Two Mathematical Models of Knowledge Distillation

Audrey Xie
Stanford University

Ludwig Schmidt
Stanford University

John Duchi
Stanford University

Abstract

Many hypotheses compete to explain the successes of knowledge distillation. To help address this, we propose and analyze a mathematical model of distillation, which suggests that distillation's performance comes not from obtaining better models but from easier to optimize landscapes. For generalized linear models trained with stochastic gradient descent, we prove that distillation fits performant student models asymptotically more quickly than non-distilled models. In rank-1 matrix approximation, we characterize conditions on the target matrix under which gradient descent with distillation converges strictly faster than training on the supervised objective. The theory helps delineate the ways distillation provides benefits (i.e., in optimization speed, not in generalization), and experiments on real datasets corroborate the theoretical predictions.

1 INTRODUCTION

Knowledge distillation transfers the predictive ability of a larger teacher model to a smaller student model (Ba and Caruana, 2014; Hinton et al., 2015) by training the student on teacher predictions rather than ground-truth labels. It provides an important tool for obtaining performant models with reduced inference costs (Romero et al., 2015; Kim and Rush, 2016; Sanh et al., 2019; Beyer et al., 2022; Rivière et al., 2024; Sreenivas et al., 2024). Two perspectives compete to explain distillation's effectiveness: one attributes its benefits to the student's improved generalization ability (Mobahi et al., 2020; Menon et al., 2020; Dao et al., 2021; Das and Sanghavi, 2023; Pareek et al., 2024; Ildiz et al., 2025); the other emphasizes its algorithmic effects during optimization (Phuong and Lampert, 2019; Safaryan et al., 2023; Panigrahi et al., 2025).

Recent evidence suggests that there is still a gap in our understanding of knowledge distillation. In particular, Busbridge et al. (2025) show empirically that, given sufficient data, a student distilled from an optimally chosen teacher does not outperform the same model trained directly on ground-truth labels. This raises a fundamental question: if distillation does not improve the asymptotic solution, why does it often accelerate learning in practice?

In this work, we provide a theoretical explanation by analyzing knowledge distillation in two canonical settings: generalized linear models (GLMs) fit with maximum likelihood via stochastic gradient descent and rank-1 matrix approximation. Our analysis shows that although a student trained with distillation ultimately converges to the same solution as one trained with ground-truth supervision, distillation improves the efficiency of gradient descent algorithms. We test these theoretical predictions through experiments on synthetic data, regression tasks, and language model pretraining. Our contributions follow:

- **Online SGD for GLMs.** We model the teacher as the population risk minimizer over the full covariates and the student as having fewer parameters. We prove that distillation and direct supervision converge to the same solution (Theorem 3.1), but that distillation yields asymptotically more efficient SGD iterates, with the efficiency gap proportional to the label variance (Theorem 3.2).
- **Finite-dataset SGD.** Extending these results, we show that both trajectories converge to the same point on finite datasets (Theorem 3.3), and that, with high probability, distillation iterates have smaller asymptotic covariance (Theorem 3.4).
- **Rank-1 matrix approximation.** We model the teacher as the best rank-r Frobenius approxima-

Proceedings of the 29th International Conference on Artificial Intelligence and Statistics (AISTATS) 2026, Tangier, Morocco. PMLR: Volume 300.

tion to a target matrix. We prove that gradient descent applied to either the target or the teacher approximation converges to the same set of global optima (Theorem 4.1). Yet distillation converges faster under the sufficient condition that the matrix dimension scales inversely with the target error tolerance, raised to a spectrum-dependent constant (Theorem 4.2).

- **Experiments.** We test our theoretical predictions on synthetic data, regression with Fashion-MNIST (Xiao et al., 2017), and GPT-2 trained on C4 (Radford et al., 2019; Dao et al., 2021).

2 PRELIMINARIES

We review knowledge distillation and formalize it for maximum likelihood estimation in generalized linear models and low-rank matrix approximation.

2.1 Knowledge Distillation

Knowledge distillation trains student models with teacher-generated labels. For a teacher model trained with standard (ground truth) labels y, distillation fits a student model by minimizing a per-sample loss taken to be a convex combination of the losses between the student's predicted label y_S and either the teacher's predicted label y_T or the standard label y,

$$\xi \cdot \ell(y_T, y_S) + (1 - \xi) \cdot \ell(y, y_S)$$

where ℓ is a loss function (e.g., squared error, negative log likelihood) and $\xi \in [0, 1]$. The constant ξ influences the degree to which the student attempts to match the teacher's label, and in our analyses we take $\xi = 1$, the *pure knowledge distillation* setting.

To formalize our model for knowledge distillation, we first define the relevant spaces. Let the input-label space be $\mathcal{X} \times \mathcal{Y}$, and let the student input space $\mathcal{V}$ be identified with a subspace of $\mathcal{X}$. Denote the teacher and student model parameter spaces by $\mathcal{T}$ and $\mathcal{S}$, respectively. The teacher model, which we assume is given, minimizes an objective $L_T : \mathcal{T} \to \mathbb{R}$.

The teacher generates distilled labels via the mapping $h : \mathcal{T} \times \mathcal{X} \to \mathcal{Y}'$, which represents the prediction on input $x \in \mathcal{X}$. The student model then minimizes an objective function $L_S : \mathcal{S} \to \mathbb{R}$, which typically aggregates per-sample losses on the domain $\mathcal{S} \times \mathcal{V} \times (\mathcal{Y} \cup \mathcal{Y}')$ over the dataset.

2.1.1 Related Work

Along with studying distillation's influence on student generalization[1], both theoretical and empirical works have also analyzed its algorithmic effects. Phuong and Lampert (2019) consider distillation applied to linear models under gradient flow, the continuous time version of gradient descent. They show that distillation biases the optimization trajectory towards favorable, low generalization error solutions and that data geometry influences the rate of convergence.

Panigrahi et al. (2025) identify a mechanism by which progressive distillation, where labels are produced by increasingly capable teacher models, accelerates the student's training. They prove sample complexity bounds for sparse parity and empirically test this phenomenon on language models for probabilistic context-free grammars and other natural language modeling tasks.

Safaryan et al. (2023) analyze stochastic gradient descent under self-distillation and under distillation with unbiased compression operators. They show that distillation acts as a form of partial variance reduction, with the degree of reduction governed by the teacher's proximity to the optimum, and they lift this observation to upper bounds on the expected optimization error. In contrast, our analysis also considers students with strictly fewer parameters than the teacher, and we quantify the relative efficiency of knowledge distillation over supervised learning. Moreover, in the setting of rank-1 matrix approximation, we establish both upper and lower bounds, thereby demonstrating a concrete convergence gap.

2.2 MLE in Generalized Linear Models

We analyze stochastic gradient descent for generalized linear models (Nelder and Wedderburn, 1972) fit with maximum likelihood estimation. In a GLM, conditional on the covariate $x \in \mathbb{R}^p$, the response $y \in \mathbb{R}$ follows a distribution in the exponential family. Specifically, the conditional density takes the form,

$$p_\theta(y|x) = \exp\left(\frac{y(x^T\theta) - b(x^T\theta)}{a(\phi)} + c(y, \phi)\right).$$

where $\theta \in \mathbb{R}^p$ is the true parameter, $\phi > 0$ is a constant, and a, b, and c are known functions. The function $a(\phi)$ controls the dispersion, which we assume $a(\phi) = 1$. As written, the model uses the canonical link function, $g(\mathbb{E}[y|x]) = x^T\theta$, with inverse $h = g^{-1}$ mapping the linear predictor to the conditional mean of the response.

We fit the GLM by maximum likelihood estimation, so the per-sample loss function is the negative log-likelihood. With a slight overload of notation, we define the teacher ℓ_T and student ℓ_S per-sample losses

[1] See Appendix A for additional related works.

as,

$$\ell_T(\theta; x, y) = -\log p_\theta(x|y),$$
$$\ell_S(\gamma; v, y) = -\log p_\gamma(v|y),$$

where θ, $x \in \mathbb{R}^p$ and γ, $v \in \mathbb{R}^d$ for $0 < d < p$. Data points $(x, y) \sim^{iid} \mathcal{P}$ come from a distribution over the input-label space, and $\{(x_i, y_i\}_{i=1}^n$ forms finite sample of size n. We define the projection onto the top-d coordinates as $\Pi = [I_d\ 0]$, so the student observes $v = \Pi x$. The teacher's population and empirical objectives are

$$L_T(\theta) = \mathbb{E}[\ell_T(\theta; x, y)],$$
$$L_T^n(\theta) = \frac{1}{n}\sum_{i=1}^n \ell_T(\theta; x_i, y_i),$$

and the student's corresponding objectives are

$$L_S(\gamma) = \mathbb{E}[\ell_S(\gamma; v, y)],$$
$$L_S^n(\gamma) = \frac{1}{n}\sum_{i=1}^n \ell_S(\gamma; v_i, y_i).$$

Let θ^* (or $\hat{\theta}$) denote a teacher model minimizing its objective, population risk L_T (or empirical risk L_T^n). Under knowledge distillation, the student observes the teacher's prediction $y' = h(x^T\theta^*)$ (or $y' = h(x^T\hat{\theta})$) as opposed to the standard label y. The distillation loss retains the form of ℓ_S, but we replace the observed response with the teacher's prediction. In the population setting, this prediction satisfies $y' = \mathbb{E}[y|x]$, the conditional expectation of the response.

2.2.1 Notation

We use α to denote the student estimator fit with true labels y, and β for the one fit with distillation labels y'. When referring to either estimator in a context where distinction is irrelevant, we use γ. Let the input-label pair be $z = (v, y)$ and $z' = (v, y')$. Use $L'_S(\gamma)$ and $L'^n_S(\gamma)$ to denote the student population and empirical objectives with distillation labels.

2.3 Low Rank Approximation of a Matrix

We then study the problem of finding a low-rank approximation of a positive semi-definite matrix $M \in S_+^n$. Specifically, we seek the rank-r matrix (for $0 < r < n$) that minimizes the Frobenius norm of its difference from M. The solution corresponds to truncating the spectral decomposition of M to its top-r eigenvalues. Burer and Monteiro (2003) introduce a nonconvex formulation that implicitly enforces both the rank and positive semi-definiteness constraints by optimizing over the factorized form XX^T for "tall and skinny" matrices X. We define the teacher L_T and student L_S objective functions as,

$$L_T(X) = \frac{1}{4}||XX^T - M||_F^2,$$
$$L_S(x) = \frac{1}{4}||xx^T - M||_F^2$$

where $X \in \mathbb{R}^{n\times r}$ and $x \in \mathbb{R}^n$. The teacher X^* minimizes L_T, and generates targets via the mapping $h : \mathbb{R}^{n\times r} \to S_+^n$ defined by $h(X) = XX^T$. In the distillation setting, we replace the original matrix M with $h(X^*)$ in the student's objective.

Note that the student objective L_S has critical points at $\pm\sqrt{\lambda_i}v_i$, where (λ_i, v_i) are eigenvalue-eigenvector pairs of either M or $h(X^*)$, depending on the target. Gradient descent avoids spurious critical points, since the Hessian at such points has a direction of negative curvature. This follows from results by Lee et al. (2016); Panageas and Piliouras (2017), which we review in our proof of Theorem 4.1.

2.3.1 Notation

Eigenvalues of M are ordered $\lambda_1 > \lambda_2 \geq \ldots \geq \lambda_n \geq 0$. The spectral gap $\Delta_i := \lambda_1 - \lambda_i$ for $i = 2, \ldots, n$. Let $M' = h(X^*)$ denote the distillation target. We use w to denote the student estimator with target matrix M, and u for the student with teacher-generated target M'. When referring to both, we use x. Let $L'_S(x)$ denote the objective function with target matrix M'.

3 GENERALIZED LINEAR MODELS

In this setting, we aim to recover the unique maximizer γ^* of the likelihood function of a generalized linear model. We study how knowledge distillation influences stochastic gradient descent in this problem.

We defer all proofs from this section to Appendices B and C.

3.1 Online Setting

In this section, we focus on the following online stochastic gradient descent (SGD) update rule. At step t, given a sample $(x_t, \tilde{y}_t) \sim \mathcal{P}$ with $v_t = \Pi x_t$ and $\tilde{y}_t \in \{y_t, y'_t\}$, we update

$$\gamma_t = \gamma_{t-1} - \eta_t \nabla \ell_S(\gamma_{t-1}; v_t, \tilde{y}_t) \tag{1}$$

starting from $\gamma_0 \in \mathbb{R}^d$. We then analyze the dynamics of the averaged iterate,

$$\bar{\gamma}_t = \frac{1}{t}\sum_{k=0}^{t-1} \gamma_k$$

which achieves the optimal rate of $O(1/\sqrt{n})$ (Polyak and Juditsky, 1992). To establish consistency and asymptotic normality of $\bar{\gamma}_t$, we rely on the following assumptions. We state a simplified version of the assumption below, the full technical conditions are deferred to Appendix B.1.

Assumption 3.0.1 (Smoothness and convexity)**.** *The population losses L_S, L'_S are strongly convex around their minimizers α^* and β^*, respectively. The per-sample losses $\ell_S(\alpha; z)$ and $\ell_S(\beta; z')$ are twice differentiable with locally Lipschitz gradients and Hessians, and their gradients have bounded variance.*

Assumption 3.0.2 (Step sizes)**.** *The sequence of step sizes satisfy $(\eta_t - \eta_{t+1})/\eta_t = o(\eta_t)$, $\eta_t > 0$ for all t, and*

$$\sum_{t=1}^{\infty} \eta_t \cdot t^{-1/2} < \infty.$$

Theorem 3.1 (Population convergence)**.** *Under Assumptions 3.0.1 and 3.0.2, let $\bar{\alpha}_t$ denote the averaged iterate from SGD with gradient estimates $\nabla \ell_S(\alpha_t; v_t, y_t)$, and let $\bar{\beta}_t$ denote the averaged iterate from SGD with $\nabla \ell_S(\beta_t; v_t, h(x_t^T \theta^*))$. Then both sequences converge almost surely to*

$$\gamma^* := \operatorname*{argmin}_{\gamma \in \mathbb{R}^d} \; L_S(\gamma)$$

Having established that supervised learning and knowledge distillation converge to the same limiting parameter, we now show that the SGD iterates obtained from knowledge distillation, $\bar{\beta}_t$, are asymptotically more efficient than the ones obtained from supervised learning, $\bar{\alpha}_t$.

Theorem 3.2. *Suppose Assumptions 3.0.1 and 3.0.2 hold. Let γ^* be the minimizer of L_S, and define the Hessian at γ^**

$$H = \mathbb{E}[h'(v^T \gamma^*) v v^T]$$

where h is the inverse link function and $v = \Pi x$. Define the second moments of the gradient noise at γ^*,

$$C(\gamma^*) = \mathbb{E}[(h(v^T \gamma^*) - y)^2 v v^T],$$
$$C'(\gamma^*) = \mathbb{E}[(h(v^T \gamma^*) - h(x^T \theta^*))^2 v v^T]$$

Let $\bar{\alpha}_t$ be the averaged iterate under supervised learning, obtained from SGD with gradient estimates $\nabla \ell_S(\alpha_t; v_t, y_t)$. Let $\bar{\beta}_t$ be the averaged iterate under knowledge distillation, obtained from SGD with $\nabla \ell_S(\beta_t; v_t, h(x_t^T \theta^*))$. Then as $t \to \infty$,

$$\sqrt{t}(\bar{\alpha}_t - \gamma^*) \xrightarrow{d} \mathcal{N}(0, H^{-1} C(\gamma^*) H^{-1}).$$

and

$$\sqrt{t}(\bar{\beta}_t - \gamma^*) \xrightarrow{d} \mathcal{N}(0, H^{-1} C'(\gamma^*) H^{-1}).$$

Moreover, the asymptotic covariance under knowledge distillation is no larger than under supervised learning,

$$C(\gamma^*) - C'(\gamma^*) \succeq \mathbb{E}[\mathit{Var}(y|x) v v^T] \succeq 0,$$

which holds with equality only if the label variance is zero.

By Theorem 2 of Polyak and Juditsky (1992), both $\bar{\alpha}_t$ and $\bar{\beta}_t$ converge almost surely γ^* and are asymptotically normal. Assumptions 3.0.1 and 3.0.2 provide the technical conditions under which this classical stochastic approximation result applies.

The key intuition is that projecting onto the teacher's manifold before projecting onto the student's manifold does not change the final point. For example, in linear regression (identity link) with Gaussian noise, the maximum likelihood estimator corresponds to the orthogonal projection of the response onto the span of the covariates. Projecting first onto the span of the full p-dimensional covariates and then onto the span of the truncated d-dimensional covariates produces the same results as projecting directly onto the truncated covariates.

Knowledge distillation improves efficiency, $\bar{\beta}_t$ has a smaller asymptotic covariance than $\bar{\alpha}_t$, by removing the label variance component from the stochastic gradient estimate. The efficiency gain is most pronounced when the response has high variance, either due to label noise or the intrinsic variability of the distribution, while the benefit diminishes when the labels are nearly deterministic and individual samples already provide reliable information.

3.2 Finite Dataset Setting

The online SGD setting in the previous section samples a new observation at each step t, analogous to training with a single pass over the data. Here, we instead assume access to a finite dataset $D = \{(x_i, y_i)\}_{i=1}^n \sim \mathcal{P}^n$, where pairs are drawn i.i.d. from $\mathcal{P}$. Once realized, we treat D as fixed and compute each SGD update (Eq. 1) using a sample $(x_t, y_t) \sim U(D)$ chosen uniformly from D.

Under standard regularity conditions, the averaged iterates $\bar{\gamma}_t$ are consistent and asymptotically normal. Their behavior parallels the conclusions of Theorems 3.1 and 3.2, except that the empirical Hessian and gradient noise second moments replace their population counterparts.

We present simplified versions of the assumptions below. The full technical conditions are deferred to Appendix C.1.

Assumption 3.2.1 (Smoothness and convexity)**.** *The empirical losses L_S^n, $L_S'^n$ are strongly convex around*

their minimizers $\hat{\alpha}$ and $\hat{\beta}$, respectively. The per-sample losses $\ell_S(\alpha; z)$ and $\ell_S(\beta; z')$ are twice differentiable with locally Lipschitz gradients and Hessians, and their gradients have bounded growth.

Assumption 3.2.2 (Concentration)**.** *The covariates and labels satisfy standard regularity conditions, ensuring concentration of empirical quantities around their population counterparts.*

In particular, the conditional variance of y given x is uniformly bounded away from zero and per-sample gradient covariance matrices are uniformly bounded around γ^*.

Theorem 3.3 (Convergence under finite dataset)**.** *Under Assumptions 3.2.1 and 3.0.2, suppose we have a fixed dataset $D = \{(x_i, y_i)\}_{i=1}^n$, and at each iteration t the sample (x_t, y_t) is drawn uniformly at random from D (with replacement).*

Let $\bar{\alpha}_t$ denote the averaged iterate from SGD with gradient estimates $\nabla \ell_S(\alpha_t; v_t, y_t)$, and let $\bar{\beta}_t$ denote the averaged iterate from SGD with $\nabla \ell_S(\beta_t; v_t, h(x_t^T \hat{\theta}))$. Then both sequences converge almost surely to

$$\hat{\gamma} := \operatorname*{argmin}_{\gamma \in \mathbb{R}^d} \; L_S^n(\gamma).$$

As the number of SGD steps tends to infinity, both supervised learning $\bar{\alpha}_t$ and knowledge distillation $\bar{\beta}_t$ tend to the same minimizer of the empirical loss.

Theorem 3.4. *Suppose Assumptions 3.2.1, 3.2.2, and 3.0.2 hold. Let $\hat{\gamma}$ be the minimizer of the empirical loss L_S^n, and define the empirical Hessian at $\hat{\gamma}$,*

$$H_n = \frac{1}{n} \sum_{i=1}^n h'(v_i^T \hat{\gamma}) v_i v_i^T$$

where h is the inverse link function and $v_i = \Pi x_i$. Define the empirical second moments of gradient noise at $\hat{\gamma}$,

$$C_n(\hat{\gamma}) = \frac{1}{n} \sum_{i=1}^n (h(v_i^T \hat{\gamma}) - y_i)^2 v_i v_i^T,$$

$$C_n'(\hat{\gamma}) = \frac{1}{n} \sum_{i=1}^n (h(v_i^T \hat{\gamma}) - h(x_i^T \hat{\theta}))^2 v_i v_i^T$$

where $\hat{\theta}$ is the teacher estimator.

Let $\bar{\alpha}_t$ denote the averaged iterate under supervised learning, obtained from SGD with gradient estimates $\nabla \ell_S(\alpha_t; v_t, y_t)$. Let $\bar{\beta}_t$ denote the averaged iterate under knowledge distillation, obtained from SGD with $\nabla \ell_S(\beta_t; v_t, h(x_t^T \hat{\theta}))$. Then as $t \to \infty$,

$$\sqrt{t}(\bar{\alpha}_t - \hat{\gamma}) \xrightarrow{d} \mathcal{N}(0, H_n^{-1} C_n(\hat{\gamma}) H_n^{-1})$$

and

$$\sqrt{t}(\bar{\beta}_t - \hat{\gamma}) \xrightarrow{d} \mathcal{N}(0, H_n^{-1} C_n'(\hat{\gamma}) H_n^{-1}).$$

Moreover, for sufficiently large n,

$$C_n(\hat{\gamma}) - C_n'(\hat{\gamma}) \succeq 0$$

with high probability.

Analogous to the online SGD (population) setting, almost sure convergence of $\bar{\alpha}_t$ and $\bar{\beta}_t$ to $\hat{\gamma}$, together with their asymptotic normality, follows from Theorem 2 of Polyak and Juditsky (1992).

In the population case, the relative efficiency gap between knowledge distillation and supervised learning increases with the label variance. Theorem 3.4 establishes this relationship in the finite-dataset setting as well. Since the empirical second moments of the gradient noise are evaluated at the empirical minimizer $\hat{\gamma}$ rather than the population minimizer γ^*, the argument alternates between applying local Lipschitz continuity of the empirical second moments and using concentration bounds to control their deviation from the population counterparts at γ^*.

Thus, the efficiency of knowledge distillation from the population setting carries over to finite datasets. Under high label variance, distillation provides a tangible improvement, while in low-variance regimes its effect is negligible.

4 LOW RANK APPROXIMATION

In this section, we study the problem of finding the best rank-1 approximation of a positive semi-definite matrix $M \in S_+^n$. We formulate this as a nonconvex optimization problem (Burer and Monteiro, 2003) and investigate how knowledge distillation influences the convergence behavior of gradient descent.

We defer all proofs from this section to Appendix D.

4.1 Convergence Analysis

We analyze the convergence of gradient descent with constant step size η. At each step, we perform the update

$$x_{t+1} = x_t - \eta \nabla \widetilde{L}_S(x_t).$$

where $\widetilde{L}_S \in \{L_S, L_S'\}$. While the gradients ∇L_S and $\nabla L_S'$ differ, the update takes the same general form.

Theorem 4.1. *Let the step size be*

$$0 < \eta < \frac{1}{L \cdot (\lambda_1 + 1)}$$

where λ_1 is the largest eigenvalue of the target matrix M and $L \geq 4$ is a fixed constant. Sample $n_0 \sim \mathcal{N}(0, k\lambda_1/n \cdot I_n)$ with some $k \in (0,1)$, and condition on $||n_0||_2 \leq \sqrt{\lambda_1}$. Set the initial points of the gradient descent trajectories to be equal, $w_0 = u_0 = n_0$.

Let w_t denote the iterate from gradient descent on standard labels, L_S, and let u_t denote the iterate from distillation labels, L'_S. Then both sequences converge almost surely to

$$x^* \in \operatorname*{argmin}_{x \in \mathbb{R}^n} L_S(x).$$

When $r > 1$, the set of critical points of L_S and L'_S is strictly larger than the set of all global minimizers. In fact, their gradients are zero at any point of the form $\pm\sqrt{\lambda_i}v_i$ for $i \in [n]$ (or $i \in [r]$). To establish that the sequences $\{w_t\}$ and $\{u_t\}$ converge to a minimizer rather than a strict saddle, we show that an open convex set is forward-invariant under the gradient updates and then apply Theorem 3 of Panageas and Piliouras (2017).

Consequently, the limiting matrices satisfy $w^*(w^*)^T = u^*(u^*)^T$, even though the limit points w^* and u^* themselves may differ. We now analyze the gradient descent dynamics over a finite number of steps, focusing on the number of iterations required to achieve a prescribed error tolerance.

Definition 4.1.1. *For a tolerance $\varepsilon > 0$, define the first step when the gradient descent sequence $\{x_t\}$ enters the $(\lambda_1 \cdot \varepsilon)$-ball around the optimum x^* as*

$$T(x) := \min\{t : ||x_t - x^*||_2^2 < \lambda_1 \cdot \varepsilon\}$$

Definition 4.1.2. *For an iterate $x_t = (x_{1,t}, x_{2:n,t})$ expressed in the eigenbasis of M, define*

$$\textsc{Im}(x_t) := \frac{||x_{2:n,t}||_2^2}{x_{1,t}^2}.$$

Equivalently, $\sqrt{\textsc{Im}(x_t)}$ is the tangent of the angle between x_t and the top eigenvector, so $\textsc{Im}(x_t)$ quantifies how misaligned the iterate is with the leading direction.

Lemma 4.1.1 (Supervised learning[2])**.** *Given $\varepsilon > 0$, under the previous step size and random initialization conditions, the number of steps required by supervised learning satisfies a lower bound of the form*

$$T(w) \geq \frac{1}{2\log(1/(1-\eta\Delta_n))} \log\left(\frac{\textsc{Im}(x_0)}{\varepsilon}\right).$$

[2]Appendix D.2 contains a detailed statement of the lemma together with its proof.

Lemma 4.1.2 (Knowledge distillation[3])**.** *Given $\varepsilon > 0$, under the same conditions, the number of iterations required by knowledge distillation is bounded above by*

$$T(u) \leq \max\{t_a, t_b, t_\xi\} + t_\delta,$$

where $t_a, t_b, t_\xi, t_\delta$ depend logarithmically on the initialization $\textsc{Im}(x_0)$, tolerance, and spectrum.

The lower and upper bounds on $T(w)$ and $T(u)$ allow us to establish conditions under which gradient descent with the knowledge distillation objective converges faster than supervised learning.

Theorem 4.2. *Given $\varepsilon > 0$ and teacher matrix rank r, let the step size be*

$$0 < \eta < \frac{1}{L \cdot (\lambda_1 + 1)}$$

where λ_1 is the largest eigenvalue of the target matrix M and $L \geq 4$ is a fixed constant. Sample $n_0 \sim \mathcal{N}(0, k\lambda_1/n \cdot I_n)$ with some $k \in (0,1)$. Set the initial points of the gradient descent trajectories to be equal, $w_0 = u_0 = n_0$.

Assume furthermore that $\Delta_n < \ell\lambda_1$ and $(1 + \eta\lambda_1/2)(1 - \eta\Delta_n) > 1$ with $\ell = L/(L+1)$. Then there exists $c > 0$ (depending on the spectrum of M) such that for all $n = \Omega(\varepsilon^{-c})$, we have $T(w) - T(u) \geq 0$ with constant probability over the initialization.

The requirement that the dimension n grows as $\Omega(\varepsilon^{-c})$ may seem undesirable, but it also not unexpected. The top spectral gap Δ_2 controls the asymptotic rate of gradient descent, leading to a $O(1/\Delta_2)$ dependence. Knowledge distillation cannot improve this rate unless the teacher itself has rank $r = 1$, in which case the dependence reduces to $O(1/\lambda_1)$. Lemma 4.1.2 shows that the main advantage of knowledge distillation lies in accelerating the decay of coordinates $r+1$ through n. As the target tolerance ε decreases, maintaining this advantage requires increasing mass in these trailing coordinates.

5 EXPERIMENTS

We test our theory on both synthetic and real data. The code for our experiments can be found at https://github.com/0aax/two-models-kd.

5.1 Synthetic Experiments

Generalized linear models. For generalized linear models, our theory predicts that distillation improves

[3]Appendix D.3 contains a detailed statement of the lemma together with its proof.

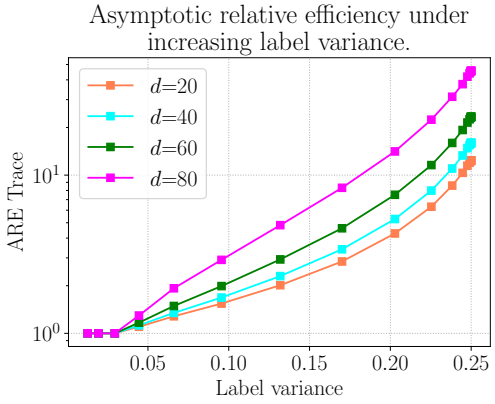

Figure 1: Asymptotic relative efficiency, given by the ratio of asymptotic covariance traces (supervised learning over knowledge distillation), as a function of label variance. The teacher dimension is fixed at $p = 1000$, and the student dimension is d. Each point represents an average over 100 sampled datasets and true parameters.

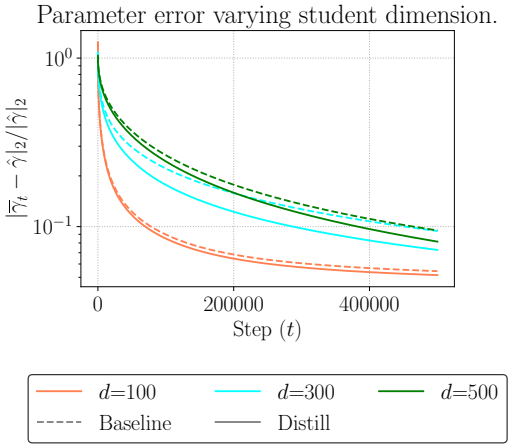

Figure 2: (Multinomial logistic regression, Fashion-MNIST) Normalized parameter error of SGD iterates under knowledge distillation and standard supervision for student dimensions $d \in \{100, 300, 500\}$. The initial step size η_0, optimal for standard supervision, is used for both methods. Results are averaged over 10 trials.

efficiency by removing label variance from the gradient estimates. To test this prediction, we construct a logistic regression dataset with $n = 1000$ samples and dimension $p = 100$ covariates, drawn i.i.d. from $\mathcal{N}(0, 1/p \cdot I_p)$. Labels follow the Bernoulli distribution with mean $\sigma(x_i^T \theta^*)$, where the scale of the true parameter $\theta^* = \kappa \cdot n_0$ (with $n_0 \sim \mathcal{N}(0, I_p)$, $\kappa \in [10^{-1}, 10^{2.5}]$) controls the label variance. We apply L-BFGS (Liu and Nocedal, 1989) to obtain the empirical minimizers $\hat{\alpha}$ and $\hat{\beta}$, and use them to compute the asymptotic covariance matrices of the supervised and distilled estimators. Figure 1 shows that distillation achieves higher relative efficiency as label variance and student dimension $d \in \{20, 40, 60, 80\}$ increase.

We then ask whether the asymptotic efficiency translates into fewer optimization steps. We construct the dataset in the manner detailed above, with $p = 10$ and $\kappa = 1.0$, and fit the teacher $\hat{\theta}$ using L-BFGS. Both supervised and distilled students start at the same point $\alpha_0 = \beta_0 = 0$ and we run SGD with step sizes $\eta_t = \eta_0/(1 + 10^{-3} t^{0.6})$ where $\eta_0 \in \{10^{-3}, 10^{-2.5}, \ldots, 10^{-0.5}, 1.0\}$, consistent with Assumption 3.0.2. Figure 4 (Appendix E.1) shows that, even when both methods use the same η_0 (optimal for standard supervision), knowledge distillation converges faster. Furthermore, this gap becomes more significant when distillation uses its own optimal η_0.

Rank-1 matrix approximation. Our analysis in Section 4 shows that distillation accelerates convergence when the teacher truncates a "heavy tail" of eigenvalues (i.e. when the removed eigenvalues are, in sum, relatively large compared to the top-r retained eigenvalues). To test this, we construct a diagonal target matrix of size $n = 500$ with a controlled spectrum, $M = \mathrm{diag}(\lambda_1, s_r, s_n)$, for arithmetic sequences s_r and s_n. We obtain the teacher matrix by truncating to the top-r eigenvalues, equivalent to zeroing out entries $r + 1$ through n on the diagonal.

Gradient descent starts at $x_0 \sim \mathcal{N}(0, 1/n \cdot I_n)$ and we measure progress via the distance to the global optimum $\pm\sqrt{\lambda_1} e_1$. We sweep step sizes $\eta \in \{10^{-2}, 10^{-1.5}, 10^{-1}, 10^{-0.5}, 1.0\}$ and select one which minimizes parameter error for supervised learning. In all settings, this is $\eta = 10^{-0.5}$. Figure 5 (Appendix E.2) shows that distillation accelerates convergence when the teacher rank approaches the student's rank and when it truncates relatively large eigenvalues.

5.2 Regression on Real Data

We next test our predictions on real-world data. On Fashion-MNIST (Xiao et al., 2017), we fit multinomial logistic regression models by flattening each 28×28 image decomposing it into its principal components. The teacher observes to the full set of components, while the student observes only a subset of the components ($d \in \{100, 300, 500\}$). Both supervised and distilled students start at the same initialization $\alpha_0^c = \beta_0^c = n_0^c$, where $n_0^c \sim \mathcal{N}(0, 1/d \cdot I_d)$ and $c = 1, \ldots, 10$. The batch size is $b = 100$ and step size $\eta_t = \eta_0/(1 + 10^{-3} t^{0.6})$, consistent with Assumption 3.0.2. We sweep step sizes

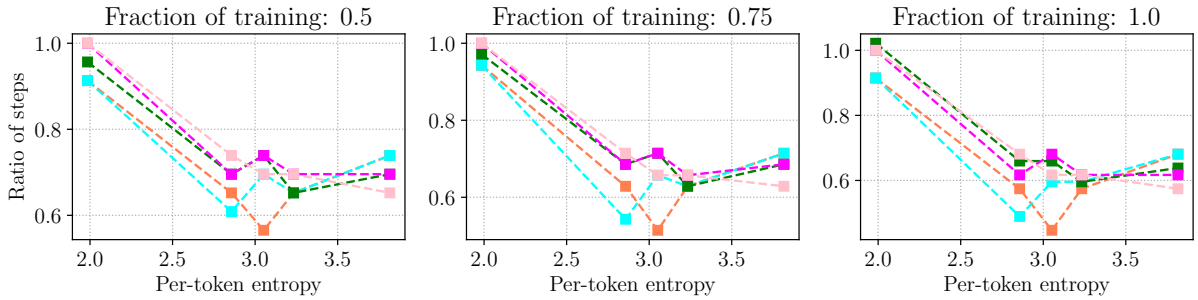

Figure 3: (Language model pretraining) Ratio of training steps required by knowledge distillation versus standard supervision across subsets of increasing entropy, evaluated at 0.5, 0.75, and 1.0 fractions of training steps. At each point, we record the maximum of the two validation losses, and measure how many steps each student needs to reach this loss. This is repeated over five trials.

$\eta_0 \in \{10^{-1}, 10^{-0.75}, 10^{-0.5}, 10^{-0.25}, 1.0\}$ and select the one which has the lowest parameter error for standard supervision. This ends up being: $\eta_0 = 10^{-0.75}$ for $d = 100$, $\eta_0 = 10^{-0.5}$ for $d = 300$, and $\eta_0 = 10^{-0.5}$ for $d = 500$. Figure 2 shows the normalized distance to the optimum over training steps. Distillation consistently converges in fewer steps across all student dimensions. We find the optimal step size η_0 for distillation to be larger in some cases than the optimal for standard supervision.[4]

5.3 Next Token Prediction

Finally, we move beyond simple regression and matrix approximation to study language model pretraining. While our GLM analysis suggests that distillation benefits setting with high label variance, since we cannot directly observe label variance in next-token prediction, we use per-token entropy of a model's predicted distribution as a proxy. We train a GPT-2 medium model on a subset of the C4 dataset (Raffel et al., 2020) and compute per-token entropy to partition the remaining data into subsets of increasing entropy.

For each subset, we train a GPT-2 small student model with different learning rates and select the model with the lowest validation loss as the supervised learning baseline. We then train the distillation student model with the same training setup as the supervised learning baseline. The only difference in the two setups being the training objective, the baseline minimizes cross-entropy with ground-truth labels, while the distilled student minimizes the KL-divergence $D_{\mathrm{KL}}(P_T||P_S)$ between teacher P_T and student P_S distributions, where the teacher GPT-2 medium model trains on the same subset.[5]

[4]Figure 6 (Appendix E.3) shows error under optimal η_0 for both methods.

[5]Appendix E.4 contains additional experimental details.

Figure 3 shows the ratio of training steps required by the student trained with knowledge distillation over the supervised learning baseline to reach a particular validation loss on each of the subsets that are increasing in entropy. Across all subsets, distillation reaches the target loss in fewer steps, with the improvement largest in high-entropy subsets.

6 CONCLUSION

In this work, we analyze knowledge distillation in two settings, generalized linear models fit with maximum likelihood estimation and rank-1 matrix approximation. For generalized linear models, we show that distillation and standard supervision converge to the same solution, but that distillation produces asymptotically more efficient SGD iterates in both online and finite-dataset settings. In rank-1 matrix approximation, we prove that both approaches converge to the same set of global optima and identified conditions under which distillation converges strictly faster. Our experiments on synthetic data, regression on real data, and language model pretraining corroborate our theory and suggests its practical relevance.

The limitations of our current analysis provide several avenues for future work. We focused on two setting with relatively benign geometry, extending this analysis to more complex optimization landscapes, where gradient descent may converge to suboptimal points, would more closely resemble practice. Another direction continues analyzing the GLM setting, but establishes non-asymptotic results. Lastly, extending empirical evaluations across a broader range of data domains may provide additional insights into how the underlying data distribution influences distillation.

7 ACKNOWLEDGMENTS

This work was partially supported by the Office for Naval Research under grant N00014-22-1-2669, NSF AI Institute for Foundations of Machine Learning (IFML), and Open Philanthropy. The authors would also like to thank Suhas Kotha and Gregory Xie for helpful feedback on the paper.

References

Jimmy Ba and Rich Caruana. Do deep nets really need to be deep? In Zoubin Ghahramani, Max Welling, Corinna Cortes, Neil D. Lawrence, and Kilian Q. Weinberger, editors, *Advances in Neural Information Processing Systems 27: Annual Conference on Neural Information Processing Systems 2014, December 8-13 2014, Montreal, Quebec, Canada*, pages 2654–2662, 2014. URL https://papers.nips.cc/paper_files/paper/2014/hash/b0c355a9dedccb50e5537e8f2e3f0810-Abstract.html.

Lucas Beyer, Xiaohua Zhai, Amélie Royer, Larisa Markeeva, Rohan Anil, and Alexander Kolesnikov. Knowledge distillation: A good teacher is patient and consistent. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2022, New Orleans, LA, USA, June 18-24, 2022*, pages 10915–10924. IEEE, 2022. doi: 10.1109/CVPR52688.2022.01065. URL https://doi.org/10.1109/CVPR52688.2022.01065.

Samuel Burer and Renato D. C. Monteiro. A nonlinear programming algorithm for solving semidefinite programs via low-rank factorization. *Math. Program.*, 95(2):329–357, 2003. doi: 10.1007/S10107-002-0352-8. URL https://doi.org/10.1007/s10107-002-0352-8.

Dan Busbridge, Amitis Shidani, Floris Weers, Jason Ramapuram, Etai Littwin, and Russell Webb. Distillation scaling laws. In Aarti Singh, Maryam Fazel, Daniel Hsu, Simon Lacoste-Julien, Felix Berkenkamp, Tegan Maharaj, Kiri Wagstaff, and Jerry Zhu, editors, *Forty-second International Conference on Machine Learning, ICML 2025, Vancouver, BC, Canada, July 13-19, 2025*, Proceedings of Machine Learning Research. PMLR / OpenReview.net, 2025. URL https://proceedings.mlr.press/v267/busbridge25a.html.

Tri Dao, Govinda M. Kamath, Vasilis Syrgkanis, and Lester Mackey. Knowledge distillation as semiparametric inference. In *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*. OpenReview.net, 2021. URL https://openreview.net/forum?id=m4UCf24r0Y.

Rudrajit Das and Sujay Sanghavi. Understanding self-distillation in the presence of label noise. In Andreas Krause, Emma Brunskill, Kyunghyun Cho, Barbara Engelhardt, Sivan Sabato, and Jonathan Scarlett, editors, *International Conference on Machine Learning, ICML 2023, 23-29 July 2023, Honolulu, Hawaii, USA*, volume 202 of *Proceedings of Machine Learning Research*, pages 7102–7140. PMLR, 2023. URL https://proceedings.mlr.press/v202/das23d.html.

Geoffrey E. Hinton, Oriol Vinyals, and Jeffrey Dean. Distilling the knowledge in a neural network. *CoRR*, abs/1503.02531, 2015. URL http://arxiv.org/abs/1503.02531.

Jordan Hoffmann, Sebastian Borgeaud, Arthur Mensch, Elena Buchatskaya, Trevor Cai, Eliza Rutherford, Diego de Las Casas, Lisa Anne Hendricks, Johannes Welbl, Aidan Clark, Tom Hennigan, Eric Noland, Katie Millican, George van den Driessche, Bogdan Damoc, Aurelia Guy, Simon Osindero, Karen Simonyan, Erich Elsen, Jack W. Rae, Oriol Vinyals, and Laurent Sifre. Training compute-optimal large language models. *CoRR*, abs/2203.15556, 2022. doi: 10.48550/ARXIV.2203.15556. URL https://doi.org/10.48550/arXiv.2203.15556.

Muhammed Emrullah Ildiz, Halil Alperen Gozeten, Ege Onur Taga, Marco Mondelli, and Samet Oymak. High-dimensional analysis of knowledge distillation: Weak-to-strong generalization and scaling laws. In *The Thirteenth International Conference on Learning Representations, ICLR 2025, Singapore, April 24-28, 2025*. OpenReview.net, 2025. URL https://openreview.net/forum?id=1xzqz73hvL.

Yoon Kim and Alexander M. Rush. Sequence-level knowledge distillation. In Jian Su, Xavier Carreras, and Kevin Duh, editors, *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing, EMNLP 2016, Austin, Texas, USA, November 1-4, 2016*, pages 1317–1327. The Association for Computational Linguistics, 2016. doi: 10.18653/V1/D16-1139. URL https://doi.org/10.18653/v1/d16-1139.

B. Laurent and P. Massart. Adaptive estimation of a quadratic functional by model selection. *The Annals of Statistics*, 28(5):1302 – 1338, 2000. doi: 10.1214/aos/1015957395. URL https://doi.org/10.1214/aos/1015957395.

Jason D. Lee, Max Simchowitz, Michael I. Jordan, and Benjamin Recht. Gradient descent only converges to minimizers. In Vitaly Feldman, Alexander Rakhlin, and Ohad Shamir, editors, *Proceedings of the 29th Conference on Learning Theory, COLT 2016, New York, USA, June 23-26, 2016*, JMLR

Workshop and Conference Proceedings, pages 1246–1257. JMLR.org, 2016. URL http://proceedings.mlr.press/v49/lee16.html.

Dong C. Liu and Jorge Nocedal. On the limited memory BFGS method for large scale optimization. *Math. Program.*, 45(1-3):503–528, 1989. doi: 10.1007/BF01589116. URL https://doi.org/10.1007/BF01589116.

Aditya Krishna Menon, Ankit Singh Rawat, Sashank J. Reddi, Seungyeon Kim, and Sanjiv Kumar. Why distillation helps: a statistical perspective. *CoRR*, abs/2005.10419, 2020. URL https://arxiv.org/abs/2005.10419.

Hossein Mobahi, Mehrdad Farajtabar, and Peter L. Bartlett. Self-distillation amplifies regularization in hilbert space. In Hugo Larochelle, Marc'Aurelio Ranzato, Raia Hadsell, Maria-Florina Balcan, and Hsuan-Tien Lin, editors, *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*, 2020. URL https://proceedings.neurips.cc/paper/2020/hash/2288f691b58edecadcc9a8691762b4fd-Abstract.html.

J. A. Nelder and R. W. M. Wedderburn. Generalized linear models. *Journal of the Royal Statistical Society. Series A (General)*, 135(3):370–384, 1972. ISSN 00359238, 23972327. URL http://www.jstor.org/stable/2344614.

Ioannis Panageas and Georgios Piliouras. Gradient descent only converges to minimizers: Non-isolated critical points and invariant regions. In Christos H. Papadimitriou, editor, *8th Innovations in Theoretical Computer Science Conference, ITCS 2017, January 9-11, 2017, Berkeley, CA, USA*, volume 67 of *LIPIcs*, pages 2:1–2:12. Schloss Dagstuhl - Leibniz-Zentrum für Informatik, 2017. doi: 10.4230/LIPICS.ITCS.2017.2. URL https://doi.org/10.4230/LIPIcs.ITCS.2017.2.

Abhishek Panigrahi, Bingbin Liu, Sadhika Malladi, Andrej Risteski, and Surbhi Goel. Progressive distillation induces an implicit curriculum. In *The Thirteenth International Conference on Learning Representations, ICLR 2025, Singapore, April 24-28, 2025*. OpenReview.net, 2025. URL https://openreview.net/forum?id=wPMRwmytZe.

Divyansh Pareek, Simon S. Du, and Sewoong Oh. Understanding the gains from repeated self-distillation. In Amir Globersons, Lester Mackey, Danielle Belgrave, Angela Fan, Ulrich Paquet, Jakub M. Tomczak, and Cheng Zhang, editors, *Advances in Neural Information Processing Systems 38: Annual Conference on Neural Information Processing Systems 2024, NeurIPS 2024, Vancouver, BC, Canada, December 10 - 15, 2024*, 2024. URL http://papers.nips.cc/paper_files/paper/2024/hash/0eb1ac7551ddbae575415aa5183a88be-Abstract-Conference.html.

Mary Phuong and Christoph Lampert. Towards understanding knowledge distillation. In Kamalika Chaudhuri and Ruslan Salakhutdinov, editors, *Proceedings of the 36th International Conference on Machine Learning, ICML 2019, 9-15 June 2019, Long Beach, California, USA*, volume 97 of *Proceedings of Machine Learning Research*, pages 5142–5151. PMLR, 2019. URL http://proceedings.mlr.press/v97/phuong19a.html.

B. T. Polyak and A. B. Juditsky. Acceleration of stochastic approximation by averaging. *SIAM Journal on Control and Optimization*, 30(4):838–855, 1992. doi: 10.1137/0330046. URL https://doi.org/10.1137/0330046.

Alec Radford, Jeff Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. Language models are unsupervised multitask learners. 2019. URL https://cdn.openai.com/better-language-models/language_models_are_unsupervised_multitask_learners.pdf.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. Exploring the limits of transfer learning with a unified text-to-text transformer. *J. Mach. Learn. Res.*, 21:140:1–140:67, 2020. URL https://jmlr.org/papers/v21/20-074.html.

Morgane Rivière, Shreya Pathak, Pier Giuseppe Sessa, Cassidy Hardin, Surya Bhupatiraju, Léonard Hussenot, Thomas Mesnard, Bobak Shahriari, Alexandre Ramé, Johan Ferret, Peter Liu, Pouya Tafti, Abe Friesen, Michelle Casbon, Sabela Ramos, Ravin Kumar, Charline Le Lan, Sammy Jerome, Anton Tsitsulin, Nino Vieillard, Piotr Stanczyk, Sertan Girgin, Nikola Momchev, Matt Hoffman, Shantanu Thakoor, Jean-Bastien Grill, Behnam Neyshabur, Olivier Bachem, Alanna Walton, Aliaksei Severyn, Alicia Parrish, Aliya Ahmad, Allen Hutchison, Alvin Abdagic, Amanda Carl, Amy Shen, Andy Brock, Andy Coenen, Anthony Laforge, Antonia Paterson, Ben Bastian, Bilal Piot, Bo Wu, Brandon Royal, Charlie Chen, Chintu Kumar, Chris Perry, Chris Welty, Christopher A. Choquette-Choo, Danila Sinopalnikov, David Weinberger, Dimple Vijaykumar, Dominika Rogozinska, Dustin Herbison, Elisa Bandy, Emma Wang, Eric Noland, Erica Moreira, Evan Senter, Evgenii Eltyshev, Francesco Visin, Gabriel Rasskin, Gary Wei, Glenn

Cameron, Gus Martins, Hadi Hashemi, Hanna Klimczak-Plucinska, Harleen Batra, Harsh Dhand, Ivan Nardini, Jacinda Mein, Jack Zhou, James Svensson, Jeff Stanway, Jetha Chan, Jin Peng Zhou, Joana Carrasqueira, Joana Iljazi, Jocelyn Becker, Joe Fernandez, Joost van Amersfoort, Josh Gordon, Josh Lipschultz, Josh Newlan, Ju-yeong Ji, Kareem Mohamed, Kartikeya Badola, Kat Black, Katie Millican, Keelin McDonell, Kelvin Nguyen, Kiranbir Sodhia, Kish Greene, Lars Lowe Sjösund, Lauren Usui, Laurent Sifre, Lena Heuermann, Leticia Lago, and Lilly McNealus. Gemma 2: Improving open language models at a practical size. *CoRR*, abs/2408.00118, 2024. doi: 10.48550/ARXIV.2408.00118. URL https://doi.org/10.48550/arXiv.2408.00118.

H. Robbins and D. Siegmund. A convergence theorem for non negative almost supermartingales and some applications. In Jagdish S. Rustagi, editor, *Optimizing Methods in Statistics*, pages 233–257. Academic Press, 1971. ISBN 978-0-12-604550-5. doi: https://doi.org/10.1016/B978-0-12-604550-5.50015-8. URL https://www.sciencedirect.com/science/article/pii/B9780126045505500158.

Adriana Romero, Nicolas Ballas, Samira Ebrahimi Kahou, Antoine Chassang, Carlo Gatta, and Yoshua Bengio. Fitnets: Hints for thin deep nets. In Yoshua Bengio and Yann LeCun, editors, *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*, 2015. URL http://arxiv.org/abs/1412.6550.

Mher Safaryan, Alexandra Peste, and Dan Alistarh. Knowledge distillation performs partial variance reduction. In Alice Oh, Tristan Naumann, Amir Globerson, Kate Saenko, Moritz Hardt, and Sergey Levine, editors, *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023*, 2023. URL http://papers.nips.cc/paper_files/paper/2023/hash/ee1f0da706829d7f198eac0edaacc338-Abstract-Conference.html.

Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. Distilbert, a distilled version of BERT: smaller, faster, cheaper and lighter. *CoRR*, abs/1910.01108, 2019. URL http://arxiv.org/abs/1910.01108.

Sharath Turuvekere Sreenivas, Saurav Muralidharan, Raviraj Joshi, Marcin Chochowski, Mostofa Patwary, Mohammad Shoeybi, Bryan Catanzaro, Jan Kautz, and Pavlo Molchanov. LLM pruning and distillation in practice: The minitron approach. *CoRR*, abs/2408.11796, 2024. doi: 10.48550/ARXIV.2408.11796. URL https://doi.org/10.48550/arXiv.2408.11796.

Joel A. Tropp. An introduction to matrix concentration inequalities, 2015. URL https://arxiv.org/abs/1501.01571.

Han Xiao, Kashif Rasul, and Roland Vollgraf. Fashion-mnist: a novel image dataset for benchmarking machine learning algorithms. *CoRR*, abs/1708.07747, 2017. URL http://arxiv.org/abs/1708.07747.

Checklist

1. For all models and algorithms presented, check if you include:
   (a) A clear description of the mathematical setting, assumptions, algorithm, and/or model. Yes.
   (b) An analysis of the properties and complexity (time, space, sample size) of any algorithm. Not Applicable.
   (c) (Optional) Anonymized source code, with specification of all dependencies, including external libraries. Yes.

2. For any theoretical claim, check if you include:
   (a) Statements of the full set of assumptions of all theoretical results. Yes.
   (b) Complete proofs of all theoretical results. Yes.
   (c) Clear explanations of any assumptions. Yes.

3. For all figures and tables that present empirical results, check if you include:
   (a) The code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL). Yes.
   (b) All the training details (e.g., data splits, hyperparameters, how they were chosen). Yes.
   (c) A clear definition of the specific measure or statistics and error bars (e.g., with respect to the random seed after running experiments multiple times). Yes.
   (d) A description of the computing infrastructure used. (e.g., type of GPUs, internal cluster, or cloud provider). Yes.

4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets, check if you include:
   (a) Citations of the creator If your work uses existing assets. Yes.

(b) The license information of the assets, if applicable. Not Applicable.

(c) New assets either in the supplemental material or as a URL, if applicable. Not Applicable.

(d) Information about consent from data providers/curators. Not Applicable.

(e) Discussion of sensible content if applicable, e.g., personally identifiable information or offensive content. Not Applicable.

5. If you used crowdsourcing or conducted research with human subjects, check if you include:

(a) The full text of instructions given to participants and screenshots. Not Applicable.

(b) Descriptions of potential participant risks, with links to Institutional Review Board (IRB) approvals if applicable. Not Applicable.

(c) The estimated hourly wage paid to participants and the total amount spent on participant compensation. Not Applicable.

Appendix

A RELATED WORK CONTINUED

Knowledge distillation has shown repeated success in training small performant models using larger, more capable models. Prior theoretical analyses that have sought to explain why, have done so primarily from the perspective that distillation yields student models with better generalization. Menon et al. (2020) connects teacher predictions in classification to modeling the Bayes class probabilities. They show that empirical risk with distillation labels (Bayes-distilled risk) has lower variance relative to standard labels, and the student minimizing the Bayes-distilled risk inherits the variance reduction, leading to better generalization. Dao et al. (2021) views distillation as a plug-in method for semiparameteric inference, where the teacher predictions serve as a plug-in estimate of the Bayes class probabilities. They establish excess risk guarantees for the student that depend on teacher quality, and introduce loss correction to address teacher bias and cross-fitting to mitigate overfitting from data reuse.

Mobahi et al. (2020) analyze self-distillation, where teacher and student are the same, in a Hilbert space with ℓ_2 regularization. They show that repeated distillation shrinks basis coefficients more aggressively, thus acting as a strong regularizer that can reduce overfitting and improve generalization. Das and Sanghavi (2023) provide another explanation for self-distillation. They consider a student that minimizes a dual objective on both original and teacher-produced labels, and characterize settings where distillation produces students with higher accuracy in ridge regression and binary logistic regression. Pareek et al. (2024) generalizes this to multistep self-distillation.

Ildiz et al. (2025) analyze knowledge distillation in high-dimensional ridgeless linear regression with Gaussian covariates. They show that in the underparameterized regime ($n > p$) the optimal teacher coincides with the ground truth parameter, so distillation cannot outperform direct supervision, while in the overparameterized regime ($n < p$) an optimally chosen teacher reduces the risk of the student.

While these works study the student model defined as the minimizer of the objective under knowledge distillation, our focus is instead on the successive iterates generated by descent algorithms that approximate this minimizer.

B Detailed assumptions and deferred proofs for Section 3.1

B.1 Assumptions

In this section, we present the formal assumptions used in Theorems 3.1 and 3.2. These assumptions apply to both the standard supervision objectives, $\ell_S(\alpha; z)$ and $L_S(\alpha)$, and the knowledge distillation objectives, $\ell_S(\beta; z')$ and $L'_S(\beta)$. To streamline the presentation, we introduce the notation $\ell_S(\gamma; \tilde{z})$ and $\widetilde{L}_S(\gamma)$, with $\tilde{z} \in \{z, z'\}$, to state the assumptions that apply to both settings.

Assumption B.0.1. *The per-sample loss $\ell_S(\gamma; \tilde{z})$ and population loss $\widetilde{L}_S(\gamma) = \mathbb{E}[\ell_S(\gamma; \tilde{z})]$ where $\tilde{z} = (v, \tilde{y})$, $v = \Pi x$, and $\tilde{y} \in \{y, y'\}$ satisfies the following:*

(i) At the unique optimal point $\gamma^* := \operatorname{argmin}_\gamma \widetilde{L}_S(\gamma)$, for a constant $m > 0$, the Hessian $\nabla^2 \widetilde{L}_S(\gamma^*) \succeq m \cdot I_d$ is positive definite.

(ii) The map $\gamma \mapsto \ell_S(\gamma; \tilde{z})$ is twice continuously differentiable for every $\tilde{z}$. There exists a random variable $M(\tilde{z})$ with $\mathbb{E}[M(\tilde{z})] < \infty$ and a constant $r > 0$ such that for all $\gamma \in B(\gamma^*, r)$,

$$||\nabla \ell_S(\gamma; \tilde{z}) - \nabla \ell_S(\gamma^*; \tilde{z})||_2 \leq M(\tilde{z}) \cdot ||\gamma - \gamma^*||_2$$
$$||\nabla^2 \ell_S(\gamma; \tilde{z}) - \nabla^2 \ell_S(\gamma^*; \tilde{z})||_{op} \leq M(\tilde{z}) \cdot ||\gamma - \gamma^*||_2.$$

(iii) There exists a constant $C > 0$ such that the per-sample gradient $\nabla \ell_S(\gamma; \tilde{z})$, is bounded $\mathbb{E}[||\nabla \ell_S(\gamma; \tilde{z})||_2^2] < C \cdot (1 + ||\gamma||_2^2)$.

We now specialize Theorem 2 of Polyak and Juditsky (1992) to our setting of GLMs with the population objective and state it as a corollary.

Corollary B.0.1 (Polyak and Juditsky (1992), GLMs with objective $\widetilde{L}_S$)**.** *Consider the stochastic gradient descent update with sample* $(v_t, \tilde{y}_t)$,

$$\gamma_t = \gamma_{t-1} - \eta_t \nabla \ell_S(\gamma_{t-1}; v_t, \tilde{y}_t).$$

Assume B.0.1 and 3.0.2 are true, then $\bar{\gamma}_t \to \gamma^*$ *almost surely, and*

$$\sqrt{t}(\bar{\gamma}_t - \gamma^*) \xrightarrow{d} \mathcal{N}(0, \widetilde{H}^{-1}\widetilde{C}(\gamma^*)\widetilde{H}^{-1})$$

with second moment of the gradient noise at the optimum $\widetilde{C}(\gamma^*) = \mathbb{E}[\nabla \ell_S(\gamma^*; v, \tilde{y})(\nabla \ell_S(\gamma^*; v, \tilde{y}))^T]$ *and Hessian* $\widetilde{H} = \nabla^2 \widetilde{L}_S(\gamma^*)$.

Proof. Begin by considering the recursion

$$\gamma_t = \gamma_{t-1} - \eta_t g_t$$

where $\xi_t = \nabla \ell_S(\gamma_{t-1}; v_t, \tilde{y}_t) - \nabla \tilde{L}_S(\gamma_{t-1})$ and $g_t = \nabla \tilde{L}_S(\gamma_{t-1}) + \xi_t$. Note that this is the stochastic optimization scheme analyzed in Theorem 2 of Polyak and Juditsky (1992). Now we show that the original assumptions are satisfied. For the assumptions needed to prove almost sure convergence of SGD using the theorem by Robbins and Siegmund (1971), note that the gradient noise $\{\xi_t\}_{t\geq 1}$ is a martingale-difference sequence with respect to the natural filtration of samples $\{\tilde{z}_t\}_{t\geq 1}$ since the gradient is computed with an i.i.d. sample $\tilde{z}_t$,

$$\begin{aligned}\mathbb{E}[\xi_t|\mathcal{F}_{t-1}] &= \mathbb{E}[\nabla \ell_S(\gamma_{t-1}; \tilde{z}_t)|\mathcal{F}_{t-1}] - \nabla \widetilde{L}_S(\gamma_{t-1}) \\ &= 0 \quad a.s.\end{aligned}$$

for all t. Furthermore, for all t

$$\begin{aligned}\mathbb{E}[||\xi_t||_2^2|\mathcal{F}_{t-1}] + ||\nabla \widetilde{L}_S(\gamma_{t-1})||_2^2 &\leq \mathbb{E}[||\xi_t||_2^2 + ||\nabla \widetilde{L}_S(\gamma_{t-1})||_2^2|\mathcal{F}_{t-1}] \\ &\leq 2\mathbb{E}[||\nabla \ell_S(\gamma_{t-1}; \tilde{z}_t)||_2^2|\mathcal{F}_{t-1}] \\ &\leq 2C \cdot (1 + ||\gamma_{t-1}||_2^2) \end{aligned} \tag{2}$$

where the last inequality follows from Assumption B.0.1.

Now consider the function $V(u) = ||u||_2^2$ which for all u, $z \in \mathbb{R}^d$ has a 2-Lipschitz continuous gradient

$$||\nabla V(u) - \nabla V(z)||_2 \leq 2 \cdot ||u - v||_2.$$

Let $\Delta_t = \gamma_t - \gamma^*$ and $\overline{R}(u) = \nabla \widetilde{L}_S(u + \gamma^*)$. Then the function $V_t = V(\Delta_t)$ after one update can be upper bounded using a second order Taylor expansion and the fact that V has a Lipschitz-continuous gradient,

$$\begin{aligned}V_t &\leq V_{t-1} + \nabla V_{t-1}^T(-\eta_t(\overline{R}(\Delta_{t-1}) + \xi_t)) + 2||\Delta_t - \Delta_{t-1}||_2^2 \\ &\leq V_{t-1} - \eta_t \nabla V_{t-1}^T \overline{R}(\Delta_{t-1}) - \eta_t \nabla V_{t-1}^T \xi_t + 2\eta_t^2 ||\overline{R}(\Delta_{t-1}) + \xi_t||_2^2.\end{aligned}$$

Taking the expectation conditioned on $\mathcal{F}_{t-1}$,

$$\begin{aligned}\mathbb{E}[V_t|\mathcal{F}_{t-1}] &\leq V_{t-1} - \eta_t \nabla V_{t-1}^T \overline{R}(\Delta_{t-1}) + 2\eta_t^2 \mathbb{E}\left[||\overline{R}(\Delta_{t-1}) + \xi_t||_2^2|\mathcal{F}_{t-1}\right] \\ &\leq V_{t-1} - \eta_t \nabla V_{t-1}^T \overline{R}(\Delta_{t-1}) + 2\eta_t^2 \left(||\overline{R}(\Delta_{t-1})||_2^2 + \mathbb{E}\left[||\xi_t||_2^2|\mathcal{F}_{t-1}\right] \right) \\ &\leq V_{t-1} - \eta_t \nabla V_{t-1}^T \overline{R}(\Delta_{t-1}) + 2C\eta_t^2(1 + ||\gamma_{t-1}||_2^2) \\ &\leq V_{t-1}(1 + \eta_t^2 K) + \eta_t^2 K - \eta_t \nabla V_{t-1}^T \overline{R}(\Delta_{t-1}) \end{aligned} \tag{3}$$

where we use the fact that $\{\xi_t\}_{t\geq 1}$ is a martingale-difference sequence, the inequality from Eq. (2), and we gather constants into a suitable $K > 0$. Then we note that the terms $(1 + \eta_t^2 K)$, $\eta_t^2 K \geq 0$ and by convexity of $\widetilde{L}_S$

$$\begin{aligned}\eta_t \nabla V_{t-1}^T \overline{R}(\Delta_{t-1}) &= 2\eta_t(\gamma_{t-1} - \gamma^*)^T \nabla \widetilde{L}_S(\gamma_{t-1}) \\ &\geq 2\eta_t(\widetilde{L}(\gamma_{t-1}) - \widetilde{L}_S(\gamma^*)) \\ &\geq 0.\end{aligned}$$

Then with the step size assumptions which ensure $\sum_{i=1}^{\infty} \eta_t = \infty$ and $\sum_{i=1}^{\infty} \eta_t^2 < \infty$ we obtain by the theorem of Robbins and Siegmund (1971) that $V_t \to V_\infty$ where V_∞ is bounded.

We now conclude this section by showing that V_t in fact converges to 0. Begin by showing $\widetilde{L}_S$ is strongly convex in a neighborhood around γ^*. Consider $\gamma \in B(\gamma^*, r)$, the function is locally Lipschitz around γ^*,

$$\lambda_{min}(\nabla^2 \widetilde{L}_S(\gamma)) \geq m - \mathbb{E}[M(\tilde{z})] \cdot ||\gamma - \gamma^*||_2.$$

Then taking the smaller of the two radii r and $m/(2\mathbb{E}[M(\tilde{z})])$, for all $\gamma \in B(\gamma^*, \min\{r, m/(2\mathbb{E}[M(\tilde{z})])\}$ we have $\lambda_{min}(\nabla^2 \widetilde{L}_S(\gamma)) \geq m/2$. Continuing to work within this neighborhood,

$$\begin{aligned}(\gamma - \gamma^*)^T \nabla \widetilde{L}_S(\gamma) &= (\gamma - \gamma^*)^T \int_0^1 \nabla^2 \widetilde{L}_S(\gamma^* + t(\gamma - \gamma^*))(\gamma - \gamma^*)\, dt \\ &\geq \frac{m}{2} \cdot ||\gamma - \gamma^*||_2^2.\end{aligned}$$

Now we showed that V_t converges to V_∞ which is bounded, thus, for any $\varepsilon > 0$, there exists R such that $\Pr[\sup_t ||\Delta_t||_2 < R] \geq 1 - \varepsilon$. Define the stopping time $\tau_R = \inf\{t \geq 1 : ||\Delta_t||_2 \leq R\}$, then on the event that $\hat{\gamma}$ is within radius R of τ, $\{t < \tau_R\}$ we have from Eq. (3) and the bound above

$$\mathbb{E}[V_t I(t < \tau_R)] \leq \mathbb{E}\left[\left(1 - \eta_t \frac{m}{2} + \eta_t^2 K\right) \cdot V_{t-1} I(t - 1 < \tau_R)\right] + \eta_t^2 K.$$

Polyak and Juditsky (1992) then conclude that $\mathbb{E}[V_t I(t < \tau_R)]$ is upper bounded by the step size η_t multiplied by a constant. As the step size converges to zero, this implies the almost sure convergence of the algorithm $||\Delta_t||_2^2 \to 0$.

Now what remains is to prove that the distribution of error from the process $\{\Delta_t\}_{t \geq 1}$ is asymptotically normal. The argument by Polyak and Juditsky (1992) proceeds by defining a process $\{\Delta_t'\}_{t \geq 1}$ which approximates the true sequence of errors $\{\Delta_t\}_{t \geq 1}$ with a linear update. Then by their Theorem 1, which applies to linear problems, the distribution of error in the proxy process is asymptotically normal, and it suffices to show that these two processes are asymptotically equivalent.

For our setting of GLMs, define the proxy process

$$\begin{aligned}\Delta_t' &= \Delta_{t-1}' - \eta_t G \Delta_{t-1}' + \eta_t \xi_t, \qquad \Delta_0' = \Delta_0 \\ \overline{\Delta}_t &= \frac{1}{t} \sum_{i=0}^{t-1} \Delta_i'\end{aligned}$$

where $G = \nabla^2 \widetilde{L}_S(\gamma^*)$. The matrix G satisfies for any $\gamma \in B(\gamma^*, r)$,

$$\begin{aligned}||\nabla \widetilde{L}_S(\gamma) - G(\gamma - \gamma^*)||_2 &\leq \left(\int_0^1 ||\nabla^2 \widetilde{L}_S(\gamma^* + t(\gamma - \gamma^*)) - \nabla^2 \widetilde{L}_S(\gamma^*)||_{op}\, dt\right) \cdot ||\gamma - \gamma^*||_2 \\ &\leq \mathbb{E}[M(\tilde{z})] \cdot ||\gamma - \gamma^*||_2^2\end{aligned}$$

which follows from the locally Lipschitz continuous Hessian around γ^*.

The remaining assumptions for Theorem 2 are regarding the error decomposition $\xi_t = \xi_t(0) + \zeta_t(\Delta_{t-1})$ where

$$\begin{aligned}\xi_t(0) &= \nabla \ell_S(\gamma^*; \tilde{z}_t) - \nabla \widetilde{L}_S(\gamma^*) = \nabla \ell_S(\gamma^*; \tilde{z}_t) \\ \zeta_t(\Delta_{t-1}) &= [\nabla \ell_S(\Delta_{t-1} + \gamma^*; \tilde{z}_t) - \nabla \ell_S(\gamma^*; \tilde{z}_t)] + [\nabla \widetilde{L}_S(\gamma^*) - \nabla \widetilde{L}_S(\Delta_{t-1} + \gamma^*)].\end{aligned}$$

that satisfy the following

$$\mathbb{E}[\xi_t(0) | \mathcal{F}_{t-1}] = 0 \quad a.s. \tag{4}$$

$$\mathbb{E}[\xi_t(0) \xi_t(0)^T | \mathcal{F}_{t-1}] \xrightarrow{P} S \quad \text{as } t \to \infty; \quad S > 0 \tag{5}$$

$$\sup_t \mathbb{E}[||\xi_t(0)||_2^2 I(||\xi_t(0)||_2 > C) | \mathcal{F}_{t-1}] \xrightarrow{P} 0 \quad \text{as } C \to \infty \tag{6}$$

$$\exists\, T, \forall t \geq T, \mathbb{E}[||\zeta_t(\Delta_{t-1})||_2^2 | \mathcal{F}_{t-1}] \leq \delta(\Delta_{t-1}) \quad a.s. \tag{7}$$

with $\delta(\Delta_t) \to 0$ as $\Delta_t \to 0$. We now prove the properties (4-7). Updates are performed with an i.i.d. sample $\tilde{z}_t$, making the expectation of $\xi_t(0)$ conditioned on the previous samples zero and furthermore,

$$\mathbb{E}[\xi_t(0)\xi_t(0)^T|\mathcal{F}_{t-1}] = \mathbb{E}[\nabla\ell_S(\gamma^*;\tilde{z})\nabla\ell_S(\gamma^*;\tilde{z})^T] =: S,$$

thus, meeting Properties (4) and (5). Now observe that for all t conditional on previous samples,

$$\mathbb{E}[||\xi_t(0)||_2^2 I(||\xi_t(0)||_2 > C)|\mathcal{F}_{t-1}] = \mathbb{E}[||\xi_1(0)||_2^2 I(||\xi_1(0)||_2 > C)].$$

We know that $\mathbb{E}[||\xi_1(0)||_2^2] = \mathbb{E}[||\nabla\ell_S(\gamma^*;\tilde{z})||_2^2] < C \cdot (1 + ||\gamma^*||_2^2)$ by Assumption B.0.1. Now Property (6) is equivalent to

$$\sup_t \mathbb{E}[||\xi_t(0)||_2^2 I(||\xi_t(0)||_2 > 0)|\mathcal{F}_{t-1}] = \mathbb{E}[||\xi_1(0)||_2^2 I(||\xi_1(0)||_2 > C)]$$

which converges to zero in probability as $C \to \infty$ since the expectation of $||\xi_1(0)||_2^2$ is finite. For Property (7) note that we assume the gradients of ℓ_S are locally Lipschitz, thus,

$$\begin{aligned}\mathbb{E}[||\zeta_t(\Delta_{t-1})||_2^2|\mathcal{F}_{t-1}] &\leq \mathbb{E}[||\nabla\ell_S(\gamma_{t-1};\tilde{z}_t) - \nabla\ell_S(\gamma^*;\tilde{z}_t)||_2^2|\mathcal{F}_{t-1}] + ||\nabla\widetilde{L}_S(\gamma^*) - \nabla\widetilde{L}_S(\gamma_{t-1})||_2^2 \\ &\leq 2\mathbb{E}[M(\tilde{z})]^2 \cdot ||\Delta_{t-1}||_2^2\end{aligned}$$

which tends to zero once $\Delta_t \to 0$. □

B.2 Proof of Theorem 3.1

Proof. By Corollary B.0.1 the SGD sequence $\bar{\alpha}_t \to^{a.s.} \alpha^*$, likewise, $\bar{\beta}_t \to^{a.s.} \beta^*$. What remains is to show $\alpha^* = \beta^*$. By definition, the teacher estimator recovers the true parameter θ^*. Consequently, the distillation label corresponding to a sample (x, y) is $y' = h(x^T\theta^*) = \mathbb{E}[y|x]$. Both limit points α^* and β^* must satisfy the first order optimality conditions for their respective objectives L_S and L'_S which correspond to the following conditions,

$$\mathbb{E}[h(v^T\alpha^*)v] = \mathbb{E}[yv], \quad \mathbb{E}[h(v^T\beta^*)v] = \mathbb{E}[y'v].$$

The following equality $\mathbb{E}[yv] = \mathbb{E}[\mathbb{E}[y|x]v] = \mathbb{E}[y'v]$ from taking iterated expectations means that points α^* and β^* must satisfy the same optimality condition. Then strict convexity implies that the optimal point is unique, thus showing that α^* and β^* which satisfy the same condition for optimality must be the same point. □

B.3 Proof of Theorem 3.2

Proof. Using Theorem 3.1, the Hessians of the population losses evaluated at the optimal points are $H(\gamma^*) = \nabla^2 L_S(\gamma^*)$ and $H'(\gamma^*) = \nabla^2 L'_S(\gamma^*)$. Let $C(\gamma^*) = \mathbb{E}[\nabla\ell_S(\gamma^*;z)(\nabla\ell_S(\gamma^*;z))^T]$ and $C'(\gamma^*) = \mathbb{E}[\nabla\ell_S(\gamma^*;z')(\nabla\ell_S(\gamma^*;z'))^T]$ be the second moment of the gradient noise at the optimum γ^*. Then with Assumptions B.0.1 and 3.0.2, we can apply Corollary B.0.1 which states that the averaged SGD iterates are asymptotically normal. Precisely,

$$\begin{aligned}\sqrt{t}(\bar{\alpha}_t - \gamma^*) &\xrightarrow{d} \mathcal{N}(0, H(\gamma^*)^{-1}C(\gamma^*)H(\gamma^*)^{-1}), \\ \sqrt{t}(\bar{\beta}_t - \gamma^*) &\xrightarrow{d} \mathcal{N}(0, H'(\gamma^*)^{-1}C'(\gamma^*)H'(\gamma^*)^{-1}).\end{aligned}$$

The Hessians are equal,

$$H(\gamma^*) = \mathbb{E}[h'(v^T\gamma^*)vv^T] = H'(\gamma^*)$$

The expression for the gradients are $\nabla\ell_S(\gamma^*;z) = (h(v^T\gamma^*) - y)v$ and $\nabla\ell_S(\gamma^*;z') = (h(v^T\gamma^*) - y')v$. Let $\omega = h(v^T\gamma^*) - y'$ and $\delta = y' - y$. We can then write the gradient covariance $C(\gamma^*)$ as,

$$\begin{aligned}C(\gamma^*) &= \mathbb{E}[(h(v^T\gamma^* - y))^2vv^T] \\ &= \mathbb{E}[((h(v^T\gamma^*) - y') + (y' - y))^2vv^T] \\ &= \mathbb{E}[\omega^2vv^T] + 2\mathbb{E}[\omega\delta vv^T] + \mathbb{E}[\delta^2vv^T].\end{aligned}$$

The cross term is equal to zero, to see this take the conditional expectation

$$\begin{aligned}\mathbb{E}[\omega\delta vv^T] &= \mathbb{E}[\mathbb{E}[y' - y|x] \cdot (h(v^T\gamma^*) - y')vv^T] \\ &= \mathbb{E}[\mathbb{E}[h(x^T\theta^*) - y)|x] \cdot (h(v^T\gamma^*) - h(x^T\theta^*))vv^T]\end{aligned}$$

where for the last equality we simply write the definition of the distillation label. Making use of the fact that the teacher estimator is well-specified and thus $\mathbb{E}[y|x] = h(x^T\theta^*)$, the term above is equal to zero. The covariance is now $C(\gamma^*) = \mathbb{E}[\omega^2 vv^T] + \mathbb{E}[\delta^2 vv^T]$. The last expression has the familiar form $\mathbb{E}[\delta^2 vv^T] = \mathbb{E}[\mathrm{Var}(y|x)vv^T]$. Now observe that the gradient covariance of the SGD sequence with knowledge distillation is,

$$C'(\gamma^*) = \mathbb{E}[(h(v^T\gamma^*) - y')^2 vv^T].$$

We can write $C(\gamma^*) = C'(\gamma^*) + \mathbb{E}[\mathrm{Var}(y|x)vv^T]$, and it is now clear that $C(\gamma^*) \succeq C'(\gamma^*)$ since the variance component is always positive semi-definite. □

C Detailed assumptions and deferred proofs for Section 3.2

C.1 Assumptions

We now present the formal assumptions in Theorems C.2 and C.3. All assumptions apply to both the standard supervision objectives, $\ell_S(\alpha; z)$ and $L_S^n(\alpha)$, as well as the knowledge distillation objectives, $\ell_S(\beta; z')$ and $L'^n_S(\beta)$. We unify the presentation by introducing the notation $\ell_S(\gamma; \tilde{z})$ and $\widetilde{L}_S^n(\gamma)$, with $\tilde{z} \in \{z, z'\}$. All assumptions are stated in this general form.

Assumption C.0.1. *The per-sample loss $\ell_S(\gamma; \tilde{z})$ and empirical loss $\widetilde{L}_S^n(\gamma) = 1/n \cdot \sum_{i=1}^n \ell_S(\gamma; \tilde{z}_i)$ where $\tilde{z} = (v, \tilde{y})$, $v = \Pi x$, and $\tilde{y} \in \{y, y'\}$ satisfies the following:*

(i) At the unique optimal point $\hat{\gamma} := \mathrm{argmin}_\gamma \widetilde{L}_S^n(\gamma)$, for a constant $m > 0$, the Hessian $\nabla^2 \widetilde{L}_S^n(\hat{\gamma}) \succeq m \cdot I_d$ is positive definite.

(ii) The map $\gamma \mapsto \ell_S(\gamma; \tilde{z})$ is twice continuously differentiable for every z. There exists constants r, $M > 0$ such that for all $\gamma \in B(\hat{\gamma}, r)$,

$$\begin{aligned}||\nabla\ell_S(\gamma; \tilde{z}) - \nabla\ell_S(\hat{\gamma}; \tilde{z})||_2 &\leq M \cdot ||\gamma - \hat{\gamma}||_2 \\ ||\nabla^2\ell_S(\gamma; \tilde{z}) - \nabla^2\ell_S(\hat{\gamma}; \tilde{z})||_{op} &\leq M \cdot ||\gamma - \hat{\gamma}||_2.\end{aligned}$$

(iii) There exists a constant $C > 0$ such that the per-sample gradient $\nabla\ell_S(\gamma; \tilde{z})$, is bounded $||\nabla\ell_S(\gamma; \tilde{z})||_2^2 \leq C \cdot (1 + ||\gamma||_2^2)$.

We now state Theorem 2 of Polyak and Juditsky (1992), specialized to the setting of GLMs with the empirical loss, as the following corollary.

Corollary C.0.1 (Polyak and Juditsky (1992), GLMs with objective $\widetilde{L}_S^n$)**.** *Consider the stochastic gradient descent update with sample $(v_t, \tilde{y}_t)$,*

$$\gamma_t = \gamma_{t-1} - \eta_t \nabla\ell_S(\gamma_{t-1}; v_t, \tilde{y}_t).$$

Assume C.0.1 and 3.0.2 are true, then $\bar{\gamma}_t \to \hat{\gamma}$ almost surely, and

$$\sqrt{t}(\bar{\gamma}_t - \hat{\gamma}) \xrightarrow{d} \mathcal{N}(0, \widetilde{H}_n^{-1}\widetilde{C}_n(\hat{\gamma})\widetilde{H}_n^{-1})$$

with second moment of the gradient noise at the optimum $\widetilde{C}_n(\hat{\gamma}) = 1/n \cdot \sum_{i=1}^n \nabla\ell_S(\hat{\gamma}; v_i, \tilde{y}_i)(\nabla\ell_S(\hat{\gamma}; v_i, \tilde{y}_i))^T$ and empirical Hessian $\widetilde{H}_n = 1/n \cdot \sum_{i=1}^n \nabla^2\ell_S(\hat{\gamma}; v_i, \tilde{y}_i)$.

Proof. We begin by writing the stochastic gradient descent update in the form analyzed in Theorem 2 of Polyak and Juditsky (1992),

$$\gamma_t = \gamma_{t-1} - \eta_t g_t$$

where $\xi_t = \nabla \ell_S(\gamma_{t-1}; v_t, \tilde{y}_t) - \nabla \widetilde{L}_S^n(\gamma_{t-1})$ and $g_t = \widetilde{L}_S^n(\gamma_{t-1}) + \xi_t$. Now we show that the assumptions of the original theorem hold, starting with the set of assumptions for proving almost sure convergence of SGD using the theorem by Robbins and Siegmund (1971). Begin by noting that the gradient noise sequence $\{\xi_t\}_{t\geq 1}$ is a martingale-difference sequence with respect to the natural filtration of samples $\{\tilde{z}_t\}_{t\geq 1}$, for all t,

$$\begin{aligned}\mathbb{E}[\xi_t|\mathcal{F}_{t-1}] &= \mathbb{E}[\nabla \ell_S(\gamma_{t-1}; \tilde{z}_t)|\mathcal{F}_{t-1}] - \nabla \widetilde{L}_S^n(\gamma_{t-1}) \\ &= 0 \quad a.s.\end{aligned}$$

where the expectation is taken over independent $\tilde{z}$ sampled uniformly from the finite dataset. Moreover, for all t

$$\begin{aligned}\mathbb{E}[||\xi_t||_2^2|\mathcal{F}_{t-1}] + ||\nabla \widetilde{L}_S^n(\gamma_{t-1})||_2^2 &= \frac{1}{n}\sum_{i=1}^n ||\nabla \ell_S(\gamma_{t-1}; \tilde{z}_i) - \nabla \widetilde{L}_S^n(\gamma_{t-1})||_2^2 + ||\nabla \widetilde{L}_S^n(\gamma_{t-1})||_2^2 \\ &\leq \frac{2}{n}\sum_{i=1}^n ||\nabla \ell_S(\gamma_{t-1}; \tilde{z}_i)||_2^2 \\ &\leq 2C \cdot (1 + ||\gamma_{t-1}||_2^2)\end{aligned}$$

where the final inequality follows from Assumption C.0.1.

Let $V(u) = ||u||_2^2$. This function has a 2-Lipschitz continuous gradient, for all u, $z \in \mathbb{R}^d$,

$$||\nabla V(u) - \nabla V(z)||_2^2 \leq 2 \cdot ||u - z||_2.$$

Define the error $\Delta_t = \gamma_t - \hat{\gamma}$ and $\overline{R}(u) = \nabla \widetilde{L}_S^n(u + \hat{\gamma})$. Then we can derive an upper bound for $V_t = V(\Delta_t)$ by a second order Taylor expansion and the fact that V has a 2-Lipschitz continuous gradient

$$\begin{aligned}V_t &\leq V_{t-1} + \nabla V_{t-1}^T(-\eta_t(\overline{R}(\Delta_{t-1}) + \xi_t)) + 2||\Delta_t - \Delta_{t-1}||_2^2 \\ &\leq V_{t-1} - \eta_t \nabla V_{t-1}^T \overline{R}(\Delta_{t-1}) - \eta_t \nabla V_{t-1}^T \xi_t + 2\eta_t^2 ||\overline{R}(\Delta_{t-1}) + \xi_t||_2^2.\end{aligned}$$

Now taking expectations conditioned on the previous history, recall that the gradient noise is a martingale-difference sequence thus

$$\begin{aligned}\mathbb{E}[V_t|\mathcal{F}_{t-1}] &\leq V_{t-1} - \eta_t \nabla V_{t-1}^T \overline{R}(\Delta_{t-1}) + 2\eta_t^2 \left(||\overline{R}(\Delta_{t-1})||_2^2 + \mathbb{E}[||\xi_t||_2^2|\mathcal{F}_{t-1}]\right) \\ &\leq V_{t-1} - \eta_t \nabla V_{t-1}^T \overline{R}(\Delta_{t-1}) + 2C\eta_t^2(1 + ||\gamma_{t-1}||_2^2) \\ &\leq V_{t-1}(1 + \eta_t^2 K) + \eta_t^2 K - \eta_t \nabla V_{t-1}^T \overline{R}(\Delta_{t-1}) \qquad (8)\end{aligned}$$

where the second to last inequality follows from Assumption C.0.1 and we gather constants into a suitable $K > 0$. To apply the theorem by Robbins and Siegmund (1971) which proves V_t converges to a bounded random variable, we must show these coefficients are nonnegative. This is clear as $(1 + \eta_t^2 K)$, $\eta_t^2 K \geq 0$ and by convexity of $\widetilde{L}_S^n$ we have

$$\begin{aligned}\eta_t \nabla V_{t-1}^T \overline{R}(\Delta_{t-1}) &= 2\eta_t(\gamma_{t-1} - \hat{\gamma})^T \nabla \widetilde{L}_S^n(\gamma_{t-1}) \\ &\geq 2\eta_t(\widetilde{L}_S^n(\gamma_{t-1}) - \widetilde{L}_S^n(\hat{\gamma})) \\ &\geq 0\end{aligned}$$

since $\hat{\gamma}$ minimizes $\widetilde{L}_S^n$. With our step size assumptions $\sum_{i=1}^\infty \eta_t = \infty$ and $\sum_{i=1}^\infty \eta_t^2 < \infty$ we can apply the theorem by Robbins and Siegmund (1971) to get that $V_t \to V_\infty$ where V_∞ is bounded.

We now conclude this section of this proof by showing V_t actually converges to 0. The idea is to show that the bound in Eq. (8) contracts V_t. We show that $\widetilde{L}_S^n$ is strongly convex in a neighborhood around $\hat{\gamma}$. For all $\gamma \in B(\hat{\gamma}, r)$ the smallest eigenvalue of the Hessian is,

$$\lambda_{min}(\nabla \widetilde{L}_S^n) \geq m - M \cdot ||\gamma - \hat{\gamma}||_2.$$

Define the ball with the minimum of the two radii r and $m/(2M)$, and the Hessian at points in this set is bounded below by $m/2 \cdot I_d$. It follows that for all $\gamma \in B(\hat{\gamma}, \min\{r, m/(2M)\})$,

$$\begin{aligned}(\gamma - \hat{\gamma})^T \nabla \widetilde{L}_S^n(\gamma) &= (\gamma - \hat{\gamma})^T \int_0^1 \nabla^2 \widetilde{L}_S^n(\hat{\gamma} + t(\gamma - \hat{\gamma}))(\gamma - \hat{\gamma})\, dt \\ &\geq \frac{m}{2} \cdot ||\gamma - \hat{\gamma}||_2^2.\end{aligned}$$

Earlier we proved that V_t converges to V_∞ which is bounded, therefore, for any $\varepsilon > 0$, there exists R such that $\Pr[\sup_t ||\Delta_t||_2 < R] \geq 1 - \varepsilon$. Let $\tau_R = \inf\{t \geq 1 : ||\Delta_t||_2 \leq R\}$ be the stopping time. Then on $\{t < \tau_R\}$ we have from Eq. (8) and bound above

$$\mathbb{E}[V_t I(t < \tau_R)] \leq \mathbb{E}\left[\left(1 - \eta_t \frac{m}{2} + \eta_t^2 K\right) \cdot V_{t-1} I(t - 1 < \tau_R)\right] + \eta_t^2 K.$$

The proof by Polyak and Juditsky (1992) concludes that $\mathbb{E}[V_t I(t < \tau_R)]$ is upper bounded by $K \cdot \eta_t$. Thus, as the step size converges to zero, V_t converges to zero, the error $||\Delta_t||_2^2 \to 0$.

In the second part of the proof we show that the distribution of error from the process $\{\Delta_t\}_{t \geq 1}$ is asymptotically normal. Polyak and Juditsky (1992) prove this by defining a proxy process $\{\Delta'_t\}_{t \geq 1}$ which is a linear approximation of the true process $\{\Delta_t\}_{t \geq 1}$. Then they apply their Theorem 1 to Δ'_t, which shows that the distribution of errors is asymptotically normal. Lastly, they conclude with showing that the two process are asymptotically equivalent.

For our setting, the proxy process is defined as

$$\begin{aligned}
\Delta'_t &= \Delta'_{t-1} - \eta_t G_n \Delta'_{t-1} + \eta_t \xi_t, \quad \Delta'_0 = \Delta_0 \\
\overline{\Delta}_t &= \frac{1}{t} \sum_{i=0}^{t-1} \Delta'_i
\end{aligned}$$

where $G_n = \nabla^2 \widetilde{L}_S^n(\hat{\gamma})$. Then using the fact that the locally Lipschitz-continuous Hessian around $\hat{\gamma}$, the matrix satisfies for all $\gamma \in B(\hat{\gamma}, r)$,

$$\begin{aligned}
||\nabla \widetilde{L}_S^n(\gamma) - G_n(\gamma - \hat{\gamma})||_2 &\leq \left(\int_0^1 ||\nabla^2 \widetilde{L}_S^n(\hat{\gamma} + t(\gamma - \hat{\gamma})) - \nabla^2 \widetilde{L}_S^n(\hat{\gamma})||_{op} \, dt\right) \cdot ||\gamma - \hat{\gamma}||_2 \\
&\leq M \cdot ||\gamma - \hat{\gamma}||_2^2.
\end{aligned}$$

The final assumptions used in the proof of Theorem 2 revolve around the error decomposition $\xi_t = \xi_t(0) + \zeta_t(\Delta_{t-1})$ where

$$\begin{aligned}
\xi_t(0) &= \nabla \ell_S(\hat{\gamma}; \tilde{z}_t) - \nabla \widetilde{L}_S^n(\hat{\gamma}) \\
\zeta_t(\gamma_{t-1}) &= [\nabla \ell_S(\gamma_{t-1}; \tilde{z}_t) - \nabla \ell_S(\hat{\gamma}; \tilde{z}_t)] + [\nabla \widetilde{L}_S^n(\hat{\gamma}) - \nabla \widetilde{L}_S^n(\gamma_{t-1})].
\end{aligned}$$

which satisfy the following properties

$$\mathbb{E}[\xi_t(0) | \mathcal{F}_{t-1}] = 0 \quad a.s. \tag{9}$$

$$\mathbb{E}[\xi_t(0)\xi_t(0)^T | \mathcal{F}_{t-1}] \xrightarrow{P} S \quad \text{as } t \to \infty; \quad S > 0 \tag{10}$$

$$\sup_t \mathbb{E}[||\xi_t(0)||_2^2 I(||\xi_t(0)||_2 > C) | \mathcal{F}_{t-1}] \xrightarrow{P} 0 \quad \text{as } C \to \infty \tag{11}$$

$$\exists \, T, \forall t \geq T, \mathbb{E}[||\zeta_t(\Delta_{t-1})||_2^2 | \mathcal{F}_{t-1}] \leq \delta(\Delta_{t-1}) \quad a.s. \tag{12}$$

with $\delta(\Delta_t) \to 0$ as $\Delta_t \to 0$. Given that each update is performed with $\tilde{z}_t$ sampled uniformly from the finite dataset, we have

$$\begin{aligned}
\mathbb{E}[\xi_t(0) | \mathcal{F}_{t-1}] &= \mathbb{E}[\nabla \ell_S(\hat{\gamma}; \tilde{z})] - \nabla \widetilde{L}_S^n(\hat{\gamma}) \\
&= 0 \quad a.s.
\end{aligned}$$

Furthermore, conditioned on the previous samples

$$\mathbb{E}[\xi_t(0)\xi_t(0)^T | \mathcal{F}_{t-1}] = \frac{1}{n} \sum_{i=1}^n \nabla \ell_S(\hat{\gamma}; \tilde{z}_i)(\nabla \ell_S(\hat{\gamma}; \tilde{z}_i))^T =: S$$

which proves Properties (9) and (10). Again, using the fact that updates are performed with independent samples, for all t

$$\mathbb{E}[||\xi_t(0)||_2^2 I(||\xi_t(0)||_2 > C) | \mathcal{F}_{t-1}] = \mathbb{E}[||\xi_t(0)||_2^2 I(||\xi_1(0)||_2 > C)].$$

By Assumption C.0.1 the expectation is finite $\mathbb{E}[||\xi_1(0)||_2^2] < C \cdot (1 + ||\hat{\gamma}||_2^2)$. This turns Property (11) into the equivalent statement

$$\sup_t \mathbb{E}[||\xi_t(0)||_2^2 I(||\xi_t(0)||_2 > C)|\mathcal{F}_{t-1}] = \mathbb{E}[||\xi_1(0)||_2^2 I(||\xi_1(0)||_2 > C)]$$

which converges to zero in probability as $C \to \infty$. Lastly,

$$\begin{aligned}\mathbb{E}[||\zeta_t(\gamma_{t-1})||_2^2|\mathcal{F}_{t-1}] &\leq \mathbb{E}[||\nabla \ell_S(\gamma_{t-1}; \tilde{z}_t) - \nabla \ell_S(\hat{\gamma}; \tilde{z}_t)||_2^2|\mathcal{F}_{t-1}] + ||\nabla \widetilde{L}_S^n(\hat{\gamma}) - \nabla \widetilde{L}_S^n(\gamma_{t-1})||_2^2 \\ &\leq 2M^2 \cdot ||\gamma - \hat{\gamma}||_2^2\end{aligned}$$

which converges to zero once $\Delta_t \to 0$, thus, proving Property (12). □

We also use the following assumptions to ensure concentration occurs.

Assumption C.0.2. *The covariates and labels are such that:*

(i) There exists a constant $\mu > 0$ such that the matrix,

$$\mathbb{E}[Var(y|x)vv^T] \succeq \mu I_d.$$

(ii) There exists constants r, $b > 0$ such that for all $\gamma \in B(\gamma^*, r)$,

$$\sup_{\gamma \in B} ||\nabla \ell_S(\gamma; \tilde{z})||_2 \leq b. \tag{13}$$

C.2 Proof of Theorem 3.3

Proof. By Corollary C.0.1, the sequences $\bar{\alpha}_t \to^{a.s.} \hat{\alpha}$ and $\bar{\beta}_t \to^{a.s.} \hat{\beta}$. This convergence ensures that the subsequent analysis on $\hat{\alpha}$ and $\hat{\beta}$ are valid. Now what's left is to show that $\hat{\alpha} = \hat{\beta}$. The gradient of the per-sample loss $\nabla \ell_T(\theta; x, y) = (h(x^T\theta) - y)x$. Since the teacher estimator $\hat{\theta}$ minimizes the empirical loss, we have

$$\frac{1}{n}\sum_{i=1}^n y_i x_i = \frac{1}{n}\sum_{i=1}^n h(x_i^T\hat{\theta})x_i. \tag{14}$$

The gradient of the per-sample loss of the student is $\nabla \ell_S(\gamma; v, y) = (h(v^T\gamma) - y)v$. Both points $\hat{\alpha}$ and $\hat{\beta}$ satisfies the first-order optimality conditions involving their respective labels, which corresponds to the following equations,

$$\begin{aligned}\frac{1}{n}\sum_{i=1}^n y_i v_i &= \frac{1}{n}\sum_{i=1}^n h(v_i^T\hat{\alpha})v_i, \\ \frac{1}{n}\sum_{i=1}^n h(x_i^T\hat{\theta})v_i &= \frac{1}{n}\sum_{i=1}^n h(v_i^T\hat{\beta})v_i.\end{aligned}$$

The objective L_S^n is strictly convex and thus has a unique optimal point. Furthermore, Eq. 14 shows both $\hat{\alpha}$ and $\hat{\beta}$ satisfy the first-order optimality conditions on either labels. This implies the points are the same. □

C.3 Proof of Theorem 3.4

Proof. Recall that both methods share an optimal point by Theorem 3.3. The Hessians at that optimal point $\hat{\gamma}$ are defined as $H_n(\hat{\gamma}) = 1/n \cdot \sum_{i=1}^n \nabla^2 \ell_S(\hat{\gamma}; z_i)$ and $H_n'(\hat{\gamma}) = 1/n \cdot \sum_{i=1}^n \nabla^2 \ell_S(\hat{\gamma}; z_i')$. The empirical second moments of the gradient noise at $\hat{\gamma}$ are,

$$\begin{aligned}C_n(\hat{\gamma}) &= \frac{1}{n} \cdot \sum_{i=1}^n \nabla \ell_S(\hat{\gamma}; z_i)(\nabla \ell_S(\hat{\gamma}; z_i))^T \\ C_n'(\hat{\gamma}) &= \frac{1}{n} \cdot \sum_{i=1}^n \nabla \ell_S(\hat{\gamma}; z_i')(\nabla \ell_S(\hat{\gamma}; z_i'))^T.\end{aligned}$$

Corollary C.0.1 states under Assumptions C.0.1 and 3.0.2, that the averaged iterates are asymptotically normal,

$$\sqrt{t}(\bar{\alpha}_t - \hat{\gamma}) \xrightarrow{d} \mathcal{N}(0, [H_n(\hat{\gamma})]^{-1} C_n(\hat{\gamma}) [H_n(\hat{\gamma})]^{-1})$$
$$\sqrt{t}(\bar{\beta}_t - \hat{\gamma}) \xrightarrow{d} \mathcal{N}(0, [H'_n(\hat{\gamma})]^{-1} C'_n(\hat{\gamma}) [H'_n(\hat{\gamma})]^{-1}).$$

Observe that the Hessians are equal,

$$H_n(\hat{\gamma}) = \frac{1}{n} \sum_{i=1}^{n} h'(v_i^T \hat{\gamma}) v_i v_i^T = H'_n(\hat{\gamma}),$$

going forward we refer to both of them as H_n. The theorem is now equivalent to showing $C_n(\hat{\gamma}) \succeq C'_n(\hat{\gamma})$. At a high level, we go about proving this is by bounding the deviation of the empirical quantities, namely the parameter estimates as well as the empirical second moments, from their corresponding population quantities.

To show the difference in covariances is positive semi-definite, we want to bound the minimum eigenvalue of $C_n(\hat{\gamma}) - C'_n(\hat{\gamma})$ and show that it is nonnegative. We know that their population counterpart $C(\gamma^*) - C'(\gamma^*) = \mathbb{E}[\mathrm{Var}(y|x) v v^T]$ is positive definite under Assumption 3.2.2. For any symmetric matrices A, $B \in \mathbb{R}^{d \times d}$ we can lower bound the sum of one's eigenvalues by $\lambda_{min}(A) \geq \lambda_{min}(B) + \lambda_{min}(A - B)$. The difference $A - B$ is still symmetric, which means $\lambda_{min}(A - B) \geq -\lambda_{max}(|A - B|) = -||A - B||_{op}$. We apply this identity to write the difference between the empirical second moments evaluated at the minimizers of the empirical objective, in terms of the deviations from their respective population counterparts,

$$\lambda_{min}\left[C_n(\hat{\gamma}) - C'_n(\hat{\gamma})\right] \geq \lambda_{min}\left[C(\gamma^*) - C'(\gamma^*)\right] - ||C_n(\hat{\gamma}) - C(\gamma^*)||_{op} - ||C'_n(\hat{\gamma}) - C'(\gamma^*)||_{op}.$$

Note that the minimum eigenvalue of $C(\gamma^*) - C'(\gamma^*)$ is at least $\mu > 0$. Let $R = \min\{r, m/(2M)\}$ where r is the radius from our assumptions regarding the function's local Lipschitz properties. The function L_S^n is strongly convex around $\hat{\gamma}$, consider $\gamma \in B(\hat{\gamma}, R)$ then

$$\begin{aligned}
\lambda_{min}(\nabla^2 L_S^n(\gamma)) &\geq m - ||\nabla^2 L_S^n(\hat{\gamma}) - \nabla^2 L_S^n(\gamma)||_{op} \\
&\geq m - M \cdot ||\hat{\gamma} - \gamma||_2 \\
&\geq m/2.
\end{aligned}$$

Suppose $\gamma \notin B(\hat{\gamma}, R)$ and let $d = ||\gamma - \hat{\gamma}||_2$, then

$$\begin{aligned}
(\gamma - \hat{\gamma})^T \nabla L_S^n(\gamma) &= (\gamma - \hat{\gamma})^T \int_0^1 \nabla^2 L_S^n(\hat{\gamma} + t(\gamma - \hat{\gamma}))(\gamma - \hat{\gamma}) \, dt \\
&\geq (\gamma - \hat{\gamma})^T \int_0^{R/d} \frac{m}{2}(\gamma - \hat{\gamma}) \, dt \\
&\geq \frac{mR}{2d} ||\gamma - \hat{\gamma}||_2^2.
\end{aligned}$$

Then by Cauchy-Schwarz the gradient norm at γ is $||\nabla L_S^n(\gamma)||_2 \geq mR/2$. This implies that if $||\nabla L_S^n(\gamma)||_2 < mR/2$, then $\gamma \in B(\hat{\gamma}, R)$. We show for sufficiently large n, the gradient norm at γ^* is less than $mR/2$. We use Assumption C.0.1 to bound each coordinate of $\nabla L_S^n(\gamma^*)$ via Hoeffding's inequality, we get for all $t \geq 0$

$$\Pr[|(\nabla L_S^n(\gamma^*))_i| \geq t] \leq 2 \exp\left(-\frac{nt^2}{2b^2}\right).$$

Then a union bound over all coordinates gives us, with probability $\geq 1 - \delta$,

$$||\nabla L_S^n(\gamma^*)||_\infty \leq b\sqrt{\frac{2\log(2d/\delta)}{n}}.$$

And the ℓ_2 bound follows, $||\nabla L_S^n(\gamma^*)||_2 \leq \sqrt{d} \cdot ||\nabla L_S^n(\gamma^*)||_\infty$. We then take n large enough so that the gradient norm is less than $mR/2$, which implies $\gamma^* \in B(\hat{\gamma}, R)$. Consider the term involving $C_n(\hat{\gamma}) - C(\gamma^*)$,

$$||C_n(\hat{\gamma}) - C(\gamma^*)||_{op} \leq ||C_n(\hat{\gamma}) - C_n(\gamma^*)||_{op} + ||C_n(\gamma^*) - C(\gamma^*)||_{op}.$$

By Assumptions C.0.1 and C.0.2, we know that the per-sample gradients are locally Lipschitz and bounded in $B(\gamma^*, r)$, it follows that

$$\begin{aligned}||C_n(\hat{\gamma}) - C_n(\gamma^*)||_{op} &\leq \frac{1}{n}\sum_{i=1}^{n} ||\nabla\ell_S(\hat{\gamma}; z_i)(\nabla\ell_S(\hat{\gamma}; z_i))^T - \nabla\ell_S(\gamma^*; z_i)(\nabla\ell_S(\gamma^*; z_i))^T||_{op} \\ &\leq \frac{1}{n}\sum_{i=1}^{n} ||\nabla\ell_S(\hat{\gamma}; z) - \nabla\ell_S(\gamma^*; z)||_2(||\nabla\ell_S(\hat{\gamma}; z)||_2 + ||\nabla\ell_S(\gamma^*; z)||_2) \\ &\leq 2M \cdot b||\hat{\gamma} - \gamma^*||_2.\end{aligned}$$

Now we derive a bound on the second term using the matrix Bernstein inequality (Tropp, 2015). For sample $(x_i, y_i) \sim \mathcal{P}$ where $z_i = (v_i, y_i)$ and $v_i = \Pi x_i$, let

$$M_i := \nabla\ell_S(\gamma^*; z_i)(\nabla\ell_S(\gamma^*; z_i))^T - C(\gamma^*).$$

Then $\mathbb{E}[M_i] = 0$ by definition. Under boundedness (Assumption C.0.2),

$$\begin{aligned}||C(\gamma^*)||_{op} &\leq \mathbb{E}[||\nabla\ell_S(\gamma^*; z)(\nabla\ell_S(\gamma^*; z))^T||_{op}] \\ &= \mathbb{E}[||\nabla\ell_S(\gamma^*; z)||_2^2] \leq b^2.\end{aligned}$$

We then have an upper bound on the variance proxy via the triangle inequality,

$$\begin{aligned}||\mathbb{E}[M_i^2]||_{op} &\leq \mathbb{E}[||M_i||_{op}^2] \\ &\leq \mathbb{E}[(||\nabla\ell_S(\gamma^*; z)||_2^2 + ||C(\gamma^*)||_{op})^2] \\ &\leq 4b^4.\end{aligned}$$

Using these quantities along with the matrix Bernstein inequality, for all $t \geq 0$

$$\begin{aligned}\Pr\left[||C_n(\gamma^*) - C(\gamma^*)||_{op} \geq t\right] &= \Pr\left[\left|\left|\frac{1}{n}\cdot\sum_{i=1}^{n} M_i\right|\right|_{op} \geq t\right] \\ &\leq 2d \cdot \exp\left(-\frac{nt^2/2}{4b^4 + 2b^2t/3}\right).\end{aligned}$$

So with probability $\geq 1 - \delta$, we have

$$||C_n(\gamma^*) - C(\gamma^*)||_{op} \leq b^2\left(\sqrt{\frac{8\cdot\log(2d/\delta)}{n}} + \frac{2\cdot\log(2d/\delta)}{3n}\right).$$

A sufficiently large n means that $||C_n(\hat{\gamma}) - C(\gamma^*)||_{op} \leq \varepsilon'$ with high probability. Given the same assumptions but for $\ell_S(\gamma; z')$, L'_S, and L'^n_S, the prior argument applies verbatim. Thus,

$$\begin{aligned}||C'_n(\hat{\gamma}) - C'(\gamma^*)||_{op} &\leq ||C'_n(\hat{\gamma}) - C'_n(\gamma^*)||_{op} + ||C'_n(\gamma^*) - C'(\gamma^*)||_{op} \\ &\leq \varepsilon'\end{aligned}$$

with high probability. Let $\varepsilon' = \mu/4$. We then have with high probability,

$$||C_n(\hat{\gamma}) - C(\gamma^*)||_{op} + ||C'_n(\hat{\gamma}) - C'(\gamma^*)||_{op} \leq \mu/2$$

which means the minimum eigenvalue of $C_n(\hat{\gamma}) - C'_n(\hat{\gamma})$ is at least $\mu/2$. Lastly, we can conclude that for sufficiently large n,

$$H_n^{-1}C_n(\hat{\gamma})H_n^{-1} \succeq H_n^{-1}C'_n(\hat{\gamma})H_n^{-1}$$

holds with high probability. □

D Detailed statements and deferred proofs for Section 4

We introduce the notation $\widetilde{M} \in \{M, M'\}$ to simplify the presentation of our results when the argument for either target matrix is identical.

To begin, we may assume that $\widetilde{M}$ is diagonal. To see this, consider the eigendecomposition $\widetilde{M} = U\Lambda U^T$. Since $\widetilde{M}$ is symmetric, U is an orthogonal matrix. For the rotated iterates $y_t = U^T x_t$, the update rule is

$$y_{t+1} = y_t - \eta(y_t y_t^T - \Lambda)y_t.$$

Thus, the dynamics are equivalent up to a rotation, so it is no loss of generality to work in the eigenbasis of $\widetilde{M}$. Now let's make some general observations about the gradient descent dynamics.

Lemma D.0.1 (Bounded magnitude). *Under initialization $x_0 \sim \mathcal{N}(0, k \cdot \lambda_1/n \cdot I_n)$ and for step size $\eta < 1/(L \cdot (\lambda_1 + 1))$ where $0 < k < 1$ is a chosen constant and $L \geq 4$, elements of the gradient descent sequence given by the update:*

$$x_{t+1} = x_t - \eta(x_t x_t^T - \widetilde{M})x_t,$$

have squared norm most λ_1 with probability $1 - k$ over the initialization.

Proof. We first consider the event that the squared norm at initialization is less than λ_1. By Markov's inequality we have $\Pr[||x_0||_2^2 \geq \lambda_1] \leq \mathbb{E}[||x_0||_2^2]/\lambda_1 = k$ since $\mathbb{E}[||x_0||_2^2] = k \cdot \lambda_1$. It follows that $\Pr[||x_0||_2^2 < \lambda_1] > 1 - k$.

We have the following gradient update for the squared norm:

$$\begin{aligned} ||x_{t+1}||_2^2 &= \sum_{i=1}^{n} (1 - \eta(||x_t||_2^2 - \lambda_i))^2 x_{i,t}^2 \\ &\leq (1 - \eta(||x_t||_2^2 - \lambda_1))^2 ||x_t||_2^2 \end{aligned}$$

where the last inequality follows from taking the update with the largest possible eigenvalue, which is λ_1. Note that by choice of η we have $(1 - \eta(||x_t||_2^2 - \lambda_i)) \geq 0$ for all $i = 1, \ldots, n$. Now we proceed via induction conditioned on the event that the initial squared norm is less than λ_1. Since the base case is satisfied, moving to the inductive step, assume $||x_t||_2^2 \leq \lambda_1$. Consider the following function f and its derivative,

$$\begin{aligned} f(z) &= (1 - \eta(z - \lambda_1))^2 z, \\ f'(z) &= (1 - \eta(z - \lambda_1))^2 - 2\eta(1 - \eta(z - \lambda_1))z \\ &= (1 - \eta(z - \lambda_1))(1 - \eta(3z - \lambda_1)) \geq 0 \end{aligned}$$

The derivative is positive for all $z \in [0, \lambda_1]$ since $\eta < 1/(L \cdot (\lambda_1 + 1))$. It follows that in this interval, $f(z) \leq f(\lambda_1) = \lambda_1$. From the inductive hypothesis, $||x_{t+1}||_2^2 \leq f(||x_t||_2^2) \leq \lambda_1$, so we can conclude that $||x_t||_2^2 \leq \lambda_1$ for all $t \geq 1$. □

D.1 Proof of Theorem 4.1

Proof. The convergence of GD sequences w_t and u_t to their respective local minima is a result of Theorem 3 in Panageas and Piliouras (2017), which relaxes the globally Lipschitz assumption required in Lee et al. (2016). This is necessary for the simple reason that ∇L_S and $\nabla L_S'$ are not globally Lipschitz. To demonstrate this for ∇L_S consider $x_1 = (\kappa + 1) \cdot v$ and $x_2 = \kappa \cdot v$ where $v \in \mathbb{R}^n$ has unit norm. Clearly, $||x_1 - x_2||_2 = 1$ so now derive,

$$\begin{aligned} \nabla L_S(x_1) - \nabla L_S(x_2) &= (||x_1||_2^2 I - M)x_1 - (||x_2||_2^2 I - M)x_2 \\ &= ((\kappa + 1)^2 I - M)(\kappa + 1) \cdot v - (\kappa^2 I - M)\kappa \cdot v \\ &= (3\kappa^2 + 3\kappa + 1) \cdot v - Mv, \end{aligned}$$

then by the reverse triangle inequality we have,

$$||\nabla L_S(x_1) - \nabla L_S(x_2)||_2 \geq |(3\kappa^2 + 3\kappa + 1) - ||Mv||_2|.$$

The term $||Mv||_2$ is a constant and κ can be made arbitrarily large, making it impossible for $||\nabla L_S(x_1) - \nabla L_S(x_2)||_2 \leq L||x_1 - x_2||_2$ for a constant $L \geq 0$.

The functions ∇L_S and $\nabla L'_S$ are locally Lipschitz on the open ball of radius $\sqrt{\lambda_1 + \varepsilon}$ for a small constant $0 < \varepsilon < 1$,

$$\mathcal{B} = \{x : ||x||_2 < \sqrt{\lambda_1 + \varepsilon}\}.$$

This set is forward invariant under the two maps $g(x) = x - \eta \nabla L_S(x)$ and $g'(x) = x - \eta \nabla L'_S(x)$ because we've assumed that at initialization, $||x_0||_2 \leq \sqrt{\lambda_1}$, so directly applying Lemma D.0.1 allows us to conclude $g(\mathcal{B}) \subseteq \mathcal{B}$ and $g'(\mathcal{B}) \subseteq \mathcal{B}$. On this set,

$$\begin{aligned}
\sup_{x \in \mathcal{B}} ||\nabla^2 L_S(x)||_{op} &= \sup_{x \in \mathcal{B}} ||2xx^T + ||x||_2^2 \cdot I - M||_{op} \\
&\leq \sup_{x \in \mathcal{B}} ||2xx^T + ||x||_2^2 \cdot I||_{op} + ||M||_{op} \\
&\leq 4\lambda_1 + 3 < \infty.
\end{aligned}$$

In the last inequality, we use the fact that $x \in \mathcal{B}$ implies $||x||_2 < \sqrt{\lambda_1 + 1}$. Since the target matrix in L'_S is the truncated eigendecomposition of M to the top-r eigenvalues, $||M'||_{op} = \lambda_1$. Consequently, the supremum of the operator norm of $\nabla^2 L'_S(x)$ over this set is still upper bounded by $4\lambda_1 + 3$. Then by Theorem 3 in Panageas and Piliouras (2017), the set of initial conditions $x \in \mathcal{B}$ so that either of the gradient descent trajectories with step size $0 < \eta < 1/(4\lambda_1 + 3)$ converges to a strict saddle point is of measure zero.

Now if we derive the gradient of our objectives, $\nabla L_S(x) = (xx^T - M)x$ and $\nabla L'_S(x) = (xx^T - M')x$. It follows that the set of all critical points is $\{\pm\sqrt{\lambda_i} e_i\}_{i=1}^n \cup \{0\}$ for L_S and $\{\pm\sqrt{\lambda_i} e_i\}_{i=1}^r \cup \{0\}$ for L'_S. All points except for $\sqrt{\lambda_1} e_1$ and $-\sqrt{\lambda_1} e_1$ are strict saddles. To see this, compute the Hessians and evaluate them at each critical point,

$$\begin{aligned}
\nabla^2 L_S(\pm\sqrt{\lambda_i} e_i) &= \lambda_i \cdot I + 2\lambda_i e_i e_i^T - M \\
\nabla^2 L'_S(x) &= \lambda_i \cdot I + 2\lambda_i e_i e_i^T - M'.
\end{aligned}$$

Along the direction e_1, they are $e_1^T \nabla^2 L_S(\pm\sqrt{\lambda_i} e_i) e_1 = (\lambda_i - \lambda_1) < 0$ and $e_1^T \nabla^2 L'_S(\pm\sqrt{\lambda_i} e_i) e_1 = (\lambda_i - \lambda_1) < 0$ when $i \neq 1$. Also note that the Hessians at 0 are simply $-M$ and $-M'$ respectively, and both have a negative eigenvalue $-\lambda_1$. Thus, all critical points except $\sqrt{\lambda_1} e_1$ and $-\sqrt{\lambda_1} e_1$ are strict saddles, and consequently, gradient descent avoids them. At $x = \pm\sqrt{\lambda_1} e_1$, the Hessians are both positive semi-definite, so these points are local minima (and also global minima since $\lambda_1 > \lambda_2$) of both objectives. □

D.2 Full statement and proof of Lemma 4.1.1

Lemma D.0.2 (Supervised learning lower bound)**.** *Given $\varepsilon > 0$, let the step size be*

$$0 < \eta < \frac{1}{L \cdot (\lambda_1 + 1)}$$

where λ_1 is the largest eigenvalue of target matrix M and $L \geq 4$. Assume the initial point w_0 satisfies $||w_0||_2 \leq \sqrt{\lambda_1}$. Then $T(w)$ has the following lower bound,

$$T(w) \geq \frac{1}{2\log(1/(1 - \eta\Delta_n))} \log\left(\frac{(1-\varepsilon) \cdot ||w_{2:n,0}||_2^2 / w_{1,0}^2}{\varepsilon}\right).$$

Proof of Lemma D.0.2. We begin by introducing some notation. For a vector w_t, its orthogonal projection onto the i-th eigenvector of M, v_i, is $P_{v_i} w_t$ where $P_{v_i} = v_i v_i^T$. As we're working with a diagonal M, $P_{v_i} w_t = w_{i,t}$ and we have the following one step update:

$$w_{i,t+1} = (1 - \eta(||w_t||_2^2 - \lambda_i)) \cdot w_{i,t}. \tag{15}$$

In general, any vector w_t can be decomposed as,

$$w_t = P_{v_1} w_t + P_{v_2,\ldots,v_n} w_t$$

where P_{v_1} and $P_{v_2,\ldots,v_n}$ are orthogonal projections onto the eigenvector(s) v_1 and $\{v_2,\ldots,v_n\}$ respectively. Now we can define the following quantity:

$$c_t = \frac{||P_{v_2,\ldots,v_n} w_t||_2}{||P_{v_1} w_t||_2} = \frac{||w_{2:n,t}||_2}{|w_{1,t}|}$$

where the last equality is again from working with a diagonal M. This value measures the tangent of the angle between w_t (projected onto the eigenvectors orthogonal to v_1) and w^*. When the direction of w_t is perfectly aligned with w^* it is zero. Let's now consider the sequence $\{c_t^2\}_{t\geq 1}$, it satisfies

$$c_{t+1}^2 \geq \left(\frac{1-\eta(||w_t||_2^2 - \lambda_n)}{1-\eta(||w_t||_2^2 - \lambda_1)}\right)^2 \cdot c_t^2,$$

where the lower bound is from Eq. 15 with the smallest eigenvalue λ_n. Adding and subtracting by $\eta\lambda_1$ in the numerator decouples the overall magnitude from the update,

$$c_{t+1}^2 \geq \left(1 - \eta\frac{\lambda_1 - \lambda_n}{1-\eta(||w_t||_2^2 - \lambda_1)}\right)^2 \cdot c_t^2 \tag{16}$$

$$\geq (1-\eta\Delta_n)^2 \cdot c_t^2 \tag{17}$$

where we define $\Delta_n := \lambda_1 - \lambda_n$. The lower bound uses $||w_t||_2^2 \leq \lambda_1$ for all $t \geq 1$ from Lemma D.0.1, which means the fraction $1/(1-\eta(||w_t||_2^2 - \lambda_1)) \leq 1$.

We now connect the quantity c_t to the main goal of lower bounding $T(w)$ for a given tolerance $\varepsilon > 0$. Observe that $||w_{2:n,t}||_2^2 = w_{1,t}^2 a_t^2$. As the eigenvectors of M are orthogonal (here they are also the basis vectors $e_1,\ldots,e_n$), the distance from w_t to w^* can be written as,

$$\begin{aligned} ||w_t - w^*||_2^2 &= (\sqrt{\lambda_1} - w_{1,t})^2 + ||w_{2:n,t}||_2^2 \\ &= (\sqrt{\lambda_1} - w_{1,t})^2 + w_{1,t}^2 c_t^2. \end{aligned}$$

Consider function,

$$g(z) = (\sqrt{\lambda_1} - z)^2 + z^2 \cdot c_t^2,$$

treating c_t^2 as fixed. It's strictly convex, so by taking the derivative $g'(z) = 2z \cdot c_t^2 - 2(\sqrt{\lambda_1} - z)$ we find that the function is minimized at $z = \sqrt{\lambda_1}/(1+c_t^2)$. From this we get the lower bound,

$$||w_t - w^*||_2^2 \geq \lambda_1 \frac{c_t^2}{1+c_t^2}.$$

The distance from the optimum can therefore be estimated from just angular information. By Eq. 17, iterating the bound shows that $c_t^2 \geq (1-\eta\Delta_n)^{2t} \cdot c_0^2$. The quantity c_t^2 must be at least $\varepsilon/(1-\varepsilon)$ in order for $||w_t - w^*||_2^2$ to be at least $\lambda_1 \cdot \varepsilon$. Solving for t at the desired value of c_t yields a lower bound on the steps,

$$T(w) \geq \frac{1}{2\log(1/(1-\eta\Delta_n))} \log\left(\frac{(1-\varepsilon)\cdot c_0^2}{\varepsilon}\right).$$

Note that the notation used in the main body of this paper is $\textsc{Im}(w_0) = c_0^2$. □

D.3 Full statement and proof of Lemma 4.1.2

Lemma D.0.3 (Knowledge distillation upper bound). *Given $\varepsilon > 0$, let the step size be*

$$0 < \eta < \frac{1}{L\cdot(\lambda_1+1)}$$

where λ_1 is the largest eigenvalue of target matrix M' and $L \geq 4$. Assume the initial point u_0 satisfies $||u_0||_2 \leq \sqrt{\lambda_1}$. Let $\theta, \xi, \gamma \in (0,1)$ be fixed constants, and define

$$\ell := \frac{L}{L+1}, \quad \kappa := \frac{1-\xi}{1+\gamma\varepsilon}.$$

Define the following quantities,

$$t_a := \frac{1}{2\log(1/(1-\ell\eta\Delta_2))}\log\left(\frac{||u_{2:r,0}||_2^2/u_{1,0}^2}{\varepsilon_a}\right),$$
$$t_b := \frac{1}{2\log(1/(1-\ell\eta\lambda_1))}\log\left(\frac{||u_{r+1:n,0}||_2^2/u_{1,0}^2}{\varepsilon_b}\right),$$
$$t_\xi := \frac{1}{2\log(1+\eta\xi\lambda_1)}\log\left(\frac{(1-\xi)\lambda_1}{u_{1,0}^2}\right),$$
$$t_\delta := \frac{1}{\log(1/(1-2\eta\kappa\lambda_1))}\log\left(\frac{1-\kappa}{\varepsilon_\delta - \kappa^{-1}(\varepsilon_a+\varepsilon_b)}\right).$$

Here, the error tolerances are given by

$$\varepsilon_a := (\gamma\theta)\cdot\varepsilon, \quad \varepsilon_b := (1-\theta)\gamma\cdot\varepsilon, \quad \varepsilon_\delta := (1-\gamma)\cdot\varepsilon.$$

Then $T(u)$ has the following upper bound

$$T(u) \le \max\{t_a, t_b, t_\xi\} + t_\delta.$$

Proof of Lemma D.0.3. Similar to the proof of the lower bound, we begin by observing that any vector u_t has the decomposition,

$$u_t = P_{v_1}u_t + P_{v_2,\ldots,v_r}u_t + P_{v_{r+1},\ldots,v_n}u_t$$

where P_{v_1}, $P_{v_2,\ldots,v_r}$, and $P_{v_{r+1},\ldots,v_n}$ are orthogonal projections onto the eigenvector(s) v_1, $\{v_2,\ldots,v_r\}$, and $\{v_{r+1},\ldots,v_n\}$ respectively. Note that $v_{r+1},\ldots,v_n$ correspond to the eigenvalue 0 and span the nullspace. Now we define the following quantities:

$$a_t = \frac{||P_{v_2,\ldots,v_r}u_t||_2}{||P_{v_1}u_t||_2} = \frac{||u_{2:r,t}||_2}{|u_{1,t}|},$$
$$b_t = \frac{||P_{v_{r+1},\ldots,v_n}u_t||_2}{||P_{v_1}u_t||_2} = \frac{||u_{r+1:n,t}||_2}{|u_{1,t}|}$$

which represent tangent of the angle between u_t, projected onto the different subspaces, and u^*. The reason we introduce these two quantities is that the distance $||u_t - u^*||_2^2$ can be upper bounded by two monotonically decreasing sequences,

$$\begin{aligned}||u_t - u^*||_2^2 &= (\sqrt{\lambda_1} - u_{1,t})^2 + ||u_{2:n,t}||_2^2\\ &= (\sqrt{\lambda_1} - u_{1,t})^2 + u_{1,t}^2(a_t^2 + b_t^2)\\ &\le (\lambda_1 - u_{1,t}^2) + \lambda_1(a_t^2+b_t^2).\end{aligned}$$

The last inequality follows from the identity $(\sqrt{\lambda_1} - u_{1,t})^2 \le (\sqrt{\lambda_1} - |u_{1,t}|)(\sqrt{\lambda_1} + |u_{1,t}|)$. Additionally, Lemma D.0.1 shows the magnitude $||u_t||_2^2 \le \lambda_1$ is bounded for all $t \ge 1$ with high probability over the initialization, the first coordinate $u_{1,t}$ also inherits this bound. We can upper bound $\{a_t^2\}_{t\ge1}$ with a monotonically decreasing sequence,

$$a_{t+1}^2 \le \left(\frac{1-\eta(||u_t||_2^2 - \lambda_2)}{1-\eta(||u_t||_2^2-\lambda_1)}\right)^2 \cdot a_t^2 \tag{18}$$
$$= \left(1 - \eta\frac{\lambda_1-\lambda_2}{1-\eta(||u_t||_2^2-\lambda_1)}\right)^2 \cdot a_t^2 \tag{19}$$
$$\le (1-\eta\ell\Delta_2)^2 \cdot a_t^2 \tag{20}$$

where we define $\Delta_2 := \lambda_1 - \lambda_2$. The constant $\ell := L/(L+1)$ appears due to the upper bound $(1-\eta(||u_t||_2^2 - \lambda_1)) \leq 1 + \eta\lambda_1 \leq (L+1)/L$ since the step size is chosen as $\eta < 1/(L \cdot (\lambda_1 + 1))$ and we can drop the additional 1 in the denominator. A similar argument shows the sequence $\{b_t\}_{t \geq 1}$ satisfies,

$$b_{t+1}^2 = \left(\frac{1 - \eta||u_t||_2^2}{1 - \eta(||u_t||_2^2 - \lambda_1)}\right)^2 \cdot b_t^2 \tag{21}$$

$$= \left(1 - \eta\frac{\lambda_1}{1 - \eta(||u_t||_2^2 - \lambda_1)}\right)^2 \cdot b_t^2 \tag{22}$$

$$\leq (1 - \eta\ell\lambda_1)^2 \cdot b_t^2. \tag{23}$$

The key detail is that the rate depends on the top eigenvalue λ_1 instead of the spectral gap Δ_i. Consequently, the b_t terms (potentially) decay at a much faster rate when the magnitude of the eigenvalues is large, but the spectrum is tightly clustered.

Let ε_a, $\varepsilon_b > 0$ be given tolerances. Solve for the number of steps such that $a_t \leq \varepsilon_a$ using Eq. 20,

$$\begin{aligned} t_a &:= \min\{t : a_t^2 \leq \varepsilon_a\} \\ &\leq \frac{1}{2\log(1/(1 - \eta\ell\Delta_2))} \log\left(\frac{a_0^2}{\varepsilon_a}\right). \end{aligned}$$

Likewise, using Eq. 23 the number of steps such that $b_t \leq \varepsilon_b$ is

$$\begin{aligned} t_b &= \min\{t : b_t^2 \leq \varepsilon_b\} \\ &\leq \frac{1}{2\log(1/(1 - \eta\ell\lambda_1))} \log\left(\frac{b_0^2}{\varepsilon_b}\right). \end{aligned}$$

Now we analyze the growth of the first coordinate $u_{1,t}$. It has the following update equation,

$$u_{1,t+1} = (1 - \eta(||u_t||_2^2 - \lambda_1))u_{1,t}. \tag{24}$$

The analysis consists of two phases, namely, the initial growth of the squared norm $||u_t||_2^2$ to $(1-\xi)\lambda_1$ and then a final convergence stage of $u_{1,t}^2$ to λ_1. For the latter, we specifically derive an update for $\delta_t := \lambda_1 - u_{1,t}^2$ which corresponds to the error in magnitude at step t. Starting with the full magnitude, while $||u_t||_2^2 < (1-\xi)\lambda_1$ the update on the full magnitude is lower bounded by the update on the first coordinate,

$$||u_{t+1}||_2^2 \geq (1 + \eta\xi\lambda_1)^2 u_{1,t}^2.$$

It follows that $||u_t||_2^2 \geq (1 - \eta\xi\lambda_1)^{2t} u_{1,0}^2$. Then solving for t leads to the upper bound on steps to reach $(1-\xi)\lambda_1$,

$$\begin{aligned} t_\xi &:= \min\{t : ||u_t||_2^2 \geq (1-\xi)\lambda_1\} \\ &\leq \frac{1}{2\log(1 + \eta\xi\lambda_1)} \log\left(\frac{(1-\xi)\lambda_1}{u_{1,0}^2}\right). \end{aligned}$$

Now we analyze the final convergence of the first coordinate $u_{1,t}^2$ to target magnitude λ_1, this requires deriving a one-step update for δ_t. At a high level, we know δ_t is monotonically decreasing because the magnitude $||u_t||_2^2$ is at most λ_1. Simply inspecting Eq. 24, shows that the coefficient $(1 - \eta(||u_t||_2^2 - \lambda_1))$ is always nonnegative. In fact, it is always positive unless the magnitude is exactly λ_1, which means $u_{1,t}^2$ is increasing and δ_t is decreasing. To derive the actual update equation, observe that the sum of the magnitude gap at time t, δ_t, and the growth of the first coordinate from step t to $t+1$ is exactly δ_{t+1},

$$\begin{aligned} \delta_{t+1} &= (\lambda_1 - u_{1,t}^2) + (u_{1,t}^2 - u_{1,t+1}^2) \\ &= \delta_t + u_{1,t}^2 - u_{1,t+1}^2. \end{aligned} \tag{25}$$

Now we can focus on the growth captured by $u_{1,t}^2 - u_{1,t+1}^2$. Using Eq. 24, the coordinate at $t+1$ can be written in terms of $u_{1,t}$ as follows,

$$\begin{aligned} u_{1,t}^2 - u_{1,t+1}^2 &= u_{1,t}^2 - (1 - \eta(||u_t||_2^2 - \lambda_1))^2 \cdot u_{1,t}^2 \\ &= u_{1,t}^2 - \left[1 - 2\eta(||u_t||_2^2 - \lambda_1) + \eta^2(||u_t||_2^2 - \lambda_1)^2\right] \cdot u_{1,t}^2 \\ &= \left[2\eta(||u_t||_2^2 - \lambda_1) - \eta^2(||u_t||_2^2 - \lambda_1)^2\right] \cdot u_{1,t}^2 \\ &\leq 2\eta(||u_t||_2^2 - \lambda_1) \cdot u_{1,t}^2. \end{aligned}$$

From Lemma D.0.1, we know that the magnitude $||u_t||_2^2 \leq \lambda_1$ is bounded which makes the term involving η^2 nonpositive. Dropping it results in the final inequality. Now write the expansion,

$$\begin{aligned} 2\eta(||u_t||_2^2 - \lambda_1) \cdot u_{1,t}^2 &= 2\eta(u_{1,t}^2 - \lambda_1)u_{1,t}^2 + 2\eta u_{1,t}^2 ||u_{2:n,t}||_2^2 \\ &= -2\eta u_{1,t}^2 \cdot \delta_t + 2\eta u_{1,t}^2 ||u_{2:n,t}||_2^2 \\ &= -2\eta u_{1,t}^2 \cdot \delta_t + 2\eta u_{1,t}^4 (a_t^2 + b_t^2) \end{aligned}$$

where the last equality uses the identity $||u_{2:n,t}||_2^2 = u_{1,t}^2(a_t^2 + b_t^2)$ for the angular quantities defined earlier. Lastly, $2\eta u_{1,t}^4(a_t^2 + b_t^2) \leq 2\eta\lambda_1^2(a_t^2 + b_t^2)$ so that the bounding sequence is cleanly decreasing. Substituting this into Eq. 25, we have

$$\delta_{t+1} \leq (1 - 2\eta u_{1,t}^2)\delta_t + 2\eta\lambda_1^2(a_t^2 + b_t^2).$$

This is the final convergence stage so we want to upper bound the number of additional steps, after $\max\{t_a, t_b, t_\xi\}$, in order for $\delta_t \leq \lambda_1 \cdot \varepsilon_\delta$ for a given tolerance $\varepsilon_\delta > 0$. To be precise,

$$t_\delta := \min\{t \geq \max\{t_a, t_b, t_\xi\} : \delta_t \leq \lambda_1 \cdot \varepsilon_\delta\}.$$

After $T_1 := \max\{t_a, t_b, t_\xi\}$, the top coordinate $u_{1,t}^2$ occupies most of the magnitude. This is from the identity $u_{1,t}^2 = ||u_t||_2^2/(1 + a_t^2 + b_t^2)$. For $t \geq \max\{t_a, t_b, t_\xi\}$, we've already proven that $a_t^2 + b_t^2 \leq \varepsilon_a + \varepsilon_b$ as well as $||u_t||_2^2 \geq (1 - \xi) \cdot \lambda_1$. To simplify notation, let $\kappa := (1 - \xi)/(1 + \varepsilon_a + \varepsilon_b)$. The first coordinate is then at least $u_{1,t}^2 \geq \kappa \cdot \lambda_1$ for all $t \geq T_1$ and the difference at time T_1 is $\delta_{T_1} \leq (1 - \kappa) \cdot \lambda_1$. Now we can write an upper bound for δ_t that holds when $t \geq \max\{t_a, t_b, t_\xi\}$,

$$\begin{aligned} \delta_{t+1} &\leq (1 - 2\eta\kappa\lambda_1)\delta_t + 2\eta\lambda_1^2(a_t^2 + b_t^2) \\ &\leq (1 - 2\eta\kappa\lambda_1)^{t'+1}\delta_{T_1} + 2\eta\lambda_1^2 \cdot (\varepsilon_a + \varepsilon_b) \sum_{i=0}^{t'} (1 - 2\eta\kappa\lambda_1)^i \end{aligned}$$

where $t' := t - T_1$. For the last inequality we unroll the recursion and use the fact that $(a_t^2 + b_t^2)$ is nonincreasing and has the value $(a_{T_1}^2 + b_{T_1}^2) \leq \varepsilon_a + \varepsilon_b$ at step T_1. Replacing the finite sum with an infinite sum to remove the dependence on t', we arrive at the final upper bound

$$\begin{aligned} \delta_t &\leq (1 - 2\eta\kappa\lambda_1)^{t'}\delta_{T_1} + 2\eta\lambda_1^2 \cdot (\varepsilon_a + \varepsilon_b) \cdot \frac{1}{1 - (1 - 2\eta\kappa\lambda_1)} \\ &= (1 - 2\eta\kappa\lambda_1)^{t'}\delta_{T_1} + \frac{\lambda_1 \cdot (\varepsilon_a + \varepsilon_b)}{\kappa}, \end{aligned}$$

from which we get the step upper bound,

$$t_\delta \leq \frac{1}{\log(1/(1 - 2\eta\kappa\lambda_1))} \log\left(\frac{1 - \kappa}{\varepsilon_\delta - (\varepsilon_a + \varepsilon_b)/\kappa}\right).$$

Then at step $t = \max\{t_a, t_b, t_\xi\} + t_\delta$, the following is true,

$$\begin{aligned} a_t^2 + b_t^2 &\leq \varepsilon_a + \varepsilon_b \\ \delta_t &\leq \lambda_1 \cdot \varepsilon_\delta, \end{aligned}$$

and consequently,

$$||u_t - u^*||_2^2 \leq \lambda_1 \cdot (\varepsilon_a + \varepsilon_b + \varepsilon_\delta).$$

Lastly, $\varepsilon_a = (\gamma\theta) \cdot \varepsilon$, $\varepsilon_b = (1 - \theta)\gamma \cdot \varepsilon$, and $\varepsilon_\delta = (1 - \gamma) \cdot \varepsilon$ for constants θ, ξ, $\gamma \in (0, 1)$. Therefore, the distance is bounded $||u_t - u^*||_2^2 \leq \lambda_1 \cdot \varepsilon$. The overall upper bound is then $T(u) \leq \max\{t_a, t_b, t_\xi\} + t_\delta$. □

D.4 Proof of Theorem 4.2

Proof of Theorem 4.2. From Lemmas D.0.2 and D.0.3, we have a lower bound on $T(w)$,

$$T(w) \geq \frac{1}{2\log(1/(1-\eta\Delta_n))} \log\left(\frac{(1-\varepsilon)\cdot c_0^2}{\varepsilon}\right)$$

and an upper bound on $T(u)$ (taking $\xi = 1/2$),

$$T(u) \leq \max\{t_a, t_b, t_\xi\} + t_\delta.$$

The quantities $T(w)$ and $T(u)$ correspond to the number of steps needed to reach a given error tolerance by supervised learning and knowledge distillation, respectively. We want to understand conditions on the original matrix M, such that the difference $T(w) - T(u)$ is positive. Recall that we initialize $n_0 \sim \mathcal{N}(0, k\cdot\lambda_1/n \cdot I_n)$ with some $k \in (0,1)$ and set $w_0 = u_0 = n_0$. For each realization of n_0, define

$$k_a = \frac{r}{n}||n_{2:r,0}||_2^2, \quad k_b = \frac{(n-r)}{n}||n_{r+1:n,0}||_2^2, \quad k_1 = \frac{1}{n}n_{1,0}^2.$$

We previously defined

$$a_t = \frac{||u_{2:r,t}||_2}{|u_{1,t}|}, \quad b_t = \frac{||u_{r+1:n,t}||_2}{|u_{1,t}|}, \quad c_t = (a_t^2 + b_t^2)^{1/2}.$$

At initialization we have $a_0^2 = k_a/k_1 \cdot r$ and $b_0^2 = k_b/k_1 \cdot (n-r)$. Also note that by the concentration of the chi-square random variable (Laurent and Massart, 2000), one can identify fixed intervals such that a_0^2 and b_0^2 fall within these intervals with constant probability. The endpoints of these intervals replace the constants k_a, k_b, and k_1 that appear as the multiplicative factors of ε^{-c} in the matrix dimension lower bound. Importantly, this substitution does not change the exponent c nor does it affect the conclusion that $n = \Omega(\varepsilon^{-c})$. Now define the following quantities

$$\begin{aligned}
d_w &:= 2\log(1/(1-\eta\Delta_n)) \\
d_a &:= 2\log(1/(1-\ell\eta\Delta_2)) \\
d_b &:= 2\log(1/(1-\ell\eta\lambda_1)) \\
d_\xi &= 2\log(1+\eta\xi\lambda_1) \\
d_\delta &:= \log(1/(1-2\eta\kappa\lambda_1)).
\end{aligned}$$

We consider the different cases of the maximum. The first case is when $t_a \geq t_b, t_\xi$, then the difference is,

$$T(w) - T(u) \geq \frac{d_a d_\delta - d_w(d_a + d_\delta)}{d_a d_\delta d_w} \log\left(\frac{1}{\varepsilon}\right) + \frac{\log(n)}{d_w} - \frac{\log(r)}{d_a} + R_a$$

where the remainder R_a does not depend on matrix dimension n,

$$R_a = \frac{\log((k_a+k_b)/k_1)}{d_w} - \frac{\log(k_a/(k_1\theta\cdot\gamma))}{d_a} - \frac{1}{d_\delta}\log\left(\frac{1}{1-\gamma(1+\kappa^{-1})}\right).$$

This leads to the following constraint on the dimension n,

$$n \geq \exp\left(\frac{d_w(d_a+d_\delta) - d_a d_\delta}{d_a d_\delta}\log\left(\frac{1}{\varepsilon}\right) + \frac{d_w}{d_a}\log(r) - d_w R_a\right).$$

The second case is when $t_b \geq t_a, t_\xi$.

$$T(w) - T(u) \geq \frac{d_b d_\delta - d_w(d_b + d_\delta)}{d_b d_\delta d_w} \log\left(\frac{1}{\varepsilon}\right) + \frac{d_b - d_w}{d_b d_w}\log(n) + R_b$$

where the n-independent constant is

$$R_b = \frac{\log((k_a+k_b)/k_1)}{d_w} - \frac{\log(k_b/(k_1(1-\theta)\cdot\gamma))}{d_b} - \frac{1}{d_\delta}\log\left(\frac{1}{1-\gamma(1+\kappa^{-1})}\right).$$

By assumption $d_b - d_w > 0$, thus the dimension n must then be at least,

$$n \geq \exp\left(\frac{d_w(d_b + d_\delta) - d_b d_\delta}{d_\delta(d_b - d_w)} \log\left(\frac{1}{\varepsilon}\right) - R_b \frac{d_b d_w}{d_b - d_w}\right).$$

The final case is $t_\xi \geq t_a, t_b$,

$$T(w) - T(u) \geq \frac{d_\delta - d_w}{d_\delta d_w} \log\left(\frac{1}{\varepsilon}\right) + \frac{d_\xi - d_w}{d_\xi d_w} \log(n) + R_\xi$$

with dimension-independent constant

$$R_\xi = \frac{\log((k_a + k_b)/k_1)}{d_w} - \frac{\log(1/k_1)}{d_\xi} - \frac{1}{d_\delta} \log\left(\frac{1}{1 - \gamma(1 + \kappa^{-1})}\right).$$

As we've assumed $d_\xi - d_w > 0$, it follows that the dimension n must be,

$$n \geq \exp\left(\frac{d_\xi(d_w - d_\delta)}{d_\delta(d_\xi - d_w)} \log\left(\frac{1}{\varepsilon}\right) - R_\xi \frac{d_\xi d_w}{d_\xi - d_w}\right).$$

Now we can conclude by taking,

$$c = \max\left\{\frac{d_w(d_a + d_\delta) - d_a d_\delta}{d_a d_\delta}, \frac{d_w(d_b + d_\delta) - d_b d_\delta}{d_\delta(d_b - d_w)}, \frac{d_\xi(d_w - d_\delta)}{d_\delta(d_\xi - d_w)}\right\},$$

and required matrix dimension is $\Omega(\varepsilon^{-c})$. □

E Additional experiments and details

E.1 Logistic regression on synthetic data

Table 1 contains the initial step size that minimizes parameter error after 10,000 steps of SGD for both knowledge distillation and standard supervision. Figure 4 plots the per-step parameter error for both methods, with the experimental setup previously described in Section 5.1.

Step size η_0, logistic regression on synthetic data.				
	$d = 3$	$d = 5$	$d = 7$	$d = 9$
Standard supervision η_0	$10^{-1.5}$	0.1	0.1	0.1
Distillation η_0	1.0	1.0	1.0	1.0

Table 1: Initial step size η_0 for logistic regression on synthetic data.

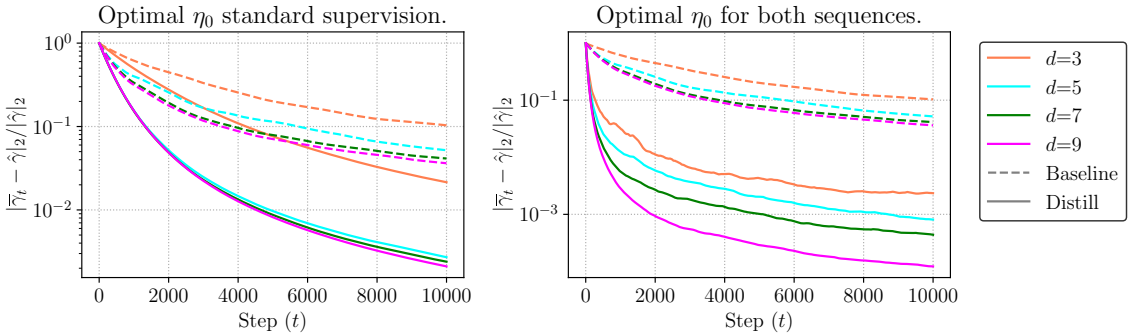

Figure 4: (Logistic regression, synthetic data) Distance from optimal point $\hat{\gamma}$ for stochastic gradient descent iterates under knowledge distillation and standard supervision. All results are averaged over 100 trials. On the left, both trajectories use η_0 that is optimal for standard supervision. On the right, the step η_0 is chosen optimally for both methods.

E.2 Rank-1 matrix approximation

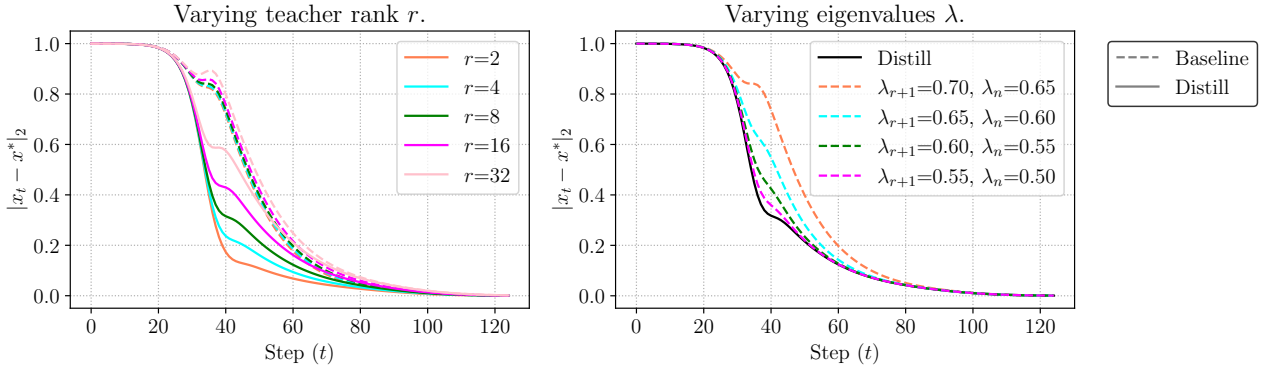

Figure 5: (Rank-1 approximation) Parameter error of gradient descent iterates under knowledge distillation and supervised learning, averaged over 100 trials. In the plot on the left, the eigenvalue spectrum is fixed ($\lambda_r = 0.7$, $\lambda_n = 0.65$) and teacher ranks are varied. In the plot on the right, the teacher rank is fixed ($r = 8$) and the "tail" spectrum λ_r through λ_n is varied.

We construct the target matrix with top eigenvalue $\lambda_1 = 1$, and the arithmetic sequence s_r consists of equally spaced values from $\lambda_2 = 0.8$ to $\lambda_r = 0.75$. Across all teacher ranks r and eigenvalues λ, the step size that achieves the lowest parameter error is $\eta = 10^{-0.5}$. Figure 5 plots the error for gradient descent iterates under standard supervision and knowledge distillation.

E.3 Multinomial logistic regression on Fashion-MNIST

We report the initial step size η_0 that minimizes parameter error after 500,000 steps of SGD for both knowledge distillation and standard supervision (Table 2). Using these optimal step sizes, we plot the per-step parameter error for both methods (Figure 6). The optimal η_0 is the same for both methods when the student dimension is $d \in \{100, 300\}$, however, when $d = 500$, distillation achieves a lower parameter error with a larger initial step size.

Step size η_0, multinomial logistic regression.			
	$d = 100$	$d = 300$	$d = 500$
Standard supervision η_0	$10^{-0.75}$	$10^{-0.5}$	$10^{-0.5}$
Distillation η_0	$10^{-0.75}$	$10^{-0.5}$	$10^{-0.25}$

Table 2: Initial steps size η_0 for multinomial logistic regression on the Fashion-MNIST dataset.

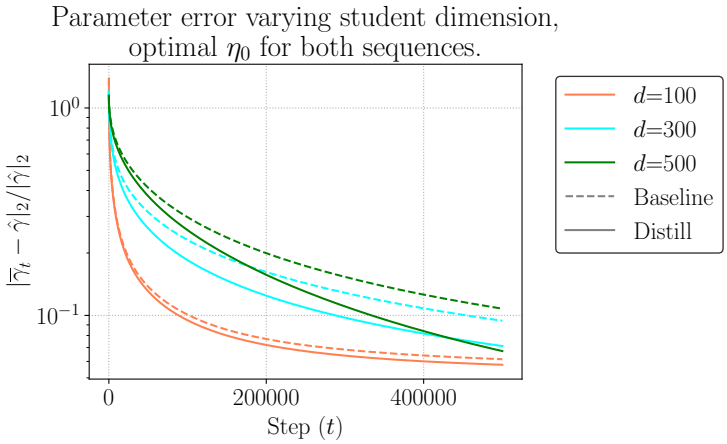

Figure 6: (Multinomial logistic regression, Fashion-MNIST) Normalized parameter error, averaged over 10 trials, of SGD iterates under knowledge distillation and standard supervision for student dimensions $d \in \{100, 300, 500\}$. The initial step size η_0 is chosen optimally for both methods.

E.4 Experiment details: Language model pretraining

We draw 87.3 billion tokens from the C4 dataset as our initial pool. To estimate entropy, we train a GPT-2 medium model on 7.1 billion tokens from the dataset and compute the per-sample entropy as the average token-level entropy from its predicted distributions. We use a learning rate of $\eta = 5 \cdot 10^{-4}$ to train the teacher. Based on these estimates, we partition the remaining data into 13 subsets of increasing entropy, each with a train/validation split of 6.1 billion and 30 million tokens, respectively. For each subset, we perform compute-optimal training of the student model (Hoffmann et al., 2022), specifically a GPT-2 small model trained on 2.48 billion tokens. We sweep learning rates $\eta \in \{10^{-4}, 10^{-3.5}, 10^{-3}, 10^{-2.5}, 10^{-2}\}$ and select the model with the lowest final validation loss as our supervised learning baseline. We find that the optimal learning rate for all subsets is $\eta = 10^{-3}$.

The distillation labels are probabilities P_T from a GPT-2 medium model, trained on the full train split. The distillation student model is trained on the same 2.48 billion tokens, with the same learning rate η and training

setup as the baseline, but has the KL-divergence $D_{\mathrm{KL}}(P_T \,||\, P_S)$ as the objective function where P_S is the student's predicted distribution. All models are trained with a batch size of 512, weight decay of 0.01, gradient clipping of 1.0, learning rate warmup ratio of 0.01, and cosine learning rate schedule. All experiments were conducted on a univeristy-managed computing cluster, using either 4x NVIDIA A100 GPUs (80GB) or 4x NVIDIA RTX 3090 (24GB).