Supplemental: Nonparametric Object and Parts Modeling with Lie Group Dynamics

1 Inference

1.1 Linear Gaussian Conditionals

Consider the multivariate Gaussian

$$\mathrm{N}\left(\begin{pmatrix} Cx_1 + u \\ x_2 \end{pmatrix} \mid \begin{pmatrix} \mu_1 \\ \mu_2 \end{pmatrix}, \begin{pmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{12}^\top & \Sigma_{22} \end{pmatrix}\right) \tag{1}$$

where $x_1, u, \mu_1 \in \mathbb{R}^{D_1}$, $x_2, \mu_2 \in \mathbb{R}^{D_2}$ and covariance $\Sigma \in \mathbb{R}^{D_1+D_2}$ has blocks $\Sigma_{11} \in \mathbb{R}^{D_1 \times D_1}, \Sigma_{12} \in \mathbb{R}^{D_1 \times D_2}, \Sigma_{21} \in \mathbb{R}^{D_2 \times D_1}, \Sigma_{22} \in \mathbb{R}^{D_2 \times D_2}$. Then, because Gaussian conditionals are Gaussian (see **[?]**, Ch. 4), it follows that the conditional $Cx_1 + u \mid x_2$ is Gaussian:

$$Cx_1 + u \mid x_2 \sim \mathrm{N}\left(Cx_1 + u \mid \mu', \Sigma'\right) \tag{2}$$

$$\mu' = \mu_1 + \Sigma_{12}\Sigma_{22}^{-1}\left(x_2 - \mu_2\right) \tag{3}$$

$$\Sigma' = \Sigma_{11} - \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21} \tag{4}$$

And, by transformation of random variables, the conditional $x_1 \mid x_2$ is Gaussian with parameters:

$$x_1 \mid x_2 \sim \mathrm{N}\left(x_1 \mid \mu'', \Sigma''\right) \tag{5}$$

$$\mu'' = C^{-1}\left(\mu' - u\right) \tag{6}$$

$$\Sigma'' = C^{-1}\Sigma' C^{-\top} \tag{7}$$

1.2 Concentrated Gaussian Priors with Gaussian Likelihoods

In this section, we show that Concentrated Gaussian priors on the Lie Group $\mathrm{SE}(D)$ coupled with multivariate Gaussian observation models have Gaussian conditionals for the translation component.

Let $a, b, c, \mu \in \mathrm{SE}(D)$ where each contain a rotation component R and translation component d.

$$a = \begin{pmatrix} R_a & d_a \\ 0 & 1 \end{pmatrix} \qquad b = \begin{pmatrix} R_b & d_b \\ 0 & 1 \end{pmatrix} \qquad c = \begin{pmatrix} R_c & d_c \\ 0 & 1 \end{pmatrix} \qquad \mu = \begin{pmatrix} R_\mu & d_\mu \\ 0 & 1 \end{pmatrix} \tag{8}$$

These can be viewed as linear operators on homogeneous coordinates. Let $y \in \mathbb{R}^D$ be a point and $E \in \mathbb{R}^{D \times D}$ be a covariance matrix. For vector v, let $\tilde{v}$ be the projection of v into homogeneous coordinates (append 1). For covariance Σ, let $\tilde{\Sigma}$ be the projection of Σ into homogeneous coordinates (append a 0 row and column).

Consider the following distribution, for Σ a covariance in the tangent plane about μ:

$$p(b \mid y, a, b) \propto \mathrm{N_L}\left(b \mid \mu, \Sigma\right) \mathrm{N}\left(\tilde{y} \mid abc\tilde{0}, (abc)\,\tilde{E}\,(abc)^\top\right) \tag{9}$$

$$= \mathrm{N}\left(\mathrm{Log}_\mu b \mid 0, \Sigma\right) \mathrm{N}\left(a^{-1}\tilde{y} \mid bc\tilde{0}, (bc)\,\tilde{E}\,(bc)^\top\right) \tag{10}$$

$$= \mathrm{N}\left(\log\left(\mu^{-1}b\right) \mid 0, \Sigma\right) \mathrm{N}\left(a^{-1}\tilde{y} \mid bc\tilde{0}, (bc)\,\tilde{E}\,(bc)^\top\right) \tag{11}$$

$$= \mathrm{N}\left(\begin{pmatrix} V^{-1}_{\mu^{-1}b} d_{\mu^{-1}b} \\ \phi_{\mu^{-1}b} \end{pmatrix} \mid 0, \Sigma\right) \mathrm{N}\left(R_a^\top (y - d_a) \mid d_b + R_b d_c\,(R_b R_c)\,E\,(R_b R_c)^\top\right) \tag{12}$$

$$= \mathrm{N}\left(\begin{pmatrix} V^{-1}_{\mu^{-1}b} \left(R_\mu^\top (d_b - d_\mu)\right) \\ \phi_{\mu^{-1}b} \end{pmatrix} \mid 0, \Sigma\right) \mathrm{N}\left(R_a^\top (y - d_a) \mid d_b + R_b d_c\,(R_b R_c)\,E\,(R_b R_c)^\top\right) \tag{13}$$

Homogeneous coordinates are used up to Eqn. (11), then dropped in Eqn. (12). Observe that Eqn. (12) is of the form:

$$\mathrm{N}\left(\begin{pmatrix} Cd_b + u \\ \phi \end{pmatrix} \mid \begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{pmatrix}\right) \mathrm{N}\left(z \mid d_b + g, \Lambda\right) \tag{14}$$

where

$$C = V^{-1}_{\mu^{-1}b} R_\mu^\top \qquad u = -Cd_\mu \qquad z = R_a^\top (y - d_a) \tag{15}$$

$$g = R_b d_c \qquad \Lambda = (R_b R_c)\,E\,(R_b R_c)^\top \qquad \phi = \phi_{\mu^{-1}b} \tag{16}$$

The conditional $p(d_b \mid R_b, y, a, b)$ is proportional to Eqn. (14), which is of the form of Eqn. (1), and $C, u, z, g, \Lambda, \phi$ are all computable given R_b, y, a, b (and C is invertible), hence $p(d_b \mid R_b, a, b) = \mathrm{N}\left(d_b \mid \mu', \Sigma'\right)$ for some μ', Σ'. Then,

$$p(d_b \mid R_b, y, a, b) \propto \mathrm{N}\left(d_b \mid \mu', \Sigma'\right) \mathrm{N}\left(z \mid d_b + g, \Lambda\right) \tag{17}$$

$$\propto \mathrm{N}\left(d_b \mid \mu'', \Sigma''\right) \tag{18}$$

where Eqn. (18) follows from Eqn. (17) because it is a linear Gaussian system, hence is itself proportional to a Gaussian with some mean and covariance μ'', Σ'' (see [**?**], Ch. 4).

1.3 Translation Full Conditionals

In the following, let $x_{t-1}, x_t, x_{t+1}, \{\omega_k, \theta_{(t-1)k}, \theta_{tk}, \theta_{(t+1)k}\}_{k=1}^K, I \in \mathrm{SE}(D)$ with rotation and translation components defined similarly to Eqn. (8). Let $\{y_{tn}\}_{n=1}^{N_t} \in \mathbb{R}^D$. Let $\{E_k\}_{k=1}^K \in \mathbb{R}^{D\times D}$ be observation covariances in $\mathbb{R}^D$ and $Q, W, \{S_k\}_{k=1}^K$ be covariances in the Lie algebra $\mathfrak{se}(D)$. Let $\{z_{tn}\}_{n=1}^{N_t}$ be assignments of observations to one of K instantiated components.

The full conditional body frame translation update is of the form:

$$p(d_{x_t} \mid R_{x_t}, x_{t-1}, x_{t+1}, Q, \{\omega_k, \theta_{tk}\}_{k=1}^K, \{y_{tn}, z_{tn}\}_{n=1}^{N_t} \tag{19}$$

$$\propto \mathrm{N_L}\left(x_t \mid x_{t-1}, Q\right) \mathrm{N_L}\left(x_{t+1} \mid x_t, Q\right) \prod_{n=1}^{N_t} \mathrm{N}\left(\tilde{y}_{tn} \mid x_t\omega_k\theta_{tk}\tilde{0}, (x_t\omega_k\theta_{tk})\,\tilde{E}_k\,(x_t\omega_k\theta_{tk})^\top\right)^{\mathbb{I}(z_{tn}=k)} \tag{20}$$

The full conditional for the k^{th} canonical part translation update is of the form:

$$p(d_{\omega_k} \mid R_{\omega_k}, W_k, \{x_t, \theta_{tk}, \{y_{tn}\}_{n=1}^{N_t}\}_{t=1}^T, E_k) \tag{21}$$

$$\propto \mathrm{N_L}\left(\omega_k \mid I, W\right) \prod_{t=1}^{T} \prod_{n=1}^{N_t} \mathrm{N}\left(\tilde{y}_{tn} \mid x_t\omega_k\theta_{tk}\tilde{0}, (x_t\omega_k\theta_{tk})\,\tilde{E}_k\,(x_t\omega_k\theta_{tk})^\top\right)^{\mathbb{I}(z_{tn}=k)} \tag{22}$$

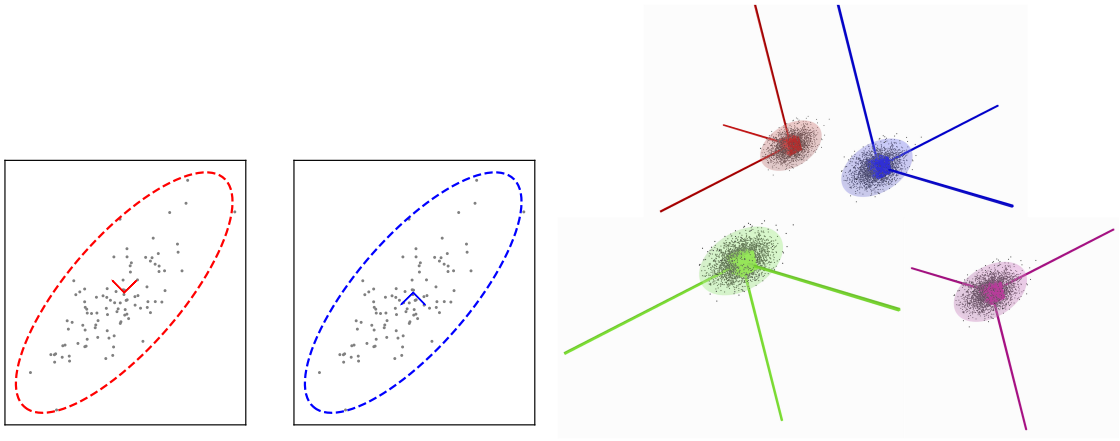

Figure 1: Likelihoods are invariant to two rotation symmetries in SE(2) (left) and four rotation symmetries in SE(3) (right). Notice that the colored observation covariances cover the same volume when drawn at fixed standard deviations, implying that mahalanobis distances of part covariances to observations will be equal for each symmetric rotation. While dynamics will typically favor one mode over others, the slice sampler sometimes locks onto the wrong mode. The remedy is a fixed number of MCMC proposals which, given a mode, can enumerate and propose all other modes

In both of the above cases, the concentrated Gaussians have Gaussian conditionals for the translation component, and combine with a product of Gaussian likelihoods, yielding a Gaussian posterior for translations d_{x_t}, d_{ω_k} (per Sections 1.1, 1.2).

Suppose θ_{tk} has dynamics:

$$\theta_{tk} = \begin{pmatrix} R_{\theta_{tk}} & d_{\theta_{tk}} \\ 0 & 1 \end{pmatrix} = \begin{pmatrix} \text{Exp}_{R_{\theta_{(t-1)k}}} \phi_{tk} & A\, d_{\theta_(t-1)k} + B\, m_{tk} \\ 0 & 1 \end{pmatrix} \tag{23}$$

where

$$(m_{tk}, \phi_{tk}) \sim \text{N}(0, S_k) \tag{24}$$

Then $p(d_{\theta_{tk}} \mid R_{\theta_{tk}}, \theta_{(t-1)k})$ is of the form Eqn. (1) with $C = B, u = A\, d_{\theta_{(t-1)k}}$. The conditional $p(d_{\theta_{tk}} \mid R_{\theta_{tk}}, \theta_{(t+1)k})$ has a similar Gaussian form. Hence, full conditional translation updates for $d_{\theta_t k}$ have a similar structure to the above object and canonical part translation updates and are themselves Gaussian.

1.4 Rotation Full Conditionals

We perform univariate slice sampling [**?**] to sample from the rotation full conditionals of body x_t, canonical part ω_k and part transformation θ_{tk}. This is straightforward because the Lie algebraic coordinates of each transformation decompose into a set of univariate coordinates corresponding to translation, and a set corresponding to rotation. Given this decomposition sampling is straightforward: each rotation coordinate is sampled, holding all others fixed. The task is furthered simplified because the bounds of $\pm\pi$ can be imposed.

One complication is that the distribution is multi-modal because the observation likelihood is invariant to 180° rotations. There are two such modes in SE(2) and four in SE(3). Given one mode, all others can be enumerated by inverting any subset of the columns of the sampled rotation matrix such that the determinant remains $+1$ (as opposed to -1 for an inversion of an odd number of columns). Figure 1 visually demonstrates these symmetries for SE(2) and SE(3) Although the dynamics will typically penalize one mode over others, it sometimes happens that the slice sampler locks onto a particular mode. The

Figure 2: Example ground-truth part segmentations.

solution is simple: we propose a fixed number of MCMC samples, one for each enumerated mode. This is of minimal cost because there is only one other mode in $\mathrm{SE}(2)$ and three other modes in $\mathrm{SE}(3)$.

When sampling rotation full conditionals, we use characteristic width $w = 0.01\pi$ and a maximum of 10 doubling iterations. Ten samples are drawn, then the MCMC proposals for rotation symmetries are proposed starting from the final sample.

1.5 Conjugate Posteriors

We show that driving noise covariance Q for body frame of reference x_t is a product of an Inverse Wishart prior with a product of multivariate Gaussian likelihoods, yielding analytic sampling updates by conjugacy. The same reasoning holds for part transformation driving noise covariances $\{S_k\}_{k=1}^K$.

The posterior distributions for Q is:

$$p(Q \mid x_{1:T}) \propto \mathrm{IW}(Q \mid \cdot) \prod_{t=1}^{T} \mathrm{N}_{\mathrm{L}}\left(x_t \mid x_{t-1}, Q\right) \tag{25}$$

$$= \mathrm{IW}(Q \mid \cdot) \prod_{t=1}^{T} \mathrm{N}\left(\begin{pmatrix} V^{-1}_{x_{t-1}^{-1}x_t} d_{x_{t-1}^{-1}x_t} \\ \phi_{x_{t-1}^{-1}x_t} \end{pmatrix} \mid 0, Q\right) \tag{26}$$

The terms inside the product are all computable given $x_{1:T}$, so this is an Inverse-Wishart multiplied by a product of Gaussians. In this case, the posterior is conjugate to the prior, yielding Inverse Wishart updates (see [**?**], Appendix A). The same form and reasoning applies for S_k, hence samples can also be analytically drawn for each S_k.

Part observation covariances E_k have the form:

$$p(E_k \mid \omega_k, \{x_t, \theta_{tk}, \{y_{tn}, z_{tn}\}_{n=1}^{N_t}\}_{t=1}^{T}) \propto \mathrm{IW}(E_k \mid \cdot) \prod_{t=1}^{T} \mathrm{N}\left((x_t \omega_k \theta_{tk})^{-1} \tilde{y}_{tn} \mid \tilde{0}, \tilde{E}_k\right)^{\mathbb{I}(z_{tn}=k)} \tag{27}$$

As above, this posterior is also Inverse Wishart.

1.6 Ground-Truth Segmentations

Figure 2 shows example ground-truth segmentations for each dataset used for quantitative comparison. Ground-truth was hand-labeled, and the number of parts were chosen at the granularity supported by the dataset (e.g. `marmoset` has head, body and tail but not hands or feet because they were not visible from the top-down RGB-D views).

1.7 Data-Dependent Priors

Results in the paper were computed by using data-dependent priors that are similar in spirit to those used for static Dirichlet Process Mixture Models. All Inverse Wishart priors (for Q, $\{S_k, E_k\}_{k=1}^{\infty}$) were set to ten

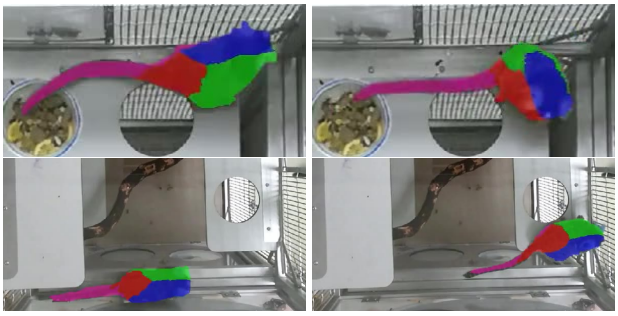

Figure 3: Example of how our model can infer results on multiple videos of the same type, but different instances, of an object in motion. These results were computed on RGB-D data but are visualized in 2D.

degrees of freedom, making the prior weak in the sense that it accounts for 10 pseudo-observations (among tens to thousands of observations incorporated into the posterior).

The Inverse Wishart scatter matrix prior for Q was set so that the expected per-timestep body rotation was 0.25 radians ($\approx 15°$) and expected per-timestep body translation was the mean absolute difference between time-adjacent pairs of observation sets.

The Inverse Wishart scatter matrix prior for S_k was set so that the expected per-timestep part rotation was 0.025 radians ($\approx 1.5°$) and expected per-timestep part translation was the mean absolute difference between time-adjacent pairs of observation sets (expected translation for parts and body are the same under the prior).

The Inverse Wishart scatter matrix prior for part observation covariances E_k was set to 0.1 times the mean observation set variance.

The prior for the initial body transformation was set to identity mean rotation with mean translation equal to the mean of the first observation set. The initial body transformation covariance was set diagonal and broad, so that π radians were within one standard deviation of rotation covariance, and body translation variances were set equal to the variance of the first observation set.

Canonical part transformations ω_k were set to identity mean transformation with π radians being within one standard deviation of rotation covariance, and canonical part translation variances set equal to the variance of the first observation set.

2 Generalization Across Videos

We demonstrate that our model can reason about the motion and parts of different instances of the same type of object across multiple videos. This is accomplished by assuming that the number of parts and the canonical part transformations, ω_k, are shared by similar objects, but that the motion parameters are distinct. In this experiment, we sample all parameters (including number of parts) from a video containing one instance of an object. In the second video, we restrict sampling to associations z_{tn}, body transformations x_t and part transformations θ_{tk}. Figure 3 shows RGB-D data projected into 2D for two videos; all model parameters are initially sampled in the video of the top row, then body and part transformations are sampled in the second video.

We note that part assignments correspond reasonably across videos. By reasoning in 3D, our model accommodates scale changes within and across videos, such as when the object is closer or further from the camera. While we do see some migration of part locations on the torso, this is due to the proximity of the respective ω_k's combined with sufficiently free motion dynamics. Regardless, torso parts remain associated to the torso, and the tail is consistently segmented.

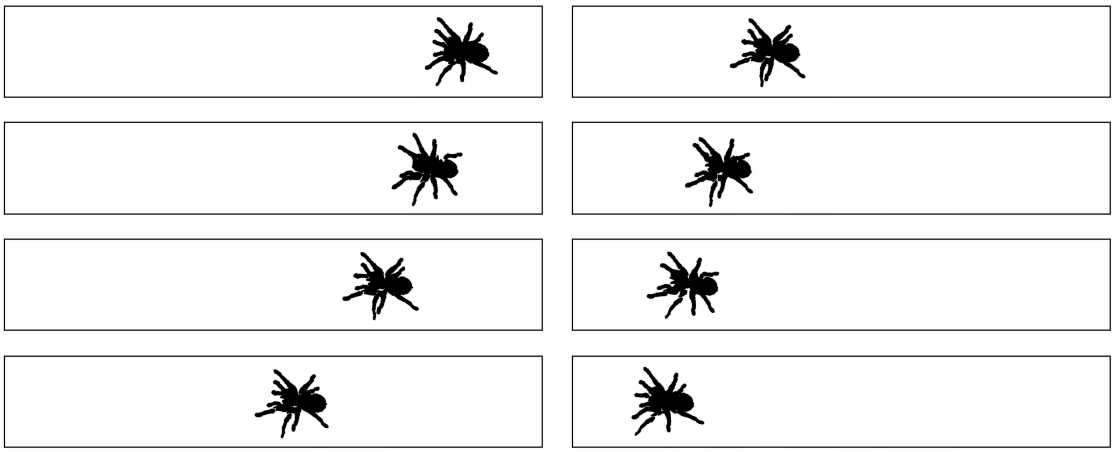

Figure 4: Novel body and part motions sampled from our model after being fitted to `spider`. Body frame is subjected to constant velocity while part transformations are sampled. See supplemental video for more views.

Figure 5: Original spider video for comparison with synthetic results.

3 Synthesized Part Motions

In Figure 4, we sample new part motions from the model after all parameters have been sampled from the `spider` dataset. Specifically, we generate new body transformations, in which the spider is subject to constant velocity and no rotation. Part transformations θ_{tk} are seeded with inference results then resampled from their full conditionals. Observations are taken from a single frame of the original video, projected into their respective part coordinate systems according to the inferred part assignments, then reprojected to new world coordinates using the newly synthesized body and part transformations at each time. We stress that these are novel part motions and that they can be generated for arbitrary durations and body paths. Video of these results is available in the video supplemental.

We observe that parts close to the spider's center exhibit relative stability, and the legs demonstrate the expected rhythmic walking motion. The pedipalps (the two front appendages) display implausible rotations, however. This is because these parts undergo foreshortening and occlusion in the original dataset. Since occlusion is not explicitly handled by our model in $\mathrm{SE}(2)$, inference permitted large rotations to explain observations on the pedipalps as they go from visible to not visible and vice versa. Nevertheless, the spider and it's basic walking motion remains recognizable.

4 Stable Dynamics

Stable Motion Dynamics

Random Walk

The random walk model is used to approximate the dynamics of a moving object in many tracking applications.

$$x_t = Ax_{t-1} + n_t \tag{28}$$
$$n_t \sim \mathrm{N}(n_t \mid 0, \Sigma_n) \tag{29}$$
$$(n_t, n_s) \sim \mathrm{N}(n_t \mid 0, \Sigma_n)\,\mathrm{N}(n_s \mid 0, \Sigma_n) \quad \forall s \neq t \tag{30}$$

where, depending on the components of the state vector x_t (e.g. position only, position and velocity, or higher-ordered terms), the matrix A encodes linear dynamics used to predict the current kinematic state x_t from the kinematic state x_{t-1} at the previous time step. The statistics of n_t are used to explain deviations from a deterministic trajectory.

For the purpose of inference, suitable tracking results can often be obtained even with a very approximate model of the dynamics. This is especially true for cases where the results are likelihood dominated *i.e.*, objects are easily distinguishable based on their appearance).

It is well known that the random walk model is unstable and for the method described in the manuscript the instability can significantly degrade results as detected parts may be kinematically ambiguous. Here, we describe a stabilized random walk model that is a special case of models encountered in linear dynamical control problems for stabilizing unstable plants.

It is straightforward to express the covariance of x_t as a function of the covariance x_{t-1}

$$\Sigma_t = A\Sigma_{t-1}A^{\mathrm{T}} + \Sigma_n \tag{31}$$

where instability depends upon the eigen-values of the matrix A. If the system were *stable* then we could solve for stationary covariance matrix, Σ (where the dependence on t has been dropped) via the following expression

$$\Sigma = A\Sigma A^{\mathrm{T}} + \Sigma_n \tag{32}$$

The expression above is a special case of the well known Algebraic Ricatti Equation. The existence of a solution depends on whether the eigenvalues of A are less than 1.

For the simple random walk dynamical model, A takes the form

$$A = \begin{bmatrix} 1 & 0 & 0 & \cdots & 0 \\ 0 & 1 & 0 & \cdots & 0 \\ 0 & 0 & 1 & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \cdots & 1 \end{bmatrix} \tag{33}$$

where the dimension of A depends on the number of motion terms used for x_t and each element of x_t corresponds to position of the ith dimension. All of the eigenvalues of A equal 1 with multiplicity equal to the dimension of x_t. Furthermore, this form of A results in unstable dynamics and there is no solution to Eqn. (32).

As described in the manuscript, we adopt a modified form of the random walk model, which we refer

to as *stabilized* random walk.

$$x_t = \begin{bmatrix} \sqrt{\alpha} & 0 & 0 & 0 \\ 0 & \sqrt{\alpha} & 0 & 0 \\ 0 & 0 & \ddots & 0 \\ 0 & 0 & 0 & \sqrt{\alpha} \end{bmatrix} x_{t-1} + \begin{bmatrix} \sqrt{1-\alpha} & 0 & 0 & 0 \\ 0 & \sqrt{1-\alpha} & 0 & 0 \\ 0 & 0 & \ddots & 0 \\ 0 & 0 & 0 & \sqrt{1-\alpha} \end{bmatrix} n_t \tag{34}$$

(35)

where $0 < \alpha < 1$.

$$\Sigma_t = \alpha \Sigma_{t-1} + (1-\alpha)\,\Sigma_n \tag{36}$$

for which the solution to Eqn. (32) becomes

$$\Sigma = \Sigma_n \tag{37}$$

This approach can be extended to more complex linear dynamical models and is equivalent to determining a gain on the state feedback that models the observed motion, although, obtaining solutions can result in additional complexity. For our purposes, the simple form of A was sufficient. For any $0 < \alpha < 1$ the marginal covariance of the position vector will be equal to Σ_n and as $\alpha \to 1$ the trajectories become increasingly smooth. This behavior is useful for approximating the motion of parts about a centroid without inducing instability.

References

[1] Andrew Gelman, John B Carlin, Hal S Stern, David B Dunson, Aki Vehtari, and Donald B Rubin. *Bayesian data analysis*. Chapman and Hall/CRC, 2013.

[2] Kevin P Murphy. *Machine learning: a probabilistic perspective*. MIT press, 2012.

[3] Radford M Neal et al. Slice sampling. *The annals of statistics*, 31(3):705–767, 2003.