Gradient descent : theory and practice

Pierre Ablin

Machine learning task

Finite Sum Training Problem

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

Today: assume that f is differentiable and L-smooth

! ∇f exists

Iterative minimization

Finite Sum Training Problem

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

Usually cannot solve this in closed form : $\quad w^* = \ldots$

Idea: start from initial guess w^0 and try to find a new, better point. Iterative process $\quad w^0 \rightarrow w^1 \rightarrow \ldots$

Gradient descent : basic idea

Given w^0, look for w^1 as $w^1 = w^0 + d$ where d is a small displacement.

Ideally: $$d \in \arg\min_{d \in \mathbb{R}^p} f(w^0 + d)$$

Just as hard as the original problem :(

Solution: $$d \in \arg\min_{\|d\| \leq \varepsilon} f(w^0 + d)$$

Q: as ε goes to 0, what is the limit of d ?

Gradient descent algorithm

Init : Select initial guess $w^0, \ \rho > 0$

For t = 0, 1, ..., T:

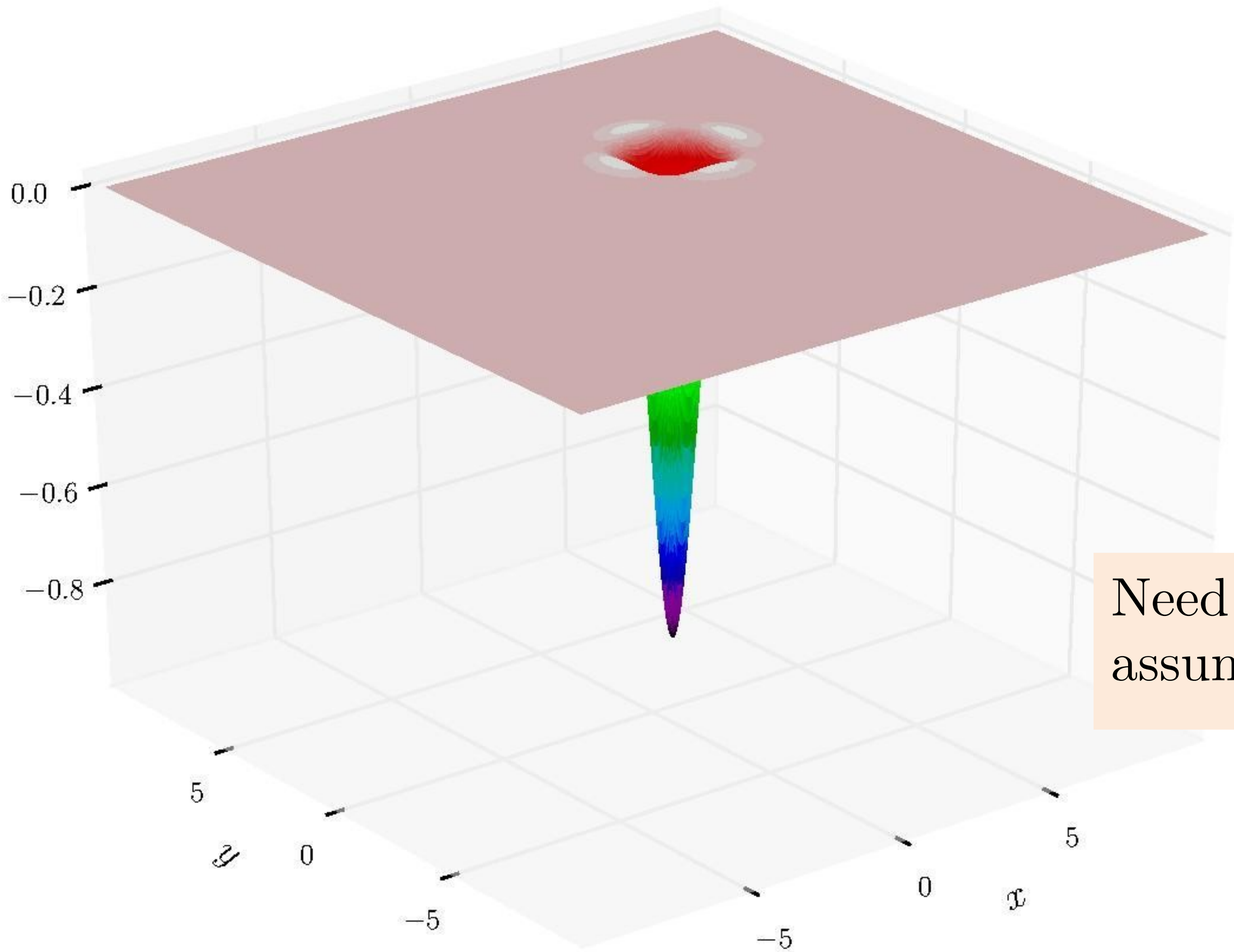
- Update $w^{t+1} = w^t - \rho \nabla f(w^t)$

Return : w^{T+1}

Questions :

- Does it converge? In which sense?
- At which speed?
- Choice of ρ ?

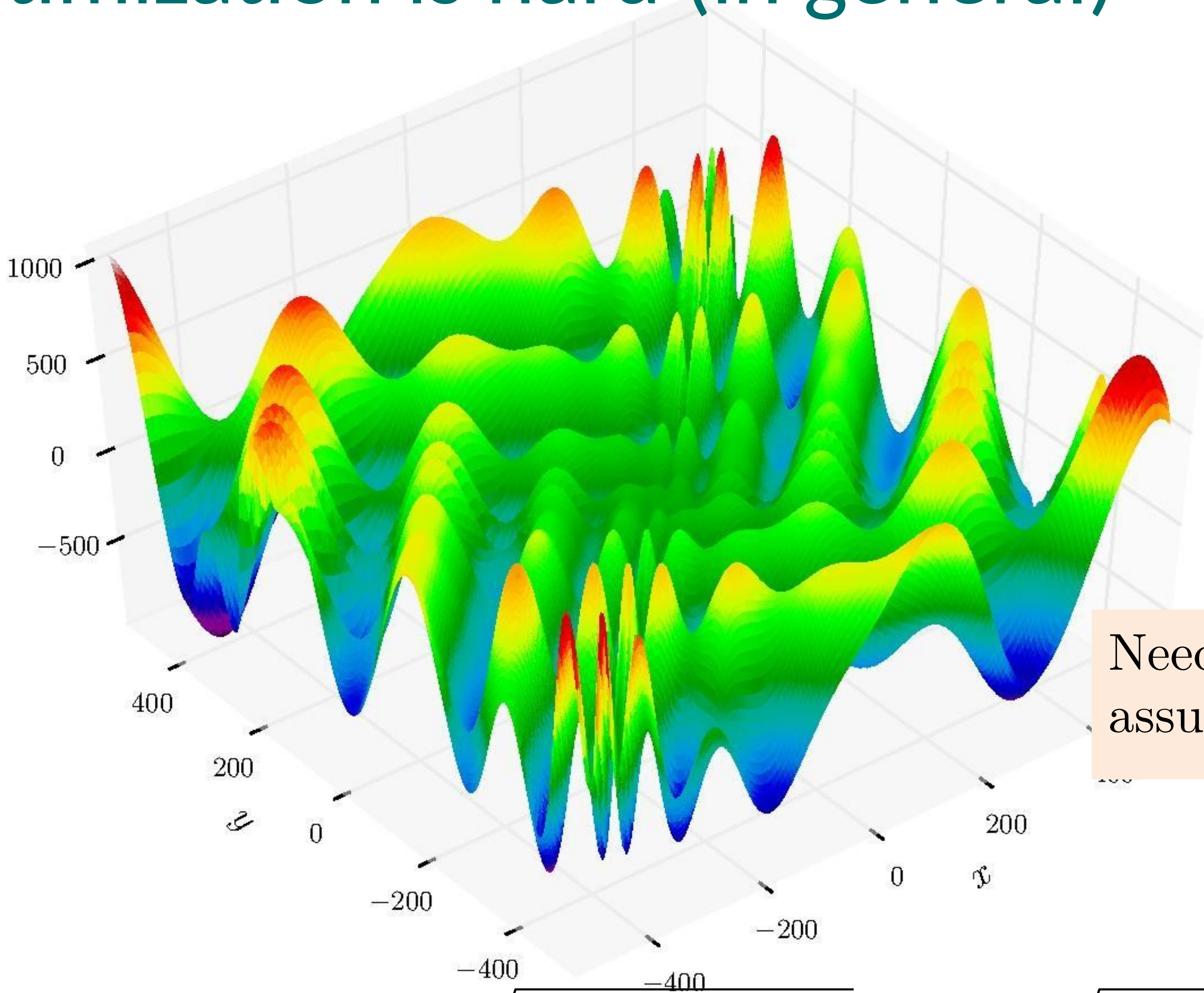
Optimization is hard (in general)



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!

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

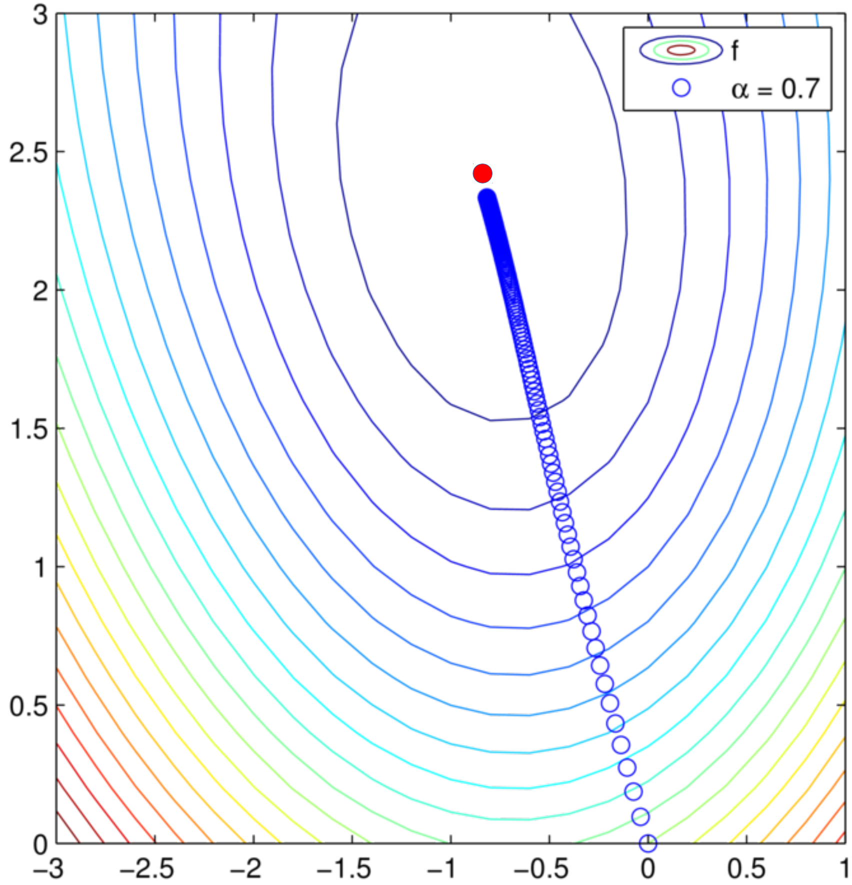
Gradient Descent Example

A Logistic Regression problem using the fourclass labelled data from LIBSVM

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

Logistic Regression

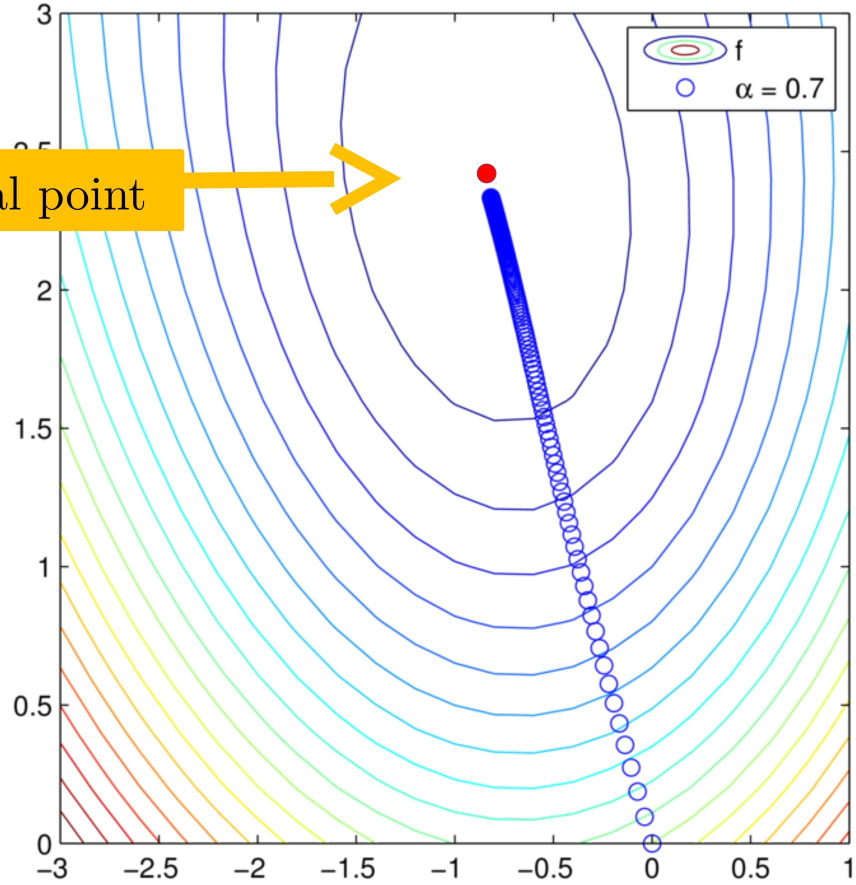
$$\min_{w\in\mathbf{R}^d} \frac{1}{n}\sum_{i=1}^{n} \ln(1+e^{-y^i\langle w,x^i\rangle})+\lambda||w||_2^2$$



Can we prove that this always works?

Gradient Descent Example

Optimal point



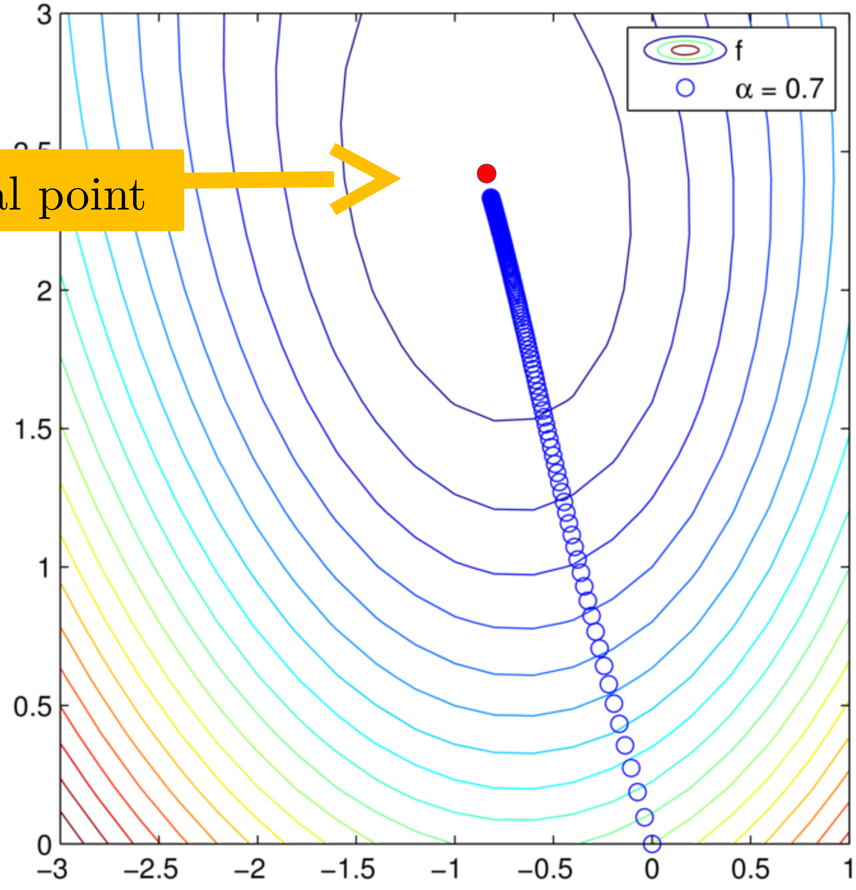
A Logistic Regression problem using the fourclass labelled data from LIBSVM
$(n,\ d) = (862,2)$

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

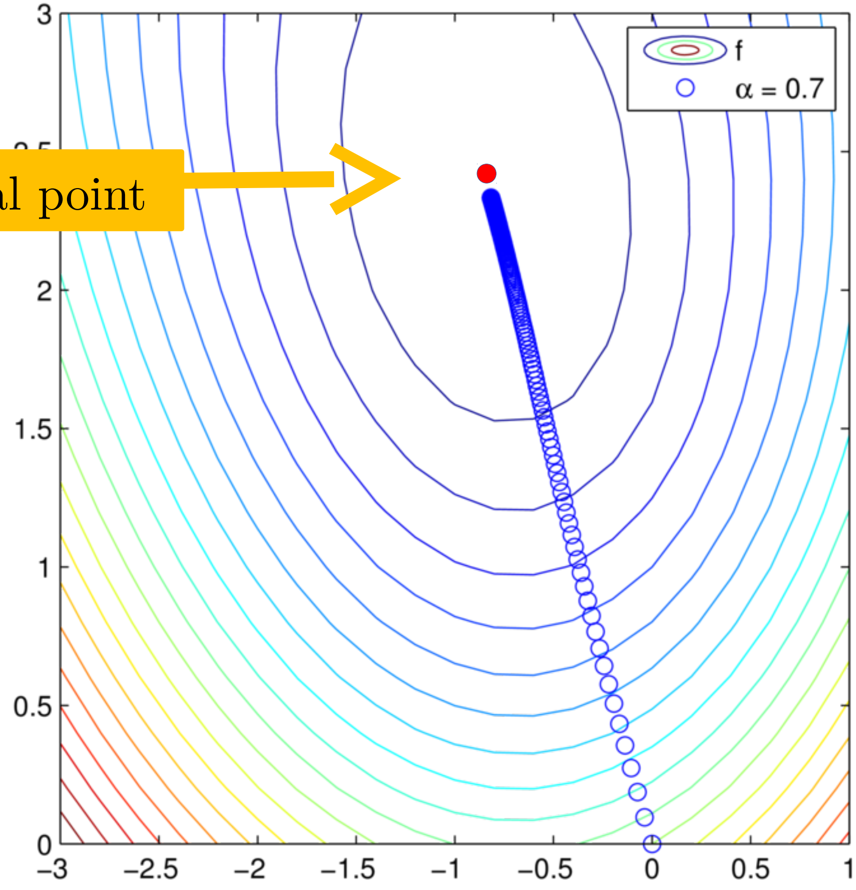
$$\min_{w \in \mathbf{R}^d} \frac{1}{n} \sum_{i=1}^{n} \ln(1 + e^{-y^i \langle w, x^i \rangle}) + \lambda ||w||_2^2$$

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

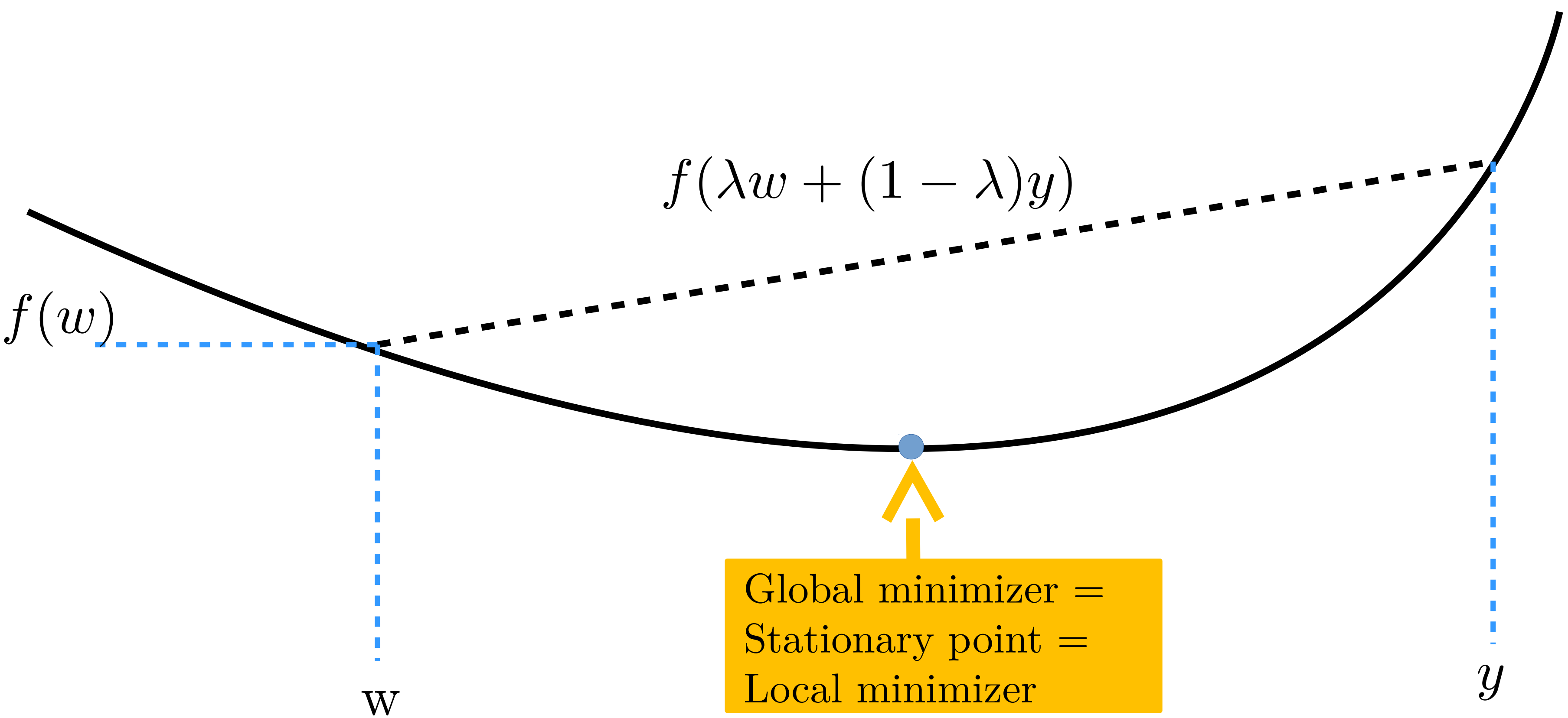
Specialize

Convex and smooth training problems

Convexity

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

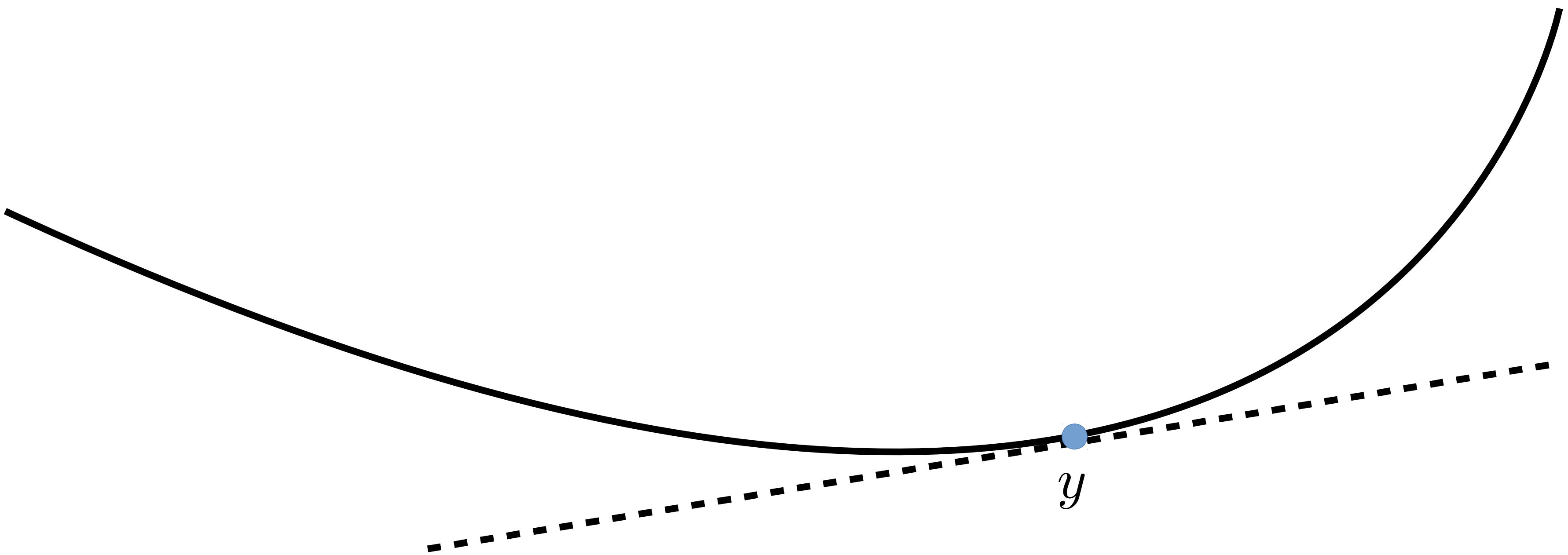
$$f(\lambda w + (1-\lambda)y) \leq \lambda f(w) + (1-\lambda) f(y), \quad \forall w, y \in C, \lambda \in [0, 1]$$



Convexity

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

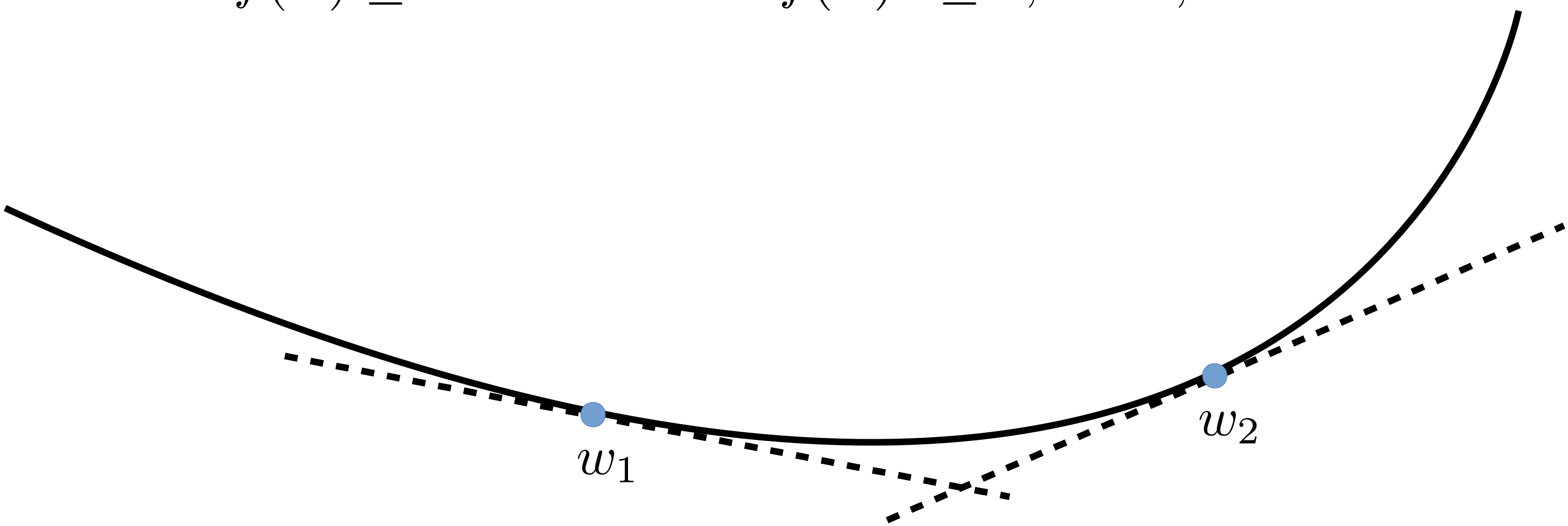
$$f(w) \geq f(y) + \langle \nabla f(y), w - y \rangle$$



Convexity

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

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



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

Main Advantage of Convexity

Nice Property

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

Main Advantage of Convexity

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

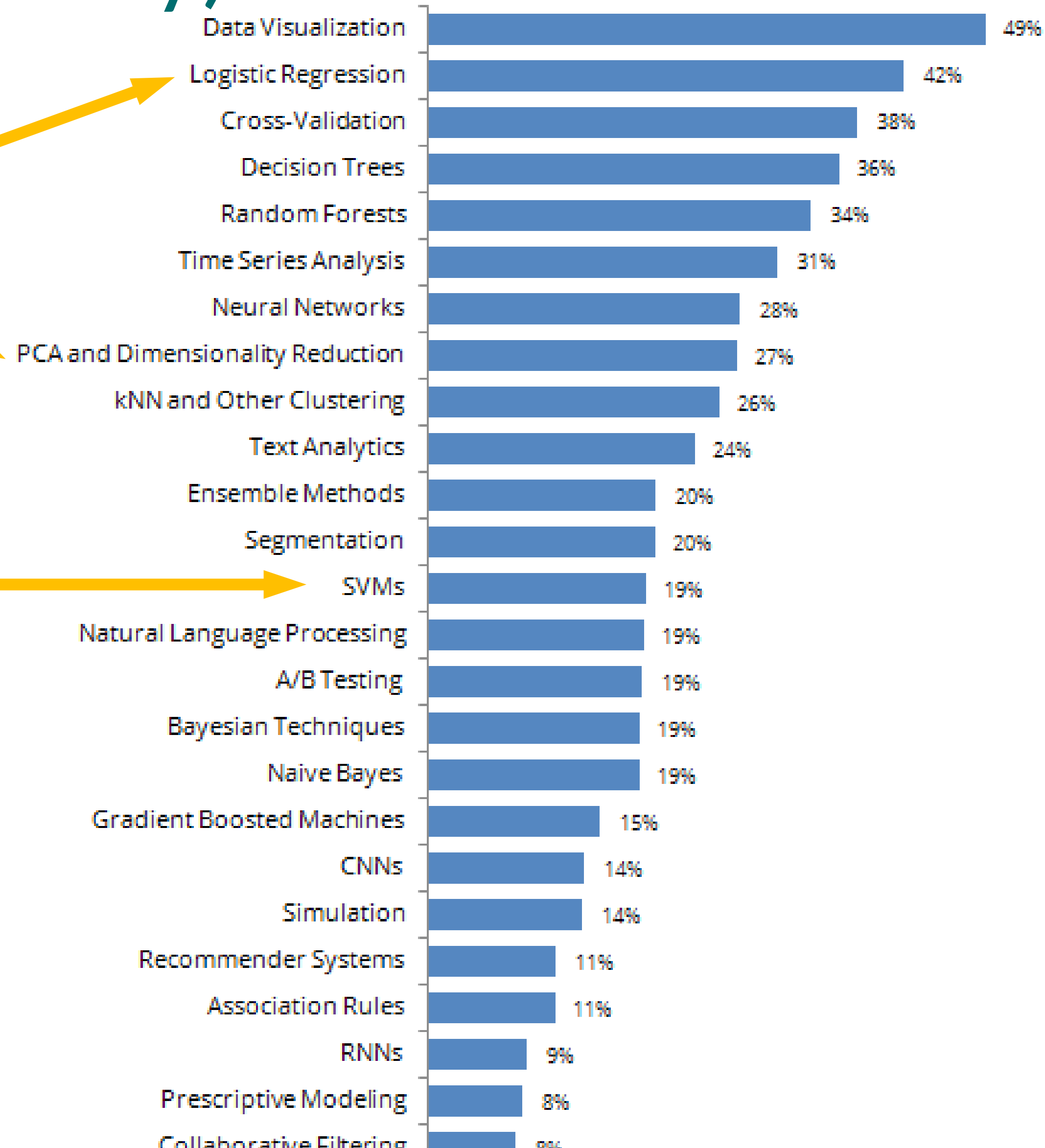
$$\text{If } \; f(w) \geq f(y) + \langle \nabla f(y), w - y \rangle, \quad \forall w, y \in \mathbb{R}^d$$

then Nice Property holds

PROOF: Choose $y = w^*$

Data science methods most used (Kaggle 2017 survey)

Convex Optimization problems



Convexity: Examples

Extended-value extension:

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

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

Norms and squared norms:

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

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

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 L-smooth if

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

Smoothness

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

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

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

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

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

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

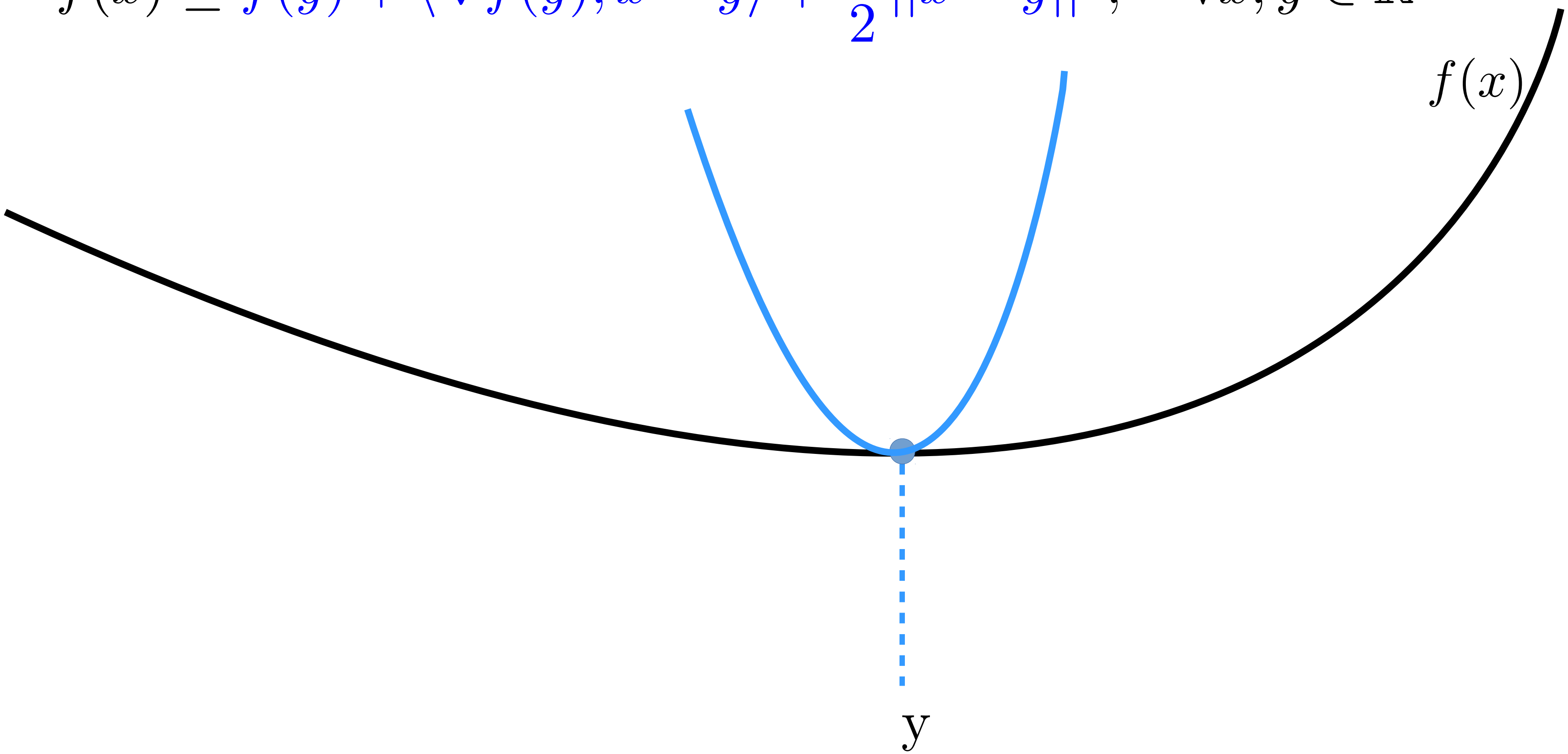
EXE: determine the smoothness constants of

$$f(w) := \tfrac{1}{2}||Xw - b||_2^2 \ \text{ for } \ X \in \mathbb{R}^{n\times p}, \ \ b \in \mathbb{R}^n$$

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: Examples

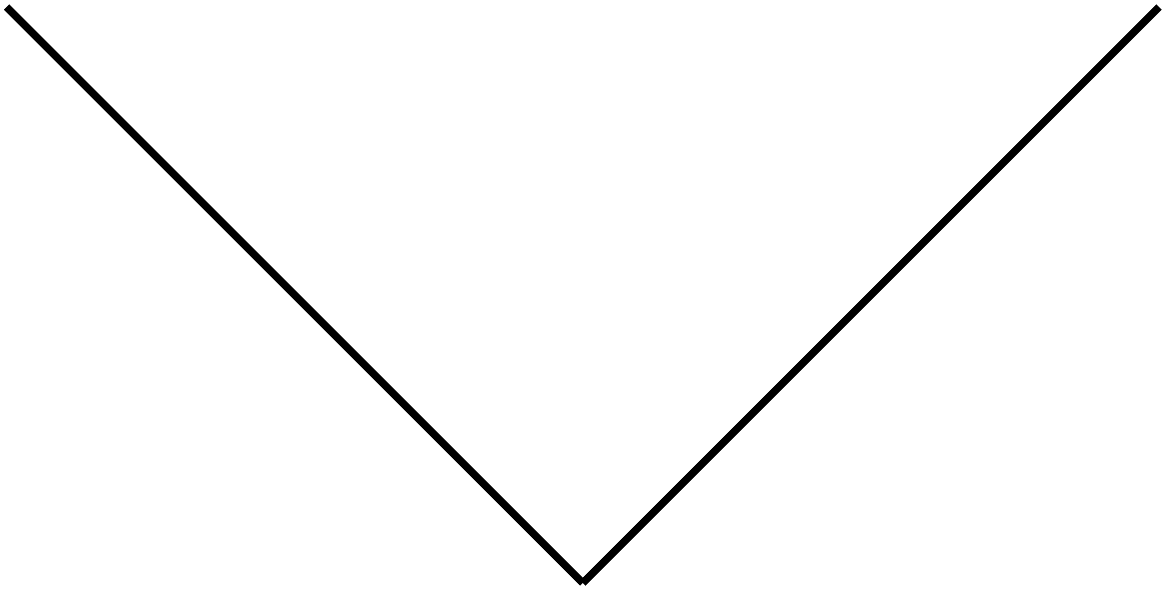
Convex quadratics: $x \mapsto x^\top A x + b^\top x + c$

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

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

Smoothness: Convex counter-example

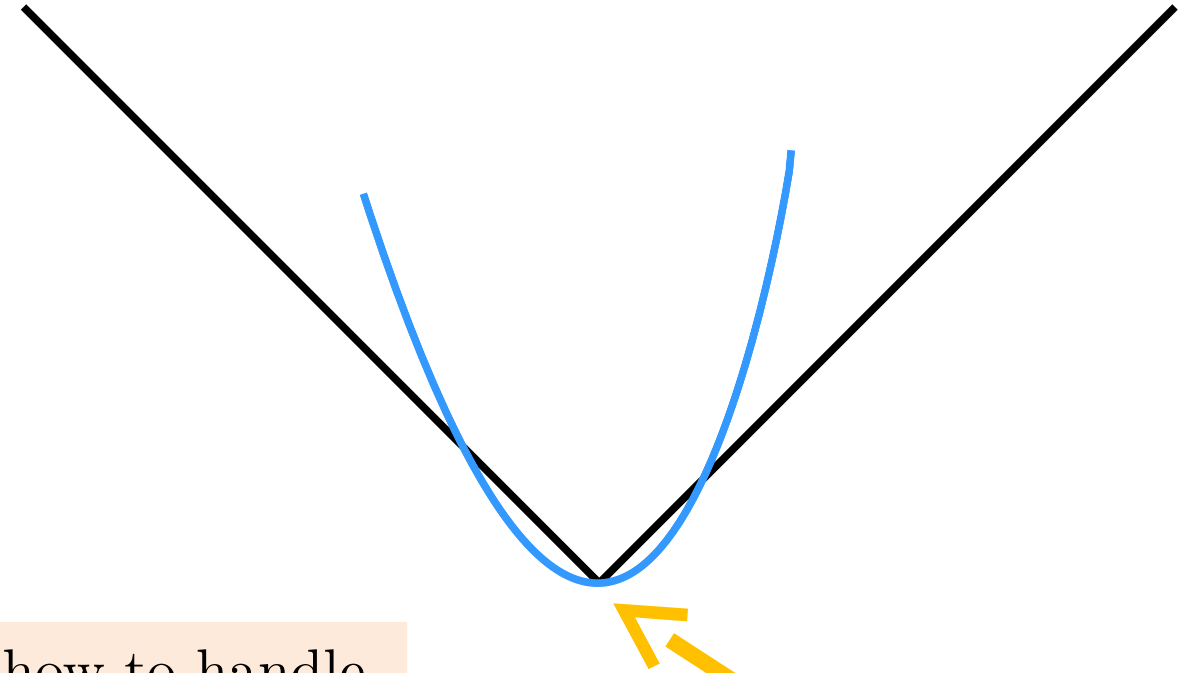
$$f(w) = ||w||_1 = \sum_{i=1}^{n} |w_i|$$



We'll see how to handle this problem next class

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 using Smoothness

$$f(w) \leq f(w^0) + \langle \nabla f(w^0), w - w^0 \rangle + \frac{L}{2}||w - w^0||^2$$

Q: what is the minimizer of the upper bound in w?

Insight into Gradient Descent using Smoothness

$$\boxed{f(w) \leq f(w^0) + \langle \nabla f(w^0), w - w^0 \rangle + \frac{L}{2}||w - w^0||^2}$$

Minimizing the upper bound in w we get:

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

A gradient descent step !

$$w = w^0 - \frac{1}{L} \nabla f(w^0)$$

Insight into Gradient Descent using Smoothness

$$\boxed{f(w) \leq f(w^0) + \langle \nabla f(w^0), w - w^0 \rangle + \frac{L}{2}||w - w^0||^2}$$

Minimizing the upper bound in w we get:

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

Smoothness Lemma (EXE):

If f is L-smooth, show that

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

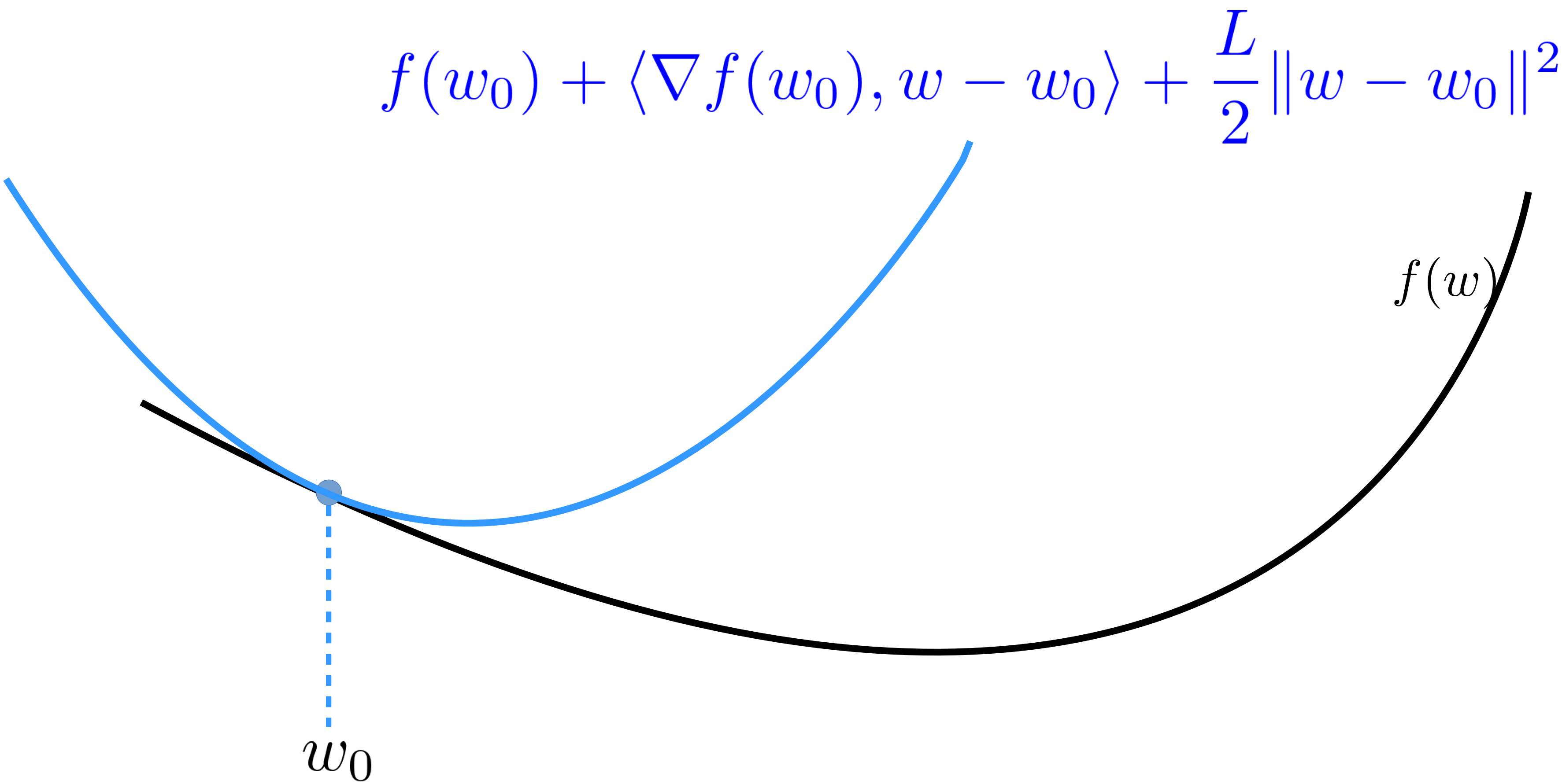
$$f(w^*) - f(w) \leq -\frac{1}{2L}||\nabla f(w)||_2^2, \quad \forall w \in \mathbb{R}^n$$

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

$$w = w^0 - \frac{1}{L}\nabla f(w^0)$$

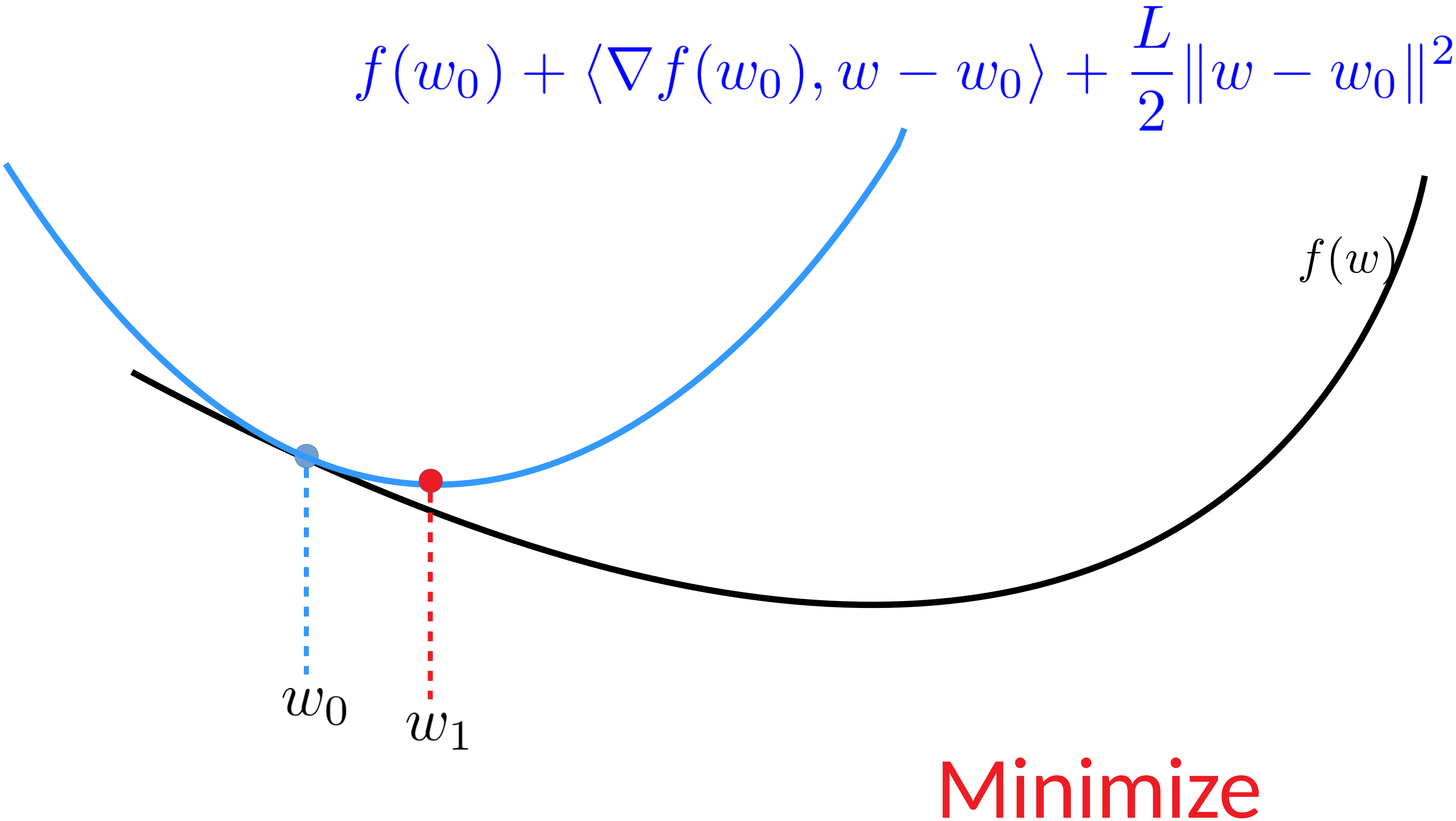
A gradient descent step !

Gradient descent as a majorization-minimization procedure

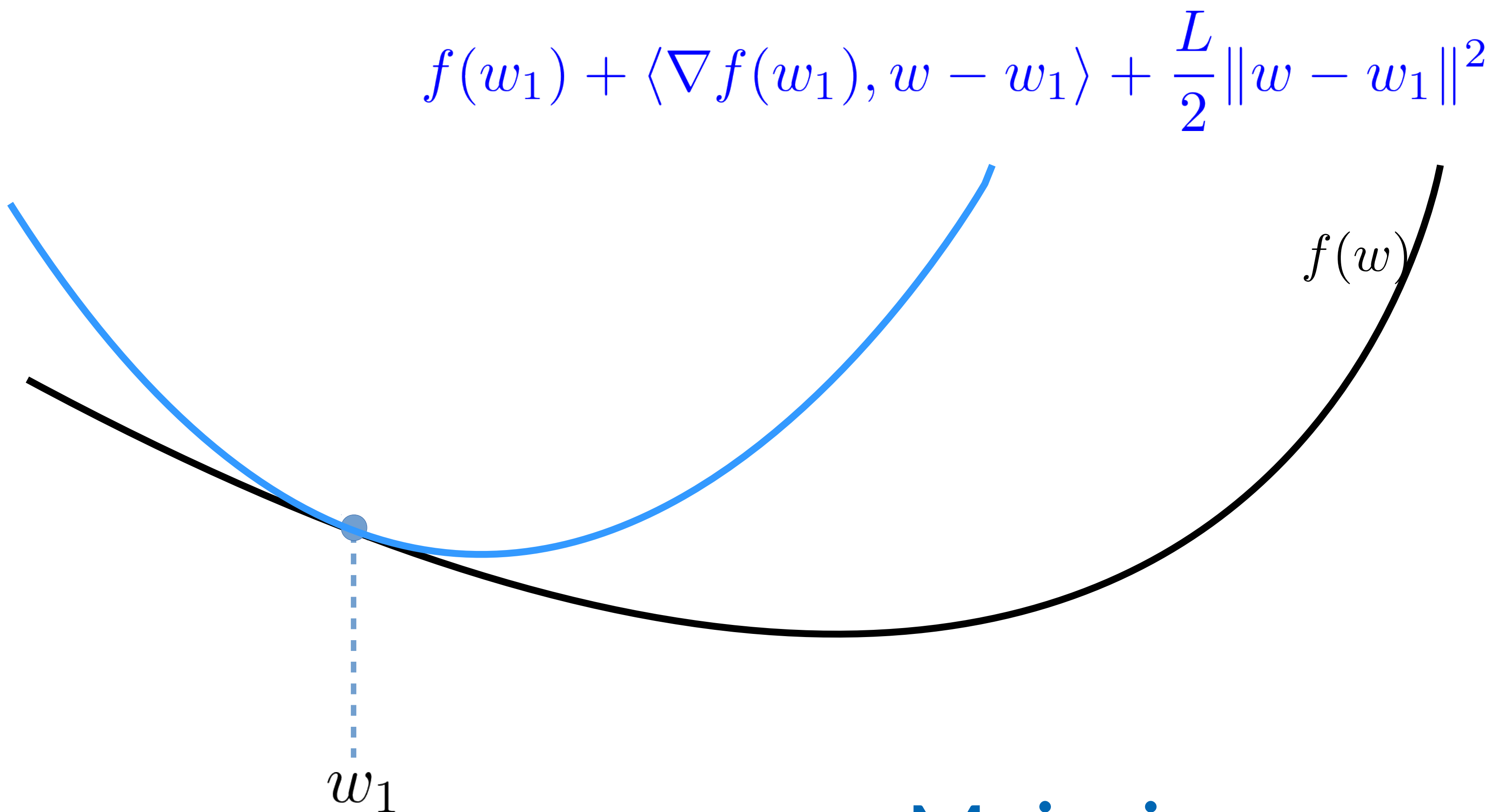


Majorize

Gradient descent as a majorization-minimization procedure

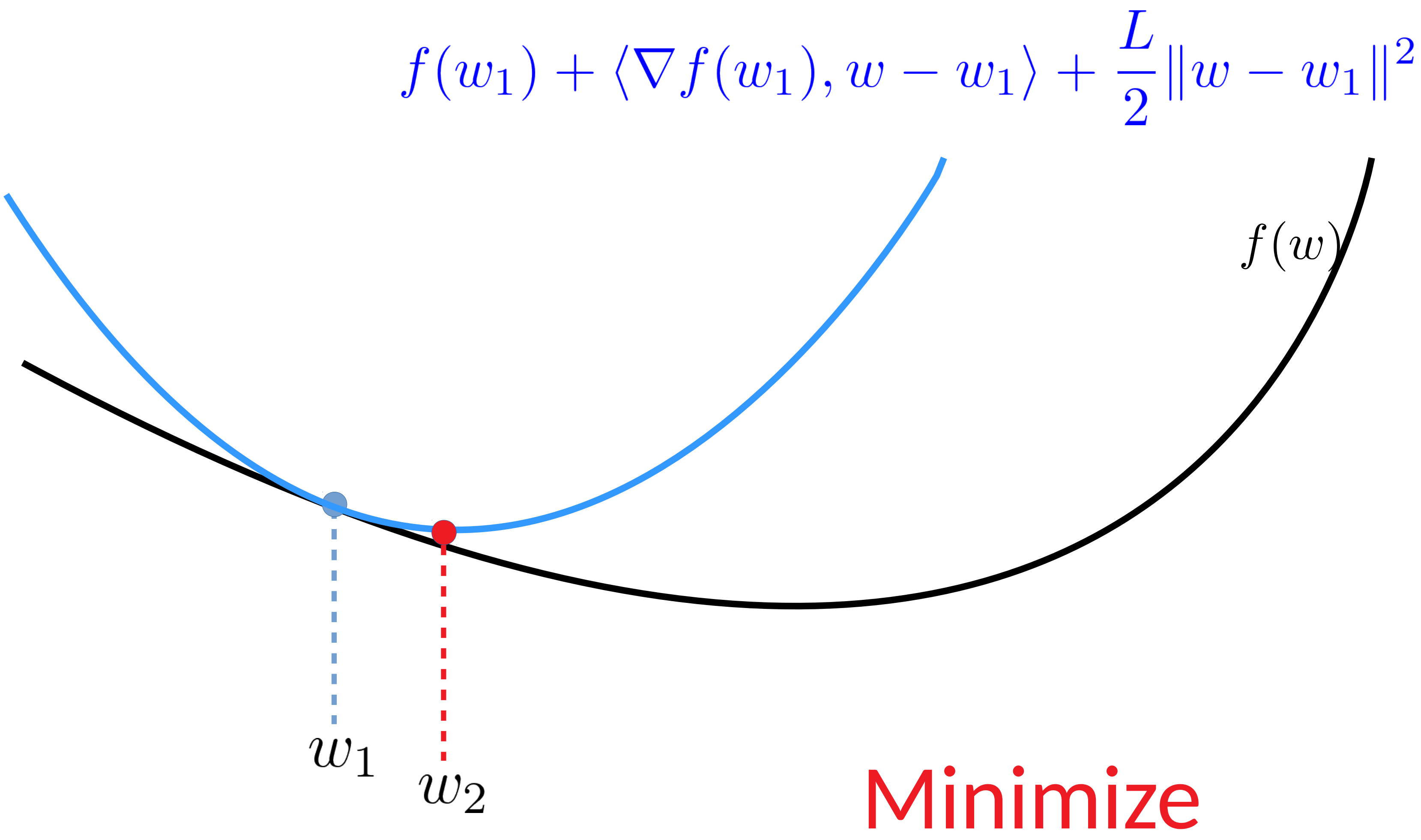


Gradient descent as a majorization-minimization procedure

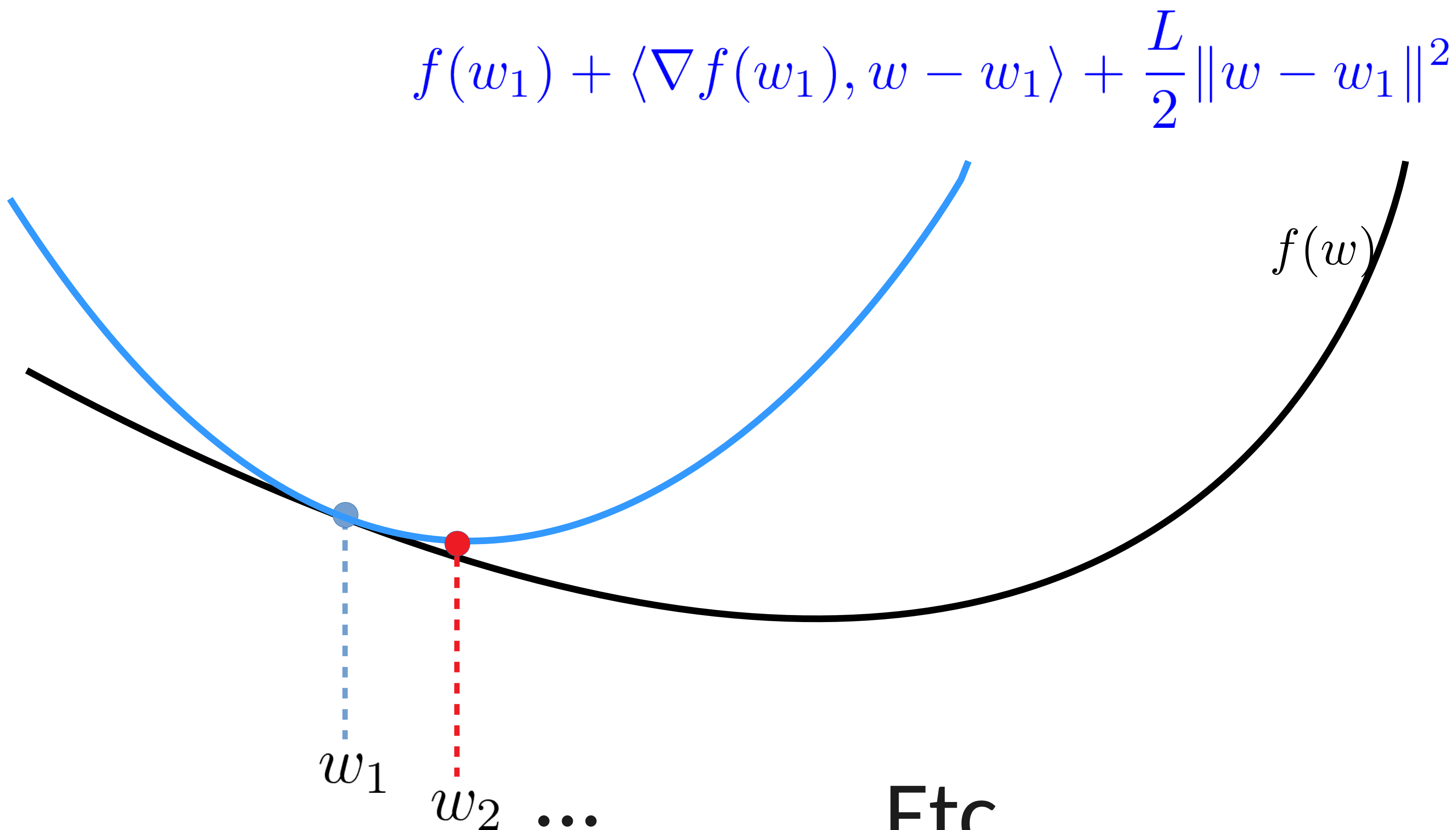


Majorize

Gradient descent as a majorization-minimization procedure



Gradient descent as a majorization-minimization procedure



Etc...

Convergence analysis

A note on convergence

Theorem: Let f convex and L-smooth. Then, the iterates w^t of gradient descent with step $\rho \in [0, 2/L[$ verify

$$\lim_{t \to +\infty} w^t = w^* \text{ with } w^* \in \arg\min f$$

A note on convergence

Theorem: Let f convex and L-smooth. Then, the iterates w^t of gradient descent with step $\rho \in [0, 2/L[$ verify

$$\lim_{t \to +\infty} w^t = w^* \text{ with } w^* \in \arg\min f$$

No convergence rate, i.e. a bound on $\|w^t - w^*\|$

A note on convergence

Theorem: Let f convex and L-smooth. Then, the iterates w^t of gradient descent with step $\rho \in [0, 2/L[$ verify

$$\lim_{t \to +\infty} w^t = w^* \text{ with } w^* \in \arg\min f$$

No convergence rate, i.e. a bound on $\|w^t - w^*\|$

Proof: Uses difficult Cauchy sequences arguments (show that any subsequence of w^t go to w^*)

Convergence rates: smooth case

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

Gradient descent with step $\rho = \frac{1}{L}$:

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

$$\sum_{t=0}^{T} \|\nabla f(w^t)\|^2 \leq 2L(f(w^0) - f(w^*)), \;\; \forall T > 0$$

Q: what does it mean?

Convergence rates: smooth case

Theorem : if f is L-smooth, the iterates of gradient descent verify

$$\nabla f(w^t) \to 0$$

$$\inf_{t \leq T} \|\nabla f(w^t)\|^2 \leq \frac{2L}{T}(f(w^0) - f(w^*))$$

Convergence speed

Slow convergence

Say $2L(f(w^0) - f(w^*)) = 1$

In order to have $\inf_{t \leq T} \|\nabla f(w^t)\|^2 \leq 10^{-4}$

Need 10^4 iterations...

Convergence rates: smooth + convex case

Fundamental lemma:

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

If f convex:

$$f(w^t) - f^* \leq \langle \nabla f(w^t), w^t - w^* \rangle$$

Together:

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

Convergence rates: smooth + convex case

If f is convex and L-smooth :

$$f(w^t) - f^* \leq \frac{L}{2T}\|w^0 - w^*\|^2$$

Proof:

Convergence rates: smooth + convex case

If f is convex and L-smooth :

$$f(w^t) - f^* \leq \frac{L}{2T}\|w^0 - w^*\|^2$$

Proof:

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

Convexity gives

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

Convergence rates: smooth + convex case

If f is convex and L-smooth :

$$f(w^t) - f^* \leq \frac{L}{2T}\|w^0 - w^*\|^2$$

Proof:

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

Convexity gives

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

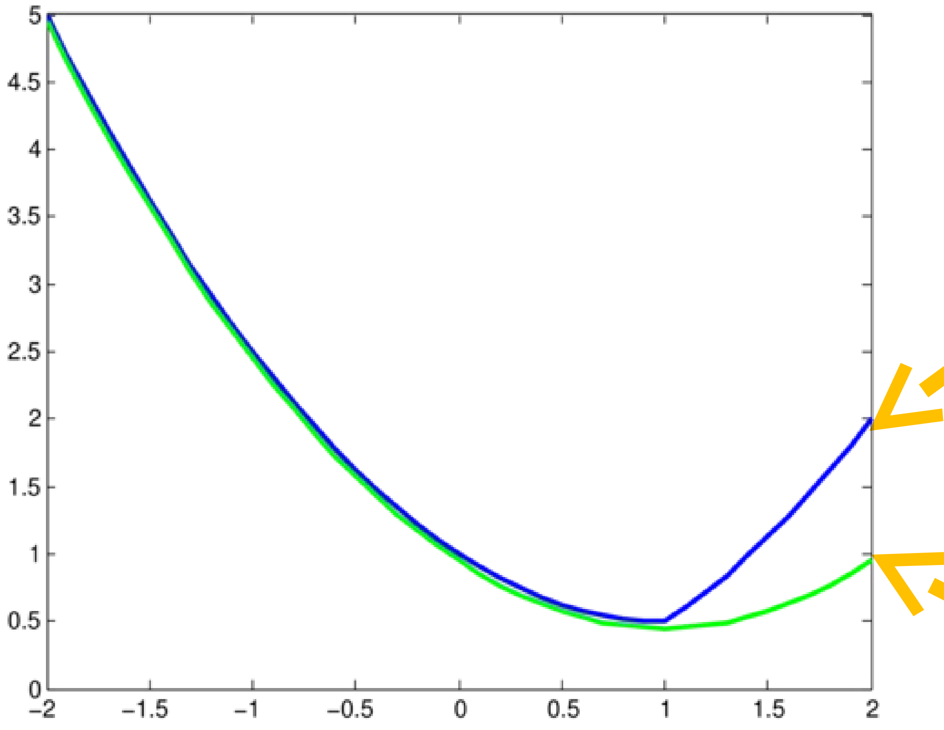
Rest on the board

Faster convergence rates: strongly convex functions

Strong convexity

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

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



Hinge loss + L2
$\max\{0, 1 - w\} + \frac{1}{2}||w||_2^2$

Quadratic lower bound

Strong convexity: equivalent definitions

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

If f is twice differentiable :

$$\nabla^2 f(x) \succeq \mu Id$$

Equivalently: "The eigenvalues of $\nabla^2 f(x)$ are all bounded below by μ"

EXE: determine the strong convexity constant of

$$f(w) := \frac{1}{2}||Xw - b||_2^2 \text{ for } X \in \mathbb{R}^{n \times p}, \;\; b \in \mathbb{R}^n$$

Convergence GD strongly convex

Theorem

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

$$||w^t - w^*||_2^2 \leq \left(1 - \frac{\mu}{L}\right)^t ||w^0 - w^*||_2^2$$

Where

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

$$\Rightarrow \text{for } \frac{||w^T - w^*||_2^2}{||w^0 - w^*||_2^2} \leq \epsilon \text{ we need } T \geq \frac{L}{\mu}\log\left(\frac{1}{\epsilon}\right) = O\left(\log\left(\frac{1}{\epsilon}\right)\right)$$

Convergence GD strongly convex

Theorem

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

$$f(w^t) - f(w^*) \leq (1 - \frac{\mu}{L})^t (f(w^0) - f(w^*))$$

Where

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

Convergence GD strongly convex

Theorem

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

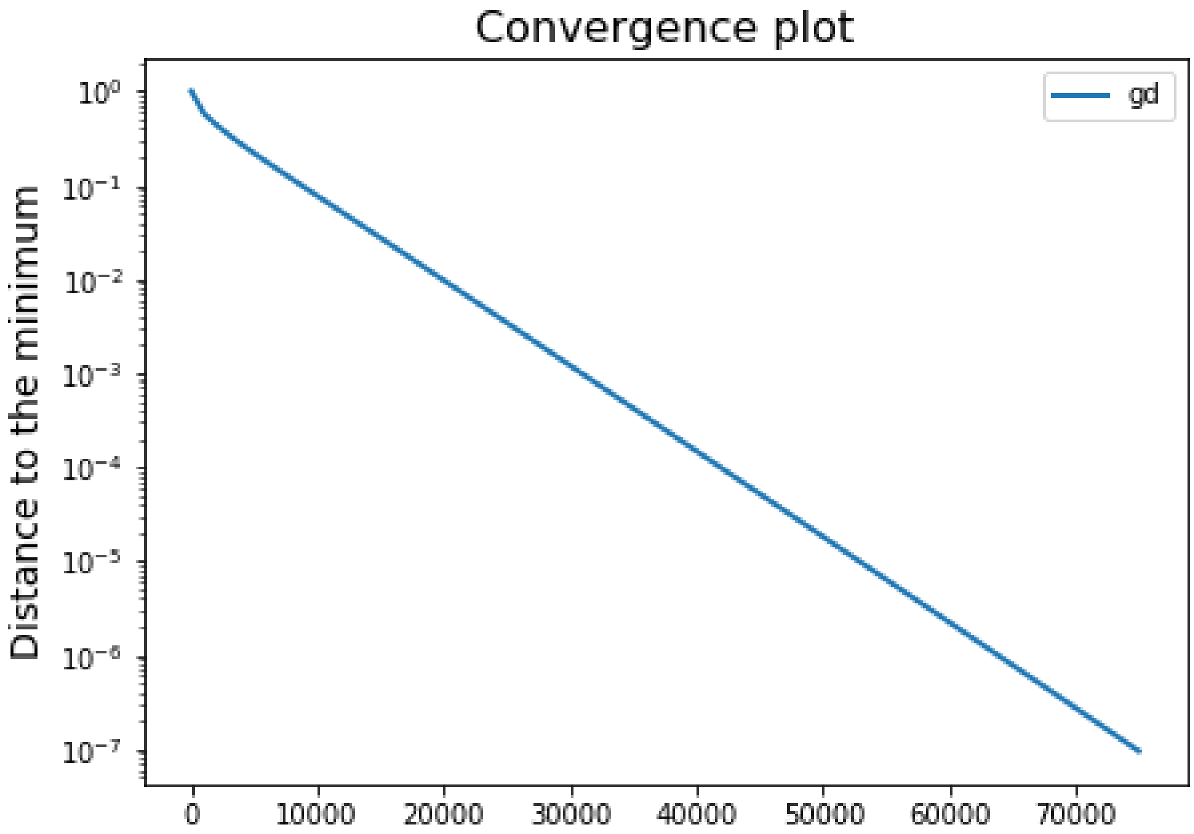
Conditioning $\kappa = \dfrac{\mu}{L}$

$$f(w^t) - f(w^*) \leq (1 - \frac{\mu}{L})^t (f(w^0) - f(w^*))$$

Where

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

Gradient Descent Example: logistic regression



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

Proof Convergence GD strongly convex + smooth

Smoothness

Proof:

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

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

Proof Convergence GD strongly convex + smooth

Smoothness

Proof:

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

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

Polyak-Lojasiewicz (PL) inequality :

Q: show that strong convexity => PL

$$\|\nabla f(w)\|^2 \geq 2\mu(f(w) - f(w^*)), \quad \forall w$$

Proof Convergence GD strongly convex + smooth

Smoothness

Proof:

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

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

Polyak-Lojasiewicz (PL) inequality :

$$\|\nabla f(w)\|^2 \geq 2\mu(f(w) - f(w^*)), \quad \forall w$$

$$f(w^{t+1}) \leq f(w^t) - \frac{\mu}{L}(f(w^t) - f(w^*))$$

$$f(w^t) - f(w^*) \leq (1 - \frac{\mu}{L})^t (f(w^0) - f(w^*))$$

Nesterov acceleration

Smooth + convex case: can we do better ?

If f is convex and L-smooth :

$$f(w^t) - f^* \leq \frac{L}{2T}\|w^0 - w^*\|^2$$

Convergence in 1/T

Smooth + convex case: can we do better ?

If f is convex and L-smooth :

$$f(w^t) - f^* \leq \frac{L}{2T} \|w^0 - w^*\|^2$$

Convergence in 1/T

Nesterov acceleration : another first order algorithm with faster convergence: **convergence in 1 / T^2**

Smooth + convex case: can we do better ?

If f is convex and L-smooth :

$$f(w^t) - f^* \leq \frac{L}{2T} \|w^0 - w^*\|^2$$

Convergence in 1/T

Nesterov acceleration : another first order algorithm with faster convergence: **convergence in 1 / T^2**

Optimal complexity! No first order algorithm can have a convergence faster than 1 / T^2

Nesterov acceleration: principle

Idea: extrapolation

Init : Select initial guess w^0, y^0 and $\alpha^0 = 1$

For t = 0, 1, ..., T:

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

$$\alpha^{t+1} = \frac{1 + \sqrt{1 + 4\alpha^t}}{2}$$

- Extrapolate

$$w^{t+1} = y^{t+1} + \frac{\alpha^t - 1}{\alpha^{t+1}}(y^{t+1} - y^t)$$

Return : w^{T+1}

Nesterov acceleration: principle

Idea: extrapolation

Init : Select initial guess w^0, y^0 and $\alpha^0 = 1$

For t = 0, 1, ..., T:

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

$$\alpha^{t+1} = \frac{1 + \sqrt{1 + 4\alpha^t}}{2}$$

- Extrapolate

$$w^{t+1} = y^{t+1} + \frac{\alpha^t - 1}{\alpha^{t+1}}(y^{t+1} - y^t)$$

Return : w^{T+1}

Go a bit “further” than y^{t+1}

Nesterov acceleration: convergence result

Theorem : if f is convex and L-smooth, the iterates of Nesterov acceleration verify

$$f(w^t) - f^* \leq \frac{2\|y^0 - x^*\|^2}{t(t+1)} = O(\frac{1}{t^2})$$

Examples of smooth machine learning problems

Least squares

Data: $x_1, \dots x_n \in \mathbb{R}^p$, and $y_1, \dots, y_n \in \mathbb{R}$

Assumption: There exists w^* such that

$$y_i \simeq \langle x_i, w^* \rangle$$

Optimization problem: $\min_w f(w) = \frac{1}{n} \sum_{i=1}^{n} (\langle x_i, w \rangle - y_i)^2$

Q: show that we can rewrite $f(w) = \frac{1}{n} \|Xw - y\|^2$

Is the problem convex, smooth? Compute the associated constants

Ridge regression

Problem :

$$\min_w f(w) = \frac{1}{n}\sum_{i=1}^{n}(\langle x_i, w\rangle - y_i)^2$$

Has infinitely many solutions when n < p. Bad conditioning, and very sensitive to X.

Solution : regularize !

$$\min_w f(w) = \frac{1}{n}\sum_{i=1}^{n}(\langle x_i, w\rangle - y_i)^2 + \frac{\lambda}{2}\|w\|^2$$

Q: Is the problem convex, smooth? Compute the associated constants

Logistic regression

Data: $x_1, \dots x_n \in \mathbb{R}^p$, and $y_1, \dots, y_n \in \{-1, +1\}$

Assumption: There exists w^* such that

$$y_i \simeq \mathrm{sign}(\langle x_i, w^* \rangle)$$

Optimization problem:

$$\min_w f(w) = \frac{1}{n} \sum_{i=1}^{n} \log(1 + \exp(-y_i \langle x_i, w \rangle))$$

Q: Is the problem convex, smooth? Compute the associated constants

Regularized logistic regression

Data: $x_1, \dots x_n \in \mathbb{R}^p$, and $y_1, \dots, y_n \in \{-1, +1\}$

Assumption: There exists w^* such that

$$y_i \simeq \text{sign}(\langle x_i, w^* \rangle)$$

Optimization problem:

$$\min_w f(w) = \frac{1}{n} \sum_{i=1}^{n} \log(1 + \exp(-y_i \langle x_i, w \rangle)) + \frac{\lambda}{2} \|w\|^2$$

Q: Is the problem convex, smooth? Compute the associated constants