THE EVOLUTION STRATEGIES PARADOX: BETTER MODELS FROM WORSE PERTURBATIONS

Anonymous authors
Paper under double-blind review

ABSTRACT

How is it possible for perturbation-based methods like Evolution Strategies (ES) to optimize billion-parameter large language models (LLMs)? Classical evolutionary methods are highly dependent on finding good candidates within generated populations. However, when using ES to fine-tune LLMs, a phenomenon can occur where all offspring models consistently perform worse than the parent model, yet with each iteration the parent models consistently improve. This paradox occurs across models, tasks, and even modalities. This paper first characterizes this ES paradox, then examines how the local reward landscape and properties of the ES estimator contribute to it: in high-dimensional spaces, exploration leads to worse performance, but still highlights good directions; aggregating those directions then allows making local improvements. These findings provide guidance for taking advantage of perturbation-based methods in LLM fine-tuning and in high-dimensional optimization in general.

1 INTRODUCTION

Optimization using population-based search, i.e. evolutionary computation, relies on generating and selecting high-quality candidates. Genetic algorithms optimize by discovering beneficial recombinations and mutations and selecting candidates with higher fitness (Zames, 1981). Particle swarm optimization pressures particles to search for better positions while maintaining the previously best found positions (Eberhart & Kennedy, 1995). Canonical evolution strategies (ES) sorts offspring candidates from best to worst, then aggregates the top candidates to form a new parent solution (Vent, 1975). In each of these methods, the generation of higher-fitness candidate solutions is crucial to the optimization process.

For ES specifically, this story has changed over time. With the introduction of natural evolution strategies (NES, Wierstra et al., 2011), which updates the parent model using a weighted sum of all generated candidates, ES was shown to potentially behave as a gradient approximator instead of working through selection of candidates with an improved fitness. NES was a precursor to ensuing large-scale ES runs on reinforcement learning (RL) problems such as Atari (Salimans et al., 2017). These runs inspired theoretical analysis that found that ES with isotropic Gaussian noise is a gradient estimate: with high enough population sizes (in the tens of thousands), correlation between the ES update direction and network gradient would approach 3% while ES performance on a per-iteration basis would match that of backpropagation (Zhang et al., 2017). The conclusion was that ES operated as a gradient estimator that naturally favors robustness (Lehman et al., 2018).

Recently, a new opportunity for ES has emerged: scaling to the vastly larger optimization problem of fine-tuning Large Language Models (LLMs; Qiu et al., 2026). This work shifted from centered rank rewards to z-score normalized rewards, but more surprisingly, it brought the population size down from thousands to just 30 (Qiu et al., 2026; Hayes et al., 2026). With such small populations in a space of billions of parameters, accurate gradient estimation appears infeasible. It was further demonstrated that large pretrained models have a rich search landscape with many directions for improvement (Liang et al., 2026; Gan & Isola, 2026). Taken together, these results suggest that in LLM-scale spaces, ES is not based on gradient estimation, but relies on finding good candidates among sampled perturbations and moving towards them, like other evolutionary methods do.

However, the results in this paper contradict that hypothesis as well: ES is found to make consistent progress even though all offspring candidates are worse than the parent (Figure 1). This paradox is

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

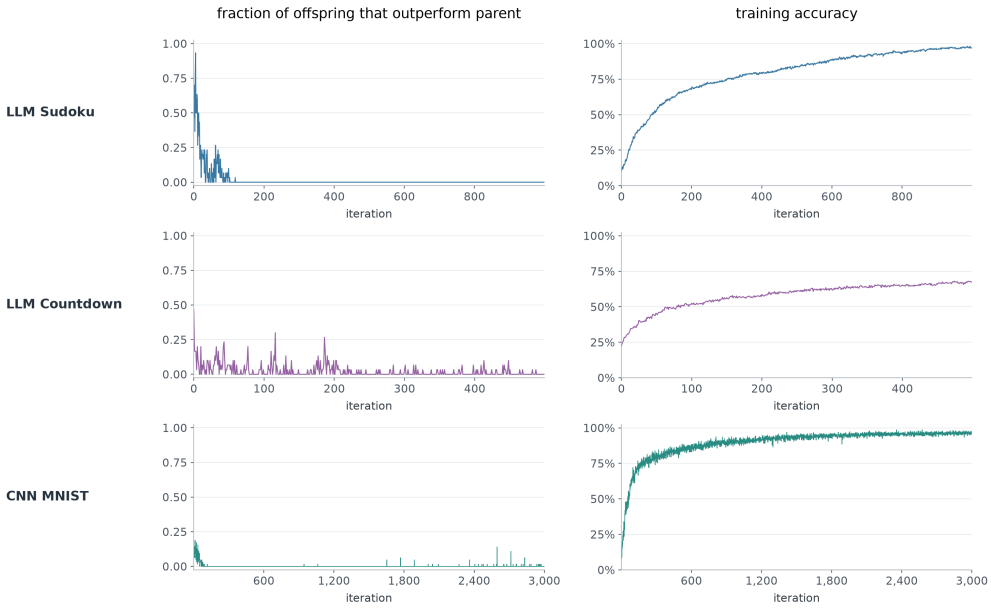

Figure 1: The Evolution Strategies paradox across three examples: Qwen 2.5 3B on GSM8K, a 60K parameter CNN on MNIST, and Qwen 2.5 7B on Sudoku. Left plots show the fraction of offspring models better than the parent model; right plots show the training curves (i.e. the performance of the parent models over time). In each case, offspring models perform worse than the parent model for nearly all of training, while the parent models consistently improve.

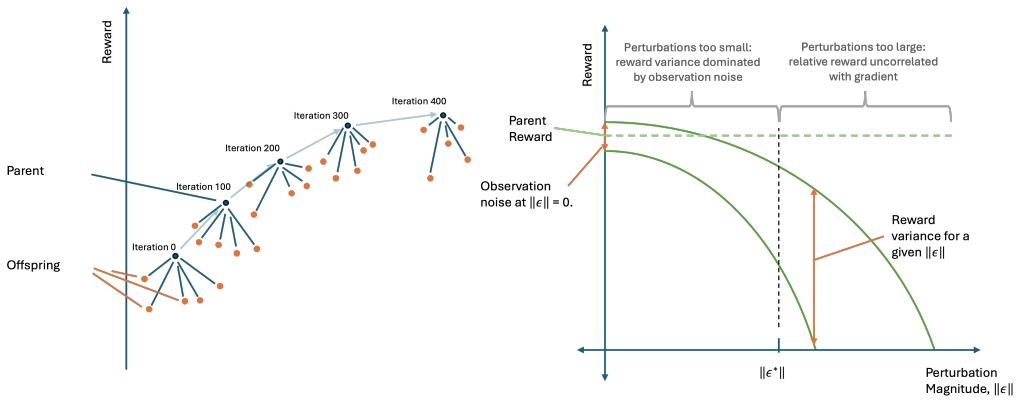

Figure 2: Left: A visual representation of the ES paradox over iterations in a training run. Right: A visual representation of reward vs. perturbation magnitude, demonstrating why the offspring are worse than the parent. Perturbations happen at a fixed size; if they are too small, observation noise drowns out differences between good and bad directions, and if they are too big, the correlation between perturbation reward and direction quality breaks down. The paradox occurs when the magnitude is at a distance where all perturbations have a reward worse than the current model, but they still differentiate the good directions from the bad.

observed across a range of ES optimization tasks, suggesting that it is a general property of high-dimensional ES optimization. A further analysis reveals that it results from three mechanisms (1) the curvature of the search landscape results in decreased performance with increased perturbation

size; (2) low-performing perturbations still differentiate good and bad directions in parameter space; and (3) the ES update follows the good directions while reducing noise (Figure 2).

The remainder of the paper is as follows: Section 2 details the ES estimator and the experimental setup in LLM fine-tuning under which the paradox is observed. The next three sections then focus on each of the three mechanisms. Section 3 runs a landscape analysis, finding that increased perturbation size is strongly tied to a decrease in reward. Section 4 demonstrates that even at larger perturbations where all offspring are worse than the parent, good directions can still be identified. Section 5 uses an analytical model of ES to show that the updates aggregate those good directions while reducing noise, making the paradox likely. These findings thus demonstrate a previously unknown mechanism that allows scaling perturbation-based optimization to modern LLMs.

2 The Evolution Strategies Paradox

This work uses Evolution Strategies (ES) as implemented by Qiu et al. (2026), specifically including isotropic Gaussian perturbations, low population size ($N = 30$), and z-score normalized rewards. For each individual ES update, a population of N perturbations is drawn from isotropic Gaussian noise $\epsilon_i \sim \mathcal{N}(0, I_d)$ and evaluated at $\theta_t + \sigma\epsilon_i$ to obtain reward R_i. The ES update is then a sum of these rewards, weighted by their z-score:

$$\Delta\theta \;=\; \theta_{t+1} - \theta_t \;=\; \frac{\alpha}{N}\sum_{i=1}^{N} Z_i\,\epsilon_i, \qquad Z_i \;=\; \frac{R_i - \mu_R}{\sigma_R}, \tag{1}$$

where α is the step size, μ_R and σ_R are the mean and standard deviation of this population's rewards, and weights Z_i are the z-score normalized rewards. The basic ES training algorithm is shown in Algorithm 1, though for LLM fine-tuning examples, a slightly modified version for computation and memory constraints is used, as described by Qiu et al. (2026).

The ES paradox occurs in event A_t when all offspring at iteration t perform worse than their parent but the update still improves the parent:

$$A_t = \Big\{ R\big(\theta_t + \sigma\epsilon_t^{(i)}\big) < R(\theta_t) \;\; \forall\, i \in \{1, \ldots, N\} \Big\}, \qquad \mathbb{E}_{\epsilon_t}\big[\, R(\theta_{t+1}) - R(\theta_t) \,\big|\, A_t \,\big] > 0. \tag{2}$$

In training runs, the paradox is tracked by recording the fraction of perturbations with rewards better than the current parent on training reward. When this metric stays at or close to zero for extended periods of time while reward increases, it is assumed the paradox is occurring. In this paper, the performance on training data is treated as the full reward landscape; however, the effect can still occur when measured with held-out reward. Monitoring this parent reward adds a marginal inference cost to ES training runs, as one additional model evaluation must be run for each population. Figure 1 displays this metric alongside training progress in fine-tuning in the three example tasks analyzed in this paper. Each training run starts with significant fractions of perturbations outperforming the parent model, but partway into training this metric falls to zero. The effect is most prominent with a fixed batch, though it can occur in shuffled batches; larger batch sizes make the effect stronger, as shown in Figure 5.

This paper's analysis focuses on LLM fine-tuning tasks, where it was first discovered. Two such tasks are analyzed: a 500-iteration training run on a 4×4 Sudoku task, and a 500-iteration training run on a countdown task (Pan et al., 2025), using harnesses and grading designed by Zhao et al. (2025). In each case, Qwen 2.5 7B was used as the LLM, evaluated at temperature zero; ES was run with hyperparameters (N = 30, σ = 0.001, $\alpha = 5 \times 10^{-4}$). Fine-tuning was done with a dataset of 800 problems, with the batch size set to the entire dataset to demonstrate a strong effect (Figure 5). Although the analysis focused on these two specific tasks, the paradox has been observed in other fine-tuning tasks as well, such as GSM8K.

In addition to LLM fine-tuning, the ES paradox occurs in other domains of ES optimization of neural networks. One such domain is included in this paper: training of a randomly initialized 60,000-parameter Convolutional Neural Network (CNN) on MNIST. The paradox was reproduced with ES parameters N = 16, σ = 0.04, and α = 0.02. The effect also increased with batch size for MNIST, occurring in at least 50% of iterations for all batch sizes greater than 20. Thus, the

paradox can be seen as a general phenomenon across models, tasks, and modalities. The next three sections analyze the mechanisms that produce it, and demonstrate that it is a side effect of effective optimization in high-dimensional search spaces.

Algorithm 1 Basic ES Algorithm

Require: Pretrained LLM with initial parameters $\boldsymbol{\theta}_0$, reward function $R(\cdot)$, total iterations T, population size N, noise scale σ, learning rate α.

for $t = 1$ to T **do** ▷ outer ES iterations
  for $n = 1$ to N **do**
    Sample noise $\boldsymbol{\varepsilon}_n \sim \mathcal{N}(0, \boldsymbol{I})$
    Compute reward for perturbed parameters: $R_n \leftarrow R(\boldsymbol{\theta}_{t-1} + \sigma \cdot \boldsymbol{\varepsilon}_n)$
  end for
  Normalize R_n
  Update model parameters as $\boldsymbol{\theta}_t \leftarrow \boldsymbol{\theta}_{t-1} + \alpha \cdot \frac{1}{N} \sum_{n=1}^{N} R_n \boldsymbol{\varepsilon}_n$
end for

3 Why Are Offspring Worse Than the Parent?

As a first step to understand how the ES paradox emerges, this section examines the relationship between perturbation size and reward in fine-tuning tasks. For both Sudoku and Countdown, checkpoints were taken every 100 iterations and 300 perturbations were applied at different magnitudes. These perturbed models were then evaluated on the original fixed training set of 800 problems, producing a distribution of model rewards at each magnitude. The magnitudes were chosen to match the average magnitude of perturbations with $\sigma = \{1e-4, 2.5e-4, 5e-4, 7.5e-4, 1e-3, 1.25e-3, 1.5e-3, 1.75e-3, 2e-3\}$. Figure 3 displays these tests at the base model and at iterations 100 and 500 in the training run, with box-and-whisker plots displayed at each perturbation magnitude.

With the base model, the mean reward is largely constant, but its variation increases approximately linearly as perturbation size increases. By iteration 100, perturbation magnitude has a significant, consistent, negative effect on the reward. By iteration 500, this effect is even more consistent and pronounced. In Sudoku in particular, the rewards at magnitude matching the σ used for training (1e-3) are multiple standard deviations below those of the parent model. Therefore, the perturbation reward distribution strongly depends on the perturbation size (σ). Small σ produces valid options for model improvement, whereas large σ nearly guarantees that no better models will be produced.

If large σ values in the range where the paradox occurs (e.g. $\sigma = 1e-3$) lead to worse performance, why not use σ in the range of $(1e-4, 2.5e-4)$ where improvement is possible? Indeed in Figure 3, when σ is small enough, the distribution sometimes exceeds parent performance, suggesting that opportunities for improvement exist. However, we see when σ is set to these values, ES progress is significantly less, as shown in figure 7. The improvements are lost in the noise, as illustrated in Figure 2b and discussed in detail in Section 5. However, even though large σ values result in degraded performance, it turns out they can be used to identify good directions reliably, as discussed in the next section.

4 How Are Good Directions Identified?

Section 3 leads to the question: When all models perform worse than the current model, does their relative performance still indicate which directions are more promising? Although it is not possible to directly measure the reward gradient for model checkpoints, it is possible to evaluate the relative performance of perturbations across magnitudes where improving directions do and do not occur. If significant correlations are seen across magnitudes, regardless of performance, then relative performance is indicative of good directions regardless of all perturbation offspring performing worse than the parent.

To test this hypothesis, the checkpoints from Sudoku and Countdown runs were again used. The set of 300 perturbation directions was evaluated on each task's respective 800-problem dataset at

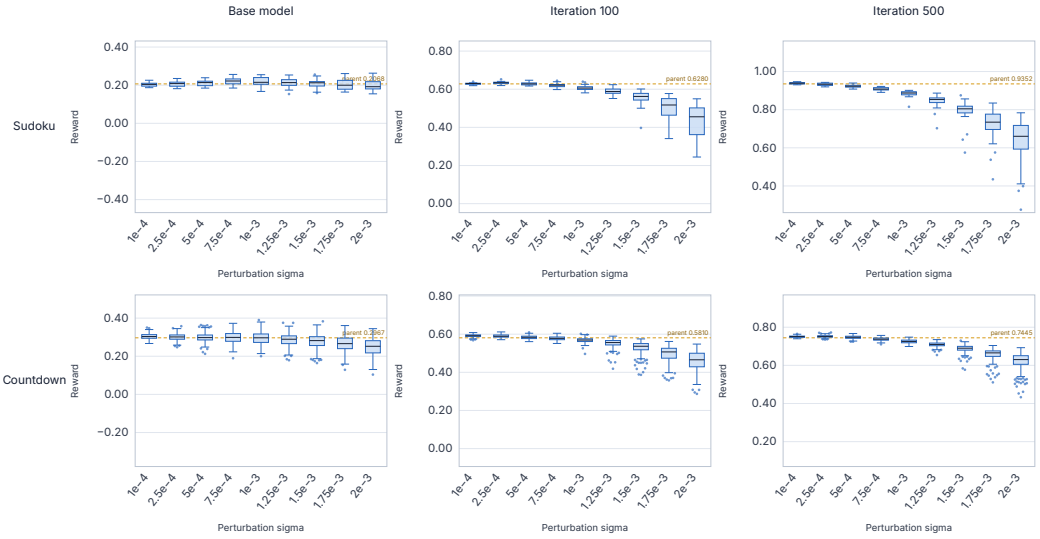

Figure 3: Reward distributions at different perturbation magnitudes. Checkpoints in Countdown and Sudoku training runs were perturbed 300 times at each specified magnitude, and distributions of rewards were plotted. As the magnitudes increase, the means decrease and variance increases. Thus, the offspring are likely to perform worse than the parent.

magnitudes corresponding to $\sigma = \{1e-4, 2.5e-4, 5e-4, 7.5e-4, 1e-3, 1.25e-3, 1.5e-3, 1.75e-3, 2e-3\}$. A correlation matrix was then constructed between the relative rewards of perturbations at these magnitudes. If the perturbation rewards at large σ contain signal, then they should correlate with the rewards of the same perturbations at small σ. Indeed, the rewards are significantly correlated between large and small perturbations, even when the rewards of large perturbations are significantly lower than those of the small perturbations. While greater magnitude differences reduce the measured correlation, the correlations still remain positive in most cases.

These correlations likely emerge due to a special property of ES with isotropic Gaussian noise in high-dimensional settings. This noise occupies a thin shell in what has been referred to as a "soap bubble" effect (Huszár, 2017). In a high-dimensional space, all perturbations are nearly identical in norm, effectively canceling out any isotropic curvature in the reward landscape, while leaving differences due to the gradient and anisotropic curvature. In other words, the magnitude of noise is identical in all directions, and thus relative performance allows signal to still emerge.

One way to take advantage of the good directions at large σ would be to simply follow the best perturbation, instead of doing the aggregation through a weighted sum. To evaluate this ablation, a second test was implemented. The perturbation rewards were sampled on the standard (large) σ (1e-3), the best of them identified, and the perturbation was scaled down to the size of an average ES update (i.e. a small σ). A 500-iteration training run was then run with this process. The training curves along with the fraction of perturbations better than the parent are displayed in Figure 8. While this best-of-N selection allows making progress, it is still outperformed by the actual ES update. This result suggests that an additional mechanism of aggregation contributes further to the ES performance, as will be discussed next.

5 How Do the Updates Improve Performance?

Sections 3 and 4 demonstrated that model performance decreases with increasing perturbation; however, even when all offspring are worse than the parent, good search directions can be identified among them. This section demonstrates how the ES aggregation mechanism adds value. The in-



Figure 4: Correlation between rewards at various perturbation magnitudes σ at iterations 200 and 500. The σ range at which the paradox occurs, i.e. the best of the 30 offspring has a lower reward than the parent, is labeled with red arrows. The rewards at small and large σ correlate significantly. Thus, the same good directions can be identified at various perturbation magnitudes, even at which all the perturbations are worse than the parent.

sights are drawn from an analytical model of the ES update in a reward landscape with observation noise and quadratic loss. In this model, the ES update can be decomposed into a *signal* component along the gradient and a *noise* component orthogonal to it. The model shows that the ES update holds the signal constant while reducing noise, and the optimal perturbation magnitude makes the paradox likely. These results are then validated empirically with existing training runs.

In the subsections below, the update decomposition is first defined in a linear reward landscape. A realistic observation noise term is then introduced, and the model extended to include curvature. In the end, the optimal perturbation magnitude is determined in the model.

5.1 Signal and Noise Components of the ES Update

Recall the definition of ES update in Equation 18. Let $g = \nabla R(\theta_t)$ be the gradient in the reward landscape and $\hat{g} = g/\|g\|$ its normalized direction. Each perturbation ϵ_i can be divided into a component along the gradient, $u_i\,\hat{g}$, and a perpendicular component, $\epsilon_i^{\perp}$, as

$$\epsilon_i \;=\; u_i\,\hat{g} \;+\; \epsilon_i^{\perp}, \qquad u_i = \hat{g}^{\top}\epsilon_i, \qquad \epsilon_i^{\perp} \sim \mathcal{N}\big(0,\, I_d - \hat{g}\hat{g}^{\top}\big). \tag{3}$$

Substituting this decomposition into Equation 18,

$$\Delta\theta_{\parallel} \;=\; \alpha\Big(\frac{1}{N}\sum_{i=1}^{N} Z_i u_i\Big)\hat{g}, \qquad \Delta\theta_{\perp} \;=\; \frac{\alpha}{N}\sum_{i=1}^{N} Z_i\,\epsilon_i^{\perp}. \tag{4}$$

Under linear reward, $R(\theta) = R(\theta_t) + g^{\top}(\theta - \theta_t)$. Then by Equation 26 and $\hat{g} = g/\|g\|$, $R_i = R(\theta_t) + \sigma\|g\|\,u_i$. Given the assumption that the sampled reward mean and standard deviation match their population statistics, $\mu_R = R(\theta_t)$ and $\sigma_R = \sigma\|g\|$. An individual perturbation's z-score is then

$$Z_i \;=\; \frac{\sigma\|g\|\,u_i}{\sigma\|g\|} \;=\; u_i. \tag{5}$$

Both σ and $\|g\|$ cancel, leaving the $Z_i = u_i$. The parallel component in Equation 27 is $\alpha \cdot \frac{1}{N}\sum_i u_i^2$ with $\sum_i u_i^2 \sim \chi^2_N$, so the expected value of its magnitude is α, and the variance of its magnitude is $2\alpha^2/N$. For the orthogonal component, $Z_i = u_i$ is independent of $\epsilon_i^\perp$. Further, $\mathbb{E}[u_i^2] = 1$ and $\mathbb{E}\|\epsilon_i^\perp\|^2 = d-1$. Hence:

$$\text{signal:} \qquad \mathbb{E}[\hat{g}^\top \Delta\theta] \;=\; \alpha, \tag{6}$$

$$\text{noise:} \qquad \sqrt{\mathbb{E}\|\Delta\theta_\perp\|^2} \;=\; \alpha\sqrt{\frac{d-1}{N}}. \tag{7}$$

$$\mathbb{E}[\Delta R] \;=\; \alpha\|g\|, \qquad \mathrm{Var}[\Delta R] \;=\; \frac{2\alpha^2\|g\|^2}{N}. \tag{8}$$

Notice that the signal component of the ES update is constant and unaffected by either the gradient or the population size. The noise term however increases with dimensionality and decreases with population size. This result demonstrates that the ES estimator works by reducing noise, not by modifying the signal component. Additionally, in high-dimensional spaces, noise dominates, implying that the change in reward results from a minuscule portion of the update step. For full derivations see B.

5.2 Including Observation Noise

In the ES fine-tuning tasks, the models are evaluated at temperature 0, and the model outputs are therefore technically deterministic. However, when exploring along specified directions, noise occurs in the reward landscape at much higher frequencies than the magnitudes of perturbations used, as shown in Figure 9. Similar to the nugget term in Gaussian process models (Pepelyshev, 2010), this noise can be incorporated in the analytical model as observation noise.

The noise in the decomposition above is due to inaccurate gradient direction. The next step is to include observation noise in the model, i.e. inaccuracy in evaluating the quality of offspring:

$$R_i \;=\; R(\theta_t + \sigma\epsilon_i) \;+\; \sigma_\xi\,\xi_i, \qquad \xi_i \sim \mathcal{N}(0,1), \quad \xi_i \perp \epsilon_i. \tag{9}$$

Plugging in this reward function into the ES update (equation 27) produces the expected signal, noise, and reward as

$$\mathbb{E}[\hat{g}^\top \Delta\theta] = \frac{\alpha\sigma\|g\|}{\sqrt{\sigma^2\|g\|^2 + \sigma_\xi^2}}, \quad \sqrt{\mathbb{E}\|\Delta\theta_\perp\|^2} = \alpha\sqrt{\frac{d-1}{N}}, \quad \mathbb{E}[\Delta R] = \frac{\alpha\sigma\|g\|^2}{\sqrt{\sigma^2\|g\|^2 + \sigma_\xi^2}} \tag{10}$$

Notice that with noise, the expected signal and reward increase with increasing σ. When observation noise is included, the variance in the reward for small σ is mostly produced by noise. Larger σ increases reward variance, making the ES estimate more accurate. For full derivations see B.

5.3 Including Landscape Curvature

In the third step, the analytical model is extended to include curvature Q, with no assumptions made on its properties. The reward can then be written as

$$R(\theta) = R(\theta_t) + g^\top(\theta - \theta_t) - \tfrac{1}{2}(\theta - \theta_t)^\top Q\,(\theta - \theta_t) \tag{11}$$

The expected signal, noise, and reward then are

$$\text{signal:} \qquad \mathbb{E}[\hat{g}^\top \Delta\theta] \;=\; \frac{\alpha\sigma\|g\|}{\sigma_R} \;=\; \frac{\alpha\|g\|}{\sqrt{\|g\|^2 \;+\; \dfrac{\sigma^2}{2}\,\mathrm{tr}[Q^2] \;+\; \dfrac{\sigma_\xi^2}{\sigma^2}}}. \tag{12}$$

$$\text{noise:} \qquad \sqrt{\mathbb{E}\|\Delta\theta_\perp\|^2} \;\approx\; \alpha\sqrt{\frac{d-1}{N}} \tag{13}$$

$$\mathbb{E}[\Delta R] \;\approx\; \underbrace{\frac{\alpha\|g\|^2}{\sqrt{\|g\|^2 \;+\; \frac{\sigma^2}{2}\,\mathrm{tr}[Q^2] \;+\; \frac{\sigma_\xi^2}{\sigma^2}}}}_{\text{signal gain}} \;-\; \underbrace{\frac{\alpha^2}{2N}\,\mathrm{tr}\,Q}_{\text{noise penalty}} \tag{14}$$

where the terms that are negligible for noise and reward have been dropped. With curvature, the signal gain creates a tradeoff between increasing perturbation size to avoid observation noise σ_ξ^2 and the trace of the Hessian $\mathrm{tr}[Q^2]$. This tradeoff has been well documented with perturbation-based methods [cite][cite][cite]. However, in high dimensions, the ES estimate has two unique properties. First, the value of the marginal population member is decreasing in N, and only shows up in the penalty for curvature. Thus, the expected gain in reward plateaus once the population size reduces noise enough so that it is no longer affected by curvature. Second, the update magnitude is expected to decrease while population size increases, meaning that the update magnitude decreases at a rate of roughly $\frac{1}{\sqrt{N}}$, while the reward improvement increases then plateaus.

To test these claims, checkpoints were taken at every 100 iterations from the Sudoku 500 iteration fine-tuning run, as well as from the MNIST training run. With each checkpoint, 30 ES updates were drawn from independent sets of N perturbations, where N is the designated population size. The reward of these updates was measured and confidence intervals drawn for the mean reward update. Figure 10 shows these results for fine-tuning, and figure 6 for MNIST. Notice that, as predicted by the noise-curvature model, the expected reward update increases and then plateaus with population size. Further, the update distributions in these figures show that the update sizes are highly consistent, and even as they decrease significantly with population, the reward improvement remains the same. This result demonstrates that the ES update is holding the signal constant while reducing noise.

These plots also illuminate why ES is able to work on billion-parameter models with population sizes of only 30: once the noise is reduced enough to avoid the curvature penalty, ES achieves its maximum possible improvement on each step. For full derivations see B.

5.4 Is the Paradox Expected?

Testing alternative σ that avoided the paradox revealed that perturbation magnitudes that invoked the paradox outperformed smaller perturbations that did not. The noise-curvature model offers a tradeoff between observation noise and curvature. If we hold α, g, Q and σ_ξ fixed, the expected improvement in Equation 48 is maximized by minimizing $f(\sigma^2) = \sigma_\xi^2/\sigma^2 + \frac{\sigma^2}{2}\,\mathrm{tr}[Q^2]$ over σ^2, which gives

$$\sigma_*^4 \;=\; \frac{2\,\sigma_\xi^2}{\mathrm{tr}[Q^2]}, \qquad \sigma_* \;=\; \left(\frac{2\,\sigma_\xi^2}{\mathrm{tr}[Q^2]}\right)^{1/4} \tag{15}$$

To characterize the population at this optimal sigma, using the fact that $\mathbb{E}[\epsilon_i^\top Q \epsilon_i] = \mathrm{tr}\,Q$, its difference in the reward from its parent are:

$$\mu_R - R(\theta_t) \;=\; -\frac{\sigma^2}{2}\,\mathrm{tr}\,Q, \qquad \mu_R - R(\theta_t)\Big|_{\sigma_*} \;=\; -\,\sigma_\xi\sqrt{\frac{r_{\text{eff}}}{2}}, \qquad r_{\text{eff}} := \frac{(\mathrm{tr}\,Q)^2}{\mathrm{tr}[Q^2]}, \tag{16}$$

where $r_{\text{eff}} \in [1, d]$ is the participation ratio (effective rank) of the Hessian, equal to d for isotropic curvature and to r for a rank-r Hessian with equal eigenvalues. In units of the population spread,

$$\frac{R(\theta_t) - \mu_R}{\sigma_R}\Big|_{\sigma_*} \;=\; \frac{\frac{\sigma^2}{2}\,\mathrm{tr}\,Q}{\sqrt{\sigma^2\|g\|^2 \;+\; \frac{\sigma^4}{2}\,\mathrm{tr}[Q^2] \;+\; \sigma_\xi^2}} \;=\; \frac{\mathrm{tr}\,Q}{2\sqrt{\frac{1}{\sigma^2}\|g\|^2 \;+\; \frac{1}{2}\,\mathrm{tr}[Q^2] \;+\; \frac{\sigma_\xi^2}{\sigma^4}}}, \tag{17}$$

where σ_R is given by equation 41 in Appendix B.4. Equation 54 quantifies how the rewards of perturbations distribute compared to the parent reward $R(\theta_t)$. A larger value of equation 54 corresponds to a higher probability for the perturbation to be worse than the parent.

Since the value of equation 54 is positively correlated with σ, a smaller σ decreases the probability for generating worse populations. In the mean time, a smaller σ ($\ll \sigma_*$) also reduces the expected

improvement of ES updates according to equation 48. As a result, smaller σ decreases the chance for the paradox to happen. For larger σ, although the chance for producing worse perturbations is increased, the expected improvement is also reduced (if $\gg \sigma_*$), resulting in less happenings of the paradox. The results in Figure 7 and Figure 11 empirically show that if a σ value provides a larger improvement for ES updates, its chance to trigger the paradox also increases. Therefore, the paradox is a byproduct of a highly effective ES run.

6 Discussion and Future Work

Small Population Performance of ES One of the most dramatic features of the new line of ES work in LLM fine-tuning has been the dramatically smaller population size. Previously it had been assumed that ES had fundamental scaling limitations with respect to high dimensional spaces [cite], yet within fine-tuning it has been competitive with, if not outperforming gradient-based reinforcement learning methods. This work now sheds some light on this phenomenon, demonstrating that once the noise component of the ES update is small enough to avoid incurring a significant cost from curvature, the update is able to achieve the maximum improvement in reward - even if the noise component is orders of magnitude larger than the signal component.

The surprising Robustness of LLMs to Noise in Parameter Space From the update reward improvement estimate results in 10, per step update is seen to plateau at approximately a population size of 30. Given the analysis in section 5, this would imply each ES update contains a signal component of maximum magnitude α, while the noise component contains a magnitude of $\alpha\sqrt{\frac{d-1}{N}} = \alpha\sqrt{\frac{7B}{30}} \approx \alpha * 15,000$. While fine-tuning, for every unit distance ES moves the network in direction of model improvement, it moves 15,000 unit distances in an irrelevant direction. This demonstrates a surprising fact that LLMs can have their parameters dramatically modified in random directions without affecting performance.

Future Work Results and the analytical model suggest that there may be ways to further optimize the hyperparameters of ES. Section 5 presented an optimal perturbation size conditional on observation noise and curvature. The expected reward presented in said section may also be useful for analytically deriving an optimal σ. Future work may look at how to identify these properties during a run, or whether estimation of said optimums is too noisy to do online, but may be aggregated across fine-tuning runs. Additional work could also modify the ES algorithm itself to take advantage of this new paradigm where improving offspring are no longer necessary for population-based optimization.

7 Conclusion

The ES paradox presents a counterintuitive conclusion: that ES can learn while consistently generating worse candidates than its parent. This paper demonstrates why this is possible: even when all offspring perform worse than their parent, their relative performance contains valuable information for improving directions. The ES update then aggregates this signal while reducing the noise to a point where improvement can be made. An analytical model was then demonstrated this was possible with a simple yet reasonable model of the reward landscape: observation noise in reward estimation, along with reward landscape curvature. In explaining the paradox this model also illuminated another property of the ES update in high dimensions: that expected improvement was constant regardless of population size, while update magnitude decreased with population size. This illustrated why population sizes of 30 may be a sweet spot for LLM fine-tuning with ES, as that appears to be the population size at which curvature no longer punishes the noise component of updates and allows ES to achieve its maximum improvement. Together, these results provide insight for how ES is able to function effectively on billion-parameter LLMs, and how improvements to population-based methods may be made in the future.

AI USE STATEMENT

In this work, we used generative AI tools for helping implement methods and provide critical ingredients for proving mathematical claims. We have reviewed all AI-assisted work. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

REFERENCES

Russell Eberhart and James Kennedy. A new optimizer using particle swarm theory. In *MHS'95. Proceedings of the sixth international symposium on micro machine and human science*, pp. 39–43. Ieee, 1995.

Yulu Gan and Phillip Isola. Neural thickets: Diverse task experts are dense around pretrained weights. In *Forty-third International Conference on Machine Learning*, 2026.

Conor F. Hayes, Elliot Meyerson, Kajetan Schweighofer, Roberto Dailey, Babak Hodjat, Risto Miikkulainen, and Xin Qiu. Beyond the best guess: Improving llm solution coverage with evolution strategies, 2026. URL https://arxiv.org/abs/2608.12679.

Ferenc Huszár. Gaussian distributions are soap bubbles. inFERENCe (blog), November 2017. URL https://www.inference.vc/high-dimensional-gaussian-distributions-are-soap-bubble/. Accessed: 2026-09-17.

Joel Lehman, Jay Chen, Jeff Clune, and Kenneth O Stanley. Es is more than just a traditional finite-difference approximator. In *Proceedings of the genetic and evolutionary computation conference*, pp. 450–457, 2018.

Qiyao Liang, Jinyeop Song, Yizhou Liu, Jeff Gore, Ila Fiete, Risto Miikkulainen, and Xin Qiu. The blessing of dimensionality in llm fine-tuning: A variance-curvature perspective, 2026. URL https://arxiv.org/abs/2602.00170.

Jiayi Pan, Junjie Zhang, Xingyao Wang, Lifan Yuan, Hao Peng, and Alane Suhr. Tinyzero. https://github.com/Jiayi-Pan/TinyZero, 2025. Accessed: 2025-01-24.

Andrey Pepelyshev. The role of the nugget term in the gaussian process method, 2010. URL https://arxiv.org/abs/1005.4385.

Xin Qiu, Yulu Gan, Conor F. Hayes, Qiyao Liang, Yinggan Xu, Roberto Dailey, Elliot Meyerson, Babak Hodjat, and Risto Miikkulainen. Evolution strategies at scale: Llm fine-tuning beyond reinforcement learning, 2026. URL https://arxiv.org/abs/2509.24372.

Tim Salimans, Jonathan Ho, Xi Chen, Szymon Sidor, and Ilya Sutskever. Evolution strategies as a scalable alternative to reinforcement learning, 2017. URL https://arxiv.org/abs/1703.03864.

W Vent. Rechenberg, ingo, evolutionsstrategie—optimierung technischer systeme nach prinzipien der biologischen evolution. 170 s. mit 36 abb. frommann-holzboog-verlag. stuttgart 1973. broschiert, 1975.

Daan Wierstra, Tom Schaul, Tobias Glasmachers, Yi Sun, and Jürgen Schmidhuber. Natural evolution strategies, 2011. URL https://arxiv.org/abs/1106.4487.

G Zames. Genetic algorithms in search, optimization and machine learning. *Inf Tech J*, 3(1):301, 1981.

Xingwen Zhang, Jeff Clune, and Kenneth O. Stanley. On the relationship between the openai evolution strategy and stochastic gradient descent, 2017. URL https://arxiv.org/abs/1712.06564.

Siyan Zhao, Devaansh Gupta, Qinqing Zheng, and Aditya Grover. d1: Scaling reasoning in diffusion large language models via reinforcement learning, 2025. URL https://arxiv.org/abs/2504.12216.

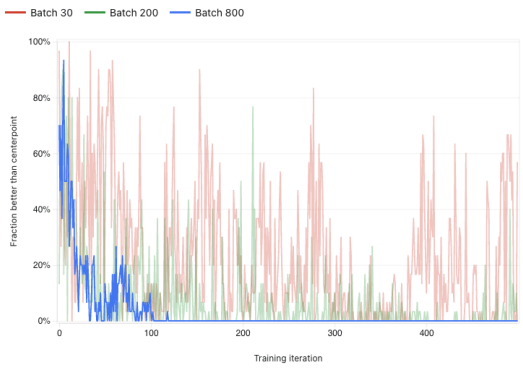

Figure 5: Fraction of perturbations better than the parent in the Sudoku fine-tuning task over iterations in the training run. Larger batch size leads to less noise in the reward landscape, and makes the paradox more prominent.

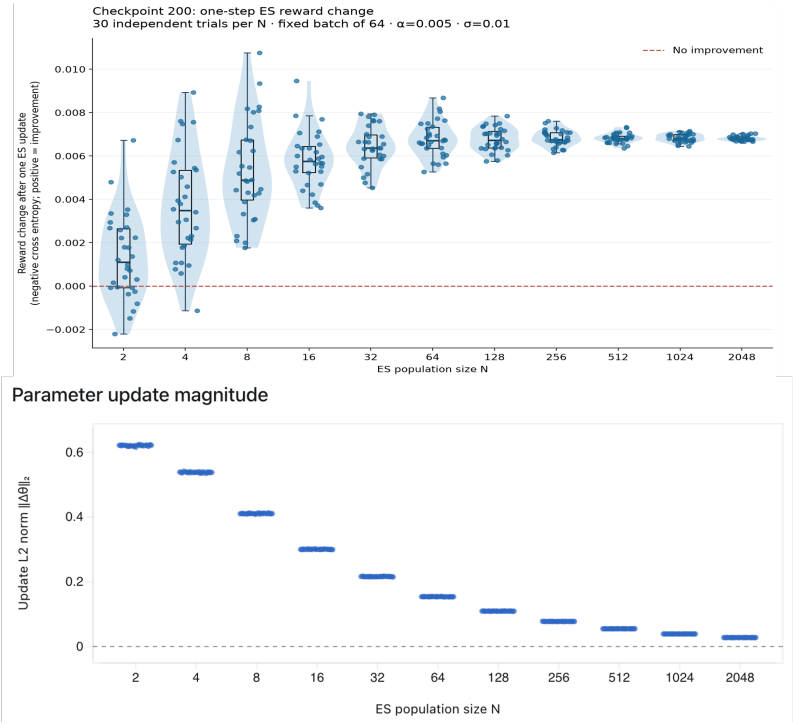

Figure 6: Top: Distributions of reward improvements from ES updates for different population sizes for CNN on MNIST. Bottom: norm of ES updates in MNIST. All ES updates were sampled from iteration 200 of a 500-iteration training run on MNIST. Update reward improvement increases in population size then plateaus between population sizes 32 and 64. At the same time, norm of the ES update decreases with population size.

A Complementary Figures

This section includes the figures that complement the main paper.

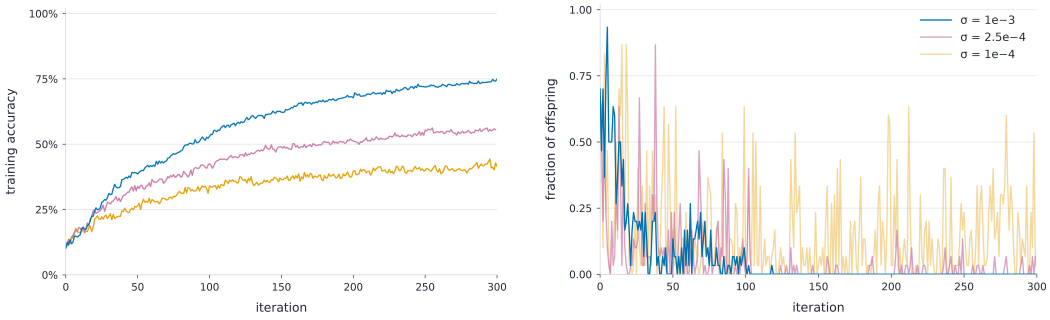

Figure 7: Left: A comparison in fine-tuning performance on Sudoku with perturbation magnitudes where the paradox occurs ($\sigma = 1e-10$) vs perturbation magnitudes where offspring produce viable improvement ($\sigma \in 1e-4, 2.5e-4$). Right: The fraction of better offspring compared to parent under different perturbation magnitudes. Even though the smaller perturbation magnitudes produce better candidates, the larger perturbations produce better directions and thus a better overall training run.

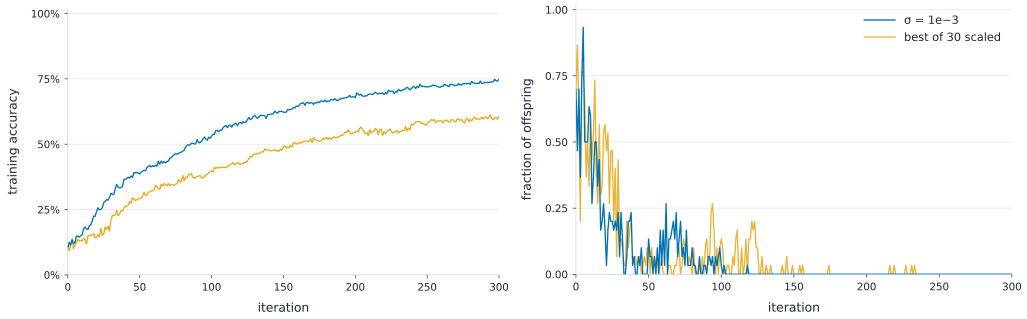

Figure 8: Left: Training performance of scaled best-of-30 perturbation selection compared with the actual aggregated ES update. Right: Fraction of the perturbations with the best-of-30 method that are better than the parent. When scaled from a size where the best perturbation performs worse to the size of a typical ES update, the model is able to improve. However, the aggregated ES update direction still significantly outperforms this best-of-30 direction, suggesting that aggregation contributes further to performance.

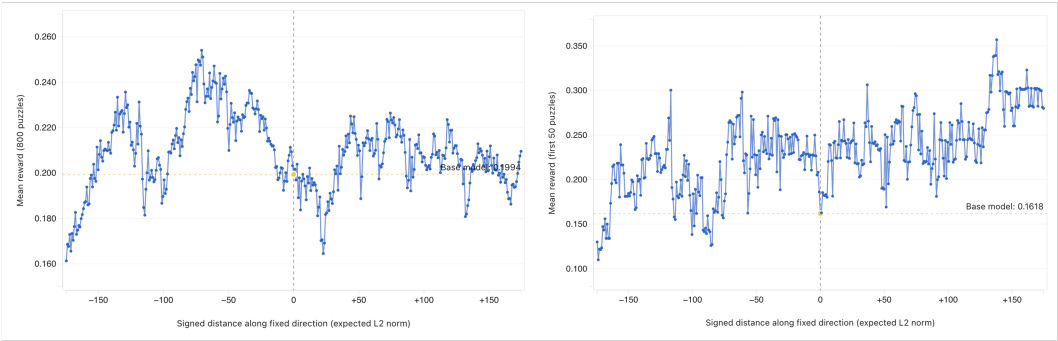

Figure 9: Average accuracy of Qwen 2.5 7B across 800 Sudoku tasks along a random direction from -2σ to $+2\sigma$. Left: Landscape on a set of 800 problems. Right: Landscape on a set of 50 problems. A significant portion of the variance occurs at frequencies much higher than the perturbation size used in ES, and can thus be treated as observation noise.

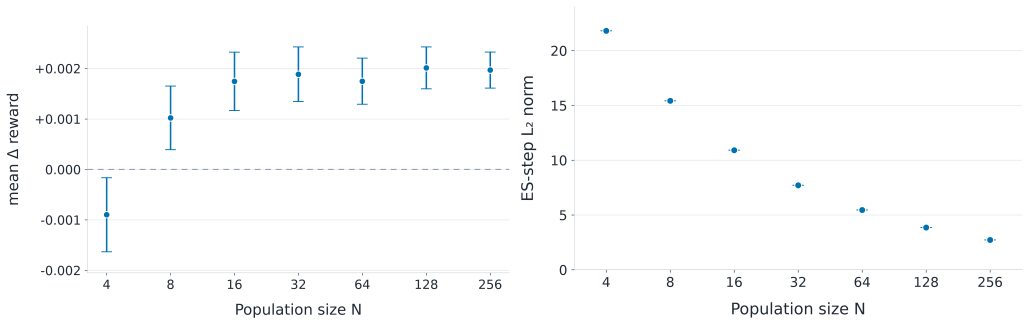

Figure 10: Left: Mean and confidence intervals of improvement from ES updates at different population sizes for the Sudoku LLM fine-tuning task. Right: Norm of ES update for various population size. Distributions were wide so for clarity only means and CI's are shown. The model was Qwen 2.5 7B. Update reward improvement increases in population size then plateaus between 16 and 32. At the same time, the norm of the ES update decreases with population size.

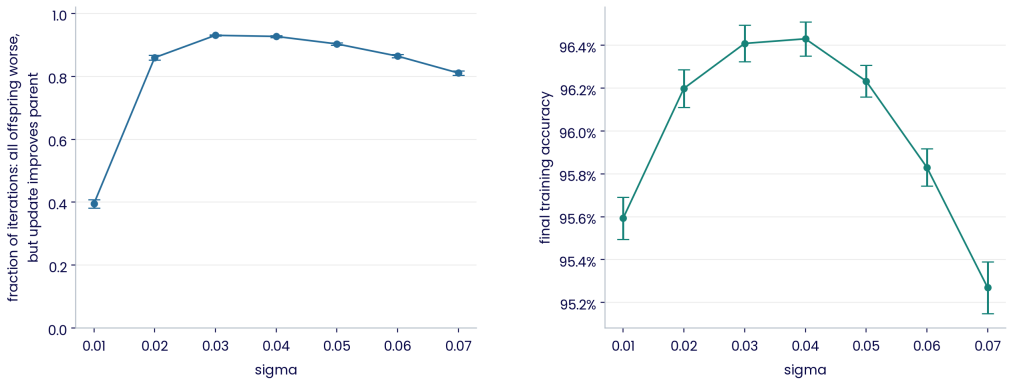

Figure 11: ES training of CNN on MNIST with various sigmas to 3,000 iterations. Each training run is rerun 16 times to produce confidence intervals. Left: Fraction of iterations where all offspring are worse vs σ. Right: Final training accuracy. The optimal σ for training occurs where the paradox is likely to occur.

B Signal, Noise, and the Perturbation Scale in z-Scored Evolution Strategies

This section starts with the basic properties of the ES estimator with z-score normalized rewards Qiu et al. (2026), under linear and quadratic reward landscapes. Moving beyond the original derivation, it also examines the noise and signal decomposition of every ES update, as well as the implied optimal σ from the curvature and observation noise.

B.1 The z-scored ES update

At iteration t, ES draws N perturbation directions $\epsilon_i \sim \mathcal{N}(0, I_d)$, evaluates the perturbed models $\theta_t + \sigma\epsilon_i$ to obtain rewards R_i, and moves the centre by the reward-weighted average of the perturbations,

$$\Delta\theta \;=\; \theta_{t+1} - \theta_t \;=\; \frac{\alpha}{N}\sum_{i=1}^{N} Z_i\,\epsilon_i, \tag{18}$$

where α is the step size and the weights Z_i are the population-standardized rewards,

$$Z_i \;=\; \frac{R_i - \mu_R}{\sigma_R}, \qquad \mu_R = \frac{1}{N}\sum_{j=1}^{N} R_j, \qquad \sigma_R^2 = \frac{1}{N}\sum_{j=1}^{N}(R_j - \mu_R)^2. \tag{19}$$

The sample statistics (μ_R, σ_R) in equation 19 are replaced by their population counterparts, $\mu_R = \mathbb{E}[R_i]$ and $\sigma_R^2 = \mathrm{Var}[R_i]$, in the derivations below.

Perturbation and update magnitudes. Before specifying any reward landscape, two magnitudes are fixed by the sampling and normalization alone. The norm of each perturbation follows a scaled chi distribution, $\|\sigma\epsilon_i\| \sim \sigma\chi_d$, so its first two moments are exact:

$$\mathbb{E}\|\sigma\epsilon_i\| \;=\; \sigma\,\frac{\sqrt{2}\,\Gamma\!\left(\frac{d+1}{2}\right)}{\Gamma\!\left(\frac{d}{2}\right)} \;=\; \sigma\sqrt{d}\left(1 - \frac{1}{4d} + O(d^{-2})\right), \tag{20}$$

$$\mathrm{Var}\,\|\sigma\epsilon_i\| \;=\; \sigma^2\Big(d - \big(\mathbb{E}\|\epsilon_i\|\big)^2\Big) \;\xrightarrow{d\to\infty}\; \frac{\sigma^2}{2}. \tag{21}$$

The standard deviation of the perturbation length tends to the constant $\sigma/\sqrt{2}$, so its relative spread is $1/\sqrt{2d}$.

The standardized weights satisfy, by construction and irrespective of the rewards,

$$\sum_{i=1}^{N} Z_i \;=\; 0, \qquad \sum_{i=1}^{N} Z_i^2 \;=\; N. \tag{22}$$

The first identity removes any constant offset in the rewards. The second fixes the update length. Because each Z_i depends only on its own perturbation, the weighted perturbations $Z_i\epsilon_i$ are identically distributed, and the landscape enters the moments of $\|\Delta\theta\|^2$ through only two scalars: the squared drift $\|m\|^2$ with $m = \mathbb{E}[Z_i\epsilon_i]$, which satisfies $\|m\| \le 1$ by Cauchy–Schwarz, and the inflation $\kappa = \mathbb{E}[Z_i^2\|\epsilon_i\|^2] - d$, which the derivations below show lies in $[0, 4]$. Expanding $\|\Delta\theta\|^2$ into its N diagonal terms $Z_i^2\|\epsilon_i\|^2$ and $N(N-1)$ cross terms $Z_iZ_j\,\epsilon_i^\top\epsilon_j$, the diagonal sum is pinned to $Nd + O(N)$ by equation 22 while each cross term has mean $\|m\|^2$ and variance d, giving (Appendix C)

$$\mathbb{E}\|\Delta\theta\|^2 \;=\; \frac{\alpha^2 d}{N} + \frac{\alpha^2\kappa}{N} + \alpha^2\Big(1 - \frac{1}{N}\Big)\|m\|^2 \;=\; \frac{\alpha^2 d}{N}\left(1 + O\big(N/d\big)\right), \tag{23}$$

$$\mathrm{Var}\,\|\Delta\theta\|^2 \;=\; \frac{2\alpha^4 d}{N^2}\left(1 + O\big(N/d\big)\right), \qquad \frac{\sqrt{\mathrm{Var}\,\|\Delta\theta\|^2}}{\mathbb{E}\|\Delta\theta\|^2} \;=\; \sqrt{\frac{2}{d}}\left(1 + O\big(N/d\big)\right). \tag{24}$$

The length of a z-scored ES step is therefore set by (α, d, N), with coefficient of variation $\sqrt{2/d}$ so that $\sqrt{\mathbb{E}\|\Delta\theta\|^2}$ is accurate to relative order $1/d$; the rewards determine its direction, through m, and contribute to its length only at relative order N/d.

Signal and noise. Let $g = \nabla R(\theta_t)$ be the local gradient and $\hat{g} = g/\|g\|$ its direction. We decompose every update into the component along the gradient and the component orthogonal to it,

$$\Delta\theta \;=\; \underbrace{\Delta\theta_{\parallel}}_{\text{signal}} \;+\; \underbrace{\Delta\theta_{\perp}}_{\text{noise}}, \qquad \Delta\theta_{\parallel} = (\hat{g}^{\top}\Delta\theta)\,\hat{g}, \qquad \Delta\theta_{\perp} = (I_d - \hat{g}\hat{g}^{\top})\,\Delta\theta, \tag{25}$$

and characterize the signal by its expected projection $\mathbb{E}[\hat{g}^{\top}\Delta\theta]$ and the noise by its root-mean-square magnitude $\sqrt{\mathbb{E}\|\Delta\theta_{\perp}\|^2}$.

To carry out the decomposition, split each perturbation into its gradient projection and its orthogonal remainder,

$$\epsilon_i \;=\; u_i\,\hat{g} \;+\; \epsilon_i^{\perp}, \qquad u_i = \hat{g}^{\top}\epsilon_i \sim \mathcal{N}(0,1), \qquad \epsilon_i^{\perp} \sim \mathcal{N}\big(0,\, I_d - \hat{g}\hat{g}^{\top}\big), \tag{26}$$

with u_i and $\epsilon_i^{\perp}$ independent. Substituting into equation 18,

$$\Delta\theta_{\parallel} \;=\; \alpha\Big(\frac{1}{N}\sum_{i=1}^{N} Z_i u_i\Big)\hat{g}, \qquad \Delta\theta_{\perp} \;=\; \frac{\alpha}{N}\sum_{i=1}^{N} Z_i\,\epsilon_i^{\perp}. \tag{27}$$

In every case considered below, $\mathbb{E}[Z_i\epsilon_i^{\perp}] = 0$, because each term of Z_i is either odd in $\epsilon_i^{\perp}$ or independent of it. The cross terms between different i in $\|\Delta\theta_{\perp}\|^2$ therefore vanish, and

$$\mathbb{E}[\hat{g}^{\top}\Delta\theta] \;=\; \alpha\,\mathbb{E}[Z_i u_i], \qquad \mathbb{E}\|\Delta\theta_{\perp}\|^2 \;=\; \frac{\alpha^2}{N}\,\mathbb{E}\big[Z_i^2\|\epsilon_i^{\perp}\|^2\big]. \tag{28}$$

The three cases below add one ingredient at a time: a linear landscape with exactly observed rewards, a linear landscape with observation noise, and a quadratic landscape with observation noise.

B.2 Linear landscape without observation noise

Let the reward be locally linear, $R(\theta) = R(\theta_t) + g^{\top}(\theta - \theta_t)$, and suppose each reward is observed exactly, $R_i = R(\theta_t + \sigma\epsilon_i)$. Then $R_i = R(\theta_t) + \sigma\|g\|\,u_i$, so $\mu_R = R(\theta_t)$, $\sigma_R = \sigma\|g\|$, and the z-score collapses to the gradient projection of the perturbation,

$$Z_i \;=\; \frac{\sigma\|g\|\,u_i}{\sigma\|g\|} \;=\; u_i. \tag{29}$$

Both σ and $\|g\|$ cancel. The parallel coefficient in equation 27 is $\alpha\cdot\frac{1}{N}\sum_i u_i^2$ with $\sum_i u_i^2 \sim \chi^2_N$, so it has mean α and variance $2\alpha^2/N$. For the orthogonal component, $Z_i = u_i$ is independent of $\epsilon_i^{\perp}$ and $\mathbb{E}[u_i^2] = 1$, $\mathbb{E}\|\epsilon_i^{\perp}\|^2 = d-1$. Hence

$$\text{signal:} \qquad \mathbb{E}[\hat{g}^{\top}\Delta\theta] \;=\; \alpha, \tag{30}$$

$$\text{noise:} \qquad \sqrt{\mathbb{E}\|\Delta\theta_{\perp}\|^2} \;=\; \alpha\sqrt{\frac{d-1}{N}}. \tag{31}$$

In the absence of observation noise, the z-scored ES step therefore carries a signal of expected length exactly α along the gradient, independent of both the perturbation scale and the gradient magnitude, surrounded by isotropic noise of length $\alpha\sqrt{(d-1)/N}$. The expected cosine between the step and the gradient is approximately $\sqrt{N/d}$. This is the best case: none of the settings below improves on it.

Reward change. On a linear landscape the reward change of the centre is $\Delta R := R(\theta_{t+1}) - R(\theta_t) = g^{\top}\Delta\theta = \|g\|\,\hat{g}^{\top}\Delta\theta$, so it depends only on the parallel coefficient:

$$\mathbb{E}[\Delta R] \;=\; \alpha\|g\|, \qquad \mathrm{Var}[\Delta R] \;=\; \frac{2\alpha^2\|g\|^2}{N}. \tag{32}$$

Each step gains $\alpha\|g\|$ in expectation, independent of σ, d and N. The orthogonal noise moves the centre sideways but, on a linear landscape, sideways movement has no reward cost; N enters only through the variance, and the step improves the reward with probability $\approx \Phi(\sqrt{N/2})$.

B.3 LINEAR LANDSCAPE WITH OBSERVATION NOISE

We now let each observed reward be the reward of the perturbed model plus an additive observation noise,

$$R_i \;=\; R(\theta_t + \sigma\epsilon_i) \;+\; \sigma_\xi\,\xi_i, \qquad \xi_i \sim \mathcal{N}(0,1), \quad \xi_i \perp \epsilon_i, \tag{33}$$

with $\sigma_\xi > 0$ the observation-noise scale. Keeping the linear landscape, the centred reward is $R_i - \mu_R = \sigma\|g\|\,u_i + \sigma_\xi\xi_i$, and

$$\sigma_R^2 \;=\; \sigma^2\|g\|^2 + \sigma_\xi^2. \tag{34}$$

The z-score is now

$$Z_i \;=\; \frac{\sigma\|g\|\,u_i + \sigma_\xi\,\xi_i}{\sigma_R} \;=:\; w_i \sim \mathcal{N}(0,1), \qquad \mathrm{Corr}(w_i,u_i) \;=\; \frac{\sigma\|g\|}{\sigma_R} \;=:\; c \in [0,1], \tag{35}$$

a unit-variance Gaussian that is only partially correlated with the gradient projection. For the parallel coefficient, $\mathbb{E}[w_i u_i] = c$ and, for a correlated standard-normal pair, $\mathbb{E}[w_i^2 u_i^2] = 1 + 2c^2$, so the coefficient has mean αc and variance $\alpha^2(1+c^2)/N$. For the orthogonal component, w_i is independent of $\epsilon_i^\perp$ and $\mathbb{E}[w_i^2] = 1$, so its expected squared norm is unchanged from the noiseless case. Hence

$$\text{signal:} \qquad \mathbb{E}[\hat{g}^\top\Delta\theta] \;=\; \alpha c \;=\; \frac{\alpha\sigma\|g\|}{\sqrt{\sigma^2\|g\|^2+\sigma_\xi^2}}, \tag{36}$$

$$\text{noise:} \qquad \sqrt{\mathbb{E}\|\Delta\theta_\perp\|^2} \;=\; \alpha\sqrt{\frac{d-1}{N}}. \tag{37}$$

Observation noise attenuates the expected signal by the factor c and leaves the noise magnitude exactly unchanged. It also makes the realized signal less reliable: the standard deviation of the parallel coefficient is $\alpha\sqrt{(1+c^2)/N}$, so the projection of a single step onto the gradient is positive with high probability only when $c \gtrsim 1/\sqrt{N}$. Rewriting the attenuation as

$$c \;=\; \frac{\|g\|}{\sqrt{\|g\|^2 + \sigma_\xi^2/\sigma^2}} \tag{38}$$

shows that it is monotonically increasing in σ, approaching 1 as $\sigma\|g\|/\sigma_\xi \to \infty$ and recovering equation 30–equation 31. On an exactly linear landscape this trade has no cost: Gaussian smoothing of a linear function is exact and introduces no bias, so the perturbation scale should be made as large as possible. The only thing that stops it is the breakdown of linearity.

Reward change. As before $\Delta R = \|g\|\,\hat{g}^\top\Delta\theta$, now with the attenuated parallel coefficient:

$$\mathbb{E}[\Delta R] \;=\; \alpha c\,\|g\| \;=\; \frac{\alpha\sigma\|g\|^2}{\sqrt{\sigma^2\|g\|^2+\sigma_\xi^2}}, \qquad \mathrm{Var}[\Delta R] \;=\; \frac{\alpha^2\|g\|^2(1+c^2)}{N}. \tag{39}$$

Observation noise scales the expected gain by c and makes individual steps less reliable, with $\mathbb{P}[\Delta R > 0] \approx \Phi\big(c\sqrt{N/(1+c^2)}\big)$, but the orthogonal noise remains free.

B.4 QUADRATIC LANDSCAPE WITH OBSERVATION NOISE

To capture that breakdown, add curvature: $R(\theta) = R(\theta_t) + g^\top(\theta-\theta_t) - \frac{1}{2}(\theta-\theta_t)^\top Q\,(\theta-\theta_t)$, with Q a general symmetric matrix (neither diagonal nor isotropic is assumed). Writing $q_i = \epsilon_i^\top Q\epsilon_i - \mathrm{tr}\,Q$ for the centred curvature fluctuation, which has $\mathbb{E}[q_i] = 0$ and $\mathrm{Var}[q_i] = 2\,\mathrm{tr}[Q^2]$, the centred reward contains three pieces,

$$R_i - \mu_R \;=\; \sigma\|g\|\,u_i \;-\; \frac{\sigma^2}{2}\,q_i \;+\; \sigma_\xi\,\xi_i. \tag{40}$$

These are mutually uncorrelated: q_i is even in ϵ_i while u_i is odd, and ξ_i is independent of both. Consequently

$$\sigma_R^2 \;=\; \sigma^2\|g\|^2 \;+\; \frac{\sigma^4}{2}\,\mathrm{tr}[Q^2] \;+\; \sigma_\xi^2. \tag{41}$$

Signal. By equation 28,

$$\mathbb{E}[Z_i u_i] \;=\; \frac{1}{\sigma_R}\Big(\sigma\|g\|\,\mathbb{E}[u_i^2] \;-\; \frac{\sigma^2}{2}\,\mathbb{E}[q_i u_i] \;+\; \sigma_\xi\,\mathbb{E}[\xi_i u_i]\Big) \;=\; \frac{\sigma\|g\|}{\sigma_R}, \tag{42}$$

since $q_i u_i$ is odd in ϵ_i and ξ_i is independent of u_i. Hence the signal magnitude is

$$\text{signal:}\qquad \mathbb{E}[\hat{g}^\top \Delta\theta] \;=\; \frac{\alpha\sigma\|g\|}{\sigma_R} \;=\; \frac{\alpha\|g\|}{\sqrt{\|g\|^2 \;+\; \dfrac{\sigma^2}{2}\operatorname{tr}[Q^2] \;+\; \dfrac{\sigma_\xi^2}{\sigma^2}}}. \tag{43}$$

We write $\rho := \sigma\|g\|/\sigma_R$ for this attenuation factor in general; it equals 1 and c in the two linear cases. The curvature fluctuation contributes nothing to the expected projection but enters σ_R in exactly the same way as observation noise. From the point of view of the z-score it is an *endogenous* noise source, generated by the algorithm's own exploration, that scales as σ^4 while the gradient term scales as σ^2.

Noise. Because q_i depends on $\epsilon_i^\perp$, Z_i is no longer independent of $\epsilon_i^\perp$, and we evaluate $\mathbb{E}[Z_i^2\|\epsilon_i^\perp\|^2] = \mathbb{E}[Z_i^2\|\epsilon_i\|^2] - \mathbb{E}[Z_i^2 u_i^2]$. Using the Gaussian moments $\mathbb{E}[u_i^2\|\epsilon_i\|^2] = d+2$, $\mathbb{E}[q_i^2\|\epsilon_i\|^2] = 2(d+4)\operatorname{tr}[Q^2]$, $\mathbb{E}[u_i^4] = 3$, and $\mathbb{E}[q_i^2 u_i^2] = 2\operatorname{tr}[Q^2] + 8\,\hat{g}^\top Q^2 \hat{g}$, with all cross terms vanishing by parity,

$$\mathbb{E}\big[Z_i^2\|\epsilon_i\|^2\big] \;=\; d + \frac{2\sigma^2\|g\|^2 + 2\sigma^4\operatorname{tr}[Q^2]}{\sigma_R^2}, \tag{44}$$

$$\mathbb{E}\big[Z_i^2 u_i^2\big] \;=\; 1 + \frac{2\sigma^2\|g\|^2 + 2\sigma^4\,\hat{g}^\top Q^2 \hat{g}}{\sigma_R^2}. \tag{45}$$

Subtracting and writing $P_\perp = I_d - \hat{g}\hat{g}^\top$, so that $\operatorname{tr}[P_\perp Q^2 P_\perp] = \operatorname{tr}[Q^2] - \hat{g}^\top Q^2 \hat{g}$ is the curvature energy orthogonal to the gradient,

$$\text{noise:}\qquad \sqrt{\mathbb{E}\|\Delta\theta_\perp\|^2} \;=\; \alpha\sqrt{\frac{d-1}{N}}\,\sqrt{1 + \frac{2\sigma^4\operatorname{tr}[P_\perp Q^2 P_\perp]}{(d-1)\,\sigma_R^2}} \;\approx\; \alpha\sqrt{\frac{d-1}{N}}, \tag{46}$$

where the curvature correction is bounded, $2\sigma^4\operatorname{tr}[P_\perp Q^2 P_\perp]/\sigma_R^2 \le 4$, by equation 41. Curvature therefore adds at most an $O(1/d)$ relative amount to the noise magnitude; its main effect on the off-gradient component is to redistribute it toward the high-curvature eigendirections of Q, as in the anisotropic covariance term $2\alpha^2\sigma^4 Q^2/(N\sigma_R^2)$. The parallel coefficient has variance $\alpha^2\big(\mathbb{E}[Z_i^2 u_i^2] - \sigma^2\|g\|^2/\sigma_R^2\big)/N$, which reduces to $\alpha^2(1+c^2)/N$ when $Q = 0$.

Reward change. On a quadratic landscape $\Delta R = g^\top\Delta\theta - \frac{1}{2}\Delta\theta^\top Q\,\Delta\theta$, and the second term is where the orthogonal noise is finally charged. Since ΔR is quadratic in $\Delta\theta$, its expectation needs only the first two moments of the update, $\mathbb{E}[\Delta\theta^\top Q\Delta\theta] = \operatorname{tr}[Q\operatorname{Cov}(\Delta\theta)] + \mathbb{E}[\Delta\theta]^\top Q\,\mathbb{E}[\Delta\theta]$. The same moment calculation that gives equation 44 gives the full covariance,

$$\operatorname{Cov}[\Delta\theta] \;=\; \frac{\alpha^2}{N}\,I_d \;+\; \frac{\alpha^2}{N\sigma_R^2}\Big(\sigma^2\,gg^\top + 2\sigma^4 Q^2\Big), \tag{47}$$

and hence, with $\lambda_g := \hat{g}^\top Q\hat{g}$ the curvature along the gradient,

$$\mathbb{E}[\Delta R] \;=\; \underbrace{\alpha\rho\|g\|}_{\text{signal gain}} \;-\; \underbrace{\frac{\alpha^2}{2N}\operatorname{tr}Q}_{\text{diffusion tax}} \;-\; \underbrace{\frac{\alpha^2\rho^2}{2}\Big(1+\frac{1}{N}\Big)\lambda_g}_{\text{overshoot}} \;-\; \underbrace{\frac{\alpha^2\sigma^4}{N\sigma_R^2}\operatorname{tr}[Q^3]}_{\text{anisotropic}}. \tag{48}$$

The first two terms dominate. The diffusion tax is the orthogonal noise, of squared length $\alpha^2 d/N$, charged at the mean curvature $\operatorname{tr}Q/d$; it is the population mean gap $\frac{\sigma^2}{2}\operatorname{tr}Q$ with σ^2 replaced by α^2/N, so the centre pays each step the same curvature cost as a perturbation of scale $\alpha/\sqrt{N}$. The overshoot term is the usual gradient-descent one and matters only when $\alpha\rho\lambda_g$ is comparable to $\|g\|$; the last term is smaller still. Keeping signal and diffusion, a step improves the expected reward iff $\alpha < 2N\rho\|g\|/\operatorname{tr}Q$, and the one-step gain is maximized at $\alpha = N\rho\|g\|/\operatorname{tr}Q$, where it equals $N\rho^2\|g\|^2/(2\operatorname{tr}Q)$: population size buys improvement linearly, gradient correlation quadratically. Because the signal gain shrinks with $\|g\|$ along a run while the diffusion tax does not, progress stalls when $\rho\|g\| = \alpha\operatorname{tr}Q/(2N)$.

Table 1: Per-update signal $\mathbb{E}[\hat{g}^\top \Delta\theta]$, noise $\sqrt{\mathbb{E}\|\Delta\theta_\perp\|^2}$, and expected reward change $\mathbb{E}[\Delta R]$ under the three landscape models, from equation 30–equation 32, equation 36–equation 39, and equation 43–equation 48. Throughout, $\rho = \sigma\|g\|/\sigma_R$ is the gradient correlation of the population rewards, equal to 1 and c in the two linear cases; $\lambda_g = \hat{g}^\top Q \hat{g}$; and $\delta_Q = 2\sigma^4 \operatorname{tr}[P_\perp Q^2 P_\perp]/((d-1)\sigma_R^2) \le 4/(d-1)$. The $O(1/N)$ and $\operatorname{tr}[Q^3]$ terms of equation 48 are omitted.

Landscape	**Signal**	**Noise**	**Expected reward change**
Linear, exact rewards	α	$\alpha\sqrt{(d-1)/N}$	$\alpha\|g\|$
Linear, observation noise	$\alpha\rho$	$\alpha\sqrt{(d-1)/N}$	$\alpha\rho\|g\|$
Quadratic, observation noise	$\alpha\rho$	$\alpha\sqrt{(d-1)/N}\,(1+\delta_Q)^{1/2}$	$\alpha\rho\|g\| - \frac{\alpha^2}{2N}\operatorname{tr} Q - \frac{\alpha^2\rho^2}{2}\lambda_g$

Summary of the three cases. Table 1 collects the signal and noise magnitudes and the expected reward change for the three cases. In every case the noise magnitude is $\alpha\sqrt{(d-1)/N}$, exactly in the linear cases and up to a bounded $O(1/d)$ relative correction in the quadratic case, and is insensitive to σ. The signal magnitude is α times a dimensionless attenuation factor $\sigma\|g\|/\sigma_R \in [0,1]$: equal to 1 without observation noise, reduced by observation noise, and reduced further by the curvature contribution to σ_R on a quadratic landscape. Consistent with equation 23, the choice of σ does not change how far ES moves per iteration, only what fraction of that movement is directed, and the perturbation-scale problem reduces to maximizing the attenuation factor in equation 43. The reward-change column adds the one place where the noise is not free: on a curved landscape the isotropic diffusion is taxed at the mean curvature, at a rate that falls as $1/N$.

B.5 Optimal perturbation scale

The denominator of equation 43 makes the competition explicit: the observation-noise term σ_ξ^2/σ^2 favours large σ, the curvature term $\frac{\sigma^2}{2}\operatorname{tr}[Q^2]$ favours small σ, and $\|g\|^2$ is the irreducible part. Neither term is small in general. Although the per-direction curvature contribution $2\sigma^4 u^\top Q^2 u$ is typically tiny for any single unit vector u, the quantity that enters σ_R is the full trace $\operatorname{tr}[Q^2] = \sum_{k=1}^d \lambda_k^2$, summed over all d eigendirections, and is extensive in the parameter dimension.

Holding α, g, Q and σ_ξ fixed, the signal equation 43 is maximized by minimizing $f(\sigma^2) = \sigma_\xi^2/\sigma^2 + \frac{\sigma^2}{2}\operatorname{tr}[Q^2]$ over σ^2, which gives

$$\boxed{\sigma_*^4 \;=\; \frac{2\,\sigma_\xi^2}{\operatorname{tr}[Q^2]}, \qquad \sigma_* \;=\; \left(\frac{2\,\sigma_\xi^2}{\operatorname{tr}[Q^2]}\right)^{1/4}} \tag{49}$$

At σ_* the curvature-induced contribution to σ_R^2 equals the exogenous one, $\frac{\sigma_*^4}{2}\operatorname{tr}[Q^2] = \sigma_\xi^2$, so the population reward variance is

$$\sigma_R^2\big|_{\sigma_*} \;=\; \sigma_*^2\|g\|^2 + 2\sigma_\xi^2, \tag{50}$$

and the largest achievable signal is

$$\max_\sigma \mathbb{E}[\hat{g}^\top \Delta\theta] \;=\; \frac{\alpha\sigma_*\|g\|}{\sqrt{\sigma_*^2\|g\|^2 + 2\sigma_\xi^2}} \;=\; \frac{\alpha\|g\|}{\sqrt{\|g\|^2 + \sigma_\xi\sqrt{2\operatorname{tr}[Q^2]}}}. \tag{51}$$

Below σ_* the update is observation-noise limited; above it, curvature limited.

What the optimal population looks like. Two further statistics characterize the population at σ_*. First, its rewards are correlated with the gradient at the centre point by exactly the attenuation factor: since $\operatorname{Cov}(u_i, R_i) = \sigma\|g\|$ and $\operatorname{Var}[u_i] = 1$,

$$\rho_* \;:=\; \operatorname{Corr}\left(\hat{g}^\top \epsilon_i,\, R_i\right)\Big|_{\sigma_*} \;=\; \frac{\sigma_*\|g\|}{\sqrt{\sigma_*^2\|g\|^2 + 2\sigma_\xi^2}} \;=\; \frac{\|g\|}{\sqrt{\|g\|^2 + \sigma_\xi\sqrt{2\operatorname{tr}[Q^2]}}}, \tag{52}$$

so that the expected signal equation 51 is $\alpha\rho_*$, and the fraction of population reward variance attributable to the gradient is ρ_*^2, with the remainder split equally between observation noise and curvature. The second form makes the curvature dependence explicit: at the optimum, curvature never contributes more than σ_ξ^2 to the population variance, but a larger $\mathrm{tr}[Q^2]$ forces a smaller σ_* to hold it there, and the gradient variance $\sigma_*^2\|g\|^2$ shrinks with it while the non-gradient variance stays at $2\sigma_\xi^2$. Second, the population does not sit at the centre reward. Only the curvature term has nonzero mean, $\mathbb{E}[\epsilon_i^\top Q\epsilon_i] = \mathrm{tr}\, Q$, so

$$\mu_R - R(\theta_t) \;=\; -\frac{\sigma^2}{2}\,\mathrm{tr}\, Q, \qquad \mu_R - R(\theta_t)\Big|_{\sigma_*} \;=\; -\,\sigma_\xi\sqrt{\frac{r_{\mathrm{eff}}}{2}}, \qquad r_{\mathrm{eff}} := \frac{(\mathrm{tr}\, Q)^2}{\mathrm{tr}[Q^2]}, \tag{53}$$

where $r_{\mathrm{eff}} \in [1, d]$ is the participation ratio (effective rank) of the Hessian, equal to d for isotropic curvature and to r for a rank-r Hessian with equal eigenvalues. In units of the population spread equation 50,

$$\frac{R(\theta_t) - \mu_R}{\sigma_R}\Big|_{\sigma_*} \;=\; \sqrt{\frac{r_{\mathrm{eff}}}{2}}\cdot\frac{\sigma_\xi}{\sqrt{\sigma_*^2\|g\|^2 + 2\sigma_\xi^2}} \;=\; \frac{\sqrt{r_{\mathrm{eff}}}}{2}\,\sqrt{1-\rho_*^2}. \tag{54}$$

Unless $r_{\mathrm{eff}} \lesssim 4$, the optimally scaled population therefore lies several standard deviations below the centre point: nearly every perturbed model is worse than the unperturbed one. This is not a defect but the price of exploring at scale σ_* on a curved landscape, and it is the same for every member; the update depends only on the spread σ_R and the gradient share ρ_*^2, not on where the population sits. The gap in equation 54 grows as the gradient share falls, so along a training run it is expected to be small while $\|g\|$ is large and to become pronounced as the centre approaches a ridge. Because equation 53 is independent of the observation-noise scale while σ_R is not, the gap in σ_R units is less pronounced whenever σ_ξ is larger, for instance when fewer prompts are averaged into each R_i. Away from σ_* the same quantities read $\rho = \sigma\|g\|/\sigma_R$ and $(R(\theta_t) - \mu_R)/\sigma_R = \sqrt{r_{\mathrm{eff}}\varphi/2}$ with $\varphi = \frac{\sigma^4}{2}\,\mathrm{tr}[Q^2]/\sigma_R^2$ the curvature share of the population variance, so a measured gap together with an estimate of φ yields r_{eff} directly.

Three remarks on equation 49. First, the dependence on the noise level is weak: σ_* grows only as $\sigma_\xi^{1/2}$, so even a large increase in observation noise justifies only a modest increase in the perturbation scale. Second, the dependence on curvature is through $\mathrm{tr}[Q^2]$, which grows with d; for a billion-parameter model even small typical curvatures accumulate to a large trace, which is the familiar dimension penalty of finite-difference estimators appearing inside the z-score normalization. Third, the optimal signal equation 51 degrades with the geometric mean $\sigma_\xi\sqrt{\mathrm{tr}[Q^2]}$ rather than with either quantity alone: noise and curvature are not separately harmful but jointly so, since one can always be traded against the other by moving σ.

C Derivation of the Update-Norm Moments

This appendix derives equation 23–equation 24. In general the weights depend on the perturbations: Z_i is a function of the reward at $\theta_t + \sigma\epsilon_i$, and hence of ϵ_i itself. What can be said without a landscape model is that each Z_i depends only on its own perturbation (and its own observation noise), so the products $Z_i\epsilon_i$ are identically distributed across i and, under the population-statistics simplification, independent. Write

$$m \;:=\; \mathbb{E}[Z_i\epsilon_i] \in \mathbb{R}^d, \qquad \kappa \;:=\; \mathbb{E}\big[Z_i^2\|\epsilon_i\|^2\big] - d \tag{55}$$

for the two landscape-dependent statistics of a single weighted perturbation. The first is the mean weighted perturbation, which is the direction the update drifts in; the second measures how much the weighting inflates the squared length of a perturbation beyond its unweighted value d. Expanding $\|\Delta\theta\|^2$ into diagonal and cross terms and taking expectations,

$$\begin{aligned}\mathbb{E}\|\Delta\theta\|^2 \;&=\; \frac{\alpha^2}{N^2}\sum_{i=1}^{N}\mathbb{E}\big[Z_i^2\|\epsilon_i\|^2\big] \;+\; \frac{\alpha^2}{N^2}\sum_{i\neq j}\mathbb{E}[Z_i\epsilon_i]^\top\mathbb{E}[Z_j\epsilon_j] \\ &=\; \frac{\alpha^2 d}{N} \;+\; \frac{\alpha^2\kappa}{N} \;+\; \alpha^2\Big(1-\frac{1}{N}\Big)\|m\|^2.\end{aligned} \tag{56}$$

This identity is exact. The drift term is bounded independently of the landscape: for any unit vector v, Cauchy–Schwarz gives $(v^\top m)^2 = \mathbb{E}[Z_i\, v^\top \epsilon_i]^2 \le \mathbb{E}[Z_i^2]\,\mathbb{E}[(v^\top \epsilon_i)^2] = 1$, so $\|m\| \le 1$ and the drift contributes at most α^2. The inflation κ does depend on the landscape; the derivations below evaluate it for linear and quadratic rewards and find $0 \le \kappa \le 4$. Both corrections are therefore smaller than the leading term $\alpha^2 d/N$ by a factor of order N/d:

$$\mathbb{E}\|\Delta\theta\|^2 \;=\; \frac{\alpha^2 d}{N}\Big(1 + O\big(N/d\big)\Big). \tag{57}$$

For the variance, expand $\|\Delta\theta\|^2$ the same way. The cross terms $Z_i Z_j\, \epsilon_i^\top \epsilon_j$ for $i \neq j$ are zero-mean up to the drift, and each has variance $\mathbb{E}[Z_i^2 Z_j^2 (\epsilon_i^\top \epsilon_j)^2] = d\,(1+O(1/d))$; there are $N(N-1)$ of them and they are pairwise uncorrelated up to $O(\|m\|^2)$ terms. The diagonal terms $Z_i^2\|\epsilon_i\|^2$ fluctuate with variance $O(d)$ each, but there are only N of them and, under the exact constraint equation 22, their $O(d)$ fluctuations cancel against those of the cross terms. The result is

$$\operatorname{Var}\|\Delta\theta\|^2 \;=\; \frac{2\alpha^4 d}{N^2}\Big(1 + O\big(N/d\big)\Big), \qquad \frac{\sqrt{\operatorname{Var}\|\Delta\theta\|^2}}{\mathbb{E}\|\Delta\theta\|^2} \;=\; \sqrt{\frac{2}{d}}\Big(1 + O\big(N/d\big)\Big), \tag{58}$$

the same leading mean and variance as $\frac{\alpha^2}{N}\chi^2_d$, the law that equation 18 would follow if the weights were independent of the perturbations. The coefficient of variation $\sqrt{2/d}$ is what makes $\sqrt{\mathbb{E}\|\Delta\theta\|^2}$ an accurate proxy for the typical length: the two agree to relative order $1/d$.[1]

[1]The cancellation between diagonal and cross-term fluctuations requires the exact sample normalization $\sum_i Z_i^2 = N$. Under the population-statistics simplification the weights are only normalized in expectation, $\sum_i Z_i^2 = N + O(\sqrt{N})$, and this fluctuation alone contributes $2\alpha^4 d^2/N^3$ to the variance, giving a coefficient of variation of order $1/\sqrt{N}$ rather than $1/\sqrt{d}$. The algorithm uses sample statistics, so equation 24 is the relevant statement; the population-statistics simplification is used below only for expectations, where it is harmless at the orders reported.