We thank anonymous reviewers for their valuable comments and suggestions.

Comment 1: Training time (Reviewer #1)
Response: We have included the training time in the paper (L318) that "Within one million timesteps, the training wall-clock time for our TRGPPO is 33 min; for PPO, 32 min". We use several techniques to allow efficient optimization, including the problem reduction, the DNN-approximation and problem discretization (described in Sec 5.1).

Comment 2: Concerns of the experiment evaluation, random number and baseline (Reviewer #1 & #2 & #3)
Response: We used the averaged top 10 reward following the setting in [24], which could somewhat reflect the algorithm's ability on searching good solution but we agree it's somehow unreliable. However, we have also plotted the learning curves in Fig. 3, which could help infer the stability of the algorithm. Per your suggestion, we have made several revisions, listed as follows: 1) *report the averaged reward over all episodes of training*; 2) *compare with the baseline of adaptive KL regularization of PPO* (and clarify the related description in the introduction); 3) *run a hyperparameter sweep for ϵ of PPO over [0.1,0.6] with step 0.05*; 4) *increase the number of random seeds to 10*. We normalized the scores for each environment so that the random policy gave a score of 0 and the best score was set to 1. The averaged normalized scores (over 60 runs with all episodes of training for each algorithm, on 6 environments) are as follows: ***TRGPPO*: <u>0.629</u>**; *PPO($\epsilon = 0.2$,default)*: <u>0.441</u>; *PPO($\epsilon = 0.25$, optimal PPO)*: <u>0.484</u>; *PPO-adaptiveKL*: <u>0.422</u>. We will add more details of the results in the final version.

Comment 3: Concerns about Lemma 2 (L113) and several typos (Reviewer #1)
Response: Thanks for your careful reading. The correct form of the LHS of the equation in Lemma 2 should be $\mathbb{E}_{\pi_{t+1}}[\pi_{t+1}(a)|\pi_0]$. This typo does not affect the correctness of the lemma and the remaining theoretical results in the manuscript. We will rectify all the typos in the final version.

Comment 4: The existence of $\pi^{\text{PPO}}_{\text{new}}$ (L235) (Reviewer #1)
Response: The problem is that how we can find $\pi^{\text{PPO}}_{\text{new}} \in \Pi^{\text{PPO}}_{\text{new}}$ that achieves minimum KL divergence on all states s_t, which can be formalized as $\min_{\pi\in\Pi^{\text{PPO}}_{\text{new}}} \left(D^{s_1}_{\text{KL}}(\pi_{\text{old}},\pi),\dots,D^{s_T}_{\text{KL}}(\pi_{\text{old}},\pi)\right)$. Note that $\pi(\cdot|s_t)$ is a conditional probability and theoretically the optimal solution on different states are independent from each other. Thus the problem can be optimized by independently solving $\min_{\pi(\cdot|s_t)\in\{\pi(\cdot|s_t):\pi\in\Pi^{\text{PPO}}_{\text{new}}\}} D_{\text{KL}}\left(\pi_{\text{old}}(\cdot|s_t),\pi(\cdot|s_t)\right)$ for each s_t. The final $\pi^{\text{PPO}}_{\text{new}}$ is obtained by integrating these independent optimal solutions $\pi^{\text{PPO}}_{\text{new}}(\cdot|s_t)$ on different state s_t. We have provided detail in Appendix D and we will add more explanation in the final version.

Comment 5: Concerns about the details of and the reproducibility of the experiment (Reviewer #1)
Response: We used Gaussian and Gibbs policy for continuous and discrete tasks respectively, parametrized by a DNN. For baseline, we used the setting recommended in the original paper. We have also submitted a link of the source code as supplementary (L48 in the paper). We will add more details and release our code in the final version.

Comment 6: How is Eq. (4) transformed into Eq. (5) in supplementary? (Reviewer #2)
Response: To be brief, *let's number the equations in Eq. (4) by (a)-(d)*. First, by (a)(b), we have $\lambda \neq 0$, since if $\lambda = 0$ then $\nu = 0$ (by (a)), which contradicts (b). Second, by (c) and $\lambda \neq 0$, we have $\sum_{a\in\mathcal{A}} p'_a \log(p'_a/p_a) = \delta$. Third, taking (a) into (d), we have $p'_a/p_a = \nu/\lambda = (1-p'_{a_t})/(1-p_{a_t})$ for $a \neq a_t$. Then, taking this equation into $\sum_{a\in\mathcal{A}} p'_a \log(p'_a/p_a) = \delta$, we obtain Eq. (5). We will add more details in the final version.

Comment 7: The performance on Humanoid (Reviewer #2)
Response: One possible explanation is that the larger clipping range of our TRGPPO may make it suffer from the noisy estimated advantage values, especially at the later training phase where the advantage values are large and noisy. This issue could be addressed using our adaptive clipping scheme by taking the trade-off between exploration and stability into account. In particular, in the revised version, we have implemented two variants of TRGPPO: linearly decaying ϵ from 0.2 to 0.1 (named by *TRGPPO-decay*) or clipping the clipping ranges (named by *TRGPPO-clipping*), i.e., $l^{\delta,\epsilon,\epsilon_t}_{s,a} = clip(l^{\delta}_{s,a},\epsilon_t,\epsilon)$, $u^{\delta,\epsilon,\epsilon_t}_{s,a} = clip(l^{\delta}_{s,a},1/\epsilon,1/\epsilon_t)$, where $0<\epsilon<1$ and $\epsilon_t = \epsilon t/T$ are parameters to control the level of the clipping ranges, t and T are the current and total training iterations respectively. Both these two methods could improve the reward and sample efficiency. The averaged episode rewards over all episodes of training on Humanoid are as follows: *TRGPPO-decay*: <u>3013.3</u>; ***TRGPPO-clipping*: <u>3148.1</u>**; *PPO*: <u>2944.2</u>. The timesteps ($\times 10^3$) to hit the threshold are as follows: *TRGPPO-decay*: <u>7514</u>; ***TRGPPO-clipping*: <u>7132</u>**; *PPO*: <u>9088</u>.

Comment 8: Prove that TRGPPO converges to the optimal policy (Reviewer #2)
Response: Thanks for your suggestion. It's interesting to prove such convergence property.However, there seems does not exist closed-form of our clipping range in Eq. (5), making it hard to measure the improvement of $\Delta^{\text{TRGPPO}}_{\pi_0,t}$ by the explicit form of $\mathbb{E}_{\pi_{t+1}}[\pi_{t+1}(a_{\text{opt}})|\pi_t]$ (see Eq. (3)). Alternatively, we plan to work on this by analyzing the corresponding bound for each term in Eq.(3).

Comment 9: Some mathematical formulae in the paper could be better formatted. (Reviewer #3)
Response: Thanks for your comment. We will polish mathematic notions and align expressions in the final version.