A. Online Appendix

This appendix details the recursive update of $\mathbf{L}$ and theoretical guarantees described in the main paper:

> Paul Ruvolo and Eric Eaton. ELLA: An Efficient Lifelong Learning Algorithm. In *Proceedings of the 30th International Conference on Machine Learning*, Atlanta, GA, 2013.

A.1. Recursive Update of L

A naïve algorithm for updating the latent model component matrix $\mathbf{L}$ whenever new data are received is to invert the matrix $\frac{1}{T}A + \lambda I$. The computational complexity of this update is $O(d^3k^3)$. However, it is possible to speedup the computation by exploiting the fact that the matrix $\mathbf{A}$ is only updated by adding or subtracting a low-rank matrix. The updates to $\mathbf{A}$ at each iteration have the form:

$$\begin{aligned}
\mathbf{A} &\leftarrow \mathbf{A} - \left(\mathbf{s}^{(t)}\mathbf{s}^{(t)\top}\right) \otimes \mathbf{D}^{(t)} \\
&= \mathbf{A} - \left(\mathbf{s}^{(t)} \otimes \mathbf{D}^{(t)\frac{1}{2}}\right)\left(\mathbf{s}^{(t)} \otimes \mathbf{D}^{(t)\frac{1}{2}}\right)^\top \\
\mathbf{A} &\leftarrow \mathbf{A} + \left(\mathbf{s}^{(t)\prime}\mathbf{s}^{(t)\prime\top}\right) \otimes \mathbf{D}^{(t)\prime} \\
&= \mathbf{A} + \left(\mathbf{s}^{(t)\prime} \otimes \mathbf{D}^{(t)^\prime\frac{1}{2}}\right)\left(\mathbf{s}^{(t)\prime} \otimes \mathbf{D}^{(t)^\prime\frac{1}{2}}\right)^\top ,
\end{aligned}$$

where we use tick marks to denote the updated versions of $\mathbf{D}^{(t)}$ and $\mathbf{s}^{(t)}$ after receiving the new training data, and $\mathbf{D}^{(t)\frac{1}{2}}$ is the matrix square-root of $\mathbf{D}^{(t)}$. The updates to $\mathbf{A}$ consist of adding or subtracting an outer-product of a matrix of size $(d \times k)$-by-d, which implies that each update has rank at most d. If we have already computed the eigenvalue decomposition of the old $\mathbf{A}$, we can compute the eigenvalue decomposition of the updated value of $\mathbf{A}$ in $O(d^3k^2)$ using the recursive decomposition algorithm proposed by Yu (1991). Given the eigenvalue decomposition of the updated value of $\mathbf{A} = \mathbf{U}\mathbf{\Sigma}\mathbf{U}^\top$, we can compute the new value of $\mathbf{L}$ by considering the resulting linear system in canonical form:

$$\begin{aligned}
\text{vec}\,(\mathbf{L}) &= \mathbf{U}\boldsymbol{\psi} &(11)\\
\psi_i &= \frac{\left(\frac{1}{T}\mathbf{U}^\top\mathbf{b}\right)_i}{\lambda + \frac{1}{T}\sigma_{i,i}} . &(12)
\end{aligned}$$

Computing the vector $\boldsymbol{\psi}$ requires multiplying a $(d\times k)$-by-$(d\times k)$ matrix by a vector of size $(d\times k)$ for a complexity of $O(d^2k^2)$. To complete the computation of $\mathbf{L}$ requires another matrix multiplication with the same size input matrices yielding another $O(d^2k^2)$. Combining the recursive computation of the eigenvalue decomposition and the computation of $\mathbf{L}$ yields a computational complexity of $O(d^3k^2)$ for the update step — a factor of k speedup from the naïve implementation.

A.2. Convergence Proof

In this section, we present complete proofs for the three results on the convergence of ELLA (previously described in Section 3.6 of the main paper):

1. The latent model component matrix, $\mathbf{L}_T$, becomes increasingly stable as the number of tasks T increases.
2. The value of the surrogate cost function, $\hat{g}_T(\mathbf{L}_T)$, and the value of the true empirical cost function, $g_T(\mathbf{L}_T)$, converge almost surely (a.s.) as the number of tasks learned goes to infinity.
3. $\mathbf{L}_T$ converges asymptotically to a stationary point of the expected loss g.

These three convergence results are given below as Propositions 1–3.

These results are based on the following assumptions:

A. The tuples $\left(\mathbf{D}^{(t)}, \boldsymbol{\theta}^{(t)}\right)$ are drawn *i.i.d.* from a distribution with compact support (bounding the entries of $\mathbf{D}^{(t)}$ and $\boldsymbol{\theta}^{(t)}$).

B. For all $\mathbf{L}$, $\mathbf{D}^{(t)}$, and $\boldsymbol{\theta}^{(t)}$, the smallest eigenvalue of $\mathbf{L}_\gamma^\top\mathbf{D}^{(t)}\mathbf{L}_\gamma$ is at least κ (with $\kappa > 0$), where γ is the subset of non-zero indices of the vector $\mathbf{s}^{(t)} = \arg\min_{\mathbf{s}} \|\boldsymbol{\theta}^{(t)} - \mathbf{L}\mathbf{s}\|^2_{\mathbf{D}^{(t)}}$. In this case, the non-zero elements of the unique minimizing $\mathbf{s}^{(t)}$ are given by: $\mathbf{s}^{(t)}{}_\gamma = \left(\mathbf{L}_\gamma^\top\mathbf{D}^{(t)}\mathbf{L}_\gamma\right)^{-1}\left(\mathbf{L}_\gamma^\top\mathbf{D}^{(t)}\boldsymbol{\theta}^{(t)} - \mu\boldsymbol{\epsilon}_\gamma\right)$, where $\boldsymbol{\epsilon}_\gamma$ is a vector containing the signs of the non-zero entries of $\mathbf{s}^{(t)}$.

Claim 1: $\exists\ c_1 \in \mathbb{R}$ *such that no element of* $\mathbf{L}_T$ *has magnitude greater than* c_1, $\forall\ T \in \{1 \ldots \infty\}$.

Proof: Consider the solution $\mathbf{L}_T = \mathbf{0}$. Since each $\boldsymbol{\theta}^{(t)}$ and $\mathbf{D}^{(t)}$ are both bounded by Assumption (A), the loss incurred on Equation 5 for the tth task when $\mathbf{L}_T = \mathbf{0}$ is $\boldsymbol{\theta}^{(t)\top}\mathbf{D}^{(t)}\boldsymbol{\theta}^{(t)}$, which is bounded by Assumption (A). The part of Equation 5 consisting of the average loss over tasks can be no larger than the maximum loss on a single task (which as we just showed is bounded). Therefore, $\hat{g}_T(\mathbf{0})$ must be bounded by some constant independent of T. Provided $\lambda > 0$, there must exist a constant c_1 to bound the maximum entry in $\mathbf{L}_T$ or else the regularization term would necessarily cause $\hat{g}_T(\mathbf{L}_T)$ to exceed $\hat{g}_T(\mathbf{0})$. ■

Claim 2: $\exists\ c_2 \in \mathbb{R}$ *such that the maximum magnitude of the entries of* $\mathbf{s}^{(t)}$ *is bounded by* c_2, $\forall\ T \in \{1 \ldots \infty\}$.

Proof: The value of $\mathbf{s}^{(t)}$ is given by the solution to Equation 3. We can use a similar argument as we did in Claim 1 to show that the magnitude of the entries of $\mathbf{s}^{(t)}$ must be bounded (i.e., by considering $\mathbf{s}^{(t)} = \mathbf{0}$ and showing the loss of this solution is bounded by Assumption (A)). ■

Proposition 1: $\mathbf{L}_T - \mathbf{L}_{T-1} = O\left(\frac{1}{T}\right)$.

Proof: First, we show that $\hat{g}_T - \hat{g}_{T-1}$ is Lipschitz with constant $O\left(\frac{1}{T}\right)$:

$$\begin{aligned}\hat{g}_T(\mathbf{L}) - \hat{g}_{T-1}(\mathbf{L}) &= \frac{1}{T}\ell(\mathbf{L}, \mathbf{s}^{(T)}, \boldsymbol{\theta}^{(T)}, \mathbf{D}^{(T)}) \\ &\quad + \frac{1}{T}\sum_{t=1}^{T-1}\ell(\mathbf{L}, \mathbf{s}^{(t)}, \boldsymbol{\theta}^{(t)}, \mathbf{D}^{(t)}) \\ &\quad - \frac{1}{T-1}\sum_{t=1}^{T-1}\ell(\mathbf{L}, \mathbf{s}^{(t)}, \boldsymbol{\theta}^{(t)}, \mathbf{D}^{(t)}) \\ &= \frac{1}{T}\ell(\mathbf{L}, \mathbf{s}^{(T)}, \boldsymbol{\theta}^{(T)}, \mathbf{D}^{(t)}) \\ &\quad - \frac{1}{T(T-1)}\sum_{t=1}^{T-1}\ell(\mathbf{L}, \mathbf{s}^{(t)}, \boldsymbol{\theta}^{(t)}, \mathbf{D}^{(t)})\end{aligned}$$

If ℓ is Lipschitz in its first argument with a constant independent of T, then $\hat{g}_T - \hat{g}_{T-1}$ has a Lipschitz constant $O\left(\frac{1}{T}\right)$. This is true since $\hat{g}_T - \hat{g}_{T-1}$ is equal to the difference of two terms: the first of which is ℓ divided by T, and the second is an average over $T-1$ terms (which can have Lipschitz constant no greater than the largest Lipschitz constant of the functions being averaged) which is then normalized by T. We can easily see that ℓ is Lipschitz with constant $O(1)$ since it is a quadratic function over a compact region with all coefficients bounded. Therefore, $\hat{g}_T - \hat{g}_{T-1}$ is Lipschitz with constant $O\left(\frac{1}{T}\right)$.

Let ξ_T be the Lipschitz constant of $\hat{g}_T - \hat{g}_{T-1}$. We have:

$$\begin{aligned}\hat{g}_{T-1}(\mathbf{L}_T) - \hat{g}_{T-1}(\mathbf{L}_{T-1}) &= \hat{g}_{T-1}(\mathbf{L}_T) - \hat{g}_T(\mathbf{L}_T) \\ &\quad + \hat{g}_T(\mathbf{L}_T) - \hat{g}_T(\mathbf{L}_{T-1}) \\ &\quad + \hat{g}_T(\mathbf{L}_{T-1}) - \hat{g}_{T-1}(\mathbf{L}_{T-1}) \\ &\leq \hat{g}_{T-1}(\mathbf{L}_T) - \hat{g}_T(\mathbf{L}_T) \\ &\quad + \hat{g}_T(\mathbf{L}_{T-1}) - \hat{g}_{T-1}(\mathbf{L}_{T-1}) \\ &\leq \xi_T\|\mathbf{L}_T - \mathbf{L}_{T-1}\|_{\mathsf{F}} \ . \qquad (13)\end{aligned}$$

Additionally, since $\mathbf{L}_{T-1}$ minimizes $\hat{g}_{T-1}$ and the L_2 regularization term ensures that the minimum eigenvalue of the Hessian of $\hat{g}_{T-1}$ is lower-bounded by 2λ, we have that $\hat{g}_{T-1}(\mathbf{L}_T) - \hat{g}_{T-1}(\mathbf{L}_{T-1}) \geq 2\lambda\|\mathbf{L}_T - \mathbf{L}_{T-1}\|_{\mathsf{F}}^2$. Combining these two inequalities, we have:

$$\|\mathbf{L}_T - \mathbf{L}_{T-1}\|_{\mathsf{F}} \quad \leq \quad \frac{\xi_T}{2\lambda} = O\left(\frac{1}{T}\right) \ .$$

Therefore, $\mathbf{L}_T - \mathbf{L}_{T-1} = O\left(\frac{1}{T}\right)$. ■

Before stating our next proposition, we define the function:

$$\alpha(\mathbf{L}, \boldsymbol{\theta}^{(t)}, \mathbf{D}^{(t)}) = \arg\min_{\mathbf{s}} \ell(\mathbf{L}, \mathbf{s}, \boldsymbol{\theta}^{(t)}, \mathbf{D}^{(t)}) \ . \qquad (14)$$

For brevity we will also use the notation $\boldsymbol{\alpha}_{\mathbf{L},\boldsymbol{\theta}^{(t)},\mathbf{D}^{(t)}} = \alpha(\mathbf{L}, \boldsymbol{\theta}^{(t)}, \mathbf{D}^{(t)})$. We define the following lemma to support the proof of the next proposition:

Lemma 1:

A. $\min_{\mathbf{s}} \ell(\mathbf{L}, \mathbf{s}, \boldsymbol{\theta}^{(t)}, \mathbf{D}^{(t)})$ is continuously differentiable in $\mathbf{L}$ with $\nabla_{\mathbf{L}} \min_{\mathbf{s}} \ell(\mathbf{L}, \mathbf{s}, \boldsymbol{\theta}^{(t)}, \mathbf{D}^{(t)}) = -2\mathbf{D}^{(t)}\left(\boldsymbol{\theta}^{(t)} - \mathbf{L}\alpha(\mathbf{L}, \boldsymbol{\theta}^{(t)}, \mathbf{D}^{(t)})\right)\alpha(\mathbf{L}, \boldsymbol{\theta}^{(t)}, \mathbf{D}^{(t)})^{\top}$.

B. g is continuously differentiable with $\nabla g(\mathbf{L}) = 2\lambda\mathbf{I} + \mathbb{E}_{\boldsymbol{\theta}^{(t)},\mathbf{D}^{(t)}}\left[\nabla_{\mathbf{L}} \min_{\mathbf{s}} \ell(\mathbf{L}, \mathbf{s}^{(t)}, \boldsymbol{\theta}^{(t)}, \mathbf{D}^{(t)})\right]$.

C. $\nabla_{\mathbf{L}} g(\mathbf{L})$ is Lipschitz on the space of latent components $\mathbf{L}$ that obey Claim (1).

Proof: To prove Part (A), we apply a corollary to Theorem 4.1 as stated in (Bonnans & Shapiro, 1998) (originally shown in (Danskin, 1967)). As applied to our problem, this corollary states that if ℓ is continuously differentiable in $\mathbf{L}$ (which it clearly is) and has a unique minimizer $\mathbf{s}^{(t)}$ regardless of $\boldsymbol{\theta}^{(t)}$ and $\mathbf{D}^{(t)}$ (which is guaranteed by Assumption (B)), then $\nabla_{\mathbf{L}} \min_{\mathbf{s}} \ell(\mathbf{L}, \mathbf{s}, \boldsymbol{\theta}^{(t)}, \mathbf{D}^{(t)})$ exists and is equal to $\nabla_{\mathbf{L}} \ell(\mathbf{L}, \alpha(\mathbf{L}, \boldsymbol{\theta}^{(t)}, \mathbf{D}^{(t)}), \boldsymbol{\theta}^{(t)}, \mathbf{D}^{(t)})$. Following some simple algebra, we arrive at the specific form of the gradient listed as Part (A). Part (B) can be proven immediately since by Assumption (A) the tuple $\left(\mathbf{D}^{(t)}, \boldsymbol{\theta}^{(t)}\right)$ is drawn from a distribution with compact support.

To prove Part (C), we first show that $\alpha(\mathbf{L}, \boldsymbol{\theta}^{(t)}, \mathbf{D}^{(t)})$ is Lipschitz in $\mathbf{L}$ with constant independent of $\boldsymbol{\theta}^{(t)}$ and $\mathbf{D}^{(t)}$. Part (C) will follow once α has been shown to be Lipschitz due to the form of the gradient of g with respect to $\mathbf{L}$. The function α is continuous in its arguments since ℓ is continuous in its arguments and by Assumption (B) has a unique minimizer. Next, we define the function $\rho(\mathbf{L}, \mathbf{D}^{(t)}, \boldsymbol{\theta}^{(t)}, j) = \left(\mathbf{D}^{(t)}\mathbf{l}_j\right)^{\top}\left(\boldsymbol{\theta}^{(t)} - \mathbf{L}\boldsymbol{\alpha}_{\mathbf{L},\boldsymbol{\theta}^{(t)},\mathbf{D}^{(t)}}\right)$, where $\mathbf{l}_j$ represents the jth column of $\mathbf{L}$, and state the following facts about $\rho(\mathbf{L}, \mathbf{D}^{(t)}, \boldsymbol{\theta}^{(t)}, j)$:

$$\begin{aligned}\left|\rho(\mathbf{L}, \mathbf{D}^{(t)}, \boldsymbol{\theta}^{(t)}, j)\right| &= \mu, \text{ iff } \left(\boldsymbol{\alpha}_{\mathbf{L},\boldsymbol{\theta}^{(t)},\mathbf{D}^{(t)}}\right)_j \neq 0 \\ \left|\rho(\mathbf{L}, \mathbf{D}^{(t)}, \boldsymbol{\theta}^{(t)}, j)\right| &< \mu, \text{ iff } \left(\boldsymbol{\alpha}_{\mathbf{L},\boldsymbol{\theta}^{(t)},\mathbf{D}^{(t)}}\right)_j = 0 \ . \qquad (15)\end{aligned}$$

Let γ be the set of indices j such that $\left|\rho(\mathbf{L}, \mathbf{D}^{(t)}, \boldsymbol{\theta}^{(t)}, j)\right| = \mu$. Since $\rho(\mathbf{L}, \mathbf{D}^{(t)}, \boldsymbol{\theta}^{(t)}, j)$ is continuous in $\mathbf{L}$, $\mathbf{D}^{(t)}$, and $\boldsymbol{\theta}^{(t)}$, there must exist an open neighborhood around $\left(\mathbf{L}, \mathbf{D}^{(t)}, \boldsymbol{\theta}^{(t)}\right)$ called V such that for all $\left(\mathbf{L}', \mathbf{D}^{(t)\prime}, \boldsymbol{\theta}^{(t)\prime}\right) \in V$ and $j \notin \gamma$, $\left|\rho(\mathbf{L}', \mathbf{D}^{(t)\prime}, \boldsymbol{\theta}^{(t)\prime}, j)\right| < \mu$. By Equation 15, we can conclude that $\left(\boldsymbol{\alpha}_{\mathbf{L}',\boldsymbol{\theta}^{(t)\prime},\mathbf{D}^{(t)\prime}}\right)_j = 0, \ \forall \ j \notin \gamma$.

Next, we define a new loss function:

$$\bar{\ell}(\mathbf{L}_\gamma, \mathbf{s}_\gamma, \boldsymbol{\theta}, \mathbf{D}) \quad = \quad \|\boldsymbol{\theta} - \mathbf{L}_\gamma\mathbf{s}_\gamma\|_{\mathbf{D}}^2 + \mu\|\mathbf{s}_\gamma\|_1 \ .$$

By Assumption (B) we are guaranteed that $\bar{\ell}$ is strictly convex with a Hessian lower-bounded by κ. Based on

this, we can conclude that:

$$\begin{aligned}&\overline{\ell}(\mathbf{L}_\gamma, \boldsymbol{\alpha}_{\mathbf{L}',\boldsymbol{\theta}^{(t)\prime},\mathbf{D}^{(t)\prime}{}_\gamma}, \boldsymbol{\theta}^{(t)}, \mathbf{D}^{(t)}) \\ &\quad - \overline{\ell}(\mathbf{L}_\gamma, \boldsymbol{\alpha}_{\mathbf{L},\boldsymbol{\theta}^{(t)},\mathbf{D}^{(t)}{}_\gamma}, \boldsymbol{\theta}^{(t)}, \mathbf{D}^{(t)}) \\ &\quad \geq \kappa \|\boldsymbol{\alpha}_{\mathbf{L}',\boldsymbol{\theta}^{(t)\prime},\mathbf{D}^{(t)\prime}{}_\gamma} - \boldsymbol{\alpha}_{\mathbf{L},\boldsymbol{\theta}^{(t)},\mathbf{D}^{(t)}{}_\gamma}\|_2^2 \ . \end{aligned} \tag{16}$$

By Assumption (A) and Claim (1), $\overline{\ell}$ is Lipschitz in its second argument, $\mathbf{s}_\gamma$, with constant equal to

$$e_1\|\mathbf{L}_\gamma - \mathbf{L}'_\gamma\|_\mathsf{F} + e_2\|\boldsymbol{\theta}' - \boldsymbol{\theta}\|_2 + e_3\|\mathbf{D}' - \mathbf{D}\|_\mathsf{F}$$

(where e_1, e_2, e_3 are all constants independent of $\mathbf{L}_\gamma, \mathbf{L}'_\gamma, \boldsymbol{\theta}, \boldsymbol{\theta}', \mathbf{D}$, and $\mathbf{D}'$). Combining this fact with Equation 16, we obtain:

$$\begin{aligned}&\|\boldsymbol{\alpha}_{\mathbf{L}',\boldsymbol{\theta}^{(t)\prime},\mathbf{D}^{(t)\prime}} - \boldsymbol{\alpha}_{\mathbf{L},\boldsymbol{\theta}^{(t)},\mathbf{D}^{(t)}}\| = \\ &\quad \left\| \left(\boldsymbol{\alpha}_{\mathbf{L}',\boldsymbol{\theta}^{(t)\prime},\mathbf{D}^{(t)\prime}}\right)_\gamma - \left(\boldsymbol{\alpha}_{\mathbf{L},\boldsymbol{\theta}^{(t)},\mathbf{D}^{(t)}}\right)_\gamma \right\| \\ &\quad \leq \frac{e_1\|\mathbf{L}'_\gamma - \mathbf{L}_\gamma\|_\mathsf{F}}{\kappa} \\ &\qquad + \frac{e_2\|\boldsymbol{\theta}^{(t)\prime} - \boldsymbol{\theta}^{(t)}\|_2}{\kappa} \\ &\qquad + \frac{e_3\|\mathbf{D}^{(t)\prime} - \mathbf{D}^{(t)}\|_\mathsf{F}}{\kappa} \ . \end{aligned}$$

Therefore, α is locally-Lipschitz. Additionally, since the domain of α is compact by Assumption (A) and Claim (1), this implies that α is uniformly Lipschitz, and we can conclude that ∇g is Lipschitz as well. ■

Proposition 2:

1. $\hat{g}_T(\mathbf{L}_T)$ converges a.s.
2. $\hat{g}_T(\mathbf{L}_T) - g_T(\mathbf{L}_T)$ converges a.s. to 0
3. $\hat{g}_T(\mathbf{L}_T) - g(\mathbf{L}_T)$ converges a.s. to 0
4. $g(\mathbf{L}_T)$ converges a.s.

Proof: We begin by defining the stochastic process:

$$u_T \quad = \quad \hat{g}_T(\mathbf{L}) \ .$$

The basic proof outline is to show that this stochastic positive process (since the loss can never be negative) is a quasi-martingale and by a theorem in (Fisk, 1965) the stochastic process converges almost surely.

$$\begin{aligned}u_{T+1} - u_T &= \hat{g}_{T+1}(\mathbf{L}_{T+1}) - \hat{g}_T(\mathbf{L}_T) \\ &= \hat{g}_{T+1}(\mathbf{L}_{T+1}) - \hat{g}_{T+1}(\mathbf{L}_T) \\ &\quad + \hat{g}_{T+1}(\mathbf{L}_T) - \hat{g}_T(\mathbf{L}_T) \\ &= \hat{g}_{T+1}(\mathbf{L}_{T+1}) - \hat{g}_{T+1}(\mathbf{L}_T) \\ &\quad + \frac{\min_{\mathbf{s}^{(T+1)}} \ell(\mathbf{L}_T, \mathbf{s}^{(T+1)}, \boldsymbol{\theta}^{(T+1)}, \mathbf{D}^{(T+1)})}{T+1} \\ &\quad - \frac{g_T(\mathbf{L}_T)}{T+1} \\ &\quad + \frac{g_T(\mathbf{L}_T) - \hat{g}_T(\mathbf{L}_T)}{T+1} \ , \end{aligned} \tag{17}$$

where we made use of the fact that:

$$\begin{aligned}\hat{g}_{T+1}(\mathbf{L}_T) = &\frac{1}{T+1}\min_{\mathbf{s}^{(T+1)}} \ell(\mathbf{L}_T, \mathbf{s}^{(T+1)}, \boldsymbol{\theta}^{(T+1)}, \mathbf{D}^{(T+1)}) \\ &+ \frac{T}{T+1}\hat{g}_T(\mathbf{L}_T) \ . \end{aligned}$$

We now need to show that the sum of the positive variations in Equation 17 from $T = 1$ to $T = \infty$ is bounded. Note that the term on the first line of Equation 17 is guaranteed to be negative since $\mathbf{L}_{T+1}$ minimizes $\hat{g}_{T+1}$. Additionally, since $\hat{g}_T$ is always at least as large as g_T, the term on the last line is also guaranteed to be negative. Therefore, if we are interested in bounding the positive variations, we focus on the terms on the middle two lines.

$$\begin{aligned}&\mathbb{E}[u_{T+1} - u_T | \mathcal{G}_T] \leq \\ &\quad \frac{\mathbb{E}[\min_{\mathbf{s}^{(T+1)}} \ell(\mathbf{L}_T, \mathbf{s}^{(T+1)}, \boldsymbol{\theta}^{(T+1)}, \mathbf{D}^{(T+1)}) | \mathcal{I}_T]}{T+1} \\ &\quad - \frac{g_T(\mathbf{L}_T)}{T+1} \\ &\quad = \frac{g(\mathbf{L}_T) - g_T(\mathbf{L}_T)}{T+1} \\ &\quad \leq \frac{\|g - g_T\|_\infty}{T+1} \ , \end{aligned} \tag{18}$$

where $\mathcal{I}_T$ represents all of the information up to time T (i.e. all the previous $\boldsymbol{\theta}^{(t)}$'s and $\mathbf{D}^{(t)}$'s) and $\|\cdot\|_\infty$ is the infinity norm of a function (e.g. the maximum of the absolute value of the function). If we are able to show that $\sum_{t=1}^{\infty} \frac{\|g - g_t\|_\infty}{t+1} < \infty$ then we will have proven that the stochastic process u_T is a quasi-martingale that converges almost surely. In order to prove this, we apply the following corollary of the Donsker theorem ((Van der Vaart, 2000) Chapter 19.2, lemma 19.36, example 19.7):

> Let $\mathcal{G} = \{g_{\boldsymbol{\theta}} : \mathcal{X} \to \mathbb{R}, \boldsymbol{\theta} \in \boldsymbol{\Theta}\}$ be a set of measurable functions indexed by a bounded subset $\boldsymbol{\Theta}$ of $\mathbb{R}^d$. Suppose that there exists a constant K such that:
>
> $$|g_{\boldsymbol{\theta}_1}(x) - g_{\boldsymbol{\theta}_2}(x)| \leq K\|\boldsymbol{\theta}_1 - \boldsymbol{\theta}_2\|_2$$
>
> for every $\boldsymbol{\theta}_1, \boldsymbol{\theta}_2 \in \boldsymbol{\Theta}$ and $x \in \mathcal{X}$. Then, $\mathcal{G}$ is P-Donsker and for any $g \in \mathcal{G}$, we define $\mathbb{P}_n g$, $\mathbb{P}g$, and $\mathbb{G}_n g$ as:
>
> $$\begin{aligned}\mathbb{P}_n g &= \frac{1}{n}\sum_{i=1}^{n} g(X_i) \\ \mathbb{P}g &= \mathbb{E}_X[g(X)] \\ \mathbb{G}_n g &= \sqrt{n}(\mathbb{P}_n g - \mathbb{P}g) \ . \end{aligned}$$
>
> If $\mathbb{P}g^2 \leq \delta^2$ and $\|g\|_\infty < M$ and the random elements are Borel measurable, then:
>
> $$\mathbb{E}[\sup_{g \in G} |\mathbb{G}_n g|] = O(1) \ .$$

In order to apply this lemma to our analysis, consider a set of functions $\mathcal{H}$ indexed by possible latent component matrices, $\mathbf{L}$. Consider the domain of each of the functions in $\mathcal{H}$ to be all possible tuples $\left(\mathbf{D}^{(t)}, \boldsymbol{\theta}^{(t)}\right)$. We define $h_{\mathbf{L}}(\mathbf{D}^{(t)}, \boldsymbol{\theta}^{(t)}) = \min_{\mathbf{s}} \ell(\mathbf{L}, \mathbf{s}, \boldsymbol{\theta}^{(t)}, \mathbf{D}^{(t)})$. First, the expected value of h^2 is bounded for all $h \in \mathcal{H}$ since the value of ℓ is bounded on the set of $\mathbf{L}$ that conform to Claim (1). Second, $\|h\|_\infty$ again is bounded given Claim (1) and Assumption (A). Therefore, we can state that:

$$\mathbb{E}\left[\sqrt{T}\left\| \left(\frac{1}{T}\sum_{t=1}^{T} \min_{\mathbf{s}} \ell\left(\mathbf{L}, \mathbf{s}, \boldsymbol{\theta}^{(\mathbf{t})}, \mathbf{D}^{(\mathbf{t})}\right)\right) - \mathbb{E}\left[\min_{\mathbf{s}} \ell\left(\mathbf{L}, \mathbf{s}, \boldsymbol{\theta}^{(\mathbf{t})}, \mathbf{D}^{(\mathbf{t})}\right)\right] \right\|_\infty\right] = O(1)$$

$$\implies \mathbb{E}\left[\|g_T(\mathbf{L}) - g(\mathbf{L})\|_\infty\right] = O\left(\frac{1}{\sqrt{T}}\right) \ .$$

Therefore, $\exists\ c_3 \in \mathbb{R}$ such that $\mathbb{E}\left[\|g_T - g\|_\infty\right] < \frac{c_3}{\sqrt{T}}$:

$$\begin{aligned}\sum_{t=1}^{\infty} \mathbb{E}\left[\mathbb{E}\left[u_{t+1} - u_t | \mathcal{I}_t\right]^+\right] &\leq \sum_{t=1}^{\infty} \frac{\mathbb{E}\left[\|g_t - g\|_\infty\right]}{t+1} \\ &< \sum_{t=1}^{\infty} \frac{c_3}{t^{\frac{3}{2}}} = O(1) \ ,\end{aligned}$$

where a superscripted + takes on value 0 for negative numbers and the value of the number otherwise. Therefore, the sum of the positive variations of u_T is bounded. By applying a theorem due to (Fisk, 1965) this implies that u_T is a quasi-martingale and converges almost surely. This proves the first part of Proposition 2.

Next, we show that u_T being a quasi-martingale implies the almost sure convergence of the fourth line of Equation 17. To see this we note that since u_T is a quasi-martingale and the sum of its negative variations is bounded, and since the term on the fourth line of Equation 17, $\frac{g_T(\mathbf{L}_T) - \hat{g}_T(\mathbf{L}_T)}{T+1}$, is guaranteed to be negative, the sum of that term from 1 to infinity must be bounded:

$$\sum_{t=1}^{\infty} \frac{\hat{g}_t(\mathbf{L}_t) - g_t(\mathbf{L}_t)}{t+1} < \infty \ . \tag{19}$$

To complete the proof of Part (B) of Proposition 2, consider the following lemma: Let a_n, b_n be two real sequences such that for all n, $a_n \geq 0, b_n \geq 0, \sum_{n=1}^{\infty} a_n = \infty, \sum_{n=1}^{\infty} a_n b_n < \infty, \exists K > 0$ s.t.$|b_{n+1} - b_n| < K a_n$. Then, $\lim_{n\to+\infty} b_n = 0$.

If we define $a_t = \frac{1}{t+1}$ and $b_t = \hat{g}_t(\mathbf{L}_t) - g_t(\mathbf{L}_t)$, then clearly these are both positive sequences, and $\sum_{t=1}^{\infty} a_t = \infty$. We just showed that $\sum_{t=1}^{\infty} \frac{\hat{g}_t(\mathbf{L}_t) - g_t(\mathbf{L}_t)}{t+1} < \infty$ which is equivalent to $\sum_{t=1}^{\infty} a_n b_n < \infty$. Since g_T and $\hat{g}_T$ are bounded and Lipschitz with constant independent of T, and $\mathbf{L}_{T+1} - \mathbf{L}_T = O\left(\frac{1}{T}\right)$ we have all of the assumptions verified, which implies that:

$$\lim_{T\to\infty} \hat{g}_T(\mathbf{L}_T) - g_T(\mathbf{L}_T) \to 0, \text{ a.s.}$$

Now we have established Part (B) of this proposition that $g_T(\mathbf{L}_T)$ and $\hat{g}_T(\mathbf{L}_T)$ converge almost surely to the same limit. Additionally, by the Glivenko-Cantelli theorem we have that $\lim_{T\to\infty} \|g - g_T\|_\infty = 0$, which implies that g must converge almost surely. By transitivity, $\lim_{T\to\infty} \hat{g}_T(\mathbf{L}_T) - g(\mathbf{L}_T) = 0$. We have now shown Parts (C) and (D) of Proposition 2. ■

Proposition 3: The distance between $\mathbf{L}_T$ and the set of all stationary points of g converges a.s. to 0 as $t \to \infty$.

Proof: Before proceeding, we show that $\nabla_{\mathbf{L}} \hat{g}_T$ is Lipschitz with constant independent of T. Since $\hat{g}_T$ is quadratic its gradient is linear which implies that it is Lipschitz. Additionally, since $\mathbf{s}^{(t)}$, $\mathbf{D}^{(t)}$, and $\boldsymbol{\theta}^{(t)}$ are all bounded and the summation over task losses is normalized by T, it follows that $\hat{g}_T$ has a Lipschitz constant independent of T.

Next, we define an arbitrary non-zero matrix $\mathbf{U}$ of the same dimensionality as $\mathbf{L}$. Since $\hat{g}_T$ upper-bounds g_T, we have:

$$\begin{aligned}\hat{g}_T(\mathbf{L}_T + \mathbf{U}) &\geq g_T(\mathbf{L}_T + \mathbf{U}) \\ \lim_{T\to\infty} \hat{g}_T(\mathbf{L}_T + \mathbf{U}) &\geq \lim_{T\to\infty} g(\mathbf{L}_T + \mathbf{U}) \ ,\end{aligned}$$

where to get the second inequality we took the limit of both sides and replaced g_T with g (which are equivalent as $T \to \infty$). Let $h_T > 0$ be a sequence of positive real numbers that converges to 0. If we take the first-order Taylor expansion on both sides of the inequality and use the fact that ∇g and $\nabla \hat{g}$ are both Lipschitz with constant independent of T, we get:

$$\lim_{T\to\infty} \left\{\hat{g}_T(\mathbf{L}_T) + \mathrm{Tr}\left(h_T \mathbf{U}^\top \nabla \hat{g}_T(\mathbf{L}_T)\right) + O(h_T \mathbf{U})\right\} \geq \lim_{T\to\infty} \left\{g(\mathbf{L}_T) + \mathrm{Tr}\left(h_T \mathbf{U}^\top \nabla g(\mathbf{L}_T)\right) + O(h_T \mathbf{U})\right\} \ .$$

Since $\lim_{T\to\infty} \hat{g}_T(\mathbf{L}_T) - g(\mathbf{L}_T) = 0$ a.s. and $\lim_{T\to\infty} h_T = 0$, we have:

$$\lim_{T\to\infty} \mathrm{Tr}\left(\frac{1}{\|\mathbf{U}\|_{\mathsf{F}}} \mathbf{U}^\top \nabla \hat{g}_T(\mathbf{L}_T)\right) \geq \lim_{T\to\infty} \mathrm{Tr}\left(\frac{1}{\|\mathbf{U}\|_{\mathsf{F}}} \mathbf{U}^\top \nabla g(\mathbf{L}_T)\right) \ .$$

Since this inequality has to hold for every $\mathbf{U}$, we require that $\lim_{T\to\infty} \nabla \hat{g}_T(\mathbf{L}_T) = \lim_{T\to\infty} \nabla g(\mathbf{L}_T)$. Since $\mathbf{L}_T$ minimizes $\hat{g}_T$, we require that $\nabla \hat{g}_T(\mathbf{L}_T) = \mathbf{0}$, where $\mathbf{0}$ is the zero-vector of appropriate dimensionality. This implies that $\nabla g(\mathbf{L}_T) = \mathbf{0}$, which is a first-order condition for $\mathbf{L}_T$ to be a stationary point of g. ■