Reviewer # 6, the motivation is coming from the many significant applications of multi-armed bandit (MAB) models, e.g. clinical trials, ad placement, adaptive routing etc. For instance in a clinical trials setting it is natural to try to model the assertion: ‘if a particular drug was successful the n-th time it’s used, then it is very likely that it will be successful again the $n+1$-th time it’s used’. This assertion requires precisely the notion of Markovian dependence, and it can not be captured using the i.i.d. model. Through our work we increase the expressive power of MAB models and make it possible to express assertions like the aforementioned. Our work is technical and follows the typical approach of mathematical research where one generalizes results in broader mathematical situations. Also in the survey [BCB12] one can find references to works where the objective is regret minimization rather than best arm identification, involving rested and restless Markovian MAB models which have hundreds of citations. Moreover we disagree with the statement that the exponential family of stochastic matrices seems artificial. This is because many well established works in i.i.d. MAB, such as [GK16], as well as the work of Cappé et al on KL-UCB strategies, build on the notion of an exponential family of probability distributions. An exponential family of stochastic matrices is a generalization of an exponential family of finitely supported probability distributions, and so we essentially assume the same type of structure as many other major works that deal with the i.i.d. case.

Reviewer # 6, we strongly believe that the value of this work is coming precisely from its technical depth and the new tools introduced. The reason that Markovian bandits are much less studied compared to i.i.d. bandits is that the tools that probability theory has to offer in the case of Markovian dependence are so much less developed compared to the i.i.d. tools. We’re the first to take the notion of an exponential family of stochastic matrices, initially introduced in the context of large deviations for Markov chains, and import it in the learning theory community. Moreover, in order to use it as the foundation of our Markovian MAB model we develop many technical properties of the family. One of the most important contributions of this work is the Chernoff type bound for Markov chains, we discuss it further in the next paragraph but here we note that reviewer # 6 incorrectly states that this is an asymptotic concentration result, while our result is non asymptotic. For the Markovian MAB lower bound our contribution is two fold. We first add the technical pieces needed in order to extend the standard change of measure argument to the Markovian setting, i.e. the regeneration argument for Markov chains. Moreover, we simplify the argument of [GK16] where they invoke some ‘transportation Lemma’, by observing that this so called ‘transportation Lemma’ is nothing more than a special case of the well known data processing inequality from information theory. The upper bound analysis indeed follows [GK16] quite closely, but even at this part we are able to fix a broken argument from [GK16]. In particular after contacting the second author of [GK16] we have verified that the last part of the proof of their Proposition 12 is wrong, and their C has to depend on δ rendering their whole result shaky.

Reviewer # 3, for the Markovian bandit problem we take $S \subset \mathbb{R}$ and in L.177 we take $\phi(x) = x$, so the rewards that we observe are the states. Moreover the learning algorithm does not need to know the whole generator P, but only some scalar quantity related to it. For instance when P is positive $\max_{x,y,z} \frac{P(x,z)}{P(y,z)}$ is all we need, while when the arms are i.i.d. we do not need to know something. In any case the generator P is being tilted exponentially by $\theta_1, \ldots, \theta_K$ in order to produce the K Markovian arms, and the algorithm is agnostic of all the tilts and of all the stationary means.

Reviewer # 4, analyticity of the PF eigenvalue, left and right eigenvectors is guaranteed because the eigenvalue is simple and we impose the two normalization constraints on the eigenvectors so that they are unique and the ‘only determined up to a scalar factor’ issue that [Lax07] mentions doesn’t occur. Reviewer # 4, thank you for your detailed comments!

Regarding our Chernoff bound for Markov chains, roughly speaking large deviations dictate that the true rate of exponential decay is given by a relative entropy representing an information projection. This relative entropy already implicitly conveys information about the eigengap of P, the mixing time of the chain, it is ϕ specific etc. Any other bound that has a different rate of exponential decay is suboptimal and leads to a Pinsker type inequality like the one in L.171-172 (reviewer # 3, this inequality holds under the assumptions of Theorem 3.3 from [Lez98]). Therefore our bound has the optimal exponent, and the only way of potentially improving it is by introducing a better constant in the prefactor or by generalizing it. Reviewer # 4, L.163-164 are meant to say that when our assumptions are active and the assumptions of some other work are active, then our bound is better. For instance, we can consider the important case of a strictly positive P (no matter ϕ). Then Theorem 3.3 from [Lez98] and our Theorem 1 are both in action, with ours dominating. Reviewer # 4, your observation is correct, in Theorem 1 we should have first fixed ϕ and then fix P, but this does not alter the optimality statement that we made above. To the best of our knowledge (following citations and consulting local experts) this constitutes a novel and optimal non asymptotic concentration result for Markov chains. Bounds of this type are of fundamental importance, not only in learning theory, e.g. MCMC, Markov decision processes etc, but also in general in the world of applied probability. For instance the work of [Gil93], which presents a suboptimal bound for reversible Markov chains accompanied with computer science applications, is coming from FOCS, (best TCS conference), and bounds inspired by the work of [Gil93] based on matrix perturbation theory without accompanying applications [Din95, Lez98, LP04], are coming from the Annals of Applied Probability (one of the best probability theory journals). In this work we’re not only developing an optimal Chernoff bound for Markov chains using new techniques, but we also apply this in order to resolve the complexity of the best Markovian arm problem.