COS 324: Lecture 8

Gradient Descent

Elad Hazan Yoram Singer

Admin

- HW4

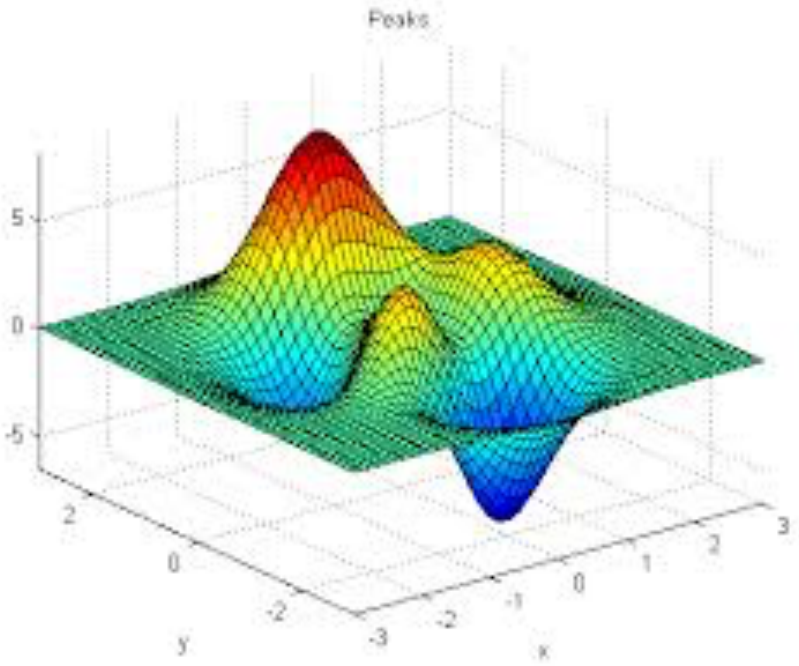
Recap

- Online learning, RWM
- statistical learning, motivated efficient algorithms/optimization
- Perceptron
- Started convex analysis
- Today: convex optimization and gradient descent

Mathematical optimization

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

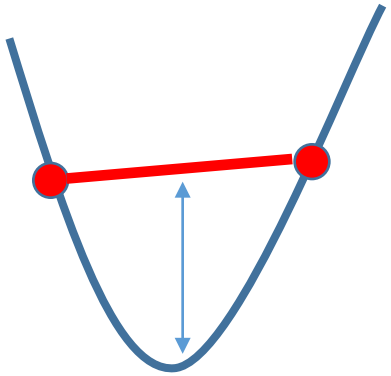
Output: point $x \in K$, such that $f(x) \leq f(y)\ \forall\ y \in K$



Convex Functions and Sets

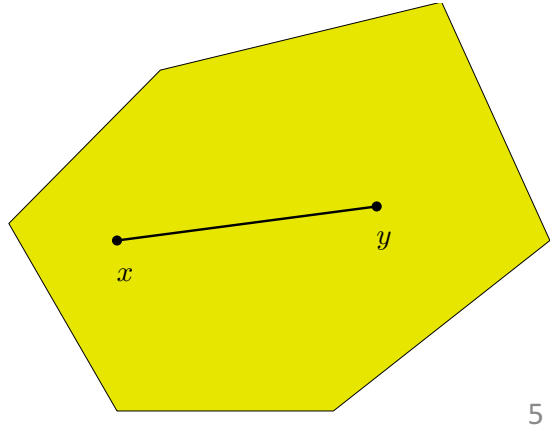
A function $f : \mathbb{R}^n \to \mathbb{R}$ is convex if for $x, y \in \mathrm{dom} f$ and any $a \in [0, 1]$,

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



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

- 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

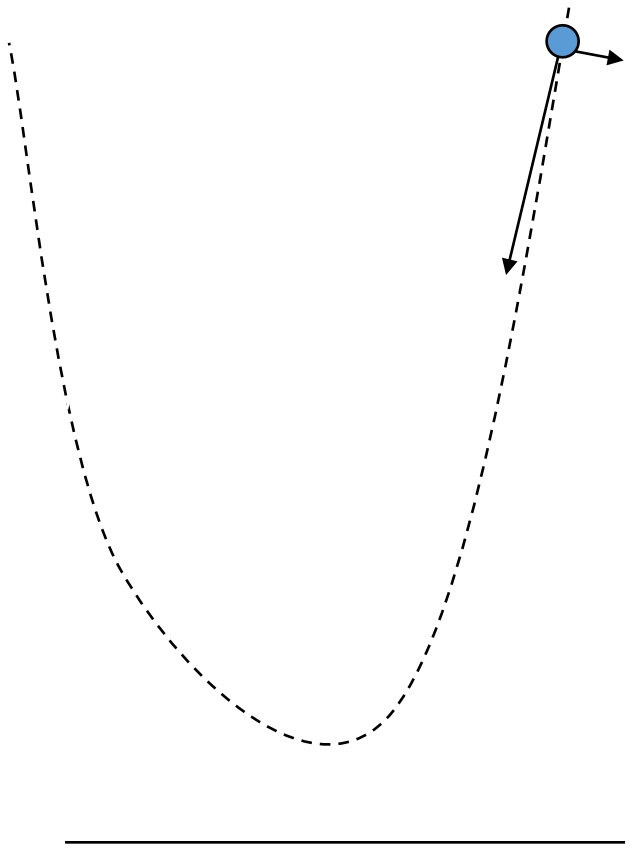
$$\forall y \in \mathrm{B_r(x)}.\, f(y) \geq f(x)$$

Calculus reminder: gradient

- Gradient = the direction of steepest descent, which is the derivative in each coordinate:

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

- Example: $f(x) = \log(w^\top x),\ f(x) = \max\{0, 1 - w^\top x\}$

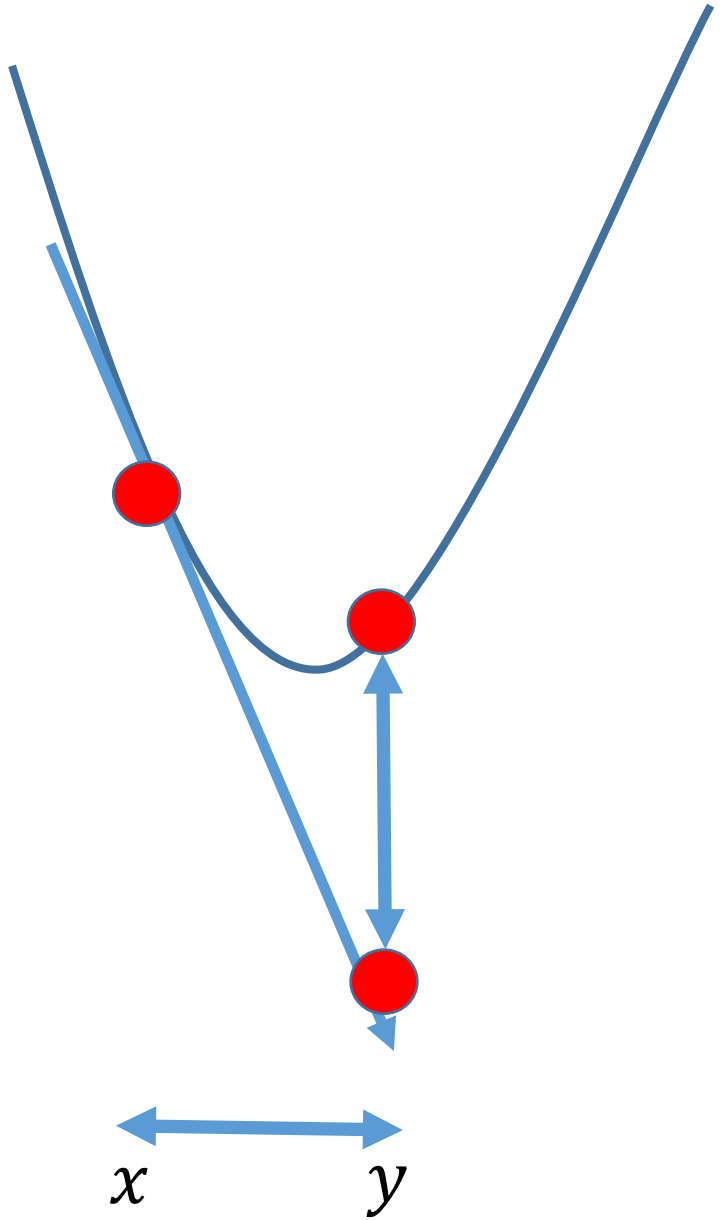


Convexity

- Alternative definition:

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

(assumes differentiability, o/w subgradient)
(another alternative: second derivative is non-negative in 1D)



Lipschitzness

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

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

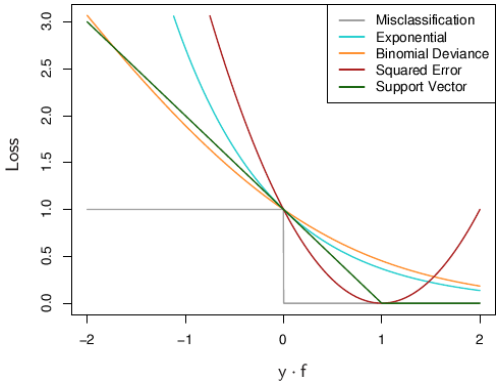
Note: for convex functions, suffices that the gradient is bounded (why?)

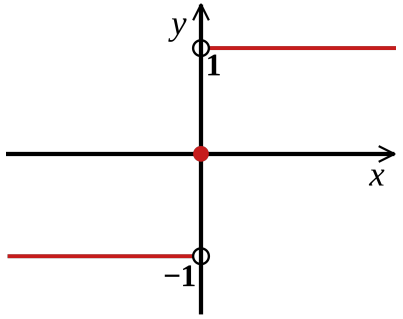
$$\forall x \in K \quad |\nabla f(x)| \leq G$$

Recall for convex functions:

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

Is sign Lipschitz? Hinge-loss?





Optimality conditions

x^* is the minimum of convex function f iff

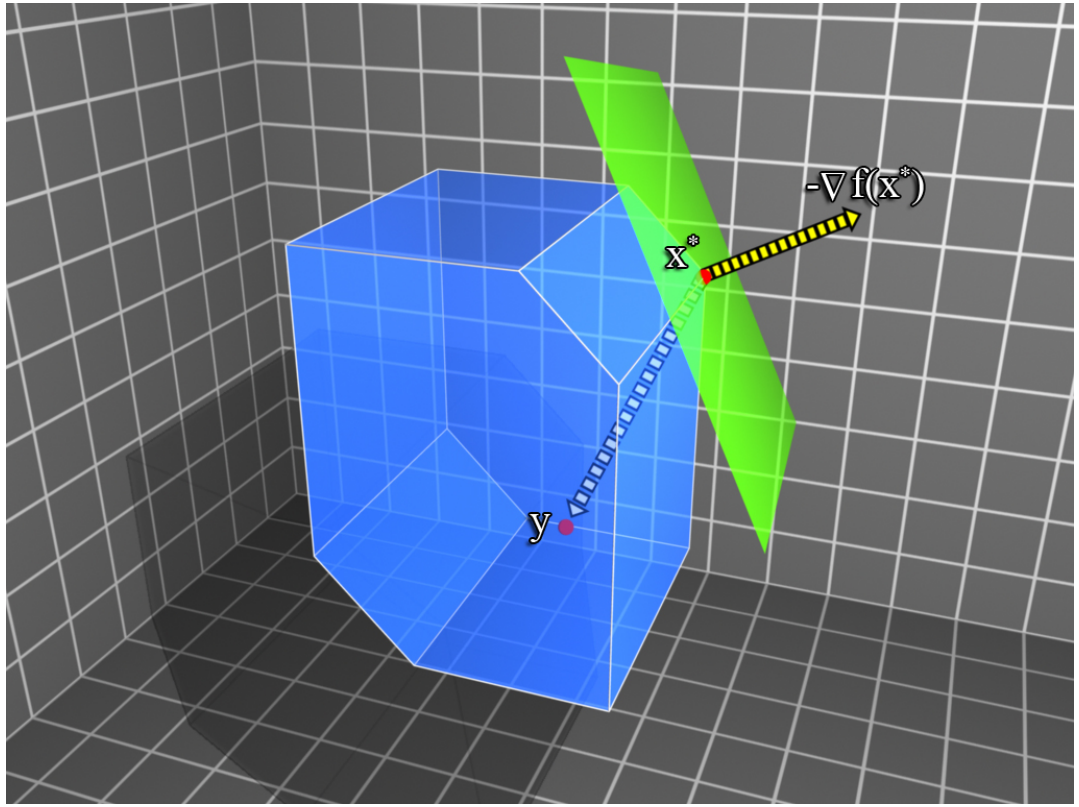
$$|\nabla f(x^*)| = 0$$

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:

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



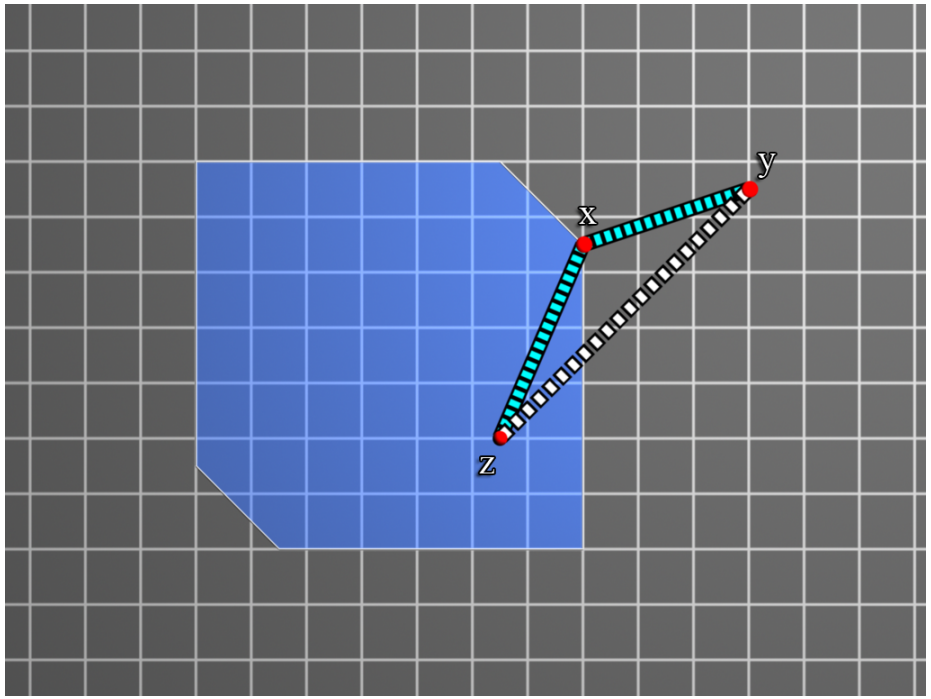
Projections

For projections over convex sets, defined as

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

We have the Pythagorean theorem:

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



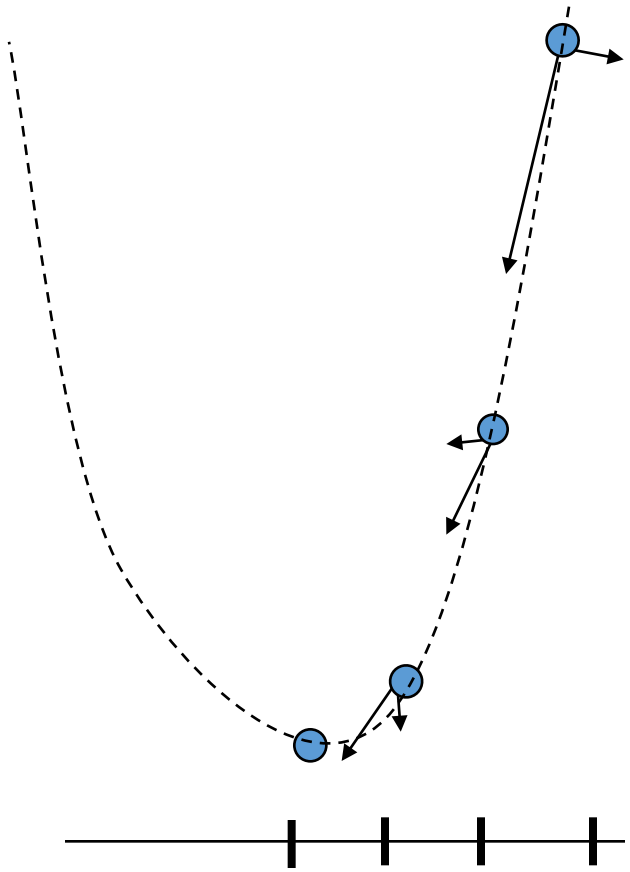
Greedy optimization: gradient descent

- 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 – unconstrained

Let:

- G = upper bound on norm of gradients

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

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

- D = distance from initial point to optimum

$$|x_1 - x^*| \leq D$$

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

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

Proof:

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

Observation:

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

Proof:

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

Observation:

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

Thus:

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

And hence:

$$f\left(\frac{1}{T}\sum_t x_t\right) - f(x^*) \leq \frac{1}{T}\sum_t [f(x_t) - f(x^*)] \leq \frac{1}{T}\sum_t \nabla f(x_t)(x_t - x^*)$$

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

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

gradient descent

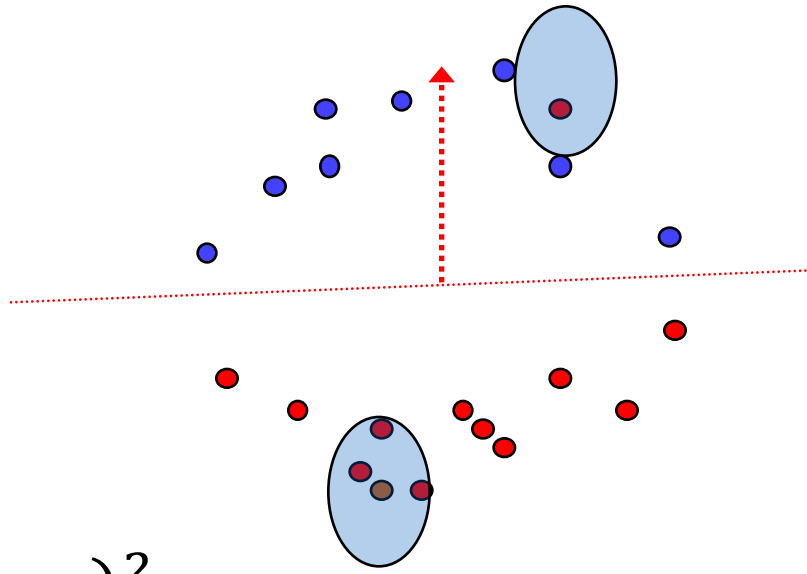
Theorem: for step size $\eta = \frac{D}{G\sqrt{T}}$

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

Thus, to get ϵ-approximate solution, apply $\frac{\mathrm{D}^2 G^2}{\epsilon^2}$ gradient iterations.

GD for linear classification

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

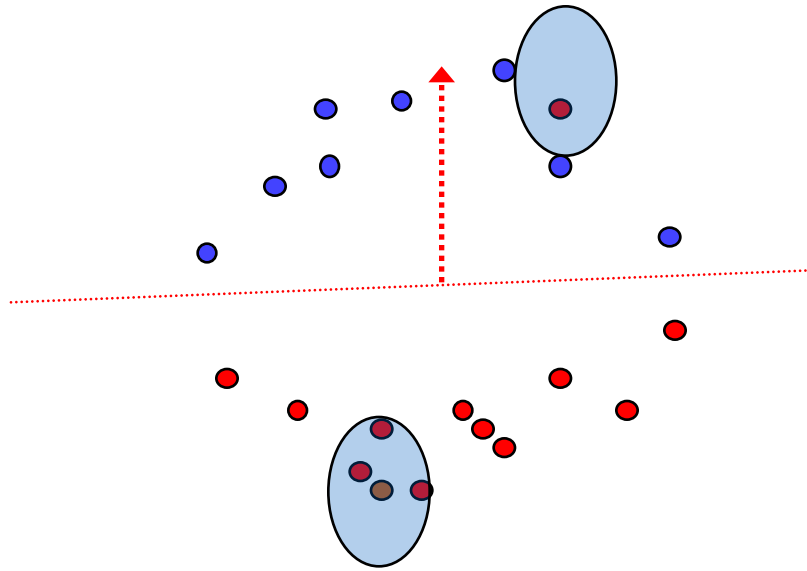


1. Ridge / linear regression $\ell(w^\top x_i, y_i) = (w^\top x_i - y_i)^2$
2. SVM $\ell(w^\top x_i, y_i) = \max\{0, 1 - y_i\ w^\top x_i\}$
3. Logistic regression $\ell(w^\top x_i, y_i) = \log(1 + e^{y_i\, w^\top x_i})$

GD for linear classification

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

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



- 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??

Summary

- Mathematical optimization for linear classification
- Convex relaxations
- Gradient descent algorithm
- GD applied to linear classification