Investigating the “Mpempba Effect” in Deep Learning and Imaging

LazyingArt

1 Background and Motivation

The term “mpempba effect” does not correspond to any well-established concept in the literature. It is likely referring to the Mpemba effect, a phenomenon from thermodynamics. The Mpemba effect describes the counterintuitive situation where a system that starts *hotter* can cool to a target temperature faster than an identical system that starts *colder*, under the same conditions scitechdaily.com. First observed by Aristotle and later documented by Erasto Mpemba in water freezing experiments, this effect has been confirmed in various physical systems beyond water scitechdaily.com. In essence, starting from a more extreme initial state can paradoxically speed up relaxation to equilibrium.

In recent years, there is growing interest in drawing analogies between the Mpemba effect and optimization in deep learning. Large-scale neural network training often employs staged learning-rate schedules that warm up (increase to a high learning rate), stay high for a while, then cool down (decay) arxiv.org emergentmind.com. This Warmup–Stable–Decay (WSD) strategy has been empirically successful in domains like vision and language models emergentmind.com emergentmind.com. Researchers have begun to connect this practice to a Mpemba-like effect: the “preheated” high-learning-rate phase sets the stage for a faster convergence during the later low-learning-rate phase arxiv.org emergentmind.com. In other words, a network trained with a higher “temperature” (learning rate) during a plateau can descend faster once cooling begins, analogous to the hotter system cooling faster in the Mpemba effect. Recent work by Liu et al. (2025) explicitly draws this parallel, showing that an optimal high plateau (a *strong Mpemba point*) can eliminate the slowest training modes and yield accelerated loss decrease during decay arxiv.org. This provides a theoretical justification for why a high constant learning rate phase often outperforms a prematurely decayed or lower learning rate schedule emergentmind.com.

Outside of learning-rate schedules, the spirit of the Mpemba effect echoes through other practices. Simulated annealing in optimization, for example, uses a high initial “temperature” (allowing random exploration) before cooling, to avoid local minima and achieve better solutions. Similarly, the phenomenon of super-convergence in neural networks demonstrates that extremely large learning rates (with appropriate scheduling) can drastically speed up training arxiv.org. These approaches align with the idea that an initially more volatile or high-energy state can ultimately lead to faster or better convergence than a conservative start.

However, it is important to note that the Mpemba effect analogy in deep learning is an emerging concept, not a firmly established law. The term “mpempba effect” itself does not appear in academic databases, so it likely refers to this nascent analogy or a misidentification of Mpemba. So far, only a few theoretical papers and discussions (e.g., on large-language model training) have explicitly used the Mpemba terminology arxiv.org emergentmind.com. There is no known “Mpempba effect” specific to imaging in the literature under that name. This suggests an opportunity to explore whether such counterintuitive behavior might manifest in imaging or deep learning contexts under different guises. For example, one might ask: Could an image reconstruction algorithm that starts with a very rough (noisy) initial guess reach a clear image faster than one that starts closer to the final image? Could a neural network trained with an initially unstable high noise or augmentation regime end up more robust than one trained slowly from the start? These questions remain open, as no direct analog of the Mpemba effect has been documented in imaging pipelines yet.

Significance: If a Mpemba-like effect does exist in deep learning or computational imaging, uncovering it could challenge our intuition about training dynamics and signal priors. It might lead to new optimization strategies (e.g., tailored learning rate or noise schedules) that achieve better results faster. It could also improve our understanding of model robustness – for instance, whether a "heated" training phase confers any resilience or simply speeds up fitting. Given the above, we propose to formally investigate this concept in the context of deep learning optimization and an imaging task.

2 Hypothesis and Objectives

Hypothesis: We hypothesize that an analog of the Mpemba effect can occur in deep learning and imaging optimization. In particular, an initially "hot" or high-energy training condition (such as a high learning rate, strong regularization, or high noise regime) will allow a model to converge faster or to a better optimum once standard training conditions are applied, compared to a model that begins training under more conservative (cooler) conditions. In other words, a neural network or imaging algorithm that is briefly subjected to a more extreme initial phase will eventually outperform (in speed of convergence or final error) an identical process that did not receive this preheating. We expect to observe a crossover in performance: initially, the conservatively trained model might have an advantage (e.g., lower loss early on), but the aggressively trained model will catch up and surpass it, achieving the target accuracy or reconstruction quality in less time – mirroring the Mpemba effect.

To break this down, the study will address the following objectives:

- **O1.** Determine if a warmup-high-decay training schedule in deep neural networks indeed yields faster convergence during the decay phase than a low-and-slow schedule, consistent with the Mpemba effect analogy emergentmind.com. We will identify if there is an optimal "high plateau" value that maximizes this speed-up (analogous to a strong Mpemba point arxiv.org).

- **O2.** Investigate whether a similar phenomenon can be observed in an imaging context (such as image denoising or deblurring). For example, test if starting an iterative reconstruction with a very noisy initial image or large initial update steps can produce a final clear image faster than starting from a mildly noisy initial state.

- **O3.** Understand the trade-offs involved. We will examine not only convergence speed but also final performance and stability. A key question is whether the Mpemba-like faster convergence might come at the cost of converging to a different kind of solution (e.g., a sharper minimum with worse generalization arxiv.org). Ensuring model robustness and accuracy in the end results is as important as speed.

By fulfilling these objectives, the study aims to clarify whether the "mpempba effect" is a real and useful concept in machine learning, or simply a coincidental metaphor. If the hypothesis is supported, it could inform better training protocols and inspire new research into non-monotonic optimization behaviors (statistical anomalies where going to an extreme early yields a better outcome later).

3 Methodology (Proposed Experiments)

To test the above hypothesis, we propose a two-part experimental approach: one in a deep learning training scenario and one in an imaging restoration scenario. Both parts will compare "hot-start" versus "cold-start" conditions, while keeping all else equal, to see if the hot-start can indeed pull ahead in later stages.

1. Deep Learning Optimization Experiment

We will use a standard image classification task (e.g., CIFAR-10 with a ResNet) as a representative deep learning problem. Two training regimens will be compared:

- **Regimen A (Conservative baseline):** A conventional learning rate schedule without a prolonged high plateau. For instance, a small warm-up to a moderate learning rate, or even a continuously decaying learning rate from the start. This simulates a "colder" start (lower effective temperature throughout training).

- **Regimen B (Mpemba-inspired):** A Warmup–Stable–Decay schedule emergentmind.com with a high plateau learning rate held for an extended duration, before decaying. After a brief warm-up (to avoid divergence), the learning rate will be kept at a substantially higher value than Regimen A for a significant portion of training, then decayed to the same final value as A. This is the "hotter" start condition.

Both regimens will use the same model architecture, dataset, number of epochs, and final learning rate value to ensure a fair comparison. The key difference is the presence of a high-LR plateau in Regimen B versus an earlier or lower LR decay in Regimen A. We will run multiple trials of each to account for variability. Throughout training, we will log metrics such as training loss, validation loss, and accuracy.

Key observations: We will look for a crossover point in the learning curves of A vs B. Initially, Regimen A (with lower LR) may have smoother or faster initial loss reduction (since B's high LR might cause more oscillation early on). But our hypothesis predicts that once we enter the decay phase, Regimen B's loss will drop faster and potentially fall below Regimen A's loss, reaching low error levels sooner emergentmind.com. We will identify if/when such curve crossing happens (analogous to the temperature curves crossing in the classical Mpemba effect). We will also experiment with different plateau heights in Regimen B to find if an optimal value exists that maximizes the speed-up, as theory suggests a "strong Mpemba point" where convergence acceleration is highest arxiv.org.

2. Imaging Restoration Experiment

To generalize the inquiry, we will examine a signal/image reconstruction task where an algorithm iteratively improves an image. A suitable testbed is the Deep Image Prior (DIP) denoising method or a classical iterative deblurring algorithm. The idea is to see if an initially more extreme distortion or update leads to faster eventual cleanup of the image:

- We take a degraded input image (e.g., a noisy or blurry image) that we want to restore. In **Method A**, we initialize the process in a typical way (e.g., start the DIP with the actual noisy image as input, or start an iterative solver from the noisy image). In **Method B**, we intentionally make the starting point more extreme – for example, add an extra burst of noise or perturbation (a "hot start") before letting the algorithm proceed, or use a very large initial gradient step in the first few iterations of the solver. After this initial difference, both Method A and B follow the same algorithm and parameters.

- Concretely, for DIP denoising, Method B might begin by injecting additional noise for a short initial period or using a higher learning rate for the network optimization in the first few hundred iterations, then reverting to the normal learning rate. For a traditional deblurring iterative algorithm, Method B could start with a deliberately over-sharpened or over-smoothed guess and a large first update step, whereas Method A starts from the raw blurry image with standard step sizes.

We will measure the quality of the image reconstruction over time (iterations). Quality can be quantified by metrics like Peak Signal-to-Noise Ratio (PSNR) or Structural Similarity (SSIM) against the ground truth image. The focus is on whether the Method B (hot-start) curve overtakes Method A in terms of reconstruction quality at a certain point, achieving a given quality level in fewer iterations.

Controls: In both experiments, careful controls will be in place. We will ensure that the total training time or iterations is the same across conditions so that any speed-up is not due to simply running longer. We also need to ensure stability: the hot-start conditions (high LR or heavy noise) will be chosen such that the process does not diverge or irreversibly damage the model. Pilot runs will determine safe yet aggressive "hot" settings (e.g., how high the learning rate plateau can go without blowing up the loss). If necessary, we will incorporate a brief warm-up for the extremely hot starts as done in practice emergentmind.com.

4 Expected Results and Implications

If our hypothesis is correct, we expect to see clear evidence of a Mpemba-like advantage in the results:

- **Faster Convergence After Heating:** In the neural network training experiment, Regimen B should converge to low loss noticeably faster once the decay phase kicks in, compared to Regimen A. For example, B might reach, say, 90% accuracy in half the number of epochs that A requires. We anticipate observing the Mpemba crossover – initially, B might lag (higher loss due to high LR instability), but eventually it will not only catch up to A but converge to a lower loss before A does emergentmind.com. This would mirror the physical Mpemba effect where the hotter sample overtakes the cooler one in cooling rate. An optimal plateau LR might be identified where this effect is strongest, aligning with theoretical predictions of a "strong Mpemba point" arxiv.org.

- **Image Reconstruction Improvement:** In the imaging task, we expect the method with the more aggressive start to achieve a given quality threshold faster. For instance, Method B might remove a certain amount of noise or blur in fewer iterations than Method A, despite starting in a worse initial state. A successful outcome would be that by the end, both methods can reach similar final quality, but the path B takes is non-monotonic – it might initially produce a more distorted image (due to the extra noise or large step) but then improves rapidly and surpasses Method A's performance at some point. This would be a novel demonstration of a Mpemba-like behavior in an imaging algorithm.

- **The Role of Signal Priors:** We might find that the effect is more pronounced when the problem has a multi-scale or multi-modal landscape. For example, the DIP network might initially explore a wider set of image structures when driven with high learning rate or noise (capturing coarse features quickly) before focusing on details, thus speeding up overall denoising. This would tie into the idea of leveraging a signal prior: the structure of the network or algorithm favors natural images, so a strong initial jolt might quickly align the solution with the natural image manifold (skipping some shallow local minima that a slow approach might get stuck in).

- **Generalization and Quality Trade-offs:** We will carefully check if the hot-start accelerated convergence comes with any downsides. One concern is that reaching lower training loss faster could mean finding a sharper minimum (lower loss but more curvature) which might harm generalization arxiv.org. In the classification task, we will evaluate test-set accuracy: it is possible that Regimen B, while faster, could overfit or ultimately yield similar accuracy to A. If we see that both regimens reach similar final accuracy, but B does so faster, that is a pure win. If B reaches lower training loss but worse test accuracy than A, it suggests the Mpemba-like strategy might be driving the model to overfit or to a less generalizable solution – an important insight for practice. In imaging, a similar check is needed: does the hot-start method produce any artifacts or overshoot the optimal solution (e.g., over-sharpening the image)? Ideally, we expect no loss in final quality; at best, maybe an improvement if the aggressive start helps avoid bad local minima.

Implications: A confirmed Mpemba effect in deep learning would provide a principled explanation for why carefully designed high-learning-rate phases improve training efficiency arxiv.org. It could guide practitioners in choosing learning rate schedules more scientifically, rather than by trial-and-error. For instance, one could aim for that optimal "preheat" intensity and duration to minimize training time without sacrificing performance arxiv.org. In imaging and other domains, it might open new ways to speed up algorithms: e.g., deliberately "shake up" a system at the start to let it settle faster. This challenges the conventional wisdom that one should always start training gently – instead, sometimes **starting hotter gets you there faster**.

On the other hand, if the expected effect does not materialize (or is very weak), that outcome is also valuable. It would imply that the Mpemba analogy might be limited to certain conditions (perhaps the highly structured loss landscapes of large language models, or only under specific types of nonlinearity). It would caution against over-applying the thermodynamic metaphor. Either way, the study will advance understanding of optimization behavior and could identify new statistical anomalies in training dynamics that merit further theoretical analysis.

5 Evaluation Strategy

We will rigorously evaluate the experimental outcomes to test the hypothesis. Our evaluation plan includes:

- **Convergence Metrics:** Measure the number of training iterations (or wall-clock time) needed to reach specific performance milestones (e.g., 80%, 90% accuracy for classification, or a PSNR of X for image denoising). A direct comparison will show if the hot-start method achieves milestones faster. We will plot and compare learning curves for each regimen/method. A clear Mpemba effect would be visualized by the curves crossing: the initially slower (hot-start) curve dropping below the initially faster (cold-start) curve at some point.

- **Final Performance:** Compare the final outcomes after a fixed training budget. We will report final validation accuracy (and loss) for the deep learning models, and final image quality metrics for the imaging task. This checks if the hot-start approach matches or exceeds the baseline in eventual performance, or if it trades off quality for speed.
- **Statistical Significance:** All experiments will be repeated (e.g., 5 runs with different random seeds for network initialization and data shuffling) to account for variability. We will use statistical tests (t-tests or Wilcoxon signed-rank on paired runs) to assess whether differences in convergence speed or final metrics are significant and not due to chance. For example, we will test if the average epoch at which 90% accuracy is reached is significantly lower for Regimen B than for Regimen A.
- **Trajectory Analysis:** Beyond simple metrics, we will analyze the behavior of the training dynamics:
  - In the deep learning experiment, we can track parameters like the gradient norms or the model's weight distribution during training. We might also periodically estimate the curvature (e.g., largest Hessian eigenvalue) to see if the hot-start phase indeed increases curvature initially and then allows it to drop – a sign of sharper valleys being navigated emergentmind.com. This connects to the theory that high "temperature" lets the optimizer traverse flat directions (rivers) more effectively emergentmind.com.
  - In the imaging experiment, we can visually inspect intermediate outputs. Does the hot-start method initially produce a very grainy or distorted intermediate (due to the extra noise or large step), which then rapidly clarifies? Such a pattern would qualitatively support the hypothesis. We will also examine if any artifacts remain in the final output of the hot-start method.
- **Robustness Checks:** To evaluate model robustness (one of the contexts of interest), we can test the trained models on slightly perturbed data or adversarial examples to see if the training regimen had any effect (positive or negative) on robustness. It is possible that a high LR plateau could act as a regularizer arxiv.org, potentially improving generalization or robustness to noise, but it could also bias towards sharper minima arxiv.org. Similarly, for imaging, we can test the methods on multiple images or different noise levels to ensure the phenomenon is consistent and not a one-off.
- **Failure Criteria:** We will consider the possibility that the Mpemba-like effect might not appear under certain conditions. If, for instance, the tasks are too simple (convex problems) or the "temperature" difference is not enough, we might see no significant difference. In that case, we would iteratively adjust the experimental setup – e.g., try a more complex model or task, or increase the disparity between cold and hot starts – until we either observe the effect or reach a conclusion that it likely does not manifest in that domain. Documenting these negative or threshold findings is important to delineate the boundaries of the effect.

Throughout the evaluation, citation of prior work and theory will guide interpretation. For example, if we observe faster optimization but poorer generalization in Regimen B, we will relate it to the known tendency of high learning rates to find sharp minima arxiv.org and discuss possible regularization remedies. If we observe a clear speed-up, we will compare the measured speed-up to theoretical predictions (e.g., does it match the exponential speed-ups seen in some thermodynamic models arxiv.org?). All findings will be contextualized with respect to the initial physics analogy and contemporary research.

By the end of the evaluation, we expect to conclude one of two things: either (a) we have demonstrated a Mpemba-like acceleration effect in the chosen deep learning and imaging scenarios (thus coining a valid new insight for those fields), or (b) the effect was not observed, suggesting that what works in thermodynamics might require very specific conditions to appear in algorithmic processes. In either case, the results will be illuminating. The study will either provide evidence to support the Mpemba effect as a useful concept in deep learning/imaging, or it will clarify misconceptions and guide researchers to where they should (or should not) look for such counterintuitive behaviors in complex models.

References

1. Mpemba effect definition and history – SciTechDaily (2025) scitechdaily.com.

2. Mpemba effect analogy in LLM training – Liu et al., *arXiv* (2025) arxiv.org arxiv.org.

3. WSD scheduling and Mpemba effect in practice – *Emergent Mind* summary (2025) emergentmind.com.

4. Super-convergence via high learning rates – Smith & Topin (2017) arxiv.org.

5. Mpemba effect and generalization trade-off – Liu et al., *arXiv* (2025) arxiv.org.