Informe narrativo de análisis y experimentos

Camilo Humberto Perez Fleita

16 de noviembre de 2025

Análisis detallado de la función

$$f(x, y) = x^2 e^y + y^2 e^x.$$

Enlace al repositorio en GitHub

`github.com/perezcam/OptModelsProject`

Desde una primera observacion se ve que siempre toma valores no negativos: cada término es un cuadrado multiplicado por una exponencial positiva, de modo que $f(x, y) \geq 0$ para todo (x, y). Esa simple observación ya sugiere que el valor mínimo absoluto será 0, y sólo puede alcanzarse cuando ambos cuadrados valen cero, es decir, en $(0, 0)$. Más adelante confirmaremos que el origen no sólo hace $f = 0$, sino que también es un punto donde la funcion es estrictamente convexa en un intervalo del punto.
Véase la Fig. 3 para un mapa de contornos cerca del origen que ilustra el mínimo local y la curvatura positiva.

Prefiero presentar el análisis como un recorrido. Primero, miro la suavidad: f está construida con polinomios y exponenciales, de manera que es tan suave como se quiera (derivable todas las veces).El gradiente, es

$$\nabla f(x, y) = \begin{bmatrix} 2x\, e^y + y^2 e^x \\ x^2 e^y + 2y\, e^x \end{bmatrix}.$$

Obtencion de la *matriz Hessiana* para mayor analisis,

$$H(x, y) = \begin{bmatrix} 2e^y + y^2 e^x & 2x\, e^y + 2y\, e^x \\ 2x\, e^y + 2y\, e^x & x^2 e^y + 2e^x \end{bmatrix}.$$

En el origen, $e^0 = 1$, las entradas cruzadas se anulan y queda $H(0, 0) = \mathrm{diag}(2, 2)$, que tiene curvaturas positivas en todas las direcciones. Esa foto local confirma que $(0, 0)$ es un mínimo local y, como ya sabíamos que $f \geq 0$ y alcanza 0 justamente ahí, también es el mínimo global.

Hay un concepto que pude aprender en la investigacion q resulta util para el analisis el cual es : **coercividad**. Una función es *coerciva* cuando, si uno se va *muy lejos* en cualquier dirección, el valor de la función crece sin límite. Aquí eso *no* ocurre. Si caminamos por la diagonal del tercer cuadrante, poniendo $x = y = -t$ y dejando que t crezca, aparece

$$f(-t, -t) = t^2 e^{-t} + t^2 e^{-t} = 2t^2 e^{-t} \longrightarrow 0 \quad \text{cuando } t \to \infty.$$

Es decir, marcharse hacia $(-\infty, -\infty)$ no hace explotar la función, sino más bien la aplasta hasta casi cero. Esa falta de coercividad es importante porque explica varios comportamientos

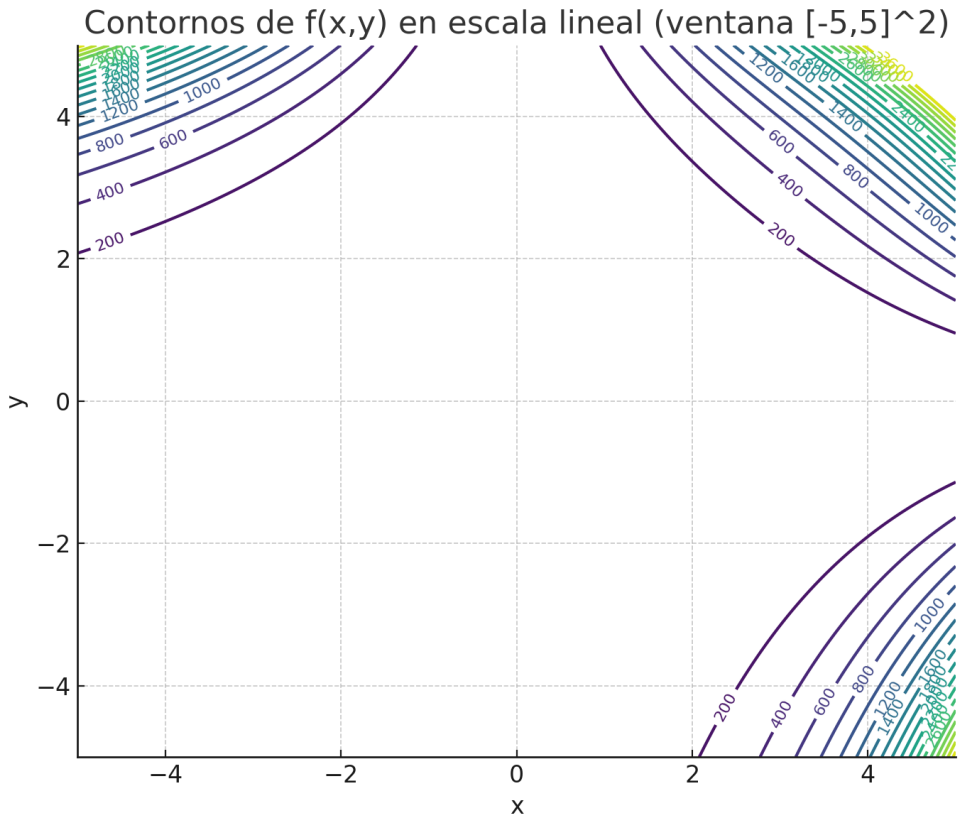

Figura 1: Curvas de nivel de $f(x, y)$ en escala lineal en $[-5, 5]^2$. Cada línea representa $f(x, y) = c$. Al estar en escala real, se aprecia directamente cómo cambian los niveles cerca del origen.

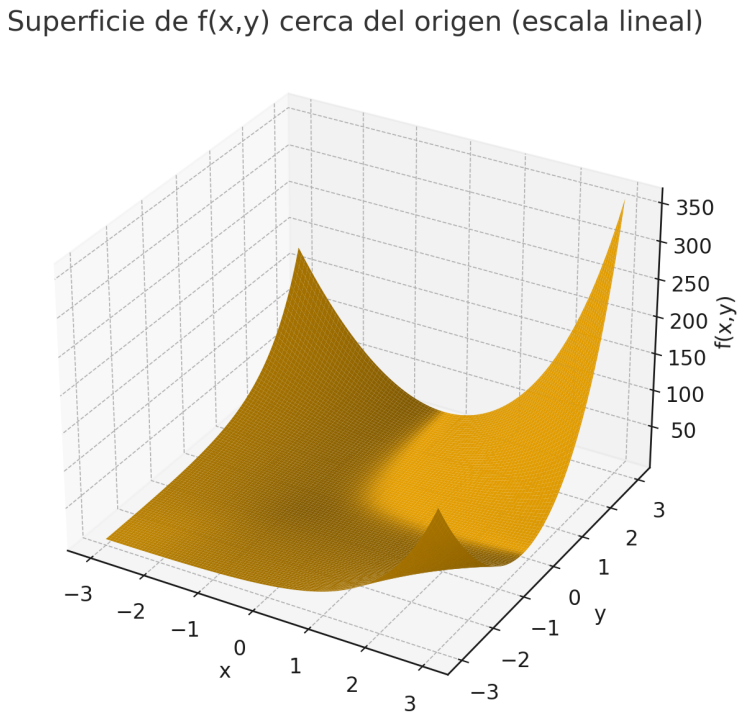

Figura 2: Superficie $z = f(x, y)$ cerca del origen (escala lineal). Vista directa de la función como gráfica en $\mathbb{R}^3$, útil para intuir la curvatura local sin transformaciones.

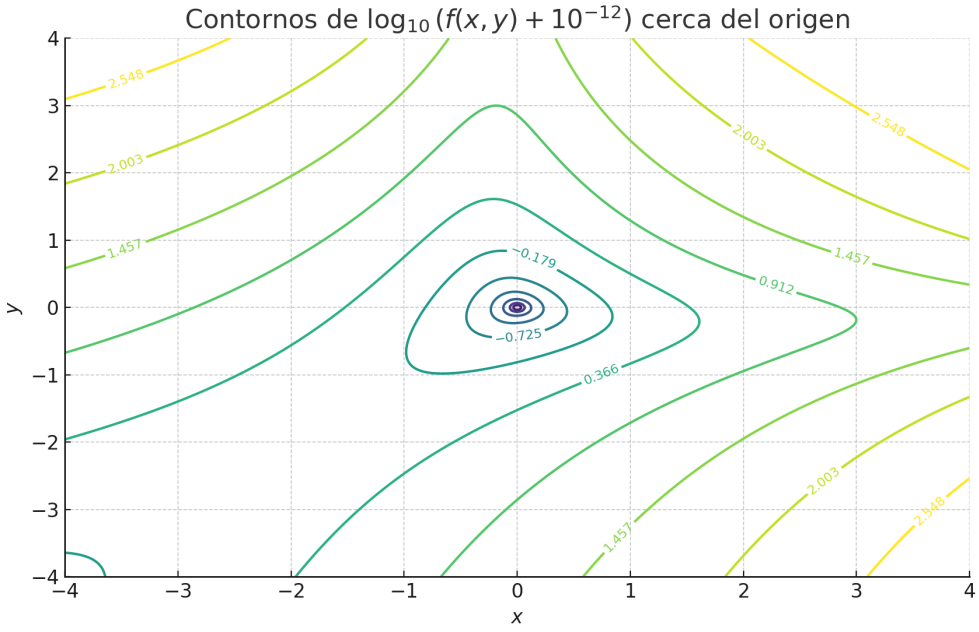

Figura 3: Curvas de nivel de $\log_{10}(f(x,y)+10^{-12})$ cerca del origen. La escala logarítmica permite ver niveles muy pequeños y moderados sin saturación.

de los algoritmos en los experimentos: hay zonas inmensas, en el tercer cuadrante, donde f y su pendiente son diminutas, y es fácil que un método declare "listo" sin estar cerca del origen.

Otra cuestión que ayuda a intuir cómo se moverán los métodos es la convexidad. En el origen el cuenco es perfecto, pero lejos de ahí la curvatura no siempre acompaña: la Hessiana puede tener una dirección de curvatura negativa; por ejemplo, en $(1,-1)$ tiene un autovalor mínimo de aproximadamente $-0{,}216$, lo cual indica que el paisaje no es globalmente convexo.

Valor esperado en el rango $[-100,100]$

Para medir qué tan "grande" es f de media cuando x e y se eligen al azar de manera uniforme en $[-100,100]$, calculé el valor esperado tanto *analíticamente* como por *simulación Monte Carlo*.

La cuenta exacta es sencilla de averiguar si usamos la independencia. Si $X, Y \sim \mathrm{Unif}[-a,a]$ con $a = 100$, entonces

$$\mathbb{E}[f(X,Y)] = \mathbb{E}[X^2 e^Y] + \mathbb{E}[Y^2 e^X] = 2\,\mathbb{E}[X^2]\,\mathbb{E}[e^Y],$$

y se tiene $\mathbb{E}[X^2] = a^2/3$ y $\mathbb{E}[e^Y] = \frac{1}{2a}\,(e^a - e^{-a}) = \frac{\sinh(a)}{a}$. Juntando todo,

$$\mathbb{E}[f(X,Y)] = \frac{2a}{3}\,\sinh(a).$$

Para $a = 100$ esto da

$$\mathbb{E}[f(X,Y)] \approx 8{,}96 \times 10^{44}.$$

La simulación con 400 pares (X_i, Y_i) extraídos uniformemente en $[-100,100]^2$ entregó una media muestral de

$$\bar{f} \approx 1{,}15 \times 10^{45},$$

con desviación típica grande (del orden de $9{,}43 \times 10^{45}$) y un intervalo de confianza aproximado al 95 % de

$$[\,2{,}27 \times 10^{44},\ 2{,}08 \times 10^{45}\,],$$

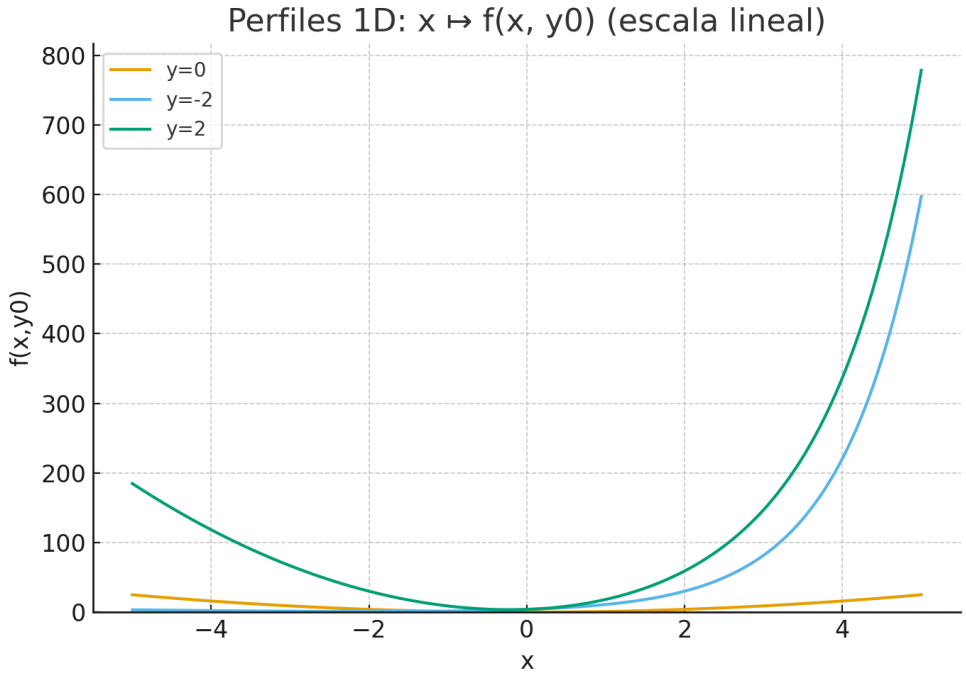

Figura 4: Cortes 1D: para $y_0 \in \{0, -2, 2\}$, se fija $y = y_0$ y se muestra $x \mapsto f(x, y_0)$ en $x \in [-5, 5]$. Sirven para ver cómo varía f cuando sólo cambia x.

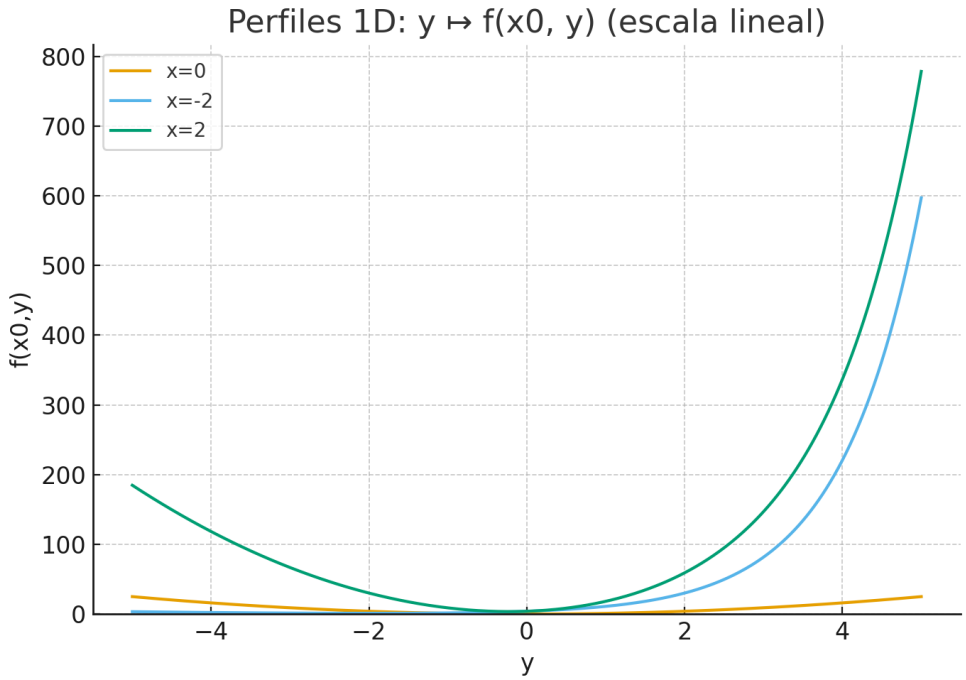

Figura 5: Cortes 1D: para $x_0 \in \{0, -2, 2\}$, se fija $x = x_0$ y se muestra $y \mapsto f(x_0, y)$ en $y \in [-5, 5]$. Complementan la vista anterior aislando la dependencia en y.

dentro del cual cae el valor exacto. La enorme variabilidad no es un error: responde a que al tomar x o y cerca de 100 las exponenciales aumentan f brutalmente, mientras que si ambos quedan muy negativos la función se vuelve microscópica.

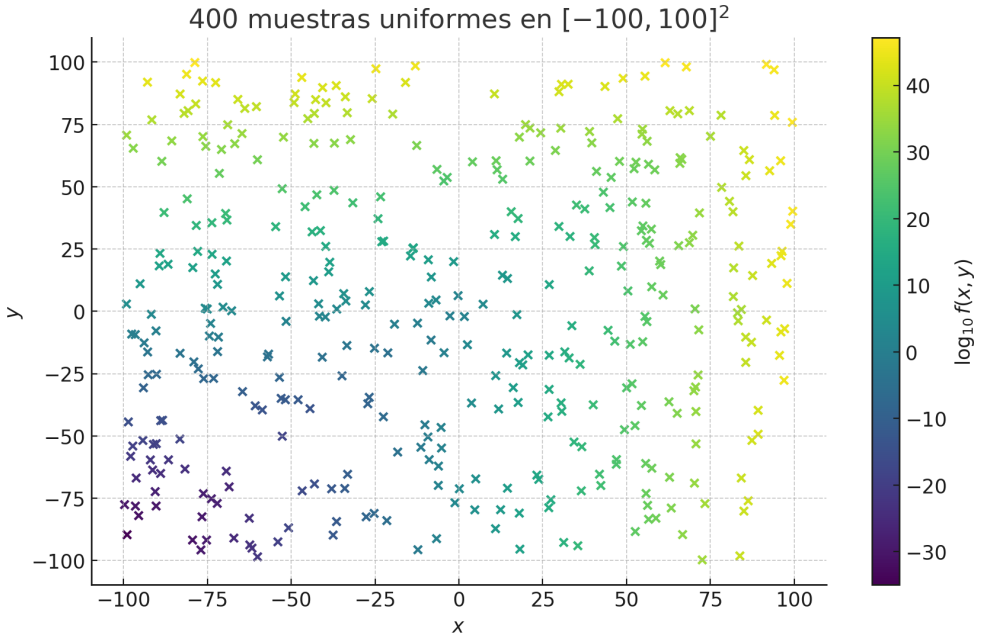

Figura 6: Muestras uniformes en $[-100, 100]^2$ coloreadas por $\log_{10} f(x, y)$. Se distinguen regiones con valores muy pequeños frente a otras con crecimiento pronunciado.

Las Figs. 6 y 7 contextualizan el cálculo: se observa la concentración de valores muy pequeños cuando ambas coordenadas son negativas, y la fuerte cola hacia valores grandes cuando alguna coordenada es positiva y de gran magnitud.

En estas visualizaciones se emplea una escala logarítmica base diez porque los valores de f abarcan muchos órdenes de magnitud: cuando alguna coordenada es positiva y grande, las exponenciales inflan f; cuando ambas son negativas y grandes en valor absoluto, f se acerca a cero. La transformación $z = \log_{10} f$ comprime esa amplitud y permite ver a la vez valores muy pequeños y muy grandes sin que la gráfica se sature. Elegir base diez facilita la lectura en términos de *órdenes de magnitud* (un incremento de 1 en z equivale a multiplicar f por 10). La base, de todos modos, es irrelevante para la forma cualitativa: $\log_{10} f = \dfrac{\ln f}{\ln 10}$ difiere de $\ln f$ sólo por un factor constante, de manera que cambiar de base reescala el eje pero no altera la estructura de los niveles ni el orden relativo de los puntos. La transformación es válida porque $f(x, y) \geq 0$ para todo (x, y); en los contornos cercanos al origen se añade un término pequeño (10^{-12}) únicamente para evitar la indeterminación en $\log(0)$ y así visualizar regiones donde f es prácticamente nula sin afectar la interpretación. En el histograma se utiliza un resguardo numérico análogo para esquivar subdesbordamientos computacionales, lo cual no modifica las conclusiones estadísticas y mantiene la lectura por órdenes de magnitud. Es importante notar que esta escala sólo se usa para *visualizar*; los criterios de parada y los métodos de optimización se aplican sobre f y sus derivadas, no sobre el valor transformado.

Quise ver cómo se comportan dos estrategias de optimización en este caso. La primera es el descenso por gradiente con búsqueda de paso, que se guía por la pendiente local: en cada punto mira hacia dónde baja más rápido y avanza un paso en esa dirección, ajustando ese paso con una regla simple: si al dar el paso sube, lo acorta hasta que baje. El método rinde bien cuando el gradiente está suficientemente alineado con la dirección hacia el mínimo; en cambio, en zonas donde el gradiente es muy pequeño o la curvatura es desfavorable, la dirección de descenso

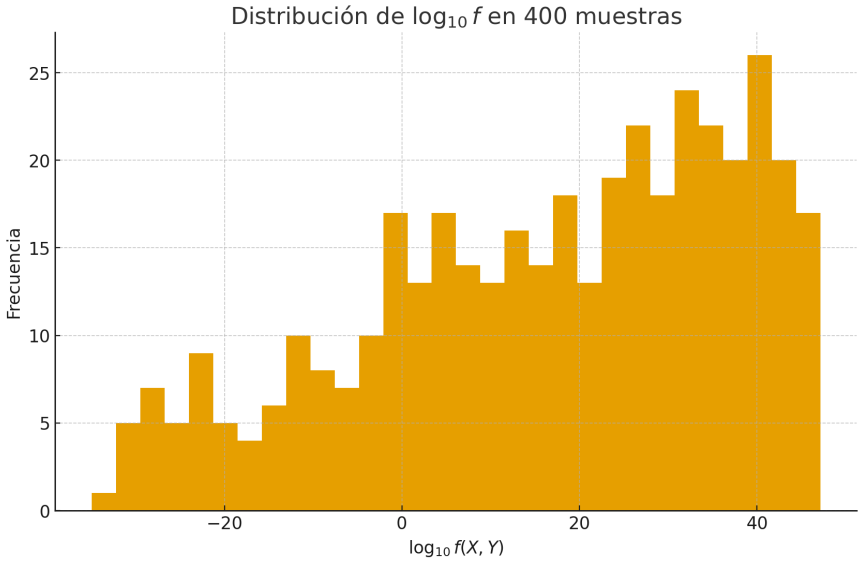

Figura 7: Histograma de $\log_{10} f$ para las 400 muestras. La asimetría refleja valores muy pequeños (ambas coordenadas negativas) y valores muy grandes (alguna coordenada positiva grande).

resulta poco eficaz y el progreso por iteración es limitado.

La segunda es una variante de Newton amortiguada: además del gradiente, utiliza la curvatura (Hessiano) para proponer un paso basado en un modelo cuadrático local. Como la función no es bien comportada en todo el dominio, se aplica una regularización al Hessiano para estabilizar la dirección y evitar pasos en direcciones problemáticas; además, se incluye una búsqueda de paso que reduce la longitud del movimiento si no se obtiene disminución. Este esquema suele rendir mejor porque, una vez dentro de la región cercana al mínimo, el modelo cuadrático aproxima bien a la función y el método alcanza la solución en pocas iteraciones

Para que la comparación sea clara, tomé 400 puntos iniciales (x_0, y_0) *al azar*, cada coordenada uniforme en $[-100, 100]$. En cada punto corrí ambos métodos, declarando "éxito" cuando la norma de la pendiente bajaba por debajo de 10^{-8}. Para evitar números imposibles (por ejemplo, intentar evaluar e^{800}), la búsqueda de paso recorta saltos que sacan las coordenadas demasiado lejos; si aún así un método se acerca a regiones numéricamente peligrosas, lo anoto como fallo por desborde. Puse un máximo de 3000 iteraciones para el descenso por gradiente y de 1000 para Newton amortiguado.

Observaciones del experimento

Lo que muestran los datos concuerda con el analisis previo. Newton amortiguado terminó con éxito en el 94.5 % de los casos (378/400). Si me quedo sólo con los que realmente se movieron (es decir, excluyo los que ya empezaban en una zona tan plana que la pendiente inicial era menor al umbral), el número típico de pasos fue doce, con un cuartil inferior de tres y un cuartil superior alrededor de veintiséis. Los pocos fallos de Newton se debieron casi siempre a intentar entrar en zonas de exponenciales gigantes (20 desbordes) y, en dos ocasiones, a agotar las iteraciones permitidas sin rebajar lo suficiente la pendiente.

En el otro extremo, el descenso por gradiente no funciono muy bien: sólo marcó 13.5 % de éxitos (54/400), y *todos* ellos fueron casos en los que ya desde el inicio la pendiente era tan diminuta que el método declaró "listo" sin dar ni un paso. Cada vez que tuvo que avanzar de verdad, se quedó corto y acabó agotando las 3000 iteraciones.

Lo más interesante es dónde terminan. Cuando empiezo en el primer cuadrante, por ejemplo en $(100, 100)$, Newton no busca el origen a cualquier precio: descubre que moverse hacia el *tercer*

cuadrante hace caer f muy rapido porque las exponenciales se hunden, y en unos cien pasos queda en una posicion muy baja, con un valor de la función del orden de $4{,}9 \times 10^{-19}$, concretamente en torno a $(-52{,}8, -200)$. El descenso por gradiente, por su parte, se queda patinando: después de 3000 intentos apenas logra un punto con $f \approx 7{,}6 \times 10^{-9}$ sin realmente consolidar la bajada, y no cumple el criterio de parada.

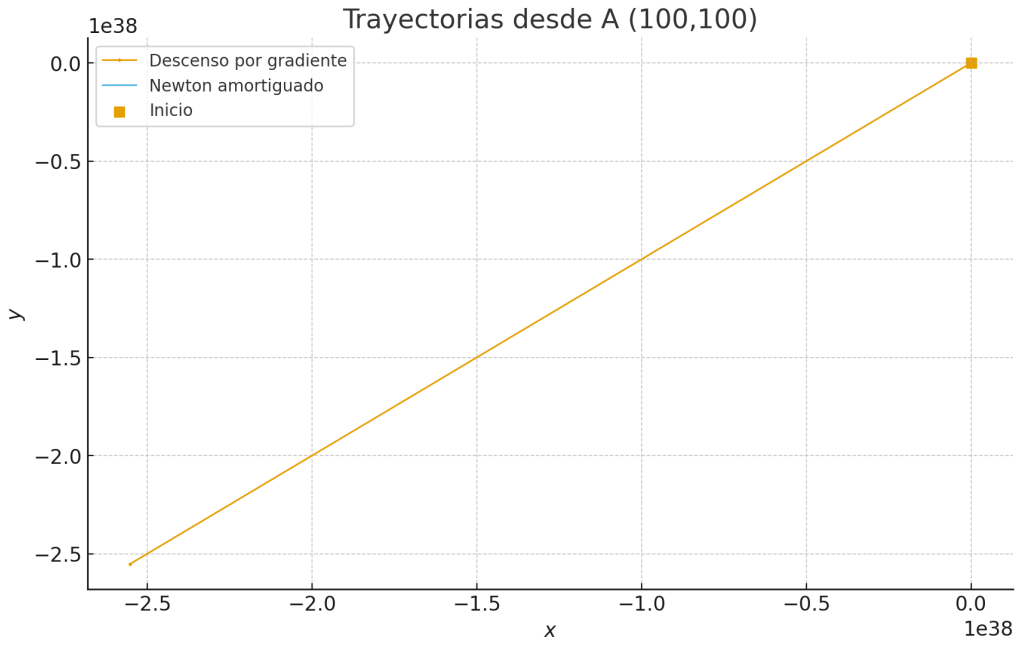

Figura 8: Trayectorias en el plano (x, y) desde $(100, 100)$ para descenso por gradiente y Newton amortiguado.

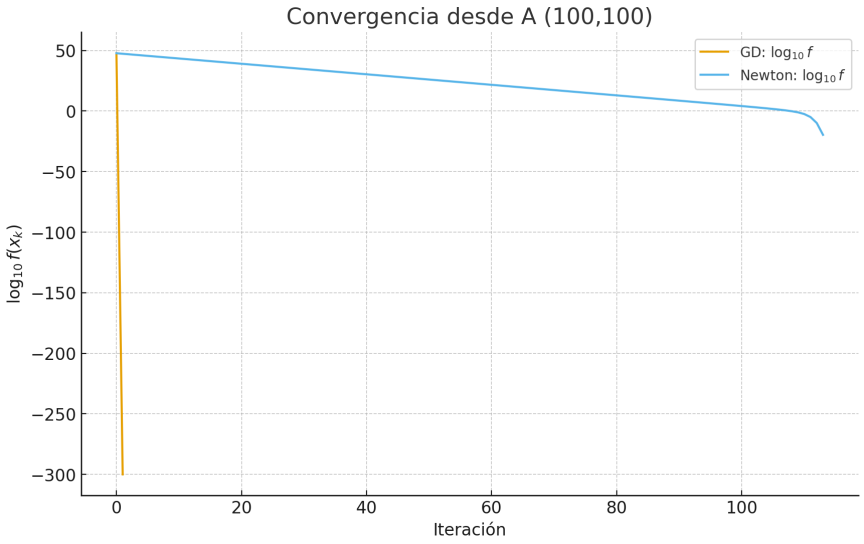

Figura 9: Convergencia de $f(x_k)$ (escala $\log_{10}$) desde $(100, 100)$.

Si el punto inicial es $(100, -100)$ o $(-100, 100)$, el método de Newton aprovecha la asimetría que introducen los términos e^x y e^y combinados con x^2 y y^2 y, en aproximadamente tres iteraciones, alcanza una región de valores extremadamente pequeños cerca de $(-191{,}3, -200)$, con $f \approx 3{,}4 \times 10^{-79}$. Si el inicio es $(-100, -100)$, tanto Newton como el criterio basado en la norma del gradiente se detienen de inmediato: se tiene $f \approx 7{,}4 \times 10^{-40}$ y el gradiente ya está por debajo

del umbral. Esto refleja la ausencia de coercividad: existen zonas muy alejadas del origen en las que f es prácticamente cero, de modo que el "mínimo práctico" se obtiene al desplazarse hacia coordenadas muy negativas y no exclusivamente en el origen.

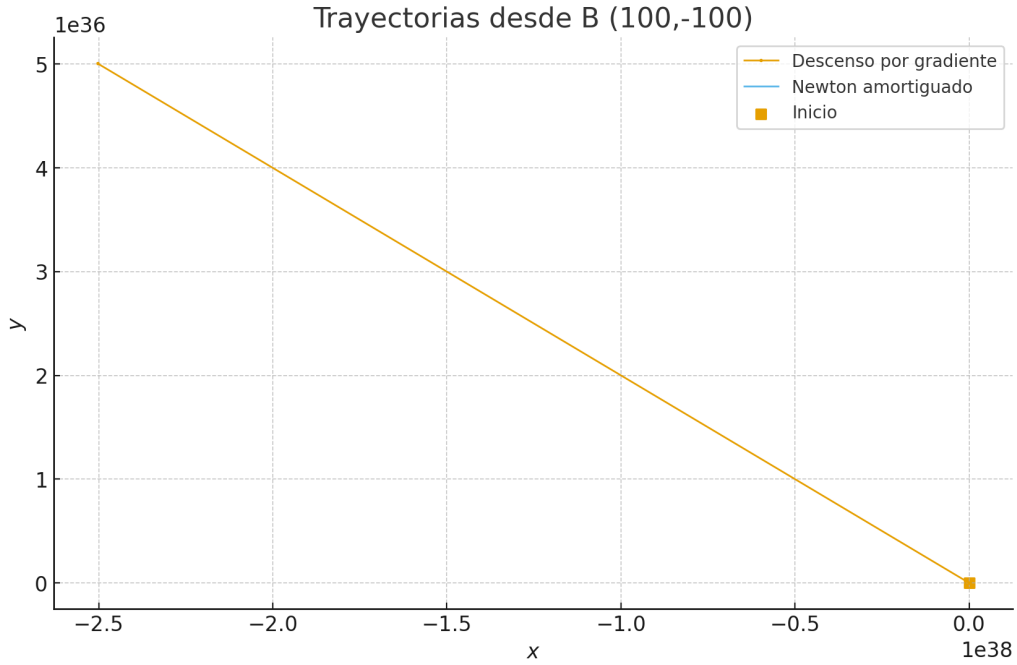

Figura 10: Trayectorias desde $(100, -100)$: Newton amortiguado alcanza rápidamente una región de f muy pequeña.

Las Figs. 8–13 ilustran las trayectorias y el ritmo de reducción de la función bajo ambos métodos en los distintos escenarios descritos.

Si clasifico los éxitos de Newton por cuadrantes, se ve el mismo patrón con otro ángulo. En el primer cuadrante, aproximadamente un cuarenta por ciento de las trayectorias terminan cayendo a esa meseta del tercer cuadrante con alguna coordenada por debajo de -150, mientras que cerca de un tercio sí logra alcanzar, al final, el entorno del origen. En el tercer cuadrante ocurre lo esperable: casi nadie se mueve hacia el origen porque la función ya es insignificante; el método necesita pocas iteraciones y la pendiente ya es desde un inicio diminuta.

Conclusion

Para esta función, es preferible utilizar Newton amortiguado. Desempeña mejor por dos motivos principales: primero, cerca del minimizador el modelo cuadrático que incorpora la curvatura (Hessiano) es preciso y conduce a una convergencia rápida en pocas iteraciones; segundo, en regiones donde el gradiente es muy pequeño (particularmente en el tercer cuadrante), el descenso por gradiente pierde eficacia, mientras que Newton, gracias a la regularización del Hessiano y a la búsqueda de paso, mantiene pasos efectivos y reduce la función con mayor rapidez. Aun así, conviene incluir salvaguardas numéricas: si el paso propuesto conduce a evaluaciones exponenciales muy grandes, se debe reducir la longitud del paso para evitar desbordamientos

En síntesis, f es suave, no negativa y alcanza su mínimo global en el origen. Sin embargo, al no ser coerciva, existen regiones alejadas hacia $(-\infty, -\infty)$ donde f toma valores cercanos a cero y el gradiente presenta magnitud muy reducida. Esta geometría explica tanto la elevada variabilidad del valor esperado al muestrear en $[-100, 100]^2$ como las diferencias observadas entre métodos. El esquema de Newton amortiguado con búsqueda de paso resulta más adecuado: aprovecha el modelo cuadrático local para converger con rapidez en la vecindad del mínimo y mantiene

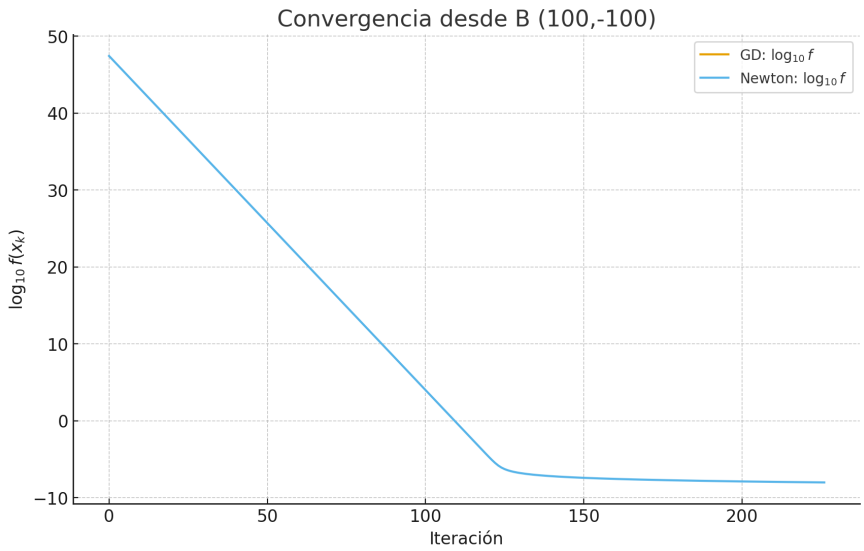

Figura 11: Convergencia de $f(x_k)$ (escala $\log_{10}$) desde $(100, -100)$.

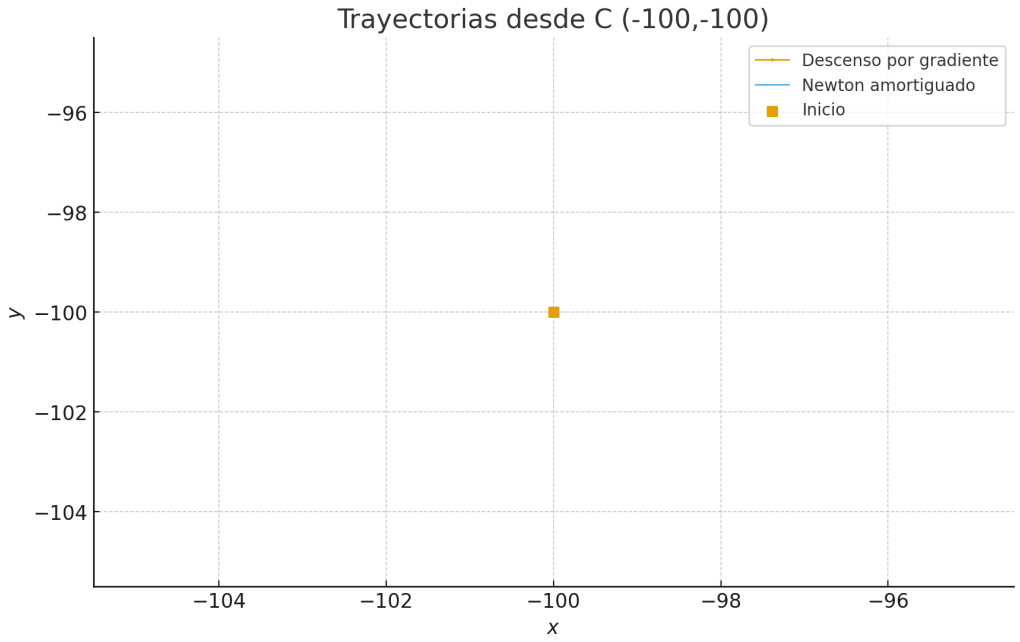

Figura 12: Trayectorias desde $(-100, -100)$: ambas estrategias se detienen casi de inmediato por gradiente diminuto.

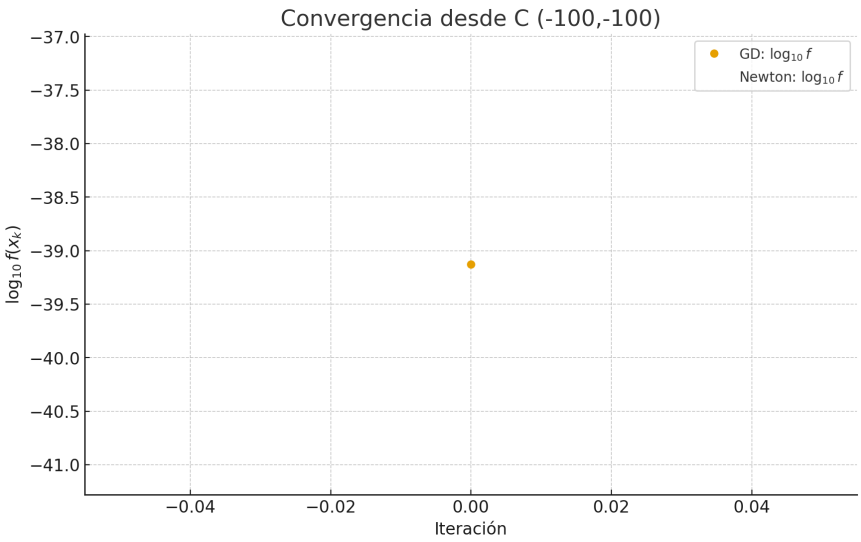

Figura 13: Convergencia de $f(x_k)$ (escala $\log_{10}$) desde $(-100, -100)$.

eficacia en zonas de gradiente pequeño. Por contraste, el descenso por gradiente sólo muestra buen desempeño cuando la condición inicial ya satisface de facto el criterio de parada.

Con el objetivo de hacer mas comprensible la interpretacion de los datos obtenidos y la comparacion realizada entre los algoritmos utilice un mapa de Basins y el metodo de las cajas Un mapa de basins (o mapa de destinos) colorea el plano de inicios según el *punto al que termina* un algoritmo al optimizar. Cada celda representa un punto de partida y su color indica la *clase de destino*: fallo (-1), final "otro" (0), vecindad del origen (1) o meseta negativa (2).

En este problema, el mapa permite *ver la geometría práctica* que encuentra Newton amortiguado: la zona verde marca inicios que *convergen al origen*, mientras que las manchas amarillas son *trayectorias que se van* a regiones muy negativas donde f es casi nula. El bloque morado en el tercer cuadrante corresponde a *fallos de paso* en zonas superplanas: la condición de descenso no acepta pasos, aunque f ya sea diminuta.

En resumen, el mapa cuenta *qué regiones del plano "pertenecen" a cada destino* y deja ver *sensibilidades del algoritmo*: cómo la búsqueda de paso se vuelve exigente en áreas planas y cómo la curvatura puede empujar hacia la meseta lejos del origen.

Cajas comparativas (GD vs. Newton amortiguado):

Las tres cajas comparan, para los dos métodos, (i) el número de iteraciones k, (ii) la longitud total del recorrido (`path_len`) y (iii) el mayor valor absoluto visitado en coordenadas (`max_abs`). La línea central es la mediana; la caja va de P25 a P75; los puntos fuera son atípicos.

Lectura corta: en k, GD suele quedarse en 0 (se "declara listo" sin moverse en zonas planas), mientras que Newton recorre más y, en varios casos, pega en el límite de iteraciones. En `path_len`, GD tiende a 0 por lo mismo; Newton muestra recorridos efectivos. En `max_abs` ambos exploran rangos parecidos en términos de magnitud, con más dispersión cuando GD intenta moverse en zonas complicadas.

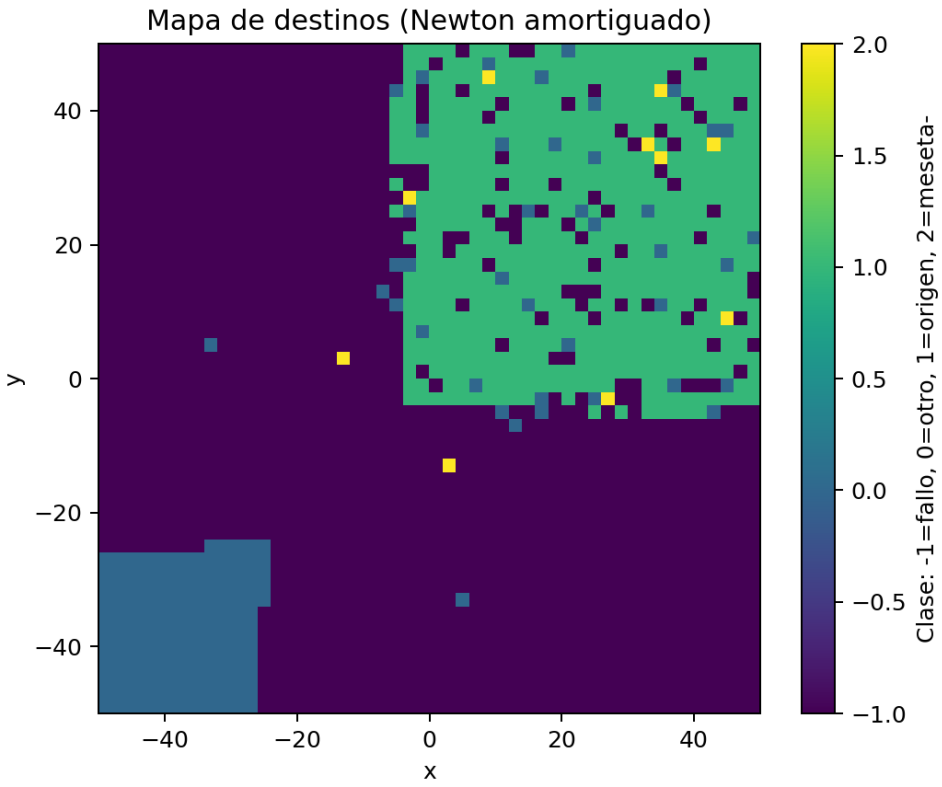

Figura 14: Mapa de basins (Newton amortiguado) en la ventana $[-50, 50]^2$. La barra de color codifica: -1 fallo, 0 otro, 1 origen, 2 meseta negativa.

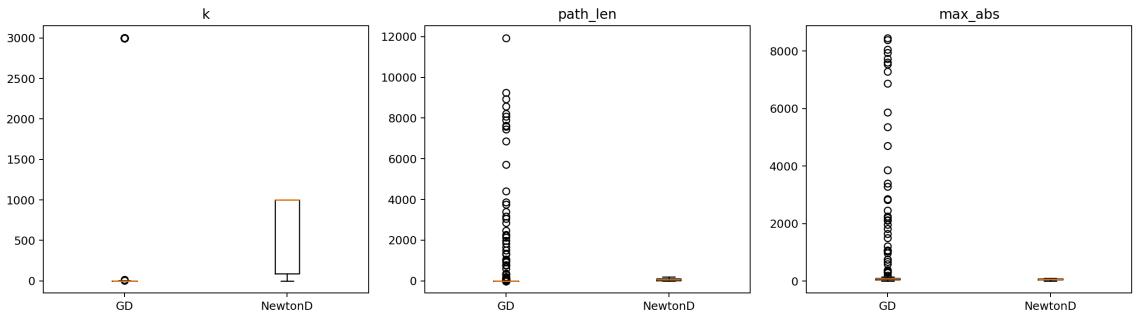

Figura 15: Boxplots de k, `path_len` y `max_abs` para GD y Newton amortiguado.