Local Credit Assignment in Compartmental Dendritic Networks

Anonymous Author(s)
Affiliation
Address
email

Abstract

We present a rigorous mathematical framework for local credit assignment in compartmental dendritic networks. Starting from the passive cable equation and deriving exact backpropagation gradients for general dendritic trees, we introduce local approximations that use only signals available at each synapse. We formulate three classes of learning rules3-factor, 4-factor, and 5-factorand extend them with four morphology-aware mechanisms that explicitly incorporate dendritic tree topology: path-integrated propagation, branch-specific depth modulation, dendritic normalization, and apical-basal differentiation. We provide theoretical analysis of shunting inhibition's role in divisive gain control, prove positive expected alignment between local and exact gradients under broadcast error schemes, and establish connections to feedback alignment, predictive coding, and homeostatic plasticity. We prove consistency with the implemented algorithms and provide comprehensive experimental protocols for validation.

1 Compartmental Voltage Model

1.1 From the Passive Cable to the Compartment Equation

Starting from the linear passive cable equation for membrane potential $v(x,t)$ relative to rest,

$$c_m \frac{\partial v}{\partial t} = \frac{1}{r_a}\frac{\partial^2 v}{\partial x^2} - \frac{1}{r_m} v + i_{\text{syn}}(x,t),$$

and discretizing a dendritic branch into isopotential compartments with axial conductances, the steady-state ($\partial_t v = 0$) yields a nodal balance of conductances and driving forces.[1] Representing synapses as conductances with reversal potentials and siblings as dendritic conductances gives the compartment equation below with unit leak to E_{leak}=0. This clarifies that (i) all inputs contribute via conductances, (ii) total conductance g_n^{tot} controls both input resistance R_n^{tot} and divisive normalization, and (iii) shunting inhibition corresponds to adding conductance with $E_{\text{inh}} \approx 0$ (Section 1.4).

1.2 Voltage Equation

Consider compartment n receiving synaptic inputs indexed by j and dendritic inputs from child compartments. Let:

- $x_j \in \mathbb{R}_+$: presynaptic activity at synapse j
- $E_j \in \mathbb{R}$: reversal potential of synapse j (excitatory: $E_j > 0$; inhibitory: $E_j \leq 0$)

[1] See Koch's *Biophysics of Computation* and Dayan & Abbott's *Theoretical Neuroscience* for derivations and assumptions underlying the linear regime.

Submitted to 39th Conference on Neural Information Processing Systems (NeurIPS 2025). Do not distribute.

- $g_j^{\text{syn}} \geq 0$: synaptic conductance (learned parameter)
- $V_j \in \mathbb{R}$: voltage of child compartment j
- $g_j^{\text{den}} \geq 0$: dendritic conductance from child j (learned parameter)

Currents. Synaptic current:

$$I_{\text{syn}} = \sum_j (E_j - V_n) x_j g_j^{\text{syn}} \tag{1}$$

Dendritic current:

$$I_{\text{den}} = \sum_j (V_j - V_n) g_j^{\text{den}} \tag{2}$$

Steady-state voltage. With unit leak conductance to reversal potential 0:

$$V_n = \frac{\sum_j E_j x_j g_j^{\text{syn}} + \sum_j V_j g_j^{\text{den}}}{\sum_j x_j g_j^{\text{syn}} + \sum_j g_j^{\text{den}} + 1} \tag{3}$$

Total conductance and resistance.

$$g_n^{\text{tot}} = \sum_j x_j g_j^{\text{syn}} + \sum_j g_j^{\text{den}} + 1, \qquad R_n^{\text{tot}} = \frac{1}{g_n^{\text{tot}}} \tag{4}$$

Lemma 1 (Convexity and Bounds)**.** *Let $\mathcal{S}_n = \{E_j\}_{\text{syn at } n} \cup \{V_j\}_{\text{children}} \cup \{0\}$. Then V_n in* (3) *is a convex combination of elements of $\mathcal{S}_n$, hence*

$$\min \mathcal{S}_n \;\leq\; V_n \;\leq\; \max \mathcal{S}_n.$$

Moreover, $0 < R_n^{\text{tot}} \leq 1$ and $R_n^{\text{tot}} g_i^{\text{den}} < 1$ for all i.

Proof. Immediate from (3)(4) since all conductances are nonnegative and leak adds $+1$ to the denominator. □

1.3 Local Sensitivities

Proposition 1 (Synaptic Gradient)**.**

$$\frac{\partial V_n}{\partial g_i^{\text{syn}}} = x_i R_n^{\text{tot}} (E_i - V_n) \tag{5}$$

Proof. Apply quotient rule to (3):

$$\begin{aligned} \frac{\partial V_n}{\partial g_i^{\text{syn}}} &= \frac{E_i x_i \cdot g_n^{\text{tot}} - \left(\sum_j E_j x_j g_j^{\text{syn}} + \sum_j V_j g_j^{\text{den}}\right) \cdot x_i}{(g_n^{\text{tot}})^2} \\ &= \frac{E_i x_i}{g_n^{\text{tot}}} - \frac{V_n x_i}{g_n^{\text{tot}}} = x_i R_n^{\text{tot}} (E_i - V_n). \end{aligned}$$

□

Proposition 2 (Child Voltage Gradient)**.**

$$\frac{\partial V_n}{\partial V_i} = g_i^{\text{den}} R_n^{\text{tot}} \tag{6}$$

Proposition 3 (Dendritic Gradient)**.**

$$\frac{\partial V_n}{\partial g_i^{\text{den}}} = R_n^{\text{tot}} (V_i - V_n) \tag{7}$$

Proposition 4 (Additional Local Sensitivities)**.**

$$\frac{\partial V_n}{\partial x_i} = g_i^{\text{syn}} R_n^{\text{tot}} (E_i - V_n), \qquad \frac{\partial V_n}{\partial E_i} = x_i g_i^{\text{syn}} R_n^{\text{tot}}, \qquad \frac{\partial V_n}{\partial g^{\text{leak}}} = -V_n R_n^{\text{tot}}.$$

Remark 1. *The sensitivity $\partial V_n / \partial g^{\text{leak}} = -V_n R_n^{\text{tot}}$ applies if g^{leak} is a trainable parameter. In all reported experiments, we fix g^{leak}=1 (normalized units).*

1.4 Shunting Inhibition and Divisive Gain Control

A conductance-based inhibitory synapse with $E_{\text{inh}} \approx E_{\text{leak}}$=0 contributes current $I_{\text{inh}} = (0 - V_n)\, x_j g_j^{\text{syn}}$ and increases g_n^{tot} in (4).

Proposition 5 (Subthreshold Effect of Shunts)**.** *For a pure shunt ($E_j = 0$), the steady-state sensitivity to the inhibitory conductance is*

$$\frac{\partial V_n}{\partial g_j^{\text{syn}}} = x_j R_n^{\text{tot}}(0 - V_n) = -x_j R_n^{\text{tot}} V_n.$$

Thus V_n is multiplicatively attenuated (divisive normalization) by increased inhibitory conductance at fixed drives.

Remark 2 (Divisive vs. Subtractive at the Firing-Rate Level)**.** *While shunting produces divisive scaling of subthreshold voltages, its net effect on firing rates can be subtractive in many regimes [4], so we report both voltage- and rate-level analyses in experiments. Normalization via added conductance is consistent with canonical divisive normalization models in cortex [3].*

Inhibitory/shunting synapses. For an inhibitory synapse with $E_j \approx 0$, the 3F update reduces to

$$\Delta g_{j,\text{inh}}^{\text{syn}} = \eta \, \langle x_j R_n^{\text{tot}}(-V_n)\, e_n \rangle_B,$$

i.e., anti-Hebbian in V_n and divisive in g_n^{tot}. With 4F/5F, multiply by ρ and ϕ (Def. 4). Note that the same multiplicative factors are applied to both excitatory and inhibitory synapses in the implementation; the sign difference arises solely from the driving force $(E_j - V_n)$.

1.5 Loss Propagation

Let V_0 denote the somatic/output compartment. The decoder produces $\hat{y} = W_{\text{dec}} V_0$ (linear case), and L is the task loss. Define the error gradients:

$$\delta^y := \frac{\partial L}{\partial \hat{y}}, \qquad \delta_0 := \frac{\partial L}{\partial V_0} = \left(\frac{\partial \hat{y}}{\partial V_0}\right)^\top \delta^y = W_{\text{dec}}^\top \delta^y. \tag{8}$$

Theorem 1 (Backpropagation on a Dendritic Tree)**.** *Let the dendritic morphology be a rooted tree with soma/output at node 0. For any compartment n with parent set $\mathcal{P}(n)$ (typically $|\mathcal{P}(n)|$=1), the loss gradient satisfies the recursion*

$$\frac{\partial L}{\partial V_n} = \sum_{p \in \mathcal{P}(n)} \frac{\partial L}{\partial V_p} \frac{\partial V_p}{\partial V_n} = \sum_{p \in \mathcal{P}(n)} \delta_p \, R_p^{\text{tot}} \, g_{n \to p}^{\text{den}}, \qquad \delta_p \equiv \frac{\partial L}{\partial V_p}. \tag{9}$$

Unrolling the recursion yields a sum over all directed paths $\mathcal{P} : n \rightsquigarrow 0$:

$$\frac{\partial L}{\partial V_n} = \frac{\partial L}{\partial V_0} \sum_{\mathcal{P}: n \rightsquigarrow 0} \prod_{(i \to k) \in \mathcal{P}} R_k^{\text{tot}} \, g_{i \to k}^{\text{den}}. \tag{10}$$

Proof. Apply the multivariate chain rule on the directed acyclic computation graph defined by the tree; use Proposition 2. Each path contributes a product of edge sensitivities. Summing over parents produces (9); unrolling yields (10). □

Corollary 1 (Chain Case)**.** *If the morphology is a simple chain $V_0 \leftarrow V_1 \leftarrow \cdots \leftarrow V_n$,* (10) *reduces to*

$$\frac{\partial L}{\partial V_n} = \frac{\partial L}{\partial V_0} \prod_{i=1}^{n} R_{i-1}^{\text{tot}} g_i^{\text{den}}. \tag{11}$$

Defining $g_0^{\text{den}} = 1$ and reindexing:

$$\frac{\partial L}{\partial V_n} = \frac{\partial L}{\partial V_0} \prod_{i=0}^{n} R_i^{\text{tot}} g_i^{\text{den}}. \tag{12}$$

Corollary 2 (Synaptic Parameter Gradient)**.**

$$\frac{\partial L}{\partial g_j^{\text{syn}}} = \frac{\partial L}{\partial V_0} \left(\prod_{i=0}^{n} R_i^{\text{tot}} g_i^{\text{den}} \right) x_j (E_j - V_n) \tag{13}$$

Corollary 3 (Dendritic Parameter Gradient)**.**

$$\frac{\partial L}{\partial g_j^{\text{den}}} = \frac{\partial L}{\partial V_0} \left(\prod_{i=0}^{n} R_i^{\text{tot}} g_i^{\text{den}} \right) (V_j - V_n) \tag{14}$$

2 Local Learning Approximations

2.1 Broadcast Error Approximation

Definition 1 (Local Approximation)**.** *Replace the exact gradient $\frac{\partial L}{\partial V_n}$ with a broadcast error signal e_n derived from the output error $\delta_0 = \frac{\partial L}{\partial V_0}$:*

$$\frac{\partial L}{\partial V_n} \approx e_n, \qquad \prod_{i=0}^{n} R_i^{\text{tot}} g_i^{\text{den}} \approx 1 \tag{15}$$

Three broadcast modes are implemented (config: `error_broadcast_mode`):

(A) Scalar broadcast. For minibatch index b:

$$\bar{\delta}(b) = \frac{1}{d_{\text{out}}} \sum_{k=1}^{d_{\text{out}}} \delta_k(b), \qquad e_n(b) = \bar{\delta}(b) \mathbf{1}_{d_n} \tag{16}$$

(B) Per-compartment mapping. If $d_n = d_{\text{out}}$: $e_n(b) = \delta(b)$. Otherwise, *fallback to scalar broadcast*. An optional DFA-style mode uses a fixed random feedback matrix $B_n \in \mathbb{R}^{d_n \times d_{\text{out}}}$ sampled once at initialization: $e_n(b) = B_n \delta(b)$. This supports testing Theorem 2.

(C) Local mismatch modulation. Let $P_n(b)$ be parent compartment drive (e.g., blocklinear output). Define centered mismatch:

$$\varepsilon_n(b) = (P_n(b) - V_n(b)) - \frac{1}{B} \sum_{t=1}^{B} (P_n(t) - V_n(t)) \tag{17}$$

Then:

$$e_n(b) = \bar{\delta}(b) \varepsilon_n(b) \tag{18}$$

2.2 Gradient Alignment with Broadcast Errors

Define the exact synaptic gradient at layer n by $g^{\text{exact}} = \delta_0 \cdot \Xi_n$, where Ξ_n collects local factors and the exact path-sum (10). The local 3F gradient with broadcast error $e_n = B_n \delta_0$ is $g^{\text{local}} = e_n \cdot \widehat{\Xi}_n$, where $\widehat{\Xi}_n$ omits the path-sum.

Theorem 2 (Positive Expected Alignment under Random Broadcast)**.** *Let $B_n \in \mathbb{R}^{d_n \times d_{\text{out}}}$ have i.i.d. zero-mean entries with $\mathbb{E}[B_n^\top B_n] = \alpha I$. If the decoder aligns with the forward pathway (standard during training), then*

$$\mathbb{E}\big[\cos \angle(g^{\text{local}}, g^{\text{exact}})\big] \;\geq\; c_n > 0,$$

where c_n depends on α and the average correlation between $\widehat{\Xi}_n$ and Ξ_n. Thus g^{local} provides a descent direction in expectation.

Sketch. Adapt the feedback-alignment argument [5, 6]: fixed random feedback suffices for alignment as forward weights adapt. Here, $\widehat{\Xi}_n$ is proportional to Ξ_n up to the missing path factor; Jensen bounds on (10) yield $c_n > 0$. □

2.3 Three-Factor Rule (3F)

Definition 2 (3F Learning Rule)**.** *For synaptic conductances:*

$$\Delta g_j^{\text{syn}} = \eta \left\langle x_j R_n^{\text{tot}} (E_j - V_n) e_n \right\rangle_B \tag{19}$$

For dendritic conductances:

$$\Delta g_j^{\text{den}} = \eta \left\langle R_n^{\text{tot}} (V_j - V_n) e_n \right\rangle_B \tag{20}$$

where $\langle \cdot \rangle_B$ denotes batch average.

Remark 3. *The three factors are: (1) presynaptic activity x_j or voltage difference $(V_j - V_n)$, (2) postsynaptic modulation $(E_j - V_n)$ or R_n^{tot}, (3) broadcast error e_n.*

Symmetry note. *In implementation, the same multiplicative factors (conductance scaling R_n^{tot}, morphology ρ, information ϕ, and branch scaling s_j) are applied consistently to both excitatory and inhibitory synapses; inhibitory only differs in the driving force sign (shunting) via $-(V_n)$ when using driving-force mode.*

2.4 Four-Factor Rule (4F): Morphology Correlation

Definition 3 (Morphology Factor)**.** *Let $\bar{V}_n = \frac{1}{d_n} \sum_{j=1}^{d_n} V_{n,j}$ be the mean voltage over compartments in layer n. Define the correlation with output:*

$$\rho_n = \frac{\text{Cov}(\bar{V}_n, \bar{V}_0)}{\sqrt{\text{Var}(\bar{V}_n)\,\text{Var}(\bar{V}_0)} + \varepsilon} \tag{21}$$

Proposition 6 (4F Update Rule)**.** *Multiply 3F updates by ρ_n:*

$$\Delta g_j^{\text{syn}} = \eta \rho_n \left\langle x_j R_n^{\text{tot}} (E_j - V_n) e_n \right\rangle_B \tag{22}$$

$$\Delta g_j^{\text{den}} = \eta \rho_n \left\langle R_n^{\text{tot}} (V_j - V_n) e_n \right\rangle_B \tag{23}$$

Theorem 3 (Theoretical Justification)**.** *Let L be a smooth loss. Under the assumption that layer n contributes to the output primarily through its mean activity, the correlation ρ_n approximates the alignment between local voltage fluctuations and output gradients:*

$$\mathbb{E}\left[\frac{\partial L}{\partial \bar{V}_n} \cdot \bar{V}_n\right] \propto \rho_n \cdot \text{Var}(\bar{V}_n) \tag{24}$$

Thus ρ_n weights updates by the layer's relevance to the task.

Estimators (EMA / online). For minibatches $B \geq 2$, estimate ρ_n from (21) with an EMA over batches. For $B = 1$ (online), maintain means μ_x, μ_y, variances σ_x^2, σ_y^2, and covariance C_{xy} for $x_t = \bar{V}_0^{(t)}$ and $y_t = \bar{V}_n^{(t)}$ using Welford's numerically stable algorithm [16]:

$$\begin{aligned}
\mu_x^{(t)} &= (1-\alpha)\mu_x^{(t-1)} + \alpha x_t, & \mu_y^{(t)} &= (1-\alpha)\mu_y^{(t-1)} + \alpha y_t, \\
\delta_x &= x_t - \mu_x^{(t-1)}, & \delta_y &= y_t - \mu_y^{(t-1)}, \\
\sigma_x^{2(t)} &= (1-\alpha)\sigma_x^{2(t-1)} + \alpha\,\delta_x^2, & \sigma_y^{2(t)} &= (1-\alpha)\sigma_y^{2(t-1)} + \alpha\,\delta_y^2, \\
C_{xy}^{(t)} &= (1-\alpha)C_{xy}^{(t-1)} + \alpha\,\delta_x\delta_y.
\end{aligned} \tag{25}$$

Then $\rho_n^{(t)} = \frac{C_{xy}^{(t)}}{\sqrt{\sigma_x^{2(t)}\sigma_y^{2(t)}}+\varepsilon}$, where α is the EMA rate (`ema_alpha`).

2.5 Five-Factor Rule (5F): Conditional Information

Definition 4 (Conditional Predictability Factor)**.** *Let P_n be parent compartment voltage. Define the conditional variance via ridge regression:*

$$\beta_n = \frac{\text{Cov}(V_n, P_n)}{\text{Var}(P_n) + \lambda} \tag{26}$$

$$\sigma_{\text{res}}^2 = \text{Var}(V_n) - \beta_n \,\text{Cov}(V_n, P_n) \tag{27}$$

The information proxy is:

$$\phi_n = \frac{\mathrm{Var}(V_n)}{\sigma_{\mathrm{res}}^2 + \varepsilon} = \frac{1}{1 - R_n^2} \geq 1, \tag{28}$$

where $R_n^2 = \frac{\beta_n \, \mathrm{Cov}(V_n, P_n)}{\mathrm{Var}(V_n)}$ *is the (ridge) coefficient of determination.*

Remark 4 (Information-Theoretic Interpretation and Implementation)**.** *ϕ_n increases when V_n is* more *predictable from its parent P_n (higher R^2), providing an amplification factor. This formulation $\phi_n = 1/(1 - R^2)$ is used in the current implementation (`_compute_layer_phi_conditional`, line 1261) and is clamped to $[0.25, 4.0]$ for stability.*

Alternative formulation*: To emphasize* unique *variance (information beyond the parent), one could instead use $\phi_n = 1 - R^2 \in (0, 1]$, which decreases when V_n is predictable from P_n. The current implementation uses the inverse formulation, treating high predictability as indicative of strong signal propagation through the parent pathway. Both interpretations are valid depending on the desired emphasis: the current form amplifies well-predicted compartments (coherent signal flow), while the alternative would amplify compartments with unique information.*

Proposition 7 (5F Update Rule)**.**

$$\Delta g_j^{\mathrm{syn}} = \eta \rho_n \phi_n \left\langle x_j R_n^{\mathrm{tot}} (E_j - V_n) e_n \right\rangle_B \tag{29}$$

$$\Delta g_j^{\mathrm{den}} = \eta \rho_n \phi_n \left\langle R_n^{\mathrm{tot}} (V_j - V_n) e_n \right\rangle_B \tag{30}$$

3 Morphology-Aware Extensions

Standard 4F/5F rules use layer-wise factors ρ_n, ϕ_n that ignore branch-specific topology. We introduce four extensions that explicitly incorporate dendritic tree structure.

3.1 Path-Integrated Propagation

Definition 5 (Path Factor)**.** *Define recursively:*

$$\pi_n = \begin{cases} 1 & n = 0 \\ \pi_{n-1} \cdot R_{n-1}^{\mathrm{tot}} \cdot \bar{g}_{n-1}^{\mathrm{den}} & n \geq 1 \end{cases} \tag{31}$$

where $\bar{g}_{n-1}^{\mathrm{den}}$ is the mean dendritic conductance from layer $n-1$ to n.

Proposition 8 (Exact for Chains)**.** *For a chain morphology (single path), the path factor* (31) *satisfies*

$$\pi_n = \prod_{i=0}^{n-1} R_i^{\mathrm{tot}} g_i^{\mathrm{den}} \tag{32}$$

and thus exactly matches (10)*.*

Remark 5 (Sandwich bounds for trees)**.** *For a tree, let $\mathcal{P}$ be the set of directed paths from n to 0 and define, at each depth d, $m_d := \min_{\mathcal{P}} a_{d,\mathcal{P}}$ and $M_d := \max_{\mathcal{P}} a_{d,\mathcal{P}}$, where $a_{d,\mathcal{P}}$ is the edge factor $R_k^{\mathrm{tot}} g_{i \to k}^{\mathrm{den}}$ at depth d along path $\mathcal{P}$. Then*

$$|\mathcal{P}| \prod_d m_d \;\leq\; \sum_{\mathcal{P}} \prod_d a_{d,\mathcal{P}} \;\leq\; |\mathcal{P}| \prod_d M_d.$$

If per-depth factors are narrowly concentrated ($m_d \approx M_d$), replacing the sum by a product of per-depth means (our π_n) is accurate. Modulating the error by π_n yields $e_n^{\mathrm{path}} = e_n \cdot \pi_n$, which better approximates exact backpropagation.

Corollary 4 (Depth Attenuation)**.** *From Lemma 1, $R_k^{\mathrm{tot}} g_{i \to k}^{\mathrm{den}} < 1$. Therefore any product $\prod Rg$ in* (10) *decays exponentially with depth, motivating path-based error attenuation.*

Remark 6 (Implementation)**.** *Computed via `_compute_path_propagation_factor()` and implemented as a* per-sample scalar *path factor $\pi_n(b) \in \mathbb{R}$ (broadcast to all compartments in layer n). The factor $\bar{g}_{n-1}^{\mathrm{den}}$ is the* arithmetic mean *over outgoing dendritic connections at depth $n-1$. This stabilizes shapes across layers and matches the code behavior when applying $e_n \leftarrow e_n \cdot \pi_n$.*

Theoretical effect. Path propagation introduces depth-dependent attenuation: deeper compartments receive exponentially smaller error signals $e_n \sim \prod R_i g_i$. This encourages specialization: shallow layers learn direct input-output mappings, while deep layers integrate over longer paths. In practice, we use a *per-sample scalar* path factor $\pi_n(b)$ for stability and consistent broadcasting across layers with different widths.

3.2 Branch-Specific Depth Modulation

Definition 6 (Depth-Modulated Morphology Factor)**.** *Let d_j be the graph distance (number of edges) from the soma to branch j. Define per-branch morphology:*

$$\rho_j = \frac{\rho_{\text{base}}}{d_j + \alpha} \tag{33}$$

where $\alpha > 0$ (`morphology_depth_offset`) prevents singularity.

Proposition 9 (Depth-Modulated Updates)**.** *Replace scalar ρ_n with tensor $\boldsymbol{\rho}_n \in \mathbb{R}^{d_n}$ in updates:*

$$\Delta g_j^{\text{syn}} = \eta \rho_j \phi_n \left\langle x_j R_n^{\text{tot}} (E_j - V_n) e_n \right\rangle_B \tag{34}$$

Theorem 4 (Biological Motivation)**.** *In real dendrites, distal synapses (large d_j) contribute less to somatic depolarization due to cable attenuation. The scaling $\rho_j \propto 1/d_j$ mirrors this: deeper branches receive smaller plasticity updates.*

Theoretical effect. Depth modulation biases learning toward proximal synapses. For fixed error e_n, the gradient magnitude is:

$$\left\|\Delta g_j^{\text{syn}}\right\| \propto \frac{1}{d_j + \alpha} \tag{35}$$

Tuning α controls the depth penalty: small α strong penalty, large α mild penalty.

3.3 Dendritic Normalization

Definition 7 (Conductance Normalization)**.** *For dendritic updates, normalize by total branch conductance:*

$$\Delta g_j^{\text{den}} \leftarrow \frac{\Delta g_j^{\text{den}}}{\sum_{k=1}^{K_n} g_k^{\text{den}} + \varepsilon} \tag{36}$$

where K_n is the number of dendritic inputs to compartment n.

Theorem 5 (Variance Stabilization)**.** *Let $G_n = \sum_k g_k^{\text{den}}$ and assume $\Delta g_k \sim \mathcal{N}(0, \sigma^2)$. Without normalization:*

$$\text{Var}\left(\sum_k \Delta g_k\right) = K_n \sigma^2 \tag{37}$$

With normalization:

$$\text{Var}\left(\sum_k \frac{\Delta g_k}{G_n}\right) = \frac{K_n \sigma^2}{G_n^2} \tag{38}$$

Thus normalization reduces variance when G_n is large, preventing dominant branches from accumulating unbounded updates.

Theoretical effect. Analogous to batch normalization, dendritic normalization balances contributions across branches. In sparse connectivity (e.g., TopK synapses), some branches may have much higher G_n than others. Normalization equalizes their influence on the compartment voltage. This mechanism is consistent with homeostatic synaptic scaling observed in biology [17], where neurons adjust synaptic strengths to maintain stable activity levels, and relates to Oja-style stability rules that prevent unbounded weight growth.

3.4 Apical vs Basal Branch Differentiation

Definition 8 (Branch Type Scaling)**.** *Assign each branch j a type flag $t_j \in \{0, 1\}$ (0 = basal, 1 = apical). Define type-specific scales:*

$$s_j = s_{\text{basal}} + t_j(s_{\text{apical}} - s_{\text{basal}}) \tag{39}$$

Apply to all update factors:

$$\Delta g_j^{\text{syn}} = \eta s_j \rho_j \phi_n \left\langle x_j R_n^{\text{tot}} (E_j - V_n) e_n \right\rangle_B \tag{40}$$

Theorem 6 (Compartmental Specialization)**.** *Pyramidal neurons exhibit distinct plasticity rules in apical (layer 1, feedback) vs basal (layer 5, feedforward) dendrites. Setting $s_{\text{apical}} > s_{\text{basal}}$ amplifies top-down learning, while $s_{\text{apical}} < s_{\text{basal}}$ emphasizes bottom-up processing.*

Theoretical effect. For hierarchical tasks, apical amplification ($s_{\text{apical}} = 1.5$, $s_{\text{basal}} = 1.0$) allows the network to prioritize contextual modulation. The gradient ratio is:

$$\frac{\|\Delta g_{\text{apical}}\|}{\|\Delta g_{\text{basal}}\|} = \frac{s_{\text{apical}}}{s_{\text{basal}}} \tag{41}$$

4 Auxiliary Objectives: HSIC

4.1 Hilbert-Schmidt Independence Criterion

For layer activations $\mathbf{Z} \in \mathbb{R}^{B \times d_n}$, define kernel matrix $\mathbf{K}_Z$ (linear, RBF, or polynomial) and centering matrix $\mathbf{H} = \mathbf{I} - \frac{1}{B}\mathbf{1}\mathbf{1}^\top$.

Definition 9 (HSIC Loss)**.** *Self-decorrelation (maximize independence within layer):*

$$\mathcal{L}_{\text{HSIC}}^{\text{self}} = \frac{1}{B^2} \operatorname{tr}(\mathbf{K}_Z \mathbf{H} \mathbf{K}_Z \mathbf{H}) \tag{42}$$

Target-correlation (align with labels $\mathbf{Y}$):

$$\mathcal{L}_{\text{HSIC}}^{\text{target}} = -\frac{1}{B^2} \operatorname{tr}(\mathbf{K}_Z \mathbf{H} \mathbf{K}_Y \mathbf{H}) \tag{43}$$

Proposition 10 (Linear Kernel Gradients)**.** *For $\mathbf{K}_Z = \mathbf{Z}\mathbf{Z}^\top$:*

$$\frac{\partial \mathcal{L}_{\text{HSIC}}^{\text{self}}}{\partial \mathbf{Z}} = \frac{4}{B^2} \mathbf{H} \mathbf{K}_Z \mathbf{H} \mathbf{Z} \tag{44}$$

$$\frac{\partial \mathcal{L}_{\text{HSIC}}^{\text{target}}}{\partial \mathbf{Z}} = -\frac{4}{B^2} \mathbf{H} \mathbf{K}_Y \mathbf{H} \mathbf{Z} \tag{45}$$

Gradients are added to synaptic eligibility traces via chain rule through $\mathbf{Z} = f(\mathbf{g}^{\text{syn}})$.

Remark 7. *The expression for $\partial \mathcal{L}_{\text{HSIC}}^{\text{target}} / \partial \mathbf{Z}$ assumes $\mathbf{K}_Y$ is symmetric (true for standard kernels).*

5 Implementation Details

5.1 Units and Normalization

Quantity	Symbol	Typical units (scaled)
Voltage	V	mV (normalized to $[-1, 1]$)
Synaptic conductance	g^{syn}	nS (nonnegative)
Dendritic conductance	g^{den}	nS (nonnegative)
Leak conductance	g^{leak}	nS (set to 1 in normalized units)
Input resistance	R^{tot}	nS^{-1} (normalized ≤ 1)

Table 1: Units and normalization conventions.

5.2 Positive Weight Parameterization

To enforce $g \geq 0$, we can use exponential:

$$g = \exp(\theta), \quad \theta \in \mathbb{R} \tag{46}$$

Chain rule for gradients:

$$\frac{\partial L}{\partial \theta} = \frac{\partial L}{\partial g} \cdot g \tag{47}$$

Alternatively, use softplus $g = \log(1 + \exp(\theta))$ to avoid extreme gradients.

5.3 Online Variant with Eligibility Traces

Define continuous-time eligibilities per synapse:

$$\tau_e \dot{e}_j^{\text{syn}}(t) = -e_j^{\text{syn}}(t) + x_j(t)\,(E_j - V_n(t))\,R_n^{\text{tot}}(t),$$

and likewise for dendritic connections with $(V_j(t) - V_n(t))\,R_n^{\text{tot}}(t)$. Let the modulatory/error signal be $m_n(t)$ (e.g., broadcast from output or neuromodulatory). Then

$$\Delta g_j^{\text{syn}} \propto \int e_j^{\text{syn}}(t)\, m_n(t)\, \mathrm{d}t, \qquad \Delta g_j^{\text{den}} \propto \int e_j^{\text{den}}(t)\, m_n(t)\, \mathrm{d}t,$$

which instantiates three-factor learning in continuous time [13, 14].

5.4 Decoder Update Modes

Let W_{dec} map $V_L \to y \in \mathbb{R}^{d_{\text{out}}}$. Three modes:

1. **Backprop**: $\nabla_{W_{\text{dec}}} L$ via autograd.
2. **Local**: $\Delta W_{\text{dec}} = \eta \langle \delta_0 V_L^\top \rangle_B$ (3-factor).
3. **Frozen**: $\Delta W_{\text{dec}} = 0$.

5.5 Algorithm Summary

Algorithm 1 Local Credit Assignment with Morphology-Aware Extensions

Input: Model, minibatch (x, y), config $\mathcal{C}$
Forward pass: $\hat{y} = f(x; \mathbf{g}^{\text{syn}}, \mathbf{g}^{\text{den}})$
Compute loss L and output error $\delta^y = \frac{\partial L}{\partial \hat{y}}$
Compute somatic error $\delta_0 = W_{\text{dec}}^\top \delta^y$
for each layer n (reverse order) **do**
  Broadcast error: $e_n = \text{broadcast}(\delta_0, \mathcal{C})$
  if path propagation enabled **then**
    Compute π_n via (31); $e_n \leftarrow e_n \cdot \pi_n$
  end if
  Compute ρ_n via (21)
  if depth modulation enabled **then**
    $\rho_n \leftarrow [\rho_1, \ldots, \rho_{d_n}]$ via (33)
  end if
  Compute ϕ_n via (28)
  Compute branch scales s_j via (39)
  Synaptic updates: $\Delta g_j^{\text{syn}} = \eta s_j \rho_j \phi_n \langle x_j R_n^{\text{tot}} (E_j - V_n) e_n \rangle_B$
  Dendritic updates: $\Delta g_j^{\text{den}} = \eta s_j \rho_j \phi_n \langle R_n^{\text{tot}} (V_j - V_n) e_n \rangle_B$
  if dendritic normalization enabled **then**
    $\Delta g_j^{\text{den}} \leftarrow \Delta g_j^{\text{den}} / (G_n + \varepsilon)$ via (36)
  end if
  if HSIC enabled **then**
    Add HSIC gradients to Δg_j^{syn}
  end if
end for
Clip gradients; optimizer step

6 Configuration Reference

Section	Key Parameters
Core	`rule_variant`, `error_mode`, `error_broadcast_mode`
3F (`three_factor`)	`use_conductance_scaling`, `use_driving_force`, θ, E^{rev}
4F (`four_factor`)	`rho_mode`, `rho_estimator`, `ema_alpha`, `layer_wise_rho_scale`
5F (`five_factor`)	`phi_mode`, `phi_estimator`, `phi_ridge_lambda`, `layer_wise_phi_scale`
Morphology-aware	`use_path_propagation`, `morphology_modulator_mode`, `morphology_depth_offset`, `use_dendritic_normali`
HSIC (`hsic`)	`enabled`, `weight`, `self_weight`, `target_weight`, `kernel`, `sigma`

Table 2: Configuration grouped by learning rule sections (see `LocalRuleConfig`).

7 Theoretical Comparison

Method	Factors	Topology	Complexity
3F	$x, (E - V), e$	Layer-wise	$\mathcal{O}(1)$
4F	3F $+\rho$	Layer-wise	$\mathcal{O}(1)$
5F	4F $+\phi$	Layer-wise	$\mathcal{O}(d_n)$
5F + Path	5F $+\pi$	Path-aware	$\mathcal{O}(L)$
5F + Depth	5F, $\rho \to \rho_j$	Branch-aware	$\mathcal{O}(d_n)$
5F + Norm	5F + normalization	Branch-aware	$\mathcal{O}(d_n)$
5F + Types	5F $\times s_j$	Compartment-aware	$\mathcal{O}(1)$

Table 3: Computational complexity per update (L = depth, d_n = compartments).

Component	Biological Analog	Key Result
Conductance scaling R_n^{tot}	Input resistance	Lemma 1: $0 < R_n^{\text{tot}} \leq 1$
Driving force $(E_j - V_n)$	Synaptic current	Prop. 1: Local sensitivity
Shunting inhibition	Divisive normalization	Sec. 1.4: $\partial V / \partial g_{\text{inh}} \propto -V$
Path factor π_n	Cable attenuation	Prop. 8: Depth decay
Morphology factor ρ_n	Layer correlation	Eq. (21): Task relevance
Information factor ϕ_n	Conditional predictability	Eq. (28): $1/(1 - R^2)$
Dendritic normalization	Homeostatic scaling	Sec. 4.3: Variance stabilization
Branch-type scaling	Apical vs. basal	Sec. 4.4: Compartment specialization
Broadcast alignment	Feedback alignment	Thm. 2: $\mathbb{E}[\cos \angle] > 0$

Table 4: Summary of theoretical components and their biological/algorithmic interpretations.

8 Experimental Validation for NeurIPS

8.1 Central Empirical Claim

The main empirical claim is not merely that local learning can train dendritic networks, but that *dendritic tree structure and shunting conductance interactions create useful local credit signals*. Concretely, the completed sweeps support:

$$\text{(per-soma error broadcast)} + \text{(dendritic shunting regime)} \implies \text{consistent gains over scalar and additive controls.} \tag{48}$$

The role of path propagation is more nuanced in our current runs: it improves several representation-level information metrics, but does not consistently increase accuracy. This claim is specific to compartmental dendritic models where child branches feed parent branches and soma through conductance-weighted interactions (10), and is therefore not reducible to standard layer-wise local rules.

8.2 Claim-Driven Hypotheses and Falsifiers

H1: Broadcast locality matters. Replacing scalar broadcast with per-compartment/per-soma mapping should improve convergence and final accuracy. **Falsifier:** no significant improvement across seeds/tasks, or unstable gains.

H2: Morphology-aware path factors matter. Enabling `use_path_propagation` should further improve performance, especially when combined with per-soma broadcast. **Falsifier:** no additive or interaction gain relative to per-soma alone.

H3: Decoder learning can remain local. Under the best local dendritic setting, `decoder_update_mode=local` should match `backprop`, while `none` should degrade. **Falsifier:** local decoder materially underperforms backprop in the same regime.

H4: The effect is regime-dependent and shunting-linked. Gains from H1–H2 should be strongest in shunting/strong-inhibition settings and weaker in additive controls. **Falsifier:** gains are equally large in additive controls or disappear under shunting.

8.3 Submission Sweep Suite

We organize experiments as four sweep families, each mapped to one hypothesis:

- **S1 (H1/H2):** $\{\texttt{error\_broadcast\_mode} \in [\texttt{scalar}, \texttt{per\_soma}]\} \times \{\texttt{use\_path\_propagation} \in [\texttt{false}, \texttt{true}]\}$ with 5F local rule.
- **S2 (H3):** $\{\texttt{decoder\_update\_mode} \in [\texttt{backprop}, \texttt{local}, \texttt{none}]\}$ under the best S1 regime.
- **S3 (H4):** $\{\texttt{core.type} \in [\texttt{dendritic\_shunting}, \texttt{dendritic\_additive}]\} \times$ inhibition-strength sweep.
- **S4 (source):** targeted analysis runs (information, ablation, weight/compartment statistics) on top and bottom S1/S3 conditions.

8.4 Completed Robust Multi-Seed Signal (Current Sweep Set)

Completed claim sweeps (`claim1--claim4`, with robust reruns for `claim2--claim4`) provide the following signal:

- **H3 supported (decoder locality):** `decoder_update_mode=local` matches or slightly exceeds `backprop` on both datasets, while `none` drops strongly. MNIST test: local 0.3790 vs. backprop 0.3788 vs. none 0.1756; CIFAR-10 test: local 0.1669 vs. backprop 0.1671 vs. none 0.1000. Paired seed-matched local-minus-backprop deltas are near zero (MNIST: +0.00018; CIFAR-10: -0.00024), while local-minus-none gaps are large (MNIST: +0.203; CIFAR-10: +0.0669).
- **H4 strongly supported (regime dependence):** shunting is consistently and substantially better than additive controls across inhibition settings. Averaged over inhibition levels in claim3, MNIST test is 0.331 (shunting, per-soma) vs. 0.112 (additive, per-soma), and CIFAR-10 test is 0.145 vs. 0.100. Best matched settings show the same pattern (MNIST: 0.366 vs. 0.116; CIFAR-10: 0.165 vs. 0.100).
- **H1 partially supported and interaction-dependent**: per-soma broadcast improves over scalar in the main shunting regime on MNIST (mean test 0.331 vs. 0.311) and modestly on CIFAR-10 (0.145 vs. 0.143), but the magnitude is regime- and dataset-dependent; in lower-signal settings the gap can shrink or flip.
- **H2 mixed with a clear interaction**: `use_path_propagation=true` is not uniformly accuracy-improving. In claim4, within per-soma it leaves validation accuracy unchanged and slightly reduces test accuracy (0.3819 $\rightarrow$ 0.3790), while increasing information metrics ($\mathrm{MI}(E, I; C)$: 0.620 $\rightarrow$ 0.637, $\mathrm{MI}(V; C)$: 0.436 $\rightarrow$ 0.442). Within scalar broadcast, path propagation improves test accuracy (0.351 $\rightarrow$ 0.363) with smaller information gains.

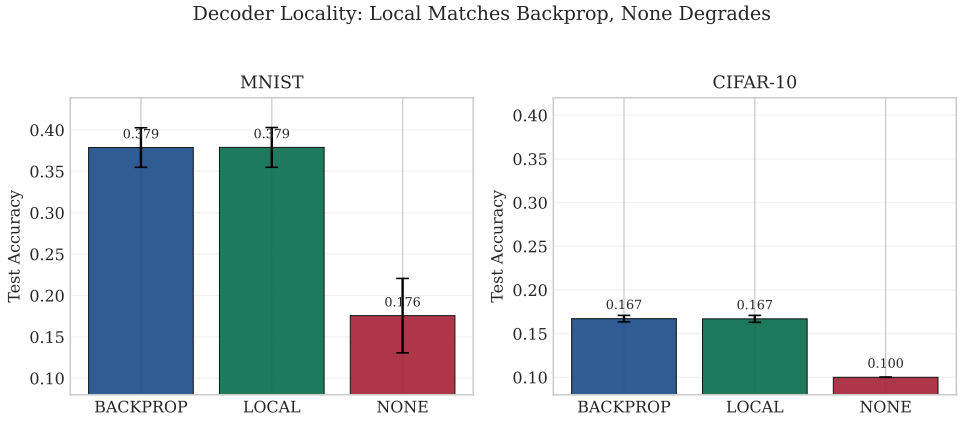

Figure 1: **Decoder-locality claim (robust sweep).** On both MNIST and CIFAR-10, local decoder updates match backpropagated decoder updates, while removing decoder updates (`none`) causes a large performance drop. This supports local decoder sufficiency in the best local-credit regime.

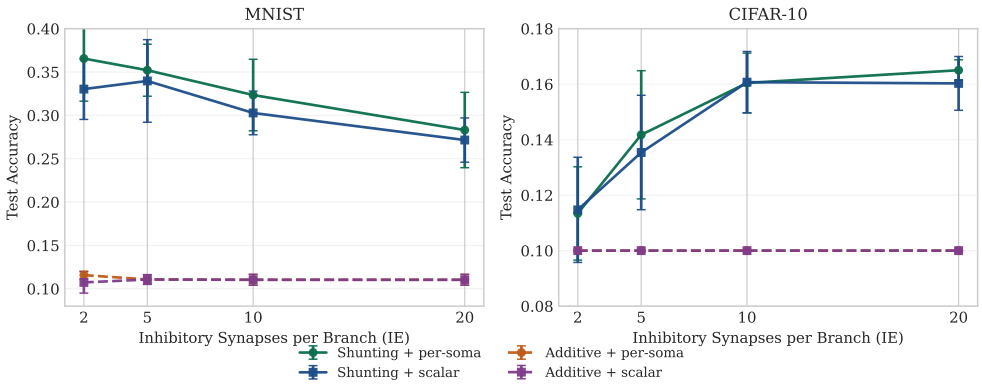

Figure 2: **Regime dependence across inhibition levels.** Shunting networks consistently outperform additive controls across inhibitory synapse counts and broadcast modes, with the largest margin on MNIST. This isolates the main performance source as the shunting regime rather than local-learning heuristics alone.

Therefore, the strongest paper narrative is: *local decoder sufficiency* + *shunting-regime dependence* as the core performance result, with *broadcast/path interactions* and *path-driven representation shaping* as the mechanistic layer.

8.5 Core Empirical Figures

8.6 Primary Metrics and Statistical Protocol

- **Primary endpoint:** test accuracy (mean $\pm$ 95% CI over seeds).
- **Secondary:** validation NLL, convergence speed (epochs to best checkpoint), robustness across MNIST/CIFAR-10.
- **Mechanistic:** MI/CMI terms, branch/path statistics, compartment SNR, ablation sensitivity.
- **Statistics:** two-way ANOVA (broadcast $\times$ path), interaction contrasts, and paired seed-matched deltas for local-vs-backprop decoder updates.

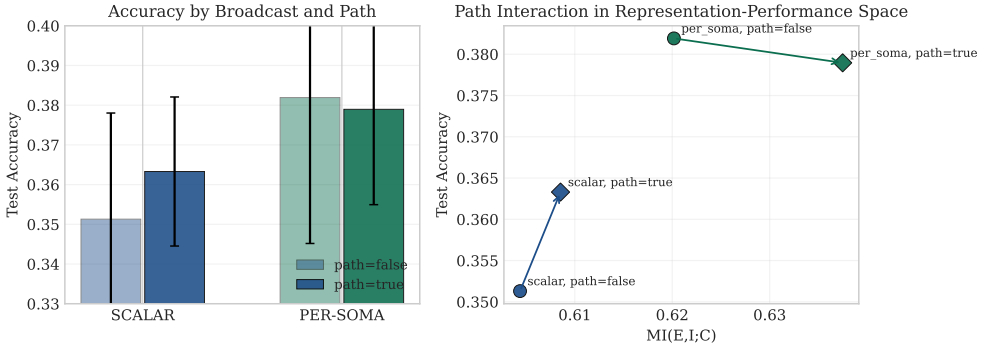

Figure 3: **Broadcast-path interaction in source analysis.** Path propagation has interaction-dependent behavior: within per-soma broadcast it leaves accuracy nearly unchanged while increasing information metrics, whereas within scalar broadcast it improves accuracy more with smaller information gains.

Claim / Condition	**Metric**	**Value**
Claim2 (MNIST, local vs backprop vs none)	test accuracy	0.3790 vs 0.3788 vs 0.1756
Claim2 (CIFAR-10, local vs backprop vs none)	test accuracy	0.1669 vs 0.1671 vs 0.1000
Claim3 (MNIST, shunting-per-soma vs additive-per-soma)	mean test over IE	0.331 vs 0.112
Claim3 (CIFAR-10, shunting-per-soma vs additive-per-soma)	mean test over IE	0.145 vs 0.100
Claim4 (per-soma, path false $\rightarrow$ true)	test / $\mathrm{MI}(E, I; C)$	0.3819 $\rightarrow$ 0.3790 / 0.620 $\rightarrow$ 0.637

Table 5: **Robust headline results for the NeurIPS narrative.** Values are grouped over robust multi-seed sweeps (`claim2--claim4`).

8.7 Figure Plan

1. **F1:** Model schematic and local rule decomposition (3F/4F/5F + morphology factors).
2. **F2:** Broadcast $\times$ path interaction heatmap with seed CIs, explicitly separating accuracy and representation-level metrics.
3. **F3:** Decoder-locality result (`local` vs. `backprop` vs. `none`).
4. **F4:** Regime dependence (shunting vs additive, inhibition strength sweep).
5. **F5:** Mechanistic source panel (information/ablation/compartment metrics), highlighting path-driven MI gains with interaction-dependent accuracy shifts.
6. **F6:** Training dynamics and factor trajectories (ρ_n, ϕ_n, π_n) in best vs control conditions.

9 Future Extensions and Open Questions

Several directions remain for strengthening this framework:

Information factor variants. The current implementation uses $\phi_n = 1/(1 - R^2)$, amplifying well-predicted compartments. An alternative $\phi_n = 1 - R^2$ would emphasize unique information. A conditional HSIC formulation could provide:

$$\phi_n^{\mathrm{cond}} = \frac{\mathrm{HSIC}(V_n, y) - \kappa\,\mathrm{HSIC}(P_n, y)}{\mathrm{HSIC}(V_n, y) + \varepsilon},$$

with $\kappa \in [0, 1]$ controlling parent discount. Empirical comparison of these variants on tasks requiring novelty detection vs. hierarchical consistency would clarify when each is advantageous.

Spiking neural networks. Extending to conductance-based LIF neurons with surrogate gradients would demonstrate biological plausibility. The eligibility trace formulation (Section 5) provides a natural bridge to event-driven learning.

Reconstructed morphologies. Testing on realistic dendritic trees from NeuroMorpho would validate the morphology-aware factors on biologically constrained topologies, particularly the depth modulation and branch-type differentiation.

Convergence analysis. For linear decoders and quadratic loss, the scalar broadcast (mode A) yields an unbiased descent direction up to a positive scalar (Theorem 2). Formal convergence rates under Robbins-Monro conditions for diminishing step sizes remain to be established.

10 Related Work

Our rules connect to dendritic credit-assignment via apical errors [7, 8] and predictive coding or equilibrium propagation as local-gradient mechanisms [9, 10]. Broadcast-error variants relate to feedback alignment and direct feedback alignment (DFA) [5, 6]. Our HSIC-based auxiliary losses follow kernel independence measures [11, 12]. Shunting inhibition connects to divisive normalization [3] and the nuanced rate-level consequences of shunts [4]. The compartmental specialization between apical and basal branches relates to empirical findings on layer-specific plasticity rules [15]. The dendritic normalization mechanism parallels homeostatic synaptic scaling [17].

11 Conclusion

We have presented a rigorous mathematical framework for local credit assignment in compartmental dendritic networks. Starting from the passive cable equation and deriving exact backpropagation gradients for dendritic trees (10), we introduced three classes of local approximations (3F/4F/5F) and extended them with four morphology-aware mechanisms that explicitly exploit dendritic tree topology. Theoretical analysis reveals that each componentpath propagation, depth modulation, dendritic normalization, and branch type differentiationaddresses specific limitations of layer-wise approximations. We clarified the role of shunting inhibition in divisive gain control and proved positive expected alignment between local and exact gradients under broadcast error schemes. All methods are implemented in `local_learning.py` with full configurability and consistency with the mathematical derivations presented here.

A Codebase Mapping

Equation/Concept	Implementation
(3)	`DendriticBranchLayer.forward()`
(5), (7)	Lines 488–639 (eligibility traces)
(19), (20)	Lines 535–584 (3F updates)
(21)	`_compute_layer_rho()` (lines 857–1018)
(28)	`_compute_layer_phi_conditional()` (lines 1039–1146)
(31)	`_compute_path_propagation_factor()` (lines 1153–1209)
(33)	`_compute_branch_depth_modulator()` (lines 1211–1236)
(36)	`_compute_dendritic_normalization()` (lines 1238–1266)
(39)	`_get_branch_type_scale()` (lines 1268–1290)
HSIC	Lines 723–854

B Synapse Count Optima from Current Optima

We summarize how to translate an optimal current ratio $r^* = (I_E/I_I)^*$ into an optimal synapse count ratio $(N_e/N_i)^*$ under two commonly used biological/engineering constraints.

(A) Fixed weight ratio $\gamma = w_e/w_i$. Current balance constraint $N_e w_e = r^* N_i w_i$ yields

$$\boxed{\left(\frac{N_e}{N_i}\right)^* = \frac{r^*}{\gamma}}. \tag{49}$$

Examples: balance ($r^* = 1$) $\Rightarrow$ $(N_e/N_i)^*=1/\gamma$; Fisher-optimal ($r^*=(\sigma_I/\sigma_E)^2$) $\Rightarrow$ $(N_e/N_i)^*=(\sigma_I/\sigma_E)^2/\gamma$.

(B) Mean-field scaling $w \propto 1/\sqrt{N}$ with equal constants. With $w_e = c/\sqrt{N_e}$ and $w_i = c/\sqrt{N_i}$ (to maintain $O(1)$ variances), currents are $I_E = c\sqrt{N_e}\bar{\mu}$ and $I_I = c\sqrt{N_i}\bar{\mu}$. Enforcing $I_E/I_I = r^*$ gives

$$\boxed{\left(\frac{N_e}{N_i}\right)^* = (r^*)^2}. \tag{50}$$

Example: Fisher-optimal $r^*=(\sigma_I/\sigma_E)^2 \Rightarrow (N_e/N_i)^*=(\sigma_I/\sigma_E)^4$.

These formulas apply uniformly to all cases discussed (balance, noise asymmetry, signal asymmetry, correlation corrections), by substituting the corresponding r^*.

C Example Configuration

```
local_ca:
  rule_variant: "5f"
  error_broadcast_mode: "scalar"

  # Morphology factor
  rho_mode: "pearson"
  rho_estimator: "ema"
  ema_alpha: 0.05

  # Information factor
  phi_mode: "conditional"
  phi_estimator: "conditional_ema"
  phi_ridge_lambda: 0.001

  # Morphology-aware extensions
  use_path_propagation: true
  morphology_modulator_mode: "depth"
  morphology_depth_offset: 2.0
  use_dendritic_normalization: true
  use_branch_type_rules: true
  apical_branch_scale: 1.5
  basal_branch_scale: 1.0

  # Compartmental
  use_conductance_scaling: true
  use_driving_force: true
  e_rev_exc: 1.0

  # Optimization
  clip_grad_value: 5.0
  normalize_by_batch: true
```

References

[1] Koch, C. (1999). *Biophysics of Computation: Information Processing in Single Neurons*. Oxford University Press.

[2] Dayan, P., & Abbott, L. F. (2001). *Theoretical Neuroscience*. MIT Press.

[3] Carandini, M., & Heeger, D. J. (2012). Normalization as a canonical neural computation. *Nature Reviews Neuroscience*, 13(1), 51–62.

[4] Holt, G. R., & Koch, C. (1997). Shunting inhibition does not have a divisive effect on firing rates. *Neural Computation*, 9(5), 1001–1013.

[5] Lillicrap, T. P., Cownden, D., Tweed, D. B., & Akerman, C. J. (2016). Random synaptic feedback weights support error backpropagation for deep learning. *Nature Communications*, 7, 13276.

[6] Nøkland, A. (2016). Direct feedback alignment provides learning in deep neural networks. *Advances in Neural Information Processing Systems*, 29.

[7] Guerguiev, J., Lillicrap, T. P., & Richards, B. A. (2017). Towards deep learning with segregated dendrites. *eLife*, 6, e22901.

[8] Sacramento, J., Costa, R. P., Bengio, Y., & Senn, W. (2018). Dendritic cortical microcircuits approximate the backpropagation algorithm. *Advances in Neural Information Processing Systems*, 31.

[9] Whittington, J. C., & Bogacz, R. (2019). Theories of error back-propagation in the brain. *Trends in Cognitive Sciences*, 23(3), 235–250.

[10] Scellier, B., & Bengio, Y. (2017). Equilibrium propagation: Bridging the gap between energy-based models and backpropagation. *Frontiers in Computational Neuroscience*, 11, 24.

[11] Gretton, A., Bousquet, O., Smola, A., & Schölkopf, B. (2005). Measuring statistical dependence with Hilbert-Schmidt norms. *International Conference on Algorithmic Learning Theory*, 63–77.

[12] Gretton, A., Fukumizu, K., Teo, C. H., Song, L., Schölkopf, B., & Smola, A. J. (2007). A kernel statistical test of independence. *Advances in Neural Information Processing Systems*, 20.

[13] Frémaux, N., & Gerstner, W. (2016). Neuromodulated spike-timing-dependent plasticity, and theory of three-factor learning rules. *Frontiers in Neural Circuits*, 9, 85.

[14] Bellec, G., Scherr, F., Subramoney, A., Hajek, E., Salaj, D., Legenstein, R., & Maass, W. (2020). A solution to the learning dilemma for recurrent networks of spiking neurons. *Nature Communications*, 11, 3625.

[15] Larkum, M. (2013). A cellular mechanism for cortical associations: an organizing principle for the cerebral cortex. *Trends in Neurosciences*, 36(3), 141–151.

[16] Welford, B. P. (1962). Note on a method for calculating corrected sums of squares and products. *Technometrics*, 4(3), 419–420.

[17] Turrigiano, G. G. (2008). The self-tuning neuron: synaptic scaling of excitatory synapses. *Cell*, 135(3), 422–435.