Dựa trên bài báo gốc về CRAIG — *"Coresets for Data-efficient Training of Machine Learning Models"* của Mirzasoleiman et al., các thử nghiệm chính đã được thực hiện nhằm kiểm chứng hiệu quả của CRAIG trong việc chọn tập dữ liệu con để tăng tốc huấn luyện mô hình mà vẫn đạt độ chính xác tương đương.

Để đề xuất các thử nghiệm số tương tự cho biến thể CRAIG kết hợp Convex Hull, trước tiên ta mô tả ngắn gọn các thử nghiệm số của thuật toán gốc CRAIG như sau

1. Kiến trúc mô hình & tác vụ

- **Convex setting**: Logistic regression với hàm mất mát L2-regularized logistic loss.
  - Dataset: covtype.binary (581,012 mẫu, 54 chiều), Ijcnn1 (49,990 train / 91,701 test).
- **Non-convex setting**: Huấn luyện mạng neural
  - Mạng nhỏ: MLP 1 tầng ẩn (100 neurons) với sigmoid + softmax, trên MNIST.
  - Mạng lớn: ResNet-20 trên CIFAR-10.

2. Phương pháp so sánh

- CRAIG vs:
  - Toàn bộ tập dữ liệu.
  - Tập con ngẫu nhiên cùng kích thước (5%, 10%, 20%, …)
- Optimizers:
  - SGD
  - SVRG
  - SAGA
- Được áp dụng trên cả tập đầy đủ và tập con được chọn bởi CRAIG.

3. Các chỉ số đánh giá

- **Loss residual** (khoảng cách tổn thất so với hội tụ tối ưu).
- **Test accuracy** (cho mạng neural).
- **Gradient approximation error** ($\approx\|\nabla f(V) - \nabla f(S)\|$).
- **Thời gian huấn luyện** (wall-clock time).
- **Speedup** (tính theo tỷ lệ $|V|/|S|$).

4. Kích thước tập con

- Đánh giá CRAIG với các tỷ lệ chọn dữ liệu: 1%, 2%, 3%, 5%, 10%, 20%, … đến 90%.
- Trong các thí nghiệm mạng nơ-ron, tập con CRAIG được cập nhật mỗi epoch hoặc mỗi 5 epoch.

5. Cách thực hiện CRAIG

- Với mô hình convex: tiền xử lý các khoảng cách gradient dựa trên $\|x_i - x_j\|$.
- Với mô hình deep: dùng gradient tại input layer cuối ($\frac{\partial L}{\partial z_L}$) để đánh giá sự khác biệt gradient, giúp chọn coreset hiệu quả và nhẹ tính toán hơn.

Đề xuất thử nghiệm cho CRAIG biến thể:

Đề xuất 1:

Giữ nguyên setup trên để so sánh công bằng, chỉ thay thế thuật toán CRAIG gốc bằng biến thể CRAIG + Convex Hull:

Thử nghiệm	Mô tả
Logistic Regression (covtype, Ijcnn1)	Đánh giá loss residual và thời gian hội tụ với biến thể CRAIG.
MLP trên MNIST	Đánh giá accuracy và training time với tập con CRAIG biến thể.
ResNet-20 trên CIFAR10	Test accuracy khi chọn tập con 1%-20% và cập nhật mỗi epoch hoặc mỗi 5 epoch.
Gradient estimation error	So sánh $\|\nabla f(V) - \nabla f(S\_variant)\|$ giữa CRAIG gốc và biến thể.
Generalization	Đánh giá hiệu quả tổng quát (accuracy trên test set) khi giảm lượng dữ liệu học.
Speedup vs Accuracy tradeoff	Plot tốc độ huấn luyện (speedup) theo từng kích thước tập con và accuracy.

Sử dụng cùng bộ mã thử nghiệm gốc, thay thế hàm chọn coreset bằng thuật toán CRAIG biến thể và đo các chỉ số tương tự.

Đề xuất 2: Để **làm nổi bật ưu thế của biến thể CRAIG + Convex Hull**. Mục tiêu là chứng minh rằng biến thể này **chọn tập con hiệu quả hơn về mặt thông tin và hình học**, từ đó **giảm đáng kể lượng dữ liệu cần thiết mà vẫn duy trì chất lượng mô hình**, đặc biệt trong các bài toán có cấu trúc hình học rõ rệt như phân loại biên quyết định.

Tổng quan chiến lược thử nghiệm

CRAIG + Convex Hull giả định rằng việc chọn các điểm nằm trên bao lồi (Convex Hull) giúp bao phủ tốt hơn không gian đặc trưng (feature space) – tức là chọn các điểm "đa dạng và rìa" hơn so với CRAIG thuần. Ta cần thử nghiệm để:

- Xác thực **hiệu quả biểu diễn hình học** của tập con.
- Đo **tốc độ hội tụ**.
- So sánh về **số lượng mẫu cần thiết để đạt độ chính xác mục tiêu**.
- Kiểm tra khả năng **tổng quát hóa (generalization)**.

Thiết kế thử nghiệm số cụ thể

1. Bài toán phân loại tuyến tính (Logistic Regression)

Dataset:

- Iris, MNIST (2-class: 3 vs 8), Ijcnn1, Covtype.binary

- Kích thước vừa đủ để hình dung phân bố.

Biến kiểm soát:

Biến	Mô tả
CRAIG	Phiên bản gốc.
CRAIG + Convex Hull	Chọn điểm từ bao lồi theo ∇fi(w).
Random subset	So sánh cơ sở.
Full dataset	Chuẩn vàng (upper bound).

Thử nghiệm:

- Vary subset size: 1%, 3%, 5%, 10%, 20%, 30%.
- Tối ưu hóa bằng SGD/SAGA.
- **Đo:**
  - Training loss residual (trên epoch).
  - Test accuracy.
  - Gradient approximation error: $\| \nabla f(w) - \sum_{j \in S} \gamma_j \nabla f(w_j) \|$

- **Convex Hull Coverage**: đo tỉ lệ các điểm trong full dataset rơi vào bao lồi của tập con.

2. Non-convex: Deep Neural Networks

Dataset:

- MNIST (full 10-class), CIFAR-10, Fashion-MNIST.

Mạng:

- MNIST: 1 Hidden Layer MLP (100 neurons)
- CIFAR: ResNet-20

Biến kiểm soát:

Biến	Mô tả
CRAIG	Original greedy coreset.
CRAIG + Convex Hull	Áp dụng convex hull trong không gian logit hoặc feature cuối cùng.
GradMatch, GLISTER	Các baseline khác trong data subset selection.

Thử nghiệm:

- Epoch-level coreset selection.
- Thử update coreset mỗi 5 epochs.

- Các size: 5%, 10%, 20%.

Đo lường:

- Accuracy vs # data used.
- Time-to-convergence (đến ngưỡng 95% accuracy).
- Sharpness/flatness của loss landscape trên coreset.

Thử nghiệm bổ sung nổi bật tính hình học

3. Phân tích không gian biểu diễn

- Ánh xạ tập con lên PCA để kiểm chứng:
  - CRAIG + Convex Hull chọn được các điểm "rìa" hoặc biểu diễn tốt toàn bộ tập.
- Đo độ phủ hình học: so sánh thể tích bao lồi của subset với tập gốc.

Kỳ vọng kết quả thể hiện ưu thế của CRAIG + Convex Hull

Tiêu chí	CRAIG	CRAIG + Con Hull	Ghi chú
Gradient approx. error	Trung bình	Thấp hơn	Bao phủ tốt hơn biên gradient
Accuracy tại 5–10% data	90–93%	94–96%	Giữ accuracy cao hơn với ít dữ liệu
Convex hull coverage	20–30%	60–80%	Nhiều điểm nằm trong hull của subset hơn
Training speed	Nhanh	Nhanh hơn	Vì ít sample hơn nhưng gradient đại diện tốt hơn