Accelerated First-Order Methods
for Large-Scale Non-Negative Linear Programs

Lorenzo Orecchia
Boston University CS

Based on joint work with Zeyuan Allen-Zhu (Princeton & IAS) and Jelena Diakonikolas (BU CS).

Intro

- **Primal-Dual View of Accelerated Methods**
  - Limited Generality. We are not trying to explain all accelerated methods.
  - Goal is to synthesize important features and deploy them to problems that do not fit standard formulations

- **Applications**
  - Fast Approximate Solvers for Packing and Covering LPs (and SDPs)

- **Open problems:**
  - Connection to discretization methods
  - Application to implicit problems

A Primal-Dual View of Accelerated Methods

Convex Optimization in the Blackbox Model

COMPUTATIONAL MODEL: $\min_{x \in X} f(x)$

f convex, differentiable

$X \subseteq \mathrm{R}^n$ compact, convex set

$$x \in X \longrightarrow \boxed{\text{BLACKBOX}} \longrightarrow (f(x), \nabla f(x))$$

GOAL: minimize number of queries $x^{(1)}, x^{(2)}, \dots, x^{(t)}, \dots, x^{(T)}$ to obtain

$$f(x_{out}) \leq f(x^\star) + \epsilon$$

ANALYSIS: each algorithm must present

- A feasible solution x_{out} with an UPPER BOUND: $f(x_{out}) \leq B$ **PRIMAL SIDE**
- A LOWER BOUND to optimum: $f(x^*) \geq B - \epsilon$ **DUAL SIDE**

Primal-Dual Approach

ANALYSIS: each algorithm must present

- A feasible solution x_{out} with an UPPER BOUND: $f(x_{out}) \leq B$ **PRIMAL SIDE**
- A LOWER BOUND to optimum: $f(x^*) \geq B - \epsilon$ **DUAL SIDE**

SEPARATION OF CONCERNS: algorithm will iteratively maintain

$$x^{(1)}, x^{(2)}, \dots, x^{(t)} \longmapsto U_t, \quad U_t \geq f(x_{out})$$ **CURRENT UPPER BOUND**

$$x^{(1)}, x^{(2)}, \dots, x^{(t)} \longmapsto L_t, \quad f(x^*) \geq L_t$$ **CURRENT LOWER BOUND**

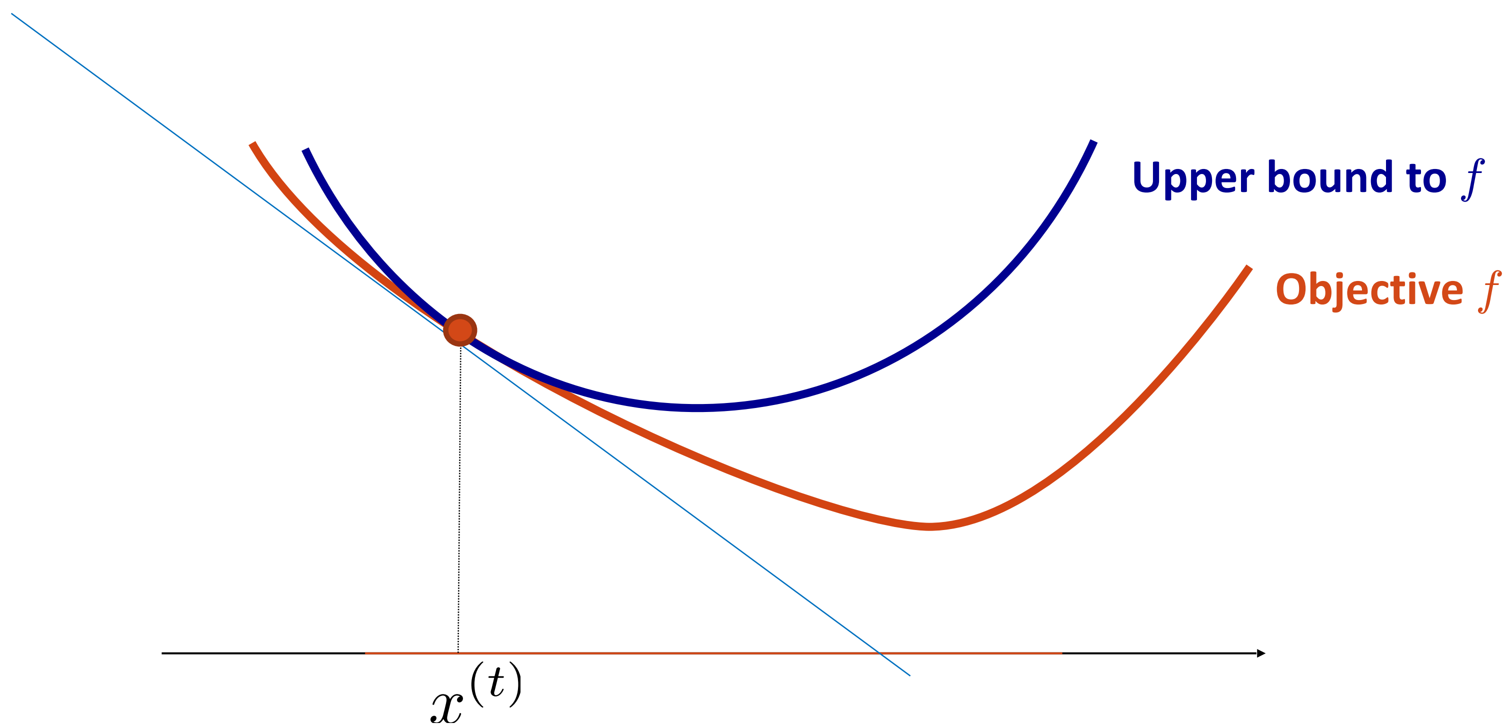
GOAL: after T iterations, duality gap is small

$$U_T - L_T \leq \epsilon$$

Smooth Functions: Primal Side

SMOOTHNESS CONDITION: $\forall x, y \in X, \quad \|\nabla f(y) - \nabla f(x)\|_* \leq L \cdot \|y - x\|$

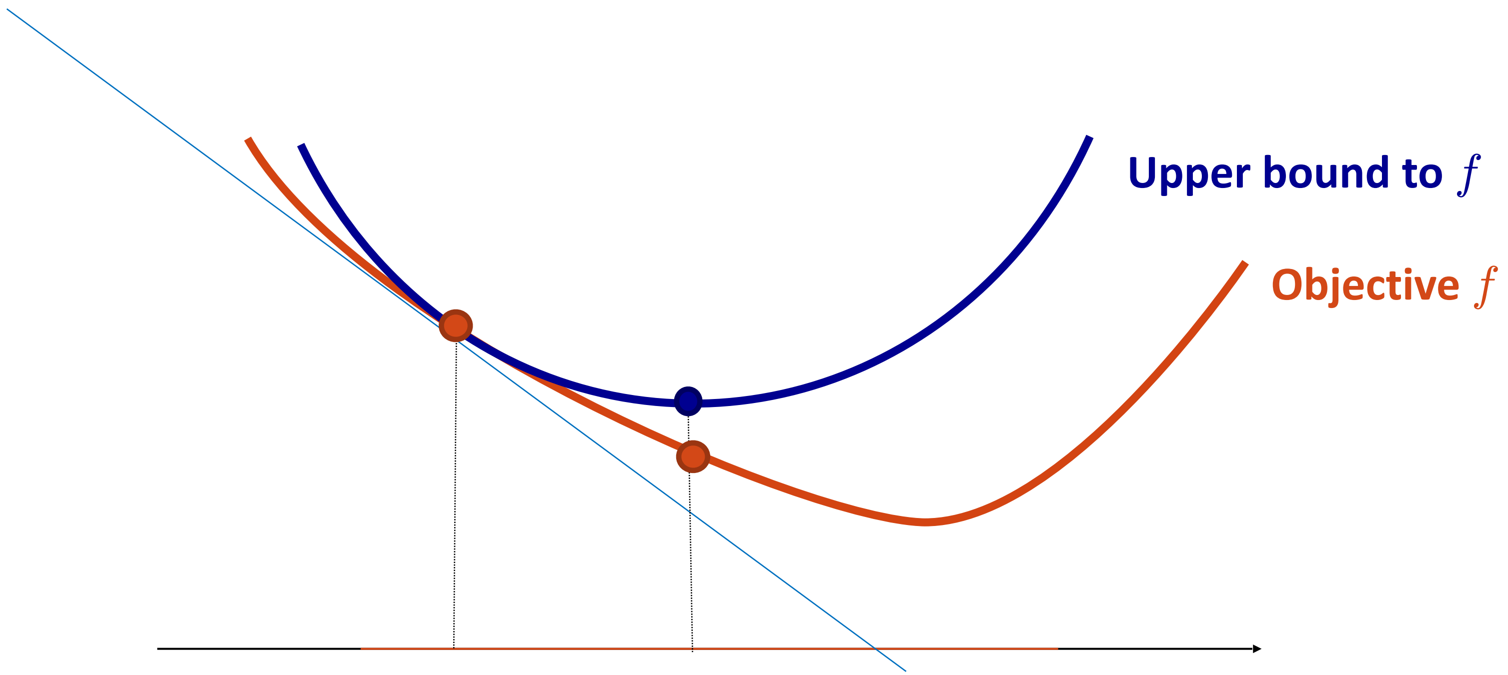
$$\forall x, y \in X, \quad f(y) \leq f(x) + \nabla f(x)^T (y - x) + \frac{L}{2} \cdot \|y - x\|^2$$



Smooth Functions: Primal Side

SMOOTHNESS CONDITION: $\forall x, y \in X, \quad \|\nabla f(y) - \nabla f(x)\|_* \leq L \cdot \|y - x\|$

$$\forall x, y \in X, \quad f(y) \leq f(x) + \nabla f(x)^T (y - x) + \frac{L}{2} \cdot \|y - x\|^2$$



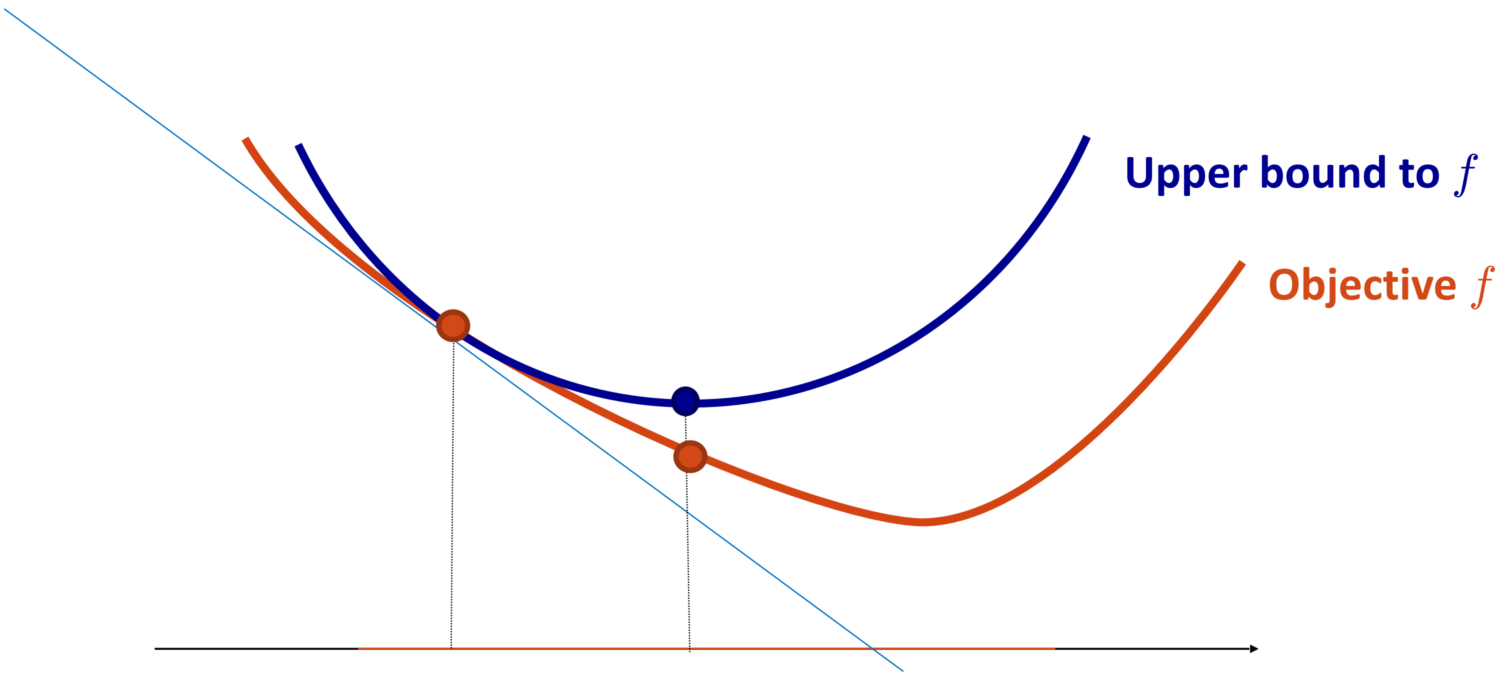
$$x^{(t)} \longrightarrow y^{(t)} = \mathrm{Grad}(x^{(t)});$$

UPPER BOUND: $U_t = f(y^{(t)})$

Smooth Functions: Primal Side

SMOOTHNESS CONDITION: $\forall x, y \in X, \quad \|\nabla f(y) - \nabla f(x)\|_* \leq L \cdot \|y - x\|$

$$\forall x, y \in X, \quad f(y) \leq f(x) + \nabla f(x)^T (y - x) + \frac{L}{2} \cdot \|y - x\|^2$$

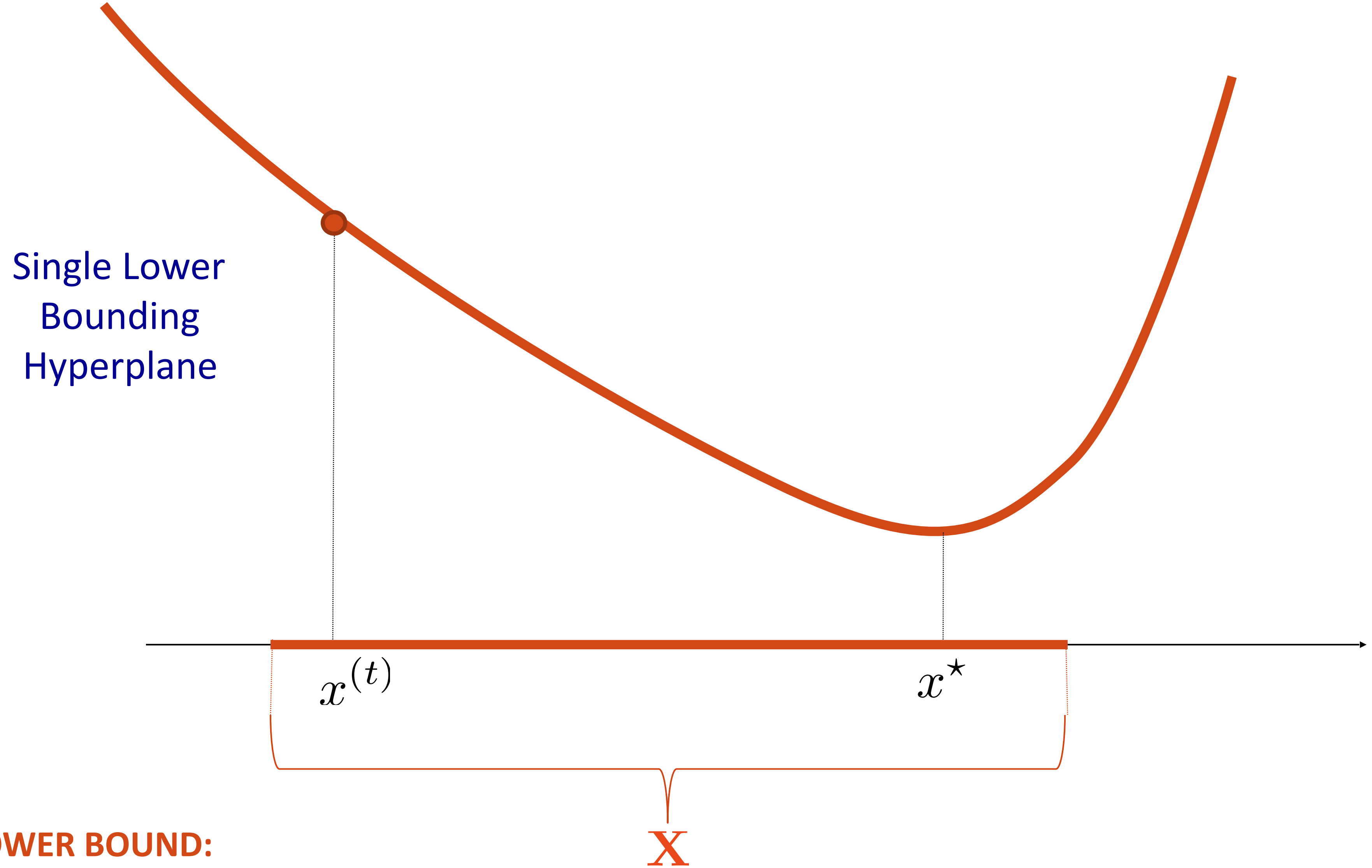


$$x^{(t)} \longrightarrow y^{(t)} = \mathrm{Grad}(x^{(t)});$$

UPPER BOUND: $U_t = f(y^{(t)}) \leq f(x^{(t)}) - \frac{\|\nabla f(x^{(t)})\|_*^2}{2L}$

for unconstrained problems

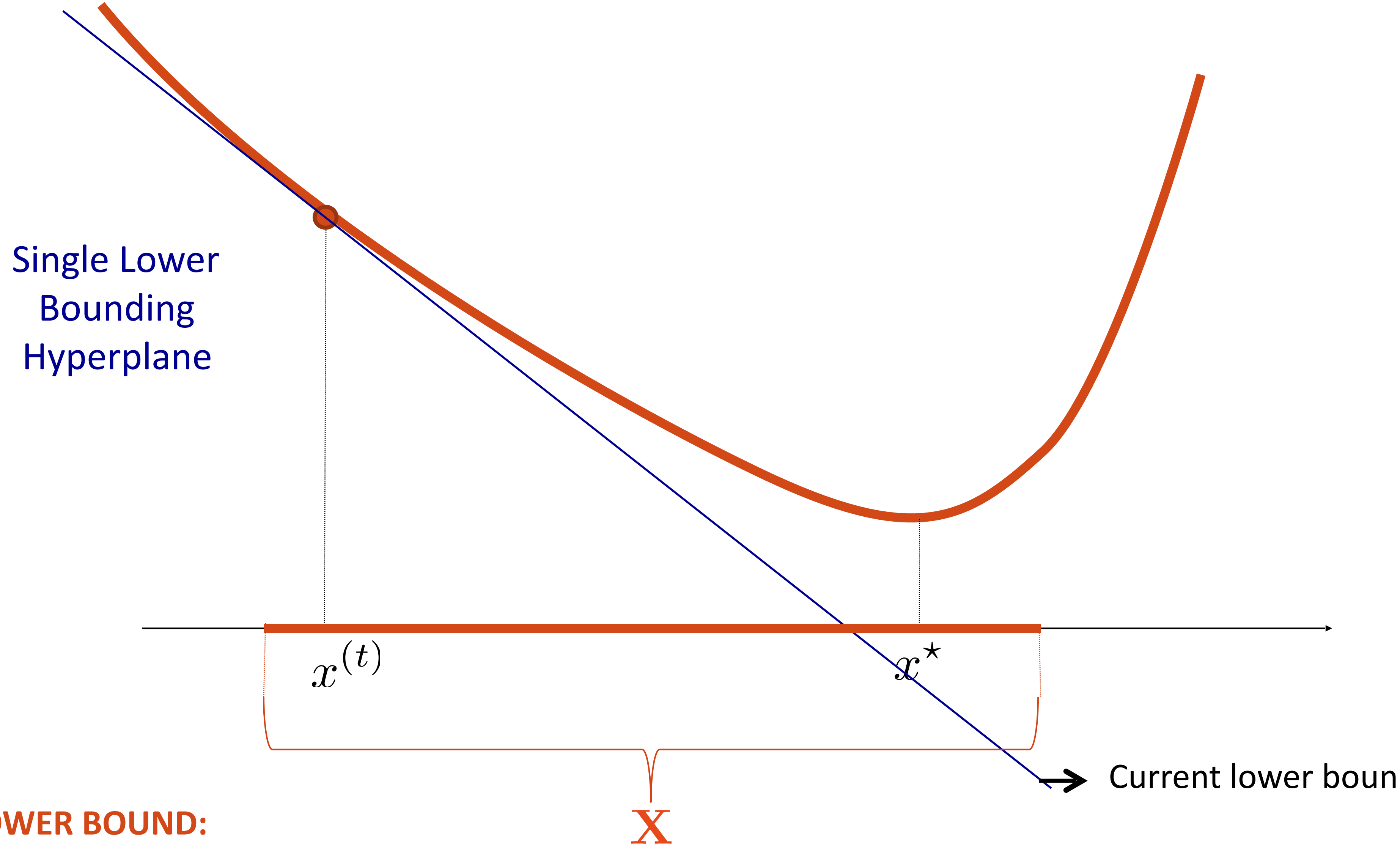
Dual Side



LOWER BOUND:

$$L_t = \min_{x \in X} f(x^{(t)}) + \langle \nabla f(x^{(t)}), x - x^{(t)} \rangle \geq f(x^{(t)}) - \|\nabla f(x^{(t)}\|_* \cdot \mathrm{diam}(X)$$

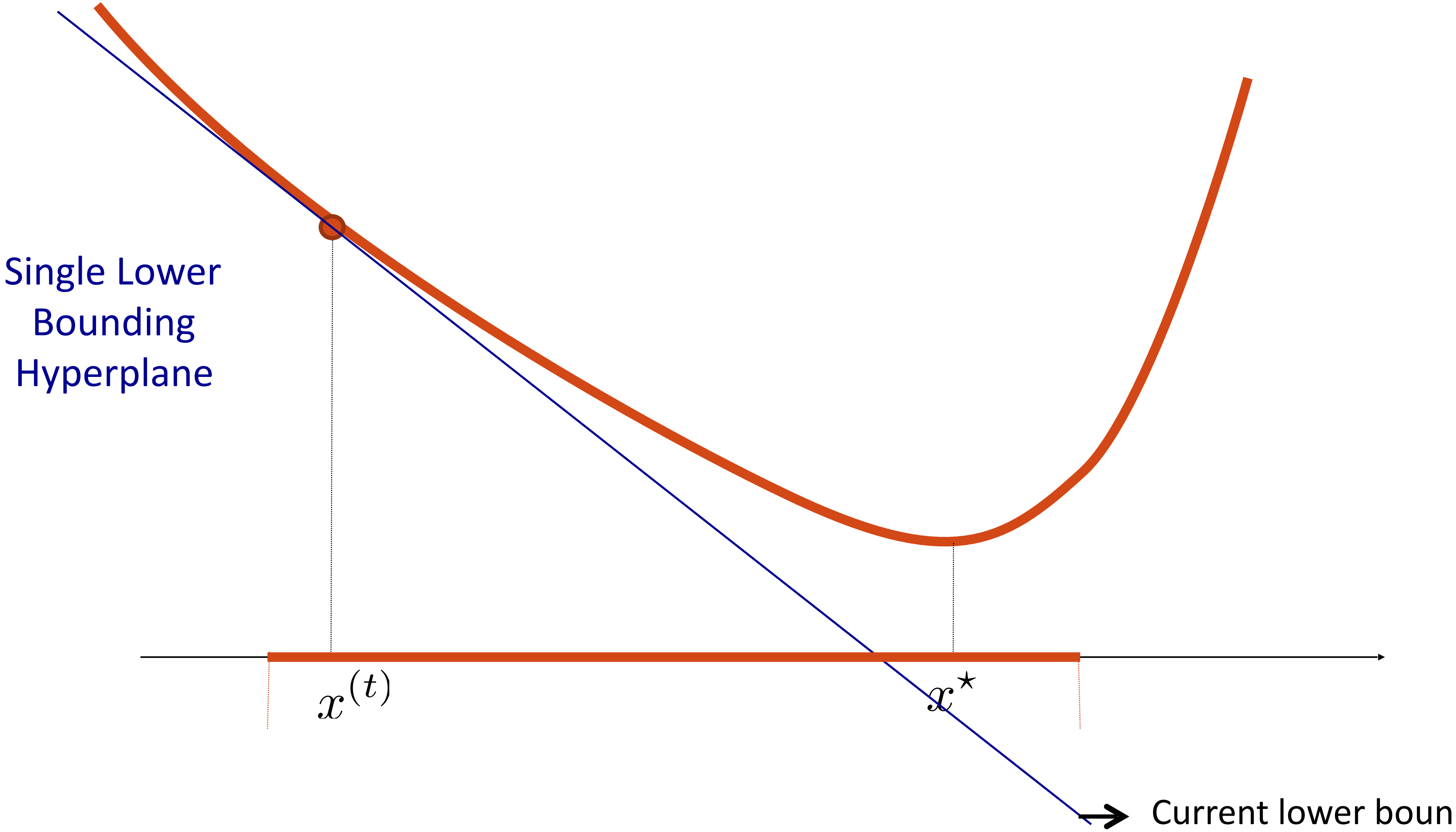
Dual Side



LOWER BOUND:

$$L_t = \min_{x \in X} f(x^{(t)}) + \langle \nabla f(x^{(t)}), x - x^{(t)} \rangle \geq f(x^{(t)}) - \|\nabla f(x^{(t)}\|_* \cdot \mathrm{diam}(X)$$

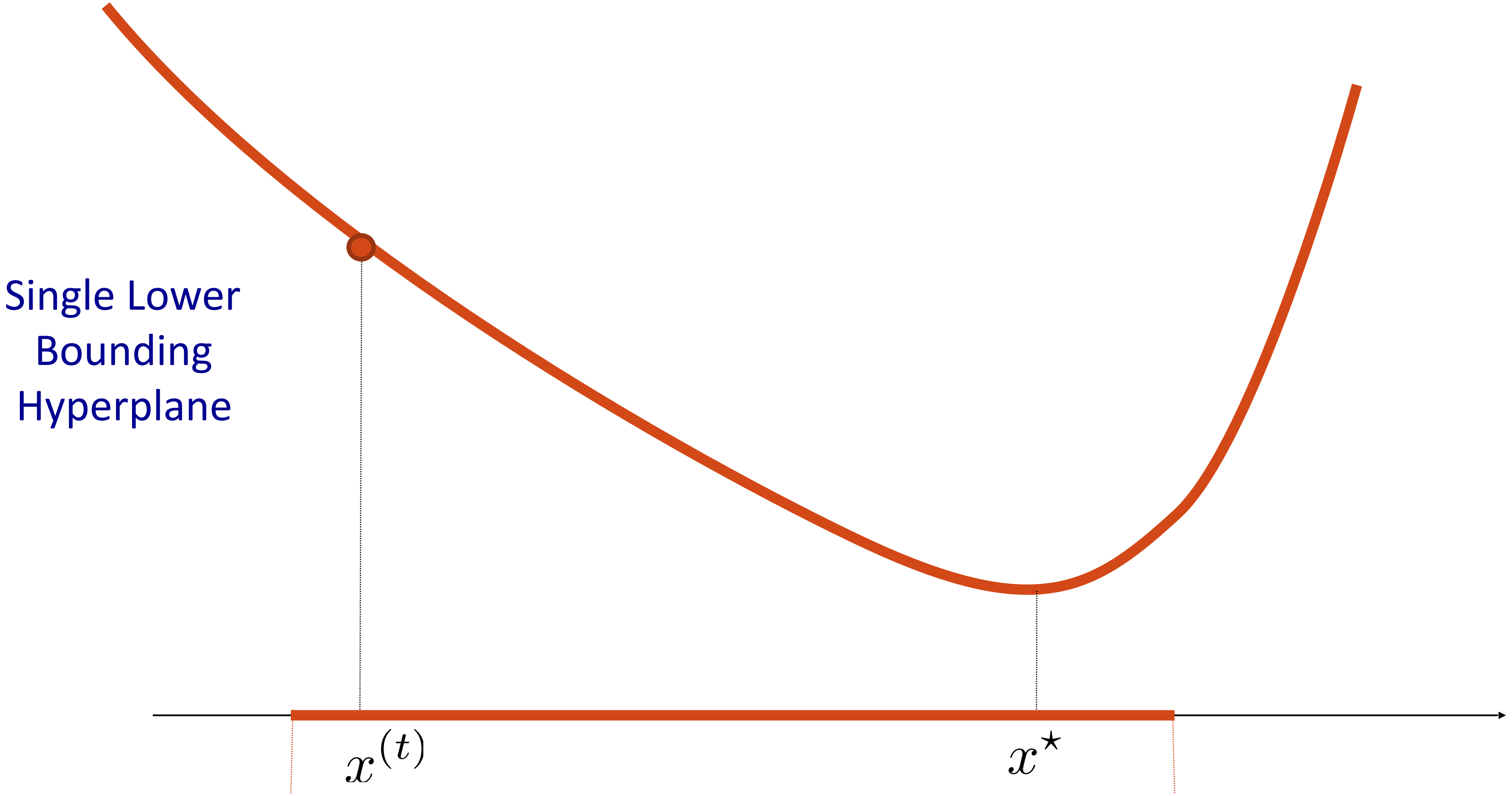
Dual Side



LOWER BOUND:

$$L_t = \min_{x \in X} f(x^{(t)}) + \langle \nabla f(x^{(t)}), x - x^{(t)} \rangle \geq f(x^{(t)}) - \|\nabla f(x^{(t)}\|_* \cdot \mathrm{diam}(X)$$

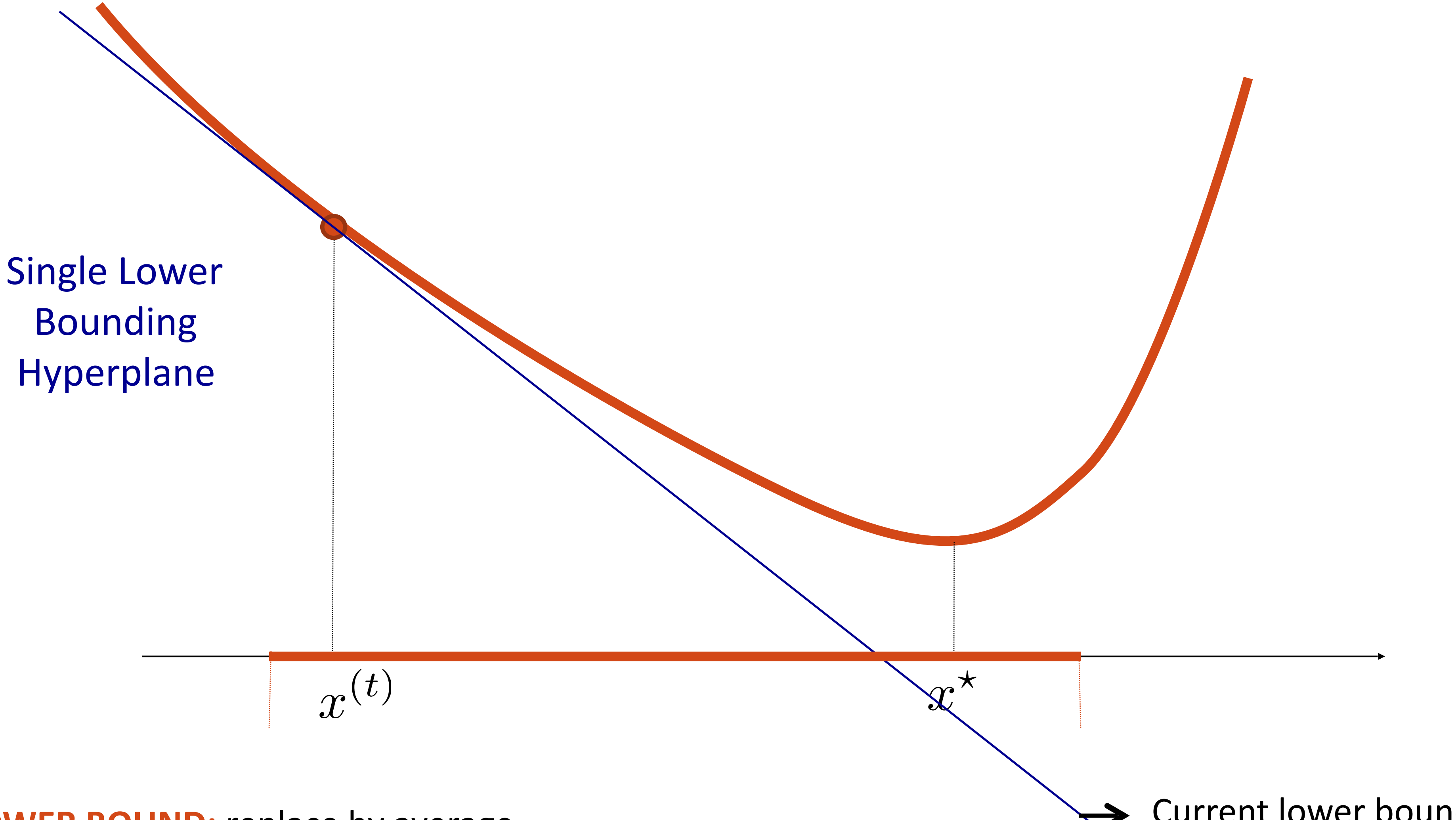
Dual Side



LOWER BOUND: replace by average → Current lower boun

$$L_t \geq \frac{1}{t}\left[\sum_{i=1}^{t} f(x^{(t)}) - \|\nabla f(x^{(t)})\|_* \cdot \mathrm{diam}(X)\right]$$

Dual Side



LOWER BOUND: replace by average

$$L_t \geq \frac{1}{t}\left[\sum_{i=1}^{t} f(x^{(t)}) - \|\nabla f(x^{(t)})\|_* \cdot \mathrm{diam}(X)\right]$$

Gradient Descent for Smooth Functions

UPPER BOUND: $U_t = f(y^{(t)})$

LOWER BOUND: $L_t \geq \frac{1}{t}\left[\sum_{i=1}^{t} f(x^{(t)}) - \|\nabla f(x^{(t)}\|_* \cdot \mathrm{diam}(X)\right]$

Gradient Descent for Smooth Functions

UPPER BOUND: $$U_t = f(y^{(t)}) \leq f(x^{(t)}) - \frac{\|\nabla f(x^{(t)})\|_*^2}{2L}$$

for unconstrained problems

LOWER BOUND: $$L_t \geq \frac{1}{t}\left[\sum_{i=1}^{t} f(x^{(t)}) - \|\nabla f(x^{(t)}\|_* \cdot \mathrm{diam}(X)\right]$$

Gradient Descent for Smooth Functions

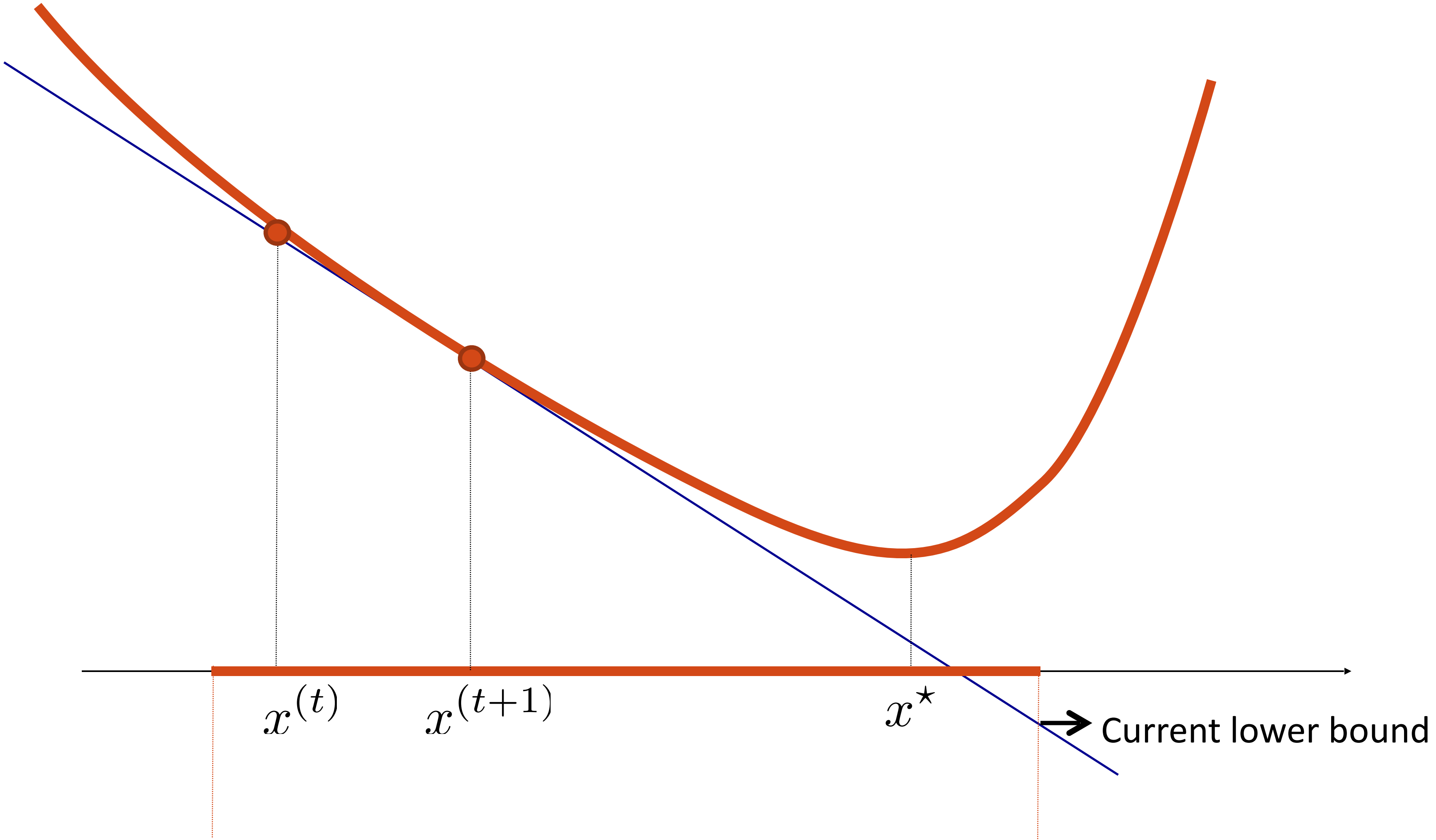
UPPER BOUND: $$U_t = f(y^{(t)}) \leq f(x^{(t)}) - \frac{\|\nabla f(x^{(t)})\|_*^2}{2L}$$ for unconstrained problems

LOWER BOUND: $$L_t \geq \frac{1}{t}\left[\sum_{i=1}^{t} f(x^{(t)}) - \|\nabla f(x^{(t)}\|_* \cdot \mathrm{diam}(X)\right]$$

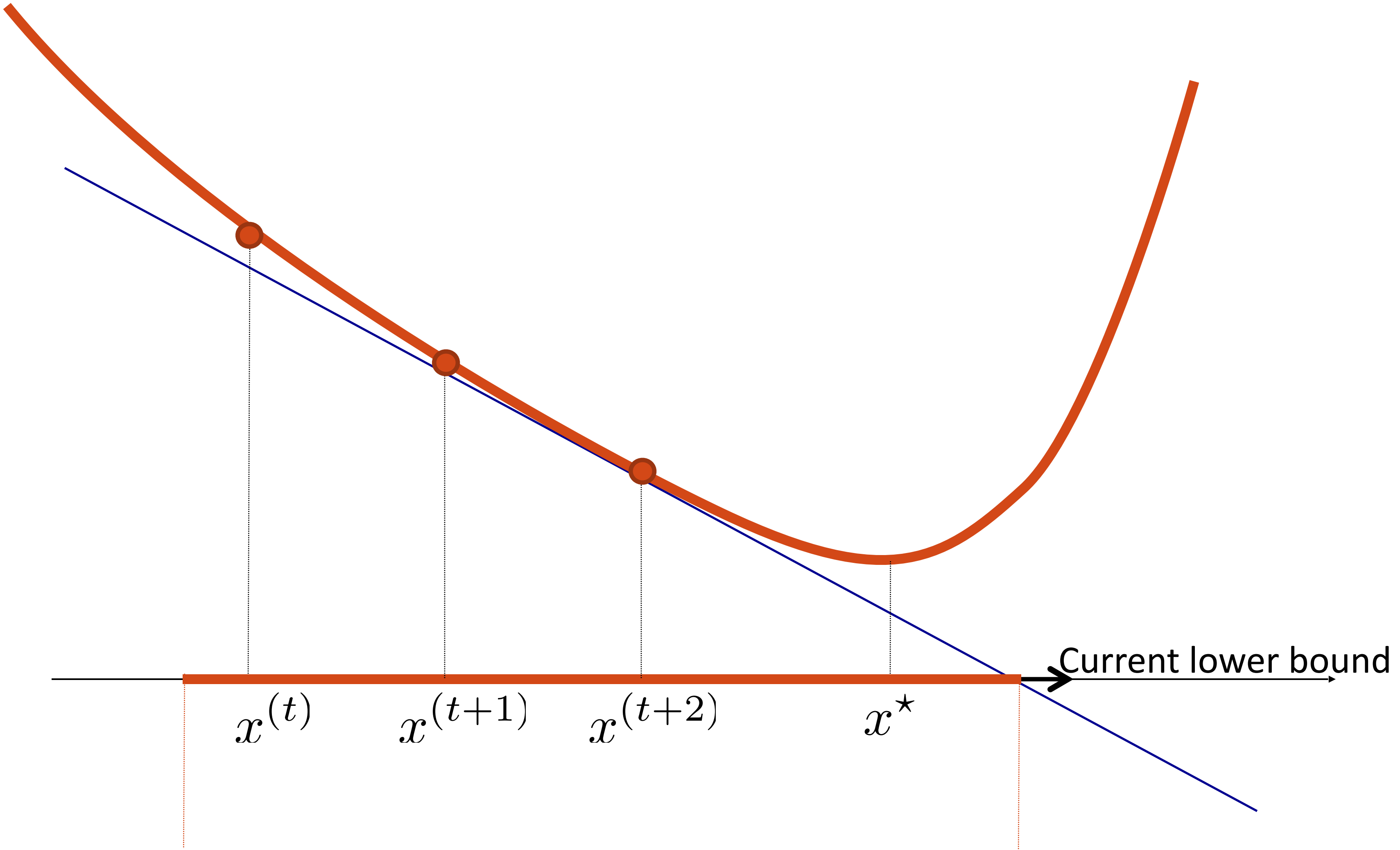
GRADIENT DESCENT STEP: $$x^{(t+1)} = y^{(t)} = \mathrm{Grad}(x^{(t)})$$

DUALITY GAP: $$U_{t+1} - L_{t+1} = \frac{t}{t+1} \cdot (U_t - L_t) = \frac{(U_0 - L_0)}{t+1}$$

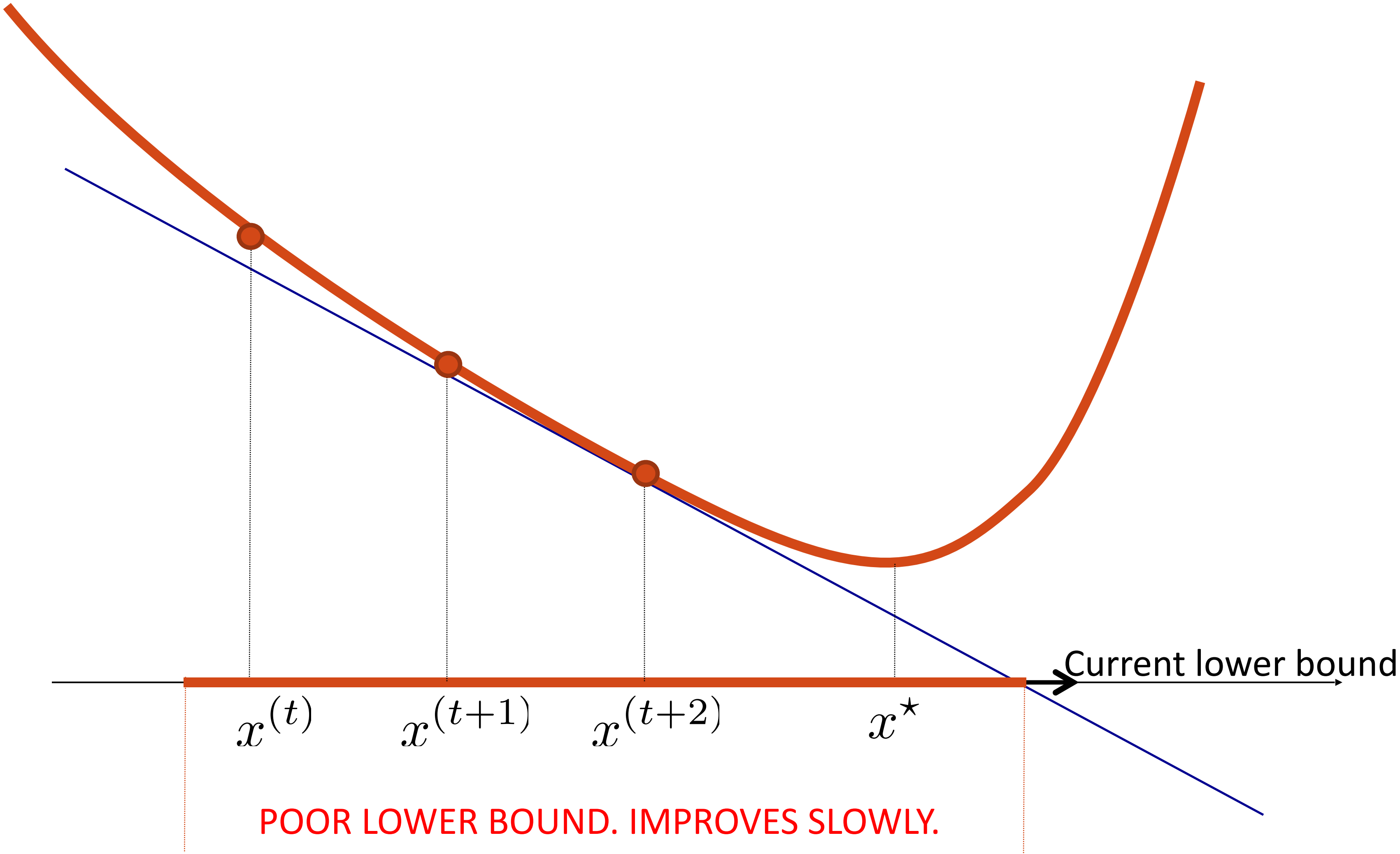
Limitations of Gradient Descent



Limitations of Gradient Descent



Limitations of Gradient Descent

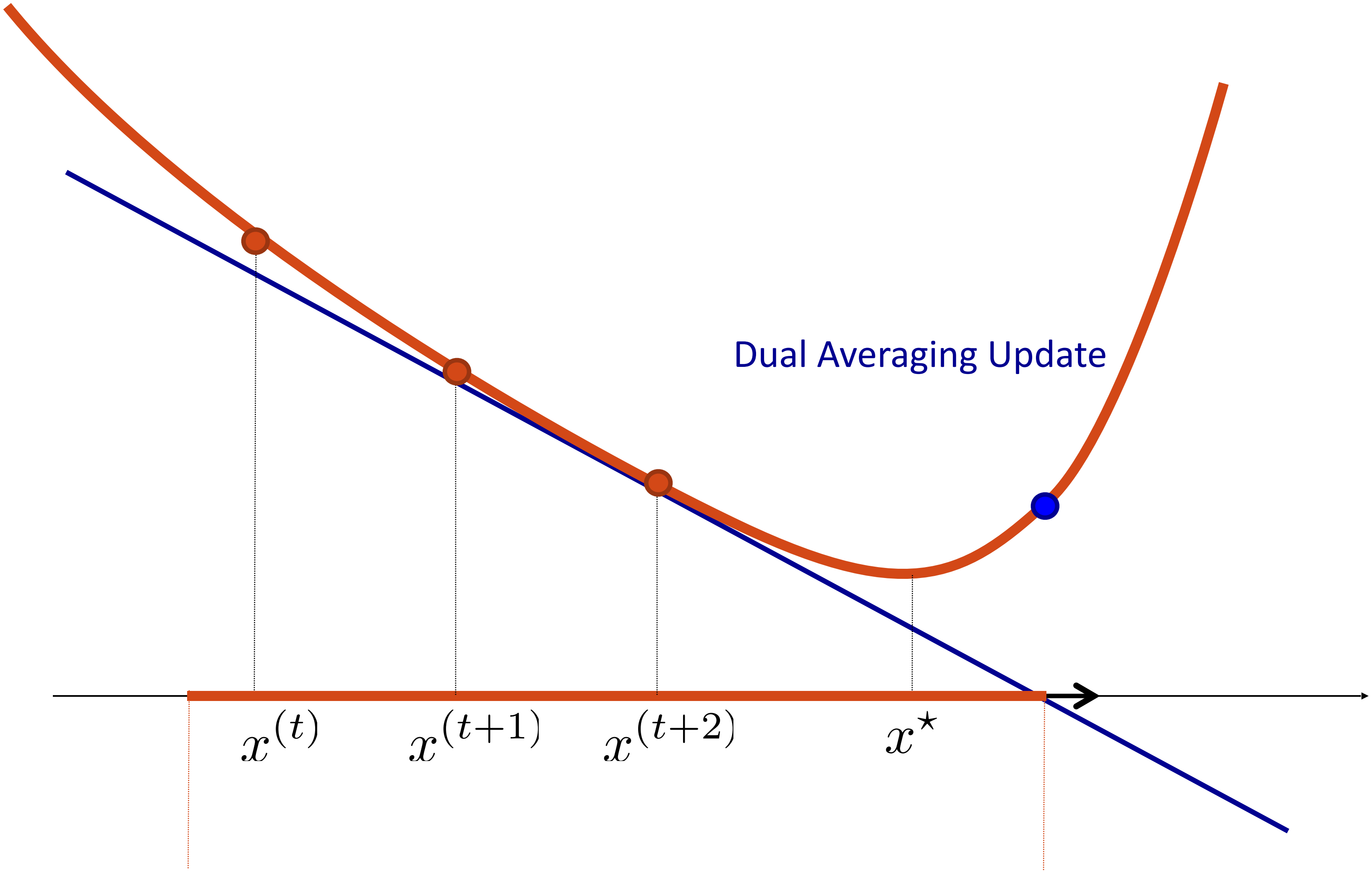


Limitations of Gradient Descent

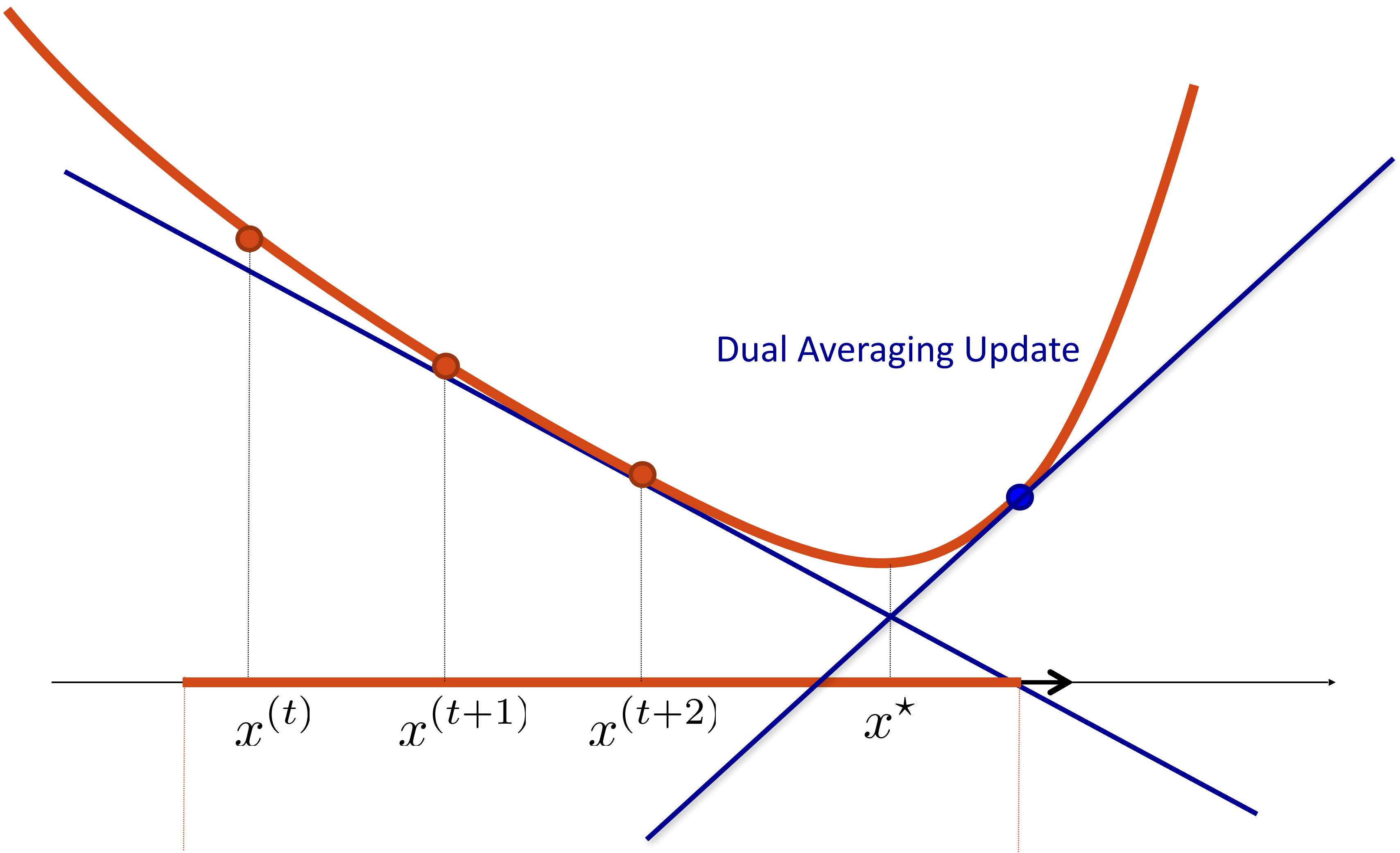
- Gradient Descent **attempts to construct best possible upper bound**

- Gradient Descent **does NOT attempt to construct a good lower bound**.
  It uses a single lower bounding hyperplane.

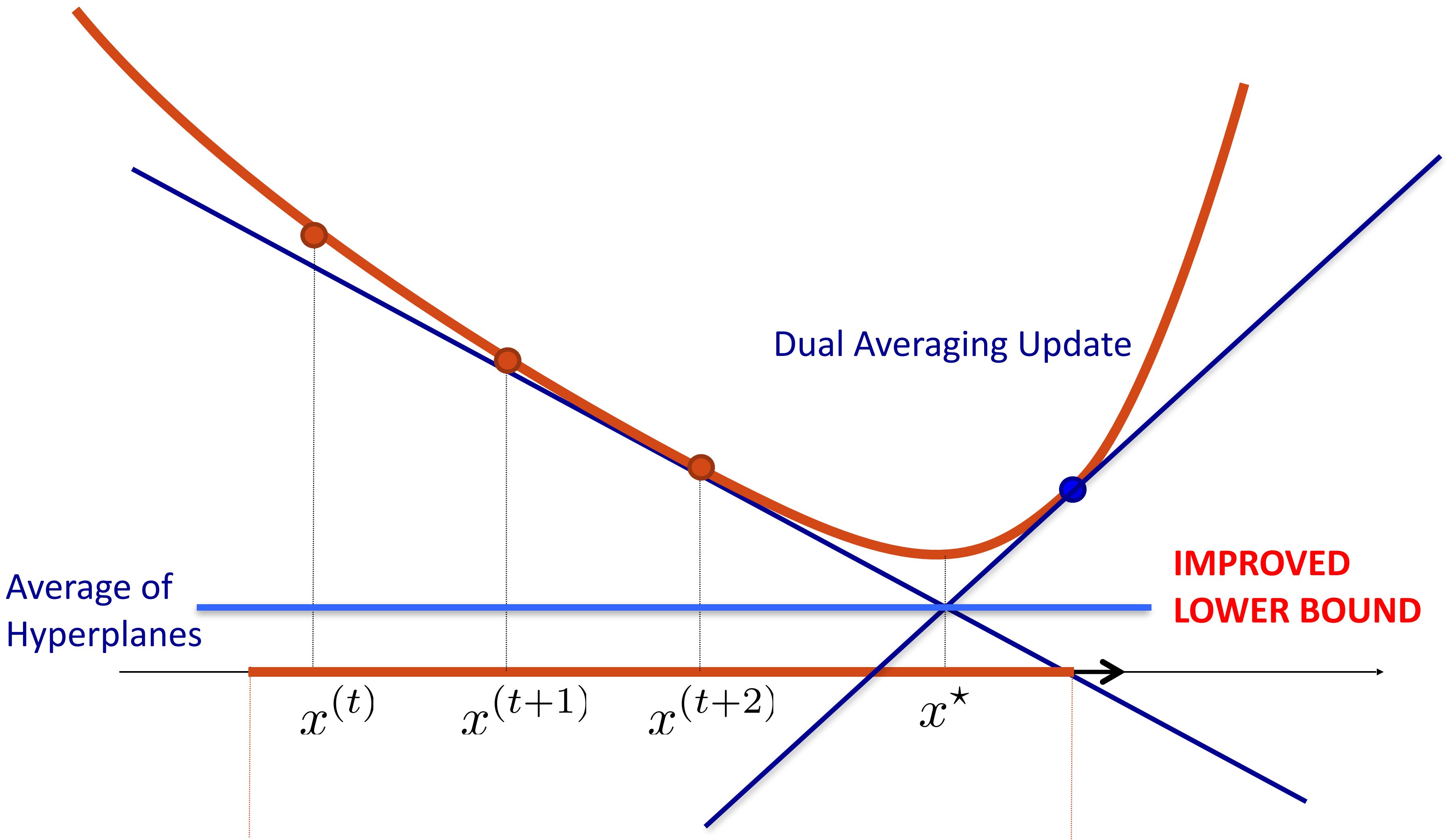
Dual Strategy: Conceptual Example



Dual Strategy: Conceptual Example

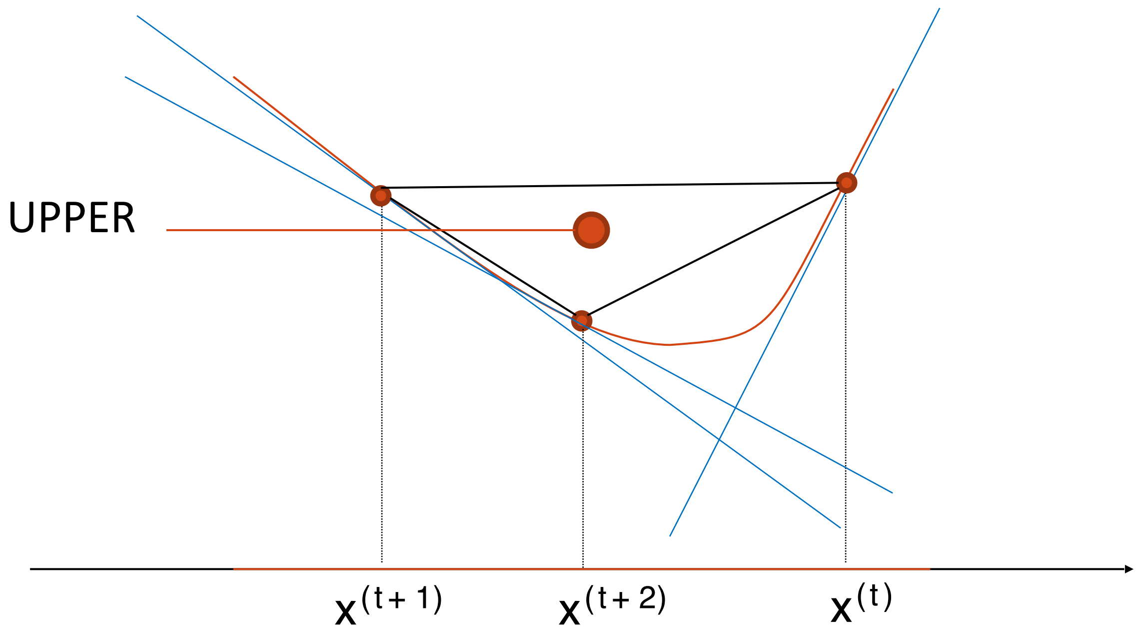


Dual Strategy: Conceptual Example



Non-Smooth Functions: Primal Side

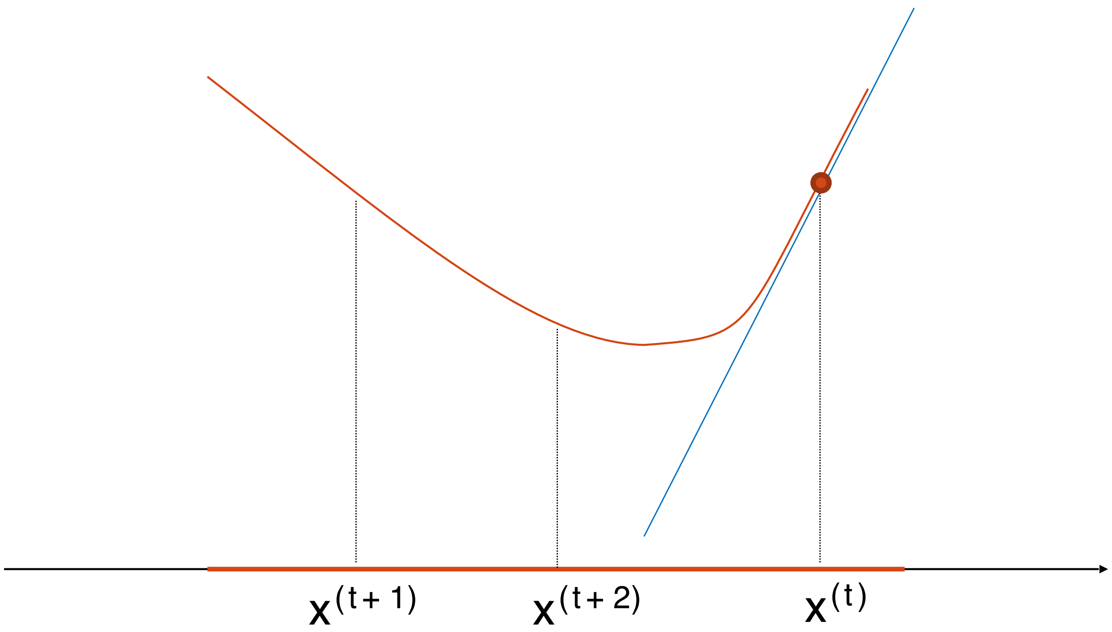
ASSUMPTION: f convex, differentiable, ρ-Lipschitz $\forall x \in X,\ \|\nabla f(x)\|_* \leq \rho$



UPPER BOUND: $$U_T = \frac{1}{T}\left(\sum_{t=1}^{T} f(x^{(t)})\right) \geq f(x^*)$$

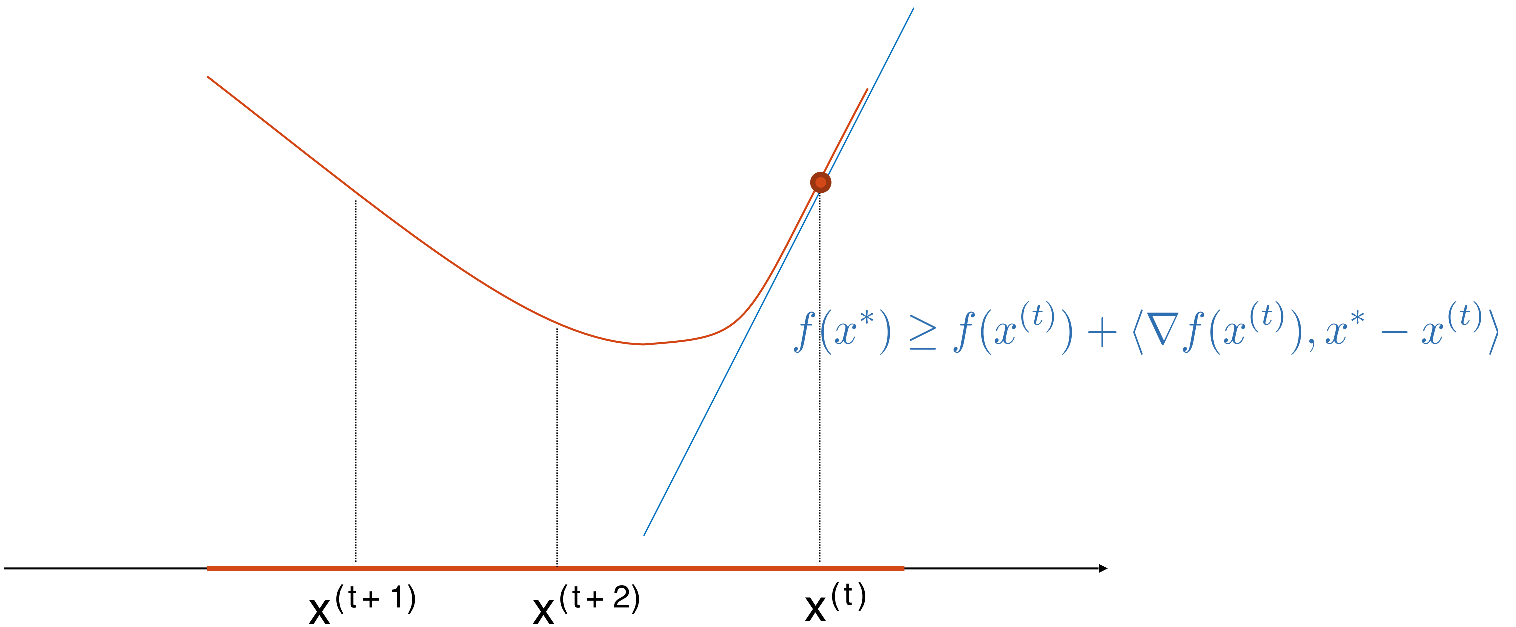
Non-Smooth Functions: Dual Side

ASSUMPTION: f convex, differentiable, ρ-Lipschitz $\forall x \in X,\ \|\nabla f(x)\|_* \leq \rho$



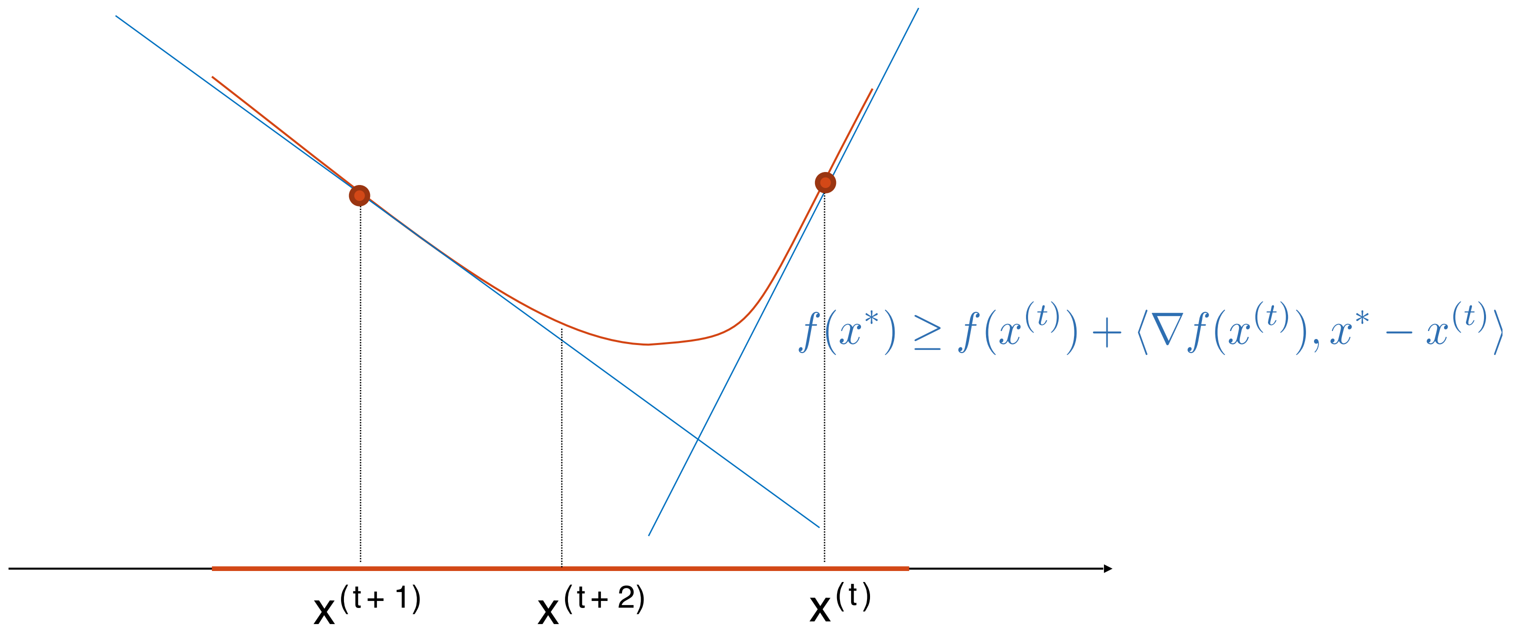
Non-Smooth Functions: Dual Side

ASSUMPTION: f convex, differentiable, ρ-Lipschitz $\forall x \in X,\ \|\nabla f(x)\|_* \leq \rho$



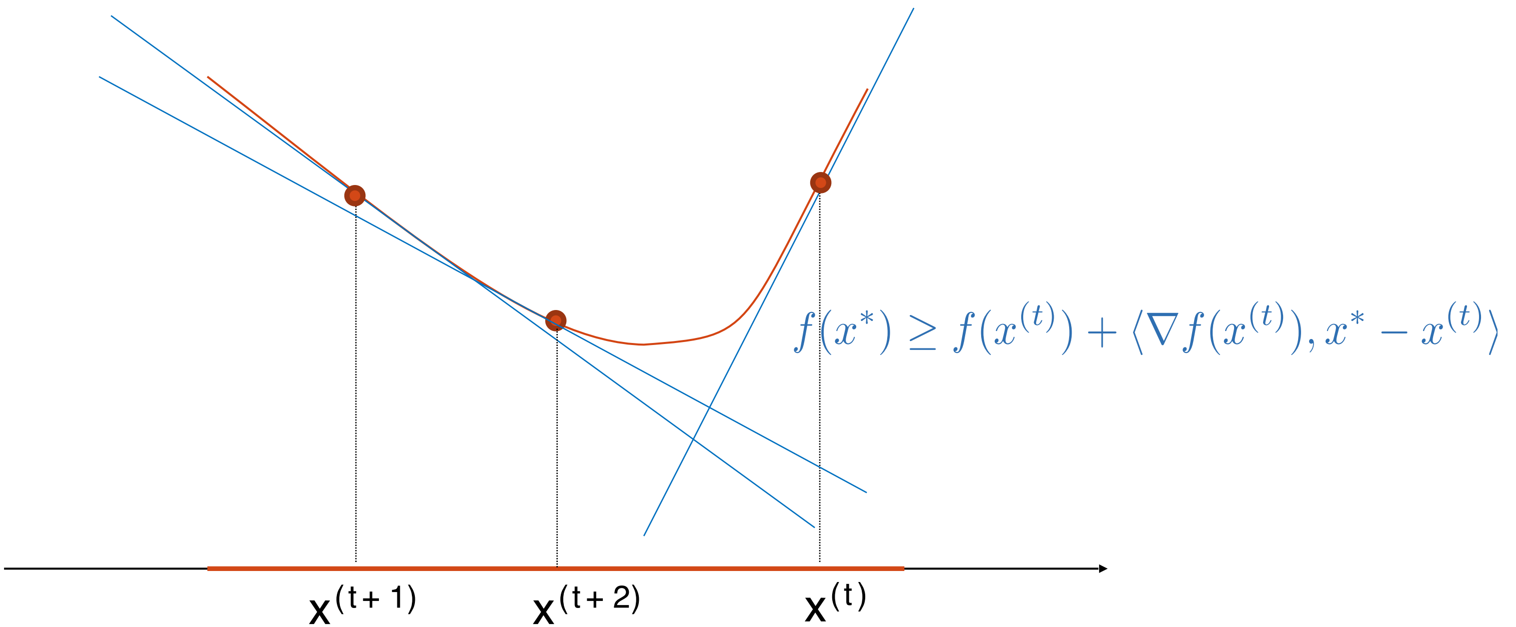
Non-Smooth Functions: Dual Side

ASSUMPTION: f convex, differentiable, ρ-Lipschitz $\forall x \in X,\ \|\nabla f(x)\|_* \leq \rho$



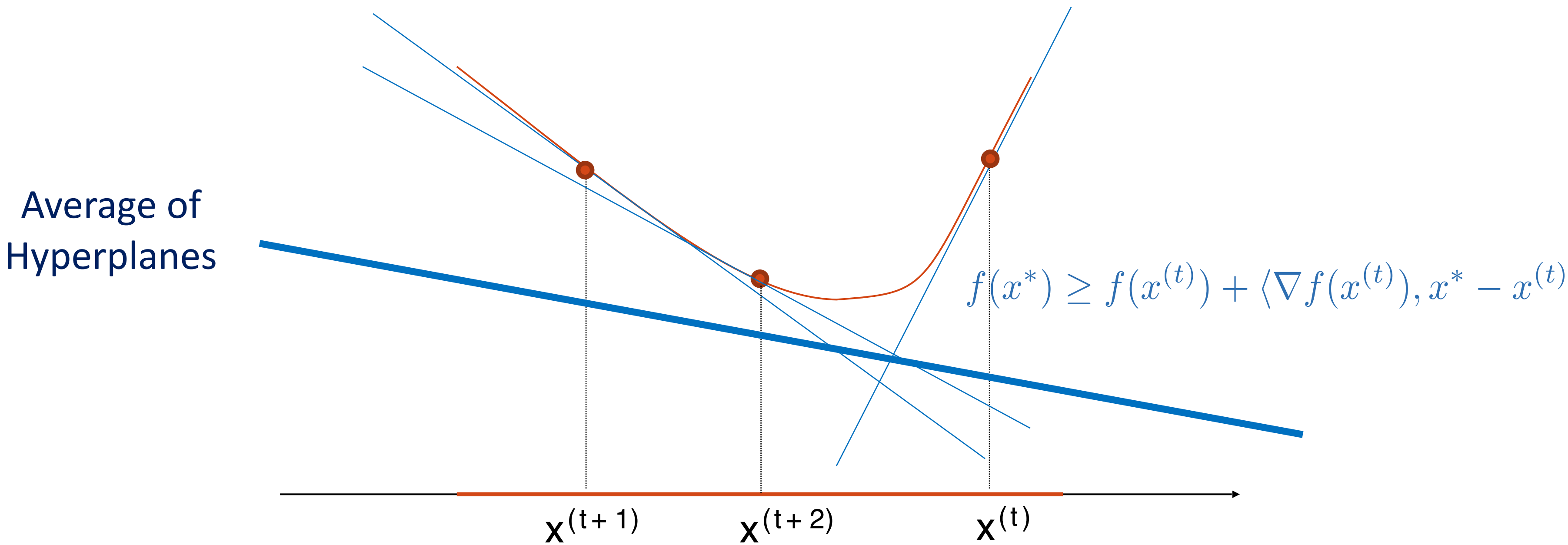
Non-Smooth Functions: Dual Side

ASSUMPTION: f convex, differentiable, ρ-Lipschitz $\forall x \in X, \ \|\nabla f(x)\|_* \leq \rho$



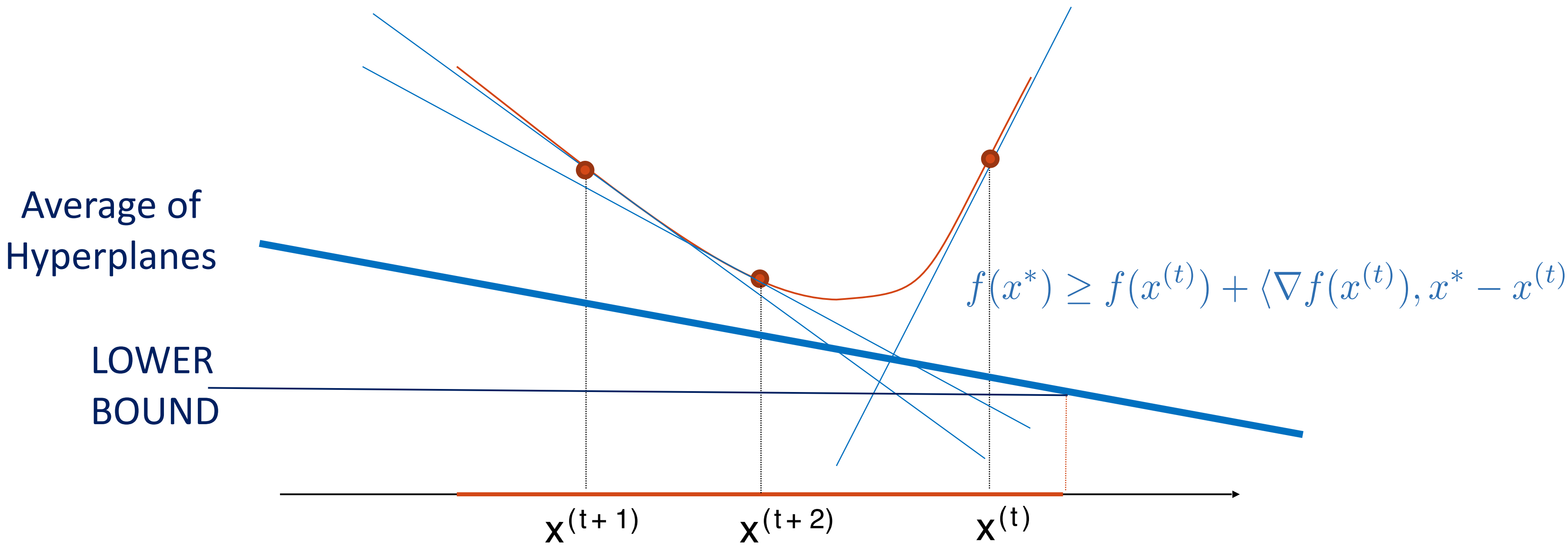
Non-Smooth Functions: Dual Side

ASSUMPTION: f convex, differentiable, ρ-Lipschitz $\forall x \in X, \; \|\nabla f(x)\|_* \leq \rho$



Non-Smooth Functions: Dual Side

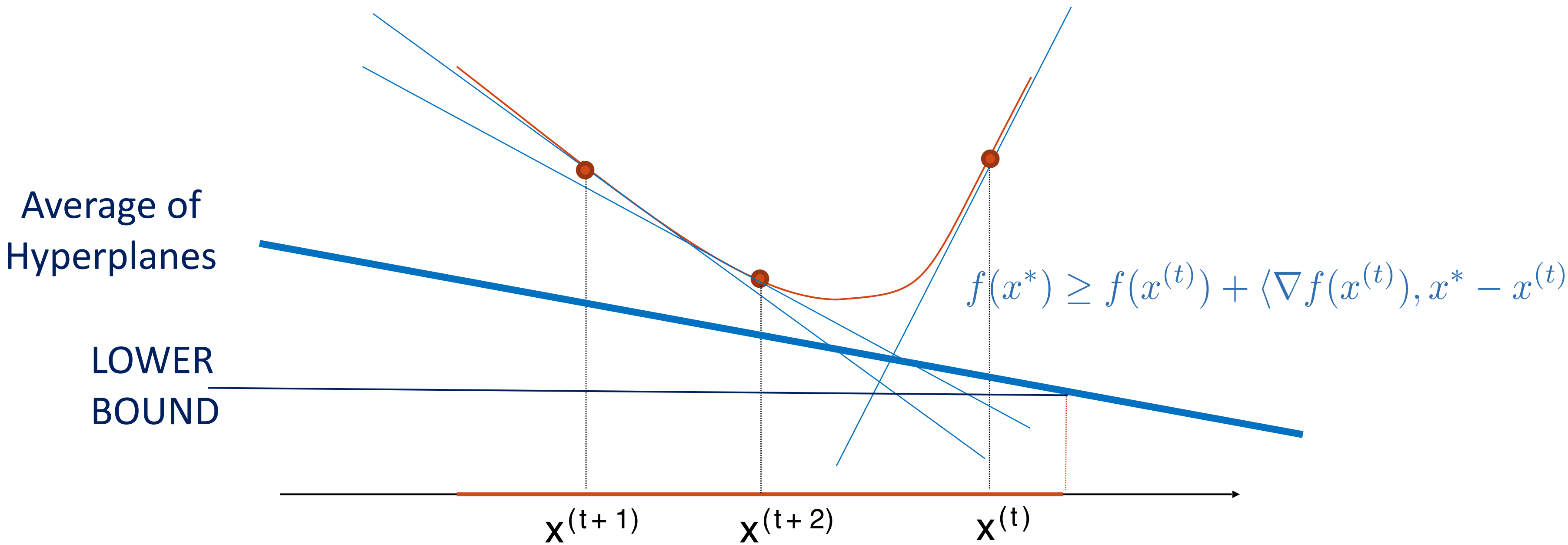
ASSUMPTION: f convex, differentiable, ρ-Lipschitz $\forall x \in X,\ \|\nabla f(x)\|_* \leq \rho$



$$f(x^*) \geq \frac{1}{T} \min_{x \in X} \left[\sum_{t=1}^{T} f(x^{(t)}) + \langle \nabla f(x^{(t)}), (x - x^{(t)}) \rangle \right]$$

Non-Smooth Functions: Dual Side

ASSUMPTION: f convex, differentiable, ρ-Lipschitz $\forall x \in X,\ \|\nabla f(x)\|_* \leq \rho$



$$f(x^*) \geq \frac{1}{T} \min_{x \in X} \left[\sum_{t=1}^{T} f(x^{(t)}) + \langle \nabla f(x^{(t)}), (x - x^{(t)}) \rangle \right]$$

LOWER BOUND: $$L_t = \frac{1}{T} \left[\sum_{t=1}^{T} f(x^{(t)}) + \min_{x \in X} \langle \sum_{t=1}^{T} \nabla f(x^{(t)}), (x - x^{(t)}) \rangle \right]$$

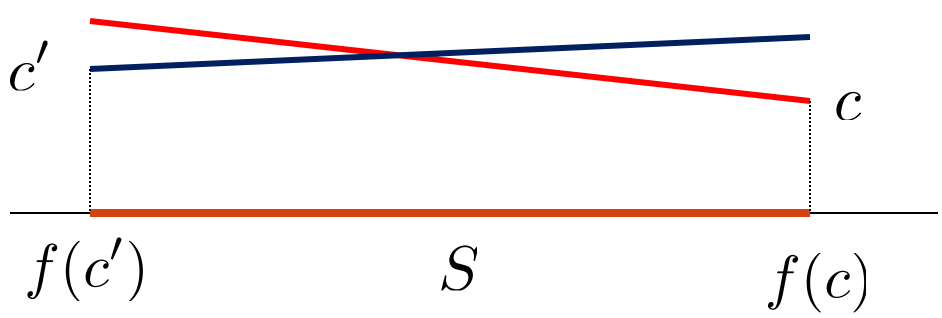
Smoothing by Regularization

Consider a convex set $S \subseteq R^n$ and a linear optimization problem:

$$f(c) = \arg\min_{x \in S} c^T x.$$

The optimal solution $f(c)$ may be very unstable under perturbation of c :

$$\|c' - c\| \leq \delta \quad \text{and} \quad \|f(c') - f(c)\| >> \delta$$



Example: Regularization Helps Stability

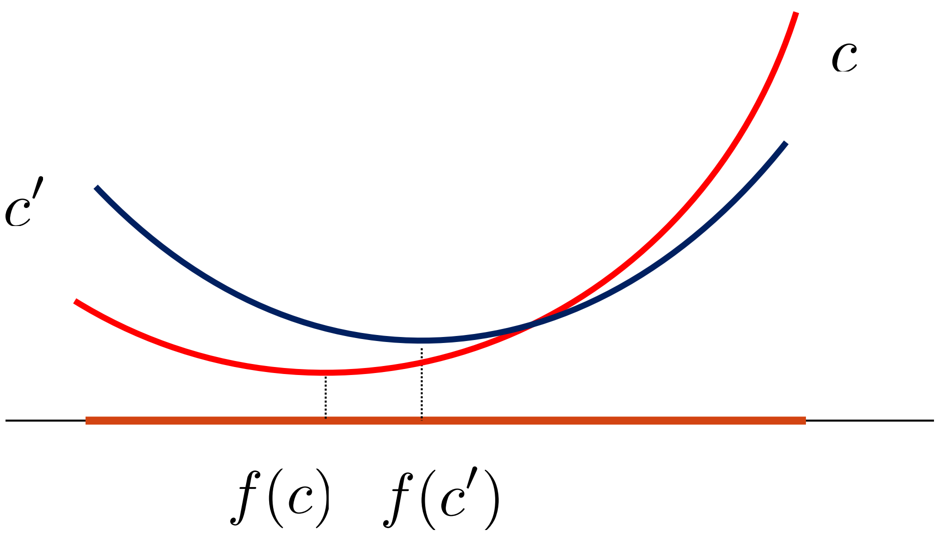
Consider a convex set $S \subseteq R^n$ and a **regularized** linear optimization problem

$$f(c) = \arg\min_{x \in S} c^T x + F(x)$$

where F is σ-strongly convex.

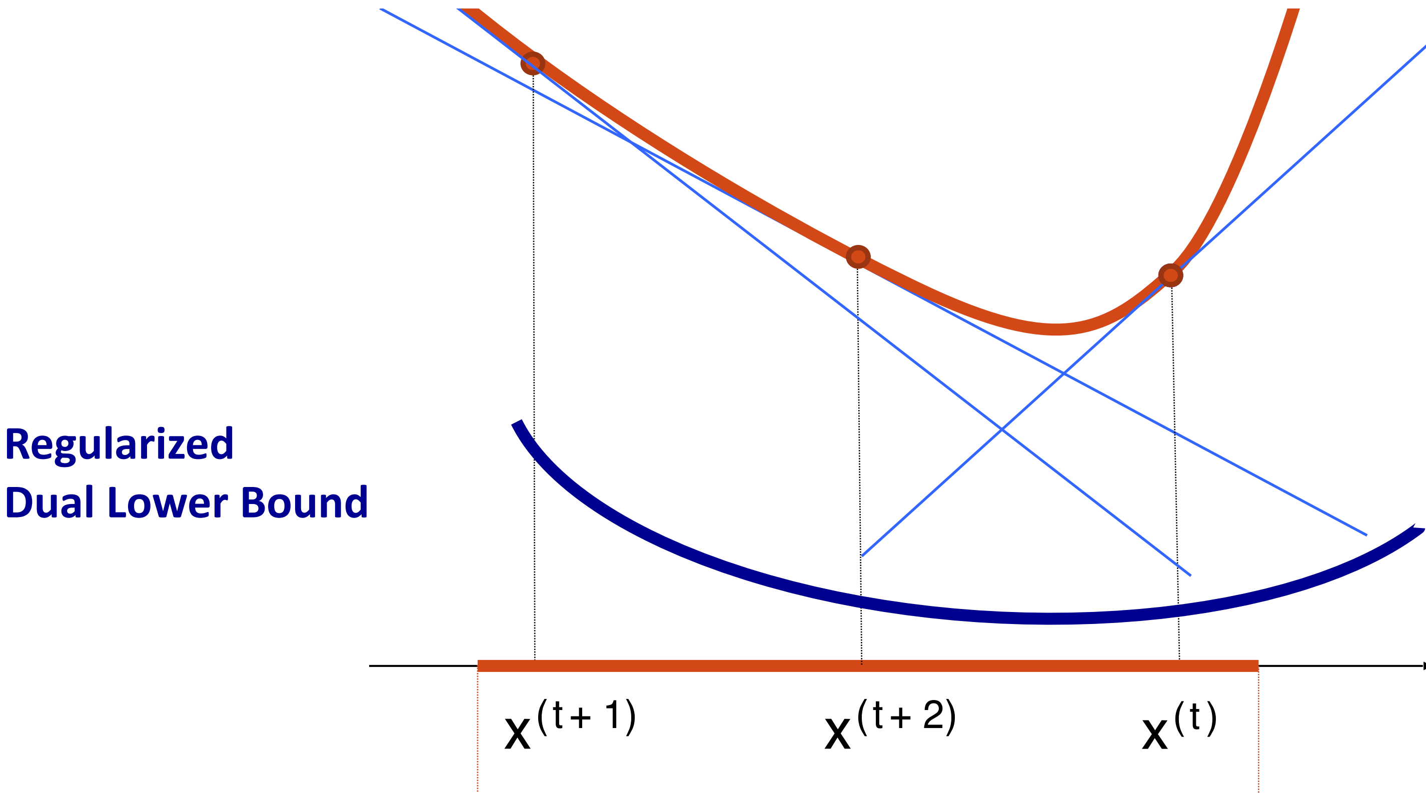
$\|c' - c\| \leq \delta$ implies $\|f(c') - f(c)\| >> \delta$

Then:



Non-Smooth Functions: Dual Side

ASSUMPTION: f convex, differentiable, ρ-Lipschitz $\forall x \in X,\ \|\nabla f(x)\|_* \leq \rho$



$$f(x^*) \geq \frac{1}{T}\left[\sum_{t=1}^{T} f(x^{(t)}) + \min_{x\in X}\langle \sum_{t=1}^{T} \nabla f(x^{(t)}), (x - x^{(t)})\rangle + \boxed{F(x)}\right]$$

Dual Averaging for Non-Smooth Functions

LOWER BOUND:

$$L_t \geq \frac{1}{T}\left[\sum_{t=1}^{T} f(x^{(t)}) + \min_{x\in X}\langle \sum_{t=1}^{T} \nabla f(x^{(t)}), (x - x^{(t)})\rangle + F(x)\right]$$

REGULARIZATION YIELDS:

$$L_{t+1} \geq \frac{t}{t+1}\cdot L_t + \frac{1}{t+1}\cdot\left(f(x_{t+1}) + \langle \nabla f(x_{t+1}), z_t - x_{t+1}\rangle - \frac{1}{2\sigma}\|\nabla f(x_{t+1})\|_*^2\right)$$

Dual Averaging for Non-Smooth Functions

LOWER BOUND:

$$L_t \geq \frac{1}{T}\left[\sum_{t=1}^{T} f(x^{(t)}) + \min_{x \in X}\langle \sum_{t=1}^{T} \nabla f(x^{(t)}), (x - x^{(t)})\rangle + F(x)\right]$$

REGULARIZATION YIELDS:

$$L_{t+1} \geq \frac{t}{t+1}\cdot L_t + \frac{1}{t+1}\cdot\left(f(x_{t+1}) + \langle \nabla f(x_{t+1}), z_t - x_{t+1}\rangle - \frac{1}{2\sigma}\|\nabla f(x_{t+1})\|_*^2\right)$$

Next iterate

x_{t+1}

Regularizer

Strong

Convexity

σ

Dual Averaging for Non-Smooth Functions

LOWER BOUND:

$$L_t \geq \frac{1}{T}\left[\sum_{t=1}^{T} f(x^{(t)}) + \min_{x \in X}\langle \sum_{t=1}^{T} \nabla f(x^{(t)}), (x - x^{(t)})\rangle + F(x)\right]$$

REGULARIZATION YIELDS:

$$L_{t+1} \geq \frac{t}{t+1} \cdot L_t + \frac{1}{t+1} \cdot \left(f(x_{t+1}) + \langle \nabla f(x_{t+1}), z_t - x_{t+1} \rangle - \frac{1}{2\sigma} \|\nabla f(x_{t+1})\|_*^2 \right)$$

Dual Averaging for Non-Smooth Functions

LOWER BOUND:

$$L_t \geq \frac{1}{T}\left[\sum_{t=1}^{T} f(x^{(t)}) + \min_{x \in X} \langle \sum_{t=1}^{T} \nabla f(x^{(t)}), (x - x^{(t)}) \rangle + F(x)\right]$$

REGULARIZATION YIELDS:

$$L_{t+1} \geq \frac{t}{t+1} \cdot L_t + \frac{1}{t+1} \cdot \left(f(x_{t+1}) + \langle \nabla f(x_{t+1}), z_t - x_{t+1} \rangle - \frac{1}{2\sigma} \|\nabla f(x_{t+1})\|_*^2 \right)$$

Dual Averaging/Mirror Descent Step

$$z_t = \arg\min_{x \in X} \sum_{i=1}^{t} \langle \nabla f(x_i), x - x_i \rangle + F(x).$$

Dual Averaging for Non-Smooth Functions

LOWER BOUND:

$$L_t \geq \frac{1}{T}\left[\sum_{t=1}^{T} f(x^{(t)}) + \min_{x\in X}\langle\sum_{t=1}^{T}\nabla f(x^{(t)}), (x - x^{(t)})\rangle + F(x)\right]$$

REGULARIZATION YIELDS:

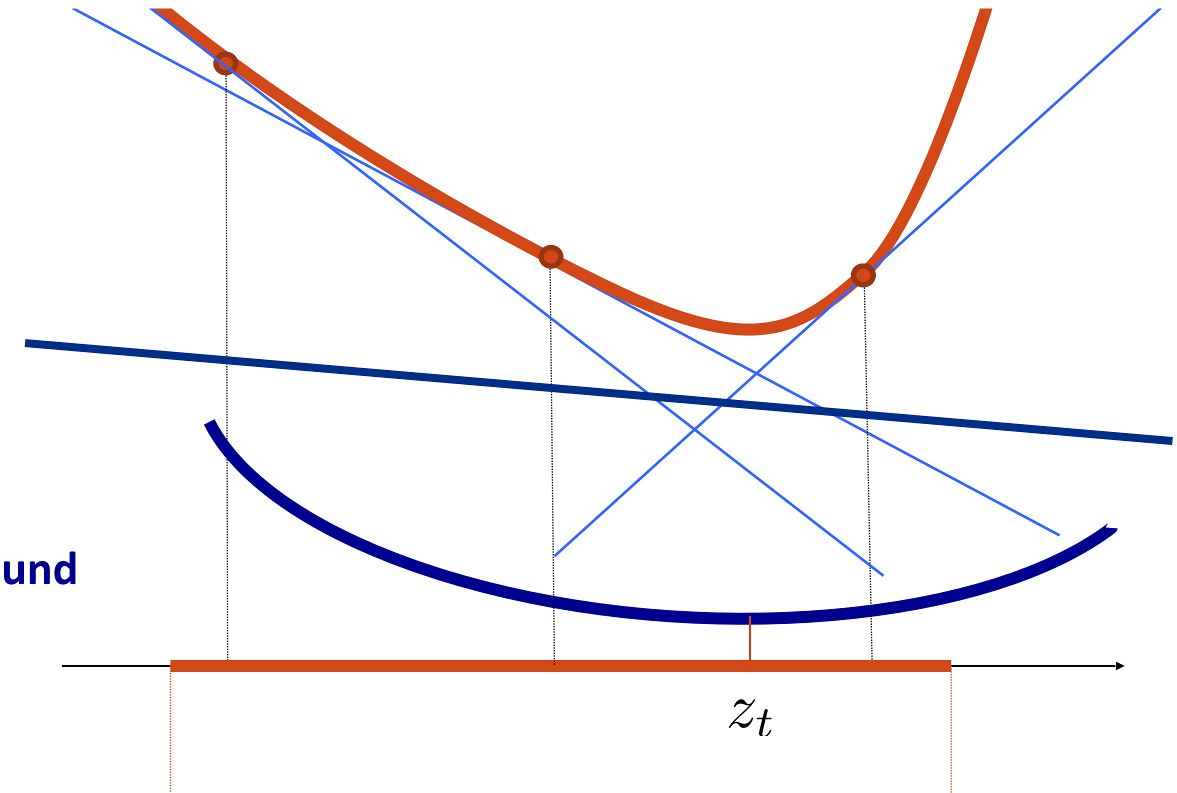
$$L_{t+1} \geq \frac{t}{t+1}\cdot L_t + \frac{1}{t+1}\cdot\left(f(x_{t+1}) + \langle\nabla f(x_{t+1}), z_t - x_{t+1}\rangle - \frac{1}{2\sigma}\|\nabla f(x_{t+1})\|_*^2\right)$$

Dual Averaging/Mirror Descent Step

$$z_t = \arg\min_{x\in X}\sum_{i=1}^{t}\langle\nabla f(x_i), x - x_i\rangle + F(x) .= \mathrm{Prox}_{z_{t-1}}^{F}(\nabla f(x_t))$$

Non-Smooth Functions: Progress on Dual Side

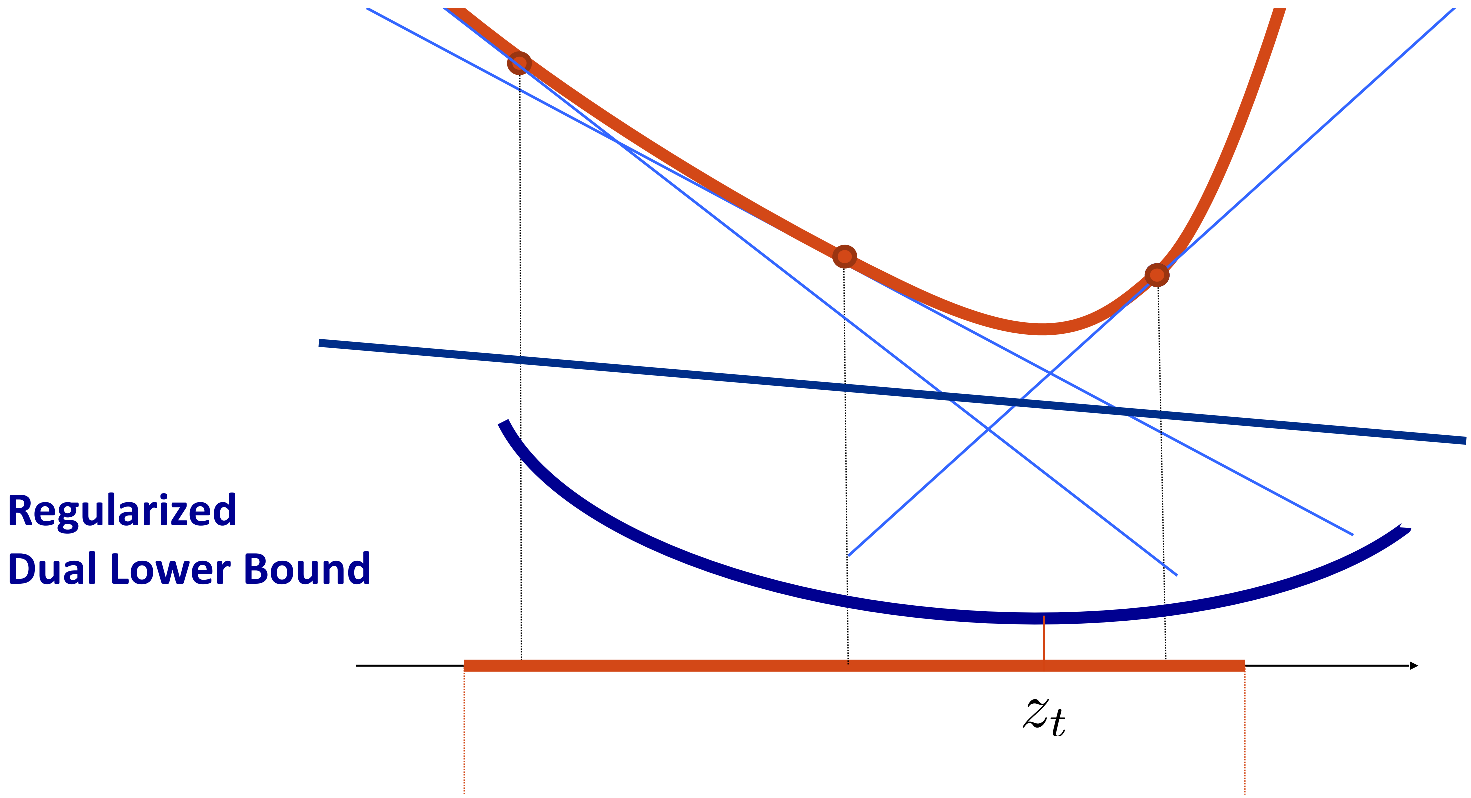
ASSUMPTION: f convex, differentiable, ρ-Lipschitz $\forall x \in X, \; \|\nabla f(x)\|_* \leq \rho$



Regularized Dual Lower Bound

Non-Smooth Functions: Progress on Dual Side

ASSUMPTION: f convex, differentiable, ρ-Lipschitz $\forall x \in X, \ \|\nabla f(x)\|_* \leq \rho$



$$z_t = \arg\min_{x \in X} \sum_{i=1}^{t} \langle \nabla f(x_i), x - x_i \rangle + F(x) = \mathrm{Prox}^F_{z_{t-1}}(\nabla f(x_t))$$

Convergence Analysis

UPPER BOUND:

$$U_T = \frac{1}{T}\left(\sum_{t=1}^{T} f(x^{(t)})\right)$$

LOWER BOUND

$$L_t = \frac{1}{T}\left[\sum_{t=1}^{T} f(x^{(t)}) + \min_{x\in X}\langle\sum_{t=1}^{T} \nabla f(x^{(t)}), (x - x^{(t)})\rangle + F(x)\right]$$

PROGRESS IN ONE ITERATION:

$$U_{t+1} - L_{t+1} =$$

$$\frac{t}{t+1}\cdot(U_t - L_t) - \frac{1}{t+1}\left(\langle\nabla f(x_{t+1}), z_t - x_{t+1}\rangle - \frac{1}{2\sigma}\|\nabla f(x_{t+1})\|_*^2\right)$$

Convergence Analysis

UPPER BOUND: $$U_T = \frac{1}{T}\left(\sum_{t=1}^{T} f(x^{(t)})\right)$$

LOWER BOUND $$L_t = \frac{1}{T}\left[\sum_{t=1}^{T} f(x^{(t)}) + \min_{x\in X}\langle\sum_{t=1}^{T} \nabla f(x^{(t)}), (x - x^{(t)})\rangle + F(x)\right]$$

PROGRESS IN ONE ITERATION:

$$U_{t+1} - L_{t+1} =$$
$$\frac{t}{t+1}\cdot(U_t - L_t) - \frac{1}{t+1}\left(\langle\nabla f(x_{t+1}), z_t - x_{t+1}\rangle - \frac{1}{2\sigma}\|\nabla f(x_{t+1})\|_*^2\right)$$

DUAL AVERAGING/MIRROR DESCENT: $x_{t+1} = z_t = \mathrm{Prox}_{z_{t-1}}^{F}(\nabla f(x_t))$

Convergence Analysis

UPPER BOUND: $$U_T = \frac{1}{T}\left(\sum_{t=1}^{T} f(x^{(t)})\right)$$

LOWER BOUND $$L_t = \frac{1}{T}\left[\sum_{t=1}^{T} f(x^{(t)}) + \min_{x\in X}\langle \sum_{t=1}^{T} \nabla f(x^{(t)}), (x - x^{(t)})\rangle + F(x)\right]$$

PROGRESS IN ONE ITERATION:

$$U_{t+1} - L_{t+1} =$$
$$\frac{t}{t+1}\cdot(U_t - L_t) - \frac{1}{t+1}\cancel{\left(\langle \nabla f(x_{t+1}), z_t - x_{t+1}\rangle - \frac{1}{2\sigma}\|\nabla f(x_{t+1})\|_*^2\right)}$$

DUAL AVERAGING/MIRROR DESCENT: $x_{t+1} = z_t = \mathrm{Prox}_{z_{t-1}}^{F}(\nabla f(x_t))$

Convergence Analysis

UPPER BOUND: $$U_T = \frac{1}{T}\left(\sum_{t=1}^{T} f(x^{(t)})\right)$$

LOWER BOUND $$L_t = \frac{1}{T}\left[\sum_{t=1}^{T} f(x^{(t)}) + \min_{x\in X}\langle\sum_{t=1}^{T}\nabla f(x^{(t)}), (x - x^{(t)})\rangle + F(x)\right]$$

PROGRESS IN ONE ITERATION:

$$U_{t+1} - L_{t+1} = \frac{t}{t+1}\cdot(U_t - L_t) + \frac{1}{t+1}\frac{\|\nabla f(x_{t+1})\|_*^2}{2\sigma}$$

DUAL AVERAGING/MIRROR DESCENT: $x_{t+1} = z_t = \mathrm{Prox}_{z_{t-1}}^{F}(\nabla f(x_t))$

Convergence Analysis

UPPER BOUND:
$$U_T = \frac{1}{T}\left(\sum_{t=1}^{T} f(x^{(t)})\right)$$

LOWER BOUND
$$L_t = \frac{1}{T}\left[\sum_{t=1}^{T} f(x^{(t)}) + \min_{x\in X}\langle\sum_{t=1}^{T}\nabla f(x^{(t)}), (x - x^{(t)})\rangle + F(x)\right]$$

PROGRESS IN ONE ITERATION:

$$U_{t+1} - L_{t+1} = \frac{t}{t+1}\cdot(U_t - L_t) + \frac{1}{t+1}\frac{\|\nabla f(x_{t+1})\|_*^2}{2\sigma}$$

DUAL AVERAGING/MIRROR DESCENT: $x_{t+1} = z_t = \mathrm{Prox}_{z_{t-1}}^{F}(\nabla f(x_t))$

CONVERGENCE:

$$U_T - L_T \leq \frac{U_0 - L_0}{T} + \frac{\sum_{t=1}^{T}\|\nabla f(x_{t+1})\|_*^2}{2\sigma\cdot T} \leq \frac{U_0 - L_0}{T} + \frac{\rho}{\sigma}$$

Summary of Upper and Lower Bounds

UPPER BOUNDS	LOWER BOUNDS
$U_T = \frac{1}{T}\left(\sum_{t=1}^{T} f(x^{(t)})\right)$ Average of function values	$L_t \geq \frac{1}{t}\left[\sum_{i=1}^{t} f(x^{(t)}) - \lVert \nabla f(x^{(t)} \rVert_* \cdot \mathrm{diam}(X)\right]$ Minimum of average hyperplane

Summary of Upper and Lower Bounds

UPPER BOUNDS	LOWER BOUNDS
$U_T = \frac{1}{T}\left(\sum_{t=1}^{T} f(x^{(t)})\right)$ Average of function values	$L_t \geq \frac{1}{t}\left[\sum_{i=1}^{t} f(x^{(t)}) - \lVert\nabla f(x^{(t)}\rVert_* \cdot \mathrm{diam}(X)\right]$ Minimum of average hyperplane
$U_t = f(y^{(t)}) \leq f(x^{(t)}) - \frac{\lVert\nabla f(x^{(t)})\rVert_*^2}{2L}$ Function value after gradient step $x^{(t+1)} = y^{(t)} = \mathrm{Grad}(x^{(t)})$	

Summary of Upper and Lower Bounds

UPPER BOUNDS	LOWER BOUNDS
$U_T = \frac{1}{T}\left(\sum_{t=1}^{T} f(x^{(t)})\right)$	$L_t \geq \frac{1}{t}\left[\sum_{i=1}^{t} f(x^{(t)}) - \|\nabla f(x^{(t)}\|_* \cdot \mathrm{diam}(X)\right]$
Average of function values	Minimum of average hyperplane
$U_t = f(y^{(t)}) \leq f(x^{(t)}) - \frac{\|\nabla f(x^{(t)})\|_*^2}{2L}$	$L_t = \frac{1}{T}\left[\sum_{t=1}^{T} f(x^{(t)}) + \min_{x \in X} \langle \sum_{t=1}^{T} \nabla f(x^{(t)}), (x - x^{(t)}) \rangle + F(x)\right]$
Function value after gradient step	Regularized minimum of average hyperplane
$x^{(t+1)} = y^{(t)} = \mathrm{Grad}(x^{(t)})$	$x_{t+1} = z_t = \mathrm{Prox}_{z_{t-1}}^{F}(\nabla f(x_t))$

Nesterov's Acceleration

Use better strategy on **both primal and dual side**:

UPPER BOUND: $U_t = f(y^{(t)}) \leq f(x^{(t)}) - \frac{\|\nabla f(x^{(t)})\|_*^2}{2L}$

LOWER BOUND: $L_t = \frac{1}{A_t}\left[\sum_{t=1}^{T} \alpha_t f(x^{(t)}) + \min_{x \in X} \langle \sum_{t=1}^{T} \alpha_t \nabla f(x^{(t)}), (x - x^{(t)}) \rangle + F(x)\right]$

Nesterov's Acceleration

Use better strategy on **both primal and dual side**:

UPPER BOUND: $$U_t = f(y^{(t)}) \leq f(x^{(t)}) - \frac{\|\nabla f(x^{(t)})\|_*^2}{2L}$$

LOWER BOUND: $$L_t = \underbrace{\frac{1}{A_t} \left[\sum_{t=1}^{T} \alpha_t f(x^{(t)}) + \min_{x \in X} \langle \sum_{t=1}^{T} \alpha_t \nabla f(x^{(t)}), (x - x^{(t)}) \rangle + F(x) \right]}$$

Non-uniform distribution $\frac{\vec{\alpha}_t}{A_t}$

Nesterov's Acceleration

Use better strategy on **both primal and dual side**:

UPPER BOUND: $$U_t = f(y^{(t)}) \leq f(x^{(t)}) - \frac{\|\nabla f(x^{(t)})\|_*^2}{2L}$$

LOWER BOUND: $$L_t = \frac{1}{A_t}\left[\underline{\sum_{t=1}^{T} \alpha_t f(x^{(t)}) + \min_{x\in X}\langle \sum_{t=1}^{T} \alpha_t \nabla f(x^{(t)}), (x - x^{(t)})\rangle} + F(x)\right]$$

Non-uniform distribution $\frac{\vec{\alpha_t}}{A_t}$

PROGRESS IN ONE ITERATION:

Nesterov's Acceleration

Use better strategy on **both primal and dual side**:

UPPER BOUND: $U_t = f(y^{(t)}) \leq f(x^{(t)}) - \frac{\|\nabla f(x^{(t)})\|_*^2}{2L}$

LOWER BOUND: $L_t = \frac{1}{A_t}\left[\sum_{t=1}^{T} \alpha_t f(x^{(t)}) + \min_{x \in X} \langle \sum_{t=1}^{T} \alpha_t \nabla f(x^{(t)}), (x - x^{(t)}) \rangle + F(x)\right]$

Non-uniform distribution $\frac{\vec{\alpha}_t}{A_t}$

PROGRESS IN ONE ITERATION:

$$
\begin{aligned}
U_{t+1} - L_{t+1} \leq \frac{A_t}{A_{t+1}}(U_t - L_t) + \\
&\frac{1}{A_{t+1}} \langle \nabla f(x_{t+1}), \alpha_{t+1} z_t + A_t y_t - x_{t+1} \rangle \\
&- \frac{1}{A_{t+1}} \left(\frac{1}{2L} - \frac{\alpha_{t+1}^2}{2\sigma} \right) \cdot \|\nabla f(x_{t+1})\|^2.
\end{aligned}
$$

Accelerating Non-Negative LPs

Non-Negative Linear Programs (NNLPs)

Linear Programs where objective and constraints are non-negative.
<u>Feasibility formulation</u>:

$$
\begin{aligned}
Ax &\geq a, \\
Bx &\leq b, \\
x &\geq 0.
\end{aligned}
$$

where $A \geq 0, B \geq 0$.

Many applications:

- resource allocation,
- packing LPs,
- covering LPs,
- **mixed packing-covering LPs**

Variations:

1. Explicit: constraint matrices are given explicitly.
2. Implicit: exponential number of constraints with efficient separation oracle.

Non-Negative Linear Programs (NNLPs)

NNLP can always be written as

$$\begin{aligned} Cx &\geq 1, \\ Px &\leq 1, \\ x &\geq 0. \end{aligned}$$

where $P \geq 0, C \geq 0.$

Notions of approximation is **multiplicative**: find x such that

$$\max_i (Px)_i \leq (1+\epsilon) \cdot \min_j (Cx)_j .$$

Computational models: sequential, parallel, distributed.

Running time depends on sparsity N of P and C.

What's special about NNLPs?

Non-smooth optimization problem with Lipschitz parameter

$$\underbrace{\rho = \max\{\|P\|_{1\to\infty}, \|C\|_{1\to\infty}}_{\text{Largest Entry of P and C}}\}$$

What's special about NNLPs?

Non-smooth optimization problem with Lipschitz parameter

WIDTH:

$$\underbrace{\rho = \max\{\|P\|_{1\to\infty}, \|C\|_{1\to\infty}\}}_{\text{Largest Entry of P and C}}$$

What's special about NNLPs?

Non-smooth optimization problem with Lipschitz parameter

WIDTH: $$\underbrace{\rho = \max\{\|P\|_{1\to\infty}, \|C\|_{1\to\infty}\}}_{\text{Largest Entry of P and C}}$$

In general, non-smooth optimization requires:

$$\frac{\rho^2}{\epsilon^2} \text{ gradient computations}$$

For general LPs, we can exploit the minmax structure. This requires:

$$\frac{\rho}{\epsilon} \text{ gradient computations}$$

What's special about NNLPs?

Non-smooth optimization problem with Lipschitz parameter

WIDTH: $$\underbrace{\rho = \max\{\|P\|_{1\to\infty}, \|C\|_{1\to\infty}}_{\text{Largest Entry of P and C}}\}$$

In general, non-smooth optimization requires:

$$\frac{\rho^2}{\epsilon^2}$$ gradient computations

For general LPs, we can exploit the minmax structure. This requires:

$$\frac{\rho}{\epsilon}$$ gradient computations

UNESCAPABLE WIDTH DEPENDENCE?

Width-Independent Algorithms

Non-smooth optimization problem with Lipschitz parameter

WIDTH: $$\underbrace{\rho = \max\{\|P\|_{1\to\infty}, \|C\|_{1\to\infty}}_{\text{Largest Entry of P and C}}\}$$

In general, non-smooth optimization requires:

$\frac{\rho^2}{\epsilon^2}$ gradient computations $\longrightarrow$ $\tilde{O}\left(\frac{1}{\epsilon^2}\right)$ Young['01]

For general LPs, we can exploit the minmax structure. This requires:

$\frac{\rho}{\epsilon}$ gradient computations $\longrightarrow$ $\tilde{O}\left(\frac{1}{\epsilon}\right)$ **OUR WORK**

Width-Independent Algorithms

Non-smooth optimization problem with Lipschitz parameter

WIDTH: $$\underbrace{\rho = \max\{\|P\|_{1\to\infty}, \|C\|_{1\to\infty}\}}_{\text{Largest Entry of P and C}}$$

In general, non-smooth optimization requires:

$\frac{\rho^2}{\epsilon^2}$ gradient computations $\longrightarrow$ $\tilde{O}\left(\frac{1}{\epsilon^2}\right)$ Young['01]

For general LPs, we can exploit the minmax structure. This requires:

$\frac{\rho}{\epsilon}$ gradient computations $\longrightarrow$ $\tilde{O}\left(\frac{1}{\epsilon}\right)$ **OUR WORK**

KEY CONTRIBUTION: Accelerating Width-Independent Algorithms

Explaining Width Independence

What's special about non-negative LPs?

Consider **saddle point formulation** for packing LP:

$$\min_{x \geq 0} \max_{y \geq 0} \langle y, Ax \rangle - \langle 1, x \rangle - \langle 1, y \rangle$$

Standard Regularization/Smoothing by entropy:

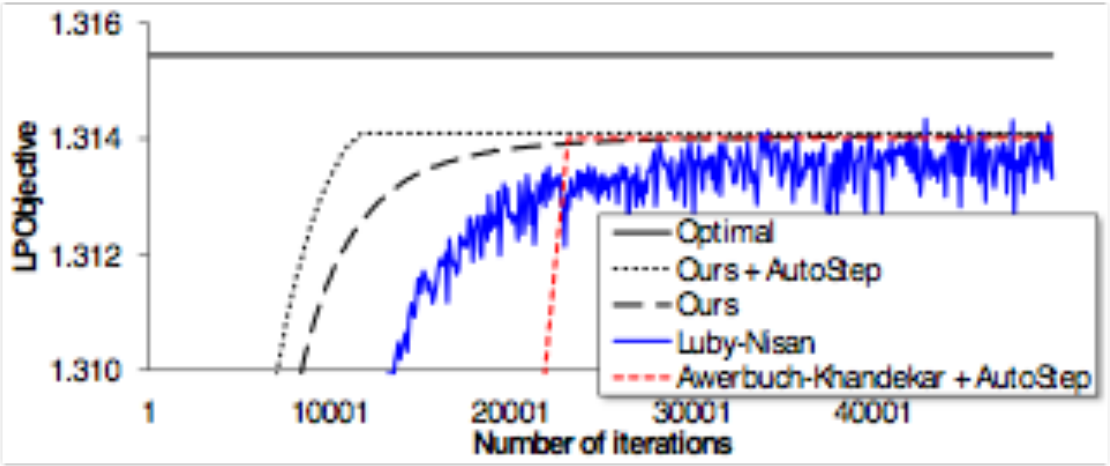
$$f_\mu(x) = \max_{y \geq 0} \langle A^T - 1, x \rangle - \langle 1, y \rangle + \mu H(x)$$

Taylor Series:

when smoothness is bad, gradient is large

Running Time Bounds: Parallel Algorithms

Problem	Paper	Total Work	Number of Iterations[a]	Notes
p/c LP	[LN93]	$\frac{\log^2 N}{\varepsilon^4} \times (N \log n)$	$\frac{\log^2 N}{\varepsilon^4}$	
p/c LP	[BBR97, BBR04]	$\frac{\log^3 N}{\varepsilon^4} \times N$	$\frac{\log^3 N}{\varepsilon^4}$	
p/c LP	[You01]	$\frac{\log^3 N}{\varepsilon^4} \times N$	$\frac{\log^3 N}{\varepsilon^4}$	mixed p/c
p/c LP	[AK08a]	$\frac{\log^4 N}{\varepsilon^5} \times N$	$\frac{\log^4 N}{\varepsilon^5}$	stateless
p/c LP	[this paper]	$\frac{\log^2 N}{\varepsilon^3} \times N$	$\frac{\log^2 N}{\varepsilon^3}$	semi-stateless



THE END – THANK YOU