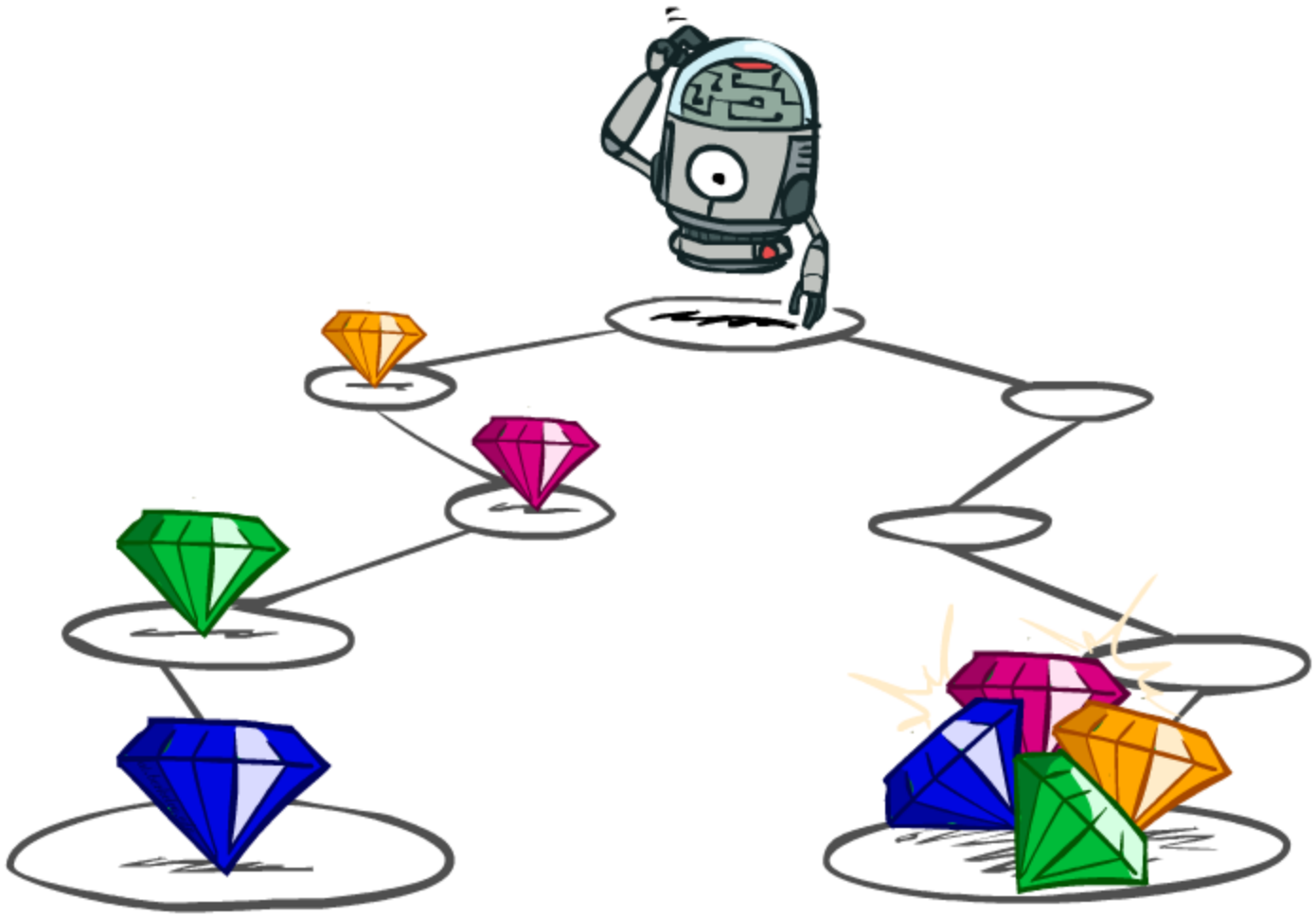
یادگیری تقویتی
(Reinforcement Learning)

1. فرایند تصمیم مارکوف (Markov Decision Processes)
2. الگوریتم تکرار مقدار (Value Iteration)
3. الگوریتم تکرار سیاست (Policy Iteration)
4. یادگیری تقویتی (Reinforcement Learning)

سودمندی (Utility) یک دنباله از عملیات

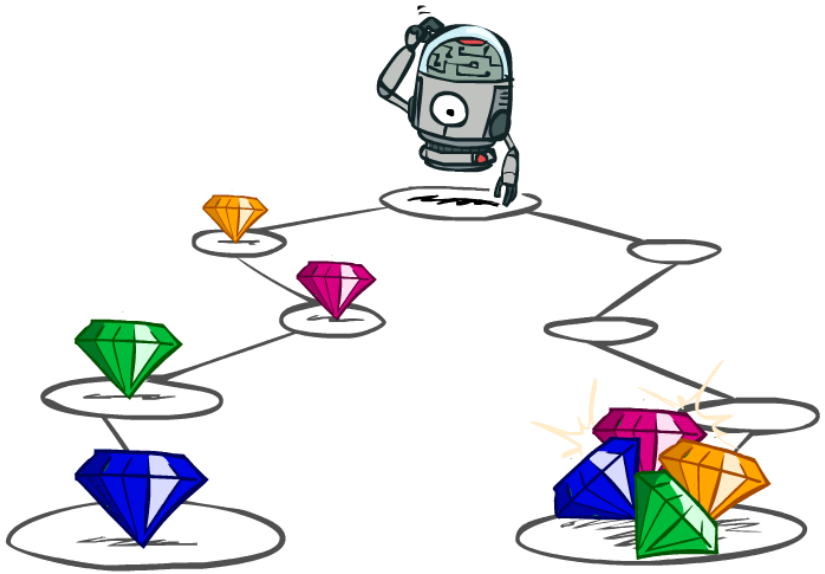


سودمندی یک دنباله از عملیات

- عامل باید کدام دنباله از پاداش‌ها را ترجیح دهد؟

- کمتر یا بیشتر؟ [1, 2, 2] or [2, 3, 4]

- حالا یا بعدا؟ [0, 0, 1] or [1, 0, 0]



کاهش پاداش‌ها (discounting)

- منطقی است که عامل بخواهد مجموع پاداش‌ها رو بیشینه کند.
- همین‌طور منطقی است که پاداش‌های الان رو به پاداش‌های آینده ترجیح دهد.
- یکی از راه‌حل‌ها این است که ارزش پاداش‌ها با گذشت زمان به‌صورت نمایی کاهش پیدا کند.

1

ارزش کنونی

γ

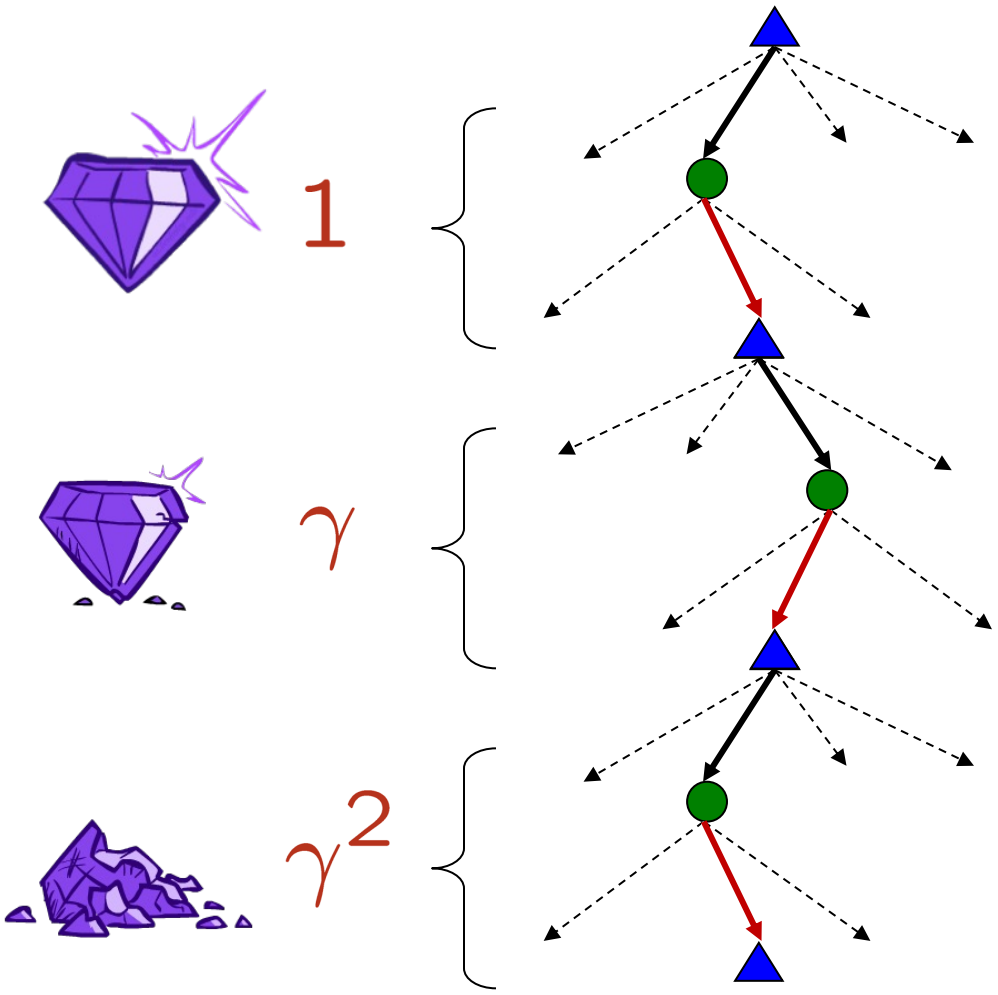
ارزش پس از یک مرحله

γ^2

ارزش پس از دو مرحله

✓ ضریب کاهش = γ

کاهش پاداش‌ها



❑چگونگی کاهش پاداش‌ها

- هر بار که یک سطح پایین‌تر می‌رویم (یعنی یک گام جلوتر در زمان)، یک‌بار ضریب کاهش را در پاداش ضرب می‌کنیم.

❑چرا باید مقدار پاداش را کاهش دهیم؟

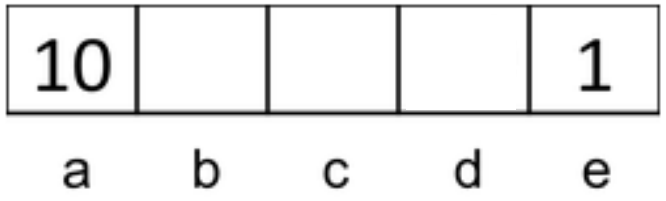
- چون گرفتن پاداش در حال حاضر بهتر از گرفتن همان پاداش در آینده است.
- همچنین می‌توان این‌طور نگاه کرد که در هر گام، با احتمال γ-۱ ممکن است فرآیند تمام شود.
- کاهش مقدار پاداش به همگرایی الگوریتم‌ها کمک می‌کند

❑مثال: ضریب کاهش پاداش= 0.5

- U([1,2,3]) = 1*1 + 0.5*2 + 0.25*3
- U([1,2,3]) < U([3,2,1])

$$U([r_0, r_1, r_2, \ldots]) = r_0 + \gamma r_1 + \gamma^2 r_2 \cdots$$

کوئیز: کاهش پاداش‌ها



❑داده‌های مسئله:

- اعمال: شرق، غرب و خروج (فقط در حالت‌های a و e قابل انجام است)
- با انجام یک عمل قطعا به حالت مشخصی می‌رویم (Transitions deterministic)

➢کوئیز۱: برای $\gamma = 1$, سیاست بهینه چیست؟

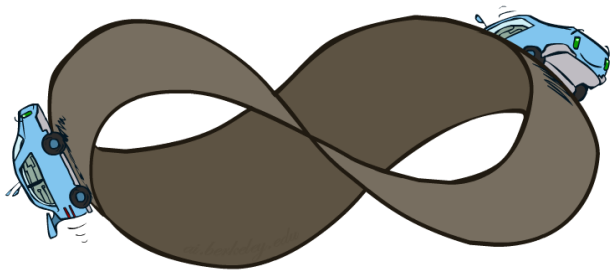
➢کوئیز۲: برای $\gamma = 0.1$, سیاست بهینه چیست؟

➢کوئیز ۳: در حالتd، γ چه مقداری باشد تا به سمت شرق یا غرب حرکت شود سودمندی یکسان باشد؟

$1=10\,\gamma^2$ ⟶ $\gamma = \sqrt{\frac{1}{10}} \approx 0.316$

سودمندی نامتناهی؟!

❑ مسئله: اگر بازی تا ابد ادامه پیدا کند، آیا پاداش بی‌نهایت دریافت خواهیم کرد؟



- راه‌حل‌ها:

1. افق متناهی
   - مشابه جستجوی عمقی محدود
   - عامل فقط برای تعداد مشخصی از مراحل (مثلاً 100 مرحله) تصمیم‌گیری می‌کند.

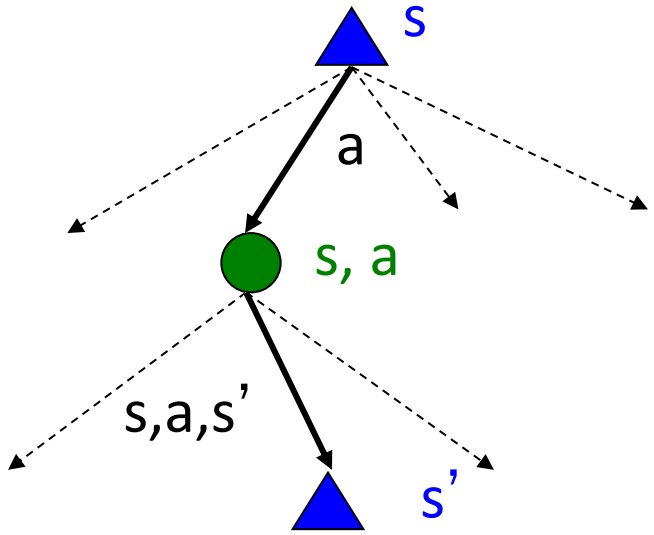
2. استفاده از ضریب کاهش پاداش: $0 < \gamma < 1$
   - ضریب γ کوچکتر = افق محدودتر = تمرکز بر پاداش‌های نزدیک‌تر

$$U([r_0, \ldots r_\infty]) = \sum_{t=0}^{\infty} \gamma^t r_t \leq R_{\max}/(1-\gamma)$$

سری هندسی با قدر نسبت γ : $1 + \gamma + \gamma^2 + \cdots = \frac{1}{1-\gamma}$

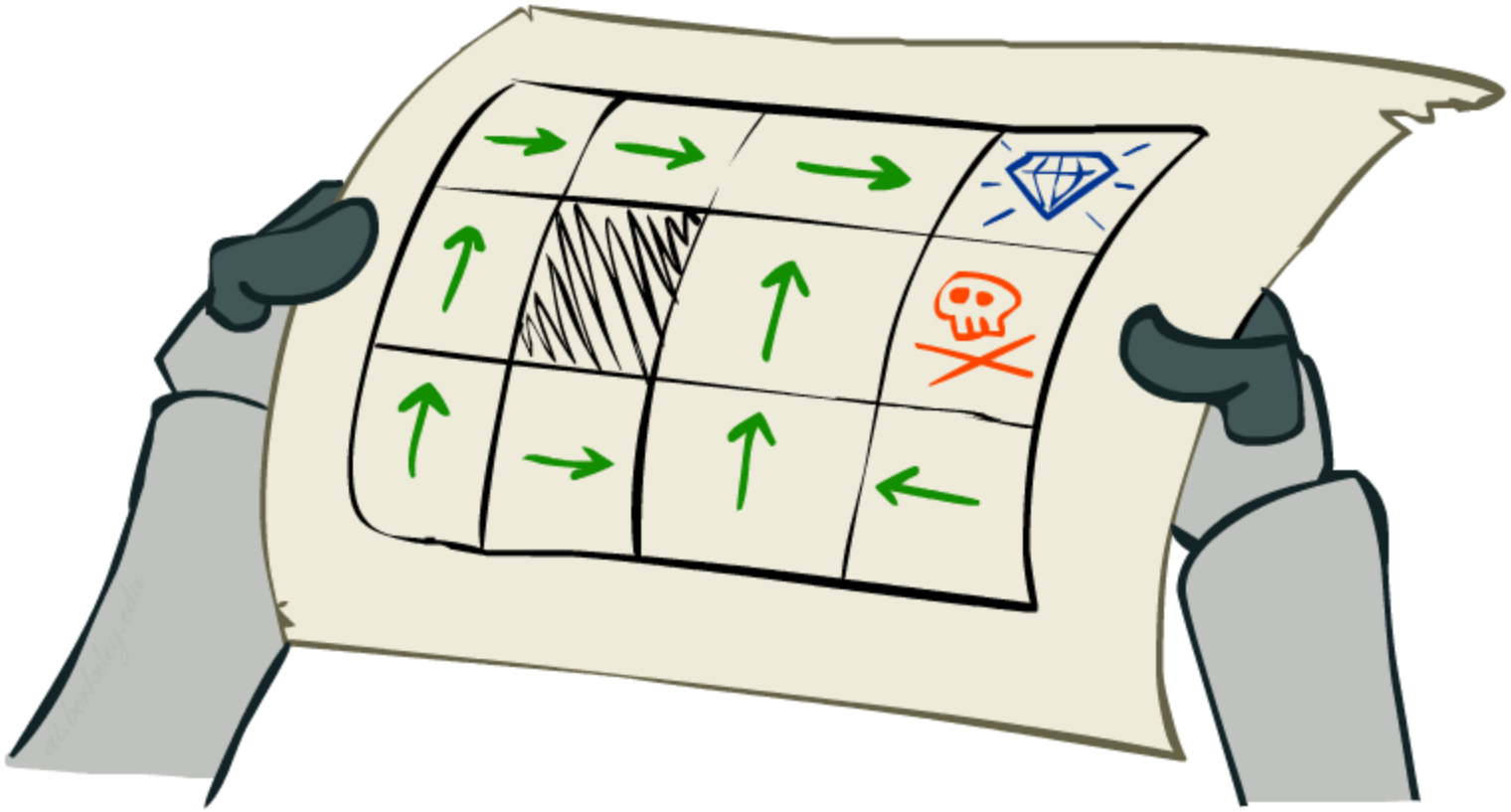
3. اطمینان حاصل می‌کنیم که تحت هر سیاستی، در نهایت به یک حالت پایانی خواهیم رسید.
   - مانند حالت «جوش» در مسابقه‌ی اتومبیل رانی

مرور: فرایند تصمیم مارکوف

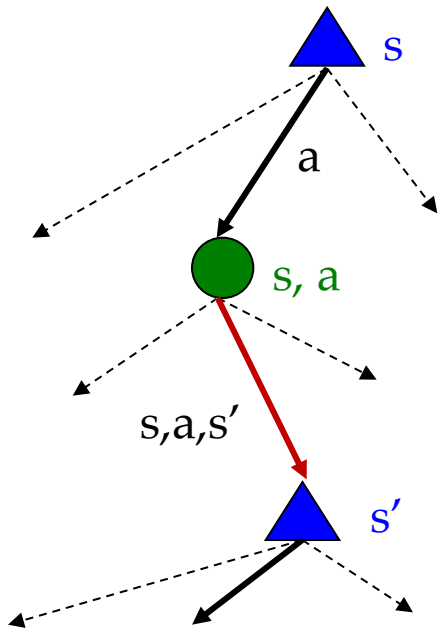


- یک MDP به صورت زیر تعریف می‌شود:
  - یک مجموعه از حالت‌ها S
  - یک مجموعه از اعمال A
  - یک تابع تغییر حالت $T(s, a, s')$ (یا $P(s'|s, a)$)
  - یک تابع پاداش $R(s, a, s')$ (و ضریب کاهش پاداش γ)
  - یک حالت شروع s_0
- دو کمیت مهم MDP تا اینجا:
  - **سیاست (Policy):** انتخاب یک عمل برای هر حالت
  - **سودمندی (Utility):** مجموع (کاهش یافته) پاداش‌ها

حل مسائل MDP



کمیت‌های بهینه



s is a *state*

(s, a) is a *q-state*

(s,a,s') is a *transition*

- ارزش (یا سودمندی) یک حالت s

  $V^*(s)$ = سودمندی مورد انتظار با شروع از s و بهینه عمل کردن.

- ارزش (یا سودمندی) یک حالت q

  $Q^*(s,a)$ = سودمندی مورد انتظار با شروع از s و انتخاب عمل a و بهینه عمل کردن

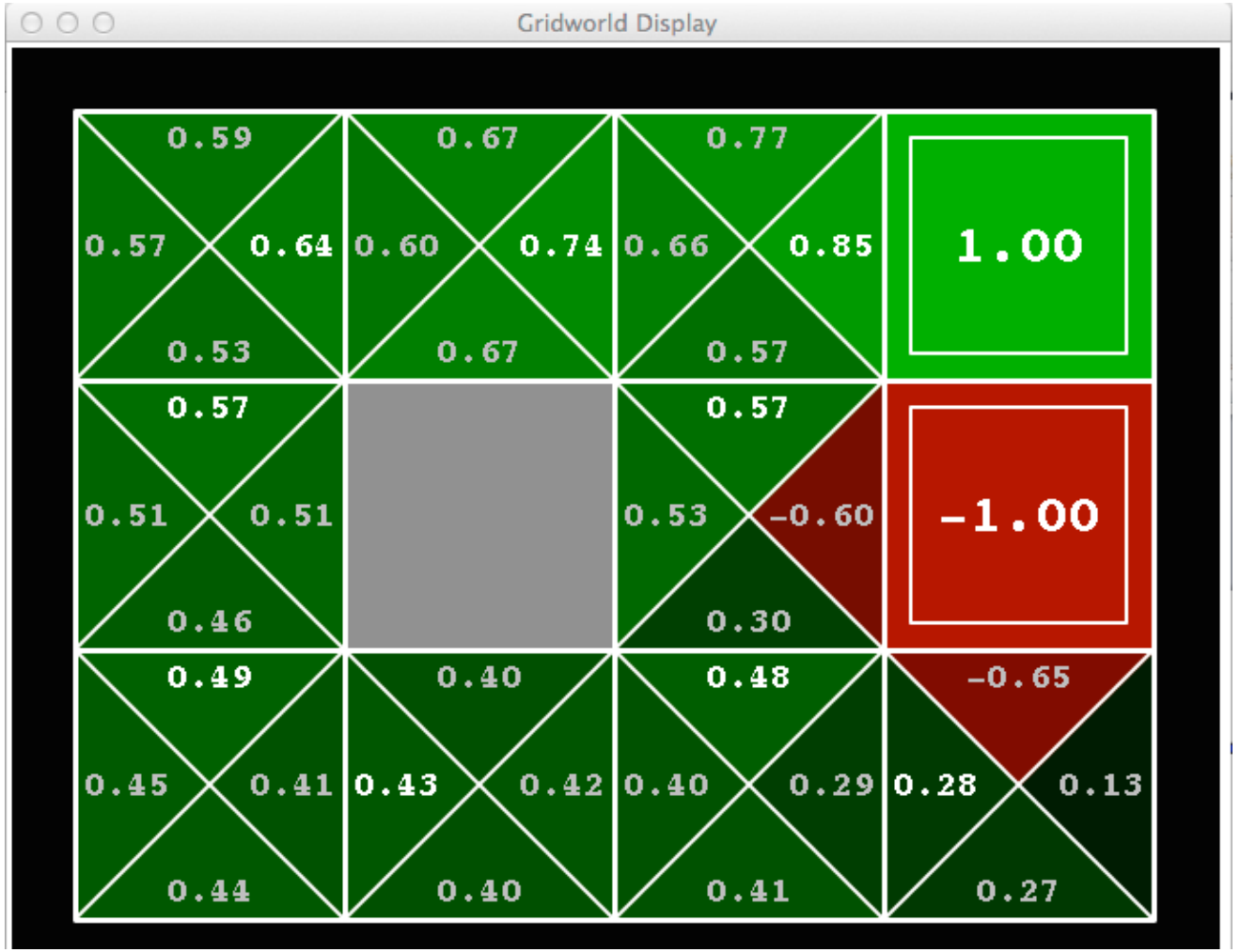
- سیاست بهینه

  $\pi^*(s)$ = عمل بهینه در حالت s

V* Values

نویز = ۰/۲
ضریب کاهش = ۰/۹
پاداش مرحله‌ای = ۰

Q* Values



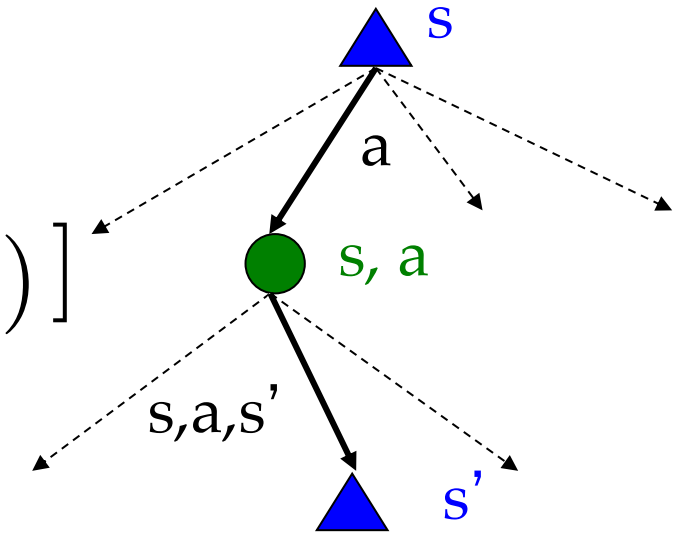
نویز = ۰/۲
ضریب کاهش = ۰/۹
پاداش مرحله‌ای = ۰

محاسبه V* و Q*

- تعریف بازگشتی V*:

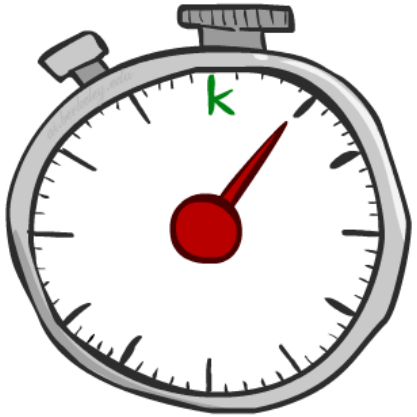
$$V^*(s) = \max_a Q^*(s,a)$$

$$Q^*(s,a) = \sum_{s'} T(s,a,s')\left[R(s,a,s') + \gamma V^*(s')\right]$$

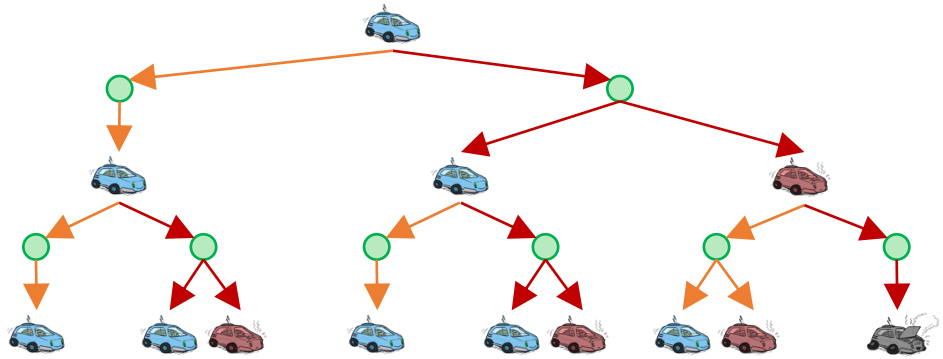


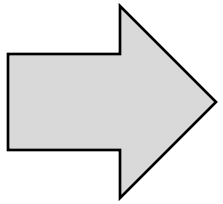
$$V^*(s) = \max_a \sum_{s'} T(s,a,s')\left[R(s,a,s') + \gamma V^*(s')\right]$$

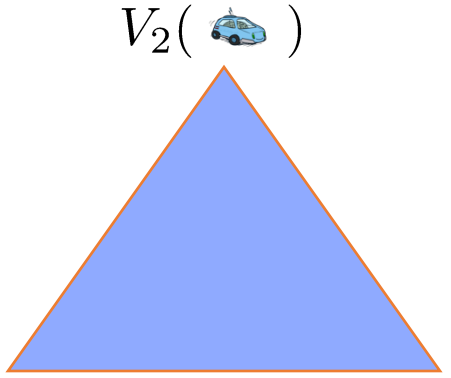
ارزش‌های با زمان محدود



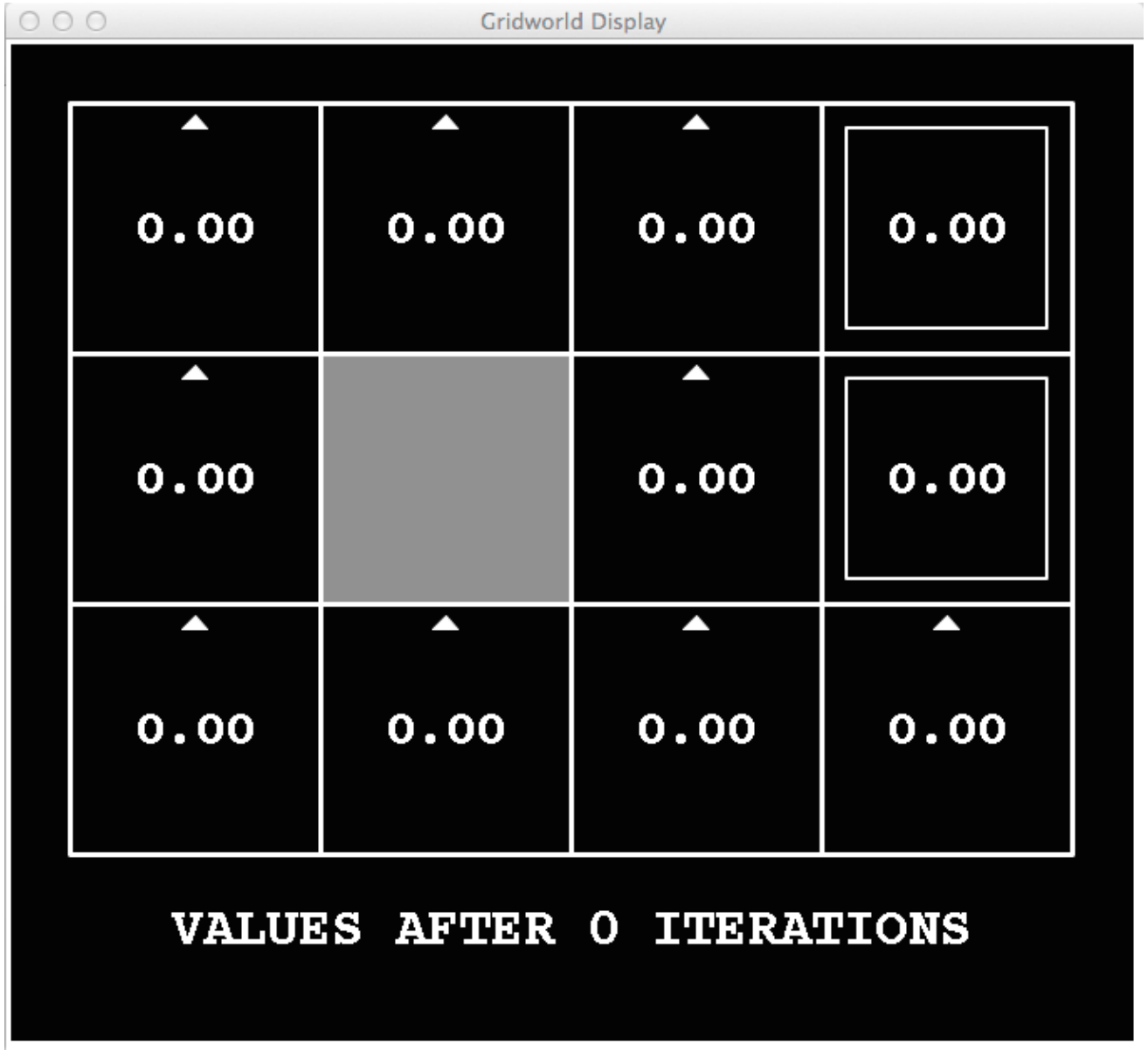
- ایده‌ی کلیدی: ارزش‌های با زمان محدود!
- تعریف $V_k(s)$ به عنوان ارزش بهینه s اگر بازی پس از k مرحله‌ی دیگر خاتمه یابد.





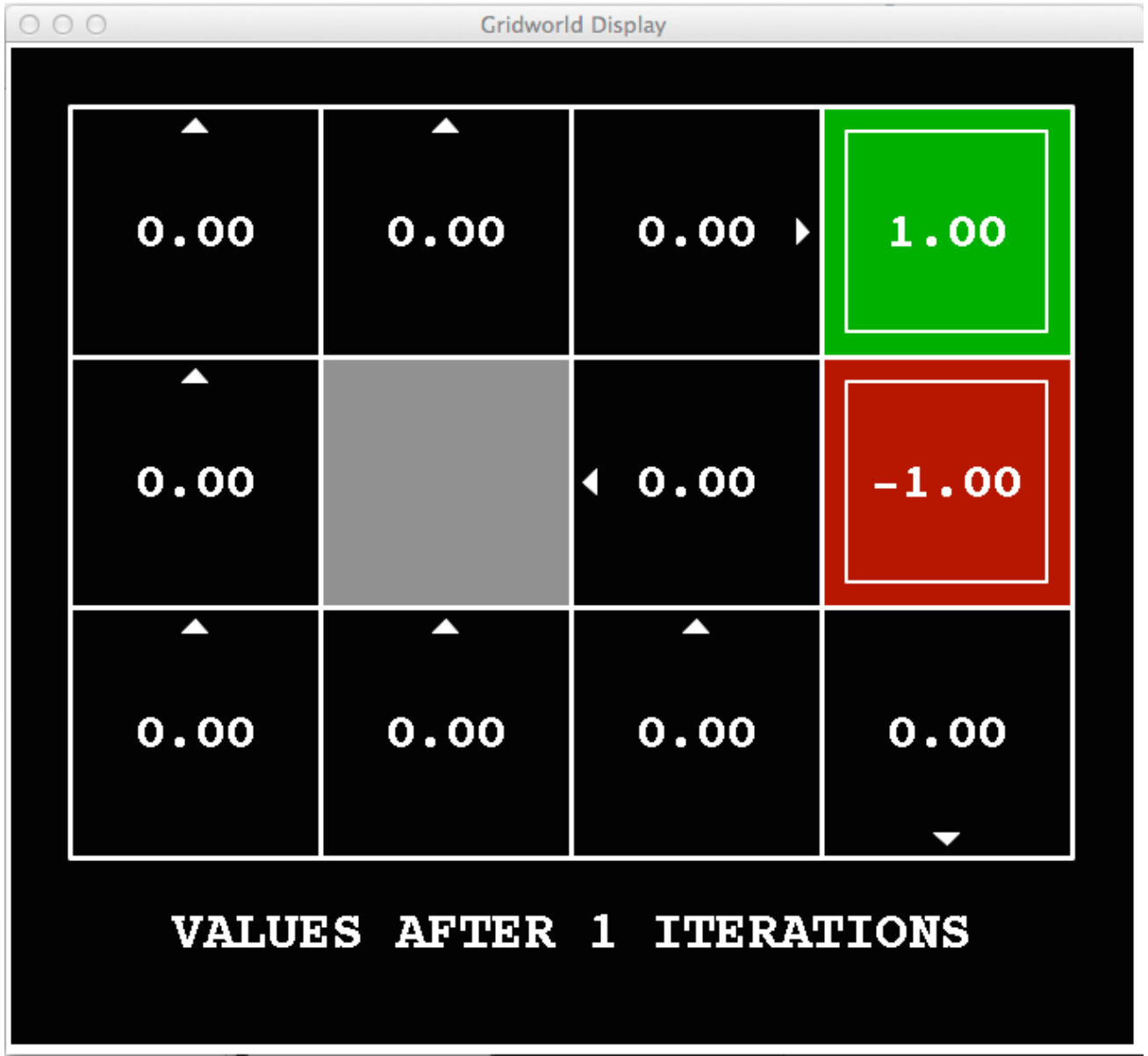


k=0



نویز = ۰/۲
ضریب کاهش = ۰/۹
پاداش مرحله‌ای = ۰

k=1



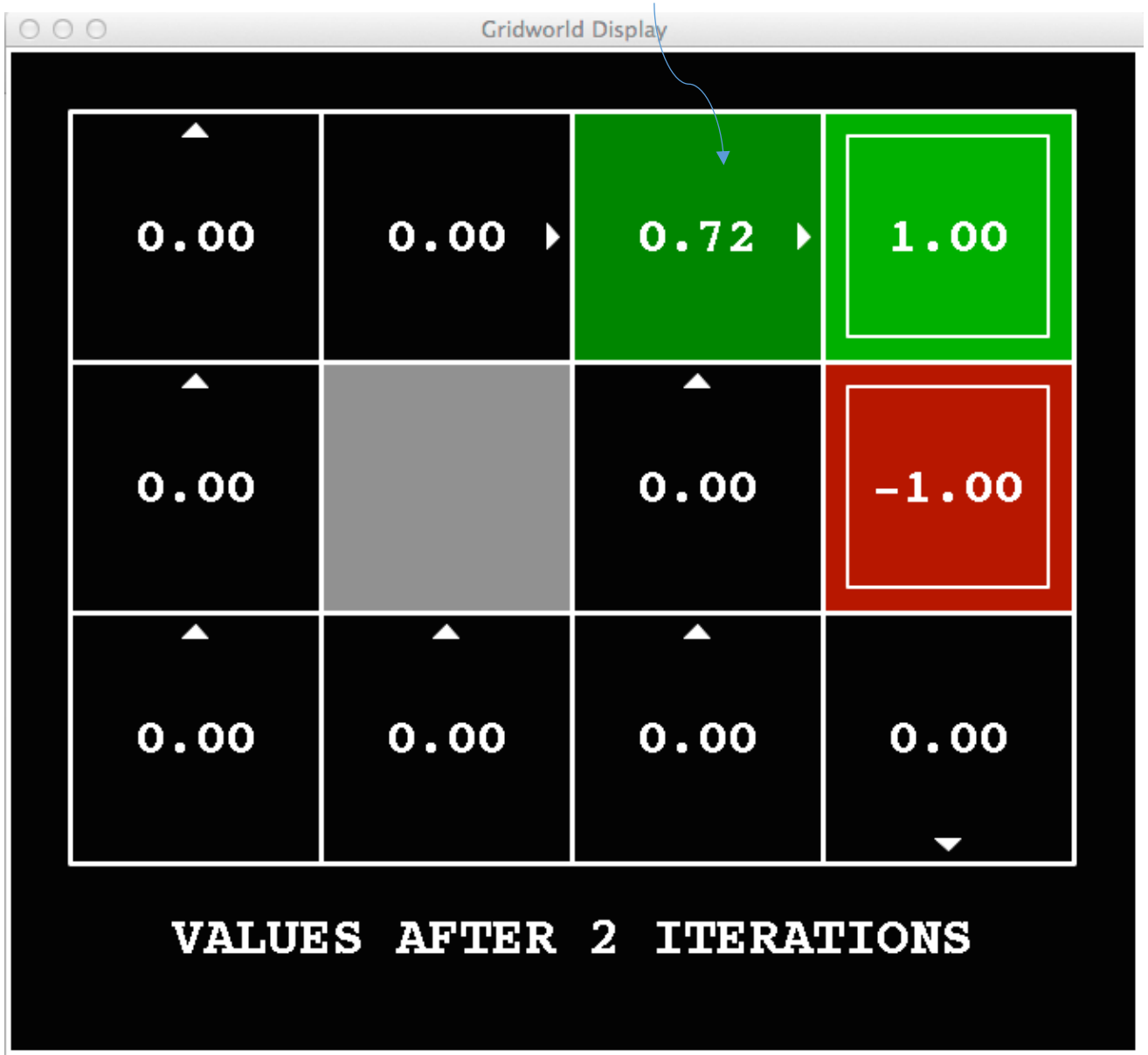
نویز = ۰/۲
ضریب کاهش = ۰/۹
پاداش مرحله‌ای = ۰

$$V^*(s) = \max_a \sum_{s'} T(s,a,s')[R(s,a,s') + \gamma V^*(s')]$$

$$V^{(2)}(s) = 0.8 \cdot (0 + 0.9 \cdot 1.00) + 0.1 \cdot (0 + 0.9 \cdot 0.00) + 0.1 \cdot (0 + 0.9 \cdot 0.00) = 0.72$$

For action = right

k=2



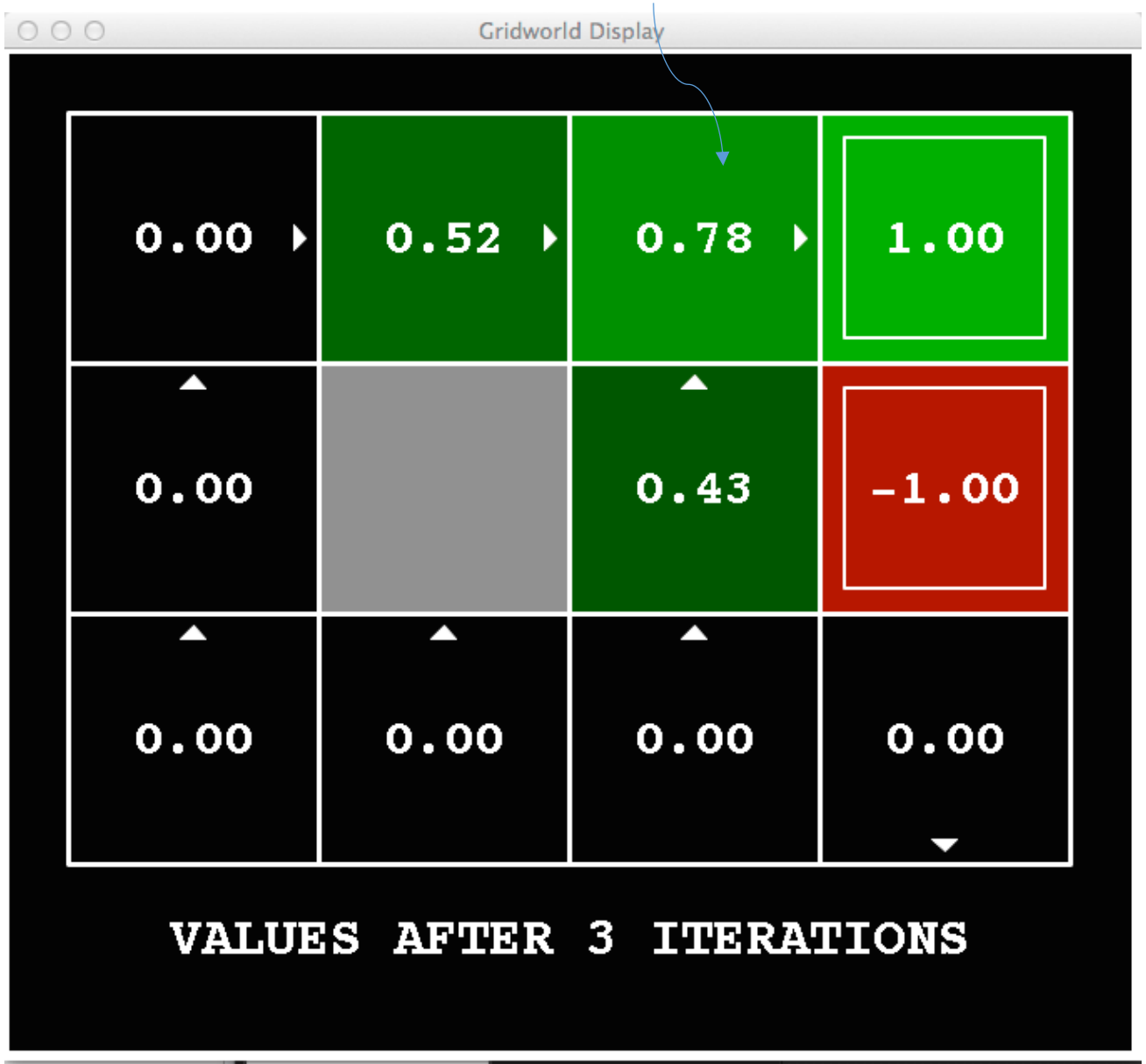
نویز = ۰/۲
ضریب کاهش = ۰/۹
پاداش مرحله‌ای = ۰

$$V^*(s) = \max_a \sum_{s'} T(s,a,s')[R(s,a,s') + \gamma V^*(s')]$$

$$V^{(3)}(s) = 0.8 \cdot (0 + 0.9 \cdot 1.00) + 0.1 \cdot (0 + 0.9 \cdot 0.72) + 0.1 \cdot (0 + 0.9 \cdot 0.00) = 0.7848$$

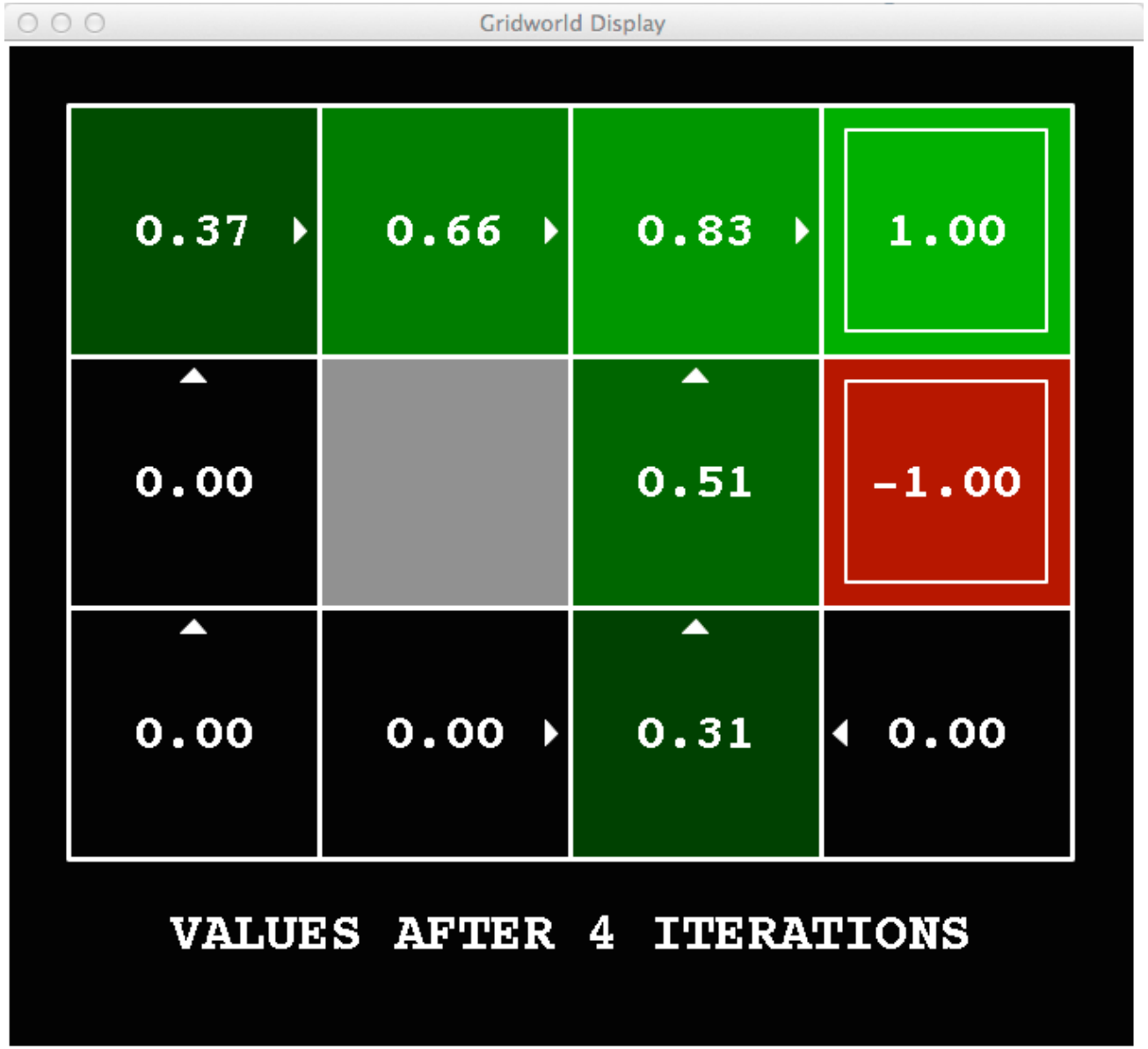
For action = right

k=3



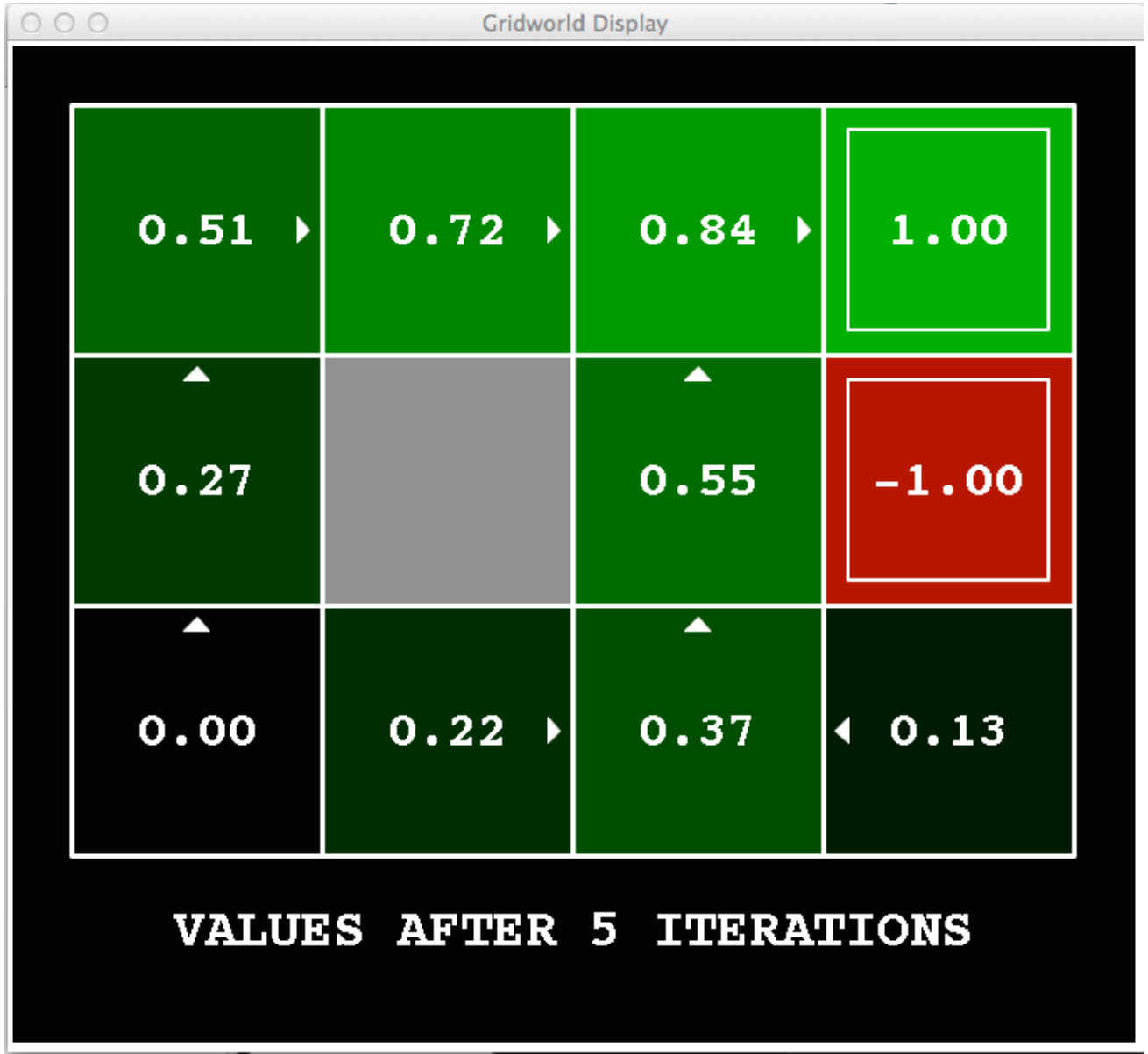
نویز = ۰/۲
ضریب کاهش = ۰/۹
پاداش مرحله‌ای = ۰

k=4



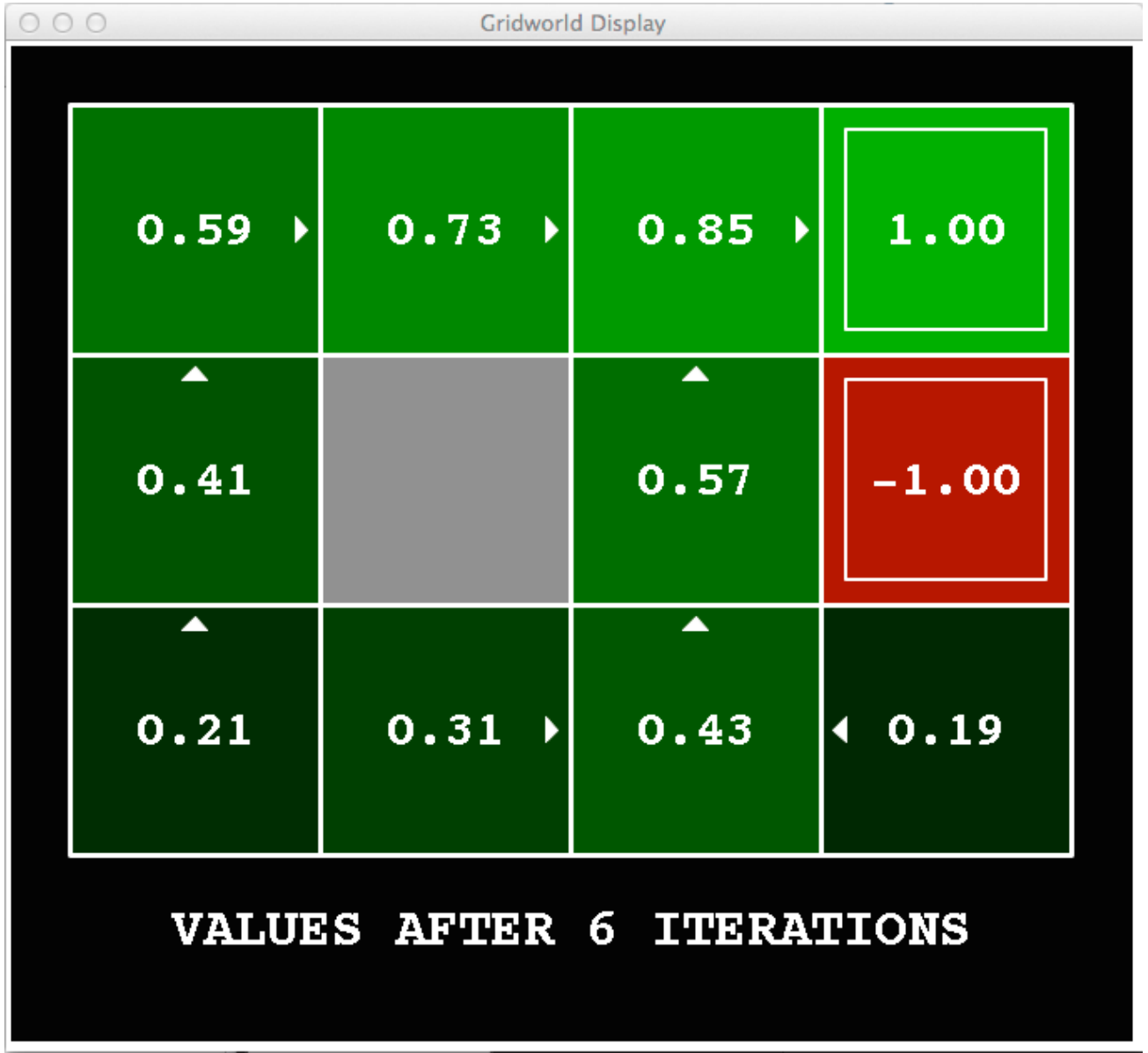
نویز = ۰/۲
ضریب کاهش = ۰/۹
پاداش مرحله‌ای = ۰

k=5



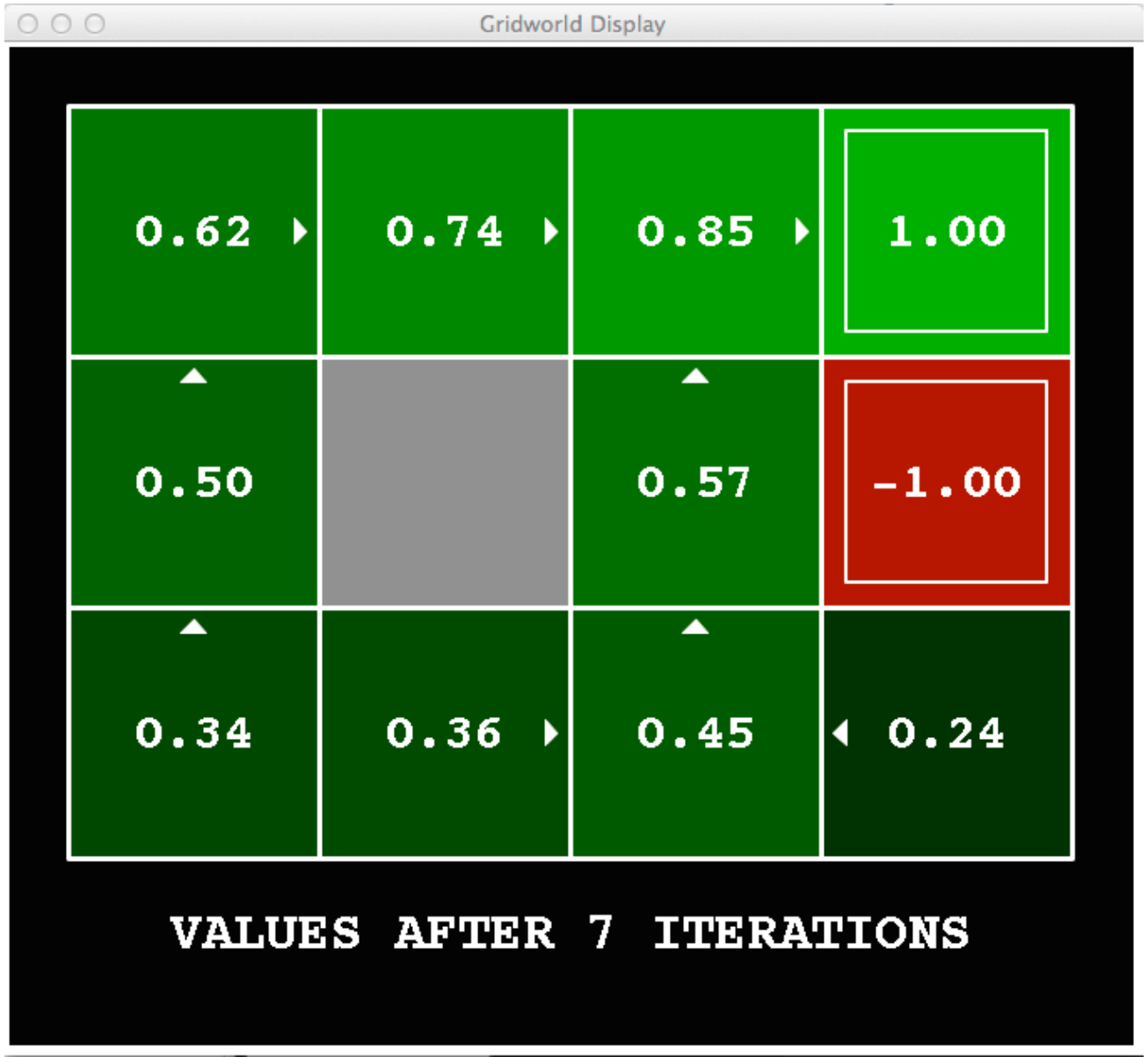
نویز = ۰/۲
ضریب کاهش = ۰/۹
پاداش مرحله‌ای = ۰

k=6



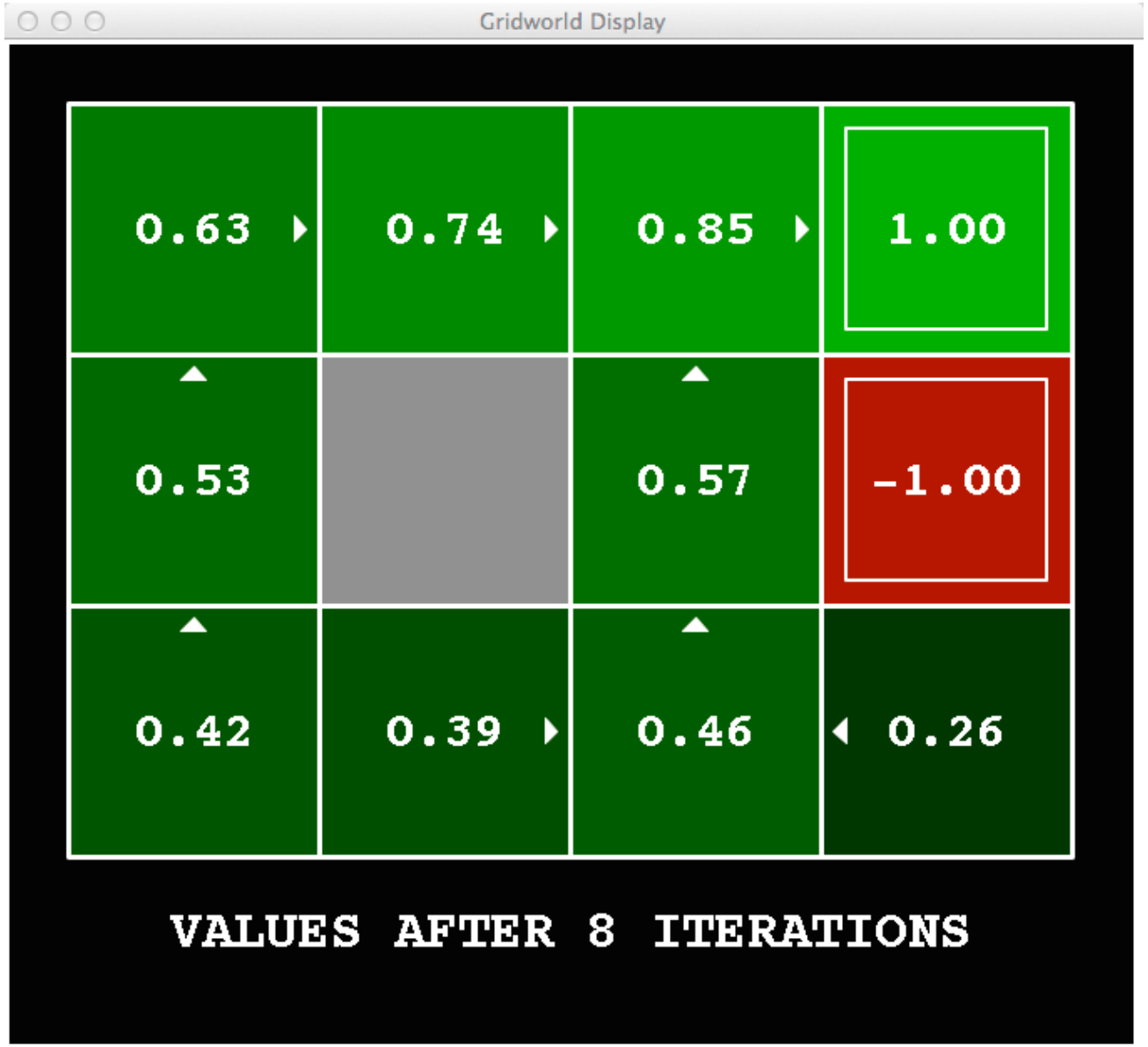
نویز = ۰/۲
ضریب کاهش = ۰/۹
پاداش مرحله‌ای = ۰

k=7



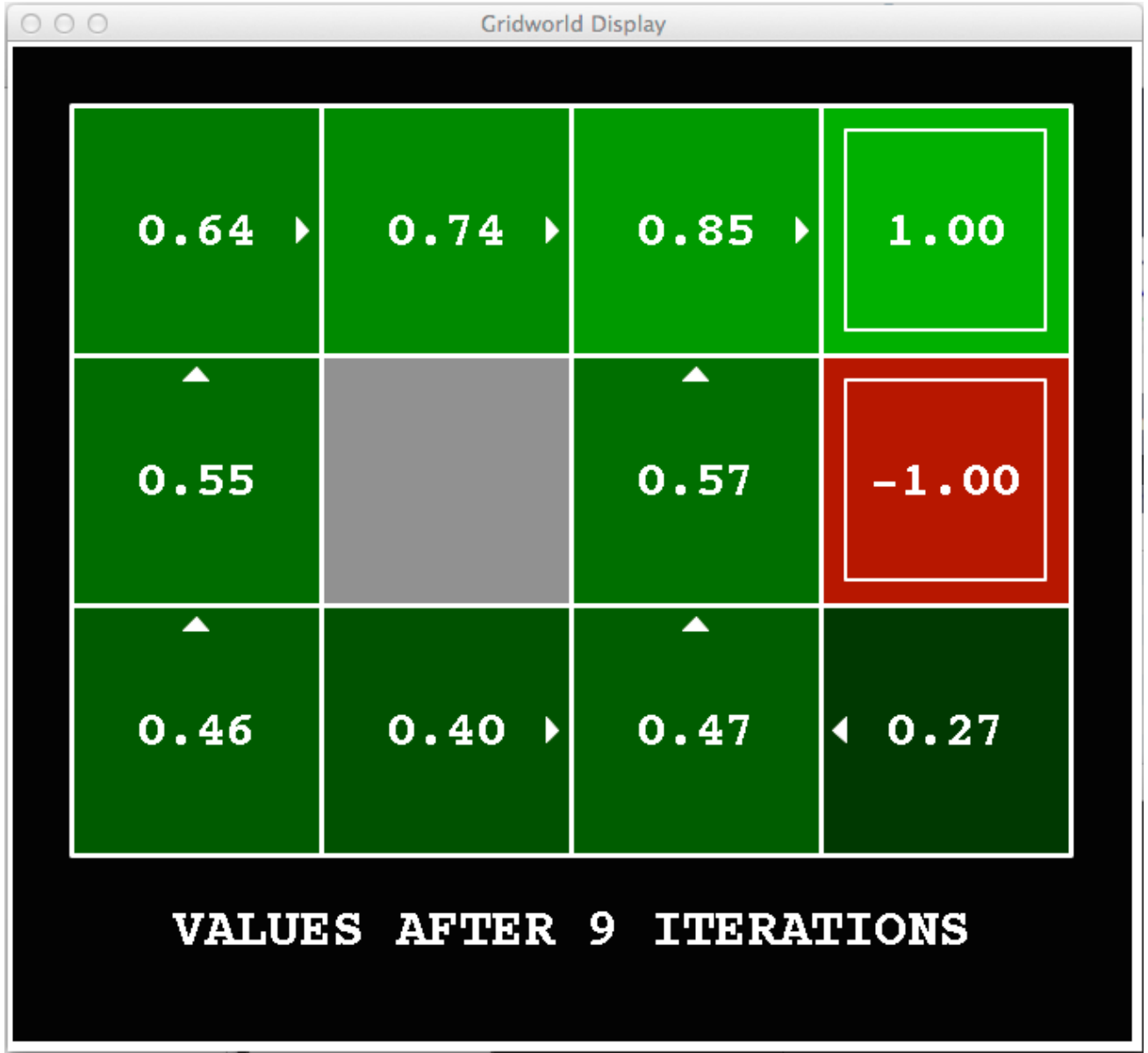
نویز = ۰/۲
ضریب کاهش = ۰/۹
پاداش مرحله‌ای = ۰

k=8



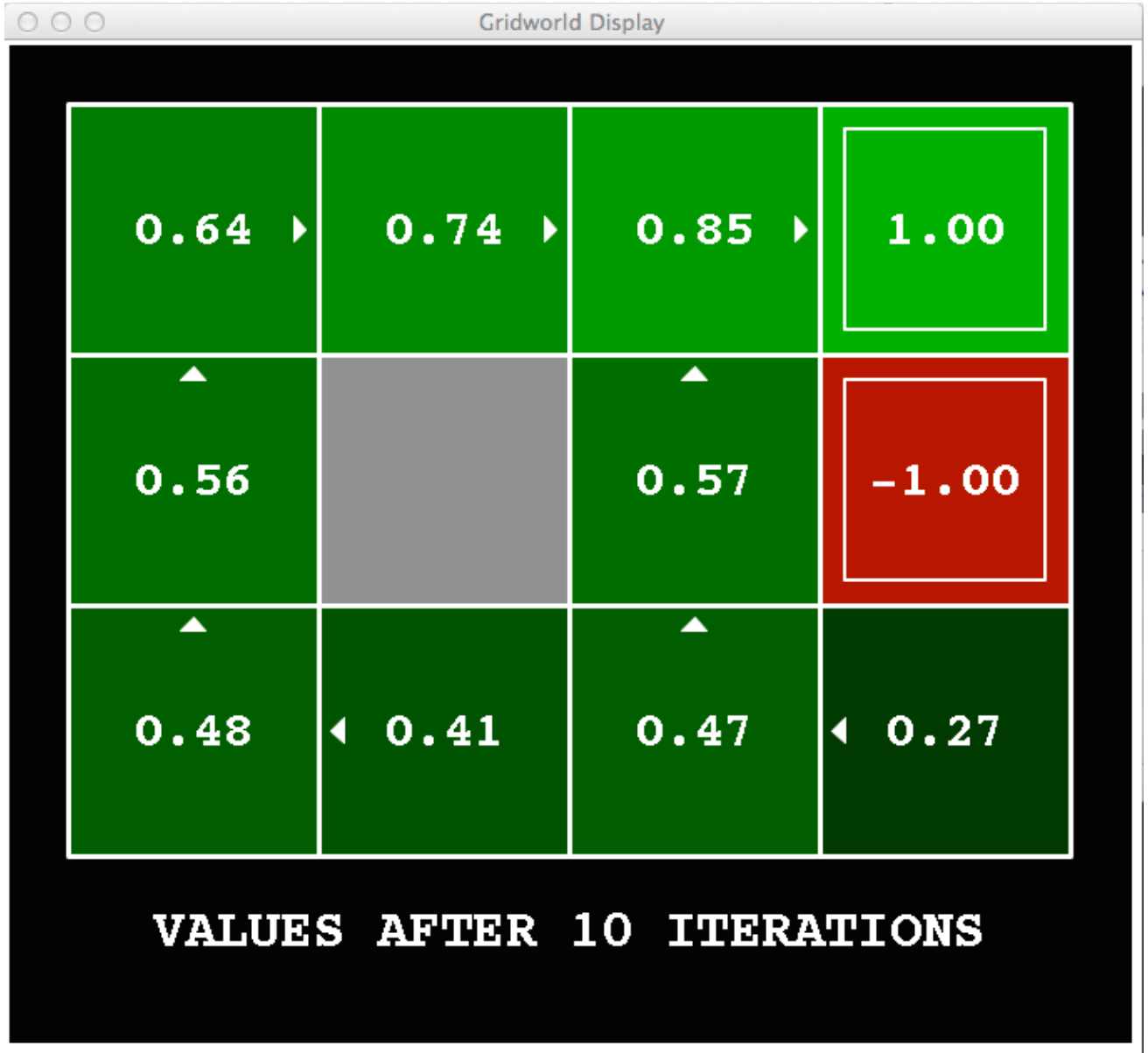
نویز = ۰/۲
ضریب کاهش = ۰/۹
پاداش مرحله‌ای = ۰

k=9



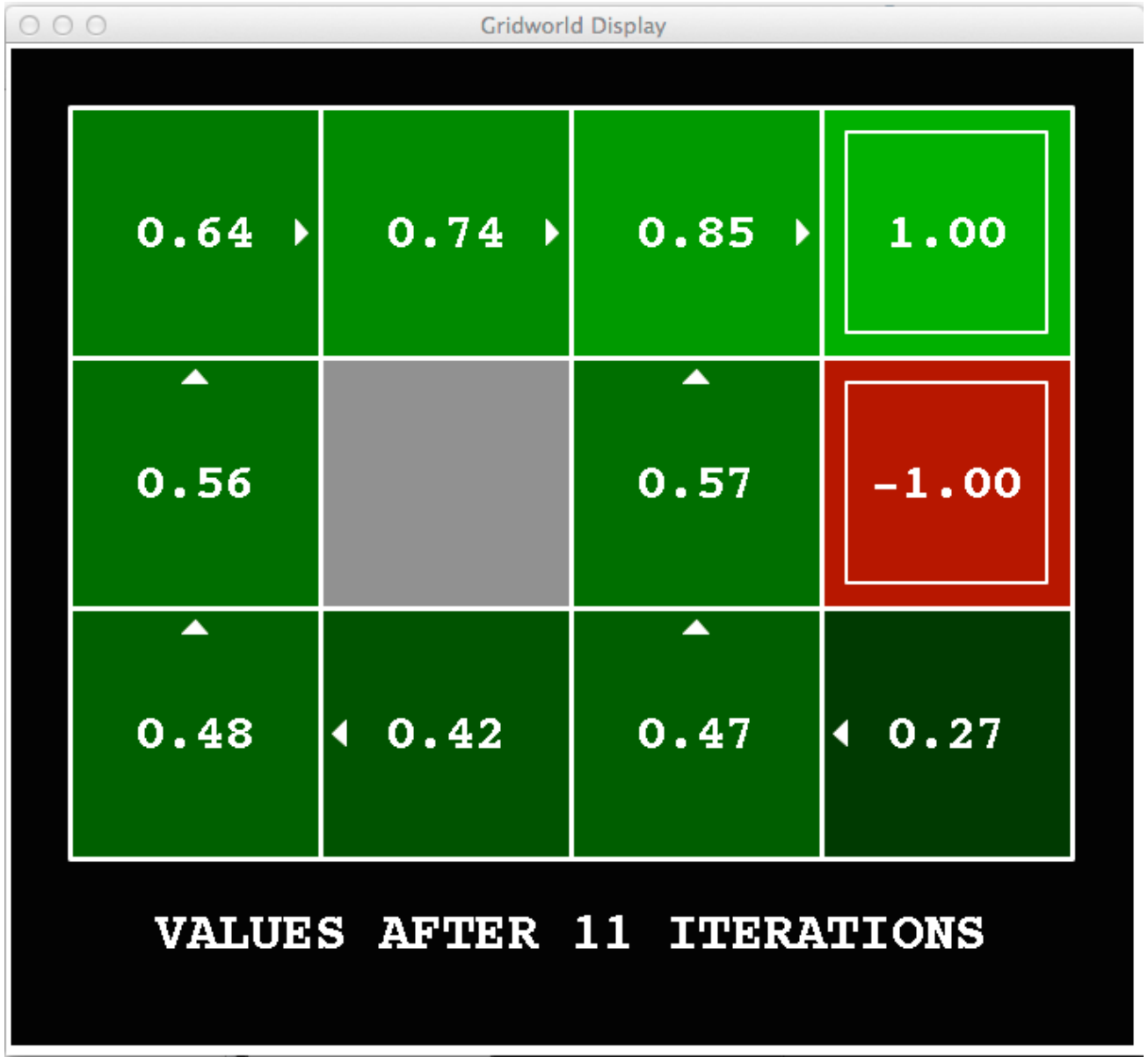
نویز = ۰/۲
ضریب کاهش = ۰/۹
پاداش مرحله‌ای = ۰

k=10



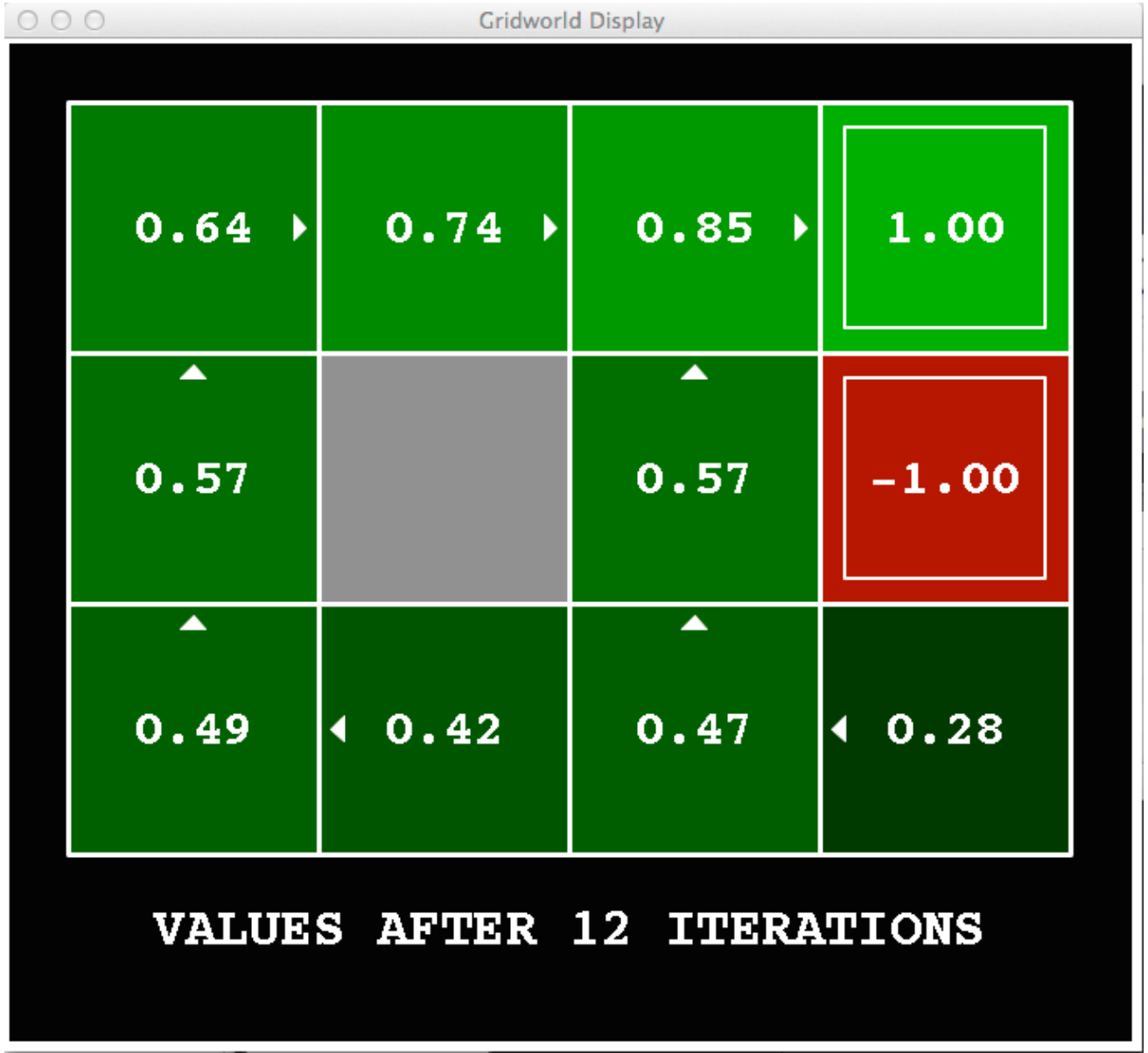
نویز = ۰/۲
ضریب کاهش = ۰/۹
پاداش مرحله‌ای = ۰

k=11



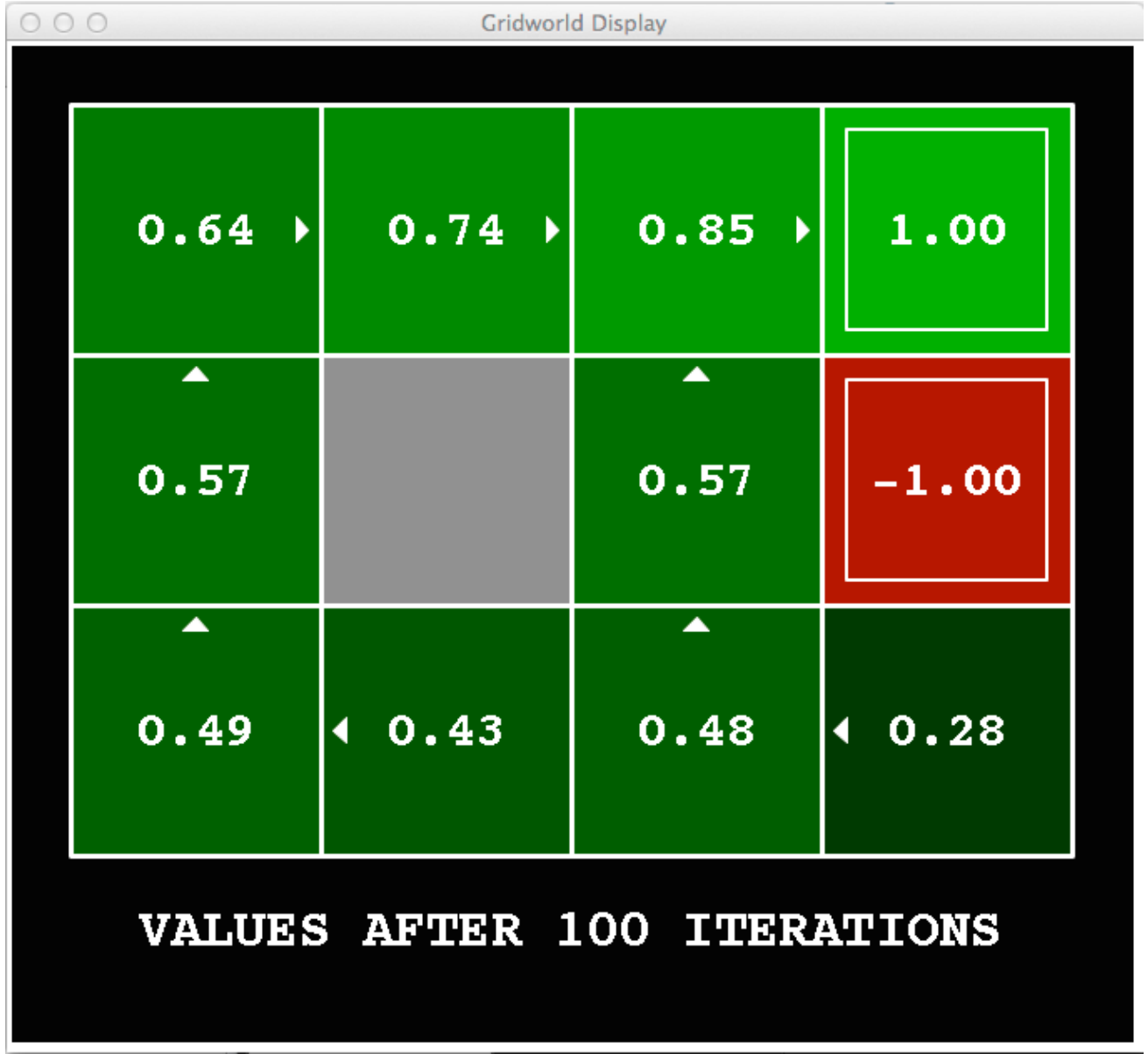
نویز = ۰/۲
ضریب کاهش = ۰/۹
پاداش مرحله‌ای = ۰

k=12

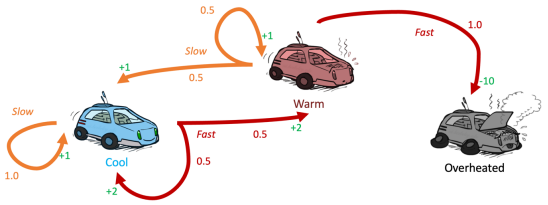


نویز = ۰/۲
ضریب کاهش = ۰/۹
پاداش مرحله‌ای = ۰

k=100



نویز = ۰/۲
ضریب کاهش = ۰/۹
پاداش مرحله‌ای = ۰



محاسبه‌ی ارزش‌های با زمان محدود

$V_4(\ )$ $V_4(\ )$ $V_4(\ )$

$V_3(\ )$ $V_3(\ )$ $V_3(\ )$

$V_2(\ )$ $V_2(\ )$ $V_2(\ )$

$V_1(\ )$ $V_1(\ )$ $V_1(\ )$

$V_0(\ )$ $V_0(\ )$ $V_0(\ )$