Optimization for Datascience

Convexity, Smoothness and the Gradient Method

Robert M. Gower

TELECOM ParisTech

Today we will

- **Lecture:** Basic theory and exercises on convexity, smoothness, strong convexity and convergence proofs
- **Exercises lists:**

  complexity_rates_exe

  exe_convexity_smoothness

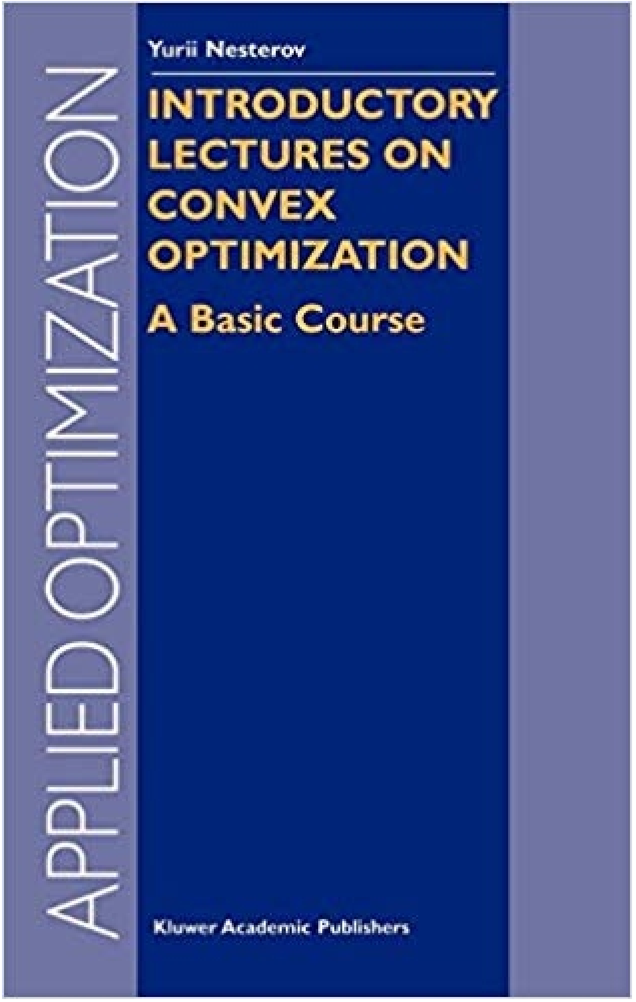
  ridge_reg_exe

References for todays class

Yurii Nestorov (2004)
Introductory Lectures on Convex Programming

Chapter 1 and Section 2.1

Free pdf online !



Solving the Finite Sum Training Problem

Optimization Sum of Terms

A Datum Function

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

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

Finite Sum Training Problem

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

Optimization Sum of Terms

A Datum Function

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

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

Finite Sum Training Problem

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

How to solve unconstrained optimization?

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 $t = 1, 2, 3, \ldots, T$

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

Output w^{T+1}

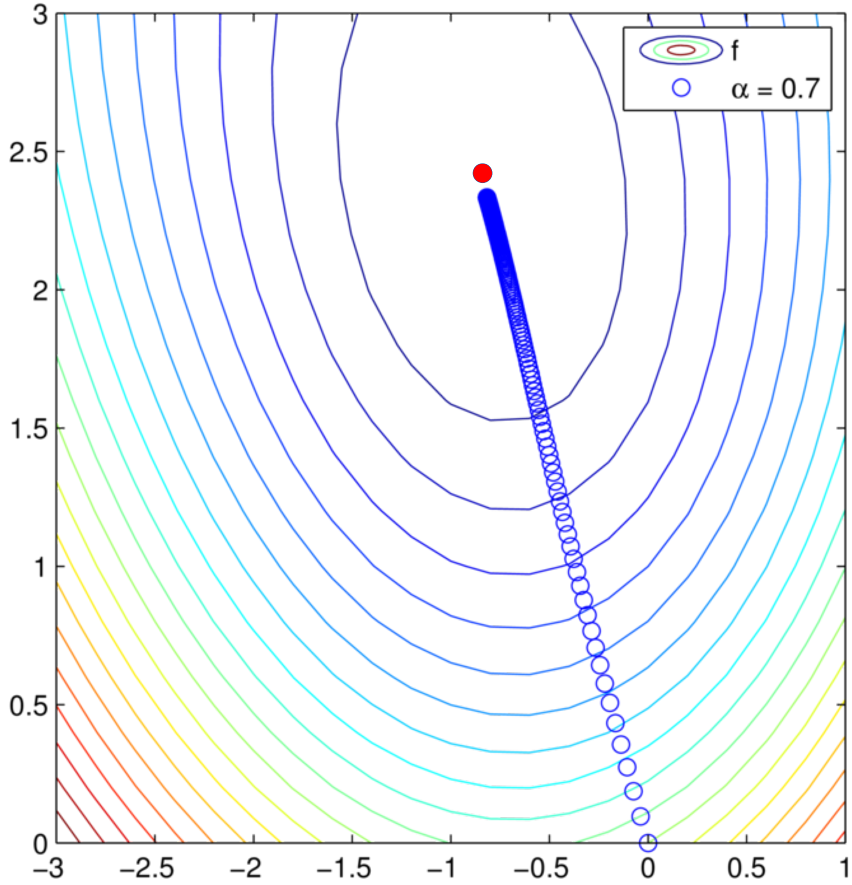
Gradient Descent Example

A Logistic Regression problem using the fourclass labelled data from LIBSVM

$(n,\ d)=(862{,}2)$

Logistic Regression

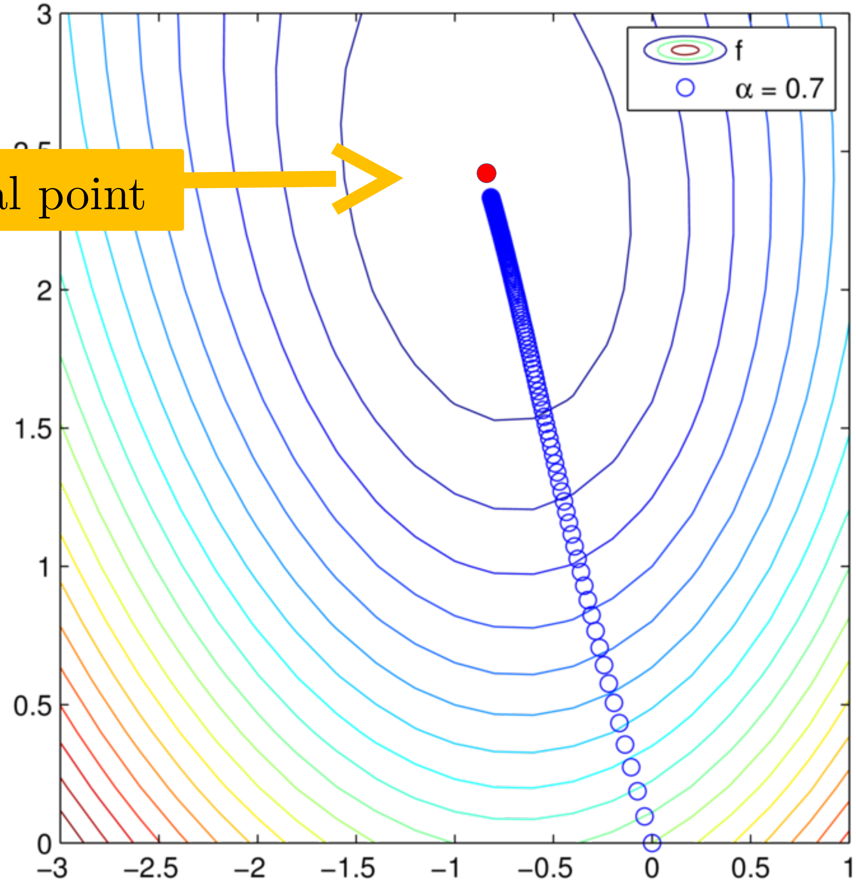
$$\min_{w\in\mathbf{R}^d} \frac{1}{n}\sum_{i=1}^{n}\ln(1+e^{-y^i\langle w,x^i\rangle})+\lambda||w||_2^2$$



Can we prove that this always works?

Gradient Descent Example

Optimal point



A Logistic Regression problem using the fourclass labelled data from LIBSVM

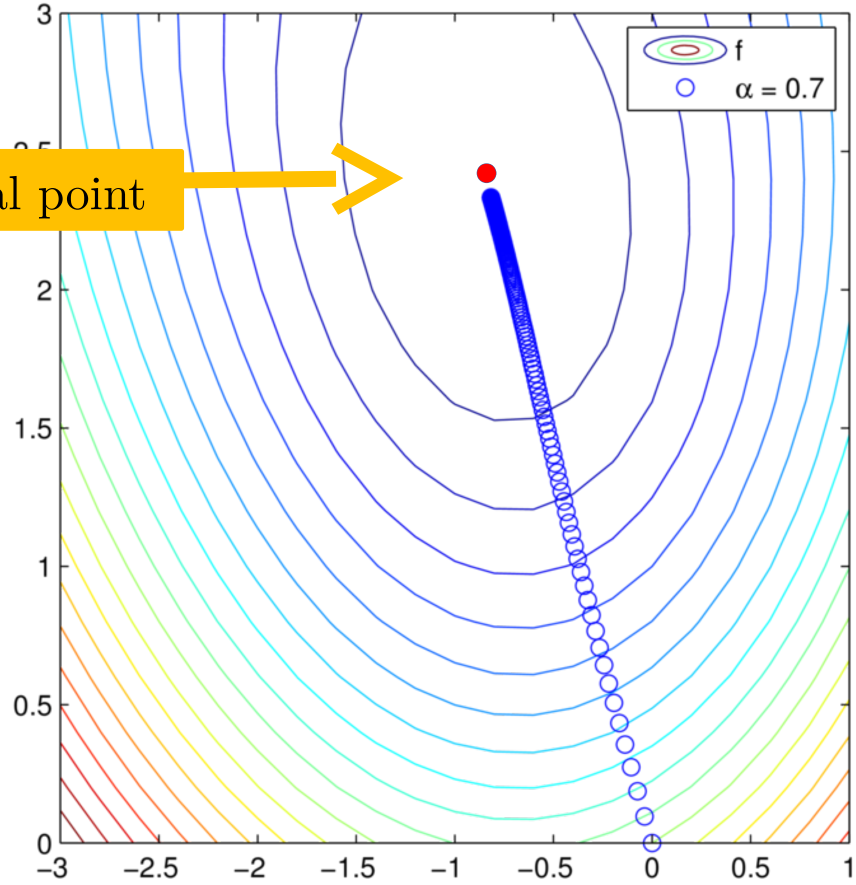
$(n,\ d) = (862{,}2)$

Logistic Regression

$$\min_{w \in \mathbf{R}^d} \frac{1}{n} \sum_{i=1}^{n} \ln(1 + e^{-y^i \langle w, x^i \rangle}) + \lambda ||w||_2^2$$

Can we prove that this always works?

Gradient Descent Example



Optimal point

A Logistic Regression problem using the fourclass labelled data from LIBSVM
$(n,\ d) = (862{,}2)$

Logistic Regression

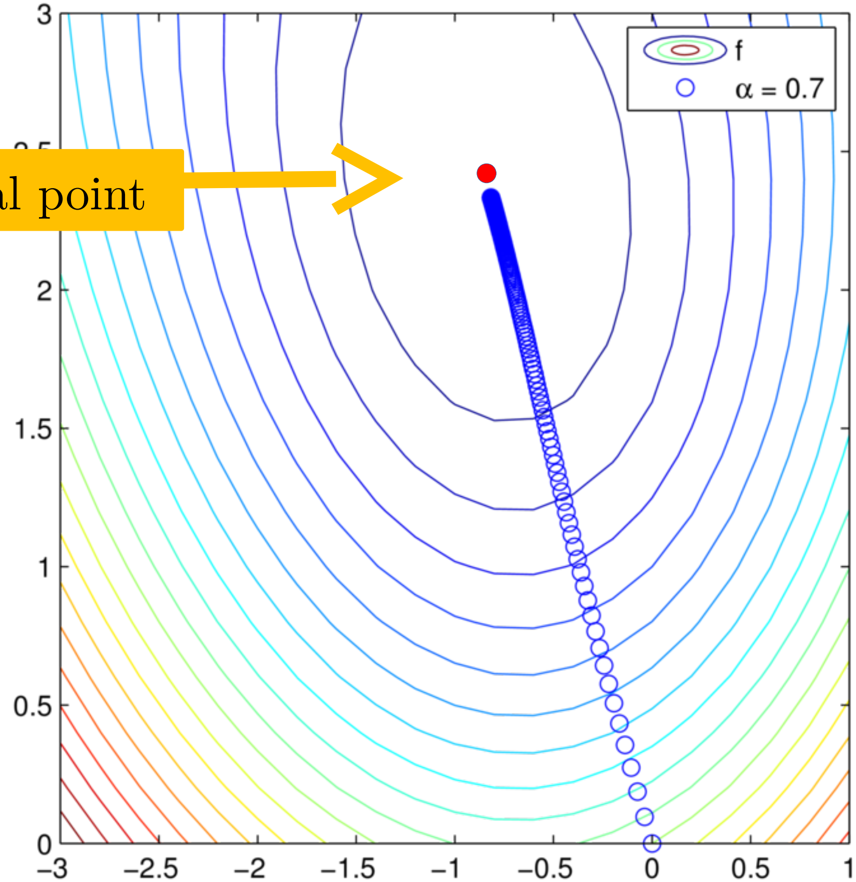
$$\min_{w \in \mathbf{R}^d} \frac{1}{n} \sum_{i=1}^{n} \ln(1 + e^{-y^i \langle w, x^i \rangle}) + \lambda ||w||_2^2$$

Can we prove that this always works?

No! There is no universal optimization method. The "no free lunch" of Optimization

Gradient Descent Example

Optimal point



A Logistic Regression problem using the fourclass labelled data from LIBSVM

$(n,\ d) = (862{,}2)$

Logistic Regression

$$\min_{w \in \mathbf{R}^d} \frac{1}{n} \sum_{i=1}^{n} \ln(1 + e^{-y^i \langle w, x^i \rangle}) + \lambda ||w||_2^2$$

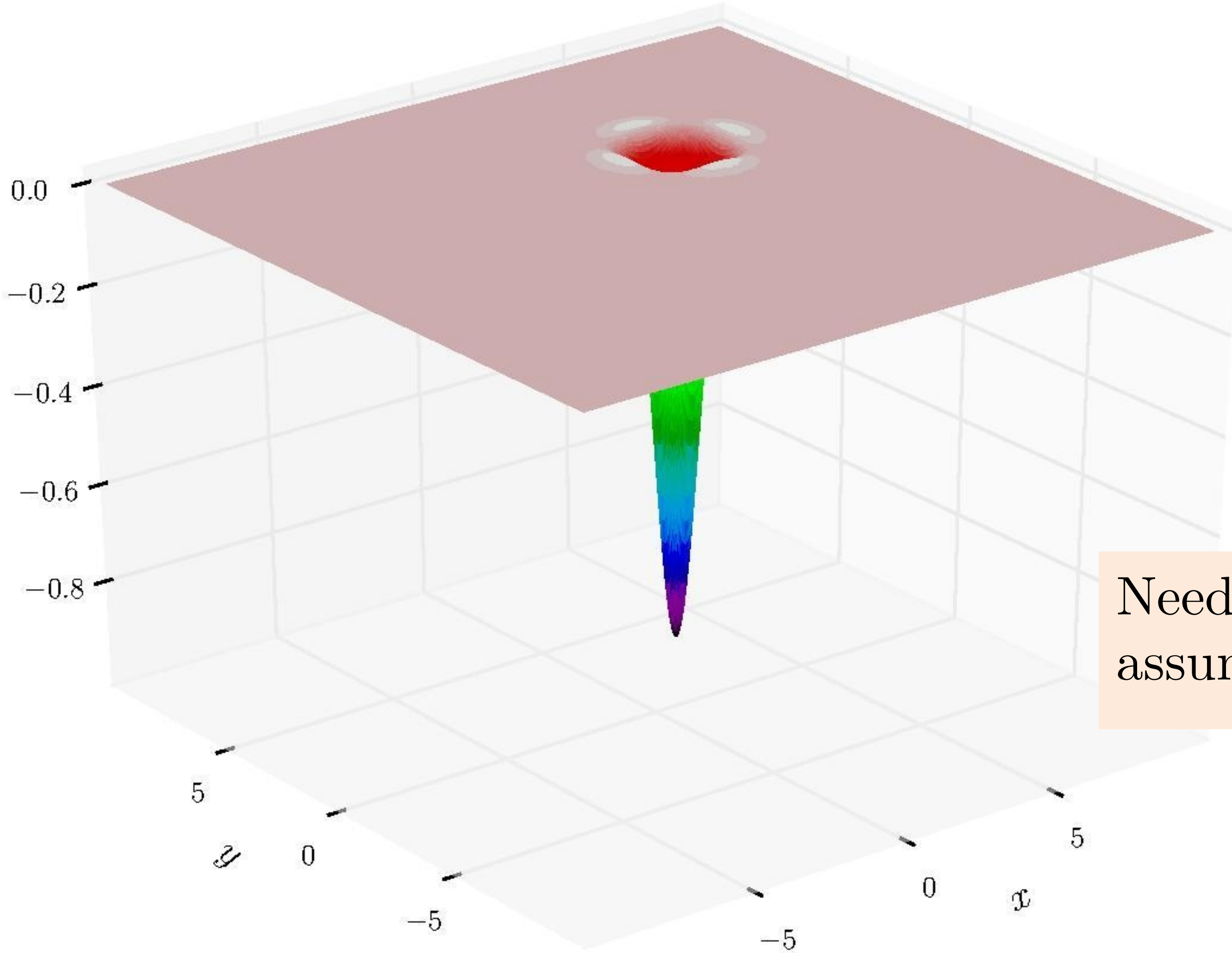
Can we prove that this always works?

No! There is no universal optimization method. The "no free lunch" of Optimization

Specialize

Convex and smooth training problems

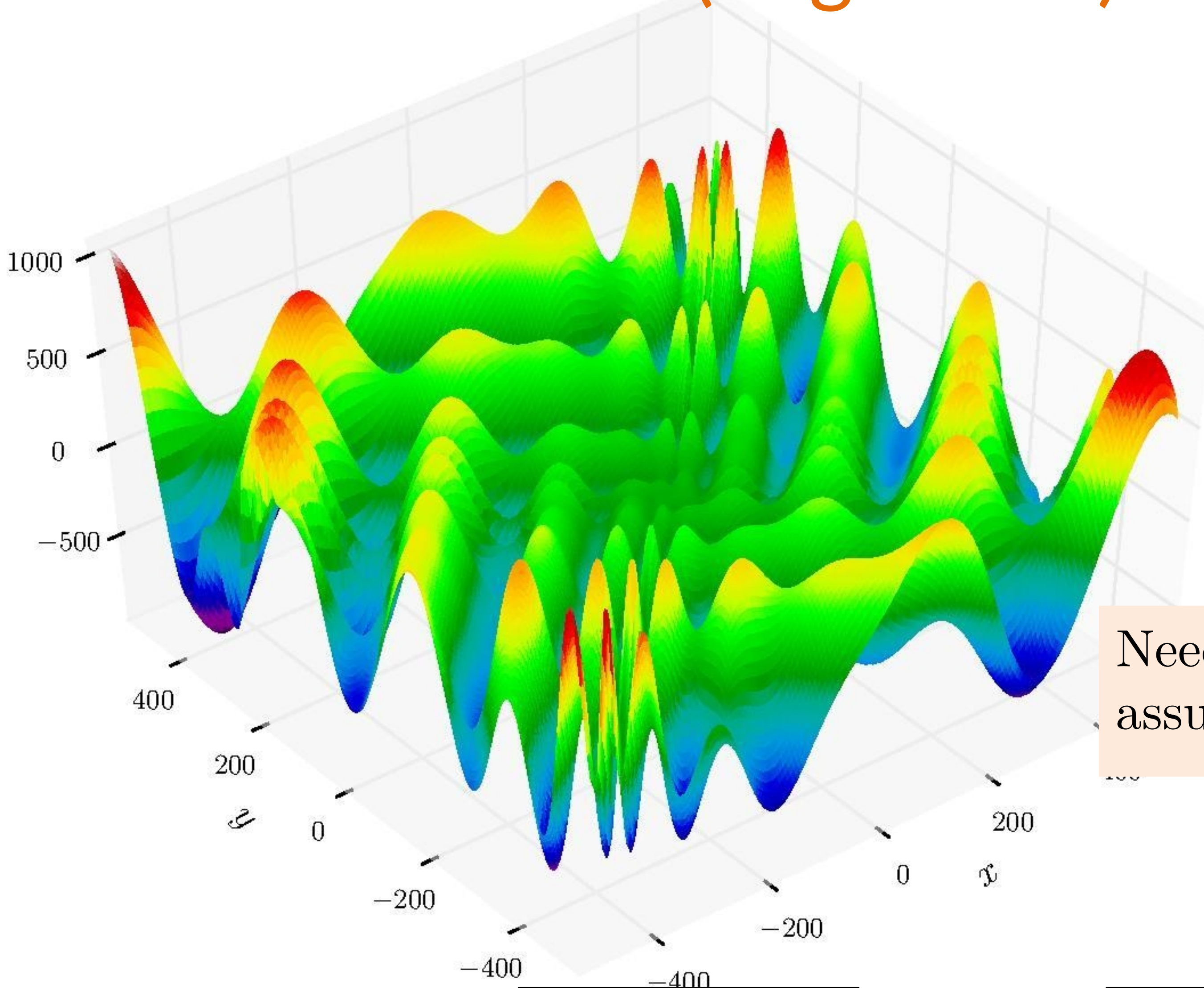
Optimization is hard (in general)



Need assumptions!

$$f(x, y) = -\cos(x)\cos(y)\exp\left(-(x-\pi)^2 - (y-\pi)^2\right)$$

Optimization is hard (in general)

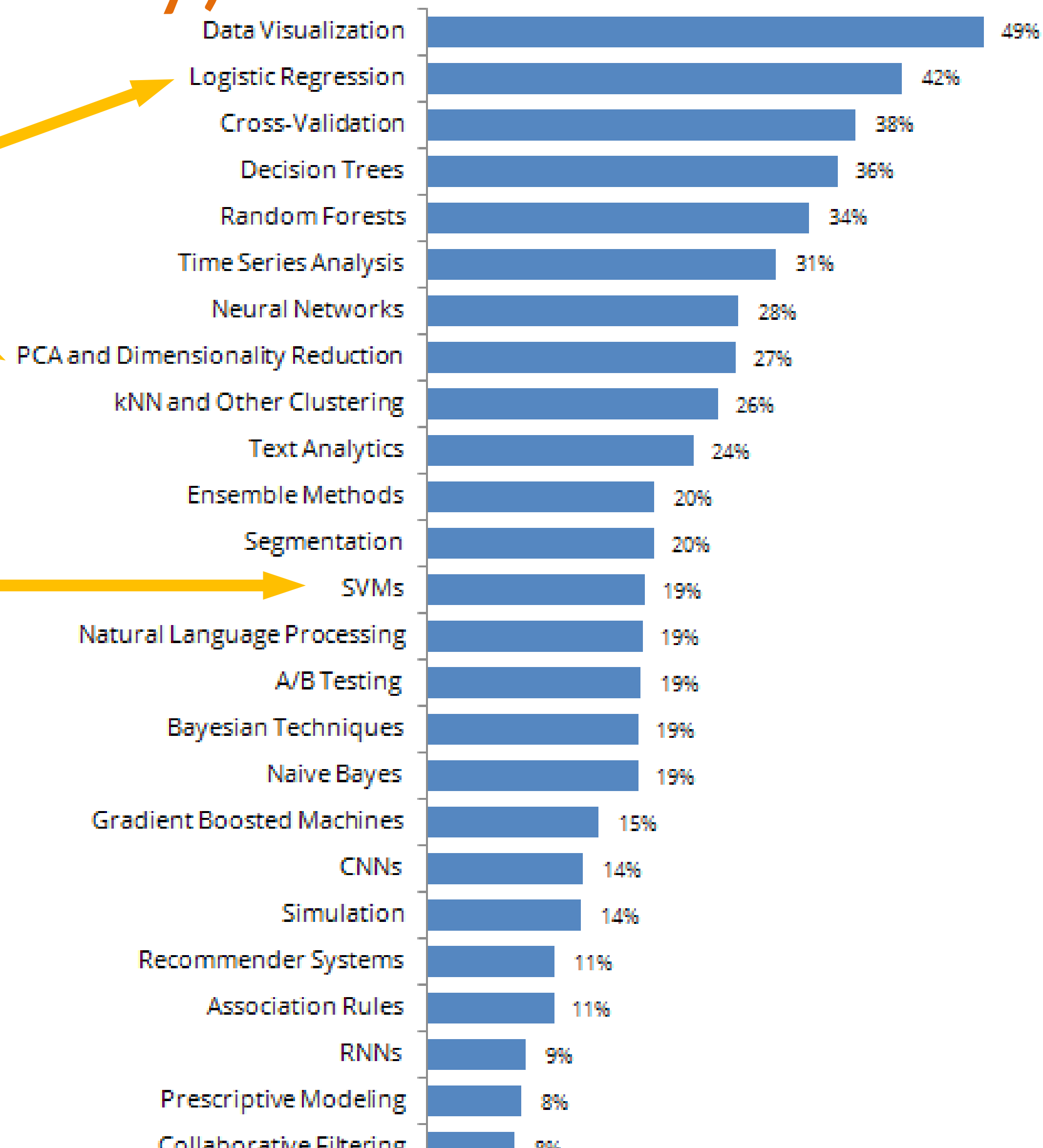


Need assumptions!

$$f(x,y) = -(y+47)\sin\sqrt{\left|\frac{x}{2}+(y+47)\right|} - x\sin\sqrt{\left|\frac{x}{2}-(y+47)\right|}$$

Data science methods most used (Kaggle 2017 survey)

Convex Optimization problems



Main assumption

Nice property

$$\text{If } \nabla f(w^*) = 0 \quad \text{then} \quad f(w^*) \leq f(w), \quad \forall w \in \mathbb{R}^d$$

All stationary points are global minima

Lemma: Convexity => Nice property

$$\text{If } \; f(w) \geq f(y) + \langle \nabla f(y), w - y \rangle, \quad \forall w, y \in \mathbb{R}^d$$

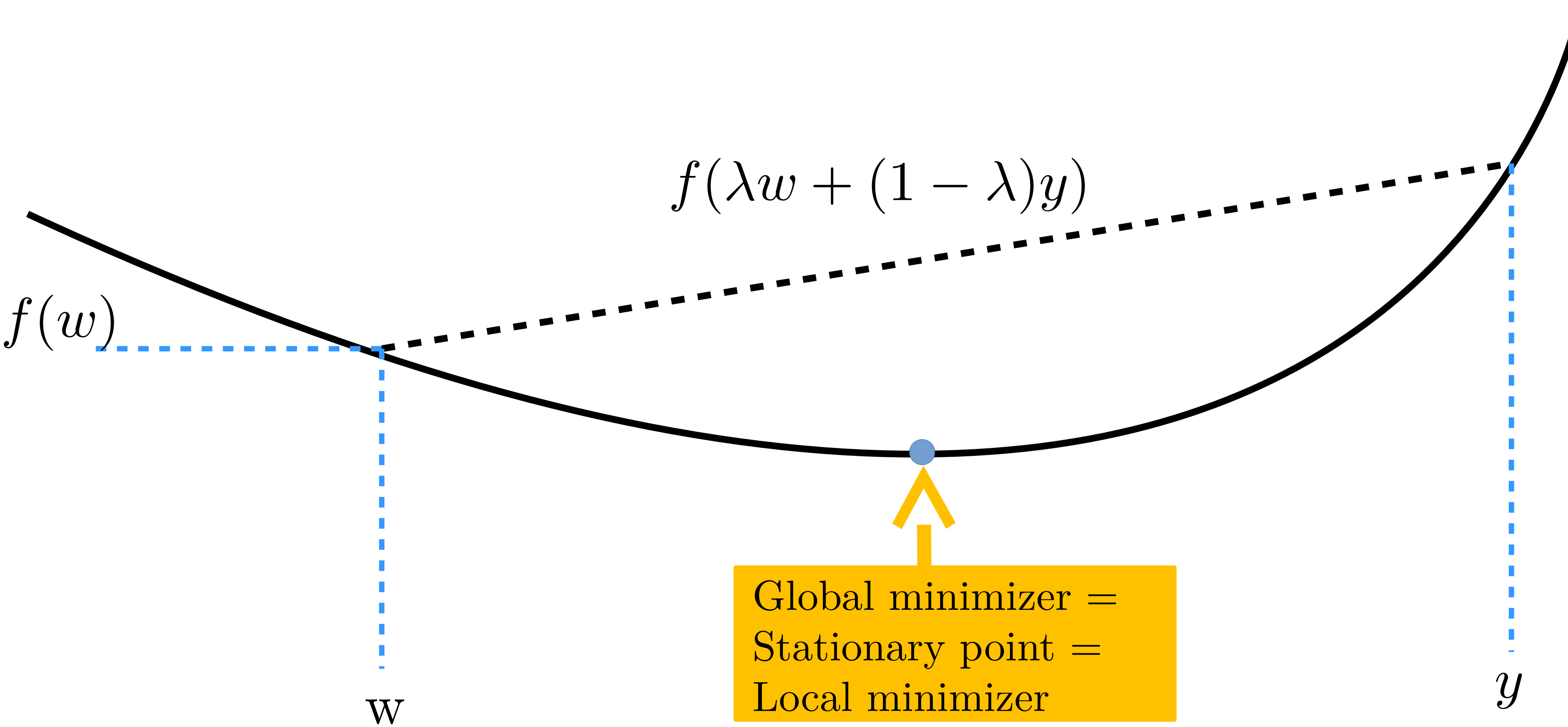
then nice property holds

PROOF: Choose $y = w^*$

Convexity

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

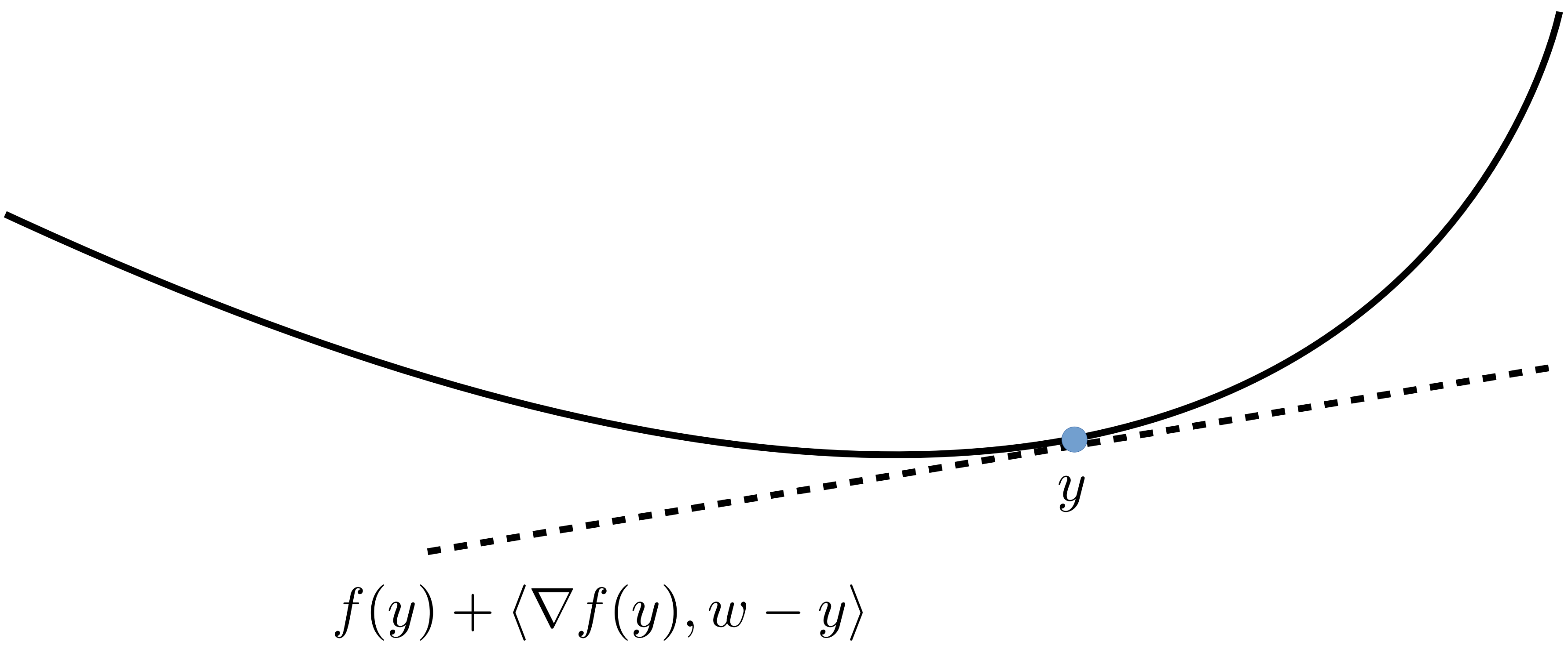
$$f(\lambda w + (1-\lambda)y) \leq \lambda f(w) + (1-\lambda) f(y), \quad \forall w, y \in C, \lambda \in [0, 1]$$



Convexity: First derivative

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

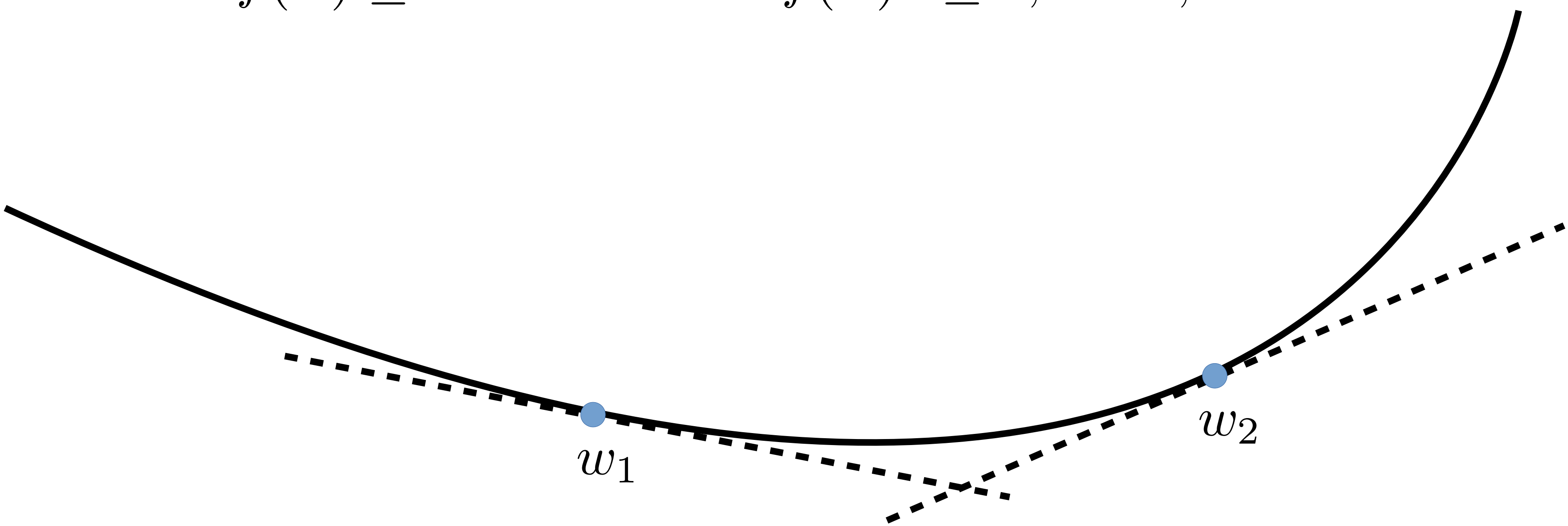
$$f(w) \geq f(y) + \langle \nabla f(y), w - y \rangle$$



Convexity: Second derivative

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

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



$$w_1 \leq w_2 \quad \Rightarrow f'(w_1) \leq f'(w_2)$$

Convexity: Examples

Extended-value extension:

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

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

Norms and squared norms:

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

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

Proof is an exercise!

Negative log and logistic:

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

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

Hinge loss

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

Negatives log determinant, exponentiation ... etc

Smoothness

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

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

Smoothness

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

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

If a twice differentiable $f : \mathbb{R}^n \to \mathbb{R} \cup \{\infty\}$ is L–smooth then

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

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

Smoothness

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

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

If a twice differentiable $f : \mathbb{R}^n \to \mathbb{R} \cup \{\infty\}$ is L–smooth then

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

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

EXE: Using that $$\sigma_{\max}(X)^2||d||_2^2 \geq ||X^\top d||_2^2$$

Show that $\frac{1}{2}||X^\top w - b||_2^2$ is $\sigma_{\max}(X)^2$–smooth

Smoothness: Examples

Convex quadratics: $x \mapsto x^\top A x + b^\top x + c$

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

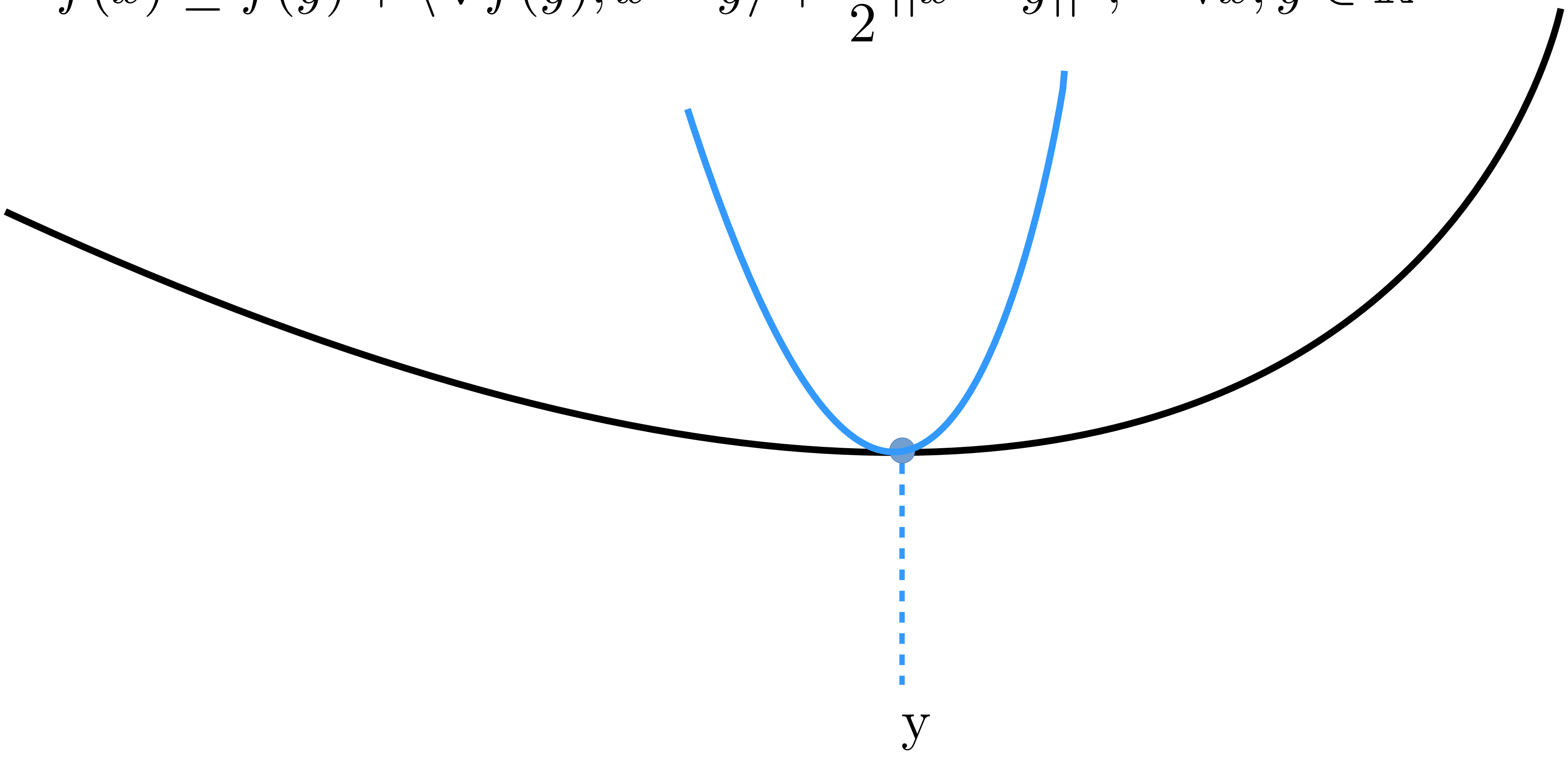
Trigonometric: $x \mapsto \cos(x), \sin(x)$

Proof is an exercise!

Important consequences of Smoothness

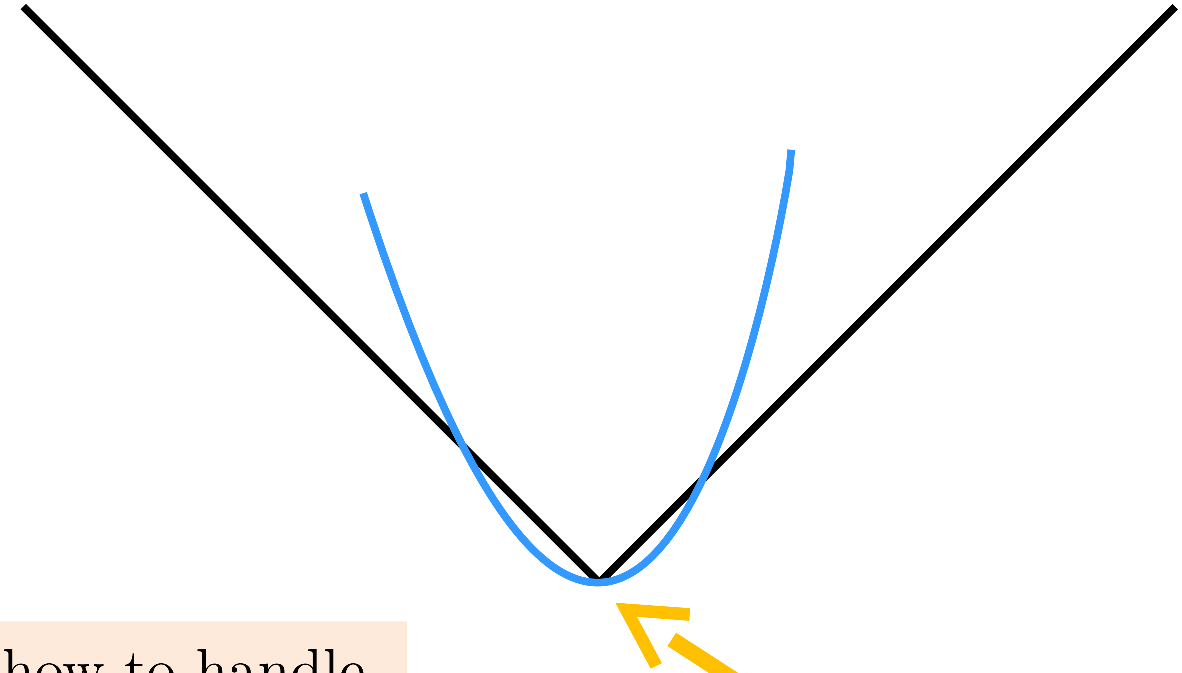
If $f : \mathbb{R}^n \to \mathbb{R} \cup \{\infty\}$ is L–smooth then

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



Smoothness: Convex counter-example

$$f(w) = ||w||_1 = \sum_{i=1}^{n} |w_i|$$



We'll see how to handle this problem next class

Does not fit.
Not smooth

Insight into Gradient Descent

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

Minimizing the upper bound in w we get:

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

Insight into Gradient Descent

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

Minimizing the upper bound in w we get:

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

$$\Downarrow$$

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

Insight into Gradient Descent

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

Minimizing the upper bound in w we get:

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

A gradient descent step !

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

Insight into Gradient Descent

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

Minimizing the upper bound in w we get:

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

EXE: **If f is L-smooth, show that**

$$f(y - \tfrac{1}{L}\nabla f(y)) - f(y) \leq -\frac{1}{2L}||\nabla f(y)||_2^2, \ \forall y$$

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

where $f(w^*) \leq f(w), \quad \forall w \in \mathbb{R}^n$

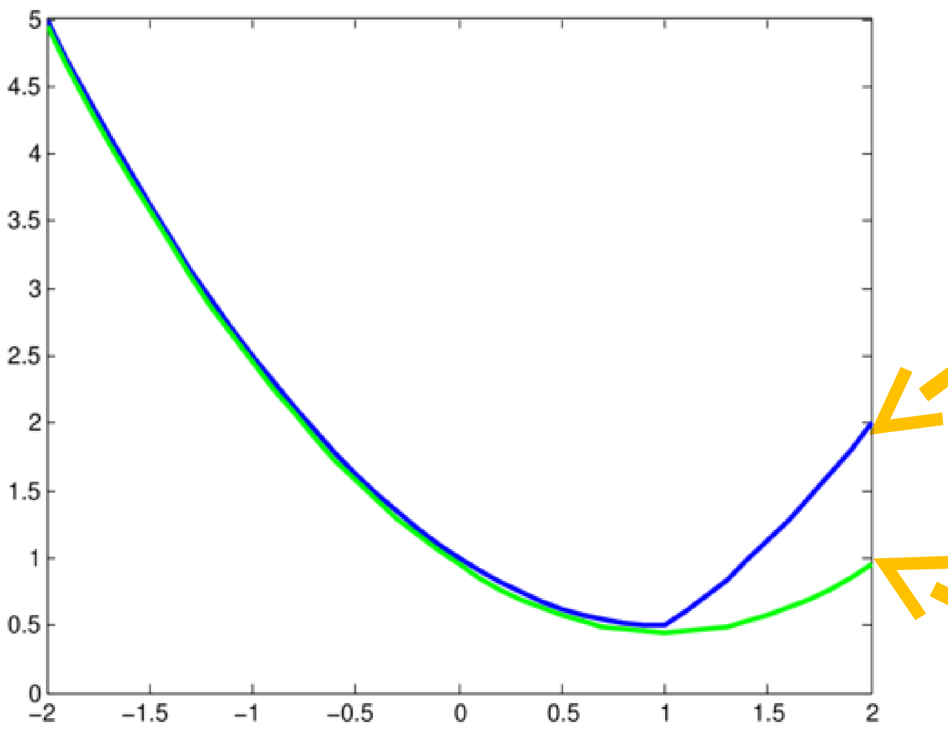
A gradient descent step !

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

Strong convexity

We say $f : \mathbb{R}^n \to \mathbb{R} \cup \{\infty\}$ is μ–strongly convex if

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



Hinge loss + L2

$$\max\{0, 1 - w\} + \frac{1}{2}||w||_2^2$$

Quadratic lower bound

Strong convexity

We say $f : \mathbb{R}^n \to \mathbb{R} \cup \{\infty\}$ is μ–strongly convex if

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

$$d^\top \nabla^2 f(w) d \geq \mu ||d||^2, \quad \forall d \in \mathbb{R}^n$$

EXE: Using that

$$\sigma_{\min}(X)^2 ||d||_2^2 \leq ||X^\top d||_2^2$$

Show that

$$\tfrac{1}{2}||X^\top w - b||_2^2 \text{ is } \sigma_{\min}(X)^2\text{–strongly convex}$$

Convergence GD strongly convex

Theorem

Let f be μ-strongly convex and L-smooth.

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

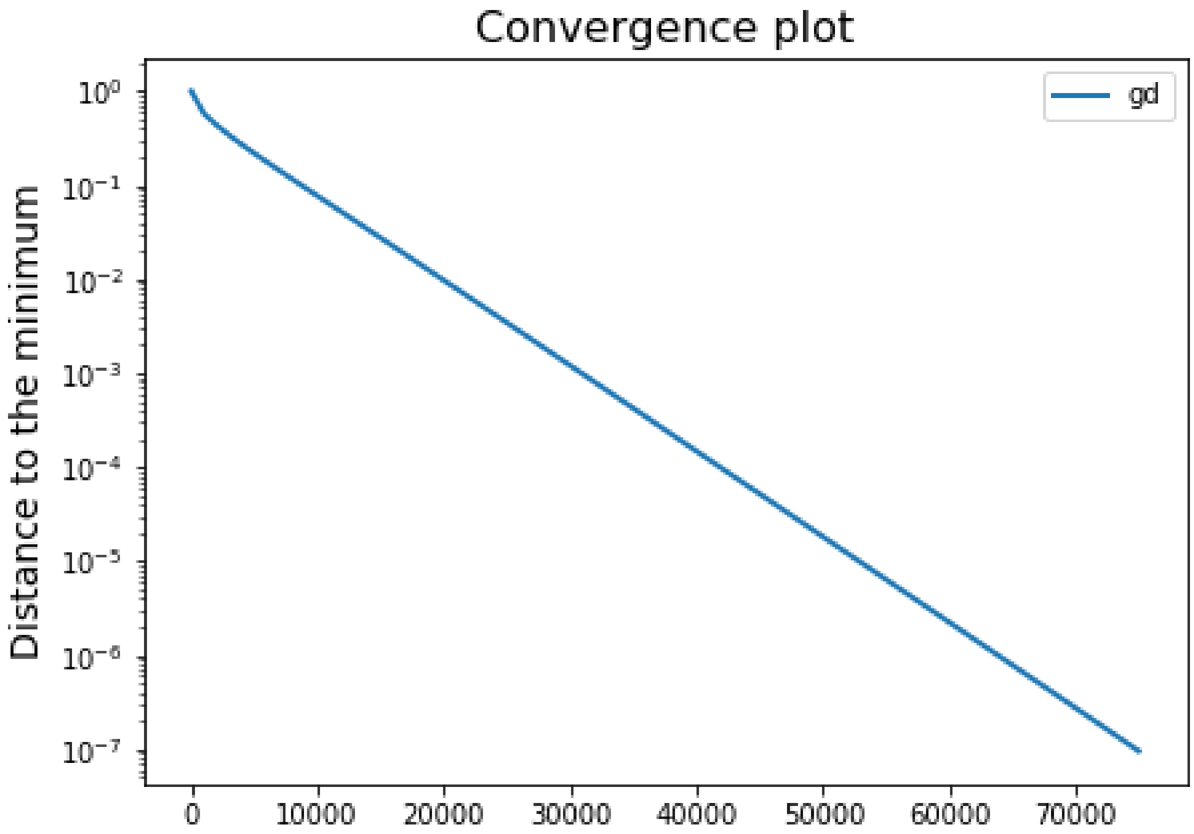
Where

$$w^{t+1} = w^t - \frac{1}{L}\nabla f(w^t), \quad \text{for } t = 1, \ldots, T$$

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

EXE: Solve the questions in complexity_rates_exe.pdf

Gradient Descent Example: logistic



$$y\text{–axis} = \frac{||w^t - w^*||_2^2}{||w^1 - w^*||_2^2} \quad \Rightarrow \quad \log\left(\frac{||w^t - w^*||_2^2}{||w^1 - w^*||_2^2}\right) \leq t \log\left(1 - \frac{\mu}{L}\right)$$

Proof Convergence GD strongly convex + smooth

Proof on board

$$||w^{t+1} - w^*||_2^2 = ||w^t - w^* - \frac{1}{L}\nabla f(w^t)||_2^2$$

$$= ||w^t - w^*||_2^2 + \frac{2}{L}\langle \nabla f(w^t), w^* - w^t\rangle + \frac{1}{L^2}||\nabla f(w^t)||_2^2$$

Now smoothness gives

$$f(w^*) - f(w) \leq -\frac{1}{2L}||\nabla f(w)||_2^2$$

$$\Downarrow$$

$$||\nabla f(w)||_2^2 \leq 2L(f(w) - f(w^*))$$

And strong convexity gives

$$f(w^*) \geq f(w) + \langle \nabla f(w), w^* - w\rangle + \frac{\mu}{2}||w - w^*||^2$$

$$\Downarrow$$

$$\langle \nabla f(w), w^* - w\rangle \leq -(f(w) - f(w^*)) - \frac{\mu}{2}||w - w^*||^2$$

Convergence GD for smooth + convex

Theorem

Let f be convex and L-smooth.

$$f(w^t) - f(w^*) \le \frac{2L||w^1 - w^*||_2^2}{t-1} = O\left(\frac{1}{t}\right).$$

Where

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

$\Rightarrow$ for $\dfrac{f(w^T) - f(w^*)}{||w^1 - w^*||_2^2} \le \epsilon$ we need $T \ge \dfrac{2L}{\epsilon} = O\left(\dfrac{1}{\epsilon}\right)$

Convex and Smooth Properties

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

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

Proof

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

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

Proof

$$f(y) - f(x) = \overbrace{f(y) - f(z)}^{\text{Use convexity}} + \overbrace{f(z) - f(x)}^{\text{Use smoothness}}$$

Convex and Smooth Properties

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

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

Proof

$$f(y) - f(x) = \overbrace{f(y) - f(z)}^{\text{Use convexity}} + \overbrace{f(z) - f(x)}^{\text{Use smoothness}}$$

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

Convex and Smooth Properties

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

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

Proof

$$\begin{aligned} f(y) - f(x) &= \overbrace{f(y) - f(z)}^{\text{Use convexity}} + \overbrace{f(z) - f(x)}^{\text{Use smoothness}} \\ &\leq \langle \nabla f(y), y - z \rangle + \langle \nabla f(x), z - x \rangle + \frac{L}{2} ||z - x||^2 \end{aligned}$$

Then minimize in z and insert back in minima.

Proof of GD smooth + convex theorem

$$||w^{t+1} - w^*||_2^2 = ||w^t - w^* - \frac{1}{L}\nabla f(w^t)||_2^2$$

Use co-coercivity

$$= ||w^t - w^*||_2^2 + \frac{2}{L}\overbrace{\langle \nabla f(w^t), w^* - w^t\rangle} + \frac{1}{L^2}||\nabla f(w^t)||_2^2$$

Co-coercivity

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

With $y = w^*$ gives $\quad \langle \nabla f(w), w^* - w\rangle \leq -\frac{1}{L}||\nabla f(w)||_2$

Inserting above show decreasing

$$||w^{t+1} - w^*||_2^2 \leq ||w^t - w^*||_2^2 - \frac{1}{L^2}||\nabla f(w^t)||_2^2$$

smoothness gives

$$f(w^{t+1}) - f^* \leq f(w^t) - f^* - \frac{1}{2L}||\nabla f(w^t)||_2^2$$

Combine with convexity

$$\begin{aligned} f(w^t) - f(w^*) &\leq \langle \nabla f(w^t), w^t - w^*\rangle \\ &\leq ||\nabla f(w^t)||_2 ||w^t - w^*||_2 \end{aligned}$$

Acceleration and lower bouds

The Accelerated gradient method

$$\min_{w \in \mathbb{R}^d} f(w)$$

Accelerated gradient

Set $w^1 = 0 = y^1, \kappa = L/\mu$

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

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

$$w^{t+1} = \left(1 + \frac{\sqrt{\kappa}-1}{\sqrt{\kappa}+1}\right) y^{t+1} - \frac{\sqrt{\kappa}-1}{\sqrt{\kappa}+1} w^t$$

Output w^{T+1}

The Accelerated gradient method

$$\min_{w \in \mathbb{R}^d} f(w)$$

Accelerated gradient

Set $w^1 = 0 = y^1, \kappa = L/\mu$

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

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

$$w^{t+1} = \left(1 + \frac{\sqrt{\kappa}-1}{\sqrt{\kappa}+1}\right) y^{t+1} - \frac{\sqrt{\kappa}-1}{\sqrt{\kappa}+1} w^t$$

Output w^{T+1}

Weird extrapolation, but it works

Convergence lower bounds strongly convex

Theorem (Nesterov)

For any optimization algorithm where

$$w^{t+1} \in w^t + \mathrm{span}\left(\nabla f(w^1), \nabla f(w^2), \ldots, \nabla f(w^t)\right)$$

There exists a function $f(w)$ that is L–smooth and μ–strongly convex such that

$$f(w^T) - f(w^*) \geq \frac{\mu}{2}\left(1 - \frac{2}{\sqrt{\kappa + 1}}\right)^{2(T-1)} ||w^1 - w^*||_2^2$$

$$= O\left(\left(1 - \frac{1}{\sqrt{\kappa}}\right)^{2T}\right).$$

Accelerated gradient has this rate

Yuri Nesterov (1998), Springer Publishing, **Introductory Lectures on Convex Optimization: A Basic Course**

Convergence lower bounds strongly convex

Theorem (Nesterov)

For any optimization algorithm where

$$w^{t+1} \in w^t + \text{span}\left(\nabla f(w^1), \nabla f(w^2), \ldots, \nabla f(w^t)\right)$$

There exists a function $f(w)$ that is L–smooth and μ–strongly convex such that

$$f(w^T) - f(w^*) \geq \frac{\mu}{2}\left(1 - \frac{2}{\sqrt{\kappa + 1}}\right)^{2(T-1)} ||w^1 - w^*||_2^2$$

$$= O\left(\left(1 - \frac{1}{\sqrt{\kappa}}\right)^{2T}\right).$$

Accelerated gradient has this rate

Yuri Nesterov (1998), Springer Publishing, **Introductory Lectures on Convex Optimization: A Basic Course**

Convergence lower bounds convex

Theorem (Nesterov)

For any optimization algorithm where

$$w^{t+1} \in w^t + \mathrm{span}\left(\nabla f(w^1), \nabla f(w^2), \ldots, \nabla f(w^t)\right)$$

There exists a function $f(w)$ that is L–smooth and convex

such that

$$\min_{i=1,\ldots,T} f(w^i) - f(w^*) \geq \frac{3L\|w^1 - w^*\|_2^2}{32(T+1)^2} = O\left(\frac{1}{T^2}\right).$$

Yuri Nesterov (1998), Springer Publishing, **Introductory Lectures on Convex Optimization: A Basic Course**

Convergence lower bounds convex

Theorem (Nesterov)

For any optimization algorithm where

$$w^{t+1} \in w^t + \mathrm{span}\left(\nabla f(w^1), \nabla f(w^2), \ldots, \nabla f(w^t)\right)$$

There exists a function $f(w)$ that is L–smooth and convex such that

$$\min_{i=1,\ldots,T} f(w^i) - f(w^*) \geq \frac{3L||w^1 - w^*||_2^2}{32(T+1)^2} = O\left(\frac{1}{T^2}\right).$$

Yuri Nesterov (1998), Springer Publishing, **Introductory Lectures on Convex Optimization: A Basic Course**

Exercises !

Solve ridge_reg_exe.pdf

Exercises !

Solve ridge_reg_exe.pdf