Activation Steering with a Feedback Controller

Dung V. Nguyen[1*], **Hieu M. Vu**[3*], **Nhi Y. Pham**[2*], **Lei Zhang**[1†], **Tan M. Nguyen**[1*]

[1]Department of Mathematics, National University of Singapore
[2]VinUniversity
[3]Independent
dungnv@u.nus.edu, vmhieu17@gmail.com, nhiiyennphamm@gmail.com,
{matzhlei,tanmn}@nus.edu.sg

Abstract

Controlling the behaviors of large language models (LLM) is fundamental to their safety alignment and reliable deployment. However, existing steering methods are primarily driven by empirical insights and lack theoretical performance guarantees. In this work, we develop a control-theoretic foundation for activation steering by showing that popular steering methods correspond to the proportional (P) controllers, with the steering vector serving as the feedback signal. Building on this finding, we propose Proportional-Integral-Derivative (PID) Steering, a principled framework that leverages the full PID controller for activation steering in LLMs. The proportional (P) term aligns activations with target semantic directions, the integral (I) term accumulates errors to enforce persistent corrections across layers, and the derivative (D) term mitigates overshoot by counteracting rapid activation changes. This closed-loop design yields interpretable error dynamics and connects activation steering to classical stability guarantees in control theory. Moreover, PID Steering is lightweight, modular, and readily integrates with state-of-the-art steering methods. Extensive experiments across multiple LLM families and benchmarks demonstrate that PID Steering consistently outperforms existing approaches, achieving more robust and reliable behavioral control.

Introduction

Large language models (LLMs) have demonstrated remarkable capabilities across diverse domains, yet ensuring that their outputs align with desired behaviors remains a central challenge (Dang et al. 2025; Sclar et al. 2023; Kotha, Springer, and Raghunathan 2023; Luo et al. 2025; Houlsby et al. 2019). Common post-training approaches (Wei et al. 2021; Ouyang et al. 2022) have proven effective for improving alignment. However, these techniques demand substantial computational resources (Houlsby et al. 2019) and require weight updates with new training data, which can unintentionally degrade fluency or performance on unrelated tasks (Templeton et al. 2024; Kotha, Springer, and Raghunathan 2023; Luo et al. 2025).

An increasingly popular alternative is *activation steering*, which modifies a model's internal activations directly at inference time, avoiding costly retraining (Vu and Nguyen 2025; Li et al. 2024; Turner et al. 2023, 2024; Lee et al. 2024; Rimsky et al. 2024; Rodriguez et al. 2025). This approach has been employed both to probe internal representations (Geiger et al. 2024; von Rütte et al. 2024; Vu and Nguyen 2025) and to enable fine-grained behavioral control (Vu and Nguyen 2025; Rodriguez et al. 2025; Turner et al. 2024; Zou et al. 2023a; Rimsky et al. 2024; Li et al. 2024). Recent work demonstrates that steering along carefully chosen low-dimensional directions can effectively alter model behavior (Turner et al. 2024; Rimsky et al. 2024; Arditi et al. 2024; Zou et al. 2023a; Vu and Nguyen 2025), highlighting its potential as a lightweight yet powerful alignment strategy.

Steering through the Lens of Dynamical Systems. Recent methods leverage the geometric structure of the activation space (Marks and Tegmark 2024; Park, Choe, and Veitch 2024) using linear algebraic techniques (Turner et al. 2024; Zou et al. 2023a; Rimsky et al. 2024; Arditi et al. 2024; Vu and Nguyen 2025) to compute the steering vectors. While effective, these works oversimplify the complex, dynamic behavior arising from the auto-regressive nature of LLMs. When viewed through this dynamical lens, activation steering can be interpreted as guiding the model's trajectory through activation space, from a region encoding one concept to another, analogous to steering a dynamical system from one state to a desired target state.

Contribution. Building on the aforementioned dynamical system insight, our work departs from the prevailing algebraic framing and instead adopts a control-theoretic perspective on activation steering. Although recent studies (Soatto et al. 2023; Kong et al. 2024; Luo et al. 2023) have begun exploring this direction, their focus has primarily remained at the level of the token-level generation proceses, treating high-level behaviors as control signals. In contrast, we take into account the internal mechanisms of LLMs by modeling the layer-wise construction of feature directions (Bricken et al. 2023; Park, Choe, and Veitch 2024) as a dynamical system. These feature directions are then used as steering vectors (Turner et al. 2024; Zou et al. 2023a; Rimsky et al. 2024; Arditi et al. 2024;

[*]These authors contributed equally.

[†]Corresponding authors.

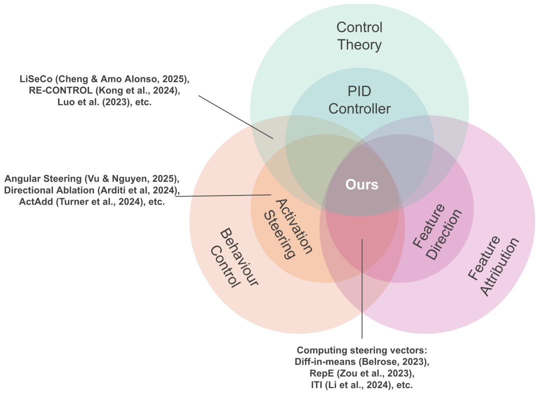

Control
Theory
PID
Controller
LiSeCo (Cheng & Amo Alonso, 2025),
RE-CONTROL (Kong et al., 2024),
Luo et al. (2023), etc.
Angular Steering (Vu & Nguyen, 2025),
Directional Ablation (Arditi et al, 2024),
ActAdd (Turner et al., 2024), etc.
Ours
Activation
Steering
Behaviour
Control
Feature
Direction
Feature
Attribution
Computing steering vectors:
Diff-in-means (Belrose, 2023),
RepE (Zou et al., 2023),
ITI (Li et al., 2024), etc.

Figure 1: Our paper connects LLM Behavior Control, Feature Attribution for LLM and Control Theory. Specifically, we apply a PID-Controller to compute the steering vector for activation steering.

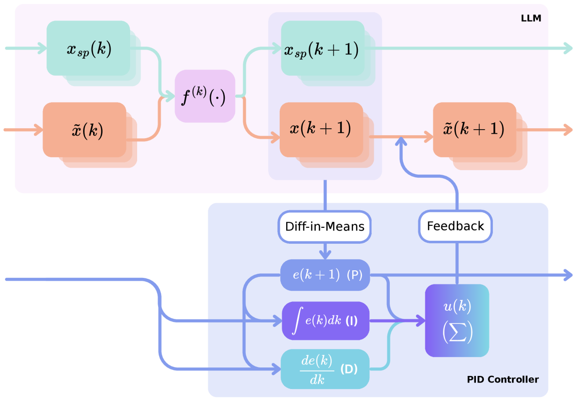


Figure 2: PID Steering: To compute the steering vector $u(k)$: a PID controller is applied at every layer $f^{(k)}(\cdot)$, using the diff-in-means between 2 contrastive data $x_{sp}(k)$ and $x(k)$ as the error signal $e(k)$.

Vu and Nguyen 2025). Specifically, we show that existing steering methods relying on difference-of-means feature directions (Rimsky et al. 2024), including Activation Addition (ActAdd) (Turner et al. 2024), Directional Ablation (Arditi et al. 2024), and Mean Activation Transport (Mean-AcT) (Rodriguez et al. 2025), can be interpreted as instances of a *proportional (P) controller*, thus suffering from the steady-state error due to the disturbance to the state of the system (Åström and Hägglund 1995a). This new perspective enables the application of principled control-theoretic strategies for extracting effective feature directions and computing steering vectors, thereby offering stronger robustness and performance guarantees for activation steering methods. An overview of our approach is shown in Fig. 1 and 2. In this paper, we use the terms feature direction and steering vector interchangeably, noting that steering vectors represent a practical application of feature directions in activation steering. Our contribution is three-fold:

1. **Control-Theoretic Formulation for Feature Direction:** We develop a new control-theoretic framework for constructing feature directions/steering vectors along the layers of an LLM.
2. **PID-Based Steering:** We propose the novel *Proportional-Integral-Derivative (PID) Steering*, a control-theoretic framework for computing feature directions using a PID controller to reduce the steady-state error inherent in existing activation steering methods (see Fig. 2).
3. **Unified Theoretical Framework:** We demonstrate that common activation steering methods correspond to proportional (P) controllers. This connection enables a theoretical analysis that highlights PID Steering's advantages in reducing steady-state error and oscillations

We comprehensively validate our PID Steering across diverse *modalities* (text and image), *downstream applications* (toxicity mitigation, jailbreaking attack, and image style control), *steering paradigms* (ActAdd, Mean-AcT, and Angular Steering (Vu and Nguyen 2025)), *model families* (Qwen2.5 (Yang et al. 2024), Gemma2 (Gemma Team et al. 2024), Llama3 (Llama Team 2024), SDXL-Lightning (Lin, Wang, and Yang 2024), and Flux (Labs 2024)), and *model scales* (3B-14B for language models and 3.5B-12B for diffusion models).

Organization. We organize the paper as follows: Section reviews background; Section links activation steering to P control and introduces PID Steering; Section presents its theoretical analysis; Section provides empirical validation; Appendix discusses related work; and Section concludes. Proofs, derivations, and additional experiments are in the Appendix.

Background

Transformers Decoder-only transformers take an input token sequence $\boldsymbol{q} = [q_1, \dots, q_n]$ and map it to initial embeddings $\boldsymbol{x}(1) = [\boldsymbol{x}_1(1), \dots, \boldsymbol{x}_n(1)]^\top = \text{Embed}(\boldsymbol{q})$. The embeddings are then propagated through K layers. At each layer k, the residual activation $\boldsymbol{x}_i(k)$ for token p_i is updated by self-attention and an MLP block, with normalization applied before (and sometimes after) these modules:

$$\boldsymbol{x}_{i,\text{post-attn}}(k) = \boldsymbol{x}_i(k) + \text{SelfAttn}^{(k)}(\text{Norm}(\boldsymbol{x}_i(k)))$$

$$\boldsymbol{x}_i(k+1) = \boldsymbol{x}_{i,\text{post-attn}}(k) + \text{MLP}^{(k)}(\text{Norm}(\boldsymbol{x}_{i,\text{post-attn}}(k))).$$

In this paper, for notational brevity, we summarize the layered processing above as $\boldsymbol{x}_i(k+1) = f_i^{(k)}(\boldsymbol{x}(k))$, $i = 1, \dots, n$, where $f_i^{(k)}$ encapsulates both the Self-Attention mechanism and Multi-Layer Perceptron at layer k. Finally, the output activations from the last layer, $\boldsymbol{x}_i(L+1)$, are decoded over the model's vocabulary to get the next token $y_i = \text{Decode}(\boldsymbol{x}_i(L+1))$ for subsequent generation.

Activation Steering Features such as behaviors or concepts are hypothesized to align with (approximately) orthogonal directions in activation space (Park, Choe, and Veitch 2024; Bereska and Gavves 2024; Elhage et al. 2022). Activation steering leverages this by modifying hidden states at inference to amplify or suppress specific features (Bayat et al. 2025; Konen et al. 2024; Li et al. 2024; Marks et al. 2025; Templeton et al. 2024). Recent approaches operationalize this idea by constructing feature directions, which act as *steering vectors* $\boldsymbol{r}$ for adjusting hidden states. These steering vectors are computed as layerwise differences in mean activations between datasets with contrasting concepts (e.g., harmful vs. harmless), a *difference-in-means* approach (Rimsky et al. 2024), shown to effectively isolate salient feature directions (Turner et al. 2023, 2024; Arditi et al. 2024).

Applying the Steering Vectors Two popular activation steering approaches that use steering vectors are: *Activation Addition* (Turner et al. 2024), and *Directional Ablation* (Arditi et al. 2024). Both methods modify the token activation $\boldsymbol{x}(k)$ using the steering vector $\boldsymbol{r}(k)$ at layer k such that the activation expresses the target concept or behavior. By setting $\boldsymbol{x}(1, \boldsymbol{q}) = \text{Embed}(\boldsymbol{q})$ and $\boldsymbol{r}(1) = 0$, these methods apply the steering vectors $\boldsymbol{r}(k)$ to the activation $\boldsymbol{x}(k)$, $k = [K]$, at each layer via a steering function ρ_{steer} as follows:

$$\boldsymbol{x}(k-1, \boldsymbol{q}) = \rho_{\text{steer}}(\boldsymbol{x}(k-1, \boldsymbol{q}), \boldsymbol{r}(k-1)), \quad \text{for } \boldsymbol{q} \in \mathcal{D}_{\text{source}} \tag{1}$$

$$\boldsymbol{x}(k, \boldsymbol{q}) = f^{(k)}(\boldsymbol{x}(k-1, \boldsymbol{q})), \quad \text{for } \boldsymbol{q} \in \mathcal{D}_{\text{source}} \cup \mathcal{D}_{\text{target}}. \tag{2}$$

We discuss here the details on how to design the steering function ρ_{steer} for each method.

Activation Addition (ActAdd). ActAdd and sets $\rho_{\text{steer}}(\boldsymbol{x}(k), \boldsymbol{r}(k)) = \boldsymbol{x}(k) + \alpha \boldsymbol{r}(k)$, where the coefficient α controls the strength of the effect.

Directional Ablation (DirAblate). DirAblate removes the feature by projecting the token activation onto the orthogonal complement, $\rho_{\text{steer}}(\boldsymbol{x}(k), \boldsymbol{r}(k)) = \boldsymbol{x}(k) - \boldsymbol{r}(k)\, \boldsymbol{r}(k)^\top\, \boldsymbol{x}(k)$.

Computing the Steering Vectors **Non-sequential Mapping.** Let us use the jailbreaking task as an example. In this task, we apply activation steering to force the LLM to respond to harmful prompts (Vu and Nguyen 2025; Arditi et al. 2024). In order to compute the steering vectors, i.e., refusal direction, for each layer $k \in [K]$ and post-instruction token position $i \in I$, we calculate the mean activation $\boldsymbol{\mu}_{i,\text{target}}(k)$ for harmless prompts from $\mathcal{D}^{(\text{train})}_{\text{target}}$ and $\boldsymbol{\mu}_{i,\text{source}}(k)$ for harmful prompts from $\mathcal{D}^{(\text{train})}_{\text{source}}$:

$$\boldsymbol{\mu}_{i,\text{target}}(k) = \frac{1}{|\mathcal{D}^{(\text{train})}_{\text{target}}|} \sum_{\boldsymbol{q} \in \mathcal{D}^{(\text{train})}_{\text{target}}} \boldsymbol{x}_i(k, \boldsymbol{q}), \tag{3}$$

$$\boldsymbol{\mu}_{i,\text{source}}(k) = \frac{1}{|\mathcal{D}^{(\text{train})}_{\text{source}}|} \sum_{\boldsymbol{q} \in \mathcal{D}^{(\text{train})}_{\text{source}}} \boldsymbol{x}_i(k, \boldsymbol{q}). \tag{4}$$

We then compute the difference-in-means vectors, $\boldsymbol{r}_i(k) = \boldsymbol{\mu}_{i,\text{target}}(k) - \boldsymbol{\mu}_{i,\text{source}}(k)$, and use them as steering vectors. Optionally, among the difference-in-means vector $\boldsymbol{r}_i(k)$ for each post-instruction token position $i \in I$ at layer k, we can select the single most effective vector $\boldsymbol{r}(k) = \text{Select}(\{\boldsymbol{r}_i(k)\}_{i \in I})$ from this set by evaluating each candidate vector over validation sets $\mathcal{D}^{(\text{val})}_{\text{source}}$ and $\mathcal{D}^{(\text{val})}_{\text{target}}$.

Sequential Mapping. A non-sequential mapping neglects the causal dependency across activations, where outputs from one layer are passed to the next, i.e., $\boldsymbol{x}_i(k+1) = f_i^{(k)}(\boldsymbol{x}(k))$. Consequently, any intervention applied at one layer must be accounted for before introducing an intervention at the subsequent layer. To capture this causal structure, *Mean Activation Transport (Mean-AcT)* in (Rodriguez et al. 2025) estimates the steering vectors incrementally at each layer as follows:

$$\boldsymbol{x}_i(k-1, \boldsymbol{q}) = \rho_{\text{steer}}(\boldsymbol{x}_i(k-1, \boldsymbol{q}), \boldsymbol{r}(k-1)), \quad \text{for } \boldsymbol{q} \in \mathcal{D}_{\text{source}} \tag{5}$$

$$\boldsymbol{x}_i(k, \boldsymbol{q}) = f_i^{(k)}(\boldsymbol{x}(k-1, \boldsymbol{q})), \text{for } \boldsymbol{q} \in \mathcal{D}_{\text{source}} \cup \mathcal{D}_{\text{target}} \tag{6}$$

$$\boldsymbol{\mu}_{\text{target}}(k) = \frac{1}{|\mathcal{D}^{(\text{train})}_{\text{target}}|} \sum_{i \in I, \boldsymbol{q} \in \mathcal{D}^{(\text{train})}_{\text{target}}} \boldsymbol{x}_i(k, \boldsymbol{q}), \tag{7}$$

$$\boldsymbol{\mu}_{\text{source}}(k) = \frac{1}{|\mathcal{D}^{(\text{train})}_{\text{source}}|} \sum_{i \in I, \boldsymbol{q} \in \mathcal{D}^{(\text{train})}_{\text{source}}} \boldsymbol{x}_i(k, \boldsymbol{q}) \tag{8}$$

$$\boldsymbol{r}(k) = \boldsymbol{\mu}_{\text{target}}(k) - \boldsymbol{\mu}_{\text{source}}(k). \tag{9}$$

Like ActAdd, Mean-AcT sets $\rho_{\text{steer}}(\boldsymbol{x}(k), \boldsymbol{r}(k)) = \boldsymbol{x}(k) + \alpha \boldsymbol{r}(k)$.

Proportional–Integral–Derivative Controller

Proportional-Integral-Derivative (PID) control is a feedback mechanism extensively used in control systems (Minorsky 1922). It is valued for its simplicity, robustness, and effectiveness in a broad range of applications, from industrial automation to robotics and aerospace systems (Visioli 2006; Borase et al. 2021). The core idea behind PID control is to compute a control signal based on the error between a target reference signal and the actual output of a system. Specifically, consider a continuous-time dynamical system governed by a state space model

$$\dot{\boldsymbol{x}}(t) = g(\boldsymbol{x}(t), \boldsymbol{u}(t), t), \qquad \boldsymbol{y}(t) = h(\boldsymbol{x}(t), \boldsymbol{u}(t), t), \tag{10}$$

where $\boldsymbol{x}(t) \in \mathbb{R}^d$ denotes the state variable, $\boldsymbol{u}(t) \in \mathbb{R}^m$ is the control variable, and $\boldsymbol{y}(t) \in \mathbb{R}^{d'}$ represents the measured output signal. Here, $g : \mathbb{R}^d \times \mathbb{R}^m \to \mathbb{R}^d$ specifies the system dynamics, and $h : \mathbb{R}^d \times \mathbb{R}^m \to \mathbb{R}^{d'}$ is an output mapping. A PID controller applies the control variable $\boldsymbol{u}(t)$ to minimize the discrepancy between a target reference, or also known as the setpoint in the literature of PID control, $\boldsymbol{y}_{sp}(t)$ and the actual output $\boldsymbol{y}(t)$. This discrepancy, called control error, is defined as

$$\boldsymbol{e}(t) = \boldsymbol{y}_{sp}(t) - \boldsymbol{y}(t). \tag{11}$$

In a PID controller, the control variable $\boldsymbol{u}(t)$ is composed of the proportional (P), integral (I), and derivative (D) terms and given by:

$$\boldsymbol{u}(t) = K_p \boldsymbol{e}(t) + K_i \int_0^t \boldsymbol{e}(\tau) d\tau + K_d \frac{d\boldsymbol{e}(t)}{dt}, \tag{12}$$

where $K_p, K_i, K_d \geq 0$ are the proportional, integral, and derivation gains, respectively. In PID control design, the P, I, and D play different roles: *Proportional term (P)* outputs a correction proportional to the current error $\boldsymbol{e}_t$, but alone leaves a steady-state offset; *Integral term (I)* accumulates past errors to remove residual bias, ensuring offsets are corrected even as proportional effects fade; and *Derivative term (D)* responds to the error's rate of change, damping rapid growth to improve stability and reduce overshoot.

State-Feedback PID Controller. A special case of the PID controller is obtained by choosing the measured output $\boldsymbol{y}(t)$ to be the state variable $\boldsymbol{x}(t)$ in Eqn. 10, yielding the following state-space model

$$\dot{\boldsymbol{x}}(t) = g(\boldsymbol{x}(t), \boldsymbol{u}(t), t), \qquad \boldsymbol{y}(t) = \boldsymbol{x}(t). \tag{13}$$

The control error then becomes the state tracking error, $\boldsymbol{e}(t) = \boldsymbol{x}_{sp}(t) - \boldsymbol{x}(t)$, and the system is controlled through feedback of the state (Åström and Murray 2021).

Steering with a Feedback Controller

In this section, we will formulate popular activation steering methods, such as ActAdd, DirAblate, and Mean-AcT, as a state-feedback P controller. Based on this new interpretation, we propose PID Steering, a novel steering method that uses a PID controller.

Activation Steering as a P Controller

We consider the state-feedback PID controller given in Eqn. 13 and the continuous steering vector $\boldsymbol{r}(t)$ in which we replace the layer index k by the time index t. Substituting the state tracking error $\boldsymbol{e}(t)$ by the difference-in-means vector $\boldsymbol{r}(t)$ and using the P controller whose system dynamics is governed by $g(\boldsymbol{x}(t), \boldsymbol{u}(t), t) = f(\rho_{\text{steer}}(\boldsymbol{x}(t), \boldsymbol{u}(t)), t) - \boldsymbol{x}(t)$, we obtain

$$\dot{\boldsymbol{x}}(t) = f(\rho_{\text{steer}}(\boldsymbol{x}(t), K_p \boldsymbol{r}(t)), t) - \boldsymbol{x}(t). \tag{14}$$

We discretize Eqn. 14 using Euler method (Euler 1768; Hairer, Wanner, and Nørsett 1993) to obtain

$$\boldsymbol{x}(k) - \boldsymbol{x}(k-1) = f^{(k)}(\rho_{\text{steer}}(\boldsymbol{x}(k-1), K_p \boldsymbol{r}(k-1))) - \boldsymbol{x}(k-1)$$

or equivalently,

$$\boldsymbol{x}(k) = f^{(k)}(\rho_{\text{steer}}(\boldsymbol{x}(k-1), K_p \boldsymbol{r}(k-1))), \tag{15}$$

where $f^{(k)}(\cdot) = f(\cdot, k)$, a function depending on index k.

Comparing Eqn. 15 with Eqn. 1 and 2 shows that applying the steering vectors as in Section is equivalent to implementing the P controller, where $f^{(k)}$ is the k-th layer in an LLM, $\boldsymbol{u}(t) = K_p \boldsymbol{r}(t)$ is the new steering vector. Thus, activation steering computes the expected state tracking error.

$$\boldsymbol{r}(t) = \bar{\boldsymbol{e}}(t) = \mathbb{E}_{\boldsymbol{q}_{sp} \in \mathcal{D}^{(\text{train})}_{\text{target}}}[\boldsymbol{x}_{sp}(t, \boldsymbol{q}_{sp})] - \mathbb{E}_{\boldsymbol{q} \in \mathcal{D}^{(\text{train})}_{\text{source}}}[\boldsymbol{x}(t, \boldsymbol{q})]. \tag{16}$$

This expected state tracking error, i.e., the difference-in-means vector $\boldsymbol{r}(t)$, can be computed non-sequentially or sequentially, as explained in Section . When $\boldsymbol{r}(t)$ is computed non-sequentially and $\rho_{\text{steer}}(\boldsymbol{x}(k), \boldsymbol{u}(k)) = \boldsymbol{x}(k) + \alpha \boldsymbol{u}(k)$ or $\boldsymbol{x}(k) - \boldsymbol{u}(k)\ \boldsymbol{u}(k)^\top\ \boldsymbol{x}(k)$, we obtain ActAdd or DirAblate, respectively. When $\boldsymbol{r}(t)$ is computed sequentially and $\rho_{\text{steer}}(\boldsymbol{x}(k), \boldsymbol{u}(k)) = \boldsymbol{x}(k) + \alpha \boldsymbol{u}(k)$, we attain Mean-AcT.

Limitations of P Controller. There is always a steady state error in P control. The error decreases with increasing gain, but the tendency towards oscillation also increases. Since activation steering methods, i.e., ActAdd, DirAblate, and Mean-Act, are P controllers, they share the same limitations. We informally state our theoretical guarantees that P-control activation steering methods cannot alleviate the steady state error in Proposition 1 below and provide detailed proofs in Appendix .

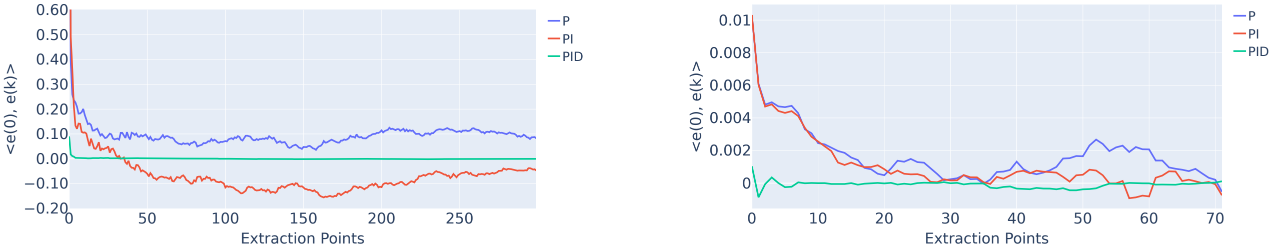
Proposition 1 (Steady-state error of P-control steering) *P-control activation steering ensures input-to-state stability (ISS) for an appropriate range of K_p. However, there still exists a steady-state error due to the disturbance $\boldsymbol{w}(k)$ to the state of the system. In the best case, when $\boldsymbol{w}(k)$ converges to $\boldsymbol{w}$, under a mild condition, the expected error, i.e., the difference-in-means, $\boldsymbol{r}(k) = \bar{\boldsymbol{e}}(k)$ eventually converges to a steady state $\bar{\boldsymbol{e}}_{ss} \propto \boldsymbol{w}$. Therefore, $\bar{\boldsymbol{e}}_{ss} \neq 0$ if $\boldsymbol{w} \neq 0$.*

We further provide empirical evidence to validate Proposition 1 in Figure 3 below. To archive this, we apply Sequential P-control activation steering (P Steering) on a randomly intitialized model with 150 layers deep, and pretrained Qwen2.5-3B-Instruct. We use $\langle \bar{e}(0), \bar{e}(t) \rangle$ as metric since it is a scalar measure of the energy retained along $\bar{e}(0)$. If this quantity fails to decay to zero, e.g., under noise, it indicates a persistent component of the initial error, i.e., an undesired dynamic (see Appendix for further explanation). It can be seen that, for both the randomly initialized and the pretrained models, the errors do not vanish completely. These results confirm that P-control activation steering ensures stability but admits a persistent steady-state error due to the disturbance.

Proportional–Integral–Derivative (PID) Steering

Overview To overcome the steady-state error inherent in P-control activation steering, we extend the method by adding integral (I) and derivative (D) terms to the steering vectors. PID Steering thus (i) reacts immediately to errors via the P term for greater responsiveness, (ii) removes steady-state offsets with the I term, ensuring convergence to the desired set point, and (iii) anticipates error trends through the D term, improving stability and reducing overshoot. Together, these properties yield the following advantages:

- **Generalization.** PID Steering extends P-control methods like ActAdd, DirAblate, and Mean-AcT by adding integral and derivative components.
- **Methodological Agnosticism.** Our PID framework can be applied across different activation steering techniques, including ActAdd, DirAblate, and Mean-AcT.
- **Stability.** We theoretical prove and empirical demonstrate that PID Steering reduces steady-state error and overshoot in P-controllers, improving existing steering methods.
- **Interpretability.** Derived from classical feedback control (Minorsky 1922), the framework inherits the simplicity and interpretability that underpin the wide use of PID controllers.



(a) Randomly Initialized LLama3

(b) Pretrained Qwen2.5-3B-Instruct

Figure 3: Scalar errors across time step of randomly initialized model after applying P, PI, and PID controller.

Computing the Steering Direction Using a PID Feedback Controller Following Section , we consider the state-feedback PID controller in Eqn. 13, replacing the state tracking error $\boldsymbol{e}(t)$ with the difference-in-means vector $\boldsymbol{r}(t)$. With the PID controller governed by the system dynamics $g(\boldsymbol{x}(t), \boldsymbol{u}(t), t) = f(\rho_{\text{steer}}(\boldsymbol{x}(t), \boldsymbol{u}(t)), t) - \boldsymbol{x}(t)$, we obtain

$$\dot{\boldsymbol{x}}(t) = f(\rho_{\text{steer}}(\boldsymbol{x}(t), K_p\boldsymbol{r}(t) + K_i \int_0^t \boldsymbol{r}(\tau)d\tau \tag{17}$$

$$+ K_d \frac{d\boldsymbol{r}(t)}{dt}), t) - \boldsymbol{x}(t). \tag{18}$$

Eqn. 17 defines the continuous-time model of PID Steering, whose steering vector is given by:

$$\boldsymbol{u}(t) = K_p\boldsymbol{r}(t) + K_i \int_0^t \boldsymbol{r}(\tau)d\tau + K_d \frac{d\boldsymbol{r}(t)}{dt}. \tag{19}$$

In order to obtain the discrete-time formulation of PID Steering, we first discretize the sytem dynamics $\dot{\boldsymbol{x}}(t) = f(\rho_{\text{steer}}(\boldsymbol{x}(t), \boldsymbol{u}(t)), t) - \boldsymbol{x}(t)$ using Euler method (Euler 1768; Hairer, Wanner, and Nørsett 1993), same as in Section , and attain

$$\boldsymbol{x}(k) = f^{(k)}(\rho_{\text{steer}}(\boldsymbol{x}(k-1), \boldsymbol{u}(k-1))), \tag{20}$$

Next, we discretize $\boldsymbol{u}(t)$ given in Eqn. 19 to obtain $\boldsymbol{u}(k)$ using Lemma 1 below.

Lemma 1 (Discretizing PID steering vector) *Consider the continuous PID steering vector defined in Eqn. 19. The discrete-time PID steering vector is given by:*

$$\boldsymbol{u}(k) = K_p\boldsymbol{r}(k) + K_i \sum_{j=0}^{k-1} \boldsymbol{r}(j) + K_d(\boldsymbol{r}(k) - \boldsymbol{r}(k-1)). \tag{21}$$

Proof of Lemma 1 is in Appendix . With Eqn. 20 and Lemma 1, we now define PID Steering.

Definition 1 (PID Steering) *Given a large language model whose layers are $\{f^{(k)}\}_{k=1}^{K}$ and a steering function ρ_{steer}, PID Steering constructs the steering vectors as follows:*

$$\boldsymbol{u}(k) = K_p\boldsymbol{r}(k) + K_i \sum_{j=0}^{k-1} \boldsymbol{r}(j) + K_d(\boldsymbol{r}(k) - \boldsymbol{r}(k-1)), \tag{22}$$

where for non-sequential mapping,

$$\boldsymbol{r}(k) = \mathbb{E}_{\boldsymbol{q}_{sp} \in \mathcal{D}_{target}^{(train)}}[\boldsymbol{x}_{sp}(k, \boldsymbol{q}_{sp})] - \mathbb{E}_{\boldsymbol{q} \in \mathcal{D}_{source}^{(train)}}[\boldsymbol{x}(k, \boldsymbol{q})],$$

and for sequential mapping,

$$\tilde{\boldsymbol{x}}(k) = f^{(k)}\left(\rho_{steer}\big(\boldsymbol{x}(k-1), \boldsymbol{u}(k-1)\big)\right), \tag{23}$$

$$\boldsymbol{r}(k) = \mathbb{E}_{\boldsymbol{q}_{sp} \in \mathcal{D}_{target}^{(train)}}[\boldsymbol{x}_{sp}(k, \boldsymbol{q}_{sp})] - \mathbb{E}_{\boldsymbol{q} \in \mathcal{D}_{source}^{(train)}}[\tilde{\boldsymbol{x}}(k, \boldsymbol{q})].$$

Theoretical Analysis of PID Steering

This section provides theoretical evidence for our claims: (i) adding integral action (PI) reduces steady-state error that remains under pure P-control (Proposition 3); and (ii) adding a derivative term (PID) preserves bias removal while mitigating oscillations/overshoot (Proposition 1 and 2). We denote $\boldsymbol{K}_p := K_p\boldsymbol{I}$, $\boldsymbol{K}_i := K_i\boldsymbol{I}$, and $\boldsymbol{K}_d := K_d\boldsymbol{I}$. Detailed proofs are provided in Appendix .

Dynamics of the Average Error Across Layers

To formalize the problem, we consider N pairs of prompts/input tokens from two contrastive datasets, e.g. harmful and harmless, $\{(\boldsymbol{q}_i^+, \boldsymbol{q}_i^-)\}_{i=1}^N$ with corresponding activations $\boldsymbol{x}_i^{\pm}(k) \in \mathbb{R}^d$ at layer k. A steering input $\boldsymbol{u}(k)$ perturbs the undesired branch:

$$\boldsymbol{x}_i^-(k+1) \;=\; f_i^{(k)}\big(\boldsymbol{x}_i^-(k) + \boldsymbol{u}(k)\big). \tag{24}$$

Let $\bar{\boldsymbol{e}}(k) := \frac{1}{N}\sum_{i=1}^N(\boldsymbol{x}_i^+(k) - \boldsymbol{x}_i^-(k))$, the error dynamics of activation steering is then given by Proposition 2 below.

Proposition 2 (Error dynamics of activation steering) *The error dynamics $\bar{\boldsymbol{e}}(k)$ in activation steering is of the form:*

$$\bar{\boldsymbol{e}}(k+1) = \bar{\boldsymbol{A}}(k)\,\bar{\boldsymbol{e}}(k) - \bar{\boldsymbol{A}}(k)\,\boldsymbol{u}(k) + \boldsymbol{w}(k), \tag{25}$$

where $\bar{\boldsymbol{A}}(k)$ is the mean local Jacobian of $f_i^{(k)}$ at $\boldsymbol{x}_i^+(k)$ and the disturbance term $\boldsymbol{w}(k)$ collects heterogeneity. See Appendix for detailed proof and explanations of the terms.

Our control objective is to drive $\bar{\boldsymbol{e}}(k)$ to zero with input-to-state stability (ISS) for disturbed discrete system Eqn. 25 (Jiang, Sontag, and Wang 1999).

Stability of the Error Dynamics: Roles and Caveats of PI and PID Control

In the following stability analysis, we consider the orthogonal decomposition for the disturbance $\boldsymbol{w}(k) = \boldsymbol{w}^{\parallel}(k) + \boldsymbol{w}^{\perp}(k)$, where $\boldsymbol{w}^{\parallel}(k) \in \operatorname{Im}\bar{\boldsymbol{A}}(k)$ and $\boldsymbol{w}^{\perp}(k) \in (\operatorname{Im}\bar{\boldsymbol{A}}(k))^{\perp}$.

PI Control The following proposition provides a theoretical guarantee of PI Steering's steady-state error reduction.

Proposition 3 (Stabilizing the PI loop) *Let* $\boldsymbol{M}_p(k) = \bar{\boldsymbol{A}}(k)(\boldsymbol{I} - \boldsymbol{K}_p)$*, and denote* $\|\boldsymbol{K}_i\| =: h$*. Assume* $\sup_k \|\bar{\boldsymbol{A}}(k)\| \leq M < \infty$ *and* $\sup_k \|\boldsymbol{M}_p(k)\| \leq q < 1$*. If* $q + Mh < 1$*, then the PI closed-loop control is ISS. Furthermore, the integral part exactly cancels the matched disturbance component* $\boldsymbol{w}^{\parallel}$*. The remaining error is due only to the unmatched component* $\boldsymbol{w}^{\perp}$*, which cannot be compensated. Full proof and term explanations provided in Appendix*

Limitations of PI control. Overshoot is common under PI: the closed loop oscillates about the setpoint before settling (Åström and Hägglund 1995b, Ch. 3, §3.3, pp. 68-69), and large overshoot can arise with a high integral gain $\boldsymbol{K}_i$. In our steering setting, we explain this by scalarizing the dynamics along a reference direction . The scalarized integral state accumulates past error, pushing the trajectory beyond the setpoint; when the scalarized error changes sign, the integral discharges and the error subsequently approaches zero. See Fig. 5 for an illustration and Appendix for the formal derivation.

PID Control The derivative action counteracts PI-induced oscillations near the setpoint by responding to decreases in the scalarized error, while preserving the integral term's bias-removal role, as shown in Theorems 1 and 2. For detailed proofs and explanations, see Appendix .

Theorem 1 (Stabilizing the PID loop) *Let* $\boldsymbol{M}_p(k) = \bar{\boldsymbol{A}}(k)(\boldsymbol{I} - \boldsymbol{K}_p)$*, and denote* $\|\boldsymbol{K}_i\| =: h$*,* $\|\boldsymbol{K}_d\| =: \ell$*. Assume* $\sup_k \|\bar{\boldsymbol{A}}(k)\| \leq M < \infty$ *and* $\sup_k \|\boldsymbol{M}_p(k)\| \leq q < 1$*. If* $q + Mh < 1$ *(stable PI loop), then there exists* $\ell > 0$ *such that the PID closed-loop control is ISS. Therefore, the integral part in PID design still cancels the matched disturbance component* $\boldsymbol{w}^{\parallel}$*.*

Theorem 2 (PID reduces the first-overshoot amplitude) *Let the first overshoot occur at index* k_0 *with amplitude* A_0 *(definition in Eqn. 73). Then, the first-overshoot amplitude under PID Steering,* A_0^{PID}*, satisfies* $A_0^{\mathrm{PID}} \leq A_0^{\mathrm{PI}}$*, where* A_0^{PI} *denotes the corresponding amplitude under PI Steering.*

To support the theory, we present empirical evidence in Fig. 3. PI and PID controllers clearly improve over P-only control: PI removes steady-state error but causes large overshoot, while adding the derivative term mitigates overshoot and enables faster, cleaner convergence to zero.

Controlling the Steering Effect

In this section, we demonstrate the applicability and effectiveness of PID-Steering by using it as a drop-in replacement for the steering vector computation step across multiple steering frameworks.

Toxicity Mitigation

We evaluate the effectiveness of PID Steering for toxic language mitigation in comparison to sequential steering methods, specifically Linear-AcT and Mean-AcT (Rodriguez et al. 2025), by closely following their experimental setup. We apply PID-Steering into Mean-AcT and call it PID-AcT.

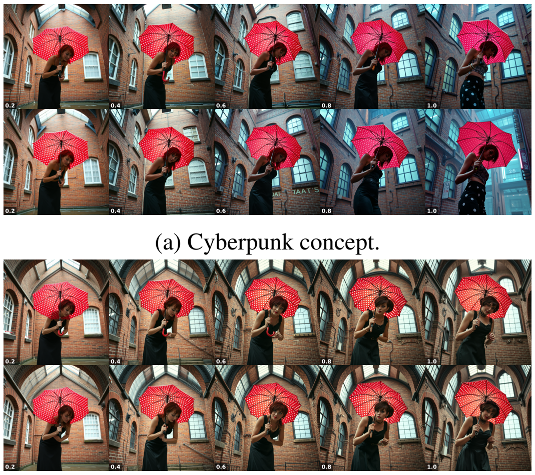
Experimental Setup. Our evaluation is conducted on Gemma2-2B (Gemma Team et al. 2024) and Llama3-8B (Llama Team 2024), using 1,000 randomly sampled prompts from the RealToxicityPrompts dataset (Gehman et al. 2020). Toxicity is quantified with a ROBERTA-based classifier (Logacheva et al. 2022), following the methodology of Suau et al. (2024). We also assess toxicity in a zero-shot setting by employing Llama3-8B-Instruct as an LLM-as-a-judge (Zheng et al. 2023).

To measure general utility of the intervened models, we report: (i) perplexity (PPL) on a fixed set of 20k Wikipedia sentences, (ii) PPL of model-generated outputs evaluated with Mistral-7B (Jiang et al. 2023), and (iii) 5-shot MMLU (Hendrycks et al. 2021) accuracy.

Results. PID-AcT cuts toxicity by up to 8× while preserving utility. As shown in Table 1, it lowers scores by **8.1×** on Gemma2-2B and **7.3×** on LLaMA3-8B, consistently outperforming Mean-AcT and Linear-AcT. It achieves the lowest toxicity under both classifier and LLM-judge evaluations, while maintaining utility: MMLU drops ≤1% and perplexity rises only modestly (≤6%). Unlike sequential methods that plateau, PID's integral-derivative dynamics deliver stronger, more stable mitigation without harming performance.

Image Generation Styles Control

We study activation steering in diffusion models using FLUX.1.Schnell's denoising transformer (Labs 2024), built on T5-XXL encoders (Raffel et al. 2020) and requiring just 4 diffusion steps. **Experimental Setup.** Following (Rodriguez et al. 2025), we intervene on all normalization layers after most residual blocks in FLUX. Style/concept expression is measured by a CLIP zero-shot classifier with two labels (`A picture of a {style/concept}''` vs. `A`



(a) Cyberpunk concept.

(b) Steampunk concept.

Figure 4: Qualitative results of activation steering in FLUX-Schnell across two style concepts with the prompt *"Lady bent over with red polka dot umbrella inside a brick building."*

Table 1: Toxicity mitigation results for Gemma-2B and Llama-8B, averaged over 10 runs. Lower is better for toxicity and perplexity; higher is better for MMLU. Best and second-best exclude the Original baseline.

		Seq.	**CLS Tox. (%) ↓**	**0-shot Tox. (%) ↓**	**PPL Wikipedia ↓**	**PPL Mistral-7B ↓**	**MMLU ↑**
Gemma2-2B	Original	–	$4.13_{\pm 0.43}$	$12.85_{\pm 0.94}$	$14.40_{\pm 0.20}$	$6.05_{\pm 0.51}$	$53.03_{\pm 0.60}$
	Mean-Act		$1.12_{\pm 0.23}$	$5.20_{\pm 0.42}$	$\underline{14.53_{\pm 0.21}}$	$6.81_{\pm 0.19}$	$\mathbf{51.74_{\pm 0.55}}$
	Linear-Act		$0.95_{\pm 0.36}$	$5.37_{\pm 0.80}$	$14.75_{\pm 0.22}$	$7.24_{\pm 0.24}$	$51.63_{\pm 0.50}$
	Mean-Act	✓	$\underline{0.68_{\pm 0.21}}$	$\underline{3.23_{\pm 0.44}}$	$\underline{14.92_{\pm 0.25}}$	$6.97_{\pm 0.74}$	$\mathbf{51.80_{\pm 0.55}}$
	Linear-Act	✓	$1.00_{\pm 0.27}$	$4.13_{\pm 0.89}$	$14.98_{\pm 0.22}$	$\underline{7.13_{\pm 0.70}}$	$\underline{51.47_{\pm 0.50}}$
	PID-Act	✓	$\mathbf{0.51_{\pm 0.21}}$	$\mathbf{2.90_{\pm 0.55}}$	$15.22_{\pm 0.24}$	$7.02_{\pm 0.65}$	$51.30_{\pm 0.52}$
Llama3-8B	Original	–	$5.30_{\pm 0.35}$	$15.24_{\pm 0.40}$	$9.17_{\pm 0.18}$	$5.18_{\pm 0.20}$	$65.33_{\pm 0.42}$
	Mean-Act		$1.78_{\pm 0.33}$	$6.56_{\pm 0.54}$	$9.36_{\pm 0.28}$	$5.45_{\pm 0.34}$	$64.35_{\pm 0.39}$
	Linear-Act		$1.87_{\pm 0.39}$	$6.55_{\pm 0.21}$	$9.35_{\pm 0.17}$	$5.56_{\pm 0.33}$	$64.55_{\pm 0.33}$
	Mean-Act	✓	$\underline{1.21_{\pm 0.41}}$	$\underline{5.09_{\pm 0.64}}$	$9.83_{\pm 0.21}$	$5.71_{\pm 0.33}$	$64.22_{\pm 0.40}$
	Linear-Act	✓	$1.68_{\pm 0.48}$	$6.47_{\pm 0.38}$	$\underline{9.48_{\pm 0.19}}$	$\underline{5.46_{\pm 0.44}}$	$\underline{64.49_{\pm 0.38}}$
	PID-Act	✓	$\mathbf{0.72_{\pm 0.49}}$	$\mathbf{4.36_{\pm 0.81}}$	$\mathbf{9.56_{\pm 0.20}}$	$\mathbf{6.08_{\pm 0.37}}$	$\mathbf{64.50_{\pm 0.36}}$

`picture of something''`), and content preservation by CLIPScore (Hessel et al. 2021). Training uses 2,048 COCO Captions (Chen et al. 2015) prompts augmented with *cyberpunk*/*steampunk* modifiers from LLaMA-8B-Instruct (source = unmodified p, target = modified q). Evaluation samples 512 validation prompts to generate images across intervention strengths.

Results. In Fig. 4, raising intervention strength from 0 to 1 yields a smooth progression of stylistic traits, e.g., neon hues for *cyberpunk*, mechanical textures for steampunk, while preserving core content. At moderate strengths, style is pronounced yet faithful, and even at high strengths semantic alignment largely holds. Quantitatively (Figs. 6), style expression measured by a zero-shot classifier increases monotonically, with PID-AcT surpassing Mean-AcT, especially at mid strengths (0.4-0.8). CLIPScore reveals the trade-off:Both Mean-AcT and PID-AcT exhibit a steady decline. While PID-AcT drops slightly more, the difference is marginal.

Related Works

Recent works increasingly frame large language models (LLMs) as *dynamical systems*, where generation is a trajectory in latent space. This view shifts activation steering from heuristic nudging to principled *control*: rather than biasing outputs without guarantees, controllers enforce constraints on trajectories with formal assurances (Cheng and Amo Alonso 2024). In our PID-steering framework, this distinction is key: we treat the model as a plant with hidden states evolving under controlled interventions.

Closed-loop activation control. Cheng and Amo Alonso (2024) propose *Linear Semantic Control* (LiSeCo), which projects activations into safe subspaces at each decoding step via a closed-form controller. This yields lightweight, guaranteed control of simple attributes (e.g., toxicity, sentiment). However, the linearity assumption only approximates LLM embeddings, guarantees are local rather than global, and long-horizon stability remains unaddressed.

Dynamic representation editing. Kong et al. (2024) introduce *RE-CONTROL*, which learns a value function on hidden states and applies gradient-based interventions at test time. This dynamic approach generalizes steering into a Bellman-optimal control problem, balancing alignment with fluency. Still, accuracy of the learned value function is critical, test-time optimization adds overhead, and local interventions may not guarantee global alignment.

Together, these works move activation steering from heuristics to control theory. Soatto et al. (2023) prove fundamental controllability (but under strong assumptions), Luo et al. unify prompt strategies as open-loop control (without guarantees), Cheng and Amo Alonso (2024) derive closed-form activation control (limited to linear approximations), and Kong et al. (2024) extend to dynamic optimal control (with overhead and approximation risks).

Concluding Remarks

We introduced PID Steering, a control-theoretic approach to activation steering that models layer-wise representations as a dynamical system. This framework unifies prior methods, offers robustness guarantees, and leverages PID dynamics for computing steering vectors. Across language and diffusion models, PID Steering achieves stronger and more stable performance than existing approaches in toxicity mitigation, jailbreak prevention, and style control, while preserving model utility. Our results highlight control theory as a principled foundation for developing reliable and generalizable steering methods. A limitation of our work is the use of "stability-first, one-gain-at-a-time" analytical strategy to find controller gains: it clarifies the role of each component but may miss optimal choices and can overlook broader feasible regions. To address this, numerical methods, for example, LMI-based computations, can be employed. We leave these for future work.

Supplement to "Activation Steering with a Feedback Controller"

Theoretical Proofs

Discretized PID controller

Implementing a continuous-time controller on digital hardware, such as PID, requires discretizing its derivative and integral terms (Åström and Hägglund 1995b, p.95)

Lemma 1 (Discretizing PID steering vector) *Consider the continuous PID steering vector defined in Eqn. 19. The discrete-time PID steering vector is given by:*

$$\boldsymbol{u}(k) = K_p\boldsymbol{r}(k) + K_i \sum_{j=0}^{k-1} \boldsymbol{r}(j) + K_d(\boldsymbol{r}(k) - \boldsymbol{r}(k-1)). \tag{21}$$

Proof. We follow the discretization procedure for PID controllers in (Åström and Hägglund 1995b, Sec. 3.6, Ch. 3). For simplicity, the sampling period is normalized to $h = 1$.

Proportional term in Eqn. 19.

$$P(t) = \boldsymbol{K}_p\, \boldsymbol{r}(t).$$

The discrete-time form is obtained by substituting sampled variables for their continuous counterparts:

$$P(k) = \boldsymbol{K}_p\, \boldsymbol{r}(k). \tag{26}$$

Integral term in Eqn. 19.

$$I(t) = \boldsymbol{K}_i \int_0^t \boldsymbol{r}(\tau)\, d\tau \qquad \Rightarrow \qquad \frac{dI}{dt} = \boldsymbol{K}_i\, \boldsymbol{r}(t).$$

Using forward Euler with $h = 1$,

$$I(k+1) - I(k) = \boldsymbol{K}_i\, \boldsymbol{r}(k).$$

Hence

$$I(k+1) = I(k) + \boldsymbol{K}_i\, \boldsymbol{r}(k),$$

which is equivalent to

$$I(k) = I(0) + \boldsymbol{K}_i \sum_{j=0}^{k-1} \boldsymbol{r}(j) = \boldsymbol{K}_i \sum_{j=0}^{k-1} \boldsymbol{r}(j), \tag{27}$$

since $I(0) = 0$.

Derivative term in Eqn. 19.

$$D(t) = \boldsymbol{K}_d\, \frac{d\boldsymbol{r}(t)}{dt}.$$

Approximating the derivative by the backward Euler difference with $h = 1$ gives

$$D(k) = \boldsymbol{K}_d\, \big(\boldsymbol{r}(k) - \boldsymbol{r}(k-1)\big). \tag{28}$$

Combining equation 26, equation 27, and equation 28 yields

$$\boldsymbol{u}(k) = \boldsymbol{K}_p\, \boldsymbol{r}(k) \;+\; \boldsymbol{K}_i \sum_{j=0}^{k-1} \boldsymbol{r}(j) \;+\; \boldsymbol{K}_d\big(\boldsymbol{r}(k) - \boldsymbol{r}(k-1)\big).$$

□

Background on Input-to-state Stability & Notations

Background on Input-to-state Stability (ISS) In our proofs, the input-to-state stability (ISS) of a system can be established either through the definition of an ISS system in (Jiang, Sontag, and Wang 1999, Def. 2.1) or via the use of an ISS-Lyapunov function as in (Jiang, Sontag, and Wang 1999, Def. 2.2, Prop. 2.3). We also rely on the definition of a Lyapunov function and the difference Lyapunov equation for linear discrete-time homogeneous dynamical systems in (Gajic and Qureshi 2008, Ch. 1, p. 8). The existence of a solution to the Lyapunov equation, together with its bound, is stated in (Gajic and Qureshi 2008, Ch. 4, p. 110).

For reference, we briefly note that input-to-state stability (ISS) extends the classical notion of Lyapunov by explicitly accounting for external inputs: the state remains bounded and eventually whenever the input is bounded. A Lyapunov function provides an energy-like certificate for stability, while the associated Lyapunov equation offers a constructive method for obtaining such functions in linear settings. These notions are central for analyzing stability and will be used throughout our proofs.

Conventions and assumptions (used throughout). Let $\|\cdot\|$ denote the Euclidean norm on $\mathbb{R}^d$; for a matrix $M \in \mathbb{R}^{d\times d}$ we also write $\|M\|$ for the operator norm induced by the Euclidean norm, i.e. $\|M\| \;:=\; \sup_{\|x\|=1} \|Mx\|$ (the spectral norm) (Horn and Johnson 2012, pp. 343–346). We assume (i) $\sup_k \|\bar{\boldsymbol{A}}(k)\| < \infty$; (ii) $\boldsymbol{w}(k)$ is bounded (for a signal $\boldsymbol{w}$ we set $\|\boldsymbol{w}\|_\infty := \sup_{k\geq 0} \|\boldsymbol{w}(k)\|$) ; (iii) the controller gains are static and time-invariant scalar multiples of the identity. We use the standard meaning of the classes $\mathcal{K}$ and $\mathcal{KL}$ as in (Jiang, Sontag, and Wang 1999).

Dynamics of the Average Error Across Layers

To formalize the problem setup, we consider N pairs of contrastive prompt/input tokens $\{(\boldsymbol{q}_i^+, \boldsymbol{q}_i^-)\}_{i=1}^N$, where $\boldsymbol{q}_i^+$ carries the desired property and $\boldsymbol{q}_i^-$ represents the opposite. For discrete time (layer) k, let $\boldsymbol{x}_i^\pm(k) \in \mathbb{R}^d$ denote the corresponding activation vectors. The layer-to-layer evolution is

$$\boldsymbol{x}_i(k+1) \;=\; f_i^{(k)}\big(\boldsymbol{x}(k)\big),\;\; i = 1, \ldots, N, \tag{29}$$

with $f_i^{(k)} : \mathbb{R}^d \to \mathbb{R}^d$ differentiable on the operating region. A steering input $u(t)$ is applied on the undesired branch:

$$\boldsymbol{x}_i^-(k+1) \;=\; f_i^{(k)}\big(\boldsymbol{x}_i^-(k) + \boldsymbol{u}(k)\big). \tag{30}$$

Defining $\bar{\boldsymbol{x}}^\pm(k) \;:=\; \frac{1}{N}\sum_{i=1}^N \boldsymbol{x}_i^\pm(k)$, we track the per-pair and average errors as

$$\boldsymbol{e}_i(k) := \boldsymbol{x}_i^+(k) - \boldsymbol{x}_i^-(k), \tag{31}$$

$$\bar{\boldsymbol{e}}(k) := \bar{\boldsymbol{x}}^+(k) - \bar{\boldsymbol{x}}^-(k), \tag{32}$$

$$\tilde{\boldsymbol{e}}_i(k) = \boldsymbol{e}_i(k) - \bar{\boldsymbol{e}}(k). \tag{33}$$

Furthermore, we define $\boldsymbol{A}_i(k)$ as the Jacobian of $f_i^{(k)}$ at $\boldsymbol{x}_i^+(k)$:

$$\boldsymbol{A}_i(k) := J_{f_i^{(k)}}\big(\boldsymbol{x}_i^+(k)\big), \tag{34}$$

$$\bar{\boldsymbol{A}}(t) := \frac{1}{N}\sum_{i=1}^{N} \boldsymbol{A}_i(k), \tag{35}$$

$$\tilde{\boldsymbol{A}}_i(k) = \boldsymbol{A}_i(k) - \bar{\boldsymbol{A}}(k). \tag{36}$$

The dynamic of the average error $\bar{\boldsymbol{e}}(k)$ is then given by Proposition 2.

Proposition 2 (Error dynamics of activation steering) *The error dynamics $\bar{\boldsymbol{e}}(k)$ in activation steering is of the form:*

$$\bar{\boldsymbol{e}}(k+1) = \bar{\boldsymbol{A}}(k)\,\bar{\boldsymbol{e}}(k) - \bar{\boldsymbol{A}}(k)\,\boldsymbol{u}(k) + \boldsymbol{w}(k), \tag{25}$$

where $\bar{\boldsymbol{A}}(k)$ is the mean local Jacobian of $f_i^{(k)}$ at $\boldsymbol{x}_i^+(k)$ and the disturbance term $\boldsymbol{w}(k)$ collects heterogeneity. See Appendix for detailed proof and explanations of the terms.

Proof. The evolution of the average error $\bar{\boldsymbol{e}}(k)$ through layers can be described as follows:

$$\bar{\boldsymbol{e}}(k+1) = \bar{\boldsymbol{x}}^+(k+1) - \bar{\boldsymbol{x}}^-(k+1) \tag{37}$$

$$= \frac{1}{N}\sum_{i=1}^{N}\Big[f_i^{(k)}\big(\boldsymbol{x}_i^+(k)\big) - f_i^{(k)}\big(\boldsymbol{x}_i^-(k) + \boldsymbol{u}(k)\big)\Big]. \tag{38}$$

Linearizing $f_i^{(k)}$ around $\boldsymbol{x}_i^+(k)$, we obtain

$$f_i^{(k)}\big(\boldsymbol{x}_i^+(k)+\delta\big) \approx f_i^{(k)}\big(\boldsymbol{x}_i^+(k)\big) + J_{f_i^{(k)}}\big(\boldsymbol{x}_i^+(k)\big)\cdot\delta, \tag{39}$$

where $J_{f_i^{(k)}}$ denotes the Jacobian of $f_i^{(k)}$.

Setting $\delta = -\,\boldsymbol{e}_i(k) + \boldsymbol{u}(k)$ yields

$$f_i^{(k)}\big(\boldsymbol{x}_i^+(k)+\delta\big) \approx f_i^{(k)}\big(\boldsymbol{x}_i^+(k)\big) \;+\; \boldsymbol{A}_i(k)\,\big(\boldsymbol{e}_i(k)+\boldsymbol{u}(k)\big),$$

Insert this into Eqn.37 we obtain

$$\bar{\boldsymbol{e}}(k+1) = \frac{1}{N}\sum_{i=1}^{N}\boldsymbol{A}_i(k)\,\boldsymbol{e}_i(k) - \bar{\boldsymbol{A}}(k)\,\boldsymbol{u}(k). \tag{40}$$

Recall that

$$\boldsymbol{e}_i(k) = \bar{\boldsymbol{e}}(k) + \tilde{\boldsymbol{e}}_{(i)}(k), \text{ then } \frac{1}{N}\sum_{i=1}^{N}\tilde{\boldsymbol{e}}_{(i)}(k) = 0,$$

and

$$\boldsymbol{A}_{(i)}(t) = \bar{\boldsymbol{A}}(k) + \tilde{\boldsymbol{A}}_{(i)}(k), \text{ then } \frac{1}{N}\sum_{i=1}^{N}\tilde{\boldsymbol{A}}_{(i)}(k) = 0.$$

Therefore,

$$\begin{aligned}\bar{\boldsymbol{e}}(k+1) &= \frac{1}{N}\sum_{i=1}^{N}\boldsymbol{A}_i(k)\,\boldsymbol{e}_i(k) - \bar{\boldsymbol{A}}(k)\,\boldsymbol{u}(k)\\ &= \frac{1}{N}\sum_{i=1}^{N}\bar{\boldsymbol{A}}(k)\,\bar{\boldsymbol{e}}_i(k) - \bar{\boldsymbol{A}}(k)\,\boldsymbol{u}(k)\\ &+ \frac{1}{N}\sum_{i=1}^{N}\tilde{\boldsymbol{e}}_{(i)}(k)\tilde{\boldsymbol{A}}_{(i)}(k)\\ &\underbrace{+\,\bar{\boldsymbol{A}}(k)\frac{1}{N}\sum_{i=1}^{N}\tilde{\boldsymbol{e}}_{(i)}(k) + \bar{\boldsymbol{e}}(k)\frac{1}{N}\sum_{i=1}^{N}\tilde{\boldsymbol{A}}_{(i)}(k)}_{=0}\end{aligned} \tag{41}$$

We then obtain the final state-space model for the dynamics of $\bar{e}(t)$ as

$$\bar{\boldsymbol{e}}(k+1) = \bar{\boldsymbol{A}}(k)\,\bar{\boldsymbol{e}}(k) - \bar{\boldsymbol{A}}(k)\,\boldsymbol{u}(k) + \boldsymbol{w}(k), \tag{42}$$

where

$$\boldsymbol{w}(k) = \frac{1}{N}\sum_{i=1}^{N}\tilde{\boldsymbol{A}}_i(k)\,\tilde{\boldsymbol{e}}_i(k),$$

which acts as a time-dependent exogeneous disturbance to the model □

Proportional (P) Control

Consider proportional control with

$$\boldsymbol{u}(k) = K_P\bar{\boldsymbol{e}}(k), \quad (K_I = K_D = 0).$$

The dynamics Eqn. 25 then become

$$\bar{\boldsymbol{e}}(k) = M_P(k)\bar{\boldsymbol{e}}(k) + \boldsymbol{w}(k), \tag{43}$$

where $M_P(k) = \bar{\boldsymbol{A}}(k)(I - K_P)$.

With a suitable choice of K_P, the system can be made input-to-state stable (ISS); that is, there exist a $\mathcal{KL}$-function β and a $\mathcal{K}$-function γ such that, for all disturbance $\boldsymbol{w}$ with bounded sup norm and all initial states $\bar{\boldsymbol{e}}(0)$,

$$\|\bar{\boldsymbol{e}}(k)\| \;\le\; \beta(\|\bar{\boldsymbol{e}}(0)\|, k) \;+\; \gamma(\|\boldsymbol{w}\|_\infty), \quad k \in \mathbb{Z}_{\ge 0}, \tag{44}$$

see (Jiang, Sontag, and Wang 1999, Def. 2.1).

In particular, the error decays from the initial condition and remains bounded under bounded disturbances.

Proposition 1 (Steady-state error of P-control steering) *P-control activation steering ensures input-to-state stability (ISS) for an appropriate range of K_p. However, there still exists a steady-state error due to the disturbance $\boldsymbol{w}(k)$ to the state of the system. In the best case, when $\boldsymbol{w}(k)$ converges to $\boldsymbol{w}$, under a mild condition, the expected error, i.e., the difference-in-means, $\boldsymbol{r}(k) = \bar{\boldsymbol{e}}(k)$ eventually converges to a steady state $\bar{\boldsymbol{e}}_{ss} \propto \boldsymbol{w}$. Therefore, $\bar{\boldsymbol{e}}_{ss} \neq 0$ if $\boldsymbol{w} \neq 0$.*

Proof. Assume $\sup_k \|\bar{\boldsymbol{A}}(k)\| \le M < \infty$, $K_P = pI$ with $p > 0$. Since $M_P(k) = \bar{\boldsymbol{A}}(k)(I - K_P) = \bar{\boldsymbol{A}}(k)(1-p)I$, by sub-multiplicative property of matrix norm we have

$$\|M_P(t)\| \le \|\bar{A}(t)\|\;\|(1-p)I\| \le M|1-p| =: q. \tag{45}$$

For $p \in \left(1 - \frac{1}{M}, 1 + \frac{1}{M}\right)$, we have $q < 1$.

Expanding recursively,

$$\bar{e}(k) = M_P(k-1) \cdots M_P(0)\bar{e}(0) \tag{46}$$

$$+ \sum_{j=0}^{k-1} M_P(k-1) \cdots M_P(j+1) w(j). \tag{47}$$

Hence,

$$\|\bar{e}(k)\| \leq q^k \|\bar{e}(0)\| + \sum_{j=0}^{k-1} q^{k-1-j} \|w(j)\| \tag{48}$$

$$\leq q^k \|\bar{e}(0)\| + \frac{1-q^k}{1-q} \|w\|_\infty \tag{49}$$

$$\leq q^k \|\bar{e}(0)\| + \frac{1}{1-q} \|w\|_\infty. \tag{50}$$

Since $q < 1$, we can set $\beta(s, k) = q^k s$, which is a $\mathcal{KL}$-function (decaying to zero as $k \to \infty$), and $\gamma(s) = \frac{1}{1-q} s$, which is a $\mathcal{K}$-function, satisfying Eqn. 44. Therefore, the system is ISS.

However, there exists a steady-state error due to the disturbance $\boldsymbol{w}(k)$. In the best case, when $\bar{\boldsymbol{A}}(k)$ converges to $\bar{\boldsymbol{A}}$ and $\boldsymbol{w}(k)$ converges to $\boldsymbol{w}$, the error $\bar{\boldsymbol{e}}(k)$ eventually converges to a steady state given by

$$\bar{\boldsymbol{e}}_{ss} = (I - \bar{\boldsymbol{A}}(1 - pI))^{-1} \boldsymbol{w}.$$

Therefore, $\bar{\boldsymbol{e}}_{ss} \neq 0$ if $\boldsymbol{w} \neq 0$. □

Remark 1 (Convergence rate versus K_P.) *From Ineq. 48, smaller q yields faster convergence. Because*

$$q(p) = M|1-p| = \begin{cases} M(1-p), & p \in \left(1 - \frac{1}{M}, 1\right), \\ M(p-1), & p \in \left[1, 1 + \frac{1}{M}\right), \end{cases}$$

we have $\frac{d}{dp}q(p) = -M < 0$ *for* $p < 1$ *and* $\frac{d}{dp}q(p) = M > 0$ *for* $p > 1$. *Therefore the contraction factor* $q(p)$ *is minimized at*

$$p^\star = 1 \quad \Longrightarrow \quad q^\star = 0,$$

and increases as p moves away from 1 *within the admissible interval.*

Proportional-Integral (PI) Control

To reduce the steady-state error, the proportional controller is extended with an integral action, resulting in a proportional-integral (PI) control law:

$$\boldsymbol{u}(k) = K_p \bar{\boldsymbol{e}}(k) + K_I \boldsymbol{s}(k), \boldsymbol{s}(k+1) = \boldsymbol{s}(k) + \bar{\boldsymbol{e}}(k), (K_D = 0). \tag{51}$$

The dynamics Eqn. 25 then become

$$\bar{\boldsymbol{e}}(k+1) = \bar{\boldsymbol{A}}(k)(I - K_P)\bar{\boldsymbol{e}}(k) - \bar{\boldsymbol{A}}(k) K_I \boldsymbol{s}(k) + \boldsymbol{w}(k). \tag{52}$$

We use the following orthogonal decomposition for $\boldsymbol{w}(k)$:

$$\boldsymbol{w}(k) = \boldsymbol{w}^{\|}(k) + \boldsymbol{w}^{\perp}(k),$$

where $\boldsymbol{w}^{\|}(k) \in \operatorname{Im} \bar{\boldsymbol{A}}(k)$ and $\boldsymbol{w}^{\perp}(k) \in (\operatorname{Im} \bar{\boldsymbol{A}}(k))^{\perp}$.

The impact of $\boldsymbol{w}^{\|}(k)$ on the error can be eliminated by PI control, as discussed below. On the other hand, P-only control is not able to do so, because keeping $\bar{\boldsymbol{e}}(k) = 0$ requires $\boldsymbol{u}(k) = 0$, leaving no component in $\boldsymbol{u}(k)$ that can compensate for $\boldsymbol{w}^{\|}(k)$.

Since $\boldsymbol{w}^{\|}(k) \in \operatorname{Im} \bar{\boldsymbol{A}}(k)$, it can be expressed as

$\boldsymbol{w}^{\|}(k) = \bar{\boldsymbol{A}}(k) K_I s^*(k). \iff \boldsymbol{s}^*(k) = K_I^{-1} \bar{\boldsymbol{A}}(k)^\dagger \boldsymbol{w}^{\|}(k)$

Let $\tilde{\boldsymbol{s}}(k) = \boldsymbol{s}(k) - \boldsymbol{s}^*(k)$ and $d(k) = s^*(k+1) - s^*(k)$. Therefore,

$$\tilde{\boldsymbol{s}}(k+1) = \tilde{\boldsymbol{s}}(k) + \bar{\boldsymbol{e}}(k) - \boldsymbol{d}(k) \tag{53}$$

Insert $\boldsymbol{s}(k) = \boldsymbol{s}^*(k) + \tilde{\boldsymbol{s}}(k)$ and $\boldsymbol{w}(k) = \boldsymbol{w}^{\|}(k) + \boldsymbol{w}^{\perp}(k) = \bar{\boldsymbol{A}}(k) K_I s^*(k) + \boldsymbol{w}^{\perp}(k)$ into Eqn. 52,

$$\begin{aligned} \bar{\boldsymbol{e}}(k+1) &= \bar{\boldsymbol{A}}(k)(I - \boldsymbol{K}_p)\bar{\boldsymbol{e}}(k) - \bar{\boldsymbol{A}}(k)\boldsymbol{K}_i \boldsymbol{s}^*(k) \\ &\quad - \bar{\boldsymbol{A}}(k)\boldsymbol{K}_i \tilde{\boldsymbol{s}}(k) + \bar{\boldsymbol{A}}(k)\boldsymbol{K}_i s^*(k) + \boldsymbol{w}^{\perp}(k) \\ &= \bar{\boldsymbol{A}}(k)(I - \boldsymbol{K}_p)\bar{\boldsymbol{e}}(k) - \bar{\boldsymbol{A}}(k)\boldsymbol{K}_i \tilde{\boldsymbol{s}}(k) + \boldsymbol{w}^{\perp}(k) \end{aligned} \tag{54}$$

We introduce the lifted state $\tilde{\zeta}_{\mathrm{PI}}(k) = \begin{bmatrix} \bar{\boldsymbol{e}}(k) \\ \tilde{\boldsymbol{s}}(k) \end{bmatrix}$ with its dynamic derived from Eqn. 53-54 as follow

$$\tilde{\zeta}_{\mathrm{PI}}(k+1) \;=\; \boldsymbol{M}_i(t)\, \tilde{\zeta}_{\mathrm{PI}}(k) \;+\; \tilde{\boldsymbol{w}}_{\mathrm{PI}}(k), \tag{55}$$

where

$$\boldsymbol{M}_i(k) = \begin{bmatrix} \boldsymbol{M}_p(k) & -\boldsymbol{G}(k) \\ I & I \end{bmatrix},$$

with $\boldsymbol{M}_p(k) = \bar{\boldsymbol{A}}(p)(I - \boldsymbol{K}_p)$, $\boldsymbol{G}(k) = \bar{\boldsymbol{A}}(k)\boldsymbol{K}_i$ and

$$\tilde{\boldsymbol{w}}_{\mathrm{PI}}(k) = \begin{bmatrix} \boldsymbol{w}^{\perp}(k) \\ -d(k) \end{bmatrix},$$

Proposition 3 (Stabilizing the PI loop) *Let* $\boldsymbol{M}_p(k) = \bar{\boldsymbol{A}}(k)(\boldsymbol{I} - \boldsymbol{K}_p)$, *and denote* $\|\boldsymbol{K}_i\| =: h$. *Assume* $\sup_k \|\bar{\boldsymbol{A}}(k)\| \leq M < \infty$ *and* $\sup_k \|\boldsymbol{M}_p(k)\| \leq q < 1$. *If* $q + Mh < 1$, *then the PI closed-loop control is ISS. Furthermore, the integral part exactly cancels the matched disturbance component* $\boldsymbol{w}^{\|}$. *The remaining error is due only to the unmatched component* $\boldsymbol{w}^{\perp}$, *which cannot be compensated. Full proof and term explanations provided in Appendix*

Proof. Using the sub-multiplicativity of the induced matrix norm and the triangle inequality, and noting that $\|\boldsymbol{M}_p(k)\| \leq q$, $\|\boldsymbol{G}(k)\| = \|\bar{\boldsymbol{A}}(k)\boldsymbol{K}_i\| \leq \|\bar{\boldsymbol{A}}(k)\|\, \|K_I\| \leq Mh$, we obtain

$$\begin{aligned} \|\bar{\boldsymbol{e}}(k+1)\| &\leq \|\boldsymbol{M}_p(k)\|\, \|\bar{\boldsymbol{e}}(k)\| + \|\boldsymbol{G}(k)\|\, \|\tilde{\boldsymbol{s}}(k)\| + \|\boldsymbol{w}^{\perp}(k)\| \\ &\leq q\, \|\bar{\boldsymbol{e}}(k)\| + Mh\, \|\tilde{\boldsymbol{s}}(k)\| + \|\boldsymbol{w}^{\perp}(k)\|, \end{aligned} \tag{56}$$

$$\|\tilde{\boldsymbol{s}}(k+1)\| \leq \|\bar{\boldsymbol{e}}(k)\| + \|\tilde{\boldsymbol{s}}(k)\| + \|d(k)\|. \tag{57}$$

Introduce

$$z(k) := \begin{bmatrix} \|\bar{\boldsymbol{e}}(k)\| \\ \|\tilde{\boldsymbol{s}}(k)\| \end{bmatrix}, \tag{58}$$

$$H := \begin{bmatrix} q & Mh \\ 1 & 1 \end{bmatrix}, \tag{59}$$

$$v(k) := \begin{bmatrix} \|\boldsymbol{w}^{\perp}(k)\| \\ \|d(k)\| \end{bmatrix}. \tag{60}$$

Then Eqn. 56- 57 give the comparison system

$$z(k+1) \ \le \ H\,z(k) + v(k). \tag{61}$$

Expanding Eqn. 61 recursively yields

$$z(k) \ \le \ H^k z(0) + \sum_{i=0}^{k-1} H^{\,k-1-i} v(i). \tag{62}$$

Consider the characteristic equation of H:

$$(\lambda - q)(\lambda - 1) - Mh = 0 \iff \lambda^2 - (q+1)\lambda + (q+Mh) = 0.$$

Since $q + Mh < 1$, the maximal root $\lambda^\star$ satisfies $\lambda^\star < 1$, hence the spectral radius $\rho(H) < 1$.

Let $r := \rho(H) < 1$ be the spectral radius of H. By the Gelfand formula for induced (operator) norms,

$$\lim_{k\to\infty} \|H^k\|^{1/k} \ = \ r \qquad \text{(Horn and Johnson 2012, p. 349).}$$

Fix any $\rho \in (r, 1)$. Then, by the definition of the limit, there exists $N \in \mathbb{N}$ such that

$$\|H^k\|^{1/k} \ \le \ \rho \quad \text{for all } k \ge N \ \implies \ \|H^k\| \ \le \ \rho^k \quad \forall\, k \ge N.$$

Define the constant

$$C \ := \ \max\Big\{1, \ \max_{0 \le k \le N} \|H^k\|\,\rho^{-k}\Big\}.$$

Then:

- If $k \ge N$, we have $\|H^k\|\,\rho^{-k} \le 1 \le C$, hence $\|H^k\| \le C\,\rho^k$.
- If $0 \le k \le N$, we have $\|H^k\|\,\rho^{-k} \le \max_{0 \le k \le N} \|H^k\|\,\rho^{-k} \le C$, hence $\|H^k\| \le C\,\rho^k$.

Therefore,

$$\|H^k\| \ \le \ C\,\rho^k \qquad \text{for all } k \ge 0. \tag{63}$$

Applying Eqn. 63 to Eqn. 62 gives

$$\begin{aligned} \|z(k)\| &\le \|H^k\|\,\|z(0)\| + \sum_{i=0}^{k-1} \|H^{k-1-i}\|\,\|v(i)\| \\ &\le C\rho^k \|z(0)\| + C \sum_{i=0}^{k-1} \rho^{k-1-i} \|v(i)\| \\ &\le C\rho^k \|z(0)\| + \frac{C}{1-\rho}\,\|v\|_\infty, \end{aligned} \tag{64}$$

where $\|v\|_\infty := \sup_{i \ge 0} \|v(i)\|$.

By construction,

$$\|\tilde{\zeta}_{PI}(k)\| = \left\| \begin{bmatrix} \bar{\boldsymbol{e}}(k) \\ \tilde{\boldsymbol{s}}(k) \end{bmatrix} \right\| = \big(\|\bar{\boldsymbol{e}}(k)\|^2 + \|\tilde{\boldsymbol{s}}(k)\|^2\big)^{1/2} = \|z(k)\|.$$

Combining this identity with Eqn. 64, the ISS estimate follows with

$$\beta(s,k) := C\rho^k\,s \ \in \ \mathcal{KL}, \qquad \gamma(s) := \frac{C}{1-\rho}\,s \ \in \ \mathcal{K},$$

which proves that the PI closed loop Eqn. 55 is ISS. □

The integral part exactly cancels the matched disturbance component $\boldsymbol{w}^{\parallel}$. The remaining error is due only to the unmatched component $\boldsymbol{w}^{\perp}$, which cannot be compensated, and to the variation rate $d(k)$ when $\bar{\boldsymbol{A}}(k)$ and $\boldsymbol{w}(k)$ change over time. In the best scenario, if a steady state exists, i.e., $\bar{\boldsymbol{A}}(k) \to \bar{\boldsymbol{A}}$ and $\boldsymbol{w}(k) \to \boldsymbol{w}$ with $\boldsymbol{w} \in \operatorname{Im} \bar{\boldsymbol{A}}$, then $\boldsymbol{w}^{\perp} \equiv 0$, $d \equiv 0$, and thus $\bar{\boldsymbol{e}}(k) \to 0$.

Remark 2 (Convergence rate versus K_i) *From proposition 3, the convergence rate of $\tilde{\zeta}_{PI}(t)$ depends on ρ: the smaller ρ, the faster the convergence. We also adopt the convention (as in the proof) that $\rho \in (r, 1)$. Equivalently, we examined $r(h) = \rho(H)$ and proved that with $h = \frac{(1-q)^2}{4M}$ this quantity is minimized.*

Proof Consider the characteristic polynomial of H:

$$\lambda^2 - (q+1)\lambda + (q+Mh) = 0.$$

Its discriminant is

$$\Delta(h) = (q+1)^2 - 4(q+Mh) = (q-1)^2 - 4Mh.$$

If $\Delta(h) \ge 0$ (i.e., $0 \le h \le \frac{(1-q)^2}{4M}$), then

$$r(h) = \frac{q+1+\sqrt{\Delta(h)}}{2},$$

and $r(h)$ decreases as h increases.

If $\Delta(h) < 0$ (i.e., $\frac{(1-q)^2}{4M} < h < \frac{1-q}{M}$), then

$$r(h) = \sqrt{q+Mh},$$

and $r(h)$ decreases as h decreases.

Hence $r(h)$ can achieve its best (smallest) value at

$$h = \frac{(1-q)^2}{4M}, \tag{65}$$

for which the error converges to zero the fastest. □

Nevertheless, as we discuss in the next section, in some situations such a large value of h may become a practical obstacle for PI control.

Oveshoot Mechanism

Phenomenon. A common issue in standard PI settings is *overshooting*: the closed loop oscillates around the setpoint before settling (see Åström and Hägglund (1995b, Ch. 3, §3.3)). In our terms, the integral part accumulates past error and can push the output beyond the setpoint; subsequent sign changes of the error gradually "discharge" the integral, producing a decaying oscillation. The big overshoot is undesirable when we prefer a more stable response near zero. Below we analyze the same mechanism for our PI steering setting.

Citing an observation from Vu and Nguyen (2025): in the *absence* of steering, the cosine similarity between error vectors at different layers is consistently *positive*, i.e.,

$$\cos\angle\big(\bar{\boldsymbol{e}}(i),\, \bar{\boldsymbol{e}}(j)\big) > 0 \quad \text{for all layers } i, j,$$

so the layerwise errors share (approximately) the same direction. Consequently, with $\{\bar{\boldsymbol{x}}^{+}(k)\}$ serving as the trajectory setpoints and $\{\bar{\boldsymbol{x}}^{-}(k)\}$ the system output, an *overshoot event* occurs when the instantaneous error reverses its initial orientation, namely when

$$\langle \bar{\boldsymbol{e}}(k),\, \bar{\boldsymbol{e}}(0) \rangle < 0.$$

We now introduce the definitions used below.

Scalarization along a direction. Let

$$v \;:=\; \frac{\bar{e}(0)}{\|\bar{e}(0)\|} \tag{66}$$

and project onto v: (67)

$$\boldsymbol{e}_v(k) := v^\top \bar{\boldsymbol{e}}(k), \quad \boldsymbol{s}_v(k) := v^\top \tilde{\boldsymbol{s}}(k). \tag{68}$$

From the PI loop dynamics Eqn. 55 we obtain the scalar PI pair

$$\boldsymbol{e}_v(k+1) = a(k)\,\boldsymbol{e}_v(k) \;-\; b(k)\,\boldsymbol{s}_v(k) \;+\; \boldsymbol{w}_v^\perp(k), \tag{69}$$

$$\boldsymbol{s}_v(k+1) = \boldsymbol{s}_v(k) \;+\; \boldsymbol{e}_v(k) \;-\; d_v(k), \tag{70}$$

with $a(k) = v^\top \bar{\boldsymbol{A}}(k)(I - \boldsymbol{K}_p)v = v^\top \boldsymbol{M}_p(k)v$, $b(k) = v^\top \bar{\boldsymbol{A}}(k)\boldsymbol{K}_i v = v^\top G(k)v$, and projected disturbances $\boldsymbol{w}_v^\perp(k) := v^\top \boldsymbol{w}^\perp(k)$, $d_v(k) := v^\top d(k)$. Empirically (and consistently with the angular-steering observation in our setup), we have $v^\top \bar{\boldsymbol{A}}(k)v \geq 0$ for all k; together with the gain $\boldsymbol{K}_i = hI$ with $h \geq 0$, this implies $b(k) \geq 0$.

Also, the assumption $q := \sup_k \|\boldsymbol{M}_p(k)\|$ and $M := \sup_k \|\bar{\boldsymbol{A}}(k)\|$ yeilds

$$a(k) \leq q < 1, \qquad 0 \leq b(k) \leq Mh, \tag{71}$$

Since the system Eqn. 55 is ISS, so is the system Eqn. 69-Eqn. 70. In other words, both $\boldsymbol{e}_v(k)$ and $\boldsymbol{s}_v(k)$ decay. Recall that

$$\boldsymbol{s}_v(k) = \sum_{i=0}^{k-1} \boldsymbol{e}_v(i).$$

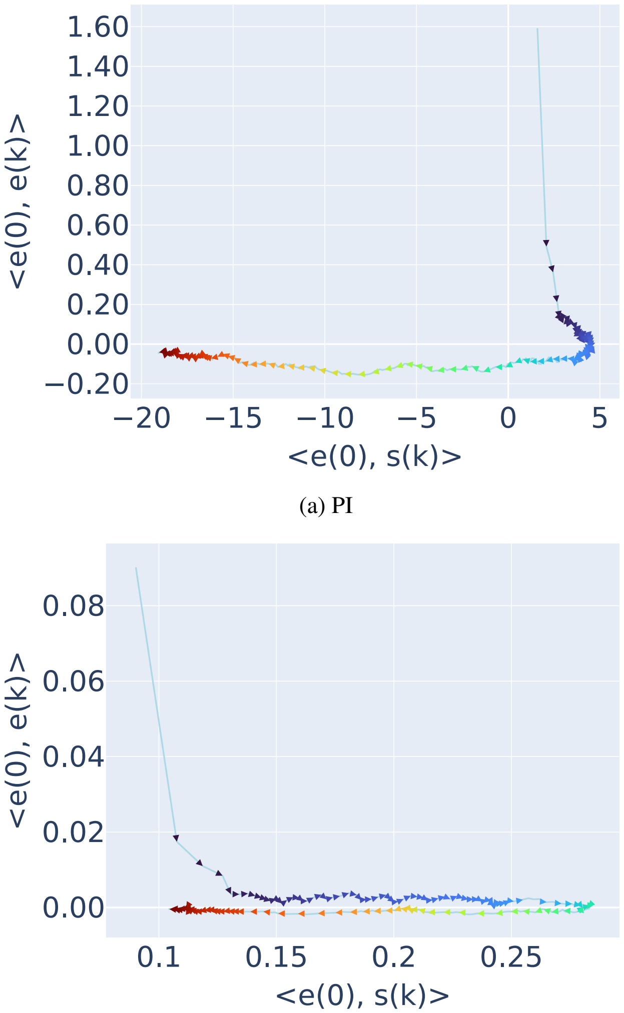
Hence, $\boldsymbol{s}_v(k)$ can only decrease when $\boldsymbol{e}_v(k) < 0$, which is precisely the moment when overshoot occurs. These overshooting and decaying phenomena are observed in empirical simulation, see Fig. 5. Below, we define the overshoot in our setting.

Definition 2 (Overshoot and its amplitude) *We say an overshoot occurs from time k_a to $k_a + m$ if $\boldsymbol{e}_v(k) < 0$ $\forall k = k_a, k_a + 1, ..., k_a + m - 1$ and $\boldsymbol{e}_v(k) \geq 0$ for $k = k_a - 1, k_a + m$. Its amplitude is defined as*

$$A_a \;:=\; \max_{kt_a \leq i \leq k_a+m-1} |\boldsymbol{e}_v(i)|\,. \tag{72}$$

In standard PID settings illustrated in Åström and Hägglund (1995b, Ch. 3, §3.3)), it is observed that the overshoot amplitude decays over time. This decay is also consistent with the ISS property of the closed loop: as both $\boldsymbol{e}_v(t)$ and $\boldsymbol{s}_v(t)$ are driven down, subsequent oscillations tend to diminish in magnitude. In our simulation (see Fig. 5), the first overshoot appear to be representative. Hence, while we are not yet able to provide a formal proof, the empirical evidence and ISS intuition justify the first overshoot which is typically the dominant one and serves as a representative indicator of oscillatory behavior.

This assumption is for Proposition 4. Suppose that $a(t) \geq 0$; equivalently, $p \in (1 - \frac{1}{M}, 1]$, which is the result of proposition 1. This assumption is expected to entail no loss of generality relative to the $|a(t)| \leq q < 1$ assumption.



(a) PI

(b) PID

Figure 5: Scalar errors across time step of randomly initialized model after applying PI and PID controller. Colors from blue to red denote the time (layer) dimension.

Proposition 4 (Agressive PI gain leads to a large first overshoot) *Let k_0 be the first sign-change time, i.e.,*

$$\boldsymbol{e_v}(j) \geq 0 \;\; \forall\, j = 0, 1, \ldots, k_0 - 1, \qquad \boldsymbol{e_v}(k_0) < 0,$$

and let k_1 be the first time the trajectory returns to the non-negative side,

$$\boldsymbol{e_v}(k_1) \geq 0, \qquad \boldsymbol{e_v}(i) < 0 \;\; \forall\, i = k_0, k_0 + 1, \ldots, k_1 - 1.$$

As in Eqn.72, the first overshoot amplitude is

$$A_0 \;=\; \max_{k_0 \leq i \leq k_1 - 1} |\boldsymbol{e_v}(i)| \;=\; |\boldsymbol{e_v}(i_{max})|. \tag{73}$$

Assume $\sup_k \|\bar{\boldsymbol{A}}(k)\| \leq M < \infty$ and $\sup_k \|\boldsymbol{M}_p(k)\| \leq q < 1$. Denote $\|\boldsymbol{K}_i\| =: h$. and given $q + Mh < 1$. There-

fore,

$$A_0 \;\le\; Mh\Big(\frac{1}{1-q}+\frac{1}{(1-q)^2}\Big)\,\boldsymbol{e_v}(0) \tag{74}$$

$$+\Big(\frac{Mh}{1-q}(k_0-1)+\frac{Mh}{1-q}\Big)\|\boldsymbol{d}\|_\infty \tag{75}$$

$$+\,\frac{Mh(k_0-1)+1}{1-q}\|\boldsymbol{w}\|_\infty; \tag{76}$$

Proof. Before the first crossing ($j \le k_0-1$) we have $\boldsymbol{s}_v(j)\ge 0$, hence from Eqn. 69

$$\boldsymbol{e}_v(j+1) = a(j)\boldsymbol{e}_v(j) - b(j)\boldsymbol{s}_v(j) + \boldsymbol{w}_v^{\perp}(j) \tag{77}$$

$$\le\; a(j)\boldsymbol{e}_v(j) + \boldsymbol{w}_v^{\perp}(j) \tag{78}$$

$$\le\; q\,\boldsymbol{e}_v(j) + \|\boldsymbol{w}\|_\infty, \tag{79}$$

so by induction $\boldsymbol{e}_v(j)\le q^j\boldsymbol{e}_v(0)+\frac{1}{1-q}\|\boldsymbol{w}\|_\infty$. Summing Eqn. 70,

$$\boldsymbol{s}_v(k_0-1) \;=\; \sum_{i=0}^{k_0-2}\boldsymbol{e}_v(i) \;-\; \sum_{i=0}^{k_0-2}\boldsymbol{d}_v(i) \tag{80}$$

$$\le\; \frac{\boldsymbol{e}_v(0)}{1-q} \;+\; (k_0-1)\|\boldsymbol{d}\|_\infty + \frac{k_0-1}{1-q}\|\boldsymbol{w}\|_\infty. \tag{81}$$

Since $a(k)\ge 0$ and $\boldsymbol{e}_v(k_0-1)\ge 0$, at the crossing step,

$$\begin{aligned}
|\boldsymbol{e}_v(k_0)| \;&=\; -\boldsymbol{e}_v(k_0)\\
&\le\; b(k_0-1)\,\boldsymbol{s}_v(k_0-1) \;+\; \|\boldsymbol{w}\|_\infty\\
&\le\; Mh\bigg(\frac{\boldsymbol{e}_v(0)}{1-q}+(k_0-1)\|\boldsymbol{d}\|_\infty+\frac{k_0-1}{1-q}W_\infty\bigg)+\|\boldsymbol{w}\|_\infty.\\
&= \frac{Mh}{1-q}\boldsymbol{e}_v(0)+Mh(k_0-1)\|\boldsymbol{d}\|_\infty\\
&+(Mh\frac{k_0-1}{1-q}+1)\|\boldsymbol{w}\|_\infty
\end{aligned} \tag{82}$$

Assume that $\boldsymbol{d}_v(k)$ is small enough s.t during the overshoot time, $\boldsymbol{s}_v$ is nonincreasing (since $\boldsymbol{e}_v<0$ a.e. on $[k_0,k_1-1]$), so $\boldsymbol{s}_v(i)\le\boldsymbol{s}_v(k_0-1)$ for $i\in[k_0,k_1-1]$. Using Eqn. 69 again and unrolling m steps from k_0,

$$\begin{aligned}
|\boldsymbol{e}_v(k_0+m)| \;&\le\; q^m|\boldsymbol{e}_v(k_0)| \;+\; \sum_{k=0}^{m-1}q^k\big(Mh\,\boldsymbol{s}_v(k_0-1)+\|\boldsymbol{w}\|_\infty\big)\\
&\le\; |\boldsymbol{e}_v(k_0)| \;+\; \frac{Mh\,\boldsymbol{s}_v(k_0-1)+\|\boldsymbol{w}\|_\infty}{1-q}.
\end{aligned} \tag{83}$$

Taking the maximum over $m\in\{0,1,\dots,k_1-k_0\}$ and substituting Ineq. 80 and Ineq. 82 into Ineq. 83 yields

$$A_0 \;\le Mh\Big(\frac{1}{1-q}+\frac{1}{(1-q)^2}\Big)\,\boldsymbol{e}_v(0) \tag{84}$$

$$+\;\Big(\frac{Mh}{1-q}(k_0-1)+\frac{Mh}{1-q}\Big)\|\boldsymbol{d}\|_\infty \tag{85}$$

$$+\,\frac{Mh(k_0-1)+1}{1-q}\|\boldsymbol{w}\|_\infty \tag{86}$$

□

Consequently, the right-hand side of Ineq.84 is *monotone increasing in* h (via the factor Mh) and increases as q decreases (through the factors $\frac{1}{1-q}$). In particular, more aggressive PI leads to a larger first–overshoot amplitude.

Remark 3 ("Fast-PI" specialization.) *With the tuning used in our analysis in remark 2, $h=\frac{(1-q)^2}{4M}$ (the value that minimizes the comparison-system rate), so $Mh=\frac{(1-q)^2}{4}$. Plugging into equation 84 gives*

$$\boldsymbol{A}_0 \;\le \Big(\frac{1-q}{4}+\frac{1}{4}\Big)\,\boldsymbol{e}_v(0) \tag{87}$$

$$+\;\Big(\frac{1-q}{4}(k_0-1)+\frac{1}{4}\Big)\|\boldsymbol{d}\|_\infty \;+\; \frac{k_0}{1-q}\|\boldsymbol{w}\|_\infty. \tag{88}$$

In particular, in the disturbance-free case ($\|\boldsymbol{w}\|_\infty = \|\boldsymbol{d}\|_\infty=0$) we obtain

$$\boldsymbol{A}_0 \;\le\; \Big(\frac{1-q}{4}+\frac{1}{4}\Big)\,\boldsymbol{e}_v(0),$$

so stronger proportional action (smaller q) comes with a larger first-overshoot envelope, even though the closed-loop settles faster.

PID Control

Stability of PID closed loop We consider the PID update $\boldsymbol{u}(k)\;=\;\boldsymbol{K}_p\,\bar{\boldsymbol{e}}(k)\;+\;\boldsymbol{K}_i\,\boldsymbol{s}(k)\;+\;\boldsymbol{K}_d\big(\bar{\boldsymbol{e}}(k)-\bar{\boldsymbol{e}}(k-1)\big)$, and define the auxiliary matrices

$$\boldsymbol{M}_p(k) := \bar{\boldsymbol{A}}(k)\big(I-\boldsymbol{K}_p\big), \tag{89}$$

$$\boldsymbol{G}(k) := \bar{\boldsymbol{A}}(k)\boldsymbol{K}_i, \tag{90}$$

$$\boldsymbol{H}(k) := \bar{\boldsymbol{A}}(k)\boldsymbol{K}_d, \tag{91}$$

together with the error increment

$$\Delta\bar{\boldsymbol{e}}(k) := \bar{\boldsymbol{e}}(k)-\bar{\boldsymbol{e}}(k-1), \qquad \Delta\bar{\boldsymbol{e}}(-1)=0$$

Using the plant relation, we obtain

$$\begin{aligned}
\Delta\bar{\boldsymbol{e}}(k+1) \;&=\; \big(\boldsymbol{M}_p(k)-I\big)\bar{\boldsymbol{e}}(k)\\
&-\;\boldsymbol{G}(k)\tilde{\boldsymbol{s}}(k)\;-\;\boldsymbol{H}(k)\Delta\bar{\boldsymbol{e}}(k)\;+\;\boldsymbol{w}^{\perp}(k),
\end{aligned} \tag{92}$$

Introduce the lifted state from the auxiliary PI state in Eqn. 55

$$\tilde{\zeta}_{\mathrm{PID}}(k) \;:=\; \begin{bmatrix}\bar{\boldsymbol{e}}(k)\\ \tilde{\boldsymbol{s}}(k)\\ \Delta\bar{\boldsymbol{e}}(k)\end{bmatrix},$$

Then the closed-loop evolution reads

$$\tilde{\zeta}_{\mathrm{PID}}(k+1) \;=\; \boldsymbol{M}_d(k)\,\tilde{\zeta}_{\mathrm{PID}}(k) \;+\; \tilde{\boldsymbol{w}}_{\mathrm{PID}}(k), \tag{93}$$

where

$$\boldsymbol{M}_d(k) \;:=\; \begin{bmatrix}\boldsymbol{M}_p(k) & -\boldsymbol{G}(k) & -\boldsymbol{H}(k)\\ I & I & 0\\ \boldsymbol{M}_p(k)-I & -\boldsymbol{G}(k) & -\boldsymbol{H}(k)\end{bmatrix}, \tag{94}$$

$$\tilde{\boldsymbol{w}}_{\mathrm{PID}}(k) \;:=\; \begin{bmatrix}\boldsymbol{w}^{\perp}(k)\\ -d(k)\\ \boldsymbol{w}^{\perp}(k)\end{bmatrix}. \tag{95}$$

Theorem 1 (Stabilizing the PID loop) *Let $\boldsymbol{M}_p(k) = \bar{\boldsymbol{A}}(k)(\boldsymbol{I}-\boldsymbol{K}_p)$, and denote $\|\boldsymbol{K}_i\| =: h$, $\|\boldsymbol{K}_d\| =: \ell$. Assume $\sup_k \|\bar{\boldsymbol{A}}(k)\| \le M < \infty$ and $\sup_k \|\boldsymbol{M}_p(k)\| \le q < 1$. If $q + Mh < 1$ (stable PI loop), then there exists $\ell > 0$ such that the PID closed-loop control is ISS. Therefore, the integral part in PID design still cancels the matched disturbance component $\boldsymbol{w}^{\parallel}$.*

Proof. We establish the ISS for system 93 using the method of ISS-Lyapunov function, see (Jiang, Sontag, and Wang 1999, Def. 2.2, Prop. 2.3). It then suffices to construct a candidate ISS-Lyapunov function $\boldsymbol{V}_{\text{PID}}(k)$ satisfying that there exist class $\mathcal{K}_\infty$ functions $\alpha_1, \alpha_2, \alpha_3$ and a class $\mathcal{K}$ function σ such that

$$\alpha_1(\|\tilde{\zeta}_{\text{PID}}(k)\|) \le \boldsymbol{V}_{\text{PID}}(\tilde{\zeta}_{\text{PID}}(k)) \le \alpha_2(\|\tilde{\zeta}_{\text{PID}}(k)\|), \tag{96}$$

and

$$\boldsymbol{V}(\tilde{\zeta}_{\text{PID}}(k+1)) - \boldsymbol{V}(\tilde{\zeta}_{\text{PID}}(k)) \le -\alpha_3(\|\tilde{\zeta}_{\text{PID}}(k)\|) + \sigma(\|\boldsymbol{w}\|). \tag{97}$$

Step 1: Candidtate $\boldsymbol{V}_{PID}(k)$ and PI-closed loop baseline

Define

$$\boldsymbol{V}_{\text{PID}}(k) := \boldsymbol{V}_{\text{PI}}\big(\tilde{\zeta}_{\text{PI}}(k), k\big) + r\,\|\Delta\bar{\boldsymbol{e}}(k)\|^2, \qquad r > 0, \tag{98}$$

where $\boldsymbol{V}_{\text{PI}}(\tilde{\zeta}_{\text{PI}}, k) = \tilde{\zeta}_{\text{PI}}^\top \boldsymbol{P}(k)\tilde{\zeta}_{\text{PI}}$ with $\boldsymbol{P}(k) = \boldsymbol{P}(k)^\top \succ 0$ and there exists some $\mu_{\text{PI}} > 0$ such that

$$\boldsymbol{M}_i(k)^\top \boldsymbol{P}(k)\,\boldsymbol{M}_i(k) - \boldsymbol{P}(k) \le -\mu_{\text{PI}} I, \qquad \forall k, \tag{99}$$

Regarding the existence of such $\boldsymbol{V}_{\text{PI}}$, recall the homogeneous PI-loop $\tilde{\zeta}_{\text{PI}}(k+1) = \boldsymbol{M}_i(k)\tilde{\zeta}_{\text{PI}}(k)$ with $\boldsymbol{M}_i(k) = \begin{bmatrix} \boldsymbol{M}_p(k) & -\boldsymbol{G}(k) \\ I & I \end{bmatrix}$, being asymptotically stable. Suppose there is $\boldsymbol{Q}(k) = \boldsymbol{Q}(k)^\top \succeq 0$ bounded so that the pair $(\boldsymbol{M}_i(k), \sqrt{\boldsymbol{Q}(k)})$ is observable for all k, hence the difference Lyapunov equation

$$\boldsymbol{M}_i^\top(k)\boldsymbol{P}(k+1)\boldsymbol{M}_i(k) - \boldsymbol{P}(k) = -\boldsymbol{Q}(k)$$

admits a unique positive definite solution $\boldsymbol{P}(k) = \boldsymbol{P}^\top(k) \succ 0$ for all k, and a uniform bound $\|\boldsymbol{P}\|_\infty := \sup_k \|\boldsymbol{P}(k)\| < \infty$ (see (Gajic and Qureshi 2008, Ch. 4, p. 110))

Step 2: Condition (i) as in Eqn. 96

We write $\boldsymbol{V}_{\text{PID}}$ as a quadratic form

$$\boldsymbol{V}_{\text{PID}}(k) = \tilde{\zeta}_{\text{PID}}(k)^\top\,\boldsymbol{P}_*(k)\,\tilde{\zeta}_{\text{PID}}(k), \tag{100}$$

$$\boldsymbol{P}_*(k) := \begin{bmatrix} \boldsymbol{P}(k) & 0 \\ 0 & rI \end{bmatrix}. \tag{101}$$

Clearly $\boldsymbol{P}_*(k) = \boldsymbol{P}_*(k)^\top$ and $\boldsymbol{P}_*(k) \succ 0$ because $\boldsymbol{P}(k) \succ 0$ and $rI \succ 0$; hence $\boldsymbol{P}_*(k)$ is symmetric positive definite for all k.

By the spectral theorem, there exists an orthogonal matrix $U_*(k)$ and a diagonal $\Lambda_*(k) = \text{diag}(\lambda_1(k), \ldots, \lambda_{n_*}(k))$ with positive entries such that $\boldsymbol{P}_*(k) = U_*(k)\Lambda_*(k)U_*(k)^\top$. Moreover, because $\boldsymbol{P}_*(k)$ is block diagonal, its eigenvalues are precisely the union of the eigenvalues of $\boldsymbol{P}(k)$ and the repeated eigenvalue r. Using the uniform bounds already established for $\boldsymbol{P}(k)$ (there exists $\underline{\lambda} > 0$ with $\lambda_{\min}(\boldsymbol{P}(k)) \ge \underline{\lambda}$ and $\lambda_{\max}(\boldsymbol{P}(k)) \le \|\boldsymbol{P}\|_\infty := \sup_k \|\boldsymbol{P}(k)\| < \infty$), we obtain the k-independent bounds

$$\lambda_{\min}\big(\boldsymbol{P}_*(k)\big) \ge \underline{\lambda}_* := \min\{\underline{\lambda}, r\} > 0, \tag{102}$$

$$\lambda_{\max}\big(\boldsymbol{P}_*(k)\big) \le \overline{\lambda}_* := \max\{|\boldsymbol{P}\|_\infty, r\} < \infty. \tag{103}$$

For every vector z and every symmetric positive definite M, $\lambda_{\min}(M)\|z\|^2 \le z^\top M z \le \lambda_{\max}(M)\|z\|^2$. Applying this to $M = \boldsymbol{P}_*(k)$ and $z = \tilde{\zeta}_{\text{PID}}(k)$ in Eqn. 100 yields

$$\underline{\lambda}_*\,\big\|\tilde{\zeta}_{\text{PID}}(k)\big\|^2 \le \boldsymbol{V}_{\text{PID}}(k) \le \overline{\lambda}_*\,\big\|\tilde{\zeta}_{\text{PID}}(k)\big\|^2.$$

Therefore, choosing the class-$\mathcal{K}_\infty$ functions

$$\alpha_1(\boldsymbol{s}) := \underline{\lambda}_*\,\boldsymbol{s}^2, \qquad \alpha_2(\boldsymbol{s}) := \overline{\lambda}_*\,\boldsymbol{s}^2,$$

we obtain the desired bound

$$\alpha_1\Big(\big\|\tilde{\zeta}_{\text{PID}}(k)\big\|\Big) \le \boldsymbol{V}_{\text{PID}}(k) \le \alpha_2\Big(\big\|\tilde{\zeta}_{\text{PID}}(k)\big\|\Big), \tag{104}$$

which establishes condition (i) in 96.

Step 3: Condition (ii) as in Eqn. 97.

From Eqn. 98,

$$\Delta\boldsymbol{V}_{\text{PID}}(k) = \underbrace{\Delta\boldsymbol{V}_{\text{PI}}\Big(\tilde{\zeta}_{\text{PI}}(k)\Big)\Big|_{\text{PID}}}_{\text{PI part under PID update}} \tag{105}$$

$$+\; r\big(\|\Delta\bar{\boldsymbol{e}}(k+1)\|^2 - \|\Delta\bar{\boldsymbol{e}}(k)\|^2\big). \tag{106}$$

Bounding the PI part under the PID update.

Under PI rule ($\boldsymbol{K}_d = 0$),

$$\tilde{\zeta}_{\text{PI}}(k+1)\big|_{\text{PI}} = \boldsymbol{M}_i(k)\,\tilde{\zeta}_{\text{PI}}(k) + \tilde{\boldsymbol{w}}_{\text{PI}}(k), \tag{107}$$

$$\boldsymbol{M}_i(k) := \begin{bmatrix} \boldsymbol{M}_p(k) & -\boldsymbol{G}(k) \\ I & I \end{bmatrix} \tag{108}$$

then

$$\begin{aligned} \Delta\boldsymbol{V}_{\text{PI}}\Big(\tilde{\zeta}_{\text{PI}}(k)\Big)\Big|_{\text{PI}} = {} & \tilde{\zeta}_{\text{PI}}(k)^\top\big(\boldsymbol{M}_i(k)^\top\boldsymbol{P}(k{+}1)\boldsymbol{M}_i(k) \\ & - \boldsymbol{P}(k)\big)\tilde{\zeta}_{\text{PI}}(k) \\ & + 2\,\tilde{\zeta}_{\text{PI}}(k)^\top\boldsymbol{M}_i(k)^\top\boldsymbol{P}(k{+}1)\,\tilde{\boldsymbol{w}}_{\text{PI}}(k) \\ & + \tilde{\boldsymbol{w}}_{\text{PI}}(k)^\top\boldsymbol{P}(k{+}1)\,\tilde{\boldsymbol{w}}_{\text{PI}}(k) \end{aligned}$$

By Eqn. 99, Cauchy-Schwarz inequality and Young's inequality,

$$\Delta\boldsymbol{V}_{\text{PI}}\Big(\tilde{\zeta}_{\text{PI}}(k)\Big)\Big|_{\text{PI}} \le -\,\mu_{\text{PI}}\,\|\tilde{\zeta}_{\text{PI}}(k)\|^2 \tag{109}$$

$$+\,2\|\boldsymbol{M}_i\|_\infty\|\boldsymbol{P}\|_\infty\|\tilde{\zeta}_{\text{PI}}(k)\|\;\|\tilde{\boldsymbol{w}}_{\text{PI}}(k)\| \tag{110}$$

$$+\,\|\boldsymbol{P}\|_\infty\|\tilde{\boldsymbol{w}}_{\text{PI}}(k)\|^2 \tag{111}$$

$$\le -(\mu_{\text{PI}} - \varepsilon_1)\,\|\tilde{\zeta}_{\text{PI}}(k)\|^2 \tag{112}$$

$$+\,\Big(\frac{\|\boldsymbol{M}_i\|_\infty^2\|\boldsymbol{P}\|_\infty^2}{\varepsilon_1} + \|\boldsymbol{P}\|_\infty\Big)\|\tilde{\boldsymbol{w}}_{\text{PI}}(k)\|^2 \tag{113}$$

$$= -\,\mu_{\text{PI}}^*\,\|\tilde{\zeta}_{\text{PI}}(k)\|^2 + C_1\|\tilde{\boldsymbol{w}}_{\text{PI}}(k)\|^2, \tag{114}$$

for any $\varepsilon_1 > 0$.

Under PID rule ($\boldsymbol{K}_d \neq 0$),

$$\tilde{\zeta}_{\mathrm{PI}}(k+1)\big|_{\mathrm{PID}} \;=\; \boldsymbol{M}_i(k)\,\tilde{\zeta}_{\mathrm{PI}}(k) \;-\; \delta(k) \;+\; \boldsymbol{w}_{\mathrm{PI}}(k),$$

with the "perturbation"

$$\delta(k) := \begin{bmatrix} H(k)\,\Delta\bar{\boldsymbol{e}}(k) \\ 0 \end{bmatrix}$$

Hence,

$$\begin{aligned}\Delta\boldsymbol{V}_{\mathrm{PI}}\big(\tilde{\zeta}_{\mathrm{PI}}(k)\big)\Big|_{\mathrm{PID}} &= \boldsymbol{V}_{\mathrm{PI}}\big(\tilde{\zeta}_{\mathrm{PI}}(k+1)\big|_{\mathrm{PI}}\big) - \boldsymbol{V}_{\mathrm{PI}}\big(\tilde{\zeta}_{\mathrm{PI}}(k)\big)\\ &= \boldsymbol{V}_{\mathrm{PI}}\big(\tilde{\zeta}_{\mathrm{PI}}(k+1)\big|_{\mathrm{PI}}\big) - \boldsymbol{V}_{\mathrm{PI}}\big(\tilde{\zeta}_{\mathrm{PI}}(k)\big)\\ &\quad + 2\big(\tilde{\zeta}_{\mathrm{PI}}(k+1)\big|_{\mathrm{PI}}\big)^{\top}\boldsymbol{P}\,\delta(k)\\ &\quad + \delta(k)^{\top}\boldsymbol{P}\,\delta(k)\end{aligned} \tag{115}$$

Bounding each term in Eqn. 115

- $\boldsymbol{V}_{\mathrm{PI}}\big(\tilde{\zeta}_{\mathrm{PI}}(k+1)\big|_{\mathrm{PI}}\big) - \boldsymbol{V}_{\mathrm{PI}}\big(\tilde{\zeta}_{\mathrm{PI}}(k)\big) \leq -\mu^*_{\mathrm{PI}}\|\tilde{\zeta}_{\mathrm{PI}}(k)\|^2 + C_1\|\tilde{\boldsymbol{w}}_{\mathrm{PI}}(k)\|$
- Applying Young's inequality for inner product, there exists $\varepsilon > 0$ s.t

$$2\big(\tilde{\zeta}_{\mathrm{PI}}(k+1)\big|_{\mathrm{PI}}\big)^{\top}\boldsymbol{P}\delta(k) \leq \varepsilon\big\|\tilde{\zeta}_{\mathrm{PI}}(k+1)\big|_{\mathrm{PI}}\big\|^2 \tag{116}$$

$$+ \tfrac{1}{\varepsilon}\delta(k)^{\top}\boldsymbol{P}\delta(k) \tag{117}$$

$$\leq \varepsilon\|\boldsymbol{P}\|\big\|\tilde{\zeta}_{\mathrm{PI}}(k+1)\big|_{\mathrm{PI}}\big\|^2 \tag{118}$$

$$+ \tfrac{\|\boldsymbol{P}\|}{\varepsilon}M^2\ell^2\|\Delta\bar{\boldsymbol{e}}(k)\|^2 \tag{119}$$

Since $\big\|\tilde{\zeta}_{\mathrm{PI}}(k+1)\big|_{\mathrm{PI}}\big\|^2 \leq 2\|\boldsymbol{M}_i\|^2_{\infty}\|\tilde{\zeta}_{\mathrm{PI}}(k)\|^2 + 2\|\tilde{\boldsymbol{w}}_{\mathrm{PI}}(k)\|^2$

$$\begin{aligned}\Rightarrow 2\big(\tilde{\zeta}_{\mathrm{PI}}(k+1)\big|_{\mathrm{PI}}\big)^{\top}\boldsymbol{P}\delta(k) &\leq 2\varepsilon\|\boldsymbol{P}\|\|\boldsymbol{M}_i\|^2_{\infty}\|\tilde{\zeta}_{\mathrm{PI}}(k)\|^2\\ &\quad + 2\varepsilon\|\boldsymbol{P}\|\|\tilde{\boldsymbol{w}}_{\mathrm{PI}}(k)\|^2\\ &\quad + \tfrac{\|\boldsymbol{P}\|}{\varepsilon}M^2\ell^2\|\Delta\bar{\boldsymbol{e}}(k)\|^2\\ &= 2\varepsilon\|\boldsymbol{P}\|\|\boldsymbol{M}_i\|^2_{\infty}\|\tilde{\zeta}_{\mathrm{PI}}(k)\|^2\\ &\quad + \tfrac{\|\boldsymbol{P}\|}{\varepsilon}M^2\ell^2\|\Delta\bar{\boldsymbol{e}}(k)\|^2\\ &\quad + C_2\|\tilde{\boldsymbol{w}}_{\mathrm{PI}}(k)\|^2,\end{aligned}$$

where $C_2 = 2\varepsilon\|\boldsymbol{P}\|$

- $\delta(k)^{\top}\boldsymbol{P}\delta(k) \leq \|\boldsymbol{P}\|\|\delta(k)\|^2 \leq \|\boldsymbol{P}\|M^2\ell^2\|\Delta\bar{\boldsymbol{e}}(k)\|^2$

Therefore,

$$\begin{aligned}\Delta\boldsymbol{V}_{\mathrm{PI}}\big(\tilde{\zeta}_{\mathrm{PI}}(k)\big)\Big|_{\mathrm{PID}} &\leq -\mu^*_{\mathrm{PI}}\|\tilde{\zeta}_{\mathrm{PI}}(k)\|^2 + C_1\|\tilde{\boldsymbol{w}}_{\mathrm{PI}}(k)\|^2\\ &\quad + 2\varepsilon\|\boldsymbol{P}\|\|\boldsymbol{M}_i\|^2_{\infty}\|\tilde{\zeta}_{\mathrm{PI}}(k)\|^2\\ &\quad + \tfrac{\|\boldsymbol{P}\|}{\varepsilon}M^2\ell^2\|\Delta\bar{\boldsymbol{e}}(k)\|^2\\ &\quad + C_2\|\tilde{\boldsymbol{w}}_{\mathrm{PI}}(k)\|^2\\ &\quad + \|\boldsymbol{P}\|M^2\ell^2\|\Delta\bar{\boldsymbol{e}}(k)\|^2\\ &= -(\mu^*_{\mathrm{PI}} - 2\varepsilon\|\boldsymbol{P}\|\|\boldsymbol{M}_i\|^2_{\infty})\|\tilde{\zeta}_{\mathrm{PI}}(k)\|^2\\ &\quad + \|\boldsymbol{P}\|M^2\ell^2\Big(\tfrac{1}{\varepsilon}+1\Big)\|\Delta\bar{\boldsymbol{e}}(k)\|^2\\ &\quad + C_3\|\tilde{\boldsymbol{w}}_{\mathrm{PI}}(k)\|^2,\end{aligned} \tag{120}$$

where $C_3 = C_1 + C_2$.

Bounding the increment term.

From Eqn. 92 and applying the inequality $(x+y+z)^2 \leq 3(x^2+y^2+z^2)$,

$$\begin{aligned}\|\Delta\bar{\boldsymbol{e}}(k+1)\|^2 \;\leq\; & 3\Big(\|\boldsymbol{M}_p(k) - I\|^2_{\infty} + Mh\Big)\,\|\tilde{\zeta}_{\mathrm{PI}}(k)\|^2\\ & +\; 3M^2\ell^2\,\|\Delta\bar{\boldsymbol{e}}(k)\|^2 \;+\; 3\|\tilde{\boldsymbol{w}}_{\mathrm{PID}}(k)\|^2,\end{aligned}$$

and so

$$\begin{aligned}& r\big(\|\Delta\bar{\boldsymbol{e}}(k+1)\|^2 - \|\Delta\bar{\boldsymbol{e}}(k)\|^2\big)\\ &\leq\; 3r\Big(\|\boldsymbol{M}_p(k) - I\|^2_{\infty} + Mh\Big)\,\|\tilde{\zeta}_{\mathrm{PI}}(k)\|^2\\ &-\; r\big(1 - 3M^2\ell^2\big)\|\Delta\bar{\boldsymbol{e}}(k)\|^2 \;+\; 3r\|\tilde{\boldsymbol{w}}_{\mathrm{PID}}(k)\|^2.\end{aligned} \tag{121}$$

Combination

Combining Eqn. 105, Ineq. 120, and Ineq. 121,

$$\begin{aligned}&\Delta\boldsymbol{V}_{\mathrm{PID}}(k) \leq\\ &-\Big(r(1-3M^2\ell^2) - \|\boldsymbol{P}\|_{\infty}M^2\ell^2\Big(\tfrac{1}{\varepsilon}+1\Big)\Big)\,\|\Delta\bar{\boldsymbol{e}}(k)\|^2\\ &\quad + C\,\|\tilde{\boldsymbol{w}}_{\mathrm{PI}}(k)\|^2,\end{aligned}$$

where $C = C_3 + r$. Define

$$S(r,\varepsilon) := \mu^*_{\mathrm{PI}} - 2\varepsilon\|\boldsymbol{P}\|_{\infty}\,\|\boldsymbol{M}_i\|^2_{\infty} \tag{122}$$

$$- 3r\big(\|\boldsymbol{M}_p(k) - I\|^2_{\infty} + Mh\big), \tag{123}$$

$$T(r,\varepsilon,\ell) := r(1-3M^2\ell^2) - \|\boldsymbol{P}\|_{\infty}M^2\ell^2\Big(\tfrac{1}{\varepsilon}+1\Big). \tag{124}$$

ISS of the PID loop follows if $S(r,\varepsilon) > 0$ and $T(r,\varepsilon,\ell) > 0$.

Feasible choices.

We are free to choose any $\varepsilon > 0$ and $r > 0$ such that $S(r,\varepsilon) > 0$. One convenient selection is

$$\begin{aligned}\varepsilon &= \frac{\mu^*_{\mathrm{PI}}}{8\,\|\boldsymbol{P}\|_{\infty}\,\|\boldsymbol{M}_i\|^2_{\infty}},\\ r &= \frac{\mu^*_{\mathrm{PI}}}{8\big(\|\boldsymbol{M}_p(k) - I\|^2_{\infty} + Mh\big)} \quad\Rightarrow\quad S(r,\varepsilon) = \tfrac{3}{8}\mu^*_{\mathrm{PI}} > 0.\end{aligned}$$

With ε, r fixed as above, pick $\ell > 0$ small enough to satisfy $T(r,\varepsilon,\ell) > 0$, namely

$$\ell^2 \;<\; \frac{r}{\big(\,\|\boldsymbol{P}\|_{\infty}\big(\tfrac{1}{\varepsilon}+1\big) + 3r\,\big)M^2},$$

Under these choices, $\Delta\boldsymbol{V}_{\mathrm{PID}}(k) \leq -\alpha_3\|\zeta_{\mathrm{PI}}(k)\|^2 - \alpha_4\|\Delta\bar{\boldsymbol{e}}(k)\|^2 + \beta\,\|\boldsymbol{w}_{\mathrm{PI}}(k)\|^2$ for some $\alpha_3, \alpha_4, \beta > 0$, which satisfies condition (ii) as in Eqn. 97 and proves ISS of the PID closed loop. □

Overshooting under PID law of control Developing from Sec. , we introduce scalar PID recursion along v:

$$\boldsymbol{e}_v(k+1) = a(k)\,\boldsymbol{e}_v(k) - b(k)\,\boldsymbol{s}_v(k) - c(k)\,\Delta\boldsymbol{e}_v(k) + \boldsymbol{w}_v^{\perp}(k), \tag{125}$$

$$\boldsymbol{s}_v(k+1) = \boldsymbol{s}_v(k) + \boldsymbol{e}_v(k) - d_v(k), \tag{126}$$

where

$$a(k) := v^\top \boldsymbol{M}_p(k) v, \tag{127}$$

$$b(k) := v^\top \boldsymbol{G}(k) v, \tag{128}$$

$$c(k) := v^\top \boldsymbol{H}(k) v, \tag{129}$$

$$\Delta \boldsymbol{e}_v(k) = v^\top \Delta \bar{\boldsymbol{e}}(k), \quad \boldsymbol{w}_v^\perp(k) = v^\top \boldsymbol{w}^\perp(k).$$

By construction $a(k) \le q < 1$, $b(k) \le Mh$, $c(k) \le M\ell$ with $M := \sup_k \|\bar{\boldsymbol{A}}(k)\|$, $h := \|\boldsymbol{K}_i\|$ and $\ell := \|\boldsymbol{K}_d\|$.

We now impose an additional requirement on the derivative gain $\boldsymbol{K}_d$ so that, without the effect of noise, the PID update secures the monotonic decrease of $\boldsymbol{e}_v(k)$ before the first negative peak of scalar error $\boldsymbol{e}_v(k)$.

Remark 4 (Pre-overshoot monotonic decrease of scalar errors) *Assume the setting of Proposition 4 and further suppose the scalar error trajectory before the first largest overshoot under PID law is* smooth *in the sense that there exists* $R \ge 1$ *such that*

$$\frac{\boldsymbol{e}_v(k-1)}{\boldsymbol{e}_v(k)} \le R \qquad \textit{for all } k = 1, 2, \ldots, i_{max} - 1, \tag{130}$$

where $\boldsymbol{A}_0 := \max_{k_0 \le i \le k_1} |\boldsymbol{e}_v(i)| = |\boldsymbol{e}_v(i_{max})|$ *from Eqn.73.*

Assume that $\mathbf{w}_v^\perp \equiv 0$ *and* $\boldsymbol{d}_v \equiv 0$.

If, in addition, the derivative gain satisfies

$$l = \|\boldsymbol{K}_d\| \le \frac{1-q}{(R-1)M}, \tag{131}$$

then under PID law

$$\boldsymbol{e}_v(k+1) \le \boldsymbol{e}_v(k) \qquad \textit{for all } k = 0, 1, \ldots, i_{max} - 1.$$

Proof. Before k_0, we have $\boldsymbol{e}_v(k) > 0$, so $\boldsymbol{s}_v(k) \ge 0$ (since $\boldsymbol{s}_v$ accumulates $\boldsymbol{e}_v$ and $\boldsymbol{s}_v(0) = 0$). For $k_0 \le k \le i_{max} - 1$, $\boldsymbol{s}_v(k) \ge 0$ proved in remark 5

Hence

$$\begin{aligned}
\boldsymbol{e}_v(k+1) &= a(k)\boldsymbol{e}_v(k) - b(k)\boldsymbol{s}_v(k) - c(k)\big(\boldsymbol{e}_v(k) - \boldsymbol{e}_v(k-1)\big) \\
&\le a(k)\boldsymbol{e}_v(k) + c(k)\big(\boldsymbol{e}_v(k-1) - \boldsymbol{e}_v(k)\big) \\
&\le \Big[a(k) + c(k)\,(R-1)\Big]\boldsymbol{e}_v(k) \\
&\le \Big[q + (R-1)M\ell\Big]\boldsymbol{e}_v(k) \le \boldsymbol{e}_v(k),
\end{aligned}$$

where the last inequality is exactly Eqn. 131. □

Note for R: In practice, one may estimate a conservative R from PI-law traces and use a small safety factor

Adding Disturbance: With bounded disturbances, the scalar update reads

$$\boldsymbol{e}_v(k+1) \le \big[q + (R-1)c_{\max}\big]\,\boldsymbol{e}_v(k) + |\boldsymbol{w}_v^\perp(k)|,$$

so the same one-step monotonicity conclusion holds whenever

$$|\boldsymbol{w}_v^\perp(k)| \le \Big(1 - \big[q + (R-1)c_{\max}\big]\Big)\,\boldsymbol{e}_v(k)$$

for all pre-first-largest-overshooting steps.

If this smallness condition on disturbances fails at some step, one-step monotonicity may be lost, but the ISS bounds proved earlier still guarantee geometric decay up to a disturbance-dependent radius.

We now show that **before the first negative peak, the integral state is positive**.

Remark 5 *Assume the setting of Remark. 4. Hence,*

$$\boldsymbol{s}_v(k) > 0 \qquad \textit{for all } k = k_0, \ldots, i_{\max} - 1.$$

Proof. We argue by contradiction. Suppose there exists the first $\tau \in [k_0, i_{\max} - 1]$ such that $\boldsymbol{s}_v(\tau) \le 0$. Then $\boldsymbol{s}_v(\tau - 1) > 0$, and since we are on the first negative lobe, $\boldsymbol{e}_v(\tau) < 0$. Compute the one-step change of $\boldsymbol{e}_v$:

$$\begin{aligned}
\boldsymbol{e}_v(\tau{+}1) - \boldsymbol{e}_v(\tau) &= (a(\tau) - 1)\boldsymbol{e}_v(\tau) - b(\tau)\boldsymbol{s}_v(\tau) \\
&= (1 - a(\tau))\,|\boldsymbol{e}_v(\tau)| + b(\tau)\,(-\boldsymbol{s}_v(\tau)) > 0,
\end{aligned}$$

because $a(\tau) \le q < 1$, $\boldsymbol{e}_v(\tau) < 0$ and $\boldsymbol{s}_v(\tau) \le 0$. Hence $\boldsymbol{e}_v(\tau{+}1) > \boldsymbol{e}_v(\tau)$. By the same reasoning, as long as both $\boldsymbol{e}_v(k) < 0$ and $\boldsymbol{s}_v(k) \le 0$ hold, we have

$$\boldsymbol{e}_v(k{+}1) - \boldsymbol{e}_v(k) \ge (1-q)\,|\boldsymbol{e}_v(k)| + b_{\min}\,(-\boldsymbol{s}_v(k)) > 0,$$

where $b_{\min} := \inf_k b(k) > 0$. Meanwhile $\boldsymbol{s}_v(k{+}1) = \boldsymbol{s}_v(k) + \boldsymbol{e}_v(k) \le \boldsymbol{s}_v(k)$ on that interval, so $\boldsymbol{s}_v(k)$ is non-increasing; equivalenty $-\boldsymbol{s}_v(k)$ is non-decreasing. If $\boldsymbol{e}_v$ stayed negative forever, then $\sum_{k=0}^{N} \boldsymbol{e}_v(\tau{+}k) \to -\infty$, so $-\boldsymbol{s}_v(k)$ would grow without bound and the increments $\boldsymbol{e}_v(k{+}1) - \boldsymbol{e}_v(k)$ would eventually be arbitrarily large, forcing $\boldsymbol{e}_v$ to cross 0 in finite time. This contradicts the choice of $i_{\max}$ as the first negative peak. Therefore such τ cannot exist and $\boldsymbol{s}_v(k) > 0$ for all $k = k_0, \ldots, i_{\max} - 1$. □

Theorem 2 (PID reduces the first-overshoot amplitude) *Let the first overshoot occur at index* k_0 *with amplitude* A_0 *(definition in Eqn. 73). Then, the first-overshoot amplitude under PID Steering,* A_0^{PID}*, satisfies* $A_0^{\text{PID}} \le A_0^{\text{PI}}$*, where* A_0^{PI} *denotes the corresponding amplitude under PI Steering.*

Proof. Under the PI law we have

$$A_0^{\text{PI}} = -\,a(i_{max}{-}1)\,\boldsymbol{e}_v(i_{max}{-}1) + b(i_{max}{-}1)\,\boldsymbol{s}_v(i_{max}{-}1) \tag{132}$$

$$-\,\boldsymbol{w}_v^\perp(i_{max}{-}1), \tag{133}$$

while under the PID law

$$A_0^{\text{PID}} = -\,a(i_{max}{-}1)\,\boldsymbol{e}_v(i_{max}{-}1) \tag{134}$$

$$+\,b(i_{max}{-}1)\,\boldsymbol{s}_v(i_{max}{-}1) \tag{135}$$

$$+\,c(i_{max}{-}1)\,\Delta\boldsymbol{e}_v(i_{max}{-}1) \tag{136}$$

$$-\,\boldsymbol{w}_v^\perp(i_{max}{-}1), \tag{137}$$

Due to the monotone decrease before this first largest overshooting condition stated in the previous part and the fact that $c(k) > 0$, we have $c(k_0{-}1)\,\Delta\boldsymbol{e}_v(k_0{-}1) < 0$. Therefore,

$$A_0^{PID} \le b(i_{max} - 1)\,\boldsymbol{s}_v(i_{max} - 1) \tag{138}$$

$$-\,a(i_{max} - 1)\,\boldsymbol{e}_v(i_{max} - 1) \tag{139}$$

$$-\,\boldsymbol{w}_v^\perp(i_{max} - 1) = A_0^{PI}. \tag{140}$$

□

Remark 4 is neccessary because monotonic decrease of $e_v(t)$ before the first peak is both a key technical property for proving Theorem 2 and a desirable feature of PID control itself. Indeed, as noted by Åström and Hägglund (1995b, p.70), poorly tuned derivative gains may produce non-monotonicity, in which case reducing only the first overshoot does not translate into improved overall behavior.

Additional Experimental Results

Qualitative Examples of Concept Steering

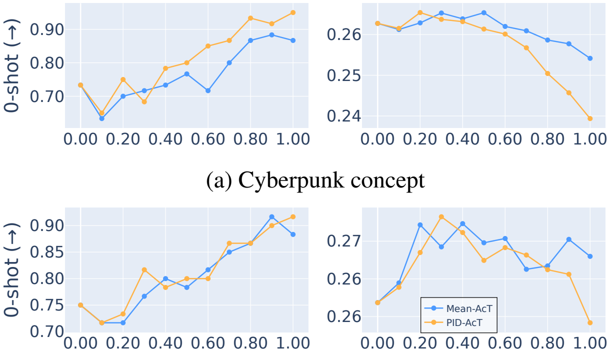
Fig. 7 and 8 show that varying the intervention strength $\alpha \in [0, 1]$ produces a smooth and controllable progression of stylistic traits in the generated images. At low strengths ($\alpha \approx 0.2$), subtle cues emerge, such as faint neon accents for the *cyberpunk* style or mild metallic shading for *steampunk*, while the overall image remains close to the original prompt. At moderate strengths ($\alpha \approx 0.5$), stylistic features become more salient: cyberpunk generations exhibit vivid neon lighting and futuristic cityscapes, whereas steampunk outputs show prominent brass textures, gears, and industrial motifs. Importantly, in this regime, the central semantic content of the prompt (i.e. objects, entities, and spatial composition) is preserved with high fidelity. At high intervention strengths ($\alpha \geq 0.8$), stylistic traits dominate the visual appearance, often saturating the scene with strong color palettes or dense textures, yet semantic alignment to the original prompt remains largely intact, indicating that the steering primarily affects style without eroding core content.

Jailbreaking Large Language Models

We evaluate our method on ActAdd within the Angular Steering framework (Vu and Nguyen 2025) on the jailbreaking task, which seeks to override a model's refusal behavior and elicit harmful outputs. For full results table, please refer to Tab2.

Experimental Setup. Following (Vu and Nguyen 2025), we replace DIM with our method and baselines RePE (Zou et al. 2023a) and ITI (Li et al. 2024). Refusal directions are built from 80% of ADVBENCH (Zou et al. 2023b) and 512 harmless ALPACA (Taori et al. 2023) samples, with the remaining 20% for evaluation. General LM ability is tested on TINYBENCHMARKS (Maia Polo et al. 2024). We evaluate across Gemma2, LLaMA3, and Qwen2.5 models (3B–14B).

Results. PID Steering consistently outperforms DIM and scales robustly across models and metrics (see Tab. 2). On Qwen2.5-14B and LLaMA3.1-8B, PID achieves the largest ASR reductions of **92.7%** and **94.9%**, exceeding DIM by 1.5-2 points, while maintaining almost the same performance, with marginal cost, on TinyBenchmarks. Smaller models also see consistent gains: +2.0 ASR on Qwen2.5-3B and +1.3 on LLaMA3.2-3B. In contrast, ITI and RePE fail to scale, collapsing on larger models with ASR values of 33.7 and 25.4, respectively, on Qwen2.5-14B. A full version of Table 2 which also studies Qwen2.5-7B and Llama3.2-3B is provided in Appendix **??**.



(a) Cyberpunk concept

(b) Steampunk concept

Figure 6: 0-shot and CLIPScore results for ‘cyperpunk‘ and ‘steampunk‘ concept.

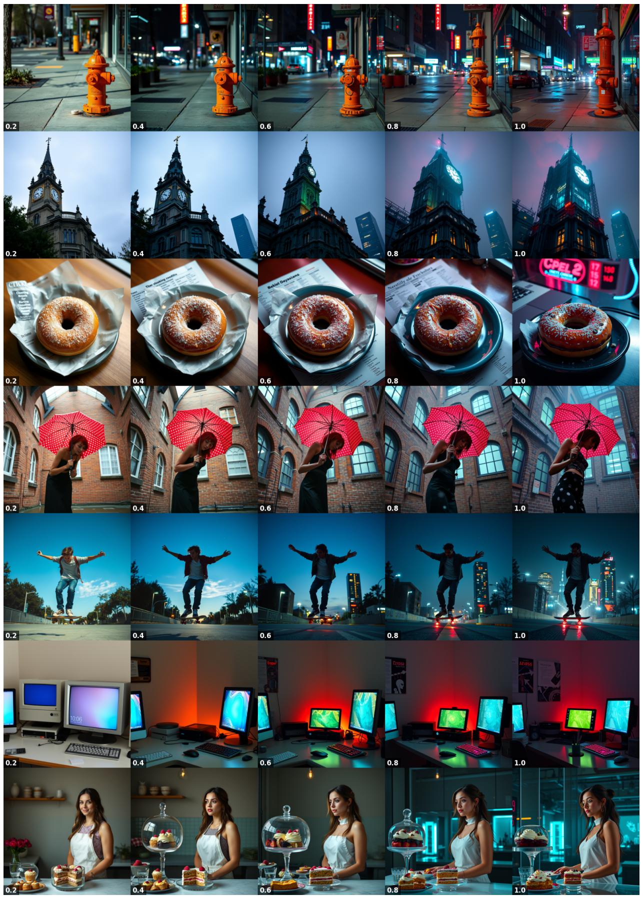


Figure 7: concept *cyberpunk*.

References

Arditi, A.; Obeso, O. B.; Syed, A.; Paleka, D.; Rimsky, N.; Gurnee, W.; and Nanda, N. 2024. Refusal in Language Models Is Mediated by a Single Direction. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*.

Table 2: Comparison of Original, DIM, ITI, RePE, and PID across models on ASR and general benchmarks. Bold = best, underline = second-best within each model (ASR column).

	Method	ASR↑	tinyArc↑	tinyGSM8k strict↑	tinyMMLU↑	tinyTruthQA↑	tinyHellaSwag↑	tinyWinoGrande↑
Qwen2.5-3B Instruct	*Original*	–	62.29	17.64	68.03	56.43	73.18	70.65
	DIM	74.03	61.95	14.80	66.11	54.95	72.40	69.85
	ITI	70.19	61.28	15.57	66.62	54.75	72.71	70.12
	RePE	68.44	61.05	14.60	65.70	54.30	72.03	69.40
	PID	**76.07**	61.20	16.01	67.29	54.10	72.59	69.72
Qwen2.5-7B Instruct	*Original*	–	68.36	81.68	72.57	56.41	78.87	75.19
	DIM	96.15	65.15	80.81	71.19	55.22	78.14	74.42
	ITI	84.61	65.76	79.48	71.23	55.63	78.36	74.75
	RePE	80.32	65.00	78.90	70.60	55.00	77.73	74.15
	PID	**96.46**	66.61	80.78	71.22	55.52	78.28	74.58
Qwen2.5-14B Instruct	*Original*	–	73.96	90.12	74.60	64.50	82.70	73.77
	DIM	90.38	72.74	87.01	74.30	63.01	81.94	72.93
	ITI	33.65	73.15	89.27	74.55	64.03	82.240	73.31
	RePE	25.42	72.40	86.20	73.90	63.20	81.52	72.60
	PID	**92.65**	72.13	88.96	74.52	63.60	82.60	73.04
Llama3.2-3B Instruct	*Original*	–	55.86	59.40	63.48	50.19	75.91	58.63
	DIM	88.46	54.24	58.63	61.68	49.78	75.10	57.94
	ITI	76.92	53.67	57.77	61.85	49.95	75.22	58.16
	RePE	70.15	53.40	57.00	61.10	49.50	74.75	57.53
	PID	**89.76**	53.93	57.26	62.01	50.19	75.07	57.83
Llama3.1-8B Instruct	*Original*	–	65.33	63.21	62.02	54.39	82.51	65.56
	DIM	93.26	62.01	60.57	60.96	54.17	81.73	64.81
	ITI	79.80	64.26	61.85	61.37	54.33	82.01	65.21
	RePE	70.42	61.40	60.00	60.20	53.70	81.35	64.45
	PID	**94.85**	62.30	61.99	61.54	54.24	81.87	64.93
Gemma2-9B Instruct	*Original*	–	69.31	83.19	76.60	55.07	82.31	72.34
	DIM	77.88	68.21	80.14	72.29	51.86	81.45	71.51
	ITI	35.57	68.32	81.47	75.33	53.13	81.70	71.82
	RePE	28.64	67.50	79.20	71.10	51.10	81.20	71.15
	PID	**79.50**	67.91	79.24	74.89	52.49	81.59	71.42

Åström, K.; and Hägglund, T. 1995a. *PID Controllers: Theory, Design, and Tuning*. ISA - The Instrumentation, Systems and Automation Society. ISBN 1-55617-516-7.

Åström, K.; and Hägglund, T. 1995b. *PID Controllers: Theory, Design, and Tuning*. ISA - The Instrumentation, Systems and Automation Society. ISBN 1-55617-516-7.

Åström, K. J.; and Murray, R. 2021. *Feedback systems: an introduction for scientists and engineers*. Princeton university press.

Bayat, R.; Rahimi-Kalahroudi, A.; Pezeshki, M.; Chandar, S.; and Vincent, P. 2025. Steering Large Language Model Activations in Sparse Spaces. arXiv:2503.00177.

Bereska, L.; and Gavves, E. 2024. Mechanistic Interpretability for AI Safety – A Review. arXiv:2404.14082.

Borase, R. P.; Maghade, D.; Sondkar, S.; and Pawar, S. 2021. A review of PID control, tuning methods and applications. *International Journal of Dynamics and Control*, 9(2): 818–827.

Bricken, T.; Templeton, A.; Batson, J.; Chen, B.; Jermyn, A.; Conerly, T.; Turner, N.; Anil, C.; Denison, C.; Askell, A.; Lasenby, R.; Wu, Y.; Kravec, S.; Schiefer, N.; Maxwell, T.; Joseph, N.; Hatfield-Dodds, Z.; Tamkin, A.; Nguyen, K.; McLean, B.; Burke, J. E.; Hume, T.; Carter, S.; Henighan, T.; and Olah, C. 2023. Towards Monosemanticity: Decomposing Language Models With Dictionary Learning. *Transformer Circuits Thread*.

Chen, X.; Fang, H.; Lin, T.-Y.; Vedantam, R.; Gupta, S.; Dollar, P.; and Zitnick, C. L. 2015. Microsoft COCO Captions: Data Collection and Evaluation Server. arXiv:1504.00325.

Cheng, E.; and Amo Alonso, C. 2024. Linearly Controlled Language Generation with Performative Guarantees. *arXiv preprint arXiv:2405.15454*.

Dang, H.-T.; Pham, T.; Thanh-Tung, H.; and Inoue, N. 2025. On Effects of Steering Latent Representation for Large Lan-

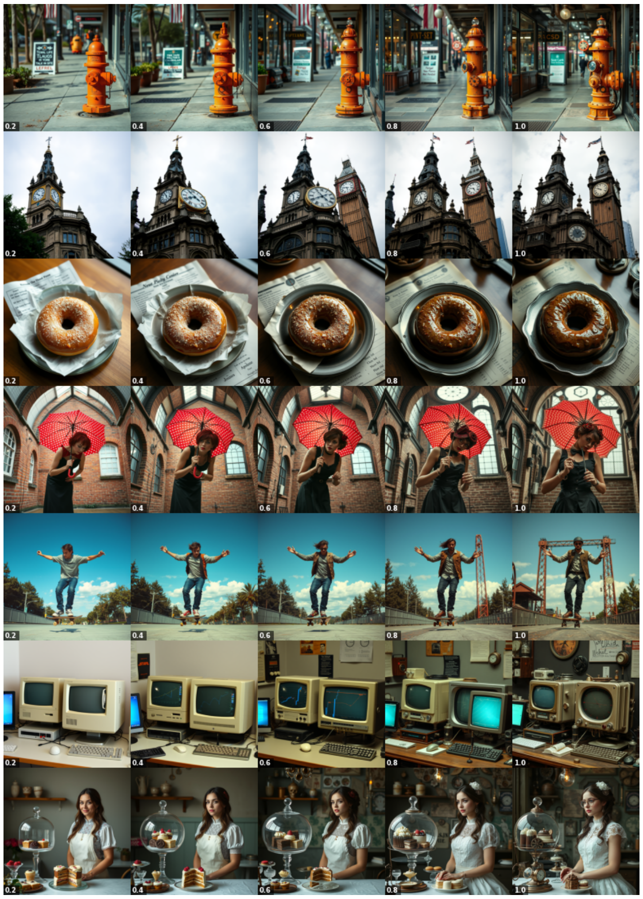


Figure 8: Concept *steampunk*

guage Model Unlearning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 23733–23742.

Elhage, N.; Hume, T.; Olsson, C.; Schiefer, N.; Henighan, T.; Kravec, S.; Hatfield-Dodds, Z.; Lasenby, R.; Drain, D.; Chen, C.; Grosse, R.; McCandlish, S.; Kaplan, J.; Amodei, D.; Wattenberg, M.; and Olah, C. 2022. Toy Models of Superposition. *Transformer Circuits Thread*.

Euler, L. 1768. *Institutionum calculi integralis*. Number v. 1 in Institutionum calculi integralis. imp. Acad. imp. Saènt.

Gajic, Z.; and Qureshi, M. T. J. 2008. *Lyapunov matrix equation in system stability and control*. Courier Corporation.

Gehman, S.; Gururangan, S.; Sap, M.; Choi, Y.; and Smith, N. A. 2020. RealToxicityPrompts: Evaluating Neural Toxic Degeneration in Language Models. In Cohn, T.; He, Y.; and Liu, Y., eds., *Findings of the Association for Computational Linguistics: EMNLP 2020*, 3356–3369. Online: Association for Computational Linguistics.

Geiger, A.; Wu, Z.; Potts, C.; Icard, T.; and Goodman, N. 2024. Finding alignments between interpretable causal variables and distributed neural representations. In *Causal Learning and Reasoning*, 160–187. PMLR.

Gemma Team, G.; Riviere, M.; Pathak, S.; Sessa, P. G.; Hardin, C.; Bhupatiraju, S.; Hussenot, L.; Mesnard, T.; Shahriari, B.; Ramé, A.; et al. 2024. Gemma 2: Improving open language models at a practical size. *arXiv preprint arXiv:2408.00118*.

Hairer, E.; Wanner, G.; and Nørsett, S. P. 1993. *Solving ordinary differential equations I: Nonstiff problems*. Springer.

Hendrycks, D.; Burns, C.; Basart, S.; Zou, A.; Mazeika, M.; Song, D.; and Steinhardt, J. 2021. Measuring Massive Multitask Language Understanding. *Proceedings of the International Conference on Learning Representations (ICLR)*.

Hessel, J.; Holtzman, A.; Forbes, M.; Le Bras, R.; and Choi, Y. 2021. CLIPScore: A Reference-free Evaluation Metric for Image Captioning. In Moens, M.-F.; Huang, X.; Specia, L.; and Yih, S. W.-t., eds., *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, 7514–7528. Online and Punta Cana, Dominican Republic: Association for Computational Linguistics.

Horn, R. A.; and Johnson, C. R. 2012. *Matrix analysis*. Cambridge university press.

Houlsby, N.; Giurgiu, A.; Jastrzebski, S.; Morrone, B.; De Laroussilhe, Q.; Gesmundo, A.; Attariyan, M.; and Gelly, S. 2019. Parameter-efficient transfer learning for NLP. In *International conference on machine learning*, 2790–2799. PMLR.

Jiang, A. Q.; Sablayrolles, A.; Mensch, A.; Bamford, C.; Chaplot, D. S.; de las Casas, D.; Bressand, F.; Lengyel, G.; Lample, G.; Saulnier, L.; Lavaud, L. R.; Lachaux, M.-A.; Stock, P.; Scao, T. L.; Lavril, T.; Wang, T.; Lacroix, T.; and Sayed, W. E. 2023. Mistral 7B. arXiv:2310.06825.

Jiang, Z.-P.; Sontag, E.; and Wang, Y. 1999. Input-to-state stability for discrete-time nonlinear systems. *IFAC Proceedings Volumes*, 32(2): 2403–2408. 14th IFAC World Congress 1999, Beijing, Chia, 5-9 July.

Konen, K.; Jentzsch, S.; Diallo, D.; Schütt, P.; Bensch, O.; El Baff, R.; Opitz, D.; and Hecking, T. 2024. Style Vectors for Steering Generative Large Language Models. In Graham, Y.; and Purver, M., eds., *Findings of the Association for Computational Linguistics: EACL 2024*, 782–802. St. Julian's, Malta: Association for Computational Linguistics.

Kong, L.; Wang, H.; Mu, W.; Du, Y.; Zhuang, Y.; Zhou, Y.; Song, Y.; Zhang, R.; Wang, K.; and Zhang, C. 2024. Aligning Large Language Models with Representation Editing: A Control Perspective. *arXiv preprint arXiv:2406.05954*.

Kotha, S.; Springer, J. M.; and Raghunathan, A. 2023. Understanding catastrophic forgetting in language models via implicit inference. *arXiv preprint arXiv:2309.10105*.

Labs, B. F. 2024. FLUX. https://github.com/black-forest-labs/flux.

Lee, B. W.; Padhi, I.; Ramamurthy, K. N.; Miehling, E.; Dognin, P.; Nagireddy, M.; and Dhurandhar, A. 2024. Programming refusal with conditional activation steering. *arXiv preprint arXiv:2409.05907*.

Li, K.; Patel, O.; Viégas, F.; Pfister, H.; and Wattenberg, M. 2024. Inference-Time Intervention: Eliciting Truthful Answers from a Language Model. arXiv:2306.03341.

Lin, S.; Wang, A.; and Yang, X. 2024. SDXL-Lightning: Progressive Adversarial Diffusion Distillation. arXiv:2402.13929.

Llama Team, A. . M. 2024. The Llama 3 Herd of Models. arXiv:2407.21783.

Logacheva, V.; Dementieva, D.; Ustyantsev, S.; Moskovskiy, D.; Dale, D.; Krotova, I.; Semenov, N.; and Panchenko, A. 2022. ParaDetox: Detoxification with Parallel Data. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 6804–6818. Dublin, Ireland: Association for Computational Linguistics.

Luo, Y.; Tang, Y.; Shen, C.; Zhou, Z.; and Dong, B. 2023. Prompt Engineering Through the Lens of Optimal Control. *arXiv preprint arXiv:2310.14201*.

Luo, Y.; Yang, Z.; Meng, F.; Li, Y.; Zhou, J.; and Zhang, Y. 2025. An empirical study of catastrophic forgetting in large language models during continual fine-tuning. *IEEE Transactions on Audio, Speech and Language Processing*.

Maia Polo, F.; Weber, L.; Choshen, L.; Sun, Y.; Xu, G.; and Yurochkin, M. 2024. tinyBenchmarks: evaluating LLMs with fewer examples. *arXiv preprint arXiv:2402.14992*.

Marks, S.; Rager, C.; Michaud, E. J.; Belinkov, Y.; Bau, D.; and Mueller, A. 2025. Sparse Feature Circuits: Discovering and Editing Interpretable Causal Graphs in Language Models. arXiv:2403.19647.

Marks, S.; and Tegmark, M. 2024. The Geometry of Truth: Emergent Linear Structure in Large Language Model Representations of True/False Datasets. In *First Conference on Language Modeling*.

Minorsky, N. 1922. Directional stability of automatically steered bodies. *Journal of the American Society for Naval Engineers*, 34(2): 280–309.

Ouyang, L.; Wu, J.; Jiang, X.; Almeida, D.; Wainwright, C.; Mishkin, P.; Zhang, C.; Agarwal, S.; Slama, K.; Ray, A.; et al. 2022. Training language models to follow instructions with human feedback. *Advances in neural information processing systems*, 35: 27730–27744.

Park, K.; Choe, Y. J.; and Veitch, V. 2024. The Linear Representation Hypothesis and the Geometry of Large Language Models. arXiv:2311.03658.

Raffel, C.; Shazeer, N.; Roberts, A.; Lee, K.; Narang, S.; Matena, M.; Zhou, Y.; Li, W.; and Liu, P. J. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *J. Mach. Learn. Res.*, 21(1).

Rimsky, N.; Gabrieli, N.; Schulz, J.; Tong, M.; Hubinger, E.; and Turner, A. 2024. Steering Llama 2 via Contrastive Activation Addition. In Ku, L.-W.; Martins, A.; and Srikumar, V., eds., *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 15504–15522. Bangkok, Thailand: Association for Computational Linguistics.

Rodriguez, P.; Blaas, A.; Klein, M.; Zappella, L.; Apostoloff, N.; marco cuturi; and Suau, X. 2025. Controlling Language and Diffusion Models by Transporting Activations. In *The Thirteenth International Conference on Learning Representations*.

Sclar, M.; Choi, Y.; Tsvetkov, Y.; and Suhr, A. 2023. Quantifying Language Models' Sensitivity to Spurious Features in Prompt Design or: How I learned to start worrying about prompt formatting. *arXiv preprint arXiv:2310.11324*.

Soatto, S.; Tabuada, P.; Chaudhari, P.; and Liu, T. Y. 2023. Taming AI Bots: Controllability of Neural States in Large Language Models. *arXiv preprint arXiv:2305.18449*.

Suau, X.; Delobelle, P.; Metcalf, K.; Joulin, A.; Apostoloff, N.; Zappella, L.; and Rodríguez, P. 2024. Whispering experts: neural interventions for toxicity mitigation in language models. In *Proceedings of the 41st International Conference on Machine Learning*, ICML'24. JMLR.org.

Taori, R.; Gulrajani, I.; Zhang, T.; Dubois, Y.; Li, X.; Guestrin, C.; Liang, P.; and Hashimoto, T. B. 2023. Stanford Alpaca: An Instruction-following LLaMA model. https://github.com/tatsu-lab/stanford_alpaca.

Templeton, A.; Conerly, T.; Marcus, J.; Lindsey, J.; Bricken, T.; Chen, B.; Pearce, A.; Citro, C.; Ameisen, E.; Jones, A.; Cunningham, H.; Turner, N. L.; McDougall, C.; MacDiarmid, M.; Freeman, C. D.; Sumers, T. R.; Rees, E.; Batson, J.; Jermyn, A.; Carter, S.; Olah, C.; and Henighan, T. 2024. Scaling Monosemanticity: Extracting Interpretable Features from Claude 3 Sonnet. *Transformer Circuits Thread*.

Turner, A.; Ringer, S.; Shah, R.; Critch, A.; Krakovna, V.; and Hubinger, E. 2023. Activation Addition: Steering Language Models Without Optimization. *arXiv preprint arXiv:2308.10248*.

Turner, A. M.; Thiergart, L.; Leech, G.; Udell, D.; Vazquez, J. J.; Mini, U.; and MacDiarmid, M. 2024. Steering Language Models With Activation Engineering. arXiv:2308.10248.

Visioli, A. 2006. *Practical PID control*. Springer.

von Rütte, D.; Anagnostidis, S.; Bachmann, G.; and Hofmann, T. 2024. A Language Model's Guide Through Latent Space. arXiv:2402.14433.

Vu, H. M.; and Nguyen, T. M. 2025. Angular Steering: Behavior Control via Rotation in Activation Space. *Advances in Neural Information Processing Systems*.

Wei, J.; Bosma, M.; Zhao, V. Y.; Guu, K.; Yu, A. W.; Lester, B.; Du, N.; Dai, A. M.; and Le, Q. V. 2021. Finetuned language models are zero-shot learners. *arXiv preprint arXiv:2109.01652*.

Yang, A.; Yang, B.; Zhang, B.; Hui, B.; Zheng, B.; Yu, B.; Li, C.; Liu, D.; Huang, F.; Wei, H.; et al. 2024. Qwen2. 5 technical report. *arXiv preprint arXiv:2412.15115*.

Zheng, L.; Chiang, W.-L.; Sheng, Y.; Zhuang, S.; Wu, Z.; Zhuang, Y.; Lin, Z.; Li, Z.; Li, D.; Xing, E. P.; Zhang, H.; Gonzalez, J. E.; and Stoica, I. 2023. Judging LLM-as-a-judge with MT-bench and Chatbot Arena. In *Proceedings of the 37th International Conference on Neural Information Processing Systems*, NIPS '23. Red Hook, NY, USA: Curran Associates Inc.

Zou, A.; Phan, L.; Chen, S.; Campbell, J.; Guo, P.; Ren, R.; Pan, A.; Yin, X.; Mazeika, M.; Dombrowski, A.-K.; Goel, S.; Li, N.; Byun, M. J.; Wang, Z.; Mallen, A.; Basart, S.; Koyejo, S.; Song, D.; Fredrikson, M.; Kolter, J. Z.; and

Hendrycks, D. 2023a. Representation Engineering: A Top-Down Approach to AI Transparency. arXiv:2310.01405.

Zou, A.; Wang, Z.; Kolter, J. Z.; and Fredrikson, M. 2023b. Universal and Transferable Adversarial Attacks on Aligned Language Models. arXiv:2307.15043.

Reproducibility Checklist

Instructions for Authors:

This document outlines key aspects for assessing reproducibility. Please provide your input by editing this `.tex` file directly.

For each question (that applies), replace the "yes" text with your answer.

Example: If a question appears as

```
\question{Proofs of all novel claims
are included} {(yes/partial/no)}
yes
```

you would change it to:

```
\question{Proofs of all novel claims
are included} {(yes/partial/no)}
yes
```

Please make sure to:

- Replace ONLY the "yes" text and nothing else.
- Use one of the options listed for that question (e.g., **yes**, **no**, **partial**, or **NA**).
- **Not** modify any other part of the `\question` command or any other lines in this document.

You can `\input` this .tex file right before `\end{document}` of your main file or compile it as a stand-alone document. Check the instructions on your conference's website to see if you will be asked to provide this checklist with your paper or separately.

1. General Paper Structure

1.1. Includes a conceptual outline and/or pseudocode description of AI methods introduced (yes/partial/no/NA) yes

1.2. Clearly delineates statements that are opinions, hypothesis, and speculation from objective facts and results (yes/no) yes

1.3. Provides well-marked pedagogical references for less-familiar readers to gain background necessary to replicate the paper (yes/no) yes

2. Theoretical Contributions

2.1. Does this paper make theoretical contributions? (yes/no) yes

If yes, please address the following points:

2.2. All assumptions and restrictions are stated clearly and formally (yes/partial/no) yes

2.3. All novel claims are stated formally (e.g., in theorem statements) (yes/partial/no) yes

2.4. Proofs of all novel claims are included (yes/partial/no) yes

2.5. Proof sketches or intuitions are given for complex and/or novel results (yes/partial/no) yes

2.6. Appropriate citations to theoretical tools used are given (yes/partial/no) yes

2.7. All theoretical claims are demonstrated empirically to hold (yes/partial/no/NA) yes

2.8. All experimental code used to eliminate or disprove claims is included (yes/no/NA) yes

3. Dataset Usage

3.1. Does this paper rely on one or more datasets? (yes/no) yes

If yes, please address the following points:

3.2. A motivation is given for why the experiments are conducted on the selected datasets (yes/partial/no/NA) NA

3.3. All novel datasets introduced in this paper are included in a data appendix (yes/partial/no/NA) NA

3.4. All novel datasets introduced in this paper will be made publicly available upon publication of the paper with a license that allows free usage for research purposes (yes/partial/no/NA) NA

3.5. All datasets drawn from the existing literature (potentially including authors' own previously published work) are accompanied by appropriate citations (yes/no/NA) yes

3.6. All datasets drawn from the existing literature (potentially including authors' own previously published work) are publicly available (yes/partial/no/NA) yes

3.7. All datasets that are not publicly available are described in detail, with explanation why publicly available alternatives are not scientifically satisficing (yes/partial/no/NA) yes

4. Computational Experiments

4.1. Does this paper include computational experiments? (yes/no) no

If yes, please address the following points:

4.2. This paper states the number and range of values tried per (hyper-) parameter during development of the paper, along with the criterion used for selecting the final parameter setting (yes/partial/no/NA) NA

4.3. Any code required for pre-processing data is included in the appendix (yes/partial/no) NA

4.4. All source code required for conducting and analyzing the experiments is included in a code appendix (yes/partial/no) NA

4.5. All source code required for conducting and analyzing the experiments will be made publicly available upon publication of the paper with a license that allows free usage for research purposes (yes/partial/no) NA

4.6. All source code implementing new methods have comments detailing the implementation, with references to the paper where each step comes from (yes/partial/no) NA

4.7. If an algorithm depends on randomness, then the method used for setting seeds is described in a way sufficient to allow replication of results (yes/partial/no/NA) NA

4.8. This paper specifies the computing infrastructure used for running experiments (hardware and software), including GPU/CPU models; amount of memory; operating system; names and versions of relevant software libraries and frameworks (yes/partial/no) NA

4.9. This paper formally describes evaluation metrics used and explains the motivation for choosing these metrics (yes/partial/no) NA

4.10. This paper states the number of algorithm runs used to compute each reported result (yes/no) NA

4.11. Analysis of experiments goes beyond single-dimensional summaries of performance (e.g., average; median) to include measures of variation, confidence, or other distributional information (yes/no) NA

4.12. The significance of any improvement or decrease in performance is judged using appropriate statistical tests (e.g., Wilcoxon signed-rank) (yes/partial/no) NA

4.13. This paper lists all final (hyper-)parameters used for each model/algorithm in the paper's experiments (yes/partial/no/NA) NA