Hay resultados similares para paso constante y paso $\alpha^k \to 0$ $\sum \alpha^k = \infty$.
El siguiente resultado explica por qué los métodos de gradiente tienden a converger a un único punto (que es estacionario).

Prop

Sea $f \in C^1$ y x^k una sucesión generada por un método de gradiente / $f(x^{k+1}) \leq f(x^k)$ y todo punto límite es estacionario. Además, $\exists\, s > 0, c > 0$ tq $\alpha^k < s$, $\|d^k\| \leq c\|\nabla f(x^k)\|$ $\forall k$.

Si x^* es un mínimo local aislado de f, entonces existe un abierto S tal que si $x^{k_0} \in S \Rightarrow x^k \in S$ $\forall k \geq k_0$ y $x^k \to x^*$

(Las condiciones $f(x^{k+1}) \leq f(x^k)$ y $\alpha^k \leq s$ se satisfacen usando Armijo, y $\|d^k\| \leq c\|\nabla f(x^k)\|$ si D^k tiene valores propios acotados)

Tasa de convergencia

Vamos a hacer un análisis local, para sucesiones $x^k \to x^*$.
Usaremos una función de error $e(x) \geq 0$ / $e(x^*) = 0$.
Ejemplos típicos: $e(x) = \|x - x^*\|$ o $e(x) = |f(x) - f(x^*)|$

Decimos que un método converge linealmente si

$$\limsup_{k \to +\infty} \frac{e(x^{k+1})}{e(x^k)} \leq \beta < 1$$

(esto implica que $e(x^k)$ decrece más rápido que β^k)

Cuando $\limsup_{k \to +\infty} \frac{e(x^{k+1})}{e(x^k)} = 0$ decimos que la convergencia es superlineal

Un caso particular es la convergencia cuadrática, que es cuando

$$\limsup_{k \to +\infty} \frac{e(x^{k+1})}{e^2(x^k)} \leq \text{const} < \infty$$

Análisis del modelo cuadrático

Supongamos que $f(x) = \frac{1}{2} x^T Q x$ con $Q > 0$

Entonces el mínimo se da en $x^* = 0$, y es $f(x^*) = 0$.

Tenemos: $\nabla f(x) = Qx$ $\quad \nabla^2 f(x) = Q$

El método práctico para el máximo descenso queda:

$$x^{k+1} = x^k - \alpha^k \nabla f(x^k) = (I - \alpha^k Q) x^k$$

$$\Rightarrow \|x^{k+1}\|^2 = x^{k^T} (I - \alpha^k Q)^2 x^k$$

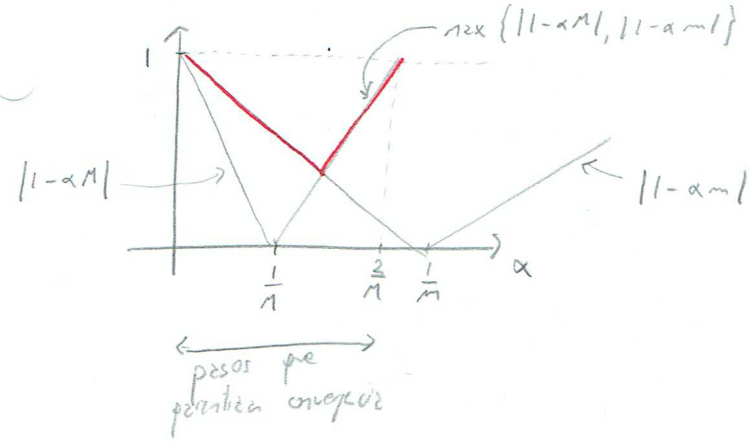
Como $z^T A z \leq \lambda_{max}(A) \|z\|^2$

$$\Rightarrow \|x^{k+1}\|^2 \leq \lambda_{max}((I - \alpha^k Q)^2) \|x^k\|^2$$

Se puede ver que los valores propios de $(I - \alpha^k Q)^2$ son $(1 - \alpha^k \lambda_i)^2$, donde λ_i son los valores propios de Q.

Si m y M son los valores propios más chico y más grande de Q, entonces

$$\frac{\|x^{k+1}\|}{\|x^k\|} \leq \max\{|1 - \alpha^k m|, |1 - \alpha^k M|\}$$



El mínimo se alcanza en $\alpha = \frac{2}{m+M}$

y vale $\frac{M-m}{M+m}$

Para el α óptimo: $\frac{e(x^{k+1})}{e(x^k)} \leq \frac{M-m}{M+m} < 1 \Rightarrow$ Conv. lineal

Al cociente $\frac{M}{m}$ se le llama número de condición de Q. $\quad cond(Q) = \frac{M}{m} \geq 1$

Si $cond(Q) \approx 1 \Rightarrow$ Convergencia rápida, si $cond(Q) \gg 1 \Rightarrow$ Convergencia lenta

Optimización con restricciones

Estudiemos $\min_{x \in X} f(x)$ $\quad$ f diferenciable

X convexo.

Un punto $x \in \mathbb{R}^n$ que cumple las restricciones decimos que es feasible (factible)

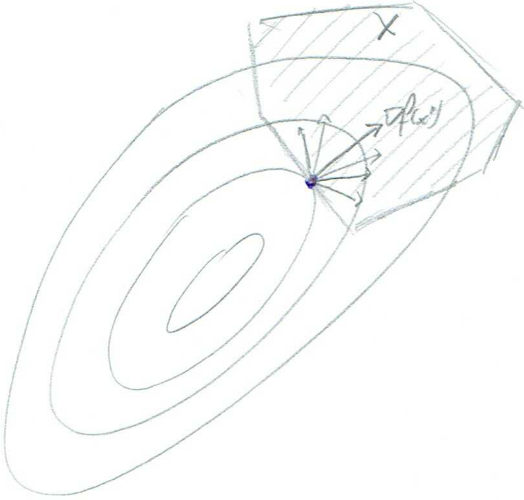
Condiciones de Optimalidad

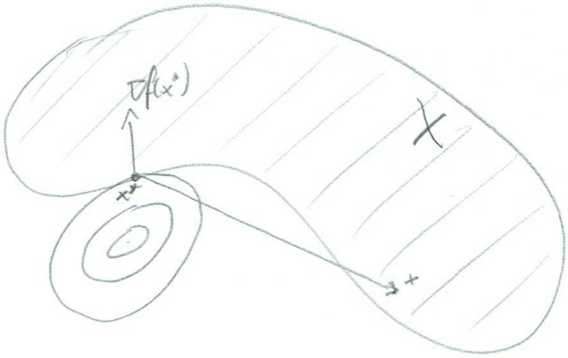
Prop

(a) Si x^* es un mínimo local de f en X, entonces $\nabla f(x^*)^T (x - x^*) \geq 0 \quad \forall x \in X$

(b) Si además f es convexa, entonces la condición (a) es suficiente.

Como X es convexo, las direcciones factibles (desde x^*) son de la forma $(x - x^*)$ con $x \in X$.





La condición (a) no es cierta.
x^* es mínimo, pero X no es convexo.

Ejemplo: Proyección a un conjunto convexo.

Dado X convexo, la proyección de un punto $z \in \mathbb{R}^n$ a X es el que minimiza:

$$\min_{x \in X} f(x) = \|z - x\|^2$$

Llamemos $P_X(z)$ a esta proyección.

El gradiente de f es $\nabla f(x) = -2(z-x)$. Evaluando en el óptimo $P_X(z)$:

$\nabla f(P_X(z)) = -2(z - P_X(z))$. Entonces la condición de optimalidad queda:

$$-2(z - P_X(z))\cdot(x - P_X(z)) \geq 0 \quad \forall x \in X$$

(es necesaria y suficiente, ya que $\|z-x\|^2$ es convexa)

Obs:

Si X es un subespacio (que es convexo), la condición se traduce en $(z - P_X(z)) \in X^{\perp}$

Sean ahora $x, y \in \mathbb{R}^n$, y consideremos sus proyecciones a X: $P_X(x)$, $P_X(y)$

Sabemos que $(w - P_X(x))^T (x - P_X(x)) \leq 0 \quad \forall w \in X$. En particular para $w = P_X(y)$

$(P_X(y) - P_X(x))^T (x - P_X(x)) \leq 0$ De manera similar:

$(P_X(x) - P_X(y))^T (y - P_X(y)) \leq 0$ Sumando:

$$(P_X(y) - P_X(x))^T (x - P_X(x) - y + P_X(y)) \leq 0$$

Cauchy-Schwartz

$$\Rightarrow \|P_X(y) - P_X(x)\|^2 \leq (P_X(y) - P_X(x))^T (y - x) \leq \|P_X(y) - P_X(x)\| \cdot \|y - x\|$$

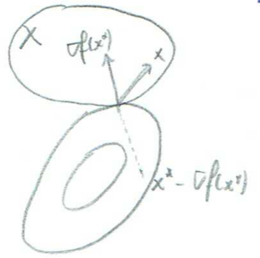
$$\Rightarrow \|P_X(y) - P_X(x)\| \leq \|y - x\|$$

La proyección es no expansiva

Volviendo a $\min_{x \in X} f(x)$ con X convexo, otra manera de escribir la condición de optimalidad $\langle \nabla f(x^*), x - x^* \rangle \geq 0 \quad \forall x \in X$ es $x^* = P_X(x^* - \nabla f(x^*))$

En efecto, $\nabla f(x^*)^T (x - x^*) \geq 0 \iff ((x^* - \nabla f(x^*)) - x^*)^T (x - x^*) \leq 0 \quad \forall x \in X$

Esto último sucede si y sólo si x^* es la proyección de $x^* - \nabla f(x^*)$ en X.



La condición $x^* = P_X(x^* - \nabla f(x^*))$ se denomina condición de estacionariedad

Algoritmos para optimización con restricciones

$\min f(x)$ f diferenciable
s.t. $x \in X$ X convexo, cerrado

De nuevo, la estrategia será
$$x^{k+1} = x^k + \alpha^k d^k$$
pero ahora además de ser de descenso, d^k tiene que ser factible.

Recordemos que al ser X convexo, las direcciones factibles d^k son de la forma $d^k = x - x^k$ con $x \in X$

Tenemos un resultado de convergencia general, similar al de opt. sin restricciones. Específicamente, si d^k es gradient related y α^k según Armijo, todo punto límite de $\{x^k\}$ es estacionario.

Veamos algunos algoritmos en particular.

Método de Frank-Wolfe (o Método de gradiente condicional)

Hay que elegir una dirección $d^k = \bar{x}^k - x^k$, y que sea de descenso: $\nabla f(x^k)^T(\bar{x}^k - x^k) < 0$

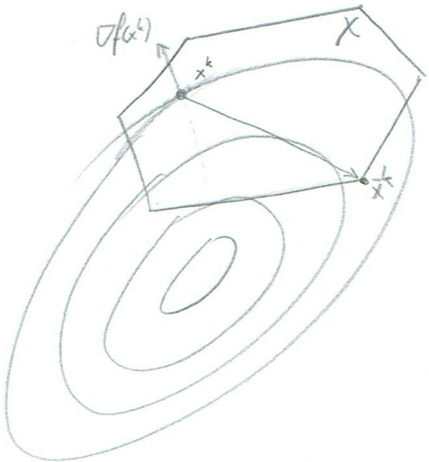
Si no existe $\bar{x}^k \in X$ / $\nabla f(x^k)^T(\bar{x}^k - x^k) < 0$, quiere decir que $\nabla f(x^k)^T(\bar{x}^k - x^k) \geq 0$ $\forall \bar{x}^k \in X$
que es la cond. de optimalidad.

En Frank-Wolfe elegimos la dirección para minimizar $\nabla f(x^k)^T(\bar{x}^k - x^k)$

Es decir: $\bar{x}^k = \underset{x \in X}{\operatorname{argmin}} \nabla f(x^k)^T (x - x^k)$

Obs: elegir $\bar{x}^k$ (y por lo tanto la dirección d^k) es un problema lineal con restricción a X.

El paso α^k lo buscamos con Armijo.



Tenemos garantía de convergencia.

Cuando X es un poliedro, usualmente los puntos $\bar{x}^k$ son vértices, y eso puede enlentecer la convergencia.

Cuando X es más complejo o cuando hay muchas restricciones, el método funciona mejor.

Método de gradiente proyectado

Es también de la forma $x^{k+1} = x^k + \alpha^k(\bar{x}^k - x^k)$

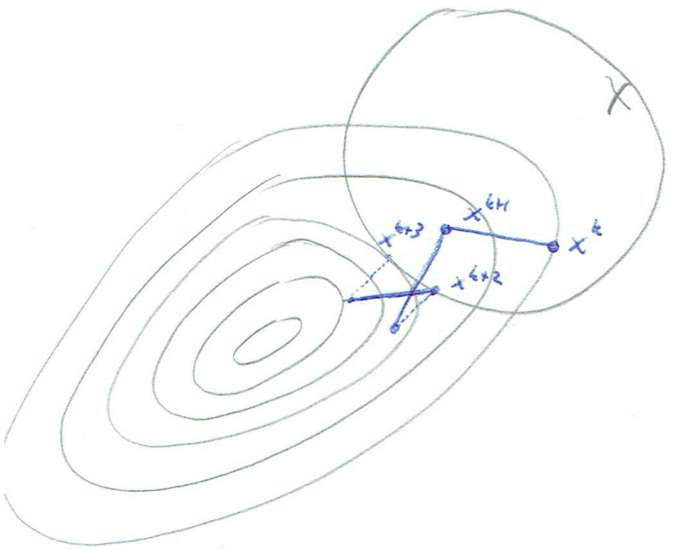
con $\bar{x}^k = P_X(x^k - s^k \nabla f(x^k))$ y $\alpha^k \in (0,1]$

Es decir, tomamos la dirección del gradiente, y si nos lleva fuera de X lo proyectamos, y usamos ese punto para la dirección.

Hay que elegir s^k y α^k, los dos se puede ver como pasos.

En particular si tomamos $\alpha^k = 1 \; \forall k$, el método queda:

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



El algoritmo se estanca si $\bar{x}^k = x^k$.
En este caso, tendríamos:

$$x^k = P_X(x^k - s^k \nabla f(x^k))$$

que es la condición de estacionaridad

De la definición de proyección, $\bar{x}^k$ es: $\bar{x}^k = \underset{x \in X}{\arg\min} \; \| x - x^k + s^k \nabla f(x^k) \|^2$

$$\Rightarrow \bar{x}^k = \underset{x \in X}{\arg\min} \; \|x - x^k\|^2 + \|s^k \nabla f(x^k)\|^2 + 2 s^k \nabla f(x^k)^T (x - x^k)$$

$$= \underset{x \in X}{\arg\min} \; \underbrace{\|x - x^k\|^2}_{\text{término cuadrático extra}} + \underbrace{2 s^k \nabla f(x^k)^T (x - x^k)}_{\text{igual que en Frank-Wolfe}}$$

La diferencia del Método de gradiente condicional con el de gradiente proyectado es el término cuadrático.