DeltaNet's Challenge

Vishal Pandey
pandeyvishal.mlprof@gmail.com
Independent Researcher

Abstract

Linear attention mechanisms can be derived from quadratic Lagrangian objectives, where coefficients emerge as dual variables. I extend this view to DeltaNet, a variant of linear attention with adaptive forgetting. We show that its recurrence arises naturally from a dynamic quadratic regularization, yielding multiplicative survival factors for memory traces. This note provides both a mathematical derivation and an intuitive interpretation of DeltaNet from a variational perspective.

1 Introduction

Attention mechanisms form the backbone of modern sequence models. While standard softmax attention incurs quadratic complexity, *linear attention* replaces the kernel with inner products, yielding efficient recurrences. A useful observation is that vanilla linear attention can be obtained by solving a quadratic optimization problem via a Lagrangian.

DeltaNet modifies the vanilla recurrence with an adaptive forgetting mechanism:

$$S_t = S_{t-1} - \beta_t S_{t-1} k_t k_t^\top + \beta_t v_t k_t^\top, \quad o_t = S_t q_t, \tag{1}$$

where $\beta_t \in (0, 1]$ is a step-size. The goal of this note is to derive the Lagrangian underlying this update and provide intuition for its dynamics.

2 Vanilla Linear Attention via Lagrangian

The vanilla update is

$$S_t = S_{t-1} + v_t k_t^\top, \quad o_t = S_t q_t. \tag{2}$$

Unrolling yields

$$o_t = \sum_{i=1}^{t} (k_i^\top q_t) v_i. \tag{3}$$

This can be recovered by maximizing

$$L(\alpha) = \sum_{i=1}^{t} \alpha_i (k_i^\top q_t) - \tfrac{1}{2} \alpha^\top I_t \alpha, \tag{4}$$

where I_t is the $t \times t$ identity and $\alpha = (\alpha_1, \ldots, \alpha_t)$. The solution is $\alpha_i = k_i^\top q_t$, matching the recurrence coefficients.

3 DeltaNet as a Constrained Optimization

DeltaNet introduces a correction: before adding $v_t k_t^\top$, we subtract redundant content already aligned with k_t. Unrolling Eq. (1) gives

$$o_t = \sum_{i=1}^{t} \Big(\beta_i \prod_{j=i+1}^{t} (1 - \beta_j \, k_j^\top k_i) \Big) (k_i^\top q_t) v_i. \tag{5}$$

Thus the coefficient in front of each v_i is a survival factor, decayed by subsequent projections.

Lagrangian Formulation

We define a time-varying quadratic form:

$$L(\alpha) = \sum_{i=1}^{t} \alpha_i (k_i^\top q_t) - \tfrac{1}{2} \alpha^\top M_t \alpha, \tag{6}$$

where M_t encodes interaction penalties between directions. For DeltaNet, we set

$$M_t = I_t + \sum_{j=1}^{t} \tfrac{1}{\beta_j} k_j k_j^\top. \tag{7}$$

The stationarity condition $\nabla_\alpha L(\alpha) = 0$ gives

$$\alpha_t = \beta_t \Big(s_t - \langle k_t, \sum_{i<t} \alpha_i k_i \rangle \Big), \quad s_t = k_t^\top q_t, \tag{8}$$

which reproduces the DeltaNet recurrence. Hence, DeltaNet is the optimizer of a constrained quadratic program with dynamic projection penalties.

4 Intuition

- **Vanilla:** Each memory v_i survives with weight $k_i^\top q_t$.
- **DeltaNet:** Contributions are adaptively corrected: if a direction k_t is already represented, future coefficients decay.
- **Optimization View:** The penalty matrix M_t enforces "do not overuse already represented key directions" with strength $1/\beta_t$.

5 Conclusion

We have shown that DeltaNet admits a principled Lagrangian formulation with a dynamic quadratic penalty, connecting it to online Newton-like methods. This perspective unifies linear attention and DeltaNet under a common variational framework, suggesting new avenues for designing recurrent attention mechanisms.