Second-Order Delta-Space Saliency: Gradient-Free, Regime-Aware Model Compression from Training Trajectories

Anonymous authors
Paper under double-blind review

Abstract

Which weights of a trained network can be pruned or quantised is usually decided from the loss landscape at convergence, using gradients on calibration data or extra Hessian-vector products. We show that the training trajectory already contains this information and that it can be read from checkpoints alone. Along a gradient-descent path the second difference of the parameters satisfies an exact identity, $a_t = -\eta \bar{H}_t v_{t-1}$, with a path-averaged Hessian. We characterise what this gives: the curvature the optimiser experiences along its own path (a Rayleigh quotient), globally and per block, with a bias under minibatch noise bounded by $1/(\eta\Delta)$ that vanishes with the checkpoint stride Δ; and what it does not: the Hessian diagonal per parameter, which we show is not identifiable. Second-Order Delta-Space Saliency (SODSS) builds on this. It detects second-order events in parameter, representation and token-embedding space with a scale-free statistic (the second difference of log-speed), segments training into temporal regimes, and scores each weight by a regime-weighted path-Fisher saliency $\theta_{N,j}^2 \sum_t w_t v_{t,j}^2$, which we prove is an Optimal-Brain-Damage saliency with the empirical Fisher averaged over the path and which reduces to Synaptic Intelligence without the θ^2 factor. The score needs no gradients, data or Hessian products and is computed in one streaming pass with $O(d)$ memory. On SGD-trained networks (an MLP and a CNN) one-shot global pruning with SODSS is within a point of magnitude pruning up to 95% sparsity, beats gradient-based Fisher and SNIP criteria throughout, and is clearly best at 98%; the event detector recovers learning-rate schedule changes and representational phase changes without being told the schedule. Under AdamW the raw score fails, because the premise $v_t \propto g_t$ does not hold; undoing the preconditioner with the second-moment state that checkpoints already contain restores it, and the corrected score then beats magnitude pruning on an AdamW-trained Transformer at every sparsity. We also report what does not work: per-parameter second-order statistics do not help, and regime weighting has no measurable effect.

1 Introduction

Compressing a trained network by pruning (Han et al., 2015; LeCun et al., 1990) or reducing precision (Han et al., 2016; Dong et al., 2019) requires an importance score per parameter. Most criteria are computed at the end point of training: the weight magnitude $|\theta_j|$, a first-order sensitivity $|g_j\theta_j|$ (Lee et al., 2019; Molchanov et al., 2019), or a second-order quantity such as the empirical Fisher or a Hessian estimate (LeCun et al., 1990; Hassibi & Stork, 1993; Singh & Alistarh, 2020; Frantar & Alistarh, 2022). Second-order criteria are the most accurate but require gradients on calibration data and, for Hessian-based methods, additional Hessian-vector products or factorised approximations (Martens & Grosse, 2015; Yao et al., 2020).

This paper takes a different view of where importance information lives. A training trajectory $\theta_0, \ldots, \theta_N$ is routinely checkpointed and then discarded. Its first differences $v_t =$

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.

$\theta_t - \theta_{t-1}$ are, under gradient descent, scaled gradients; its second differences $a_t = v_t - v_{t-1}$ are, as we show, scaled Hessian-vector products. The trajectory therefore contains first- and second-order information about the loss landscape along the path the optimiser actually took, at no additional cost.

We make this idea precise and ask three questions. (1) What curvature information does a trajectory carry? We prove an exact identity $a_t = -\eta \bar{H}_t v_{t-1}$ with a path-averaged Hessian, derive the resulting Rayleigh quotient along the path, and characterise the bias introduced by minibatch noise: it is bounded by $1/(\eta\Delta)$ for snapshot stride Δ, so sparse checkpointing improves the estimate. We also give a negative result: per-coordinate regression of a_t on v_{t-1} does not recover the Hessian diagonal, because off-diagonal terms dominate. (2) Can trajectories be segmented into meaningful phases? We detect second-order events with a scale-free statistic, the second difference of log-speed, in parameter space, in a probe representation space measured by centred kernel alignment (Kornblith et al., 2019), and in the token-embedding space of language models, and we classify the spans between events into four temporal regimes. (3) Does a per-parameter trajectory score prune well? We define the SODSS saliency $s_j = \theta_{N,j}^2 \sum_t w_t v_{t,j}^2$ with regime weights w_t and prove that it is an Optimal-Brain-Damage saliency with a path-averaged empirical Fisher; dropping θ^2 recovers Synaptic Intelligence (Zenke et al., 2017),.

Contributions. (i) An exact second-order identity for trajectories, its Rayleigh-quotient corollary, a noise-bias bound explaining why checkpoint stride helps, and evidence that the Hessian diagonal is not identifiable per coordinate (§3.2, §4.1). (ii) A scale-free multi-space event detector and regime classifier that recover schedule changes and representational phase changes from checkpoints alone (§3.3, §4.2). (iii) The SODSS saliency: gradient-free, data-free, optimiser-agnostic, streaming $O(d)$, with its relation to Fisher/OBD, Synaptic Intelligence and Movement pruning, and a budget-driven three-way rule (§3.5, §3.6). (iv) A proof-of-concept evaluation on three architectures and two optimisers against magnitude, random, gradient-based and trajectory-based baselines, with ablations and clearly stated negative results, including the failure of the raw score under AdamW and its repair from optimiser state (§4).

2 Related work

Pruning criteria. Magnitude pruning (Han et al., 2015) remains a strong baseline (Gale et al., 2019; Blalock et al., 2020). First-order criteria score $|g_j\theta_j|$ (Lee et al., 2019; Molchanov et al., 2019), GraSP preserves gradient flow (Wang et al., 2020), SynFlow is data-free but needs forward/backward passes (Tanaka et al., 2020). Second-order criteria descend from Optimal Brain Damage and Surgeon (LeCun et al., 1990; Hassibi & Stork, 1993) through the empirical Fisher (Theis et al., 2018; Singh & Alistarh, 2020) to layer-wise reconstruction for LLMs (Frantar & Alistarh, 2022; 2023; Sun et al., 2024); Hessian-aware mixed precision (Dong et al., 2019; 2020) assigns bit-widths from Hessian spectra. All compute their statistics at the final checkpoint with backpropagation (or activations) on calibration data; SODSS needs neither.

Importance from training dynamics. Synaptic Intelligence (Zenke et al., 2017) accumulates $\omega_j = \sum_t -g_{t,j}\Delta\theta_{t,j}$ during training; Memory Aware Synapses (Aljundi et al., 2018) and EWC (Kirkpatrick et al., 2017) use output sensitivity and the diagonal Fisher; Movement pruning (Sanh et al., 2020) accumulates $-\sum_t g_{t,j}\theta_{t,j}$ during fine-tuning. Dataset cartography (Swayamdipta et al., 2020) categorises examples by their dynamics; we do the analogous thing for parameters, show that SI and Movement are functions of the trajectory alone, and show that the θ_N^2 factor turning SI into an OBD saliency is what makes it competitive one-shot.

Curvature, phases and NAS. The coupling of step size and sharpness along the path is studied by Cohen et al. (2021) and Jastrzębski et al. (2020), and Gur-Ari et al. (2018) show gradients concentrate in a small top-eigenspace; our Rayleigh estimator measures curvature in exactly the direction of motion. Critical periods (Achille et al., 2019), the early phase

(Frankle et al., 2020), rewinding (Frankle & Carbin, 2019) and grokking (Power et al., 2022; Nanda et al., 2023) document distinct phases; our detector is a checkpoint-only way to locate their boundaries. Training-free NAS (Mellor et al., 2021) shares only the premise that partial training signals are informative; we do not search over architectures.

3 Method

3.1 Setting and notation

Let $\theta_t \in \mathbb{R}^d$ be the prunable parameters (weight matrices and convolution kernels; biases and normalisation parameters are kept) at training step t, and let a trajectory be recorded at stride Δ: snapshots $\theta_0, \theta_\Delta, \theta_{2\Delta}, \dots$, re-indexed as $\theta_0, \dots, \theta_N$. We also record the mean training loss L_t over each stride window, the learning rate η_t, and optionally a probe representation $H_t \in \mathbb{R}^{B\times k}$ (activations of a fixed probe batch at a chosen layer) and, for language models, the token-embedding matrix G_t. Define the discrete velocity and acceleration

$$v_t = \theta_t - \theta_{t-1}, \qquad a_t = v_t - v_{t-1} = \theta_t - 2\theta_{t-1} + \theta_{t-2}, \tag{1}$$

where a_t is the backward second difference at t. The goal is a specification $c_j \in \{\text{Keep}, \text{Quantise}, \text{Prune}\}$, $j = 1, \dots, d$, under a bit budget, computed from $\{\theta_t, L_t, \eta_t, H_t, G_t\}$ without gradients or data.

3.2 What second differences of a trajectory measure

Proposition 1 (Exact second-order identity). Let $\theta_t = \theta_{t-1} - \eta\nabla L(\theta_{t-1})$ be full-batch gradient descent with constant step η and twice continuously differentiable L. Then for every $t \geq 2$,

$$a_t = -\eta\, \bar{H}_t\, v_{t-1}, \qquad \bar{H}_t = \int_0^1 \nabla^2 L\big(\theta_{t-2} + s\, v_{t-1}\big)\, ds. \tag{2}$$

Proof. $a_t = v_t - v_{t-1} = -\eta[\nabla L(\theta_{t-1}) - \nabla L(\theta_{t-2})]$, and the bracket equals $\int_0^1 \nabla^2 L(\theta_{t-2} + sv_{t-1})\, v_{t-1}\, ds$ by the fundamental theorem of calculus applied to $s \mapsto \nabla L(\theta_{t-2} + sv_{t-1})$. □

The identity is exact with the path-averaged Hessian (the mean-value theorem, often invoked here, does not hold for vector-valued maps and is not needed).

Corollary 1 (Curvature along the path). $r_t := -\dfrac{\langle a_t, v_{t-1}\rangle}{\eta\,\|v_{t-1}\|^2} = \dfrac{v_{t-1}^\top \bar{H}_t v_{t-1}}{\|v_{t-1}\|^2} \in [\lambda_{\min}(\bar{H}_t), \lambda_{\max}(\bar{H}_t)]$ is the Rayleigh quotient of the path-averaged Hessian in the direction of motion. For a coordinate block B, $r_t^B := -\langle a_t^B, v_{t-1}^B\rangle/(\eta\|v_{t-1}^B\|^2) = (v^B)^\top(\bar{H}v)^B/\|v^B\|^2$ is the block-restricted curvature along the path.

Proposition 2 (Noise bias under SGD and the effect of stride). Let SGD use gradient estimates $\nabla L(\theta) + \xi$ with zero-mean noise ξ, independent across steps, $\mathbb{E}\|\xi\|^2 = \sigma^2$, and let snapshots be taken every Δ steps so that the effective step is $\eta\Delta$. Let $\bar{g}$ be the mean full-batch gradient over the window and $\hat{r}_t = -\langle a_t, v_{t-1}\rangle/(\eta\Delta\|v_{t-1}\|^2)$ the estimate computed from the sub-sampled trajectory. Then, to first order in η and with the ratio of expectations in place of the expectation of the ratio,

$$\hat{r}_t \;\approx\; \frac{v_{t-1}^\top \bar{H}_t v_{t-1}}{\|v_{t-1}\|^2} \;+\; \beta_\Delta, \qquad \beta_\Delta = \frac{\sigma^2}{\eta\Delta\,\big(\Delta\|\bar{g}\|^2 + \sigma^2\big)} \;\leq\; \frac{1}{\eta\Delta}. \tag{3}$$

In the noise-dominated case ($\sigma^2 \gg \Delta\|\bar{g}\|^2$) $\beta_\Delta \to 1/(\eta\Delta)$; in the signal-dominated case $\beta_\Delta \to \sigma^2/(\eta\Delta^2\|\bar{g}\|^2)$; in both cases the bias is positive and $O(1/\Delta)$.

Proof sketch. Over a window of Δ steps the displacement is $v_{t-1} = -\eta(\Delta\bar{g} + \Xi_{t-1})$ with $\Xi = \sum_s \xi_s$, $\mathbb{E}\|\Xi\|^2 = \Delta\sigma^2$, and to first order $a_t = -\eta\Delta\bar{H}_t v_{t-1} - \eta(\Xi_t - \Xi_{t-1})$. The noise contributes to $-\langle a_t, v_{t-1}\rangle$ only through $\mathbb{E}\langle \Xi_t - \Xi_{t-1}, -\eta(\Delta\bar{g} + \Xi_{t-1})\rangle = \eta\Delta\sigma^2$, while $\mathbb{E}\|v_{t-1}\|^2 = \eta^2(\Delta^2\|\bar{g}\|^2 + \Delta\sigma^2)$. Dividing by $\eta\Delta\,\mathbb{E}\|v_{t-1}\|^2$ gives β_Δ. Full derivation in Appendix A. □

Thus the raw per-step acceleration under SGD is dominated by a noise term of order $1/\eta$ (20 at $\eta = 0.05$, larger than any eigenvalue of interest), but sub-sampling at stride Δ shrinks the bias as $1/\Delta$: sparse checkpointing, which is what is available in practice, is the right regime for trajectory curvature. Both the identity and the bound are verified in §4.1.

Remark 1 (The Hessian diagonal is not identifiable per coordinate). Regressing $a_{t,j}$ on $v_{t-1,j}$ over a window estimates $\sum_k \bar{H}_{jk}\, \overline{v_k v_j}/\overline{v_j^2}$, which equals $\bar{H}_{jj}$ only if the off-diagonal contributions average out. Because v concentrates in the top eigenspace of H (Gur-Ari et al., 2018), they do not: in §4.1 the per-coordinate estimate is uncorrelated with $\mathrm{diag}(H)$ even for noise-free GD. Second-order information in a trajectory is therefore available at the level of directions and blocks, not individual weights. This is why SODSS uses second-order statistics for temporal decisions (events, regimes) and a first-order-squared statistic for per-parameter decisions.

3.3 Second-order events in three spaces

Thresholding $\|a_t - a_{t-1}\|$ against a fixed constant, as in earlier trajectory-analysis designs, is not scale-free: the threshold's meaning depends on d, η, Δ and the parameter scale, and under SGD noise the relative third difference $\|a_t - a_{t-1}\|/\|v_{t-1}\|$ is $\approx \sqrt{6}$ for every step regardless of any event. We instead detect events in the log-speed series. For a sequence of velocity magnitudes s_t let $\ell_t = \log s_t$ and define the event score

$$e_t = \big|\ell_t - 2\ell_{t-1} + \ell_{t-2}\big|, \tag{4}$$

the second-order delta of log-speed. A learning-rate drop by a factor c produces $e_t \approx |\log c|$ independent of scale; a plateau produces $e_t \approx 0$; a phase transition that changes the rate of parameter motion produces a localised spike. Events are the indices whose robust z-score $(e_t - \text{median})/(1.4826\,\text{MAD})$ exceeds $z^\star$ (default 3.5), after non-maximum suppression with a minimum gap of 5 snapshots and a cap of 8 events, which keeps the number of spans small enough for the per-span statistics below to be meaningful. The same operator is applied to three speed series: parameter space, $s_t = \|v_t\|_2$ (per-block speeds $\|v_t^B\|$ attribute events to layers); representation (semantic) space, $s_t = 1 - \mathrm{CKA}(H_t, H_{t-1})$ with linear CKA (Kornblith et al., 2019) on the probe activations, which is invariant to orthogonal transformations and isotropic scaling, unlike the centroid of an embedding set (nearly constant under LayerNorm and zero-mean initialisation, and blind to all structure); and lexical space, $s_t = \|G_t - G_{t-1}\|_F$ for the token-embedding matrix of language models. Let I_θ, I_H, I_G be the resulting event sets and κ the fraction of parameter-space events that have a representation-space event within δ snapshots (default $\delta = 2$).

3.4 Temporal regimes

The union of events partitions $\{0, \dots, N\}$ into spans $S_i = [b_i, b_{i+1})$. For each span we compute three scale-free statistics on the loss series $L_t/\bar{L}_{S_i}$: the ordinary-least-squares slope λ_i, the standard deviation σ_i of the residuals about that linear trend (the coefficient of variation around the trend, so that a steadily decreasing loss is not mistaken for a noisy one), and the cross-space coincidence κ_i of the event that opens the span. Spans are labelled by priority-ordered rules:

$$R_i = \begin{cases} \text{ConvergenceAttractor (CA)} & |\lambda_i| < \tau_\lambda \text{ and } \sigma_i < \tau_\sigma, \\ \text{BoundaryRedefinition (BR)} & \text{opening event coincident across spaces } (\kappa_i \geq \tfrac{1}{2}), \\ \text{MonotonicRefinement (MR)} & \lambda_i < -\tau_\lambda \text{ and } \sigma_i < \tau'_\sigma, \\ \text{InterruptedRefinement (IR)} & \text{otherwise,} \end{cases} \tag{5}$$

with $\tau_\lambda = 0.002$ per snapshot, $\tau_\sigma = 0.05$, $\tau'_\sigma = 0.15$, and regime weights $w(\mathrm{BR}) = 1.0$, $w(\mathrm{MR}) = 0.7$, $w(\mathrm{IR}) = 0.5$, $w(\mathrm{CA}) = 0.2$ assigned to every snapshot in the span, $w_t = w(R_{i(t)})$. These labels are interpretable summaries of the dynamics (§4.2); their quantitative effect on pruning is evaluated by ablation and found to be small (§4.4).

3.5 Per-parameter saliency

Definition 1 (SODSS saliency). $s_j = \theta_{N,j}^2 \, \mathrm{PF}_j, \qquad \mathrm{PF}_j = \sum_{t=1}^{N} w_t \, v_{t,j}^2.$

Proposition 3 (Path-Fisher is the OBD saliency with a path-averaged Fisher). For SGD with stride $\Delta = 1$ and step sizes η_t, $v_{t,j} = -\eta_{t-1} g_{t-1,j}$ where g is the minibatch gradient. Hence $\mathrm{PF}_j = \sum_t w_t \eta_{t-1}^2 g_{t-1,j}^2$, so that for constant η and $w \equiv 1$, $\mathrm{PF}_j/(N\eta^2) = \bar{F}_{jj}$ is the diagonal of the empirical Fisher averaged along the path, and $s_j = N\eta^2 \cdot \bar{F}_{jj}\theta_{N,j}^2$ is, up to the constant, the Optimal-Brain-Damage saliency $\frac{1}{2} H_{jj}\theta_j^2$ with $H \approx \bar{F}$. Moreover $\mathrm{PF}_j/\eta = \sum_t -g_{t-1,j}\Delta\theta_{t,j} = \omega_j$ is exactly the Synaptic-Intelligence path integral, and $\sum_t v_{t,j}\theta_{t-1,j}/\eta = -\sum_t g_{t-1,j}\theta_{t-1,j}$ is the Movement-pruning score.

Proof. Substitute $v_t = -\eta_{t-1} g_{t-1}$ into each definition. □

The path-averaged Fisher weights coordinates by how much the optimiser actually moved them throughout training and is unaffected by the end-point gradient vanishing at convergence; empirically it correlates with $\mathrm{diag}(H)$ (Spearman ≈ 0.65 under SGD, §4.1), consistent with the Fisher–Gauss–Newton relation (Kunstner et al., 2019). With stride $\Delta > 1$, v_t sums Δ updates and PF is a smoothed version of the same quantity. Under Adam (Kingma & Ba, 2015), $v_t = -\eta_t \hat{m}_t/(\sqrt{\hat{v}_t} + \epsilon)$ and PF is the energy of the preconditioned update, i.e. how far the optimiser moved the weight in its own metric; the score stays well defined and is tested in §4.3. The factor $\theta_{N,j}^2$ is essential: without it the score is SI, which prunes markedly worse one-shot (Table 2).

3.6 Budgeted three-way specification

Fixed thresholds on a normalised score are brittle (a prune threshold below the score's own lower bound is never reached). We instead rank by s_j and choose fractions (k, p) to meet a bit budget: the top k fraction is kept in full precision, the bottom p fraction pruned, the rest quantised (INT8, symmetric per-tensor scale), for $32k + 8(1-k-p)$ average bits per stored weight; sparsity and bits are reported separately. A short masked fine-tune, optionally distilling from the uncompressed model (Hinton et al., 2015), can follow; we report one-shot and fine-tuned results.

3.7 Streaming implementation and complexity

All per-parameter statistics are sums over t ($\sum v_t^2$, $\sum v_t\theta_{t-1}$, and for the ablations $\sum a_t^2$, $\sum a_t v_{t-1}$) and are accumulated online in $O(d)$ memory, together with the scalar series $\|v_t\|$, $\|v_t^B\|$, $1 - \mathrm{CKA}(H_t, H_{t-1})$ and L_t. Since regime labels are known only after all events are, per-parameter sums are kept per window of W snapshots and combined with the weights at the end ($O(dN/W)$ memory; 15 vectors of size d for $W = 20$, $N \le 300$). Time is one $O(Nd)$ pass over the checkpoints with no forward or backward passes. Algorithm 1 (Appendix C) summarises the pipeline.

4 Experiments

We ask in turn whether a trajectory carries the curvature information claimed in §3.2 (§4.1), whether the event detector finds meaningful phase boundaries (§4.2), and whether the saliency prunes well without gradients (§4.3–4.4). All experiments are deliberately small (CPU only, under two hours in total) so that every number is reproducible from the supplementary code; scale is discussed in §5; hyperparameters are in Appendix B.

Testbeds. (i) MNIST / LeNet-300-100, plain SGD, step schedule (learning rate ×0.1 at 50% and 80% of training), three seeds; (ii) CIFAR-10 / 5-layer CNN, plain SGD, same schedule; (iii) character-level Transformer on Tiny Shakespeare with AdamW, warm-up and cosine decay, i.e. a preconditioned optimiser with no discrete schedule events, and all three

spaces (parameters, final-layer hidden states of a probe batch, token embeddings). Snapshot stride $\Delta = 10$ (20 for CIFAR-10).

Baselines. Gradient-free (checkpoints only): random, magnitude (Han et al., 2015), Synaptic Intelligence in trajectory form $\sum_t v_{t,j}^2$ (Zenke et al., 2017), the Movement score in trajectory form $\sum_t v_{t,j}\theta_{t-1,j}$ (Sanh et al., 2020), and the original layer-level rule (layers ranked by event count, weights within a layer by magnitude). Gradient-based reference (training data and backpropagation at the final checkpoint): empirical Fisher $\times\theta^2$ (diagonal OBD) and SNIP $|g\theta|$ (Lee et al., 2019). All are applied as one-shot global unstructured pruning at matched sparsity, optionally followed by a short masked fine-tune.

4.1 Does the trajectory carry curvature?

We train a $784-32-10$ MLP on 5,000 MNIST images, recording every step, and compare trajectory quantities with exact Hessian computations (Figure 1; full numbers in Table 3, Appendix C). Three findings. (1) The identity is exact. Under full-batch GD the trajectory Rayleigh quotient $\hat{r}_t$ agrees with the exact $v^\top Hv/\|v\|^2$ (one Hessian-vector product at each of the last ten snapshots) to within 1.5% (0.0223 ± 0.0001 vs. 0.0226 ± 0.0001), globally and per layer block, at every stride tested. (2) Noise bias behaves as Proposition 2 predicts. Under minibatch SGD the per-step estimate is dominated by noise ($\hat{r} = 20.9 \pm 1.1$ against a true value of 0.75 ± 0.21 at $\eta = 0.05$, where the bound is $1/\eta = 20$), and the bias falls with the snapshot stride exactly as the $1/(\eta\Delta)$ bound predicts ($20.1, 9.3, 3.5, 1.5, 0.5, 0.04$ for $\Delta = 1, 2, 5, 10, 20, 50$ against bounds $20, 10, 4, 2, 1, 0.4$), dropping below the bound once the signal dominates; at $\Delta = 50$ the estimate is within 0.05 of the exact value. Checkpoints taken every few dozen steps, the usual practice, are therefore the right input. We also note that the curvature along the path is far below $\lambda_{\max}(H) \approx 3.6$–$6$: after the initial phase GD moves in low-curvature directions, consistent with Gur-Ari et al. (2018) and Cohen et al. (2021). (3) The diagonal is not identifiable per coordinate. Per-coordinate regression $\hat{H}_{jj}$ has Spearman correlation ≈ 0 with the Hutchinson estimate of $\mathrm{diag}(H)$ even for noise-free GD, and is negatively correlated under SGD (Remark 1). By contrast the first-order-squared path-Fisher statistic $\sum_t v_{t,j}^2$ correlates with $\mathrm{diag}(H)$ at $\rho \approx 0.65$ under SGD (0.38 under GD), which is the empirical basis for using it as the per-parameter criterion.

4.2 Events and regimes

Figure 2 and Table 5 (appendix) show what the detector finds. On MNIST both learning-rate drops are recovered in parameter space and, independently, in representation space (CKA drift falls by two orders of magnitude at each drop), within one snapshot of the true event across seeds; the only other events are at the start of training. On CIFAR-10 the parameter-space detector finds both drops (snapshots 81 and 128 for drops after 78 and 125) but the representation detector fires nowhere, since the CKA drift changes smoothly through the drops (Figure 4). The AdamW Transformer has no discrete schedule events: the detector marks the end of warm-up and the final collapse of the update size as the cosine schedule reaches zero, and the representation detector marks where hidden states stop changing (snapshots 130–150), labelled as a convergence attractor. The lexical space adds mid-training events (66, 102, 111) absent from the other spaces. Regime labels are plausible in all runs.

4.3 One-shot pruning

Table 1 gives the main result (curves in Figure 3, Appendix C). MNIST (SGD). Up to 80% sparsity every informative criterion is within a point of the dense model (97.5%). One-shot, magnitude is best at 90% (95.7%) with SODSS two points behind (93.5%) and ahead of both gradient-based references (Fisher 89.0%, SNIP 86.5%); at 95% SODSS, Movement and magnitude are within 1.5 points; at 98% SODSS is clearly best (65.6 ± 4.8% against 48–55% for every other criterion, gradient-based ones included). After 300 fine-tuning steps every non-random criterion recovers to 96–97% at 90% sparsity; the trajectory form of the Movement score is then the strongest at all sparsities (93.9% at 98%), with SODSS,

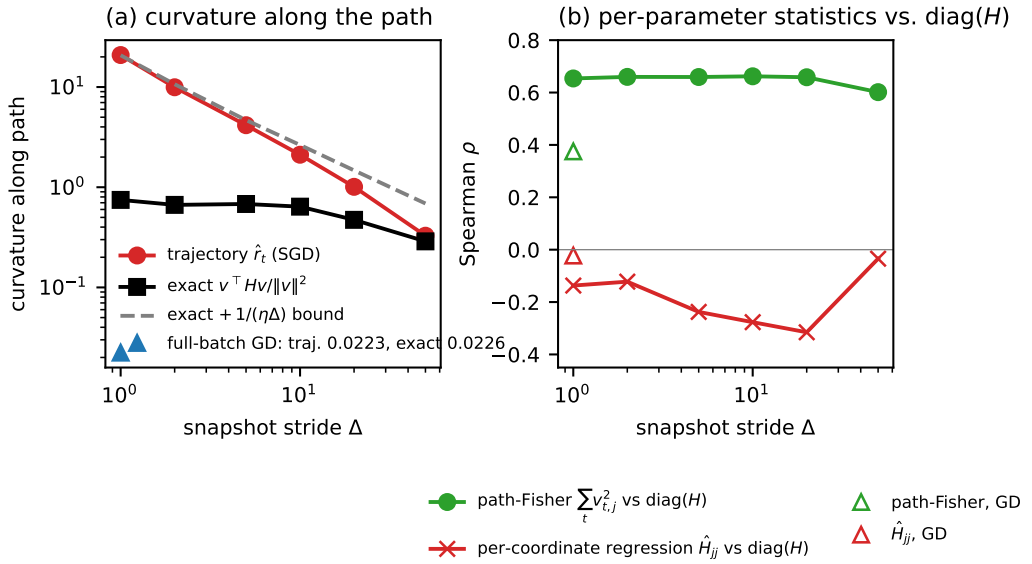

Figure 1: Curvature from checkpoints alone (MLP, MNIST subset). (a) Trajectory Rayleigh quotient under SGD vs. the exact value and the $1/(\eta\Delta)$ bound of Prop. 2; the GD point shows the identity holding exactly. (b) Spearman correlation with diag(H): path-Fisher is informative, per-coordinate regression is not.

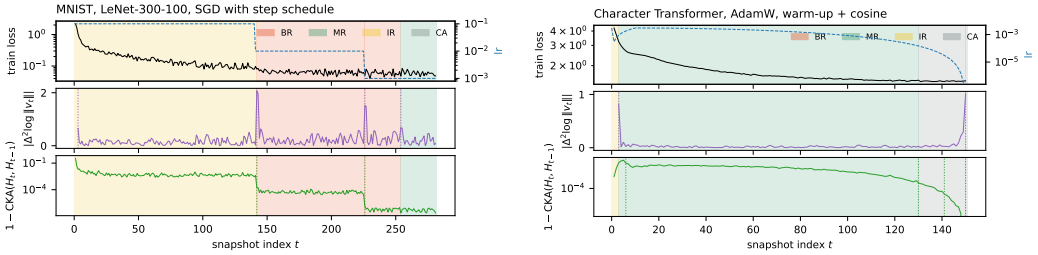

Figure 2: Events and regimes recovered from checkpoints (seed 0). Top: loss (black), learning rate (dashed); middle: parameter-space event score with detected events (dotted); bottom: representation drift $1 - \mathrm{CKA}(H_t, H_{t-1})$ with its events; shading: regimes. Left, MNIST: both learning-rate drops (snapshots 141, 225) are found in both spaces without access to the schedule and open BR spans. Right, AdamW Transformer with cosine decay: events at the end of warm-up and of decay; representation events where the representation freezes (CA).

magnitude and Fisher within 0.3 points of each other. CIFAR-10 (SGD). The same pattern holds on the CNN (dense 67.6%): one-shot, SODSS is within 1.7 points of magnitude at 90% (59.7% vs. 61.4%), ahead of Fisher (50.9%) and SNIP (37.8%), and clearly best at 95% (51.2% vs. 38.6% magnitude, 43.6% Fisher, 50.2% Movement); after fine-tuning SODSS, magnitude and Movement are within a point of each other at 90% and SODSS leads magnitude by 1.7 points at 95%. Character Transformer (AdamW). Here the premise of Proposition 3, $v_t \propto g_t$, does not hold, and the result is negative: magnitude pruning is best at every sparsity (1.86 nats/char at 50% against 1.64 dense), the path-Fisher saliency is far behind (2.90), and so is the gradient-based end-point Fisher (2.77) and SNIP (3.00): Fisher-type saliency as such, not just its trajectory form, fails on this Adam-trained model. Only the trajectory-curvature ratio (Table 2) matches magnitude (1.86). The repair follows from the theory: multiplying v_t by the AdamW denominator $\sqrt{\hat{v}_t} + \epsilon$ saved in the checkpoint

Table 1: One-shot global unstructured pruning at matched sparsity. MNIST: test accuracy (%, mean±std over 3 seeds); CIFAR-10: test accuracy (%); character LM: validation loss (nats/char, lower is better). Bold: best gradient-free criterion. Gradient-based criteria (bottom two rows) use the training data and backpropagation at the final checkpoint and are shown as reference.

Criterion	MNIST (acc. %)			CIFAR-10 (acc. %)			Char-LM, AdamW (loss)		
	90%	95%	98%	80%	90%	95%	50%	70%	80%
one-shot									
SODSS	93.5±0.3	87.8±0.3	65.6±4.8	64.7	59.7	51.2	2.898	4.033	4.445
SODSS, Adam-corrected	–	–	–	–	–	–	1.755	2.182	2.735
Synaptic Intelligence (traj.)	84.2±0.1	68.2±4.7	50.6±9.3	57.3	53.5	36.5	5.038	4.307	4.222
Movement score (traj.)	95.4±0.5	88.2±1.6	54.6±12.1	60.0	57.7	50.2	3.700	3.825	3.817
Magnitude	95.7±0.7	86.7±2.4	48.1±11.6	65.6	61.4	38.6	1.862	2.629	3.410
Random	14.7±3.3	10.6±0.5	10.5±0.6	10.0	10.0	10.0	3.563	3.926	4.061
Original layer-level rule	38.8±40.5	37.1±38.2	23.6±19.1	10.0	10.0	10.0	4.206	4.174	4.174
Fisher $\times\theta^2$ (gradients)	89.0±0.2	74.0±2.6	49.7±4.1	62.3	50.9	43.6	2.766	3.543	3.482
SNIP $\lvert g\theta\rvert$ (gradients)	86.5±2.2	65.0±4.8	48.9±3.4	56.3	37.8	25.0	3.002	3.373	3.491
fine-tuned									
SODSS	96.9±0.0	96.0±0.1	92.7±0.4	67.3	66.4	63.9	1.682	1.997	2.451
SODSS, Adam-corrected	–	–	–	–	–	–	1.664	1.747	1.871
Synaptic Intelligence (traj.)	95.4±0.1	93.5±0.4	89.4±0.3	65.8	63.5	60.1	3.070	3.516	3.811
Movement score (traj.)	97.3±0.0	96.6±0.0	93.9±0.1	67.8	66.8	65.1	2.039	2.195	2.316
Magnitude	97.1±0.0	96.1±0.1	92.4±0.2	67.7	66.1	62.2	1.664	1.774	1.982
Random	84.4±1.1	40.6±4.9	15.2±3.9	25.7	10.0	10.0	2.120	2.455	2.512
Original layer-level rule	39.9±40.4	39.7±40.1	39.1±39.3	10.0	10.0	10.0	4.005	4.174	4.174
Fisher $\times\theta^2$ (gradients)	96.6±0.1	95.4±0.3	91.4±0.4	66.5	65.7	62.0	1.711	2.171	2.305
SNIP $\lvert g\theta\rvert$ (gradients)	96.3±0.1	94.8±0.2	90.9±0.4	66.8	64.9	61.8	1.771	2.228	2.361

Table 2: Ablation of the SODSS saliency (same protocol as Table 1).

Criterion	MNIST (acc. %)			CIFAR-10 (acc. %)			Char-LM, AdamW (loss)		
	90%	95%	98%	80%	90%	95%	50%	70%	80%
one-shot									
SODSS	93.5±0.3	87.8±0.3	65.6±4.8	64.7	59.7	51.2	2.898	4.033	4.445
SODSS, no regime weights	93.5±0.3	87.8±0.3	65.7±4.7	64.7	59.4	51.3	2.896	4.027	4.436
SODSS, first half only	93.5±0.3	87.7±0.2	65.8±5.0	64.8	59.5	51.1	2.858	4.051	4.459
SODSS, second half only	91.4±1.2	80.4±2.0	50.9±5.9	62.0	51.2	40.6	2.926	4.162	4.476
SODSS without θ^2	84.1±0.1	68.3±4.6	50.7±9.2	57.3	53.3	36.8	5.039	4.306	4.222
Traj. curvature ratio $\times\theta^2$	93.5±0.6	72.7±5.6	38.2±4.5	63.3	54.8	23.0	1.858	2.808	3.545
Acceleration energy $\times\theta^2$	92.3±0.5	84.3±0.6	56.1±5.6	63.6	58.2	49.4	2.901	4.119	4.471
fine-tuned									
SODSS	96.9±0.0	96.0±0.1	92.7±0.4	67.3	66.4	63.9	1.682	1.997	2.451
SODSS, no regime weights	96.9±0.0	96.0±0.1	92.7±0.4	67.4	66.5	64.0	1.682	1.998	2.451
SODSS, first half only	96.9±0.0	96.0±0.1	92.7±0.3	67.3	66.4	64.0	1.683	1.996	2.401
SODSS, second half only	96.7±0.1	95.5±0.3	92.3±0.5	67.1	65.8	62.9	1.683	2.146	2.815
SODSS without θ^2	95.4±0.0	93.5±0.3	89.3±0.3	65.6	63.6	60.3	3.071	3.515	3.811
Traj. curvature ratio $\times\theta^2$	96.9±0.1	95.8±0.1	90.7±0.6	67.5	65.6	60.3	1.666	1.782	2.031
Acceleration energy $\times\theta^2$	96.8±0.0	95.8±0.2	92.3±0.5	67.0	66.2	63.3	1.687	2.162	2.529

recovers the accumulated first moment, and the corrected score is the best criterion at every sparsity (1.76, 2.18, 2.74 at 50, 70, 80% against 1.86, 2.63, 3.41 for magnitude), still with no gradients or data at analysis time; after fine-tuning all four are within 0.02 nats at 50%. Original layer-level rule. Ranking whole layers by event count is unstable because it can remove an entire small layer: $39 \pm 40\%$ at 90% sparsity on MNIST and chance accuracy on CIFAR-10, where it prunes the output layer first.

Three-way compression. Table 4 (Appendix C) applies the budgeted keep/INT8/prune rule on MNIST: at 6.4 average bits per stored weight every informative ranking is within 0.1 points of the dense model; at 1.7 bits magnitude retains most and SODSS is second.

4.4 Ablations

Table 2 isolates the ingredients of the score. θ^2 matters most. Removing it (path-Fisher alone, i.e. Synaptic Intelligence) costs 9 points at 90% and 20 points at 95% on MNIST and is worse on the Transformer at every sparsity: the OBD form $\bar{F}_{jj}\theta_j^2$, not the path integral

by itself, is what prunes well. Regime weights have no measurable effect. With or without w_t every entry agrees to within 0.1 points. Restricting the accumulation to the first half of training changes nothing either, whereas using only the second half is worse (80% vs. 88% at 95% sparsity): the informative motion happens early, and the regime weights, which down-weight late spans, are simply too mild to matter. We report this as a negative result about regime weighting; the regimes remain useful as a description of the run. Second-order per-parameter statistics do not help under SGD. The trajectory-regressed diagonal curvature and the acceleration energy match SODSS at moderate sparsity, where θ^2 dominates the ranking, and degrade sharply at 95–98% (73% and 38% vs. 88% and 66%), in line with Remark 1. Under AdamW, however, the curvature ratio is the one trajectory statistic that survives the preconditioner (§5).

5 Discussion and limitations

What the trajectory gives and what it does not. The results separate two kinds of second-order information. Direction-level curvature (Rayleigh quotients, globally and per block) is available exactly from checkpoints and, under SGD noise, becomes accurate at moderate strides. Per-coordinate curvature is not available: the diagonal is not identifiable from the path, and criteria that assume it (our "trajectory-regressed diagonal" ablation) degrade at high sparsity. The practical consequence is a division of labour: second-order statistics for when (events, regimes, curvature diagnostics) and squared first-order statistics for which (per-parameter saliency).

Optimisers. Proposition 3 needs $v_t \propto g_t$. Under Adam the per-coordinate normalisation makes $\sum_t v_{t,j}^2$ nearly uninformative and the raw score falls well behind magnitude (Table 1). The repair uses the second-moment state that standard checkpoints already contain: $u_t = v_t(\sqrt{\hat{v}_t}+\epsilon)$ is proportional to the accumulated first moment, and $\theta_N^2 \sum_t u_{t,j}^2$ beats magnitude at every sparsity on the Transformer. Without optimiser state, the curvature ratio, invariant to a diagonal preconditioner, is the fallback.

Regimes. The detector reliably finds schedule and representational phase changes and the labels describe a run well, but their effect as importance weights is nil: when a weight moved matters less than how much it moved.

Scale and assumptions. Our experiments are proof-of-concept scale (up to 0.6M prunable weights, CPU training). The method's cost, one $O(Nd)$ pass, is not the obstacle; the missing evidence is behaviour on ImageNet-scale CNNs and billion-parameter LMs against SparseGPT and Wanda, and with structured sparsity. Proposition 1 is stated for plain gradient descent; under momentum or Adam v_t is an averaged (and preconditioned) gradient, the saliency remains defined and works empirically under AdamW, but the curvature corollaries then refer to the preconditioned Hessian. The regime rules use three fixed scale-free thresholds and are not learned.

6 Conclusion

A training trajectory is a usable substitute for gradients and Hessian-vector products in model compression. We characterised exactly which second-order information it contains (direction-level curvature, exactly; the diagonal, not at all), bounded the effect of SGD noise, and turned this into SODSS: a scale-free multi-space event detector, a temporal-regime segmentation, and a gradient-free per-parameter saliency that is an OBD saliency with a path-averaged Fisher. On SGD-trained networks it is within a point of magnitude pruning up to 95% sparsity, beats gradient-based Fisher and SNIP one-shot, and is clearly best at 98%; under AdamW the raw score fails, and the optimiser-state correction makes it the best criterion on the Transformer.

Reproducibility statement

All experiments run on CPU in under two hours in total. The analysis code (event detection, regimes, saliency, pruning and quantisation, Hessian validation utilities) and the experiment scripts for every table and figure are provided as supplementary material with fixed seeds; Appendix B lists every hyperparameter, Appendix A contains the full proofs, and the raw JSON results used to generate all tables are included.

Ethics statement

This work concerns model compression and training-dynamics analysis. It uses only standard public datasets (MNIST, CIFAR-10, the Tiny Shakespeare corpus) and involves no human subjects or personal data. Compression methods lower the cost of deploying models and are dual-use in the same broad sense as all efficiency research; we see no specific ethical concern beyond that.

AI use statement

(This section is required and does not count toward the page limit.) In this work we used a generative AI tool (Anthropic Claude, used through Claude Code) for the following tasks with required disclosure: it critiqued an earlier draft of this manuscript and identified the defects in its importance rule; it proposed the reformulated per-parameter score, the log-speed event detector, the CKA-based representation signal and the optimiser-state correction for Adam, i.e. it contributed to the conceptual framework and to formulating and refining hypotheses; it formulated and wrote the proofs of Propositions 1–3 and Corollary 1; it designed the experimental methodology and implemented all experiment code; and it produced a first interpretation of the results. We did not use generative AI to generate synthetic datasets, to clean data, or to produce any reported number: every number in the paper was produced by running the released code on the stated public datasets. Tasks with recommended disclosure for which we also used the tool: literature search and analysis, figure and table generation, code editing, and drafting and editing the text for structure and readability. We reviewed all AI-assisted work: the authors checked each proof line by line, the curvature estimator was tested against exact Hessian-vector products (Table 3), the code was read and executed by the authors, the related-work claims were checked against the cited papers, and every claim in the text was checked against the raw results. We take responsibility for the final content of this work, including text, claims and artifacts produced with the aid of generative AI.

References

Alessandro Achille, Matteo Rovere, and Stefano Soatto. Critical learning periods in deep networks. In International Conference on Learning Representations (ICLR), 2019.

Rahaf Aljundi, Francesca Babiloni, Mohamed Elhoseiny, Marcus Rohrbach, and Tinne Tuytelaars. Memory aware synapses: Learning what (not) to forget. In European Conference on Computer Vision (ECCV), 2018.

Davis Blalock, Jose Javier Gonzalez Ortiz, Jonathan Frankle, and John Guttag. What is the state of neural network pruning? In Proceedings of Machine Learning and Systems (MLSys), 2020.

Jeremy M. Cohen, Simran Kaur, Yuanzhi Li, J. Zico Kolter, and Ameet Talwalkar. Gradient descent on neural networks typically occurs at the edge of stability. In International Conference on Learning Representations (ICLR), 2021.

Zhen Dong, Zhewei Yao, Amir Gholami, Michael W. Mahoney, and Kurt Keutzer. HAWQ: Hessian aware quantization of neural networks with mixed-precision. In IEEE/CVF International Conference on Computer Vision (ICCV), 2019.

Zhen Dong, Zhewei Yao, Daiyaan Arfeen, Amir Gholami, Michael W. Mahoney, and Kurt Keutzer. HAWQ-V2: Hessian aware trace-weighted quantization of neural networks. In Advances in Neural Information Processing Systems (NeurIPS), 2020.

Jonathan Frankle and Michael Carbin. The lottery ticket hypothesis: Finding sparse, trainable neural networks. In International Conference on Learning Representations (ICLR), 2019.

Jonathan Frankle, David J. Schwab, and Ari S. Morcos. The early phase of neural network training. In International Conference on Learning Representations (ICLR), 2020.

Elias Frantar and Dan Alistarh. Optimal brain compression: A framework for accurate post-training quantization and pruning. In Advances in Neural Information Processing Systems (NeurIPS), 2022.

Elias Frantar and Dan Alistarh. SparseGPT: Massive language models can be accurately pruned in one-shot. In International Conference on Machine Learning (ICML), 2023.

Trevor Gale, Erich Elsen, and Sara Hooker. The state of sparsity in deep neural networks. In arXiv preprint arXiv:1902.09574, 2019.

Guy Gur-Ari, Daniel A. Roberts, and Ethan Dyer. Gradient descent happens in a tiny subspace. arXiv preprint arXiv:1812.04754, 2018.

Song Han, Jeff Pool, John Tran, and William J. Dally. Learning both weights and connections for efficient neural networks. In Advances in Neural Information Processing Systems (NeurIPS), 2015.

Song Han, Huizi Mao, and William J. Dally. Deep compression: Compressing deep neural networks with pruning, trained quantization and Huffman coding. In International Conference on Learning Representations (ICLR), 2016.

Babak Hassibi and David G. Stork. Second order derivatives for network pruning: Optimal brain surgeon. In Advances in Neural Information Processing Systems (NeurIPS), 1993.

Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. Distilling the knowledge in a neural network. arXiv preprint arXiv:1503.02531, 2015.

Michael F. Hutchinson. A stochastic estimator of the trace of the influence matrix for Laplacian smoothing splines. In Communications in Statistics - Simulation and Computation, volume 18, pp. 1059–1076, 1989.

Stanisław Jastrzębski, Maciej Szymczak, Stanislav Fort, Devansh Arpit, Jacek Tabor, Kyunghyun Cho, and Krzysztof Geras. The break-even point on optimization trajectories of deep neural networks. In International Conference on Learning Representations (ICLR), 2020.

Andrej Karpathy. The unreasonable effectiveness of recurrent neural networks. http://karpathy.github.io/2015/05/21/rnn-effectiveness/, 2015.

Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In International Conference on Learning Representations (ICLR), 2015.

James Kirkpatrick, Razvan Pascanu, Neil Rabinowitz, Joel Veness, Guillaume Desjardins, Andrei A. Rusu, Kieran Milan, John Quan, Tiago Ramalho, Agnieszka Grabska-Barwinska, Demis Hassabis, Claudia Clopath, Dharshan Kumaran, and Raia Hadsell. Overcoming catastrophic forgetting in neural networks. Proceedings of the National Academy of Sciences, 114(13):3521–3526, 2017.

Simon Kornblith, Mohammad Norouzi, Honglak Lee, and Geoffrey Hinton. Similarity of neural network representations revisited. In International Conference on Machine Learning (ICML), 2019.

Alex Krizhevsky. Learning multiple layers of features from tiny images. Technical report, University of Toronto, 2009.

Frederik Kunstner, Philipp Hennig, and Lukas Balles. Limitations of the empirical Fisher approximation for natural gradient descent. Advances in Neural Information Processing Systems (NeurIPS), 2019.

Yann LeCun, John S. Denker, and Sara A. Solla. Optimal brain damage. In Advances in Neural Information Processing Systems (NeurIPS), 1990.

Yann LeCun, Léon Bottou, Yoshua Bengio, and Patrick Haffner. Gradient-based learning applied to document recognition. In Proceedings of the IEEE, volume 86, pp. 2278–2324, 1998.

Namhoon Lee, Thalaiyasingam Ajanthan, and Philip H. S. Torr. SNIP: Single-shot network pruning based on connection sensitivity. In International Conference on Learning Representations (ICLR), 2019.

James Martens and Roger Grosse. Optimizing neural networks with Kronecker-factored approximate curvature. In International Conference on Machine Learning (ICML), 2015.

Joseph Mellor, Jack Turner, Amos Storkey, and Elliot J. Crowley. Neural architecture search without training. In International Conference on Machine Learning (ICML), 2021.

Pavlo Molchanov, Arun Mallya, Stephen Tyree, Iuri Frosio, and Jan Kautz. Importance estimation for neural network pruning. In IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019.

Neel Nanda, Lawrence Chan, Tom Lieberum, Jess Smith, and Jacob Steinhardt. Progress measures for grokking via mechanistic interpretability. In International Conference on Learning Representations (ICLR), 2023.

Alethea Power, Yuri Burda, Harri Edwards, Igor Babuschkin, and Vedant Misra. Grokking: Generalization beyond overfitting on small algorithmic datasets. In arXiv preprint arXiv:2201.02177, 2022.

Victor Sanh, Thomas Wolf, and Alexander M. Rush. Movement pruning: Adaptive sparsity by fine-tuning. In Advances in Neural Information Processing Systems (NeurIPS), 2020.

Sidak Pal Singh and Dan Alistarh. WoodFisher: Efficient second-order approximation for neural network compression. In Advances in Neural Information Processing Systems (NeurIPS), 2020.

Mingjie Sun, Zhuang Liu, Anna Bair, and J. Zico Kolter. A simple and effective pruning approach for large language models. In International Conference on Learning Representations (ICLR), 2024.

Swabha Swayamdipta, Roy Schwartz, Nicholas Lourie, Yizhong Wang, Hannaneh Hajishirzi, Noah A. Smith, and Yejin Choi. Dataset cartography: Mapping and diagnosing datasets with training dynamics. In Empirical Methods in Natural Language Processing (EMNLP), 2020.

Hidenori Tanaka, Daniel Kunin, Daniel L. K. Yamins, and Surya Ganguli. Pruning neural networks without any data by iteratively conserving synaptic flow. In Advances in Neural Information Processing Systems (NeurIPS), 2020.

Lucas Theis, Iryna Korshunova, Alykhan Tejani, and Ferenc Huszár. Faster gaze prediction with dense networks and Fisher pruning. In arXiv preprint arXiv:1801.05787, 2018.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In Advances in Neural Information Processing Systems (NeurIPS), 2017.

Chaoqi Wang, Guodong Zhang, and Roger Grosse. Picking winning tickets before training by preserving gradient flow. In International Conference on Learning Representations (ICLR), 2020.

Zhewei Yao, Amir Gholami, Kurt Keutzer, and Michael W. Mahoney. PyHessian: Neural networks through the lens of the Hessian. In IEEE International Conference on Big Data, 2020.

Friedemann Zenke, Ben Poole, and Surya Ganguli. Continual learning through synaptic intelligence. In International Conference on Machine Learning (ICML), 2017.

A Proofs

A.1 Proof of Proposition 1 and Corollary 1

For gradient descent $\theta_t = \theta_{t-1} - \eta \nabla L(\theta_{t-1})$ we have $v_t = -\eta \nabla L(\theta_{t-1})$ and therefore

$$a_t = v_t - v_{t-1} = -\eta\big[\nabla L(\theta_{t-1}) - \nabla L(\theta_{t-2})\big].$$

Let $\phi(s) = \nabla L(\theta_{t-2} + s\, v_{t-1})$ for $s \in [0,1]$, so that $\phi(0) = \nabla L(\theta_{t-2})$ and $\phi(1) = \nabla L(\theta_{t-1})$ because $\theta_{t-2} + v_{t-1} = \theta_{t-1}$. Since $L \in C^2$, ϕ is C^1 with $\phi'(s) = \nabla^2 L(\theta_{t-2} + s v_{t-1})\, v_{t-1}$, and the fundamental theorem of calculus gives $\phi(1) - \phi(0) = \int_0^1 \phi'(s)\, ds = \bar{H}_t v_{t-1}$, where $\bar{H}_t = \int_0^1 \nabla^2 L(\theta_{t-2} + s v_{t-1}) ds$ is symmetric as an average of symmetric matrices. Substituting proves (2).

For the corollary, $-\langle a_t, v_{t-1}\rangle / (\eta \|v_{t-1}\|^2) = v_{t-1}^\top \bar{H}_t v_{t-1} / \|v_{t-1}\|^2$, which is the Rayleigh quotient of the symmetric matrix $\bar{H}_t$ at v_{t-1} and hence lies in $[\lambda_{\min}(\bar{H}_t), \lambda_{\max}(\bar{H}_t)]$ by the Courant–Fischer theorem. For a block B, $-a_t^B/\eta = (\bar{H}_t v_{t-1})^B = \bar{H}_{BB} v^B + \bar{H}_{B\bar{B}} v^{\bar{B}}$, so $r_t^B = (v^B)^\top \bar{H}_{BB} v^B / \|v^B\|^2 + (v^B)^\top \bar{H}_{B\bar{B}} v^{\bar{B}} / \|v^B\|^2$: the block Rayleigh quotient plus a cross-block term. We report r_t^B as the block-restricted curvature along the path and verify it against $(v^B)^\top (Hv)^B / \|v^B\|^2$ computed with one Hessian-vector product in Table 3.

A.2 Proof of Proposition 2

Write the minibatch gradient at step s as $\nabla L(\theta_s) + \xi_s$ with $\mathbb{E}\xi_s = 0$, $\mathbb{E}\|\xi_s\|^2 = \sigma^2$, and ξ_s independent across s and independent of the past trajectory. Over a window of Δ steps ending at snapshot $t-1$,

$$v_{t-1} = -\eta \sum_{s \in W_{t-1}} \big(\nabla L(\theta_s) + \xi_s\big) = -\eta\big(\Delta \bar{g}_{t-1} + \Xi_{t-1}\big), \qquad \Xi_{t-1} = \sum_{s \in W_{t-1}} \xi_s,$$

with $\bar{g}_{t-1}$ the mean full-batch gradient over the window and $\mathbb{E}\|\Xi_{t-1}\|^2 = \Delta\sigma^2$. Similarly $v_t = -\eta(\Delta \bar{g}_t + \Xi_t)$, hence

$$a_t = v_t - v_{t-1} = -\eta\Delta(\bar{g}_t - \bar{g}_{t-1}) - \eta(\Xi_t - \Xi_{t-1}).$$

To first order in η the change in the window-mean gradient is $\bar{g}_t - \bar{g}_{t-1} = \bar{H}_t v_{t-1} + O(\eta^2)$ for a suitable path-averaged Hessian $\bar{H}_t$ (apply the argument of Proposition 1 to the displacement between the two window centres, which equals v_{t-1} up to $O(\eta)$). Therefore

$$-\frac{\langle a_t, v_{t-1}\rangle}{\eta\Delta} = v_{t-1}^\top \bar{H}_t v_{t-1} + \frac{1}{\Delta}\langle \Xi_t - \Xi_{t-1}, v_{t-1}\rangle + O(\eta^2).$$

Take expectations over ξ. Since Ξ_t is independent of v_{t-1} and zero-mean, $\mathbb{E}\langle \Xi_t, v_{t-1}\rangle = 0$. For the second term, $v_{t-1} = -\eta(\Delta \bar{g}_{t-1} + \Xi_{t-1})$ and $\mathbb{E}\langle \Xi_{t-1}, \bar{g}_{t-1}\rangle = 0$, so $\mathbb{E}\langle -\Xi_{t-1}, v_{t-1}\rangle = \eta\, \mathbb{E}\|\Xi_{t-1}\|^2 = \eta\Delta\sigma^2$. The noise contribution to the numerator is thus $+\eta\Delta\sigma^2/\Delta = \eta\sigma^2$, and $\mathbb{E}\|v_{t-1}\|^2 = \eta^2(\Delta^2\|\bar{g}_{t-1}\|^2 + \Delta\sigma^2)$. Replacing the expectation of the ratio by the ratio of expectations,

$$\hat{r}_t \approx \frac{v_{t-1}^\top \bar{H}_t v_{t-1}}{\|v_{t-1}\|^2} + \frac{\eta\sigma^2}{\eta^2(\Delta^2\|\bar{g}\|^2 + \Delta\sigma^2)} = \frac{v_{t-1}^\top \bar{H}_t v_{t-1}}{\|v_{t-1}\|^2} + \frac{\sigma^2}{\eta\Delta(\Delta\|\bar{g}\|^2 + \sigma^2)},$$

which is (3). The bound $\beta_\Delta \le 1/(\eta\Delta)$ follows from $\sigma^2 \le \Delta\|\bar{g}\|^2 + \sigma^2$, and the two limiting cases are immediate. The bias is positive because the noise in v_{t-1} is anti-correlated with the noise increment in a_t: an unusually large noisy step is followed, on average, by a correction of the opposite sign, which looks like positive curvature.

A.3 Proof of Proposition 3

With stride $\Delta = 1$ and SGD, $v_t = \theta_t - \theta_{t-1} = -\eta_{t-1} g_{t-1}$ where g_{t-1} is the minibatch gradient used at step $t-1$. Then $\mathrm{PF}_j = \sum_t w_t v_{t,j}^2 = \sum_t w_t \eta_{t-1}^2 g_{t-1,j}^2$. For $w \equiv 1$ and constant η, $\mathrm{PF}_j = \eta^2 \sum_t g_{t-1,j}^2 = N\eta^2 \bar{F}_{jj}$ with $\bar{F} = \frac{1}{N}\sum_t g_{t-1} g_{t-1}^\top$ the empirical Fisher averaged along the path (its diagonal). The OBD saliency for removing weight j under a diagonal Hessian approximation is $\frac{1}{2} H_{jj}\theta_j^2$; with $H \approx \bar{F}$ this is $\frac{1}{2}\bar{F}_{jj}\theta_j^2 = s_j/(2N\eta^2)$. The Synaptic-Intelligence importance is $\omega_j = \sum_t -g_{t-1,j}\,\Delta\theta_{t,j} = \sum_t -g_{t-1,j}(-\eta g_{t-1,j}) = \eta \sum_t g_{t-1,j}^2 = \mathrm{PF}_j/\eta$. The Movement score is $-\sum_t g_{t-1,j}\theta_{t-1,j} = \frac{1}{\eta}\sum_t v_{t,j}\theta_{t-1,j}$.

A.4 On Remark 1

Per-coordinate least squares of $a_{t,j}$ on $v_{t-1,j}$ over a window gives $\hat{H}_{jj} = -\sum_t a_{t,j} v_{t-1,j} / (\eta \sum_t v_{t-1,j}^2)$. Substituting (2) coordinate-wise, $-a_{t,j}/\eta = \sum_k \bar{H}_{jk} v_{t-1,k}$, so $\hat{H}_{jj} = \bar{H}_{jj} + \sum_{k\neq j} \bar{H}_{jk} \frac{\sum_t v_{t-1,k} v_{t-1,j}}{\sum_t v_{t-1,j}^2}$. The second term vanishes only if the velocity coordinates are uncorrelated over the window. Because the velocity lives predominantly in a low-dimensional top eigenspace (Gur-Ari et al., 2018), its coordinates are strongly correlated and the off-diagonal sum is of the same order as $\bar{H}_{jj}$ (and is not sign-definite), which is what Table 3 shows: Spearman correlation with the Hutchinson diagonal is ≈ 0 even for noise-free GD.

B Hyperparameters and protocol

Models and training. MNIST / LeNet-300-100 (LeCun et al., 1998): MLP 784–300–100–10 with ReLU (266,200 prunable weights), plain SGD without momentum, batch 128, 6 epochs (2,814 steps), learning rate 0.1 multiplied by 0.1 at 50% and 80% of training, snapshot stride $\Delta = 10$ (282 snapshots), probe representation: second hidden layer on 512 fixed test images. Three seeds. CIFAR-10 / CNN (Krizhevsky, 2009): three 3×3 convolutions ($32, 64, 128$ channels, each followed by ReLU and 2×2 max-pooling) and two linear layers ($2048{-}256{-}10$), ≈ 0.62M prunable weights, random horizontal flips, plain SGD, batch 128, 8 epochs, learning rate 0.05 with the same step schedule, $\Delta = 20$, probe: penultimate layer on 256 fixed test images. One seed. Character Transformer (Vaswani et al., 2017; Karpathy, 2015): 2 pre-LN blocks, $d = 128$, 4 heads, MLP width 512, context 128, vocabulary 65, tied-free output head (≈ 0.42M prunable weights including token and position embeddings and the head), AdamW (weight decay 0.01, gradient clipping 1.0), batch 32, 1,500 steps, linear warm-up over 100 steps then cosine decay from 3×10^{-3}, $\Delta = 10$, probe: final-layer hidden states of 16 fixed validation sequences (2,048 tokens), lexical space: the token-embedding matrix. Validation loss is measured on 20 fixed batches of 64 sequences from the held-out 10% of the corpus. One seed.

Event detection and regimes. Robust-z threshold $z^\star = 3.5$, minimum gap 5 snapshots, at most 8 events per space, coincidence window $\delta = 2$ snapshots, minimum span length 3; regime thresholds $\tau_\lambda = 0.002$, $\tau_\sigma = 0.05$, $\tau'_\sigma = 0.15$; regime weights $w = (1.0, 0.7, 0.5, 0.2)$ for (BR, MR, IR, CA). None of these were tuned on test data; they were fixed once on a short MNIST run with a single learning-rate drop.

Pruning protocol. Global unstructured pruning over all prunable weights at sparsity s: the sd lowest-scoring weights are zeroed and masked. One-shot numbers are measured immediately. Fine-tuned numbers are after 300 (MNIST) / 400 (CIFAR-10) masked SGD steps with momentum 0.9 and learning rate 0.01 / 0.005, or 200 masked AdamW steps at 10^{-3} (LM), on the training set. Gradient-based baselines use 8 batches of 512 (MNIST), 8 of

256 (CIFAR-10) or 8 of 32 sequences (LM) of training data at the final checkpoint: empirical Fisher $\times\theta^2$ uses $\frac{1}{n}\sum_b g_b^2\theta^2$ and SNIP uses $|\bar{g}\theta|$. The "Original SODSS rule" baseline re-implements the layer-level score $0.5 + \min(0.1 \cdot \#\text{events in layer}, 0.5)$ with per-layer event detection, ranking layers by that score and weights within a layer by magnitude.

Three-way compression. Weights ranked by s_j; top k kept in FP32, bottom p pruned, remainder fake-quantised to INT8 with a symmetric per-tensor scale $\max|w|/127$; budgets $(k, p) \in \{(0.10, 0.50), (0.05, 0.70), (0.02, 0.85)\}$, i.e. 6.4, 3.6 and 1.7 average bits per weight excluding the zero-bit accounting of pruned weights.

Curvature validation. MLP $784-32-10$ with tanh (25,408 prunable weights) on the first 5,000 MNIST training images, learning rate 0.05; full-batch GD for 600 steps or SGD (batch 128) for 3,000 steps, recorded at every step. Exact Rayleigh quotients use one Hessian-vector product of the full-batch loss on 2,000 images; $\mathrm{diag}(H)$ uses Hutchinson's estimator (Hutchinson, 1989) with 200 Rademacher probes; $\lambda_{\max}$ uses 30 power iterations. Trajectory estimates use the last 60 sub-sampled snapshots at stride Δ (window of 60Δ steps) for the per-coordinate regression and the last three sub-sampled snapshots for the Rayleigh quotient.

Compute and determinism. All runs use CPU only (8 cores); the complete set of experiments takes under two hours. The Adam-corrected Transformer row was obtained from a re-run of the same seed with optimiser-state recording; the shared criteria reproduced to within 0.01 nats (multi-threaded CPU training is not bit-deterministic).

C Algorithm and additional results

Algorithm 1 SODSS

Require: checkpoints $\theta_0..\theta_N$ (stride Δ), losses L_t, optional probe activations H_t, embeddings G_t; budget (k, p)

1: for $t = 1..N$ do
2: $v_t \leftarrow \theta_t - \theta_{t-1}$; accumulate window sums of v_t^2; record $\|v_t\|$, $\|v_t^B\|$, $1-\mathrm{CKA}(H_t, H_{t-1})$, $\|G_t - G_{t-1}\|_F$
3: end for
4: events $I_\theta, I_H, I_G \leftarrow$ robust-z + NMS on $e_t = |\Delta^2 \log s_t|$ for each speed series (Eq. 4)
5: spans and regimes $R_i \leftarrow$ rules of §3.4; $w_t \leftarrow w(R_{i(t)})$
6: $\mathrm{PF}_j \leftarrow \sum_t w_t v_{t,j}^2$; $s_j \leftarrow \theta_{N,j}^2 \mathrm{PF}_j$
7: rank by s_j: keep top k, prune bottom p, quantise the rest; optional short masked fine-tune

C.1 Pruning curves, curvature validation and three-way compression

C.2 Detected events per run

Table 5 lists the detected event indices and regime segmentation for seed 0 of each testbed. Across the three MNIST seeds the parameter-space events at the learning-rate drops are at snapshots $\{142, 142, 143\}$ and $\{226, 226, 226\}$ (true drops after snapshots 141 and 225), and the representation-space events at $\{142, 142, 145\}$ and $\{226, 226, 241\}$.

C.3 Momentum and snapshot stride

Table 6 repeats the MNIST experiment with momentum SGD and with snapshot strides of 5 and 50 instead of 10. With momentum 0.9 (where v_t is an exponentially averaged gradient) the pattern is unchanged: SODSS is 1–3 points behind magnitude at 90–95% and ahead at 98% (51.4% vs. 47.6%), and ahead of the gradient-based Fisher throughout. Coarser snapshots help rather than hurt: at $\Delta = 50$ (57 snapshots) SODSS reaches 94.9% / 91.4% at 90 / 95% sparsity against 93.9% / 88.0% at $\Delta = 10$ and 93.3% / 85.8% at $\Delta = 5$,

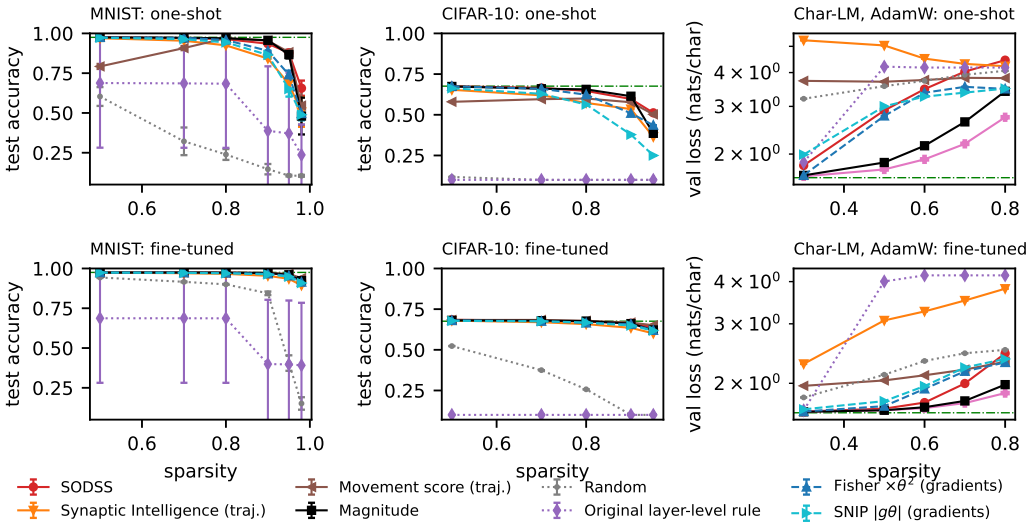

Figure 3: Sparsity–quality curves for one-shot global pruning (top) and after a short masked fine-tune (bottom). Solid: gradient-free criteria; dashed: gradient-based references; dash-dot green: uncompressed model. Error bars: std over seeds.

Table 3: Curvature along the path from the trajectory alone, mean±std over the last 10 sub-sampled snapshots. Exact Rayleigh quotient $v^\top Hv/\|v\|^2$ (one Hessian-vector product at each snapshot) versus the trajectory estimate $\hat{r}_t = -\langle a_t, v_{t-1}\rangle/(\eta\Delta\|v_{t-1}\|^2)$, for minibatch SGD at increasing snapshot stride Δ and for full-batch GD. The SGD bias is bounded by $1/(\eta\Delta)$ (Prop. 2). Right: Spearman correlation with the Hutchinson estimate of $\mathrm{diag}(H)$ for the per-parameter path-Fisher statistic and for per-coordinate regression.

setting	Δ	exact	trajectory	$1/(\eta\Delta)$	ρ(path-Fisher, diag H)	$\rho(\hat{H}_{jj}$, diag $H)$
SGD (b=128)	1	0.75±0.21	20.86±1.11	20.00	0.65	-0.14
SGD (b=128)	2	0.67±0.09	9.95±0.69	10.00	0.66	-0.12
SGD (b=128)	5	0.68±0.12	4.16±0.19	4.00	0.66	-0.24
SGD (b=128)	10	0.64±0.21	2.11±0.11	2.00	0.66	-0.28
SGD (b=128)	20	0.47±0.11	1.01±0.06	1.00	0.66	-0.32
SGD (b=128)	50	0.29±0.07	0.33±0.02	0.40	0.60	-0.03
full-batch GD	1	0.0226±0.0001	0.0223±0.0001	–	0.38	-0.02
full-batch GD	2	0.0229±0.0003	0.0226±0.0002	–	0.38	-0.03
full-batch GD	5	0.0239±0.0008	0.0233±0.0007	–	0.42	-0.02

consistent with the noise averaging of Proposition 2; magnitude and the gradient-based reference do not depend on Δ.

C.4 Where the sparsity goes

Table 7 reports the per-layer sparsity induced by global pruning at 90% on MNIST for three criteria. Magnitude and SODSS both concentrate sparsity in the 784×300 input layer, which holds 88% of the weights, but SODSS keeps more of the output layer, whose weights move the most per parameter during training.

C.5 Why the original importance rule could not prune

The importance rule of the earlier design was $\tilde{\iota}_j = \frac{1}{2} + \gamma \sum_i w(R_i)|S_i|/N + \min(\beta|I_\theta^{(j)}|, \beta_{\max})$ with $\gamma = 10^{-3}$, $\beta = 0.1$, $\beta_{\max} = 0.5$, normalised by its maximum and thresholded at 0.7 (keep) and 0.4 (prune). The regime term does not depend on j and is at most 10^{-3}, so it is a constant offset that cancels under normalisation; the score therefore reduces to the

Table 4: Three-way compression (keep FP32 / INT8 / prune) on MNIST LeNet-300-100 at fixed budgets, one-shot test accuracy (%), mean±std over 3 seeds. Budget is given as (kept fraction / pruned fraction) and the resulting average bits per weight.

Criterion	0.1/0.5 (6.4 bits)	0.05/0.7 (3.6 bits)	0.02/0.85 (1.7 bits)
SODSS	97.5±0.0	97.1±0.1	95.6±0.2
Magnitude	97.5±0.0	97.2±0.1	97.0±0.1
Random	60.4±6.0	32.2±8.7	19.4±2.8
Fisher $\times\theta^2$ (gradients)	97.5±0.0	97.2±0.2	94.0±0.1

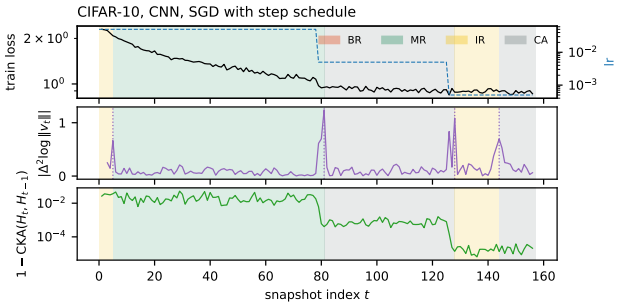

Figure 4: Events and regimes for the CIFAR-10 CNN (seed 0), same layout as Figure 2. The learning-rate drops (after snapshots 78 and 125) are detected at snapshots 81 and 128 in parameter space.

per-layer event count. Moreover $\tilde{\iota}_j \in [0.5, 1.001]$ implies $\iota_j \geq 0.4995 > 0.4$, so the prune branch is unreachable for any input. The budgeted ranking of §3.6 removes both problems; the "Original SODSS rule" baseline in Table 1 implements the layer-level ranking that the old rule effectively computed, with magnitude as the within-layer tie-breaker, so that it can be evaluated at a matched sparsity.

C.6 Sensitivity of the event detector

The detector has three parameters: the robust-z threshold $z^\star$, the minimum gap, and the maximum number of events. On MNIST, the two learning-rate drops are the two largest event scores by a factor of more than 5 over the next candidate ($e \approx 2.1$ and 1.5 against a median of 0.14 and MAD of 0.10, i.e. $z \approx 19$ and 13), so any $z^\star \in [3, 10]$ recovers them; lowering $z^\star$ below 3 adds events at the start of training and during the noisy high-learning-rate phase. The cap of 8 events was never reached in any run.

Table 5: Detected second-order events and regimes (seed 0). Snapshot indices; learning-rate drops are after snapshots 141 and 225 (MNIST) and 78 and 125 (CIFAR-10); the LM uses warm-up + cosine and has no discrete schedule events.

run	parameter-space events	representation events	regimes (start–end: label)
MNIST / LeNet-300-100	[3, 142, 226, 254]	[142, 226]	0–3: IR; 3–142: IR; 142–226: BR; 226–254: BR; 254–282: MR
CIFAR-10 / CNN	[5, 81, 128, 144]	[]	0–5: IR; 5–81: MR; 81–128: CA; 128–144: IR; 144–157: CA
Char-Transformer (AdamW)	[3, 150]	[6, 130, 141, 150] (lexical: [3, 66, 102, 111, 150])	0–3: IR; 3–6: MR; 6–130: MR; 130–141: CA; 141–150: CA; 150–151: CA

Table 6: Robustness on MNIST (seed 0, one-shot test accuracy %): plain SGD versus SGD with momentum 0.9, and snapshot strides $\Delta \in \{2, 10, 50\}$. Magnitude and the gradient-based Fisher reference are unaffected by Δ and shown for comparison.

setting	criterion	90%	95%	98%
default (SGD, $\Delta = 10$)	SODSS	93.9	88.0	60.9
default (SGD, $\Delta = 10$)	Magnitude	96.3	88.3	33.4
default (SGD, $\Delta = 10$)	Fisher $\times\theta^2$ (gradients)	89.3	71.8	44.0
SGD + momentum 0.9, $\Delta = 10$	SODSS	95.1	88.0	51.4
SGD + momentum 0.9, $\Delta = 10$	Magnitude	96.5	90.6	47.6
SGD + momentum 0.9, $\Delta = 10$	Fisher $\times\theta^2$ (gradients)	91.2	77.1	44.7
SGD, $\Delta = 5$	SODSS	93.3	85.8	57.9
SGD, $\Delta = 5$	Magnitude	96.3	88.3	33.4
SGD, $\Delta = 5$	Fisher $\times\theta^2$ (gradients)	89.3	71.8	44.0
SGD, $\Delta = 50$	SODSS	94.9	91.4	58.1
SGD, $\Delta = 50$	Magnitude	96.3	88.3	33.4
SGD, $\Delta = 50$	Fisher $\times\theta^2$ (gradients)	89.3	71.8	44.0

Table 7: Per-layer sparsity (%) induced by global pruning at 90% on MNIST LeNet-300-100, mean over 3 seeds.

Criterion	fc1.weight	fc2.weight	fc3.weight
SODSS	93.1	67.8	16.6
Magnitude	94.2	59.4	19.3
Fisher $\times\theta^2$ (gradients)	93.1	68.4	17.1