# Algoritmos de Optimización

No siempre se puede encontrar soluciones cerradas (casi nunca).

Se consideran entonces distintos algoritmos. Una clase grande corresponde a métodos iterativos.

Ejemplos:
- "Biseccion" (Golden section search)
- Descenso (gradiente, Newton)
- Proximal / Punto fijo
- Punto interior
- Stochastic GD

## Métodos de descenso / de gradiente

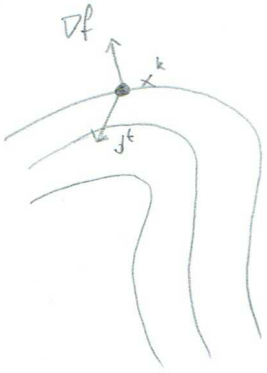
Consideremos un problema de optimización sin restricciones, $\min_{x \in \mathbb{R}^n} f(x)$

Empezamos de un punto $x^0$, y pensamos sucesivamente $x^1, x^2, x^3, \dots$

- Si $f(x^{k+1}) < f(x^k)$ $\forall k$ decimos que es un método de descenso.

Tenemos $x^{k+1} = x^k + \alpha^k d^k$, donde $d^k \in \mathbb{R}^n$ es una dirección y $\alpha^k$ el tamaño del paso, positivo.

Un ejemplo puede ser $d^k = -\nabla f(x^k)$, o más en general, $d^k$ / $\langle \nabla f(x^k), d^k \rangle < 0$. A estos se los llama métodos de gradiente.




Esta condición garantiza que $f$ decrece localmente en esa dirección. Es decir, $\exists \delta > 0$ /

$$f(x^k + \alpha^k d^k) < f(x^k) \quad \text{si } \alpha^k \in (0, \delta)$$

El método es entonces:

Dado un punto inicial $x^0$

Repetir

- Elegir dirección $d^k$
- Elegir peso $\alpha^k$
- Hacer $x^{k+1} = x^k + \alpha^k d^k$

Hasta condición de parada

¿Cómo elegir dirección $d^k$?

¿Cómo elegir el peso $\alpha^k$?

¿Qué condición de parada usar?

## Elección de dirección

Escribimos genéricamente $x^{k+1} = x^k - \alpha^k D^k \nabla f(x^k)$ con $D^k \in S_{++}^n$

La condición $\nabla f(x^k)^T d^k < 0$ se traduce en $\nabla f(x^k)^T D^k \nabla f(x^k) > 0$

que se cumple pues $D^k$ es definida positiva.

## $D^k = I$

Corresponde a la dirección de máximo descenso, pero puede tener convergencia lenta, dependiendo de la geometría de las curvas de nivel.

## Newton

$$D^k = \left(\nabla^2 f(x^k)\right)^{-1}$$

Sale de considerar la aproximación de 2° orden: $f^k(x) \approx f(x^k) + \nabla f(x^k)(x - x^k) + \frac{1}{2}(x - x^k)\nabla^2 f(x^k)(x - x^k)$

↳ $x^{k+1} = x^k - \left(\nabla^2 f(x^k)\right)^{-1} \nabla f(x^k)$ (corresponde a $\alpha^k = 1$)

$$x^{k+1} = x^k - \alpha^k \left(\nabla^2 f(x^k)\right)^{-1} \nabla f(x^k)$$

## Diagonal Scaling

$$D^k = \begin{pmatrix} d_1^k & & 0 \\ & d_2^k & \\ 0 & & \ddots \\ & & & d_n^k \end{pmatrix}$$

Cuando $d_i^k \cong \left( \frac{\partial^2 f(x^k)}{\partial x_i^2} \right)^{-1}$

es una aproximación a Newton.

## Elección del paso $\alpha^k$

### Line search

Encontrar el $\alpha^k$ que minimiza $f$:

$$\alpha^k \;/\; f(x^k + \alpha^k d^k) = \min_{\alpha \geq 0} f(x^k + \alpha d^k)$$

### Limited Line search

Se fija un $s > 0$ y se busca el mejor $\alpha^k \in [0, s]$

$$\alpha^k \;/\; f(x^k + \alpha^k d^k) = \min_{\alpha \in [0,s]} f(x^k + \alpha d^k)$$

### Regla de Armijo

Encontrar el mejor $\alpha^k$ con line search suele ser costoso. Una alternativa es probar con un paso inicial $s$, e ir reduciéndolo por un factor hasta encontrar un punto donde $f$ disminuye. Esto puede tener problemas de convergencia, pero con una modificación funciona bien.

Fijamos dos parámetros, $0 < \sigma < 1$ y $0 < \beta < 1$.

Empezamos con $s (= 1)$ y vamos reduciendo por un factor $\beta$ ($\beta^m s$) hasta que

$$f(x^k) - f(x^k + \beta^m s d^k) \geq -\sigma \beta^m s \nabla f(x^k)^T d^k$$

Otra forma de verlo:

$t = 1$

Mientras

$f(x^k + t d^k) > f(x^k) + \sigma t \nabla f(x^k)^T d^k$

$t = \beta t$

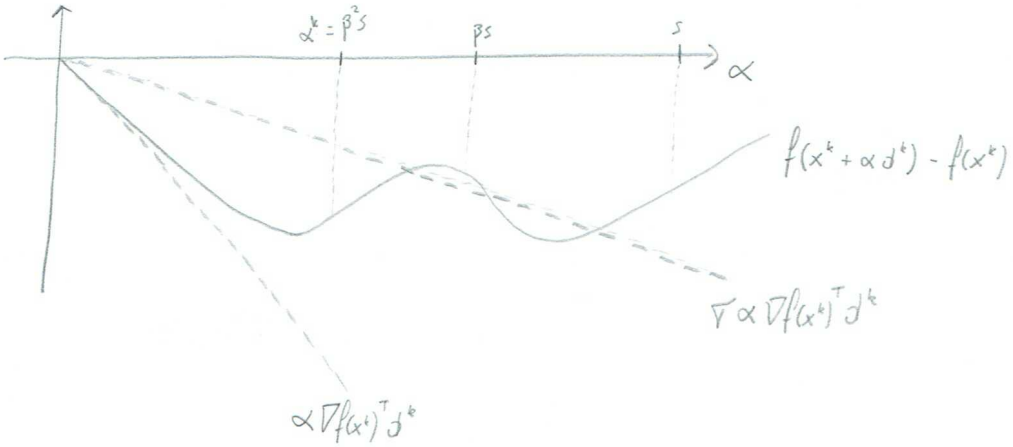
## Interpretación (Armijo)

$$\underbrace{f(x^k) - f(x^k + \beta^m s d^k)}_{\text{ganancia al moverse a } x^k+\beta^m s d^k} \geq \underbrace{-\sigma \beta^m s \nabla f(x^k)^T d^k}_{\text{ganancia si fuera lineal}}$$

nos conformamos con una fracción de la ganancia si fuera lineal

Usualmente: $\sigma \approx 0.1$ $(\sigma \in [10^{-5}, 10^{-1}])$

$\beta \in \left[\frac{1}{10}, \frac{1}{2}\right]$




## Paso constante

Se selecciona un paso $\alpha^k = s$ constante $\forall k$.

Si es muy grande, no se puede asegurar convergencia.
Si es muy chico, puede llevar a convergencia muy lenta.

(No se puede asegurar que sea de descenso en todo momento)

## Paso decreciente

Hacemos $\alpha^k \to 0$, pero que $\sum \alpha^k = \infty$

(tampoco se puede garantizar descenso en cada iteración)

Si $x^k \to \bar{x}$, pero no tan grandes

$x^n \approx x^m \approx \bar{x}$

$x^m \approx x^n - \left(\sum_{k=n}^{m-1} \alpha^k\right) \nabla f(\bar{x})$

$\Rightarrow \bar{x}$ estacionario $(\nabla f(\bar{x}) = 0)$

## Condición de parada

Como buscamos puntos críticos, una condición típica es

$$\|\nabla f(x^k)\| \leq \varepsilon$$

O similar. Por ejemplo $\dfrac{\|\nabla f(x^k)\|}{\|\nabla f(x^0)\|} \leq \varepsilon$ o $\|d^k\| \leq \varepsilon$

## Análisis de convergencia

Decimos que las direcciones $d^k$ son gradient related si para toda subsucesión $\{x^k\}_{k\in K}$ que converge a un punto no estacionario, la subsucesión correspondiente $\{d^k\}_{k\in K}$ es acotada y $\limsup_{\substack{k\to+\infty \\ k\in K}} \nabla f(x^k)^T d^k < 0$

En particular, si $d^k = -D^k \nabla f(x^k)$, alcanza pedir que los valores propios de $D^k$ estén alejados de 0: $c_1\|z\|^2 \leq z^T D^k z \leq c_2\|z\|^2 \quad \forall z \in \mathbb{R}^n$

## Prop

Sea $x^k$ generada por un método de gradiente $x^{k+1} = x^k + \alpha^k d^k$, con $d^k$ gradient related y $\alpha^k$ según la regla de Armijo.

Entonces todo punto límite de $x^k$ es un punto estacionario

$\exists$ subsucesión de $x^k$ convergente a él

## Dem

Sea $\{x^k\}_{k\in K}$ subsucesión convergente a $\bar{x}$, y supongamos que $\nabla f(\bar{x}) \neq 0$.

$\{f(x^k)\}$ es monótona decreciente, y $\{f(x^k)\}_{k\in K} \to f(\bar{x})$ porque $f$ es continua, por lo tanto $\{f(x^k)\} \to f(\bar{x})$ (toda la sucesión) $\Rightarrow f(x^k) - f(x^{k+1}) \to 0$

De la regla de Armijo: $f(x^k) - f(x^{k+1}) \geq -\sigma \alpha^k \nabla f(x^k)^T d^k \Rightarrow \alpha^k \nabla f(x^k)^T d^k \to 0$

Como $d^k$ es gradient related, y supusimos $\bar{x}$ no estacionario:

$$\limsup_{\substack{k\to+\infty \\ k\in K}} \nabla f(x^k)^T d^k < 0 \quad \Rightarrow \quad \alpha^k \to 0$$

Como en Armijo en cada iteración se empieza con paso $s$, si $\alpha^k \to 0$ entonces el paso inicial es reducido al menos una vez a partir de un $\bar{k}$:

$$f(x^k) - f\left(x^k + \frac{\alpha^k}{\beta} d^k\right) < -\sigma \frac{\alpha^k}{\beta} \nabla f(x^k)^T d^k \qquad \forall k \geq \bar{k}$$

Le llamamos $\bar{\alpha}^k = \frac{\alpha^k}{\beta}$, que es el paso anterior al aceptado.

Reescribimos:

$$\frac{f(x^k) - f(x^k + \bar{\alpha}^k d^k)}{\bar{\alpha}^k} < -\sigma \nabla f(x^k)^T d^k$$

Por Teorema de Valor Medio, $\exists\, \tilde{\alpha}^k \in [0, \bar{\alpha}^k]$ /

$$\frac{f(x^k) - f(x^k + \bar{\alpha}^k d^k)}{\bar{\alpha}^k} = -\nabla f(x^k + \tilde{\alpha}^k d^k)^T d^k$$

Entonces: $-\nabla f(x^k + \tilde{\alpha}^k d^k)^T d^k < -\sigma \nabla f(x^k)^T d^k$

Como $d^k$ acotado (por ser gradient related) $\Rightarrow \exists\, d^{k_\ell} \to \bar{d}$

Tomando límite según esta subsucesión:

$$-\nabla f(\bar{x})^T \bar{d} < -\sigma \nabla f(\bar{x})^T \bar{d}$$

$$\Rightarrow \quad 0 \leq (1-\sigma) \nabla f(\bar{x})^T \bar{d}$$

$$\text{como } \sigma < 1 \Rightarrow \nabla f(\bar{x})^T \bar{d} \geq 0$$

Lo cual es absurdo, pues contradice que $d^k$ sea gradient related.