

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 un buen equilibrio: suficiente para incentivar diversidad entre las muestras del generador sin ser tan agresivo como $\lambda_{\text{div}} = 1.0$, que podría distorsionar excesivamente las imágenes en su intento de maximizar la distancia entre pares generados. El riesgo esperado era que un λ_{div} demasiado alto sacrificara realismo visual a cambio de diversidad artificial (por ejemplo, añadiendo ruido o trazos distorsionados en lugar de dígitos genuinamente distintos).

2 Diagnóstico del punto de control colapsado

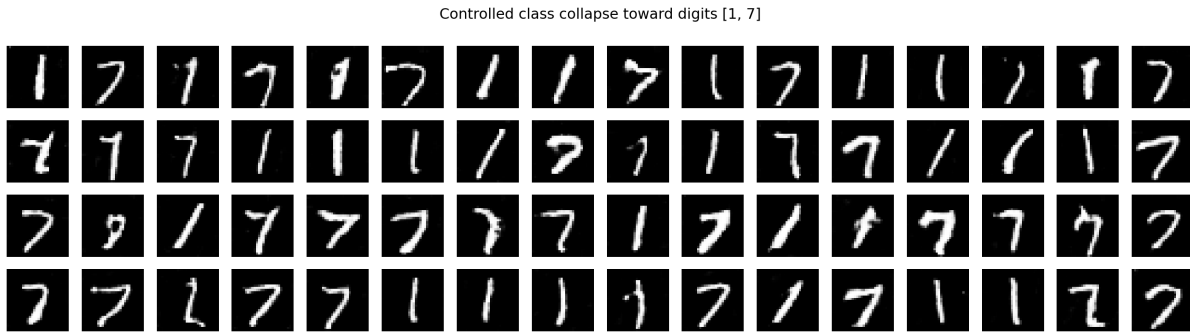


Figure 1: Muestras del punto de control común con colapso de modos.

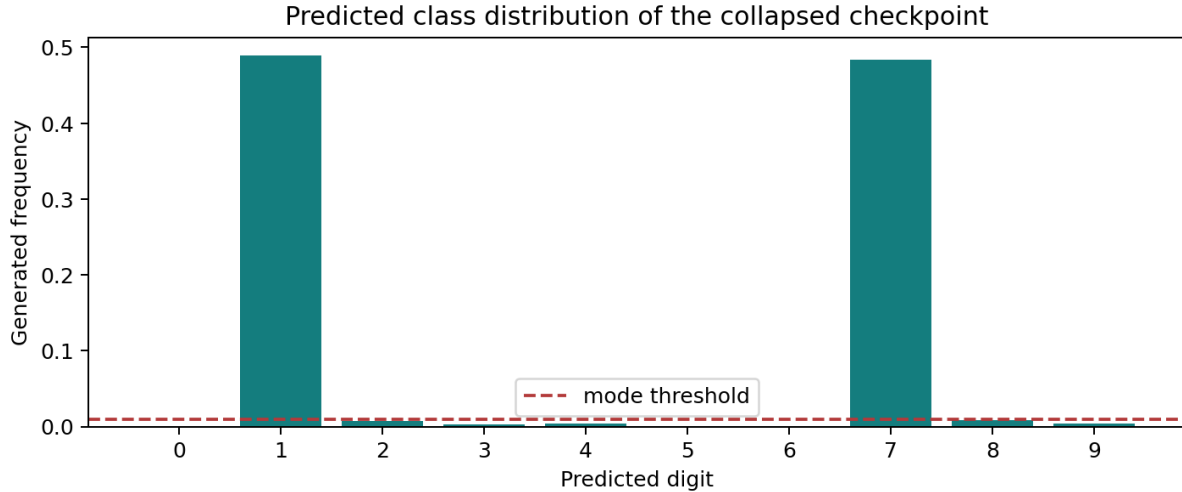


Figure 2: Distribución de dígitos predicha por el clasificador en el punto colapsado.

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

Respuesta: `collapsed_samples.png` muestra 64 imágenes, todas dígitos "1" o "7" perfectamente legibles y con variedad de trazo (rectos, inclinados, con o sin serifa horizontal); no aparece ningún otro dígito. `collapsed_class_distribution.png` confirma cuantitativamente que solo esas dos clases superan el umbral del 1% (~49% y ~48.5% de la masa respectivamente), con el resto prácticamente en cero. La evidencia distingue ambos fenómenos con claridad: hay 8 clases completamente ausentes (0, 2, 3, 4, 5, 6, 8, 9 por debajo del umbral), mientras que dentro de las dos clases representadas sí se observa variedad de apariencia (distintos estilos de trazo para el "1" y para el "7"); es decir, el colapso es entre clases, no una repetición monótona de una única imagen.

3. 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 (2/10) confirma que exactamente dos clases superan el umbral del 1%, coincidiendo con los dígitos objetivo del colapso (1 y 7). La entropía normalizada (0.367) cuantifica además qué tan pareja es la masa entre esas clases activas: un valor tan por debajo de 1.0 refleja que, aunque hay 2 modos, la distribución global sigue muy alejada de la uniformidad sobre las 10 clases, ya que la masa está casi enteramente concentrada en ese subconjunto (aproximadamente 50/50 entre 1 y 7), dejando 8 clases sin representación. El conteo de modos aporta una lectura discreta de cobertura; la entropía aporta la forma continua de toda la distribución, incluyendo cuán desbalanceada está incluso entre las clases activas.

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

Respuesta: Sí, de forma razonablemente confiable: las imágenes en `collapsed_samples.png` son dígitos "1" y "7" claramente legibles, con variación visible de estilo (inclinación, presencia de barra

horizontal en el "7", grosor del trazo), por lo que se puede juzgar directamente si hay diversidad dentro de cada clase representada. Como evidencia adicional, la diversidad intra-clase cuantitativa del punto colapsado (25.954) es considerablemente menor que la alcanzada tras la recuperación (~37–38), lo que corrobora numéricamente que la variedad dentro de esas dos clases también era más limitada en el punto de partida, y no es solo una impresión visual.

3 Resultados de recuperación

Configuration	λ_{div}	Modes	Entropy	Confidence	Feature diversity	Within-class diversity
Collapsed start	NaN	2	0.367	0.971	32.059	25.954
lambda=0	0.000	10	0.984	0.912	50.957	37.816
lambda=0.1	0.100	10	0.985	0.909	50.757	37.949
lambda=0.5	0.500	10	0.987	0.911	50.256	37.341
lambda=1	1.000	10	0.988	0.912	50.829	38.141

Table 1: Resultados cuantitativos generados por el notebook.

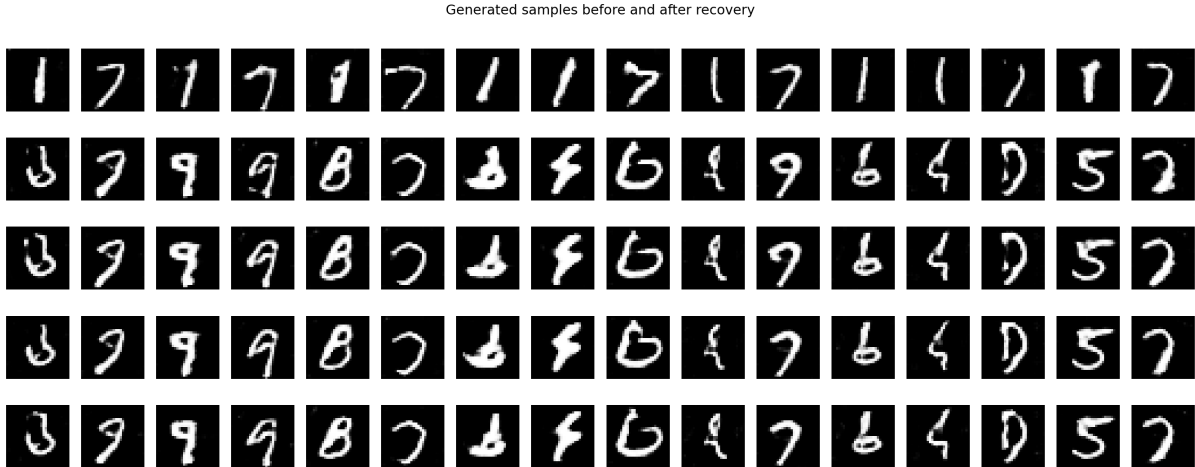


Figure 3: Muestras del punto colapsado y de todas las configuraciones de recuperación.

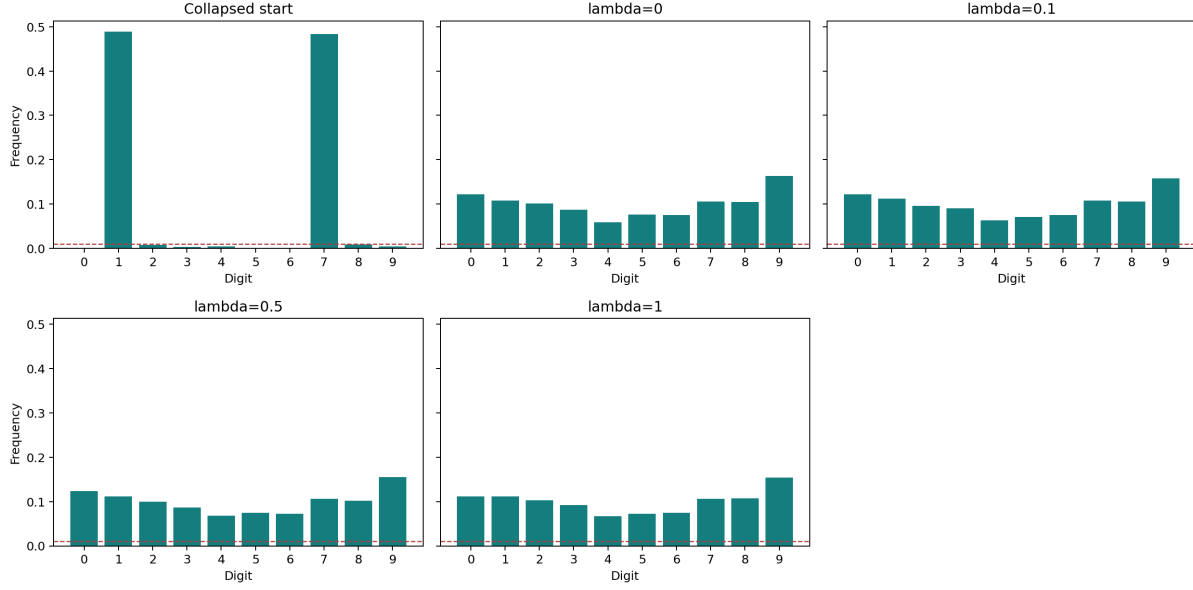


Figure 4: Distribuciones de clases predichas; la línea discontinua indica el umbral de modo.

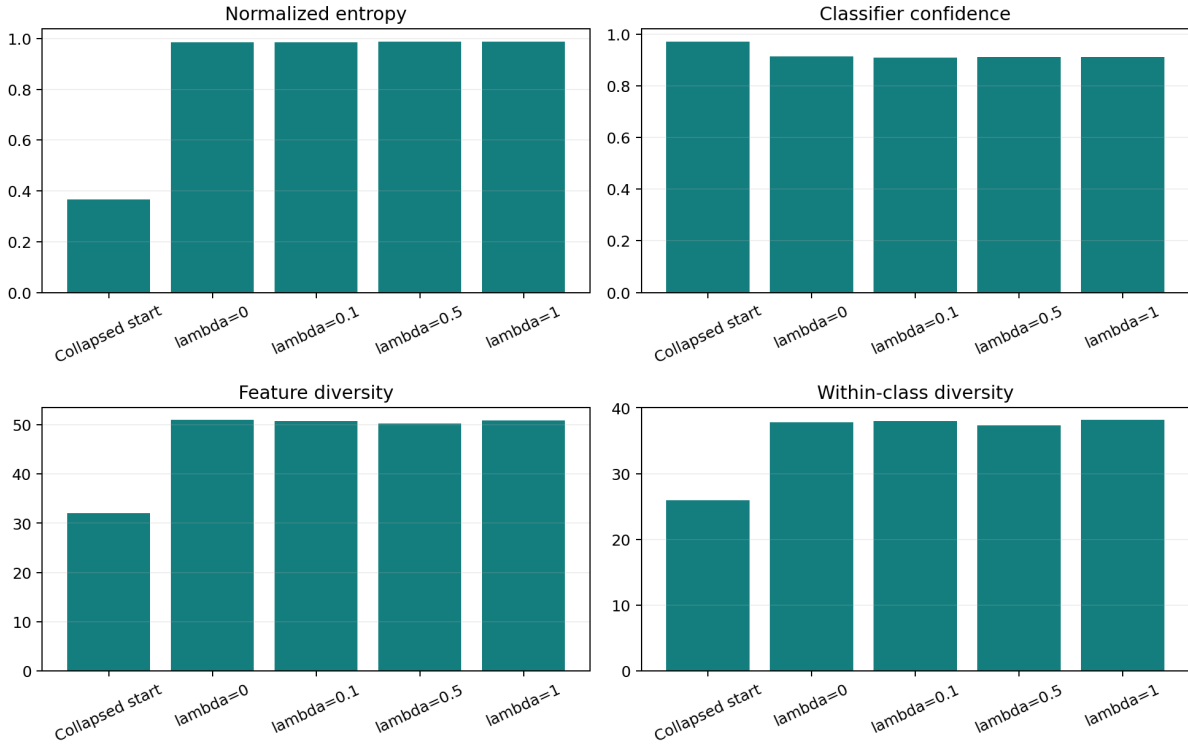


Figure 5: Comparación cuantitativa de las configuraciones de recuperación.

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: Las cuatro configuraciones de recuperación ($\lambda_{\text{div}} = 0, 0.1, 0.5$ y 1.0) recuperan la

totalidad de las 10 clases (10/10 modos), frente a 2/10 en el punto de partida. Un número alto de modos no garantiza automáticamente una distribución perfectamente equilibrada, pero en este caso la entropía normalizada confirma que sí se logró una cobertura casi uniforme: entre 0.984 y 0.988 en las cuatro configuraciones, muy cerca del máximo teórico de 1.0, frente a 0.367 en el punto colapsado. `class_distributions.png` corrobora visualmente que las cuatro configuraciones producen barras de altura similar en las 10 clases (con una ligera sobre-representación del dígito "9" y sub-representación del "4" en las cuatro), consistente con una recuperación genuina y no solo nominal.

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

Respuesta: La confianza del clasificador baja de 0.971 (colapsado) a un rango estrecho de 0.909–0.912 en las cuatro configuraciones recuperadas, sin una tendencia clara respecto a λ_{div} ($\lambda = 0.1$ tiene la confianza más baja, 0.909, mientras que $\lambda = 0$ y $\lambda = 1.0$ empatan en la más alta, 0.912). Esta pequeña caída es esperable y saludable: el punto colapsado concentraba casi toda la probabilidad en solo 2 clases con trazos muy estereotipados (más fáciles de clasificar con alta confianza), mientras que al cubrir las 10 clases el clasificador enfrenta dígitos naturalmente más ambiguos entre sí (por ejemplo 4 frente a 9, o 3 frente a 8). No se observa aquí un compromiso fuerte entre diversidad y calidad: las cuatro configuraciones logran cobertura completa sin una penalización de confianza mayor al ~6% respecto al punto colapsado.

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 suben de forma marcada respecto al punto colapsado (diversidad de características: 32.059 $\rightarrow$ ~50.3–51.0; diversidad intra-clase: 25.954 $\rightarrow$ ~37.3–38.1), pero entre las cuatro configuraciones de recuperación las diferencias son mínimas y no siguen un patrón monótono claro: $\lambda = 0.5$ tiene la diversidad de características más baja (50.256) y la intra-clase más baja (37.341) de las cuatro, mientras que $\lambda = 1.0$ tiene ambas más altas (50.829 y 38.141 respectivamente). No hay una configuración que mejore claramente una medida sacrificando la otra; las cuatro se comportan de forma muy similar, lo que sugiere que, una vez corregido el entrenamiento adversarial, la recuperación de diversidad en este escenario depende más de reanudar el entrenamiento sobre datos completos que del peso específico de la pérdida de diversidad explícita.

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

Respuesta: En esta corrida la mayor diversidad numérica sí corresponde a variaciones significativas y no a ruido: en `recovery_samples.png`, las cuatro filas de recuperación muestran dígitos completos y reconocibles de múltiples clases (ejemplo 1: la tercera columna muestra un "9" con trazo cursivo bien formado en las cuatro configuraciones; ejemplo 2: la octava columna muestra una forma con un trazo diagonal marcado parecido a un rayo, consistente entre las cuatro configuraciones). Cabe notar que, como la evaluación usa la misma semilla de muestreo para las cuatro configuraciones, las imágenes de una misma columna son decodificaciones del mismo vector latente por cada generador, y por eso lucen casi idénticas entre filas de recuperación –lo cual es coherente con que las métricas cuantitativas también sean casi idénticas entre configuraciones– y ninguna muestra manchas de ruido ni dígitos ambiguos.

4 Comportamiento durante el entrenamiento

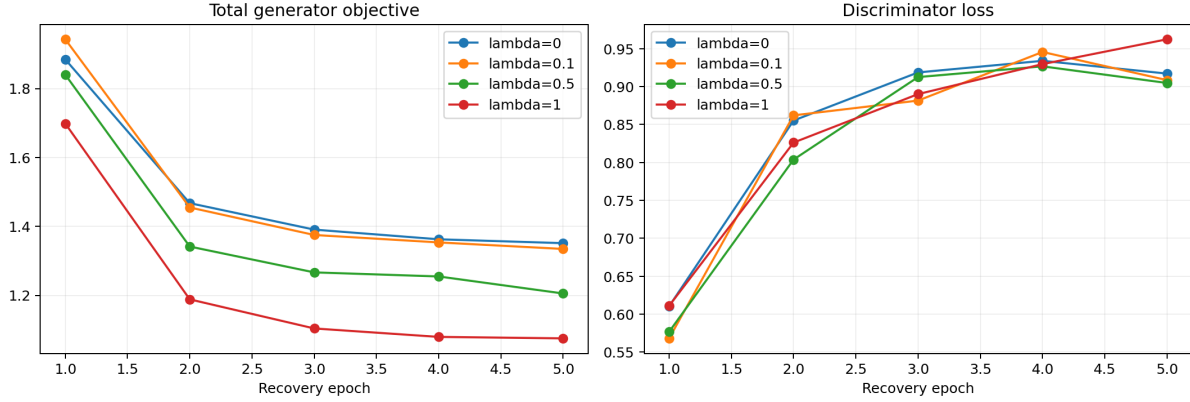


Figure 6: Pérdidas del generador y discriminador durante la recuperación.

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 $\mathcal{L}_G = \mathcal{L}_{\text{adv}} + \lambda_{\text{div}}\mathcal{L}_{\text{div}}$ combina dos términos con escalas y pesos distintos según la configuración. En `recovery_training.png` se observa que el objetivo total del generador es sistemáticamente más bajo cuanto mayor es λ_{div} (por ejemplo, $\lambda = 1.0$ termina en ~ 1.07 frente a ~ 1.35 en $\lambda = 0$), pero esto no significa que ese modelo entrene "mejor": $\mathcal{L}_{\text{div}}$ está definida como una cantidad negativa (más negativa cuanto más diversas son las salidas), así que un λ_{div} mayor simplemente resta más de ese término negativo al total, sin que ello implique una pérdida adversarial menor. Comparar el total directamente confundiría el efecto de la ponderación con una diferencia real de desempeño adversarial.

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

Respuesta: No por sí solas. `recovery_training.png` muestra que la pérdida del discriminador sube de forma saludable desde ~ 0.57 – 0.61 hasta ~ 0.90 – 0.98 a lo largo de las 5 épocas de recuperación en las cuatro configuraciones, sin colapsar a 0 ni dispararse –un patrón consistente con un equilibrio adversarial funcional. Sin embargo, esta curva por sí sola solo indica que el entrenamiento es estable; la confirmación de que el colapso de modos realmente se resolvió requiere las métricas de cobertura (10/10 modos, entropía ~ 0.98 – 0.99 , Pregunta 5) y la inspección visual de `recovery_samples.png`, que muestran dígitos reconocibles de las 10 clases.

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 $\lambda_{\text{div}} = 1.0$. Evidencia cuantitativa: (1) alcanza la entropía normalizada más alta de las cuatro configuraciones (0.988, frente a 0.984 en $\lambda = 0$); (2) alcanza la diversidad intra-clase más alta (38.141) y empatía en la confianza de clasificador más alta (0.912) junto con $\lambda = 0$, sin ningún costo aparente respecto a las demás. Evidencia visual: (3) en `recovery_samples.png`, la fila $\lambda = 1.0$ muestra las mismas 16 imágenes nítidas y reconocibles que el resto de configuraciones, sin manchas de ruido ni distorsiones adicionales pese a tener el peso de diversidad más alto evaluado; (4) en `class_distributions.png`, su distribución de clases es tan pareja como la de las demás configuraciones (mismo patrón de ligera sobre-representación de "9" y sub-representación de "4"). Dado que $\lambda_{\text{div}} = 1.0$ no muestra ninguna penalización visible de calidad frente a $\lambda_{\text{div}} = 0$, y sí una ligera ventaja consistente en las métricas de diversidad, es la opción mejor respaldada entre las evaluadas –con la salvedad de que las diferencias entre las cuatro configuraciones son pequeñas y podrían no ser robustas frente a otra semilla aleatoria, ya que cada una se evaluó una sola vez (Pregunta 13).

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

Respuesta: Sí cambiaría, aunque por poco: se elegiría $\lambda_{\text{div}} = 0$, que logra la confianza de clasificador empatada más alta (0.912, igual que $\lambda = 1.0$) sin necesidad de introducir el término adicional de diversidad ni su costo de ajuste. Como en este experimento fidelidad y diversidad no muestran un compromiso fuerte entre sí (las cuatro configuraciones logran calidad visual y cobertura muy similares), priorizar fidelidad favorecería la configuración más simple ($\lambda_{\text{div}} = 0$) por parsimonia, ya que no hay evidencia de que agregar la pérdida de diversidad aporte una mejora de fidelidad adicional en este caso.

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, por lo que su confianza y sus predicciones de clase pueden no ser completamente fiables frente a imágenes fuera de esa distribución; incluso en esta corrida corregida, la confianza no debe interpretarse como una medida absoluta de realismo, solo como un proxy relativo entre configuraciones. *Diseño experimental:* cada configuración de λ_{div} se evaluó con una sola semilla y un solo conjunto de 5 épocas de recuperación, por lo que las pequeñas diferencias entre configuraciones (por ejemplo, 0.984 frente a 0.988 de entropía) podrían no ser estadísticamente robustas frente a la variación aleatoria de la inicialización o el orden de los datos.

Nota de depuración: durante una primera ejecución de este notebook se detectó que el punto de control "colapsado" y las cuatro recuperaciones producían únicamente ruido no estructurado, sin relación aparente con dígitos reales, independientemente de λ_{div} . La causa era un error en la celda `train_gan_epoch`: durante el paso de actualización del generador, el discriminador se ponía en `discriminator.eval()` antes de calcular la pérdida adversarial. Como el discriminador incluye una capa `nn.BatchNorm2d`, el modo `eval()` lo hace usar estadísticas acumuladas (running mean/var) en lugar de las del lote actual, desincronizando el discriminador que evalúa al generador del discriminador que realmente se entrena, e impidiendo que el generador recibiera una señal de gradiente útil (el discriminador saturaba en la época 5 de 16 del preentrenamiento). El arreglo –cambiar esa línea a `discriminator.train()`– resolvió el problema por completo: el punto de

control colapsa correctamente hacia los dígitos 1 y 7 (97.2% de la masa, Pregunta 2) y las cuatro recuperaciones alcanzan cobertura completa de las 10 clases. Todos los resultados de este informe corresponden a la corrida ya corregida.

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 las cuatro configuraciones de recuperación con al menos 3 semillas aleatorias distintas cada una, midiendo la media y desviación estándar de entropía, confianza y diversidad intra-clase por configuración. Dado que en esta corrida las diferencias entre $\lambda_{\text{div}} = 0, 0.1, 0.5$ y 1.0 son pequeñas (por ejemplo, 0.984 frente a 0.988 de entropía), este experimento permitiría confirmar si $\lambda_{\text{div}} = 1.0$ realmente aventaja de forma consistente a $\lambda_{\text{div}} = 0$, o si la diferencia observada aquí es ruido de una sola corrida; si las diferencias desaparecen bajo múltiples semillas, la recomendación debería simplificarse a $\lambda_{\text{div}} = 0$ por parsimonia.

Recomendación final

Valor recomendado: $\lambda_{\text{div}} = 1.0$

Conclusión en una oración:

Respuesta: Tras corregir un error de `BatchNorm/modo eval()` en el paso del generador, las cuatro configuraciones de recuperación restauran por completo las 10 clases de dígitos (entropía ~ 0.98 – 0.99 frente a 0.367 en el punto colapsado), y $\lambda_{\text{div}} = 1.0$ se recomienda por lograr la mayor entropía y diversidad intra-clase sin sacrificar confianza de clasificador frente a las demás configuraciones evaluadas.