Convex Optimization

(1) Convexification of the 0-1 loss

A problem for classification

$$\hat{f}_{\mathcal{F}} \in \underset{f \in \mathcal{F}}{\arg\min} \left\{ R_n(f) = \frac{1}{n} \sum_{i=1}^{n} \mathbb{1}\{ f(X_i) \neq Y_i \} \right\}$$

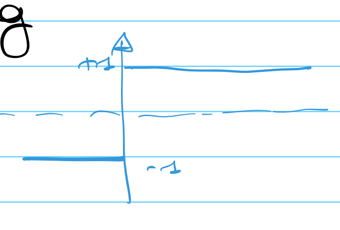
⇝ How to compute this ?

The function $f \mapsto R_n(f)$ is discontinuous and takes $\underbrace{N_{\mathcal{F}}(x_1, \dots, x_n)}_{n^{VC(\mathcal{F})} \,=\, \text{large}}$ values !

(ex: if $VC(\mathcal{F}) > 3$)

Alternative point of view :

Look for $f = \underbrace{\mathrm{sgn}}_{} \circ g$

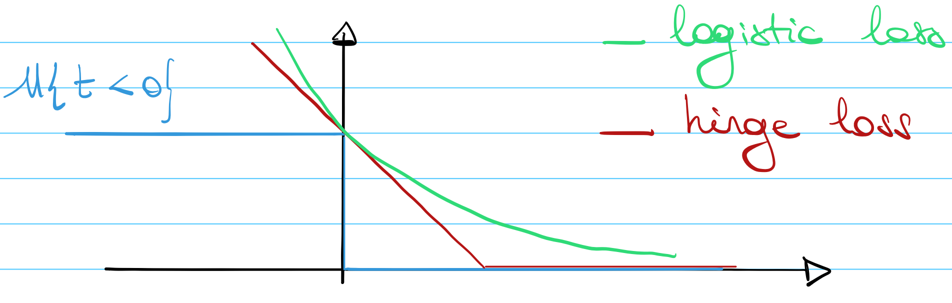


$\rightsquigarrow \quad \mathcal{R}_P(f) = \mathbb{E}_P\left[\mathbb{1}\{f(X) \neq Y\}\right]$

$= \mathbb{E}_P\left[\mathbb{1}\{\operatorname{sgn}(g(X)) \neq Y\}\right]$

$= \mathbb{E}_P\left[\mathbb{1}\{g(X)Y < 0\}\right]$

$\rightsquigarrow$ Still discontinuous ...



$\rightsquigarrow$ Replace the 0-1 loss by a convex surrogate.

$\hookrightarrow$ convex functions can be optimized efficiently.

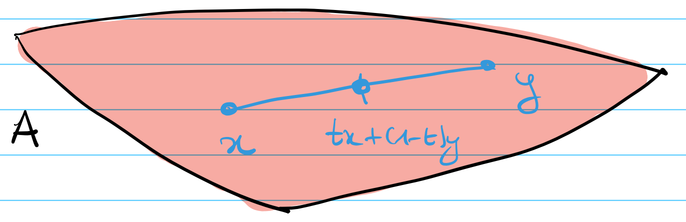
Logistic Regression (Later today)

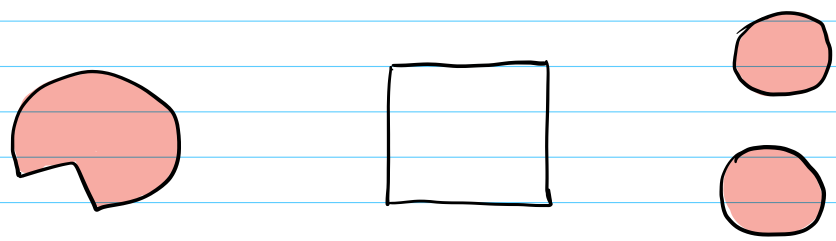
Support Vector Machines (Ch. 3)

② Convex functions :

- **Convex Sets** : $A \subset \mathbb{R}^d$

$\forall x, y \in A, \; t \in [0,1], \quad tx + (1-t)y \in A$



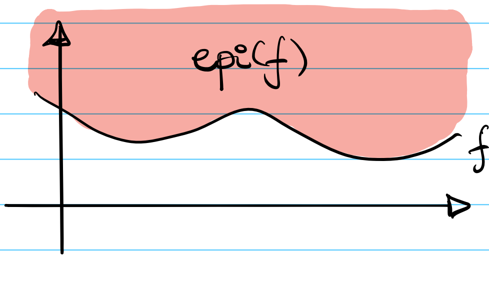


Not convex

- **Epigraph** of $f : A \to \mathbb{R}$

$$\left[\; epi(f) = \{ (x,y) \in A \times \mathbb{R} \; : \; y \geq f(x) \} \; \right]$$

$\subseteq \mathbb{R}^{d+1}$



- **Convex function:** A convex set
  $f : A \to \mathbb{R}$

f is convex if $\text{epi}(f)$ is convex.

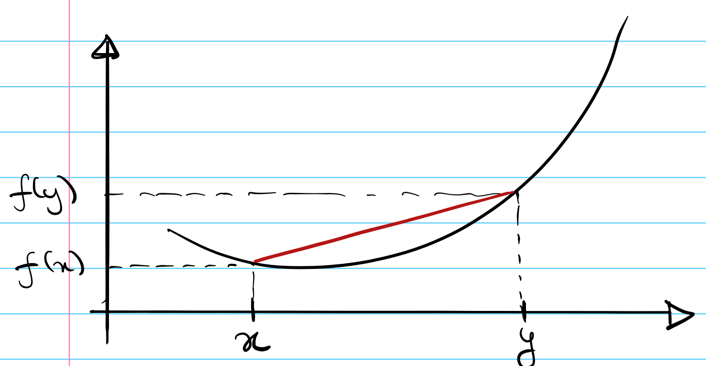
$\leadsto$ We have $(x_1, f(x_1)), (x_2, f(x_2)) \in \text{epi}(f)$.

So $t(x_1, f(x_1)) + (1-t)(x_2, f(x_2)) \in \text{epi}(f)$

$\leadsto$ $f(tx_1 + (1-t)x_2) \leq t f(x_1) + (1-t) f(x_2)$

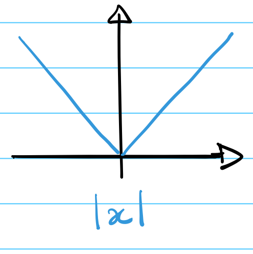
$\Longleftrightarrow$ f convex if $\forall\, x, y \in A,\ t \in [0,1]$

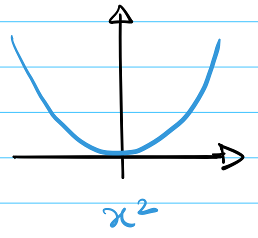
$$f(tx + (1-t)y) \leq t f(x) + (1-t) f(y)$$

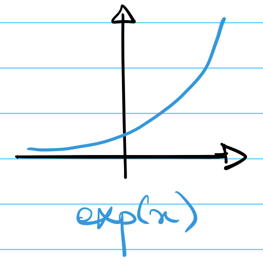


Chord is above the graph.

Examples:



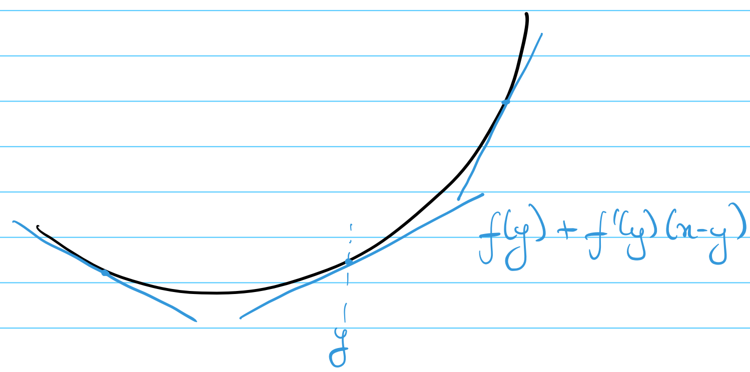




Properties of convex functions :

- If f is differentiable + convex
  - $d=1$ f' is non decreasing
  - $d\geq 2$ $\langle \nabla f(x) - \nabla f(y), x-y \rangle \geq 0$

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



- If f is twice differentiable + convex
  - $d=1$ $f'' \geq 0$
  - $d \geq 2$ $\forall x$, $\underbrace{\nabla^2 f(x) \succeq 0}$.

symmetric positive definite matrix H

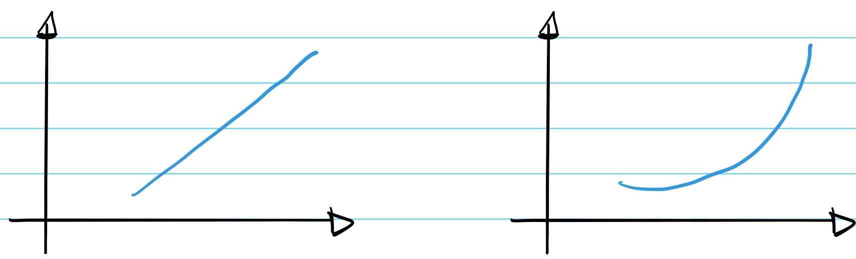
$\forall u \in \mathbb{R}^d$ $u^T H u \geq 0$

$\rightsquigarrow$ Let $g_u : t \mapsto f(x+tu)$

g_u cvx : $g_u'(t) = \langle \nabla f(x+tu), u \rangle$

$g_u''(t) = \langle \nabla^2 f(x+tu)u, u \rangle$

$\rightsquigarrow$ $g_u''(0) = u^T \nabla^2 f(x) u \geq 0$.



Both functions are convex ...

But one is "more" convex.

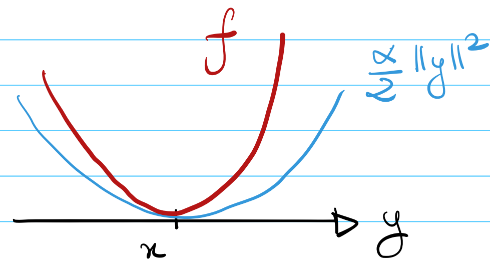
- α-strongly convex function

$\forall x, y$, $t \in [0,1]$

$$f(tx+(1-t)y) \leq tf(x) + (1-t)f(y) - \frac{\alpha}{2} t(1-t)\|x-y\|^2$$

f differentiable $\Rightarrow$ $\left[f(y) \geq f(x) + \langle \nabla f(x), y-x \rangle + \frac{\alpha}{2} \|x-y\|^2 \right]$

$$\begin{cases} f(x) = 0 \\ \nabla f(x) = 0 \end{cases}$$



f is sufficiently curved

f twice differentiable $\Rightarrow$ $\boxed{\nabla^2 f(x) \succcurlyeq \alpha \, \mathrm{Id}}$

$\rightsquigarrow$ $H \succcurlyeq \alpha \, \mathrm{Id} \iff u^T H u \geq \alpha \|u\|^2 \quad \forall u \in \mathbb{R}^d$

$\hookrightarrow u^T H u = \sum_i \lambda_i \langle u, e_i \rangle^2$

$(e_1, \dots, e_d)$ orthonormal basis of eigenvectors
$\lambda_1, \dots, \lambda_d$ associated eigenvalues

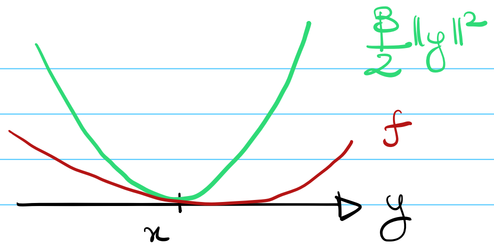
$\rightsquigarrow \lambda_i \geq \alpha$.

- β- smooth function:

$$\forall x, y \qquad \|\nabla f(x) - \nabla f(y)\| \leq \beta \|x - y\|$$

$$\iff \forall x, y \qquad f(y) \leq f(x) + \langle \nabla f(x), y - x \rangle + \frac{\beta}{2} \|y - x\|^2$$

twice differentiable $\iff$ $\boxed{\nabla^2 f(x) \preccurlyeq \beta \, \mathrm{Id}}$



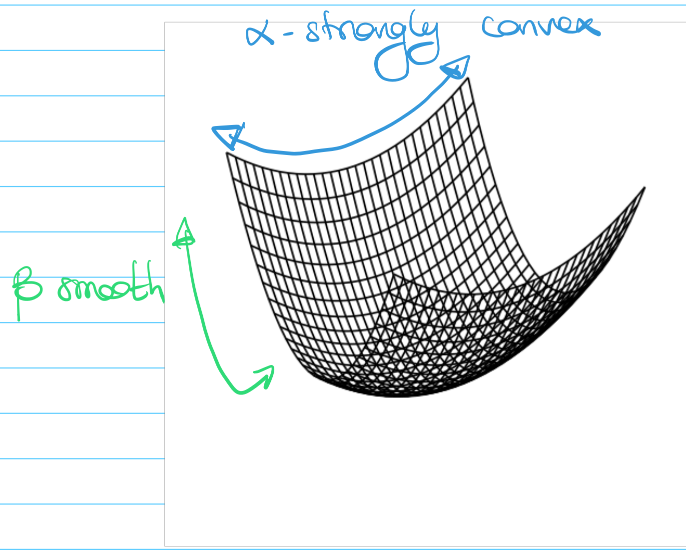
f is not too curved

Example : A function α-strongly convex and β-smooth is

$\alpha \leq \beta$

$$f_{\alpha,\beta} : x \in \mathbb{R}^2 \longmapsto \frac{\alpha}{2} x_1^2 + \frac{\beta}{2} x_2^2 .$$

$$\nabla^2 f_{\alpha,\beta}(x) = \begin{pmatrix} \alpha & 0 \\ 0 & \beta \end{pmatrix} \begin{cases} \succeq \alpha \, \mathrm{Id} \\ \preceq \beta \, \mathrm{Id} \end{cases}$$



③ Gradient Descent

GOAL : Find the minimum of a convex function f.

- Start at a point x.
- Make a step of size r in some direction h

$$f(x+rh) \simeq f(x) + r\langle \nabla f(x), h\rangle$$

$\rightsquigarrow$ Largest decrease for $h = -\nabla f(x)$.

Def : $\begin{cases} f : \mathbb{R}^d \longrightarrow \mathbb{R} \\ x_0 \quad \text{initialization} \\ s > 0 \quad \text{step size} \end{cases}$

GRADIENT DESCENT $\quad \boxed{x^{t+1} = x^t - s\nabla f(x^t)}$

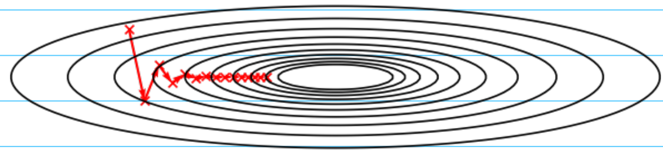
THEOREM : Let f α-strongly convex and β-smooth.

Take $s < 1/\beta$. Let $x^* = \arg\min f$.

Gradient $\Rightarrow$ Descent

$$f(x^T) - f(x^*) \leq \underbrace{(1-\alpha s)^T}_{\leq \exp(-\alpha s T)} (f(x^0) - f(x^*))$$

Example GD on $f_{\alpha,\beta}$



proof:

$$f(x^{t+1}) = f(x^t - s\nabla f(x^t))$$

$$\overset{\beta\text{-smooth}}{\leq} f(x^t) + \underbrace{\langle \nabla f(x^t), -s\nabla f(x^t)\rangle}_{-s\|\nabla f(x^t)\|^2} + s^2 \frac{\beta}{2}\|\nabla f(x^t)\|^2$$

$$= f(x^t) + s\Big(\underbrace{\frac{s\beta}{2} - 1}_{<0}\Big)\|\nabla f(x^t)\|^2$$

Polyak - Łojasiewicz inequality:

PL $$\left[\; f(x) - f(x^*) \leq \frac{1}{2\alpha}\|\nabla f(x)\|^2 \;\right]$$

proof: $f(x^*) \geq f(x) + \langle \nabla f(x), x^* - x\rangle + \frac{\alpha}{2}\|x^* - x\|^2$

$\leadsto$ $f(x) - f(x^*) \leq \langle \nabla f(x), x - x^*\rangle - \frac{\alpha}{2}\|x^* - x\|^2$

$$\Rightarrow \left[\; f(x^{t+1}) - f(x^*) \leq (f(x^t) - f(x^*))\Big(\underbrace{1 + 2\alpha s\Big(\frac{s\beta}{2} - 1\Big)}_{\leq\, 1 - \alpha s}\Big) \;\right]$$

$s < \frac{1}{\beta}$

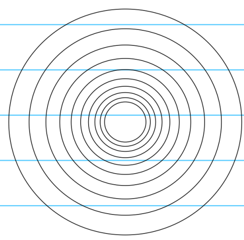
$\blacksquare$

[Gradient Descent has a Linear rate of CV.]

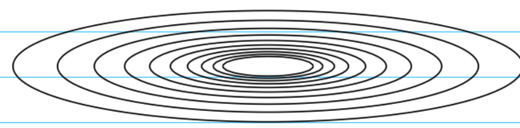
- If $s = 1/\beta$: $f(x^T) - f(x^*) \le \exp\left(-\frac{\alpha}{\beta} T\right)\left(f(x^0) - f(x^*)\right)$

$\kappa = \frac{\beta}{\alpha} \ge 1$ is the condition number.

$\kappa \gg 1$ = ill-conditioned problem
= harder to optimize



$\kappa = 1$



$\kappa \gg 1$

⇝ To get a precision ε

$$\varepsilon = \exp\left(-\frac{T}{\kappa}\right) \rightsquigarrow T = \log(\varepsilon^{-1})/\kappa.$$

iterations

Excess risk: $R_P(\hat{f}_{\mathcal{F}}) - R_P(f_P^*)$ at least of order $\frac{1}{\sqrt{n}}$.

$$\rightsquigarrow \boxed{T \simeq \frac{\log(n)}{\kappa} \text{ iterations.}}$$

⟿ What if f is only smooth? $(\alpha = 0)$

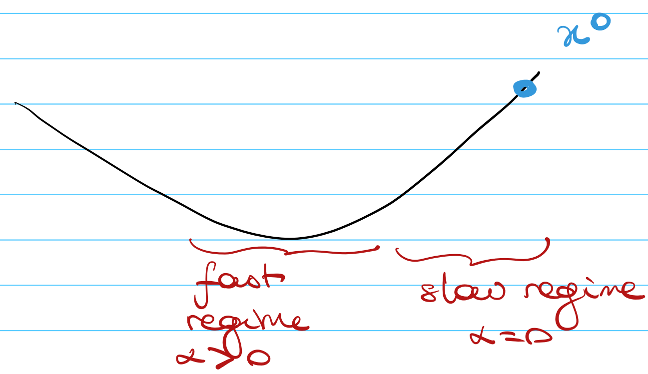
THEOREM: Let f be a convex β-smooth function.
Take $s = 1/\beta$. Let $x^* = \operatorname{argmin} f$.

Gradient ⟹ Descent

$$f(x^T) - f(x^*) \leq \frac{\beta \|x^0 - x^T\|^2}{2T}$$

⟿ MUCH slower

$T = \frac{\beta}{\varepsilon}$ iterations for error ε.



⟹ Gradient Descent will naturally adapt to the degree of convexity of f.

4 Newton's Method

SECOND ORDER APPROXIMATION:

$$f(x+h) \approx P_{f,x}(h) = f(x) + \langle \nabla f(x), h \rangle + \frac{1}{2} h^T \nabla^2 f(x) h$$

↳ min attained at $h = (\nabla^2 f(x))^{-1} \nabla f(x)$

Def: $\begin{cases} f : \mathbb{R}^d \longrightarrow \mathbb{R} \\ x_0 \text{ initialization} \end{cases}$

NEWTON'S METHOD $\left[x^{t+1} = x^t - (\nabla^2 f(x^t))^{-1} \nabla f(x^t) \right]$

THEOREM: Let f be:

- α strongly convex
- β smooth
- $\forall x, y \quad \|\nabla^2 f(x) - \nabla^2 f(y)\|_{op} \leq \gamma \|x - y\|$
- Good initialization: $\|x^0 - x^*\| \leq \alpha / \gamma$

Newton's method $\Longrightarrow$ $\left[f(x^T) - f(x^*) \leq \frac{\beta \alpha}{\gamma} 2^{-2^T + 1} \right]$

[NEWTON'S METHOD HAS A QUADRATIC RATE OF CV.]

$T = O(\underbrace{\log\log \varepsilon^{-1}}_{\text{super small}})$ iterations to get an ε-approximation.

$\rightsquigarrow$ $\log\log n$ iterations.

However: ① Compute $(\nabla^2 f(x))^{-1}$

$\longrightarrow$ $O(d^3)$ operations (naively)

Each iteration is very costly if d is large.

② Method breaks down if no α-strong convexity.

5 Logistic Regression

Recall: Binary classification

$\rightsquigarrow$ classifier of the form $f = \mathrm{sgn} \circ g$

$$\mathcal{R}_P(f) = \mathbb{E}_P[\mathbb{1}\{f(X) \neq Y\}]$$
$$= \mathbb{E}_P[\mathbb{1}\{\mathrm{sgn}(g(X)) \neq Y\}]$$
$$= \mathbb{E}_P[\mathbb{1}\{g(X)Y < 0\}]$$

We replace $\mathbb{1}\{t<0\}$ by a convex function.

Logistic loss: $t \longmapsto \log(1+e^{-t})$

$$= -\log(\sigma(t))$$

$$\sigma(t) = \frac{1}{1+e^{-t}} \in [0,1]$$

New loss: $\ell_{\log}(y,y') = \log(1+e^{-yy'})$

$$\rightsquigarrow \tilde{R}_n(g) = \frac{1}{n}\sum_{i=1}^{n} \log(1+\exp(-Y_i g(X_i)))$$

$$= \frac{1}{n}\sum_{i=1}^{n}\Big(Z_i \log(1+\exp(-g(X_i)))$$

$Z_i = \mathbb{1}\{Y_i = 1\}$
$= \begin{cases} 1 & \text{if } Y_i = 1 \\ 0 & \text{if } Y_i = -1 \end{cases}$

$$+ (1-Z_i)\log(1+\exp(g(X_i)))\Big)$$

Link with Maximum Likelihood Estimation

Statistical model: Set $\mathcal{G}$. Fix $g \in \mathcal{G}$.

Assume Y is obtained in the following way:

$$\begin{cases} Y = 1 & \text{with probability } \sigma(g(X)) = p_g(X,1) \\ Y = -1 & \text{otherwise.} \end{cases} \qquad = \frac{1}{1+e^{-g(x)}}$$

$\hookrightarrow$ with proba $p_g(x,0)$

$\rightsquigarrow P_g$ is the joint distribution of (X,Y).

$\rightsquigarrow$ We observe $(X_1,Y_1) \ldots (X_n,Y_n)$ with distribution $P_{g_0} \longrightarrow$ Estimate g_0.

Likelihood of $(X_1,Y_1) \ldots (X_n,Y_n)$ in g:

$$\prod_{i=1}^{n} p_g(X_i,Y_i) = \prod_{i=1}^{n} \sigma(g(X_i))^{\overbrace{\mathbb{1}\{Y_i=0\}}^{Z_i}} (1-\sigma(g(X_i)))^{\overbrace{\mathbb{1}\{Y_i=1\}}^{1-Z_i}}$$

Log - Likelihood:

$$\sum_{i=1}^{n} z_i \log(\sigma(g(x_i))) + (1-z_i)\log(1-\sigma(g(x_i)))$$

$$= -\tilde{R}_n(g)$$

$$\boxed{\text{Max Likelihood} = \text{Min Empirical Risk}}$$

$\rightsquigarrow$ $\tilde{g}_{\mathcal{G}}$ satisfies strong theoretical properties.

Example: Two different variety of raisins.

Eight Geometric Features

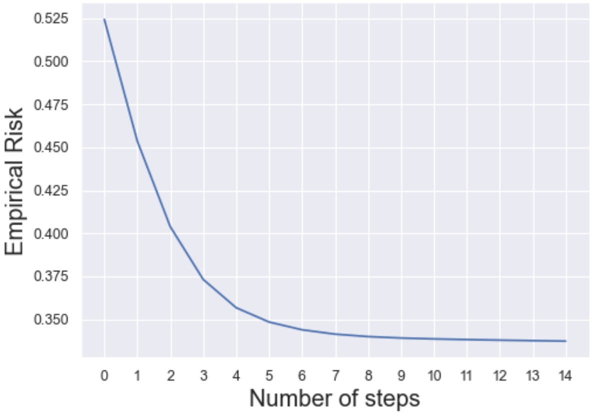
(area, perimeter, ...)

$\mathcal{G} = \{\text{Linear classifiers}\}$

$\rightsquigarrow$ Optimization with gradient descent.

Decision boundary
Kecimen
Besni

85% accuracy



Empirical Risk
0.525
0.500
0.475
0.450
0.425
0.400
0.375
0.350
0 1 2 3 4 5 6 7 8 9 10 11 12 13 14
Number of steps