Corrections for

REINFORCEMENT LEARNING AND OPTIMAL CONTROL

by Dimitri P. Bertsekas

Athena Scientific

Last Updated: 11/19/2023

ERRATA

p. 29 (-2) Change "the artificial termination state t," to "the artificial termination state, denoted by (k,t),"

p. 30 (+7) Change $k = 0, \ldots, N-1$ to $k = 1, \ldots, N-1$

p. 62 (+7) Change $\min_{u \in U_k(x_k)}$ to $\min_{u \in U_k(x_k^s)}$

p. 65 (+3) Change "function approximation" to "function"

p. 66 (+24) Change "shortest problem" to "shortest path problem"

p. 67 (+1) Change "shortest problem" to "shortest path problem"

p. 76 (+3) Change the expression to

$$\sum_{k=0}^{N-1} \lambda_k \left(\sum_{i=1}^{n} c^i u_k^i - b \right),$$

p. 113 The stability argument given here should be slightly modified by adding over $k \in [1, K]$ (rather than over $k \in [0, K]$). Then in Eq. (2.40) $H_0(x_0)$ should be replaced by

$$g_0(x_0, u_0) + H_1(x_1).$$

This is the optimal cost of transfer from x_0 to $x_\ell = 0$ (i.e., the first ℓ-stage problem solved by MPC). Since this transfer is feasible by the constrained controllability condition, the above expression is finite and the stability condition is satisfied.

p. 114 (+16) Change "Moreover, it is easily verified that the base heuristic is not sequentially consistent. For example," to "Moreover, it is easily verified that the base heuristic is not sequentially consistent when $\ell > 2$. For example, when $\ell = 3$,"

p. 116 (+2) Change "If the initial state lies within the set $(-1, 1)$ the constrained controllability condition is satisfied for sufficiently large λ" to "If the initial state lies within the set $(-1, 1)$ it can be driven to 0 within a sufficiently large number of steps ℓ"

p. 118 (+16) Delete the sentence starting with "By contrast if we remove from X_k the boundary points ..."

p. 121 (+15) An added footnote: Further insight into the reasons for the effectiness of approximation in value space schemes and for the beneficial role of longer multistep lookahead can be obtained through research that appeared in the subsequent books by the author:

Rollout, Policy Iteration, and Distributed Reinforcement Learning, Athena Scientific, 2020.
Lessons from AlphaZero for Optimal, Model Predictive, and Adaptive Control, Athena Scientific, 2022.

and the author's paper "Newton's Method for Reinforcement Learning and Model Predictive Control," Results in Control and Optimization, Vol. 7, 2022, pp. 100-121.

These sources show that the cost functions $J_{k,\tilde{\pi}}$ of the one-step lookahead policy are the result of a Newton iteration starting from the cost function approximations $\tilde{J}_{k+1}$ [suitably enhanced by $(\ell-1)$ DP iterations in the case of ℓ-step lookahead]. This interpretation also applies to the infinite horizon problems discussed in Chapters 4-6, and shows that truncated rollout has a similar beneficial effect: it acts as an economical way to increase the length of lookahead.

p. 144 (+17) Delete "so that when a component gradient is reevaluated at a new point, the preceding gradient of the same component is discarded from the sum of gradients of Eq. (3.11)."

p. 159 (-9) Change

$$E(L_1, \ldots, L_{m+1}) = \tfrac{1}{2}\big(y - F(L_1, \ldots, L_{m+1}, x)\big)^2,$$

to

$$E(L_1, \ldots, L_{m+1}) = \tfrac{1}{2}\big\|y - F(L_1, \ldots, L_{m+1}, x)\big\|^2,$$

p. 167 Change the proposition statement as follows:

Proposition 3.5.1: (Least Squares Property of Conditional Probabilities) Let $\xi(x)$ be any prior distribution of x, so that the joint distribution of (c, x) is

$$\zeta(c, x) = \xi(x)p(c \,|\, x).$$

For a pair of classes (c, c'), define $z(c, c')$ by

$$z(c, c') = \begin{cases} 1 & \text{if } c = c', \\ 0 & \text{otherwise,} \end{cases}$$

and for a fixed class c and any function h of (c, x), consider

$$E\Big\{ \big(z(c, c') - h(c, x)\big)^2 \Big\},$$

the expected value with respect to the distribution $\zeta(c', x)$ of the random variable $\big(z(c, c') - h(c, x)\big)^2$. Then $p(c \,|\, x)$ minimizes this expected value over all functions $h(c, x)$, i.e., for all functions h and all classes c, we have

$$E\Big\{ \big(z(c, c') - p(c \,|\, x)\big)^2 \Big\} \leq E\Big\{ \big(z(c, c') - h(c, x)\big)^2 \Big\}. \tag{3.36}$$

p. 168 Change the page to read as follows:

The proof of the proposition may be found in textbooks that deal with Bayesian least squares estimation (see, e.g., [BeT08], Section 8.3).†

---

† For a quick proof argument, fix c, and for any scalar y, consider for a given x the conditional expected value $E\big\{(z(c, c') - y)^2 \mid x\big\}$. Here the random variable $z(c, c')$ takes the value 1 with probability $p(c \,|\, x)$ and the value 0 with probability $1 - p(c \,|\, x)$, so we have

$$E\Big\{ \big(z(c, c') - y\big)^2 \;\Big|\; x \Big\} = p(c \,|\, x)(y - 1)^2 + \big(1 - p(c \,|\, x)\big) y^2.$$

We minimize this expression with respect to y, by setting to 0 its derivative, i.e.,

$$0 = 2p(c \,|\, x)(y - 1) + 2\big(1 - p(c \,|\, x)\big) y = 2\big(- p(c \,|\, x) + y \big).$$

We thus obtain the minimizing value of y, namely $y^* = p(c \,|\, x)$, so that

$$E\Big\{ \big(z(c, c') - p(c \,|\, x)\big)^2 \mid x \Big\} \leq E\Big\{ \big(z(c, c') - y\big)^2 \mid x \Big\}, \qquad \text{for all scalars } y.$$

We set $y = h(c, x)$ in the above expression and obtain

$$E\Big\{ \big(z(c, c') - p(c \,|\, x)\big)^2 \mid x \Big\} \leq E\Big\{ \big(z(c, c') - h(c, x)\big)^2 \mid x \Big\}.$$

Since this is true for all x, we also have

$$\sum_x \xi(x) E\Big\{ \big(z(c, c') - p(c \,|\, x)\big)^2 \;\Big|\; x \Big\} \leq \sum_x \xi(x) E\Big\{ \big(z(c, c') - h(c, x)\big)^2 \;\Big|\; x \Big\},$$

showing that Eq. (3.36) holds for all functions h and all classes c.

The proposition states that $p(c \,|\, x)$ is the function of (c, x) that minimizes

$$E\Big\{\big(z(c,c') - h(c,x)\big)^2\Big\} \tag{3.37}$$

over all functions h of (c, x), *for any prior distribution of x and class c.* This suggests that we can obtain approximations to the probabilities $p(c \,|\, x)$, $c = 1, \ldots, m$, by minimizing an empirical/simulation-based approximation of the expected value (3.37).

p. 177 (+8) Change μ^k to μ_k

p. 186 (+6) Change "cost 0" to "cost $g(i, u, j)$"

p. 187-188 The conversion of the discounted problem to an equivalent SSP problem needs correction. The cost per stage of the equivalent SSP problem at state i when control u is applied should be

$$E\big\{g(i,u,j)\big\} = \sum_{j=1}^{n} p_{ij}(u)g(i,u,j)$$

(regardless of whether the next state is $j = 1, \ldots, n$ or the artificial termination state t) and not $g(i, u, j)$.

p. 191 (+3) Change "minimizing" to "maximizing"

p. 199 (Eq. (4.25)) Change $\tau_{\ell j}$ to $\tau_{\ell j}(t)$

p. 203 (+9) Change "Prop. 4.3.2" to "Prop. 4.3.3"

p. 205 (+13) Change "for all states $i = 2, \ldots, n$" to "from all states $i = 2, \ldots, n-1$"

p. 225 (+2) Change j_k to i_{k+1}

p. 226 (-15 and -12) Change $c = 0$ to $c \leq 0$.

p. 232 (+14) Change "Here ϵ" to "Here δ"

p. 232 (+15) Change "Also δ" to "Also ϵ"

p. 232 (+18) Change "cases $\delta = 0$" to "cases $\epsilon = 0$"

p. 245 (+2) (1st printing of the book) Change "(i, u)" to "(i^s, u^s)"

p. 257 (+5 and +9) The summation should be over j not i

p. 260 (Eq. (5.42)) The limit should be as $q \to \infty$

p. 263 (+15) Change $\tilde{J}_\mu(i)$ to $\tilde{J}_\mu$

p. 266 (+13) Change C to C_λ C_λ

p. 267 (line 4 of the caption of Fig. 5.5.1) Change $T^{(\lambda)}$ to $T_\mu^{(\lambda)}$

p. 267 (-10) Change ΠJ_μ to $\Pi(J_\mu)$

p. 276 (line 4 of the caption of Fig. 5.7.3) Change $\mu(r)$ to $\tilde{\mu}(r)$

p. 279 (+10) Change "Section 1.3" to "Section 3.1.3"

p. 283 (-8) Change r to r^k

p. 287 (Eqs. (5.79) and (5.80)) Change $=$ to $\in$

p. 287 (+19) Change i^s to i^q

p. 287 (+20) Change is to iq

p. 321 (+11) Change $\tilde{J}_1$ to $\tilde{J}$