Bayesian Learning and Inference in Recurrent Switching Linear Dynamical Systems

Scott W. Linderman* Columbia University

Matthew J. Johnson* Harvard and Google Brain

Andrew C. Miller Harvard University

Ryan P. Adams Harvard and Google Brain

David M. Blei Columbia University

Liam Paninski Columbia University

1 Stochastic Variational Inference

The main paper introduces a Gibbs sampling algorithm for the recurrent SLDS and its siblings, but it is straightforward to derive a mean field variational inference algorithm as well. From this, we can immediately derive a stochastic variational inference (SVI) [Hoffman et al., 2013] algorithm for conditionally independent time series.

We use a structured mean field approximation on the augmented model,

$$p(z_{1:T}, x_{1:T}, \omega_{1:T}, \theta \,|\, y_{1:T}) \approx q(z_{1:T})\, q(x_{1:T})\, q(\omega_{1:T})\, q(\theta; \eta).$$

The first three factors will not be explicitly parameterized; rather, as with Gibbs sampling, we leverage standard message passing algorithms to compute the necessary expectations with respect to these factors. Moreover, $q(\omega_{1:T})$ further factorizes as,

$$q(\omega_{1:T}) = \prod_{t=1}^{T} \prod_{k=1}^{K-1} q(\omega_{t,k}).$$

To be concrete, we also expand the parameter factor,

$$q(\theta; \eta) = \prod_{k=1}^{K} q(R_k, r_k \,|\, \eta_{\mathsf{rec},k})\, q(A_k, b_k, B_k; \eta_{\mathsf{dyn},k}) \times q(C_k, d_k, D_k; \eta_{\mathsf{obs},k}).$$

The algorithm proceeds by alternating between optimizing $q(x_{1:T})$, $q(z_{1:T})$, $q(\omega_{1:T})$, and $q(\theta)$.

Updating $q(x_{1:T})$. Fixing the factor on the discrete states $q(z_{1:T})$, the optimal variational factor on the continuous states $q(x_{1:T})$ is determined by,

$$\ln q(x_{1:T}) = \psi(x_1) + \sum_{t=1}^{T-1} \psi(x_t, x_{t+1}) + \sum_{t=1}^{T-1} \psi(x_t, z_{t+1}, \omega_t) + \sum_{t=1}^{T} \psi(x_t; y_t) + c.$$

where

$$\psi(x_1) = \mathbb{E}_{q(\theta)q(z)} \ln p(x_1 \,|\, z_1, \theta) \quad (1)$$

$$\psi(x_t, x_{t+1}) = \mathbb{E}_{q(\theta)q(z)} \ln p(x_{t+1} \,|\, x_t, z_t, \theta), \quad (2)$$

$$\psi(x_t, z_{t+1}) = \mathbb{E}_{q(\theta)q(z)q(\omega)} \ln p(z_{t+1} \,|\, x_t, z_t, \omega_t, \theta), \quad (3)$$

Because the densities $p(x_1 \,|\, z_1, \theta)$ and $p(x_{t+1} \,|\, x_t, z_t, \theta)$ are Gaussian exponential families, the expectations in Eqs. (1)-(2) can be computed efficiently, yielding Gaussian potentials with natural parameters that depend on both $q(\theta)$ and $q(z_{1:T})$. Furthermore, each $\psi(x_t; y_t)$ is itself a Gaussian potential. As in the Gibbs sampler, the only non-Gaussian potential comes from the logistic stick breaking model, but once again, the Pólya-gamma augmentation scheme comes to the rescue. After augmentation, the potential as a function of x_t is,

$$\mathbb{E}_{q(\theta)q(z)q(\omega)} \ln p(z_{t+1} \,|\, x_t, z_t, \omega_t, \theta) = -\frac{1}{2} \nu_{t+1}^{\mathsf{T}} \Omega_t \nu_{t+1} + \nu_{t+1}^{\mathsf{T}} \kappa(z_{t+1}) + c.$$

Since $\nu_{t+1} = R_{z_t} x_t + r_{z_t}$ is linear in x_t, this is another Gaussian potential. As with the dynamics and observation potentials, the recurrence weights, (R_k, r_k), also have matrix normal factors, which are conjugate after augmentation. We also need access to $\mathbb{E}_q[\omega_{t,k}]$; we discuss this computation below.

After augmentation, the overall factor $q(x_{1:T})$ is a Gaussian linear dynamical system with natural parameters computed from the variational factor on the

dynamical parameters $q(\theta)$, the variational parameter on the discrete states $q(z_{1:T})$, the recurrence potentials $\{\psi(x_t, z_t, z_{t+1})\}_{t=1}^{T-1}$, and the observation model potentials $\{\psi(x_t; y_t)\}_{t=1}^{T}$.

Because the optimal factor $q(x_{1:T})$ is a Gaussian linear dynamical system, we can use message passing to perform efficient inference. In particular, the expected sufficient statistics of $q(x_{1:T})$ needed for updating $q(z_{1:T})$ can be computed efficiently.

Updating $q(\omega_{1:T})$. We have,

$$\begin{aligned}\ln q(\omega_{t,k}) &= \mathbb{E}_q \ln p(z_{t+1} \mid \omega_t, x_t) + c \\ &= -\frac{1}{2}\mathbb{E}_q[\nu_{t+1}^2]\omega_{t,k} \\ &\quad + \mathbb{E}_{q(z_{1:T})} \ln p_{\mathrm{PG}}(\omega_{t,k} \mid \mathbb{I}[z_{t+1} \geq k], 0) + c\end{aligned}$$

While the expectation with respect to $q(z_{1:T})$ makes this challenging, we can approximate it with a sample, $\hat{z}_{1:T} \sim q(z_{1:T})$. Given a fixed value $\hat{z}_{1:T}$ we have,

$$q(\omega_{t,k}) = p_{\mathrm{PG}}(\omega_{t,k} \mid \mathbb{I}[\hat{z}_{t+1} \geq k], \mathbb{E}_q[\nu_{t+1}^2]).$$

The expected value of the distribution is available in closed form:

$$\mathbb{E}_q[\omega_{t,k}] = \frac{\mathbb{I}[\hat{z}_{t+1} \geq k]}{2\mathbb{E}_q[\nu_{t+1}^2]} \tanh\left(\tfrac{1}{2}\mathbb{E}_q[\nu_{t+1}^2]\right).$$

Updating $q(z_{1:T})$. Similarly, fixing $q(x_{1:T})$ the optimal factor $q(z_{1:T})$ is proportional to

$$\exp\left\{\psi(z_1) + \sum_{t=1}^{T-1}\psi(z_t, x_t, z_{t+1}) + \sum_{t=1}^{T}\psi(z_t)\right\},$$

where

$$\begin{aligned}\psi(z_1) &= \mathbb{E}_{q(\theta)} \ln p(z_1 \mid \theta) + \mathbb{E}_{q(\theta)q(x)} \ln p(x_1 \mid z_1, \theta) \\ \psi(z_t, x_t, z_{t+1}) &= \mathbb{E}_{q(\theta)q(x_{1:T})} \ln p(z_{t+1} \mid z_t, x_t) \\ \psi(z_t) &= \mathbb{E}_{q(\theta)q(x)} \ln p(x_{t+1} \mid x_t, z_t, \theta)\end{aligned}$$

The first and third densities are exponential families; these expectations can be computed efficiently. The challenge is the recurrence potential,

$$\psi(z_t, x_t, z_{t+1}) = \mathbb{E}_{q(\theta),q(x)} \ln \pi_{\mathsf{SB}}(\nu_{t+1}).$$

Since this is not available in closed form, we approximate this expectation with Monte Carlo over x_t, R_k, and r_k. The resulting factor $q(z_{1:T})$ is an HMM with natural parameters that are functions of $q(\theta)$ and $q(x_{1:T})$, and the expected sufficient statistics required for updating $q(x_{1:T})$ can be computed efficiently by message passing in the same manner.

Updating $q(\theta)$. To compute the expected sufficient statistics for the mean field update on η, we can also use message passing, this time in both factors $q(x_{1:T})$ and $q(z_{1:T})$ separately. The required expected sufficient statistics are of the form

$$\begin{gathered}\mathbb{E}_{q(z)}\mathbb{I}[z_t = i, z_{t+1} = j], \quad \mathbb{E}_{q(z)}\mathbb{I}[z_t = i], \\ \mathbb{E}_{q(z)}\mathbb{I}[z_t = k]\mathbb{E}_{q(x)}\left[x_t x_{t+1}^{\mathsf{T}}\right], \\ \mathbb{E}_{q(z)}\mathbb{I}[z_t = k]\mathbb{E}_{q(x)}\left[x_t x_t^{\mathsf{T}}\right], \quad \mathbb{E}_{q(z)}\mathbb{I}[z_1 = k]\mathbb{E}_{q(x)}[x_1],\end{gathered} \tag{4}$$

where $\mathbb{I}[\,\cdot\,]$ denotes an indicator function. Each of these can be computed easily from the marginals $q(x_t, x_{t+1})$ and $q(z_t, z_{t+1})$ for $t = 1, 2, \ldots, T-1$, and these marginals can be computed in terms of the respective graphical model messages.

Given the conjugacy of the augmented model, the dynamics and observation factors will be MNIW distributions as well. These allow closed form expressions for the required expectations,

$$\begin{gathered}\mathbb{E}_q[A_k], \quad \mathbb{E}_q[b_k], \quad \mathbb{E}_q[A_k B_k^{-1}], \quad \mathbb{E}_q[b_k B_k^{-1}], \quad \mathbb{E}_q[B_k^{-1}], \\ \mathbb{E}_q[C_k], \quad \mathbb{E}_q[d_k], \quad \mathbb{E}_q[C_k D_k^{-1}], \quad \mathbb{E}_q[d_k D_k^{-1}], \quad \mathbb{E}_q[D_k^{-1}].\end{gathered}$$

Likewise, the conjugate matrix normal prior on (R_k, r_k) provides access to

$$\mathbb{E}_q[R_k], \qquad \mathbb{E}_q[R_k R_k^{\mathsf{T}}], \qquad \mathbb{E}_q[r_k].$$

Stochastic Variational Inference. Given multiple, conditionally independent observations of time series, $\{y_{1:T_p}^{(p)}\}_{p=1}^{P}$ (using the same notation as in the basketball experiment), it is straightforward to derive a stochastic variational inference (SVI) algorithm [Hoffman et al., 2013]. In each iteration, we sample a random time series; run message passing to compute the optimal local factors, $q(z_{1:T_p}^{(p)})$, $q(x_{1:T_p}^{(p)})$, and $q(\omega_{1:T_p}^{(p)})$; and then use expectations with respect to these local factors as unbiased estimates of expectations with respect to the complete dataset when updating the global parameter factor, $q(\theta)$. Given a single, long time series, we can still derive efficient SVI algorithms that use subsets of the data, as long as we are willing to accept minor, controllable bias [Johnson and Willsky, 2014, Foti et al., 2014].

2 Initialization

Given the complexity of these models, it is important to initialize the parameters and latent states with reasonable values. We used the following initialization procedure: (i) use probabilistic PCA or factor analysis to initialize the continuous latent states, $x_{1:T}$, and the observation, C, D, and d; (ii) fit an AR-HMM to $x_{1:T}$ in order to initialize the discrete latent states, $z_{1:T}$,

and the dynamics models, $\{A_k, Q_k, b_k\}$; and then (iii) greedily fit a decision list with logistic regressions at each node in order to determine a permutation of the latent states most amenable to stick breaking. In practice, the last step alleviates the undesirable dependence on ordering that arises from the stick breaking formulation.

As mentioned in Section 4, one of the less desirable features of the logistic stick breaking regression model is its dependence on the ordering of the output dimensions; in our case, on the permutation of the discrete states $\{1, 2, \ldots, K\}$. To alleviate this issue, we first do a greedy search over permutations by fitting a decision list to $(x_t, z_t), z_{t+1}$ pairs. A decision list is an iterative classifier of the form,

$$z_{t+1} = \begin{cases} o_1 & \text{if } \mathbb{I}[p_1] \\ o_2 & \text{if } \mathbb{I}[\neg p_1 \wedge p_2] \\ o_3 & \text{if } \mathbb{I}[\neg p_1 \wedge \neg p_2 \wedge p_3] \\ \vdots & \\ o_K & \text{o.w.,} \end{cases}$$

where $(o_1, \ldots, o_K)$ is a permutation of $(1, \ldots, K)$, and $p_1, \ldots, p_k$ are predicates that depend on (x_t, z_t) and evaluate to true or false. In our case, these predicates are given by logistic functions,

$$p_j = \sigma(r_j^\mathsf{T} x_t) > 0.$$

We fit the decision list using a greedy approach: to determine o_1 and r_1, we use maximum a posterior estimation to fit logistic regressions for each of the K possible output values. For the k-th logistic regression, the inputs are $x_{1:T}$ and the outputs are $y_t = \mathbb{I}[z_{t+1} = k]$. We choose the best logistic regression (measured by log likelihood) as the first output. Then we remove those time points for which $z_{t+1} = o_1$ from the dataset and repeat, fitting $K - 1$ logistic regressions in order to determine the second output, o_2, and so on.

After iterating through all K outputs, we have a permutation of the discrete states. Moreover, the predicates $\{r_k\}_{k=1}^{K-1}$ serve as an initialization for the recurrence weights, R, in our model.

3 Bernoulli-Lorenz Details

The Pólya-gamma augmentation makes it easy to handle discrete observations in the rSLDS, as illustrated in the Bernoulli-Lorenz experiment. Since the Bernoulli likelihood is given by,

$$\begin{aligned} p(y_t \mid z_t, \theta) &= \prod_{n=1}^{N} \text{Bern}(\sigma(c_n^\mathsf{T} x_t + d_n) \\ &= \prod_{n=1}^{N} \frac{(e^{c_n^\mathsf{T} x_t + d_n})^{y_{t,n}}}{1 + e^{c_n^\mathsf{T} x_t + d_n}}, \end{aligned}$$

we see that it matches the form of (7) with,

$$\nu_{t,n} = c_n^\mathsf{T} x_t + d_n, \quad b(y_{t,n}) = 1, \quad \kappa(y_{t,n}) = y_{t,n} - \frac{1}{2}.$$

Thus, we introduce an additional set of Pólya-gamma auxiliary variables,

$$\xi_{t,n} \sim \text{PG}(1, 0),$$

to render the model conjugate. Given these auxiliary variables, the observation potential is proportional to a Gaussian distribution on x_t,

$$\psi(x_t, y_t) \propto \mathcal{N}(C x_t + d \mid \Xi_t^{-1} \kappa(y_t), \Xi_t^{-1}),$$

with

$$\begin{aligned} \Xi_t &= \text{diag}\left([\xi_{t,1}, \ldots, \xi_{t,N}]\right), \\ \kappa(y_t) &= [\kappa(y_{t,1}), \ldots, \kappa(y_{t,N})]. \end{aligned}$$

Again, this admits efficient message passing inference for $x_{1:T}$. In order to update the auxiliary variables, we sample from their conditional distribution, $\xi_{t,n} \sim \text{PG}(1, \nu_{t,n})$.

This augmentation scheme also works for binomial, negative binomial, and multinomial obsevations as well [Polson et al., 2013].

4 Basketball Details

For completeness, Figures 1 and 2 show all $K = 30$ inferred states of the rAR-HMM (ro) for the basketball data.

References

Nicholas Foti, Jason Xu, Dillon Laird, and Emily Fox. Stochastic variational inference for hidden Markov models. In *Advances in Neural Information Processing Systems*, pages 3599–3607, 2014.

Matthew D Hoffman, David M Blei, Chong Wang, and John William Paisley. Stochastic variational inference. *Journal of Machine Learning Research*, 14(1): 1303–1347, 2013.

Matthew J. Johnson and Alan S. Willsky. Stochastic variational inference for Bayesian time series models. In *Proceedings of the 31st International Conference on Machine Learning (ICML-14)*, pages 1854–1862, 2014.

Nicholas G Polson, James G Scott, and Jesse Windle. Bayesian inference for logistic models using Pólya-gamma latent variables. *Journal of the American Statistical Association*, 108(504):1339–1349, 2013.

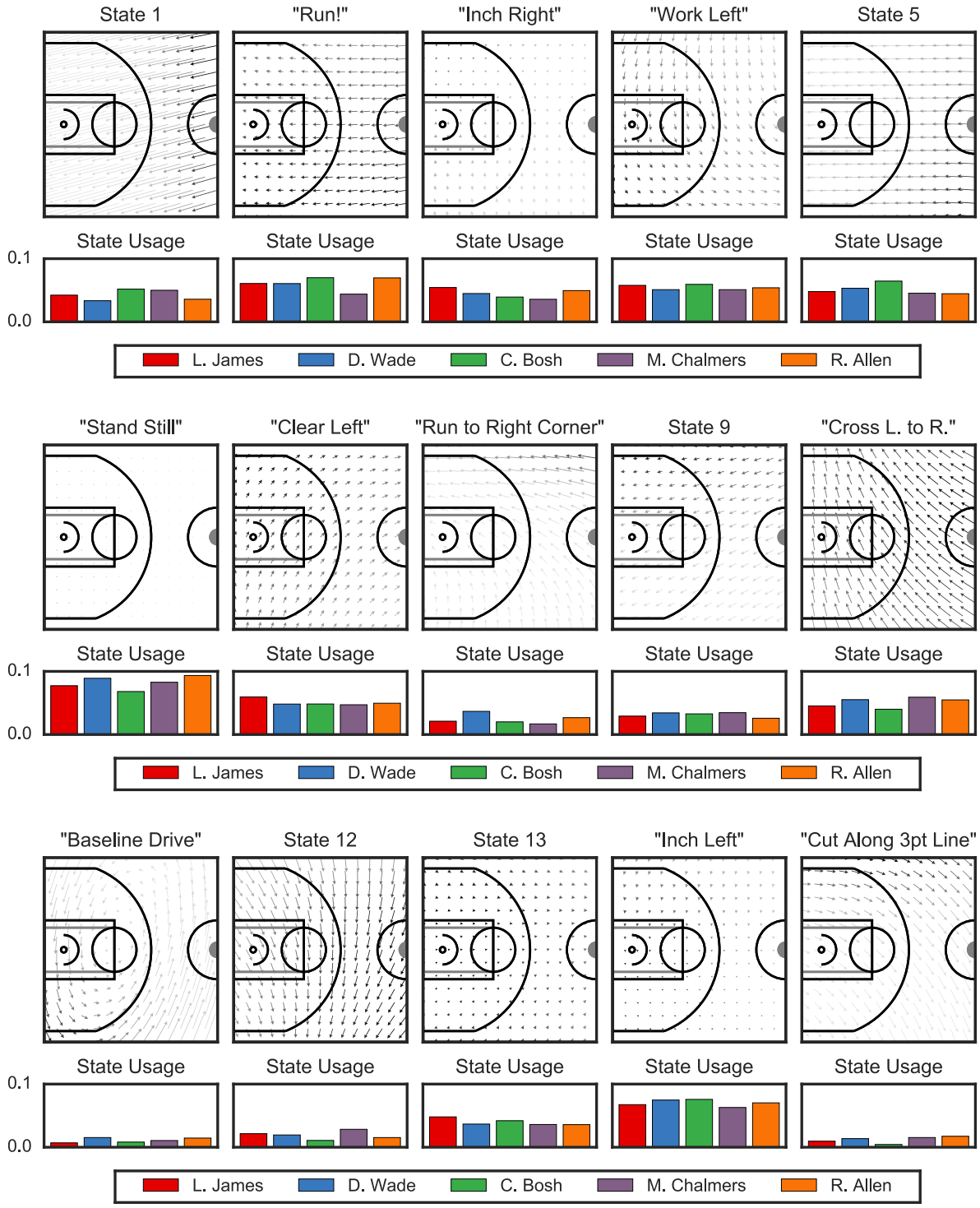

Figure 1: All of the inferred basketball states

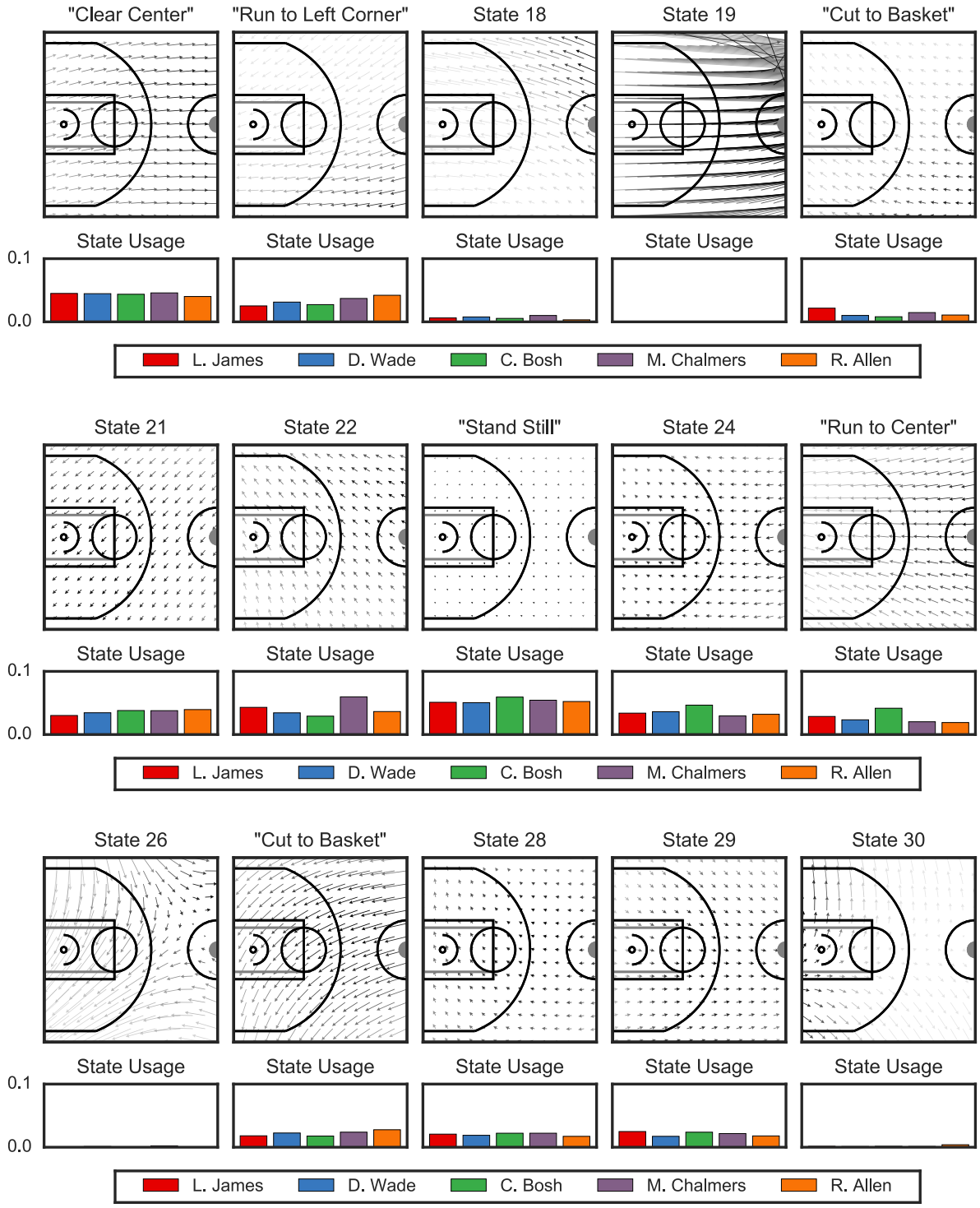

Figure 2: All of the inferred basketball states