Deep Reinforcement Learning Empowered Rate
Selection of XP-HARQ

DaDWWuJihlahitiFeng, ZibengShi, Hongjiang Lei, Guanghua Yang, and Shaodah Mafa

Abstract - The complex transmission mechanism of crosspacket hybrid automatic repeat requests (XPAHARQ) binders its optimal raust emstdesignes Ton. overcome othise difficulty fiethlisy, lether attempts to use the deeph reinforcement learning (DRL) to DBIke the state selection problemonf XP-HARQL over-lear related efading channelsaltimarticular, the long term average throughput (V/FAT) ish maximpized (by Aproperlynchoosized the incremental information rate-for each HARQ round on the basis of the outdated chamtel states information ((CSI) available at the transmitter. (The rate selection to robblem is sfirst converted into se Marko v decision process (MDP); ewhich is Machosolveds by papitalizing Don), the ialgorithm of I deep by deterministing policy hogradienth (DDPG) cwith torioritisted experience dieptay. DThtGsimulation icresults! finally iccorrobodate three superiority of stiles proposed o XPEHARQ tischeme rovery the conventional HARQAwith sincremental tredundancy (HARQ-IR) ariththe:XP:HARQewithdomlyys(4tisfic@HESIand the XP-HARQ with only statistical CSI.

Index Terms—Cross-packet hybrid automatic repeat request

Thdex Terms—Cross-packet hybrid automatic repeat request (XP:HARQ): deep heinforcement dearning (DRL); outdated chantel hith) rulation, rate selection. learning (DRL), outdated channel state information, rate selection.

I. INTRODUCTION

Hybrid automatic repeat request (HARQ) is one of the keyl veenhougheratist repeatpasterest offering rename afailse knystonim plowevethatnis wareald so Esstariany reliable taranse prissions tallow transmission bearful, is brene utially averaged for thm mige of lange temporission delatared ich is umfaverable ferriul@jirro thesorteraceliablineandalomeletra avurgenmueed cotions (of Pal-4cX) bit thereby granishis in dilementant sherehia cours betreenth surabiselo meetlavible URES teguisuiseins mechanismethate coulds be promotigate able to mark diverge HRROS marris emercial directhis sletters on of or any thrus with packetal HARRICAXP-AlbAR On that is not exploration acompressing PTHARQ: with high apectral enflorance Artor scale files price information bits are introduced in retransmissions such that surplus introless are surses care substrutially exploited: Hence, and 62261160650, in part by Chongqing Key Laboratory of Mobile CoThiss weide two softpedited log sparts by: National Naptrah Sci2022 (Foundation by Chinangelderg Grants: 62H71200,li621E420d, F64974080 Fand 6226h160650; Grpatt B926Adrigtfing 0809, Liabopatoty by Mobile i Chasimaication Ap Feirl Boloigy Russderr (Grifint unduption ct-2022/04 rain t pZid 289 1G008gd 00g (BakViC and Applied Basic NResearth Foundation ander Grant 2023Ad 5550 109001 implant Gyr sZhubbil 9BaNé 1 Srid) 3 Appliedir Busict Research Science attend under Grant ZH2201j7003210050P,WCaimpaStAjRthenMajoGTaletst P09g7/i2082 GAIng dong BioViritinT Soi(Ei NG): an 22929QNO(IS de3; s poulding partitly orth Z Isoign Shia) d Technology, Development Jund. Macag S&Ri, under Gratits a 1087/2022/AFJ and SKItt-IOTSC(UM):2021-2023. (Gibelliguntdingsystthms Zhinigushi.) and ErDaneWringJiahdinaFieng, UiZhengit Shi, Zindha Guañ g 9070, Yang hine with the iSchool (bf84meltige621Systems) Science jinds Högine26029, jdinad dJmixbreitys h Zhjubai echii 9070g h Ching@j(u+mails.cn0x8a@ stu2021.jnu.edu.cn; jiahiH@ktu2020:jritled@bnnzhengsli@jnuLedu.crf; }bbybilg@@nnedu.cri):ations Tellhikelojsyváth Changqing Keyi kabibý Mobiles Conndufficiations (fechnolbigyrs& Chongqing Uni0065ity OhiPasts-ands Telledhjifmquications, Gilongqing 408065;dChinhf(e-inathdeBtj@eqbptyeduadn).ratory of Internet of Things forSlStodart Matrix the State KewfLaboratory of Anternethol Things of br Smart CitynUniversitylufiMa)cau, Macau, China (e-mail: shaodanma@um.edu.mo).

itf ishiginecessaryxitywail, for the end of the tetransmissions of the Burent message before the delivery to be the intext decided age especially saides benign channel gonditions. As a consequence; shdospectiallyefficiencyedf HARQ is boostedceseanwhilevtlic average transmission delays is is ethered the current message beRecently, districtly attenues to the sXPg-HARQ is the mendee stelli ign tilein flah cym Sletierak, effortschave cheencepalle speatoral ráficije revaluateFlahR Optimálity sdesigm XPnHARQ tšehemes a In t?anMohammedellabisereduceramined the long term average thrRughptity(IfTAT)nofesXPeHiARQowithewXiRhHAcRDroughput impreviement (gained aby XPSHARQ) will over the de I to 62 juant whe tayer: codatel scheme at a developed toll implement XIP-HARQ solguaranten tije inhutskofnthedenkoder ewith the same eldnigth; where epuncturing a nilimixing operations were proveriged wither punicturing three swere then poptimized with dynamic Brighting mang vier [13].dThe adaptive two dutation can't god in gracheme was furtherdintroducted temboost-HhARCTAT gofaXR-H-ARQ inpfRs In [1]; the coffective that he it you feel was a palyzed for huffer-limited XPrHARQ. However-the-performance energies of tXP-HARQ lim(?)pp[2]u[2]d[?]twlerel obtained byogonducting Monte-Carlo asimulations and dacked in heightful analysis. wife fill this ivacambycethcomostsfundamental geNDrHahReQmetfic, hamely, tout age probability a citas de rixed in AcReedy for ranfoly XPd HARQfiver independent Rangeighfading channels indehavith www.chcfullftXne-llfsArsityi.of[XP-HARQ [wassprovedbtHiowelvbs; evenlunders udbatsi füpld «bännelatiodel athe butage änalysis üs too/costspleto to fluither assist the optimal design dofn XP(HARQ), from the mention rinder melecomplicated balding channels rived in Totaldriss the aboxPistueRWe resert to the data driven leep feilifogodmentelearnfilg (DRLs) forhtheloptimali design of XP-HARQ over correlated Farting schemes suits should be sotticed that nordy racfew, worksouttempted abysdevise the conventional HARQ selsemethousing in the Design of holds. Harring in the htensidaRLnesshodressemsoliedubingsphilicycleasadesigned to millimizehthes age abonformatione(Aed): tfoo HARQt systems. the following th wfas/IPvFrAgeQ towna ximize the lthrolighpuh vianoptirhizing the Incremental redundancy bifs: "Unforkmately ritheed x tensilow is 6 the DRIvenicithods HoAgeQesah HAAQ singethes Das Lnovethbebn Pepoitedlaifhisidetter an@Rhizes.thel LiFAT swiechdaptiveoliate selectionic by eccouside in in prior that each of the control of t (GSI) ATBQ optiteization problem desprintity erformatite de als cu problem: of DMarkov alegisitim processle(MfaR); dBtyo taking ninte abeculatotheligoritimizous (static land) that lone is paces, atherdrobilem cis thien solvedrbyrusing/DDPGewith pirioritizeth experience et plays By ground act In A IMO at the Carlo has mutation be the reproposed TXPs HARQuschemiz is thro Vell All be a superior verthe seleventional HARiQevith intrestental redundancy (HARQ-IR) and the XPe

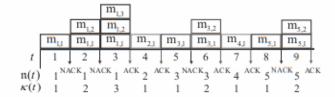


Fig. 11. After example left the NP-MIARQAR Generowith K ith 3K = 3.

btARQzwithroplyolstatisticalicSt. Furtheraneleastaispfolulehthaf the time constation among fading channels does not lead to a significant impact at point he dt TAT spathe, proposed IXP-HARQ soberdeby using DDPG with prioritized experience replay. ByThendesetinfgthis detterais octilined in a sfollows: o SestionX ?? HANDQcesh thee system enhodelse Sections? 20 tlevelops eatiDRAL Ehtp@@eveit]ratecselection| algorithm_for(NR-B/ARQ)Thedsime MRtdd AdSQltsváth presented tintiSed tiGlS1??FSedtiom3?efintally concludes this hetterne correlation among fading channels does not lead to a significant impact upon the LTAT of the proposed XP-HARSY STEMMODEL

This verter fethisidetter journalisation of the manifestion vs. tetrochiconithe xpetrarocile acoptiono? Endovelore reiland missions of that message or algorithm section defined as The systelatabate, the lading the XPdHARQ tiansmission three has finallytherelarder this deft berformance metrics, and the rate selection problem.

II. System Model

A. Typis Harrer considers a point-to-point communication system, in which XP-HARQ is adopted to enable the As shown in Fig. ??, an example is used to illustrate the retransmissions of the amessage TAR of art, this section transmission mechanism of the XP-HAR of a to avoid network delineates the system model including the XP-HARO congestion in unlavorable propagation environment, the number transmission mechanism the channel model, performance ber of transmissions of XP-HARO is limited up to K. For notational simplicity, let $\mathbf{n}(t) \in \mathbb{Z}^n$ and $\kappa(t) \in [1, K]$ be the functions that map the time slot t to the current HARQ cycle and the current transmission round, respectively. In the initial transmission round of the a(t)-thrillARQ cycle, theilmessage the trais renesided rased and award the NP with R Gransmission ratewBik Thengeseived isignal annual leads pagation environment, the number of transmissions of XP-HARQ is limited up to K. $Y \otimes U$ notational interest the V tet V V V V and where $h[t_0,K]$ denotes the entirest characteristic of the instability by the normal HARO could would the courrest transmission sounds, for pheticomolex tadditive Gaussian inoise TAWGNI thaving t) zero Hinear Ω and Ω was ances one σ^{Ω}_{n} and is P_{i} and Ω average draws and bower in the antian HARO with Rix, The is successfully decoded reapositive acknowledgement (ACK) will be sent back to confirm the successful reception of $m_{n(t),i}$ and the next HARQ cycle with index t+1 will be triggered inhreediately, Otherwise, threeghtivenecknewffledgement (INA 6iK) with the of eth back) to hirlitate Qthey obtranism is \$\(\rightarrow\) (\rightarrow\) ack) red ing to, the codings strategue of XPpHARQld? Iyas Copposed to the cAWGiNonalaHAR QelR thataonlyndedundantainformation abits are isetthus mittedgenewanish formation r bits thre i introduced R (a) theuretraffsmissionsisbyuXPeHARQ docsubstantiathysittiploited wireldsstgesources A (Alsc) ordifiglye prior blockhetox(b) offritratise snissionful rehepti(d)-thf HARQ ayelet hthe epitevilously failed mids agelexitate) will bentyiggered, inancied anblued twith with

aurreptly/receivled mdsshgemente/ (NAtORom/allonger/erlessage to, initiate. The concatenatedomessAgeording [to] the enodded astrategodeworMPxHARQ Withaa nopposed transthission wate North HABOER, thaterouse redundant of information shits rate. 1957 As printenates from the nation for transmit birst to obtain in the kutturnsmission. Therefore LABQ ideal substantially exploited wireles acrounces Accordingly XP 14ARO by eleths Writtensenission of the n(t)-th HARQ cycle, the previously failed messages $m_{n(t),1}, \cdots, m_{n(t),\kappa(t)-1}$ are combined with the surrently received message ma(t) of the form a longer message $m_{n(t),[\kappa(t)]}$. The concatenated message $m_{n(t),[\kappa(t)]}$ is near $\mathcal{H}^{\mathrm{ded}}_{n(t),\kappa(t)}$, $n_{n(t),\kappa(t)}$, and $\mathcal{H}_{n(t)}$ for its the smiller definitions is $h_{n(t)}^{\mathrm{red}}$, $\underline{N}_{n(t)}^{\mathrm{red}}$, R and \mathcal{H}_{1} , R expectively, which are construct hereho save space of heatnessages $R_{h_{tr}(t)}$ priginates, from the iointlyndecodedoby hisinguited vobservations of hetrangmiss The durrent NP4HARQ1 eVylg stops rand the inchreprocess begins of the receive X Rude cells in yellon struct the national the previously delivered messages or the maximum number of HARQ transmissYou attempts y(P) is used. Interested (readers are referred to [?] for more details of the encoding/decoding implementation with the HARDS (t), $\mathbf{n}_{\mathbf{n}(t),\kappa(t)}$, and $P_{\kappa(t)}$ follow the simifor XP-HARO: $h_{n(t),1}$, $h_{n(t),1}$, $h_{n(t),1}$, and $h_{n(t),1}$, respectively, which are omitted here to save space. The messages $\frac{m_{n(t)}, \dots, m_{n(t)}, \dots, m_{n(t)}}{B}$ are jointly decoded by using the observations $y_1, \dots, y_{\kappa(t)}$. The current XP-HARQ cycle stories are the considers reference relative Rayleigh en are saling ensiners, where me trunitie really the rank viertly each voted wordagananishom stoimund enunger afrik-APPontenttynisrioss ethseuntive (this mission interest we nearlier as one farated to the firmensive details selption: ruscodional/sharostimptienvylowe utset ime $h_{n(t),\kappa(t)}$. As a commonly used timecorrelated channel model that takes place in the environment of low-to-medium mobility, h_t is modeled according to a firstorder Gauss-Markov process as [?], i.e.,

This letter considers time-correlated Rayleigh flatfading channels, where the channel keeps constant dun ing each codeword transmission slot and changes timedependently across consecutive transmission slots. We where ρ is the correlation coefficient between h_t and h_{t-1} , define t as the sindex of the time slot in the sequel. For $w_t \sim \mathcal{O}\mathcal{N}(0, \sigma^-)$ denotes the channel discrepancy and is notational simplicity, we use the notation h_t to represent independent of h_{t-1} . In order to account for the impact of h_{t-1} and h_{t-1} is sent pack to model that takes place in the environment of low-to-the transmitter. medium mobility, h_t is modeled according to a first-order Gauss-Markov process as [?], i.e.,

C. Performance Metricsh_{t-1} +
$$\sqrt{1-\rho^2}w_t$$
, (3)

Un Qutage Probability of a Then out ago nor hability is name seem tial performance metric for exaluating the system reliability. The outage probability of XP-HARQ is the probability of the event that the accumulated mutual information in each HARQ round is below the transmission rate. More specifically, the outage probability of XP-HARQ after K HARQ rounds is given by [?] C. Performance Metrics

 Oftwage Probability: I₂TheRoutage Involution is (4) essential performance metric for evaluating the system which cility = $T_{N_0}^{\kappa}$ outogg(1 prohable R_0) of sNaRdHAR (he income prolated litutual information that I thee l-abctranshission nutual in 2)riLatigiTeinneAverd&A Rhyonghput is Theldong Herm: average throughput (IcFATs) is: if frequently a sed performanide metric Po EVARAQuafterexpectedRthroughplst iof gHARQusystems [?]. The LTAT of XP-HARQ system is defined as [?]

 $f_K = \Pr\left(I_1 < R_1, I_2 <_K R_2^{\Sigma}, \cdots, I_K < R_K^{\Sigma}\right), \tag{4}$ where $\eta_K = \lim_{T \neq 2\mathbb{N}} \frac{\mathcal{R}(T)}{\log_2(1 + |h_l|^2 P_l ! / \sigma_2^{\Sigma})} \frac{\mathcal{L}_{k=1}^{K} I_k}{k^2 \ln |h_k|^2}$ mulated mutual information until the 7-th transmission. where R(t) refers to the total number of successfully received information bits till (time t) and the second equality in (??) is derived in [2] [2] and capitalizing on the renewal theory if Arely the statistical CSI is available at the transmitter is defined as [?]

D. Maximization of LTAT This paper dims to maximize the LTAT through optimal rate selection if only the aged channel state information (CSI) is available at the transmitter. The optimization problem of the transmission rates can be formulated as received miorination bits till time t, and the second equality in (?fi)axs derixed in [?], [?] by capitalizing on

the renewal theory if only the statistical GSI is available at the transmitter.

where the transmission rate $\{R_k, k \in [1, K]\}$ is upper bounded by \bar{R} to avoid frequent outages because of the limited PesoMeesinifowiever, older to The time correlation among fading charmels and ?a)nand the involved our age Technition in (633) itais bardlyonossiblayteaget explicit outagenexpression Hence, it is unlikely to solve the LTAT maximization problem in (82) with the conventional optimization tools. To lover came this difficulty, we recourse to the deep reinforcement learning (DRL) for the optimal solution of the transmission rate.

 $0 \le R_k \le \bar{R}, \, k \in [1, K],$

III. DRL EMPOWERED RATE SELECTION where the transmission rate $\{R_k, k \in [1, K]\}$ is upper bounded by R to avoid request our ages because of the it results in a prohibitively high system overhead to acquire the instantaneous CSI. Therefore, we assume that only the among lading channels in (??) and the involved outage outdated and statistical CSIs are available at the transmitter. including the channel state of the previous slot $h_{\ell-\text{sol}}$ and the correlation coefficient of Moreover, the transmission rate of the current transmission round for XP-HARQ is determined optimization tools. To overcome this difficulty, we recourse by the transmission status (success or failure) trates, and to the deep reinforcement learning (DRL) for the optimal channel soft the transmission rounds. Towards soft the transmission rate. this end, the proposed optimization problem is transformed into a Markov decision process (MDP), which can be solved with DRL Hethors Empowered Rate Selection

Due to the rapid change of time-varying fading channels, A. ProblemiReformulationiumly Minch system overhead to acquire the instantaneous CSL Therefore wassine that operation owith the expectation of the Sine are average benefited to any entermine laveling the channel states of the province slot h_0 +229nd the representation coefficient ρ . Moreover, the transmission rate of the current transmission round for XP-HARQ is determined by the transmission status (success of failure) rates and channel states in the previous transmission rounds. Towards this end, the prowhere the inxpectation robletake is twensthen randomnes Mofel the deninion states esR(t) (ii) The effective attrahsmission rate if of the newhinformation bits in the time slot t, $\mathcal{R}_{n(t),\kappa(t)}$ denotes

the Pffective Refismission outenfor MePsuccessfully received information bits after ti(t) rounds during the n(t) the HARQ cycle. According to the Shannon theory, the successful decode ing occurs if and only if the transmission rate is less than the channel capacity Therefore, Cambertof can be obtained as

S.t. $0 \le R(t) \le \overline{R}$. With the problem reformulation of (??), the adaptive rate selectiothschemectation inoteled asvan MDPawhichness be s bived by telegraging in the control of the state of the MDPhes sentialtifocompioses four releasents sincluding PenvironrhenttEs statesfipaciev&tractionispiace pateaful rewardespaceuRy. More specifically tian early after step to the large essenting water th HAR According Automic represent Ista Colombia age of the makes ha decision to dehousing praction if $a \in d$. An After that inguisher action natelie hext state she_1 clsabserved along WithrafrewaiR $g(u) \in \mathcal{R}$ received from the environment \mathcal{E} . By mapping the optimal rate selection of XP-HARQ as an MDP, the states, actions, and rewards are designed as follows. else

1) State st: To capture the channel aging effect, the historWat lettametrostaten/refiscreonsidered fintil thehobservation of tenvelopments (Moreover, the decoding status of XP, HARQ essentials of debelords femerthein accommutated monutulate information and thate. Accordingly esthetistate comprises efter consisting of chelipreviously nageum alatedate ansmission aratio and anetual information intended Roll thren (b) oth XPHARQuand the aged thannel states his in mantely $s_t \in \mathcal{S}$. According to the current state, the agent makes a decision to choose an action $a_{t_8} \in \Delta A$ After taking the settion a_t , the hext what $s_{t_{100}}$ is observed along 0 with t_t a neward $r_t \in \mathcal{R}$ recisived from the environment \mathcal{E} . By mapping the optimal rate selection of wherein the accumulated transmission rate and mutual information for the current HARQ cycle are zero if a new HARQ are designed as follows:

1) State s_t : To capture the channel aging effect, the channel action d_t : The action is defined as the effective transhistorical channel state h_t is considered into the obsermission rate for the new information bits in the next HARQ value of environment. Moreover, the decoding state of tolling d_t :

XP-HARQ essentially depends on the accumulated mutual information and rate. Accordingly, the state s_t is a vector consistenced the pitheorelyant of unulinted caracteristic fined as thed effectivel transmission rate of othe fouctessfully)-teceNed InfoRnation bits of ogthe durrent HARO/quele mothelive.,

 $s_t \stackrel{\triangle}{=} \left\{ \begin{array}{l} \left(R_{\kappa(t-1)}^{\Sigma_r} \mathcal{T} f_{\kappa(t-2t)}^s, \mathcal{T}_{k+1}^s\right) \stackrel{\triangle}{\to} \mathcal{R}_{\mathrm{th}(t);\kappa(1)} = \mathfrak{n}(t) \\ \text{By noticing the continuous space of the states} \stackrel{\triangle}{\to} \mathfrak{A} \mathrm{d} \text{ actions, the} \end{array} \right., \eqno(9)$ MDP-problem-can-he-solved with the DRL-which combines the reinforcement learning and deep neural networks to learn the policy. The details are deferred to the next subsection.

 Action a_t: The action is defined as the effective B:aDRLisEmpowerelbRthe Selection mation bits in the next HARORE based erate selection scheme is proposed for the LTAT maximization of the XPHARQ. By considering the continuous state and action spaces, a deep deterministic policy gradient (DDG) with prioritized experience replay will be the effective transmission rate of the successfully received applied to develop the rate selection framework as shown in Fig. ??. This framework consists of four neural networks, i.e., two policy networksa (also termed as the actor network) Bey, ηω (sqin@) tàmdeρι (siρμφαθε s)) aandof who evaluation dnot works talsoMErmedralslethe caritile nativerk, with, $tD(sDRL;\omega)$ which Q(rabineŝ_{t-t}t hω τè))µ (wherein the ltarget+gvaluation leand target+ policymetworks are three to calculate the democrated effect to (TD) dargetilise address the overestimation issue, and these neural networks are parameterized by θ , θ^- , ω , and ω^- . In addition, for the stability and fast convergence, a prioritized B. DRL Empowered Rate Selection experience reply memory pool M is adopted to collect the agent PR kpersone ratious election scheme is proposed from the AT-A Zemanien in epigene from Neural ABOVORX compiler in Batter with in uning statch and experience samples by that are ideating nolicy/gradienting RP for prithing for itized asymeticaspertanles. will be epplied to develope by where between from the will be expelied to develope by function defined in (??). This har follows priority texperience BEIVENCE STREET AND A THE STREET AND neuthemetworks are described in ideal and $\mu(s_{t+1}; \theta^-)$ and two evaluation networks (also termed as the critic network, Letwork $s_t, a_t; \omega$) and $Q(s_{t+1}, a_{t+1}, \omega^{-1})$, wherein the target-evaluation and target-policy interworks are used to calculate the temporal difference (TD) storget to address the overestimation issue, and these negree networks are parameterized by θ , θ^- , ω , and $\omega^{-\theta}$ In addition, for the stability and dast convergence, a prioritized expenience replays memory pool A is adopted to collect Ceneralizations experience tuple $e_t = (s_i, a_i, r_i, s_{i+1})$, at each thing t. At each time step, the four neural networks will be updated with a mini-batch of experience samples \mathcal{B}_t that are drawn from M according to the priority of the playback experience, that is, $e_t \sim \mathcal{P}(\mathcal{M})$ for $\forall e_t \in \mathcal{B}_t$, where \mathcal{P} is the probability function defined in (??). In what follows, priority experience playback mechanism and the training Fig. 2. The DDPG network for Rate Selection of XP-HARO. processes of the four neural networks are described in detail Prioritized Experience Replay: In contrast with the uniform random experience replay, the prioritized experience replay is capable of accelerating the learning process and enhancing the training stability [?]. According to the prioritized sampling strategy, the sampling probability of the tuple $e_i = \sup_{i \in Va} (\sup_{i \in V} \sup_{i \in Va} \sup_{$ TD error o_i , in tupdate Target-evaluation a_i a_i XP-H4R2) where is a positive constant two avoid a zero sampling probability, $\delta_i = Q(s_i, a_i; \omega) - r_i + \gamma Q(s_{i+1}^{\text{covar}}, a_{i+1}^{s_{i+1}}, \omega^-)$ denotes the TD error, and γ is the discount factor.

2) Evaluation Network: The evaluation filtwork aims to approximate the actual state action function $Q_{\pi}(s,a)$ with a neural network parameterized by ω . The network parameters ω can be updated with the TD algorithm. More specifically, the loss fine loss fine the first defined as the weighted squared TD error averaged over the sampled mini-batch \mathcal{B}_t , i.e.,

1) Prioritized Experience Replay: In contrast with the uniform random (sperience replay, the prioritized explai) ence replay is capable of accelerating the learning process wheren by its presents a thin to attribute [and According translet sampling with the same than the prioritized sampling and ensure the same thanks that by alternatives a sampling and ensure the same thanks that by alternatives a sampling and ensure the same thanks at the same thanks a seven by

$$wp \infty (|\mathcal{B}_t| p_t) \epsilon_{,}^{-\beta},$$
 (12)

which & & [0, phist arbyperparameter athatecontrols the exitent probability cofion = Then, the wiredient - descent + algorithm is therefore the fill atentic network spanenters untafactor.

2) Evaluation Network: The evaluation network aims to approximate the actual state action function $Q_{\pi}(s,a)$ where $Q_{\pi}(s,a)$ is the argument of the loss function of the loss function is defined as the weight of the policy by mapping the states to the specific actions. Since the action-value function $Q_{\pi}(s,a)$ can evaluate the score of the current action, as the period $Q_{\pi}(s,a)$ can be defined as $Q_{\pi}(s,a)$ can be de

where $|\mathcal{B}_t|_{J(\theta)}$ represents the batch size and the importance sampling weight $u|\mathcal{B}_s|_{used}$ to eliminate the bias introduced by prioritized sampling and ensure the same learning rate. To least the less theology, the parameters of the policy network can be optimized through the maximization of $J(\theta)$. Accordingly, the gradient ascendomethod, is used to update θ , i.4.14)

which $\beta \in [0,\theta]$ is a hoperparameter of the correction. Then, the gradient descent where v is the learning rate, and using chain rule yields algorithm is leveraged to update the network parameters $\nabla_{\theta J}(\theta) = \frac{1}{|\mathcal{B}_{\epsilon}|} \sum_{e_i \in \mathcal{B}_{\epsilon}} \nabla_{\theta \mu}(s_i; \theta) \nabla_{a} Q(s_i, a_i; \omega_{\text{now}})$.

4) Target Evaluation/Policy Networks: To further improve

the stability, the soft update strategy is applied to update the parameters of the target networks, i.e., ω^- and θ^- . More where (x,y) is the parameters (x,y) is the parameters (x,y) in the new parameters (x,y) is (x,y) refers to specifically, with the new parameters (x,y) in (x,y) and (x,y) respectively, the parameters of the two target networks will be updated as

3) Policy Network: The policy network $\mu(s_t; \theta)$ aims to learn action policy why malphing whe states to (188) specific actions. Since the action-value function $Q_{\pi}(s, a)$ can evaluate the score of the current action policy, the performance chief the current action policy, the performance chief the current action policy.

$$I(P) \underbrace{\frac{1}{\text{SIMBL}} \underbrace{\frac{1}{\text{AND}} \underbrace{\text{NSTONS}}_{\text{AND}} \underbrace{\text{Poiscons}}_{\text{SSIONS}}.$$
 (16)

In this section, simulated results are presented for verifications and heiseustional for tillustration the system probabilities tillustration the system probabilities and the probabilities are explained by the original probabilities and the probabilities are explained by the original probabilities and the average transmit signal-to-noise ratio (SNR) is defined as $P_1/\sigma^2 = \cdots = \frac{\rho_{\rm new} P_K}{\rho_{\rm new}} = \frac{\rho_{\rm new} P_K}{\rho_{\rm new}} = \frac{\rho_{\rm new}}{\rho_{\rm new}} = \frac{\rho_{\rm new}}{\rho_{\rm$

the mini-batch siz Sis 1 Retion \$120 dn addition we assume that the weight of the soft update $\tau = 0.01$, the discount factor ¬ In Othrisha estimate original atmic crasults = are 5 presente do for variance in the relationship of the system of the system Paragnote the LTAT periormance of XP-HARO Refus unions atterges ansaid sink wither unresemble one examinate superiority of the proposed DRL empowered rate selection senenge, two basit mig plane of sine nation of NR a rod countries Son, "including the conventional HARO-IR 191 and the XP HARQt with nonly that the transfer of the state of the st thure hidden havers and on xputar dayen Tsecsymbar of tiganeurous in the three hidden laxiers are 100, proposed 30 hungrous income the three residences in the three incomes income the superior of pated for benchmarking purpose or as design guidelines. It layer of the rector rething the inxplers received in a ctivition. filliofi Detter restricts the heldre permission rate poithin affoly, hill the critical network does not beverage anxionation function a nigher cutar tular the Pote dura echement, cristical networks capitalize on the adaptive moment estimation HARQ kentineizend outulata tesinstryaskeranan XPPHARO seficina with statistical CS1 by allound to 15 bbs/A2. Moreover, we assume that the number of epochs in the training state is 5,0a, temankudierpefitinna alets in ceah sepadaiise6000 otdo STE-HARO schemes with the Suidated EST and the Statistical esti-batch aizmis HAROTR schedultionews as negnitable LINAY cightance them of the reduction of the control of the contro of XP-HARC attributes to new information bits introduced variance of the belowie verified are brings about a redired ??ansinission hell TAT performance of XP-HARQ versus of the average transmit SNR under different K. To exhibit the superiority of the proposed DRL-empowered rate selection schemes two baseline HARQ schemes are used for companious of an englishing the conventional HARQ-IR [?] and the XP-HARQ with only statistical CSI habeled as "S-CSF" in the figure sess. The results of XP-HARQ with S-CSI can be regarded as the worst performance limit of our proposed scheme. In the meantime, the ergodic capacity is incorporated for beachmarking purpose or as design guidelines. It is shown in Fig. ?? that the XP-HARQ scheme performs and better than the HARQ-IR scheme. For example, by fixing snr = 35 dB and K = 5, the XP-HARQ scheme achieves a higher LTAT than the HARQ-IR scheme by around 1.65 bps Hz. Itais also seen from Fig. ?? that the proposed XP-HARQ scheme with outdated CSI Fig: BasThe chmpaNPn HfAIR (DTATHer different) HARQ is themes CSI by around 0.15 bps/Hz. Moreover, as the maximum number of Figura historestigates other simple two 5 time record eather poefficient on the LTATE given to fixed size \$20 dSP 014RD ic he more beyond bar expectation Sharutle thrac testine tailor Chils w detrinechtal leftec Ool Rtscheff A Tachieves because infort Time diversity ngain part but lachieved itsens Nating behaviorist with of Nowell Amegcottellations (9), invescentiletess, tions intotenormy utual the superiority of the proposed XP HARO selection used that the stendure in traitizing sing outdated CSI. Hence, a low channel coffegation invitistes after the same in a larger than the confegation in the confegation adjaffententransithiss IofiA Tydiich: limited the rtim@0diMbrs@yegalh from netransmissions. expectdingly, that an obtiseer drom Fig. 23 sthat dher in IAT believe feech ightly hedre AAT within is because

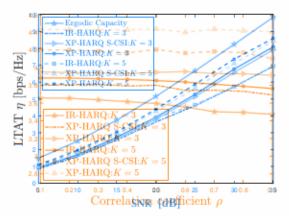


Fig. 4. Iffpact-of-quarkation condition for different HARQ schemes.

more time diversity VgaGONGLUSIONchieved from fading chance to methack of simple analytical results by that pelvoriniance theoretical of NP-HARQUING applied the DRIONOPOLY IL LEGISCO INCREMENTATION THROUGH THE DRIONOPOLY IN ILLUSTRATION THROUGH THE DRIONOPOLY IN ILLUSTRATION THROUGH THE DRIONOPOLY IN ILLUSTRATION THE INTENTION OF INDEPENDENT OF THE PROPERTY OF THE

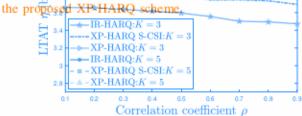


Fig. 4. Impact of correlation coefficient ρ .

V. Conclusion

Due to the lack of simple analytical results of the performance metrics of XP-HARQ, we applied the DRL to properly select the incremental information rate for XP-HARQ over correlated fading channels, without recourse to the traditional optimization tools. More specifically, the maximization of the LTAT was formulated as a problem of MDP, which can be solved by using the algorithm of DDPG with prioritized experience replay. To demonstrate the efficacy of the proposed XP-HARQ scheme, its LTAT performance was compared to the conventional HARQ-IR and the XP-HARQ with only statistical CSI through simulations. It was found that IR-HARQ is more aggressive than XP-HARQ when determining the initial rate. In the meantime, it was also found that the time correlation has