COS 324: Lecture 7

Introduction to convex optimization

Elad Hazan Yoram Singer

PRINCETON UNIVERSITY

Admin

- Survey:
  - Mic
  - Examples
  - Theory vs. implementation standoff
  - Too slow/fast/just-right standoff (and also easy/hard)
  - Ex1 Q2 typo
- HW3

Recap

- Online learning, RWM
- Definition + fundamental theorem of statistical learning, motivated efficient algorithms/optimization
- Perceptron

Agenda

- convex relaxations
- convex optimization
- (perhaps) Gradient descent

    - But first – some examples of learning problems that fit our model!

Definition: learning from examples w.r.t. hypothesis class

A learning problem: $L = (X, Y, H)$

- X = Domain of examples (emails, pictures, documents, …)
- Y = label space (usually, binary Y={-1,1} or {0,1})
- D = distribution over (X,Y) (the world)
- Data access model: learner can obtain i.i.d samples from D
- H = class of hypothesis: $H \subseteq \{X \mapsto Y\}$
- Goal: produce hypothesis h$\in H$ with low *generalization error*

$$err(h) = E_{(x,y)\sim D}\,[h(x) \neq y]$$

(agnostic) PAC learnability

Learning problem $L = (X, Y, H)$ is (agnostically) PAC-learnable if there exists a learning algorithm (i.e.ERM) s.t. for every $\delta, \epsilon > 0$, there exists $\mathrm{m} = f(\epsilon, \delta, H) < \infty$, s.t. after observing S examples, for $|S| = m$, returns a hypothesis $h \in H$, such that with probability at least

$$1 - \delta$$

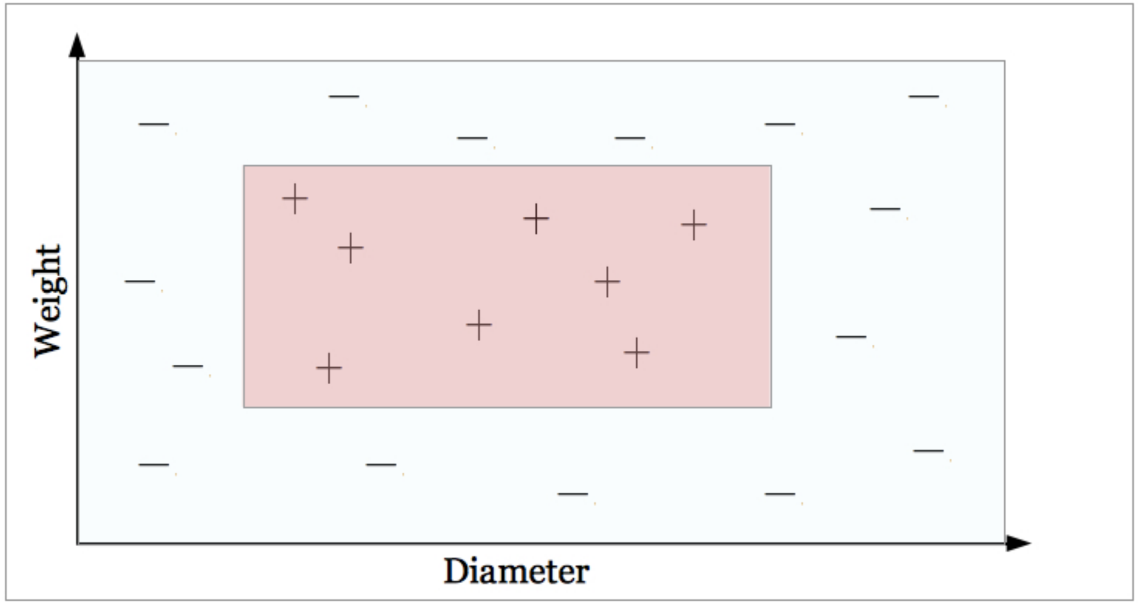
it holds that

$$err(h) \leq \min_{h^* \in H} err(h^*) + \epsilon$$

Equals zero for realizable concepts

Examples

- Apple factory:
  - Apples are sweet (box) or sour (for export)
  - Features of apples: weight and diameter
  - Weight, diameter are distributed uniformly at random in a certain range



  - X,Y = ?
  - Reasonable hypothesis class?
  - Realizable?

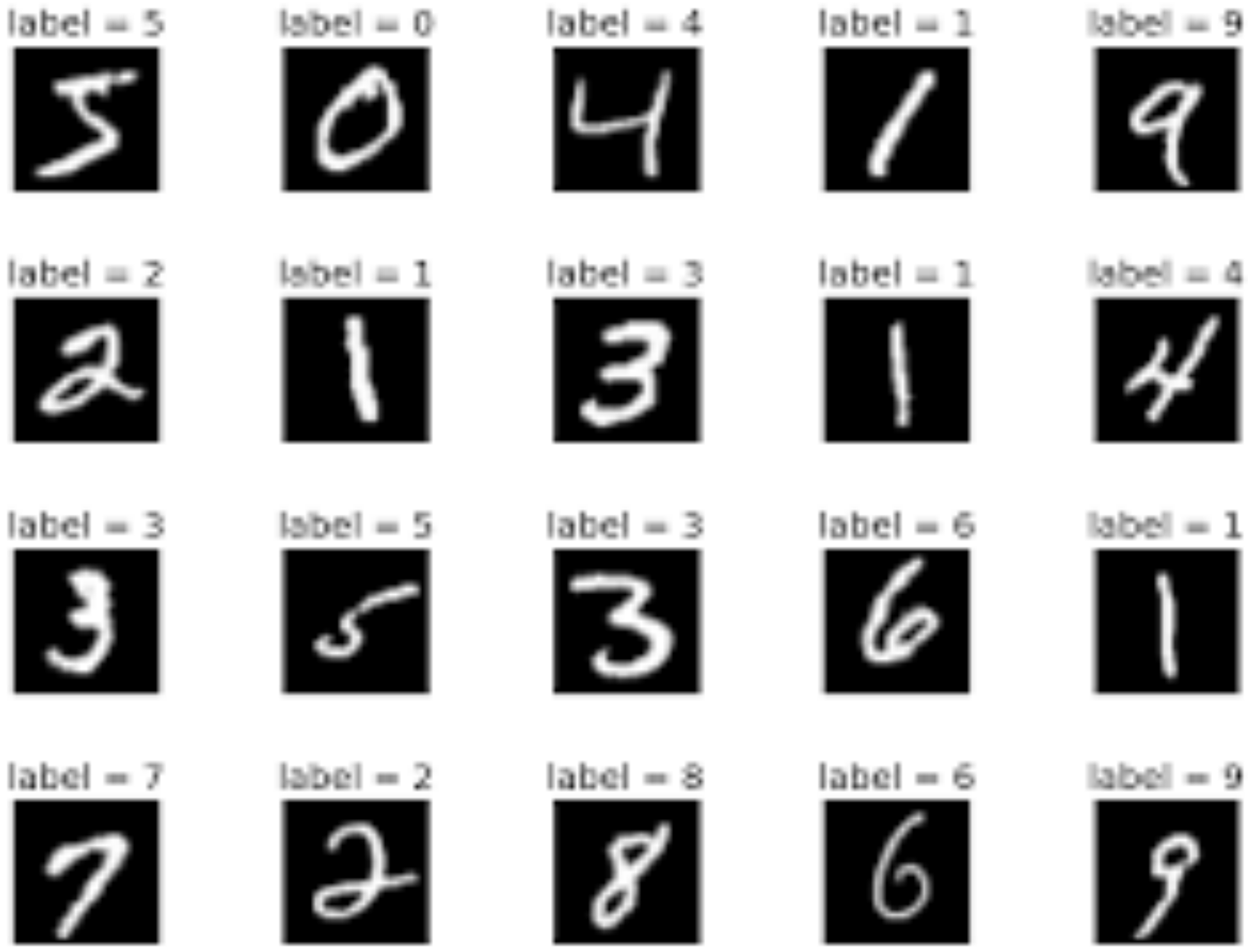
Examples: MPG prediction

- X,Y,H = ?

mpg	cylinders	displacement	horsepower	weight	acceleration	modelyear	maker
good	4	low	low	low	high	75to78	asia
bad	6	medium	medium	medium	medium	70to74	america
bad	4	medium	medium	medium	low	75to78	europe
bad	8	high	high	high	low	70to74	america
bad	6	medium	medium	medium	medium	70to74	america
bad	4	low	medium	low	medium	70to74	asia
bad	4	low	medium	low	low	70to74	asia
bad	8	high	high	high	low	75to78	america
:	:	:	:	:	:	:	:
:	:	:	:	:	:	:	:
:	:	:	:	:	:	:	:
bad	8	high	high	high	low	70to74	america
good	8	high	medium	high	high	79to83	america
bad	8	high	high	high	low	75to78	america
good	4	low	low	low	low	79to83	america
bad	6	medium	medium	medium	high	75to78	america
good	4	medium	low	low	low	79to83	america
good	4	low	low	medium	high	79to83	america
bad	8	high	high	high	low	70to74	america
good	4	low	medium	low	medium	75to78	europe
bad	5	medium	medium	medium	medium	75to78	europe

Examples

- Character recognition
  - X,Y = ?
  - Reasonable hypothesis class?
  - Realizable?

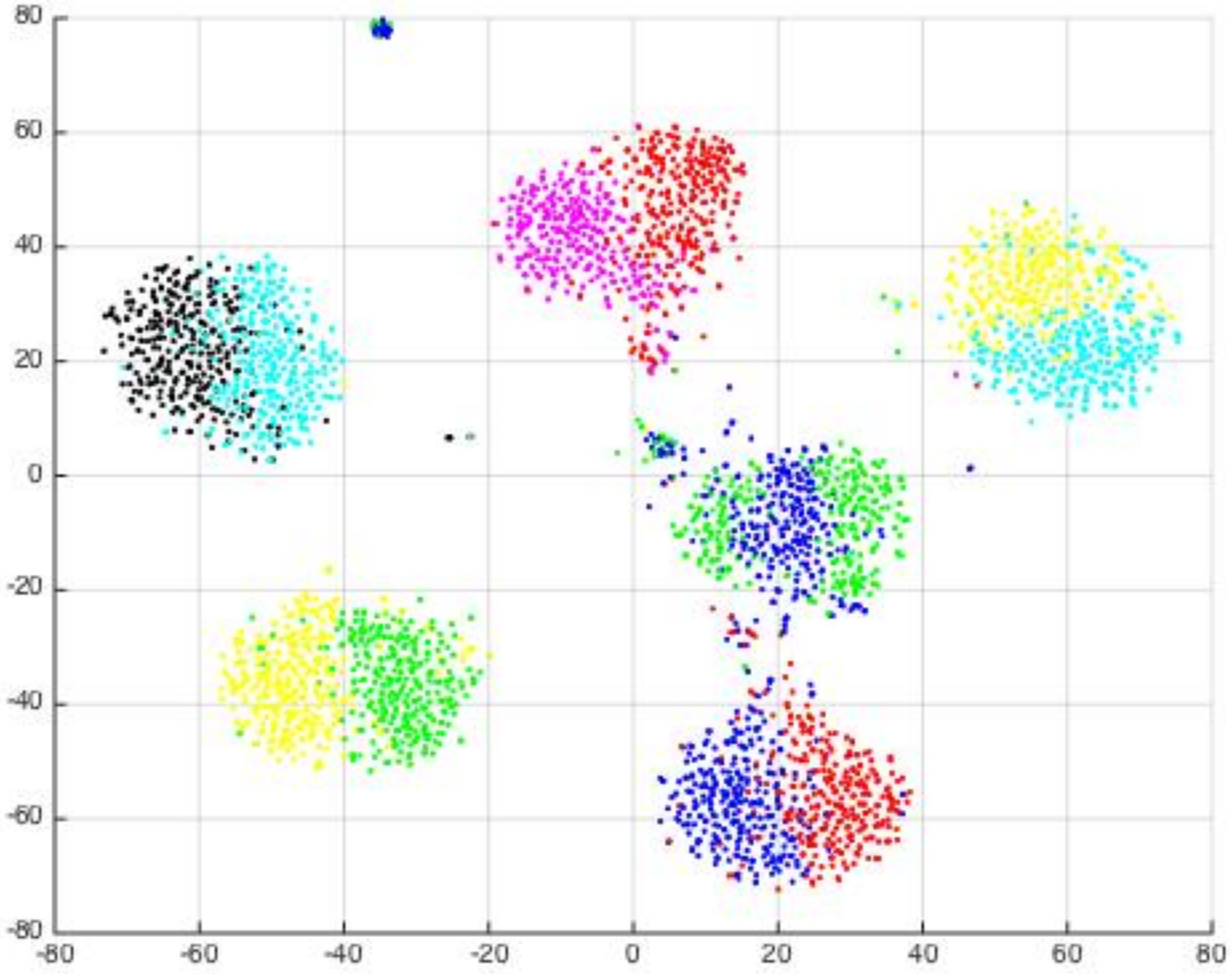


Examples

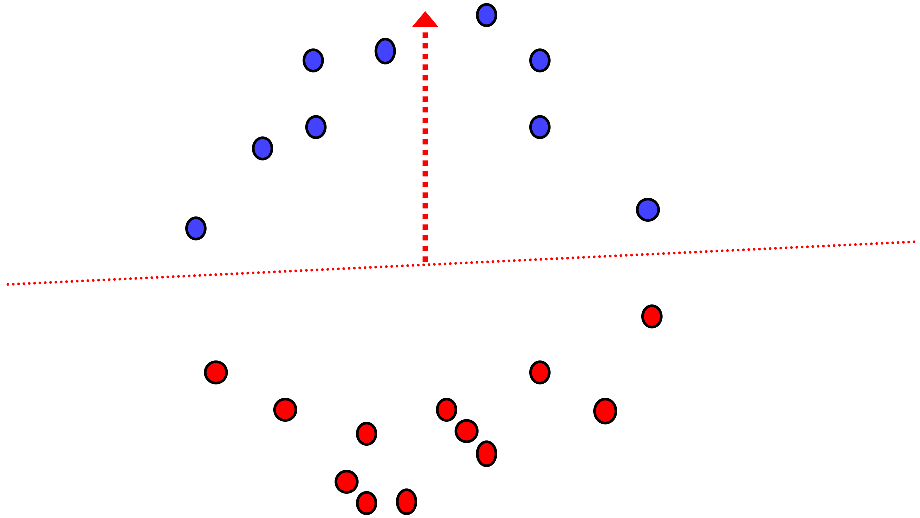
- Spam detection:
  - X,Y = ?
  - Reasonable hypothesis class?
  - Realizable?
- Chair classification
- Gene association w. diseases
- …



Empirically: the world is many times linearly-separable



Linear classifiers



Domain = vectors over Euclidean space R^{d}

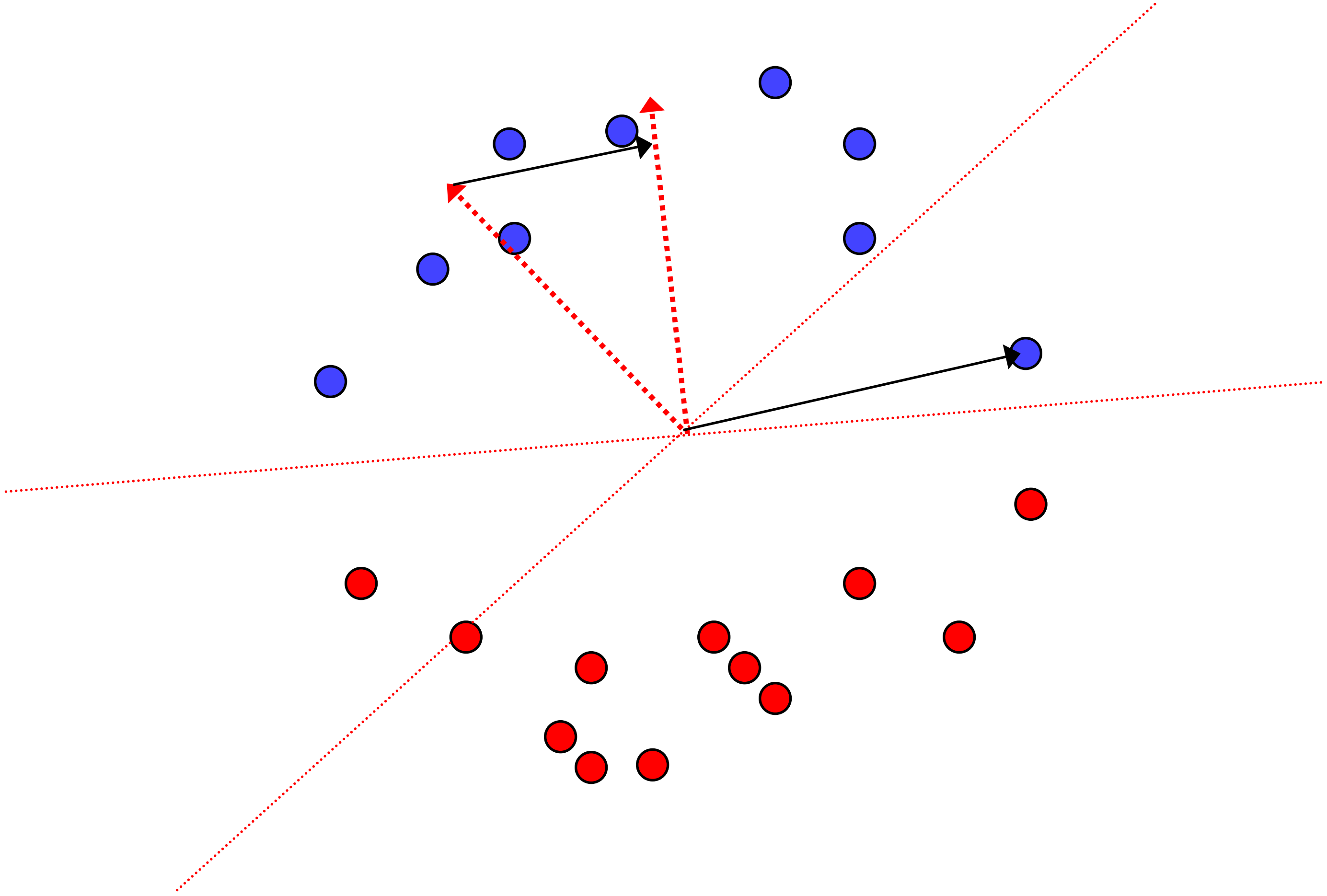
Hypothesis class: all hyperplanes that classify according to:

$$h(x) = sign(w^{\top} x - b)$$

(we usually ignore b – the bias, it is 0 almost w.l.o.g.)

The Perceptron Algorithm

[Rosenblatt 1957, Novikoff 1962, Minsky&Papert 1969]



The Perceptron Algorithm

Iterately:

1. Find vector x_i for which $\text{sign}\ (w^T x_i) \neq y_i$
2. Add x_i to w:

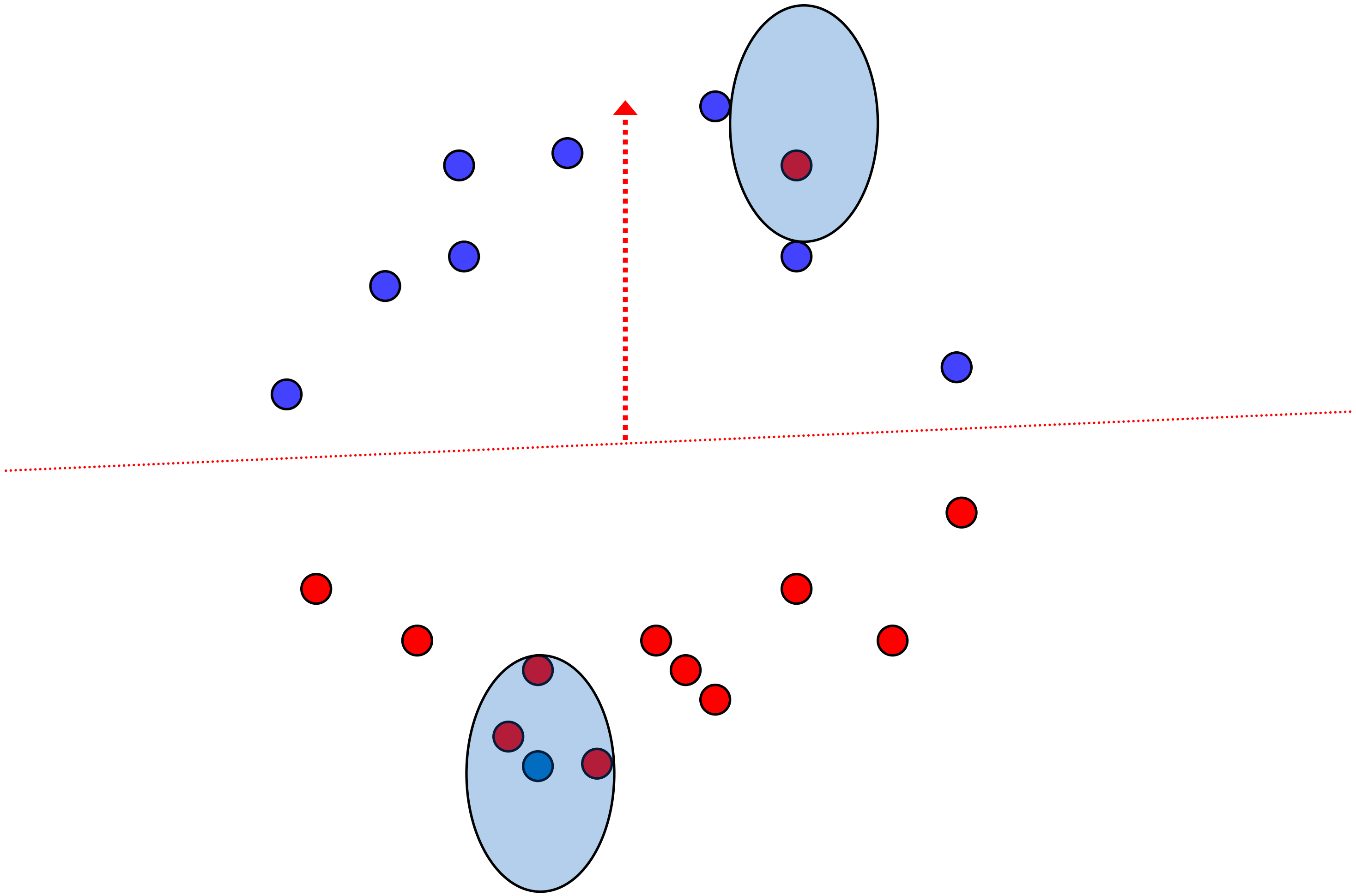
$$w_{t+1} \leftarrow w_t + \ y_i x_i$$

The Perceptron Algorithm

Reminder: Thm [Novikoff 1962]: for data with margin ϵ, perceptron returns

separating hyperplane in $\frac{1}{\epsilon^2}$ iterations

Noise?



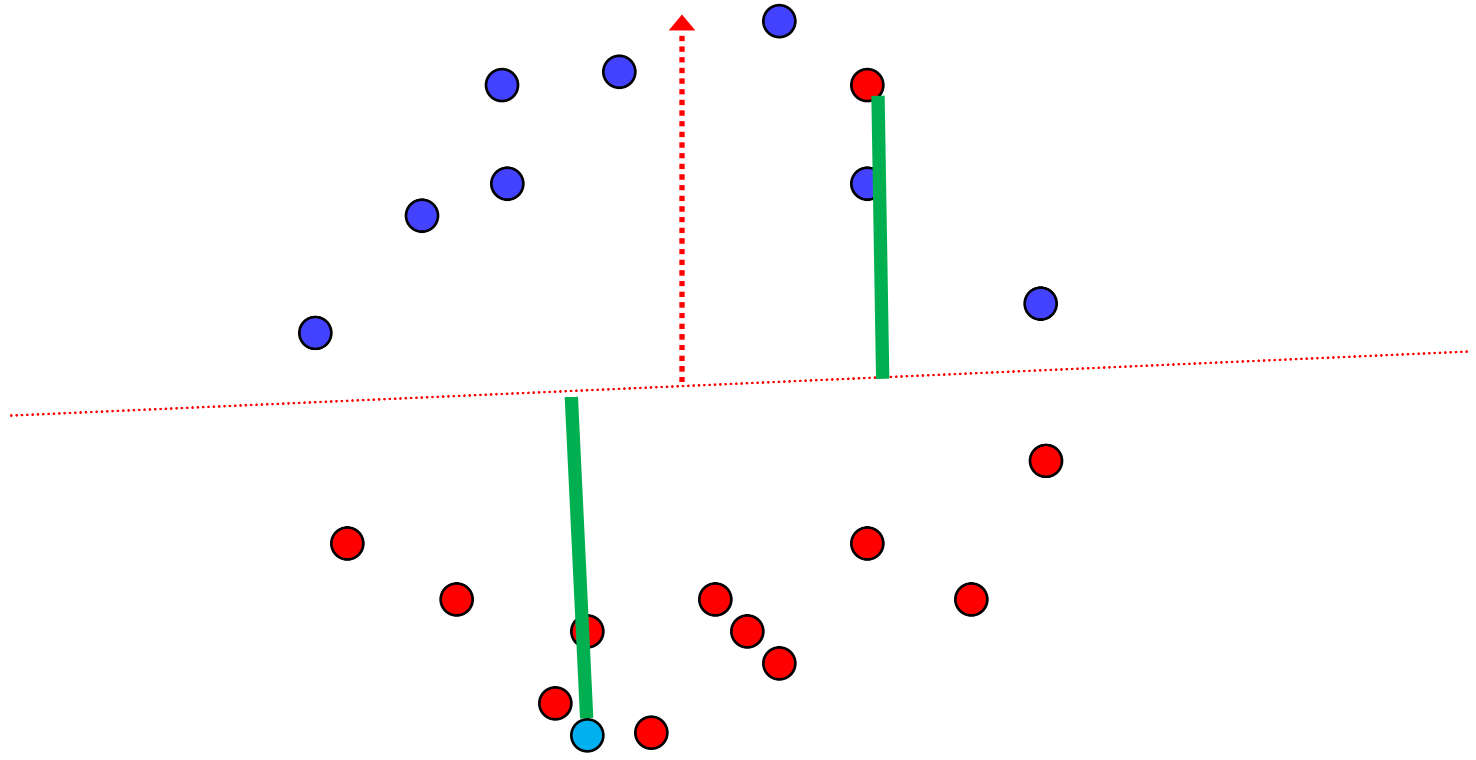
ERM for noisy linear separators?

Given a sample $S = \{(x_1, y_1), \ldots, (x_m, y_m)\}$, find hyperplane (through the origin w.l.o.g) such that:

$$w = \arg\min_{|w| \leq 1} |\{i \quad \text{s.}t.\ sign\ (w^T x_i) \neq y_i\}|$$

- NP-hard!
- → convex relaxation + optimization!

Noise – minimize sum of weighted violations



Soft-margin SVM (support vector machines)

Given a sample $S = \{(x_1, y_1), \dots, (x_m, y_m)\}$, find hyperplane (through the origin w.l.o.g) such that:

$$w = \underset{|\mathrm{w}| \leq 1}{\operatorname{argmin}} \left\{ \frac{1}{m} \sum_i \max\{0, 1 - y_i w^\top x_i\} \right\}$$

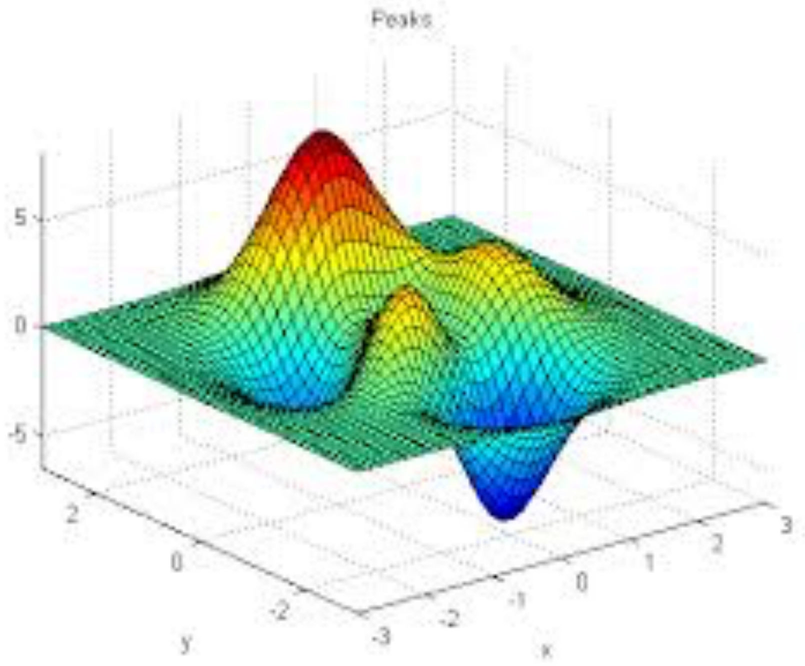
- Efficiently solvable by greedy algorithm – gradient descent
- More general methodology: convex optimization

- Next few lectures: optimization theory & algorithms!

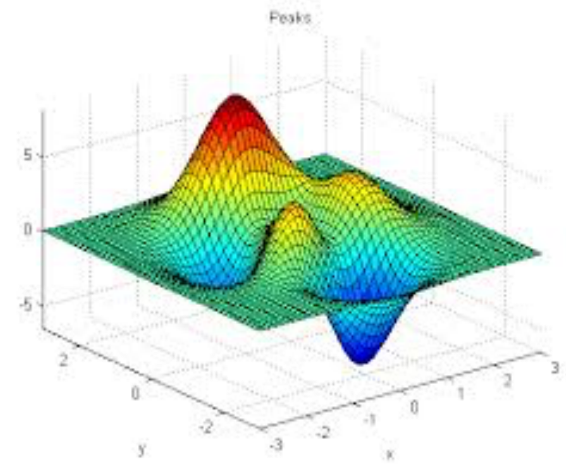
Mathematical optimization

Input: function $f: K \mapsto R$, for $K \subseteq R^d$

Output: point $x \in K$, such that $f(x) \leq f(y) \; \forall \, y \in K$



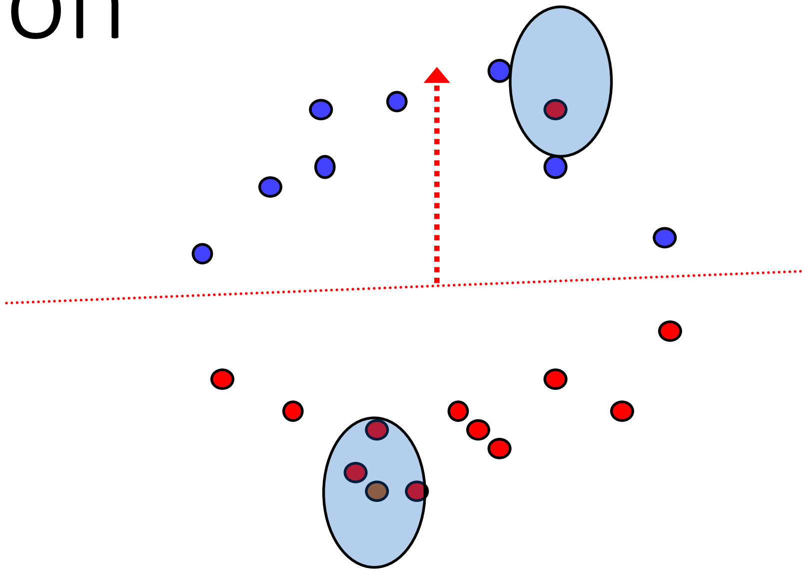
Mathematical optimization



- Continuous functions (back to calculus, derivatives, differentiability, …)
- Vs. combinatorial optimization as in graph algorithms (strong connection)
- Studied since early 1900’s , lots of work in soviet union (central optimization, resource allocation, military applications, etc.)
- Special cases: linear programming, convex optimization, max flow in graphs

Efficient (poly-time) algorithms

Optimization for linear classification

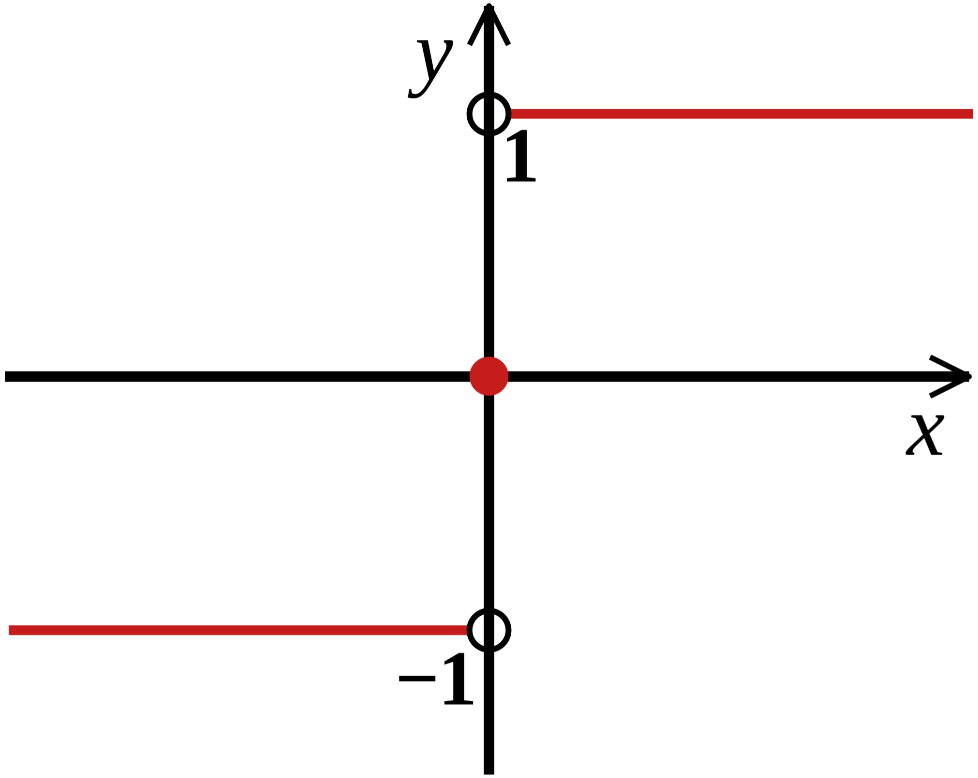


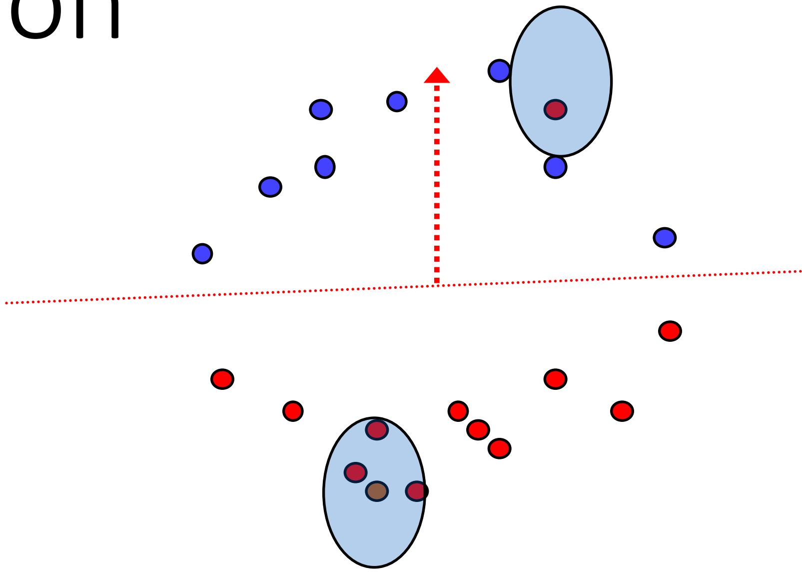
Given a sample $S = \{(x_1, y_1), \dots, (x_m, y_m)\}$, find hyperplane (through the origin w.l.o.g) such that:

$$w = \arg\min_{|w| \leq 1} \# \text{ of mistakes}$$

Optimization for linear classification

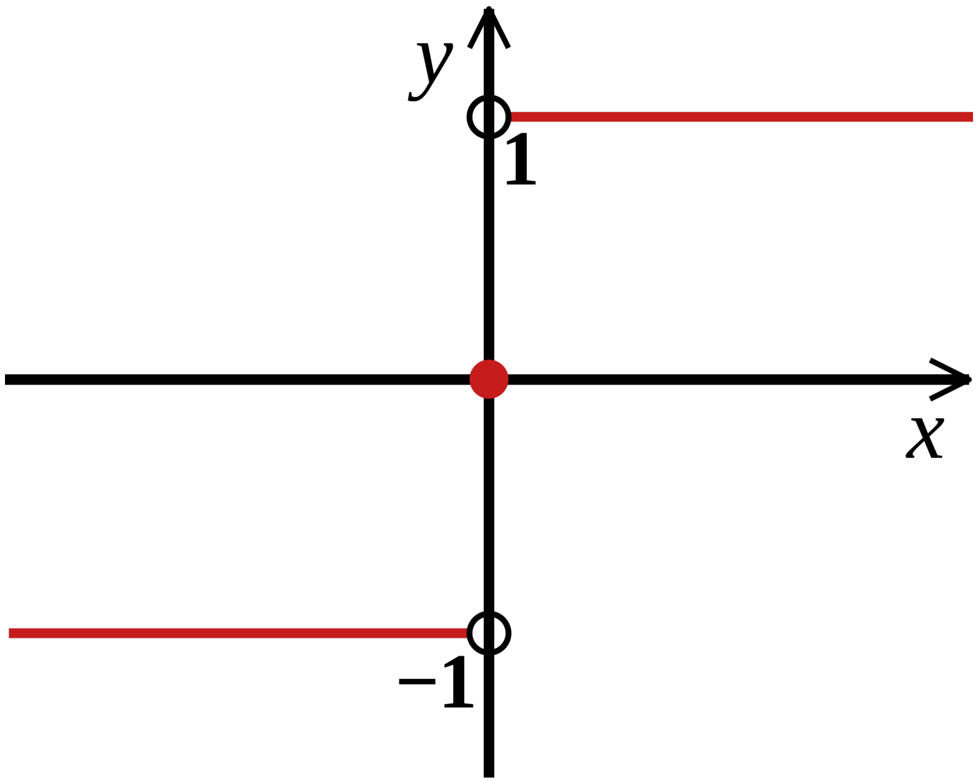
$$w = \arg\min_{|w|\leq 1} |\{i \;\; \mathrm{s.t.}\; sign\,(w^T x_i) \neq y_i\}|$$

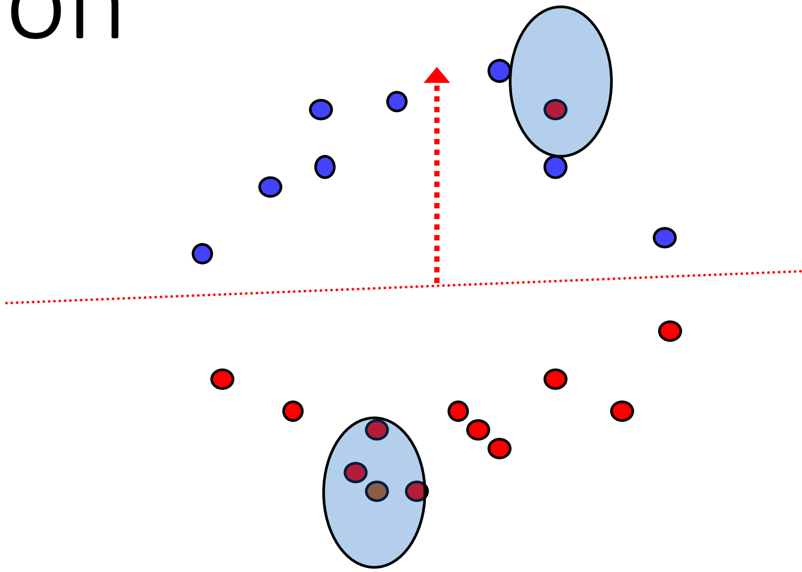




Optimization for linear classification

$$w = \arg\min_{|w|\leq 1} E_i[\ell(w,(x_i,y_i))]$$

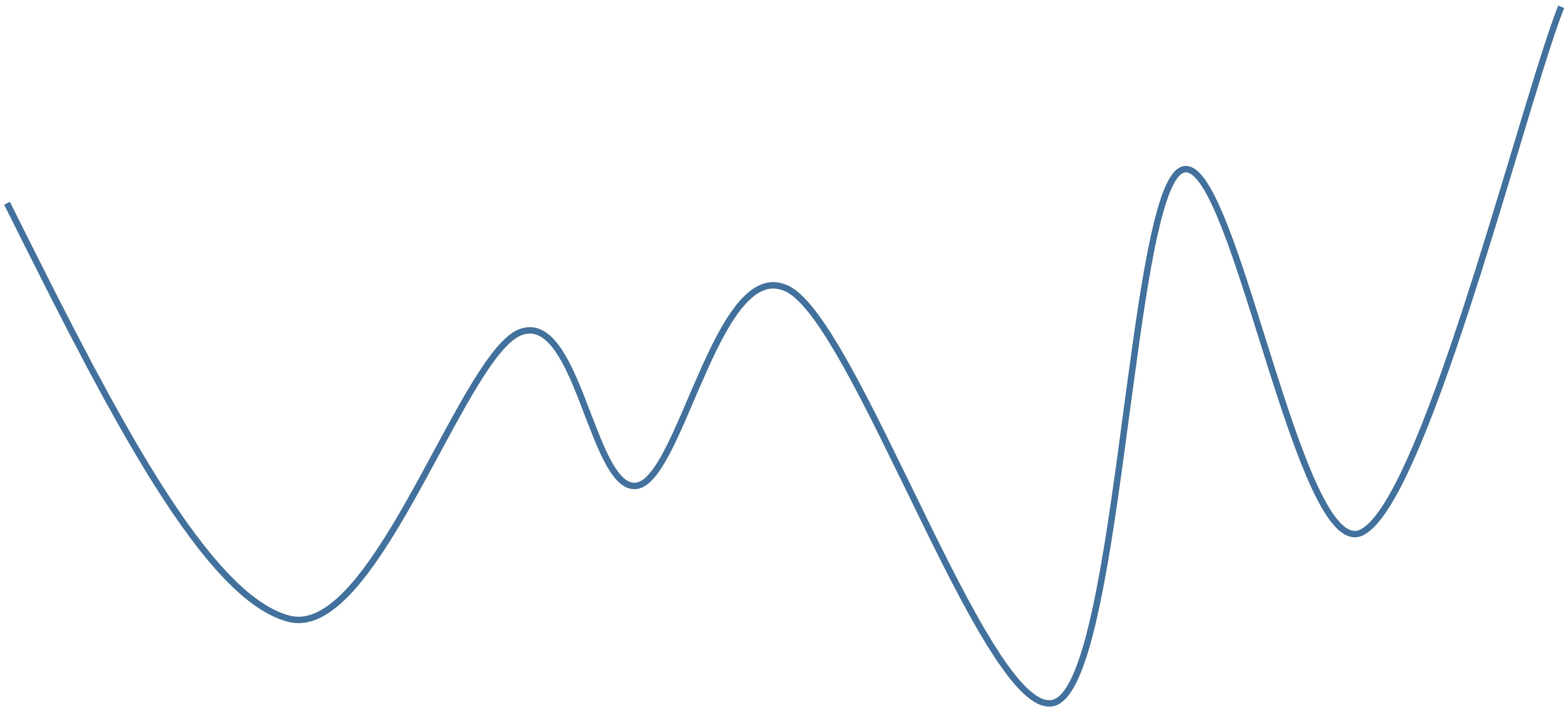




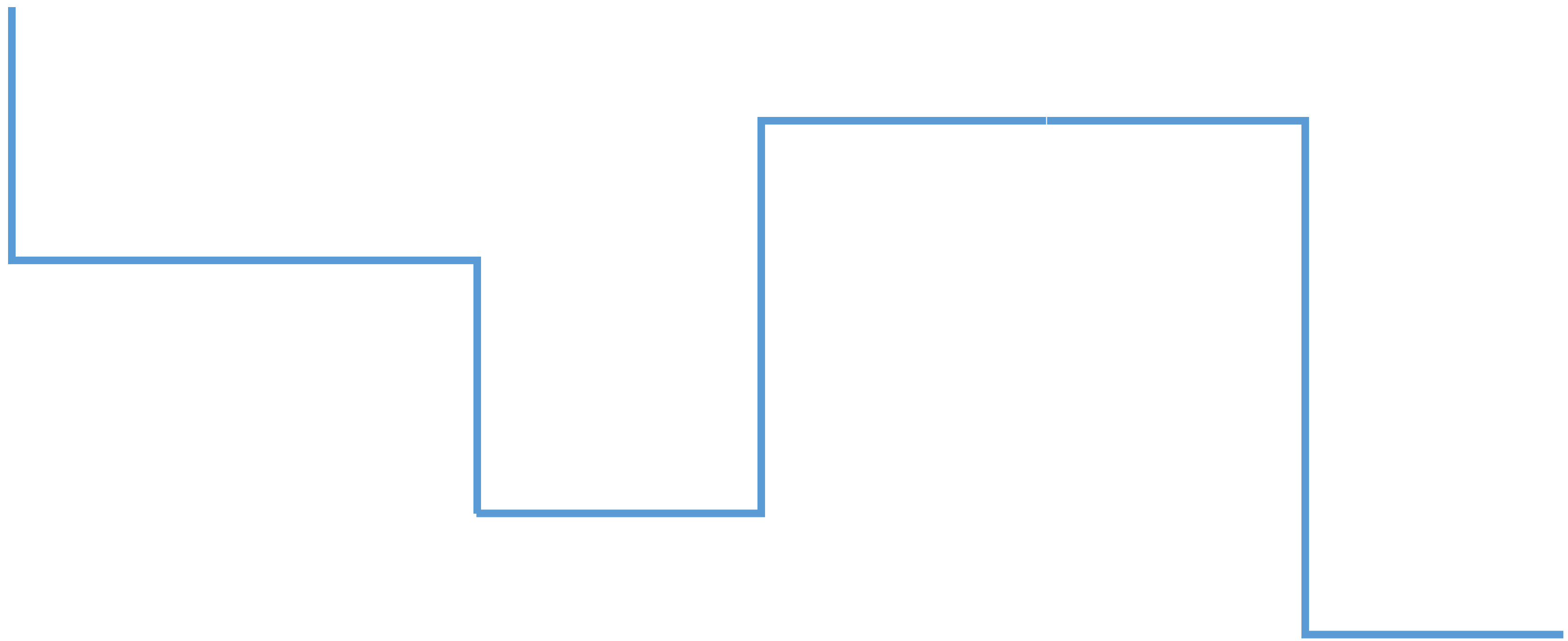
Loss function:
$\ell\big(w,(x_i,y_i)\big) = sign(y_i w^\top x_i)$

Mathematical optimization:
$$\min f(w)$$
For f(w) $= E\ [\ell(w,(x_i,y_i))]$

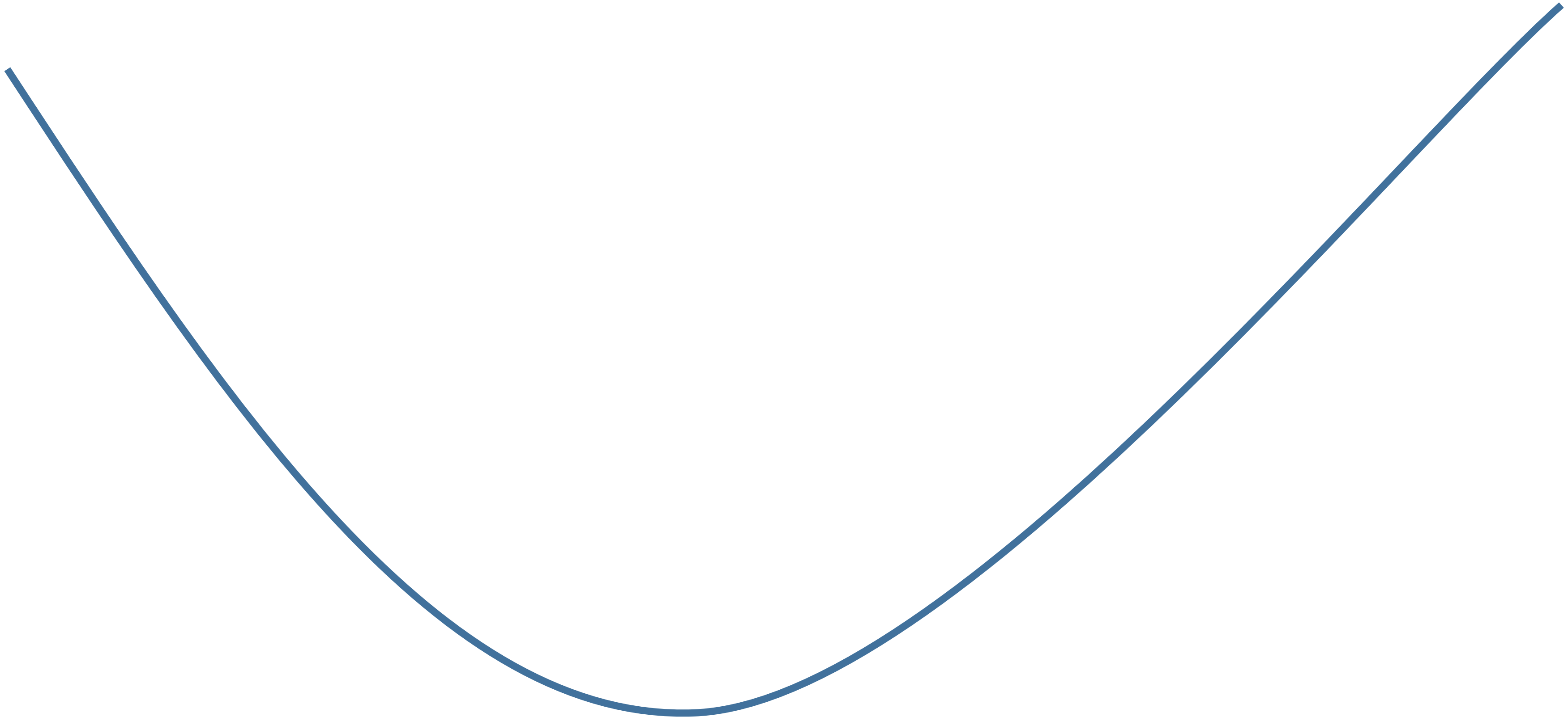
Minimization can be hard



Sum of signs → hard

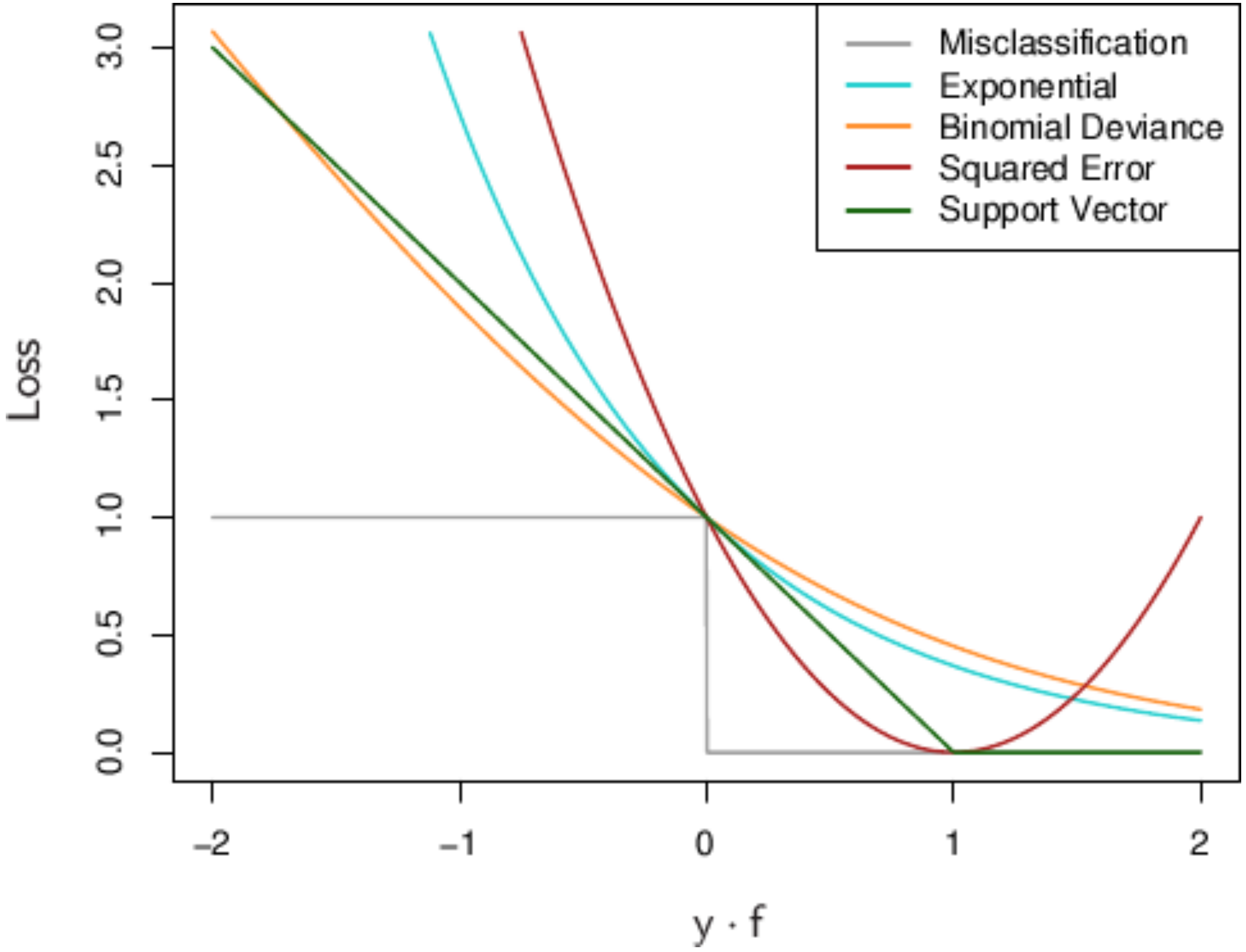


Convex functions: local → global

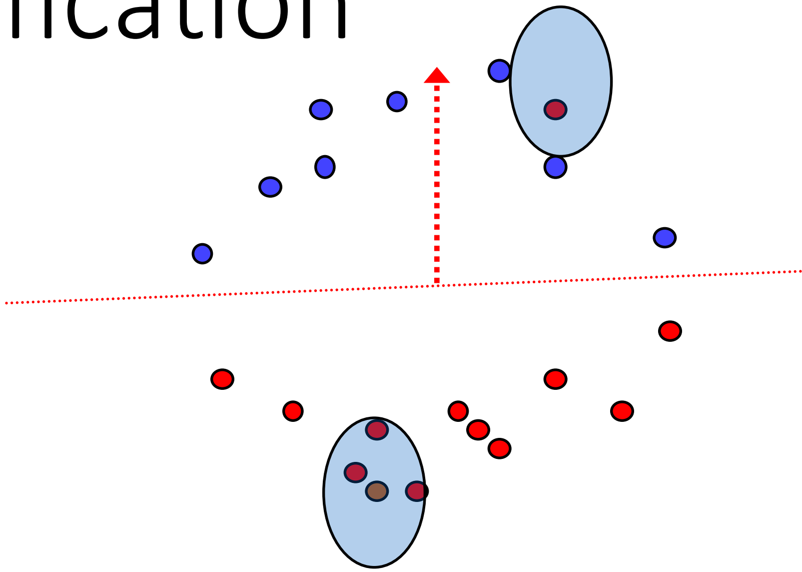


Sum of convex functions → also convex

Convex relaxation for 0-1 loss



Convex relaxation for linear classification



$$w = \arg\min_{|w|\le 1} |\{i \ \ s.t.\ sign\,(w^T x_i) \neq y_i\}|$$

$$w = \arg\min_{|w|\le 1} \ell(w^\top x_i, y_i) \text{ such as:}$$

1. Ridge / linear regression $\ell(w^\top x_i, y_i) = (w^\top x_i - y_i)^2$
2. SVM $\ell(w^\top x_i, y_i) = \max\{0, 1 - y_i\ w^\top x_i\}$
3. Logistic regression $\ell(w^\top x_i, y_i) = \log(1 + e^{y_i w^\top x_i})$

Small recap

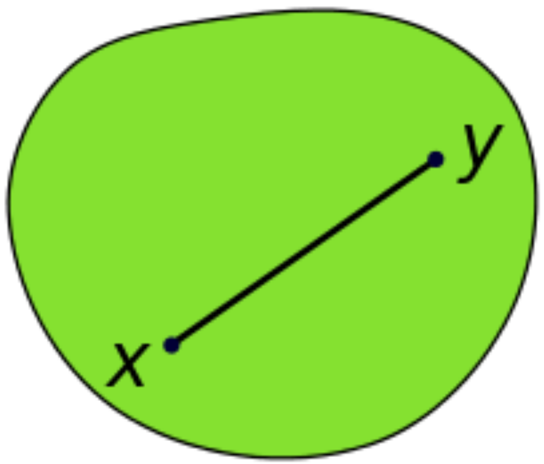
- Finding linear classifiers: formulated as mathematical optimization
- Convexity: property that allows local greedy algorithms
- Formulate convex relaxations to linear classification

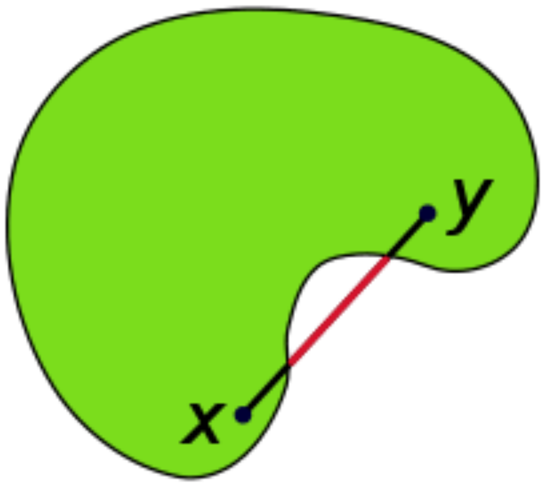
Next:

- Convex analysis
- Algorithms for convex optimization

Convexity

A set $f: K \subseteq R^d$ is convex if and only if for every $\mathrm{x}, \mathrm{y} \in K$, the segment $[\mathrm{x}, \mathrm{y}] \in K$ is also in K. That is, for every $\alpha \in [0,1]$, the **convex combination** $\alpha\, x + (1 - \alpha)\, y$ is in K.



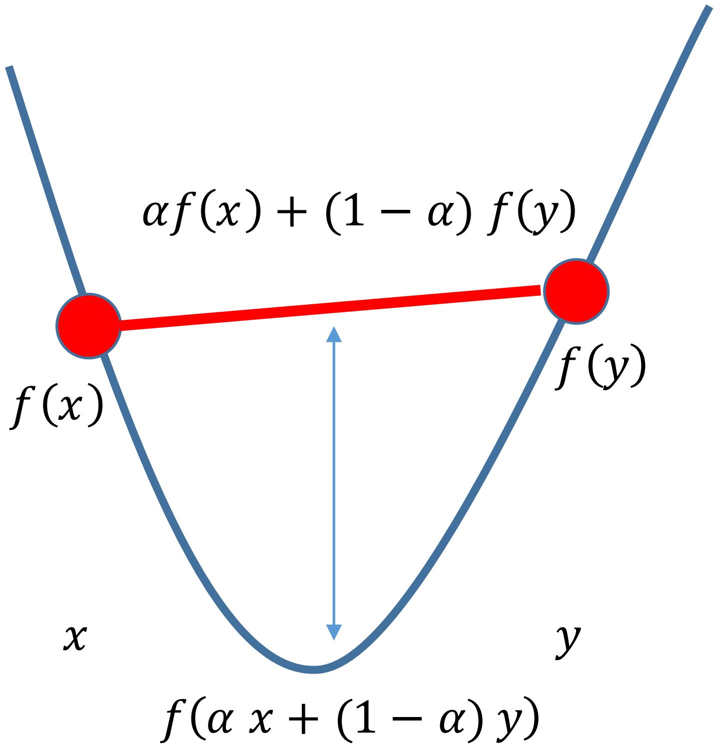


Convexity

A function $f: R^d \mapsto R$ is convex if and only if for every $\alpha \in [0,1]$:

$$f(\alpha\, x + (1-\alpha)\, y) \leq \alpha f(x) + (1-\alpha)\, f(y)$$

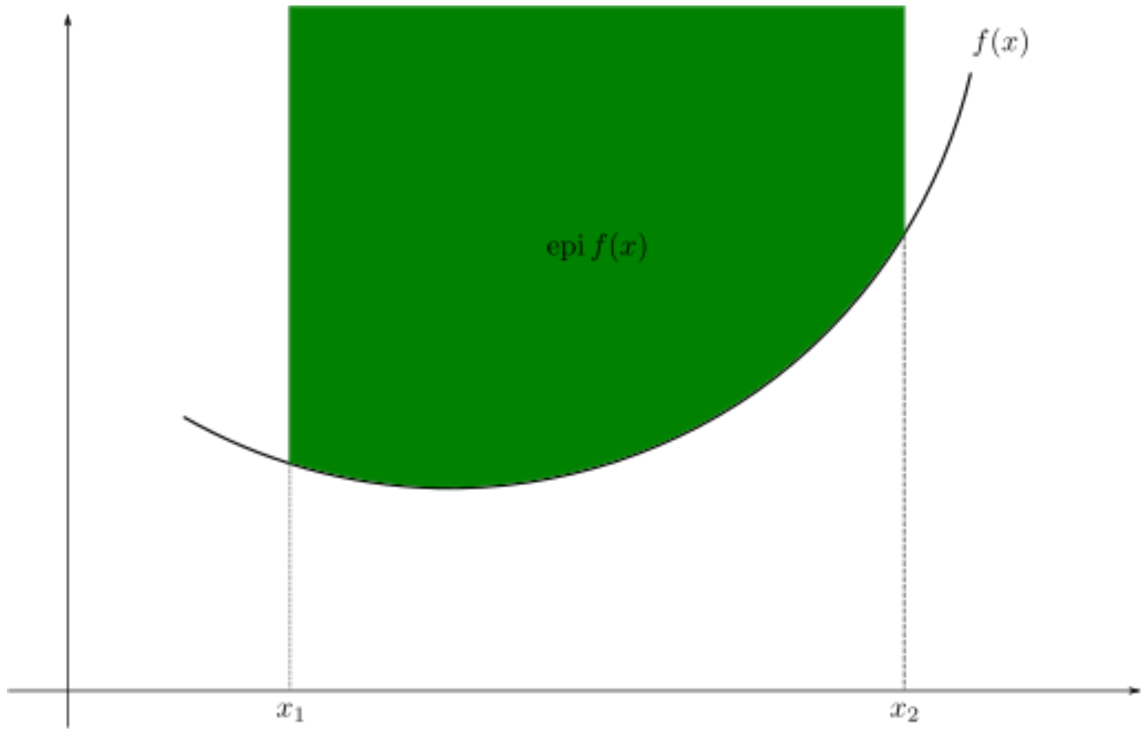
- Informally: smiley ☺



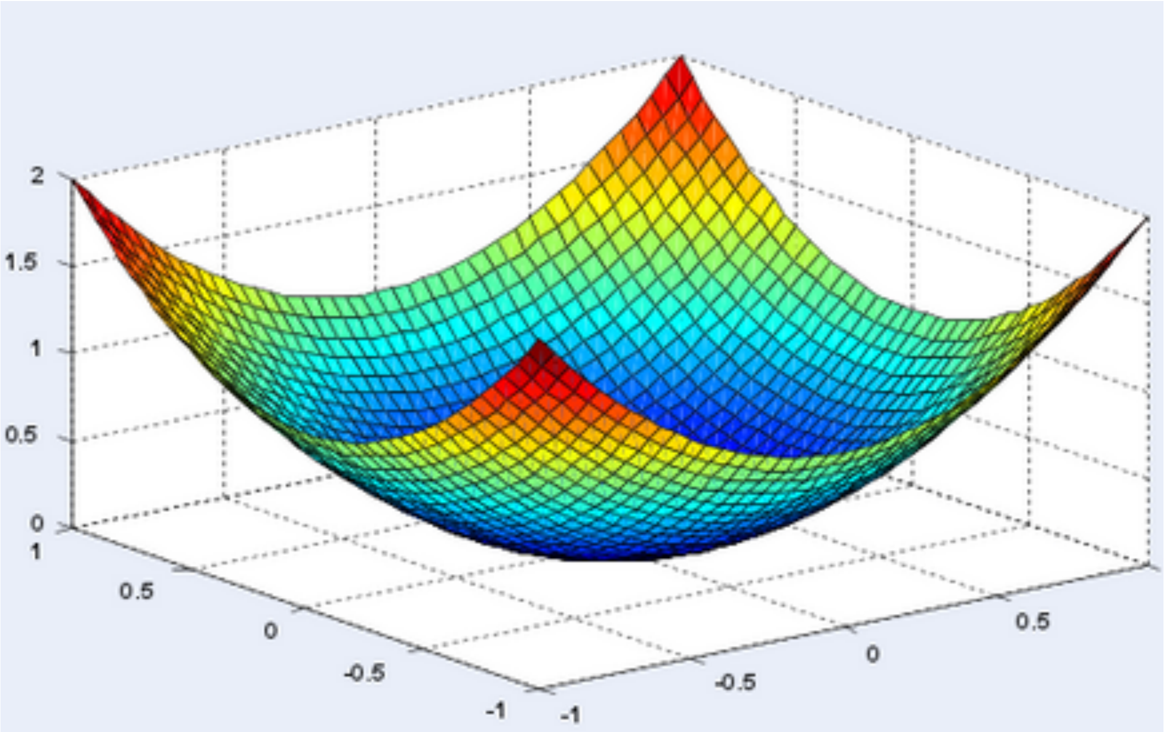
Epigraph

A function $f: R^d \mapsto R$ is convex if and only if its epigraph is a convex set:

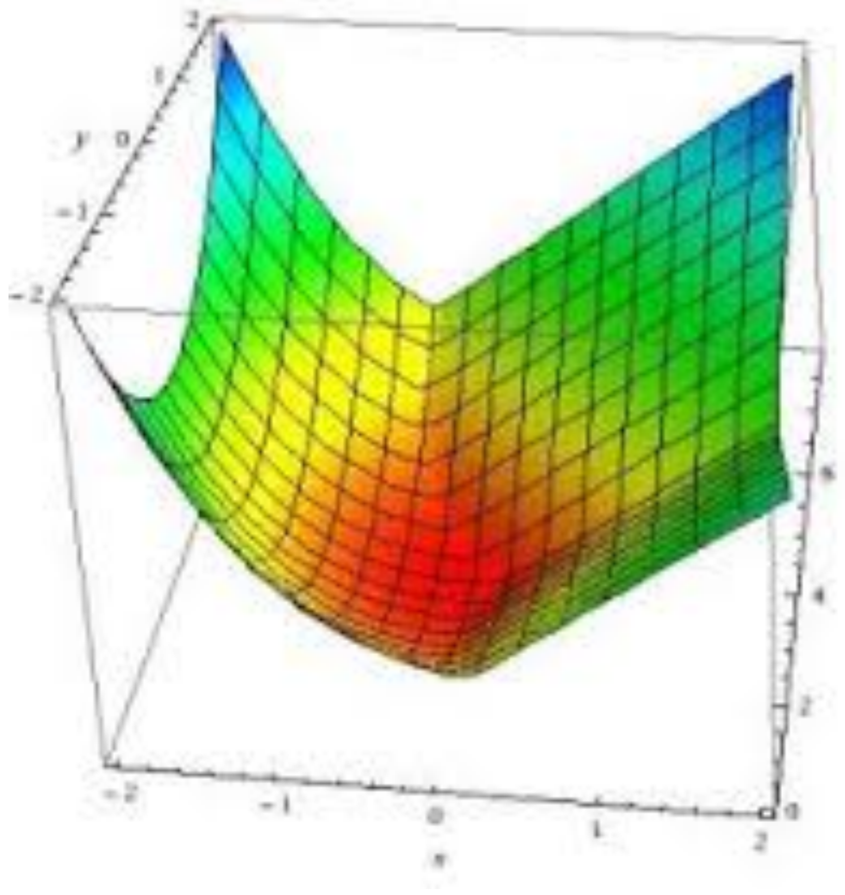
Epigraph(f) $= \{(x, y) | f(x) \leq y\}$



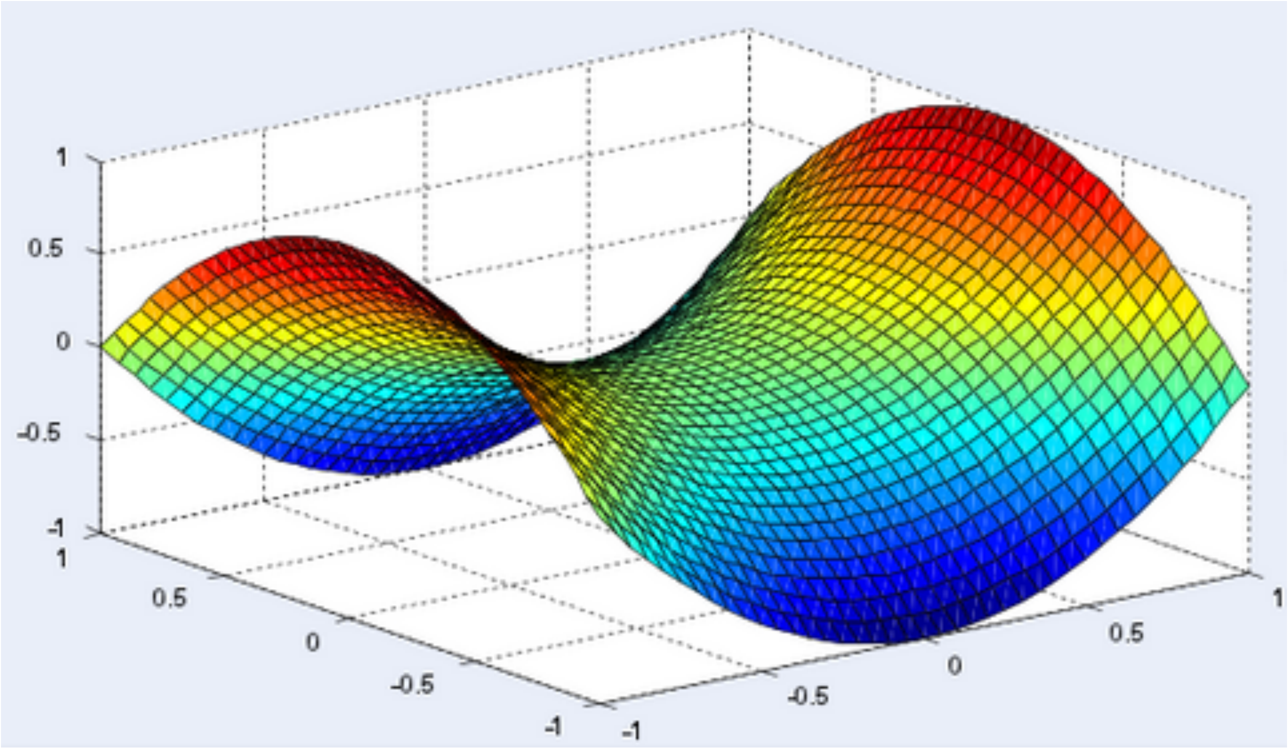
Convex and general functions



Convex and general functions



Convex and general functions



Convexity: local → global

- Theorem: for f convex, every local minimum is a global minimum

- Global minimum = smallest point according to f
- Local minimum: everyone around the point is larger.
- Formally:

$$\mathrm{B_r(x)} = \{y\colon\ |x - y| \leq r\ \}$$

- X is local min if exists r>0 such that

$$\forall y \in \mathrm{B_r(x)}.\, f(y) \geq f(x)$$

Theorem: f convex, every local minimum is a global minimum

- local min: x, exists r>0 such that
$$\forall y \in \mathrm{B_r(x)}.\, f(y) \geq f(x)$$
- Thus for every v, there exists some very very small $\alpha > 0$, such that $\mathrm{x} + \alpha(v - x) \in B_r(x)$, and thus
$$f(x) \leq f\big(x + \alpha(v - x)\big)$$
$$= f\big((1-\alpha)x + \alpha v\big)$$
$$\leq (1-\alpha)f(x) + \alpha f(v)$$
- Thus,
$$\alpha f(x) \leq \alpha f(v)$$
- This holds for every v, and thus x is a global minimum.

Summary

- Motivation: linear classification with noise is NP-hard
- Thus we have convex relaxation (i.e. SVM), for which we have efficient algorithms
- Started the theory of mathematical & convex optimization