Statistical Limits of Generalization Part II: Linear Realizability

Sham Kakade and Wen Sun

CS 6789: Foundations of Reinforcement Learning

Announcements

- HW1 will be posted tonight
  - Tentatively: due Sept 28 (check ED posting)
- Chapters 1-5 substantially updated.
  - Feedback/questions/finding typos appreciated!

Today:

Today:

- Recap: lower bounds for generalization
  - agnostic learning is not possible in RL
    (unless we pay an exponential in H dependence)
  - Linear realizability of $Q^\star$ is still statistically hard

Today:

- Recap: lower bounds for generalization
  - agnostic learning is not possible in RL
    (unless we pay an exponential in H dependence)
  - Linear realizability of $Q^\star$ is still statistically hard

- Today:
  - Finish up lower bound when we know $Q^\star$ is linearly realizable.
  - New: linear Bellman completeness.

Part-2: Linear Realizability

What if we impose linearity assumptions?

Let’s look at the most natural assumptions.

RL with Linearly Realizable Q*-Function Approximation

(Does there exist a sample efficient algo?)

RL with Linearly Realizable Q*-Function Approximation

(Does there exist a sample efficient algo?)

- Suppose we have a feature map: $\overrightarrow{\phi}(s, a) \in R^d$.

RL with Linearly Realizable Q*-Function Approximation
(Does there exist a sample efficient algo?)

- Suppose we have a feature map: $\overrightarrow{\phi}(s, a) \in R^d$.
- (A1: Linearly Realizable Q*): Assume for all $s, a, h \in [H]$, there exists $w_1^\star, \ldots w_H^\star \in R^d$ s.t.

$$Q_h^\star(s, a) = w_h^\star \cdot \phi(s, a)$$

RL with Linearly Realizable Q*-Function Approximation

(Does there exist a sample efficient algo?)

- Suppose we have a feature map: $\overrightarrow{\phi}(s, a) \in R^d$.
- (A1: Linearly Realizable Q*): Assume for all $s, a, h \in [H]$, there exists $w_1^\star, \ldots w_H^\star \in R^d$ s.t.

$$Q_h^\star(s, a) = w_h^\star \cdot \phi(s, a)$$

- (A2: Large Suboptimality Gap): for all $a \neq \pi^\star(s)$,

$$V_h^\star(s) - Q_h^\star(s, a) \geq \text{constant}$$

Linearly Realizability is Not Sufficient for RL

Linearly Realizability is Not Sufficient for RL

Theorem:

Linearly Realizability is Not Sufficient for RL

Theorem:

- [Weisz, Amortila, Szepesvári '21]:
  There exists an MDP and a ϕ satisfying A1 s.t any online RL algorithm (with knowledge of ϕ) requires $\Omega(\min(2^d, 2^H))$ samples to output the value $V^\star(s_0)$ up to constant additive error (with prob. ≥ 0.9).

Linearly Realizability is Not Sufficient for RL

Theorem:

- [Weisz, Amortila, Szepesvári '21]:
  There exists an MDP and a ϕ satisfying A1 s.t any online RL algorithm (with knowledge of ϕ) requires $\Omega(\min(2^d, 2^H))$ samples to output the value $V^\star(s_0)$ up to constant additive error (with prob. ≥ 0.9).
- [Wang, Wang, K. '21]:
  Let's make the problem even easier, where we also assume A2 (large gap)
  The lower bound holds even with **both** A1 and A2.

Linearly Realizability is Not Sufficient for RL

Theorem:

- [Weisz, Amortila, Szepesvári '21]:
  There exists an MDP and a ϕ satisfying A1 s.t any online RL algorithm (with knowledge of ϕ) requires $\Omega(\min(2^d, 2^H))$ samples to output the value $V^\star(s_0)$ up to constant additive error (with prob. ≥ 0.9).
- [Wang, Wang, K. '21]:
  Let's make the problem even easier, where we also assume A2 (large gap)
  The lower bound holds even with **both** A1 and A2.

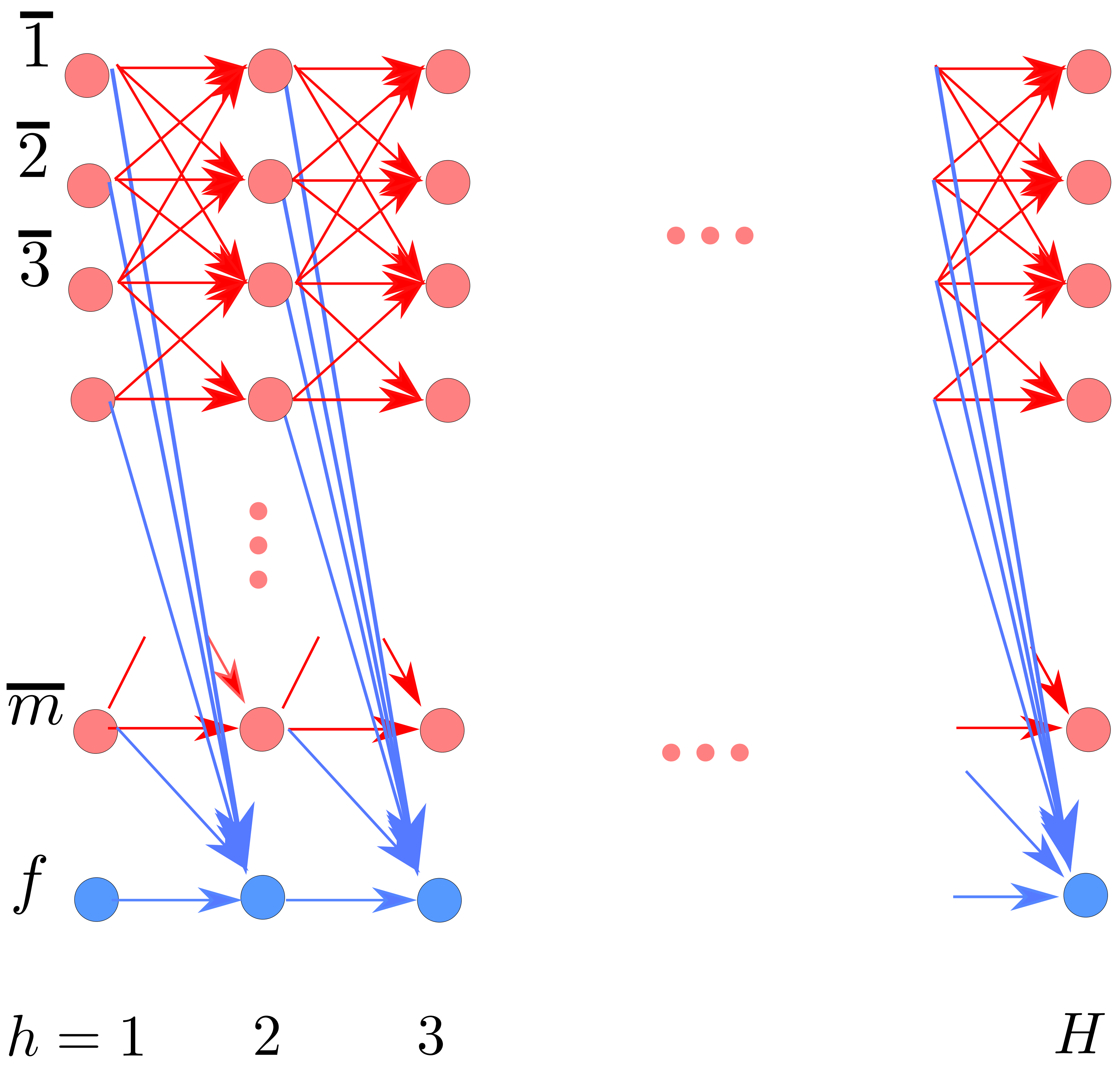
Comments: An exponential separation between online RL vs simulation access.

[Du, K., Wang, Yang '20]: A1+A2+simulator access (input: any s, a; output: $s' \sim P(\cdot \mid s, a), r(s, a)$)

$\Longrightarrow$ there is sample efficient approach to find an ϵ-opt policy.

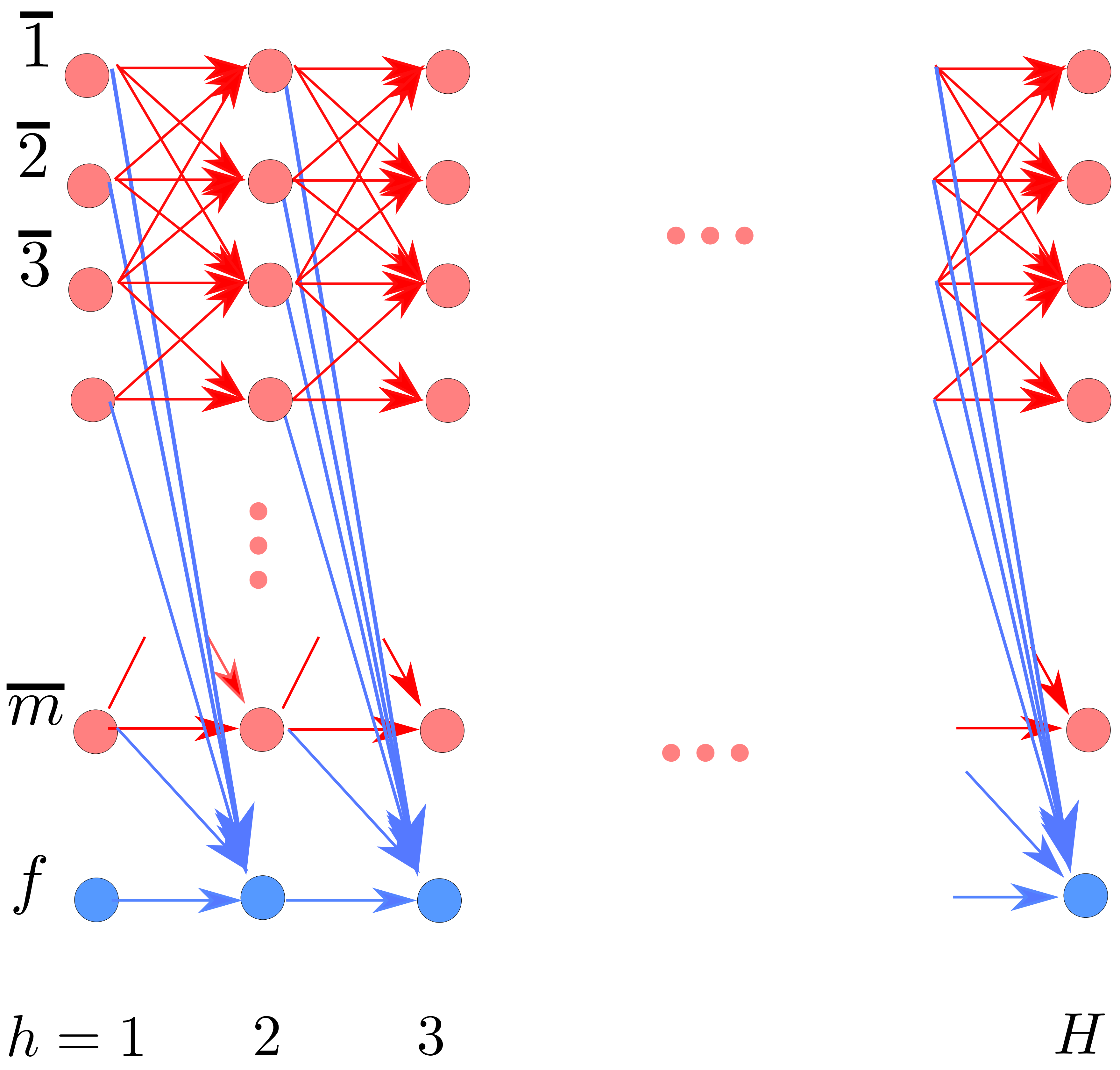
Construction Sketch: a Hard MDP Family

(A ``leaking complete graph'')

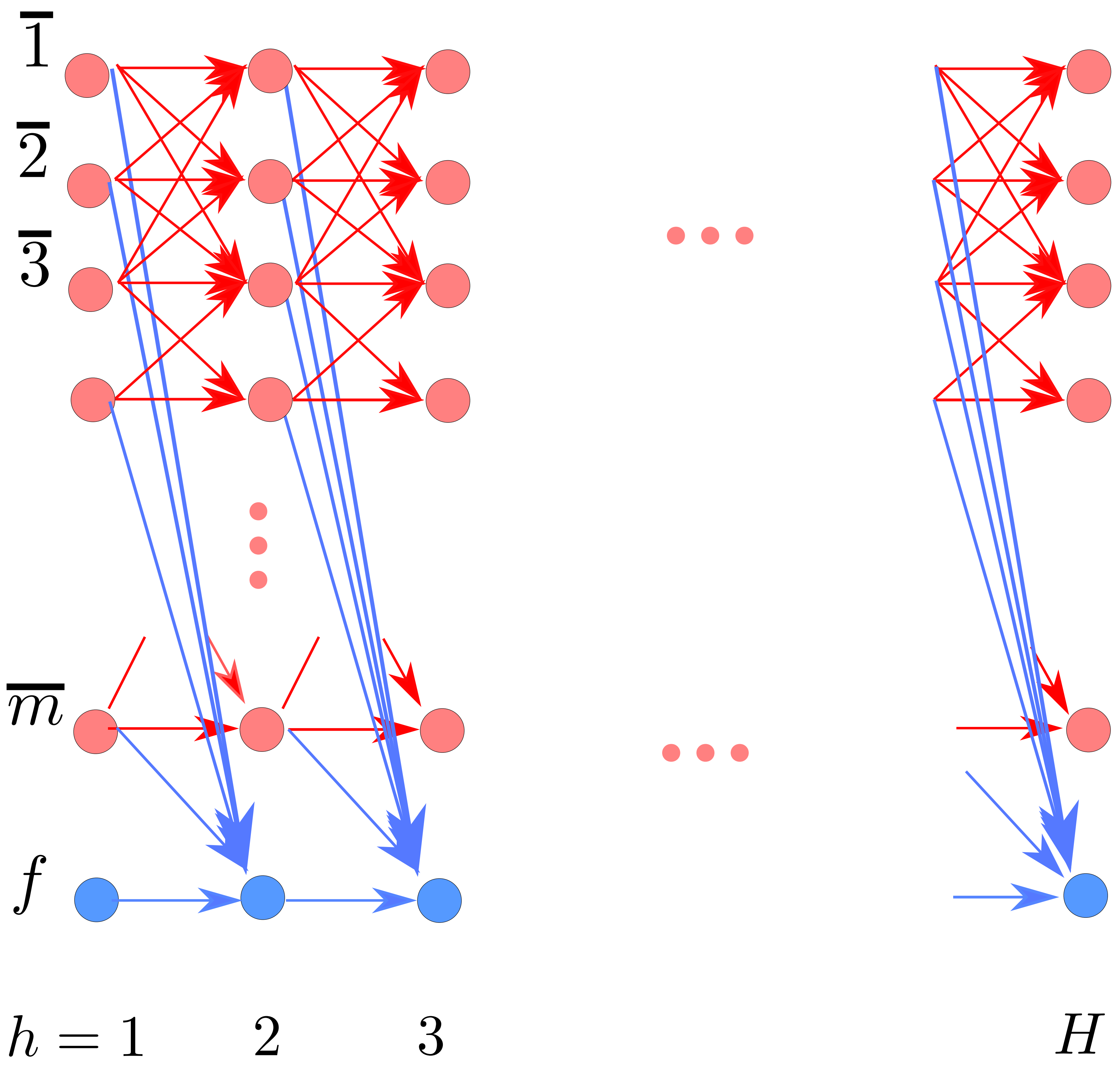


Construction Sketch: a Hard MDP Family

(A ``leaking complete graph'')



- m is an integer (we will set $m \approx 2^d$)



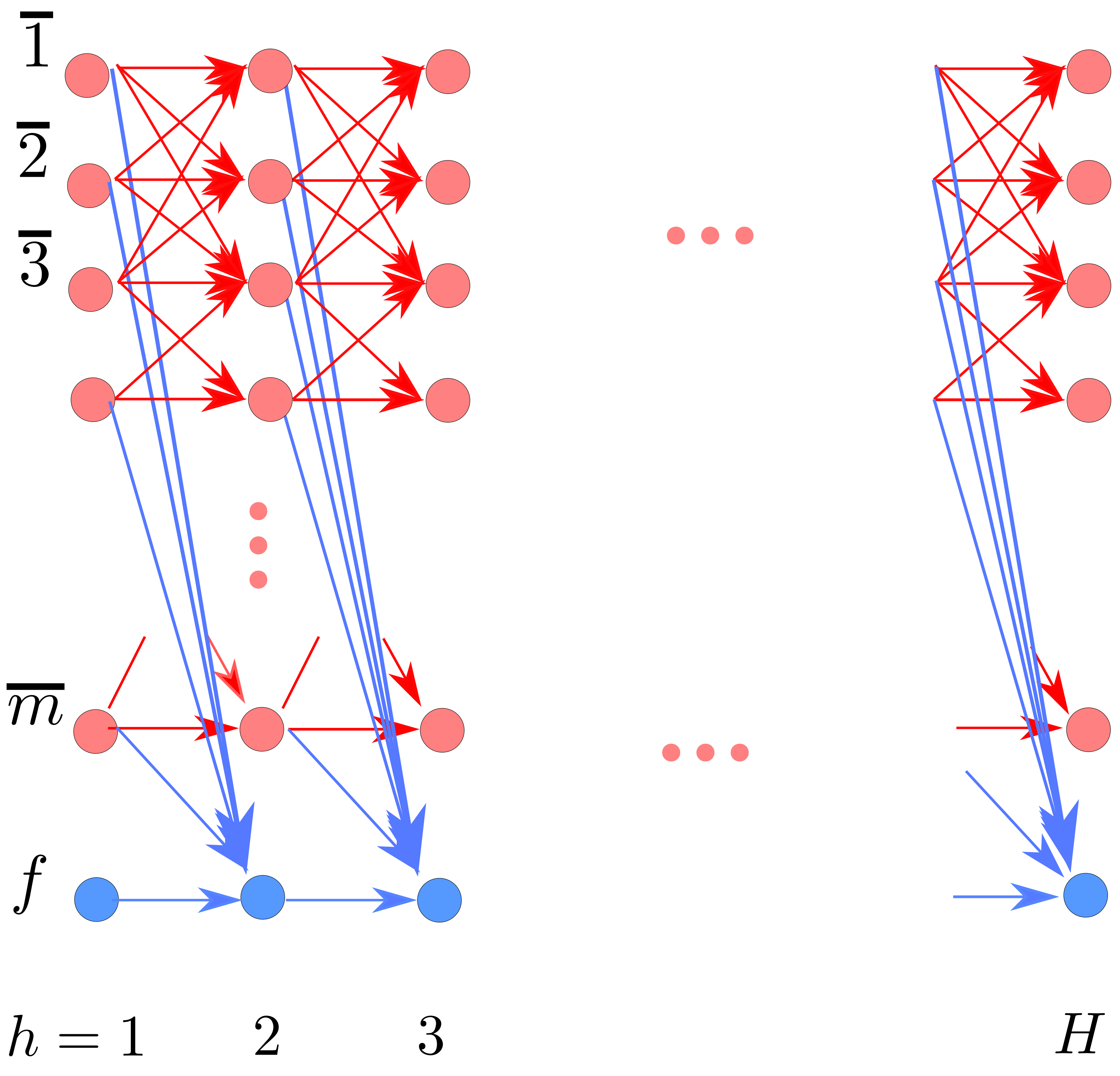
Construction Sketch: a Hard MDP Family

(A ``leaking complete graph'')

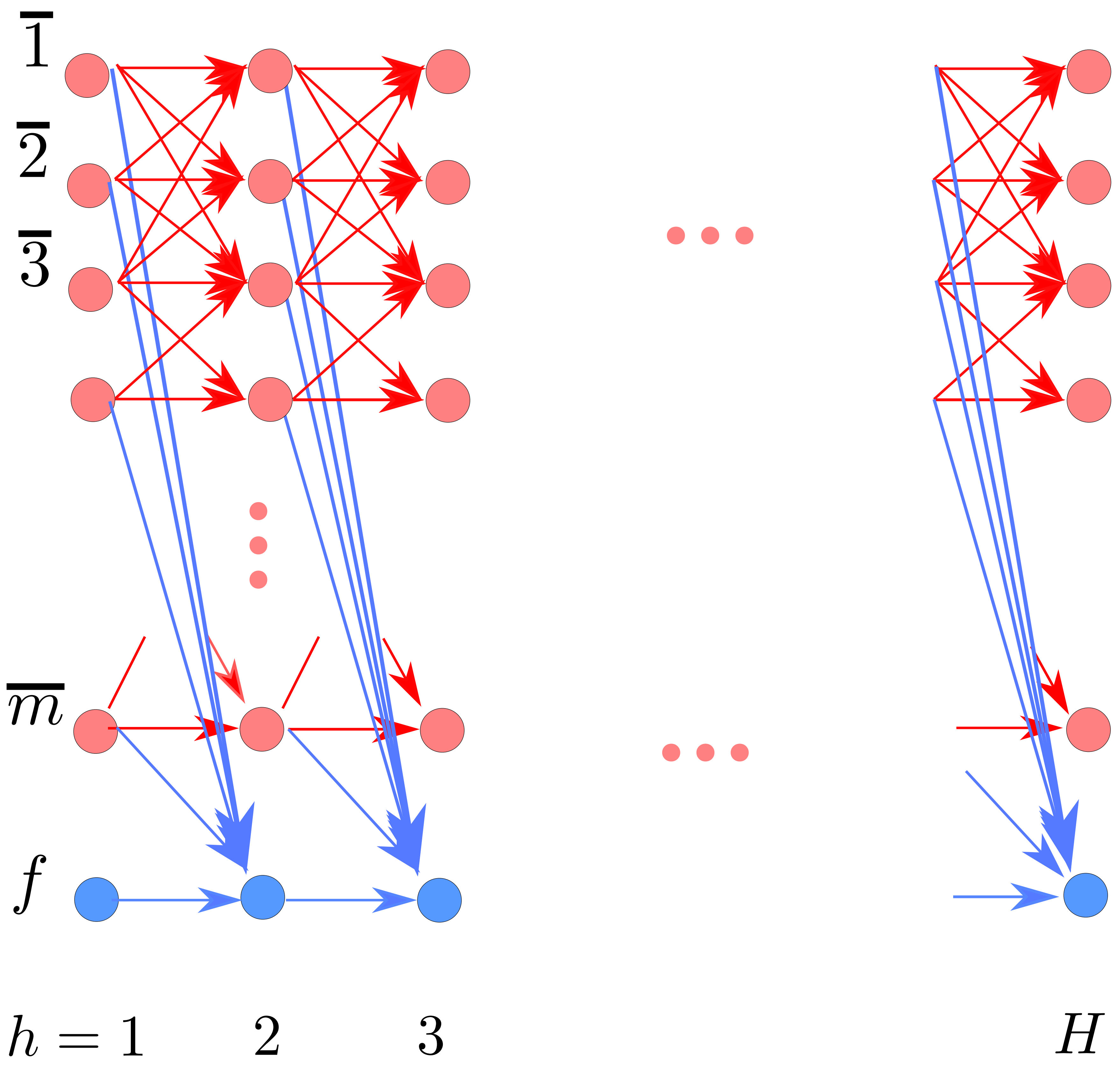
- m is an integer (we will set $m \approx 2^d$)
- the state space: $\{\bar{1}, \cdots, \bar{m}, f\}$

Construction Sketch: a Hard MDP Family

(A ``leaking complete graph'')



- m is an integer (we will set $m \approx 2^d$)
- the state space: $\{\bar{1}, \cdots, \bar{m}, f\}$
- call the special state f a "terminal state".



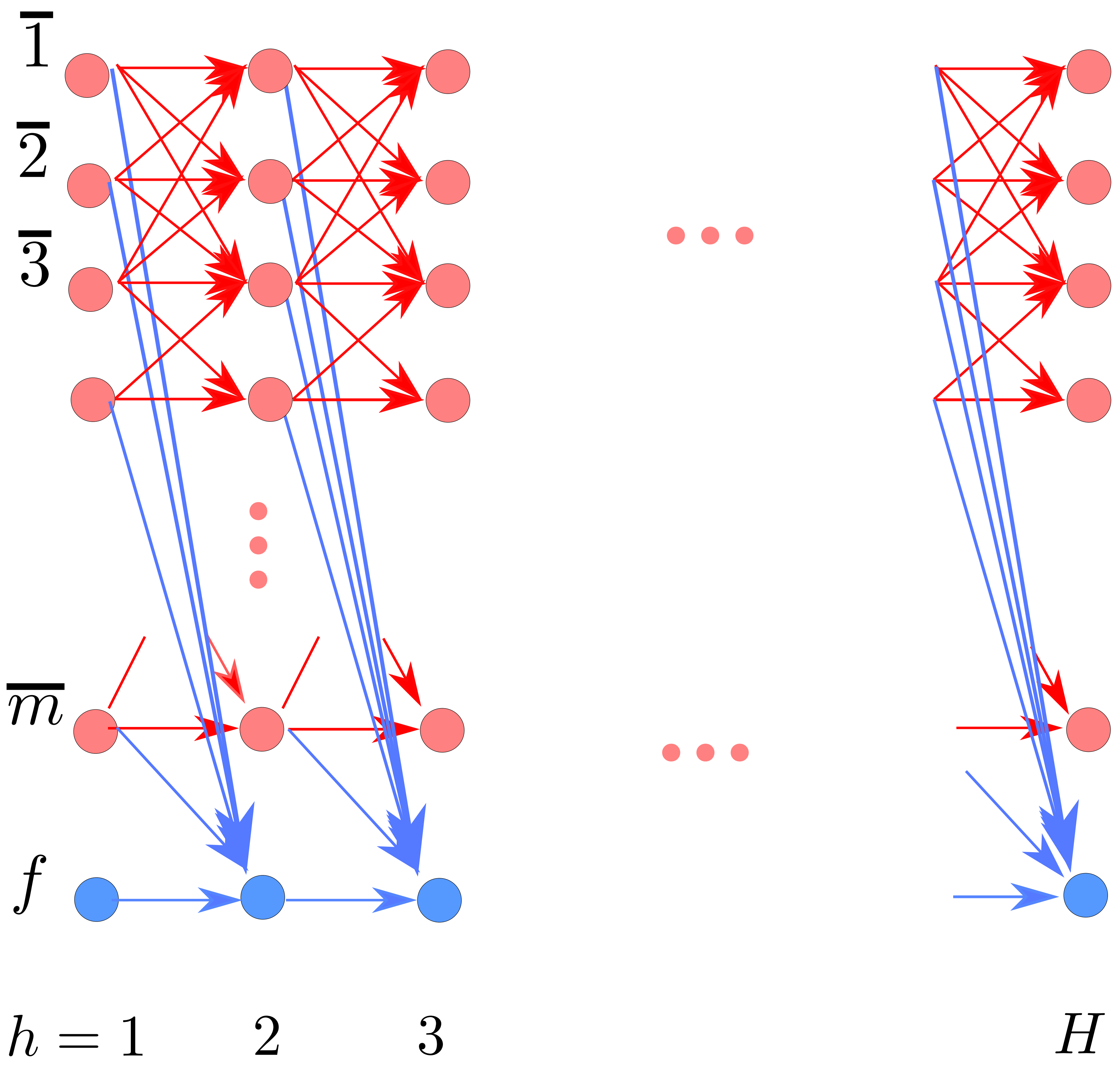
Construction Sketch: a Hard MDP Family

(A ``leaking complete graph'')

- m is an integer (we will set $m \approx 2^d$)
- the state space: $\{\bar{1}, \cdots, \bar{m}, f\}$
- call the special state f a "terminal state".
- at state $\bar{i}$, the feasible actions set is $[m] \backslash \{i\}$

  at f, the feasible action set is $[m-1]$.

  i.e. there are $m-1$ feasible actions at each state.



Construction Sketch: a Hard MDP Family

(A ``leaking complete graph'')

- m is an integer (we will set $m \approx 2^d$)
- the state space: $\{\bar{1}, \cdots, \bar{m}, f\}$
- call the special state f a "terminal state".
- at state $\bar{i}$, the feasible actions set is $[m] \backslash \{i\}$

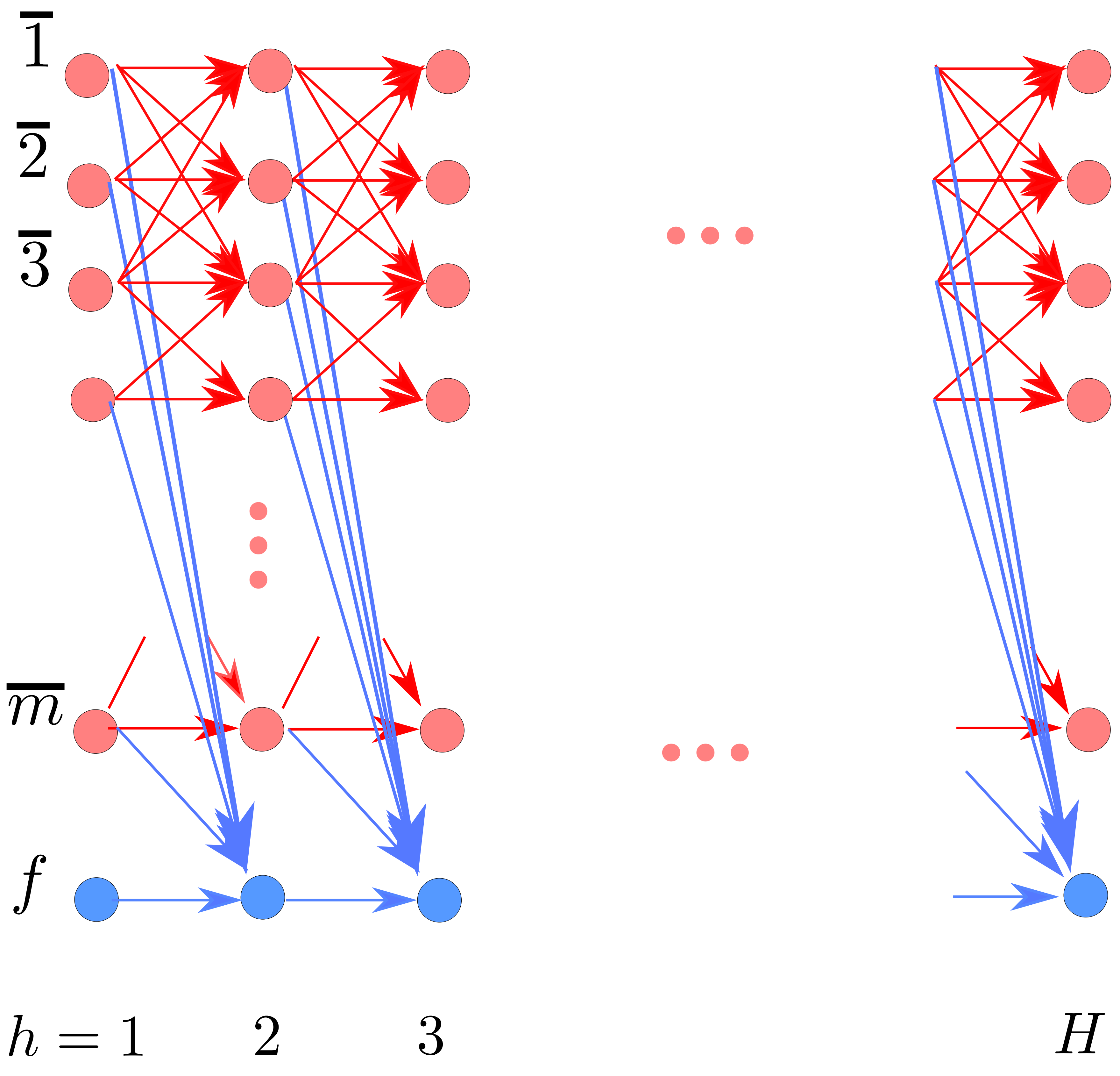
  at f, the feasible action set is $[m-1]$.

  i.e. there are $m-1$ feasible actions at each state.
- each MDP in this family is specified by an index $a^* \in [m]$ and denoted by $\mathscr{M}_{a^*}$.

  i.e. there are m MDPs in this family.

Construction Sketch: a Hard MDP Family

(A ``leaking complete graph'')



- m is an integer (we will set $m \approx 2^d$)
- the state space: $\{\bar{1}, \cdots, \bar{m}, f\}$
- call the special state f a "terminal state".
- at state $\bar{i}$, the feasible actions set is $[m] \backslash \{i\}$

  at f, the feasible action set is $[m-1]$.

  i.e. there are $m-1$ feasible actions at each state.
- each MDP in this family is specified by an index $a^* \in [m]$ and denoted by $\mathscr{M}_{a^*}$.

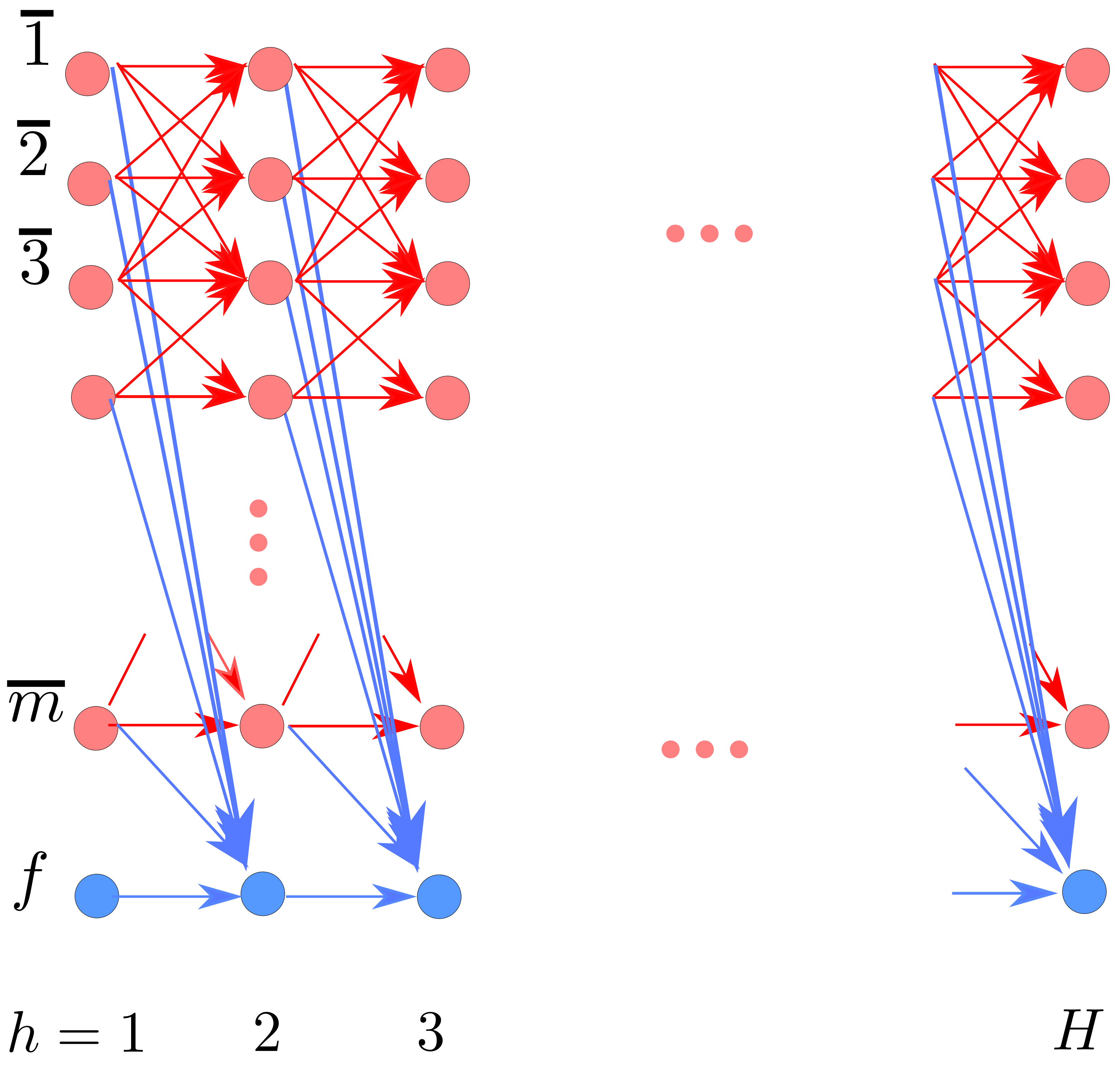
  i.e. there are m MDPs in this family.

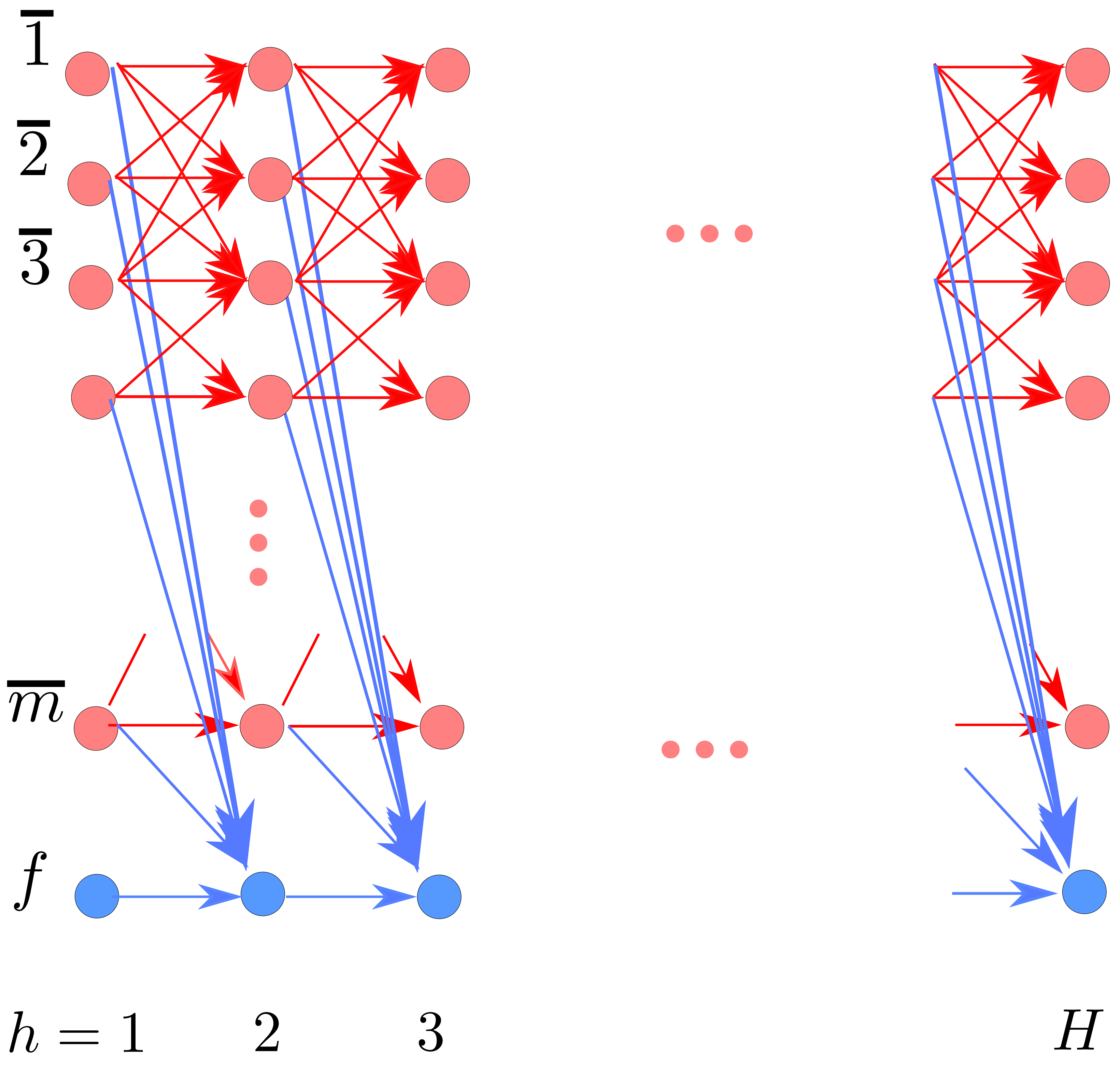
Lemma: For any $\gamma > 0$, there exist $m = \lfloor \exp(\frac{1}{8}\gamma^2 d) \rfloor$ unit vectors $\{v_1, \cdots, v_m\}$ in R^d s.t. $\forall i, j \in [m]$ and $i \neq j$, $|\langle v_i, v_j \rangle| \leq \gamma$.

We will set $\gamma = 1/4$.

(proof: Johnson-Lindenstrauss)

The construction, continued





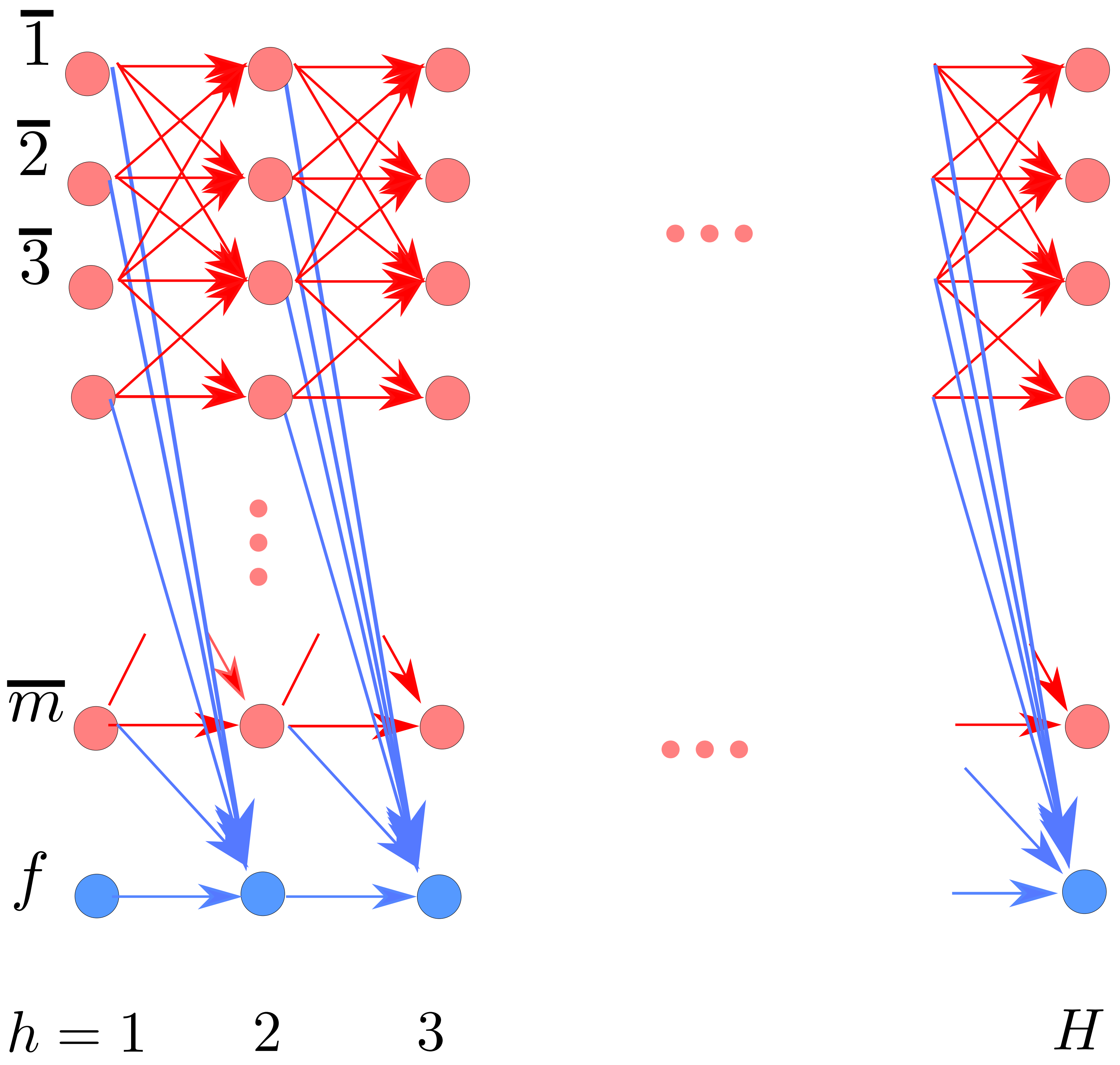
The construction, continued

- Transitions: $s_0 \sim$ Uniform($[m]$).

  $\Pr[f \,|\, \overline{a_1}, a^*] = 1$

  $$\Pr[\,\cdot\,|\,\overline{a_1}, a_2] = \begin{cases} \overline{a_2} : \left\langle v(a_1), v(a_2) \right\rangle + 2\gamma \\ f : 1 - \left\langle v(a_1), v(a_2) \right\rangle - 2\gamma \end{cases}, (a_2 \neq a^*, a_2 \neq a_1)$$

  $\Pr[f \,|\, f, \,\cdot\,] = 1.$



The construction, continued

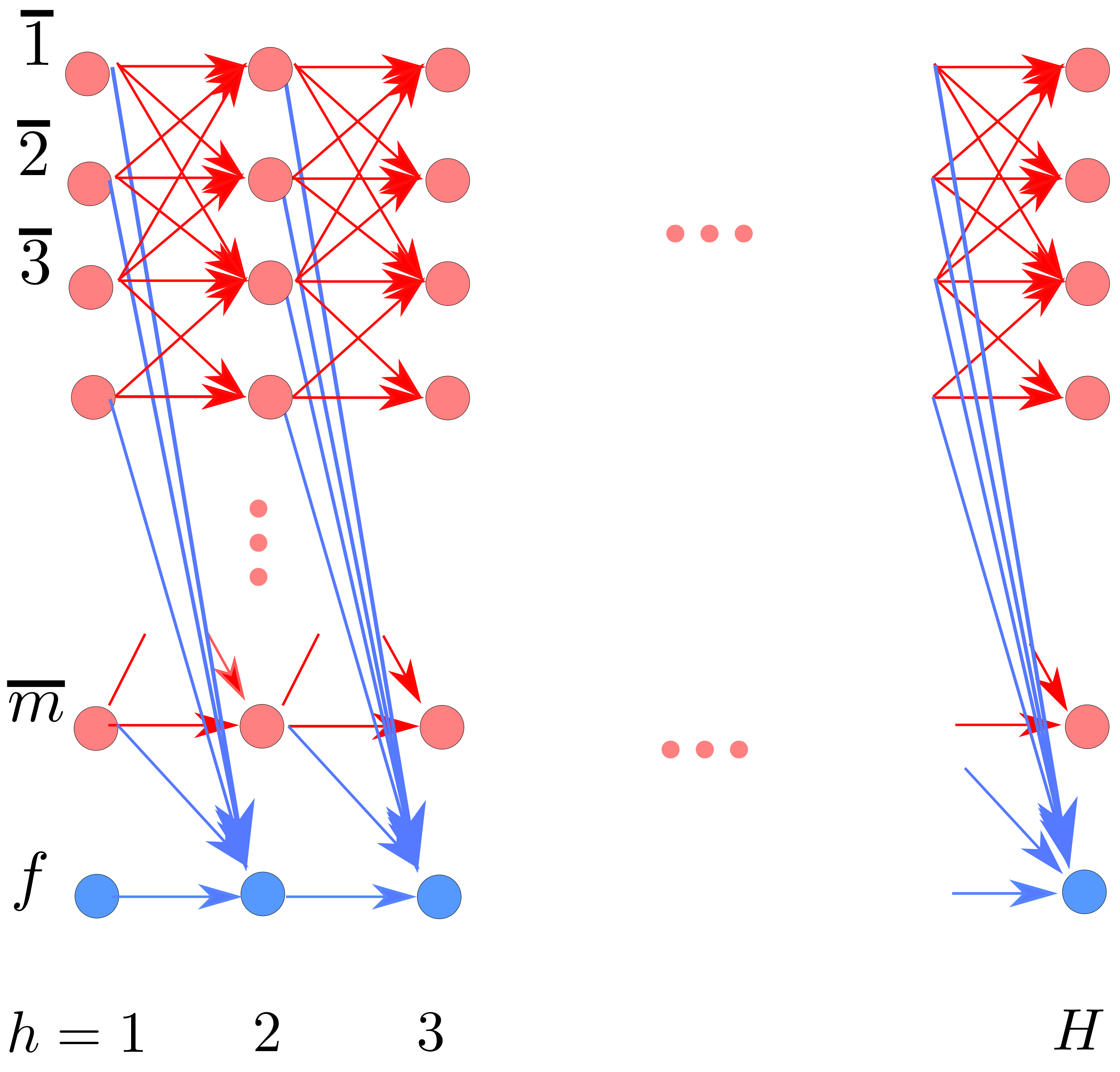
- Transitions: $s_0 \sim$ Uniform($[m]$).

  $\Pr[f \,|\, \overline{a_1}, a^*] = 1$

  $$\Pr[\,\cdot\,|\,\overline{a_1}, a_2] = \begin{cases} \overline{a_2} : \left\langle v(a_1), v(a_2) \right\rangle + 2\gamma \\ f : 1 - \left\langle v(a_1), v(a_2) \right\rangle - 2\gamma \end{cases}, (a_2 \neq a^*, a_2 \neq a_1)$$

  $$\Pr[f \,|\, f, \cdot\,] = 1.$$

- After taking action a_2, the next state is either $\overline{a_2}$ or f. This MDP looks like a ``leaking complete graph''



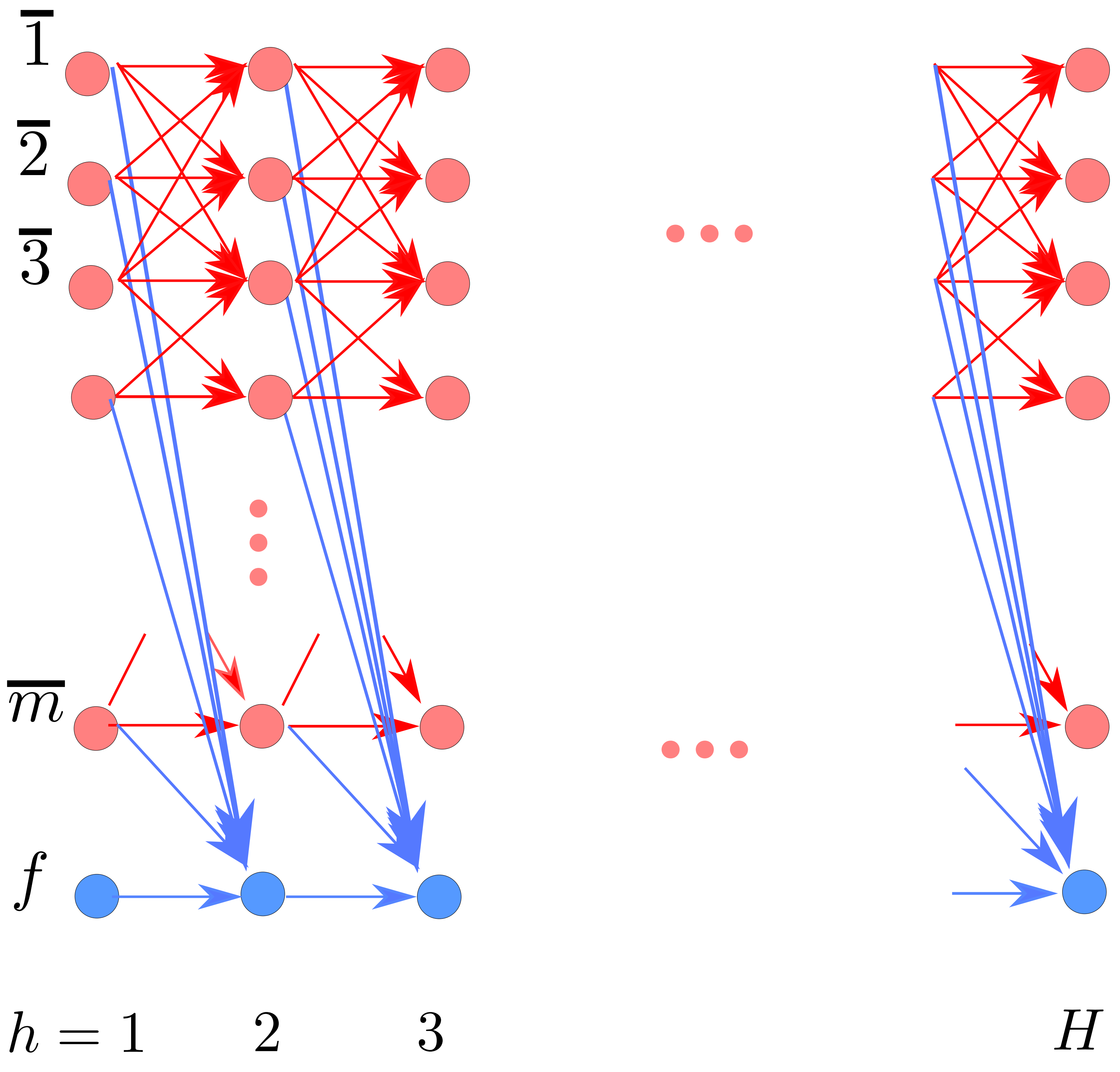
The construction, continued

- Transitions: $s_0 \sim$ Uniform($[m]$).

  $\Pr[f \,|\, \overline{a_1}, a^*] = 1$

  $$\Pr[\,\cdot\,|\,\overline{a_1}, a_2] = \begin{cases} \overline{a_2} : \left\langle v(a_1), v(a_2) \right\rangle + 2\gamma \\ f : 1 - \left\langle v(a_1), v(a_2) \right\rangle - 2\gamma \end{cases}, (a_2 \neq a^*, a_2 \neq a_1)$$

  $\Pr[f \,|\, f, \cdot\,] = 1.$
- After taking action a_2, the next state is either $\overline{a_2}$ or f. This MDP looks like a ``leaking complete graph''
- It is possible to visit any other state (except for $\overline{a^*}$); however, there is at least $1 - 3\gamma = 1/4$ probability of going to the terminal state f.



The construction, continued

- Transitions: $s_0 \sim$ Uniform($[m]$).

  $\Pr[f \mid \overline{a_1}, a^*] = 1$

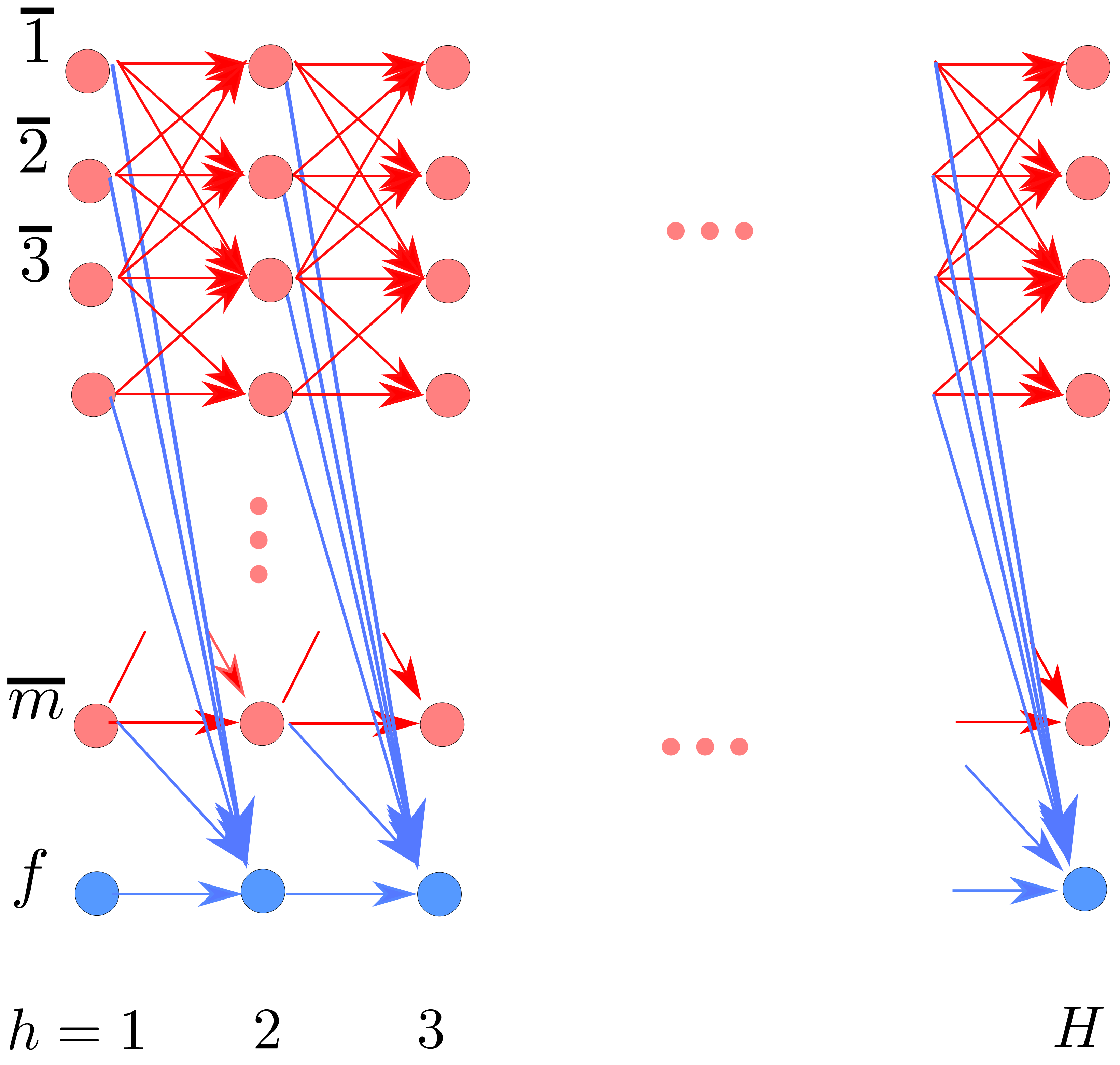
  $$\Pr[\,\cdot \mid \overline{a_1}, a_2] = \begin{cases} \overline{a_2} : \left\langle v(a_1), v(a_2) \right\rangle + 2\gamma \\ f : 1 - \left\langle v(a_1), v(a_2) \right\rangle - 2\gamma \end{cases}, (a_2 \neq a^*, a_2 \neq a_1)$$

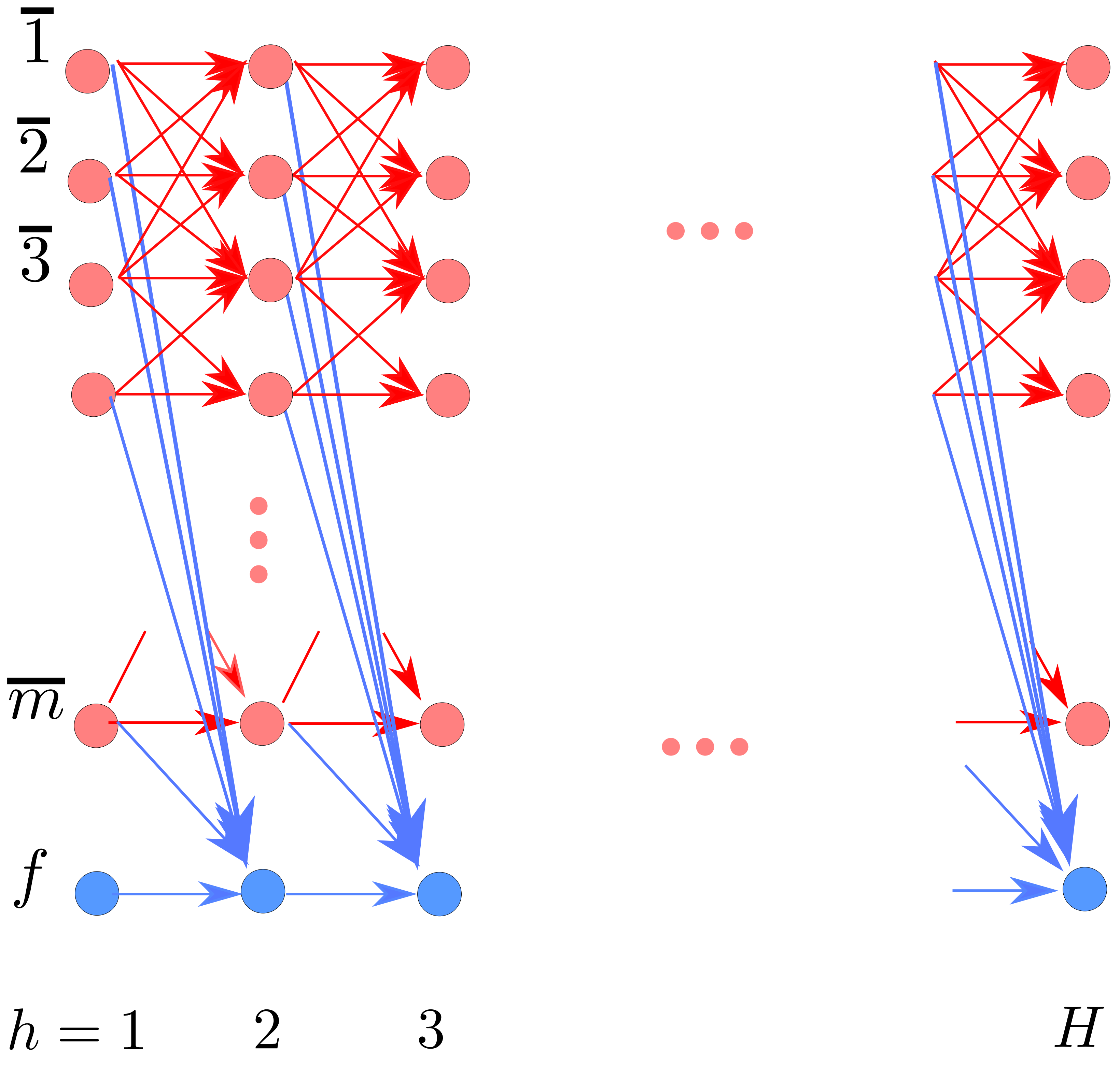
  $$\Pr[f \mid f, \cdot\,] = 1.$$

- After taking action a_2, the next state is either $\overline{a_2}$ or f. This MDP looks like a "leaking complete graph"
- It is possible to visit any other state (except for $\overline{a^*}$); however, there is at least $1 - 3\gamma = 1/4$ probability of going to the terminal state f.
- The transition probabilities are indeed valid, because

  $$0 < \gamma \leq \left\langle v(a_1), v(a_2) \right\rangle + 2\gamma \leq 3\gamma < 1.$$

The construction, continued





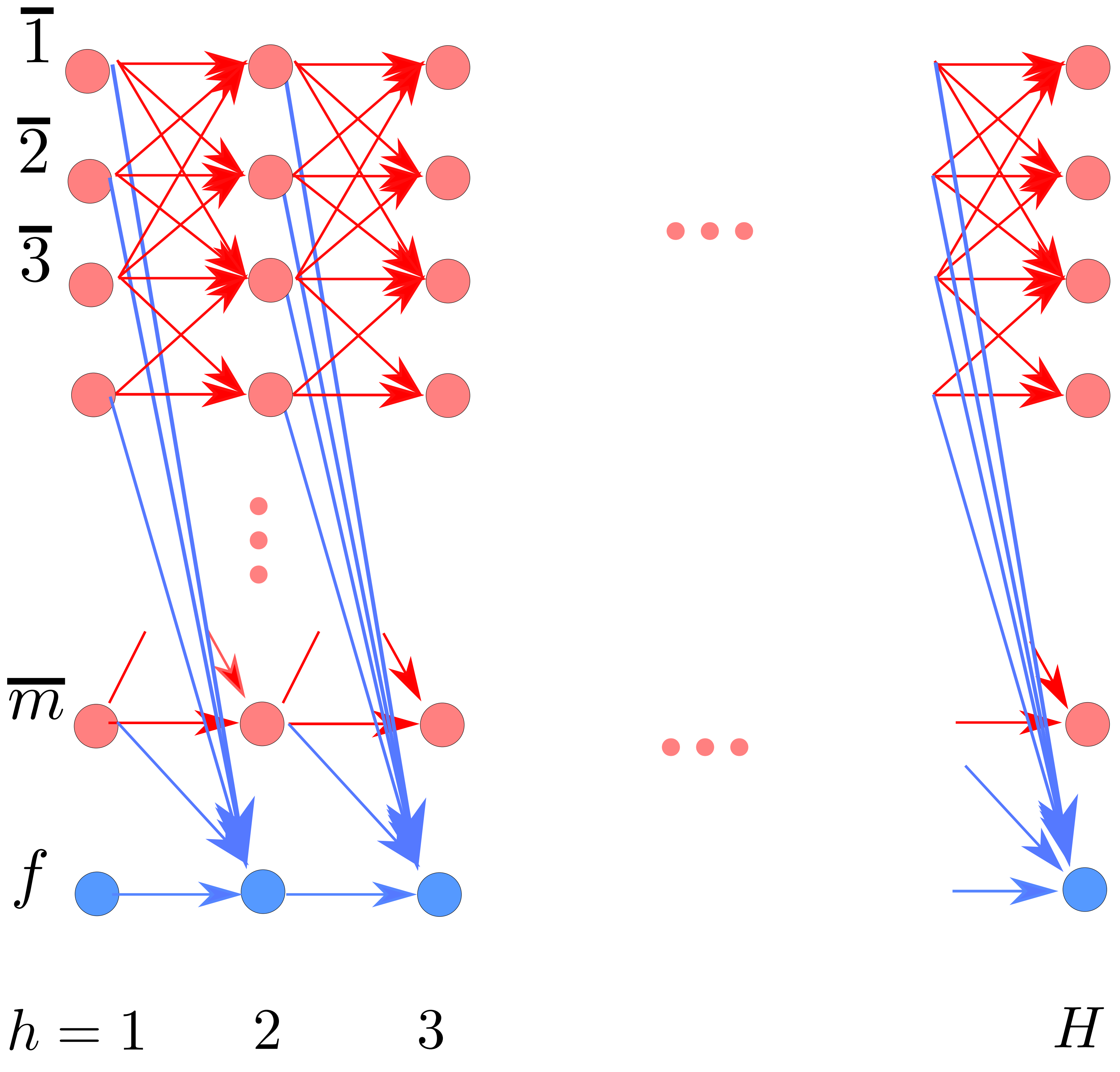
The construction, continued

- Features: of dimension d defined as:

$$\phi(\overline{a_1}, a_2) := \left(\left\langle v(a_1), v(a_2) \right\rangle + 2\gamma \right) \cdot v(a_2), \quad \forall a_1 \neq a_2$$

$$\phi(f, \cdot\,) := \mathbf{0}$$

note: the feature map does not depend of a^*.



The construction, continued

- Features: of dimension d defined as:

$$\phi(\overline{a_1}, a_2) := \left(\left\langle v(a_1), v(a_2) \right\rangle + 2\gamma \right) \cdot v(a_2), \quad \forall a_1 \neq a_2$$

$$\phi(f, \cdot) := \mathbf{0}$$

note: the feature map does not depend of a^*.

- Rewards:

for $1 \leq h < H$,

$$R_h(\overline{a_1}, a^*) := \left\langle v(a_1), v(a^*) \right\rangle + 2\gamma,$$

$$R_h(\overline{a_1}, a_2) := -2\gamma \left[\left\langle v(a_1), v(a_2) \right\rangle + 2\gamma \right], \, a_2 \neq a^*, a_2 \neq a_1$$

$$R_h(f, \cdot) \; := 0.$$

for $h = H$,

$$r_H(s, a) := \langle \phi(s, a), v(a^*) \rangle$$

Verifying the Assumptions: Realizability and the Large Gap

Verifying the Assumptions: Realizability and the Large Gap

Lemma: For all (s, a), we have $Q_h^*(s, a) = \langle \phi(s, a), v(a^*) \rangle$ and the "gap" is $\geq \gamma/4$.

Verifying the Assumptions: Realizability and the Large Gap

Lemma: For all (s, a), we have $Q_h^*(s, a) = \langle \phi(s, a), v(a^*) \rangle$ and the "gap" is $\geq \gamma/4$.

Proof: throughout $a_2 \neq a^*$

Verifying the Assumptions: Realizability and the Large Gap

Lemma: For all (s, a), we have $Q_h^*(s, a) = \langle \phi(s, a), v(a^*) \rangle$ and the "gap" is $\geq \gamma/4$.

Proof: throughout $a_2 \neq a^*$

- First, let's verify $Q^\pi(s, a) = \langle \phi(s, a), v(a^*) \rangle$ is the value of the policy $\pi(\overline{a}) = a^\star$. By induction, we can show:

$$Q_h^\pi(\overline{a_1}, a_2) = \left(\left\langle v(a_1), v(a_2) \right\rangle + 2\gamma \right) \cdot \left\langle v(a_2), v(a^*) \right\rangle,$$

$$Q_h^\pi(\overline{a_1}, a^*) = \left\langle v(a_1), v(a^*) \right\rangle + 2\gamma$$

Verifying the Assumptions: Realizability and the Large Gap

Lemma: For all (s, a), we have $Q_h^*(s, a) = \langle \phi(s, a), v(a^*) \rangle$ and the "gap" is $\geq \gamma/4$.

Proof: throughout $a_2 \neq a^*$

- First, let's verify $Q^\pi(s, a) = \langle \phi(s, a), v(a^*) \rangle$ is the value of the policy $\pi(\overline{a}) = a^\star$. By induction, we can show:

$$Q_h^\pi(\overline{a_1}, a_2) = \left(\left\langle v(a_1), v(a_2) \right\rangle + 2\gamma \right) \cdot \left\langle v(a_2), v(a^*) \right\rangle,$$

$$Q_h^\pi(\overline{a_1}, a^*) = \left\langle v(a_1), v(a^*) \right\rangle + 2\gamma$$

- Proving optimality: for $a_2 \neq a^*, a_1$

$$Q_h^\pi(\overline{a_1}, a_2) \leq 3\gamma^2, \quad Q_h^\pi(\overline{a_1}, a^*) = \left\langle v(a_1), v(a^*) \right\rangle + 2\gamma \geq \gamma > 3\gamma^2$$

$\implies$ π is optimal

Verifying the Assumptions: Realizability and the Large Gap

Lemma: For all (s, a), we have $Q_h^*(s, a) = \langle \phi(s, a), v(a^*) \rangle$ and the "gap" is $\geq \gamma/4$.

Proof: throughout $a_2 \neq a^*$

- First, let's verify $Q^\pi(s, a) = \langle \phi(s, a), v(a^*) \rangle$ is the value of the policy $\pi(\overline{a}) = a^\star$.
  By induction, we can show:

$$Q_h^\pi(\overline{a_1}, a_2) = \left(\left\langle v(a_1), v(a_2) \right\rangle + 2\gamma \right) \cdot \left\langle v(a_2), v(a^*) \right\rangle,$$

$$Q_h^\pi(\overline{a_1}, a^*) = \left\langle v(a_1), v(a^*) \right\rangle + 2\gamma$$

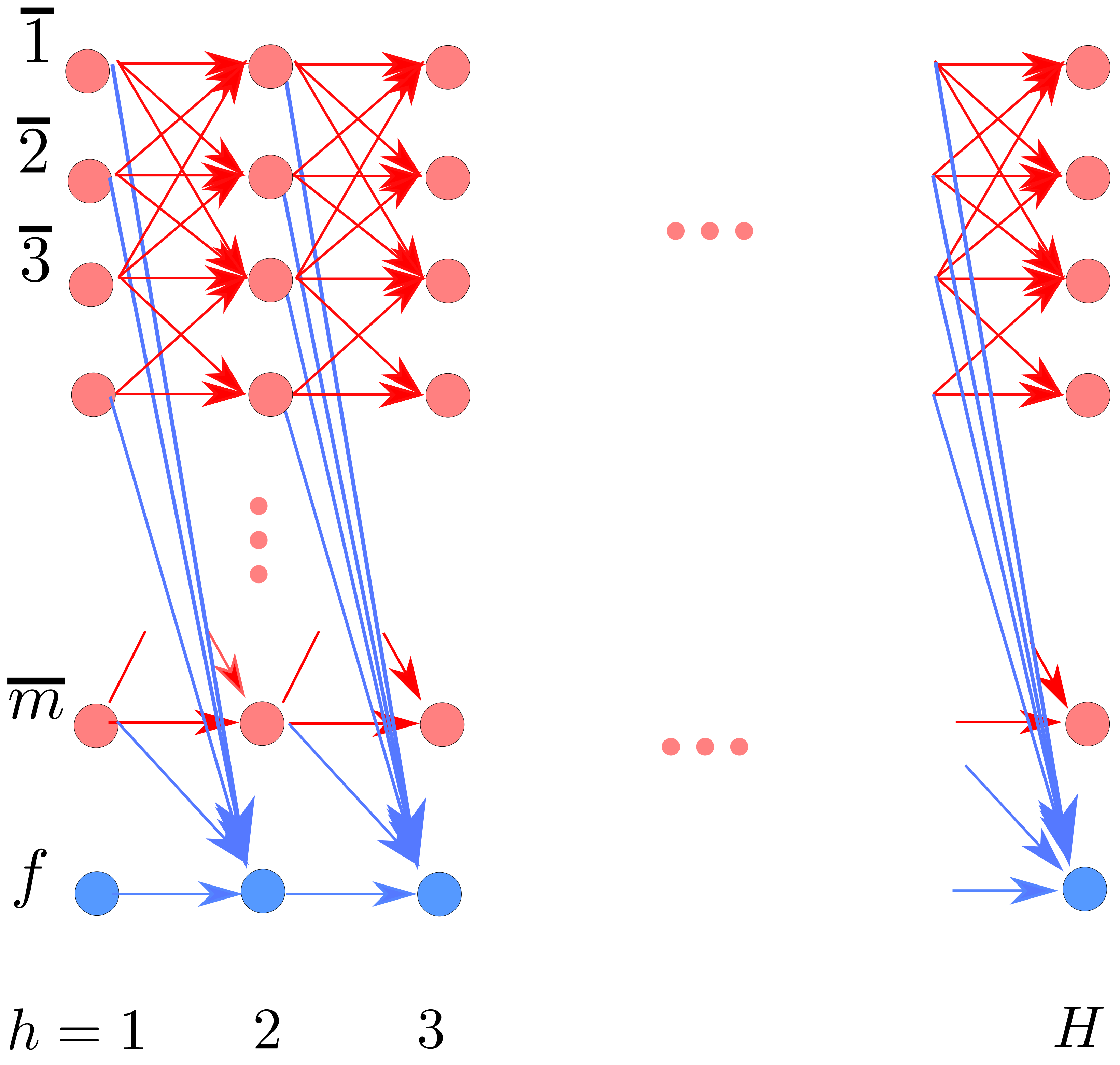
- Proving optimality: for $a_2 \neq a^*, a_1$

$$Q_h^\pi(\overline{a_1}, a_2) \leq 3\gamma^2, \quad Q_h^\pi(\overline{a_1}, a^*) = \left\langle v(a_1), v(a^*) \right\rangle + 2\gamma \geq \gamma > 3\gamma^2$$

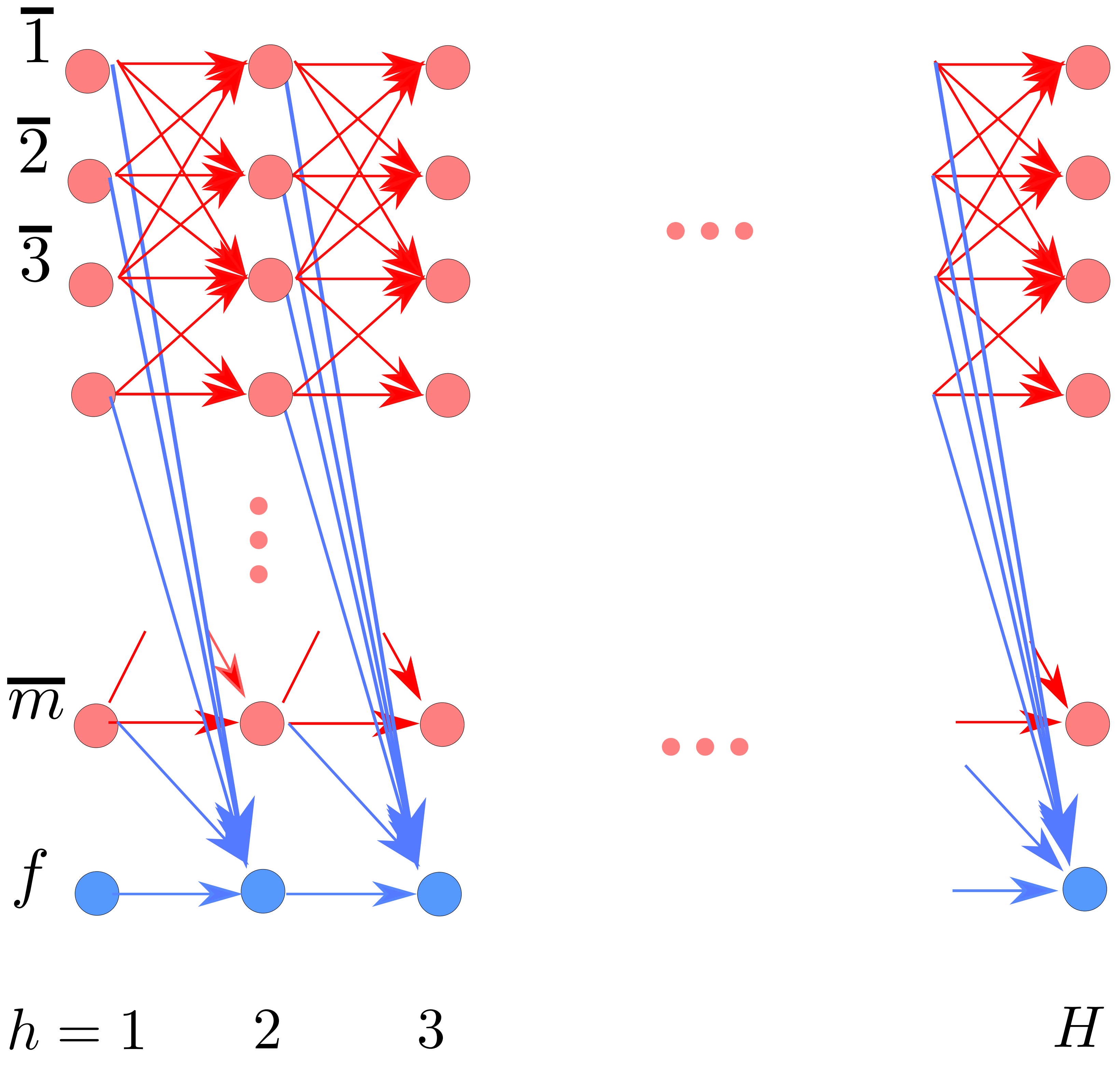
$\implies$ π is optimal

- Proving the large gap: for $a_2 \neq a^*$

$$V_h^*(\overline{a_1}) - Q_h^*(\overline{a_1}, a_2) = Q_h^\pi(\overline{a_1}, a^*) - Q_h^\pi(\overline{a_1}, a_2) > \gamma - 3\gamma^2 \geq \frac{1}{4}\gamma\,.$$

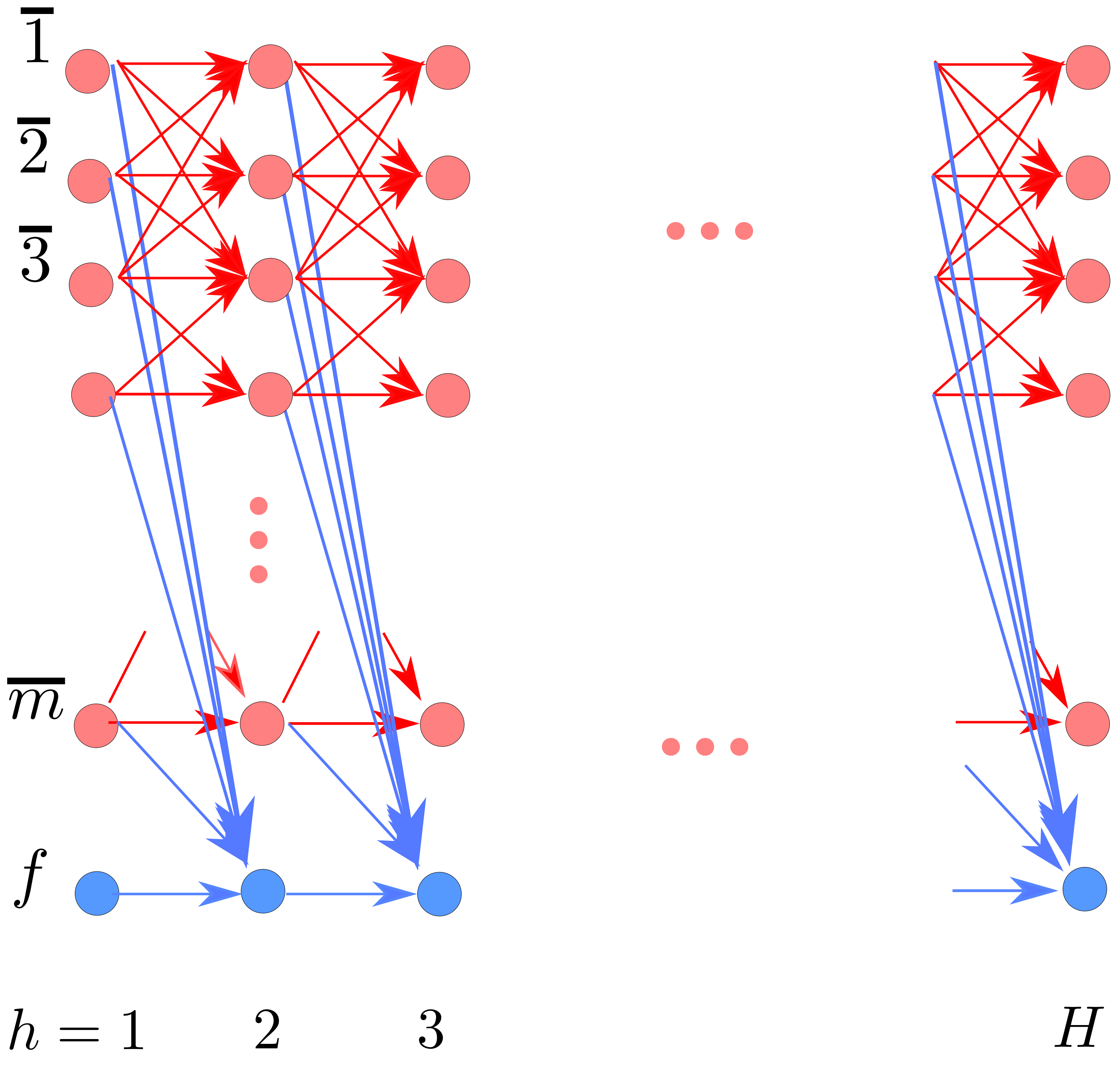


The information theoretic proof:



The information theoretic proof:

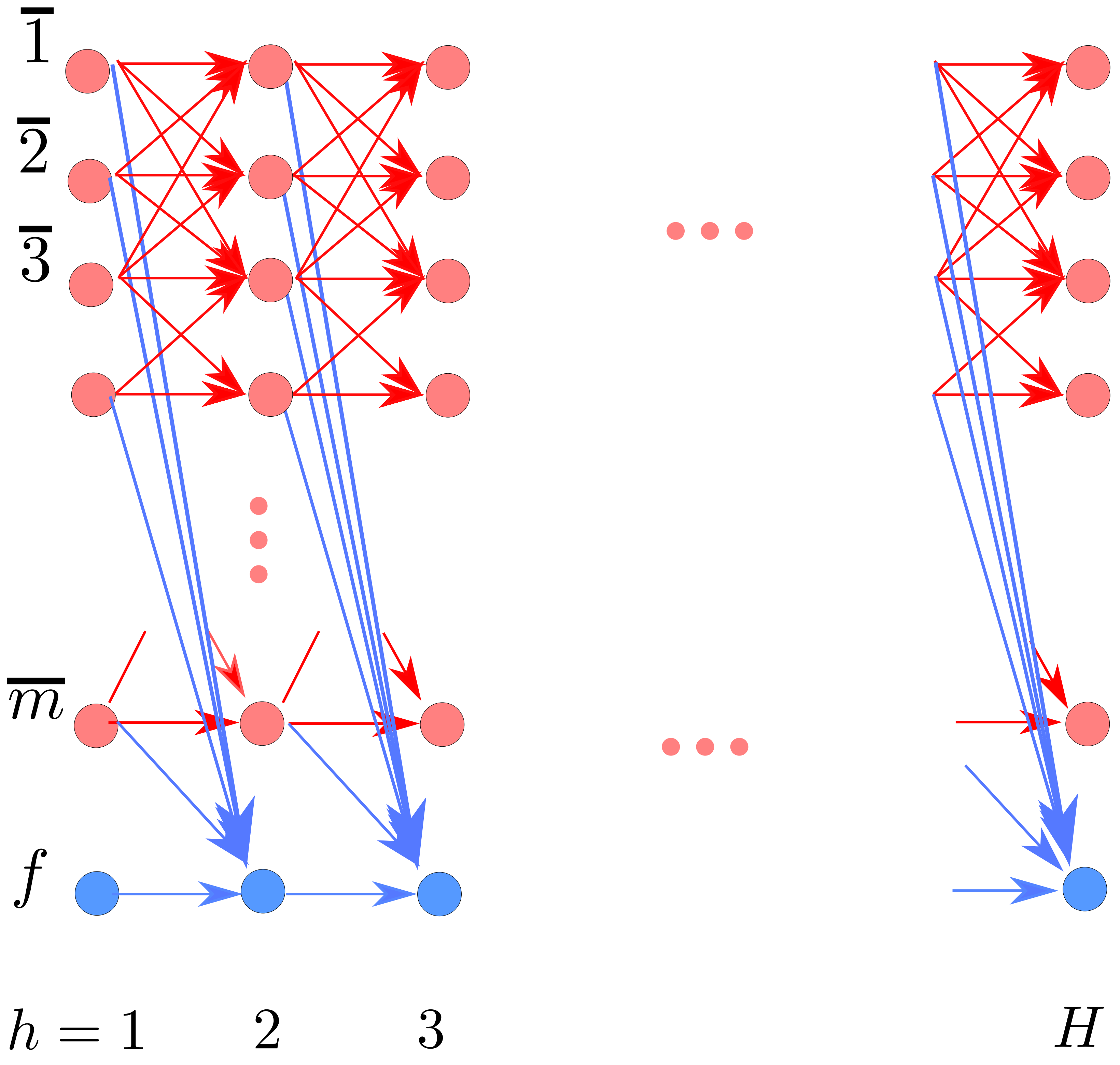
Proof: When is info revealed about $\mathscr{M}_{a^*}$, indexed by a^*?



The information theoretic proof:

Proof: When is info revealed about $\mathscr{M}_{a^*}$, indexed by a^*?

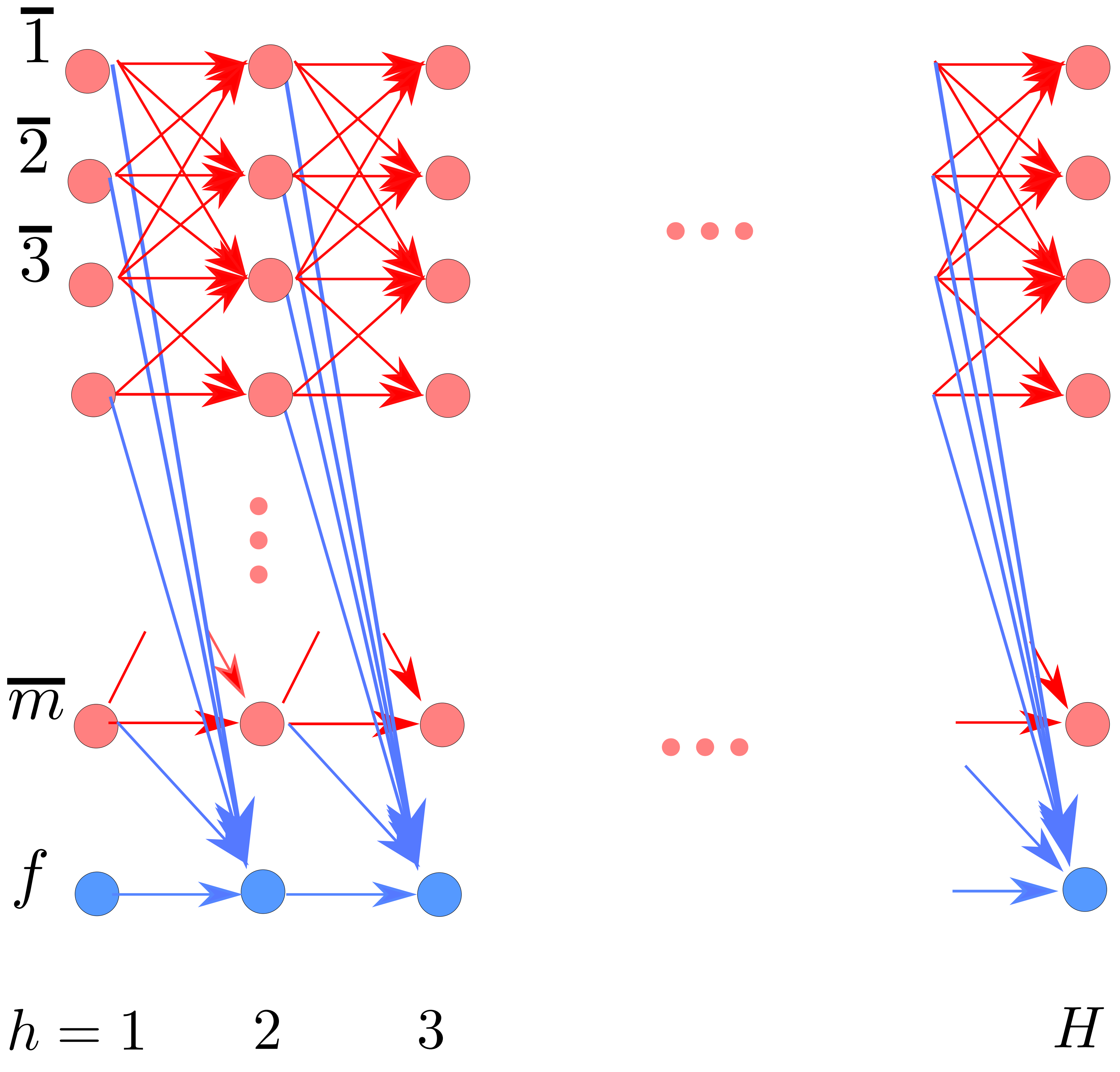
- Features: The construction of ϕ does not depend on $a^\star$.



The information theoretic proof:

Proof: When is info revealed about $\mathscr{M}_{a^*}$, indexed by a^*?

- Features: The construction of ϕ does not depend on $a^\star$.
- Transitions: if we take a^*, only then does the dynamics leak info about a^* (but there $O(2^d)$ actions)



The information theoretic proof:

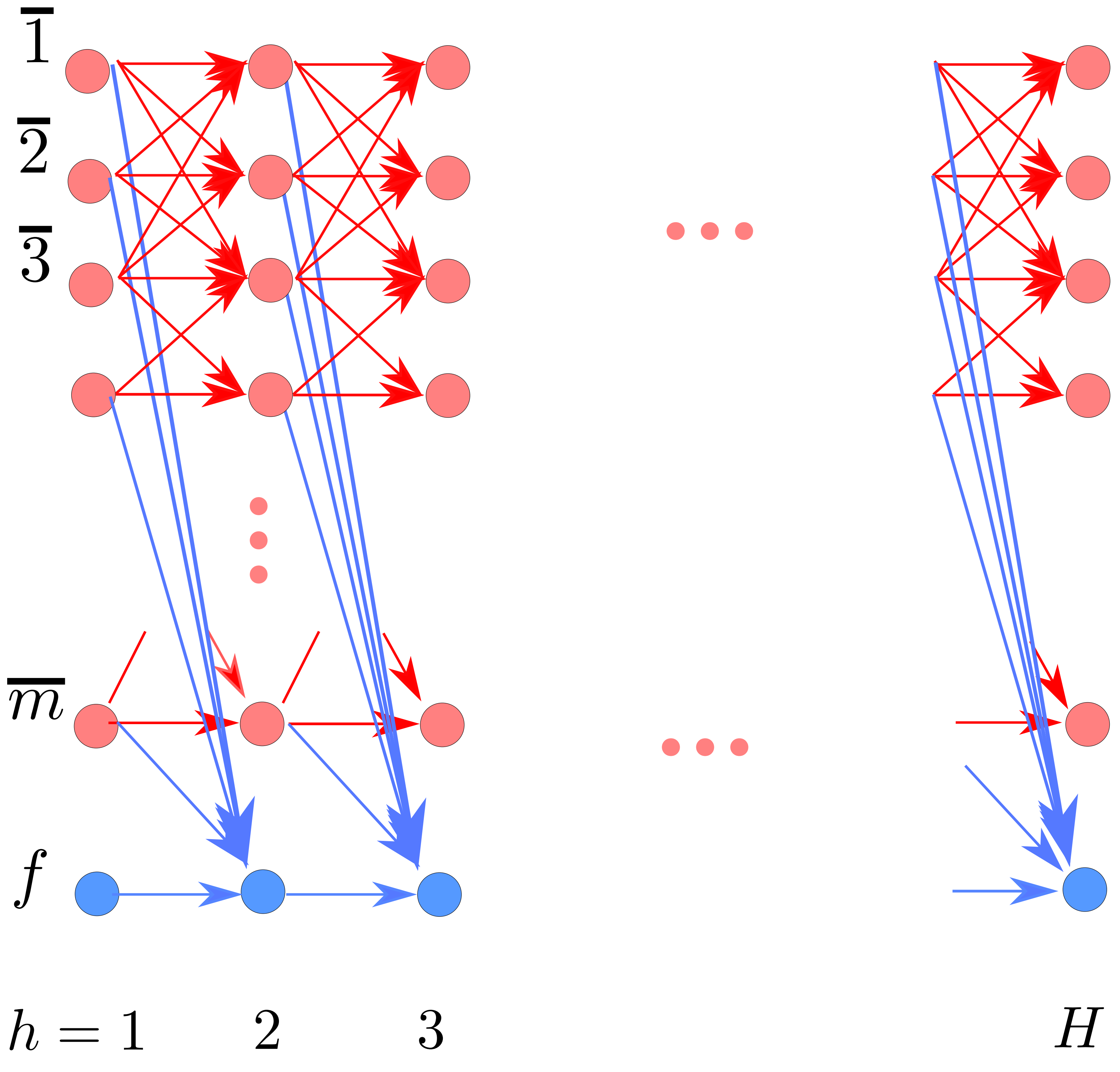
Proof: When is info revealed about $\mathscr{M}_{a^*}$, indexed by a^*?

- Features: The construction of ϕ does not depend on $a^\star$.
- Transitions: if we take a^*, only then does the dynamics leak info about a^* (but there $O(2^d)$ actions)
- Rewards: two cases which leak info about $a^\star$

  (1) if we take a^* at any h, then reward leaks info about a^* (but there $m = O(2^d)$ actions)

  (2) also, if we terminate at $s_H \neq f$, then the reward r_H leaks info about on a^*

  - But there is always at least $1/4$ chance of moving to f
  - So need at least $O((4/3)^H)$ trajectories to hit $s_H \neq f$

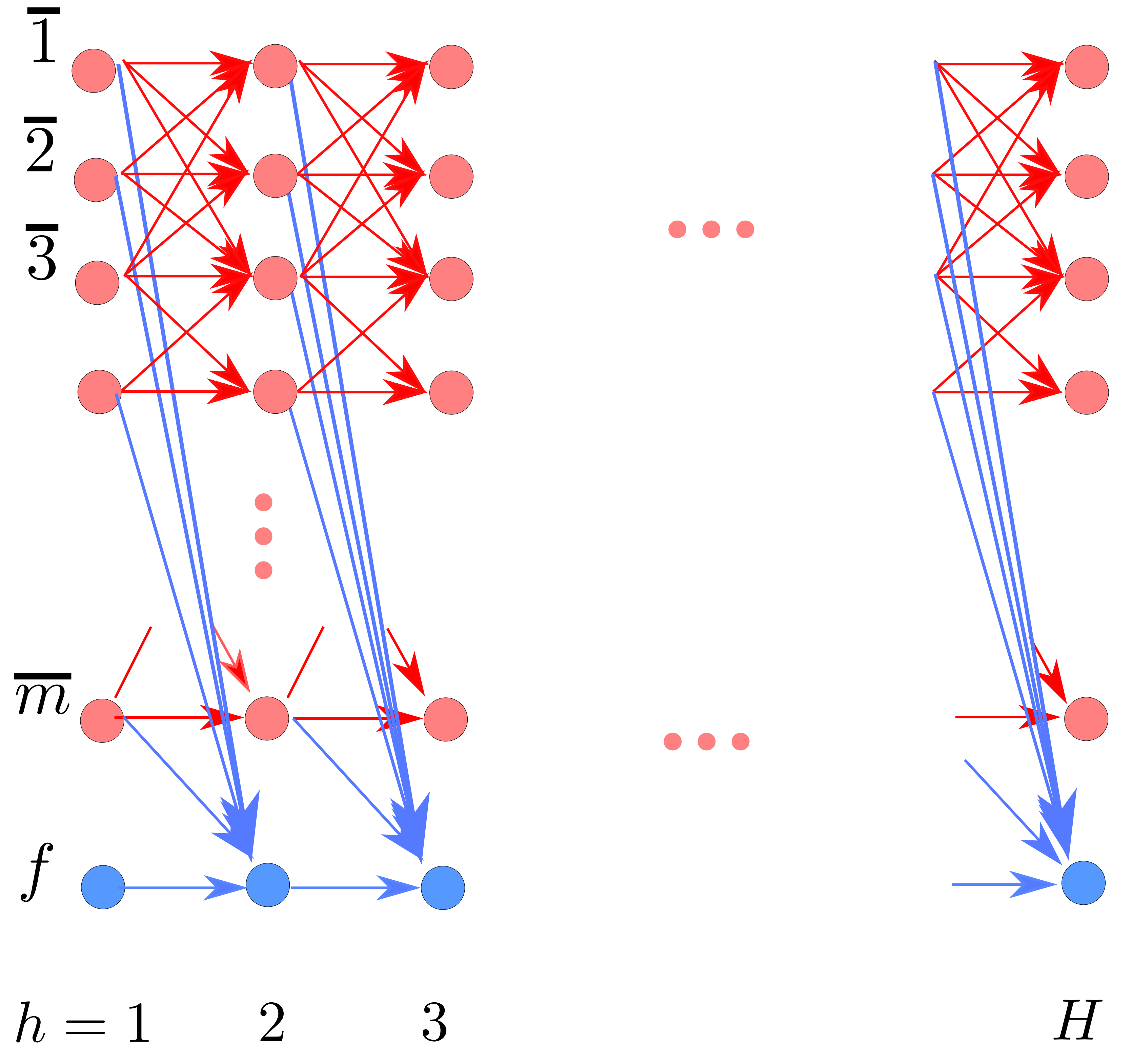


The information theoretic proof:

Proof: When is info revealed about $\mathscr{M}_{a*}$, indexed by $a*$?

- Features: The construction of ϕ does not depend on $a^\star$.
- Transitions: if we take $a*$, only then does the dynamics leak info about $a*$ (but there $O(2^d)$ actions)
- Rewards: two cases which leak info about $a^\star$
  (1) if we take $a*$ at any h, then reward leaks info about $a*$ (but there $m = O(2^d)$ actions)
  (2) also, if we terminate at $s_H \neq f$, then the reward r_H leaks info about on $a*$
  - But there is always at least $1/4$ chance of moving to f
  - So need at least $O((4/3)^H)$ trajectories to hit $s_H \neq f$

$\Longrightarrow$ need $\Omega(\min(2^d, 2^H))$ samples to discover $\mathscr{M}_{a*}$.



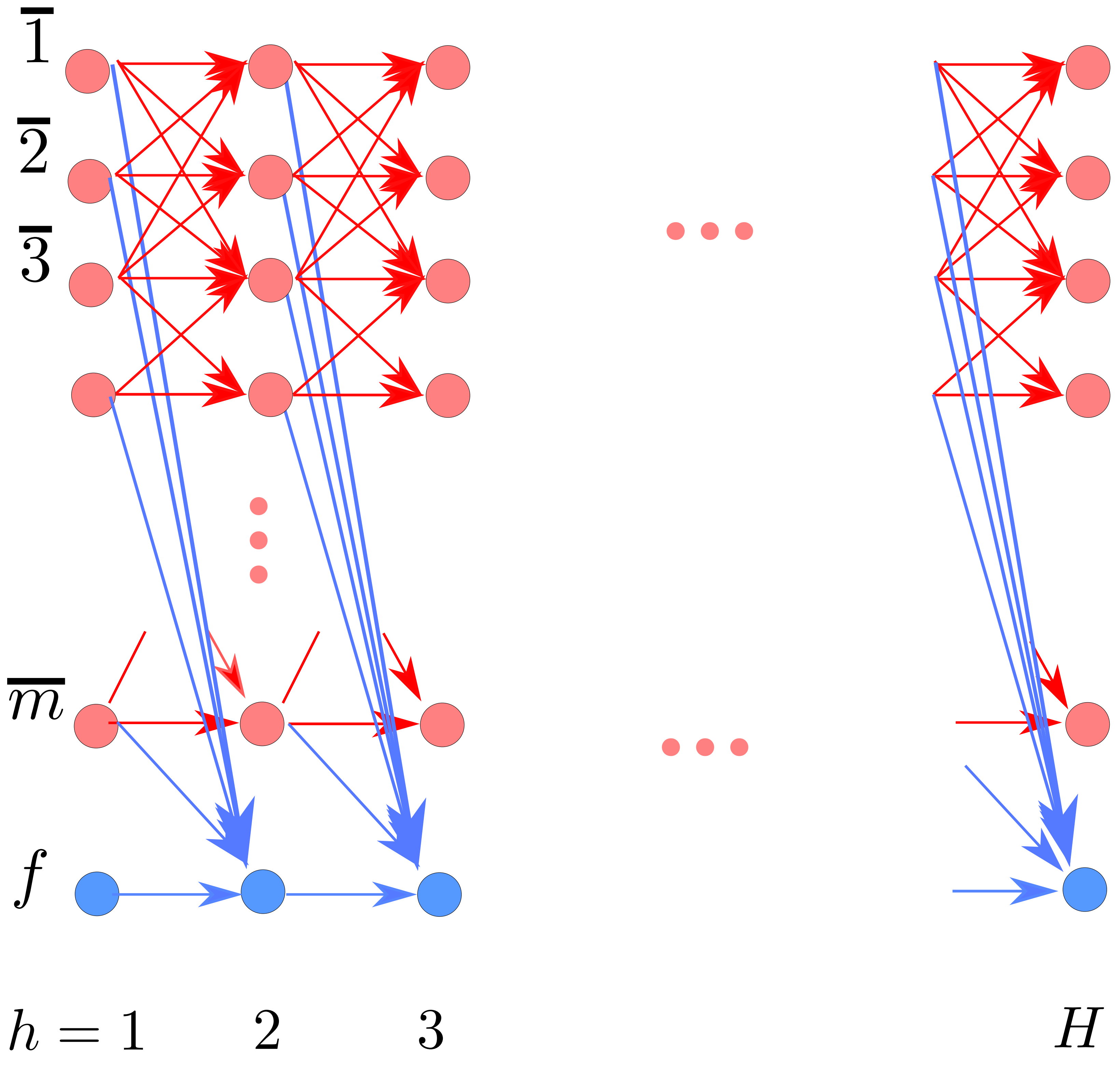
The information theoretic proof:

Proof: When is info revealed about $\mathscr{M}_{a^*}$, indexed by a^*?

- Features: The construction of ϕ does not depend on $a^\star$.
- Transitions: if we take a^*, only then does the dynamics leak info about a^* (but there $O(2^d)$ actions)
- Rewards: two cases which leak info about $a^\star$
  (1) if we take a^* at any h, then reward leaks info about a^* (but there $m = O(2^d)$ actions)
  (2) also, if we terminate at $s_H \neq f$, then the reward r_H leaks info about on a^*
  - But there is always at least $1/4$ chance of moving to f
  - So need at least $O((4/3)^H)$ trajectories to hit $s_H \neq f$

$\Longrightarrow$ need $\Omega(\min(2^d, 2^H))$ samples to discover $\mathscr{M}_{a^*}$.

Caveats: Haven't handled the state $\overline{a}^*$ cafefully.



The information theoretic proof:

Proof: When is info revealed about $\mathscr{M}_{a^*}$, indexed by a^*?

- Features: The construction of ϕ does not depend on $a^\star$.
- Transitions: if we take a^*, only then does the dynamics leak info about a^* (but there $O(2^d)$ actions)
- Rewards: two cases which leak info about $a^\star$
  (1) if we take a^* at any h, then reward leaks info about a^* (but there $m = O(2^d)$ actions)
  (2) also, if we terminate at $s_H \neq f$, then the reward r_H leaks info about on a^*
  - But there is always at least $1/4$ chance of moving to f
  - So need at least $O((4/3)^H)$ trajectories to hit $s_H \neq f$

$\implies$ need $\Omega(\min(2^d, 2^H))$ samples to discover $\mathscr{M}_{a^*}$.

Caveats: Haven't handled the state $\overline{a}^*$ cafefully.

Open Problem: Can we prove a lower bound with $A = 2$ actions?

<u>Part-3:</u> Discussion

RL is different from SL.

+ we have seen negative results.

How do we obtain positive results?

How should we approach generalization in RL?

How should we approach generalization in RL?

- We have seen that:

How should we approach generalization in RL?

- We have seen that:
  - agnostic learning is not possible in RL
    (unless we pay an exponential in H dependence)

How should we approach generalization in RL?

- We have seen that:
  - agnostic learning is not possible in RL
    (unless we pay an exponential in H dependence)
  - simple linear realizability assumptions are also not sufficient

How should we approach generalization in RL?

- We have seen that:
  - agnostic learning is not possible in RL

    (unless we pay an exponential in H dependence)
  - simple linear realizability assumptions are also not sufficient
- What next?

How should we approach generalization in RL?

- We have seen that:
  - agnostic learning is not possible in RL

    (unless we pay an exponential in H dependence)
  - simple linear realizability assumptions are also not sufficient
- What next?
  - Structural Assumptions: Need even stronger assumptions. We start this study (**today**) with the stronger linear Bellman completeness. More examples of this in “Part 2”.

How should we approach generalization in RL?

- We have seen that:
  - agnostic learning is not possible in RL
    (unless we pay an exponential in H dependence)
  - simple linear realizability assumptions are also not sufficient
- What next?
  - Structural Assumptions: Need even stronger assumptions. We start this study (**today**) with the stronger linear Bellman completeness. More examples of this in "Part 2".
  - Distribution Dependent Results: We will see examples of this approach when we consider approximate dynamic programming. And more refined bounds when we consider policy gradient methods.

How should we approach generalization in RL?

- We have seen that:
  - agnostic learning is not possible in RL
    (unless we pay an exponential in H dependence)
  - simple linear realizability assumptions are also not sufficient
- What next?
  - Structural Assumptions: Need even stronger assumptions. We start this study (**today**) with the stronger linear Bellman completeness. More examples of this in "Part 2".
  - Distribution Dependent Results: We will see examples of this approach when we consider approximate dynamic programming. And more refined bounds when we consider policy gradient methods.
  - Imitation learning and behavior cloning: models where the agent has input from, effectively, a teacher.