WHEN ERRORS CAN BE BENEFICIAL: A CATEGORIZATION OF IMPERFECT REWARDS FOR POLICY GRADIENT

Anonymous authors
Paper under double-blind review

ABSTRACT

Training language models via reinforcement learning often relies on imperfect *proxy* rewards, since *ground truth* rewards that precisely define the intended behavior are rarely available. Standard metrics for assessing the quality of proxy rewards, such as ranking accuracy, treat incorrect rewards as strictly harmful. In this work, however, we highlight that not all deviations from the ground truth are equal. By theoretically analyzing which outputs attract probability during policy gradient optimization, we categorize reward errors according to their effect on the increase in ground truth reward. The analysis establishes that reward errors, though conventionally viewed as harmful, can also be benign or even beneficial by preventing the policy from stalling around outputs with mediocre ground truth reward. We then present two practical implications of our theory. First, for reinforcement learning from human feedback (RLHF), we develop reward model evaluation metrics that account for the harmfulness of reward errors. Compared to standard ranking accuracy, these metrics typically correlate better with the performance of a language model after RLHF, yet gaps remain in robustly evaluating reward models. Second, we provide insights for reward design in settings with verifiable rewards. A key theme underlying our results is that the effectiveness of a proxy reward function depends heavily on its interaction with the initial policy and learning algorithm.

1 INTRODUCTION

Training language models via reinforcement learning commonly relies on imperfect *proxy* rewards. This reliance is largely unavoidable since specifying *ground truth* rewards that exactly capture the intended behavior is rarely feasible. For example, in reinforcement learning from human feedback (RLHF) (Ouyang et al., 2022), learned reward models serve as proxies for a ground truth reward that is assumed to govern human preferences. Moreover, so-called verifiable rewards (Lambert et al., 2024; Guo et al., 2025) are often also imperfect proxies: rewards for mathematical reasoning usually do not verify intermediate steps and can suffer from incorrect parsing (Huang et al., 2025b; Tao et al., 2025), and rewards for code generation are based on unit tests, which are inherently incomplete (Liu et al., 2023; Guo et al., 2025; Olmo et al., 2025).

Despite the widespread use of proxy rewards, there is limited understanding of how discrepancies between the proxy and ground truth rewards affect the reinforcement learning process. Arguably the most well-known potential consequence of such discrepancies is *reward hacking*, where maximizing the proxy reward results in poor ground truth performance (Amodei et al., 2016; Skalse et al., 2022; Pang et al., 2023; Gao et al., 2023; Fluri et al., 2025; Laidlaw et al., 2025). Reflecting this concern, standard metrics for assessing the quality of a proxy reward function, including ranking accuracy (Lambert et al., 2025) and mean squared error (Huang et al., 2025a), consider deviations from the ground truth reward as strictly harmful. However, this coarse treatment overlooks the possibility that different erroneous proxy rewards may influence the learning process in distinct ways.

In this work, we highlight that not all reward errors are equal, or even necessarily harmful. Formally, *reward error* refers to cases where $r_{\mathrm{P}}(x, y) \neq r_{\mathrm{G}}(x, y)$, for a proxy reward function r_{P}, ground truth reward function r_{G}, and one or more input-output pairs (x, y). Focusing on *policy gradient*—the predominant approach for training language models via reinforcement learning—we theoretically characterize the effect of reward errors on the increase in ground truth reward (Section 3). In particular, we categorize reward errors into *harmful*, *benign*, or *beneficial*. Harmful errors, aside from causing reward hacking, can slow down the rate at which the ground truth reward increases. Benign errors

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

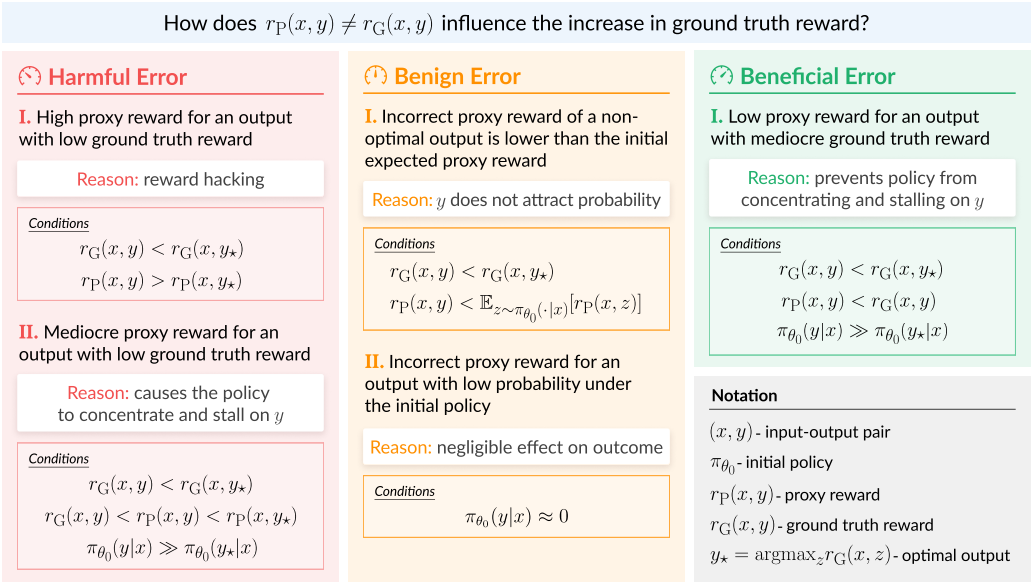

Figure 1: **Reward error categorization overview.** We categorize *reward errors*—cases where the proxy and ground truth rewards disagree on one or more input-output pairs—according to their effect on the increase in ground truth reward under policy gradient optimization (Section 3). Aside from being harmful, we prove that reward errors can also be benign or even beneficial. This categorization depends on the interplay between the proxy reward function, initial policy, and policy gradient updates. Sections 4 and 5 explore practical implications of our theory for reward model evaluation in RLHF and reward design in settings with verifiable rewards.

have negligible impact on optimization. Lastly, and perhaps most strikingly, we prove that some errors are beneficial, accelerating the increase in ground truth reward beyond what is achieved when optimizing the ground truth reward directly. See Figure 1 for an overview of our categorization.

We derive this categorization by analyzing which outputs attract probability during policy gradient optimization for linear softmax policies. The main theoretical result, which underlies the existence of beneficial reward errors, shows that outputs with mediocre reward can attract probability and stall policy gradient in their vicinity for an arbitrarily long time. Thus, assigning low proxy reward to outputs with mediocre ground truth reward can accelerate the increase in ground truth reward by steering the policy away from such outputs. The analysis is corroborated by controlled experiments.

As a practical application of our theory, we consider the problem of reward model evaluation for RLHF (Section 4). Reward models are primarily evaluated through ranking accuracy (Lambert et al., 2025; Zhou et al., 2025; Liu et al., 2025b; Frick et al., 2025; Malik et al., 2025), which treats all incorrect rankings of outputs as equally harmful. Based on our categorization of reward errors, we develop ranking accuracy variants that account for the harmfulness of an incorrect ranking to policy gradient optimization. Experiments show that these harm-aware metrics typically correlate better with the performance of a language model after RLHF, across datasets and model families (Llama (Grattafiori et al., 2024), OLMo (Olmo et al., 2025), and Qwen (Yang et al., 2025)). Yet, despite these gains, the correlation can still be weak, showcasing challenges in robustly evaluating reward models.

Beyond reward model evaluation, we explore implications of our theory for reward design in settings with verifiable rule-based rewards (Section 5). Namely, we demonstrate that rewarding partially correct outputs (*cf.* (Pyatkin et al., 2025; Sun et al., 2026)) can be detrimental for optimization if the initial policy is noticeably more likely to produce partially correct outputs than fully correct ones.

A central theme of this work is that the effectiveness of proxy rewards cannot be assessed solely by how much they deviate from the ground truth. It is essential to consider the interplay between the proxy reward function, initial policy, and learning algorithm. Our results take a step towards characterizing principles for proxy reward evaluation and design that account for this interplay. We hope that they will inspire further research in this direction, given the growing usage of reinforcement learning for training language models in complex environments.

Related work. We discuss related work throughout and defer an extended account to Appendix A.

2 PRELIMINARIES

We use π_θ to denote a language model, parameterized by θ, that maps an input (*i.e.*, prompt) $x \in \mathcal{X}$ to a distribution over outputs $y \in \mathcal{Y}$, where $\mathcal{X}$ and $\mathcal{Y}$ are the input and output spaces, respectively. We will also refer to π_θ as a *policy*.

2.1 REINFORCEMENT LEARNING FOR LANGUAGE MODELS

Reinforcement learning is a key component of language model post-training pipelines (Olmo et al., 2025; Yang et al., 2025). In many cases, it is difficult to specify a *ground truth* reward function $r_{\mathrm{G}} : \mathcal{X} \times \mathcal{Y} \to [-1, 1]$ that exactly defines the intended objective. Thus, a *proxy* reward function $r_{\mathrm{P}} : \mathcal{X} \times \mathcal{Y} \to [-1, 1]$ is used instead.[1] A prominent example is reinforcement learning from human feedback (RLHF) (Ouyang et al., 2022), where r_{P} is a learned reward model and r_{G} is a reward function assumed to capture human preferences. Perhaps less obvious examples arise in environments with so-called verifiable rewards, such as mathematical reasoning and code generation. An ideal reward function would distinguish fully correct outputs from partially correct or incorrect ones. Yet, rewards in math environments typically do not verify intermediate steps and can suffer from incorrect parsing (Lambert et al., 2024; Huang et al., 2025b; Tao et al., 2025), and rewards for code generation rely on unit tests, which are fundamentally limited (Guo et al., 2025; Olmo et al., 2025).

Given a proxy reward function r_{P}, the language model π_θ is usually optimized via *policy gradient* methods (*e.g.*, PPO (Schulman et al., 2017), RLOO (Ahmadian et al., 2024), and GRPO (Shao et al., 2024)). This amounts to maximizing through gradient updates the expected proxy reward $\mathbb{E}_{x\sim\mathcal{D},y\sim\pi_\theta(\cdot|x)}[r_{\mathrm{P}}(x, y)]$, where $\mathcal{D}$ is a distribution over inputs. Optionally, a KL regularization term may be included to penalize divergence from some reference policy (*cf.* Ziegler et al. (2019); Stiennon et al. (2020); Ouyang et al. (2022)). The premise is that, if r_{P} is a good proxy for r_{G}, increasing expected proxy reward should also increase the expected ground truth reward $\mathbb{E}_{x\sim\mathcal{D},y\sim\pi_\theta(\cdot|x)}[r_{\mathrm{G}}(x, y)]$.

2.2 EVALUATING THE QUALITY OF PROXY REWARD FUNCTIONS

Assessing the quality of a proxy reward function r_{P} is a problem of both practical and theoretical interest (Gleave et al., 2021; Wulfe et al., 2022; Skalse et al., 2024; Karwowski et al., 2024; Lambert et al., 2025). This problem is particularly salient in the context of RLHF, where *ranking accuracy* (Definition 1) is the primary metric used to evaluate reward models (Lambert et al., 2025; Liu et al., 2025b; Frick et al., 2025; Zhou et al., 2025; Malik et al., 2025). A main reason for its adoption in practice is that it does not require direct access to the ground truth r_{G}, relying instead on preference labels that can be obtained through human annotation.

Definition 1. For a set $\mathcal{S}$ containing preference examples $(x, y^+, y_1^-, \ldots, y_K^-)$, where $x \in \mathcal{X}$ is an input and $y^+, y_1^-, \ldots, y_K^- \in \mathcal{Y}$ are outputs satisfying $r_{\mathrm{G}}(x, y^+) > \max_{k\in[K]} r_{\mathrm{G}}(x, y_k^-)$, the *ranking accuracy* of r_{P} is given by:

$$\mathrm{Acc}(r_{\mathrm{P}}; \mathcal{S}) := \frac{1}{|\mathcal{S}|} \sum\nolimits_{(x,y^+,\{y_k^-\}_{k\in[K]})\in\mathcal{S}} \mathbb{1}\left[\max\nolimits_{k\in[K]} r_{\mathrm{P}}(x, y_k^-) < r_{\mathrm{P}}(x, y^+)\right],$$

where $\mathbb{1}[\cdot]$ is an indicator function and $[K] := \{1, \ldots, K\}$.

Ranking accuracy, along with other metrics for evaluating proxy rewards in the literature (Gleave et al., 2021; Wulfe et al., 2022; Skalse et al., 2024; Huang et al., 2025a), rests on a common assumption: deviations of r_{P} from r_{G} are strictly harmful. Moreover, it treats all incorrect rankings of outputs as equal. Our aim is to show that this view is too coarse, as it disregards the fact that different *reward errors* (*i.e.*, cases where the proxy and ground truth rewards disagree on one or more input-output pairs) affect the policy optimization process in distinct ways.

3 CATEGORIZATION OF REWARD ERRORS

Reward errors are conventionally viewed as *harmful* due to the risk of reward hacking—if an output with low ground truth reward is assigned a high proxy reward, maximizing the proxy reward may lead to poor ground truth performance (Amodei et al., 2016; Skalse et al., 2022; Pang et al., 2023;

[1] As typical in language model applications, we consider rewards defined over complete outputs.

Gao et al., 2023; Karwowski et al., 2024; Fluri et al., 2025). However, we prove that aside from being harmful, reward errors can also be *benign* or even *beneficial*. Specifically, by theoretically analyzing which outputs attract probability during policy gradient optimization, we categorize reward errors according to their effect on the increase in ground truth reward; see Figure 1 for an overview.

Section 3.1 presents the technical setting, after which Section 3.2 characterizes two types of benign reward errors by showing that outputs with low proxy reward or low probability under the initial policy negligibly affect the outcome of policy gradient. Section 3.3 then delivers our main theoretical result, establishing the existence of beneficial reward errors. It identifies that outputs with mediocre proxy reward can attract probability and trap policy gradient for an arbitrarily long time. Thus, assigning low proxy reward to outputs with mediocre ground truth reward can accelerate the increase in ground truth reward by steering the policy away from such outputs. Lastly, the attraction to mediocre outputs reveals an additional type of harmful reward error (Section 3.4). Namely, suppose that an output with low ground truth reward is incorrectly assigned a higher proxy reward. Even if the proxy reward is not high enough to induce reward hacking, this error may still slow down the increase in ground truth reward. Controlled experiments support our theoretical findings.

3.1 Technical Setting

Policy parameterization. We consider bandit environments with a single input (*i.e.*, context) $x \in \mathcal{X}$, finite set of outputs $\mathcal{Y}$, and *linear softmax policies* (Agarwal et al., 2021; Mei et al., 2023a; Razin et al., 2024; Lin et al., 2025; Foster et al., 2025). Each output $y \in \mathcal{Y}$ is associated with a feature vector $\phi(y) \in \mathbb{R}^D \setminus \{0\}$ and the policy π_θ is defined by $\pi_\theta(y) := \exp(\langle \phi(y), \theta \rangle) / \big(\sum_{z \in \mathcal{Y}} \exp(\langle \phi(z), \theta \rangle)\big)$, where $\theta \in \mathbb{R}^D$. For conciseness, in this section we omit the input $x \in \mathcal{X}$ from our notation. Despite the apparent simplicity of linear softmax policies, they induce non-concave proxy and ground truth objectives for $r_\mathrm{P}, r_\mathrm{G} : \mathcal{Y} \to [-1, 1]$:

$$V_\mathrm{P}(\theta) := \mathbb{E}_{y \sim \pi_\theta}[r_\mathrm{P}(y)] \ , \ \ V_\mathrm{G}(\theta) := \mathbb{E}_{y \sim \pi_\theta}[r_\mathrm{G}(y)] \, ,$$

whose optimization properties are not well understood (Mei et al., 2020b; Agarwal et al., 2021; Li et al., 2021; Mei et al., 2023a; Lin et al., 2025).

Optimization. Language model post-training is often done with small learning rates (Xu et al., 2024; Lambert et al., 2024; Olmo et al., 2025). Accordingly, and following prior theoretical analyses of gradient-based optimization with small learning rates (*e.g.*, Saxe et al. (2014); Woodworth et al. (2020); Vardi & Shamir (2021); Chou et al. (2023); Bietti et al. (2025)), we analyze *gradient flow*:

$$\frac{d}{dt}\theta_t = \nabla V_\mathrm{P}(\theta_t) \ , \ \ t \geq 0 \, , \tag{1}$$

where θ_t denotes the parameters at time t of training and π_{θ_0} is the initial policy.

Limitations. The technical setting involves three main simplifications: *(i)* it considers linear softmax policies; *(ii)* it assumes access to exact gradients of the expected proxy reward V_P; and *(iii)* it focuses on bandit environments with a single input. As Sections 4 and 5 demonstrate empirically, although these simplifications abstract away complexities that may stem from using a neural network architecture, sample-based estimates of $\nabla V_\mathrm{P}(\theta)$, or the interaction between inputs, they allow deriving insights that apply to practical settings.

3.2 Benign Errors

3.2.1 Proxy Reward Is Lower Than the Initial Expected Proxy Reward

We begin by making a straightforward, yet important observation: outputs with proxy reward lower than the initial expected proxy reward, $V_\mathrm{P}(\theta_0)$, generally do not attract probability. This becomes apparent by examining the optimization dynamics for the logit $\langle \phi(y), \theta_t \rangle$ of an output $y \in \mathcal{Y}$.

Proposition 1 (Proof deferred to Appendix D.2)**.** *Suppose that we maximize via gradient flow the expected proxy reward V_P (Equation (1)). For any output $y \in \mathcal{Y}$ and time $t \geq 0$ it holds that:*

$$\frac{d}{dt}\langle \phi(y), \theta_t \rangle = \underbrace{\pi_{\theta_t}(y) A_\mathrm{P}(y; \theta_t) \cdot \|\phi(y)\|^2}_{\text{contribution due to } y} + \sum\nolimits_{z \in \mathcal{Y} \setminus \{y\}} \underbrace{\pi_{\theta_t}(z) A_\mathrm{P}(z; \theta_t) \cdot \langle \phi(z), \phi(y) \rangle}_{\text{contribution due to other output}} \, ,$$

where $A_\mathrm{P}(z; \theta_t) := r_\mathrm{P}(z) - V_\mathrm{P}(\theta_t)$ is the advantage of $z \in \mathcal{Y}$ under r_P and π_{θ_t}.

Since $V_{\mathrm{P}}(\theta_t)$ is non-decreasing under gradient flow, if $r_{\mathrm{P}}(y) < V_{\mathrm{P}}(\theta_0)$, then the contribution of y to its own logit dynamics is negative for all $t \geq 0$, and is more negative the lower $r_{\mathrm{P}}(y)$ is compared to $V_{\mathrm{P}}(\theta_t)$. The contribution due to other outputs counteracts this negative push only if the proxy reward function and output features are misaligned, in the sense that there exist outputs with reward higher than $V_{\mathrm{P}}(\theta_0)$ whose features have a large inner product with $\phi(y)$. Thus, as long as the feature geometry reasonably accords with the proxy rewards, an output y with $r_{\mathrm{P}}(y) < V_{\mathrm{P}}(\theta_0)$ will not attract probability. This implies that the following reward error type is benign.

Benign Error I: Incorrect proxy reward is lower than the initial expected proxy reward. If an output y is assigned an incorrect proxy reward $r_{\mathrm{P}}(y) \neq r_{\mathrm{G}}(y)$ satisfying $r_{\mathrm{P}}(y) < V_{\mathrm{P}}(\theta_0)$, then it will generally not attract probability. As a result, unless y achieves maximal ground truth reward, this reward error usually does not harm the increase in ground truth reward.

3.2.2 Incorrect Proxy Reward for Improbable Outputs

The logit dynamics in Proposition 1 further reveals that the contribution of an output y is dampened by its probability $\pi_{\theta_t}(y)$. This suggests that the optimization trajectory is not significantly affected by incorrect rewards assigned to outputs with low probability under the initial policy π_{θ_0}.

Proposition 2 formalizes this prospect. Namely, suppose that r_{P} and r_{G} differ only on a set of outputs with low initial probability. Proposition 2 shows that policies learned by maximizing these reward functions remain close in total variation distance throughout training. This is in line with common intuition, by which low probability outputs have little influence since they are unlikely to be sampled when constructing sample-based gradient estimates. However, we prove a stronger claim: even in an idealized setting with access to exact gradients—where all outputs factor into the update—the contribution of low probability outputs is still negligible. This is a consequence of the optimization dynamics induced by policies that produce a distribution over outputs via the softmax function.

Proposition 2 (Proof deferred to Appendix D.3)**.** *Suppose that the proxy reward function r_{P} and ground truth reward function r_{G} differ only on a set of outputs $\mathcal{Z} \subseteq \mathcal{Y}$, i.e., $r_{\mathrm{P}}(y) = r_{\mathrm{G}}(y)$ for all $y \in \mathcal{Y} \setminus \mathcal{Z}$. Denote by θ_t and θ_t^{G} the parameters at time $t \geq 0$ when maximizing via gradient flow (Equation (1)) the expected reward with respect to r_{P} and r_{G}, respectively, starting from the same initial parameters $\theta_0 = \theta_0^{\mathrm{G}}$. For any time $T \geq 0$ and $\epsilon > 0$, if $\pi_{\theta_0}(\mathcal{Z}) \leq \frac{2}{\Delta_{\mathcal{Z}} \exp(10B^2T)} \cdot \epsilon$, with $B := \max_{y \in \mathcal{Y}} \|\phi(y)\|$ and $\Delta_{\mathcal{Z}} := \max_{y \in \mathcal{Z}} |r_{\mathrm{P}}(y) - r_{\mathrm{G}}(y)|$, then for all $t \in [0, T]$:*

$$\mathrm{TV}\left(\pi_{\theta_t}, \pi_{\theta_t^{\mathrm{G}}}\right) \leq \epsilon,$$

where $\mathrm{TV}(\pi_{\theta_t}, \pi_{\theta_t^{\mathrm{G}}}) := \frac{1}{2}\|\pi_{\theta_t} - \pi_{\theta_t^{\mathrm{G}}}\|_1$ is the total variation distance between π_{θ_t} and $\pi_{\theta_t^{\mathrm{G}}}$.

Proposition 2 yields an additional type of benign reward error.

Benign Error II: Incorrect proxy reward for an output with low probability under π_{θ_0}. If an output y is improbable under π_{θ_0}, then assigning it an incorrect proxy reward $r_{\mathrm{P}}(y) \neq r_{\mathrm{G}}(y)$ negligibly affects the outcome of policy gradient.

3.3 Beneficial Errors That Prevent Attraction to Mediocre Outputs

Section 3.2 showed that outputs with low proxy reward or low probability under the initial policy do not attract probability. Where then does probability mass go during policy gradient? It is helpful to first consider the case of orthonormal feature vectors.[2] In this setting, the logit dynamics of an output $y \in \mathcal{Y}$ simplifies to $\frac{d}{dt}\langle \phi(y), \theta_t \rangle = \pi_{\theta_t}(y) A_{\mathrm{P}}(y; \theta_t)$ (Proposition 1). As evident from this equation, the logit of y grows whenever its advantage $A_{\mathrm{P}}(y; \theta_t) = r_{\mathrm{P}}(y) - V_{\mathrm{P}}(\theta_t)$ is positive. Though, since the advantage is scaled by the output probability $\pi_{\theta_t}(y)$, policy gradient can be drawn to outputs with mediocre proxy reward at the expense of outputs with higher proxy reward, if the former are more likely under the policy. This phenomenon was empirically observed in Mei et al. (2020a). However, the analysis of Mei et al. (2020a) only showed that if the policy already assigns high probability to a suboptimal output, escaping that region takes a long time, leaving open the question of when and to what extent the policy becomes highly concentrated on such outputs to begin with.

[2] For conciseness, we defer treatment of non-orthonormal feature vectors to Appendix C.

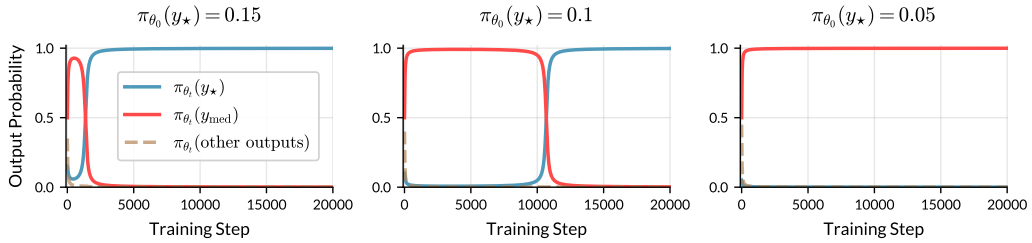

Figure 2: **Attraction to mediocre outputs can impede policy gradient optimization.** We plot the evolution of output probabilities during policy gradient in settings corresponding to Theorem 1: a linear softmax policy with orthonormal output features, trained using exact gradients of the expected ground truth reward. The ground truth reward r_{G} assigns a maximal reward of 1 to $y_\star$, a mediocre reward of 0.8 to y_{med}, and a low reward of -1 to all remaining outputs. In line with Theorem 1, when y_{med} receives mediocre reward, the policy initially concentrates its probability mass on y_{med} and stagnates. The lower $\pi_{\theta_0}(y_\star)$ is (left to right columns), the longer the policy is delayed by an attraction to y_{med}. By contrast, Figure 6 shows that when y_{med} is assigned low reward, the policy directly increases the probability of $y_\star$. Appendix E provides experiments demonstrating similar trends with sample-based gradients instead of exact gradients. Appendix F includes further implementation details.

We prove that the attraction to mediocre outputs can be severe, stalling optimization for an arbitrarily long time (Theorem 1). As we establish below, this implies that the following reward error type is beneficial for improving the rate at which the ground truth reward increases.

Beneficial Error I: Low proxy reward for an output with mediocre ground truth reward. Let y_{med} be an output with mediocre ground truth reward, *i.e.*, $V_{\mathrm{G}}(\theta_0) < r_{\mathrm{G}}(y_{\mathrm{med}}) < r_{\mathrm{G}}(y_\star)$ where $y_\star := \operatorname{argmax}_{y \in \mathcal{Y}} r_{\mathrm{G}}(y)$ and $V_{\mathrm{G}}(\theta_0) := \mathbb{E}_{y \sim \pi_{\theta_0}}[r_{\mathrm{G}}(y)]$. If the probability of y_{med} under π_{θ_0} is higher than that of $y_\star$, then assigning low proxy reward to y_{med} can accelerate the increase in ground truth reward by steering the policy away from it.

Formally, we compare the optimization trajectory across two settings. In the first setting, we maximize directly a ground truth reward function r_{G} that assigns mediocre reward to an output y_{med}. In the second setting, we maximize a proxy reward function r_{P} that is identical to r_{G}, except that it assigns low reward to y_{med}. Theorem 1 establishes that, if $\pi_{\theta_0}(y_\star)$ is sufficiently small relative to $\pi_{\theta_0}(y_{\mathrm{med}})$, the time required to achieve high ground truth reward when maximizing r_{G} directly can be arbitrarily larger than the time required when maximizing r_{P}. This gap in optimization time arises because, when maximizing r_{G}, the policy initially concentrates its probability mass on y_{med} and stalls in its vicinity. By contrast, when maximizing r_{P}, the policy directly shifts its probability mass towards $y_\star$.

Theorem 1 provides an informal version of the result; see Appendix C for the detailed theorem statement. Figure 2 corroborates Theorem 1 by demonstrating empirically the effect of mediocre outputs: the smaller $\pi_{\theta_0}(y_\star)$ is relative to $\pi_{\theta_0}(y_{\mathrm{med}})$, the longer policy gradient is delayed by an attraction to y_{med} (when y_{med} is assigned mediocre proxy reward).

Theorem 1 (Informal version of Theorem 2)**.** *Assume that the feature vectors are orthonormal and that $\mathcal{Y}$ consists of an output with maximal ground truth reward $y_\star = \operatorname{argmax}_{y \in \mathcal{Y}} r_{\mathrm{G}}(y)$, a mediocre output y_{med} with ground truth reward satisfying $V_{\mathrm{G}}(\theta_0) = \mathbb{E}_{y \sim \pi_{\theta_0}}[r_{\mathrm{G}}(y)] < r_{\mathrm{G}}(y_{\mathrm{med}}) < r_{\mathrm{G}}(y_\star)$, and outputs $\mathcal{Y}_{\mathrm{bad}} \subset \mathcal{Y}$ that have a ground truth reward lower than $V_{\mathrm{G}}(\theta_0)$. Let r_{P} be a proxy reward function that is identical to r_{G}, except that it assigns y_{med} minimal reward. Then, for any $\epsilon \in \big(0, r_{\mathrm{G}}(y_\star) - r_{\mathrm{G}}(y_{\mathrm{med}})\big)$, the time it takes to achieve ϵ-optimal expected ground truth reward when maximizing r_{G} directly can be arbitrarily larger than the time it takes when maximizing r_{P}. The smaller $\pi_{\theta_0}(y_\star)$ is, the larger this gap in optimization time becomes.*

3.4 Harmful Errors Beyond Reward Hacking

To conclude the categorization we revisit the most well-known type of reward error (*cf.* Amodei et al. (2016); Skalse et al. (2022); Fluri et al. (2025)).

Harmful Error I: High proxy reward for an output with low ground truth reward. Let y_{bad} be an output with low ground truth reward and denote $y_\star := \operatorname{argmax}_{y \in \mathcal{Y}} r_{\mathrm{G}}(y)$. If $r_{\mathrm{P}}(y_{\mathrm{bad}}) > r_{\mathrm{P}}(y_\star)$, then maximizing r_{P} can lead to reward hacking.

While assigning high proxy reward to an output y_{bad} with low ground truth reward is obviously undesirable, the attraction to mediocre outputs characterized in Section 3.3 reveals a more subtle harmful reward error. Even if the proxy reward assigned to y_{bad} is not high enough to cause reward hacking, it may still impede optimization. This follows as a corollary of Theorem 1; see Corollary 1 in Appendix C for the formal statement.

Harmful Error II: Mediocre proxy reward for an output with low ground truth reward. Let y_{bad} be an output with low ground truth reward whose probability under π_{θ_0} is higher than the probability of $y_\star := \operatorname{argmax}_{y \in \mathcal{Y}} r_{\text{G}}(y)$. Then, assigning a mediocre proxy reward to y_{bad}, satisfying $V_{\text{P}}(\theta_0) < r_{\text{P}}(y_{\text{bad}}) < r_{\text{P}}(y_\star)$ where $V_{\text{P}}(\theta_0) := \mathbb{E}_{y \sim \pi_{\theta_0}}[r_{\text{P}}(y)]$, can slow down the increase in ground truth reward by causing the policy to stall in the vicinity of y_{bad}.

4 Harm-Aware Accuracy for Reward Model Evaluation

To demonstrate the potential utility of our theory, we consider the problem of reward model evaluation for RLHF. Existing benchmarks primarily evaluate reward models via *ranking accuracy* (see Section 2). A limitation of ranking accuracy, uncovered in Section 3, is that it treats all incorrect rankings of outputs as equally harmful, while some deviations from the ground truth reward are benign or even beneficial for policy gradient optimization. Indeed, recent work showed that ranking accuracy often does not correlate well with the performance of a language model after RLHF (Chen et al., 2024; Wen et al., 2025; Razin et al., 2025).

Motivated by this observation, we develop ranking accuracy variants that account for the harmfulness of an incorrect ranking to policy gradient. Compared to standard ranking accuracy, our harm-aware variants usually correlate better with language model performance, across datasets and model families (Llama, OLMo, and Qwen). However, our experiments also show that the correlation can still be weak, highlighting persistent challenges in robustly evaluating reward models. For brevity, we defer some experiments and implementation details to Appendices E and F, respectively.

4.1 Harm-Aware Ranking Accuracy

Let $\mathcal{S}$ be a dataset of preference examples $(x, y^+, y_1^-, \ldots, y_K^-)$, where x is an input and the output y^+ is preferred over $y_1^-, \ldots, y_K^-$ according to a ground truth reward function r_{G}. In practice, we typically do not have access to r_{G}. It is therefore impossible to determine for every incorrect ranking whether it is harmful or not based on the categorization in Section 3. Nevertheless, the categorization allows us to identify certain types of incorrect rankings that are unlikely to be harmful.

For a reward model r_{P}, let $V_{\text{P}}(x; \theta) := \mathbb{E}_{y \sim \pi_\theta(\cdot|x)}[r_{\text{P}}(x, y)]$ be the expected proxy reward achieved by a language model π_θ given an input x. If r_{P} ranks a dispreferred output y_k^- above y^+, but $r_{\text{P}}(x, y_k^-) < V_{\text{P}}(x; \theta)$, then this error may be benign since y_k^- does not attract probability (Benign Error I); it may even be beneficial when y^+ is of mediocre quality (Beneficial Error I). Accordingly, we define the *harm-aware ranking accuracy (HAcc)* by making a simple modification to ranking accuracy: HAcc does not penalize incorrect rankings whenever all dispreferred outputs receive proxy reward below an empirical estimate of $V_{\text{P}}(x; \theta)$, denoted $\bar{V}_{\text{P}}(x; \theta)$. That is, we define:

$$\text{HAcc}(r_{\text{P}}; \mathcal{S}) := \frac{1}{|\mathcal{S}|} \sum\nolimits_{(x, y^+, \{y_k^-\}_{k \in [K]}) \in \mathcal{S}} \mathbb{1}\left[\max_{k \in [K]} r_{\text{P}}(x, y_k^-) < \max\left\{ r_{\text{P}}(x, y^+), \bar{V}_{\text{P}}(x; \theta) \right\} \right].$$

Furthermore, reward errors involving outputs that are improbable under π_θ are also benign (Benign Error II). Thus, in addition to ranking accuracy (Acc) and HAcc, we consider weighted variants (Acc-W and HAcc-W) that multiply the contribution of an example by $\bar{\pi}_\theta(y^+|x) \prod_{k=1}^K \bar{\pi}_\theta(y_k^-|x) / Z(\mathcal{S}; \pi_\theta)$ instead of $1/|\mathcal{S}|$, where $\bar{\pi}_\theta(y|x) := \pi_\theta(y|x)^{1/|y|}$ and $Z(\mathcal{S}; \pi_\theta)$ is the normalization constant ensuring weights sum to one. The usage of $\bar{\pi}_\theta(y|x)$ is equivalent to length normalizing the log probability of y. We apply this transformation for numerical stability and to avoid biasing against long sequences.

4.2 Comparison of Ranking Accuracy Variants

Main setting. We use four language models of different families and types: Llama-3.2-3B-Instruct, Llama-3.2-1B-Instruct (Grattafiori et al., 2024), OLMo-2-1B-SFT (OLMo et al., 2024), and Qwen3-1.7B-Base (Yang et al., 2025). Each language model is trained via the RLOO (Kool et al., 2019;

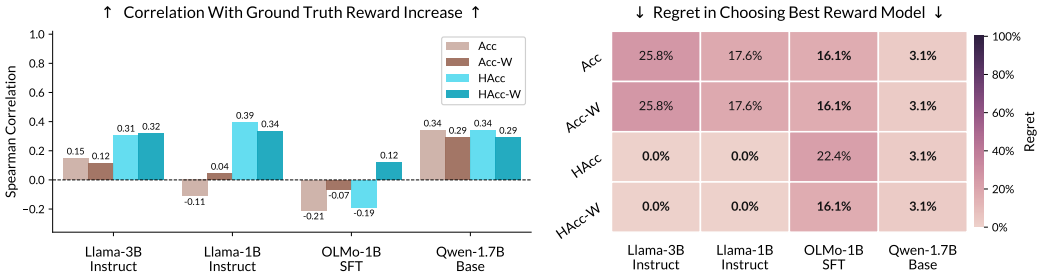

Figure 3: **Harm-aware ranking accuracy variants are more predictive of which reward model leads to better language model performance.** For each language model, we run policy gradient (specifically, RLOO) using 13 different reward models on prompts from the UltraFeedback dataset, and compute (per language and reward model) the mean ground truth reward increase based on three separate runs. Compared to standard ranking accuracy and its weighted counterpart (Acc and Acc-W), the harm-aware variants (HAcc and HAcc-W) correlate better with ground truth reward increase and lead to lower regret in selecting the best reward model (see Section 4.2 for the definition of regret). Yet, the correlation remains below 0.4 and can even be negative, highlighting that robust reward model evaluation remains an open challenge. See Section 4 for further details.

Ahmadian et al., 2024) policy gradient method over prompts from the UltraFeedback dataset (Cui et al., 2024), using 13 reward models chosen to span a diverse range of RewardBench2 scores (Malik et al., 2025) (see Table 1 in Appendix F for the full list). We then compare ranking accuracy with its harm-aware variants (computed for the initial language models) based on their predictiveness of which reward model leads to better language model performance. Following prior work (Gao et al., 2023; Coste et al., 2024; Tang et al., 2024; Balashankar et al., 2024; Chen et al., 2024; Wen et al., 2025), we simulate ground truth rewards with a reward model (ArmoRM (Wang et al., 2024)), excluded from the ones used for training. This ground truth model is used for evaluating both language model performance and reward model accuracy. Unless stated otherwise, rewards and accuracy values are measured on examples from the policy gradient training set.[3]

Additional settings. Appendix E includes experiments showing similar trends: *(i)* when computing reward model accuracy values on RewardBench2 instead of the policy gradient training set; *(ii)* over the WildChat-IF dataset (Lambert et al., 2024) instead of UltraFeedback; *(iii)* using a different ground truth reward model; and *(iv)* when evaluating language model performance by win-rates according to a frontier (GPT) judge model, as opposed to by ground truth reward increase.

Result I: Harm-aware ranking accuracy variants are more predictive of which reward model leads to better language model performance. For each language model and ranking accuracy variant, Figure 3 presents the Spearman correlation between reward model accuracy and ground truth reward increase. We also report the regret incurred when choosing the most accurate reward model, defined as $100(R_{\text{best}} - R_{\text{chosen}})/R_{\text{best}}$, where R_{best} is the largest ground truth reward increase due to policy gradient, across all reward models, and R_{chosen} is the increase under the chosen reward model. Compared to standard ranking accuracy, we find that HAcc typically correlates better with ground truth reward increase and leads to lower regret. On the other hand, we observe that weighting examples by output probabilities is not consistently helpful. This suggests that the contribution of low probability outputs to standard ranking accuracy is not a primary cause for its limited correlation with language model performance (for the considered dataset and models).

Result II: Robust reward model evaluation remains an open challenge. Despite the improvement provided by harm-aware ranking accuracy, the correlation between a ranking accuracy variant and ground truth reward increase remains below 0.4, and in some settings is even negative. As we discuss in Appendix B, this indicates that robustly evaluating reward models may require looking beyond adjustments to ranking accuracy.

5 When Should Partially Correct Outputs Be Rewarded?

Aside from reward model evaluation for RLHF (Section 4), we explore implications of our theory for reward design in settings with verifiable rule-based rewards. In tasks such as instruction following or code generation, an output must satisfy a set of constraints or unit tests to be considered correct (Guo et al., 2025; Olmo et al., 2025). This raises the question of whether partially correct outputs should be rewarded. For example, if an output satisfies one out of two desired constraints, should it receive some partial reward (*e.g.*, 0.5) or should all non-fully correct outputs receive zero reward?

[3] As reported in Appendix E, we found the reward increase on test examples to be nearly identical.

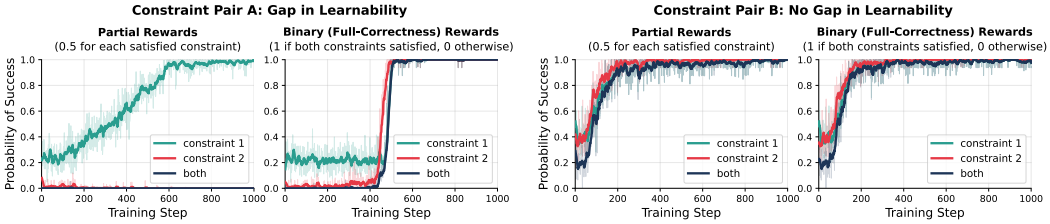

Figure 4: **Rewarding partially correct outputs can impede policy gradient optimization.** We train Qwen3-1.7B using GRPO on two instruction following datasets, where the prompts in each dataset include a pair of constraints from IFBench that an output must satisfy to be considered correct (all prompts within a dataset share the same constraint pair). Plotted are probabilities of satisfying the constraints, averaged over ten training steps for clarity; unsmoothed values are shown in semi-transparent lines. As suggested by our theory (Section 3), when the initial probability of satisfying one constraint is noticeably higher than the probability of satisfying the other, rewarding partial correctness can cause the policy to learn only the easier constraint. In this case, rewarding only fully correct outputs can lead to faster learning of both constraints. However, when both constraints are initially satisfied with similar probability, both reward designs tend to work well. See Section 5 for further details.

Since the goal is to produce fully correct outputs, it is natural to view the ground truth reward function as giving a reward of 1 to fully correct outputs and 0 to all others. The question above then becomes a matter of proxy reward design. One might expect rewarding partially correct outputs to always be beneficial because it provides a richer learning signal. However, our analysis (Section 3) suggests that when the initial policy is noticeably more likely to produce partially correct outputs than fully correct ones, rewarding partial correctness can cause it to stall on these mediocre outputs (Harmful Error II). Thus, as long as there is a non-zero probability of producing fully correct outputs at initialization, in such cases it may be more effective to only reward full correctness. We demonstrate this prospect empirically using instruction following tasks from IFBench (Pyatkin et al., 2025).

Setting. We construct instruction following datasets by taking prompts from UltraFeedback and appending to them two instructions, each specifying a constraint from IFBench (*e.g.*, "the response must start with a verb"). All prompts within a dataset share the same constraints, and the datasets differ in which constraints they contain. We then train Qwen3-1.7B on these datasets using GRPO (Shao et al., 2024) with two proxy reward designs. In the first, we reward partial correctness by giving a reward of 0.5 for each satisfied constraint. In the second, the reward is binary: fully correct outputs receive a reward of 1, and all other outputs receive a reward of 0. See Appendix E for experiments with additional language models and Appendix F for further implementation details.

Results. Figure 4 shows that, if the initial probability of satisfying one constraint noticeably exceeds the probability of satisfying the other, rewarding partial correctness may cause the policy to learn only the easier constraint, while binary (full-correctness) rewards can enable the policy to learn both constraints. If instead the constraints are initially satisfied with similar probability, both reward designs tend to work well. Taken together, these findings yield a practical insight: rewarding partially correct outputs is not universally beneficial or harmful. Choosing suitable proxy rewards in verifiable settings requires considering both the task structure and language model capabilities.

6 Conclusion

Imperfect proxy rewards are widely used for language model post-training, and reinforcement learning more broadly. Yet, the understanding of how discrepancies between the proxy and ground truth rewards affect the learning process remains limited. Such reward errors are conventionally viewed as harmful due to the risk of reward hacking. In this work, however, we proved that aside from being harmful, reward errors can also be benign or even beneficial for policy gradient optimization. Specifically, our theoretical analysis categorizes prominent reward error types into harmful, benign, or beneficial, and characterizes the mechanisms behind their distinct effects on optimization.

We then demonstrated that this categorization yields practical insights. For RLHF, it motivated harm-aware reward model evaluation metrics that are more predictive of language model performance than standard ranking accuracy. For environments with verifiable rewards, it revealed that rewarding partially correct outputs can impede learning rather than help it. A key theme underlying these results is that the effectiveness of proxy rewards depends heavily on their interplay with the initial policy and learning algorithm. As elaborated in Appendix B, we hope that our work will inspire further research on proxy reward evaluation and design that accounts for this interplay.

AI USE STATEMENT

We used generative AI tools to assist with code implementation, checking the mathematical derivations, which were done manually, and light editing of the manuscript. Furthermore, in one set of experiments (Figure 11), we followed the common practice of using a judge model to evaluate language model performance by win-rates. We did not use generative AI tools for any other purposes, thoroughly reviewed all AI-assisted work, and take full responsibility for the content of this paper.

REPRODUCIBILITY STATEMENT

Section 3.1 specifies the setting of our theoretical analysis, including the policy parameterization and optimization procedure. The statements of Propositions 1 and 2 are self-contained and rely only on this setting. For our main theoretical result, Section 3.3 gives an informal statement (Theorem 1), while Appendix C provides the detailed formal statement (Theorem 2) together with corollaries and extensions. Complete proofs for all theoretical results can be found in Appendix D. All information required to reproduce our experimental results is provided in Sections 4 and 5 and Appendix F. We will publicly release our code implementation, which is attached as supplementary material.

REFERENCES

Alekh Agarwal, Sham M Kakade, Jason D Lee, and Gaurav Mahajan. On the theory of policy gradient methods: Optimality, approximation, and distribution shift. *The Journal of Machine Learning Research*, 22 (1):4431–4506, 2021.

Arash Ahmadian, Chris Cremer, Matthias Gallé, Marzieh Fadaee, Julia Kreutzer, Olivier Pietquin, Ahmet Üstün, and Sara Hooker. Back to basics: Revisiting reinforce style optimization for learning from human feedback in llms. *arXiv preprint arXiv:2402.14740*, 2024.

Zafarali Ahmed, Nicolas Le Roux, Mohammad Norouzi, and Dale Schuurmans. Understanding the impact of entropy on policy optimization. In *International conference on machine learning*, pp. 151–160. PMLR, 2019.

Dario Amodei, Chris Olah, Jacob Steinhardt, Paul Christiano, John Schulman, and Dan Mané. Concrete problems in ai safety. *arXiv preprint arXiv:1606.06565*, 2016.

John Asmuth, Michael L Littman, and Robert Zinkov. Potential-based shaping in model-based reinforcement learning. In *AAAI*, pp. 604–609, 2008.

Ananth Balashankar, Ziteng Sun, Jonathan Berant, Jacob Eisenstein, Michael Collins, Adrian Hutter, Jong Lee, Chirag Nagpal, Flavien Prost, Aradhana Sinha, et al. Infalign: Inference-aware language model alignment. *arXiv preprint arXiv:2412.19792*, 2024.

Christopher Berner, Greg Brockman, Brooke Chan, Vicki Cheung, Przemysław Dębiak, Christy Dennison, David Farhi, Quirin Fischer, Shariq Hashme, Chris Hesse, et al. Dota 2 with large scale deep reinforcement learning. *arXiv preprint arXiv:1912.06680*, 2019.

Alberto Bietti, Joan Bruna, and Loucas Pillaud-Vivien. On learning gaussian multi-index models with gradient flow part i: General properties and two-timescale learning. *Communications on Pure and Applied Mathematics*, 78(12):2354–2435, 2025.

Yanjun Chen, Dawei Zhu, Yirong Sun, Xinghao Chen, Wei Zhang, and Xiaoyu Shen. The accuracy paradox in rlhf: When better reward models don't yield better language models. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, 2024.

Ching-An Cheng, Andrey Kolobov, and Adith Swaminathan. Heuristic-guided reinforcement learning. *Advances in Neural Information Processing Systems*, 34:13550–13563, 2021.

Hao-Tien Lewis Chiang, Aleksandra Faust, Marek Fiser, and Anthony Francis. Learning navigation behaviors end-to-end with autorl. *IEEE Robotics and Automation Letters*, 4(2):2007–2014, 2019.

Hung-Hsu Chou, Johannes Maly, and Holger Rauhut. More is less: inducing sparsity via overparameterization. *Information and Inference: A Journal of the IMA*, 12(3):1437–1460, 2023.

Thomas Coste, Usman Anwar, Robert Kirk, and David Krueger. Reward model ensembles help mitigate overoptimization. In *International Conference on Learning Representations*, 2024.

Ganqu Cui, Lifan Yuan, Ning Ding, Guanming Yao, Wei Zhu, Yuan Ni, Guotong Xie, Zhiyuan Liu, and Maosong Sun. Ultrafeedback: Boosting language models with high-quality feedback. In *International Conference on Machine Learning*, 2024.

Falcon Dai and Matthew Walter. Maximum expected hitting cost of a markov decision process and informativeness of rewards. *Advances in Neural Information Processing Systems*, 2019.

Rati Devidze, Parameswaran Kamalaruban, and Adish Singla. Exploration-guided reward shaping for reinforcement learning under sparse rewards. *Advances in Neural Information Processing Systems*, 2022.

Lukas Fluri, Leon Lang, Alessandro Abate, Patrick Forré, David Krueger, and Joar Skalse. The perils of optimizing learned reward functions: Low training error does not guarantee low regret. In *International Conference on Machine Learning*, 2025.

Dylan J Foster, Zakaria Mhammedi, and Dhruv Rohatgi. Is a good foundation necessary for efficient reinforcement learning? the computational role of the base model in exploration. In *Proceedings of Thirty Eighth Conference on Learning Theory*, 2025.

Evan Frick, Tianle Li, Connor Chen, Wei-Lin Chiang, Anastasios N Angelopoulos, Jiantao Jiao, Banghua Zhu, Joseph E Gonzalez, and Ion Stoica. How to evaluate reward models for rlhf. In *International Conference on Learning Representations*, 2025.

Leo Gao, John Schulman, and Jacob Hilton. Scaling laws for reward model overoptimization. In *International Conference on Machine Learning*, pp. 10835–10866. PMLR, 2023.

Shivam Garg, Samuele Tosatto, Yangchen Pan, Martha White, and Rupam Mahmood. An alternate policy gradient estimator for softmax policies. In *Proceedings of The 25th International Conference on Artificial Intelligence and Statistics*, pp. 6630–6689, 2022.

Adam Gleave, Michael Dennis, Shane Legg, Stuart Russell, and Jan Leike. Quantifying differences in reward functions. In *International Conference on Learning Representations*, 2021.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, et al. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*, 2024.

Marek Grzes. Reward shaping in episodic reinforcement learning. In *Proceedings of the 16th Conference on Autonomous Agents and MultiAgent Systems*, 2017.

Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, Xiao Bi, et al. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *arXiv preprint arXiv:2501.12948*, 2025.

Abhishek Gupta, Aldo Pacchiano, Yuexiang Zhai, Sham Kakade, and Sergey Levine. Unpacking reward shaping: Understanding the benefits of reward engineering on sample complexity. *Advances in Neural Information Processing Systems*, 2022.

Daniel Hennes, Dustin Morrill, Shayegan Omidshafiei, Remi Munos, Julien Perolat, Marc Lanctot, Audrunas Gruslys, Jean-Baptiste Lespiau, Paavo Parmas, Edgar Duenez-Guzman, et al. Neural replicator dynamics. *arXiv preprint arXiv:1906.00190*, 2019.

Audrey Huang, Adam Block, Qinghua Liu, Nan Jiang, Akshay Krishnamurthy, and Dylan J Foster. Is best-of-n the best of them? coverage, scaling, and optimality in inference-time alignment. In *International Conference on Machine Learning*, 2025a.

Yu Huang, Zixin Wen, Yuejie Chi, Yuting Wei, Aarti Singh, Yingbin Liang, and Yuxin Chen. On the learning dynamics of rlvr at the edge of competence. *arXiv preprint arXiv:2602.14872*, 2026.

Yuzhen Huang, Weihao Zeng, Xingshan Zeng, Qi Zhu, and Junxian He. Pitfalls of rule-and model-based verifiers–a case study on mathematical reasoning. *arXiv preprint arXiv:2505.22203*, 2025b.

Jacek Karwowski, Oliver Hayman, Xingjian Bai, Klaus Kiendlhofer, Charlie Griffin, and Joar Skalse. Goodhart's law in reinforcement learning. In *International Conference on Learning Representations*, 2024.

Sara Klein, Simon Weissmann, and Leif Döring. Beyond stationarity: Convergence analysis of stochastic softmax policy gradient methods. In *International Conference on Learning Representations*, 2024.

Wouter Kool, Herke van Hoof, and Max Welling. Buy 4 reinforce samples, get a baseline for free! *Deep Reinforcement Learning Meets Structured Prediction ICLR Workshop*, 2019.

Cassidy Laidlaw, Shivam Singhal, and Anca Dragan. Correlated proxies: A new definition and improved mitigation for reward hacking. In *International Conference on Learning Representations*, 2025.

Nathan Lambert, Jacob Morrison, Valentina Pyatkin, Shengyi Huang, Hamish Ivison, Faeze Brahman, Lester James V Miranda, Alisa Liu, Nouha Dziri, Shane Lyu, et al. Tulu 3: Pushing frontiers in open language model post-training. *arXiv preprint arXiv:2411.15124*, 2024.

Nathan Lambert, Valentina Pyatkin, Jacob Morrison, LJ Miranda, Bill Yuchen Lin, Khyathi Chandu, Nouha Dziri, Sachin Kumar, Tom Zick, Yejin Choi, et al. Rewardbench: Evaluating reward models for language modeling. In *Findings of the Association for Computational Linguistics: NAACL*, 2025.

Adam Laud and Gerald DeJong. The influence of reward on the speed of reinforcement learning: An analysis of shaping. In *Proceedings of the 20th International Conference on Machine Learning (ICML)*, pp. 440–447, 2003.

Gen Li, Yuting Wei, Yuejie Chi, Yuantao Gu, and Yuxin Chen. Softmax policy gradient methods can take exponential time to converge. In *Conference on Learning Theory*. PMLR, 2021.

Xuechen Li, Tianyi Zhang, Yann Dubois, Rohan Taori, Ishaan Gulrajani, Carlos Guestrin, Percy Liang, and Tatsunori B. Hashimoto. Alpacaeval: An automatic evaluator of instruction-following models. https://github.com/tatsu-lab/alpaca_eval, 5 2023.

Max Qiushi Lin, Jincheng Mei, Matin Aghaei, Michael Lu, Bo Dai, Alekh Agarwal, Dale Schuurmans, Csaba Szepesvari, and Sharan Vaswani. Rethinking the global convergence of softmax policy gradient with linear function approximation. *arXiv preprint arXiv:2505.03155*, 2025.

Chris Yuhao Liu, Liang Zeng, Yuzhen Xiao, Jujie He, Jiacai Liu, Chaojie Wang, Rui Yan, Wei Shen, Fuxiang Zhang, Jiacheng Xu, Yang Liu, and Yahui Zhou. Skywork-reward-v2: Scaling preference data curation via human-ai synergy. *arXiv preprint arXiv:2507.01352*, 2025a.

Jiacai Liu, Wenye Li, and Ke Wei. Elementary analysis of policy gradient methods. *arXiv preprint arXiv:2404.03372*, 2024.

Jiate Liu, Yiqin Zhu, Kaiwen Xiao, QIANG FU, Xiao Han, Yang Wei, and Deheng Ye. RLTF: Reinforcement learning from unit test feedback. *Transactions on Machine Learning Research*, 2023.

Yantao Liu, Zijun Yao, Rui Min, Yixin Cao, Lei Hou, and Juanzi Li. Rm-bench: Benchmarking reward models of language models with subtlety and style. In *International Conference on Learning Representations*, 2025b.

Saumya Malik, Valentina Pyatkin, Sander Land, Jacob Morrison, Noah A Smith, Hannaneh Hajishirzi, and Nathan Lambert. Rewardbench 2: Advancing reward model evaluation. *arXiv preprint arXiv:2506.01937*, 2025.

Md Rayhanul Masud, Azmine Toushik Wasi, Salman Rahman, and Md Rizwan Parvez. Reward engineering for reinforcement learning in software tasks. *arXiv preprint arXiv:2601.19100*, 2026.

Maja J Mataric. Reward functions for accelerated learning. In *Machine learning proceedings 1994*, pp. 181–189. Elsevier, 1994.

Jincheng Mei, Chenjun Xiao, Bo Dai, Lihong Li, Csaba Szepesvári, and Dale Schuurmans. Escaping the gravitational pull of softmax. *Advances in Neural Information Processing Systems*, 2020a.

Jincheng Mei, Chenjun Xiao, Csaba Szepesvari, and Dale Schuurmans. On the global convergence rates of softmax policy gradient methods. In *International Conference on Machine Learning*, pp. 6820–6829. PMLR, 2020b.

Jincheng Mei, Yue Gao, Bo Dai, Csaba Szepesvari, and Dale Schuurmans. Leveraging non-uniformity in first-order non-convex optimization. In *International Conference on Machine Learning*, pp. 7555–7564. PMLR, 2021.

Jincheng Mei, Bo Dai, Alekh Agarwal, Mohammad Ghavamzadeh, Csaba Szepesvári, and Dale Schuurmans. Ordering-based conditions for global convergence of policy gradient methods. *Advances in Neural Information Processing Systems*, 2023a.

Jincheng Mei, Zixin Zhong, Bo Dai, Alekh Agarwal, Csaba Szepesvari, and Dale Schuurmans. Stochastic gradient succeeds for bandits. In *International Conference on Machine Learning*, pp. 24325–24360. PMLR, 2023b.

Andrew Y Ng, Daishi Harada, and Stuart Russell. Policy invariance under reward transformations: Theory and application to reward shaping. In *ICML*, volume 99, pp. 278–287, 1999.

Team OLMo, Pete Walsh, Luca Soldaini, Dirk Groeneveld, Kyle Lo, Shane Arora, Akshita Bhagia, Yuling Gu, Shengyi Huang, Matt Jordan, et al. 2 olmo 2 furious. *arXiv preprint arXiv:2501.00656*, 2024.

Team Olmo, Allyson Ettinger, Amanda Bertsch, Bailey Kuehl, David Graham, David Heineman, Dirk Groeneveld, Faeze Brahman, Finbarr Timbers, Hamish Ivison, et al. Olmo 3. *arXiv preprint arXiv:2512.13961*, 2025.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. Training language models to follow instructions with human feedback. *Advances in Neural Information Processing Systems*, 2022.

Richard Yuanzhe Pang, Vishakh Padmakumar, Thibault Sellam, Ankur Parikh, and He He. Reward gaming in conditional text generation. In *The 61st Annual Meeting Of The Association For Computational Linguistics*, 2023.

Adam Paszke, Sam Gross, Soumith Chintala, Gregory Chanan, Edward Yang, Zachary DeVito, Zeming Lin, Alban Desmaison, Luca Antiga, and Adam Lerer. Automatic differentiation in pytorch. In *NIPS-W*, 2017.

Xue Bin Peng, Ze Ma, Pieter Abbeel, Sergey Levine, and Angjoo Kanazawa. Amp: Adversarial motion priors for stylized physics-based character control. *ACM Transactions on Graphics (ToG)*, 40(4):1–20, 2021.

Valentina Pyatkin, Saumya Malik, Victoria Graf, Hamish Ivison, Shengyi Huang, Pradeep Dasigi, Nathan Lambert, and Hannaneh Hajishirzi. Generalizing verifiable instruction following. In *Advances in Neural Information Processing Systems*, 2025.

Yuval Ran-Milo, Yotam Alexander, Shahar Mendel, and Nadav Cohen. Outcome-based rl provably leads transformers to reason, but only with the right data. *arXiv preprint arXiv:2601.15158*, 2026.

Jette Randløv and Preben Alstrøm. Learning to drive a bicycle using reinforcement learning and shaping. In *International Conference on Machine Learning*, 1999.

Noam Razin, Hattie Zhou, Omid Saremi, Vimal Thilak, Arwen Bradley, Preetum Nakkiran, Joshua M. Susskind, and Etai Littwin. Vanishing gradients in reinforcement finetuning of language models. In *International Conference on Learning Representations*, 2024.

Noam Razin, Zixuan Wang, Hubert Strauss, Stanley Wei, Jason D Lee, and Sanjeev Arora. What makes a reward model a good teacher? an optimization perspective. In *Advances in Neural Information Processing Systems*, 2025.

Noam Razin, Yong Lin, Jiarui Yao, and Sanjeev Arora. Why is your language model a poor implicit reward model? In *International Conference on Learning Representations*, 2026.

Andrew M Saxe, James L McClelland, and Surya Ganguli. Exact solutions to the nonlinear dynamics of learning in deep linear neural networks. In *International Conference on Learning Representations*, 2014.

Tom Schaul, Diana Borsa, Joseph Modayil, and Razvan Pascanu. Ray interference: a source of plateaus in deep reinforcement learning. *arXiv preprint arXiv:1904.11455*, 2019.

John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. Proximal policy optimization algorithms. *arXiv preprint arXiv:1707.06347*, 2017.

Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang, Mingchuan Zhang, YK Li, Y Wu, et al. Deepseekmath: Pushing the limits of mathematical reasoning in open language models. *arXiv preprint arXiv:2402.03300*, 2024.

Satinder Singh, Richard L Lewis, Andrew G Barto, and Jonathan Sorg. Intrinsically motivated reinforcement learning: An evolutionary perspective. *IEEE Transactions on Autonomous Mental Development*, 2(2):70–82, 2010.

Joar Skalse, Nikolaus Howe, Dmitrii Krasheninnikov, and David Krueger. Defining and characterizing reward gaming. *Advances in Neural Information Processing Systems*, 2022.

Joar Skalse, Lucy Farnik, Sumeet Ramesh Motwani, Erik Jenner, Adam Gleave, and Alessandro Abate. Starc: A general framework for quantifying differences between reward functions. In *International Conference on Learning Representations*, 2024.

Jonathan Sorg, Richard L Lewis, and Satinder Singh. Reward design via online gradient ascent. *Advances in Neural Information Processing Systems*, 23, 2010.

Nisan Stiennon, Long Ouyang, Jeffrey Wu, Daniel Ziegler, Ryan Lowe, Chelsea Voss, Alec Radford, Dario Amodei, and Paul F Christiano. Learning to summarize with human feedback. In *Advances in Neural Information Processing Systems*, 2020.

Yiyou Sun, Yuhan Cao, Pohao Huang, Haoyue Bai, Hannaneh Hajishirzi, Nouha Dziri, and Dawn Song. Rl grokking recipe: How does rl unlock and transfer new algorithms in llms? In *International Conference on Learning Representations*, 2026.

Gokul Swamy, Sanjiban Choudhury, Wen Sun, Zhiwei Steven Wu, and J Andrew Bagnell. All roads lead to likelihood: The value of reinforcement learning in fine-tuning. In *International Conference on Learning Representations*, 2026.

Yunhao Tang, Daniel Zhaohan Guo, Zeyu Zheng, Daniele Calandriello, Yuan Cao, Eugene Tarassov, Rémi Munos, Bernardo Ávila Pires, Michal Valko, Yong Cheng, et al. Understanding the performance gap between online and offline alignment algorithms. *arXiv preprint arXiv:2405.08448*, 2024.

Leitian Tao, Ilia Kulikov, Swarnadeep Saha, Tianlu Wang, Jing Xu, Sharon Li, Jason E Weston, and Ping Yu. Hybrid reinforcement: When reward is sparse, it's better to be dense. *arXiv preprint arXiv:2510.07242*, 2025.

Gal Vardi and Ohad Shamir. Implicit regularization in relu networks with the square loss. In *Conference on Learning Theory*, pp. 4224–4258. PMLR, 2021.

Haoxiang Wang, Wei Xiong, Tengyang Xie, Han Zhao, and Tong Zhang. Interpretable preferences via multi-objective reward modeling and mixture-of-experts. *arXiv preprint arXiv:2406.12845*, 2024.

Xueru Wen, Jie Lou, Yaojie Lu, Hongyu Lin, Xing Yu, Xinyu Lu, Ben He, Xianpei Han, Debing Zhang, and Le Sun. Rethinking reward model evaluation: Are we barking up the wrong tree? In *International Conference on Learning Representations*, 2025.

Ronald J Williams. Simple statistical gradient-following algorithms for connectionist reinforcement learning. *Machine learning*, 8(3):229–256, 1992.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, et al. Huggingface's transformers: State-of-the-art natural language processing. *arXiv preprint arXiv:1910.03771*, 2019.

Blake Woodworth, Suriya Gunasekar, Jason D Lee, Edward Moroshko, Pedro Savarese, Itay Golan, Daniel Soudry, and Nathan Srebro. Kernel and rich regimes in overparametrized models. In *Conference on Learning Theory (COLT)*, pp. 3635–3673, 2020.

Blake Wulfe, Ashwin Balakrishna, Logan Ellis, Jean Mercat, Rowan McAllister, and Adrien Gaidon. Dynamics-aware comparison of learned reward functions. In *International Conference on Learning Representations*, 2022.

Shusheng Xu, Wei Fu, Jiaxuan Gao, Wenjie Ye, Weilin Liu, Zhiyu Mei, Guangju Wang, Chao Yu, and Yi Wu. Is dpo superior to ppo for llm alignment? a comprehensive study. *arXiv preprint arXiv:2404.10719*, 2024.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, et al. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*, 2025.

Qiying Yu, Zheng Zhang, Ruofei Zhu, Yufeng Yuan, Xiaochen Zuo, Yu Yue, Weinan Dai, Tiantian Fan, Gaohong Liu, Lingjun Liu, et al. Dapo: An open-source llm reinforcement learning system at scale. *arXiv preprint arXiv:2503.14476*, 2025.

Zeyu Zheng, Junhyuk Oh, and Satinder Singh. On learning intrinsic rewards for policy gradient methods. *Advances in neural information processing systems*, 31, 2018.

Enyu Zhou, Guodong Zheng, Binghai Wang, Zhiheng Xi, Shihan Dou, Rong Bao, Wei Shen, Limao Xiong, Jessica Fan, Yurong Mou, et al. Rmb: Comprehensively benchmarking reward models in llm alignment. In *International Conference on Learning Representations*, 2025.

Daniel M Ziegler, Nisan Stiennon, Jeffrey Wu, Tom B Brown, Alec Radford, Dario Amodei, Paul Christiano, and Geoffrey Irving. Fine-tuning language models from human preferences. *arXiv preprint arXiv:1909.08593*, 2019.

A Related Work

Reinforcement learning with proxy rewards. Manually handcrafted and automatically learned proxy rewards have a long history across reinforcement learning applications (*e.g.*, Mataric (1994); Randløv & Alstrøm (1999); Singh et al. (2010); Sorg et al. (2010); Zheng et al. (2018); Berner et al. (2019); Chiang et al. (2019); Peng et al. (2021); Devidze et al. (2022)), far too broad to fully cover here. In the context of language models, which are the focus of this work, proxy rewards are widely used both for alignment with human preferences (*i.e.*, RLHF; (Ouyang et al., 2022; Grattafiori et al., 2024; Razin et al., 2026; Swamy et al., 2026)) and for verifiable tasks such as mathematical problem solving, code generation, and instruction following (Lambert et al., 2024; Guo et al., 2025; Olmo et al., 2025; Huang et al., 2025b; Tao et al., 2025; Pyatkin et al., 2025; Masud et al., 2026). Yet, despite the ubiquity of imperfect proxy rewards in practice, there is limited theoretical understanding of what makes a good proxy reward function. We take a step towards addressing this gap.

What makes a good proxy reward function? This question has roots in the reward shaping literature (Ng et al., 1999). There, prior theoretical work characterized reward transformations that preserve the order of policies (Ng et al., 1999; Asmuth et al., 2008; Grzes, 2017; Skalse et al., 2024) and demonstrated that reward shaping can improve the efficiency of stylized learning algorithms or value-based methods in long-horizon environments (Laud & DeJong, 2003; Dai & Walter, 2019; Cheng et al., 2021; Devidze et al., 2022; Gupta et al., 2022). Furthermore, Gleave et al. (2021); Wulfe et al. (2022); Skalse et al. (2024) derived measures for comparing two reward functions (*e.g.*, a proxy and ground truth reward function) that account for said reward transformations, which leave policy ordering unchanged. Similarly to our work, these lines of research can be interpreted as identifying cases where deviating from a ground truth reward function is in some sense benign or beneficial. However, they disregard the impact of such deviations on policy gradient optimization. As Section 3 proves, reward errors that modify policy ordering can still be benign or beneficial. Conversely, reward errors that preserve policy ordering may be harmful and slow down optimization.

Perhaps most closely related to our work is Razin et al. (2025). While reward models for RLHF are mainly evaluated via ranking accuracy (*cf.* Section 2.2), Razin et al. (2025) established that other aspects can determine how good a reward model is. In particular, they showed that when the reward model induces low reward variance, policy gradient suffers from a flat objective landscape that hinders optimization. Our technical approach resembles that of Razin et al. (2025), which also analyzes the rate of ground truth reward increase under proxy reward functions. Though, unlike Razin et al. (2025), we consider how different types of reward errors influence optimization beyond their effect on reward variance (see Remark 1 in Appendix C for further discussion on this point). This allows us to develop harm-aware ranking accuracy variants and provide insights for reward design in settings with verifiable rewards.

Theoretical analyses of policy gradient optimization. For softmax policies, *i.e.*, policies that produce a distribution over outputs via the softmax function, the reward maximization objective is non-concave even under tabular or linear parameterizations (Agarwal et al., 2021). Characterizing the complex optimization dynamics of policy gradient methods in such settings has therefore remained an active area of research. Existing analyses showed that optimization can fail due to the gradient vanishing in long-horizon environments with sparse rewards (Agarwal et al., 2021; Li et al., 2021), when the policy is nearly deterministic (Ahmed et al., 2019; Hennes et al., 2019; Schaul et al., 2019; Mei et al., 2020a;b; Agarwal et al., 2021; Garg et al., 2022), or, more generally, when the reward variance is low (Razin et al., 2024; 2025). Particularly relevant to our work, Mei et al. (2020a) observed that policy gradient can be attracted to mediocre outputs at the expense of higher-quality ones, but left open the question of when and to what extent this phenomenon can impede optimization. Theorem 1 shows that the attraction to mediocre outputs can be severe, stalling optimization for an arbitrarily long time. Our results also complement existing convergence guarantees (Mei et al., 2020b; 2021; 2023b; Klein et al., 2024; Liu et al., 2024; Lin et al., 2025). These guarantees depend on $\inf_{t \geq 0} \pi_{\theta_t}(y_\star)$—the infimal probability assigned to the optimal output during training. We identify that the attraction to mediocre outputs can drive this quantity arbitrarily low.

B Limitations and Future Work

Theoretical analysis: beyond bandit environments. Our theory (Section 3) focuses on bandit environments since, in language model applications, rewards are often available only for complete

outputs. Extending the analysis to longer-horizon settings, corresponding to multi-turn dialogue or environments with process rewards, may reveal additional ways in which reward errors shape learning. It would also be valuable to understand how reward errors interact across inputs, especially given recent evidence that learnability via policy gradient can depend on the data composition (Ran-Milo et al., 2026; Huang et al., 2026).

Towards robust reward model evaluation. Although the harm-aware ranking accuracy variants proposed in Section 4 improve upon standard ranking accuracy, our results suggest that even these metrics capture only part of what makes an effective reward model. We outline two potential barriers towards robust reward model evaluation, which future work can attempt to address. First, output rankings provide only coarse information about the ground truth reward function. This seems to place an inherent ceiling on metrics that solely rely on such data. Second, reward model evaluation benchmarks often lack coverage of relevant inputs and outputs. This limitation is reflected both in recent efforts to revisit existing benchmarks (Zhou et al., 2025; Wen et al., 2025; Malik et al., 2025) and in our experiments, where evaluating reward models on RewardBench2 is markedly less predictive of language model performance than evaluating them on examples drawn from the policy gradient training set (Figure 9 in Appendix E).

Reward design in verifiable settings. We identified a drawback of rewarding partially correct outputs: it can hinder optimization by causing the policy to stall on such suboptimal outputs (Section 5). However, there are cases where rewarding partial correctness is beneficial. Namely, Sun et al. (2026) showed that when the initial policy produces fully correct outputs with near-zero probability, by first using partial rewards and then switching to binary (full-correctness) rewards, one can achieve better performance than using binary rewards from the start. We therefore view analyzing adaptive proxy reward schemes that vary across inputs and training steps as a promising direction for future work.

C Detailed Theorem Statements and Influence of Feature Similarity on Attraction to Mediocre Outputs

In this appendix, we present Theorem 2—the detailed statement of Theorem 1—which establishes the existence of beneficial reward errors. In particular, Theorem 2 shows that mediocre outputs can attract probability and stall policy gradient optimization for an arbitrarily long time. Thus, assigning low proxy reward to outputs with mediocre ground truth reward can be beneficial for accelerating the increase in ground truth reward. As a corollary (Corollary 1), we also obtain that assigning mediocre proxy reward to outputs with low ground truth reward can impede optimization, even if it does not induce reward hacking. This formalizes Harmful Error II from Section 3.4. Both Theorem 2 and Corollary 1 consider orthonormal output feature vectors. In Appendix C.1, we relax this assumption and analyze how the similarity between output features can exacerbate or attenuate the attraction to mediocre outputs. We refer the reader to Section 3.1 for the technical setting of the analysis.

Our aim is to characterize the phenomenon where mediocre outputs attract probability at the expense of higher-quality outputs. Thus, we consider a setting in which $\mathcal{Y}$ contains an output y_{med} with ground truth reward lower than $r_{\text{G}}(y_\star)$, where $y_\star = \operatorname{argmax}_{y\in\mathcal{Y}} r_{\text{G}}(y)$ is the optimal output, but higher than the ground truth reward of all remaining outputs $\mathcal{Y}_{\text{bad}} \subset \mathcal{Y}$. For simplicity, we assume all outputs in $\mathcal{Y}_{\text{bad}}$ have the same ground truth reward. Our analysis extends straightforwardly to the case where outputs in $\mathcal{Y}_{\text{bad}}$ have different rewards, as long as they are sufficiently lower than $r_{\text{G}}(y_{\text{med}})$.

Assumption 1 (Reward structure)**.** *The set of outputs can be decomposed as* $\mathcal{Y} = \{y_\star, y_{\text{med}}\} \cup \mathcal{Y}_{\text{bad}}$ *such that* $r_{\text{G}}(y_\star) > r_{\text{G}}(y_{\text{med}}) > 0 \geq \max_{y\in\mathcal{Y}_{\text{bad}}} r_{\text{G}}(y)$*. Furthermore, all outputs in* $\mathcal{Y}_{\text{bad}} \subset \mathcal{Y}$ *have the same ground truth reward, i.e.,* $r_{\text{G}}(y) = r_{\text{G}}(y')$ *for all* $y, y' \in \mathcal{Y}_{\text{bad}}$*.*

Suppose that we can maximize the ground truth reward directly (*i.e.*, $r_{\text{P}} = r_{\text{G}}$). In this case, as discussed in Section 3.3, an output y_{med} with mediocre (ground truth and proxy) reward can attract probability if $r_{\text{G}}(y_{\text{med}}) > V_{\text{G}}(\theta_0) = \mathbb{E}_{y\sim\pi_{\theta_0}}[r_{\text{G}}(y)]$ and the probability of y_{med} under π_{θ_0} is higher than the probability of $y_\star$. This condition is formalized by assuming that $\pi_{\theta_0}(y_\star)$ is not too high and that $\pi_{\theta_0}(\mathcal{Y}_{\text{bad}})$ is neither too low (otherwise, $V_{\text{G}}(\theta_0)$ will be above $r_{\text{G}}(y_{\text{med}})$) nor too high (otherwise, $\pi_{\theta_0}(y_{\text{med}})$ may not be sufficiently higher than $\pi_{\theta_0}(y_\star)$). See Lemma 14 in Appendix D for a proof that Assumptions 1 and 2 ensure $r_{\text{G}}(y_{\text{med}}) > V_{\text{G}}(\theta_0)$.

Assumption 2 (Initial policy)**.** *The initial policy* π_{θ_0} *satisfies the following conditions.*

- *The initial probability of $y_\star$ is not too high:*

$$\pi_{\theta_0}(y_\star) < M \cdot \min\left\{0.01, \left(\frac{21}{1100(\Delta_1+\Delta_2)}\right)^{7/3}, \Delta_2^2\right\}^{13/14},$$

  where $\Delta_1 := r_{\mathrm{G}}(y_\star) - r_{\mathrm{G}}(y_{\mathrm{med}})$, $\Delta_2 := r_{\mathrm{G}}(y_{\mathrm{med}}) - r_{\mathrm{G}}(y_{\mathrm{bad}})$ for some $y_{\mathrm{bad}} \in \mathcal{Y}_{\mathrm{bad}}$, and $M := \min\{1, 0.05 \cdot r_{\mathrm{G}}(y_{\mathrm{med}})^{2/7}(\Delta_1+\Delta_2)^{-1}\}$.

- *The initial probability of $\mathcal{Y}_{\mathrm{bad}}$ is neither too low nor too high:*

$$\max\left\{\frac{2\pi_{\theta_0}(y_\star)^{7/13}}{\Delta_2 M^{7/13}}, 0.05 \cdot r_{\mathrm{G}}(y_{\mathrm{med}})\right\} \le \pi_{\theta_0}(\mathcal{Y}_{\mathrm{bad}}) \le 0.1 \cdot r_{\mathrm{G}}(y_{\mathrm{med}}).$$

Under Assumptions 1 and 2, we prove that incorrectly assigning low reward to y_{med} is beneficial for increasing the ground truth reward. Specifically, we compare the optimization trajectory when maximizing r_{G} directly with the trajectory obtained when maximizing a proxy reward function r_{P} that is identical to r_{G}, except that it assigns low reward to the mediocre output y_{med}. Theorem 2 establishes that the time required to achieve high ground truth reward when maximizing r_{G} directly can be arbitrarily larger than when maximizing r_{P}. This gap in optimization time arises because, when maximizing r_{G}, the policy first concentrates its probability mass on y_{med} and stalls in its vicinity. By contrast, when maximizing r_{P}, the policy directly shifts its probability mass towards $y_\star$.

Theorem 2 (Formal version of Theorem 1)**.** *Suppose that the feature vectors $\{\phi(y)\}_{y\in\mathcal{Y}}$ are orthonormal and that the ground truth reward function r_{G} and initial policy parameters θ_0 uphold Assumptions 1 and 2. Furthermore, let r_{P} be a proxy reward function identical to r_{G}, except that it assigns y_{med} a low reward $r_{\mathrm{P}}(y_{\mathrm{med}}) = \min_{y\in\mathcal{Y}_{\mathrm{bad}}} r_{\mathrm{G}}(y)$. For any $\epsilon \in (0, \Delta_1)$, where $\Delta_1 := r_{\mathrm{G}}(y_\star) - r_{\mathrm{G}}(y_{\mathrm{med}})$, the following hold.*

- ***Case I: Maximizing** r_{G}. If gradient flow is used to maximize the expected reward with respect to r_{G} (Equation (1) with V_{G} in place of V_{P}), then the initial time at which $V_{\mathrm{G}}(\theta_t) \ge r_{\mathrm{G}}(y_\star) - \epsilon$, denoted $t_\star$, is lower bounded as follows:*

$$t_\star \ge \frac{2\sqrt{2}M^{14/13}\Delta_1 r_{\mathrm{G}}(y_{\mathrm{med}})(1-e^{-\sqrt{2}(\Delta_1-\epsilon)})}{(\Delta_1+\Delta_2)(1+8\Delta_1)} \cdot \pi_{\theta_0}(y_\star)^{-14/13} = \Omega\Big(\pi_{\theta_0}(y_\star)^{-14/13}\Big),$$

  where $\Delta_2 := r_{\mathrm{G}}(y_{\mathrm{med}}) - r_{\mathrm{G}}(y_{\mathrm{bad}})$ for some $y_{\mathrm{bad}} \in \mathcal{Y}_{\mathrm{bad}}$ and M is as defined in Assumption 2.

- ***Case II: Maximizing** r_{P}. In contrast, denote by $t_\star^{\mathrm{no-med}}$ the initial time at which $V_{\mathrm{G}}(\theta_t) \ge r_{\mathrm{G}}(y_\star) - \epsilon$ when gradient flow is used to maximize the expected reward with respect to r_{P}. Then, $t_\star^{\mathrm{no-med}}$ is upper bounded by a quantity that grows asymptotically slower as $\pi_{\theta_0}(y_\star) \to 0$ compared to the lower bound on $t_\star$ in Case I:*

$$t_\star^{\mathrm{no-med}} \le \frac{(r_{\mathrm{G}}(y_\star)+1)^2}{\epsilon^2(r_{\mathrm{P}}(y_\star) - V_{\mathrm{P}}(\theta_0))} \cdot \pi_{\theta_0}(y_\star)^{-1} = \mathcal{O}\big(\pi_{\theta_0}(y_\star)^{-1}\big),$$

  Furthermore, at any time t greater than or equal to the right-hand side in the inequality above, we have $\pi_{\theta_t}(y_\star) \ge 1 - \epsilon/(\Delta_1+\Delta_2)$.

Thus, for any arbitrarily large $T \ge 0$, if $\pi_{\theta_0}(y_\star)$ is sufficiently small, then $t_\star - t_\star^{\mathrm{no-med}} \ge T$.

Proof sketch (full proof in Appendix D.4). Consider first the case of maximizing r_{G} directly. The logit dynamics of an output y in this case is given by $\frac{d}{dt}\langle\phi(y), \theta_t\rangle = \pi_{\theta_t}(y)A_{\mathrm{G}}(y;\theta_t)$, where $A_{\mathrm{G}}(y;\theta_t) = r_{\mathrm{G}}(y) - V_{\mathrm{G}}(\theta_t)$ is the advantage of y under r_{G} and π_{θ_t} (*cf.* Proposition 1). If $\pi_{\theta_0}(y_{\mathrm{med}})$ is substantially higher than $\pi_{\theta_0}(y_\star)$, the logit of the mediocre output y_{med} initially grows much faster than that of the optimal output $y_\star$, despite y_{med} having a smaller advantage than $y_\star$. This causes the probability of $y_\star$ to decrease. By carefully analyzing the optimization trajectory, we show that $\pi_{\theta_t}(y_\star)$ continues to decrease at least until $\pi_{\theta_t}(y_{\mathrm{med}})$ is close to one, in the sense that $\pi_{\theta_t}(y_{\mathrm{med}}) \ge 1 - \Theta(\pi_{\theta_0}(y_\star)^{14/13})$. We then use the fact that the gradient vanishes whenever the policy is highly concentrated on a single output (*cf.* Mei et al. (2020a); Agarwal et al. (2021); Razin et al. (2024; 2025)) to prove that escaping the vicinity of y_{med} requires at least $\Omega(\pi_{\theta_0}(y_\star)^{-14/13})$ time, and so $t_\star$ is at least of that order. By contrast, when maximizing r_{P}, the advantage of all

outputs except $y_\star$ is negative throughout training since r_P assigns them a low reward. Thus, the policy directly shifts its probability mass towards $y_\star$ and achieves ϵ-optimal ground truth reward at time $t_\star^{\mathrm{no-med}}$, which we prove is at most $\mathcal{O}(\pi_{\theta_0}(y_\star)^{-1})$. This implies that the gap between $t_\star$ and $t_\star^{\mathrm{no-med}}$ becomes arbitrarily large as $\pi_{\theta_0}(y_\star) \to 0$. □

Remark 1. Recently, Razin et al. (2024; 2025) identified that if the reward function r used for training induces low reward variance for the initial policy, *i.e.*, if $\mathrm{Var}_{y\sim\pi_{\theta_0}}[r(y)] \approx 0$, then policy gradient suffers from slow optimization (see Theorem 1 in Razin et al. (2025)). It is therefore natural to wonder whether r_P facilitates faster optimization than r_G in Theorem 2 due to inducing a higher initial reward variance. Interestingly, the opposite is true: the initial reward variance is actually higher under r_G. The slow optimization under r_G is not caused by low initial reward variance, but rather by a collapse in reward variance during training, as the policy becomes concentrated on the mediocre output y_med. Although r_P initially induces lower reward variance than r_G, it guides the policy directly towards $y_\star$, avoiding this intermediate variance collapse.

Formal statement behind Harmful Error II. As a corollary of Theorem 2, by swapping the proxy and ground truth reward functions, we obtain that assigning mediocre proxy reward to an output with low ground truth reward can be harmful due to impeding optimization, even if it does not induce reward hacking. That is, Corollary 1 formalizes Harmful Error II from Section 3.4.

Corollary 1. *Consider the setting and assumptions of Theorem 2, with Assumptions 1 and 2 interpreted with r_P in place of r_G. Furthermore, let r_G be identical to r_P, except that it assigns y_med a low reward $r_\mathrm{G}(y_\mathrm{med}) = \min_{y\in\mathcal{Y}_\mathrm{bad}} r_\mathrm{P}(y)$. For any $\epsilon \in (0, \Delta_1)$, where $\Delta_1 := r_\mathrm{P}(y_\star) - r_\mathrm{P}(y_\mathrm{med})$, the following hold.*

- ***Case I: Maximizing** r_P. If gradient flow is used to maximize the expected reward with respect to r_P (Equation (1)), then the initial time at which $V_\mathrm{G}(\theta_t) \geq r_\mathrm{G}(y_\star) - \epsilon$, denoted $t_\star$, is lower bounded as follows:*

$$t_\star \geq \frac{2\sqrt{2}M^{14/13}\Delta_1 r_\mathrm{P}(y_\mathrm{med})(1 - e^{-\sqrt{2}(\Delta_1-\epsilon)})}{(\Delta_1+\Delta_2)(1+8\Delta_1)} \cdot \pi_{\theta_0}(y_\star)^{-14/13} = \Omega\Big(\pi_{\theta_0}(y_\star)^{-14/13}\Big),$$

  where $\Delta_2 := r_\mathrm{P}(y_\mathrm{med}) - r_\mathrm{P}(y_\mathrm{bad})$ for some $y_\mathrm{bad} \in \mathcal{Y}_\mathrm{bad}$ and M is as defined in Assumption 2.

- ***Case II: Maximizing** r_G. In contrast, denote by $t_\star^{\mathrm{no-med}}$ the initial time at which $V_\mathrm{G}(\theta_t) \geq r_\mathrm{G}(y_\star) - \epsilon$ when gradient flow is used to maximize the expected reward with respect to r_G (Equation (1) with V_G in place of V_P). Then, $t_\star^{\mathrm{no-med}}$ is upper bounded by a quantity that grows asymptotically slower as $\pi_{\theta_0}(y_\star) \to 0$ compared to the lower bound on $t_\star$ in Case I:*

$$t_\star^{\mathrm{no-med}} \leq \frac{(r_\mathrm{G}(y_\star)+1)^2}{\epsilon^2(r_\mathrm{G}(y_\star) - V_\mathrm{G}(\theta_0))} \cdot \pi_{\theta_0}(y_\star)^{-1} = \mathcal{O}\big(\pi_{\theta_0}(y_\star)^{-1}\big),$$

  Furthermore, at any time t greater than or equal to the right-hand side in the inequality above, we have $\pi_{\theta_t}(y_\star) \geq 1 - \epsilon/(\Delta_1+\Delta_2)$.

Thus, for any arbitrarily large $T \geq 0$, if $\pi_{\theta_0}(y_\star)$ is sufficiently small, then $t_\star - t_\star^{\mathrm{no-med}} \geq T$.

Proof. The result follows by applying Theorem 2 while swapping the roles of r_G and r_P. For Case I (*i.e.*, when maximizing r_P), Theorem 2 yields a lower bound on the time until $V_\mathrm{P}(\theta_t) \geq r_\mathrm{P}(y_\star) - \epsilon$. Since $r_\mathrm{G}(y_\star) = r_\mathrm{P}(y_\star)$ and $V_\mathrm{G}(\theta) \leq V_\mathrm{P}(\theta)$ for all θ (recall, r_G is identical to r_P, except that it assigns a lower reward to y_med), this immediately implies that the same lower bound holds for $t_\star$—the time until $V_\mathrm{G}(\theta_t) \geq r_\mathrm{G}(y_\star) - \epsilon$. For Case II (*i.e.*, when maximizing r_G), Theorem 2 implies that $\pi_{\theta_t}(y_\star) \geq 1 - \epsilon/(\Delta_1+\Delta_2)$ at any time t greater than or equal to the desired upper bound on $t_\star^{\mathrm{no-med}}$. Hence, because ground truth rewards lie within the interval $[r_\mathrm{G}(y_\mathrm{bad}), r_\mathrm{G}(y_\star)]$ and $\Delta_1 + \Delta_2 = r_\mathrm{G}(y_\star) - r_\mathrm{G}(y_\mathrm{bad})$, where $y_\mathrm{bad} \in \mathcal{Y}_\mathrm{bad}$, at any such time $V_\mathrm{G}(\theta_t) \geq r_\mathrm{G}(y_\star) - \epsilon$. This establishes the upper bound on $t_\star^{\mathrm{no-med}}$. □

C.1 Influence of Feature Similarity

We now analyze how the similarity (or dissimilarity) of output features affects the extent to which mediocre outputs impede optimization. Namely, under conditions analogous to those of Theorem 2,

we characterize how the conclusion of Theorem 2 changes when $\phi(y_\star)$ and $\phi(y_{\text{med}})$ are not orthogonal (recall that $y_\star$ is the output with maximal ground truth reward and y_{med} is an output with mediocre ground truth reward). The analysis reveals that when $\langle \phi(y_\star), \phi(y_{\text{med}}) \rangle < 0$, the mediocre output y_{med} delays optimization even further and can prevent the policy from ever reaching the optimal output $y_\star$. Conversely, when $\langle \phi(y_\star), \phi(y_{\text{med}}) \rangle > 0$, attraction to y_{med} does not necessarily impede optimization and, if $\phi(y_\star)$ and $\phi(y_{\text{med}})$ are extremely similar, assigning y_{med} a mediocre proxy reward may be required for enabling convergence to $y_\star$. Appendices C.1.1 and C.1.2 deliver the formal results for the cases of negative and positive inner products, respectively.

C.1.1 Negative Inner Product Between $\phi(y_\star)$ and $\phi(y_{\text{med}})$

We consider the ground truth reward function r_{G} described in Assumption 1 (*i.e.*, the ground truth from the orthonormal features setting considered in Theorem 2) and make the following Assumption 3 on the initial policy probabilities. Assumption 3 is identical to Assumption 2 up to constants that depend on the norms of feature vectors. Similar to Assumption 2, Assumption 3 guarantees that $r_{\text{G}}(y_{\text{med}}) > V_{\text{G}}(\theta_0)$ (see Lemma 15) and that $\pi_{\theta_0}(y_{\text{med}})$ is substantially higher than $\pi_{\theta_0}(y_\star)$.

Assumption 3 (Initial policy)**.** *The initial policy π_{θ_0} satisfies the following conditions.*

- *The initial probability of $y_\star$ is not too high:*

$$\pi_{\theta_0}(y_\star) < \min\left\{ M' \cdot \min\left\{ 0.01, \left(\frac{21\|\phi(y_{\text{med}})\|^2}{1100 B^2 (\Delta_1 + \Delta_2)} \right)^{7/3} \right\}^{13/14}, \frac{\|\phi(y_{\text{med}})\|}{\|\phi(y_\star)\|} \pi_{\theta_0}(y_{\text{med}}) \right\},$$

  where $\Delta_1 := r_{\text{G}}(y_\star) - r_{\text{G}}(y_{\text{med}})$, $\Delta_2 := r_{\text{G}}(y_{\text{med}}) - r_{\text{G}}(y_{\text{bad}})$ for some $y_{\text{bad}} \in \mathcal{Y}_{\text{bad}}$, $B := \max_{y \in \mathcal{Y}} \|\phi(y)\|$, and $M' := \min\{1, \|\phi(y_{\text{med}})\|^2 r_{\text{G}}(y_{\text{med}})^{2/7} (40 B^2 (\Delta_1 + \Delta_2))^{-1}\}$.

- *The initial probability of $\mathcal{Y}_{\text{bad}}$ is neither too low nor too high:*

$$\max\left\{ \frac{2\pi_{\theta_0}(y_\star)^{7/13}}{\Delta_2 M'^{7/13}}, 0.05 \cdot r_{\text{G}}(y_{\text{med}}) \right\} \leq \pi_{\theta_0}(\mathcal{Y}_{\text{bad}}) \leq 0.1 \cdot r_{\text{G}}(y_{\text{med}}).$$

We focus here on the case of $\langle \phi(y_\star), \phi(y_{\text{med}}) \rangle < 0$. To isolate the effect of $\langle \phi(y_\star), \phi(y_{\text{med}}) \rangle$ on the optimization dynamics from interactions between other feature vectors, we require that, for any output in $\mathcal{Y}_{\text{bad}}$, its feature vector is orthogonal to the feature vectors of all other outputs. We also make the mild assumption that $\|\phi(y_{\text{med}})\| \geq \|\phi(y_{\text{bad}})\|$ for all $y_{\text{bad}} \in \mathcal{Y}_{\text{bad}}$.

Assumption 4 (Output feature vectors)**.** *The feature vectors of $y_\star$ and y_{med} have negative inner product, i.e., $\langle \phi(y_\star), \phi(y_{\text{med}}) \rangle < 0$, and $\langle \phi(y_{\text{bad}}), \phi(y) \rangle = 0$ for all $y_{\text{bad}} \in \mathcal{Y}_{\text{bad}}, y \in \mathcal{Y} \setminus \{y_{\text{bad}}\}$. Furthermore, $\|\phi(y_{\text{med}})\| \geq \|\phi(y_{\text{bad}})\|$ for all $y_{\text{bad}} \in \mathcal{Y}_{\text{bad}}$.*

If we maximize r_{G} directly, y_{med} attracts probability at the expense of $y_\star$ even more strongly than in the case of orthonormal feature vectors. The reason for this stronger attraction is that, as long as the advantage $A_{\text{G}}(y_{\text{med}}; \theta_t) := r_{\text{G}}(y_{\text{med}}) - V_{\text{G}}(\theta_t)$ is positive, $\langle \phi(y_\star), \phi(y_{\text{med}}) \rangle$ pushes the logit of $y_\star$ downwards (see Proposition 1). As Theorem 3 shows, this yields a larger separation between $t_\star$—the time needed to achieve high ground truth reward when maximizing r_{G} directly—and the corresponding time $t_\star^{\text{no-med}}$ when maximizing a proxy reward function r_{P} that assigns low reward to y_{med}. In particular, the lower bound in Theorem 3 on $t_\star$ grows asymptotically faster as $\pi_{\theta_0}(y_\star) \to 0$ compared to the lower bound in Theorem 2. On the other hand, the upper bound on $t_\star^{\text{no-med}}$ can remain roughly the same since $A_{\text{P}}(y_{\text{med}}; \theta_t) < 0$ throughout optimization when y_{med} is assigned a low proxy reward, in which case $\langle \phi(y_\star), \phi(y_{\text{med}}) \rangle$ only pushes the logit of $y_\star$ further upwards.

Theorem 3. *Suppose that the ground truth reward function r_{G}, initial policy parameters θ_0, and output feature vectors $\{\phi(y)\}_{y \in \mathcal{Y}}$ uphold Assumptions 1, 3, and 4. Let r_{P} be a proxy reward function identical to r_{G}, except that it assigns y_{med} a low reward $r_{\text{P}}(y_{\text{med}}) = \min_{y \in \mathcal{Y}_{\text{bad}}} r_{\text{G}}(y)$. For any $\epsilon \in (0, \Delta_1)$, where $\Delta_1 := r_{\text{G}}(y_\star) - r_{\text{G}}(y_{\text{med}})$, the following hold.*

- ***Case I: Maximizing** r_{G}. If gradient flow is used to maximize the expected reward with respect to r_{G} (Equation (1) with V_{G} in place of V_{P}), then the initial time at which $V_{\text{G}}(\theta_t) \geq r_{\text{G}}(y_\star) - \epsilon$, denoted $t_\star$, is lower bounded as follows:*

$$t_\star \geq \frac{r_{\text{G}}(y_{\text{med}}) M'^{14/13+\alpha} (1 - e^{-2(\Delta_1 - \epsilon)})}{4 B^2 (\Delta_1 + \Delta_2) K} \cdot \pi_{\theta_0}(y_\star)^{-14/13-\alpha} = \Omega\left(\pi_{\theta_0}(y_\star)^{-14/13-\alpha} \right),$$

where $\Delta_2 := r_{\mathrm{G}}(y_{\mathrm{med}}) - r_{\mathrm{G}}(y_{\mathrm{bad}})$ *for some* $y_{\mathrm{bad}} \in \mathcal{Y}_{\mathrm{bad}}$. *The constants* B *and* M' *are as defined in Assumption 3,* $\alpha := -\langle \phi(y_\star), \phi(y_{\mathrm{med}}) \rangle / \left(26(\|\phi(y_{\mathrm{med}})\|^2 - 0.5\langle \phi(y_\star), \phi(y_{\mathrm{med}}) \rangle)\right) \in (0, 1/13)$, *and*

$$K := \left(\frac{26}{r_{\mathrm{G}}(y_{\mathrm{med}})^2}\right)^{13\alpha/7} \left(1 + \frac{\|\phi(y_{\mathrm{med}})\|^2 \pi_{\theta_0}(y_{\mathrm{med}})^2}{8(\|\phi(y_\star)\|^2 - \langle \phi(y_\star), \phi(y_{\mathrm{med}}) \rangle) A_{\mathrm{G}}(y_\star; \theta_0)}\right).$$

- ***Case II: Maximizing*** r_{P}. *In contrast, denote by* $t_\star^{\mathrm{no-med}}$ *the initial time at which* $V_{\mathrm{G}}(\theta_t) \geq r_{\mathrm{G}}(y_\star) - \epsilon$ *when gradient flow is used to maximize the expected reward with respect to* r_{P}. *Then,* $t_\star^{\mathrm{no-med}}$ *is upper bounded by a quantity that grows asymptotically slower as* $\pi_{\theta_0}(y_\star) \to 0$ *compared to the lower bound on* $t_\star$ *in Case I:*

$$t_\star^{\mathrm{no-med}} \leq \frac{(r_{\mathrm{G}}(y_\star) + 1)^2}{\epsilon^2 \left(r_{\mathrm{P}}(y_\star) - V_{\mathrm{P}}(\theta_0)\right) \|\phi(y_\star)\|^2} \cdot \pi_{\theta_0}(y_\star)^{-1} = \mathcal{O}\left(\pi_{\theta_0}(y_\star)^{-1}\right),$$

Thus, for any arbitrarily large $T > 0$, *if* $\pi_{\theta_0}(y_\star)$ *is sufficiently small, then* $t_\star - t_\star^{\mathrm{no-med}} \geq T$.

Proof sketch (full proof in Appendix D.5). The proof follows a line similar to that of Theorem 2. The key difference is that, due to the negative inner product between $\phi(y_\star)$ and $\phi(y_{\mathrm{med}})$, the logit dynamics of $y_\star$ now contains an additional term that depends on $\langle \phi(y_\star), \phi(y_{\mathrm{med}}) \rangle$. Specifically, when maximizing r_{G} directly (Case I), the logit dynamics of $y_\star$ is given by (*cf.* Proposition 1):

$$\frac{d}{dt}\langle \phi(y_\star), \theta_t \rangle = \pi_{\theta_t}(y_\star) A_{\mathrm{G}}(y_\star; \theta_t) \cdot \|\phi(y_\star)\|^2 + \pi_{\theta_t}(y_{\mathrm{med}}) A_{\mathrm{G}}(y_{\mathrm{med}}; \theta_t) \cdot \langle \phi(y_\star), \phi(y_{\mathrm{med}}) \rangle .$$

Thus, as long as $A_{\mathrm{G}}(y_{\mathrm{med}}; \theta_t) > 0$, the second term on the right-hand side is negative and pushes the probability of $y_\star$ further downwards compared to the case of orthonormal feature vectors. We show that this implies that the policy will become even more concentrated on y_{med}, and so will take a longer time to escape the vicinity of y_{med}. On the other hand, when maximizing r_{P} (Case II), the advantage $A_{\mathrm{P}}(y_{\mathrm{med}}; \theta_t)$ is negative throughout optimization since r_{P} assigns y_{med} a low reward. Hence, the contribution of y_{med} to the logit dynamics of $y_\star$ can only push the logit of $y_\star$ further upwards compared to the orthonormal feature vectors case. This straightforwardly allows deriving an upper bound on $t_\star^{\mathrm{no-med}}$ that is roughly the same as the one in Theorem 2. □

When $\langle \phi(y_\star), \phi(y_{\mathrm{med}}) \rangle < 0$ **the policy may fail to ever assign high probability to** $y_\star$**.** Theorem 3 indicates that the lower the inner product between $\phi(y_\star)$ and $\phi(y_{\mathrm{med}})$ is, the longer the policy stalls in the vicinity of the mediocre output y_{med} when maximizing r_{G}. One may wonder if the policy always escapes the attraction of y_{med} and reaches $y_\star$ as optimization progresses. We show that the answer is no. For certain initial policies and output feature vectors, Proposition 3 proves that the policy never achieves a ground truth reward above $r_{\mathrm{G}}(y_{\mathrm{med}})$, and thus never assigns high probability to $y_\star$. Notably, this failure occurs despite the existence of policy parameters that assign $y_\star$ a probability arbitrarily close to one. Proposition 3 is obtained by adapting Proposition 13 from Lin et al. (2025) to our setting.

Proposition 3. *Suppose that the ground truth reward function* r_{G} *upholds Assumption 1. Furthermore, let* $y_{\mathrm{bad}} \in \mathcal{Y}_{\mathrm{bad}}$, *define* $\Delta_1 := r_{\mathrm{G}}(y_\star) - r_{\mathrm{G}}(y_{\mathrm{med}})$ *and* $\Delta_2 := r_{\mathrm{G}}(y_{\mathrm{med}}) - r_{\mathrm{G}}(y_{\mathrm{bad}})$, *and assume that the initial policy parameters* θ_0 *and output feature vectors* $\{\phi(y)\}_{y \in \mathcal{Y}}$ *satisfy the following conditions, alongside Assumption 4.*

- $\pi_{\theta_0}(y_{\mathrm{med}}) > \pi_{\theta_0}(y)$ *for all* $y \in \mathcal{Y} \setminus \{y_{\mathrm{med}}\}$;
- $\pi_{\theta_0}(y_\star) \leq \frac{\Delta_2 \pi_{\theta_0}(\mathcal{Y}_{\mathrm{bad}})}{\Delta_1}$;
- *and* $\theta_0 = C \cdot (\phi(y_{\mathrm{bad}}) - \phi(y_\star))$ *for some* $C > \max\left\{0, -\ln(\zeta)/\|\phi(y_{\mathrm{bad}}) - \phi(y_\star)\|^2\right\}$, *where*

$$\zeta := \frac{\ln \pi_{\theta_0}(y_{\mathrm{med}}) - \ln \pi_{\theta_0}(y_{\mathrm{bad}})}{\ln \pi_{\theta_0}(y_{\mathrm{med}}) - \ln \pi_{\theta_0}(y_\star)} \cdot \frac{V_{\mathrm{G}}(\theta_0) - r_{\mathrm{G}}(y_{\mathrm{bad}})}{r_{\mathrm{G}}(y_\star) - V_{\mathrm{G}}(\theta_0)} > 0 .$$

If gradient flow is used to maximize the expected reward with respect to r_{G} *(Equation (1) with* V_{G} *in place of* V_{P}*), then* $V_{\mathrm{G}}(\theta_t) < r_{\mathrm{G}}(y_{\mathrm{med}})$ *for all* $t \geq 0$.

Proof sketch (full proof in Appendix D.6). At initialization, the policy favors y_{bad} over $y_\star$ in the sense that $\pi_{\theta_0}(y_{\text{bad}}) > \pi_{\theta_0}(y_\star)$ and $\pi_{\theta_0}(y_\star)/\pi_{\theta_0}(y_{\text{bad}}) < \zeta$. The proof follows by showing that the suppressive effect of y_{med} on $y_\star$, which stems from y_{med} having a positive initial advantage and the inner product of $\phi(y_\star)$ and $\phi(y_{\text{med}})$ being negative, is strong enough to ensure that the policy maintains this relationship between y_{bad} and $y_\star$ throughout optimization. In other words, the policy never assigns substantially more probability to $y_\star$ than to y_{bad}. This in turn implies that the expected ground truth reward remains below $r_{\text{G}}(y_{\text{med}})$ for all $t \geq 0$. □

Remark 2. Consider the policy parameters $\theta(\beta) = \beta \cdot \phi(y_\star)$, for $\beta \in \mathbb{R}$. Under the conditions of Proposition 3 on $\{\phi(y)\}_{y\in\mathcal{Y}}$, taking $\beta \to \infty$ implies $\pi_{\theta(\beta)}(y_\star) \to 1$ since $\langle \phi(y_\star), \phi(y_{\text{med}})\rangle < 0$ and $\langle \phi(y_\star), \phi(y)\rangle = 0$ for all $y \in \mathcal{Y}_{\text{bad}}$. Thus, as mentioned in the text preceding Proposition 3, this result shows that even though there exist policy parameters that assign $y_\star$ a probability arbitrarily close to one, policy gradient fails to reach them.

Remark 3. For completeness, we provide an example of output feature vectors satisfying Assumption 4 for which the conditions in Proposition 3 hold. Consider for simplicity a case with only three outputs, $y_\star$, y_{med}, and y_{bad}. Let

$$\phi(y_\star) = \mathbf{e}_1 \quad , \quad \phi(y_{\text{med}}) = -\rho \cdot \mathbf{e}_1 + \sqrt{1-\rho^2} \cdot \mathbf{e}_2 \quad , \quad \phi(y_{\text{bad}}) = \lambda \cdot \mathbf{e}_3 \,,$$

for $0 < \lambda^2 < \rho < 1$ and orthonormal vectors $\mathbf{e}_1, \mathbf{e}_2, \mathbf{e}_3 \in \mathbb{R}^D$. In this case, the inner products between the feature vectors are given by:

$$\langle \phi(y_\star), \phi(y_{\text{med}})\rangle = -\rho < 0 \quad , \quad \langle \phi(y_{\text{bad}}), \phi(y_{\text{med}})\rangle = \langle \phi(y_{\text{bad}}), \phi(y_\star)\rangle = 0 \,.$$

Thus, taking $\theta_0 = C \cdot (\phi(y_{\text{bad}}) - \phi(y_\star))$ for some $C > 0$, the initial logits assigned to $y_\star$, y_{med}, and y_{bad} by θ_0 are:

$$\langle \phi(y_\star), \theta_0\rangle = -C \quad , \quad \langle \phi(y_{\text{med}}), \theta_0\rangle = C\rho \quad , \quad \langle \phi(y_{\text{bad}}), \theta_0\rangle = C\lambda^2 \,.$$

This implies that $\pi_{\theta_0}(y_{\text{med}}) > \pi_{\theta_0}(y_{\text{bad}}) > \pi_{\theta_0}(y_\star)$. In particular, y_{med} has the highest initial probability, as required by the first condition in Proposition 3. For the second and third conditions, notice that:

$$\zeta := \frac{\ln \pi_{\theta_0}(y_{\text{med}}) - \ln \pi_{\theta_0}(y_{\text{bad}})}{\ln \pi_{\theta_0}(y_{\text{med}}) - \ln \pi_{\theta_0}(y_\star)} \cdot \frac{V_{\text{G}}(\theta_0) - r_{\text{G}}(y_{\text{bad}})}{r_{\text{G}}(y_\star) - V_{\text{G}}(\theta_0)} = \frac{\rho - \lambda^2}{\rho + 1} \cdot \frac{V_{\text{G}}(\theta_0) - r_{\text{G}}(y_{\text{bad}})}{r_{\text{G}}(y_\star) - V_{\text{G}}(\theta_0)} > 0$$

and

$$\frac{\pi_{\theta_0}(y_\star)}{\pi_{\theta_0}(y_{\text{bad}})} = \exp\big(-C(1+\lambda^2)\big) \,.$$

Hence, for any $C > \max\big\{0, -\ln(\zeta)/\|\phi(y_{\text{bad}}) - \phi(y_\star)\|^2, \ln(\Delta_1/\Delta_2)\big\}$ we get that:

$$\pi_{\theta_0}(y_\star) \leq \frac{\Delta_2 \pi_{\theta_0}(y_{\text{bad}})}{\Delta_1} \,,$$

meaning all conditions in Proposition 3 are satisfied.

C.1.2 Positive Inner Product Between $\phi(y_\star)$ and $\phi(y_{\text{med}})$

Moving to the case where $\langle \phi(y_\star), \phi(y_{\text{med}})\rangle > 0$, we again consider the ground truth reward function described in Assumption 1 and impose the same assumptions on the initial policy as in the negative inner product case (*i.e.*, Assumption 3). Furthermore, we make the following assumption on the structure of output features, analogous to Assumption 4.

Assumption 5 (Output feature vectors)**.** *The feature vectors of $y_\star$ and y_{med} have positive inner product, i.e., $\langle \phi(y_\star), \phi(y_{\text{med}})\rangle > 0$, and $\langle \phi(y_{\text{bad}}), \phi(y)\rangle = 0$ for all $y_{\text{bad}} \in \mathcal{Y}_{\text{bad}}, y \in \mathcal{Y} \setminus \{y_{\text{bad}}\}$.*

Suppose that we maximize r_{G} directly. In contrast to the case of $\langle \phi(y_\star), \phi(y_{\text{med}})\rangle < 0$, when the advantage $A_{\text{G}}(y_{\text{med}}; \theta_t) = r_{\text{G}}(y_{\text{med}}) - V_{\text{G}}(\theta_t)$ is positive, the mediocre output y_{med} contributes a positive term $\pi_{\theta_t}(y_{\text{med}}) A_{\text{G}}(y_{\text{med}}; \theta_t) \cdot \langle \phi(y_\star), \phi(y_{\text{med}})\rangle$ to the logit dynamics of $y_\star$ (see Proposition 1). Intuitively, this implies that the attraction to y_{med} need not impede optimization: even if probability mass initially moves towards y_{med}, this movement can simultaneously facilitate the growth of $\pi_{\theta_t}(y_\star)$. However, once $V_{\text{G}}(\theta_t)$ exceeds $r_{\text{G}}(y_{\text{med}})$, the advantage of y_{med} becomes negative. As a result, the contribution of y_{med} to the logit of $y_\star$ reverses sign and starts pushing it downwards. This creates an intricate dynamics, where to enable the growth of $\pi_{\theta_t}(y_\star)$ towards one,

the contribution of $y_\star$ to its own logit dynamics, $\pi_{\theta_t}(y_\star)A_{\mathrm{G}}(y_\star;\theta_t)\cdot\|\phi(y_\star)\|^2$, needs to overcome the negative contribution of y_{med}.

To enable a clean characterization, we focus on a regime where $\|\phi(y_\star)\|^2 > \langle\phi(y_\star),\phi(y_{\mathrm{med}})\rangle$ and $\langle\phi(y_\star),\phi(y_{\mathrm{med}})\rangle > \|\phi(y_{\mathrm{med}})\|^2$, leaving treatment of other regimes to future work. As established in Theorem 4, under this condition the mediocre output y_{med} does not harm optimization. Specifically, the time needed to achieve high ground truth reward when maximizing r_{G} directly has the same asymptotic dependence on $\pi_{\theta_0}(y_\star)$ as the upper bounds in Theorems 2 and 3, which apply to cases where y_{med} is assigned low proxy reward. Furthermore, if $\phi(y_\star)$ and $\phi(y_{\mathrm{med}})$ are highly similar, in the sense that their inner product is above a certain threshold, then assigning y_{med} a mediocre reward can be necessary for optimization to succeed. In such cases, we prove that if one maximizes a proxy reward function r_{P} that assigns low reward to y_{med}, the policy fails to achieve near-optimal ground truth reward, and thus never assigns high probability to $y_\star$.

Remark 4. Theorem 4 implies that, when the feature vectors of $y_\star$ and y_{med} are highly similar, the type of reward error caused by assigning y_{med} a low proxy reward changes from beneficial to harmful. By the same reasoning, in the case where y_{med} has low ground truth reward, assigning it a mediocre proxy reward is beneficial rather than harmful.

Theorem 4. *Suppose that the ground truth reward function r_{G}, initial policy parameters θ_0, and output feature vectors $\{\phi(y)\}_{y\in\mathcal{Y}}$ uphold Assumptions 1, 3, and 5. Furthermore, assume that $\|\phi(y_{\mathrm{med}})\|^2 < \langle\phi(y_\star),\phi(y_{\mathrm{med}})\rangle < \|\phi(y_\star)\|^2 < 100\|\phi(y_{\mathrm{med}})\|^2$, and let r_{P} be a proxy reward function identical to r_{G}, except that it assigns y_{med} a low reward $r_{\mathrm{P}}(y_{\mathrm{med}}) \le \min_{y\in\mathcal{Y}_{\mathrm{bad}}} r_{\mathrm{G}}(y)$. For any $\epsilon\in(0,\Delta_1)$, where $\Delta_1 := r_{\mathrm{G}}(y_\star) - r_{\mathrm{G}}(y_{\mathrm{med}})$, the following hold.*

- ***Case I: Maximizing** r_{G}. If gradient flow is used to maximize the expected reward with respect to r_{G} (Equation (1) with V_{G} in place of V_{P}), then the initial time at which $V_{\mathrm{G}}(\theta_t) \ge r_{\mathrm{G}}(y_\star) - \epsilon$, denoted $t_\star$, is upper bounded as follows:*

$$t_\star \le \frac{\left(r_{\mathrm{G}}(y_\star)+1\right)^2}{\epsilon^2\Delta_1\|\phi(y_\star)-\phi(y_{\mathrm{med}})\|^2}\cdot\pi_{\theta_0}(y_\star)^{-1} = \mathcal{O}\big(\pi_{\theta_0}(y_\star)^{-1}\big).$$

- ***Case II: Maximizing** r_{P}. In contrast, if gradient flow is used to maximize the expected reward with respect to r_{P} and $\langle\phi(y_\star),\phi(y_{\mathrm{med}})\rangle$ is sufficiently high, in the sense that*

$$\langle\phi(y_\star),\phi(y_{\mathrm{med}})\rangle > \frac{\pi_{\theta_0}(y_\star)A_{\mathrm{P}}(y_\star;\theta_0)\|\phi(y_\star)\|^2 - \pi_{\theta_0}(y_{\mathrm{med}})A_{\mathrm{P}}(y_{\mathrm{med}};\theta_0)\|\phi(y_{\mathrm{med}})\|^2}{\pi_{\theta_0}(y_\star)A_{\mathrm{P}}(y_\star;\theta_0) - \pi_{\theta_0}(y_{\mathrm{med}})A_{\mathrm{P}}(y_{\mathrm{med}};\theta_0)},$$

then $V_{\mathrm{G}}(\theta_t) < r_{\mathrm{G}}(y_\star) - (r_{\mathrm{G}}(y_\star) - r_{\mathrm{G}}(y_{\mathrm{med}}))\pi_{\theta_0}(y_{\mathrm{med}})$ for all $t \ge 0$.

Proof sketch (full proof in Appendix D.7). When maximizing r_{G} directly (Case I), the logit dynamics of $y_\star$ is given by (*cf.* Proposition 1):

$$\frac{d}{dt}\langle\phi(y_\star),\theta_t\rangle = \pi_{\theta_t}(y_\star)A_{\mathrm{G}}(y_\star;\theta_t)\cdot\|\phi(y_\star)\|^2 + \pi_{\theta_t}(y_{\mathrm{med}})A_{\mathrm{G}}(y_{\mathrm{med}};\theta_t)\cdot\langle\phi(y_\star),\phi(y_{\mathrm{med}})\rangle\ .$$

The advantage of the mediocre output y_{med} is initially positive, *i.e.*, $A_{\mathrm{G}}(y_{\mathrm{med}};\theta_0) > 0$. Thus, since $\langle\phi(y_\star),\phi(y_{\mathrm{med}})\rangle > 0$, whenever y_{med} attracts probability it also pushes the logit of $y_\star$ upwards. In particular, the conditions on the output feature vectors ensure that $\pi_{\theta_t}(y_\star)$ is monotonically non-decreasing, as opposed to the orthonormal feature vectors case (Theorem 2) where $\pi_{\theta_t}(y_\star)$ initially decreases due to the increase in the logit of y_{med}. As a result, the policy does not become highly concentrated on y_{med} and $t_\star$ can be upper bounded by a quantity that has the same asymptotic dependence on $\pi_{\theta_0}(y_\star)$ as the upper bound in Theorem 2, which applies to cases where y_{med} is assigned low proxy reward. On the other hand, when maximizing r_{P} (Case II), the advantage of y_{med} is negative throughout optimization since r_{P} assigns y_{med} a low reward. Hence, y_{med} contributes negatively to the logit dynamics of $y_\star$. We show that this suppressive effect of y_{med} on $y_\star$ prevents the policy from assigning $y_\star$ a higher probability than y_{med}. Specifically, $\pi_{\theta_t}(y_\star)/\pi_{\theta_t}(y_{\mathrm{med}})$ is monotonically non-increasing for all $t \ge 0$, and initially $\pi_{\theta_0}(y_\star) < \pi_{\theta_0}(y_{\mathrm{med}})$. This in turn implies that the expected ground truth reward is bounded away from $r_{\mathrm{G}}(y_\star)$ even as $t\to\infty$. □

C.1.3 Empirical Demonstration

Figure 5 empirically demonstrates how the similarity between output features affects the extent to which mediocre outputs impede optimization. Consistent with our theoretical analysis (Theorems 2,

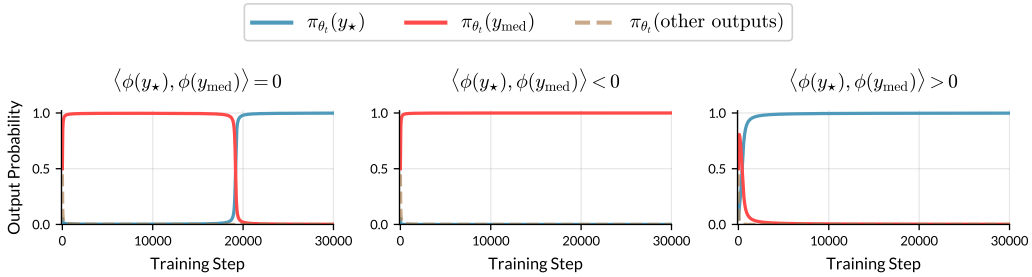

Figure 5: **Feature similarity affects whether mediocre outputs impede policy gradient optimization.** Plotted is the evolution of output probabilities during policy gradient in settings corresponding to Theorems 2, 3, and 4 (left to right plots). We train linear softmax policies using exact gradients of the expected ground truth reward r_{G}, which assigns a maximal reward of 1 to $y_\star$, a mediocre reward of 0.8 to y_{med}, and a low reward of -1 to the remaining outputs. In all experiments, initial policy probabilities and feature vector norms are identical, with $\pi_{\theta_0}(y_\star) = 0.05$, $\pi_{\theta_0}(y_{\text{med}}) = 0.5$, $\|\phi(y_\star)\| = 1.5$, and $\|\phi(y_{\text{med}})\| = 1$. The feature vectors of other outputs (*i.e.*, not $y_\star$ and y_{med}) have unit norm and are mutually orthogonal, as well as orthogonal to $\phi(y_\star)$ and $\phi(y_{\text{med}})$. What varies across the experiments is the value of $\langle\phi(y_\star),\phi(y_{\text{med}})\rangle$, which is equal to 0 in the leftmost plot, $-3/2\sqrt{2}$ in the middle plot, and $3/2\sqrt{2}$ in the rightmost plot. In line with our theory, when $\phi(y_\star)$ and $\phi(y_{\text{med}})$ are orthogonal or have a negative inner product, the policy initially concentrates its probability mass on y_{med} and stagnates, with the attraction to y_{med} being stronger in the negative inner product case. By contrast, when $\phi(y_\star)$ and $\phi(y_{\text{med}})$ have a positive inner product, the attraction to y_{med} is milder and does not obstruct the policy from assigning high probability to $y_\star$. See Appendix F for additional implementation details.

3, and 4), when $\phi(y_\star)$ and $\phi(y_{\text{med}})$ have a negative inner product, attraction to y_{med} is more severe than in the orthogonal case, causing the policy to stall near y_{med} for longer. Conversely, when $\phi(y_\star)$ and $\phi(y_{\text{med}})$ have a positive inner product, the attraction to y_{med} is milder and does not obstruct the policy from assigning high probability to $y_\star$.

D Deferred Proofs

D.1 Notation

We use $A_{\text{G}}(y;\theta) := r_{\text{G}}(y) - V_{\text{G}}(\theta)$ to denote the advantage of an output $y \in \mathcal{Y}$ under the ground truth reward function r_{G} and the policy π_θ. Similarly, we let $A_{\text{P}}(y;\theta) := r_{\text{P}}(y) - V_{\text{P}}(\theta)$ denote the advantage of y under r_{P} and π_θ. Furthermore, $\|\cdot\|$ denotes the Euclidean norm and $\|\cdot\|_1$ the ℓ_1 norm. For matrices, $\|\cdot\|_2$ stands for the spectral norm.

D.2 Proof of Proposition 1

Fix an output $y \in \mathcal{Y}$. By Equation (1) and the chain rule we have that

$$\frac{d}{dt}\langle\phi(y),\theta_t\rangle = \left\langle\phi(y),\tfrac{d}{dt}\theta_t\right\rangle = \langle\phi(y),\nabla V_{\text{P}}(\theta_t)\rangle.$$

By Lemma 10 we know that $\nabla V_{\text{P}}(\theta) = \sum_{z\in\mathcal{Y}} \pi_\theta(z) A_{\text{P}}(z;\theta)\cdot\phi(z)$. Plugging this gradient expression into the logit dynamics above yields:

$$\frac{d}{dt}\langle\phi(y),\theta_t\rangle = \left\langle\phi(y),\sum\nolimits_{z\in\mathcal{Y}} \pi_{\theta_t}(z) A_{\text{P}}(z;\theta_t)\cdot\phi(z)\right\rangle = \sum\nolimits_{z\in\mathcal{Y}} \pi_{\theta_t}(z) A_{\text{P}}(z;\theta_t)\langle\phi(z),\phi(y)\rangle.$$

Separating the term corresponding to $z = y$ from the rest concludes the proof:

$$\frac{d}{dt}\langle\phi(y),\theta_t\rangle = \pi_{\theta_t}(y) A_{\text{P}}(y;\theta_t)\cdot\|\phi(y)\|^2 + \sum\nolimits_{z\in\mathcal{Y}\setminus\{y\}} \pi_{\theta_t}(z) A_{\text{P}}(z;\theta_t)\cdot\langle\phi(z),\phi(y)\rangle.$$

□

D.3 Proof of Proposition 2

If $\Delta_{\mathcal{Z}} = 0$, then $r_{\text{P}} = r_{\text{G}}$, and so trivially $\theta_t = \theta_t^{\text{G}}$ and $\text{TV}\left(\pi_{\theta_t}, \pi_{\theta_t^{\text{G}}}\right) = 0$ for all $t \geq 0$. Hence, in the remainder of the proof we assume $\Delta_{\mathcal{Z}} > 0$. We begin by showing that $\pi_{\theta_T}(\mathcal{Z})$ remains low until time T, regardless of which reward function is used for optimization.

Lemma 1. *Suppose that gradient flow is used to maximize the expected reward with respect to $r : \mathcal{Y} \to [-1, 1]$, starting from θ_0. Then, for all $t \in [0, T]$ it holds that:*

$$\pi_{\theta_t}(\mathcal{Z}) \leq \pi_{\theta_0}(\mathcal{Z}) \cdot \exp\big(4B^2T\big) \leq \frac{2}{\Delta_{\mathcal{Z}} \exp(6B^2T)} \cdot \epsilon\,.$$

Proof. For simplicity of notation, let $f_y(\theta) := \langle \phi(y), \theta \rangle$ denote the logit of $y \in \mathcal{Y}$ under the policy parameters θ. By Proposition 1, the fact that rewards are bounded within $[-1, 1]$, and the Cauchy-Schwarz inequality, we can upper bound $\frac{d}{dt} f_y(\theta_t)$, for any $y \in \mathcal{Y}$ and $t \in [0, T]$, as follows:

$$\begin{aligned}
\left|\frac{d}{dt} f_y(\theta_t)\right| &= \left|\left\langle \phi(y), \tfrac{d}{dt}\theta_t \right\rangle\right| \\
&= \left|\sum\nolimits_{z \in \mathcal{Y}} \pi_{\theta_t}(z)(r(z) - V(\theta_t))\langle \phi(y), \phi(z)\rangle\right| \\
&\leq 2 \sum\nolimits_{z \in \mathcal{Y}} \pi_{\theta_t}(z) |\langle \phi(y), \phi(z)\rangle| \\
&\leq 2 \sum\nolimits_{z \in \mathcal{Y}} \pi_{\theta_t}(z) \|\phi(y)\| \|\phi(z)\| \\
&\leq 2B^2 \sum\nolimits_{z \in \mathcal{Y}} \pi_{\theta_t}(z) \\
&= 2B^2,
\end{aligned} \tag{2}$$

where $V(\theta) := \mathbb{E}_{z \sim \pi_\theta}[r(z)]$ is the expected reward with respect to r and π_θ, and recall that $B = \max_{y \in \mathcal{Y}} \|\phi(y)\|$. This leads to an upper bound on $\frac{d}{dt}\pi_{\theta_t}(y)$:

$$\begin{aligned}
\frac{d}{dt}\pi_{\theta_t}(y) &= \pi_{\theta_t}(y) \frac{d}{dt} \ln \pi_{\theta_t}(y) \\
&= \pi_{\theta_t}(y) \Big(\tfrac{d}{dt} f_y(\theta_t) - \tfrac{d}{dt} \ln\Big(\sum\nolimits_{u \in \mathcal{Y}} \exp(f_u(\theta_t)) \Big) \Big) \\
&= \pi_{\theta_t}(y) \big(\tfrac{d}{dt} f_y(\theta_t) - \mathbb{E}_{u \sim \pi_{\theta_t}} \big[\tfrac{d}{dt} f_u(\theta_t) \big] \big) \\
&\leq \pi_{\theta_t}(y) \big(\big| \tfrac{d}{dt} f_y(\theta_t) \big| + \big| \mathbb{E}_{u \sim \pi_{\theta_t}} \big[\tfrac{d}{dt} f_u(\theta_t) \big] \big| \big) \\
&\leq \pi_{\theta_t}(y) \big(\big| \tfrac{d}{dt} f_y(\theta_t) \big| + \mathbb{E}_{u \sim \pi_{\theta_t}} \big[\big| \tfrac{d}{dt} f_u(\theta_t) \big| \big] \big) \\
&\leq 4B^2 \pi_{\theta_t}(y).
\end{aligned}$$

Here, in the first and second inequalities we apply the triangle inequality, and the last inequality is from Equation (2). Summing over all $y \in \mathcal{Z}$, we therefore have that for all $t \in [0, T]$:

$$\frac{d}{dt}\pi_{\theta_t}(\mathcal{Z}) \leq 4B^2 \pi_{\theta_t}(\mathcal{Z}).$$

By Grönwall's inequality, this implies that for all $t \in [0, T]$:

$$\pi_{\theta_t}(\mathcal{Z}) \leq \pi_{\theta_0}(\mathcal{Z}) \cdot \exp(4B^2 t) \leq \pi_{\theta_0}(\mathcal{Z}) \cdot \exp(4B^2 T) \leq \frac{2}{\Delta_{\mathcal{Z}} \exp(6B^2T)} \cdot \epsilon,$$

where the last transition is by the assumption that $\pi_{\theta_0}(\mathcal{Z}) \leq \frac{2}{\Delta_{\mathcal{Z}} \exp(10B^2T)} \cdot \epsilon$ □

With Lemma 1 in place, we turn to bound the deviation between θ_t and θ_t^{G} at time $t \in [0, T]$. Examining their time derivatives, by the triangle inequality we get that

$$\big\| \tfrac{d}{dt}\theta_t^{\mathrm{G}} - \tfrac{d}{dt}\theta_t \big\| = \|\nabla V_{\mathrm{G}}(\theta_t^{\mathrm{G}}) - \nabla V_{\mathrm{P}}(\theta_t)\| \leq \underbrace{\|\nabla V_{\mathrm{G}}(\theta_t^{\mathrm{G}}) - \nabla V_{\mathrm{G}}(\theta_t)\|}_{I_1} + \underbrace{\|\nabla V_{\mathrm{G}}(\theta_t) - \nabla V_{\mathrm{P}}(\theta_t)\|}_{I_2}.$$

For I_1, Lemma 3 shows that ∇V_{G} is $6B^2$-Lipschitz, for $B = \max_{y \in \mathcal{Y}} \|\phi(y)\|$. Thus,

$$I_1 \leq 6B^2 \big\| \theta_t^{\mathrm{G}} - \theta_t \big\|.$$

For I_2, by examining the gradient expression derived in Lemma 10 we can see that

$$\begin{aligned}\nabla V_{\mathrm{G}}(\theta_t) - \nabla V_{\mathrm{P}}(\theta_t) &= \mathbb{E}_{y\sim\pi_{\theta_t}}[(r_{\mathrm{G}}(y) - V_{\mathrm{G}}(\theta_t) - r_{\mathrm{P}}(y) + V_{\mathrm{P}}(\theta_t)) \cdot \phi(y)] \\ &= \mathbb{E}_{y\sim\pi_{\theta_t}}[(r_{\mathrm{G}}(y) - r_{\mathrm{P}}(y)) \cdot \phi(y)] - (V_{\mathrm{G}}(\theta_t) - V_{\mathrm{P}}(\theta_t)) \cdot \bar{\phi}_{\theta_t} \\ &= \mathbb{E}_{y\sim\pi_{\theta_t}}[(r_{\mathrm{G}}(y) - r_{\mathrm{P}}(y)) \cdot \phi(y)] - \mathbb{E}_{y\sim\pi_{\theta_t}}\left[(r_{\mathrm{G}}(y) - r_{\mathrm{P}}(y)) \cdot \bar{\phi}_{\theta_t}\right] \\ &= \mathbb{E}_{y\sim\pi_{\theta_t}}\left[(r_{\mathrm{G}}(y) - r_{\mathrm{P}}(y)) \cdot \left(\phi(y) - \bar{\phi}_{\theta_t}\right)\right],\end{aligned}$$

where $\bar{\phi}_{\theta_t} := \mathbb{E}_{z\sim\pi_{\theta_t}}[\phi(z)]$. By the triangle inequality and the fact that $r_{\mathrm{G}}(y) = r_{\mathrm{P}}(y)$ for all $y \in \mathcal{Y} \setminus \mathcal{Z}$, this leads to the following upper bound on I_2:

$$\begin{aligned}I_2 &= \|\nabla V_{\mathrm{G}}(\theta_t) - \nabla V_{\mathrm{P}}(\theta_t)\| \\ &\le \mathbb{E}_{y\sim\pi_{\theta_t}}\left[|r_{\mathrm{G}}(y) - r_{\mathrm{P}}(y)| \cdot \left\|\phi(y) - \bar{\phi}_{\theta_t}\right\|\right] \\ &\le 2B\Delta_{\mathcal{Z}}\pi_{\theta_t}(\mathcal{Z}),\end{aligned}$$

where $\Delta_{\mathcal{Z}} = \max_{y\in\mathcal{Z}}|r_{\mathrm{G}}(y) - r_{\mathrm{P}}(y)|$. Thus, applying the bound from Lemma 1 on $\pi_{\theta_t}(\mathcal{Z})$, we arrive at:

$$I_2 \le \frac{4B\epsilon}{\exp(6B^2T)}.$$

Combining the upper bounds on I_1 and I_2, we get:

$$\left\|\tfrac{d}{dt}\theta_t^{\mathrm{G}} - \tfrac{d}{dt}\theta_t\right\| \le 6B^2\left\|\theta_t^{\mathrm{G}} - \theta_t\right\| + \frac{4B\epsilon}{\exp(6B^2T)}. \tag{3}$$

Since $h(t) := \theta_t^{\mathrm{G}} - \theta_t$ is absolutely continuous with respect to t over $[0,T]$, so is $\|h(t)\|$. As a result, for almost every $t \in [0,T]$,

$$\frac{d}{dt}\left\|\theta_t^{\mathrm{G}} - \theta_t\right\| \le \left\|\tfrac{d}{dt}\theta_t^{\mathrm{G}} - \tfrac{d}{dt}\theta_t\right\|.$$

Thus, by Equation (3), for almost every $t \in [0,T]$ it holds that

$$\frac{d}{dt}\|\theta_t^{\mathrm{G}} - \theta_t\| \le 6B^2\|\theta_t^{\mathrm{G}} - \theta_t\| + \frac{4B\epsilon}{\exp(6B^2T)}.$$

Grönwall's inequality, along with the fact that $\theta_0^{\mathrm{G}} = \theta_0$, then yields for all $t \in [0,T]$:

$$\|\theta_t^{\mathrm{G}} - \theta_t\| \le \frac{4B\epsilon}{\exp(6B^2T)} \cdot \frac{\exp(6B^2T) - 1}{6B^2} \le \frac{2\epsilon}{3B} \le \frac{\epsilon}{B}.$$

Finally, since $\mathrm{TV}(\pi_\theta, \pi_{\theta'}) \le B\|\theta - \theta'\|$ (Lemma 2) for any $\theta, \theta' \in \mathbb{R}^D$, we may conclude that for all $t \in [0,T]$:

$$\mathrm{TV}\left(\pi_{\theta_t}, \pi_{\theta_t^{\mathrm{G}}}\right) \le B\|\theta_t^{\mathrm{G}} - \theta_t\| \le \epsilon.$$

□

D.3.1 Auxiliary Lemmas for the Proof of Proposition 2

Lemma 2. *For any $\theta, \theta' \in \mathbb{R}^D$ it holds that $\mathrm{TV}(\pi_\theta, \pi_{\theta'}) \le B\|\theta - \theta'\|$, where $B = \max_{y\in\mathcal{Y}}\|\phi(y)\|$.*

Proof. Define the curve $\theta(s) := \theta' + s \cdot (\theta - \theta')$ for $s \in [0,1]$. By the fundamental theorem of calculus it holds that

$$\pi_\theta - \pi_{\theta'} = \int_0^1 J(\theta(s))(\theta - \theta')\, ds,$$

where $J(\xi)$ is the Jacobian of π_ξ at $\xi \in \mathbb{R}^D$, with rows $J(\xi)_{y,:} = (\nabla\pi_\xi(y))^\top$. Taking ℓ_1 norms and using the induced operator norm $\|\cdot\|_{2\to1}$ gives

$$\|\pi_\theta - \pi_{\theta'}\|_1 \le \int_0^1 \|J(\theta(s))(\theta - \theta')\|_1\, ds \le \|\theta - \theta'\| \cdot \int_0^1 \|J(\theta(s))\|_{2\to1}\, ds. \tag{4}$$

It therefore suffices to show that $\|J(\xi)\|_{2\to 1} \le 2B$ for all $\xi \in \mathbb{R}^D$. For linear softmax policies,

$$\nabla \pi_\xi(y) = \pi_\xi(y)(\phi(y) - \bar{\phi}_\xi),$$

where $\bar{\phi}_\xi := \sum_{y\in\mathcal{Y}} \pi_\xi(y)\phi(y)$. Furthermore, from the definition of norm $\|\cdot\|_{2\to 1}$, we have:

$$\|J(\xi)\|_{2\to 1} = \sup_{\|v\|=1} \|J(\xi)v\|_1 = \sup_{\|v\|=1} \sum\nolimits_{y\in\mathcal{Y}} |\langle \nabla \pi_\xi(y), v\rangle|,$$

and we can upper bound $|\langle \nabla \pi_\xi(y), v\rangle|$, for $y \in \mathcal{Y}$ and $v \in \mathbb{R}^D$ with $\|v\| = 1$, by:

$$\begin{aligned} |\langle \nabla \pi_\xi(y), v\rangle| &= \left|\pi_\xi(y)\left\langle \phi(y) - \bar{\phi}_\xi, v\right\rangle\right| \\ &\le \pi_\xi(y) \cdot \left\|\phi(y) - \bar{\phi}_\xi\right\| \cdot \|v\| \\ &\le \pi_\xi(y) \cdot 2B. \end{aligned}$$

The first inequality is from the Cauchy-Schwarz inequality and the second is due to $\|\phi(y)\| \le B$ for all $y \in \mathcal{Y}$. Summing over all $y \in \mathcal{Y}$ then yields:

$$\sum\nolimits_{y\in\mathcal{Y}} |\langle \nabla \pi_\xi(y), v\rangle| \le \sum\nolimits_{y\in\mathcal{Y}} 2B\pi_\xi(y) = 2B.$$

Since this holds for all $v \in \mathbb{R}^D$ with $\|v\| = 1$, we get the desired bound on $\|J(\xi)\|_{2\to 1}$:

$$\|J(\xi)\|_{2\to 1} \le 2B.$$

Plugging this into Equation (4) concludes the proof:

$$\mathrm{TV}\big(\pi_\theta, \pi_{\theta'}\big) = \frac{1}{2}\|\pi_\theta - \pi_{\theta'}\|_1 \le \frac{1}{2}\|\theta - \theta'\| \cdot \int_0^1 2B\, ds = B\|\theta - \theta'\|.$$

□

Lemma 3. *For any $\theta, \theta' \in \mathbb{R}^D$ it holds that $\|\nabla V_{\mathrm{G}}(\theta) - \nabla V_{\mathrm{G}}(\theta')\| \le 6B^2\|\theta - \theta'\|$, where $B = \max_{y\in\mathcal{Y}}\|\phi(y)\|$.*

Proof. By Lemma 10, we may write $\nabla V_{\mathrm{G}}(\theta)$ as:

$$\nabla V_{\mathrm{G}}(\theta) = \mathbb{E}_{y\sim\pi_\theta}[A_{\mathrm{G}}(y;\theta)\phi(y)] = \mathbb{E}_{y\sim\pi_\theta}[r_{\mathrm{G}}(y)\phi(y)] - V_{\mathrm{G}}(\theta)\cdot\bar{\phi}_\theta,$$

where $\bar{\phi}_\theta := \mathbb{E}_{y\sim\pi_\theta}[\phi(y)]$. Furthermore, notice that for linear softmax policies and $y \in \mathcal{Y}$,

$$\nabla \pi_\theta(y) = \pi_\theta(y)(\phi(y) - \bar{\phi}_\theta).$$

Now, for convenience of notation denote $a(\theta) := \mathbb{E}_{y\sim\pi_\theta}[r_{\mathrm{G}}(y)\phi(y)]$ and $b(\theta) := \bar{\phi}_\theta$, and let $J_a(\theta), J_b(\theta) \in \mathbb{R}^{D\times D}$ be the Jacobians of $a(\theta)$ and $b(\theta)$ at θ, respectively. We have:

$$\begin{aligned} J_a(\theta) &= \sum\nolimits_{y\in\mathcal{Y}} r_{\mathrm{G}}(y)\phi(y)(\nabla\pi_\theta(y))^\top \\ &= \sum\nolimits_{y\in\mathcal{Y}} r_{\mathrm{G}}(y)\phi(y)\pi_\theta(y)(\phi(y) - \bar{\phi}_\theta)^\top \\ &= \mathbb{E}_{y\sim\pi_\theta}\big[r_{\mathrm{G}}(y)\phi(y)(\phi(y) - \bar{\phi}_\theta)^\top\big]. \end{aligned}$$

Similarly,

$$J_b(\theta) = \sum\nolimits_{y\in\mathcal{Y}} \phi(y)\nabla\pi_\theta(y)^\top = \mathbb{E}_{y\sim\pi_\theta}\big[\phi(y)(\phi(y) - \bar{\phi}_\theta)^\top\big].$$

Thus, the Hessian $\nabla^2 V_{\mathrm{G}}(\theta)$ is given by

$$\begin{aligned} \nabla^2 V_{\mathrm{G}}(\theta) &= J_a(\theta) - b(\theta)\nabla V_{\mathrm{G}}(\theta)^\top - V_{\mathrm{G}}(\theta)\cdot J_b(\theta) \\ &= \mathbb{E}_{y\sim\pi_\theta}\big[r_{\mathrm{G}}(y)\phi(y)(\phi(y) - \bar{\phi}_\theta)^\top\big] - \bar{\phi}_\theta\nabla V_{\mathrm{G}}(\theta)^\top - \mathbb{E}_{y\sim\pi_\theta}\big[V_{\mathrm{G}}(\theta)\phi(y)(\phi(y) - \bar{\phi}_\theta)^\top\big] \\ &= \mathbb{E}_{y\sim\pi_\theta}\big[A_{\mathrm{G}}(y;\theta)\phi(y)(\phi(y) - \bar{\phi}_\theta)^\top\big] - \bar{\phi}_\theta\nabla V_{\mathrm{G}}(\theta)^\top. \end{aligned}$$

To upper bound the spectral norm of the Hessian, *i.e.*, $\|\nabla^2 V_{\mathrm{G}}(\theta)\|_2$, we bound the two terms separately. For the first term, $\mathbb{E}_{y\sim\pi_\theta}[A_{\mathrm{G}}(y;\theta)\phi(y)(\phi(y)-\bar{\phi}_\theta)^\top]$, we have:

$$\begin{aligned}\left\|\mathbb{E}_{y\sim\pi_\theta}\left[A_{\mathrm{G}}(y;\theta)\phi(y)(\phi(y)-\bar{\phi}_\theta)^\top\right]\right\|_2 &\le \mathbb{E}_{y\sim\pi_\theta}\left[\left\|A_{\mathrm{G}}(y;\theta)\phi(y)(\phi(y)-\bar{\phi}_\theta)^\top\right\|_2\right] \\ &\le \mathbb{E}_{y\sim\pi_\theta}\left[2\|\phi(y)\|\|\phi(y)-\bar{\phi}_\theta\|\right] \\ &\le 4B^2. \end{aligned} \tag{5}$$

The first inequality is from $|A_{\mathrm{G}}(y;\theta)| \le 2$ and the second inequality is from $B = \max_{y\in\mathcal{Y}}\|\phi(y)\|$. For the second term, $\bar{\phi}_\theta \nabla V_{\mathrm{G}}(\theta)^\top$, we have:

$$\left\|\bar{\phi}_\theta \nabla V_{\mathrm{G}}(\theta)^\top\right\|_2 = \left\|\bar{\phi}_\theta\right\|\|\nabla V_{\mathrm{G}}(\theta)\| \le B\cdot\|\mathbb{E}_{y\sim\pi_\theta}[A_{\mathrm{G}}(y;\theta)\phi(y)]\| \le 2B^2. \tag{6}$$

Combining Equation (5) and Equation (6), we obtain:

$$\left\|\nabla^2 V_{\mathrm{G}}(\theta)\right\|_2 \le 4B^2 + 2B^2 = 6B^2,$$

and so ∇V_{G} is $6B^2$-Lipschitz. □

D.4 Proof of Theorem 2

For convenience, we start by introducing the following helper variable:

$$\gamma := \left(\frac{\pi_{\theta_0}(y_\star)}{M}\right)^{14/13},$$

where recall that $M := \min\{1, 0.05\cdot r_{\mathrm{G}}(y_{\mathrm{med}})^{2/7}(\Delta_1+\Delta_2)^{-1}\}$, with $\Delta_1 := r_{\mathrm{G}}(y_\star) - r_{\mathrm{G}}(y_{\mathrm{med}})$ and $\Delta_2 := r_{\mathrm{G}}(y_{\mathrm{med}}) - r_{\mathrm{G}}(y_{\mathrm{bad}})$ for some $y_{\mathrm{bad}} \in \mathcal{Y}_{\mathrm{bad}}$. We can then straightforwardly state the conditions of Assumption 2 in terms of γ.

Lemma 4. *Under Assumptions 1 and 2, the following properties hold:*

1. $\gamma \le \min\left\{0.01, \left(\frac{21}{1100(\Delta_1+\Delta_2)}\right)^{7/3}, \Delta_2^2\right\}$;
2. $\pi_{\theta_0}(y_\star) \le \min\left\{\gamma^{13/14}, \frac{r_{\mathrm{G}}(y_{\mathrm{med}})^{2/7}}{20(\Delta_1+\Delta_2)}\cdot\gamma^{13/14}\right\}$; *and*
3. $\max\left\{\frac{2\sqrt{\gamma}}{\Delta_2}, 0.05\cdot r_{\mathrm{G}}(y_{\mathrm{med}})\right\} \le \pi_{\theta_0}(\mathcal{Y}_{\mathrm{bad}}) \le 0.1\cdot r_{\mathrm{G}}(y_{\mathrm{med}})$.

Proof. All three conditions are obtained directly from Assumption 2 by substituting the definition of M and noticing that $\pi_{\theta_0}(y_\star) = M\gamma^{13/14}$. Note that Assumption 1 on the ground truth reward structure ensures that Δ_1 and Δ_2 are positive, so all quantities in the conditions above are well-defined. □

D.4.1 Proof of Case I

Suppose that gradient flow is used to maximize the expected reward with respect to r_{G}. We define t_γ as the initial time at which $V_{\mathrm{G}}(\theta_t)$ reaches $r_{\mathrm{G}}(y_{\mathrm{med}}) - \gamma$, *i.e.*:

$$t_\gamma := \min\{t \ge 0 : V_{\mathrm{G}}(\theta_t) \ge r_{\mathrm{G}}(y_{\mathrm{med}}) - \gamma\},$$

and analyze the time interval $[0, t_\gamma]$. In Lemma 5 and Proposition 4, we show that throughout this time interval, $\pi_{\theta_t}(y_{\mathrm{med}})$ monotonically increases while $\pi_{\theta_t}(\mathcal{Y}_{\mathrm{bad}})$ and $\pi_{\theta_t}(y_\star)$ monotonically decrease. Moreover, we prove that until t_γ the policy becomes highly concentrated on y_{med}, which in turn implies through Lemma 11 that $V_{\mathrm{G}}(\theta_t)$ remains stuck near y_{med} for a long time. This will yield the desired lower bound on $t_\star$—the initial time at which $V_{\mathrm{G}}(\theta_t)$ is ϵ-optimal.

Note that if $V_{\mathrm{G}}(\theta_t) < r_{\mathrm{G}}(y_{\mathrm{med}})$ for all $t \ge 0$, then the lower bound on $t_\star$ trivially holds as $V_{\mathrm{G}}(\theta_t)$ never reaches $r_{\mathrm{G}}(y_\star) - \epsilon$. Thus, throughout the proof of this case we can assume that there exists a time at which $V_{\mathrm{G}}(\theta_t) \ge r_{\mathrm{G}}(y_{\mathrm{med}})$.

Towards showing that $\pi_{\theta_t}(y_{\mathrm{med}})$ increases over $[0, t_\gamma]$ while $\pi_{\theta_t}(y_\star)$ decreases, the following lemma proves that $\frac{d}{dt}\pi_{\theta_t}(\mathcal{Y}_{\mathrm{bad}}) < 0$ when $V_{\mathrm{G}}(\theta_t) < r_{\mathrm{G}}(y_{\mathrm{med}})$.

Lemma 5. *At any time $t \geq 0$, if $V_{\mathrm{G}}(\theta_t) < r_{\mathrm{G}}(y_{\mathrm{med}})$, then $\frac{d}{dt}\pi_{\theta_t}(\mathcal{Y}_{\mathrm{bad}}) < 0$.*

Proof. For any $z \in \mathcal{Y}_{\mathrm{bad}}$, by Lemma 13:

$$\frac{d}{dt}\pi_{\theta_t}(z) = \pi_{\theta_t}(z)\Big(\pi_{\theta_t}(z)A_{\mathrm{G}}(z;\theta_t) - \sum\nolimits_{y\in\mathcal{Y}} \pi_{\theta_t}(y)^2 A_{\mathrm{G}}(y;\theta_t)\Big),$$

Summing the equation above over all $z \in \mathcal{Y}_{\mathrm{bad}}$, we arrive at:

$$\sum_{z\in\mathcal{Y}_{\mathrm{bad}}} \frac{d}{dt}\pi_{\theta_t}(z) = \sum_{z\in\mathcal{Y}_{\mathrm{bad}}} \pi_{\theta_t}(z)^2 A_{\mathrm{G}}(z;\theta_t) - \Big(\sum_{z\in\mathcal{Y}_{\mathrm{bad}}} \pi_{\theta_t}(z)\Big)\Big(\sum_{y\in\mathcal{Y}} \pi_{\theta_t}(y)^2 A_{\mathrm{G}}(y;\theta_t)\Big).$$

Thus,

$$\begin{aligned}
&\frac{d}{dt}\pi_{\theta_t}(\mathcal{Y}_{\mathrm{bad}}) \\
&= \sum_{z\in\mathcal{Y}_{\mathrm{bad}}} \pi_{\theta_t}(z)^2 A_{\mathrm{G}}(z;\theta_t) \\
&\quad - \Big(\sum_{z\in\mathcal{Y}_{\mathrm{bad}}} \pi_{\theta_t}(z)\Big)\Big(\sum_{z\in\mathcal{Y}_{\mathrm{bad}}} \pi_{\theta_t}(z)^2 A_{\mathrm{G}}(z;\theta_t) + \pi_{\theta_t}(y_{\mathrm{med}})^2 A_{\mathrm{G}}(y_{\mathrm{med}};\theta_t) + \pi_{\theta_t}(y_\star)^2 A_{\mathrm{G}}(y_\star;\theta_t)\Big) \\
&= \underbrace{\Big(\sum\nolimits_{z\in\mathcal{Y}_{\mathrm{bad}}} \pi_{\theta_t}(z)^2 A_{\mathrm{G}}(z;\theta_t)\Big)\Big(1 - \sum\nolimits_{z\in\mathcal{Y}_{\mathrm{bad}}} \pi_{\theta_t}(z)\Big)}_{I_1} \\
&\quad - \underbrace{\Big(\sum\nolimits_{z\in\mathcal{Y}_{\mathrm{bad}}} \pi_{\theta_t}(z)\Big)\Big(\pi_{\theta_t}(y_{\mathrm{med}})^2 A_{\mathrm{G}}(y_{\mathrm{med}};\theta_t) + \pi_{\theta_t}(y_\star)^2 A_{\mathrm{G}}(y_\star;\theta_t)\Big)}_{I_2}.
\end{aligned}$$

When maximizing r_{G} via gradient flow, $V_{\mathrm{G}}(\theta_t)$ is monotonically increasing since

$$\frac{d}{dt}V_{\mathrm{G}}(\theta_t) = \|\nabla V_{\mathrm{G}}(\theta_t)\|^2 \geq 0,$$

for all $t \geq 0$. Hence, $A_{\mathrm{G}}(z;\theta_t) \leq 0$ for all $z \in \mathcal{Y}_{\mathrm{bad}}$ and $t \geq 0$, because outputs in $\mathcal{Y}_{\mathrm{bad}}$ have minimal ground truth reward. Furthermore, as long as $V_{\mathrm{G}}(\theta_t) < r_{\mathrm{G}}(y_{\mathrm{med}})$, it holds that $A_{\mathrm{G}}(y_{\mathrm{med}};\theta_t) > 0$ and $A_{\mathrm{G}}(y_\star;\theta_t) > 0$. This implies that $I_1 < 0$ and $I_2 > 0$, so $\frac{d}{dt}\pi_{\theta_t}(\mathcal{Y}_{\mathrm{bad}}) = I_1 - I_2 < 0$. □

Using Lemma 5, we can now prove that until t_γ the probability of y_{med} monotonically increases while the probability of $y_\star$ monotonically decreases, and characterize how small $\pi_{\theta_t}(y_\star)$ becomes. This is a core component in the proof of Theorem 2.

Proposition 4. *For all $t \in [0, t_\gamma]$ it holds that:*

- $\frac{d}{dt}\pi_{\theta_t}(y_\star) < 0$*; and*
- $\frac{d}{dt}\pi_{\theta_t}(y_{\mathrm{med}}) > 0$.

Furthermore,

$$\min\nolimits_{t\in[0,t_\gamma]} \pi_{\theta_t}(y_\star) = \pi_{\theta_{t_\gamma}}(y_\star) \leq \frac{\pi_{\theta_0}(y_{\mathrm{med}})^2\gamma}{8A_{\mathrm{G}}(y_\star;\theta_0)}.$$

Proof. Notice that when $\frac{d}{dt}\pi_{\theta_t}(y_\star) < 0$, we must have $\frac{d}{dt}\pi_{\theta_t}(y_{\mathrm{med}}) > 0$. This is because $\frac{d}{dt}\pi_{\theta_t}(y_\star) + \frac{d}{dt}\pi_{\theta_t}(y_{\mathrm{med}}) + \frac{d}{dt}\pi_{\theta_t}(\mathcal{Y}_{\mathrm{bad}}) = \frac{d}{dt}\pi_{\theta_t}(\mathcal{Y}) = \frac{d}{dt}1 = 0$, and so when $\frac{d}{dt}\pi_{\theta_t}(y_\star) < 0$:

$$\frac{d}{dt}\pi_{\theta_t}(y_{\mathrm{med}}) = -\frac{d}{dt}\pi_{\theta_t}(y_\star) - \frac{d}{dt}\pi_{\theta_t}(\mathcal{Y}_{\mathrm{bad}}) > -\frac{d}{dt}\pi_{\theta_t}(\mathcal{Y}_{\mathrm{bad}}) > 0,$$

where the second inequality follows from Lemma 5. Hence, for proving monotonicity of $\pi_{\theta_t}(y_\star)$ and $\pi_{\theta_t}(y_{\mathrm{med}})$, it suffices to show that $\frac{d}{dt}\pi_{\theta_t}(y_\star) < 0$ for all $t \in [0, t_\gamma]$.

To do this, we let $t_{\sqrt{\gamma}}$ be the initial time at which $V_{\mathrm{G}}(\theta_t) \geq r_{\mathrm{G}}(y_{\mathrm{med}}) - \sqrt{\gamma}$, *i.e.*:

$$t_{\sqrt{\gamma}} := \min\{t \geq 0 : V_{\mathrm{G}}(\theta_t) \geq r_{\mathrm{G}}(y_{\mathrm{med}}) - \sqrt{\gamma}\}.$$

From Lemma 14, $V_{\mathrm{G}}(\theta_0) < r_{\mathrm{G}}(y_{\mathrm{med}}) - \sqrt{\gamma}$. Hence, $t_{\sqrt{\gamma}} > 0$. Furthermore, since $V_{\mathrm{G}}(\theta_t)$ is continuous in t, at $t_{\sqrt{\gamma}}$ it holds that $V_{\mathrm{G}}(\theta_{t_{\sqrt{\gamma}}}) = r_{\mathrm{G}}(y_{\mathrm{med}}) - \sqrt{\gamma}$. The proof proceeds by showing that $\frac{d}{dt}\pi_{\theta_t}(y_\star) < 0$ up to time $t_{\sqrt{\gamma}}$ and upper bounding $\pi_{\theta_{t_{\sqrt{\gamma}}}}(y_\star)$ by $\pi_{\theta_0}(y_{\mathrm{med}})^2\gamma/8A_{\mathrm{G}}(y_\star;\theta_0)$. We will then establish that $\pi_{\theta_t}(y_\star)$ continues decreasing until time t_γ, from which it follows that the upper bound on $\pi_{\theta_{t_{\sqrt{\gamma}}}}(y_\star)$ transfers to $\pi_{\theta_{t_\gamma}}(y_\star)$.

Step 1: monotonicity in $[0, t_{\sqrt{\gamma}}]$. We show that $\frac{d}{dt}\pi_{\theta_t}(y_\star) < 0$ and hence $\frac{d}{dt}\pi_{\theta_t}(y_{\mathrm{med}}) > 0$ for all $t \in [0, t_{\sqrt{\gamma}}]$. Assume by way of contradiction that there exists a time $t' \le t_{\sqrt{\gamma}}$ at which $\frac{d}{dt}\pi_{\theta_t}(y_\star)\big|_{t=t'} \ge 0$ and denote by τ the initial such time, *i.e.*:

$$\tau := \min\big\{t \in [0, t_{\sqrt{\gamma}}] : \tfrac{d}{dt}\pi_{\theta_t}(y_\star) \ge 0\big\}.$$

For all $t \in [0, \tau)$, we therefore have that $\frac{d}{dt}\pi_{\theta_t}(y_\star) < 0$ and $\frac{d}{dt}\pi_{\theta_t}(y_{\mathrm{med}}) > 0$, and so $\pi_{\theta_\tau}(y_\star) \le \pi_{\theta_0}(y_\star)$ and $\pi_{\theta_\tau}(y_{\mathrm{med}}) \ge \pi_{\theta_0}(y_{\mathrm{med}})$. We now upper bound $\pi_{\theta_\tau}(\mathcal{Y}_{\mathrm{bad}})^2$ under this assumption. Notice that

$$\begin{aligned} A_{\mathrm{G}}(y_{\mathrm{med}};\theta_\tau) &= (1-\pi_{\theta_\tau}(y_{\mathrm{med}}))r_{\mathrm{G}}(y_{\mathrm{med}}) - \pi_{\theta_\tau}(y_\star)r_{\mathrm{G}}(y_\star) - \sum\nolimits_{z\in\mathcal{Y}_{\mathrm{bad}}} \pi_{\theta_\tau}(z)r_{\mathrm{G}}(z) \\ &\ge -\Delta_1\pi_{\theta_\tau}(y_\star) + \Delta_2\pi_{\theta_\tau}(\mathcal{Y}_{\mathrm{bad}}). \end{aligned}$$

Thus,

$$\pi_{\theta_\tau}(\mathcal{Y}_{\mathrm{bad}}) \le \frac{A_{\mathrm{G}}(y_{\mathrm{med}};\theta_\tau) + \Delta_1\pi_{\theta_\tau}(y_\star)}{\Delta_2} \le \frac{1.3A_{\mathrm{G}}(y_{\mathrm{med}};\theta_\tau)}{\Delta_2}.$$

The second inequality is from $\Delta_1\pi_{\theta_\tau}(y_\star) \le 2\pi_{\theta_0}(y_\star) \le 2\gamma^{13/14} \le 0.3\sqrt{\gamma} \le 0.3A_{\mathrm{G}}(y_{\mathrm{med}};\theta_\tau)$ for $\gamma \le 0.01$ (Lemma 4). Then, we have:

$$\begin{aligned} \pi_{\theta_\tau}(\mathcal{Y}_{\mathrm{bad}})^2 &\le \frac{1.69A_{\mathrm{G}}(y_{\mathrm{med}};\theta_\tau)^2}{\Delta_2^2} \\ &< \frac{7\pi_{\theta_0}(y_{\mathrm{med}})^2A_{\mathrm{G}}(y_{\mathrm{med}};\theta_\tau)}{8\Delta_2} \cdot \frac{2A_{\mathrm{G}}(y_{\mathrm{med}};\theta_\tau)}{\Delta_2\pi_{\theta_0}(y_{\mathrm{med}})^2} \\ &\le \frac{7\pi_{\theta_0}(y_{\mathrm{med}})^2A_{\mathrm{G}}(y_{\mathrm{med}};\theta_\tau)}{8\Delta_2} \cdot \frac{2A_{\mathrm{G}}(y_{\mathrm{med}};\theta_0)}{\Delta_2\pi_{\theta_0}(y_{\mathrm{med}})^2}, \end{aligned}$$

where the second inequality is due to $2\times 7 > 1.69\times 8$ and the third inequality is from $A_{\mathrm{G}}(y_{\mathrm{med}};\theta_\tau) \le A_{\mathrm{G}}(y_{\mathrm{med}};\theta_0)$ (under gradient flow $V_{\mathrm{G}}(\theta_t)$ is monotonically non-decreasing). Focusing on the second term on the right-hand side, we prove that $\frac{2A_{\mathrm{G}}(y_{\mathrm{med}};\theta_0)}{\Delta_2\pi_{\theta_0}(y_{\mathrm{med}})^2} \le 1$. This follows from:

$$\begin{aligned} A_{\mathrm{G}}(y_{\mathrm{med}};\theta_0) &= (1-\pi_{\theta_0}(y_{\mathrm{med}}))r_{\mathrm{G}}(y_{\mathrm{med}}) - \pi_{\theta_0}(y_\star)r_{\mathrm{G}}(y_\star) - \sum\nolimits_{z\in\mathcal{Y}_{\mathrm{bad}}} \pi_{\theta_0}(z)r_{\mathrm{G}}(z) \\ &\le (1-\pi_{\theta_0}(y_{\mathrm{med}}))r_{\mathrm{G}}(y_{\mathrm{med}}) + \pi_{\theta_0}(\mathcal{Y}_{\mathrm{bad}}) \\ &\le 0.2r_{\mathrm{G}}(y_{\mathrm{med}}) + 0.1r_{\mathrm{G}}(y_{\mathrm{med}}) \\ &= 0.3r_{\mathrm{G}}(y_{\mathrm{med}}) \\ &< \frac{0.8^2r_{\mathrm{G}}(y_{\mathrm{med}})}{2} \\ &\le \frac{\Delta_2\pi_{\theta_0}(y_{\mathrm{med}})^2}{2}. \end{aligned}$$

Here, the first inequality is from $r_{\mathrm{G}}(z) \ge -1$ for $z \in \mathcal{Y}_{\mathrm{bad}}$ and $r_{\mathrm{G}}(y_\star) \ge 0$; the second inequality is from Items 2 and 3 in Lemma 4, which imply $\pi_{\theta_0}(y_{\mathrm{med}}) \ge 0.8$; and the last inequality is from $r_{\mathrm{G}}(y_{\mathrm{bad}}) \le 0$ and $\pi_{\theta_0}(y_{\mathrm{med}}) \ge 0.8$. Therefore, $\frac{2A_{\mathrm{G}}(y_{\mathrm{med}};\theta_0)}{\Delta_2\pi_{\theta_0}(y_{\mathrm{med}})^2} \le 1$, which yields

$$\pi_{\theta_\tau}(\mathcal{Y}_{\mathrm{bad}})^2 \le \frac{7\pi_{\theta_0}(y_{\mathrm{med}})^2A_{\mathrm{G}}(y_{\mathrm{med}};\theta_\tau)}{8\Delta_2}. \tag{7}$$

We now upper bound $\frac{d}{dt}\pi_{\theta_t}(y_\star)\big|_{t=\tau}$, based on the expression derived in Lemma 13:

$$\begin{aligned}
\frac{d}{dt}\pi_{\theta_t}(y_\star)\Big|_{t=\tau} &= \pi_{\theta_\tau}(y_\star)\Big(\pi_{\theta_\tau}(y_\star)A_{\mathrm{G}}(y_\star;\theta_\tau) - \sum\nolimits_{y\in\mathcal{Y}}\pi_{\theta_\tau}(y)^2 A_{\mathrm{G}}(y;\theta_\tau)\Big)\\
&= \pi_{\theta_\tau}(y_\star)\Big(\big(\pi_{\theta_\tau}(y_\star)-\pi_{\theta_\tau}(y_\star)^2\big)A_{\mathrm{G}}(y_\star;\theta_\tau) - \pi_{\theta_\tau}(y_{\mathrm{med}})^2 A_{\mathrm{G}}(y_{\mathrm{med}};\theta_\tau)\\
&\qquad - \sum\nolimits_{z\in\mathcal{Y}_{\mathrm{bad}}}\pi_{\theta_\tau}(z)^2 A_{\mathrm{G}}(z;\theta_\tau)\Big)\\
&\le \pi_{\theta_\tau}(y_\star)\Big(\big(\pi_{\theta_\tau}(y_\star)-\pi_{\theta_\tau}(y_\star)^2\big)A_{\mathrm{G}}(y_\star;\theta_\tau) - \pi_{\theta_\tau}(y_{\mathrm{med}})^2 A_{\mathrm{G}}(y_{\mathrm{med}};\theta_\tau)\\
&\qquad + (\Delta_2-\sqrt{\gamma})\sum\nolimits_{z\in\mathcal{Y}_{\mathrm{bad}}}\pi_{\theta_\tau}(z)^2\Big),
\end{aligned}$$

where the inequality is by $-A_{\mathrm{G}}(z;\theta_\tau)\le -A_{\mathrm{G}}(z;\theta_{t_{\sqrt{\gamma}}})\le r_{\mathrm{G}}(y_{\mathrm{med}})-\sqrt{\gamma}-r_{\mathrm{G}}(y_{\mathrm{bad}})=\Delta_2-\sqrt{\gamma}$. We continue upper bounding $\frac{d}{dt}\pi_{\theta_t}(y_\star)\big|_{t=\tau}$ as follows:

$$\begin{aligned}
\frac{d}{dt}\pi_{\theta_t}(y_\star)\Big|_{t=\tau} &\le \pi_{\theta_\tau}(y_\star)\Big(\big(\pi_{\theta_\tau}(y_\star)-\pi_{\theta_\tau}(y_\star)^2\big)A_{\mathrm{G}}(y_\star;\theta_\tau) - \pi_{\theta_\tau}(y_{\mathrm{med}})^2 A_{\mathrm{G}}(y_{\mathrm{med}};\theta_\tau)\\
&\qquad + (\Delta_2-\sqrt{\gamma})\sum\nolimits_{z\in\mathcal{Y}_{\mathrm{bad}}}\pi_{\theta_\tau}(z)^2\Big)\\
&\le \pi_{\theta_\tau}(y_\star)\Big(\big(\pi_{\theta_0}(y_\star)-\pi_{\theta_0}(y_\star)^2\big)A_{\mathrm{G}}(y_\star;\theta_\tau) - \pi_{\theta_0}(y_{\mathrm{med}})^2 A_{\mathrm{G}}(y_{\mathrm{med}};\theta_\tau)\\
&\qquad + (\Delta_2-\sqrt{\gamma})\pi_{\theta_\tau}(\mathcal{Y}_{\mathrm{bad}})^2\Big)\\
&\le \pi_{\theta_\tau}(y_\star)\Big(\big(\pi_{\theta_0}(y_\star)-\pi_{\theta_0}(y_\star)^2\big)A_{\mathrm{G}}(y_\star;\theta_\tau) - (1/8)\pi_{\theta_0}(y_{\mathrm{med}})^2 A_{\mathrm{G}}(y_{\mathrm{med}};\theta_\tau)\Big)\\
&\le \pi_{\theta_\tau}(y_\star)\Big(\big(\pi_{\theta_0}(y_\star)-\pi_{\theta_0}(y_\star)^2\big)A_{\mathrm{G}}(y_\star;\theta_0) - (1/8)\pi_{\theta_0}(y_{\mathrm{med}})^2\sqrt{\gamma}\Big).
\end{aligned}$$

The second inequality is from $\pi_{\theta_\tau}(y_\star)-\pi_{\theta_\tau}(y_\star)^2\le\pi_{\theta_0}(y_\star)-\pi_{\theta_0}(y_\star)^2$ (this holds since $\pi_{\theta_\tau}(y_\star)\le\pi_{\theta_0}(y_\star)<0.5$), $\pi_{\theta_\tau}(y_{\mathrm{med}})\ge\pi_{\theta_0}(y_{\mathrm{med}})$, and $\sum_{z\in\mathcal{Y}_{\mathrm{bad}}}\pi_{\theta_t}(z)^2\le\big(\sum_{z\in\mathcal{Y}_{\mathrm{bad}}}\pi_{\theta_t}(z)\big)^2$. The third inequality is from $(\Delta_2-\sqrt{\gamma})\pi_{\theta_\tau}(\mathcal{Y}_{\mathrm{bad}})^2\le\Delta_2\pi_{\theta_\tau}(\mathcal{Y}_{\mathrm{bad}})^2\le\frac{7}{8}\cdot\pi_{\theta_0}(y_{\mathrm{med}})^2A_{\mathrm{G}}(y_{\mathrm{med}};\theta_\tau)$ (Equation (7)). The last inequality is from the fact that $V_{\mathrm{G}}(\theta_t)$ is monotonically increasing, so $A_{\mathrm{G}}(y_\star;\theta_\tau)\le A_{\mathrm{G}}(y_\star;\theta_0)$ and $A_{\mathrm{G}}(y_{\mathrm{med}};\theta_\tau)\ge A_{\mathrm{G}}(y_{\mathrm{med}};\theta_{t_{\sqrt{\gamma}}})=\sqrt{\gamma}$.

From Lemma 4, we further obtain that:

$$\pi_{\theta_0}(y_\star)\le\frac{r_{\mathrm{G}}(y_{\mathrm{med}})^{2/7}\gamma^{13/14}}{20(\Delta_1+\Delta_2)}\le\frac{\sqrt{\gamma}\cdot\gamma^{3/7}}{20A_{\mathrm{G}}(y_\star;\theta_0)}<\frac{0.8^2\sqrt{\gamma}}{8A_{\mathrm{G}}(y_\star;\theta_0)}\le\frac{\pi_{\theta_0}(y_{\mathrm{med}})^2\sqrt{\gamma}}{8A_{\mathrm{G}}(y_\star;\theta_0)}.$$

Here, the first inequality is from Item 2 in Lemma 4; the second inequality is from $r_{\mathrm{G}}(y_{\mathrm{med}})\le 1$ and $A_{\mathrm{G}}(y_\star;\theta_0)\le\Delta_1+\Delta_2$; the third inequality is from $\gamma^{3/7}<1.6$; and the last inequality is from $\pi_{\theta_0}(y_{\mathrm{med}})=1-\pi_{\theta_0}(y_\star)-\pi_{\theta_0}(\mathcal{Y}_{\mathrm{bad}})\ge 1-\gamma^{13/14}-0.1r_{\mathrm{G}}(y_{\mathrm{med}})\ge 0.8$. Therefore,

$$\big(\pi_{\theta_0}(y_\star)-\pi_{\theta_0}(y_\star)^2\big)A_{\mathrm{G}}(y_\star;\theta_0)-(1/8)\pi_{\theta_0}(y_{\mathrm{med}})^2\sqrt{\gamma}<0,$$

which implies that $\frac{d}{dt}\pi_{\theta_t}(y_\star)\big|_{t=\tau}<0$. However, this contradicts the definition of τ, according to which $\frac{d}{dt}\pi_{\theta_t}(y_\star)\big|_{t=\tau}\ge 0$. Hence, $\frac{d}{dt}\pi_{\theta_t}(y_\star)<0$ and $\frac{d}{dt}\pi_{\theta_t}(y_{\mathrm{med}})>0$ for all $t\in[0,t_{\sqrt{\gamma}}]$.

Step 2: upper bound on $\pi_{\theta_{t_{\sqrt{\gamma}}}}(y_\star)$ and monotonicity in $[t_{\sqrt{\gamma}},t_\gamma]$. Next, we prove that $\pi_{\theta_t}(y_\star)$ keeps decreasing and $\pi_{\theta_t}(y_{\mathrm{med}})$ keeps increasing on the time interval $[t_{\sqrt{\gamma}},t_\gamma]$. This requires deriving upper bounds on $\frac{d}{dt}\pi_{\theta_t}(y_\star)$ and $\frac{d}{dt}\pi_{\theta_t}(y_{\mathrm{med}})$ for $t\in[0,t_{\sqrt{\gamma}}]$ that allow quantifying how small the probability of $y_\star$ becomes until time $t_{\sqrt{\gamma}}$.

Step 2.1: upper bound on $\frac{d}{dt}\pi_{\theta_t}(y_\star)$ in $[0,t_{\sqrt{\gamma}}]$. In this step, we prove that for all $t\in[0,t_{\sqrt{\gamma}}]$:

$$\frac{d}{dt}\pi_{\theta_t}(y_\star)\le -0.3\cdot\pi_{\theta_t}(y_\star)\pi_{\theta_t}(y_{\mathrm{med}})^2A_{\mathrm{G}}(y_{\mathrm{med}};\theta_t).$$

First, by Lemma 13, for $t \in [0, t_{\sqrt{\gamma}}]$ it holds that:

$$\begin{aligned}
&\frac{d}{dt}\pi_{\theta_t}(y_\star) \\
&= \pi_{\theta_t}(y_\star)\Big(\pi_{\theta_t}(y_\star)A_{\mathrm{G}}(y_\star;\theta_t) - \sum\nolimits_{y\in\mathcal{Y}} \pi_{\theta_t}(y)^2 A_{\mathrm{G}}(y;\theta_t)\Big) \\
&= \pi_{\theta_t}(y_\star)\Big(\big(\pi_{\theta_t}(y_\star) - \pi_{\theta_t}(y_\star)^2\big)A_{\mathrm{G}}(y_\star;\theta_t) \\
&\qquad\qquad - \pi_{\theta_t}(y_{\mathrm{med}})^2 A_{\mathrm{G}}(y_{\mathrm{med}};\theta_t) - \sum_{z\in\mathcal{Y}_{\mathrm{bad}}} \pi_{\theta_t}(z)^2 A_{\mathrm{G}}(z;\theta_t)\Big) \qquad (8) \\
&\le \pi_{\theta_t}(y_\star)\Bigg(\underbrace{\begin{aligned}&\big(\pi_{\theta_0}(y_\star) - \pi_{\theta_0}(y_\star)^2\big)A_{\mathrm{G}}(y_\star;\theta_t) \\ &\quad - \pi_{\theta_t}(y_{\mathrm{med}})^2 A_{\mathrm{G}}(y_{\mathrm{med}};\theta_t) - \sum_{z\in\mathcal{Y}_{\mathrm{bad}}} \pi_{\theta_t}(z)^2 A_{\mathrm{G}}(z;\theta_t)\end{aligned}}_{G}\Bigg),
\end{aligned}$$

where the inequality is due to $\pi_{\theta_t}(y_\star)$ being monotonically decreasing in $[0, t_{\sqrt{\gamma}}]$ and $\pi_{\theta_0}(y_\star) < 0.5$ (the function $g(p) = p - p^2$ is monotonically increasing over $[0, 0.5]$). We proceed by establishing that $G \le -0.3\pi_{\theta_t}(y_{\mathrm{med}})^2 A_{\mathrm{G}}(y_{\mathrm{med}};\theta_t)$ through the following computations:

$$\begin{aligned}
&G + 0.3\pi_{\theta_t}(y_{\mathrm{med}})^2 A_{\mathrm{G}}(y_{\mathrm{med}};\theta_t) \\
&= -0.7\pi_{\theta_t}(y_{\mathrm{med}})^2 A_{\mathrm{G}}(y_{\mathrm{med}};\theta_t) + \big(\pi_{\theta_0}(y_\star) - \pi_{\theta_0}(y_\star)^2\big)A_{\mathrm{G}}(y_\star;\theta_t) - \sum_{z\in\mathcal{Y}_{\mathrm{bad}}} \pi_{\theta_t}(z)^2 A_{\mathrm{G}}(z;\theta_t) \\
&\le -0.7\pi_{\theta_t}(y_{\mathrm{med}})^2 A_{\mathrm{G}}(y_{\mathrm{med}};\theta_t) + \pi_{\theta_0}(y_\star)A_{\mathrm{G}}(y_\star;\theta_t) - \sum_{z\in\mathcal{Y}_{\mathrm{bad}}} \pi_{\theta_t}(z)^2 A_{\mathrm{G}}(z;\theta_t) \\
&= -0.7\pi_{\theta_t}(y_{\mathrm{med}})^2 (r_{\mathrm{G}}(y_{\mathrm{med}}) - V_{\mathrm{G}}(\theta_t)) + \pi_{\theta_0}(y_\star)A_{\mathrm{G}}(y_\star;\theta_t) + \sum_{z\in\mathcal{Y}_{\mathrm{bad}}} \pi_{\theta_t}(z)^2 (V_{\mathrm{G}}(\theta_t) - r_{\mathrm{G}}(z)) \\
&= -0.7\pi_{\theta_t}(y_{\mathrm{med}})^2 \Big((1 - \pi_{\theta_t}(y_{\mathrm{med}}))r_{\mathrm{G}}(y_{\mathrm{med}}) - \pi_{\theta_t}(y_\star)r_{\mathrm{G}}(y_\star) - \sum\nolimits_{z\in\mathcal{Y}_{\mathrm{bad}}} \pi_{\theta_t}(z) r_{\mathrm{G}}(y_{\mathrm{bad}})\Big) \\
&\quad + \pi_{\theta_0}(y_\star)A_{\mathrm{G}}(y_\star;\theta_t) \\
&\quad + \sum\nolimits_{z\in\mathcal{Y}_{\mathrm{bad}}} \pi_{\theta_t}(z)^2 \big((1 - \pi_{\theta_t}(z))(-r_{\mathrm{G}}(z)) + \pi_{\theta_t}(y_\star)r_{\mathrm{G}}(y_\star) + \pi_{\theta_t}(y_{\mathrm{med}})r_{\mathrm{G}}(y_{\mathrm{med}})\big) \\
&\quad + \sum\nolimits_{z\in\mathcal{Y}_{\mathrm{bad}}} \pi_{\theta_t}(z)^2 \Big(\sum\nolimits_{z'\in\mathcal{Y}_{\mathrm{bad}}\setminus\{z\}} \pi_{\theta_t}(z') r_{\mathrm{G}}(z')\Big) \\
&= -0.7\pi_{\theta_t}(y_{\mathrm{med}})^2 \big(\pi_{\theta_t}(\mathcal{Y}_{\mathrm{bad}})\Delta_2 - \pi_{\theta_t}(y_\star)\Delta_1\big) + \pi_{\theta_0}(y_\star)A_{\mathrm{G}}(y_\star;\theta_t) \\
&\quad + \sum\nolimits_{z\in\mathcal{Y}_{\mathrm{bad}}} \pi_{\theta_t}(z)^2 \big(\pi_{\theta_t}(y_{\mathrm{med}})\Delta_2 + \pi_{\theta_t}(y_\star)(\Delta_1 + \Delta_2)\big) \\
&\le -0.7\pi_{\theta_t}(y_{\mathrm{med}})^2 \big(\pi_{\theta_t}(\mathcal{Y}_{\mathrm{bad}})\Delta_2 - \pi_{\theta_t}(y_\star)\Delta_1\big) + \pi_{\theta_0}(y_\star)A_{\mathrm{G}}(y_\star;\theta_t) \\
&\quad + \pi_{\theta_t}(\mathcal{Y}_{\mathrm{bad}})^2 \big(\pi_{\theta_t}(y_{\mathrm{med}})\Delta_2 + \pi_{\theta_t}(y_\star)(\Delta_1 + \Delta_2)\big) \\
&\le \underbrace{\pi_{\theta_t}(\mathcal{Y}_{\mathrm{bad}})\Delta_2\big[-0.7\pi_{\theta_t}(y_{\mathrm{med}})^2 + \pi_{\theta_t}(\mathcal{Y}_{\mathrm{bad}})\pi_{\theta_t}(y_{\mathrm{med}})\big]}_{I_1} \\
&\quad + \underbrace{\pi_{\theta_0}(y_\star)\big(0.7\pi_{\theta_t}(y_{\mathrm{med}})^2\Delta_1 + \pi_{\theta_t}(\mathcal{Y}_{\mathrm{bad}})^2(\Delta_1 + \Delta_2) + A_{\mathrm{G}}(y_\star;\theta_t)\big)}_{I_2}.
\end{aligned}$$

The first inequality is from $1 - \pi_{\theta_0}(y_\star) \le 1$; the second inequality is due to $\sum_{z\in\mathcal{Y}_{\mathrm{bad}}} \pi_{\theta_t}(z)^2 \le \pi_{\theta_t}(\mathcal{Y}_{\mathrm{bad}})^2$; and the last inequality is from $\pi_{\theta_t}(y_\star) \le \pi_{\theta_0}(y_\star)$. We now upper bound I_1 and I_2 separately. Specifically, we will prove the following two inequalities:

$$\begin{aligned}
I_1 &\le -0.2\pi_{\theta_t}(\mathcal{Y}_{\mathrm{bad}})\pi_{\theta_t}(y_{\mathrm{med}})^2\Delta_2, \\
I_2 &\le 0.2\pi_{\theta_t}(\mathcal{Y}_{\mathrm{bad}})\pi_{\theta_t}(y_{\mathrm{med}})^2\Delta_2.
\end{aligned}$$

For I_1, note that it suffices to show that $0.5\pi_{\theta_t}(y_{\mathrm{med}}) \ge \pi_{\theta_t}(\mathcal{Y}_{\mathrm{bad}})$. Indeed, from Lemma 4 we know that $\pi_{\theta_0}(\mathcal{Y}_{\mathrm{bad}}) < 0.2$ and $\gamma < 0.05$, so $\pi_{\theta_0}(y_\star) \le \gamma^{13/14} < 0.13$ and $\pi_{\theta_0}(y_{\mathrm{med}}) = 1 -$

$\pi_{\theta_0}(y_\star) - \pi_{\theta_0}(\mathcal{Y}_{\text{bad}}) > 0.67$. In turn, this implies that $\pi_{\theta_t}(\mathcal{Y}_{\text{bad}}) \le \pi_{\theta_0}(\mathcal{Y}_{\text{bad}}) < 0.5\pi_{\theta_0}(y_{\text{med}}) \le 0.5\pi_{\theta_t}(y_{\text{med}})$. Hence, $I_1 \le -0.2\pi_{\theta_t}(\mathcal{Y}_{\text{bad}})\pi_{\theta_t}(y_{\text{med}})^2\Delta_2$.

As for I_2, we have:

$$0.7\pi_{\theta_t}(y_{\text{med}})^2\Delta_1 + \pi_{\theta_t}(\mathcal{Y}_{\text{bad}})^2(\Delta_1+\Delta_2) + A_{\text{G}}(y_\star;\theta_t) \le 0.7\Delta_1 + 0.1(\Delta_1+\Delta_2) + A_{\text{G}}(y_\star;\theta_0) \le 2(\Delta_1+\Delta_2).$$

Here, the first inequality is from $\pi_{\theta_t}(\mathcal{Y}_{\text{bad}})^2 \le \pi_{\theta_0}(\mathcal{Y}_{\text{bad}})^2 \le \pi_{\theta_0}(\mathcal{Y}_{\text{bad}}) \le 0.1$ and $V_{\text{G}}(\theta_t) \ge V_{\text{G}}(\theta_0)$, and the second inequality is from $A_{\text{G}}(y_\star;\theta_0) \le \Delta_1+\Delta_2$. Meanwhile, since $A_{\text{G}}(y_{\text{med}};\theta_t) \ge \sqrt{\gamma}$, it follows that:

$$\begin{aligned} A_{\text{G}}(y_{\text{med}};\theta_t) &= (1-\pi_{\theta_t}(y_{\text{med}}))r_{\text{G}}(y_{\text{med}}) - \pi_{\theta_t}(y_\star)r_{\text{G}}(y_\star) - \sum\nolimits_{z\in\mathcal{Y}_{\text{bad}}} \pi_{\theta_t}(z)r_{\text{G}}(z) \\ &= -\Delta_1\pi_{\theta_t}(y_\star) + \Delta_2\pi_{\theta_t}(\mathcal{Y}_{\text{bad}}). \end{aligned}$$

Thus, $-\Delta_1\pi_{\theta_t}(y_\star) + \Delta_2\pi_{\theta_t}(\mathcal{Y}_{\text{bad}}) \ge \sqrt{\gamma}$, which implies that

$$\pi_{\theta_t}(\mathcal{Y}_{\text{bad}}) \ge \frac{\sqrt{\gamma} + \Delta_1\pi_{\theta_t}(y_\star)}{\Delta_2} \ge \frac{\sqrt{\gamma}}{\Delta_2}. \tag{9}$$

Since $\pi_{\theta_0}(y_{\text{med}}) > 0.67$ (as we showed above when bounding I_1), from Lemma 4 we know that $\gamma \le \pi_{\theta_0}(y_{\text{med}})^{14/3}/(10(\Delta_1+\Delta_2))^{7/3}$. This leads to:

$$\pi_{\theta_0}(y_\star) \le \gamma^{13/14} \le \frac{\sqrt{\gamma}\pi_{\theta_0}(y_{\text{med}})^2}{10(\Delta_1+\Delta_2)},$$

which yields:

$$I_2 \le 2(\Delta_1+\Delta_2)\pi_{\theta_0}(y_\star) \le 0.2\pi_{\theta_t}(\mathcal{Y}_{\text{bad}})\pi_{\theta_t}(y_{\text{med}})^2\Delta_2.$$

Therefore, $I_1 + I_2 \le 0$ and we can conclude that $G \le -0.3\pi_{\theta_t}(y_{\text{med}})^2 A_{\text{G}}(y_{\text{med}};\theta_t)$. Going back to Equation (8), we obtain the desired upper bound on $\frac{d}{dt}\pi_{\theta_t}(y_\star)$ for $t \in [0, t_{\sqrt{\gamma}}]$:

$$\frac{d}{dt}\pi_{\theta_t}(y_\star) \le -0.3 \cdot \pi_{\theta_t}(y_\star)\pi_{\theta_t}(y_{\text{med}})^2 A_{\text{G}}(y_{\text{med}};\theta_t).$$

Step 2.2: upper bound on $\frac{d}{dt}\pi_{\theta_t}(y_{\text{med}})$ in $[0, t_{\sqrt{\gamma}}]$. Step 1 showed that $\frac{d}{dt}\pi_{\theta_t}(y_{\text{med}}) > 0$ for all $t \in [0, t_{\sqrt{\gamma}}]$. Here, we prove that $\pi_{\theta_t}(y_{\text{med}})$ does not grow too fast by establishing that for any such t:

$$\frac{d}{dt}\pi_{\theta_t}(y_{\text{med}}) \le 2.1\pi_{\theta_t}(y_{\text{med}})^2(1-\pi_{\theta_t}(y_{\text{med}}))A_{\text{G}}(y_{\text{med}};\theta_t).$$

We first upper bound $\frac{d}{dt}\pi_{\theta_t}(y_{\text{med}})$, based on the expression derived in Lemma 13:

$$\begin{aligned} &\frac{d}{dt}\pi_{\theta_t}(y_{\text{med}}) \\ &= \pi_{\theta_t}(y_{\text{med}})\Big(\pi_{\theta_t}(y_{\text{med}})A_{\text{G}}(y_{\text{med}};\theta_t) - \sum\nolimits_{y\in\mathcal{Y}} \pi_{\theta_t}(y)^2 A_{\text{G}}(y;\theta_t)\Big) \\ &= \pi_{\theta_t}(y_{\text{med}})\Big(\big(\pi_{\theta_t}(y_{\text{med}}) - \pi_{\theta_t}(y_{\text{med}})^2\big)A_{\text{G}}(y_{\text{med}};\theta_t) - \pi_{\theta_t}(y_\star)^2 A_{\text{G}}(y_\star;\theta_t) \\ &\qquad\qquad - \sum\nolimits_{z\in\mathcal{Y}_{\text{bad}}} \pi_{\theta_t}(z)^2 A_{\text{G}}(z;\theta_t)\Big) \\ &\le \pi_{\theta_t}(y_{\text{med}})\Big(\big(\pi_{\theta_t}(y_{\text{med}}) - \pi_{\theta_t}(y_{\text{med}})^2\big)A_{\text{G}}(y_{\text{med}};\theta_t) - \sum\nolimits_{z\in\mathcal{Y}_{\text{bad}}} \pi_{\theta_t}(z)^2 A_{\text{G}}(z;\theta_t)\Big) \\ &= \pi_{\theta_t}(y_{\text{med}})\Big(\big(\pi_{\theta_t}(y_{\text{med}}) - \pi_{\theta_t}(y_{\text{med}})^2\big)A_{\text{G}}(y_{\text{med}};\theta_t) + (\Delta_2 - A_{\text{G}}(y_{\text{med}};\theta_t)) \sum_{z\in\mathcal{Y}_{\text{bad}}} \pi_{\theta_t}(z)^2\Big) \\ &\le \pi_{\theta_t}(y_{\text{med}})\Big(\big(\pi_{\theta_t}(y_{\text{med}}) - \pi_{\theta_t}(y_{\text{med}})^2\big)A_{\text{G}}(y_{\text{med}};\theta_t) + (\Delta_2 - A_{\text{G}}(y_{\text{med}};\theta_t))\pi_{\theta_t}(\mathcal{Y}_{\text{bad}})^2\Big) \\ &\le \pi_{\theta_t}(y_{\text{med}})\Big(\big(1-\pi_{\theta_t}(y_{\text{med}})\big)\big(\pi_{\theta_t}(y_{\text{med}})A_{\text{G}}(y_{\text{med}};\theta_t) + \underbrace{(\Delta_2 - A_{\text{G}}(y_{\text{med}};\theta_t))\pi_{\theta_t}(\mathcal{Y}_{\text{bad}})}_{G}\big)\Big). \end{aligned} \tag{10}$$

The first inequality is from $A_{\mathrm{G}}(y_\star;\theta_t) \geq 0$; the second inequality is due to $\sum_{z\in\mathcal{Y}_{\mathrm{bad}}} \pi_{\theta_t}(z)^2 \leq \pi_{\theta_t}(\mathcal{Y}_{\mathrm{bad}})^2$; and the last inequality is by $\pi_{\theta_t}(\mathcal{Y}_{\mathrm{bad}}) \leq 1 - \pi_{\theta_t}(y_{\mathrm{med}})$. Next, we upper bound the term G from the inequality above, proving that $G \leq 1.1\pi_{\theta_t}(y_{\mathrm{med}})A_{\mathrm{G}}(y_{\mathrm{med}};\theta_t)$. Equivalently, we show that $1.1\pi_{\theta_t}(y_{\mathrm{med}})A_{\mathrm{G}}(y_{\mathrm{med}};\theta_t) - G \geq 0$. This follows from the following computations:

$$
\begin{aligned}
&1.1\pi_{\theta_t}(y_{\mathrm{med}})A_{\mathrm{G}}(y_{\mathrm{med}};\theta_t) - G \\
&= 1.1\pi_{\theta_t}(y_{\mathrm{med}})A_{\mathrm{G}}(y_{\mathrm{med}};\theta_t) - (\Delta_2 - A_{\mathrm{G}}(y_{\mathrm{med}};\theta_t))\pi_{\theta_t}(\mathcal{Y}_{\mathrm{bad}}) \\
&= 1.1\pi_{\theta_t}(y_{\mathrm{med}})\big(r_{\mathrm{G}}(y_{\mathrm{med}}) - V_{\mathrm{G}}(\theta_t)\big) - \pi_{\theta_t}(\mathcal{Y}_{\mathrm{bad}})\big(V_{\mathrm{G}}(\theta_t) - r_{\mathrm{G}}(y_{\mathrm{bad}})\big) \\
&= 1.1\pi_{\theta_t}(y_{\mathrm{med}})\Big((1-\pi_{\theta_t}(y_{\mathrm{med}}))r_{\mathrm{G}}(y_{\mathrm{med}}) - \pi_{\theta_t}(y_\star)r_{\mathrm{G}}(y_\star) - \sum_{z\in\mathcal{Y}_{\mathrm{bad}}} \pi_{\theta_t}(z)r_{\mathrm{G}}(z)\Big) \\
&\quad - \pi_{\theta_t}(\mathcal{Y}_{\mathrm{bad}})\Big((1-\pi_{\theta_t}(\mathcal{Y}_{\mathrm{bad}}))(-r_{\mathrm{G}}(y_{\mathrm{bad}})) + \pi_{\theta_t}(y_\star)r_{\mathrm{G}}(y_\star) + \pi_{\theta_t}(y_{\mathrm{med}})r_{\mathrm{G}}(y_{\mathrm{med}})\Big) \\
&= 1.1\pi_{\theta_t}(y_{\mathrm{med}})\big(\pi_{\theta_t}(\mathcal{Y}_{\mathrm{bad}})\Delta_2 - \pi_{\theta_t}(y_\star)\Delta_1\big) - \pi_{\theta_t}(\mathcal{Y}_{\mathrm{bad}})\Big(\pi_{\theta_t}(y_{\mathrm{med}})\Delta_2 + \pi_{\theta_t}(y_\star)(\Delta_1+\Delta_2)\Big) \\
&= \pi_{\theta_t}(\mathcal{Y}_{\mathrm{bad}})\big(1.1\pi_{\theta_t}(y_{\mathrm{med}})\Delta_2 - \pi_{\theta_t}(y_{\mathrm{med}})\Delta_2\big) \\
&\quad - \pi_{\theta_t}(y_\star)\big(1.1\pi_{\theta_t}(y_{\mathrm{med}})\Delta_1 + \pi_{\theta_t}(\mathcal{Y}_{\mathrm{bad}})(\Delta_1+\Delta_2)\big) \\
&= \underbrace{0.1\pi_{\theta_t}(\mathcal{Y}_{\mathrm{bad}})\Delta_2\pi_{\theta_t}(y_{\mathrm{med}})}_{I_1} - \underbrace{\pi_{\theta_t}(y_\star)\big(1.1\pi_{\theta_t}(y_{\mathrm{med}})\Delta_1 + \pi_{\theta_t}(\mathcal{Y}_{\mathrm{bad}})(\Delta_1+\Delta_2)\big)}_{I_2}.
\end{aligned}
$$

It remains to show that $I_2 \leq 0.1\pi_{\theta_t}(y_{\mathrm{med}})\pi_{\theta_t}(\mathcal{Y}_{\mathrm{bad}})\Delta_2 = I_1$. Focusing on I_2, we have:

$$
\begin{aligned}
1.1\pi_{\theta_t}(y_{\mathrm{med}})\Delta_1 + \pi_{\theta_t}(\mathcal{Y}_{\mathrm{bad}})(\Delta_1+\Delta_2) &< 1.1\pi_{\theta_t}(y_{\mathrm{med}})\Delta_1 + 0.4\pi_{\theta_t}(y_{\mathrm{med}})(\Delta_1+\Delta_2) \\
&\leq 1.5\pi_{\theta_t}(y_{\mathrm{med}})(\Delta_1+\Delta_2),
\end{aligned}
$$

where the first inequality follows from Lemma 4, which along with the monotonicity of $\pi_{\theta_t}(\mathcal{Y}_{\mathrm{bad}})$ (Lemma 5) and $\pi_{\theta_t}(y_{\mathrm{med}})$, implies that $\pi_{\theta_t}(\mathcal{Y}_{\mathrm{bad}}) \leq \pi_{\theta_0}(\mathcal{Y}_{\mathrm{bad}}) < 0.4\pi_{\theta_0}(y_{\mathrm{med}}) \leq 0.4\pi_{\theta_t}(y_{\mathrm{med}})$. Furthermore, because $\gamma \leq \left(\frac{1}{50(\Delta_1+\Delta_2)}\right)^{7/3}$ (Lemma 4) and $\pi_{\theta_t}(y_\star)$ monotonically decreases over $[0, t_{\sqrt{\gamma}}]$, we can bound $\pi_{\theta_t}(y_\star)$ as:

$$
\pi_{\theta_t}(y_\star) \leq \pi_{\theta_0}(y_\star) \leq \gamma^{13/14} \leq \frac{\sqrt{\gamma}}{50(\Delta_1+\Delta_2)}.
$$

By Equation (9), $\pi_{\theta_t}(\mathcal{Y}_{\mathrm{bad}}) \geq \frac{\sqrt{\gamma}}{\Delta_2}$ whenever $A_{\mathrm{G}}(y_{\mathrm{med}};\theta_t) \geq \sqrt{\gamma}$. Using this fact, we obtain:

$$
I_2 \leq 1.5\pi_{\theta_t}(y_{\mathrm{med}})\pi_{\theta_t}(y_\star)(\Delta_1+\Delta_2) \leq 0.1\pi_{\theta_t}(y_{\mathrm{med}})\pi_{\theta_t}(\mathcal{Y}_{\mathrm{bad}})\Delta_2 = I_1.
$$

Thus, $G \leq 1.1\pi_{\theta_t}(y_{\mathrm{med}})A_{\mathrm{G}}(y_{\mathrm{med}};\theta_t)$ for G defined in Equation (10) and we can conclude that:

$$
\frac{d}{dt}\pi_{\theta_t}(y_{\mathrm{med}}) \leq 2.1\pi_{\theta_t}(y_{\mathrm{med}})^2(1-\pi_{\theta_t}(y_{\mathrm{med}}))A_{\mathrm{G}}(y_{\mathrm{med}};\theta_t).
$$

Step 2.3: upper bound on $\pi_{\theta_{t_{\sqrt{\gamma}}}}(y_\star)$. To summarize, we showed that for all $t \in [0, t_{\sqrt{\gamma}}]$:

$$
\begin{aligned}
\frac{d}{dt}\pi_{\theta_t}(y_\star) &\leq -0.3 \cdot \pi_{\theta_t}(y_\star)\pi_{\theta_t}(y_{\mathrm{med}})^2 A_{\mathrm{G}}(y_{\mathrm{med}};\theta_t), \\
\frac{d}{dt}\pi_{\theta_t}(y_{\mathrm{med}}) &\leq 2.1 \cdot \pi_{\theta_t}(y_{\mathrm{med}})^2(1-\pi_{\theta_t}(y_{\mathrm{med}}))A_{\mathrm{G}}(y_{\mathrm{med}};\theta_t).
\end{aligned}
$$

Since both sides of the first inequality are negative and both sides of the second inequality are positive, we can divide them to get:

$$
\frac{\frac{d}{dt}\pi_{\theta_t}(y_\star)}{\frac{d}{dt}\pi_{\theta_t}(y_{\mathrm{med}})} \leq -\frac{1}{7} \cdot \frac{\pi_{\theta_t}(y_\star)}{1-\pi_{\theta_t}(y_{\mathrm{med}})}.
$$

We can rewrite this inequality as:

$$
\frac{\frac{d}{dt}\pi_{\theta_t}(y_\star)}{\pi_{\theta_t}(y_\star)} \leq \frac{1}{7} \cdot \frac{-\frac{d}{dt}\pi_{\theta_t}(y_{\mathrm{med}})}{1-\pi_{\theta_t}(y_{\mathrm{med}})} = \frac{1}{7} \cdot \frac{\frac{d}{dt}(1-\pi_{\theta_t}(y_{\mathrm{med}}))}{1-\pi_{\theta_t}(y_{\mathrm{med}})},
$$

from which it follows that:

$$\frac{d}{dt}\ln \pi_{\theta_t}(y_\star) \leq \frac{1}{7}\cdot\frac{d}{dt}\ln(1-\pi_{\theta_t}(y_{\text{med}})).$$

Integrating both sides then leads to:

$$\ln \pi_{\theta_t}(y_\star)\Big|_0^{t_{\sqrt{\gamma}}} \leq \frac{1}{7}\cdot \ln(1-\pi_{\theta_t}(y_{\text{med}}))\Big|_0^{t_{\sqrt{\gamma}}}.$$

Thus,

$$\frac{\pi_{\theta_{t_{\sqrt{\gamma}}}}(y_\star)}{\pi_{\theta_0}(y_\star)} \leq \left(\frac{1-\pi_{\theta_{t_{\sqrt{\gamma}}}}(y_{\text{med}})}{1-\pi_{\theta_0}(y_{\text{med}})}\right)^{1/7},$$

and so,

$$\pi_{\theta_{t_{\sqrt{\gamma}}}}(y_\star) \leq \frac{\pi_{\theta_0}(y_\star)(1-\pi_{\theta_{t_{\sqrt{\gamma}}}}(y_{\text{med}}))^{1/7}}{(1-\pi_{\theta_0}(y_{\text{med}}))^{1/7}}. \tag{11}$$

Next, we prove that $1-\pi_{\theta_{t_{\sqrt{\gamma}}}}(y_{\text{med}}) \leq \frac{\sqrt{\gamma}+\pi_{\theta_0}(y_\star)}{r_{\text{G}}(y_{\text{med}})}$. Assume by way of contradiction that this is not the case. We would therefore have that:

$$\begin{aligned} V_{\text{G}}(\theta_{t_{\sqrt{\gamma}}}) &= \pi_{\theta_{t_{\sqrt{\gamma}}}}(y_{\text{med}})r_{\text{G}}(y_{\text{med}}) + \pi_{\theta_{t_{\sqrt{\gamma}}}}(y_\star)r_{\text{G}}(y_\star) + \sum\nolimits_{z\in\mathcal{Y}_{\text{bad}}}\pi_{\theta_{t_{\sqrt{\gamma}}}}(z)r_{\text{G}}(z) \\ &< \left(1-\frac{\sqrt{\gamma}+\pi_{\theta_0}(y_\star)}{r_{\text{G}}(y_{\text{med}})}\right)r_{\text{G}}(y_{\text{med}}) + \pi_{\theta_0}(y_\star) \leq r_{\text{G}}(y_{\text{med}}) - \sqrt{\gamma}, \end{aligned}$$

where the inequality is by $r_{\text{G}}(z)\leq 0$ for all $z\in\mathcal{Y}_{\text{bad}}$ and $r_{\text{G}}(y_\star)\leq 1$. However, this contradicts the definition of $t_{\sqrt{\gamma}}$, which is the initial time at which $V_{\text{G}}(\theta_t) \geq r_{\text{G}}(y_{\text{med}}) - \sqrt{\gamma}$. Hence, it must be that:

$$1-\pi_{\theta_{t_{\sqrt{\gamma}}}}(y_{\text{med}}) \leq \frac{\sqrt{\gamma}+\pi_{\theta_0}(y_\star)}{r_{\text{G}}(y_{\text{med}})}.$$

Plugging this upper bound into Equation (11), we get:

$$\begin{aligned} \pi_{\theta_{t_{\sqrt{\gamma}}}}(y_\star) &\leq \frac{\pi_{\theta_0}(y_\star)\big(\sqrt{\gamma}+\pi_{\theta_0}(y_\star)\big)^{1/7}}{\big[r_{\text{G}}(y_{\text{med}})(1-\pi_{\theta_0}(y_{\text{med}}))\big]^{1/7}} \\ &\leq \frac{\pi_{\theta_0}(y_\star)(1.3\sqrt{\gamma})^{1/7}}{(0.05\cdot r_{\text{G}}(y_{\text{med}})^2)^{1/7}} \\ &= \pi_{\theta_0}(y_\star)\cdot\left(\frac{26\sqrt{\gamma}}{r_{\text{G}}(y_{\text{med}})^2}\right)^{1/7}, \end{aligned} \tag{12}$$

where the second inequality is from $1-\pi_{\theta_0}(y_{\text{med}}) \geq \pi_{\theta_0}(\mathcal{Y}_{\text{bad}}) \geq 0.05\cdot r_{\text{G}}(y_{\text{med}})$ and $\pi_{\theta_0}(y_\star)\leq \gamma^{13/14}\leq 0.3\sqrt{\gamma}$ (see Lemma 4).

Step 2.4: completing the proof of monotonicity in $[t_{\sqrt{\gamma}}, t_\gamma]$ Finally, we conclude the proof of Proposition 4 by considering the optimization dynamics over $[t_{\sqrt{\gamma}}, t_\gamma]$ and showing that $\pi_{\theta_t}(y_\star)$ continues to decrease during this time interval.

Assume by way of contradiction that there exists $t'\in[0,t_\gamma]$ such that $\frac{d}{dt}\pi_{\theta_t}(y_\star)\big|_{t=t'} \geq 0$. Let τ be the initial time in $[0,t_\gamma]$ at which $\frac{d}{dt}\pi_{\theta_t}(y_\star)\geq 0$, *i.e.*:

$$\tau := \min\big\{t\in[0,t_\gamma] : \tfrac{d}{dt}\pi_{\theta_t}(y_\star)\geq 0\big\}.$$

By definition, $\frac{d}{dt}\pi_{\theta_t}(y_\star) < 0$, and so $\frac{d}{dt}\pi_{\theta_t}(y_{\text{med}}) > 0$, for all $t\in[0,\tau)$. We also know that if τ exists, then $\tau \geq t_{\sqrt{\gamma}}$ since $\frac{d}{dt}\pi_{\theta_t}(y_\star) < 0$ for all $t\in[0,t_{\sqrt{\gamma}}]$. This implies that $\pi_{\theta_\tau}(y_\star) \leq \pi_{\theta_{t_{\sqrt{\gamma}}}}(y_\star)$ and $\pi_{\theta_\tau}(y_{\text{med}}) \geq \pi_{\theta_0}(y_{\text{med}})$. Next, we establish that under the assumption above

$$\pi_{\theta_\tau}(\mathcal{Y}_{\text{bad}})^2 \leq \frac{7\pi_{\theta_0}(y_{\text{med}})^2 A_{\text{G}}(y_{\text{med}};\theta_\tau)}{8\Delta_2}.$$

To see this, notice that:

$$\begin{aligned} A_{\mathrm{G}}(y_{\mathrm{med}};\theta_\tau) &= (1-\pi_{\theta_\tau}(y_{\mathrm{med}}))r_{\mathrm{G}}(y_{\mathrm{med}}) - \pi_{\theta_\tau}(y_\star)r_{\mathrm{G}}(y_\star) - \sum\nolimits_{z\in\mathcal{Y}_{\mathrm{bad}}} \pi_{\theta_\tau}(z)r_{\mathrm{G}}(z) \\ &= -\Delta_1\pi_{\theta_\tau}(y_\star) + \Delta_2\pi_{\theta_\tau}(\mathcal{Y}_{\mathrm{bad}}). \end{aligned}$$

Thus,

$$\pi_{\theta_\tau}(\mathcal{Y}_{\mathrm{bad}}) = \frac{A_{\mathrm{G}}(y_{\mathrm{med}};\theta_\tau) + \Delta_1\pi_{\theta_\tau}(y_\star)}{\Delta_2} \le \frac{1.3A_{\mathrm{G}}(y_{\mathrm{med}};\theta_\tau)}{\Delta_2}.$$

The inequality above follows from Item 2 in Lemma 4 since

$$\pi_{\theta_0}(y_\star) \le \frac{r_{\mathrm{G}}(y_{\mathrm{med}})^{2/7}\gamma^{13/14}}{20(\Delta_1+\Delta_2)} \le \frac{r_{\mathrm{G}}(y_{\mathrm{med}})^{2/7}\gamma^{13/14}}{20\Delta_1} \le \frac{0.3\gamma^{13/14}}{\Delta_1(26/r_{\mathrm{G}}(y_{\mathrm{med}})^2)^{1/7}},$$

which combined with Equation (12) yields:

$$\Delta_1\pi_{\theta_\tau}(y_\star) \le \Delta_1\pi_{\theta_{t_{\sqrt{\gamma}}}}(y_\star) \le \Delta_1\pi_{\theta_0}(y_\star)\cdot(26\sqrt{\gamma}/r_{\mathrm{G}}(y_{\mathrm{med}})^2)^{1/7} \le 0.3\gamma \le 0.3A_{\mathrm{G}}(y_{\mathrm{med}};\theta_\tau).$$

It therefore holds that:

$$\begin{aligned} \pi_{\theta_\tau}(\mathcal{Y}_{\mathrm{bad}})^2 &\le \frac{1.69A_{\mathrm{G}}(y_{\mathrm{med}};\theta_\tau)^2}{\Delta_2^2} \\ &< \frac{7\pi_{\theta_0}(y_{\mathrm{med}})^2A_{\mathrm{G}}(y_{\mathrm{med}};\theta_\tau)}{8\Delta_2}\cdot\frac{2A_{\mathrm{G}}(y_{\mathrm{med}};\theta_\tau)}{\Delta_2\pi_{\theta_0}(y_{\mathrm{med}})^2} \\ &\le \frac{7\pi_{\theta_0}(y_{\mathrm{med}})^2A_{\mathrm{G}}(y_{\mathrm{med}};\theta_\tau)}{8\Delta_2}\cdot\frac{2A_{\mathrm{G}}(y_{\mathrm{med}};\theta_0)}{\Delta_2\pi_{\theta_0}(y_{\mathrm{med}})^2}, \end{aligned}$$

where the third inequality is from $A_{\mathrm{G}}(y_{\mathrm{med}};\theta_\tau) \le A_{\mathrm{G}}(y_{\mathrm{med}};\theta_0)$ (the advantage of an output is monotonically non-increasing since $V_{\mathrm{G}}(\theta_t)$ is non-decreasing under gradient flow). We now upper bound the second term on the right-hand side in the inequality above by one. Examining $A_{\mathrm{G}}(y_{\mathrm{med}};\theta_0)$:

$$\begin{aligned} A_{\mathrm{G}}(y_{\mathrm{med}};\theta_0) &= (1-\pi_{\theta_0}(y_{\mathrm{med}}))r_{\mathrm{G}}(y_{\mathrm{med}}) - \pi_{\theta_0}(y_\star)r_{\mathrm{G}}(y_\star) - \sum\nolimits_{z\in\mathcal{Y}_{\mathrm{bad}}} \pi_{\theta_0}(z)r_{\mathrm{G}}(z) \\ &\le (1-\pi_{\theta_0}(y_{\mathrm{med}}))r_{\mathrm{G}}(y_{\mathrm{med}}) + \pi_{\theta_0}(\mathcal{Y}_{\mathrm{bad}}) \\ &\le 0.2r_{\mathrm{G}}(y_{\mathrm{med}}) + 0.1r_{\mathrm{G}}(y_{\mathrm{med}}) \\ &= 0.3r_{\mathrm{G}}(y_{\mathrm{med}}) \\ &< \frac{0.8^2r_{\mathrm{G}}(y_{\mathrm{med}})}{2} \\ &\le \frac{21\Delta_2\pi_{\theta_0}(y_{\mathrm{med}})^2}{44}. \end{aligned}$$

Here, the first inequality is from $r_{\mathrm{G}}(z) \ge -1$; the second inequality is from Lemma 4 since $\pi_{\theta_0}(\mathcal{Y}_{\mathrm{bad}}) \le 0.1r_{\mathrm{G}}(y_{\mathrm{med}}) \le 0.1$ and $\pi_{\theta_0}(y_{\mathrm{med}}) \ge 0.8$; and the last inequality is from $r_{\mathrm{G}}(y_{\mathrm{bad}}) \le 0$ and $\pi_{\theta_0}(y_{\mathrm{med}}) \ge 0.8$. This implies that $\frac{2A_{\mathrm{G}}(y_{\mathrm{med}};\theta_0)}{\Delta_2\pi_{\theta_0}(y_{\mathrm{med}})^2} \le 1$, and so:

$$\pi_{\theta_\tau}(\mathcal{Y}_{\mathrm{bad}})^2 \le \frac{7\pi_{\theta_0}(y_{\mathrm{med}})^2A_{\mathrm{G}}(y_{\mathrm{med}};\theta_\tau)}{8\Delta_2}. \tag{13}$$

Let us now consider $\frac{d}{dt}\pi_{\theta_t}(y_\star)\big|_{t=\tau}$ and obtain a contradiction by showing that it is negative. We start with the following computations:

$$
\begin{aligned}
&\frac{d}{dt}\pi_{\theta_t}(y_\star)\Big|_{t=\tau} \\
&= \pi_{\theta_\tau}(y_\star)\Big(\pi_{\theta_\tau}(y_\star)A_{\mathrm{G}}(y_\star;\theta_\tau) - \sum_y \pi_{\theta_\tau}(y)^2 A_{\mathrm{G}}(y;\theta_\tau)\Big) \\
&= \pi_{\theta_\tau}(y_\star)\Big(\big(\pi_{\theta_\tau}(y_\star) - \pi_{\theta_\tau}(y_\star)^2\big)A_{\mathrm{G}}(y_\star;\theta_\tau) - \pi_{\theta_\tau}(y_{\mathrm{med}})^2 A_{\mathrm{G}}(y_{\mathrm{med}};\theta_\tau) \\
&\qquad\qquad - \sum\nolimits_{z\in\mathcal{Y}_{\mathrm{bad}}} \pi_{\theta_\tau}(z)^2 A_{\mathrm{G}}(z;\theta_\tau)\Big) \\
&\le \pi_{\theta_\tau}(y_\star)\Big(\big(\pi_{\theta_{t_{\sqrt{\gamma}}}}(y_\star) - \pi_{\theta_{t_{\sqrt{\gamma}}}}(y_\star)^2\big)A_{\mathrm{G}}(y_\star;\theta_\tau) - \pi_{\theta_0}(y_{\mathrm{med}})^2 A_{\mathrm{G}}(y_{\mathrm{med}};\theta_\tau) \\
&\qquad\qquad - \sum\nolimits_{z\in\mathcal{Y}_{\mathrm{bad}}} \pi_{\theta_\tau}(z)^2 A_{\mathrm{G}}(z;\theta_\tau)\Big) \\
&\le \pi_{\theta_\tau}(y_\star)\Big(\big(\pi_{\theta_{t_{\sqrt{\gamma}}}}(y_\star) - \pi_{\theta_{t_{\sqrt{\gamma}}}}(y_\star)^2\big)A_{\mathrm{G}}(y_\star;\theta_0) - \pi_{\theta_0}(y_{\mathrm{med}})^2 A_{\mathrm{G}}(y_{\mathrm{med}};\theta_\tau) \\
&\qquad\qquad + (\Delta_2-\gamma)\sum\nolimits_{z\in\mathcal{Y}_{\mathrm{bad}}} \pi_{\theta_\tau}(z)^2\Big) \\
&\le \pi_{\theta_\tau}(y_\star)\Big(\big(\pi_{\theta_{t_{\sqrt{\gamma}}}}(y_\star) - \pi_{\theta_{t_{\sqrt{\gamma}}}}(y_\star)^2\big)A_{\mathrm{G}}(y_\star;\theta_0) - \pi_{\theta_0}(y_{\mathrm{med}})^2 A_{\mathrm{G}}(y_{\mathrm{med}};\theta_\tau) \\
&\qquad\qquad + (\Delta_2-\gamma)\pi_{\theta_\tau}(\mathcal{Y}_{\mathrm{bad}})^2\Big) \\
&\le \pi_{\theta_\tau}(y_\star)\Big(\big(\pi_{\theta_{t_{\sqrt{\gamma}}}}(y_\star) - \pi_{\theta_{t_{\sqrt{\gamma}}}}(y_\star)^2\big)A_{\mathrm{G}}(y_\star;\theta_0) - (1/8)\pi_{\theta_0}(y_{\mathrm{med}})^2 A_{\mathrm{G}}(y_{\mathrm{med}};\theta_\tau)\Big) \\
&\le \pi_{\theta_\tau}(y_\star)\Big(\big(\pi_{\theta_{t_{\sqrt{\gamma}}}}(y_\star) - \pi_{\theta_{t_{\sqrt{\gamma}}}}(y_\star)^2\big)A_{\mathrm{G}}(y_\star;\theta_0) - (1/8)\pi_{\theta_0}(y_{\mathrm{med}})^2\gamma\Big) \\
&\le \pi_{\theta_\tau}(y_\star)\Big(\pi_{\theta_{t_{\sqrt{\gamma}}}}(y_\star)A_{\mathrm{G}}(y_\star;\theta_0) - (1/8)\pi_{\theta_0}(y_{\mathrm{med}})^2\gamma\Big)
\end{aligned}
$$

The first inequality is from $\pi_{\theta_\tau}(y_\star) - \pi_{\theta_\tau}(y_\star)^2 \le \pi_{\theta_{t_{\sqrt{\gamma}}}}(y_\star) - \pi_{\theta_{t_{\sqrt{\gamma}}}}(y_\star)^2$ (this holds since $\pi_{\theta_\tau}(y_\star) \le \pi_{\theta_{t_{\sqrt{\gamma}}}}(y_\star) \le \pi_{\theta_0}(y_\star) < 0.5$); the second inequality is from $A_{\mathrm{G}}(y_\star;\theta_\tau) \le A_{\mathrm{G}}(y_\star;\theta_0)$ and $-A_{\mathrm{G}}(z;\theta_\tau) \le \Delta_2 - \gamma$ for all $z \in \mathcal{Y}_{\mathrm{bad}}$; the third inequality is due to $\sum_{z\in\mathcal{Y}_{\mathrm{bad}}} \pi_{\theta_\tau}(z)^2 \le \pi_{\theta_\tau}(\mathcal{Y}_{\mathrm{bad}})^2$; the fourth inequality is from Equation (13); the fifth inequality is from $A_{\mathrm{G}}(y_{\mathrm{med}};\theta_\tau) \ge A_{\mathrm{G}}(y_{\mathrm{med}};\theta_{t_\gamma}) = \gamma$; and the last inequality is by $1 - \pi_{\theta_{t_{\sqrt{\gamma}}}}(y_\star) \le 1$.

Now, by Equation (12) and Item 2 in Lemma 4, we have

$$
\begin{aligned}
\pi_{\theta_{t_{\sqrt{\gamma}}}}(y_\star) &\le \pi_{\theta_0}(y_\star)\cdot(26\sqrt{\gamma}/r_{\mathrm{G}}(y_{\mathrm{med}})^2)^{1/7} \\
&\le \frac{\gamma^{13/14} r_{\mathrm{G}}(y_{\mathrm{med}})^{2/7}}{20(\Delta_1+\Delta_2)}\cdot\left(\frac{26\sqrt{\gamma}}{r_{\mathrm{G}}(y_{\mathrm{med}})^2}\right)^{1/7} \\
&< \frac{0.8^2\gamma^{13/14} r_{\mathrm{G}}(y_{\mathrm{med}})^{2/7}}{8\cdot 26^{1/7}(\Delta_1+\Delta_2)}\cdot\left(\frac{26\sqrt{\gamma}}{r_{\mathrm{G}}(y_{\mathrm{med}})^2}\right)^{1/7} \\
&\le \frac{\pi_{\theta_0}(y_{\mathrm{med}})^2\gamma}{8A_{\mathrm{G}}(y_\star;\theta_0)}.
\end{aligned}
$$

Therefore:

$$
\begin{aligned}
\frac{d}{dt}\pi_{\theta_t}(y_\star)\Big|_{t=\tau} &\le \pi_{\theta_\tau}(y_\star)\Big(\pi_{\theta_{t_{\sqrt{\gamma}}}}(y_\star)A_{\mathrm{G}}(y_\star;\theta_0) - (1/8)\pi_{\theta_0}(y_{\mathrm{med}})^2\gamma\Big) \\
&< \pi_{\theta_\tau}(y_\star)\left(\frac{\pi_{\theta_0}(y_{\mathrm{med}})^2\gamma}{8A_{\mathrm{G}}(y_\star;\theta_0)}\cdot A_{\mathrm{G}}(y_\star;\theta_0) - (1/8)\pi_{\theta_0}(y_{\mathrm{med}})^2\gamma\right) \\
&= 0,
\end{aligned}
$$

which contradicts the fact that τ is the initial time at which $\frac{d}{dt}\pi_{\theta_t}(y_\star)\big|_{t=\tau} \ge 0$. Hence, it must be that $\frac{d}{dt}\pi_{\theta_t}(y_\star) < 0$ for all $t \in [0, t_\gamma]$. Additionally, by Lemma 5 we know that $\pi_{\theta_t}(\mathcal{Y}_{\mathrm{bad}})$ is also decreasing in the time interval $[0, t_\gamma]$, and so $\frac{d}{dt}\pi_{\theta_t}(y_{\mathrm{med}}) > 0$ over $[0, t_\gamma]$.

To conclude, we showed that $\frac{d}{dt}\pi_{\theta_t}(y_\star) < 0$ and $\frac{d}{dt}\pi_{\theta_t}(y_{\text{med}}) > 0$ for all $t \in [0, t_\gamma]$. Thus, $\min_{t\in[0,t_\gamma]} \pi_{\theta_t}(y_\star) = \pi_{\theta_{t_\gamma}}(y_\star)$. Furthermore, since

$$\pi_{\theta_{t_{\sqrt{\gamma}}}}(y_\star) \leq \frac{\pi_{\theta_0}(y_{\text{med}})^2\gamma}{8A_{\text{G}}(y_\star;\theta_0)}$$

and $t_{\sqrt{\gamma}} < t_\gamma$, it holds that

$$\pi_{\theta_{t_\gamma}}(y_\star) < \pi_{\theta_{t_{\sqrt{\gamma}}}}(y_\star) \leq \frac{\pi_{\theta_0}(y_{\text{med}})^2\gamma}{8A_{\text{G}}(y_\star;\theta_0)},$$

completing the proof of Proposition 4. □

Based on Proposition 4, we can now complete the proof of Case I.

Proof of Theorem 2, Case I. By Proposition 4, as long as $V_{\text{G}}(\theta_t) \leq r_{\text{G}}(y_{\text{med}}) - \gamma$, we know that $\pi_{\theta_t}(y_\star)$ is monotonically decreasing and $\pi_{\theta_t}(y_{\text{med}})$ is monotonically increasing. Recall that t_γ denotes the initial time at which $V_{\text{G}}(\theta_t) \geq r_{\text{G}}(y_{\text{med}}) - \gamma$. Since $V_{\text{G}}(\theta_t)$ is continuous in t, at time t_γ it must be that $V_{\text{G}}(\theta_{t_\gamma}) = r_{\text{G}}(y_{\text{med}}) - \gamma$ and t_γ is the initial time at which this equality holds.

We first prove that at t_γ the policy is highly concentrated on y_{med} by upper bounding $1 - \pi_{\theta_{t_\gamma}}(y_{\text{med}})$. This follows from

$$V_{\text{G}}(\theta_{t_\gamma}) = \pi_{\theta_{t_\gamma}}(y_{\text{med}})r_{\text{G}}(y_{\text{med}}) + \pi_{\theta_{t_\gamma}}(y_\star)r_{\text{G}}(y_\star) + \sum\nolimits_{z\in\mathcal{Y}_{\text{bad}}} \pi_{\theta_{t_\gamma}}(z)r_{\text{G}}(z) = r_{\text{G}}(y_{\text{med}}) - \gamma,$$

which implies:

$$\begin{aligned} 1 - \pi_{\theta_{t_\gamma}}(y_{\text{med}}) &= \frac{1}{r_{\text{G}}(y_{\text{med}})}\Big(\gamma + \pi_{\theta_{t_\gamma}}(y_\star)r_{\text{G}}(y_\star) + \sum\nolimits_{z\in\mathcal{Y}_{\text{bad}}} \pi_{\theta_{t_\gamma}}(z)r_{\text{G}}(z)\Big) \\ &\leq \frac{1}{r_{\text{G}}(y_{\text{med}})}\left(\gamma + \frac{\gamma}{8\Delta_1}\right) \\ &= \frac{(1+8\Delta_1)\gamma}{8\Delta_1 r_{\text{G}}(y_{\text{med}})}. \end{aligned} \tag{14}$$

The inequality is due to $\pi_{\theta_{t_\gamma}}(y_\star) \leq \frac{\pi_{\theta_0}(y_{\text{med}})^2\gamma}{8A_{\text{G}}(y_\star;\theta_0)} \leq \frac{\gamma}{8\Delta_1}$ (Proposition 4).

Now, the fact that the policy assigns high probability to y_{med} at t_γ allows us to lower bound $t_\star$—the initial time at which $V_{\text{G}}(\theta_t) \geq r_{\text{G}}(y_\star) - \epsilon$. For any time $T \geq t_\gamma$, we have:

$$\begin{aligned} \|\theta_T - \theta_{t_\gamma}\| &= \left\| \int_{t_\gamma}^{T} \frac{d}{dt}\theta_t dt \right\| \\ &\leq \int_{t_\gamma}^{T} \left\| \frac{d}{dt}\theta_t \right\| dt \\ &= \int_{t_\gamma}^{T} \|\nabla V_{\text{G}}(\theta_t)\| dt \\ &\leq \int_{t_\gamma}^{T} 2(\Delta_1 + \Delta_2)(1 - \pi_{\theta_t}(y_{\text{med}}))dt, \end{aligned} \tag{15}$$

where the last inequality is by Lemma 11, with $B = 1$ since the feature vectors are orthonormal. We proceed by upper bounding the rate at which $1 - \pi_{\theta_t}(y_{\text{med}})$ increases (*i.e.*, the rate at which $\pi_{\theta_t}(y_{\text{med}})$ decreases). From the chain rule we get:

$$\begin{aligned} \frac{d}{dt}(1 - \pi_{\theta_t}(y_{\text{med}})) &= \left\langle \nabla(1 - \pi_{\theta_t}(y_{\text{med}})), \frac{d}{dt}\theta_t \right\rangle \\ &= \langle \nabla(1 - \pi_{\theta_t}(y_{\text{med}})), \nabla V_{\text{G}}(\theta_t) \rangle \\ &= \left\langle \nabla(\pi_{\theta_t}(y_\star) + \sum\nolimits_{z\in\mathcal{Y}_{\text{bad}}} \pi_{\theta_t}(z)), \nabla V_{\text{G}}(\theta_t) \right\rangle. \end{aligned}$$

For all $y \in \mathcal{Y}$, the gradient of $\pi_{\theta_t}(y)$ with respect to θ_t can be computed as follows:

$$\nabla \pi_{\theta_t}(y) = \pi_{\theta_t}(y)\Big(\phi(y) - \sum\nolimits_{z\in\mathcal{Y}} \pi_{\theta_t}(z)\phi(z)\Big) = \pi_{\theta_t}(y)\big(\phi(y) - \bar{\phi}_{\theta_t}\big),$$

where $\bar{\phi}_{\theta_t} := \sum_{z\in\mathcal{Y}} \pi_{\theta_t}(z)\phi(z)$. Thus:

$$\begin{aligned}
&\frac{d}{dt}(1 - \pi_{\theta_t}(y_{\mathrm{med}})) \\
&= \langle \nabla \pi_{\theta_t}(y_\star), \nabla V_{\mathrm{G}}(\theta_t)\rangle + \sum\nolimits_{z\in\mathcal{Y}_{\mathrm{bad}}} \langle \nabla \pi_{\theta_t}(z), \nabla V_{\mathrm{G}}(\theta_t)\rangle \\
&= \pi_{\theta_t}(y_\star)\left\langle \phi(y_\star) - \bar{\phi}_{\theta_t}, \nabla V_{\mathrm{G}}(\theta_t)\right\rangle + \sum\nolimits_{z\in\mathcal{Y}_{\mathrm{bad}}} \pi_{\theta_t}(z)\left\langle \phi(z) - \bar{\phi}_{\theta_t}, \nabla V_{\mathrm{G}}(\theta_t)\right\rangle \\
&\le (1 - \pi_{\theta_t}(y_{\mathrm{med}}))\sup\nolimits_{y\in\mathcal{Y}} \|\phi(y) - \bar{\phi}_{\theta_t}\|\|\nabla V_{\mathrm{G}}(\theta_t)\|. \qquad (16)
\end{aligned}$$

Notice that since the feature vectors are orthonormal, for any $y \in \mathcal{Y}$:

$$\begin{aligned}
\|\phi(y) - \bar{\phi}_{\theta_t}\|^2 &= 1 - 2\pi_{\theta_t}(y) + \sum\nolimits_{z\in\mathcal{Y}} \pi_{\theta_t}(z)^2 \\
&= (1 - \pi_{\theta_t}(y))^2 + \sum\nolimits_{z\in\mathcal{Y}\setminus\{y\}} \pi_{\theta_t}(z)^2 \\
&\le (1 - \pi_{\theta_t}(y))^2 + \left(\sum\nolimits_{z\in\mathcal{Y}\setminus\{y\}} \pi_{\theta_t}(z)\right)^2 \\
&= 2(1 - \pi_{\theta_t}(y))^2.
\end{aligned}$$

Thus, $\|\phi(y) - \bar{\phi}_{\theta_t}\| \le \sqrt{2}(1 - \pi_{\theta_t}(y)) \le \sqrt{2}$. Plugging this bound into Equation (16) and applying Lemma 11 with $B = 1$ to upper bound $\|\nabla V_{\mathrm{G}}(\theta_t)\|$ then yields:

$$\frac{d}{dt}(1 - \pi_{\theta_t}(y_{\mathrm{med}})) \le 2\sqrt{2}(\Delta_1 + \Delta_2)(1 - \pi_{\theta_t}(y_{\mathrm{med}}))^2.$$

This implies that

$$\frac{d}{dt}\left(\frac{1}{1 - \pi_{\theta_t}(y_{\mathrm{med}})}\right) = -\frac{\frac{d}{dt}(1 - \pi_{\theta_t}(y_{\mathrm{med}}))}{(1 - \pi_{\theta_t}(y_{\mathrm{med}}))^2} \ge -2\sqrt{2}(\Delta_1 + \Delta_2).$$

Integrating both sides from t_γ to t, for any $t \ge t_\gamma$, leads to:

$$\frac{1}{1 - \pi_{\theta_t}(y_{\mathrm{med}})} - \frac{1}{1 - \pi_{\theta_{t_\gamma}}(y_{\mathrm{med}})} \ge -2\sqrt{2}(\Delta_1 + \Delta_2)(t - t_\gamma).$$

Hence, for any $t \ge t_\gamma$ such that

$$2\sqrt{2}(\Delta_1 + \Delta_2)(1 - \pi_{\theta_{t_\gamma}}(y_{\mathrm{med}}))(t - t_\gamma) < 1,$$

it holds that

$$1 - \pi_{\theta_t}(y_{\mathrm{med}}) \le \frac{1 - \pi_{\theta_{t_\gamma}}(y_{\mathrm{med}})}{1 - 2\sqrt{2}(\Delta_1 + \Delta_2)(1 - \pi_{\theta_{t_\gamma}}(y_{\mathrm{med}}))(t - t_\gamma)}.$$

Consequently, for any $T \ge t_\gamma$ satisfying

$$2\sqrt{2}(\Delta_1 + \Delta_2)(1 - \pi_{\theta_{t_\gamma}}(y_{\mathrm{med}}))(T - t_\gamma) < 1,$$

we can plug this upper bound into Equation (15) and obtain:

$$\begin{aligned}
\|\theta_T - \theta_{t_\gamma}\| &\le 2(\Delta_1 + \Delta_2)\int_{t_\gamma}^{T} (1 - \pi_{\theta_t}(y_{\mathrm{med}}))dt \\
&= 2(\Delta_1 + \Delta_2)\int_0^{T - t_\gamma} \frac{1 - \pi_{\theta_{t_\gamma}}(y_{\mathrm{med}})}{1 - 2\sqrt{2}(\Delta_1 + \Delta_2)(1 - \pi_{\theta_{t_\gamma}}(y_{\mathrm{med}}))s} ds \\
&= \frac{1}{\sqrt{2}} \ln\left(\frac{1}{1 - 2\sqrt{2}(\Delta_1 + \Delta_2)(1 - \pi_{\theta_{t_\gamma}}(y_{\mathrm{med}}))(T - t_\gamma)}\right).
\end{aligned}$$

By Lemma 12, V_{G} is 1-Lipschitz with respect to θ, and so for any $T \geq t_\gamma$:

$$|V_{\mathrm{G}}(\theta_T) - V_{\mathrm{G}}(\theta_{t_\gamma})| \leq \|\theta_T - \theta_{t_\gamma}\| \leq \frac{1}{\sqrt{2}} \ln\left(\frac{1}{1 - 2\sqrt{2}(\Delta_1 + \Delta_2)(1 - \pi_{\theta_{t_\gamma}}(y_{\mathrm{med}}))(T - t_\gamma)}\right). \tag{17}$$

If $t_\star = \infty$, then the desired lower bound holds trivially. Hence, assume $t_\star < \infty$. By the definition of $t_\star$ and continuity of $V_{\mathrm{G}}(\theta_t)$, $V_{\mathrm{G}}(\theta_{t_\star}) = r_{\mathrm{G}}(y_\star) - \epsilon$. In particular, $t_\star > t_\gamma$ since $V_{\mathrm{G}}(\theta_{t_\star}) = r_{\mathrm{G}}(y_\star) - \epsilon > r_{\mathrm{G}}(y_{\mathrm{med}}) > V_{\mathrm{G}}(\theta_{t_\gamma})$. Therefore,

$$V_{\mathrm{G}}(\theta_{t_\star}) - V_{\mathrm{G}}(\theta_{t_\gamma}) = r_{\mathrm{G}}(y_\star) - \epsilon - r_{\mathrm{G}}(y_{\mathrm{med}}) + \gamma = \Delta_1 - \epsilon + \gamma \geq \Delta_1 - \epsilon.$$

If $2\sqrt{2}(\Delta_1 + \Delta_2)(1 - \pi_{\theta_{t_\gamma}}(y_{\mathrm{med}}))(t_\star - t_\gamma) \geq 1$, then directly

$$t_\star - t_\gamma \geq \frac{1}{2\sqrt{2}(\Delta_1 + \Delta_2)(1 - \pi_{\theta_{t_\gamma}}(y_{\mathrm{med}}))} \geq \frac{1 - e^{-\sqrt{2}(\Delta_1 - \epsilon)}}{2\sqrt{2}(\Delta_1 + \Delta_2)(1 - \pi_{\theta_{t_\gamma}}(y_{\mathrm{med}}))},$$

since $1 - e^{-\sqrt{2}(\Delta_1 - \epsilon)} \leq 1$. Otherwise, we apply the bound in Equation (17) with $T = t_\star$ to obtain:

$$\Delta_1 - \epsilon \leq V_{\mathrm{G}}(\theta_{t_\star}) - V_{\mathrm{G}}(\theta_{t_\gamma}) \leq \frac{1}{\sqrt{2}} \ln\left(\frac{1}{1 - 2\sqrt{2}(\Delta_1 + \Delta_2)(1 - \pi_{\theta_{t_\gamma}}(y_{\mathrm{med}}))(t_\star - t_\gamma)}\right),$$

which implies

$$t_\star - t_\gamma \geq \frac{1 - e^{-\sqrt{2}(\Delta_1 - \epsilon)}}{2\sqrt{2}(\Delta_1 + \Delta_2)(1 - \pi_{\theta_{t_\gamma}}(y_{\mathrm{med}}))}.$$

Therefore, in either case:

$$\begin{aligned}
t_\star &\geq t_\star - t_\gamma \\
&\geq \frac{1 - e^{-\sqrt{2}(\Delta_1 - \epsilon)}}{2\sqrt{2}(\Delta_1 + \Delta_2)(1 - \pi_{\theta_{t_\gamma}}(y_{\mathrm{med}}))} \\
&\geq \frac{2\sqrt{2}\Delta_1 r_{\mathrm{G}}(y_{\mathrm{med}})(1 - e^{-\sqrt{2}(\Delta_1 - \epsilon)})}{(\Delta_1 + \Delta_2)(1 + 8\Delta_1)\gamma} \\
&\geq \frac{2\sqrt{2}M^{14/13}\Delta_1 r_{\mathrm{G}}(y_{\mathrm{med}})(1 - e^{-\sqrt{2}(\Delta_1 - \epsilon)})}{(\Delta_1 + \Delta_2)(1 + 8\Delta_1)} \cdot \pi_{\theta_0}(y_\star)^{-14/13} \\
&= \Omega\Big(\pi_{\theta_0}(y_\star)^{-14/13}\Big),
\end{aligned}$$

where the second inequality follows from Equation (14). This completes the proof for Case I of Theorem 2. □

D.4.2 Proof of Case II

Suppose that gradient flow is used to maximize the expected reward with respect to r_{P}, which assigns y_{med} a low reward $r_{\mathrm{P}}(y_{\mathrm{med}}) = \min_{y \in \mathcal{Y}_{\mathrm{bad}}} r_{\mathrm{G}}(y)$. In this case, we show that $\pi_{\theta_t}(y_\star)$ increases already from $t = 0$ and that $t_\star^{\mathrm{no-med}} = \mathcal{O}(\pi_{\theta_0}(y_\star)^{-1})$, where recall that $t_\star^{\mathrm{no-med}}$ is the initial time at which $V_{\mathrm{G}}(\theta_t) \geq r_{\mathrm{G}}(y_\star) - \epsilon$ when maximizing r_{P}.

Proof of Theorem 2, Case II. Notice that $V_{\mathrm{G}}(\theta) \geq V_{\mathrm{P}}(\theta)$ for all $\theta \in \mathbb{R}^D$ since $r_{\mathrm{G}}(y) \geq r_{\mathrm{P}}(y)$ for all $y \in \mathcal{Y}$. Thus, when $V_{\mathrm{P}}(\theta_t) \geq r_{\mathrm{G}}(y_\star) - \epsilon$ it also holds that $V_{\mathrm{G}}(\theta_t) \geq r_{\mathrm{G}}(y_\star) - \epsilon$. We can therefore focus on V_{P} and consider the time it takes until it reaches a value of $r_{\mathrm{G}}(y_\star) - \epsilon$. In particular, denote by $t_{\mathrm{P}}^{\mathrm{no-med}}$ the initial time at which $V_{\mathrm{P}}(\theta_t) \geq r_{\mathrm{G}}(y_\star) - \epsilon$, *i.e.*:

$$t_{\mathrm{P}}^{\mathrm{no-med}} := \min\{t \geq 0 : V_{\mathrm{P}}(\theta_t) \geq r_{\mathrm{G}}(y_\star) - \epsilon\}.$$

The discussion above implies that $t_\star^{\mathrm{no-med}} \leq t_{\mathrm{P}}^{\mathrm{no-med}}$, and so it suffices to upper bound $t_{\mathrm{P}}^{\mathrm{no-med}}$. For simplicity of notation, denote $\rho := \frac{r_{\mathrm{G}}(y_\star) + 1 - \epsilon}{r_{\mathrm{G}}(y_\star) + 1}$. Our first claim is that when $\pi_{\theta_t}(y_\star) \geq \rho$, we

must have $V_{\mathrm{P}}(\theta_t) \geq r_{\mathrm{G}}(y_\star) - \epsilon$. This can be seen via the following calculation:

$$\begin{aligned} V_{\mathrm{P}}(\theta_t) &\geq \pi_{\theta_t}(y_\star) r_{\mathrm{P}}(y_\star) - (1 - \pi_{\theta_t}(y_\star)) \\ &= \pi_{\theta_t}(y_\star)(r_{\mathrm{G}}(y_\star) + 1) - 1 \\ &\geq r_{\mathrm{G}}(y_\star) - \epsilon. \end{aligned}$$

This implies that $\pi_{\theta_t}(y_\star) \leq \rho$ for all $t \in [0, t_{\mathrm{P}}^{\mathrm{no-med}}]$.

We prove that

$$T := \frac{1}{(1-\rho)^2 A_{\mathrm{P}}(y_\star; \theta_0)} \left(\frac{1}{\pi_{\theta_0}(y_\star)} - \frac{1}{\rho} \right) \geq t_{\mathrm{P}}^{\mathrm{no-med}},$$

from which the desired upper bound on $t_\star^{\mathrm{no-med}}$ immediately follows by $t_\star^{\mathrm{no-med}} \leq t_{\mathrm{P}}^{\mathrm{no-med}}$ and substituting the value of ρ.

Assume by way of contradiction that $t_{\mathrm{P}}^{\mathrm{no-med}} > T$. Let us lower bound the rate at which $\pi_{\theta_t}(y_\star)$ increases over $[0, t_{\mathrm{P}}^{\mathrm{no-med}}]$. Notice first that $V_{\mathrm{P}}(\theta_0) > r_{\mathrm{P}}(y)$ for all $y \in \mathcal{Y} \setminus \{y_\star\}$ since all such y have the same proxy reward value, which is lower than $r_{\mathrm{P}}(y_\star) = r_{\mathrm{G}}(y_\star)$. Now, consider the time derivative of $\pi_{\theta_t}(y_\star)$, starting from the expression derived in Lemma 13:

$$\begin{aligned} &\frac{d}{dt} \pi_{\theta_t}(y_\star) \\ &= \pi_{\theta_t}(y_\star) \Big(\pi_{\theta_t}(y_\star) A_{\mathrm{P}}(y_\star; \theta_t) - \sum\nolimits_{y \in \mathcal{Y}} \pi_{\theta_t}(y)^2 A_{\mathrm{P}}(y; \theta_t) \Big) \\ &= \pi_{\theta_t}(y_\star) \Big(\big(\pi_{\theta_t}(y_\star) - \pi_{\theta_t}(y_\star)^2 \big) A_{\mathrm{P}}(y_\star; \theta_t) \\ &\qquad - \pi_{\theta_t}(y_{\mathrm{med}})^2 A_{\mathrm{P}}(y_{\mathrm{med}}; \theta_t) - \sum_{z \in \mathcal{Y}_{\mathrm{bad}}} \pi_{\theta_t}(z)^2 A_{\mathrm{P}}(z; \theta_t) \Big) \\ &\geq \pi_{\theta_t}(y_\star)^2 (1 - \pi_{\theta_t}(y_\star)) A_{\mathrm{P}}(y_\star; \theta_t). \end{aligned}$$

Here, the inequality is from $A_{\mathrm{P}}(y; \theta_t) \leq A_{\mathrm{P}}(y; \theta_0) \leq 0$ for all $y \in \mathcal{Y} \setminus \{y_\star\}$. As we proved above, $\pi_{\theta_t}(y_\star) \leq \rho$ for all $t \in [0, t_{\mathrm{P}}^{\mathrm{no-med}}]$. Thus for such t it holds that:

$$\frac{d}{dt} \pi_{\theta_t}(y_\star) \geq (1-\rho) A_{\mathrm{P}}(y_\star; \theta_t) \pi_{\theta_t}(y_\star)^2.$$

Next, we show that $A_{\mathrm{P}}(y_\star; \theta_t) \geq (1-\rho) A_{\mathrm{P}}(y_\star; \theta_0)$. To prove this, observe that since $r_{\mathrm{P}}(y_\star) = r_{\mathrm{G}}(y_\star)$ and $V_{\mathrm{P}}(\theta_0) \geq -1$,

$$\frac{r_{\mathrm{G}}(y_\star) - \epsilon - V_{\mathrm{P}}(\theta_0)}{r_{\mathrm{P}}(y_\star) - V_{\mathrm{P}}(\theta_0)} = 1 - \frac{\epsilon}{r_{\mathrm{P}}(y_\star) - V_{\mathrm{P}}(\theta_0)} \leq 1 - \frac{\epsilon}{r_{\mathrm{P}}(y_\star) + 1} = \frac{r_{\mathrm{G}}(y_\star) + 1 - \epsilon}{r_{\mathrm{G}}(y_\star) + 1} = \rho.$$

Using this fact, we have that:

$$\frac{A_{\mathrm{P}}(y_\star; \theta_t)}{A_{\mathrm{P}}(y_\star; \theta_0)} = \frac{r_{\mathrm{P}}(y_\star) - V_{\mathrm{P}}(\theta_t)}{r_{\mathrm{P}}(y_\star) - V_{\mathrm{P}}(\theta_0)} = 1 - \frac{V_{\mathrm{P}}(\theta_t) - V_{\mathrm{P}}(\theta_0)}{r_{\mathrm{P}}(y_\star) - V_{\mathrm{P}}(\theta_0)} \geq 1 - \frac{r_{\mathrm{G}}(y_\star) - \epsilon - V_{\mathrm{P}}(\theta_0)}{r_{\mathrm{P}}(y_\star) - V_{\mathrm{P}}(\theta_0)} \geq 1 - \rho.$$

Therefore, we can plug in this bound to get:

$$\frac{d}{dt} \pi_{\theta_t}(y_\star) \geq (1-\rho)^2 A_{\mathrm{P}}(y_\star; \theta_0) \pi_{\theta_t}(y_\star)^2.$$

Dividing both sides by $\pi_{\theta_t}(y_\star)^2$, integrating from 0 to t, and rearranging terms results in the following lower bound:

$$\pi_{\theta_t}(y_\star) \geq \frac{\pi_{\theta_0}(y_\star)}{1 - (1-\rho)^2 A_{\mathrm{P}}(y_\star; \theta_0) \pi_{\theta_0}(y_\star) \cdot t}.$$

At time $t = T$ for

$$T = \frac{1}{(1-\rho)^2 A_{\mathrm{P}}(y_\star; \theta_0)} \left(\frac{1}{\pi_{\theta_0}(y_\star)} - \frac{1}{\rho} \right),$$

this lower bound on $\pi_{\theta_T}(y_\star)$ implies that $\pi_{\theta_T}(y_\star) \geq \rho$, and so $V_{\mathrm{P}}(\theta_T) \geq r_{\mathrm{G}}(y_\star) - \epsilon$, in contradiction to our assumption that $t_{\mathrm{P}}^{\mathrm{no-med}} > T$. Therefore,

$$t_\star^{\mathrm{no-med}} \leq t_{\mathrm{P}}^{\mathrm{no-med}} \leq T \leq \frac{(r_{\mathrm{G}}(y_\star) + 1)^2}{\epsilon^2 (r_{\mathrm{P}}(y_\star) - V_{\mathrm{P}}(\theta_0))} \cdot \pi_{\theta_0}(y_\star)^{-1} = \mathcal{O}\big(\pi_{\theta_0}(y_\star)^{-1}\big).$$

Moreover, notice that $V_{\mathrm{P}}(\theta_t)$ is monotonically non-decreasing when maximizing r_{P} via gradient flow since $\frac{d}{dt}V_{\mathrm{P}}(\theta_t) = \|\nabla V_{\mathrm{P}}(\theta_t)\|^2 \geq 0$ for all $t \geq 0$. Thus, for all $t \geq T \geq t_{\mathrm{P}}^{\mathrm{no-med}}$:

$$\epsilon \geq r_{\mathrm{G}}(y_\star) - V_{\mathrm{P}}(\theta_t) = (1 - \pi_{\theta_t}(y_\star))(\Delta_1 + \Delta_2),$$

from which we can conclude that for all $t \geq T$:

$$\pi_{\theta_t}(y_\star) \geq 1 - \frac{\epsilon}{\Delta_1 + \Delta_2}.$$

□

D.5 Proof of Theorem 3

For convenience of notation, we start by introducing the following helper variables:

$$\gamma := \left(\frac{\pi_{\theta_0}(y_\star)}{M'}\right)^{14/13} \quad , \quad s := \langle \phi(y_\star), \phi(y_{\mathrm{med}}) \rangle \quad , \quad h(s) := \frac{\|\phi(y_{\mathrm{med}})\|^2 - s}{7\big(\|\phi(y_{\mathrm{med}})\|^2 - s/2\big)}, \tag{18}$$

where recall that $M' := \min\big\{1, \|\phi(y_{\mathrm{med}})\|^2 r_{\mathrm{G}}(y_{\mathrm{med}})^{2/7}(40B^2(\Delta_1 + \Delta_2))^{-1}\big\}$, with $\Delta_1 := r_{\mathrm{G}}(y_\star) - r_{\mathrm{G}}(y_{\mathrm{med}})$, $\Delta_2 := r_{\mathrm{G}}(y_{\mathrm{med}}) - r_{\mathrm{G}}(y_{\mathrm{bad}})$ for some $y_{\mathrm{bad}} \in \mathcal{Y}_{\mathrm{bad}}$, and $B := \max_{y \in \mathcal{Y}} \|\phi(y)\|$. We can then straightforwardly state the conditions of Assumption 3 in terms of γ.

Lemma 6. *Under Assumptions 1 and 3, the following properties hold:*

1. $\gamma \leq \min\left\{0.01, \left(\frac{21\|\phi(y_{\mathrm{med}})\|^2}{1100B^2(\Delta_1+\Delta_2)}\right)^{7/3}\right\}$;
2. $\pi_{\theta_0}(y_\star) \leq \min\left\{\gamma^{13/14}, \frac{\|\phi(y_{\mathrm{med}})\|^2 r_{\mathrm{G}}(y_{\mathrm{med}})^{2/7}}{40B^2(\Delta_1+\Delta_2)} \cdot \gamma^{13/14}, \frac{\|\phi(y_{\mathrm{med}})\|}{\|\phi(y_\star)\|}\pi_{\theta_0}(y_{\mathrm{med}})\right\}$; *and*
3. $\max\left\{\frac{2\sqrt{\gamma}}{\Delta_2}, 0.05 \cdot r_{\mathrm{G}}(y_{\mathrm{med}})\right\} \leq \pi_{\theta_0}(\mathcal{Y}_{\mathrm{bad}}) \leq 0.1 \cdot r_{\mathrm{G}}(y_{\mathrm{med}})$.

Proof. All three conditions are obtained directly from Assumption 3 by substituting the definition of M' and noticing that $\pi_{\theta_0}(y_\star) = M'\gamma^{13/14}$. Note that Assumption 1 on the ground truth reward structure ensures Δ_1 and Δ_2 are positive, so all quantities in the conditions above are well-defined. □

D.5.1 Proof of Case I

Suppose that gradient flow is used to maximize the expected reward with respect to r_{G}. We define t_γ as the initial time at which $V_{\mathrm{G}}(\theta_t)$ reaches $r_{\mathrm{G}}(y_{\mathrm{med}}) - C\gamma^{13/14+h(s)/2}$, *i.e.*:

$$t_\gamma := \min\Big\{t \geq 0 : V_{\mathrm{G}}(\theta_t) \geq r_{\mathrm{G}}(y_{\mathrm{med}}) - C\gamma^{13/14+h(s)/2}\Big\}, \tag{19}$$

where $C := \big(26/r_{\mathrm{G}}(y_{\mathrm{med}})^2\big)^{13\alpha/7}$ for $\alpha := -s/\big(26(\|\phi(y_{\mathrm{med}})\|^2 - 0.5s)\big) \in (0, 1/13)$. As in the proof of Theorem 2, the key step is to show that on $[0, t_\gamma]$ the probability of $y_\star$ decreases while the probability of y_{med} increases. We begin with a monotonicity lemma showing that, whenever $V_{\mathrm{G}}(\theta_t) < r_{\mathrm{G}}(y_{\mathrm{med}})$, a decrease in $\pi_{\theta_t}(y_\star)$ implies an increase in $\pi_{\theta_t}(y_{\mathrm{med}})$. We then prove that $\pi_{\theta_t}(y_\star)$ indeed decreases throughout $[0, t_\gamma]$, which in turn yields monotonic growth of $\pi_{\theta_t}(y_{\mathrm{med}})$ on this interval. The negative inner product between $\phi(y_\star)$ and $\phi(y_{\mathrm{med}})$ strengthens this effect relative to the case of orthonormal features (Theorem 2) by pulling probability away from $y_\star$ more aggressively and concentrating the policy more strongly around y_{med}. Combining this stronger concentration with the gradient-norm bound from Lemma 11 will allow establishing the desired lower bound on $t_\star$.

Note that if $V_{\mathrm{G}}(\theta_t) < r_{\mathrm{G}}(y_{\mathrm{med}})$ for all $t \geq 0$, then the lower bound on $t_\star$ holds trivially, since $V_{\mathrm{G}}(\theta_t)$ never reaches $r_{\mathrm{G}}(y_\star) - \epsilon$. Thus, throughout the proof of this case, we may assume that there exists a time at which $V_{\mathrm{G}}(\theta_t) \geq r_{\mathrm{G}}(y_{\mathrm{med}})$.

Lemma 7. *At any time $t \geq 0$ such that $V_{\mathrm{G}}(\theta_t) < r_{\mathrm{G}}(y_{\mathrm{med}})$, if $\frac{d}{dt}\pi_{\theta_t}(y_\star) < 0$, then $\frac{d}{dt}\pi_{\theta_t}(y_{\mathrm{med}}) > 0$.*

Proof. Fix any time $t \geq 0$ such that $V_{\mathrm{G}}(\theta_t) < r_{\mathrm{G}}(y_{\mathrm{med}})$. By Lemma 13, instantiated with $r = r_{\mathrm{G}}$ and $y = y_\star$, we have

$$\frac{d}{dt}\pi_{\theta_t}(y_\star) = \pi_{\theta_t}(y_\star)\left\langle \phi(y_\star) - \sum\nolimits_{y\in\mathcal{Y}} \pi_{\theta_t}(y)\phi(y), \sum\nolimits_{y\in\mathcal{Y}} \pi_{\theta_t}(y) A_{\mathrm{G}}(y;\theta_t)\phi(y)\right\rangle,$$

Since $\langle \phi(z), \phi(y)\rangle = 0$ for all $z \in \mathcal{Y}_{\mathrm{bad}}, y \in \mathcal{Y}\setminus\{z\}$ (Assumption 4), we may write:

$$\begin{aligned}
\frac{d}{dt}\pi_{\theta_t}(y_\star) &= \pi_{\theta_t}(y_\star)\Big[\pi_{\theta_t}(y_\star)A_{\mathrm{G}}(y_\star;\theta_t)\big(\|\phi(y_\star)\|^2 - \|\phi(y_\star)\|^2\pi_{\theta_t}(y_\star) - s\pi_{\theta_t}(y_{\mathrm{med}})\big) \\
&\qquad + \pi_{\theta_t}(y_{\mathrm{med}})A_{\mathrm{G}}(y_{\mathrm{med}};\theta_t)\big(s - \|\phi(y_{\mathrm{med}})\|^2\pi_{\theta_t}(y_{\mathrm{med}}) - s\pi_{\theta_t}(y_\star)\big) \\
&\qquad - \sum\nolimits_{z\in\mathcal{Y}_{\mathrm{bad}}} \pi_{\theta_t}(z)^2 A_{\mathrm{G}}(z;\theta_t)\|\phi(z)\|^2\Big] \\
&\geq \pi_{\theta_t}(y_\star)\Big[\pi_{\theta_t}(y_\star)A_{\mathrm{G}}(y_\star;\theta_t)\big(\|\phi(y_\star)\|^2 - \|\phi(y_\star)\|^2\pi_{\theta_t}(y_\star) - s\pi_{\theta_t}(y_{\mathrm{med}})\big) \\
&\qquad + \pi_{\theta_t}(y_{\mathrm{med}})A_{\mathrm{G}}(y_{\mathrm{med}};\theta_t)\big(s - \|\phi(y_{\mathrm{med}})\|^2\pi_{\theta_t}(y_{\mathrm{med}}) - s\pi_{\theta_t}(y_\star)\big)\Big].
\end{aligned}$$

Here, the inequality is due to $A_{\mathrm{G}}(z;\theta_t) \leq 0$ for all $z \in \mathcal{Y}_{\mathrm{bad}}$. Thus, when $\frac{d}{dt}\pi_{\theta_t}(y_\star) < 0$, it must hold that:

$$\begin{aligned}
&\pi_{\theta_t}(y_\star)A_{\mathrm{G}}(y_\star;\theta_t)\big(\|\phi(y_\star)\|^2 - \|\phi(y_\star)\|^2\pi_{\theta_t}(y_\star) - s\pi_{\theta_t}(y_{\mathrm{med}})\big) \\
&< \pi_{\theta_t}(y_{\mathrm{med}})A_{\mathrm{G}}(y_{\mathrm{med}};\theta_t)\big(-s + \|\phi(y_{\mathrm{med}})\|^2\pi_{\theta_t}(y_{\mathrm{med}}) + s\pi_{\theta_t}(y_\star)\big),
\end{aligned}$$

or equivalently,

$$\frac{\pi_{\theta_t}(y_\star)A_{\mathrm{G}}(y_\star;\theta_t)}{\pi_{\theta_t}(y_{\mathrm{med}})A_{\mathrm{G}}(y_{\mathrm{med}};\theta_t)} < \frac{-s + \|\phi(y_{\mathrm{med}})\|^2\pi_{\theta_t}(y_{\mathrm{med}}) + s\pi_{\theta_t}(y_\star)}{\|\phi(y_\star)\|^2 - \|\phi(y_\star)\|^2\pi_{\theta_t}(y_\star) - s\pi_{\theta_t}(y_{\mathrm{med}})}, \tag{20}$$

since $\|\phi(y_\star)\|^2 - \|\phi(y_\star)\|^2\pi_{\theta_t}(y_\star) - s\pi_{\theta_t}(y_{\mathrm{med}}) > 0$ and $\pi_{\theta_t}(y_{\mathrm{med}})A_{\mathrm{G}}(y_{\mathrm{med}};\theta_t) > 0$. We continue by upper bounding the right-hand side in the inequality above via the following claim:

$$\frac{-s + \|\phi(y_{\mathrm{med}})\|^2\pi_{\theta_t}(y_{\mathrm{med}}) + s\pi_{\theta_t}(y_\star)}{\|\phi(y_\star)\|^2 - \|\phi(y_\star)\|^2\pi_{\theta_t}(y_\star) - s\pi_{\theta_t}(y_{\mathrm{med}})} \leq \frac{\|\phi(y_{\mathrm{med}})\|^2 - \|\phi(y_{\mathrm{med}})\|^2\pi_{\theta_t}(y_{\mathrm{med}}) - s\pi_{\theta_t}(y_\star)}{\|\phi(y_\star)\|^2\pi_{\theta_t}(y_\star) - s + s\pi_{\theta_t}(y_{\mathrm{med}})}. \tag{21}$$

Observe that the denominators of both sides are positive. Hence, upon multiplying both sides by the denominators and subtracting common terms, to show that Equation (21) holds it suffices to prove:

$$s^2(1 - \pi_{\theta_t}(y_{\mathrm{med}}) - \pi_{\theta_t}(y_\star)) \leq \|\phi(y_\star)\|^2\|\phi(y_{\mathrm{med}})\|^2(1 - \pi_{\theta_t}(y_{\mathrm{med}}) - \pi_{\theta_t}(y_\star)).$$

Indeed, this is true because $s^2 = \langle \phi(y_\star), \phi(y_{\mathrm{med}})\rangle^2 \leq \|\phi(y_{\mathrm{med}})\|^2\|\phi(y_\star)\|^2$. Combining Equation (20) and Equation (21), we thus get:

$$\begin{aligned}
\frac{\pi_{\theta_t}(y_\star)A_{\mathrm{G}}(y_\star;\theta_t)}{\pi_{\theta_t}(y_{\mathrm{med}})A_{\mathrm{G}}(y_{\mathrm{med}};\theta_t)} &< \frac{-s + \|\phi(y_{\mathrm{med}})\|^2\pi_{\theta_t}(y_{\mathrm{med}}) + s\pi_{\theta_t}(y_\star)}{\|\phi(y_\star)\|^2 - \|\phi(y_\star)\|^2\pi_{\theta_t}(y_\star) - s\pi_{\theta_t}(y_{\mathrm{med}})} \\
&\leq \frac{\|\phi(y_{\mathrm{med}})\|^2 - \|\phi(y_{\mathrm{med}})\|^2\pi_{\theta_t}(y_{\mathrm{med}}) - s\pi_{\theta_t}(y_\star)}{\|\phi(y_\star)\|^2\pi_{\theta_t}(y_\star) - s + s\pi_{\theta_t}(y_{\mathrm{med}})},
\end{aligned}$$

which implies:

$$\begin{aligned}
&\pi_{\theta_t}(y_{\mathrm{med}})A_{\mathrm{G}}(y_{\mathrm{med}};\theta_t)(\|\phi(y_{\mathrm{med}})\|^2 - \|\phi(y_{\mathrm{med}})\|^2\pi_{\theta_t}(y_{\mathrm{med}}) - s\pi_{\theta_t}(y_\star)) \\
&- \pi_{\theta_t}(y_\star)A_{\mathrm{G}}(y_\star;\theta_t)(\|\phi(y_\star)\|^2\pi_{\theta_t}(y_\star) - s + s\pi_{\theta_t}(y_{\mathrm{med}})) > 0.
\end{aligned} \tag{22}$$

Next, we analyze the time derivative of $\pi_{\theta_t}(y_{\mathrm{med}})$, which is (see Lemma 13):

$$\begin{aligned}
&\frac{d}{dt}\pi_{\theta_t}(y_{\mathrm{med}}) \\
&= \pi_{\theta_t}(y_{\mathrm{med}})\Big[\pi_{\theta_t}(y_{\mathrm{med}})A_{\mathrm{G}}(y_{\mathrm{med}};\theta_t)\big(\|\phi(y_{\mathrm{med}})\|^2 - \|\phi(y_{\mathrm{med}})\|^2\pi_{\theta_t}(y_{\mathrm{med}}) - s\pi_{\theta_t}(y_\star)\big) \\
&\qquad + \pi_{\theta_t}(y_\star)A_{\mathrm{G}}(y_\star;\theta_t)\big(s - \|\phi(y_\star)\|^2\pi_{\theta_t}(y_\star) - s\pi_{\theta_t}(y_{\mathrm{med}})\big) \\
&\qquad - \sum\nolimits_{z\in\mathcal{Y}_{\mathrm{bad}}} \pi_{\theta_t}(z)^2 A_{\mathrm{G}}(z;\theta_t)\|\phi(z)\|^2\Big] \\
&\geq \pi_{\theta_t}(y_{\mathrm{med}})\Big[\pi_{\theta_t}(y_{\mathrm{med}})A_{\mathrm{G}}(y_{\mathrm{med}};\theta_t)\big(\|\phi(y_{\mathrm{med}})\|^2 - \|\phi(y_{\mathrm{med}})\|^2\pi_{\theta_t}(y_{\mathrm{med}}) - s\pi_{\theta_t}(y_\star)\big) \\
&+ \pi_{\theta_t}(y_\star)A_{\mathrm{G}}(y_\star;\theta_t)\big(s - \|\phi(y_\star)\|^2\pi_{\theta_t}(y_\star) - s\pi_{\theta_t}(y_{\mathrm{med}})\big)\Big] \\
&> 0.
\end{aligned}$$

Here, the first inequality is due to $A_{\mathrm{G}}(z;\theta_t) \le 0$ for all $z \in \mathcal{Y}_{\mathrm{bad}}$, and the second inequality is from Equation (22). Putting it all together, we have shown that if $V_{\mathrm{G}}(\theta_t) < r_{\mathrm{G}}(y_{\mathrm{med}})$ and $\frac{d}{dt}\pi_{\theta_t}(y_\star) < 0$, then necessarily $\frac{d}{dt}\pi_{\theta_t}(y_{\mathrm{med}}) > 0$. □

Using Lemma 7, we can now prove that until t_γ the probability of y_{med} monotonically increases while the probability of $y_\star$ monotonically decreases, and characterize how small $\pi_{\theta_t}(y_\star)$ becomes. This is a core component in the proof of Theorem 3.

Proposition 5. *For all $t \in [0, t_\gamma]$ it holds that:*

- $\frac{d}{dt}\pi_{\theta_t}(y_\star) < 0$*; and*
- $\frac{d}{dt}\pi_{\theta_t}(y_{\mathrm{med}}) > 0$.

Furthermore,

$$\min_{t\in[0,t_\gamma]} \pi_{\theta_t}(y_\star) = \pi_{\theta_{t_\gamma}}(y_\star) \le \frac{C\|\phi(y_{\mathrm{med}})\|^2 \pi_{\theta_0}(y_{\mathrm{med}})^2 \gamma^{\frac{13}{14}+\frac{h(s)}{2}}}{8(\|\phi(y_\star)\|^2 - s) A_{\mathrm{G}}(y_\star;\theta_0)},$$

where $C := \left(26/r_{\mathrm{G}}(y_{\mathrm{med}})^2\right)^{13\alpha/7}$, as introduced in Equation (19), and $s := \langle \phi(y_\star), \phi(y_{\mathrm{med}})\rangle$ and $h(s) := \frac{\|\phi(y_{\mathrm{med}})\|^2 - s}{7\left(\|\phi(y_{\mathrm{med}})\|^2 - s/2\right)}$ are as defined in Equation (18).

Proof. Since t_γ is the initial time at which $V_{\mathrm{G}}(\theta_t) \ge r_{\mathrm{G}}(y_{\mathrm{med}}) - C\gamma^{13/14+h(s)/2}$, Lemma 7 implies that for all $t \in [0, t_\gamma]$, if $\frac{d}{dt}\pi_{\theta_t}(y_\star) < 0$, then $\frac{d}{dt}\pi_{\theta_t}(y_{\mathrm{med}}) > 0$. Hence, for establishing that $\frac{d}{dt}\pi_{\theta_t}(y_\star) < 0$ and $\frac{d}{dt}\pi_{\theta_t}(y_{\mathrm{med}}) > 0$ over $[0, t_\gamma]$, it suffices to prove that $\frac{d}{dt}\pi_{\theta_t}(y_\star) < 0$.

To that end, we let $t_{\sqrt{\gamma}}$ be the initial time at which $V_{\mathrm{G}}(\theta_t) \ge r_{\mathrm{G}}(y_{\mathrm{med}}) - \sqrt{\gamma}$, *i.e.*:

$$t_{\sqrt{\gamma}} := \min\{t \ge 0 : V_{\mathrm{G}}(\theta_t) \ge r_{\mathrm{G}}(y_{\mathrm{med}}) - \sqrt{\gamma}\}.$$

From Lemma 15, $V_{\mathrm{G}}(\theta_0) < r_{\mathrm{G}}(y_{\mathrm{med}}) - \sqrt{\gamma}$, so $t_{\sqrt{\gamma}} > 0$. Furthermore, since $V_{\mathrm{G}}(\theta_t)$ is continuous in t, at $t_{\sqrt{\gamma}}$ it holds that $V_{\mathrm{G}}(\theta_{t_{\sqrt{\gamma}}}) = r_{\mathrm{G}}(y_{\mathrm{med}}) - \sqrt{\gamma}$. The proof proceeds by showing that $\frac{d}{dt}\pi_{\theta_t}(y_\star) < 0$ up to time $t_{\sqrt{\gamma}}$ and upper bounding $\pi_{\theta_{t_{\sqrt{\gamma}}}}(y_\star)$. Lastly, we will establish that $\pi_{\theta_t}(y_\star)$ continues decreasing until time t_γ.

Step 1: monotonicity in $[0, t_{\sqrt{\gamma}}]$. We show that $\frac{d}{dt}\pi_{\theta_t}(y_\star) < 0$, and so $\frac{d}{dt}\pi_{\theta_t}(y_{\mathrm{med}}) > 0$, for all $t \in [0, t_{\sqrt{\gamma}}]$. Assume by way of contradiction that there exists a time $t' \le t_{\sqrt{\gamma}}$ at which $\frac{d}{dt}\pi_{\theta_t}(y_\star)\big|_{t=t'} \ge 0$ and denote by τ the initial such time, *i.e.*:

$$\tau := \inf\left\{t \in [0, t_{\sqrt{\gamma}}] : \tfrac{d}{dt}\pi_{\theta_t}(y_\star) \ge 0\right\}.$$

For all $t \in [0, \tau)$, we therefore have that $\frac{d}{dt}\pi_{\theta_t}(y_\star) < 0$ and $\frac{d}{dt}\pi_{\theta_t}(y_{\mathrm{med}}) > 0$, which imply $\pi_{\theta_\tau}(y_\star) \le \pi_{\theta_0}(y_\star)$ and $\pi_{\theta_\tau}(y_{\mathrm{med}}) \ge \pi_{\theta_0}(y_{\mathrm{med}})$, respectively.

We now show that $\frac{d}{dt}\pi_{\theta_t}(\mathcal{Y}_{\mathrm{bad}}) < 0$ for all $t \in [0, \tau)$. By Lemma 13, the time derivative of $\pi_{\theta_t}(z)$ for $z \in \mathcal{Y}_{\mathrm{bad}}$ can be written as:

$$\begin{aligned}\frac{d}{dt}\pi_{\theta_t}(z) = \pi_{\theta_t}(z)^2 A_{\mathrm{G}}(z;\theta_t)\|\phi(z)\|^2 \\ - \pi_{\theta_t}(z)\Big(\sum\nolimits_{z'\in\mathcal{Y}_{\mathrm{bad}}} \pi_{\theta_t}(z')^2 A_{\mathrm{G}}(z';\theta_t)\|\phi(z')\|^2 \\ + \pi_{\theta_t}(y_{\mathrm{med}}) A_{\mathrm{G}}(y_{\mathrm{med}};\theta_t)\left(\|\phi(y_{\mathrm{med}})\|^2 \pi_{\theta_t}(y_{\mathrm{med}}) + s\pi_{\theta_t}(y_\star)\right) \\ + \pi_{\theta_t}(y_\star) A_{\mathrm{G}}(y_\star;\theta_t)\left(\|\phi(y_\star)\|^2 \pi_{\theta_t}(y_\star) + s\pi_{\theta_t}(y_{\mathrm{med}})\right)\Big).\end{aligned}$$

Summing over all $z \in \mathcal{Y}_{\text{bad}}$, we have:

$$\frac{d}{dt}\pi_{\theta_t}(\mathcal{Y}_{\text{bad}}) = \underbrace{\Big(\sum\nolimits_{z\in\mathcal{Y}_{\text{bad}}} \pi_{\theta_t}(z)^2 A_{\text{G}}(z;\theta_t)\|\phi(z)\|^2\Big)\big(1-\pi_{\theta_t}(\mathcal{Y}_{\text{bad}})\big)}_{I_1}$$
$$- \pi_{\theta_t}(\mathcal{Y}_{\text{bad}})\Big(\pi_{\theta_t}(y_{\text{med}})A_{\text{G}}(y_{\text{med}};\theta_t)(\|\phi(y_{\text{med}})\|^2\pi_{\theta_t}(y_{\text{med}}) + s\pi_{\theta_t}(y_\star))$$
$$+ \pi_{\theta_t}(y_\star)A_{\text{G}}(y_\star;\theta_t)(\|\phi(y_\star)\|^2\pi_{\theta_t}(y_\star) + s\pi_{\theta_t}(y_{\text{med}}))\Big),$$

where we denote by I_2 the second term in the expression above (excluding the minus sign). Since all outputs in $\mathcal{Y}_{\text{bad}}$ have the minimal ground truth reward, we have $A_{\text{G}}(z;\theta_t) \le 0$ for all $z \in \mathcal{Y}_{\text{bad}}$. Together with $1-\pi_{\theta_t}(\mathcal{Y}_{\text{bad}}) \ge 0$, this gives $I_1 \le 0$. We now show that $I_2 > 0$. It suffices to prove that:

$$\begin{aligned} 0 < {} & \pi_{\theta_t}(y_{\text{med}})A_{\text{G}}(y_{\text{med}};\theta_t)(\|\phi(y_{\text{med}})\|^2\pi_{\theta_t}(y_{\text{med}}) + s\pi_{\theta_t}(y_\star)) \\ & + \pi_{\theta_t}(y_\star)A_{\text{G}}(y_\star;\theta_t)(\|\phi(y_\star)\|^2\pi_{\theta_t}(y_\star) + s\pi_{\theta_t}(y_{\text{med}})). \end{aligned} \tag{23}$$

If $\|\phi(y_\star)\|^2\pi_{\theta_t}(y_\star) + s\pi_{\theta_t}(y_{\text{med}}) \ge 0$, which is equivalent to the second term on the right-hand side being non-negative, then:

$$\begin{aligned} & \pi_{\theta_t}(y_{\text{med}})A_{\text{G}}(y_{\text{med}};\theta_t)(\|\phi(y_{\text{med}})\|^2\pi_{\theta_t}(y_{\text{med}}) + s\pi_{\theta_t}(y_\star)) \\ & \quad + \pi_{\theta_t}(y_\star)A_{\text{G}}(y_\star;\theta_t)(\|\phi(y_\star)\|^2\pi_{\theta_t}(y_\star) + s\pi_{\theta_t}(y_{\text{med}})) \\ & \ge \pi_{\theta_t}(y_{\text{med}})A_{\text{G}}(y_{\text{med}};\theta_t)(\|\phi(y_{\text{med}})\|^2\pi_{\theta_t}(y_{\text{med}}) + s\pi_{\theta_t}(y_\star)) \\ & \ge \pi_{\theta_t}(y_{\text{med}})A_{\text{G}}(y_{\text{med}};\theta_t)(\|\phi(y_{\text{med}})\|^2\pi_{\theta_0}(y_{\text{med}}) + s\pi_{\theta_0}(y_\star)) \\ & \ge \pi_{\theta_t}(y_{\text{med}})A_{\text{G}}(y_{\text{med}};\theta_t)\|\phi(y_{\text{med}})\|\big(\|\phi(y_{\text{med}})\|\pi_{\theta_0}(y_{\text{med}}) - \|\phi(y_\star)\|\pi_{\theta_0}(y_\star)\big) \\ & > 0, \end{aligned}$$

where the second inequality is from $\pi_{\theta_t}(y_\star) \le \pi_{\theta_0}(y_\star)$ and $\pi_{\theta_t}(y_{\text{med}}) \ge \pi_{\theta_0}(y_{\text{med}})$ for all $t \in [0,\tau)$, the third inequality is from Cauchy-Schwarz since $-s \le \|\phi(y_\star)\|\|\phi(y_{\text{med}})\|$, and the last inequality is from Item 2 in Lemma 6. Overall, this implies that if $\|\phi(y_\star)\|^2\pi_{\theta_t}(y_\star) + s\pi_{\theta_t}(y_{\text{med}}) \ge 0$, then $I_1 - I_2 < 0$.

We now consider the case of $\|\phi(y_\star)\|^2\pi_{\theta_t}(y_\star) + s\pi_{\theta_t}(y_{\text{med}}) < 0$, noting that this inequality is equivalent to the second term in Equation (23) being negative. We show that Equation (23) holds by proving the following equivalent inequality:

$$\frac{\pi_{\theta_t}(y_\star)A_{\text{G}}(y_\star;\theta_t)}{\pi_{\theta_t}(y_{\text{med}})A_{\text{G}}(y_{\text{med}};\theta_t)} \le \frac{\|\phi(y_{\text{med}})\|^2\pi_{\theta_t}(y_{\text{med}}) + s\pi_{\theta_t}(y_\star)}{-\|\phi(y_\star)\|^2\pi_{\theta_t}(y_\star) - s\pi_{\theta_t}(y_{\text{med}})}.$$

To see that this inequality holds, we lower bound the right-hand side by:

$$\frac{\|\phi(y_{\text{med}})\|^2\pi_{\theta_t}(y_{\text{med}}) + s\pi_{\theta_t}(y_\star)}{-\|\phi(y_\star)\|^2\pi_{\theta_t}(y_\star) - s\pi_{\theta_t}(y_{\text{med}})} \ge \frac{-s + \|\phi(y_{\text{med}})\|^2\pi_{\theta_t}(y_{\text{med}}) + s\pi_{\theta_t}(y_\star)}{\|\phi(y_\star)\|^2 - \|\phi(y_\star)\|^2\pi_{\theta_t}(y_\star) - s\pi_{\theta_t}(y_{\text{med}})}.$$

Since both denominators are positive in this case, multiplying both sides of the inequality by them and canceling common terms, we arrive at the following equivalent form of the inequality above:

$$s^2\pi_{\theta_t}(y_{\text{med}}) \le \|\phi(y_\star)\|^2\|\phi(y_{\text{med}})\|^2\pi_{\theta_t}(y_{\text{med}}),$$

which straightforwardly holds because $s^2 = \langle\phi(y_\star),\phi(y_{\text{med}})\rangle^2 \le \|\phi(y_{\text{med}})\|^2\|\phi(y_\star)\|^2$. Together with Equation (20), this implies that:

$$\begin{aligned} \frac{\pi_{\theta_t}(y_\star)A_{\text{G}}(y_\star;\theta_t)}{\pi_{\theta_t}(y_{\text{med}})A_{\text{G}}(y_{\text{med}};\theta_t)} &< \frac{-s + \|\phi(y_{\text{med}})\|^2\pi_{\theta_t}(y_{\text{med}}) + s\pi_{\theta_t}(y_\star)}{\|\phi(y_\star)\|^2 - \|\phi(y_\star)\|^2\pi_{\theta_t}(y_\star) - s\pi_{\theta_t}(y_{\text{med}})} \\ &\le \frac{\|\phi(y_{\text{med}})\|^2\pi_{\theta_t}(y_{\text{med}}) + s\pi_{\theta_t}(y_\star)}{-\|\phi(y_\star)\|^2\pi_{\theta_t}(y_\star) - s\pi_{\theta_t}(y_{\text{med}})}. \end{aligned} \tag{24}$$

Thus, Equation (23) holds, and so $I_2 > 0$, also when $\|\phi(y_\star)\|^2\pi_{\theta_t}(y_\star) + s\pi_{\theta_t}(y_{\text{med}}) < 0$. Altogether, we may conclude that $\frac{d}{dt}\pi_{\theta_t}(\mathcal{Y}_{\text{bad}}) = I_1 - I_2 < 0$ for all $t \in [0,\tau)$.

Next, we prove that $\pi_{\theta_\tau}(\mathcal{Y}_{\text{bad}})^2 \leq \frac{7\|\phi(y_{\text{med}})\|^2\pi_{\theta_0}(y_{\text{med}})^2 A_{\text{G}}(y_{\text{med}};\theta_\tau)}{8\max_{z\in\mathcal{Y}_{\text{bad}}}\|\phi(z)\|^2\Delta_2}$. We can write $A_{\text{G}}(y_{\text{med}};\theta_\tau)$ as:

$$\begin{aligned} A_{\text{G}}(y_{\text{med}};\theta_\tau) &= (1-\pi_{\theta_\tau}(y_{\text{med}}))r_{\text{G}}(y_{\text{med}}) - \pi_{\theta_\tau}(y_\star)r_{\text{G}}(y_\star) - \sum\nolimits_{z\in\mathcal{Y}_{\text{bad}}}\pi_{\theta_\tau}(z)r_{\text{G}}(z) \\ &= -\Delta_1\pi_{\theta_\tau}(y_\star) + \Delta_2\pi_{\theta_\tau}(\mathcal{Y}_{\text{bad}}). \end{aligned}$$

Thus,

$$\pi_{\theta_\tau}(\mathcal{Y}_{\text{bad}}) = \frac{A_{\text{G}}(y_{\text{med}};\theta_\tau) + \Delta_1\pi_{\theta_\tau}(y_\star)}{\Delta_2} \leq \frac{1.3A_{\text{G}}(y_{\text{med}};\theta_\tau)}{\Delta_2}.$$

The inequality is due to $\Delta_1\pi_{\theta_\tau}(y_\star) \leq 2\pi_{\theta_0}(y_\star) \leq 2\gamma^{13/14} \leq 0.3\sqrt{\gamma} \leq 0.3A_{\text{G}}(y_{\text{med}};\theta_\tau)$, where $2\gamma^{13/14} \leq 0.3\sqrt{\gamma}$ since $\gamma \leq 0.01$. Then,

$$\begin{aligned} \pi_{\theta_\tau}(\mathcal{Y}_{\text{bad}})^2 &\leq \frac{1.69A_{\text{G}}(y_{\text{med}};\theta_\tau)^2}{\Delta_2^2} \\ &< \frac{7\|\phi(y_{\text{med}})\|^2\pi_{\theta_0}(y_{\text{med}})^2 A_{\text{G}}(y_{\text{med}};\theta_\tau)}{8\max_{z\in\mathcal{Y}_{\text{bad}}}\|\phi(z)\|^2\Delta_2} \cdot \frac{2\max_{z\in\mathcal{Y}_{\text{bad}}}\|\phi(z)\|^2 A_{\text{G}}(y_{\text{med}};\theta_\tau)}{\|\phi(y_{\text{med}})\|^2\Delta_2\pi_{\theta_0}(y_{\text{med}})^2} \\ &\leq \frac{7\|\phi(y_{\text{med}})\|^2\pi_{\theta_0}(y_{\text{med}})^2 A_{\text{G}}(y_{\text{med}};\theta_\tau)}{8\max_{z\in\mathcal{Y}_{\text{bad}}}\|\phi(z)\|^2\Delta_2} \cdot \frac{2A_{\text{G}}(y_{\text{med}};\theta_0)}{\Delta_2\pi_{\theta_0}(y_{\text{med}})^2}, \end{aligned}$$

where the second inequality is by $1.69\times 8 < 2\times 7$ and the last inequality is by $\max_{z\in\mathcal{Y}_{\text{bad}}}\|\phi(z)\|^2 \leq \|\phi(y_{\text{med}})\|^2$ (Assumption 4) and $A_{\text{G}}(y_{\text{med}};\theta_\tau) \leq A_{\text{G}}(y_{\text{med}};\theta_0)$. Hence, it suffices to prove that $\frac{2A_{\text{G}}(y_{\text{med}};\theta_0)}{\Delta_2\pi_{\theta_0}(y_{\text{med}})^2} \leq 1$. Indeed, it holds that:

$$\begin{aligned} A_{\text{G}}(y_{\text{med}};\theta_0) &= (1-\pi_{\theta_0}(y_{\text{med}}))r_{\text{G}}(y_{\text{med}}) - \pi_{\theta_0}(y_\star)r_{\text{G}}(y_\star) - \sum\nolimits_{z\in\mathcal{Y}_{\text{bad}}}\pi_{\theta_0}(z)r_{\text{G}}(z) \\ &\leq (1-\pi_{\theta_0}(y_{\text{med}}))r_{\text{G}}(y_{\text{med}}) + \pi_{\theta_0}(\mathcal{Y}_{\text{bad}}) \\ &\leq 0.2r_{\text{G}}(y_{\text{med}}) + 0.1r_{\text{G}}(y_{\text{med}}) \\ &< \frac{0.8^2 r_{\text{G}}(y_{\text{med}})}{2} \\ &\leq \frac{\Delta_2\pi_{\theta_0}(y_{\text{med}})^2}{2}, \end{aligned}$$

where the first inequality is from $r_{\text{G}}(z) \geq -1$; the second inequality is from Lemma 6, according to which $\pi_{\theta_0}(\mathcal{Y}_{\text{bad}}) \leq 0.1r_{\text{G}}(y_{\text{med}}) \leq 0.1$ and $\pi_{\theta_0}(y_{\text{med}}) \geq 0.8$; and the last inequality is from $r_{\text{G}}(y_{\text{bad}}) \leq 0$ and $\pi_{\theta_0}(y_{\text{med}}) \geq 0.8$. Thus, $\frac{2A_{\text{G}}(y_{\text{med}};\theta_0)}{\Delta_2\pi_{\theta_0}(y_{\text{med}})^2} \leq 1$, from which it follows that:

$$\pi_{\theta_\tau}(\mathcal{Y}_{\text{bad}})^2 \leq \frac{7\|\phi(y_{\text{med}})\|^2\pi_{\theta_0}(y_{\text{med}})^2 A_{\text{G}}(y_{\text{med}};\theta_\tau)}{8\max_{z\in\mathcal{Y}_{\text{bad}}}\|\phi(z)\|^2\Delta_2}, \tag{25}$$

as desired.

We proceed by analyzing $\frac{d}{dt}\pi_{\theta_t}(y_\star)\big|_{t=\tau}$, based on the expression derived in Lemma 13:

$$
\begin{aligned}
&\frac{d}{dt}\pi_{\theta_t}(y_\star)\Big|_{t=\tau} \\
&= \pi_{\theta_\tau}(y_\star)\Big[\pi_{\theta_\tau}(y_\star)A_{\mathrm{G}}(y_\star;\theta_\tau)\big(\|\phi(y_\star)\|^2 - \|\phi(y_\star)\|^2\pi_{\theta_\tau}(y_\star) - s\pi_{\theta_\tau}(y_{\mathrm{med}})\big) \\
&\qquad + \pi_{\theta_\tau}(y_{\mathrm{med}})A_{\mathrm{G}}(y_{\mathrm{med}};\theta_\tau)\big(s - \|\phi(y_{\mathrm{med}})\|^2\pi_{\theta_\tau}(y_{\mathrm{med}}) - s\pi_{\theta_\tau}(y_\star)\big) \\
&\qquad - \textstyle\sum_{z\in\mathcal{Y}_{\mathrm{bad}}} \pi_{\theta_\tau}(z)^2 A_{\mathrm{G}}(z;\theta_\tau)\|\phi(z)\|^2\Big] \\
&\le \pi_{\theta_\tau}(y_\star)\Big[\pi_{\theta_\tau}(y_\star)A_{\mathrm{G}}(y_\star;\theta_\tau)\big(\|\phi(y_\star)\|^2 - \|\phi(y_\star)\|^2\pi_{\theta_\tau}(y_\star) - s\pi_{\theta_\tau}(y_{\mathrm{med}})\big) \\
&\qquad + \pi_{\theta_\tau}(y_{\mathrm{med}})A_{\mathrm{G}}(y_{\mathrm{med}};\theta_\tau)\big(s - \|\phi(y_{\mathrm{med}})\|^2\pi_{\theta_\tau}(y_{\mathrm{med}}) - s\pi_{\theta_\tau}(y_\star)\big) \\
&\qquad + (\Delta_2 - \sqrt{\gamma})\textstyle\sum_{z\in\mathcal{Y}_{\mathrm{bad}}} \pi_{\theta_\tau}(z)^2\|\phi(z)\|^2\Big] \\
&\le \pi_{\theta_\tau}(y_\star)\Big[(\|\phi(y_\star)\|^2 - s)\pi_{\theta_\tau}(y_\star)A_{\mathrm{G}}(y_\star;\theta_\tau)\big(1 - \pi_{\theta_\tau}(y_\star)\big) \\
&\qquad - \|\phi(y_{\mathrm{med}})\|^2\pi_{\theta_\tau}(y_{\mathrm{med}})^2 A_{\mathrm{G}}(y_{\mathrm{med}};\theta_\tau) \\
&\qquad + (\Delta_2 - \sqrt{\gamma})\textstyle\sum_{z\in\mathcal{Y}_{\mathrm{bad}}} \pi_{\theta_\tau}(z)^2\|\phi(z)\|^2\Big] \\
&\le \pi_{\theta_\tau}(y_\star)\Big[(\|\phi(y_\star)\|^2 - s)\pi_{\theta_0}(y_\star)A_{\mathrm{G}}(y_\star;\theta_0)\big(1 - \pi_{\theta_0}(y_\star)\big) \\
&\qquad - \|\phi(y_{\mathrm{med}})\|^2\pi_{\theta_\tau}(y_{\mathrm{med}})^2 A_{\mathrm{G}}(y_{\mathrm{med}};\theta_\tau) \\
&\qquad + \max_{z\in\mathcal{Y}_{\mathrm{bad}}}\|\phi(z)\|^2(\Delta_2 - \sqrt{\gamma})\pi_{\theta_\tau}(\mathcal{Y}_{\mathrm{bad}})^2\Big]
\end{aligned}
$$

The first inequality is from $-A_{\mathrm{G}}(z;\theta_\tau) \le -A_{\mathrm{G}}(z;\theta_{t_{\sqrt{\gamma}}}) \le \Delta_2 - \sqrt{\gamma}$; the second inequality is from $s \le 0$ and $\pi_{\theta_\tau}(y_{\mathrm{med}}) \le 1 - \pi_{\theta_\tau}(y_\star)$; and the third inequality is from $A_{\mathrm{G}}(y_\star;\theta_\tau) \le A_{\mathrm{G}}(y_\star;\theta_0)$, $\pi_{\theta_\tau}(y_\star) \le \pi_{\theta_0}(y_\star)$, $\pi_{\theta_\tau}(y_{\mathrm{med}}) \ge \pi_{\theta_0}(y_{\mathrm{med}})$ and $\sum_{z\in\mathcal{Y}_{\mathrm{bad}}}\pi_{\theta_\tau}(z)^2 \le \pi_{\theta_\tau}(\mathcal{Y}_{\mathrm{bad}})^2$. We continue bounding $\frac{d}{dt}\pi_{\theta_t}(y_\star)\big|_{t=\tau}$ as follows:

$$
\begin{aligned}
&\frac{d}{dt}\pi_{\theta_t}(y_\star)\Big|_{t=\tau} \\
&\le \pi_{\theta_\tau}(y_\star)\Big[(\|\phi(y_\star)\|^2 - s)\pi_{\theta_0}(y_\star)A_{\mathrm{G}}(y_\star;\theta_0)\big(1 - \pi_{\theta_0}(y_\star)\big) \\
&\qquad - \|\phi(y_{\mathrm{med}})\|^2\pi_{\theta_\tau}(y_{\mathrm{med}})^2 A_{\mathrm{G}}(y_{\mathrm{med}};\theta_\tau) \\
&\qquad + \max_{z\in\mathcal{Y}_{\mathrm{bad}}}\|\phi(z)\|^2(\Delta_2 - \sqrt{\gamma})\pi_{\theta_\tau}(\mathcal{Y}_{\mathrm{bad}})^2\Big] \\
&\le \pi_{\theta_\tau}(y_\star)\Big[(\|\phi(y_\star)\|^2 - s)\pi_{\theta_0}(y_\star)A_{\mathrm{G}}(y_\star;\theta_0)\big(1 - \pi_{\theta_0}(y_\star)\big) \\
&\qquad - (\|\phi(y_{\mathrm{med}})\|^2/8)\pi_{\theta_\tau}(y_{\mathrm{med}})^2 A_{\mathrm{G}}(y_{\mathrm{med}};\theta_\tau)\Big] \\
&\le \pi_{\theta_\tau}(y_\star)\Big[(\|\phi(y_\star)\|^2 - s)\pi_{\theta_0}(y_\star)A_{\mathrm{G}}(y_\star;\theta_0)\big(1 - \pi_{\theta_0}(y_\star)\big) \\
&\qquad - (\|\phi(y_{\mathrm{med}})\|^2/8)\pi_{\theta_\tau}(y_{\mathrm{med}})^2\sqrt{\gamma}\Big].
\end{aligned}
\tag{26}
$$

Here, the second inequality is from Equation (25) and $\pi_{\theta_0}(y_{\mathrm{med}}) \le \pi_{\theta_\tau}(y_{\mathrm{med}})$, according to which

$$
\begin{aligned}
(\Delta_2 - \sqrt{\gamma})\pi_{\theta_\tau}(\mathcal{Y}_{\mathrm{bad}})^2 &\le \Delta_2\pi_{\theta_\tau}(\mathcal{Y}_{\mathrm{bad}})^2 \\
&\le \frac{7\|\phi(y_{\mathrm{med}})\|^2\pi_{\theta_0}(y_{\mathrm{med}})^2 A_{\mathrm{G}}(y_{\mathrm{med}};\theta_\tau)}{8\max_{z\in\mathcal{Y}_{\mathrm{bad}}}\|\phi(z)\|^2} \\
&\le \frac{7\|\phi(y_{\mathrm{med}})\|^2\pi_{\theta_\tau}(y_{\mathrm{med}})^2 A_{\mathrm{G}}(y_{\mathrm{med}};\theta_\tau)}{8\max_{z\in\mathcal{Y}_{\mathrm{bad}}}\|\phi(z)\|^2},
\end{aligned}
$$

and the third inequality is from the fact that $V_{\mathrm{G}}(\theta_t)$ is monotonically non-decreasing when maximizing it via gradient flow, so $A_{\mathrm{G}}(y_{\mathrm{med}};\theta_\tau) \ge A_{\mathrm{G}}(y_{\mathrm{med}};\theta_{t_{\sqrt{\gamma}}})$.

Now, by Lemma 6, since $\|\phi(y_\star)\|^2 > 0 > s$ we get:

$$\begin{aligned}\pi_{\theta_0}(y_\star) &\le \frac{\|\phi(y_{\text{med}})\|^2 r_{\text{G}}(y_{\text{med}})^{2/7}\gamma^{13/14}}{40B^2(\Delta_1+\Delta_2)} \\ &\le \frac{\|\phi(y_{\text{med}})\|^2\sqrt{\gamma}\cdot\gamma^{3/7}}{40B^2 A_{\text{G}}(y_\star;\theta_0)} \\ &< \frac{0.8^2\|\phi(y_{\text{med}})\|^2\sqrt{\gamma}}{16B^2 A_{\text{G}}(y_\star;\theta_0)} \\ &\le \frac{\|\phi(y_{\text{med}})\|^2\pi_{\theta_0}(y_{\text{med}})^2\sqrt{\gamma}}{8(\|\phi(y_\star)\|^2-s)A_{\text{G}}(y_\star;\theta_0)}.\end{aligned}$$

The first inequality is from Item 2 in Lemma 6; the second inequality is from $r_{\text{G}}(y_{\text{med}}) \le 1$ and $A_{\text{G}}(y_\star;\theta_0) \le \Delta_1+\Delta_2$; the third inequality is from $\gamma^{3/7} \le 1$; and the last inequality is from $\pi_{\theta_0}(y_{\text{med}}) = 1-\pi_{\theta_0}(y_\star)-\pi_{\theta_0}(\mathcal{Y}_{\text{bad}}) \ge 1-\gamma^{13/14}-0.1r_{\text{G}}(y_{\text{med}}) \ge 0.8$ and $\|\phi(y_\star)\|^2 - s \le 2B^2$. Plugging this upper bound on $\pi_{\theta_0}(y_\star)$ into Equation (26), and using $\pi_{\theta_\tau}(y_{\text{med}}) \ge \pi_{\theta_0}(y_{\text{med}})$ together with $1-\pi_{\theta_0}(y_\star) < 1$, yields $\frac{d}{dt}\pi_{\theta_t}(y_\star)\big|_{t=\tau} < 0$. This contradicts the fact that τ is defined as the initial time at which $\frac{d}{dt}\pi_{\theta_t}(y_\star)\big|_{t=\tau} \ge 0$. Thus, it must be that $\frac{d}{dt}\pi_{\theta_t}(y_\star) < 0$ for all $t \in [0, t_{\sqrt{\gamma}}]$.

Step 2: upper bound on $\pi_{\theta_{t_{\sqrt{\gamma}}}}(y_\star)$ and monotonicity in $[t_{\sqrt{\gamma}}, t_\gamma]$. Next, we prove that $\pi_{\theta_t}(y_\star)$ keeps decreasing and $\pi_{\theta_t}(y_{\text{med}})$ keeps increasing for all $t \in [t_{\sqrt{\gamma}}, t_\gamma]$. To do so, we first derive upper bounds on $\frac{d}{dt}\pi_{\theta_t}(y_\star)$ and $\frac{d}{dt}\pi_{\theta_t}(y_{\text{med}})$ for $t \in [0, t_{\sqrt{\gamma}}]$ that allow quantifying how small the probability of $y_\star$ becomes until $t_{\sqrt{\gamma}}$.

Step 2.1: upper bound on $\frac{d}{dt}\pi_{\theta_t}(y_\star)$ in $[0, t_{\sqrt{\gamma}}]$. In this step, we prove that for all $t \in [0, t_{\sqrt{\gamma}}]$:

$$\frac{d}{dt}\pi_{\theta_t}(y_\star) \le -0.3(\|\phi(y_{\text{med}})\|^2 - s)\cdot\pi_{\theta_t}(y_\star)\pi_{\theta_t}(y_{\text{med}})^2 A_{\text{G}}(y_{\text{med}};\theta_t).$$

First, for $t \in [0, t_{\sqrt{\gamma}}]$, based on the expression derived in Lemma 13 we upper bound $\frac{d}{dt}\pi_{\theta_t}(y_\star)$ as follows:

$$\begin{aligned}&\frac{d}{dt}\pi_{\theta_t}(y_\star) \\ &= \pi_{\theta_t}(y_\star)\Big[\pi_{\theta_t}(y_\star)A_{\text{G}}(y_\star;\theta_t)\big(\|\phi(y_\star)\|^2 - \|\phi(y_\star)\|^2\pi_{\theta_t}(y_\star) - s\pi_{\theta_t}(y_{\text{med}})\big) \\ &\qquad + \pi_{\theta_t}(y_{\text{med}})A_{\text{G}}(y_{\text{med}};\theta_t)\big(s - \|\phi(y_{\text{med}})\|^2\pi_{\theta_t}(y_{\text{med}}) - s\pi_{\theta_t}(y_\star)\big) \\ &\qquad - \sum\nolimits_{z\in\mathcal{Y}_{\text{bad}}}\pi_{\theta_t}(z)^2 A_{\text{G}}(z;\theta_t)\|\phi(z)\|^2\Big] \\ &\le \pi_{\theta_t}(y_\star)\Big[(\|\phi(y_\star)\|^2 - s)\pi_{\theta_t}(y_\star)A_{\text{G}}(y_\star;\theta_t)\big(1-\pi_{\theta_t}(y_\star)\big) \\ &\qquad - (\|\phi(y_{\text{med}})\|^2 - s)\pi_{\theta_t}(y_{\text{med}})^2 A_{\text{G}}(y_{\text{med}};\theta_t) \\ &\qquad - \sum\nolimits_{z\in\mathcal{Y}_{\text{bad}}}\pi_{\theta_t}(z)^2 A_{\text{G}}(z;\theta_t)\|\phi(z)\|^2\Big] \\ &\le \pi_{\theta_t}(y_\star)\Big[(\|\phi(y_\star)\|^2 - s)\pi_{\theta_0}(y_\star)A_{\text{G}}(y_\star;\theta_t)\big(1-\pi_{\theta_0}(y_\star)\big) \\ &\qquad - (\|\phi(y_{\text{med}})\|^2 - s)\pi_{\theta_t}(y_{\text{med}})^2 A_{\text{G}}(y_{\text{med}};\theta_t) \\ &\qquad - \sum\nolimits_{z\in\mathcal{Y}_{\text{bad}}}\pi_{\theta_t}(z)^2 A_{\text{G}}(z;\theta_t)\|\phi(z)\|^2\Big].\end{aligned} \tag{27}$$

The first inequality is from $\pi_{\theta_t}(y_{\text{med}}) \le 1-\pi_{\theta_t}(y_\star)$ and the last inequality is by $\pi_{\theta_t}(y_\star) \le \pi_{\theta_0}(y_\star) < 0.5$, which implies $\pi_{\theta_t}(y_\star)(1-\pi_{\theta_t}(y_\star)) \le \pi_{\theta_0}(y_\star)(1-\pi_{\theta_0}(y_\star))$. Denote by G the term inside the square brackets in the last inequality above. We now prove that $G \le -0.3(\|\phi(y_{\text{med}})\|^2 - s)\pi_{\theta_t}(y_{\text{med}})^2 A_{\text{G}}(y_{\text{med}};\theta_t)$, which is equivalent to

$$G + 0.3(\|\phi(y_{\text{med}})\|^2 - s)\pi_{\theta_t}(y_{\text{med}})^2 A_{\text{G}}(y_{\text{med}};\theta_t) \le 0.$$

Towards doing so, we perform the following sequence of computations:

$$
\begin{aligned}
&G + 0.3(\|\phi(y_{\text{med}})\|^2 - s)\pi_{\theta_t}(y_{\text{med}})^2 A_{\text{G}}(y_{\text{med}};\theta_t) \\
&= -0.7(\|\phi(y_{\text{med}})\|^2 - s)\pi_{\theta_t}(y_{\text{med}})^2 A_{\text{G}}(y_{\text{med}};\theta_t) \\
&\quad + (\|\phi(y_\star)\|^2 - s)\pi_{\theta_0}(y_\star) A_{\text{G}}(y_\star;\theta_t)\big(1 - \pi_{\theta_0}(y_\star)\big) - \sum\nolimits_{z\in\mathcal{Y}_{\text{bad}}} \pi_{\theta_t}(z)^2 A_{\text{G}}(z;\theta_t)\|\phi(z)\|^2 \\
&\le -0.7(\|\phi(y_{\text{med}})\|^2 - s)\pi_{\theta_t}(y_{\text{med}})^2 A_{\text{G}}(y_{\text{med}};\theta_t) \\
&\quad + (\|\phi(y_\star)\|^2 - s)\pi_{\theta_0}(y_\star) A_{\text{G}}(y_\star;\theta_t) - \sum\nolimits_{z\in\mathcal{Y}_{\text{bad}}} \max\nolimits_{z\in\mathcal{Y}_{\text{bad}}} \|\phi(z)\|^2 \pi_{\theta_t}(z)^2 A_{\text{G}}(z;\theta_t) \\
&= -0.7(\|\phi(y_{\text{med}})\|^2 - s)\pi_{\theta_t}(y_{\text{med}})^2 A_{\text{G}}(y_{\text{med}};\theta_t) \\
&\quad + (\|\phi(y_\star)\|^2 - s)\pi_{\theta_0}(y_\star) A_{\text{G}}(y_\star;\theta_t) \\
&\quad + \sum\nolimits_{z\in\mathcal{Y}_{\text{bad}}} \max\nolimits_{z\in\mathcal{Y}_{\text{bad}}} \|\phi(z)\|^2 \pi_{\theta_t}(z)^2 (V_{\text{G}}(\theta_t) - r_{\text{G}}(z)) \\
&= -0.7(\|\phi(y_{\text{med}})\|^2 - s)\pi_{\theta_t}(y_{\text{med}})^2 \Big((1 - \pi_{\theta_t}(y_{\text{med}}))r_{\text{G}}(y_{\text{med}}) - \pi_{\theta_t}(y_\star) r_{\text{G}}(y_\star) \\
&\qquad\qquad - r_{\text{G}}(y_{\text{bad}}) \sum\nolimits_{z\in\mathcal{Y}_{\text{bad}}} \pi_{\theta_t}(z)\Big) \\
&\quad + (\|\phi(y_\star)\|^2 - s)\pi_{\theta_0}(y_\star) A_{\text{G}}(y_\star;\theta_t) \\
&\quad + \sum\nolimits_{z\in\mathcal{Y}_{\text{bad}}} \max\nolimits_{z\in\mathcal{Y}_{\text{bad}}} \|\phi(z)\|^2 \pi_{\theta_t}(z)^2 \Big((1 - \pi_{\theta_t}(z))(-r_{\text{G}}(z)) + \pi_{\theta_t}(y_\star) r_{\text{G}}(y_\star) \\
&\qquad\qquad + \pi_{\theta_t}(y_{\text{med}}) r_{\text{G}}(y_{\text{med}}) + \sum\nolimits_{z'\in\mathcal{Y}_{\text{bad}}\setminus\{z\}} \pi_{\theta_t}(z') r_{\text{G}}(z')\Big) \\
&= -0.7(\|\phi(y_{\text{med}})\|^2 - s)\pi_{\theta_t}(y_{\text{med}})^2 \big(\pi_{\theta_t}(\mathcal{Y}_{\text{bad}})\Delta_2 - \pi_{\theta_t}(y_\star)\Delta_1\big) \\
&\quad + (\|\phi(y_\star)\|^2 - s)\pi_{\theta_0}(y_\star) A_{\text{G}}(y_\star;\theta_t) \\
&\quad + \sum\nolimits_{z\in\mathcal{Y}_{\text{bad}}} \max\nolimits_{z\in\mathcal{Y}_{\text{bad}}} \|\phi(z)\|^2 \pi_{\theta_t}(z)^2 \big(\pi_{\theta_t}(y_{\text{med}})\Delta_2 + \pi_{\theta_t}(y_\star)(\Delta_1 + \Delta_2)\big),
\end{aligned}
$$

where the inequality is from $1 - \pi_{\theta_0}(y_\star) \le 1$.

Continuing from the last expression above, we get:

$$
\begin{aligned}
&G + 0.3(\|\phi(y_{\text{med}})\|^2 - s)\pi_{\theta_t}(y_{\text{med}})^2 A_{\text{G}}(y_{\text{med}};\theta_t) \\
&\le -0.7(\|\phi(y_{\text{med}})\|^2 - s)\pi_{\theta_t}(y_{\text{med}})^2 \big(\pi_{\theta_t}(\mathcal{Y}_{\text{bad}})\Delta_2 - \pi_{\theta_t}(y_\star)\Delta_1\big) \\
&\quad + (\|\phi(y_\star)\|^2 - s)\pi_{\theta_0}(y_\star) A_{\text{G}}(y_\star;\theta_t) \\
&\quad + \max\nolimits_{z\in\mathcal{Y}_{\text{bad}}} \|\phi(z)\|^2 \pi_{\theta_t}(\mathcal{Y}_{\text{bad}})^2 \big(\pi_{\theta_t}(y_{\text{med}})\Delta_2 + \pi_{\theta_t}(y_\star)(\Delta_1 + \Delta_2)\big) \\
&\le \pi_{\theta_t}(\mathcal{Y}_{\text{bad}})\Big[-0.7(\|\phi(y_{\text{med}})\|^2 - s)\pi_{\theta_t}(y_{\text{med}})^2 \Delta_2 \\
&\qquad\qquad + \max\nolimits_{z\in\mathcal{Y}_{\text{bad}}} \|\phi(z)\|^2 \big(\pi_{\theta_t}(\mathcal{Y}_{\text{bad}})\pi_{\theta_t}(y_{\text{med}})\Delta_2\big)\Big] \\
&\quad + \pi_{\theta_0}(y_\star)\Big[0.7(\|\phi(y_{\text{med}})\|^2 - s)\pi_{\theta_t}(y_{\text{med}})^2 \Delta_1 + (\|\phi(y_\star)\|^2 - s) A_{\text{G}}(y_\star;\theta_t) \\
&\qquad\qquad + \max\nolimits_{z\in\mathcal{Y}_{\text{bad}}} \|\phi(z)\|^2 (\Delta_1 + \Delta_2)\pi_{\theta_t}(\mathcal{Y}_{\text{bad}})^2\Big] \\
&\le \underbrace{\pi_{\theta_t}(\mathcal{Y}_{\text{bad}})\Delta_2\Big[-0.7(\|\phi(y_{\text{med}})\|^2 - s)\pi_{\theta_t}(y_{\text{med}})^2 + \|\phi(y_{\text{med}})\|^2 \pi_{\theta_t}(\mathcal{Y}_{\text{bad}})\pi_{\theta_t}(y_{\text{med}})\Big]}_{I_1} \\
&\quad + \pi_{\theta_0}(y_\star)\Big[0.7(\|\phi(y_{\text{med}})\|^2 - s)\pi_{\theta_t}(y_{\text{med}})^2 \Delta_1 + \max\nolimits_{z\in\mathcal{Y}_{\text{bad}}} \|\phi(z)\|^2 \pi_{\theta_t}(\mathcal{Y}_{\text{bad}})^2 (\Delta_1 + \Delta_2) \\
&\qquad\qquad + (\|\phi(y_\star)\|^2 - s) A_{\text{G}}(y_\star;\theta_t)\Big].
\end{aligned}
$$

Here, the first inequality is due to $\sum_{z\in\mathcal{Y}_{\text{bad}}} \pi_{\theta_t}(z)^2 \le \pi_{\theta_t}(\mathcal{Y}_{\text{bad}})^2$; the second inequality is from $\pi_{\theta_t}(y_\star) \le \pi_{\theta_0}(y_\star)$; and the last inequality is by Assumption 4. We denote the second term in the last inequality by I_2, *i.e.*, I_2 is equal to $\pi_{\theta_0}(y_\star)$ multiplied by the term inside the last square brackets.

Next, we show that:

$$I_1 \le -0.2(\|\phi(y_{\text{med}})\|^2 - s)\pi_{\theta_t}(\mathcal{Y}_{\text{bad}})\pi_{\theta_t}(y_{\text{med}})^2\Delta_2,$$
$$I_2 \le 0.2(\|\phi(y_{\text{med}})\|^2 - s)\pi_{\theta_t}(\mathcal{Y}_{\text{bad}})\pi_{\theta_t}(y_{\text{med}})^2\Delta_2,$$

from which it follows that $G + 0.3(\|\phi(y_{\text{med}})\|^2 - s)\pi_{\theta_t}(y_{\text{med}})^2 A_{\text{G}}(y_{\text{med}};\theta_t) \le 0$, and so $G \le -0.3(\|\phi(y_{\text{med}})\|^2 - s)\pi_{\theta_t}(y_{\text{med}})^2 A_{\text{G}}(y_{\text{med}};\theta_t)$. Plugging this inequality into Equation (27) will then yield the desired upper bound on $\frac{d}{dt}\pi_{\theta_t}(y_\star)$.

For I_1, it suffices to show that $0.5(\|\phi(y_{\text{med}})\|^2 - s)\pi_{\theta_t}(y_{\text{med}}) \ge \|\phi(y_{\text{med}})\|^2\pi_{\theta_t}(\mathcal{Y}_{\text{bad}})$. From Lemma 6, we have $\pi_{\theta_0}(\mathcal{Y}_{\text{bad}}) < 0.2$ and $\gamma < 0.05$, so $\pi_{\theta_0}(y_\star) \le \gamma^{13/14} < 0.13$ and $\pi_{\theta_0}(y_{\text{med}}) = 1 - \pi_{\theta_0}(y_\star) - \pi_{\theta_0}(\mathcal{Y}_{\text{bad}}) > 0.67$. This implies that $\pi_{\theta_t}(\mathcal{Y}_{\text{bad}}) \le \pi_{\theta_0}(\mathcal{Y}_{\text{bad}}) < 0.4\pi_{\theta_0}(y_{\text{med}}) \le 0.4\pi_{\theta_t}(y_{\text{med}})$. Since $s \le 0$, we further obtain:

$$\|\phi(y_{\text{med}})\|^2\pi_{\theta_t}(\mathcal{Y}_{\text{bad}}) < 0.4\pi_{\theta_t}(y_{\text{med}})\|\phi(y_{\text{med}})\|^2 < 0.5(\|\phi(y_{\text{med}})\|^2 - s)\pi_{\theta_t}(y_{\text{med}}),$$

which implies that $I_1 \le -0.2(\|\phi(y_{\text{med}})\|^2 - s)\pi_{\theta_t}(\mathcal{Y}_{\text{bad}})\pi_{\theta_t}(y_{\text{med}})^2\Delta_2$.

For I_2, we have:

$$\begin{aligned}
&0.7(\|\phi(y_{\text{med}})\|^2 - s)\pi_{\theta_t}(y_{\text{med}})^2\Delta_1 + \max_{z\in\mathcal{Y}_{\text{bad}}}\|\phi(z)\|^2\pi_{\theta_t}(\mathcal{Y}_{\text{bad}})^2(\Delta_1+\Delta_2)\\
&\quad + (\|\phi(y_\star)\|^2 - s)A_{\text{G}}(y_\star;\theta_t)\\
&\le 0.7(\|\phi(y_{\text{med}})\|^2 - s)\Delta_1 + 0.1\max_{z\in\mathcal{Y}_{\text{bad}}}\|\phi(z)\|^2(\Delta_1+\Delta_2) + (\|\phi(y_\star)\|^2 - s)A_{\text{G}}(y_\star;\theta_0)\\
&\le 2(B^2 - s)(\Delta_1+\Delta_2),
\end{aligned}$$

where the first inequality is from $\pi_{\theta_t}(\mathcal{Y}_{\text{bad}})^2 \le \pi_{\theta_0}(\mathcal{Y}_{\text{bad}})^2 \le \pi_{\theta_0}(\mathcal{Y}_{\text{bad}}) \le 0.1$ and $V_{\text{G}}(\theta_t) \ge V_{\text{G}}(\theta_0)$, and the second inequality is from $A_{\text{G}}(y_\star;\theta_0) \le \Delta_1 + \Delta_2$. Meanwhile, as $A_{\text{G}}(y_{\text{med}};\theta_t) \ge \sqrt{\gamma}$ for $t \in [0, t_{\sqrt{\gamma}}]$, we get:

$$\begin{aligned}
A_{\text{G}}(y_{\text{med}};\theta_t) &= (1 - \pi_{\theta_t}(y_{\text{med}}))r_{\text{G}}(y_{\text{med}}) - \pi_{\theta_t}(y_\star)r_{\text{G}}(y_\star) - \sum\nolimits_{z\in\mathcal{Y}_{\text{bad}}}\pi_{\theta_t}(z)r_{\text{G}}(z)\\
&= -\Delta_1\pi_{\theta_t}(y_\star) + \Delta_2\pi_{\theta_t}(\mathcal{Y}_{\text{bad}}).
\end{aligned}$$

Thus, $-\Delta_1\pi_{\theta_t}(y_\star) + \Delta_2\pi_{\theta_t}(\mathcal{Y}_{\text{bad}}) \ge \sqrt{\gamma}$, and rearranging this inequality gives

$$\pi_{\theta_t}(\mathcal{Y}_{\text{bad}}) \ge \frac{\sqrt{\gamma} + \Delta_1\pi_{\theta_t}(y_\star)}{\Delta_2} \ge \frac{\sqrt{\gamma}}{\Delta_2}. \tag{28}$$

Here, the second inequality uses $\Delta_1\pi_{\theta_t}(y_\star) \ge 0$. By the bound on γ in Item 1 of Lemma 6, together with $\pi_{\theta_0}(y_{\text{med}}) > 0.67$ (shown above when bounding I_1), it holds that

$$\gamma^{3/7} \le \frac{21\|\phi(y_{\text{med}})\|^2}{1100B^2(\Delta_1+\Delta_2)} \le \frac{\|\phi(y_{\text{med}})\|^2\pi_{\theta_0}(y_{\text{med}})^2}{10B^2(\Delta_1+\Delta_2)}.$$

Moreover, notice that since $s < 0$ and $\|\phi(y_{\text{med}})\|^2 \le B^2$, we have

$$\frac{\|\phi(y_{\text{med}})\|^2}{B^2} \le \frac{\|\phi(y_{\text{med}})\|^2 - s}{B^2 - s}.$$

The last two inequalities thus give:

$$\pi_{\theta_0}(y_\star) \le \gamma^{13/14} = \sqrt{\gamma}\cdot\gamma^{3/7} \le \frac{(\|\phi(y_{\text{med}})\|^2 - s)\sqrt{\gamma}\pi_{\theta_0}(y_{\text{med}})^2}{10(B^2 - s)(\Delta_1+\Delta_2)},$$

which leads to:

$$I_2 \le 2(B^2 - s)(\Delta_1+\Delta_2)\pi_{\theta_0}(y_\star) \le 0.2(\|\phi(y_{\text{med}})\|^2 - s)\pi_{\theta_t}(\mathcal{Y}_{\text{bad}})\pi_{\theta_t}(y_{\text{med}})^2\Delta_2.$$

Combining the upper bounds on I_1 and I_2, we have shown that $I_1 + I_2 \le 0$, and so $G \le -0.3(\|\phi(y_{\text{med}})\|^2 - s)\pi_{\theta_t}(y_{\text{med}})^2 A_{\text{G}}(y_{\text{med}};\theta_t)$. Plugging this upper bound on G into Equation (27) establishes that for all $t \in [0, t_{\sqrt{\gamma}}]$:

$$\frac{d}{dt}\pi_{\theta_t}(y_\star) \le -0.3(\|\phi(y_{\text{med}})\|^2 - s)\cdot\pi_{\theta_t}(y_\star)\pi_{\theta_t}(y_{\text{med}})^2 A_{\text{G}}(y_{\text{med}};\theta_t),$$

as desired.

Step 2.2: upper bound on $\frac{d}{dt}\pi_{\theta_t}(y_{\mathrm{med}})$ in $[0, t_{\sqrt{\gamma}}]$. In Step 1, we showed that $\frac{d}{dt}\pi_{\theta_t}(y_{\mathrm{med}}) > 0$ for $t \in [0, t_{\sqrt{\gamma}}]$. In this step, we prove that $\pi_{\theta_t}(y_{\mathrm{med}})$ does not grow too fast by establishing that:

$$\frac{d}{dt}\pi_{\theta_t}(y_{\mathrm{med}}) \le 2.1(\|\phi(y_{\mathrm{med}})\|^2 - s/2)\pi_{\theta_t}(y_{\mathrm{med}})^2(1-\pi_{\theta_t}(y_{\mathrm{med}}))A_{\mathrm{G}}(y_{\mathrm{med}};\theta_t).$$

We begin upper bounding $\frac{d}{dt}\pi_{\theta_t}(y_{\mathrm{med}})$ using the expression derived in Lemma 13:

$$\begin{aligned}
&\frac{d}{dt}\pi_{\theta_t}(y_{\mathrm{med}}) \\
&= \pi_{\theta_t}(y_{\mathrm{med}})\Big[\pi_{\theta_t}(y_{\mathrm{med}})A_{\mathrm{G}}(y_{\mathrm{med}};\theta_t)\big(\|\phi(y_{\mathrm{med}})\|^2 - \|\phi(y_{\mathrm{med}})\|^2\pi_{\theta_t}(y_{\mathrm{med}}) - s\pi_{\theta_t}(y_\star)\big) \\
&\qquad + \pi_{\theta_t}(y_\star)A_{\mathrm{G}}(y_\star;\theta_t)\big(s - \|\phi(y_\star)\|^2\pi_{\theta_t}(y_\star) - s\pi_{\theta_t}(y_{\mathrm{med}})\big) \\
&\qquad - \sum\nolimits_{z\in\mathcal{Y}_{\mathrm{bad}}}\pi_{\theta_t}(z)^2A_{\mathrm{G}}(z;\theta_t)\|\phi(z)\|^2\Big] \\
&\le \pi_{\theta_t}(y_{\mathrm{med}})\Big[\pi_{\theta_t}(y_{\mathrm{med}})A_{\mathrm{G}}(y_{\mathrm{med}};\theta_t)\big(\|\phi(y_{\mathrm{med}})\|^2 - \|\phi(y_{\mathrm{med}})\|^2\pi_{\theta_t}(y_{\mathrm{med}}) - s\pi_{\theta_t}(y_\star)\big) \\
&\qquad - \sum\nolimits_{z\in\mathcal{Y}_{\mathrm{bad}}}\pi_{\theta_t}(z)^2A_{\mathrm{G}}(z;\theta_t)\|\phi(z)\|^2\Big] \\
&\le \pi_{\theta_t}(y_{\mathrm{med}})\Big[\pi_{\theta_t}(y_{\mathrm{med}})A_{\mathrm{G}}(y_{\mathrm{med}};\theta_t)\big(\|\phi(y_{\mathrm{med}})\|^2 - \|\phi(y_{\mathrm{med}})\|^2\pi_{\theta_t}(y_{\mathrm{med}}) - s\pi_{\theta_t}(y_\star)\big) \\
&\qquad + \max\nolimits_{z\in\mathcal{Y}_{\mathrm{bad}}}\|\phi(z)\|^2(\Delta_2 - A_{\mathrm{G}}(y_{\mathrm{med}};\theta_t))\sum\nolimits_{z\in\mathcal{Y}_{\mathrm{bad}}}\pi_{\theta_t}(z)^2\Big] \\
&\le \pi_{\theta_t}(y_{\mathrm{med}})\Big[\pi_{\theta_t}(y_{\mathrm{med}})A_{\mathrm{G}}(y_{\mathrm{med}};\theta_t)\big(\|\phi(y_{\mathrm{med}})\|^2 - \|\phi(y_{\mathrm{med}})\|^2\pi_{\theta_t}(y_{\mathrm{med}}) - s\pi_{\theta_t}(y_\star)\big) \\
&\qquad + \max\nolimits_{z\in\mathcal{Y}_{\mathrm{bad}}}\|\phi(z)\|^2(\Delta_2 - A_{\mathrm{G}}(y_{\mathrm{med}};\theta_t))\pi_{\theta_t}(\mathcal{Y}_{\mathrm{bad}})^2\Big].
\end{aligned} \tag{29}$$

Here, the first inequality is from $s - \|\phi(y_\star)\|^2\pi_{\theta_t}(y_\star) - s\pi_{\theta_t}(y_{\mathrm{med}}) \le 0$ and the last inequality is from $\sum_{z\in\mathcal{Y}_{\mathrm{bad}}}\pi_{\theta_t}(z)^2 \le \pi_{\theta_t}(\mathcal{Y}_{\mathrm{bad}})^2$. Our next claim is that $1 - \pi_{\theta_t}(y_{\mathrm{med}}) \ge 2\pi_{\theta_t}(y_\star)$ for all $t \in [0, t_{\sqrt{\gamma}}]$. Indeed, by Lemma 16, for $t \in [0, t_{\sqrt{\gamma}}]$ it holds that:

$$1 - \pi_{\theta_t}(y_{\mathrm{med}}) \ge \frac{\sqrt{\gamma}}{1 + r_{\mathrm{G}}(y_{\mathrm{med}})} \ge \frac{\sqrt{\gamma}}{2}.$$

Moreover, Items 2 and 3 in Lemma 6 give $\pi_{\theta_t}(y_\star) \le \pi_{\theta_0}(y_\star) \le \gamma^{13/14}$. Under the conditions on γ in Lemma 6, it holds that $\sqrt{\gamma} \ge 4\gamma^{13/14}$. Thus, $1 - \pi_{\theta_t}(y_{\mathrm{med}}) \ge 2\pi_{\theta_t}(y_\star)$, from which we conclude:

$$\|\phi(y_{\mathrm{med}})\|^2 - \|\phi(y_{\mathrm{med}})\|^2\pi_{\theta_t}(y_{\mathrm{med}}) - s\pi_{\theta_t}(y_\star) \le (\|\phi(y_{\mathrm{med}})\|^2 - s/2)(1 - \pi_{\theta_t}(y_{\mathrm{med}})).$$

Returning to Equation (29), this leads to:

$$\begin{aligned}
\frac{d}{dt}\pi_{\theta_t}(y_{\mathrm{med}}) &\le \pi_{\theta_t}(y_{\mathrm{med}})\Big[(\|\phi(y_{\mathrm{med}})\|^2 - s/2)\pi_{\theta_t}(y_{\mathrm{med}})(1-\pi_{\theta_t}(y_{\mathrm{med}}))A_{\mathrm{G}}(y_{\mathrm{med}};\theta_t) \\
&\qquad + \max\nolimits_{z\in\mathcal{Y}_{\mathrm{bad}}}\|\phi(z)\|^2(\Delta_2 - A_{\mathrm{G}}(y_{\mathrm{med}};\theta_t))\pi_{\theta_t}(\mathcal{Y}_{\mathrm{bad}})^2\Big] \\
&\le \pi_{\theta_t}(y_{\mathrm{med}})(1-\pi_{\theta_t}(y_{\mathrm{med}}))\Big[(\|\phi(y_{\mathrm{med}})\|^2 - s/2)\pi_{\theta_t}(y_{\mathrm{med}})A_{\mathrm{G}}(y_{\mathrm{med}};\theta_t) \\
&\qquad + \underbrace{\max\nolimits_{z\in\mathcal{Y}_{\mathrm{bad}}}\|\phi(z)\|^2(\Delta_2 - A_{\mathrm{G}}(y_{\mathrm{med}};\theta_t))\pi_{\theta_t}(\mathcal{Y}_{\mathrm{bad}})}_{G}\Big]
\end{aligned} \tag{30}$$

where the second inequality is from $\pi_{\theta_t}(\mathcal{Y}_{\text{bad}}) \le 1-\pi_{\theta_t}(y_{\text{med}})$. We now show that $1.1(\|\phi(y_{\text{med}})\|^2 - s/2)\pi_{\theta_t}(y_{\text{med}})A_{\text{G}}(y_{\text{med}};\theta_t) - G \ge 0$:

$$
\begin{aligned}
&1.1(\|\phi(y_{\text{med}})\|^2 - s/2)\pi_{\theta_t}(y_{\text{med}})A_{\text{G}}(y_{\text{med}};\theta_t) - G \\
&= 1.1(\|\phi(y_{\text{med}})\|^2 - s/2)\pi_{\theta_t}(y_{\text{med}})A_{\text{G}}(y_{\text{med}};\theta_t) \\
&\quad - \max_{z\in\mathcal{Y}_{\text{bad}}} \|\phi(z)\|^2 \pi_{\theta_t}(\mathcal{Y}_{\text{bad}})(\Delta_2 - A_{\text{G}}(y_{\text{med}};\theta_t)) \\
&\ge 1.1\|\phi(y_{\text{med}})\|^2 \pi_{\theta_t}(y_{\text{med}})\big(r_{\text{G}}(y_{\text{med}}) - V_{\text{G}}(\theta_t)\big) \\
&\quad - \max_{z\in\mathcal{Y}_{\text{bad}}} \|\phi(z)\|^2 \pi_{\theta_t}(\mathcal{Y}_{\text{bad}})\big(V_{\text{G}}(\theta_t) - r_{\text{G}}(y_{\text{bad}})\big) \\
&= 1.1\|\phi(y_{\text{med}})\|^2 \pi_{\theta_t}(y_{\text{med}})\big(\pi_{\theta_t}(\mathcal{Y}_{\text{bad}})\Delta_2 - \pi_{\theta_t}(y_\star)\Delta_1\big) \\
&\quad - \max_{z\in\mathcal{Y}_{\text{bad}}} \|\phi(z)\|^2 \pi_{\theta_t}(\mathcal{Y}_{\text{bad}})\big(\pi_{\theta_t}(y_{\text{med}})\Delta_2 + \pi_{\theta_t}(y_\star)(\Delta_1+\Delta_2)\big) \\
&= \pi_{\theta_t}(\mathcal{Y}_{\text{bad}})\Big(1.1\|\phi(y_{\text{med}})\|^2 \pi_{\theta_t}(y_{\text{med}})\Delta_2 - \max_{z\in\mathcal{Y}_{\text{bad}}} \|\phi(z)\|^2 \pi_{\theta_t}(y_{\text{med}})\Delta_2\Big) \\
&\quad - \pi_{\theta_t}(y_\star)\big(1.1\|\phi(y_{\text{med}})\|^2 \pi_{\theta_t}(y_{\text{med}})\Delta_1 + \max_{z\in\mathcal{Y}_{\text{bad}}} \|\phi(z)\|^2 \pi_{\theta_t}(\mathcal{Y}_{\text{bad}})(\Delta_1+\Delta_2)\big) \\
&\ge \underbrace{0.1\pi_{\theta_t}(\mathcal{Y}_{\text{bad}})\pi_{\theta_t}(y_{\text{med}})\Delta_2\|\phi(y_{\text{med}})\|^2}_{I_1} \\
&\quad - \underbrace{\pi_{\theta_t}(y_\star)\|\phi(y_{\text{med}})\|^2\big(1.1\pi_{\theta_t}(y_{\text{med}})\Delta_1 + \pi_{\theta_t}(\mathcal{Y}_{\text{bad}})(\Delta_1+\Delta_2)\big)}_{I_2}.
\end{aligned}
$$

Here, the first inequality is from $s \le 0$ and the last inequality is from Assumption 4.

We proceed by showing that $I_2 \le 0.1\|\phi(y_{\text{med}})\|^2\pi_{\theta_t}(y_{\text{med}})\pi_{\theta_t}(\mathcal{Y}_{\text{bad}})\Delta_2 = I_1$. From Lemma 6, we have $\pi_{\theta_t}(\mathcal{Y}_{\text{bad}}) \le \pi_{\theta_0}(\mathcal{Y}_{\text{bad}}) < 0.4\pi_{\theta_0}(y_{\text{med}}) \le 0.4\pi_{\theta_t}(y_{\text{med}})$. Therefore,

$$
\begin{aligned}
1.1\pi_{\theta_t}(y_{\text{med}})\Delta_1 + \pi_{\theta_t}(\mathcal{Y}_{\text{bad}})(\Delta_1+\Delta_2) &< 1.1\pi_{\theta_t}(y_{\text{med}})\Delta_1 + 0.4\pi_{\theta_t}(y_{\text{med}})(\Delta_1+\Delta_2) \\
&\le 1.5\pi_{\theta_t}(y_{\text{med}})(\Delta_1+\Delta_2).
\end{aligned}
$$

As shown in Equation (28), $\pi_{\theta_t}(\mathcal{Y}_{\text{bad}}) \ge \frac{\sqrt{\gamma}}{\Delta_2}$ for $t \in [0, t_{\sqrt{\gamma}}]$. Furthermore, by Lemma 6 we know that

$$
\gamma \le \left(\frac{\|\phi(y_{\text{med}})\|^2}{50B^2(\Delta_1+\Delta_2)}\right)^{7/3},
$$

and

$$
\pi_{\theta_0}(y_\star) \le \gamma^{13/14} \le \frac{\sqrt{\gamma}\|\phi(y_{\text{med}})\|^2}{50B^2(\Delta_1+\Delta_2)}.
$$

Thus, we arrive at:

$$
\begin{aligned}
I_2 &\le 1.5B^2\pi_{\theta_t}(y_{\text{med}})\pi_{\theta_0}(y_\star)(\Delta_1+\Delta_2) \\
&\le 0.03\|\phi(y_{\text{med}})\|^2\pi_{\theta_t}(y_{\text{med}})\pi_{\theta_t}(\mathcal{Y}_{\text{bad}})\Delta_2 \\
&\le 0.1\|\phi(y_{\text{med}})\|^2\pi_{\theta_t}(y_{\text{med}})\pi_{\theta_t}(\mathcal{Y}_{\text{bad}})\Delta_2 \\
&= I_1,
\end{aligned}
$$

which implies that $G \le 1.1(\|\phi(y_{\text{med}})\|^2 - s/2)\pi_{\theta_t}(y_{\text{med}})A_{\text{G}}(y_{\text{med}};\theta_t)$. Going back to Equation (30), the desired upper bound on $\frac{d}{dt}\pi_{\theta_t}(y_{\text{med}})$ follows:

$$
\begin{aligned}
\frac{d}{dt}\pi_{\theta_t}(y_{\text{med}}) &\le \pi_{\theta_t}(y_{\text{med}})(1-\pi_{\theta_t}(y_{\text{med}}))\Big[(\|\phi(y_{\text{med}})\|^2 - s/2)\pi_{\theta_t}(y_{\text{med}})A_{\text{G}}(y_{\text{med}};\theta_t) + G\Big] \\
&\le \pi_{\theta_t}(y_{\text{med}})(1-\pi_{\theta_t}(y_{\text{med}}))\Big[(\|\phi(y_{\text{med}})\|^2 - s/2)\pi_{\theta_t}(y_{\text{med}})A_{\text{G}}(y_{\text{med}};\theta_t) \\
&\qquad + 1.1(\|\phi(y_{\text{med}})\|^2 - s/2)\pi_{\theta_t}(y_{\text{med}})A_{\text{G}}(y_{\text{med}};\theta_t)\Big] \\
&= 2.1(\|\phi(y_{\text{med}})\|^2 - s/2)\pi_{\theta_t}(y_{\text{med}})^2(1-\pi_{\theta_t}(y_{\text{med}}))A_{\text{G}}(y_{\text{med}};\theta_t).
\end{aligned}
$$

Step 2.3: upper bound on $\pi_{\theta_{t_{\sqrt{\gamma}}}}(y_\star)$. To summarize, we showed that for all $t \in [0, t_{\sqrt{\gamma}}]$:

$$
\begin{aligned}
\frac{d}{dt}\pi_{\theta_t}(y_\star) &\le -0.3(\|\phi(y_{\text{med}})\|^2 - s)\cdot\pi_{\theta_t}(y_\star)\pi_{\theta_t}(y_{\text{med}})^2 A_{\text{G}}(y_{\text{med}};\theta_t), \\
\frac{d}{dt}\pi_{\theta_t}(y_{\text{med}}) &\le 2.1(\|\phi(y_{\text{med}})\|^2 - s/2)\cdot\pi_{\theta_t}(y_{\text{med}})^2(1-\pi_{\theta_t}(y_{\text{med}}))A_{\text{G}}(y_{\text{med}};\theta_t).
\end{aligned}
$$

Since both sides of the first inequality are negative and both sides of the second inequality are positive, we can divide them to obtain:

$$\frac{\frac{d}{dt}\pi_{\theta_t}(y_\star)}{\frac{d}{dt}\pi_{\theta_t}(y_{\text{med}})} \leq -\frac{\|\phi(y_{\text{med}})\|^2 - s}{7(\|\phi(y_{\text{med}})\|^2 - s/2)} \cdot \frac{\pi_{\theta_t}(y_\star)}{1-\pi_{\theta_t}(y_{\text{med}})}$$
$$= -h(s) \cdot \frac{\pi_{\theta_t}(y_\star)}{1-\pi_{\theta_t}(y_{\text{med}})},$$

where recall that $h(s) = \frac{\|\phi(y_{\text{med}})\|^2 - s}{7(\|\phi(y_{\text{med}})\|^2 - s/2)}$ (Equation (18)). We can rewrite this inequality as:

$$\frac{\frac{d}{dt}\pi_{\theta_t}(y_\star)}{\pi_{\theta_t}(y_\star)} \leq h(s) \cdot \frac{-\frac{d}{dt}\pi_{\theta_t}(y_{\text{med}})}{1-\pi_{\theta_t}(y_{\text{med}})} = h(s) \cdot \frac{\frac{d}{dt}(1-\pi_{\theta_t}(y_{\text{med}}))}{1-\pi_{\theta_t}(y_{\text{med}})},$$

from which it follows that:

$$\frac{d}{dt}\ln \pi_{\theta_t}(y_\star) \leq h(s) \cdot \frac{d}{dt}\ln(1-\pi_{\theta_t}(y_{\text{med}})).$$

Integrating both sides over $[0, t_{\sqrt{\gamma}}]$ gives:

$$\ln \pi_{\theta_t}(y_\star)\Big|_0^{t_{\sqrt{\gamma}}} \leq h(s) \cdot \ln(1-\pi_{\theta_t}(y_{\text{med}}))\Big|_0^{t_{\sqrt{\gamma}}}.$$

Therefore,

$$\frac{\pi_{\theta_{t_{\sqrt{\gamma}}}}(y_\star)}{\pi_{\theta_0}(y_\star)} \leq \left(\frac{1-\pi_{\theta_{t_{\sqrt{\gamma}}}}(y_{\text{med}})}{1-\pi_{\theta_0}(y_{\text{med}})}\right)^{h(s)},$$

and hence

$$\pi_{\theta_{t_{\sqrt{\gamma}}}}(y_\star) \leq \frac{\pi_{\theta_0}(y_\star)(1-\pi_{\theta_{t_{\sqrt{\gamma}}}}(y_{\text{med}}))^{h(s)}}{(1-\pi_{\theta_0}(y_{\text{med}}))^{h(s)}}. \tag{31}$$

Next, we prove that $1-\pi_{\theta_{t_{\sqrt{\gamma}}}}(y_{\text{med}}) \leq \frac{\sqrt{\gamma}+\pi_{\theta_0}(y_\star)}{r_{\text{G}}(y_{\text{med}})}$. To see this, assume by way of contradiction that $1-\pi_{\theta_{t_{\sqrt{\gamma}}}}(y_{\text{med}}) > \frac{\sqrt{\gamma}+\pi_{\theta_0}(y_\star)}{r_{\text{G}}(y_{\text{med}})}$. Then, we would have:

$$\begin{aligned} V_{\text{G}}(\theta_{t_{\sqrt{\gamma}}}) &= \pi_{\theta_{t_{\sqrt{\gamma}}}}(y_{\text{med}})r_{\text{G}}(y_{\text{med}}) + \pi_{\theta_{t_{\sqrt{\gamma}}}}(y_\star)r_{\text{G}}(y_\star) + \pi_{\theta_{t_{\sqrt{\gamma}}}}(\mathcal{Y}_{\text{bad}})r_{\text{G}}(y_{\text{bad}}) \\ &< \left(1 - \frac{\sqrt{\gamma}+\pi_{\theta_0}(y_\star)}{r_{\text{G}}(y_{\text{med}})}\right) r_{\text{G}}(y_{\text{med}}) + \pi_{\theta_0}(y_\star) \\ &\leq r_{\text{G}}(y_{\text{med}}) - \sqrt{\gamma}, \end{aligned}$$

where $y_{\text{bad}} \in \mathcal{Y}_{\text{bad}}$, and so $r_{\text{G}}(y_{\text{bad}}) \leq 0$. This contradicts the definition of $t_{\sqrt{\gamma}}$ as the initial time at which $V_{\text{G}}(\theta_t) \geq r_{\text{G}}(y_{\text{med}}) - \sqrt{\gamma}$.

Plugging the upper bound on $1-\pi_{\theta_{t_{\sqrt{\gamma}}}}(y_{\text{med}})$ into Equation (31) then gives:

$$\begin{aligned} \pi_{\theta_{t_{\sqrt{\gamma}}}}(y_\star) &\leq \frac{\pi_{\theta_0}(y_\star)\left(\sqrt{\gamma}+\pi_{\theta_0}(y_\star)\right)^{h(s)}}{\left[r_{\text{G}}(y_{\text{med}})(1-\pi_{\theta_0}(y_{\text{med}}))\right]^{h(s)}} \\ &\leq \frac{\pi_{\theta_0}(y_\star)(1.3\sqrt{\gamma})^{h(s)}}{(0.05 r_{\text{G}}(y_{\text{med}})^2)^{h(s)}} \\ &= \pi_{\theta_0}(y_\star) \cdot \left(\frac{26\sqrt{\gamma}}{r_{\text{G}}(y_{\text{med}})^2}\right)^{h(s)}. \end{aligned} \tag{32}$$

where the second inequality is from $1-\pi_{\theta_0}(y_{\text{med}}) \geq \pi_{\theta_0}(\mathcal{Y}_{\text{bad}}) \geq 0.05 \cdot r_{\text{G}}(y_{\text{med}})$ and $\pi_{\theta_0}(y_\star) \leq \gamma^{13/14} \leq 0.3\sqrt{\gamma}$ (see Lemma 6). From Equation (32) and Item 2 in Lemma 6 we then get:

$$\begin{aligned}
\pi_{\theta_{t_{\sqrt{\gamma}}}}(y_\star) &\leq \pi_{\theta_0}(y_\star) \cdot (26\sqrt{\gamma}/r_{\text{G}}(y_{\text{med}})^2)^{h(s)} \\
&\leq \frac{\|\phi(y_{\text{med}})\|^2 \gamma^{13/14} r_{\text{G}}(y_{\text{med}})^{2/7}}{40B^2(\Delta_1+\Delta_2)} \cdot \left(\frac{26\sqrt{\gamma}}{r_{\text{G}}(y_{\text{med}})^2}\right)^{h(s)} \\
&< \frac{0.8^2\|\phi(y_{\text{med}})\|^2 \gamma^{13/14} r_{\text{G}}(y_{\text{med}})^{2/7}}{16 \cdot 26^{1/7} B^2(\Delta_1+\Delta_2)} \cdot \left(\frac{26\sqrt{\gamma}}{r_{\text{G}}(y_{\text{med}})^2}\right)^{h(s)} \\
&\leq \frac{C\|\phi(y_{\text{med}})\|^2 \pi_{\theta_0}(y_{\text{med}})^2 \gamma^{\frac{13}{14}+\frac{h(s)}{2}}}{8(\|\phi(y_\star)\|^2 - s) A_{\text{G}}(y_\star;\theta_0)}.
\end{aligned} \tag{33}$$

Here, the last inequality uses $h(s) = 1/7 + 13\alpha/7$, and hence $r_{\text{G}}(y_{\text{med}})^{2/7}(26/r_{\text{G}}(y_{\text{med}})^2)^{h(s)} = 26^{1/7}C$, together with $\pi_{\theta_0}(y_{\text{med}}) \geq 0.8$, $\|\phi(y_\star)\|^2 - s \leq 2B^2$, and $A_{\text{G}}(y_\star;\theta_0) \leq \Delta_1 + \Delta_2$.

Step 2.4: completing the proof of monotonicity in $[t_{\sqrt{\gamma}}, t_\gamma]$ Finally, we conclude the proof of Proposition 5. We first show that $t_{\sqrt{\gamma}} < t_\gamma$, *i.e.*, we prove that the interval $[t_{\sqrt{\gamma}}, t_\gamma]$ is non-empty. Recall that $h(s) = 1/7 + 13\alpha/7$ and $C = (26/r_{\text{G}}(y_{\text{med}})^2)^{13\alpha/7}$. Therefore,

$$\frac{C\gamma^{13/14+h(s)/2}}{\sqrt{\gamma}} = C\gamma^{3/7+h(s)/2} = \sqrt{\gamma}\left(\frac{26\sqrt{\gamma}}{r_{\text{G}}(y_{\text{med}})^2}\right)^{13\alpha/7}.$$

Moreover, by Item 3 in Lemma 6,

$$\frac{2\sqrt{\gamma}}{\Delta_2} \leq \pi_{\theta_0}(\mathcal{Y}_{\text{bad}}) \leq 0.1 r_{\text{G}}(y_{\text{med}}),$$

and since $\Delta_2 = r_{\text{G}}(y_{\text{med}}) - r_{\text{G}}(y_{\text{bad}}) \leq r_{\text{G}}(y_{\text{med}}) + 1 \leq 2$, for $y_{\text{bad}} \in \mathcal{Y}_{\text{bad}}$, we get $\sqrt{\gamma} \leq 0.1 r_{\text{G}}(y_{\text{med}})$. Using also the fact that $13\alpha/7 \in (0, 1/7)$ and $r_{\text{G}}(y_{\text{med}}) \leq 1$, this yields

$$\begin{aligned}
\frac{C\gamma^{13/14+h(s)/2}}{\sqrt{\gamma}} &\leq 0.1 r_{\text{G}}(y_{\text{med}}) \left(\frac{2.6}{r_{\text{G}}(y_{\text{med}})}\right)^{13\alpha/7} \\
&= 0.1 \cdot 2.6^{13\alpha/7} r_{\text{G}}(y_{\text{med}})^{1-13\alpha/7} < 1.
\end{aligned}$$

Hence, $C\gamma^{13/14+h(s)/2} < \sqrt{\gamma}$. Thus, the threshold $r_{\text{G}}(y_{\text{med}}) - C\gamma^{13/14+h(s)/2}$ is strictly larger than $r_{\text{G}}(y_{\text{med}}) - \sqrt{\gamma}$, and by continuity of $V_{\text{G}}(\theta_t)$ together with the definitions of $t_{\sqrt{\gamma}}$ and t_γ, it follows that $t_{\sqrt{\gamma}} < t_\gamma$.

Now, we can consider the optimization dynamics over $[t_{\sqrt{\gamma}}, t_\gamma]$ and show that $\pi_{\theta_t}(y_\star)$ continues to decrease during this time interval. Assume by way of contradiction that there exists $t' \in [0, t_\gamma]$ such that $\frac{d}{dt}\pi_{\theta_t}(y_\star)\big|_{t=t'} \geq 0$. Let τ be the initial time in $[0, t_\gamma]$ at which $\frac{d}{dt}\pi_{\theta_t}(y_\star) \geq 0$, *i.e.*:

$$\tau := \min\{t \in [0, t_\gamma] : \tfrac{d}{dt}\pi_{\theta_t}(y_\star) \geq 0\}.$$

By definition, $\frac{d}{dt}\pi_{\theta_t}(y_\star) < 0$, and so $\frac{d}{dt}\pi_{\theta_t}(y_{\text{med}}) > 0$ (Lemma 7), for all $t \in [0, \tau)$. We also know that if τ exists, then $\tau \geq t_{\sqrt{\gamma}}$ since $\frac{d}{dt}\pi_{\theta_t}(y_\star) < 0$ for all $t \in [0, t_{\sqrt{\gamma}}]$. This implies that $\pi_{\theta_\tau}(y_\star) \leq \pi_{\theta_{t_{\sqrt{\gamma}}}}(y_\star)$ and $\pi_{\theta_\tau}(y_{\text{med}}) \geq \pi_{\theta_0}(y_{\text{med}})$.

Now, by the same arguments made in Step 1, it follows that $\frac{d}{dt}\pi_{\theta_t}(\mathcal{Y}_{\text{bad}}) \leq 0$ for all $t \in [0, \tau)$. We continue by showing that $\pi_{\theta_\tau}(\mathcal{Y}_{\text{bad}})^2 \leq \frac{7\|\phi(y_{\text{med}})\|^2 \pi_{\theta_0}(y_{\text{med}})^2 A_{\text{G}}(y_{\text{med}};\theta_\tau)}{8 \max_{z \in \mathcal{Y}_{\text{bad}}} \|\phi(z)\|^2 \Delta_2}$. To see this, notice that:

$$\begin{aligned}
A_{\text{G}}(y_{\text{med}};\theta_\tau) &= (1-\pi_{\theta_\tau}(y_{\text{med}}))r_{\text{G}}(y_{\text{med}}) - \pi_{\theta_\tau}(y_\star)r_{\text{G}}(y_\star) - \sum\nolimits_{z \in \mathcal{Y}_{\text{bad}}} \pi_{\theta_\tau}(z) r_{\text{G}}(z) \\
&= -\Delta_1 \pi_{\theta_\tau}(y_\star) + \Delta_2 \pi_{\theta_\tau}(\mathcal{Y}_{\text{bad}}).
\end{aligned}$$

Thus,

$$\pi_{\theta_\tau}(\mathcal{Y}_{\text{bad}}) = \frac{A_{\text{G}}(y_{\text{med}};\theta_\tau) + \Delta_1 \pi_{\theta_\tau}(y_\star)}{\Delta_2} \leq \frac{1.3 A_{\text{G}}(y_{\text{med}};\theta_\tau)}{\Delta_2},$$

where the inequality is from Item 2 in Lemma 6 since $\pi_{\theta_0}(y_\star) \le \frac{\|\phi(y_{\text{med}})\|^2 r_{\text{G}}(y_{\text{med}})^{2/7}\gamma^{13/14}}{40B^2(\Delta_1+\Delta_2)} \le \frac{r_{\text{G}}(y_{\text{med}})^{2/7}\gamma^{13/14}}{40(\Delta_1+\Delta_2)} \le \frac{0.3\gamma^{13/14}}{\Delta_1(26/r_{\text{G}}(y_{\text{med}})^2)^{1/7}}$, which combined with Equation (32) yields $\Delta_1\pi_{\theta_\tau}(y_\star) \le \Delta_1\pi_{\theta_{t_{\sqrt{\gamma}}}}(y_\star) \le \Delta_1\pi_{\theta_0}(y_\star)\cdot(26\sqrt{\gamma}/r_{\text{G}}(y_{\text{med}})^2)^{h(s)} \le 0.3C\gamma^{13/14+h(s)/2} \le 0.3A_{\text{G}}(y_{\text{med}};\theta_\tau)$.

Squaring both sides of the inequality, we then have:

$$\begin{aligned}
\pi_{\theta_\tau}(\mathcal{Y}_{\text{bad}})^2 &\le \frac{1.69A_{\text{G}}(y_{\text{med}};\theta_\tau)^2}{\Delta_2^2} \\
&< \frac{7\|\phi(y_{\text{med}})\|^2\pi_{\theta_0}(y_{\text{med}})^2A_{\text{G}}(y_{\text{med}};\theta_\tau)}{8\max_{z\in\mathcal{Y}_{\text{bad}}}\|\phi(z)\|^2\Delta_2}\cdot\frac{2\max_{z\in\mathcal{Y}_{\text{bad}}}\|\phi(z)\|^2A_{\text{G}}(y_{\text{med}};\theta_\tau)}{\|\phi(y_{\text{med}})\|^2\Delta_2\pi_{\theta_0}(y_{\text{med}})^2} \\
&\le \frac{7\|\phi(y_{\text{med}})\|^2\pi_{\theta_0}(y_{\text{med}})^2A_{\text{G}}(y_{\text{med}};\theta_\tau)}{8\max_{z\in\mathcal{Y}_{\text{bad}}}\|\phi(z)\|^2\Delta_2}\cdot\frac{2A_{\text{G}}(y_{\text{med}};\theta_0)}{\Delta_2\pi_{\theta_0}(y_{\text{med}})^2}.
\end{aligned}$$

The second inequality is from $1.69\times 8 < 2\times 7$ and the last inequality is by $\max_{z\in\mathcal{Y}_{\text{bad}}}\|\phi(z)\|^2 \le \|\phi(y_{\text{med}})\|^2$ (Assumption 4) and $A_{\text{G}}(y_{\text{med}};\theta_\tau) \le A_{\text{G}}(y_{\text{med}};\theta_0)$. We now prove that the second term on the right-hand side of the inequality above is at most 1. To do so, we upper bound $A_{\text{G}}(y_{\text{med}};\theta_0)$:

$$\begin{aligned}
A_{\text{G}}(y_{\text{med}};\theta_0) &= (1-\pi_{\theta_0}(y_{\text{med}}))r_{\text{G}}(y_{\text{med}}) - \pi_{\theta_0}(y_\star)r_{\text{G}}(y_\star) - \sum\nolimits_{z\in\mathcal{Y}_{\text{bad}}}\pi_{\theta_0}(z)r_{\text{G}}(z) \\
&\le (1-\pi_{\theta_0}(y_{\text{med}}))r_{\text{G}}(y_{\text{med}}) + \pi_{\theta_0}(\mathcal{Y}_{\text{bad}}) \\
&\le 0.2r_{\text{G}}(y_{\text{med}}) + 0.1r_{\text{G}}(y_{\text{med}}) \\
&= 0.3r_{\text{G}}(y_{\text{med}}) \\
&< \frac{0.8^2 r_{\text{G}}(y_{\text{med}})}{2} \\
&\le \frac{\Delta_2\pi_{\theta_0}(y_{\text{med}})^2}{2}.
\end{aligned}$$

The first inequality is due to $r_{\text{G}}(z) \ge -1$ for all $z \in \mathcal{Y}_{\text{bad}}$; the second inequality is from Lemma 6 since $\pi_{\theta_0}(\mathcal{Y}_{\text{bad}}) \le 0.1r_{\text{G}}(y_{\text{med}}) \le 0.1$ and $\pi_{\theta_0}(y_{\text{med}}) \ge 0.8$; and the last inequality is from $r_{\text{G}}(y_{\text{bad}}) \le 0$ and $\pi_{\theta_0}(y_{\text{med}}) \ge 0.8$. Therefore, $\frac{2A_{\text{G}}(y_{\text{med}};\theta_0)}{\Delta_2\pi_{\theta_0}(y_{\text{med}})^2} \le 1$, which leads to

$$\pi_{\theta_\tau}(\mathcal{Y}_{\text{bad}})^2 \le \frac{7\|\phi(y_{\text{med}})\|^2\pi_{\theta_0}(y_{\text{med}})^2A_{\text{G}}(y_{\text{med}};\theta_\tau)}{8\max_{z\in\mathcal{Y}_{\text{bad}}}\|\phi(z)\|^2\Delta_2}. \tag{34}$$

Let us now consider $\frac{d}{dt}\pi_{\theta_t}(y_\star)\big|_{t=\tau}$ and obtain a contradiction by showing that it is negative. We start with the following computations, based on the expression derived in Lemma 13:

$$\begin{aligned}
\frac{d}{dt}\pi_{\theta_t}(y_\star)\Big|_{t=\tau} &= \pi_{\theta_\tau}(y_\star)\Big[\pi_{\theta_\tau}(y_\star)A_{\text{G}}(y_\star;\theta_\tau)\big(\|\phi(y_\star)\|^2 - \|\phi(y_\star)\|^2\pi_{\theta_\tau}(y_\star) - s\pi_{\theta_\tau}(y_{\text{med}})\big) \\
&\qquad + \pi_{\theta_\tau}(y_{\text{med}})A_{\text{G}}(y_{\text{med}};\theta_\tau)\big(s - \|\phi(y_{\text{med}})\|^2\pi_{\theta_\tau}(y_{\text{med}}) - s\pi_{\theta_\tau}(y_\star)\big) \\
&\qquad - \sum\nolimits_{z\in\mathcal{Y}_{\text{bad}}}\pi_{\theta_\tau}(z)^2A_{\text{G}}(z;\theta_\tau)\|\phi(z)\|^2\Big] \\
&\le \pi_{\theta_\tau}(y_\star)\Big[\pi_{\theta_\tau}(y_\star)A_{\text{G}}(y_\star;\theta_\tau)\big(\|\phi(y_\star)\|^2 - \|\phi(y_\star)\|^2\pi_{\theta_\tau}(y_\star) - s\pi_{\theta_\tau}(y_{\text{med}})\big) \\
&\qquad + \pi_{\theta_\tau}(y_{\text{med}})A_{\text{G}}(y_{\text{med}};\theta_\tau)\big(s - \|\phi(y_{\text{med}})\|^2\pi_{\theta_\tau}(y_{\text{med}}) - s\pi_{\theta_\tau}(y_\star)\big) \\
&\qquad + \max_{z\in\mathcal{Y}_{\text{bad}}}\|\phi(z)\|^2\big(\Delta_2 - \gamma^{\frac{13}{14}+\frac{h(s)}{2}}\big)\sum\nolimits_{z\in\mathcal{Y}_{\text{bad}}}\pi_{\theta_\tau}(z)^2\Big] \\
&\le \pi_{\theta_\tau}(y_\star)\Big[(\|\phi(y_\star)\|^2 - s)\pi_{\theta_\tau}(y_\star)A_{\text{G}}(y_\star;\theta_\tau)\big(1-\pi_{\theta_\tau}(y_\star)\big) \\
&\qquad - \|\phi(y_{\text{med}})\|^2\pi_{\theta_\tau}(y_{\text{med}})^2A_{\text{G}}(y_{\text{med}};\theta_\tau) \\
&\qquad + \max_{z\in\mathcal{Y}_{\text{bad}}}\|\phi(z)\|^2(\Delta_2 - \gamma^{\frac{13}{14}+\frac{h(s)}{2}})\sum\nolimits_{z\in\mathcal{Y}_{\text{bad}}}\pi_{\theta_\tau}(z)^2\Big].
\end{aligned}$$

The first inequality is from $-A_{\text{G}}(z;\theta_\tau) = \Delta_2 - A_{\text{G}}(y_{\text{med}};\theta_\tau) \le \Delta_2 - \gamma^{\frac{13}{14}+\frac{h(s)}{2}}$ for all $z \in \mathcal{Y}_{\text{bad}}$ (this holds since $\tau \le t_\gamma$ and $C \ge 1$) and the second inequality is from $s \le 0$ and $\pi_{\theta_\tau}(y_{\text{med}}) \le$

$1-\pi_{\theta_\tau}(y_\star)$. We continue upper bounding $\frac{d}{dt}\pi_{\theta_t}(y_\star)\big|_{t=\tau}$ as follows:

$$\begin{aligned}\frac{d}{dt}\pi_{\theta_t}(y_\star)\Big|_{t=\tau} &\le \pi_{\theta_\tau}(y_\star)\Big[(\|\phi(y_\star)\|^2-s)\pi_{\theta_{t_{\sqrt{\gamma}}}}(y_\star)A_{\mathrm{G}}(y_\star;\theta_0)\big(1-\pi_{\theta_{t_{\sqrt{\gamma}}}}(y_\star)\big)\\ &\qquad -\|\phi(y_{\mathrm{med}})\|^2\pi_{\theta_\tau}(y_{\mathrm{med}})^2A_{\mathrm{G}}(y_{\mathrm{med}};\theta_\tau)\\ &\qquad +\max_{z\in\mathcal{Y}_{\mathrm{bad}}}\|\phi(z)\|^2\big(\Delta_2-\gamma^{\frac{13}{14}+\frac{h(s)}{2}}\big)\pi_{\theta_\tau}(\mathcal{Y}_{\mathrm{bad}})^2\Big]\\ &\le \pi_{\theta_\tau}(y_\star)\Big[(\|\phi(y_\star)\|^2-s)\pi_{\theta_{t_{\sqrt{\gamma}}}}(y_\star)A_{\mathrm{G}}(y_\star;\theta_0)\big(1-\pi_{\theta_{t_{\sqrt{\gamma}}}}(y_\star)\big)\\ &\qquad -(\|\phi(y_{\mathrm{med}})\|^2/8)\pi_{\theta_\tau}(y_{\mathrm{med}})^2A_{\mathrm{G}}(y_{\mathrm{med}};\theta_\tau)\Big]\\ &\le \pi_{\theta_\tau}(y_\star)\Big[(\|\phi(y_\star)\|^2-s)\pi_{\theta_{t_{\sqrt{\gamma}}}}(y_\star)A_{\mathrm{G}}(y_\star;\theta_0)\big(1-\pi_{\theta_{t_{\sqrt{\gamma}}}}(y_\star)\big)\\ &\qquad -(C\|\phi(y_{\mathrm{med}})\|^2/8)\pi_{\theta_\tau}(y_{\mathrm{med}})^2\gamma^{\frac{13}{14}+\frac{h(s)}{2}}\Big].\end{aligned}$$

The first inequality is from $A_{\mathrm{G}}(y_\star;\theta_\tau)\le A_{\mathrm{G}}(y_\star;\theta_0)$, $\pi_{\theta_\tau}(y_\star)\le\pi_{\theta_{t_{\sqrt{\gamma}}}}(y_\star)$, $\pi_{\theta_\tau}(y_{\mathrm{med}})\ge\pi_{\theta_{t_{\sqrt{\gamma}}}}(y_{\mathrm{med}})$, and $\sum_{z\in\mathcal{Y}_{\mathrm{bad}}}\pi_{\theta_\tau}(z)^2\le\pi_{\theta_\tau}(\mathcal{Y}_{\mathrm{bad}})^2$; the second inequality is from

$$\begin{aligned}\max_{z\in\mathcal{Y}_{\mathrm{bad}}}\|\phi(z)\|^2(\Delta_2-\gamma^{\frac{13}{14}+\frac{h(s)}{2}})\pi_{\theta_\tau}(\mathcal{Y}_{\mathrm{bad}})^2 &\le \max_{z\in\mathcal{Y}_{\mathrm{bad}}}\|\phi(z)\|^2\Delta_2\pi_{\theta_\tau}(\mathcal{Y}_{\mathrm{bad}})^2\\ &\le \frac{7\|\phi(y_{\mathrm{med}})\|^2\pi_{\theta_\tau}(y_{\mathrm{med}})^2A_{\mathrm{G}}(y_{\mathrm{med}};\theta_\tau)}{8},\end{aligned}$$

which is implied by Equation (34), and $\pi_{\theta_\tau}(y_{\mathrm{med}})\ge\pi_{\theta_0}(y_{\mathrm{med}})$; and the last inequality is from $A_{\mathrm{G}}(y_{\mathrm{med}};\theta_\tau)\ge A_{\mathrm{G}}(y_{\mathrm{med}};\theta_{t_\gamma})=C\gamma^{\frac{13}{14}+\frac{h(s)}{2}}$.

Therefore, using the upper bound on $\pi_{\theta_{t_{\sqrt{\gamma}}}}(y_\star)$ from Equation (33):

$$\begin{aligned}&\frac{d}{dt}\pi_{\theta_t}(y_\star)\Big|_{t=\tau}\\ &\le \pi_{\theta_\tau}(y_\star)\Big[(\|\phi(y_\star)\|^2-s)\pi_{\theta_{t_{\sqrt{\gamma}}}}(y_\star)A_{\mathrm{G}}(y_\star;\theta_0)\big(1-\pi_{\theta_{t_{\sqrt{\gamma}}}}(y_\star)\big)\\ &\qquad -(C\|\phi(y_{\mathrm{med}})\|^2/8)\pi_{\theta_\tau}(y_{\mathrm{med}})^2\gamma^{\frac{13}{14}+\frac{h(s)}{2}}\Big]\\ &< \pi_{\theta_\tau}(y_\star)\Big[(C\|\phi(y_{\mathrm{med}})\|^2/8)\pi_{\theta_0}(y_{\mathrm{med}})^2\gamma^{\frac{13}{14}+\frac{h(s)}{2}}-(C\|\phi(y_{\mathrm{med}})\|^2/8)\pi_{\theta_\tau}(y_{\mathrm{med}})^2\gamma^{\frac{13}{14}+\frac{h(s)}{2}}\Big]\\ &= \frac{C\pi_{\theta_\tau}(y_\star)\|\phi(y_{\mathrm{med}})\|^2\gamma^{\frac{13}{14}+\frac{h(s)}{2}}}{8}\Big(\pi_{\theta_0}(y_{\mathrm{med}})^2-\pi_{\theta_\tau}(y_{\mathrm{med}})^2\Big)\\ &\le 0,\end{aligned}$$

where the last inequality is from $\pi_{\theta_\tau}(y_{\mathrm{med}})\ge\pi_{\theta_0}(y_{\mathrm{med}})$. Thus, the strict inequality above implies $\frac{d}{dt}\pi_{\theta_t}(y_\star)\big|_{t=\tau}<0$, which contradicts the fact that τ is the initial time at which $\frac{d}{dt}\pi_{\theta_t}(y_\star)\big|_{t=\tau}\ge 0$. Hence, it must be that $\frac{d}{dt}\pi_{\theta_t}(y_\star)<0$ for all $t\in[0,t_\gamma]$. Additionally, by Lemma 7, we also have $\frac{d}{dt}\pi_{\theta_t}(y_{\mathrm{med}})>0$ for all $t\in[0,t_\gamma]$.

To conclude, we showed that $\frac{d}{dt}\pi_{\theta_t}(y_\star)<0$ and $\frac{d}{dt}\pi_{\theta_t}(y_{\mathrm{med}})>0$ for all $t\in[0,t_\gamma]$. Thus, $\min_{t\in[0,t_\gamma]}\pi_{\theta_t}(y_\star)=\pi_{\theta_{t_\gamma}}(y_\star)$. Furthermore, since $t_{\sqrt{\gamma}}<t_\gamma$, by Equation (33) it follows that

$$\pi_{\theta_{t_\gamma}}(y_\star)<\pi_{\theta_{t_{\sqrt{\gamma}}}}(y_\star)\le\frac{C\|\phi(y_{\mathrm{med}})\|^2\pi_{\theta_0}(y_{\mathrm{med}})^2\gamma^{\frac{13}{14}+\frac{h(s)}{2}}}{8(\|\phi(y_\star)\|^2-s)A_{\mathrm{G}}(y_\star;\theta_0)},$$

completing the proof of Proposition 5. □

Based on Proposition 5, we can now complete the proof of Case I.

Proof of Theorem 3, Case I. By Proposition 5, as long as $V_{\mathrm{G}}(\theta_t)\le r_{\mathrm{G}}(y_{\mathrm{med}})-C\gamma^{\frac{13}{14}+\frac{h(s)}{2}}$, we know that $\pi_{\theta_t}(y_\star)$ is monotonically decreasing and $\pi_{\theta_t}(y_{\mathrm{med}})$ is monotonically increasing. Recall

that t_γ denotes the initial time at which $V_{\mathrm{G}}(\theta_t) \geq r_{\mathrm{G}}(y_{\mathrm{med}}) - C\gamma^{\frac{13}{14}+\frac{h(s)}{2}}$. Since $V_{\mathrm{G}}(\theta_t)$ is continuous in t, at time t_γ it must be that $V_{\mathrm{G}}(\theta_{t_\gamma}) = r_{\mathrm{G}}(y_{\mathrm{med}}) - C\gamma^{\frac{13}{14}+\frac{h(s)}{2}}$ and t_γ is the initial time at which this equality holds.

We first prove that at t_γ the policy is highly concentrated on y_{med} by upper bounding $1 - \pi_{\theta_{t_\gamma}}(y_{\mathrm{med}})$. This follows from

$$\begin{aligned} r_{\mathrm{G}}(y_{\mathrm{med}}) - C\gamma^{\frac{13}{14}+\frac{h(s)}{2}} &= V_{\mathrm{G}}(\theta_{t_\gamma}) \\ &= r_{\mathrm{G}}(y_{\mathrm{med}})\pi_{\theta_{t_\gamma}}(y_{\mathrm{med}}) + \pi_{\theta_{t_\gamma}}(y_\star) r_{\mathrm{G}}(y_\star) + \sum\nolimits_{z\in\mathcal{Y}_{\mathrm{bad}}} \pi_{\theta_{t_\gamma}}(z) r_{\mathrm{G}}(z), \end{aligned}$$

which implies:

$$\begin{aligned} 1 - \pi_{\theta_{t_\gamma}}(y_{\mathrm{med}}) &= \frac{1}{r_{\mathrm{G}}(y_{\mathrm{med}})}\left(C\gamma^{\frac{13}{14}+\frac{h(s)}{2}} + \pi_{\theta_{t_\gamma}}(y_\star) r_{\mathrm{G}}(y_\star) + \sum_{z\in\mathcal{Y}_{\mathrm{bad}}} \pi_{\theta_{t_\gamma}}(z) r_{\mathrm{G}}(z) \right) \\ &\leq \frac{1}{r_{\mathrm{G}}(y_{\mathrm{med}})}\left(C\gamma^{\frac{13}{14}+\frac{h(s)}{2}} + \frac{C\|\phi(y_{\mathrm{med}})\|^2 \pi_{\theta_0}(y_{\mathrm{med}})^2 \gamma^{\frac{13}{14}+\frac{h(s)}{2}}}{8(\|\phi(y_\star)\|^2 - s) A_{\mathrm{G}}(y_\star;\theta_0)} \right) \\ &= \frac{K\gamma^{\frac{13}{14}+\frac{h(s)}{2}}}{r_{\mathrm{G}}(y_{\mathrm{med}})}. \end{aligned} \tag{35}$$

The inequality is due to $\pi_{\theta_{t_\gamma}}(y_\star) \leq \frac{C\|\phi(y_{\mathrm{med}})\|^2 \pi_{\theta_0}(y_{\mathrm{med}})^2 \gamma^{\frac{13}{14}+\frac{h(s)}{2}}}{8(\|\phi(y_\star)\|^2 - s) A_{\mathrm{G}}(y_\star;\theta_0)}$ from Proposition 5, together with $r_{\mathrm{G}}(y_\star) \leq 1$ and $\sum_{z\in\mathcal{Y}_{\mathrm{bad}}} \pi_{\theta_{t_\gamma}}(z) r_{\mathrm{G}}(z) \leq 0$, and the last equality is by the definition of K in Theorem 3.

Now, the fact that the policy assigns high probability to y_{med} at t_γ allows us to lower bound $t_\star$—the initial time at which $V_{\mathrm{G}}(\theta_t) \geq r_{\mathrm{G}}(y_\star) - \epsilon$. For any time $T \geq t_\gamma$, we have:

$$\begin{aligned} \|\theta_T - \theta_{t_\gamma}\| &= \left\| \int_{t_\gamma}^{T} \frac{d}{dt}\theta_t dt \right\| \\ &\leq \int_{t_\gamma}^{T} \left\| \frac{d}{dt}\theta_t \right\| dt \\ &= \int_{t_\gamma}^{T} \|\nabla V_{\mathrm{G}}(\theta_t)\| dt \\ &\leq \int_{t_\gamma}^{T} 2B(\Delta_1 + \Delta_2)(1 - \pi_{\theta_t}(y_{\mathrm{med}})) dt, \end{aligned} \tag{36}$$

where the last inequality is by Lemma 11. We proceed by upper bounding the rate at which $1 - \pi_{\theta_t}(y_{\mathrm{med}})$ increases (*i.e.*, the rate at which $\pi_{\theta_t}(y_{\mathrm{med}})$ decreases). From the chain rule we obtain:

$$\begin{aligned} \frac{d}{dt}(1 - \pi_{\theta_t}(y_{\mathrm{med}})) &= \left\langle \nabla(1 - \pi_{\theta_t}(y_{\mathrm{med}})), \frac{d}{dt}\theta_t \right\rangle \\ &= \langle \nabla(1 - \pi_{\theta_t}(y_{\mathrm{med}})), \nabla V_{\mathrm{G}}(\theta_t) \rangle \\ &= \left\langle \nabla(\pi_{\theta_t}(y_\star) + \sum\nolimits_{z\in\mathcal{Y}_{\mathrm{bad}}} \pi_{\theta_t}(z)), \nabla V_{\mathrm{G}}(\theta_t) \right\rangle. \end{aligned}$$

For all $y \in \mathcal{Y}$, the gradient of $\pi_{\theta_t}(y)$ with respect to θ_t can be computed as follows:

$$\nabla \pi_{\theta_t}(y) = \pi_{\theta_t}(y)\Big(\phi(y) - \sum\nolimits_{z\in\mathcal{Y}} \pi_{\theta_t}(z)\phi(z)\Big) = \pi_{\theta_t}(y)\big(\phi(y) - \bar{\phi}_{\theta_t}\big),$$

where $\bar{\phi}_{\theta_t} := \sum_{z\in\mathcal{Y}} \pi_{\theta_t}(z)\phi(z)$. This yields:

$$\begin{aligned} \frac{d}{dt}(1 - \pi_{\theta_t}(y_{\mathrm{med}})) &\leq \sum\nolimits_{y\in\mathcal{Y}\setminus\{y_{\mathrm{med}}\}} \pi_{\theta_t}(y) \big| \langle \phi(y) - \bar{\phi}_{\theta_t}, \nabla V_{\mathrm{G}}(\theta_t) \rangle \big| \\ &\leq \sum\nolimits_{y\in\mathcal{Y}\setminus\{y_{\mathrm{med}}\}} \pi_{\theta_t}(y) \|\phi(y) - \bar{\phi}_{\theta_t}\| \|\nabla V_{\mathrm{G}}(\theta_t)\| \\ &\leq \max\nolimits_{y\in\mathcal{Y}\setminus\{y_{\mathrm{med}}\}} \|\phi(y) - \bar{\phi}_{\theta_t}\| \|\nabla V_{\mathrm{G}}(\theta_t)\| \cdot \sum\nolimits_{y\in\mathcal{Y}\setminus\{y_{\mathrm{med}}\}} \pi_{\theta_t}(y) \\ &\leq 4B^2(\Delta_1 + \Delta_2)(1 - \pi_{\theta_t}(y_{\mathrm{med}}))^2. \end{aligned}$$

Here, the second inequality is due to the Cauchy-Schwarz inequality; the third inequality is due to $\|\phi(z) - \bar{\phi}_{\theta_t}\| \leq \max_{y \in \mathcal{Y} \setminus \{y_{\text{med}}\}} \|\phi(y) - \bar{\phi}_{\theta_t}\|$ for all $z \in \mathcal{Y} \setminus \{y_{\text{med}}\}$; and the last inequality is from $\max_{y \in \mathcal{Y} \setminus \{y_{\text{med}}\}} \|\phi(y) - \bar{\phi}_{\theta_t}\| \leq 2B$ and Lemma 11.

This implies that:

$$\frac{d}{dt}\left(\frac{1}{1-\pi_{\theta_t}(y_{\text{med}})}\right) = -\frac{\frac{d}{dt}(1-\pi_{\theta_t}(y_{\text{med}}))}{(1-\pi_{\theta_t}(y_{\text{med}}))^2} \geq -4B^2(\Delta_1+\Delta_2).$$

Integrating both sides from t_γ to t, for any $t \geq t_\gamma$, leads to:

$$\frac{1}{1-\pi_{\theta_t}(y_{\text{med}})} - \frac{1}{1-\pi_{\theta_{t_\gamma}}(y_{\text{med}})} \geq -4B^2(\Delta_1+\Delta_2)(t-t_\gamma).$$

Hence, for any $t \geq t_\gamma$, such that

$$4B^2(\Delta_1+\Delta_2)(1-\pi_{\theta_{t_\gamma}}(y_{\text{med}}))(t-t_\gamma) < 1,$$

it holds that

$$1-\pi_{\theta_t}(y_{\text{med}}) \leq \frac{1-\pi_{\theta_{t_\gamma}}(y_{\text{med}})}{1-4B^2(\Delta_1+\Delta_2)(1-\pi_{\theta_{t_\gamma}}(y_{\text{med}}))(t-t_\gamma)}.$$

Thus, for any $T \geq t_\gamma$ satisfying $4B^2(\Delta_1+\Delta_2)(1-\pi_{\theta_{t_\gamma}}(y_{\text{med}}))(T-t_\gamma) < 1$, we can plug this upper bound into Equation (36) to get:

$$\begin{aligned}
\|\theta_T - \theta_{t_\gamma}\| &\leq 2B(\Delta_1+\Delta_2)\int_{t_\gamma}^{T}(1-\pi_{\theta_t}(y_{\text{med}}))dt \\
&= 2B(\Delta_1+\Delta_2)\int_0^{T-t_\gamma}\frac{1-\pi_{\theta_{t_\gamma}}(y_{\text{med}})}{1-4B^2(\Delta_1+\Delta_2)(1-\pi_{\theta_{t_\gamma}}(y_{\text{med}}))u}du \qquad (37)\\
&= \frac{1}{2B}\ln\left(\frac{1}{1-4B^2(\Delta_1+\Delta_2)(1-\pi_{\theta_{t_\gamma}}(y_{\text{med}}))(T-t_\gamma)}\right).
\end{aligned}$$

By Lemma 12, V_{G} is B-Lipschitz with respect to θ, and so:

$$|V_{\text{G}}(\theta_T) - V_{\text{G}}(\theta_{t_\gamma})| \leq B\|\theta_T - \theta_{t_\gamma}\| \leq \frac{1}{2}\ln\left(\frac{1}{1-4B^2(\Delta_1+\Delta_2)(1-\pi_{\theta_{t_\gamma}}(y_{\text{med}}))(T-t_\gamma)}\right).$$

If $t_\star = \infty$, then the desired lower bound holds trivially. Hence, assume $t_\star < \infty$. By the definition of $t_\star$ and continuity of $V_{\text{G}}(\theta_t)$, $V_{\text{G}}(\theta_{t_\star}) = r_{\text{G}}(y_\star) - \epsilon$. In particular, $t_\star > t_\gamma$ since $V_{\text{G}}(\theta_{t_\star}) = r_{\text{G}}(y_\star) - \epsilon > r_{\text{G}}(y_{\text{med}}) > V_{\text{G}}(\theta_{t_\gamma})$. Therefore,

$$V_{\text{G}}(\theta_{t_\star}) - V_{\text{G}}(\theta_{t_\gamma}) = r_{\text{G}}(y_\star) - \epsilon - r_{\text{G}}(y_{\text{med}}) + C\gamma^{\frac{13}{14}+\frac{h(s)}{2}} \geq \Delta_1 - \epsilon.$$

If $4B^2(\Delta_1+\Delta_2)(1-\pi_{\theta_{t_\gamma}}(y_{\text{med}}))(t_\star - t_\gamma) \geq 1$, then directly

$$t_\star - t_\gamma \geq \frac{1}{4B^2(\Delta_1+\Delta_2)(1-\pi_{\theta_{t_\gamma}}(y_{\text{med}}))} \geq \frac{1-e^{-2(\Delta_1-\epsilon)}}{4B^2(\Delta_1+\Delta_2)(1-\pi_{\theta_{t_\gamma}}(y_{\text{med}}))},$$

since $1 - e^{-2(\Delta_1-\epsilon)} \leq 1$. Otherwise, we apply the bound in Equation (37) to get:

$$\Delta_1 - \epsilon \leq V_{\text{G}}(\theta_{t_\star}) - V_{\text{G}}(\theta_{t_\gamma}) \leq \frac{1}{2}\ln\left(\frac{1}{1-4B^2(\Delta_1+\Delta_2)(1-\pi_{\theta_{t_\gamma}}(y_{\text{med}}))(t_\star-t_\gamma)}\right),$$

which implies

$$t_\star - t_\gamma \geq \frac{1-e^{-2(\Delta_1-\epsilon)}}{4B^2(\Delta_1+\Delta_2)(1-\pi_{\theta_{t_\gamma}}(y_{\text{med}}))}.$$

Therefore, in either case:

$$\begin{aligned}
t_\star &\geq t_\star - t_\gamma \\
&\geq \frac{1 - e^{-2(\Delta_1 - \epsilon)}}{4B^2(\Delta_1 + \Delta_2)(1 - \pi_{\theta_{t_\gamma}}(y_{\mathrm{med}}))} \\
&\geq \frac{r_{\mathrm{G}}(y_{\mathrm{med}})(1 - e^{-2(\Delta_1 - \epsilon)})}{4B^2(\Delta_1 + \Delta_2)K\gamma^{\frac{13}{14} + \frac{h(s)}{2}}} \\
&= \frac{r_{\mathrm{G}}(y_{\mathrm{med}})M'^{14/13+\alpha}(1 - e^{-2(\Delta_1 - \epsilon)})}{4B^2(\Delta_1 + \Delta_2)K} \cdot \pi_{\theta_0}(y_\star)^{-14/13-\alpha} \\
&= \Omega\Big(\pi_{\theta_0}(y_\star)^{-14/13-\alpha}\Big).
\end{aligned}$$

where the third inequality follows from Equation (35). This completes the proof for Case I of Theorem 3. □

D.5.2 Proof of Case II

Suppose that gradient flow is used to maximize the expected reward with respect to r_{P}, which assigns y_{med} a low reward $r_{\mathrm{P}}(y_{\mathrm{med}}) = \min_{y \in \mathcal{Y}_{\mathrm{bad}}} r_{\mathrm{G}}(y)$. Similarly to the proof of Case II in Theorem 2, we show that in this case $\pi_{\theta_t}(y_\star)$ increases already from $t = 0$ and that $t_\star^{\mathrm{no-med}} = \mathcal{O}(\pi_{\theta_0}(y_\star)^{-1})$, where recall that $t_\star^{\mathrm{no-med}}$ is the initial time at which $V_{\mathrm{G}}(\theta_t) \geq r_{\mathrm{G}}(y_\star) - \epsilon$ when maximizing r_{P}.

Proof of Theorem 3, Case II. Notice that $V_{\mathrm{G}}(\theta) \geq V_{\mathrm{P}}(\theta)$ for all $\theta \in \mathbb{R}^D$ since $r_{\mathrm{G}}(y) \geq r_{\mathrm{P}}(y)$ for all $y \in \mathcal{Y}$. Thus, when $V_{\mathrm{P}}(\theta_t) \geq r_{\mathrm{G}}(y_\star) - \epsilon$ it also holds that $V_{\mathrm{G}}(\theta_t) \geq r_{\mathrm{G}}(y_\star) - \epsilon$. We can therefore focus on V_{P} and consider the time it takes until it reaches a value of $r_{\mathrm{G}}(y_\star) - \epsilon$. In particular, denote by $t_{\mathrm{P}}^{\mathrm{no-med}}$ the initial time at which $V_{\mathrm{P}}(\theta_t) \geq r_{\mathrm{G}}(y_\star) - \epsilon$, *i.e.*:

$$t_{\mathrm{P}}^{\mathrm{no-med}} := \min\{t \geq 0 : V_{\mathrm{P}}(\theta_t) \geq r_{\mathrm{G}}(y_\star) - \epsilon\}.$$

The discussion above implies that $t_\star^{\mathrm{no-med}} \leq t_{\mathrm{P}}^{\mathrm{no-med}}$, and so it suffices to upper bound $t_{\mathrm{P}}^{\mathrm{no-med}}$. For simplicity of notation, denote $\rho := \frac{r_{\mathrm{G}}(y_\star)+1-\epsilon}{r_{\mathrm{G}}(y_\star)+1}$. Observe that when $\pi_{\theta_t}(y_\star) \geq \rho$, it holds that $V_{\mathrm{P}}(\theta_t) \geq r_{\mathrm{G}}(y_\star) - \epsilon$. This is because:

$$\begin{aligned}
V_{\mathrm{P}}(\theta_t) &\geq \pi_{\theta_t}(y_\star) r_{\mathrm{P}}(y_\star) - (1 - \pi_{\theta_t}(y_\star)) \\
&= \pi_{\theta_t}(y_\star)(r_{\mathrm{G}}(y_\star) + 1) - 1 \\
&\geq r_{\mathrm{G}}(y_\star) - \epsilon.
\end{aligned}$$

Thus, $\pi_{\theta_t}(y_\star) \leq \rho$ for all $t \in [0, t_{\mathrm{P}}^{\mathrm{no-med}}]$.

We prove that

$$T := \frac{1}{(1-\rho)^2 A_{\mathrm{P}}(y_\star; \theta_0) \|\phi(y_\star)\|^2} \left(\frac{1}{\pi_{\theta_0}(y_\star)} - \frac{1}{\rho} \right) \geq t_{\mathrm{P}}^{\mathrm{no-med}},$$

from which the desired upper bound on $t_\star^{\mathrm{no-med}}$ immediately follows by $t_\star^{\mathrm{no-med}} \leq t_{\mathrm{P}}^{\mathrm{no-med}}$ and substituting the value of ρ.

Assume by way of contradiction that $t_{\mathrm{P}}^{\mathrm{no-med}} > T$. Let us lower bound the rate at which $\pi_{\theta_t}(y_\star)$ increases over $[0, t_{\mathrm{P}}^{\mathrm{no-med}}]$. Notice first that $V_{\mathrm{P}}(\theta_0) > r_{\mathrm{P}}(y)$ for all $y \in \mathcal{Y} \setminus \{y_\star\}$ since all such y have the same proxy reward value, which is lower than $r_{\mathrm{P}}(y_\star) = r_{\mathrm{G}}(y_\star)$. Now, consider the time

derivative of $\pi_{\theta_t}(y_\star)$, starting from the expression derived in Lemma 13:

$$\begin{aligned}
&\frac{d}{dt}\pi_{\theta_t}(y_\star) \\
&= \pi_{\theta_t}(y_\star)\Big[\pi_{\theta_t}(y_\star)A_{\mathrm{P}}(y_\star;\theta_t)\big(\|\phi(y_\star)\|^2 - \|\phi(y_\star)\|^2\pi_{\theta_t}(y_\star) - s\pi_{\theta_t}(y_{\mathrm{med}})\big) \\
&\qquad + \pi_{\theta_t}(y_{\mathrm{med}})A_{\mathrm{P}}(y_{\mathrm{med}};\theta_t)\big(s - \|\phi(y_{\mathrm{med}})\|^2\pi_{\theta_t}(y_{\mathrm{med}}) - s\pi_{\theta_t}(y_\star)\big) \\
&\qquad - \sum\nolimits_{z\in\mathcal{Y}_{\mathrm{bad}}} \pi_{\theta_t}(z)^2 A_{\mathrm{P}}(z;\theta_t)\|\phi(z)\|^2\Big] \\
&\geq \pi_{\theta_t}(y_\star)\Big[\pi_{\theta_t}(y_\star)A_{\mathrm{P}}(y_\star;\theta_t)\big(\|\phi(y_\star)\|^2 - \|\phi(y_\star)\|^2\pi_{\theta_t}(y_\star) - s\pi_{\theta_t}(y_{\mathrm{med}})\big) \\
&\qquad + \pi_{\theta_t}(y_{\mathrm{med}})A_{\mathrm{P}}(y_{\mathrm{med}};\theta_t)\big(s - \|\phi(y_{\mathrm{med}})\|^2\pi_{\theta_t}(y_{\mathrm{med}}) - s\pi_{\theta_t}(y_\star)\big)\Big] \\
&\geq \pi_{\theta_t}(y_\star)\Big[\pi_{\theta_t}(y_\star)A_{\mathrm{P}}(y_\star;\theta_t)\big(\|\phi(y_\star)\|^2 - \|\phi(y_\star)\|^2\pi_{\theta_t}(y_\star)\big) \\
&\qquad + \pi_{\theta_t}(y_{\mathrm{med}})A_{\mathrm{P}}(y_{\mathrm{med}};\theta_t)\big(s - \|\phi(y_{\mathrm{med}})\|^2\pi_{\theta_t}(y_{\mathrm{med}}) - s\pi_{\theta_t}(y_\star)\big)\Big] \\
&\geq \pi_{\theta_t}(y_\star)^2\|\phi(y_\star)\|^2 A_{\mathrm{P}}(y_\star;\theta_t)\big(1 - \pi_{\theta_t}(y_\star)\big).
\end{aligned}$$

Here, the first inequality is from $A_{\mathrm{P}}(z;\theta_t) \leq 0$ for all $z \in \mathcal{Y}_{\mathrm{bad}}$; the second inequality is from $s \leq 0$; and the last inequality is from $A_{\mathrm{P}}(y_{\mathrm{med}};\theta_t) \leq A_{\mathrm{P}}(y_{\mathrm{med}};\theta_0) \leq 0$ together with

$$s - \|\phi(y_{\mathrm{med}})\|^2\pi_{\theta_t}(y_{\mathrm{med}}) - s\pi_{\theta_t}(y_\star) = s(1 - \pi_{\theta_t}(y_\star)) - \|\phi(y_{\mathrm{med}})\|^2\pi_{\theta_t}(y_{\mathrm{med}}) \leq 0,$$

so the y_{med} term is non-negative. Next, we show that $A_{\mathrm{P}}(y_\star;\theta_t) \geq (1-\rho)A_{\mathrm{P}}(y_\star;\theta_0)$. Observe that since $r_{\mathrm{P}}(y_\star) = r_{\mathrm{G}}(y_\star)$ and $V_{\mathrm{P}}(\theta_0) \geq -1$,

$$\frac{r_{\mathrm{G}}(y_\star) - \epsilon - V_{\mathrm{P}}(\theta_0)}{r_{\mathrm{P}}(y_\star) - V_{\mathrm{P}}(\theta_0)} = 1 - \frac{\epsilon}{r_{\mathrm{P}}(y_\star) - V_{\mathrm{P}}(\theta_0)} \leq 1 - \frac{\epsilon}{r_{\mathrm{P}}(y_\star) + 1} = \frac{r_{\mathrm{G}}(y_\star) + 1 - \epsilon}{r_{\mathrm{G}}(y_\star) + 1} = \rho.$$

Using this fact, we have that:

$$\frac{A_{\mathrm{P}}(y_\star;\theta_t)}{A_{\mathrm{P}}(y_\star;\theta_0)} = \frac{r_{\mathrm{P}}(y_\star) - V_{\mathrm{P}}(\theta_t)}{r_{\mathrm{P}}(y_\star) - V_{\mathrm{P}}(\theta_0)} = 1 - \frac{V_{\mathrm{P}}(\theta_t) - V_{\mathrm{P}}(\theta_0)}{r_{\mathrm{P}}(y_\star) - V_{\mathrm{P}}(\theta_0)} \geq 1 - \frac{r_{\mathrm{G}}(y_\star) - \epsilon - V_{\mathrm{P}}(\theta_0)}{r_{\mathrm{P}}(y_\star) - V_{\mathrm{P}}(\theta_0)} \geq 1 - \rho.$$

Therefore, along with $1 - \pi_{\theta_t}(y_\star) \geq 1 - \rho$ for $t \in [0, t_{\mathrm{P}}^{\mathrm{no-med}}]$ (as proven above), this leads to:

$$\frac{d}{dt}\pi_{\theta_t}(y_\star) \geq (1-\rho)^2 A_{\mathrm{P}}(y_\star;\theta_0)\|\phi(y_\star)\|^2\pi_{\theta_t}(y_\star)^2.$$

Dividing both sides of the inequality above by $\pi_{\theta_t}(y_\star)^2$, integrating from 0 to t, and rearranging terms results in the following lower bound:

$$\pi_{\theta_t}(y_\star) \geq \frac{\pi_{\theta_0}(y_\star)}{1 - (1-\rho)^2 A_{\mathrm{P}}(y_\star;\theta_0)\|\phi(y_\star)\|^2\pi_{\theta_0}(y_\star)\cdot t}.$$

At time $t = T$ for

$$T = \frac{1}{(1-\rho)^2 A_{\mathrm{P}}(y_\star;\theta_0)\|\phi(y_\star)\|^2}\left(\frac{1}{\pi_{\theta_0}(y_\star)} - \frac{1}{\rho}\right),$$

this lower bound on $\pi_{\theta_t}(y_\star)$ implies that $\pi_{\theta_t}(y_\star) \geq \rho$, in contradiction to our assumption that $t_{\mathrm{P}}^{\mathrm{no-med}} > T$. Thus,

$$t_\star^{\mathrm{no-med}} \leq t_{\mathrm{P}}^{\mathrm{no-med}} \leq T \leq \frac{\big(r_{\mathrm{G}}(y_\star) + 1\big)^2}{\epsilon^2\big(r_{\mathrm{P}}(y_\star) - V_{\mathrm{P}}(\theta_0)\big)\|\phi(y_\star)\|^2} \cdot \pi_{\theta_0}(y_\star)^{-1} = \mathcal{O}\big(\pi_{\theta_0}(y_\star)^{-1}\big).$$

□

D.6 Proof of Proposition 3

The core of the proof is to show that for all $t \geq 0$:

$$\frac{\pi_{\theta_t}(y_\star)}{\pi_{\theta_t}(y_{\mathrm{bad}})} < \zeta \leq \frac{\Delta_2}{\Delta_1}. \tag{38}$$

That is, the policy never assigns substantially higher probability to $y_\star$ relative to y_{bad}. We first establish that this bound implies the desired conclusion. Towards this, we start by showing that $V_{\text{G}}(\theta_0) \leq r_{\text{G}}(y_{\text{med}})$. Indeed:

$$\begin{aligned} r_{\text{G}}(y_{\text{med}}) - V_{\text{G}}(\theta_0) &= r_{\text{G}}(y_{\text{med}}) - \pi_{\theta_0}(y_\star) r_{\text{G}}(y_\star) - \pi_{\theta_0}(y_{\text{med}}) r_{\text{G}}(y_{\text{med}}) - \sum_{z \in \mathcal{Y}_{\text{bad}}} \pi_{\theta_0}(z) r_{\text{G}}(z) \\ &= r_{\text{G}}(y_{\text{med}}) - \pi_{\theta_0}(y_\star) r_{\text{G}}(y_\star) - \pi_{\theta_0}(y_{\text{med}}) r_{\text{G}}(y_{\text{med}}) - \pi_{\theta_0}(\mathcal{Y}_{\text{bad}}) r_{\text{G}}(y_{\text{bad}}) \\ &= (\pi_{\theta_0}(y_\star) + \pi_{\theta_0}(\mathcal{Y}_{\text{bad}})) r_{\text{G}}(y_{\text{med}}) - \pi_{\theta_0}(y_\star) r_{\text{G}}(y_\star) - \pi_{\theta_0}(\mathcal{Y}_{\text{bad}}) r_{\text{G}}(y_{\text{bad}}) \\ &= -\Delta_1 \pi_{\theta_0}(y_\star) + \Delta_2 \pi_{\theta_0}(\mathcal{Y}_{\text{bad}}) \\ &\geq 0, \end{aligned} \tag{39}$$

where the inequality is due to $\Delta_1 \pi_{\theta_0}(y_\star) \leq \Delta_2 \pi_{\theta_0}(\mathcal{Y}_{\text{bad}})$ (see the second assumption in the proposition statement). Next, we prove that $\zeta \leq \Delta_2 / \Delta_1$. Since $V_{\text{G}}(\theta_0) \leq r_{\text{G}}(y_{\text{med}})$ (Equation (39)), it holds that

$$\frac{V_{\text{G}}(\theta_0) - r_{\text{G}}(y_{\text{bad}})}{r_{\text{G}}(y_\star) - V_{\text{G}}(\theta_0)} \leq \frac{r_{\text{G}}(y_{\text{med}}) - r_{\text{G}}(y_{\text{bad}})}{r_{\text{G}}(y_\star) - r_{\text{G}}(y_{\text{med}})} = \frac{\Delta_2}{\Delta_1}. \tag{40}$$

Furthermore, because $\theta_0 = C(\phi(y_{\text{bad}}) - \phi(y_\star))$ for some $C > 0$ and $\langle \phi(y_{\text{bad}}), \phi(y_\star) \rangle = 0$, we have that $\langle \phi(y_\star), \theta_0 \rangle = -C \|\phi(y_\star)\|^2 \leq 0 \leq C \|\phi(y_{\text{bad}})\|^2 = \langle \phi(y_{\text{bad}}), \theta_0 \rangle$. That is, the logit of $y_\star$ is lower than that of y_{bad} at initialization, and so $\pi_{\theta_0}(y_\star) \leq \pi_{\theta_0}(y_{\text{bad}})$. Together with the first assumption in the proposition statement, which gives $\pi_{\theta_0}(y_{\text{med}}) > \pi_{\theta_0}(y_{\text{bad}})$, this implies

$$\frac{\ln \pi_{\theta_0}(y_{\text{med}}) - \ln \pi_{\theta_0}(y_{\text{bad}})}{\ln \pi_{\theta_0}(y_{\text{med}}) - \ln \pi_{\theta_0}(y_\star)} \leq 1. \tag{41}$$

Combining Equations (40) and (41) with the definition of ζ, we then obtain

$$\zeta = \frac{\ln \pi_{\theta_0}(y_{\text{med}}) - \ln \pi_{\theta_0}(y_{\text{bad}})}{\ln \pi_{\theta_0}(y_{\text{med}}) - \ln \pi_{\theta_0}(y_\star)} \cdot \frac{V_{\text{G}}(\theta_0) - r_{\text{G}}(y_{\text{bad}})}{r_{\text{G}}(y_\star) - V_{\text{G}}(\theta_0)} \leq \frac{\Delta_2}{\Delta_1}.$$

Now, assuming Equation (38) holds for all $t \geq 0$, we lower bound $r_{\text{G}}(y_\star) - V_{\text{G}}(\theta_t)$ as follows:

$$\begin{aligned} r_{\text{G}}(y_\star) - V_{\text{G}}(\theta_t) &= r_{\text{G}}(y_\star)(1 - \pi_{\theta_t}(y_\star)) - \pi_{\theta_t}(y_{\text{med}}) r_{\text{G}}(y_{\text{med}}) - \sum\nolimits_{z \in \mathcal{Y}_{\text{bad}}} \pi_{\theta_t}(z) r_{\text{G}}(z) \\ &\geq r_{\text{G}}(y_\star)(1 - \pi_{\theta_t}(y_\star)) - \pi_{\theta_t}(y_{\text{med}}) r_{\text{G}}(y_{\text{med}}) - \pi_{\theta_t}(y_{\text{bad}}) r_{\text{G}}(y_{\text{bad}}) \\ &\geq r_{\text{G}}(y_\star)(1 - \pi_{\theta_t}(y_\star)) - (1 - \pi_{\theta_t}(y_\star) - \pi_{\theta_t}(y_{\text{bad}})) r_{\text{G}}(y_{\text{med}}) \\ &\quad - \pi_{\theta_t}(y_{\text{bad}}) r_{\text{G}}(y_{\text{bad}}) \\ &= \Delta_1 (1 - \pi_{\theta_t}(y_\star)) + \Delta_2 \pi_{\theta_t}(y_{\text{bad}}) \\ &> \Delta_1 (1 - \pi_{\theta_t}(y_\star)) + \Delta_1 \pi_{\theta_t}(y_\star) \\ &= \Delta_1. \end{aligned}$$

Here, the first inequality is from $r_{\text{G}}(z) \leq 0$ for all $z \in \mathcal{Y}_{\text{bad}}$; the second inequality is from $\pi_{\theta_t}(y_{\text{med}}) \leq 1 - \pi_{\theta_t}(y_\star) - \pi_{\theta_t}(y_{\text{bad}})$; and the last inequality is from Equation (38). Thus, Equation (38) implies that for all $t \geq 0$,

$$V_{\text{G}}(\theta_t) < r_{\text{G}}(y_\star) - \Delta_1 = r_{\text{G}}(y_{\text{med}}).$$

It remains to prove Equation (38). Let $f_y(\theta) := \langle \phi(y), \theta \rangle$ denote the logit of $y \in \mathcal{Y}$ under $\theta \in \mathbb{R}^D$. At initialization,

$$\begin{aligned} \frac{\pi_{\theta_0}(y_\star)}{\pi_{\theta_0}(y_{\text{bad}})} &= \exp\left(f_{y_\star}(\theta_0) - f_{y_{\text{bad}}}(\theta_0) \right) \\ &= \exp\left(\langle \phi(y_\star) - \phi(y_{\text{bad}}), \theta_0 \rangle \right) \\ &= \exp\left(-C \|\phi(y_{\text{bad}}) - \phi(y_\star)\|^2 \right) \\ &< \exp(\ln(\zeta)) = \zeta. \end{aligned}$$

Both the third equality and the last inequality are due to the last assumption in the proposition statement, *i.e.*, due to $\theta_0 = C \cdot (\phi(y_{\text{bad}}) - \phi(y_\star))$ for some $C > \max\left\{0, -\ln(\zeta) / \|\phi(y_{\text{bad}}) - \phi(y_\star)\|^2\right\}$.

Now, assume by way of contradiction that there exists a time t' at which $\frac{\pi_{\theta_{t'}}(y_\star)}{\pi_{\theta_{t'}}(y_{\text{bad}})} \geq \zeta$ and let τ denote the initial such time, *i.e.*:

$$\tau := \min\left\{t \geq 0 : \frac{\pi_{\theta_t}(y_\star)}{\pi_{\theta_t}(y_{\text{bad}})} \geq \zeta\right\}.$$

Then, for all $t \in [0, \tau)$ it holds that $\pi_{\theta_t}(y_\star)/\pi_{\theta_t}(y_{\text{bad}}) < \zeta$. We now show that $\pi_{\theta_\tau}(y_\star)/\pi_{\theta_\tau}(y_{\text{bad}}) < \zeta$, in contradiction to the definition of τ as the initial time at which this ratio is at least ζ. We will do so by proving that $\pi_{\theta_t}(y_\star)/\pi_{\theta_t}(y_{\text{bad}})$ is decreasing for all $t \in [0, \tau)$, which, along with the fact that $\pi_{\theta_0}(y_\star)/\pi_{\theta_0}(y_{\text{bad}}) < \zeta$, implies that $\pi_{\theta_\tau}(y_\star)/\pi_{\theta_\tau}(y_{\text{bad}}) < \zeta$.

Notice that, since $\pi_{\theta_t}(y_\star)/\pi_{\theta_t}(y_{\text{bad}}) = \exp\left(f_{y_\star}(\theta_t) - f_{y_{\text{bad}}}(\theta_t)\right)$:

$$\text{sign}\left(\frac{d}{dt}\left(\frac{\pi_{\theta_t}(y_\star)}{\pi_{\theta_t}(y_{\text{bad}})}\right)\right) = \text{sign}\left(\frac{d}{dt}(f_{y_\star}(\theta_t) - f_{y_{\text{bad}}}(\theta_t))\right).$$

It therefore suffices to show that $\frac{d}{dt}(f_{y_\star}(\theta_t) - f_{y_{\text{bad}}}(\theta_t)) < 0$ for all $t \in [0, \tau)$.

The time derivative of $f_y(\theta_t)$ for an output $y \in \mathcal{Y}$ is given by:

$$\frac{d}{dt} f_y(\theta_t) = \left\langle \phi(y), \tfrac{d}{dt}\theta_t \right\rangle = \langle \phi(y), \nabla V_{\text{G}}(\theta_t) \rangle = \sum\nolimits_{z \in \mathcal{Y}} \pi_{\theta_t}(z) A_{\text{G}}(z; \theta_t) \langle \phi(y), \phi(z) \rangle ,$$

where the last equality is by Lemma 10. Thus, for all $t \in [0, \tau)$ we have that:

$$\begin{aligned}
&\frac{d}{dt}\left(f_{y_\star}(\theta_t) - f_{y_{\text{bad}}}(\theta_t)\right) \\
&= \sum\nolimits_{y \in \mathcal{Y}} \left(\langle \phi(y_\star), \phi(y) \rangle - \langle \phi(y_{\text{bad}}), \phi(y) \rangle\right) \pi_{\theta_t}(y) A_{\text{G}}(y; \theta_t) \\
&= \sum\nolimits_{y \in \mathcal{Y}} \langle \phi(y_\star) - \phi(y_{\text{bad}}), \phi(y) - \phi(y_{\text{med}}) + \phi(y_{\text{med}}) \rangle \, \pi_{\theta_t}(y) A_{\text{G}}(y; \theta_t) \\
&= \sum\nolimits_{y \in \mathcal{Y}} \langle \phi(y_\star) - \phi(y_{\text{bad}}), \phi(y) - \phi(y_{\text{med}}) \rangle \, \pi_{\theta_t}(y) A_{\text{G}}(y; \theta_t) \\
&\quad + \langle \phi(y_\star) - \phi(y_{\text{bad}}), \phi(y_{\text{med}}) \rangle \sum\nolimits_{y \in \mathcal{Y}} \pi_{\theta_t}(y) A_{\text{G}}(y; \theta_t).
\end{aligned}$$

Because

$$\sum\nolimits_{y \in \mathcal{Y}} \pi_{\theta_t}(y) A_{\text{G}}(y; \theta_t) = \sum\nolimits_{y \in \mathcal{Y}} \pi_{\theta_t}(y)(r_{\text{G}}(y) - V_{\text{G}}(\theta_t)) = V_{\text{G}}(\theta_t) - V_{\text{G}}(\theta_t) = 0,$$

we may write:

$$\begin{aligned}
&\frac{d}{dt}\left(f_{y_\star}(\theta_t) - f_{y_{\text{bad}}}(\theta_t)\right) \\
&= \sum\nolimits_{y \in \mathcal{Y}} \langle \phi(y_\star) - \phi(y_{\text{bad}}), \phi(y) - \phi(y_{\text{med}}) \rangle \, \pi_{\theta_t}(y) A_{\text{G}}(y; \theta_t) \\
&= \langle \phi(y_\star) - \phi(y_{\text{bad}}), \phi(y_\star) - \phi(y_{\text{med}}) \rangle \, \pi_{\theta_t}(y_\star) A_{\text{G}}(y_\star; \theta_t) \\
&\quad + \langle \phi(y_\star) - \phi(y_{\text{bad}}), \phi(y_{\text{bad}}) - \phi(y_{\text{med}}) \rangle \, \pi_{\theta_t}(y_{\text{bad}}) A_{\text{G}}(y_{\text{bad}}; \theta_t) \\
&\quad + \sum\nolimits_{z \in \mathcal{Y}_{\text{bad}} \setminus \{y_{\text{bad}}\}} \langle \phi(y_\star) - \phi(y_{\text{bad}}), \phi(z) - \phi(y_{\text{med}}) \rangle \, \pi_{\theta_t}(z) A_{\text{G}}(z; \theta_t).
\end{aligned} \tag{42}$$

Notice that the contribution of outputs in $\mathcal{Y}_{\text{bad}} \setminus \{y_{\text{bad}}\}$ is non-positive. Indeed, for every $z \in \mathcal{Y}_{\text{bad}} \setminus \{y_{\text{bad}}\}$, Assumption 4 implies $\langle \phi(y_\star) - \phi(y_{\text{bad}}), \phi(z) - \phi(y_{\text{med}}) \rangle = -\langle \phi(y_\star), \phi(y_{\text{med}}) \rangle > 0$, and since all outputs in $\mathcal{Y}_{\text{bad}}$ have the same minimal ground truth reward (Assumption 1), $A_{\text{G}}(z; \theta_t) < 0$ for all $z \in \mathcal{Y}_{\text{bad}}$. Using Equation (42), we therefore get

$$\begin{aligned}
&\frac{d}{dt}\left(f_{y_\star}(\theta_t) - f_{y_{\text{bad}}}(\theta_t)\right) \\
&\leq \langle \phi(y_\star) - \phi(y_{\text{bad}}), \phi(y_\star) - \phi(y_{\text{med}}) \rangle \, \pi_{\theta_t}(y_\star) A_{\text{G}}(y_\star; \theta_t) \\
&\quad + \langle \phi(y_\star) - \phi(y_{\text{bad}}), \phi(y_{\text{bad}}) - \phi(y_{\text{med}}) \rangle \, \pi_{\theta_t}(y_{\text{bad}}) A_{\text{G}}(y_{\text{bad}}; \theta_t).
\end{aligned}$$

Next, using $A_{\mathrm{G}}(y_\star;\theta_t) = r_{\mathrm{G}}(y_\star) - V_{\mathrm{G}}(\theta_t)$ and $A_{\mathrm{G}}(y_{\mathrm{bad}};\theta_t) = r_{\mathrm{G}}(y_{\mathrm{bad}}) - V_{\mathrm{G}}(\theta_t)$, the right-hand side of the equation above factors as

$$\begin{aligned}
&\frac{d}{dt}\big(f_{y_\star}(\theta_t) - f_{y_{\mathrm{bad}}}(\theta_t)\big) \\
&\leq -\langle \phi(y_\star) - \phi(y_{\mathrm{bad}}), \phi(y_{\mathrm{bad}}) - \phi(y_{\mathrm{med}})\rangle\, \pi_{\theta_t}(y_{\mathrm{bad}})(V_{\mathrm{G}}(\theta_t) - r_{\mathrm{G}}(y_{\mathrm{bad}})) \\
&\quad \cdot \left(- \frac{\langle \phi(y_\star) - \phi(y_{\mathrm{med}}), \phi(y_\star) - \phi(y_{\mathrm{bad}})\rangle}{\langle \phi(y_{\mathrm{bad}}) - \phi(y_{\mathrm{med}}), \phi(y_\star) - \phi(y_{\mathrm{bad}})\rangle} \cdot \frac{\pi_{\theta_t}(y_\star)}{\pi_{\theta_t}(y_{\mathrm{bad}})} \cdot \frac{r_{\mathrm{G}}(y_\star) - V_{\mathrm{G}}(\theta_t)}{V_{\mathrm{G}}(\theta_t) - r_{\mathrm{G}}(y_{\mathrm{bad}})} + 1 \right). \qquad (43)
\end{aligned}$$

It remains to show that the right-hand side of Equation (43) is negative. By the first assumption in the proposition statement, $\pi_{\theta_0}(y_{\mathrm{med}}) > \pi_{\theta_0}(y_{\mathrm{bad}})$. This implies that $f_{y_{\mathrm{med}}}(\theta_0) > f_{y_{\mathrm{bad}}}(\theta_0)$, or equivalently, $\langle \phi(y_{\mathrm{med}}) - \phi(y_{\mathrm{bad}}), \theta_0\rangle > 0$. Substituting $\theta_0 = C(\phi(y_{\mathrm{bad}}) - \phi(y_\star))$ and dividing by $C > 0$ gives

$$\langle \phi(y_{\mathrm{bad}}) - \phi(y_{\mathrm{med}}), \phi(y_\star) - \phi(y_{\mathrm{bad}})\rangle = \langle \phi(y_{\mathrm{bad}}) - \phi(y_\star), \phi(y_{\mathrm{med}}) - \phi(y_{\mathrm{bad}})\rangle > 0.$$

Furthermore,

$$\begin{aligned}
\langle \phi(y_\star) - \phi(y_{\mathrm{med}}), \phi(y_\star) - \phi(y_{\mathrm{bad}})\rangle &= \langle \phi(y_\star) - \phi(y_{\mathrm{bad}}), \phi(y_\star) - \phi(y_{\mathrm{bad}})\rangle \\
&\quad + \langle \phi(y_{\mathrm{bad}}) - \phi(y_{\mathrm{med}}), \phi(y_\star) - \phi(y_{\mathrm{bad}})\rangle \\
&\geq \langle \phi(y_{\mathrm{bad}}) - \phi(y_{\mathrm{med}}), \phi(y_\star) - \phi(y_{\mathrm{bad}})\rangle \\
&> 0.
\end{aligned}$$

These are exactly the numerator and denominator of the inner product terms appearing in

$$\zeta = \frac{\langle \phi(y_{\mathrm{bad}}) - \phi(y_{\mathrm{med}}), \phi(y_\star) - \phi(y_{\mathrm{bad}})\rangle}{\langle \phi(y_\star) - \phi(y_{\mathrm{med}}), \phi(y_\star) - \phi(y_{\mathrm{bad}})\rangle} \cdot \frac{V_{\mathrm{G}}(\theta_0) - r_{\mathrm{G}}(y_{\mathrm{bad}})}{r_{\mathrm{G}}(y_\star) - V_{\mathrm{G}}(\theta_0)},$$

and both are positive. Thus, for $t \in [0, \tau)$, the bound $\pi_{\theta_t}(y_\star)/\pi_{\theta_t}(y_{\mathrm{bad}}) < \zeta$ (Equation (38)) implies

$$\frac{\langle \phi(y_\star) - \phi(y_{\mathrm{med}}), \phi(y_\star) - \phi(y_{\mathrm{bad}})\rangle}{\langle \phi(y_{\mathrm{bad}}) - \phi(y_{\mathrm{med}}), \phi(y_\star) - \phi(y_{\mathrm{bad}})\rangle} \cdot \frac{\pi_{\theta_t}(y_\star)}{\pi_{\theta_t}(y_{\mathrm{bad}})} < \frac{V_{\mathrm{G}}(\theta_0) - r_{\mathrm{G}}(y_{\mathrm{bad}})}{r_{\mathrm{G}}(y_\star) - V_{\mathrm{G}}(\theta_0)}.$$

Since under gradient flow V_{G} is monotonically non-decreasing, we have $V_{\mathrm{G}}(\theta_0) \leq V_{\mathrm{G}}(\theta_t)$, and so

$$\frac{r_{\mathrm{G}}(y_\star) - V_{\mathrm{G}}(\theta_t)}{V_{\mathrm{G}}(\theta_t) - r_{\mathrm{G}}(y_{\mathrm{bad}})} \leq \frac{r_{\mathrm{G}}(y_\star) - V_{\mathrm{G}}(\theta_0)}{V_{\mathrm{G}}(\theta_0) - r_{\mathrm{G}}(y_{\mathrm{bad}})}.$$

Combining the last two inequalities shows that the product inside the parentheses in Equation (43) is less than 1. Since $\langle \phi(y_\star) - \phi(y_{\mathrm{bad}}), \phi(y_{\mathrm{bad}}) - \phi(y_{\mathrm{med}})\rangle > 0$, $\pi_{\theta_t}(y_{\mathrm{bad}}) > 0$, and $V_{\mathrm{G}}(\theta_t) > r_{\mathrm{G}}(y_{\mathrm{bad}})$, we conclude from Equation (43) that

$$\frac{d}{dt}\big(f_{y_\star}(\theta_t) - f_{y_{\mathrm{bad}}}(\theta_t)\big) < 0. \qquad (44)$$

Equation (44) implies that $f_{y_\star}(\theta_t) - f_{y_{\mathrm{bad}}}(\theta_t)$ is non-increasing, and so $\pi_{\theta_t}(y_\star)/\pi_{\theta_t}(y_{\mathrm{bad}})$ is non-increasing, for all $t \in [0, \tau)$. Thus,

$$\frac{\pi_{\theta_\tau}(y_\star)}{\pi_{\theta_\tau}(y_{\mathrm{bad}})} \leq \frac{\pi_{\theta_0}(y_\star)}{\pi_{\theta_0}(y_{\mathrm{bad}})} < \zeta,$$

in contradiction to our assumption that $\pi_{\theta_\tau}(y_\star)/\pi_{\theta_\tau}(y_{\mathrm{bad}}) \geq \zeta$. This establishes Equation (38), concluding the proof. □

D.7 Proof of Theorem 4

For convenience of notation, throughout the proof we denote $s := \langle \phi(y_\star), \phi(y_{\mathrm{med}})\rangle$.

D.7.1 Proof of Case I

Suppose that gradient flow is used to maximize the expected reward with respect to r_{G}. We begin by proving that $\pi_{\theta_t}(\mathcal{Y}_{\mathrm{bad}})$ is monotonically non-increasing (Lemma 8) and $\pi_{\theta_t}(y_\star)$ is monotonically non-decreasing (Lemma 9) throughout optimization. We then upper bound $t_\star$—the initial time at which $V_{\mathrm{G}}(\theta_t) \geq r_{\mathrm{G}}(y_\star) - \epsilon$—by lower bounding the growth rate of $\pi_{\theta_t}(y_\star)$ over the interval $[0, t_\star)$.

Lemma 8. *At any time $t \geq 0$, it holds that $\frac{d}{dt}\pi_{\theta_t}(\mathcal{Y}_{\mathrm{bad}}) \leq 0$.*

Proof. Suppose first that $V_{\mathrm{G}}(\theta_t) \leq r_{\mathrm{G}}(y_{\mathrm{med}})$. Under Assumption 5, for every $z \in \mathcal{Y}_{\mathrm{bad}}$, the feature vector $\phi(z)$ is orthogonal to the feature vectors of all other outputs. Therefore, by Lemma 13 we have for all $z \in \mathcal{Y}_{\mathrm{bad}}$:

$$\begin{aligned}\frac{d}{dt}\pi_{\theta_t}(z) = \pi_{\theta_t}(z)^2 A_{\mathrm{G}}(z;\theta_t)\|\phi(z)\|^2 \\ - \pi_{\theta_t}(z)\Big(\sum\nolimits_{z'\in\mathcal{Y}_{\mathrm{bad}}} \pi_{\theta_t}(z')^2 A_{\mathrm{G}}(z';\theta_t)\|\phi(z')\|^2 \\ + \pi_{\theta_t}(y_{\mathrm{med}})A_{\mathrm{G}}(y_{\mathrm{med}};\theta_t)\big(\|\phi(y_{\mathrm{med}})\|^2\pi_{\theta_t}(y_{\mathrm{med}}) + s\pi_{\theta_t}(y_\star)\big) \\ + \pi_{\theta_t}(y_\star)A_{\mathrm{G}}(y_\star;\theta_t)\big(\|\phi(y_\star)\|^2\pi_{\theta_t}(y_\star) + s\pi_{\theta_t}(y_{\mathrm{med}})\big)\Big).\end{aligned}$$

Summing over all $z \in \mathcal{Y}_{\mathrm{bad}}$, we have:

$$\begin{aligned}\frac{d}{dt}\pi_{\theta_t}(\mathcal{Y}_{\mathrm{bad}}) = \underbrace{\Big(\sum\nolimits_{z\in\mathcal{Y}_{\mathrm{bad}}} \pi_{\theta_t}(z)^2 A_{\mathrm{G}}(z;\theta_t)\|\phi(z)\|^2\Big)\big(1-\pi_{\theta_t}(\mathcal{Y}_{\mathrm{bad}})\big)}_{I_1} \\ - \pi_{\theta_t}(\mathcal{Y}_{\mathrm{bad}})\Big(\pi_{\theta_t}(y_{\mathrm{med}})A_{\mathrm{G}}(y_{\mathrm{med}};\theta_t)\big(\|\phi(y_{\mathrm{med}})\|^2\pi_{\theta_t}(y_{\mathrm{med}}) + s\pi_{\theta_t}(y_\star)\big) \\ + \pi_{\theta_t}(y_\star)A_{\mathrm{G}}(y_\star;\theta_t)\big(\|\phi(y_\star)\|^2\pi_{\theta_t}(y_\star) + s\pi_{\theta_t}(y_{\mathrm{med}})\big)\Big),\end{aligned}$$

where we denote by I_2 the second term on the right-hand side of the equation above (excluding the minus sign). Since all outputs in $\mathcal{Y}_{\mathrm{bad}}$ have the minimal ground truth reward, we have $A_{\mathrm{G}}(z;\theta_t) \leq 0$ for all $z \in \mathcal{Y}_{\mathrm{bad}}$. Together with $A_{\mathrm{G}}(y_{\mathrm{med}};\theta_t) \geq 0$ and $A_{\mathrm{G}}(y_\star;\theta_t) \geq 0$, this gives $I_1 \leq 0$ and $I_2 \geq 0$. This implies that $\frac{d}{dt}\pi_{\theta_t}(\mathcal{Y}_{\mathrm{bad}}) \leq 0$ when $V_{\mathrm{G}}(\theta_t) \leq r_{\mathrm{G}}(y_{\mathrm{med}})$.

Now, consider the case where $V_{\mathrm{G}}(\theta_t) > r_{\mathrm{G}}(y_{\mathrm{med}})$. In this case, $I_1 \leq 0$ still holds. Furthermore,

$$\begin{aligned}V_{\mathrm{G}}(\theta_t) - r_{\mathrm{G}}(y_{\mathrm{med}}) &= (\pi_{\theta_t}(y_{\mathrm{med}}) - 1)r_{\mathrm{G}}(y_{\mathrm{med}}) + \pi_{\theta_t}(y_\star)r_{\mathrm{G}}(y_\star) + \sum\nolimits_{z\in\mathcal{Y}_{\mathrm{bad}}} \pi_{\theta_t}(z)r_{\mathrm{G}}(z) \\ &\leq (\pi_{\theta_t}(y_{\mathrm{med}}) - 1)r_{\mathrm{G}}(y_{\mathrm{med}}) + \pi_{\theta_t}(y_\star)r_{\mathrm{G}}(y_\star) \\ &\leq -\pi_{\theta_t}(y_\star)r_{\mathrm{G}}(y_{\mathrm{med}}) + \pi_{\theta_t}(y_\star)r_{\mathrm{G}}(y_\star) \\ &= \pi_{\theta_t}(y_\star)\Delta_1,\end{aligned} \tag{45}$$

where the first inequality is from $r_{\mathrm{G}}(z) \leq 0$ for all $z \in \mathcal{Y}_{\mathrm{bad}}$ and the second inequality is because $1 - \pi_{\theta_t}(y_{\mathrm{med}}) \geq \pi_{\theta_t}(y_\star)$ and $r_{\mathrm{G}}(y_{\mathrm{med}}) > 0$. Thus, $|A_{\mathrm{G}}(y_{\mathrm{med}};\theta_t)| = V_{\mathrm{G}}(\theta_t) - r_{\mathrm{G}}(y_{\mathrm{med}}) \leq \pi_{\theta_t}(y_\star)\Delta_1$. Moreover,

$$\begin{aligned}A_{\mathrm{G}}(y_\star;\theta_t) &= r_{\mathrm{G}}(y_\star) - \pi_{\theta_t}(y_\star)r_{\mathrm{G}}(y_\star) - \pi_{\theta_t}(y_{\mathrm{med}})r_{\mathrm{G}}(y_{\mathrm{med}}) - \sum\nolimits_{z\in\mathcal{Y}_{\mathrm{bad}}} \pi_{\theta_t}(z)r_{\mathrm{G}}(z) \\ &\geq r_{\mathrm{G}}(y_\star) - \pi_{\theta_t}(y_\star)r_{\mathrm{G}}(y_\star) - (1-\pi_{\theta_t}(y_\star))r_{\mathrm{G}}(y_{\mathrm{med}}) \\ &= (1-\pi_{\theta_t}(y_\star))\Delta_1.\end{aligned} \tag{46}$$

We can therefore lower bound $I_2/\pi_{\theta_t}(\mathcal{Y}_{\mathrm{bad}})$ as:

$$\begin{aligned}&\pi_{\theta_t}(y_{\mathrm{med}})A_{\mathrm{G}}(y_{\mathrm{med}};\theta_t)\big(\|\phi(y_{\mathrm{med}})\|^2\pi_{\theta_t}(y_{\mathrm{med}}) + s\pi_{\theta_t}(y_\star)\big) \\ &\quad + \pi_{\theta_t}(y_\star)A_{\mathrm{G}}(y_\star;\theta_t)\big(\|\phi(y_\star)\|^2\pi_{\theta_t}(y_\star) + s\pi_{\theta_t}(y_{\mathrm{med}})\big) \\ &\geq \pi_{\theta_t}(y_\star)(1-\pi_{\theta_t}(y_\star))\Delta_1\big(\|\phi(y_\star)\|^2\pi_{\theta_t}(y_\star) + s\pi_{\theta_t}(y_{\mathrm{med}})\big) \\ &\quad - \pi_{\theta_t}(y_{\mathrm{med}})\pi_{\theta_t}(y_\star)\Delta_1\big(\|\phi(y_{\mathrm{med}})\|^2\pi_{\theta_t}(y_{\mathrm{med}}) + s\pi_{\theta_t}(y_\star)\big) \\ &\geq \pi_{\theta_t}(y_\star)(1-\pi_{\theta_t}(y_\star))\Delta_1\big((\|\phi(y_\star)\|^2 - s)\pi_{\theta_t}(y_\star) + (s - \|\phi(y_{\mathrm{med}})\|^2)\pi_{\theta_t}(y_{\mathrm{med}})\big) \\ &\geq 0.\end{aligned}$$

Here, the first inequality is from $|A_{\mathrm{G}}(y_{\mathrm{med}};\theta_t)| \leq \pi_{\theta_t}(y_\star)\Delta_1$ and $A_{\mathrm{G}}(y_\star;\theta_t) \geq (1-\pi_{\theta_t}(y_\star))\Delta_1$, along with the fact that the multipliers of these advantages are non-negative. The second inequality is from $1 - \pi_{\theta_t}(y_\star) \geq \pi_{\theta_t}(y_{\mathrm{med}})$, combined with $s > 0$ and $\|\phi(y_{\mathrm{med}})\|^2 \geq 0$. The last inequality

follows from the assumption in Theorem 4 that $\|\phi(y_{\text{med}})\|^2 < s < \|\phi(y_\star)\|^2$. Thus, $I_2 \geq 0$, which means $\frac{d}{dt}\pi_{\theta_t}(\mathcal{Y}_{\text{bad}}) \leq 0$.

Overall, we showed that $\frac{d}{dt}\pi_{\theta_t}(\mathcal{Y}_{\text{bad}}) \leq 0$ both when $V_{\text{G}}(\theta_t) \leq r_{\text{G}}(y_{\text{med}})$ and when $V_{\text{G}}(\theta_t) > r_{\text{G}}(y_{\text{med}})$, and so $\pi_{\theta_t}(\mathcal{Y}_{\text{bad}})$ is non-increasing for all $t \geq 0$. □

Lemma 9. *At any time $t \geq 0$, it holds that $\frac{d}{dt}\pi_{\theta_t}(y_\star) \geq 0$.*

Proof. Let $\ell(\theta) := \langle \phi(y_\star), \theta\rangle - \langle \phi(y_{\text{med}}), \theta\rangle = \langle \phi(y_\star) - \phi(y_{\text{med}}), \theta\rangle$ be the logit difference between $y_\star$ and y_{med} under $\theta \in \mathbb{R}^D$. We prove that $\ell(\theta_t)$ is non-decreasing for all $t \geq 0$. We can write the time derivative of $\ell(\theta_t)$ as:

$$\begin{aligned}\frac{d}{dt}\ell(\theta_t) &= \left\langle \phi(y_\star) - \phi(y_{\text{med}}), \tfrac{d}{dt}\theta_t \right\rangle \\ &= \|\phi(y_\star)\|^2 \pi_{\theta_t}(y_\star) A_{\text{G}}(y_\star;\theta_t) + s\pi_{\theta_t}(y_{\text{med}}) A_{\text{G}}(y_{\text{med}};\theta_t) - s\pi_{\theta_t}(y_\star) A_{\text{G}}(y_\star;\theta_t) \\ &\quad - \|\phi(y_{\text{med}})\|^2 \pi_{\theta_t}(y_{\text{med}}) A_{\text{G}}(y_{\text{med}};\theta_t) \\ &= (\|\phi(y_\star)\|^2 - s)\pi_{\theta_t}(y_\star) A_{\text{G}}(y_\star;\theta_t) + (s - \|\phi(y_{\text{med}})\|^2)\pi_{\theta_t}(y_{\text{med}}) A_{\text{G}}(y_{\text{med}};\theta_t).\end{aligned}$$

If $V_{\text{G}}(\theta_t) \leq r_{\text{G}}(y_{\text{med}})$ (*i.e.*, $A_{\text{G}}(y_{\text{med}};\theta_t) \geq 0$), since $\|\phi(y_\star)\|^2 > s > \|\phi(y_{\text{med}})\|^2$ and $A_{\text{G}}(y_\star;\theta_t) \geq 0$, it follows directly from the expression above that $\frac{d}{dt}\ell(\theta_t) \geq 0$. On the other hand, if $V_{\text{G}}(\theta_t) > r_{\text{G}}(y_{\text{med}})$, Equations (45) and (46) hold, as established in the proof of Lemma 8. Thus, in this case it also holds that $\frac{d}{dt}\ell(\theta_t) \geq 0$ since:

$$\begin{aligned}\frac{d}{dt}\ell(\theta_t) &\geq (\|\phi(y_\star)\|^2 - s)\pi_{\theta_t}(y_\star)(1 - \pi_{\theta_t}(y_\star))\Delta_1 - (s - \|\phi(y_{\text{med}})\|^2)\pi_{\theta_t}(y_{\text{med}})\pi_{\theta_t}(y_\star)\Delta_1 \\ &\geq \pi_{\theta_t}(y_\star)(1 - \pi_{\theta_t}(y_\star))\Delta_1(\|\phi(y_\star)\|^2 - 2s + \|\phi(y_{\text{med}})\|^2) \\ &\geq 0,\end{aligned}$$

where the second inequality is from $\pi_{\theta_t}(y_{\text{med}}) \leq 1 - \pi_{\theta_t}(y_\star)$, and the last inequality is from $s \leq \|\phi(y_\star)\|\|\phi(y_{\text{med}})\| \leq (\|\phi(y_\star)\|^2 + \|\phi(y_{\text{med}})\|^2)/2$. Thus, $\ell(\theta_t)$ is monotonically non-decreasing for all $t \geq 0$.

Now, notice that

$$\pi_{\theta_t}(y_\star) = (1 - \pi_{\theta_t}(\mathcal{Y}_{\text{bad}})) \cdot \frac{\pi_{\theta_t}(y_\star)}{\pi_{\theta_t}(y_\star) + \pi_{\theta_t}(y_{\text{med}})}.$$

The factor $1 - \pi_{\theta_t}(\mathcal{Y}_{\text{bad}})$ is monotonically non-decreasing by Lemma 8, since $\pi_{\theta_t}(\mathcal{Y}_{\text{bad}})$ is non-increasing. Furthermore, the factor $\pi_{\theta_t}(y_\star)/(\pi_{\theta_t}(y_\star) + \pi_{\theta_t}(y_{\text{med}}))$ is also non-decreasing. To see this, let $\beta_t := \pi_{\theta_t}(y_\star)/\pi_{\theta_t}(y_{\text{med}})$. By the softmax parametrization of the policy, $\ell(\theta_t) = \ln \beta_t$, and so $\beta_t = \exp(\ell(\theta_t))$. Since both $\ell(\theta_t)$ and the exponential function are non-decreasing, β_t is non-decreasing. Furthermore,

$$\frac{\pi_{\theta_t}(y_\star)}{\pi_{\theta_t}(y_\star) + \pi_{\theta_t}(y_{\text{med}})} = \frac{\beta_t}{1 + \beta_t},$$

and the map $\beta \mapsto \beta/(1+\beta)$ is non-decreasing for $\beta \geq 0$. Thus, $\pi_{\theta_t}(y_\star)/(\pi_{\theta_t}(y_\star) + \pi_{\theta_t}(y_{\text{med}}))$ is non-decreasing in t. Overall, the above implies that $\pi_{\theta_t}(y_\star)$ is non-decreasing, and so $\frac{d}{dt}\pi_{\theta_t}(y_\star) \geq 0$, for all $t \geq 0$. □

We now complete the proof of Case I.

Proof of Theorem 4, Case I. Denote

$$\rho := \frac{r_{\text{G}}(y_{\text{med}}) + \Delta_1 - \epsilon + 1}{r_{\text{G}}(y_\star) + 1},$$

and notice that when $\pi_{\theta_t}(y_\star) \geq \rho$ it must be that $V_{\text{G}}(\theta_t) \geq r_{\text{G}}(y_{\text{med}}) + \Delta_1 - \epsilon = r_{\text{G}}(y_\star) - \epsilon$ since:

$$\begin{aligned}V_{\text{G}}(\theta_t) &\geq \pi_{\theta_t}(y_\star) r_{\text{G}}(y_\star) - (1 - \pi_{\theta_t}(y_\star)) \\ &= \pi_{\theta_t}(y_\star)(r_{\text{G}}(y_\star) + 1) - 1 \\ &\geq r_{\text{G}}(y_\star) - \epsilon,\end{aligned}$$

where the first inequality is from $r_{\mathrm{G}}(y) \geq -1$ for all $y \in \mathcal{Y}$ and the second inequality follows from the definition of ρ. This implies that $\pi_{\theta_t}(y_\star) < \rho$ for all $t \in [0, t_\star)$. In particular, Lemma 15 implies that $V_{\mathrm{G}}(\theta_0) < r_{\mathrm{G}}(y_{\mathrm{med}}) < r_{\mathrm{G}}(y_\star) - \epsilon$, and so $\pi_{\theta_0}(y_\star) < \rho$ and $t_\star > 0$.

Next, we lower bound the time derivative of $\pi_{\theta_t}(y_\star)$ for $t \in [0, t_\star)$, based on the expression derived in Lemma 13:

$$\begin{aligned}
\frac{d}{dt}\pi_{\theta_t}(y_\star) &= \pi_{\theta_t}(y_\star)\Big[\pi_{\theta_t}(y_\star)A_{\mathrm{G}}(y_\star;\theta_t)\big(\|\phi(y_\star)\|^2 - s\pi_{\theta_t}(y_{\mathrm{med}}) - \|\phi(y_\star)\|^2\pi_{\theta_t}(y_\star)\big) \\
&\qquad + \pi_{\theta_t}(y_{\mathrm{med}})A_{\mathrm{G}}(y_{\mathrm{med}};\theta_t)\big(s - \|\phi(y_{\mathrm{med}})\|^2\pi_{\theta_t}(y_{\mathrm{med}}) - s\pi_{\theta_t}(y_\star)\big) \\
&\qquad - \sum\nolimits_{z\in\mathcal{Y}_{\mathrm{bad}}} \pi_{\theta_t}(z)^2 A_{\mathrm{G}}(z;\theta_t)\|\phi(z)\|^2\Big] \\
&\geq \pi_{\theta_t}(y_\star)\Big[\pi_{\theta_t}(y_\star)A_{\mathrm{G}}(y_\star;\theta_t)\big(\|\phi(y_\star)\|^2 - s\pi_{\theta_t}(y_{\mathrm{med}}) - \|\phi(y_\star)\|^2\pi_{\theta_t}(y_\star)\big) \\
&\qquad + \pi_{\theta_t}(y_{\mathrm{med}})A_{\mathrm{G}}(y_{\mathrm{med}};\theta_t)\big(s - \|\phi(y_{\mathrm{med}})\|^2\pi_{\theta_t}(y_{\mathrm{med}}) - s\pi_{\theta_t}(y_\star)\big)\Big] \\
&= \pi_{\theta_t}(y_\star)\Big[\pi_{\theta_t}(y_\star)A_{\mathrm{G}}(y_\star;\theta_t)\big(\|\phi(y_\star)\|^2(1-\pi_{\theta_t}(y_\star)) - s\pi_{\theta_t}(y_{\mathrm{med}})\big) \\
&\qquad + \pi_{\theta_t}(y_{\mathrm{med}})A_{\mathrm{G}}(y_{\mathrm{med}};\theta_t)\big(s(1-\pi_{\theta_t}(y_\star)) - \|\phi(y_{\mathrm{med}})\|^2\pi_{\theta_t}(y_{\mathrm{med}})\big)\Big] \\
&\geq \pi_{\theta_t}(y_\star)^2\Delta_1\Big[\|\phi(y_\star)\|^2(1-\pi_{\theta_t}(y_\star))^2 - 2s\pi_{\theta_t}(y_{\mathrm{med}})(1-\pi_{\theta_t}(y_\star)) \\
&\qquad + \|\phi(y_{\mathrm{med}})\|^2\pi_{\theta_t}(y_{\mathrm{med}})^2\Big].
\end{aligned} \tag{47}$$

Here, the first inequality is from $A_{\mathrm{G}}(z;\theta_t) \leq 0$ for all $z \in \mathcal{Y}_{\mathrm{bad}}$. The second inequality is from $A_{\mathrm{G}}(y_\star;\theta_t) \geq (1-\pi_{\theta_t}(y_\star))\Delta_1$ (as in Equation (46)), $A_{\mathrm{G}}(y_{\mathrm{med}};\theta_t) \geq -\pi_{\theta_t}(y_\star)\Delta_1$ (which follows from Equation (45) when $V_{\mathrm{G}}(\theta_t) > r_{\mathrm{G}}(y_{\mathrm{med}})$ and is trivial otherwise), and the fact that $\|\phi(y_\star)\|^2(1-\pi_{\theta_t}(y_\star)) - s\pi_{\theta_t}(y_{\mathrm{med}})$ and $s(1-\pi_{\theta_t}(y_\star)) - \|\phi(y_{\mathrm{med}})\|^2\pi_{\theta_t}(y_{\mathrm{med}})$ are non-negative, since $\pi_{\theta_t}(y_{\mathrm{med}}) \leq 1-\pi_{\theta_t}(y_\star)$ and $\|\phi(y_{\mathrm{med}})\|^2 < s < \|\phi(y_\star)\|^2$. To simplify the bracketed term on the right-hand side of Equation (47), we can view the expression

$$\|\phi(y_\star)\|^2(1-\pi_{\theta_t}(y_\star))^2 - 2s\pi_{\theta_t}(y_{\mathrm{med}})(1-\pi_{\theta_t}(y_\star)) + \|\phi(y_{\mathrm{med}})\|^2\pi_{\theta_t}(y_{\mathrm{med}})^2$$

as a function of $\pi_{\theta_t}(y_{\mathrm{med}})$ over the interval $[0, 1-\pi_{\theta_t}(y_\star)]$. Its derivative with respect to $\pi_{\theta_t}(y_{\mathrm{med}})$ is $-2s(1-\pi_{\theta_t}(y_\star)) + 2\|\phi(y_{\mathrm{med}})\|^2\pi_{\theta_t}(y_{\mathrm{med}})$, which is at most $-2(s - \|\phi(y_{\mathrm{med}})\|^2)(1-\pi_{\theta_t}(y_\star)) \leq 0$, since $\pi_{\theta_t}(y_{\mathrm{med}}) \leq 1-\pi_{\theta_t}(y_\star)$ and $s > \|\phi(y_{\mathrm{med}})\|^2$. Thus, the expression is minimized at $\pi_{\theta_t}(y_{\mathrm{med}}) = 1-\pi_{\theta_t}(y_\star)$. This implies that:

$$\begin{aligned}
\frac{d}{dt}\pi_{\theta_t}(y_\star) &\geq \pi_{\theta_t}(y_\star)^2(1-\pi_{\theta_t}(y_\star))^2\Delta_1(\|\phi(y_\star)\|^2 + \|\phi(y_{\mathrm{med}})\|^2 - 2s) \\
&\geq (1-\rho)^2\Delta_1(\|\phi(y_\star)\|^2 + \|\phi(y_{\mathrm{med}})\|^2 - 2s)\pi_{\theta_t}(y_\star)^2 \\
&= (1-\rho)^2\Delta_1\|\phi(y_\star) - \phi(y_{\mathrm{med}})\|^2\pi_{\theta_t}(y_\star)^2,
\end{aligned} \tag{48}$$

where the second inequality is from $\pi_{\theta_t}(y_\star) < \rho$ for all $t \in [0, t_\star)$.

Dividing both sides of Equation (48) by $\pi_{\theta_t}(y_\star)^2$, integrating from 0 to t, and rearranging terms yields:

$$\pi_{\theta_t}(y_\star) \geq \frac{\pi_{\theta_0}(y_\star)}{1 - (1-\rho)^2\Delta_1\|\phi(y_\star) - \phi(y_{\mathrm{med}})\|^2\pi_{\theta_0}(y_\star)\cdot t}.$$

Since $\pi_{\theta_t}(y_\star) < \rho$ for all $t \in [0, t_\star)$, we get that for all such t:

$$t \leq \frac{1}{(1-\rho)^2\|\phi(y_\star) - \phi(y_{\mathrm{med}})\|^2\Delta_1}\left(\frac{1}{\pi_{\theta_0}(y_\star)} - \frac{1}{\rho}\right).$$

Since this holds for all $t \in [0, t_\star)$, we get

$$t_\star \leq \frac{1}{(1-\rho)^2\|\phi(y_\star) - \phi(y_{\mathrm{med}})\|^2\Delta_1}\cdot\frac{1}{\pi_{\theta_0}(y_\star)}.$$

Plugging in the definition of ρ gives

$$t_\star \leq \frac{(r_{\mathrm{G}}(y_\star)+1)^2}{\epsilon^2\|\phi(y_\star) - \phi(y_{\mathrm{med}})\|^2\Delta_1}\cdot\frac{1}{\pi_{\theta_0}(y_\star)},$$

which completes the proof of Case I. □

D.7.2 Proof of Case II

Suppose that gradient flow is used to maximize the expected reward with respect to r_{P}, and that $\langle \phi(y_\star), \phi(y_{\mathrm{med}}) \rangle$ is sufficiently high, in the sense that

$$\langle \phi(y_\star), \phi(y_{\mathrm{med}}) \rangle > \frac{\pi_{\theta_0}(y_\star) A_{\mathrm{P}}(y_\star; \theta_0) \|\phi(y_\star)\|^2 - \pi_{\theta_0}(y_{\mathrm{med}}) A_{\mathrm{P}}(y_{\mathrm{med}}; \theta_0) \|\phi(y_{\mathrm{med}})\|^2}{\pi_{\theta_0}(y_\star) A_{\mathrm{P}}(y_\star; \theta_0) - \pi_{\theta_0}(y_{\mathrm{med}}) A_{\mathrm{P}}(y_{\mathrm{med}}; \theta_0)}. \tag{49}$$

Proof of Theorem 4, Case II. We prove that, due to the high inner product between $\phi(y_\star)$ and $\phi(y_{\mathrm{med}})$, the logit difference $\langle \phi(y_\star), \theta_t \rangle - \langle \phi(y_{\mathrm{med}}), \theta_t \rangle$ is monotonically non-increasing with respect to t. Consequently, the ratio $\pi_{\theta_t}(y_\star)/\pi_{\theta_t}(y_{\mathrm{med}})$ is bounded by its initial value for all $t \geq 0$, and the policy never assigns $y_\star$ sufficiently high probability to reach an expected ground truth reward above the constant $r_{\mathrm{G}}(y_\star) - (r_{\mathrm{G}}(y_\star) - r_{\mathrm{G}}(y_{\mathrm{med}}))\pi_{\theta_0}(y_{\mathrm{med}})$.

We first note that $A_{\mathrm{P}}(y_{\mathrm{med}}; \theta_0) < 0$ because y_{med} has a minimal proxy reward of $r_{\mathrm{P}}(y_{\mathrm{med}}) \leq \min_{y \in \mathcal{Y}_{\mathrm{bad}}} r_{\mathrm{G}}(y)$, while $y_\star$ has strictly larger proxy reward and π_{θ_0} has full support as a softmax policy. Moreover, $A_{\mathrm{P}}(y_\star; \theta_0) > 0$ since $y_\star$ has the unique maximal proxy reward.

Now, let $\ell(\theta) := \langle \phi(y_\star), \theta \rangle - \langle \phi(y_{\mathrm{med}}), \theta \rangle = \langle \phi(y_\star) - \phi(y_{\mathrm{med}}), \theta \rangle$ be the logit difference between $y_\star$ and y_{med} under $\theta \in \mathbb{R}^D$. We prove that $\ell(\theta_t)$ is non-increasing for all $t \geq 0$. We can write the time derivative of $\ell(\theta_t)$ as:

$$\begin{aligned}
\frac{d}{dt}\ell(\theta_t) &= \left\langle \phi(y_\star) - \phi(y_{\mathrm{med}}), \tfrac{d}{dt}\theta_t \right\rangle \\
&= \|\phi(y_\star)\|^2 \pi_{\theta_t}(y_\star) A_{\mathrm{P}}(y_\star; \theta_t) + s\pi_{\theta_t}(y_{\mathrm{med}}) A_{\mathrm{P}}(y_{\mathrm{med}}; \theta_t) - s\pi_{\theta_t}(y_\star) A_{\mathrm{P}}(y_\star; \theta_t) \\
&\quad - \|\phi(y_{\mathrm{med}})\|^2 \pi_{\theta_t}(y_{\mathrm{med}}) A_{\mathrm{P}}(y_{\mathrm{med}}; \theta_t) \\
&= (\|\phi(y_\star)\|^2 - s)\pi_{\theta_t}(y_\star) A_{\mathrm{P}}(y_\star; \theta_t) + (s - \|\phi(y_{\mathrm{med}})\|^2)\pi_{\theta_t}(y_{\mathrm{med}}) A_{\mathrm{P}}(y_{\mathrm{med}}; \theta_t) \\
&\leq (\|\phi(y_\star)\|^2 - s)\pi_{\theta_t}(y_\star) A_{\mathrm{P}}(y_\star; \theta_0) + (s - \|\phi(y_{\mathrm{med}})\|^2)\pi_{\theta_t}(y_{\mathrm{med}}) A_{\mathrm{P}}(y_{\mathrm{med}}; \theta_0) \\
&= \pi_{\theta_t}(y_{\mathrm{med}}) \big((\|\phi(y_\star)\|^2 - s) A_{\mathrm{P}}(y_\star; \theta_0) \exp(\ell(\theta_t)) + (s - \|\phi(y_{\mathrm{med}})\|^2) A_{\mathrm{P}}(y_{\mathrm{med}}; \theta_0) \big).
\end{aligned} \tag{50}$$

Here, the first inequality is from $V_{\mathrm{P}}(\theta_t)$ being monotonically non-decreasing when maximizing it via gradient flow, together with the positivity of $\|\phi(y_\star)\|^2 - s$ and $s - \|\phi(y_{\mathrm{med}})\|^2$, and the last equality is by the definition of $\ell(\theta_t)$ since $\pi_{\theta_t}(y_\star) = \pi_{\theta_t}(y_{\mathrm{med}}) \cdot \exp(\ell(\theta_t))$. Since $A_{\mathrm{P}}(y_\star; \theta_0) > 0 > A_{\mathrm{P}}(y_{\mathrm{med}}; \theta_0)$ and $\|\phi(y_{\mathrm{med}})\|^2 < s < \|\phi(y_\star)\|^2$, the denominator and numerator of the right-hand side below are both positive. Therefore, when

$$\exp(\ell(\theta_t)) < \frac{-(s - \|\phi(y_{\mathrm{med}})\|^2) A_{\mathrm{P}}(y_{\mathrm{med}}; \theta_0)}{(\|\phi(y_\star)\|^2 - s) A_{\mathrm{P}}(y_\star; \theta_0)}, \tag{51}$$

we get that $\frac{d}{dt}\ell(\theta_t) < 0$.

Equation (49) implies Equation (51) at $t = 0$. Indeed, the denominator on the right-hand side of Equation (49) is positive because $A_{\mathrm{P}}(y_\star; \theta_0) > 0 > A_{\mathrm{P}}(y_{\mathrm{med}}; \theta_0)$. Therefore, multiplying both sides of Equation (49) by this denominator and rearranging gives

$$\pi_{\theta_0}(y_\star) A_{\mathrm{P}}(y_\star; \theta_0) \big(\|\phi(y_\star)\|^2 - s \big) < -\pi_{\theta_0}(y_{\mathrm{med}}) A_{\mathrm{P}}(y_{\mathrm{med}}; \theta_0) \big(s - \|\phi(y_{\mathrm{med}})\|^2 \big).$$

Since $\pi_{\theta_0}(y_{\mathrm{med}}) > 0$, $A_{\mathrm{P}}(y_\star; \theta_0) > 0$, and $\|\phi(y_\star)\|^2 - s > 0$, we may divide both sides by

$$\pi_{\theta_0}(y_{\mathrm{med}}) A_{\mathrm{P}}(y_\star; \theta_0) \big(\|\phi(y_\star)\|^2 - s \big)$$

to obtain

$$\frac{\pi_{\theta_0}(y_\star)}{\pi_{\theta_0}(y_{\mathrm{med}})} < \frac{-(s - \|\phi(y_{\mathrm{med}})\|^2) A_{\mathrm{P}}(y_{\mathrm{med}}; \theta_0)}{(\|\phi(y_\star)\|^2 - s) A_{\mathrm{P}}(y_\star; \theta_0)}.$$

Finally, $\pi_{\theta_0}(y_\star)/\pi_{\theta_0}(y_{\mathrm{med}}) = \exp(\ell(\theta_0))$, so Equation (51) holds at $t = 0$. We prove that Equation (51) holds for all $t \geq 0$ by contradiction. Assume that there exists a time $t' > 0$ at which

$$\exp(\ell(\theta_{t'})) \geq \frac{-(s - \|\phi(y_{\mathrm{med}})\|^2) A_{\mathrm{P}}(y_{\mathrm{med}}; \theta_0)}{(\|\phi(y_\star)\|^2 - s) A_{\mathrm{P}}(y_\star; \theta_0)},$$

and denote by τ the initial such time, *i.e.*:

$$\tau := \min\left\{ t \geq 0 : \exp(\ell(\theta_t)) \geq \frac{-(s - \|\phi(y_{\text{med}})\|^2) A_{\text{P}}(y_{\text{med}}; \theta_0)}{(\|\phi(y_\star)\|^2 - s) A_{\text{P}}(y_\star; \theta_0)} \right\}.$$

By the definition of τ, for all $t \in [0, \tau)$ Equation (51) holds. Thus, for all $t \in [0, \tau)$, by Equation (50) $\frac{d}{dt}\ell(\theta_t) < 0$, and so $\ell(\theta_\tau) < \ell(\theta_0)$. This implies that:

$$\exp(\ell(\theta_\tau)) < \exp(\ell(\theta_0)) < \frac{-(s - \|\phi(y_{\text{med}})\|^2) A_{\text{P}}(y_{\text{med}}; \theta_0)}{(\|\phi(y_\star)\|^2 - s) A_{\text{P}}(y_\star; \theta_0)},$$

in contradiction to the definition of τ. Thus, for all $t \geq 0$, Equation (51) holds and $\frac{d}{dt}\ell(\theta_t) < 0$.

Now, examining $\pi_{\theta_t}(y_\star)$, we can see that:

$$\begin{aligned}
\pi_{\theta_t}(y_\star) &\leq \frac{\pi_{\theta_t}(y_\star)}{\pi_{\theta_t}(y_{\text{med}}) + \pi_{\theta_t}(y_\star)} \\
&= \frac{\pi_{\theta_t}(y_\star)/\pi_{\theta_t}(y_{\text{med}})}{1 + \pi_{\theta_t}(y_\star)/\pi_{\theta_t}(y_{\text{med}})} \\
&\leq \frac{\pi_{\theta_0}(y_\star)/\pi_{\theta_0}(y_{\text{med}})}{1 + \pi_{\theta_0}(y_\star)/\pi_{\theta_0}(y_{\text{med}})} \\
&= \frac{\pi_{\theta_0}(y_\star)}{\pi_{\theta_0}(y_\star) + \pi_{\theta_0}(y_{\text{med}})},
\end{aligned}$$

where the second inequality is due to $\pi_{\theta_t}(y_\star)/\pi_{\theta_t}(y_{\text{med}}) = \exp(\ell(\theta_t))$ and the fact that $\ell(\theta_t) \leq \ell(\theta_0)$ for all $t \geq 0$. Thus,

$$1 - \pi_{\theta_t}(y_\star) \geq \frac{\pi_{\theta_0}(y_{\text{med}})}{\pi_{\theta_0}(y_\star) + \pi_{\theta_0}(y_{\text{med}})}.$$

Furthermore, since $r_{\text{G}}(z) < r_{\text{G}}(y_{\text{med}})$, for all $z \in \mathcal{Y}_{\text{bad}}$, we know that

$$V_{\text{G}}(\theta_t) < \pi_{\theta_t}(y_\star) r_{\text{G}}(y_\star) + (1 - \pi_{\theta_t}(y_\star)) r_{\text{G}}(y_{\text{med}}).$$

We may therefore conclude that for all $t \geq 0$:

$$\begin{aligned}
r_{\text{G}}(y_\star) - V_{\text{G}}(\theta_t) &> (1 - \pi_{\theta_t}(y_\star)) r_{\text{G}}(y_\star) - (1 - \pi_{\theta_t}(y_\star)) r_{\text{G}}(y_{\text{med}}) \\
&= (r_{\text{G}}(y_\star) - r_{\text{G}}(y_{\text{med}}))(1 - \pi_{\theta_t}(y_\star)) \\
&\geq \frac{(r_{\text{G}}(y_\star) - r_{\text{G}}(y_{\text{med}})) \pi_{\theta_0}(y_{\text{med}})}{\pi_{\theta_0}(y_\star) + \pi_{\theta_0}(y_{\text{med}})} \\
&\geq (r_{\text{G}}(y_\star) - r_{\text{G}}(y_{\text{med}})) \pi_{\theta_0}(y_{\text{med}}),
\end{aligned}$$

where the second inequality is from the lower bound on $1 - \pi_{\theta_t}(y_\star)$ above, and the last inequality is from $\pi_{\theta_0}(y_\star) + \pi_{\theta_0}(y_{\text{med}}) \leq 1$. □

D.8 Auxiliary Lemmas

In this appendix, we include auxiliary lemmas that are used throughout the proofs in Appendix D.

Lemma 10. *Let $r : \mathcal{Y} \to [-1, 1]$. For a linear softmax policy parameterized by $\theta \in \mathbb{R}^D$, denote the expected reward with respect to r by $V(\theta) := \mathbb{E}_{y \sim \pi_\theta}[r(y)]$. Then:*

$$\nabla V(\theta) = \sum\nolimits_{y \in \mathcal{Y}} \pi_\theta(y) A(y; \theta) \cdot \phi(y),$$

where $A(y; \theta) := r(y) - V(\theta)$ is the advantage of $y \in \mathcal{Y}$ under r and π_θ.

Proof. Since the probability of an output $y \in \mathcal{Y}$ under π_θ is given by

$$\pi_\theta(y) = \frac{\exp(\langle \phi(y), \theta \rangle)}{\sum_{z \in \mathcal{Y}} \exp(\langle \phi(z), \theta \rangle)},$$

the gradient of $\pi_\theta(y)$ with respect to θ can be written as:

$$\nabla \pi_\theta(y) = \pi_\theta(y) \left(\phi(y) - \sum\nolimits_{z \in \mathcal{Y}} \pi_\theta(z) \cdot \phi(z) \right).$$

Hence:

$$\begin{aligned}\nabla V(\theta) &= \sum\nolimits_{y\in\mathcal{Y}} r(y)\nabla\pi_\theta(y)\\ &= \sum\nolimits_{y\in\mathcal{Y}} \pi_\theta(y)r(y)\cdot\phi(y) - \Big(\sum\nolimits_{y\in\mathcal{Y}} \pi_\theta(y)r(y)\Big)\Big(\sum\nolimits_{z\in\mathcal{Y}} \pi_\theta(z)\cdot\phi(z)\Big)\\ &= \sum\nolimits_{y\in\mathcal{Y}} \pi_\theta(y)r(y)\cdot\phi(y) - V(\theta)\Big(\sum\nolimits_{z\in\mathcal{Y}} \pi_\theta(z)\cdot\phi(z)\Big)\\ &= \sum\nolimits_{y\in\mathcal{Y}} \pi_\theta(y)A(y;\theta)\cdot\phi(y).\end{aligned}$$

□

Lemma 11. *For a linear softmax policy parameterized by $\theta \in \mathbb{R}^D$ and ground truth reward function r_{G} upholding Assumption 1, it holds that:*

$$\|\nabla V_{\mathrm{G}}(\theta)\| \le 2B(\Delta_1+\Delta_2)(1-\pi_\theta(y_{\mathrm{med}}))\,,$$

where $\Delta_1 := r_{\mathrm{G}}(y_\star) - r_{\mathrm{G}}(y_{\mathrm{med}})$, $\Delta_2 := r_{\mathrm{G}}(y_{\mathrm{med}}) - r_{\mathrm{G}}(y_{\mathrm{bad}})$ for some $y_{\mathrm{bad}} \in \mathcal{Y}_{\mathrm{bad}}$, and $B := \max_{y\in\mathcal{Y}}\|\phi(y)\|$.

Proof. For a linear softmax policy π_θ, by Lemma 10 we have that:

$$\nabla V_{\mathrm{G}}(\theta) = \sum\nolimits_{y\in\mathcal{Y}} \pi_\theta(y)A_{\mathrm{G}}(y;\theta)\cdot\phi(y).$$

Since all outputs in $\mathcal{Y}_{\mathrm{bad}}$ have the same reward of $r_{\mathrm{G}}(y_{\mathrm{bad}})$, we can write

$$\begin{aligned}V_{\mathrm{G}}(\theta) &= \pi_\theta(y_\star)r_{\mathrm{G}}(y_\star) + \pi_\theta(y_{\mathrm{med}})r_{\mathrm{G}}(y_{\mathrm{med}}) + \pi_\theta(\mathcal{Y}_{\mathrm{bad}})r_{\mathrm{G}}(y_{\mathrm{bad}})\\ &= r_{\mathrm{G}}(y_{\mathrm{med}}) + \pi_\theta(y_\star)\Delta_1 - \pi_\theta(\mathcal{Y}_{\mathrm{bad}})\Delta_2.\end{aligned}$$

Hence the three possible advantage values are:

$$\begin{aligned}A_{\mathrm{G}}(y_\star;\theta) &= r_{\mathrm{G}}(y_\star) - V_{\mathrm{G}}(\theta) = (1-\pi_\theta(y_\star))\Delta_1 + \pi_\theta(\mathcal{Y}_{\mathrm{bad}})\Delta_2,\\ A_{\mathrm{G}}(y_{\mathrm{med}};\theta) &= r_{\mathrm{G}}(y_{\mathrm{med}}) - V_{\mathrm{G}}(\theta) = -\pi_\theta(y_\star)\Delta_1 + \pi_\theta(\mathcal{Y}_{\mathrm{bad}})\Delta_2,\\ A_{\mathrm{G}}(z;\theta) &= r_{\mathrm{G}}(y_{\mathrm{bad}}) - V_{\mathrm{G}}(\theta) = -\pi_\theta(y_\star)\Delta_1 - (1-\pi_\theta(\mathcal{Y}_{\mathrm{bad}}))\Delta_2, \qquad \text{for all } z\in\mathcal{Y}_{\mathrm{bad}}.\end{aligned}$$

Substituting these expressions into the gradient formula gives:

$$\begin{aligned}\nabla V_{\mathrm{G}}(\theta) = \pi_\theta(y_\star)A_{\mathrm{G}}(y_\star;\theta)\phi(y_\star) + \pi_\theta(y_{\mathrm{med}})A_{\mathrm{G}}(y_{\mathrm{med}};\theta)\phi(y_{\mathrm{med}})\\ + \sum\nolimits_{z\in\mathcal{Y}_{\mathrm{bad}}} \pi_\theta(z)A_{\mathrm{G}}(z;\theta)\phi(z).\end{aligned}$$

Therefore, by the triangle inequality and the definition of B,

$$\begin{aligned}\|\nabla V_{\mathrm{G}}(\theta)\| &\le B\Big(\pi_\theta(y_\star)A_{\mathrm{G}}(y_\star;\theta) + \pi_\theta(y_{\mathrm{med}})|A_{\mathrm{G}}(y_{\mathrm{med}};\theta)| + \sum\nolimits_{z\in\mathcal{Y}_{\mathrm{bad}}} \pi_\theta(z)|A_{\mathrm{G}}(z;\theta)|\Big)\\ &= B\big(\pi_\theta(y_\star)A_{\mathrm{G}}(y_\star;\theta) + \pi_\theta(y_{\mathrm{med}})|-\pi_\theta(y_\star)\Delta_1 + \pi_\theta(\mathcal{Y}_{\mathrm{bad}})\Delta_2|\big)\\ &\quad + B\big(\pi_\theta(\mathcal{Y}_{\mathrm{bad}})(\pi_\theta(y_\star)\Delta_1 + (1-\pi_\theta(\mathcal{Y}_{\mathrm{bad}}))\Delta_2)\big).\end{aligned}$$

Applying the triangle inequality, collecting the coefficients of Δ_1 and Δ_2, and using $\pi_\theta(y_\star) + \pi_\theta(y_{\mathrm{med}}) + \pi_\theta(\mathcal{Y}_{\mathrm{bad}}) = 1$, we can further upper bound the above expression as:

$$\begin{aligned}&\|\nabla V_{\mathrm{G}}(\theta)\|\\ &\le B\big(\pi_\theta(y_\star)((1-\pi_\theta(y_\star))\Delta_1 + \pi_\theta(\mathcal{Y}_{\mathrm{bad}})\Delta_2) + \pi_\theta(y_{\mathrm{med}})(\pi_\theta(y_\star)\Delta_1 + \pi_\theta(\mathcal{Y}_{\mathrm{bad}})\Delta_2)\big)\\ &\quad + B\big(\pi_\theta(\mathcal{Y}_{\mathrm{bad}})(\pi_\theta(y_\star)\Delta_1 + (1-\pi_\theta(\mathcal{Y}_{\mathrm{bad}}))\Delta_2)\big)\\ &= B\big(2\pi_\theta(y_\star)(1-\pi_\theta(y_\star))\Delta_1 + 2\pi_\theta(\mathcal{Y}_{\mathrm{bad}})(1-\pi_\theta(\mathcal{Y}_{\mathrm{bad}}))\Delta_2\big).\end{aligned}$$

Since $1-\pi_\theta(y_\star) \le 1$ and $1-\pi_\theta(\mathcal{Y}_{\mathrm{bad}}) \le 1$, we may conclude:

$$\begin{aligned}\|\nabla V_{\mathrm{G}}(\theta)\| &\le 2B(\pi_\theta(y_\star)\Delta_1 + \pi_\theta(\mathcal{Y}_{\mathrm{bad}})\Delta_2)\\ &\le 2B(\Delta_1+\Delta_2)(\pi_\theta(y_\star) + \pi_\theta(\mathcal{Y}_{\mathrm{bad}}))\\ &= 2B(\Delta_1+\Delta_2)(1-\pi_\theta(y_{\mathrm{med}})).\end{aligned}$$

□

Lemma 12. *For a linear softmax policy parameterized by $\theta \in \mathbb{R}^D$ and a ground truth reward function $r_{\mathrm{G}} : \mathcal{Y} \to [-1, 1]$, let $B := \max_{y\in\mathcal{Y}} \|\phi(y)\|$. Then, the expected ground truth reward $V_{\mathrm{G}}(\theta) = \mathbb{E}_{y\sim\pi_\theta}[r_{\mathrm{G}}(y)]$ is B-Lipschitz with respect to the Euclidean norm, i.e., for any $\theta, \theta' \in \mathbb{R}^D$:*

$$|V_{\mathrm{G}}(\theta) - V_{\mathrm{G}}(\theta')| \le B\|\theta - \theta'\|.$$

Proof. By Lemma 10, for $r = r_{\mathrm{G}}$ we have:

$$\nabla V_{\mathrm{G}}(\theta) = \sum\nolimits_{y\in\mathcal{Y}} \pi_\theta(y) A_{\mathrm{G}}(y;\theta) \cdot \phi(y).$$

Since $\sum_{y\in\mathcal{Y}} \pi_\theta(y) A_{\mathrm{G}}(y;\theta) = 0$, we may rewrite this as:

$$\nabla V_{\mathrm{G}}(\theta) = \sum\nolimits_{y\in\mathcal{Y}} \pi_\theta(y) A_{\mathrm{G}}(y;\theta) \cdot \big(\phi(y) - \bar{\phi}_\theta\big),$$

where $\bar{\phi}_\theta := \sum_{z\in\mathcal{Y}} \pi_\theta(z)\phi(z)$. Applying the triangle inequality followed by the Cauchy-Schwarz inequality gives:

$$\begin{aligned}
\|\nabla V_{\mathrm{G}}(\theta)\| &\le \sum\nolimits_{y\in\mathcal{Y}} \pi_\theta(y) |A_{\mathrm{G}}(y;\theta)| \big\|\phi(y) - \bar{\phi}_\theta\big\| \\
&\le \Big(\sum\nolimits_{y\in\mathcal{Y}} \pi_\theta(y) A_{\mathrm{G}}(y;\theta)^2\Big)^{1/2} \Big(\sum\nolimits_{y\in\mathcal{Y}} \pi_\theta(y) \big\|\phi(y) - \bar{\phi}_\theta\big\|^2\Big)^{1/2} \\
&= \mathrm{Var}_{y\sim\pi_\theta}[r_{\mathrm{G}}(y)]^{1/2} \Big(\mathbb{E}_{y\sim\pi_\theta}\big[\|\phi(y)\|^2\big] - \big\|\bar{\phi}_\theta\big\|^2\Big)^{1/2} \\
&\le B,
\end{aligned}$$

where the last inequality follows from the fact that $r_{\mathrm{G}}(y) \in [-1, 1]$ implies $\mathrm{Var}_{y\sim\pi_\theta}[r_{\mathrm{G}}(y)] \le 1$, and from $\mathbb{E}_{y\sim\pi_\theta}\big[\|\phi(y)\|^2\big] - \|\bar{\phi}_\theta\|^2 \le \mathbb{E}_{y\sim\pi_\theta}\big[\|\phi(y)\|^2\big] \le B^2$. Therefore, by the mean value theorem,

$$|V_{\mathrm{G}}(\theta) - V_{\mathrm{G}}(\theta')| \le \sup\nolimits_{\xi\in[\theta,\theta']} \|\nabla V_{\mathrm{G}}(\xi)\| \, \|\theta - \theta'\| \le B\|\theta - \theta'\|.$$

□

Lemma 13. *Let $r : \mathcal{Y} \to [-1, 1]$. For a linear softmax policy parameterized by $\theta \in \mathbb{R}^D$, denote the expected reward with respect to r by $V(\theta) := \mathbb{E}_{y\sim\pi_\theta}[r(y)]$. Suppose that gradient flow is used to maximize the expected reward with respect to r (Equation (1) with V in place of V_{P}). Then, for any output $y \in \mathcal{Y}$,*

$$\frac{d}{dt}\pi_{\theta_t}(y) = \pi_{\theta_t}(y) \left\langle \phi(y) - \sum\nolimits_{z\in\mathcal{Y}} \pi_{\theta_t}(z)\phi(z), \sum\nolimits_{z\in\mathcal{Y}} \pi_{\theta_t}(z) A(z;\theta_t)\phi(z) \right\rangle,$$

where $A(z;\theta_t) := r(z) - V(\theta_t)$ is the advantage of $z \in \mathcal{Y}$ under r and π_{θ_t}. Moreover, if the feature vectors $\{\phi(z)\}_{z\in\mathcal{Y}}$ are orthonormal, then

$$\frac{d}{dt}\pi_{\theta_t}(y) = \pi_{\theta_t}(y)\Big(\pi_{\theta_t}(y) A(y;\theta_t) - \sum\nolimits_{z\in\mathcal{Y}} \pi_{\theta_t}(z)^2 A(z;\theta_t)\Big).$$

Proof. By the chain rule and the gradient flow dynamics,

$$\frac{d}{dt}\pi_{\theta_t}(y) = \big\langle \nabla\pi_{\theta_t}(y), \tfrac{d}{dt}\theta_t \big\rangle = \big\langle \nabla\pi_{\theta_t}(y), \nabla V(\theta_t) \big\rangle.$$

Since π_θ is a linear softmax policy, we have

$$\nabla\pi_{\theta_t}(y) = \pi_{\theta_t}(y)\Big(\phi(y) - \sum\nolimits_{z\in\mathcal{Y}} \pi_{\theta_t}(z)\phi(z)\Big).$$

Moreover, by Lemma 10,

$$\nabla V(\theta_t) = \sum\nolimits_{z\in\mathcal{Y}} \pi_{\theta_t}(z) A(z;\theta_t)\phi(z).$$

Substituting these two expressions into the derivative identity above gives the first claim. If, in addition, the feature vectors are orthonormal, then $\langle\phi(y), \phi(z)\rangle = \mathbb{1}[y = z]$ for all $y, z \in \mathcal{Y}$, and therefore

$$\begin{aligned}
&\Big\langle \phi(y) - \sum\nolimits_{z\in\mathcal{Y}} \pi_{\theta_t}(z)\phi(z), \sum\nolimits_{u\in\mathcal{Y}} \pi_{\theta_t}(u) A(u;\theta_t)\phi(u) \Big\rangle \\
&= \pi_{\theta_t}(y) A(y;\theta_t) - \sum_{z\in\mathcal{Y}} \pi_{\theta_t}(z)^2 A(z;\theta_t).
\end{aligned}$$

Plugging this back into the general expression yields the orthonormal-features formula. □

Lemma 14. *Under Assumptions 1 and 2, it holds that* $V_{\mathrm{G}}(\theta_0) = \mathbb{E}_{y\sim\pi_{\theta_0}}[r_{\mathrm{G}}(y)] < r_{\mathrm{G}}(y_{\mathrm{med}}) - \sqrt{\gamma}$, *for* $\gamma := \pi_{\theta_0}(y_\star)^{14/13} M^{-14/13}$ *with* M *defined as in Assumption 2.*

Proof. Fix some $y_{\mathrm{bad}} \in \mathcal{Y}_{\mathrm{bad}}$. Under Assumptions 1 and 2, Lemma 4 establishes bounds on γ, $\pi_{\theta_0}(y_\star)$, and $\pi_{\theta_0}(\mathcal{Y}_{\mathrm{bad}})$. Based on these bounds, we can show that $V_{\mathrm{G}}(\theta_0) < r_{\mathrm{G}}(y_{\mathrm{med}}) - \sqrt{\gamma}$ as follows:

$$\begin{aligned}
V_{\mathrm{G}}(\theta_0) &= r_{\mathrm{G}}(y_{\mathrm{med}})\pi_{\theta_0}(y_{\mathrm{med}}) + r_{\mathrm{G}}(y_\star)\pi_{\theta_0}(y_\star) + \sum\nolimits_{z\in\mathcal{Y}_{\mathrm{bad}}} r_{\mathrm{G}}(z)\pi_{\theta_0}(z) \\
&\le (1 - \pi_{\theta_0}(\mathcal{Y}_{\mathrm{bad}}))r_{\mathrm{G}}(y_{\mathrm{med}}) + \gamma^{13/14} + \pi_{\theta_0}(\mathcal{Y}_{\mathrm{bad}})r_{\mathrm{G}}(y_{\mathrm{bad}}) \\
&= r_{\mathrm{G}}(y_{\mathrm{med}}) - \pi_{\theta_0}(\mathcal{Y}_{\mathrm{bad}})\Delta_2 + \gamma^{13/14} \\
&\le r_{\mathrm{G}}(y_{\mathrm{med}}) - 2\sqrt{\gamma} + \gamma^{13/14} \\
&< r_{\mathrm{G}}(y_{\mathrm{med}}) - \sqrt{\gamma},
\end{aligned}$$

The first inequality is from $r_{\mathrm{G}}(y_\star) \le 1$ and $\pi_{\theta_0}(y_\star) \le \gamma^{13/14}$ (Item 2 in Lemma 4); the second inequality is from $\pi_{\theta_0}(\mathcal{Y}_{\mathrm{bad}}) \ge 2\sqrt{\gamma}/\Delta_2$ (Item 3 in Lemma 4); and the last inequality is due to $\sqrt{\gamma} > \gamma^{13/14}$ for $\gamma \le 0.01$ (Item 1 in Lemma 4). □

Lemma 15. *Under Assumptions 1 and 3, it holds that* $V_{\mathrm{G}}(\theta_0) = \mathbb{E}_{y\sim\pi_{\theta_0}}[r_{\mathrm{G}}(y)] < r_{\mathrm{G}}(y_{\mathrm{med}}) - \sqrt{\gamma}$, *for* $\gamma := \pi_{\theta_0}(y_\star)^{14/13} M'^{-14/13}$ *with* M' *defined as in Assumption 3.*

Proof. Fix some $y_{\mathrm{bad}} \in \mathcal{Y}_{\mathrm{bad}}$. Under Assumptions 1 and 3, Lemma 6 establishes bounds on γ, $\pi_{\theta_0}(y_\star)$, and $\pi_{\theta_0}(\mathcal{Y}_{\mathrm{bad}})$. The proof of this lemma is analogous to that of Lemma 14, using the bounds provided by Lemma 6 in place of those from Lemma 4. Based on these bounds, we can show that $V_{\mathrm{G}}(\theta_0) < r_{\mathrm{G}}(y_{\mathrm{med}}) - \sqrt{\gamma}$ as follows:

$$\begin{aligned}
V_{\mathrm{G}}(\theta_0) &= r_{\mathrm{G}}(y_{\mathrm{med}})\pi_{\theta_0}(y_{\mathrm{med}}) + r_{\mathrm{G}}(y_\star)\pi_{\theta_0}(y_\star) + \sum\nolimits_{z\in\mathcal{Y}_{\mathrm{bad}}} r_{\mathrm{G}}(z)\pi_{\theta_0}(z) \\
&\le (1 - \pi_{\theta_0}(\mathcal{Y}_{\mathrm{bad}}))r_{\mathrm{G}}(y_{\mathrm{med}}) + \gamma^{13/14} + \pi_{\theta_0}(\mathcal{Y}_{\mathrm{bad}})r_{\mathrm{G}}(y_{\mathrm{bad}}) \\
&= r_{\mathrm{G}}(y_{\mathrm{med}}) - \pi_{\theta_0}(\mathcal{Y}_{\mathrm{bad}})\Delta_2 + \gamma^{13/14} \\
&\le r_{\mathrm{G}}(y_{\mathrm{med}}) - 2\sqrt{\gamma} + \gamma^{13/14} \\
&< r_{\mathrm{G}}(y_{\mathrm{med}}) - \sqrt{\gamma}.
\end{aligned}$$

The first inequality is from $r_{\mathrm{G}}(y_\star) \le 1$ and $\pi_{\theta_0}(y_\star) \le \gamma^{13/14}$ (Item 2 in Lemma 6); the second inequality is from $\pi_{\theta_0}(\mathcal{Y}_{\mathrm{bad}}) \ge 2\sqrt{\gamma}/\Delta_2$ (Item 3 in Lemma 6); and the last inequality is due to $\sqrt{\gamma} > \gamma^{13/14}$ for $\gamma \le 0.01$ (Item 1 in Lemma 6). □

Lemma 16. *Under Assumption 1, for all* $\gamma > 0$, *if* $V_{\mathrm{G}}(\theta_t) \le r_{\mathrm{G}}(y_{\mathrm{med}}) - \gamma$, *then* $\pi_{\theta_t}(y_{\mathrm{med}}) \le 1 - \frac{\gamma}{1+r_{\mathrm{G}}(y_{\mathrm{med}})}$.

Proof. We will proceed by contradiction. Suppose to the contrary that $\pi_{\theta_t}(y_{\mathrm{med}}) > 1 - \frac{\gamma}{1+r_{\mathrm{G}}(y_{\mathrm{med}})}$. Then,

$$\begin{aligned}
V_{\mathrm{G}}(\theta_t) &= \pi_{\theta_t}(y_{\mathrm{med}})r_{\mathrm{G}}(y_{\mathrm{med}}) + \pi_{\theta_t}(y_\star)r_{\mathrm{G}}(y_\star) + \sum_{z\in\mathcal{Y}_{\mathrm{bad}}} \pi_{\theta_t}(z)r_{\mathrm{G}}(z) \\
&> \left(1 - \frac{\gamma}{1 + r_{\mathrm{G}}(y_{\mathrm{med}})}\right) r_{\mathrm{G}}(y_{\mathrm{med}}) - \pi_{\theta_t}(y_\star) - \pi_{\theta_t}(\mathcal{Y}_{\mathrm{bad}}) \\
&> \left(1 - \frac{\gamma}{1 + r_{\mathrm{G}}(y_{\mathrm{med}})}\right) r_{\mathrm{G}}(y_{\mathrm{med}}) - \frac{\gamma}{1 + r_{\mathrm{G}}(y_{\mathrm{med}})} \\
&= r_{\mathrm{G}}(y_{\mathrm{med}}) - \gamma.
\end{aligned}$$

Here, the first inequality is from our assumption on $\pi_{\theta_t}(y_{\mathrm{med}})$, $r_{\mathrm{G}}(y_{\mathrm{med}}) > 0$ (Assumption 1), and $r_{\mathrm{G}}(y) \ge -1$ for all $y \in \mathcal{Y}$; the second inequality is due to $\sum_{y\in\mathcal{Y}\setminus\{y_{\mathrm{med}}\}} \pi_{\theta_t}(y) < \frac{\gamma}{1+r_{\mathrm{G}}(y_{\mathrm{med}})}$ also by assumption. This contradicts the premise that $V_{\mathrm{G}}(\theta_t) \le r_{\mathrm{G}}(y_{\mathrm{med}}) - \gamma$. Hence, $\pi_{\theta_t}(y_{\mathrm{med}}) \le 1 - \frac{\gamma}{1+r_{\mathrm{G}}(y_{\mathrm{med}})}$, as desired. □

E Additional Experimental Results

E.1 Empirical Demonstration of Theoretical Results With Linear Softmax Policies

Figure 6 supplements Figure 2 by comparing the evolution of output probabilities during policy gradient in settings corresponding to Theorem 1 when the mediocre output y_{med} receives either mediocre or low proxy reward.

Figure 7 corroborates Figure 2 by presenting the results of an identical experiment, except that the policy is trained using REINFORCE (Williams, 1992), *i.e.*, sample-based gradients of the expected proxy reward, instead of exact gradients. The results show that outputs with mediocre proxy reward can impede optimization regardless of whether one uses exact or sample-based gradients.

E.2 Comparison of Ranking Accuracy Variants

Listed below are experimental results omitted from Section 4.

- Figure 8 supplements Figure 3 by reporting the correlation with ground truth reward increase and regret of ranking accuracy variants, when measuring reward increase on test examples from the UltraFeedback dataset as opposed to examples from the policy gradient training set. The results are analogous to Figure 3 since, as can be seen in Tables 2, 4, 6, and 8, the reward increase on test examples is nearly identical to that on training examples.
- Figure 9 supplements Figure 3 by considering ranking accuracy variants computed on RewardBench2 instead of on UltraFeedback examples from the policy gradient training set. Across language models, accuracy values computed on RewardBench2 are substantially less predictive of which reward model leads to better language model performance, as indicated by the negative correlation and high regret. This highlights the importance of evaluating reward models on prompts close in distribution to those used for policy gradient training.
- Figure 10 supports the experiments of Figure 3 by showing that similar trends persist when using: *(i)* a different dataset for policy gradient training and reward model evaluation (WildChat-IF-On-Policy-8B (Lambert et al., 2024) instead of UltraFeedback); or *(ii)* a different ground truth reward model (Skywork-Reward-V2-Llama-3.1-8B (Liu et al., 2025a) instead of ArmoRM). It is worth noting that on the WildChat-IF dataset, HAcc improves upon Acc in both correlation and regret, but the correlation remains negative. We attribute this finding to the fact that outputs in WildChat-IF are substantially less probable under the initial language model compared to outputs in UltraFeedback, and are thus less indicative of policy gradient outcomes. Specifically, for WildChat-IF the mean and median log probabilities of outputs under the Llama-3.2-1B-Instruct model used in Figure 10 are -346.8 and -309.3, respectively, whereas for UltraFeedback they are substantially higher: -259.4 and -224.5, respectively. This highlights the importance of evaluating reward models on outputs that are probable under the language model being trained.
- Figure 11 supports the experiments of Figure 3 by showing that similar trends persist when evaluating language model performance through win-rate against the initial language model, according to a frontier (GPT) judge model, instead of through ground truth reward increase.
- Tables 2 to 15 include the numerical results based on which Figures 3, 8, 9, 10, and 11 were generated. Specifically, for each setting (*i.e.*, language model, reward model, and dataset), they report the increase in ground truth reward due to policy gradient training (or win-rate against the initial language model), as well as reward model accuracy values.

E.3 When Should Partially Correct Outputs Be Rewarded?

Listed below are experimental results omitted from Section 5.

- Figures 12 and 13 present analogous results to Figure 4 for the Llama-3.2-3B-Instruct and OLMo-2-1B-Instruct language models (instead of Qwen3-1.7B), respectively.
- Figure 14 demonstrates that the probability of initially satisfying a constraint is not the sole factor determining its ease of learnability. In particular, it shows that when one constraint is initially satisfied with substantially higher probability than the other, it is possible for both constraints to

be learned quickly under partial rewards. In such cases, under binary rewards (*i.e.*, rewarding only fully correct outputs), the constraints can either also be learned quickly or fail to be learned, especially when the initial probability of satisfying both constraints is near-zero.

- Table 16 specifies the IFBench constraint pairs used in each of the experiments corresponding to Figures 4, 12, 13, and 14.

F Additional Implementation Details

In this appendix, we provide implementation details omitted from the main text and Appendix E. Code for reproducing our results, based on the PyTorch (Paszke et al., 2017), Hugging Face TRL (Wolf et al., 2019), and open-instruct (Lambert et al., 2024; Olmo et al., 2025) libraries, is attached as supplementary material and will be made publicly available.

F.1 Empirical Demonstration of Theoretical Results With Linear Softmax Policies

Output features. The feature vectors are of dimension $D = 5$ and the number of outputs is 5 in all experiments. For Figures 2 and 7, we use standard basis vectors as output features. Specifically, $\phi(y_\star) = (1, 0, 0, 0, 0)$, $\phi(y_{\text{med}}) = (0, 1, 0, 0, 0)$, and the remaining three outputs have feature vectors $(0, 0, 1, 0, 0)$, $(0, 0, 0, 1, 0)$, and $(0, 0, 0, 0, 1)$. For Figure 5, the feature vectors of outputs other than $y_\star$ and y_{med} are the same as in the experiments of Figures 2 and 7. The feature vectors of $y_\star$ and y_{med} are set as follows.

- In the case of $\langle \phi(y_\star), \phi(y_{\text{med}}) \rangle = 0$:
$$\phi(y_\star) = (3/2, 0, 0, 0, 0) \quad , \quad \phi(y_{\text{med}}) = (0, 1, 0, 0, 0)\,.$$
- In the case of $\langle \phi(y_\star), \phi(y_{\text{med}}) \rangle < 0$:
$$\phi(y_\star) = (3/2, 0, 0, 0, 0) \quad , \quad \phi(y_{\text{med}}) = (-1/\sqrt{2}, 1/\sqrt{2}, 0, 0, 0)\,.$$
- In the case of $\langle \phi(y_\star), \phi(y_{\text{med}}) \rangle > 0$:
$$\phi(y_\star) = (3/2, 0, 0, 0, 0) \quad , \quad \phi(y_{\text{med}}) = (1/\sqrt{2}, 1/\sqrt{2}, 0, 0, 0)\,.$$

Note that the norm of $\phi(y_\star)$ is taken to be larger than that of $\phi(y_{\text{med}})$ to accord with the conditions of Theorem 4. We observed trends analogous to those in Figure 5 when setting the first coordinate of $\phi(y_\star)$ to 1 instead of $3/2$, in which case all feature vectors have unit norm.

Initial policy. For the experiments of Figures 2 and 7, we set the initial policy weights θ_0 such that $\pi_{\theta_0}(y_\star)$ is either 0.05, 0.1, or 0.15, $\pi_{\theta_0}(y_{\text{med}}) = 0.5$, and the remaining probability mass is uniformly divided between the remaining three outputs. Since the feature vectors in these experiments are standard basis vectors, we can achieve this by setting the coordinate of θ_0 associated with each output $y \in \mathcal{Y}$ to the logarithm of the desired probability. For the experiments of Figure 5, we set θ_0 such that $\pi_{\theta_0}(y_\star) = 0.05$, $\pi_{\theta_0}(y_{\text{med}}) = 0.5$, and the remaining probability mass is uniformly divided between the remaining three outputs. This is done by using a linear equations solver to find a θ_0 satisfying $\langle \phi(y), \theta_0 \rangle = \ln \pi(y)$ for all $y \in \mathcal{Y}$, where $\pi(y)$ is the desired initial probability of y.

Policy gradient optimization. In the experiments of Figures 2 and 5, we train linear softmax policies by running gradient ascent over the expected proxy or ground truth rewards using exact gradients. In Figure 7, we use REINFORCE (Williams, 1992). Namely, at each training step, we sample an output y from the current policy and update the parameters via $\theta_{t+1} = \theta_t + \eta \cdot r_{\text{P}}(y) \cdot \nabla_\theta \ln \pi_{\theta_t}(y)$, where $r_{\text{P}}(y) \cdot \nabla_\theta \ln \pi_{\theta_t}(y)$ is an unbiased estimate of $\nabla V_{\text{P}}(\theta_t)$ and $\eta > 0$ is the learning rate. We set the learning rate to 0.1 in all experiments. Each figure caption specifies the reward assignments in the corresponding experiments.

Hardware. Standard laptop CPU.

F.2 Comparison of Ranking Accuracy Variants

Language models. We consider as initial policies four language models that cover different model families and types (*i.e.*, pretrained, supervised finetuned, and instruction-tuned): Llama-3.2-3B-Instruct, Llama-3.2-1B-Instruct, OLMo-2-1B-SFT, and Qwen3-1.7B-Base. When processing inputs to the language models we use their default chat templates.

Reward models. Table 1 lists the reward models used in the experiments of Section 4 and Appendix E.2. The reward models were chosen so that they cover a range of RewardBench2 scores (Malik et al., 2025). When using Skywork-Reward-V2-Llama-3.1-8B as the ground truth reward model (Figure 10), we exclude it from the reward models used for training and include instead the ArmoRM (Wang et al., 2024) model.

Table 1: Listed are the reward models used in the experiments of Section 4 and Appendix E.2. The RewardBench2 scores are taken from the official leaderboard (Malik et al., 2025). The two models without a RewardBench2 score were chosen and positioned in the table based on their original RewardBench score (Lambert et al., 2025).

Reward Model	RewardBench2 Score
Skywork-Reward-V2-Llama-3.1-8B	84.1
Skywork-Reward-V2-Qwen3-8B	78.4
Skywork-Reward-V2-Qwen3-4B	75.5
Llama-3.1-8B-Instruct-RM-RB2	72.8
Skywork-Reward-V2-Qwen3-1.7B	68.2
GRM-Llama3.2-3B-rewardmodel-ft	–
Llama-3-OffsetBias-RM-8B	64.8
Skywork-Reward-V2-Llama-3.2-1B	64.4
Skywork-Reward-V2-Qwen3-0.6B	61.2
RM-Mistral-7B	59.6
internlm2-1_8b-reward	39.0
llama-3-tulu-2-8b-uf-mean-rm	–
RM-Gemma-2B	30.6

Reward normalization. To ensure fair comparison across reward models, which can differ in the scale of rewards they produce, we normalize rewards in each batch of training by subtracting the mean and dividing by the standard deviation.

Data. We take the binarized version of UltraFeedback[4] (Cui et al., 2024) and filter out examples in which the prompt or one of the outputs exceeds 512 tokens. We then select a subset of 10000 examples from the training set (`train_prefs` split), relabel output preferences in the selected subset and the test set (`test_prefs` split) using the ground truth reward model (ArmoRM or Skywork-Reward-V2-Llama-3.1-8B, depending on the experiment), and filter out examples where both outputs have the same ground truth reward. For the ArmoRM ground truth reward model, the resulting training and test sets contain 9808 and 1406 examples, respectively, while for the Skywork-Reward-V2-Llama-3.1-8B ground truth reward model, the resulting training and test sets contain 9930 and 1422 examples, respectively. For experiments on the WildChat-IF-On-Policy-8B dataset[5] (Figure 10), which does not have a test set, we follow the same procedure using 3000 randomly selected examples from the dataset and split the examples remaining after filtering into a training set of 1969 examples and a test set of 985 examples.

Policy gradient optimization. Our RLOO implementation is based on the RLOOTrainer class from the TRL framework, which uses the Adam optimizer. We adopt RLOO because it is more resource efficient than PPO (Schulman et al., 2017) and has shown competitive results (Ahmadian et al., 2024). We also experimented with GRPO (Shao et al., 2024) and observed analogous results. We set the learning rate to 1e-7, batch size to 32, KL regularization coefficient to 0.01, and the `num_mini_batches` hyperparameter to 1. Optimization is carried out for two full passes over the prompts in the training set, where for each prompt in a batch we sample four outputs.

Generation hyperparameters. For both training and evaluation, we generate outputs from the policies (*i.e.*, language models) using a temperature of 1 and a maximum output length of 512 tokens.

Ground truth reward increase policy evaluation. We evaluate the ground truth reward achieved by a policy based on fixed (randomly selected) subsets of 500 prompts from the policy gradient training and test sets. We sample 10 outputs from the policy for each prompt, compute their ground truth rewards, and then average the rewards across all outputs and prompts. As noted in Section 4, we mainly consider the ground truth reward increase on training examples as it directly reflects policy

[4] https://huggingface.co/datasets/HuggingFaceH4/ultrafeedback_binarized

[5] https://huggingface.co/datasets/allenai/tulu-3-wildchat-if-on-policy-8b

gradient optimization, which is the focus of our work, without conflating it with generalization. Nonetheless, as reported in Appendix E, we found the reward increase on test examples to be nearly identical. For each combination of reward model and initial policy, the reported ground truth reward increase is the mean across three separate runs with different random seeds.

Win-rate policy evaluation. For Figure 11, instead of using a ground truth reward model for policy evaluation, we adopt the AlpacaEval framework (Li et al., 2023) to compute win-rates against the initial policy according to a GPT judge model. The win-rates are computed for a single policy gradient run per combination of reward model and initial policy, based on the same 500 prompts used for evaluation via ground truth reward increase. We sample one output from the initial policy and one from the final policy, for each prompt, and ask the judge model to select the better output using the format provided by AlpacaEval. For all models, except Qwen3-1.7B-Base, `gpt-4-1106-preview` served as the judge model. For Qwen3-1.7B-Base we used `gpt-4.1-2025-04-14` as the judge since the win-rates for this model were computed after the deprecation of `gpt-4-1106-preview`.

Reward model evaluation. For each reward model and initial policy in Figures 3, 8, 10, and 11, we compute all ranking-accuracy variants (Acc, Acc-W, HAcc, and HAcc-W) on the same 500 training prompts used for policy evaluation, where output preferences are relabeled using the ground truth reward model. In Figure 9, accuracy values are computed on RewardBench2, after excluding examples from the "Ties" subset and relabeling preferences using the ground truth reward model. Lastly, to compute HAcc and HAcc-W, we estimate the mean proxy reward achieved by an initial policy for each prompt using 10 sampled outputs.

Computational cost of harm-aware ranking accuracy variants. Unlike standard ranking accuracy, the proposed harm-aware variants depend on the language model (*i.e.*, policy) π_θ. This dependence is desirable: our analysis, together with prior work (Razin et al., 2025), shows that the effectiveness of a reward model is inherently policy-dependent. The drawback is a modest computational overhead. Namely, these variants require sampling from π_θ to estimate $\bar{V}_{\mathrm{P}}(x;\theta)$ and computing the probabilities of outputs in the dataset $\mathcal{S}$. In practice, this overhead is small relative to the cost of running policy gradient and is further reduced by reusing samples and probabilities when evaluating multiple reward models.

Hardware. All experiments ran on Nvidia H100 GPUs with 80GB memory. For the Llama-3.2-1B-Instruct and OLMo-2-1B-SFT models, we used two GPUs per policy gradient run. For the Llama-3.2-3B-Instruct and Qwen3-1.7B-Base models, we used four GPUs per policy gradient run. Evaluations were carried out on a single GPU.

F.3 When Should Partially Correct Outputs Be Rewarded?

Language models. We consider as initial policies three instruction-tuned language models: Qwen3-1.7B (thinking disabled), OLMo-2-1B-Instruct, and Llama-3.2-3B-Instruct. When processing inputs to the language models we use their default chat templates.

Data. We take the binarized version of UltraFeedback and randomly sample 4000 prompts that do not exceed 512 tokens from the `train_prefs` split. Then, we construct instruction following datasets by appending to each prompt a pair of constraints from IFBench (Pyatkin et al., 2025) using the following template:

```
{original prompt} {constraint 1} {constraint 2}
```

All prompts within a dataset share the same pair of constraints and the datasets differ in which constraints they contain. For each initial policy, we experimented with multiple pairs of constraints and report the results for pairs exhibiting representative behaviors.

Reward design. We consider two reward schemes. The "partial rewards" scheme corresponds to giving a reward of 0.5 for each constraint satisfied. Thus, an output receives reward 0, 0.5, or 1 depending on whether it satisfies none, exactly one, or both constraints. The "binary (full-correctness) rewards" scheme corresponds to assigning a reward of 1 to fully correct outputs, satisfying both constraints, and 0 to all other outputs.

Policy gradient optimization. We rely on the open-instruct framework and train in bfloat16 mixed precision using its default DAPO objective implementation (Yu et al., 2025). Under our configuration, which adopts symmetric clipping (with $\varepsilon = 0.2$), disables dynamic sampling, sets

`non_stop_penalty` to false, and does not use KL regularization, the objective reduces to the GRPO objective with token-level normalization. Each batch consists of 16 unique prompts. The number of sampled outputs per prompt varies across language models: 96 for Qwen3-1.7B, 64 for Llama-3.2-3B-Instruct, and 32 for OLMo-2-1B-Instruct. Optimization is carried out for four full passes over the training prompts using the Adam optimizer and a linear decay learning rate schedule with warmup ratio 0.03 and base learning rate 1e-6. We chose these relatively large numbers of outputs per prompt together with this learning rate schedule to promote stable training. Note that for the learning rate decay schedule we set the number of epochs to eight, but ended up stopping early after four epochs (1000 training steps) since we observed that the probabilities of success tend to plateau by that point.

Policy evaluation. The probability of satisfying a constraint (or both constraints jointly) is estimated at each training step by the fraction of outputs in a batch that satisfy the constraint (or both constraints).

Generation hyperparameters. We generate outputs from the policies (*i.e.*, language models) using a temperature of 1 and a maximum output length of 512 tokens.

Hardware. Each experiment ran on two Nvidia H100 GPUs with 80GB memory.

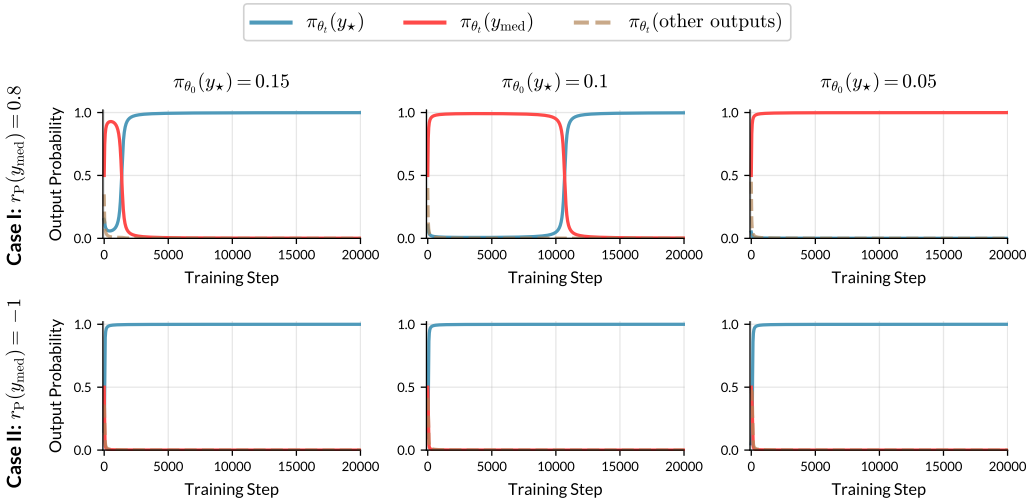

Figure 6: **Attraction to mediocre outputs can impede policy gradient optimization.** This figure supplements Figure 2 by comparing the evolution of output probabilities during policy gradient in settings corresponding to Theorem 1 when the mediocre output y_{med} receives either mediocre or low proxy reward. The ground truth reward r_{G} assigns a maximal reward of 1 to $y_\star$, a mediocre reward of 0.8 to y_{med}, and a low reward of -1 to all remaining outputs. The top row presents the experiments presented in Figure 2, where $r_{\text{P}} = r_{\text{G}}$. Bottom row shows experiments in identical settings, except that $r_{\text{P}}(y_{\text{med}}) = -1$. In line with Theorem 1, when y_{med} receives mediocre proxy reward, the policy initially concentrates its probability mass on y_{med} and stagnates. By contrast, when y_{med} is assigned low proxy reward, the policy directly increases the probability of $y_\star$. The gap in optimization time between the two cases becomes larger as $\pi_{\theta_0}(y_\star)$ decreases (left to right). See Appendix F for additional implementation details.

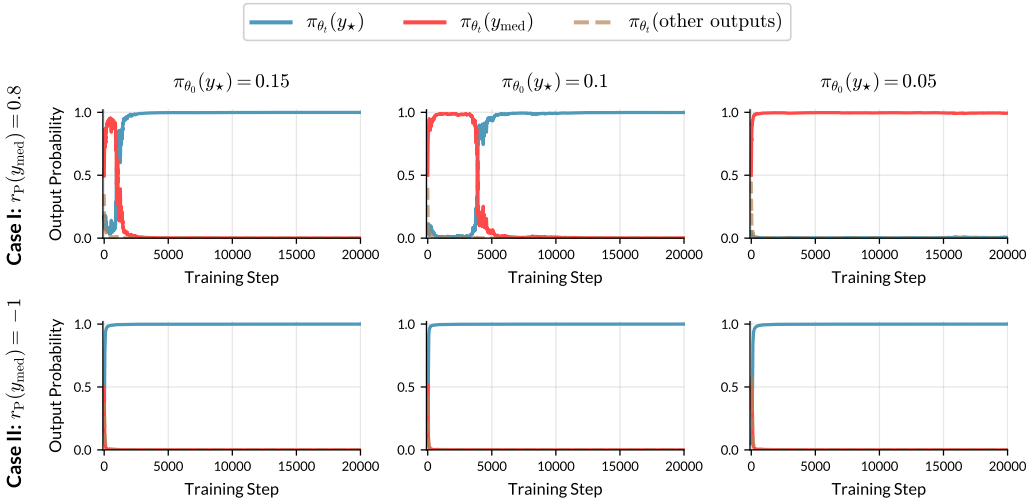

Figure 7: **Attraction to outputs with mediocre proxy reward can impede policy gradient optimization (with sample-based gradients).** This figure presents the results of an experiment identical to that of Figure 2, except that the policy is trained using REINFORCE (Williams, 1992), *i.e.*, sample-based gradients of the expected proxy reward, instead of exact gradients. As when training with exact gradients of the expected proxy reward, outputs with mediocre proxy reward can impede optimization. See Appendix F for additional implementation details.

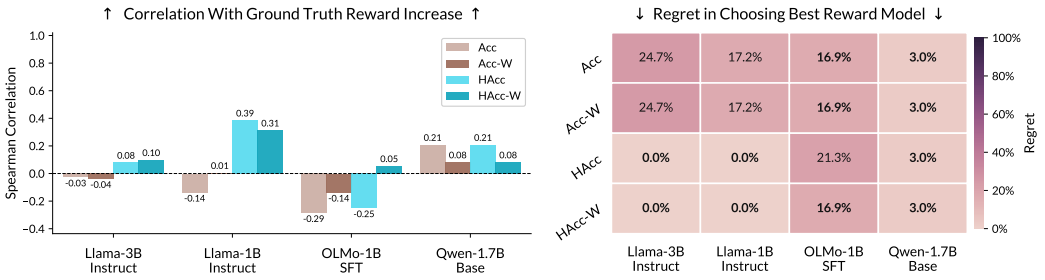

Figure 8: **Harm-aware ranking accuracy variants are more predictive of which reward model leads to better language model performance.** This figure supplements Figure 3 by reporting the same metrics, but based on the reward increase over test examples from the UltraFeedback dataset. The results are analogous to Figure 3, where reward increase is measured on UltraFeedback examples used for policy gradient training. Specifically, as can be seen in Tables 2, 4, 6, and 8, the reward increase on test examples is nearly identical to that on training examples. Additional implementation details are provided in Appendix F.

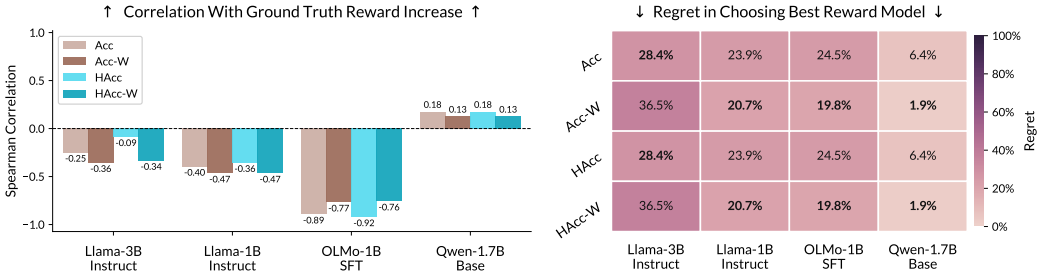

Figure 9: **Ranking accuracy variants computed on prompts differing from those used for policy gradient training are less predictive of which reward model leads to better language model performance.** This figure supplements Figure 3 by considering ranking accuracy variants computed on RewardBench2 instead of on UltraFeedback examples from the policy gradient training set. Across language models, accuracy values computed on RewardBench2 are substantially less predictive of which reward model leads to better language model performance, as indicated by the negative correlation and high regret. This highlights a pitfall of evaluating reward models based on prompts that differ from those used for training the language model. See Appendix F for additional implementation details.

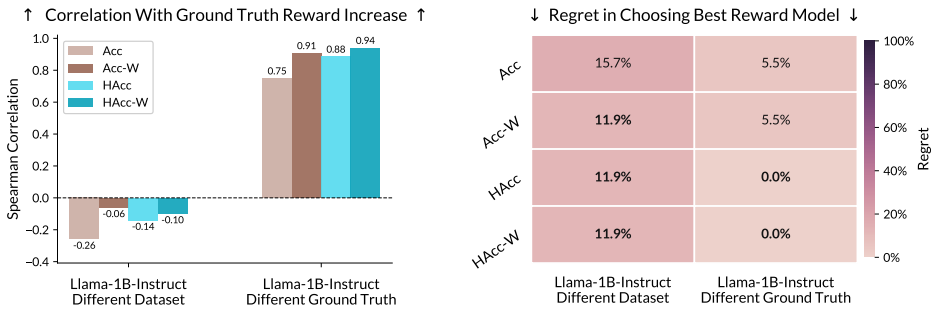

Figure 10: **Harm-aware ranking accuracy variants are more predictive of which reward model leads to better language model performance.** This figure supports the experiments of Figure 3 by showing the same trends persist when using: *(i)* a different dataset for policy gradient training and reward model evaluation (WildChat-IF-On-Policy-8B (Lambert et al., 2024) instead of UltraFeedback); or *(ii)* a different ground truth reward model for evaluating language model performance and accuracy values (Skywork-Reward-V2-Llama-3.1-8B (Liu et al., 2025a) instead of ArmoRM). Both sets of experiments were conducted with the Llama-3.2-1B-Instruct language model. See Appendix F for additional implementation details.

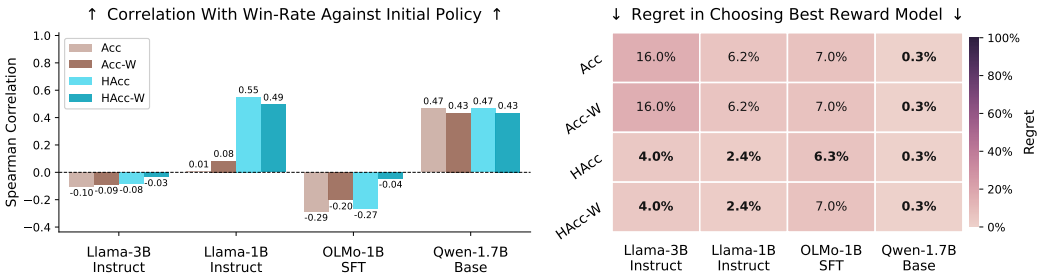

Figure 11: **Harm-aware ranking accuracy variants are more predictive of which reward model leads to better language model performance.** This figure supports Figure 3 by showing that similar trends persist when evaluating language model performance through win-rate against the initial language model, according to a frontier (GPT) judge model, instead of through ground truth reward increase. See Appendix F for additional implementation details.

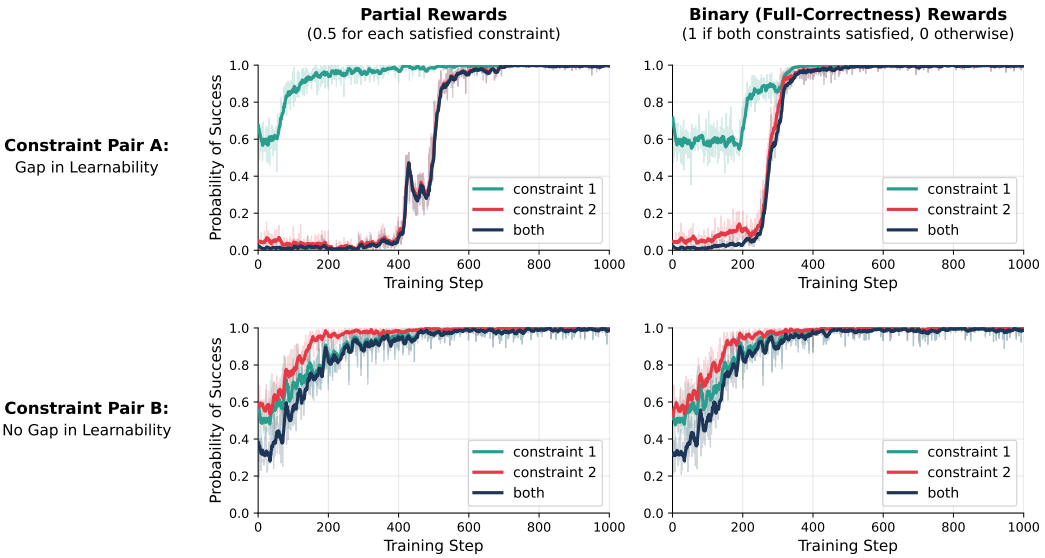

Figure 12: **Rewarding partially correct outputs can impede policy gradient optimization.** This figure presents the results of an experiment analogous to that of Figure 4, where the language model is Llama-3.2-3B-Instruct instead of Qwen3-1.7B. Similarly to Figure 4, when the initial probability of satisfying one constraint is noticeably higher than the probability of satisfying the other, rewarding partial correctness can cause the policy to stall on satisfying only the easier constraint. In this case, rewarding only fully correct outputs can lead to faster learning of both constraints. However, when both constraints are initially satisfied with similar probability, both reward designs tend to work well. See Appendix F for further implementation details.

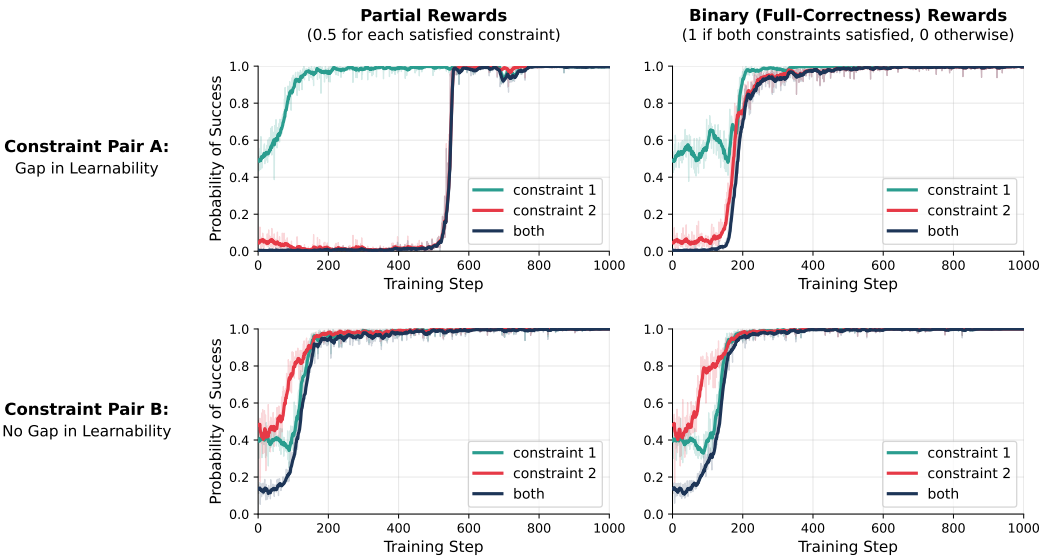

Figure 13: **Rewarding partially correct outputs can impede policy gradient optimization.** This figure presents the results of an experiment analogous to that of Figure 4, where the language model is OLMo-2-1B-Instruct instead of Qwen3-1.7B. Similarly to Figure 4, when the initial probability of satisfying one constraint is noticeably higher than the probability of satisfying the other, rewarding partial correctness can cause the policy to stall on satisfying only the easier constraint. In this case, rewarding only fully correct outputs can lead to faster learning of both constraints. However, when both constraints are initially satisfied with similar probability, both reward designs tend to work well. See Appendix F for further implementation details.

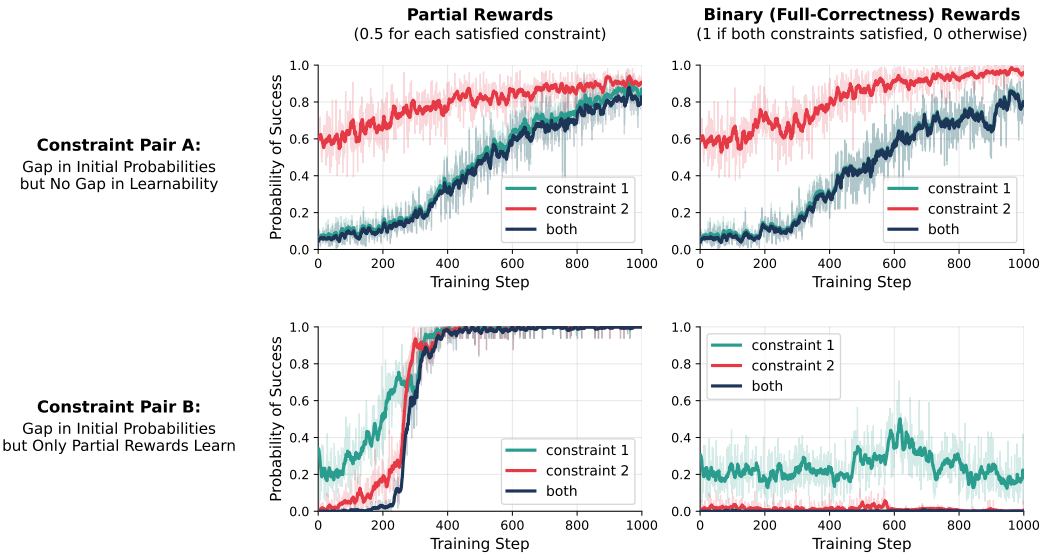

Figure 14: **The probability of initially satisfying a constraint is not the sole factor determining its ease of learnability.** This figure supplements Figure 4 by demonstrating that when one constraint is initially satisfied with substantially higher probability than the other, it is possible for both constraints to be learned at a similar rate under partial rewards. In such cases, under binary rewards, we find that the constraints can either also be learned relatively in tandem (top row) or fail to be learned (bottom row). The latter often occurs when the initial probability of satisfying both constraints is near-zero. These experiments were conducted with the Qwen3-1.7B language model and the same training procedure as in Figure 4, but used different constraint pairs.

Table 2: Numerical results underlying Figures 3 and 8 for the Llama-3.2-3B-Instruct language model. Specifically, reported is the ground truth reward increase (GT Increase) on train and test prompts (mean and standard deviation over three separate runs), along with accuracy values for all considered reward models. We measure accuracy values on examples from the policy gradient training set, with preference labels provided by the ArmoRM ground truth reward model.

Language Model: Llama-3.2-3B-Instruct						
	Dataset: UltraFeedback					
Reward Model (used for training)	GT Increase (train)	GT Increase (test)	Acc	Acc-W	HAcc	HAcc-W
GRM-Llama3.2-3B-rewardmodel-ft	**0.374** $\pm$0.011	**0.369** $\pm$0.002	0.850	0.842	**0.930**	**0.925**
internlm2-1_8b-reward	0.220 $\pm$0.018	0.223 $\pm$0.015	0.828	0.802	0.892	0.859
Llama-3.1-8B-Instruct-RM-RB2	0.268 $\pm$0.005	0.270 $\pm$0.004	0.810	0.791	0.880	0.853
Llama-3-OffsetBias-RM-8B	0.278 $\pm$0.007	0.278 $\pm$0.003	**0.864**	**0.850**	0.926	0.914
llama-3-tulu-2-8b-uf-mean-rm	0.273 $\pm$0.014	0.293 $\pm$0.008	0.792	0.765	0.856	0.842
RM-Gemma-2B	0.329 $\pm$0.013	0.334 $\pm$0.006	0.778	0.764	0.878	0.867
RM-Mistral-7B	0.290 $\pm$0.016	0.293 $\pm$0.008	0.840	0.794	0.890	0.847
Skywork-Reward-V2-Llama-3.1-8B	0.238 $\pm$0.010	0.243 $\pm$0.007	0.796	0.739	0.890	0.874
Skywork-Reward-V2-Llama-3.2-1B	0.286 $\pm$0.001	0.292 $\pm$0.005	0.806	0.779	0.920	0.883
Skywork-Reward-V2-Qwen3-0.6B	0.277 $\pm$0.002	0.281 $\pm$0.007	0.802	0.766	0.904	0.879
Skywork-Reward-V2-Qwen3-1.7B	0.272 $\pm$0.004	0.276 $\pm$0.005	0.826	0.793	0.906	0.872
Skywork-Reward-V2-Qwen3-4B	0.268 $\pm$0.008	0.274 $\pm$0.008	0.844	0.809	0.912	0.885
Skywork-Reward-V2-Qwen3-8B	0.281 $\pm$0.002	0.285 $\pm$0.004	0.834	0.800	0.924	0.894

Table 3: For the Llama-3.2-3B-Instruct language model, this table reports reward model accuracy values computed on examples from RewardBench2, with preference labels provided by the ArmoRM ground truth reward model. These results were used for producing Figure 9.

Language Model: Llama-3.2-3B-Instruct				
	Dataset: RewardBench2			
Reward Model (used for training)	Acc	Acc-W	HAcc	HAcc-W
GRM-Llama3.2-3B-rewardmodel-ft	0.698	0.646	0.753	0.684
internlm2-1_8b-reward	0.467	0.467	0.555	0.571
Llama-3.1-8B-Instruct-RM-RB2	**0.739**	0.714	**0.767**	0.733
Llama-3-OffsetBias-RM-8B	0.704	0.693	0.735	0.721
llama-3-tulu-2-8b-uf-mean-rm	0.602	0.623	0.626	0.633
RM-Gemma-2B	0.387	0.408	0.442	0.439
RM-Mistral-7B	0.653	0.630	0.684	0.650
Skywork-Reward-V2-Llama-3.1-8B	0.720	**0.725**	0.751	**0.742**
Skywork-Reward-V2-Llama-3.2-1B	0.672	0.638	0.718	0.665
Skywork-Reward-V2-Qwen3-0.6B	0.647	0.644	0.681	0.665
Skywork-Reward-V2-Qwen3-1.7B	0.690	0.672	0.721	0.688
Skywork-Reward-V2-Qwen3-4B	0.712	0.700	0.732	0.717
Skywork-Reward-V2-Qwen3-8B	0.720	0.707	0.744	0.724

Table 4: Numerical results underlying Figures 3 and 8 for the Llama-3.2-1B-Instruct language model. Specifically, reported is the ground truth reward increase (GT Increase) on train and test prompts (mean and standard deviation over three separate runs), along with accuracy values for all considered reward models. We measure accuracy values on examples from the policy gradient training set, with preference labels provided by the ArmoRM ground truth reward model.

Language Model: Llama-3.2-1B-Instruct						
	Dataset: UltraFeedback					
Reward Model (used for training)	GT Increase (train)	GT Increase (test)	Acc	Acc-W	HAcc	HAcc-W
GRM-Llama3.2-3B-rewardmodel-ft	**0.457** $\pm$0.015	**0.465** $\pm$0.014	0.850	0.835	**0.896**	**0.895**
internlm2-1_8b-reward	-1.945 $\pm$0.100	-1.960 $\pm$0.105	0.828	0.801	0.850	0.820
Llama-3.1-8B-Instruct-RM-RB2	0.348 $\pm$0.010	0.359 $\pm$0.012	0.810	0.796	0.838	0.822
Llama-3-OffsetBias-RM-8B	0.376 $\pm$0.006	0.385 $\pm$0.010	**0.864**	**0.852**	0.888	0.876
llama-3-tulu-2-8b-uf-mean-rm	0.421 $\pm$0.012	0.433 $\pm$0.021	0.792	0.777	0.826	0.809
RM-Gemma-2B	0.403 $\pm$0.008	0.425 $\pm$0.012	0.778	0.775	0.864	0.874
RM-Mistral-7B	0.402 $\pm$0.011	0.418 $\pm$0.009	0.840	0.819	0.864	0.854
Skywork-Reward-V2-Llama-3.1-8B	0.362 $\pm$0.005	0.364 $\pm$0.009	0.796	0.742	0.822	0.773
Skywork-Reward-V2-Llama-3.2-1B	0.390 $\pm$0.004	0.398 $\pm$0.004	0.806	0.779	0.876	0.837
Skywork-Reward-V2-Qwen3-0.6B	0.384 $\pm$0.008	0.396 $\pm$0.008	0.802	0.773	0.840	0.807
Skywork-Reward-V2-Qwen3-1.7B	0.390 $\pm$0.010	0.400 $\pm$0.010	0.826	0.796	0.852	0.814
Skywork-Reward-V2-Qwen3-4B	0.383 $\pm$0.010	0.388 $\pm$0.014	0.844	0.823	0.862	0.835
Skywork-Reward-V2-Qwen3-8B	0.389 $\pm$0.014	0.393 $\pm$0.005	0.834	0.809	0.850	0.820

Table 5: For the Llama-3.2-1B-Instruct language model, this table reports reward model accuracy values computed on examples from RewardBench2, with preference labels provided by the ArmoRM ground truth reward model. These results were used for producing Figure 9.

Language Model: Llama-3.2-1B-Instruct				
	Dataset: RewardBench2			
Reward Model (used for training)	Acc	Acc-W	HAcc	HAcc-W
GRM-Llama3.2-3B-rewardmodel-ft	0.696	0.642	0.714	0.654
internlm2-1_8b-reward	0.467	0.462	0.523	0.530
Llama-3.1-8B-Instruct-RM-RB2	**0.739**	0.711	**0.746**	0.716
Llama-3-OffsetBias-RM-8B	0.704	0.693	0.710	0.699
llama-3-tulu-2-8b-uf-mean-rm	0.602	0.621	0.612	0.623
RM-Gemma-2B	0.387	0.417	0.427	0.445
RM-Mistral-7B	0.653	0.631	0.663	0.637
Skywork-Reward-V2-Llama-3.1-8B	0.720	**0.727**	0.729	**0.730**
Skywork-Reward-V2-Llama-3.2-1B	0.672	0.634	0.693	0.646
Skywork-Reward-V2-Qwen3-0.6B	0.647	0.649	0.662	0.657
Skywork-Reward-V2-Qwen3-1.7B	0.690	0.682	0.698	0.685
Skywork-Reward-V2-Qwen3-4B	0.712	0.707	0.718	0.709
Skywork-Reward-V2-Qwen3-8B	0.720	0.716	0.726	0.719

Table 6: Numerical results underlying Figures 3 and 8 for the OLMo-2-1B-SFT language model. Specifically, reported is the ground truth reward increase (GT Increase) on train and test prompts (mean and standard deviation over three separate runs), along with accuracy values for all considered reward models. We measure accuracy values on examples from the policy gradient training set, with preference labels provided by the ArmoRM ground truth reward model.

Language Model: OLMo-2-1B-SFT						
	Dataset: UltraFeedback					
Reward Model (used for training)	GT Increase (train)	GT Increase (test)	Acc	Acc-W	HAcc	HAcc-W
GRM-Llama3.2-3B-rewardmodel-ft	0.218 $\pm$0.011	0.198 $\pm$0.014	0.850	0.827	**0.880**	0.861
internlm2-1_8b-reward	**0.281** $\pm$0.006	**0.251** $\pm$0.011	0.828	0.793	0.854	0.809
Llama-3.1-8B-Instruct-RM-RB2	0.212 $\pm$0.004	0.191 $\pm$0.006	0.810	0.778	0.828	0.789
Llama-3-OffsetBias-RM-8B	0.236 $\pm$0.009	0.209 $\pm$0.009	**0.864**	**0.856**	0.878	**0.863**
llama-3-tulu-2-8b-uf-mean-rm	0.260 $\pm$0.006	0.234 $\pm$0.002	0.792	0.759	0.814	0.782
RM-Gemma-2B	0.275 $\pm$0.004	0.242 $\pm$0.006	0.778	0.760	0.814	0.800
RM-Mistral-7B	0.266 $\pm$0.006	0.233 $\pm$0.004	0.840	0.798	0.846	0.809
Skywork-Reward-V2-Llama-3.1-8B	0.213 $\pm$0.002	0.192 $\pm$0.007	0.796	0.748	0.818	0.769
Skywork-Reward-V2-Llama-3.2-1B	0.234 $\pm$0.006	0.212 $\pm$0.004	0.806	0.763	0.826	0.776
Skywork-Reward-V2-Qwen3-0.6B	0.238 $\pm$0.001	0.216 $\pm$0.005	0.802	0.757	0.824	0.771
Skywork-Reward-V2-Qwen3-1.7B	0.241 $\pm$0.001	0.219 $\pm$0.009	0.826	0.784	0.840	0.794
Skywork-Reward-V2-Qwen3-4B	0.225 $\pm$0.009	0.206 $\pm$0.002	0.844	0.807	0.856	0.814
Skywork-Reward-V2-Qwen3-8B	0.226 $\pm$0.008	0.204 $\pm$0.006	0.834	0.794	0.846	0.802

Table 7: For the OLMo-2-1B-SFT language model, this table reports reward model accuracy values computed on examples from RewardBench2, with preference labels provided by the ArmoRM ground truth reward model. These results were used for producing Figure 9.

Language Model: OLMo-2-1B-SFT				
	Dataset: RewardBench2			
Reward Model (used for training)	Acc	Acc-W	HAcc	HAcc-W
GRM-Llama3.2-3B-rewardmodel-ft	0.691	0.651	0.708	0.658
internlm2-1_8b-reward	0.467	0.465	0.526	0.527
Llama-3.1-8B-Instruct-RM-RB2	**0.739**	0.715	**0.750**	0.721
Llama-3-OffsetBias-RM-8B	0.704	0.703	0.708	0.705
llama-3-tulu-2-8b-uf-mean-rm	0.602	0.622	0.615	0.624
RM-Gemma-2B	0.387	0.414	0.402	0.418
RM-Mistral-7B	0.653	0.645	0.659	0.647
Skywork-Reward-V2-Llama-3.1-8B	0.720	0.735	0.733	0.742
Skywork-Reward-V2-Llama-3.2-1B	0.672	0.653	0.683	0.658
Skywork-Reward-V2-Qwen3-0.6B	0.647	0.667	0.666	0.675
Skywork-Reward-V2-Qwen3-1.7B	0.690	0.703	0.704	0.708
Skywork-Reward-V2-Qwen3-4B	0.712	0.723	0.727	0.730
Skywork-Reward-V2-Qwen3-8B	0.720	**0.738**	0.733	**0.744**

Table 8: Numerical results underlying Figures 3 and 8 for the Qwen3-1.7B-Base language model. Specifically, reported is the ground truth reward increase (GT Increase) on train and test prompts (mean and standard deviation over three separate runs), along with accuracy values for all considered reward models. We measure accuracy values on examples from the policy gradient training set, with preference labels provided by the ArmoRM ground truth reward model.

Language Model: Qwen3-1.7B-Base						
	Dataset: UltraFeedback					
Reward Model (used for training)	GT Increase (train)	GT Increase (test)	Acc	Acc-W	HAcc	HAcc-W
GRM-Llama3.2-3B-rewardmodel-ft	**2.257** ± 0.009	**2.264** ± 0.006	0.850	0.859	0.850	0.859
internlm2-1_8b-reward	-0.969 ± 0.038	-0.964 ± 0.052	0.828	0.807	0.832	0.807
Llama-3.1-8B-Instruct-RM-RB2	2.113 ± 0.007	2.106 ± 0.006	0.810	0.817	0.810	0.817
Llama-3-OffsetBias-RM-8B	2.187 ± 0.050	2.197 ± 0.056	**0.864**	**0.865**	**0.866**	**0.868**
llama-3-tulu-2-8b-uf-mean-rm	2.206 ± 0.006	2.207 ± 0.004	0.792	0.794	0.794	0.795
RM-Gemma-2B	2.011 ± 0.074	2.026 ± 0.065	0.778	0.790	0.778	0.790
RM-Mistral-7B	2.207 ± 0.006	2.205 ± 0.004	0.840	0.843	0.840	0.843
Skywork-Reward-V2-Llama-3.1-8B	2.179 ± 0.008	2.177 ± 0.007	0.796	0.756	0.796	0.756
Skywork-Reward-V2-Llama-3.2-1B	2.213 ± 0.004	2.216 ± 0.004	0.806	0.793	0.806	0.793
Skywork-Reward-V2-Qwen3-0.6B	2.200 ± 0.007	2.213 ± 0.008	0.802	0.768	0.802	0.768
Skywork-Reward-V2-Qwen3-1.7B	2.203 ± 0.009	2.207 ± 0.005	0.826	0.803	0.826	0.803
Skywork-Reward-V2-Qwen3-4B	2.201 ± 0.008	2.199 ± 0.003	0.844	0.821	0.844	0.821
Skywork-Reward-V2-Qwen3-8B	2.214 ± 0.008	2.216 ± 0.008	0.834	0.814	0.834	0.814

Table 9: For the Qwen3-1.7B-Base language model, this table reports reward model accuracy values computed on examples from RewardBench2, with preference labels provided by the ArmoRM ground truth reward model. These results were used for producing Figure 9.

Language Model: Qwen3-1.7B-Base				
	Dataset: RewardBench2			
Reward Model (used for training)	Acc	Acc-W	HAcc	HAcc-W
GRM-Llama3.2-3B-rewardmodel-ft	0.691	0.648	0.691	0.648
internlm2-1_8b-reward	0.467	0.470	0.469	0.470
Llama-3.1-8B-Instruct-RM-RB2	**0.739**	0.714	**0.739**	0.714
Llama-3-OffsetBias-RM-8B	0.704	0.697	0.705	0.697
llama-3-tulu-2-8b-uf-mean-rm	0.602	0.652	0.602	0.652
RM-Gemma-2B	0.387	0.421	0.387	0.421
RM-Mistral-7B	0.653	0.652	0.653	0.652
Skywork-Reward-V2-Llama-3.1-8B	0.720	0.729	0.720	0.729
Skywork-Reward-V2-Llama-3.2-1B	0.672	0.655	0.672	0.655
Skywork-Reward-V2-Qwen3-0.6B	0.647	0.669	0.647	0.669
Skywork-Reward-V2-Qwen3-1.7B	0.690	0.700	0.690	0.700
Skywork-Reward-V2-Qwen3-4B	0.712	0.723	0.712	0.723
Skywork-Reward-V2-Qwen3-8B	0.720	**0.732**	0.720	**0.732**

Table 10: Numerical results underlying Figure 10 for the Llama-3.2-1B-Instruct language model trained on the WildChat-IF-On-Policy-8B dataset. Specifically, reported is the ground truth reward increase (GT Increase) on train and test prompts (mean and standard deviation over three separate runs), along with accuracy values for all considered reward models. We measure accuracy values on examples from the policy gradient training set, with preference labels provided by the ArmoRM ground truth reward model.

Language Model: Llama-3.2-1B-Instruct						
	Dataset: WildChat-IF-On-Policy-8B					
Reward Model (used for training)	GT Increase (train)	GT Increase (test)	Acc	Acc-W	HAcc	HAcc-W
GRM-Llama3.2-3B-rewardmodel-ft	0.361 ± 0.033	0.338 ± 0.035	0.838	0.841	0.852	0.847
internlm2-1_8b-reward	-0.689 ± 0.924	-0.693 ± 0.915	0.740	0.732	0.762	0.748
Llama-3.1-8B-Instruct-RM-RB2	0.297 ± 0.049	0.277 ± 0.049	0.848	0.826	0.858	0.832
Llama-3-OffsetBias-RM-8B	0.360 ± 0.049	0.329 ± 0.035	0.846	0.824	0.850	0.826
llama-3-tulu-2-8b-uf-mean-rm	0.380 ± 0.050	0.350 ± 0.041	0.782	0.777	0.792	0.779
RM-Gemma-2B	$\mathbf{0.391} \pm 0.057$	$\mathbf{0.369} \pm 0.051$	0.738	0.735	0.776	0.773
RM-Mistral-7B	0.356 ± 0.045	0.324 ± 0.041	0.838	0.827	0.848	0.834
Skywork-Reward-V2-Llama-3.1-8B	0.306 ± 0.054	0.278 ± 0.048	0.856	0.840	0.868	0.842
Skywork-Reward-V2-Llama-3.2-1B	0.329 ± 0.054	0.318 ± 0.041	0.854	0.845	0.872	0.863
Skywork-Reward-V2-Qwen3-0.6B	0.321 ± 0.064	0.302 ± 0.051	0.818	0.832	0.836	0.850
Skywork-Reward-V2-Qwen3-1.7B	0.329 ± 0.058	0.299 ± 0.044	0.844	0.836	0.850	0.838
Skywork-Reward-V2-Qwen3-4B	0.330 ± 0.039	0.297 ± 0.035	**0.876**	0.864	0.878	0.864
Skywork-Reward-V2-Qwen3-8B	0.344 ± 0.042	0.315 ± 0.036	0.874	**0.869**	**0.880**	**0.870**

Table 11: Numerical results underlying Figure 10 for the Llama-3.2-1B-Instruct language model. Specifically, reported is the ground truth reward increase (GT Increase) on train and test prompts (mean and standard deviation over three separate runs), along with accuracy values for all considered reward models. We measure accuracy values on examples from the policy gradient training set, with preference labels provided by the Skywork-Reward-V2-Llama-3.1-8B ground truth reward model.

Language Model: Llama-3.2-1B-Instruct						
	Dataset: UltraFeedback (Ground Truth: Skywork-Reward-V2-Llama-3.1-8B)					
Reward Model (used for training)	GT Increase (train)	GT Increase (test)	Acc	Acc-W	HAcc	HAcc-W
ArmoRM-Llama3-8B-v0-1	0.420 ± 0.105	0.389 ± 0.107	0.810	0.779	0.834	0.811
GRM-Llama3.2-3B-rewardmodel-ft	0.466 ± 0.084	0.462 ± 0.103	0.808	0.805	0.842	0.837
internlm2-1_8b-reward	-1.768 ± 0.221	-1.824 ± 0.247	0.774	0.721	0.802	0.745
Llama-3.1-8B-Instruct-RM-RB2	0.416 ± 0.082	0.406 ± 0.078	0.862	0.804	0.878	0.822
Llama-3-OffsetBias-RM-8B	0.440 ± 0.062	0.417 ± 0.072	0.826	0.814	0.852	0.836
llama-3-tulu-2-8b-uf-mean-rm	0.414 ± 0.146	0.397 ± 0.152	0.810	0.767	0.826	0.776
RM-Gemma-2B	0.330 ± 0.176	0.338 ± 0.178	0.740	0.703	0.832	0.785
RM-Mistral-7B	0.471 ± 0.116	0.453 ± 0.119	0.824	0.801	0.848	0.816
Skywork-Reward-V2-Llama-3.2-1B	$\mathbf{0.533} \pm 0.095$	$\mathbf{0.522} \pm 0.106$	0.878	0.886	**0.920**	**0.914**
Skywork-Reward-V2-Qwen3-0.6B	0.503 ± 0.101	0.498 ± 0.101	0.852	0.850	0.898	0.890
Skywork-Reward-V2-Qwen3-1.7B	0.510 ± 0.101	0.489 ± 0.097	0.856	0.864	0.904	0.907
Skywork-Reward-V2-Qwen3-4B	0.503 ± 0.078	0.479 ± 0.086	0.894	0.881	0.912	0.899
Skywork-Reward-V2-Qwen3-8B	0.504 ± 0.080	0.488 ± 0.078	**0.896**	**0.893**	0.908	0.908

Table 12: Numerical results underlying Figure 11 for the Llama-3.2-3B-Instruct language model. Specifically, for each reward model, the table reports the win-rate of the language model after policy gradient against the initial language model, as judged by a frontier GPT model, along with the accuracy values for the reward model. We measure accuracy values on examples from the policy gradient training set, with preference labels provided by the ArmoRM ground truth reward model.

Language Model: Llama-3.2-3B-Instruct					
	Dataset: UltraFeedback				
Reward Model (used for training)	Win-Rate	Acc	Acc-W	HAcc	HAcc-W
GRM-Llama3.2-3B-rewardmodel-ft	0.738	0.850	0.842	**0.930**	**0.925**
internlm2-1_8b-reward	0.704	0.828	0.802	0.892	0.859
Llama-3.1-8B-Instruct-RM-RB2	0.715	0.810	0.791	0.880	0.853
Llama-3-OffsetBias-RM-8B	0.646	**0.864**	**0.850**	0.926	0.914
llama-3-tulu-2-8b-uf-mean-rm	0.702	0.792	0.765	0.856	0.842
RM-Gemma-2B	**0.769**	0.778	0.764	0.878	0.867
RM-Mistral-7B	0.726	0.840	0.794	0.890	0.847
Skywork-Reward-V2-Llama-3.1-8B	0.677	0.796	0.739	0.890	0.874
Skywork-Reward-V2-Llama-3.2-1B	0.720	0.806	0.779	0.920	0.883
Skywork-Reward-V2-Qwen3-0.6B	0.735	0.802	0.766	0.904	0.879
Skywork-Reward-V2-Qwen3-1.7B	0.713	0.826	0.793	0.906	0.872
Skywork-Reward-V2-Qwen3-4B	0.724	0.844	0.809	0.912	0.885
Skywork-Reward-V2-Qwen3-8B	0.702	0.834	0.800	0.924	0.894

Table 13: Numerical results underlying Figure 11 for the Llama-3.2-1B-Instruct language model. Specifically, for each reward model, the table reports the win-rate of the language model after policy gradient against the initial language model, as judged by a frontier GPT model, along with the accuracy values for the reward model. We measure accuracy values on examples from the policy gradient training set, with preference labels provided by the ArmoRM ground truth reward model.

Language Model: Llama-3.2-1B-Instruct					
	Dataset: UltraFeedback				
Reward Model (used for training)	Win-Rate	Acc	Acc-W	HAcc	HAcc-W
GRM-Llama3.2-3B-rewardmodel-ft	0.737	0.850	0.835	**0.896**	**0.895**
internlm2-1_8b-reward	0.009	0.828	0.801	0.850	0.820
Llama-3.1-8B-Instruct-RM-RB2	0.701	0.810	0.796	0.838	0.822
Llama-3-OffsetBias-RM-8B	0.708	**0.864**	**0.852**	0.888	0.876
llama-3-tulu-2-8b-uf-mean-rm	0.722	0.792	0.777	0.826	0.809
RM-Gemma-2B	0.733	0.778	0.775	0.864	0.874
RM-Mistral-7B	**0.755**	0.840	0.819	0.864	0.854
Skywork-Reward-V2-Llama-3.1-8B	0.695	0.796	0.742	0.822	0.773
Skywork-Reward-V2-Llama-3.2-1B	0.741	0.806	0.779	0.876	0.837
Skywork-Reward-V2-Qwen3-0.6B	0.725	0.802	0.773	0.840	0.807
Skywork-Reward-V2-Qwen3-1.7B	0.704	0.826	0.796	0.852	0.814
Skywork-Reward-V2-Qwen3-4B	0.712	0.844	0.823	0.862	0.835
Skywork-Reward-V2-Qwen3-8B	0.716	0.834	0.809	0.850	0.820

Table 14: Numerical results underlying Figure 11 for the OLMo-2-1B-SFT language model. Specifically, for each reward model, the table reports the win-rate of the language model after policy gradient against the initial language model, as judged by a frontier GPT model, along with the accuracy values for the reward model. We measure accuracy values on examples from the policy gradient training set, with preference labels provided by the ArmoRM ground truth reward model.

Language Model: OLMo-2-1B-SFT					
	Dataset: UltraFeedback				
Reward Model (used for training)	Win-Rate	Acc	Acc-W	HAcc	HAcc-W
GRM-Llama3.2-3B-rewardmodel-ft	0.616	0.850	0.827	**0.880**	0.861
internlm2-1_8b-reward	0.628	0.828	0.793	0.854	0.809
Llama-3.1-8B-Instruct-RM-RB2	0.607	0.810	0.778	0.828	0.789
Llama-3-OffsetBias-RM-8B	0.611	**0.864**	**0.856**	0.878	**0.863**
llama-3-tulu-2-8b-uf-mean-rm	0.627	0.792	0.759	0.814	0.782
RM-Gemma-2B	0.642	0.778	0.760	0.814	0.800
RM-Mistral-7B	**0.657**	0.840	0.798	0.846	0.809
Skywork-Reward-V2-Llama-3.1-8B	0.610	0.796	0.748	0.818	0.769
Skywork-Reward-V2-Llama-3.2-1B	0.611	0.806	0.763	0.826	0.776
Skywork-Reward-V2-Qwen3-0.6B	0.634	0.802	0.757	0.824	0.771
Skywork-Reward-V2-Qwen3-1.7B	0.624	0.826	0.784	0.840	0.794
Skywork-Reward-V2-Qwen3-4B	0.610	0.844	0.807	0.856	0.814
Skywork-Reward-V2-Qwen3-8B	0.586	0.834	0.794	0.846	0.802

Table 15: Numerical results underlying Figure 11 for the Qwen3-1.7B-Base language model. Specifically, for each reward model, the table reports the win-rate of the language model after policy gradient against the initial language model, as judged by a frontier GPT model, along with the accuracy values for the reward model. We measure accuracy values on examples from the policy gradient training set, with preference labels provided by the ArmoRM ground truth reward model.

Language Model: Qwen3-1.7B-Base					
	Dataset: UltraFeedback				
Reward Model (used for training)	Win-Rate	Acc	Acc-W	HAcc	HAcc-W
GRM-Llama3.2-3B-rewardmodel-ft	0.957	0.850	0.859	0.850	0.859
internlm2-1_8b-reward	0.191	0.828	0.807	0.832	0.807
Llama-3.1-8B-Instruct-RM-RB2	0.963	0.810	0.817	0.810	0.817
Llama-3-OffsetBias-RM-8B	0.962	**0.864**	**0.865**	**0.866**	**0.868**
llama-3-tulu-2-8b-uf-mean-rm	0.936	0.792	0.794	0.794	0.795
RM-Gemma-2B	0.919	0.778	0.790	0.778	0.790
RM-Mistral-7B	0.955	0.840	0.843	0.840	0.843
Skywork-Reward-V2-Llama-3.1-8B	0.957	0.796	0.756	0.796	0.756
Skywork-Reward-V2-Llama-3.2-1B	0.961	0.806	0.793	0.806	0.793
Skywork-Reward-V2-Qwen3-0.6B	0.951	0.802	0.768	0.802	0.768
Skywork-Reward-V2-Qwen3-1.7B	0.946	0.826	0.803	0.826	0.803
Skywork-Reward-V2-Qwen3-4B	0.959	0.844	0.821	0.844	0.821
Skywork-Reward-V2-Qwen3-8B	**0.965**	0.834	0.814	0.834	0.814

Table 16: This table specifies the IFBench constraint pairs used in each of the experiments corresponding to Figures 4, 12, 13, and 14. We include the constraint name according to the IFBench taxonomy (*e.g.*, `words:start_verb`), followed by the instruction that is appended to the prompt.

Language Model	Constraint Pair	Constraint 1	Constraint 2
Qwen3-1.7B Figure 4	Pair A	`words:start_verb` The response must start with a verb.	`words:alphabet` Each word must start with the next letter of the alphabet, looping back to 'A' after 'Z'.
	Pair B	`count:numbers` Include exactly 13 numbers in the response.	`format:newline` Write each word on a new line.
Llama-3.2-3B-Instruct Figure 12	Pair A	`count:pronouns` The response should include at least 5 pronouns.	`words:last_first` The last word of each sentence must become the first word of the next sentence.
	Pair B	`words:start_verb` The response must start with a verb.	`count:pronouns` The response should include at least 5 pronouns.
OLMo-2-1B-Instruct Figure 13	Pair A	`count:pronouns` The response should include at least 5 pronouns.	`sentence:increment` Each sentence must contain exactly 5 more words than the previous one.
	Pair B	`format:quotes` Include quotes within quotes within quotes, at least 3 levels deep, alternating between double quotes and single quotes.	`words:repeats` The response should not repeat any word more than 10 times.
Qwen3-1.7B Figure 14	Pair A	`count:punctuation` Use every standard punctuation mark at least once, including semicolons, colons, and the interrobang (?!).	`count:person_names` Mention at least 7 different person names in the response, from this list of person names: Emma, Liam, Sophia...
	Pair B	`words:start_verb` The response must start with a verb.	`custom:multiples` Count from 10 to 50 but only print multiples of 7.