LEARNING BEYOND FULL IMITATION: TASK-PRESERVING KNOWLEDGE DISTILLATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Knowledge distillation transfers knowledge by encouraging a student to match a teacher's predicted class probabilities. These probabilities express not only confidence in the correct class, but also relations among incorrect alternatives. Yet closer imitation does not necessarily yield a better student. Our key observation is that a student may already distinguish the correct class more sharply than its teacher, so further imitation can require giving back discrimination it has acquired. Our main result is an exact separation between full imitation and conditional learning. When the correct class's score advantage over each alternative must be preserved, full teacher-to-student KL minimization is blocked exactly when the student assigns no more probability than the teacher to every incorrect class. Crucially, the teacher's relative probabilities among incorrect classes remain fully learnable. We characterize the exact price of this transfer: a minimum increase in correct-class log-odds that compensates for the largest conditional-probability mismatch. Label fitting and conditional matching can therefore be completed even as full teacher KL diverges. This separation motivates task-preserving knowledge distillation (TPKD), which keeps the label gradient intact and minimally corrects the conditional gradient so that its output update preserves the label step's gains against every incorrect alternative. The corrected conditional direction retains more than half of the original first-order conditional descent at the same step size, with a tight bound. For a fixed positive conditional target and sufficiently small constant output steps, label and conditional errors vanish together. Experiments trace this learning from exact head updates to ordinary network training. TPKD reaches 88.05% accuracy on CIFAR-100 and 93.81% on CLINC150, improving over standard distillation by 0.47 and 0.35 percentage points across three seeds.

Keywords: knowledge distillation, task preservation, conditional knowledge transfer, constrained optimization, learning dynamics.

1 INTRODUCTION

Knowledge distillation (KD) transfers knowledge from a teacher to a student through the teacher's predictive distribution (Hinton et al., 2015). Besides confidence in the correct class, this distribution expresses how plausible the incorrect alternatives are relative to one another. These relative probabilities can encode similarities and confusions absent from hard labels, making them an important source of transferable knowledge (Zhao et al., 2022).

Progress in KD has been driven largely by empirical advances in objectives and training procedures, while theoretical analyses remain comparatively limited (Phuong & Lampert, 2019; Menon et al., 2021; Harutyunyan et al., 2023). A closer imitator is not necessarily a better student (Cho & Hariharan, 2019; Stanton et al., 2021). Existing theory studies capacity, teacher quality, regularization, and data geometry or optimization bias. We study a complementary question: when further imitation conflicts with the student's task progress, what can the teacher still teach?

Our key observation is that the student is not merely a passive copy of the teacher. Label learning may make it more discriminative on an example, assigning lower probability than the teacher to every incorrect class. Moving its full prediction closer to the teacher then requires giving back some of that discrimination. Yet the teacher may still describe a different relative ordering among the incorrect

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.

alternatives. Reproducing the teacher's complete prediction and learning the structure within it are therefore different objectives.

This leads to the central question of this work: *What can a student still learn from its teacher without sacrificing any of its existing task discrimination, and at what cost?*

Preserving all existing task discrimination requires more than keeping the answer correct. For example, a smaller score lead can make a previously accepted prediction require human confirmation (Liang et al., 2024). Since different confusions may require different acceptance thresholds, we preserve each correct-versus-incorrect score gap—its margin—while allowing the teacher to reshape relations among incorrect classes.

Our answer is an exact separation between full imitation and conditional learning. We prove that full imitation is blocked precisely when the student assigns no more probability than the teacher to every incorrect class. Its current prediction is then already the closest point to the teacher within the entire task-preserving region. Crucially, this does not mean that the teacher has nothing left to teach. The student can still recover the teacher's entire conditional distribution over incorrect classes, not merely their ranking, without decreasing any correct-class margin.

Learning this structure has an exact compensation cost. Matching a different conditional distribution raises some incorrect alternatives relative to others. Preserving their margins requires a corresponding increase in correct-class confidence. *The price is greater confidence, not weaker discrimination.* We derive the minimum increase in correct-class log-odds needed for exact, task-preserving transfer: the largest teacher-to-student conditional-probability ratio determines the cost, and a continuous safe path attains it. This yields an important consequence: the student can approach perfect label fitting while exactly matching the teacher's conditional structure, even as full teacher KL diverges. Learning the teacher's knowledge need not mean reproducing its confidence.

This separation suggests how to keep learning. We propose Task-Preserving Knowledge Distillation (TPKD). **Our main contributions are: Separation and transfer cost.** To our knowledge, we provide the first joint characterization of exact full-KL blocking, complete conditional reachability, and minimum confidence compensation under per-competitor margin preservation. We prove a necessary-and-sufficient condition for blocking and derive the attainable minimum correct-class log-odds cost. **A protected update with tight retention.** TPKD keeps the label gradient intact and adds the nearest safe conditional direction. It preserves every margin gain of the label step from the same start and with the same step size. The correction retains over half the first-order conditional descent with a tight bound, and a nonzero, aligned conditional component whenever error remains. **Joint learning and parameter-space guarantees.** For fixed positive targets and sufficiently small constant output steps, label and conditional errors vanish together. For nonuniform targets, this differs from CE's uniform conditional limit. Exact fixed-feature head updates and computable backpropagation conditions connect the output construction to network parameters.

Experiments test the mechanism and its use in full training. Across the evaluated visual CE states, blocking rises from 50.00% to 58.90%, with conditional knowledge still unlearned. All 64 exact-head batches gain additional conditional knowledge while retaining label-step margins; native updates carry this gain into the network and accumulate it over successive iterations. Full training reaches 88.05% on CIFAR-100 and 93.81% on CLINC150, exceeding standard KD by 0.47 and 0.35 percentage points across three seeds. Removing the CE gradient or replacing conditional learning with protected full imitation reduces accuracy in both domains, while removing projection gives closely matched performance.

2 Related work

Distillation targets. Classical KD matches softened teacher probabilities (Hinton et al., 2015), while DKD separates target-class and non-target supervision (Zhao et al., 2022). These approaches identify useful teacher signals; our analysis asks which signals remain attainable when existing task discrimination must be preserved.

Theoretical perspectives. KD theory addresses capacity mismatch (Mirzadeh et al., 2020), teacher quality and statistical supervision (Menon et al., 2021), self-distillation as regularization (Mobahi et al., 2020), and data geometry or optimization-induced deviations (Phuong & Lampert, 2019;

Nagarajan et al., 2023). We instead characterize what remains learnable, and at what cost, while preserving the student's acquired class comparisons.

Preservation and gradient coordination. MPT reduces prediction regressions during model updates through margin calibration and dual-source distillation (Ricci et al., 2026). PCGrad projects conflicting gradients (Yu et al., 2020); DeepKD decouples momentum updates and filters non-target knowledge (Huang et al., 2025); DTO-KD balances task and distillation gradients (Hayder et al., 2026). Our analysis first characterizes blocking and transfer cost in the task-preserving output region. TPKD then keeps the complete label gradient and minimally corrects each example's conditional increment, with tight bounds on retained learning.

3 What to learn and what to preserve

Learning targets. Labels specify the correct answer; the teacher also describes relations among alternatives. We distinguish label fitting from two teacher-learning objectives: reproducing the full prediction or learning only its conditional structure among incorrect classes.

Consider an example with label $y \in \{1, \dots, K\}$, $K \geq 3$. The student produces logits $z \in \mathbb{R}^K$ and probabilities $p = \mathrm{softmax}(z)$, where $p_j = e^{z_j} / \sum_{k=1}^{K} e^{z_k}$; the teacher supplies probabilities t. Both distributions are strictly positive. Define their correct-class confidences $\rho = p_y$, $\tau = t_y$, and conditional probabilities $r_j = p_j/(1-\rho)$, $q_j = t_j/(1-\tau)$ for $j \neq y$. The vectors r, q each sum to one over incorrect classes. With coordinate y first and the other class indices aligned, $p = (\rho, (1-\rho)r)$, $t = (\tau, (1-\tau)q)$. For positive distributions a, b, let $\mathrm{KL}(a\|b) = \sum_j a_j \log(a_j/b_j)$; $\mathrm{kl}_{\mathrm{Ber}}(\tau\|\rho)$ denotes the KL between $(\tau, 1-\tau)$ and $(\rho, 1-\rho)$. Label fitting minimizes $L_{\mathrm{CE}}(p, y) = -\log \rho$, full imitation minimizes $L_T(p) = \mathrm{KL}(t\|p)$, and conditional learning minimizes $L_{\mathrm{cond}}(p) = \mathrm{KL}(q\|r)$. The teacher loss decomposes as (Zhao et al., 2022)

$$L_T(p) = \mathrm{kl}_{\mathrm{Ber}}(\tau\|\rho) + (1-\tau)L_{\mathrm{cond}}(p). \tag{1}$$

Full imitation matches both teacher confidence and conditional structure. Increasing ρ with r fixed improves label fitting without changing these relations; label progress alone therefore does not measure conditional learning. We use natural logarithms and Euclidean inner products, norms and projections. Logit arguments mean evaluation at $p = \mathrm{softmax}(z)$, with y, t fixed.

Preserving learned comparisons. The student may change its prediction, but should not give back its acquired advantage over any competitor. For each $j \neq y$, define the margin $M_j(p) = \log(p_y/p_j) = z_y - z_j$. Let $\Delta_m^\circ = \{a \in \mathbb{R}^m : a_j > 0,\ \sum_{j=1}^m a_j = 1\}$ be the positive probability simplex. From reference p, the task-preserving predictions are

$$\mathcal{S}(p) = \{p' \in \Delta_K^\circ : M_j(p') \geq M_j(p) \text{ for all } j \neq y\}. \tag{2}$$

This protects individual comparisons as well as the label loss. For nonnegative competitor weights $\omega = (\omega_j)_{j \neq y}$, define $L_\omega(z, y) = \log(1 + \sum_{j \neq y} \omega_j e^{-M_j(z)})$. Preserving all margins is equivalent to not increasing any L_ω: unit weights give CE; a single unit weight isolates one competitor (Appendix A.1). CE alone can improve while one comparison worsens; Appendix A.6 gives its constrained imitation optimum.

How learning and protection interact. Since $p_j = (1-\rho)r_j$, we have $M_j(p) = \log[\rho/(1-\rho)] - \log r_j$. Matching a different conditional target raises some r_j, shrinking its margin at fixed ρ. Increasing correct-class confidence can compensate. Conditional learning leaves this confidence free; full imitation also seeks the teacher's confidence. We next determine which goal remains attainable within $\mathcal{S}(p)$, and at what minimum cost.

4 When imitation stops, what can the teacher still teach?

4.1 Why full imitation can become blocked

Our first result identifies when closer imitation must undo task progress. Once the student has suppressed every incorrect class at least as strongly as the teacher, its current prediction is already the best full imitation permitted by task preservation.

Theorem 1 (Full-imitation blocking). *For positive p, t, the inequalities $p_j \le t_j$ for all $j \neq y$ hold if and only if p uniquely minimizes L_T over $\mathcal{S}(p)$. For any $p' \in \mathcal{S}(p)$, define $\Delta M_j = M_j(p') - M_j(p)$. Under the blocking condition $p_j \le t_j$ for all $j \neq y$,*

$$L_T(p') - L_T(p) = \mathrm{KL}(p\|p') + \sum_{j\neq y}(t_j - p_j)\Delta M_j \ge \mathrm{KL}(p\|p'). \tag{3}$$

The teacher asks for at least as much probability on each incorrect class, while protection permits only nonnegative margin gains. Every term on the right of Eq. (3) is therefore nonnegative: any change from p strictly increases full imitation error. Yet conditional knowledge can remain unlearned ($r \neq q$) throughout a nonempty region of blocked states (Appendix A.2).

4.2 Learning the remaining structure, and its price

Blocking full imitation does not exhaust the teacher's knowledge. The student can still match the teacher's relative probabilities among incorrect classes. Raising a competitor's conditional probability requires enough extra correct-class confidence to preserve its margin. The next result gives the minimum increase needed for complete transfer.

For a candidate confidence $a \in (0,1)$, let $p' = (a, (1-a)q)$ match the full conditional target. Define $\mathrm{logit}(a) = \log[a/(1-a)]$, the log-odds of the correct class, $\mathcal{R}_\infty = \max_{j\neq y} q_j/r_j$, and $D_\infty(q\|r) = \log \mathcal{R}_\infty$. Let $a_{\min}$ be the smallest confidence permitting task-preserving transfer.

Theorem 2 (Minimum compensation for exact transfer). *The prediction p' belongs to $\mathcal{S}(p)$ if and only if*

$$\mathrm{logit}(a) - \mathrm{logit}(\rho) \ge D_\infty(q\|r), \qquad a_{\min} = \frac{\rho\mathcal{R}_\infty}{1-\rho+\rho\mathcal{R}_\infty} < 1. \tag{4}$$

A continuous path attains $a_{\min}$ while every margin remains nondecreasing. If $r \neq q$, then $a_{\min} > \rho$ and conditional KL decreases strictly along the path to zero.

The most underestimated competitor sets the price. To attain it, for $s \in [0,1]$ define $r(s) = (1-s)r + sq$, choose $a(s)$ by $\mathrm{logit}\, a(s) = \mathrm{logit}\,\rho + \log(1-s+s\mathcal{R}_\infty)$, and set $p(s) = (a(s), (1-a(s))r(s))$. This moves the whole conditional distribution toward the teacher while compensating just enough at each point (Appendices A.3–A.4); Appendix A.7 gives optimal partial transfer at smaller budgets.

A concrete example. In Figure 1, $y = 1$, $p = (0.90, 0.07, 0.03)$ and $t = (0.40, 0.25, 0.35)$. Both incorrect-class probabilities are below the teacher's, so full imitation is blocked. Yet the student favors class 2 over class 3 (7:3), whereas the teacher favors class 3 (5:7). At the transfer endpoint $p^* = (a_{\min}, (1-a_{\min})q)$, $a_{\min} = 35/37 \approx 0.946$: the class-3 margin stays fixed and the class-2 margin grows. The plot uses conditional coordinate $x = r_3$ and confidence increment $\Delta = \mathrm{logit}(a) - \mathrm{logit}(\rho)$. Appendix A.4 works through the calculation.

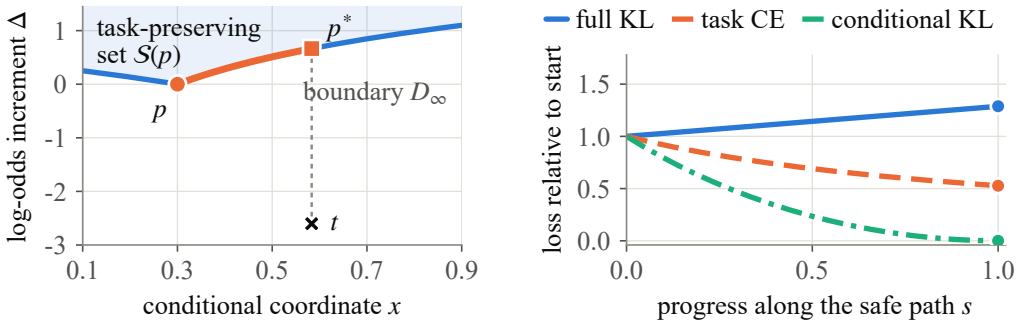


Figure 1: **Full imitation can worsen while conditional knowledge is learned.** The safe path attains the minimum confidence $a_{\min} \approx 0.946$. Label CE and conditional KL fall while full KL rises; losses are normalized by their initial values.

The separation is strongest near perfect label fitting. For $\varepsilon \in (0,1)$, define $p^{(\varepsilon)} = (1-\varepsilon, \varepsilon q)$, whose conditional distribution is $r^{(\varepsilon)} = q$. Then, as $\varepsilon \downarrow 0$,

$$L_{\mathrm{CE}}(p^{(\varepsilon)}, y) \to 0, \qquad \mathrm{KL}(q \| r^{(\varepsilon)}) = 0, \qquad \mathrm{KL}(t \| p^{(\varepsilon)}) \to \infty. \tag{5}$$

Learning the label and the teacher's conditional structure therefore need not make the student a closer full imitator. Theorem 2 identifies the minimum total confidence compensation required to complete this transfer. **We now turn to the local learning problem: how can each update acquire conditional knowledge while preserving all the progress of the corresponding label step?** TPKD addresses this question by making the smallest necessary correction to the combined learning direction.

5 Task-preserving knowledge distillation

5.1 Keep label learning intact and correct the teacher signal

The separation suggests a simple design: keep the progress of ordinary label learning and correct only the additional teacher signal that would interfere with it. The correction should be minimal, so protection does not unnecessarily discard teacher knowledge.

Let $e_j \in \mathbb{R}^K$ have one in coordinate j and zeros elsewhere. The two logit gradients are $h = \nabla_z L_{\mathrm{CE}} = p - e_y$ and $u = \nabla_z L_{\mathrm{cond}} = (0, r-q)$, where the zero occupies coordinate y. Let $\mathbf{1} \in \mathbb{R}^K$ denote the all-ones vector. Conditional gradients lie in the subspace $\mathcal{L}_y$. Directions that can be subtracted without reducing any margin form the safe cone $\mathcal{K}_y$:

$$\mathcal{L}_y = \{x \in \mathbb{R}^K : x_y = 0,\ \mathbf{1}^\top x = 0\}, \qquad \mathcal{K}_y = \{x \in \mathbb{R}^K : x_j - x_y \geq 0 \text{ for all } j \neq y\}.$$

A nonzero direction in $\mathcal{L}_y$ cannot belong to $\mathcal{K}_y$: changing only the relative wrong-class scores must favor some competitor. Thus $\mathcal{L}_y \cap \mathcal{K}_y = \{0\}$. Write $\Pi_{\mathcal{K}_y}$ for Euclidean projection onto $\mathcal{K}_y$, let d be the corrected conditional direction and v the complete update direction. With output step size $\eta \geq 0$, TPKD uses

$$d = \Pi_{\mathcal{K}_y}(u), \qquad \boxed{v = h + \tfrac{1}{2} d, \qquad z^+ = z - \eta v.} \tag{6}$$

Equivalently, v is the nearest direction to $h + u/2$ that preserves every label-step margin gain. The projection costs $O(K \log K)$ per example (Appendix B.1).

The reference is the progress that the label step would have achieved on its own. Starting from the same logits and using the same step size, adding the corrected conditional signal preserves every margin gain of that label step:

$$\begin{aligned} M_j(z - \eta v) &\geq M_j(z - \eta h) \geq M_j(z), \qquad j \neq y, \\ L_{\mathrm{CE}}(z - \eta v) &\leq L_{\mathrm{CE}}(z - \eta h) \leq L_{\mathrm{CE}}(z). \end{aligned} \tag{7}$$

The same ordering holds for every weighted label loss L_ω with $\omega \geq 0$. Thus TPKD preserves not only the student's existing discrimination, but also the additional discrimination that the corresponding CE step would have gained.

Training uses ordinary backpropagation. Let $\theta \in \mathbb{R}^P$ collect the P network parameters; subscript i indexes the N batch examples. Define the output Jacobian $J_i = \partial z_i / \partial \theta \in \mathbb{R}^{K \times P}$. Holding v_i fixed during differentiation with $\mathrm{stopgrad}$, we inject it through the surrogate loss

$$L_{\mathrm{sur}} = \frac{1}{N} \sum_{i=1}^{N} \langle \mathrm{stopgrad}(v_i), z_i \rangle, \qquad \nabla_\theta L_{\mathrm{sur}} = \frac{1}{N} \sum_{i=1}^{N} J_i^\top v_i. \tag{8}$$

For teacher logits z^T, let $z^T_{\neg y}$ denote the non-target coordinates. The fixed teacher supplies its original $q = \mathrm{softmax}(z^T_{\neg y})$. Figure 2 summarizes the two learning signals and their combination.

5.2 Protection retains real conditional learning

What does the correction change? It caps the teacher's strongest requests to raise competing classes and turns the clipped mass into correct-class compensation. Let $\widetilde{q}$ be the retained teacher masses and

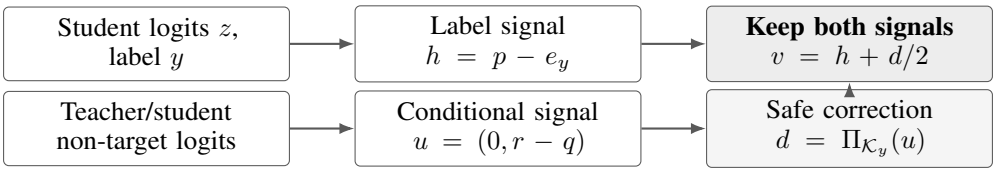


Figure 2: **TPKD keeps the label signal intact.** Only the teacher-conditional direction is corrected. The combined direction is passed to ordinary backpropagation through Eq. (8).

$\ell = -d_y \geq 0$ the compensation. With $[a]_+ = \max(a, 0)$, ℓ uniquely solves $\ell = \sum_{j \neq y} [q_j - r_j - \ell]_+$, giving

$$\widetilde{q}_j = \min(q_j, r_j + \ell), \qquad d = (-\ell, r - \widetilde{q}), \qquad \sum_{j \neq y} \widetilde{q}_j = 1 - \ell. \tag{9}$$

Freezing the clipping at the current output gives a local loss whose gradient equals d at that output: fit the normalized clipped teacher and increase correct-class log-odds (Appendix B.6). The extra margin over the CE endpoint is $\frac{\eta}{2}[r_j - q_j + \ell]_+$: the most demanding competitors keep exactly the CE margin; the others gain more.

The important question is whether protection preserves genuine conditional learning or merely increases confidence in the correct class. To distinguish these effects, define the symmetric confidence direction $a_0 = (-1, 1/(K-1), \ldots, 1/(K-1))$ and the remaining component $w_{\text{cond}} = d - \ell a_0$. Then $d = \ell a_0 + w_{\text{cond}}$, $w_{\text{cond}} \in \mathcal{L}_y$, $w_{\text{cond}} \perp a_0$. A step along $-a_0$ increases every correct-class margin equally and leaves the relative probabilities among incorrect classes unchanged. The component w_{cond} changes those relations. The next theorem shows that this component remains nonzero and aligned with the teacher's conditional signal whenever conditional error remains.

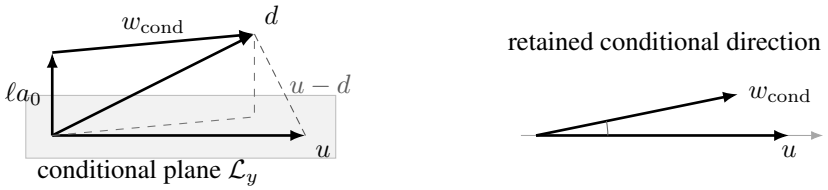
Theorem 3 (Retained conditional learning)**.** *For $u \in \mathcal{L}_y$, define $\gamma_K = K/[2(K-1)]$ and $c_K = 2\sqrt{\gamma_K}/(1+\gamma_K)$. Then*

$$u^\top d = \|d\|^2 \geq \gamma_K \|u\|^2, \qquad \|w_{\text{cond}}\|^2 \geq \gamma_K \|d\|^2 \geq \gamma_K^2 \|u\|^2, \tag{10}$$

$$\cos \angle(u, w_{\text{cond}}) \geq c_K \geq \frac{2\sqrt{2}}{3} \qquad (u \neq 0). \tag{11}$$

Both norm bounds are tight and can be attained simultaneously.

Here $u^\top d$ measures first-order conditional loss reduction along $-d$. The first bound retains more than half the reduction along $-u$ at the same step size; the remaining bounds ensure a nonzero, aligned conditional component whenever $r \neq q$ (Figure 3). Appendix B quantifies the removed signal and compensation, and establishes optimality under local quadratic and fixed-displacement budgets.



(a) $d = \ell a_0 + w_{\text{cond}}$ (b) view within $\mathcal{L}_y$

Figure 3: **Protection preserves a genuine conditional-learning direction.** The corrected direction combines confidence compensation with a pure conditional component (a), which remains closely aligned with the original signal (b). The four-class example is constructed in Appendix B.5.

5.3 Both kinds of learning can be completed together

Can the learning retained in each step accumulate until both objectives are achieved? We prove that it can: for a fixed positive teacher-conditional target and sufficiently small constant output steps,

label fitting and conditional matching converge together. To track their joint progress, define the potential $\Phi(z) = 4L_{\mathrm{CE}}(z, y) + L_{\mathrm{cond}}(z)$ and its logit gradient $g_\Phi = \nabla_z \Phi = 4h + u$. A subscript k denotes evaluation at output z_k, for example $p_k = \mathrm{softmax}(z_k)$ and $v_k = v(z_k)$.

Theorem 4 (Joint descent and completed learning)**.** *The complete direction obeys the tight alignment bound*

$$g_\Phi^\top v \geq \|v\|^2 + \frac{\gamma_K}{4}\|g_\Phi\|^2, \qquad \cos\angle(g_\Phi, v) \geq \sqrt{\gamma_K}. \tag{12}$$

For fixed positive q, finite initial logits, and $z_{k+1} = z_k - \eta v_k$ with $0 < \eta \leq 4/5$,

$$\Phi(z_{k+1}) \leq \Phi(z_k) - \frac{\eta\gamma_K}{4}\|g_{\Phi,k}\|^2 - \eta(1 - 5\eta/4)\|v_k\|^2. \tag{13}$$

Consequently, $p_{k,y} \to 1$, $r_k \to q$, and $v_k \to 0$.

The retained conditional signal therefore supports completed learning, not just a favorable local direction. Joint and component gradient residuals have vanishing time-averaged squared norms (Appendix C.2); Appendix C.4 gives conditions for a complete step to lower its own conditional KL. For comparison, define the uniform non-target vector by $\nu_j = 1/(K-1)$ for $j \neq y$. CE output gradient flow, or fixed steps $0 < \eta \leq 1/20$, fits the label but drives $r \to \nu$. For nonuniform q, both rules fit the label, but TPKD retains the teacher's conditional structure rather than a uniform distribution (Appendix C.3).

5.4 Turning output progress into parameter learning

The output construction becomes useful for training through two connections: exact execution by a classification head, and conditional learning beyond compensation under ordinary backpropagation.

Proposition 1 (Exact head update)**.** *If the frozen batch feature matrix has full row rank, a unique minimum-Frobenius-norm head displacement realizes Eq. (6) exactly for every example. It therefore inherits the step's margin and loss orderings in Eq. (7) (Appendix D.1).*

Learning beyond compensation. Compare the parameter steps from Eq. (8) for TPKD and $h_i + \ell_i a_{0,i}/2$: the same confidence compensation without direct conditional adjustment. Backpropagation can distort these signals. Let χ_c, χ_Φ be its largest-to-smallest squared stretch ratios on the spans of the batch-concatenated pairs $(u_i, w_{\mathrm{cond},i})$ and $(g_{\Phi,i}, v_i)$, respectively. Zero minimum stretch gives an infinite ratio. Define $\kappa_\Phi = (1 + \sqrt{\gamma_K})/(1 - \sqrt{\gamma_K})$ and $\kappa_c = \kappa_\Phi^2$.

Theorem 5 (Conditional learning through network parameters)**.** *At a differentiable parameter state, for sufficiently small equal positive parameter steps:* (i) *if some $u_i \neq 0$ and $\chi_c < \kappa_c$, TPKD has a nonzero conditional parameter increment over compensation only and achieves strictly lower batch-average conditional KL;* (ii) *if both joint batch vectors are nonzero and $\chi_\Phi < \kappa_\Phi$, TPKD strictly decreases the batch mean of Φ.*

The first conclusion guarantees a useful conditional increment, not merely extra confidence; the second guarantees progress of the complete update. These results establish both a parameter implementation and, under the stated conditions, genuine learning from the retained teacher signal. Appendix D gives the equivalent formulas, proofs, quantitative bounds, computation and measured-angle refinements.

6 Experiments

The mechanism study uses CIFAR-100 (Krizhevsky, 2009), with a VOLO-D2 teacher and PiT-B student (Yuan et al., 2023; Heo et al., 2021), under the main training protocol. CE20/40/60 denote label-only students after 20/40/60 epochs. Margins measure the correct class's advantage over each competitor; conditional KL measures the error in learning the teacher's relations among incorrect classes.

Blocking in ordinary training. We first test whether ordinary label training can block full imitation while leaving conditional knowledge unlearned. Each CE state is evaluated on 2,000 fixed images under eight views; an image–view pair is blocked when the student's probability for every incorrect class is no greater than the teacher's. Blocking rises from 50.00% to 58.90%, yet every blocked pair has positive conditional error (Table 1). Thus the theoretical obstruction occurs in ordinary training while teacher knowledge remains available.

Table 1: **Full-imitation blocking leaves conditional knowledge unlearned.** Conditional error and transfer cost are means within the blocked set (nats).

Student state	Blocked (%)	Remaining conditional KL	Minimum log-odds cost
CE20	50.00	1.696	7.520
CE40	55.29	1.766	7.632
CE60	58.90	1.803	7.569

Protection without discarding knowledge. Exact head updates on frozen features isolate whether the prescribed correction protects the task without discarding conditional learning. At CE20 and CE60, we compare TPKD with CE, compensation only, unprojected learning, and CE plus full KL from the same start. Compensation only keeps the correct-class push but removes changes among incorrect classes. TPKD gains conditional knowledge beyond CE on all 64 batches and all 4,096 example states while preserving every CE-step margin to numerical precision. Compensation only has CE's conditional endpoint; unprojected learning loses at least one CE-step margin on every example. TPKD therefore combines the two desired effects. Minimum retention and complete-update cosine exceed the 100-class bounds of approximately 50.51% and 0.7107 (Theorems 3–4); all joint-descent checks pass (Table 2).

Table 2: **Exact head steps learn conditional structure while preserving label-step progress.**

State	Gain (10^{-3})	KL change (10^{-3})	Min. retained (%)	Min. cosine	Φ change (10^{-3})
CE20	1.195	-1.359	52.13	0.7220	-3.302
CE60	1.213	-1.283	52.13	0.7221	-1.831

KL denotes conditional KL. Gain is CE minus TPKD at the endpoint; changes are TPKD endpoint minus common start. Loss columns are means; minima are over individual examples. Retention is $u^\top d/\|u\|^2$; cosine compares v with g_Φ. Each state: 32×64 examples; output step 0.01.

Conditional learning through the network. We next test whether this conditional increment remains useful under ordinary full-network optimization. At CE20 and CE60, 16 native-optimizer pairs per state compare TPKD with compensation only from identical model, optimizer and random states. For $K = 100$, Theorem 5 gives distortion-ratio limits of approximately 34.96 for conditional learning and 5.91 for joint descent. All 32 batches pass both geometric tests, and every native pair has lower conditional KL under TPKD (Table 3). These paired gains isolate conditional learning beyond compensation.

Table 3: **The conditional advantage survives ordinary backpropagation.**

State	Conditional test	Joint descent test	Positive pairs	Mean gain (10^{-3})
CE20	16/16	16/16	16/16	8.252
CE60	16/16	16/16	16/16	9.886

Gain is compensation-only endpoint conditional KL minus TPKD endpoint conditional KL.

Accumulation over successive updates. Finally, we test whether the single-step advantage accumulates. From CE60, CE, compensation only and TPKD follow the same data sequence for 128 native-optimizer iterations. On 2,048 fixed observation images separate from the update images, TPKD lowers conditional KL from 1.800 to 1.557. Its endpoint KL is 0.244 nats below CE's and 0.320 below compensation only's: the extra conditional learning persists over successive updates. Protocols are in Appendix E.

6.1 Full training in vision and text

TPKD is not limited to vision: its update uses class probabilities and labels, not modality-specific features. We therefore also evaluate full training on CLINC150 text intent classification (Larson et al., 2019), with BERT-large as teacher and BERT-Mini as student (Devlin et al., 2019; Turc et al., 2019).

Table 4 compares CE and nine distillation methods (Hinton et al., 2015; Zhao et al., 2022; Roth et al., 2024; Yang et al., 2025; Hayder et al., 2026) using shared within-domain protocols and final-epoch evaluation fixed in advance over three seeds (Appendix E.1).

TPKD reaches 88.05% on CIFAR-100 and 93.81% on CLINC150, improving over standard KD by 0.47 and 0.35 percentage points and over CE by 0.68 and 0.56 points. Retained conditional learning thus benefits full training in both domains.

Table 4: **Full-training accuracy (%).** Mean ± sample standard deviation over three seeds.

Method	CIFAR-100 60 epochs	CLINC150 4 epochs	Method	CIFAR-100 60 epochs	CLINC150 4 epochs
CE	87.37 ± 0.07	93.24 ± 0.28	DP-S	83.65 ± 0.01	93.44 ± 0.28
KD	87.58 ± 0.07	93.46 ± 0.34	XE-KL	84.17 ± 0.05	93.41 ± 0.22
KL-Dist	82.26 ± 0.05	93.38 ± 0.19	DHKD	85.83 ± 0.10	93.29 ± 0.38
DKD	84.53 ± 0.09	93.59 ± 0.20	DTO-KD*	87.81 ± 0.11	93.46 ± 0.23
DP-U	83.14 ± 0.04	93.38 ± 0.19	**TPKD**	$\mathbf{88.05 \pm 0.12}$	$\mathbf{93.81 \pm 0.27}$

*Text DTO-KD includes multi-layer feature distillation. Full configurations are in Appendix E.1.

6.2 Which components make the difference?

Independent label learning. Matching the conditional target makes $d = 0$, but the CE gradient continues fitting the label. Removing it lowers accuracy by 2.72 points in vision and 2.21 in text (Table 5).

Learning the right teacher object. Safe full KL keeps CE and projection but uses $p - t$ instead of the conditional gradient. In a blocked state, its projected teacher increment is zero, whereas TPKD's remains nonzero when $r \neq q$ (Appendix A.5). Accuracy falls to 87.21% and 93.33%, close to CE. Protecting full imitation alone does not recover the benefit of conditional learning.

The cost of protection. Without projection, accuracy is 87.83% and 93.83%: TPKD is 0.22 points higher in vision and differs by only 0.02 points in text. The correction thus provides the demonstrated margin protection while retaining closely matched predictive performance.

Table 5: **Component ablations.** Accuracy (%), mean ± sample standard deviation over three seeds.

Condition	Direction	CIFAR-100	CLINC150
CE	h	87.37 ± 0.07	93.24 ± 0.28
Without CE gradient	$d/2$	85.33 ± 0.05	91.60 ± 0.07
Safe full-KL	$h + \Pi_{\mathcal{K}_y}(p - t)/2$	87.21 ± 0.04	93.33 ± 0.29
Without projection	$h + u/2$	87.83 ± 0.03	93.83 ± 0.23
TPKD	$h + d/2$	88.05 ± 0.12	93.81 ± 0.27

7 Conclusion

Full imitation can become blocked before the teacher's knowledge is exhausted. We prove that its conditional structure remains fully transferable under task preservation and determine the exact confidence compensation required. TPKD implements this separation by retaining the label gradient and adding the nearest safe conditional direction. Tight retention and joint-learning results explain why genuine conditional learning survives; parameter-space results connect it to network updates. Mechanism experiments locate blocking in ordinary training and follow conditional gains through successive updates, while full training and ablations demonstrate their contribution in vision and text. Learning from a teacher need not mean reproducing its full prediction: conditional knowledge can be acquired without surrendering label-learning progress.

AI USE STATEMENT

No generative AI tools were used in conducting this research or preparing this manuscript.

REPRODUCIBILITY STATEMENT

Appendices A–D prove the results stated in the main text; Appendix E specifies the reported experiments. The source bundle includes the reported accuracy summaries, table-generation scripts and analytical checks.

REFERENCES

Jang Hyun Cho and Bharath Hariharan. On the efficacy of knowledge distillation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 4794–4802, 2019.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pp. 4171–4186, 2019.

Hrayr Harutyunyan, Ankit Singh Rawat, Aditya Krishna Menon, Seungyeon Kim, and Sanjiv Kumar. Supervision complexity and its role in knowledge distillation. In *International Conference on Learning Representations*, 2023. URL https://arxiv.org/abs/2301.12245.

Zeeshan Hayder, Ali Cheraghian, Lars Petersson, Mehrtash Harandi, and Richard Hartley. DTO-KD: Dynamic trade-off optimization for effective knowledge distillation. In *International Conference on Learning Representations*, 2026.

Byeongho Heo, Sangdoo Yun, Dongyoon Han, Sanghyuk Chun, Junsuk Choe, and Seong Joon Oh. Rethinking spatial dimensions of vision transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 11936–11945, 2021.

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

Haiduo Huang, Jiangcheng Song, Yadong Zhang, and Pengju Ren. DeepKD: A deeply decoupled and denoised knowledge distillation trainer. In *Advances in Neural Information Processing Systems*, volume 38, pp. 27138–27167, 2025. doi: 10.52202/085713-0915.

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

Stefan Larson, Anish Mahendran, Joseph J. Peper, Christopher Clarke, Andrew Lee, Parker Hill, Jonathan K. Kummerfeld, Kevin Leach, Michael A. Laurenzano, Lingjia Tang, and Jason Mars. An evaluation dataset for intent classification and out-of-scope prediction. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pp. 1311–1316, 2019.

Hengyue Liang, Le Peng, and Ju Sun. Selective classification under distribution shifts. *Transactions on Machine Learning Research*, 2024. ISSN 2835-8856. URL https://openreview.net/forum?id=dmxMGW6J7N.

Aditya K. Menon, Ankit Singh Rawat, Sashank Reddi, Seungyeon Kim, and Sanjiv Kumar. A statistical perspective on distillation. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pp. 7632–7642, 2021.

Seyed Iman Mirzadeh, Mehrdad Farajtabar, Ang Li, Nir Levine, Akihiro Matsukawa, and Hassan Ghasemzadeh. Improved knowledge distillation via teacher assistant. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pp. 5191–5198, 2020.

Hossein Mobahi, Mehrdad Farajtabar, and Peter L. Bartlett. Self-distillation amplifies regularization in Hilbert space. In *Advances in Neural Information Processing Systems*, volume 33, pp. 3351–3361, 2020.

Vaishnavh Nagarajan, Aditya K. Menon, Srinadh Bhojanapalli, Hossein Mobahi, and Sanjiv Kumar. On student-teacher deviations in distillation: Does it pay to disobey? In *Advances in Neural Information Processing Systems*, volume 36, pp. 5961–6000, 2023. doi: 10.52202/075280-0261. URL https://proceedings.neurips.cc/paper_files/paper/2023/hash/12d286282e1be5431ea05262a21f415c-Abstract-Conference.html.

Neal Parikh and Stephen Boyd. Proximal algorithms. *Foundations and Trends in Optimization*, 1(3): 127–239, 2014.

Mary Phuong and Christoph H. Lampert. Towards understanding knowledge distillation. In *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pp. 5142–5151, 2019.

Simone Ricci, Niccolò Biondi, Federico Pernici, and Alberto Del Bimbo. Mitigating negative flips via margin preserving training. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 8721–8730, 2026. doi: 10.1609/aaai.v40i11.37825.

Karsten Roth, Lukas Thede, Almut Sophia Koepke, Oriol Vinyals, Olivier Hénaff, and Zeynep Akata. Fantastic gains and where to find them: On the existence and prospect of general knowledge transfer between any pretrained model. In *International Conference on Learning Representations*, 2024.

Samuel Stanton, Pavel Izmailov, Polina Kirichenko, Alexander A. Alemi, and Andrew Gordon Wilson. Does knowledge distillation really work? In *Advances in Neural Information Processing Systems*, volume 34, pp. 6906–6919, 2021.

Iulia Turc, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Well-read students learn better: On the importance of pre-training compact models. *arXiv preprint arXiv:1908.08962*, 2019.

Penghui Yang, Chen-Chen Zong, Sheng-Jun Huang, Lei Feng, and Bo An. Dual-head knowledge distillation: Enhancing logits utilization with an auxiliary head. In *Proceedings of the 31st ACM SIGKDD Conference on Knowledge Discovery and Data Mining V.2*, pp. 3530–3541, 2025.

Tianhe Yu, Saurabh Kumar, Abhishek Gupta, Sergey Levine, Karol Hausman, and Chelsea Finn. Gradient surgery for multi-task learning. In *Advances in Neural Information Processing Systems*, volume 33, pp. 5824–5836, 2020.

Li Yuan, Qibin Hou, Zihang Jiang, Jiashi Feng, and Shuicheng Yan. VOLO: Vision outlooker for visual recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(5): 6575–6586, 2023. doi: 10.1109/TPAMI.2022.3206108.

Borui Zhao, Quan Cui, Renjie Song, Yiyu Qiu, and Jiajun Liang. Decoupled knowledge distillation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 11953–11962, 2022.

A PROOFS SUPPORTING TASK PRESERVATION AND THE SEPARATION

This appendix proves the equivalence in Section 3, Theorems 1–2, and the two component consequences stated in Section 6.2.

A.1 WEIGHTED TASK LOSSES AND PER-MARGIN PRESERVATION

For nonnegative competitor weights ω, let $L_\omega(z, y) = \log(1 + \sum_{j \neq y} \omega_j e^{-M_j(z)})$. The all-ones choice is cross-entropy; a single nonzero weight isolates one competitor.

Proof of the equivalence in Section 3. For logits $z, z' \in \mathbb{R}^K$, if $M_j(z') \geq M_j(z)$ for all $j \neq y$, then $e^{-M_j(z')} \leq e^{-M_j(z)}$; any $\omega_j \geq 0$ preserves the termwise inequality, and summing and using the monotonicity of $\log$ gives $L_\omega(z', y) \leq L_\omega(z, y)$. Conversely, for any $j \neq y$ take $\omega_j = 1$ and all other weights zero, which gives $\log(1 + e^{-M_j(z')}) \leq \log(1 + e^{-M_j(z)})$; since $M \mapsto \log(1 + e^{-M})$ is strictly decreasing, $M_j(z') \geq M_j(z)$. Taking z' and the reference to be $z - \eta v$ and $z - \eta h$ yields the ordering of TPKD against the label base under every competitor weighting.

A.2 PROOF OF THEOREM 1

Let $\Delta M_j = M_j(p') - M_j(p)$. Since $\sum_j (t_j - p_j) = 0$, the common quantity $\log(p_y / p'_y)$ may be subtracted from every term without changing the sum, so

$$\begin{aligned} \mathrm{KL}(t\|p') - \mathrm{KL}(t\|p) - \mathrm{KL}(p\|p') &= \sum_j (t_j - p_j) \log \frac{p_j}{p'_j} \\ &= \sum_{j \neq y} (t_j - p_j) \Big[\log \frac{p_j}{p'_j} - \log \frac{p_y}{p'_y} \Big] = \sum_{j \neq y} (t_j - p_j)\, \Delta M_j. \end{aligned} \tag{A.1}$$

This gives the identity in Eq. (3) before imposing the sign conditions. When $p' \in \mathcal{S}(p)$ and $p_j \leq t_j$ for all non-target coordinates, the right-hand side is nonnegative; for strictly positive probabilities $\mathrm{KL}(p\|p')$ is strictly positive when $p' \neq p$, so p is the unique minimizer and Eq. (3) follows. Conversely, if $p_j > t_j$ for some j, let $z' = z - \eta e_j$ and $p' = \mathrm{softmax}(z')$: this leaves the other margins unchanged and increases the j-th, so $p' \in \mathcal{S}(p)$; since $\partial_{z_j} \mathrm{KL}(t\|\,\mathrm{softmax}(z)) = p_j - t_j$, the derivative of the full KL in η at zero equals $-(p_j - t_j) < 0$, and p is not a minimizer.

Strict blocking with residual conditional error. For any fixed positive $r \neq q$ and $\tau < 1$, every ρ with

$$0 < 1 - \rho < (1 - \tau) \min_{j \neq y} q_j / r_j$$

is strictly blocked and has $\mathrm{KL}(q\|r) > 0$. Thus blocking with unlearned conditional structure has nonempty interior.

A.3 PROOF OF THEOREM 2

With $\mathcal{R}_\infty = \max_{j \neq y} q_j / r_j$, the smallest feasible correct-class probability is

$$a_{\min} = \frac{\rho\, \mathcal{R}_\infty}{1 - \rho + \rho\, \mathcal{R}_\infty} < 1. \tag{A.2}$$

For $p' = (a, (1 - a)q)$, the j-th margin constraint $M_j(p') \geq M_j(p)$ is equivalent to

$$\log \frac{a}{(1 - a) q_j} \geq \log \frac{\rho}{(1 - \rho) r_j} \iff \mathrm{logit}(a) - \mathrm{logit}(\rho) \geq \log \frac{q_j}{r_j}.$$

Taking the maximum over j gives $D_\infty(q\|r)$; solving $\mathrm{logit}(a) = \mathrm{logit}(\rho) + \log \mathcal{R}_\infty$ for a gives Eq. (4). Normalization $\sum_j q_j = \sum_j r_j = 1$ implies $\mathcal{R}_\infty \geq 1$ with equality if and only if $q = r$; positivity ensures $\mathcal{R}_\infty < \infty$, hence $a_{\min} < 1$.

A.4 THE MINIMUM-COMPENSATION PATH

For $s \in [0,1]$, define

$$r(s) = (1-s)r + sq, \qquad a(s) = \frac{\rho(1-s+s\mathcal{R}_\infty)}{1-\rho+\rho(1-s+s\mathcal{R}_\infty)}. \tag{A.3}$$

Proof. Write $l_j = q_j/r_j \le \mathcal{R}_\infty$. By Eq. (A.3), $r_j(s)/r_j = 1-s+sl_j$ and $\operatorname{logit} a(s) - \operatorname{logit}\rho = \log(1-s+s\mathcal{R}_\infty)$, so

$$M_j(p(s)) - M_j(p) = \log(1-s+s\mathcal{R}_\infty) - \log(1-s+sl_j) \ \ge 0, \tag{A.4}$$

$$\frac{d}{ds} M_j(p(s)) = \frac{\mathcal{R}_\infty - l_j}{(1-s+s\mathcal{R}_\infty)(1-s+sl_j)} \ \ge 0. \tag{A.5}$$

With $\Delta_j = q_j - r_j$ we have $q_j = r_j(s) + (1-s)\Delta_j$ and $\sum_j \Delta_j = 0$, hence

$$\frac{d}{ds}\,\mathrm{KL}(q\|r(s)) = -\sum_j \frac{q_j\Delta_j}{r_j(s)} = -(1-s)\sum_j \frac{\Delta_j^2}{r_j(s)}. \tag{A.6}$$

If $q \neq r$ this derivative is strictly negative for $s < 1$; at $s = 1$, $r(1) = q$ and $a(1) = a_{\min}$.

The numerical illustration in Figure 1. The correct label is class 1, with $p = (0.90, 0.07, 0.03)$ and $t = (0.40, 0.25, 0.35)$. Since $0.07 < 0.25$ and $0.03 < 0.35$, Theorem 1 applies. The conditional distributions are nevertheless different:

$$\rho = \frac{9}{10}, \quad \tau = \frac{2}{5}, \qquad r = \left(\frac{7}{10}, \frac{3}{10}\right), \qquad q = \left(\frac{5}{12}, \frac{7}{12}\right).$$

The largest relative mismatch is on class 3:

$$\mathcal{R}_\infty = \max\left\{\frac{25}{42}, \frac{35}{18}\right\} = \frac{35}{18}, \qquad D_\infty(q\|r) = \log\frac{35}{18} \approx 0.664976.$$

Inserting this value into Eq. (4) gives

$$a_{\min} = \frac{(9/10)(35/18)}{1/10 + (9/10)(35/18)} = \frac{35}{37}, \qquad p^* = \left(\frac{35}{37}, \frac{5}{222}, \frac{7}{222}\right).$$

The endpoint's non-target ratio is 5:7, exactly the teacher's. Its correct-versus-incorrect probability ratios are

$$\frac{p_1}{p_2} = \frac{90}{7} \ \longrightarrow\ \frac{p_1^*}{p_2^*} = 42, \qquad \frac{p_1}{p_3} = 30 \ \longrightarrow\ \frac{p_1^*}{p_3^*} = 30.$$

Thus the class-2 margin increases and the class-3 margin is unchanged. Along the path,

$$r_2(s) = \frac{7}{10} - \frac{17s}{60}, \quad r_3(s) = \frac{3}{10} + \frac{17s}{60}, \qquad a(s) = \frac{\frac{9}{10}(1+\frac{17s}{18})}{\frac{1}{10} + \frac{9}{10}(1+\frac{17s}{18})}.$$

The left plot uses $x = r_3(s)$ and $\Delta = \operatorname{logit}(a(s)) - \operatorname{logit}(\rho)$. For an arbitrary positive conditional vector $(1-x, x)$, its safe boundary is $\Delta = D_\infty((1-x,x)\|r)$. The teacher is at $x = 7/12$ and $\Delta = \operatorname{logit}(2/5) - \operatorname{logit}(9/10) = \log(2/27) \approx -2.60269$, below this region.

The right plot divides each loss by its own initial value. Substitution into the loss definitions gives

	p	p^*
L_{CE}	$-\log(9/10)$	$-\log(35/37)$
L_{cond}	$\frac{5}{12}\log\frac{25}{42} + \frac{7}{12}\log\frac{35}{18}$	0
L_T	$\frac{2}{5}\log\frac{4}{9} + \frac{1}{4}\log\frac{25}{7} + \frac{7}{20}\log\frac{35}{3}$	$\frac{2}{5}\log\frac{74}{175} + \frac{3}{5}\log\frac{111}{10}$

Numerically, label CE changes from approximately 0.105361 to 0.055570, conditional KL from 0.171739 to zero, and full teacher KL from 0.853727 to 1.099879. These are evaluations of the analytic path, whose margin monotonicity and conditional descent were established above.

A.5 THE TWO COMPONENT CONSEQUENCES USED IN THE ABLATION

For the control without h, $r = q$ implies $u = (0, r - q) = 0$ and hence $d = \Pi_{\mathcal{K}_y}(0) = 0$. This need not imply that $p_y = 1$, because r is independent of the correct-class probability. The ordinary label gradient has $h_y = p_y - 1$, so it remains nonzero at every positive prediction with $p_y < 1$.

For the safe full-KL control, define the polar cone $\mathcal{K}_y^\circ = \{g \in \mathbb{R}^K : g^\top x \le 0 \text{ for every } x \in \mathcal{K}_y\}$. Put $g = p - t$ and suppose $p_j \le t_j$ for $j \ne y$. For every $x \in \mathcal{K}_y$, using $\mathbf{1}^\top g = 0$,

$$g^\top x = \sum_{j \ne y} (t_j - p_j)(x_y - x_j) \le 0.$$

Thus g belongs to the polar cone $\mathcal{K}_y^\circ$ and $\Pi_{\mathcal{K}_y}(g) = 0$. If $r \ne q$, Theorem 3 instead gives $\|\Pi_{\mathcal{K}_y}(u)\|^2 \ge \gamma_K \|u\|^2 > 0$. This proves the loss of the full-KL teacher increment at a blocked state, while the conditional increment remains available.

A.6 THE IMITATION OPTIMUM WHEN ONLY CE IS PROTECTED

The comparison in Section 3 can be solved exactly. Preserving only the label cross-entropy requires $p'_y \ge \rho$. Write $p' = (a, (1-a)x)$, where x is a conditional probability vector. The KL decomposition gives

$$\mathrm{KL}(t \| p') = \mathrm{kl}_{\mathrm{Ber}}(\tau \| a) + (1 - \tau)\,\mathrm{KL}(q \| x).$$

For any feasible a, the second term is uniquely minimized by $x = q$. Since

$$\frac{\partial}{\partial a}\,\mathrm{kl}_{\mathrm{Ber}}(\tau \| a) = \frac{a - \tau}{a(1-a)},$$

the unique optimum over $a \ge \rho$ is $a_* = \max(\rho, \tau)$. Thus

$$p^*_{\mathrm{CE}} = (a_*, (1 - a_*)q), \qquad \min_{p'_y \ge \rho} \mathrm{KL}(t \| p') = \mathrm{kl}_{\mathrm{Ber}}(\tau \| a_*). \tag{A.7}$$

This solution matches the conditional target without lowering the correct-class probability. It can still reduce an individual competitor margin, which the stronger requirement in Eq. (2) preserves. For the example in Figure 1, $p^*_{\mathrm{CE}} = (0.9, 1/24, 7/120)$ has unchanged CE, but its third-class probability exceeds the initial 0.03 and hence its third-class margin is smaller.

A.7 THE FINITE-COMPENSATION FRONTIER

Theorem 2 gives the budget needed for complete conditional transfer. With a smaller budget, the best partial transfer also has a closed form. Fix $B \ge 0$ as the maximum allowed increase in correct-class log-odds. The attainable conditional distributions are

$$\mathcal{Q}_B(r) = \left\{ x \in \mathbb{R}^{K-1}_{>0} : \sum_{j \ne y} x_j = 1, \quad x_j \le e^B r_j \ (j \ne y) \right\}. \tag{A.8}$$

Indeed, a prediction $(a, (1-a)x)$ preserves the jth margin exactly when $x_j / r_j \le \exp(\mathrm{logit}(a) - \mathrm{logit}(\rho))$. Any x in Eq. (A.8) is attained safely with $\mathrm{logit}(a) = \mathrm{logit}(\rho) + B$.

Define the least remaining conditional error by

$$\mathcal{V}(B) = \min_{x \in \mathcal{Q}_B(r)} \mathrm{KL}(q \| x).$$

Let $\lambda > 0$ enforce normalization. The unique minimizer x^* is positive and satisfies

$$x^*_j = \min\{e^B r_j, q_j / \lambda\}, \qquad \sum_{j \ne y} x^*_j = 1, \tag{A.9}$$

Moreover,

$$\mathcal{V}(0) = \mathrm{KL}(q \| r), \qquad \mathcal{V}(B) = 0 \iff B \ge D_\infty(q \| r). \tag{A.10}$$

The function $\mathcal{V}$ is nonincreasing and convex in B.

Proof. The closure of the feasible set is compact and contains the positive vector r. Since q is positive, the objective diverges when any coordinate approaches zero; a positive minimizer therefore exists. Strict convexity of $-\sum_j q_j \log x_j$ gives uniqueness. For $B > 0$, let $\mu_j \geq 0$ be the multipliers for $x_j \leq e^B r_j$. The KKT conditions are

$$-q_j/x_j + \lambda + \mu_j = 0, \qquad \mu_j \geq 0, \qquad \mu_j(x_j - e^B r_j) = 0.$$

At least one coordinate is uncapped because the caps sum to $e^B > 1$; hence $\lambda = q_j/x_j > 0$ there. Uncapped coordinates equal q_j/λ, while capped coordinates equal $e^B r_j$, yielding Eq. (A.9). At $B = 0$ the only feasible vector is r; any $0 < \lambda \leq \min_j q_j/r_j$ gives the same formula. Equation (A.10) follows from $\mathrm{KL}(q\|x) = 0$ exactly at $x = q$ and the feasibility of that vector.

The sets $\mathcal{Q}_B(r)$ grow with B, so $\mathcal{V}$ is nonincreasing. For convexity, define $Y = (Y_j)_{j \neq y}$ by $Y_j = \log(x_j/r_j)$ and use the equivalent problem

$$\mathcal{V}(B) = \min_{Y} \left\{ \mathrm{KL}(q\|r) - q^\top Y : \sum_j r_j e^{Y_j} \leq 1, \quad Y_j \leq B \right\}. \tag{A.11}$$

The normalization constraint is active at an optimum: otherwise a coordinate below its cap can be increased, strictly reducing the objective, since $q_j > 0$ and the caps sum to at least one. The objective is affine and the constraints are jointly convex in (Y, B). A convex combination of feasible minimizers at two budgets is feasible at their averaged budget and has the averaged objective, proving convexity of $\mathcal{V}$.

B Proofs supporting the safe correction and retained learning

These proofs support the update in Section 5.1, the clipped representation in Section 5.2, and Theorem 3. We use Euclidean projection and its Moreau decomposition (Parikh & Boyd, 2014).

B.1 The projection and task ordering

The general star-cone projection is useful for both TPKD and safe full KL. For an input $x \in \mathbb{R}^K$, sort its non-target coordinates as $x_{(1)} \leq \cdots \leq x_{(K-1)}$. With an empty prefix sum for $k = 0$,

$$\zeta = \min_{0 \leq k \leq K-1} \frac{x_y + \sum_{j=1}^{k} x_{(j)}}{k+1}, \qquad (\Pi_{\mathcal{K}_y} x)_y = \zeta, \quad (\Pi_{\mathcal{K}_y} x)_j = \max(x_j, \zeta). \tag{B.1}$$

For TPKD, $x = u$ has $u_y = 0$, so $\zeta \leq 0$. Computing the sorted prefixes costs $O(K \log K)$. To prove the formula, fix the projected target coordinate at ζ. The other optimal coordinates are $\max(x_j, \zeta)$, reducing the problem to

$$F(\zeta) = \tfrac{1}{2}(\zeta - x_y)^2 + \tfrac{1}{2} \sum_{j \neq y} [\zeta - x_j]_+^2.$$

Its derivative $F'(\zeta) = \zeta - x_y + \sum_{j \neq y} [\zeta - x_j]_+$ is continuous and strictly increasing. If k coordinates are below its root, then $(k+1)\zeta = x_y + \sum_{j \leq k} x_{(j)}$. The minimum-prefix expression selects precisely such a consistent root, including ties, and strict convexity gives uniqueness. This formula does not assume $x_y = 0$ when used for the full-KL gradient.

Nearest correction of the complete direction. Let $w \in \mathbb{R}^K$ be a candidate complete direction. The update v in Eq. (6) is the unique solution of

$$\underset{w \in \mathbb{R}^K:\ w - h \in \mathcal{K}_y}{\text{minimize}} \quad \frac{1}{2}\|w - (h + u/2)\|^2.$$

Indeed, the constraint preserves every margin gain of the corresponding label step, and translation by h followed by the positive homogeneity of projection onto the cone gives $v = h + \Pi_{\mathcal{K}_y}(u)/2$.

Full task ordering. From $h = (1 - \rho)(-1, r)$ we get $h_j - h_y = (1 - \rho)(1 + r_j) \geq 0$ for every $j \neq y$, so $h \in \mathcal{K}_y$. Since $d \in \mathcal{K}_y$ by construction and $v - h = d/2 \in \mathcal{K}_y$, for every $\eta \geq 0$,

$$M_j(z - \eta v) \geq M_j(z - \eta h) \geq M_j(z).$$

The cross-entropy ordering in Eq. (7) follows from $L_{\mathrm{CE}}(z, y) = \log(1 + \sum_{j \neq y} e^{-M_j(z)})$.

B.2 Clipped teacher and the tight compensation bounds

For the compensation in Eq. (9), let $\mathrm{TV}(q,r) = \frac{1}{2}\|q-r\|_1$, where $\|\cdot\|_1$ sums absolute coordinates. The tight size bounds are $\mathrm{TV}(q,r)/(K-1) \le \ell \le \mathrm{TV}(q,r)/2$: compensation is controlled by the mismatch between the two conditional distributions.

Proof. From the threshold form of the projection, Eq. (B.1), $d_y = -\ell$ and $d_j = \max(r_j - q_j, -\ell)$, and from $\mathbf{1}^\top d = 0$ (Appendix B.3),

$$\ell = \sum_{j\neq y} [q_j - r_j - \ell]_+, \qquad d_j = r_j - \min(q_j,\ r_j + \ell).$$

The function $g(a) = a - \sum_j [q_j - r_j - a]_+$ is continuous and strictly increasing. If $q \neq r$ then $g(0) = -\mathrm{TV}(q,r) < 0$ and $g(\mathrm{TV}(q,r)) > 0$; if $q = r$ the unique root is zero. Hence ℓ is unique. Summing the clipped mass gives $\sum_j \widetilde{q}_j = 1 - \ell$.

Let $T = \mathrm{TV}(q,r)$ and let $k \le K-2$ be the number of coordinates with positive deviation ($q_j > r_j$). The root equation gives

$$T = \ell + \sum_{q_j > r_j} \min(q_j - r_j,\ \ell) \ \le\ (k+1)\ell \ \le\ (K-1)\ell.$$

If $T > 0$, the set $A = \{j : q_j - r_j > \ell\}$ is nonempty; with $m = |A|$,

$$T \ \ge\ \sum_{j\in A}(q_j - r_j) = (m+1)\ell \ \ge\ 2\ell.$$

This proves the two compensation bounds. One positive deviation a shared as negative deviation by the other coordinates gives $\ell = T/2$; $K-2$ equal positive deviations with the last coordinate carrying all the negative deviation gives $\ell = T/(K-1)$. Sufficiently small perturbations of the uniform distribution make both examples strictly positive, so both bounds are tight, and $\ell < 1/2$.

Finally $d_j - d_y = [r_j - q_j + \ell]_+$. Substituting this into the margin change relative to the CE endpoint gives the extra-margin formula in Section 5.2.

B.3 Proof of Theorem 3

Let $n = K-1$. The Moreau decomposition gives $u = d + e$ with $e \in \mathcal{K}_y^\circ$, $d^\top e = 0$, and $\|u\|^2 = \|d\|^2 + \|e\|^2$. Vectors of the polar cone $\mathcal{K}_y^\circ$ have the form $(e)_y = \Lambda$, $(e)_j = -\lambda_j$, $\lambda_j \ge 0$, $\Lambda = \sum_{j\neq y}\lambda_j$. Put $Q = \sum_j \lambda_j^2 \le \Lambda^2$, so $\|e\|^2 = Q + \Lambda^2$. From $d^\top e = 0$ we have $u^\top e = \|e\|^2$; since $u_y = 0$ and $\sum_{j\neq y} u_j = 0$, the (λ_j) may be centered inside the inner product:

$$\|e\|^4 = (u^\top e)^2 = \Big(\sum_{j\neq y} u_j\big(\lambda_j - \tfrac{\Lambda}{n}\big)\Big)^2 \le \|u\|^2\left(Q - \frac{\Lambda^2}{n}\right) \le \frac{n-1}{2n}\,\|u\|^2\,(Q+\Lambda^2), \qquad \text{(B.2)}$$

where the last inequality is equivalent to $Q \le \Lambda^2$. If $e \neq 0$, dividing by $\|e\|^2$ gives $\|e\|^2 \le \frac{n-1}{2n}\|u\|^2$ and hence $\|d\|^2 \ge \frac{n+1}{2n}\|u\|^2 = \gamma_K \|u\|^2$; the case $e = 0$ is trivial. Finally $u^\top d = (d+e)^\top d = \|d\|^2$, which gives Eq. (10).

Pure conditional component. The polar representation gives $\mathbf{1}^\top e = 0$, and with $\mathbf{1}^\top u = 0$ we get $\mathbf{1}^\top d = 0$. Let $\ell = -d_y$; then $\sum_{j\neq y} d_j = \ell$, and $w_{\mathrm{cond}} = d - \ell a_0$ satisfies $(w_{\mathrm{cond}})_y = 0$ and $\mathbf{1}^\top w_{\mathrm{cond}} = 0$, i.e. $w_{\mathrm{cond}} \in \mathcal{L}_y$; moreover $a_0^\top d = \ell(1 + 1/n) = \ell\,\|a_0\|^2$ shows $w_{\mathrm{cond}} \perp a_0$. Since $u \perp a_0$, $u^\top w_{\mathrm{cond}} = u^\top d = \|d\|^2$. For $u \neq 0$, Cauchy–Schwarz gives

$$\|w_{\mathrm{cond}}\|^2 \ge \frac{(u^\top w_{\mathrm{cond}})^2}{\|u\|^2} = \frac{\|d\|^4}{\|u\|^2} \ge \gamma_K \|d\|^2 \ge \gamma_K^2 \|u\|^2 .$$

Tightness. Take $a > 0$, $u_y = 0$, one non-target coordinate $u_{j_*} = -a$, and the remaining $n-1$ coordinates equal to $a/(n-1)$. By Eq. (B.1), $\zeta = -a/2$, and the projection is $d_y = d_{j_*} = -a/2$, $d_j = a/(n-1)$ for $j \neq y, j_*$. A direct computation gives $\|d\|^2 / \|u\|^2 = (n+1)/(2n) = \gamma_K$ and $w_{\mathrm{cond}} = \gamma_K u$, so both bounds are attained simultaneously; for sufficiently small a this u is the difference of two strictly positive conditional distributions.

B.4 THE ALIGNMENT PART OF THEOREM 3

Keep $u = d + e$ and the polar representation. Since $u_y = 0$, $(e)_y = -d_y = \ell = \sum_j \lambda_j$, so $\|e\|^2 = \ell^2 + \sum_j \lambda_j^2 \le 2\ell^2$. Write $\gamma = \gamma_K$. From $w_{\text{cond}} \perp a_0$, $u \perp a_0$, and $\|a_0\|^2 = (n+1)/n = 2\gamma$,

$$\|d\|^2 = \|w_{\text{cond}}\|^2 + 2\gamma\ell^2, \qquad u^\top w_{\text{cond}} = \|d\|^2 .$$

Therefore

$$\begin{aligned}\gamma \|u\|^2 = \gamma \|d\|^2 + \gamma \|e\|^2 &\le \gamma \|d\|^2 + 2\gamma\ell^2 \\ &= \gamma \|d\|^2 + \|d\|^2 - \|w_{\text{cond}}\|^2 = (1+\gamma)\, u^\top w_{\text{cond}} - \|w_{\text{cond}}\|^2 , \quad \text{(B.3)}\end{aligned}$$

Completing the square in this inequality gives the equivalent form

$$\left\| w_{\text{cond}} - \frac{1+\gamma}{2} u \right\|^2 \le \frac{(1-\gamma)^2}{4} \|u\|^2 . \quad \text{(B.4)}$$

For $u \neq 0$ we have $w_{\text{cond}} \neq 0$, and $(1+\gamma) u^\top w_{\text{cond}} \ge \gamma \|u\|^2 + \|w_{\text{cond}}\|^2 \ge 2\sqrt{\gamma}\, \|u\|\, \|w_{\text{cond}}\|$ gives Eq. (11). Since $\gamma \in (1/2, 3/4]$ and $2\sqrt{\gamma}/(1+\gamma)$ is increasing on $(0, 1]$, the uniform lower bound is $2\sqrt{2}/3$.

B.5 THE GEOMETRIC EXAMPLE IN FIGURE 3

The illustration uses $K = 4$, with the correct class first, $r = (0.20, 0.30, 0.50)$ and $q = (0.30, 0.32, 0.38)$. These give

$$u = (0, -0.10, -0.02, 0.12), \qquad d = (-0.05, -0.05, -0.02, 0.12), \qquad \ell = 0.05.$$

With $a_0 = (-1, 1/3, 1/3, 1/3)$, the pure conditional component is

$$w_{\text{cond}} = (0, -1/15, -11/300, 31/300).$$

It is orthogonal to a_0 and forms an angle of approximately $11.54°$ with u. Panel (a) uses an oblique view to display both orthogonal components; panel (b) shows the angle within the conditional plane. This is a concrete illustration of the decomposition and alignment in Theorem 3.

B.6 A LOCAL LOSS REPRESENTATION OF THE CLIPPED TEACHER

The clipped direction in Eq. (9) has an exact local loss interpretation. Fix a reference output z_0, compute ℓ and $\widetilde{q}$ there, and hold both fixed. Since $\sum_j \widetilde{q}_j = 1 - \ell > 0$, the vector $\bar{q} = \widetilde{q}/(1-\ell)$ is a positive conditional distribution. Define

$$\Lambda_{z_0}(z) = (1-\ell)\, \mathrm{KL}(\bar{q} \| r(z)) - \ell \operatorname{logit} \rho(z). \quad \text{(B.5)}$$

Then

$$\nabla_z \Lambda_{z_0}(z)|_{z=z_0} = d(z_0). \quad \text{(B.6)}$$

The two terms learn the normalized clipped teacher and increase correct-class log-odds, respectively, with weights determined by the retained and clipped probability masses.

Proof. Using $\operatorname{logit} \rho(z) = z_y - \log \sum_{j \neq y} e^{z_j}$ and differentiating with the reference quantities frozen gives

$$\partial_{z_y} \Lambda_{z_0} = -\ell, \qquad \partial_{z_j} \Lambda_{z_0} = (1-\ell)(r_j(z) - \bar{q}_j) + \ell r_j(z) = r_j(z) - \widetilde{q}_j .$$

At $z = z_0$ these are precisely the coordinates of $d(z_0)$ in Eq. (9).

B.7 THE EXACT SIGNAL REMOVED AND THE SIZE OF COMPENSATION

The retention bound can be refined using the coordinates actually clipped. Let $\lambda = (\lambda_j)_{j \neq y}$ collect the clipped masses, defined by

$$\lambda_j = [q_j - r_j - \ell]_+, \qquad e = u - d = (\ell, -\lambda), \qquad \sum_{j \neq y} \lambda_j = \ell.$$

Moreau orthogonality gives the exact amount removed from the conditional first-order signal:

$$\|u\|^2 - u^\top d = \|u\|^2 - \|d\|^2 = \ell^2 + \sum_{j \neq y} \lambda_j^2. \tag{B.7}$$

For $u \neq 0$, let $m = |\{j : \lambda_j > 0\}|$. Then

$$\ell^2(1 + 1/m) \leq \|u\|^2 - \|d\|^2 \leq 2\ell^2. \tag{B.8}$$

The lower bound follows from Cauchy–Schwarz applied to the m positive coordinates of λ; the upper bound follows from their nonnegativity and sum ℓ.

The extra task-side change is also controlled by the current conditional mismatch. Write $T = \mathrm{TV}(q, r) = \frac{1}{2}\|q - r\|_1$ and let $\Delta M_j^{\mathrm{extra}}$ be the TPKD margin minus the corresponding CE-step margin. Because $r_j - q_j \leq T$ and $\ell \leq T/2$, Eq. (9) gives, for every $\eta \geq 0$,

$$0 \leq \Delta M_j^{\mathrm{extra}} = \frac{\eta}{2}[r_j - q_j + \ell]_+ \leq \frac{3\eta T}{4}. \tag{B.9}$$

Let $z' = z - \eta h$ and $\rho' = \mathrm{softmax}(z')_y$. Convexity of CE and the margin ordering imply

$$\begin{aligned} 0 \leq L_{\mathrm{CE}}(z') - L_{\mathrm{CE}}(z' - \eta d/2) &\leq \frac{\eta}{2} h(z')^\top d \\ &= \frac{\eta}{2}(1 - \rho') \sum_{j \neq y} r_j(z')(d_j - d_y) \\ &\leq \frac{3\eta}{4}(1 - \rho')T \leq \frac{3\eta}{4}(1 - \rho)T. \end{aligned} \tag{B.10}$$

Thus, at a fixed output step size, the additional confidence effect vanishes as the conditional mismatch vanishes.

B.8 Optimal conditional progress under two local budgets

The nearest-safe projection also maximizes conditional progress under two complementary comparisons.

Local quadratic model. Let $L > 0$ be a smoothness bound for L_{cond} and let x be an output displacement, so that $z^+ = z - x$. Smoothness gives

$$L_{\mathrm{cond}}(z) - L_{\mathrm{cond}}(z - x) \geq Q_u(x), \qquad Q_u(x) = u^\top x - \frac{L}{2}\|x\|^2.$$

Completing the square yields

$$Q_u(x) = \frac{\|u\|^2}{2L} - \frac{L}{2}\|x - u/L\|^2.$$

The unconstrained maximizer is u/L; over $x \in \mathcal{K}_y$ it is d/L, by positive homogeneity of cone projection. Consequently,

$$\max_{x \in \mathcal{K}_y} Q_u(x) = \frac{\|d\|^2}{2L}, \qquad \max_x Q_u(x) = \frac{\|u\|^2}{2L}. \tag{B.11}$$

For $u \neq 0$, the ratio of these optimal guaranteed decreases is $\|d\|^2/\|u\|^2 \geq \gamma_K$, with equality in the tightness construction of Appendix B.3.

Fixed displacement length. For $u \neq 0$ and a budget $S > 0$,

$$\max_{x \in \mathcal{K}_y,\, \|x\| \leq S} u^\top x = S\|d\|, \qquad x^* = S\frac{d}{\|d\|}. \tag{B.12}$$

To prove this, use $u = d + e$, where $e \in \mathcal{K}_y^\circ$: for any feasible x,

$$u^\top x = d^\top x + e^\top x \leq \|d\|\|x\| \leq S\|d\|.$$

The stated x^* attains equality. Without the cone constraint the optimum is $S\|u\|$, so the safe-to-unconstrained ratio is $\|d\|/\|u\| \geq \sqrt{\gamma_K}$, again tight. This is the comparison at equal displacement length; Eq. (10) compares the original and projected directions at equal step size.

C PROOFS SUPPORTING JOINT FITTING AND THE CE COMPARISON

We prove Theorem 4, including its residual rate, and the uniform-limit comparison stated in Section 5.3.

C.1 COMPLETE-UPDATE ALIGNMENT IN THEOREM 4

Write $u = d + e$ for the Moreau decomposition. Then $d^\top e = 0$, $e \in \mathcal{K}_y^\circ$, and Theorem 3 implies

$$\|e\|^2 \le \frac{1-\gamma_K}{\gamma_K} \|d\|^2 .$$

We have $h \in \mathcal{K}_y$, so $h^\top e \le 0$. Also, using the CE form $h = (1-\rho)(-1, r)$,

$$h^\top d = (1-\rho) \sum_{j \ne y} r_j (d_j - d_y) \ge 0.$$

Set $g_\Phi = 4h + u$ and abbreviate $\gamma = \gamma_K$. Expanding gives

$$\begin{aligned} g_\Phi^\top v - \|v\|^2 - \frac{\gamma}{4} \|g_\Phi\|^2 &= (3-4\gamma) \|h\|^2 + 2(1-\gamma) h^\top d + (1-2\gamma) h^\top e \\ &\quad + \frac{1}{4} \left[(1-\gamma) \|d\|^2 - \gamma \|e\|^2 \right] \ge 0. \end{aligned} \tag{C.1}$$

Every term is nonnegative because $1/2 < \gamma \le 3/4$. The arithmetic–geometric mean inequality gives

$$g_\Phi^\top v \ge \|v\|^2 + \frac{\gamma}{4} \|g_\Phi\|^2 \ge \sqrt{\gamma} \|g_\Phi\| \|v\| ,$$

which proves Eq. (12).

Sharpness. Fix any strictly positive pair r, q realizing the equality construction in Appendix B.3; for example, perturb the uniform conditional distribution by a sufficiently small one-negative-coordinate, equal-positive-remainder difference. Keep this $u = (0, r - q)$ fixed and let the student correct-class probability ρ tend to one. Then $h \to 0$, $g_\Phi \to u$ and $v \to d/2$, whence

$$\cos \angle(g_\Phi, v) \longrightarrow \frac{u^\top d}{\|u\| \|d\|} = \frac{\|d\|}{\|u\|} = \sqrt{\gamma}.$$

Thus no larger uniform angle constant holds for strictly positive probabilities. This construction also makes the gap in Eq. (C.1) tend to zero.

Batch form. Let col denote vertical stacking. For $W_\Phi = \mathrm{col}(g_{\Phi,i})/\sqrt{N}$ and $V = \mathrm{col}(v_i)/\sqrt{N}$, summing the per-example inequality yields

$$W_\Phi^\top V \ge \|V\|^2 + \frac{\gamma}{4} \|W_\Phi\|^2 \ge \sqrt{\gamma} \|W_\Phi\| \|V\| . \tag{C.2}$$

C.2 DESCENT, CONVERGENCE AND THE RESIDUAL RATE

The logit Hessians of the cross-entropy and conditional KL are softmax covariance matrices, with spectral norm at most $1/2$. Therefore $\Phi = 4L_{\mathrm{CE}} + L_{\mathrm{cond}}$ is $5/2$-smooth. For the specified output step,

$$\begin{aligned} \Phi(z - \eta v) &\le \Phi(z) - \eta g_\Phi^\top v + \frac{5}{4} \eta^2 \|v\|^2 \\ &\le \Phi(z) - \frac{\eta \gamma_K}{4} \|g_\Phi\|^2 - \eta \left(1 - \frac{5}{4}\eta\right) \|v\|^2 . \end{aligned} \tag{C.3}$$

This proves Eq. (13). When $0 < \eta \le 4/5$, the last term is nonpositive. For any integer $T \ge 1$, summing over $k = 0, \dots, T-1$ and using $\Phi \ge 0$ gives

$$\sum_{k<T} \|g_{\Phi,k}\|^2 \le \frac{4\Phi(z_0)}{\eta \gamma_K}. \tag{C.4}$$

Thus $g_{\Phi,k} \to 0$. Since $u_{k,y} = 0$ and $(g_{\Phi,k})_y = -4(1-\rho_k)$, we obtain $\rho_k \to 1$. The identity $h_k = (1-\rho_k)(-1, r_k)$ then gives $h_k \to 0$, so $u_k = g_{\Phi,k} - 4h_k \to 0$ and $r_k \to q$. Finally $\|d_k\| \le \|u_k\|$ implies $v_k \to 0$.

Joint and component residuals. Dividing Eq. (C.4) by T gives

$$\frac{1}{T}\sum_{k<T}\|g_{\Phi,k}\|^2 \le \frac{4\Phi(z_0)}{\eta\gamma_K T}. \tag{C.5}$$

The two learning signals inherit quantitative bounds. Since $h = (1-\rho)(-1, r)$ and $(g_\Phi)_y = -4(1-\rho)$,

$$\|h\|^2 = (1-\rho)^2(1+\|r\|^2) \le 2(1-\rho)^2 \le \frac{1}{8}\|g_\Phi\|^2.$$

Also $u = g_\Phi - 4h$ implies $\|u\|^2 \le 2\|g_\Phi\|^2 + 32\|h\|^2 \le 6\|g_\Phi\|^2$. Hence

$$\frac{1}{T}\sum_{k<T}\left(\|h_k\|^2 + \frac{1}{12}\|u_k\|^2\right) \le \frac{5\Phi(z_0)}{2\eta\gamma_K T}. \tag{C.6}$$

These bounds quantify the vanishing time-averaged squared gradient residuals of the joint, label and conditional signals.

C.3 The CE conditional limit stated in Section 5.3

Let $n = K-1$, $\mathbf{1}_n = (1,\dots,1) \in \mathbb{R}^n$, and $\nu = \mathbf{1}_n/n$. From finite initial logits, CE gradient flow and the fixed-step rule with $0 < \eta \le 1/20$ have limits $p_y \to 1$ and $r \to \nu$. This differs from the TPKD conditional limit only when $q \ne \nu$.

Proof. Let $z_{\neg y} \in \mathbb{R}^n$ contain the non-target logits, set $a = z_{\neg y} - (\mathbf{1}_n^\top z_{\neg y})\mathbf{1}_n/n$, and define on the zero-sum subspace

$$F(a) = \log\sum_j e^{a_j} - \frac{1}{n}\sum_j a_j - \log n = \mathrm{KL}(\nu\|r), \qquad \nabla F = r - \nu.$$

Let $t \ge 0$ denote flow time, with dots denoting time derivatives. Along the cross-entropy gradient flow, $\dot z_y = 1-\rho$ and $\dot z_j = -(1-\rho)r_j$, so

$$\dot a = -(1-\rho)\nabla F, \qquad \dot F = -(1-\rho)\,\|r-\nu\|^2,$$

which gives the conditional dynamics of CE. Let $s(t) = \int_0^t (1-\rho(\tau))\,d\tau$. If $s(\infty) < \infty$, every logit has finite total variation, all logits converge to finite values, and $1-\rho$ converges to a positive number, contradicting the finiteness of the integral; hence $s(t) \to \infty$. In the effective time s, $da/ds = -\nabla F$. The sublevel sets of F on the zero-sum subspace are bounded and its Hessian $\mathrm{diag}(r) - rr^\top$, where $\mathrm{diag}(r)$ is the diagonal matrix with diagonal r, is positive definite on that subspace, so on the compact sublevel set containing the whole trajectory there is $\mu > 0$ such that F is μ-strongly convex. The unique minimizer is $a = 0$, hence $r \to \nu$. Moreover $z_y = z_y(0) + s(t) \to \infty$ while all non-target logits are nonincreasing, so $\rho \to 1$.

In the discrete case let $s_k = \eta(1-\rho_k)$, so $a_{k+1} = a_k - s_k\nabla F(a_k)$. The gradient of F is $\frac{1}{2}$-Lipschitz, hence

$$F(a_{k+1}) \le F(a_k) - s_k(1 - s_k/4)\,\|\nabla F(a_k)\|^2.$$

For $0 < \eta \le 1/20$ the sequence stays in the same compact sublevel set. If $\sum_k s_k < \infty$, the same finite-total-variation argument gives a positive limit of $1-\rho_k$, a contradiction; hence $\sum_k s_k = \infty$. By strong convexity on that compact set and $\|\nabla F\|^2 \ge 2\mu F$,

$$F(a_{k+1}) \le \big[1 - 2\mu s_k(1 - s_k/4)\big]F(a_k),$$

so $F(a_k) \to 0$ and $r_k \to \nu$. At the same time $z_{k,y} = z_{0,y} + \sum_{j<k} s_j \to \infty$ and all non-target logits are nonincreasing, so $\rho_k \to 1$.

C.4 When a complete step lowers its own conditional error

The effect of CE on the teacher's conditional objective has the exact form

$$u^\top h = (1-\rho)(r-q)^\top r = \frac{1-\rho}{2}\left(\|r-q\|^2 + \|r\|^2 - \|q\|^2\right). \tag{C.7}$$

This follows by expanding $\|r-q\|^2$. The projected conditional contribution adds $\|d\|^2/2$, so for the complete direction $v = h + d/2$,

$$u^\top v = u^\top h + \tfrac{1}{2}\|d\|^2 \geq -\|h\|\|u\| + \frac{\gamma_K}{2}\|u\|^2. \tag{C.8}$$

In particular, $\|u\| > 0$ and $\|h\|/\|u\| < \gamma_K/2$ imply $c = u^\top v > 0$. The conditional KL has logit Hessian $\mathrm{diag}(r) - rr^\top$ on the non-target coordinates, with spectral norm at most $1/2$. Thus

$$L_{\mathrm{cond}}(z - \eta v) \leq L_{\mathrm{cond}}(z) - \eta c + \frac{\eta^2}{4}\|v\|^2.$$

Consequently the sufficient conditions

$$\|u\| > 0, \qquad \frac{\|h\|}{\|u\|} < \frac{\gamma_K}{2}, \qquad 0 < \eta < \frac{4u^\top v}{\|v\|^2} \tag{C.9}$$

give $L_{\mathrm{cond}}(z - \eta v) < L_{\mathrm{cond}}(z)$. The complete update then reduces its own conditional error while retaining the label-step margin gains in Eq. (7).

C.5 General positive conditional weights

For positive coefficients β, w, consider the family

$$v_\beta = h + \beta d, \qquad \Phi_w = wL_{\mathrm{CE}} + L_{\mathrm{cond}}.$$

For every $\beta > 0$, $v_\beta - h = \beta d \in \mathcal{K}_y$, so the label-step margin ordering holds for all $\eta \geq 0$. Moreover, v_β is the unique minimizer of $\frac{1}{2}\|a - (h + \beta u)\|^2$ over candidate directions $a \in \mathbb{R}^K$ with $a - h \in \mathcal{K}_y$.

A sufficient joint-learning condition is

$$4w\beta\gamma_K > 1. \tag{C.10}$$

To make the resulting constants explicit, set

$$c_{w,\beta} = \frac{w + \beta\gamma_K - \sqrt{(w - \beta\gamma_K)^2 + 1}}{2} > 0, \qquad C_\beta = \max\{1, \beta^2\}.$$

For a fixed positive conditional target and finite initial logits, every constant step

$$0 < \eta \leq \frac{c_{w,\beta}}{(w+1)C_\beta} \tag{C.11}$$

satisfies

$$\Phi_w(z - \eta v_\beta) \leq \Phi_w(z) - \frac{\eta c_{w,\beta}}{2}\left(\|h\|^2 + \|u\|^2\right). \tag{C.12}$$

It follows that

$$\frac{1}{T}\sum_{k<T}\left(\|h_k\|^2 + \|u_k\|^2\right) \leq \frac{2\Phi_w(z_0)}{\eta c_{w,\beta} T}, \qquad p_{k,y} \to 1, \quad r_k \to q, \quad v_{\beta,k} \to 0.$$

Proof. Write $H = \|h\|$ and $S = \|u\|$. The identities $h^\top d \geq 0$ and $u^\top d \geq \gamma_K\|u\|^2$ give

$$\nabla\Phi_w^\top v_\beta = w\|h\|^2 + w\beta h^\top d + u^\top h + \beta u^\top d \geq wH^2 - HS + \beta\gamma_K S^2.$$

The matrix of this quadratic form is $\left(\begin{smallmatrix} w & -1/2 \\ -1/2 & \beta\gamma_K \end{smallmatrix}\right)$. It is positive definite under Eq. (C.10), with smallest eigenvalue $c_{w,\beta}$, so the last expression is at least $c_{w,\beta}(H^2 + S^2)$. The potential is $(w+1)/2$-smooth and

$$\|v_\beta\|^2 \leq 2H^2 + 2\beta^2 S^2 \leq 2C_\beta(H^2 + S^2).$$

The descent lemma therefore gives

$$\Phi_w(z - \eta v_\beta) \leq \Phi_w(z) - \eta\left[c_{w,\beta} - \frac{(w+1)C_\beta\eta}{2}\right](H^2 + S^2),$$

which implies Eq. (C.12). Summation and $\Phi_w \geq 0$ show that $h_k, u_k \to 0$. The identity $h_{k,y} = p_{k,y} - 1$ gives $p_{k,y} \to 1$, while $u_k = (0, r_k - q)$ gives $r_k \to q$; finally $\|d_k\| \leq \|u_k\|$ yields $v_{\beta,k} \to 0$.

D PROOFS SUPPORTING EXACT REALIZATION AND PARAMETER LEARNING

This appendix proves Proposition 1, Theorem 5, and the measured-angle test stated in Section 5.4.

D.1 EXACT HEAD REALIZATION

For N frozen features of dimension D, absorb the classifier bias into the last feature column. Let $X \in \mathbb{R}^{N\times D}$ be the feature matrix, $A \in \mathbb{R}^{D\times K}$ the classifier weights, and $Z = XA$ the batch logits. Let V_{row} have rows $v_i^\top$, $X^\dagger$ denote the pseudoinverse, I_N the $N \times N$ identity, and ΔA the head-weight displacement. The exact construction in Proposition 1 is

$$\Delta A = -\eta X^\dagger V_{\mathrm{row}}, \qquad X(A + \Delta A) = Z - \eta V_{\mathrm{row}}. \tag{D.1}$$

Full row rank gives $XX^\dagger = I_N$, so $X(A - \eta X^\dagger V_{\mathrm{row}}) = Z - \eta V_{\mathrm{row}}$. Let $H \in \mathbb{R}^{D\times K}$ satisfy $XH = 0$. Every other feasible displacement is $-\eta X^\dagger V_{\mathrm{row}} + H$. The columns of $X^\dagger V_{\mathrm{row}}$ lie in the column space of $X^\top$ and are orthogonal to the columns of H; writing $\|\cdot\|_F$ for the Frobenius norm gives

$$\| - \eta X^\dagger V_{\mathrm{row}} + H\|_F^2 = \eta^2 \|X^\dagger V_{\mathrm{row}}\|_F^2 + \|H\|_F^2.$$

This proves the unique minimum-norm claim and the exact inheritance of the same step's output orderings.

Repeated realization on fixed features. Let the same full-row-rank X be used at every iteration, with fixed positive conditional targets for its rows. For $A_{k+1} = A_k - \eta X^\dagger V_{\mathrm{row},k}$,

$$Z_{k+1} = XA_{k+1} = Z_k - \eta V_{\mathrm{row},k}.$$

Thus every row follows the output iteration of Theorem 4. With finite initial logits and $0 < \eta \le 4/5$, each row has $p_{i,k,y_i} \to 1$ and $r_{i,k} \to q_i$.

Computational cost. For $N \le D$, forming and factorizing $XX^\top$ and applying $X^\dagger V_{\mathrm{row}}$ costs $O(N^2D + N^3 + NDK)$. With fixed X, its pseudoinverse can be reused, leaving $O(NDK)$ work per head displacement.

Parameter steps and batch objectives. Let $g_{\mathrm{T}} = N^{-1}\sum_i J_i^\top v_i$ and $g_{\mathrm{C}} = N^{-1}\sum_i J_i^\top (h_i + \ell_i a_{0,i}/2)$ be the respective parameter gradients of TPKD and compensation only. Define their conditional increment $\delta g = g_{\mathrm{T}} - g_{\mathrm{C}}$, the batch-average conditional loss $\overline{L}_{\mathrm{cond}}(\theta) = N^{-1}\sum_i L_{\mathrm{cond}}(z_i(\theta))$, and the batch-average joint potential $\Psi(\theta) = N^{-1}\sum_i \Phi(z_i(\theta))$. Under the respective conditions of Theorem 5, for sufficiently small $\eta > 0$ its conclusions are

$$\begin{aligned} &\text{(i)} \quad \delta g \neq 0, \qquad \overline{L}_{\mathrm{cond}}(\theta - \eta g_{\mathrm{T}}) < \overline{L}_{\mathrm{cond}}(\theta - \eta g_{\mathrm{C}}), \\ &\text{(ii)} \quad \Psi(\theta - \eta g_{\mathrm{T}}) < \Psi(\theta). \end{aligned} \tag{D.2}$$

D.2 THE ANGLE–SPECTRUM ARGUMENT

Batch signals and restricted spectra. Here col denotes vertical concatenation. For the batch Jacobian $J = \mathrm{col}(J_i)/\sqrt{N}$, define $G = JJ^\top$ and the normalized signal stacks

$$(U, R) = \frac{1}{\sqrt{N}}\big(\mathrm{col}(u_i), \mathrm{col}(w_{\mathrm{cond},i})\big), \qquad (W, V) = \frac{1}{\sqrt{N}}\big(\mathrm{col}(g_{\Phi,i}), \mathrm{col}(v_i)\big).$$

The first pair describes conditional learning and the second joint descent. Equation (8) yields

$$\delta g = J^\top R/2, \qquad \nabla_\theta \overline{L}_{\mathrm{cond}} = J^\top U, \qquad \nabla_\theta \Psi = J^\top W, \qquad g_{\mathrm{T}} = J^\top V.$$

For either pair let Q have orthonormal columns spanning its subspace. The smallest and largest eigenvalues of $Q^\top GQ$ are the squared-stretch extrema m, M for $J^\top$. Use (m_c, M_c) for $\mathrm{span}\{U, R\}$ and (m_Φ, M_Φ) for $\mathrm{span}\{W, V\}$. The batch map in Section 5.4 is $J^\top/\sqrt{N}$; its squared stretches are $m/N, M/N$, with the same ratio. Consequently $\chi_c = M_c/m_c$ and $\chi_\Phi = M_\Phi/m_\Phi$ when their minima are positive, and the ratios are infinite otherwise. The limits used in Theorem 5 are

$$\kappa_c = \left(\frac{1 + \sqrt{\gamma_K}}{1 - \sqrt{\gamma_K}}\right)^2, \qquad \kappa_\Phi = \frac{1 + \sqrt{\gamma_K}}{1 - \sqrt{\gamma_K}}. \tag{D.3}$$

With $K = 100$, these give $\kappa_c \approx 34.957650$ and $\kappa_\Phi \approx 5.912499$.

Let $a, b \neq 0$ have cosine c, and suppose the compression of $G = JJ^\top$ to their span has spectrum in $[m, M]$, $m > 0$. For $-1 < c < 1$, the unit bisectors

$$e_\pm = \frac{a/\|a\| \pm b/\|b\|}{\sqrt{2(1 \pm c)}}$$

are orthonormal. Symmetry of G cancels the cross terms and gives

$$\frac{a^\top G b}{\|a\|\|b\|} = \tfrac{1}{2}\left[(1+c)e_+^\top G e_+ - (1-c)e_-^\top G e_-\right] \geq \tfrac{1}{2}[(1+c)m - (1-c)M]. \tag{D.4}$$

Thus a positive measured cosine certifies $a^\top G b > 0$ whenever $M/m < (1+c)/(1-c)$. For $c = 1$, a, b are positively collinear and the lower bound is $m\|a\|\|b\| > 0$. This proves the measured-angle refinement described in Section 5.4. Using a universal lower bound on c gives the class-dependent tests instead.

D.3 Conditional increment and finite-step advantage

Assume $U \neq 0$ and let the conditional signal subspace have spectral bounds $0 < m_c \leq M_c$. Let $\delta g = g_\mathrm{T} - g_\mathrm{C} = J^\top R/2$. Averaging the per-example squared-norm bounds gives $\|R\|^2 \geq \gamma_K^2\|U\|^2$. To transfer the angle bound to the stacked signals, average the stronger inequality proved in Appendix B.4:

$$(1+\gamma_K)U^\top R \geq \gamma_K\|U\|^2 + \|R\|^2 \geq 2\sqrt{\gamma_K}\|U\|\|R\|. \tag{D.5}$$

This yields $R \neq 0$ and $\cos\angle(U, R) \geq c_K$. Therefore

$$\|\delta g\|^2 = \tfrac{1}{4}R^\top G R \geq \tfrac{m_c}{4}\|R\|^2 \geq \tfrac{m_c\gamma_K^2}{4}\|U\|^2 > 0. \tag{D.6}$$

Moreover, with $\beta_c = (1+c_K)m_c - (1-c_K)M_c$, Eq. (D.4) yields

$$t_c := \left\langle \nabla_\theta \overline{L}_\mathrm{cond}, \delta g \right\rangle = \tfrac{1}{2}U^\top G R \geq \tfrac{\beta_c}{4}\|U\|\|R\|.$$

The threshold $M_c/m_c < (1+c_K)/(1-c_K)$ makes $t_c > 0$. Since $c_K = 2\sqrt{\gamma_K}/(1+\gamma_K)$, this threshold is exactly κ_c in Eq. (D.3). At a differentiable parameter state,

$$\overline{L}_\mathrm{cond}(\theta - \eta g_\mathrm{T}) - \overline{L}_\mathrm{cond}(\theta - \eta g_\mathrm{C}) = -\eta t_c + o(\eta) < 0$$

for sufficiently small positive η, proving the conditional conclusion in Eq. (D.2). More explicitly, if the loss gradient is L_c-Lipschitz on both step segments, the difference is at most $-\eta t_c + L_c\eta^2(\|g_\mathrm{T}\|^2 + \|g_\mathrm{C}\|^2)/2$; choosing $\eta \leq t_c/[L_c(\|g_\mathrm{T}\|^2 + \|g_\mathrm{C}\|^2)]$ gives an upper bound of $-\eta t_c/2$.

D.4 Joint-potential descent

Assume $W, V \neq 0$ and let the joint signal subspace have spectral bounds $0 < m_\Phi \leq M_\Phi$. The batch form of Eq. (12) gives $\cos\angle(W, V) \geq s_K = \sqrt{\gamma_K}$. Applying Eq. (D.4) on $\mathrm{span}\{W, V\}$ gives

$$t_\Phi := \langle \nabla_\theta \Psi, g_\mathrm{T}\rangle = W^\top G V \geq \tfrac{1}{2}[(1+s_K)m_\Phi - (1-s_K)M_\Phi]\|W\|\|V\|.$$

Thus $M_\Phi/m_\Phi < \kappa_\Phi$ makes $t_\Phi > 0$, and $\Psi(\theta - \eta g_\mathrm{T}) = \Psi(\theta) - \eta t_\Phi + o(\eta) < \Psi(\theta)$ for sufficiently small positive η. This proves Eq. (D.2).

An explicit finite-step bound. Put

$$E = \frac{1}{N}\sum_i \left(\|h_i\|^2 + \frac{1}{12}\|u_i\|^2\right), \qquad \mu_\Phi = \frac{1}{2}[(1+\sqrt{\gamma_K})m_\Phi - (1-\sqrt{\gamma_K})M_\Phi].$$

Under $M_\Phi/m_\Phi < \kappa_\Phi$, $\mu_\Phi > 0$. The output inequalities give

$$(4h+u)^\top v \geq 4\|h\|^2 - \|h\|\|u\| + \tfrac{1}{4}\|u\|^2 \geq \|h\|^2 + \tfrac{1}{12}\|u\|^2,$$

where $\|h\|\|u\| \leq 3\|h\|^2 + \|u\|^2/12$ suffices. Hence $\|W\|\|V\| \geq W^\top V \geq E$, while $\|V\|^2 \leq 8E$. It follows that

$$t_\Phi \geq \mu_\Phi E, \qquad \|g_\mathrm{T}\|^2 = V^\top G V \leq M_\Phi\|V\|^2 \leq 8M_\Phi E.$$

If $\nabla_\theta\Psi$ is L_Ψ-Lipschitz on the current step segment, with $L_\Psi > 0$, then

$$\Psi(\theta - \eta g_\mathrm{T}) \leq \Psi(\theta) - \eta\mu_\Phi E + 4L_\Psi M_\Phi \eta^2 E.$$

Consequently,

$$0 < \eta \leq \frac{\mu_\Phi}{8M_\Phi L_\Psi} \quad\Longrightarrow\quad \Psi(\theta - \eta g_\mathrm{T}) \leq \Psi(\theta) - \frac{\eta\mu_\Phi}{2}E. \tag{D.7}$$

D.5 Computing the two-dimensional certificates

For either output pair, orthonormalize its span to obtain $Q = [\xi_1, \xi_2]$, omitting a dependent column. Compute $b_i = J^\top \xi_i$ by vector–Jacobian products; the compressed matrix has entries $(Q^\top G Q)_{ij} = b_i^\top b_j$. Its smallest and largest eigenvalues are the required m, M. No full Jacobian or kernel matrix is formed. The direct conditional and joint inner products are $\langle J^\top U, J^\top R/2\rangle$ and $\langle J^\top W, J^\top V\rangle$, respectively. Each pair uses its own subspace, spectrum and measured angle.

E Experimental protocols

This appendix specifies the training and measurements reported in Section 6.

E.1 Models, data, optimization and evaluation

Vision. All visual experiments use CIFAR-100 with 50,000 training images and 10,000 test images, a VOLO-D2 teacher and a PiT-B student. The student combines a publicly pretrained backbone with a classifier fitted on features of the 50,000 training images by standardized logistic regression with regularization parameter $C = 0.01$. All full-training methods use the same initial assets. Inputs use eight fixed augmented views: pad each 32-pixel image by four pixels, crop and horizontally flip, then bicubic resize to 224×224.

Full training uses 60 epochs, effective batch size 64 (microbatch 32), native automatic mixed precision (AMP), and SGD with momentum 0.9 and constant learning rate 10^{-4}, without warmup or decay. Weight decay is 10^{-3}, except for KL-Dist and XE-KL, whose recorded baseline configurations use 10^{-4}. Every method is evaluated at epoch 60.

Text. CLINC150 has 150 in-domain classes. We combine its original 15,000 training examples and original 3,000 validation examples into one fixed 18,000-example training pool. The official 4,500-example in-domain test set remains separate. All methods use this same split and a fixed final-epoch evaluation.

The teacher is the existing five-epoch BERT-large. The student uses a publicly pretrained BERT-Mini backbone and a shared random classifier, initialized with truncated normal weights (standard deviation 0.02, limits ± 0.04) and zero bias. The head is trained from the first update. Inputs have maximum length 128 with fixed padding and an attention mask; the classifier receives the pretrained pooler output.

Training uses four epochs, 1,125 optimizer iterations, effective batch size 64 (microbatch 32), FP32 and disabled TF32. Google AdamWeightDecay uses moment-decay coefficients $(\beta_1, \beta_2) = (0.9, 0.999)$ and numerical stabilizer $\epsilon = 10^{-6}$, weight decay 0.01, no bias correction, and global gradient-norm clipping at 1.0. Bias and LayerNorm parameters are excluded from weight decay. With zero-indexed iteration k, the shared learning-rate schedule is

$$\eta_k = 3 \times 10^{-4} \begin{cases} k/112, & k < 112, \\ 1 - k/1125, & 112 \le k < 1125. \end{cases}$$

Thus 3×10^{-4} is the base rate of the common warmup and decay schedule.

Method-specific objectives and temperatures. The comparisons align data, teacher–student pair, initialization, training budget and evaluation within each domain. The loss used by each baseline retains its method-specific definition and prescribed scale. In particular, temperature is part of that definition, rather than a common training-budget parameter. The following choices are fixed in the reported runs.

KD and DKD use temperature $T = 4$, matching the KD and DKD settings in the official DKD implementation (Zhao et al., 2022). Its configuration declares `KD.TEMPERATURE=4` and `DKD.T=4.0`.[1] For DKD, target and non-target weights are 1 and 8, respectively; the recorded method warmup is retained (two epochs in text). Standard KD uses distillation coefficient 1 in the shared training setup.

[1] Official configuration: https://github.com/megvii-research/mdistiller/blob/master/mdistiller/engine/cfg.py.

DHKD uses temperature 2 for its binary-KL objective, following its released training commands (Yang et al., 2025); the official ImageNet command explicitly specifies `-BinaryKL_T 2`.[2] This is the temperature of DHKD's binary-KL loss, rather than a replacement of its objective by the KD loss.

The recorded KL-Dist, XE-KL, DP-U, DP-S and DTO-KD objectives use unit logit scale ($T = 1$ where a temperature argument is present). TPKD also uses the original, unit-temperature conditional probabilities, with coefficient $1/2$. CE has no distillation temperature. Text DTO-KD retains its multi-layer feature-distillation implementation. These objective settings are unchanged across the three seeds; the five component controls use the TPKD settings and alter only the direction shown in Table 5.

Accuracy reporting. All full-training entries use seeds 42, 43 and 44. Within each domain, the comparisons share initial assets, training objects, data-order rules and the final-epoch evaluation. Tables 4–5 report test accuracy in percent as the mean and sample standard deviation over $n = 3$ seeds (denominator $n - 1$). The final epoch is fixed before training; no validation, observation or best-test checkpoint is used for selection.

KL-Dist and DP-U on CLINC150. In the reported CLINC150 runs, DP-U selected the teacher target at a rate of 100% in every epoch for all three seeds (42, 43 and 44), making its training objective identical to that of KL-Dist in these runs. Under the shared initialization, data order and optimization settings, we verified that the final logits were elementwise identical between the two methods for each seed. This explains their identical per-seed accuracies and the same mean and sample standard deviation of $93.38 \pm 0.19\%$ in Table 4.

E.2 Blocking and exact head updates

Blocking (Table 1). CE20, CE40 and CE60 are fixed seed-42 CE states under the visual protocol. Each uses 2,000 predetermined images with eight fixed views. The blocking criterion is $p_j \le t_j$ for every $j \neq y$. Conditional KL and D_∞ are evaluated from log probabilities within the blocked set. Positive conditional error is checked at tolerance 10^{-12}; all 16,000 image–view pairs enter the measurement.

Exact updates (Table 2). Freeze the backbone features and absorb the head bias into the feature matrix. The head, direction calculation and SVD pseudoinverse use FP64. Each candidate starts from the same model and batch with output step $\eta = 0.01$. The row-stacked direction specifies each example's output displacement, so it is not divided by batch size. The full-row-rank solve uses the pseudoinverse directly without ridge regularization.

The five directions are CE h, compensation only $h + \ell a_0/2$, unprojected conditional learning $h + u/2$, TPKD $h + d/2$, and CE plus full KL $h + (p - t)$. The last uses the original teacher probabilities with coefficient 1. Two states and 32 batches of 64 examples per state give 320 candidate steps, with two repeat checks. Let z^+_{CE} and z^+_{TPKD} denote the respective endpoint logits. The paired gain is $G_{\mathrm{CE}} = L_{\mathrm{cond}}(z^+_{\mathrm{CE}}) - L_{\mathrm{cond}}(z^+_{\mathrm{TPKD}})$; table losses and changes are means over examples.

Retention is $u^\top d/\|u\|^2$ and the complete-update cosine is $\cos \angle(g_\Phi, v)$. We also verify the alignment residual $R_h = g_\Phi^\top v - \|v\|^2 - \gamma_K \|g_\Phi\|^2/4$ and the finite-step descent bound in Eq. (13). A tolerance of 10^{-12} is used for inequalities and margin-order checks. The largest output execution error and CE-reference margin discrepancy are both 1.42×10^{-14}, supporting the numerical-precision statement in Section 6.

E.3 Backpropagation conditions and native paired updates

Table 3 uses sixteen fixed batches at each of CE20 and CE60. TPKD uses $h + d/2$; compensation only uses $h + \ell a_0/2$, preserving the common confidence contribution while deleting the pure conditional component. Batch-average losses and the $1/\sqrt{64}$ normalization in Appendix D are used throughout.

Network and vector–Jacobian calculations use FP32, with FP64 geometric calculations and diagnostic microbatch size eight. The two restricted spectra are computed as in Appendix D.5, with relative

[2] Official DHKD training commands: https://github.com/penghui-yang/DHKD.

rank tolerance 10^{-10} and spectral tolerance 10^{-12}. For 100 classes, the theoretical constants are $\gamma_{100} = 100/198$, $\sqrt{\gamma_{100}} \approx 0.7107$, $\kappa_c \approx 34.96$ and $\kappa_\Phi \approx 5.91$.

Each native candidate restores the same model, optimizer history, AMP scaler, buffers and random state, then executes the usual training iteration with microbatch size 32 and learning rate 10^{-4}. Let θ_{C}^{+} and θ_{T}^{+} denote the resulting compensation-only and TPKD parameter endpoints. Their paired gain is

$$G_{\mathrm{native}} = \overline{L}_{\mathrm{cond}}(\theta_{\mathrm{C}}^{+}) - \overline{L}_{\mathrm{cond}}(\theta_{\mathrm{T}}^{+}).$$

FP32 paired replay evaluates both endpoint losses consistently. The geometric tests establish a useful conditional parameter direction; the native pairs measure its additional benefit under the training optimizer.

E.4 The 128-iteration continuation

CE, compensation only and TPKD continue from the same CE60 state for 128 native-optimizer iterations at constant learning rate 10^{-4}. They follow the same sequence of 8,192 update images, recomputing their directions from their own current student states. A fixed 2,048-image observation set is disjoint from these updates. Both sets are drawn from the original training pool, and observation is used only to measure conditional learning. We record conditional KL at iterations 0, 1, 8, 32, 64 and 128. Section 6 reports its starting value and TPKD's endpoint advantages.

E.5 The five component controls

All five rows in Table 5 use the full-training protocols above. CE (h) and TPKD ($h + d/2$) use the corresponding main-table runs. Removing projection gives $h + u/2$. Removing the CE gradient gives $d/2$: it retains coefficient $1/2$, true-label indexing of the non-target classes, and the same safe cone. Safe full KL gives $h + \Pi_{\mathcal{K}_y}(p - t)/2$, using the full unit-temperature gradient and the generic projection in Eq. (B.1). Every control recomputes its direction at each iteration. The initialization, teacher, data order, optimization schedule and endpoint selection are unchanged.

E.6 Implementation of the prescribed direction

Keep the teacher fixed and detached. Compute $q = \mathrm{softmax}(z_{\neg y}^{T})$ and $r = \mathrm{softmax}(z_{\neg y})$ directly on non-target logits, form $u = (0, r - q)$, project to d, and set $v = h + d/2$. Detach v in Eq. (8); its batch-averaged surrogate supplies the complete parameter data gradient, including the label contribution. Log-softmax values are used for conditional losses and compensation statistics, avoiding division by small $1 - p_y$ values. The native optimizer then applies its usual update to the student parameters.