PROYECTO DE INVESTIGACIÓN

Proyecto Interno ■ Proyecto Semilla ❑ Proyecto Junior ❑ Proyecto Multi e Inter Disciplinario ❑

Investigación Básica ■ Investigación Aplicada ❑

DEPARTAMENTO(S) Y/O INSTITUTO:
1. CENTRO DE MODELIZACIÓN MATEMÁTICA MODEMAT
2. DEPARTAMENTO DE MATEMÁTICA

LÍNEA(S) DE INVESTIGACIÓN (verificable en el SAEW):
1. OPTIMIZACIÓN Y CONTROL
2. BIOLOGÍA DE ORGANISMOS

CAMPO DEL CONOCIMIENTO (Ver Anexo A: Detalle de los campos del conocimiento)		
Campo amplio	Campo detallado	Campo especifico
Ciencias Físicas, Ciencias Naturales, Matemáticas y Estadísticas	Matemáticas y estadística	Matemáticas

DISCIPLINA CIENTÍFICA *(Marque X, solamente una opción)*	
Ciencias Naturales y Exactas	X
Ingeniería y Tecnologías	
Ciencias Médicas	
Ciencias Agrícolas	
Ciencias Sociales	
Humanidades	

OBJETIVO SOCIOECONÓMICO *(Marque X, solamente una opción)*	
Exploración y explotación del medio terrestre	
Ambiente	
Exploración y explotación del espacio	
Transporte, telecomunicaciones y otras infraestructuras	
Energía	
Producción y tecnología industrial	
Salud	
Agricultura	
Educación	
Cultura, ocio, religión y medios de comunicación	
Sistemas políticos y sociales, estructuras y procesos	
Defensa	
Avance general del conocimiento: I+D financiada con los Fondos Generales de Universidades (FGU)	X
Avance general del conocimiento: I+D financiados con otras fuentes	

Alcance Territorial (*Marque X, solamente una opción*)			
Institucional	X	Nacional	

Institucional	X	Nacional	
Parroquial		Internacional	
Cantonal		No definido	
Provincial			

Proyecto de Investigación

Título (mínimo 10 palabras):
RESOLUCIÓN NUMÉRICA DE PROBLEMAS DISPERSOS GENERALIZADOS CON APLICACIONES EN MODELOS ECOLÓGICOS DE PREDICCIÓN DE LA PRESENCIA DE ESPECIES EN AREAS NATURALES

Resumen del proyecto (máximo 200 palabras)

Consideramos problemas del tipo LASSO (*least absolute shrinkage and selection operator*) generalizados, que son modelos de regresión muy utilizados en *estadística*, *aprendizaje automático* e imágenes, para realizar predicciones de una variable desconocida en función de información existente. En particular, los modelos de *máxima entropía* (maxent) usados para estimar una distribución de probabilidad de la presencia de especies en territorios naturales son un caso particular de estos problemas, los cuales se formulan como un problema de optimización de la forma:

$$\min_x f(x) + \beta\|Cx\|_1,$$

donde f representa el costo del ajuste de los datos y el termino $\beta\|Cx\|_1$ corresponde a una penalización para inducir soluciones poco densas, i.e. con un número grande de entradas nulas en el rango de C. En particular, si C representa al gradiente discreto, se promueve que la solución varíe poco en un sentido local. Esta característica es muy utilizada en la reconstrucción de imágenes y señales.

Debido a que estos problemas son no diferenciables por la presencia del termino de penalización y la presencia del operador C, su resolución numérica no es trivial y requiere de herramientas del análisis convexo para su resolución. Proponemos extender el método desarrollado en [1], que considera $C=I$ *(I=identidad)*, para el caso general en que C es una matriz rectangular. Esta elección implica retos matemáticos significativos, tanto teóricos como numéricos debido a la correlación de las variables por la matriz C. En este caso, no se tiene una interpretación por componentes de las cantidades asociadas a la variable óptima, como en el caso de [1]

Palabras clave (4-6):

Lasso generalizado, optimización no diferenciable, algoritmos de descenso, modelos de máxima entropía

2 Objetivos, limitaciones, hipótesis y resultados esperados de esta propuesta de investigación

2.1 Objetivos

2.1.1 Objetivo General

Desarrollar un método numérico para la resolución del problema de tipo LASSO generalizado para su aplicación en la resolución de modelos de máxima entropía

2.1.2 Objetivos Específicos

a. Estudiar el problema de tipo LASSO generalizado: existencia y condiciones de optimalidad útiles en el diseño del método numérico.
b. Extender la noción de direcciones ortantes para el caso generalizado
c. Extender el algoritmo OESOM al caso generalizado
d. Implementar el método en los lenguajes: R y Python, y realizar experimentos numéricos

e. Aplicar el método en la resolución numérica de problemas de máxima entropía relacionados con la predicción de la presencia de especies en áreas naturales.

2.2 Limitaciones (Aspectos que quedan fuera del alcance del Proyecto de Investigación)

a. No se incluye el desarrollo software
b. Se considerarán datos disponibles de una especie ecuatoriana.

2.3 Hipótesis (Responden al problema de investigación)

a. El algoritmo OESOM c.f. [] puede ser extendido al problema tipo LASSO generalizado
b. Se puede extender la noción de direcciones ortantes al caso general, mediante una noción más general a la de ortogonalidad.
c. El método puede ser utilizado para resolver problemas de máxima entropía.

2.3 Detalle de los resultados esperados (con relación a los objetivos)

a. Desarrollo e implementación del código de un método numérico
b. Un artículo científico que será remitido a una revista ISI/SCOPUS
c. Presentación de los resultados en un congreso o seminario nacional o internacional
d. Tesis de grado

3 Relevancia de la propuesta de investigación y su relación con la(s) líneas de investigación

Este proyecto se enmarca en tres de las líneas de investigación del departamento:

1. OPTIMIZACION MATEMATICA Y CONTROL: esta investigación continua con la investigación del algoritmo de optimización desarrollada en []. La extensión no es trivial debido a que la presencia del operador C correlaciona las variables del vector solución y por tanto el análisis por componentes que se aplicó en [1] ya no es válido y es necesario estudiar el problema de optimización generalizado.
2. BIOLOGIA DE ORGANISMOS: por cuanto se aplicará en la predicción de presencia de especies en áreas naturales del ecuador.
3. MODELIZACION MATEMATICA Y CALCULO CIENTIFICO: ya que se plantea el desarrollo de un algoritmo aplicado a modelos ecológicos que involucran experimentación numérica.

4 Impacto de la investigación

4.1 Impacto Social (máximo 250 palabras)
4.2 Impacto Económico (máximo 250 palabras)
4.3 Impacto Político (máximo 250 palabras)

4.4 Impacto Científico (máximo 250 palabras)

El desarrollo de métodos numéricos contribuirá al conocimiento en la resolución problemas tipo LASSO generalizados y sus aplicaciones. En particular, este tipo de métodos tienen un potencial de uso en aprendizaje automático como son las máquinas de soporte vectoriales (SVM) y reconstrucción de imágenes: reducción de ruido e inpainting.

4.5 Otro Impacto (máximo 250 palabras)
Se espera que los métodos y aplicaciones que se investigaran en este proyecto sean de utilidad para la colaboración futura con instituciones encargadas del estudio del medio ambiente, en la formulación de proyectos conjuntos para el estudio de poblaciones de especies endémicas ecuatorianas, en particular de

aquellas en peligro de extinción. El levantamiento y predicción de presencia de especies en áreas naturales tendría impacto en la creación de políticas de conservación basadas en métodos cuantitativos.

5 Productos esperados

Tipo de Producto:	Marcar con una "X"
a. Publicaciones científicas y/o patente (obligatorio);	**X**
b. Disertación a la comunidad politécnica;	X
c. Trabajo de titulación de acuerdo a lo que establece el Reglamento de Régimen Académico y la Normativa Interna de la EPN;	X
d. Aplicación tecnológica construida o implementada;	
e. Perfil de proyecto de mayor impacto científico, técnico, pedagógico o de innovación.	X

6 Descripción, metodología y diseño del proyecto

6.1 Descripción, metodología y diseño del proyecto (Máximo dos carillas)

Muchos problemas de regresión estadística y procesamiento de imágenes (ver [4] y [5]) se formulan como el siguiente problema de optimización:

$$\min_x f(x) + \beta \|Cx\|_1 .$$

Aquí, *f* representa un costo asociado al ajuste de datos: una regresión logarítmica en los métodos de máxima entropía, o puede ser un costo asociado a la fidelidad de una imagen procesada respecto a la original. Por otra parte, el término $\beta\|Cx\|_1$ representa el costo de la estructura de la solución la cual en el caso de la norma-1 induce soluciones que poseen muchas entradas nulas en el rango de *C*. Si por ejemplo *C* corresponde a la identidad, la solución tendrá muchas entradas nulas (dependiendo del tamaño de β) y por tanto recibe el nombre de solución "sparse", cf. [10],[1]. En el caso generalizado, *C* puede representar el gradiente discreto de x y en consecuencia la solución tendrá un comportamiento de variación acotada ya que la penalización induce entradas nulas de su gradiente. En ambos casos, dichas estructuras son utilizadas en diversas aplicaciones en medicina, ecología ([4][7][8]) , procesamiento de imágenes ([5]) y señales ([6]), entre otros.

La naturaleza no diferenciable de dichos problemas conlleva a que su análisis y resolución numérica implican otras dificultades comparados con aquellos problemas cuyas funciones de costo son diferenciables. De hecho, en la ausencia de derivadas, los métodos de primer orden son la primera opción en la resolución. Además, en el caso generalizado, la presencia de la matriz *C* produce una correlación de las variables, por lo que el análisis por componentes realizado en [1] ya no es válido.

El método OESOM ([1]) tiene la particularidad de ser un método de segundo orden que, bajo ciertas condiciones, es equivalente al método de Newton semi-suave. Existen pocos métodos con consideren este tipo de información, ya que al no ser diferenciables no existe segunda derivada en forma de matriz Hessiana. LA información que se considera en [1] está determinada por el salto que ocurre en la norma-1 cerca del origen y que es aproximada tomando en cuenta una regularización para añadir información sin regularizar el problema original.

Continuando con la investigación realizada en [1], nos enfocamos en extender el método OESOM que considera *C=I (identidad)* para el problema generalizado, donde *C* es una matriz rectangular, por ejemplo: el gradiente discreto, como fue mencionado anteriormente. De esta manera, la extensión del algoritmo OESOM a este caso debería contemplar 3 problemas fundamentales a ser superados:

1. Extensión de la noción de direcciones de descenso ortantes: la correlación de las variables por C impide el uso de direcciones ortantes introducidas en [1]
2. Determinación de un subgradiente de norma mínima de la función objetivo.
3. Información generalizada de segundo orden introducida en [1] en la presencia de C.

En base al estudio de la solución de estos tres problemas sería posible extender el método OESOM al caso generalizado y realizar un análisis de las propiedades del método. A diferencia de otros métodos para resolver estos problemas (ver [5], [6] y [10]), el método OESOM considera información del salto que se produce en la no diferenciabilidad de la norma-1, lo cual permite acelerar el método de segundo orden, ver [1].

Una vez formulado el método generalizado, éste será implementado en Python y R para ser sujeto de pruebas numéricas que nos permitan recabar evidencia numérica sobre la eficiencia del método en diferentes tipos de problemas. En particular problemas asociados a:

i) *eliminación de ruido en imágenes* usando el método LAD (mínima desviación absoluta), similar al problema de variación acotada, utilizado cuando la hipótesis de que el ruido en la imagen no sigue una distribución gaussiana cf. [5]. Recopilaremos algunas imágenes para aplicar el algoritmo a estos problemas como test para medir la eficiencia del método.

ii) De manera análoga, profundizando en la aplicación del método propuesto, se considerarán *problemas de máxima entropía* aplicados en la predicción de la presencia de especies en áreas naturales. Estos problemas modelan la probabilidad de presencia de una especie basado en datos de las condiciones ambientales cuando la presencia es positiva cf. [3][4][7]. En Ecuador, la institución que se encarga de la recolección de dichos datos es la Secretaría de la Biodiversidad. Usando los datos recolectados por dicha entidad, implementaremos el método en el lenguaje R para comparar con el desempeño del software maxent ([4][8]) para resolver el modelo de máxima entropía para predecir la presencia de una especie en un área determinadas por dichos estudios.

Bibliografía

[1] De Los Reyes, J. C., Loayza, E., & Merino, P. (2017). Second-order orthant-based methods with enriched Hessian information for sparse ℓ1-optimization. *Computational Optimization and Applications*, *67*(2), 225-258.

[2] Duan, J., Soussen, C., Brie, D., Idier, J., Wan, M., & Wang, Y. P. (2016). Generalized LASSO with under-determined regularization matrices. *Signal processing*, *127*, 239-246.

[3] Elith, J. et al. (2011) A statistical explanation of MaxEnt for ecologists. *Diversity and Distributions*, 17 43-57. DOI: 10.1111/j.1472-4642.2010.00725.x

[4] Fithian, W., & Hastie, T. (2013). Finite-sample equivalence in statistical models for presence-only data. *The annals of applied statistics*, *7*(4), 1917.

[5] Fu, H., Ng, M. K., Nikolova, M., & Barlow, J. L. (2006). Efficient minimization methods of mixed l2-l1 and l1-l1 norms for image restoration. *SIAM Journal on Scientific computing*, *27*(6), 1881-1902.

[6] Loris, I., & Verhoeven, C. (2012). Iterative algorithms for total variation-like reconstructions in seismic tomography. *GEM-International Journal on Geomathematics*, *3*(2), 179-208.

[7] Merow, C., Smith, M. J., & Silander Jr, J. A. (2013). A practical guide to MaxEnt for modeling species' distributions: what it does, and why inputs and settings matter. *Ecography*, *36*(10), 1058-1069.

[8] Phillips, S. J., Anderson, R. P., Dudík, M., Schapire, R. E., & Blair, M. E. (2017). Opening the black box: an open-source release of Maxent. *Ecography*, *40*(7), 887-893.

[9] Tibshirani, R. J. (2011). *The solution path of the generalized lasso*. Stanford University press.

[10] Zhu, Y. (2017). An augmented ADMM algorithm with application to the generalized lasso problem. *Journal of Computational and Graphical Statistics*, *26*(1), 195-204.

7 Infraestructura, equipos y fondos adicionales.

7.1 Infraestructura y equipos

- *Indicar la infraestructura y equipos **disponibles** para la ejecución del proyecto, con la ubicación actual de los mismos*

Infraestructura	Equipos	
	Nombre del Equipo	**Ubicación del Equipo**
Laboratorio de Cálculo Científico	HPC-Modemat (Quinde Cluster)	Laboratorio de Cálculo Científico: Datacenter del Centro de Modelización Matemática MODEMAT

7.2 Breve justificación del equipo requerido

- *Justificar la infraestructura y equipos **solicitados** para la ejecución del proyecto e indicar el departamento en el cual se ubicará dicho equipamiento.*

No aplica

7.3 Fondos Adicionales

- *Otros fondos de otros organismos (si los hubiere)*

No aplica