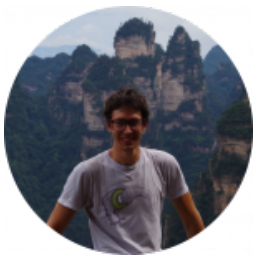
Potential-based reward shaping as a tool to safely incorporate auxiliary information

Anna Harutyunyan
AI Lab, *VU Brussel*

work with Tim Brys, Peter Vrancx, Ann Nowé



(as well as Sam Devlin, Matt Taylor, Halit Bener Suay, Sonia Chernova)

SequeL seminar
INRIA Lille
March 10, 2017

AI Lab at VUB

AI Lab at VUB

AI Lab at VUB

- Founded in 1983 by Luc Steels as the first AI lab in mainland Europe
  - focus in evolution of language and construction grammars

AI Lab at VUB

- Founded in 1983 by Luc Steels as the first AI lab in mainland Europe
  - focus in evolution of language and construction grammars
- Later (late 90s?), our group split off as a ML offshoot

AI Lab at VUB

- Founded in 1983 by Luc Steels as the first AI lab in mainland Europe
  - focus in evolution of language and construction grammars
- Later (late 90s?), our group split off as a ML offshoot
  - Led by Ann Nowe and Bernard Manderick
  - 11 PhD students, 5-7 postdocs
  - Multi-objective / multi-agent RL, deep RL, game theory, computational biology
  - Applications to wind turbine control, smart grids, HIV treatment strategies, etc

This talk

will advocate for potential-based reward shaping as a suitable paradigm for integrating auxiliary information into RL, in particular when the information is in the form of behavioral advice, or expert demonstration data.

The world is full of (questionable) reward functions

- There are many sources of reward complementary to the main objective

The world is full of (questionable) reward functions

- There are many sources of reward complementary to the main objective

  Transfer old policies, similar policies

The world is full of (questionable) reward functions

- There are many sources of reward complementary to the main objective

  Transfer old policies, similar policies
  Unsupervised novelty, auxilary tasks

The world is full of (questionable) reward functions

- There are many sources of reward complementary to the main objective

Transfer	old policies, similar policies
Unsupervised	novelty, auxilary tasks
Supervised	domain knowledge, expert data, online feedback

The world is full of (questionable) reward functions

- There are many sources of reward complementary to the main objective

Transfer	old policies, similar policies
Unsupervised	novelty, auxilary tasks
Supervised	domain knowledge, expert data, online feedback

- How to safely combine them with the native reward?

Reward shaping[1]

[1] Skinner BF, 1938. The behavior of organisms: An experimental analysis.

Reward shaping[1]

In RL, a naive reward shaping scheme augments the MDP:

$$M = (S, A, P, \gamma, R) \rightarrow M' = (S, A, P, \gamma, R + F),$$

where we hope that M' is somehow easier.

[1] Skinner BF, 1938. The behavior of organisms: An experimental analysis.

Reward shaping[1]

In RL, a naive reward shaping scheme augments the MDP:

$$M = (S, A, P, \gamma, R) \rightarrow M' = (S, A, P, \gamma, R + F),$$

where we hope that M' is somehow easier.

- F can be wrong
- F can be right, but have inadvertent consequences

[1] Skinner BF, 1938. The behavior of organisms: An experimental analysis.

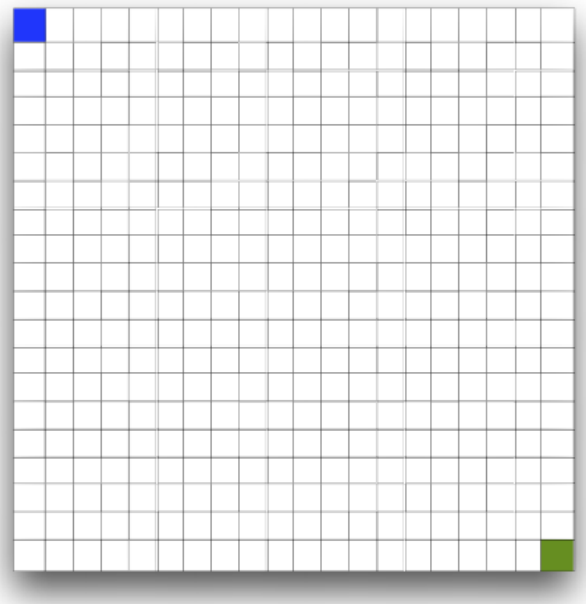
Reward shaping[1]

In RL, a naive reward shaping scheme augments the MDP:

$$M = (S, A, P, \gamma, R) \rightarrow M' = (S, A, P, \gamma, R + F),$$

where we hope that M' is somehow easier.

- F can be wrong
- F can be right, but have inadvertent consequences



[1] Skinner BF, 1938. The behavior of organisms: An experimental analysis.

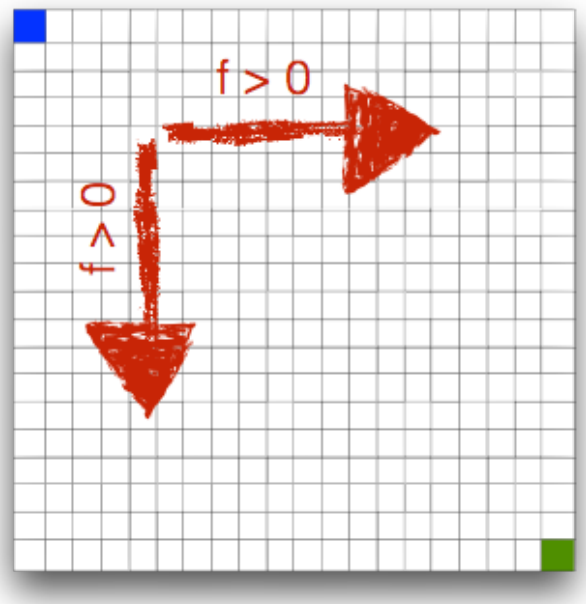
Reward shaping[1]

In RL, a naive reward shaping scheme augments the MDP:

$$M = (S, A, P, \gamma, R) \rightarrow M' = (S, A, P, \gamma, R + F),$$

where we hope that M' is somehow easier.

- F can be wrong
- F can be right, but have inadvertent consequences



[1] Skinner BF, 1938. The behavior of organisms: An experimental analysis.

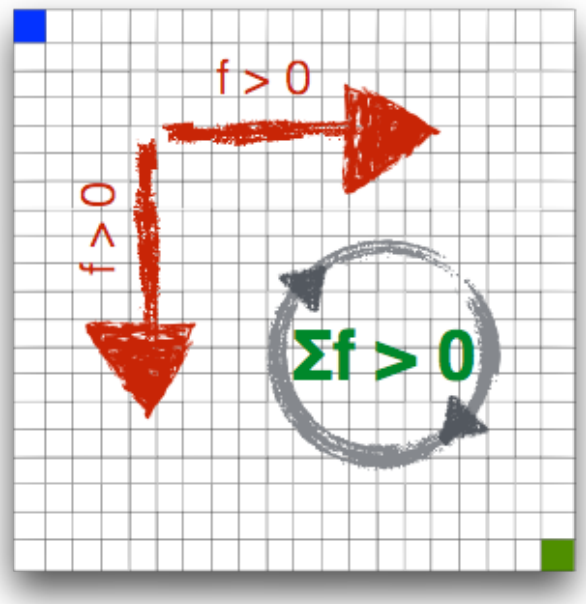
Reward shaping[1]

In RL, a naive reward shaping scheme augments the MDP:

$$M = (S, A, P, \gamma, R) \rightarrow M' = (S, A, P, \gamma, R + F),$$

where we hope that M' is somehow easier.

- F can be wrong
- F can be right, but have inadvertent consequences



[1] Skinner BF, 1938. The behavior of organisms: An experimental analysis.

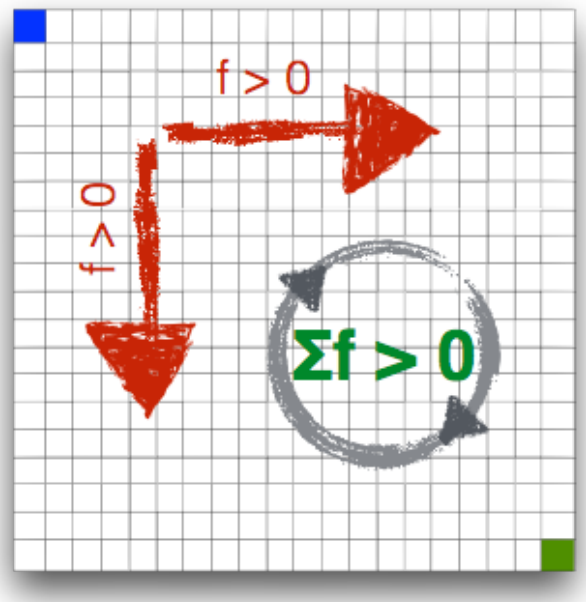
Reward shaping[1]

In RL, a naive reward shaping scheme augments the MDP:

$$M = (S, A, P, \gamma, R) \rightarrow M' = (S, A, P, \gamma, R + F),$$

where we hope that M' is somehow easier.

- F can be wrong
- F can be right, but have inadvertent consequences



So we want to solve the original M, but use F to solve it quicker.

[1] Skinner BF, 1938. The behavior of organisms: An experimental analysis.

Potential-based reward shaping (PBRS)[2]

[2] Ng, Harada and Russel. "Policy Invariance Under Reward Transformations". ICML, 1999

Potential-based reward shaping (PBRS)[2]

Definition

F is **potential-based**, if $\exists \Phi : S \to \mathbb{R}$, s.t. $\forall s, s' \in S, a \in A$:

$$F(s, a, s') = \gamma \Phi(s') - \Phi(s)$$

[2] Ng, Harada and Russel. "Policy Invariance Under Reward Transformations". ICML, 1999

Potential-based reward shaping (PBRS)[2]

Definition

F is **potential-based**, if $\exists \Phi : S \to \mathbb{R}$, s.t. $\forall s, s' \in S, a \in A$:

$$F(s, a, s') = \gamma \Phi(s') - \Phi(s)$$

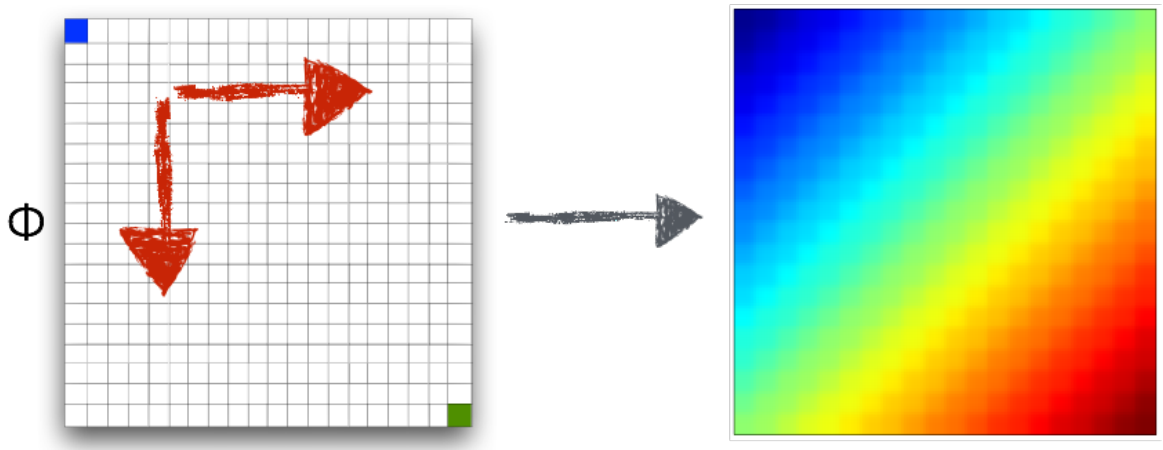
Theorem

That F is potential-based is sufficient and necessary (when M is completely unknown) to guarantee that the optimal policies of $M = (S, A, P, \gamma, R)$ and $M' = (S, A, P, \gamma, R + F)$ are the same.

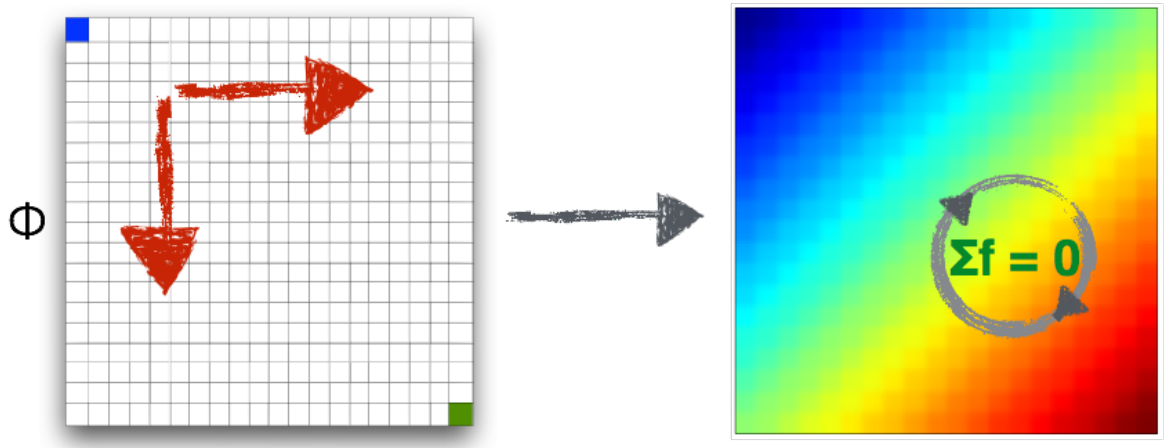
[2] Ng, Harada and Russel. "Policy Invariance Under Reward Transformations". ICML, 1999

Potential functions encode how good it is to be in a state

Potential functions encode how good it is to be in a state



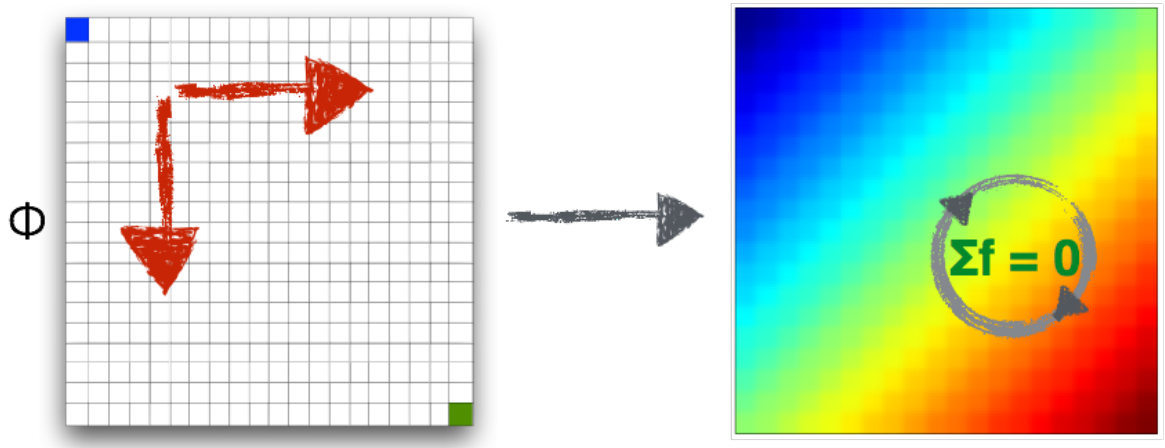
Potential functions encode how good it is to be in a state



Policy invariance

- No positive reward cycles

Potential functions encode how good it is to be in a state



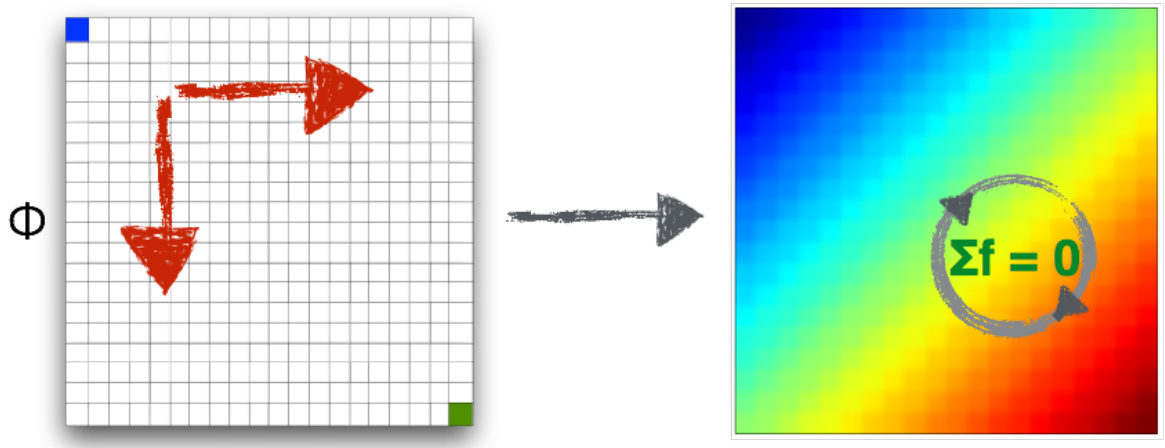
Policy invariance

- No positive reward cycles
- The relationship of value functions is just a constant shift:

$$\begin{aligned} Q^*_{M'}(s,a) &= Q^*_M(s,a) - \Phi(s) \\ V^*_{M'}(s) &= V^*_M(s) - \Phi(s) \end{aligned}$$

This allows to extend Φ to be over state-actions

Potential functions encode how good it is to be in a state



Policy invariance

- No positive reward cycles
- The relationship of value functions is just a constant shift:

$$\begin{aligned} Q^*_{M'}(s,a) &= Q^*_M(s,a) - \Phi(s) \\ V^*_{M'}(s) &= V^*_M(s) - \Phi(s) \end{aligned}$$

  This allows to extend Φ to be over state-actions
- Ideal potential function = optimal value function

PBRS vs. Q-value initialization

- PBRS fell out of fashion for a time, as it was discovered that it's exactly equivalent to Q-value initialization, given the same stream of experience.[3]

[3] Wiewiora, E. (2003). Potential-based shaping and Q-value initialization are equivalent. J. Artif. Intell. Res.(JAIR), 19, 205-208.

PBRS vs. Q-value initialization

- PBRS fell out of fashion for a time, as it was discovered that it's exactly equivalent to Q-value initialization, given the same stream of experience.[3]
- When in a **tabular** setting with a **fixed** potential function, initialization is feasible, and might be simpler

[3] Wiewiora, E. (2003). Potential-based shaping and Q-value initialization are equivalent. J. Artif. Intell. Res.(JAIR), 19, 205-208.

PBRS vs. Q-value initialization

- PBRS fell out of fashion for a time, as it was discovered that it's exactly equivalent to Q-value initialization, given the same stream of experience.[3]
- When in a **tabular** setting with a **fixed** potential function, initialization is feasible, and might be simpler
- but it gets trickier with function approximation, and impossible with potential functions that change over time,

[3] Wiewiora, E. (2003). Potential-based shaping and Q-value initialization are equivalent. J. Artif. Intell. Res.(JAIR), 19, 205-208.

PBRS vs. Q-value initialization

- PBRS fell out of fashion for a time, as it was discovered that it's exactly equivalent to Q-value initialization, given the same stream of experience.[3]
- When in a **tabular** setting with a **fixed** potential function, initialization is feasible, and might be simpler
- but it gets trickier with function approximation, and impossible with potential functions that change over time,
- PBRS still provides the only Bellman-consistent mechanism of combining value functions.

[3] Wiewiora, E. (2003). Potential-based shaping and Q-value initialization are equivalent. J. Artif. Intell. Res.(JAIR), 19, 205-208.

This talk: From information to potentials

- PBRS is theoretically justified, but requires an extra abstraction – the potential function

This talk: From information to potentials

- PBRS is theoretically justified, but requires an extra abstraction – the potential function
- In the rest of the talk, we'll cover a few ways to obtain a potential function from whatever information is available. In particular:

This talk: From information to potentials

- PBRS is theoretically justified, but requires an extra abstraction – the potential function
- In the rest of the talk, we'll cover a few ways to obtain a potential function from whatever information is available. In particular:

  Advice is trivial to express a reward function. We describe a trick that for an arbitrary reward function, obtains the corresponding potential function whose induced shaping reward reflects the desired one exactly.

This talk: From information to potentials

- PBRS is theoretically justified, but requires an extra abstraction – the potential function
- In the rest of the talk, we'll cover a few ways to obtain a potential function from whatever information is available. In particular:

  Advice is trivial to express a reward function. We describe a trick that for an arbitrary reward function, obtains the corresponding potential function whose induced shaping reward reflects the desired one exactly.

  Expert trajectories. We construct a potential function centered around the demonstrated state-action pairs with generalization based on state similarity

From rewards to potentials: Bellman equations to the rescue[4]

- By definition of PBRS we have that the shaping reward

$$F = \gamma P^{\pi} \Phi - \Phi.$$

- Normally, we know Φ, and calculate F. What if we know F and would like to obtain Φ?

[4] Harutyunyan, A., Devlin, S., Vrancx, P., and Nowé, A. Expressing Arbitrary Reward Functions as Potential-Based Advice. In AAAI 2015.

From rewards to potentials: Bellman equations to the rescue[4]

- By definition of PBRS we have that the shaping reward

$$F = \gamma P^{\pi} \Phi - \Phi.$$

- Normally, we know Φ, and calculate F. What if we know F and would like to obtain Φ?
- Well, $\Phi = (I - \gamma P^{\pi})^{-1}(-F)$, which is exactly something we can estimate in RL!

[4] Harutyunyan, A., Devlin, S., Vrancx, P., and Nowé, A. Expressing Arbitrary Reward Functions as Potential-Based Advice. In AAAI 2015.

From rewards to potentials: Bellman equations to the rescue[4]

- By definition of PBRS we have that the shaping reward

$$F = \gamma P^{\pi} \Phi - \Phi.$$

- Normally, we know Φ, and calculate F. What if we know F and would like to obtain Φ?
- Well, $\Phi = (I - \gamma P^{\pi})^{-1}(-F)$, which is exactly something we can estimate in RL!
- The reason this helps is because
  - evaluation is easier
  - Φ will be helpful long before convergence

[4] Harutyunyan, A., Devlin, S., Vrancx, P., and Nowé, A. Expressing Arbitrary Reward Functions as Potential-Based Advice. In AAAI 2015.

Empirically it works

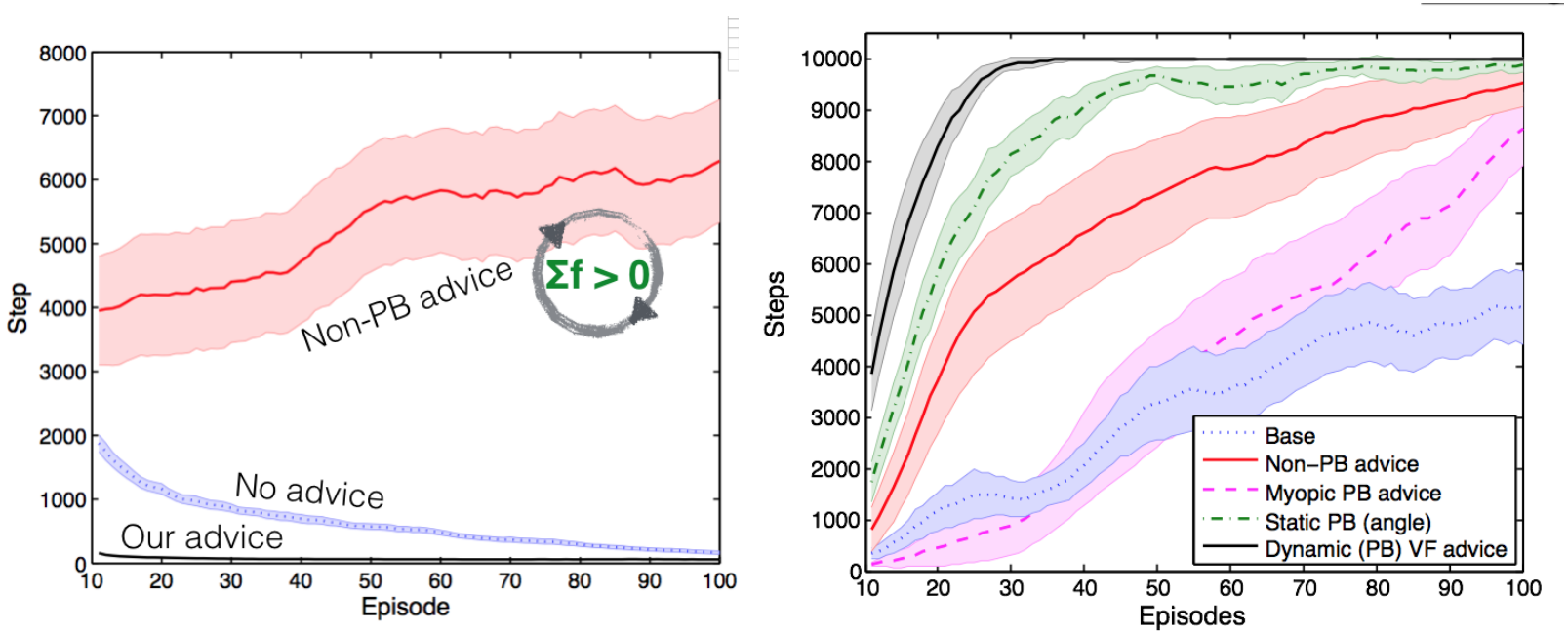


Figure: **Left:** The gridworld example. **Right:** Cartpole.

Learning from online feedback

A nice corollary is that it's possible to soundly learn from sporadic (e.g. online) feedback.

Learning from online feedback

A nice corollary is that it's possible to soundly learn from sporadic (e.g. online) feedback.
It wasn't before because:

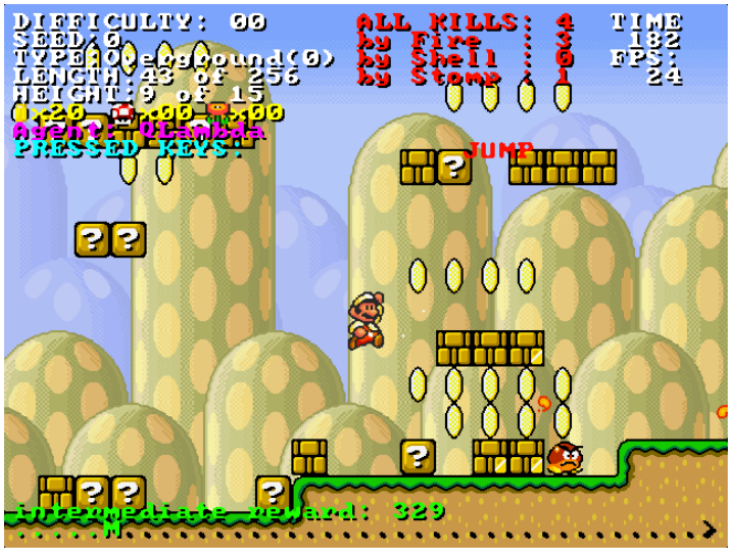
- it is very easy to create positive reward cycles with ad hoc feedback, so non-PB is dangerous

Learning from online feedback

A nice corollary is that it's possible to soundly learn from sporadic (e.g. online) feedback.
It wasn't before because:

- it is very easy to create positive reward cycles with ad hoc feedback, so non-PB is dangerous
- A naive translation of $\Phi = R^A$ does not work either:
  - Let $R^A(s, a) = 1$ at an approved action and 0 elsewhere
  - Then $F(s, a, s') = \Phi(s', a') - \Phi(s, a) = -1$
  - The desired behavior got a *negative* shaping reward

Advising Mario online[5]



- 12 actions
- 7000 state features per action
- The advisor clicked a button if it approved of Mario's behavior, and did nothing otherwise (a very frustrating setting)

Variant	Advice phase	Cumulative
Baseline	-376±51	470±83
Non-expert	**401±54**	**677±60**
Expert	**402±62**	**774±47**

[5] Harutyunyan, A., Brys, T., Vrancx, P., and Nowé, A. Shaping mario with human advice (demonstration). In Proceedings of AAMAS 2015.

Discussion

- Works well when Φ isn't too hard to learn
  - e.g. when the advice is sparse, and has the same sign
- Stability may be an issue, since the potential function must be learnt on-policy, while in control the behavior typically changes.
  - A policy iteration setting or updating on two timescales is needed.

Learning from demonstration (LfD)

LfD setting

Turn the learning problem into a classification task:

- Given examples $\{(s_i, a_i)\}_{i=0}^{K}$ of actions in states, and assuming they are optimal, the goal is to learn the underlying mapping $\pi : S \to A$, which by assumption will be optimal.

Learning from demonstration (LfD)

LfD setting

Turn the learning problem into a classification task:

- Given examples $\{(s_i, a_i)\}_{i=0}^{K}$ of actions in states, and assuming they are optimal, the goal is to learn the underlying mapping $\pi : S \to A$, which by assumption will be optimal.

The assumption is practical, but problematic, as the agent may never surpass the teacher. We'd like to be able to improve through learning. But how to incorporate?

Reinforcement learning from demonstration (RLfD) through shaping[6]

Let us consider tasks where both demonstrations and an environment reward signal are available.

- In any goal-directed task (however complex) this can be done by having a reward of 1 at the goal, and a reward of 0 elsewhere. This will be extremely difficult to learn, but optimizing it will be sure to do the right thing.

In such tasks, we propose to incorporate demonstration data into the reward function through PBRS.

[6] Brys, T., Harutyunyan, A., Suay, H. B., Chernova, S., Taylor, M. E., and Nowé, A. Reinforcement Learning from Demonstration through Shaping. In IJCAI 2015

Similarity-based shaping

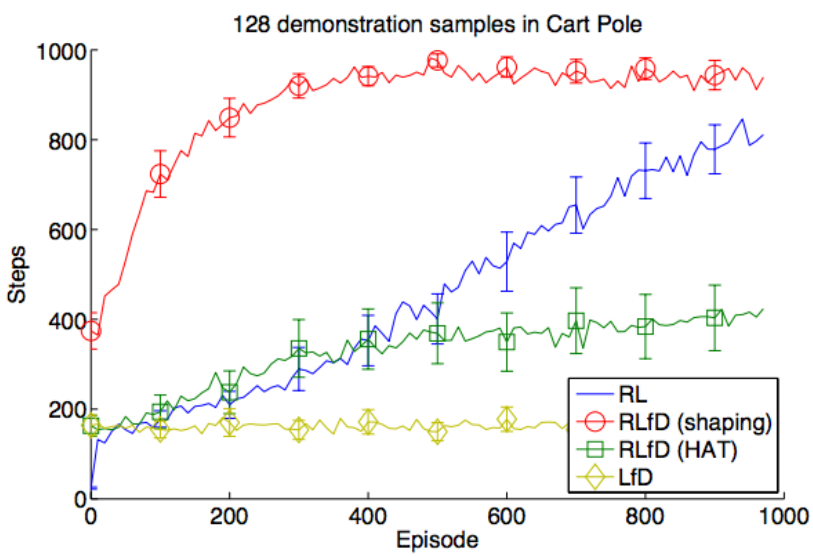
Construct a potential function centered around the demonstrated state-action pairs, with generalization w.r.t. a covariance matrix over states. Simple form:

$$\Sigma = \sigma I,$$

$$g(s, a, s^d, a^d, \Sigma) = \begin{cases} e^{-\frac{1}{2}(s-s^d)^T \Sigma^{-1}(s-s^d)}, & \text{if } a = a^d \\ 0, & \text{otherwise} \end{cases}$$

$$\Phi(s, a) = \max_{s^d, a^d} g(s, a, s^d, a^d, \Sigma)$$

Empirical studies: learning curves



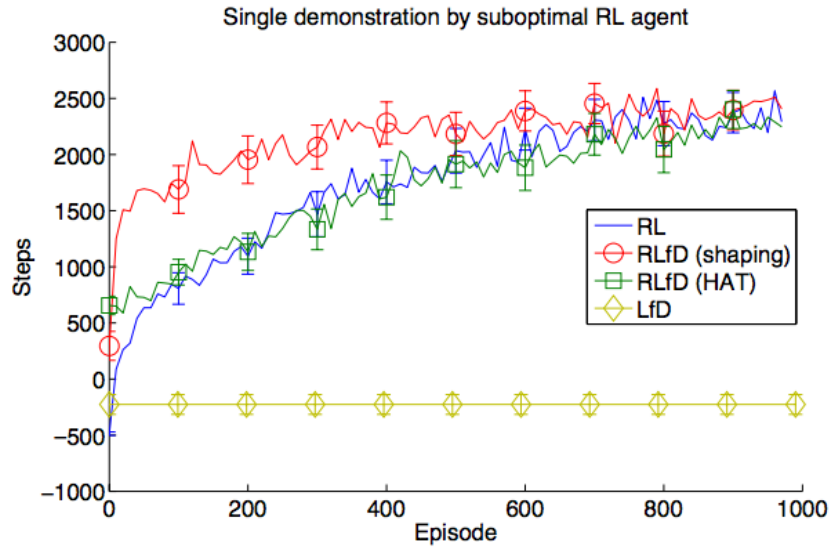
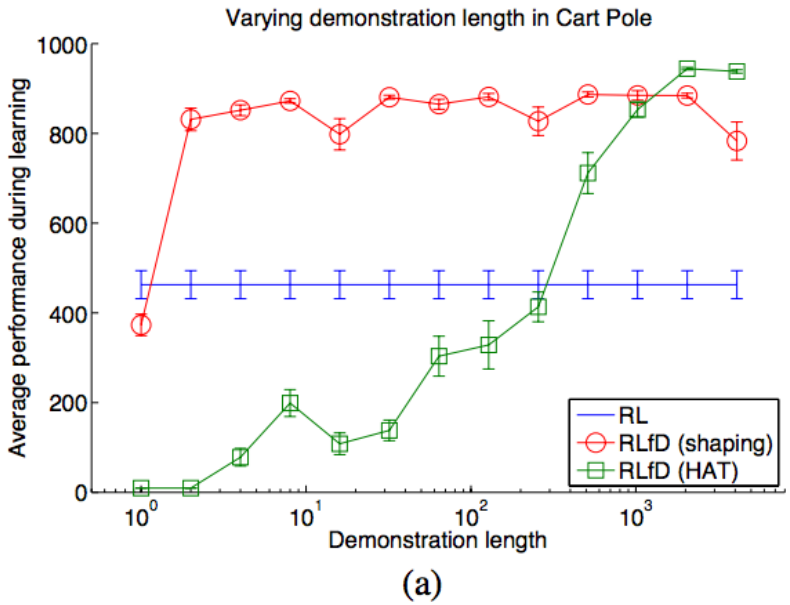
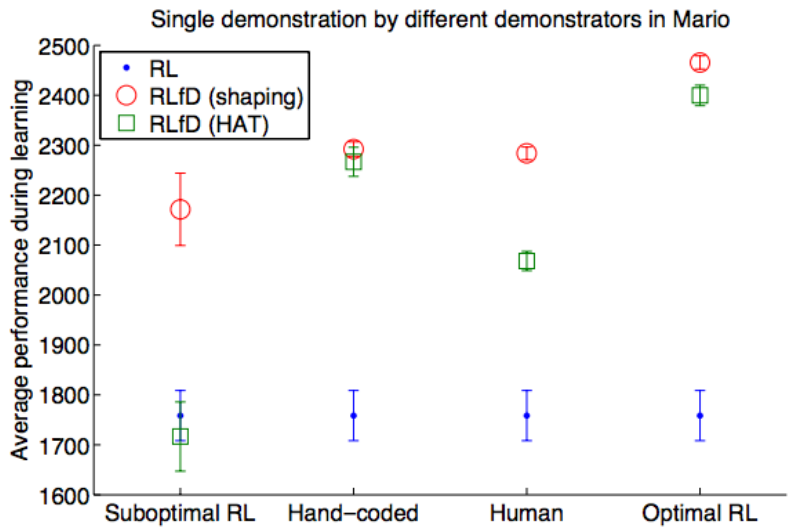


Figure: **Left:** Cartpole, **Right:** Mario

Empirical studies: effect of demonstration length (left) and quality (right)





Discussion

- Works surprisingly well with a very small number (consistent!) demonstrations
- Can be robust to the quality of the demonstrations
- Performance depends on the feature covariance kernel: in our experiments, it was of fixed width, which was a parameter
- Computationally expensive – search for a max over demonstration pairs at every iteration, but speedup tricks are possible

Inverse RL through PBRS [7]

[7] Suay, H. B., Brys, T., Taylor, M. E., and Chernova, S. Learning from demonstration for shaping through inverse reinforcement learning. In AAMAS 2016.

Inverse RL through PBRS [7]

Inverse RL

Given examples $\{(s_i, a_i)\}_{i=0}^{K}$ of expert actions, infer the *reward function* that the expert policy has optimized, and solve the MDP w.r.t. it

[7] Suay, H. B., Brys, T., Taylor, M. E., and Chernova, S. Learning from demonstration for shaping through inverse reinforcement learning. In AAMAS 2016.

Inverse RL through PBRS [7]

Inverse RL

Given examples $\{(s_i, a_i)\}_{i=0}^{K}$ of expert actions, infer the *reward function* that the expert policy has optimized, and solve the MDP w.r.t. it

- This work doesn't assume access to the model, and solves the MDP from samples.

[7] Suay, H. B., Brys, T., Taylor, M. E., and Chernova, S. Learning from demonstration for shaping through inverse reinforcement learning. In AAMAS 2016.

Inverse RL through PBRS [7]

Inverse RL

Given examples $\{(s_i, a_i)\}_{i=0}^{K}$ of expert actions, infer the *reward function* that the expert policy has optimized, and solve the MDP w.r.t. it

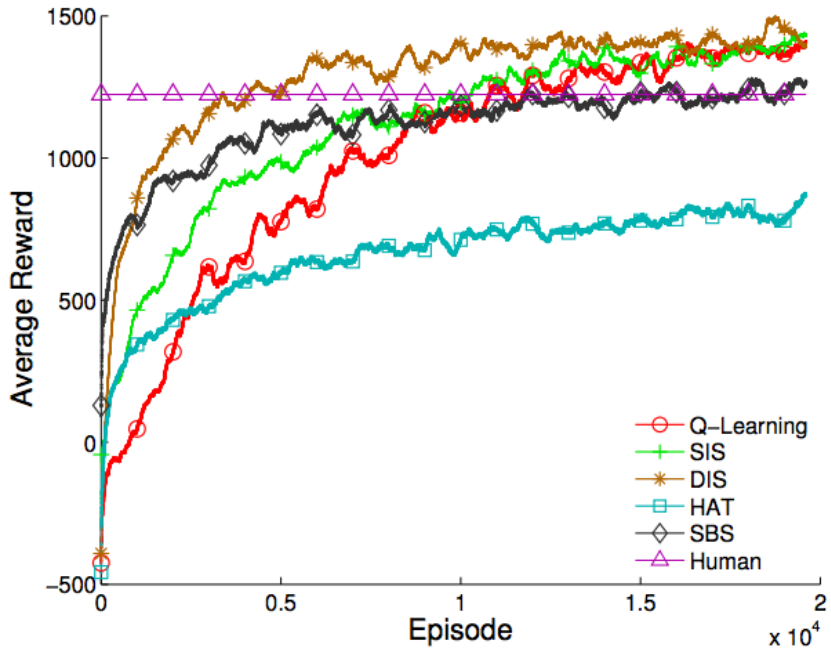
- This work doesn't assume access to the model, and solves the MDP from samples.

Same problem of relying on the expert being optimal. We could try the same trick as before:

1. Infer the expert reward function
2. **Learn the potential function that expresses it**, as discussed
3. Use PBRS to incorporate it safely

[7] Suay, H. B., Brys, T., Taylor, M. E., and Chernova, S. Learning from demonstration for shaping through inverse reinforcement learning. In AAMAS 2016.

Empirically (Mario)



SBS Φ is similarity-based

HAT another RLfD benchmark

SIS $\Phi = R$ (f of state)

DIS Φ is learnt w.r.t. to $-R$

There is benefit to learning a value function, rather than treating R as a myopic potential.

Outro

- We advocate using potentials when trying to modify the reward function.
- It adds an extra safeguard abstraction between the agent and imperfect domain knowledge
- We discussed a few ways to obtain this abstraction

[8] https://openai.com/blog/faulty-reward-functions/

Outro

- We advocate using potentials when trying to modify the reward function.
- It adds an extra safeguard abstraction between the agent and imperfect domain knowledge
- We discussed a few ways to obtain this abstraction

Philosophically

Dense reward function are dangerous – we don't know what loopholes the agent will find. E.g. Atari games are full of these.[8] Having a minimalistic reward signal, and many overlapping potential fields that guide the agent around the environment seems like an appealing way to think about learning.

[8] https://openai.com/blog/faulty-reward-functions/

What's missing

- A lot of room for theoretical work

[9] Ng. Shaping and policy search in reinforcement learning. PhD thesis. 2003.

What's missing

- A lot of room for theoretical work
  - Can we have theoretically grounded prescriptions for "good" potentials that guarantee improvement in learning given some assumptions on the reward structure?

[9] Ng. Shaping and policy search in reinforcement learning. PhD thesis. 2003.

What's missing

- A lot of room for theoretical work
  - Can we have theoretically grounded prescriptions for "good" potentials that guarantee improvement in learning given some assumptions on the reward structure?
  - Can we quantify the effect of PBRS beyond it being an exploration boost?

[9] Ng. Shaping and policy search in reinforcement learning. PhD thesis. 2003.

What's missing

- A lot of room for theoretical work
  - Can we have theoretically grounded prescriptions for "good" potentials that guarantee improvement in learning given some assumptions on the reward structure?
  - Can we quantify the effect of PBRS beyond it being an exploration boost?
    - E.g. there is intuition that it reduces the planning horizon

Theorem (PBRS reduces planning horizon[9])

Let M be the MDP (S, A, T, γ, R). Let F be a PB reward function w.r.t. Φ, s.t. $||\Phi - V^*_M||_\infty < \epsilon$, and let $\hat{\pi}$ be the optimal policy for $\hat{M} = (S, A, T, \gamma', R + F)$ for some $\gamma' < \gamma$. Then:

$$V^{\hat{\pi}}_M \geq V^*_M - O_\gamma((\gamma - \gamma')\epsilon)$$

[9] Ng. Shaping and policy search in reinforcement learning. PhD thesis. 2003.

Thank you for your attention!

Questions?

References

- Ng, Harada and Russel. "Policy Invariance Under Reward Transformations". ICML, 1999
- Wiewiora, E. (2003). Potential-based shaping and Q-value initialization are equivalent. J. Artif. Intell. Res.(JAIR), 19.
- Harutyunyan, A., Devlin, S., Vrancx, P., and Nowé, A. Expressing Arbitrary Reward Functions as Potential-Based Advice. In AAAI 2015
- Harutyunyan, A., Brys, T., Vrancx, P., and Nowé, A. Shaping mario with human advice (Demonstration). In AAMAS 2015
- Brys, T., Harutyunyan, A., Suay, H. B., Chernova, S., Taylor, M. E., and Nowé, A. Reinforcement Learning from Demonstration through Shaping. In IJCAI 2015
- Suay, H. B., Brys, T., Taylor, M. E., and Chernova, S. Learning from demonstration for shaping through inverse reinforcement learning. In AAMAS 2016
- Ng. Shaping and policy search in reinforcement learning. PhD thesis. 2003