COS 324: Lecture 9

Stochastic Gradient Descent

Elad Hazan Yoram Singer

Admin

- HW4

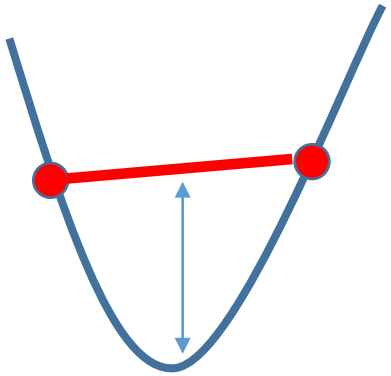
Recap

- Started convex analysis
- Gradient descent + analysis
- Today: faster optimization through randomization: stochastic gradient descent
  - Part I: how to deal with constraints?
  - Part 2: SGD

Convex Functions and Sets

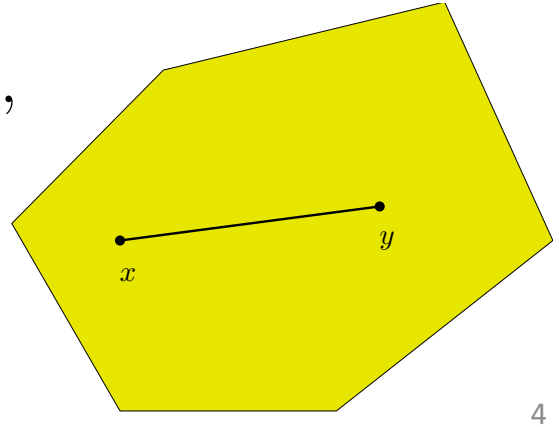
A function $f : \mathbb{R}^n \to \mathbb{R}$ is convex if for $x, y \in \mathrm{dom} f$ and any $a \in [0, 1]$,

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



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

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

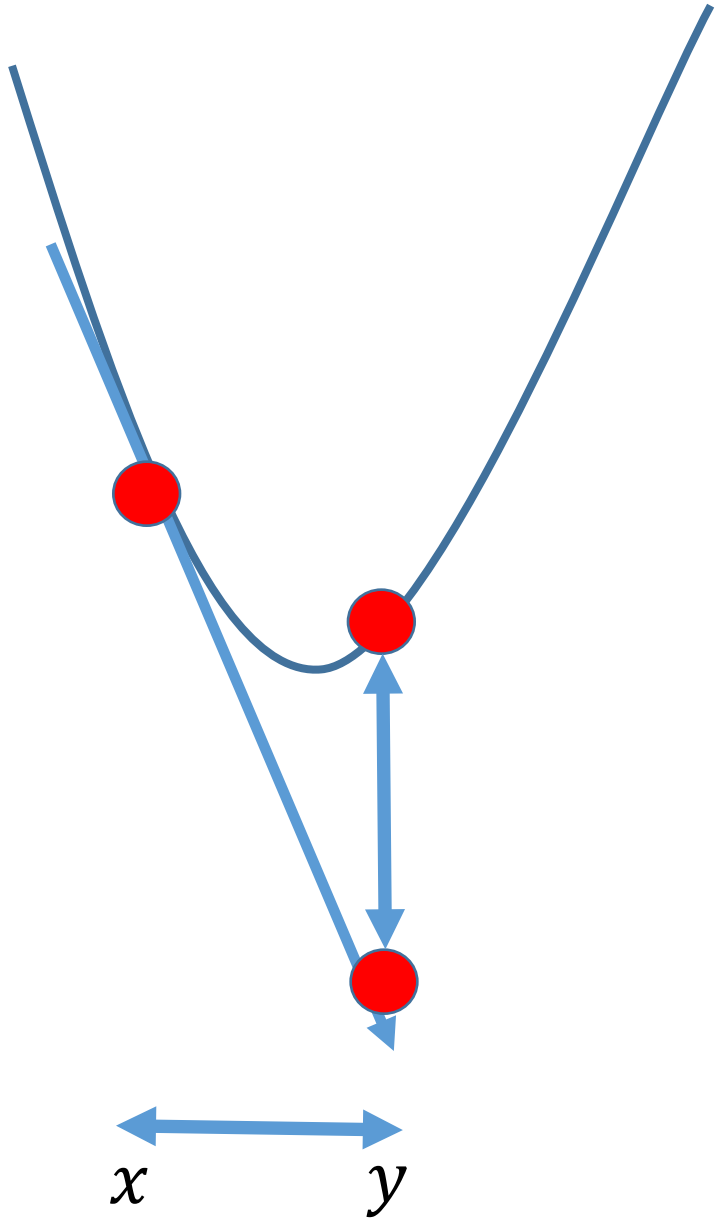


Convexity: local → global

Convexity

- Alternative definition:

$$\mathrm{f(y)} \geq \mathrm{f(x)} + \nabla f(x)^{\mathsf{T}}(y - x)$$



Lipschitzness

f is G-Lipschitz if for every $x, y \in K$, we have

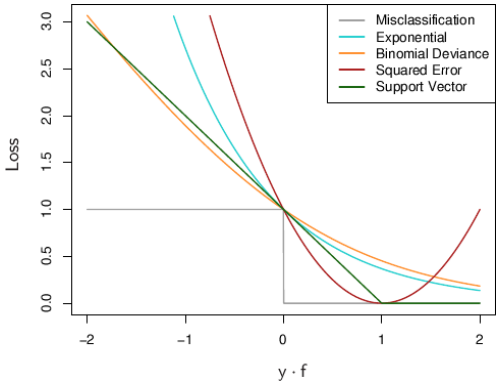
$$|f(x) - f(y)| \leq G\,|x - y|$$

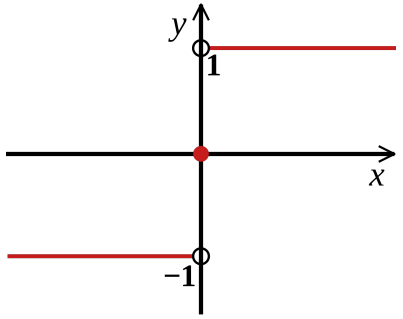
Note: for convex functions, suffices that the gradient is bounded (why?)

$$\forall x \in K \quad |\nabla f(x)| \leq G$$

Recall for convex functions:

$$f(x) - f(y) \leq \nabla f(x)^{\mathsf{T}}(x - y) \leq |\nabla f(x)||x - y| \leq G|x - y|$$





Optimality conditions

x^* is the minimum of convex function f iff

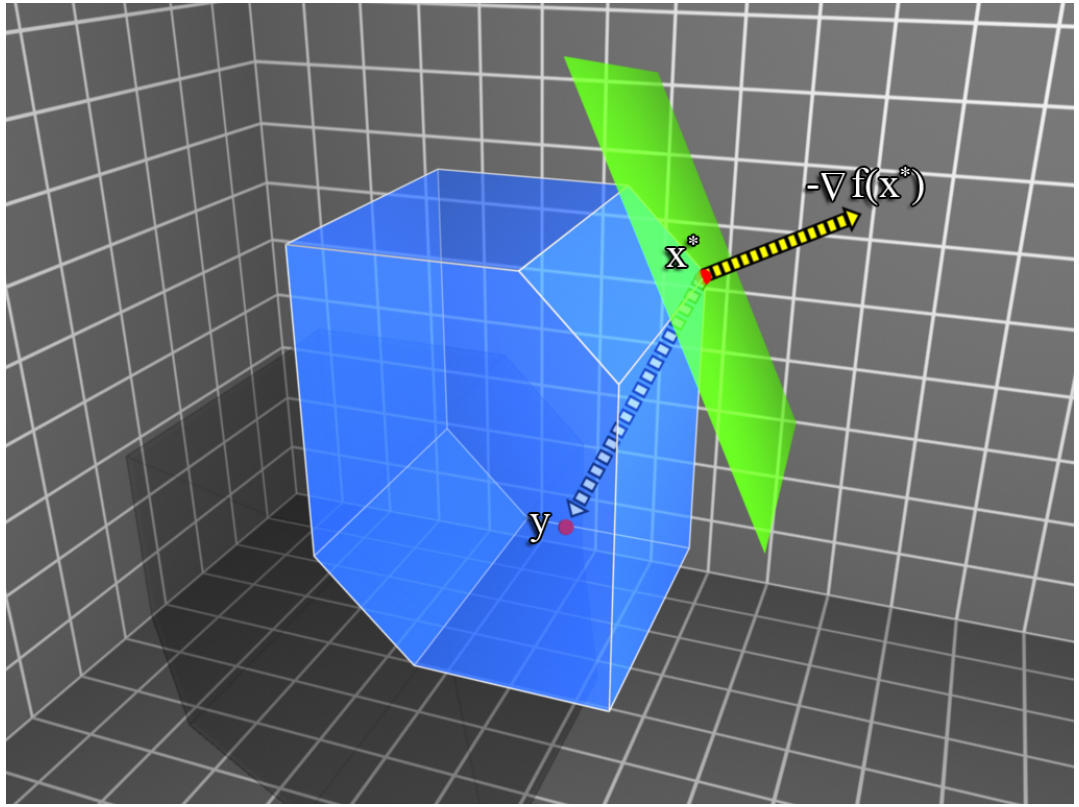
$$|\nabla f(x^*)| = 0$$

If we have a constrained set K, then it is optimum iff

$$\prod_K [x^* - \nabla f(x^*)] = x^*$$

Here $\prod_K$ denotes the projection operation, defined as:

$$\prod_K [y] = \arg\min_{x \in K} |x - y|$$



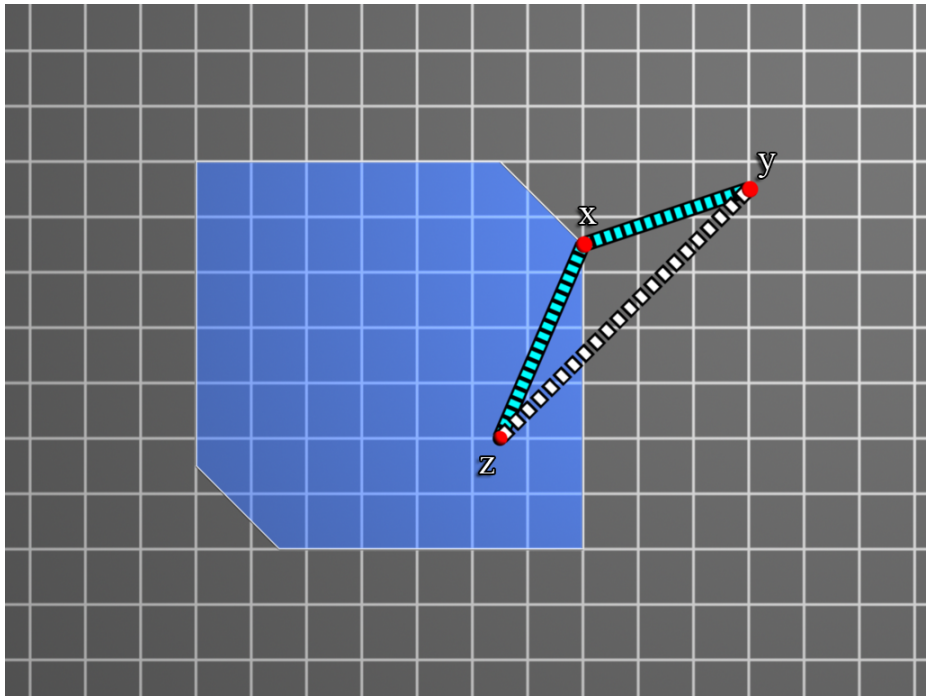
Projections

For projections over convex sets, defined as

$$\prod_{K}[y] = \arg\min_{x \in K}|x - y|$$

We have the Pythagorean theorem:

$$|y - x|^2 \leq |y - z|^2$$



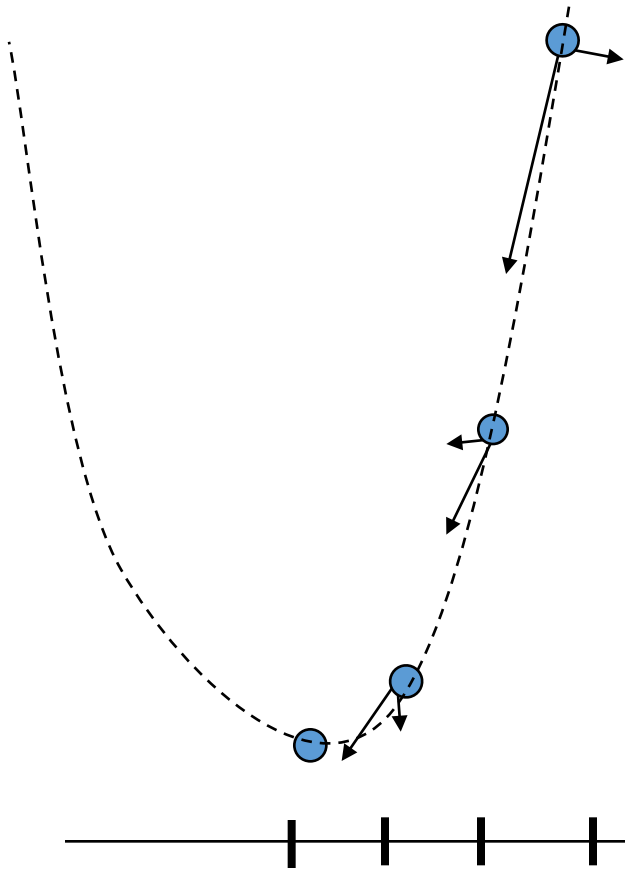
Greedy optimization: gradient descent

- Move in the direction of steepest descent, which is:

$$-[\nabla f(x)]_i = -\frac{\partial}{\partial x_i} f(x)$$

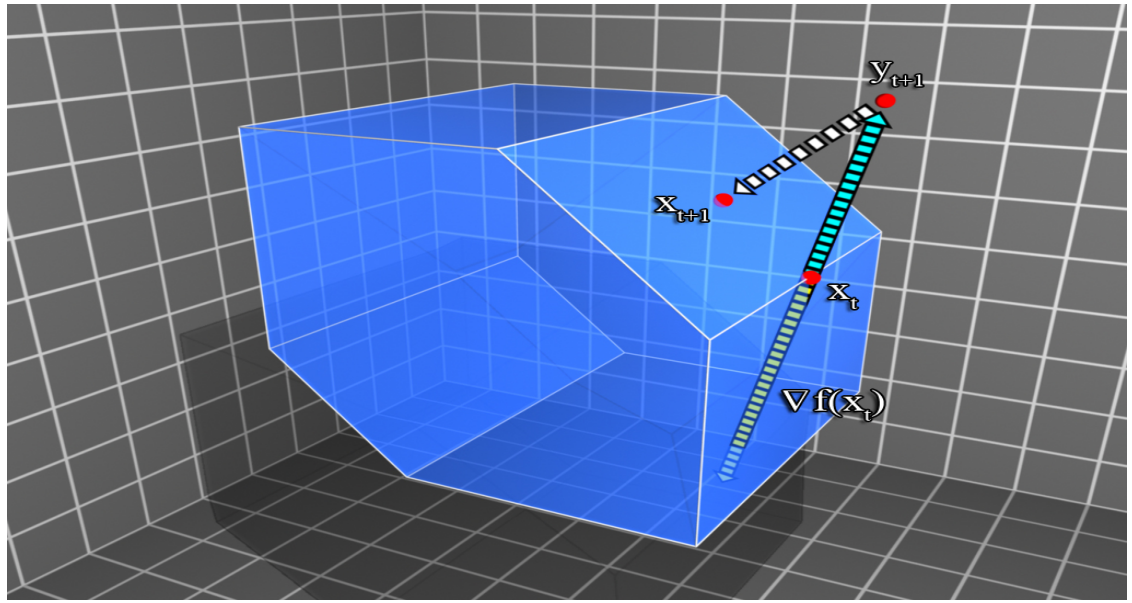
$$x_{t+1} \leftarrow x_t - \eta \nabla f(x_t)$$

"step size" or "Learning rate"



gradient descent – constrained set

$$y_{t+1} \leftarrow x_t - \eta \nabla f(x_t)$$
$$x_{t+1} = \arg\min_{x \in K} |y_{t+1} - x|$$



gradient descent – constrained set

Let:

- G = upper bound on norm of gradients

$$|\nabla f(x_t)| \leq G$$

$$y_{t+1} \leftarrow x_t - \eta \nabla f(x_t)$$
$$x_{t+1} = \arg\min_{x \in K} |y_{t+1} - x|$$

- D = diameter of constraint set

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

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

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

online gradient descent

Sequence of functions $f_1, f_2, \ldots, f_T$. Let:

- G = upper bound on norm of gradients

$$|\nabla f_t(x_t)| \leq G$$

$$y_{t+1} \leftarrow x_t - \eta \nabla f_t(x_t)$$
$$x_{t+1} = \arg\min_{x \in K} |y_{t+1} - x|$$

- D = diameter of constraint set

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

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

$$\sum_t f_t(x_t) \leq \min_{x^* \in K} \sum_t f_t(x^*) + 2DG\sqrt{T}$$

Proof:

$$y_{t+1} \leftarrow x_t - \eta \nabla f_t(x_t)$$
$$x_{t+1} = \arg\min_{x \in K} |y_{t+1} - x|$$

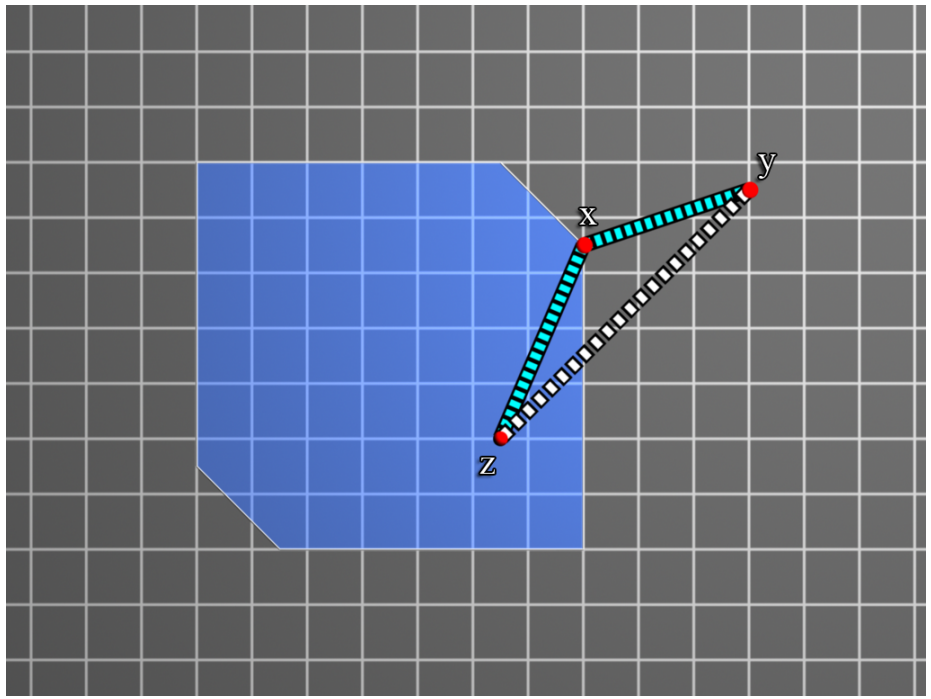
1. Observation 1:

$$|x^* - y_{t+1}|^2 = |x^* - x_t|^2 - 2\eta \nabla f_t(x_t)(x_t - x^*) + |\nabla f_t(x_t)|^2$$

2. Observation 2:

$$|x^* - x_{t+1}|^2 \leq |x^* - y_{t+1}|^2$$

This is the Pythagorean theorem:



Proof:

$$y_{t+1} \leftarrow x_t - \eta \nabla f_t(x_t)$$
$$x_{t+1} = \arg\min_{x \in K} |y_{t+1} - x|$$

1. Observation 1:

$$|x^* - y_{t+1}|^2 = |x^* - x_t|^2 - 2\eta \nabla f_t(x_t)(x_t - x^*) + \eta^2 |\nabla f_t(x_t)|^2$$

2. Observation 2:

$$|x^* - x_{t+1}|^2 \leq |x^* - y_{t+1}|^2$$

Thus:

$$|x^* - x_{t+1}|^2 \leq |x^* - x_t|^2 - 2\eta \nabla f_t(x_t)(x_t - x^*) + \eta^2 G^2$$

And hence:

$$\sum_t f_t(x_t) - \sum_t f_t(x^*) \leq \sum_t \nabla f_t(x_t)(x_t - x^*)$$

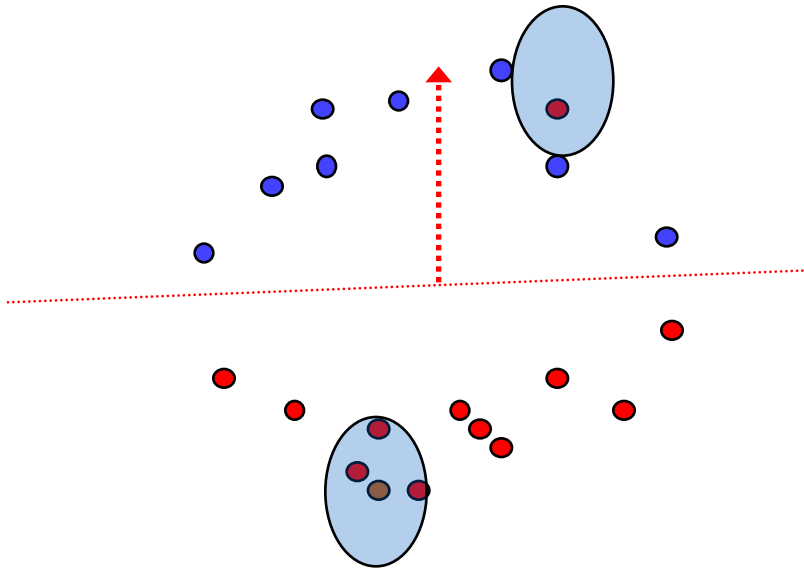
$$\leq \sum_t \frac{1}{2\eta}(|x^* - x_t|^2 - |x^* - x_{t+1}|^2) + \frac{\eta}{2} T G^2$$

$$\leq \frac{1}{2\eta} D^2 + \frac{\eta}{2} T G^2 \leq DG\sqrt{T}$$

GD for linear classification

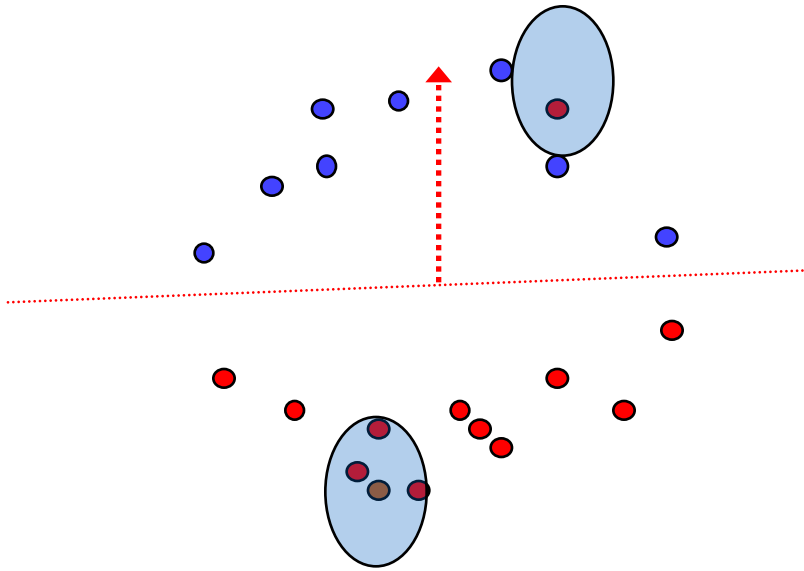
$$\min_{|w|\leq 1} \frac{1}{m} \sum_i \ell(w^\top x_i, y_i)$$

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



- 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??

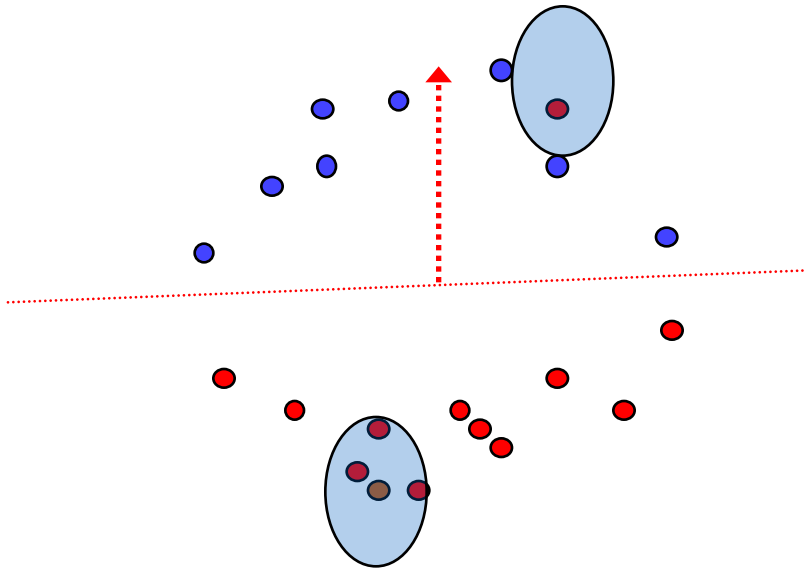
- Which example?
- --> uniformly at random…

- Why would this work?

SGD for linear classification

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

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



Has expectation = full gradient

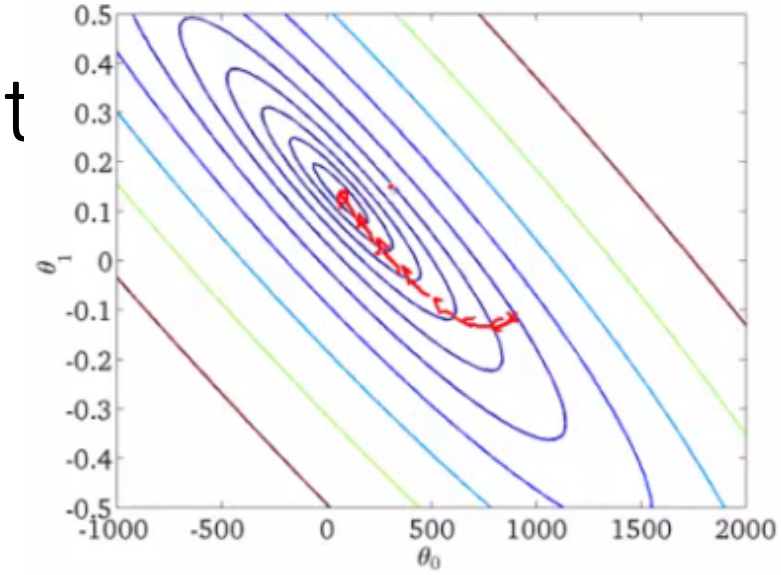
- Uniformly at random?! $i_t \sim U[1, \ldots, m]$
- Each iteration is much faster O(md) → O(d), convergence??

Crucial for SGD: linearity of expectation and derivatives

Let f(w) = $\frac{1}{m}\sum_i \ell_i(w)$, then for $i_t \sim U[1, \dots, m]$ 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

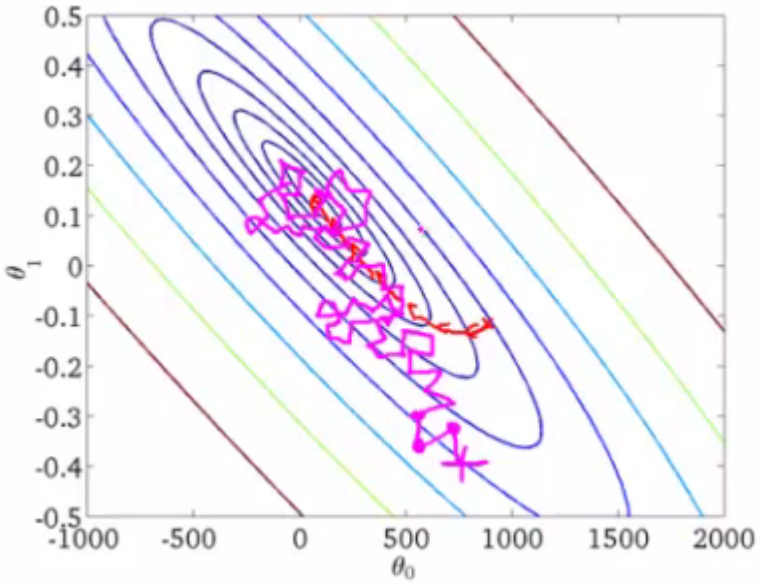


- Move in a random direction, whose expectation is the steepest descent:

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

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

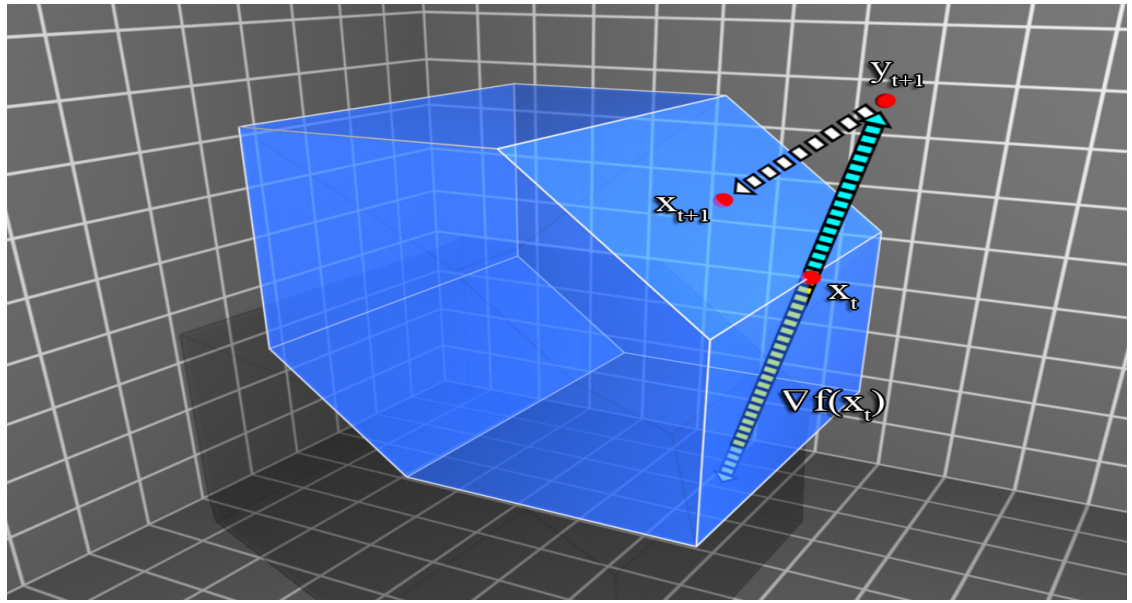
$$x_{t+1} \leftarrow x_t - \eta \widetilde{\nabla f}(x_t)$$



Stochastic gradient descent – constrained case

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

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



Stochastic gradient descent – constrained set

Let:

- G = upper bound on norm of gradient estimators

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

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

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

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

- D = diameter of constraint set

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

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

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

$$y_{t+1} \leftarrow x_t - \eta \widetilde{\nabla f(x_t)}$$
$$\mathrm{E}\left[\widetilde{\nabla f(x_t)}\right] = \nabla f(x_t)$$
$$x_{t+1} = \arg\min_{x \in K} |y_{t+1} - x|$$

Proof:

- We have proved: (for any sequence of f_t)

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

- Let $f_t(x) = \widetilde{\nabla f(x_t)}^\top x = \tilde{\nabla}_t^\top x$, then

$$\left(\frac{1}{T}\sum_t \tilde{\nabla}_t^\top x_t\right] \leq \min_{x^* \in K} \frac{1}{T}\sum_t \tilde{\nabla}_t^\top x^* + \frac{DG}{\sqrt{T}}$$

- By property of expectation:

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

Recap

$$y_{t+1} \leftarrow x_t - \eta \widetilde{\nabla f(x_t)}$$
$$\mathrm{E}\left[\widetilde{\nabla f(x_t)}\right] = \nabla f(x_t)$$
$$x_{t+1} = \arg\min_{x \in K} |y_{t+1} - x|$$

- SGD convergence:

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

- Faster per-iteration step O(d) instead of O(md)

- Faster convergence? (in terms of T)

Convexity

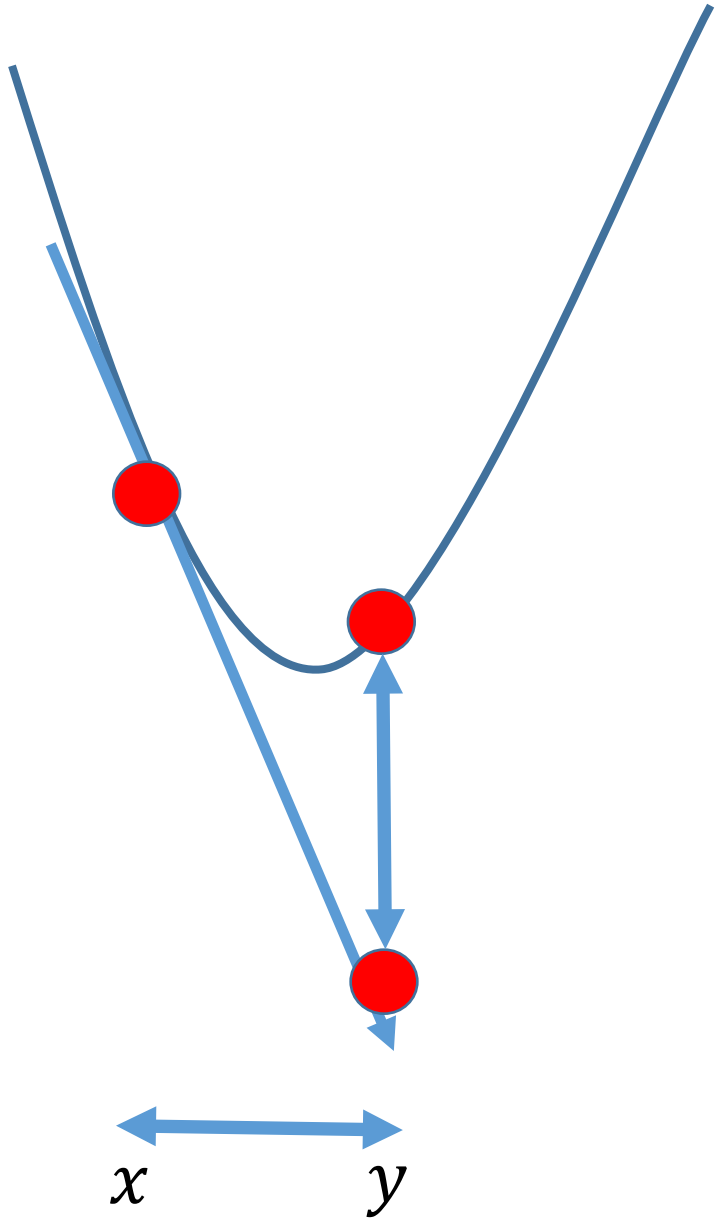
- definition:

$$\mathrm{f(y)} \geq \mathrm{f(x)} + \nabla f(x)^{\mathsf{T}}(y - x)$$

- if twice differentiable:
  second derivative is non-negative (in 1D)

- Higher dimensions:

$$\nabla^2 f(x) \succcurlyeq 0$$



Strong convexity and smoothness

Convexity:

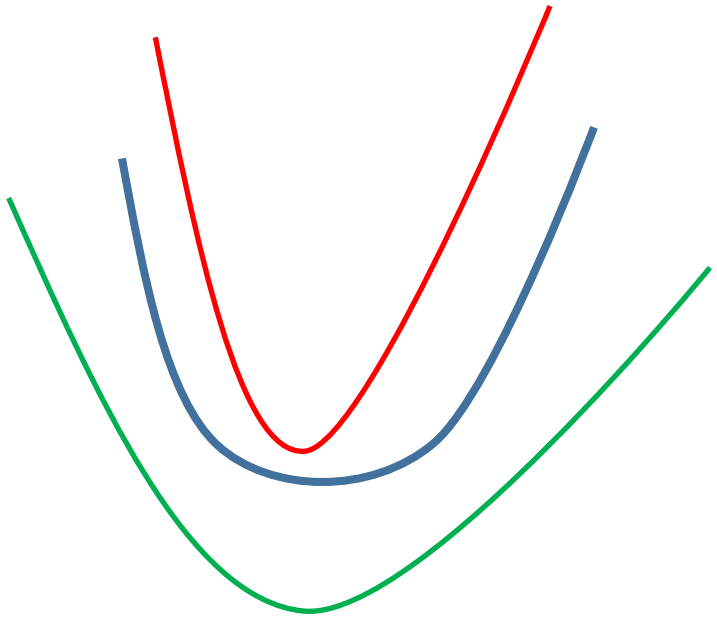
$$\nabla^2 f(x) \succcurlyeq 0$$

Strong convexity:

$$\nabla^2 f(x) \succcurlyeq \alpha I$$

Smoothness:

$$\nabla^2 f(x) \preccurlyeq \beta I$$



GD convergence:

$f(x_T) - \min_{x^* \in K} f(x^*) \leq O\left(e^{-\frac{\alpha}{\beta}t}\right)$ (as opposed to $\frac{DG}{\sqrt{T}}$)

Summary

- (online) gradient descent algorithm with constraints
- Stochastic GD
- Conditions that imply faster convergence