Randomized Kaczmarz EM for Large-Scale State-Space Models

Jaafar Almutawa
Department of Mathematics, Bahrain Polytechnic, Bahrain
Email: jaafar111@yahoo.com

***Abstract*—The Expectation–Maximization (EM) algorithm is a standard tool for estimating linear state-space models, but its reliance on dense covariance matrices makes it difficult to use in large dimensions. We show that, for systems with sparse or banded dynamics and simple noise structures, EM can be run without ever forming these matrices. The key step is to replace the usual smoother with randomized Kaczmarz iterations in information form, reducing the covariance memory cost to $\mathcal{O}(n)$ and yielding a total footprint of $\mathcal{O}(Nn)$ for the state trajectory, versus $\mathcal{O}(Nn^2)$ for the classical smoother. The overall behavior of EM is preserved: the likelihood still improves up to a tolerance set by the iterative solver, and shrinking this tolerance recovers the classical updates. Experiments on large advection–diffusion models illustrate the payoff: RK-EM matches standard EM at moderate sizes and continues to operate at $n = 65{,}536$, where traditional methods run out of memory. When the system is sparse and reasonably well observed, this matrix-free version offers a practical alternative to ensemble smoothers, with no localization or inflation parameters to tune.**

***Index Terms*—Expectation–Maximization, Randomized Kaczmarz, Large-scale systems.**

I. Introduction

High-dimensional state-space models appear in many modern applications, including geophysical forecasting, sensor networks, and electrochemical systems, where the latent state may contain tens of thousands of variables. The Expectation–Maximization (EM) algorithm is a natural choice for maximum-likelihood (ML) estimation in linear dynamical systems [2], [3], but its usability declines sharply once the state dimension grows: the Rauch–Tung–Striebel (RTS) smoother propagates dense $n \times n$ covariance matrices, giving an $\mathcal{O}(n^2)$ memory cost that quickly becomes the limiting factor.

Several strategies have been proposed to reduce this burden. Classical EM-based formulations [3], [4] remain statistically efficient but still rely on full covariance updates. Recent scalable variants—gradient-based [5] or precision-based sampling [1]—either depend on dense matrix operations or target different regimes. Subspace methods [6] and prediction-error techniques [7] scale more comfortably but do not offer EM's likelihood guarantees, while Bayesian and kernel approaches [8], [9] often introduce dense operators of their own. Ensemble Kalman methods [10] reduce memory by replacing covariance propagation with ensembles of size $N_e \ll n$, but they rely on tuning choices such as localization and inflation [11] and do not directly optimize the likelihood. Randomized Kaczmarz (RK) methods [12]–[14] are effective for large linear systems but have not been integrated into EM in a way that preserves its monotonicity or ML interpretation.

EM remains attractive because it provides stable ML updates for latent-variable models, naturally handles missing data, and avoids direct optimization of nonconvex likelihoods; its main limitation in high dimensions is the covariance-based smoothing step. We address that limitation with a matrix-free EM algorithm that targets sparse or banded dynamics and diagonal or block-diagonal noise. The smoother is reformulated in the information domain and solved via RK iterations, giving $\mathcal{O}(n)$ peak memory while retaining the ML objective, linear-memory scaling, and clear convergence guarantees.

Our contributions are:

1) **Matrix-free smoothing via randomized Kaczmarz.** We replace the RTS smoother with RK iterations on information-form equations, eliminating dense covariance storage.
2) **Quasi-monotone likelihood ascent.** We prove that inexact RK solves cause only a tolerance-dependent deviation from classical EM ascent, with constants that are dimension-independent under localized interactions.
3) **Scalability.** We contrast RK-EM with localized EnKS-EM (Table II) and demonstrate scalability to $n = 65{,}536$ on linearized advection–diffusion and reaction–diffusion models where classical EM is infeasible.

The paper proceeds as follows. Section II reviews EM and RK. Section III introduces the matrix-free formulation. Section IV presents convergence and complexity results. Section V gives practical guidance, Section VI provides experiments, and Section VII concludes.

II. Background

We consider the discrete-time linear state-space model

$$x_{t+1} = Ax_t + Bu_t + w_t, \qquad w_t \sim \mathcal{N}(0, Q), \tag{1}$$

$$y_t = Cx_t + Du_t + e_t, \qquad e_t \sim \mathcal{N}(0, R), \tag{2}$$

with $x_t \in \mathbb{R}^n$, $u_t \in \mathbb{R}^m$, $y_t \in \mathbb{R}^p$, and $x_0 \sim \mathcal{N}(\mu_0, P_0)$. Given $\{(u_t, y_t)\}_{t=1}^N$, the goal is to estimate $\theta = \{A, B, C, D, Q, R, \mu_0, P_0\}$ by maximizing the log-likelihood $\log p(Y_N \mid U_N, \theta)$.

The EM algorithm [15] alternates between computing the expected complete-data log-likelihood

$$\mathbf{Q}(\theta, \theta^{(k)}) = \mathbb{E}\Big[\log p(Y_N, X_N \mid \theta) \mid Y_N, \theta^{(k)}\Big], \tag{3}$$

and updating $\theta^{(k+1)} = \arg\max_\theta \mathbf{Q}(\theta, \theta^{(k)})$. For linear–Gaussian models, $\mathbf{Q}$ depends only on three sufficient statistics [2]:

$$S_1 = \sum_{t=1}^{N} \mathbb{E}[x_t \mid Y_N], \tag{4}$$

$$S_2 = \sum_{t=1}^{N} \mathbb{E}[x_t x_t^\top \mid Y_N], \tag{5}$$

$$S_3 = \sum_{t=1}^{N} \mathbb{E}[x_t x_{t-1}^\top \mid Y_N], \tag{6}$$

obtained via a forward Kalman filter and a Rauch–Tung–Striebel (RTS) smoother [16]. The limitation is that smoothing requires the dense posterior covariances $P_{t|N} \in \mathbb{R}^{n\times n}$: even if the full sequence $\{P_{t|N}\}$ is not stored, the backward pass must maintain at least one dense covariance matrix and its Cholesky factor, yielding an $\mathcal{O}(n^2)$ memory footprint.

Randomized Kaczmarz (RK) is attractive in this setting because it requires only row-wise access to the system matrix and never forms or stores it explicitly. Starting from $x^{(\ell)}$, one randomly selects a row $a_j^\top$ of $\mathcal{A}x = b$ with probability proportional to $\|a_j\|^2$ and performs the update

$$x^{(\ell+1)} = x^{(\ell)} + \frac{b_j - a_j^\top x^{(\ell)}}{\|a_j\|^2} a_j.$$

This stores only a handful of vectors (an $\mathcal{O}(n)$ footprint) and converges linearly in expectation,

$$\mathbb{E}\Big[\|x^{(\ell)} - x^*\|^2\Big] \le (1 - \kappa^{-2})^\ell \|x^{(0)} - x^*\|^2,$$

with $\kappa = \|\mathcal{A}\|_F / \sigma_{\min}(\mathcal{A})$ [12]. RK supports warm starts and admits block, greedy [13], [14], and preconditioned variants. In this paper we use RK to replace the covariance-based RTS smoother within the E-step and analyze how its approximation error propagates through the EM iterations.

III. Matrix-Free EM Algorithm

Rather than storing full posterior covariances, we perform smoothing in the *information form* [17]. The information matrix $Y_{t|N}$ and information vector $\eta_{t|N}$ are computed by the backward recursion ($t = N-1, \dots, 1$):

$$Y_{t|N} = Y_{t|t} + A^\top \Lambda_{t+1} A, \tag{7}$$

$$\eta_{t|N} = \eta_{t|t} + A^\top \Lambda_{t+1}\, \eta_{t+1|N}, \tag{8}$$

$$\Lambda_{t+1} := (Q + Y_{t+1|N}^{-1})^{-1}, \tag{9}$$

where $\Lambda_{t+1} v$ is evaluated matrix-free by an RK solve against $(I + Q\, Y_{t+1|N})$, which is sparse under our assumptions. The recursion is initialized at $t = N$ with $Y_{N|N} = P_{N|N}^{-1}$ and $\eta_{N|N} = P_{N|N}^{-1} \hat{x}_{N|N}$, and the filtered information quantities $Y_{t|t} = P_{t|t}^{-1}$, $\eta_{t|t} = P_{t|t}^{-1} \hat{x}_{t|t}$ are produced by the forward Kalman pass. The smoothed mean is then the unique solution of

$$Y_{t|N}\, \hat{x}_{t|N} = \eta_{t|N}. \tag{10}$$

When A is sparse or banded and Q^{-1} is diagonal or block-diagonal, $Y_{t|N}$ inherits this sparsity, so matrix–vector products $Y_{t|N}x$ reduce to local stencil operations and RK accesses only individual rows. These two structural conditions are precisely what enables per-step memory $\mathcal{O}(n)$ rather than $\mathcal{O}(n^2)$, avoiding any dense $n \times n$ factorization.

The linear system (10), which replaces the classical RTS smoothing step within the E-step of EM, is solved approximately using the randomized Kaczmarz (RK) method:

$$\hat{x}_{t|N}^{\mathrm{RK}} \approx \arg\min_x \|Y_{t|N} x - \eta_{t|N}\|,$$

and iterations stop once the relative residual $\|\mathbf{r}^{(\ell)}\| / \|\eta_{t|N}\|$ falls below a tolerance ϵ_{RK}, where $\mathbf{r}^{(\ell)} = Y_{t|N} x^{(\ell)} - \eta_{t|N}$.

During the backward sweep $t = N, \dots, 1$, the algorithm proceeds in analogy with the RTS smoother:

1) solves $Y_{t|N} x = \eta_{t|N}$ via RK (warm-starting from $\hat{x}_{t|t}$),
2) updates the required sufficient statistics,
3) stores only $\hat{x}_{t|N}^{\mathrm{RK}}$ for use in the M-step.

This requires keeping N state vectors of length n but avoids storing any dense $n \times n$ covariance matrices.

The smoothed state sequence enables the M-step statistics to be accumulated on the fly. This keeps memory proportional to n and avoids storing second-order moments in full matrix form.

The statistic S_1 is updated incrementally as $S_1 \leftarrow S_1 + \hat{x}_{t|N}^{\mathrm{RK}}$, requiring only one n-vector of storage.

When each $x_t[i]$ depends only on a neighborhood $\mathcal{N}(i)$ of size $w = \mathcal{O}(1)$, the cross-moment needed for row i of A is

$$\sum_{t=1}^{N} \hat{x}_{t|N}^{\mathrm{RK}}[i]\, \hat{x}_{t-1|N}^{\mathrm{RK}}[\mathcal{N}(i)]^\top.$$

These per-stencil quantities require only $\mathcal{O}(nw^2)$ memory and avoid forming the full outer product. If A is dense, S_3 is inherently $n \times n$, so the $\mathcal{O}(n)$ savings do not apply.

For diagonal Q, only the diagonal entries of S_2 are needed:

$$[S_2]_{ii} \approx \sum_{t=1}^{N} \left(\hat{x}_{t|N}^{\mathrm{RK}}[i]^2 + [P_{t|N}]_{ii} \right).$$

The diagonal variance $[P_{t|N}]_{ii}$ is estimated by probe vectors:

$$[P_{t|N}]_{ii} \;\approx\; \frac{1}{N_s} \sum_{s=1}^{N_s} [\tilde{z}^{(s)}]_i^2, \quad \tilde{z}^{(s)} \approx Y_{t|N}^{-1} \epsilon^{(s)},\ \epsilon^{(s)} \sim \mathcal{N}(0, I),$$

in the spirit of Hutchinson's stochastic diagonal/trace estimator [18], where $\tilde{z}^{(s)}$ is the approximate RK solution satisfying $\|\tilde{z}^{(s)} - Y_{t|N}^{-1}\epsilon^{(s)}\| \le \epsilon_{\mathrm{RK}}\, \kappa_{\mathrm{spec}}(Y_{t|N})\, \|Y_{t|N}^{-1}\epsilon^{(s)}\|$, with $\kappa_{\mathrm{spec}}(Y_{t|N}) := \sigma_{\max}(Y_{t|N}) / \sigma_{\min}(Y_{t|N})$. This introduces a per-entry bias of order $\mathcal{O}(\epsilon_{\mathrm{RK}}\, \kappa_{\mathrm{spec}}(Y_{t|N})\, [P_{t|N}]_{ii})$, which is absorbed into the overall tolerance budget of Theorem 1. Only $\mathcal{O}(n)$ memory is required for these estimates.

The resulting per-component memory breakdown for a representative configuration is given in Table I.

The M-step, which updates the parameters by maximizing the expected complete-data log-likelihood, solves the normal equations

$$\Gamma \begin{bmatrix} A^\top \\ B^\top \end{bmatrix} = S_3^\top,\ \Gamma = \sum_{t=1}^{N-1} \mathbb{E}[\tilde{z}_t \tilde{z}_t^\top \mid Y_N],\ \tilde{z}_t = \begin{bmatrix} x_{t-1} \\ u_{t-1} \end{bmatrix}, \quad (11)$$

but forming Γ directly requires $\mathcal{O}(n^2)$ memory. To preserve linear scaling, we solve row-wise systems using RK. For each state index $i = 1, \ldots, n$, we consider the reduced system

$$\Gamma_i \begin{bmatrix} a_i[\mathcal{N}(i)] \\ b_i \end{bmatrix} = s_i^{(\mathrm{loc})}, \quad (12)$$

where Γ_i is the local $(w+m) \times (w+m)$ normal matrix constructed using only neighborhood quantities.

A matrix-free evaluation of $\Gamma_i v$ for $v \in \mathbb{R}^{w+m}$ is given by

$$\Gamma_i v = \sum_{t=1}^{N-1} \begin{bmatrix} \hat{x}_{t-1|N}[\mathcal{N}(i)] \\ u_{t-1} \end{bmatrix} \left(\hat{x}_{t-1|N}[\mathcal{N}(i)]^\top v_x + u_{t-1}^\top v_u \right) + \sum_{t=1}^{N-1} \begin{bmatrix} \mathrm{diag}(P_{t-1|N}[\mathcal{N}(i), \mathcal{N}(i)]) v_x \\ 0 \end{bmatrix}, \quad (13)$$

which uses only the stored smoothed states and diagonal variance estimates. The RK method solves (12) with tolerance $\epsilon_M \approx 10\epsilon_{\mathrm{RK}}$, and warm starts are taken from the previous EM iteration. Since $w = \mathcal{O}(1)$, per-row storage remains $\mathcal{O}(w+m)$, yielding total memory $\mathcal{O}(n)$.

Assumption 1 (Persistent Excitation)**.** *There exists $\alpha_\Gamma > 0$ such that $\sigma_{\min}(\Gamma_i) \geq \alpha_\Gamma$ for all $i = 1, \ldots, n$ and all EM iterates.*

Under Assumption 1, Γ_i is positive definite and the system (12) is consistent, so Proposition 1 applies with $\kappa_\Gamma = \sigma_{\max}(\Gamma_i)/\alpha_\Gamma$. The off-diagonal entries of $P_{t-1|N}[\mathcal{N}(i), \mathcal{N}(i)]$ appearing in (13) are set to zero when only diagonal variance estimates are available; by Lemma 1 the resulting bias per row of A is bounded by $\mathcal{O}(C_{\mathrm{loc}}\, w\, \sigma_{\max}(P_\infty)\, e^{-1/\rho})$ (from Lemma 1 applied to off-diagonal entries at graph distance ≥ 1 within the neighbourhood $\mathcal{N}(i)$), which is negligible when ρ is small.

Similar row-wise updates, obtained as the analog of (11) for the measurement equation, are carried out for (C, D) and for the diagonal components of Q and R.

Table I summarizes the memory usage for a typical configuration with $n = 4096$, $N = 500$, and stencil width $w = 5$. All components scale linearly in n when w is fixed, and the dominant cost becomes storing the state trajectory $\{\hat{x}_{t|N}\}$.

For comparison, the classical RTS smoother requires at least one dense $n \times n$ covariance matrix, i.e., $8n^2$ bytes: about 134 MB at $n = 4096$ and about 34 GB at $n = 65{,}536$ for a single covariance matrix. Storing the full set of $N = 500$ smoothing covariances would already require roughly 67 GB and 17 TB, respectively, making traditional EM infeasible at such scales.

TABLE I
PEAK MEMORY USAGE FOR $n = 4096$, $N = 500$.

Component	Memory (MB)
Smoothed states ($N \times n$ doubles)	15.6
RK auxiliary vectors	0.16
Local stencil accumulators	0.82
Diagonal variance summaries	0.03
Sparse LU / FFT plans	0.05
Miscellaneous overhead	0.50
Total peak RSS	**17 MB**

IV. Convergence Theory and Complexity

This section characterizes how the inexact RK solves used in the E-step influence the ascent behavior of EM. Throughout, the RK iterator for $Y_{t|N} x = \eta_{t|N}$ is terminated once the relative residual reaches a prescribed tolerance.

Assumption 2 (RK Approximation Error)**.** *For each t, let $\hat{x}^*_{t|N}$ denote the exact information-form smoother solution. Assume the RK solve returns an estimate $\hat{x}^{\mathrm{RK}}_{t|N}$ satisfying*

$$\|\hat{x}^{\mathrm{RK}}_{t|N} - \hat{x}^*_{t|N}\| \ \leq\ \delta_x \, \|\hat{x}^*_{t|N}\|,\ \delta_x \ \leq\ \kappa_{\mathrm{spec}}(Y_{t|N})\, \epsilon_{\mathrm{RK}}, \quad (14)$$

whenever the RK residual obeys $\|Y_{t|N} x^{(\ell)} - \eta_{t|N}\| / \|\eta_{t|N}\| \leq \epsilon_{\mathrm{RK}}$.

Assumption 3 (Independent RK randomness across time)**.** *The random row selections used by the RK solver at different smoothing times t are generated from independent streams, so that the errors $\{\xi_t = \hat{x}^{\mathrm{RK}}_{t|N} - \hat{x}^*_{t|N}\}$ form a martingale-difference sequence with respect to the natural filtration.*

Remark 1. *Assumption 3 is algorithmic: it is enforced trivially by the implementation (independent pseudo-random streams per time step). It does not preclude statistical dependence through the filtered quantities $Y_{t|N}, \eta_{t|N}$, which enter the conditional mean.*

The behavior of the approximation error in large systems depends strongly on the structure of the covariance operator. In many high-dimensional yet *locally coupled* models—where each component of x_t interacts with only a modest number of neighbors—posterior correlations decay rapidly along the sparsity graph, keeping the effective dimensionality of the covariance small even when n is large.

Lemma 1 (Localization on Lattice Dependency Graphs)**.** *Let the smoothing covariance P_∞ correspond to a system whose state-transition matrix induces a sparsity graph isomorphic to a sublattice of $\mathbb{Z}^d$ with bounded degree Δ. Suppose P_∞ exhibits exponentially decaying off-diagonal correlations along this graph, i.e.,*

$$|[P_\infty]_{ij}| \leq C_{\mathrm{loc}}\, \sigma_{\max}(P_\infty) \exp\big(-\mathrm{dist}(i,j)/\rho\big), \quad (15)$$

where $\mathrm{dist}(\cdot,\cdot)$ is the graph distance and $\rho > 0$ is a local dependency radius. Then the local effective rank per unit volume satisfies

$$\bar{r}_{\mathrm{loc}} := \frac{1}{n} \frac{\mathrm{Tr}(P_\infty)}{\sigma_{\max}(P_\infty)} \leq 1,\ \max_i \sum_j \frac{|[P_\infty]_{ij}|}{\sigma_{\max}(P_\infty)} \leq C_d\, \rho^d, \quad (16)$$

so that, although $\mathrm{Tr}(P_\infty) = \Theta(n)$*, each row of* P_∞ *is effectively supported on* $\mathcal{O}(\rho^d)$ *sites independent of* n*.*

Proof. Under the exponential-decay assumption, each row satisfies $\sum_j |[P_\infty]_{ij}| \le C_{\mathrm{loc}}\,\sigma_{\max}(P_\infty)\sum_{k=0}^{\infty} V_d(k)\,e^{-k/\rho}$, where $V_d(k) \le c_d(k+1)^{d-1}$ counts the number of lattice sites at graph distance k from i. The finite geometric-polynomial series evaluates to a constant $C_d\,\rho^d$ depending only on d, C_{loc}, and the lattice geometry. Hence every row of P_∞ has bounded ℓ_1 mass (normalized by $\sigma_{\max}(P_\infty)$) independent of n; this is the localization property used throughout the paper. □

The following structural conditions are sufficient to guarantee such decay.

Assumption 4 (Localized smoothing regime)**.** *Consider a system where:*

- *the transition matrix* A *has local sparsity width* $w = \mathcal{O}(1)$*,*
- *the process noise is uncorrelated,* $Q = q^2 I$*,*
- *the pair* (C, A) *is* ν*-step detectable, i.e.,*$\sigma_{\min}\Big(\sum_{k=0}^{\nu-1}(A^k)^\top C^\top C A^k\Big) \ge \alpha > 0$ *for some* $\nu = \mathcal{O}(w/\gamma)$ *and* $\alpha > 0$*, and*
- *the dynamics are contractive,* $\mathrm{sr}(A) \le 1-\gamma$ *for some* $\gamma > 0$*, where* $\mathrm{sr}(\cdot)$ *denotes spectral radius.*

with dependency radius $\rho = \mathcal{O}(w/\gamma)$*. We assume that the smoothing covariance* P_∞ *then satisfies the exponential-decay bound of Lemma 1 with that* ρ*; a rigorous derivation from mixing and detectability is left to future work.*

The combination of Assumption 2 and the localization properties above yields a controlled relaxation of EM monotonicity: the likelihood is allowed a small decrease set by the RK tolerance. Intuitively, Theorem 1 says that each EM step behaves like the classical one up to an error proportional to ϵ_{RK}; in the general case this error carries a $\sqrt{n}$ prefactor, but in the localized regime (Assumption 4) the dimension factor is replaced by the constant ρ^d, so the guarantee no longer degrades with n.

Throughout, $\sigma_x^2 := \max_t \mathrm{Tr}(P_{t|N})/n$ denotes the average per-component smoothing variance; under stationarity $\sigma_x^2 \le \sigma_{\max}(P_\infty)$, and localization does *not* reduce σ_x^2. Dimension-independence in the localized regime instead enters through the stable-rank bound of Lemma 1 (Step 3 of the proof).

Theorem 1 (Quasi-Monotone Likelihood Ascent)**.** *Under Assumptions 2 and 1, and assuming that the RK errors* $\{\xi_t\}_{t=1}^N$ *form a martingale-difference sequence with respect to the filtration* $\mathcal{F}_t$ *generated by the random row selections up to time* t*, with probability at least* $1-\delta$*,*

$$\log p(Y_N \mid \theta^{(k+1)}) \ge \log p(Y_N \mid \theta^{(k)}) - C(\theta^{(k)})\,\epsilon_{\mathrm{RK}}, \quad (17)$$

where a general bound (up to problem-dependent constants) is

$$C(\theta^{(k)}) \le \frac{C_0 N\sqrt{n}\,\|A\|^2\sigma_{\max}(P_\infty)}{\sigma_{\min}(Q)\,\lambda_*\,\alpha_\Gamma}\sqrt{\log(2N/\delta)},$$
$$\lambda_* := \min_{1\le t\le N}\lambda_{\min}(Y_{t|N}), \quad (18)$$

If Assumption 4 holds (localized covariance regime), then

$$C(\theta^{(k)}) \le \frac{C_0 N\rho^{\,d}\,\|A\|^2\sigma_{\max}(P_\infty)}{\sigma_{\min}(Q)\,\lambda_*\,\alpha_\Gamma}\sqrt{\log(2N/\delta)}, \quad (19)$$

where the dependency radius ρ *is independent of* n*. Hence the bound becomes insensitive to the ambient dimension in locally coupled regimes.*

Proof of Theorem 1. Throughout, $L_i := \sup_{\theta\in\Theta_0}\|\nabla_{S_i}\mathbf{Q}(\theta,\theta^{(k)})\|_{\mathrm{op}}$ denotes the operator Lipschitz constant of $\mathbf{Q}$ with respect to the i-th sufficient statistic on a compact set Θ_0 containing the EM iterates; these are finite under Assumption 1. We write C_0 for a generic numerical constant absorbing Lipschitz factors, $\|A\|^2/\sigma_{\min}(Q)$, the Azuma–Hoeffding factor, and the bound on $\lambda_{\max}(Y_{t|N})$ from Step 3.

Step 1: Q-function decomposition. The EM identity gives

$$L(\theta^{(k+1)}) - L(\theta^{(k)}) = \big[\mathbf{Q}(\theta^{(k+1)},\theta^{(k)}) - \mathbf{Q}(\theta^{(k)},\theta^{(k)})\big] - D_{\mathrm{KL}}\big(p(X|Y,\theta^{(k)})\,\|\,p(X|Y,\theta^{(k+1)})\big).$$

Since $D_{\mathrm{KL}} \ge 0$, it suffices to lower-bound the bracketed term. Writing $\mathbf{Q} = \mathbf{Q}_{\mathrm{RK}} + (\mathbf{Q}-\mathbf{Q}_{\mathrm{RK}})$ and using Lipschitz continuity, $|\mathbf{Q}-\mathbf{Q}_{\mathrm{RK}}| \le \sum_i L_i\|\tilde S_i - S_i\|_F =: \Delta Q$, so

$$L(\theta^{(k+1)}) - L(\theta^{(k)}) \ge [\mathbf{Q}_{\mathrm{RK}}(\theta^{(k+1)},\theta^{(k)}) - \mathbf{Q}(\theta^{(k)},\theta^{(k)})] - 2\Delta Q,$$

where the factor 2 accounts for errors at both $\theta^{(k+1)}$ and $\theta^{(k)}$.

Step 2: Sufficient-statistic concentration. Let $\xi_t := \hat x_{t|N}^{\mathrm{RK}} - \hat x_{t|N}^{*}$, so $\|\xi_t\| \le \delta_x\|\hat x_{t|N}^{*}\|$ by Assumption 2. Set $\Phi := S_2 = \sum_t(P_{t|N} + \hat x_{t|N}\hat x_{t|N}^\top)$ and $\Psi := S_3 = \sum_t(\hat x_{t|N}\hat x_{t-1|N}^\top + P_{t,t-1|N})$; the cross-covariance $P_{t,t-1|N}$ is estimated by probe vectors at the same tolerance as the diagonal variance, and its contribution is absorbed into C_0.

Let $\mathcal{F}_t$ denote the σ-algebra generated by all RK row selections at times $1,\dots,t$, and define the centered increments

$$Z_t := (\xi_t\hat x_{t|N}^{*\top} + \hat x_{t|N}^{*}\xi_t^\top + \xi_t\xi_t^\top) - \mathbb{E}[\,\cdot \mid \mathcal{F}_{t-1}].$$

Under Assumption 3, $\{Z_t\}$ is a matrix martingale-difference sequence with $\|Z_t\|_{\mathrm{op}} \le 2\delta_x\sigma_x^2$ a.s. and $\sum_t \mathbb{E}[Z_t^2 \mid \mathcal{F}_{t-1}] \preceq N\delta_x^2\sigma_x^4\,I$, where $\sigma_x^2 := \max_t \mathrm{Tr}(P_{t|N})/n$. The matrix Azuma–Hoeffding inequality [19, Thm. 7.1] then gives, with probability $\ge 1-\delta$,

$$\Big\|\sum_t Z_t\Big\|_{\mathrm{op}} \le 2\delta_x\sigma_x^2\sqrt{2N\log(2n/\delta)}.$$

Converting to Frobenius norm via $\|\cdot\|_F \le \sqrt{r}\,\|\cdot\|_{\mathrm{op}}$ with r the stable rank of the error matrix yields

$$\Big\|\sum_t Z_t\Big\|_F \le 2\delta_x\sigma_x^2\sqrt{2rN\log(2n/\delta)},$$

where $r \le n$ in general and $r \le C_d\rho^d$ under Lemma 1. Analogous bounds hold for Ψ and Γ.

Step 3: Combining the bounds. From Step 1 and the Frobenius bound above, $\Delta Q \le C_0 N\delta_x\sqrt{r}\,\|A\|^2\sigma_x^2/\sigma_{\min}(Q)$. Assumption 2 gives $\delta_x \le \kappa_{\mathrm{spec}}(Y_{t|N})\,\epsilon_{\mathrm{RK}}$, and under the standard non-degeneracy condition $Q \succeq q_{\min}^2 I$ (which inherits a process-noise floor to $P_{t|N}$, so $\lambda_{\max}(Y_{t|N}) \le 1/p_{\min}$

with $p_{\min} := \min_t \lambda_{\min}(P_{t|N}) > 0$), together with $\sigma_x^2 \le \sigma_{\max}(P_\infty)$, absorbing numerical constants into C_0 yields (18) with $\lambda_* := p_{\min}$ (taking $r = n$). Under Assumption 4, $r = C_d\rho^d$ replaces n and gives (19). □

Remark 2. *The constant $C(\theta^{(k)})$ grows with N and with the conditioning of the information operator, and improves markedly when local coupling limits long-range correlations. As $\epsilon_{\mathrm{RK}} \to 0$, the inequality collapses to classical EM monotonicity.*

RK Iteration Complexity

The number of RK iterations needed to reach a relative residual ϵ depends on the conditioning of the system matrix.

Proposition 1 (RK Iteration Bound)**.** *For $\mathcal{A}x = b$, define the Frobenius condition number $\kappa_F^2 = \|\mathcal{A}\|_F^2/\sigma_{\min}^2(\mathcal{A})$, which satisfies $\kappa_F \le \sqrt{n}\,\kappa_{\mathrm{spec}}$ where $\kappa_{\mathrm{spec}} = \sigma_{\max}(\mathcal{A})/\sigma_{\min}(\mathcal{A})$. The RK method requires $\mathcal{O}\big(\kappa_F^2 \log(1/\epsilon)\big)$ expected iterations to achieve relative accuracy ϵ [12]; for a square nonsingular $\mathcal{A}$ this is at most $\mathcal{O}(n\,\kappa_{\mathrm{spec}}^2 \log(1/\epsilon))$.*

Application to Smoothing.: For the information-form system $Y_{t|N}x = \eta_{t|N}$ the RK convergence rate is governed by the Frobenius condition number $\kappa_F(Y_{t|N})$. When $Y_{t|N}$ is s-sparse per row (as under Assumption 4), each row has ℓ_2 norm bounded by $\sqrt{s}\,\|Y_{t|N}\|_{\mathrm{op}}$ (up to a constant), so $\|Y_{t|N}\|_F^2 \le s\,\|Y_{t|N}\|_{\mathrm{op}}^2$ and $\kappa_F^2(Y_{t|N}) = \mathcal{O}(s\,\kappa_{\mathrm{spec}}^2(Y_{t|N})) = \mathcal{O}(s/\gamma^2)$ independently of n. When observations constrain only a small subset of directions, $\sigma_{\min}(Y_{t|N})$ shrinks and RK convergence slows.

Total Computational Cost

For a system with local sparsity width $w = \mathcal{O}(1)$, an EM iteration requires:

- Forward filtering: $\mathcal{O}(Nnw^2)$,
- RK-based smoothing: $\mathcal{O}\big(N\,\kappa^2 \log(1/\epsilon_{\mathrm{RK}})\,nw\big)$,
- Row-wise M-step updates: $\mathcal{O}\big(n\,\kappa_\Gamma^2 \log(1/\epsilon_M)\,Nw^2\big)$.

In locally coupled regimes where $\kappa_F(Y_{t|N})$ and $\kappa_F(\Gamma_i)$ remain bounded in n, the overall complexity is $\mathcal{O}(Nn)$ with $\mathcal{O}(Nn)$ state-trajectory memory and $\mathcal{O}(n)$ covariance memory, compared to $\mathcal{O}(Nn^2)$ for classical EM. In poorly observed or strongly correlated regimes, κ may grow with n, and the cost approaches that of dense EM—while still preserving linear memory usage.

V. Practical Guidance

The bound in (18) suggests how to select the RK residual tolerance ϵ_{RK}. To limit the per-iteration likelihood drop to a user-specified level τ (e.g., 0.01% of the current log-likelihood magnitude), it suffices to enforce

$$\epsilon_{\mathrm{RK}} \;\le\; \frac{\tau\,\sigma_{\min}(Q)\,\lambda_{\min}(Y_{t|N})}{C_0 N\sqrt{n}\|A\|^2\sigma_{\max}(P_\infty)\sqrt{\log(2N/\delta)}}. \tag{20}$$

In practice we use an adaptive coarse-to-fine schedule: early EM iterations run with a loose tolerance to move quickly in parameter space, and ϵ_{RK} is tightened by one order of magnitude whenever a significant likelihood drop is observed. The M-step tolerance $\epsilon_M = 10\,\epsilon_{\mathrm{RK}}$ is typically adequate, since its contribution to the likelihood error is of higher order.

Two diagnostics are useful:

- **Slow RK convergence.** If RK exceeds $\sim 10^4$ iterations per solve, $Y_{t|N}$ is likely ill-conditioned. Estimating $\kappa(Y_{t|N})$ by power iteration and applying mild Tikhonov regularization, $\tilde{Y}_{t|N} = Y_{t|N} + \lambda I$ with $\lambda = 10^{-5}\|Y_{t|N}\|_{\mathrm{op}}$, usually restores stability.
- **Unphysical parameters.** Unstable estimates ($\mathrm{sr}(A) > 1$ or near-singular Q, R) indicate poor observability or model mismatch. Soft penalties on $\log\det Q$, $\log\det R$, or a projection of A onto the stable set during the M-step remove such pathologies.

Table II summarizes when RK-EM or localized EnKS-EM is preferable.

TABLE II
Qualitative comparison of RK-EM and localized EnKS–EM.

Scenario	RK-EM	EnKS–EM
Banded A, diagonal Q, R	✓	✓
Adequate observability ($p/\rho^d \gtrsim 1$)	✓ ✓	✓
Very sparse sensing ($p \ll n$)	✓	✓ ✓
Maximum-likelihood guarantees desired	✓	–
No tuning of localization/inflation	✓	–
Embarrassingly parallel updates	–	✓
High robustness under ill-conditioning	–	✓

VI. Numerical Experiments

We assess RK-EM on large linear state-space models and compare against classical EM. Rather than matrix estimation error, we report the normalized $\mathcal{H}_2$ error (FRE) of the identified transfer function $G(z) = C(zI - A)^{-1}$ on a dense frequency grid, which measures input–output fidelity more interpretably in high dimensions. Identifiability is ensured by structural constraints (banded or sparse A). All experiments use double precision, $\epsilon_{\mathrm{RK}} = 10^{-4}$, $\epsilon_M = 10^{-3}$, on a dual-socket Intel Xeon E5-2680 v4 with 128 GB RAM.

The primary benchmark is a one-dimensional stochastic advection–diffusion equation,

$$\frac{\partial c}{\partial t} = \alpha\frac{\partial^2 c}{\partial x^2} - v\frac{\partial c}{\partial x} + \sigma_w\xi(x,t),$$

with periodic boundaries, discretized using central differences for diffusion and an upwind scheme for advection. This yields a tridiagonal state-transition matrix with periodic wrap-around and a set of $p = 10$ uniformly spaced sensors. At a moderate dimension ($n = 100$), RK-EM and classical EM produce nearly identical likelihoods, and their transfer-function fits differ by only about half a percent. Table III summarizes the results averaged over 20 trials; RK-EM reduces memory by

roughly a factor of two while retaining a faithful reconstruction of $G(z)$.

TABLE III
TRANSFER-FUNCTION ACCURACY FOR $n = 100$, $p = 10$, $N = 500$.

Method	$\mathrm{LL}_{\mathrm{final}}$	FRE (%)	Mem (MB)
EM	-678.2	3.21	3.45
RK ($w = 2$)	-680.5	3.74	1.82

Varying the assumed stencil width w_{fit} confirms a clear asymmetry: severe under-specification ($w_{\mathrm{fit}} = 0$) raises FRE to 46%, while mild over-parameterization ($w_{\mathrm{fit}} = w_{\mathrm{true}} + 1$) reduces FRE slightly (from 4.10% to 3.74%). Choosing $w_{\mathrm{fit}} = w_{\mathrm{true}} + 1$ is safe in practice.

Next, we examine how performance scales as the state dimension grows from $n = 50$ to $n = 1024$. Classical EM quickly becomes memory-limited, whereas RK-EM increases linearly and remains well below practical memory thresholds. The transfer-function error increases only modestly, indicating that the essential spectral structure is preserved even when the underlying optimization relies on iterative solves. The results appear in Table IV.

TABLE IV
SCALING WITH STATE DIMENSION ($p = 10$, $N = 500$).

n	EM Mem	EM FRE	RK Mem	RK FRE
50	1.2 MB	3.54%	0.8 MB	3.82%
100	3.5 MB	3.21%	1.8 MB	3.74%
256	18.3 MB	3.46%	4.2 MB	3.95%
512	68.1 MB	3.58%	8.1 MB	4.12%
1024	272 MB	3.67%	16.5 MB	4.29%

To test the method on a more challenging operator, we repeat the study on a Fisher–KPP linearization, which yields a substantially stiffer system. Even though the condition number increases by roughly a factor of four, RK-EM remains stable, requiring additional iterations but introducing only a small increase in FRE (Table V). This suggests that the method tolerates moderate ill-conditioning without the help of preconditioners.

TABLE V
ADVECTION–DIFFUSION VS. REACTION–DIFFUSION ($n = 1024$).

System	$\kappa(A)$	FRE (%)	RK iters/step
Advection–diffusion	12.3	4.29	320
Reaction–diffusion	48.7	4.88	480

Finally, we explore the extreme large-scale regime. Classical EM cannot run beyond a few thousand states due to its n^2 memory footprint. RK-EM remains feasible even at $n = 65{,}536$, where a single dense covariance matrix alone occupies tens of gigabytes (see Table VI caption), yet recovers the system's transfer function with only a small spectral error (Table VI).

TABLE VI
LARGE-SCALE TRANSFER-FUNCTION ACCURACY ($p = 10$, $N = 500$). EM MEMORY REPORTS THE SINGLE-COVARIANCE FOOTPRINT; STORING ALL N SMOOTHING COVARIANCES WOULD MULTIPLY THIS BY N (67 GB AT $n = 4{,}096$; 17 TB AT $n = 65{,}536$).

n	EM Mem (1 cov.)	RK Mem	FRE (%)
4,096	0.13 GB	17 MB	2.8%
65,536	34 GB	260 MB	2.5%

VII. CONCLUSIONS

This paper presented a matrix-free EM algorithm that replaces the covariance-based smoother with an information-form scheme solved by randomized Kaczmarz iterations, reducing memory to linear growth in the state dimension for models with sparse or banded structure. Theoretical bounds show that the inexact solves introduce only controllable deviation from classical EM, with standard monotonicity recovered as tolerances tighten. Numerical experiments confirm that the algorithm retains EM-level accuracy while remaining feasible at scales where traditional implementations fail, including problems with more than 6×10^4 states.

REFERENCES

[1] J. C. C. Chan, A. Poon, and D. Zhu, "High-dimensional conditionally Gaussian state space models with missing data," *J. Econometrics*, 2023.

[2] R. H. Shumway and D. S. Stoffer, "An approach to time series smoothing and forecasting using the EM algorithm," *J. Time Series Anal.*, vol. 3, no. 4, pp. 253–264, 1982.

[3] Z. Ghahramani and G. E. Hinton, "Parameter estimation for linear dynamical systems," Tech. Rep. CRG-TR-96-2, U. Toronto, 1996.

[4] S. Gibson and B. Ninness, "Robust maximum-likelihood estimation of multivariable dynamic systems," *Automatica*, vol. 41, no. 10, pp. 1667–1682, 2005.

[5] B. A. Surya, "Maximum likelihood recursive state estimation: An incomplete-information based approach," *Automatica*, vol. 168, 2024.

[6] P. Van Overschee and B. De Moor, *Subspace Identification for Linear Systems*, Kluwer, 1996.

[7] L. Ljung, *System Identification: Theory for the User*, 2nd ed., Prentice Hall, 1999.

[8] G. Pillonetto et al., "Kernel methods in system identification, machine learning and Gaussian processes," *Automatica*, vol. 50, no. 3, pp. 607–628, 2014.

[9] J. You and C. Yu, "Sparse plus low-rank identification for dynamical latent-variable graphical AR models," *Automatica*, vol. 155, 2023.

[10] G. Evensen, *Data Assimilation: The Ensemble Kalman Filter*, Springer, 2009.

[11] P. L. Houtekamer and H. L. Mitchell, "A sequential ensemble Kalman filter for atmospheric data assimilation," *Mon. Weather Rev.*, vol. 129, no. 1, pp. 123–137, 2001.

[12] T. Strohmer and R. Vershynin, "A randomized Kaczmarz algorithm with exponential convergence," *J. Fourier Anal. Appl.*, vol. 15, pp. 262–278, 2009.

[13] Z.-Z. Bai and W.-T. Wu, "On greedy randomized Kaczmarz method for solving large sparse linear systems," *SIAM J. Sci. Comput.*, vol. 40, no. 1, pp. A592–A606, 2018.

[14] L. Wen et al., "A greedy average block Kaczmarz method for the large scaled consistent system of linear equations," *AIMS Math.*, vol. 7, no. 4, pp. 6792–6806, 2022.

[15] A. P. Dempster, N. M. Laird, and D. B. Rubin, "Maximum likelihood from incomplete data via the EM algorithm," *J. R. Stat. Soc. B*, vol. 39, no. 1, pp. 1–38, 1977.

[16] H. E. Rauch, F. Tung, and C. T. Striebel, "Maximum likelihood estimates of linear dynamic systems," *AIAA J.*, vol. 3, no. 8, pp. 1445–1450, 1965.

[17] T. Kailath, A. H. Sayed, and B. Hassibi, *Linear Estimation*, Prentice Hall, 2000.

[18] M. F. Hutchinson, "A stochastic estimator of the trace of the influence matrix," *Commun. Stat. Simul. Comput.*, vol. 18, pp. 1059–1076, 1989.

[19] J. A. Tropp, "An introduction to matrix concentration inequalities," *Found. Trends Mach. Learn.*, vol. 8, pp. 1–230, 2015.