MÉTODOS DE PRIMER ORDEN?

ANÁLISIS DE CONVERGENCIA??

Universidad de Buenos Aires

Facultad de Ciencias Exactas y Naturales
Departamento de Matemática

Tesis de Licenciatura
Director de Tesis: Dr. Pablo Amster
Septiembre 2018 – version 0.1

ABSTRACT

Aca va a ir el abstract cuando lo tengamos

We have seen that computer programming is an art,
because it applies accumulated knowledge to the world,
because it requires skill and ingenuity, and especially
because it produces objects of beauty.

— knuth:1974 [knuth:1974]

AGRADECIMIENTOS

Agradecimientos para todos

CONTENTS

LIST OF FIGURES

LIST OF TABLES

LISTINGS

ACRÓNIMOS

Part I

Introducción

1

INTRODUCCIÓN

2

INTUICIÓN

Usemos un caso modelo para ejemplificar porque no es probable que los metodos de primer orden (entre ellos *gradient descent*) convergan a puntos silla. Sea $f : \mathbb{R}^n \to \mathbb{R}^n$ dada por $f(x) = \frac{1}{2}x^T Hx$ con $H = \mathbf{diag}\,(\lambda_1, \ldots, \lambda_n)$; supongamos además que $\lambda_1, \ldots, \lambda_k > 0$ y $\lambda_{k+1}, \ldots, \lambda_n < 0$.

Si usamos en la base canónica de $\mathbb{R}^n$ $\left\{e^1, \ldots, e^n\right\}$ entonces:

$$f(x) = f(x^1, \ldots, x^n) = \frac{1}{2}\left(\lambda_1 {x_1}^2 + \cdots + \lambda_n {x_n}^2\right)$$

Por lo tanto:

$$\nabla f(x) \;=\; \lambda_i x_i e^i = 0 \iff x = x_1 e^1 = 0$$

Y tenemos que en el único punto crítico el Hessiano de f es $\nabla^2 f(0) = H$.

Recordemos que si $g(x) = x - \alpha \nabla f(x)$ entonces *gradient descent* está dado por la iteración $x_{t+1} = \;\; g(x_t) \;\; := g^t(x_0)$ con $t \in \mathbb{N}$ y $x_0 \in \mathbb{R}^n$, y en este caso esta representado por:

$$\begin{aligned} x_{t+1} =& g(x_t) \\ =& x_t - \alpha \nabla f(x_t) \\ =& (1 - \alpha\lambda_i)\, x_{i\,t} e^i \\ =& (1 - \alpha\lambda_i) \left\langle x_t, e^i \right\rangle e^i \end{aligned}$$

Por lo tanto por inducción es fácil probar que:

$$x_{t+1} = (1 - \alpha\lambda_i)^t \left\langle x_o, e^i \right\rangle e^i$$

Sea $L = \max_i |\lambda_i|$ y supongamos que $\alpha < \frac{1}{L}$, luego:

$$\begin{aligned} 1 - \alpha\lambda_i < 1 \quad \text{Si } i \leq k \\ 1 - \alpha\lambda_i > 1 \quad \text{Si } i > k \end{aligned}$$

Con lo que concluímos que:

$$\lim_t x_t = \begin{cases} 0 & \text{Si } x \in E_s := \left\langle e^1, \ldots, e^k \right\rangle \\ \infty & \text{Si no} \end{cases}$$

Finalmente, si $k < n$ entonces concluímos que:

$$P_{\mathbb{R}^n}(\left\{ x \in \mathbb{R}^n \;/\; \lim_t g^t(x) = 0 \right\}) = |E_s| = 0$$

Para notar este fenómeno en un ejemplo no cuadrático consideremos $f(x,y) = \frac{1}{2}x^2 + \frac{1}{4}y^4 - \frac{1}{2}y^2$, reproduciendo los calculos anteriores:

$$\begin{aligned} \nabla f = & \qquad (x, y^3 - y) \\ g = & \quad ((1-\alpha)x, (1+\alpha)y - \alpha y^3) \\ \nabla^2 f = & \qquad \begin{pmatrix} 1 & 0 \\ 0 & 3y^2 - 1 \end{pmatrix} \end{aligned} \tag{1}$$

De lo que vemos que los puntos críticos son:

$$z_1 = (0,0) \qquad z_2 = (0,1) \qquad z_3 = (0,-1)$$

Y del critério del Hessiano concluímos que z_2, z_3 son mínimos locales mientras que z_1 es un punto silla. De la intuición previa, como en z_1 el autovector asociado al autovalor positivo es e^1 podemos intuir que:

Lema 2.0.1 *Para $f(x,y) = \frac{1}{2}x^2 + \frac{1}{4}y^4 - \frac{1}{2}y^2$ resulta que $E_s = \langle t * e^1 \ / \ t \in \mathbb{R} \rangle := W_s$*

Asumiendo el resultado por un momento, dado que $\dim_{\mathbb{R}^2}(E_s) = 1 < 2$ entonces $P_{\mathbb{R}^2}(E_s) = 0$ que es lo que queríamos verificar. Demostremos el lema ahora:

Demostración Del lema Sea $x_0 \in \mathbb{R}^n$ y g la iteración de *gradient descent* dada por 2, luego:

$$(x_t, y_t) = g^t(x,y) = \begin{pmatrix} (1-\alpha)^t x_0 \\ g_y^t(y_0) \end{pmatrix} \underset{(t\to\infty)}{\longrightarrow} \begin{pmatrix} 0 \\ \lim_t g_y^t(y_0) \end{pmatrix}$$

Por lo que todo depende de y_0. Analizando $\dfrac{dg_y}{dy} = 1 + \alpha - 3\alpha y^2$ notemos que:

$$\begin{aligned} \left| \frac{dg_y}{dy} \right| < 1 \iff & \quad \left| 1 + \alpha - 3\alpha y^2 \right| < 1 \\ \iff & \quad -1 < 1 + \alpha - 3\alpha y^2 < 1 \\ \iff & \quad -2 - \alpha < -3\alpha y^2 < -\alpha \\ \iff & \quad \sqrt{\frac{2+\alpha}{3\alpha}} > |y| > \sqrt{\frac{1}{3}} \\ \iff & \quad \sqrt{\frac{1 + \frac{2}{\alpha}}{3}} > |y| > \sqrt{\frac{1}{3}} \end{aligned}$$

Por lo que por el Teorema de Punto Fijo de Banach:

$$\lim_t g_y^t(y_0) = \begin{cases} 1 & \text{Si } \sqrt{\frac{1+\frac{2}{\alpha}}{3}} > y_0 > \sqrt{\frac{1}{3}} \\ -1 & \text{Si } \sqrt{\frac{1+\frac{2}{\alpha}}{3}} < -y_0 < \sqrt{\frac{1}{3}} \end{cases}$$

Si analizamos simplemente los signos de g y $\frac{dg_y}{dy}$ en los otros intervalos podemos conluir que:

$$\lim_t g_y^t(y_0) = \begin{cases} -\infty & \text{Si } y_0 > \sqrt{\frac{1+\frac{2}{\alpha}}{3}} \\ 1 & \text{Si } \sqrt{\frac{1+\frac{2}{\alpha}}{3}} > y_0 > 0 \\ -1 & \text{Si } -\sqrt{\frac{1+\frac{2}{\alpha}}{3}} < y_0 < 0 \\ \infty & \text{Si } y_0 < -\sqrt{\frac{1+\frac{2}{\alpha}}{3}} \end{cases}$$

Dedujimos entonces que $(x,y) \in E_s \iff (x,y) = (t,0)\ t \in \mathbb{R} \iff (x,y) \in W_s$. ∎

Part II

El teorema y aplicaciones

En esta parte vamos a demostrar el resultado principal referido a la convergencia a mínimos de los diferentes algoritmos de primer orden usados en Machine Learning

3

TEOREMA DE LA VARIEDAD ESTABLE Y LOS PUNTOS FIJOS INESTABLES

3.1 RESULTADOS PREVIOS

Por el resto del documento, $g : \chi \to \chi$ y χ es una $d-$variedad sin borde.

Esto quizas deberia ir en prerequisitos cuando lo tengamos

Definición Dada una variedad de dimensión d χ y el espacio de medida $\left(\mathbb{R}^d, \mathcal{B}, \mu\right)$, decimos que $E \subset \chi$ tiene *medida cero* si existe un atlas $\mathcal{A} = \left\{U_i, \phi^i\right\}_{i\in\mathbb{N}}$ tal que $\mu\left(\phi^i\left(E \cap U_i\right)\right) = 0$. En este caso usamos el abuso de notación $\mu(E) = 0$.

Lema 3.1.1 *Sea $E \subset \chi$ tal que $\mu(E) = 0$; si* $\det(Dg(x)) \neq 0$ *para todo $x \in \chi$, luego $\mu\left(g^{-1}(E)\right) = 0$*

Demostración Sea $h = g^{-1}$ y $\left(V_i, \psi^i\right)$ una colección de cartas en el dominio de g, si verificamos que $\mu\left(h\left(E\right) \cap V_i\right) = 0$ para todo $i \in \mathbb{N}$ entonces:

$$\mu(h(E)) = \mu\left(\bigcup_{i\in\mathbb{N}} h(E) \cap V_i\right) \leqslant \sum_{i\in\mathbb{N}} \mu\left(h(E) \cap V_i\right) = 0$$

Sin pérdida de generalidad podemos asumir que $h(E) \subseteq V$ con $(V, \psi) \in \left\{\left(V_i, \phi^i\right)\right\}$ una carta determinada. Sea $\mathcal{A} := \left\{\left(U_i, \varphi^i\right)\right\}$ un atlas de χ y notemos $E_i = E \cap U_i$; luego $E = \bigcup\limits_{i\in\mathbb{N}} E_i = \bigcup\limits_{i\in\mathbb{N}} {\varphi^i}^{-1} \circ \varphi^i\left(E_i\right)$ por lo que:

$$\begin{aligned} \mu\left(\psi \circ h(E)\right) = \;& \mu\left(\psi \circ h\left(\bigcup_{i\in\mathbb{N}} {\varphi^i}^{-1} \circ \varphi^i\left(E_i\right)\right)\right) \\ \leqslant \;& \sum_{i\in\mathbb{N}} \mu\left(\psi \circ h \circ {\varphi^i}^{-1}\left(\varphi^i(E_i)\right)\right) \end{aligned}$$

Por hipótesis $\varphi^i(E_i)$ es de medida cero, luego como g es difeomorfismo local por **??** entonces $\psi \circ h \circ {\varphi^i}^{-1} \in C^1$. Como si $f \in C^1(\mathbb{R}^d)$ entonces es localmente Lipshitz, ergo f preserva la medida, concluímos que $\mu\left(\psi \circ h \circ {\varphi^i}^{-1}\left(\varphi^i(E_i)\right)\right) = 0$ para todo $i \in \mathbb{N}$. ∎

Uso Teorema de la funcion inversa en variedades y que localmente Lipshitz preserva medida

3.2 PUNTOS FIJOS INESTABLES

Definición Sea:

$$\mathcal{A}_g^* := \left\{x \;:\; g(x) = x \quad \max_i \left|\lambda_i\left(Dg(x)\right) > 1\right|\right\}$$

El conjunto de puntos fijos de g cuyo diferencial en ese punto tiene algún autovalor mayor que 1. A este conjunto lo llamaremos el conjunto de *puntos fijos inestables*

Este teorema debería ir en prerequisitos

Teorema 3.2.1 *Sea x^* un punto fijo de $g \in C^r(\chi)$ un difeomorfismo local. Supongamos que $E = E_s \oplus E_u$ donde*

$$\begin{aligned} E_s = & \quad \langle \{v_i \ / \ Dg(x^*)v_i = \lambda_i v_i \quad , \quad \lambda_i \leqslant 1\}\rangle \\ E_u = & \quad \langle \{v_i \ / \ Dg(x^*)v_i = \lambda_i v_i \quad , \quad \lambda_i > 1\}\rangle \end{aligned}$$

Entonces existe $W^{cs}_{loc} \hookrightarrow \chi$ un embedding C^r local tangente a E_s en x^* llamado la variedad local estable central que cumple que existe $B \ni x^*$ entorno tal que $g\left(W^{cs}_{loc}\right) \cap B \subseteq W^{cs}_{loc}$ y $\bigcap\limits_{k\in\mathbb{N}} g^{-k}(B) \subseteq W^{cs}_{loc}$

Con todos estos resultados demostremos el teorema principal:

Teorema 3.2.2 *Sea $g \in C^1(\chi)$ tal que $\det(Dg(x)) \neq 0$ para todo $x \in \chi$, luego el conjunto de puntos iniciales que convergen por g a un punto fijo inestable tiene medida cero, i. e.:*

$$\mu\left(\left\{x_0 : \ \lim_k g^k(x_0) \in \mathcal{A}^*_g\right\}\right) = 0$$

Demostración Para cada $x^* \in \mathcal{A}^*_g$ por 3.2.2 existe B_{x^*} un entorno abierto; es más, $\bigcup\limits_{x^*\in\mathcal{A}^*_g} B_{x^*}$ forma un cubrimiento abierto del cual existe un subcubrimiento numerable pues X es variedad, i. e.

$$\bigcup_{x^*\in\mathcal{A}^*_g} B_{x^*} = \bigcup_{i\in\mathbb{N}} B_{x^*_i}$$

Usamos que en una variedad se cumple la propiedad de Lindeloff

Primero si $x_0 \in \chi$ sea:

$$\begin{aligned} x_k =& g^k(x_0) \\ =& \underbrace{g \circ \cdots \circ g}_{k \text{ veces}}(x_0) \end{aligned}$$

la sucesión del flujo de g evaluado en x_0, entonces si $W := \left\{x_0 : \ \lim_k x_k \in \mathcal{A}^*_g\right\}$ queremos ver que $\mu(W) = 0$.

Sea $x_0 \in W$, luego como $x_k \to x^* \in \mathcal{A}^*_g$ entonces existe $T \in \mathbb{N}$ tal que para todo $t \geqslant T$, $x_t \in \bigcup\limits_{i\in\mathbb{N}} B_{x^*_i}$ por lo que $x_t \in B_{x^*_i}$ para algún $x^*_i \in \mathcal{A}^*_g$ y $t \geqslant T$. Afirmo que:

Pablo: Hace falta demostrar esto??

Lema 3.2.3 $x_t \in \bigcap\limits_{k\in\mathbb{N}} g^{-k}(B_{x^*_i})$ *para todo $t \geqslant T$*

Si notamos $S_i \triangleq \bigcap_{k\in\mathbb{N}} g^{-k}(B_{x_i^*})$, entonces por 3.2.1 sabemos por un lado que es una subvariedad de W^{cs}_{loc} y por el otro que $\dim(S_i) \leqslant \dim(W^{cs}_{loc}) = \dim(E_s) < d-1$ [1]; por lo que $\mu(S_i) = 0$.

Finalmente como $x_T \in S_i$ para algún T entonces $x_0 \in \bigcup_{k\in\mathbb{N}} g^{-k}(S_i)$ por lo que $W \subseteq \bigcup_{i\in\mathbb{N}} \bigcup_{k\in\mathbb{N}} g^{-k}(S_i)$. Concluímos:

Usamos que la dimension de la variedad es la de su tangente

Usamos que una subvariedad de dimension menor tiene medida 0

$$\begin{aligned}
\mu\left(W\right) \leqslant & \quad \mu\left(\bigcup_{i\in\mathbb{N}} \bigcup_{k\in\mathbb{N}} g^{-k}(S_i)\right) \\
\leqslant & \quad \sum_{i\in\mathbb{N}} \sum_{k\in\mathbb{N}} \mu\left(g^{-k}(S_i)\right) \\
\overset{3.1.1}{=} & \quad 0
\end{aligned}$$

$\blacksquare$

Para finalizar veamos un caso simple que nos encontraremos seguido:

Corolario 3.2.4 *Bajo las mismas hipótesis que en 3.2.2 si agregamos que* $\chi* \subseteq \mathcal{A}^*_g$ *entonces* $\mu(W_g) = 0$

Demostración Como $\chi^* \subseteq \mathcal{A}^*_g$ entonces $W_g \subseteq W$, luego $\mu(W_g) \leqslant \mu(W) \overset{3.2.2}{=} 0$. $\blacksquare$

1 Por que???

4

APLICACIÓNES

4.1 GRADIENT DESCENT

Como una aplicación del teorema en 3.2.2 demostremos que *gradient descent* tiene probabilidad cero de converger a puntos silla. Consideremos *gradient descent* con *learning rate* α:

$$x_{k+1} = g(x_k) \triangleq x_k - \alpha \nabla f(x_k) \tag{2}$$

Hipótesis 1 Asumamos que $f \in \mathcal{C}^2$ y $\left\|\nabla^2 f(x)\right\|_2 \leqslant L$

Proposición 4.1.1 *Todo punto silla estricto de f es un punto fijo inestable de g, i. e.$\chi^* \subseteq \mathcal{A}_g^*$.*

Demostración Es claro que un punto crítico de f es punto fijo de g; si $x^* \in \chi^*$ entonces $Dg(x^*) = Id - \alpha\nabla^2 f(x^*)$ y entonces los autovalores de Dg son $\{1 - \alpha\lambda_i \ : \ \lambda_i \in \{\mu \ : \ \nabla^2 f(x^*)v = \mu v \quad \text{para algún } v \neq 0\}\}$. Como $x^* \in \chi^*$ existe $\lambda_{j^*} < 0$ por lo que $1 - \alpha\lambda_{j^*} > 1$; concluímos que $x^* \in \mathcal{A}_g^*$. ∎

Usamos que f(A) tiene autovalores $f(\{\lambda_i\})$

Proposición 4.1.2 *Bajo 4.1 y $\alpha < \frac{1}{L}$ entonces* $\det(Dg(x)) \neq 0$.

Demostración Como ya sabemos $Dg(x) = Id - \alpha\nabla^2 f(x)$ por lo que:

$$\det(Dg(x)) = \prod_{i \in \{1,\ldots,d\}} (1 - \alpha\lambda_i)$$

Luego por 4.1 tenemos que $\alpha < \frac{1}{|\lambda_i|}$ y entonces $1 - \alpha\lambda_i > 0$ para todo $i \in \{1, \ldots, d\}$; concluímos que $\det(Dg(x)) > 0$. ∎

Corolario 4.1.3 *Gradient descent converge a mínimos Sea g dada por Gradient descent en 2, bajo 4.1 y $\alpha < \frac{1}{L}$ se tiene que $\mu(W_g) = 0$.*

Demostración Por 4.1.1 y 4.1.2 tenemos que vale 3.2.4 y concluímos que $\mu(W_g) = 0$. ∎

4.2 PUNTO PRÓXIMO

El algoritmo de punto próximo esta dado por la iteración:

$$x_{k+1} = g(x_k) \triangleq \arg\min_{z \in \chi} f(z) + \frac{1}{2\alpha}\|x_k - z\|_2^2 \tag{3}$$

Proposición 4.2.1 *Bajo 4.1 y $\alpha < \frac{1}{L}$ entonces vale:*

1. $\det(Dg(x)) \neq 0$
2. $\chi^* \subseteq \mathcal{A}_g^*$

Probamos esto? Me parece un poco claro

Demostración Veamos primero el siguiente lema:

Lema 4.2.2 *Bajo 4.1, $\alpha < \frac{1}{L}$ y $x \in \chi$ entonces $f(z) + \frac{1}{2\alpha}\|x-z\|_2^2$ es estrictamente convexa, por lo que $g \in \mathcal{C}^1(\chi)$*

Por lo tanto por 4.2.2 podemos tomar límite, i. e.

$$
\begin{array}{ccccc}
x_{k+1} = & g(x_k) = & \arg\min_{z\in\chi} f(z) + \frac{1}{2\alpha}\|x_k - z\|_2^2 \\
\downarrow & \downarrow & \downarrow \\
x = & g(x) = & \arg\min_{z\in\chi} f(z) + \frac{1}{2\alpha}\|x - z\|_2^2
\end{array}
$$

$$
\begin{aligned}
&\Longleftrightarrow \nabla_z\left(f(z) + \frac{1}{2\alpha}\|x-z\|^2\right)(g(x)) = 0\\
&\Longleftrightarrow \nabla f(g(x)) - \frac{1}{\alpha}(x - g(x)) = 0\\
&\Longleftrightarrow g(x) + \alpha\nabla f(g(x)) = x
\end{aligned}
$$

Finalmente por diferenciación implicita obtenemos:

$$
\begin{aligned}
& Dg(x) + \alpha\nabla^2 f(g(x))Dg(x) = Id\\
\Longrightarrow \quad & Dg(x) = \left(Id + \alpha\nabla^2 f(g(x))\right)^{-1}
\end{aligned}
$$

Luego si $x^* \in \chi^*$ entonces $Dg(x^*) = \left(Id + \alpha\nabla^2 f(x^*)\right)^{-1}$ y tiene autovalores $\left\{\frac{1}{1+\alpha\lambda_i}\right\}$ con λ_i autovalores de $\nabla^2 f(x^*)$. Por lo tanto $x^* \in \mathcal{A}_g^*$ y para $\alpha < \frac{1}{L}$ se tiene que $\det(Dg(x)) \neq 0$. ∎

Corolario 4.2.3 *Sea g dado por el algoritmo de punto próximo con ecuación 3, bajo 4.1 y $\alpha < \frac{1}{L}$ se tiene que $\mu(W_g) = 0$.*

Demostración Por 4.2.1 tenemos que vale 3.2.4 y concluímos que $\mu(W_g) = 0$. ∎

4.3 DESCENSO POR COORDENADAS

Sea $S_1, \ldots, S_b$ una partición disjunta de $\{1, \ldots, d\}$ donde d y b son parámetros del método.

Consideremos el algoritmo 1:

Algorithmus 1 : Descenso por coordenadas

Input: $f \in C^1$, $\alpha > 0$, $x_0 \in \chi$
for $k \in \mathbb{N}$ **do**
  for *block* $i = 1, \ldots, b$ **do**
    for *index* $j \in S_i$ **do**
      $y_k^{S_0} = x_k$ e $y_k^{S_i} = \left(x_{k+1}^{S_1}, \ldots, x_{k+1}^{S_i}, x_k^{S_{i+1}}, \ldots, x_k^{S_b}\right)$
      $x_{k+1}^j \leftarrow x_k^j - \alpha \dfrac{\partial f}{\partial x_j}\left(y_k^{S_{i-1}}\right)$
    end
  end
end

Luego si definimos $g_i(x) = x - \alpha \sum_{j \in S_i} e_j^T \nabla f(x)$ entonces:

Lema 4.3.1 *La iteración de Descenso por coordenadas esta dada por:*

$$x_{k+1} = g(x_k) \stackrel{\triangle}{=} g_d \circ g_{d-1} \circ \cdots \circ g_1(x) \tag{4}$$

Lema 4.3.2 *Si g está dada por 4 entonces si notamos $P_S = \sum_{i \in S} e_i e_i^T$ entonces:*

$$Dg(x_k) = \prod_{i \in \{1,\ldots,b\}} \left(Id - \alpha P_{b-i+1} \nabla^2 f(y_k^{S_{b-i}})\right) \tag{5}$$

Demostración Notemos primero que:

$$Dg_i(x) = Id - \alpha P_{S_i} \nabla^2 f(x)$$

Por lo tanto:

$$\begin{aligned}
Dg(x_k) = & \quad D\left(g_b \circ \ldots g_1\right)(x_k) \\
= & \quad \left(Id - \alpha P_{S_b} \nabla^2 f\right)\left(\underbrace{g_{b-1} \circ \ldots g_1(x_k)}_{y_k^{S_{b-1}}}\right) D\left(g_{b-1} \circ \ldots g_1\right)(x_k) \\
\vdots & \\
= & \quad \prod_{i \in \{1,\ldots,b\}} \left(Id - \alpha P_{b-i+1} \nabla^2 f(y_k^{S_{b-i}})\right)
\end{aligned}$$

∎

Observación Sea $f \in C^2$ y notemos $\nabla^2 f|_S$ a la submatriz que resulta de quedarme con filas y columnas indexadas por S. Sea $\max_{i \in \{1,\ldots,b\}} \left\|\nabla^2 f(x)|_{S_i}\right\| = L_b$

Proposición 4.3.3 *Bajo 9 y $\alpha < \frac{1}{L_b}$ se tiene que* $\det(Dg(x)) \neq 0$

Demostración Basta probar que cada término de 5 es invertible, para eso:

$$\begin{aligned}\chi_{Dg_i(x)}(\lambda) &= \det\left(\lambda Id_d - Id_d - \alpha P_{S_{b-i+1}} \nabla^2 f(x)\right)\\ &= (\lambda-1)^{n-|S_i|}\prod_{j\in S_i}\left(\lambda - 1 + \alpha\frac{\partial^2 f}{\partial x_j^2}(x)\right)\end{aligned}$$

Luego si $\alpha < \frac{1}{L_{max}}$ entonces $\lambda - 1 + \alpha\frac{\partial^2 f}{\partial x_j^2}(x) > 0$ para todo $j \in S_i$, $i \in \{1,\ldots,b\}$ por lo que todos los autovalores son positivos y $Dg_i(x)$ es invertible para todo i. ∎

Proposición 4.3.4 *Bajo 9 y* $\alpha < \frac{1}{L_{max}}$ *se tiene que* $\chi^* \subseteq \mathcal{A}_g^*$

Demostración Sea $x^* \in \chi^*$, $H = \nabla^2 f(x^*)$, $J = Dg(x^*) = \prod_{i\leq b}\left(Id_n - \alpha P_{S_{b-i+1}} H\right)$ e y_0 el autovector correspondiente al menor autovalor de H. Vamos a probar que $\left\|J^t y_0\right\|_2 \geq c(1+\epsilon)^t$ por lo que $\left\|J^t\right\|_2 \geq c(1+\epsilon)^t$, luego por el teorema de Gelfand

Usamos que el radio espectral es el limite de cualquier norma matricial

$$\rho(J) = \lim_{t\to\infty}\left\|J^t\right\|^{1/t} \geq \lim_{t\to\infty} c^{1/t}(1+\epsilon) = 1+\epsilon$$

Y concluímos que $\chi^* \subseteq \mathcal{A}_g^*$.

En pos de eso fijemos $t \geq 1$ una iteración , $y_t = J^t x_0$, $z_1 = y_t$ y definamos $z_{i+1} = (Id - \alpha P_{S_i} H)\, z_i = z_i - \alpha \sum_{j\in S_i}\left(e_j^T H z_i\right) e_j$. Luego $y_{t+1} = z_{b+1}$, afirmo:

Esta demo es horrenda, hay que pensar una mejor y pionerla en el Anexo

Afirmación 4.3.5 *Sea* $y_t \in Ran(H)$, *luego existe* $i \in \{1,\ldots,b\}$ *y* $\delta > 0$ *tal que* $\alpha\sum_{j\in S_i}\left|e_j^T H z_i\right| \geq \delta\left\|z_i\right\|_2$

Lema 4.3.6 *Existe* $\epsilon > 0$ *tal que para todo* $t \in \mathbb{N}$:

$$y_{t+1}^T H y_{t+1} \leq (1+\epsilon) y_t^T H y_t$$

Demostración Manteniendo la notación previa a la afirmación:

$$\begin{aligned}
z_{i+1}^T H z_{i+1} \leq & \left[z_i^T - \alpha \sum_{j \in S_i} \left(e_j^T H z_i \right) e_j^T \right] H \left[z_i - \alpha \sum_{j \in S_i} \left(e_j^T H z_i \right) e_j \right] \\
= & \ z_i^T H z_i - \alpha \sum_{j \in S_i} \left(z_i^T H e_j \right) \left(e_j^T H z_i \right) - \alpha \sum_{j \in S_i} \left(e_j^T H z_i \right) \left(e_j^T H z_i \right) \\
& + \alpha^2 \left(\sum_{j \in S_i} \left(e_j^T H z_i \right) e_j \right)^T H \left(\sum_{j \in S_i} \left(e_j^T H z_i \right) e_j \right) \\
(\|H_{S_i}\|_2 \leq L_b) < & \ z_i^T H z_i - 2\alpha \sum_{j \in S_i} \left(e_j^T H z_i \right)^2 + \alpha^2 L_b \left\| \sum_{j \in S_i} \left(e_j^T H z_i \right) e_j \right\|_2^2 \\
= & \ z_i^T H z_i - \alpha \left(2 - \alpha L_b \right) \left\| \sum_{j \in S_i} \left(e_j^T H z_i \right) e_j \right\|_2^2 \\
(\alpha L_b < 1) < & \ z_i^T H z_i - \alpha \left\| \sum_{j \in S_i} \left(e_j^T H z_i \right) e_j \right\|_2^2
\end{aligned}$$

Luego juntando todo probamos que $z_i^T H z_i$ es decreciente y cumple la cota:

$$z_{i+1}^T H z_{i+1} < z_i^T H z_i - \alpha \left\| \sum_{j \in S_i} \left(e_j^T H z_i \right) e_j \right\|_2^2 \tag{6}$$

Por otro lado sabemos que para todo w vale:

$$w^T H w \geq \lambda_{min}(H) \|w\|_2^2 \geq -L_b \|w\|_2^2 \tag{7}$$

Usamos Cauchy Schwartz

Luego si usamos 4.3.5, 7 y Cauchy-Schwartz existe $i \in \{1, \ldots, b\}$ y $\delta > 0$ tal que:

$$\begin{aligned}
z_{i+1}^T H z_{i+1} < & \ z_i^T H z_i - \alpha \sum_{j \in S_i} \left(e_j^T H z_i \right)^2 \\
< & \ z_i^T H z_i - \frac{\alpha}{d} \left(\sum_{j \in S_i} \left| e_j^T H z_i \right| \right)^2 \\
< & \ z_i^T H z_i - \frac{\delta^2}{d\alpha} \|z_i\|_2^2 \\
< & \left(1 + \frac{\delta^2}{d\alpha L_b} \right) z_i^T H z_i
\end{aligned}$$

Tomando $\epsilon = \dfrac{\delta^2}{d\alpha L_b}$ probamos que $y_{t+1}^T H y_{t+1} \leq (1+\epsilon) y_t^T H y_t$ para $y_t \in Ran(H)$.

Si $y_t = y_N + y_R$ con $y_N \in Ker(H)$, $y_R \in Ran(H)$ entonces $y_t^T H y_t = y_R^T H y_R$ y $y_{t+1} = J y_t = y_N + J y_R$ por lo que $y_{t+1}^T H y_{t+1} = (J y_R)^T H (J y_R)$. Concluímos:

$$y_{t+1}^T H y_{t+1} = (J y_R)^T H (J y_R) \leq (1+\epsilon) y_R^T H y_R = (1+\epsilon) y_t^T H y_t$$

∎

Volviendo a la demostración general logramos probar que dado y_0 autovector de norma 1 de H con menor autovalor $\lambda < 0$ (pues $x^* \in \chi^*$) vale que:

$$\lambda_{min}(H) \|y_t\|_2^2 \leq y_t^T H y_t \leq (1+\epsilon)^t y_0^T H y_0 \leq (1+\epsilon)^t \lambda$$

Luego:

$$\|y_t\|_2^2 \geq \left(1 + \underbrace{\epsilon}_{} < \frac{1}{2}\right)^{\frac{t}{2}} \frac{\lambda}{\lambda_{min}(H)} \geq \frac{\lambda}{\lambda_{min}(H)} \left(1 + \frac{\epsilon}{4}\right)^t$$

Que era lo que queríamos demostrar con $c = \dfrac{\lambda}{\lambda_{min}(H)}$ y $\tilde{\epsilon} = \frac{\epsilon}{4}$.

∎

Corolario 4.3.7 *Sea g dado por el algoritmo de descenso por coordenadas con ecuación 4, bajo 9 y $\alpha < \frac{1}{L_b}$ se tiene que $\mu(W_g) = 0$.*

Demostración Por 4.3.3 y 4.3.4 tenemos que vale 3.2.4 y concluímos que $\mu(W_g) = 0$. ∎

Part III

Apéndice

APÉNDICE A