Memory by Design: Probabilistic Sequence Layers

Dense uncertainty as a recurrent memory state

Paper: Matthew Dowling, Hyungju Jeon,
Cristina Savin, Il Memming Park

Independent paper reading

arXiv:2605.31163v3 Revised 18 August 2026

Reading: 14 September 2026 | 25–30 minutes + backup

A memory update must decide where to write

Question

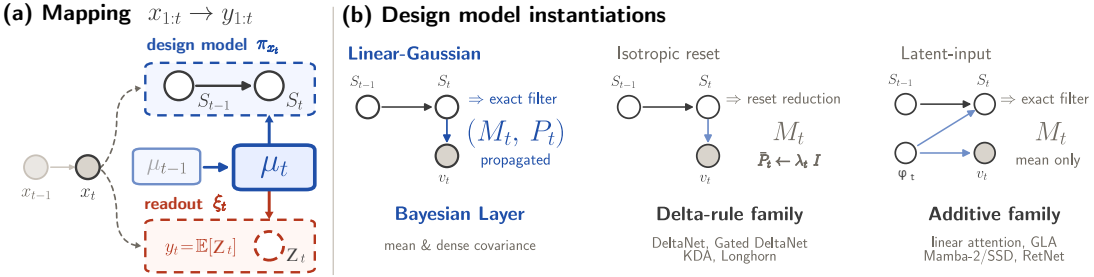
When two keys overlap, can a new association be written without repeatedly damaging an established one?

- A mean state stores what is currently remembered.
- A covariance state records which directions are resolved.
- Bayesian Layer (BL) uses both to shape the next write.

Evidence: improved recall in reported settings, with extra state and language-model perplexity costs.

Paper, Secs. 1–2, pp. 1–5; Table 2, p. 9; Tables 6–7, p. 29.

The design model separates writing from reading



- Input selects an auxiliary probabilistic memory model.
- Filtering updates a belief; a separate query produces the output.
- The input sequence itself has no likelihood in this design model.

Original Fig. 1, p. 2; Sec. 2, pp. 2–3.

A Gaussian belief makes memory updates tractable

$$\mathrm{vec}(\mathbf{S}_t) \sim \mathcal{N}\big(\mathrm{vec}(\mathbf{M}_t), \mathbf{I}_m \otimes \mathbf{P}_t\big)$$

Persistent state

$\mathbf{M}_t \in \mathbb{R}^{D\times m}$: stored associations.

$\mathbf{P}_t \in \mathbb{R}^{D\times D}$: shared key-space covariance.

Input-dependent quantities

Key/query: $\mathbf{k}_t, \mathbf{q}_t \in \mathbb{R}^D$.

Value: $\mathbf{v}_t \in \mathbb{R}^m$.

Dynamics $\mathbf{A}_t$; variances ℓ_t^2, r_t^2.

$$\mathbf{S}_t \mid \mathbf{S}_{t-1} \sim \mathcal{N}(\mathbf{A}_t\mathbf{S}_{t-1}, \ell_t^2\mathbf{I}), \qquad \mathbf{v}_t \mid \mathbf{S}_t \sim \mathcal{N}(\mathbf{S}_t^\top\mathbf{k}_t, r_t^2\mathbf{I}_m)$$

Sec. 2.1, Definition 1, pp. 3–4. Gaussian matrix notation follows the paper.

The innovation follows an uncertainty-shaped key

$$\bar{\mathbf{M}}_t = \mathbf{A}_t \mathbf{M}_{t-1}, \qquad \bar{\mathbf{P}}_t = \mathbf{A}_t \mathbf{P}_{t-1} \mathbf{A}_t^\top + \ell_t^2 \mathbf{I},$$

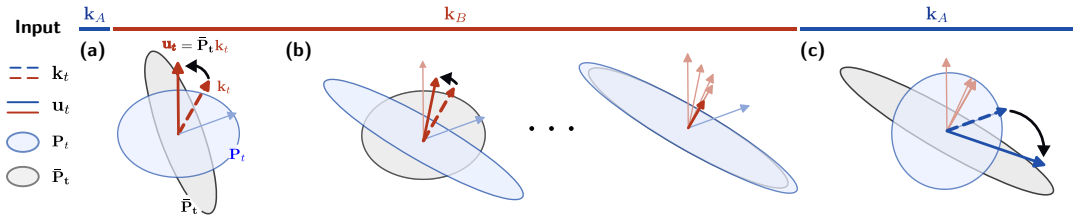
$$\mathbf{u}_t = \bar{\mathbf{P}}_t \mathbf{k}_t, \qquad \beta_t = (r_t^2 + \mathbf{k}_t^\top \mathbf{u}_t)^{-1}.$$

$$\mathbf{M}_t = \bar{\mathbf{M}}_t + \beta_t \mathbf{u}_t (\mathbf{v}_t - \bar{\mathbf{M}}_t^\top \mathbf{k}_t)^\top, \qquad \mathbf{P}_t = \bar{\mathbf{P}}_t - \beta_t \mathbf{u}_t \mathbf{u}_t^\top.$$

- Covariance determines direction, gain, and future uncertainty.
- Readout: $\mathbf{y}_t = \mathbf{M}_t^\top \mathbf{q}_t$; query does not drive the write.
- β_t is a precision; the scalar write fraction is $g_t = \beta_t \mathbf{k}_t^\top \mathbf{u}_t$.

Proposition 1, Eqs. (9)–(13), p. 4; error-correction form, Appendix B, p. 15.

Resolved directions receive less interference



$$\mathbf{k}_A = \rho \mathbf{k}_B + \sqrt{1-\rho^2}\mathbf{k}_\perp$$

$$\mathbf{u}_t \approx \rho \bar{\sigma}_B^2 \mathbf{k}_B + \sqrt{1-\rho^2}\bar{\sigma}_\perp^2 \mathbf{k}_\perp, \qquad \bar{\sigma}_B^2 \ll \bar{\sigma}_\perp^2$$

After repeated B writes, a colliding A write is directed toward the unresolved residual. This approximation assumes near-diagonal covariance in the B/residual basis.

Original Fig. 2; Eq. (15), Sec. 2.3, p. 5.

Resetting covariance recovers the Delta-rule core

Reset $\bar{\mathbf{P}}_t \leftarrow \lambda_t \mathbf{I}$ at every step:

$$\mathbf{M}_t = \mathbf{A}_t \mathbf{M}_{t-1} + \eta_t \mathbf{k}_t \big(\mathbf{v}_t - (\mathbf{A}_t \mathbf{M}_{t-1})^\top \mathbf{k}_t\big)^\top, \quad \eta_t = \frac{\lambda_t}{r_t^2 + \lambda_t \|\mathbf{k}_t\|^2}.$$

Core recurrence	Choice of dynamics
DeltaNet	$\mathbf{A}_t = \mathbf{I}$
Gated DeltaNet	$\mathbf{A}_t = \alpha_t \mathbf{I}$
KDA / Kimi Linear	$\mathbf{A}_t = \mathrm{diag}(\boldsymbol{\alpha}_t)$

The residual correction remains; covariance history is discarded. Full architectures also contain components outside this correspondence.

Table 1; Proposition 2, Eq. (17), p. 6.

Additive cores use a different observation model

$$\phi_t \sim \mathcal{N}(0, \lambda_t \mathbf{I}), \qquad \mathbf{v}_t \mid \phi_t \sim \mathcal{N}(\phi_t, r_t^2 \mathbf{I}),$$
$$\mathbf{S}_t = \mathbf{A}_t \mathbf{S}_{t-1} + \mathbf{k}_t \phi_t^\top \quad \Longrightarrow \quad \mathbf{M}_t = \mathbf{A}_t \mathbf{M}_{t-1} + \omega_t \mathbf{k}_t \mathbf{v}_t^\top.$$

- $\omega_t = \lambda_t/(\lambda_t + r_t^2)$; observation measures the *new write coefficient*.
- Dynamics recover Linear Attention, RetNet, diagonal GLA, and the SSD/Mamba-2 core.
- The mean recurrence closes without covariance feedback. This does not imply zero posterior uncertainty.

Sec. 3.2, pp. 6–7; Appendix A, pp. 14–15. Last qualification: reading-side derivation, backup.

Repeated writes predict a positive gain floor

Assume unit keys, $\mathbf{A}_t = \mathbf{I}$, constant $r, \ell > 0$, and sustained writes at A.

$$x_{t+1} = \frac{r^2 x_t}{r^2 + x_t} + \ell^2, \qquad x_t = \bar{\sigma}^2_{A,t}$$

$$x_\infty = \frac{\ell^2 + \sqrt{\ell^4 + 4r^2\ell^2}}{2}, \qquad g_\infty = \frac{x_\infty}{r^2 + x_\infty}$$

- With $r^2 = \ell^2 = 0.05$, $g_\infty \approx 0.618$; the innovation residual still vanishes.
- Cross-direction variance increment tends to $(1 - \rho^2)\ell^2$.
- These are conditional asymptotic results, not universal retention guarantees.

Appendix E.1–E.2, pp. 17–19; Eq. (42), p. 19.

Dense covariance adds state and arithmetic

One head, G noise groups

State: $O(Dm + GD^2)$.

Per token: $O(Dm + GD^2)$ with structured dynamics.

Linear in sequence length; quadratic in key dimension.

Implementation

Covariance pass produces gains.

Affine mean update uses chunkwise machinery.

Chunk boundaries remain sequential; filter arithmetic uses fp32.

Retrofit: four heads with dense 256×256 covariance blocks in each of four replaced layers. No throughput or peak-memory comparison table is reported.

Sec. 2.2, p. 4; Appendix D, pp. 16–17; Discussion, pp. 10–11.

The evaluations test different levels of the claim

Setting	Question	Repetitions
Deterministic collision	Does the algebra predict recall?	Exact trajectory
Learned collision	Does the mechanism survive training?	3 seeds
Random-key recall	Does it survive unstructured crowding?	3 seeds
MQAR variants	Can tokens form associations?	1 retained run
WikiText from scratch	Does recall help token prediction?	3 seeds
340M retrofit	Can pretrained retrieval improve?	1 seed

Matching backbone, state budget, data tokens, and actual hardware cost are different controls.

Table 3, p. 17; experimental details, Appendices E–G.

The stated collision recurrence preserves B

$D = 16$, $m = 6$, $\rho = 0.92$; 12 seed steps, 40 B writes, 60 A writes.
$\mathbf{P}_0 = 3\mathbf{I}$, $r^2 = \ell^2 = 0.05$, $\mathbf{A} = \mathbf{I}$.

Pairwise probability p^{AB}	BL	Reset/Delta	Linear Attn.
After B writes ($t = 52$)	0.73106	0.73106	≈ 1
After A writes ($t = 112$)	0.72309	0.35138	0.00054

- Pairwise probability is not full-vocabulary accuracy.
- Independent calculation agrees with the appendix values.
- v3 discrepancy: Fig. 3(a)'s BL endpoint appears near 0.60; it does not match the stated 0.72309 plateau.

Appendix E.3–E.4, pp. 20–22. Table independently recomputed from the reported setup; Fig. 3(a), p. 7.

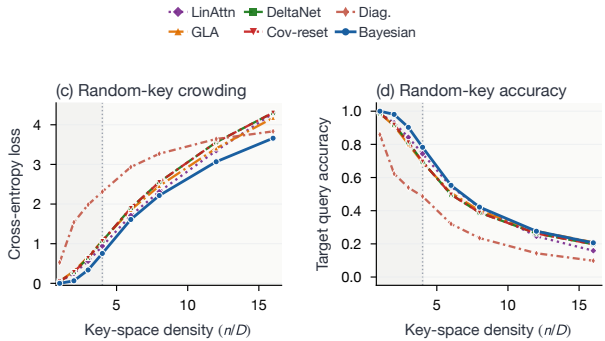
Learned recall survives stronger interference in BL



- Train: $n_f \leq 8$, $\rho \in [0.60, 0.80]$; addresses bypass learned projections.
- Left: accuracy to $n_f = 256$ at $\rho = 0.80$; right: margin at $n_f = 64$.
- v3 metric/reset descriptions conflict; read panel labels literally.

Original Fig. 3(c,d), p. 7, cropped with shared legend; Sec. 4.1, p. 8; Appendix E.5, pp. 23–26.

Random-key crowding exposes a finite memory budget

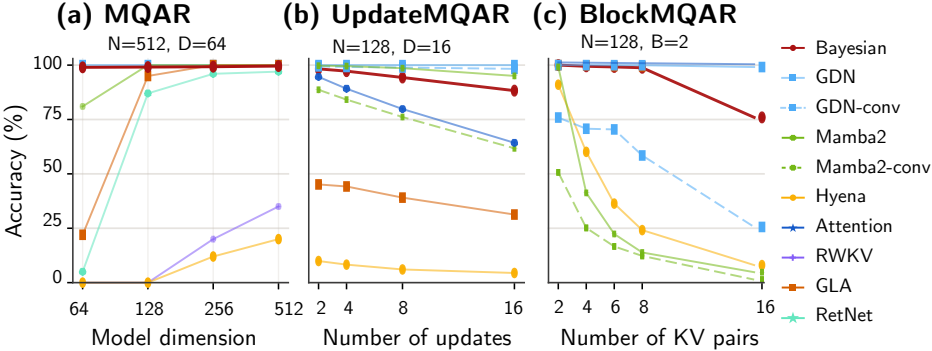


- Frozen projection: 64 raw dimensions to $D = 16$.
- Train: $n/D \leq 4$; test: $6, 8, 12, 16$.
- BL reduces loss; all methods degrade at high density.

The rank argument concerns exact linear value reconstruction. It is not a general accuracy bound for the entire multilayer classifier.

Original Fig. 7, p. 26; Appendix E.5.4, p. 25. Qualification: reading-side analysis.

MQAR gains depend on the convolution configuration



- Base: BL approaches the best curves at larger widths.
- Update at 16 writes: BL $\approx 88\%$; GDN-conv/Mamba2 $\approx 60\%$.
- Conv-free GDN remains higher ($\approx 98\%$) and also leads BL on Block-MQAR at 16 pairs.

Original Fig. 4, p. 8; Sec. 4.2, pp. 8–9; Appendix F, pp. 26–27. Approximate curve readings.

Recall improves while overall perplexity rises

Model	Overall PPL ↓	AR-hit PPL ↓	Other PPL ↓
BL	20.01 ± 0.15	$\mathbf{18.15 \pm 1.08}$	20.03 ± 0.15
GDN	$\mathbf{18.75 \pm 0.07}$	23.72 ± 1.76	$\mathbf{18.69 \pm 0.08}$
Mamba-2	21.36 ± 0.27	424.8 ± 120.4	20.58 ± 0.19

- WikiText-103; matched recurrent-state budgets; 520M training tokens; 3 seeds.
- AR-hit: pair count ≤ 16 in the 130M-token corpus; most recent repeat within 1,024 tokens in the same window.
- BL vs. GDN: 23.48% lower AR-hit PPL, but 6.72% higher overall PPL.

Table 2, p. 9; Appendix G.1, pp. 29–30. Mean $\pm$ SD; changes recomputed from table means.

The farthest AR-hit bucket has substantial seed variation

Distance to previous pair	BL PPL ↓	GDN PPL ↓
1–64 tokens	7.21 ± 0.52	10.09 ± 1.07
65–256 tokens	13.91 ± 1.00	19.57 ± 2.07
257–1,024 tokens	92.1 ± 8.9	99.9 ± 0.7

- BL has the lowest mean in every distance bucket.
- At 257–1,024 tokens, the gap is not resolved relative to seed variation.
- These are BPE-token perplexities on AR-hit positions, not word-level WikiText scores.

Table 2, p. 9; the authors' qualification in Appendix G.1, p. 30.

Four Bayesian layers are distilled into a 340M GDN

Retrofit

24-layer GDN; replace indices $\{5, 11, 17, 23\}$.
Four heads, head dimension 256, hidden width 1,024.

1. Phase A: parallel layer matching, $\approx$ 100M tokens per layer.
2. Phase B: joint teacher-logit KL and boundary MSE, $\approx$ 300M tokens.

Train at length 2,048; one seed; reported data exposure $\approx$ 400M tokens. GDN-FT continues the same four-layer mask for 400M tokens with next-token CE.

Sec. 4.3, pp. 9–10; Appendix G, pp. 27–28. The objectives differ between BL and GDN-FT.

NIAH gains persist against continued training

Context	GDN-zero	GDN-FT	BL	BL−FT (points)
2,048	49.52	48.46	**50.81**	+2.35
4,096	31.08	31.58	**34.02**	+2.44
8,192	16.81	16.06	**20.00**	+3.94
16,384	7.85	8.36	**8.81**	+0.45

- Mean over six NIAH tasks; CWE excluded. Not the full RULER suite.
- At 16k, absolute retrieval remains poor in all three models.
- At 2k, `single_2` regresses from 96.50 to 89.00 vs. GDN-zero.

Table 6 (Appendix Table A1), p. 29. Higher is better; differences recomputed.

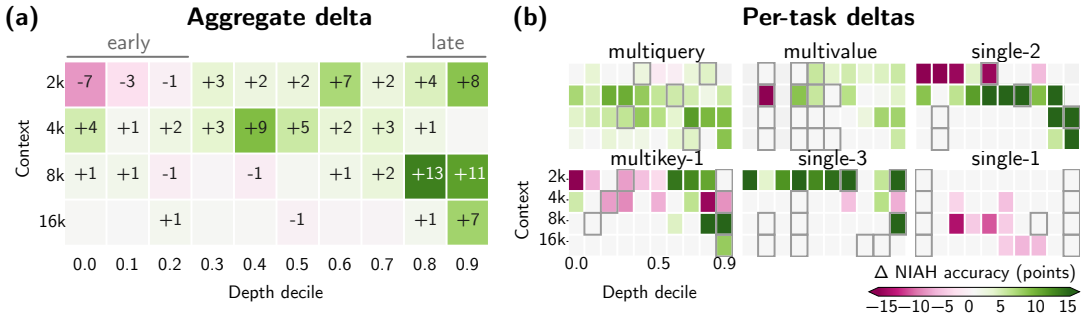
Retrieval gains carry a persistent perplexity penalty

Corpus at 8k	GDN-zero	BL	Relative PPL rise
WikiText-103	14.57	14.97	2.75%
SlimPajama-heldout	12.15	12.43	2.30%
PG-19 validation	19.29	19.81	2.70%

- The penalty persists from 2k to 16k; no overall PPL crossover.
- Paper summary: 2.5–2.7%; SlimPajama table values imply 2.30–2.38% across lengths.
- GDN-FT stays close to GDN-zero. Tokenizers differ from the from-scratch experiment.

Table 7 (Appendix Table A2), p. 29. Lower PPL is better; percentages use rounded table entries.

Gains vary with task and needle position



- Columns are needle-depth deciles, not network layers; green favors BL over GDN-zero.
- Middle positions at 4k and recent positions at 8k improve; early 2k can regress.
- 208 of 240 task/length/decile cells have fewer than 30 BL samples.

Original Fig. 5, p. 10; sample counts, Appendix G, p. 29. Matched-control heatmap: Fig. 8, p. 28.

The contribution is an uncertainty-adaptive write rule

1. **Mechanism:** covariance history shapes write direction and gain, with analytically testable collision dynamics.
2. **Evidence:** stronger recall in reported tasks; overall language-model PPL can worsen.
3. **Scope:** dense-state cost, small-scale evidence, single-seed retrofit, and v3 reporting inconsistencies remain relevant.

Discussion question

How much recall gain survives when covariance structure, actual throughput, and training objective are controlled together?

Paper, Sec. 5, pp. 10–11; Tables 2, 6–7. Discussion question: reading-side analysis.

Source and reading scope

Matthew Dowling, Hyungju Jeon, Cristina Savin, Il Memming Park.

Memory by Design: Probabilistic Sequence Layers.
arXiv:2605.31163v3, revised 18 August 2026.

https://arxiv.org/abs/2605.31163v3

- Read the versioned author TeX, all main sections, and Appendices A–G.1; cross-checked the 30-page PDF.
- Figures are original vector assets; Fig. 3(c,d) is a documented crop with its shared legend.
- Deterministic recurrence and table differences checked independently. No training or benchmark evaluation rerun.

Prepared 14 September 2026. Detailed caveats accompany the Chinese reading notes.

Backup: Kronecker closure keeps covariance compact

$$\widetilde{\mathbf{A}}_t = \mathbf{I}_m \otimes \mathbf{A}_t, \quad \mathbf{C}_t = \mathbf{I}_m \otimes \mathbf{k}_t^\top, \quad \mathbf{R}_t = r_t^2 \mathbf{I}_m$$

$$\widetilde{\mathbf{A}}_t (\mathbf{I}_m \otimes \mathbf{P}_{t-1}) \widetilde{\mathbf{A}}_t^\top + \ell_t^2 \mathbf{I}_{Dm} = \mathbf{I}_m \otimes \bar{\mathbf{P}}_t$$

$$\text{Posterior covariance} = \mathbf{I}_m \otimes \left(\bar{\mathbf{P}}_t - \frac{\bar{\mathbf{P}}_t \mathbf{k}_t \mathbf{k}_t^\top \bar{\mathbf{P}}_t}{r_t^2 + \mathbf{k}_t^\top \bar{\mathbf{P}}_t \mathbf{k}_t} \right).$$

- Shared column dynamics/noise preserve this structure exactly.
- Grouped observation noise requires one covariance per group.

Appendix C, pp. 15–16; grouped noise, Appendix D, p. 17.

Backup: a two-dimensional example of write steering

$$\mathbf{k}_B = \begin{pmatrix} 1 \\ 0 \end{pmatrix}, \qquad \mathbf{k}_A = \begin{pmatrix} 0.8 \\ 0.6 \end{pmatrix}, \qquad \bar{\mathbf{P}} = \begin{pmatrix} 0.05 & 0 \\ 0 & 2 \end{pmatrix}$$

$$\mathbf{u} = \bar{\mathbf{P}}\mathbf{k}_A = \begin{pmatrix} 0.04 \\ 1.2 \end{pmatrix}.$$

- B is well resolved; the orthogonal component is uncertain.
- A triggers the write, but its direction largely avoids B.
- This example uses covariance principal axes. In arbitrary fixed coordinates, off-diagonal entries represent that geometry.

Teaching example constructed for this reading, not a paper experiment; illustrates Eq. (15), p. 5.

Backup: distillation matches layers, then model outputs

$$\mathcal{L}_A^{(\ell)} = \frac{\|y_\ell^{BL} - y_\ell^{GDN}\|_2^2}{\mathrm{Var}(y_\ell^{GDN}) + 10^{-6}} + 0.05(1 - \cos(y_\ell^{BL}, y_\ell^{GDN}))$$

$$\mathcal{L}_B = \lambda_{KL}\,\mathrm{KL}(p_{GDN}\|p_{BL}) + \frac{\lambda_b}{|\mathcal{I}|}\sum_{\ell\in\mathcal{I}} \frac{\|y_\ell^{BL} - y_\ell^{GDN}\|_2^2}{\mathrm{Var}(y_\ell^{GDN}) + 10^{-6}}$$

- y_ℓ: matched layer-boundary outputs; frozen teacher GDN.
- Phase B: $(\lambda_{KL}, \lambda_b) = (0.8, 0.2)$ then $(0.9, 0.1)$.
- GDN-FT uses next-token CE. The equal-token control does not isolate architecture from objective choice.

Appendix G, pp. 27–28. Loss normalization reproduced as specified in the paper.

Backup: reporting differences affect reproduction

- **Fig. 3(a):** visible BL endpoint differs from the appendix's independently reproducible 0.72309.
- **Fig. 3(c):** axis is accuracy; main text describes margin. The appendix also uses a different softmax definition.
- **Learned reset:** main text describes a learned scalar; Appendix E.5.5 fixes reset process variance to 0.05.
- **MQAR:** distinguish GDN from GDN-conv. Imported Base curves under the modified generator need clarification.

These differences limit quantitative claims from the affected experiments; the recurrence derivation can be checked separately.

Fig. 3, p. 7; Secs. 4.1–4.2, pp. 8–9; Appendices E.3–E.5 and F, pp. 20–27.

Backup: two algebraic qualifications

Latent-input posterior uncertainty need not be zero.

$$c_t = \frac{\lambda_t r_t^2}{\lambda_t + r_t^2} > 0, \quad \mathbf{C}_t = \mathbf{A}_t \mathbf{C}_{t-1} \mathbf{A}_t^\top + c_t \mathbf{k}_t \mathbf{k}_t^\top .$$

The mean update does not need $\mathbf{C}_t$, so the layer may omit it.

Use the exact gain formula for asymptotic expansion.

$$\ell^2 \gg r^2 : \quad g_\infty = 1 - \frac{r^2}{\ell^2} + O\left(\frac{r^4}{\ell^4}\right).$$

Appendix E.2 prints a factor 1/2 in the leading correction. Its exact formula implies the expansion above.

Reading-side derivations from Sec. 3.2 / Appendix A and Eq. (42), p. 19.