

# Convexity, Smoothness and the Gradient Method

Master 2 Data Science, Univ. Paris Saclay

**Robert M. Gower**



# Solving the Finite Sum Training Problem

# Optimization Sum of Terms

## A Datum Function

$$f_i(w) := \ell(h_w(x^i), y^i) + \lambda R(w)$$

$$\begin{aligned} \frac{1}{n} \sum_{i=1}^n \ell(h_w(x^i), y^i) + \lambda R(w) &= \frac{1}{n} \sum_{i=1}^n (\ell(h_w(x^i), y^i) + \lambda R(w)) \\ &= \frac{1}{n} \sum_{i=1}^n f_i(w) \end{aligned}$$

## Finite Sum Training Problem

$$\min_{w \in \mathbf{R}^d} \frac{1}{n} \sum_{i=1}^n f_i(w) =: f(w)$$

# The Training Problem

Solving the *training problem*:

$$\min_{w \in \mathbf{R}^d} \frac{1}{n} \sum_{i=1}^n f_i(w)$$

Reference method: Gradient descent

$$\nabla \left( \frac{1}{n} \sum_{i=1}^n f_i(w) \right) = \frac{1}{n} \sum_{i=1}^n \nabla f_i(w)$$

## Gradient Descent Algorithm

Set  $w^0 = 0$ , choose  $\alpha > 0$ .

for  $t = 1, 2, 3, \dots, T$

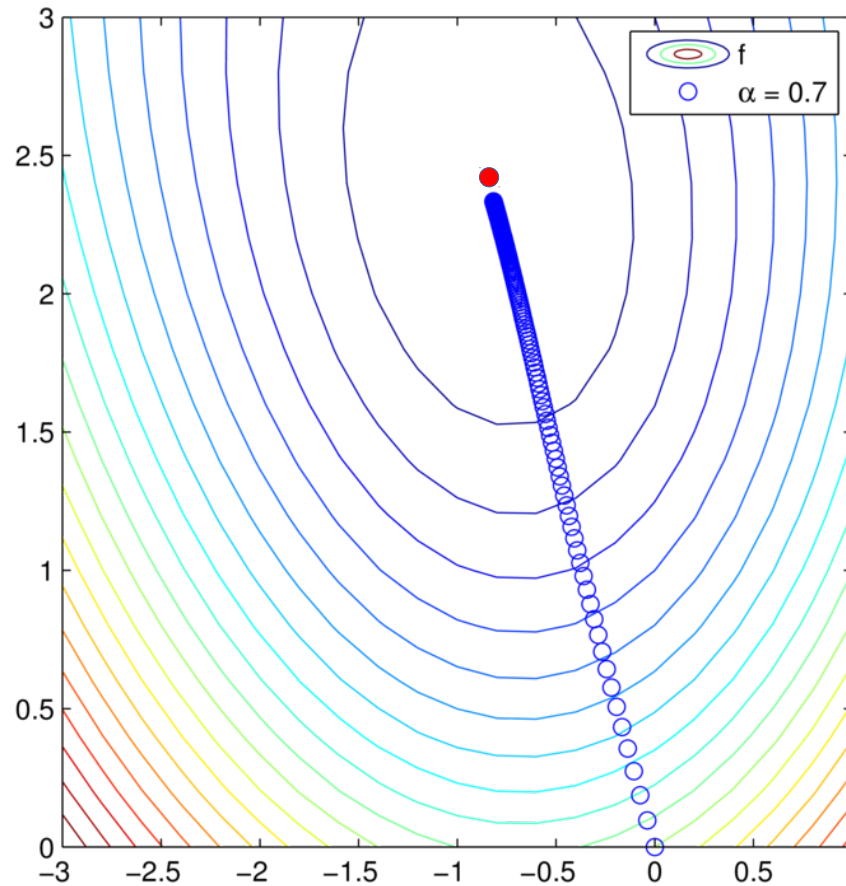
$$w^{t+1} = w^t - \frac{\alpha}{n} \sum_{i=1}^n \nabla f_i(w^t)$$

Output  $w^{T+1}$

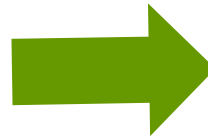
# Gradient Descent Example

A Logistic Regression problem using the fourclass labelled data from LIBSVM

$(n, d) = (862, 2)$



Can we prove  
that this  
always works!?



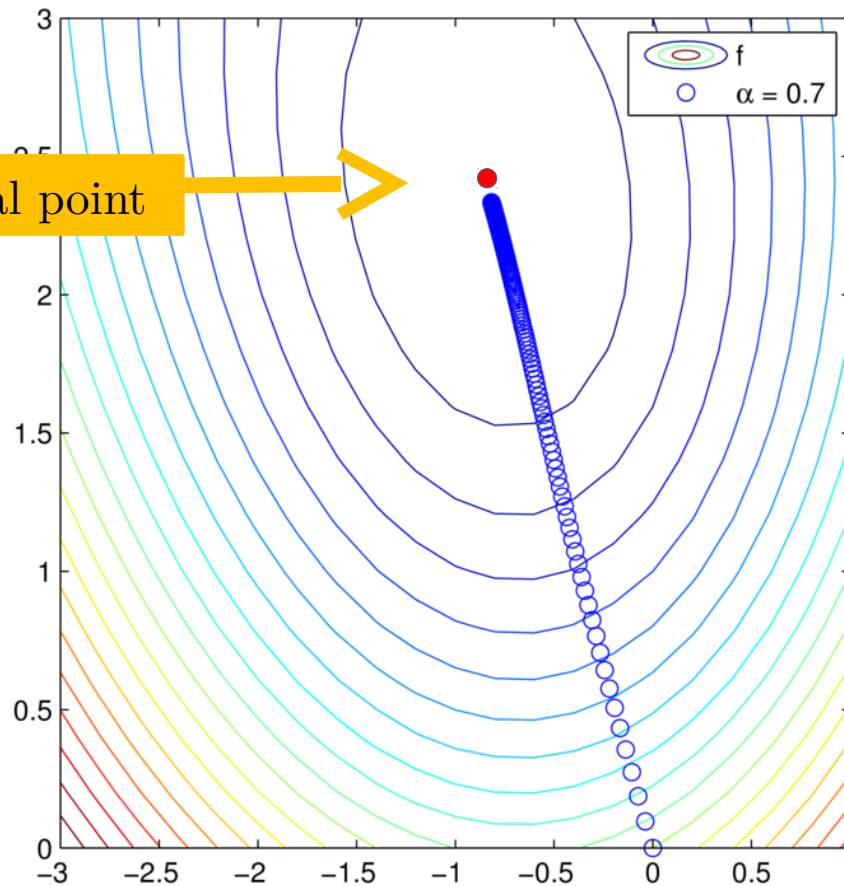
Convex  
training  
problems

# Gradient Descent Example

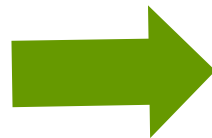
Optimal point

A Logistic Regression problem using the fourclass labelled data from LIBSVM

$(n, d) = (862, 2)$



Can we prove  
that this  
always works!?

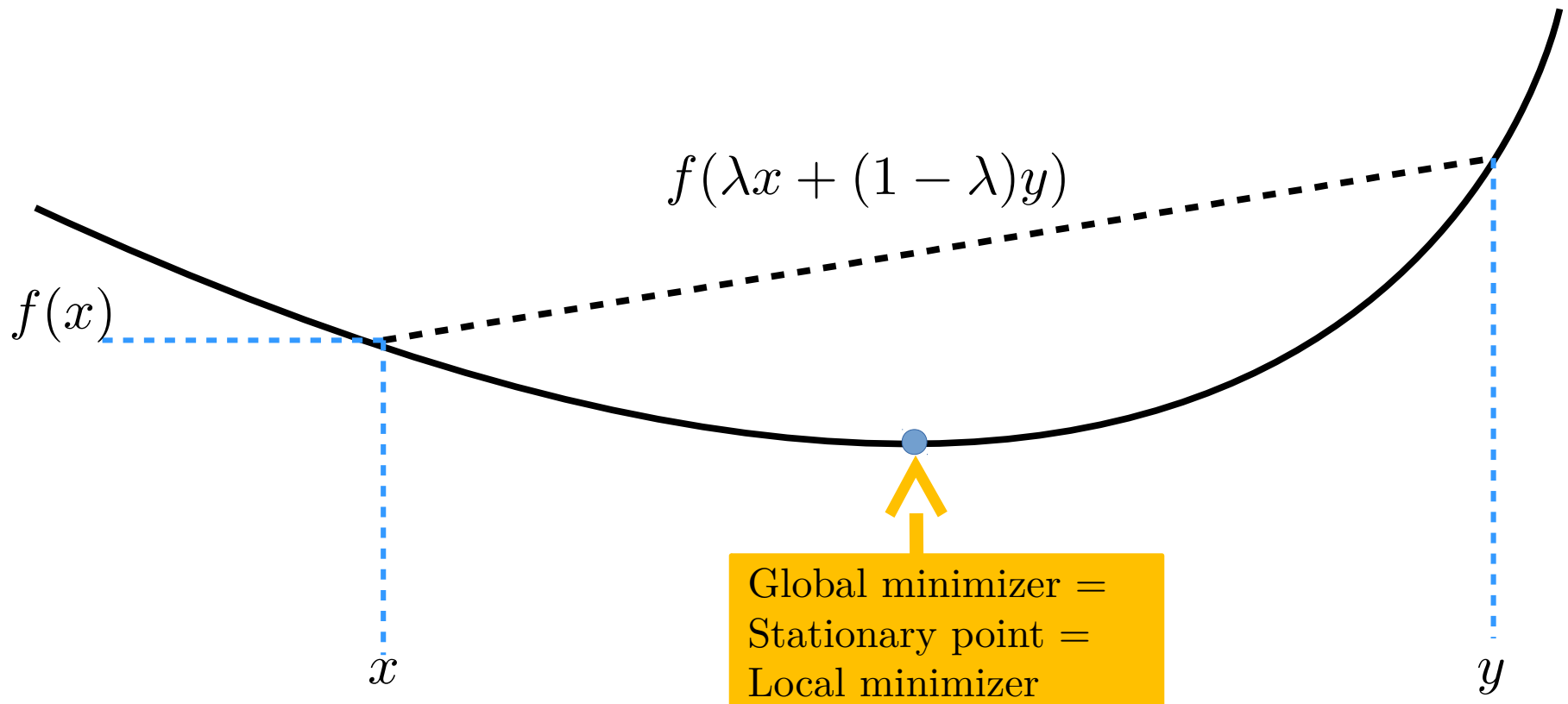


Convex  
training  
problems

# Convexity

We say  $f : \text{dom}(f) \subset \mathbb{R}^n \rightarrow \mathbb{R}$  is convex if  $\text{dom}(f)$  is convex and

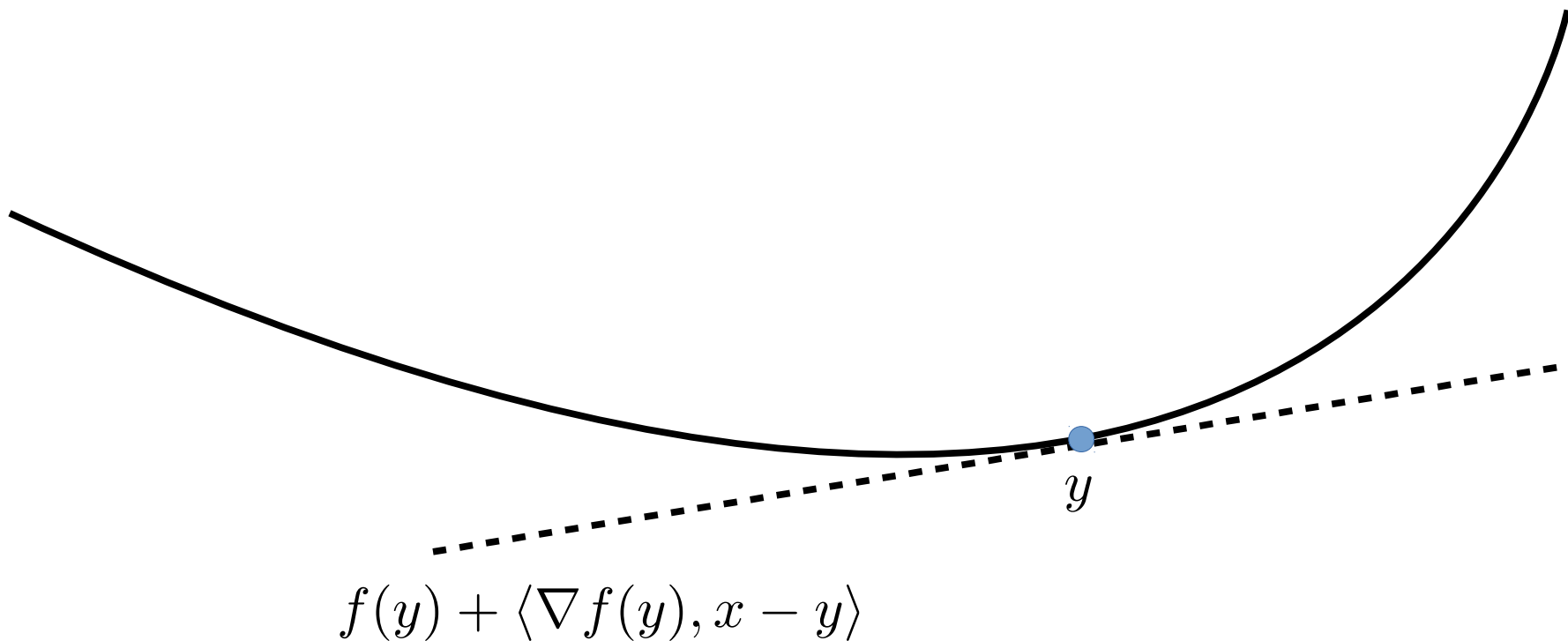
$$f(\lambda x + (1 - \lambda)y) \leq \lambda f(x) + (1 - \lambda)f(y), \quad \forall x, y \in C, \lambda \in [0, 1]$$



# Convexity: First derivative

A differential function  $f : \text{dom}(f) \subset \mathbb{R}^n \rightarrow \mathbb{R}$  is convex iff

$$f(x) \geq f(y) + \langle \nabla f(y), x - y \rangle$$

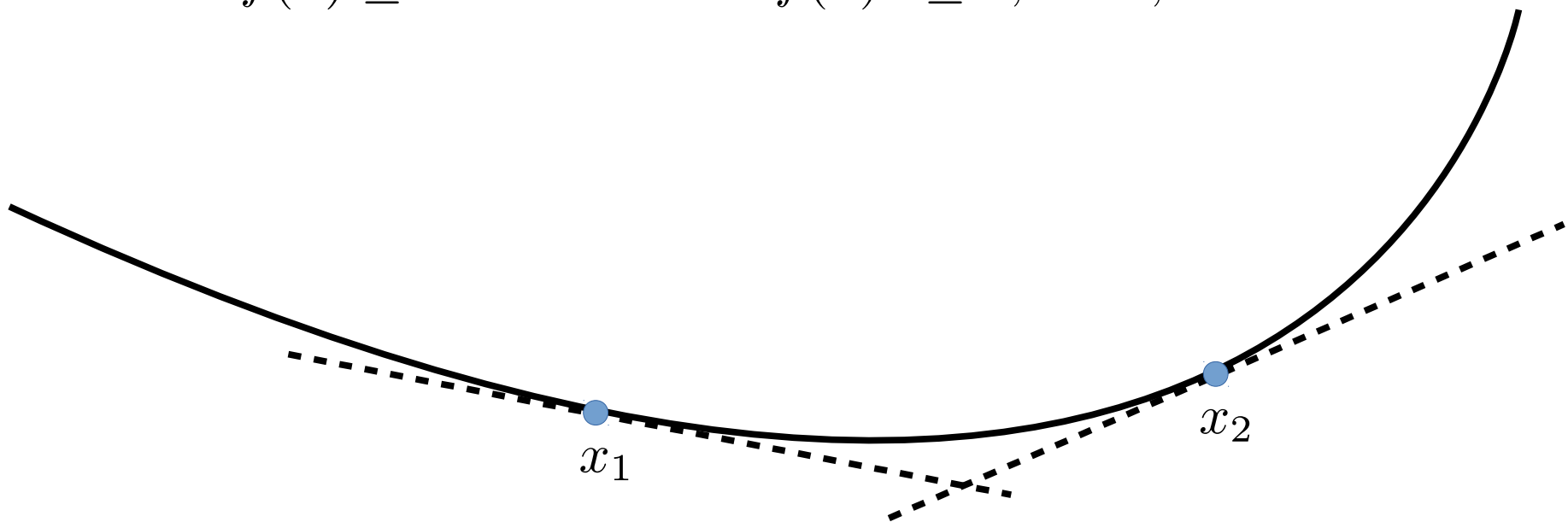




# Convexity: Second derivative

A twice differential function  $f : \text{dom}(f) \subset \mathbb{R}^n \rightarrow \mathbb{R}$  is convex iff

$$\nabla^2 f(x) \succeq 0 \quad \Leftrightarrow \quad v^\top \nabla^2 f(x) v \geq 0, \quad \forall x, v \in \mathbb{R}^n$$



$$x_1 \leq x_2 \quad \Rightarrow \quad f'(x_1) \leq f'(x_2)$$

# Convexity: Examples

Extended-value extension:

$$f : \mathbb{R}^n \rightarrow \mathbb{R} \cup \{\infty\}$$

$$f(x) = \infty, \quad \forall x \notin \text{dom}(f)$$

Norms and squared norms:

$$x \mapsto \|x\|$$

$$x \mapsto \|x\|^2$$

Proof is an  
exercise!

Negative log and logisitc:

$$x \mapsto -\log(x)$$

$$x \mapsto \log \left( 1 + e^{-y \langle a, x \rangle} \right)$$

Hinge loss

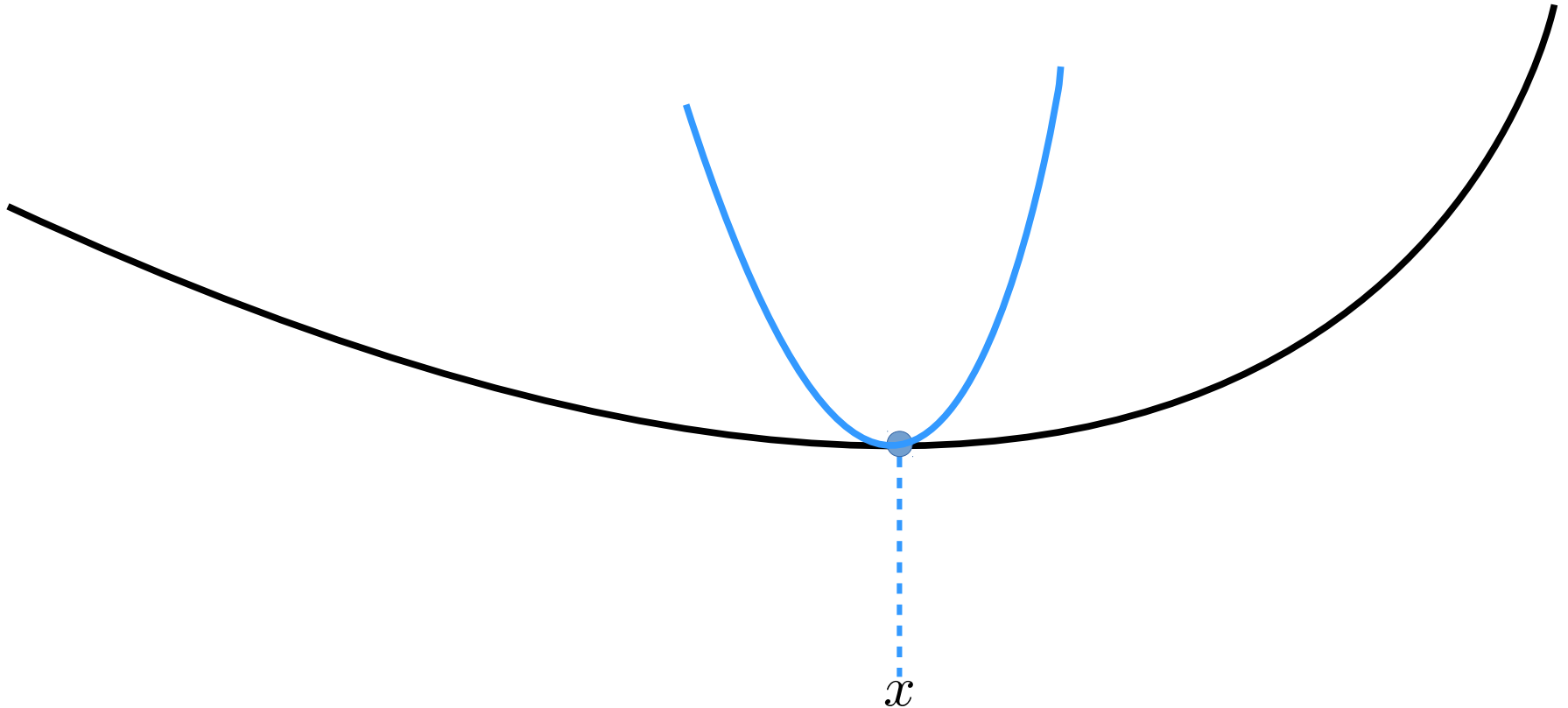
$$x \mapsto \max\{0, 1 - yx\}$$

Negatives log determinant, exponentiation ... etc

# Smoothness

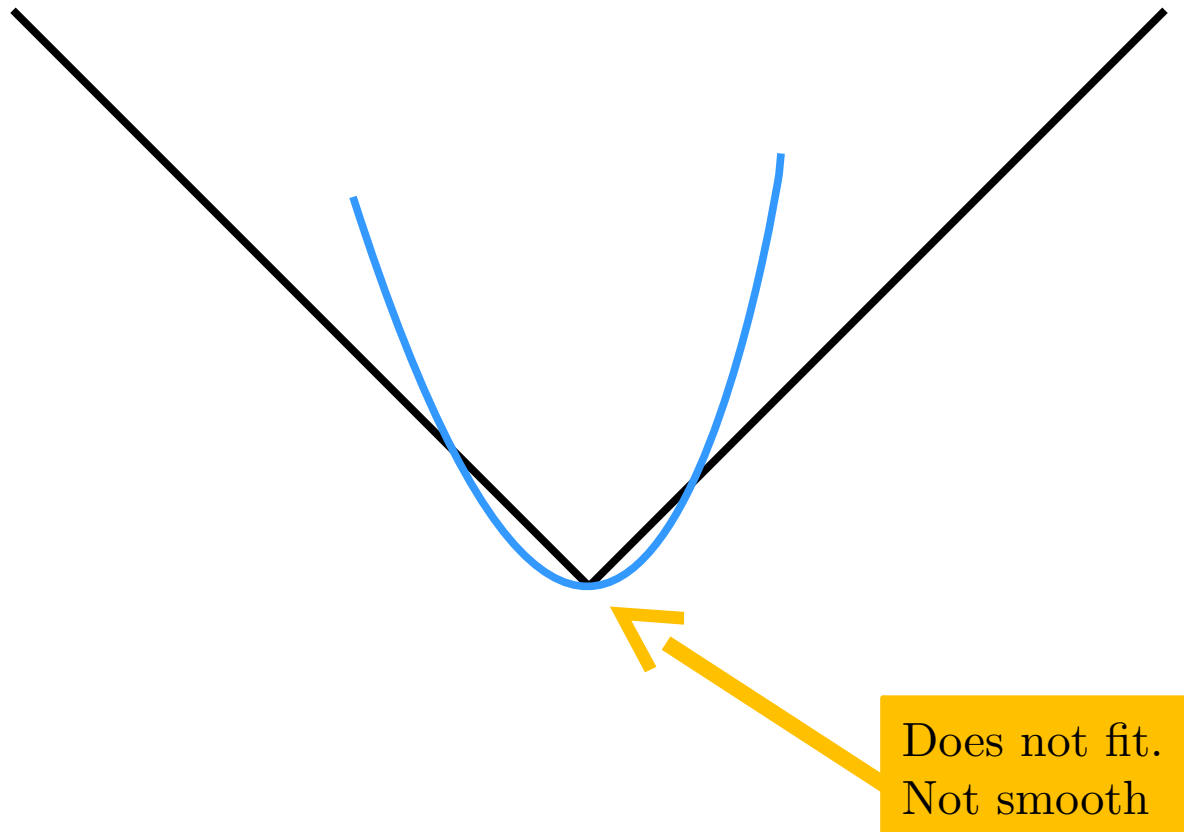
We say  $f : \mathbb{R}^n \rightarrow \mathbb{R} \cup \{\infty\}$  is smooth if

$$f(x) \leq f(y) + \langle \nabla f(y), x - y \rangle + \frac{L}{2} \|x - y\|^2, \quad \forall x, y \in \mathbb{R}^n$$



# Smoothness: Convex counter-example

$$f(x) = \|x\|_1 = \sum_{i=1}^n |x_i|$$



# Smoothness Equivalence

We say  $f : \mathbb{R}^n \rightarrow \mathbb{R} \cup \{\infty\}$  is  $L$ -smooth if

$$\|\nabla f(x) - \nabla f(y)\| \leq L\|x - y\|, \quad \forall x, y \in \mathbb{R}^n$$

$$\nabla^2 f(x) \preceq L \cdot I, \quad \forall x \in \mathbb{R}^n$$

$$f(x) \leq f(y) + \langle \nabla f(y), x - y \rangle + \frac{L}{2}\|x - y\|^2, \quad \forall x, y \in \mathbb{R}^n$$

Proof on board

# Insight into Gradient Descent

$$f(x) \leq f(y) + \langle \nabla f(y), x - y \rangle + \frac{L}{2} \|x - y\|^2, \quad \forall x, y \in \mathbb{R}^n$$

Minimizing the upper bound in  $x$  we get:

$$\nabla_x \left( f(y) + \langle \nabla f(y), x - y \rangle + \frac{L}{2} \|x - y\|^2 \right) = \nabla f(y) + L(x - y) = 0$$



$$x = y - \frac{1}{L} \nabla f(y)$$

A gradient  
descent step !

# Smoothness: Examples

Convex quadratics:

$$x \mapsto x^T A x + b^T x + c$$

Logistic:

$$x \mapsto \log \left( 1 + e^{-y \langle a, x \rangle} \right)$$

Trigonometric:

$$x \mapsto \cos(x), \sin(x)$$

Proof is an  
exercise!

# Smoothness Properties

If  $f : \mathbb{R}^n \rightarrow \mathbb{R} \cup \{\infty\}$  is smooth then

$$f(x - \frac{1}{L} \nabla f(x)) - f(x) \leq -\frac{1}{2L} \|\nabla f(x)\|_2^2, \quad \forall x \in \mathbb{R}^n$$

$$f(x^*) - f(x) \leq -\frac{1}{2L} \|\nabla f(x)\|_2^2, \quad \forall x \in \mathbb{R}^n$$

Proof on board



# Convex and Smooth Properties

If  $f : \mathbb{R}^n \rightarrow \mathbb{R} \cup \{\infty\}$  convex and  $L$ -smooth then

$$f(y) - f(x) \leq \langle \nabla f(y), y - x \rangle - \frac{1}{2L} \|\nabla f(y) - \nabla f(x)\|_2^2$$

**Co-coercivity**

$$\langle \nabla f(y) - \nabla f(x), y - x \rangle \geq \frac{1}{L} \|\nabla f(x) - \nabla f(y)\|_2^2$$

Proof on board

# Convergence GD I

## Theorem

Let  $f$  be convex and  $L$ -smooth.

$$f(x^T) - f(x^1) \leq \frac{2L\|x^1 - x^*\|_2^2}{T-1} = O\left(\frac{1}{T}\right).$$

Where

$$x^{t+1} = x^t - \frac{1}{L} \nabla f(x^t)$$

Proof on board

$$\Rightarrow \text{for } \frac{f(x^T) - f(x^*)}{\|x^1 - x^*\|_2^2} \leq \epsilon \text{ we need } T \geq \frac{2L}{\epsilon} = O\left(\frac{1}{\epsilon}\right)$$

# Strong convexity

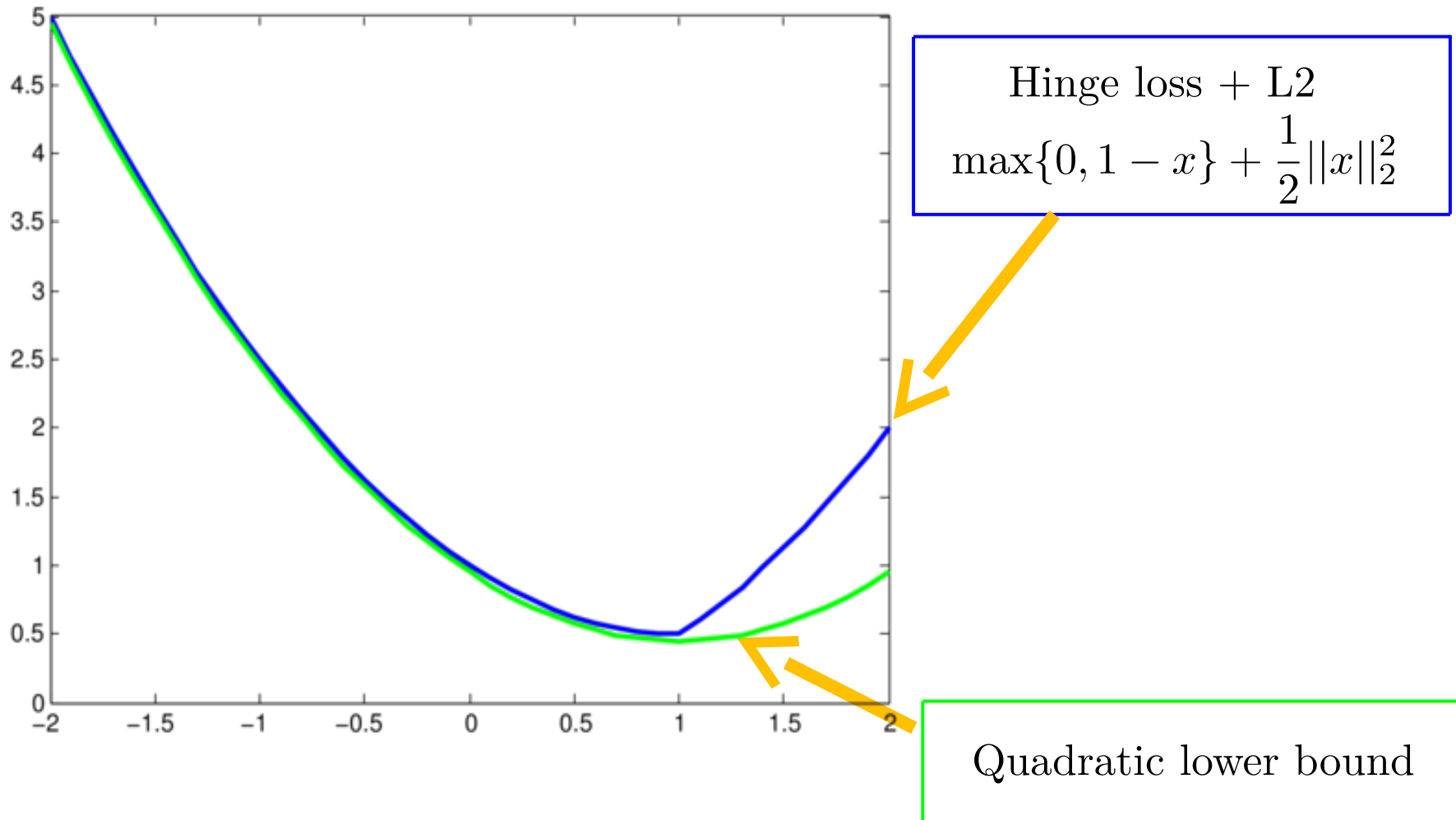
We say  $f : \mathbb{R}^n \rightarrow \mathbb{R} \cup \{\infty\}$  is  $\mu$ -strongly convex if

$$\|\nabla f(x) - \nabla f(y)\| \geq \mu \|x - y\|, \quad \forall x, y \in \mathbb{R}^n$$

$$\nabla^2 f(x) \succeq \mu \cdot I, \quad \forall x \in \mathbb{R}^n$$

$$f(x) \geq f(y) + \langle \nabla f(y), x - y \rangle + \frac{\mu}{2} \|x - y\|^2, \quad \forall x, y \in \mathbb{R}^n$$

# Example of Strong Convexity



# Strong Convexity Properties

If  $f : \mathbb{R}^n \rightarrow \mathbb{R} \cup \{\infty\}$  is  $\mu$ -strongly convex then

$$\|\nabla f(x)\|_2^2 \geq 2\mu(f(x) - f(x^*)), \quad \forall x \in \mathbb{R}^n$$

This is known as the *Polyak-Lojasiewicz* inequality.

Proof on board

# Convergence GD II

## Theorem

Let  $f$  be  $\mu$ -convex and  $L$ -smooth.

$$\|x^T - x^*\|_2^2 \leq \left(1 - \frac{\mu}{L}\right)^T \|x^1 - x^*\|_2^2$$

Where

$$x^{t+1} = x^t - \frac{1}{L} \nabla f(x^t)$$

Proof on board

$$\Rightarrow \text{for } \frac{\|x^T - x^*\|_2^2}{\|x^1 - x^*\|_2^2} \leq \epsilon \text{ we need } T \geq \frac{L}{\mu} \log \left( \frac{1}{\epsilon} \right) = O \left( \log \left( \frac{1}{\epsilon} \right) \right)$$