Método del gradiente conjugado para optimizar la función de costo en la utilización de regresión logística binaria para clasificación estelar

Andrés Bolaños

***Abstract*— El gradiente conjugado es un método utilizado para la optimización de una función de varias variables, ya sea para encontrar mínimos o máximos. La regresión logística binaria puede ser empleada para clasificar elementos a partir de las probabilidades que estos posean de pertencer a un cierto grupo, dadas variables independientes predictorias previas. Al momento de utilizarse este recurso estadístico, se puede calcular una función de costo a partir del error, misma que busca ser optimizada. Debido a que es una función que depende de dos variables, el gradiente conjugado es un algoritmo idóneo para conseguir este propósito. La regresión puede después emplearse para la clasificación estelar, que analiza grandes cantidades de cuerpos celestes y busca categorizarlos a partir de los datos espectrales que se posean. El objetivo de este documento es explicar estos temas y cómo se conectan entre sí.**

I. PALABRAS CLAVE

Regresión logística binaria, Gradiente conjugado, Clasificación estelar, Optimización, Función convexa, Costo, Error

II. INTRODUCCIÓN

La motivación para este texto es observar aplicaciones reales en el mundo para los métodos númericos aprendidos en clase. En la actualidad, existen muchas herramientas para desarrollar soluciones que permitan avanzar más en los campos de la ciencia y hacer crecer el conocimiento humano. Temas como *Machine Learning* y *Deep Learning* han permitido a varios científicos descubrir grandes hallazgos y formular inferencias para avanzar en los estudios.

En este documento se busca mostrar un ejemplo de la optimización numérica utilizada en astrofísica y astronomía. Se provee información previa de ciertos temas teóricos para poder saber cómo es que se emplean cada uno de ellos en esta aplicación. Para ser más específicos, es un caso de optmización sin restricciones en una función convexa y diferenciable.

Estamos en una época donde se puede hablar de *Big Data* y algoritmos eficientes, computacionalmente viables, que permiten maniobrar dichos datos para obtener información posterior. Esto es muy útil para la investigación astronómica. Varios algoritmos han sido utilizados en esta ciencia para la clasificación estelar como *Random Forest* y *Support Vector Machines*. La regresión logística es un recurso estadístico que ha funcionado en muchas otras áreas, particularmente bioestadística, pero que tiene muy poco uso en astronomía.

Al usar la regresión logística, se obtiene una función multivariable que determina el costo a partir del error que se obtiene después de verificar todos los resultados. Esta función multivariable puede ser optimizada para encontrar los dos coeficientes que minimizan el costo. Para esta optimización pueden ser utilzados varios métodos. En este ejemplo, se utiliza el método del gradiente conjugado para optimizar dicha función.

El gradiente conjugado es un método iterativo que permite encontrar las coordenadas que brindan el mínimo de la función si esta es diferenciable y convexa. Se apoya del gradiente para saber en qué dirección ir, es básicamente el método del descenso más pronunciado, que utiliza la función alpha y el gradiente, pero en lugar de únicamente optimizar alpha, recalcula el gradiente para realizar menos iteraciones utlizando una proyección.

III. MARCO TEÓRICO

A. *Optimización en astronomía y astrofísica*

La optimización puede ser descrita como: encontrar la mejor manera de realizar algo. Este enunciado determina que debe encontrarse una combinación de factores que permite obtener un resultado utilizando estos recursos de la mejor manera posible. Este es un arte de dos pasos, el primero de estos es transcribir el problema y colocar todas las barreras necesarias. Y el segundo es encontrar esa combinación eficiente para dicho problema. En astrofísica, la información de interés rara vez puede ser observada. Una razón para esto es que la data cruda como series de tiempo o imágenes astronómicas (galaxias, nebulosas, fondo de microondas cosmológico, ...) solo son explotables una vez corregida una variedad de distorsiones (fuentes molestas de primer plano, perturbaciones atmosféricas, resolución limitada del telescopio, ruido instrumental fuentes, etc.).

La información astrofísica a ser detectada o estimada a menudo no está directamente relacionada con los datos. Por ejemplo, información como la masa de un planeta extrasolar, la distancia a su estrella anfitriona, o incluso simplemente su misma existencia a veces puede ser revelada solo por medios sofisticados de algoritmos de detección y estimación en series temporales o imágenes. Una última razón es que la astronomía moderna lleva a construir instrumentos que dan acceso a observaciones de escala cada vez mayor, como el satélite Gaia y la extracción de información debe ser estadística y automatizada. En consecuencia, la información astrofísica siempre resulta de un complejo proceso de extracción más o menos supervisado. La forma general de optimización está dada por:

$$\min_{x \in C} F(x)$$

B. *Optimización convexa y diferenciable sin restricciones*

1) Convexidad: La convexidad de una curva o una superficie, es la zona que se asemeja al exterior de una circunferencia o una superficie esférica, es decir, que tiene su parte sobresaliente dirigida al observador. Es el concepto opuesto a la ’concavidad’.

Una parte C de un espacio vectorial real es convexa si para cada par de puntos de C, el segmento que los une está totalmente incluido en C; es decir, un conjunto es convexo si se puede ir de cualquier punto a cualquier otro en línea recta, sin salir del mismo.

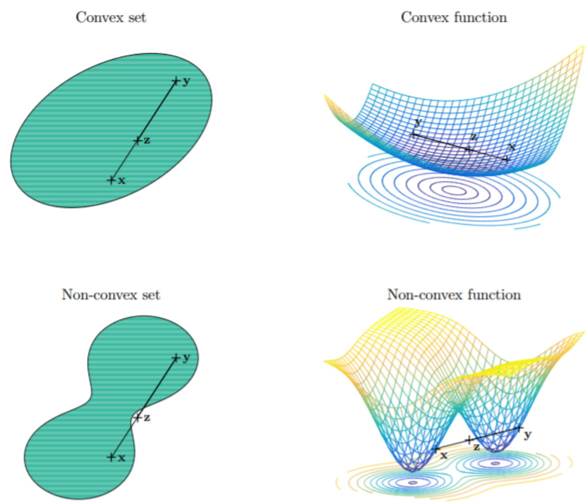

Fig. 1. Ilustración de un set convexo (arriba) y uno no convexo (abajo)

2) Función diferenciable: Junto con la convexidad, la diferenciabilidad es otra noción importante que permite distinguir entre los diferentes tipos de problemas de optimización. El concepto de función diferenciable es una generalización para el cálculo en varias variables del concepto más simple de función derivable. En esencia una función diferenciable admite derivadas en cualquier dirección y puede aproximarse al menos hasta primer orden por una aplicación afín.

La formulación rigurosa de esta idea intuitiva sin embargo es algo más complicada y requiere de conocimientos de álgebra lineal. Debe notarse que aunque una función de varias variables admita derivadas parciales según cada una de sus variables no necesariamente eso implica que sea una función diferenciable.

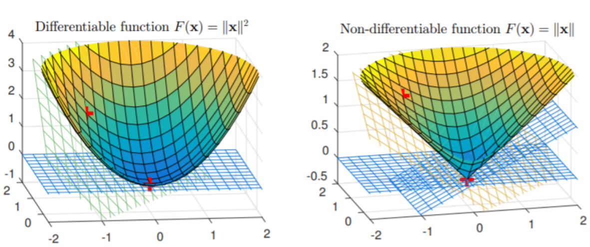

Fig. 2. Ejemplificación de una función no diferenciable y una diferenciable

C. *Gradiente conjugado*

Utilizando este método podemos encontrar mínimos en una función multivariable. El gradiente negativo de la función en un punto indica la dirección de mayor decrecimiento. Es una dirección correcta para buscar el mínimo de la función.

Proceso:

1) Definición de alpha: Dado un punto x y una dirección s (dada por el gradiente) se obtiene la siguiente función de una variable definida por:

$$\Phi(\alpha) = f(x + \alpha s)$$

2) Optimización de alpha: Ahora se optimiza $\Phi(\alpha)$ con alguno de los métodos de optimización de una variable, en este ejemplo se utilizará el método de la proporción áurea.

3) Salto: Alpha representa el salto que se dará en cada una de la iteraciones. Este salto se hará en la dirección que indique s, quien es determinada a partir del gradiente negativo valuado en x

4) Nueva X: La nueva X es calculada a partir de la anterior de la siguiente manera:

$$x_{n+1} = x_n + a_n s_n$$

5) Nueva S: Se calcula una nueva dirección, calculando la proyección. El gradiente en el punto x_n y en x_{n+1} se utilizan para conseguir una constante que modifica la nueva dirección.

6) Algoritmo: El algoritmo para utilizar el gradiente conjugado es el siguiente:

$\boldsymbol{x}_0 =$ initial guess
$\boldsymbol{g}_0 = \nabla f(\boldsymbol{x}_0)$
$\boldsymbol{s}_0 = -\boldsymbol{g}_0$
for $k = 0, 1, 2, \ldots$
  Choose α_k to minimize $f(\boldsymbol{x}_k + \alpha_k \boldsymbol{s}_k)$
  $\boldsymbol{x}_{k+1} = \boldsymbol{x}_k + \alpha_k \boldsymbol{s}_k$
  $\boldsymbol{g}_{k+1} = \nabla f(\boldsymbol{x}_{k+1})$
  $\beta_{k+1} = (\boldsymbol{g}_{k+1}^T \boldsymbol{g}_{k+1})/(\boldsymbol{g}_k^T \boldsymbol{g}_k)$
  $\boldsymbol{s}_{k+1} = -\boldsymbol{g}_{k+1} + \beta_{k+1} \boldsymbol{s}_k$
end

D. *Regresión logística binaria*

La regresión logística es un tipo de análisis de regresión utilizado para predecir el resultado de una variable categórica (una variable que puede adoptar un número limitado de categorías) en función de las variables independientes o predictoras. Es útil para modelar la probabilidad de un evento ocurriendo como función de otros factores. El análisis de regresión logística se enmarca en el conjunto de Modelos Lineales Generalizados (GLM por sus siglas en inglés) que usa como función de enlace la función logit. Las probabilidades que describen el posible resultado de un único ensayo se modelan, como una función de variables explicativas, utilizando una función logística.

La función para resolver el modelo de regresión logística se define como:

$$y_p = \sigma(w^t x + b)$$

Utilizando la función de Sigmoid para que los valores de las probabilidades queden dentro de 0 y 1:

$$\sigma(Z) = \frac{1}{1 + e^{-Z}}$$

Ahora definiendo la función de costo para este problema obtenemos:

$$J(\beta_0, \beta_1) = \frac{\sum_1^m y^i \log(y_p^i) + (1 - y^i) \log(1 - y_p^i)}{m}$$

donde,

$$\beta_0 = w^t$$
$$\beta_1 = b$$

Vemos que esta función de costo tiene una forma convexa y es diferenciable.

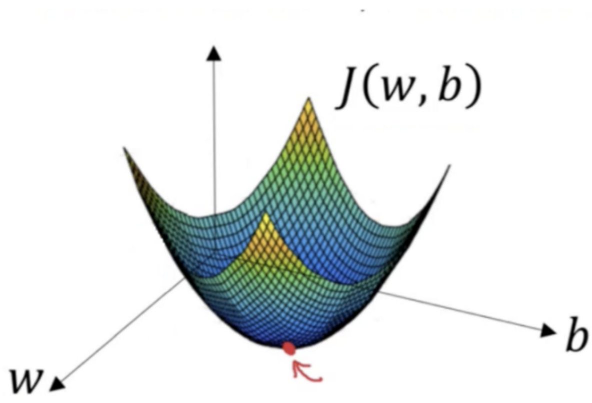

Fig. 3. Figura donde se demuestra que la función es convexa y posee un mínimo

E. *Clasificación estelar*

La clasificación estelar es la clasificación de las estrellas en función de sus características espectrales. La radiación electromagnética procedente de la estrella es analizada mediante su división por un prisma o por una red de difracción en un espectro, mostrando así el arcoíris de colores (espectro electromagnético visual) entremezclados con líneas de absorción. Cada línea indica un ion de un determinado elemento químico, junto con la intensidad de la línea que determina la abundancia de ese ion. La abundancia relativa de los diferentes iones varía con la temperatura de la fotosfera. La clase espectral de una estrella es un código corto que resume el estado de ionización, dando una medida objetiva de la temperatura de la fotosfera y la densidad.

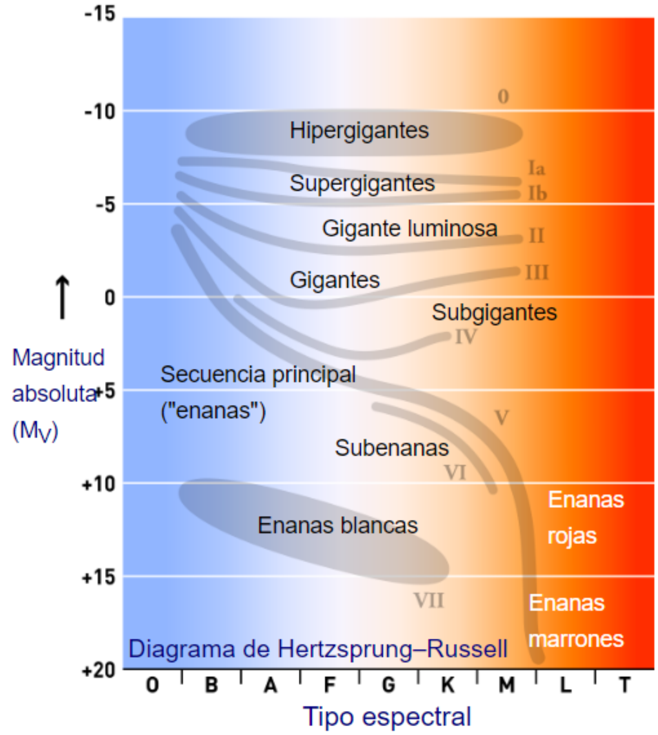

Fig. 4. Ejemplo de clasificación estelar

IV. Aplicación

A. *Explicación*

Teniendo todas las piezas juntas, podemos comenzar a hablar acerca de la aplicación que tiene este algoritmo en la regresión logística para clasificar estrellas.

El gradiente conjugado es un método que permite encontrar los máximos y mínimos de una función definida por varias variables. Utiliza el gradiente negativo para buscar dónde es que está el descenso más pronunciado, y posteriormente, va recalculando las direcciones para no pasar por lugares donde ya lo hizo. Optimiza alpha, que es el tamaño del salto en la dirección dada por el gradiente. Esta optimización puede ser hecha con cualquier método de optimización de una variable, para este ejemplo en específico se ha utilizado el método de *Golden Ratio*.

La regresión logística puede ser utilizada para clasificar elementos a partir de características conocidas. Teniendo un set de datos, podemos entrenar a nuestro algoritmo para conseguir predicciones de pertenencia. En este caso particular, se utiliza la regresión logística binaria, que utiliza condiciones de probabilidad de acierto o fallo (0 y 1).

Utilizando este modelo estadístico, se puede calcular su pérdida total. Esta función está definda por el error de cada uno de los datos presentados en el set. Cuando hacemos la sumatoria de todas las pérdidas de cada dato, obtenemos el costo total del modelo. Este costo total termina siendo definido por las dos constantes que acompañan a x en la ecuación de Sigmoid.

Por tanto, esta función puede ser optimizada al encontrar su mínimo. Si encontramos esa combinación de valores que brindan el costo más pequeño, podemos asegurar de que el modelo está utilizando la menor cantidad posible de error.

Teniendo los valores que minimizan el error y la función de regresión logística binaria junto con el set de datos a utilizar, se puede comenzar con la clasficación estelar.

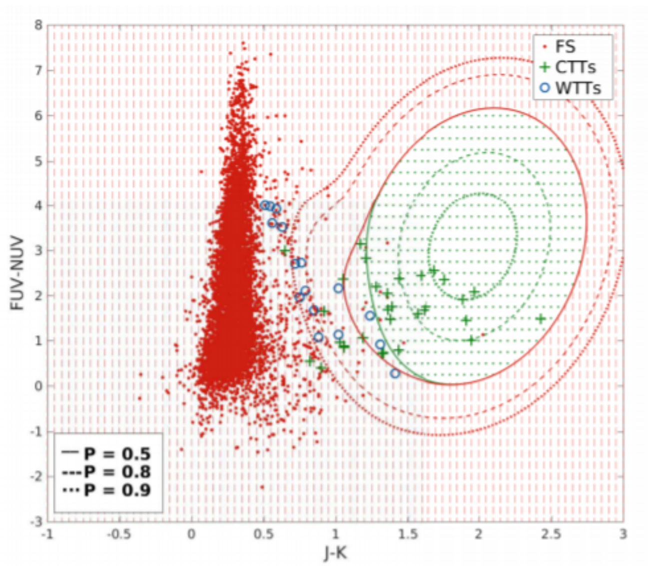

Fig. 5. Resultados del ajuste de regresión logística de la muestra de FS y TT de GdC2015. A cada punto del plano se le ha asignado la clase más probable con un código de color: líneas rojas verticales para FS y puntos verdes para CTT. También se han trazado tres líneas de probabilidad para separar las regiones de probabilidad de 0.5, 0.8 y 0.9 para FS (rojo) y CTT (verde).

B. Funcionamiento del algoritmo

Para hacer que el algoritmo pueda funcionar en este ejemplo, se debe definir la función en Python junto con su gradiente. Así también, la función alpha que es la misma función original evaluada en un término $x_n + a_n s_n$. Después, únicamente es necesario mandar a llamar al método del gradiente conjugado para obtener las dos constantes que permiten minimizar el error.

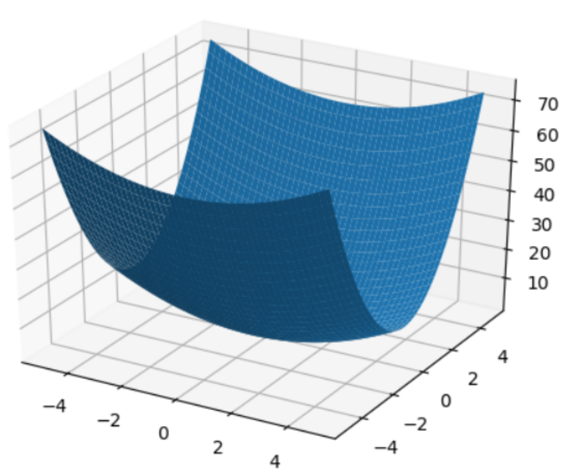

Fig. 6. Gráfica en 3d de la función de costo

Utilizando librerías de Python, es posible graficar los movimientos que dio el algoritmo a través de las curvas de nivel de la función. A continuación, se presenta una comparación entre el método del descenso más pronunciado versus el gradiente conjugado.

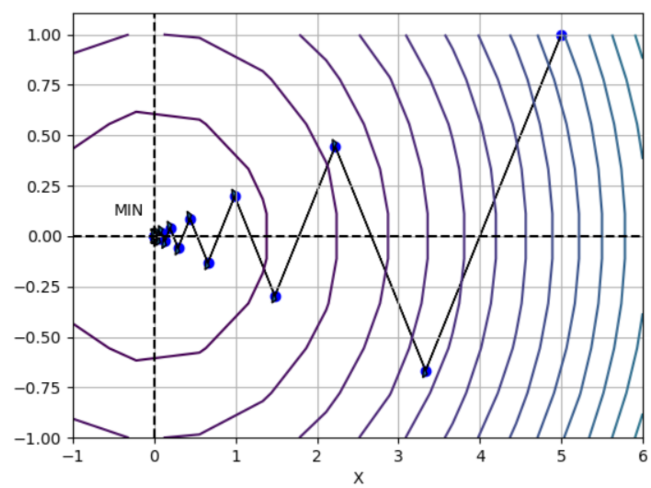

Fig. 7. Curvas de nivel utilizando el descenso más pronunciado

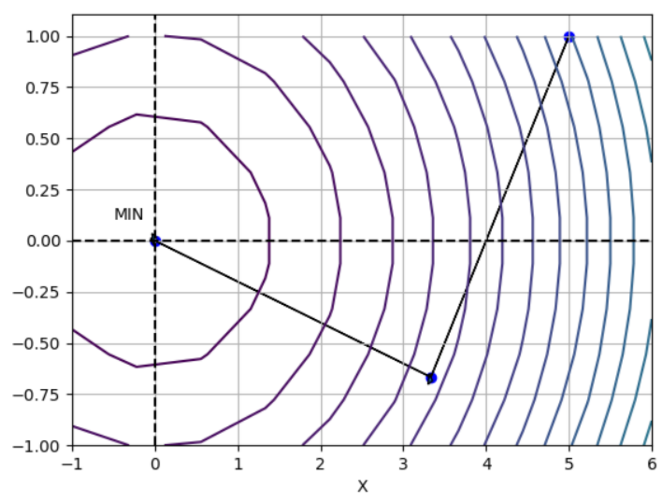

Fig. 8. Curvas de nivel utilizando el gradiente conjugado

Vemos cómo es que utilizando el gradiente conjugado puede realizarse la misma operación en menos iteraciones.

El gradiente conjugado es una alternativa con ventajas al método de Newton, ya que no necesita el cálculo explícito de la segunda derivada (Hessiano).

Al igual que el método del descenso más pronunciado, es necesario calcular el gradiente de la función. Pero es más eficiente ya que el gradiente se modifica para no buscar en direcciones ya recorridas.

C. Código

```python
def conjugado(xk,f,tol):
    #Error igual a algo mayor para que cumpla la condición
    #inicial
    e=1
    #Arreglo con todos los puntos
    X=[]
    #Arreglo con las direcciones calculadas
    S=[]
    #XK recibe el primer valor de x
    X.append(xk)
    #el gradiente que servirá para calcular la nueva dirección
    gk = gradiente(xk)
    #Nuestra primera dirección, es el negativo del gradiente
    sk=np.dot(-1.0,gradiente(xk))
    while(e>tol):
        #se optimiza alpha, utilizando Golden Ratio de python
        ak1=sp.optimize.golden(alpha,args=(xk,sk))
        #Obtengo mi nuevo paso, al multiplicar la dirección
        # y la magnitud alpha
        sk1=np.dot(ak1,sk)
        #Guarda la dirección en el arreglo
        S.append(sk1)
        #Se calcula la nueva x
        xk1=xk+np.dot(ak1,sk1)
        #Se guarda en el arreglo de X
        X.append(xk1)
        #Se calcula el error
        e=np.linalg.norm(xk1-xk)
        #Mi nueva x es ahora la x del algoritmo
        xk=xk1
        #mi nuevo gradiente al evaluar la función en la nueva x
        gk1 = gradiente(xk)
        #la magnitud de cambio bk es calculada utilizando g0 y g1
        bk1 = np.dot(np.transpose(gk1),gk1)/np.dot(np.transpose(gk), gk)
        #obtengo una nueva dirección
        sk1 = np.dot(-1.0,gk1) + np.dot(bk1,sk)
        #mi dirección anterior se convierte en la nueva para los cálculos
        #del algoritmo
        sk = sk1
    #esta función regresa el vector de coordenadas, los arreglos
    #de dirección y de
    #coordenadas anteriores
    return xk,X,S
```

REFERENCES

[1] Stephane Canu, Rémi Flamary, David Mary. Introduction to optimization with applications in astronomy and astrophysics. 2016. ffhal-01346134f

[2] https://www.internalpointers.com/post/cost-function-logistic-regression

[3] Bombal, R. Marín, Vera: Problemas de Análisis matemático: Cálculo Diferencial, 1988, ed. AC, ISBN 84-7288-101-6.

[4] Rauw, G., Manfroid, J., Gosset, E., Nazé, Y., Sana, H.; De Becker, M.; Foellmi, C.; Moffat, A. F. J. (2007): Early-type stars in the core of the young open cluster Westerlund 2. Astronomy and Astrophysics 463 (3): 981-991.

[5] https://towardsdatascience.com/understanding-logistic-regression-using-a-simple-example-163de52ea900

[6] https://medium.com/technology-nineleaps/logistic-regression-gradient-descent-optimization-part-1-ed320325a67e

[7] https://arxiv.org/pdf/1805.09551.pdf

[8] S. Bourguignon, D. Mary, and E. Slezak. Sparsity-based denoising of hyperspectral astrophysical data with colored noise: Application to the muse instrument. In Hyperspectral Image and Signal Processing: Evolution in Remote Sensing (WHISPERS), 2010 2nd Workshop on, pages 1–4. IEEE, 2010.

[9] https://hea-www.harvard.edu/AstroStat/slog/groundtruth.info/AstroStat/slog/tag/logistic-regression/index.html

[10] Agnello, B. C. Kelly, T. Treu, and P. J. Marshall. Data mining for gravitationally lensed quasars. Monthly Notices of the Royal Astronomical Society, 448(2):1446–1462, 2015.

[11] N. M. Ball and R. J. Brunner. Data mining and machine learning in astronomy. International Journal of Modern Physics D, 19(07):1049–1106, 2010.

[12] R. E. Carrillo, J. D. McEwen, and Y. Wiaux. Purify: a new approach to radiointerferometric imaging. Monthly Notices of the Royal Astronomical Society, 439(4):3591–3604, 2014.

[13] S. Bubeck. Convex optimization: Algorithms and complexity. Foundations and Trends in Machine Learning, 8(3-4):231–357, 2015.

[14] J. Christensen-Dalsgaard. Lecture notes on stellar oscillations, http: // astro. phys. au. dk/ jcd/ oscilnotes/ , 2003.

[15] N. Komodakis and J.-C. Pesquet. Playing with duality: An overview of recent primal? dual approaches for solving large-scale optimization problems. Signal Processing Magazine, IEEE, 32(6):31–54, 2015.

[16] J.-L. Starck, F. Murtagh, and J. Fadili. Sparse Image and Signal Processing: Wavelets and Related Geometric Multiscale Analysis. Cambridge University Press, 2015.

[17] S. J. Wright, R. D. Nowak, and M. A. Figueiredo. Sparse reconstruction by separable approximation. Signal Processing, IEEE Transactions on, 57(7): 2479–2493, 2009.

[18] D. G. York, J. Adelman, J. E. Anderson Jr, S. F. Anderson, J. Annis, N. A. Bahcall, J. Bakken, R. Barkhouser, S. Bastian, E. Berman, et al. The sloan digital sky survey: Technical summary. The Astronomical Journal, 120(3): 1579, 2000.

[19] Cristianini N, Shawe-Taylor J (2000) An Introduction to Support Vector Machines: and other Kernel-based learning methods. Cambridge University Press

[20] Kurcz A, Bilicki M, Solarz A, Krupa M, Pollo A, et al (2016) Towards automatic classification of all WISE sources. AA 592:A25