Novelty Assessment Report

Paper: SGD-Based Knowledge Distillation with Bayesian Teachers: Theory and Guidelines
PDF URL: https://openreview.net/pdf?id=YZGBnZbMYN
Venue: ICLR 2026 Conference Submission
Year: 2026
Report Generated: 2026-01-01

Abstract

Knowledge Distillation (KD) is a central paradigm for transferring knowledge from a large teacher network to a typically smaller student model, often by leveraging soft probabilistic outputs. While KD has shown strong empirical success in numerous applications, its theoretical underpinnings remain only partially understood. In this work, we adopt a Bayesian perspective on KD to rigorously analyze the convergence behavior of students trained with Stochastic Gradient Descent (SGD). We study two regimes: (i) when the teacher provides the exact Bayes Class Probabilities (BCPs); and (ii) supervision with noisy approximations of the BCPs. Our analysis shows that learning from BCPs yields variance reduction and removes neighborhood terms in the convergence bounds compared to one-hot supervision. We further characterize how the level of noise affects generalization and accuracy. Motivated by these insights, we advocate the use of Bayesian deep learning models, which typically provide improved estimates of the BCPs, as teachers in KD. Consistent with our analysis, we experimentally demonstrate that students distilled from Bayesian teachers not only achieve higher accuracies (up to +4.27\ %), but also exhibit more stable convergence (up to 30\% less noise), compared to students distilled from deterministic teachers.

Core Task Landscape

This paper addresses: **Convergence Analysis of Knowledge Distillation with Probabilistic Supervision**
A total of **10 papers** were analyzed and organized into a taxonomy with **9 categories**.

Taxonomy Overview

The research landscape has been organized into the following main categories:

- **Theoretical Convergence Analysis and Optimization Dynamics**
- **Divergence Measures and Loss Functions**
- **Student-Teacher Dynamics and Deviations**
- **Training Strategies and Curriculum Learning**
- **Application-Specific Implementations**

Complete Taxonomy Tree

- Convergence Analysis of Knowledge Distillation with Probabilistic Supervision Survey Taxonomy
- Theoretical Convergence Analysis and Optimization Dynamics
  - Bayesian and Probabilistic Teacher Models ★ (1 papers)
  - [0] SGD-Based Knowledge Distillation with Bayesian Teachers: Theory and Guidelines (Anon et al., 2026) View paper
  - Stochastic Optimization Methods (2 papers)
  - [3] Distributed distillation for on-device learning (Ilai Bistritz, 2020) View paper
  - [4] Analysis of an Idealized Stochastic Polyak Method and its Application to Black-Box Model Distillation (Gower, 2025) View paper
  - Variance Reduction and Regularization Effects (2 papers)
  - [5] Knowledge Distillation Performs Partial Variance Reduction (Safaryan, 2023) View paper
  - [8] Stochastic gradient descent with random label noises: doubly stochastic models and inference stabilizer (Haoyi Xiong, 2023) View paper
- Divergence Measures and Loss Functions
  - Kullback-Leibler Divergence Variants (1 papers)
  - [1] Rethinking kullback-leibler divergence in knowledge distillation for large language models (Wu, 2025) View paper
  - Alternative Probability Distillation Methods (1 papers)
  - [6] Probability distillation: A caveat and alternatives (Chin-Wei Huang, 2020) View paper
- Student-Teacher Dynamics and Deviations (1 papers)
  - [7] On student-teacher deviations in distillation: does it pay to disobey? (Nagarajan, 2023) View paper
- Training Strategies and Curriculum Learning (1 papers)
  - [9] Annealing Knowledge Distillation (Aref Jafari, 2021) View paper
- Application-Specific Implementations
  - Lightweight and Resource-Constrained Scenarios (1 papers)
  - [2] A Lightweight and Small Sample Bearing Fault Diagnosis Algorithm Based on Probabilistic Decoupling Knowledge Distillation and Meta-Learning (Hao Luo, 2024) View paper
  - Sequential and Temporal Data Applications (1 papers)
  - [10] Three-way decision-based experience replay mechanism for online time series forecasting (Jing Wang, n.d.) View paper

Narrative

Core task: convergence analysis of knowledge distillation with probabilistic supervision. The field structure reflects a multi-faceted investigation into how student models learn from teacher distributions rather than hard labels. The taxonomy organizes work into theoretical convergence analysis and optimization dynamics, divergence measures and loss functions, student-teacher dynamics and

deviations, training strategies and curriculum learning, and application-specific implementations. Theoretical branches examine the mathematical foundations of distillation convergence, including Bayesian and probabilistic teacher models that provide uncertainty-aware supervision. Divergence measures explore how different loss functions—such as KL divergence variants studied in Rethinking KL Divergence[1]—affect learning dynamics. Student-teacher dynamics investigate mismatches and deviations, as seen in Student Teacher Deviations[7], while training strategies address curriculum design and annealing approaches like Annealing Distillation[9]. Application branches demonstrate these principles in domains such as fault diagnosis, exemplified by Lightweight Bearing Fault[2].

Particularly active lines of work contrast deterministic versus probabilistic teacher supervision, examining trade-offs between convergence guarantees and practical performance. Distributed settings, explored in Distributed Distillation[3], raise questions about how decentralized training affects convergence properties. Variance reduction techniques, such as those in Partial Variance Reduction[5] and Stochastic Polyak Distillation[4], address optimization stability. The original paper, SGD Bayesian Distillation[0], sits within the theoretical convergence branch focusing on Bayesian teacher models. It shares thematic ground with works analyzing probabilistic supervision but emphasizes rigorous SGD convergence analysis under Bayesian uncertainty, contrasting with more heuristic approaches in Probability Distillation Caveat[6] or noise-focused studies like Random Label Noises[8]. This positioning highlights a growing interest in formal guarantees for distillation with uncertain teachers.

Related Works in Same Category

No sibling papers were found in the same taxonomy leaf. A taxonomy-subtopic-level comparison will be produced instead.

Taxonomy-Level Summary

The original leaf focuses on theoretical convergence guarantees when knowledge distillation uses Bayesian or probabilistic teacher models that provide exact class probabilities. The sibling subtopics address complementary aspects of convergence analysis: one examines the optimization algorithms (stochastic gradient descent variants) used to train distilled models, while the other investigates the implicit regularization and variance reduction properties that emerge during distillation training.

Similarities: - All three subtopics contribute to theoretical understanding of knowledge distillation convergence - Each provides formal analysis with mathematical guarantees or characterizations - All are concerned with training dynamics and optimization behavior rather than architectural choices - Each subtopic excludes divergence-based or loss function analysis, suggesting these belong to a separate category

Differences: - Bayesian and Probabilistic Teacher Models focuses on the teacher's probabilistic nature and how exact Bayes probabilities affect convergence, while siblings focus on optimization mechanics - Stochastic Optimization Methods analyzes algorithmic variants (SGD, momentum, etc.) and their convergence rates, not the probabilistic structure of supervision - Variance Reduction and Regularization Effects examines emergent properties (implicit regularization, noise effects) rather than explicit teacher model characteristics or optimization algorithms - The original leaf is teacher-centric (analyzing supervision quality), while Stochastic Optimization is algorithm-centric and Variance Reduction is property-centric

Suggested Search Directions: - Convergence analysis combining Bayesian teachers with specific SGD variants - How probabilistic supervision affects variance reduction properties during distillation - Theoretical connections between Bayesian teacher uncertainty and implicit regularization in student models

Sibling Subtopics

- **Stochastic Optimization Methods** (leaves: 1, papers: 2)
- Scope: Analysis of stochastic gradient descent variants and optimization algorithms for knowledge distillation with convergence rate characterization.
- Exclude: Bayesian-specific methods belong to Bayesian and Probabilistic Teacher Models; variance reduction mechanisms belong to Variance Reduction and Regularization Effects.
- **Variance Reduction and Regularization Effects** (leaves: 1, papers: 2)
- Scope: Studies examining implicit regularization, variance reduction properties, and noise effects in knowledge distillation optimization.
- Exclude: Explicit optimization algorithms belong to Stochastic Optimization Methods; student-teacher deviation analysis belongs to Student-Teacher Dynamics and Deviations.

Contributions Analysis

This paper presents **3 main contributions**, each analyzed against relevant prior work:

Contribution 1: Convergence analysis of SGD-based knowledge distillation with Bayesian class probability supervision

Description: The authors provide a theoretical analysis of how students trained via SGD converge when supervised with exact Bayes Class Probabilities versus noisy approximations. They show that learning from BCPs yields variance reduction and removes neighborhood terms in convergence bounds compared to one-hot supervision.

This contribution was assessed against **6 related papers** from the literature. Papers with potential prior art are analyzed in detail with textual evidence; others receive brief assessments.

---

1. Towards understanding knowledge distillation

URL: View paper

Brief Assessment

Understanding Distillation[21] analyzes linear distillation with gradient flow and focuses on data geometry and optimization bias. It does not address SGD convergence with probabilistic supervision or variance reduction in the sense studied by the original paper.

---

2. Probabilistic Self-supervised Learning via Scoring Rules Minimization

URL: View paper

Brief Assessment

Probabilistic Self-supervised Learning[24] focuses on self-supervised learning with scoring rules for representation learning, not on convergence analysis of SGD in knowledge distillation with probabilistic supervision from teacher models.

---

3. Evidential Knowledge Distillation

URL: View paper

Brief Assessment

Evidential Distillation[25] focuses on second-order Dirichlet distributions for knowledge distillation in classification tasks, not on SGD convergence analysis with Bayesian class probabilities. The candidate provides PAC-Bayesian bounds on expected risk rather than analyzing SGD dynamics under probabilistic supervision.

---

4. Stochastic gradient descent with random label noises: doubly stochastic models and inference stabilizer

URL: View paper

Brief Assessment

Random Label Noises[8] analyzes SGD dynamics with unbiased label noise as implicit regularization, focusing on inference stability. It does not address knowledge distillation, Bayesian class probabilities, or teacher-student frameworks that are central to the original paper's contribution.

---

5. Agree to disagree: Adaptive ensemble knowledge distillation in gradient space

URL: View paper

Brief Assessment

Adaptive Ensemble Distillation[22] focuses on multi-teacher ensemble distillation using gradient-space optimization and multi-objective optimization techniques, not on convergence analysis of SGD with probabilistic supervision or Bayesian class probabilities.

---

6. Classification accuracy improvement of the optical diffractive deep neural network by employing a knowledge distillation and stochastic gradient descent Î²-Lasso joint â¦

URL: View paper

Brief Assessment

Optical Diffractive Accuracy[23] focuses on optical diffractive deep neural networks with knowledge distillation for image classification, not on theoretical convergence analysis of SGD with probabilistic supervision or Bayesian class probabilities.

Contribution 2: Advocacy for Bayesian deep learning models as teachers in knowledge distillation

Description: Motivated by their theoretical findings that teacher calibration affects student performance, the authors propose using Bayesian neural networks as teachers because they provide better-calibrated probability estimates that more faithfully approximate the true BCPs.

This contribution was assessed against **10 related papers** from the literature. Papers with potential prior art are analyzed in detail with textual evidence; others receive brief assessments.

---

1. Bayesian knowledge distillation for online action detection

URL: View paper

Brief Assessment

Bayesian Action Detection[13] focuses on online action detection in video streams using Bayesian teachers for temporal feature selection and uncertainty quantification, not on general knowledge distillation frameworks or the theoretical analysis of teacher calibration effects on student performance in classification tasks.

---

2. Bayeskd: Bayesian knowledge distillation for compact llms in constrained fine-tuning scenarios

URL: View paper

Brief Assessment

BayesKD Compact LLMs[19] focuses on using Bayesian optimization for hyperparameter search in knowledge distillation for compressed LLMs, not on using Bayesian neural networks as teacher models to provide better-calibrated probability estimates as advocated in the original paper.

---

3. A statistical perspective on distillation

URL: View paper

Prior Art Analysis

Statistical Perspective Distillation[18] demonstrates that the original paper's advocacy for Bayesian neural networks as teachers is not novel. The candidate paper establishes the same core principle: that teachers providing better approximations of the Bayes class probabilities (BCPs) improve student performance. Specifically, Statistical Perspective Distillation[18] proves that a 'Bayes teacher' providing true class-probabilities can lower the variance of the student objective (Lemma 1), and establishes a bias-variance tradeoff showing that teachers with better probability estimates yield better students (Proposition 3). The candidate paper explicitly states that 'good probability modelling can aid student generalisation' and that 'models producing good probabilities will also be accurate, models that are accurate do not necessarily offer good probabilities.' This directly anticipates the original paper's motivation for using Bayesian models, which are known for better-calibrated probability estimates.

Evidence

Evidence 1 - **Rationale**: Both papers establish that teachers with better probability estimates (approximating BCPs) improve student performance. Statistical Perspective Distillation[18] provides the theoretical foundation (bias-variance tradeoff) that the original paper uses to justify Bayesian teachers. - **Original**: based on our analysis, which indicates that the effectiveness ofkd depends on how well the teacher approximates the bcp (i.e., how well-calibrated the teacher is), we advocate the use of bayesian deep learning models as teachers gawlikowski et al. (2023) inkd. bayesian deep learning brings forth a k... - **Candidate**: our core observation is that a 'bayes teacher' providing the true classprobabilities canlower the varianceof the student objective, and thus improve performance. we then establish abias-variance tradeoff that quantifies the utility of teachers that approximate the bayes class-probabilities. this prov...

---

4. Modeling rapid language learning by distilling Bayesian priors into artificial neural networks

URL: View paper

Brief Assessment

Rapid Language Learning[11] focuses on distilling Bayesian priors into neural networks for language learning tasks, not on using Bayesian models as teachers in knowledge distillation frameworks. The paper addresses a different problem domain (rapid language acquisition) with a different methodology (meta-learning from Bayesian-sampled tasks).

---

5. Bayes conditional distribution estimation for knowledge distillation based on conditional mutual information

URL: View paper

Brief Assessment

Conditional Mutual Information[17] focuses on training teachers to maximize conditional mutual information to capture contextual information, not on using Bayesian neural networks as teachers for better calibration.

---

6. Uncertainty-based knowledge distillation for Bayesian deep neural network compression

URL: View paper

Brief Assessment

The candidate paper (Uncertainty-based Compression[14]) is not available for comparison. Without access to the full text, it is impossible to assess whether it demonstrates prior work on using Bayesian neural networks as teachers in knowledge distillation.

---

7. Knowledge Distillation and Its Application to Network Traffic Classification

URL: View paper

Brief Assessment

Traffic Classification Distillation[20] focuses on network traffic classification applications and does not discuss Bayesian neural networks as teachers. The candidate's contributions center on flow-packet hybrid classification, knowledge explaining distillation with superfeatures, and federated learning aggregation—none of which address Bayesian teacher models or teacher calibration.

---

8. Efficient Uncertainty Estimation via Distillation of Bayesian Large Language Models

URL: View paper

Brief Assessment

Efficient Uncertainty Distillation[12] focuses on distilling Bayesian LLMs into deterministic student models for efficient uncertainty estimation, not on advocating Bayesian models as teachers to improve student calibration in general knowledge distillation frameworks.

---

9. Bayesian evidential deep learning for online action detection

URL: View paper

Brief Assessment

Evidential Action Detection[16] uses Bayesian neural networks as teachers in a teacher-student framework, but focuses on online action detection with evidential deep learning for uncertainty quantification, not on the theoretical analysis of how teacher calibration affects student performance in general knowledge distillation settings.

---

10. Bayesian knowledge distillation: A bayesian perspective of distillation with uncertainty quantification

URL: View paper

Brief Assessment

Bayesian Perspective Distillation[15] focuses on establishing a Bayesian framework for knowledge distillation by formulating a teacher-informed prior for student parameters, rather than advocating for Bayesian neural networks as teachers to improve calibration. The original paper's contribution centers on using BNNs as teachers specifically because they provide better-calibrated probability estimates, while the candidate uses Bayesian modeling to interpret the distillation mechanism itself.

Contribution 3: Characterization of interpolation property and gradient noise under BCP supervision

Description: The authors prove that when students are supervised with true BCPs, the optimization task satisfies the interpolation property, meaning the minimizer matches true BCPs at each sample. They also characterize how gradient noise depends on the quality of BCP estimates.

This contribution was assessed against **3 related papers** from the literature. Papers with potential prior art are analyzed in detail with textual evidence; others receive brief assessments.

---

1. Local Linear Neuro-Fuzzy Models: Advanced Aspects

URL: View paper

Brief Assessment

Local Linear Neuro-Fuzzy[28] discusses interpolation characteristics in the context of neuro-fuzzy models and noise amplification, but does not address the specific theoretical characterization of interpolation property and gradient noise under Bayesian class probability supervision in knowledge distillation settings.

---

2. Randomness and Interpolation Improve Gradient Descent

URL: View paper

Brief Assessment

Randomness Interpolation Gradient[26] focuses on accelerating SGD convergence through Newton interpolation and noise regularization techniques for CNNs. It does not address Bayes class probabilities, probabilistic supervision, or the theoretical characterization of interpolation properties in knowledge distillation contexts.

---

3. SEGA: Shaping Semantic Geometry for Robust Hashing under Noisy Supervision

URL: View paper

Brief Assessment

SEGA Semantic Hashing[27] focuses on robust hashing under noisy supervision in retrieval tasks, not on analyzing SGD convergence properties under Bayesian class probability supervision or characterizing interpolation properties in knowledge distillation contexts.

Appendix: Text Similarity Detection

Textual similarity detection checked 19 papers and found 2 similarity segment(s) across 2 paper(s).

The following **2 paper(s)** were detected to have high textual similarity with the original paper. These may represent different versions of the same work, duplicate submissions, or papers with substantial textual overlap. Readers are advised to verify these relationships independently.

1. Towards understanding knowledge distillation

Detected in: Contribution: contribution_1

⚠ **Note**: This paper shows substantial textual similarity with the original paper. It may be a different version, a duplicate submission, or contain significant overlapping content. Please review carefully to determine the nature of the relationship.

---

2. A statistical perspective on distillation

Detected in: Contribution: contribution_2

⚠ **Note**: This paper shows substantial textual similarity with the original paper. It may be a different version, a duplicate submission, or contain significant overlapping content. Please review carefully to determine the nature of the relationship.

References

- [0] SGD-Based Knowledge Distillation with Bayesian Teachers: Theory and Guidelines View paper

- [1] Rethinking kullback-leibler divergence in knowledge distillation for large language models View paper
- [2] A Lightweight and Small Sample Bearing Fault Diagnosis Algorithm Based on Probabilistic Decoupling Knowledge Distillation and Meta-Learning View paper
- [3] Distributed distillation for on-device learning View paper
- [4] Analysis of an Idealized Stochastic Polyak Method and its Application to Black-Box Model Distillation View paper
- [5] Knowledge Distillation Performs Partial Variance Reduction View paper
- [6] Probability distillation: A caveat and alternatives View paper
- [7] On student-teacher deviations in distillation: does it pay to disobey? View paper
- [8] Stochastic gradient descent with random label noises: doubly stochastic models and inference stabilizer View paper
- [9] Annealing Knowledge Distillation View paper
- [10] Three-way decision-based experience replay mechanism for online time series forecasting View paper
- [11] Modeling rapid language learning by distilling Bayesian priors into artificial neural networks View paper
- [12] Efficient Uncertainty Estimation via Distillation of Bayesian Large Language Models View paper
- [13] Bayesian knowledge distillation for online action detection View paper
- [14] Uncertainty-based knowledge distillation for Bayesian deep neural network compression View paper
- [15] Bayesian knowledge distillation: A bayesian perspective of distillation with uncertainty quantification View paper
- [16] Bayesian evidential deep learning for online action detection View paper
- [17] Bayes conditional distribution estimation for knowledge distillation based on conditional mutual information View paper
- [18] A statistical perspective on distillation View paper
- [19] Bayeskd: Bayesian knowledge distillation for compact llms in constrained fine-tuning scenarios View paper
- [20] Knowledge Distillation and Its Application to Network Traffic Classification View paper
- [21] Towards understanding knowledge distillation View paper
- [22] Agree to disagree: Adaptive ensemble knowledge distillation in gradient space View paper
- [23] Classification accuracy improvement of the optical diffractive deep neural network by employing a knowledge distillation and stochastic gradient descent Î²-Lasso joint â¦ View paper
- [24] Probabilistic Self-supervised Learning via Scoring Rules Minimization View paper
- [25] Evidential Knowledge Distillation View paper
- [26] Randomness and Interpolation Improve Gradient Descent View paper
- [27] SEGA: Shaping Semantic Geometry for Robust Hashing under Noisy Supervision View paper
- [28] Local Linear Neuro-Fuzzy Models: Advanced Aspects View paper