Chapter 3

Theoretical Limits of Machine Learning under Noisy and Limited Data

Gökhan Halimoğlu[1]

Abstract

Machine learning has achieved remarkable success across many scientific and engineering domains. However, the increasing reliance on data driven methods has also raised important questions regarding the theoretical limits of learning systems operating under noisy and limited data conditions. While modern algorithms often demonstrate impressive predictive performance, their reliability depends strongly on the statistical assumptions and structural properties of the data used during training.

This chapter examines the fundamental limitations of machine learning when observations are affected by measurement noise, limited sample size, and violations of classical statistical assumptions. Particular attention is given to the physical and mathematical structure of noise, including electronic fluctuations, quantization effects, drift phenomena, and heavy tailed variability. These characteristics challenge simplified modeling assumptions frequently adopted in machine learning, such as independent and identically distributed observations or Gaussian noise structures.

The discussion further explores how noise and data scarcity influence generalization behavior. Highly flexible models may achieve low training error while memorizing noise rather than learning meaningful signal patterns. Such phenomena highlight the importance of understanding the interaction between model complexity, data availability, and uncertainty in the data generating process. In addition, the chapter analyzes potential risks associated with simulation based data augmentation and the generation of synthetic datasets. When the statistical structure of simulated data does not accurately reflect the underlying system, models may learn artifacts of the simulation process instead of genuine physical relationships.

1 Dr. Gökhan Halimoğlu, Istanbul Kultur University, Vocational School, Department of Air Conditioning and Refrigeration Technology, g.halimoglu@iku.edu.tr, 0000-0002-0419-7723.

https://doi.org/10.58830/ozgur.pub1236.c5004

Finally, the role of physical priors and domain knowledge in stabilizing learning systems is discussed. Incorporating structural constraints derived from known physical principles can reduce the effective hypothesis space and improve robustness in noisy environments. From this perspective, machine learning should be viewed not as a universal replacement for theoretical modeling, but as a complementary tool within a broader scientific methodology.

2.1. Why Do Limits Matter?

Machine learning has advanced rapidly over the past decade and has produced impressive results across many application domains. However, a large portion of the literature primarily focuses on model performance, while giving less systematic attention to limitations, assumptions, and failure conditions. This tendency creates an implicit success bias: published studies often emphasize high accuracy scores and improved metrics, whereas the conditions under which a model loses reliability receive comparatively limited discussion. In contrast, real world problems differ substantially from controlled benchmark datasets and idealized experimental settings.

In real systems, data are often limited. This limitation is not only related to sample size. Physical and operational constraints such as measurement duration, hardware capacity, energy budget, cost, and accessibility also shape the data generation process. In fields such as particle detection, biomedical sensing, industrial monitoring, or financial time series analysis, data are not abstract mathematical objects; they are physical outputs produced under specific conditions. Therefore, the assumption that "more data" is always available or feasible is often unrealistic in practice.

Similarly, in many machine learning models, noise is treated as an unwanted error term. Yet in numerous physical systems, noise is an inherent component of the measurement process. Thermal fluctuations, microscopic electronic behavior, quantization effects, environmental variations, and long term drift all contribute to the observed signal. In this sense, noise is not merely a disturbance to be removed, but part of the system's natural dynamics. Models that ignore the structural characteristics of noise may achieve strong statistical performance while producing physically misleading interpretations.

This raises a fundamental question: when does machine learning fail, and why? Increasing model complexity does not automatically guarantee better generalization. When data are limited, when distributional assumptions are violated, or when noise carries structural properties, a model may learn

variations instead of the underlying signal. At this point, the distinction between training performance and physical validity becomes critical.

The aim of this chapter is to examine the limits of machine learning within a systematic framework. By analyzing the mathematical and physical nature of noise, the statistical consequences of limited data, the violation of distributional assumptions, and the risks of simulation based approaches, we seek to highlight the need for more cautious, physically informed, and methodologically robust learning strategies.

2.1.1. Structural Origins of Success Bias in Machine Learning Research

As emphasized in the introductory discussion, there exists a noticeable imbalance between the visibility of performance gains and the visibility of model limitations in machine learning research. This imbalance does not arise solely from individual research choices; it is also shaped by the structural characteristics of scientific production and evaluation. Within academic publishing, novelty, measurable improvement, and statistical performance gains often serve as primary evaluation criteria. As a result, demonstrating superior metrics can become more central than systematically analyzing the conditions under which a model becomes unstable or unreliable.

Model comparisons are frequently conducted on widely accepted benchmark datasets. Over time, these datasets evolve into reference standards, and models are ranked within this fixed experimental framework. While such comparisons are useful, they assess model behavior under a specific distributional setting. The response of the same model to distributional shifts or alternative data generation mechanisms is often left unexplored. Consequently, performance may be interpreted as an intrinsic property of the model rather than a context dependent outcome.

A similar issue arises in hyperparameter optimization and architectural refinement. Small numerical improvements are commonly reported as evidence of methodological progress. However, it is not always clear whether these gains reflect genuine structural advancement or dataset specific variation. When data are limited, even modest differences in evaluation metrics may fall within the range of statistical uncertainty. Without careful analysis, such differences can be overstated.

Success bias also influences the visibility of negative results. Instances of failed generalization, unstable behavior, or sensitivity to minor perturbations are less frequently reported. This creates a gap between the controlled scenarios represented in the literature and the variability encountered in real

systems. Models that perform well under optimized conditions may behave unpredictably when confronted with constraints that were not present during development.

This gap becomes particularly significant in domains where data generation is governed by physical and operational constraints. Performance evaluation is often carried out under the implicit assumption that sufficient and stable data are available. In many real world systems, however, data are inherently constrained in both quantity and structure. Recognizing the structural roots of success bias therefore leads naturally to a more fundamental question: why is data in real systems almost always limited, and how do these limitations shape the learning process?

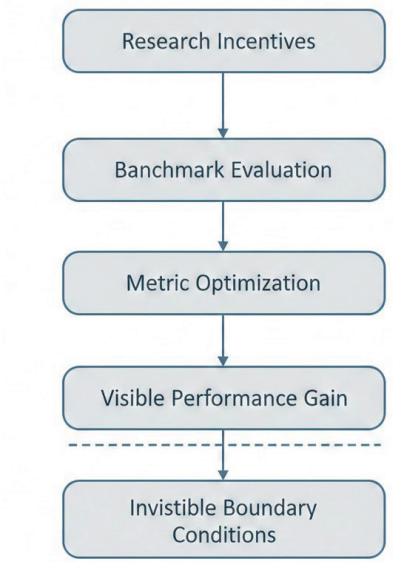

Figure 2.1. Structural Drivers of Success Bias in Machine Learning Research.

This conceptual diagram illustrates how structural elements within academic research and evaluation processes can collectively contribute to success bias. Rather than emerging solely from individual reporting choices, performance visibility is shaped by benchmark centered evaluation, metric optimization practices, and selective reporting mechanisms. The dashed boundary emphasizes the often overlooked conditions under which model reliability deteriorates, highlighting the gap between visible performance gains and unexamined boundary constraints.

2.1.2. Noise as a Physical Reality Rather Than an Undesired Error

In many machine learning formulations, the error term is treated as an abstract and idealized random component. The observed output is typically expressed as:

$$y = f(x) + \varepsilon$$

In this representation, ε is often assumed to be zero mean, independent, and drawn from a specific distribution, most commonly Gaussian. This assumption simplifies mathematical analysis and provides a convenient theoretical foundation for learning algorithms. However, in real systems, noise rarely follows such an idealized structure.

If one has worked with physical measurement systems, it becomes evident that noise is an intrinsic part of the measurement chain. Thermal fluctuations in electronic circuits, quantization limits in sensors, micro variations in environmental conditions, and long term drift effects are naturally embedded in the observed signal. Yet this phenomenon is not limited to physics based systems. In economic time series, market microstructure effects introduce inherent volatility. In financial data, liquidity driven jumps are part of the system's behavior. In biomedical measurements, inter individual physiological variability shapes the recorded signal. In social sciences, sampling uncertainty and contextual variability influence observed outcomes. These variations are often labeled as "error," but they are, in fact, structural features of the data generating process.

Noise, therefore, should not be interpreted as an external disturbance simply added to an otherwise clean system. It is frequently a direct consequence of the internal dynamics of the process under investigation. This distinction is not merely conceptual; it has methodological implications. When noise depends on measurement conditions, time, or system parameters, it cannot be adequately modeled as a simple independent random variable. In such cases, the error term may be more appropriately represented as:

$$\varepsilon = g(x, t, \theta)$$

indicating that noise may interact with observed variables, temporal structure, or intrinsic system parameters. Under these circumstances, the classical assumption of independent error becomes insufficient.

Recognizing noise as a physical or structural component of the system alters how learning outcomes should be interpreted. Methods that rely on simplified distributional assumptions may generate a misleading sense of robustness in environments where noise carries structure. A model may achieve low training error, yet still fail to capture the underlying system behavior in a physically meaningful way. This issue becomes particularly pronounced in systems characterized by temporal dependence, cross channel correlations, or resolution limits, where the boundary between signal and noise is not sharply defined.

Moreover, noise is not always devoid of information. Variations observed in measurement systems can encode indirect knowledge about system stability, environmental conditions, or operational regimes. Aggressive preprocessing steps designed to suppress noise may therefore remove patterns that are informative for modeling. The objective should not be the elimination of noise, but the accurate characterization of its structure.

Viewing noise as a structural property of the data generating mechanism, rather than as a purely unwanted perturbation, provides a more realistic foundation for modeling. Otherwise, learning algorithms risk optimizing idealized assumptions instead of representing the actual system. This perspective naturally leads to a deeper examination of the mathematical properties of noise and the limitations of common distributional assumptions, which will be discussed in the following section.

2.1.3. Structural Limits and Assumption Breakdown

The discussions presented in this section indicate that failures in machine learning systems are rarely accidental. Learning algorithms produce meaningful and reliable outcomes only under specific statistical and structural assumptions. When these assumptions weaken or are violated, performance metrics may cease to reflect the actual behavior of the model. In particular, when independence, distributional stability, or bounded variation assumptions break down, generalization capacity may deteriorate even if model complexity increases.

This effect becomes more pronounced in limited data regimes. Instead of capturing the underlying structure, the model may internalize incidental variation, widening the gap between apparent training success and physical interpretability. In such cases, failure does not arise from a purely technical

deficiency but from a misalignment between the model and the data generating process. The learning procedure remains confined within abstract statistical assumptions rather than representing the true dynamics of the system.

For this reason, understanding the limits of machine learning cannot be reduced to a discussion of algorithmic capacity alone. The decisive factor lies in the nature of variability observed in the data. Noise is frequently modeled as small and symmetric perturbations; however, in many physical and operational systems it exhibits structural properties, temporal dependence, or heavy tailed behavior. When these characteristics are ignored, the discrepancy between apparent predictive success and genuine system behavior may widen.

At this point, the discussion naturally leads to a more fundamental question: What is the mathematical and physical nature of noise? A substantial portion of the limitations observed in learning systems originates from the structural character of uncertainty within the data. The next chapter therefore examines noise in greater depth, addressing electronic noise mechanisms, quantization effects, drift and systematic bias, the implicit assumption of noise free modeling, the limitations of Gaussian approximations, the role of heavy tailed distributions, and the distinction between outliers and intrinsic noise. This technical analysis provides a deeper explanation for the failure mechanisms outlined above.

2.2. Mathematical and Physical Nature of Noise

The limitations discussed in the previous section ultimately converge on a common source: the structural character of uncertainty embedded in data. If machine learning systems fail under noisy and limited conditions, it becomes essential to examine the mathematical and physical foundations of that noise itself. Noise is not merely a statistical residual; it reflects measurable processes, resolution constraints, electronic fluctuations, discretization effects, and systematic biases inherent to real systems.

Importantly, such uncertainty does not arise from a single homogeneous mechanism. Different physical and operational processes generate distinct forms of variability, each carrying its own statistical structure and theoretical implications. Understanding the theoretical limits of learning therefore requires moving beyond abstract error terms and toward a structured analysis of these heterogeneous noise sources as physical and mathematical phenomena.

2.2.1.Electronic Noise

Electronic noise represents one of the most fundamental sources of uncertainty in physical measurement systems. It originates from microscopic

processes inherent to electronic components and cannot be eliminated entirely, only characterized and controlled.

Thermal noise, for instance, arises from the random motion of charge carriers within resistive elements. Its power spectral density is proportional to temperature and bandwidth, indicating that uncertainty is intrinsically linked to physical conditions rather than modeling imperfections. Similarly, shot noise emerges from the discrete nature of electric charge and becomes particularly relevant in low current or high sensitivity systems.

In many measurement environments, the observed signal can be expressed as:

$$y(t) = s(t) + \varepsilon_{el}(t)$$

Where $\varepsilon_{el}(t)$ reflects electronic fluctuations. While often approximated as Gaussian for analytical convenience, the distributional form depends on operating conditions, bandwidth constraints, and device architecture. Thus, electronic noise is not an abstract modeling artifact but a physically grounded stochastic process.

For machine learning systems trained on such measurements, ignoring the physical structure of electronic noise may lead to overconfidence in apparent signal patterns that partially originate from instrumentation effects.

2.2.2. Quantization Noise

Quantization noise arises from the discretization of continuous signals during analog to digital conversion. Every digital system imposes finite resolution, meaning that measured values are mapped onto discrete levels.

If x denotes a continuous signal and $Q(x)$ its quantized representation, the quantization error can be written as:

$$\varepsilon(q) = Q(x) - \chi$$

Under high resolution assumptions, this error is often modeled as uniformly distributed within a bounded interval. However, this approximation relies on specific regularity conditions. In low resolution systems or in signals with structured patterns, quantization effects may correlate with the signal itself, violating independence assumptions.

From a learning perspective, quantization introduces systematic discretization artifacts. Models trained on discretized data may inadvertently learn resolution induced patterns rather than intrinsic system behavior.

2.2.3. Drift and Systematic Bias

Unlike random fluctuations, drift reflects slow temporal variation in system parameters. It may originate from temperature changes, hardware aging, calibration shifts, or environmental transitions.

A simple representation may be written as:

$$y(t) = s(t) + \varepsilon(t) + d(t)$$

where $d(t)$ represents a slowly varying drift component.

Systematic bias differs from stochastic noise in that its expectation is non zero:

$$\text{Å}[\varepsilon] \neq 0$$

In such cases, the assumption of unbiased error collapses. For learning systems, drift and bias can produce misleading trends, causing models to interpret structural shifts as meaningful predictive signals.

2.2.4. The Implicit "Noise Free" Assumption in Modeling

Many learning formulations implicitly assume that noise is small, independent, and removable. Even when not stated explicitly, optimization objectives often treat the residual term as symmetric and statistically well behaved.

This "noise free" assumption becomes embedded in loss functions, regularization strategies, and validation procedures. However, when noise carries structure temporal dependence, cross channel correlation, or heavy tailed deviations such simplifications distort model evaluation.

The theoretical limits of learning are therefore partly determined by how closely the assumed noise model matches the physical reality of the data generating mechanism.

2.2.5. Why the Gaussian Assumption Often Fails

Gaussian noise assumptions are mathematically convenient due to their stability under linear transformations and their central role in classical statistical theory. Many generalization results rely implicitly on finite variance and concentration inequalities that hold under sub Gaussian conditions.

Yet empirical data in numerous domains exhibit skewness, kurtosis excess, or tail behavior inconsistent with Gaussian decay. In systems affected by rare but significant perturbations, the probability of extreme deviations may decay polynomially rather than exponentially.

When variance is unstable or dominated by rare events, classical concentration results weaken. Consequently, empirical risk minimization may not approximate expected risk as reliably as theory suggests.

2.2.6. Heavy Tailed Distributions

Heavy tailed distributions describe situations in which extreme deviations occur more frequently than predicted by Gaussian models. In contrast to exponential decay in normal distributions, heavy tailed behavior often follows polynomial decay of the form as:

$$P\left(|x| > x\right) \sim x^{-\alpha}, \alpha > 0$$

When the tail index $\alpha \leq 2$, the variance may be infinite; when $\alpha \leq 1$, even the mean may be undefined. Under such conditions, classical statistical assumptions finite variance, stable concentration around the mean, rapid convergence of sample averages no longer hold in their standard form.

In learning systems, heavy tailed variability can significantly alter optimization dynamics and generalization behavior. Rare but high magnitude deviations may dominate gradients, distort empirical risk estimates, or produce unstable parameter updates. Consequently, theoretical guarantees derived under sub Gaussian or bounded noise assumptions may substantially underestimate real world variability.

Heavy tails therefore do not merely represent statistical anomalies; they redefine the scale at which uncertainty operates.

2.2.7. Outliers Are Not Necessarily Noise

Extreme observations are often treated as undesirable perturbations and removed through filtering or trimming procedures. However, not every outlier is noise. An outlier may reflect:

- A rare but valid operating regime
- A transition between system states
- A structural anomaly
- The emergence of a new distributional phase

Automatically classifying extreme values as measurement error risks discarding meaningful information. In systems characterized by regime shifts, heavy tailed dynamics, or structural instability, extreme points may contain precisely the signals that reveal underlying constraints.

The distinction between intrinsic noise and structurally meaningful deviation is therefore not purely statistical; it is contextual and system dependent. Learning algorithms that indiscriminately suppress outliers may improve short term metrics while obscuring deeper structural behavior.stability, extreme points may contain precisely the signals that reveal underlying constraints.

2.2.8. Structural Implications for Learning Limits

The phenomena discussed throughout this section electronic fluctuations, quantization effects, drift, systematic bias, heavy tailed variability, and the ambiguous status of outliers collectively demonstrate that uncertainty in real systems is heterogeneous and structured. Noise does not arise from a single, uniform mechanism, nor does it consistently conform to classical assumptions of independence, Gaussian decay, or bounded variance.

When machine learning models rely on simplified noise formulations, theoretical guarantees regarding stability and generalization may rest on fragile premises. Mischaracterizing the structure of uncertainty does not merely affect parameter estimation; it alters the interpretation of performance metrics and the credibility of generalization bounds.

Within the broader theme of *Theoretical Limits of Machine Learning under Noisy and Limited Data*, these observations clarify a central point: the limits of learning are inseparable from the structure of noise. To model uncertainty inaccurately is to misinterpret the theoretical boundaries of machine learning itself.

2.3. Violation of the IID Assumption

Many theoretical guarantees in machine learning rely on a fundamental statistical assumption: that the observed data are independent and identically distributed (IID). Under this assumption, each observation is treated as a random sample drawn independently from the same underlying probability distribution. In other words, every data point is assumed to carry the same statistical characteristics and to be unaffected by the presence or order of other observations.

The IID framework provides a convenient foundation for theoretical analysis. Many classical results in statistical learning theory including convergence guarantees, risk bounds, and generalization theorems are derived under the assumption that training samples represent unbiased realizations of a stable data generating process. When this condition holds, empirical observations can be interpreted as reliable approximations of the underlying distribution, allowing models to infer patterns that extend beyond the observed dataset.

However, the IID assumption is fundamentally an idealization. In many practical contexts, the process that generates the data is influenced by physical mechanisms, environmental factors, or operational constraints. These influences introduce dependencies and structural variations that challenge the assumption that observations are independent realizations from a single stationary distribution.

As a result, the IID framework should not be interpreted as a universal description of real world data, but rather as a mathematical simplification that facilitates theoretical development.

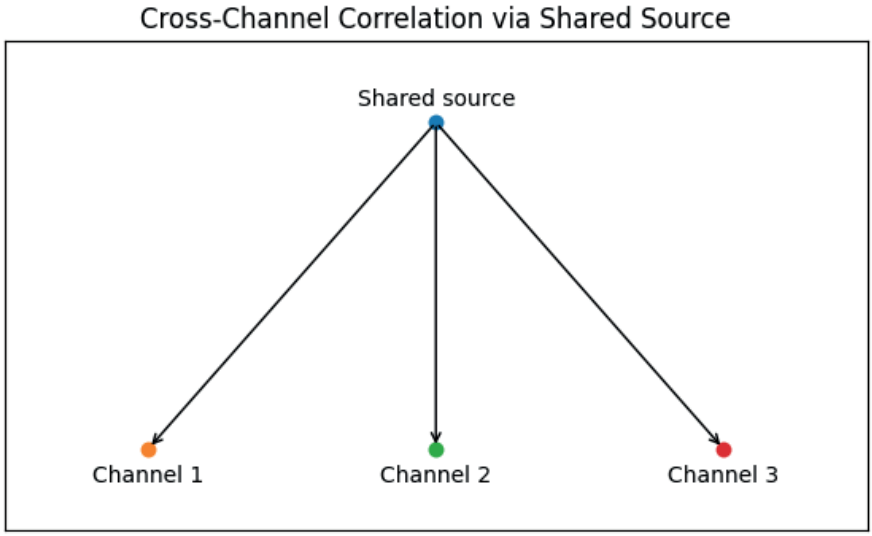

Figure 2.2. Illustration of cross channel correlation caused by a shared physical or environmental source. When multiple measurement channels respond to the same underlying process, their outputs become statistically dependent, violating the independence assumption frequently used in machine learning models.

2.3.1. Hidden Assumptions in Machine Learning Algorithms

Although the IID assumption is often explicitly stated in theoretical analyses, it also appears implicitly in many practical machine learning workflows. Training procedures, cross validation strategies, and performance evaluation metrics typically rely on the premise that the training and testing data originate from the same statistical distribution

For instance, common validation techniques assume that randomly partitioning a dataset into training and testing subsets provides an unbiased estimate of model performance. This assumption holds only when the underlying data are IID. If the data exhibit temporal structure, environmental drift, or hidden correlations, the random partitioning process may produce training and test sets that share similar dependencies. In such situations, the measured performance may overestimate the true predictive capability of the model.

Furthermore, many theoretical generalization guarantees rely on the idea that the empirical distribution observed during training approximates the true data distribution. When the IID assumption is violated, this approximation may break down. The learned model may perform well on the training data

yet fail to maintain similar performance when applied to data collected under slightly different conditions.

Consequently, the IID assumption plays a deeper role than simply simplifying mathematical derivations. It shapes how machine learning systems are evaluated, interpreted, and deployed.

2.3.2. Temporal Dependence

One of the most common sources of IID violation arises from temporal dependence. In many real world systems, data are collected sequentially over time rather than sampled independently from a static distribution. In such cases, successive observations may be influenced by the previous state of the system.

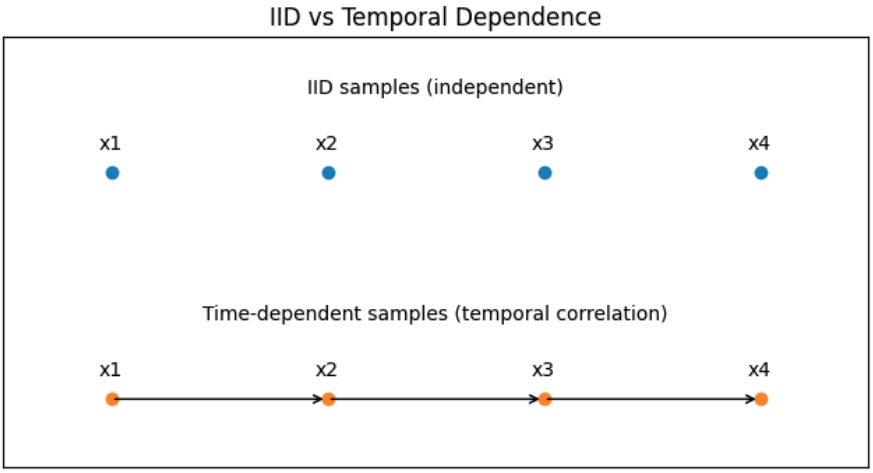

Figure 2.3. Illustration of the IID assumption versus temporally dependent observations. While many machine learning methods assume that samples are independent and identically distributed, real world measurements are often sequentially correlated, violating the independence assumption.

Time series data provide a clear example of this phenomenon. Measurements collected from sensors, financial markets, monitoring systems, or experimental apparatus often exhibit temporal correlations. These correlations may appear as trends, seasonal cycles, periodic oscillations, or slow drift in system behavior.

When temporal dependence is present, the assumption that each observation is independent of the previous one no longer holds. Instead, observations become linked through underlying system dynamics. The statistical properties

of the data may therefore evolve gradually over time, reflecting changes in system conditions or measurement environments.

For machine learning algorithms trained under the assumption of independence, temporal dependence can introduce significant challenges. Models may inadvertently learn patterns that are specific to a particular time interval rather than capturing stable relationships that persist across different periods.

2.3.3. Cross Channel Correlations

Another important source of IID violation emerges in systems where multiple measurement channels are recorded simultaneously. In complex observational environment such as sensor networks, particle detectors, industrial monitoring systems, or biomedical instrumentatio different channels often respond to the same underlying physical processes.

For example, in multi sensor measurement systems, environmental disturbances such as temperature changes or electromagnetic interference may affect multiple channels simultaneously, introducing correlations across observations.

Because these channels share common environmental influences or system dynamics, their outputs may become correlated. For example, a temperature fluctuation affecting a measurement device may simultaneously influence multiple sensors. Similarly, variations in power supply, mechanical vibrations, or electromagnetic interference may propagate across different components of an experimental setup.

When such cross channel dependencies exist, the assumption that individual observations are independent becomes questionable. Even if each channel produces a distinct signal, the underlying noise sources or environmental influences may introduce shared variability across measurements.

For machine learning models, ignoring these dependencies can lead to misleading conclusions. Apparent predictive relationships between variables may arise not from meaningful causal connections, but from shared background influences that simultaneously affect multiple channels.

2.3.4. Environmental Influences

Beyond temporal and cross channel dependencies, external environmental factors can also introduce significant deviations from the IID assumption. Real world data collection rarely occurs under perfectly controlled and unchanging conditions. Instead, measurements are often influenced by slowly varying

environmental parameters such as temperature fluctuations, humidity changes, hardware aging, calibration shifts, or operational adjustments.

These factors can gradually alter the statistical properties of the observed data. For example, sensor sensitivity may drift over time as components degrade, or measurement noise may increase due to environmental disturbances. In such cases, the distribution from which the data are drawn is no longer stationary.

Distributional shifts of this kind create additional challenges for machine learning models. A model trained under one set of environmental conditions may encounter systematically different data when deployed in another context. If the learning algorithm assumes that all observations originate from a single stable distribution, it may struggle to adapt to these changes.

Understanding the influence of environmental variability is therefore essential when evaluating the reliability of machine learning systems applied to real world data.

2.3.5. Summary

Taken together, these factors demonstrate that the IID assumption represents a theoretical idealization rather than a universal property of empirical data. Temporal dependencies, cross channel correlations, and environmental influences introduce structural relationships that violate the assumption of independent and identically distributed observations.

When machine learning models are trained under conditions where independence or distributional stability does not hold, the theoretical guarantees derived from statistical learning theory may become fragile. Performance estimates obtained under IID assumptions may no longer accurately reflect how the model will behave in realistic operational environments.

Recognizing and analyzing these deviations is therefore crucial for understanding the practical limits of machine learning. In noisy and data constrained systems, the reliability of learning algorithms depends not only on model architecture or optimization strategies, but also on how closely the underlying data generating process aligns with the assumptions upon which the learning framework is built.

2.4. Impact of Noise on Generalization

A central objective of machine learning is the ability to generalize beyond the data used during training. Models are typically evaluated using performance metrics measured on a training set, and improvements in these metrics are

often interpreted as evidence of successful learning. However, in the presence of noise, such interpretations may become misleading.

Training error measures how well a model fits the observed data, but it does not necessarily reflect how accurately the model captures the underlying structure of the system. When noise is present in the data, a learning algorithm may reduce training error by adapting to random fluctuations rather than identifying stable patterns. As a result, low training error does not always imply strong predictive performance on new observations.

Noise therefore introduces a fundamental limitation on generalization. Even when a model has sufficient expressive capacity, the information contained in noisy observations may be insufficient to reliably distinguish signal from random variation. In such cases, the effective amount of usable information in the dataset becomes smaller than the nominal sample size.

This limitation becomes particularly visible in the phenomenon known as *noise memorization*. Highly flexible models, including modern deep neural networks, are capable of fitting datasets even when labels contain substantial randomness. Empirical studies have demonstrated that deep networks can eventually achieve near perfect training accuracy even when the labels in the dataset are partially randomized. In such cases, the model is not learning meaningful structure but rather memorizing the idiosyncratic patterns present in the training data.

In such situations, increasing model complexity does not necessarily improve the model's ability to capture the underlying signal. Instead, the model may begin to adapt to random fluctuations present in the training data, effectively memorizing noise rather than learning the true structure of the system. This effect is illustrated conceptually in Figure 2.4.

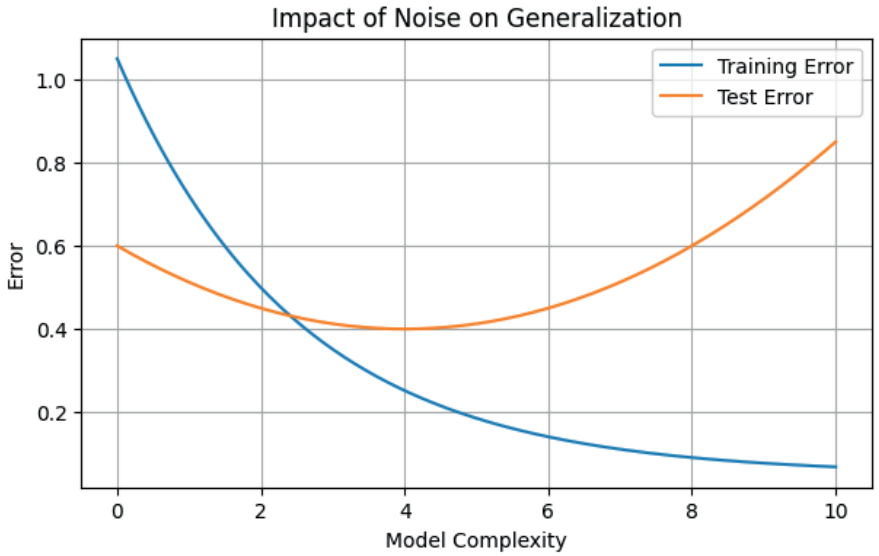

Figure 2.4. Conceptual illustration of the relationship between model complexity and generalization error in the presence of noise. While training error decreases as model capacity increases, test error may eventually rise when the model begins to memorize noise rather than capturing the underlying signal.

This behavior is often referred to as *noise fitting*. Instead of capturing the underlying data generating process, the model adapts to irregularities introduced by noise. While such models may appear highly accurate during training, their predictions can become unstable or unreliable when applied to new data.

Understanding the interaction between noise and generalization is therefore essential for interpreting model performance. In noisy environments, improvements in training accuracy must be evaluated cautiously, as they may reflect the model's ability to fit noise rather than its ability to learn the true structure of the system.

2.5. Theoretical Risks of Simulation and Data Augmentation

When real world data are limited, a common strategy is to generate additional samples through simulation or data augmentation techniques. The underlying idea is that artificially increasing the size of the training dataset can improve model performance and stabilize the learning process. However, this approach introduces its own theoretical challenges.

A key requirement for synthetic data generation is that the simulated samples preserve the statistical properties of the original data distribution. Methods such as Kernel Density Estimation (KDE) or Empirical Cumulative Distribution Functions (ECDF) attempt to approximate the observed distribution and generate additional samples consistent with that approximation. While such approaches can be useful under certain conditions, they rely on the assumption that the estimated distribution accurately reflects the true underlying process.

In practice, this assumption may be fragile. If the simulation procedure fails to reproduce the correct variance structure or dependency patterns present in the real data, the generated samples may introduce artificial regularities or distortions. In such situations, the model may learn features of the simulation process rather than characteristics of the original system.

Another potential issue arises when the generated data lack sufficient diversity. In generative modeling literature, similar phenomena are sometimes described as *mode collapse*, where simulated samples concentrate around limited regions of the distribution and fail to represent the full range of variability observed in the real data.

When these effects occur, the learning algorithm may become increasingly adapted to the properties of the synthetic dataset. Instead of improving generalization, the model may effectively learn the structure of the simulation procedure itself. This risk becomes particularly relevant when the original dataset is small, as the simulation model is then estimated from limited information.

Consequently, while simulation and data augmentation can be powerful tools, their use requires careful consideration of the assumptions underlying the generation process. Without preserving the true statistical and structural characteristics of the data, synthetic samples may inadvertently amplify the very limitations they are intended to mitigate.

2.6. Integrating Physical Priors into Learning

The discussions in the previous sections have emphasized that learning under noisy and limited data is fundamentally constrained by the structure of uncertainty in the observed system. When data are scarce and noise is present, the flexibility of machine learning models can become a source of instability rather than an advantage. In such situations, incorporating prior knowledge about the system can play a critical role in stabilizing the learning process.

In many scientific and engineering applications, the systems being modeled are governed by well established physical principles. Measurement processes,

conservation laws, and structural relationships often impose constraints on how variables can interact. However, standard machine learning approaches are typically designed to operate with minimal assumptions about the data generating mechanism, relying primarily on statistical patterns extracted from observations.

While this data driven flexibility can be useful in settings where little domain knowledge is available, it may also allow models to learn relationships that are inconsistent with the known behavior of the system. When training data are limited or noisy, the absence of structural constraints can lead the learning algorithm to interpret random fluctuations as meaningful patterns.

Introducing physical priors into the learning process provides a way to restrict the space of admissible solutions. Instead of allowing the model to explore all mathematically possible mappings between inputs and outputs, prior knowledge constrains the hypothesis space to configurations that remain consistent with the underlying physical system.

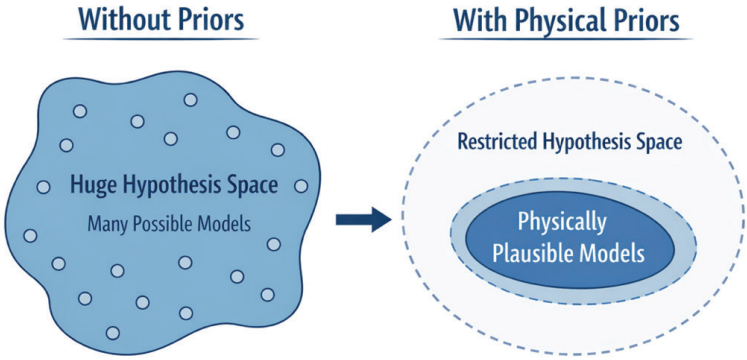

Figure 2.5. Conceptual illustration of the effect of physical priors on the hypothesis space of machine learning models. Without structural constraints, the learning algorithm can explore a very large set of possible mappings between inputs and outputs. Incorporating physical priors restricts this hypothesis space to solutions that remain consistent with the known structure of the system, thereby reducing the risk of fitting noise or spurious correlations.

This restriction acts as a form of regularization, reducing the tendency of the model to overfit noise or spurious correlations.

Another important consideration concerns the choice of input features. In purely data driven models, algorithms may exploit correlations that arise accidentally within the training dataset. Such correlations may appear predictive but often lack a meaningful causal or physical interpretation. As a result, models built on such features may perform well during training but fail to remain stable when conditions change.

Incorporating physically meaningful features can therefore improve the robustness of learning systems. When the model relies on variables that reflect genuine properties of the system, its predictions become less sensitive to incidental variations and measurement noise.

Physical knowledge can be integrated into machine learning models in several ways. Some constraints are hard constraints, meaning that the model must satisfy them exactly. Examples include conservation laws, symmetry relations, or structural dependencies that cannot be violated. Other constraints are soft constraints, which allow limited deviations but penalize violations during training. Such constraints are typically implemented through additional regularization terms in the loss function.

By introducing these forms of prior knowledge, the effective flexibility of the learning algorithm is reduced. Although this reduction may appear to limit the expressive capacity of the model, it often leads to more reliable learning in practice. When the hypothesis space is constrained by physically meaningful relationships, the model becomes less prone to fitting noise and more likely to capture stable patterns in the data.

From this perspective, incorporating physical priors should not be viewed as restricting machine learning models, but rather as guiding them toward physically plausible solutions. Especially in scientific and engineering contexts, combining data driven methods with domain knowledge can provide a more robust foundation for learning under noisy and limited data conditions.

2.7. Discussion: Where Should Machine Learning Stop?

The rapid success of machine learning across many domains has led to an increasing tendency to frame diverse problems as learning tasks. Advances in computational power, the availability of large datasets, and the impressive performance of modern algorithms have reinforced the perception that sufficiently complex models can extract meaningful patterns from almost any type of data. While this perspective has driven significant progress, it also

raises an important methodological question: where should the application of machine learning appropriately stop?

Not every analytical problem is naturally suited to a machine learning formulation.

This distinction can be illustrated conceptually in Figure 2.6.

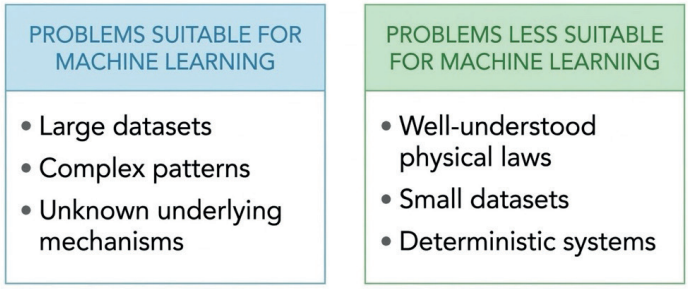

Figure 2.6. Conceptual illustration of problem types with respect to the suitability of machine learning methods. Machine learning is particularly effective in settings characterized by large datasets, complex patterns, and unknown mechanisms, whereas problems governed by well understood physical laws or limited data may be more appropriately addressed using theoretical or statistical approaches.

In some cases, the underlying mechanisms governing a system are already well understood through theoretical models, physical laws, or established statistical frameworks. When such knowledge exists, replacing these models with purely data driven approaches may introduce unnecessary complexity and reduce interpretability without providing meaningful gains in predictive capability.

A related misconception arises when machine learning is implicitly treated as a substitute for statistical reasoning. Although modern learning algorithms build upon statistical principles, they do not replace the need for careful modeling assumptions, uncertainty analysis, or hypothesis testing. Statistical thinking remains essential for interpreting model outputs, evaluating robustness, and understanding the limitations of empirical results.

These issues become particularly visible in the academic literature, where methodological pitfalls occasionally appear in studies that rely heavily on predictive performance metrics while paying less attention to the underlying assumptions of the learning process. High accuracy on benchmark datasets may not always translate into reliable understanding of the system being studied.

Benchmark performance therefore does not necessarily imply scientific validity. A model may perform well on a dataset while still failing to capture the causal mechanisms governing the underlying system. When noise, limited data, or distributional shifts are present, models can appear successful while capturing artifacts of the data rather than genuine structural relationships.

Recognizing these limitations does not diminish the value of machine learning. Instead, it highlights the importance of applying learning methods within an appropriate methodological framework. Machine learning is most powerful when used as a complementary tool one that operates alongside theoretical knowledge, statistical reasoning, and domain expertise.

From this perspective, the central challenge is not whether machine learning should be used, but how and where it should be integrated into the broader process of scientific and analytical inquiry. Careful consideration of model assumptions, data limitations, and domain knowledge is therefore essential for ensuring that machine learning contributes meaningfully to understanding complex systems rather than merely optimizing predictive performance.

References

Arpit, D., et al. (2017). A closer look at memorization in deep networks. ICML.

Belkin, M., et al. (2019). Reconciling modern machine learning and bias–variance trade-off. PNAS.

Breiman, L. (2001). Statistical modeling: The two cultures. Statistical Science.

Clauset, A., Shalizi, C., & Newman, M. (2009). Power law distributions in empirical data. SIAM Review, 51(4), 661–703.

Goodfellow, I., Bengio, Y., & Courville, A. (2016). Deep learning. MIT Press.

Karniadakis, G., et al. (2021). Physics informed machine learning. Nature Reviews Physics.

Kay, S. M. (1993). Fundamentals of statistical signal processing: Estimation theory. Prentice Hall.

Lipton, Z. C., & Steinhardt, J. (2019). Troubling trends in machine learning scholarship. Queue, 17(1), 45–77.

Middleton, D. (1999). Non-Gaussian noise models. IEEE Transactions on Information Theory, 45(4), 1129–1149.

Mohri, M., Rostamizadeh, A., & Talwalkar, A. (2018). Foundations of machine learning. MIT Press.

Papoulis, A., & Pillai, S. (2002). Probability, random variables and stochastic processes. McGraw-Hill.

Pineau, J., et al. (2021). Improving reproducibility in machine learning research. JMLR, 22(164), 1–20.

Raissi, M., Perdikaris, P., & Karniadakis, G. (2019). Physics informed neural networks. Journal of Computational Physics, 378.

Rudin, C. (2019). Stop explaining black box ML models. Nature Machine Intelligence.

Sculley, D., et al. (2018). Winner's curse? On pace, progress, and empirical rigor. ICLR Workshop.

Shalev-Shwartz, S., & Ben-David, S. (2014). Understanding machine learning. Cambridge University Press.

Shorten, C., & Khoshgoftaar, T. (2019). A survey on image data augmentation. Journal of Big Data, 6(1).

Taleb, N. N. (2020). Statistical consequences of fat tails. STEM Academic Press.

Willard, J., et al. (2020). Integrating physics based modeling with ML. Nature Reviews Earth & Environment.

Zhang, C., et al. (2017). Understanding deep learning requires rethinking generalization. ICLR.