Interpretable Gradient Descent for the Kalman Gain

When first-order methods beat the Riccati equation

Alex Olshevsky & Mohamed-Ali Belabbas

OUTLINE

This Talk

Can we recover the Kalman gain without solving the Riccati equation?

— **Gradient formula:** observability $\times$ orthogonality violation

— **Convergence:** geometric rate under (A, CA) observable

— **Sparse GD:** $O(mnp^2)$ per iteration vs. $O(n^3)$ for Riccati

— **Experiments:** up to 34$\times$ speedup on large sparse systems

BACKGROUND

Tracking Linear State-Space Models

Consider a discrete-time system with hidden state and noisy observations:

$$x_{t+1} = A\,x_t + w_t, \qquad y_{t+1} = C\,x_{t+1} + v_{t+1}$$

- $x_t \in \mathbb{R}^n$ — hidden state
- $y_t \in \mathbb{R}^p$ — observation
- $w_t \sim (0, Q_w),\ Q_w \succ 0$ — process noise
- $v_t \sim (0, R_v),\ R_v \succ 0$ — measurement noise

Neither x_t nor the noises are directly observed. Goal is to estimate x_t from current and past y_t.

BACKGROUND

The Kalman Filter: Predict, then Correct

A linear filter producing a state estimate $\hat{x}_t$:

$$
\begin{aligned}
\hat{x}_{t+1}^{-} &= A\,\hat{x}_t && \text{(predict)} \\
\delta_{t+1} &= y_{t+1} - C\,\hat{x}_{t+1}^{-} && \text{(innovation)} \\
\hat{x}_{t+1} &= \hat{x}_{t+1}^{-} + L\,\delta_{t+1} && \text{(correct)}
\end{aligned}
$$

Can we do gradient descent on L to compute the optimal Kalman gain?

BACKGROUND

The Gain Controls Everything

The **gain matrix** $L \in \mathbb{R}^{n \times p}$ determines filter behavior. The estimation error $e_t = x_t - \hat{x}_t$ evolves as:

$$e_{t+1} = \underbrace{(I - LC)A}_{=:\, F(L)} \; e_t + \eta_t(L)$$

We call $\mathcal{S} = \{L : \rho(F(L)) < 1\}$ the set of **stabilizing gains**.

Classical approach: solve the **discrete algebraic Riccati equation** for the optimal error covariance P, extract the optimal gain L_{KF}. Exact, $O(n^3)$.

BACKGROUND

The Orthogonality Principle

A fundamental property of the optimal Kalman gain: at $L = L_{\mathrm{KF}}$, the estimation error and the innovation are **uncorrelated** in steady state.

Define the steady-state cross-covariance:

$$\Gamma(L) \; := \; \lim_{t \to \infty} \mathbb{E}\left[e_t \, \delta_t^\top\right]$$

The orthogonality principle states: $\Gamma(L_{\mathrm{KF}}) = 0$.

Recall: $e_t = x_t - \hat{x}_t$

Recall: $\delta_t = y_t - C\hat{x}_t^-$

PRIOR WORK

Fazel, Ge, Kakade, Mesbahi (2018): Policy Gradient for LQR

The **Linear Quadratic Regulator** (LQR) minimizes $C = \sum (x_t^\top Q\, x_t + u_t^\top R\, u_t)$ for $x_{t+1} = Ax_t + Bu_t$ over linear feedback $u_t = -Kx_t$. The optimal K^* is classically found via the Riccati equation. Fazel et al. asked: what if we skip Riccati and run **gradient descent directly on K**?

The cost $C(K)$ is non-convex, but satisfies **gradient domination** (Polyak–Łojasiewicz):

$$C(K) - C(K^*) \;\leq\; \frac{\|\Sigma_{K^*}\|}{\sigma_{\min}(\Sigma_0)^2\, \sigma_{\min}(R)}\, \|\nabla C(K)\|_F^2$$

This guarantees global convergence of GD to K^*. Natural question: can we do the same for estimation?

MORE PRIOR WORK

From Control to Estimation

LQR: find the best *control* gain. Kalman filtering: find the best *estimation* gain. There is a classical duality between the two.

Dualizing Fazel's results does not work. As observed by Talebi, Taghavaei, and Mesbahi (2023), key nonsingularity assumptions in the LQR analysis are not preserved under the duality transformation. A direct analysis is needed.

Umenberger, Simchowitz, Perdomo, Zhang, and Tedrake (2022) study the harder problem where the matrix A is unknown. They find a complex structure of the optimization landscape and get an $O(1/T)$ convergence rate via a regularizer.

SETUP

The Innovation Loss

Recall the innovation $\delta_{t+1} = y_{t+1} - C\hat{x}^{-}_{t+1}$, the one-step-ahead prediction error. We optimize over the gain L using the cost:

$$J_{\text{innov}}(L) \;=\; \lim_{t\to\infty} \mathbb{E}\left[\|\delta_{t+1}\|^2\right]$$

— Measures the steady-state magnitude of the prediction error

— Depends only on **observable data** — no access to the hidden state needed

— Minimized at the optimal Kalman gain L_{KF}

THIS PRESENTATION

Goals

- When does gradient descent on the innovation loss recover L_{KF}?
- Can it be cheaper than solving the Riccati equation?

A SUBTLETY

What Can the Innovation See?

The error and innovation dynamics in steady state:

$$e_{t+1} = F(L)\, e_t + \text{noise}, \qquad \delta_{t+1} = CA\, e_t + \text{noise}$$

Recall $F(L) = (I - LC)A$.

Observability of $(F(L), CA)$ is clearly a problem: singular directions of CA which are also eigenvectors of $F(L)$ are zeroed out by these dynamics.

MAIN RESULT 1

The Gradient Has a Clean Decomposition

$$\nabla_L J_{\text{innov}}(L) = -2 \underbrace{W_o(L)}_{\text{observability}} \quad \underbrace{\Gamma(L)}_{\text{orthogonality violation}}$$

$W_o(L)$ — Observability Gramian

$W_o(L)$ is Gramian of the pair $\big(F(L),\ CA\big)$

Formally, $W_0 = \sum_m (F(L)^T)^m (CA)^T (CA) F(L)$.

Can the observations see the estimation error?

$\Gamma(L)$ — Cross-Covariance

$\Gamma(L) = \lim_{t\to\infty} \mathbb{E}[e_t\, \delta_t^\top]$

A SUBTLETY

Spurious Stationary Points Exist

COUNTEREXAMPLE

$$x_{t+1} = \begin{pmatrix} 0 & 1 \\ 0 & 0 \end{pmatrix} x_t + w_t, \qquad y_t = \begin{pmatrix} 1 & 0 \end{pmatrix} x_t + v_t, \qquad w_t \sim \mathcal{N}(0, I), \;\; v_t \sim \mathcal{N}(0, 1).$$

(A, C) is **observable** ✓

$(F(L), CA)$ is **not observable** ✗

The cost landscape has a **flat direction**: every $L = (\ell_1, 0)^\top$ is stationary.

Gradient descent initialized there stays forever and the Kalman gain is not recovered.

Convergence of Gradient Descent

Assume (A, CA) is observable. Let $L(0) \in \mathcal{S}$ be any stabilizing initialization, and define the level set

$$\mathcal{S}_0 = \{L \in \mathcal{S} : J_{\text{innov}}(L) \leq J_{\text{innov}}(L(0))\}$$

For

$$\dot{L} = -\nabla J_{\text{innov}}(L)$$

we have

$$J_{\text{innov}}(L(t)) - J_{\text{innov}}(L_{\text{KF}}) \;\leq\; e^{-4(\kappa^2/c)\,t}\;\big(J_{\text{innov}}(L(0)) - J_{\text{innov}}(L_{\text{KF}})\big)$$

where

$\kappa = \inf_{L \in \mathcal{S}_0} \lambda_{\min}\big(W_o(L)\big)$

$c = \sup_{L \in \mathcal{S}_0 \setminus \{L_{\text{KF}}\}} \frac{J(L) - J(L_{\text{KF}})}{\|\Gamma(L)\|_F^2}$

MAIN RESULT 2

The Level Set $\mathcal{S}_0$ Is Compact

For the rate to be meaningful, we need $\kappa > 0$ and $c < \infty$. Both follow from the **compactness of $\mathcal{S}_0$**, which in turn requires showing that J_{innov} is coercive:

$$J_{\text{innov}}(L) \to \infty \quad \text{as} \quad \|L\| \to \infty \;\text{ or }\; \rho(F(L)) \to 1^-$$

This is where the assumption that (A, CA) is observable is essential. Recall the counterexample: when (A, CA) is not observable, the level sets are *not* compact and the cost has flat directions.

Proving coercivity turns out to be the hardest part of the analysis. The difficulty: as $\|L\| \to \infty$, the system can have modes that asymptotically become unobservable, and one must show the noise cannot "hide within" these modes.

PROOF SKETCH

Why the Cost Is Coercive

Goal: show $J_{\text{innov}}(L) \to \infty$ as $\|L\|_F \to \infty$.

As $\|L\| \to \infty$, the effective noise $\eta(L)$ grows as $\|L\|^2$. For the cost to stay bounded, all of this noise must land in directions **invisible** to CA.

But the correction $L \cdot CA$ in $F(L) = A - L{\cdot}CA$ acts **through** CA: it can only deflect directions that CA already sees.

Hidden directions evolve under the free dynamics A unimpeded. Observability of (A, CA) guarantees A eventually rotates every direction into CA's field of view, so noise has **nowhere to hide permanently**.

Formally: the smallest p with $CA^pL^* \neq 0$ is the “relative degree defined as the number of steps before the hidden direction becomes visible. An induction on p yields the contradiction.

COMPUTATIONAL IMPLICATIONS

From Theory to Computation

What does gradient descent actually buy us?

Both W_o and Γ solve Lyapunov equations. Solving those directly is $O(n^3)$. But because $\rho(F(L)) < 1$, both admit convergent **Neumann series**:

$$W_o(L) = \sum_{j=0}^{\infty} (F^\top)^j (CA)^\top (CA)\, F^j, \qquad \Gamma(L) = (I - LC) Q_w C^\top - L R_v + \sum_{t=0}^{\infty} F^{t+1}\, Q_\eta\, (F^\top)^t (CA)^\top$$

If $\|F^k\| \leq c\, r^k$ for some $r < 1$, truncating at term m introduces error $O(r^{2m})$

COMPUTATIONAL IMPLICATIONS

Complexity Under Sparsity

Assume:

— A sparse with $\|A\|_{\ell_0} = O(n)$ (banded, graph Laplacian, PDE discretization)

— $C \in \mathbb{R}^{p\times n}$ sparse with $\|C\|_{\ell_0} = O(n)$ (few sensors, $p \ll n$)

— $Q_w = q\, I_n,\;\; R_v = r\, I_p$

THEOREM

Under the above assumptions, a single gradient step can be computed in $O(mnp^2)$ time using only sparse matrix–vector products, where m is a truncation horizon for the Neumann series.

No dense $n \times n$ matrices are ever formed or stored.

Key idea: the gradient $-2\, W_o(L)\,\Gamma(L)$ has convergent Neumann series whose terms involve only $F(L)$, $F(L)^\top$, C, L applied to $n \times p$ blocks.

EXPERIMENTS

Four Test Systems

Damped Heat Diffusion

$A = \gamma(I + \alpha L_\Delta)$, 2D grid, $n = N^2$. 20% sensors.

Advection–Diffusion

$A = \gamma(I + \alpha L_\Delta + \beta G)$, upwind advection. Non-normal dynamics.

Graph Diffusion

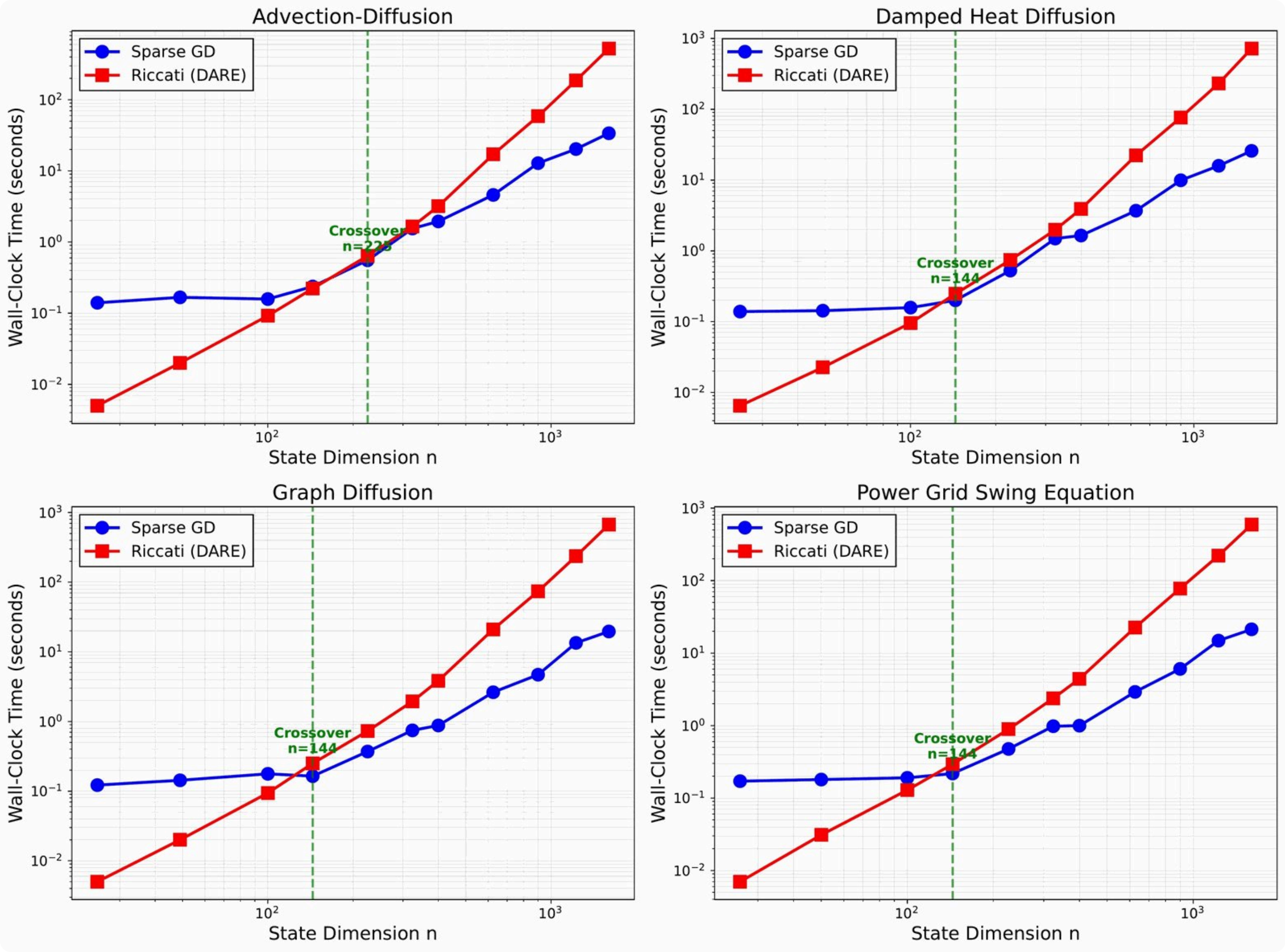
$A = (1 - \gamma)I - \alpha L_G$, random graph, avg degree 6. 10% sensors.

Power Grid Swing Equation

Block dynamics $x = [\delta; \omega]$, $n = 2n_{\text{bus}}$. 20% sensors.

All have sparse A with $O(n)$ nonzeros. Dimensions: $n = 25$ to 1600.

Sparse GD vs. Riccati: Wall-Clock Time



Log-log scale. Green dashed line marks the crossover dimension where Sparse GD becomes faster.

EXPERIMENTS

Sparse GD vs. Riccati: Wall-Clock Time

System	Crossover	Speedup (n=1600)	GD	Riccati
Graph diffusion	n ≈ 144	**34×**	20 s	670 s
Power grid	n ≈ 144	**28×**	21 s	593 s
Heat diffusion	n ≈ 144	**28×**	26 s	718 s
Advection–diffusion	n ≈ 225	**15×**	34 s	524 s

Gap widens with n: $O(n)$ vs. $O(n^3)$ scaling.

Riccati: SciPy `solve_discrete_are` (Schur method, same as MATLAB `idare`).

SUMMARY

Takeaways

- The gradient of the innovation loss decomposes as **observability × orthogonality violation**
- Under observability of (A, CA), GD converges globally to L_{KF} at an interpretable geometric rate
- Observability of (A, C) alone is *not* sufficient: spurious stationary points exist
- The decomposition enables **sparse gradient computation**: $O(mnp^2)$ per iteration vs. $O(n^3)$ to solve the full Riccati equation
- In the sparse, large-scale regime: up to **34× faster** than the state-of-the-art Riccati solver

LOOKING FORWARD

Open Questions

- **Model-free variant** — learn the gain from data alone, without knowing A, C, Q_w, R_v
- **Time-varying systems** — can GD track a moving Kalman gain?
- **Natural gradient for filtering** — the analog of Fazel et al.'s natural PG and Gauss–Newton results
- **Weakening the assumption** — performance bounds when (A, CA) observable does not hold.