Continual Learning and Memory Models Aksel Kretsinger-Walters, adk2164
Gradient Episodic Memory for Continual Learning adk2164@columbia.edu
COMS 6998: Spring 2026 **Due:** 02/05/2026 11:59 PM ET

Gradient Episodic Memory for Continual Learning

Motivation. In continual learning, models trained sequentially on tasks often suffer from catastrophic forgetting, where updates for new tasks overwrite parameters critical for previous tasks, motivating methods that explicitly constrain learning dynamics to preserve past performance.

Problem Setting. The model observes a sequence of tasks and receives training data for one task at a time, with the goal of minimizing loss on the current task while ensuring that performance on all previously learned tasks does not degrade.

Core Idea. Gradient Episodic Memory (GEM) frames continual learning as a constrained optimization problem, where parameter updates for the current task are restricted to directions that do not increase the loss on stored examples from previous tasks.

Episodic Memory. GEM maintains a small episodic memory buffer containing representative samples from each past task, which serves as a proxy for the original task distributions when enforcing constraints during learning.

Gradient Constraints. At each update step, gradients are computed for the current task and for each previous task using their episodic memory samples, and the update is required to satisfy non-increase constraints on all previous task losses.

Geometric Interpretation. If the gradient for the current task has a negative dot product with a past task gradient, it would increase that task's loss, and GEM resolves this by projecting the gradient onto the closest feasible direction that satisfies all constraints.

Optimization Formulation. The constrained update is obtained by solving a quadratic program that minimizes the distance between the original gradient and a modified gradient subject to linear inequality constraints defined by past task gradients.

Learning Dynamics. By explicitly shaping gradient directions, GEM balances plasticity and stability, allowing learning on new tasks while guaranteeing that empirical loss on stored past examples does not increase.

Comparison to Regularization. Unlike parameter regularization methods such as EWC, which penalize parameter movement based on importance measures, GEM operates directly in gradient space and enforces task-level performance constraints.

Limitations. GEM incurs additional computational cost due to storing episodic memories and solving a quadratic program at each update, and its guarantees depend on the representativeness of the stored memory samples.

Key Takeaway. Gradient Episodic Memory demonstrates that catastrophic forgetting can be mitigated by explicitly constraining optimization trajectories, reframing continual learning as a problem of feasible gradient geometry rather than parameter anchoring.