# A. Theoretical Analysis

## A.1. Convergence Analysis

**Theorem A.1.** *(Bottou et al., 2018) If the step size $\eta_t \equiv \eta \leq \frac{1}{Lc_g}$, then a fixed resolution solution satisfies*

$$\mathbb{E}[\|\mathbf{w}_{t+1} - \mathbf{w}_\star\|_2^2] \leq \gamma^t[\mathbb{E}[\|\mathbf{w}_0 - \mathbf{w}_\star\|_2^2] - \beta] + \beta$$

*where $\gamma = 1 - 2\eta\mu$, $\beta = \frac{\eta\sigma_g^2}{2\mu}$, and $\mathbf{w}_\star$ is the optimal solution.*

*Proof.* For a single step update,

$$\begin{aligned}\|\mathbf{w}_{t+1} - \mathbf{w}_\star\|_2^2 &= \|\mathbf{w}_t - \eta_t \mathbf{g}(\mathbf{w}_t; \xi_t) - \mathbf{w}_\star\|_2^2 \\ &= \|\mathbf{w}_t - \mathbf{w}_\star\|_2^2 + \|\eta_t \mathbf{g}(\mathbf{w}_t; \xi_t)\|_2^2 - 2\eta_t \mathbf{g}(\mathbf{w}_t; \xi_t)(\mathbf{w}_t - \mathbf{w}_\star)\end{aligned} \tag{9}$$

by the law of total expectation

$$\begin{aligned}\mathbb{E}[\mathbf{g}(\mathbf{x}_t; \xi_t)(\mathbf{w}_t - \mathbf{w}_\star)] &= \mathbb{E}[\mathbb{E}[\mathbf{g}(\mathbf{w}_t; \xi_t)(\mathbf{w}_t - \mathbf{w}_\star)|\xi_{<t}]] \\ &= \mathbb{E}[(\mathbf{w}_t - \mathbf{w}_\star)\mathbb{E}[\mathbf{g}(\mathbf{w}_t; \xi_t)|\xi_{<t}]] \\ &= \mathbb{E}[(\mathbf{w}_t - \mathbf{w}_\star)^\top \nabla f(\mathbf{w}_t)]\end{aligned} \tag{10}$$

From strong convexity,

$$\langle \nabla f(\mathbf{w}_t) - \nabla f(\mathbf{w}_\star), \mathbf{w}_t - \mathbf{w}_\star \rangle = \langle \nabla f(\mathbf{w}_t), \mathbf{w}_t - \mathbf{w}_\star \rangle \geq \mu \|\mathbf{w}_t - \mathbf{w}_\star\|_2^2 \tag{11}$$

which implies $\mathbb{E}[(\mathbf{w}_t - \mathbf{w}_\star)^\top \nabla f(\mathbf{w}_t)] \geq \mu \mathbb{E}[\|\mathbf{w}_t - \mathbf{w}_\star\|_2^2]$ as $\nabla f(\mathbf{w}_\star) = 0$. Putting it all together yields

$$\mathbb{E}[\|\mathbf{w}^{t+1} - \mathbf{w}_\star\|_2^2] \leq (1 - 2\eta_t\mu)\mathbb{E}[\|\mathbf{w}_t - \mathbf{w}_\star\|_2^2] + (\eta_t \sigma_g)^2 \tag{12}$$

As $\eta_t = \eta$, we complete the contraction, by setting $\beta = \frac{(\eta\sigma_g)^2}{(2\eta\mu)}$

$$\mathbb{E}[\|\mathbf{w}_{t+1} - \mathbf{w}_\star\|_2^2] - \beta \leq (1 - 2\eta_t\mu)(\mathbb{E}[\|\mathbf{w}_t - \mathbf{w}_\star\|_2^2] - \beta) \tag{13}$$

Repeat the iterations

$$\mathbb{E}[\|\mathbf{w}_{t+1} - \mathbf{w}_\star\|_2^2] - \beta \leq (1 - 2\eta\mu)^t(\mathbb{E}[\|\mathbf{w}_0 - \mathbf{w}_\star\|_2^2] - \beta) \tag{14}$$

Rearranging the terms, we get

$$\mathbb{E}[\|\mathbf{w}_{t+1} - \mathbf{w}_\star\|_2^2] \leq (1 - 2\eta\mu)^t \mathbb{E}[\|\mathbf{w}_0 - \mathbf{w}_\star] - ((1 - 2\eta\mu)^t + 1)\frac{(\eta\sigma_g)^2}{(2\eta\mu)} \tag{15}$$

□

**Theorem A.2.** *If the step size $\eta_t \equiv \eta \leq \frac{1}{Lc_g}$, then `MRTL` solution satisfies*

$$\begin{aligned}\mathbb{E}[\|\mathbf{w}_t^{(r)} - \mathbf{w}^\star\|_2^2] \leq \gamma^t(\|P\|_{op}^2)^r[\mathbb{E}[\|\mathbf{w}_0^{(1)} - \mathbf{w}^{(1),\star}\|_2^2] \\ - \gamma^t\|P\|_2^2\beta + \gamma^{t_2}(\|P\|_{op}^2\beta - \beta) + O(1)\end{aligned}$$

*where $\gamma = 1 - 2\eta\mu$, $\beta = \frac{\eta\sigma_g^2}{2\mu}$, and $\|P\|_{op}$ is the operator norm of the interpolation operator $P$.*

Consider a two resolution case where $R = 2$ and $\mathbf{w}_\star^{(2)} = \mathbf{w}_\star$. Let $t_r$ be the total number of iterations of resolution $r$. Based on Eqn. (12), for a fixed resolution algorithm, after $t_1 + t_2$ number of iterations,

$$\mathbb{E}[\|\mathbf{w}_{t_1+t_2} - \mathbf{w}_\star\|_2^2] - \beta \leq (1 - 2\eta\mu)^{t_1+t_2}(\mathbb{E}[\|\mathbf{w}_0 - \mathbf{w}^\star\|_2^2] - \beta)$$

For multiresolution, where we train on resolution $r = 1$ first, we have

$$E[\|\mathbf{w}_{t_1}^{(1)} - \mathbf{w}_\star^{(1)}\|_2^2] - \beta \leq (1 - 2\eta\mu)^{t_1}(\mathbb{E}[\|\mathbf{w}_0^{(1)} - \mathbf{w}_\star^{(1)}\|_2^2] - \beta)$$

At resolution $r = 2$, we have

$$\mathbb{E}[\|\mathbf{w}_{t_2}^{(2)} - \mathbf{w}_\star\|_2^2] - \beta \leq (1 - 2\eta\mu)^{t_2}(\mathbb{E}[\|\mathbf{w}_0^{(2)} - \mathbf{w}_\star\|_2^2] - \beta) \tag{16}$$

Using interpolation, we have $\mathbf{w}_0^{(2)} = P\mathbf{w}_{t_1}^{(1)}$. Given the spatial autocorrelation assumption, we have

$$\|\mathbf{w}_\star^{(2)} - P\mathbf{w}_\star^{(1)}\|_2 \leq \epsilon$$

By the definition of operator norm and triangle inequality,

$$\mathbb{E}[\|\mathbf{w}_0^{(2)} - \mathbf{w}_\star^{(2)}\|_2^2 \leq \mathbb{E}[\|P\mathbf{w}_{t_1}^{(1)} - \mathbf{w}_\star^{(2)}\|_2^2] \leq \|P\|_{op}^2\mathbb{E}[\|\mathbf{w}_{t_1}^{(1)} - \mathbf{w}_\star^{(1)}\|_2^2] + \epsilon^2$$

Combined with eq. (16), we have

$$\mathbb{E}[\|\mathbf{w}_{t_2}^{(2)} - \mathbf{w}_\star\|_2^2] - \beta \leq (1 - 2\eta\mu)^{t_2}(\|P\|_{op}^2\mathbb{E}[\|\mathbf{w}_{t_1}^{(1)} - \mathbf{w}_\star^{(1)}\|_2^2] + \epsilon^2 - \beta) \tag{17}$$

$$= (1 - 2\eta\mu)^{t_1+t_2}\|P\|_{op}^2(\mathbb{E}[\|\mathbf{w}_0^{(1)} - \mathbf{w}_\star^{(1)}\|_2^2] - \beta) + (1 - 2\eta\mu)^{t_2}(\|P\|_{op}^2\beta + \epsilon^2 - \beta) \tag{18}$$

If we initialize $\mathbf{w}_0$ and $\mathbf{w}_0^{(1)}$ such that $\|\mathbf{w}_0^{(1)} - \mathbf{w}_\star^{(1)}\|_2^2 = \|\mathbf{w}_0 - \mathbf{w}_\star\|_2^2$, we have `MRTL` solution

$$\mathbb{E}[\|\mathbf{w}_{t_1+t_2}^{'} - \mathbf{w}_\star\|_2^2] - \alpha \leq (1 - 2\eta\mu)^{t_1+t_2}\|P\|_{op}^2(\mathbb{E}[\|\mathbf{w}_0^{'} - \mathbf{w}_\star\|_2^2] - \alpha) \tag{19}$$

for some $\alpha$ that completes the contraction. Repeat the resolution iterates in Eqn. (18), we reach our conclusion. □

### A.2. Computational Complexity Analysis

In this section, we analyze the computational complexity for `MRTL` (Algorithm 1). Assuming that $\nabla f$ is Lipschitz continuous, we can view gradient-based optimization as a fixed-point iteration operator $F$ with a contraction constant of $\gamma \in (0, 1)$ (note that *stochastic* gradient descent converges to a noise ball instead of a fixed point).

$$\mathbf{w} \leftarrow F(\mathbf{w}), \quad F := I - \eta\nabla f, \|F(\mathbf{w}) - F(\mathbf{w}')\| \leq \gamma\|\mathbf{w} - \mathbf{w}'\|.$$

Let $\mathbf{w}_\star^{(r)}$ be the optimal estimator at resolution $r$. Suppose for each resolution $r$, we use the following finegrain criterion:

$$\|\mathbf{w}_t^{(r)} - \mathbf{w}_{t-1}^{(r)}\| \leq \frac{C_0 D_r}{\gamma(1-\gamma)}. \tag{20}$$

where $t_r$ is the number of iterations taken at level $r$. The algorithm terminates when the estimation error reaches $\frac{C_0 R}{(1-\gamma)^2}$. The following main theorem characterizes the speed-up gained by multiresolution learning `MRTL` w.r.t. the contraction factor $\gamma$ and the terminal estimation error $\epsilon$.

**Theorem A.3.** *Suppose the fixed point iteration operator (gradient descent) for the optimization algorithm has a contraction factor (Lipschitz constant) of $\gamma$, the multiresolution learning procedure is faster than that of the fixed resolution algorithm by a factor of $\log \frac{1}{(1-\gamma)\epsilon}$, with $\epsilon$ as the terminal estimation error.*

We prove several useful Lemmas before proving the main Theorem A.3. The following lemma analyzes the computational cost of the *fixed-resolution* algorithm.

**Lemma A.4.** *Given a fixed point iteration operator with a contraction factor $\gamma$, the computational complexity of a fixed-resolution training for a $p$-order tensor with rank $K$ is*

$$\mathcal{C} = \mathcal{O}\left(\frac{1}{|\log\gamma|} \cdot \log\left(\frac{1}{(1-\gamma)\epsilon}\right) \cdot \frac{Kp}{(1-\gamma)^2\epsilon}\right). \tag{21}$$

*Proof.* At a high level, we can prove this by choosing a small enough resolution $r$ such that the approximation error is bounded with a fixed number of iterations. Let $\mathbf{w}_\star^{(r)}$ be the optimal estimate at resolution $r$ and $\mathbf{w}_t$ be the estimate at step $t$. Then

$$\|\mathbf{w}_\star - \mathbf{w}_t\| \leq \|\mathbf{w}_\star - \mathbf{w}_\star^{(r)}\| + \|\mathbf{w}_\star^{(r)} - \mathbf{w}_t\| \leq \epsilon. \tag{22}$$

We pick a fixed resolution $r$ small enough such that

$$\|\mathbf{w}_\star - \mathbf{w}_\star^{(r)}\| \leq \frac{\epsilon}{2}, \tag{23}$$

then using the termination criteria $\|\mathbf{w}_\star - \mathbf{w}_\star^{(r)}\| \leq \frac{C_0 R}{(1-\gamma)^2}$ gives $D_r = \Omega((1-\gamma)^2\epsilon)$ where $D_r$ is the discretization size at resolution $r$. Initialize $\mathbf{w}_0 = 0$ and apply $F$ to $\mathbf{w}$ for $t$ times such that

$$\frac{\gamma^t}{2(1-\gamma)}\|F(\mathbf{w}_0)\| \leq \frac{\epsilon}{2}. \tag{24}$$

As $\mathbf{w}_0 = 0$, $\|F(\mathbf{w}_0)\| \leq 2C$, we obtain that

$$t \leq \frac{1}{|\log\gamma|} \cdot \log\left(\frac{2C}{(1-\gamma)\epsilon}\right), \tag{25}$$

Note that for an order $p$ tensor with rank $K$, the computational complexity of every iteration in `MRTL` is $\mathcal{O}(Kp/D_r)$ with $D_r$ as the discretization size. Hence, the computational complexity of the fixed resolution training is

$$\begin{aligned}\mathcal{C} &= \mathcal{O}\left(\frac{1}{|\log\gamma|} \cdot \log\left(\frac{1}{(1-\gamma)\epsilon}\right) \cdot \frac{Kp}{D_r}\right) \\ &= \mathcal{O}\left(\frac{1}{|\log\gamma|} \cdot \log\left(\frac{1}{(1-\gamma)\epsilon}\right) \cdot \frac{Kp}{(1-\gamma)^2\epsilon}\right). \quad \square\end{aligned}$$

Given a spatial discretization $r$, we can construct an operator $F_r$ that learns discretized tensor weights. The next lemma relates the estimation error with resolution. The following lemma relates the estimation error with resolution:

**Lemma A.5.** *(Nash, 2000) For each resolution level $r = 1, \cdots, R$, there exists a constant $C_1$ and $C_2$, such that the fixed point iteration with discretization size $D_r$ has an estimation error:*

$$\|F(\mathbf{w}) - F^{(r)}(\mathbf{w})\| \leq (C_1 + \gamma C_2\|\mathbf{w}\|)D_r \tag{26}$$

*Proof.* See (Nash, 2000) for details.

We have obtained the discretization error for the fixed point operation at any resolution. Next we analyze the number of iterations $t_r$ needed at each resolution $r$ before finegraining.

**Lemma A.6.** *For every resolution $r = 1, \ldots, R$, there exists a constant $C'$ such that the number of iterations $t_r$ before finegraining satisfies:*

$$t_r \leq C'/\log|\gamma| \tag{27}$$

*Proof.* According to the fixed point iteration definition, we have for each resolution $r$:

$$\begin{aligned}\|F_r(\mathbf{w}_{t_r}) - \mathbf{w}_{t_r}^{(r)})\| &\leq \gamma^{t_r-1}\|F_r(\mathbf{w}_0^{(r)}) - \mathbf{w}_0^{(r)}\| &(28)\\ &\leq \gamma^{t_r-1}\frac{C_0 D_r}{1-\gamma} &(29)\\ &\leq C'\gamma^{t_r-1} &(30)\end{aligned}$$

using the definition of the finegrain criterion. □

By combining Lemmas A.6 and the computational cost per iteration, we can compute the total computational cost for our MRTL algorithm, which is proportional to the total number of iterations for all resolutions:

$$\begin{aligned}
\mathcal{C}_{\texttt{MRTL}} &= \mathcal{O}\left(\frac{1}{|\log\gamma|}\left[(D_r/Kp)^{-1} + (2D_r/Kp)^{-1} + (4D_r/Kp)^{-1} + \cdots\right]\right) \\
&= \mathcal{O}\left(\frac{1}{|\log\gamma|}\left(\frac{Kp}{D_r}\right)\left[1 + \frac{1}{2} + \frac{1}{4} + \cdots\right]\right) \\
&= \mathcal{O}\left(\frac{1}{|\log\gamma|}\left(\frac{Kp}{D_r}\right)\left[\frac{1-(\frac{1}{2})^n}{1-\frac{1}{2}}\right]\right) \\
&= \mathcal{O}\left(\frac{1}{|\log\gamma|}\left(\frac{Kp}{(1-\gamma)^2\epsilon}\right)\right), \qquad (31)
\end{aligned}$$

where the last step uses the termination criterion in (20). Comparing with the complexity analysis for the fixed resolution algorithm in Lemma A.4, we complete the proof. □

## B. Experiment Details

**Basketball** We list implementation details for the basketball dataset. We focus only on half-court possessions, where all players have crossed into the half court as in (Yue et al., 2014). The ball must also be inside the court and be within a 4 foot radius of the ballhandler. We discard any passing/turnover events and do not consider frames with free throws.

For the ball handler location $\{D_r^1\}$, we discretize the half-court into resolutions $4 \times 5, 8 \times 10, 20 \times 25, 40 \times 50$. For the relative defender locations, at the full resolution, we choose a $12 \times 12$ grid around the ball handler where the ball handler is located at $(6, 2)$ (more space in front of the ball handler than behind him/her). We also consider a smaller grid around the ball handler for the defender locations, assuming that defenders that are far away from the ball handler do not influence shooting probability. We use $6 \times 6, 12 \times 12$ for defender positions.

Let us denote the pair of resolutions as $(D_r^1, D_r^2)$. We train the full-rank model at resolutions $(4 \times 5, 6 \times 6), (8 \times 10, 6 \times 6), (8 \times 10, 12 \times 12)$ and the low-rank model at resolutions $(8 \times 10, 12 \times 12), (20 \times 25, 12 \times 12), (40 \times 50, 12 \times 12)$.

There is a notable class imbalance in labels (88% of data points have zero labels) so we use weighted cross entropy loss using the inverse of class counts as weights. For the low-rank model, we use tensor rank $K = 20$. The performance trend of MRTL is similar across a variety of tensor ranks. $K$ should be chosen appropriately to the desired level of approximation.

**Climate** We describe the data sources used for climate. The precipitation data comes from the PRISM group (PRISM Climate Group, 2013), which provides estimates monthly estimates at 1/24$^o$ spatial resolution across the continental U.S from 1895 to 2018. For oceanic data we use the EN4 reanalysis product (Good et al., 2013), which provides monthly estimates for ocean salinity and temperature at 1$^o$ spatial resolution across the globe from 1900 to the present (see Fig. 3). We constrain our spatial analysis to the range [-180$^o$W, 0$^o$W] and [-20$^o$S, 60$^o$N], which encapsulates the area around North America and a large portion of South America.

The ocean data is non-stationary, with the variance of the data increasing over time. This is likely due to improvement in observational measurements of ocean temperature and salinity over time, which reduce the amount of interpolation needed to generate an estimate for a given month. After detrending and deseasonalizing, we split the train, validation, and test sets using random consecutive sequences so that their samples come from a similar distribution.

We train the full-rank model at resolutions $4 \times 9$ and $8 \times 18$ and the low-rank model at resolutions $8 \times 18$, $12 \times 27$, $24 \times 54$, $40 \times 90$, $60 \times 135$, and $80 \times 180$. For finegraining criteria, we use a patience factor of 4, i.e. training was terminated when a finegraining criterion was reached a total of 4 times. Both validation loss and gradient statistics were relatively noisy during training (possibly due to a small number of samples), leading to early termination without the patience factor.

During finegraining, the weights were upsampled to the higher resolution using bilinear interpolation and then scaled by the ratio of the number of inputs for the higher resolution to the number of inputs for the lower resolution (as described in Section 4) to preserve the magnitude of the prediction.

**Details** We trained the basketball dataset on 4 RTX 2080 Ti GPUs, while the climate dataset experiments were performed on a separate workstation with 1 RTX 2080 Ti GPU. The computation times of the fixed-resolution and `MRTL` model were compared on the same setup for all experiments.

### B.1. Hyperparameters

| Hyperparameter | Basketball | Climate |
|---|---|---|
| Batch size | $32 - 1024$ | $8 - 128$ |
| Full-rank learning rate $\eta$ | $10^{-3} - 10^{-1}$ | $10^{-4} - 10^{-1}$ |
| Full-rank regularization $\lambda$ | $10^{-5} - 10^{0}$ | $10^{-4} - 10^{-1}$ |
| Low-rank learning rate $\eta$ | $10^{-5} - 10^{-1}$ | $10^{-4} - 10^{-1}$ |
| Low-rank regularization $\lambda$ | $10^{-5} - 10^{0}$ | $10^{-4} - 10^{-1}$ |
| Spatial regularization $\sigma$ | $0.03 - 0.2$ | $0.03 - 0.2$ |
| Learning rate decay $\gamma$ | $0.7 - 0.95$ | $0.7 - 0.95$ |

*Table 2.* Search range for `Opt` hyperparameters

Table 2 show the search ranges of all hyperparameters considered. We performed separate random searches over this search space for `MRTL`, fixed-resolution model, and the randomly initialized low-rank model. We also separate the learning rate $\eta$ and regularization coefficient $\lambda$ between the full-rank and low-rank models.

### B.2. Accuracy and Convergence

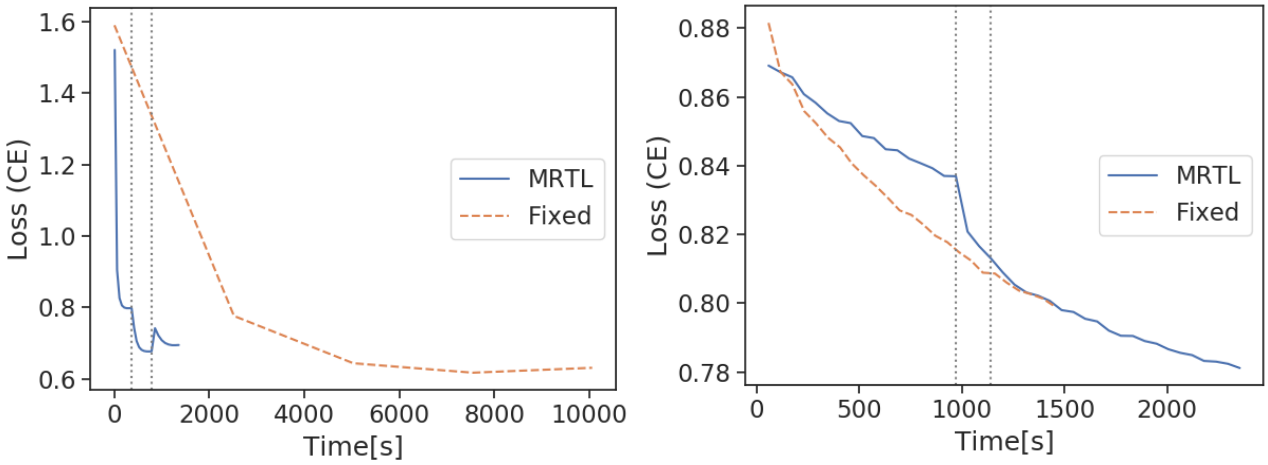


*Figure 10.* Basketball: Loss curves of `MRTL` vs. the fixed-resolution model for the full rank (left) and low rank model (right). The vertical lines indicate finegraining to the next resolution.

Fig. 10 shows the loss curves of `MRTL` vs. the fixed resolution model for the full rank and low rank case. They show a similar convergence trend, where the fixed-resolution model is much slower than `MRTL`.

### B.3. Finegraining Criteria

Table 3 lists the results for the different finegraining criteria. In the classification case, we see that validation loss reaches much faster convergence than other gradient-based criteria in the full-rank case, while the gradient-based criteria are faster for the low-rank model. All criteria can reach similar F1 scores. For the regression case, all stopping criteria converge to a similar loss in roughly the same amount of time for the full-rank model. For the low-rank model, validation loss appears to converge more quickly and to a lower loss value.

### B.4. Random initialization

Fig. 11 shows all latent factors after training `MRTL` vs a randomly initialized low-rank model for ballhandler position. We can see clearly that full-rank initialization produces spatially coherent factors while random initialization can produce some

*Table 3.* Runtime and prediction performance comparison of different finegraining criteria

| Dataset | Model | Full-Rank | | | Low-Rank | | |
|---|---|---|---|---|---|---|---|
| | | Time [s] | Loss | F1 | Time [s] | Loss | F1 |
| Basketball | Validation loss | **1230** ±74.1 | 0.699 ±0.00237 | 0.607 ±0.00182 | 2009 ±715 | 0.868 ±0.0399 | 0.475 ±0.0121 |
| | Gradient norm | 7029 ±759 | 0.703 ±0.00216 | **0.610** ±0.00149 | **912** ±281 | 0.883 ±0.00664 | 0.476 ±0.00270 |
| | Gradient variance | 7918 ±1949 | 0.701 ±0.00333 | 0.609 ±0.00315 | 933 ±240 | 0.883 ±0.00493 | **0.476** ±0.00197 |
| | Gradient entropy | 8715 ±957 | 0.697 ±0.00551 | 0.597 ±0.00737 | 939 ±259 | 0.886 ±0.00248 | 0.475 ±0.00182 |
| Climate | Validation loss | 1.04 ±0.115 | **0.0448** ±0.0108 | - | **37.4** ±28.7 | **0.0284** ±0.00171 | - |
| | Gradient norm | 1.11 ±0.0413 | 0.0506 ±0.00853 | - | 59.1 ±16.9 | 0.0301 ±0.00131 | - |
| | Gradient variance | 1.14 ±0.0596 | 0.0458 ±0.00597 | - | 62.9 ±14.4 | 0.0305 ±0.00283 | - |
| | Gradient entropy | **0.984** ±0.0848 | 0.0490 ±0.0144 | - | 48.4 ±21.1 | 0.0331 ±0.00949 | - |

uninterpretable factors (e.g. the latent factors for $k = 3, 4, 5, 7, 19, 20$ are not semantically meaningful). Fig. 12 shows latent factors for the defender position spatial mode, and we can draw similar conclusions about random initialization.

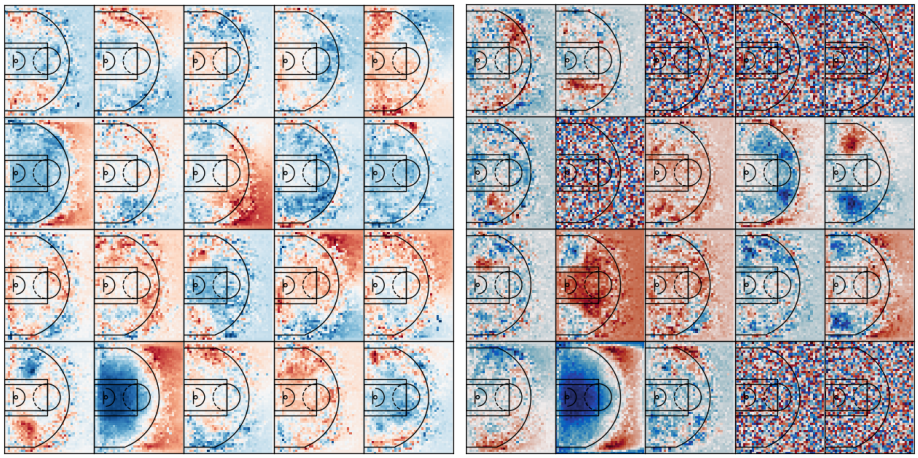

*Figure 11.* Basketball: Latent factors of ball handler position after training `MRTL` (left) and a low-rank model using random initialization (right). The factors have been normalized to (-1,1) so that reds are positive and blues are negative. The latent factors are numbered left to right, top to bottom.

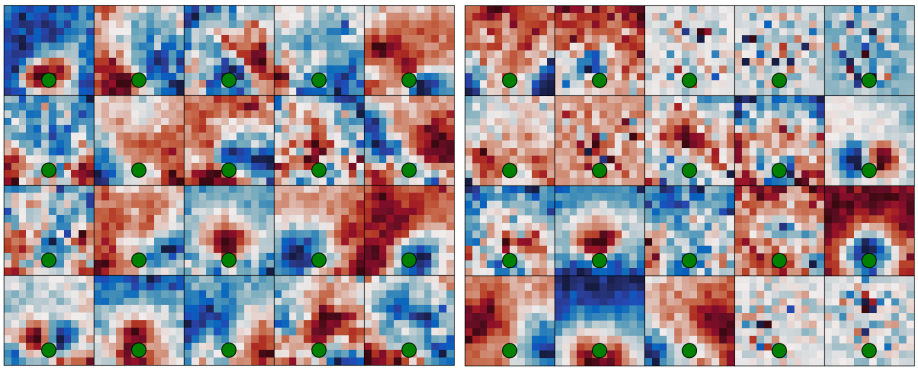

*Figure 12.* Basketball: Latent factors of relative defender positions after training `MRTL` (left) and a low-rank model using random initialization (right). The factors have been normalized to (-1,1) so that reds are positive and blues are negative. The green dot represents the ballhandler at (6, 2). The latent factors are numbered left to right, top to bottom.