

How to Make GDN2 More Stable?

Su Jianlin

2026-10-09

We know that the core of DeltaNet is the update matrix of the form $\mathbf{I} - \eta \mathbf{k} \mathbf{k}^\top$ (see “A Brief History of Linear Attention: From Imitation, to Innovation, to Giving Back”). When $\eta \leq 2$ and L2 normalization is imposed on $\mathbf{k}$, its cumulative product has good numerical stability (see “Why Does DeltaNet Need L2 Normalize?” and “The Core Inverse Matrix Elements of DeltaNet Are Always in $[-1, 1]$ ”). This is the key to its stable training, and subsequent works such as GDN and KDA have followed this setting.

Later, GDN2 attempted to generalize the update matrix to $\mathbf{I} - \eta \mathbf{k} \mathbf{j}^\top$; a similar choice also appears in RWKV7. However, this generalization may bring some new instability issues, and this article gives a brief analysis of them.

1 Problem Description

The general form of GDN2 is

$$\mathbf{S}_t = \mathbf{S}_{t-1} \mathbf{D}_t (\mathbf{I} - \eta_t \mathbf{k}_t \mathbf{j}_t^\top) + \eta_t \mathbf{v}_t \mathbf{k}_t^\top \quad (1)$$

where $\mathbf{k}_t, \mathbf{j}_t \in \mathbb{R}^d$, $\eta_t \geq 0$, and $\mathbf{D}_t$ is a diagonal matrix whose diagonal elements are usually restricted to $[0, 1]$, representing the forgetting matrix. KDA corresponds to the special case $\mathbf{j}_t = \mathbf{k}_t$, and GDN further corresponds to the special case $\mathbf{D}_t = \gamma_t \mathbf{I}$.

When we expand the recursion, $\mathbf{S}_t$ involves a cumulative product of $\mathbf{D}_t (\mathbf{I} - \eta_t \mathbf{k}_t \mathbf{j}_t^\top)$, so to guarantee numerical stability, we should hope that

$$\|\mathbf{D}_t (\mathbf{I} - \eta_t \mathbf{k}_t \mathbf{j}_t^\top)\|_2 \leq 1 \quad (2)$$

where $\|\cdot\|_2$ is the spectral norm. Since $\mathbf{D}_t$ is a diagonal matrix whose diagonal entries are already in $[0, 1]$, it satisfies $\|\mathbf{D}_t\|_2 \leq 1$; therefore a sufficient condition for the above is

$$\|\mathbf{I} - \eta_t \mathbf{k}_t \mathbf{j}_t^\top\|_2 \leq 1 \quad (3)$$

Unfortunately, as we shall see next, when $\mathbf{k}_t, \mathbf{j}_t$ are not collinear (i.e., the regime GDN2 hopes to extend to), the above is impossible to satisfy.

2 The Norm of the Spectrum

For simplicity, the discussion below omits the subscript t and assumes that $\mathbf{k}, \mathbf{j}$ are already normalized (otherwise $\|\mathbf{k}\|, \|\mathbf{j}\|$ can always be absorbed into η). We have

$$\|\mathbf{I} - \eta \mathbf{k} \mathbf{j}^\top\|_2 = \frac{\eta + \sqrt{\eta^2 - 4\eta \mathbf{k}^\top \mathbf{j} + 4}}{2} \quad (4)$$

This result is perhaps the most tedious part of this article, so we defer its proof to the end. Since $\mathbf{k}, \mathbf{j}$ are already normalized, we have $\mathbf{k}^\top \mathbf{j} \in [-1, 1]$; and when $\mathbf{k}^\top \mathbf{j} < 1$ and $\eta > 0$, we have

$$\frac{\eta + \sqrt{\eta^2 - 4\eta \mathbf{k}^\top \mathbf{j} + 4}}{2} > \frac{\eta + \sqrt{\eta^2 - 4\eta + 4}}{2} = \frac{\eta + |\eta - 2|}{2} = \max(\eta, 2) - 1 \geq 1 \quad (5)$$

That is, as long as $\mathbf{k} \neq \mathbf{j}$ and $\eta > 0$, no matter how you tune it, the spectral norm of $\mathbf{I} - \eta \mathbf{k} \mathbf{j}^\top$ is strictly greater than 1! More precisely,

$$\|\mathbf{I} - \eta \mathbf{k} \mathbf{j}^\top\|_2 \leq 1 \iff \mathbf{k} = \mathbf{j} \text{ and } \eta \leq 2 \text{ (or } \eta = 0) \quad (6)$$

But if $\mathbf{k} = \mathbf{j}$, this degenerates back to KDA, losing the generalization that GDN2 is meant to provide. And when $\mathbf{k} \neq \mathbf{j}$, $\|\mathbf{I} - \eta \mathbf{k} \mathbf{j}^\top\|_2$ is necessarily greater than 1, which means the stability of GDN2 is not guaranteed. How can we break this impasse?

3 A Guarantee Strategy

Fortunately, we still have the forgetting-gate matrix $\mathbf{D}$. Since it is restricted to $[0, 1]$, the spectral norm of their product $\mathbf{D}(\mathbf{I} - \eta \mathbf{k} \mathbf{j}^\top)$ will most likely be less than 1, so training in practice does not necessarily fail. However, this is only probabilistic, not a theoretical guarantee.

The most ideal approach would be to directly compute the spectral norm of $\mathbf{D}(\mathbf{I} - \eta \mathbf{k} \mathbf{j}^\top)$ and then apply spectral clipping to it; however, for a general $\mathbf{D}$, the spectral norm of this matrix has no general closed-form solution.

Fortunately, the spectral norm of $\mathbf{I} - \eta \mathbf{k} \mathbf{j}^\top$ has an analytic solution, so we can settle for second best: use its spectral norm to clip $\mathbf{D}$:

$$\mathbf{D}(\mathbf{I} - \eta \mathbf{k} \mathbf{j}^\top) \rightarrow \min(\mathbf{D}, 1/\rho)(\mathbf{I} - \eta \mathbf{k} \mathbf{j}^\top) \quad (7)$$

where $\rho = \|\mathbf{I} - \eta \mathbf{k} \mathbf{j}^\top\|_2$. In this way, we only modify the diagonal entries of $\mathbf{D}$ that are greater than $1/\rho$. Since in practice η is usually not large, on average $1/\rho$ will not be noticeably smaller than 1, so the scope of the change is narrow—that is, with as small and simple a modification as possible, the stability of GDN2 is guaranteed.

It is worth pointing out that both ρ and $\min(\mathbf{D}, 1/\rho)$ are merely token-wise operations that can be kept independent of the RNN operator, and ρ also has a closed form, so after kernel fusion they basically do not affect efficiency. Hence this counts as a modification with extremely low mental burden.

4 Proof

Finally, we supply the proof of Equation (4). Without loss of generality, assume $\mathbf{k}, \mathbf{j}$ are not collinear. Orthogonalize them as $\mathbf{u}_1 = \mathbf{j}$ and $\mathbf{u}_2 = \frac{\mathbf{k} - c\mathbf{j}}{\sqrt{1-c^2}}$, where $c = \mathbf{k}^\top \mathbf{j}$. Let $\mathbf{U}_1 = [\mathbf{u}_1, \mathbf{u}_2]$, extend it to a complete orthogonal basis $\mathbf{U} = [\mathbf{U}_1, \mathbf{U}_2]$, and write $\mathbf{A} = \mathbf{I} - \eta \mathbf{k} \mathbf{j}^\top$.

For a vector $\mathbf{x}$ orthogonal to both $\mathbf{k}$ and $\mathbf{j}$, we have $\mathbf{A}\mathbf{x} = \mathbf{x}$ and $\mathbf{x}^\top \mathbf{A} = \mathbf{x}^\top$, hence $\mathbf{A}\mathbf{U}_2 = \mathbf{U}_2$ and $\mathbf{U}_2^\top \mathbf{A} = \mathbf{U}_2^\top$, and therefore

$$\mathbf{U}^\top \mathbf{A} \mathbf{U} = \begin{pmatrix} \mathbf{U}_1^\top \\ \mathbf{U}_2^\top \end{pmatrix} \mathbf{A} \begin{pmatrix} \mathbf{U}_1 & \mathbf{U}_2 \end{pmatrix} = \begin{pmatrix} \mathbf{U}_1^\top \mathbf{A} \mathbf{U}_1 & \mathbf{U}_1^\top \mathbf{A} \mathbf{U}_2 \\ \mathbf{U}_2^\top \mathbf{A} \mathbf{U}_1 & \mathbf{U}_2^\top \mathbf{A} \mathbf{U}_2 \end{pmatrix} = \begin{pmatrix} \mathbf{U}_1^\top \mathbf{A} \mathbf{U}_1 & \mathbf{0} \\ \mathbf{0} & \mathbf{I}_{d-2} \end{pmatrix} \quad (8)$$

Since orthogonal transformations do not change singular values, the non-trivial singular values of $\mathbf{A}$ can only come from the matrix $\mathbf{U}_1^\top \mathbf{A} \mathbf{U}_1$, which computes directly to

$$\mathbf{B} \triangleq \mathbf{U}_1^\top \mathbf{A} \mathbf{U}_1 = \begin{pmatrix} 1 - \eta c & 0 \\ -\eta \sqrt{1 - c^2} & 1 \end{pmatrix} \quad (9)$$

For the singular values of a 2×2 matrix, there is a rather clever direct computation method. Let the two singular values be $\sigma_1 \geq \sigma_2$; using the relations between singular values, the Frobenius norm, and the determinant, we get

$$\underbrace{\sigma_1^2 + \sigma_2^2 = \|\mathbf{B}\|_F^2 = \eta^2 - 2\eta c + 2, \quad \sigma_1 \sigma_2 = |\det(\mathbf{B})| = |1 - \eta c|}_{\Downarrow} \quad (10)$$

$$\sigma_1 \pm \sigma_2 = \sqrt{\sigma_1^2 + \sigma_2^2 \pm 2\sigma_1 \sigma_2} = \sqrt{\eta^2 - 2\eta c + 2 \pm 2|1 - \eta c|} \quad (11)$$

We are mainly interested in σ_1 , which equals the average of the two ‘ $\pm$ ’ expressions; at this point the absolute value can be dropped (both signs are computed anyway), and simplifying finally yields

$$\sigma_1 = \frac{\eta + \sqrt{\eta^2 - 4\eta c + 4}}{2} \quad (12)$$

We have already shown that this is no less than 1, so σ_1 is the spectral norm. Q.E.D.

5 Summary

This article analyzed the potential stability issues that arise when GDN2 generalizes the update matrix from $\mathbf{I} - \eta \mathbf{k} \mathbf{k}^\top$ to $\mathbf{I} - \eta \mathbf{k} \mathbf{j}^\top$, and discussed the corresponding mitigation strategy.

*When reposting, please include the URL of this article: <https://kerue.fm/archives/11945>
For more detailed reposting terms, please refer to: “Scientific Spaces FAQ”*