = \mathbf{x}_t + \eta_t \mathbf{g}_t$$

$$\mathbf{x}_{t+1} = \mathrm{Proj}_{\mathcal{K}}(\mathbf{y}_{t+1})$$



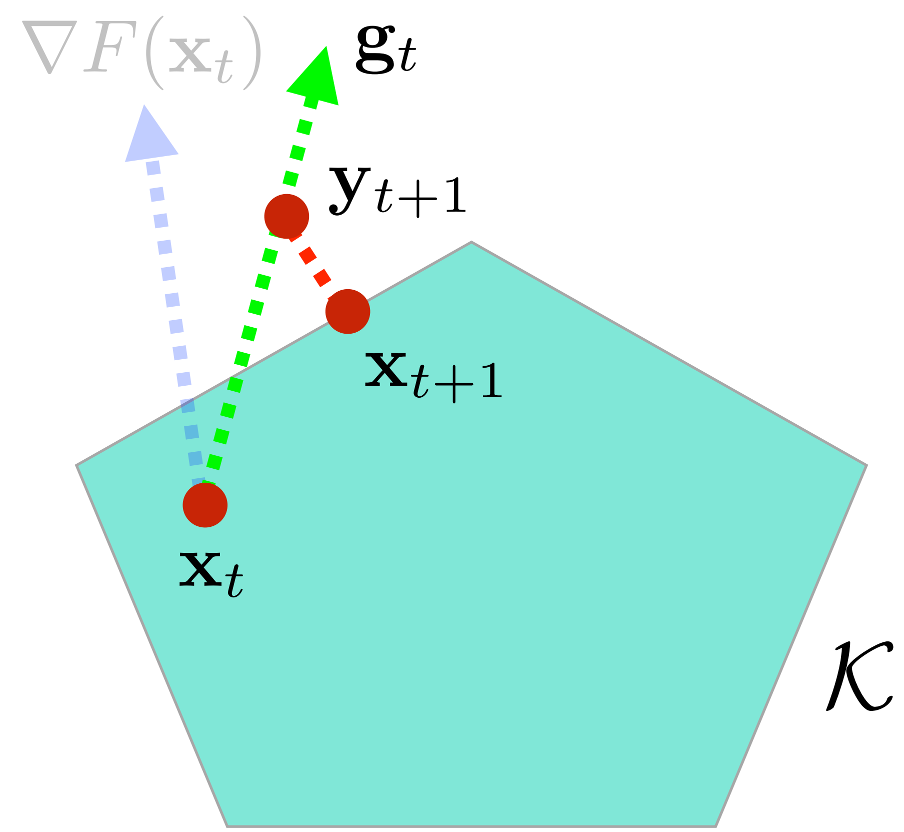
Do projected gradient methods lead to provably good solutions for continuous submodular maximization with general convex constraints?

Projected Gradient Methods

- Repeat for T iterations:

$$\mathbf{y}_{t+1} = \mathbf{x}_t + \eta_t \mathbf{g}_t$$

$$\mathbf{x}_{t+1} = \mathrm{Proj}_{\mathcal{K}}(\mathbf{y}_{t+1})$$



Do projected gradient methods lead to provably good solutions for continuous submodular maximization with general convex constraints?

- Does it converge?

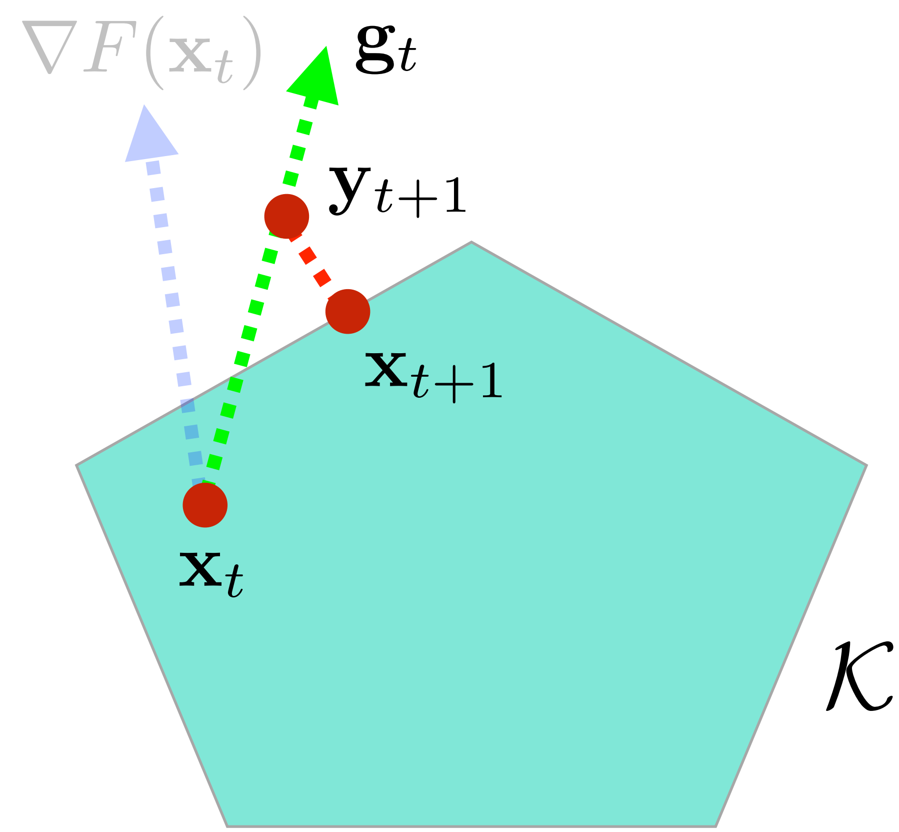
- Quality of the final solution?

Projected Gradient Methods

- Repeat for T iterations:

$$\mathbf{y}_{t+1} = \mathbf{x}_t + \eta_t \mathbf{g}_t$$

$$\mathbf{x}_{t+1} = \mathrm{Proj}_{\mathcal{K}}(\mathbf{y}_{t+1})$$



Do projected gradient methods lead to provably good solutions for continuous submodular maximization with general convex constraints?

- Does it converge?
- Quality of the final solution?

- Broader context of non-convex opt:
- Convergence to stationary points
- Structure matters

Stationary Points and Their Quality

- Let's first characterize the fixed point of gradient ascent

Stationary Points and Their Quality

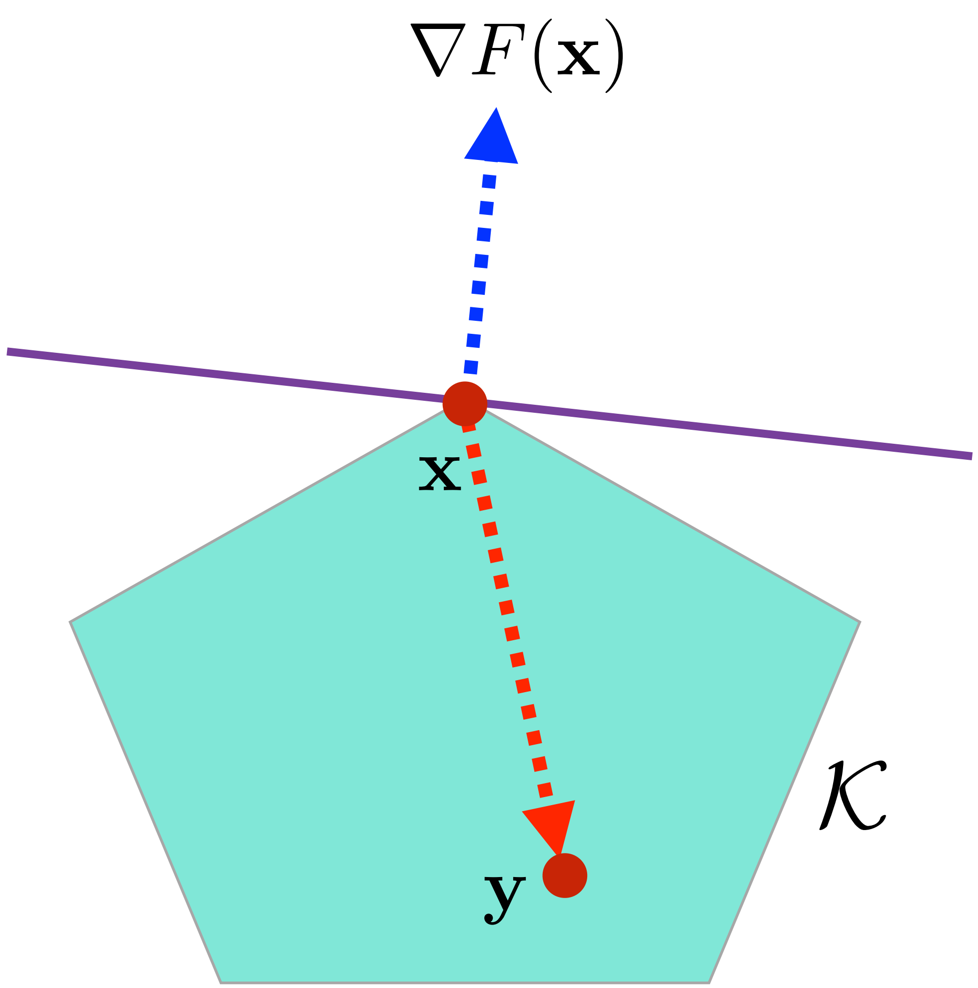
- Let's first characterize the fixed point of gradient ascent
- A vector $\mathbf{x}$ is a **stationary point** of a function F over the set $\mathcal{K}$ if

$$\max_{\mathbf{y}\in\mathcal{K}} \langle \nabla F(\mathbf{x}), \mathbf{y}-\mathbf{x}\rangle \ \leq 0$$

Stationary Points and Their Quality

- Let's first characterize the fixed point of gradient ascent
- A vector $\mathbf{x}$ is a **stationary point** of a function F over the set $\mathcal{K}$ if

$$\max_{\mathbf{y}\in\mathcal{K}} \langle \nabla F(\mathbf{x}), \mathbf{y} - \mathbf{x} \rangle \leq 0$$

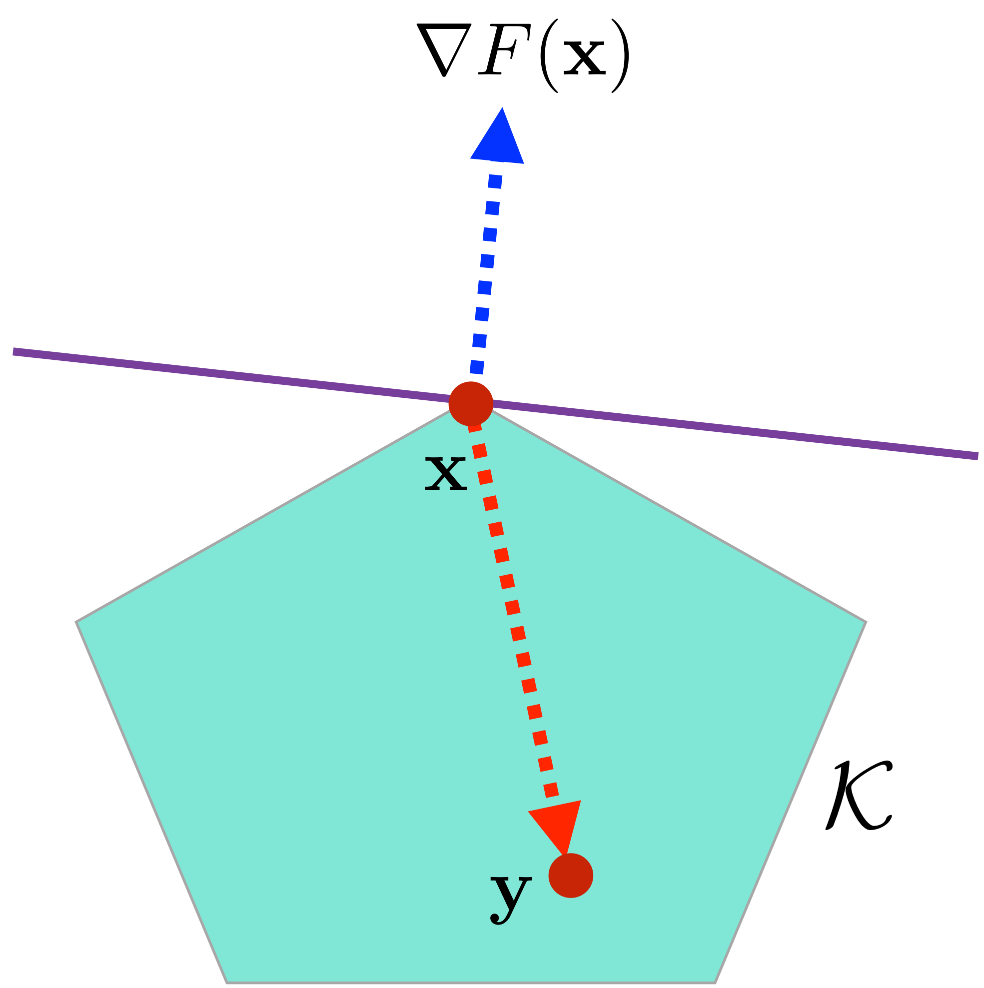


Yale Penn

Stationary Points and Their Quality

- Let's first characterize the fixed point of gradient ascent
- A vector $\mathbf{x}$ is a **stationary point** of a function F over the set $\mathcal{K}$ if

$$\max_{\mathbf{y} \in \mathcal{K}} \langle \nabla F(\mathbf{x}), \mathbf{y} - \mathbf{x} \rangle \leq 0$$

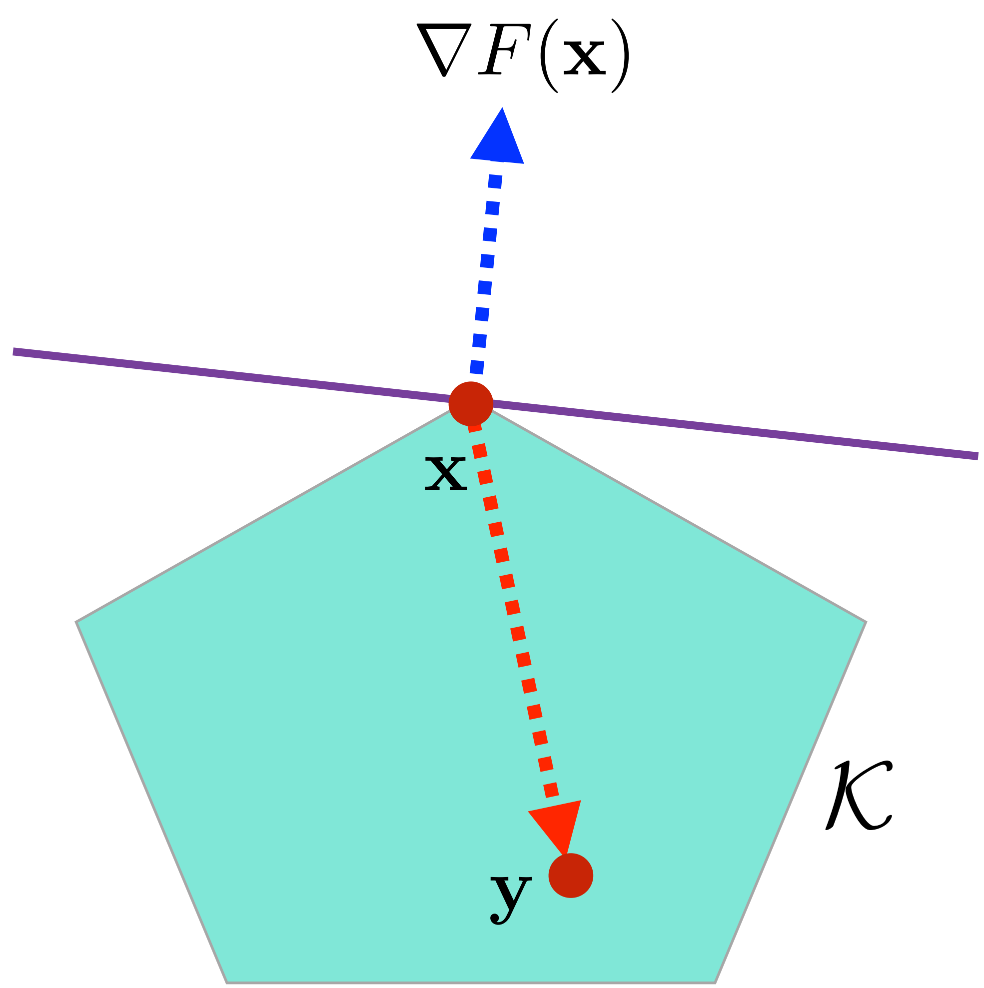


Intuition (first-order):

Stationary Points and Their Quality

- Let's first characterize the fixed point of gradient ascent
- A vector $\mathbf{x}$ is a **stationary point** of a function F over the set $\mathcal{K}$ if

$$\max_{\mathbf{y}\in\mathcal{K}} \langle \nabla F(\mathbf{x}), \mathbf{y}-\mathbf{x}\rangle \ \leq 0$$



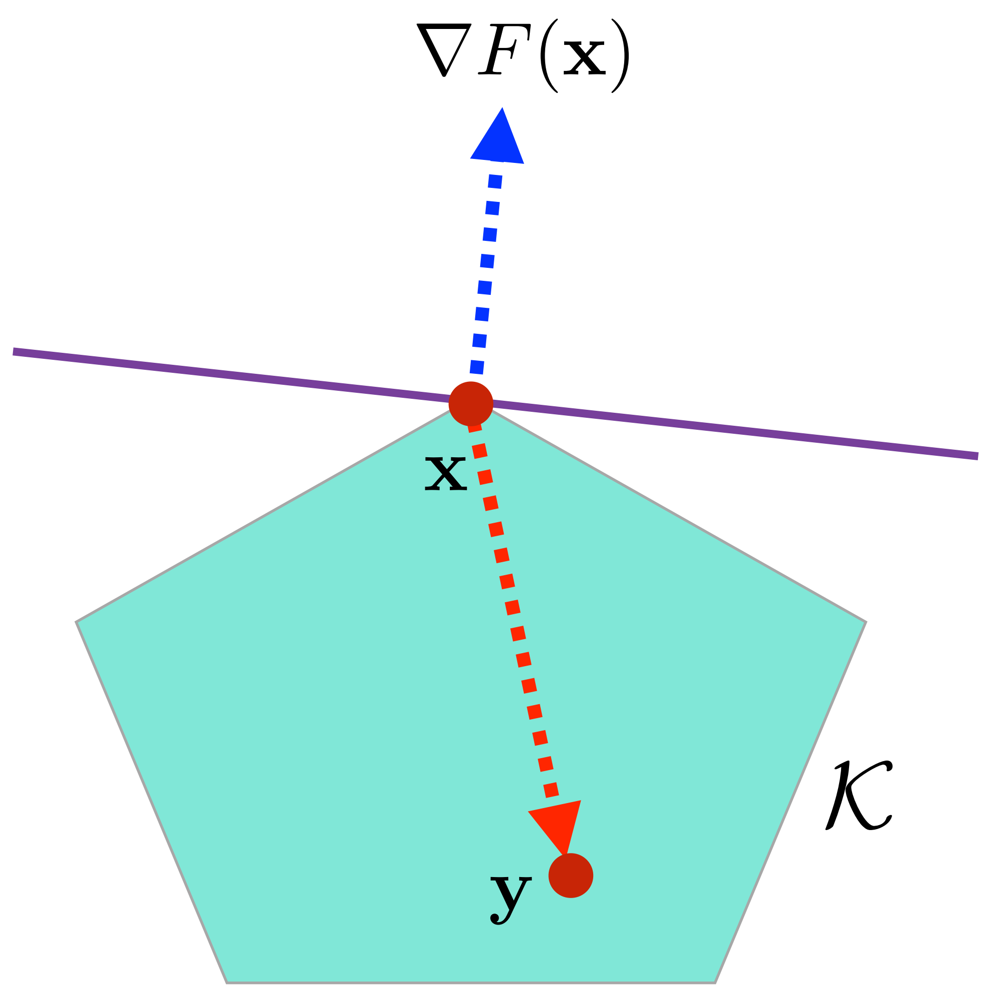
Intuition (first-order):

$F(\mathbf{y})$

Stationary Points and Their Quality

- Let's first characterize the fixed point of gradient ascent
- A vector $\mathbf{x}$ is a **stationary point** of a function F over the set $\mathcal{K}$ if

$$\max_{\mathbf{y} \in \mathcal{K}} \langle \nabla F(\mathbf{x}), \mathbf{y} - \mathbf{x} \rangle \leq 0$$



Intuition (first-order):

$$F(\mathbf{y}) = F(\mathbf{x} + (\mathbf{y} - \mathbf{x}))$$

$$\approx F(\mathbf{x}) + \langle \nabla F(\mathbf{x}), \mathbf{y} - \mathbf{x} \rangle$$

Stationary Points and Their Quality

- Let's first characterize the fixed point of gradient ascent
- A vector $\mathbf{x} \in \mathcal{K}$ is a **stationary point** of a function F over the set $\mathcal{K}$ if

$$\max_{\mathbf{y} \in \mathcal{K}} \langle \nabla F(\mathbf{x}), \mathbf{y} - \mathbf{x} \rangle \leq 0$$

Stationary Points and Their Quality

- Let's first characterize the fixed point of gradient ascent

- A vector $\mathbf{x} \in \mathcal{K}$ is a **stationary point** of a function F over the set $\mathcal{K}$ if

$$\max_{\mathbf{y} \in \mathcal{K}} \langle \nabla F(\mathbf{x}), \mathbf{y} - \mathbf{x} \rangle \leq 0$$

[Hassani, Soltanolkotabi, Karbasi]

If $\mathbf{x} \in \mathcal{K}$ is a stationary point of a monotone and DR-submodular function F, we have:

$$F(\mathbf{x}) \geq \frac{1}{2}\mathrm{OPT}$$

"Gradient Methods for Submodular Maximization", NIPS'17

Stationary Points and Their Quality

- Let's first characterize the fixed point of gradient ascent
- A vector $\mathbf{x} \in \mathcal{K}$ is a **stationary point** of a function F over the set $\mathcal{K}$ if

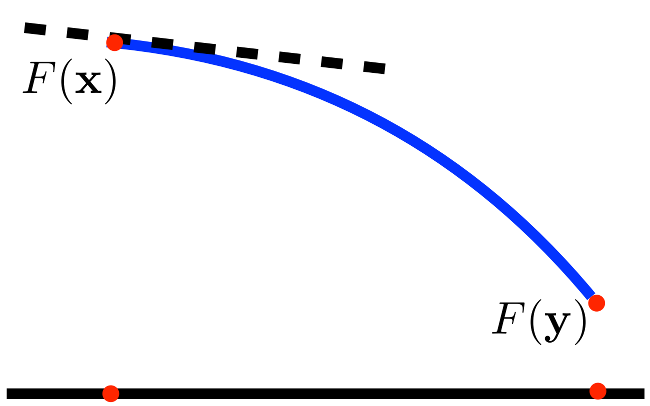
$$\max_{\mathbf{y} \in \mathcal{K}} \langle \nabla F(\mathbf{x}), \mathbf{y} - \mathbf{x} \rangle \ \leq 0$$

[Hassani, Soltanolkotabi, Karbasi]

If $\mathbf{x} \in \mathcal{K}$ is a stationary point of a monotone and DR-submodular function F, we have:

$$F(\mathbf{x}) \geq \frac{1}{2} \mathrm{OPT}$$

"Gradient Methods for Submodular Maximization", NIPS'17



concave: $F(\mathbf{y}) - F(\mathbf{x}) \leq \langle \nabla F(\mathbf{x}), \mathbf{y} - \mathbf{x} \rangle$

monotone DR-submodular: $F(\mathbf{y}) - 2F(\mathbf{x}) \leq \langle \nabla F(\mathbf{x}), \mathbf{y} - \mathbf{x} \rangle$

Can We Reach Stationary Points?

[Hassani, Soltanolkotabi, Karbasi]

For the (stochastic) projected gradient ascent algorithm, we have:

$$\mathbb{E}[F(\mathbf{x}_T)] \geq \frac{1}{2}\mathrm{OPT} - (\frac{R^2 L}{2T} + \frac{R\sigma}{\sqrt{T}})$$

"Gradient Methods for Submodular Maximization", NIPS'17

Can We Reach Stationary Points?

[Hassani, Soltanolkotabi, Karbasi]

For the (stochastic) projected gradient ascent algorithm, we have:

$$\mathbb{E}[F(\mathbf{x}_T)] \geq \frac{1}{2}\mathrm{OPT} - \left(\frac{R^2 L}{2T} + \frac{R\sigma}{\sqrt{T}}\right)$$

variance of stochastic gradients

diameter of $\mathcal{K}$ smoothness of F

"Gradient Methods for Submodular Maximization", NIPS'17

Can We Reach Stationary Points?

[Hassani, Soltanolkotabi, Karbasi]

For the (stochastic) projected gradient ascent algorithm, we have:

$$\mathbb{E}[F(\mathbf{x}_T)] \geq \frac{1}{2}\mathrm{OPT} - \left(\frac{R^2 L}{2T} + \frac{R\sigma}{\sqrt{T}}\right)$$

variance of stochastic gradients

diameter of $\mathcal{K}$

smoothness of F

"Gradient Methods for Submodular Maximization", NIPS'17

- To achieve $\frac{1}{2}\mathrm{OPT} - \epsilon$ we require $T = \mathcal{O}\left(\frac{R^2 L}{\epsilon} + \frac{R^2\sigma^2}{\epsilon^2}\right)$

Can We Reach Stationary Points?

[Hassani, Soltanolkotabi, Karbasi]

For the (stochastic) projected gradient ascent algorithm, we have:

$$\mathbb{E}[F(\mathbf{x}_T)] \geq \frac{1}{2}\text{OPT} - \left(\frac{R^2 L}{2T} + \frac{R\sigma}{\sqrt{T}}\right)$$

variance of stochastic gradients

diameter of $\mathcal{K}$ smoothness of F

"Gradient Methods for Submodular Maximization", NIPS'17

- To achieve $\frac{1}{2}\text{OPT} - \epsilon$ we require $T = \mathcal{O}\left(\frac{R^2 L}{\epsilon} + \frac{R^2\sigma^2}{\epsilon^2}\right)$
- Can stochastic gradient ascent do better than 1/2?

Can We Reach Stationary Points?

[Hassani, Soltanolkotabi, Karbasi]

For the (stochastic) projected gradient ascent algorithm, we have:

$$\mathbb{E}[F(\mathbf{x}_T)] \geq \frac{1}{2}\mathrm{OPT} - \left(\frac{R^2 L}{2T} + \frac{R\sigma}{\sqrt{T}}\right)$$

variance of stochastic gradients

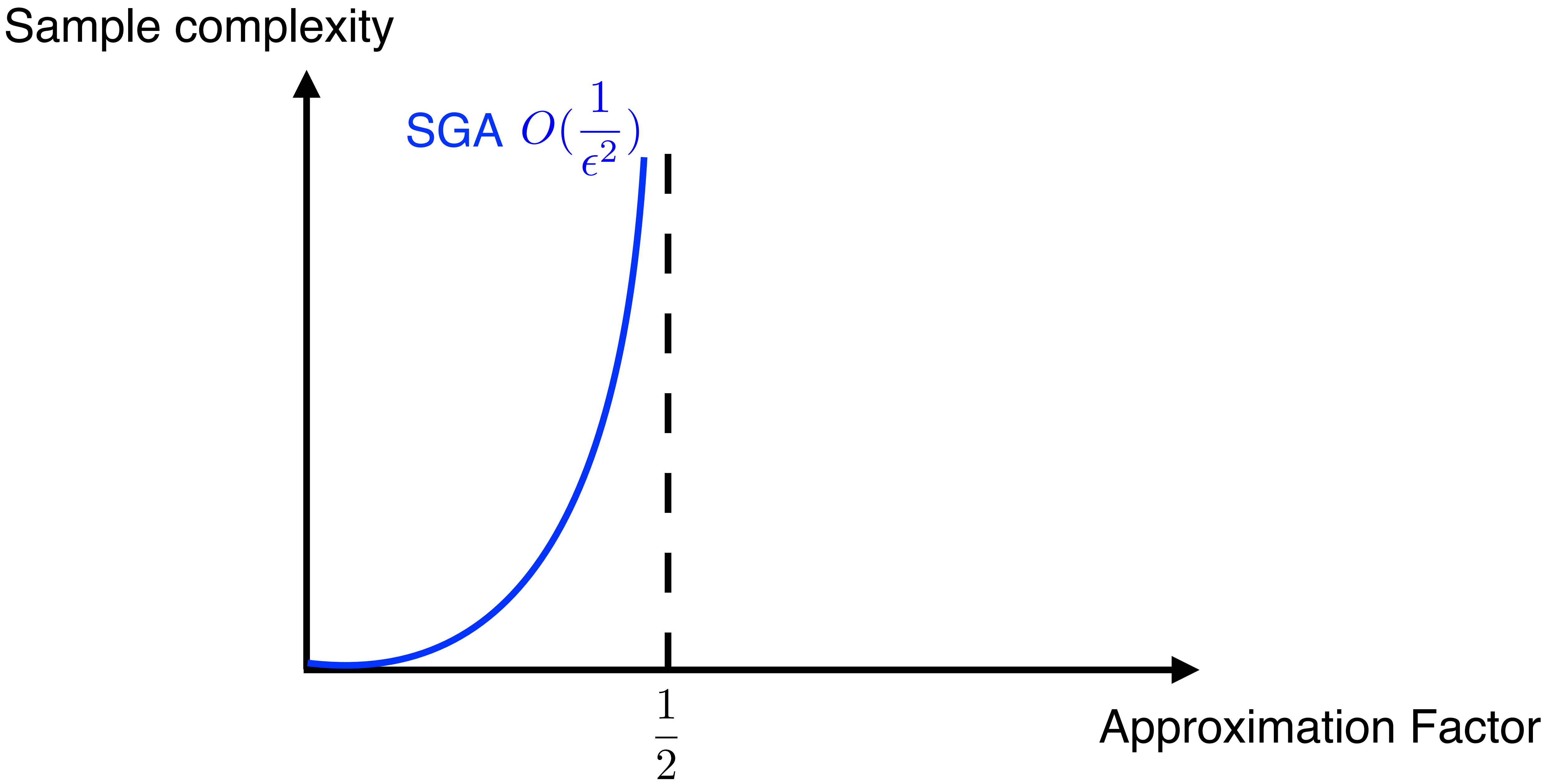
diameter of $\mathcal{K}$ smoothness of F

"Gradient Methods for Submodular Maximization", NIPS'17

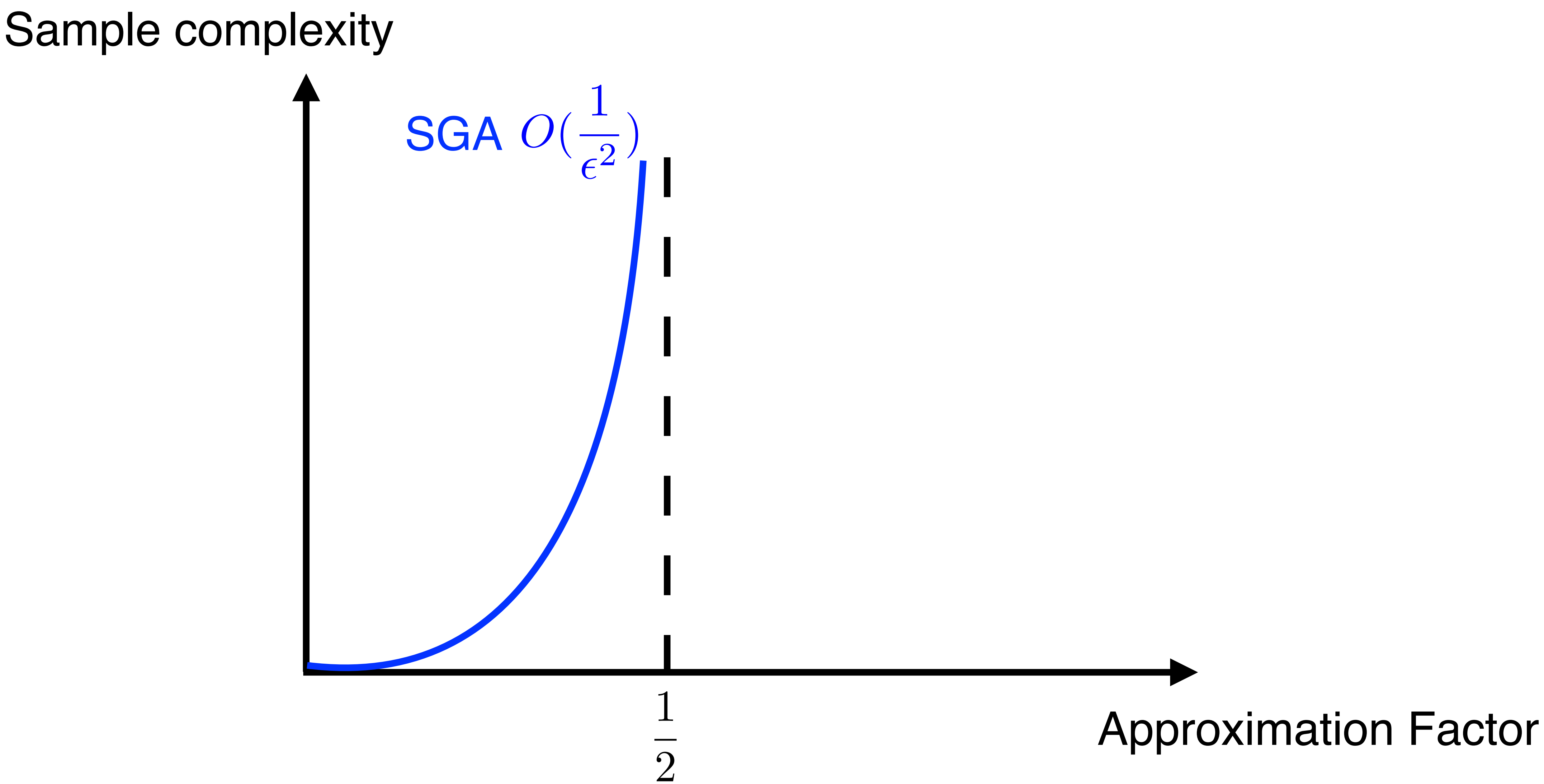
- To achieve $\frac{1}{2}\mathrm{OPT} - \epsilon$ we require $T = \mathcal{O}\left(\frac{R^2 L}{\epsilon} + \frac{R^2\sigma^2}{\epsilon^2}\right)$
- Can stochastic gradient ascent do better than 1/2?
  - No. There exists a DR-Submodular function that attains OPT/2 at a stationary point which is also a *local maximum.*

Yale Penn

Maximizing DR-submodular Functions

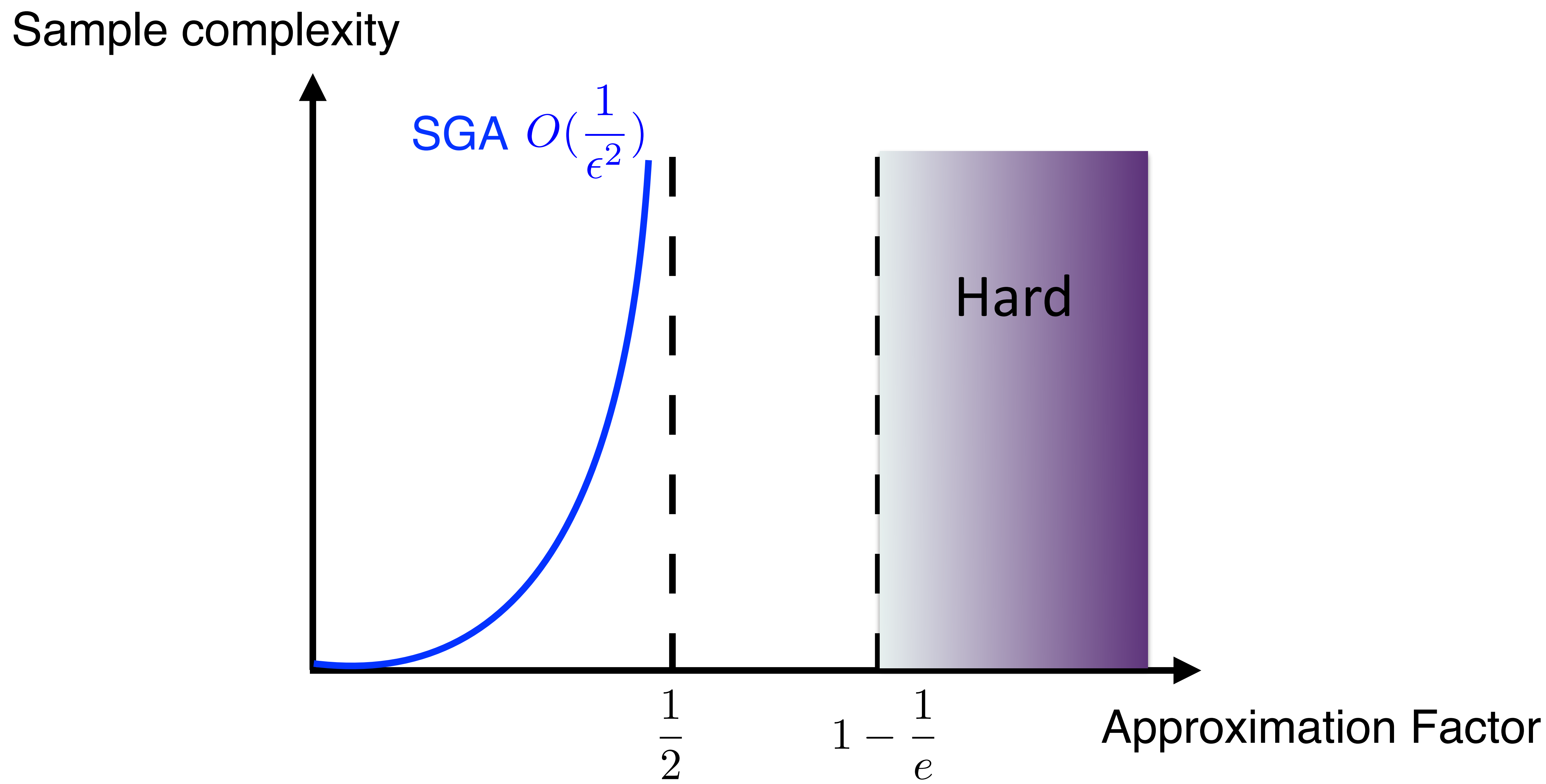


Maximizing DR-submodular Functions



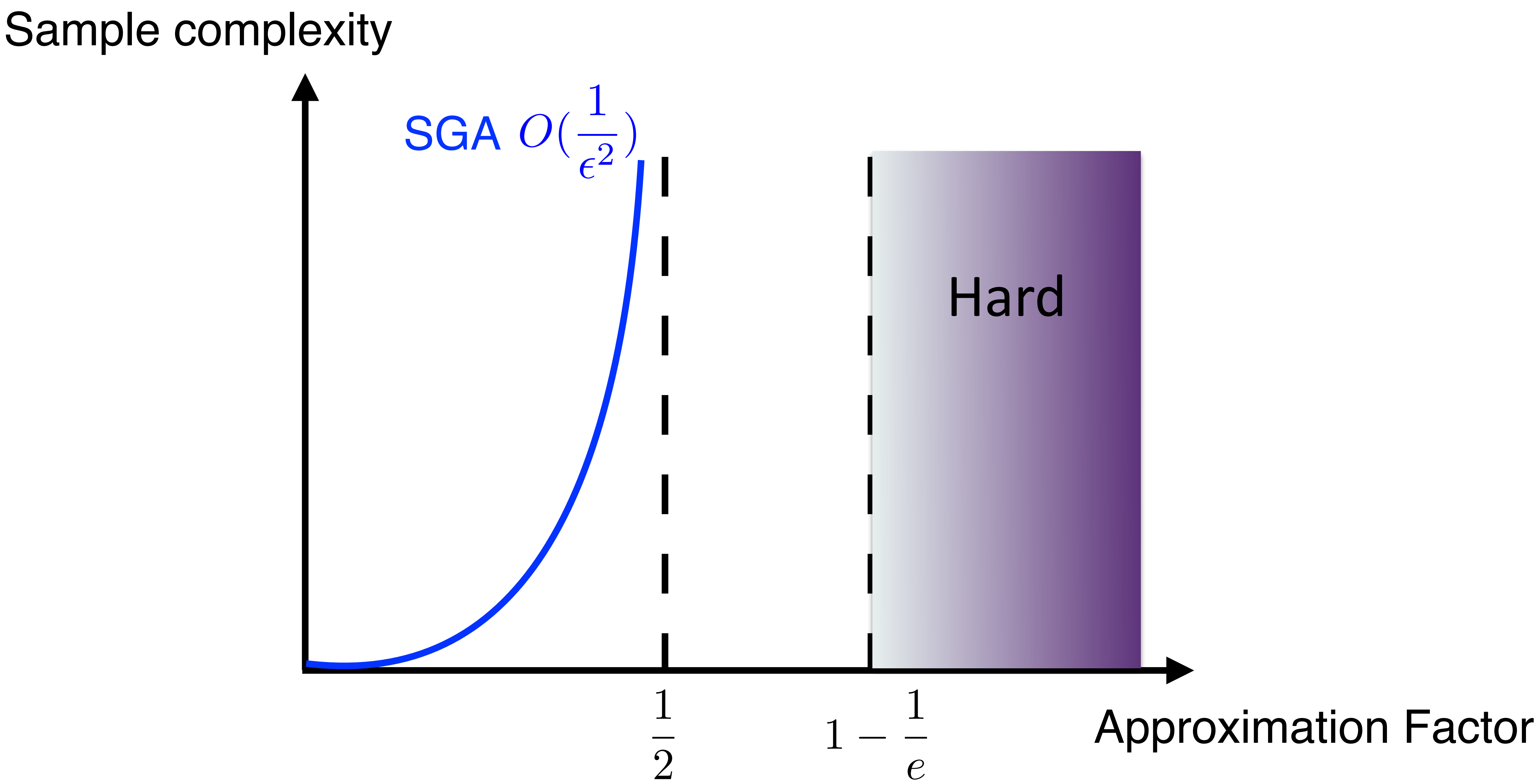
- Can we achieve better than 1/2?

Maximizing DR-submodular Functions



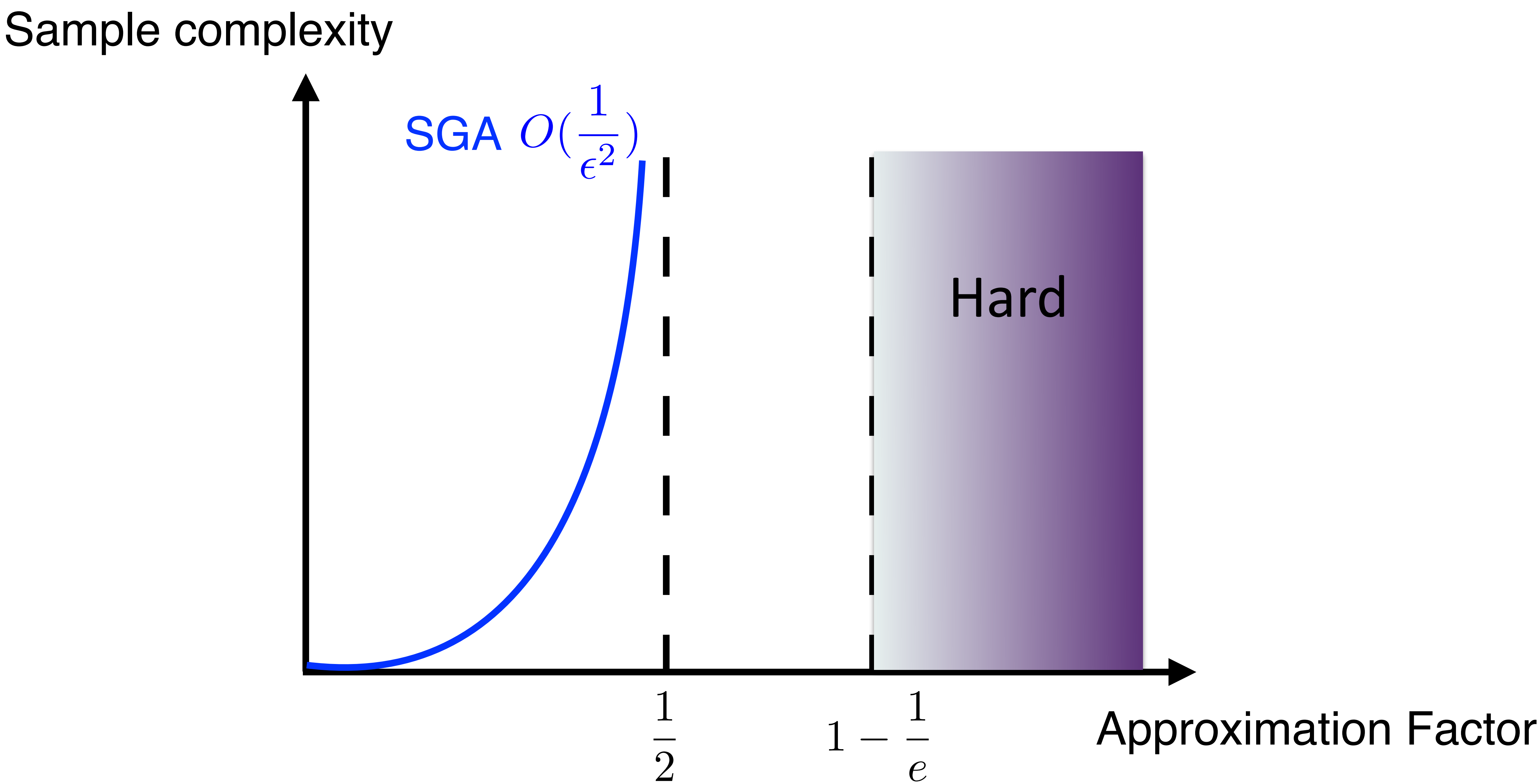
- Can we achieve better than 1/2?

Maximizing DR-submodular Functions



- Can we achieve better than 1/2? ▸ Yes

Maximizing DR-submodular Functions



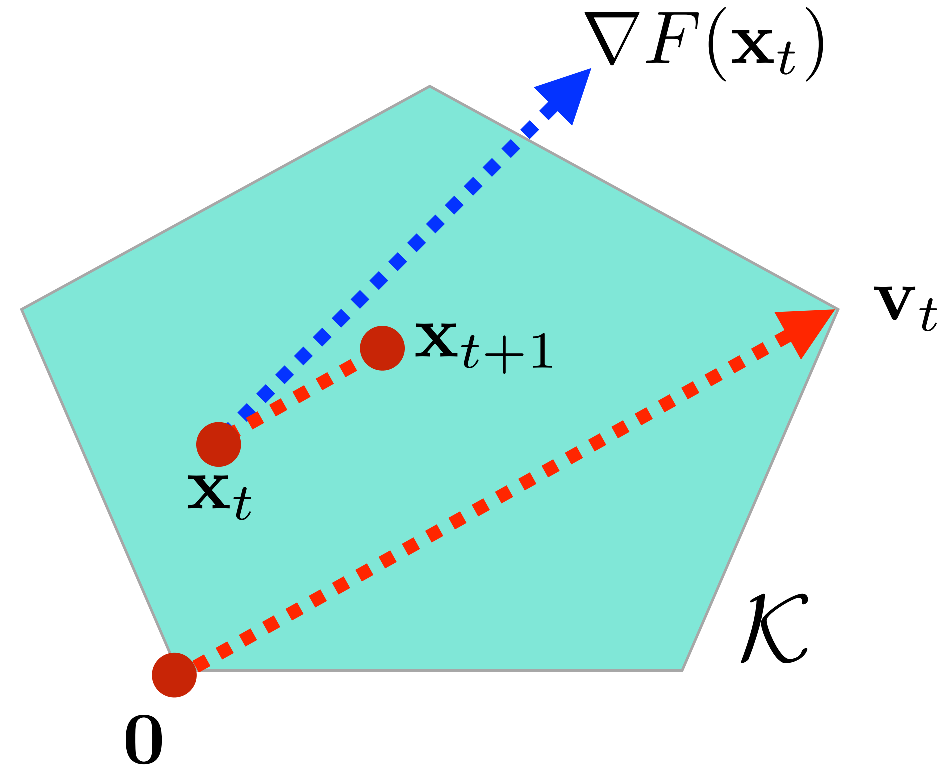
- Can we achieve better than 1/2? ▸ Yes
- How? ▸ Smart initialization + conditional gradient methods

Continuous Greedy

- Always initialize at $\mathbf{x}_0 = \mathbf{0}$
- Repeat for T iterations:

$$\mathbf{v}_t = \arg\max_{\mathbf{v} \in \mathcal{K}} \langle \nabla F(\mathbf{x}_t), \mathbf{v} \rangle$$

$$\mathbf{x}_{t+1} = \mathbf{x}_t + \frac{1}{T} \mathbf{v}_t$$



"Maximizing a monotone submodular function subject to a matroid constraint", Calinescu, Chekuri, Pal, Vondrák, 2011

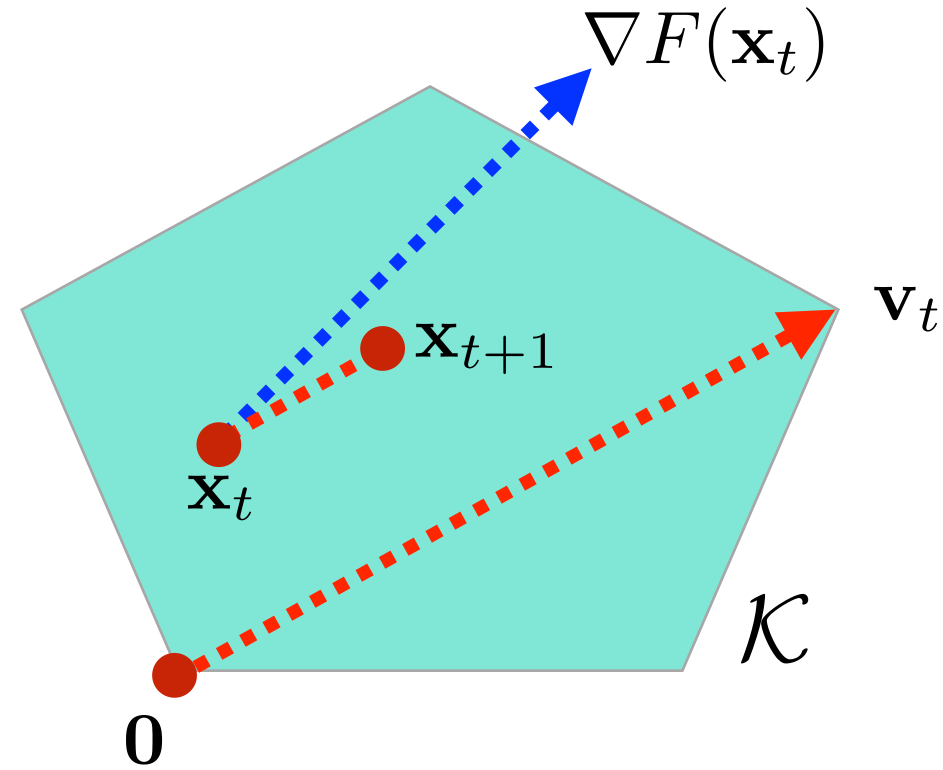
"Guaranteed non-convex optimization: Submodular maximization over continuous domains", Bian, Mirzasoleiman, Buhmann, Krause, 2017

Continuous Greedy

- Always initialize at $\mathbf{x}_0 = \mathbf{0}$
- Repeat for T iterations:

$$\mathbf{v}_t = \arg\max_{\mathbf{v} \in \mathcal{K}} \langle \nabla F(\mathbf{x}_t), \mathbf{v} \rangle$$

$$\mathbf{x}_{t+1} = \mathbf{x}_t + \frac{1}{T} \mathbf{v}_t$$



[Bian, Mirzasoleiman, Buhmann, Krause]

For the continuous greedy algorithm, we have:

$$F(\mathbf{x}_T) \geq (1 - \frac{1}{e})\mathrm{OPT} - \frac{LR}{2T}$$

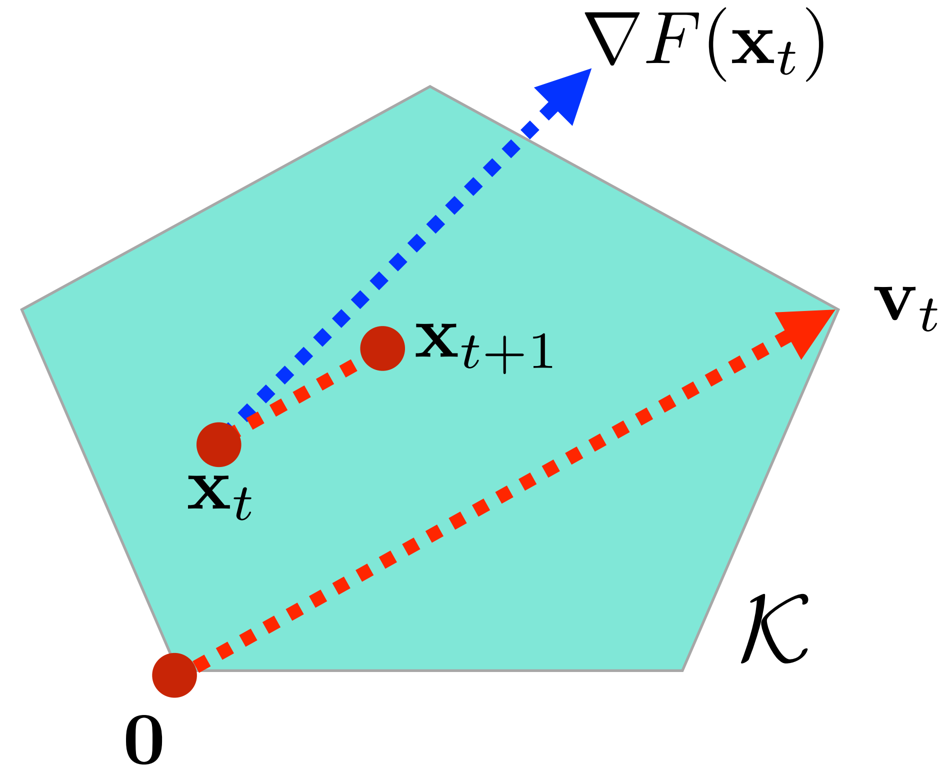
"Guaranteed non-convex optimization: Submodular maximization over continuous domains", AISTATS'17

Continuous Greedy

- Always initialize at $\mathbf{x}_0 = \mathbf{0}$
- Repeat for T iterations:

$$\mathbf{v}_t = \arg\max_{\mathbf{v}\in\mathcal{K}} \langle \nabla F(\mathbf{x}_t), \mathbf{v}\rangle$$

$$\mathbf{x}_{t+1} = \mathbf{x}_t + \frac{1}{T}\mathbf{v}_t$$



[Bian, Mirzasoleiman, Buhmann, Krause]

For the continuous greedy algorithm, we have:

$$F(\mathbf{x}_T) \geq (1 - \frac{1}{e})\mathrm{OPT} - \frac{LR}{2T}$$

"Guaranteed non-convex optimization: Submodular maximization over continuous domains", AISTATS'17

- The continuous greedy is a variant of the Frank-Wolfe method

Yale Penn

Stochastic Continuous Greedy

- Stochastic optimization: cannot compute the ascent direction

$$\mathbf{v}_t = \text{argmax}_{\mathbf{v} \in \mathcal{K}} \{ \langle \nabla F(\mathbf{x}_t), \mathbf{v} \rangle \}$$

Stochastic Continuous Greedy

- Stochastic optimization: cannot compute the ascent direction

$$\mathbf{v}_t = \text{argmax}_{\mathbf{v} \in \mathcal{K}} \{ \langle \nabla F(\mathbf{x}_t), \mathbf{v} \rangle \}$$

- Naive idea: use the stochastic gradient $\mathbb{E}[\mathbf{g}_t] = \nabla F(\mathbf{x}_t)$

Stochastic Continuous Greedy

- Stochastic optimization: cannot compute the ascent direction

$$\mathbf{v}_t = \text{argmax}_{\mathbf{v}\in\mathcal{K}}\{\langle \nabla F(\mathbf{x}_t), \mathbf{v}\rangle\}$$

- Naive idea: use the stochastic gradient $\mathbb{E}[\mathbf{g}_t] = \nabla F(\mathbf{x}_t)$
- Does not work: $\mathbb{E}[\text{argmax}_{\mathbf{v}\in\mathcal{K}}\{\langle \mathbf{g}_t, \mathbf{v}\rangle\}] \neq \text{argmax}_{\mathbf{v}\in\mathcal{K}}\{\langle \mathbb{E}[\mathbf{g}_t], \mathbf{v}\rangle\}$

Stochastic Continuous Greedy

- Stochastic optimization: cannot compute the ascent direction

$$\mathbf{v}_t = \text{argmax}_{\mathbf{v} \in \mathcal{K}} \{ \langle \nabla F(\mathbf{x}_t), \mathbf{v} \rangle \}$$

Stochastic Continuous Greedy

- Stochastic optimization: cannot compute the ascent direction

$$\mathbf{v}_t = \text{argmax}_{\mathbf{v}\in\mathcal{K}}\{\langle \nabla F(\mathbf{x}_t), \mathbf{v}\rangle\}$$

- Idea: Reduce variance through momentum

Stochastic Continuous Greedy

- Stochastic optimization: cannot compute the ascent direction

$$\mathbf{v}_t = \text{argmax}_{\mathbf{v} \in \mathcal{K}} \{\langle \nabla F(\mathbf{x}_t), \mathbf{v} \rangle\}$$

- Idea: Reduce variance through momentum
- Replace $\nabla F(\mathbf{x}_t)$ by weighted average of $\{\mathbf{g}_s\}_{s=0}^{t}$

$$\boxed{\mathbf{d}_t = (1 - \rho_t)\mathbf{d}_{t-1} + \rho_t \mathbf{g}_t}$$

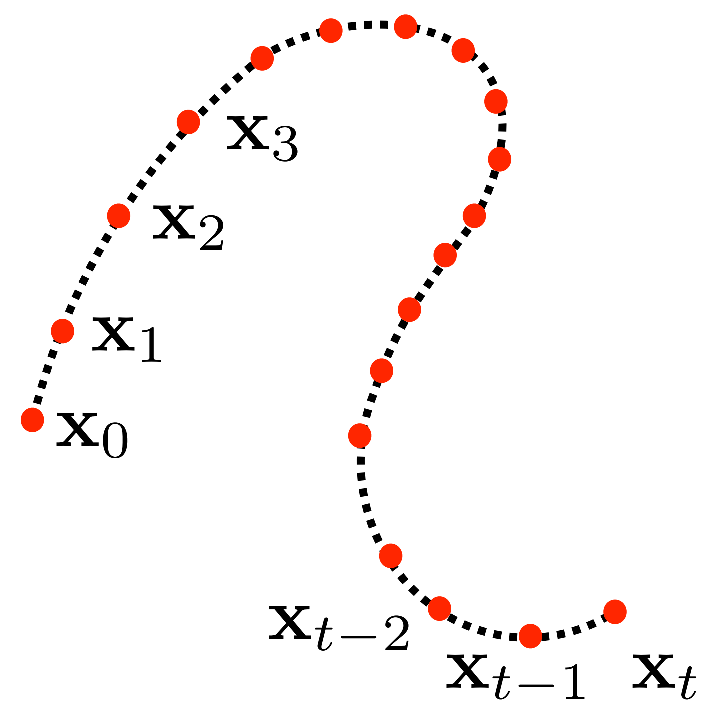
Stochastic Continuous Greedy

- Stochastic optimization: cannot compute the ascent direction

$$\mathbf{v}_t = \text{argmax}_{\mathbf{v}\in\mathcal{K}}\{\langle\nabla F(\mathbf{x}_t), \mathbf{v}\rangle\}$$

- Idea: Reduce variance through momentum
- Replace $\nabla F(\mathbf{x}_t)$ by weighted average of $\{\mathbf{g}_s\}_{s=0}^{t}$

$$\boxed{\mathbf{d}_t = (1-\rho_t)\mathbf{d}_{t-1} + \rho_t\mathbf{g}_t}$$



Stochastic Continuous Greedy

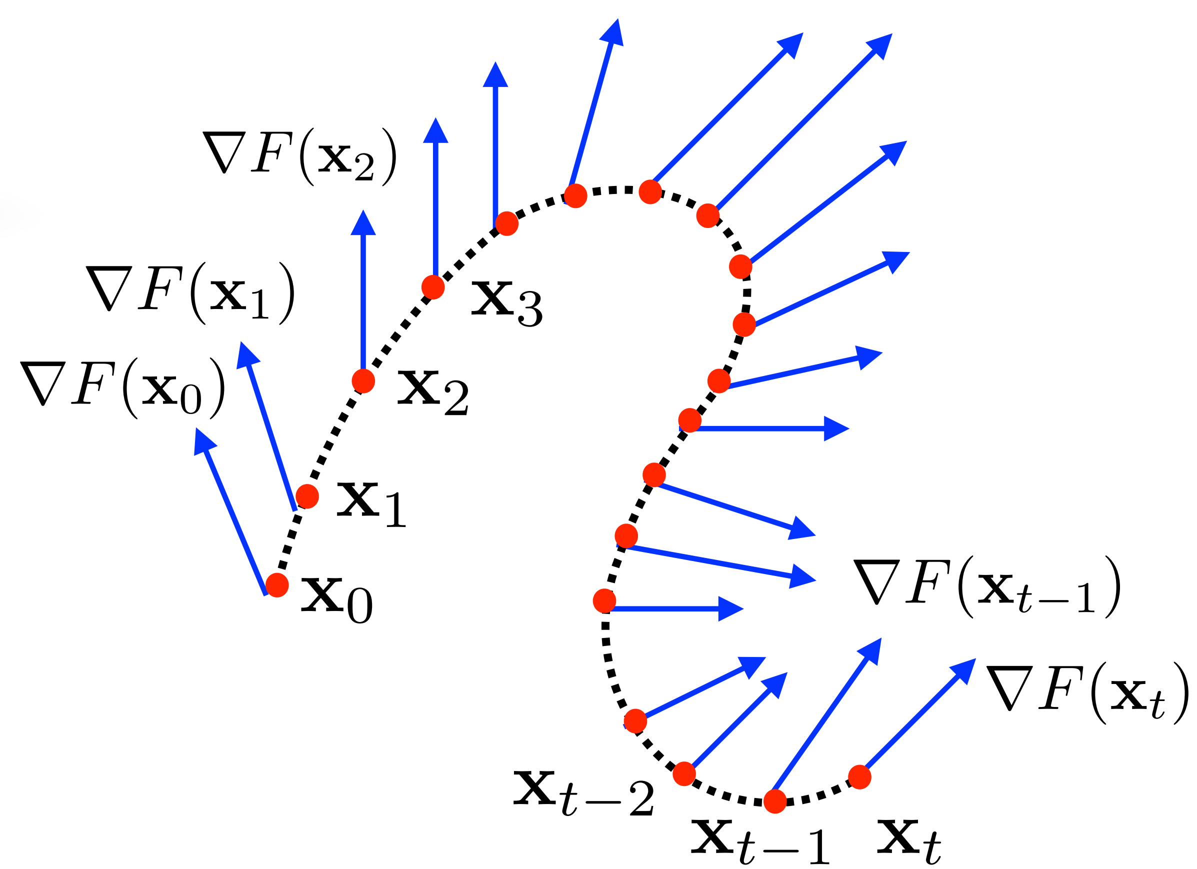
- Stochastic optimization: cannot compute the ascent direction

$$\mathbf{v}_t = \mathrm{argmax}_{\mathbf{v} \in \mathcal{K}} \{ \langle \nabla F(\mathbf{x}_t), \mathbf{v} \rangle \}$$

- Idea: Reduce variance through momentum

- Replace $\nabla F(\mathbf{x}_t)$ by weighted average of $\{\mathbf{g}_s\}_{s=0}^{t}$

$$\mathbf{d}_t = (1 - \rho_t)\mathbf{d}_{t-1} + \rho_t \mathbf{g}_t$$



Stochastic Continuous Greedy

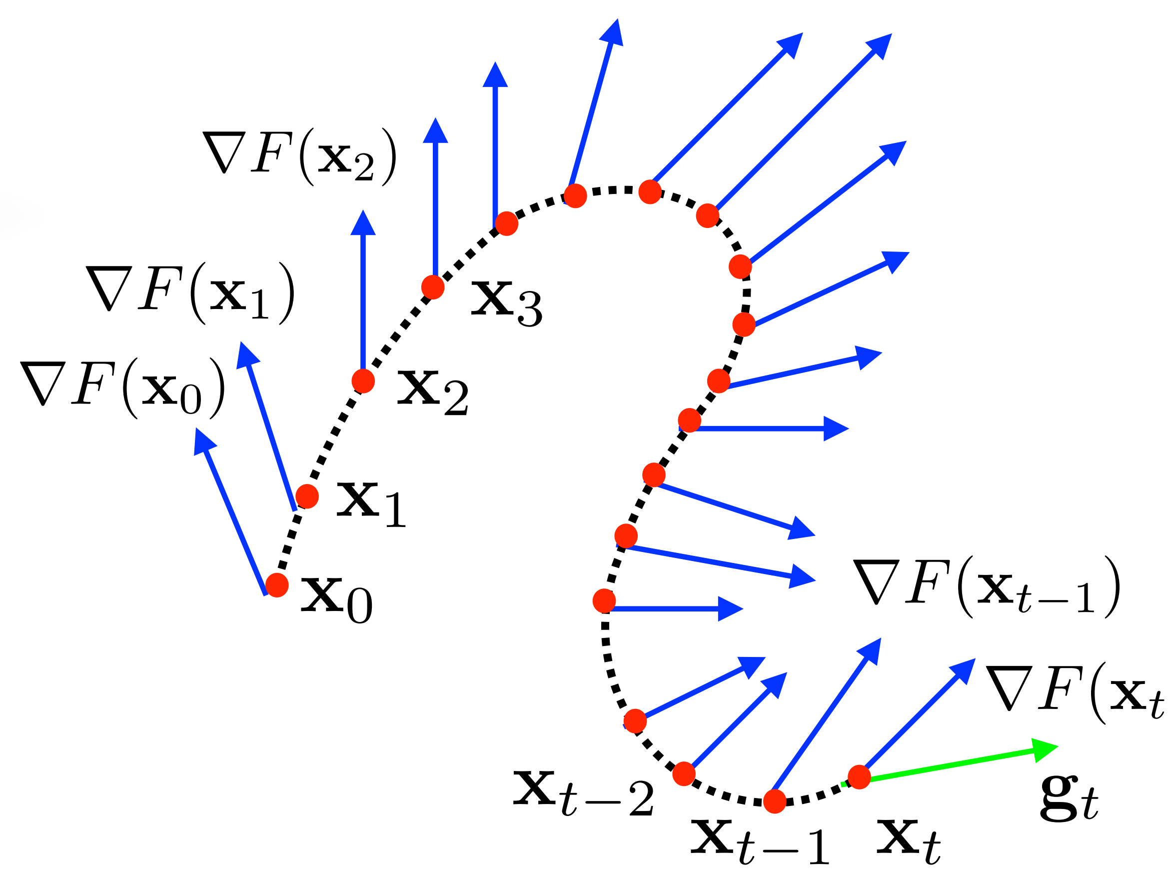
- Stochastic optimization: cannot compute the ascent direction

$$\mathbf{v}_t = \mathrm{argmax}_{\mathbf{v}\in\mathcal{K}}\{\langle \nabla F(\mathbf{x}_t), \mathbf{v}\rangle\}$$

- Idea: Reduce variance through momentum

- Replace $\nabla F(\mathbf{x}_t)$ by weighted average of $\{\mathbf{g}_s\}_{s=0}^{t}$

$$\boxed{\mathbf{d}_t = (1-\rho_t)\mathbf{d}_{t-1} + \rho_t \mathbf{g}_t}$$



Stochastic Continuous Greedy

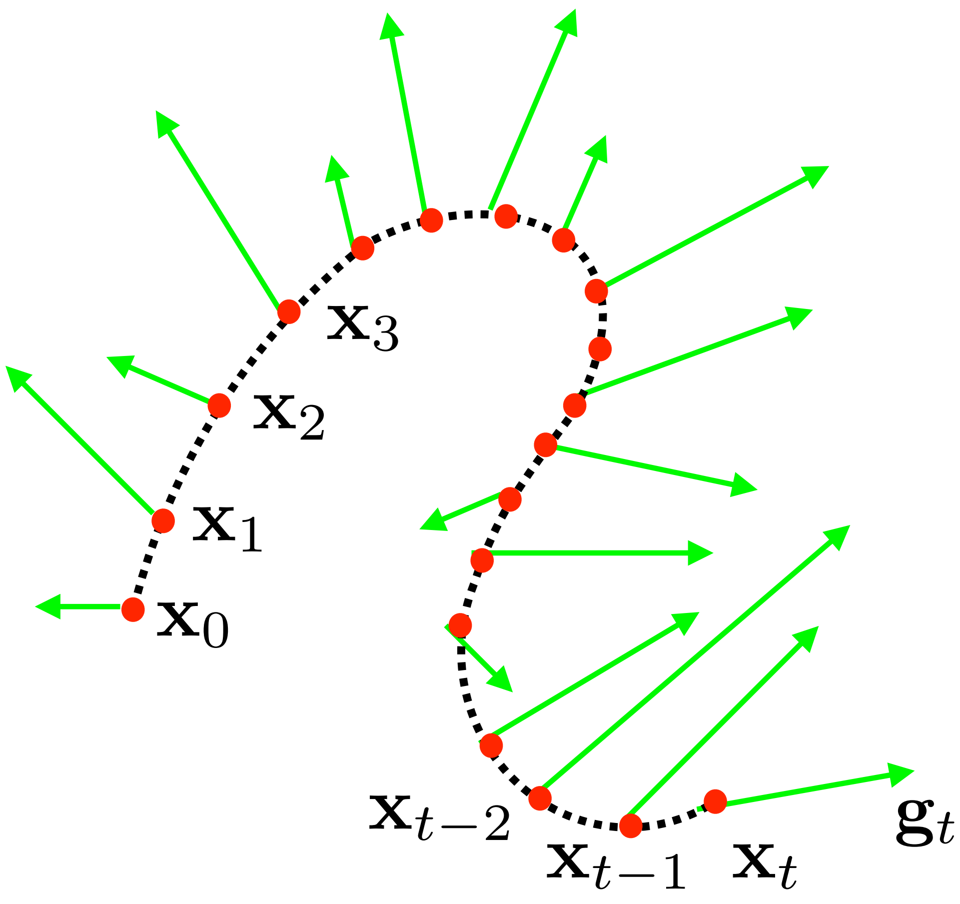
- Stochastic optimization: cannot compute the ascent direction

$$\mathbf{v}_t = \text{argmax}_{\mathbf{v} \in \mathcal{K}} \{ \langle \nabla F(\mathbf{x}_t), \mathbf{v} \rangle \}$$

- Idea: Reduce variance through momentum

- Replace $\nabla F(\mathbf{x}_t)$ by weighted average of $\{\mathbf{g}_s\}_{s=0}^{t}$

$$\boxed{\mathbf{d}_t = (1-\rho_t)\mathbf{d}_{t-1} + \rho_t \mathbf{g}_t}$$



Yale Penn

Stochastic Continuous Greedy

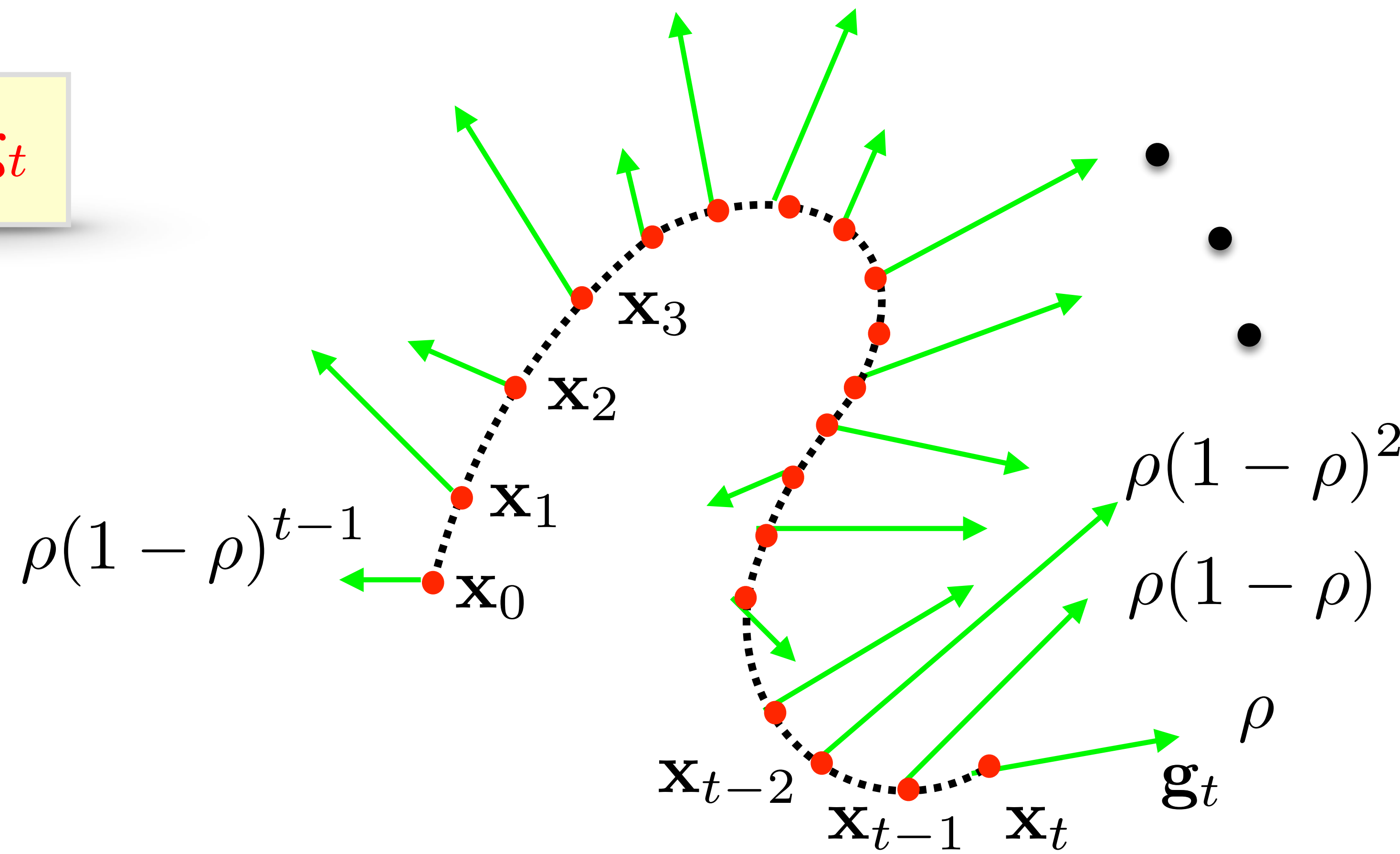
- Stochastic optimization: cannot compute the ascent direction

$$\mathbf{v}_t = \operatorname{argmax}_{\mathbf{v}\in\mathcal{K}}\{\langle \nabla F(\mathbf{x}_t), \mathbf{v}\rangle\}$$

- Idea: Reduce variance through momentum

- Replace $\nabla F(\mathbf{x}_t)$ by weighted average of $\{\mathbf{g}_s\}_{s=0}^{t}$

$$\boxed{\mathbf{d}_t = (1-\rho_t)\mathbf{d}_{t-1} + \rho_t \mathbf{g}_t}$$



Stochastic Continuous Greedy

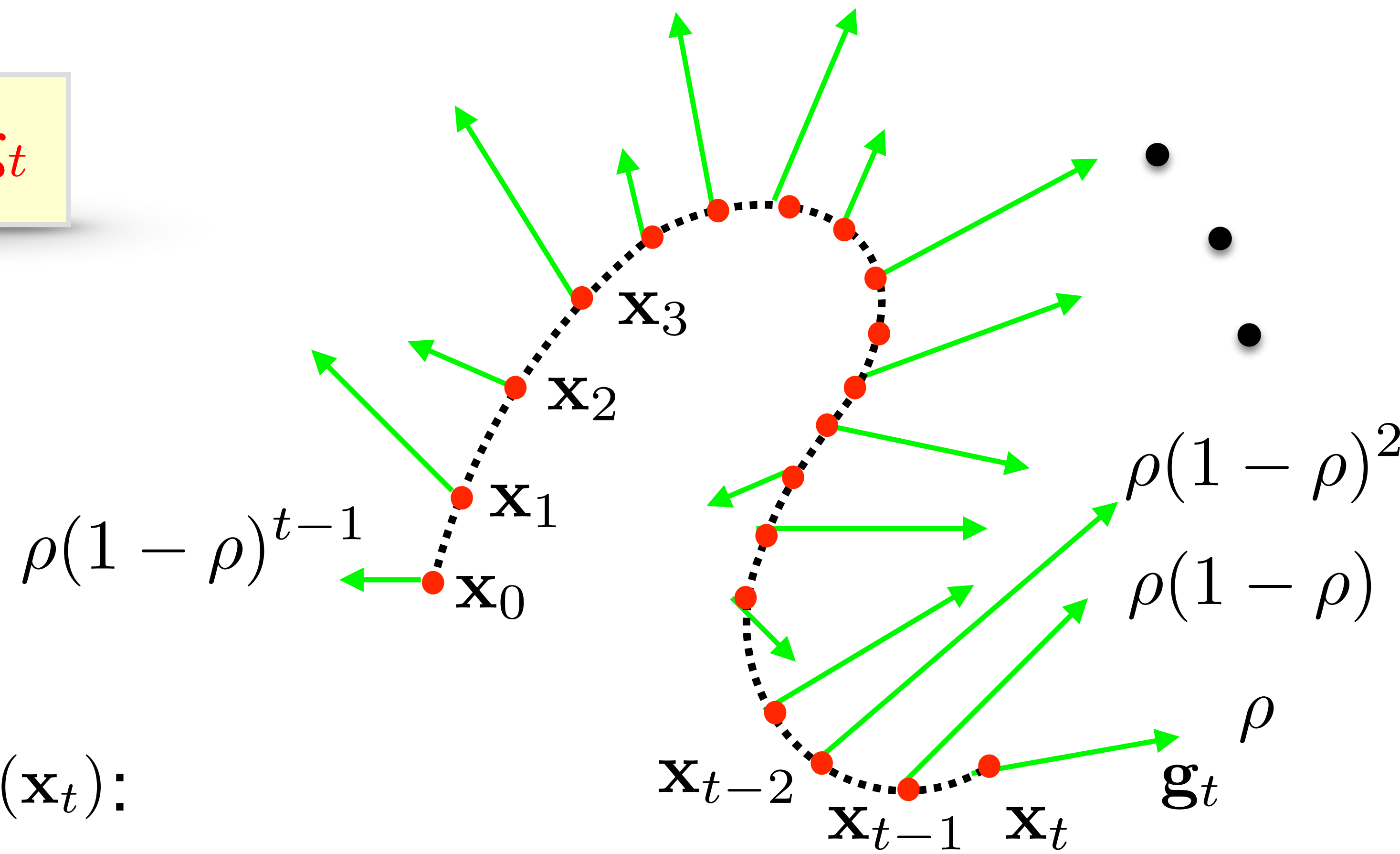
- Stochastic optimization: cannot compute the ascent direction

$$\mathbf{v}_t = \text{argmax}_{\mathbf{v}\in\mathcal{K}}\{\langle\nabla F(\mathbf{x}_t), \mathbf{v}\rangle\}$$

- Idea: Reduce variance through momentum

- Replace $\nabla F(\mathbf{x}_t)$ by weighted average of $\{\mathbf{g}_s\}_{s=0}^{t}$

$$\boxed{\mathbf{d}_t = (1-\rho_t)\mathbf{d}_{t-1} + \rho_t \mathbf{g}_t}$$



- $\mathbf{d}_t$ is a good approximation for $\nabla F(\mathbf{x}_t)$:
  - Computationally affordable $\Rightarrow$ single gradient evaluation
  - Reduces the noise of gradient approximation

Yale Penn

Stochastic Continuous Greedy

[Mokhtari, Hassani, Karbasi]

Assuming smoothness of F, by letting $\rho_t = \dfrac{4}{(t+8)^{2/3}}$, we have

$$\mathbb{E}[||\nabla F(\mathbf{x}_t) - \mathbf{d}_t||^2] \leq \frac{Q}{(t+9)^{2/3}}$$

where $Q := \max\left\{\, 5||\nabla F(\mathbf{x}_0) - \mathbf{d}_0||^2 \,,\, 16\sigma^2 + 2L^2R^2 \,\right\}$.

"Stochastic Conditional Gradient Methods: From Convex Minimization to Submodular Maximization", AISTATS '18, JMLR '20

Yale

Penn

Stochastic Continuous Greedy

[Mokhtari, Hassani, Karbasi]

Assuming smoothness of F, by letting $\rho_t = \frac{4}{(t+8)^{2/3}}$, we have

$$\mathbb{E}[||\nabla F(\mathbf{x}_t) - \mathbf{d}_t||^2] \leq \frac{Q}{(t+9)^{2/3}}$$

where $Q := \max\{\, 5||\nabla F(\mathbf{x}_0) - \mathbf{d}_0||^2 \,,\, 16\sigma^2 + 2L^2R^2 \,\}$.

"Stochastic Conditional Gradient Methods: From Convex Minimization to Submodular Maximization", AISTATS '18, JMLR '20

$$\mathbf{v}_t = \text{argmax}_{\mathbf{v}\in\mathcal{K}}\{\langle \nabla F(\mathbf{x}_t), \mathbf{v}\rangle\} \quad \longrightarrow \quad \mathbf{v}_t = \text{argmax}_{\mathbf{v}\in\mathcal{K}}\{\langle \mathbf{d}_t, \mathbf{v}\rangle\}$$

Stochastic Continuous Greedy

- Always initialize at $\mathbf{x}_0 = \mathbf{0}$
- Repeat for T iterations:

$$\mathbf{v}_t = \arg\max_{\mathbf{v} \in \mathcal{K}} \langle \mathbf{d}_t, \mathbf{v} \rangle$$

$$\mathbf{x}_{t+1} = \mathbf{x}_t + \frac{1}{T}\mathbf{v}_t$$

$$\mathbf{d}_{t+1} = (1 - \rho_t)\,\mathbf{d}_t + \rho_t\,\mathbf{g}_t$$

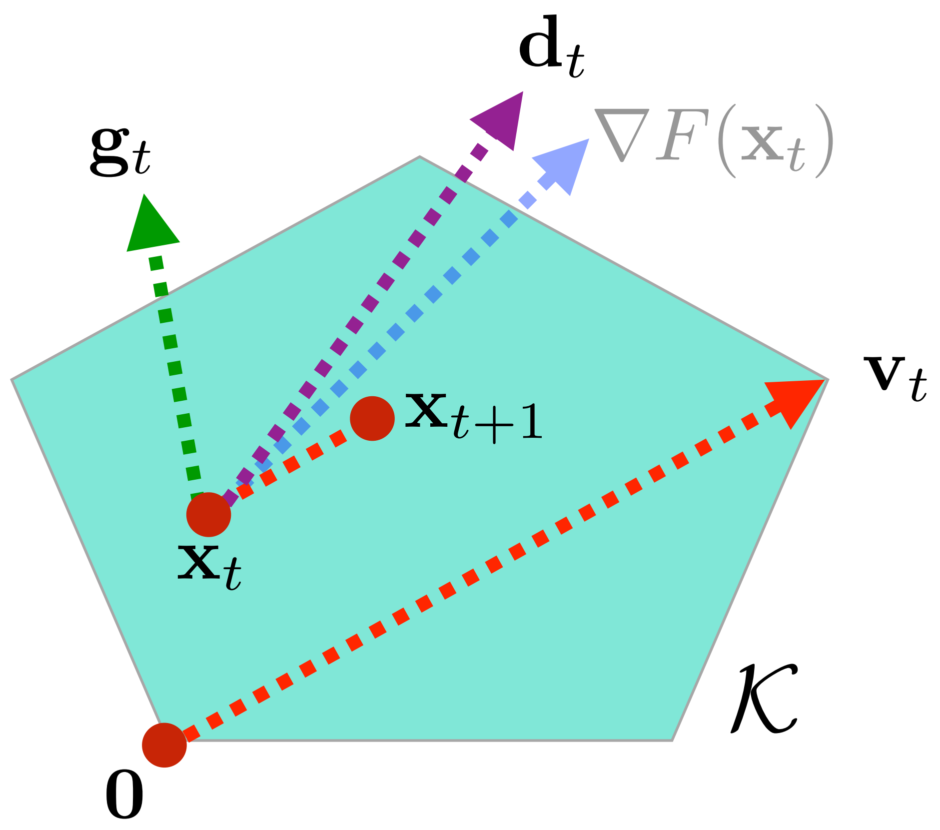
Stochastic Continuous Greedy

- Always initialize at $\mathbf{x}_0 = \mathbf{0}$
- Repeat for T iterations:

$$\mathbf{v}_t = \arg\max_{\mathbf{v}\in\mathcal{K}} \langle \mathbf{d}_t, \mathbf{v} \rangle$$

$$\mathbf{x}_{t+1} = \mathbf{x}_t + \frac{1}{T}\mathbf{v}_t$$

$$\mathbf{d}_{t+1} = (1-\rho_t)\,\mathbf{d}_t + \rho_t\,\mathbf{g}_t$$



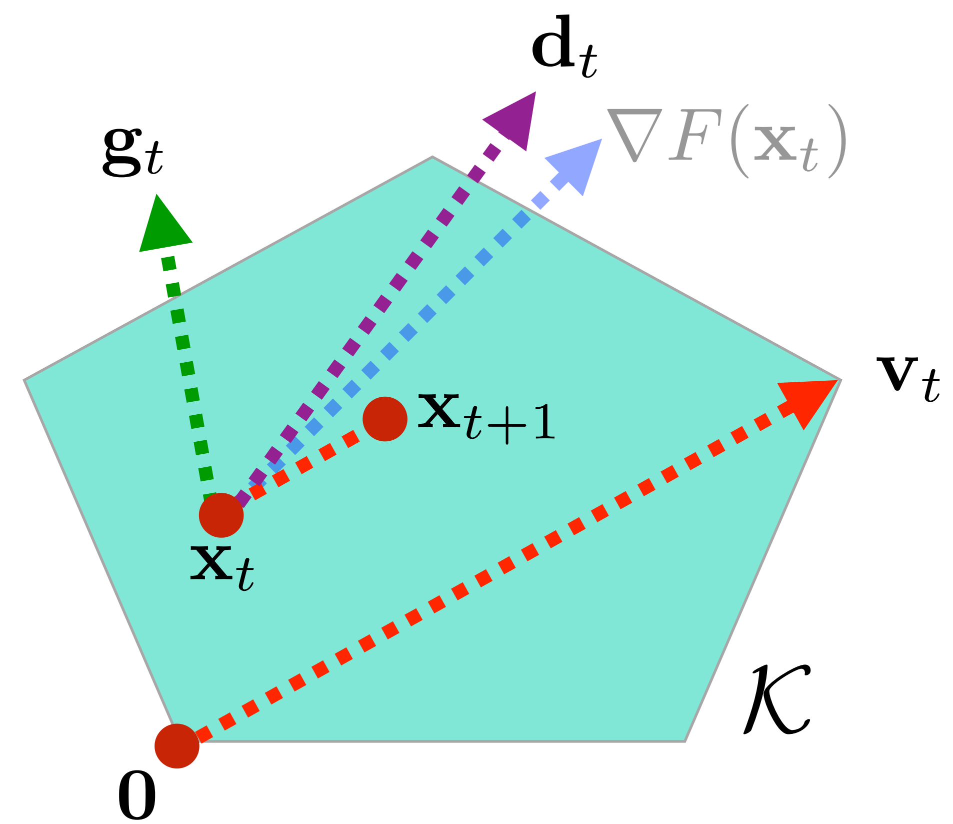
Stochastic Continuous Greedy

- Always initialize at $\mathbf{x}_0 = \mathbf{0}$
- Repeat for T iterations:

$$\mathbf{v}_t = \arg\max_{\mathbf{v}\in\mathcal{K}} \langle \mathbf{d}_t, \mathbf{v}\rangle$$

$$\mathbf{x}_{t+1} = \mathbf{x}_t + \frac{1}{T}\mathbf{v}_t$$

$$\mathbf{d}_{t+1} = (1-\rho_t)\,\mathbf{d}_t + \rho_t\,\mathbf{g}_t$$



[Mokhtari, Hassani, Karbasi]

For the stochastic continuous greedy (SCG) algorithm, we have:

$$\mathbb{E}[F(\mathbf{x}_T)] \geq (1-\frac{1}{e})\mathrm{OPT} - \frac{15RQ^{\frac{1}{2}}}{T^{\frac{1}{3}}} - \frac{LR^2}{2T}$$

"Stochastic Conditional Gradient Methods: From Convex Minimization to Submodular Maximization", AISTATS '18, JMLR '20

Yale

Penn

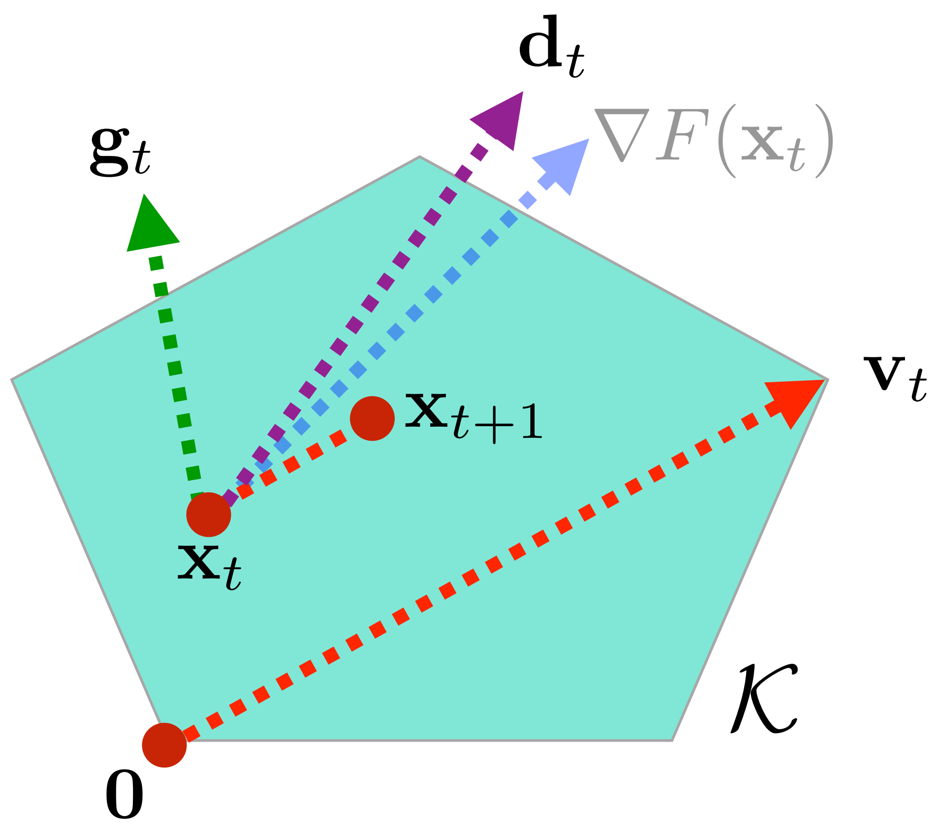
Stochastic Continuous Greedy

- Always initialize at $\mathbf{x}_0 = \mathbf{0}$
- Repeat for T iterations:

$$\mathbf{v}_t = \arg\max_{\mathbf{v}\in\mathcal{K}} \langle \mathbf{d}_t, \mathbf{v} \rangle$$

$$\mathbf{x}_{t+1} = \mathbf{x}_t + \frac{1}{T}\mathbf{v}_t$$

$$\mathbf{d}_{t+1} = (1-\rho_t)\,\mathbf{d}_t + \rho_t\,\mathbf{g}_t$$



[Mokhtari, Hassani, Karbasi]

For the stochastic continuous greedy (SCG) algorithm, we have:

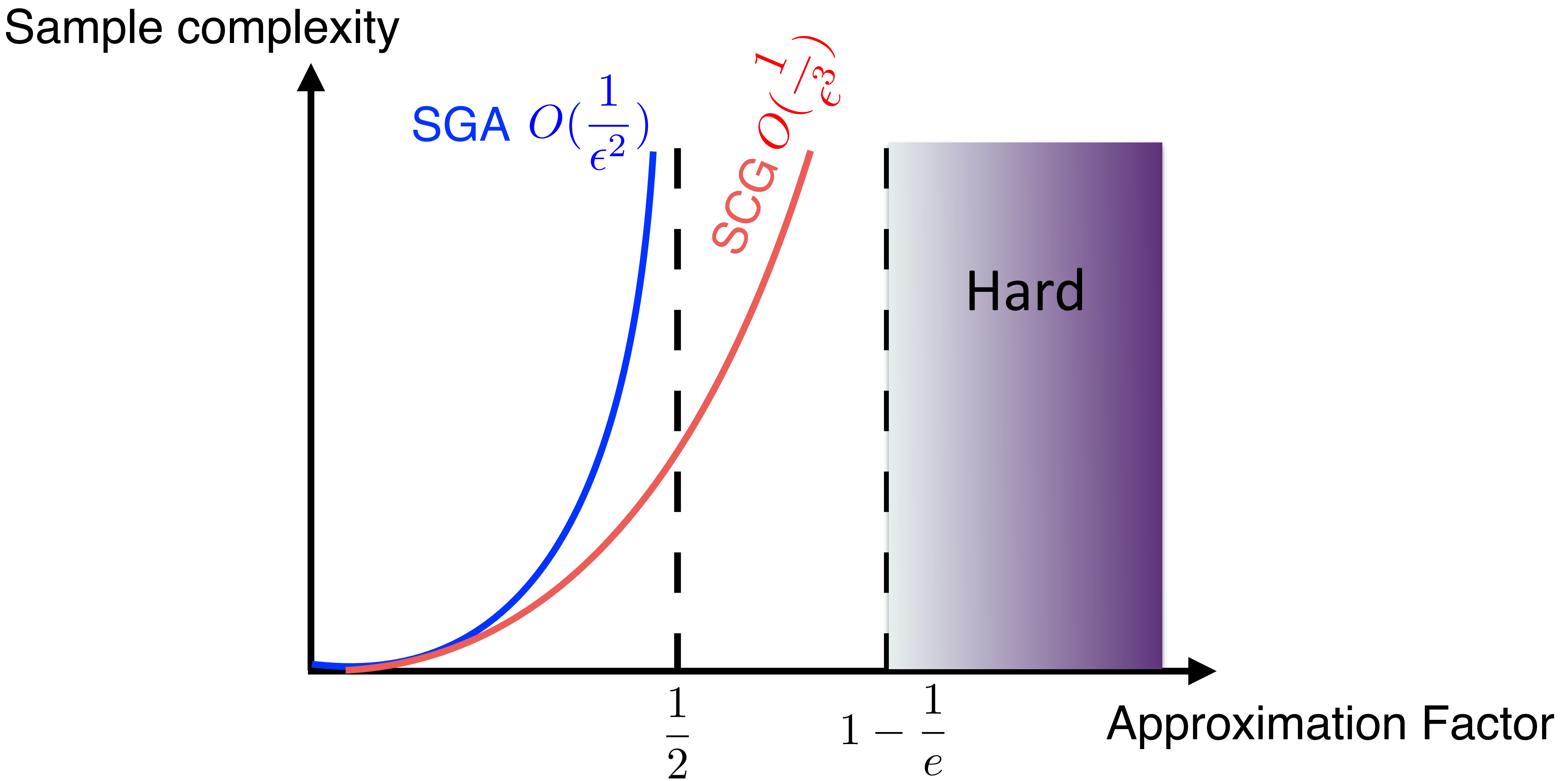
$$\mathbb{E}[F(\mathbf{x}_T)] \geq (1-\frac{1}{e})\mathrm{OPT} - \frac{15RQ^{\frac{1}{2}}}{T^{\frac{1}{3}}} - \frac{LR^2}{2T}$$

"Stochastic Conditional Gradient Methods: From Convex Minimization to Submodular Maximization", AISTATS '18, JMLR '20

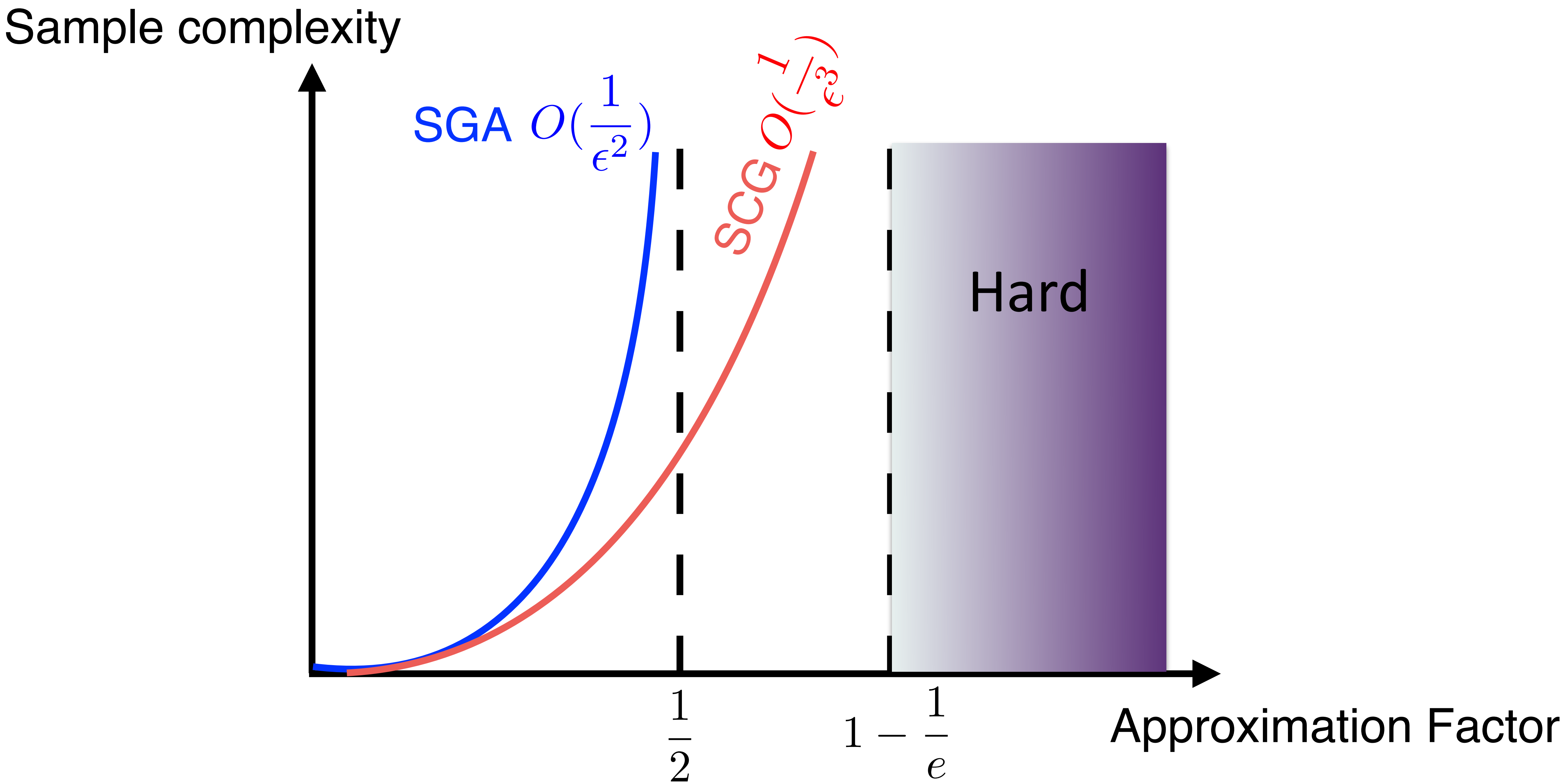
- To achieve $(1-\frac{1}{e})\mathrm{OPT} - \epsilon$, SCG requires $T = O(\frac{1}{\epsilon^3})$

Yale Penn

Maximizing DR-submodular Functions



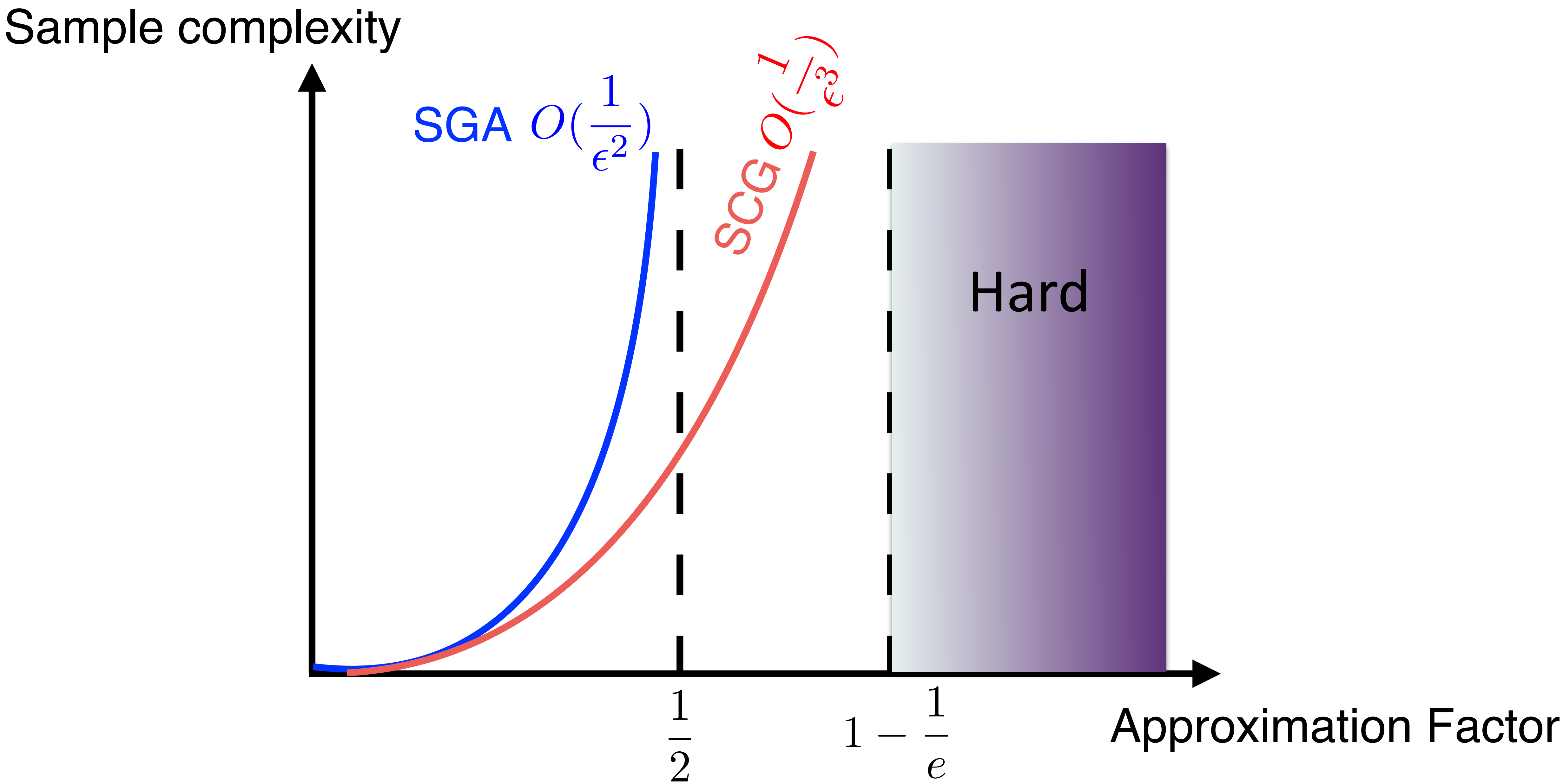
Maximizing DR-submodular Functions



- Optimal sample complexity of stochastic first order methods:

convex: $O(\frac{1}{\epsilon^2})$ non-convex: $O(\frac{1}{\epsilon^3})$

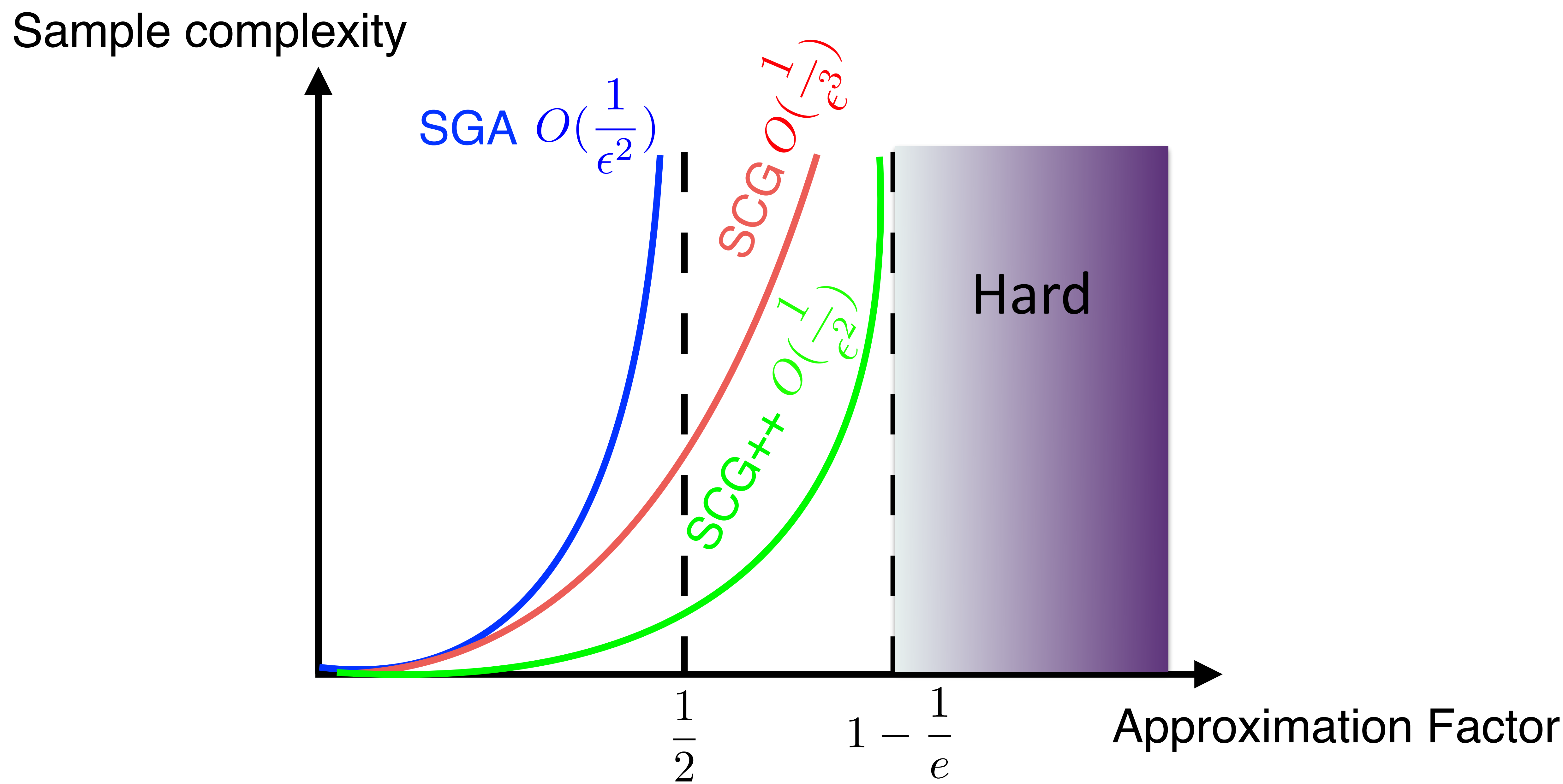
Maximizing DR-submodular Functions



[Hassani, Karbasi, Mokhtari, Shen]

Stochastic continuous greedy++ (SCG++) achieves $(1 - \frac{1}{e})\mathrm{OPT} - \epsilon$ with optimal sample complexity $O(\frac{1}{\epsilon^2})$.

Maximizing DR-submodular Functions



[Hassani, Karbasi, Mokhtari, Shen]

Stochastic continuous greedy++ (SCG++) achieves $(1-\frac{1}{e})\mathrm{OPT}-\epsilon$ with optimal sample complexity $O(\frac{1}{\epsilon^2})$.

Submodular Maximization

- We consider the following optimization problem: