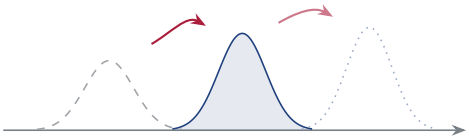
14

Policy Gradient Methods



The methods of the previous chapters learn values and derive policies from them by greediness. One may instead search the policies directly: parametrize them, measure their performance, and climb the gradient. The gradient of the expected return with respect to the policy parameters has a remarkably simple form—the policy gradient theorem—which can be estimated from sampled trajectories by weighting the scores of the actions taken by the returns that followed. This chapter proves the theorem, explains the variance-reduction devices that make it usable (baselines, critics), and studies the tabular softmax case in detail, where the natural *gradient turns out to be a multiplicative-weights update on the advantages and converges to the optimum at a rate that does not depend on the size of the state space—the theoretical core of the trust-region methods used in practice. The chapter ends where the distributional viewpoint takes over: the gradient of* any *functional of the return distribution can be written by the same likelihood-ratio trick, and for exponential utilities it takes a Bellman form through exponentially tilted occupation measures, so that risk-sensitive policies can be trained by the same machinery, at the price of variance.*

14.1 Parametrized policies and the performance criterion

We work in a discounted MDP $\mathcal{M} = (\mathcal{S}, \mathcal{A}, P, r, \gamma)$ with finite $\mathcal{S}$ and $\mathcal{A}$ and rewards in $[0, R_{\max}]$, and, when the horizon matters (Section 14.4), in the finite-horizon MDPs of Chapter 3. A *parametrized policy* is a family $(\pi_\theta)_{\theta \in \Theta}$, $\Theta \subset \mathbb{R}^d$ open, of stationary randomized policies such that $\theta \mapsto \pi_\theta(a \mid s)$ is differentiable and positive for all (s, a), so that the *score function* $\nabla_\theta \ln \pi_\theta(a \mid s)$ is defined. Three families cover most uses:

- the *tabular softmax*, $\pi_\theta(a \mid s) \propto e^{\theta(s,a)}$ with $\theta \in \mathbb{R}^{\mathcal{S} \times \mathcal{A}}$, whose score is $\nabla_{\theta(s,a')} \ln \pi_\theta(a \mid s) = \mathbb{1}\{a = a'\} - \pi_\theta(a' \mid s)$: every randomized stationary policy with positive probabilities is of this form, and deterministic policies are reached in the limit $\|\theta\| \to \infty$;
- the *log-linear* policies $\pi_\theta(a \mid s) \propto e^{\theta^\top \varphi(s,a)}$ on features φ, and the Gaussian policies $\pi_\theta(\cdot \mid s) = \mathcal{N}(m_\theta(s), \sigma_\theta(s)^2)$ for continuous actions;
- neural networks with a softmax output, as in session H06, where a network with a hidden layer of 128 units maps the four-dimensional state of the cart-pole to the probabilities of its two actions.

The *performance* of θ from an initial distribution μ is

$$J(\theta) \coloneqq V^{\pi_\theta}(\mu) = \sum_s \mu(s) V^{\pi_\theta}(s) = \mathbb{E}_\mu^{\pi_\theta}\Big[\sum_{t \geq 1} \gamma^{t-1} R_t\Big],$$

and policy gradient methods perform stochastic gradient *ascent* on J. Why search the policy rather

than the value? Because the action space may be continuous or huge, so that the maximization in the Bellman operator is itself a problem; because the optimal policy may have to be stochastic (in games, or under partial observability); because J is a smooth function of θ whereas greedy policies jump; and because a good parametrization encodes prior knowledge about the shape of good policies (Example 14.6: two parameters for the machine of Chapter 13). The price is that J is not concave, so that gradient ascent may stop at stationary points that are not optimal—and that the natural estimates of the gradient have a large variance.

Criteria, and two regimes. The chapter uses three criteria, each where it serves best. The discounted performance J above is the setting of the theory: the policy gradient theorem, the natural gradient and the trust-region bound are discounted statements. The algorithms are run on *episodes* of length H, so that estimates are written in the finite-horizon form, with time-indexed policies when needed; the discounted and episodic forms are related by truncation and by the weights γ^{t-1} (Exercise 14.4). The risk criteria of Section 14.4 are finite-horizon functionals of the return, as in Chapter 5. The chapter also runs through the two regimes of Chapter 13 ("One operator, four regimes"). *With the model*, the gradient is an exact quantity: d_μ^π and Q^π are computed, and gradient ascent is a *planning* method, an alternative to value and policy iteration whose convergence can be studied as such—this is what the theorems and the worked examples of Section 14.3 do, and what the machine, the regulator and the hike illustrate. *From episodes*, the same gradient is estimated by sampling, and the estimate's variance becomes the whole problem: REINFORCE, actor–critic and the proximal methods are the learning side, and every worked example runs both so that the two can be compared.

Occupation measures. Recall the discounted occupation measure of π from μ, (8.1) in Chapter 8: $d_\mu^\pi(s) = (1-\gamma)\sum_{t\geq 1}\gamma^{t-1}\mathbb{P}_\mu^\pi(S_t = s)$, $d_\mu^\pi(s,a) = d_\mu^\pi(s)\pi(a \mid s)$, and $V^\pi(\mu) = \frac{1}{1-\gamma}\sum_{s,a} d_\mu^\pi(s,a)r(s,a)$ with $r(s,a) = \sum_{s'} P(s' \mid s,a)r(s,a,s')$ the mean one-step reward. The comparison of two policies is governed by the advantage $A^\pi(s,a) = Q^\pi(s,a) - V^\pi(s)$ of one under the occupation measure of the other.

Lemma 14.1 (Performance difference, Kakade and Langford, 2002). For all policies $\pi, \pi' \in \Pi_{\mathrm{SR}}$ and $\mu \in \mathcal{M}_1(\mathcal{S})$,

$$V^{\pi'}(\mu) - V^\pi(\mu) \;=\; \frac{1}{1-\gamma}\sum_{s,a} d_\mu^{\pi'}(s,a)\,A^\pi(s,a) \;=\; \frac{1}{1-\gamma}\,\mathbb{E}_{s\sim d_\mu^{\pi'},\; a\sim\pi'(\cdot|s)}\big[A^\pi(s,a)\big].$$

In a finite-horizon MDP, for Markov policies π, π' and $A_t^\pi \coloneqq Q_t^\pi - V_t^\pi$, $V_1^{\pi'}(\mu) - V_1^\pi(\mu) = \sum_{t\leq H}\mathbb{E}_\mu^{\pi'}\big[A_t^\pi(S_t,A_t)\big]$.

Proof. Along a trajectory of π' from $S_1 \sim \mu$, $A^\pi(S_t,A_t) = Q^\pi(S_t,A_t) - V^\pi(S_t)$ and $\mathbb{E}[Q^\pi(S_t,A_t) \mid S_t,A_t] = \mathbb{E}[R_t + \gamma V^\pi(S_{t+1}) \mid S_t,A_t]$ by the Bellman equation for Q^π, whatever the policy that chose A_t. Hence, by (8.1),

$$\frac{1}{1-\gamma}\sum_{s,a} d_\mu^{\pi'}(s,a)A^\pi(s,a) = \mathbb{E}_\mu^{\pi'}\Big[\sum_{t\geq 1}\gamma^{t-1}\big(R_t + \gamma V^\pi(S_{t+1}) - V^\pi(S_t)\big)\Big] = V^{\pi'}(\mu) - V^\pi(\mu),$$

the sum telescoping: the terms $\gamma^t V^\pi(S_{t+1})$ cancel, $\gamma^t V^\pi(S_{t+1}) \to 0$, and the expectation of $V^\pi(S_1)$ is $V^\pi(\mu)$. The finite-horizon identity is proved in the same way, with $\gamma = 1$ and the telescoping sum stopping at $V_{H+1}^\pi \equiv 0$. ∎

The lemma is the tool of this chapter: it expresses the improvement brought by a new policy in terms of the advantages of the old one, weighted by where the *new* policy goes. It also gives a two-line proof of Theorem 8.5 and of policy improvement (Exercise 14.1).

14.2 The policy gradient theorem

The gradient of an expectation with respect to a parameter of the sampling law is the expectation of the quantity times the score of the sample. In our setting the law of a trajectory (Theorem 1.14) factorizes into the policy's probabilities, which depend on θ, and the kernel's, which do not.

Lemma 14.2 (Score-function identity). Let f be a bounded function of the trajectory $\tau = (S_1, A_1, \dots, S_H, A_H, S_{H+1})$ of a finite-horizon MDP with time-indexed policy $(\pi_{\theta,t})_{t \le H}$ (a stationary policy being the case $\pi_{\theta,t} = \pi_\theta$). Then

$$\nabla_\theta \, \mathbb{E}_\mu^{\pi_\theta}\big[f(\tau)\big] \;=\; \mathbb{E}_\mu^{\pi_\theta}\Big[f(\tau) \sum_{t=1}^{H} \nabla_\theta \ln \pi_{\theta,t}(A_t \mid S_t)\Big].$$

Proof. By Theorem 1.14, $\mathbb{P}_\mu^{\pi_\theta}(\tau) = \mu(s_1) \prod_{t \le H} \pi_{\theta,t}(a_t \mid s_t) P_t(s_{t+1} \mid s_t, a_t)$ on the finite set Ω_H, so $\nabla_\theta \mathbb{P}_\mu^{\pi_\theta}(\tau) = \mathbb{P}_\mu^{\pi_\theta}(\tau) \sum_t \nabla_\theta \ln \pi_{\theta,t}(a_t \mid s_t)$ (the kernel factors do not depend on θ), and differentiating the finite sum $\sum_\tau f(\tau) \mathbb{P}_\mu^{\pi_\theta}(\tau)$ gives the claim. ∎

The identity holds for every f, a fact that Section 14.4 will exploit. For the return $f(\tau) = W_H$, two refinements are possible: the score at time t is uncorrelated with the rewards *before* t and with any function of S_t alone, so that the return can be replaced by the reward-to-go minus a baseline. In the discounted setting, the same argument organized through the Bellman equation gives the classical statement.

Theorem 14.3 (Policy gradient theorem, Sutton et al., 2000). For every $\theta \in \Theta$, writing $\pi = \pi_\theta$,

$$\nabla J(\theta) \;=\; \frac{1}{1-\gamma} \sum_{s,a} d_\mu^\pi(s) \, \nabla_\theta \pi_\theta(a \mid s) \, Q^\pi(s,a) \;=\; \frac{1}{1-\gamma} \, \mathbb{E}_{s \sim d_\mu^\pi,\; a \sim \pi(\cdot \mid s)}\big[\nabla_\theta \ln \pi_\theta(a \mid s) \, Q^\pi(s,a)\big].$$

Moreover Q^π may be replaced by $Q^\pi - b$ for any function $b : \mathcal{S} \to \mathbb{R}$, in particular by the advantage A^π. In a finite-horizon MDP, the same holds in the form $\nabla J(\theta) = \sum_{t=1}^{H} \mathbb{E}_\mu^\pi\big[\nabla_\theta \ln \pi_{\theta,t}(A_t \mid S_t) \, \big(Q_t^\pi(S_t, A_t) - b_t(S_t)\big)\big]$.

Proof. Differentiate the Bellman equations $V^\pi(s) = \sum_a \pi_\theta(a \mid s) Q^\pi(s,a)$ and $Q^\pi(s,a) = r(s,a) + \gamma \sum_{s'} P(s' \mid s,a) V^\pi(s')$ (the values are differentiable in θ, being rational functions of the $\pi_\theta(a \mid s)$ by Theorem 7.11):

$$\nabla V^\pi(s) = \sum_a \nabla \pi_\theta(a \mid s) Q^\pi(s,a) + \sum_a \pi_\theta(a \mid s) \gamma \sum_{s'} P(s' \mid s,a) \nabla V^\pi(s') = g(s) + \gamma \sum_{s'} P^\pi(s,s') \nabla V^\pi(s'),$$

with $g(s) \coloneqq \sum_a \nabla \pi_\theta(a \mid s) Q^\pi(s,a)$. Iterating, $\nabla V^\pi = \sum_{t \ge 1} \gamma^{t-1} (P^\pi)^{t-1} g$, and averaging against μ gives $\nabla J = \sum_s \big(\sum_t \gamma^{t-1} \mathbb{P}_\mu^\pi(S_t = s)\big) g(s) = \frac{1}{1-\gamma} \sum_s d_\mu^\pi(s) g(s)$. The second form uses $\nabla \pi_\theta = \pi_\theta \nabla \ln \pi_\theta$. The baseline can be subtracted because $\sum_a \nabla \pi_\theta(a \mid s) b(s) = b(s) \nabla \sum_a \pi_\theta(a \mid s) = 0$. The finite-horizon statement follows from the same computation with the time-indexed Bellman equations of Corollary 3.11, the sum over t replacing the geometric series. ∎

The theorem says that the gradient is an average, over the states visited and the actions taken, of the score of the action weighted by its value: actions better than average have their probability increased. Its power is that it involves no derivative of the dynamics, which is unknown, and no derivative of the occupation measure, which does depend on θ: this is the content of the proof.

Example 14.4 (A one-state problem, by hand). Everything in this chapter can be computed on a bandit: one state, three actions with Bernoulli rewards of means $r = (0.7, 0.5, 0.2)$, $\gamma = 0$, so that $J(\theta) = \sum_a \pi_\theta(a) r(a)$ for the tabular softmax $\pi_\theta(a) \propto e^{\theta(a)}$ (`figures/scripts/pg_one_state.py`). *The gradient.* With d_μ^π trivial and $Q^\pi = r$, Theorem 14.3 with the score $\mathbb{1}\{a' = a\} - \pi(a)$ gives $\partial J / \partial \theta(a) = \pi(a)\big(r(a) - J(\theta)\big)$: the probability of an action moves in proportion to its advantage $r(a) - J$ *and* to its current probability. At $\theta = 0$, $J = 0.467$ and the gradient is $(0.078, 0.011, -0.089)$; at $\theta = (0, 0, 5)$, where the worst action has probability 0.987 and the best 0.0066, $J = 0.205$ and the gradient is $(0.0033, 0.0020, -0.0053)$, of norm 0.0065—eighteen times smaller, although the policy is much worse: a good action of small probability is promoted slowly, and near every deterministic policy, optimal or not, the gradient vanishes. *The plateau.* Gradient ascent $\theta_{k+1} = \theta_k + \eta \nabla J(\theta_k)$ with $\eta = 1$ gives probability $\frac{1}{2}$ to the best action after 7 steps from $\theta = 0$ and after 136 steps from $\theta = (0, 0, 5)$ (2 and 29 steps with $\eta = 5$); from the bad start the gap $J^\star - J(\theta_k)$ is still 0.476 after 100 steps with $\eta = 1$, and J is nondecreasing along all four runs. *The sampled gradient.* REINFORCE from one sample, $\hat{g} = \nabla \ln \pi_\theta(A)\,(R - b)$ with $A \sim \pi_\theta$ and $R \sim \mathrm{Ber}(r(A))$, has $\mathbb{E}[\|\hat{g}\|^2] = \sum_a \pi(a) \|e_a - \pi\|^2 \big(r(a)(1 - r(a)) + (r(a) - b)^2\big)$, hence an explicit variance (trace of the covariance). At $\theta = 0$ it is 0.297 without baseline and 0.152 with $b = J$, which is also the optimal baseline $b^\star$ of Exercise 14.3 there (all scores have the same norm), against $\|\nabla J\|^2 = 0.014$: a signal-to-noise ratio of one needs 11 samples. At $\theta = (0, 0, 5)$ the variance is 0.0157 ($b = 0$), 0.0103 ($b = J$) and 0.0063 ($b^\star = 0.596$, far from $J = 0.205$: the optimal baseline weights the rewards by the squared norms of the scores, which are large for the rare actions), against $\|\nabla J\|^2 = 4.2 \cdot 10^{-5}$: 150 to 370 samples per unit of signal-to-noise ratio. On the plateau the gradient is not only small, it is drowned. *The natural gradient.* The update (14.2) reads $\pi_{k+1}(a) \propto \pi_k(a) e^{\eta (r(a) - J_k)}$, and since J_k does not depend on a the tilts compose: $\pi_K(a) \propto \pi_0(a) e^{\eta K r(a)}$, a closed form. The best action reaches probability $\frac{1}{2}$ after 3 steps from $\theta = 0$ and 11 from $\theta = (0, 0, 5)$ with $\eta = 1$ (1 and 3 with $\eta = 5$): the factor $\pi(a)$ is gone, and the gap decays geometrically, at the rate $e^{-\eta(r(1) - r(2))} = 0.82$ per step—$2.7 \cdot 10^{-2}$ at $K = 10$, $9 \cdot 10^{-6}$ at $K = 50$ from the uniform start, whereas the bound of Theorem 14.13, $\ln 3/(\eta K) + R_{\max}/K$, is 0.21 and 0.042. From $\theta = (0, 0, 5)$ the theorem does not apply (π_0 is not uniform), but its proof does, with $\mathrm{KL}\big(\pi^\star \,\big\|\, \pi_0\big) = \ln(1/\pi_0(1)) = 5.0$ in place of $\ln A$: the bound becomes 0.30 at $K = 20$, where the gap is $6.9 \cdot 10^{-3}$. The $1/K$ rate is the regret rate of Hedge, paid once; what the bandit shows is that the mechanism of the natural gradient is multiplicative, and that of the plain gradient is not.

REINFORCE. An unbiased estimate of the finite-horizon gradient from one trajectory is

$$\hat{g} \;=\; \sum_{t=1}^{H} \nabla_\theta \ln \pi_{\theta,t}(A_t \mid S_t)\,\big(W_{t:H} - b_t(S_t)\big),$$

the reward-to-go $W_{t:H}$ being an unbiased estimate of $Q_t^\pi(S_t, A_t)$ given (S_t, A_t) (Exercise 14.2 shows directly why the rewards before t may be dropped). This is the REINFORCE algorithm of Williams (1992), Algorithm 14.1. In the discounted setting one truncates at a horizon H with γ^H negligible and weights the score at time t by γ^{t-1}; in practice the factor γ^{t-1} is almost always dropped, and the estimate then follows a direction that is the gradient of no function of θ at all (Nota and Thomas, 2020): states visited late in the episode are given the same weight as the first ones, whereas J discounts them. The difference is small when the episodes are short compared with $1/(1 - \gamma)$ and matters when they are not (Exercise 14.4).

Remark 14.5 (Policies on observations). If the agent sees only observations Y_t of a hidden state (Section 5.7) and uses a policy $\pi_\theta(a \mid Y_t)$, or a policy of the whole observation history, the REINFORCE estimate with $\nabla_\theta \ln \pi_{\theta,t}(A_t \mid Y_t)$ in place of $\nabla_\theta \ln \pi_{\theta,t}(A_t \mid S_t)$ remains unbiased: the

Algorithm 14.1: REINFORCE with a baseline (episodic)

Input: parametrized policy π_θ, step sizes (η_k), baseline b (e.g. a running average of returns, or a learned value function), batch size N

for $k = 1, 2, \ldots$ **do**
  run N episodes with π_θ; for each episode i compute the rewards-to-go $W^i_{t:H}$
  $\hat{g} \leftarrow \frac{1}{N} \sum_{i \le N} \sum_{t \le H} \nabla_\theta \ln \pi_{\theta,t}(A^i_t \mid S^i_t) \left(W^i_{t:H} - b(S^i_t)\right)$
  $\theta \leftarrow \theta + \eta_k \hat{g}$; update the baseline

proof of Lemma 14.2 uses only the product form of the law of the trajectory (5.4), in which θ enters through the policy factors alone, not the Markov property of what the agent sees. That property is needed, on the contrary, by the forms of the gradient written with $Q^\pi(s,a)$ and by temporal-difference and Q-learning on observations: when two states share an observation (*aliasing*), the observation process is not Markov and the Bellman equation on observations fails. The best memoryless policy may then have to be stochastic, and be arbitrarily better than every deterministic memoryless one (Singh et al., 1994)—one of the reasons, announced above, for searching among stochastic policies.

Stochastic gradient ascent on a non-concave function converges, under the Robbins–Monro conditions, for smooth J and as long as the iterates remain bounded, to the set of stationary points (Borkar, 2008, Chap. 2) (Theorem 2.21 treated the convex case); which stationary point is reached depends on the initialization, an issue that Section 14.3 examines in the tabular case. The practical issue is the variance of $\hat{g}$, which grows with the horizon and with the variability of the returns.

Baselines and critics. The baseline $b(s)$ does not change the expectation of the estimate but changes its variance, and the choice $b = V^\pi$—the advantage form—is close to optimal in the sense of Greensmith et al. (2004) (Exercise 14.3). Since V^π is unknown, one *learns* it: an *actor–critic* method maintains a parametrized value function V_w, the critic, trained by TD(0) or by regression on the observed returns (Chapters 12 and 13), and uses it both as a baseline and, replacing the reward-to-go by the one-step target $R_t + \gamma V_w(S_{t+1})$, to reduce the variance of the actor's gradient at the price of a bias when $V_w \neq V^\pi$. The session-H06 script, described in the hands-on box, is of this kind. The interpolation between the one-step target and the full return through geometric weights, the generalized advantage estimation of Schulman et al. (2016), is the TD(λ) idea of Remark 12.12 applied to the advantage that weights the score. How large is the variance being fought? For REINFORCE with returns in $[0, HR_{\max}]$ and scores bounded by G, the estimate $\hat{g}$ has a second moment of order $H^2 \cdot (HR_{\max})^2 G^2$ in the worst case: the sum of H scores and the return each scale with the horizon, and the same computation in the discounted case gives $(1-\gamma)^{-4}$; baselines and critics remove the part of the return that the actions do not control, which is where most of this variance sits (Greensmith et al., 2004). The convergence of actor–critic methods, in which the critic must track the changing policy, is analyzed as a two-time-scale stochastic approximation by Konda and Tsitsiklis (2003): the critic moves on the fast scale and the actor sees an almost converged critic. Two theoretical remarks close the paragraph. First, if the critic is linear in the score, $Q_w(s,a) = w^\top \nabla_\theta \ln \pi_\theta(a \mid s)$, and fitted by least squares under $d^\pi_\mu \otimes \pi$, then the policy gradient computed with Q_w in place of Q^π is *exact*: this is the *compatible function approximation* of Sutton et al. (2000) and Konda and Tsitsiklis (2003) (Exercise 14.5). Second, for continuous actions a deterministic policy $a = m_\theta(s)$ has a gradient of its own, $\nabla J = \frac{1}{1-\gamma} \mathbb{E}_{d^\pi_\mu} [\nabla_\theta m_\theta(s)\, \nabla_a Q^\pi(s,a)|_{a=m_\theta(s)}]$, which needs a differentiable critic but no sampling over actions (Silver et al., 2014); it underlies the deep deterministic policy gradient of Lillicrap et al. (2016). The simplest actor–critic is written out in Algorithm 14.2: the temporal-difference error δ_t of the critic is, when $V_w \approx V^\pi$, an unbiased estimate of the advantage $A^\pi(S_t, A_t)$ given (S_t, A_t), and the actor moves along the score weighted by it.

Algorithm 14.2: One-step actor–critic (online)

Input: parametrized policy π_θ (actor), parametrized value V_w (critic), step sizes $\eta_\theta \ll \eta_w$

for *each episode* **do**
  observe S_1
  for $t = 1, 2, \ldots$ *until the episode ends* **do**
    draw $A_t \sim \pi_\theta(\cdot \mid S_t)$; observe R_t, S_{t+1}
    $\delta_t \leftarrow R_t + \gamma V_w(S_{t+1}) - V_w(S_t)$ ▷ $V_w(S_{t+1}) = 0$ *if* S_{t+1} *is terminal*
    $w \leftarrow w + \eta_w\, \delta_t\, \nabla_w V_w(S_t)$ ▷ *critic: TD(0), Algorithm 12.1*
    $\theta \leftarrow \theta + \eta_\theta\, \gamma^{t-1}\, \delta_t\, \nabla_\theta \ln \pi_\theta(A_t \mid S_t)$ ▷ *actor;* γ^{t-1} *usually dropped*

Two worked examples close the section, both with the model: the first on a problem of Chapter 13 whose exact answer is known, with a policy class that encodes the shape of the optimal policy; the second on a continuous action, where the deterministic gradient appears as a limit.

Example 14.6 (The machine with a parametrized threshold policy). Take the machine replacement problem on the discretized model of Example 13.8 (wear grid of step 0.02 on $[0, 60]$, 3 001 states, cost x per year while keeping, replacement cost 4, exponential increments of mean 1, $\gamma = 0.9$); costs are minimized, so we *descend* on J. The optimal policy is the threshold rule at $x^\star = 2.12$ (Theorem 3.22), of cost 17.17 from a new machine. The class

$$\pi_\theta(\text{replace} \mid x) \;=\; \sigma\Big(\frac{x - \theta_1}{w}\Big), \qquad w = e^{\theta_2}, \quad \theta = (\theta_1, \theta_2) \in \mathbb{R}^2,$$

with σ the logistic function, encodes this knowledge: a location θ_1 and a width w, the threshold rules being the limit $w \to 0$. With the model, everything is exact (script `replacement_pg.py`): $J(\theta) = V^{\pi_\theta}(0)$ is one linear solve on the grid (the wear only increases, so the matrix is triangular up to the rank-one replacement row), the occupation measure $d_0^{\pi_\theta}$ one transposed solve, and Theorem 14.3 gives $\nabla J(\theta) = \frac{1}{1-\gamma} \sum_x d_0^{\pi_\theta}(x)\, \nabla_\theta \pi_\theta(\text{replace} \mid x)\, \big(Q^{\pi_\theta}(x, \text{replace}) - Q^{\pi_\theta}(x, \text{keep})\big)$, which agrees with central finite differences to five decimals. Figure 14.1 (left) shows J along θ_1 for $w \in \{1, 0.3, 0.1\}$: unimodal in each case, with minima 19.45 at $\theta_1 = 3.10$, 17.51 at 2.29 and 17.22 at 2.14. The best location tends to $x^\star$, and the cost of the smoothing, $J - 17.17$, is 2.27, 0.34 and 0.04, shrinking by almost an order of magnitude with each reduction of the width: a smooth policy is never exactly optimal, and the sharper the cheaper. But the gradient shrinks at the same rate: at $\theta_1 = 2.12$ its norm is 4.4, 0.9, 0.10 and 0.008 for $w = 1$, 0.3, 0.1 and 0.03, proportional to w^2 like the excess cost. This is the plateau of this class. It is not where the policy is bad—a late threshold, $\theta_1 = 20$ with $w = 1$, costs 59.6 and has a gradient of norm 3.7, because the wear reaches every threshold in the end and the smooth policy keeps the score alive—but where the policy becomes deterministic: the phenomenon of Section 14.3, in a class with two parameters.

Gradient descent on $(\theta_1, \ln w)$ with the constant step $\eta = 0.5$ from $(\theta_1, w) = (6, 1)$, a machine replaced too late ($J = 24.01$), first *widens* the policy ($w = 2.44$ at $k = 1$: a wider policy replaces earlier) and then moves the location: $(2.42, 0.33)$ at $k = 5$, $(2.16, 0.14)$ at $k = 10$, $(2.12, 0.05)$ at $k = 50$ and $(2.12, 0.024)$ at $k = 200$, with $J = 17.59$, 17.25, 17.185 and 17.177. From $(0.5, 1)$, replaced too early ($J = 26.75$), the location overshoots to 3.23 at $k = 1$ and the iterates join the same path, ending at $(2.12, 0.024)$ and $J = 17.177$ (Figure 14.1, right). The cost is within 0.1 of the optimum after 9 and 7 iterations; the remaining 0.003 is the plateau, on which the width decreases like $k^{-1/2}$ since $\partial J / \partial \ln w \propto w^2$. From episodes, the REINFORCE estimate at $(6, 1)$ from $N = 100$ episodes of length 50 ($\gamma^{50} = 0.005$) has, over 200 replications, standard deviations 1.4 and 3.2 on the two components of the exact gradient $(2.50, -1.78)$, and 0.22 and 0.47 with V^{π_θ} as a baseline: the first step has the right sign on θ_1 in 94% of the replications without a baseline, always with it.

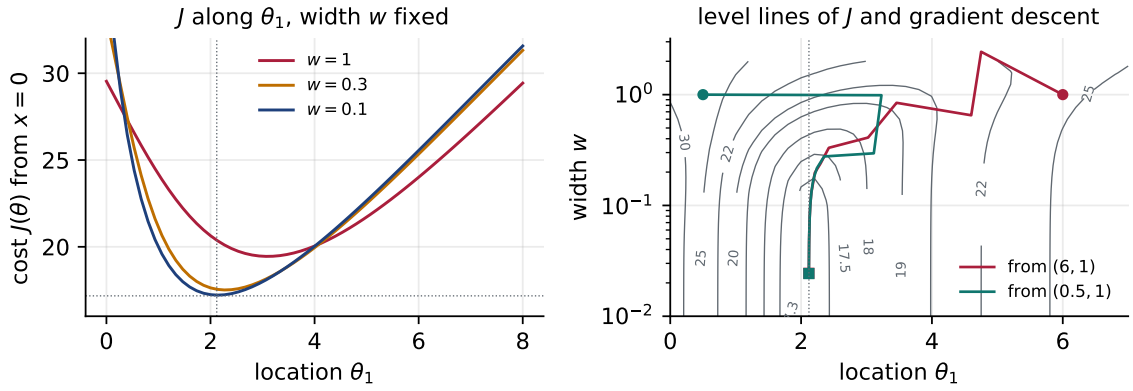


Figure 14.1. The machine with a parametrized threshold policy (Example 14.6). Left: the cost J from a new machine as a function of the location θ_1 for three widths; dotted, the optimal threshold $x^\star = 2.12$ and the optimal cost 17.17. Right: level lines of J in the (θ_1, w) plane (logarithmic width) and the 200 iterates of gradient descent with step $\eta = 0.5$ on $(\theta_1, \ln w)$ from $(6, 1)$ and from $(0.5, 1)$ (circles: starts; squares: ends).

What has this bought over Chapter 13? Two parameters instead of a value on 3 001 states; a structure, the threshold, imposed rather than rediscovered from the shape of $V^\star$, which the classes of Example 13.8 had to bend at $x^\star$; and an objective that is smooth in θ where the greedy threshold of a fitted value jumps from one grid cell to the next. The price: a smooth policy is never exactly optimal, the sharp ones sit on a plateau, and the gradient, exact here, becomes from episodes an estimate whose noise the baseline must tame.

Example 14.7 (The linear-quadratic regulator with one policy parameter). Return to the discounted regulator of Example 13.20: $S_{t+1} = S_t + A_t + U_{t+1}$ with $U \sim \mathcal{N}(0, \sigma^2)$, reward $-qe^2 - a^2$ where $e = s - \bar{s}$, and $q = 1$, $\gamma = 0.9$, $\sigma = 1$, started at $e_0 = 2$. The policy class is Gaussian with a single parameter, $a = -g\,e + \sigma_e \xi$ with $\xi \sim \mathcal{N}(0, 1)$: a linear feedback of gain g, plus an exploration noise of level σ_e that we keep fixed, so that the score is $\partial_g \ln \pi_g(a \mid s) = -(a + ge)\, e/\sigma_e^2$. The closed loop $E_{t+1} = (1 - g)E_t + \sigma_e \xi_t + U_{t+1}$ is stable if and only if $\gamma(1 - g)^2 < 1$, that is $g \in (1 - \gamma^{-1/2}, 1 + \gamma^{-1/2}) = (-0.054, 2.054)$; outside, $J = -\infty$. Inside, the expected reward at e is $-(q + g^2)e^2 - \sigma_e^2$ and the Bellman equation of π_g is solved by $V^{\pi_g}(e) = -P(g)e^2 - c$ with $P(g) = (q + g^2)/(1 - \gamma(1 - g)^2)$, as in Exercise 13.5(c), and $c = \big(\sigma_e^2 + \gamma P(g)(\sigma^2 + \sigma_e^2)\big)/(1 - \gamma)$: the exploration noise costs σ_e^2 per step directly, and $\gamma P(g)\sigma_e^2$ through the state it perturbs. Hence

$$J(g, \sigma_e) \;=\; -P(g)\, D(\sigma_e) - \frac{\sigma_e^2}{1 - \gamma}, \qquad D(\sigma_e) \;\coloneqq\; e_0^2 + \frac{\gamma(\sigma^2 + \sigma_e^2)}{1 - \gamma},$$

and $\partial_g J = -P'(g) D(\sigma_e) = 2\big[\gamma(1 - g)(q + g) - g\big] D(\sigma_e)/(1 - \gamma(1 - g)^2)^2$ vanishes at the single point $g^\star = \gamma p^\star/(1 + \gamma p^\star) = 0.588$, the optimal gain of Section 13.5, whatever σ_e: exploration costs, but does not move the optimum. Exact policy evaluation on the grid of Example 13.20 reproduces J to within $4 \cdot 10^{-2}$ for $g \in \{0.3, g^\star, 1, 1.5\}$ and $\sigma_e \in \{0, 0.3, 1\}$ (`figures/scripts/lq_policy_gradient.py`). Three things can now be checked.

The score-function form of Theorem 14.3. With $Q^{\pi_g}(e, a) = -qe^2 - a^2 - \gamma\big[P(g)((e+a)^2 + \sigma^2) + c\big]$, integrating the score against Q^{π_g} over $a \sim \pi_g(\cdot \mid e)$ leaves $2e^2\big(\gamma P(g)(1-g) - g\big)$, and $\frac{1}{1-\gamma}\mathbb{E}_{d_\mu^{\pi_g}}[e^2] = \sum_t \gamma^{t-1}\mathbb{E}[E_t^2] = D(\sigma_e)/(1 - \gamma(1 - g)^2)$ by the recursion $\mathbb{E}[E_{t+1}^2] = (1 - g)^2\mathbb{E}[E_t^2] + \sigma^2 + \sigma_e^2$; the

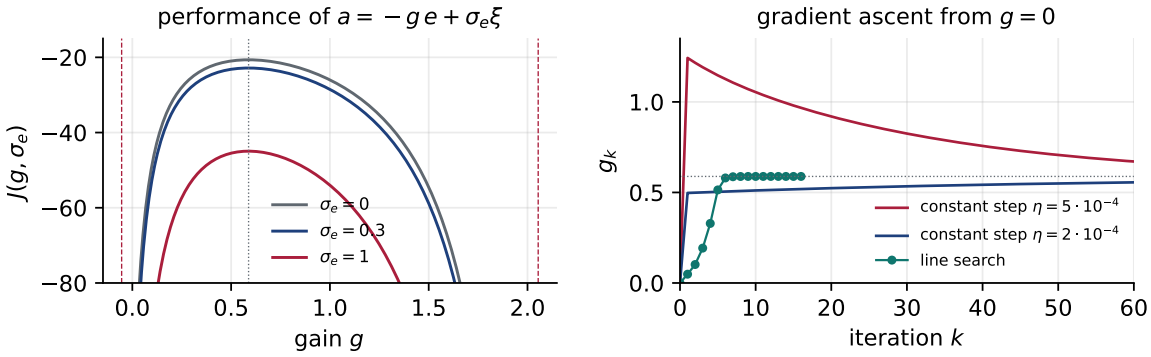


Figure 14.2. The regulator with one policy parameter (Example 14.7). Left: $J(g, \sigma_e)$ for three exploration levels on the stable range, whose boundaries (dashed) are where $J = -\infty$; the optimal gain $g^\star = 0.588$ (dotted) does not depend on σ_e. Right: gradient ascent from $g = 0$ with $\sigma_e = 0.3$, with two constant steps and a backtracking line search.

theorem thus gives $2(\gamma P(g)(1-g) - g)D(\sigma_e)/(1-\gamma(1-g)^2)$, which is $\partial_g J$ by the identity $P(g)(1-\gamma(1-g)^2) = q + g^2$. Computed numerically—occupation measure on the grid, Gauss–Hermite quadrature over the action—the score-function form agrees with $\partial_g J$ to 10^{-3} relatively (45.85 against 45.87 at $g = 0.3$, $\sigma_e = 0.3$). The REINFORCE estimate of the same gradient from one episode of length 100, with the baseline V^{π_g}, has at $g = 0.3$ a standard deviation of 161, 167, 425 and 1 372 for $\sigma_e = 1$, 0.3, 0.1 and 0.03 (estimated from 4 000 episodes): it grows like $1/\sigma_e$, as the score does, while the gradient stays near 45; without the baseline the figures are 397, 461, 1 163 and 4 017. A batch of $N = 100$ episodes at $\sigma_e = 0.3$ estimates the gradient 45.9 with a standard deviation of 17; at $\sigma_e = 0.03$, with one of 137, three times the gradient.

The limit $\sigma_e \to 0$. The deterministic policy gradient of Silver et al. (2014) for $m_g(e) = -ge$ is $\frac{1}{1-\gamma}\mathbb{E}_d\big[\partial_g m_g(e)\,\partial_a Q(e,a)|_{a=m_g(e)}\big] = \frac{1}{1-\gamma}\mathbb{E}_d\big[2e^2(\gamma P(g)(1-g) - g)\big]$: the same integrand. For a Gaussian policy and a Q quadratic in a, Stein's identity $\mathbb{E}[\xi f(\sigma_e\xi)] = \sigma_e\mathbb{E}[f'(\sigma_e\xi)]$ turns the score-function integrand into the deterministic one exactly, and the two gradients differ only through the occupation measure, by the factor $D(\sigma_e)/D(0) = 1 + \gamma\sigma_e^2/((1-\gamma)D(0))$: 1.69, 1.06 and 1.007 for $\sigma_e = 1$, 0.3 and 0.1. The limit is the deterministic gradient; the price of sampling the action is the $1/\sigma_e$ above, the price of not sampling it is that $\partial_a Q$, hence a differentiable critic, is needed.

Gradient ascent with the model. J is concave in g on the stable range (its second derivative is at most -62 there), so that the stationary point is the global maximum: the one-dimensional case of the gradient domination property proved by Fazel et al. (2018) for linear-quadratic regulators of any dimension, where J is not concave but every stationary point in the stable set is optimal and gradient descent converges at a linear rate. But the curvature ranges from $-9 \cdot 10^4$ at $g = 0$, near the unstable boundary, to -79 at $g^\star$: with $\sigma_e = 0.3$, the constant step $\eta = 10^{-3}$ sends $g_1 = 2.49$ out of the stable range, where $J = -\infty$; $\eta = 5 \cdot 10^{-4}$ overshoots to $g_1 = 1.24$ and then needs 172 iterations to come within 10^{-3} of $g^\star$; $\eta = 2 \cdot 10^{-4}$ needs 276 (Figure 14.2). This is planning: the gradient is exact and J itself is computable, so that a backtracking line search costs nothing, and it reaches $g^\star$ to six digits in 16 iterations. From episodes, neither is available: the step must be fixed in advance, the gradient is the REINFORCE estimate with the variance above, and an iterate thrown beyond 2.054 by one noisy step diverges before anything signals it.

14.3 Softmax policies, natural gradients and mirror descent

The tabular softmax is the case where everything can be computed, and it displays both the weakness of plain gradient ascent and the remedy. The section is a planning section: the gradients are exact, and the question is how fast gradient ascent reaches the optimum. It proceeds in four steps. The gradient of the softmax has two damping factors, the probability of the action and the occupation of the state, which make plain ascent slow (a theorem, and two examples). The *natural* gradient removes both and becomes, at each state, a multiplicative update on the advantages; that one update has five readings, collected in Table 14.1, and a convergence rate free of the size of the state space. Regularizing toward the previous policy, the same device, stabilizes approximate dynamic programming (error averaging). And in the general parametrized case the update becomes a trust region, whose guarantee rests on a bound that we prove and whose practical forms are TRPO and PPO.

Proposition 14.8 (Softmax policy gradient). For the tabular softmax parametrization, writing $\pi = \pi_\theta$,

$$\frac{\partial J}{\partial \theta(s,a)}(\theta) \;=\; \frac{1}{1-\gamma}\, d_\mu^\pi(s)\, \pi(a \mid s)\, A^\pi(s,a).$$

Proof. By Theorem 14.3 with the score $\partial \ln \pi(a' \mid s')/\partial\theta(s,a) = \mathbb{1}\{s'=s\}(\mathbb{1}\{a'=a\} - \pi(a \mid s))$: $\frac{\partial J}{\partial \theta(s,a)} = \frac{1}{1-\gamma} d_\mu^\pi(s) \sum_{a'} \pi(a' \mid s)(\mathbb{1}\{a'=a\} - \pi(a \mid s)) Q^\pi(s,a') = \frac{1}{1-\gamma} d_\mu^\pi(s)\pi(a \mid s)\big(Q^\pi(s,a) - V^\pi(s)\big)$. ∎

Two features of this formula shape the behavior of gradient ascent. The factor $\pi(a \mid s)$ makes the gradient small for actions of small probability: a good action that the current policy rarely takes is slowly promoted, and near-deterministic policies are plateaus—the gradient vanishes as $\|\theta\| \to \infty$ towards any deterministic policy, optimal or not. The factor $d_\mu^\pi(s)$ makes the gradient small at states that the current policy rarely visits: the parameters there barely move, however bad the initial policy is at those states. At any finite θ, on the other hand, $\nabla J(\theta) = 0$ with μ of full support forces $A^{\pi_\theta} \equiv 0$, hence optimality: the danger is convergence towards infinity, and global convergence is nevertheless guaranteed when the initial distribution has full support. The theorem is a planning statement: the gradient is exact, computed from the model.

Theorem 14.9 (Global convergence of softmax policy gradient, Agarwal et al., 2021a; Mei et al., 2020). Suppose $\mu(s) > 0$ for all s and let $\theta_{k+1} = \theta_k + \eta \nabla J(\theta_k)$ with a constant step $\eta \le (1-\gamma)^3/(8R_{\max})$, the inverse of the smoothness constant of J in the softmax parametrization. Then $V^{\pi_{\theta_k}}(s) \to V^\star(s)$ for every s, and $V^\star(\mu) - V^{\pi_{\theta_k}}(\mu) \le C/k$ for a constant C depending on μ, γ, the MDP and θ_0.

The substance of the proof (Agarwal et al., 2021a, Thm. 10) is to exclude convergence to a suboptimal deterministic policy at infinity; the rate is due to Mei et al. (2020). But the constant C can be enormous: Li et al. (2021) exhibit MDPs on which softmax gradient ascent needs $S^{2^{\Omega(1/(1-\gamma))}}$ iterations to escape its plateaus, a time exponential in the effective horizon. Example 14.10 shows a milder but typical failure: with μ concentrated on a single start state—outside the scope of the theorem, which assumes μ of full support—gradient ascent stalls, over the 10^4 iterations shown (and the 10^5 of Example 14.21), near a suboptimal policy, because the states that this policy visits only with vanishing probability keep essentially their initial, uniform decision rule.

Example 14.10 (Policy gradient on the windy hike: the mean). We return to the windy hike of Example 3.6—14 steps, the 20 cells outside the pit plus the sink, 4 actions, start state S, one unit of reward per step spent at the bar—with time-indexed softmax policies $\pi_{\theta,t}(a \mid s) \propto e^{\theta_t(s,a)}$, $14 \times 21 \times 4$ parameters initialized at 0 (the uniform policies), and maximize the expected return $U_0(\theta) = \mathbb{E}_S^{\pi_\theta}[W_{14}]$, whose optimum 5.09 is attained by the paths of row 3 (Example 5.10); the child's path along row 2 is worth 5.02. The gradient is the finite-horizon form of Proposition 14.8, $\partial U_0/\partial\theta_t(s,a) = d_t^\pi(s)\,\pi_t(a \mid s)\,A_t^\pi(s,a)$ with $d_t^\pi(s) = \mathbb{P}_S^\pi(S_t = s)$, computed exactly from the model by one backward and one forward recursion on the 21 states (checked against finite differences in `figures/scripts/hike_pg.py`); the numbers below are printed by `figures/scripts/hike_pg_mean.py`.

With the model (Figure 14.3, left). Gradient ascent with step $\eta = 3$ first sits on a plateau—the uniform policy reaches the bar so rarely that all advantages are tiny, and $U_0(\theta_k)$ is still 0.03 at $k = 100$—then climbs within a few dozen iterations to 4.82 at $k = 300$ (95% of the optimum is passed at $k = 404$) and settles on a policy following row 2: 4.85 at $k = 1\,000$, 4.93 at $k = 3\,000$ and at $k = 10^4$, where the child's 5.02 is not even reached. The mechanism is the factor $d_t^\pi(s)$. Once the policy heads east along row 2 (at $k = 10^4$, $\pi_1(\mathrm{N} \mid \mathrm{S}) = 1.000$ and a trajectory visits rows 3–4 with probability 0.002), the cells of rows 3 and 4 carry a visitation probability of at most $3.7 \cdot 10^{-4}$ at any time; their decision rules have remained close to uniform—none gives an action a probability above 1/2, and those of row 4 are within 0.013 of the uniform rule—so that going north from row 2 looks bad, and the exact gradient on those cells, below $3 \cdot 10^{-6}$ in supremum norm (and $3.3 \cdot 10^{-5}$ over all cells), does not move them: the iterates are on a plateau that this run does not leave (Example 14.21 continues it to 10^6 iterations). This is an observation over a finite budget, not a proof that the plateau is never left: Theorem 14.9, which would exclude a suboptimal limit, assumes the discounted criterion and an initial law of full support, and neither hypothesis holds here (Exercise 14.13(a) restores the second). Natural gradient ascent with $\eta = 1$—the update (14.2) applied at every pair (t, s), which removes both damping factors—passes 95% of the optimum at $k = 6$, reaches 5.09 at $k = 85$ and 5.0947, the optimum to four decimals, from $k = 300$ on, with the greedy path on row 3: the μ-independence of Theorem 14.13, proved there for the discounted criterion but whose mechanism is the same here, is not a technicality.

From episodes (right). Replace now the exact gradient by the REINFORCE estimate of Algorithm 14.1 from N episodes per iteration (same step $\eta = 3$, the batch mean of the returns as baseline), for $N \in \{10, 100, 1\,000\}$, 3 000 iterations and 5 independent runs, the value $U_0(\theta_k)$ of each iterate being computed exactly from the model. With $N = 1\,000$ the five runs follow the exact ascent closely: they pass 95% of the optimum (4.84) after 365 to 467 iterations (428 on average, against 404 with the exact gradient) and end between 4.86 and 4.99 (mean 4.89; the exact gradient gives 4.93 at $k = 3\,000$). With $N = 100$, one run of five passes 4.84 (at $k = 348$) and the final values range from 3.72 to 4.86 (mean 4.28); with $N = 10$, none does and the final values range from 2.09 to 4.62 (mean 3.29). The noise does not lengthen the initial plateau—half of the optimum is passed, on average over the runs that pass it, at $k = 116$, 95 and 90 for $N = 1\,000$, 100 and 10, against 113 with the exact gradient, though two runs with $N = 10$ never do—but it cuts the ascent short. With $\eta = 3$ and noisy gradients the parameters grow fast, the final policies are deterministic on the cells they visit (the mean of $\max_a \pi_t(a \mid s)$ under d_t^π is above 0.98 for every run), and the runs freeze: twelve of the fifteen gain less than 0.03 between $k = 1\,000$ and $k = 3\,000$, at values such as 4.25, 3.72 or 2.09 where the exact gradient is between 10^{-4} and $3 \cdot 10^{-3}$ in supremum norm. These are the plateaus near deterministic policies of Proposition 14.8, which the noise reaches by chance and cannot leave; a smaller step delays them, at the price of more iterations. The noise also helps, in its way: three of the five runs with $N = 10$ have been pushed onto row 3 (their final policies visit rows 3–4 with probability 0.66 to 0.81, against 0.002

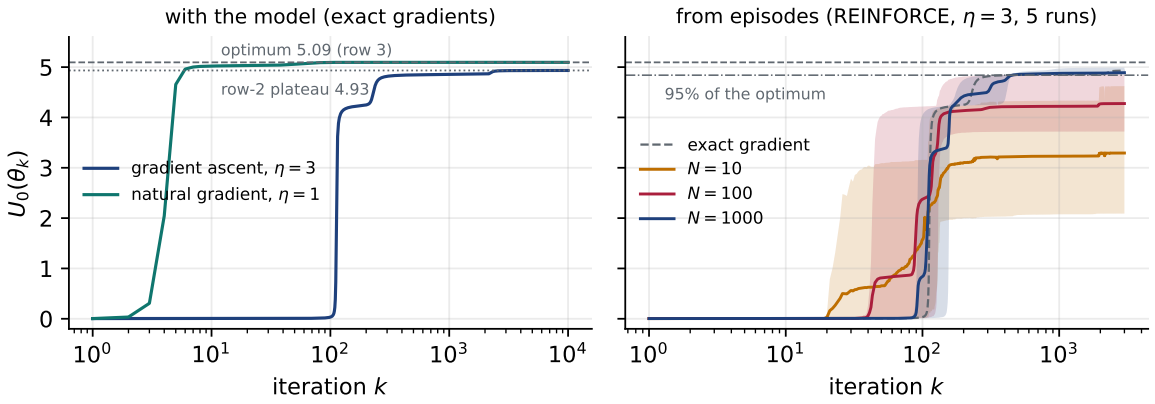


Figure 14.3. Expected return $U_0(\theta_k)$ of the iterates on the windy hike with time-indexed softmax policies (Example 14.10). Left: with the model, gradient ascent ($\eta = 3$) and natural gradient ascent ($\eta = 1$) over 10^4 iterations; the dashed line is the optimum 5.09 (paths of row 3), the dotted line the row-2 plateau 4.93 on which gradient ascent settles. Right: REINFORCE from N episodes per iteration, $\eta = 3$, mean (line) and range (band) over 5 runs; the exact-gradient ascent is dashed for reference and the level of 95% of the optimum dash-dotted.

for the exact gradient), the row of the optimum, but with decision rules there that are far from refined (2.09, 2.32 and 4.62 at $k = 3\,000$); and one run with $N = 1\,000$ ends at 4.99, above the exact-gradient plateau, with a better row-2 policy.

The natural gradient. The two damping factors disappear if the gradient is measured in the geometry of the policy rather than of the parameter. The *Fisher information* of the parametrization, under the state distribution d_μ^π, is $F(\theta) = \mathbb{E}_{s\sim d_\mu^\pi, a\sim\pi_\theta(\cdot|s)}\left[\nabla \ln \pi_\theta(a \mid s) \nabla \ln \pi_\theta(a \mid s)^\top\right]$, and the *natural gradient* (Amari, 1998; Kakade, 2002) is $\widetilde{\nabla} J(\theta) = F(\theta)^+ \nabla J(\theta)$, with F^+ the pseudo-inverse—F is often singular (for the tabular softmax, adding a constant to $\theta(s, \cdot)$ does not change the policy, and the whole block of a state with $d_\mu^\pi(s) = 0$ vanishes), and $F^+\nabla J$ is then the minimum-norm solution of $Fw = \nabla J$, the others differing from it by an element of $\ker F$, which leaves the policy unchanged to first order on the states visited: the direction of steepest ascent when distances between parameters are measured by the Kullback–Leibler divergence between the policies they define, to second order. The last words rest on one expansion, used again for the trust regions below: for a twice differentiable parametrization and a fixed state s, writing $F_s(\theta) = \mathbb{E}_{a\sim\pi_\theta(\cdot|s)}[\nabla \ln \pi_\theta(a \mid s) \nabla \ln \pi_\theta(a \mid s)^\top]$, so that $F(\theta) = \mathbb{E}_{s\sim d_\mu^\pi}[F_s(\theta)]$,

$$\mathrm{KL}(\pi_\theta(\cdot \mid s) \,\|\, \pi_{\theta'}(\cdot \mid s)) \;=\; \tfrac{1}{2}(\theta' - \theta)^\top F_s(\theta)(\theta' - \theta) + o(\|\theta' - \theta\|^2), \tag{14.1}$$

because expanding $\ln \pi_{\theta'}$ around θ to second order, the first-order term $\mathbb{E}_{\pi_\theta}[\nabla \ln \pi_\theta] = \nabla \sum_a \pi_\theta(a \mid s)$ vanishes and $\mathbb{E}_{\pi_\theta}[\nabla^2 \ln \pi_\theta] = -F_s(\theta)$ (differentiate $\sum_a \pi_\theta(a \mid s) = 1$ twice); the other orientation $\mathrm{KL}(\pi_{\theta'}(\cdot \mid s) \,\|\, \pi_\theta(\cdot \mid s))$ has the same expansion, the two differing at third order only.

Proposition 14.11 (Natural policy gradient for the softmax is a multiplicative-weights update). For the tabular softmax, one natural gradient step $\theta_{k+1} = \theta_k + \eta \widetilde{\nabla} J(\theta_k)$ yields, up to an additive constant in each $\theta(s, \cdot)$ that does not affect the policy,

$$\pi_{k+1}(a \mid s) \;=\; \frac{\pi_k(a \mid s) \exp\left(\frac{\eta}{1-\gamma} A^{\pi_k}(s, a)\right)}{Z_k(s)}, \qquad Z_k(s) \coloneqq \sum_{a'} \pi_k(a' \mid s) \exp\left(\frac{\eta}{1-\gamma} A^{\pi_k}(s, a')\right), \tag{14.2}$$

for every state s with $d_\mu^{\pi_k}(s) > 0$. Equivalently, $\pi_{k+1}(\cdot \mid s)$ maximizes $\eta \sum_a p(a) Q^{\pi_k}(s,a)/(1-\gamma) - \mathrm{KL}(p \,\|\, \pi_k(\cdot \mid s))$ over $p \in \mathcal{M}_1(\mathcal{A})$: the update is a *mirror descent* step (an ascent, here) with the negative entropy as mirror map.

Proof. The score at (s,a) has nonzero entries only in the block $\theta(s,\cdot)$, where it equals $e_a - \pi(\cdot \mid s)$; hence the matrix $F_s(\theta)$ of (14.1) vanishes outside that block, where it equals $\sum_a \pi(a \mid s)(e_a - \pi_s)(e_a - \pi_s)^\top = \mathrm{diag}(\pi_s) - \pi_s \pi_s^\top$, $\pi_s = \pi(\cdot \mid s)$, and $F(\theta) = \sum_s d_\mu^\pi(s) F_s(\theta)$ is block diagonal with blocks $d_\mu^\pi(s) F_s$. By Proposition 14.8 the block of the gradient is $\frac{1}{1-\gamma} d_\mu^\pi(s)\ \mathrm{diag}(\pi_s) A^\pi(s,\cdot)$. Now $d_\mu^\pi(s) F_s\, x = d_\mu^\pi(s)\ \mathrm{diag}(\pi_s)\big(x - (\pi_s^\top x)\mathbf{1}\big)$, so $x = \frac{1}{1-\gamma} A^\pi(s,\cdot)$ satisfies $d_\mu^\pi(s) F_s x = \frac{1}{1-\gamma} d_\mu^\pi(s)\ \mathrm{diag}(\pi_s) A^\pi(s,\cdot)$ because $\pi_s^\top A^\pi(s,\cdot) = 0$; the pseudo-inverse solution differs from x by an element of the kernel of F_s, which is spanned by $\mathbf{1}$. Adding ηx to $\theta(s,\cdot)$ multiplies $\pi(a \mid s)$ by $e^{\eta A^\pi(s,a)/(1-\gamma)}$ and renormalizes. For the mirror-descent form, the maximizer of $\eta \langle q, p\rangle - \mathrm{KL}(p \,\|\, \pi_s)$ over the simplex is $p(a) \propto \pi_s(a) e^{\eta q(a)}$ (Theorem C.11), and Q^{π_k} and A^{π_k} differ by a constant in a. ∎

The mirror-descent form deserves a statement of its own, since the same one-stage problem recurs, with other reference laws, in Chapters 15 and 16.

Corollary 14.12 (Regularized greedy steps are exponential tilts). For $q \in \mathbb{R}^{\mathcal{A}}$, a reference law $p_0 \in \mathcal{M}_1(\mathcal{A})$ with full support and $\lambda > 0$,

$$\arg\max_{p \in \mathcal{M}_1(\mathcal{A})} \big\{ \langle p, q\rangle - \lambda\, \mathrm{KL}(p \,\|\, p_0) \big\} = \frac{p_0\, e^{q/\lambda}}{\sum_a p_0(a) e^{q(a)/\lambda}}, \qquad \max_{p \in \mathcal{M}_1(\mathcal{A})} \{ \cdots \} = \lambda \ln \sum_a p_0(a) e^{q(a)/\lambda},$$

the β-mean of $q(A)$, $A \sim p_0$, at rate $\beta = 1/\lambda$ (Theorem C.11). One natural gradient step (14.2) is this problem with $q = Q^{\pi_k}(s,\cdot)$, $p_0 = \pi_k(\cdot \mid s)$ and $\lambda = (1-\gamma)/\eta$: natural policy gradient is regularized policy iteration with a moving reference, the step size being the inverse of the regularization. With p_0 uniform the same problem is the soft Bellman step of Chapter 15, and with p_0 an external reference policy the regularized Bellman step of Chapter 16.

The vocabulary used above deserves three lines (Bubeck, 2015, Chap. 4). A gradient step $p_{k+1} = \arg\max_p \{\langle p, q\rangle - \frac{1}{2\eta}\|p - p_k\|^2\}$ measures the distance to the current point by the squared Euclidean norm. *Mirror descent* replaces this proximity term by the Bregman divergence $D_\Phi(p, p_k) = \Phi(p) - \Phi(p_k) - \langle \nabla\Phi(p_k), p - p_k\rangle$ of a strictly convex *mirror map* Φ (Definition C.13); for the negative entropy $\Phi(p) = \sum_a p(a) \ln p(a)$ on the simplex, $D_\Phi(p, p_k) = \mathrm{KL}(p \,\|\, p_k)$, and the step is the exponential tilt of Corollary 14.12. *Dual averaging* is the variant in which the step is taken from a fixed reference point with the *sum* of all past gradients, $p_{k+1} = \arg\max_p \{\langle p, \sum_{j \le k} q_j\rangle - \frac{1}{\eta} D_\Phi(p, p_0)\}$; with the entropy the two variants coincide (the tilts compose), which is the closed form $\pi_{k+1} \propto \exp(\frac{1}{\lambda}\sum_{j \le k} Q_j)$ met below.

The update (14.2) is the *multiplicative weights* rule of online learning (Freund and Schapire, 1999), the Hedge algorithm of Freund and Schapire (1997), applied at each state to the advantages—a connection that Chapter 17 develops. Its two features are exactly those the plain gradient lacked: the occupation measure $d_\mu^\pi(s)$ has disappeared, so that rarely visited states are updated as firmly as frequent ones, and the multiplicative form promotes a good action at a rate that does not vanish with its current probability. The consequence is a convergence guarantee that involves neither μ nor the size of the state space.

Theorem 14.13 (Convergence of the natural policy gradient, Agarwal et al., 2021a ♦). Let π_0 be the uniform policy and π_{k+1} be given by (14.2) at every state (the closed form is applied

Table 14.1. Five readings of the natural policy gradient step (14.2). Each row is the same update seen from a different field; the columns give what is maximized, the proximity term that keeps the new policy close to the old one, and where the reading is used.

Reading	What is maximized	Proximity term	Where
Natural gradient (parameter space)	the linearized J, $\nabla J^\top(\theta-\theta_k)$	$\frac{1}{2}(\theta-\theta_k)^\top F(\theta_k)(\theta-\theta_k)$, the KL to second order	Proposition 14.11, (14.1)
Multiplicative weights, Hedge (one state)	$\sum_a p(a)A^{\pi_k}(s,a)$	$\mathrm{KL}(p \,\|\, \pi_k(\cdot \mid s))$	Theorem 14.13, Chapter 17
Mirror descent, dual averaging	the same	Bregman divergence of the entropy, i.e. the KL; from a fixed reference with the sum of past gradients	Definition C.13, (14.4)
Regularized policy iteration	$\langle p, Q^{\pi_k}(s,\cdot)\rangle$	$\lambda\,\mathrm{KL}(p \,\|\, \pi_k)$ with $\lambda=(1-\gamma)/\eta$	Corollary 14.12, Chapters 15 and 16
Trust region (any parametrization)	the surrogate $L_{\pi_k}(\pi_\theta)$	$\mathrm{KL}(\pi_k \,\|\, \pi_\theta) \le \delta_{\mathrm{TR}}$ (constraint), or clipping (PPO)	Theorem 14.16, Algorithm 14.3

also where $d_\mu^{\pi_k}(s)=0$, where the natural gradient itself would prescribe no change), for some $\eta>0$. Then for every $K\ge 1$ and every μ,

$$V^\star(\mu)-V^{\pi_K}(\mu) \;\le\; \frac{\ln A}{\eta K}+\frac{R_{\max}}{(1-\gamma)^2 K}$$

($A=|\mathcal{A}|$ is the number of actions; A^π is the advantage).

The proof has two steps, both resting on the performance difference lemma. First, the update improves the value at *every* state, by an amount that the normalizer Z_k measures. Second, the gap to the optimum is written along the occupation measure of the optimal policy, where the update is a Hedge step whose regret telescopes through the divergences to $\pi^\star$; the normalizers that appear there are paid for by the improvement of the first step.

Proof. Write $V^k=V^{\pi_k}$, $A^k=A^{\pi_k}$, and $\pi^\star$ for an optimal deterministic policy with occupation measure $d^\star=d_\mu^{\pi^\star}$.

Step 1: monotone improvement. For every state s, by Lemma 14.1 with initial law δ_s and by (14.2), which gives $\frac{\eta}{1-\gamma}A^k(s',a)=\ln\frac{\pi_{k+1}(a|s')}{\pi_k(a|s')}+\ln Z_k(s')$,

$$\begin{aligned} V^{k+1}(s)-V^k(s) &= \frac{1}{1-\gamma}\mathbb{E}_{s'\sim d_s^{\pi_{k+1}}}\Big[\sum_a \pi_{k+1}(a\mid s')A^k(s',a)\Big] \\ &= \frac{1}{\eta}\,\mathbb{E}_{s'\sim d_s^{\pi_{k+1}}}\Big[\mathrm{KL}(\pi_{k+1}(\cdot\mid s')\,\|\,\pi_k(\cdot\mid s'))+\ln Z_k(s')\Big]. \end{aligned}$$

Both terms inside the expectation are nonnegative: the divergence by definition, and $\ln Z_k(s')\ge \frac{\eta}{1-\gamma}\sum_a \pi_k(a\mid s')A^k(s',a)=0$ by Jensen's inequality. Since $d_s^{\pi_{k+1}}(s)\ge 1-\gamma$ (the start state carries the first term of (8.1)),

$$V^{k+1}(s)-V^k(s) \;\ge\; \frac{1-\gamma}{\eta}\ln Z_k(s) \;\ge\; 0 \qquad \text{for every } s. \tag{14.3}$$

Step 2: the potential. By Lemma 14.1 with $\pi' = \pi^\star$ and the same identity,

$$V^\star(\mu) - V^k(\mu) = \frac{1}{1-\gamma}\mathbb{E}_{s\sim d^\star}\Big[\sum_a \pi^\star(a \mid s) A^k(s,a)\Big] = \frac{1}{\eta}\mathbb{E}_{s\sim d^\star}\Big[\sum_a \pi^\star(a \mid s)\ln\frac{\pi_{k+1}(a \mid s)}{\pi_k(a \mid s)} + \ln Z_k(s)\Big]$$
$$= \frac{1}{\eta}\mathbb{E}_{s\sim d^\star}\Big[\,\mathrm{KL}\big(\pi^\star(\cdot \mid s)\,\big\|\,\pi_k(\cdot \mid s)\big) - \mathrm{KL}\big(\pi^\star(\cdot \mid s)\,\big\|\,\pi_{k+1}(\cdot \mid s)\big)\,\Big] + \frac{1}{\eta}\mathbb{E}_{s\sim d^\star}\big[\ln Z_k(s)\big],$$

and by (14.3), $\frac{1}{\eta}\mathbb{E}_{d^\star}[\ln Z_k] \le \frac{1}{1-\gamma}\big(V^{k+1}(d^\star) - V^k(d^\star)\big)$. Summing over $k = 0, \dots, K-1$, the divergences telescope and the values telescope:

$$\sum_{k=0}^{K-1}\big(V^\star(\mu) - V^k(\mu)\big) \le \frac{1}{\eta}\mathbb{E}_{d^\star}\big[\,\mathrm{KL}\big(\pi^\star(\cdot \mid s)\,\big\|\,\pi_0(\cdot \mid s)\big)\,\big] + \frac{V^K(d^\star) - V^0(d^\star)}{1-\gamma} \le \frac{\ln A}{\eta} + \frac{R_{\max}}{(1-\gamma)^2},$$

using $\mathrm{KL}\big(\pi^\star(\cdot \mid s)\,\big\|\,\text{uniform}\big) = \ln A$ for a deterministic $\pi^\star$ and $0 \le V \le R_{\max}/(1-\gamma)$. Finally $V^k(\mu)$ is nondecreasing in k by (14.3), so $V^\star(\mu) - V^K(\mu)$ is at most the average of the K terms of the left-hand side. ∎

The bound does not depend on S, on μ, or on the transition kernel: it is the regret bound of Hedge (Chapter 17) paid once per state along the optimal occupation measure, plus a term that accounts for the fact that the advantages change with the policy. Larger steps are better, and indeed with $\eta \to \infty$ the update becomes policy iteration (the greedy action gets all the mass), which converges in finitely many steps; with increasing step sizes the rate is linear (Xiao, 2022; Lan, 2023), and with an entropy regularization—the topic of Chapter 15, where the log-sum-exp appears as a β-mean over actions—the convergence is linear at a dimension-free rate, equal to γ for the largest admissible step (Cen et al., 2022; Mei et al., 2020). What the exact statement hides is that the advantages must be *known*; with estimated advantages, the statistical error enters through the occupation measure of the optimal policy, via the mismatch coefficient $\|d_\mu^{\pi^\star}/\mu\|_\infty$—the single-policy concentrability coefficient $C^\star$ of Definition 13.3, in which the initial law μ plays both roles, that of the initial law and that of the sampling distribution—and the distribution mismatch that the exact analysis avoided returns (Agarwal et al., 2021a, Sect. 6).

Example 14.14 (The natural gradient on the two-state chain, against its bound). Take the two-state chain of Example 7.12 ($\gamma = 0.8$), give state G a second action with transition probabilities $(\frac{1}{3}, \frac{2}{3})$ to (G, B) instead of $(\frac{2}{3}, \frac{1}{3})$, and shift every reward by +1 so that they lie in $[0, 2]$ ($R_{\max} = 2$); state B keeps its single action and μ is uniform (`figures/scripts/npg_two_state.py`). The mean rewards are $r(\mathrm{G}, 1) = \frac{5}{3}$, $r(\mathrm{G}, 2) = \frac{4}{3}$ and $r(\mathrm{B}) = \frac{1}{2}$; the original action is optimal, with $V^\star(\mathrm{G}) = \frac{20}{13} + 5 = 6.538$, $V^\star(\mathrm{B}) = 5.192$, $V^\star(\mu) = 5.865$ and $Q^\star(\mathrm{G}, 2) - V^\star(\mathrm{G}) = -0.692$. The only decision is at G, and with two actions the update (14.2) has a closed form: writing $p_k = \pi_k(1 \mid \mathrm{G})$ and $\Delta_k = Q^{\pi_k}(\mathrm{G}, 1) - Q^{\pi_k}(\mathrm{G}, 2) = \frac{1}{3} + \frac{\gamma}{3}\big(V^{\pi_k}(\mathrm{G}) - V^{\pi_k}(\mathrm{B})\big)$, the advantages are $(1 - p_k)\Delta_k$ and $-p_k\Delta_k$, so that the log-odds of action 1 increase by $\eta\Delta_k/(1-\gamma) = 5\Delta_k$ at each iteration. Since $V^\pi(\mathrm{G}) - V^\pi(\mathrm{B}) \ge 0.735$ for every policy, $\Delta_k \ge 0.53$ and the odds are multiplied by at least $e^{2.6} = 14$ per iteration: from the uniform policy ($\Delta_0 = 0.6$, gap $V^\star(\mu) - V^{\pi_0}(\mu) = 0.865$), $p_1 = 0.953$, $p_2 = 0.9984$, and the gap is 0.093, $3.3 \cdot 10^{-3}$, $1.0 \cdot 10^{-4}$, $3.2 \cdot 10^{-6}$ after 1, 2, 3, 4 iterations, $3.6 \cdot 10^{-15}$ after 10 and 0 to machine precision from $k = 12$ on, successive gaps being in the ratio $e^{-5\Delta_\infty} = 0.031$—a linear rate. The bound of Theorem 14.13, $\ln 2/k + R_{\max}/((1-\gamma)^2 k) = 50.7/k$, equals 50.7, 25.3, 5.1 and 1.0 at $k = 1, 2, 10, 50$: true, and useless here, because it is dimension-free and worst-case—its second term alone exceeds the whole range $[0, 10]$ of the values up to $k = 5$—whereas the chain has one decision, an advantage gap bounded away from zero, and a geometric convergence that no $1/k$ bound captures (the $\ln A/(\eta k)$ term alone, 0.069 at $k = 10$, is thirteen orders of magnitude above the truth). Plain gradient ascent (Proposition 14.8) with the

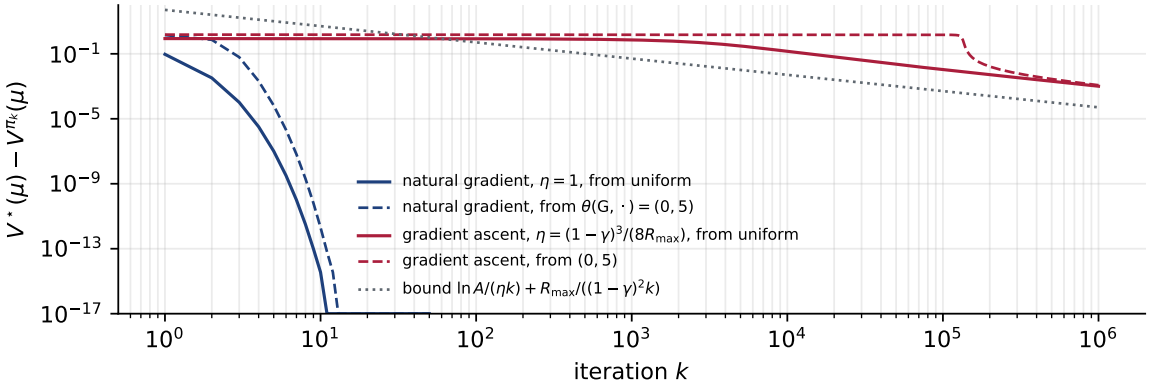


Figure 14.4. The two-state chain with a second action at G (Example 14.14, $\gamma = 0.8$, μ uniform): gap $V^\star(\mu) - V^{\pi_k}(\mu)$ of the natural gradient with $\eta = 1$ (blue) and of plain gradient ascent with the step $(1-\gamma)^3/(8R_{\max})$ of Theorem 14.9 (crimson), from the uniform policy (solid) and from $\theta(\mathrm{G}, \cdot) = (0, 5)$ (dashed), against the bound of Theorem 14.13 (dotted). The natural gradient converges linearly and reaches machine precision in a dozen iterations; gradient ascent follows a $1/k$ law and is slowed further by the bad start.

step $\eta = (1-\gamma)^3/(8R_{\max}) = 5 \cdot 10^{-4}$ of Theorem 14.9 reaches a gap of 10^{-1} after 13 487 iterations and 10^{-2} after 106 623, and k times the gap stays near 1 000 from 10^5 to 10^6 iterations: the C/k rate of the theorem, the gradient being proportional to $p_k(1-p_k)$, which vanishes as the policy becomes deterministic. With $\eta = 1$, two thousand times the guaranteed step and still monotone on this chain, the same gaps are reached after 7 and 52 iterations, 10^{-4} after 5 003 and 10^{-6} after 500 003: the rate, not the constant, is the limitation. From the bad initialization $\theta(\mathrm{G}, \cdot) = (0, 5)$, where $p_0 = 0.0067$ and the gap is 1.52, the natural gradient gives $p_1 = 0.087$, $p_2 = 0.59$, $p_3 = 0.97$, a gap below 10^{-2} at $k = 4$ and below 10^{-6} at $k = 7$: two iterations later than from the uniform policy, the price of a start at Kullback–Leibler distance $\ln(1/p_0) = 5.0$ from $\pi^\star$ instead of $\ln 2$—the proof of Theorem 14.13 goes through with this distance in place of $\ln A$. Gradient ascent with the guaranteed step is still at a gap of 1.49 after 10^5 iterations and needs 146 764 for a gap of 10^{-1} (75 with $\eta = 1$): the factor $\pi(a \mid s)$ of Proposition 14.8 is the plateau of Example 14.4 again (Figure 14.4).

Regularizing toward the previous policy: error averaging ♦. Corollary 14.12 read the natural gradient step as a regularized greedy step whose reference is the previous policy. The same device, applied to value iteration with approximation errors, turns regularization into a stabilizer of the approximate dynamic programming of Chapter 13: this is the idea of *dynamic policy programming* (Azar et al., 2012), of which what follows is a descendant. The *mirror-descent value iteration* of Geist et al. (2019) and Vieillard et al. (2020a)—written MD-VI(λ) here—alternates a regularized greedy step and a regularized evaluation step affected by an error $\epsilon_{k+1} \in \mathbb{R}^{\mathcal{S}\times\mathcal{A}}$ (the regression error of a fitted Q-iteration, Algorithm 13.1):

$$\begin{aligned} \pi_{k+1}(\cdot \mid s) &= \operatorname*{arg\,max}_{p \in \mathcal{M}_1(\mathcal{A})} \left\{ \langle p, Q_k(s,\cdot)\rangle - \lambda\, \mathrm{KL}(p \,\|\, \pi_k(\cdot \mid s)) \right\} \;\propto\; \pi_k(\cdot \mid s)\, e^{Q_k(s,\cdot)/\lambda}, \\ Q_{k+1} &= r + \gamma P \left[\langle \pi_{k+1}, Q_k\rangle - \lambda\, \mathrm{KL}(\pi_{k+1} \,\|\, \pi_k) \right] + \epsilon_{k+1}, \end{aligned} \tag{14.4}$$

started from the uniform policy π_0. Unrolling the closed form (Corollary 14.12), $\pi_{k+1} \propto \exp\left(\frac{1}{\lambda}\sum_{j=0}^{k} Q_j\right)$: the policy is a softmax of the *sum* of all past action values (the dual averaging of Section 14.3), which is what makes the errors average. This closed form and the averaging of the errors are already those of

Azar et al. (2012); Theorem 14.15 is the sharpened form given by Vieillard et al. (2020a).

Theorem 14.15 (Error averaging, Vieillard et al., 2020a). Let rewards lie in $[0, R_{\max}]$ as in Chapter 13, $V_{\max} = R_{\max}/(1-\gamma)$, and suppose $\|Q_k\|_\infty \le V_{\max}$ for all k (which clipping enforces). Then for every $k \ge 1$,

$$\|Q^\star - Q^{\pi_k}\|_\infty \;\le\; \frac{2}{1-\gamma}\left\|\frac{1}{k}\sum_{j=1}^{k}\epsilon_j\right\|_\infty + \frac{4\,(R_{\max} + \lambda \ln A)}{(1-\gamma)^2\,k},$$

whereas approximate value iteration without regularization satisfies, in the summed form established in the proof of Theorem 13.4, $\|Q^\star - Q^{\pi_k}\|_\infty \le \frac{2\gamma}{(1-\gamma)^2}\,(1-\gamma)\sum_{j\le k}\gamma^{k-j}\|\epsilon_j\|_\infty + \frac{2\gamma^{k+1}}{1-\gamma}V_{\max}$.

Two things change. The horizon factor multiplying the errors is $1/(1-\gamma)$ instead of $1/(1-\gamma)^2$. And the errors enter through the norm of their *average* $\frac{1}{k}\sum_j \epsilon_j$ instead of a discounted sum of their norms: errors of zero mean—the statistical noise of a regression on fresh samples—cancel at rate $1/\sqrt{k}$ rather than accumulate, only their bias survives. The price is the second term: convergence in $1/k$ instead of γ^k, the trust region slowing the policy down; the term also grows with λ. The structure is that of the natural policy gradient (Proposition 14.11), which is the policy-iteration variant of the same scheme, and the $1/k$ term is, up to horizon factors, the $\ln A/(\eta K)$ of Theorem 14.13. In deep reinforcement learning the scheme is implemented implicitly: since $\langle p, Q\rangle - \lambda\,\mathrm{KL}(p \,\|\, \pi_k) = \langle p, Q + \lambda \ln \pi_k\rangle + \lambda\mathcal{H}(p)$, regularizing toward the previous policy amounts to adding the "Munchausen" bonus $\lambda \ln \pi_k(a \mid s)$ to the reward and running an entropy-regularized (soft) update (Vieillard et al., 2020b); the analysis of such soft and averaged operators is due to Kozuno et al. (2019) (Exercise 14.8).

Trust regions: TRPO and PPO. In the general parametrized case, a natural gradient step is the maximization of the linearized performance under a constraint on the Kullback–Leibler divergence. The *surrogate objective*

$$L_{\pi_k}(\pi) \;\coloneqq\; V^{\pi_k}(\mu) + \frac{1}{1-\gamma}\mathbb{E}_{s\sim d_\mu^{\pi_k}}\Big[\sum_a \pi(a \mid s) A^{\pi_k}(s,a)\Big] \tag{14.5}$$

coincides with $V^\pi(\mu)$ to first order in $\pi - \pi_k$ (by Lemma 14.1, the difference is that the states are sampled from $d_\mu^{\pi_k}$ instead of d_μ^π), and it is a lower bound up to a penalty on the divergence between the policies.

Theorem 14.16 (Trust-region bound, Kakade and Langford, 2002; Schulman et al., 2015). For every $\pi \in \Pi_{\mathrm{SR}}$, with $\|A^{\pi_k}\|_\infty = \max_{s,a}|A^{\pi_k}(s,a)|$,

$$V^\pi(\mu) \;\ge\; L_{\pi_k}(\pi) - \frac{2\gamma\,\|A^{\pi_k}\|_\infty}{(1-\gamma)^2}\max_s \mathrm{KL}(\pi_k(\cdot \mid s) \,\|\, \pi(\cdot \mid s))\,.$$

Proof sketch. By Lemma 14.1, $V^\pi(\mu) - L_{\pi_k}(\pi) = \frac{1}{1-\gamma}\sum_s \big(d_\mu^\pi(s) - d_\mu^{\pi_k}(s)\big)\bar{A}(s)$ with $\bar{A}(s) = \sum_a \pi(a \mid s)A^{\pi_k}(s,a)$, and $|\bar{A}(s)| \le 2\|A^{\pi_k}\|_\infty\alpha$ where $\alpha = \max_s \mathrm{TV}(\pi_k(\cdot \mid s), \pi(\cdot \mid s))$, because $\sum_a \pi_k(a \mid s)A^{\pi_k}(s,a) = 0$. Couple the two processes so that they take the same action with probability $\ge 1-\alpha$ at each step: the occupation measures then differ, at time t, by at most $1-(1-\alpha)^{t-1} \le (t-1)\alpha$ in total variation, so $\sum_s |d_\mu^\pi(s) - d_\mu^{\pi_k}(s)| \le 2(1-\gamma)\sum_t \gamma^{t-1}(t-1)\alpha = 2\gamma\alpha/(1-\gamma)$. Combining, $|V^\pi(\mu) - L_{\pi_k}(\pi)| \le 4\gamma\|A^{\pi_k}\|_\infty\alpha^2/(1-\gamma)^2$, and Pinsker's inequality $\mathrm{TV}^2 \le \frac{1}{2}\mathrm{KL}(\cdot\,\|\,\cdot)$ (Theorem D.12) gives the statement. Schulman et al. (2015) state the bound with the constant $4\gamma\|A^{\pi_k}\|_\infty/(1-\gamma)^2$, through a slightly different coupling argument. Exercise 14.9 fills in the details. ∎

A word on orientations. The mirror-descent step of Proposition 14.11 penalizes KL(new ‖ old), whereas the trust-region bound uses KL(old ‖ new), the orientation in which the coupling argument and Pinsker's inequality are applied; by (14.1) both equal $\frac{1}{2}(\theta-\theta_k)^\top F_s(\theta_k)(\theta-\theta_k)$ to second order, which is why the two readings of the natural gradient agree.

Maximizing the right-hand side over π guarantees monotone improvement (the value of the right-hand side at $\pi=\pi_k$ is $V^{\pi_k}(\mu)$), but the penalty is too conservative in practice. *Trust region policy optimization* (Schulman et al., 2015) maximizes $L_{\pi_k}(\pi_\theta)$ subject to $\mathbb{E}_{d_\mu^{\pi_k}}\mathrm{KL}(\pi_k(\cdot\mid s)\,\|\,\pi_\theta(\cdot\mid s))\le\delta_{\mathrm{TR}}$, by a natural gradient step (the constraint is, to second order, $\frac{1}{2}(\theta-\theta_k)^\top F(\theta_k)(\theta-\theta_k)\le\delta_{\mathrm{TR}}$) followed by a line search that enforces the constraint and the improvement of the surrogate.

Example 14.17 (The averaged constraint buys nothing). Replacing the maximum by an average is not a technical convenience: it voids the guarantee (Agarwal et al., 2026, Sect. 12.2.2). Take three states s_0, s_1 and a sink. At s_0 the action good leads to s_1 and bad to the sink, both with reward 0; at s_1, good pays 1 and bad pays -1, both then absorbing (adding 1 to every reward, sink included, would make them nonnegative and shift every value by $1/(1-\gamma)$, changing nothing below). Let π play bad at s_0 and good at s_1, so that $V^\pi(s_0)=0$ and—this is the point—$d_{s_0}^\pi$ gives *no* mass to s_1, which π never reaches. Let π' play good at s_0 with probability α and bad at s_1 with probability one. Since s_1 carries no weight under $d_{s_0}^\pi$, the averaged divergence $\mathbb{E}_{s\sim d_{s_0}^\pi}\mathrm{KL}(\pi(\cdot\mid s)\,\|\,\pi'(\cdot\mid s))$ of the trust region, which equals $(1-\gamma)\ln\frac{1}{1-\alpha}$ (only s_0 contributes), is as small as one pleases for small α, so π' satisfies the constraint—whereas the maximum over states in Theorem 14.16 is infinite, the divergence at s_1 being $\mathrm{KL}(\delta_{\mathrm{good}}\,\|\,\delta_{\mathrm{bad}})=+\infty$. The surrogate improves, $L_\pi(\pi')=\alpha\gamma>0$, because reaching s_1 looks good when evaluated with π's continuation. But $V^{\pi'}(s_0)=-\alpha\gamma<0$ for every $\alpha>0$: the new policy is *worse*, and monotone improvement fails at every radius.

The mechanism is the one of Chapter 13: a constraint measured under the old policy's occupation says nothing about the states that policy never visits, and those are exactly the states the new policy will reach once it changes. The maximum over states in Theorem 14.16 is not conservatism, it is what makes the bound true. The objectives used for language models (Section 16.7) all constrain an average too, but under the occupation of the *new* policy, which changes the conclusion; see the discussion there.

Proximal policy optimization (Schulman et al., 2017) replaces the constraint by a clipped surrogate: with the likelihood ratio $\rho_\theta(s,a)=\pi_\theta(a\mid s)/\pi_k(a\mid s)$, it maximizes

$$\mathbb{E}_{s\sim d_\mu^{\pi_k},a\sim\pi_k}\left[\min\left(\rho_\theta A^{\pi_k},\ \mathrm{clip}(\rho_\theta,1-\epsilon_{\mathrm{clip}},1+\epsilon_{\mathrm{clip}})A^{\pi_k}\right)\right] \tag{14.6}$$

by a few epochs of stochastic gradient ascent on a batch of trajectories, which removes the incentive to move the ratio outside $[1-\epsilon_{\mathrm{clip}},1+\epsilon_{\mathrm{clip}}]$ in the direction that would improve the surrogate. PPO is first-order, simple to implement, and the workhorse of deep reinforcement learning—including the fine-tuning of language models from human feedback of Chapter 16. Algorithm 14.3 writes it out as it is run: a batch of trajectories is collected with the current policy, the advantages are estimated with a critic, and the clipped surrogate is maximized for a few epochs of stochastic gradient steps on that batch, which is where the ratio ρ_θ moves away from 1.

How far can the surrogate be trusted, and how conservative is the bound? The last example of this section measures both on a problem where everything is exact.

Example 14.18 (The surrogate against the true value along a step). Take the discounted riverswim of Example 8.16 ($N=4$, five states, $p=0.3$, $q=0.6$, the bonus 0.005 for drifting left in state 0) with $\gamma=0.95$, so that $1/(1-\gamma)=20$, $\mu=\delta_0$ and tabular softmax policies; the

Algorithm 14.3: Proximal policy optimization (clipped surrogate)

Input: actor π_θ, critic V_w, batch of T steps, epochs E, minibatch size B, clipping ϵ_{clip}, weights c_V, c_{H} of the critic loss and of the entropy bonus

for $k = 1, 2, \ldots$ **do**
  run π_{θ_k} for T steps (several episodes); store (S_t, A_t, R_t) and the ratios' denominators $\pi_{\theta_k}(A_t \mid S_t)$
  estimate the advantages $\hat{A}_t$ (generalized advantage estimation with V_w, Section 14.2) and the targets $\hat{G}_t = \hat{A}_t + V_w(S_t)$
  for *E epochs, over minibatches of B steps* **do**
    $\rho_t \leftarrow \pi_\theta(A_t \mid S_t)/\pi_{\theta_k}(A_t \mid S_t)$
    take a gradient step (Adam) on $\frac{1}{B}\sum_t \left[\min\left(\rho_t\hat{A}_t,\ \text{clip}(\rho_t, 1-\epsilon_{\text{clip}}, 1+\epsilon_{\text{clip}})\hat{A}_t\right) - c_V\left(V_w(S_t) - \hat{G}_t\right)^2 + c_{\text{H}}\mathcal{H}(\pi_\theta(\cdot \mid S_t))\right]$ in (θ, w)
  $\theta_{k+1} \leftarrow \theta$

optimal policy swims right from every state and $V^\star(0) = 6.98$. Everything below is exact (five states) and printed by `figures/scripts/surrogate_riverswim.py`. At a policy π_k we follow the natural-gradient direction $\theta = \theta_k + \eta A^{\pi_k}/(1-\gamma)$, that is, $\pi_\theta \propto \pi_k e^{\eta A^{\pi_k}/(1-\gamma)}$ by (14.2), and compare, as functions of the step size η (Figure 14.5), the true value $V^{\pi_\theta}(\mu)$, the surrogate $L_{\pi_k}(\pi_\theta)$ of (14.5), the lower bound of Theorem 14.16, and the clipped objective of PPO of (14.6) with $\epsilon_{\text{clip}} = 0.2$, shifted by $V^{\pi_k}(\mu)$ and scaled by $1/(1-\gamma)$ so that it coincides with L_{π_k} as long as no ratio is clipped.

From the uniform policy (left). Here $V^{\pi_k}(0) = 0.095$ and the swimmer is mostly in state 0: $d_\mu^{\pi_k} = (0.76, 0.18, 0.045, 0.011, 0.003)$. The advantage of swimming right is -0.0003 in state 0 (the bonus wins, barely), 0.44 in state 3 and 0.76 in state 4. The surrogate and the true value share their slope at $\eta = 0$ but part almost at once: they agree to 1% up to $\eta = 0.006$ only, to 5% up to $\eta = 0.014$. Beyond, the surrogate saturates at 0.40—it can only reallocate the advantages of π_k over the states that π_k visits, and π_k hardly visits the states where swimming right pays—while the true value climbs to 3.86 at $\eta = 1$ and to its maximum 5.77 at $\eta = 7.6$ (then decreases slowly, the tilt toward L in state 0 eventually prevailing): the improvement is 19 times the one announced. The surrogate is the first-order model of the performance under the *old* occupation measure; when the step takes the policy to states the old policy did not visit, it is blind to what happens there, in one direction or the other. The lower bound of Theorem 14.16 subtracts from the surrogate, with $\|A^{\pi_k}\|_\infty = 0.76$, 579 times the largest divergence over states, which is driven by state 4, where the tilt $e^{\eta A/(1-\gamma)}$ is the strongest and whose occupation is 0.003. The bound is maximized at $\eta = 1.2 \cdot 10^{-5}$, with a guaranteed improvement of 10^{-5} against 5.67 for the best step along the line, a factor $5 \cdot 10^5$; at $\eta = 1$ it equals $-8\,400$. The PPO objective is flat (to 10^{-3}) from $\eta = 0.47$ on: the ratio π_θ/π_k of the action R reaches 1.2 at $\eta = 0.013$, 0.023, 0.10 and 0.50 in states 4, 3, 2 and 1, and never in state 0. Where it stops, the true value is 2.66, 45% of the best improvement, and the surrogate, 0.31, is 88% below it: clipping has stopped the ascent far beyond the region where the surrogate is reliable—and far short of the best step.

From the first natural-gradient iterate (right), $\pi_1 \propto e^{A^{\pi_0}/(1-\gamma)}$, which swims right with probabilities $(0.50, 0.69, 0.98, 1, 1)$ and has $V^{\pi_1}(0) = 3.86$ and $d_\mu^{\pi_1} = (0.51, 0.13, 0.08, 0.09, 0.18)$: the swimmer now reaches the right bank, and the only sizeable advantages left are those of R in states 0 and 1, 0.19 and 0.68. The surrogate is now reliable for larger steps—within 1% of the true value up to $\eta = 0.088$, where $V = 6.29$, within 5% up to $\eta = 0.20$—and the true value reaches 6.978 at $\eta = 0.95$ and the optimum beyond. The surrogate keeps rising to 7.73, above $V^\star(0)$, a

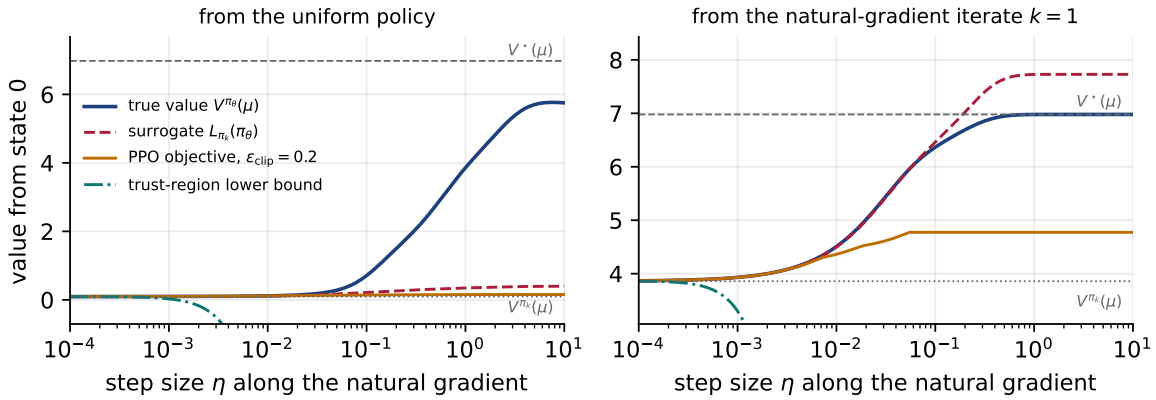


Figure 14.5. Riverswim ($\gamma = 0.95$, $\mu = \delta_0$): true value, surrogate (14.5), PPO clipped objective ($\epsilon_{\text{clip}} = 0.2$) and trust-region lower bound of Theorem 14.16 along the natural-gradient direction from the uniform policy (left) and from the first natural-gradient iterate (right), as functions of the step size η (Example 14.18). The dotted line is $V^{\pi_k}(\mu)$, the dashed line $V^\star(\mu)$; the lower bound leaves the frame before $\eta = 4 \cdot 10^{-3}$.

level no policy attains: it credits state 0, of occupation 0.51 under π_1, with the advantage 0.19 of swimming right as if the swimmer, having left state 0, still returned to it as often as under π_1 (under the new policy, $d_\mu(0) = 0.20$). Once the step changes the occupation measure, the surrogate overestimates. The bound is again useless—constant 3 040 with $\|A^{\pi_1}\|_\infty = 4.0$, maximized at $\eta = 5.7 \cdot 10^{-5}$ with a guaranteed gain of 0.002 against 3.12 for the best step, a factor $1.5 \cdot 10^3$. The PPO objective flattens at $\eta = 0.054$, when the ratio of R in state 0 reaches 1.2 (state 1 was clipped at $\eta = 0.018$, states 2 to 4 already swim right): there the true value is 5.94, 67% of the available improvement, and the surrogate agrees with it to $3 \cdot 10^{-4}$. Clipping thus stops the ascent well inside the reliable region, before the overshoot—the next PPO iteration, from the new policy, collects the rest. On this five-state example the bound of Theorem 14.16 is correct and loose by three to five orders of magnitude (its constant contains $1/(1-\gamma)^2 = 400$ and the *maximum* over states of the divergence, which is largest exactly at the states the step moves most, whatever their weight); the clipping of PPO is a trust region in disguise, measured per state-action pair rather than by a divergence, whose radius was adequate from π_1 (a maximal divergence of 0.4 over states) and far too cautious from the uniform policy (6.5)—in both cases without any guarantee, and in both cases the direction was good enough for an excess of step to cost nothing.

14.4 Policy gradients for utilities, β-means and risk measures

The score-function identity (Lemma 14.2) holds for any function of the trajectory, hence for any functional of the return: if $J_\psi(\theta) = \psi(v^{\pi_\theta})$ is a differentiable functional of the law of W_H, its gradient is computable by weighting the scores by the influence of the trajectory on ψ. We work in the finite-horizon setting of Chapter 3, with the criteria of Chapter 5, time-indexed policies $\pi_{\theta,t}$ and $J_\psi(\theta) = \psi(v^{\pi_\theta}(\mu))$, where $v^\pi(\mu)$ is the law of W_H under $\mathbb{P}_\mu^\pi$.

Proposition 14.19 (Policy gradient for utilities). Let u be a bounded utility and $J_u(\theta) =$

$\mathbb{E}_\mu^{\pi_\theta}[u(W_H)]$. Then

$$\nabla J_u(\theta) \;=\; \mathbb{E}_\mu^{\pi_\theta}\Big[\big(u(W_H) - b\big)\sum_{t=1}^{H}\nabla_\theta \ln \pi_{\theta,t}(A_t \mid S_t)\Big]$$

for every constant $b \in \mathbb{R}$, and more generally b may be replaced in the t-th term by any function $b_t(S_t)$. What is not available in general is the reward-to-go decomposition: $u(W_H)$ is not a sum over time, and no value function offers itself as a natural baseline.

Proof. Lemma 14.2 with $f = u(W_H)$, and $\mathbb{E}[\nabla \ln \pi_{\theta,t}(A_t \mid S_t)\, b_t(S_t)] = 0$ as in Exercise 14.2. ∎

For the exponential utility the situation is better: the β-mean is Bellman closed (Theorem 4.15), and its gradient has a Bellman form in which the occupation measure is replaced by its exponentially tilted version. Recall from Definition 4.14 the exponential moments $Z^\pi_{\beta,t}(s,a) = \mathbb{E}^\pi_{t,s,a}[e^{\beta W_{t:H}}]$ and $Z^\pi_{\beta,t}(s) = \sum_a \pi_t(a \mid s) Z^\pi_{\beta,t}(s,a)$, and define their forward counterparts

$$Y^\pi_{\beta,t}(s) \;\coloneqq\; \mathbb{E}^\pi_\mu\big[e^{\beta W_{1:t-1}}\mathbb{1}\{S_t = s\}\big], \qquad Y^\pi_{\beta,1} = \mu,$$

which satisfy the forward recursion $Y^\pi_{\beta,t+1}(s') = \sum_{s,a} Y^\pi_{\beta,t}(s)\pi_t(a \mid s) P_t(s' \mid s,a) e^{\beta r_t(s,a,s')}$ (Exercise 14.10). Under the tilted trajectory law $\mathbb{P}^{\pi,\beta}_\mu = e^{\beta W_H}\mathbb{P}^\pi_\mu / Z^\pi_\beta$, $Z^\pi_\beta \coloneqq \mathbb{E}^\pi_\mu[e^{\beta W_H}]$, of Remark 4.19, the law of (S_t, A_t) is the *tilted occupation measure*

$$d^{\pi,\beta}_t(s,a) \;=\; \frac{Y^\pi_{\beta,t}(s)\,\pi_t(a \mid s)\, Z^\pi_{\beta,t}(s,a)}{Z^\pi_\beta}, \tag{14.7}$$

since $\mathbb{E}[e^{\beta W_H}\mathbb{1}\{S_t = s, A_t = a\}] = \mathbb{E}[e^{\beta W_{1:t-1}}\mathbb{1}\{S_t = s\}]\pi_t(a \mid s)\mathbb{E}[e^{\beta W_{t:H}} \mid S_t = s, A_t = a]$ by the Markov property, $W_H = W_{1:t-1} + W_{t:H}$ with the two pieces conditionally independent given (S_t, A_t).

Theorem 14.20 (Policy gradient for β-means). For $\beta \neq 0$ and $J_\beta(\theta) = U_\beta(v^{\pi_\theta}(\mu)) = \frac{1}{\beta}\ln Z^{\pi_\theta}_\beta$,

$$\begin{aligned}\nabla J_\beta(\theta) \;&=\; \frac{1}{\beta}\,\mathbb{E}^{\pi_\theta,\beta}_\mu\Big[\sum_{t=1}^{H}\nabla_\theta \ln \pi_{\theta,t}(A_t \mid S_t)\Big] \\ &=\; \frac{1}{\beta\, Z^\pi_\beta}\sum_{t=1}^{H}\sum_s Y^\pi_{\beta,t}(s)\sum_a \nabla_\theta \pi_{\theta,t}(a \mid s)\,\big(Z^\pi_{\beta,t}(s,a) - Z^\pi_{\beta,t}(s)\big),\end{aligned}$$

where $\mathbb{E}^{\pi,\beta}_\mu$ is the expectation under the tilted law. For the tabular softmax, $\partial J_\beta/\partial\theta_t(s,a) = Y^\pi_{\beta,t}(s)\pi_t(a \mid s)\big(Z^\pi_{\beta,t}(s,a) - Z^\pi_{\beta,t}(s)\big)/(\beta Z^\pi_\beta)$. As $\beta \to 0$ the formulas tend to those of Theorem 14.3 in advantage form.

Proof. $\nabla \ln Z_\beta = \mathbb{E}[e^{\beta W_H}\sum_t \nabla \ln \pi_t(A_t \mid S_t)]/Z_\beta$ by Lemma 14.2, which is the tilted expectation of the sum of scores. Conditioning the t-th term on (S_t, A_t) as in (14.7) gives $\sum_{s,a} Y_{\beta,t}(s)\pi_t(a \mid s)\nabla \ln \pi_t(a \mid s) Z_{\beta,t}(s,a) = \sum_s Y_{\beta,t}(s)\sum_a \nabla\pi_t(a \mid s) Z_{\beta,t}(s,a)$, and $\sum_a \nabla\pi_t(a \mid s) = 0$ allows the subtraction of $Z_{\beta,t}(s)$. The softmax formula follows as in Proposition 14.8. For $\beta \to 0$: $Y_{\beta,t}(s) \to \mathbb{P}^\pi_\mu(S_t = s)$, $Z_\beta \to 1$ and $(Z_{\beta,t}(s,a) - Z_{\beta,t}(s))/\beta \to Q^\pi_t(s,a) - V^\pi_t(s)$. ∎

The theorem is the gradient counterpart of the exponential Bellman equation: the backward quantities are the exponential moments computed by Algorithm 4.2, the forward quantities by a mirror-image

recursion, and the gradient is assembled from both—as in the forward–backward algorithm of hidden Markov models. Two consequences follow. A *sampled* version: from N trajectories, estimate the gradient by $\frac{1}{\beta}\sum_{i\le N} w_i \sum_t \nabla \ln \pi_{\theta,t}(A_t^i \mid S_t^i)$ with the self-normalized tilting weights $w_i = e^{\beta W_H^i}/\sum_j e^{\beta W_H^j}$. This is a self-normalized form of the risk-sensitive REINFORCE of Noorani and Baras (2021), who work with the unnormalized weights $e^{\beta W}$, that is, with the gradient of $\mathbb{E}[e^{\beta W_H}]$ (see also Nass et al., 2019, and Borkar, 2001 for the actor–critic form). The estimate is consistent, biased at finite N by the normalization, and its variance is that of importance sampling with weights $e^{\beta W}$: for $\beta < 0$ the weights concentrate on the worst trajectories, whose number in a batch is small—the statistical price of risk aversion, already met in Proposition 2.29 (Exercise 14.11). And a *natural gradient* version for the softmax: dividing the gradient by the Fisher blocks $d_t^\pi(s)(\operatorname{diag}\pi_t - \pi_t\pi_t^\top)$, where $d_t^\pi(s) = \mathbb{P}_\mu^\pi(S_t = s)$, gives the update

$$\theta_t(s,a) \;\leftarrow\; \theta_t(s,a) + \eta\,\frac{d_t^{\pi,\beta}(s)}{d_t^\pi(s)}\cdot\frac{1}{\beta}\Big(\frac{Z_{\beta,t}(s,a)}{Z_{\beta,t}(s)} - 1\Big),$$

with $d_t^{\pi,\beta}(s) = Y_{\beta,t}(s)Z_{\beta,t}(s)/Z_\beta$ the tilted state distribution: a multiplicative-weights update on a tilted advantage, weighted by the likelihood ratio of the tilted to the original state distribution.

Example 14.21 (Policy gradient on the windy hike: β-means). We return to the setting of Example 14.10—the windy hike with time-indexed softmax policies, $14\times21\times4$ parameters initialized at 0, start state S—and replace the mean by the β-means $U_\beta(\theta)$ for $\beta \in \{-1, 0.5\}$, with the exact gradients of Theorem 14.20 (checked against finite differences; `figures/scripts/hike_pg.py`). The optimal β-values are 2.81 and 6.95, with paths on rows 4 and 2 (Example 5.10), against 5.09 on row 3 for the mean. Figure 14.6 shows $U_\beta(\theta_k)$ along the iterations, together with the $\beta = 0$ run of Example 14.10. Gradient ascent with step $\eta = 3$ sits on the initial plateau—the uniform policy almost never reaches the bar, so that all advantages are tiny—for some twenty iterations when $\beta = 0.5$ and for close to a thousand when $\beta = -1$, and then settles, as for $\beta = 0$, on a policy following *row* 2, the child's path: 6.89 for $\beta = 0.5$, and 1.05 for $\beta = -1$, far from the optimum 2.81. The reason is the one of Example 14.10, the factor $d_t^\pi(s)$—here the tilted visitation $Y_{\beta,t}^\pi(s)/Z_\beta^\pi$ of Theorem 14.20, which vanishes just as fast on rows 3 and 4 once the policy heads east along row 2: their decision rules stay essentially uniform, going north looks bad, and the exact gradient there is too small to move them. Continued to 10^5 iterations, it still follows row 2 for every β, with values 4.94, 6.91 and 1.06 for $\beta = 0$, 0.5 and -1 and an exact gradient of supremum norm below 10^{-5}; the gains between 10^4 and 10^5 iterations (0.010, 0.012, 0.007) refine the row-2 policy and do not move it toward another row, and the same holds at 10^6 iterations for $\beta = 0$ and $\beta = -1$ (4.944 and 1.062). This is an observation over a finite budget, not a proof that the plateau is never left (see Example 14.10). Natural gradient ascent with $\eta = 1$ removes the factor and reaches the optimal β-values in about a hundred iterations (2.811 and 6.954, as 5.095 for the mean), with the greedy paths on rows 4 and 2. Note that the risk-averse problem $\beta = -1$ is the slowest for both methods: in the exponential domain the informative trajectories are the rare ones.

The gradient of any functional of the return distribution ♦

Proposition 14.19, Theorem 14.20 and the conditional value at risk treated below are three instances of one statement. Let $\psi : \mathcal{M}_1(\mathbb{R}) \to \mathbb{R}$ be a functional of the return distribution and let $J_\psi(\theta) = \psi(\nu^{\pi_\theta}(\mu))$. What plays the role of the reward is the *influence function* of ψ, the derivative of ψ along a perturbation of ν toward a point mass:

$$\mathrm{IF}_\psi(y;\nu) \;\coloneqq\; \lim_{\epsilon\downarrow 0}\frac{\psi\big((1-\epsilon)\nu + \epsilon\delta_y\big) - \psi(\nu)}{\epsilon}, \tag{14.8}$$

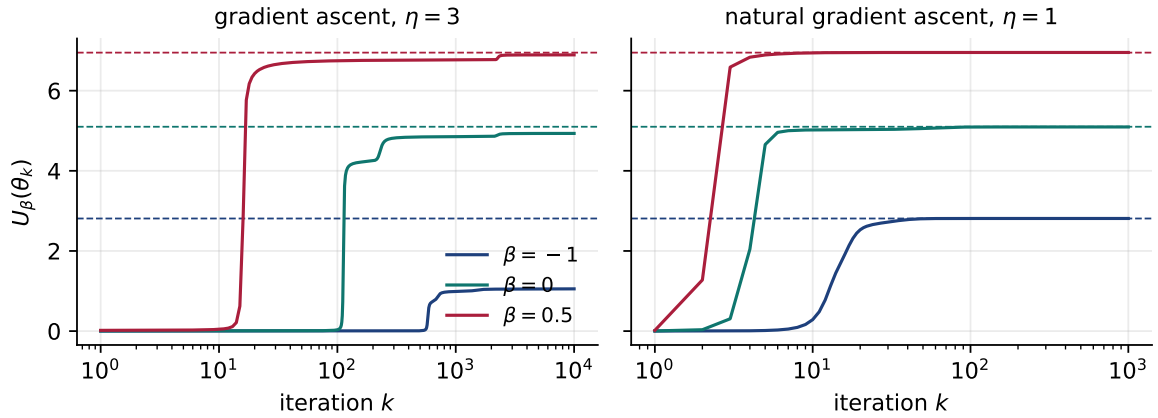


Figure 14.6. Gradient ascent (left) and natural gradient ascent (right) on the β-means of the windy hike with time-indexed softmax policies (Example 14.21); dashed lines mark the optimal β-values of Example 5.10. Over the 10 000 iterations shown, plain gradient ascent settles on the row-2 path for every β; the natural gradient finds the β-optimal paths.

the object of robust statistics (Hampel, 1974), normalized so that $\int \mathrm{IF}_\psi \, \mathrm{d}\nu = 0$.

Proposition 14.22 (Policy gradient of a distributional functional). Suppose ψ is Hadamard differentiable at $\nu^{\pi_\theta}(\mu)$—differentiable along every direction, uniformly over compact sets of directions, the derivative being the linear map $h \mapsto \int \mathrm{IF}_\psi \, \mathrm{d}h$—with influence function $\mathrm{IF}_\psi(\cdot\,; \nu)$ bounded, and that $\theta \mapsto \nu^{\pi_\theta}(\mu)$ is differentiable. Then

$$\nabla J_\psi(\theta) \;=\; \mathbb{E}_\mu^{\pi_\theta}\Big[\mathrm{IF}_\psi\big(W_H; \nu^{\pi_\theta}(\mu)\big) \sum_{t=1}^{H} \nabla_\theta \ln \pi_{\theta,t}(A_t \mid S_t)\Big].$$

Proof. By Hadamard differentiability, $\nabla_\theta \psi(\nu_\theta) = \int \mathrm{IF}_\psi(y; \nu_\theta) \, \nabla_\theta \nu_\theta(\mathrm{d}y)$, the derivative of the measure acting on the influence function. For any bounded g, Lemma 14.2 applied to $f(\tau) = g(W_H)$ gives $\nabla_\theta \int g \, \mathrm{d}\nu_\theta = \mathbb{E}_\mu^{\pi_\theta}[g(W_H) \sum_t \nabla_\theta \ln \pi_{\theta,t}]$; taking $g = \mathrm{IF}_\psi(\cdot\,; \nu_\theta)$—which is legitimate because the influence function is evaluated at the *current* θ, so that it does not itself have to be differentiated—yields the formula. ∎

The directions h in the statement are differences of laws, signed measures of total mass zero, and "differentiable" refers to a norm on them, such as the supremum norm of their distribution functions; the choice matters for the quantile and the CVaR, not for the smooth functionals below. In the setting of this chapter a simpler notion suffices: in a finite MDP with deterministic rewards, W_H takes its values in a finite set $\mathcal{W}$ whatever the policy, so that $\nu^{\pi_\theta}(\mu)$ is a point of the simplex $\mathcal{M}_1(\mathcal{W})$ depending smoothly on θ, and the proposition is the chain rule for a function ψ differentiable on that simplex in the ordinary sense, along the directions $h \in \mathbb{R}^{\mathcal{W}}$ with $\sum_y h(y) = 0$; the influence function is then its gradient, defined up to an additive constant that the normalization $\int \mathrm{IF}_\psi \, \mathrm{d}\nu = 0$ fixes.

Three readings of the same formula. For the mean, $\psi(\nu) = \int x \, \mathrm{d}\nu$ and $\mathrm{IF}(y) = y - \int x \, \mathrm{d}\nu$: one recovers the policy gradient theorem, with the mean return appearing as a baseline for free. For a utility, $\psi(\nu) = \int u \, \mathrm{d}\nu$ and $\mathrm{IF}(y) = u(y) - \int u \, \mathrm{d}\nu$, which is Proposition 14.19. For a β-mean, $\psi(\nu) = \frac{1}{\beta} \ln \int e^{\beta x} \, \mathrm{d}\nu$ and $\mathrm{IF}(y) = (e^{\beta y} - Z)/(\beta Z)$ with $Z = \int e^{\beta x} \, \mathrm{d}\nu$, which is Theorem 14.20 after the constant is dropped. The pattern is the one of Chapter 4 seen from the other side: *every* smooth functional has a policy gradient, given by its influence function, and what distinguishes the β-means is not that they have a

gradient but that the gradient itself satisfies a Bellman recursion with *one* forward and *one* backward quantity—the influence function of a β-mean is affine in $e^{\beta W_H} = \prod_t e^{\beta R_t}$, which factorizes along the trajectory (Exercise 14.15). A polynomial functional such as the variance still has a forward–backward gradient, assembled from the finitely many coupled moments of Proposition 4.1; a quantile or a CVaR needs the whole conditional law of the reward-to-go. Estimating ∇J_ψ requires estimating $\mathrm{IF}_\psi(\cdot\,; \nu_\theta)$, that is, a functional of the *current* return distribution, which is where distributional critics earn their keep (Remark 14.24).

Conditional value at risk. For the tail measures of Chapter 6, the likelihood-ratio trick applies through the variational representations. Consider CVaR_δ in the gain convention (Definition 6.9), and suppose that the maximizer in the Rockafellar–Uryasev representation (Proposition 6.10) is unique: the set of q with $\mathbb{P}_\theta(W_H < q) \le \delta \le \mathbb{P}_\theta(W_H \le q)$ is the single point $q_\theta = \mathrm{VaR}_\delta(W_H)$. This holds when the law of W_H has a density that is continuous and *positive* near q_θ, and also for a finite-support return, such as that of the hike, as soon as δ is not one of the values of its distribution function. A density alone is not enough: if it vanishes on an interval on which the distribution function equals δ, the maximizers form that interval and the CVaR, as a function of θ, can have a kink (two one-sided derivatives, given by Proposition C.15); the same happens for a finite-support return when δ is one of the values of its distribution function. The quantile itself needs more: it is differentiable in θ when the distribution function $F_\theta(q)$ is continuously differentiable in (θ, q) near (θ, q_θ) with $\partial_q F_\theta(q_\theta) > 0$, by the implicit function theorem applied to $F_\theta(q) = \delta$ (Exercise 14.15(b)), whereas for a finite-support return it is piecewise constant in θ and jumps where the uniqueness fails.

The uniqueness of the maximizer is also what makes CVaR_δ Hadamard differentiable, so that Proposition 14.23 below is indeed a case of Proposition 14.22. Let the returns lie in a fixed interval $[a, b]$ (bounded rewards), identify a law with its distribution function F, and write $\mathrm{CVaR}_\delta(F) = \max_{q \in [a,b]} \phi(q, F)$ with $\phi(q, F) = q - \frac{1}{\delta}\int_a^q F(x)\,\mathrm{d}x$ (Proposition 6.10, since $\mathbb{E}[(q - W)_+] = \int_a^q F(x)\,\mathrm{d}x$); this formula extends CVaR_δ to all bounded functions F. Let $t \downarrow 0$ and $H_t \to H$ in supremum norm. The point q_θ gives the lower bound $\mathrm{CVaR}_\delta(F + tH_t) \ge \mathrm{CVaR}_\delta(F) - \frac{t}{\delta}\int_a^{q_\theta} H_t$, and a maximizer q_t of $\phi(\cdot, F + tH_t)$ the upper bound $\mathrm{CVaR}_\delta(F + tH_t) \le \mathrm{CVaR}_\delta(F) - \frac{t}{\delta}\int_a^{q_t} H_t$; since $\phi(\cdot, F + tH_t) \to \phi(\cdot, F)$ uniformly on $[a, b]$ and the maximizer of the limit is unique, $q_t \to q_\theta$. The derivative is therefore $H \mapsto -\frac{1}{\delta}\int_a^{q_\theta} H(x)\,\mathrm{d}x$, which, for H the distribution function of a signed measure h of total mass 0 on $[a, b]$, is $\int \mathrm{IF}_\delta\,\mathrm{d}h$ after an integration by parts, with the influence function of Proposition 14.23. When the maximizers form an interval, the same two bounds give only directional derivatives, the maximum of $-\frac{1}{\delta}\int_a^q H$ over that interval: the kink above. For unbounded returns, weighted supremum norms and a weaker notion of differentiability are needed (Beutner and Zähle, 2010).

Proposition 14.23 (CVaR policy gradient, Tamar et al., 2015). Under these assumptions, with $J_\delta(\theta) = \mathrm{CVaR}_\delta(\nu^{\pi_\theta}(\mu))$—the case of Proposition 14.22 with $\mathrm{IF}_\delta(y) = \frac{1}{\delta}(y - q_\theta)\mathbb{1}\{y \le q_\theta\} + q_\theta - \mathrm{CVaR}_\delta$, where $q_\theta = \mathrm{VaR}_\delta(\nu^{\pi_\theta}(\mu))$—

$$\nabla J_\delta(\theta) \;=\; \frac{1}{\delta}\,\mathbb{E}_\mu^{\pi_\theta}\Big[\big(W_H - q_\theta\big)\mathbb{1}\{W_H \le q_\theta\}\textstyle\sum_{t=1}^{H} \nabla_\theta \ln \pi_{\theta,t}(A_t \mid S_t)\Big].$$

Proof sketch. By the Rockafellar–Uryasev representation (Proposition 6.10), $J_\delta(\theta) = \max_q\{q - \frac{1}{\delta}\mathbb{E}_\theta[(q - W_H)_+]\}$, attained at $q = q_\theta$. By the envelope theorem (Danskin's theorem, Proposition C.15: the derivative of a maximum with respect to a parameter is the partial derivative of the objective at the maximizer, when the maximizer is unique), the gradient in θ is the partial gradient of the objective at the maximizer, $-\frac{1}{\delta}\big[\nabla_\theta \mathbb{E}_\theta[(q - W_H)_+]\big]_{q=q_\theta}$, and Lemma 14.2 with $f = (q_\theta - W_H)_+ = (q_\theta - W_H)\mathbb{1}\{W_H \le q_\theta\}$ gives the formula. The differentiability in θ of $q \mapsto \mathbb{E}_\theta[(q - W_H)_+]$ and the uniqueness of the maximizer are where the assumptions are used; Tamar et al. (2015) give the details and the extension to the case

where the return is a function of the trajectory with a continuous component. ■

The sample version keeps, in a batch of N trajectories, the $\lceil \delta N \rceil$ worst returns, and applies REINFORCE to them with the weights $W_H^i - \hat{q}$, $\hat{q}$ the empirical quantile: only a fraction δ of the data is used, and the variance grows like $1/\delta$—the same phenomenon as the tilting weights above, in a cruder form. The entropic value at risk (Definition 6.13) is the supremum $\sup_{\beta>0}\{U_{-\beta}(W_H) - \ln(1/\delta)/\beta\}$ of risk-averse β-means, at the negative tilt $-\beta$, shifted by $-\ln(1/\delta)/\beta$; if the supremum is attained at a unique $\beta^\star(\theta) \in (0,\infty)$, the envelope theorem gives $\nabla \operatorname{EVaR}_\delta = \nabla J_{-\beta^\star}(\theta)$: the tilted gradient of Theorem 14.20 at the optimal tilt, which uses *all* the trajectories with weights $e^{-\beta^\star W}$ rather than a fraction δ of them (Exercise 14.12). Variance-penalized criteria $\mathbb{E}[W] - \lambda \mathbb{V}\mathrm{ar}[W]$ are handled likewise, $\nabla \mathbb{V}\mathrm{ar}[W] = \nabla \mathbb{E}[W^2] - 2\mathbb{E}[W]\nabla \mathbb{E}[W]$ being two REINFORCE gradients (Tamar et al., 2012). Their critics exist: the mean and the second moment of the reward-to-go satisfy the coupled recursions of Proposition 4.1, and, since the past and the future of a Markov policy are conditionally independent given (S_t, A_t), $\mathbb{E}[W_H^2 \mid \text{past}, S_t, A_t] = W_{1:t-1}^2 + 2W_{1:t-1}Q_t^\pi(S_t, A_t) + M_t^{(2)}(S_t, A_t)$, with $M_t^{(2)}$ the second moment of $W_{t:H}$ given (S_t, A_t). What the criterion lacks is not a gradient but the optimization structure: Example 5.16 shows that, unlike the β-means, it is not Bellman optimizable, so that no greedy or policy-iteration reading of the ascent is available.

Remark 14.24 (Distributional critics). In actor–critic methods the critic may be distributional (Section 13.9): the actor then uses the mean of the learned return distribution as its advantage—as in the distributed distributional deterministic policy gradient of Barth-Maron et al. (2018) and in the racing agent of Wurman et al. (2022) mentioned in Section 1.2, whose quantile critic serves to stabilize the learning—or a risk-sensitive statistic of it. When the statistic is a β-mean, Theorem 14.20 says which weights the actor should use. When it is a CVaR or a distorted expectation, three situations must be distinguished. The objective itself, say $J_\delta(\theta) = \operatorname{CVaR}_\delta(\nu^{\pi_\theta}(\mu))$, is well defined and, under the assumptions above, has the exact gradient of Proposition 14.23. An *exact* distributional critic delivers it: conditioning the t-th score on the history up to (S_t, A_t), the weight $\frac{1}{\delta}(W_H - q_\theta)\mathbb{1}\{W_H \le q_\theta\}$ may be replaced by its conditional expectation, a function of the reward $W_{1:t-1}$ already collected, of the quantile q_θ of the return from the start, and of the law of $W_{t:H}$ given (S_t, A_t)—the critic's output; an approximate critic gives an approximate gradient, biased as the critic of Section 14.2 is. What is *not* a gradient of J_δ is the common shortcut that weights the score by the CVaR of the critic's law at (S_t, A_t) alone, ignoring the past rewards and the global quantile: since the criterion is not Bellman optimizable (Theorem 5.5), no such per-state risk-to-go is consistent with it, and the resulting actor–critic loop is a heuristic—often a good one—rather than an ascent on J_δ, exactly as greedy selection was in Section 8.4.

Distributional viewpoint

Policy gradient methods optimize a functional of the return distribution, and the likelihood-ratio trick computes the gradient of *any* such functional from sampled trajectories: utilities, β-means, conditional and entropic values at risk. What distinguishes the exponential utilities is that their gradient has a Bellman form, with the exponential moments as backward quantities and the exponentially tilted occupation measures as forward ones—the same closure property that made them plannable in Chapter 5 and learnable in Chapter 12 makes their gradients decomposable over time. The price of risk sensitivity is statistical: the tilting weights $e^{\beta W}$ concentrate the information on rare trajectories. And the natural gradient, which removes the occupation measure from the update, is what allows a policy to discover paths it does not yet visit—for the mean and for the β-means alike.

Hands-on

Session H06 (`cartpole.py`) trains the actor–critic of Section 14.2 on the cart-pole: a shared layer of 128 units, a softmax actor over two actions, a scalar critic, the returns standardized within each episode, a smooth L^1 loss for the critic and Adam with step 0.03. Replace the standardized returns by raw returns, then remove the critic (REINFORCE with and without a running-average baseline), and compare the learning curves over several seeds: the variance reduction of the baseline is visible, the bias of the critic less so on this problem. Then implement the risk-sensitive REINFORCE of Section 14.4 with $\beta < 0$ on the episode length, taking $|\beta|$ of the order of the inverse of a typical return (episodes last up to 500 steps) and computing the weights through a log-sum-exp: a risk-averse agent prefers balancing reliably for a moderate time to occasionally balancing for a long one. On the hike, reproduce Figure 14.6 from the exact gradients (`hike_pg.py`, `hike_pg_mean.py`), then replace them by REINFORCE estimates from N episodes and observe that the noise does not so much lengthen the plateaus as freeze the ascent at premature deterministic policies (Example 14.10); try a smaller step.

Exercises

Exercise 14.1 (The performance difference lemma at work). [verify] (a) Deduce Theorem 8.5 from Lemma 14.1 (*hint:* for π_V greedy for V, A^{π_V} and $A^\star$ are related through $V - V^\star$). (b) Deduce the policy improvement theorem (Proposition 8.4). (c) Check that the finite-horizon identity of the lemma still holds when the comparison policy π' is history-dependent.

Exercise 14.2 (Causality and baselines). [verify] (a) Show that $\mathbb{E}[\nabla \ln \pi_{\theta,t}(A_t \mid S_t)\, R_k] = 0$ for $k < t$ and $\mathbb{E}[\nabla \ln \pi_{\theta,t}(A_t \mid S_t)\, b_t(S_t)] = 0$ for any function b_t; deduce the reward-to-go form of REINFORCE from Lemma 14.2. (b) Show that $\mathbb{E}[\nabla \ln \pi_{\theta,t}(A_t \mid S_t)\, W_{t:H} \mid S_t, A_t] = \nabla \ln \pi_{\theta,t}(A_t \mid S_t)\, Q_t^\pi(S_t, A_t)$ and recover the finite-horizon statement of Theorem 14.3.

Exercise 14.3 (Optimal baselines). [verify] For a single state and a one-step problem, $\hat{g} = \nabla \ln \pi_\theta(A)(R - b)$ with $A \sim \pi_\theta$. (a) Show that the variance of $\hat{g}$ (trace of the covariance) is minimized by $b^\star = \mathbb{E}[\|\nabla \ln \pi_\theta(A)\|^2 R]/\mathbb{E}[\|\nabla \ln \pi_\theta(A)\|^2]$. (b) Compare with $b = \mathbb{E}[R]$ on an example with two actions, and explain why V^π is a good but not optimal baseline (Greensmith et al., 2004). (c) Why is a baseline depending on A_t not allowed?

Exercise 14.4 (The discount factor in REINFORCE). [limits] In the discounted setting with truncation at H, show that $\sum_t \gamma^{t-1} \nabla \ln \pi_\theta(A_t \mid S_t) W_{t:H}$ with $W_{t:H} = \sum_{k=t}^{H} \gamma^{k-t} R_k$ is (up to the truncation) an unbiased estimate of ∇J, whereas the common estimate $\sum_t \nabla \ln \pi_\theta(A_t \mid S_t) W_{t:H}$, which drops the factor γ^{t-1}, is in general the gradient of no function of θ at all (Nota and Thomas, 2020): compute its expectation on a two-state example and exhibit the missing derivative of the state distribution. When does the difference matter little?

Exercise 14.5 (Compatible function approximation). [verify] Let $Q_w(s,a) = w^\top \nabla_\theta \ln \pi_\theta(a \mid s)$ and let $w^\star$ be a minimizer of $\mathbb{E}_{s \sim d_\mu^\pi, a \sim \pi}[(Q_w(s,a) - Q^\pi(s,a))^2]$. Show that $\mathbb{E}_{d_\mu^\pi \otimes \pi}[\nabla \ln \pi_\theta(a \mid s) Q_{w^\star}(s,a)] = \mathbb{E}_{d_\mu^\pi \otimes \pi}[\nabla \ln \pi_\theta(a \mid s) Q^\pi(s,a)]$, so that the policy gradient computed with the critic $Q_{w^\star}$ is exact (Sutton et al., 2000). Show also that the minimizers are the solutions of

$F(\theta)w = (1-\gamma)\nabla J(\theta)$, that they form the affine space $(1-\gamma)F(\theta)^{+}\nabla J(\theta) + \ker F(\theta)$, of which $(1-\gamma)F(\theta)^{+}\nabla J(\theta)$ is the minimum-norm element, and that all of them define the same critic $Q_{w^\star}$ on the support of $d_\mu^\pi \otimes \pi$: the compatible critic is the natural gradient (Kakade, 2002; Peters and Schaal, 2008).

Exercise 14.6 (A wrong critic). [verify] In Algorithm 14.2, suppose the critic has converged to $V_w = V^\pi + e$ for some error $e : \mathcal{S} \to \mathbb{R}$, and consider the expected actor update under $(s,a) \sim d_\mu^\pi \otimes \pi$ (the discounted weights γ^{t-1} kept). (a) Show that $\mathbb{E}[\delta_t \mid S_t = s, A_t = a] = A^\pi(s,a) + \gamma(Pe)(s,a) - e(s)$ with $(Pe)(s,a) = \sum_{s'} P(s' \mid s,a)e(s')$, and deduce that the expected update is $(1-\gamma)\nabla J(\theta)$ plus the bias $\gamma\, \mathbb{E}_{s\sim d_\mu^\pi}\big[\sum_a \nabla\pi_\theta(a \mid s)\,(Pe)(s,a)\big]$: the term $-e(s)$ is a baseline and costs nothing, the term γPe is the policy gradient of a one-step problem whose reward is the error at the next state. (b) When does the bias vanish? Show that a constant error is harmless, and more generally any e such that $(Pe)(s,\cdot)$ does not depend on the action at the states that π visits. (c) In the two-state chain of Example 14.14, take $e = c\,\mathbb{1}_{\mathrm{G}}$ and compute the bias at the uniform policy as a function of c: for which c does the expected update point *away* from the optimal action at G? (d) If the critic is the TD(0) fixed point of a linear class (Theorem 13.15), bound the bias by $\frac{2\gamma}{1-\gamma}\, G\, \|e\|_\infty$ with G a bound on the scores, and compare with the error bound of that theorem: the actor inherits the critic's approximation error once per step, amplified by the horizon.

Exercise 14.7 (Details of the natural gradient). [verify] (a) Verify that the block $d_\mu^\pi(s)F_s$ of F, with $F_s = \operatorname{diag}\pi_s - \pi_s\pi_s^\top$ on $\theta(s,\cdot)$, has kernel spanned by $\mathbf{1}$ when $d_\mu^\pi(s) > 0$ (recall $\pi_s > 0$ for the softmax), that it vanishes when $d_\mu^\pi(s) = 0$, so that the natural gradient $F^{+}\nabla J$ leaves $\theta(s,\cdot)$ unchanged at such states (the convention of Theorem 14.13 applies the closed form there instead), and that adding a constant to $\theta(s,\cdot)$ does not change the policy. (b) Show that $\ln Z_k(s) \geq 0$ in (14.3) and that $d_s^\pi(s) \geq 1-\gamma$. (c) Show that as $\eta \to \infty$ the update (14.2) tends to a greedy policy for Q^{π_k}, and that with η finite but $\eta_k \to \infty$ geometrically the convergence is linear (*hint:* Xiao, 2022). (d) Reproduce Example 14.14 and try $\eta \in \{0.3, 3, 10\}$, and an increasing step $\eta_k = 2^k$.

Exercise 14.8 (Mirror-descent value iteration). [verify] (a) Derive the closed form in (14.4) and the dual-averaging identity $\pi_{k+1} \propto \exp(\frac{1}{\lambda}\sum_{j\leq k} Q_j)$. (b) Verify the identity $\langle p, Q\rangle - \lambda\,\mathrm{KL}(p \,\|\, \pi_k) = \langle p, Q + \lambda \ln \pi_k\rangle + \lambda\mathcal{H}(p)$ behind the Munchausen implementation. (c) In a one-state problem (a bandit with $\gamma = 0$), let the evaluation errors ϵ_j be independent, centered and bounded by 1; compare the two error terms of Theorem 14.15 after k iterations, in expectation, and explain in this case why the softmax of the sum of the estimates converges to the best arm. (d) Show that with exact evaluation ($\epsilon_j = 0$) and $\lambda \to 0$, (14.4) reduces to value iteration, and with $\lambda \to \infty$ the policy does not move; relate to the step size of Proposition 14.11.

Exercise 14.9 (The trust-region bound). [verify] Complete the proof of Theorem 14.16: (a) show $|\bar{A}(s)| \leq 2\|A^{\pi_k}\|_\infty\, \mathrm{TV}(\pi_k(\cdot \mid s), \pi(\cdot \mid s))$; (b) construct the coupling of the two chains (at each step, draw a common uniform variable to decide whether the actions agree) and bound the total variation between the time-t marginals by $1 - (1-\alpha)^{t-1}$; (c) sum the geometric series and apply Pinsker's inequality; (d) show that maximizing the lower bound over π yields a nondecreasing sequence of values (*hint:* the bound is tight at $\pi = \pi_k$).

Exercise 14.10 (Tilted occupation measures). [verify] (a) Prove the forward recursion for $Y^{\pi}_{\beta,t}$ and check that $\sum_s Y^{\pi}_{\beta,t}(s) Z^{\pi}_{\beta,t}(s) = Z^{\pi}_{\beta}$ for every t. (b) Show that (14.7) is the law of (S_t, A_t) under the tilted trajectory law and that, under this law, (S_t) is a Markov chain with the kernel $P^{\pi,\beta}_t$ of Remark 4.19. (c) Show that the expected return under the tilted law, $\mathbb{E}^{\pi,\beta}_{\mu}[W_H]$, is *not* the β-mean of the return but the derivative $\Lambda'(\beta)$ of its cumulant generating function (Proposition 2.15), and compare the two numerically on the hike for $\beta = -1$ under the β-optimal policy.

Exercise 14.11 (Risk-sensitive REINFORCE). [verify] (a) Show that the self-normalized estimate $\frac{1}{\beta}\sum_i w_i \sum_t \nabla \ln \pi_t(A^i_t \mid S^i_t)$ converges almost surely to $\nabla J_\beta(\theta)$ as $N \to \infty$, and that it is biased for finite N. (b) Define the effective sample size $N_{\mathrm{eff}} = 1/\sum_i w_i^2$ and compute its expectation, to first order, for N i.i.d. returns with cumulant generating function Λ: show that $N_{\mathrm{eff}}/N \approx e^{-(\Lambda(2\beta)-2\Lambda(\beta))}$. (c) On the hike, estimate N_{eff}/N for $\beta = -1$ under the uniform policy and under the β-optimal policy, by simulation ($N = 1\,000$ episodes, 100 independent batches: report the mean and the spread), compare with the first-order formula of (b), whose cumulant generating function is computed exactly by the exponential recursion of Theorem 14.20, and comment.

Exercise 14.12 (EVaR and variance gradients). [verify] (a) Assuming the supremum in Definition 6.13 is attained at a unique $\beta^\star(\theta) > 0$, show that $\nabla_\theta \,\mathrm{EVaR}_\delta(\nu^{\pi_\theta}(\mu)) = \nabla_\theta J_{-\beta^\star}(\theta)$ (envelope theorem), and write the corresponding REINFORCE estimate. (b) For the variance, show $\nabla \mathbb{V}\mathrm{ar}[W_H] = \mathbb{E}[(W_H^2 - 2\mathbb{E}[W_H] W_H) \sum_t \nabla \ln \pi_t(A_t \mid S_t)]$ and propose an estimate from a batch of trajectories. (c) For the tabular softmax, divide each gradient by the Fisher blocks as was done after Theorem 14.20. Show that both criteria give a step $\theta_t(s,a) \leftarrow \theta_t(s,a) + \eta\, G_t(s,a)$ with a weight $G_t(s,a)$ computable by forward and backward recursions: exponential moments at the tilt $-\beta^\star$ for the first, the coupled first and second moments of Proposition 4.1 together with the forward moments $\mathbb{E}[W^j_{1:t-1} \mid S_t = s]$, $j \le 2$, for the second. What does the variance step lack that the step of a β-mean has? (*Hint:* Example 5.16; can a Markov policy be optimal for a mean–variance criterion?)

Exercise 14.13 (Numerical study: the hike). [implement] Reproduce Figure 14.6. (a) Replace the start-state distribution $\mu = \delta_{\mathrm{S}}$ by the uniform distribution over the cells (the states of rows 3 and 4 are then visited at time 1): with exact gradients, $\eta = 3$ and 10^4 iterations as in the figure, does plain gradient ascent still settle on the row-2 path from S? Relate to Theorem 14.9. (b) Replace the exact gradients by REINFORCE estimates from $N \in \{10, 100, 1000\}$ episodes per iteration (self-normalized for $\beta \neq 0$, Exercise 14.11), with a budget of 3 000 iterations and 5 independent runs (Monte Carlo: report the mean and the spread over the runs), the value $U_\beta(\theta_k)$ of each iterate being computed exactly. For $\beta = 0$, report the number of iterations needed to reach 95% of the optimal value 5.09; for $\beta = -1$, report the value reached, to be compared with the exact-gradient plateau 1.06 of Example 14.21, below 95% of the optimum 2.81. (c) Run the natural gradient with exact gradients and $\eta \in \{0.3, 1, 3, 10\}$ for $K = 1\,000$ iterations, and compare the gap to the exact optimal value with the $\ln A/(\eta K)$ term of Theorem 14.13.

Exercise 14.14 (The two-product store under a parametrized ordering policy). [implement] Take the bike store with two products of Example 13.9 (256 states, $\gamma = 0.99$, discounted profit maximized; the tabular optimum $V^\star(0,0) = 271.4$ is attained by a can-order rule with reorder

point about 2 and order-up-to level 9) and search directly in a class of *smoothed can-order rules* with four parameters $\theta = (c_1, c_2, S_1, S_2)$ and a fixed temperature $\tau = \frac{1}{2}$: in state x, product i calls for a delivery with probability $\sigma((c_i - x_i)/\tau)$, independently of the other, and a delivery takes place, with probability $p_\theta(x)$, if at least one product calls; when it does, the level of product i is drawn from the discretized Gaussian $g_i(z \mid x_i) \propto e^{-(z-S_i)^2/(2\tau^2)}$ on $z \in \{x_i, \dots, 15\}$, so that a product above its level S_i is left alone. Thus $\pi_\theta(z \mid x) = p_\theta(x) g_1(z_1 \mid x_1) g_2(z_2 \mid x_2) + (1 - p_\theta(x)) \mathbb{1}\{z = x\}$, and the score $\nabla_\theta \ln \pi_\theta(z \mid x)$ is explicit. (a) With the model (regime (i)), compute $J(\theta) = V^{\pi_\theta}(0,0)$ by one linear solve and $\nabla J(\theta)$ by Theorem 14.3, with $d^{\pi_\theta}_{(0,0)}$ and Q^{π_θ} read from the model; check against central finite differences at $\theta = (6, 6, 12, 12)$. (b) Run gradient ascent with a constant step from this poor start (reorder at 6, order up to 12): report the parameters reached, J and the loss against $V^\star(0,0)$; compare with the start $(2.5, 2.5, 9, 9)$ read from the optimal policy and with the deterministic can-order rule $(2, 9)$. (c) Replace the exact gradient by the REINFORCE estimate of Algorithm 14.1 from $N \in \{10, 100\}$ episodes of length 200 from the empty store, with the weights γ^{t-1} and the batch mean of the rewards-to-go at each time as baseline: measure the standard deviation of the estimate at the poor start, then run the ascent over 5 independent runs and report the spread of the exact value of the last iterate and of the average of the last 100 iterates. Why must the step of (b) be reduced, and where do the iterates end when it is not? (d) Compare with the value-based solution of Chapter 13: four parameters against 256 values—what is gained and what is lost?

Exercise 14.15 (Influence functions and policy gradients). [verify] (a) Compute the influence function of the mean, of a utility $\int u \, \mathrm{d}\nu$, of the variance, of a β-mean and of CVaR_δ, and check that each recovers the corresponding gradient formula of this chapter. (b) Assuming that the law has a density f continuous and positive at $q = \mathrm{VaR}_\delta$, show that the influence function of the quantile VaR_δ is $(\delta - \mathbb{1}\{y \le q\})/f(q)$, and explain why the resulting gradient estimator is badly conditioned when $f(q)$ is small. What happens when $f(q) = 0$, or when f vanishes on an interval containing q? (The formula is then undefined: the problem is nonregular, not merely noisy.) (c) Show that for a β-mean the gradient of Proposition 14.22 can be rewritten as an expectation under the tilted occupation measure (14.7), and that this is what makes Theorem 14.20 a recursion; try the same with the variance and say what fails. (d) In the windy hike, at the policy reached after 100 natural-gradient iterations for $\beta = 0$ in Example 14.10 (where $\delta = 1/4$ is not a value of the distribution function of the return, so that Proposition 14.23 applies), estimate ∇J_ψ for $\psi = \mathrm{CVaR}_{1/4}$ by Proposition 14.22 with the influence function evaluated at an empirical $\hat{\nu}$ built from $n \in \{10^2, 10^3, 10^4\}$ trajectories, and study the bias as a function of n (Monte Carlo over 200 replications, with the exact gradient as reference: the law of the return jointly with the state is computed exactly by a forward recursion on the pairs (state, accumulated reward), the rewards being integers).

Bibliographical notes

The likelihood-ratio method for gradient estimation is classical in simulation (Glynn, 1990); its use for policies is due to Williams (1992), whose REINFORCE algorithm includes the baseline, and the policy gradient theorem in the form of Theorem 14.3, with compatible function approximation, to Sutton et al. (2000), with closely related results by Marbach and Tsitsiklis (2001) and Baxter and Bartlett (2001); Konda and Tsitsiklis (2003) analyzed actor–critic methods as two-time-scale stochastic approximations. Baselines and variance reduction are studied by Greensmith et al. (2004) and generalized advantage estimation is due to Schulman et al. (2016). The performance difference lemma is due to Kakade and

Langford (2002), whose conservative policy iteration is the ancestor of the trust-region methods of Schulman et al. (2015) and Schulman et al. (2017); the deterministic policy gradient is due to Silver et al. (2014). For the linear-quadratic regulator of Example 14.7, the performance is not concave in the gain, yet gradient ascent converges globally from any stabilizing feedback (Fazel et al., 2018), a result that has made the regulator the test case of the theory of policy gradient methods.

The natural gradient comes from Amari (1998) and was brought to reinforcement learning by Kakade (2002); natural actor–critic algorithms are due to Peters and Schaal (2008) and Bhatnagar et al. (2009). The global convergence theory of policy gradient methods in the tabular case is developed by Agarwal et al. (2021a), from whom Theorem 14.13 and its proof are taken; Mei et al. (2020) give the $O(1/k)$ rate of softmax gradient ascent and the linear rate under entropy regularization, Li et al. (2021) the exponential lower bound, Cen et al. (2022) the linear convergence of regularized natural gradient, Xiao (2022) and Lan (2023) the linear convergence of policy mirror descent with increasing steps, and Bhandari and Russo (2024) a general treatment of global optimality conditions. Regularization toward the previous policy, its dual-averaging closed form and the averaging of the errors are the dynamic policy programming of Azar et al. (2012); Theorem 14.15 is the sharpened form of Vieillard et al. (2020a), with the operator analysis of Kozuno et al. (2019) and the implicit implementation of Vieillard et al. (2020b). Sutton and Barto (2018, Chap. 13) present the algorithms; Agarwal et al. (2026) the theory.

Policy gradients for risk criteria go back to Tamar et al. (2012) for variance-related criteria and to Prashanth and Ghavamzadeh (2013) for actor–critic versions; the CVaR gradient is due to Tamar et al. (2015) and Chow and Ghavamzadeh (2014); the exponential-utility gradient in REINFORCE form to Borkar (2001), Nass et al. (2019) and Noorani and Baras (2021), and its Bellman form through tilted occupation measures is the gradient counterpart of the change of measure of Remark 4.19, whose roots are in Howard and Matheson (1972). Distributional critics appear in Barth-Maron et al. (2018) and in the racing agent of Wurman et al. (2022).

The influence function of Proposition 14.22 is the central object of robust statistics (Hampel, 1974); reading the policy gradient of a distributional functional through it is folklore, and makes Proposition 14.19, Theorem 14.20 and Proposition 14.23 three instances of one computation.