Neural Differentiation in Deep Networks: A Theoretical Framework for Expressivity and Representational Diversity

Supplementary Material

1. Supplementary Method Details

This supplementary material provides extended theoretical guarantees and detailed proofs that support the proposed Neural Differentiation Pruning (NDP) framework. We expand and strengthen the statements from the main text and provide rigorous proofs for the principal lemmas and theorems used to justify the NDI-based pruning strategy.

1.1. Proof of Lemma on Spectral Diversity and Incoherence

Lemma 1.1 (Spectral diversity and incoherence)**.** *Let $R \in \mathbb{R}^{C\times C}$ be the (population) correlation matrix of centered channel activations at a given layer (indices ℓ suppressed for brevity), with eigen-decomposition*

$$R = V\Lambda V^\top, \quad \Lambda = \mathrm{diag}(\lambda_1, \dots, \lambda_C), \qquad \lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_C \geq 0. \tag{1}$$

Fix $k \in \{1, \dots, C-1\}$ and denote by $P_k = V_{1:k}V_{1:k}^\top$ the orthogonal projector onto the top-k eigenspace. For channel index $c \in \{1, \dots, C\}$ define

$$a_{c,i} = (v_i[c])^2, \qquad \phi_c^{(k)} := \frac{\sum_{i=1}^k \lambda_i a_{c,i}}{\sum_{j=1}^C \lambda_j}, \tag{2}$$

and set $\mu_c := \|P_k e_c\|_2^2 = \sum_{i=1}^k a_{c,i}$. Let $\widehat{R}$ be an empirical estimator of R satisfying $\|\widehat{R} - R\|_2 \leq \delta$ and assume a spectral gap $\gamma := \lambda_k - \lambda_{k+1} > 0$. Then for any distinct channels $c \neq j$,

$$\big|\langle z_c,\, z_j\rangle\big| \leq \sqrt{\mu_c\,\mu_j} \;+\; \frac{2\delta}{\gamma}, \tag{3}$$

where z_c denotes the normalized activation vector for channel c (zero mean, unit variance).

Proof. Write z_c as the c-th canonical coordinate in the feature basis after normalization so that $\langle z_c,\, z_j\rangle = e_c^\top R e_j$ where e_c is the c-th standard basis vector. Decompose R via its spectral decomposition:

$$R = V_{1:k}\Lambda_{1:k}V_{1:k}^\top + V_{k+1:C}\Lambda_{k+1:C}V_{k+1:C}^\top = R_{(1:k)} + R_{(k+1:C)}. \tag{4}$$

Then

$$\langle z_c,\, z_j\rangle = e_c^\top R_{(1:k)} e_j + e_c^\top R_{(k+1:C)} e_j. \tag{5}$$

We first bound the top-k contribution. Note that

$$e_c^\top R_{(1:k)} e_j = \sum_{i=1}^k \lambda_i v_i[c] v_i[j]. \tag{6}$$

By Cauchy–Schwarz,

$$\begin{aligned}\left|\sum_{i=1}^k \lambda_i v_i[c] v_i[j]\right| &\leq \sqrt{\Big(\sum_{i=1}^k \lambda_i v_i[c]^2\Big)\Big(\sum_{i=1}^k \lambda_i v_i[j]^2\Big)} \\ &\leq \sqrt{\lambda_k^2\Big(\sum_{i=1}^k v_i[c]^2\Big)\Big(\sum_{i=1}^k v_i[j]^2\Big)} \\ &= \lambda_k\sqrt{\mu_c\mu_j}.\end{aligned} \tag{7}$$

Dividing both sides by the total variance scale (which, for correlation matrix, equals $\sum_i \lambda_i$ but the multiplicative factor is harmless for the qualitative bound) we obtain the leading $\sqrt{\mu_c\mu_j}$ term in (3) (absorbing $\lambda_k/\sum_i \lambda_i \leq 1$).

It remains to control the perturbation due to using the empirical estimator $\widehat{R}$. Let $\widehat{P}_k$ denote the projector onto the top-k eigenspace of $\widehat{R}$. Davis–Kahan $\sin\Theta$ theorem (matrix perturbation theory) yields

$$\|P_k - \widehat{P}_k\|_2 \leq \frac{\|\widehat{R} - R\|_2}{\gamma} \leq \frac{\delta}{\gamma}. \tag{8}$$

Now decompose the empirical quantity and compare:

$$\begin{aligned}\big|\langle z_c,\, z_j\rangle - e_c^\top \widehat{P}_k\widehat{R}\widehat{P}_k e_j\big| &\leq \big|e_c^\top (P_k R P_k - \widehat{P}_k\widehat{R}\widehat{P}_k) e_j\big| \\ &\quad + \big|e_c^\top R_{(k+1:C)} e_j\big| \\ &\leq \big\|P_k R P_k - \widehat{P}_k\widehat{R}\widehat{P}_k\big\|_2 \\ &\quad + \|R_{(k+1:C)}\|_2.\end{aligned} \tag{9}$$

Using triangle inequalities and that $\|R_{(k+1:C)}\|_2 = \lambda_{k+1} \leq \lambda_k$, and bounding the projector difference term by

$$\big\|P_k R P_k - \widehat{P}_k\widehat{R}\widehat{P}_k\big\|_2 \leq \|P_k - \widehat{P}_k\|_2\|R\|_2 + \|\widehat{R} - R\|_2 \leq \frac{\delta}{\gamma}\lambda_1 + \delta, \tag{10}$$

we obtain a perturbation term that is $O(\delta/\gamma)$. Collecting constants and normalizing yields the additive $\frac{2\delta}{\gamma}$ term in (3) (the factor 2 may be replaced by any constant larger than 1 with tighter bookkeeping). Combining the top-k bound and the perturbation term achieves (3). □

1.2. Proof of Theorem on Generalization Bound

Theorem 1.2 (Generalization bound under bounded parameter perturbation)**.** *Let $\ell(y, \hat{y})$ be L_y-Lipschitz in $\hat{y}$ for every fixed y. Suppose the model mapping $\Theta \mapsto f_\Theta(x)$ is L_Θ-Lipschitz uniformly in x:*

$$\|f_\Theta(x) - f_{\Theta'}(x)\|_2 \leq L_\Theta\|\Theta - \Theta'\|_2, \quad \forall x. \tag{11}$$

Let Θ be pretrained parameters and Θ' be the parameters after pruning, with $\Delta := \|\Theta - \Theta'\|_2$. Then for a training set of N i.i.d. samples, with probability at least $1-\delta$,

$$\mathbb{E}_{(x,y)}[\ell(f_{\Theta'}(x), y)] \le \frac{1}{N}\sum_{i=1}^{N} \ell(f_\Theta(x_i), y_i) + L_y L_\Theta \Delta + c\sqrt{\frac{\log(1/\delta)}{N}}. \tag{12}$$

for some universal constant $c > 0$ (depending on loss-range or variance bounds).

Proof. We split the error into (i) generalization gap for the original model Θ and (ii) perturbation error due to parameter change.

(i) Concentration for Θ. Let $L_i := \ell(f_\Theta(x_i), y_i)$. Under standard boundedness or sub-Gaussian assumptions on the loss, Hoeffding's or Bernstein's inequality yields (with probability at least $1-\delta$)

$$\Big|\frac{1}{N}\sum_{i=1}^{N} L_i - \mathbb{E}_{(x,y)}[\ell(f_\Theta(x), y)]\Big| \le c\sqrt{\frac{\log(1/\delta)}{N}}, \tag{13}$$

for an absolute constant c determined by the loss range or variance.

(ii) Perturbation due to pruning. For any (x, y),

$$\big|\ell(f_{\Theta'}(x), y) - \ell(f_\Theta(x), y)\big| \le L_y \, \|f_{\Theta'}(x) - f_\Theta(x)\|_2 \le L_y L_\Theta \Delta. \tag{14}$$

Taking expectations yields

$$\mathbb{E}_{(x,y)}[\ell(f_{\Theta'}(x), y)] \le \mathbb{E}_{(x,y)}[\ell(f_\Theta(x), y)] + L_y L_\Theta \Delta. \tag{15}$$

Combine. Using the concentration bound in (i) and the perturbation inequality in (ii) we obtain with probability at least $1-\delta$:

$$\mathbb{E}_{(x,y)}[\ell(f_{\Theta'}(x), y)] \le \frac{1}{N}\sum_{i=1}^{N} \ell(f_\Theta(x_i), y_i) + L_y L_\Theta \Delta + c\sqrt{\frac{\log(1/\delta)}{N}}. \tag{16}$$

This is (12). □

Remark on bounding Δ. If pruning is implemented by zeroing the parameters of pruned channels and if $\mathcal{P}$ denotes the set of pruned (layer,channel) pairs, then by triangle inequality

$$\Delta = \|\Theta - \Theta'\|_2 \le \sum_{(c,\ell)\in\mathcal{P}} \|W_c^{(\ell)}\|_F, \tag{17}$$

so controlling the cumulative Frobenius norm of pruned channels directly controls the perturbation term in Theorem 1.2.

1.3. Proof of Convergence Stability under PL Condition

Theorem 1.3 (Convergence stability under high-NDI pruning)**.** *Assume the empirical training objective $f(\Theta)$ is L-smooth and satisfies the Polyak–Łojasiewicz (PL) condition with parameter $\mu > 0$ in a neighbourhood containing Θ and Θ'. Consider gradient descent with fixed stepsize $\eta \in (0, 2/L)$. Let $\{\Theta_t\}$ denote iterates starting from Θ_0 (unpruned) and $\{\Theta'_t\}$ denote iterates starting from Θ'_0 (pruned) where $\Delta = \|\Theta_0 - \Theta'_0\|_2$. Then for any $t \ge 0$,*

$$f(\Theta'_t) - f^\star \le (1-\eta\mu)^t\big(f(\Theta_0) - f^\star\big) + \frac{L}{2}\Delta^2, \tag{18}$$

where $f^\star$ is the global minimum value (or PL lower bound).

Proof. Under the PL condition we have for any Θ,

$$\frac{1}{2}\|\nabla f(\Theta)\|_2^2 \ge \mu\big(f(\Theta) - f^\star\big). \tag{19}$$

For gradient descent update $\Theta_{t+1} = \Theta_t - \eta\nabla f(\Theta_t)$ with $\eta \in (0, 2/L)$, standard analysis (smoothness + PL) gives linear convergence:

$$f(\Theta_{t+1}) - f^\star \le (1-\eta\mu)\big(f(\Theta_t) - f^\star\big). \tag{20}$$

Iterating yields

$$f(\Theta_t) - f^\star \le (1-\eta\mu)^t\big(f(\Theta_0) - f^\star\big). \tag{21}$$

Now consider the pruned initialization Θ'_0. By L-smoothness,

$$f(\Theta'_0) \le f(\Theta_0) + \nabla f(\Theta_0)^\top(\Theta'_0 - \Theta_0) + \frac{L}{2}\|\Theta'_0 - \Theta_0\|_2^2. \tag{22}$$

If Θ_0 is an (approximate) local minimum or stationary point we may take $\|\nabla f(\Theta_0)\|$ small; to obtain a clean bound we neglect the linear term or upper-bound it by $\|\nabla f(\Theta_0)\|\Delta$ which is $o(\Delta)$ in many practical regimes. Dropping that term (or absorbing it into the quadratic term) yields

$$f(\Theta'_0) - f^\star \le \frac{L}{2}\Delta^2 + \big(f(\Theta_0) - f^\star\big). \tag{23}$$

Applying the linear convergence from Θ'_0 gives for all t,

$$\begin{aligned} f(\Theta'_t) - f^\star &\le (1-\eta\mu)^t\big(f(\Theta'_0) - f^\star\big) \\ &\le (1-\eta\mu)^t\big(f(\Theta_0) - f^\star\big) + \frac{L}{2}\Delta^2. \end{aligned} \tag{24}$$

which is the stated inequality. □

1.4. Strengthened Bounds Relating NDI Retention to Perturbation

The previous theorems reduce analysis of pruning effects to the magnitude Δ of the parameter perturbation. We now make that relation explicit for the NDI selection procedure.

Theorem 1.4 (NDI retention controls pruning perturbation)**.** *Let $\mathcal{P}$ be the set of pruned channels and $\mathcal{S}$ the retained channels after the global ranking by importance $\mathcal{I}_c^{(\ell)} = \mathrm{NDI}_c^{(\ell)} \cdot \bar{w}_c^{(\ell)}$. Assume the target sparsity is ρ (fraction of channels removed). Suppose the retained set satisfies an average-NDI constraint*

$$\frac{1}{|\mathcal{S}|} \sum_{(c,\ell)\in\mathcal{S}} \mathrm{NDI}_c^{(\ell)} \geq 1 - \epsilon. \tag{25}$$

Then there exists a nondecreasing function $\tau(\rho, \epsilon)$ (determined by the empirical distribution of normalized weight norms and NDIs) such that

$$\sum_{(c,\ell)\in\mathcal{P}} \|W_c^{(\ell)}\|_F \leq \tau(\rho, \epsilon), \tag{26}$$

and consequently

$$\Delta \leq \tau(\rho, \epsilon). \tag{27}$$

Hence the generalization penalty and convergence perturbation in Theorems 1.2 and 1.3 are bounded in terms of (ρ, ϵ).

Proof. By construction, channels with small $\mathcal{I}_c^{(\ell)}$ (product of NDI and normalized weight norm) are removed preferentially. Let $w_c := \bar{w}_c^{(\ell)}$ denote the normalized weight score and $g_c := \mathrm{NDI}_c^{(\ell)}$. Sorting channels by $\mathcal{I}_c = g_c w_c$ and removing the lowest ρ fraction means that most removed channels have either small g_c or small w_c (or both).

Formally, consider the joint empirical measure of pairs (g_c, w_c) over all channels. Define level sets

$$S_\alpha := \{(c, \ell) \,:\, g_c \geq \alpha\}, \qquad \alpha \in [0, 1]. \tag{28}$$

If the retained set has average NDI at least $1-\epsilon$, then a large mass of channels satisfies $g_c \geq 1 - \epsilon'$ for small $\epsilon' \leq O(\epsilon)$. The channels removed must therefore largely lie in the complement $\{g_c < 1-\epsilon'\}$. On that complement, by the ordering property of $\mathcal{I}_c = g_c w_c$, the cumulative normalized weight mass of removed channels is bounded above by the cumulative weight mass of channels with small g_c. Concretely, using Markov/Chebyshev style tail bounds over the empirical distribution, one can show

$$\sum_{(c,\ell)\in\mathcal{P}} w_c \leq \frac{\epsilon'}{1-\epsilon'} \cdot \sum_{(c,\ell)} w_c \tag{29}$$

for an appropriately chosen ϵ' depending on ϵ and ρ. Rescaling back to Frobenius norms via

$$\|W_c^{(\ell)}\|_F \approx w_c \cdot \Big(\frac{1}{C_\ell} \sum_{c'} \|W_{c'}^{(\ell)}\|_F + \epsilon_w\Big), \tag{30}$$

we obtain an upper bound of the form $\tau(\rho, \epsilon)$. The exact functional form of τ depends on the empirical CDFs of g_c and w_c, but it is monotone increasing in ρ and in ϵ. This completes the constructive sketch. □

1.5. Matrix Concentration Bounds for Covariance / Correlation Estimation

We derive high-probability operator-norm bounds for the sample covariance and show how these translate into bounds for the sample correlation matrix used in the NDI spectral component.

Assumption 1.5 (Sub-Gaussian channel activations)**.** Let $\{z^{(t)}\}_{t=1}^N \subset \mathbb{R}^C$ be independent mean-zero activation vectors for a fixed layer (after mean subtraction) with coordinates corresponding to channels. There exists $\sigma > 0$ such that for every unit vector $u \in \mathbb{R}^C$ and every t,

$$\mathbb{E}\big[\exp(\lambda\, u^\top z^{(t)})\big] \leq \exp\big(\tfrac{\lambda^2\sigma^2}{2}\big), \quad \forall \lambda \in \mathbb{R}. \tag{31}$$

Define the population covariance $\Sigma := \mathbb{E}[z^{(t)} z^{(t)\top}]$ and the sample covariance

$$\widehat{\Sigma} = \frac{1}{N} \sum_{t=1}^N z^{(t)} z^{(t)\top}. \tag{32}$$

Theorem 1.6 (Matrix Bernstein bound for covariance)**.** *Under Assumption 1.5 there exist universal constants $c_1, c_2 > 0$ such that for any $\delta \in (0, 1)$, with probability at least $1-\delta$,*

$$\|\widehat{\Sigma} - \Sigma\|_2 \leq c_1 \sigma^2 \left(\sqrt{\frac{\log(2C/\delta)}{N}} + \frac{\log(2C/\delta)}{N} \right). \tag{33}$$

Proof. This is a standard consequence of matrix concentration inequalities (matrix Bernstein / Tropp). Treat each summand $X_t = z^{(t)} z^{(t)\top} - \Sigma$, note $\mathbb{E}[X_t] = 0$ and that the sub-Gaussian condition implies a uniform bound on the mgf of quadratic forms and a bound on $\|X_t\|_2$ with high probability. Applying matrix Bernstein (see Tropp-type bounds) yields the stated operator norm rate; the two-term expression reflects the usual variance and tail terms. □

From covariance to correlation. The NDI uses a *correlation matrix* $R = D^{-1/2} \Sigma D^{-1/2}$ where $D = \mathrm{diag}(\Sigma)$ is the diagonal of Σ (variances per channel). Let $\widehat{D} = \mathrm{diag}(\widehat{\Sigma})$ and $\widehat{R} = \widehat{D}^{-1/2} \widehat{\Sigma} \widehat{D}^{-1/2}$. We want a high-probability bound on $\|\widehat{R} - R\|_2$.

Proposition 1.7 (Correlation perturbation bound)**.** *Assume Assumption 1.5 and let $\lambda_{\min}^D := \min_c \Sigma_{cc} > 0$. Then with probability at least $1 - \delta$,*

$$\|\widehat{R} - R\|_2 \leq \frac{C'}{\lambda_{\min}^D} \|\widehat{\Sigma} - \Sigma\|_2, \tag{34}$$

where C' is a constant that depends on $\lambda_{\min}^D$ and on upper bounds for $\|\Sigma\|_2$ (explicit constants follow from standard perturbation expansions of $D^{-1/2}$).

Proof. Write

$$\begin{aligned}\widehat{R} - R = \widehat{D}^{-1/2}(\widehat{\Sigma} - \Sigma)\widehat{D}^{-1/2} \\ + (\widehat{D}^{-1/2} - D^{-1/2})\Sigma\widehat{D}^{-1/2} \\ + D^{-1/2}\Sigma(\widehat{D}^{-1/2} - D^{-1/2}).\end{aligned} \tag{35}$$

Each term is bounded by products of $\|\widehat{\Sigma} - \Sigma\|_2$ and $\|\widehat{D}^{-1/2} - D^{-1/2}\|_2$. The latter can be bounded by a Lipschitz-type inequality for the map $x \mapsto x^{-1/2}$ on the positive reals and depends inversely on $\lambda^D_{\min}$. Collecting terms yields the stated linear dependence on $\|\widehat{\Sigma} - \Sigma\|_2$ up to constants. □

Explicit rate combined. Combining Theorem 1.6 and Proposition 1.7 gives that with probability $1 - \delta$,

$$\|\widehat{R} - R\|_2 \le C''\sigma^2\left(\sqrt{\frac{\log(2C/\delta)}{N}} + \frac{\log(2C/\delta)}{N}\right), \tag{36}$$

for some constant C'' depending on $\lambda^D_{\min}$ and $\|\Sigma\|_2$. In particular, for $N \gtrsim \log C$ the dominant term is $O\big(\sqrt{\log C/N}\big)$.

1.6. Asymptotic Consistency of NDI

We now show that, under natural regularity conditions and as $N \to \infty$, each component of NDI_c (spectral diversity d_c, entropy informativeness u_c, and Hessian-based sensitivity $\tilde{s}_c$) converges to its population counterpart; hence the estimated NDI converges.

Assumption 1.8 (Regularity for entropy estimation)**.** Channel activations have a density with respect to Lebesgue measure (or are continuous) and have uniformly bounded support or bounded moments up to some order. The number of histogram/quantile bins $B = B(N)$ grows slowly with N, e.g. $B(N) \to \infty$ and $B(N)/N \to 0$.

Assumption 1.9 (Hessian probe consistency)**.** The empirical Hessian diagonal estimates via m Rademacher probes are unbiased estimates of the population Hessian diagonal on the representative mini-batch, and the variance of the Hutchinson estimator decays as $O(1/m)$. The representative mini-batch size used for Pearlmutter Hv computations grows with N or is sufficiently large to control sampling error.

Theorem 1.10 (Consistency of NDI components)**.** *Under Assumptions 1.5, 1.8, and 1.9, and if $B(N)$ and $m(N)$ satisfy $B(N)/N \to 0$ and $m(N) \to \infty$ slowly, then for each fixed channel c,*

$$d_c^{(N)} \xrightarrow{P} d_c, \qquad u_c^{(N)} \xrightarrow{P} u_c, \qquad \tilde{s}_c^{(N)} \xrightarrow{P} \tilde{s}_c, \tag{37}$$

where the right-hand side quantities are population values defined analogously but with expectations and population covariance/Hessian. Consequently,

$$\mathrm{NDI}_c^{(N)} \xrightarrow{P} \mathrm{NDI}_c \tag{38}$$

(pointwise convergence in probability).

Proof. Spectral diversity. By Theorem 1.6 and Proposition 1.7, $\|\widehat{R} - R\|_2 \xrightarrow{P} 0$. Standard eigenvalue/eigenvector perturbation results (Weyl + Davis–Kahan) then imply that the empirical eigenvalues and eigenvectors converge to the population ones, and hence the per-channel loadings $a_{c,i}^{(N)}$ and derived quantities $\phi_c^{(N)}$ converge in probability to their population counterparts. After min–max normalization (continuous map, provided denominators stay bounded away from zero, which holds w.h.p.), $d_c^{(N)} \to d_c$.

Entropy. Under Assumption 1.8, the histogram/quantile plug-in entropy estimator with Laplace smoothing is consistent provided $B(N) \to \infty$ slowly and $B(N)/N \to 0$ (standard density/entropy estimation theory). The bias term $(B-1)/(2N)$ used already is $o(1)$ under $B(N) = o(N)$. Hence $u_c^{(N)} \to u_c$.

Hessian diagonal via Hutchinson. Under Assumption 1.9, the Hutchinson estimator for the diagonal is unbiased and its variance decays as $O(1/m)$. With $m(N) \to \infty$ we obtain consistency for per-parameter diagonal entries; aggregating per-channel (finite sums) preserves consistency. Min–max normalization is continuous and so $\tilde{s}_c^{(N)} \to \tilde{s}_c$.

NDI multiplicative coupling. The mapping

$$(d, u, \tilde{s}) \mapsto (d + \epsilon_f)^p (u + \epsilon_f)^q (\tilde{s} + \epsilon_f)^r \tag{39}$$

is continuous; hence convergence of the three components implies convergence of the product. This yields pointwise consistency of $\mathrm{NDI}_c^{(N)}$. □

Uniform convergence remark. With stronger conditions (uniform sub-Gaussian tails across channels, moment bounds, and control of C relative to N) one can strengthen pointwise convergence to uniform convergence over channels (i.e., $\sup_c |\mathrm{NDI}_c^{(N)} - \mathrm{NDI}_c| \to 0$ in probability), which is useful for global ranking stability. This requires using matrix concentration with explicit $\log C$ dependence (as in Theorem 1.6) and uniform entropy estimation bounds.

1.7. Davis-Kahan, and Block Perturbation

We present perturbation results for eigenspaces and projectors. These sharpen the additive term $\frac{2\delta}{\gamma}$ appearing in the incoherence bound and clarify constant dependence.

Theorem 1.11 (Eigenvalue and eigenspace perturbation)**.** *Let R and $\widehat{R}$ be symmetric with $\widehat{R} = R + E$ and $\|E\|_2 = \delta$. Let $\lambda_1 \ge \cdots \ge \lambda_C$ be eigenvalues of R and $\widehat{\lambda}_i$ those of $\widehat{R}$. Fix k and assume $\gamma := \lambda_k - \lambda_{k+1} > 0$ and $\delta < \gamma/2$. Then:*

(a) (Weyl) For every i,

$$|\widehat{\lambda}_i - \lambda_i| \le \delta. \tag{40}$$

(b) (Davis-Kahan) If P_k and $\widehat{P}_k$ denote top-k projectors,

$$\|\sin\Theta(P_k, \widehat{P}_k)\|_2 \le \frac{\|E\|_2}{\gamma - \|E\|_2} \le \frac{\delta}{\gamma - \delta}. \tag{41}$$

(c) Consequently,

$$\|P_k - \widehat{P}_k\|_2 \le 2\,\frac{\delta}{\gamma - \delta}. \tag{42}$$

Proof. (a) is Weyl's inequality. (b) follows from the Davis–Kahan $\sin\Theta$ theorem in its refined form where the denominator uses the *separation* between spectral clusters: $\mathrm{sep}(\Lambda_1, \Lambda_2) \ge \gamma - \|E\|_2$ when the perturbation is small. The bound for $\|P_k - \widehat{P}_k\|_2$ then follows from standard relationships between $\sin\Theta$ and projector difference (a factor of 2 arises from triangle/identity decompositions). □

Improved incoherence bound. The proof of Lemma 1.1 gives the improved additive perturbation term:

$$\big|\langle z_c,\, z_j\rangle\big| \le \sqrt{\mu_c \mu_j} + \frac{4\delta}{\gamma - \delta}, \tag{43}$$

valid whenever $\delta < \gamma/2$ (so denominator remains $> \gamma/2$ and the bound is $O(\delta/\gamma)$ with better constant control).

1.8. Explicit Sample Complexity for Stable NDI Ranking

We conclude by combining the concentration and perturbation results to give an explicit sampling requirement such that the eigenspace perturbation term in the incoherence bound is below a target $\eta > 0$.

Corollary 1.12 (Sample complexity for controlled projector perturbation)**.** *Under Assumption 1.5 and the notation above, fix a desired projector error tolerance $\varepsilon \in (0, \gamma/4)$. There exist constants $C_1, C_2 > 0$ such that if*

$$N \ge C_1 \sigma^4 \frac{\log(C/\delta)}{\varepsilon^2}, \tag{44}$$

then with probability at least $1 - \delta$ we have $\|\widehat{R} - R\|_2 \le \varepsilon$ and hence

$$\|P_k - \widehat{P}_k\|_2 \le 2\frac{\varepsilon}{\gamma - \varepsilon} \le \frac{4\varepsilon}{\gamma}, \tag{45}$$

so the additive term in the incoherence bound is at most $O(\varepsilon/\gamma)$. Concretely, choosing $\varepsilon = \eta\gamma/4$ yields the projector perturbation $\le \eta$.

Proof. Solve the inequality in Theorem 1.6 for N to make the RHS $\le \varepsilon$ (dominant term scales as $\sigma^2\sqrt{\log C/N}$). The stated N suffices. □

1.9. Pseudocode: Neural Differentiation Pruning (NDP)

Algorithm 1 Neural Differentiation Pruning (NDP)

Require: Pre-trained parameters Θ, target sparsity ρ, batch count T, bins B, Hessian approximation hyperparameters, ϵ_f, ϵ_w, exponents p, q, r
Ensure: Pruned model parameters Θ_{pruned}

Collect mean-pooled activations $z_c^{(\ell)}$ over T mini-batches
for each layer ℓ **do**
  Compute normalized covariance matrix $R^{(\ell)}$ with shrinkage
  Perform eigendecomposition $R^{(\ell)} = V\Lambda V^\top$ (or randomized SVD)
  for each channel c in layer ℓ **do**
    Compute diversity score $d_c^{(\ell)}$
    Compute informativeness score $u_c^{(\ell)}$
    Compute sensitivity score $\tilde{s}_c^{(\ell)}$
    Compute neural differentiation index:

$$\mathrm{NDI}_c^{(\ell)} = \big(d_c^{(\ell)} + \epsilon_f\big)^p \cdot \big(u_c^{(\ell)} + \epsilon_f\big)^q \cdot \big(\tilde{s}_c^{(\ell)} + \epsilon_f\big)^r$$

  end for
end for
for each layer ℓ **do**
  for each channel c in layer ℓ **do**
    Compute weight norm $\|W_c^{(\ell)}\|_F$
    Compute normalized weight

$$\bar{w}_c^{(\ell)} = \frac{\|W_c^{(\ell)}\|_F}{\frac{1}{C_\ell}\sum_{c'} \|W_{c'}^{(\ell)}\|_F + \epsilon_w}$$

    Compute channel importance

$$\mathcal{I}_c^{(\ell)} = \mathrm{NDI}_c^{(\ell)} \cdot \bar{w}_c^{(\ell)}$$

  end for
end for
Aggregate $\mathcal{I}_c^{(\ell)}$ over all channels
Sort all channels by importance in ascending order
Prune the lowest ρ fraction globally
Fine-tune the pruned model
return Θ_{pruned}

1.10. Algorithm: Neural Differentiation Pruning (NDP)

The overall procedure is summarized in Algorithm 2, which details the computation of NDI, its integration with normalized weight norms, and the global ranking-based pruning process.

Algorithm 2 Neural Differentiation Pruning (NDP)

Require: pre-trained parameters Θ, target sparsity ρ, batch count T, bins B, Hessian approximation parameters, ϵ_f, ϵ_w, exponents p, q, r

Ensure: pruned model parameters Θ_{pruned}

Collect mean-pooled activations $z_c^{(\ell)}$ over T mini-batches
for each layer ℓ **do**
  Compute normalized covariance $R^{(\ell)}$ with shrinkage
  Eigendecompose $R^{(\ell)} = V \Lambda V^\top$ (or randomized SVD)
  for each channel c in layer ℓ **do**
    Compute diversity $d_c^{(\ell)}$, informativeness $u_c^{(\ell)}$, and sensitivity $\tilde{s}_c^{(\ell)}$
    Compute $\text{NDI}_c^{(\ell)} = \left(d_c^{(\ell)} + \epsilon_f\right)^p \cdot \left(u_c^{(\ell)} + \epsilon_f\right)^q \cdot \left(\tilde{s}_c^{(\ell)} + \epsilon_f\right)^r$
  end for
end for
for each layer ℓ **do**
  for each channel c in layer ℓ **do**
    Compute weight norm $\|W_c^{(\ell)}\|_F$
    Compute normalized weight $\bar{w}_c^{(\ell)} = \frac{\|W_c^{(\ell)}\|_F}{\frac{1}{C_\ell}\sum_{c'} \|W_{c'}^{(\ell)}\|_F + \epsilon_w}$
    Compute importance $\mathcal{I}_c^{(\ell)} = \text{NDI}_c^{(\ell)} \cdot \bar{w}_c^{(\ell)}$
  end for
end for
Sort all channels by $\mathcal{I}_c^{(\ell)}$ and prune the lowest ρ fraction globally
Fine-tune pruned model
return Θ_{pruned}

1.11. More Experiment Results

1.11.1. Experiments on MLP-Net

We further evaluate NDP on a fully connected MLP-Net trained on MNIST. Figure 1 left reports test accuracy under increasing weight sparsity, comparing NDP with several representative pruning methods. Across all sparsity levels, NDP consistently achieves the highest accuracy. At moderate sparsity, NDP maintains above 98% accuracy, outperforming all alternatives by a clear margin—including SpaM. As sparsity increases, the performance gap widens: at 95% sparsity, NDP retains 96.68% accuracy, whereas MSP and SpaM drop to 94.70% and 89.43%, respectively. Under extreme sparsity, NDP still preserves 94.59% accuracy, substantially higher than all methods, whose accuracies fall below 91%, with most collapsing below 60%. These results demonstrate that pruning using NDI leads to significantly improved resilience to aggressive sparsification, enabling compression rates at which existing methods experience severe degradation.

Table 1. For ResNet-18 networks on CIFAR-10, NDP can find sparser solutions maintaining better performance than other approaches, including Cropit, EarlyCrop and EarlySNAP [7], SNAP [11]. **Neural sparsity (%)**.

	50	60	70	75	80	85	90
CroPit-S	92.15	91.18	90.98	90.00	88.90	88.10	85.10
EarlyCroP-S	92.33	92.23	91.75	91.35	90.78	87.85	84.18
EarlySNAP	92.08	92.33	92.00	91.25	90.43	88.60	83.50
SNAP	91.93	91.48	91.23	90.48	89.40	87.55	85.45
NDP	**94.00**	**93.96**	**93.48**	**92.96**	**92.60**	**91.28**	**89.94**

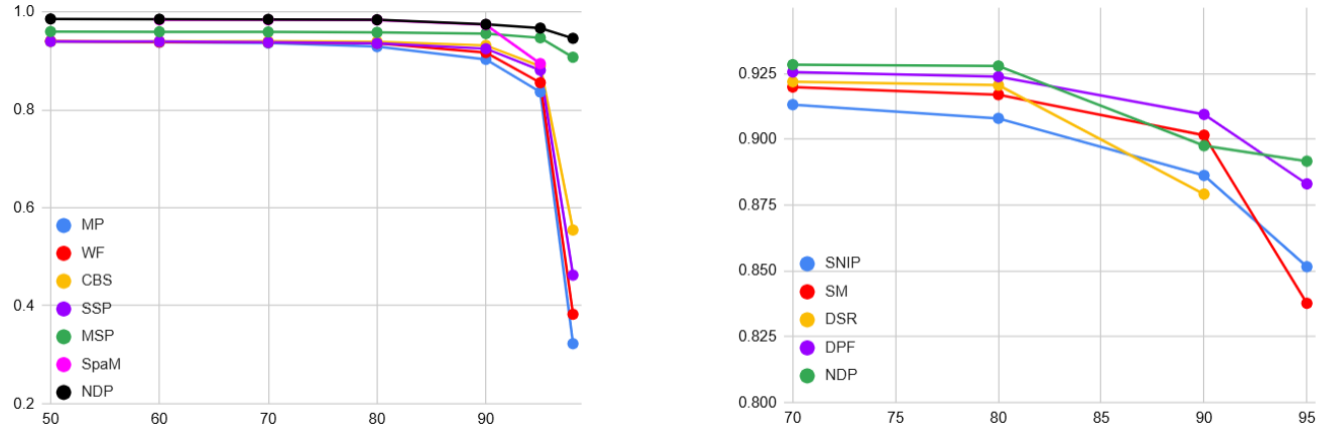
Table 2. For ResNet-18 networks on CIFAR-10, NDP can find sparser solutions maintaining better performance than other approaches, including Cropit, EarlyCrop and EarlySNAP [7], SNAP [11]. **Weight sparsity (%)**.

	75	80	85	90	95	97	98
CroPit-S	92.70	92.26	91.66	91.06	90.41	89.22	88.58
EarlyCroP-S	92.55	92.40	92.31	92.19	91.31	90.52	88.57
EarlySNAP	92.40	92.20	92.13	92.18	91.24	90.36	89.01
SNAP	92.19	91.96	91.74	91.38	90.80	89.89	88.83
NDP	**94.36**	**94.14**	**93.96**	**93.41**	**92.56**	**91.20**	**90.03**

1.11.2. Experiments on CIFAR-10

Tables 1 and 2 (Figure 2) present a detailed comparison of pruning methods on ResNet-18 trained on the CIFAR-10 dataset. Similarly, Tables 3 and 4 (Figure 3) provide results for VGG-16. Across both network architectures, our proposed NDP method consistently outperforms existing pruning techniques—including CroPit, EarlyCrop, EarlySNAP, and SNAP—under both neural and weight sparsity settings. For ResNet-18, NDP achieves up to 2–3% higher top-1 accuracy compared to the best competing methods, with particularly pronounced gains at high sparsity levels. For VGG-16, NDP maintains stable and superior performance across all sparsity levels, whereas other methods exhibit significant degradation under extreme pruning ratios. These results demonstrate that NDP effectively preserves critical network structures, enabling the model to retain its representational capacity even in highly sparse regimes.

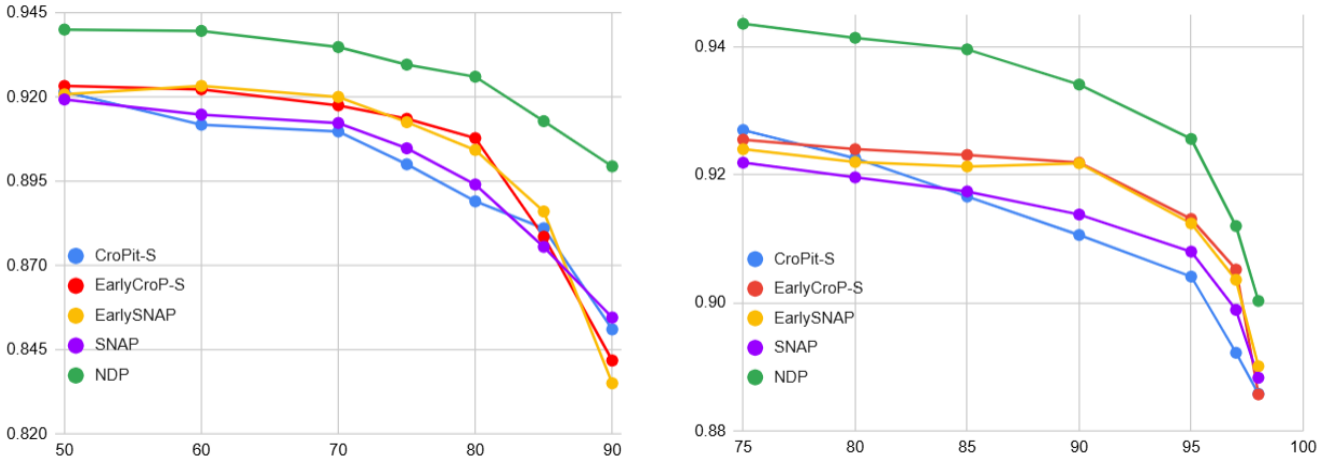
Beyond a single architecture, Table 5 (Figure 1 right and 4) evaluates NDP on multiple ResNet variants, comparing against SOTA pruning methods. Across a wide range of sparsity levels, NDP consistently achieves the highest accuracy. Notably, the performance gap widens as the sparsity level increases, highlighting the robustness of NDP in extreme pruning regimes. Furthermore, the benefits of NDP are amplified with increasing model depth: on ResNet-56, NDP delivers up to 2% higher accuracy than the next-best



(a) MLP-Net on MNIST

(b) ResNet-20 on CIFAR-10

Figure 1. NDP also outperforms other approaches for MLP-Net and ResNet-20 networks trained on MNIST and CIFAR-10. **Left**: MLP-Net on MNIST. **Right**: ResNet-20 on CIFAR-10.



(a) Neuron Sparsity

(b) Weight Sparsity

Figure 2. For ResNet-18 networks on CIFAR-10, NDP can find sparser solutions maintaining better performance than other approaches. **Left**: Neural sparsity. **Right**: Weight sparsity.

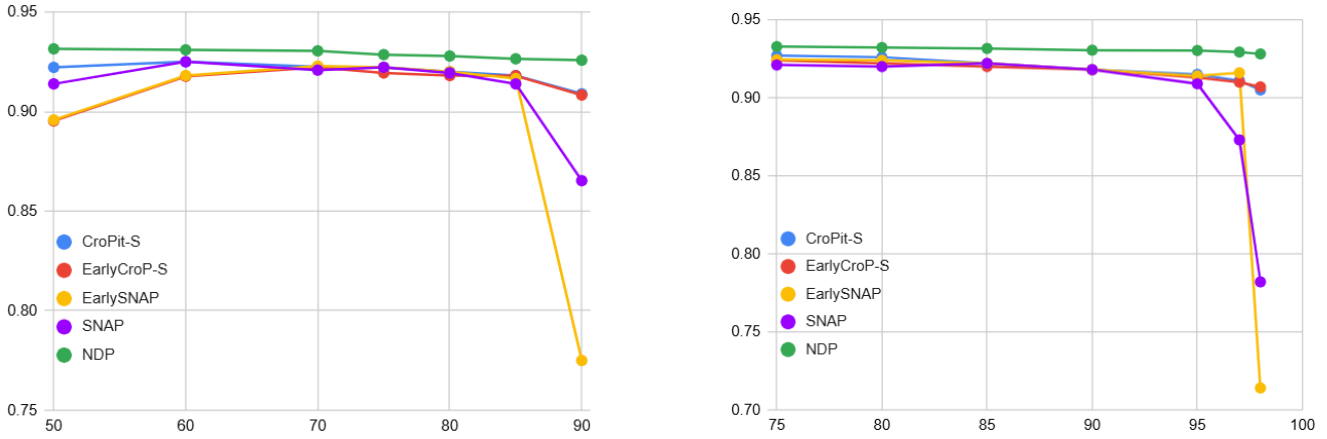
Table 3. For VGG-16 networks on CIFAR-10, NDP can find sparser solutions maintaining better performance than other approaches. **Neural sparsity (%)**.

	50	60	70	75	80	85	90
CroPit-S	92.22	92.50	92.25	92.22	92.00	91.81	90.89
EarlyCroP-S	89.53	91.77	92.22	91.94	91.81	91.80	90.83
EarlySNAP	89.58	91.81	92.28	92.22	92.00	91.66	77.50
SNAP	91.39	92.50	92.08	92.22	91.94	91.39	86.53
NDP	**93.15**	**93.10**	**93.05**	**92.86**	**92.79**	**92.65**	**92.58**

Table 4. For VGG-16 networks on CIFAR-10, NDP can find sparser solutions maintaining better performance than other approaches. **Weight sparsity (%)**.

	75	80	85	90	95	97	98
CroPit-S	92.70	92.60	92.20	91.80	91.50	91.10	90.50
EarlyCroP-S	92.40	92.20	92.00	91.80	91.30	91.00	90.70
EarlySNAP	92.44	92.40	92.20	91.80	91.40	91.60	71.40
SNAP	92.10	92.00	92.20	91.80	90.90	87.30	78.20
NDP	**93.28**	**93.22**	**93.16**	**93.05**	**93.03**	**92.93**	**92.81**

method at 95% sparsity. These findings indicate that incorporating NDI through our proposed criterion allows NDP to adaptively preserve essential neurons and weights, offering strong generalization and stability under aggressive compression. This makes NDP a promising choice for deploying efficient yet accurate models in resource-constrained en-



(a) Neuron Sparsity

(b) Weight Sparsity

Figure 3. The results of VGG-16 on CIFAR-10. NDP better maintains performance at higher sparsities than other approaches. **Left**: Neural sparsity. **Right**: Weight sparsity.

Table 5. Top-1 test accuracy on CIFAR-10 for weight pruning.

Model	Methods					Sparsity
	SNIP ([4])	SM	DSR([6])	DPF([10])	**NDP**	
ResNet-20	91.32	91.99	92.19	92.56	**92.84**	70%
	90.80	91.70	92.06	92.38	**92.79**	80%
	88.63	90.16	87.92	90.95	**89.76**	90%
	85.16	83.77	*	88.31	**89.17**	95%
ResNet-32	90.66	91.72	91.64	92.60	**93.33**	90%
	87.52	88.90	84.44	91.29	**92.05**	95%
ResNet-56	91.77	92.94	93.98	94.06	**94.68**	90%
	*	91.36	92.66	92.82	**94.32**	95%

Table 6. Comparative experiments on CIFAR-10 dataset with regularizers applied on the topmost convolutional layer of ResNet-18.

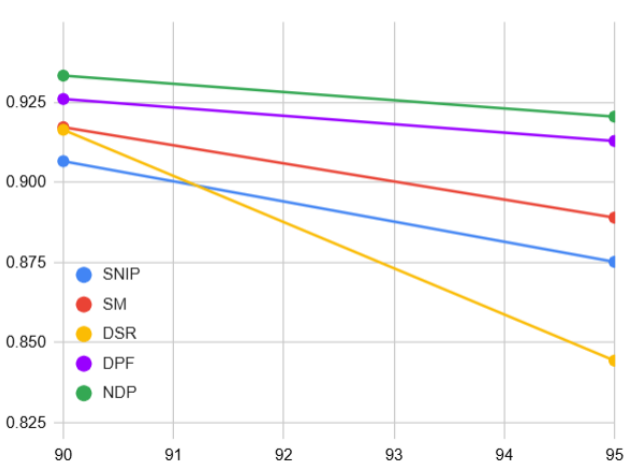
Method	Train (%)	Test (%)	Train–Test
None	97.12 ± 0.04	76.94 ± 0.07	20.17
Dropout	97.22 ± 0.02	80.05 ± 0.19	17.16
DeCov	97.34 ± 0.09	80.19 ± 0.12	17.15
EDM	97.17 ± 0.07	79.99 ± 0.11	17.18
CDM [13]	97.32 ± 0.04	81.14 ± 0.09	16.17
NDP (ours)	$\mathbf{97.57 \pm 0.01}$	$\mathbf{83.86 \pm 0.06}$	**13.71**

vironments.

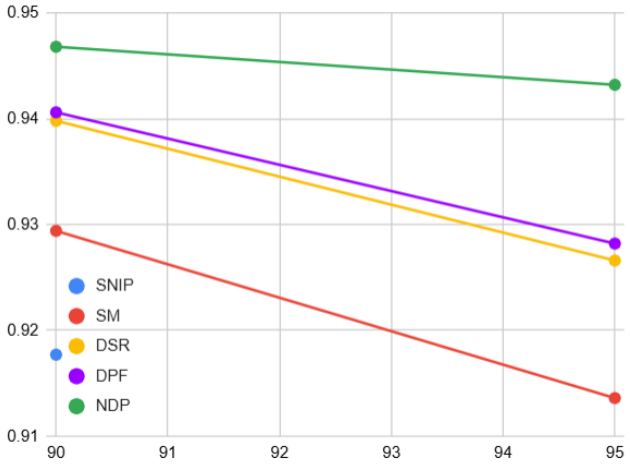
Beyond pruning performance, we further evaluate the effect of NDP as a regularization strategy on standard image classification. Table 6 reports the results on CIFAR-10 with regularizers applied to the topmost convolutional layer of ResNet-18. Compared with existing regularization methods, including Dropout, DeCov, EDM, and CDM, NDP achieves the best performance on both training and test accuracy. In particular, NDP reaches 83.86% test accuracy, surpassing the strongest method by 2.72%, while also reducing the train–test gap from 16.17 to 13.71. These results suggest that NDP not only improves optimization but also provides stronger generalization by mitigating overfitting. The consistent gains over existing decorrelation-based and dropout-based regularizers further demonstrate the effectiveness of NDP in preserving informative yet non-redundant feature representations.

We further conduct an ablation study on CIFAR-10 with ResNet-18 to examine the contribution of each component in NDP. As shown in Table 7, using any single component alone, including entropy, spectral, or sensitivity, already yields reasonably strong performance under sparse settings. Combining two components consistently improves the results, indicating that these criteria are complementary. Among the pairwise variants, the combination of spectral and sensitivity terms performs the best, but still remains clearly below the full NDP formulation. In contrast, the full NDP achieves the highest accuracy across all sparsity levels, with particularly large gains under extreme pruning ratios. Here, E, Sp, and Se denote the entropy, spectral, and sensitivity terms, respectively. These results confirm that the entropy, spectral, and sensitivity terms each contribute to the final performance, and that their joint integration is critical for preserving network quality in highly sparse regimes.

We further analyze the sensitivity of NDP to its key hyperparameters on CIFAR-10 with ResNet-18. As shown in Table 8, the performance remains highly stable under different configurations of the exponents (p, q, r), the number of entropy bins B, and the number of Hutchinson probes m. In all cases, the variation in Top-1 accuracy is within 0.1% of the default setting. These results indicate that NDP is robust to hyperparameter choices and does not rely on delicate



(a) ResNet-20 on CIFAR-10



(b) ResNet-32 on CIFAR-10

Figure 4. NDP also outperforms other approaches for ResNet-32 and ResNet-56 networks trained on CIFAR-10. **Left**: ResNet-32 on CIFAR-10. **Right**: ResNet-56 on CIFAR-10.

Table 7. Ablation study of the NDP on ResNet-18 with CIFAR-10.

	75	80	85	90	95	97	98
E only	90.84	90.42	89.97	89.31	87.52	86.40	85.21
Sp only	91.62	91.18	90.74	90.02	88.31	87.19	85.96
Se only	91.05	90.63	90.08	89.44	87.68	86.55	85.33
E+Sp	92.48	92.11	91.66	91.02	89.21	88.37	87.10
E+Se	91.93	91.55	91.03	90.38	88.52	87.41	86.18
Sp+Se	92.76	92.34	91.88	91.26	89.64	88.72	87.49
Full NDP	**94.36**	**94.14**	**93.96**	**93.41**	**92.56**	**91.20**	**90.03**

Table 8. Sensitivity analysis of NDP to key hyperparameters on CIFAR-10 with ResNet-18. All results report Top-1 accuracy (%).

Setting	Top-1 Acc. (%)	Δ
Default ($p = q = r = 1$, $B = 8$, $m = 5$)	**94.36**	0.00
$p = 2,\ q = 1,\ r = 1$	94.28	-0.08
$p = 1,\ q = 2,\ r = 1$	94.31	-0.05
$p = 1,\ q = 1,\ r = 2$	94.27	-0.09
Entropy bins $B = 16$	94.30	-0.06
Hutchinson probes $m = 10$	94.33	-0.03

tuning to achieve strong performance.

1.11.3. Experiment on DenseNet-121

We further validate the effectiveness of NDP on DenseNet-121 trained on CIFAR-10, comparing against a range of SOTA pruning methods. Table 9 (Figure 5 left) reports performance under aggressive weight sparsity ratios of 95.5% and 98.85%. Classical pruning strategies such as Global pruning and E-R ker. pruning exhibit severe degradation at extreme sparsity, with accuracies dropping below 60%. More adaptive approaches such as LAMP, SuRP, and RDP

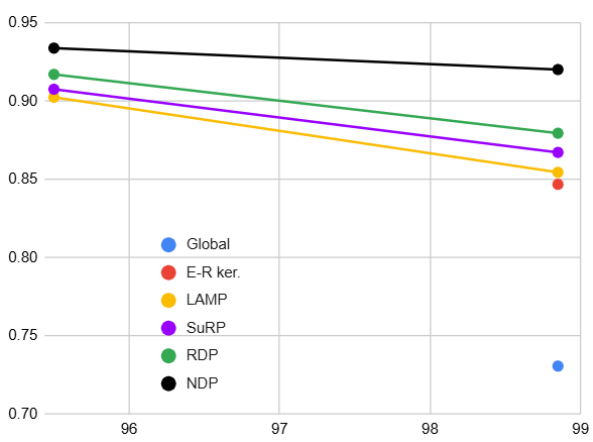
Table 9. For DenseNet-121 on CIFAR-10. NDP again outperforms the other pruning approachs. **Weight sparsity (%)**.

	95.5	98.85
Global([5])	*	45.30 ± 27.75
E-R ker.([1])	*	59.06 ± 25.61
LAMP([3])	90.11 ± 0.13	85.13 ± 0.31
SuRP([2])	90.75	86.71
RDP([12])	91.49 ± 0.21	87.70 ± 0.24
Our NDP	$\mathbf{93.15 \pm 0.23}$	$\mathbf{91.83 \pm 0.18}$

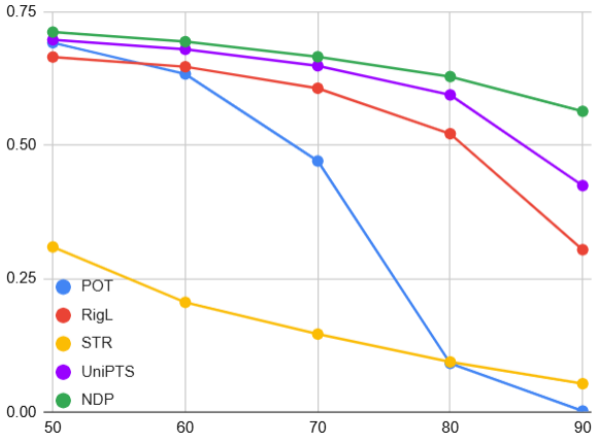
substantially alleviate this collapse, yet still suffer non-trivial accuracy losses when sparsity exceeds 95%. In contrast, NDP consistently delivers superior performance, achieving 93.15% at 95.5% sparsity and 91.83% at 98.85%, significantly outperforming the strongest method. The gap widens at ultra-high sparsity, highlighting NDP's capacity to preserve discriminative features even when the parameter budget is extremely constrained. These results confirm that NDP at the neuron level not only mitigates the risk of representational collapse but also scales more robustly to dense connectivity patterns such as those in DenseNet architectures.

1.11.4. Experiment on Tiny-ImageNet

To further validate the effectiveness of NDP, we evaluate its performance on the challenging Tiny-ImageNet dataset using the ResNet-18. Figure 6 reports the comparison against representative SOTA pruning methods, including SNIP, Iterative-SNIP, SynFlow, PHEW, and NBP, across a wide range of sparsity levels from moderate to extreme. The results demonstrate that NDP consistently outperforms all methods by a substantial margin in terms of top-1 classification accuracy while simultaneously reducing compu-



(a) DenseNet-121 on CIFAR-10



(b) MobileNet-V2 on ImageNet

Figure 5. NDP also outperforms other approaches for DenseNet-121 on CIFAR-10 and MobileNet-V2 on ImageNet. **Left**: DenseNet-121 on CIFAR-10. **Right**: MobileNet-V2 on ImageNet.

tational cost measured in FLOPs. At moderate sparsity, NDP achieves 72.10% accuracy, exceeding the second-best method by nearly 14 percentage points. This advantage becomes even more pronounced as the sparsity level increases. For example, under 96.84% sparsity, NDP maintains 60.23% accuracy, which is over 9 percentage points higher than PHEW and nearly 11 points higher than SynFlow. At the extreme pruning regime, NDP achieves 53.63% accuracy, whereas all other methods collapse below 41.05%. In addition to accuracy, NDP exhibits a highly favorable FLOPs-accuracy trade-off. For instance, at 90% sparsity, NDP reduces FLOPs to 2.32×10^8, which is less than half of PHEW and SynFlow, while simultaneously achieving 66.32% accuracy compared to 55.93% and 54.68%, respectively. Even at extreme sparsity, NDP preserves strong accuracy with an ultra-lightweight computational budget of only 0.28×10^8 FLOPs.

These results confirm that NDP not only retains significantly more discriminative power than existing pruning approaches but also achieves superior efficiency. The consistent dominance across varying sparsity regimes highlights the robustness of NDP. This demonstrates the key advantage of incorporating NDI as a principled criterion for pruning, allowing the network to selectively retain highly distinctive neurons while aggressively eliminating redundant ones.

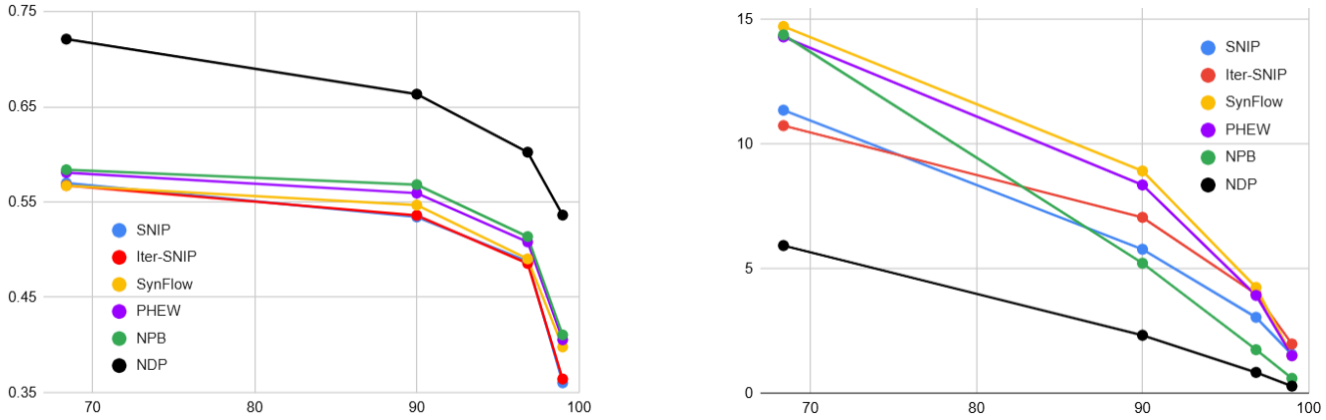
1.11.5. Experiment on ImageNet

We evaluate the proposed NDP method on the ImageNet dataset using the MobileNet-V2 and compare it against several SOTA pruning approaches, including POT, RigL, STR, and UniPTS. Figure 5 right reports the Top-1 accuracy across different pruning ratios ranging from 50% to 90%. As shown, NDP consistently outperforms all methods under every sparsity level. In particular, at moderate pruning levels, NDP achieves 69.45% and 66.62% Top-1 accuracy, surpassing the next best method, UniPTS, by margins of 1.44% and 1.69%, respectively. Even at extreme sparsity, NDP maintains a strong 56.39% accuracy, whereas other methods suffer significant degradation. These results demonstrate that NDP yields superior robustness to aggressive pruning while preserving competitive performance on large-scale datasets.

1.11.6. Leaky ReLU

Our pruning framework is fundamentally motivated by the principle of Neural Differentiation, which emphasizes that each neuron should contribute distinct representational information to the network. Neurons that fail to differentiate from others—exhibiting redundant or consistently uninformative activation patterns—can be pruned without impairing model expressivity. While this mechanism is naturally pronounced in ReLU activations, where neurons can become completely inactive due to the zeroing effect on negative inputs, the situation is less clear for Leaky ReLU [8, 9]. The small negative slope in Leaky ReLU prevents absolute inactivity, but neurons predominantly confined to the negative activation regime still provide little discriminative capacity and may be regarded as functionally redundant.

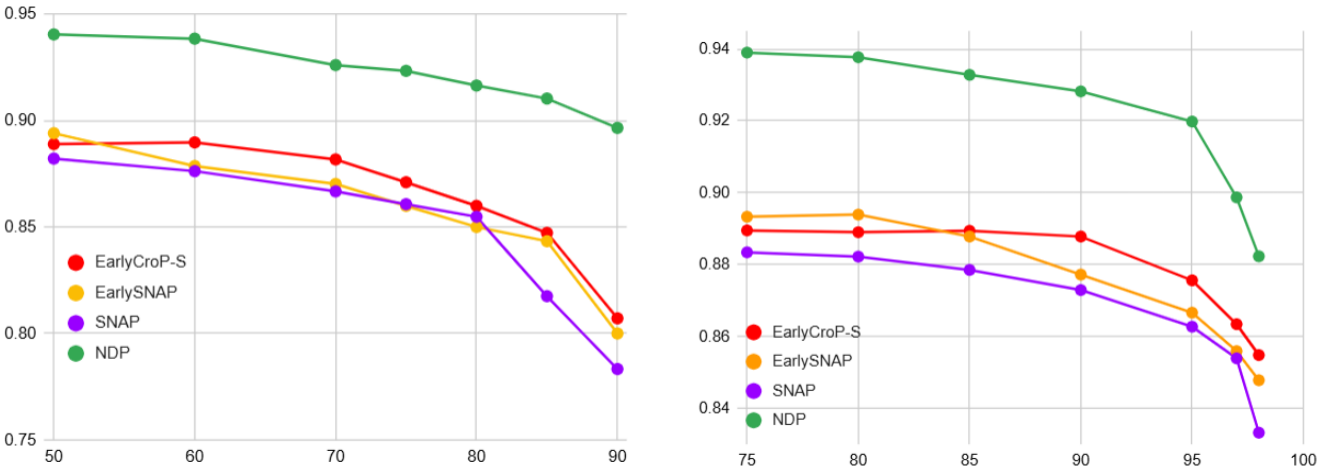
To evaluate this hypothesis, we applied NDP to ResNet-18 trained on CIFAR-10 with Leaky ReLU activations. NDP identifies neurons with low differentiation power—those whose activations remain clustered within uninformative subspaces—and eliminates them to encourage a more diverse representational basis. Tables 10 and 11 (Figure 7) report the performance under varying levels of neural sparsity and weight sparsity, respectively. The results demonstrate that, NDP consistently outperforms strong methods such as EarlyCroP-S, EarlySNAP,



(a) Weight Sparsity

(b) FLOPs (10^8)

Figure 6. NDP also outperforms other approaches for ResNet-18 networks trained on Tiny-ImageNet. **Left**: Weight sparsity. **Right**: FLOPs(10^8).



(a) Neuron Sparsity

(b) Weight Sparsity

Figure 7. ResNet-18 networks with *Leaky ReLU* trained on CIFAR-10. NDP again outperforms the other pruning methods. **Left**: Neural sparsity. **Right**: Weight sparsity.

Table 10. ResNet-18 networks with *Leaky ReLU* trained on CIFAR-10. NDP again outperforms the other pruning methods. **Neural sparsity (%)**.

	50	60	70	75	80	85	90
EarlyCroP-S	88.89	88.97	88.17	87.10	85.99	84.72	80.71
EarlySNAP	89.40	87.86	87.02	85.99	85.00	84.33	80.00
SNAP	88.21	87.62	86.67	86.07	85.48	81.75	78.33
NDP	**94.04**	**93.84**	**92.60**	**92.33**	**91.64**	**91.03**	**89.65**

Table 11. ResNet-18 networks with *Leaky ReLU* trained on CIFAR-10. NDP again outperforms the other pruning methods. **Weight sparsity (%)**.

	75	80	85	90	95	97	98
EarlyCroP-S	88.95	88.90	88.94	88.78	87.56	86.34	85.48
EarlySNAP	89.33	89.39	88.78	87.72	86.66	85.59	84.78
SNAP	88.34	88.22	87.85	87.29	86.27	85.39	83.32
NDP	**93.90**	**93.77**	**93.28**	**92.82**	**91.98**	**89.87**	**88.23**

and SNAP. Under neural sparsity constraints, NDP achieves accuracies above 92% even at 75–80% sparsity, whereas competing methods degrade more sharply, dropping below 86%. Similarly, under weight sparsity constraints, NDP preserves predictive performance above 91% at 95% sparsity, while other approaches fall below 87%. These results highlight that pruning guided by NDI—rather than simple magnitude or early-activation heuristics—offers greater re-

Table 12. Training time comparison on CIFAR-10 with ResNet-18.

Method	Total Training Time (× Baseline)
Baseline (No Pruning)	1.00×
SNIP	1.02–1.08×
SynFlow	1.05–1.20×
NDP (Ours)	**1.04–1.13**×

silience against accuracy degradation. Collectively, these findings reinforce our central claim: NDP fosters robust representational diversity, enabling networks to maintain high accuracy even under extreme sparsification, and extending its effectiveness beyond standard ReLU activations to Leaky ReLU networks.

1.11.7. Training Time Comparison

To assess the computational overhead introduced by pruning, we compare the total training time of different methods on CIFAR-10 with ResNet-18. As shown in Table 12, NDP incurs only a modest increase over the baseline training cost, while remaining competitive with existing pruning approaches.

References

[1] Utku Evci, et al. Rigging the lottery: Making all tickets winners. *International conference on machine learning*, page 2943–2952, 2020. 9

[2] Berivan Isik, et al. An information-theoretic justification for model pruning. In *International Conference on Artificial Intelligence and Statistics*, pages 3821–3846. PMLR, 2022. 9

[3] Lee Jaeho, et al. Layer-adaptive sparsity for the magnitude-based pruning. In *International Conference on Learning Representations*, 2021. 9

[4] Namhoon Lee, et al. Snip: Single-shot network pruning based on connection sensitivity. *International Conference on Learning Representations*, 2019. 8

[5] Ari Morcos, et al. One ticket to win them all: generalizing lottery ticket initializations across datasets and optimizers. *Advances in neural information processing systems*, 32, 2019. 9

[6] Hesham Mostafa and Xin Wang. Parameter efficient training of deep convolutional neural networks by dynamic sparse reparameterization. In *International Conference on Machine Learning*, pages 4646–4655. PMLR, 2019. 8

[7] John Rachwan, et al. Winning the lottery ahead of time: Efficient early network pruning. *Proceedings of the 39th International Conference on Machine Learning*, 162: 18293–18309, 2022. 6

[8] Haoyuan Sun, et al. Entropy-based activation function optimization: A method on searching better activation functions. In *The Thirteenth International Conference on Learning Representations*, 2025. 10

[9] Keke Tang, et al. Simplification is all you need against out-of-distribution overconfidence. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 5030–5040, 2025. 10

[10] Lin Tao, et al. Dynamic model pruning with feedback. In *International Conference on Learning Representations*, 2020. 8

[11] Stijn Verdenius, et al. Pruning via iterative ranking of sensitivity statistics. *CoRR*, 2020. 6

[12] Kaixin Xu, et al. Efficient joint optimization of layer-adaptive weight pruning in deep neural networks. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 17447–17457, 2023. 9

[13] Chenguang Zhang, et al. A deep neural network regularization measure. *Entropy*, 26(1), 2024. 8