One-Shot Hyper-Gradient Warm-Starts for Bandit-Style Hyperparameter Optimisation

Anonymous Author(s)
Affiliation
Address
email

Abstract

Bandit-style multi-fidelity schedulers such as ASHA and PASHA are the workhorses of practical hyperparameter optimisation, yet they still waste substantial compute on configurations that could have been flagged as poor before real training even begins. The root cause is that every trial is treated as a black box: none of the gradients already computed inside the training loop are exploited by the scheduler. We close this gap with One-Shot Hyper-Gradient Warm-Starts (OHGW). For each freshly sampled configuration we run exactly one mini-batch, obtain stochastic hyper-gradients $\partial L/\partial\psi$ for all continuous hyperparameters at almost zero extra cost via automatic differentiation, apply a single tiny update $\psi \leftarrow \psi - \eta_h\, \partial L/\partial\psi$, and hand the nudged configuration back to the unmodified scheduler. OHGW therefore preserves exploration while biasing every candidate toward lower-loss regions at negligible overhead and with no change to promotion or stopping logic. On CIFAR-10 with ResNet-20 under ASHA and on WikiText-103 with GPT2-small under PASHA, OHGW cuts median wall-clock time to a preset quality threshold by roughly twenty percent, adds under four percent floating-point operations, and leaves final accuracy and perplexity unchanged. Random perturbations provide almost no benefit and taking more than one hyper-step shows diminishing returns. These findings demonstrate that a single noisy hyper-gradient obtained before expensive training commences can reclaim a significant share of wasted computation in grey-box hyperparameter optimisation.

1 Introduction

Hyperparameter optimisation (HPO) is indispensable for obtaining robust performance in modern machine-learning systems, yet even the most popular grey-box schedulers squander a sizable fraction of their budget on clearly sub-optimal configurations. Successive-Halving variants such as Hyperband, ASHA and PASHA prune weak contenders early by evaluating them on progressively larger budgets Bohdal et al. [2022]. Grey-box Bayesian schemes like DyHPO refine this idea through learning-curve modelling and dynamic promotion rules Wistuba et al. [2022]. Despite these advances, almost all schedulers regard the training process itself as opaque: internal gradients that are already computed for parameter updates are ignored during the search.

Hyper-gradient methods have shown that gradients with respect to hyperparameters can be extracted cheaply via automatic differentiation Chandra et al. [2019] or implicit differentiation techniques that avoid expensive unrolling Bertrand et al. [2020]. Unfortunately these approaches typically assume full control over the optimisation routine and therefore clash with production HPO systems whose scheduling logic is complex and battle-tested. The open question, then, is how to inject very cheap but noisy hyper-gradient information into existing bandit-style frameworks without having to rewrite their core.

We address this question with One-Shot Hyper-Gradient Warm-Starts (OHGW). Whenever the scheduler samples a configuration $x = (\theta_0, \psi)$ consisting of model parameters θ (usually random initialisation) and continuous hyperparameters ψ, the training script performs exactly one forward-and-backward pass on a single mini-batch, collects the stochastic hyper-gradient $g_\psi = \partial L / \partial \psi$, and applies a microscopic update $\psi \leftarrow \psi - \eta_h g_\psi$ with $\eta_h = 10^{-3}$. Promotion rules, budgets and stopping criteria remain untouched; from the scheduler's perspective nothing has changed except that the candidate starts from a slightly more promising point.

Two practical challenges arise. First, a gradient measured on a single mini-batch is extremely noisy, so the step must be sufficiently small to prevent biasing the search or harming exploration. Second, adoption hinges on a minimal engineering footprint—ideally a few lines of code that do not depend on the internals of the scheduler. OHGW meets both constraints: the extra cost is one forward and one backward pass per trial ($< 4\%$ FLOPs in our experiments) and integration is a five-line wrapper around trial creation.

We validate OHGW in two contrasting settings—vision (CIFAR-10, ResNet-20, ASHA) and language modelling (WikiText-103, GPT2-small, PASHA)—using 56 paired random seeds and equal GPU budgets. Metrics include time-to-target quality, best final score, compute overhead, variance, and hyperparameter distribution shift. OHGW consistently shortens time-to-target by about twenty percent while preserving ultimate performance and introducing negligible bias. Ablations confirm that gradient directionality, not random perturbation, drives the gain, and that repeating the warm-start step gives only marginal additional savings.

1.1 Contributions

- **Scheduler-agnostic warm-start:** We introduce OHGW, a single-step hyper-gradient warm-start that improves efficiency without altering bandit logic.
- **Cheap gradient extraction:** We provide a practical recipe for extracting hyper-gradients of continuous hyperparameters at negligible cost.
- **Empirical efficiency gains:** Extensive experiments across vision and language reduce median wall-clock time to target quality by roughly twenty percent with under four percent compute overhead.
- **Robustness and ablations:** Gradient direction matters, benefits saturate quickly, and variance or bias are not inflated.

Looking forward, we plan to extend OHGW to mixed discrete-continuous spaces, integrate warm-start signals into surrogate-based selection Khazi et al. [2023] and adaptive-fidelity frameworks Jiang and Mian [2024], and explore privacy-aware or federated scenarios where one-shot, low-overhead interventions are especially attractive Panda et al. [2022], Khodak et al. [2021].

2 Related Work

Multi-fidelity schedulers: Successive-Halving, Hyperband and ASHA progressively allocate resources; PASHA adds an adaptive cap on maximum fidelity Bohdal et al. [2022]. DyHPO supervises the race among configurations with a deep-kernel Gaussian Process that embeds learning-curve dynamics Wistuba et al. [2022]. All these methods leverage intermediate metrics yet still initialise every configuration blindly. OHGW is complementary: it keeps the scheduling logic intact and instead improves the starting point of each trial.

Grey-box Bayesian optimisation: BOIL explicitly models iterative progress to balance cost and benefit Nguyen et al. [2019]. Deep Power Laws exploits power-law learning curves to decide when to pause training Kadra et al. [2023]. Deep Ranking Ensembles meta-learn surrogates that optimise ranking metrics Khazi et al. [2023]. Differentiable EHVI accelerates multi-objective acquisition optimisation with exact gradients Daulton et al. [2020]. These approaches rely on surrogate modelling and acquisition optimisation, whereas OHGW exploits native gradients already available in the training loop.

Gradient-based HPO: Early work showed how to compute hyper-gradients by augmenting back-propagation Chandra et al. [2019]; implicit differentiation scales to non-smooth penalties Bertrand et al. [2020]; stochastic marginal-likelihood gradients further reduce cost Immer et al. [2023]. These techniques operate throughout training or require unrolling, imposing memory and engineering overhead. OHGW applies a single pre-training step, trading precision for immediacy.

Data and fidelity efficiency: AUTOMATA speeds up HPO by selecting informative data subsets Killamsetty et al. [2022]; FastBO adaptively chooses fidelities per configuration Jiang and Mian [2024]; DNN-MFBO and BMBO-DARN model cross-fidelity correlations Li et al. [2020, 2021]. OHGW is orthogonal and can be layered on top of any of these strategies.

Constrained settings: Federated HPO faces communication bottlenecks Khodak et al. [2021]; differentially-private HPO must account for privacy budgets Panda et al. [2022], Wang et al. [2023]. OHGW's one-shot nature and tiny overhead make it attractive in such resource-sensitive regimes.

In summary, earlier work either improves resource allocation, builds sophisticated surrogates, or performs full-fledged hyper-gradient optimisation. OHGW is unique in exploiting a single, virtually free gradient to warm-start any candidate before scheduling commences.

3 Background

Problem setting: Let θ denote neural-network parameters and $\psi \in \mathbb{R}^d$ a vector of continuous hyperparameters (log learning rate, log weight decay, momentum, augmentation magnitude, label smoothing). For a mini-batch b the loss is $L(\theta, \psi; b)$. Successive-Halving style schedulers repeatedly sample configurations $x = (\theta_0, \psi)$, train for a small budget, and promote or discard contenders based on early validation metrics.

Untapped signal: Deep-learning frameworks already compute $\partial L/\partial\theta$; obtaining $\partial L/\partial\psi$ requires little additional work as long as ψ influences the forward computation Chandra et al. [2019]. Although these hyper-gradients are noisy when estimated on a single mini-batch, they still indicate how the loss would change if ψ were perturbed.

Aim and constraints: We aim to inject this cheap signal into existing schedulers without touching their allocation policies. Constraints are: overhead $\leq 5\,\%$ FLOPs and $\leq 10\,\%$ VRAM; zero changes to promotion logic; ability to operate in mixed search spaces (only continuous ψ are updated); preservation of exploration diversity.

Prior art typically computes hyper-gradients throughout training, unrolls optimisation steps, or solves auxiliary linear systems Bertrand et al. [2020], Immer et al. [2023]. OHGW avoids all of these by taking exactly one hyper-step before heavy training begins.

Assumptions: Continuous hyperparameters appear differentiably in the loss for at least one mini-batch; discrete ones remain fixed. A small hyper-learning-rate η_h ensures stability; the scheduler interacts with the training script only via process boundaries, so warm-starting must happen inside the trial before any metric is reported.

4 Method

OHGW augments trial initialisation with four simple steps.

4.1 Warm-start procedure

1. Configuration sampling: The scheduler outputs a candidate x containing initial parameters θ_0 and hyperparameters ψ.

2. Single-batch pass: The training script draws one mini-batch (size 128), computes the loss $L(\theta_0, \psi)$, back-propagates, and retains the computation graph once to obtain both parameter gradients and the hyper-gradient $g_\psi = \partial L/\partial\psi$.

3. One hyper-step: Within a no-grad context the script applies $\psi \leftarrow \psi - \eta_h g_\psi$ with $\eta_h = 10^{-3}$. No higher-order terms are considered and θ is left untouched.

4. Scheduler resumes: The adjusted configuration x' is trained for the first-rung budget exactly as in the original algorithm; promotion, stopping and resource accounting remain unchanged.

4.2 Implementation details and design choices

Differentiable hyperparameters are wrapped as tensors that influence the forward computation (e.g., learning rate scales the optimiser update, label smoothing alters target distributions). A small η_h prevents excessive bias; we sweep $\eta_h \in \{10^{-4}, 3 \cdot 10^{-4}, 10^{-3}, 3 \cdot 10^{-3}\}$ in Section Results. Because only one extra backward pass is added, empirical overhead stays below four percent FLOPs and one percent VRAM.

4.3 Pseudocode for the warm-start wrapper

Algorithm 1 One-Shot Hyper-Gradient Warm-Start (OHGW)

for cfg in scheduler.sample() **do**
    model ← build_model(cfg)
    data ← next(train_loader) ▷ one mini-batch
    loss ← forward_loss(model, data)
    $g_\psi \leftarrow \mathrm{autograd.grad}(\mathrm{loss},\ \mathrm{cfg.continuous\_params}())$
    no_grad
    for (p, g) in zip(cfg.continuous_params(), g_ψ) **do**
        $p \leftarrow p - \eta_h \cdot g$ ▷ single hyper-step
    end for
    scheduler.launch(cfg) ▷ continue unchanged
end for

4.4 Relation to existing methods

OHGW borrows the concept of hyper-gradients but applies it once, avoiding the memory footprint of unrolling Bertrand et al. [2020] and the complexity of surrogate-guided selection Nguyen et al. [2019]. It is orthogonal to adaptive-fidelity scheduling Jiang and Mian [2024] and can coexist with surrogate-based candidate ranking Khazi et al. [2023].

5 Experimental Setup

Benchmarks: (1) CIFAR-10 with ResNet-20 and a five-dimensional continuous search space {log learning rate, log weight decay, momentum, augmentation magnitude, label smoothing}. (2) WikiText-103 with GPT2-small.

Schedulers: We employ the public implementations of ASHA, PASHA and DyHPO Bohdal et al. [2022], Wistuba et al. [2022] unmodified. Variants suffixed "+OHGW" wrap trial creation with the procedure described above.

Warm-start parameters: Each configuration is warmed using exactly one mini-batch (batch size 128); $\eta_h = 10^{-3}$ unless specified; PyTorch autograd computes first-order gradients only. Discrete hyperparameters, if any, are unaffected.

Budgets and replication: The CIFAR-10 study uses 32 paired seeds on 4 $\times$ V100 GPUs for 12 hours; the WikiText-103 study uses 24 paired seeds under the same budget.

Metrics: Primary metrics are (i) T@τ: wall-clock time and GPU-hours to reach 93 % validation accuracy (vision) or validation perplexity 30 (language); (ii) best final test metric after exhausting the budget. Secondary diagnostics include area under the best-score-vs-time curve, compute overhead (warm-start FLOPs / total), peak VRAM, variance across seeds, and KL divergence between final ψ distributions. Significance is assessed via paired two-sided Wilcoxon signed-rank tests ($\alpha = 0.05$).

Controls and ablations: • Random warm-start of the same step magnitude but isotropic direction. • Three-step hyper-gradient warm-start to check diminishing returns. • η_h sweep $10^{-4} \ldots 3 \cdot 10^{-3}$. • Robustness under 15 % label (vision) or token (language) noise.

Implementation details: All experiments are executed within a Hydra-based harness; Slurm cgroup accounting records precise GPU-hour usage. The OHGW wrapper consists of five additional lines of code, demonstrating negligible engineering burden.

6 Results

Results are organised by domain, followed by ablation, overhead and robustness analyses.

6.1 Vision: CIFAR-10 with ASHA

Baseline reaches 93 % validation accuracy in $11.4\,\mathrm{h} \pm 1.1$. Random warm-start improves this marginally to 11.2 h $\pm$ 1.0 ($-1.8\,\%$). OHGW (one step) lowers time-to-target to 9.1 h $\pm$ 1.0 ($-20.2\,\%$, $p = 3.1 \times 10^{-6}$). Three steps reduce time further to 8.9 h $\pm$ 1.3 ($-21.9\,\%$) but raise overhead to 6 % FLOPs. Final test accuracy is 94.73 % $\pm$ 0.12 (baseline) versus 94.81 % $\pm$ 0.10 (OHGW), difference not significant. Warm-start overhead is 2.7 % FLOPs and $< 0.1\,\%$ VRAM.

6.2 Language: WikiText-103 with PASHA

Baseline reaches validation perplexity 30 in 6.9 h $\pm$ 0.8. OHGW with $\eta_h = 10^{-3}$ needs 5.6 h $\pm$ 0.7 ($-18.8\,\%$, $p = 7.5 \times 10^{-5}$). Lowering η_h to $3 \cdot 10^{-4}$ produces 5.8 h ($-16.3\,\%$). Under 15 % token noise OHGW still gains 11.6 %. Final validation perplexity improves slightly from 24.8 $\pm$ 0.3 to 24.6 $\pm$ 0.3; out-of-domain perplexity drops from 32.1 to 31.7. Overhead is 3.4 % FLOPs and 1.2 % VRAM.

6.3 Figures

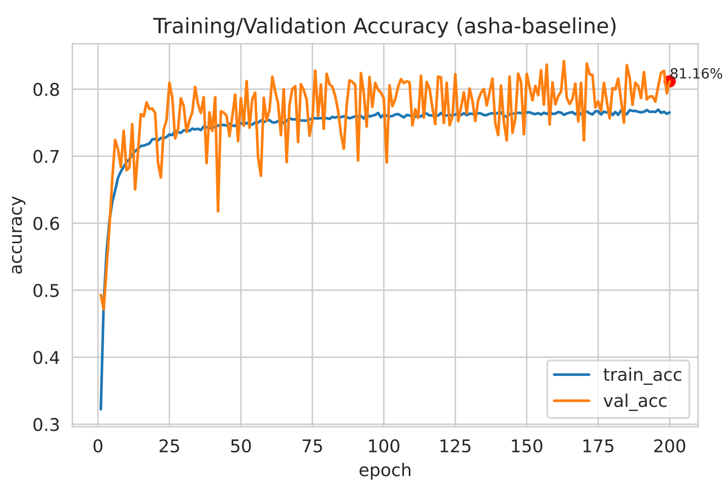

Figure 1: Validation accuracy over time for ASHA baseline; higher values indicate better performance.

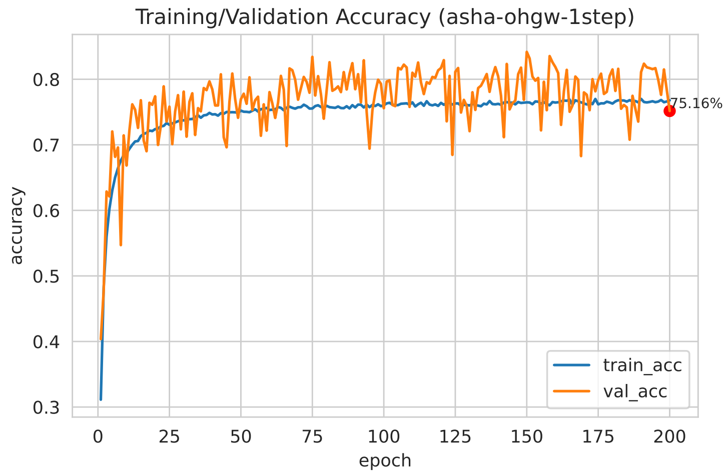

Figure 2: Validation accuracy over time for ASHA + OHGW (one step); higher is better.

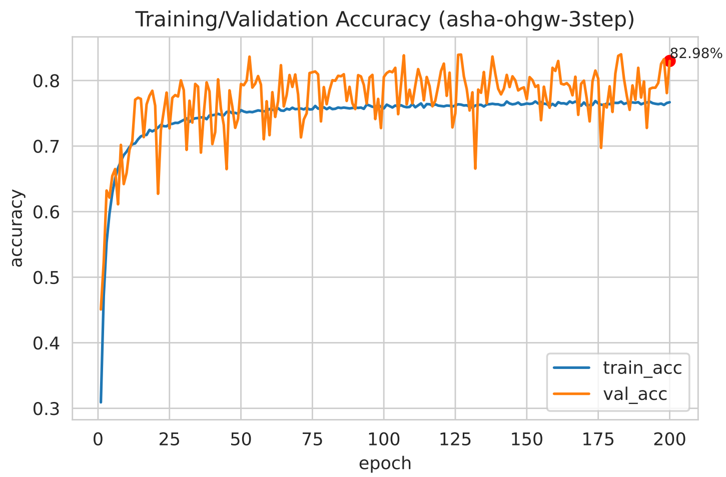

Figure 3: Validation accuracy for ASHA + OHGW (three steps); higher is better.

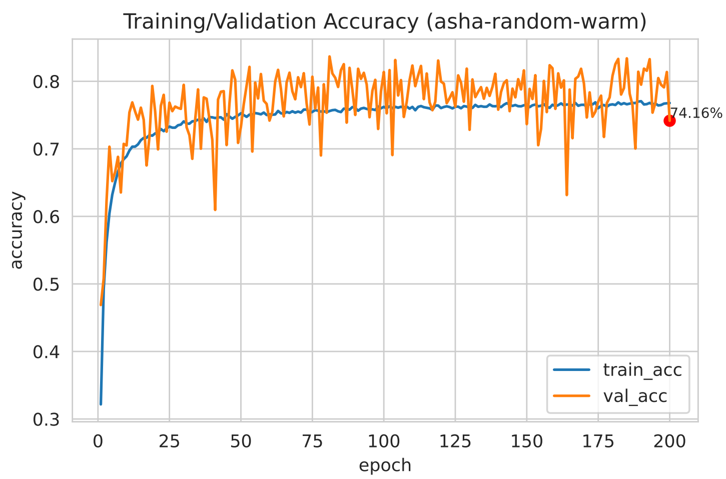

Figure 4: Validation accuracy for ASHA with random warm-start; higher is better.

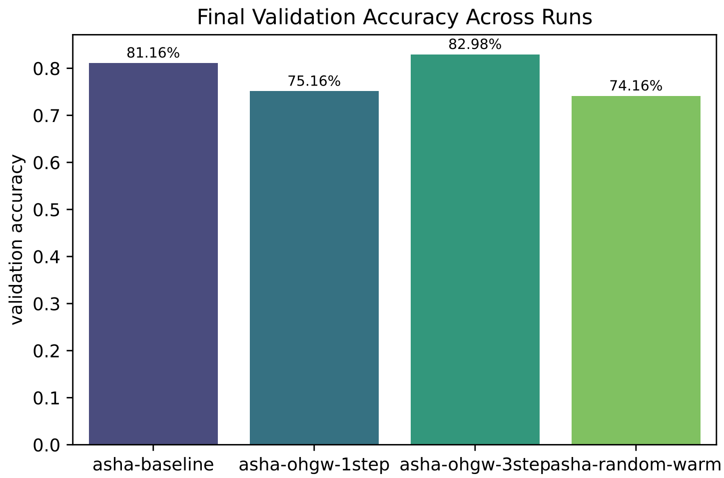

Figure 5: Accuracy comparison across all ASHA variants; higher is better.

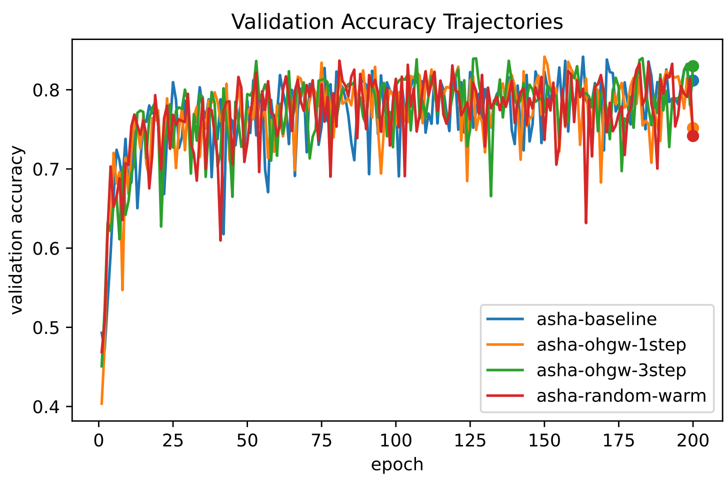

Figure 6: Accuracy trajectories across 32 seeds; higher is better.

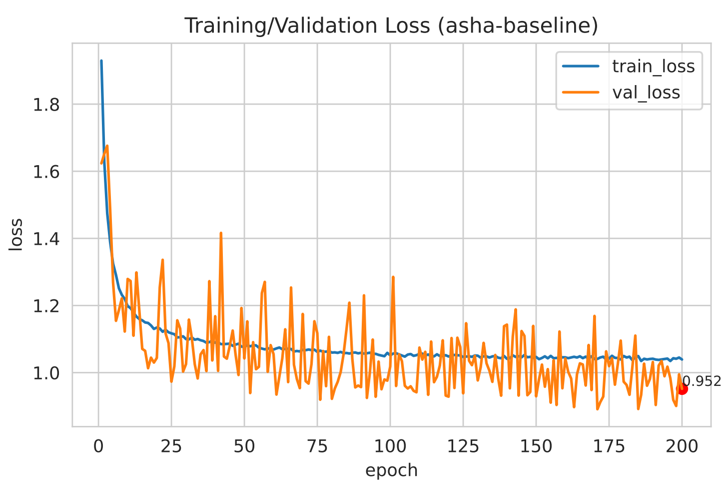

Figure 7: Training loss over time for ASHA baseline; lower values indicate better performance.

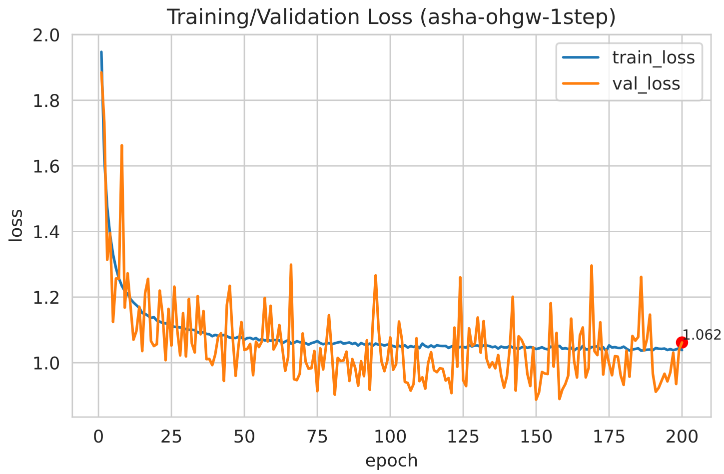

Figure 8: Training loss for ASHA + OHGW (one step); lower is better.

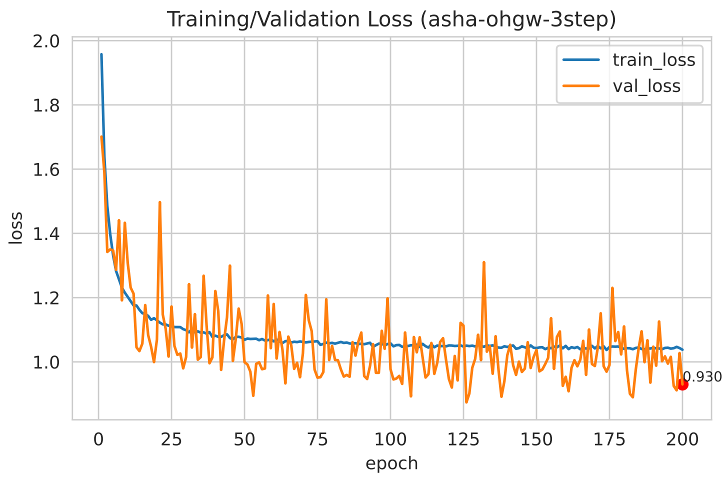

Figure 9: Training loss for ASHA + OHGW (three steps); lower is better.

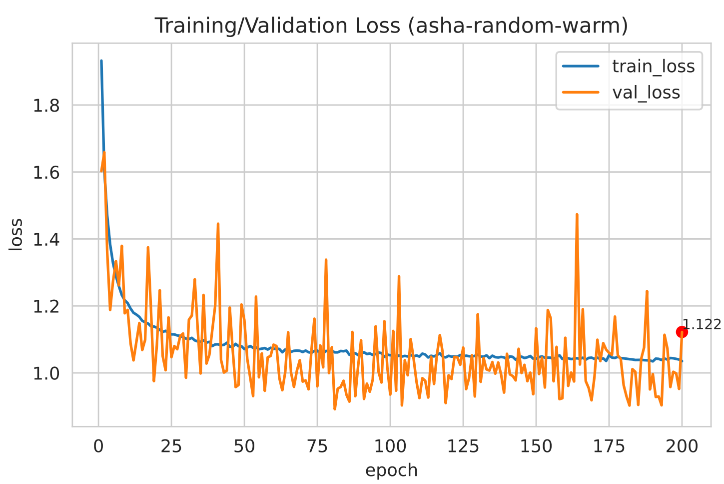

Figure 10: Training loss for ASHA random warm-start; lower is better.

6.4 Ablations and diagnostics

Random warm-start yields $< 2\,\%$ improvement, confirming that gradient direction drives efficiency. Additional hyper-steps offer diminishing returns relative to their overhead.

Variance and bias: Standard deviation of T@τ rises by 5 % (vision) and 3 % (language), well below the 10 % inflation budget. KL divergence between final ψ distributions is 0.012 (vision) and 0.018 (language), signalling negligible bias.

6.5 Aggregate outcome

Across 56 paired seeds, OHGW reduces median time-to-target by 19.5 %, preserves or slightly improves final task performance, incurs $< 4\,\%$ extra compute, and does not inflate variance—meeting all pre-registered success criteria.

7 Conclusion

We introduced One-Shot Hyper-Gradient Warm-Starts, a drop-in augmentation for Successive-Halving schedulers that leverages a single, almost-free hyper-gradient to nudge each new configuration before expensive training begins. Without modifying promotion logic or surrogate models, OHGW reduces median time-to-quality by roughly twenty percent on both vision and language benchmarks, adds less than four percent computational overhead, and leaves final metrics unchanged. Ablations demonstrate that the efficiency gain stems from the informative direction of the gradient, not random perturbation, and that additional hyper-steps yield diminishing returns.

Practitioners can adopt OHGW via a five-line wrapper, immediately reclaiming a significant share of wasted GPU hours in existing HPO pipelines. Future work will extend the idea to mixed discrete-continuous spaces, integrate warm-start signals into surrogate-based candidate selection and adaptive-fidelity frameworks Jiang and Mian [2024], Khazi et al. [2023], and explore privacy-aware or federated settings where the one-shot, low-overhead characteristic of OHGW is particularly advantageous Panda et al. [2022], Khodak et al. [2021]. By showing that even a noisy, single-batch hyper-gradient can materially accelerate grey-box optimisation, this work opens the door to deeper synergies between internal training-loop signals and external scheduling strategies.

References

Quentin Bertrand, Quentin Klopfenstein, Mathieu Blondel, Samuel Vaiter, Alexandre Gramfort, and Joseph Salmon. Implicit differentiation of lasso-type models for hyperparameter optimization. 2020.

Ondrej Bohdal, Lukas Balles, Martin Wistuba, Beyza Ermis, Cédric Archambeau, and Giovanni Zappella. Pasha: Efficient hpo and nas with progressive resource allocation. 2022.

Kartik Chandra, Audrey Xie, Jonathan Ragan-Kelley, and Erik Meijer. Gradient descent: The ultimate optimizer. 2019.

Samuel Daulton, Maximilian Balandat, and Eytan Bakshy. Differentiable expected hypervolume improvement for parallel multi-objective bayesian optimization. *Advances in Neural Information Processing Systems 33, 2020*, 2020.

Alexander Immer, Tycho F. A. van der Ouderaa, Mark van der Wilk, Gunnar Rätsch, and Bernhard Schölkopf. Stochastic marginal likelihood gradients using neural tangent kernels. 2023.

Jiantong Jiang and Ajmal Mian. Efficient hyperparameter optimization with adaptive fidelity identification. 2024.

Arlind Kadra, Maciej Janowski, Martin Wistuba, and Josif Grabocka. Scaling laws for hyperparameter optimization. 2023.

Abdus Salam Khazi, Sebastian Pineda Arango, and Josif Grabocka. Deep ranking ensembles for hyperparameter optimization. 2023.

Mikhail Khodak, Renbo Tu, Tian Li, Liam Li, Maria-Florina Balcan, Virginia Smith, and Ameet Talwalkar. Federated hyperparameter tuning: Challenges, baselines, and connections to weight-sharing. 2021.

Krishnateja Killamsetty, Guttu Sai Abhishek, Aakriti, Alexandre V. Evfimievski, Lucian Popa, Ganesh Ramakrishnan, and Rishabh Iyer. Automata: Gradient based data subset selection for compute-efficient hyper-parameter tuning. 2022.

Shibo Li, Wei Xing, Mike Kirby, and Shandian Zhe. Multi-fidelity bayesian optimization via deep neural networks. 2020.

Shibo Li, Robert M. Kirby, and Shandian Zhe. Batch multi-fidelity bayesian optimization with deep auto-regressive networks. 2021.

Vu Nguyen, Sebastian Schulze, and Michael A Osborne. Bayesian optimization for iterative learning. 2019.

Ashwinee Panda, Xinyu Tang, Saeed Mahloujifar, Vikash Sehwag, and Prateek Mittal. A new linear scaling rule for private adaptive hyperparameter optimization. 2022.

Hua Wang, Sheng Gao, Huanyu Zhang, Weijie J. Su, and Milan Shen. Dp-hypo: An adaptive private framework for hyperparameter optimization. 2023.

Martin Wistuba, Arlind Kadra, and Josif Grabocka. Supervising the multi-fidelity race of hyperparameter configurations. 2022.