A PROOFS OF TECHNICAL RESULTS

A.1 PROOFS OF SECTION 4

Proof of Proposition 4.2. We recall the expression of $\tilde{G}_q$:

$$\tilde{G}_q = \frac{1}{n}\sum_{j=1}^{n}(\tilde{R}_j - \frac{1}{n}\sum_{k=1}^{n}\tilde{R}_k)\tilde{Z}_j.$$

The expectation of $\tilde{G}_q$ is:

$$\begin{aligned}
\mathbb{E}[\tilde{G}_q] &= \mathbb{E}[\frac{1}{n}\sum_{j=1}^{n}(\tilde{R}_j - \frac{1}{n}\sum_{k=1}^{n}\tilde{R}_k)\tilde{Z}_j] \\
&= \mathbb{E}[\frac{1}{n}\sum_{j=1}^{n}\tilde{R}_j\tilde{Z}_j - \frac{1}{n^2}\sum_{j=1}^{n}\sum_{k=1}^{n}\tilde{R}_k\tilde{Z}_j] \\
&= \frac{1}{n}\sum_{j=1}^{n}\mathbb{E}[\tilde{R}_j\tilde{Z}_j] - \frac{1}{n^2}\sum_{j=1}^{n}\sum_{k=1}^{n}\mathbb{E}[\tilde{R}_k\tilde{Z}_j] \\
&= \mathbb{E}[\tilde{R}\tilde{Z}] - \mathbb{E}[\tilde{R}\tilde{Z}] = 0 \qquad \text{(by Assumption 4.1}(i)\text{)}.
\end{aligned}$$

The variance of $\tilde{G}_q$ is:

$$\begin{aligned}
&\mathrm{Var}(\tilde{G}_q) \\
&= \mathbb{E}[\tilde{G}_q^2] - (\mathbb{E}[\tilde{G}_q])^2 = \mathbb{E}[\tilde{G}_q^2] - 0 = \mathbb{E}[\tilde{G}_q^2] \\
&= \mathbb{E}\left[\left(\frac{1}{n}\sum_{j=1}^{n}\tilde{R}_j\tilde{Z}_j - \frac{1}{n^2}\sum_{j=1}^{n}\sum_{k=1}^{n}\tilde{R}_k\tilde{Z}_j\right)^2\right] \\
&= \mathbb{E}\left[\frac{1}{n^2}\left(\sum_{j=1}^{n}\tilde{R}_j\tilde{Z}_j\right)^2 + \frac{1}{n^4}\left(\sum_{j=1}^{n}\sum_{k=1}^{n}\tilde{R}_k\tilde{Z}_j\right)^2 - \frac{2}{n^3}\left(\sum_{j=1}^{n}\tilde{R}_j\tilde{Z}_j\right)\left(\sum_{j=1}^{n}\sum_{k=1}^{n}\tilde{R}_k\tilde{Z}_j\right)\right] \\
&= \underbrace{\mathbb{E}\left[\frac{1}{n^2}\left(\sum_{j=1}^{n}\tilde{R}_j\tilde{Z}_j\right)^2\right]}_{\triangleq C_1} + \underbrace{\mathbb{E}\left[\frac{1}{n^4}\left(\sum_{j=1}^{n}\sum_{k=1}^{n}\tilde{R}_k\tilde{Z}_j\right)^2\right]}_{\triangleq C_2} - \underbrace{\mathbb{E}\left[\frac{2}{n^3}\left(\sum_{j=1}^{n}\tilde{R}_j\tilde{Z}_j\right)\left(\sum_{j=1}^{n}\sum_{k=1}^{n}\tilde{R}_k\tilde{Z}_j\right)\right]}_{\triangleq C_3}.
\end{aligned}$$

Next we compute each term. We find for C_1:

$$\begin{aligned}
C_1 &= \mathbb{E}\left[\frac{1}{n^2}\left(\sum_{j=1}^{n}\tilde{R}_j\tilde{Z}_j\right)^2\right] = \frac{1}{n^2}\mathbb{E}\left[\left(\sum_{j=1}^{n}\tilde{R}_j\tilde{Z}_j\right)^2\right] \\
&= \frac{1}{n^2}\mathbb{E}\left[\sum_{j=1}^{n}\tilde{R}_j^2\tilde{Z}_j^2 + \sum_{1\le j<k\le n}\tilde{R}_j\tilde{Z}_j\tilde{R}_k\tilde{Z}_k\right] \\
&= \frac{1}{n^2}\left(\sum_{j=1}^{n}\mathbb{E}\left[\tilde{R}_j^2\tilde{Z}_j^2\right] + \sum_{1\le j<k\le n}\mathbb{E}\left[\tilde{R}_j\tilde{Z}_j\tilde{R}_k\tilde{Z}_k\right]\right) \\
&= \frac{1}{n^2}\left(\sum_{j=1}^{n}\mathbb{E}\left[\tilde{R}_j^2\tilde{Z}_j^2\right] + \sum_{1\le j<k\le n}\mathbb{E}\left[\tilde{R}_j\tilde{Z}_j\right]\mathbb{E}\left[\tilde{R}_k\tilde{Z}_k\right]\right) \quad \text{(independence between } j \text{ and } k\text{)} \\
&= \frac{1}{n^2}\left(n\mathbb{E}\left[\tilde{R}^2\tilde{Z}^2\right] + n(n-1)(\mathbb{E}[\tilde{R}\tilde{Z}])^2\right) \quad \text{(identically distributed)} \\
&= \frac{1}{n}\mathbb{E}\left[\tilde{R}^2\tilde{Z}^2\right] + \frac{(n-1)}{n}(\mathbb{E}[\tilde{R}\tilde{Z}])^2 \\
&= \frac{1}{n}\mathbb{E}[\tilde{R}^2](\mu_Z^2+\sigma_Z^2) + \frac{n-1}{n}\mu_z^2(\mathbb{E}[\tilde{R}])^2 \\
&= \mathbb{E}[\tilde{R}^2](\frac{1}{n}\mu_Z^2 + \frac{1}{n}\sigma_Z^2) + (\mathbb{E}[\tilde{R}])^2(\frac{n-1}{n}\mu_z^2).
\end{aligned}$$

We next compute for C_2:

$$C_2 = \mathbb{E}\left[\frac{1}{n^4}\left(\sum_{j=1}^{n}\sum_{k=1}^{n}\tilde{R}_k\tilde{Z}_j\right)^2\right] = \frac{1}{n^4}\mathbb{E}\left[\sum_{k=1}^{n}\sum_{k'=1}^{n}\sum_{j=1}^{n}\sum_{j'=1}^{n}\tilde{R}_k\tilde{R}_{k'}\tilde{Z}_j\tilde{Z}_{j'}\right].$$

We can decompose the quadruple sum by whether the indices are equal or not. There are four index-pattern types:

- When $k=k'$ and $j=j'$: there are n^2 such terms, each term is $\mathbb{E}[\tilde{R}^2\tilde{Z}^2] = (\mu_Z^2+\sigma_Z^2)\mathbb{E}[\tilde{R}^2]$. Total contribution to C_2 is
$$T_1 = n^2(\mu_Z^2+\sigma_Z^2)\mathbb{E}[\tilde{R}^2].$$
- When $k=k'$ and $j\neq j'$: there are $n^2(n-1)$ such terms. For any fixed k and distinct j, j', $\mathbb{E}[\tilde{R}_k^2\tilde{Z}_j\tilde{Z}_{j'}] = \mathbb{E}[\tilde{R}_k^2]\mathbb{E}[\tilde{Z}_j\tilde{Z}_{j'}] = \mu_Z^2\mathbb{E}[\tilde{R}^2]$. Total contribution to C_2 is
$$T_2 = n^2(n-1)\mu_Z^2\mathbb{E}[\tilde{R}^2].$$
- When $k\neq k'$ and $j=j'$: there are $n^2(n-1)$ such terms. For distinct k, k', $\mathbb{E}[\tilde{R}_k\tilde{R}_{k'}\tilde{Z}^2] = \mathbb{E}[\tilde{R}_k\tilde{R}_{k'}]\mathbb{E}[\tilde{Z}^2] = (\mu_Z^2+\sigma_Z^2)(\mathbb{E}[\tilde{R}])^2$. Total contribution to C_2 is
$$T_3 = n^2(n-1)(\mu_Z^2+\sigma_Z^2)(\mathbb{E}[\tilde{R}])^2.$$
- When $k\neq k'$ and $j\neq j'$: there are $n^2(n-1)^2$ such terms. For all indices different, $\mathbb{E}[\tilde{R}_k\tilde{R}_{k'}\tilde{Z}_j\tilde{Z}_{j'}] = \mu_Z^2(\mathbb{E}[\tilde{R}])^2$. Total contribution to C_2 is
$$T_4 = n^2(n-1)^2\mu_Z^2(\mathbb{E}[\tilde{R}])^2.$$

Therefore, we find

$$\begin{aligned}C_2 &= \frac{1}{n^4}(T_1+T_2+T_3+T_4)\\ &= \frac{1}{n^4}\Big[n^2(\mu_Z^2+\sigma_Z^2)\mathbb{E}[\tilde{R}^2] + n^2(n-1)\mu_Z^2\mathbb{E}[\tilde{R}^2]\\ &\quad + n^2(n-1)(\mu_Z^2+\sigma_Z^2)(\mathbb{E}[\tilde{R}])^2 + n^2(n-1)^2\mu_Z^2(\mathbb{E}[\tilde{R}])^2\Big]\\ &= \mathbb{E}[\tilde{R}^2]\left(\frac{\mu_Z^2}{n}+\frac{\sigma_Z^2}{n^2}\right) + (\mathbb{E}[\tilde{R}])^2\left(\frac{n-1}{n}\mu_Z^2 + \frac{n-1}{n^2}\sigma_Z^2\right).\end{aligned}$$

We next compute for C_3:

$$C_3 = \mathbb{E}\left[\frac{2}{n^3}\left(\sum_{j=1}^n \tilde{R}_j\tilde{Z}_j\right)\left(\sum_{j'=1}^n\sum_{k=1}^n \tilde{R}_k\tilde{Z}_{j'}\right)\right] = \frac{2}{n^3}\mathbb{E}\Big[\sum_{j=1}^n\sum_{j'=1}^n\sum_{k=1}^n \tilde{R}_j\tilde{Z}_j\tilde{R}_k\tilde{Z}_{j'}\Big].$$

We can decompose the triplet sum by whether the indies are equal or not. There are five index-pattern types:

- When $j=k=j'$: there are n such terms. For each j, $\mathbb{E}[\tilde{R}_j^2\tilde{Z}_j^2] = \mathbb{E}[\tilde{R}_j^2]\mathbb{E}[\tilde{Z}_j^2] = (\mu_Z^2+\sigma_Z^2)\mathbb{E}[\tilde{R}^2]$. Total contribution:
$$T_1 = n(\mu_Z^2+\sigma_Z^2)\mathbb{E}[\tilde{R}^2].$$
- When $j=k\neq j'$: there are $n(n-1)$ such terms. For each j and $j'\neq j$, $\mathbb{E}[\tilde{R}_j^2\tilde{Z}_j\tilde{Z}_{j'}] = \mathbb{E}[\tilde{R}_j^2]\mathbb{E}[\tilde{Z}_j\tilde{Z}_{j'}] = \mathbb{E}[\tilde{R}^2]\mathbb{E}[\tilde{Z}_j]\mathbb{E}[\tilde{Z}_{j'}] = \mathbb{E}[\tilde{R}^2]\mu_Z^2$. Total contribution is
$$T_2 = n(n-1)\mathbb{E}[\tilde{R}^2]\mu_Z^2.$$
- When $j=j'\neq k$: there are $n(n-1)$ such terms. For each j and $k\neq j$, $\mathbb{E}[\tilde{R}_j\tilde{Z}_j^2\tilde{R}_k] = \mathbb{E}[\tilde{R}_j\tilde{R}_k]\mathbb{E}[\tilde{Z}_j^2] = (\mu_Z^2+\sigma_Z^2)(\mathbb{E}[\tilde{R}])^2$. Total contribution is
$$T_3 = n(n-1)(\mu_Z^2+\sigma_Z^2)(\mathbb{E}[\tilde{R}])^2.$$
- When $k=j'\neq j$: there are $n(n-1)$ such terms. For each j and $k\neq j$, $\mathbb{E}[\tilde{R}_j\tilde{Z}_j\tilde{R}_k\tilde{Z}_k] = \mathbb{E}[\tilde{R}_j\tilde{R}_k]\mathbb{E}[\tilde{Z}_j\tilde{Z}_k] = \mu_Z^2(\mathbb{E}[\tilde{R}])^2$. Total contribution is
$$T_4 = n(n-1)\mu_Z^2(\mathbb{E}[\tilde{R}])^2.$$
- When j, j', k are all distinct: there are $n(n-1)(n-2)$ such terms. For each triple of distinct indices, $\mathbb{E}[\tilde{R}_j\tilde{Z}_j\tilde{R}_k\tilde{Z}_{j'}] = \mu_Z^2(\mathbb{E}[\tilde{R}])^2$. Total contribution is
$$T_5 = n(n-1)(n-2)\mu_Z^2(\mathbb{E}[\tilde{R}])^2.$$

Therefore, we find

$$\begin{aligned}C_3 &= \frac{2}{n^3}\left(T_1+T_2+T_3+T_4+T_5\right)\\ &= \frac{2}{n^3}\Bigg[n(\mu_Z^2+\sigma_Z^2)\mathbb{E}[\tilde{R}^2] + n(n-1)\mathbb{E}[\tilde{R}^2]\mu_Z^2\\ &\qquad + n(n-1)(\mu_Z^2+\sigma_Z^2)\left(\mathbb{E}[\tilde{R}]\right)^2 + n(n-1)\mu_Z^2\left(\mathbb{E}[\tilde{R}]\right)^2\\ &\qquad + n(n-1)(n-2)\mu_Z^2\left(\mathbb{E}[\tilde{R}]\right)^2\Bigg]\\ &= \mathbb{E}[\tilde{R}^2]\left(\frac{2}{n}\mu_Z^2+\frac{2}{n^2}\sigma_Z^2\right) + (\mathbb{E}[\tilde{R}])^2\left(\frac{2n-2}{n}\mu_Z^2+\frac{2n-2)}{n^2}\sigma_Z^2\right).\end{aligned}$$

Group terms with $\mathbb{E}[\tilde{R}^2]$ and $(\mathbb{E}[\tilde{R}])^2$ coefficients:

$$C_1 + C_2 - C_3 = \mathbb{E}[\tilde{R}^2]\left[\left(\frac{1}{n}\mu_Z^2 + \frac{1}{n}\sigma_Z^2\right) + \left(\frac{\mu_Z^2}{n} + \frac{\sigma_Z^2}{n^2}\right) - \left(\frac{2}{n}\mu_Z^2 + \frac{2}{n^2}\sigma_Z^2\right)\right]$$
$$+ (\mathbb{E}[\tilde{R}])^2\left[\frac{n-1}{n}\mu_Z^2 + \left(\frac{n-1}{n}\mu_Z^2 + \frac{n-1}{n^2}\sigma_Z^2\right) - \left(\frac{2n-2}{n}\mu_Z^2 + \frac{2n-2}{n^2}\sigma_Z^2\right)\right].$$

We simplify each bracket to obtain:

$$\mathrm{Var}(\tilde{G}_q) = C_1 + C_2 - C_3 = \frac{n-1}{n^2}\sigma_Z^2\left(\mathbb{E}[\tilde{R}^2] - (\mathbb{E}[\tilde{R}])^2\right) = \frac{n-1}{n^2}\sigma_Z^2\sigma_R^2.$$

For a given prompt, $\tilde{R}$ takes 1 with probability p and -1 with probability $1-p$, leading to its variance of $4p(1-p)$. We obtain the final variance of the per-prompt gradient estimator:

$$\mathrm{Var}(\tilde{G}_q) = \frac{\sigma_Z^2(n-1)}{n^2} \cdot 4p(1-p).$$

This completes the proof. □

Proof of Proposition 4.3. We recall the expression of $\tilde{G}_q$:

$$\tilde{G}_q = \frac{1}{n}\sum_{j=1}^{n}\left(\tilde{R}_j - \frac{1}{n-1}\sum_{\substack{k=1\\k\neq j}}^{n}\tilde{R}_k\right)\tilde{Z}_j.$$

The expectation of $\tilde{G}_q$ is:

$$\begin{aligned}
\mathbb{E}[\tilde{G}_q] &= \mathbb{E}\left[\frac{1}{n}\sum_{j=1}^{n}\left(\tilde{R}_j - \frac{1}{n-1}\sum_{\substack{k=1\\k\neq j}}^{n}\tilde{R}_k\right)\tilde{Z}_j\right] \\
&= \mathbb{E}\left[\frac{1}{n}\sum_{j=1}^{n}\tilde{R}_j\tilde{Z}_j - \frac{1}{n(n-1)}\sum_{j=1}^{n}\sum_{\substack{k=1\\k\neq j}}^{n}\tilde{R}_k\tilde{Z}_j\right] \\
&= \frac{1}{n}\sum_{j=1}^{n}\mathbb{E}[\tilde{R}_j\tilde{Z}_j] - \frac{1}{n(n-1)}\sum_{j=1}^{n}\sum_{\substack{k=1\\k\neq j}}^{n}\mathbb{E}[\tilde{R}_k\tilde{Z}_j] \\
&= \mathbb{E}[\tilde{R}\tilde{Z}] - \mathbb{E}[\tilde{R}\tilde{Z}] \qquad \text{(by Assumption 4.1}(i)) \\
&= 0.
\end{aligned}$$

The variance of $\tilde{G}_q$ is:

$$\begin{aligned}
&\mathrm{Var}(\tilde{G}_q) \\
&= \mathbb{E}[\tilde{G}_q^2] - (\mathbb{E}[\tilde{G}_q])^2 = \mathbb{E}[\tilde{G}_q^2] - 0 = \mathbb{E}[\tilde{G}_q^2] \\
&= \mathbb{E}\left[\left(\frac{1}{n}\sum_{j=1}^{n}\tilde{R}_j\tilde{Z}_j - \frac{1}{n(n-1)}\sum_{j=1}^{n}\sum_{\substack{k=1\\k\neq j}}^{n}\tilde{R}_k\tilde{Z}_j\right)^2\right] \\
&= \mathbb{E}\left[\frac{1}{n^2}\left(\sum_{j=1}^{n}\tilde{R}_j\tilde{Z}_j\right)^2 + \frac{1}{n^2(n-1)^2}\left(\sum_{j=1}^{n}\sum_{\substack{k=1\\k\neq j}}^{n}\tilde{R}_k\tilde{Z}_j\right)^2 - \frac{2}{n^2(n-1)}\left(\sum_{j=1}^{n}\tilde{R}_j\tilde{Z}_j\right)\left(\sum_{j=1}^{n}\sum_{\substack{k=1\\k\neq j}}^{n}\tilde{R}_k\tilde{Z}_j\right)\right] \\
&= \underbrace{\mathbb{E}\left[\frac{1}{n^2}\left(\sum_{j=1}^{n}\tilde{R}_j\tilde{Z}_j\right)^2\right]}_{\triangleq C_1} + \underbrace{\mathbb{E}\left[\frac{1}{n^2(n-1)^2}\left(\sum_{j=1}^{n}\sum_{\substack{k=1\\k\neq j}}^{n}\tilde{R}_k\tilde{Z}_j\right)^2\right]}_{\triangleq C_2} - \underbrace{\mathbb{E}\left[\frac{2}{n^2(n-1)}\left(\sum_{j=1}^{n}\tilde{R}_j\tilde{Z}_j\right)\left(\sum_{j=1}^{n}\sum_{\substack{k=1\\k\neq j}}^{n}\tilde{R}_k\tilde{Z}_j\right)\right]}_{\triangleq C_3}.
\end{aligned}$$

The first term C_1 is already computed in the proof of Proposition 4.2, and we have:

$$C_1 = \mathbb{E}[\tilde{R}^2](\frac{1}{n}\mu_Z^2 + \frac{1}{n}\sigma_Z^2) + (\mathbb{E}[\tilde{R}])^2(\frac{n-1}{n}\mu_z^2).$$

Next, we consider the term C_2:

$$\begin{aligned}
C_2 &= \mathbb{E}\left[\frac{1}{n^2(n-1)^2}\left(\sum_{j=1}^n \sum_{\substack{k=1\\k\neq j}}^n \tilde{R}_k \tilde{Z}_j\right)^2\right] \\
&= \mathbb{E}\left[\frac{1}{n^2(n-1)^2}\left(\left(\sum_{j=1}^n \sum_{k=1}^n \tilde{R}_k \tilde{Z}_j\right) - \left(\sum_{j=1}^n \tilde{R}_j \tilde{Z}_j\right)\right)^2\right] \\
&= \frac{1}{n^2(n-1)^2}\Bigg(\mathbb{E}\left[\left(\sum_{j=1}^n \sum_{k=1}^n \tilde{R}_k \tilde{Z}_j\right)^2\right] - 2\,\mathbb{E}\left[\left(\sum_{j=1}^n \sum_{k=1}^n \tilde{R}_k \tilde{Z}_j\right)\left(\sum_{j'=1}^n \tilde{R}_{j'} \tilde{Z}_{j'}\right)\right] + \mathbb{E}\left[\left(\sum_{j=1}^n \tilde{R}_j \tilde{Z}_j\right)^2\right]\Bigg).
\end{aligned}$$

We can utilize the computation from the proof of Proposition 4.2 to have:

$$\mathbb{E}\left[\frac{1}{n^4}\left(\sum_{j=1}^n \sum_{k=1}^n \tilde{R}_k \tilde{Z}_j\right)^2\right] = \mathbb{E}[\tilde{R}^2]\left(\frac{\mu_Z^2}{n} + \frac{\sigma_Z^2}{n^2}\right) + (\mathbb{E}[\tilde{R}])^2\left(\frac{n-1}{n}\mu_Z^2 + \frac{n-1}{n^2}\sigma_Z^2\right),$$

$$\mathbb{E}\left[\frac{2}{n^3}\left(\sum_{j=1}^n \tilde{R}_j \tilde{Z}_j\right)\left(\sum_{j'=1}^n \sum_{k=1}^n \tilde{R}_k \tilde{Z}_{j'}\right)\right] = \mathbb{E}[\tilde{R}^2]\left(\frac{2}{n}\mu_Z^2 + \frac{2}{n^2}\sigma_Z^2\right) + (\mathbb{E}[\tilde{R}])^2\left(\frac{2n-2}{n}\mu_Z^2 + \frac{2n-2}{n^2}\sigma_Z^2\right),$$

$$\mathbb{E}\left[\frac{1}{n^2}\left(\sum_{j=1}^n \tilde{R}_j \tilde{Z}_j\right)^2\right] = \mathbb{E}[\tilde{R}^2]\left(\frac{1}{n}\mu_Z^2 + \frac{1}{n}\sigma_Z^2\right) + (\mathbb{E}[\tilde{R}])^2\left(\frac{n-1}{n}\mu_Z^2\right).$$

Therefore,

$$\begin{aligned}
C_2 &= \mathbb{E}[\tilde{R}^2]\left[\frac{n}{(n-1)^2}\mu_Z^2 + \frac{1}{(n-1)^2}\sigma_Z^2 - \left(\frac{2}{(n-1)^2}\mu_Z^2 + \frac{2}{n(n-1)^2}\sigma_Z^2\right) + \frac{1}{n(n-1)^2}\mu_Z^2 + \frac{1}{n(n-1)^2}\sigma_Z^2\right] \\
&+ (\mathbb{E}[\tilde{R}])^2\left[\frac{n}{(n-1)}\mu_Z^2 + \frac{1}{(n-1)}\sigma_Z^2 - \left(\frac{2}{n-1}\mu_Z^2 + \frac{2}{n(n-1)}\sigma_Z^2\right) + \frac{1}{n(n-1)}\mu_Z^2\right] \\
&= \mathbb{E}[\tilde{R}^2]\left[\frac{n^2-2n+1}{n(n-1)^2}\mu_Z^2 + \frac{n-1}{n(n-1)^2}\sigma_Z^2\right] + (\mathbb{E}[\tilde{R}])^2\left[\frac{n^2-2n+1}{n(n-1)}\mu_Z^2 + \frac{n-2}{n(n-1)}\sigma_Z^2\right] \\
&= \mathbb{E}[\tilde{R}^2]\left[\frac{1}{n}\mu_Z^2 + \frac{1}{n(n-1)}\sigma_Z^2\right] + (\mathbb{E}[\tilde{R}])^2\left[\frac{n-1}{n}\mu_Z^2 + \frac{n-2}{n(n-1)}\sigma_Z^2\right].
\end{aligned}$$

We compute C_3 as follows:

$$\begin{aligned}
C_3 &= \mathbb{E}\left[\frac{2}{n^2(n-1)}\left(\sum_{j=1}^n \tilde{R}_j \tilde{Z}_j\right)\left(\sum_{j'=1}^n \sum_{\substack{k=1\\k\neq j'}}^n \tilde{R}_k \tilde{Z}_{j'}\right)\right] \\
&= \mathbb{E}\left[\frac{2}{n^2(n-1)}\left(\sum_{j=1}^n \tilde{R}_j \tilde{Z}_j\right)\left(\sum_{j'=1}^n \sum_{k=1}^n \tilde{R}_k \tilde{Z}_{j'} - \sum_{j'=1}^n \tilde{R}_{j'} \tilde{Z}_{j'}\right)\right] \\
&= \frac{2}{n^2(n-1)}\Bigg(\mathbb{E}\left[\left(\sum_{j=1}^n \tilde{R}_j \tilde{Z}_j\right)\left(\sum_{j'=1}^n \sum_{k=1}^n \tilde{R}_k \tilde{Z}_{j'}\right)\right] - \mathbb{E}\left[\left(\sum_{j=1}^n \tilde{R}_j \tilde{Z}_j\right)\left(\sum_{j'=1}^n \tilde{R}_{j'} \tilde{Z}_{j'}\right)\right]\Bigg).
\end{aligned}$$

We can utilize the computation of $\frac{n^3}{2}C_3$ and n^2C_1 from the proof of Proposition 4.2 to have:

$$\mathbb{E}\left[\left(\sum_{j=1}^{n}\tilde{R}_j\tilde{Z}_j\right)\left(\sum_{j'=1}^{n}\sum_{k=1}^{n}\tilde{R}_k\tilde{Z}_{j'}\right)\right] = \mathbb{E}[\tilde{R}^2]\left(n^2\mu_Z^2 + n\sigma_Z^2\right) + (\mathbb{E}[\tilde{R}])^2\left(n^2(n-1)\mu_Z^2 + n(n-1)\sigma_Z^2\right),$$

$$\mathbb{E}\left[\left(\sum_{j=1}^{n}\tilde{R}_j\tilde{Z}_j\right)\left(\sum_{j'=1}^{n}\tilde{R}_{j'}\tilde{Z}_{j'}\right)\right] = \mathbb{E}[\tilde{R}^2](n\mu_Z^2 + n\sigma_Z^2) + (\mathbb{E}[\tilde{R}])^2(n(n-1)\mu_z^2).$$

Plugging these terms to the computation of C_3 yields us:

$$\begin{aligned}
C_3 = \frac{2}{n^2(n-1)}\Big\{&\mathbb{E}[\tilde{R}^2]\left(n^2\mu_Z^2 + n\sigma_Z^2\right) + (\mathbb{E}[\tilde{R}])^2\left(n^2(n-1)\mu_Z^2 + n(n-1)\sigma_Z^2\right) \\
&- \left[\mathbb{E}[\tilde{R}^2]\left(n\mu_Z^2 + n\sigma_Z^2\right) + (\mathbb{E}[\tilde{R}])^2\left(n(n-1)\mu_Z^2\right)\right]\Big\} \\
= \mathbb{E}[\tilde{R}^2]\cdot&\frac{2(n^2-n)}{n^2(n-1)}\mu_Z^2 + (\mathbb{E}[\tilde{R}])^2\cdot\frac{2n(n-1)}{n^2(n-1)}\left((n-1)\mu_Z^2 + \sigma_Z^2\right) \\
= \mathbb{E}[\tilde{R}^2]&\left(\frac{2}{n}\mu_Z^2\right) + (\mathbb{E}[\tilde{R}])^2\left(\frac{2n-2}{n}\mu_Z^2 + \frac{2}{n}\sigma_Z^2\right).
\end{aligned}$$

We have:

$$\begin{aligned}
\mathrm{Var}(\tilde{G}_q) = C_1 + C_2 - C_3 &= \mathbb{E}[\tilde{R}^2]\left(\frac{1}{n-1}\sigma_Z^2\right) + (\mathbb{E}[\tilde{R}])^2\left(-\frac{1}{n-1}\sigma_Z^2\right) \\
&= \frac{\sigma_Z^2}{n-1}(\mathbb{E}[\tilde{R}])^2 - (\mathbb{E}[\tilde{R}])^2) \\
&= \frac{\sigma_Z^2}{n-1}\mathrm{Var}(\tilde{R}).
\end{aligned}$$

For a given prompt, $\tilde{R}$ takes 1 with probability p and -1 with probability $1-p$, leading to its variance of $4p(1-p)$. We obtain the final variance of the per-prompt gradient estimator:

$$\mathrm{Var}(\tilde{G}_q) = \frac{\sigma_Z^2}{n-1}\cdot 4p(1-p).$$

This completes the proof. □

A.2 Proofs of Section 5

Proof of Theorem 5.1. For clarity and continuity, we restate problem (6) before proceeding with the proof:

$$\begin{aligned}
\min \quad & \sum_{q\in\mathcal{B}_t} a_q\frac{n_q-1}{n_q^2} \\
\text{s.t.} \quad & \sum_{q\in\mathcal{B}_t} n_q = C \\
& L \le n_q \le U \quad \forall q \in \mathcal{B}_t.
\end{aligned} \tag{10}$$

Let $V(\{n_q\})$ be the objective function of the above problem. We compute the first and second derivatives of the objective function with respect to each coordinate n_q:

$$\frac{\partial V}{\partial n_q} = -a_q\frac{n_q-2}{n_q^3}.$$

Since $n_q \ge L \ge 3$, so for all q, $\frac{\partial V}{\partial n_q} < 0$. Thus, V is decreasing with respect to each n_q on the feasible set.

For the second derivatives:

$$\frac{\partial^2 V}{\partial n_q\partial n_{q'}} = 0 \quad \forall q \ne q', \quad \frac{\partial^2 V}{\partial n_q^2} = a_q\frac{2n_q-6}{n_q^4} \ge 0 \quad \forall q \quad (\text{Since } n_q \ge L \ge 3, \text{ and } a_q \ge 0)$$

Therefore, V is convex and decreasing in each n_q on the feasible set

$$\left\{ n \in \mathbb{R}^B : \sum_{q\in\mathcal{B}_t} n_q = C, \quad L \le n_q \le U \;\forall q \right\}.$$

Hence, the minimizer exists and is unique whenever the feasible set is nonempty $BL \le C \le BU$.

The Lagrangian function is

$$\mathcal{L} = \sum_{q\in\mathcal{B}_t} a_q \frac{n_q - 1}{n_q^2} + \lambda \left(\sum_{q\in\mathcal{B}_t} n_q - C \right) + \sum_{q\in\mathcal{B}_t} \mu_q (L - n_q) + \sum_{q\in\mathcal{B}_t} \nu_q (n_q - U)$$

where $\lambda \in \mathbb{R}$, and $\mu_q, \nu_q \ge 0$ are Lagrangian multipliers. The KKT conditions are:

$$\begin{aligned}
& - a_q \frac{n_q - 2}{n_q^3} + \lambda - \mu_q + \nu_q = 0 && \forall q, \\
& \mu_q \ge 0, \quad \nu_q \ge 0 && \forall q, \\
& \mu_q (n_q - L) = 0, \quad \nu_q (n_q - U) = 0 && \forall q, \\
& L \le n_q \le U && \forall q, \\
& \sum_{q\in\mathcal{B}_t} n_q = C.
\end{aligned}$$

We consider three cases of n_q:

- For each q with $L < n_q < U$, the KKT stationarity condition is

  $$\lambda = a_q \frac{n_q - 2}{n_q^3},$$

  where λ is the Lagrange multiplier for the sum constraint. Note that the right-hand side is decreasing in n_q.

  For $n_q = L$, the right-hand side is $a_q \frac{L-2}{L^3}$, and for $n_q = U$, it is $a_q \frac{U-2}{U^3}$. Therefore, for each q and any $\lambda \in (a_q \frac{U-2}{U^3}, a_q \frac{L-2}{L^3})$, there is at most one solution n_q to $a_q \frac{n_q-2}{n_q^3} = \lambda$ in the interior (L, U). If $\lambda \ge a_q \frac{L-2}{L^3}$ or $\lambda \le a_q \frac{U-2}{U^3}$, there is no interior solution, and the optimum for n_q must be at a bound.

- If $n_q = L$, then $\mu_q \ge 0$ and $\nu_q = 0$. According to the KKT condition, we obtain:

  $$\lambda = a_q \frac{L-2}{L^3} + \mu_q \ge a_q \frac{L-2}{L^3}.$$

- If $n_q = U$, then $\mu_q = 0$ and $\nu_q \ge 0$. According to the KKT condition, we obtain:

  $$\lambda = a_q \frac{U-2}{U^3} - \nu_q \le a_q \frac{U-2}{U^3}.$$

For a value of λ, for each coordinate, the KKT solution for n_q is defined as:

$$n_q^\star(\lambda) = \begin{cases} U & \text{if } \lambda \le a_q \frac{U-2}{U^3}, \\ \text{the unique solution to } \lambda = a_q \frac{n_q-2}{n_q^3} & \text{if } a_q \frac{U-2}{U^3} < \lambda < a_q \frac{L-2}{L^3}, \\ L & \text{if } \lambda \ge a_q \frac{L-2}{L^3}. \end{cases}$$

The coupling constraint $\sum_{q\in\mathcal{B}_t} n_q = C$ is enforced by selecting λ such that

$$S(\lambda) \triangleq \sum_{q\in\mathcal{B}_t} n_q^\star(\lambda) = C.$$

Each $n_q^\star(\lambda)$ is non-increasing in λ since $a_q \frac{n_q-2}{n_q^3}$ is decreasing and the projection preserves monotonicity. Consequently, $S(\lambda)$ is also non-increasing. In particular:

- As $\lambda \to -\infty$, $n_q^\star(\lambda) \to U$, so $S(-\infty) = BU$.
- As $\lambda \to +\infty$, $n_q^\star(\lambda) \to L$, so $S(+\infty) = BL$.

Therefore, for any feasible C with $BL \le C \le BU$, there exists a unique $\lambda^\star$ such that $S(\lambda^\star) = C$. Moreover, because S is non-increasing, finding $\lambda^\star$ can be done by bisection. If $C > BU$ or $C < BL$, the problem is infeasible. □

Proof of Theorem 5.2. For clarity and continuity, we restate Problem 8 before proceeding with the proof:

$$\begin{aligned} \min \quad & \sum_{q \in \mathcal{B}_t} a_q \frac{1}{n_q} \\ \text{s.t.} \quad & \sum_{q \in \mathcal{B}_t} n_q = C \\ & L \le n_q \le U \quad \forall q \in \mathcal{B}_t \end{aligned} \tag{11}$$

Let $V(\{n_q\})$ be the objective function of the above problem. We compute the first and second derivatives of the objective function with respect to each coordinate n_q:

$$\frac{\partial V}{\partial n_q} = -a_q \frac{1}{(n_q - 1)^2}$$

Since $n_q \ge L \ge 3$ and $a_q > 0$, we have $\frac{\partial V}{\partial n_q} \le 0$ for all q. Thus, V is decreasing with respect to each n_q on the feasible set.

For the second derivatives:

$$\frac{\partial^2 V}{\partial n_q \partial n_{q'}} = 0 \quad \forall q \ne q', \qquad \frac{\partial^2 V}{\partial n_q^2} = 2a_q \frac{1}{(n_q - 1)^3} > 0 \quad \forall q$$

Therefore, V is convex and decreasing in each n_q on the feasible set

$$\left\{ n \in \mathbb{R}^B : \sum_{q \in \mathcal{B}_t} n_q = C, \quad L \le n_q \le U \right\}.$$

Hence, the minimizer exists and is unique whenever the feasible set is nonempty ($BL \le C \le BU$).

The Lagrangian function is

$$\mathcal{L} = \sum_{q \in \mathcal{B}_t} a_q \frac{1}{n_q - 1} + \lambda \left(\sum_{q \in \mathcal{B}_t} n_q - C \right) + \sum_{q \in \mathcal{B}_t} \mu_q (L - n_q) + \sum_{q \in \mathcal{B}_t} \nu_q (n_q - U)$$

where $\lambda \in \mathbb{R}$, $\mu_q, \nu_q \ge 0$. The KKT conditions are:

$$\begin{aligned} & -a_q \frac{1}{(n_q - 1)^2} + \lambda - \mu_q + \nu_q = 0 && \forall q \\ & \mu_q \ge 0, \quad \nu_q \ge 0 && \forall q \\ & \mu_q (n_q - L) = 0, \quad \nu_q (n_q - U) = 0 && \forall q \\ & L \le n_q \le U && \forall q \\ & \sum_{q \in \mathcal{B}_t} n_q = C. \end{aligned}$$

We consider three cases of n_q:

- For each q with $L < n_q < U$, the KKT stationarity condition is

$$\lambda = a_q \frac{1}{(n_q - 1)^2},$$

where λ is the Lagrange multiplier for the sum constraint. Note that the right-hand side is decreasing in n_q since $n_q \geq L \geq 3$.

For $n_q = L$, the right-hand side is $a_q \frac{1}{(L-1)^2}$, and for $n_q = U$, it is $a_q \frac{1}{(U-1)^2}$. Therefore, for each q and any $\lambda \in (a_q \frac{1}{(U-1)^2}, a_q \frac{1}{(L-1)^2})$, there is one solution $n_q = \sqrt{\frac{a_q}{\lambda}} + 1$ to $a_q \frac{1}{(n_q-1)^2} = \lambda$ in the interior (L, U). If $\lambda \geq a_q \frac{1}{(L-1)^2}$ or $\lambda \leq a_q \frac{1}{(U-1)^2}$, there is no interior solution, and the optimum for n_q must be at a bound.

- If $n_q = L$, then $\mu_q \geq 0$ and $\nu_q = 0$. According to the KKT condition, we obtain:

$$\lambda = a_q \frac{1}{(L-1)^2} + \mu_q \geq a_q \frac{1}{(L-1)^2}.$$

- If $n_q = U$, then $\mu_q = 0$ and $\nu_q \geq 0$. According to the KKT condition, we obtain:

$$\lambda = a_q \frac{1}{(U-1)^2} - \nu_q \leq a_q \frac{1}{(U-1)^2}.$$

For a value of λ, for each coordinate, the KKT solution for n_q is defined as:

$$n_q^\star(\lambda) = \begin{cases} U & \text{if } \lambda \leq a_q \frac{1}{(U-1)^2}, \\ \sqrt{\frac{a_q}{\lambda}} + 1 & \text{if } a_q \frac{1}{(U-1)^2} < \lambda < a_q \frac{1}{(L-1)^2}, \\ L & \text{if } \lambda \geq a_q \frac{1}{(L-1)^2}. \end{cases}$$

The coupling constraint $\sum_{q \in \mathcal{B}_t} n_q = C$ is enforced by selecting λ such that

$$S(\lambda) := \sum_{q \in \mathcal{B}_t} n_q^\star(\lambda) = C.$$

Each $n_q^\star(\lambda)$ is non-increasing in λ (since $a_q \frac{1}{(n_q-1)^2}$ is decreasing and the projection preserves monotonicity), so $S(\lambda)$ is also non-increasing. In particular:

- As $\lambda \to -\infty$, $n_q^\star(\lambda) \to U$, so $S(-\infty) = BU$.
- As $\lambda \to +\infty$, $n_q^\star(\lambda) \to L$, so $S(+\infty) = BL$.

Therefore, for any feasible C with $BL \leq C \leq BU$, there exists a unique λ such that $S(\lambda) = C$. If $C > BU$ or $C < BL$, the problem is infeasible. □

B Statistical Tests for Second-Order Uncorrelation

In this section, we provide statistical tests to validate the assumptions in our paper.

B.1 First-order Correlation Test via Fisher's method

For each question q, consider the two random variables $\tilde{R}_q$ and $\tilde{Z}_q$, with n independent observations

$$\{(\tilde{R}_{q,j}, \tilde{Z}_{q,j})\}_{j=1}^n.$$

Compute per-question Pearson correlation. The sample Pearson correlation for question q is

$$\hat{\rho}_q = \frac{\sum_{j=1}^n (\tilde{R}_{q,j} - \bar{\tilde{R}}_q)(\tilde{Z}_{q,j} - \bar{Z}_q)}{\sqrt{\sum_{j=1}^n (\tilde{R}_{q,j} - \bar{R}_q)^2 \sum_{j=1}^n (\tilde{Z}_{q,j} - \bar{Z}_q)^2}},$$

where

$$\bar{R}_q = \frac{1}{n} \sum_{j=1}^n \tilde{R}_{q,j}, \qquad \bar{Z}_q = \frac{1}{n} \sum_{j=1}^n \tilde{Z}_{q,j}.$$

Compute per-question p-values. For each question q, we test the null hypothesis

$$H_{0,q} : \rho_q = 0.$$

The p-value p_q is obtained directly from the standard Pearson correlation test.

Combine p-values across questions using Fisher's method. Let Q be the total number of questions. Fisher's method combines the per-question p-values $\{p_q\}_{q=1}^{Q}$ into a single test statistic:

$$\chi^2_{\text{Fisher}} = -2 \sum_{q=1}^{Q} \ln p_q.$$

Under the global null hypothesis

$$H_0 : \rho_q = 0 \quad \forall q,$$

the statistic χ^2_{Fisher} follows a chi-squared distribution with $2Q$ degrees of freedom:

$$\chi^2_{\text{Fisher}} \sim \chi^2_{2Q}.$$

Global p-value and decision rule. The global p-value for testing H_0 across all questions is

$$p_{\text{global}} = \Pr\left(\chi^2_{2Q} \geq \chi^2_{\text{Fisher}}\right).$$

Given a significance level α (e.g., $\alpha = 0.05$), we make the following decision:

- If $p_{\text{global}} < \alpha$, we reject the global null hypothesis H_0, which indicates that at least some of the correlations ρ_q are significantly different from zero across the questions.
- If $p_{\text{global}} \geq \alpha$, we fail to reject H_0, which supports the hypothesis that the correlations ρ_q are zero for all questions at the significance level α.

We conduct the correlation test described above on a benchmark of $Q = 600$ questions, each with $n = 16$ independent rollouts. For each question q, we compute the Pearson correlation between $\tilde{R}_q$ and $\tilde{Z}_q$, obtain the corresponding p-value p_q, and aggregate across all questions using Fisher's method to compute the global p-value p_{global}.

We evaluate the policy model π_{θ_t} at four checkpoints during training of `Qwen2.5-Math-1.5B`, corresponding to 0.0, 0.5, 1.0 epochs. At each checkpoint, we report the resulting p_{global} values in Table 5. Since all global p-values exceed the chosen significance level $\alpha = 0.05$, we do not reject the null hypothesis, which supports our assumption that the correlations ρ_q are zero across all questions.

Epoch	Global p-value	
	$\tilde{Z}_j = \mathbb{1}^\top H(\tilde{o}_j)$	$\tilde{Z}_j = \lVert H(\tilde{o}_j) \rVert_2$
0.0	0.3230	0.7322
0.5	0.3050	0.1108
1.0	0.3050	0.2186

Table 5: Global p-values (p_{global}) across training epochs for `Qwen2.5-Math-1.5B`.

B.2 First-order Correlation Test via Edgington's Method

For each question q, let $\hat{\rho}_q$ denote the sample Pearson correlation computed from n independent rollouts, and let p_q be the corresponding two-sided p-value for testing the null hypothesis

$$H_{0,q} : \rho_q = 0.$$

To aggregate evidence across all Q questions, we apply Edgington's sum-of-p method.

Sum of p-values. Each per-question p_q is treated as a realization of a $\text{Uniform}(0,1)$ variable under its null hypothesis. Edgington's statistic is defined by the simple sum

$$S_{\text{Ed}} = \sum_{q=1}^{Q} p_q.$$

Null distribution. Under the global null hypothesis

$$H_0 : \rho_q = 0 \quad \forall q,$$

each $p_q \sim \text{Uniform}(0, 1)$, and therefore

$$S_{\text{Ed}} \sim \text{Irwin–Hall}(Q),$$

with mean and variance

$$\mathbb{E}[S_{\text{Ed}}] = \frac{Q}{2}, \qquad \text{Var}(S_{\text{Ed}}) = \frac{Q}{12}.$$

For large Q, S_{Ed} is well approximated by a normal distribution:

$$S_{\text{Ed}} \approx \mathcal{N}\left(\frac{Q}{2}, \frac{Q}{12}\right).$$

Global p-value and decision rule. Small values of S_{Ed} indicate joint evidence against H_0. The corresponding one-sided global p-value is

$$p_{\text{global}} = \Phi\left(\frac{S_{\text{Ed}} - Q/2}{\sqrt{Q/12}}\right),$$

where Φ denotes the standard normal CDF. Given a significance level $\alpha = 0.5$, we reject H_0 when $p_{\text{global}} < \alpha$.

We set up the experiment identically to the Fisher's method test in Appendix B.1, using the same benchmark of $Q = 600$ questions, each with $n = 16$ independent rollouts. For each checkpoint of the policy model π_{θ_t}, we compute the Edgington statistic and report the global p-value. Since all global p-values exceed the chosen significance level $\alpha = 0.05$, we do not reject the null hypothesis, which supports our assumption that the correlations ρ_q are zero across all questions.

Epoch	Global p-value	
	$\tilde{Z}_j = \mathbb{1}^\top H(\tilde{o}_j)$	$\tilde{Z}_j = \lVert H(\tilde{o}_j)\rVert_2$
0.0	0.9125	0.7894
0.5	0.8963	0.3964
1.0	0.8912	0.2148

Table 6: Global p-values (p_{global}) across training epochs for `Qwen2.5-Math-1.5B` using Edgington's method.

B.3 Equal Variance Test via Levene's Test

In the numerical experiments, we have assumed that the variance for $\tilde{Z}_q$ is constant across different prompts q. We proceed with a hypothesis test:

$$H_0 : \sigma^2_{Z_q} = \sigma^2_{Z_{q'}} \ \forall q \neq q', \qquad H_1 : \text{At least one } \sigma^2_{Z_q} \neq \sigma^2_{Z_{q'}}$$

For each question q, consider the random variable $\tilde{Z}_q$ with n_q independent observations $\{\tilde{Z}_{q,j}\}_{j=1}^{n_q}$.

Transform observations for Levene's test. Let $Y_{q,j}$ denote the absolute deviation from the per-question median:

$$Y_{q,j} = \left|\tilde{Z}_{q,j} - \text{median}(\tilde{Z}_{q,1}, \ldots, \tilde{Z}_{q,n_q})\right|.$$

Compute group means of transformed observations. The mean of the transformed observations for question q is

$$\bar{Y}_q = \frac{1}{n_q} \sum_{j=1}^{n_q} Y_{q,j},$$

and the overall mean across all questions is

$$\bar{Y} = \frac{1}{N} \sum_{q=1}^{Q} \sum_{j=1}^{n_q} Y_{q,j}, \qquad N = \sum_{q=1}^{Q} n_q.$$

Compute Levene's test statistic. The test statistic is given by

$$W = \frac{(N-Q) \sum_{q=1}^{Q} n_q (\bar{Y}_q - \bar{Y})^2}{(Q-1) \sum_{q=1}^{Q} \sum_{j=1}^{n_q} (Y_{q,j} - \bar{Y}_q)^2}.$$

Under the null hypothesis that the variances are equal across questions,

$$H_0 : \sigma^2_{Z_q} = \sigma^2_{Z_{q'}} \quad \forall q \neq q',$$

the statistic W approximately follows an F-distribution with $Q-1$ and $N-Q$ degrees of freedom $W \sim F_{Q-1,N-Q}$.

Compute p-value and decision rule. The p-value for testing H_0 is

$$p_{\text{Levene}} = \Pr(F_{Q-1,N-Q} \geq W).$$

Given a significance level α (e.g., $\alpha = 0.05$), we make the following decision:

- If $p_{\text{Levene}} < \alpha$, we reject H_0, indicating that the variances of $\tilde{Z}_q$ differ across questions.
- If $p_{\text{Levene}} \geq \alpha$, we fail to reject H_0, the hypothesis that the variances are equal across all questions, at the significance level α.

We conduct the variance homogeneity test described above on a benchmark of $Q = 600$ questions, each with $n = 16$ independent rollouts. We perform Levene's test across all questions to assess the equality of variances. We evaluate the policy model π_{θ_t} at four checkpoints during training of `Qwen2.5-Math-1.5B`, corresponding to 0.0, 0.5, 1.0 epochs. At each checkpoint, we report the resulting global p-values p_{Levene} in Table 7. Since all p_{Levene} exceed the chosen significance level $\alpha = 0.05$, we can not reject the null hypothesis, which supports our assumption that the variances $\sigma^2_{Z_q}$ are equal across all questions.

Epoch	Global p-value	
	$\tilde{Z}_j = \mathbb{1}^\top H(\tilde{o}_j)$	$\tilde{Z}_j = \lVert H(\tilde{o}_j) \rVert_2$
0.0	0.5019	0.2705
0.5	0.4132	0.4785
1.0	0.3847	0.3847

Table 7: p_{Levene} from Levene's test across training epochs for `Qwen2.5-Math-1.5B`, assessing variance homogeneity of $\tilde{Z}_q$.

B.4 Equal Variance Test via O'Brien's Test

In the numerical experiments, we have assumed that the variance for $\tilde{Z}_q$ is constant across different prompts q. We proceed with a hypothesis test:

$$H_0 : \sigma^2_{Z_q} = \sigma^2_{Z_{q'}} \ \forall q \neq q', \qquad H_1 : \text{At least one } \sigma^2_{Z_q} \neq \sigma^2_{Z_{q'}}$$

For each question q, consider the random variable $\tilde{Z}_q$ with n_q independent observations $\{\tilde{Z}_{q,j}\}_{j=1}^{n_q}$.

Transform observations for O'Brien's test. Let $Y_{q,j}$ denote O'Brien's transformation of the observations:

$$Y_{q,j} = \frac{(n_q - 1.5) n_q (\tilde{Z}_{q,j} - \bar{\tilde{Z}}_q)^2 - 0.5 s_q^2 (n_q - 1)}{(n_q - 1)(n_q - 2)},$$

where $\bar{\tilde{Z}}_q$ is the sample mean for question q, and s_q^2 is the unbiased sample variance for question q.

Compute group means of transformed observations. The mean of the transformed observations for question q is

$$\bar{Y}_q = \frac{1}{n_q} \sum_{j=1}^{n_q} Y_{q,j},$$

and the overall mean across all questions is

$$\bar{Y} = \frac{1}{N} \sum_{q=1}^{Q} \sum_{j=1}^{n_q} Y_{q,j}, \qquad N = \sum_{q=1}^{Q} n_q.$$

Compute O'Brien's test statistic. The test statistic is given by

$$W_{\text{OB}} = \frac{(N-Q)\sum_{q=1}^{Q} n_q (\bar{Y}_q - \bar{Y})^2}{(Q-1)\sum_{q=1}^{Q}\sum_{j=1}^{n_q}(Y_{q,j} - \bar{Y}_q)^2}.$$

Under the null hypothesis that the variances are equal across questions,

$$H_0 : \sigma^2_{Z_q} = \sigma^2_{Z_{q'}} \quad \forall q \neq q',$$

the statistic W_{OB} approximately follows an F-distribution with $Q-1$ and $N-Q$ degrees of freedom $W_{\text{OB}} \sim F_{Q-1,N-Q}$.

Compute p-value and decision rule. The p-value for testing H_0 is

$$p_{\text{OB}} = \Pr(F_{Q-1,N-Q} \geq W_{\text{OB}}).$$

Given a significance level α (e.g., $\alpha = 0.05$), we make the following decision:

- If $p_{\text{OB}} < \alpha$, we reject H_0, indicating that the variances of $\tilde{Z}_q$ differ across questions.
- If $p_{\text{OB}} \geq \alpha$, we fail to reject H_0, the hypothesis that the variances are equal across all questions, at the significance level α.

We conduct the variance homogeneity test described above on a benchmark of $Q = 600$ questions, each with $n = 16$ independent rollouts. We perform O'Brien's test across all questions to assess the equality of variances. We evaluate the policy model π_{θ_t} at three checkpoints during training of `Qwen2.5-Math-1.5B`, corresponding to 0.0, 0.5, 1.0 epochs. At each checkpoint, we report the resulting global p-values p_{OB} in Table 8. Since all p_{OB} exceed the chosen significance level $\alpha = 0.05$, we cannot reject the null hypothesis, which supports our assumption that the variances $\sigma^2_{Z_q}$ are equal across all questions.

Epoch	**Global p-value**	
	$\tilde{Z}_j = \mathbb{1}^\top H(\tilde{o}_j)$	$\tilde{Z}_j = \|H(\tilde{o}_j)\|_2$
0.0	0.1612	0.3009
0.5	0.1215	0.2563
1.0	0.1229	0.2420

Table 8: p_{OB} from O'Brien's test across training epochs for `Qwen2.5-Math-1.5B`, assessing variance homogeneity of $\tilde{Z}_q$.

C Additional Information on Numerical Experiments

Hyperparameters. We curate a list of important training hyperparameters for our experiment in Table 9.

Table 9: Hyperparameter configuration.

Category	Hyperparameter	Value / Setting
Optimizer	Optimizer	AdamW
	Learning rate	1×10^{-6}
	Warm-up	20 rollout steps
rollout	Prompt batch size	512
	Responses per prompt	6/8/Dynamic
Training	Mini-batch size	512
	Max generation length	10 240 tokens
	Temperature	1.0

C.1 Additional information on ablation studies

Inverse-accuracy allocation. We allocate more rollout budget to prompts with lower empirical accuracy. Concretely, letting acc_i denote the running accuracy estimate for prompt i, we set target weights $w_i \propto (1 - \mathrm{acc}_i + \epsilon)$ and normalize to meet the global budget and per-prompt bounds.

Inverse-variance allocation. We allocate more rollout budget to prompts whose answers exhibit lower variance. Letting σ_i^2 be the (running) answer variance estimate, we set $w_i \propto 1/(\sigma_i^2 + \epsilon)$ with the same normalization.

Both heuristics are implemented via a continuously relaxed, constrained optimization that enforces the total-budget and box constraints; we solve it with an online solver and then map fractional solutions to integers using the rounding heuristic.

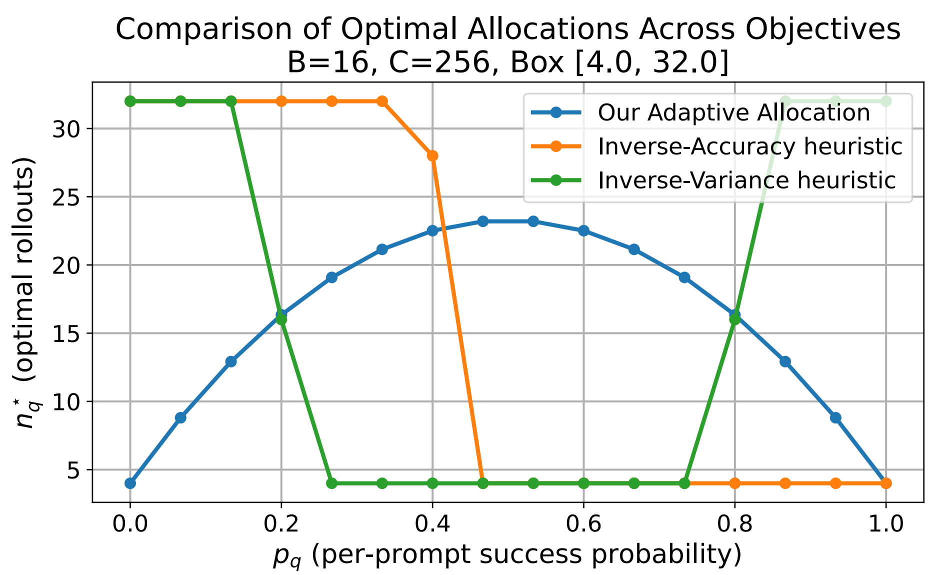


Figure 3: Comparison of optimal rollout allocations produced by different heuristics versus our proposed variance-aware allocation strategy. The figure plots the optimal number of rollouts $n_i^\star$ against prompt difficulty p_i, highlighting how our method allocates budget differently from inverse-accuracy and inverse-variance baselines.

C.2 Prompt template.

During training, we only use one prompt template for every prompt in the dataset. There are two prompt templates, one for mathematical reasoning and one for tool-augmented reasoning.

Figure 4: Prompt template for mathematical reasoning

```
Solve the following math problem step by step. The last line of your
    response should be of the form Answer: $Answer (without quotes)
    where $Answer is the answer to the problem. Do not wrap $Answer with
    \boxed{}.

current question: {{question}}

Below are two examples for format reference.
Example question 1: Solve for x: 3x - 5 = 16.

Response:
Add 5 to both sides: 3x = 21.
Divide both sides by 3: x = 7.
Answer: 7

Solve the current question. Remember to put your answer on its own line
    after "Answer:".
```

Figure 5: Prompt template for tool augmented reasoning

```
In this environment you have access to a set of tools you can use to
    assist with the user query.

You may perform multiple rounds of function calls.

In each round, you can call one or more functions.

Here are available functions in JSONSchema format:
    \n```json\n{func_schemas}\n```

In your response, you need to first think about the reasoning process in
    the mind and then conduct function calling to get the information or
    perform the actions if needed. \
The reasoning process and function calling are enclosed within <think>
    </think> and <tool_call> </tool_call> tags. \
The results of the function calls will be given back to you after
    execution, \
and you can continue to call functions until you get the final answer
    for the user's question. \
Finally, if you have got the answer, enclose it within \\boxed{{}} with
    latex format and do not continue to call functions, \
i.e., <think> Based on the response from the function call, I get the
    weather information. </think> The weather in Beijing on 2025-04-01
    is \\[ \\boxed{{20C}} \\].

For each function call, return a json object with function name and
    arguments within <tool_call></tool_call> XML tags:
<tool_call>
{{"name": <function-name>, "arguments": <args-json-object>}}
</tool_call>
```

D Algorithms

The algorithm capturing the complete flow the posterior update for the Gaussian Process is provided in Algorithm 1.

Algorithm 1 Recursive GP Posterior Update

Require: Mini-batch $\mathcal{B}_t$; rollout allocation $\{n_q\}_{q=1}^{\mathcal{B}_t}$; prior mean $m_t(\mathcal{D}) \in \mathbb{R}^Q$, kernel matrix $\Sigma \in \mathbb{R}^{Q\times Q}$;

for each $q \in \mathcal{B}_t$ **do**
  # Run n_q rollouts and observe outcomes $\tilde{R}_j \in \{-1, 1\}$
  $\bar{R}_q \leftarrow \frac{1}{n_q}\sum_{j=1}^{n_q} \tilde{R}_j$
  $\hat{g}_q \leftarrow \text{sigmoid}^{-1}\left(\text{clip}\left(\frac{\bar{R}_q+1}{2}, \epsilon, 1-\epsilon\right)\right)$
end for
$g_t^{\text{observe}} \leftarrow (\hat{g}_q)_{q\in\mathcal{B}_t}$
Partition m_t and Σ according to $\mathcal{B}_t$ and $\mathcal{B}_t^c$
$m^\star_{t,\mathcal{B}_t^c} \leftarrow m_{t,\mathcal{B}_t^c} + \Sigma_{\mathcal{B}_t^c\mathcal{B}_t}\Sigma^{-1}_{\mathcal{B}_t\mathcal{B}_t}(g_t^{\text{observe}} - m_{t,\mathcal{B}_t})$
$\Sigma^\star \leftarrow \Sigma_{\mathcal{B}_t^c\mathcal{B}_t^c} - \Sigma_{\mathcal{B}_t^c\mathcal{B}_t}\Sigma^{-1}_{\mathcal{B}_t\mathcal{B}_t}\Sigma_{\mathcal{B}_t\mathcal{B}_t^c}$
for $q = 1$ **to** Q **do**
  if $q \in \mathcal{B}_t$ **then** $m_{t+1}(x_q) \leftarrow \hat{g}_q$ **else** $m_{t+1}(x_q) \leftarrow m^\star_{t,\mathcal{B}_t^c}(x_q)$ **end if**
end for
$\hat{p}_{t+1} = \text{sigmoid}(m_{t+1}(\mathcal{D}))$
return $\{\hat{p}_{t+1}\}, m_{t+1}$

Algorithm 2 presents our heuristic rounding procedure, which maps a continuous solution to a discrete one while ensuring that the budget constraints remain satisfied.

Algorithm 2 Heuristic rounding for integer rollout allocation

Require: Solution $\{n_q^\star\}$, total budget C, bounds $\{L, U\}$, objective functions $f_q(\cdot)$ for each q

For each q, set $\hat{n}_q \leftarrow \lfloor n_q^\star \rfloor$
$C_{\text{rem}} \leftarrow C - \sum_{q\in\mathcal{B}_t} \hat{n}_q$
for each q with $\hat{n}_q < U$ **do**
  Compute incentive: $\Delta_q \leftarrow f_q(\hat{n}_q) - f_q(\hat{n}_q + 1)$
end for
while $C_{\text{rem}} > 0$ **do**
  Select $q^\star = \arg\max_{q:\hat{n}_q<U} \Delta_q$
  Set $\hat{n}_{q^\star} \leftarrow \hat{n}_{q^\star} + 1$
  Recompute $\Delta_{q^\star} \leftarrow f_{q^\star}(\hat{n}_{q^\star}) - f_{q^\star}(\hat{n}_{q^\star} + 1)$
  $C_{\text{rem}} \leftarrow C_{\text{rem}} - 1$
end while
return Integer allocation $\{\hat{n}_q\}$ with $\sum_{q\in\mathcal{B}_t} \hat{n}_q = C$ and $L \leq \hat{n}_q \leq U$ for all q

E Extension to Continuous Rewards

This section details the necessary adaptations to our predictive rollout allocation strategy for the case where the reward $R(\tilde{o}_j)$ is a real-valued random variable. All definitions, assumptions, and notation follow the main text unless otherwise stated.

E.1 Gradient Variance for Continuous Rewards

We first state the analogues of our variance propositions for the continuous reward setting. The proofs are intermediate results from proofs for binary case in Appendix A.

Proposition E.1 (Dr. GRPO gradient variance, continuous reward)**.** *Let $R(\tilde{o}_j) = \tilde{R}$ be a real-valued random variable with variance $\mathrm{Var}(\tilde{R})$. If Assumption 4.1 holds and $\mathrm{Var}(\tilde{Z}) = \sigma_Z^2$, then the variance of the per-prompt projected Dr. GRPO gradient estimator with n rollouts is*

$$\mathrm{Var}(\tilde{G}) = \frac{(n-1)\sigma_Z^2}{n^2}\mathrm{Var}(\tilde{R}).$$

Proposition E.2 (RLOO gradient variance, continuous reward)**.** *Let $R(\tilde{o}_j) = \tilde{R}$ be a real-valued random variable with variance $\mathrm{Var}(\tilde{R})$. If Assumption 4.1 holds and $\mathrm{Var}(\tilde{Z}) = \sigma_Z^2$, then the variance of the per-prompt projected RLOO gradient estimator with n rollouts is*

$$\mathrm{Var}(\tilde{G}) = \frac{\sigma_Z^2}{n-1}\mathrm{Var}(\tilde{R}).$$

E.2 Gaussian Process Prediction of Reward Variance

For continuous rewards, the per-prompt gradient variance depends on $\mathrm{Var}(\tilde{R}_q)$, which is not directly observable prior to rollout. To predict this quantity, we replace the GP model for success probability with a GP model for reward variance. Specifically, for each prompt q, we model the reward variance as $v_{q,t} = \mathrm{softplus}(g_t(x_q)) = \log(1 + \exp(g_t(x_q)))$, where g_t is a latent GP as in the main text. After observing rewards $\{\tilde{R}_{q,j}\}_{j=1}^{n_q}$, we compute the sample variance $\widehat{s}_q^2$ and set the observation for the latent variable as $\hat{g}_{q,t} = \log(\exp(\widehat{s}_q^2) - 1)$. The GP posterior update and recursive prediction steps proceed identically, replacing the sigmoid link with the softplus link.

E.3 Budget Allocation Optimization

Given predicted reward variances $\widehat{\mathrm{Var}}(\tilde{R}_q)$, we define $a_q := \sigma_{Z_q}^2 \widehat{\mathrm{Var}}(\tilde{R}_q)$. The continuous relaxation of the rollout allocation problem for Dr. GRPO becomes

$$\min\left\{\sum\nolimits_{q\in\mathcal{B}_t} a_q\frac{n_q-1}{n_q^2}\ :\ \sum\nolimits_{q\in\mathcal{B}_t} n_q = C,\ L\le n_q\le U,\ n_q\in\mathbb{R}\ \forall q\right\},$$

and for RLOO,

$$\min\left\{\sum\nolimits_{q\in\mathcal{B}_t} a_q\frac{1}{n_q-1}\ :\ \sum\nolimits_{q\in\mathcal{B}_t} n_q = C,\ L\le n_q\le U,\ n_q\in\mathbb{R}\ \forall q\right\}.$$

The optimal solutions are given by Theorems 5.1 and 5.2 in the main text, now with the updated definition of a_q. The rounding procedure described in Appendix D applies without modification.

F Empirical Validation of Importance Ratios in Partially Off-Policy Training

In the off-policy regime, importance ratios $r_{j,\tau}(\theta)$ rarely deviate from 1. This indicates that even partially off-policy training methods produce updates that are close to on-policy, a phenomenon particularly pronounced in LLM post-training. Consequently, Assumption 3.1 is unlikely to be restrictive in our setting.

To support this assumption empirically, we measure importance ratios on the response tokens of off-policy samples from our training runs. Prompt and padding tokens are excluded from this analysis. Our evaluation uses 2,560 prompts sampled across different stages of training for `Qwen2.5-Math-1.5B`, with 4 rollouts per prompt. We then collect the importance ratios for all generated tokens and compute the fraction that falls within the interval $[1-\alpha, 1+\alpha]$ for several values of α. The results are summarized in Table 10.

These results confirm that the vast majority of importance ratios remain extremely close to 1, providing strong empirical justification for the approximation $r_{j,\tau}(\theta) \approx 1$ in our analysis.

α	Percentage in $[1-\alpha, 1+\alpha]$
5e-02	97.85%
5e-03	82.46%
5e-04	71.51%

Table 10: Fraction of response tokens whose importance ratios fall within $[1-\alpha, 1+\alpha]$ for various choices of α.

G Training Evolution Comparison

In this section, we assess the robustness and stability of our method by retraining `Qwen2.5-Math-1.5B` using GRPO, RLOO, and their VIP-augmented counterparts (GRPO+VIP, RLOO+VIP) across **five random seeds**. Figures 6 and 7 report the mean and standard deviation for multiple performance metrics (*best@32*, *maj@32*, *mean@32*).

To ensure that all training trajectories are directly comparable, **every model is trained on the same dataset under identical optimization settings**: the same fixed ordering of 17k training prompts, one epoch of training, a batch size of 512, mini-batch size of 64, and rollout budget per batch of 512 * 8. As a result, each gradient step corresponds to the same amount of data and computation across all methods.

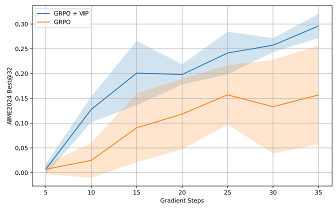
Across all seeds and evaluation checkpoints, we observe consistent and pronounced improvements from using VIP:

(i) Faster early-stage learning. VIP yields substantial gains in the early phase of training. For example, on AIME2024 *mean@32*, RLOO+VIP reaches an accuracy of **0.0316** by step 10, whereas RLOO reaches only **0.0056**—a **6$\times$ increase**. Similar trends appear in both *best@32* and *maj@32* metrics across AIME2024 and AIME2025.

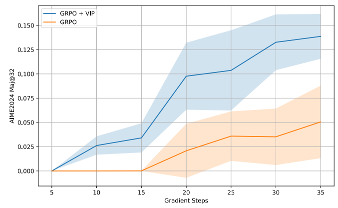
(ii) Steeper and more reliable improvement per gradient step. VIP consistently increases the slope of the learning curve. Its trajectories rise smoothly and monotonically, while the baselines (particularly GRPO on AIME2025 *best@32*) often progress slowly or temporarily plateau between steps 10–20. This shows that variance-aware allocation accelerates the effective learning rate without introducing instability.

(iii) Increased training stability. VIP reduces variance across seeds and produces smoother learning curves, reflecting more stable gradient updates. This aligns with the goal of variance-informed allocation: reducing gradient noise directly translates into more predictable and reliable optimization dynamics.

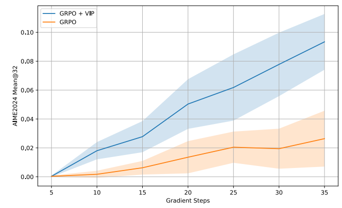
Together, these results demonstrate that VIP improves both the **speed** and the **stability** of GRPO and RLOO training, leading to faster convergence and consistently higher performance throughout the entire training trajectory.



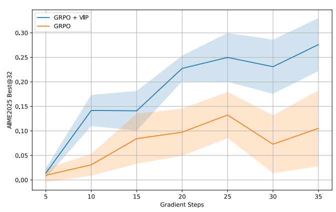
(a) AIME 2024, best@32



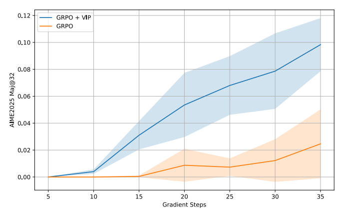
(b) AIME 2024, maj@32



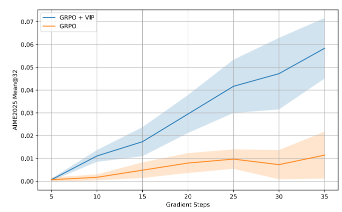
(c) AIME 2024, mean



(d) AIME 2025, best@32

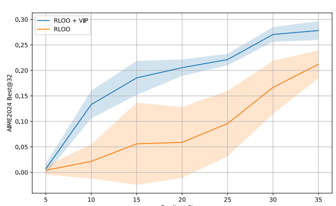


(e) AIME 2025, maj@32

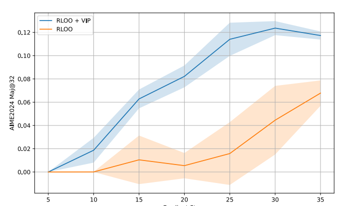


(f) AIME 2025, mean

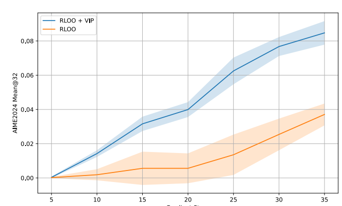
Figure 6: GRPO vs. GRPO+VIP on AIME 2024 and 2025 across different accuracy metrics.



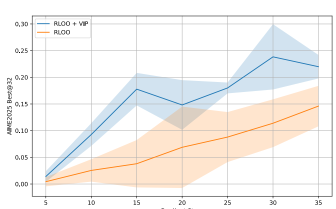
(a) AIME 2024, best@32



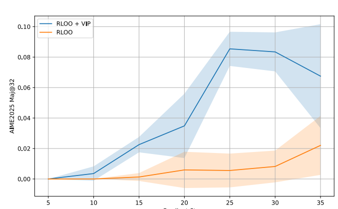
(b) AIME 2024, maj@32



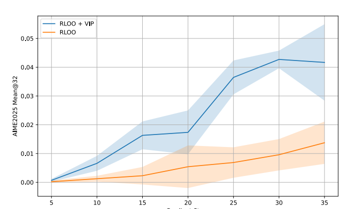
(c) AIME 2024, mean



(d) AIME 2025, best@32



(e) AIME 2025, maj@32



(f) AIME 2025, mean

Figure 7: RLOO vs. RLOO+VIP on AIME 2024 and 2025 across different accuracy metrics.