Northeastern University

Carnegie Mellon University

NEURAL INFORMATION PROCESSING SYSTEMS

Markovian Experimental Design under Concept Drift

Ayberk Yarkın Yıldız[1] Lili Su[1] Carlee Joe-Wong[2] Edmund Yeh[1] Stratis Ioannidis[1]

[1]Northeastern University [2]Carnegie Mellon University

NeurIPS 2026 · Atlanta

Motivation

- Deployed models face **concept drift**: the input and label relationship changes over time.
- Recommender systems, wireless channels, air quality, finance.
- Models are retrained continuously, on a **subset of recent data**.



Drift in the wild: cellular measurements along a 4,500 km drive (AT&T datasets; route from Chen et al., 2025).

Which samples should the learner collect, so that its model stays accurate *while it drifts*?

Classic experimental design assumes a **fixed** model.

Motivation

- Deployed models face **concept drift**: the input and label relationship changes over time.
- Recommender systems, wireless channels, air quality, finance.
- Models are retrained continuously, on a **subset of recent data**.



Drift in the wild: cellular measurements along a 4,500 km drive (AT&T datasets; route from Chen et al., 2025).

Which samples should the learner collect, so that its model stays accurate *while it drifts*?

Classic experimental design assumes a **fixed** model.

Motivation

- Deployed models face **concept drift**: the input and label relationship changes over time.
- Recommender systems, wireless channels, air quality, finance.
- Models are retrained continuously, on a **subset of recent data**.



Drift in the wild: cellular measurements along a 4,500 km drive (AT&T datasets; route from Chen et al., 2025).

Which samples should the learner collect, so that its model stays accurate *while it drifts*?

Classic experimental design assumes a **fixed** model.

Motivation

- Deployed models face **concept drift**: the input and label relationship changes over time.
- Recommender systems, wireless channels, air quality, finance.
- Models are retrained continuously, on a **subset of recent data**.

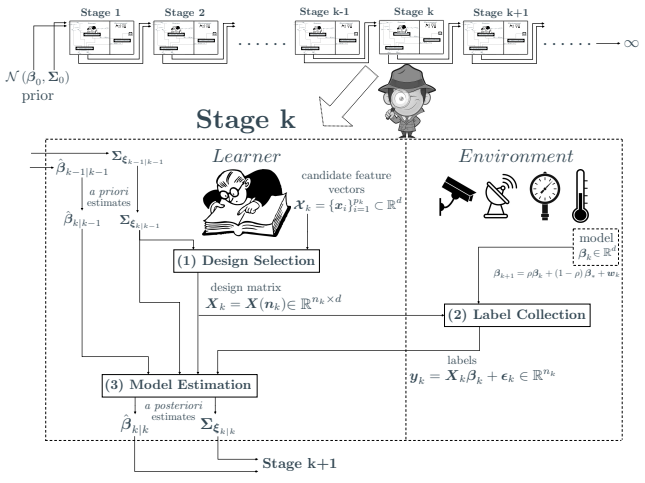
Which samples should the learner collect, so that its model stays accurate *while it drifts*?

Classic experimental design assumes a **fixed** model.



Drift in the wild: cellular measurements along a 4,500 km drive (AT&T datasets; route from Chen et al., 2025).

Markovian Experimental Design



In each stage k, the learner:

(1) **selects** a design $\boldsymbol{X}_k$ from candidate features,

(2) **collects labels** from the current model β_k,

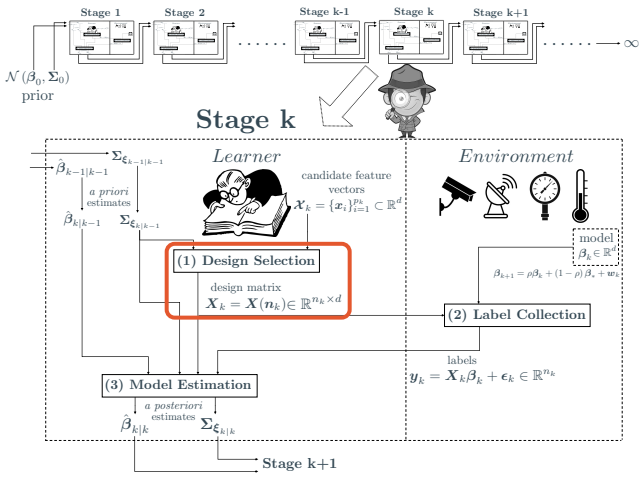
(3) **updates** its estimates $\hat{\beta}_{k|k}$, $\boldsymbol{\Sigma}_{\boldsymbol{\xi}_{k|k}}$.

Before the next stage, the model **drifts**.

This repeats over stages $k = 1, 2, \ldots$

Goal: make step (1) **optimal**, per stage or in steady state.

Markovian Experimental Design



In each stage k, the learner:

(1) **selects** a design X_k from candidate features,

(2) **collects labels** from the current model β_k,

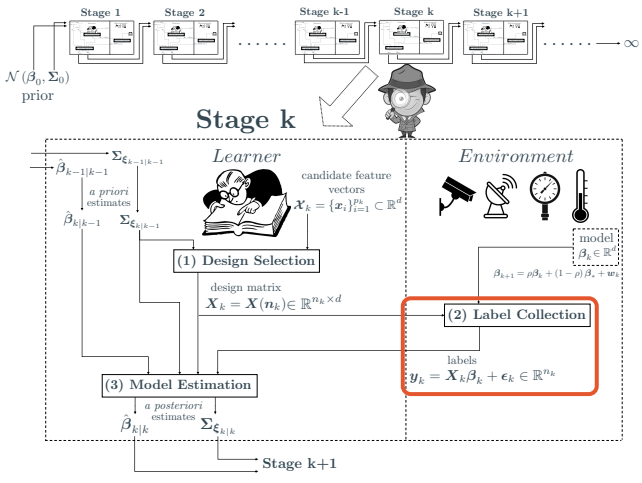
(3) **updates** its estimates $\hat{\beta}_{k|k}$, $\Sigma_{\xi_{k|k}}$.

Before the next stage, the model **drifts**.

This repeats over stages $k = 1, 2, \ldots$

Goal: make step (1) **optimal**, per stage or in steady state.

Markovian Experimental Design



In each stage k, the learner:

(1) **selects** a design $\boldsymbol{X}_k$ from candidate features,

(2) **collects labels** from the current model β_k,

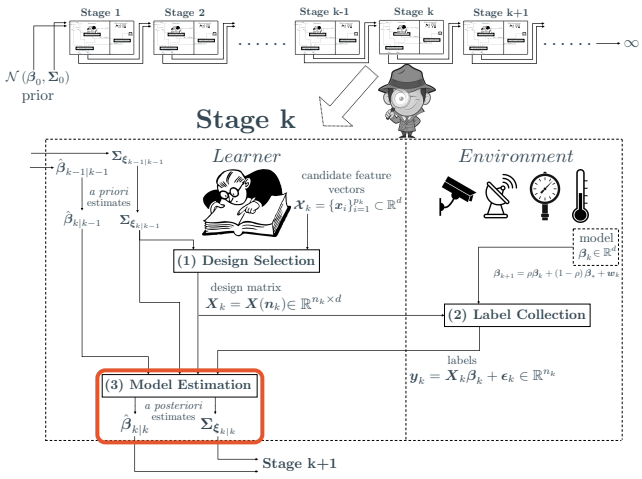
(3) **updates** its estimates $\hat{\beta}_{k|k}$, $\boldsymbol{\Sigma}_{\boldsymbol{\xi}_{k|k}}$.

Before the next stage, the model **drifts**.

This repeats over stages $k = 1, 2, \ldots$

Goal: make step (1) **optimal**, per stage or in steady state.

Markovian Experimental Design



In each stage k, the learner:

(1) **selects** a design $\boldsymbol{X}_k$ from candidate features,

(2) **collects labels** from the current model β_k,

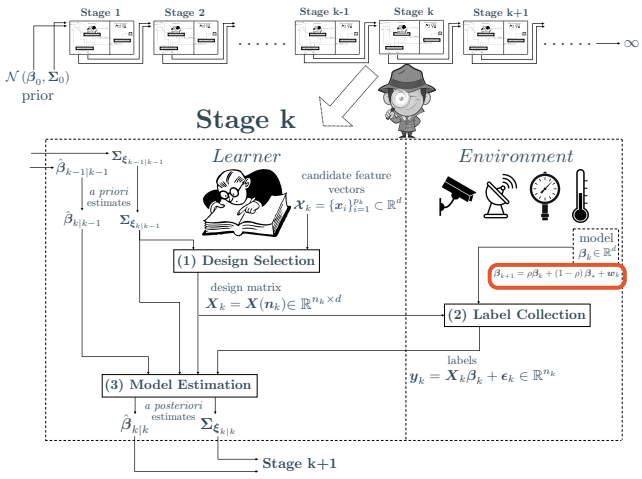
(3) **updates** its estimates $\hat{\beta}_{k|k}$, $\Sigma_{\xi_{k|k}}$.

Before the next stage, the model **drifts**.

This repeats over stages $k = 1, 2, \ldots$

Goal: make step (1) **optimal**, per stage or in steady state.

Markovian Experimental Design



In each stage k, the learner:

(1) **selects** a design $\boldsymbol{X}_k$ from candidate features,

(2) **collects labels** from the current model β_k,

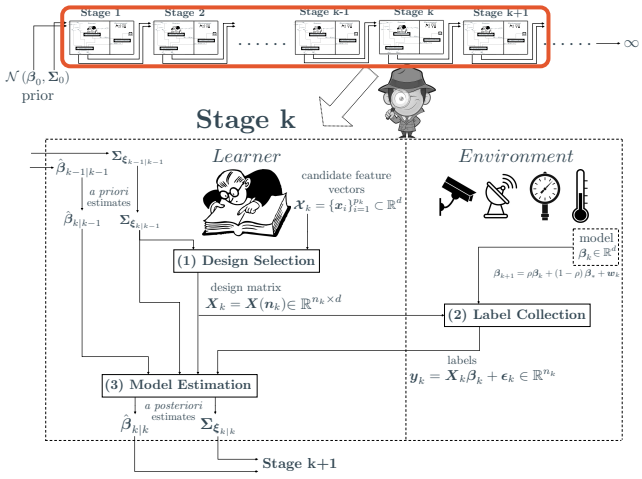
(3) **updates** its estimates $\hat{\beta}_{k|k}$, $\boldsymbol{\Sigma}_{\boldsymbol{\xi}_{k|k}}$.

Before the next stage, the model **drifts**.

This repeats over stages $k = 1, 2, \ldots$

Goal: make step (1) **optimal**, per stage or in steady state.

Markovian Experimental Design



In each stage k, the learner:

(1) **selects** a design $\boldsymbol{X}_k$ from candidate features,

(2) **collects labels** from the current model β_k,

(3) **updates** its estimates $\hat{\beta}_{k|k}$, $\Sigma_{\xi_{k|k}}$.

Before the next stage, the model **drifts**.

This repeats over stages $k = 1, 2, \ldots$

Goal: make step (1) **optimal**, per stage or in steady state.

Contributions

- **Drift model:** a mean-reverting Gaussian process around an unknown **mean fixed-point** β_*.

Contributions

- **Drift model:** a mean-reverting Gaussian process around an unknown **mean fixed-point** β_*.
- **Linear models:** Kalman filter posteriors, a consistent estimate of β_*, and a steady state given by the Discrete Algebraic Riccati Equation (DARE).

Contributions

- **Drift model:** a mean-reverting Gaussian process around an unknown **mean fixed-point** β_*.
- **Linear models:** Kalman filter posteriors, a consistent estimate of β_*, and a steady state given by the Discrete Algebraic Riccati Equation (DARE).
- **Convex** D-optimal designs, both **myopic** and **steady-state**, via an explicit solution of the DARE.

Contributions

- **Drift model:** a mean-reverting Gaussian process around an unknown **mean fixed-point** β_*.
- **Linear models:** Kalman filter posteriors, a consistent estimate of β_*, and a steady state given by the Discrete Algebraic Riccati Equation (DARE).
- **Convex** D-optimal designs, both **myopic** and **steady-state**, via an explicit solution of the DARE.
- **Non-linear models** via an extended Kalman filter; the lowest test RMSE on 7 datasets and 4 model classes.

Preliminaries

Bayesian linear regression: $\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$, $\boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \sigma^2\mathbf{I})$, prior $\beta \sim \mathcal{N}(\beta_0, \boldsymbol{\Sigma}_0)$.
Covariance of the error $\boldsymbol{\xi} = \hat{\boldsymbol{\beta}} - \boldsymbol{\beta}$, and expected prediction error at a test point $\boldsymbol{x}$:

$$\boldsymbol{\Sigma}_{\boldsymbol{\xi}} = \left(\tfrac{1}{\sigma^2}\boldsymbol{X}^\top\boldsymbol{X} + \boldsymbol{\Sigma}_0^{-1}\right)^{-1}, \qquad \mathrm{EPE}(\boldsymbol{x}) = \sigma^2 + \boldsymbol{x}^\top\boldsymbol{\Sigma}_{\boldsymbol{\xi}}\,\boldsymbol{x}.$$

D-optimal design: pick the fraction π_i of the n samples taken at each candidate $\boldsymbol{x}_i$:

$$\max_{\boldsymbol{\pi}\in\Delta}\ \log\det\left[\tfrac{n}{\sigma^2}\boldsymbol{B}(\boldsymbol{\pi}) + \boldsymbol{\Sigma}_0^{-1}\right], \qquad \boldsymbol{B}(\boldsymbol{\pi}) = \textstyle\sum_i \pi_i \boldsymbol{x}_i\boldsymbol{x}_i^\top,$$

over the simplex $\Delta = \{\boldsymbol{\pi} \geq \mathbf{0} : \boldsymbol{\pi}^\top\mathbf{1} \leq 1\}$. Concave in $\boldsymbol{\pi}$, so a convex problem, but for a **fixed** model β.

Preliminaries

Bayesian linear regression: $\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$, $\boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \sigma^2\mathbf{I})$, prior $\beta \sim \mathcal{N}(\beta_0, \boldsymbol{\Sigma}_0)$.
Covariance of the error $\boldsymbol{\xi} = \hat{\beta} - \beta$, and expected prediction error at a test point $\boldsymbol{x}$:

$$\boldsymbol{\Sigma}_{\boldsymbol{\xi}} = \left(\tfrac{1}{\sigma^2}\boldsymbol{X}^\top\boldsymbol{X} + \boldsymbol{\Sigma}_0^{-1}\right)^{-1}, \qquad \mathrm{EPE}(\boldsymbol{x}) = \sigma^2 + \boldsymbol{x}^\top\boldsymbol{\Sigma}_{\boldsymbol{\xi}}\,\boldsymbol{x}.$$

D-optimal design: pick the fraction π_i of the n samples taken at each candidate $\boldsymbol{x}_i$:

$$\max_{\boldsymbol{\pi}\in\Delta} \log\det\left[\tfrac{n}{\sigma^2}\boldsymbol{B}(\boldsymbol{\pi}) + \boldsymbol{\Sigma}_0^{-1}\right], \qquad \boldsymbol{B}(\boldsymbol{\pi}) = \textstyle\sum_i \pi_i \boldsymbol{x}_i\boldsymbol{x}_i^\top,$$

over the simplex $\Delta = \{\boldsymbol{\pi} \geq \mathbf{0} : \boldsymbol{\pi}^\top\mathbf{1} \leq 1\}$. Concave in $\boldsymbol{\pi}$, so a convex problem, but for a **fixed** model β.

Preliminaries

Bayesian linear regression: $\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$, $\boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \sigma^2\mathbf{I})$, prior $\beta \sim \mathcal{N}(\beta_0, \boldsymbol{\Sigma}_0)$.
Covariance of the error $\boldsymbol{\xi} = \hat{\beta} - \beta$, and expected prediction error at a test point $\boldsymbol{x}$:

$$\boldsymbol{\Sigma}_{\boldsymbol{\xi}} = \left(\tfrac{1}{\sigma^2}\boldsymbol{X}^\top\boldsymbol{X} + \boldsymbol{\Sigma}_0^{-1}\right)^{-1}, \qquad \mathrm{EPE}(\boldsymbol{x}) = \sigma^2 + \boldsymbol{x}^\top\boldsymbol{\Sigma}_{\boldsymbol{\xi}}\,\boldsymbol{x}.$$

D-optimal design: pick the fraction π_i of the n samples taken at each candidate $\boldsymbol{x}_i$:

$$\max_{\boldsymbol{\pi}\in\Delta} \log\det\left[\tfrac{n}{\sigma^2}\boldsymbol{B}(\boldsymbol{\pi}) + \boldsymbol{\Sigma}_0^{-1}\right], \qquad \boldsymbol{B}(\boldsymbol{\pi}) = \textstyle\sum_i \pi_i \boldsymbol{x}_i\boldsymbol{x}_i^\top,$$

over the simplex $\Delta = \{\boldsymbol{\pi} \geq \mathbf{0} : \boldsymbol{\pi}^\top\mathbf{1} \leq 1\}$. Concave in $\boldsymbol{\pi}$, so a convex problem, but for a **fixed** model β.

Preliminaries

Bayesian linear regression: $\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$, $\boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \sigma^2\mathbf{I})$, prior $\beta \sim \mathcal{N}(\beta_0, \boldsymbol{\Sigma}_0)$.
Covariance of the error $\boldsymbol{\xi} = \hat{\boldsymbol{\beta}} - \boldsymbol{\beta}$, and expected prediction error at a test point $\boldsymbol{x}$:

$$\boldsymbol{\Sigma}_{\boldsymbol{\xi}} = \left(\tfrac{1}{\sigma^2}\boldsymbol{X}^\top\boldsymbol{X} + \boldsymbol{\Sigma}_0^{-1}\right)^{-1}, \qquad \mathrm{EPE}(\boldsymbol{x}) = \sigma^2 + \boldsymbol{x}^\top\boldsymbol{\Sigma}_{\boldsymbol{\xi}}\,\boldsymbol{x}.$$

D-optimal design: pick the fraction π_i of the n samples taken at each candidate $\boldsymbol{x}_i$:

$$\max_{\boldsymbol{\pi}\in\Delta} \log\det\left[\tfrac{n}{\sigma^2}\boldsymbol{B}(\boldsymbol{\pi}) + \boldsymbol{\Sigma}_0^{-1}\right], \qquad \boldsymbol{B}(\boldsymbol{\pi}) = \textstyle\sum_i \pi_i \boldsymbol{x}_i\boldsymbol{x}_i^\top,$$

over the simplex $\Delta = \{\boldsymbol{\pi} \geq \mathbf{0} : \boldsymbol{\pi}^\top\mathbf{1} \leq 1\}$. Concave in $\boldsymbol{\pi}$, so a convex problem, but for a **fixed** model β.

Preliminaries

Bayesian linear regression: $\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$, $\boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \sigma^2\mathbf{I})$, prior $\beta \sim \mathcal{N}(\beta_0, \boldsymbol{\Sigma}_0)$.
Covariance of the error $\boldsymbol{\xi} = \hat{\beta} - \beta$, and expected prediction error at a test point $\boldsymbol{x}$:

$$\boldsymbol{\Sigma}_{\boldsymbol{\xi}} = \left(\tfrac{1}{\sigma^2}\boldsymbol{X}^\top\boldsymbol{X} + \boldsymbol{\Sigma}_0^{-1}\right)^{-1}, \qquad \mathrm{EPE}(\boldsymbol{x}) = \sigma^2 + \boldsymbol{x}^\top\boldsymbol{\Sigma}_{\boldsymbol{\xi}}\,\boldsymbol{x}.$$

D-optimal design: pick the fraction π_i of the n samples taken at each candidate $\boldsymbol{x}_i$:

$$\max_{\boldsymbol{\pi}\in\Delta} \log\det\left[\tfrac{n}{\sigma^2}\boldsymbol{B}(\boldsymbol{\pi}) + \boldsymbol{\Sigma}_0^{-1}\right], \qquad \boldsymbol{B}(\boldsymbol{\pi}) = \textstyle\sum_i \pi_i \boldsymbol{x}_i\boldsymbol{x}_i^\top,$$

over the simplex $\Delta = \{\boldsymbol{\pi} \geq \mathbf{0} : \boldsymbol{\pi}^\top\mathbf{1} \leq 1\}$. Concave in $\boldsymbol{\pi}$, so a convex problem, but for a **fixed** model β.

Modeling Drift

Mean-reverting drift:

$$\boldsymbol{\beta}_{k+1} = \rho\,\boldsymbol{\beta}_k + (1-\rho)\,\boldsymbol{\beta}_* + \boldsymbol{w}_k, \qquad \rho \in (0,1), \quad \boldsymbol{w}_k \sim \mathcal{N}(\mathbf{0}, \sigma_{\boldsymbol{w}}^2 \mathbf{I}), \quad \boldsymbol{\beta}_1 \sim \mathcal{N}(\boldsymbol{\beta}_0, \boldsymbol{\Sigma}_0).$$

Labels:

$$\boldsymbol{y}_k = \boldsymbol{X}_k \boldsymbol{\beta}_k + \boldsymbol{\epsilon}_k, \qquad \boldsymbol{\epsilon}_k \sim \mathcal{N}(\mathbf{0}, \sigma^2 \mathbf{I}).$$

- β_*: unknown **mean fixed-point**, the “status-quo” model learned when ignoring drift.

Modeling Drift

Mean-reverting drift:

$$\boldsymbol{\beta}_{k+1} = \rho\,\boldsymbol{\beta}_k + (1-\rho)\,\boldsymbol{\beta}_* + \boldsymbol{w}_k, \qquad \rho \in (0,1), \quad \boldsymbol{w}_k \sim \mathcal{N}(\mathbf{0}, \sigma_{\boldsymbol{w}}^2 \mathbf{I}), \quad \boldsymbol{\beta}_1 \sim \mathcal{N}(\boldsymbol{\beta}_0, \boldsymbol{\Sigma}_0).$$

Labels:

$$\boldsymbol{y}_k = \boldsymbol{X}_k \boldsymbol{\beta}_k + \boldsymbol{\epsilon}_k, \qquad \boldsymbol{\epsilon}_k \sim \mathcal{N}(\mathbf{0}, \sigma^2 \mathbf{I}).$$

- β_*: unknown **mean fixed-point**, the "status-quo" model learned when ignoring drift.

Modeling Drift

Mean-reverting drift:

$$\boldsymbol{\beta}_{k+1} = \rho\,\boldsymbol{\beta}_k + (1-\rho)\,\boldsymbol{\beta}_* + \boldsymbol{w}_k, \qquad \rho \in (0,1), \quad \boldsymbol{w}_k \sim \mathcal{N}(\mathbf{0}, \sigma_{\boldsymbol{w}}^2 \mathbf{I}), \quad \boldsymbol{\beta}_1 \sim \mathcal{N}(\boldsymbol{\beta}_0, \boldsymbol{\Sigma}_0).$$

Labels:

$$\boldsymbol{y}_k = \boldsymbol{X}_k \boldsymbol{\beta}_k + \boldsymbol{\epsilon}_k, \qquad \boldsymbol{\epsilon}_k \sim \mathcal{N}(\mathbf{0}, \sigma^2 \mathbf{I}).$$

- β_*: unknown **mean fixed-point**, the "status-quo" model learned when ignoring drift.

Estimation

Theorem 1: Kalman filter estimation under known β_*

For any design policy, the posterior of β_k is $\mathcal{N}(\hat{\beta}_{k|k}, \boldsymbol{\Sigma}_{\boldsymbol{\xi}_{k|k}})$, with Kalman gain $\boldsymbol{K}_k$:

$$\boldsymbol{K}_k = \boldsymbol{\Sigma}_{\boldsymbol{\xi}_{k|k-1}} \boldsymbol{X}_k^\top \left(\boldsymbol{X}_k \boldsymbol{\Sigma}_{\boldsymbol{\xi}_{k|k-1}} \boldsymbol{X}_k^\top + \sigma^2 \mathbf{I} \right)^{-1},$$

$$\hat{\beta}_{k|k-1} = \rho\, \hat{\beta}_{k-1|k-1} + (1-\rho)\, \beta_*, \qquad \boldsymbol{\Sigma}_{\boldsymbol{\xi}_{k|k-1}} = \rho^2 \boldsymbol{\Sigma}_{\boldsymbol{\xi}_{k-1|k-1}} + \sigma_{\boldsymbol{w}}^2 \mathbf{I},$$

$$\hat{\beta}_{k|k} = \hat{\beta}_{k|k-1} + \boldsymbol{K}_k \big(\boldsymbol{y}_k - \boldsymbol{X}_k \hat{\beta}_{k|k-1} \big), \qquad \boldsymbol{\Sigma}_{\boldsymbol{\xi}_{k|k}} = (\mathbf{I} - \boldsymbol{K}_k \boldsymbol{X}_k)\, \boldsymbol{\Sigma}_{\boldsymbol{\xi}_{k|k-1}}.$$

Theorem 2: Estimating an unknown β_*

If the eigenvalues of $\boldsymbol{X}_k^\top \boldsymbol{X}_k$ are bounded from above and away from zero, then $\hat{\beta}_* = \frac{1}{k} \sum_{k'=1}^{k} (\boldsymbol{X}_{k'}^\top \boldsymbol{X}_{k'})^{-1} \boldsymbol{X}_{k'}^\top \boldsymbol{y}_{k'}$ converges to β_* w.p. 1 and in expectation.

The covariance $\boldsymbol{\Sigma}_{\boldsymbol{\xi}_{k|k}}$ does **not** depend on β_*: designs need no knowledge of β_*.

Estimation

Theorem 1: Kalman filter estimation under known β_*

For any design policy, the posterior of β_k is $\mathcal{N}(\hat{\beta}_{k|k}, \Sigma_{\xi_{k|k}})$, with Kalman gain K_k:

$$K_k = \Sigma_{\xi_{k|k-1}} X_k^\top \left(X_k \Sigma_{\xi_{k|k-1}} X_k^\top + \sigma^2 \mathbf{I} \right)^{-1},$$

$$\hat{\beta}_{k|k-1} = \rho\, \hat{\beta}_{k-1|k-1} + (1-\rho)\, \beta_*, \qquad \Sigma_{\xi_{k|k-1}} = \rho^2 \Sigma_{\xi_{k-1|k-1}} + \sigma_w^2 \mathbf{I},$$

$$\hat{\beta}_{k|k} = \hat{\beta}_{k|k-1} + K_k \left(y_k - X_k \hat{\beta}_{k|k-1} \right), \qquad \Sigma_{\xi_{k|k}} = \left(\mathbf{I} - K_k X_k \right) \Sigma_{\xi_{k|k-1}}.$$

Theorem 2: Estimating an unknown β_*

If the eigenvalues of $X_k^\top X_k$ are bounded from above and away from zero, then $\hat{\beta}_* = \frac{1}{k} \sum_{k'=1}^{k} (X_{k'}^\top X_{k'})^{-1} X_{k'}^\top y_{k'}$ converges to β_* w.p. 1 and in expectation.

The covariance $\Sigma_{\xi_{k|k}}$ does **not** depend on β_*: designs need no knowledge of β_*.

Estimation

Theorem 1: Kalman filter estimation under known β_*

For any design policy, the posterior of β_k is $\mathcal{N}(\hat{\beta}_{k|k}, \boldsymbol{\Sigma}_{\boldsymbol{\xi}_{k|k}})$, with Kalman gain $\boldsymbol{K}_k$:

$$\boldsymbol{K}_k = \boldsymbol{\Sigma}_{\boldsymbol{\xi}_{k|k-1}} \boldsymbol{X}_k^\top \left(\boldsymbol{X}_k \boldsymbol{\Sigma}_{\boldsymbol{\xi}_{k|k-1}} \boldsymbol{X}_k^\top + \sigma^2 \mathbf{I} \right)^{-1},$$

$$\hat{\beta}_{k|k-1} = \rho\, \hat{\beta}_{k-1|k-1} + (1-\rho)\, \beta_*, \qquad \boldsymbol{\Sigma}_{\boldsymbol{\xi}_{k|k-1}} = \rho^2 \boldsymbol{\Sigma}_{\boldsymbol{\xi}_{k-1|k-1}} + \sigma_{\boldsymbol{w}}^2 \mathbf{I},$$

$$\hat{\beta}_{k|k} = \hat{\beta}_{k|k-1} + \boldsymbol{K}_k \left(\boldsymbol{y}_k - \boldsymbol{X}_k \hat{\beta}_{k|k-1} \right), \qquad \boldsymbol{\Sigma}_{\boldsymbol{\xi}_{k|k}} = \left(\mathbf{I} - \boldsymbol{K}_k \boldsymbol{X}_k \right) \boldsymbol{\Sigma}_{\boldsymbol{\xi}_{k|k-1}}.$$

Theorem 2: Estimating an unknown β_*

If the eigenvalues of $\boldsymbol{X}_k^\top \boldsymbol{X}_k$ are bounded from above and away from zero, then $\hat{\beta}_* = \frac{1}{k} \sum_{k'=1}^{k} (\boldsymbol{X}_{k'}^\top \boldsymbol{X}_{k'})^{-1} \boldsymbol{X}_{k'}^\top \boldsymbol{y}_{k'}$ converges to β_* w.p. 1 and in expectation.

The covariance $\boldsymbol{\Sigma}_{\boldsymbol{\xi}_{k|k}}$ does **not** depend on β_*: designs need no knowledge of β_*.

Estimation

Theorem 1: Kalman filter estimation under known β_*

For any design policy, the posterior of β_k is $\mathcal{N}(\hat{\beta}_{k|k}, \boldsymbol{\Sigma}_{\boldsymbol{\xi}_{k|k}})$, with Kalman gain $\boldsymbol{K}_k$:

$$\boldsymbol{K}_k = \boldsymbol{\Sigma}_{\boldsymbol{\xi}_{k|k-1}} \boldsymbol{X}_k^\top \left(\boldsymbol{X}_k \boldsymbol{\Sigma}_{\boldsymbol{\xi}_{k|k-1}} \boldsymbol{X}_k^\top + \sigma^2 \mathbf{I} \right)^{-1},$$

$$\hat{\beta}_{k|k-1} = \rho\, \hat{\beta}_{k-1|k-1} + (1-\rho)\, \beta_*, \qquad \boldsymbol{\Sigma}_{\boldsymbol{\xi}_{k|k-1}} = \rho^2 \boldsymbol{\Sigma}_{\boldsymbol{\xi}_{k-1|k-1}} + \sigma_{\boldsymbol{w}}^2 \mathbf{I},$$

$$\hat{\beta}_{k|k} = \hat{\beta}_{k|k-1} + \boldsymbol{K}_k \left(\boldsymbol{y}_k - \boldsymbol{X}_k \hat{\beta}_{k|k-1} \right), \qquad \boldsymbol{\Sigma}_{\boldsymbol{\xi}_{k|k}} = \left(\mathbf{I} - \boldsymbol{K}_k \boldsymbol{X}_k \right) \boldsymbol{\Sigma}_{\boldsymbol{\xi}_{k|k-1}}.$$

Theorem 2: Estimating an unknown β_*

If the eigenvalues of $\boldsymbol{X}_k^\top \boldsymbol{X}_k$ are bounded from above and away from zero, then $\hat{\beta}_* = \frac{1}{k} \sum_{k'=1}^{k} (\boldsymbol{X}_{k'}^\top \boldsymbol{X}_{k'})^{-1} \boldsymbol{X}_{k'}^\top \boldsymbol{y}_{k'}$ converges to β_* w.p. 1 and in expectation.

The covariance $\boldsymbol{\Sigma}_{\boldsymbol{\xi}_{k|k}}$ does **not** depend on β_*: designs need no knowledge of β_*.

Optimal Design Selection

Theorem 3: Optimal myopic design (STG)

At each stage k, the D-optimal fractional policy solves the convex problem

$$\max_{\boldsymbol{\pi}_k \in \Delta} \log\det\left[\tfrac{n_k}{\sigma^2}\boldsymbol{B}(\boldsymbol{\pi}_k) + \boldsymbol{\Sigma}^{-1}_{\boldsymbol{\xi}_{k|k-1}}\right].$$

Theorem 4: Convergence to steady state

Under a stationary policy $\boldsymbol{\pi}$ with a rank-d design, $\boldsymbol{\Sigma}_{\boldsymbol{\xi}_{k|k}} \to \overline{\boldsymbol{\Sigma}}_{\boldsymbol{\xi}}$, the solution of the **Discrete Algebraic Riccati Equation**:

$$\overline{\boldsymbol{\Sigma}}^{-1}_{\boldsymbol{\xi}} = \tfrac{n}{\sigma^2}\boldsymbol{B}(\boldsymbol{\pi}) + \left(\rho^2\overline{\boldsymbol{\Sigma}}_{\boldsymbol{\xi}} + \sigma^2_{\boldsymbol{w}}\mathbf{I}\right)^{-1}. \qquad \text{(DARE)}$$

Steady-state design (STS): $\max_{\boldsymbol{\pi}\in\Delta} G(\boldsymbol{\pi}) \equiv \log\det \overline{\boldsymbol{\Sigma}}^{-1}_{\boldsymbol{\xi}}(\boldsymbol{\pi})$, with $\overline{\boldsymbol{\Sigma}}_{\boldsymbol{\xi}}(\boldsymbol{\pi})$ defined only *implicitly*.

Optimal Design Selection

Theorem 3: Optimal myopic design (STG)

At each stage k, the D-optimal fractional policy solves the convex problem

$$\max_{\boldsymbol{\pi}_k \in \Delta} \log\det\left[\tfrac{n_k}{\sigma^2}\boldsymbol{B}(\boldsymbol{\pi}_k) + \boldsymbol{\Sigma}^{-1}_{\boldsymbol{\xi}_{k|k-1}}\right].$$

Theorem 4: Convergence to steady state

Under a stationary policy $\boldsymbol{\pi}$ with a rank-d design, $\boldsymbol{\Sigma}_{\boldsymbol{\xi}_{k|k}} \to \overline{\boldsymbol{\Sigma}}_{\boldsymbol{\xi}}$, the solution of the **Discrete Algebraic Riccati Equation**:

$$\overline{\boldsymbol{\Sigma}}^{-1}_{\boldsymbol{\xi}} = \tfrac{n}{\sigma^2}\boldsymbol{B}(\boldsymbol{\pi}) + \left(\rho^2\overline{\boldsymbol{\Sigma}}_{\boldsymbol{\xi}} + \sigma^2_{\boldsymbol{w}}\mathbf{I}\right)^{-1}. \qquad \text{(DARE)}$$

Steady-state design (STS): $\max_{\boldsymbol{\pi}\in\Delta} G(\boldsymbol{\pi}) \equiv \log\det \overline{\boldsymbol{\Sigma}}^{-1}_{\boldsymbol{\xi}}(\boldsymbol{\pi})$, with $\overline{\boldsymbol{\Sigma}}_{\boldsymbol{\xi}}(\boldsymbol{\pi})$ defined only *implicitly*.

Optimal Design Selection

Theorem 3: Optimal myopic design (STG)

At each stage k, the D-optimal fractional policy solves the convex problem

$$\max_{\boldsymbol{\pi}_k \in \Delta} \log\det\left[\tfrac{n_k}{\sigma^2}\boldsymbol{B}(\boldsymbol{\pi}_k) + \boldsymbol{\Sigma}^{-1}_{\boldsymbol{\xi}_{k|k-1}}\right].$$

Theorem 4: Convergence to steady state

Under a stationary policy $\boldsymbol{\pi}$ with a rank-d design, $\boldsymbol{\Sigma}_{\boldsymbol{\xi}_{k|k}} \to \overline{\boldsymbol{\Sigma}}_{\boldsymbol{\xi}}$, the solution of the **Discrete Algebraic Riccati Equation**:

$$\overline{\boldsymbol{\Sigma}}^{-1}_{\boldsymbol{\xi}} = \tfrac{n}{\sigma^2}\boldsymbol{B}(\boldsymbol{\pi}) + \left(\rho^2\overline{\boldsymbol{\Sigma}}_{\boldsymbol{\xi}} + \sigma^2_{\boldsymbol{w}}\mathbf{I}\right)^{-1}. \qquad \text{(DARE)}$$

Steady-state design (STS): $\max_{\boldsymbol{\pi}\in\Delta} G(\boldsymbol{\pi}) \equiv \log\det \overline{\boldsymbol{\Sigma}}^{-1}_{\boldsymbol{\xi}}(\boldsymbol{\pi})$, with $\overline{\boldsymbol{\Sigma}}_{\boldsymbol{\xi}}(\boldsymbol{\pi})$ defined only *implicitly*.

Optimal Design Selection

Theorem 3: Optimal myopic design (STG)

At each stage k, the D-optimal fractional policy solves the convex problem

$$\max_{\boldsymbol{\pi}_k \in \Delta} \log\det\left[\tfrac{n_k}{\sigma^2}\boldsymbol{B}(\boldsymbol{\pi}_k) + \boldsymbol{\Sigma}^{-1}_{\boldsymbol{\xi}_{k|k-1}}\right].$$

Theorem 4: Convergence to steady state

Under a stationary policy π with a rank-d design, $\boldsymbol{\Sigma}_{\boldsymbol{\xi}_{k|k}} \to \overline{\boldsymbol{\Sigma}}_{\boldsymbol{\xi}}$, the solution of the **Discrete Algebraic Riccati Equation**:

$$\overline{\boldsymbol{\Sigma}}^{-1}_{\boldsymbol{\xi}} = \tfrac{n}{\sigma^2}\boldsymbol{B}(\boldsymbol{\pi}) + \left(\rho^2\overline{\boldsymbol{\Sigma}}_{\boldsymbol{\xi}} + \sigma^2_{\boldsymbol{w}}\mathbf{I}\right)^{-1}. \qquad \text{(DARE)}$$

Steady-state design (STS): $\max_{\pi\in\Delta} G(\pi) \equiv \log\det \overline{\boldsymbol{\Sigma}}^{-1}_{\boldsymbol{\xi}}(\pi)$, with $\overline{\boldsymbol{\Sigma}}_{\boldsymbol{\xi}}(\pi)$ defined only *implicitly*.

Steady-State Guarantee

Lemma 1: Explicit DARE solution

Let $\overline{\boldsymbol{\Sigma}}_{\boldsymbol{\xi}}$ be the solution of the DARE, and let λ_i, $\boldsymbol{e}_i$ be the eigenvalues and eigenvectors of $\boldsymbol{B}(\boldsymbol{\pi})$. Then $\overline{\boldsymbol{\Sigma}}_{\boldsymbol{\xi}} = \sum_i \lambda'_i \, \boldsymbol{e}_i \boldsymbol{e}_i^\top$, with

$$\lambda'_i = \frac{2\sigma_{\boldsymbol{w}}^2 \, \frac{\sigma^2}{n}}{\sqrt{\left(\frac{\sigma^2}{n}(1-\rho^2) + \sigma_{\boldsymbol{w}}^2 \lambda_i\right)^2 + 4\lambda_i \rho^2 \sigma_{\boldsymbol{w}}^2 \frac{\sigma^2}{n}} + \frac{\sigma^2}{n}(1-\rho^2) + \sigma_{\boldsymbol{w}}^2 \lambda_i}.$$

Theorem 5: The steady-state design is convex

The steady-state objective $G(\boldsymbol{\pi}) = \log\det \overline{\boldsymbol{\Sigma}}_{\boldsymbol{\xi}}^{-1}(\boldsymbol{\pi})$ is **non-decreasing and concave**.

Key takeaway: no closed form in general, but under mean-reverting drift the DARE is explicit in the spectrum of $\boldsymbol{B}(\boldsymbol{\pi})$. We solve STS with Frank–Wolfe.

Steady-State Guarantee

Lemma 1: Explicit DARE solution

Let $\overline{\Sigma}_{\boldsymbol{\xi}}$ be the solution of the DARE, and let λ_i, $\boldsymbol{e}_i$ be the eigenvalues and eigenvectors of $\boldsymbol{B}(\boldsymbol{\pi})$. Then $\overline{\Sigma}_{\boldsymbol{\xi}} = \sum_i \lambda'_i \boldsymbol{e}_i \boldsymbol{e}_i^\top$, with

$$\lambda'_i = \frac{2\sigma_{\boldsymbol{w}}^2 \frac{\sigma^2}{n}}{\sqrt{\left(\frac{\sigma^2}{n}(1-\rho^2) + \sigma_{\boldsymbol{w}}^2 \lambda_i\right)^2 + 4\lambda_i \rho^2 \sigma_{\boldsymbol{w}}^2 \frac{\sigma^2}{n}} + \frac{\sigma^2}{n}(1-\rho^2) + \sigma_{\boldsymbol{w}}^2 \lambda_i}.$$

Theorem 5: The steady-state design is convex

The steady-state objective $G(\boldsymbol{\pi}) = \log\det \overline{\Sigma}_{\boldsymbol{\xi}}^{-1}(\boldsymbol{\pi})$ is **non-decreasing and concave**.

Key takeaway: no closed form in general, but under mean-reverting drift the DARE is explicit in the spectrum of $\boldsymbol{B}(\boldsymbol{\pi})$. We solve STS with Frank–Wolfe.

Steady-State Guarantee

Lemma 1: Explicit DARE solution

Let $\overline{\boldsymbol{\Sigma}}_{\boldsymbol{\xi}}$ be the solution of the DARE, and let λ_i, $\boldsymbol{e}_i$ be the eigenvalues and eigenvectors of $\boldsymbol{B}(\boldsymbol{\pi})$. Then $\overline{\boldsymbol{\Sigma}}_{\boldsymbol{\xi}} = \sum_i \lambda'_i \, \boldsymbol{e}_i \boldsymbol{e}_i^\top$, with

$$\lambda'_i = \frac{2\sigma_{\boldsymbol{w}}^2 \, \frac{\sigma^2}{n}}{\sqrt{\left(\frac{\sigma^2}{n}(1-\rho^2) + \sigma_{\boldsymbol{w}}^2 \lambda_i\right)^2 + 4\lambda_i \rho^2 \sigma_{\boldsymbol{w}}^2 \frac{\sigma^2}{n}} + \frac{\sigma^2}{n}(1-\rho^2) + \sigma_{\boldsymbol{w}}^2 \lambda_i}.$$

Theorem 5: The steady-state design is convex

The steady-state objective $G(\boldsymbol{\pi}) = \log\det \overline{\boldsymbol{\Sigma}}_{\boldsymbol{\xi}}^{-1}(\boldsymbol{\pi})$ is **non-decreasing and concave**.

Key takeaway: no closed form in general, but under mean-reverting drift the DARE is explicit in the spectrum of $\boldsymbol{B}(\boldsymbol{\pi})$. We solve STS with Frank–Wolfe.

Non-linear Models

For $\boldsymbol{y}_k = h(\boldsymbol{X}_k, \boldsymbol{\beta}_k) + \boldsymbol{\epsilon}_k$ (e.g., a neural network with weights $\boldsymbol{\beta}_k$), the EKF linearizes

$$h(\boldsymbol{X}_k, \boldsymbol{\beta}_k) \approx h(\boldsymbol{X}_k, \hat{\boldsymbol{\beta}}_{k|k-1}) + \boldsymbol{H}_k\big(\boldsymbol{\beta}_k - \hat{\boldsymbol{\beta}}_{k|k-1}\big), \qquad \boldsymbol{H}_k = \partial h(\boldsymbol{X}_k, \boldsymbol{\beta})/\partial \boldsymbol{\beta}\,|_{\hat{\boldsymbol{\beta}}_{k|k-1}},$$

and the Jacobian $\boldsymbol{H}_k$ replaces $\boldsymbol{X}_k$ in the filter and in both designs.

Theorem 8: EKF estimation under known β_*

$$\boldsymbol{K}_k = \boldsymbol{\Sigma}_{\boldsymbol{\xi}_{k|k-1}} \boldsymbol{H}_k^\top \big(\boldsymbol{H}_k \boldsymbol{\Sigma}_{\boldsymbol{\xi}_{k|k-1}} \boldsymbol{H}_k^\top + \sigma^2 \mathbf{I}\big)^{-1}, \qquad \begin{aligned} \hat{\boldsymbol{\beta}}_{k|k} &= \hat{\boldsymbol{\beta}}_{k|k-1} + \boldsymbol{K}_k\big(\boldsymbol{y}_k - h(\boldsymbol{X}_k, \hat{\boldsymbol{\beta}}_{k|k-1})\big), \\ \boldsymbol{\Sigma}_{\boldsymbol{\xi}_{k|k}} &= (\mathbf{I} - \boldsymbol{K}_k \boldsymbol{H}_k)\, \boldsymbol{\Sigma}_{\boldsymbol{\xi}_{k|k-1}}. \end{aligned}$$

Corollary 1: Non-linear steady state is convex

With Jacobian vectors $\boldsymbol{g}_i = \partial h(\boldsymbol{x}_i, \boldsymbol{\beta})/\partial \boldsymbol{\beta}$ fixed at $\hat{\boldsymbol{\beta}}_*$, Lemma 1 holds for $\boldsymbol{B}(\boldsymbol{\pi}; \hat{\boldsymbol{\beta}}_*) = \sum_i \pi_i \boldsymbol{g}_i \boldsymbol{g}_i^\top$: G is concave.

Non-linear Models

For $\boldsymbol{y}_k = h(\boldsymbol{X}_k, \boldsymbol{\beta}_k) + \boldsymbol{\epsilon}_k$ (e.g., a neural network with weights $\boldsymbol{\beta}_k$), the EKF linearizes

$$h(\boldsymbol{X}_k, \boldsymbol{\beta}_k) \approx h(\boldsymbol{X}_k, \hat{\boldsymbol{\beta}}_{k|k-1}) + \boldsymbol{H}_k\big(\boldsymbol{\beta}_k - \hat{\boldsymbol{\beta}}_{k|k-1}\big), \qquad \boldsymbol{H}_k = \partial h(\boldsymbol{X}_k, \boldsymbol{\beta})/\partial \boldsymbol{\beta}\,|_{\hat{\boldsymbol{\beta}}_{k|k-1}},$$

and the Jacobian $\boldsymbol{H}_k$ replaces $\boldsymbol{X}_k$ in the filter and in both designs.

Theorem 8: EKF estimation under known β_*

$$\boldsymbol{K}_k = \boldsymbol{\Sigma}_{\boldsymbol{\xi}_{k|k-1}} \boldsymbol{H}_k^\top \big(\boldsymbol{H}_k \boldsymbol{\Sigma}_{\boldsymbol{\xi}_{k|k-1}} \boldsymbol{H}_k^\top + \sigma^2 \mathbf{I}\big)^{-1}, \qquad \begin{aligned} \hat{\boldsymbol{\beta}}_{k|k} &= \hat{\boldsymbol{\beta}}_{k|k-1} + \boldsymbol{K}_k\big(\boldsymbol{y}_k - h(\boldsymbol{X}_k, \hat{\boldsymbol{\beta}}_{k|k-1})\big), \\ \boldsymbol{\Sigma}_{\boldsymbol{\xi}_{k|k}} &= (\mathbf{I} - \boldsymbol{K}_k \boldsymbol{H}_k)\, \boldsymbol{\Sigma}_{\boldsymbol{\xi}_{k|k-1}}. \end{aligned}$$

Corollary 1: Non-linear steady state is convex

With Jacobian vectors $\boldsymbol{g}_i = \partial h(\boldsymbol{x}_i, \boldsymbol{\beta})/\partial \boldsymbol{\beta}$ fixed at $\hat{\boldsymbol{\beta}}_*$, Lemma 1 holds for $\boldsymbol{B}(\boldsymbol{\pi}; \hat{\boldsymbol{\beta}}_*) = \sum_i \pi_i \boldsymbol{g}_i \boldsymbol{g}_i^\top$: G is concave.

Non-linear Models

For $\boldsymbol{y}_k = h(\boldsymbol{X}_k, \boldsymbol{\beta}_k) + \boldsymbol{\epsilon}_k$ (e.g., a neural network with weights $\boldsymbol{\beta}_k$), the EKF linearizes

$$h(\boldsymbol{X}_k, \boldsymbol{\beta}_k) \approx h(\boldsymbol{X}_k, \hat{\boldsymbol{\beta}}_{k|k-1}) + \boldsymbol{H}_k\big(\boldsymbol{\beta}_k - \hat{\boldsymbol{\beta}}_{k|k-1}\big), \qquad \boldsymbol{H}_k = \partial h(\boldsymbol{X}_k, \boldsymbol{\beta})/\partial \boldsymbol{\beta}\,|_{\hat{\boldsymbol{\beta}}_{k|k-1}},$$

and the Jacobian $\boldsymbol{H}_k$ replaces $\boldsymbol{X}_k$ in the filter and in both designs.

Theorem 8: EKF estimation under known β_*

$$\boldsymbol{K}_k = \boldsymbol{\Sigma}_{\boldsymbol{\xi}_{k|k-1}} \boldsymbol{H}_k^\top \left(\boldsymbol{H}_k \boldsymbol{\Sigma}_{\boldsymbol{\xi}_{k|k-1}} \boldsymbol{H}_k^\top + \sigma^2 \mathbf{I}\right)^{-1}, \qquad \begin{aligned} \hat{\boldsymbol{\beta}}_{k|k} &= \hat{\boldsymbol{\beta}}_{k|k-1} + \boldsymbol{K}_k\big(\boldsymbol{y}_k - h(\boldsymbol{X}_k, \hat{\boldsymbol{\beta}}_{k|k-1})\big), \\ \boldsymbol{\Sigma}_{\boldsymbol{\xi}_{k|k}} &= (\mathbf{I} - \boldsymbol{K}_k \boldsymbol{H}_k)\, \boldsymbol{\Sigma}_{\boldsymbol{\xi}_{k|k-1}}. \end{aligned}$$

Corollary 1: Non-linear steady state is convex

With Jacobian vectors $\boldsymbol{g}_i = \partial h(\boldsymbol{x}_i, \boldsymbol{\beta})/\partial \boldsymbol{\beta}$ fixed at $\hat{\boldsymbol{\beta}}_*$, Lemma 1 holds for $\boldsymbol{B}(\boldsymbol{\pi}; \hat{\boldsymbol{\beta}}_*) = \sum_i \pi_i \boldsymbol{g}_i \boldsymbol{g}_i^\top$: G is concave.

Experiments: Test RMSE

Data: synthetic and six real sets. **Models:** linear, ResNet, CNN, MLP.
Selection: STG, STS (ours) vs. greedy info-gain (GKS), Adam (ADM), random (RND).
Estimation: Kalman/EKF (AE, ours) vs. forgetting-factor RLS, static (SE), SGD.

		Beijing	S&P	Bike	Housing	AT&T DL	AT&T UL	Synth.
Linear	Best baseline	140.53	1.79	260.73	3.70	139.14	12.17	2.74
	STG-AE (ours)	138.61	**1.68**	252.97	**3.66**	138.97	11.10	2.66
	STS-AE (ours)	**133.82**	1.75	**244.18**	3.68	**132.78**	**10.51**	**2.56**
ResNet	Best baseline	93.42	0.71	175.46	1.23	127.44	6.86	1.11
	STG-AE (ours)	92.22	0.67	**174.81**	1.24	123.32	6.74	0.98
	STS-AE (ours)	**87.28**	**0.60**	177.88	**1.19**	**122.43**	**6.60**	**0.89**
CNN	Best baseline	94.98	0.71	178.41	1.24	125.86	6.42	1.37
	STG-AE (ours)	93.23	0.66	177.93	**1.20**	123.47	5.86	1.39
	STS-AE (ours)	**88.23**	**0.62**	**173.93**	1.24	**121.39**	**5.78**	**1.35**
MLP	Best baseline	92.98	0.74	178.12	1.33	126.65	7.95	3.59
	STG-AE (ours)	91.65	**0.69**	177.97	**1.28**	124.29	8.21	2.98
	STS-AE (ours)	**89.70**	0.72	**174.99**	1.31	**121.45**	**7.90**	**2.63**

Test RMSE over 100 seeds (lower is better). Best baseline: best of the other 15 pairs; bold: best.

Experiments: Test RMSE

Data: synthetic and six real sets. **Models:** linear, ResNet, CNN, MLP.
Selection: STG, STS (ours) vs. greedy info-gain (GKS), Adam (ADM), random (RND).
Estimation: Kalman/EKF (AE, ours) vs. forgetting-factor RLS, static (SE), SGD.

		Beijing	S&P	Bike	Housing	AT&T DL	AT&T UL	Synth.
Linear	Best baseline	140.53	1.79	260.73	3.70	139.14	12.17	2.74
	STG-AE (ours)	138.61	**1.68**	252.97	**3.66**	138.97	11.10	2.66
	STS-AE (ours)	**133.82**	1.75	**244.18**	3.68	**132.78**	**10.51**	**2.56**
ResNet	Best baseline	93.42	0.71	175.46	1.23	127.44	6.86	1.11
	STG-AE (ours)	92.22	0.67	**174.81**	1.24	123.32	6.74	0.98
	STS-AE (ours)	**87.28**	**0.60**	177.88	**1.19**	**122.43**	**6.60**	**0.89**
CNN	Best baseline	94.98	0.71	178.41	1.24	125.86	6.42	1.37
	STG-AE (ours)	93.23	0.66	177.93	**1.20**	123.47	5.86	1.39
	STS-AE (ours)	**88.23**	**0.62**	**173.93**	1.24	**121.39**	**5.78**	**1.35**
MLP	Best baseline	92.98	0.74	178.12	1.33	126.65	7.95	3.59
	STG-AE (ours)	91.65	**0.69**	177.97	**1.28**	124.29	8.21	2.98
	STS-AE (ours)	**89.70**	0.72	**174.99**	1.31	**121.45**	**7.90**	**2.63**

Test RMSE over 100 seeds (lower is better). Best baseline: best of the other 15 pairs; bold: best.

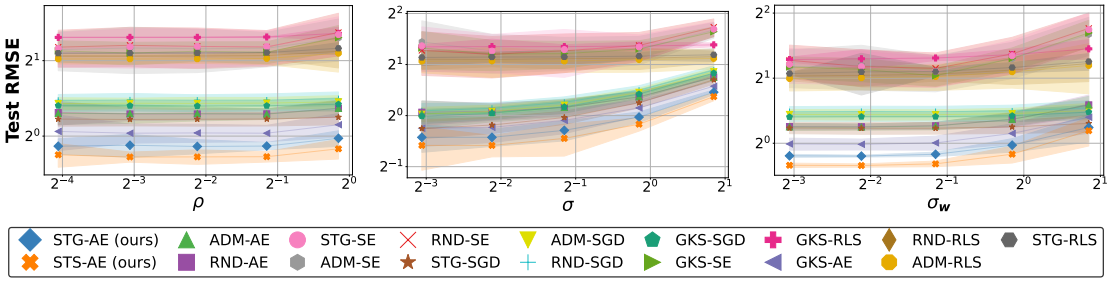
Experiments: Test RMSE

Data: synthetic and six real sets. **Models:** linear, ResNet, CNN, MLP.
Selection: STG, STS (ours) vs. greedy info-gain (GKS), Adam (ADM), random (RND).
Estimation: Kalman/EKF (AE, ours) vs. forgetting-factor RLS, static (SE), SGD.

		Beijing	S&P	Bike	Housing	AT&T DL	AT&T UL	Synth.
Linear	Best baseline	140.53	1.79	260.73	3.70	139.14	12.17	2.74
	STG-AE (ours)	138.61	**1.68**	252.97	**3.66**	138.97	11.10	2.66
	STS-AE (ours)	**133.82**	1.75	**244.18**	3.68	**132.78**	**10.51**	**2.56**
ResNet	Best baseline	93.42	0.71	175.46	1.23	127.44	6.86	1.11
	STG-AE (ours)	92.22	0.67	**174.81**	1.24	123.32	6.74	0.98
	STS-AE (ours)	**87.28**	**0.60**	177.88	**1.19**	**122.43**	**6.60**	**0.89**
CNN	Best baseline	94.98	0.71	178.41	1.24	125.86	6.42	1.37
	STG-AE (ours)	93.23	0.66	177.93	**1.20**	123.47	5.86	1.39
	STS-AE (ours)	**88.23**	**0.62**	**173.93**	1.24	**121.39**	**5.78**	**1.35**
MLP	Best baseline	92.98	0.74	178.12	1.33	126.65	7.95	3.59
	STG-AE (ours)	91.65	**0.69**	177.97	**1.28**	124.29	8.21	2.98
	STS-AE (ours)	**89.70**	0.72	**174.99**	1.31	**121.45**	**7.90**	**2.63**

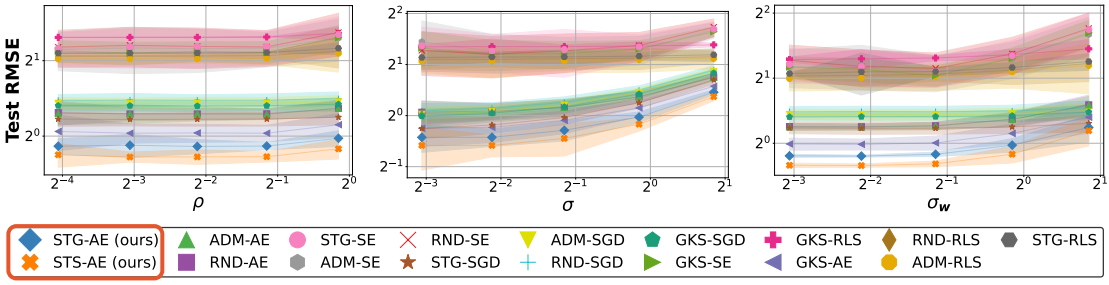
Test RMSE over 100 seeds (lower is better). Best baseline: best of the other 15 pairs; bold: best.

Experiments: Parameter Robustness



ResNet on synthetic data, varying ρ, σ or σ_w with the others at their defaults: our methods stay **lowest**, and the advantage of Kalman tracking grows as the drift increases.

Experiments: Parameter Robustness



ResNet on synthetic data, varying ρ, σ or σ_w with the others at their defaults: our methods stay **lowest**, and the advantage of Kalman tracking grows as the drift increases.

Conclusion

- **Markovian experimental design**: choosing which samples to collect while the model drifts.
- A Kalman filter (an EKF for neural networks) tracks the model, and both the myopic and the steady-state D-optimal designs are **convex**.
- The lowest test RMSE on every dataset and model.

Paper

Thank you!