Wasserstein Distributionally Robust Kalman Filtering

Ori Meiraz

Background

Let’s say we have a signal $x \in R^n$ which we do not know – called the state.

We have an observable signal $y \in R^m$ - called the output.

We aim to estimate the current state x based on y

Goal

We want to choose an estimator – given $y \in R^m$, predict $x \in R^n$.

In a simpler way:

$$\inf_{\psi \in \mathcal{L}} \sup_{\mathbb{Q} \in \mathcal{P}} \mathbb{E}^{\mathbb{Q}}[\|x - \psi(y)\|^2]$$

Where $\mathcal{L}$ denotes the family of all measurable functions from $\mathbb{R}^m$ to $\mathbb{R}^n$.

Overview of the paper

Reminders:

Type 2 Wasserstein distance:

$$W_2(\mathbb{Q}_1, \mathbb{Q}_2) \triangleq \inf_{\pi \in \Pi(\mathbb{Q}_1, \mathbb{Q}_2)} \left\{ \left(\int_{R^d \times R^d} \|z_1 - z_2\|^2 \pi(dz_1, dz_2) \right)^{\frac{1}{2}} \right\}$$

Where $\Pi(\mathbb{Q}_1, \mathbb{Q}_2)$ is the set of all probability distributions on $\mathbb{R}^d \times \mathbb{R}^d$ with marginals $\mathbb{Q}_1$ and $\mathbb{Q}_2$.

Theorem:

$$\mathbb{Q}_1 = \mathcal{N}_d(\mu_1, \Sigma_1), \mathbb{Q}_2 = \mathcal{N}_d(\mu_2, \Sigma_2), \qquad \Sigma_1, \Sigma_2 \in S_+^d \implies$$

$$W_2(\mathbb{Q}_1, \mathbb{Q}_2) = \sqrt{\|\mu_1 - \mu_2\| + Tr\left[\Sigma_1 + \Sigma_2 - 2\left(\Sigma_2^{0.5} \Sigma_1 \Sigma_2^{0.5}\right)^{0.5}\right]}$$

Reminders

Wasserstein ambiguity set:

$$\mathcal{P} = \{\mathbb{Q} \in \mathcal{N}_d : W_2(\mathbb{Q}, \mathbb{P}) \leq \rho\}$$

Theorem:

$$\inf_{\psi \in \mathcal{L}} \sup_{\mathbb{Q} \in \mathcal{P}} \mathbb{E}^{\mathbb{Q}}[\|x - \psi(x)\|^2] = \sup_{\mathbb{Q} \in \mathcal{P}} \inf_{\psi \in \mathcal{L}} \mathbb{E}^{\mathbb{Q}}[\|x - \psi(x)\|^2]$$

Reformulation

The minmax problem with the Wasserstein ambiguity set centered at $\mathbb{P} = \mathcal{N}_d(\mu, \Sigma)$, $\underline{\sigma} \stackrel{\Delta}{=} \lambda_{\min}(\Sigma) > 0$.

$$
\begin{aligned}
&\sup \ \operatorname{Tr}\left[S_{xx} - S_{xy} S_{yy}^{-1} S_{yx}\right] \qquad [5] \\
&s.t \\
&S = \begin{bmatrix} S_{xx} & S_{xy} \\ S_{yx} & S_{yy} \end{bmatrix} \in \mathbb{S}_{+}^{d}, S_{xx} \in \mathbb{S}_{+}^{n}, S_{xy} = S_{yx}^{T} \in \mathbb{R}^{n \times m} \\
&\operatorname{Tr}[S + \Sigma - 2(\Sigma^{0.5} S \Sigma^{0.5})^{0.5}] \leq \rho^2 \quad , S \succcurlyeq \underline{\sigma} I_d
\end{aligned}
$$

How does it help me?

Using the solution:

If S_{xx}^*, S_{yy}^* and S_{xy}^* is optimal in the problem above, then the affine function $\psi^*(y) = S_{xy}^*\left(S_{yy}^*\right)^{-1}\left(y - \mu_y\right) + \mu_x$ is the distributionally robust minimum mean square estimator and the normal distribution $\mathbb{Q}^* = \mathcal{N}_d(\mu, S^*)$ is the least favorable prior.

$$\mu = \begin{pmatrix} \mu_x \\ \mu_y \end{pmatrix} \quad S^* = \begin{bmatrix} S_{xx}^* & S_{xy}^* \\ {S_{xy}^*}^T & S_{yy}^* \end{bmatrix}$$

So now we know what to do

But there is a long way

Definitions

Denote

$$f(S) \triangleq \mathrm{Tr}\left[S_{xx} - S_{xy} S_{yy}^{-1} S_{yx}\right]$$

Definitions

A function $\varphi: S_+^d \to R_+$ has a unit total elasticity if

$$\varphi(S) = \langle S, \nabla\varphi(S)\rangle \quad \forall S \in S_+^d$$

Turns out that $f(S)$ has a unit total elasticity.

Uses

We can use this conclusion to replace the function with a linear approximation $\Longrightarrow$ the problem can be solved highly efficiently.

Using a Frank-Wolfe algorithm:

$$S^{k+1} = \alpha_k F\left(S^k\right) + (1 - \alpha_k) S^k \quad (S^0 = \Sigma)$$

Where

$$F(S) = \arg\max_{L \succcurlyeq \sigma I_d} \langle L, \nabla f(S) \rangle$$

$$s.t \qquad \mathrm{Tr}[L + \Sigma - 2(\Sigma^{0.5} L \Sigma^{0.5})^{0.5}] \leq \rho^2 \qquad [7b]$$

In simple words

In each iteration, the Frank-Wolfe algorithm thus maximizes a linearized objective function over the original feasible set.

In contrast to other commonly used first-order methods, the Frank-Wolfe

algorithm thus obviates the need for a potentially expensive projection step to recover feasibility.

In simpler words

To make the frank – wolfe algorithm work in practice, one needs:

i. an efficient routine for solving the direction-finding subproblem (7b)

ii. a step-size rule that offers rigorous guarantees on the algorithm's convergence rat

Algorithm 1 – Bisection algorithm to solve 7b

Input: Covariance matrix $\Sigma \succ 0$
Gradient matrix $D \triangleq \nabla f(S) \succeq 0$
Wasserstein radius $\rho > 0$
Tolerance $\varepsilon > 0$

Denote the largest eigenvalue of D by λ_1
Let v_1 be an eigenvector of λ_1
Set $LB \leftarrow \lambda_1(1 + \sqrt{v_1^\top \Sigma v_1}/\rho)$
Set $UB \leftarrow \lambda_1(1 + \sqrt{\mathrm{Tr}[\Sigma]}/\rho)$
repeat
Set $\gamma \leftarrow (UB + LB)/2$
Set $L \leftarrow \gamma^2(\gamma I_d - D)^{-1}\Sigma(\gamma I_d - D)^{-1}$
if $h(\gamma) < 0$ **then**
Set $LB \leftarrow \gamma$
else
Set $UB \leftarrow \gamma$
end if
Set $\Delta \leftarrow \gamma(\rho^2 - \mathrm{Tr}[\Sigma]) - \langle L, D\rangle$
$+\gamma^2\langle(\gamma I_d - D)^{-1}, \Sigma\rangle$
until $h(\gamma) > 0$ and $\Delta < \varepsilon$
Output: L

$$h(\gamma) \triangleq \rho^2 - \langle \Sigma, (I_d - \gamma(\gamma I_d - \nabla f(S))^{-1})^2 \rangle.$$

Theorem:
For any fixed inputs $\rho, \epsilon \in R_{++}$, $\Sigma \in \mathbb{S}^d_{++}$ and $S \in \mathbb{S}^d_{+}$, algorithm 1 outputs a feasible and ϵ-suboptimal solution to (7b)

Algorithm 2 – Frank-Wolfe algorithm to solve 5

Input: Covariance matrix $\Sigma \succ 0$
Wasserstein radius $\rho > 0$
Tolerance $\delta > 0$

Set $\underline{\sigma} \leftarrow \lambda_{\min}(\Sigma), \bar{\sigma} \leftarrow (\rho + \sqrt{\mathrm{Tr}[\Sigma]})^2$
Set $\overline{C} \leftarrow 2\bar{\sigma}^4/\underline{\sigma}^3$
Set $S^{(0)} \leftarrow \Sigma, k \leftarrow 0$
while Stopping criterion is not met **do**
Set $\alpha_k \leftarrow \frac{2}{k+2}$
Set $G \leftarrow S_{xy}^{(k)}(S_{yy}^{(k)})^{-1}$
Compute gradient $D \leftarrow \nabla f(S^{(k)})$ by
$$D \leftarrow [I_n, \ -G]^\top [I_n, \ -G]$$
Set $\varepsilon \leftarrow \alpha_k \delta \overline{C}$
Solve the subproblem (7b) by Algorithm 1
$$L \leftarrow \text{Bisection}(\Sigma, D, \rho, \varepsilon)$$
Set $S^{(k+1)} \leftarrow S^{(k)} + \alpha_k(L - S^{(k)})$
Set $k \leftarrow k + 1$
end while

Output: $S^{(k)}$

What is a Kalman filter?

Kalman filter receives a series in time, not just one x and one y.

So, we need to generalize what we found for $x_t \in R^n$, $y_t \in R^m$.

At any time $t \in \mathbb{N}$, we aim to estimate the current.

state x_t based on the output history $Y_t \triangleq (y_1, \ldots, y_t)$.

What is a Kalman filter?

Denote z_t as $z_t = \binom{x_t}{y_t}$.

The nominal distribution $\mathbb{P}^*_{z_t|x_{t-1}}$ is known and is:

$$x_t = A_t x_{t-1} + B_t v_t$$
$$y_t = C_t x_t + D_t v_t$$

for known A_t, B_t, C_t, D_t.

$v_t \sim \mathcal{N}_d(0, I_d)$ is the noise and is independent of x_t.

What is a Kalman filter

So,

$$P^*_{z_t|x_{t-1}} = \mathcal{N}_d\left(\begin{bmatrix} A_t \\ C_tA_t \end{bmatrix} x_{t-1}, \begin{bmatrix} B_t \\ C_tB_t + D_t \end{bmatrix}\begin{bmatrix} B_t \\ C_tB_t + D_t \end{bmatrix}^T\right)$$
$$\forall t \in \mathbb{N}$$

Unlike P^*, the true distribution $\mathbb{Q}$ is unknown.

Example

We are driving a car from Netanya to Haifa.

We want to know what our location (x) in every minute $(t \in \mathbb{N})$ -

$$x_t \in R$$

We only know the speed at which we are going in every minute –

$$y_t \in R$$

Accept we don’t know the true speed; we know a noisy signal.

We want to know our location from our speed.

Algorithm

We assume that the marginal distribution $\mathbb{Q}^*_{x_0}$ equals $\mathbb{P}_{x_0}$ - that is $\mathbb{Q}^*_{x_0} = \mathcal{N}(\widehat{x_0}, V_0)$.

Next, fix any t $\in \mathbb{N}$ and assume that the conditional distribution $Q^*_{x_{t-1}|\, Y_t - 1}$ of x_{t-1} given Y_{t-1} under $\mathbb{Q}^*$ has already been computed as $Q^*_{x_{t-1}|Y_{t-1}} = \mathcal{N}_n(x_{t-1}, V_{t-1})$.

The construction of $Q^*(x_t|Y_t)$is then split into a prediction step and an update step.

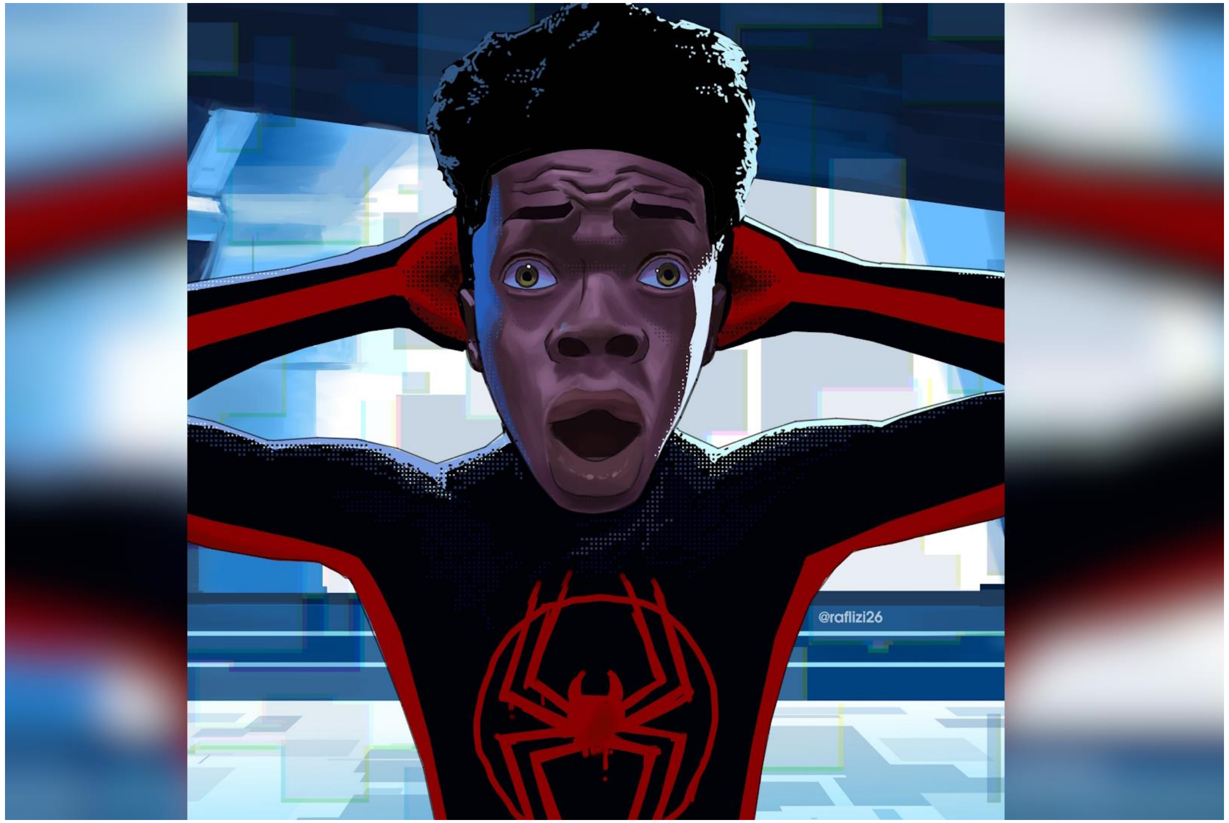
Prediction step

The prediction step combines the previous state estimate $Q^*_{x_{t-1}|\,Y_{t-1}}$ with the nominal transition kernel $P_{z_t|\,x_{t-1}}$ to generate a pseudo-nominal distribution $P_{z_t|Y_{t-1}}$ of z_t conditioned on Y_{t-1}, which is defined through

$$\mathbb{P}_{z_t|Y_{t-1}}(B|Y_{t-1}) = \int_{\mathbb{R}^n} P^*_{z_t|x_{t-1}}(B|x_{t-1})\mathbb{Q}^*_{x_{t-1}|Y_{t-1}}(dx_{t-1}|Y_{t-1})$$

For every borel set $B \subseteq \mathbb{R}^d$

memes



Prediction step

The well-known formula for the convolution of two multivariate Gaussians reveals $that\ \mathbb{P}_{z_t|Y_{t-1}} = \mathcal{N}_d(\mu_t, \Sigma_t)$ where

$$\mu_t = \begin{bmatrix} A_t \\ C_t A_t \end{bmatrix} \hat{x}_{t-1}$$

$$\Sigma_t = \begin{bmatrix} A_t \\ C_t A_t \end{bmatrix} V_{t-1} \begin{bmatrix} A_t \\ C_t A_t \end{bmatrix}^T + \begin{bmatrix} B_t \\ C_t B_t + D_t \end{bmatrix} \begin{bmatrix} B_t \\ C_t B_t + D_t \end{bmatrix}^T$$

Update step

In the update step, the pseudo-nominal a priori estimate $\mathbb{P}_{z_t|Y_{t-1}}$ is updated by the measurement y_t and robustified against model uncertainty to yield a refined a posteriori estimate $\mathbb{Q}^*_{x_t|Y_t}$. This a posteriori estimate is found by solving the minimax problem:

$$\inf_{\psi \in \mathcal{L}} \sup_{\mathbb{Q} \in \mathcal{P}_{z_t|Y_{t-1}}} \mathbb{E}^{\mathbb{Q}}[\|x_t - \psi_t(y_t)\|^2]$$

$$\mathcal{P}_{z_t|Y_{t-1}} = \left\{\mathbb{Q} \in \mathcal{N}_d : W_2\left(\mathbb{Q}, \mathbb{P}_{z_t|\, Y_{t-1}}\right) \leq \rho_t\right\}$$

Update step

Finally,

We obtain that the least favorable distribution is:

$$\mathbb{Q}_{x_t|Y_t} = \mathcal{N}_n(\widehat{x_t}, V_t)$$

$$\widehat{x_t} = S^*_{t,xy}\left(S^*_{t,yy}\right)^{-1}\left(y_t - \mu_{t,y}\right) + \mu_{t,x}$$
$$V_t = S_{t,xx} - S^*_{t,xy}\left(S^*_{t,yy}\right)^{-1} S^*_{t,yx}$$

Sum it all up

Algorithm 3 Robust Kalman filter at time t

Input: Covariance matrix $V_{t-1} \succeq 0$
State estimate $\hat{x}_{t-1}$
Wasserstein radius $\rho_t > 0$
Tolerance $\delta > 0$

Prediction:
Form the pseudo-nominal distribution
$\mathbb{P}_{z_t|Y_{t-1}} = \mathcal{N}_d(\mu_t, \Sigma_t)$ using (10)

Observation:
Observe the output y_t

Update:
Use Algorithm 2 to solve (11)
$S_t^\star \leftarrow \text{Frank-Wolfe}(\Sigma_t, \mu_t, \rho_t, \delta)$

Output: $V_t = S_{t,xx} - S_{t,xy}(S_{t,yy})^{-1}S_{t,yx}$
$\hat{x}_t = S_{t,xy}^\star(S_{t,yy}^\star)^{-1}(y_t - \mu_{t,y}) + \mu_{t,x}$

Connected literature

Connected literature

The $\mathcal{H}_\infty$ filter also targets situations in which the statistics of the noise process is uncertain and where one aims to minimize the worst case.

However, in transient operation, the desired $\mathcal{H}_\infty$-performance is lost, and the filter may diverge unless some (typically restrictive) positivity condition holds in each iteration.

Connected literature

A risk-sensitive Kalman filter is obtained by minimizing the moment-generating function instead of the mean of the squared estimation error.

This risk-sensitive Kalman filter minimizes the worst-case mean square error across all joint state-output distributions in a Kullback-Leibler (KL) ball around a nominal distribution.

Paper contribution

Paper contribution

- The paper focuses on a (nonconvex) Wasserstein ambiguity set containing only normal distributions.
- Shows that the optimal estimator and the least favorable distribution form a Nash equilibrium.
- Proves that the estimation problem is equivalent to a tractable convex program. Devises a Frank-Wolfe algorithm for this convex program.

Assessment of strength and weakness

Strength

- The paper solves gives an efficient way to solve the optimization problem.
- Given the solution, S^*, we know exactly the worst-case distribution.

Weakness

- Only considers type-2 Wasserstein distance.
- Only considers normal distributions.