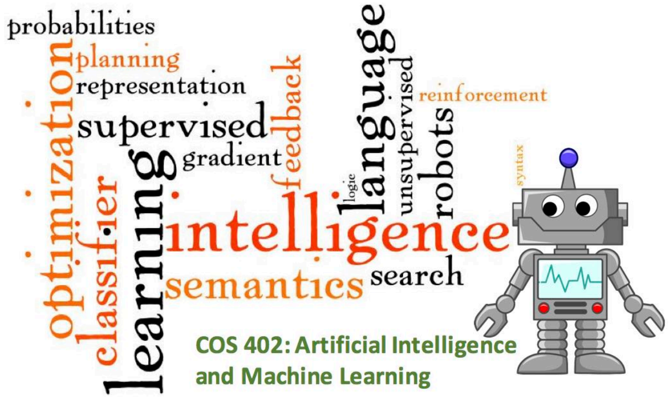


COS 402 – Machine Learning and Artificial Intelligence
Fall 2016

Lecture 6: stochastic gradient descent

Sanjeev Arora Elad Hazan

PRINCETON UNIVERSITY

Admin

- Exercise 2 (implementation) this Thu, in class
- Exercise 3 (written), this Thu, in class
- Movie – “Ex Machina” + discussion panel w. Prof. Hasson (PNI)
  Wed Oct. 5th 19:30
  tickets from Bella; room 204 COS

- Today: special guest - Dr. Yoram Singer @ Google

Recap

- Definition + fundamental theorem of statistical learning, motivated efficient algorithms/optimization
- Convexity and it’s computational importance
- Local greedy optimization – gradient descent

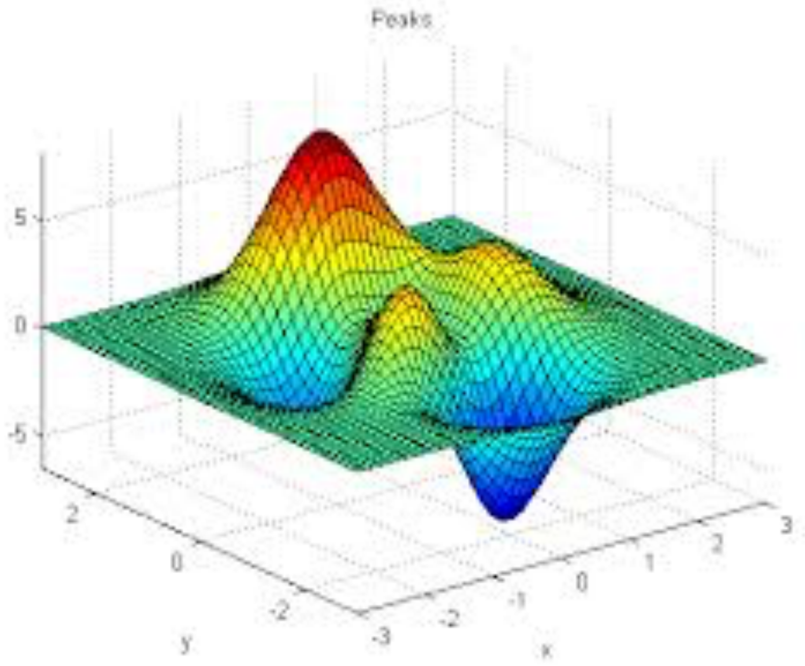
Agenda

- Stochastic gradient descent
- Dr. Singer on opt @ google & beyond

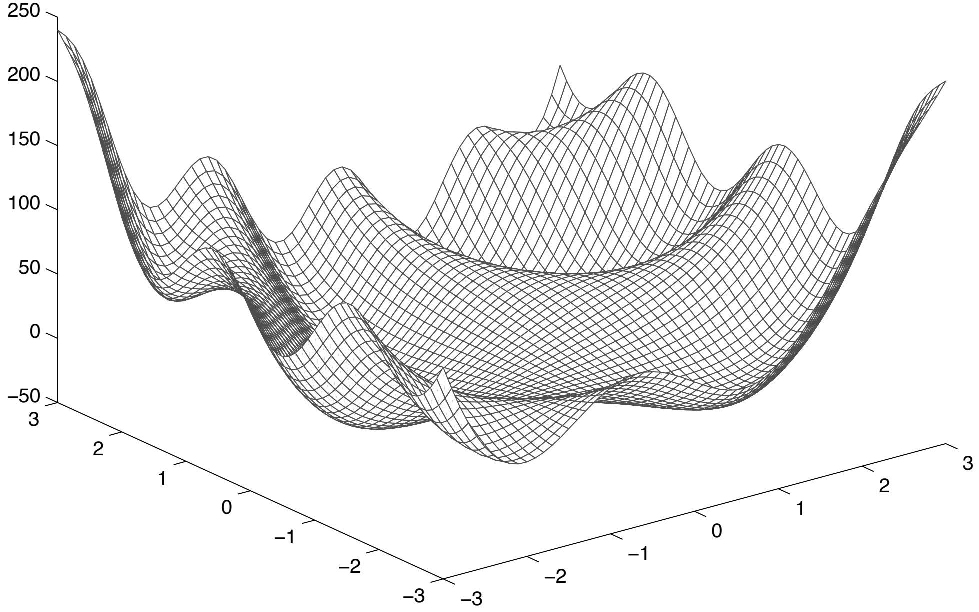
Mathematical optimization

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

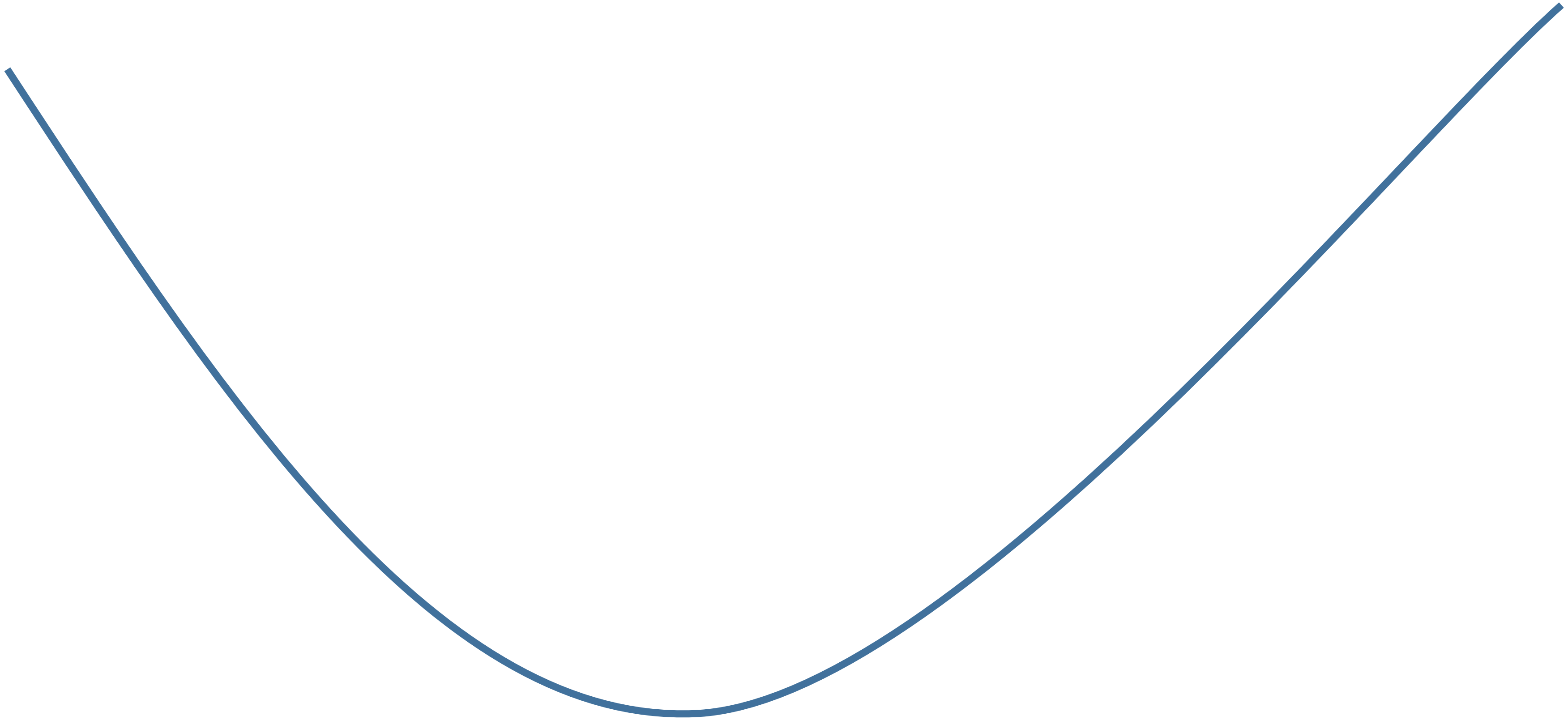
Output: point $x \in K$, such that $f(x) \leq f(y)\ \forall\ y \in K$



Prefer Convex Problems



Convex functions: local → global

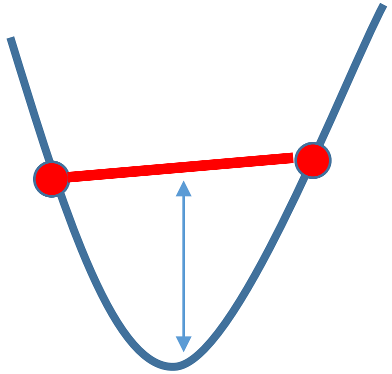


Sum of convex functions → also convex

Convex Functions and Sets

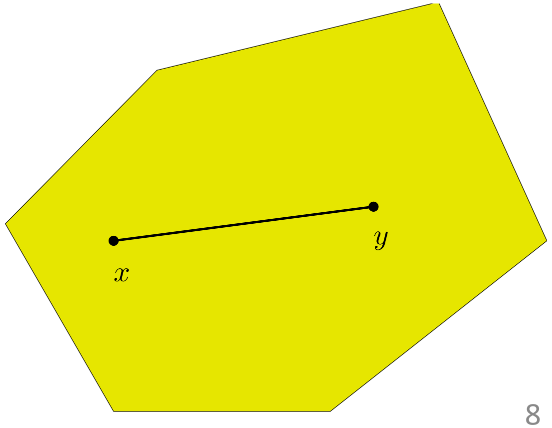
A function $f : \mathbb{R}^n \to \mathbb{R}$ is convex if for $x, y \in \mathrm{dom} f$ and any $a \in [0, 1]$,

$$f(ax + (1-a)y) \le af(x) + (1-a)f(y)$$



A set $C \subseteq \mathbb{R}^n$ is convex if for $x, y \in C$ and any $a \in [0, 1]$,

$$ax + (1-a)y \in C$$

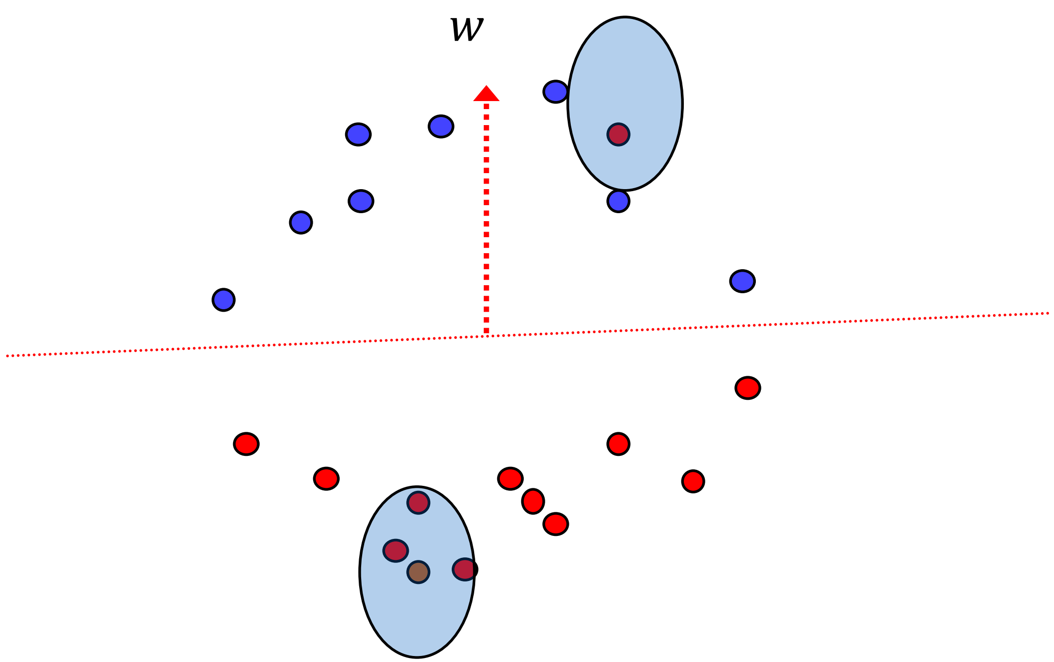


Special case: optimization for linear classification

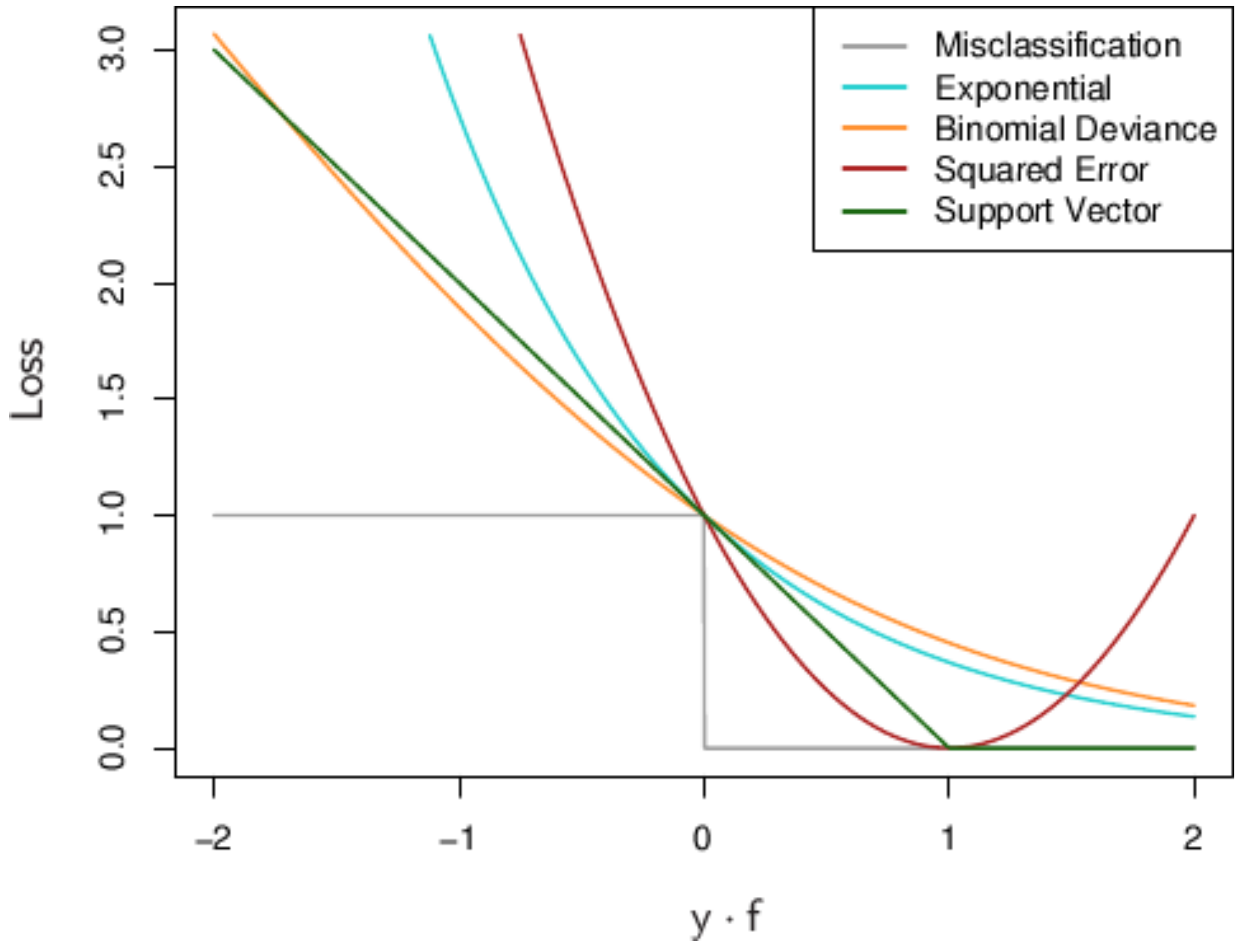
Given a sample $S = \{(x_1, y_1), \ldots, (x_m, y_m)\}$, find hyperplane (through the origin w.l.o.g) such that:

$\min_{w}$ # of mistakes

$\min_{|w| \leq 1} \ell(w^\top x_i, y_i)$ for a convex loss function



Convex relaxation for 0-1 loss



1. Ridge / linear regression

   $\ell(w, x_i, y_i) = (w^\top x_i - y_i)^2$

2. SVM

   $\ell(w, x_i, y_i) = \max\{0, 1 - y_i \; w^\top x_i\}$

i.e. for |w|=|x_i|=1,
We have:

$1 - y_i w^T x_i = \{0 \quad y_i = w^\top x_i$
$\{\leq 2 \quad y_i \neq w^T x_i$

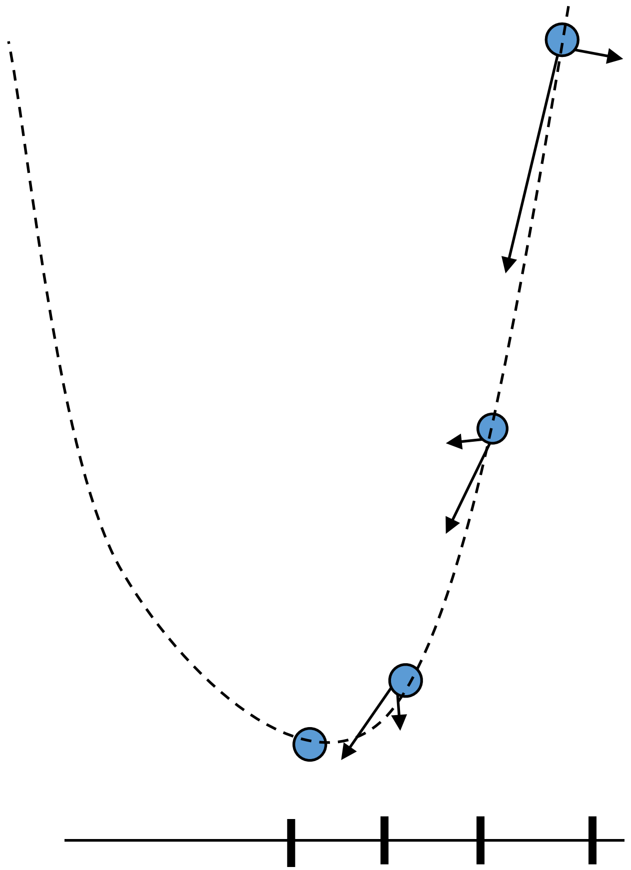
Greedy optimization: gradient descent

- Move in the direction of steepest descent:

$$x_{t+1} \leftarrow x_t - \eta \nabla f(x_t)$$

We saw: for certain step size choice,

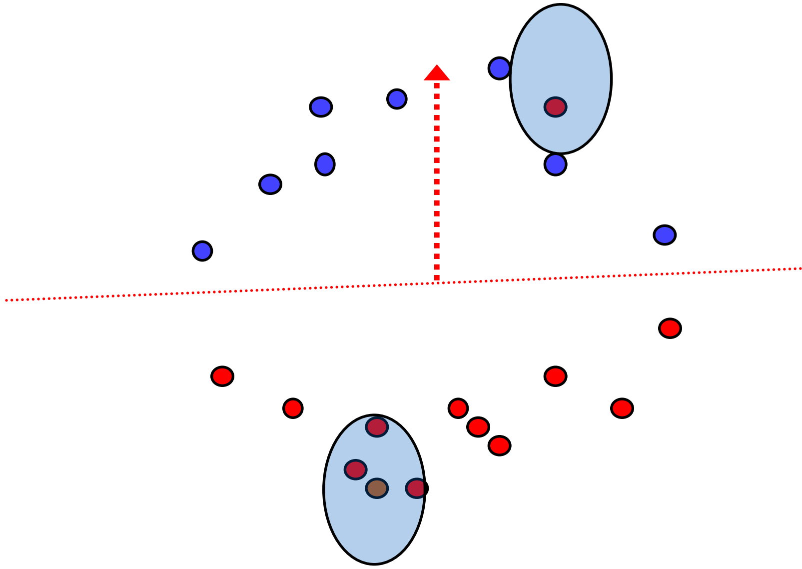
$$f\left(\frac{1}{T}\sum_t x_t\right) \leq \min_{x^* \in K} f(x^*) + \frac{1}{\sqrt{T}}$$



GD for linear classification

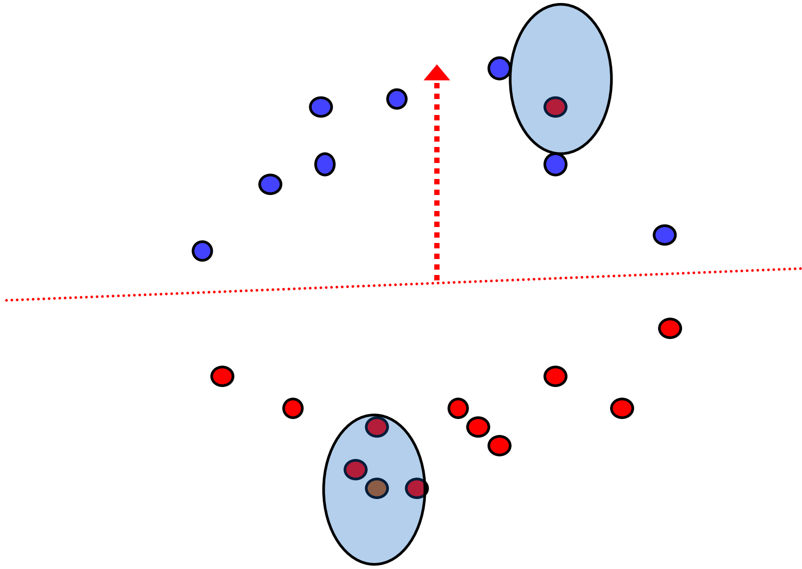
$$\min_{|w|\leq 1} \frac{1}{m} \sum_i \ell(w^\top x_i, y_i)$$

$$w_{t+1} = \mathrm{w_t} - \eta \frac{1}{m} \sum_i \ell'(w_t^\top x_i, y_i) x_i$$



- Complexity? $\frac{1}{\epsilon^2}$ iterations, each taking ~ linear time in data set
- Overall $O\left(\frac{md}{\epsilon^2}\right)$ running time, m=# of examples in R^d
- Can we speed it up??

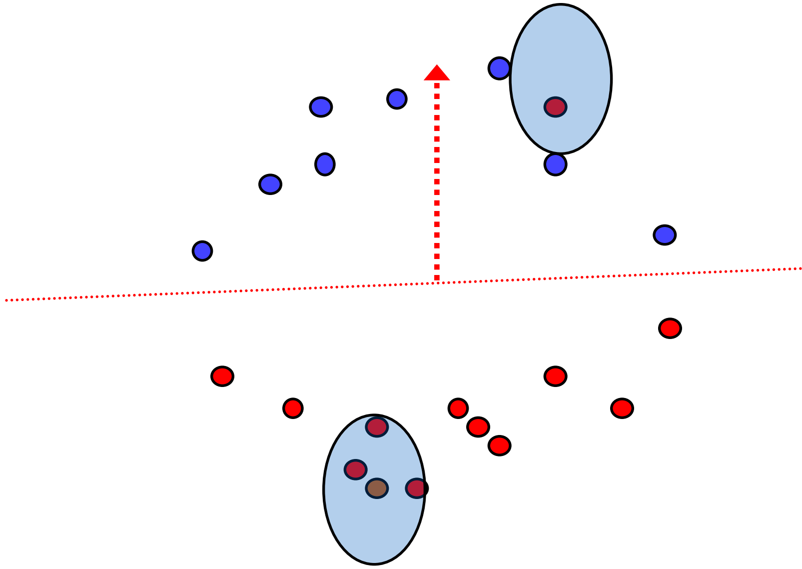
GD for linear classification



- What if we take a single example,
  and compute gradient only w.r.t it's loss??
- Which example?
- --> uniformly at random…
- Why would this work?

SGD for linear classification

$$\min_{|w|\le 1} \frac{1}{m}\sum_i \ell(w^\top x_i, y_i)$$



$$w_{t+1} = \mathrm{w_t} - \eta\, \ell'\left(w_t^\top x_{i_t}, y_{i_t}\right) x_{i_t}$$

Has expectation = full gradient

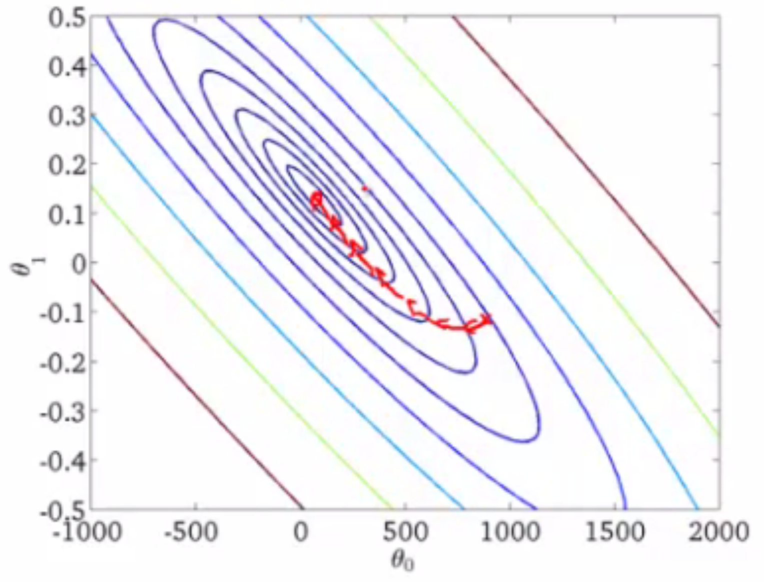
- Uniformly at random?! $i_t \sim U[1, \ldots, m]$
- Each iteration is much faster O(md) → O(d), convergence??

Crucial for SGD: linearity of expectation and derivatives

Let f(w) $= \frac{1}{m}\sum_i \ell_i(w)$, then for $i_t \sim U[1, \ldots, m]$ chosen uniformly at random, we have

$$\mathrm{E}[\nabla \ell_{i_t}(w)] = \sum_{i=1} \frac{1}{m} \nabla \ell_i(w) = \nabla \frac{1}{m} \sum_i \ell_i(w) = \nabla \ell(w)$$

Greedy optimization: gradient descent

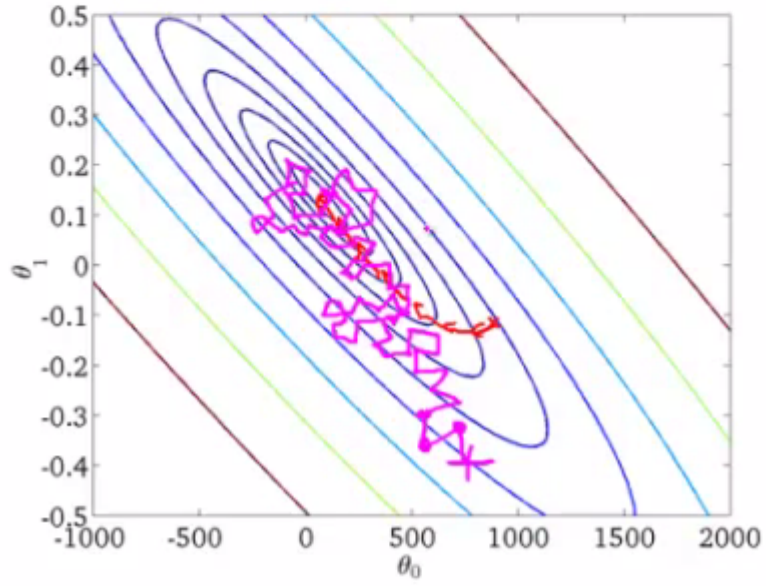


- Move in a random direction, whose expectation is the steepest descent:

- Denote by $\widetilde{\nabla f}(\mathrm{w})$ a vector random variable whose expectation is the gradient,

$$E\left[\widetilde{\nabla f}(\mathrm{w})\right] = \nabla f(w)$$

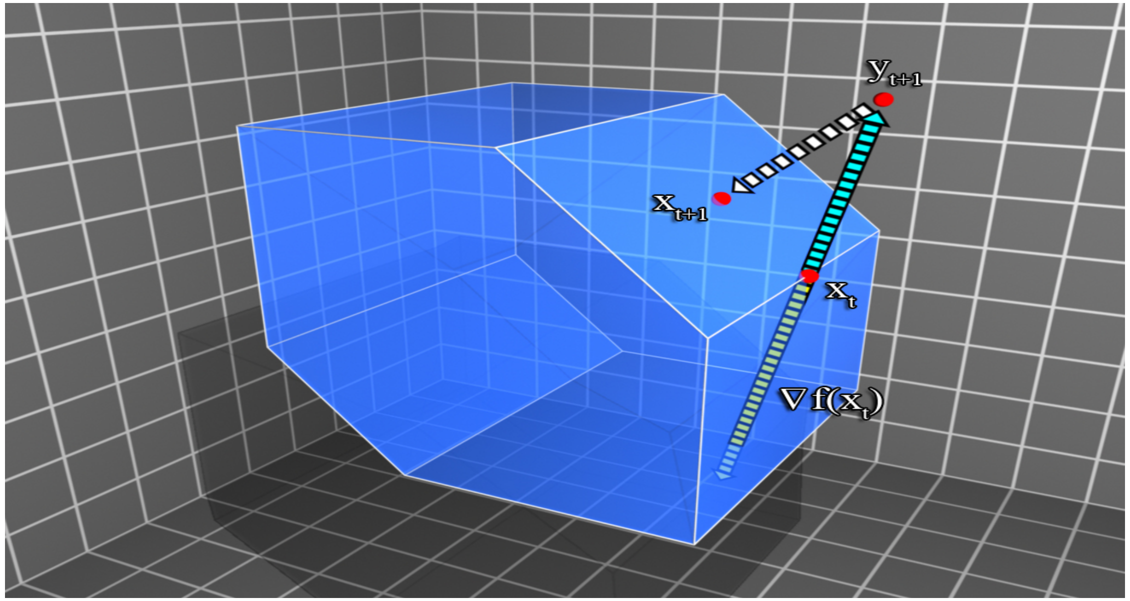
$$x_{t+1} \leftarrow x_t - \eta \widetilde{\nabla f}(x_t)$$



Stochastic gradient descent – constrained case

$$y_{t+1} \leftarrow x_t - \eta \widetilde{\nabla f(x_t)} \quad , \mathrm{E}[\widetilde{\nabla f(x_t)}] = \nabla f(x_t)$$

$$x_{t+1} = \arg\min_{x \in K} |y_{t+1} - x|$$



Stochastic gradient descent – constrained set

Let:

- G = upper bound on norm of gradient estimators

$$|\widetilde{\nabla f(x_t)}| \leq G$$

- D = diameter of constraint set

$$\forall x, y \in K \ . \ |x - y| \leq D$$

Theorem: for step size $\eta = \frac{D}{G\sqrt{T}}$

$$E[f\left(\frac{1}{T}\sum_t x_t]\right) \leq \min_{x^* \in K} f(x^*) + \frac{DG}{\sqrt{T}}$$

$$y_{t+1} \leftarrow x_t - \eta \widetilde{\nabla f(x_t)}$$

$$\mathrm{E}[\widetilde{\nabla f(x_t)}] = \nabla f(x_t)$$

$$x_{t+1} = \arg\min_{\mathrm{x} \in K} |y_{t+1} - x|$$

$$y_{t+1} \leftarrow x_t - \eta \widetilde{\nabla f(x_t)}$$

$$\mathrm{E}[\widetilde{\nabla f(x_t)}] = \nabla f(x_t)$$

$$x_{t+1} = \arg\min_{x \in K} |y_{t+1} - x|$$

Proof:

1. We have proved: (for any sequence of ∇_t)

$$\left(\frac{1}{T}\sum_t \nabla_t^\top x_t]\right) \le \min_{x^* \in K} \frac{1}{T}\sum_t \nabla_t^\top x^* + \frac{DG}{\sqrt{T}}$$

2. By property of expectation:

$$\mathrm{E}[f\left(\frac{1}{T}\sum_t x_t\right) - \min_{x^* \in K} f(x^*)] \le \left(\frac{1}{T}\sum_t \nabla f(x_t)^\top (x_t - x^*)]\right) \le \frac{DG}{\sqrt{T}}$$

Summary

- Mathematical & convex optimization
- Gradient descent algorithm, linear classification
- Stochastic gradient descent