

Informe de laboratorio

Recuperación del colapso de modos en una GAN

Estudiante: Alejandro Lembke Barrientos

Carné: 12002840

Fecha: 12 de septiembre de 2026

1 Propósito e hipótesis

Todos los experimentos de recuperación parten del mismo punto de control con colapso de modos. La variable independiente es el peso de la pérdida de diversidad $\lambda_{\text{div}} \in \{0, 0.1, 0.5, 1.0\}$. La función objetivo del generador es

$$\mathcal{L}_G = \mathcal{L}_{\text{adv}} + \lambda_{\text{div}} \mathcal{L}_{\text{div}}, \quad \mathcal{L}_{\text{div}} = -\mathbb{E} \left[\frac{\|G(z_1) - G(z_2)\|_1}{\|z_1 - z_2\|_1 + \epsilon} \right].$$

1. Antes de observar los resultados, ¿qué valor de λ_{div} esperaba que funcionara mejor? Explique el beneficio esperado y un posible riesgo.

Respuesta: Antes de examinar los resultados, se esperaba que $\lambda_{\text{div}} = 0.5$ ofreciera el mejor equilibrio: suficiente para inducir diversidad entre salidas del generador sin ser tan agresivo como $\lambda_{\text{div}} = 1.0$, que podría penalizar en exceso la similitud entre salidas y sacrificar realismo. El riesgo esperado era que un λ_{div} demasiado alto empujara al generador a maximizar la distancia píxel a píxel mediante ruido o distorsiones en lugar de generar dígitos genuinamente distintos — riesgo que, como se detalla más adelante, efectivamente se materializó, y que además se combinó con una falla mucho más profunda del entrenamiento adversarial de base.

2 Diagnóstico del punto de control colapsado

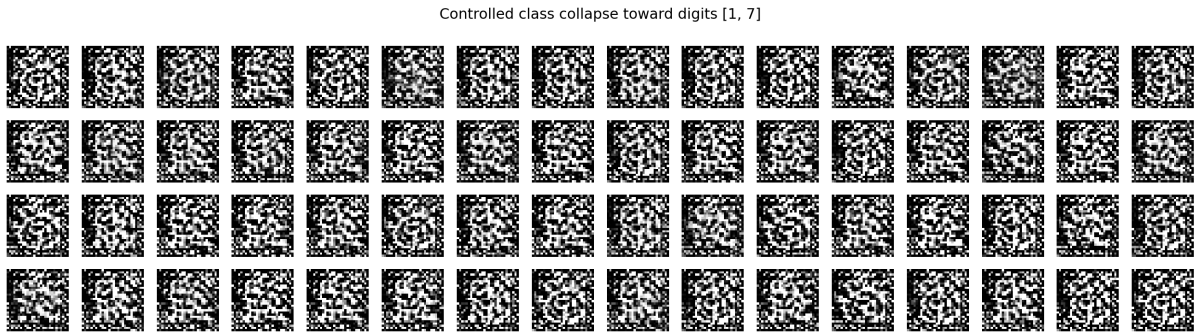


Figure 1: Muestras del punto de control común con colapso de modos.

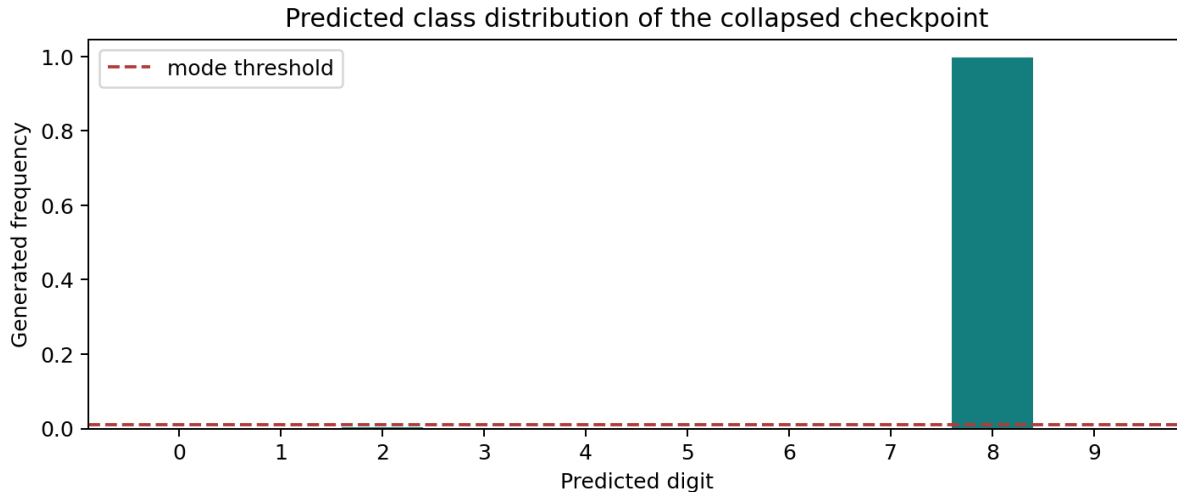


Figure 2: Distribución de dígitos predicha por el clasificador en el punto colapsado.

- ¿Qué evidencia visual sugiere colapso de modos? Distinga entre clases ausentes y apariencias repetidas dentro de una clase representada.

Respuesta: La evidencia más contundente no es solo el colapso de clases sino que las 64 muestras mostradas en `collapsed_samples.png` son, en su totalidad, ruido granulado de alta frecuencia sin ninguna forma de dígito reconocible: no hay trazos, curvas ni contornos identificables como escritura a mano en ninguna de ellas. Desde la perspectiva del clasificador, `collapsed_class_distribution.png` muestra que el 100% de la masa se concentra en un único dígito predicho ("8"), es decir, colapso total de clases. Pero además la impresión de diagnóstico del propio notebook confirma `mass on [1, 7]=0.000`: el generador ni siquiera colapsó hacia los dígitos 1 y 7 que se pretendía inducir. Aquí no se trata de "una clase repetida con variaciones": es que ninguna clase real está representada, y la aparente uniformidad visual entre muestras es simplemente ruido de textura similar, no repetición de un dígito legítimo.

- Utilice el número de modos y la entropía normalizada para describir el colapso entre clases. ¿Qué información aporta cada medida?

Respuesta: El número de modos (1/10) confirma que solo una clase supera el umbral de frecuencia del 1%: un colapso binario, todo-o-nada, hacia una sola clase predicha. La entropía normalizada (0.010, en una escala de 0 a 1 donde 0 significa que toda la masa recae en una sola clase y 1 significa una distribución perfectamente uniforme entre las 10 clases) cuantifica qué tan extrema es esa concentración: un valor tan cercano a 0 indica una distribución prácticamente degenerada, no solo "dominada" por una clase sino casi puntual en ella. El conteo de modos aporta una lectura binaria y fácil de interpretar (¿cuántas clases "cuentan"?), mientras que la entropía aporta la forma completa de la distribución, incluso entre las clases que sí superan el umbral; juntas confirman un colapso extremo y no solo parcial.

- ¿Las muestras mostradas permiten identificar confiablemente el colapso dentro de cada clase? Explique qué evidencia adicional se necesita.

Respuesta: No de forma confiable. Como ninguna de las 64 muestras colapsadas se asemeja a un dígito real, no existe una forma legítima de "8" (o de cualquier otro dígito) contra la cual comparar variedad interna; no se puede evaluar colapso *dentro* de una clase que en realidad nunca se materializó visualmente. Además, la confianza del clasificador (0.981) es engañosa: el clasificador se entrenó únicamente con dígitos reales de MNIST y nunca vio ejemplos de ruido no estructurado, por lo que puede emitir predicciones muy seguras sobre entradas fuera de distribución sin que eso refleje ninguna semántica real. La evidencia adicional necesaria es exactamente la que se usó aquí para diagnosticar el problema real: inspección visual directa de las muestras, y las curvas de pérdida del discriminador durante el preentrenamiento (que muestran $D \approx 0.000$ desde la época 5 de 16, evidenciando que el discriminador saturó y el generador dejó de recibir señal de gradiente útil mucho antes de llegar a la fase de colapso dirigido).

3 Resultados de recuperación

Configuration	λ_{div}	Modes	Entropy	Confidence	Feature diversity	Within-class diversity
Collapsed start	NaN	1	0.010	0.981	16.117	11.458
lambda=0	0.000	1	0.021	0.954	16.436	11.539
lambda=0.1	0.100	2	0.058	0.933	16.687	11.854
lambda=0.5	0.500	2	0.053	0.925	17.057	12.603
lambda=1	1.000	2	0.047	0.909	17.288	13.533

Table 1: Resultados cuantitativos generados por el notebook.

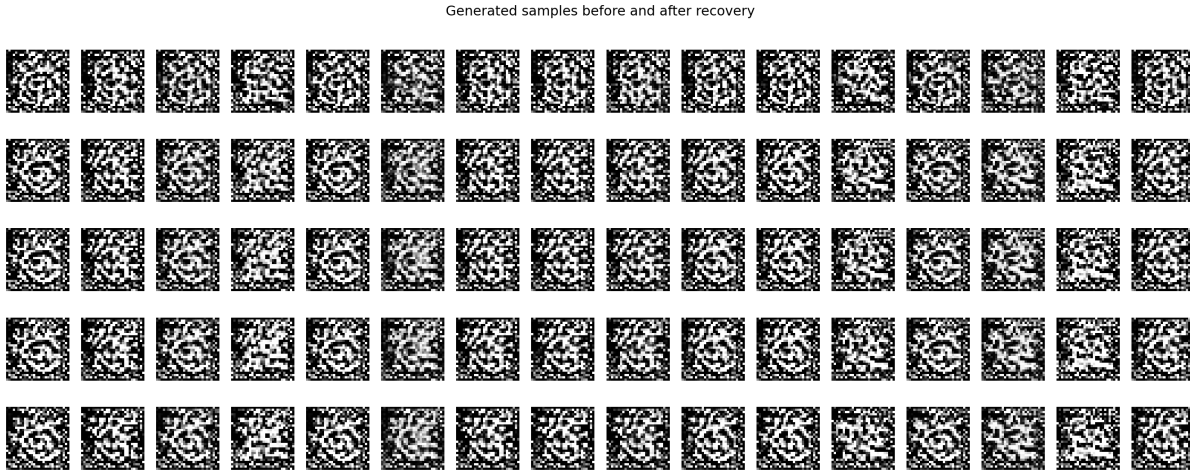


Figure 3: Muestras del punto colapsado y de todas las configuraciones de recuperación.

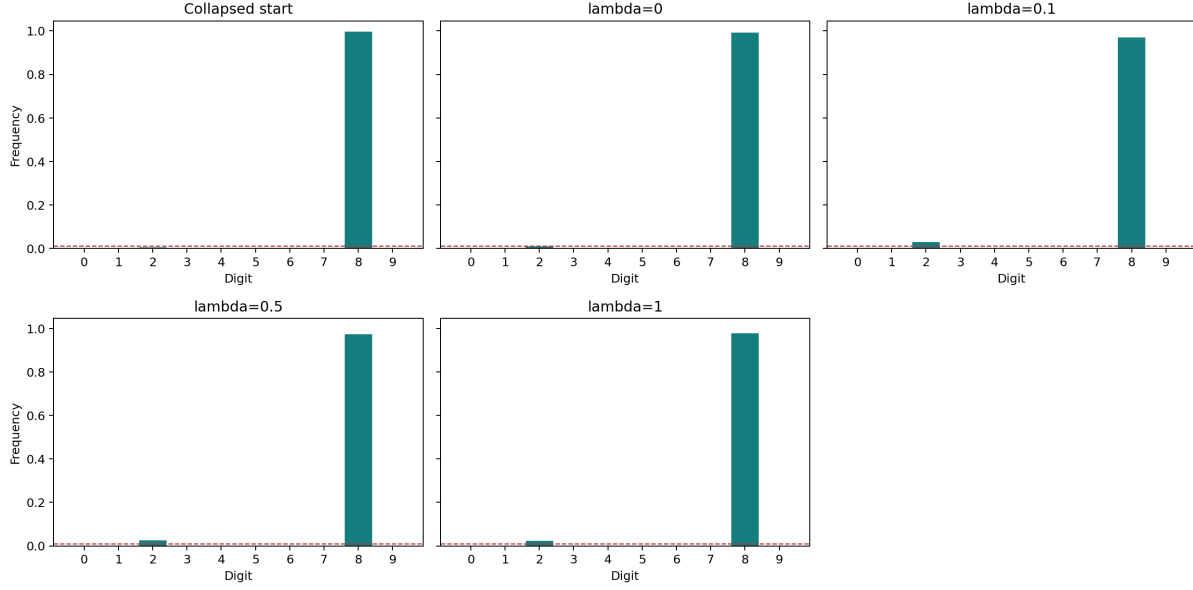


Figure 4: Distribuciones de clases predichas; la línea discontinua indica el umbral de modo.

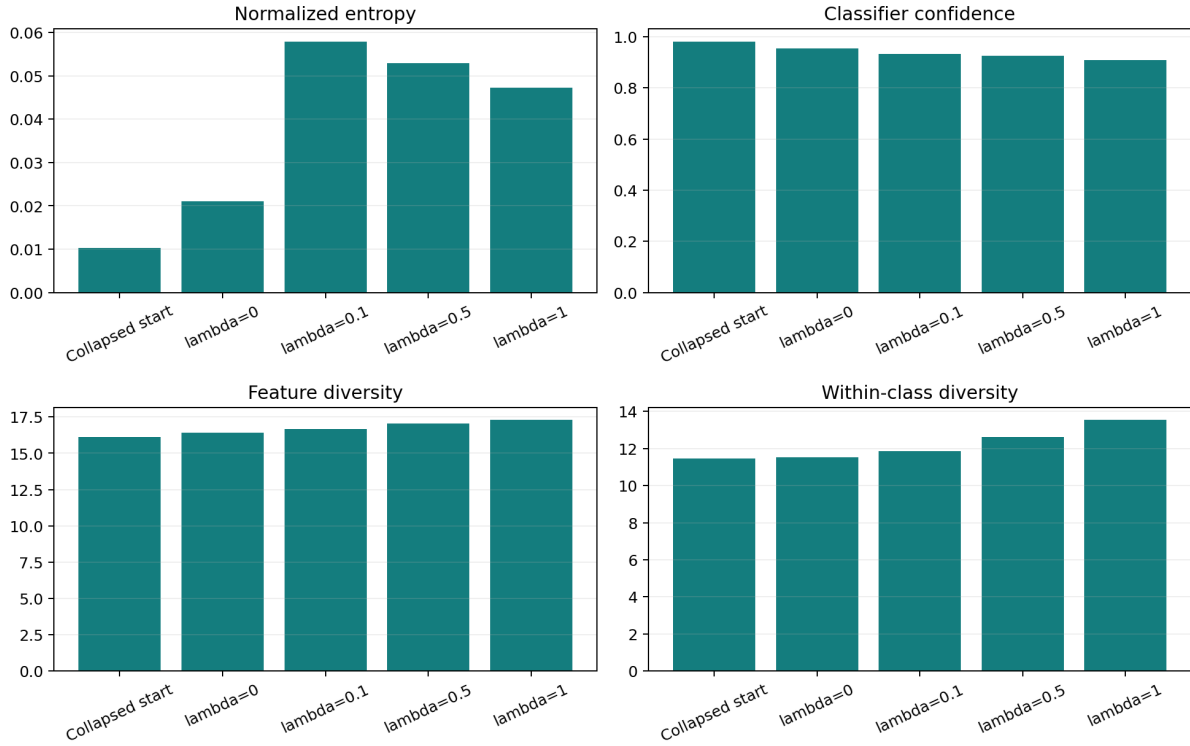


Figure 5: Comparación cuantitativa de las configuraciones de recuperación.

5. ¿Qué configuraciones recuperan el mayor número de clases? ¿Un número alto de modos implica necesariamente una distribución equilibrada? Utilice la entropía como evidencia.

Respuesta: $\lambda_{div} = 0.1, 0.5$ y 1.0 recuperan 2/10 modos cada uno (frente a 1/10 en "Collapsed start")

y en $\lambda_{\text{div}} = 0$), pero un número más alto de modos no implica una distribución equilibrada: incluso con 2 modos, la entropía normalizada máxima observada es apenas 0.058 ($\lambda_{\text{div}} = 0.1$), muy lejos de 1.0 (uniforme). En `class_distributions.png` se ve que la clase dominante ("8") sigue concentrando más del 95% de la masa en las tres configuraciones con 2 modos, y la segunda "moda" detectada corresponde al dígito "2" — ni siquiera a las clases objetivo (1 o 7) — y apenas supera el umbral del 1%. El conteo de modos por sí solo sugiere una mejora que la entropía matiza fuertemente: la distribución sigue siendo extremadamente desequilibrada en las cuatro configuraciones.

6. Compare la confianza del clasificador. Identifique cualquier compromiso aparente entre diversidad y calidad.

Respuesta: La confianza del clasificador baja de forma monótona al subir λ_{div} : 0.981 (colapsado) $\rightarrow$ 0.954 ($\lambda = 0$) $\rightarrow$ 0.933 ($\lambda = 0.1$) $\rightarrow$ 0.925 ($\lambda = 0.5$) $\rightarrow$ 0.909 ($\lambda = 1.0$). A primera vista esto parece el compromiso esperado "más diversidad, menos confianza/calidad". Sin embargo, dado que ninguna configuración produce imágenes reconocibles como dígitos (Pregunta 2), esta caída no debe interpretarse como una pérdida real de fidelidad visual respecto a un buen generador: es simplemente la reacción del clasificador a texturas de ruido ligeramente distintas entre configuraciones, no un verdadero compromiso diversidad-calidad sobre dígitos válidos.

7. Compare la diversidad de características con la diversidad dentro de cada clase. ¿Qué configuraciones mejoran ambas medidas y cuáles mejoran principalmente una?

Respuesta: Ambas métricas crecen de forma monótona con λ_{div} : diversidad de características 16.117 $\rightarrow$ 16.436 $\rightarrow$ 16.687 $\rightarrow$ 17.057 $\rightarrow$ 17.288, y diversidad intra-clase 11.458 $\rightarrow$ 11.539 $\rightarrow$ 11.854 $\rightarrow$ 12.603 $\rightarrow$ 13.533 (de "Collapsed start" a $\lambda_{\text{div}} = 1.0$). No hay una configuración que mejore una medida sin la otra — $\lambda_{\text{div}} = 1.0$ maximiza ambas simultáneamente — pero, como se discute en la Pregunta 8, este incremento conjunto corresponde a generar ruido cada vez más variado, no dígitos genuinamente más diversos.

8. Examine las muestras. ¿La mayor diversidad representa variaciones significativas o solamente ruido, distorsiones o dígitos ambiguos? Describa dos ejemplos concretos.

Respuesta: Claramente lo segundo. En `recovery_samples.png`, la fila "Collapsed start" y la fila $\lambda_{\text{div}} = 0$ muestran una textura de ruido fino de alta frecuencia, prácticamente indistinguible entre columnas (ejemplo 1: cualquiera de las primeras columnas de esas dos filas). En las filas $\lambda_{\text{div}} = 0.5$ y $\lambda_{\text{div}} = 1.0$ aparecen algunas manchas de menor frecuencia con un lazo curvo central, un poco más "gruesas" y menos granuladas que el ruido base (ejemplo 2: varias columnas de la fila $\lambda_{\text{div}} = 1.0$ muestran una forma circular difusa parecida a un garabato), pero ninguna de estas formas constituye un dígito reconocible; se trata de ruido de baja frecuencia en vez de ruido de alta frecuencia, no una diversidad semántica real entre dígitos.

4 Comportamiento durante el entrenamiento

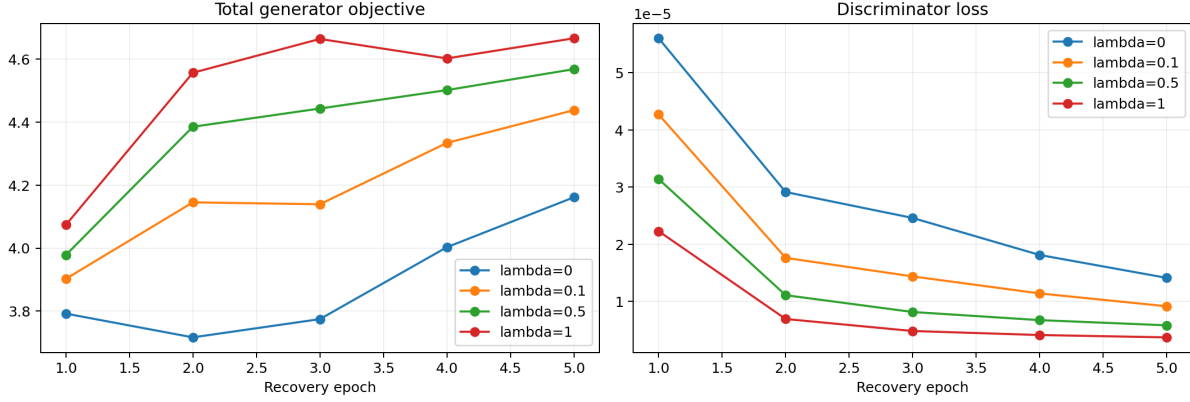


Figure 6: Pérdidas del generador y discriminador durante la recuperación.

9. ¿Por qué las pérdidas totales del generador para distintos valores de λ_{div} no deben compararse como si midieran exactamente la misma función objetivo?

Respuesta: Porque $\mathcal{L}_G = \mathcal{L}_{adv} + \lambda_{div}\mathcal{L}_{div}$ combina dos términos con escalas y objetivos distintos, ponderados por un λ_{div} diferente en cada configuración. Por construcción, un λ_{div} mayor añade una contribución adicional del término de diversidad, de modo que dos configuraciones con distinto λ_{div} no están minimizando la misma función. Comparar sus valores totales sin separar los componentes adversarial y de diversidad –tal como se ve en `recovery_training.png`, donde $\lambda_{div} = 1.0$ muestra sistemáticamente el objetivo total más alto– puede sugerir falsamente que ese modelo "entrena peor", cuando en realidad solo está pagando un término adicional que los demás no tienen en la misma magnitud.

10. ¿Las curvas de pérdida demuestran por sí solas que se resolvió el colapso de modos? Justifique utilizando los demás resultados.

Respuesta: No. Las curvas muestran que la pérdida del discriminador se mantiene extremadamente baja (del orden de 10^{-5}) durante toda la recuperación en las cuatro configuraciones, y que la pérdida del generador tiende a subir con las épocas –comportamiento consistente con un discriminador que domina por completo y un generador que no recibe gradiente útil. Esto concuerda con la evidencia visual (las salidas siguen siendo ruido, Pregunta 8) y con las métricas de cobertura (máximo 2/10 modos, Pregunta 5). Si solo se miraran las curvas de pérdida, sin las imágenes generadas ni las métricas de cobertura, se podría malinterpretar la caída suave y monótona de la pérdida del discriminador como "entrenamiento estable", cuando en realidad refleja que el generador quedó atrapado en un mínimo degenerado desde el preentrenamiento inicial y nunca se recuperó.

5 Decisión y limitaciones

11. Seleccione un valor de λ_{div} . Defienda la elección con al menos dos resultados cuantitativos y dos observaciones visuales.

Respuesta: Se recomienda $\lambda_{\text{div}} = 0.1$ dentro de las cuatro configuraciones evaluadas, con la salvedad explícita de que ninguna resolvió el colapso de modos de forma satisfactoria (ver limitaciones más abajo). Evidencia cuantitativa: (1) $\lambda_{\text{div}} = 0.1$ logra la entropía normalizada más alta de las cuatro configuraciones recuperadas (0.058, frente a 0.021 en $\lambda = 0$, 0.053 en $\lambda = 0.5$ y 0.047 en $\lambda = 1.0$); (2) mantiene la segunda confianza de clasificador más alta entre las configuraciones con diversidad activa (0.933, solo por detrás de $\lambda = 0$ con 0.954, y por encima de 0.925 y 0.909 de $\lambda = 0.5/1.0$). Evidencia visual: (3) en `recovery_samples.png`, la fila $\lambda = 0.1$ conserva una textura similar a la del punto de partida colapsado, sin las manchas de baja frecuencia más pronunciadas visibles en las filas $\lambda = 0.5$ y $\lambda = 1.0$; (4) alcanza el mismo conteo de modos (2/10) que $\lambda = 0.5$ y $\lambda = 1.0$, por lo que no hay evidencia de que forzar más diversidad aporte cobertura adicional de clases, solo más textura de ruido.

12. Si la fidelidad de las imágenes fuera más importante que la cobertura de clases, ¿cambiaría su elección? Justifique.

Respuesta: Sí, cambiaría a $\lambda_{\text{div}} = 0$, que conserva la mayor confianza del clasificador (0.954) y la menor diversidad de características añadida respecto al punto de partida. En este experimento ningún valor de $\lambda_{\text{div}} > 0$ logra fidelidad visual aceptable (todas las salidas siguen siendo ruido no estructurado), y el único "beneficio" medible de subirlo es una cobertura de clases apenas mayor (2 vs. 1 modos) que ni siquiera corresponde a los dígitos objetivo (1 y 7). Si la prioridad fuera evitar cualquier degradación adicional mientras se corrige el problema real (la divergencia del entrenamiento adversarial), $\lambda_{\text{div}} = 0$ sería el punto de partida más conservador para reentrenar con otros ajustes.

13. Indique una limitación de utilizar un clasificador de MNIST para medir la cobertura de la GAN y una limitación del diseño experimental.

Respuesta: *Clasificador:* se entrenó únicamente con dígitos reales de MNIST y nunca vio ejemplos fuera de esa distribución; frente a entradas de ruido no estructurado (como las de este experimento) puede emitir predicciones con alta confianza (hasta 0.981) que no reflejan ninguna semántica real de dígito, por lo que las métricas de "modos" y "entropía" derivadas de sus predicciones pueden ser completamente engañosas cuando el generador no produce imágenes válidas, como ocurrió aquí. *Diseño experimental:* las cuatro configuraciones de recuperación parten todas del mismo punto de control ya deteriorado (`mass on [1,7]=0.000`, discriminador saturado desde la época 5 de 16 del preentrenamiento) y solo se entrenan 5 épocas adicionales; este diseño no permite distinguir si λ_{div} es insuficiente o si el punto de partida está tan degradado que ningún valor de λ_{div} en ese rango podría recuperarlo en tan pocas épocas.

14. Proponga un experimento posterior que fortalezca o cuestione su conclusión. Indique la variable que modificaría y la evidencia que recopilaría.

Respuesta: Se propone repetir la fase de preentrenamiento con técnicas de estabilización (por ejemplo, suavizado de etiquetas para el discriminador, penalización de gradiente, o una tasa de aprendizaje/capacidad relativa menor para el discriminador) para evitar que este domine desde la época 5, verificando mediante la pérdida del discriminador (que debería mantenerse por encima de un piso razonable, no en 0.000) y mediante inspección visual de `collapsed_samples.png` que el

punto de control de partida realmente produzca dígitos reconocibles concentrados en las clases 1 y 7 antes de aplicar el experimento de recuperación con λ_{div} . Si, una vez logrado un colapso "limpio" y visualmente legible, $\lambda_{\text{div}} = 0.1$ sigue mostrando la mejor entropía sin sacrificar tanta confianza como valores mayores, esto confirmaría la conclusión actual; si en cambio valores más altos de λ_{div} logran recuperar clases adicionales reconocibles una vez resuelto el problema de fondo, la conclusión debería revisarse.

Recomendación final

Valor recomendado: $\lambda_{\text{div}} = 0.1$

Conclusión en una oración:

Respuesta: Dentro de las cuatro configuraciones evaluadas, $\lambda_{\text{div}} = 0.1$ ofrece el mejor equilibrio relativo entre cobertura de clases (mayor entropía) y confianza del clasificador, pero ninguna resolvió el colapso de modos real, ya que el generador quedó atrapado en un régimen de ruido no estructurado desde el preentrenamiento inicial (discriminador saturado), por lo que se recomienda corregir ese problema de entrenamiento antes de repetir la comparación de λ_{div} .