

Reward Shaping for Reinforcement Learning-based Placement and Power Allocation of Flying Base Stations

Fady A. Abouelghit

*Electronics and Communications Eng. Dept.
American University in Cairo
Cairo, Egypt
fady.abouelghit@aucegypt.edu*

Karim G. Seddik

*Electronics and Communications Eng. Dept.
American University in Cairo
Cairo, Egypt
kseddik@aucegypt.edu*

Abstract—Flying base stations (FBSs) carried by unmanned aerial vehicles can be repositioned in three dimensions and switched on only while demand lasts, which makes them an attractive way to extend coverage over crowded events, emergencies, and temporary hotspots. We formulate a reinforcement learning (RL) framework for FBS deployment in a two-tier network, in which an always-on terrestrial macro base station (MBS) already serves the ground users and the aerial cells share its band. A single RL policy sets the three-dimensional position, the transmit power, and the on/off state of every FBS in one action, with no variable held fixed and none delegated to a separate stage. Performance is evaluated under a realistic channel model generated by QuaDRiGa. For a single FBS, an exhaustive grid search over the same channel supplies the reference. An unshaped RL reward is compared against three shaping forms at matched settings. All four come within 1% of the grid search reference. They differ in deployment, where the aerial tier carries between 13% and 16% of connected users depending on reward form.

Index Terms—flying base station, reinforcement learning, reward shaping, QuaDRiGa, multi-tier networks.

I. INTRODUCTION

Unmanned aerial vehicles (UAVs) used as flying base stations (FBSs) add coverage and capacity on demand, and where they are placed decides how much they contribute. The horizontal position, the altitude, and the transmit power of an aerial cell together determine how many users it serves and how much interference it creates for the users it does not [1]. Selecting these variables is the FBS placement and power control problem, and it is the problem addressed in this paper.

Optimization of these variables has been approached with both search-based and learning-based methods. Reinforcement learning (RL) has proved to be a prominent approach to the FBS placement and power control problem [2]. Value-based agents move an aerial cell over a grid to cover as many users as possible [3], while actor-critic methods handle continuous position and power and fleets of cooperating cells [4]. The appeal is speed after training where a trained policy returns placements quickly, whereas a search-based approach must be repeated for every new set of users.

Three limitations recur in the literature and together they motivate this work: the terrestrial network is usually absent, comparisons are relative rather than absolute, and every study builds its own environment.

The ground network is not normally modelled in the RL placement studies surveyed. In more than half of them the aerial cell is the only serving node, which is justified in certain scenarios, e.g., a disaster in which the terrestrial network is down. It is not the case an operator faces when an FBS is used to extend a network that is already running. With a macro tier already serving most users, an FBS adds a margin rather than the whole service, so switching it off is a feasible configuration whose objective value is already close to the best achievable. Reward design, a secondary concern when the aerial tier is the entire network, therefore decides the outcome in this setting. Rewards in the surveyed works are weighted sums of coverage and energy terms with weights that are set once and neither swept nor justified, and we found no application of potential-based shaping [5] to aerial placement.

With comparative studies, learned placements are typically evaluated against the authors' own ablations, against a value-based agent such as a deep Q-network [3], or against a random or fixed placement, and only a few works compare against a heuristic [6]. Such comparisons establish how one learner ranks against another.

Every study builds its own environment. There is no shared benchmark, so results do not carry across different papers. Because each environment is written from scratch and has to stay cheap enough to train against, propagation is usually reduced to a line-of-sight probability formula [1] or a coverage disk. Detailed models such as QuaDRiGa [7] and 3GPP TR 38.901 [8] do appear, but offline. They produce a fixed dataset, or they train a surrogate that replaces the channel during learning.

We address these three limitations. Extending beyond our earlier genetic-algorithm treatment of FBS placement and resource allocation [9], we study a single-band network in which an always-on macro base station (MBS) serves the ground users and FBSs are placed above it, and we cast the

deployment problem as a Markov decision process (MDP) solved with proximal policy optimization. The agent sets the position, power, and on/off state of every FBS in one action, with no variable held fixed and none delegated to a separate stage. Where the earlier work searched a population of candidate configurations directly, the policy here is trained once and the reward that drives it becomes the object of study. Our contributions are the following.

- We build a joint RL-based optimization framework that extends to any number of FBSs, anchored in the single-FBS case by an exhaustive grid-search reference.
- We compare three reward-shaping forms at matched settings: potential-based, record, and a combination of the two, each against an unshaped reward.
- We evaluate QuaDRiGa inside the training loop rather than offline, and we report what that costs.

The rest of this paper is organized as follows. Section II reviews the related work. Section III describes the system and channel model. Section IV states the optimization problem. Section V casts it as an MDP and defines the reward forms. Section VI gives the simulation setup and the results, and Section VII concludes the paper.

II. RELATED WORK

The FBS placement and power control problem is defined by the variables the agent is allowed to set: the horizontal position, the altitude, the transmit power, and the on/off state of each aerial cell. The RL formulation reviewed formulations admit only a subset. Altitude is held constant in [3] and [10], and reduced to a choice between two levels in [11], where transmit power is not a decision variable in [3], [11] or [10]. Where more variables are admitted they are set in separate stages: the user association is fixed by a K-means step before the agent runs [4], the transmit power is set by a pricing game formulated separately from the MDP defined in [12], or a coarse placement is refined by a second pass [3]. Single-agent formulations that emit position and power in one continuous action are rare [13]. The terrestrial tier is usually absent as well, or kept out of the signal-to-interference-plus-noise ratio (SINR) by declaring the system noise-limited or by assigning the tiers orthogonal bands. Where a terrestrial network does remain active, the decision variables are association, backhaul or radio access technology rather than the aerial placement itself [11]. We instead keep the macro tier transmitting on the same band and set the position, power and on/off state of every FBS in one action.

The reward formulations in the works surveyed are weighted sums of service and cost terms, following the pattern set by early energy-aware UAV control [14] and reported as the prevailing design across the RL-for-UAV literature [2]. The weights are chosen once, with no reported sweep or justification for the values used, and the resulting placements are typically compared against another learner or a heuristic [6] rather than against a reference computed over the environment itself. Potential-based shaping is standard elsewhere in reinforcement learning and leaves the optimal policy unchanged

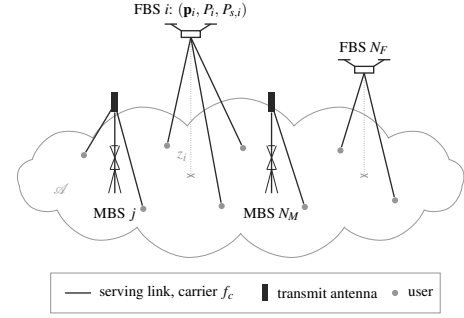


Fig. 1. Single-band two-tier system model. The macro tier has N_M sites and the aerial tier N_F FBSs. Solid lines are serving links and gray dots are users in the service area $\mathcal{A}$.

[5], but to the best of our knowledge it has not been applied to aerial placement.

Moreover, the channel models used differ widely in detail, and the more detailed ones are kept outside the training loop. Surveyed environments use a line-of-sight probability model [1] or a coverage disk [10], while ray tracing and site-specific data are computed offline to build a dataset or to train a surrogate that the agent then learns against [6]. The former do not represent spatially correlated shadowing, small-scale fading, or the line-of-sight state of an individual link, and the latter represent them only for a precomputed set of positions. Geometry-based stochastic models such as QuaDRiGa [7], [8] address these shortcomings but, to the best of our knowledge, have not been evaluated inside a training loop.

III. SYSTEM AND CHANNEL MODEL

A. Network Topology

We consider the downlink of a single-band, two-tier deployment over a two-dimensional service region $\mathcal{A} \subset \mathbb{R}^2$, shown in Fig. 1. Every transmitter uses the same carrier frequency f_c , so any active transmitter interferes with every user it does not serve.

The macro tier has N_M sites at fixed, known coordinates

$$\mathbf{m}_j = (x_{m_j}, y_{m_j}, z_{m_j}), \quad j = 1, \dots, N_M, \quad (1)$$

each transmitting continuously at a fixed power. The macro tier is a given: its positions and powers are not optimized and it is never switched off. The aerial tier has N_F FBSs at controllable three-dimensional positions

$$\mathbf{p}_i = (x_i, y_i, z_i) \in \mathcal{A} \times [z_{\min}, z_{\max}], \quad i = 1, \dots, N_F, \quad (2)$$

each with a continuous transmit power $P_i \in [P_{\min}, P_{\max}]$ and a binary on/off state $P_{s,i} \in \{0, 1\}$. The tuple $(\mathbf{p}_i, P_i, P_{s,i})$ collects the quantities to be optimized for FBS i so that the network meets a target quality of service; the resulting problem is stated in Section IV.

We collect the transmitters in a set $\mathcal{S}$ of $N_M + N_F$ elements and let $a_s \in \{0, 1\}$ mark whether transmitter s is radiating, so $a_s = 1$ for every macro site and $a_s = P_{s,i}$ for FBS i .

B. Users and Received Power

The region holds K users, with user k at $\mathbf{u}_k = (x_k, y_k, z_u)$, where $(x_k, y_k) \in \mathcal{A}$ and z_u is a common receiver height. For user k and transmitter s , a channel model returns the received power $P_{k,s}^{\text{rx}}$, which accounts for distance-dependent path loss, shadowing, small-scale fading, and the line-of-sight or non-line-of-sight state of the link. Nothing that follows constrains which model is used; the one adopted here is given in Section III-E.

C. Users Connectivity Model

Since one band is shared, the interference seen by user k on transmitter s sums over every other active transmitter,

$$\gamma_{k,s} = \frac{a_s P_{k,s}^{\text{rx}}}{\sum_{s' \in \mathcal{S} \setminus \{s\}} a_{s'} P_{k,s'}^{\text{rx}} + N_0}, \quad (3)$$

where N_0 is the receiver noise power and $a_s \in \{0, 1\}$, defined above, indicates whether transmitter s is radiating, so a switched-off FBS contributes neither signal nor interference.

Each user associates with the transmitter of largest SINR,

$$\gamma_k = \max_{s \in \mathcal{S}} \gamma_{k,s}, \quad (4)$$

and is connected when $\gamma_k \geq \gamma_{\text{th}}$ for a fixed threshold γ_{th} .

D. QoS Metrics

A configuration is scored by the number of users connected to either tier,

$$U_{\text{conn}} = \sum_{k=1}^K \mathbf{1}\{\gamma_k \geq \gamma_{\text{th}}\}, \quad (5)$$

and by the transmit power drawn by the aerial tier,

$$P_{\text{FBS}} = \sum_{i=1}^{N_F} P_{s,i} P_i. \quad (6)$$

We also record how the connected users divide between the two tiers. With a macro tier already serving U_{conn} on its own does not separate a deployment that leaves the aerial cell idle from one in which it carries traffic, so the FBS-served count is reported alongside it.

E. Channel Model

Channels are generated with QuaDRiGa [7] in the 3GPP TR 38.901 urban macro (UMa) scenario [8], a geometry-based stochastic model whose large-scale parameters are drawn from spatially correlated maps. A moving transmitter sees a time-correlated sequence of channels rather than an independent draw at each position. The channel is therefore regenerated whenever a transmitter moves.

IV. PROBLEM FORMULATION

We collect the aerial-tier decisions of Section III in the vector

$$\boldsymbol{\theta} = [\mathbf{p}_1^\top, P_1, P_{s,1}, \dots, \mathbf{p}_{N_F}^\top, P_{N_F}, P_{s,N_F}]^\top \in \Theta, \quad (7)$$

of length $5N_F$, where the feasible set Θ is the product over i of $\mathcal{A} \times [z_{\min}, z_{\max}]$ for the position, $[P_{\min}, P_{\max}]$ for the transmit power, and $\{0, 1\}$ for the on/off state. We optimize a snapshot of the network. The macro tier and the user positions are held fixed, and $\boldsymbol{\theta}$ collects every quantity the optimizer may change.

Let U_{FBS} be the number of connected users served by the aerial tier and U_{max} the number of users the network can connect at all. A configuration is summarized by three normalized quantities,

$$u(\boldsymbol{\theta}) = \frac{U_{\text{conn}}}{U_{\text{max}}}, \quad \rho(\boldsymbol{\theta}) = \frac{P_{\text{FBS}}}{N_F P_{\text{max}}}, \quad \varphi(\boldsymbol{\theta}) = \frac{U_{\text{FBS}}}{U_{\text{conn}}}, \quad (8)$$

each of which lies in $[0, 1]$: the share of users connected, the share of the aerial power budget drawn, and the share of connected users carried by the aerial tier. The objective combines a network term with an offload term,

$$\Phi(\boldsymbol{\theta}) = (1 - w_F) u(\boldsymbol{\theta})^\beta (1 - \rho(\boldsymbol{\theta}))^\gamma + w_F \varphi(\boldsymbol{\theta}), \quad (9)$$

where $\beta \geq 0$ and $\gamma \geq 0$ set the relative weight of connectivity and aerial power within the network term, and $w_F \in [0, 1]$ sets the relative weight of the offload term. These are design parameters through which the operator trades the objectives against one another. The values used in this paper are one such choice. The deployment problem is then

$$\max_{\boldsymbol{\theta} \in \Theta} \Phi(\boldsymbol{\theta}), \quad (10)$$

in which U_{conn} , U_{FBS} and P_{FBS} follow from $\boldsymbol{\theta}$ only through the channel, the SINR of eq. (3), and the association and threshold rules of Section III.

Eq. (10) has three properties that shape everything that follows. First, it is a mixed-integer nonlinear problem, mixing continuous variables with one binary variable per FBS. Second, its objective has no closed form, because the received powers come from a stochastic channel model rather than a formula, so no gradient is available and neither concavity nor continuity can be assumed. The association maximum in particular is discontinuous, since a small move can flip a user from one transmitter to another. Third, the configuration in which the aerial tier is switched off, $P_{s,i} = 0$ for all i , is feasible and attains

$$\Phi_{\text{off}} = (1 - w_F) u_0^\beta, \quad (11)$$

where u_0 is the connectivity of the macro tier alone, so every deployed configuration is measured against Φ_{off} rather than against zero, and only the margin above it is available to the search.

For a single FBS, eq. (10) can be solved by enumeration. The costly quantity is the QuaDRiGa power map, which depends only on the FBS position, so one map per grid

position covers the search. We write Φ_{grid} for the largest objective value the grid contains. Enumeration does not extend to a larger aerial tier: FBSs interfere with one another, so a joint solution has to be scored over combinations of per-FBS grid points and the count grows as M^{N_F} for M points per FBS. Eq. (10) is therefore solved by learning a policy instead.

V. REINFORCEMENT LEARNING FRAMEWORK

A. Algorithm and Policy

Eq. (10) provides no gradient and has a mixed-integer feasible set, so the search can rely only on sampled evaluations of Φ . We use proximal policy optimization (PPO) [15], an on-policy actor-critic method that emits continuous vectors and constrains each update to stay close to the current policy. The constrained update matters because every environment step costs one channel regeneration, so each update is formed from few samples.

The trained policy is not deployed as an online controller: because the problem is a snapshot, a run produces a configuration rather than a trajectory, and the deployment we report is the best configuration the policy visits.

B. State and Action

The state s_t concatenates three groups: the current configuration θ_t , min-max scaled to $[-1, 1]$, the changes in total and in FBS-served connectivity produced by the previous step, and the share of connected users carried by each tier. Each additional FBS adds five entries to the state and to the action, so the framework extends to larger aerial tiers.

The action $\mathbf{c}_t \in [-1, 1]^{5N_F}$ carries one entry per decision variable. The interval $[-1, 1]$ is the bounded output range of the policy network, and each entry is read as a relative move rather than as a new value. Let v be one of the four continuous variables of an FBS, that is, x_i, y_i, z_i or P_i , let c be the entry of $\mathbf{c}_t$ that addresses it, and let $[v_{\text{lo}}, v_{\text{hi}}]$ be the bounds that Θ imposes on v . Continuous variables are updated by

$$v \leftarrow \text{clip}(v + \lambda c(v_{\text{hi}} - v_{\text{lo}}), v_{\text{lo}}, v_{\text{hi}}), \quad (12)$$

where λ is a step size common to all variables, so one action displaces each variable by at most λ times its range. The on/off variable is set to $\mathbf{1}\{c \geq 0\}$ when $|c| \geq \frac{1}{2}$ and left unchanged otherwise, so small actions do not toggle the aerial cell. Algorithm 1 states the resulting loop.

C. Reward Shaping

The objective Φ scores configurations, not transitions, so a per-step reward has to be built from it. The obvious choice, paying Φ at every step, gives the critic a target that barely varies within an episode, since Φ occupies a narrow range above Φ_{off} ; the reward then carries little information about which direction improves the configuration. Shaping addresses this by paying for change rather than for level, and the three forms below differ in which kind of change they pay for. The unshaped choice pays the objective at every step,

$$r_t = \Phi(\theta_t), \quad (13)$$

Algorithm 1 Shaped policy search for FBS deployment

Require: shaping form $\in \{\text{unshaped, potential, record, combined}\}$, horizon T , step size λ , weight w

- 1: initialize policy parameters ψ
- 2: **for** each episode **do**
- 3: draw $\theta_0 \sim \text{Uniform}(\Theta)$ and evaluate Φ_0
- 4: $\Phi_{\text{prev}} \leftarrow \Phi_0, \Phi^* \leftarrow \Phi_0$
- 5: **for** $t = 1$ **to** T **do**
- 6: sample $\mathbf{c}_t \sim \pi_\psi(\cdot | s_{t-1})$
- 7: form θ_t from θ_{t-1} and $\mathbf{c}_t$ by eq. (12) and the sticky rule
- 8: regenerate the channel and evaluate $\Phi_t = \Phi(\theta_t) \triangleright$ dominant cost
- 9: $r_t \leftarrow$ shaping form applied to $(\Phi_t, \Phi_{\text{prev}}, \Phi^*)$, cf. eqs. (13)-(16)
- 10: $\Phi_{\text{prev}} \leftarrow \Phi_t, \Phi^* \leftarrow \max(\Phi^*, \Phi_t)$
- 11: store $(s_{t-1}, \mathbf{c}_t, r_t, s_t)$
- 12: **end for**
- 13: update ψ by PPO on the stored batch
- 14: **end for**
- 15: **return** the configuration attaining the largest Φ visited

so an episode that starts in a good configuration and never moves earns as much as one that searches its way there. We compare it against three shaped alternatives.

The first pays the step-to-step change in the objective,

$$r_t = \Phi(\theta_t) - \Phi(\theta_{t-1}), \quad (14)$$

so the return over an episode telescopes to $\Phi(\theta_T) - \Phi(\theta_0)$: the agent is paid for improving on the configuration it started from, and charged the same amount for giving that improvement back. Because the reward is the difference of Φ at consecutive configurations, this is potential-based shaping with Φ as the potential, and [5] shows that such a reward leaves the optimal policies of eq. (10) unchanged when the return is undiscounted.

The second pays for discovery. With $\Phi_t^* = \max_{\tau \leq t} \Phi(\theta_\tau)$ the best value seen so far in the episode,

$$r_t = \max(0, \Phi(\theta_t) - \Phi_{t-1}^*), \quad (15)$$

and the return telescopes to $\Phi_T^* - \Phi(\theta_0)$. The agent is paid when it improves on the best configuration found and is never charged for moving away from it. This form is deliberately not policy-invariant: it replaces ending in a good configuration with having found one, which matches deployment planning, where the configuration that gets flown is the best one found.

The third is their linear combination,

$$r_t = (\Phi(\theta_t) - \Phi(\theta_{t-1})) + w \max(0, \Phi(\theta_t) - \Phi_{t-1}^*), \quad (16)$$

with return $(\Phi(\theta_T) - \Phi(\theta_0)) + w(\Phi_T^* - \Phi(\theta_0))$, so $w \geq 0$ sets the weight of discovery relative to the potential term.

The shaped rewards of eqs. (14)-(16) are differences of Φ , and since Φ occupies a narrow range these differences are small in absolute terms, so the gradient scale would otherwise

TABLE I
SIMULATION PARAMETERS.

Parameter	Value
<i>Network and channel</i>	
Service area	2000 × 1500 m
Macro N_M , position, height, power	1 at (1100, 966) m, 25 m, 20 dBm
FBS N_F , altitude, power	1, [20, 150] m, [7, 10.5] dBm
Users K , receiver height	1000, 1.5 m
Carrier, channel	2 GHz, QuaDRiGa 38.901 UMa
Noise N_0 , threshold γ_{th}	10^{-11} mW, 5 dB
<i>Objective</i>	
(β, γ, w_F)	(1.0, 0.0, 0.4)
<i>Exhaustive reference grid</i>	
Step, altitude and power levels	20 m, 8, 8
Maps, configurations scored	61 408, 491 264
<i>PPO</i>	
Horizon T , environment steps	40, 25 600
Rollout, batch, epochs	1024, 64, 10
Learning rate, discount, GAE	3×10^{-4} , 0.95, 0.95
Clip, entropy coef., step size λ	0.2, 0.01, 0.05

depend on the shaping form. We therefore divide the reward stream handed to the optimizer by a running estimate of the standard deviation of the discounted return. The normalization sits outside the environment, so it acts on the shaped reward and never on Φ , and is enabled identically for all four reward forms.

VI. SIMULATION SETUP AND RESULTS

A. Simulation Setup

The scenario is a single macro site at the centre of a 2000×1500 m service area with one FBS above it, i.e., Section III with $N_M = N_F = 1$. Users are placed uniformly in the area, and the map is drawn once and held fixed across every run, so the runs differ in the RL agent and not in the simulated scenario. Channels are regenerated at every environment step. Table I lists the relevant parameters, and the codebase is found at [16].

The exhaustive reference of Section IV is computed on the grid in Table I. Because the channel map depends only on the FBS position, the sweep evaluates 61 408 maps and recovers the transmit-power and on/off dimensions from them by the linear rescaling of Section IV, which brings 491 264 feasible scored configurations. The largest objective value on this grid is Φ_{grid} , the reference against which every learned result is reported. We write U_{grid} for the number of users connected in that configuration and U_0 for the number connected with the aerial tier switched off, so that $u_0 = U_0/U_{max}$.

Four reward forms are compared, given respectively by eqs. (13)-(16): the unshaped reward, the potential form, the record form, and the combination at $w = 1.0$. The four runs differ in the shaping form alone; every other setting in Table I, the random seed and the user map included, is held fixed, so the comparison isolates the reward from everything else.

B. Results

Each run is summarized in two ways. The best configuration visited during training is the deployment the method

TABLE II
RESULTS BY REWARD FORM. P_1 IS THE FBS TRANSMIT POWER, IN DBM.

Reward form	Best visited		Converged policy			
	Φ (% of Φ_{grid})	P_1	U	F	Φ (% of Φ_{grid})	
unshaped	0.489 (99.2%)	9.74	715	96	0.483 (98.0%)	
potential	0.492 (99.9%)	9.82	718	104	0.488 (99.2%)	
record	0.494 (100.2%)	10.21	680	107	0.471 (95.6%)	
combined	0.492 (99.9%)	10.31	700	111	0.483 (98.1%)	

returns, since what is flown is the best configuration found. The configuration produced by the final policy, evaluated deterministically after training, is what the policy has learned to do. Table II reports both.

Training proceeds alike in all four runs. Fig. 2(a) tracks the total number of connected users and Fig. 2(b) shows the critic value losses. Every form starts near 655, passes the do-nothing level $U_0 = 694$ within some 4000 steps of the others, and settles between 685 and 704 against $U_{grid} = 722$. No run stalls, and no form is distinguishable from the others on this axis. Keeping the aerial cell switched on is a learned behaviour. Between a quarter and a third of training episodes in each run end with the cell inactive, a behaviour the unshaped, potential and combined forms have all but shed by the final quarter of training (1.9, 2.5 and 6.2% of episodes) while the record form retains it in 24.4% of training steps. Total connectivity is thus saturated once the cell is deployed at all, and cannot separate the four forms.

Measured by the best configuration visited, the forms are equivalent. Every one reaches at least 99% of Φ_{grid} at 84%-96% of the total allowable power. The record form reaches 100.2% of Φ_{grid} because the reference is quantized to the 20 m grid of Table I while the policy moves continuously and can settle between grid points. Whatever the reward, a near-optimum region is found. Reward design is therefore not what decides whether the search succeeds.

The converged objective spans nearly four percentage points of the reference, from 99.2% under the potential form to 95.6% under the record form. The forms differ less in objective value than in which users they serve. Across Table II, the FBS-served count rises from 96 users to 111 while the total falls from 718 to 680. The forms trade macro-served users for aerial-served users rather than adding users in aggregate. The deployments differ geographically as well (Fig. 3): three forms converge within a few tens of metres of the grid reference, while the record form settles some 245 m away, serving fewer users in total but assigning more of them to the aerial tier. Reward shaping alone therefore separates two deployments of practically equal objective value.

Fig. 2(b) shows where the divergence originates. The critic value loss separates the four runs by two orders of magnitude, lowest under the unshaped reward and highest under the record form. A low value loss is not a better critic here: paying the agent Φ itself gives it a target that barely moves within an episode, so the critic fits that target closely while learning little about which direction improves the configuration. The shaped

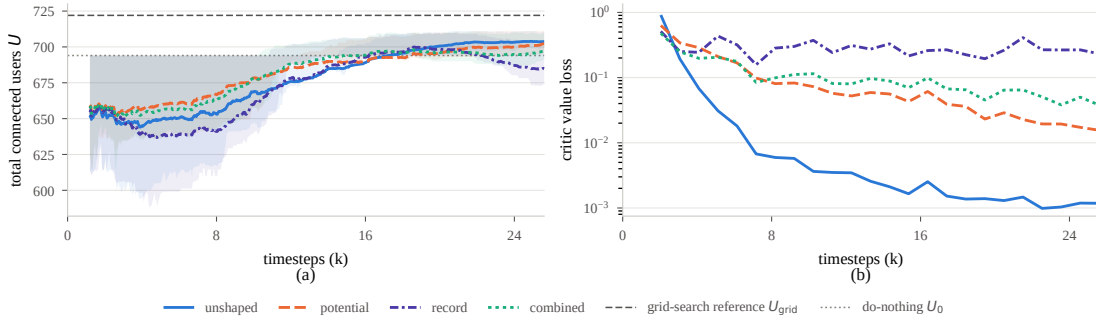


Fig. 2. Training behaviour of the four reward forms. (a) Total connected users against environment steps, with the grid-search reference U_{grid} and the do-nothing level U_0 marked. (b) Critic value loss on a logarithmic axis.

forms present a target that moves with the search, which is harder to fit and more informative.

VII. CONCLUSION

We studied FBS deployment in a single-band two-tier network in which a macro tier is already serving. A reinforcement learning policy sets the position, transmit power and on/off state of every FBS in one action, with the channel evaluated at every environment step. For a single FBS, an exhaustive grid search over the same environment supplies the reference against which the learned results are measured.

Against it we compared an unshaped reward with three shaped forms at identical settings. All four behave alike in training and all reach near-optimum regions at their best visited configuration. The reward changes which deployment the policy settles on, not how good the best one found is. The forms trade macro-served users for aerial-served users and place the aerial cell in visibly different parts of the area.

Several directions follow. The exhaustive reference disappears beyond one MBS and one FBS. Because each environment step costs one channel evaluation, an off-policy learner that reuses stored transitions would extract more from the same evaluation budget. Mobile users would turn deployment into a

tracking problem, and rate requirements could enter as explicit constraints rather than penalty terms.

REFERENCES

- [1] A. Al-Hourani, S. Kandeepan, and S. Lardner, "Optimal LAP altitude for maximum coverage," *IEEE Wireless Commun. Lett.*, vol. 3, no. 6, pp. 569–572, 2014.
- [2] N. Parvaresh, M. Kulhandjian, H. Kulhandjian, C. D'Amours, and B. Kantarci, "A tutorial on AI-powered 3D deployment of drone base stations: State of the art, applications and challenges," *Veh. Commun.*, vol. 36, p. 100474, 2022.
- [3] J. Qiu, J. Lyu, and L. Fu, "Placement optimization of aerial base stations with deep reinforcement learning," in *Proc. IEEE Int. Conf. Commun. (ICC)*, 2020, pp. 1–6.
- [4] D. Rizvi and D. Boyle, "Multi-agent reinforcement learning with action masking for UAV-enabled mobile communications," *IEEE Trans. Mach. Learn. Commun. Netw.*, vol. 3, pp. 117–132, 2025.
- [5] A. Y. Ng, D. Harada, and S. Russell, "Policy invariance under reward transformations: Theory and application to reward shaping," in *Proc. Int. Conf. Mach. Learn. (ICML)*, 1999, pp. 278–287.
- [6] J. Lyu, X. Chen, J. Zhang, and L. Fu, "Spatial deep learning for site-specific movement optimization of aerial base stations," *IEEE Trans. Wireless Commun.*, vol. 23, no. 7, pp. 7712–7727, 2024.
- [7] S. Jaeckel, L. Raschkowski, K. Börner, and L. Thiele, "QuaDRiGa: A 3-D multi-cell channel model with time evolution for enabling virtual field trials," *IEEE Trans. Antennas Propag.*, vol. 62, no. 6, pp. 3242–3256, 2014.
- [8] 3GPP, "Study on channel model for frequencies from 0.5 to 100 GHz," 3rd Generation Partnership Project (3GPP), Tech. Rep. TR 38.901, 2024.
- [9] F. A. Abouelghit and K. G. Seddik, "Genetic algorithm-based placement and resource allocation for flying base stations," in *Proc. IEEE Middle East Conf. Commun. Netw. (MeCom)*, Cairo, Egypt, Nov. 2025.
- [10] M. R. Khan, G. R. V. Premkumar, and B. Van Scoy, "Robust UAV-oriented wireless communications via multi-agent deep reinforcement learning to optimize user coverage," *Drones*, vol. 9, no. 5, p. 321, 2025.
- [11] H. Zhang, Z. Qi, J. Li, A. Aronsson, J. Bosch, and H. Holmström Olsson, "5G network on wings: A deep reinforcement learning approach to the UAV-based integrated access and backhaul," *IEEE Trans. Mach. Learn. Commun. Netw.*, vol. 2, pp. 1109–1126, 2024.
- [12] S. Fu, X. Feng, A. Sultana, and L. Zhao, "Joint power allocation and 3D deployment for UAV-BSS: A game theory based deep reinforcement learning approach," *IEEE Trans. Wireless Commun.*, vol. 23, no. 1, pp. 736–748, 2024.
- [13] M. Zhang, S. Fu, and Q. Fan, "Joint 3D deployment and power allocation for UAV-BSS: A deep reinforcement learning approach," *IEEE Wireless Commun. Lett.*, vol. 10, no. 10, pp. 2309–2312, 2021.
- [14] C. H. Liu, Z. Chen, J. Tang, J. Xu, and C. Piao, "Energy-efficient UAV control for effective and fair communication coverage: A deep reinforcement learning approach," *IEEE J. Sel. Areas Commun.*, vol. 36, no. 9, pp. 2059–2070, 2018.
- [15] J. Schulman, F. Wolski, P. Dhariwal, A. Radford, and O. Klimov, "Proximal policy optimization algorithms," arXiv:1707.06347, 2017.
- [16] F. A. Abouelghit, "rl-fbs-optimization," 2026, [Online]. Available: <https://github.com/fadyabouelghit/rl-fbs-optimization>.

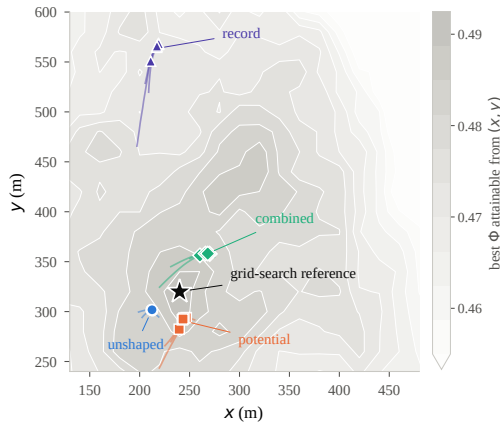


Fig. 3. Converged FBS positions over the objective landscape. Trails show the final approach of each run and the star is the grid-search reference.