36 Causality

This chapter was written by Victor Veitch and Alex D'Amour.

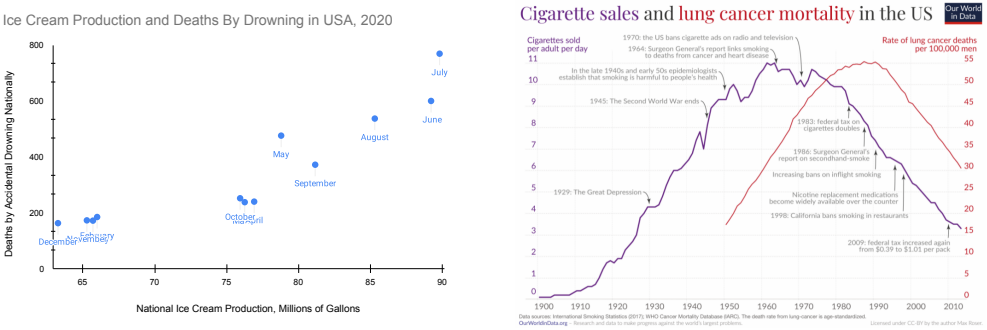
36.1 Introduction

The bulk of machine learning considers relationships between observed variables with the goal of summarizing these relationships in a manner that allows predictions on similar data. However, for many problems, our main interest is to predict how system would change if it were observed under different conditions. For instance, in healthcare, we are interested in whether a patient will recover if given a certain treatment (as opposed to whether treatment and recovery are associated in the observed data). **Causal inference** addresses how to formalize such problems, determine whether they can be solved, and, if so, how to solve them. This chapter covers the fundamentals of this subject. Code examples for the discussed methods are available at https://github.com/vveitch/causality-tutorials.

To make the gap between observed data modeling and causal inference concrete, consider the relationships depicted in Figure 36.1a and Figure 36.1b. Figure 36.1a shows the relationship between deaths by drowning and ice cream production in the United States in 1931 (the pattern holds across most years). Figure 36.1b shows the relationship between smoking and lung cancer across various countries. In each case, there is a strong positive association. Faced with this association, we might ask: could we reduce drowning deaths by banning ice cream? Could we reduce lung cancer by banning cigarettes? We intuitively understand that these interventional questions have different answers, despite the fact that the observed associations are similar. Determining the causal effect of some intervention in the world requires some such causal hypothesis about the world.

For concreteness, consider three possible explanations for the association between ice cream and drowning. Perhaps eating ice cream does cause people to drown—due to stomach cramps or similar. Or, perhaps, drownings increase demand for ice cream—the survivors eat huge quantities of ice cream to handle their grief. Or, the association may be due (at least in part) to a common cause: warm weather makes people more likely to eat ice cream and more likely to go swimming (and, hence, to drown). Under all three scenarios, we can observe exactly the same data, but the implications for an ice cream ban are very different. Hence, answering questions about what will happen under an intervention requires us to incorporate some causal knowledge of the world—e.g., which of these scenarios is plausible?

Our goal in this chapter to introduce the essentials of estimating causal effects. The high-level approach has three steps.



(a) *(b)*

Figure 36.1: Correlation is not causation. (a) Ice cream production is strongly associated with deaths by drowning. Ice cream production data from the US Department of Agriculture National Agricultural Statistics Service. Drowning data from the National Center for Health Statistics at the United States Centers for Disease Control. (b) Smoking is strongly associated with lung cancer. From ourworldindata.org/smoking-big-problem-in-brief. Used with kind permission of Max Roser.

- **Causal Estimands**: The first step is to formally define the quantities we want to estimate. These are summaries of how the world would change under intervention, rather than summaries of the world as it has already been observed. E.g., we want to formalize "The expected number of drownings in the United States if we ban ice cream".
- **Identification**: The next step is to identify the causal estimands with quantities that can, in principle, be estimated from observational data. This step involves codifying our causal knowledge of the world and translating this into a statement such as, "The causal effect is equal to the expected number of drownings after adjusting for month". This step tells us what causal questions we could answer with perfect knowledge of the observed data distribution.
- **Estimation**: Finally, we must estimate the observable quantity using a finite data sample. The form of causal estimands favors certain efficient estimation procedures that allow us to exploit non-parametric (e.g., machine learning) predictive models.

In this chapter, we'll mainly focus on the estimation of the causal effect of an intervention averaged over all members of a population, known as the **Average Treatment Effect** or **ATE**. This is the most common problem in applied causal inference work. It is in some sense the simplest problem, and will allow us to concretely explain the use and importance of the fundamental causal concepts. These causal concepts include structural causal models, causal graphical models, the do-calculus, and efficient estimation using influence function techniques. This problem is also useful for understanding the role that standard predictive modeling and machine learning play in estimating causal quantities.

36.2 Causal Formalism

In causal inference, the goal is to use data to learn about how the outcome in the world would change under intervention. In order to make such inferences, we must also make use of our causal knowledge

of the world. This requires a formalism that lets us make the notion of intervention precise and lets us encode our causal knowledge as assumptions.

36.2.1 Structural Causal Models

Consider a setting in which we observe four variables from a population of people: A_i, an indicator of whether or not person i smoked at a particular age, Y_i, an indicator of whether or not person i developed lung cancer at a later age, H_i, a "health consciousness" index that measures a person's health-consciousness (perhaps constructed from a set of survey responses about attitudes toward health), and G_i, an indicator for whether the person has a genetic predisposition towards cancer. Suppose we observe a dataset of these variables drawn independently and identically from a population, $(A_i, Y_i, H_i) \overset{\text{iid}}{\sim} P^{\text{obs}}$, where "obs" stands for "observed".

In standard practice, we model data like these using probabilistic models. Notably, there are many different ways to specify a probabilistic model for the same observed distribution. For example, we could write a probabilistic model for P^{obs} as

$$A \sim P^{\text{obs}}(A) \tag{36.1}$$

$$H|A \sim P^{\text{obs}}(H|A) \tag{36.2}$$

$$Y|A,H \sim P^{\text{obs}}(Y|H,A) \tag{36.3}$$

$$G|A,H,Y \sim P^{\text{obs}}(G|A,H,Y) \tag{36.4}$$

This is a valid factorization, and sampling variables in this order would yield valid samples from the joint distribution P^{obs}. However, this factorization does not map well to a mechanistic understanding of how these variables are causally related in the world. In particular, it is perhaps more plausible that health-consciousness H causally precedes smoking status A, since a person's health-consciousness would influence their decision to smoke.

These intuitions about causal ordering are intimately tied to the notion of intervention. Here, we will focus on a notion of intervention that can be represented in terms of "structural" models that describe mechanistic relationships between variables. The fundamental objects that we will reason about are **structural causal models**, or SCM's. SCM's resemble probabilistic models, but they encode additional assumptions (see also Section 4.7). Specifically, SCM's serve two purposes: they describe a probabilistic model *and* they provide semantics for transforming the data-generating process through intervention.

Formally, SCM's describe a mechanistic data generating process with an ordered sequence of equations that resemble assignment operations in a program. Each variable in a system is determined by combining other modeled variables (the causes) with exogenous "noise" according to some (unknown) deterministic function. For instance, a plausible SCM for P^{obs} might be

$$G \leftarrow f_G(\xi_0) \tag{36.5}$$

$$H \leftarrow f_H(\xi_1) \tag{36.6}$$

$$A \leftarrow f_A(H, \xi_2) \tag{36.7}$$

$$Y \leftarrow f_Y(G, H, A, \xi_3) \tag{36.8}$$

where the (unknown) functions f are fixed, and the variables ξ are unmeasured causes, modeled as independent random "noise" variables. Conceptually, the functions f_G, f_H, f_A, f_Y describe deter-

Author: Kevin P. Murphy. (C) MIT Press. CC-BY-NC-ND license

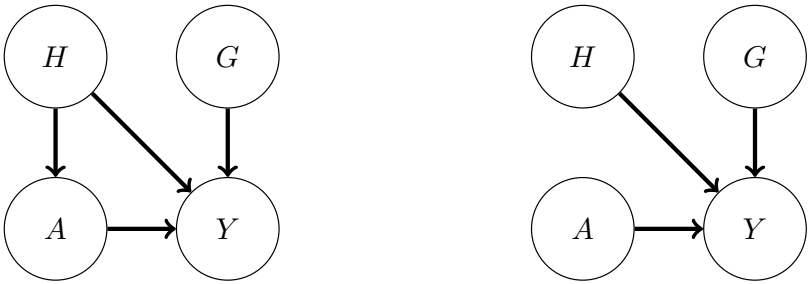


Figure 36.2: (a) Causal graph illustrating relationships between smoking A, cancer Y, health conciousness H, and genetic cancer pre-disposition G. (b) "Mutilated" causal graph illustrating relationships under an intervention on smoking A.

ministic physical relationships in the real world, while the variables ξ are hidden causes that are sufficient to distinguish each unit i in the population. Because we assume that each observed unit i is drawn at random from the population, we model ξ as random noise.

SCM's imply probabilistic models, but not the other way around. For example, our example SCM implies probabilistic model for the observed data based on the factorization $P^{\text{obs}}(G, H, A, Y) = P^{\text{obs}}(G)P^{\text{obs}}(H)P^{\text{obs}}(A \mid H)P^{\text{obs}}(Y \mid A, H)$. Thus, we could sample from the SCM in the same way we would from a probabilistic model: draw a set of noise variables ξ and evaluate each assignment operation in the SCM in order.

Beyond the probabilistic model, an SCM encodes additional assumptions about the effects of **interventions**. This can be formalized using the **do calculus** (as in the verb "to do"), which we describe in Section 36.8; But in brief, interventions are represented by replacing assignment statements. For example, if we were interested in the distribution of Y in the hypothetical scenario that smoking were eliminated, we could set the second line of the SCM to be $A \leftarrow 0$. We would denote this by $P(Y|\text{do}(A = 0), H)$. Because the f functions in the SCM are assumed to be invariant mechanistic relationships, the SCM encodes the assumption that this edited SCM generates data that we would see if we really applied this intervention in the world. Thus, the ordering of statements in an SCM are load-bearing: they imply substantive assumptions about how the world changes in response to interventions. This is in contrast to more standard probabilistic models where variables can be rearranged by applications of Bayes Rule without changing the substantive implications of the model. (See also Section 4.7.3.)

We note that structural causal model may not incorporate all possible notions of causality. For example, laws based on conserved quantities or equilibria—e.g., the ideal gas law—do not trivially map to SCMs, though these are fundamental in disciplines such as physics and economics. Nonetheless, we will confine our discussion to SCMs.

36.2.2 Causal DAGs

SCM's encode many details about the assumed generative process of a system, but often it is useful to reason about causal problems at a higher level of abstraction. In particular, it is often useful

to separate the causal structure of a problem from the particular functional form of those causal relationships. **Causal graphs** provide this level of abstraction. A causal graph specifies which variables causally affect other variables, but leaves the parametric form of the structural equations f unspecified. Given an SCM, the corresponding causal graph can be drawn as follows: for each line of the SCM, draw arrows from the variables on the right hand side to variables on the left hand side. The causal DAG for our smoking-cancer example is shown in Figure 36.2. In this way, causal DAGs are related to SCMs in the same way that probabilistic graphical models (PGMs) are related to probabilistic models.

In fact, in the same way that SCMs imply a probabilistic model, causal DAGs imply a PGM. Functionally, causal graphs behave as probabilistic graphical models (Chapter 4). They imply conditional independence relationships between the variables in the observed data in same way. They obey the Markov property: If $X \leftarrow Y \rightarrow Z$ then $X \perp\!\!\!\perp Z|Y$; recall d-separation (Section 4.2.4.1). Additionally, if $X \rightarrow Y \leftarrow Z$ then, usually, $X \not\perp\!\!\!\perp Z|Y$ (even if X and Z are marginally independent). In this case, Y is called a **collider** for X and Z.

Conceptually, the difference between causal DAGs and PGMs is that probabilistic graphical models encode our assumptions about statistical relationships, whereas causal graphs encode our (stronger) assumptions about causal relationships. Such causal relationships can be used to derive how statistical relationships would change under intervention.

Causal graphs also allow us to reason about the causal and non-causal origins of statistical dependencies in observed data without specifying a full SCM. In a causal graph, two variables—say, A and D—can be statistically associated in different ways. First, there can be a directed path from (ancestor) A to (descendant) D. In this case, A is a causal ancestor of D and interventions on A will propagate through to change D; $\mathrm{P}(D|\mathrm{do}(A = a)) \neq \mathrm{P}(D|\mathrm{do}(A = a'))$. For example, smoking is a causal ancestor of cancer in our example. Alternatively, A and D could share a common cause—there is some variable C such that there is a directed path from C to A and from C to D. If A and D are associated only through such a path then interventions on A will not change the distribution of D. However, it is still the case that $\mathrm{P}(D|A = a) \neq \mathrm{P}(D|A = a')$—observing different values of A changes our guess for the value of D. The reason is that A carries information about C, which carries information about D. For example, suppose we lived in a world where there was no effect of smoking on developing cancer (e.g., everybody vapes), there would nevertheless be an association between smoking and cancer because of the path $A \leftarrow H \rightarrow Y$. The existence of such "backdoor paths" is one core reason that statistical and causal association are not the same. Of course, more complicated variants of these associations are possible—e.g., C is itself only associated with A through a backdoor path—but this already captures the key distinction between causal and non-causal paths.

Recall that our aim in introducing SCMs and causal graphs is to enable us to formalize our causal knowledge of the world and to make precise what interventional quantities we'd like to estimate. Writing down a causal graph gives a simple formal way to encode our knowledge of the causal structure of a problem. Usefully, this causal structure is sufficient to directly reason about the implications of interventions without fully specifying the underlying SCM. The key observation is that if a variable A is intervened on then, after intervention, none of the other variables are causes of A. That is, when we replace a line of an SCM with a statement directly assigning a variable a particular value, we cut off all dependencies that variable had on its causal parents. Accordingly, in the causal graph, the intervened on variable has no parents. This leads us to the **graph surgery** notion of intervention: an intervention that sets A to a is the operation that deletes all incoming edges to A in the graph, and then conditions on $A = a$ in the resulting probability distribution (which is defined

Author: Kevin P. Murphy. (C) MIT Press. CC-BY-NC-ND license

by the conditional independence structure of the post-surgery graph). We'll use Pearl's do notation to denote this operation. $\mathrm{P}(\mathbf{X}|\mathrm{do}(A=a))$ is the distribution of $\mathbf{X}$ given $A=a$ under the mutilated graph that results from deleting all edges going into A. Similarly, $\mathbb{E}[\mathbf{X}|\mathrm{do}(A=a)] \triangleq \mathbb{E}_{\mathrm{P}(\mathbf{X}|\mathrm{do}(A=a))}[\mathbf{X}]$. Thus, we can formalize statements such as "The average effect of receiving drug A" as

$$\mathrm{ATE} = E[Y|\mathrm{do}(A=1)] - \mathbb{E}[Y|\mathrm{do}(A=0)], \tag{36.9}$$

where ATE stands for Average Treatment Effect.

For concreteness, consider our running example. We contrast the distribution that results by conditioning on A with the distribution that results from intervening on A:

$$\mathrm{P}(Y,H,G|A=a) = \mathrm{P}(Y|H,G,A=a)\mathrm{P}(G)\mathrm{P}(H|A=a) \tag{36.10}$$

$$\mathrm{P}(Y,H,G|\mathrm{do}(A=a)) = \mathrm{P}(Y|H,G,A=a)\mathrm{P}(G)\mathrm{P}(H) \tag{36.11}$$

The key difference between these two distributions is that the standard conditional distribution describes a population where health consciousness H has the distribution that we observe among individuals with smoking status $A=a$, while the interventional distribution described a population where health consciousness H follows the marginal distribution among all individuals. For example, we would expect $P(H \mid A = \texttt{smoker})$ to put more mass on lower values of H than the marginal health consciousness distribution than the marginal distribution $P(H)$, which would also include non-smokers. The intervention distribution thus incorporates a hypothesis of how smoking would affect the subpopulation individuals who tend to be too health conscious to smoke in the observed data.

36.2.3 Identification

A central challenge in causal inference is that many different SCM's can produce identical distributions of observed data. This means that, on the basis of observed data alone, we cannot uniquely identify the SCM that generated it. This is true no matter how large of a data sample is available to us.

For example, consider the setting where there is a treatment A that may or may not have an effect on outcome Y, and where both the treatment and outcome are known to be affected by some *unobserved* common binary cause U. Now, we might be interested in the causal estimand $E[Y|\mathrm{do}(A=1)]$. In general, we can't learn this quantity from the observed data. The problem is that, we can't tell apart the case where the treatment has a strong effect from the case where the treatment has no effect, but $U=1$ both causes people to tend to be treated and increases the probability of a positive outcome. The same observation shows we can't learn the (more complicated) interventional distribution $\mathrm{P}(Y|\mathrm{do}(A=1))$ (if we could learn this, then we'd get the average effect automatically).

Thus, an important part of causal inference is to augment the observed data with knowledge about the underlying causal structure of the process under consideration. Often, these assumptions can narrow the space of SCM's sufficiently so that there is only one value of the causal estimand that is compatible with the observed data. We say that the causal estimand is **identified** or **identifiable** under a given set of assumptions if those assumptions are sufficient to provide a unique answer. There are many different sets of sufficient conditions that yield identifiable causal effects; we call each set of sufficient conditions an **identification strategy**.

Given a set of assumptions about the underlying SCM, the most common way to show that a causal estimand is identified is by construction. Specifically, if the causal estimand can be written

entirely in terms of observable probability distributions, then it is identifed. We call such a function of observed distributions a **statistical estimand**. Once such a statistical estimand has been recovered, we can then construct and analyze an estimator for that quantity using standard statistical tools. As an example of a statistical estimand, in the SCM above, it can be shown the ATE as defined in Equation (36.9), is equal to the following statistical estimand

$$\text{ATE} \overset{(*)}{=} \tau^{\text{ATE}} \triangleq \mathbb{E}[\mathbb{E}[Y|H, A=1] - \mathbb{E}[Y|H, A=0]], \tag{36.12}$$

where the equality $(*)$ only holds because of some specific properties of the SCM. Note that the RHS above only involves conditional expectations between observed variables (there are no do operators), so τ^{ATE} is only a function of observable probability distributions.

There are many kinds of assumptions we might make about the SCM governing the process under consideration. For example, the following are assertions we might make about the system in our running example:

1. The probability of developing cancer is additive on the logit scale in A, G, and H (i.e., logistic regression is a well-specified model).
2. For each individual, smoking can never decrease the probability of developing cancer.
3. Whether someone smokes is influenced by their health consciousness H, but not by their genetic predisposition to cancer G.

These assumptions range from strong parametric assumptions fully specifying the form of the SCM equations, to non-parametric assumptions that only specify what the inputs to each equation are, leaving the form fully unspecified. Typically, assumptions that fully specify the parametric form are very strong, and would require far more detailed knowledge of the system under consideration than we actually have. The goal in identification arguments is to find a set of assumptions that are weak enough that they might be plausibly true for the system under consideration, but which are also strong enough to allow for identification of the causal effect.

If we are not willing to make any assumptions about the functional form of the SCM, then our assumptions are just about which variables affect (and do not affect) the other variables. In this sense, such which-affects-which assumptions are minimal. These assumptions are exactly the assumptions captured by writing down a (possibly incomplete) causal DAG, showing which variables are parents of each other variable. The graph may be incomplete because we may not know whether each possible edge is present in the physical system. For example, we might be unsure whether the gene G actually has a causal effect on health consciousness H. It is natural to ask to what extent we can identify causal effects only on the basis of partially specified causal DAGs. It turns out much progress can be made based on such non-parametric assumptions; we discuss this in detail in Section 36.8.

We will also discuss certain assumptions that cannot be encoded in a causal graph, but that are still weaker than assuming that full functional forms are known. For example, we might assume that the outcome is affected additively by the treatment and any confounders, with no interaction terms between them. These weaker assumptions can enable causal identification even when assuming the causal graph alone does not.

It is worth emphasizing that every causal identification strategy relies on assumptions that have some content that cannot be validated in the observed data. This follows directly from the ill-posedness of causal problems: if the assumptions used to identify causal quantities could be validated, that

Author: Kevin P. Murphy. (C) MIT Press. CC-BY-NC-ND license

would imply that the causal estimand was identifiable from the observed data alone. However, since we know that there are many values of the causal estimand that are compatible with observed data, it follows that the assumptions in our identification strategy must have unobservable implications.

36.2.4 Counterfactuals and the Causal Hierarchy

Structural causal models let us formalize and study a hierarchy of different kinds of query about the system under consideration. The most familiar is observational queries: questions that are purely about statistical associations (e.g., "Are smoking and lung cancer associated in the population this sample was drawn from?"). Next is interventional queries: questions about causal relationships at the population level (e.g., "How much does smoking increase the probability of cancer in a given population?"). The rest of this chapter is focused on the defintion, identification, and estimation of interventional queries. Finally, there are counterfactual queries: questions about causal relationships at the level of specific individuals, had something been different (e.g., "Would Alice have developed cancer had she not smoked?"). This causal hierarchy was popularized by [Pea09a, Ch. 1].

Inteventional queries concern the prospective effect of an intervention on an outcome; for example, if we intervene and prevent a randomly sampled individual from smoking, what is the probability they develop lung cancer? Ultimately, the probability statement here is about our uncertainity about the "noise" variables ξ in the SCM. These are the unmeasured factors specific to the randomly selected individual. The distribution is determined by the population from which that individual is sampled. Thus, interventional queries are statements about populations. Interventional queries can be written in terms of conditional distributions using do-notation, e.g. $P(Y|\text{do}(A = 0))$. In our example, this represents the distribution of lung cancer outcomes for an individual selected at random and prevented from smoking.

Counterfactual queries concern how an observed outcome might have been different had an intervention been applied in the past. Counterfactual queries are often framed in terms of attributing a given outcome to a particular cause. For example, would Alice have developed cancer had she not smoked? Did most smokers with lung cancer develop cancer because they smoked? Counterfactual queries are so called because they require a comparison of counterfactual outcomes within individuals. In the formalism of SCM's, counterfactual outcomes for an individual i are generated by running the same values of ξ_i through differently intervened SCM's. Counterfactual outcomes are often written in terms of **potential outcomes** notation. In our running smoking example, this would look like:

$$Y_i(a) \triangleq f_Y(G_i, H_i, a, \xi_{3,i}). \tag{36.13}$$

That is, $Y_i(a)$ is the outcome we would have seen had A been set to a while all of $G_i, H_i, \xi_{3,i}$ were kept fixed.

It is important to understand what distinguishes interventional and fundamentally counterfactual queries. Just because a query can be written in terms of potential outcomes does not make it a counterfactual query. For example, the average treatment effect, which is the canonical interventional query, is easy to write in potential outcomes notation:

$$\text{ATE} = \mathbb{E}[Y_i(1) - Y_i(0)]. \tag{36.14}$$

Instead, the key dividing line between counterfactual and interventional queries is whether the query requires knowing the joint distribution of potential outcomes within individuals, or whether marginal

distributions of potential outcomes across individuals will suffice. An important signature of a counterfactual query is conditioning on the value of one potential outcome. For example, "the lung cancer rate among smokers who developed cancer, had they not smoked" is a counterfactual query, and can be written as:

$$\mathbb{E}[Y_i(0) \mid Y_i(1) = 1, A_i = 1] \tag{36.15}$$

Answering this query requires knowing how individual-level cancer outcomes are related (through $\xi_{3,i}$) across the worlds where the each individual i did and did not smoke. Notably, this query cannot be rewritten using do-notation, because it requires a distinction between $Y(0)$ and $Y(1)$ while the ATE can: $\mathbb{E}[Y \mid \text{do}(A = 1)] - \mathbb{E}[Y \mid \text{do}(A = 0)]$.

Counterfactual queries require categorically more assumptions for identification than interventional ones. For identifying interventional queries, knowing the DAG structure of an SCM is often sufficient, while for counterfactual queries, some assumptions about the functional forms in the SCM are necessary. This is because only one potential outcome is ever observed for each individual, so the dependence between potential outcomes within individuals is not observable. For example, the data in our running example provide no information on how individual-level smoking and non-smoking cancer risk are related. Thus, answering a question like "Did smokers who developed cancer have lower non-smoking cancer risk than smokers who did not develop cancer?", requires additional assumptions about how characteristics encoded in ξ_i are translated to cancer outcomes. To answer this question without such assumptions, we would need to observe smokers who developed cancer in the alternate world where they did not smoke. Because they compare how individuals would have turned out under different generating processes, counterfactual queries are often referred to as "cross-world" quantities. On the other hand, interventional queries only require understanding the marginal distributions of potential outcomes $Y_i(0)$ and $Y_i(1)$ across individuals; thus, no cross-world information is necessary at the individual level.

We conclude this section by noting that counterfactual outcomes and potential outcomes notation are often conceptually useful, even if they are not used to explicitly answer counterfactual queries. Many causal queries are more intuitive to formalize in terms of potential outcomes. E.g., "Would I have smoked if I was more health conscious?" may be more intuitive than "Would a randomly sampled individual from the same population have smoked had they been subject to an intervention that made them more health concious?". In fact, some schools of causal inference use potential outcomes, rather than DAGs, as their primary conceptual building block [See IR15]. Causal graphs and potential outcomes both provide ways to formulate interventional queries and causal assumptions. Ultimately, these are mathematically equivalent. Nevertheless, practically, they have different strengths. The main advantage of potential outcomes is that counterfactual statements often map more directly to our mechanistic understanding of the world. This can make it easier to articulate causal desiderata and causal assumptions we may wish to use. On the other hand, the potential outcomes notation does not automatically distinguish between interventional and counterfactual queries. Additionally, causal graphs often give an intuitive and easy way of articulating assumptions about structural causal models involving many variables—potential outcomes get quickly unwieldly. In short: both formalizations have distinct advantages, and those advantages are simply about how easy it is to translate our causal understanding of the world into crisp mathematical assumptions.

Author: Kevin P. Murphy. (C) MIT Press. CC-BY-NC-ND license

36.3 Randomized Control Trials

We now turn to the business of estimating causal effects from data. We begin with **randomized control trials**, which are experiments designed to make the causal concerns as simple as possible.

The simplest situation for causal estimation is when there are no common causes of A and Y. The world is rarely so obliging as to make this the case. However, sometimes we can design an experiment to enforce the no-common-causes structure. In randomized control trials we assign each participant to either the treatment or control group at random. Because random assignment does not depend on any property of the units in the study, there are no causes of treatment assignment, and hence also no common causes of Y and A.

In this case, it's straightforward to see that $\mathrm{P}(Y|\mathrm{do}(A = a) = \mathrm{P}(Y|a)$. This is essentially by definition of the graph surgery: since A has no parents, the mutilated graph is the same as the original graph. Indeed, the graph surgery definition is chosen to make this true: any sensible formalization of causality should have this identification result.

It is common to use RCTs to study the average treatment effect,

$$\mathrm{ATE} = E[Y|\mathrm{do}(A = 1)] - \mathbb{E}[Y|\mathrm{do}(A = 0)]. \tag{36.16}$$

This is the expected difference between being assigned treatment and assigned no treatment for a randomly chosen member of the population. It's easy to see that in an RCT this causal quantity is identified as a parameter τ^{RCT} of the observational distribution:

$$\tau^{\mathrm{RCT}} = \mathbb{E}[Y|A = 1] - \mathbb{E}[Y|A = 0].$$

Then, a natural estimator is:

$$\hat{\tau}^{\mathrm{RCT}} \triangleq \frac{1}{n_A} \sum_{i:A_i=1} Y_i - \frac{1}{n - n_A} \sum_{i:A_i=0} Y_i, \tag{36.17}$$

where n_A is the number of units who received treatment. That is, we estimate the average treatment effect as the difference between the average outcome of the treated group and the average outcome of the untreated (control) group. [1]

Randomized control trials are the gold standard for estimating causal effects. This is because we know *by design* that there are no confounders that can produce alternative causal explanations of the data. In particular, the assumption of the triangle DAG—there are no unobserved confounders—is enforced by design. However, there are limitations. Most obviously, randomized control trials are sometimes infeasible to conduct. This could be due to expense, regulatory restrictions, or more fundamental difficulties (e.g., in developmental economics, the response of interest is sometimes collected decades after treatment). Additionally, it may be difficult to ensure that the participants in an RCT are representative of the population where the treatment will be deployed. For instance, participants in drug trials may skew younger and poorer than the population of patients who will ultimately take the drug.

1. There is a literature on efficient estimation of causal effects in RCT's going back to Fisher [Fis25] that employ more sophisticated estimators. See also Lin [Lin13a] and Bloniarz et al. [Blo+16] for more modern treatments.

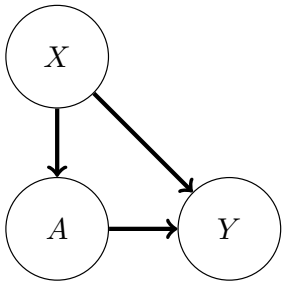


Figure 36.3: A causal DAG illustrating a situation where treatment A and outcome Y are both influenced by observed confounders X.

36.4 Confounder Adjustment

We now turn to the problem of estimating causal effects using observational (i.e., not experimental) data. The most common application of causal inference is estimating the average treatment effect (ATE) of an intervention. The ATE is also commonly called the **average causal effect**, or ACE. Here, we focus on the important special case where the treatment A is binary, and we observe the outcome Y as well as a set of common causes X that influence both A and Y.

36.4.1 Causal Estimand, Statistical Estimand, and Identification

Consider a problem where we observe treatment A, outcome Y, and covariates X, which are drawn i.i.d. from some unknown distribution P. We wish to learn the average treatment effect: the expected difference between being assigned treatment and assigned no treatment for a randomly chosen member of the population. Following the discussion in the introduction, there are three steps to learning this quantity: mathematically formalize the causal estimand, give conditions for the causal estimand to be identified as a statistical estimand, and, finally, estimate this statistical estimand from data. We now turn to the first two steps.

The average treatment effect is defined to be the difference between the average outcome if we *intervened* and set A to be 0, versus the average outcome if we intervented and set A to be 1. Using the do notation, we can write this formally as

$$\text{ATE} = \mathbb{E}[Y|\text{do}(A=1)] - \mathbb{E}[Y|\text{do}(A=0)]. \tag{36.18}$$

The next step is to articulate sufficient conditions for the ATE to be identified as a statistical estimand (a parameter of distribution P). The key issue is the possible presence of **confounders**. Confounders are "common cause" variables that affect both the treatment and outcome. When there are confounding variables in observed data, the sub-population of people who are observed to have received one level of the treatment A will differ from the rest of the population in ways that are relevant to their observed Y. For example, there is a strong positive association between horseback riding in childhood (treatment) and healthiness as an adult (outcome) [RB16]. However, both of these quantities are influenced by wealth X. The population of people who rode horses as children ($A=1$) is wealthier than the population of people who did not. Accordingly, horseback-riding population

Author: Kevin P. Murphy. (C) MIT Press. CC-BY-NC-ND license

will have better health outcomes even if there is no actual causal benefit of horseback riding for adult health.

We'll express the assumptions required for causal identification in the form of a causal DAG. Namely, we consider the simple triangle DAG in Figure 36.3, where the treatment and outcome are influenced by *observed* confounders X. It turns out that the assumption encoded by this DAG suffices for identification. To understand why this is so, recall that the target causal effect is defined according to the distribution we would see if the edge from X to A was removed (that's the meaning of do). The key insight is that because the intervention only modifies the relationship between X and A, the structural equation that generates outcomes Y given X and A, illustrated in Figure 36.3 as the $A \to Y \leftarrow X$, is the same even after the $X \to Y$ edge is removed. For example, we might believe that the physiological processes by which smoking status A and confounders X produce lung cancer Y remain the same, regardless of how the decision to smoke or not smoke was made. Secondly, because the intervention does not change the composition of the population, we would also expect the distribution of background characteristics X to be the same between the observational and intervened processes.

With these insights about invariances between observed and interventional data, we can derive a statistical estimand for the ATE as follows.

Theorem 2 (Adjustment with No Unobserved Confounders)**.** *We observe $A, Y, X \sim P$. Suppose that*

1. *(Confounders observed) The data obeys the causal structure in Figure 36.3. In particular, X contains all common causes of A and Y and no variable in X is caused by A or Y.*
2. *(Overlap) $0 < \mathrm{P}(A = 1|X = x) < 1$ for all values of x. That is, there are no individuals for whom treatment is always or never assigned.*

Then, the average treatment effect is identified as $\mathrm{ATE} = \tau$*, where*

$$\tau = \mathbb{E}[\mathbb{E}[Y|A = 1, X]] - \mathbb{E}[\mathbb{E}[Y|A = 0, X]]. \tag{36.19}$$

Proof. First, we expand the ATE using the tower property of expectation, conditioning on X. Then, we apply the invariances discussed above:

$$\begin{aligned} ATE &= \mathbb{E}[Y|\mathrm{do}(A = 1)] - \mathbb{E}[Y|\mathrm{do}(A = 0)] && (36.20)\\ &= \mathbb{E}[\mathbb{E}[Y|\mathrm{do}(A = 1), X]] - \mathbb{E}[\mathbb{E}[Y|\mathrm{do}(A = 0), X]] && (36.21)\\ &= \mathbb{E}[\mathbb{E}[Y|A = 1, X]] - \mathbb{E}[\mathbb{E}[Y|A = 0, X]] && (36.22) \end{aligned}$$

The final equality is the key to passing from a causal to observational quantity. This follows because, from the causal graph, the conditional distribution of Y given A, X is the same in both the original graph and in the mutilated graph created by removing the edge from X to A. This mutilated graph defines $\mathrm{P}(Y|\mathrm{do}(A = 1), X)$, so the equality holds.

The condition that $0 < \mathrm{P}(A = 1|X = x) < 1$ is required for the first equality (the tower property) to be well defined. □

Note that Equation (36.19) is a function of only conditional expectations and distributions that appear in the observed data distribution (in particular, it contains no "do" operators). Thus, if we

can fully characterize the observed data distribution P, we can map that distribution to a unique ATE.

It is useful to note how τ differs from the naive estimand $\mathbb{E}[Y|A=1] - \mathbb{E}[Y|A=0]$ that just reports the treatment-outcome association without adjusting for confounding. The comparison is especially clear when we write out the outer expectation in τ explicitly as an integral over X:

$$\tau = \int \mathbb{E}[Y \mid A=1, X]P(X)dX - \int \mathbb{E}[Y \mid A=0, X]P(X)dX \tag{36.23}$$

We can write the naive estimand in a similar form by applying the tower property of expectation:

$$\mathbb{E}[Y \mid A=1] - \mathbb{E}[Y \mid A=0] = \int \mathbb{E}[Y \mid A=1, X]P(X \mid A=1)dX - \int \mathbb{E}[Y \mid A=0, X]P(X \mid A=0)dX \tag{36.24}$$

The key difference is the probability distribuiton over X that is being integrated over. The observational difference in means integrates over the distinct conditional distributions of confounders X, depending on the value of A. On the other hand, in the ATE estimand τ, we integrate over the same distribution $P(X)$ for both levels of the treatment.

Overlap In addition to the assumption on the causal structure, identification requires that there is sufficient random variation in how treatments are assigned.

Definition 1. *A distribution P on A, X satisfies* **overlap** *if $0 < P(A=1|x) < 1$ for all x. It satisfies* **strict overlap** *if $\epsilon < \mathrm{P}(A=1|x) < 1-\epsilon$ for all x and some $\epsilon > 0$.*

Overlap is the requirement that any unit could have either recieved the treatment or not.

To see the necessity of overlap, consider estimating the effectiveness of a drug in a study where patient sex is a confounder, but the drug was only ever prescribed to male patients. Then, conditional on a patient being female, we would know that patient was assigned to control. Without further assumptions, it's impossible to know the effect of the drug on a population with female patients, because there would be no data to inform the expected outcome for treated female patients, that is, $\mathbb{E}[Y \mid A=1, X=\texttt{female}]$. In this case, the statistical estimand equation 36.19 would not be identifiable. In the same vein, strict overlap ensures that the conditional distributions at each stratum of X can be estimated in finite samples.

Overlap can be particularly limiting in settings where we are adjusting for a large number of covariates (in an effort to satisfy no unobserved confounding). Then, certain combinations of traits may be very highly predictive of treatment assignment, even if individual traits are not. E.g., male patients over age 70 with BMI greater than 25 are very rarely assigned the drug. If such groups represent a significant fraction of the target population, or have significantly different treatment effects, then this issue can be problematic. In this case, the strict overlap assumption puts very strong restrictions on observational studies: for an observational study to satisfy overlap, most dimensions of the confounders X would need to closely mimic the balance we would expect in an RCT [D'A+21].

36.4.2 ATE Estimation with Observed Confounders

We now return to estimating the ATE using observed—i.e., not experimental—data. We've shown that in the case where we observe all common causes of the treatment and outcome, the ATE is

Author: Kevin P. Murphy. (C) MIT Press. CC-BY-NC-ND license

causally identified with a statistical estimand τ. We now consider several strategies for estimating this quantity using a finite data sample. Broadly, these techniques are known as backdoor adjustment.[2]

Recall that the defining characteristic of a confounding variable is that it affects both treatment and outcome. Thus, an adjustment strategy may aim to account for the influence of confounders on the observed outcome, the influence of confounders on treatment, or both. We discuss each of these strategies in turn.

36.4.2.1 Outcome Model Adjustment

We begin with an approach to covariate adjustment that relies on modeling the conditional expectation of the outcome Y given treatment A and confounders X. This strategy is often referred to as g-computation or outcome adjustment.[3] To begin, we define

Definition 2. *The* **conditional expected outcome** *is the function Q given by*

$$Q(a, x) = \mathbb{E}[Y|A = a, X = x]. \tag{36.25}$$

Substituting this definition into the definition of our estimand τ, Equation (36.19), we have $\tau = \mathbb{E}[Q(1, x) - Q(0, x)]$. This suggests a procedure for estimating τ: fit a model $\hat{Q}$ for Q and then report

$$\hat{\tau}^Q \triangleq \frac{1}{n} \sum_i \hat{Q}(1, x_i) - \hat{Q}(0, x_i). \tag{36.26}$$

To fit $\hat{Q}$, recall that $E[Y|a, x] = \text{argmin}_Q \mathbb{E}[(Y - Q(A, X)^2]$. That is, the minimizer (among all functions) of the squared loss risk is the conditional expected outcome.[4] So, to approximate Q, we simply use mean squared error to fit a predictor that predicts Y from A and X.

The estimation procedure takes several steps. We first fit a model $\hat{Q}$ to predict Y. Then, for each unit i, we predict that unit's outcome had they received treatment $\hat{Q}(1, x_i)$ and we predict their outcome had they not received treatment $\hat{Q}(0, x_i)$.[5] If the unit actually did receive treatment ($a_i = 1$) then $\hat{Q}(0, x_i)$ is our guess about what would have happened in the counterfactual case that they did not. The estimated expected gain from treatment for this individual is $\hat{Q}(1, x_i) - \hat{Q}(0, x_i)$—the difference in expected outcome between being treated and not treated. Finally, we estimate the outer expectation with respect to $\mathrm{P}(X)$—the true population distribution of the confounders—using the empirical distribution $\hat{\mathrm{P}}(X) = 1/n \sum_i \delta_{x_i}$. In effect, this means we substitute the expectation (over an unknown distribution) by an average over the observed data.

Linear regression It's worth saying something more about the special case where Q is modeled as a linear function of both the treatment and all the covariates. That is, the case where we assume the identification conditions of Theorem 2 and we additionally assume that the true, causal law

2. As we discuss in Section 36.8, this backdoor adjustment references the estimand returned by the do-calculus to eliminate confounding from a backdoor path. This also generalizes the approaches discussed here to some cases where we do not observe all common causes.

3. The "g" stands for generalized, for now-inscrutable historical reasons [Rob86].

4. To be precise, this definition applies when X and Y are square-integrable, and the minimzation taken over measurable functions.

5. this interpretation is justified by the same conditions as Theorem 2

(the SCM) governing Y yields: $Q(A, X) = \mathbb{E}[Y|A, X] = \mathbb{E}[f_Y(A, X, \xi)|A, X] = \beta_0 + \beta_A A + \beta_X X$. Plugging in, we see that $Q(1, X) - Q(0, X) = \beta_A$ (and so also $\tau = \beta_A$). Then, the estimator for the average treatment effect reduces to the estimator for the regression coefficient β_A. This "fit linear regression and report the regression coefficient" remains a common way of estimating the association between two variables in practice. The expected-outcome-adjustment procedure here may be viewed as a generalization of this procedure that removes the linear parametric assumption.

36.4.2.2 Propensity Score Adjustment

Outcome model adjustment relies on modeling the relationship between the confounders and the outcome. A popular alternative is to model the relationship between the confounders and the treatment. This strategy adjusts for confounding by directly addressing sampling bias in the treated and control groups. This bias arises from the relationship between the confounders and the treatment. Intuitively, the effect of confounding may be viewed as due to the difference between $\mathrm{P}(X|A = 1)$ and $\mathrm{P}(X|A = 0)$—e.g., the population of people who rode horses as children is wealthier than the population of people who did not. When we observe all confounding variables X, this degree of over- or under-representation can be adjusted away by reweighting samples such that the confounders X have the same distribution in the treated and control groups. When the confounders are balanced between the two groups, then any differences between them must be attributable to the treatment.

A key quantity for balancing treatment and control groups is the **propensity score**, which summarises the relationship between confounders and treatment.

Definition 3. *The* **propensity score** *is the function g given by $g(x) = \mathrm{P}(A = 1|X = x)$.*

To make use of the propensity score in adjustment, we first rewrite the estimand τ in a suggestive form, leveraging the fact that $A \in \{0, 1\}$:

$$\tau = \mathbb{E}[\frac{YA}{g(X)} - \frac{Y(1-A)}{1-g(X)}]. \tag{36.27}$$

This identity can be verified by noting that $\mathbb{E}[YA|X] = \mathbb{E}[Y|A = 1, X]\mathrm{P}(A = 1|X) + 0$, rearranging for $\mathbb{E}[Y|A = 1, X]$, doing the same for $\mathbb{E}[Y|A = 0, X]$, and substituting in to Equation (36.19). Note that the identity is just a mathematical fact about the statistical estimand—it does not rely on any causal assumptions, and holds whether or not τ can be interpreted as a causal effect.

This expression suggests the **inverse probability of treatment weighted estimator**, or IPTW estimator:

$$\hat{\tau}^{\mathrm{IPTW}} \triangleq \frac{1}{n}\sum_i \frac{Y_i A_i}{\hat{g}(X_i)} - \frac{Y_i(1-A_i)}{1-\hat{g}(X_i)}. \tag{36.28}$$

Here, $\hat{g}$ is an estimate of the propensity score function. Recall from Section 14.2.1 that if a model is well-specified and the loss function is a proper scoring rule then risk minimizer $g^* = \mathrm{argmin}_g \mathbb{E}[L(A, g(X))]$ will be $g^*(X) = \mathrm{P}(A = 1|X)$. That is, we can estimate the propensity score by fitting a model that predicts A from X. Cross-entropy and squared loss are both proper scoring rules, so we may use standard supervised learning methods.

In summary, the procedure is to estimate the propensity score function (with machine learning), and then to plug the estimated propensity scores $\hat{g}(x_i)$ into Equation (36.28). The IPTW estimator

Author: Kevin P. Murphy. (C) MIT Press. CC-BY-NC-ND license

computes a difference of weighted averages between the treated and untreated group. The effect is to upweight the outcomes of units who were unlikely to be treated but who nevertheless actually, by chance, recieved treatment (and similarly for untreated). Intuitively, such units are typical for the untreated population. So, their outcomes under treatment are informative about what would have happened had a typical untreated unit received treatment.

A word of warning is in order. Although the IPTW is asymtotically valid and popular in practice, it can be very unstable in finite samples. If estimated propensity scores are extreme for some values of x (that is, very close to 0 or 1), then the corresponding IPTW weights can be very large, resulting in a high-variance estimator. In some cases, this instability can be mitigated by instead using the Hajek version of the estimator.

$$\hat{\tau}^{\text{h-IPTW}} \triangleq \sum_i Y_i A_i \frac{1/\hat{g}(X_i)}{\sum_i A_i/\hat{g}(X_i)} - \sum_i Y_i(1-A_i)\frac{1/(1-\hat{g}(X_i))}{\sum_i (1-A_i)/(1-\hat{g}(X_i))}. \tag{36.29}$$

However, extreme weights can persist even after self-normalization, either because there are truly strata of X where treatment assignment is highly imbalanced, or because the propensity score estimation method has overfit. In such cases, it is common to apply heuristics such as weight clipping.

See Khan and Ugander [KU21] for a longer discussion of inverse-propensity type estimators, including some practical improvements.

36.4.2.3 Double Machine Learning

We have seen how to estimate the average treatment effect using either the relationship between confounders and outcome, or the relationship between confounders and treatment. In each case, we follow a two step estimation procedure. First, we fit models for the expected outcome or the propensity score. Second, we plug these fitted models into a downstream estimator of the effect.

Unsurprisingly, the quality of the estimate of τ depends on the quality of the estimates $\hat{Q}$ or $\hat{g}$. This is problematic because Q and g may be complex functions that require large numbers of samples to estimate. Even though we're only interested in the 1-dimensional parameter τ, the naive estimators described thus far can have very slow rates of convergence. This leads to unreliable inference or very large confidence intervals.

Remarkably, there are strategies for combining Q and g in estimators that, in principle, do better than using either Q or g alone. The **Augmented Inverse Probability of Treatment Weighted Estimator** (**AIPTW**) is one such estimator. It is defined as

$$\hat{\tau}^{\text{AIPTW}} \triangleq \frac{1}{n}\sum_i \hat{Q}(1,X_i) - \hat{Q}(0,X_i) + A_i\frac{Y_i - \hat{Q}(1,X_i)}{\hat{g}(x_i)} - (1-A_i)\frac{Y_i - \hat{Q}(0,X_i)}{1-\hat{g}(X_i)}. \tag{36.30}$$

That is, $\hat{\tau}^{\text{AIPTW}}$ is the outcome adjustment estimator plus a stabilization term that depends on the propensity score. This estimator is a particular case of a broader class of estimators that are refered to as **semi-parametrically efficient** or **double machine-learning** estimators [Che+17e; Che+17d]. We'll use the later terminology here.

We now turn to understanding the sense in which double machine learning estimators are robust to misestimation of the **nuisance functions** Q and g. To this end, we define the **influence curve**

of τ to be the function ϕ defined by[6]

$$\phi(X_i, A_i, Y_i; Q, g, \tau) \triangleq Q(1, X_i) - Q(0, X_i) + A_i \frac{Y_i - Q(1, X_i)}{g(x_i)} - (1 - A_i) \frac{Y_i - Q(0, X_i)}{1 - g(X_i)} - \tau. \tag{36.31}$$

By design, $\hat{\tau}^{\text{AIPTW}} - \tau = \frac{1}{n} \sum_i \phi(\mathbf{X}_i; \hat{Q}, \hat{g}, \tau)$. We begin by considering what would happen if we simply knew Q and g, and didn't have to estimate them. In this case, the estimator would be $\hat{\tau}^{\text{ideal}} = \frac{1}{n} \sum_i \phi(\mathbf{X}_i; Q, g, \tau)$ and, by the central limit theorem, we would have:

$$\sqrt{n}(\hat{\tau}^{\text{ideal}} - \tau) \xrightarrow{d} \text{Normal}(0, \mathbb{E}[\phi(\mathbf{X}_i; Q, g, \tau)^2]). \tag{36.32}$$

This result characterizes the estimation uncertainity in the best possible case. If we knew Q and g, we could rely on this result for, e.g., finding confidence intervals for our estimate.

The question is: what happens when Q and g need to be estimated? For general estimators and nuisance function models, we don't expect the $\sqrt{n}$-rate of Equation (36.32) to hold. For instance, $\sqrt{n}(\hat{\tau}^{Q} - \tau)$ only converges if $\sqrt{n}\mathbb{E}[(\hat{Q} - Q)^2]^{\frac{1}{2}} \to 0$. That is, for the naive estimator we only get the $\sqrt{n}$ rate for estimating τ if we can also estimate Q at the $\sqrt{n}$ rate—a much harder task! This is the issue that the double machine learning estimator helps with.

To understand how, we decompose the error in estimating τ as follows:

$$\sqrt{n}(\hat{\tau}^{\text{AIPTW}} - \tau) \tag{36.33}$$

$$= \frac{1}{\sqrt{n}} \sum_i \phi(\mathbf{X}_i; Q, g, \tau) \tag{36.34}$$

$$+ \frac{1}{\sqrt{n}} \sum_i \phi(\mathbf{X}_i; \hat{Q}, \hat{g}, \tau) - \phi(\mathbf{X}_i; Q, g, \tau) - \mathbb{E}[\phi(\mathbf{X}; \hat{Q}, \hat{g}, \tau) - \phi(\mathbf{X}; Q, g, \tau)] \tag{36.35}$$

$$+ \sqrt{n}\mathbb{E}[\phi(\mathbf{X}; \hat{Q}, \hat{g}, \tau) - \phi(\mathbf{X}; Q, g, \tau)] \tag{36.36}$$

We recognize the first term, Equation (36.34), as $\sqrt{n}(\hat{\tau}^{\text{ideal}} - \tau)$, the estimation error in the optimal case where we know Q and g. Ideally, we'd like the error of $\hat{\tau}^{\text{AIPTW}}$ to be asymptotically equal to this ideal case—which will happen if the other two terms go to 0.

The second term, Equation (36.35), is a penalty we pay for using the same data to estimate Q, g and to compute τ. For many model classes, it can be shown that such "empirical process" terms go to 0. This can also be guaranteed in general by using different data for fitting the nuisance functions and for computing the estimator (see the next section).

The third term, Equation (36.36), captures the penalty we pay for misestimating the nuisance functions. This is where the particular form of the AIPTW is key. With a little algebra, we can show that

$$\mathbb{E}[\phi(\mathbf{X}; \hat{Q}, \hat{g}) - \phi(\mathbf{X}; Q, g)] = \mathbb{E}[\frac{1}{g(X)}(\hat{g}(X) - g(X))(\hat{Q}(1, X) - Q(1, X)) \tag{36.37}$$

$$+ \frac{1}{1 - g(X)}(\hat{g}(X) - g(X))(\hat{Q}(0, X) - Q(0, X))]. \tag{36.38}$$

6. Influence curves are the foundation of what follows, and the key to generalizing the analysis beyond the ATE. Unfortunately, going into the general mathematics would require a major digression, so we omit it. However, see references at the end of the chapter for some pointers to the relevant literature.

Author: Kevin P. Murphy. (C) MIT Press. CC-BY-NC-ND license

The important point is that estimation errors of Q and g are multiplied together. Using the Cauchy-Schwarz inequality, we find that $\sqrt{n}\mathbb{E}[\phi(\mathbf{X}; \hat{Q}, \hat{g}) - \phi(\mathbf{X}; Q, g)] \to 0$ as long as $\sqrt{n} \max_a \mathbb{E}[(\hat{Q}(a, X) - Q(a, X))^2]^{\frac{1}{2}} \mathbb{E}[(\hat{g}(X) - g(X))^2]^{\frac{1}{2}} \to 0$. That is, the misestimation penalty will vanish so long as the *product* of the misestimation errors is $o(\sqrt{n})$. For example, this means that that τ can be estimated at the (optimal) $\sqrt{n}$ rate even when the estimation error of each of Q and g only decreases as $o(n^{-1/4})$.

The upshot here is that the double machine learning estimator has the special property that the weak condition $\sqrt{n}\mathbb{E}(\hat{Q}(T, X) - Q(T, X))^2 \mathbb{E}(\hat{g}(X) - g(X))^2 \to 0$ suffices to imply that

$$\sqrt{n}(\hat{\tau}^{\text{AIPTW}} - \tau) \xrightarrow{d} \text{Normal}(0, \mathbb{E}[\phi(\mathbf{X}_i; Q, g, \tau)^2]) \tag{36.39}$$

(though strictly speaking this requires some additional technical conditions we haven't discussed). This is *not* true for the earlier estimators we discussed, which require a much faster rate of convergence for the nuisance function estimation.

The AIPTW estimator has two further nice properties that are worth mentioning. First, it is **non-parametrically efficient**. This means that this estimator has the smallest possible variance of any estimator that does not make parametric assumptions; namely, $\mathbb{E}[\phi(\mathbf{X}_i; Q, g, \tau)^2]$. This means, for example, that this estimator yields the smallest confidence intervals of any approach that does not rely on parametric assumptions. Second, it is **doubly robust**: the estimator is consistent (converges to the true τ as $n \to \infty$) as long as at least one of either $\hat{Q}$ or $\hat{g}$ is consistent.

36.4.2.4 Cross Fitting

The term Equation (36.35) in the error decomposition above is the penalty we pay for reusing the same data to both fit Q, g and to compute the estimator. For many choices of model for Q, g, this term goes to 0 quickly as n gets large and we achieve the (best case) $\sqrt{n}$ error rate. However, this property doesn't always hold.

As an alternative, we can always randomly split the available data and use one part for model fitting, and the other to compute the estimator. Effectively, this means the nuisance function estimation and estimator computation are done using independent samples. It can then be shown that the reuse penalty will vanish. However, this comes at the price of reducing the amount of data available for each of nuisance function estimation and estimator computation.

This strategy can be improved upon by a **cross fitting** approach. We divide the data into K folds. For each fold j we use the other $K - 1$ folds to fit the nuisance function models $\hat{Q}^{-j}, \hat{g}^{-j}$. Then, for each datapoint i in fold j, we take $\hat{Q}(a_i, x_i) = \hat{Q}^{-j}(a_i, x_i)$ and $\hat{g}(x_i) = \hat{g}^{-j}(x_i)$. That is, the estimated conditional outcomes and propensity score for each datapoint are predictions from a model that was not trained on that datapoint. Then, we estimate τ by plugging $\{\hat{Q}(a_i, x_i), \hat{g}(x_i)\}_i$ into Equation (36.30). It can be shown that this cross fitting procedure has the same asymptotic guarantee—the central limit theorem at the $\sqrt{n}$ rate—as described above.

36.4.3 Uncertainity Quantification

In addition to the point estimate $\hat{\tau}$ of the average treatment effect, we'd also like to report a measure of the uncertainity in our estimate. For example, in the form of a confidence interval. The asymptotic normality of $\sqrt{n}\hat{\tau}$ (Equation (36.39)) provides a means for this quantification. Namely, we could base confidence intervals and similar on the limiting variance $\mathbb{E}[\phi(\mathbf{X}_i; Q, g, \tau)^2]$. Of course, we don't actually know any of Q, g, or τ. However, it turns out that it suffices to estimate the asymptotic

variance with $\frac{1}{n}\sum_i \phi(\mathbf{X}_i; \hat{Q}, \hat{g}, \hat{\tau})^2$ [Che+17e]. That is, we can estimate the uncertainity by simply plugging in our fitted nuisance models and our point estimate of τ into

$$\hat{\mathbb{V}}[\hat{\tau}] = 1/n \sum_i \phi(\mathbf{X}_i; \hat{Q}, \hat{g}, \hat{\tau})^2. \tag{36.40}$$

This estimated variance can then be used to compute confidence intervals in the usual manner. E.g., we'd report a 95% confidence interval for τ as $\hat{\tau} \pm 1.96\sqrt{\hat{\mathbb{V}}[\hat{\tau}]/n}$.

Alternatively, we could quantify the uncertainity by bootstrapping. Note, however, that this would require refitting the nuisance functions with each bootstrap model. Depending on the model and data, this can be prohibitively computationally expensive.

36.4.4 Matching

One particularly popular approach to adjustment-based causal estimation is **matching**. Intuitively, the idea is to match each treated to unit to an untreated unit that has the same (or at least similar) values of the confounding variables and then compare the observed outcomes of the treated unit and its matched control. If we match on the full set of common causes, then the difference in outcomes is, intuitively, a noisy estimate of the effect the treatment had on that treated unit. We'll now build this up a bit more carefully. In the process we'll see that matching can be understood as, essentially, a particular kind of outcome model adjustment.

For simplicity, consider the case where X is a discrete random variable. Define $\mathcal{A}_x$ to be the set of treated units with covariate value x, and $\mathcal{C}_x$ to be the set of untreated units with covariate value x. In this case, the matching estimator is:

$$\hat{\tau}^{\text{matching}} = \sum_x \hat{\mathrm{P}}(x)\big(\frac{1}{|\mathcal{A}_x|}\sum_{i\in\mathcal{A}_x} Y_i - \frac{1}{|\mathcal{C}_x|}\sum_{j\in\mathcal{C}_x} Y_j\big), \tag{36.41}$$

where $\hat{\mathrm{P}}(x)$ is an estimator of $\mathrm{P}(X=x)$—e.g., the fraction of units with $X=x$. Now, we can rewrite $Y_i = Q(A_i, X_i) + \xi_i$ where ξ_i is a unit-specific noise term defined by the equation. In particular, we have that $\mathbb{E}[\xi_i|A_i, X_i] = 0$. Substituting this in, we have:

$$\hat{\tau}^{\text{matching}} = \sum_x \hat{\mathrm{P}}(x)\big(Q(1,x) - Q(0,x)\big) + \sum_x \frac{1}{|\mathcal{A}_x|}\sum_{i\in\mathcal{A}_x} \xi_i - \frac{1}{|\mathcal{C}_x|}\sum_{j\in\mathcal{C}_x} \xi_j. \tag{36.42}$$

We can recognize the first term as an estimator of usual target parameter τ (it will be equal to τ if $\hat{\mathrm{P}}(x) = \mathrm{P}(x)$). The second term is a difference of averages of random variables with expectation 0, and so each term will converge to 0 as long as $|\mathcal{A}_x|$ and $|\mathcal{C}_x|$ each go to infinity as we see more and more data. Thus, we see that the matching estimator is a particular way of estimating the parameter τ. The procedure can be extended to continuous covariates by introducing some notion of values of X being close, and then matching close treatment and control variables.

There are two points we should emphasize here. First, notice that the argument here has nothing to do with causal identification. Matching is a particular technique for estimating the observational parameter τ. Whether or not τ can be interpreted as an average treatment effect is determined by the conditions of Theorem 2—the particular estimation strategy doesn't say anything about this. Second, notice that in essence matching amounts to a particular choice of model for $\hat{Q}$. Namely,

Author: Kevin P. Murphy. (C) MIT Press. CC-BY-NC-ND license

$\hat{Q}(1,x) = \frac{1}{|\mathcal{A}_x|}\sum_{i\in\mathcal{A}_x} Y_i$ and similarly for $\hat{Q}(0,x)$. That is, we estimate the conditional expected outcome as a sample mean over units with the same covariate value. Whether this is a good idea depends on the quality of our model for Q. In situations where better models are possible (e.g., a machine-learning model fits the data well), we might expect to get a more accurate estimate by using the conditional expected outcome predictor directly.

There is another important case we mention in passing. In general, when using adjustment based identification, it suffices to adjust for any function $\phi(X)$ of X such that $A \perp\!\!\!\perp X|\phi(X)$. To see that adjusting for only $\phi(X)$ suffices, first notice that $g(X) = \mathrm{P}(A=1|X) = \mathrm{P}(A=1|\phi(X))$ only depends on $\phi(X)$, and then recall that can write the target parameter as $\tau = \mathbb{E}[\frac{YA}{g(X)} - \frac{Y(1-A)}{1-g(X)}]$, whence τ only depends on X through $g(X)$. That is: replacing X by a reduced version $\phi(X)$ such that $g(X) = \mathrm{P}(A=1|\phi(X))$ can't make any difference to τ. Indeed, the most popular choice of $\phi(X)$ is the propensity score itself, $\phi(X) = g(X)$. This leads to **propensity score matching**, a two step procedure where we first fit a model for the propensity score, and then run matching based on the estimated propensity score values for each unit. Again, this is just a particular estimation procedure for the observational parameter τ, and says nothing about whether it's valid to interpret τ as a causal effect.

36.4.5 Practical Considerations and Procedures

when performing causal analysis, many issues can arise in practice, some of which we discuss below.

36.4.5.1 What to adjust for

Choosing which variables to adjust for is a key detail in estimating causal effects using covariate adjustment. The criterion is clear when one has a full causal graph relating A, Y, and all covariates X to each other. Namely, adjust for all variables that are actually causal parents of A and Y. In fact, with access to the full graph, this criterion can be generalized somewhat—see Section 36.8.

In practice, we often don't actually know the full causal graph relating all of our variables. As a result, it is common to apply simple heuristics to determine which variables to adjust for. Unfortunately, these heuristics have serious limitations. However, exploring these is instructive.

A key condition in Theorem 2 is that the covariates X that we adjust for must include all the common causes. In the absence of a full causal graph, it is tempting to condition on as many observed variables as possible to try to ensure this condition holds. However, this can be problematic. For instance, suppose that M is a mediator of the effect of A on Y—i.e., M lies on one of the directed paths between A and Y. Then, conditioning on M will block this path, removing some of the causal effect. Note that this does not always result in an attenuated, or smaller-magnitude, effect estimate. The effect through a given mediator may run in the opposite direction of other causal pathways from the treatment; thus conditioning on a mediator can inflate or even flip the sign of a treatment effect. Alternatively, if C is a collider between A and Y—a variable that is caused by both—then conditioning on C will induce an extra statistical dependency between A and Y.

Both pitfalls of the "condition on everything" heuristic discussed above both involve conditioning on variables that are downstream of the treatment A. A natural response is to this is to limit conditioning to all pre-treatment variables, or those that are causally upstream of the treatment. Importantly, if there is a valid adjustment set in the observed covariates X, then there will also be a valid adjustment set among the pre-treatment covariates. This is because any open backdoor path

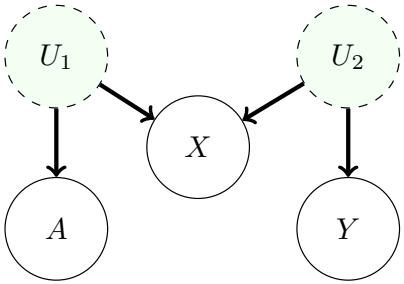


Figure 36.4: The M-bias causal graph. Here, A and Y are not confounded. However, conditioning on the covariate X opens a backdoor path, passing through U_1 and U_2 (because X is a colider). Thus, adjusting for X creates bias. This is true even though X need not be a pre-treatment variable.

between A and Y must include a parent of A, and the set of pre-treatment covariates includes these parents. However, it is still possible that conditioning on the full set of pre-treatment variables can induce new backdoor paths between A and Y through colliders. In particular, if there is a covariate D that is separately confounded with the treatment A and the outcome Y then D is a collider, and conditioning on D opens a new backdoor path. This phenomenon is known as m-bias because of the shape of the graph [Pea09c], see Figure 36.4.

A practical refinement of the pre-treatment variable heuristic is given in VanderWeele TJ [VT11]. Their heuristic suggests conditioning on all pre-treatment variables that are causes of the treatment, outcome, or both. The essential qualifier in this heuristic is that the variable is causally upstream of treatment and/or outcome. This eliminates the possibility of conditioning on covariates that are only confounded with treatment and outcome, avoiding m-bias. Notably, this heuristic requires more causal knowledge than the above heuristics, but does not require detailed knowledge of how different covariates are causally related to each other.

The VanderWeele TJ [VT11] criterion is a useful rule of thumb, but other practical considerations often arise. For example, if one has more knowledge about the causal structure among covariates, it is possible to optimize adjustment sets to minimize the variance of the resulting estimator [RS20]. One important example of reducing variance by pruning adjustment sets is the exclusion of variables that are known to only be a parent of the treatment, and not of the outcome (so called instruments, as discussed in Section 36.5).

Finally, adjustment set selection criteria operate under the assumption that there actually exists a valid adjustment set among observed covariates. When there is no set of observed covariates in X that block all backdoor paths, then any adjusted estimate will be biased. Importantly, in this case, the bias does not necessarily decrease as one conditions on more variables. For example, conditioning on an instrumental variable often results in an estimate that has higher bias, in addition to the higher variance discussed above. This phenomenon is known as bias amplification or z-bias; see Section 36.7.2. A general rule of thumb is that variables that explain away much more variation in the treatment than in the outcome can potentially amplify bias, and should be treated with caution.

Author: Kevin P. Murphy. (C) MIT Press. CC-BY-NC-ND license

36.4.5.2 Overlap

Recall that in addition to no-unobserved-confounders, identification of the average treatment effect requires overlap: the condition that $0 < P(A = 1|x) < 1$ for the population distribution P. With infinite data, any amount of overlap will suffice for estimating the causal effect. In realistic settings, even near failures can be problematic. Equation (36.39) gives an expression for the (asymptotic) variance of our estimate: $\mathbb{E}[\phi(\mathbf{X}_i; \hat{Q}, \hat{g}, \hat{\tau})^2]/n$. Notice that $\phi(\mathbf{X}_i; \hat{Q}, \hat{g}, \hat{\tau})^2$ involves terms that are proportional to $1/g(X)$ and $1/(1 - g(X)$. Accordingly, the variance of our estimator will balloon if there are units where $g(x) \approx 0$ or $g(x) \approx 1$ (unless such units are rare enough that they don't contribute much to the expectation).

In practice, a simple way to deal with potential overlap violation is to fit a model $\hat{g}$ for the treatment assignment probability—which we need to do anyways—and check that the values $\hat{g}(x)$ are not too extreme. In the case that some values are too extreme, the simplest resolution is to cheat. We can simply exclude all the data with extreme values of $\hat{g}(x)$. This is equivalent to considering the average treatment effect over only the subpopulation where overlap is satisfied. This changes the interpretation of the estimand. The restricted subpopulation ATE may or may not provide a satisfactory answer to the real-world problem at hand, and this needs to be justified based on knowledge of the real-world problem.

36.4.5.3 Choice of Estimand and Average Treatment Effect on the Treated

Usually, our goal in estimating a causal effect is qualitative. We want to know what the sign of the effect is, and whether it's large or small. The utility of the ATE is that it provides a concrete query we can use to get a handle on the qualitative question. However, it is not sacrosanct; sometimes we're better off choosing an alternative causal estimand that still answers the qualitative question but which is easier to estimate statistically. The **average treatment effect on the treated** or **ATT**,

$$\text{ATT} \triangleq \mathbb{E}_{X|A=1}[\mathbb{E}[Y|X, \text{do}(A = 1)] - E[Y|X, \text{do}(A = 0)]], \tag{36.43}$$

is one such an estimand that is frequently useful.

The ATT is useful when many members of the population are very unlikely to receive treatment, but the treated units had a reasonably high probability of receiving the control. This can happen if, e.g., we sample control units from the general population, but the treatment units all self-selected into treatment from a smaller subpopulation. In this case, it's not possible to (non-parametrically) determine the treatment effect for the control units where no similar unit took treatment. The ATT solves this obstacle by simply omitting such units from the average.

If we have the causal structure Figure 36.3, and the overlap condition $\text{P}(A = 1|X = x) < 1$ for all $X = x$ then the ATT is causally identified as

$$\tau^{\text{ATT}} = \mathbb{E}_{X|A=1}[\mathbb{E}[Y|A = 1, X] - E[Y|A = 0, X]]. \tag{36.44}$$

Note that the required overlap condition here is weaker than for identifying the ATE. (The proof is the same as Theorem 2.)

The estimation strategies for the ATE translate readily to estimation strategies for the ATT. Namely, estimate the nuisance functions the same way and then simply replace averages over all data points by averages over the treated datapoints only. In principle, it's possible to do a little

better than this by making use of the untreated datapoints as well. A corresponding double machine learning estimator is

$$\hat{\tau}^{\text{ATT}-\text{AIPTW}} \triangleq \frac{1}{n}\sum_i \frac{A_i}{\text{P}(A=1)}(Y-\hat{Q}(0,X_i)) - \frac{(1-A_i)g(X)}{\text{P}(A=1)(1-g(X))}(Y-\hat{Q}(0,X_i). \tag{36.45}$$

. The variance of this estimator can be estimated by

$$\begin{aligned}\phi^{\text{ATT}}(\mathbf{X}_i;Q,g,\tau) \triangleq \frac{1}{n}\sum_i \Bigg[&\frac{A_i}{\text{P}(A=1)}(Y-\hat{Q}(0,X_i)) \\ &- \frac{(1-A_i)g(X)}{\text{P}(A=1)(1-g(X))}(Y-\hat{Q}(0,X_i) - \frac{A\tau}{\text{P}(A=1)}\Bigg]\end{aligned} \tag{36.46}$$

$$\hat{\mathbb{V}}[\hat{\tau}^{\text{ATT}-\text{AIPTW}}] \triangleq \frac{1}{n}\sum_i \phi^{\text{ATT}}(\mathbf{X}_i;\hat{Q},\hat{g},\hat{\tau}^{\text{ATT}-\text{AIPTW}}). \tag{36.47}$$

Notice that the estimator for the ATT doesn't require estimating $Q(1,X)$. This can be a considerable advantage when the treated units are rare.

See Chernozhukov et al. [Che+17e] for details.

36.4.6 Summary and Practical Advice

We have seen a number of estimators that follow the general procedure:

1. Fit statistical or machine-learning models $\hat{Q}(a,x)$ as a predictor for Y, and/or $\hat{g}(x)$ as a predictor for A
2. Compute the predictions $\hat{Q}(0,x_i), \hat{Q}(1,x_i), \hat{g}(x_i)$ for each data point, and
3. Combine these predictions into an estimate of the average treatment effect.

Importantly, no single estimation approach is a silver bullet. For example, the double machine-learning estimator has appealing theoretical properties, such as asymptotic efficiency guarantees and a recipe for estimating uncertainity without needing to bootstrap the model fitting. However, in terms of the quality of point estimates, the double ML estimators can sometimes underperform their more naive counterparts [KS07]. In fact, there are cases where each of outcome regression, propensity weighting, or doubly robust methods will outperform the others.

One difficulty in choosing an estimator in practice is that there are fewer guardrails in causal inference than there are in standard predictive modeling. In predictive modeling, we construct a train-test split and validate our prediction models using the true labels or outcomes in the held-out dataset. However, for causal problems, the causal estimands are functionals of a different data-generating process from the one that we actually observed. As a result, it is impossible to empirically validate many aspects of causal estimation using standard techniques.

The effectiveness of a given approach is often determined by how much we trust the specification of our propensity score or outcome regression models $\hat{g}(x)$ and $\hat{Q}(a,x)$, and how well the treatment and control groups overlap in the dataset. Using flexible models for the nuisance functions g and Q can alleviate some of the concerns about model misspecification, but our freedom to use such models is

often constrained by dataset size. When we have the luxury of large data, we can use flexible models; on the other hand, when the dataset is relatively small, we may need to use a smaller parametric family or stringent regularization to obtain stable estimates of Q and g. Similarly, if overlap is poor in some regions of the covariate space, then flexible models for Q may be highly variable, and inverse propensity score weights may be large. In these cases, IPTW or AIPTW estimates may fluctuate wildly as a function of large weights. Meanwhile, outcome regression estimates will be sensitive to the specification of the Q model and its regularization, and can incur bias that is difficult to measure if the specification or regularization does not match the true outcome process.

There are a number of practical steps that we can take to sanity-check causal estimates. The simplest check is to compute many different ATE estimators (e.g., outcome regression, IPTW, doubly robust) using several comparably complex estimators of Q and g. We can then check whether they agree, at least qualitatively. If they do agree then this can provide some peace of mind (although it is not a guarantee of accuracy). If they disagree, caution is warranted, particularly in choosing the specification of the Q and g models.

It is also important to check for failures of overlap. Often, issues such as disagreement between alternative estimators can be traced back to poor overlap. A common way to do this, particularly with high-dimensional data, is to examine the estimated (ideally cross-fitted) propensity scores $\hat{g}(x_i)$. This is a useful diagnostic, even if the intention is to use an outcome regression approach that only incorporates and estimated outcome regression function $\hat{Q}(a, x_i)$. If overlap issues are relevant, it may be better to instead estimate either the average treatment effect on the treated, or the "trimmed" estimand given by discarding units with extreme propensities.

Uncertainty quantification is also an essential part of most causal analyses. This frequently take the form of an estimate of the estimator's variance, or a confidence interval. This may be important for downstream decision-making, and can also be a useful diagnostic. We can calculate variance either by bootstrapping the entire procedure (including refitting the models in each bootstrap replicate), or computing analytical variance estimates from the AIPTW estimator. Generally, large variance estimates may indicate issues with the analysis. For example, poor overlap will often (although not always) manifest as extremely large variances under either of these methods. Small variance estimates should be treated with caution, unless other checks, such as overlap checks, or stability across different Q and g models, also pass.

The previous advice only addresses the statistical problem of estimating τ from a data sample. It does not speak to whether or not τ can reasonably be interpreted as an average treatment effect. Considerable care should be devoted to whether or not the assumption that there are no unobserved confounders is reasonable. There are several methods for assessing the sensitivity of the ATE estimate to violations of this assumption. See Section 36.7. Bias due to unobserved confounding can be substantial in practice—often overwhelming bias due to estimation error—so it is wise to conduct such an analysis.

36.5 Instrumental Variable Strategies

Adjustment-based methods rely on observing all confounders affecting the treatment and outcome. In some situations, it is possible to identify interesting causal effects even when there are unobserved confounders. We now consider strategies based on **instrumental variables**. The instrumental variable graph is shown in Figure 36.5. The key ingredient is the instrumental variable Z, a variable

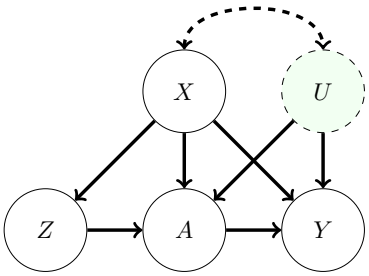


Figure 36.5: Causal graph illustrating the Instrumental Variable setup. The treatment A and outcome Y are both influenced by unobserved confounder U. Nevertheless, identification is sometimes possible due to the presence of the instrument Z. We also allow for observed covariates X that we may need to adjust for. The dashed arrow between U and X indicates a statistical dependency where we remain agnostic to the particular causal relationship.

that has a causal effect on Y only through its causal effect on A. Informally, the identification strategy is to determine to the causal effect of Z on Y, the causal effect of Z on A, and then combine these into an estimate of the causal effect of A on Y.

For this identification to strategy to work the instrument must satisfy three conditions. There are observed variables (confounders) X such that:

1. **Instrument Relevance** $Z \not\perp\!\!\!\perp A|X$: the instrument must actually affect the treatment assignment.
2. **Instrument Unconfoundedness** Any backdoor path between Z and Y is blocked by X, even conditional on A.
3. **Exclusion Restriction** All directed paths from Z to Y pass through A. That is, the instrument affects the outcome *only* through its effect on A.

(It may help conceptually to first think through the case where X is the empty set—i.e., where the only confounder is the unobserved U). These assumptions are necessary for using instrumental variables for causal identification, but they are not quite sufficient. In practice, they must be supplemented by an additional assumption that depends more closely on the details of the problem at hand. Historically, this additional assumption was usually that both the instrument-treatment and treatment-outcome relationships are linear. We'll examine some less restrictive alternatives below.

Before moving on to how to use instrumental variables for identification, let's consider how we might encounter instruments in practice. The key is that its often possible to find, and measure, variables that affect treatment and that are assigned (as if) at random. For example, suppose we are interested in measuring the effect of taking a drug A on some health outcome Y. The challenge is that whether a study participant actually takes the drug can be confounded with Y—e.g., sicker people may be more likely to take their medication, but have worse outcomes. However, the assignment of treatments to patients can be randomized and this random assignment can be viewed as an instrument. This **random assignment with non-compliance** scenario is common in practice. The random assignment—the instrument—satisfies relevance (so long as assigning the drug affects the probability of the patient taking the drug). It also satisfies unconfoundedness (because the

instrument is randomized). And, it plausibly satisfies exclusion restriction: telling (or not telling) a patient to take a drug has no effect on their health outcome except through influencing whether or not they actually take the drug. As a second example, the **judge fixed effects** research design uses the identity of the judge assigned to each criminal case to infer the effect of incarceration on some life outcome of interest (e.g., total lifetime earnings). Relevance will be satisfied so long as different judges have different propensities to hand out severe sentences. The assignment of trial judges to cases is randomized, so unconfoundedness will also be satisfied. And, exclusion restriction is also plausible: the particular identity of the judge assigned to your case has no bearing on your years-later life outcomes, except through the particular sentence that you're subjected to.

It's important to note that these assumptions require some care, particularly exclusion restriction. Relevance can be checked directly from the data, by fitting a model to predict the treatment from the instrument (or vice versa). Unconfoundedness is often satisfied by design: the instrument is randomly assigned. Even when literal random assignment doesn't hold, we often restrict to instruments where unconfoundedness is "obviously" satisfied—e.g., using number of rainy days in a month as an instrument for sun exposure. Exclusion restriction is trickier. For example, it might fail in the drug assignment case if patients who are not told to take a drug respond by seeking out alternative treatment. Or, it might fail in the judge fixed effects case if judges hand out additional, unrecorded, punishments in addition to incarceration. Assessing the plausibility of exclusion restriction requires careful consideration based on domain expertise.

We now return to the question of how to make use of an instrument once we have it in hand. As previously mentioned, getting causal identification using instrumental variables requires supplementing the IV assumptions with some additional assumption about the causal process.

36.5.1 Additive Unobserved Confounding

We first consider **additive unobserved confounding**. That is, we assume that the structural caual model for the outcome has the form:[7]

$$Y \leftarrow f(A, X) + f_U(U). \tag{36.48}$$

In words, we assume that there are no interaction effects between the treatment and the unobserved confounder—everyone responds to treatment in the same way. With this additional assumption, we see that $\mathbb{E}[Y|X, \text{do}(A = a)] - \mathbb{E}[Y|X, \text{do}(A = a')] = f(a, X) - f(a', X)$. In this setting, our goal is to learn this contrast.

Theorem 3 (Additive Confounding Identification)**.** *If the instrumental variables assumptions hold and also additive unobserved confounding holds, then there is a function $\tilde{f}(a, x)$ where*

$$\mathbb{E}[Y|x, \text{do}(A = a)] - \mathbb{E}[Y|x, \text{do}(A = a')] = \tilde{f}(a, x) - \tilde{f}(a', x), \tag{36.49}$$

for all x, a, a' and such that $\tilde{f}$ satisfies

$$\mathbb{E}[Y|z, x] = \int \tilde{f}(a, x) p(a|z, x) \mathrm{d}a. \tag{36.50}$$

Here, $p(a|z, x)$ is the conditional probability density of treatment.

7. We roll the unit-specific variables ξ into U to avoid notational overload.

In particular, if there is a unique function g that satisfies

$$\mathbb{E}[Y|z,x] = \int g(a,x)p(a|z,x)\mathrm{d}a, \tag{36.51}$$

then $g = \tilde{f}$ and this relation identifies the target causal effect.

Before giving the proof, lets understand the point of this identification result. The key insight is that both the left hand side of Equation (36.51) and $p(a|z,x)$ (appearing in the integrand) are identified by the data, since they involve only observational relationships between observed variables. So, $\tilde{f}$ is identified implicitly as one of the functions that makes Equation (36.51) true. If there is a unique such function, then this fully identifies the causal effect.

Proof. With the additive unobserved confounding assumption, the instrument unconfoundedness implies that $U \perp\!\!\!\perp Z|X$. Then, we have that:

$$\mathbb{E}[Y|Z,X] = \mathbb{E}[f(A,X)|Z,X] + \mathbb{E}[f_U(U)|Z,X] \tag{36.52}$$
$$= \mathbb{E}[f(A,X)|Z,X] + \mathbb{E}[f_U(U)|X] \tag{36.53}$$
$$= \mathbb{E}[\tilde{f}(A,X)|Z,X], \tag{36.54}$$

where $\tilde{f} = f(A,X) + \mathbb{E}[f_U(U)|X]$. Now, identifying just $\tilde{f}$ would suffice for us, because we could then identify contrasts between treatements: $f(a,x) - f(a',x) = \tilde{f}(a,x) - \tilde{f}(a',x)$. (The term $\mathbb{E}[f_U(U)|x]$ cancels out). Accordingly, we rewrite Equation (36.54) as:

$$\mathbb{E}[Y|z,x] = \int \tilde{f}(a,x)p(a|z,x)\mathrm{d}a. \tag{36.55}$$

□

It's worth dwelling briefly on how the IV assumptions come into play here. The exclusion restriction is implied by the additive unobserved confounding assumption, which we use explicilty. We also use the unconfoundedness assumption to conclude $U \perp\!\!\!\perp Z|X$. However, we do not use relevance. The role of relevance here is in ensuring that few functions solve the relation Equation (36.51). Informally, the solution g is constrained by the requirement that it hold for all values of Z. However, different values of Z only add non-trivial constraints if $p(a|z,x)$ differ depending on the value of z—this is exactly the relevance condition.

Estimation The basic estimation strategy is to fit models for $\mathbb{E}[Y|z,x]$ and $p(a|z,x)$ from the data, and then solve the implicit equation Equation (36.51) to find g consistent with the fitted models. The procedures for doing this can vary considerably depending on the particulars of the data (e.g., if Z is discrete or continuous) and the choice of modeling strategy. We omit a detailed discussion, but [see e.g. NP03; Dar+11; Har+17; SSG19; BKS19; Mua+20; Dik+20] for various concrete approaches.

It's also worth mentioning an additional nuance to the general procedure. Even if relevance holds, there will often be more than one function that satisfies Equation (36.51). So, we have only identified $\tilde{f}$ as a member of this set of functions. In practice, this ambiguity is defeated by making some additional structural assumption about $\tilde{f}$. For example, we model $\tilde{f}$ with a neural network, and then choose the network satisfying Equation (36.51) that has minimum $l2$-norm on the parameters (i.e., we pick the $l2$-regularized solution).

Author: Kevin P. Murphy. (C) MIT Press. CC-BY-NC-ND license

36.5.2 Instrument Monotonicity and Local Average Treatment Effect

We now consider an alternative assumption to additive unobserved confounding that is applicable when both the instrument and treatment are binary. It will be convenient to conceptualize the instrument as assignment-to-treatment. Then, the population divides into four subpopulations:

1. Compliers, who take the treatment if assigned to it, and who don't take the treatment otherwise.
2. Always takers, who take the treatment no matter their assignment
3. Never takers, who refuse the treatment no matter their assignment
4. Defiers, who refuse the treatment if assigned to it, and who take the treatment if not assigned.

Our goal in this setting will be to identify the average treatment effect among the compliers. The **local average treatment effect** (or **complier average treatment effect**) is defined to be[8]

$$\text{LATE} = \mathbb{E}[Y|\text{do}(A=1), \text{complier}] - \mathbb{E}[Y|\text{do}(A=0), \text{complier}]. \tag{36.56}$$

The LATE requires an additional assumption for identification. Namely, **instrument monotonicity**: being assigned (not assigned) the treatment only increases (decreases) the probability that each unit will take the treatment. Equivalently, $\text{P}(\text{defier}) = 0$.

We can then write down the identification result.

Theorem 4. *Given the instrumental variable assumptions and instrument monotonicty, the local average treatment is identified as a parameter τ^{LATE} of the observational distributional; that is,* $\text{LATE} = \tau^{\text{LATE}}$. *Namely,*

$$\tau^{\text{LATE}} = \frac{\mathbb{E}[\mathbb{E}[Y|X, Z=1] - \mathbb{E}[Y|X, Z=0]]}{\mathbb{E}[\text{P}(A=1|X, Z=1) - \text{P}(A=1|X, Z=0)]}. \tag{36.57}$$

Proof. We now show that, given the IV assumptions and monotonicity, $\text{LATE} = \tau^{\text{LATE}}$. First, notice that

$$\tau^{\text{LATE}} = \frac{\mathbb{E}[Y|\text{do}(Z=1)] - \mathbb{E}[Y|\text{do}(Z=0)]}{\text{P}(A=1|\text{do}(Z=1)) - \text{P}(A=1|\text{do}(Z=0))}. \tag{36.58}$$

This follows from backdoor adjustment, Theorem 2, applied to the numerator and denominator separately. Our strategy will be to decompose $\mathbb{E}[Y|\text{do}(Z=z)]$ into the contributions from the compliers, the units that ignore the instrument (the always/never takers), and the defiers. To that end, note that $\text{P}(\text{complier}|\text{do}(Z=z)) = \text{P}(\text{complier})$ and similarly for always/never takers and defiers—interventions on the instrument don't change the composition of the population. Then,

$$\mathbb{E}[Y|\text{do}(Z=1)] - \mathbb{E}[Y|\text{do}(Z=1)] \tag{36.59}$$

$$= \big(\mathbb{E}[Y|\text{complier}, \text{do}(Z=1)] - \mathbb{E}[Y|\text{complier}, \text{do}(Z=0)]\big)\text{P}(\text{complier}) \tag{36.60}$$

$$+ \big(\mathbb{E}[Y|\text{always/never}, \text{do}(Z=1)] - \mathbb{E}[Y|\text{always/never}, \text{do}(Z=0)]\big)\text{P}(\text{always/never}) \tag{36.61}$$

$$+ \big(\mathbb{E}[Y|\text{defier}, \text{do}(Z=1)] - \mathbb{E}[Y|\text{defier}, \text{do}(Z=0)]\big)\text{P}(\text{defier}). \tag{36.62}$$

8. We follow the econometrics literature in using "LATE" because "CATE" is already commonly used for conditional average treatment effect.

The key is the effect on the complier subpopulation, Equation (36.60). First, by definition of the complier population, we have that:

$$\mathbb{E}[Y|\text{complier}, \text{do}(Z = z)] = \mathbb{E}[Y|\text{complier}, \text{do}(A = z)]. \tag{36.63}$$

That is, the causal effect of the treatment is the same as the causal effect of the instrument in this subpopulation—this is the core reason why access to an instrument allows identification of the local average treatment effect. This means that

$$\text{LATE} = \mathbb{E}[Y|\text{complier}, \text{do}(Z = 1)] - \mathbb{E}[Y|\text{complier}, \text{do}(Z = 0)]. \tag{36.64}$$

Further, we have that $\text{P}(\text{complier}) = \text{P}(A = 1|\text{do}(Z = 1)) - \text{P}(A = 1|\text{do}(Z = 0))$. The reason is simply that, by definition of the subpopulations,

$$\text{P}(A = 1|\text{do}(Z = 1)) = \text{P}(\text{complier}) + \text{P}(\text{always taker}) \tag{36.65}$$
$$\text{P}(A = 1|\text{do}(Z = 0)) = \text{P}(\text{always taker}). \tag{36.66}$$

Now, plugging the expression for $\text{P}(\text{complier})$ and Equation (36.64) into Equation (36.60) we have that:

$$\big(\mathbb{E}[Y|\text{complier}, \text{do}(Z = 1)] - \mathbb{E}[Y|\text{complier}, \text{do}(Z = 0)]\big)\text{P}(\text{complier}) \tag{36.67}$$
$$= \text{LATE} \times (\text{P}(A = 1|\text{do}(Z = 1)) - \text{P}(A = 1|\text{do}(Z = 0))) \tag{36.68}$$

This gives us an expression for the local average treatment effect in terms of the effect of the instrument on the compliers and the probability that a unit takes the treatment when assigned/not-assigned.

The next step is to show that the remaining instrument effect decomposition terms, Equations (36.61) and (36.62), are both 0. Equation (36.61) is the causal effect of the instrument on the always/never takers. It's equal to 0 because, by definition of this subpopulation, the instrument has no causal effect in the subpopulation—they ignore the instrument! Mathematically, this is just $\mathbb{E}[Y|\text{always/never}, \text{do}(Z = 1)] = \mathbb{E}[Y|\text{always/never}, \text{do}(Z = 0)]$. Finally, Equation (36.62) is 0 by the instrument monotonicity assumption: we assumed that $\text{P}(\text{defier}) = 0$.

In totality, we now have that Equations (36.60) to (36.62) reduces to:

$$\mathbb{E}[Y|\text{do}(Z = 1)] - \mathbb{E}[Y|\text{do}(Z = 1)] \tag{36.69}$$
$$= \text{LATE} \times \big(\text{P}(A = 1|\text{do}(Z = 1)) - \text{P}(A = 1|\text{do}(Z = 0))\big) + 0 + 0 \tag{36.70}$$

Rearranging for LATE and plugging in to Equation (36.58) gives claimed identification result. □

36.5.2.1 Estimation

For estimating the local average treatment effect under the monotone instrument assumption, there is a double-machine learning approach that works with generic supervised learning approaches. Here, we want an estimator $\hat{\tau}^{\text{LATE}}$ for the parameter

$$\tau^{\text{LATE}} = \frac{\mathbb{E}[\mathbb{E}[Y|X, Z = 1] - \mathbb{E}[Y|X, Z = 0]]}{\mathbb{E}[\text{P}(A = 1|X, Z = 1) - \text{P}(A = 1|X, Z = 0)]}. \tag{36.71}$$

Author: Kevin P. Murphy. (C) MIT Press. CC-BY-NC-ND license

To define the estimator, it's convenient to introduce some additional notation. First, we define the nuisance functions:

$$\mu(z,x) = \mathbb{E}[Y|z,x] \tag{36.72}$$

$$m(z,x) = \mathrm{P}(A=1|x,z) \tag{36.73}$$

$$p(x) = \mathrm{P}(Z=1|x). \tag{36.74}$$

We also define the score ϕ by:

$$\phi_{Z\to Y}(\mathbf{X};\mu,p) \triangleq \mu(1,X) - \mu(0,X) + \frac{Z(Y-\mu(1,X))}{p(X)} - \frac{(1-Z)(Y-\mu(0,X))}{1-p(X)} \tag{36.75}$$

$$\phi_{Z\to A}(\mathbf{X};m,p) \triangleq m(1,X) - m(0,X) + \frac{Z(A-m(1,X))}{p(X)} - \frac{(1-Z)(A-m(0,X))}{1-p(X)} \tag{36.76}$$

$$\phi(\mathbf{X};\mu,m,p,\tau) \triangleq \phi_{Z\to Y}(\mathbf{X};\mu,p) - \phi_{Z\to A}(\mathbf{X};m,p)\times\tau \tag{36.77}$$

Then, the estimator is defined by a two stage procedure:

1. Fit models $\hat{\mu},\hat{m},\hat{p}$ for each of μ,m,p (using supervised machine learning).
2. Define $\hat{\tau}^{\mathrm{LATE}}$ as the solution to $\frac{1}{n}\sum_i \phi(\mathbf{X}_i;\hat{\mu},\hat{m},\hat{p},\hat{\tau}^{\mathrm{LATE}}) = 0$. That is,

$$\hat{\tau}^{\mathrm{LATE}} = \frac{\frac{1}{n}\sum_i \phi_{Z\to Y}(\mathbf{X}_i;\hat{\mu},\hat{p})}{\frac{1}{n}\sum_i \phi_{Z\to A}(\mathbf{X}_i;\hat{m},\hat{p})} \tag{36.78}$$

It may help intuitions to notice that the double machine learning estimator of the LATE is effectively the double machine learning estimator of of the average treatment effect of Z on Y divided by the double machine learning estimator of the average treatment effect of Z on A.

Similarly to Section 36.4, the nuisance functions can be estimated by:

1. fit a model $\hat{\mu}$ that predicts Y from Z,X by minimizing mean square error
2. fit a model $\hat{m}$ that predicts A from Z,X by minimizing mean cross-entropy
3. fit a model $\hat{p}$ that predicts Z from X by minimizing mean cross-entropy.

As in Section 36.4, reusing the same data for model fitting and computing the estimator can potentially cause problems. This can be avoided with use a cross-fitting procedure as described in Section 36.4.2.4. In this case, we split the data into K folds and, for each fold k, use all the but the kth fold to compute estimates $\hat{\mu}_{-k},\hat{m}_{-k},\hat{p}_{-k}$ of the nuisance parameters. Then we compute the nuisance estimates for each datapoint i in fold k by predicting the required quantity using the nuisance model fit on the other folds. That is, if unit i is in fold k, we compute $\hat{\mu}(z_i,x_i) \triangleq \hat{\mu}^{-k}(z_i,x_i)$ and so forth.

The key result is that if we use the cross-fit version of the estimator and the estimators for the nuisance functions converge to their true values in the sense that

1. $\mathbb{E}(\hat{\mu}(Z,X)-\mu(Z,X))^2 \to 0$, $\mathbb{E}(\hat{m}(Z,X)-m(Z,X))^2 \to 0$, and $\mathbb{E}(\hat{p}(X)-p(X))^2 \to 0$
2. $\sqrt{\mathbb{E}[(\hat{p}(X)-p(X))^2]} \times \left(\sqrt{\mathbb{E}[(\hat{\mu}(Z,X)-\mu(Z,X))^2]} + \sqrt{\mathbb{E}[(\hat{m}(Z,X)-m(Z,X))^2]}\right) = o(\sqrt{n})$

then (with some omitted technical conditions) we have asymptotic normality at the $\sqrt{n}$-rate:

$$\sqrt{n}(\hat{\tau}^{\text{LATE}-\text{cf}} - \tau^{\text{LATE}}) \xrightarrow{d} \text{Normal}(0, \frac{\mathbb{E}[\phi(\mathbf{X}; \mu, m, p, \tau^{\text{LATE}})^2]}{\mathbb{E}[m(1,X) - m(0,X)]^2}). \tag{36.79}$$

As with double machine learning for the confounder adjustment strategy, the key point here is that we can achieve the (optimal) $\sqrt{n}$ rate for estimating the LATE under a relatively weak condition on how well we estimate the nuisance functions—what matters is the *product* of the error in p and the errors in μ, m. So, for example, a very good model for how the instrument is assigned (p) can make up for errors in the estimation of the treatment-assignment (m) and outcome (μ) models.

The double machine learning estimator also gives a recipe for quantifying uncertainity. To that end, define

$$\hat{\tau}_{Z \to A} \triangleq \frac{1}{n} \sum_i \phi_{Z \to A}(\mathbf{X}_i; \hat{m}, \hat{p}) \tag{36.80}$$

$$\hat{\mathbb{V}}[\hat{\tau}^{\text{LATE}}] \triangleq \frac{1}{\hat{\tau}^2_{Z \to A}} \frac{1}{n} \sum_i \phi(\mathbf{X}_i; \hat{\mu}, \hat{m}, \hat{p}, \hat{\tau}^{\text{LATE}})^2. \tag{36.81}$$

Then, subject to suitable technical conditions, $\hat{\mathbb{V}}[\hat{\tau}^{\text{LATE}-\text{cf}}]$ can be used as an estimate of the variance of the estimator. More precisely,

$$\sqrt{n}(\hat{\tau}^{\text{LATE}} - \tau^{\text{LATE}}) \xrightarrow{d} \text{Normal}(0, \hat{\mathbb{V}}[\hat{\tau}^{\text{LATE}}]). \tag{36.82}$$

Then, confidence intervals or p-values can be computed using this variance in the usual way. The main extra condition required for the variance estimator to be valid is that the nuisance parameters must all converge at rate $O(n^{-1/4})$ (so an excellent estimator for one can't fully compensate for terrible estimators of the others). In fact, even this condition is unnecessary in certain special cases—e.g., when p is known exactly, which occurs when the instrument is randomly assigned. See Chernozhukov et al. [Che+17e] for technical details.

36.5.3 Two Stage Least Squares

Commonly, the IV assumptions are supplemented with the following linear model assumptions:

$$A_i \leftarrow \alpha_0 + \alpha Z_i + \delta_A X_i + \gamma_A X_i + \xi_i^A \tag{36.83}$$

$$Y_i \leftarrow \beta_0 + \beta A_i + \delta_Y X_i + \gamma_Y X_i + \xi_i^Y \tag{36.84}$$

That is, we assume that the real-world process for treatment assignment and the outcome are both linear. In this case, plugging Equation (36.83) into Equation (36.84) yields

$$Y_i \leftarrow \tilde{\beta}_0 + \beta\alpha Z_i + \tilde{\delta} X_i + \tilde{\gamma} X_i + \tilde{\xi}_i. \tag{36.85}$$

The point is that β, the average treatment effect of A on Y, is equal to the coefficient $\beta\alpha$ of the instrument in the outcome-instrument model divided by the coefficient α of the instrument in the treatment-instrument model. So, to estimate the treatment effect, we simply fit both linear models and divide the estimated coefficients. This procedure is called **two stage least squares**.

Author: Kevin P. Murphy. (C) MIT Press. CC-BY-NC-ND license

The simplicity of this procedure is seductive. However, the required linearity assumptions are hard to satisfy in practice and frequently lead to severe issues. A particularly pernicious version of this is that linear-model misspecfication together with weak relevance can yield standard errors for the estimate that are far too small. In practice, this can lead us to find large, significant estimates from two stage least squares when the truth is actually a weak or null effect. See [Rei16; You19; ASS19; Lal+21] for critical evaluations of two stage least squares in practice.

36.6 Difference in Differences

Unsurprisingly, time plays an important role in causality. Causes precede effects, and we should be able to incorporate this knowledge into causal identification. We now turn to a particular strategy for causal identification that relies on observing each unit at multiple time points. Data of this kind is sometimes called **panel data**. We'll consider the simplest case. There are two time periods. In the first period, none of the units are treated, and we observe an outcome Y_{0i} for each unit. Then, a subset of the units are treated, denoted by $A_i = 1$. In the second time period, we again observe the outcomes Y_{1i} for each unit, where now the outcomes of the treated units are affected by the treatment. Our goal is to determine the average effect receiving the treatment had on the treated units. That is, we want to know the average difference between the outcomes we actually observed for the treated units, and the outcomes we would have observed on those same units if they had not been treated. The general strategy we look at is called **difference in differences**.[9]

As a concrete motivating example, consider trying to determine the effect raising minimum wage on employment. The concern here is that, in an efficient labor market, increasing the price of workers will reduce the demand for them, thereby driving down employment. As such, it seems increasing minimum wage may hurt the people the policy is nominally intended to help. The question is: how strong is this effect in practice? Card and Krueger [CK94a] studied this effect using difference in differences. The Philadelphia metropolitan area includes regions in both Pennsylvania and New Jersey (different US states). On April 1st 1992, New Jersey raised its minimum wage from \$4.25 to \$5.05. In Pennsylvania, the wage remained constant at \$4.25. The strategy is to collect employment data from fast food restaurants (which pay many employees minimum wage) in each state before and after the change in minimum wage. In this case, for restaurant i, we have Y_{0i}, the number of full time employees in February 1992, and Y_{1i}, the number of full time employees in November 1992. The treatment is simply $A_i = 1$ if the restaurant was located in New Jersey, and $A_i = 0$ if located in Pennsylvania. Our goal is to estimate the average effect of the minimum wage hike on employment in the restaurants affected by it (i.e., the ones in New Jersey).

The assumption in classical difference-in-differences is the following structural equation:

$$Y_{ti} \leftarrow W_i + S_t + \tau A_i \mathbb{I}(t = 1) + \xi_{ti}, \tag{36.86}$$

with $\mathbb{E}[\xi_{ti}|W_i, S_t, A_i] = 0$. Here, W_i is a unit specific effect that is constant across time (e.g., the location of the restuarant or competence of the management) and S_t is a time-specific effect that applies to all units (e.g., the state of the US economy at each time). Both of these quantities are treated as unobserved, and not explicitly accounted for. The parameter τ captures the target causal effect. The (strong) assumption here is that unit, time, and treatment effects are all additive. This

9. See github.com/vveitch/causality-tutorials/blob/main/difference_in_differences.ipynb.

assumption is called **parallel trends**, because it is equivalent to assuming that, in the absence of treatment, the trend over time would be the same in both groups. It's easy to see that under this assumption, we have:

$$\tau = \mathbb{E}[Y_{1i} - Y_{0i}|A = 1] - \mathbb{E}[Y_{1i} - Y_{0i}|A = 0]. \tag{36.87}$$

That is, the estimand first computes the difference across time for both the treated and untreated group, and then computes the difference between these differences across the groups. The obvious estimator is then

$$\hat{\tau} = \frac{1}{n_A} \sum_{i:A_i=1} Y_{1i} - Y_{0i} - \frac{1}{n - n_A} \sum_{i:A_i=0} Y_{1i} - Y_{0i}, \tag{36.88}$$

where n_A is the number of treated units.

The root identification problem addressed by difference-in-differences is that $\mathbb{E}[W_i|A_i = 1] \neq \mathbb{E}[W_i|A_i = 0]$. That is, restaurants in New Jersey may be systematically different from restuarants in Pennsylvania in unobserved ways that affect employment.[10] This is why we can't simply compare average outcomes for the treated and untreated. The identification assumption is that this unit-specific effect is the only source of statistical association with treatment; in particular we assume the time-specific effect has no such issue: $\mathbb{E}[S_{1i} - S_{0i}|A_i = 1] = \mathbb{E}[S_{1i} - S_{0i}|A_i = 0]$. Unfortunately, this assumption can be too strong. For instance, administrative data shows employment in Pennsylvania falling relative to employment in New Jersey between 1993 and 1996 [AP08, §5.2]. Although this doesn't directly contradict the parallel trends assumption used for indentification, which needs to hold only in 1992, it does make it seem less credible.

To weaken the assumption, we'll look at a version that requires parallel trends to hold only after adjusting for covariates. To motivate this, we note that there were several different types of fast food restaurant included in the employment data. These vary, e.g., in the type of food they serve, and in cost per meal. Now, it seems reasonable the trend in employment may depend on the type of restuarant. For example, more expensive chains (such as Kentucky Fried Chicken) might be more affected by recessions than cheaper chains (such as McDonald's). If expensive chains are more common in New Jersey than in Pennsylvania, this effect can create a violation of parallel trends—if there's recession affecting both states, we'd expect employment to go down more in New Jersey than in Pennsylvania. However, we may find it credible that McDonald's restaurants in New Jersey have the same trend as McDonald's in Pennsylvania, and similarly for Kentucky Fried Chicken.

The next step is to give a definition of the target causal effect that doesn't depend on a parametric model, and a non-parametric statement of the identification assumption to go with it. In words, the causal estimand will be the average treatment effect on the units that received the treatment. To make sense of this mathematically, we'll introduce a new piece of notation:

$$\mathrm{P}^{A=1}(Y|\mathrm{do}(A = a)) \triangleq \int \mathrm{P}(Y|A = a, \text{parents of } Y)\mathrm{dP}(\text{parents of } Y|A = 1) \tag{36.89}$$

$$\mathbb{E}^{A=1}[Y|\mathrm{do}(A = a)] \triangleq \mathbb{E}_{\mathrm{P}^{A=1}(Y|\mathrm{do}(A=a))}[Y]. \tag{36.90}$$

In words: recall that the ordinary do operator works by replacing $\mathrm{P}(\text{parents}|A = a)$ by the marginal distribution $\mathrm{P}(\text{parents})$, thereby breaking the backdoor associations. Now, we're replacing the

10. This is similar to the issue that arises from unobserved confounding, except W_i need not be a cause of the treatment assignment.

Author: Kevin P. Murphy. (C) MIT Press. CC-BY-NC-ND license

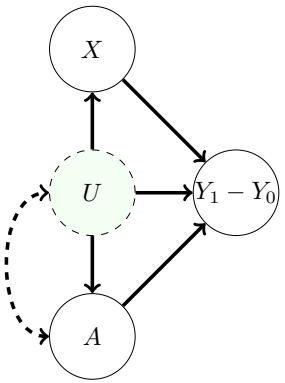


Figure 36.6: Causal graph assumed for the difference-in-differences setting. Here, the outcome of interest is the difference between the pre- and post-treatment period, $Y_1 - Y_0$. This difference is influenced by the treatment, unobserved factors U, and observed covariates X. The dashed arrow between U and A indicates a statistical dependency between the variables, but where we remain agnostic to the precise causal mechanism. For example, in the minimum wage example, U might be the average income in restaurant's neighbourhood, which is dependent on the state, and hence also the treatment.

distribution $\mathrm{P}(\text{parents}|A = a)$ by $\mathrm{P}(\text{parents}|A = 1)$, irrespective of the actual treatment value. This still breaks all backdoor associations, but is a better match for our target of estimating the treatment effect only among the treated units.

To formalize a causal estimand using the do calculus, we need to assume some partial causal structure. We'll use the graph in Figure 36.6. With this in hand, our causal estimand is the average treatment effect on the units that received the treatment, namely:

$$\mathrm{ATT}^{\mathrm{DiD}} = \mathbb{E}^{A=1}[Y_1 - Y_0|\mathrm{do}(A = 1)] - \mathbb{E}^{A=1}[Y_1 - Y_0|\mathrm{do}(A = 0)]] \tag{36.91}$$

In the minimum wage example, this is the average effect of the minimum wage hike on employment in the restaurants affected by it (i.e., the ones in New Jersey).

Finally, we formalize the identification assumption that, conditional on X, the trends in the treated and untreated groups are the same. The **conditional parallel trends** assumption is:

$$\mathbb{E}^{A=1}[Y_1 - Y_0|X, \mathrm{do}(A = 0)] = \mathbb{E}[Y_1 - Y_0|X, A = 0]. \tag{36.92}$$

In words, this says that for treated units with covariates X, the trend we would have seen had we not assigned treatment is the same as the trend we actually saw for the untreated units with covariates X. That is, if New Jersey had not raised its minimum wage, then McDonald's in New Jersey would have the same expected change in employment as McDonald's in Pennsylvania.

With this in hand, we can give the main identification result:

Theorem 5 (Difference in Differences Identification)**.** *We observe $A, Y_0, Y_1, X \sim P$. Suppose that*

1. *(Causal Structure) The data follows the causal graph in Figure 36.6.*
2. *(Conditional Parallel Trends) $\mathbb{E}^{A=1}[Y_1 - Y_0|X, \mathrm{do}(A = 0)] = \mathbb{E}[Y_1 - Y_0|X, A = 0]$.*

3. *(Overlap)* $\mathrm{P}(A=1) > 0$ *and* $\mathrm{P}(A=1|X=x) < 1$ *for all values of* x *in the sample space. That is, there are no covariate values that only exist in the treated group.*

Then, the average treatment effect on the treated is identified as $\mathrm{ATT}^{\mathrm{DiD}} = \tau^{\mathrm{DiD}}$*, where*

$$\tau^{\mathrm{DiD}} = \mathbb{E}[\mathbb{E}[Y_1 - Y_0|A=1, X] - \mathbb{E}[Y_1 - Y_0|A=0, X]|A=1]. \tag{36.93}$$

Proof. First, by unrolling definitions, we have that

$$\mathbb{E}^{A=1}[Y_1 - Y_0|\mathrm{do}(A=1), X] = \mathbb{E}[Y_1 - Y_0|A=1, X]. \tag{36.94}$$

The interpretation is the near-tautology that the average effect among the treated under treatment is equal to the actually observed average effect among the treated. Next,

$$\mathbb{E}^{A=1}[Y_1 - Y_0|\mathrm{do}(A=0), X] = \mathbb{E}[Y_1 - Y_0|A=0, X]. \tag{36.95}$$

is just the conditional parallel trends assumption. The result follows immediately.

(The overlap assumption is required to make sure all the conditional expectations are well defined). □

36.6.1 Estimation

With the identification result in hand, the next task is to estimate the observational estimand Equation (36.93). To that end, we define $\tilde{Y} \triangleq Y_1 - Y_0$. Then, we've assumed that $\tilde{Y}, X, A \overset{\mathrm{iid}}{\sim} P$ for some unknown distribution P, and our target estimand is $\mathbb{E}[\mathbb{E}[\tilde{Y}|A=1, X] - \mathbb{E}[\tilde{Y}|A=0, X]|A=1]$. We can immediately recognize this as the observational estimand that occurs in estimating the average treatment effect through adjustment, described in Section 36.4.5.3. That is, even though the causal situation and the identification argument are different between the adjustment setting and the difference in differences setting, the statistical estimation task we end up with is the same. Accordingly, we can use all of the estimation tools we developed for adjustment. That is, all of the techniques there—expected outcome modeling, propensity score methods, double machine learning, and so forth—were purely about the *statistical* task, which is the same between the two scenarios.

So, we're left with the same general recipe for estimation we saw in Section 36.4.6. Namely,

1. Fit statistical or machine-learning models $\hat{Q}(a, x)$ as a predictor for $\tilde{Y} = Y_1 - Y_0$, and/or $\hat{g}(x)$ as a predictor for A
2. Compute the predictions $\hat{Q}(0, x_i), \hat{Q}(1, x_i), \hat{g}(x_i)$ for each data point, and
3. Combine these predictions into an estimate of the average treatment effect on the treated.

The estimator in the third step can be the expected outcome model estimator, the propensity weighted estimator, the double machine learning estimator, or any other strategy that's valid in the adjustment setting.

36.7 Credibility Checks

Once we've chosen an identification strategy, fit our models, and produced an estimate, we're faced with a basic question: should we believe it? Whether the reported estimate succeeds in capturing

Author: Kevin P. Murphy. (C) MIT Press. CC-BY-NC-ND license

the true causal effect depends on whether the assumptions required for causal identification hold, the quality of the machine learning models, and the variability in the estimate due to only having access to a finite data sample. The latter two problems are already familiar from machine learning and statistical practice. We should, e.g., assess our models by checking performance on held out data, examining feature importance, and so forth. Similarly, we should report measures of the uncertainity due to finite sample (e.g., in the form of confidence intervals). Because these procedures are already familiar practice, we will not dwell on them further. However, model evaluation and uncertainity quantification are key parts of any credible causal analysis.

Assessing the validity of identification assumptions is trickier. First, there are assumptions that can in fact be checked from data. For example, overlap should be checked in analysis using backdoor adjustment or difference in differences, and relevance should be checked in the instrumental variable setting. Again, checking these conditions is absolutely necessary for a credible causal analysis. But, again, this involves only familiar data analysis, so we will not discuss it further. Next, there are the causal assumptions that cannot be verified from data; e.g., no unobserved confounding in backdoor adjustment, the exclusion restriction in IV, and conditional parallel trends in DiD. Ultimately, the validity of these assumptions must be assessed using substantive causal knowledge of the particular problem under consideration. However, it is possible to conduct some supplementary analyses that make the required judgement easier. We now discuss two such two such techniques.

36.7.1 Placebo Checks

In many situations we may be able to find a variable that can be interepreted as a "treatment" that is known to have no effect on the outcome, but which we expect to be confounded with the outcome in a very similar fashion to the true treatment of interest. For example, if we're trying to estimate the efficacy of a COVID vaccine in preventing symptomatic COVID, we might take our placebo treatment to be vaccination against HPV. We do not expect that there's any causal effect here. However, it seems plausible that latent factors that cause an individual to seek (or avoid) HPV vaccination and COVID vaccination are similar; e.g., health concientiousness, fear of needles, and so forth. Then, if our identification strategy is valid for the COVID vaccine, we'd also expect it to be to be valid for HPV vaccination. Accordingly, our estimation procedure we use for estimating the COVID effect should, when applied to HPV, yield $\hat{\tau} \approx 0$. Or, more precisely, the confidence interval should contain 0. If this does not happen, then we may suspect that there are still some confounding factors lurking that are not adequately handled by the identification procedure.

A similar procedure works when there is a variable that can be interpreted as an outcome which is known to not be affected by the treatment, but that shares confounders with the outcome we're actually interested in. For example, in the COVID vaccination case, we might take the null outcome to be symptomatic COVID within 7 days of vaccination [Dag+21]. Our knowledge of both the biological mechanism of vaccination and the amount of time it takes to develop symptoms after COVID infection (at least 2 days) lead us to conclude that it's unlikely that the treatment has a causal effect on the outcome. However, the properties of the treated people that affect how likely they are to develop symptomatic COVID are largely the same in the 7 day and, e.g., 6 month window. That includes factors such as risk aversion, baseline health, and so forth. Again, we can apply our identification strategy to estimate the causal effect of the treatment on the null outcome. If the confidence interval does not include 0, then we should doubt the credibility of the analysis.

36.7.2 Sensitivity Analysis to Unobserved Confounding

We now specialize to the case of estimating the average causal effect of a binary treatment by adjusting for confounding variables, as described in Section 36.4. In this case, causal identification is based on the assumption of 'no unobserved confounding'; i.e., the assumption that the observed covariates include all common causes of the treatment assignment and outcome. This assumption is fundamentally untestable from observed data, but its violation can induce bias in the estimation of the treatment effect—the unobserved confounding may completely or in part explain the observed association. Our aim in this part is to develop a sensitivity analysis tool to aid in reasoning about potential bias induced by unobserved confounding.

Intuitively, if we estimate a large positive effect then we might expect the real effect is also positive, even in the presence of mild unobserved confounding. For example, consider the association between smoking and lung cancer. One could argue that this association arises from a hormone that both predisposes carriers to both an increased desire to smoke and to a greater risk of lung cancer. However, the association between smoking and lung cancer is large—is it plausible that some unknown hormonal association could have a strong enough influence to explain the association? Cornfield et al. [Cor+59] showed that, for a particular observational dataset, such an umeasured hormone would need to increase the probability of smoking by at least a factor of nine. This is an unreasonable effect size for a hormone, so they conclude it's unlikely the causal effect can be explained away.

We would like a general procedure to allow domain experts to make judgments about whether plausible confounding is "mild" relative to the "large" effect. In particular, the domain expert must translate judgments about the strength of the unobserved confounding into judgments about the bias induced in the estimate of the effect. Accordingly, we must formalize what is meant by strength of unobserved confounding, and to show how to translate judgments about confounding strength into judgments about bias.

A prototypical example, due to Imbens [Imb03] (building on [RR83]), illustrates the broad approach. As above, the observed data consists of a treatment A, an outcome Y, and covariates X that may causally affect the treatment and outcome. Imbens [Imb03] then posits an additional unobserved binary confounder U for each patient, and supposes that the observed data and unobserved confounder were generated according to the following assumption, known as **Imbens' Sensitivity Model**:

$$U_i \overset{\text{iid}}{\sim} \text{Bern}(1/2) \tag{36.96}$$

$$A_i | X_i, U_i \overset{ind}{\sim} \text{Bern}(\text{sig}(\gamma X_i + \alpha U_i)) \tag{36.97}$$

$$Y_i | X_i, A_i, U_i \overset{ind}{\sim} \mathcal{N}(\tau A_i + \beta X_i + \delta U_i, \sigma^2). \tag{36.98}$$

where sig is the sigmoid function.

If we had observed U_i, we could estimate $(\hat{\tau}, \hat{\gamma}, \hat{\beta}, \hat{\alpha}, \hat{\delta}, \hat{\sigma}^2)$ from the data and report $\hat{\tau}$ as the estimate of the average treatment effect. Since U_i is not observed, it is not possible to identify the parameters from the data. Instead, we make (subjective) judgments about plausible values of α—how strongly U_i affects the treatment assignment—and δ—how strongly U_i affects the outcome. Contingent on plausible $\alpha = \alpha^*$ and $\delta = \delta^*$, the other parameters can be estimated. This yields an estimate of the treatment effect $\hat{\tau}(\alpha^*, \delta^*)$ under the presumed values of the sensitivity parameters.

The approach just outlined has a major drawback: it relies on a parametric model for the full data generating process. The assumed model is equivalent to assuming that, had U been observed, it

Author: Kevin P. Murphy. (C) MIT Press. CC-BY-NC-ND license

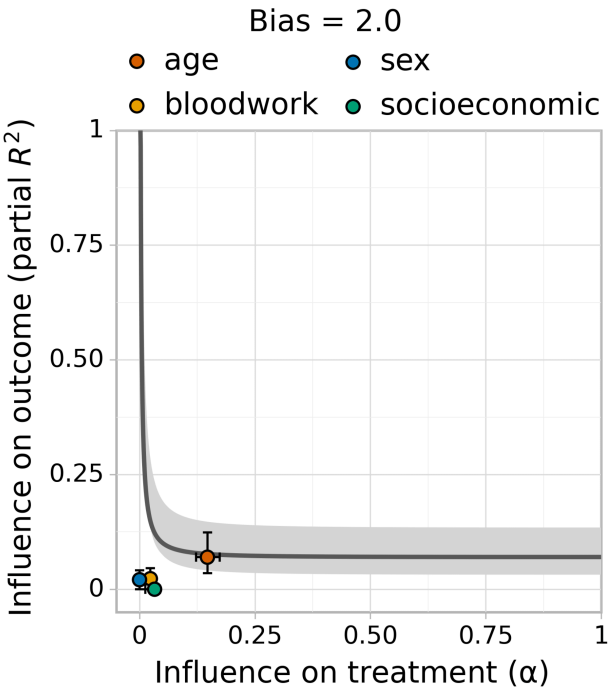


Figure 36.7: Austen plot showing how strong an unobserved confounder would need to be to induce a bias of 2 in an observational study of the effect of combination blood pressure medications on diastolic blood pressure [Dor+16]. We chose this bias to equal the nominal average treatment effect estimated from the data. We model the outcome with Bayesian Additive Regression Trees and the treatment assignment with logistic regression. The curve shows all values treatment and outcome influence that would induce a bias of 2. The colored dots show the influence strength of (groups of) observed covariates, given all other covariates. For example, an unobserved confounder with as much influence as the patient's age might induce a bias of about 2.

would have been appropriate to use logistic regression to model treatment assignment, and linear regression to model the outcome. This assumption also implies a simple, parametric model for the relationships governing the observed data. This restriction is out of step with modern practice, where we use flexible machine-learning methods to model these relationships. For example, the assumption forbids the use of neural networks or random forests, though such methods are often state-of-the-art for causal effect estimation.

Austen plots We now turn to developing an alternative an adaptation of Imbens' approach that fully decouples sensitivity analysis and modeling of the observed data. Namely, the **Austen plots** of [VZ20]. An example Austen plot is shown in Figure 36.7. The high-level idea is to posit a generative model that uses a simple, interpretable parametric form for the influence of the unobserved confounder, but that *puts no constraints on the model for the observed data.* We then use the

parametric part of the model to formalize "confounding strength" and to compute the induced bias as a function of the confounding.

Austen plots further adapt two strategies pioneered by Imbens [Imb03]. First, we find a parameterization of the model so that the sensitivity parameters, measuring strength of confounding, are on a standardized, unitless scale. This allows us to compare the strength of hypothetical unobserved confounding to the strength of observed covariates, measured from data. Second, we plot the curve of all values of the sensitivity parameter that would yield given level of bias. This moves the analyst judgment from "what are plausible values of the sensitivity parameters?" to "are sensitivity parameters this extreme plausible?"

Figure 36.7, an Austen plot for an observational study of the effect of combination medications on diastolic blood pressure, illustrates the idea. A bias of 2 would suffice to undermine the qualitative conclusion that the blood-pressure treatment is effective. Examining the plot, an unobserved confounder as strong as age could induce this amount of confounding, but no other (group of) observed confounders has so much influence. Accordingly, if a domain expert thinks an unobserved confounder as strong as age is unlikely then they may conclude that the treatment is likely effective. Or, if such a confounder is plausible, they may conclude that the study fails to establish efficacy.

Setup The data are generated independently and identically $(Y_i, A_i, X_i, U_i) \overset{\text{iid}}{\sim} P$, where U_i is not observed and P is some unknown probability distribution. The approach in Section 36.4 assumes that the observed covariates X contain all common causes of Y and A. If this 'no unobserved confounding' assumption holds, then the ATE is equal to parameter, τ, of the observed data distribution, where

$$\tau = \mathbb{E}[\mathbb{E}[Y|X, A=1] - \mathbb{E}[Y|X, A=0]]. \tag{36.99}$$

This observational parameter is then estimated from a finite data sample. Recall from Section 36.4 that this involves estimating the conditional expected outcome $Q(A, X) = \mathbb{E}[Y|A, X]$ and the propensity score $g(X) = \mathrm{P}(A=1|X)$, then plugging these into an estimator $\hat{\tau}$.

We are now concerned with the case of possible unobserved confounding. That is, where U causally affects Y and A. If there is unobserved confounding then the parameter τ is not equal to the ATE, so $\hat{\tau}$ is a biased estimate. Inference about the ATE then divides into two tasks. First, the statistical task: estimating τ as accurately as possible from the observed data. And, second, the causal (domain-specific) problem of assessing $\texttt{bias} = \texttt{ATE} - \tau$. We emphasize that our focus here is bias due to causal misidentification, not the statistical bias of the estimator. Our aim is to reason about the bias induced by unobserved confounding—the second task—in a way that imposes no constraints on the modeling choices for $\hat{Q}$, $\hat{g}$ and $\hat{\tau}$ used in the statistical analysis.

Sensitivity Model Our sensitivity analysis should impose no constraints on how the *observed* data is modeled. However, sensitivity analysis demands some assumption on the relationship between the observed data and the *unobserved* confounder. It is convenient to formalize such assumptions by specifying a probabilistic model for how the data is generated. The strength of confounding is then formalized in terms of the parameters of the model (the sensitivity parameters). Then, the bias induced by the confounding can be derived from the assumed model. Our task is to posit a generative model that both yields a useful and easily interpretable sensitivity analysis, and that avoids imposing any assumptions about the observed data.

Author: Kevin P. Murphy. (C) MIT Press. CC-BY-NC-ND license

To begin, consider the functional form of the sensitivity model used by Imbens [Imb03].

$$\text{logit}(\text{P}(A = 1|x, u)) = h(x) + \alpha u \tag{36.100}$$

$$\mathbb{E}[Y|a, x, u] = l(a, x) + \delta u, \tag{36.101}$$

for some functions h and l. That is, the propensity score is logit-linear in the unobserved confounder, and the conditional expected outcome is linear.

By rearranging Equation (36.100) to solve for u and plugging in to Equation (36.101), we see that it's equivalent to assume $\mathbb{E}[Y|t, x, u] = \tilde{l}(t, x) + \tilde{\delta}\text{logit}\text{P}(A = 1|x, u)$. That is, the unobserved confounder u only influences the outcome through the propensity score. Accordingly, by positing a distribution on $\text{P}(A = 1|x, u)$ directly, we can circumvent the need to explicitly articulate U (and h).

Definition 36.7.1. *Let $\tilde{g}(x, u) = \text{P}(A = 1|x, u)$ denote the propensity score given observed covariates x and the unobserved confounder u.*

The insight is that we can posit a sensitivity model by defining a distribution on $\tilde{g}$ directly. We choose:

$$\tilde{g}(X, U)|X \sim \text{Beta}(g(X)(1/\alpha - 1), (1 - g(X))(1/\alpha - 1)).$$

That is, the full propensity score $\tilde{g}(X, U)$ for each unit is assumed to be sampled from a Beta distribution centered at the observed propensity score $g(X)$. The sensitivity parameter α plays the same role as in Imbens' model: it controls the influence of the unobserved confounder U on treatment assignment. When α is close to 0 then $\tilde{g}(X, U)|X$ is tightly concentrated around $g(X)$, and the unobserved confounder has little influence. That is, U minimally affects our belief about who is likely to receive treatment. Conversely, when α is close to 1 then $\tilde{g}$ concentrates near 0 and 1; i.e., knowing U would let us accurately predict treatment assignment. Indeed, it can be shown that α is the change in our belief about how likely a unit was to have gotten the treatment, given that they were actually observed to be treated (or not):

$$\alpha = \mathbb{E}[\tilde{g}(X, U)|A = 1] - \mathbb{E}[\tilde{g}(X, U)|A = 0]. \tag{36.102}$$

With the $\tilde{g}$ model in hand, we define the **Austen Sensitivity Model** as follows:

$$\tilde{g}(X, U)|X \sim \text{Beta}(g(X)(1/\alpha - 1), (1 - g(X))(1/\alpha - 1)) \tag{36.103}$$

$$A|X, U \sim \text{Bern}(\tilde{g}(X, U)) \tag{36.104}$$

$$\mathbb{E}[Y|A, X, U] = Q(A, X) + \delta\big(\text{logit}\tilde{g}(X, U) - \mathbb{E}[\text{logit}\tilde{g}(X, U)|A, X]\big). \tag{36.105}$$

This model has been constructed to satisfy the requirement that the propensity score and conditional expected outcome are the g and Q actually present in the observed data:

$$\text{P}(A = 1|X) = \mathbb{E}[\mathbb{E}[T|X, U]|X] = \mathbb{E}[\tilde{g}(X, U)|X] = g(X)$$

$$\mathbb{E}[Y|A, X] = \mathbb{E}[\mathbb{E}[Y|A, X, U]|A, X] = Q(A, X).$$

The sensitivity parameters are α, controlling the dependence between the unobserved confounder the treatment assignment, and δ, controlling the relationship with the outcome.

Bias We now turn to calculating the bias induced by unobserved confounding. By assumption, X and U together suffice to render the average treatment effect identifiable as:

$$\texttt{ATE} = \mathbb{E}[\mathbb{E}[Y|A=1,X,U] - \mathbb{E}[Y|A=0,X,U]].$$

Plugging in our sensitivity model yields,

$$\texttt{ATE} = \mathbb{E}[Q(1,X) - Q(0,X)] + \delta(\mathbb{E}[\text{logit}\tilde{g}(X,U)|X,A=1] - \mathbb{E}[\text{logit}\tilde{g}(X,U)|X,A=0]).$$

The first term is the observed-data estimate τ, so

$$\texttt{bias} = \delta(\mathbb{E}[\text{logit}\tilde{g}(X,U)|X,A=1] - \mathbb{E}[\text{logit}\tilde{g}(X,U)|X,A=0]).$$

Then, by invoking Beta-Bernoulli conjugacy and standard Beta identities,[11] we arrive at,

Theorem 6. *Under the Austen sensitivity model, Equation (36.105), an unobserved confounder with influence α and δ induces bias in the estimated treatment effect equal to*

$$\texttt{bias} = \frac{\delta}{1/\alpha - 1}\mathbb{E}\big[\frac{1}{g(X)} + \frac{1}{1-g(X)}\big].$$

That is, the amount of bias is determined by the sensitivity parameters and by the *realized* propensity score. Notice that more extreme propensity scores lead to more extreme bias in response to unobserved confounding. This means, in particular, that conditioning on a covariate that affects the treatment but that does not directly affect the outcome (an instrument) will increase any bias due to unobserved confounding. This general phenomena is known as **z-bias**.

Sensitivity Parameters The Austen model provides a formalization of confounding strength in terms of the parameters α and δ and tells us how much bias is induced by a given strength of confounding. This lets us translate judgments about confounding strength to judgments about bias. However, it is not immediately obvious how to translate qualitative judgements such as "I think any unobserved confounder would be much less important than Age" to judgements about the possible values of the sensitivity parameters.

First, because the scale of δ is not fixed, it may be difficult to compare the influence of potential unobserved confounders to the influence of reference variables. To resolve this, we reexpress the outcome-confounder strength in terms of the (non-parametric) partial coefficient of determination:

$$R^2_{Y,\text{par}}(\alpha,\delta) = 1 - \frac{\mathbb{E}(Y - \mathbb{E}[Y|A,X,U])^2}{\mathbb{E}(Y - Q(A,X))^2}.$$

The key to computing the reparameterization is the following result

Theorem 7. *Under the Austen sensitivity model, Equation (36.105), the outcome influence is*

$$R^2_{Y,\text{par}}(\alpha,\delta) = \delta^2 \sum_{a=0}^{1} \frac{\mathbb{E}\big[\psi_1\big(g(X)^a(1-g(X))^{1-a}(1/\alpha - 1) + 1[A=a]\big)\big]}{\mathbb{E}[(Y - Q(A,X))^2]},$$

where ψ_1 is the trigamma function.

11. We also use the recurrence relation $\psi(x+1) - \psi(x) = 1/x$, where ψ is the digamma function.

Author: Kevin P. Murphy. (C) MIT Press. CC-BY-NC-ND license

See Veitch and Zaveri [VZ20] for the proof.

By design, α—the strength of confounding influence on on treatment assignment—is already on a fixed, unitless scale. However, because the measure is tied to the model it may be difficult to interpret, and it is not obvious how to compute reference confounding strength values from the observed data. The next result clarifies these issues.

Theorem 8. *Under the Austen sensitivity model, Equation* (36.105)*,*

$$\alpha = 1 - \frac{\mathbb{E}[\tilde{g}(X,U)(1-\tilde{g}(X,U))]}{\mathbb{E}[g(X)(1-g(X))]}.$$

See Veitch and Zaveri [VZ20] for the proof. That is, the sensitivity parameter α is measures how much more extreme the propensity scores become when we condition on U. That is, α is a measure of the extra predictive power U adds for A, above and beyond the predictive power in X. It may also be insightful to notice that

$$\alpha = R^2_{A,\text{par}} = 1 - \frac{\mathbb{E}[(A-\tilde{g}(X,U))^2]}{\mathbb{E}[(A-g(X))^2]}. \tag{36.106}$$

That is, α is just the (non-parametric) partial coefficient of determination of U on A—the same measure used for the outcome influence. (To see this, just expand the expectations conditional on $A=1$ and $A=0$).

Estimating bias In combination, Theorems 6 and 7 yield an expression for the bias in terms of α and $R^2_{Y,\text{par}}$. In practice, we can estimate the bias induced by confounding by fitting models for $\hat{Q}$ and $\hat{g}$ and replacing the expectations by means over the data.

36.7.2.1 Calibration using observed data

The analyst must make judgments about the influence a hypothetical unobserved confounder might have on treatment assignment and outcome. To calibrate such judgments, we'd like to have a reference point for how much the observed covariates influence the treatment assignment and outcome. In the sensitivity model, the degree of influence is measured by partial R^2_Y and α. We want to measure the degree of influence of an observed covariate Z given the other observed covariates $X \backslash Z$.

For the outcome, this can be measured as:

$$R^2_{Y\cdot Z|T,X\backslash Z} \triangleq 1 - \frac{\mathbb{E}(Y-Q(A,X))^2}{\mathbb{E}(Y-\mathbb{E}[Y|A,X\backslash Z])^2}.$$

In practice, we can estimate the quantity by fitting a new regression model $\hat{Q}_Z$ that predicts Y from A and $X\backslash Z$. Then we compute

$$R^2_{Y\cdot Z|T,X\backslash Z} = 1 - \frac{\frac{1}{n}\sum_i (y_i - \hat{Q}(t_i,x_i))^2}{\frac{1}{n}\sum_i (y_i - \hat{Q}_Z(t_i,x_i\backslash z_i))^2}.$$

Using Theorem 8, we can measure influence of observed covariate Z on treatment assignment given $X\backslash Z$ in an analogous fashion to the outcome. We define $g_{X\backslash Z}(X\backslash Z) = \mathrm{P}(A=1|X\backslash Z)$, then fit a

model for $g_{X\backslash Z}$ by predicting A from $X\backslash Z$, and estimate

$$\hat{\alpha}_{Z|X\backslash Z} = 1 - \frac{\frac{1}{n}\sum_i \hat{g}(x_i)(1-\hat{g}(x_i))}{\frac{1}{n}\sum_i \hat{g}_{X\backslash Z}(x_i\backslash z_i)(1-\hat{g}_{X\backslash Z}(x_i\backslash z_i))}.$$

Grouping covariates The estimated values $\hat{\alpha}_{X\backslash Z}$ and $\hat{R}^2_{Y,X\backslash Z}$ measure the influence of Z conditioned on all the other confounders. In some cases, this can be misleading. For example, if some piece of information is important but there are multiple covariates providing redundant measurements, then the estimated influence of each covariate will be small. To avoid this, group together related or strongly dependent covariates and compute the influence of the entire group in aggregate. For example, grouping income, location, and race as 'socioeconomic variables'.

36.7.2.2 Practical Use

We now have sufficient results to produce Austen plots such as Figure 36.7. At a high level, the procedure is:

1. Produce an estimate $\hat{\tau}$ using any modeling tools. As a component of this, estimate the propensity score $\hat{g}$ and conditional outcome model $\hat{Q}$
2. Pick a level of bias that would suffice to change the qualitative interpretation of the estimate (e.g., the lower bound of a 95% confidence interval).
3. Plot the values of α and $R^2_{Y,\text{par}}$ that would suffice to induce that much bias. This is the black curve on the plot. To calculate these values, use Theorems 6 and 7 together with the estimated $\hat{g}$ and $\hat{Q}$.
4. Finally, compute reference influence level for (groups of) observed covariates. In particular, this requires fitting reduced models for the conditional expected outcome and propensity that do not use the reference covariate as a feature.

In practice, an analyst only needs to do the model fitting parts themselves. The bias calculations, reference value calculations, and plotting can be done automatically with standard libraries.[12].

Austen plots are predicated on Equation (36.105). This assumption replaces the purely parametric Equation (36.98) with a version that eliminates any parametric requirements on the observed data. However, we emphasize that Equation (36.105) does, implicitly, impose some parametric assumption on the structural causal relationship between U and A, Y. Ultimately, any conclusion drawn from the sensitivity analysis depends on this assumption, which is not justified on any substantive grounds. Accordingly, such sensitivity analyses can only be used to informally guide domain experts. They do not circumvent the need to thoroughly adjust for confounding. This reliance on a structural assumption is a generic property of sensitivity analysis.[13] Indeed, there are now many sensitivity analysis models that allow the use of any machine learning model in the data analysis [e.g., RRS00; FDF19; She+11; HS13; BK19; Ros10; Yad+18; ZSB19; Sch+21a]. However, none of these are yet in

12. See github.com/vveitch/causality-tutorials/blob/main/Sensitivity_Analysis.ipynb.
13. In extreme cases, there can be so little unexplained variation in A or Y that only a very weak confounder could be compatible with the data. In this case, essentially assumption free sensitivity analysis is possible [Man90].

routine use in practice. We have presented Austen plots here not because they make an especially virtuous modeling assumption, but because they are (relatively) easy to understand and interpret.

Austen plots are most useful in situations where the conclusion from the plot would be 'obvious' to a domain expert. For instance, in Figure 36.7, we can be confident that an unobserved confounder similar to socioeconomic status would not induce enough bias to change the qualitative conclusion. By contrast, Austen plots should not be used to draw conclusions such as, "I think a latent confounder could only be 90% as strong as 'age', so there is evidence of a small non-zero effect". Such nuanced conclusions might depend on issues such as the particular sensitivity model we use, or finite-sample variation of our bias and influence estimates, or on incautious interpretation of the calibration dots. These issues are subtle, and it would be difficult resolve them to a sufficient degree that a sensitivity analysis would make an analysis credible.

Calibration using observed data The interpretation of the observed-data calibration requires some care. The sensitivity analysis requires the analyst to make judgements about the strength of influence of the unobserved confounder U, *conditional on the observed covariates* X. However, we report the strength of influence of observed covariate(s) Z, *conditional on the other observed covariates* $X \backslash Z$. The difference in conditioning sets can have subtle effects.

Cinelli and Hazlett [CH20] give an example where Z and U are identical variables in the true model, but where influence of U given A, X is larger than the influence of Z given $A, X \backslash Z$. (The influence of Z given $X \backslash Z, U$ would be the same as the influence of U given X). Accordingly, an analyst is *not* justified in a judgement such as, "I know that U and Z are very similar. I see Z has substantial influence, but the dot is below the line. Thus, U will not undo the study conclusions." In essence, if the domain expert suspects a strong interaction between U and Z then naively eyeballing the dot-vs-line position may be misleading. A particular subtle case is when U and Z are independent variables that both strongly influence A and Y. The joint influence on A creates an interaction effect between them when A is conditioned on (the treatment is a collider). This affects the interpretation of $R^2_{Y \cdot U|X,A}$. Indeed, we should generally be skeptical of sensitivity analysis interpretation when it is expected that a strong confounder has been omitted. In such cases, our conclusions may depend substantively on the particular form of our sensitivity model, or other unjustifiable assumptions.

Although the interaction problem is conceptually important, its practical significance is unclear. We often expect the opposite effect: if U and Z are dependent (e.g., race and wealth) then omitting U should increase the apparent importance of Z—leading to a conservative judgement (a dot artifically towards the top right part of the plot).

36.8 The Do Calculus

We have seen several strategies for identifying causal effects as parameters of observational distributions. Confounder adjustment (Section 36.4) relied only on the assumed causal graph (and overlap), which specified that we observe all common causes of A and Y. On the other hand, instrumental variable methods and difference-in-differences each relied on both an assumed causal graph and partial functional form assumptions about the underlying structural causal model. Because functional form assumptions can be quite difficult to justify on substantive grounds, it's natural to ask when causal identification is possible from the causal graph alone. That is, when can we be agnostic to the particular functional form of the structural causal models?

There is a general "**calculus of intervention**", known as the **do-calculus**, that gives a general recipe for determining when the causal assumptions expressed in a causal graph can be used to identify causal effects [Pea09c]. The do-calculus is a set of three rewrite rules that allows us to replace statements where we condition on variables being set by intervention, e.g. $\mathrm{P}(Y|\mathrm{do}(A=a))$, with statements involving only observational quantities, e.g. $\mathbb{E}_X[\mathrm{P}(Y|A=a,X)]$. When causal identification is possible, we can repeatedly apply the three rules to boil down our target causal parameter into an expression involving only the observational distribution.

36.8.1 The three rules

To express the rules, let X, Y, Z, and W be arbitrary disjoint sets of variables in a causal DAG G.

Rule 1 The first rule allows us to insert or delete observations z:

$$p(y|\mathrm{do}(x),z,w) = p(y|\mathrm{do}(x),w) \text{ if } (Y \perp Z|X,W)_{G_{\overline{X}}} \tag{36.107}$$

where $G_{\overline{X}}$ denotes cuting edges going into X, and $(Y \perp Z|X,W)_{G_{\overline{X}}}$ denotes conditional independence in the mutilated graph. The rule follows from d-separation in the mutilated graph. This rule just says that conditioniong on irrelevant variables leaves the distribution invariant (as we would expect).

Rule 2 The second rule allows us to replace $\mathrm{do}(z)$ with conditioning on (seeing) z. The simplest case where can do this is: if Z is a root of the causal graph (i.e., it has no causal parents) then $p(y|\mathrm{do}(z)) = p(y|z)$. The reason is that the do operator is equivalent to conditioning in the mutilated causal graph where all the edges into Z are removed, but, because Z is a root, the mutilated graph is just the original causal graph. The general form of this rule is:

$$p(y|\mathrm{do}(x),\mathrm{do}(z),w) = p(y|\mathrm{do}(x),z,w) \text{ if } (Y \perp Z|X,W)_{G_{\overline{X}\underline{Z}}} \tag{36.108}$$

where $G_{\overline{X}\underline{Z}}$ cuts edges going into X and out of Z. Intutively, we can replace $\mathrm{do}(z)$ by z as long as there are no backdoor (non-directed) paths between z and y. If there are in fact no such paths, then cutting all the edges going out of Z will mean there are no paths connecting Z and Y, so that $Y \perp\!\!\!\perp Z$. The rule just generalizes this line of reasoning to allow for extra observed and intervened variables.

Rule 3 The third rule allows us to insert or delete actions $\mathrm{do}(z)$:

$$p(y|\mathrm{do}(x),\mathrm{do}(z),w) = p(y|\mathrm{do}(x),w) \text{ if } (Y \perp Z|X,W)_{G_{\overline{X}\overline{Z^*}}} \tag{36.109}$$

where $G_{\overline{X}\overline{Z^*}}$ cuts edges going into X and Z^*, and where Z^* is the set of Z-nodes that are not ancestors of any W-node in $G_{\overline{X}}$. Intuitively, this condition corresponds to intervening on X, and checking whether the distribution of Y is invariant to *any* intervention that we could apply on Z.

36.8.2 Revisiting Backdoor Adjustment

We begin with a more general form of the adjustment formula we used in Section 36.4.

Author: Kevin P. Murphy. (C) MIT Press. CC-BY-NC-ND license

First, suppose we observe all of A's parents, call them X. For notational simplicity, we'll assume for the moment that X is discrete. Then,

$$p(Y = y|\text{do}(A = a)) = \sum_x p(Y = y|x, \text{do}(A = a))p(x|\text{do}(A = a)) \tag{36.110}$$

$$= \sum_x p(Y = y|x, A = a)p(x). \tag{36.111}$$

The first line is just a standard probability relation (marginalizing over z). We are using causal assumptions in two ways in the second line. First, $p(x|\text{do}(A = a)) = p(x)$: the treatment has no causal effect on Z, so interventions on A don't change the distribution of Z. This is rule 3, Equation (36.109). Second, $p(Y = y|z, \text{do}(A = a)) = p(Y = y|z, A = a)$. This equality holds because conditioning on the parents blocks all non-directed paths from A to Y, reducing the causal effect to be the same as the observational effect. The equality is an application of rule 2, Equation (36.108).

Now, what if we don't observe all the parents of A? The key issue is **backdoor paths**: paths between A and Y that contain an arrow into A. These paths are the general form of the problem that occurs when A and Y share a common cause. Suppose that we can find a set of variables S such that (1) no node in S is a descendant of A; and (2) S blocks every backdoor path between A and Y. Such a set is said to satisfy the **backdoor criterion**. In this case, we can use S instead of the parents of X in the adjustment formula, Equation (36.111). That is,

$$p(Y = y|\text{do}(A = a)) = \mathbb{E}_S[p(Y = y|S, A = a)]. \tag{36.112}$$

The proof follows the invocation of rules 3 and 2, in the same way as for the case where S is just the parents of A. Notice that requiring S to not contain any descendants of A means that we don't risk conditioning on any variables that mediate the effect, nor any variables that might be colliders—either would undermine the estimate.

The backdoor adjustment formula generalizes the adjust-for-parents approach and adjust-for-all-common-causes approach of Section 36.4. That's because both the parents of A and the common causes satisfy the backdoor criterion.

In practice, the full distribution $p(Y = y|\text{do}(A = a))$ is rarely used as the causal target. Instead, we try to estimate a low-dimensional parameter of this distribution, such as the average treatment effect. The adjustment formula immediately translates in the obvious way. If we define

$$\tau = \mathbb{E}_S[\mathbb{E}[Y|A = 1, S] - \mathbb{E}[Y|A = 0, S]],$$

then we have that ATE $= \tau$ whenever S satisfies the backdoor criteria. The parameter τ can then be estimated from finite data using the methods described in Section 36.4, using S in place of the common causes X.

36.8.3 Frontdoor Adjustment

Backdoor adjustment is applicable if there's at least one observed variable on every backdoor path between A and Y. As we have seen, identification is sometimes still possible even when this condition doesn't hold. Frontdoor adjustment is another strategy of this kind. Figure 36.8 shows the causal structure that allows this kind of adjustment strategy. Suppose we're interested in the effect of smoking A on developing cancer Y, but we're concerned about some latent genetic confounder U.

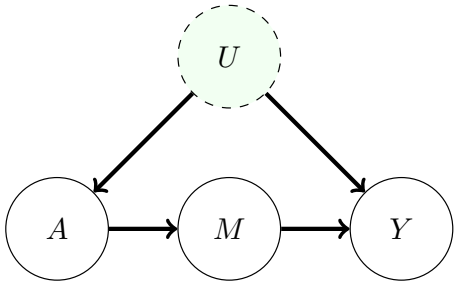


Figure 36.8: Causal graph illustrating the frontdoor criterion setup. The effect of the treatment A on outcome Y is entirely mediated by mediator M. This allows us infer the causal effect even if the treatment and outcome are confounded by U.

Suppose that all of the directed paths from A to Y pass through some set of variables M. Such variables are called **mediators**. For example, the effect of smoking on lung cancer might be entirely mediated by the amount of tar in the lungs and measured tissue damage. It turns out that if all such mediators are observed, and the mediators do not have an unobserved common cause with A or Y, then causal identification is possible. To understand why this is true, first notice that we can identify the causal effect of A on M and the causal effect of M on A, both by backdoor adjustment. Further, the mechanism of action of A on Y is: A changes M which in turn changes Y. Then, we can combine these as:

$$p(Y|\text{do}(A=a)) = \sum_m p(Y|\text{do}(M=m))p(M=m|\text{do}(A=a)) \tag{36.113}$$

$$= \sum_m \sum_{a'} p(Y|a',m)p(a')p(m|a) \tag{36.114}$$

The second line is just backdoor adjustment applied to identify each of the do expressions (note that A blocks the M-Y backdoor path through U).

Equation (36.114) is called the **front-door formula** [Pea09b, §3.3.2]. To state the result in more general terms, let us introduce a definition. We say a set of variables M satisfies the **front-door criterion** relative to an ordered pair of variables (A, Y) if (1) M intercepts all directed paths from A to Y; (2) there is no unblocked backdoor path from A to M; and (3) all backdoor paths from M to Y are blocked by A. If M satisfies this criterion, and if $p(A, M) > 0$ for all values of A and M, then the causal effect of A on Y is identifiable and is given by Equation (36.114).

Let us interpret this theorem in terms of our smoking example. Condition 1 means that smoking A should have no effect on cancer Y except via tar and tissue damage M. Conditions 2 and 3 mean that the genotype U cannot have any effect on M except via smoking A. Finally, the requirement that $p(A, M) > 0$ for all values implies that high levels of tar in the lungs must arise not only due to smoking, but also other factors (e.g., pollutants). In other words, we require $p(A = 0, M = 1) > 0$ so we can assess the impact of the mediator in the untreated setting.

We can now use the do-calculus to derive the frontdoor criterion; following [PM18b, p236]. Assuming

Author: Kevin P. Murphy. (C) MIT Press. CC-BY-NC-ND license

the causal graph G shown in Figure 36.8:

$$\begin{aligned}
p(y|\text{do}(a)) &= \sum_m p(y|\text{do}(a), m)p(m|\text{do}(a)) && \text{(probability axioms)} \\
&= \sum_m p(y|\text{do}(a), \text{do}(m))p(m|\text{do}(a)) && \text{(rule 2 using } G_{\overline{S}\underline{T}}\text{)} \\
&= \sum_m p(y|\text{do}(a), \text{do}(m))p(m|a) && \text{(rule 2 using } G_{\underline{S}}\text{)} \\
&= \sum_m p(y|\text{do}(m))p(m|a) && \text{(rule 3 using } G_{\overline{ST}^*}\text{)} \\
&= \sum_{a'}\sum_m p(y|\text{do}(m), a')p(a'|\text{do}(m))p(m|a) && \text{(probability axioms)} \\
&= \sum_{a'}\sum_m p(y|m, a')p(a'|\text{do}(m))p(m|a) && \text{(rule 2 using } G_{\underline{T}}\text{)} \\
&= \sum_{a'}\sum_m p(y|m, a')p(a')p(m|a) && \text{(rule 3 using } G_{\overline{T}^*}\text{)}
\end{aligned}$$

Estimation To estimate the causal distribution from data using the frontdoor criterion we need to estimate each of $p(y|m, a)$, $p(a)$, and $p(m|a)$. In practice, we can fit models $\hat{p}(y|m, a)$ by predicting Y from M and A, and $\hat{p}(m|a)$ by predicting M from A. Then, using the empirical distribution to estimate $p(a)$, the final estimate is:

$$\frac{1}{|A|}\sum_{a'}\sum_m \hat{p}(y|m, a')\hat{p}(m|a), \tag{36.115}$$

where $|A|$ is the number of treatments.

We usually have more modest targets than the full distribution $p(y|\text{do}(a))$. For instance, we may be content with just estimating the average treatment effect. It's straightforward to derive a formula for this using the frontdoor adjustment. Similarly to backdoor adjustment, more advanced estimators of the ATE through frontdoor effect are possible in principle. For example, we might combine fitted models for $\mathbb{E}[Y|m, a]$ and $\text{P}(M|a)$. See Fulcher et al. [Ful+20] for an approach to robust estimation via front door adjustment, as well as a generalization of the front door approach to more general settings.

36.9 Further Reading

There is an enormous and growing literature on the intersection of causality and machine learning.

First, there are many textbooks on theoretical and practical elements of causal inference. These include Pearl [Pea09c], focused on causal graphs, Angrist and Pischke [AP08], focused on econometrics, Hernán and Robins [HR20b], with roots in epidemiology, Imbens and Rubin [IR15], with origin in statistics, and Morgan and Winship [MW15], for a social sciences perspective. The introduction to causality in Shalizi [Sha22, §7] is also recommended, particularly the treatment of matching.

Double machine-learning has featured prominently in this chapter. This is a particular instantiation of non-parametric estimation. This topic has substantial theoretical and practical importance in

modern causal inference. The double machine learning work includes estimators for many commonly encountered scenarios [Che+17e; Che+17d]. Good references for a lucid explanation of how and why non-parametric estimation works include [Ken16; Ken17; FK21]. Usually, the key guarantees of non-parametric estimator are asymptotic. Generally, there are many estimators that share optimal asymptotic guarantees (e.g. the AIPTW estimator given in Equation (36.30)). Although these are asymptotically equivalent, in finite samples their behavior can be very different. There are estimators that preserve asymptotic guarantees but aim to improve performance in practical finite sample regimes [e.g., vR11].

There is also considerable interest in the estimation of heterogeneous treatment effects. The question here is: what effect would this treatment have when applied to a unit with such-and-such specific characteristics? E.g., what is the effect of this drug on women over the age of 50? The causal identification arguments used here are more-or-less the same as for the estimation of average case effects. However, the estimation problems can be substantially more involved. Some reading includes [Kün+19; NW20; Ken20; Yad+21].

There are several commonly applicable causal identification and estimation strategies beyond the ones we've covered in this chapter. **Regression discontinuity designs** rely on the presence of some sharp, arbitrary non-linearity in treatment assignment. For example, eligibility for some aid programs is determined by whether an individual has income below or above a fixed amount. The effect of the treatment can be studied by comparing units just below and just above this threshhold. **Synthetic controls** are a class of methods that try to study the effect of a treatment on a given unit by constructing a synthetic version of that unit that acts as a control. For example, to study the effect of legislation banning smoking indoors in California, we can construct a synthetic California as a weighted average of other states, with weights chosen to balance demographic characteristics. Then, we can compare the observed outcome of California with the outcome of the synthetic control, constructed as the weighted average of the outcomes of the donor states. See Angrist and Pischke [AP08] for a textbook treatment of both strategies. Closely related are methods that use time series modeling to create synthetic outcomes. For example, to study the effect of an advertising campaign beginning at time T on product sales Y_t, we might build a time series model for Y_t using data in the $t < T$ period, and then use this model to predict the values of $(\hat{Y}_t)_{t>T}$ we would have seen had the campaign not been run. We can estimate the causal effect by comparing the factual, realized Y_t to the predicted, counterfactual, $\hat{Y}_t$. See Brodersen et al. [Bro+15] for an instantiation of this idea.

In this chapter, our focus has been on using machine learning tools to estimate causal effects. There is also a growing interest in using the ideas of causality to improve machine learning tools. This is mainly aimed at building predictors that are robust when deployed in new domains [SS18b; SCS19; Arj+20; Mei18b; PBM16a; RC+18; Zha+13a; Sch+12b; Vei+21] or that do not rely on particular 'spurious' correlations in the training data [RPH21; Wu+21; Gar+19; Mit+20; WZ19; KCC20; KHL20; TAH20; Vei+21].

Author: Kevin P. Murphy. (C) MIT Press. CC-BY-NC-ND license

Index

Author: Kevin P. Murphy. (C) MIT Press. CC-BY-NC-ND license

Author: Kevin P. Murphy. (C) MIT Press. CC-BY-NC-ND license

Author: Kevin P. Murphy. (C) MIT Press. CC-BY-NC-ND license

Author: Kevin P. Murphy. (C) MIT Press. CC-BY-NC-ND license

Author: Kevin P. Murphy. (C) MIT Press. CC-BY-NC-ND license

Bibliography

[AA18] D. Amir and O. Amir. "Highlights: Summarizing agent behavior to people". In: *Proceedings of the 17th International Conference on Autonomous Agents and MultiAgent Systems*. 2018, pp. 1168–1176.

[AB08] C. Archambeau and F. Bach. "Sparse probabilistic projections". In: *NIPS*. 2008.

[AB17] M. Arjovsky and L. Bottou. "Towards principled methods for training generative adversarial networks". In: (2017).

[AB21] A. N. Angelopoulos and S. Bates. "A Gentle Introduction to Conformal Prediction and Distribution-Free Uncertainty Quantification". In: (2021). arXiv: 2107.07511 [cs.LG].

[Aba] A. Abadie. "Using Synthetic Controls: Feasibility, Data Requirements, and Methodological Aspects". In: *J. of Economic Literature* ().

[Abd+18] A. Abdolmaleki, J. T. Springenberg, Y. Tassa, R. Munos, N. Heess, and M. A. Riedmiller. "Maximum a Posteriori Policy Optimisation". In: *ICLR*. 2018.

[ABM10] J.-Y. Audibert, S. Bubeck, and R. Munos. "Best Arm Identification in Multi-Armed Bandits". In: *COLT*. 2010, pp. 41–53.

[Abn+21] S. Abnar, M. Dehghani, B. Neyshabur, and H. Sedghi. "Exploring the limits of large scale pre-training". In: *ICLR*. 2021.

[ABV21] S. Akbayrak, I. Bocharov, and B. de Vries. "Extended Variational Message Passing for Automated Approximate Bayesian Inference". In: *Entropy* 23.7 (2021).

[AC17] S. Aminikhanghahi and D. J. Cook. "A Survey of Methods for Time Series Change Point Detection". en. In: *Knowl. Inf. Syst.* 51.2 (2017), pp. 339–367.

[AC93] J. Albert and S. Chib. "Bayesian analysis of binary and polychotomous response data". In: *JASA* 88.422 (1993), pp. 669–679.

[ACB17] M. Arjovsky, S. Chintala, and L. Bottou. "Wasserstein generative adversarial networks". In: *ICML*. 2017, pp. 214–223.

[ACL16] L. Aitchison, N. Corradi, and P. E. Latham. "Zipf's Law Arises Naturally When There Are Underlying, Unobserved Variables". en. In: *PLoS Comput. Biol.* 12.12 (2016), e1005110.

[ACL89] L. Atlas, D. Cohn, and R. Ladner. "Training connectionist networks with queries and selective sampling". In: *Advances in neural information processing systems* 2 (1989).

[ACP87] S. Arnborg, D. G. Corneil, and A. Proskurowski. "Complexity of finding embeddings in a k-tree". In: *SIAM J. on Algebraic and Discrete Methods* 8 (1987), pp. 277–284.

[Ada00] L. Adamic. *Zipf, Power-laws, and Pareto - a ranking tutorial*. Tech. rep. 2000.

[Ada+20] V. Adam, S. Eleftheriadis, N. Durrande, A. Artemev, and J. Hensman. "Doubly Sparse Variational Gaussian Processes". In: *AISTATS*. 2020.

[Ade+20a] J. Adebayo, J. Gilmer, M. Muelly, I. Goodfellow, M. Hardt, and B. Kim. *Sanity Checks for Saliency Maps*. 2020. arXiv: 1810.03292 [cs.CV].

[Ade+20b] J. Adebayo, M. Muelly, I. Liccardi, and B. Kim. "Debugging tests for model explanations". In: *arXiv preprint arXiv:2011.05429* (2020).

[Adl+18] P. Adler, C. Falk, S. A. Friedler, T. Nix, G. Rybeck, C. Scheidegger, B. Smith, and S. Venkatasubramanian. "Auditing black-box models for indirect influence". In: *Knowledge and Information Systems* 54.1 (2018), pp. 95–122.

[Ado] *Taking It to the MAX: Adobe Photoshop Gets New NVIDIA AI-Powered Neural Filters*. https://blogs.nvidia.com/blog/2020/10/20/adobe-max-ai/. Accessed: 2021-08-12.

[AE+20] D. Agudelo-España, S. Gomez-Gonzalez, S. Bauer, B. Schölkopf, and J. Peters. "Bayesian Online Prediction of Change Points". In: *UAI*. Vol. 124. Proceedings of Machine Learning Research. PMLR, 2020, pp. 320–329.

[AFD01] C. Andrieu, N. de Freitas, and A. Doucet. "Robust Full Bayesian Learning for Radial Basis Networks". In: *Neural Computation* 13.10 (2001), pp. 2359–2407.

[AFG19] M. Akten, R. Fiebrink, and M. Grierson. "Learning to See: You Are What You See". In: *ACM SIGGRAPH 2019 Art Gallery*. SIGGRAPH '19. Association for Computing Machinery, 2019.

[AG11] A. Allahverdyan and A. Galstyan. "Comparative Analysis of Viterbi Training and Maximum Likelihood Estimation for HMMs". In: *NIPS*. 2011, pp. 1674–1682.

[AG13] S. Agrawal and N. Goyal. "Further Optimal Regret Bounds for Thompson Sampling". In: *AISTATS*. 2013.

[Aga+14] D. Agarwal, B. Long, J. Traupman, D. Xin, and L. Zhang. "LASER: a scalable response prediction platform for online advertising". In: *WSDM*. 2014.

[Aga+21a] A. Agarwal, N. Jiang, S. M. Kakade, and W. Sun. *Reinforcement Learning: Theory and Algorithms*. 2021.

[Aga+21b] R. Agarwal, L. Melnick, N. Frosst, X. Zhang, B. Lengerich, R. Caruana, and G. Hinton. *Neural Additive Models: Interpretable Machine Learning with Neural Nets*. 2021. arXiv: 2004.13912 [cs.LG].

[AGM14] P. Agrawal, R. Girshick, and J. Malik. "Analyzing the performance of multilayer neural networks for object recognition". In: *European conference on computer vision*. Springer. 2014, pp. 329–344.

[AH09] I. Arasaratnam and S. Haykin. "Cubature Kalman Filters". In: *IEEE Trans. Automat. Contr.* 54.6 (2009), pp. 1254–1269.

[AHE07] I. Arasaratnam, S. Haykin, and R. J. Elliott. "Discrete-Time Nonlinear Filtering Algorithms

Using Gauss–Hermite Quadrature". In: *Proc. IEEE* 95.5 (2007), pp. 953–977.

[AHG20] L. Ambrogioni, M. Hinne, and M. van Gerven. "Automatic structured variational inference". In: (2020). arXiv: 2002.00643 [stat.ML].

[AHK01] C. C. Aggarwal, A. Hinneburg, and D. A. Keim. "On the Surprising Behavior of Distance Metrics in High Dimensional Space". In: *Database Theory — ICDT 2001*. Springer Berlin Heidelberg, 2001, pp. 420–434.

[AHK12] A. Anandkumar, D. Hsu, and S. M. Kakade. "A Method of Moments for Mixture Models and Hidden Markov Models". In: *COLT*. Vol. 23. Proceedings of Machine Learning Research. PMLR, 2012, pp. 33.1–33.34.

[AHK65] K. Abend, T. J. Harley, and L. N. Kanal. "Classification of Binary Random Patterns". In: *IEEE Transactions on Information Theory* 11(4) (1965), pp. 538–544.

[Ahm+17] S. Ahmad, A. Lavin, S. Purdy, and Z. Agha. "Unsupervised real-time anomaly detection for streaming data". In: *Neurocomputing* 262 (2017), pp. 134–147.

[AHP+05] P. K. Agarwal, S. Har-Peled, et al. "Geometric approximation via coresets". In: *Combinatorial and computational geometry* 52.1-30 (2005), p. 3.

[AHS85] D. Ackley, G. Hinton, and T. Sejnowski. "A Learning Algorithm for Boltzmann Machines". In: *Cognitive Science* 9 (1985), pp. 147–169.

[AHT07] Y. Altun, T. Hofmann, and I. Tsochantaridis. "Support Vector Machine Learning for Interdependent and Structured Output Spaces". In: *Predicting Structured Data*. Ed. by G. Bakir, T. Hofmann, B. Scholkopf, A. Smola, B. Taskar, and S. Vishwanathan. MIT Press, 2007.

[Ahu+21] K. Ahuja, J. Wang, A. Dhurandhar, K. Shanmugam, and K. R. Varshney. "Empirical or Invariant Risk Minimization? A Sample Complexity Perspective". In: *ICLR*. 2021.

[AI 19] AI Artists. *Creative Tools to Generate AI Art*. 2019.

[Air+08] E. Airoldi, D. Blei, S. Fienberg, and E. Xing. "Mixed-membership stochastic blockmodels". In: *JMLR* 9 (2008), pp. 1981–2014.

[Ait18] L. Aitchison. "A unified theory of adaptive stochastic gradient descent as Bayesian filtering". In: (2018). arXiv: 1807.07540 [stat.ML].

[Ait21] L. Aitchison. "A statistical theory of cold posteriors in deep neural networks". In: *ICLR*. 2021.

[Aka74] H. Akaike. "A new look at the statistical model identification". In: *IEEE Trans. on Automatic Control* 19.6 (1974).

[AKO18] S.-I. Amari, R. Karakida, and M. Oizumi. "Fisher Information and Natural Gradient Learning of Random Deep Networks". In: (2018). arXiv: 1808.07172 [cs.LG].

[AKZK19] B. Amos, V. Koltun, and J Zico Kolter. "The Limited Multi-Label Projection Layer". In: (2019). arXiv: 1906.08707 [cs.LG].

[AL+16] J. Ala-Luhtala, N. Whiteley, K. Heine, and R. Piche. "An Introduction to Twisted Particle Filters and Parameter Estimation in Non-linear State-space Models". In: *IEEE Trans. Signal Process.* 64.18 (2016), pp. 4875–4890.

[al21] M. A. et al. *Understanding Dataset Shift and Potential Remedies*. Tech. rep. Vector Institute, 2021.

[Ale+16] A. A. Alemi, I. Fischer, J. V. Dillon, and K. Murphy. "Deep Variational Information Bottleneck". In: *ICLR*. 2016.

[Ale+17] A. A. Alemi, I. S. Fischer, J. V. Dillon, and K. P. Murphy. "Deep Variational Information Bottleneck". In: *ArXiv* abs/1612.00410 (2017).

[Ale+18] A. A. Alemi, B. Poole, I. Fischer, J. V. Dillon, R. A. Saurous, and K. Murphy. "Fixing a broken ELBO". In: *ICML*. 2018.

[Alq21] P. Alquier. "User-friendly introduction to PAC-Bayes bounds". In: (2021). arXiv: 2110.11216 [stat.ML].

[Als+19] J. Alsing, T. Charnock, S. Feeney, and B. Wandelt. "Fast likelihood-free cosmology with neural density estimators and active learning". In: *Monthly Notices of the Royal Astronomical Society* 488.3 (2019), pp. 4440–4458.

[ALS20] B. Axelrod, Y. P. Liu, and A. Sidford. "Near-optimal approximate discrete and continuous submodular function minimization". In: *Proceedings of the Fourteenth Annual ACM-SIAM Symposium on Discrete Algorithms*. SIAM. 2020, pp. 837–853.

[AM07] R. P. Adams and D. J. C. MacKay. "Bayesian Online Changepoint Detection". In: (2007). arXiv: 0710.3742 [stat.ML].

[AM+16] M. Auger-Méthé, C. Field, C. M. Albertsen, A. E. Derocher, M. A. Lewis, I. D. Jonsen, and J. Mills Flemming. "State-space models' dirty little secrets: even simple linear Gaussian models can have estimation problems". en. In: *Sci. Rep.* 6 (2016), p. 26677.

[AM74] D. Andrews and C. Mallows. "Scale mixtures of Normal distributions". In: *J. of Royal Stat. Soc. Series B* 36 (1974), pp. 99–102.

[AM89] B. D. Anderson and J. B. Moore. *Optimal Control: Linear Quadratic Methods*. Prentice-Hall International, Inc., 1989.

[Ama09] S.-I. Amari. "α-Divergence Is Unique, Belonging to Both f-Divergence and Bregman Divergence Classes". In: *IEEE Trans. Inf. Theory* 55.11 (2009), pp. 4925–4931.

[Ama98] S Amari. "Natural Gradient Works Efficiently in Learning". In: *Neural Comput.* 10.2 (1998), pp. 251–276.

[Ame+19] S. Amershi, D. Weld, M. Vorvoreanu, A. Fourney, B. Nushi, P. Collisson, J. Suh, S. Iqbal, P. N. Bennett, K. Inkpen, et al. "Guidelines for human-AI interaction". In: *Proceedings of the 2019 chi conference on human factors in computing systems*. 2019, pp. 1–13.

[Ami01] E. Amir. "Efficient Approximation for Triangulation of Minimum Treewidth". In: *UAI*. 2001.

[AMJ18a] D. Alvarez-Melis and T. S. Jaakkola. "On the robustness of interpretability methods". In: *arXiv preprint arXiv:1806.08049* (2018).

[AMJ18b] D. Alvarez-Melis and T. S. Jaakkola. "Towards robust interpretability with self-explaining neural networks". In: *arXiv preprint arXiv:1806.07538* (2018).

[Amo+16] D. Amodei, C. Olah, J. Steinhardt, P. F. Christiano, J. Schulman, and D. Mané. "Concrete Problems in AI Safety". In: *CoRR* abs/1606.06565 (2016). arXiv: 1606.06565.

[Amo+18] B. Amos, L. Dinh, S. Cabi, T. Rothörl, S. G. Colmenarejo, A. Muldal, T. Erez, Y. Tassa, N. de Freitas, and M. Denil. "Learning Awareness Models". In: *ICLR*. 2018.

[Amo22] B. Amos. "Tutorial on amortized optimization for learning to optimize over continuous domains". In: (2022). arXiv: 2202.00665 [cs.LG].

[AMO88] R. K. Ahuja, T. L. Magnanti, and J. B. Orlin. "Network flows". In: (1988).

[Ana+14] A. Anandkumar, R. Ge, D. Hsu, S. M. Kakade, and M. Telgarsky. "Tensor Decompositions for Learning Latent Variable Models". In: *JMLR* 15 (2014), pp. 2773–2832.

[And+03] C. Andrieu, N. de Freitas, A. Doucet, and M. Jordan. "An introduction to MCMC for machine learning". In: *Machine Learning* 50 (2003), pp. 5–43.

[And+20] O. M. Andrychowicz et al. "Learning dexterous in-hand manipulation". In: *Int. J. Rob. Res.* 39.1 (2020), pp. 3–20.

[Ang+18] E. Angelino, N. Larus-Stone, D. Alabi, M. Seltzer, and C. Rudin. *Learning Certifiably Optimal Rule Lists for Categorical Data*. 2018. arXiv: 1704.01701 [stat.ML].

[Ang+20] C. Angermueller, D. Dohan, D. Belanger, R. Deshpande, K. Murphy, and L. Colwell. "Model-based reinforcement learning for biological sequence design". In: *ICLR*. 2020.

[Ang+21] A. N. Angelopoulos, S. Bates, E. J. Candès, M. I. Jordan, and L. Lei. "Learn then Test: Calibrating Predictive Algorithms to Achieve Risk Control". In: (2021). arXiv: 2110.01052 [cs.LG].

[Ang88] D. Angluin. "Queries and concept learning". In: *Machine learning* 2.4 (1988), pp. 319–342.

[Ani+18] R. Anirudh, J. J. Thiagarajan, B. Kailkhura, and T. Bremer. "An Unsupervised Approach to Solving Inverse Problems using Generative Adversarial Networks". In: (2018). arXiv: 1805.07281 [cs.CV].

[Ano19] Anonymous. "Neural Tangents: Fast and Easy Infinite Neural Networks in Python". In: (2019).

[Ant+22] I. Antonoglou, J. Schrittwieser, S. Ozair, T. K. Hubert, and D. Silver. "Planning in Stochastic Environments with a Learned Model". In: *ICLR*. 2022.

[AO03] J.-H. Ahn and J.-H. Oh. "A Constrained EM Algorithm for Principal Component Analysis". In: *Neural Computation* 15 (2003), pp. 57–65.

[AOM17] M. G. Azar, I. Osband, and R. Munos. "Minimax Regret Bounds for Reinforcement Learning". In: *ICML*. 2017, pp. 263–272.

[AP08] J. D. Angrist and J.-S. Pischke. *Mostly harmless econometrics: An empiricist's companion*. Princeton university press, 2008.

[AP09] J. Angrist and J.-S. Pischke. *Mostly Harmless Econometrics*. 2009.

[AP19] M. Abadi and G. D. Plotkin. "A simple differentiable programming language". In: *Proceedings of the ACM on Programming Languages* 4.POPL (2019), pp. 1–28.

[Ara+09] A. Aravkin, B. Bell, J. Burke, and G. Pillonetto. *An L1-Laplace Robust Kalman Smoother*. Tech. rep. U. Washington, 2009.

[Ara10] A. Aravkin. *Student's t Kalman Smoother*. Tech. rep. U. Washington, 2010.

[Ara+17] A. Aravkin, J. V. Burke, L. Ljung, A. Lozano, and G. Pillonetto. "Generalized Kalman smoothing: Modeling and algorithms". In: *Automatica* 86 (2017), pp. 63–86.

[Arb+18] M. Arbel, D. Sutherland, M. Bińkowski, and A. Gretton. "On gradient regularizers for MMD GANs". In: *Advances in neural information processing systems*. 2018, pp. 6700–6710.

[Arj+19] M. Arjovsky, L. Bottou, I. Gulrajani, and D. Lopez-Paz. "Invariant Risk Minimization". In: (2019). arXiv: 1907.02893 [stat.ML].

[Arj+20] M. Arjovsky, L. Bottou, I. Gulrajani, and D. Lopez-Paz. *Invariant Risk Minimization*. 2020. arXiv: 1907.02893 [stat.ML].

[Arn+10] C. W. Arnold, S. M. El-Saden, A. A. Bui, and R. Taira. "Clinical case-based retrieval using latent topic analysis". In: *AMIA annual symposium proceedings*. Vol. 2010. American Medical Informatics Association. 2010, p. 26.

[Aro+13] S. Arora, R. Ge, Y. Halpern, D. Mimno, A. Moitra, D. Sontag, Y. Wu, and M. Zhu. "A Practical Algorithm for Topic Modeling with Provable Guarantees". In: *ICML*. 2013.

[Aro+19] S. Arora, S. S. Du, W. Hu, Z. Li, R. Salakhutdinov, and R. Wang. "On Exact Computation with an Infinitely Wide Neural Net". In: (2019). arXiv: 1904.11955 [cs.LG].

[Aro+21] R. Arora et al. *Theory of deep learning*. 2021.

[ARS13] N. S. Arora, S. Russell, and E. Sudderth. "NET-VISA: Network Processing Vertically Integrated Seismic AnalysisNET-VISA: Network Processing Vertically Integrated Seismic Analysis". In: *Bull. Seismol. Soc. Am.* 103.2A (2013), pp. 709–729.

[Aru+02] M. Arulampalam, S. Maskell, N. Gordon, and T. Clapp. "A Tutorial on Particle Filters for Online Nonlinear/Non-Gaussian Bayesian Tracking". In: *IEEE Trans. on Signal Processing* 50.2 (2002), pp. 174–189.

[Aru+17] K. Arulkumaran, M. P. Deisenroth, M. Brundage, and A. A. Bharath. "A Brief Survey of Deep Reinforcement Learning". In: *IEEE Signal Processing Magazine, Special Issue on Deep Learning for Image Understanding* (2017).

[AS17] A. Achille and S. Soatto. "On the Emergence of Invariance and Disentangling in Deep Representations". In: (2017). arXiv: 1706.01350 [cs.LG].

[AS18] A. Achille and S. Soatto. "On the Emergence of Invariance and Disentangling in Deep Representations". In: *JMLR* 18 (2018), pp. 1–34.

[AS66] S. M. Ali and S. D. Silvey. "A General Class of Coefficients of Divergence of One Distribution from Another". In: *J. R. Stat. Soc. Series B Stat. Methodol.* 28.1 (1966), pp. 131–142.

[Asa00] C. Asavathiratham. "The Influence Model: A Tractable Representation for the Dynamics of Networked Markov Chains". PhD thesis. MIT, Dept. EECS, 2000.

[ASD20] A. Agrawal, D. Sheldon, and J. Domke. "Advances in Black-Box VI: Normalizing Flows, Importance Weighting, and Optimization". In: *NIPS*. 2020.

[ASM17] A. Azuma, M. Shimbo, and Y. Matsumoto. "An Algebraic Formalization of Forward and Forward-backward Algorithms". In: (2017). arXiv: 1702.06941 [cs.LG].

[ASN20] R. Agarwal, D. Schuurmans, and M. Norouzi. "An Optimistic Perspective on Offline Reinforcement Learning". In: *ICML*. 2020.

[ASS19] I. Andrews, J. H. Stock, and L. Sun. "Weak Instruments in Instrumental Variables Regression: Theory and Practice". In: *Annual Review of Economics* 11.1 (2019).

[AT08] C. Andrieu and J. Thoms. "A tutorial on adaptive MCMC". In: *Statistical Computing* 18 (2008), pp. 343–373.

[AT20] E. Agustsson and L. Theis. "Universally Quantized Neural Compression". 2020.

[Att00] H. Attias. "A Variational Bayesian Framework for Graphical Models". In: *NIPS-12*. 2000.

[Att03] H. Attias. "Planning by Probabilistic Inference". In: *AI-Stats*. 2003.

[Aue12] J. E. Auerbach. "Automated evolution of interesting images". In: *Artificial Life 13*. 2012.

[AWR17] J. Altschuler, J. Weed, and P. Rigollet. "Near-linear time approximation algorithms for optimal transport via Sinkhorn iteration". In: *arXiv preprint arXiv:1705.09634* (2017).

[AXK17] B. Amos, L. Xu, and J. Z. Kolter. "Input convex neural networks". In: *International Conference on Machine Learning*. PMLR. 2017, pp. 146–155.

[AY19] B. Amos and D. Yarats. "The Differentiable Cross-Entropy Method". In: (2019). arXiv: 1909.12830 [cs.LG].

[AZ17] S. Arora and Y. Zhang. "Do gans actually learn the distribution? an empirical study". In: *arXiv preprint arXiv:1706.08224* (2017).

[Aze+20] E. M. Azevedo, A. Deng, J. L. Montiel Olea, J. Rao, and E. G. Weyl. "A/B Testing with Fat Tails". In: *J. Polit. Econ.* (2020), pp. 000–000.

[Azi+15] H. Azizpour, A. Sharif Razavian, J. Sullivan, A. Maki, and S. Carlsson. "From generic to specific deep representations for visual recognition". In: *Proceedings of the IEEE conference on computer vision and pattern recognition workshops*. 2015, pp. 36–45.

[BA03] D. Barber and F. Agakov. "The IM Algorithm: A Variational Approach to Information Maximization". In: *NIPS*. NIPS'03. MIT Press, 2003, pp. 201–208.

[BA05] W. Bechtel and A. Abrahamsen. "Explanation: A mechanist alternative". In: *Studies in History and Philosophy of Science Part C: Studies in History and Philosophy of Biological and Biomedical Sciences* 36.2 (2005), pp. 421–441.

[Bac09] F. Bach. "High-Dimensional Non-Linear Variable Selection through Hierarchical Kernel Learning". In: (2009). arXiv: 0909.0844 [cs.LG].

[Bac+13] F. Bach et al. "Learning with Submodular Functions: A Convex Optimization Perspective". In: *Foundations and Trends® in Machine Learning* 6.2-3 (2013), pp. 145–373.

[Bac+15a] S. Bach, A. Binder, G. Montavon, F. Klauschen, K.-R. Müller, and W. Samek. "On pixel-wise explanations for non-linear classifier decisions by layer-wise relevance propagation". In: *PloS one* 10.7 (2015), e0130140.

[Bac+15b] S. H. Bach, M. Broecheler, B. Huang, and L. Getoor. "Hinge-Loss Markov Random Fields and Probabilistic Soft Logic". In: (2015). arXiv: 1505.04406 [cs.LG].

[Bac+18] E. Bach, J. Dusart, L. Hellerstein, and D. Kletenik. "Submodular goal value of Boolean functions". In: *Discrete Applied Mathematics* 238 (2018), pp. 1–13.

[Bad+18] M. A. Badgeley, J. R. Zech, L. Oakden-Rayner, B. S. Glicksberg, M. Liu, W. Gale, M. V. McConnell, B. Percha, T. M. Snyder, and J. T. Dudley. "Deep Learning Predicts Hip Fracture using Confounding Patient and Healthcare Variables". In: *CoRR* abs/1811.03695 (2018). arXiv: 1811.03695.

[Bah+20] Y. Bahri, J. Kadmon, J. Pennington, S. Schoenholz, J. Sohl-Dickstein, and S. Ganguli. "Statistical Mechanics of Deep Learning". In: *Annu. Rev. Condens. Matter Phys.* (2020).

[Bai+15] R. Bairi, R. Iyer, G. Ramakrishnan, and J. Bilmes. "Summarization of multi-document topic hierarchies using submodular mixtures". In: *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*. 2015, pp. 553–563.

[Bai95] L. C. Baird. "Residual Algorithms: Reinforcement Learning with Function Approximation". In: *ICML*. 1995, pp. 30–37.

[Bak+17] J. Baker, P. Fearnhead, E. B. Fox, and C. Nemeth. "Control Variates for Stochastic Gradient MCMC". In: (2017). arXiv: 1706.05439 [stat.CO].

[Bal17] S. Baluja. "Learning deep models of optimization landscapes". In: *IEEE Symposium Series on Computational Intelligence (SSCI)* (2017).

[Bal+18] D. Balduzzi, S. Racaniere, J. Martens, J. Foerster, K. Tuyls, and T. Graepel. "The mechanics of n-player differentiable games". In: *International Conference on Machine Learning*. PMLR. 2018, pp. 354–363.

[Ban+05] A. Banerjee, I. S. Dhillon, J. Ghosh, and S. Sra. "Clustering on the unit hypersphere using von Mises-Fisher distributions". In: *JMLR*. 2005, pp. 1345–1382.

[Ban06] A. Banerjee. "On bayesian bounds". In: *ICML*. 2006, pp. 81–88.

[Ban+18] A. Bansal, S. Ma, D. Ramanan, and Y. Sheikh. "Recycle-gan: Unsupervised video retargeting". In: *Proceedings of the European conference on computer vision (ECCV)*. 2018, pp. 119–135.

[Bao+22] H. Bao, L. Dong, S. Piao, and F. Wei. "BEiT: BERT Pre-Training of Image Transformers". In: *International Conference on Learning Representations*. 2022.

[Baq+16] P. Baqué, T. Bagautdinov, F. Fleuret, and P. Fua. "Principled parallel mean-field inference for discrete random fields". In: *CVPR*. 2016, pp. 5848–5857.

[Bar17] D. Barber. *Evolutionary Optimization as a Variational Method*. 2017.

[Bas+01] S. Basu, T. Choudhury, B. Clarkson, and A. Pentland. *Learning Human Interactions with the Influence Model*. Tech. rep. 539. MIT Media Lab, 2001.

[Bat+12] D. Batra, P. Yadollahpour, A. Guzman-Rivera, and G. Shakhnarovich. "Diverse M-Best Solutions in Markov Random Fields". In: *ECCV*. Springer Berlin Heidelberg, 2012, pp. 1–16.

[Bau+17] D. Bau, B. Zhou, A. Khosla, A. Oliva, and A. Torralba. "Network Dissection: Quantifying Inter-

pretability of Deep Visual Representations". In: *Computer Vision and Pattern Recognition*. 2017.

[Bau+18] D. Bau, J.-Y. Zhu, H. Strobelt, B. Zhou, J. B. Tenenbaum, W. T. Freeman, and A. Torralba. "Gan dissection: Visualizing and understanding generative adversarial networks". In: *arXiv preprint arXiv:1811.10597* (2018).

[Bau+20] D. Bau, J.-Y. Zhu, H. Strobelt, A. Lapedriza, B. Zhou, and A. Torralba. "Understanding the role of individual units in a deep neural network". In: *Proceedings of the National Academy of Sciences* (2020).

[Bau+70] L. E. Baum, T. Petrie, G. Soules, and N. Weiss. "A maximization technique occuring in the statistical analysis of probabalistic functions in Markov chains". In: *The Annals of Mathematical Statistics* 41 (1970), pp. 164–171.

[Bau74] B. G. Baumgart. "Geometric modeling for computer vision." In: 1974.

[Bax00] J Baxter. "A Model of Inductive Bias Learning". In: *JAIR* (2000).

[Bay+15] A. G. Baydin, B. A. Pearlmutter, A. A. Radul, and J. M. Siskind. "Automatic differentiation in machine learning: a survey". In: (2015). arXiv: 1502.05767 [cs.SC].

[BB12] J. Bergstra and Y. Bengio. "Random Search for Hyper-Parameter Optimization". In: *JMLR* 13 (2012), pp. 281–305.

[BB15a] A. Bendale and T. Boult. "Towards Open World Recognition". In: *CVPR*. 2015.

[BB15b] J. Bornschein and Y. Bengio. "Reweighted Wake-Sleep". In: *ICLR*. 2015.

[BB17] J. Bilmes and W. Bai. "Deep Submodular Functions". In: *Arxiv* abs/1701.08939 (2017).

[BB18] W. Bai and J. Bilmes. "Greed is Still Good: Maximizing Monotone Submodular+Supermodular (BP) Functions". In: *International Conference on Machine Learning (ICML)*. http://proceedings.mlr.press/v80/bai18a.html. Stockholm, Sweden, 2018.

[BBH12] C. Blundell, J. Beck, and K. A. Heller. "Modelling reciprocating relationships with Hawkes processes". In: *Advances in Neural Information Processing Systems*. 2012, pp. 2600–2608.

[BBM07] A. Banerjee, S. Basu, and S. Merugu. "Multi-way Clustering on Relation Graphs". In: *Proc. SIAM Intl. Conf. on Data Mining (SDM)*. 2007.

[BBM10] N. Bhatnagar, C. Bogdanov, and E. Mossel. *The Computational Complexity of Estimating Convergence Time*. Tech. rep. arxiv, 2010.

[BBS09] J. O. Berger, J. M. Bernardo, and D. Sun. "The Formal Definition of Reference Priors". In: *Ann. Stat.* 37.2 (2009), pp. 905–938.

[BBS95] A. G. Barto, S. J. Bradtke, and S. P. Singh. "Learning to act using real-time dynamic programming". In: *AIJ* 72.1 (1995), pp. 81–138.

[BBV11a] R. Benassi, J. Bect, and E. Vazquez. "Bayesian optimization using sequential Monte Carlo". In: (2011). arXiv: 1111.4802 [math.OC].

[BBV11b] R. Benassi, J. Bect, and E. Vazquez. "Robust Gaussian Process-Based Global Optimization Using a Fully Bayesian Expected Improvement Criterion". en. In: *Intl. Conf. on Learning and Intelligent Optimization (LION)*. 2011, pp. 176–190.

[BC07] D. Barber and S. Chiappa. "Unified inference for variational Bayesian linear Gaussian state space models". In: *NIPS*. 2007.

[BC08] M. Bădoiu and K. L. Clarkson. "Optimal coresets for balls". In: *Computational Geometry* 40.1 (2008), pp. 14–22.

[BC14] J. Ba and R. Caruana. "Do Deep Nets Really Need to be Deep?" In: *Advances in Neural Information Processing Systems* 27 (2014).

[BC17] D. Beck and T. Cohn. "Learning Kernels over Strings using Gaussian Processes". In: *Proceedings of the Eighth International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*. Vol. 2. 2017, pp. 67–73.

[BC89] D. P. Bertsekas and D. A. Castanon. "The auction algorithm for the transportation problem". In: *Annals of Operations Research* 20.1 (1989), pp. 67–96.

[BC93] B. M. Bell and F. W. Cathey. "The iterated Kalman filter update as a Gauss-Newton method". In: *IEEE Trans. Automat. Contr.* 38.2 (Feb. 1993), pp. 294–297.

[BC94] P. Baldi and Y. Chauvin. "Smooth online learning algorithms for Hidden Markov Models". In: *Neural Computation* 6 (1994), pp. 305–316.

[BC95] S. Baluja and R. Caruana. "Removing the Genetics from the Standard Genetic Algorithm". In: *ICML*. 1995, pp. 38–46.

[BCF10] E. Brochu, V. M. Cora, and N. de Freitas. "A Tutorial on Bayesian Optimization of Expensive Cost Functions, with Application to Active User Modeling and Hierarchical Reinforcement Learning". In: (2010). arXiv: 1012.2599 [cs.LG].

[BCH20] M. Briers, M. Charalambides, and C. Holmes. "Risk scoring calculation for the current NHSx contact tracing app". In: (2020). arXiv: 2005.11057 [cs.CY].

[BCJ20] A. Buchholz, N. Chopin, and P. E. Jacob. "Adaptive Tuning Of Hamiltonian Monte Carlo Within Sequential Monte Carlo". In: *Bayesian Anal.* (2020).

[BCN18] L. Bottou, F. E. Curtis, and J. Nocedal. "Optimization Methods for Large-Scale Machine Learning". In: *SIAM Rev.* 60.2 (2018), pp. 223–311.

[BCNM06] C. Buciluǎ, R. Caruana, and A. Niculescu-Mizil. "Model compression". In: *Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining*. 2006, pp. 535–541.

[BCV13] Y. Bengio, A. Courville, and P. Vincent. "Representation learning: A review and new perspectives". In: *IEEE transactions on pattern analysis and machine intelligence* 35.8 (2013), pp. 1798–1828.

[BD+10] S. Ben-David, J. Blitzer, K. Crammer, A. Kulesza, F. Pereira, and J. W. Vaughan. "A theory of learning from different domains". In: *Mach. Learn.* 79.1 (May 2010), pp. 151–175.

[BD11] A. Bhattacharya and D. B. Dunson. "Simplex factor models for multivariate unordered categorical data". In: *JASA* (2011).

[BD87] G. Box and N. Draper. *Empirical Model-Building and Response Surfaces*. Wiley, 1987.

Author: Kevin P. Murphy. (C) MIT Press. CC-BY-NC-ND license

[BD92] D. Bayer and P. Diaconis. "Trailing the dovetail shuffle to its lair". In: *The Annals of Applied Probability* 2.2 (1992), pp. 294–313.

[BD97] S. Baluja and S. Davies. "Using Optimal Dependency-Trees for Combinatorial Optimization: Learning the Structure of the Search Space". In: *ICML*. 1997.

[BDM09] R. Burkard, M. Dell'Amico, and S. Martello. *Assignment Problems*. SIAM, 2009.

[BDM10] M. Briers, A. Doucet, and S. Maskel. "Smoothing algorithms for state-space models". In: *Annals of the Institute of Statistical Mathematics* 62.1 (2010), pp. 61–89.

[BDM17] M. G. Bellemare, W. Dabney, and R. Munos. "A Distributional Perspective on Reinforcement Learning". In: *ICML*. 2017.

[BDM18] N. Brosse, A. Durmus, and E. Moulines. "The promises and pitfalls of Stochastic Gradient Langevin Dynamics". In: *NIPS*. 2018.

[BDS18] A. Brock, J. Donahue, and K. Simonyan. "Large Scale GAN Training for High Fidelity Natural Image Synthesis". In: (2018). arXiv: 1809.11096 [cs.LG].

[Bea03] M. Beal. "Variational Algorithms for Approximate Bayesian Inference". PhD thesis. Gatsby Unit, 2003.

[Bea19] M. A. Beaumont. "Approximate Bayesian Computation". In: *Annual Review of Statistics and Its Application* 6.1 (2019), pp. 379–403.

[Béd08] M. Bédard. "Optimal acceptance rates for Metropolis algorithms: Moving beyond 0.234". In: *Stochastic Process. Appl.* 118.12 (2008), pp. 2198–2222.

[Beh+19] J. Behrmann, W. Grathwohl, R. T. Q. Chen, D. Duvenaud, and J.-H. Jacobsen. "Invertible Residual Networks". In: *ICML*. 2019.

[Bel03] A. J. Bell. "The co-information lattice". In: *ICA conference*. 2003.

[Bel+16] M. G. Bellemare, S. Srinivasan, G. Ostrovski, T. Schaul, D. Saxton, and R. Munos. "Unifying Count-Based Exploration and Intrinsic Motivation". In: *NIPS*. 2016.

[Bel+18] M. I. Belghazi, A. Baratin, S. Rajeshwar, S. Ozair, Y. Bengio, A. Courville, and D. Hjelm. "Mutual Information Neural Estimation". In: *ICML*. Ed. by J. Dy and A. Krause. Vol. 80. Proceedings of Machine Learning Research. PMLR, 2018, pp. 531–540.

[Bel+19] D. Belanger, S. Vora, Z. Mariet, R. Deshpande, D. Dohan, C. Angermueller, K. Murphy, O. Chapelle, and L. Colwell. "Biological Sequence Design using Batched Bayesian Optimization". In: *NIPS workshop on ML for the sciences*. 2019.

[Bel94] B. M. Bell. "The Iterated Kalman Smoother as a Gauss–Newton Method". In: *SIAM J. Optim.* 4.3 (Aug. 1994), pp. 626–636.

[Ben13] Y. Bengio. "Estimating or Propagating Gradients Through Stochastic Neurons". In: (2013). arXiv: 1305.2982 [cs.LG].

[Ben+20] K. Benidis et al. "Neural forecasting: Introduction and literature overview". In: (2020). arXiv: 2004.10240 [cs.LG].

[Bén+21] C. Bénard, G. Biau, S. Veiga, and E. Scornet. "Interpretable random forests via rule extraction". In: *International Conference on Artificial Intelligence and Statistics*. PMLR. 2021, pp. 937–945.

[Ben+21a] Y. Bengio, T. Deleu, E. J. Hu, S. Lahlou, M. Tiwari, and E. Bengio. "GFlowNet Foundations". In: (Nov. 2021). arXiv: 2111.09266 [cs.LG].

[Ben+21b] G. W. Benton, W. J. Maddox, S. Lotfi, and A. G. Wilson. "Loss Surface Simplexes for Mode Connecting Volumes and Fast Ensembling". In: *ICML*. 2021.

[Ber05] J. M. Bernardo. "Reference Analysis". In: *Handbook of Statistics*. Ed. by D. K. Dey and C. R. Rao. Vol. 25. Elsevier, 2005, pp. 17–90.

[Ber15] D. Bertsekas. *Convex Optimization Algorithms*. Athena Scientific, 2015.

[Ber16] D. Bertsekas. *Nonlinear Programming*. Third. Athena Scientific, 2016.

[Ber+18] R. van den Berg, L. Hasenclever, J. M. Tomczak, and M. Welling. "Sylvester normalizing flows for variational inference". In: *AISTATS*. 2018.

[Ber+19] H. Berard, G. Gidel, A. Almahairi, P. Vincent, and S. Lacoste-Julien. "A Closer Look at the Optimization Landscapes of Generative Adversarial Networks". In: *International Conference on Learning Representations*. 2019.

[Ber19] D. Bertsekas. *Reinforcement learning and optimal control*. Athena Scientific, 2019.

[Ber+21] J. Berner, P. Grohs, G. Kutyniok, and P. Petersen. "The Modern Mathematics of Deep Learning". In: (2021). arXiv: 2105.04026 [cs.LG].

[Ber85] J. Berger. "Bayesian Salesmanship". In: *Bayesian Inference and Decision Techniques with Applications: Essays in Honor of Bruno deFinetti*. Ed. by P. K. Goel and A. Zellner. North-Holland, 1985.

[Ber96] J. Bertoin. *Lévy processes*. Vol. 121. Cambridge university press Cambridge, 1996.

[Ber97] D. Bertsekas. *Parallel and Distribution Computation: Numerical Methods*. Athena Scientific, 1997.

[Ber99] A. Berchtold. "The double chain Markov model". In: *Comm. Stat. Theor. Methods* 28 (1999), pp. 2569–2589.

[Bes75] J. Besag. "Statistical analysis of non-lattice data". In: *The Statistician* 24 (1975), pp. 179–196.

[Bet13] M. Betancourt. "A General Metric for Riemannian Manifold Hamiltonian Monte Carlo". In: *Geometric Science of Information*. Springer Berlin Heidelberg, 2013, pp. 327–334.

[Bet17] M. Betancourt. "A Conceptual Introduction to Hamiltonian Monte Carlo". In: (2017). arXiv: 1701.02434 [stat.ME].

[BFH75] Y. Bishop, S. Fienberg, and P. Holland. *Discrete Multivariate Analysis: Theory and Practice*. MIT Press, 1975.

[BFY20] T. D. Barfoot, J. R. Forbes, and D. Yoon. "Exactly Sparse Gaussian Variational Inference with Application to Derivative-Free Batch Nonlinear State Estimation". In: *Intl. J. of Robotics Research* (2020).

[BG06] M. Beal and Z. Ghahramani. "Variational Bayesian Learning of Directed Graphical Models with Hidden Variables". In: *Bayesian Analysis* 1.4 (2006).

[BG13] M. J. Betancourt and M. Girolami. "Hamiltonian Monte Carlo for Hierarchical Models". In: (2013). arXiv: 1312.0906 [stat.ME].

[BG73] A. Björck and G. H. Golub. "Numerical methods for computing angles between linear subspaces". In: *Mathematics of computation* 27.123 (1973), pp. 579–594.

[BG96] A. Becker and D. Geiger. "A sufficiently fast algorithm for finding close to optimal junction trees". In: *UAI*. 1996.

[BGHM17] J. Boyd-Graber, Y. Hu, and D. Mimno. "Applications of Topic Models". In: *Foundations and Trends® in Information Retrieval* 11.2-3 (2017), pp. 143–296.

[BGM17] J Ba, R Grosse, and J Martens. "Distributed Second-Order Optimization using Kronecker-Factored Approximations". In: *ICLR*. openreview.net, 2017.

[BGS16] Y. Burda, R. Grosse, and R. Salakhutdinov. "Importance Weighted Autoencoders". In: *ICLR*. 2016.

[BGT93] C. Berrou, A. Glavieux, and P. Thitimajashima. "Near Shannon limit error-correcting coding and decoding: Turbo codes". In: *Proc. IEEE Intl. Comm. Conf.* (1993).

[BH11] M.-F. Balcan and N. J. Harvey. "Learning submodular functions". In: *Proceedings of the forty-third annual ACM symposium on Theory of computing*. 2011, pp. 793–802.

[BH20] M. T. Bahadori and D. Heckerman. "Debiasing Concept-based Explanations with Causal Analysis". In: *International Conference on Learning Representations*. 2020.

[BH92] D. Barry and J. A. Hartigan. "Product partition models for change point problems". In: *Annals of statistics* 20 (1992), pp. 260–279.

[Bha+19] A. Bhadra, J. Datta, N. G. Polson, and B. T. Willard. "Lasso Meets Horseshoe: a survey". In: *Bayesian Anal.* 34.3 (2019), pp. 405–427.

[Bha+21] K. Bhatia, N. Kuang, Y. Ma, and Y. Wang. *Statistical and computational tradeoffs in variational Bayes: a case study of inferential model selection*. Tech. rep. 2021.

[Bha+22] K. Bhatia, N. L. Kuang, Y.-A. Ma, and Y. Wang. "Statistical and Computational Trade-offs in Variational Inference: A Case Study in Inferential Model Selection". In: (July 2022). arXiv: 2207.11208 [stat.ML].

[BHB19] P. Bachman, R. D. Hjelm, and W. Buchwalter. *Learning Representations by Maximizing Mutual Information Across Views*. 2019. arXiv: 1906.00910 [cs.LG].

[BHC19] M. Binkowski, D. Hjelm, and A. Courville. "Batch weight for domain adaptation with mass shift". In: *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 2019, pp. 1844–1853.

[BHPI02] M. Bādoiu, S. Har-Peled, and P. Indyk. "Approximate clustering via core-sets". In: *Proceedings of the thiry-fourth annual ACM symposium on Theory of computing*. 2002, pp. 250–257.

[BHW16] P. G. Bissiri, C. Holmes, and S. Walker. "A General Framework for Updating Belief Distributions". In: *JRSSB* 78.5 (2016), 1103–1130.

[Bic09] D. Bickson. "Gaussian Belief Propagation: Theory and Application". PhD thesis. Hebrew University of Jerusalem, 2009.

[Bie06] G. J. Bierman. *Factorization Methods for Discrete Sequential Estimation (Dover Books on Mathematics)*. en. Illustrated edition. Dover Publications, 2006.

[Big+11] B. Biggio, G. Fumera, I. Pillai, and F. Roli. "A survey and experimental evaluation of image spam filtering techniques". In: *Pattern recognition letters* 32.10 (2011), pp. 1436–1446.

[Bil01] J. A. Bilmes. *Graphical Models and Automatic Speech Recognition*. Tech. rep. UWEETR-2001-0005. Univ. Washington, Dept. of Elec. Eng., 2001.

[Bil22] J. Bilmes. "Submodularity In Machine Learning and Artificial Intelligence". In: (2022). arXiv: 2202.00132 [cs.LG].

[Biń+18] M. Bińkowski, D. J. Sutherland, M. Arbel, and A. Gretton. "Demystifying MMD GANs". In: *ICLR*. 2018.

[Bin+18] M. Binkowski, D. J. Sutherland, M. Arbel, and A. Gretton. "Demystifying MMD GANs". In: *International Conference on Learning Representations*. 2018.

[Bin+19] E. Bingham, J. P. Chen, M. Jankowiak, F. Obermeyer, N. Pradhan, T. Karaletsos, R. Singh, P. Szerlip, P. Horsfall, and N. D. Goodman. "Pyro: Deep Universal Probabilistic Programming". In: *JMLR* 20.28 (2019), pp. 1–6.

[Biń+19] M. Bińkowski, J. Donahue, S. Dieleman, A. Clark, E. Elsen, N. Casagrande, L. C. Cobo, and K. Simonyan. "High Fidelity Speech Synthesis with Adversarial Networks". In: *International Conference on Learning Representations*. 2019.

[Bin+97] J. Binder, D. Koller, S. J. Russell, and K. Kanazawa. "Adaptive Probabilistic Networks with Hidden Variables". In: *Machine Learning* 29 (1997), pp. 213–244.

[Bis06] C. Bishop. *Pattern recognition and machine learning*. Springer, 2006.

[Bis99] C. Bishop. "Bayesian PCA". In: *NIPS*. 1999.

[Bit16] S. Bitzer. *The UKF exposed: How it works, when it works and when it's better to sample*. 2016.

[Bit+21] J. Bitterwolf, A. Meinke, M. Augustin, and M. Hein. "Revisiting out-of-distribution detection: A simple baseline is surprisingly effective". In: *ICML Workshop on Uncertainty in Deep Learning (UDL)*. 2021.

[BJ05] F. Bach and M. Jordan. *A probabilistic interpretation of canonical correlation analysis*. Tech. rep. 688. U. C. Berkeley, 2005.

[BJ+06] D. M. Blei, M. I. Jordan, et al. "Variational inference for Dirichlet process mixtures". In: *Bayesian analysis* 1.1 (2006), pp. 121–143.

[BJ06] W. Buntine and A. Jakulin. "Discrete Component Analysis". In: *Subspace, Latent Structure and Feature Selection: Statistical and Optimization Perspectives Workshop*. 2006.

[BJV97] J. S. D. Bonet, C. L. I. Jr., and P. A. Viola. "MIMIC: Finding Optima by Estimating Probability Densities". In: *NIPS*. MIT Press, 1997, pp. 424–430.

[BK01] Y. Boykov and V. Kolmogorov. "An Experimental Comparison of Min-Cut/Max-Flow Algorithms for Energy Minimization in Computer Vision". In: *Third International Workshop on Energy Minimization Methods in Computer Vision and Pattern Recognition*. 2001.

[BK10] R. Bardenet and B. Kegl. "Surrogating the surrogate: accelerating Gaussian-process-based global optimization with a mixture cross-entropy algorithm". In: *ICML*. 2010.

[BK15] D. Belanger and S. Kakade. "A Linear Dynamical System Model for Text". en. In: *ICML*. 2015, pp. 833–842.

[BK19] M. Bonvini and E. H. Kennedy. "Sensitivity Analysis via the Proportion of Unmeasured Confounding". In: *arXiv e-prints*, arXiv:1912.02793 (Dec. 2019), arXiv:1912.02793. arXiv: 1912.02793 [stat.ME].

[BKB17] O. Bastani, C. Kim, and H. Bastani. "Interpreting blackbox models via model extraction". In: *arXiv preprint arXiv:1705.08504* (2017).

[BKH16] J. L. Ba, J. R. Kiros, and G. E. Hinton. "Layer Normalization". In: (2016). arXiv: 1607.06450 [stat.ML].

[BKJ13] T. Broderick, B. Kulis, and M. Jordan. "MAD-Bayes: MAP-based asymptotic derivations from Bayes". In: *International Conference on Machine Learning*. PMLR. 2013, pp. 226–234.

[BKM16] D. M. Blei, A. Kucukelbir, and J. D. McAuliffe. "Variational Inference: A Review for Statisticians". In: *JASA* (2016).

[BKS19] A. Bennett, N. Kallus, and T. Schnabel. "Deep Generalized Method of Moments for Instrumental Variable Analysis". In: *Advances in Neural Information Processing Systems*. 2019, pp. 3564–3574.

[BL06] D. Blei and J. Lafferty. "Dynamic topic models". In: *ICML*. 2006, pp. 113–120.

[BL07a] C. M. Bishop and J. Lasserre. "Generative or discriminative? Getting the best of both worlds". In: *Bayesian Statistics 8*. 2007.

[BL07b] D. Blei and J. Lafferty. "A Correlated Topic Model of "Science"". In: *Annals of Applied Stat.* 1.1 (2007), pp. 17–35.

[BLC18] A. J. Bose, H. Ling, and Y. Cao. "Adversarial Contrastive Estimation". In: *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. 2018, pp. 1021–1032.

[Ble12] D. M. Blei. "Probabilistic topic models". In: *Commun. ACM* 55.4 (2012), pp. 77–84.

[Ble17] D. Blei. *Variational inference: foundations and innovations (Lecture)*. Simons Institute Lecture. 2017.

[BLM16] S. Boucheron, G. Lugosi, and P. Massart. *Concentration Inequalities: A Nonasymptotic Theory of Independence*. Oxford University Press, 2016.

[Blo+16] A. Bloniarz, H. Liu, C.-H. Zhang, J. S. Sekhon, and B. Yu. "Lasso adjustments of treatment effect estimates in randomized experiments". In: *Proceedings of the National Academy of Sciences* 113.27 (2016), pp. 7383–7390.

[BLS11] G. Blanchard, G. Lee, and C. Scott. "Generalizing from several related classification tasks to a new unlabeled sample". In: *NIPS*. 2011.

[BLS17] J. Ballé, V. Laparra, and E. P. Simoncelli. "End-to-end Optimized Image Compression". In: *ICLR*. 2017.

[Blu+15] C. Blundell, J. Cornebise, K. Kavukcuoglu, and D. Wierstra. "Weight Uncertainty in Neural Networks". In: *ICML*. 2015.

[BM+18] G. Barth-Maron, M. W. Hoffman, D. Budden, W. Dabney, D. Horgan, T. B. Dhruva, A. Muldal, N. Heess, and T. Lillicrap. "Distributed Distributional Deterministic Policy Gradients". In: *ICLR*. 2018.

[BM19] Y. Blau and T. Michaeli. "Rethinking Lossy Compression: The Rate-Distortion-Perception Tradeoff". In: *ICML*. 2019.

[BM+73] D. Blackwell, J. B. MacQueen, et al. "Ferguson distributions via Pólya urn schemes". In: *The annals of statistics* 1.2 (1973), pp. 353–355.

[BMK20] Z. Borsos, M. Mutny, and A. Krause. "Coresets via Bilevel Optimization for Continual Learning and Streaming". In: *Advances in Neural Information Processing Systems* 33 (2020).

[BMP19] A. Brunel, D. Mazza, and M. Pagani. "Backpropagation in the Simply Typed Lambda-Calculus with Linear Negation". In: *Proc. ACM Program. Lang.* 4.POPL (2019).

[BMR97] J. Binder, K. Murphy, and S. Russell. "Space-efficient inference in dynamic probabilistic networks". In: *IJCAI*. 1997.

[BMS11] S. Bubeck, R. Munos, and G. Stoltz. "Pure Exploration in Finitely-armed and Continuous-armed Bandits". In: *Theoretical Computer Science* 412.19 (2011), pp. 1832–1852.

[BNJ03a] D. Blei, A. Ng, and M. Jordan. "Latent Dirichlet allocation". In: *JMLR* 3 (2003), pp. 993–1022.

[BNJ03b] D. M. Blei, A. Y. Ng, and M. I. Jordan. "Latent dirichlet allocation". In: *JMLR* 3 (2003), pp. 993–1022.

[BO14] J. Bayer and C. Osendorfer. "Learning Stochastic Recurrent Networks". In: *Workshop on Advances in Variational Inference*. 2014.

[Boe+05] P.-T. de Boer, D. P. Kroese, S. Mannor, and R. Y. Rubinstein. "A Tutorial on the Cross-Entropy Method". en. In: *Ann. Oper. Res.* 134.1 (2005), pp. 19–67.

[Boh92] D. Bohning. "Multinomial logistic regression algorithm". In: *Annals of the Inst. of Statistical Math.* 44 (1992), pp. 197–200.

[Bol89] K. Bollen. *Structural Equation Models with Latent Variables*. John Wiley & Sons, 1989.

[Bon64] G. Bonnet. "Transformations des signaux aleátoires a travers les systemes non lineáires sans meḿoire". In: *Annales des Telecommuncations* 19 (1964).

[Bor] A. Borodin. "Determinantal point processes". In: *The Oxford Handbook of Random Matrix Theory*.

[Bös+17] J.-H. Böse, V. Flunkert, J. Gasthaus, T. Januschowski, D. Lange, D. Salinas, S. Schelter, M. Seeger, and Y. Wang. "Probabilistic Demand Forecasting at Scale". In: *Proceedings VLDB Endowment* 10.12 (2017), pp. 1694–1705.

[Bot+13] L. Bottou, J. Peters, J. Quiñonero-Candela, D. X. Charles, D. M. Chickering, E. Portugaly, D. Ray, P. Simard, and E. Snelson. "Counterfactual Reasoning and Learning Systems: The Example of Computational Advertising". In: *JMLR* 14 (2013), pp. 3207–3260.

[Bow+16a] S. R. Bowman, L. Vilnis, O. Vinyals, A. M. Dai, R. Jozefowicz, and S. Bengio. "Generating Sentences from a Continuous Space". In: *CONLL*. 2016.

[Bow+16b] S. R. Bowman, L. Vilnis, O. Vinyals, A. M. Dai, R. Jozefowicz, and S. Bengio. "Generating Sentences from a Continuous Space". In: *CONLL*. 2016.

[Box80] G. E. P. Box. "Sampling and Bayes' Inference in Scientific Modelling and Robustness". In: *J. of Royal Stat. Soc. Series A* 143.4 (1980), pp. 383–430.

[BP13] K. A. Bollen and J. Pearl. "Eight Myths About Causality and Structural Equation Models". In: *Handbook of Causal Analysis for Social Research*. Ed. by S. L. Morgan. Springer Netherlands, 2013, pp. 301–328.

[BP16] E. Bareinboim and J. Pearl. "Causal inference and the data-fusion problem". en. In: *Proc. Natl. Acad. Sci. U. S. A.* 113.27 (2016), pp. 7345–7352.

[BP21] T. Bricken and C. Pehlevan. "Attention Approximates Sparse Distributed Memory". In: *NIPS*. 2021.

[BPK16] S. Bulo, L. Porzi, and P. Kontschieder. "Distillation dropout". In: *ICML*. 2016.

[BPL21a] R. Balestriero, J. Pesenti, and Y. LeCun. "Learning in High Dimension Always Amounts to Extrapolation". In: (Oct. 2021). arXiv: 2110.09485 [cs.LG].

[BPL21b] A. Bardes, J. Ponce, and Y. LeCun. "Vicreg: Variance-invariance-covariance regularization for self-supervised learning". In: *arXiv preprint arXiv:2105.04906* (2021).

[BPS16] A. G. Baydin, B. A. Pearlmutter, and J. M. Siskind. "DiffSharp: An AD library for .NET languages". In: *arXiv preprint arXiv:1611.03423* (2016).

[BR05] H. Bang and J. M. Robins. "Doubly Robust Estimation in Missing Data and Causal Inference Models". In: *Biometrics* 61.4 (2005), pp. 962–973.

[BR18] B. Biggio and F. Roli. "Wild patterns: Ten years after the rise of adversarial machine learning". In: *Pattern Recognition* 84 (2018), pp. 317–331.

[BR98] S. Brooks and G. Roberts. "Assessing convergence of Markov Chain Monte Carlo algorithms". In: *Statistics and Computing* 8 (1998), pp. 319–335.

[Bra+18] J. Bradbury, R. Frostig, P. Hawkins, M. J. Johnson, C. Leary, D. Maclaurin, and S. Wanderman-Milne. *JAX: composable transformations of Python+NumPy programs*. Version 0.1.55. 2018.

[Bra96] M. Brand. *Coupled hidden Markov models for modeling interacting processes*. Tech. rep. 405. MIT Lab for Perceptual Computing, 1996.

[Bre01] L. Breiman. "Statistical modeling: The two cultures (with comments and a rejoinder by the author)". In: *Statistical science* 16.3 (2001), pp. 199–231.

[Bre+17] L. Breiman, J. H. Friedman, R. A. Olshen, and C. J. Stone. *Classification and regression trees*. Routledge, 2017.

[Bre+20a] J. Brehmer, G. Louppe, J. Pavez, and K. Cranmer. "Mining gold from implicit models to improve likelihood-free inference". en. In: *Proc. Natl. Acad. Sci. U. S. A.* 117.10 (2020), pp. 5242–5249.

[Bre+20b] R. Brekelmans, V. Masrani, F. Wood, G. Ver Steeg, and A. Galstyan. "All in the Exponential Family: Bregman Duality in Thermodynamic Variational Inference". In: *ICML*. 2020.

[Bre67] L. M. Bregman. "The relaxation method of finding the common point of convex sets and its application to the solution of problems in convex programming". In: *USSR Computational Mathematics and Mathematical Physics* 7.3 (1967), pp. 200–217.

[Bre91] Y. Brenier. "Polar factorization and monotone rearrangement of vector-valued functions". In: *Communications on pure and applied mathematics* 44.4 (1991), pp. 375–417.

[Bre92] J. Breese. "Construction of belief and decision networks". In: *Computational Intelligence* 8 (1992), 624–647.

[Bre96] L. Breiman. "Stacked regressions". In: *Mach. Learn.* 24.1 (1996), pp. 49–64.

[BRG20] R. Bai, V. Rockova, and E. I. George. "Spike-and-slab meets LASSO: A review of the spike-and-slab LASSO". In: (2020). arXiv: 2010.06451 [stat.ME].

[Bri50] G. W. Brier. "Verification of forecasts expressed in terms of probability". In: *Monthly Weather Review* 78.1 (1950), pp. 1–3.

[Bro09] G. Brown. "A new perspective on information theoretic feature selection". In: *AISTATS*. 2009.

[Bro+13] T. Broderick, N. Boyd, A. Wibisono, A. C. Wilson, and M. I. Jordan. "Streaming Variational Bayes". In: *NIPS*. 2013.

[Bro+15] K. H. Brodersen, F. Gallusser, J. Koehler, N. Remy, and S. L. Scott. "Inferring causal impact using Bayesian structural time-series models". In: *Ann. Appl. Stat.* 9.1 (2015), pp. 247–274.

[Bro18] T. Broderick. *Tutorial: Variational Bayes and Beyond*. 2018.

[Bro19] J. Brownlee. *Generative Adversarial Networks with Python*. Accessed: 2019-8-27. Machine Learning Mastery, 2019.

[Bro+20a] D. Brookes, A. Busia, C. Fannjiang, K. Murphy, and J. Listgarten. "A view of estimation of distribution algorithms through the lens of expectation-maximization". In: *GECCO*. GECCO '20. Association for Computing Machinery, 2020, pp. 189–190.

[Bro+20b] P. Brouillard, S. Lachapelle, A. Lacoste, S. Lacoste-Julien, and A. Drouin. "Differentiable Causal Discovery from Interventional Data". In: *NIPS*. July 2020.

[Bro+20c] D. Brown, R. Coleman, R. Srinivasan, and S. Niekum. "Safe imitation learning via fast bayesian reward inference from preferences". In: *International Conference on Machine Learning*. PMLR. 2020, pp. 1165–1177.

[Bro+20d] T. B. Brown et al. "Language Models are Few-Shot Learners". In: (2020). arXiv: 2005.14165 [cs.CL].

[BRS17] E. Balkanski, A. Rubinstein, and Y. Singer. "The Limitations of Optimization from Samples". In: *Proceedings of the 49th Annual ACM SIGACT Symposium on Theory of Computing*. STOC 2017. Montreal, Canada: Association for Computing Machinery, 2017, 1016–1027.

[BRSS18] N. Bou-Rabee and J. M. Sanz-Serna. "Geometric integrators and the Hamiltonian Monte Carlo method". In: *Acta Numer.* (2018).

[Bru+18] M. Brundage et al. "The Malicious Use of Artificial Intelligence: Forecasting, Prevention, and Mitigation". In: (2018). arXiv: 1802.07228 [cs.AI].

[BS17] E. Balkanski and Y. Singer. "Minimizing a Submodular Function from Samples." In: *NIPS*. 2017, pp. 814–822.

[BS18] S. Barratt and R. Sharma. "A note on the inception score". In: *arXiv preprint arXiv:1801.01973* (2018).

[BS20] E. Balkanski and Y. Singer. "A lower bound for parallel submodular minimization". In: *Proceedings of the 52nd Annual ACM SIGACT Symposium on Theory of Computing*. 2020, pp. 130–139.

[BS+20] S. Bobadilla-Suarez, C. Ahlheim, A. Mehrotra, A. Panos, and B. C. Love. "Measures of neural similarity". In: *Computational Brain & Behavior* 3.4 (2020), pp. 369–383.

[BS95a] A. J. Bell and T. J. Sejnowski. "An information maximisation approach to blind separation and blind deconvolution". In: *Neural Computation* 7.6 (1995), pp. 1129–1159.

[BS95b] A. J. Bell and T. J. Sejnowski. "An information-maximization approach to blind separation and blind deconvolution". In: *Neural computation* 7.6 (1995), pp. 1129–1159.

[BSA83] A. G. Barto, R. S. Sutton, and C. W. Anderson. "Neuronlike adaptive elements that can solve difficult learning control problems". In: *SMC* 13.5 (1983), pp. 834–846.

[BSF88] Y. Bar-Shalom and T. Fortmann. *Tracking and data association*. Academic Press, 1988.

[BSL93] Y. Bar-Shalom and X. Li. *Estimation and Tracking: Principles, Techniques and Software*. Artech House, 1993.

[BSWT11] Y. Bar-Shalom, P. K. Willett, and X. Tian. *Tracking and Data Fusion: A Handbook of Algorithms*. en. Yaakov Bar-Shalom, 2011.

[BT00] C. Bishop and M. Tipping. "Variational relevance vector machines". In: *UAI*. 2000.

[BT04] G. Bouchard and B. Triggs. "The tradeoff between generative and discriminative classifiers". In: *IASC International Symposium on Computational Statistics (COMPSTAT '04)*. 2004.

[BT12] M. Botvinick and M. Toussaint. "Planning as inference". en. In: *Trends Cogn. Sci.* 16.10 (2012), pp. 485–488.

[BT73] G. Box and G. Tiao. *Bayesian inference in statistical analysis*. Addison-Wesley, 1973.

[BTEGN09] A. Ben-Tal, L. El Ghaoui, and A. Nemirovski. *Robust optimization*. Vol. 28. Princeton University Press, 2009.

[Buc+12] N. Buchbinder, M. Feldman, J. Naor, and R. Schwartz. "A tight (1/2) linear-time approximation to unconstrained submodular maximization". In: *In FOCS* (2012).

[Buc+17] C. L. Buckley, C. S. Kim, S. McGregor, and A. K. Seth. "The free energy principle for action and perception: A mathematical review". In: *J. Math. Psychol.* 81 (2017), pp. 55–79.

[Bud+21] K. Budhathoki, D. Janzing, P. Bloebaum, and H. Ng. "Why did the distribution change?" In: *AISTATS*. Ed. by A. Banerjee and K. Fukumizu. Vol. 130. Proceedings of Machine Learning Research. PMLR, 2021, pp. 1666–1674.

[Bul11] A. D. Bull. "Convergence rates of efficient global optimization algorithms". In: *JMLR* 12 (2011), 2879–2904.

[Bul+20] S. Bulusu, B. Kailkhura, B. Li, P. K. Varshney, and D. Song. "Anomalous Example Detection in Deep Learning: A Survey". In: *IEEE Access* 8 (2020), pp. 132330–132347.

[BV04] S. Boyd and L. Vandenberghe. *Convex optimization*. Cambridge, 2004.

[BVHP18] S. Beery, G. Van Horn, and P. Perona. "Recognition in terra incognita". In: *Proceedings of the European Conference on Computer Vision (ECCV)*. 2018, pp. 456–473.

[BVW02] H. Bui, S. Venkatesh, and G. West. "Policy Recognition in the Abstract Hidden Markov Model". In: *JAIR* 17 (2002), pp. 451–499.

[BVZ01] Y. Boykov, O. Veksler, and R. Zabih. "Fast Approximate Energy Minimization via Graph Cuts". In: *IEEE PAMI* 23.11 (2001).

[BVZ99] Y. Boykov, O. Veksler, and R. Zabih. "Fast Approximate Energy Minimization via Graph Cuts". In: *ICCV (1)*. 1999, pp. 377–384.

[BW20] G. J. J. van den Burg and C. K. I. Williams. "An Evaluation of Change Point Detection Algorithms". In: (2020). arXiv: 2003.06222 [stat.ML].

[BW21] G. J. van den Burg and C. K. Williams. "On Memorization in Probabilistic Deep Generative Models". In: *NIPS*. 2021.

[BWM18] A. Buchholz, F. Wenzel, and S. Mandt. "Quasi-Monte Carlo Variational Inference". In: *ICML*. 2018.

[BWR16] M. Bauer, M. van der Wilk, and C. E. Rasmussen. "Understanding Probabilistic Sparse Gaussian Process Approximations". In: *NIPS*. 2016, pp. 1533–1541.

[BYH20] O. Bohdal, Y. Yang, and T. Hospedales. "Flexible Dataset Distillation: Learn Labels Instead of Images". In: *arXiv preprint arXiv:2006.08572* (2020).

[BYM17] D. Belanger, B. Yang, and A. McCallum. "End-to-End Learning for Structured Prediction Energy Networks". In: *ICML*. Ed. by D. Precup and Y. W. Teh. Vol. 70. Proceedings of Machine Learning Research. PMLR, 2017, pp. 429–439.

[Byr+16] R Byrd, S Hansen, J Nocedal, and Y Singer. "A Stochastic Quasi-Newton Method for Large-Scale Optimization". In: *SIAM J. Optim.* 26.2 (2016), pp. 1008–1031.

[BZ20] A. Barbu and S.-C. Zhu. *Monte Carlo Methods*. en. Springer, 2020.

[CA13] E. F. Camacho and C. B. Alba. *Model predictive control*. Springer, 2013.

[Cac+18] M. Caccia, L. Caccia, W. Fedus, H. Larochelle, J. Pineau, and L. Charlin. "Language GANs Falling Short". In: *CoRR* abs/1811.02549 (2018). arXiv: 1811.02549.

[CAII20] V. Coscrato, M. H. de Almeida Inácio, and R. Izbicki. "The NN-Stacking: Feature weighted linear stacking through neural networks". In: *Neurocomputing* (2020).

[Cal+07] G. Calinescu, C. Chekuri, M. Pál, and J. Vondrák. "Maximizing a submodular set function subject to a matroid constraint". In: *Proceedings of the 12th International Conference on Integer Programming and Combinatorial Optimization (IPCO)*. 2007, pp. 182–196.

[Cal20] O. Calin. *Deep Learning Architectures: A Mathematical Approach*. en. 1st ed. Springer, 2020.

[Cam+21] A. Campbell, Y. Shi, T. Rainforth, and A. Doucet. "Online Variational Filtering and Parameter Learning". In: *NIPS*. 2021.

[Can04] J. Canny. "GaP: a factor model for discrete data". In: *SIGIR*. 2004, pp. 122–129.

[Cao+15] Y. Cao, M. A Brubaker, D. J Fleet, and A. Hertzmann. "Efficient Optimization for Sparse Gaussian Process Regression". en. In: *IEEE PAMI* 37.12 (2015), pp. 2415–2427.

[Car03] P. Carbonetto. "Unsupervised Statistical Models for General Object Recognition". MA thesis. University of British Columbia, 2003.

[Car+15] R. Caruana, Y. Lou, J. Gehrke, P. Koch, M. Sturm, and N. Elhadad. "Intelligible models for healthcare: Predicting pneumonia risk and hospital 30-day readmission". In: *Proceedings of the 21th ACM SIGKDD international conference on knowledge discovery and data mining*. 2015, pp. 1721–1730.

[Car+19] N. Carlini, A. Athalye, N. Papernot, W. Brendel, J. Rauber, D. Tsipras, I. Goodfellow, A. Madry, and A. Kurakin. "On evaluating adversarial robustness". In: *arXiv preprint arXiv:1902.06705* (2019).

[Car+21] M. Caron, H. Touvron, I. Misra, H. Jégou, J. Mairal, P. Bojanowski, and A. Joulin. *Emerging Properties in Self-Supervised Vision Transformers*. 2021. arXiv: 2104.14294 [cs.CV].

[Car97] R. Caruana. "Multitask Learning". In: *Machine Learning* 28.1 (1997), pp. 41–75.

[Cat08] A. Caticha. *Lectures on Probability, Entropy, and Statistical Physics*. 2008. arXiv: 0808 . 0012 [physics.data-an].

[Cat+11] A. Caticha, A. Mohammad-Djafari, J.-F. Bercher, and P. Bessiére. "Entropic Inference". In: *AIP Conference Proceedings* 1305.1 (2011), pp. 20–29. eprint: https://aip.scitation.org/doi/pdf/10.1063/1.3573619.

[CB20] Y. Chen and P. Bühlmann. "Domain adaptation under structural causal models". In: *JMLR* (2020).

[CBL20] K. Cranmer, J. Brehmer, and G. Louppe. "The frontier of simulation-based inference". In: *Proceedings of the National Academy of Sciences* 117.48 (2020), pp. 30055–30062.

[CBR20] Z. Chen, Y. Bei, and C. Rudin. "Concept whitening for interpretable image recognition". In: *Nature Machine Intelligence* 2.12 (2020), pp. 772–782.

[CC84] M. Conforti and G. Cornuejols. "Submodular set functions, matroids and the greedy algorithm: tight worst-case bounds and some generalizations of the Rado-Edmonds theorem". In: *Discrete Applied Mathematics* 7.3 (1984), pp. 251–274.

[CC96] M. Cowles and B. Carlin. "Markov Chain Monte Carlo Convergence Diagnostics: A Comparative Review". In: *JASA* 91 (1996), pp. 883–904.

[CCS22] A. Corenflos, N. Chopin, and S. Särkkä. "De-Sequentialized Monte Carlo: a parallel-in-time particle smoother". In: (2022). arXiv: 2202.02264 [stat.CO].

[CDC15] C. Chen, N. Ding, and L. Carin. "On the Convergence of Stochastic Gradient MCMC Algorithms with High-Order Integrators". In: *NIPS*. 2015.

[CDS02] M. Collins, S. Dasgupta, and R. E. Schapire. "A Generalization of Principal Components Analysis to the Exponential Family". In: *NIPS-14*. 2002.

[CDS19] A. Clark, J. Donahue, and K. Simonyan. "Adversarial video generation on complex datasets". In: *arXiv preprint arXiv:1907.06571* (2019).

[Cér+12] F Cérou, P Del Moral, T Furon, and A Guyader. "Sequential Monte Carlo for rare event estimation". In: *Stat. Comput.* 22.3 (2012), pp. 795–808.

[CFG14] T. Chen, E. B. Fox, and C. Guestrin. "Stochastic Gradient Hamiltonian Monte Carlo". In: *ICML*. 2014.

[CG00] S. S. Chen and R. A. Gopinath. "Gaussianization". In: *NIPS*. 2000, pp. 423–429.

[CG15] X. Chen and A. Gupta. "Webly supervised learning of convolutional networks". In: *Proceedings of the IEEE International Conference on Computer Vision*. 2015, pp. 1431–1439.

[CG18] C. Ceylan and M. U. Gutmann. "Conditional Noise-Contrastive Estimation of Unnormalised Models". In: *International Conference on Machine Learning*. 2018, pp. 726–734.

[CG96] S. Chen and J. Goodman. "An empirical study of smoothing techniques for language modeling". In: *Proc. 34th ACL*. 1996, pp. 310–318.

[CG98] S. Chen and J. Goodman. *An empirical study of smoothing techniques for language modeling*. Tech. rep. TR-10-98. Dept. Comp. Sci., Harvard, 1998.

[CGR06] S. R. Cook, A. Gelman, and D. B. Rubin. "Validation of Software for Bayesian Models Using Posterior Quantiles". In: *J. Comput. Graph. Stat.* 15.3 (2006), pp. 675–692.

[CGS15] T. Chen, I. Goodfellow, and J. Shlens. "Net2net: Accelerating learning via knowledge transfer". In: *International Conference on Learning Representations*. 2015.

[CH20] C. Cinelli and C. Hazlett. "Making sense of sensitivity: extending omitted variable bias". In: *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 82.1 (2020), pp. 39–67. eprint: https://rss.onlinelibrary.wiley.com/doi/pdf/10.1111/rssb.12348.

[Cha12] K. M. A. Chai. "Variational Multinomial Logit Gaussian Process". In: *JMLR* 13.Jun (2012), pp. 1745–1808.

[Cha14] N. Chapados. "Effective Bayesian Modeling of Groups of Related Count Time Series". In: *ICML*. 2014.

[Cha+14] K. Chatfield, K. Simonyan, A. Vedaldi, and A. Zisserman. "Return of the devil in the details: Delving deep into convolutional nets". In: *British Machine Vision Conference*. 2014.

[Cha+17] D. Chakrabarty, Y. T. Lee, A. Sidford, and S. C.-w. Wong. "Subquadratic submodular function minimization". In: *Proceedings of the 49th Annual ACM SIGACT Symposium on Theory of Computing*. 2017, pp. 1220–1231.

[Cha+18] N. S. Chatterji, N. Flammarion, Y.-A. Ma, P. L. Bartlett, and M. I. Jordan. "On the theory of variance reduction for stochastic gradient Monte Carlo". In: *ICML*. 2018.

[Cha+19a] C. Chan, S. Ginosar, T. Zhou, and A. A. Efros. "Everybody dance now". In: *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 2019, pp. 5933–5942.

[Cha+19b] T. Chavdarova, G. Gidel, F. Fleuret, and S. Lacoste-Julien. "Reducing noise in GAN training with variance reduced extragradient". In: *Advances in Neural Information Processing Systems*. 2019, pp. 393–403.

[Cha21] S. H. Chan. *Introduction to Probability for Data Science*. Michigan Publishing, 2021.

[Che+05] G. Chechik, A. Globerson, N. Tishby, and Y. Weiss. "Information Bottleneck for Gaussian Variables". In: *JMLR* 6.Jan (2005), pp. 165–188.

[Che+15] L.-C. Chen, G. Papandreou, I. Kokkinos, K. Murphy, and A. L. Yuille. "Semantic Image Segmentation with Deep Convolutional Nets and Fully Connected CRFs". In: *ICLR*. 2015.

[Che+16] X. Chen, Y. Duan, R. Houthooft, J. Schulman, I. Sutskever, and P. Abbeel. "InfoGAN: Interpretable Representation Learning by Information Maximizing Generative Adversarial Nets". In: *NIPS*. 2016.

[Che+17] T. Che, Y. Li, R. Zhang, R. D. Hjelm, W. Li, Y. Song, and Y. Bengio. "Maximum-likelihood augmented discrete generative adversarial networks". In: *arXiv preprint arXiv:1702.07983* (2017).

[Che17] C. Chelba. *Language Modeling in the Era of Abundant Data*. AI with the best. 2017.

[Che+17a] L.-C. Chen, G. Papandreou, I. Kokkinos, K. Murphy, and A. L. Yuille. "DeepLab: Semantic Image Segmentation with Deep Convolutional Nets, Atrous Convolution, and Fully Connected CRFs". In: *IEEE PAMI* (2017).

[Che+17b] X. Chen, D. P. Kingma, T. Salimans, Y. Duan, P. Dhariwal, J. Schulman, I. Sutskever, and P. Abbeel. "Variational Lossy Autoencoder". In: *ICLR*. 2017.

[Che+17c] X. Chen, N. Mishra, M. Rohaninejad, and P. Abbeel. "PixelSNAIL: An Improved Autoregressive Generative Model". In: (2017). arXiv: 1712.09763 [cs.LG].

[Che+17d] V. Chernozhukov, D. Chetverikov, M. Demirer, E. Duflo, C. Hansen, and W. Newey. "Double/Debiased/Neyman Machine Learning of Treatment Effects". In: *American Economic Review* 107.5 (2017), pp. 261–65.

[Che+17e] V. Chernozhukov, D. Chetverikov, M. Demirer, E. Duflo, C. Hansen, W. Newey, and J. Robins. "Double/Debiased Machine Learning for Treatment and Structural parameters". In: *The Econometrics Journal* (2017).

[Che+18a] C. Chen, W. Wang, Y. Zhang, Q. Su, and L. Carin. "A convergence analysis for a class of practical variance-reduction stochastic gradient MCMC". In: *Sci. China Inf. Sci.* 62.1 (2018), p. 12101.

[Che+18b] C. Chen, O. Li, C. Tao, A. J. Barnett, J. Su, and C. Rudin. "This looks like that: deep learning for interpretable image recognition". In: *arXiv preprint arXiv:1806.10574* (2018).

[Che+18c] R. T. Q. Chen, Y. Rubanova, J. Bettencourt, and D. Duvenaud. "Neural Ordinary Differential Equations". In: *NIPS*. 2018.

[Che+18d] X. Cheng, N. S. Chatterji, P. L. Bartlett, and M. I. Jordan. "Underdamped Langevin MCMC: A non-asymptotic analysis". In: *COLT*. 2018.

[Che+19] R. T. Q. Chen, J. Behrmann, D. Duvenaud, and J.-H. Jacobsen. "Residual Flows for Invertible Generative Modeling". In: *NIPS*. 2019.

[Che+20a] M. Chen, A. Radford, R. Child, J. Wu, H. Jun, D. Luan, and I. Sutskever. "Generative pretraining from pixels". In: *International Conference on Machine Learning*. PMLR. 2020, pp. 1691–1703.

[Che+20b] M. Chen, Y. Wang, T. Liu, Z. Yang, X. Li, Z. Wang, and T. Zhao. "On computation and generalization of generative adversarial imitation learning". In: *arXiv preprint arXiv:2001.02792* (2020).

[Che+20c] T. Chen, S. Kornblith, M. Norouzi, and G. Hinton. "A simple framework for contrastive learning of visual representations". In: *ICML*. 2020.

[Che+21] N. Chen, Y. Zhang, H. Zen, R. J. Weiss, M. Norouzi, N. Dehak, and W. Chan. "WaveGrad 2: Iterative refinement for text-to-speech synthesis". In: *Interspeech 2021*. ISCA: ISCA, Aug. 2021.

[Che+22] D. Chen, D. Wang, T. Darrell, and S. Ebrahimi. "Contrastive Test-Time Adaptation". In: *CVPR*. Apr. 2022.

[Che95] Y. Cheng. "Mean shift, mode seeking, and clustering". In: *IEEE PAMI* 17.8 (1995).

[Chi14] S. Chiappa. "Explicit-Duration Markov Switching Models". In: *Foundations and Trends in Machine Learning* 7.6 (2014), pp. 803–886.

[Chi21a] R. Child. "Very Deep VAEs Generalize Autoregressive Models and Can Outperform Them on Images". In: *ICLR*. 2021.

[Chi21b] R. Child. "Very Deep VAEs Generalize Autoregressive Models and Can Outperform Them on Images". In: *ArXiv* abs/2011.10650 (2021).

[CHL+05] S. Chopra, R. Hadsell, Y. LeCun, et al. "Learning a similarity metric discriminatively, with application to face verification". In: *CVPR*. 2005, pp. 539–546.

[CHM97] D. M. Chickering, D. Heckerman, and C. Meek. "A Bayesian Approach to Learning Bayesian Networks with Local Structure". In: *UAI*. UAI'97. 1997, pp. 80–89.

[Cho02] N. Chopin. "A Sequential Particle Filter Method for Static Models". In: *Biometrika* 89.3 (2002), pp. 539–551.

[Cho11] M. J. Choi. "Trees and Beyond: Exploiting and Improving Tree-Structured Graphical Models". PhD thesis. MIT, 2011.

[Cho+15] Y. Chow, A. Tamar, S. Mannor, and M. Pavone. "Risk-Sensitive and Robust Decision-Making: A CVaR Optimization Approach". In: *NIPS*. 2015, pp. 1522–1530.

[Cho21] F. Chollet. *Deep learning with Python (second edition)*. Manning, 2021.

[Cho57] N. Chomsky. *Syntactic Structures*. Mouton, 1957.

[Chr+21] R. Christiansen, N. Pfister, M. E. Jakobsen, N. Gnecco, and J. Peters. "A causal framework for distribution generalization". In: *IEEE PAMI* (2021).

[Chu+15] J. Chung, K. Kastner, L. Dinh, K. Goel, A. Courville, and Y. Bengio. "A Recurrent Latent Variable Model for Sequential Data". In: *NIPS*. 2015.

[Chu+18] K. Chua, R. Calandra, R. McAllister, and S. Levine. "Deep Reinforcement Learning in a Handful of Trials using Probabilistic Dynamics Models". In: *NIPS*. 2018.

[Chu+19] G. Chuang, G. DeSalvo, L. Karydas, J.-F. Kagy, A. Rostamizadeh, and A Theeraphol. "Active Learning Empirical Study". In: *NeurIPS LIRE Workshop*. 2019.

[Chw+15] K. Chwialkowski, A. Ramdas, D. Sejdinovic, and A. Gretton. "Fast Two-Sample Testing with Analytic Representations of Probability Measures". In: *NIPS*. 2015.

[CI80] D. R. Cox and V. Isham. *Point processes*. Vol. 12. CRC Press, 1980.

[CJ21] A. D. Cobb and B. Jalaian. "Scaling Hamiltonian Monte Carlo inference for Bayesian neural networks with symmetric splitting". In: *UAI*. Vol. 161. Proceedings of Machine Learning Research. PMLR, 2021, pp. 675–685.

[CK05] M. Collins and T. Koo. "Discriminative Reranking for Natural Language Parsing". In: *Proc. ACL*. 2005.

[CK07] J. J. F. Commandeur and S. J. Koopman. *An Introduction to State Space Time Series Analysis (Practical Econometrics)*. en. 1st ed. Oxford University Press, 2007.

[CK21] D. Chakrabarty and S. Khanna. "Better and simpler error analysis of the Sinkhorn–Knopp algorithm for matrix scaling". In: *Mathematical Programming* 188.1 (2021), pp. 395–407.

[CK94a] D. Card and A. B. Krueger. "Minimum Wages and Employment: A Case Study of the Fast-Food Industry in New Jersey and Pennsylvania". In: *American Economic Review* 84.4 (1994), pp. 772–793.

[CK94b] C. Carter and R. Kohn. "On Gibbs sampling for state space models". In: *Biometrika* 81.3 (1994), pp. 541–553.

[CK96] C. Carter and R. Kohn. "Markov Chain Monte Carlo in Conditionally Gaussian State Space Models". In: *Biometrika* 83 (1996), pp. 589–601.

[CKK17] L. Chen, A. Krause, and A. Karbasi. "Interactive Submodular Bandit." In: *NIPS*. 2017, pp. 141–152.

[CL00] R. Chen and S. Liu. "Mixture Kalman Filters". In: *J. Royal Stat. Soc. B* (2000).

[CL07] L. Carvahlo and C. Lawrence. "Centroid estimation in discrete high-dimensional spaces with applications in biology". In: *PNAS* 105.4 (2007).

[CL11] O. Chapelle and L. Li. "An empirical evaluation of Thompson sampling". In: *NIPS*. 2011.

[CL18] Z. Chen and B. Liu. *Lifelong Machine Learning*. Synthesis Lectures on Artificial Intelligence and Machine Learning. Morgan Claypool, 2018.

[CL96] B. P. Carlin and T. A. Louis. *Bayes and Empirical Bayes Methods for Data Analysis*. Chapman and Hall, 1996.

[Cla20] P. Clavier. "Sum-Product Network in the context of missing data". en. MA thesis. KTH, 2020.

[Cla21] A. Clayton. *Bernoulli's Fallacy: Statistical Illogic and the Crisis of Modern Science*. en. Columbia University Press, 2021.

[CLD18] C. Cremer, X. Li, and D. Duvenaud. "Inference Suboptimality in Variational Autoencoders". In: *ICML*. 2018.

[Clo+19] J. R. Clough, I. Oksuz, E. Puyol-Antón, B. Ruijsink, A. P. King, and J. A. Schnabel. "Global and local interpretability for cardiac MRI classification". In: *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer. 2019, pp. 656–664.

[Clo20] Cloudera. *Causality for ML*. 2020.

[CLV19] M. Cox, T. van de Laar, and B. de Vries. "A factor graph approach to automated design of Bayesian signal processing algorithms". In: *Int. J. Approx. Reason.* 104 (2019), pp. 185–204.

[CLW18] Y. Chen, L. Li, and M. Wang. "Scalable Bilinear π-Learning Using State and Action Features". In: *ICML*. 2018, pp. 833–842.

[CM09] O. Cappé and E. Moulines. "Online EM Algorithm for Latent Data Models". In: *J. of Royal Stat. Soc. Series B* 71.3 (2009), pp. 593–613.

[CMD17] C. Cremer, Q. Morris, and D. Duvenaud. "Reinterpreting Importance-Weighted Autoencoders". In: *ICLR Workshop*. 2017.

[CMR05] O. Cappe, E. Moulines, and T. Ryden. *Inference in Hidden Markov Models*. Springer, 2005.

[CMR12] C. Cortes, M. Mohri, and A. Rostamizadeh. "Algorithms for learning kernels based on centered alignment". In: *The Journal of Machine Learning Research* 13.1 (2012), pp. 795–828.

[CMS12] D. Ciregan, U. Meier, and J. Schmidhuber. "Multi-column deep neural networks for image classification". In: *2012 IEEE conference on computer vision and pattern recognition*. IEEE. 2012, pp. 3642–3649.

[CN01] H. Choset and K. Nagatani. "Topological simultaneous localization and mapping (SLAM): toward exact localization without explicit localization". In: *IEEE Trans. Robotics and Automation* 17.2 (2001).

[CNW20] M. Collier, A. Nazabal, and C. K. I. Williams. "VAEs in the Presence of Missing Data". In: *ICML Workshop on the Art of Learning with Missing Values*. 2020.

[CO06] N. Chater and M. Oaksford. "Mental mechanisms". In: *Information sampling and adaptive cognition* (2006), pp. 210–236.

[COB18] L. Chizat, E. Oyallon, and F. Bach. "On Lazy Training in Differentiable Programming". In: (2018). arXiv: 1812.07956 [math.OC].

[Cor+12] G. Cormode, M. Garofalakis, P. J. Haas, and C. Jermaine. "Synopses for massive data: Samples, histograms, wavelets, sketches". In: *Foundations and Trends in Databases* 4.1–3 (2012), pp. 1–294.

[Cor17] G. Cormode. "Data sketching". In: *Communications of the ACM* 60.9 (2017), pp. 48–55.

[Cor+59] J. Cornfield, W. Haenszel, E. C. Hammond, A. M. Lilienfeld, M. B. Shimkin, and E. L. Wynder. "Smoking and Lung Cancer: Recent Evidence and a Discussion of Some Questions". In: *JNCI: Journal of the National Cancer Institute* 22.1 (Jan. 1959), pp. 173–203. eprint: https://academic.oup.com/jnci/article-pdf/22/1/173/2704718/22-1-173.pdf.

[Cor+87] A Corana, M Marchesi, C Martini, and S Ridella. "Minimizing Multimodal Functions of Continuous Variables with the "Simulated Annealing" Algorithm". In: *ACM Trans. Math. Softw.* 13.3 (1987), pp. 262–280.

[Cou16] Council of European Union. *General Data Protection Regulation*. 2016.

[Cou+20] J. Courts, A. Wills, T. Schön, and B. Ninness. "Variational System Identification for Nonlinear State-Space Models". In: (2020). arXiv: 2012.05072 [stat.ML].

[Cou+21] J. Courts, J. Hendriks, A. Wills, T. Schön, and B. Ninness. "Variational State and Parameter Estimation". In: *19th IFAC Symposium on System Identification SYSID 2021*. 2021.

[Cov99] T. M. Cover. *Elements of information theory*. John Wiley & Sons, 1999.

[CP20a] R. Chen and I. C. Paschalidis. *Distributionally Robust Learning*. NOW Foundations and Trends in Optimization, 2020.

[CP20b] N. Chopin and O. Papaspiliopoulos. *An Introduction to Sequential Monte Carlo*. en. 1st ed. Springer, 2020.

[CPD17] P. Constantinou and A Philip Dawid. "Extended conditional independence and applications in causal inference". en. In: *Ann. Stat.* 45.6 (2017), pp. 2618–2653.

[CPS10] C. Carvahlo, N. Polson, and J. Scott. "The horseshoe estimator for sparse signals". In: *Biometrika* 97.2 (2010), p. 465.

[Cri+02] N. Cristianini, J. Shawe-Taylor, A. Elisseeff, and J. S. Kandola. "On Kernel-Target Alignment". In: *Advances in Neural Information Processing Systems*. 2002, pp. 367–373.

[CRK19] J. M. Cohen, E. Rosenfeld, and J. Z. Kolter. "Certified adversarial robustness via randomized smoothing". In: *arXiv preprint arXiv:1902.02918* (2019).

[Cro+11] D. F. Crouse, P. Willett, K. Pattipati, and L. Svensson. "A look at Gaussian mixture reduction algorithms". In: *14th International Conference on Information Fusion*. 2011, pp. 1–8.

[CS04] I. Csiszár and P. C. Shields. "Information theory and statistics: A tutorial". In: (2004).

[CS09] Y. Cho and L. K. Saul. "Kernel Methods for Deep Learning". In: *NIPS*. 2009, pp. 342–350.

[CS18] P. Chaudhari and S. Soatto. "Stochastic gradient descent performs variational inference, converges to limit cycles for deep networks". In: *ICLR*. 2018.

[CSF16] A. W. Churchill, S. Sigtia, and C. Fernando. "Learning to Generate Genotypes with Neural Networks". In: (2016). arXiv: 1604.04153 [cs.NE].

[Csi67] I. Csiszar. "Information-Type Measures of Difference of Probability Distributions and Indirect Observations". In: *Studia Scientiarum Mathematicarum Hungarica* 2 (1967), pp. 299–318.

[CSN21] J. Coullon, L. South, and C. Nemeth. "Stochastic Gradient MCMC with Multi-Armed Bandit Tuning". In: (2021). arXiv: 2105.13059 [stat.CO].

[CT06] T. M. Cover and J. A. Thomas. *Elements of Information Theory*. 2nd edition. John Wiley, 2006.

[CT+19] M. F. Cusumano-Towner, F. A. Saad, A. K. Lew, and V. K. Mansinghka. "Gen: a general-purpose probabilistic programming system with programmable inference". In: *Proceedings of the 40th ACM SIGPLAN Conference on Programming Language Design and Implementation*. PLDI 2019. Association for Computing Machinery, 2019, pp. 221–236.

[CT91] T. M. Cover and J. A. Thomas. *Elements of Information Theory*. John Wiley, 1991.

[CTM17] M. F. Cusumano-Towner and V. K. Mansinghka. "AIDE: An algorithm for measuring the accuracy of probabilistic inference algorithms". In: *NIPS*. 2017.

[CTN17] Y. Chali, M. Tanvee, and M. T. Nayeem. "Towards abstractive multi-document summarization using submodular function-based framework, sentence compression and merging". In: *Proceedings of the Eighth International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*. 2017, pp. 418–424.

[CTS78] C. Cannings, E. A. Thompson, and M. H. Skolnick. "Probability functions in complex pedigrees". In: *Advances in Applied Probability* 10 (1978), pp. 26–61.

[CUH16] D.-A. Clevert, T. Unterthiner, and S. Hochreiter. "Fast and Accurate Deep Network Learning by Exponential Linear Units (ELUs)". In: *ICLR*. 2016.

[Cui+18] Y. Cui, Y. Song, C. Sun, A. Howard, and S. Belongie. "Large scale fine-grained categorization and domain-specific transfer learning". In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2018, pp. 4109–4118.

[Cun22] Y. L. Cun. *A path towards autonomous AI*. 2022.

[Cun83] W. H. Cunningham. "Decomposition of submodular functions". In: *Combinatorica* 3.1 (1983), pp. 53–68.

[Cut13] M. Cuturi. "Sinkhorn Distances: Lightspeed Computation of Optimal Transportation Distances". In: *NIPS*. 2013.

[CW07] C. M. Carvahlo and M. West. "Dynamic Matrix-Variate Graphical Models". In: *Bayesian Analysis* 2.1 (2007), pp. 69–98.

[CW16] T. Cohen and M. Welling. "Group Equivariant Convolutional Networks". en. In: *ICML*. 2016, pp. 2990–2999.

[CWG20] D. C. Castro, I. Walker, and B. Glocker. "Causality matters in medical imaging". en. In: *Nat. Commun.* 11.1 (2020), p. 3673.

[CWS21] J. Courts, A. G. Wills, and T. B. Schön. "Gaussian Variational State Estimation for Nonlinear State-Space Models". In: *IEEE Trans. Signal Process.* 69 (2021), pp. 5979–5993.

[CXH21] X. Chen, S. Xie, and K. He. "An empirical study of training self-supervised vision transformers". In: *arXiv preprint arXiv:2104.02057* (2021).

[CY20] G. Cormode and K. Yi. *Small Summaries for Big Data*. Cambridge University Press, 2020.

[Cza+20] J. Czarnowski, T. Laidlow, R. Clark, and A. J. Davison. "DeepFactors: Real-Time Probabilistic Dense Monocular SLAM". In: *ICRA*. 2020.

[CZG20] B. Charpentier, D. Zügner, and S. Günnemann. "Posterior network: Uncertainty estimation without ood samples via density-based pseudo-counts". In: *arXiv preprint arXiv:2006.09239* (2020).

[D'A+20] A. D'Amour et al. "Underspecification Presents Challenges for Credibility in Modern Machine Learning". In: (2020). arXiv: 2011.03395 [cs.LG].

[D'A+21] A. D'Amour, P. Ding, A. Feller, L. Lei, and J. Sekhon. "Overlap in observational studies with high-dimensional covariates". In: *Journal of Econometrics* 221.2 (2021), pp. 644–654.

[Dag+21] N. Dagan, N. Barda, E. Kepten, O. Miron, S. Perchik, M. A. Katz, M. A. Hernán, M. Lipsitch, B. Reis, and R. D. Balicer. "BNT162b2 mRNA Covid-19 Vaccine in a Nationwide Mass Vaccination Setting". In: *New England Journal of Medicine* 384.15 (2021), pp. 1412–1423. eprint: https://doi.org/10.1056/NEJMoa2101765.

[Dai+17] H. Dai, B. Dai, Y.-M. Zhang, S. Li, and L. Song. "Recurrent Hidden Semi-Markov Model". In: *ICLR*. 2017.

[Dai+18] B. Dai, A. Shaw, L. Li, L. Xiao, N. He, Z. Liu, J. Chen, and L. Song. "SBEED: Convergent Rein-

forcement Learning with Nonlinear Function Approximation". In: *ICML*. 2018, pp. 1133–1142.

[Dai+19a] B. Dai, H. Dai, A. Gretton, L. Song, D. Schuurmans, and N. He. "Kernel exponential family estimation via doubly dual embedding". In: *AISTATS*. PMLR. 2019, pp. 2321–2330.

[Dai+19b] B. Dai, Z. Liu, H. Dai, N. He, A. Gretton, L. Song, and D. Schuurmans. "Exponential family estimation via adversarial dynamics embedding". In: *Advances in Neural Information Processing Systems*. 2019, pp. 10979–10990.

[Dai+20a] C. Dai, J. Heng, P. E. Jacob, and N. Whiteley. "An invitation to sequential Monte Carlo samplers". In: (2020). arXiv: 2007.11936 [stat.CO].

[Dai+20b] Z. Dai, G. Lai, Y. Yang, and Q. V. Le. "Funnel-Transformer: Filtering out Sequential Redundancy for Efficient Language Processing". In: *NIPS*. 2020.

[Dal+04] N. Dalvi, P. Domingos, S. Sanghai, and D. Verma. "Adversarial classification". In: *Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining*. 2004, pp. 99–108.

[Dar03] A. Darwiche. "A Differential Approach to Inference in Bayesian Networks". In: *J. ACM* 50.3 (2003), pp. 280–305.

[Dar09] A. Darwiche. *Modeling and Reasoning with Bayesian Networks*. Cambridge, 2009.

[Dar+11] S. Darolles, Y. Fan, J.-P. Florens, and E. Renault. "Nonparametric instrumental regression". In: *Econometrica* 79.5 (2011), pp. 1541–1565.

[Dar80] R. A. Darton. "Rotation in Factor Analysis". In: *Journal of the Royal Statistical Society. Series D (The Statistician)* 29.3 (1980), pp. 167–194.

[Dau07] H. Daume. "Fast search for Dirichlet process mixture models". In: *AISTATS*. 2007.

[Dav+04] T. A. Davis, J. R. Gilbert, S. I. Larimore, and E. G. Ng. "A Column Approximate Minimum Degree Ordering Algorithm". In: *ACM Trans. Math. Softw.* 30.3 (2004), pp. 353–376.

[Dav+18] T. R. Davidson, L. Falorsi, N. De Cao, T. Kipf, and J. M. Tomczak. "Hyperspherical Variational Auto-Encoders". In: *UAI*. 2018.

[Daw00] A. P. Dawid. "Causal Inference Without Counterfactuals". In: *JASA* 95.450 (2000), pp. 407–424.

[Daw02] A. P. Dawid. "Influence diagrams for causal modelling and inference". In: *Intl. Stat. Review* 70 (2002). Corrections p437, pp. 161–189.

[Daw15] A. P. Dawid. "Statistical Causality from a Decision-Theoretic Perspective". In: *Annu. Rev. Stat. Appl.* 2.1 (2015), pp. 273–303.

[Daw82] A. P. Dawid. "The Well-Calibrated Bayesian". In: *JASA* 77.379 (1982), pp. 605–610.

[Dax+21] E. Daxberger, A. Kristiadi, A. Immer, R. Eschenhagen, M. Bauer, and P. Hennig. "Laplace Redux–Effortless Bayesian Deep Learning". In: *NIPS*. 2021.

[Day+95] P. Dayan, G. Hinton, R. Neal, and R. Zemel. "The Helmholtz machine". In: *Neural Networks* 9.8 (1995).

[DB+13] P. De Blasi, S. Favaro, A. Lijoi, R. H. Mena, I. Prünster, and M. Ruggiero. "Are Gibbs-type priors the most natural generalization of the Dirichlet process?" In: *IEEE PAMI* 37.2 (2013), pp. 212–229.

[DB+21] V. De Bortoli, J. Thornton, J. Heng, and A. Doucet. "Diffusion Schrödinger Bridge with Applications to Score-Based Generative Modeling". In: *NIPS*. June 2021.

[DBB20] S. Daulton, M. Balandat, and E. Bakshy. "Differentiable Expected Hypervolume Improvement for Parallel Multi-Objective Bayesian Optimization". In: *NIPS*. 2020.

[DBP19] C. Durkan, A. Bekasov, and I. M. G. Papamakarios. "Neural Spline Flows". In: *NIPS*. 2019.

[DBW20] I. A. Delbridge, D. S. Bindel, and A. G. Wilson. "Randomly Projected Additive Gaussian Processes for Regression". In: *International Conference on Machine Learning*. 2020.

[DCF+15] E. Denton, S. Chintala, R. Fergus, et al. "Deep generative image models using a Laplacian pyramid of adversarial networks". In: *NIPS*. 2015.

[DD22] S. Doyen and N. B. Dadario. "12 Plagues of AI in Healthcare: A Practical Guide to Current Issues With Using Machine Learning in a Medical Context". en. In: *Front Digit Health* 4 (May 2022), p. 765406.

[DDL97] S. DellaPietra, V. DellaPietra, and J. Lafferty. "Inducing features of random fields". In: *IEEE PAMI* 19.4 (1997).

[DE00] R. Dahlhaus and M. Eichler. "Causality and graphical models for time series". In: *Highly structured stochastic systems*. Ed. by P. Green, N. Hjort, and S. Richardson. Oxford University Press, 2000.

[DE04] J. Dow and J. Endersby. "Multinomial probit and multinomial logit: a comparison of choice models for voting research". In: *Electoral Studies* 23.1 (2004), pp. 107–122.

[Dec96] R. Dechter. "Bucket elimination: a unifying framework for probabilistic inference". In: *UAI*. 1996.

[DeG70] M. DeGroot. *Optimal Statistical Decisions*. McGraw-Hill, 1970.

[DEL20] H. M. Dolatabadi, S. Erfani, and C. Leckie. "Invertible Generative Modeling using Linear Rational Splines". In: *AISTATS*. 2020, pp. 4236–4246.

[Del+21] M. Delange, R. Aljundi, M. Masana, S. Parisot, X. Jia, A. Leonardis, G. Slabaugh, and T. Tuytelaars. "A continual learning survey: Defying forgetting in classification tasks". en. In: *IEEE PAMI* (2021).

[Den+02] D. Denison, C. Holmes, B. Mallick, and A. Smith. *Bayesian methods for nonlinear classification and regression*. Wiley, 2002.

[Den+20] Y. Deng, A. Bakhtin, M. Ott, A. Szlam, and M. Ranzato. "Residual Energy-Based Models for Text Generation". In: *International Conference on Learning Representations*. 2020.

[Dev+18] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova. "Bert: Pre-training of deep bidirectional transformers for language understanding". In: *arXiv preprint arXiv:1810.04805* (2018).

[Dev+19] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova. "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding". In: *NAACL*. 2019.

[Dev+21] L. Devlin, P. Horridge, P. L. Green, and S. Maskell. "The No-U-Turn Sampler as a Proposal Distribution in a Sequential Monte Carlo Sampler with a Near-Optimal L-Kernel". In: (2021). arXiv: 2108.02498 [stat.CO].

[Dev85] P. A. Devijver. "Baum's forward-backward algorithm revisited". In: *Pattern Recognition Letters* 3.6 (1985), pp. 369–373.

[Dex] *Dex: Research language for array processing in the Haskell/ML family*. https://github.com/google-research/dex-lang. 2019.

[DF18] E. Denton and R. Fergus. "Stochastic Video Generation with a Learned Prior". In: *ICML*. 2018.

[DF19] X. Ding and D. J. Freedman. "Learning Deep Generative Models with Annealed Importance Sampling". In: (2019). arXiv: 1906.04904 [stat.ML].

[DF21] F. D'Angelo and V. Fortuin. "Repulsive Deep Ensembles are Bayesian". In: *NIPS*. May 2021.

[DFF21] S. Dozinski, U. Feige, and M. Feldman. "Are Gross Substitutes a Substitute for Submodular Valuations?" In: *arXiv preprint arXiv:2102.13343* (2021).

[DFO20] M. Deisenroth, A. Faisal, and C. S. Ong. *Mathematics for machine learning*. Cambridge, 2020.

[DFR15] M. P. Deisenroth, D. Fox, and C. E. Rasmussen. "Gaussian Processes for Data-Efficient Learning in Robotics and Control". en. In: *IEEE PAMI* 37.2 (2015), pp. 408–423.

[DFS16] A. Daniely, R. Frostig, and Y. Singer. "Toward Deeper Understanding of Neural Networks: The Power of Initialization and a Dual View on Expressivity". In: *NIPS*. 2016, pp. 2253–2261.

[DG17] P. Dabkowski and Y. Gal. "Real time image saliency for black box classifiers". In: *NeurIPS* (2017).

[DG84] P. J. Diggle and R. J. Gratton. "Monte Carlo methods of inference for implicit statistical models". In: *Journal of the Royal Statistical Society. Series B (Methodological)* (1984), pp. 193–227.

[DGA00] A. Doucet, S. Godsill, and C. Andrieu. "On sequential Monte Carlo Sampling Methods for Bayesian Filtering". In: *Statistics and Computing* 10.3 (2000), pp. 197–208.

[DGE15] C. Doersch, A. Gupta, and A. A. Efros. "Unsupervised visual representation learning by context prediction". In: *Proceedings of the IEEE international conference on computer vision*. 2015, pp. 1422–1430.

[DGF20] Y. Dubois, J. Gordon, and A. Y. Foong. *Neural Process Family*. http://yanndubs.github.io/Neural-Process-Family/. 2020.

[DGK01] A. Doucet, N. Gordon, and V. Krishnamurthy. "Particle Filters for State Estimation of Jump Markov Linear Systems". In: *IEEE Trans. on Signal Processing* 49.3 (2001), pp. 613–624.

[DH22] F. Dellaert and S. Hutchinson. *Introducion to Robotics and Perception*. 2022.

[DHK14] A. Deshpande, L. Hellerstein, and D. Kletenik. "Approximation algorithms for stochastic boolean function evaluation and stochastic submodular set cover". In: *Proceedings of the twenty-fifth annual ACM-SIAM Symposium on Discrete Algorithms*. SIAM. 2014, pp. 1453–1466.

[Dia88] P. Diaconis. "Sufficiency as statistical symmetry". In: *Proceedings of the AMS Centennial Symposium*. 1988, pp. 15–26.

[Die00] T. G. Dietterich. "Ensemble Methods in Machine Learning". In: *Multiple Classifier Systems*. Springer Berlin Heidelberg, 2000, pp. 1–15.

[Die+07] M. Diehl, H. G. Bock, H. Diedam, and P.-B. Wieber. "Fast Direct Multiple Shooting Algorithms for Optimal Robot Control". In: *Lecture Notes in Control and Inform. Sci.* 340 (2007).

[Die10] L. Dietz. *Directed Factor Graph Notation for Generative Models*. Tech. rep. MPI, 2010.

[Die+17] A. B. Dieng, C. Wang, J. Gao, and J. Paisley. "TopicRNN: A Recurrent Neural Network with Long-Range Semantic Dependency". In: *ICLR*. 2017.

[Die+19a] A. B. Dieng, Y. Kim, A. M. Rush, and D. M. Blei. "Avoiding Latent Variable Collapse With Generative Skip Models". In: *AISTATS*. 2019.

[Die+19b] A. B. Dieng, F. J. Ruiz, D. M. Blei, and M. K. Titsias. "Prescribed generative adversarial networks". In: *arXiv preprint arXiv:1910.04302* (2019).

[Die22] S. Dieleman. *Guidance: a cheat code for diffusion models*. 2022.

[Dik+20] N. Dikkala, G. Lewis, L. Mackey, and V. Syrgkanis. "Minimax Estimation of Conditional Moment Models". In: *Advances in Neural Information Processing Systems*. 2020.

[Din+17] L. Dinh, R. Pascanu, S. Bengio, and Y. Bengio. "Sharp Minima Can Generalize For Deep Nets". In: (2017). arXiv: 1703.04933 [cs.LG].

[DJ11] A. Doucet and A. M. Johansen. "A Tutorial on Particle Filtering and Smoothing: Fifteen years later". In: *Handbook of Nonlinear Filtering*. Ed. by D Crisan and B Rozovsk. 2011.

[DJ21] K. Desai and J. Johnson. *VirTex: Learning Visual Representations from Textual Annotations*. 2021. arXiv: 2006.06666 [cs.CV].

[DJK18] J. Djolonga, S. Jegelka, and A. Krause. "Provable Variational Inference for Constrained Log-Submodular Models". In: *NeurIPS*. 2018.

[DK12] J. Durbin and S. J. Koopman. *Time Series Analysis by State Space Methods: Second Edition*. en. Revised ed. edition. Oxford University Press, 2012.

[DK14] J. Djolonga and A. Krause. "From map to marginals: Variational inference in bayesian submodular models". In: *Advances in Neural Information Processing Systems*. 2014, pp. 244–252.

[DK15a] J. Djolonga and A. Krause. "Scalable variational inference in log-supermodular models". In: *International Conference on Machine Learning*. PMLR. 2015, pp. 1804–1813.

[DK15b] G. Durrett and D. Klein. "Neural CRF Parsing". In: *Proc. ACL*. 2015.

[DKB15] L. Dinh, D. Krueger, and Y. Bengio. "NICE: Non-linear Independent Components Estimation". In: *ICLR*. 2015.

[DKD16] J. Donahue, P. Krahenbuhl, and T. Darrell. "Adversarial feature learning". In: *arXiv preprint arXiv:1605.09782* (2016).

[DKD17] J. Donahue, P. Krähenbühl, and T. Darrell. *Adversarial Feature Learning*. 2017. arXiv: 1605.09782 [cs.LG].

[DKS13] J. Dick, F. Y. Kuo, and I. H. Sloan. "High-dimensional integration: the quasi-Monte Carlo way". In: *Acta Numerica* 22 (2013), 133–288.

[DL09] P. Domingos and D. Lowd. *Markov Logic: An Interface Layer for AI*. Morgan & Claypool, 2009.

[DL10] J. V. Dillon and G. Lebanon. "Stochastic Composite Likelihood". In: *J. Mach. Learn. Res.* 11 (2010), pp. 2597–2633.

[DL13] A. Damianou and N. Lawrence. "Deep Gaussian Processes". en. In: *AISTATS*. 2013, pp. 207–215.

[DL93] P. Dagum and M. Luby. "Approximating probabilistic inference in Bayesian belief networks is NP-hard". In: *Artificial Intelligence* 60 (1993), pp. 141–153.

[DLB17] C. Dann, T. Lattimore, and E. Brunskill. "Unifying PAC and Regret: Uniform PAC Bounds for Episodic Reinforcement Learning". In: *NIPS*. 2017, pp. 5717–5727.

[DLM09] H. Daumé III, J. Langford, and D. Marcu. "Search-based Structured Prediction". In: *MLJ* 75.3 (2009), pp. 297–325.

[DLM99] B. Delyon, M. Lavielle, and E. Moulines. "Convergence of a stochastic approximation version of the EM algorithm". In: *Annals of Statistics* 27.1 (1999), pp. 94–128.

[DLR77] A. P. Dempster, N. M. Laird, and D. B. Rubin. "Maximum likelihood from incomplete data via the EM algorithm". In: *J. of the Royal Statistical Society, Series B* 34 (1977), pp. 1–38.

[DLT21] M. De Lange and T. Tuytelaars. "Continual Prototype Evolution: Learning Online from Non-Stationary Data Streams". In: *ICCV*. 2021.

[DM01] D. van Dyk and X.-L. Meng. "The Art of Data Augmentation". In: *J. Computational and Graphical Statistics* 10.1 (2001), pp. 1–50.

[DM19a] Y. Du and I. Mordatch. "Implicit Generation and Generalization in Energy-Based Models". In: (2019). arXiv: 1903.08689 [cs.LG].

[DM19b] Y. Du and I. Mordatch. "Implicit Generation and Modeling with Energy Based Models". In: *NIPS*. 2019, pp. 3608–3618.

[DM22] A. P. Dawid and M. Musio. "Effects of Causes and Causes of Effects". In: *Annu. Rev. Stat. Appl.* 9.1 (Mar. 2022), pp. 261–287.

[DMDJ12] P. Del Moral, A. Doucet, and A. Jasra. "An adaptive sequential Monte Carlo method for approximate Bayesian computation". In: *Stat. Comput.* 22.5 (2012), pp. 1009–1020.

[DMKM22] G. Duran-Martin, A. Kara, and K. Murphy. "Efficient Online Bayesian Inference for Neural Bandits". In: *AISTATS*. 2022.

[DMM17] A. P. Dawid, M. Musio, and R. Murtas. "The probability of causation". In: *Law, Probability and Risk* 16.4 (2017), pp. 163–179.

[DMP18] C. Donahue, J. McAuley, and M. Puckette. "Adversarial Audio Synthesis". In: *International Conference on Learning Representations*. 2018.

[DMV15] S. Dash, D. M. Malioutov, and K. R. Varshney. "Learning interpretable classification rules using sequential rowsampling". In: *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE. 2015, pp. 3337–3341.

[DN21] P. Dhariwal and A. Q. Nichol. "Diffusion Models Beat GANs on Image Synthesis". In: *NIPS*. May 2021.

[Dnp] .

[DNR11] D. Duvenaud, H. Nickisch, and C. E. Rasmussen. "Additive Gaussian Processes". In: *NIPS*. 2011.

[Dom+06] P. Domingos, S. Kok, H. Poon, M. Richardson, and P. Singla. "Unifying Logical and Statistical AI". In: *IJCAI*. 2006.

[Dom+19] A.-K. Dombrowski, M. Alber, C. J. Anders, M. Ackermann, K.-R. Müller, and P. Kessel. "Explanations can be manipulated and geometry is to blame". In: *arXiv preprint arXiv:1906.07983* (2019).

[Don+14] J. Donahue, Y. Jia, O. Vinyals, J. Hoffman, N. Zhang, E. Tzeng, and T. Darrell. "Decaf: A deep convolutional activation feature for generic visual recognition". In: *International conference on machine learning*. 2014, pp. 647–655.

[Don+17a] K. Dong, D. Eriksson, H. Nickisch, D. Bindel, and A. G. Wilson. "Scalable Log Determinants for Gaussian Process Kernel Learning". In: *NIPS*. 2017, pp. 6327–6337.

[Don+17b] Y. Dong, H. Su, J. Zhu, and F. Bao. "Towards interpretable deep neural networks by leveraging adversarial examples". In: *arXiv preprint arXiv:1708.05493* (2017).

[Don+21] X. Dong, J. Bao, T. Zhang, D. Chen, W. Zhang, L. Yuan, D. Chen, F. Wen, and N. Yu. "Peco: Perceptual codebook for bert pre-training of vision transformers". In: *arXiv preprint arXiv:2111.12710* (2021).

[Dor+16] V. Dorie, M. Harada, N. B. Carnegie, and J. Hill. "A flexible, interpretable framework for assessing sensitivity to unmeasured confounding". In: *Statistics in Medicine* 35.20 (2016), pp. 3453–3470. eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/sim.6973.

[Dos+21] A. Dosovitskiy et al. "An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale". In: *International Conference on Learning Representations*. 2021.

[Doy+07] K. Doya, S. Ishii, A. Pouget, and R. P. N. Rao, eds. *Bayesian Brain: Probabilistic Approaches to Neural Coding*. MIT Press, 2007.

[DR11] M. P. Deisenroth and C. E. Rasmussen. "PILCO: A Model-Based and Data-Efficient Approach to Policy Search". In: *ICML*. 2011.

[DR17] G. K. Dziugaite and D. M. Roy. "Computing nonvacuous generalization bounds for deep (stochastic) neural networks with many more parameters than training data". In: *UAI*. 2017.

[DRB19] A. B. Dieng, F. J. R. Ruiz, and D. M. Blei. "The Dynamic Embedded Topic Model". In: (2019). arXiv: 1907.05545 [cs.CL].

[DRG15] G. K. Dziugaite, D. M. Roy, and Z. Ghahramani. "Training generative neural networks via Maximum Mean Discrepancy optimization". In: *ICML*. 2015.

[Dru08] J. Drugowitsch. *Bayesian linear regression*. Tech. rep. U. Rochester, 2008.

[DS18] J. Domke and D. R. Sheldon. "Importance Weighting and Variational Inference". In: *NIPS*. Curran Associates, Inc., 2018, pp. 4474–4483.

[DS19] J. Donahue and K. Simonyan. "Large scale adversarial representation learning". In: *arXiv preprint arXiv:1907.02544* (2019).

[DSDB17] L. Dinh, J. Sohl-Dickstein, and S. Bengio. "Density estimation using Real NVP". In: *ICLR*. 2017.

[DSK16] V. Dumoulin, J. Shlens, and M. Kudlur. "A Learned Representation For Artistic Style". In: (2016).

[DSZ16] A. Datta, S. Sen, and Y. Zick. "Algorithmic transparency via quantitative input influence: Theory and experiments with learning systems". In: *2016 IEEE symposium on security and privacy (SP)*. IEEE. 2016, pp. 598–617.

[DTK16] J. Djolonga, S. Tschiatschek, and A. Krause. "Variational inference in mixed probabilistic submodular models". In: *Advances in Neural Information Processing Systems*. 2016, pp. 1759–1767.

[Du+16] N. Du, H. Dai, R. Trivedi, U. Upadhyay, M. Gomez-Rodriguez, and L. Song. "Recurrent marked temporal point processes: Embedding event history to vector". In: *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. 2016, pp. 1555–1564.

[Du+18] J. Du, S. Ma, Y.-C. Wu, S. Kar, and J. M. F. Moura. "Convergence Analysis of Distributed Inference with Vector-Valued Gaussian Belief Propagation". In: *JMLR* 18.172 (2018), pp. 1–38.

[Du+19] S. S. Du, K. Hou, B. Póczos, R. Salakhutdinov, R. Wang, and K. Xu. "Graph Neural Tangent Kernel: Fusing Graph Neural Networks with Graph Kernels". In: (2019). arXiv: 1905.13192 [cs.LG].

[Du+20] Y. Du, S. Li, J. Tenenbaum, and I. Mordatch. "Improved Contrastive Divergence Training of Energy Based Models". In: *arXiv preprint arXiv:2012.01316* (2020).

[Du+21] C. Du, Z. Gao, S. Yuan, L. Gao, Z. Li, Y. Zeng, X. Zhu, J. Xu, K. Gai, and K.-C. Lee. "Exploration in Online Advertising Systems with Deep Uncertainty-Aware Learning". In: *KDD*. KDD '21. Association for Computing Machinery, 2021, pp. 2792–2801.

[Dua+20] S. Duan, N. Watters, L. Matthey, C. P. Burgess, A. Lerchner, and I. Higgins. "A Heuristic for Unsupervised Model Selection for Variational Disentangled Representation Learning". In: *ArXiv* abs/1905.12614 (2020).

[Dua+87] S. Duane, A. Kennedy, B. Pendleton, and D. Roweth. "Hybrid Monte Carlo". In: *Physics Letters B* 195.2 (1987), pp. 216–222.

[Dub+16] A. Dubey, S. J. Reddi, B. Póczos, A. J. Smola, E. P. Xing, and S. A. Williamson. "Variance Reduction in Stochastic Gradient Langevin Dynamics". en. In: *NIPS*. Vol. 29. 2016, pp. 1154–1162.

[Dud13] J. Duda. "Asymmetric numeral systems: entropy coding combining speed of Huffman coding with compression rate of arithmetic coding". In: (2013). arXiv: 1311.2540 [cs.IT].

[Duf02] M. Duff. "Optimal Learning: Computational procedures for Bayes-adaptive Markov decision processes". PhD thesis. U. Mass. Dept. Comp. Sci., 2002.

[Dum+16] V. Dumoulin, I. Belghazi, B. Poole, O. Mastropietro, A. Lamb, M. Arjovsky, and A. Courville. "Adversarially learned inference". In: *arXiv preprint arXiv:1606.00704* (2016).

[Dum+17] V. Dumoulin, I. Belghazi, B. Poole, O. Mastropietro, A. Lamb, M. Arjovsky, and A. Courville. *Adversarially Learned Inference*. 2017. arXiv: 1606.00704 [stat.ML].

[Dur+19] C. Durkan, A. Bekasov, I. Murray, and G. Papamakarios. "Cubic-Spline Flows". In: *ICML Workshop on Invertible Neural Networks and Normalizing Flows*. 2019.

[Dur+98] R. Durbin, S. Eddy, A. Krogh, and G. Mitchison. *Biological Sequence Analysis: Probabilistic Models of Proteins and Nucleic Acids*. Cambridge University Press, 1998.

[Duv+13] D. Duvenaud, J. Lloyd, R. Grosse, J. Tenenbaum, and G. Zoubin. "Structure Discovery in Nonparametric Regression through Compositional Kernel Search". en. In: *ICML*. 2013, pp. 1166–1174.

[Duv14] D. Duvenaud. "Automatic Model Construction with Gaussian Processes". PhD thesis. Computational and Biological Learning Laboratory, University of Cambridge, 2014.

[dV04] D. P. de Farias and B. Van Roy. "On Constraint Sampling in the Linear Programming Approach to Approximate Dynamic Programming". In: *Mathematics of Operations Research* 29.3 (2004), pp. 462–478.

[DV+17] H. De Vries, F. Strub, J. Mary, H. Larochelle, O. Pietquin, and A. C. Courville. "Modulating early visual processing by language". In: *Advances in Neural Information Processing Systems*. 2017, pp. 6594–6604.

[DV75] M. D. Donsker and S. S. Varadhan. "Asymptotic evaluation of certain Markov process expectations for large time, I". In: *Communications on Pure and Applied Mathematics* 28.1 (1975), pp. 1–47.

[DV99] A. P. Dawid and V. Vovk. "Prequential probability: Principles and properties". In: *Bernoulli* 5 (1999), pp. 125–162.

[DVK17] F. Doshi-Velez and B. Kim. *Towards A Rigorous Science of Interpretable Machine Learning*. 2017. eprint: 1702.08608 (stat.ML).

[DW19] B. Dai and D. Wipf. "Diagnosing and Enhancing VAE Models". In: *ICLR*. 2019.

[DWS12] T. Degris, M. White, and R. S. Sutton. "Off-Policy Actor-Critic". In: *ICML*. 2012.

[DWW19] B. Dai, Z. Wang, and D. Wipf. "The Usual Suspects? Reassessing Blame for VAE Posterior Collapse". In: (2019). arXiv: 1912.10702 [cs.LG].

[DWW20] B. Dai, Z. Wang, and D. Wipf. "The Usual Suspects? Reassessing Blame for VAE Posterior Collapse". In: *ICML*. Ed. by H. D. Iii and A. Singh. Vol. 119. Proceedings of Machine Learning Research. PMLR, 2020, pp. 2313–2322.

[DY79] P. Diaconis and D. Ylvisaker. "Conjugate priors for exponential families". In: vol. 7. 1979, pp. 269–281.

[ECM18] P. Etoori, M. Chinnakotla, and R. Mamidi. "Automatic Spelling Correction for Resource-Scarce Languages using Deep Learning". In: *Proceedings of ACL 2018, Student Research Workshop*. Association for Computational Linguistics, 2018, pp. 146–152.

[ED05] D. Earl and M. Deem. "Parallel tempering: Theory, applications, and new perspectives". In: *Phys. Chem. Chem. Phys.* 7 (2005), p. 3910.

[Edm69] H. P. Edmundson. "New methods in automatic extracting". In: *Journal of the ACM (JACM)* 16.2 (1969), pp. 264–285.

[Edm70] J. Edmonds. "Matroids, submodular functions, and certain polyhedra". In: *Combinatorial Structures and Their Applications* (1970), pp. 69–87.

[EFL04] E. Erosheva, S. Fienberg, and J. Lafferty. "Mixed-membership models of scientific publications". In: *PNAS* 101 (2004), pp. 5220–2227.

[Efr86] B. Efron. "Why Isn't Everyone a Bayesian?" In: *The American Statistician* 40.1 (1986).

[EGW05] D. Ernst, P. Geurts, and L. Wehenkel. "Tree-Based Batch Mode Reinforcement Learning". In: *JMLR* 6 (2005), pp. 503–556.

[Eis16] J. Eisner. "Inside-Outside and Forward-Backward Algorithms Are Just Backprop (Tutorial Paper)". In: *EMNLP Workshop on Structured Prediction for NLP*. 2016.

[Eke+13] M. Ekeberg, C. Lövkvist, Y. Lan, M. Weigt, and E. Aurell. "Improved contact prediction in proteins: using pseudolikelihoods to infer Potts models". en. In: *Phys. Rev. E Stat. Nonlin. Soft Matter Phys.* 87.1 (2013), p. 012707.

[EM07] D. Eaton and K. Murphy. "Exact Bayesian structure learning from uncertain interventions". In: *AI/Statistics*. 2007.

[EMH19] T. Elsken, J. H. Metzen, and F. Hutter. "Neural Architecture Search: A Survey". In: *JMLR* 20 (2019), pp. 1–21.

[EMK06] G. Elidan, I. McGraw, and D. Koller. "Residual belief propagation: Informed scheduling for asynchronous message passing". In: *UAI*. 2006.

[Eng+18] J. Engel, K. K. Agrawal, S. Chen, I. Gulrajani, C. Donahue, and A. Roberts. "GANSynth: Adversarial Neural Audio Synthesis". In: *International Conference on Learning Representations*. 2018.

[Erh+09] D. Erhan, Y. Bengio, A. Courville, and P. Vincent. "Visualizing higher-layer features of a deep network". In: *University of Montreal* 1341.3 (2009), p. 1.

[Erm+13] S. Ermon, C. P. Gomes, A. Sabharwal, and B. Selman. "Taming the Curse of Dimensionality: Discrete Integration by Hashing and Optimization". In: *ICML*. Feb. 2013.

[ERO21] P. Esser, R. Rombach, and B. Ommer. "Taming Transformers for High-Resolution Image Synthesis". In: *CVPR*. 2021.

[Eva+18] R. Evans et al. "De novo structure prediction with deep-learning based scoring". In: (2018).

[Eve09] G. Evensen. *Data Assimilation: The Ensemble Kalman Filter*. en. 2nd ed. 2009 edition. Springer, 2009.

[EW10] El-Yaniv and Wiener. "On the Foundations of Noise-free Selective Classification". In: *JMLR* (2010).

[EW95] M. D. Escobar and M. West. "Bayesian density estimation and inference using mixtures". In: *JASA* 90.430 (1995), pp. 577–588.

[Ewe72] W. J. Ewens. "The sampling theory of selectively neutral alleles". In: *Theoretical population biology* 3.1 (1972), pp. 87–112.

[Ewe90] W. Ewens. "Population genetics theory - the past and the future". In: *Mathemetical and Statistica Developments of Evolutionary Theory*. Ed. by S.Lessard. Reidel, 1990, pp. 177–227.

[EY09] M. Elad and I. Yavnch. "A plurality of sparse representations is better than the sparsest one alone". In: *IEEE Trans. on Info. Theory* 55.10 (2009), pp. 4701–4714.

[Eyk+18] K. Eykholt, I. Evtimov, E. Fernandes, B. Li, A. Rahmati, C. Xiao, A. Prakash, T. Kohno, and D. Song. "Robust Physical-World Attacks on Deep Learning Models". In: *CVPR*. 2018.

[Eys+21] B. Eysenbach, A. Khazatsky, S. Levine, and R. Salakhutdinov. "Mismatched No More: Joint Model-Policy Optimization for Model-Based RL". In: (2021). arXiv: 2110.02758 [cs.LG].

[FA20] I. Fischer and A. A. Alemi. "CEB Improves Model Robustness". In: *Entropy* 22.10 (2020), p. 1081.

[Fad+20] S. G. Fadel, S. Mair, R. da S. Torres, and U. Brefeld. "Principled Interpolation in Normalizing Flows". In: (2020). arXiv: 2010.12059 [stat.ML].

[Fag+18] F. Faghri, D. J. Fleet, J. R. Kiros, and S. Fidler. "VSE++: Improving Visual-Semantic Embeddings with Hard Negatives". In: *BMVC*. 2018.

[FAL17] C. Finn, P. Abbeel, and S. Levine. "Model-Agnostic Meta-Learning for Fast Adaptation of Deep Networks". In: *ICML*. 2017.

[Fan+17] H. Fang, N. Tian, Y. Wang, M. Zhou, and M. A. Haile. "Nonlinear Bayesian Estimation: From Kalman Filtering to a Broader Horizon". In: (2017). arXiv: 1712.01406 [cs.SY].

[Fau+18] L. Faury, F. Vasile, C. Calauzènes, and O. Fercoq. "Neural Generative Models for Global Optimization with Gradients". In: (2018). arXiv: 1805.08594 [cs.NE].

[FB19] E. M. Feit and R. Berman. "Test & Roll: Profit-Maximizing A/B Tests". In: *Marketing Science* 38.6 (2019), pp. 1038–1058.

[FBW21] M. Finzi, G. Benton, and A. G. Wilson. "Residual Pathway Priors for Soft Equivariance Constraints". In: *NIPS*. 2021.

[FC03] P. Fearnhead and P. Clifford. "On-line inference for hidden Markov models via particle filters". en. In: *J. of Royal Stat. Soc. Series B* 65.4 (2003), pp. 887–899.

[FD07a] B. Frey and D. Dueck. "Clustering by Passing Messages Between Data Points". In: *Science* 315 (2007), 972–976.

[FD07b] B. J. Frey and D. Dueck. "Clustering by passing messages between data points". In: *science* 315.5814 (2007), pp. 972–976.

[FDF19] A. M. Franks, A. D'Amour, and A. Feller. "Flexible Sensitivity Analysis for Observational Studies Without Observable Implications". In: *Journal of the American Statistical Association* 0.0 (2019), pp. 1–33. eprint: https://doi.org/10.1080/01621459.2019.1604369.

[FDZ19] A. Fasano, D. Durante, and G. Zanella. "Scalable and Accurate Variational Bayes for High-Dimensional Binary Regression Models". In: (2019). arXiv: 1911.06743 [stat.ME].

[FE73] M. Fischler and R. Elschlager. "The representation and matching of pictorial structures". In: *IEEE Trans. on Computer* 22.1 (1973).

[Fed+18] W. Fedus, M. Rosca, B. Lakshminarayanan, A. M. Dai, S. Mohamed, and I. Goodfellow. "Many Paths to Equilibrium: GANs Do Not Need to Decrease a Divergence At Every Step". In: *International Conference on Learning Representations*. 2018.

[Fei98] U. Feige. "A threshold of ln n for approximating set cover". In: *Journal of the ACM (JACM)* (1998).

[Fel+10] P. Felzenszwalb, R. Girshick, D. McAllester, and D. Ramanan. "Object Detection with Discriminatively Trained Part Based Models". In: *IEEE PAMI* 32.9 (2010).

[Fel+19] M. Fellows, A. Mahajan, T. G. J. Rudner, and S. Whiteson. "VIREL: A Variational Inference Framework for Reinforcement Learning". In: *NeurIPS*. 2019, pp. 7120–7134.

[Fen+21] S. Y. Feng, V. Gangal, J. Wei, S. Chandar, S. Vosoughi, T. Mitamura, and E. Hovy. "A Survey of Data Augmentation Approaches for NLP". In: (2021). arXiv: 2105.03075 [cs.CL].

[Fer73] T. S. Ferguson. "A Bayesian analysis of some nonparametric problems". In: *The Annals of Statistics* (1973), pp. 209–230.

[Feu+15] M. Feurer, A. Klein, K. Eggensperger, J. Springenberg, M. Blum, and F. Hutter. "Efficient and Robust Automated Machine Learning". In: *NIPS*. 2015, pp. 2962–2970.

[FG15] N. Fournier and A. Guillin. "On the rate of convergence in Wasserstein distance of the empirical measure". In: *Probability Theory and Related Fields* 162.3 (2015), pp. 707–738.

[FG18] S. Farquhar and Y. Gal. "Towards Robust Evaluations of Continual Learning". In: (2018). arXiv: 1805.09733 [stat.ML].

[FGG97] N. Friedman, D. Geiger, and M. Goldszmidt. "Bayesian network classifiers". In: *MLJ* 29 (1997), pp. 131–163.

[FH12] P. F. Felzenszwalb and D. P. Huttenlocher. "Distance Transforms of Sampled Functions". In: *Theory of Computing* 8.19 (2012), pp. 415–428.

[FH17] N. Frosst and G. Hinton. *Distilling a Neural Network Into a Soft Decision Tree*. 2017. arXiv: 1711.09784 [cs.LG].

[FH20] E. Fong and C. Holmes. "On the marginal likelihood and cross-validation". In: *Biometrika* 107.2 (2020).

[FH75] K Fukunaga and L Hostetler. "The estimation of the gradient of a density function, with applications in pattern recognition". In: *IEEE Trans. Inf. Theory* 21.1 (1975), pp. 32–40.

[FH97] B. J. Frey and G. Hinton. "Efficient stochastic source coding and an application to a Bayesian network source model". In: *Computer Journal* (1997).

[FHDV20] J. Futoma, M. C. Hughes, and F. Doshi-Velez. "Popcorn: Partially observed prediction constrained reinforcement learning". In: *AISTATS* (2020).

[FHK03] P. Felzenszwalb, D. Huttenlocher, and J. Kleinberg. "Fast Algorithms for Large State Space HMMs with Applications to Web Usage Analysis". In: *NIPS*. 2003.

[FHL19] S. Fort, H. Hu, and B. Lakshminarayanan. "Deep Ensembles: A Loss Landscape Perspective". In: (2019). arXiv: 1912.02757 [stat.ML].

[FHM18] S. Fujimoto, H. van Hoof, and D. Meger. "Addressing Function Approximation Error in Actor-Critic Methods". In: *ICLR*. 2018.

[FHT08] J. Friedman, T. Hastie, and R. Tibshirani. "Sparse inverse covariance estimation the graphical lasso". In: *Biostatistics* 9.3 (2008), pp. 432–441.

[FI10] A. Fischer and C. Igel. "Empirical analysis of the divergence of Gibbs sampling based learning algorithms for restricted Boltzmann machines". In: *International conference on artificial neural networks*. Springer. 2010, pp. 208–217.

[Fie70] S. Fienberg. "An Iterative Procedure for Estimation in Contingency Tables". In: *Annals of Mathematical Statistics* 41.3 (1970), pp. 907–917.

[Fin+16] C. Finn, P. Christiano, P. Abbeel, and S. Levine. "A connection between generative adversarial networks, inverse reinforcement learning, and energy-based models". In: *arXiv preprint arXiv:1611.03852* (2016).

[Fis20] I. Fischer. "The Conditional Entropy Bottleneck". In: *Entropy* 22.9 (2020).

[Fis25] R. Fisher. *Statistical Methods for Research Workers*. Biological monographs and manuals. Oliver and Boyd, 1925.

[FJ02] M. A. T. Figueiredo and A. K. Jain. "Unsupervised Learning of Finite Mixture Models". In: *IEEE PAMI* 24.3 (2002), pp. 381–396.

[FJL18] R. Frostig, M. J. Johnson, and C. Leary. "Compiling machine learning programs via high-level tracing". In: *Machine Learning and Systems (MLSys)* (2018).

[FK13a] V. Feldman and P. Kothari. "Learning Coverage Functions". In: *CoRR* abs/1304.2079 (2013). arXiv: 1304.2079.

[FK13b] M Frei and H. R. Künsch. "Bridging the ensemble Kalman and particle filters". In: *Biometrika* 100.4 (2013), pp. 781–800.

[FK14] V. Feldman and P. Kothari. "Learning Coverage Functions and Private Release of Marginals". In: *Proceedings of The 27th Conference on Learning Theory*. Ed. by M. F. Balcan, V. Feldman, and C. Szepesvári. Vol. 35. Proceedings of Machine Learning Research. Barcelona, Spain: PMLR, 2014, pp. 679–702.

[FK21] A. Fisher and E. H. Kennedy. "Visually Communicating and Teaching Intuition for Influence Functions". In: *The American Statistician* 75.2 (2021), pp. 162–172. eprint: https://doi.org/10.1080/00031305.2020.1717620.

[FKH17] S. Falkner, A. Klein, and F. Hutter. "Combining Hyperband and Bayesian Optimization". In: *NIPS 2017 Bayesian Optimization Workshop*. 2017.

[FKV13] V. Feldman, P. Kothari, and J. Vondrák. "Representation, Approximation and Learning of Submodular Functions Using Low-rank Decision Trees". In: *Proceedings of the 26th Annual Conference on Learning Theory*. Ed. by S. Shalev-Shwartz and I. Steinwart. Vol. 30. Proceedings of Machine Learning Research. Princeton, NJ, USA: PMLR, 2013, pp. 711–740.

[FKV14] V. Feldman, P. Kothari, and J. Vondrák. "Nearly tight bounds on l1 approximation of self-bounding functions". In: *CoRR, abs/1404.4702* 1 (2014).

[FKV17] V. Feldman, P. Kothari, and J. Vondrák. "Tight Bounds on ℓ_1 Approximation and Learning of Self-Bounding Functions". In: *International Conference on Algorithmic Learning Theory*. PMLR. 2017, pp. 540–559.

[FKV20] V. Feldman, P. Kothari, and J. Vondrák. "Tight bounds on ℓ_1 approximation and learning of self-bounding functions". In: *Theoretical Computer Science* 808 (2020), pp. 86–98.

[FL07] P. Fearnhead and Z. Liu. "Online Inference for Multiple Changepoint Problems". In: *J. of Royal Stat. Soc. Series B* 69 (2007), pp. 589–605.

[FL11] P. Fearnhead and Z. Liu. "Efficient Bayesian analysis of multiple changepoint models with dependence across segments". In: *Statistics and Computing* 21.2 (2011), pp. 217–229.

[FL+18] V. François-Lavet, P. Henderson, R. Islam, M. G. Bellemare, and J. Pineau. "An Introduction to Deep Reinforcement Learning". In: *Foundations and Trends in Machine Learning* 11.3 (2018).

[FLA16] C. Finn, S. Levine, and P. Abbeel. "Guided Cost Learning: Deep Inverse Optimal Control via Policy Optimization". In: *ICML*. 2016, pp. 49–58.

[Fla+16] S. Flaxman, D. Sejdinovic, J. P. Cunningham, and S. Filippi. "Bayesian Learning of Kernel Embeddings". In: *UAI*. 2016.

[FLL18] J. Fu, K. Luo, and S. Levine. "Learning Robust Rewards with Adverserial Inverse Reinforcement Learning". In: *ICLR*. 2018.

[FLL19] Y. Feng, L. Li, and Q. Liu. "A Kernel Loss for Solving the Bellman Equation". In: *NeurIPS*. 2019, pp. 15430–15441.

[FLMM21] D. T. Frazier, R. Loaiza-Maya, and G. M. Martin. "Variational Bayes in State Space Models: Inferential and Predictive Accuracy". In: (June 2021). arXiv: 2106.12262 [stat.ME].

[FMM18] M. Figurnov, S. Mohamed, and A. Mnih. "Implicit Reparameterization Gradients". In: *NIPS*. 2018.

[FMP19] S. Fujimoto, D. Meger, and D. Precup. "Off-Policy Deep Reinforcement Learning without Exploration". In: *ICML*. 2019, pp. 2052–2062.

[FNG00] N. de Freitas, M. Niranjan, and A. Gee. "Hierarchical Bayesian models for regularisation in sequential learning". In: *Neural Computation* 12.4 (2000), pp. 955–993.

[FNW78] M. Fisher, G. Nemhauser, and L. Wolsey. "An analysis of approximations for maximizing submodular set functions—II". In: *Polyhedral combinatorics* (1978), pp. 73–87.

[FO20] F. Farnia and A. Ozdaglar. "Do GANs always have Nash equilibria?" In: *Proceedings of the 37th International Conference on Machine Learning*. Ed. by H. D. III and A. Singh. Vol. 119. Proceedings of Machine Learning Research. PMLR, 2020, pp. 3029–3039.

[Fon+21] D. Fontanel, F. Cermelli, M. Mancini, and B. Caputo. "On the Challenges of Open World Recognition under Shifting Visual Domains". In: *ICRA*. 2021.

[For01] G. D. Forney. "Codes on graphs: normal realizations". In: *IEEE Trans. Inf. Theory* 47.2 (2001), pp. 520–548.

[For+18a] V. Fortuin, G. Dresdner, H. Strathmann, and G. Rätsch. "Scalable Gaussian Processes on Discrete Domains". In: (2018). arXiv: 1810.10368 [stat.ML].

[For+18b] M. Fortunato et al. "Noisy Networks for Exploration". In: *ICLR*. 2018.

[For+19] N. Ford, J. Gilmer, N. Carlini, and D. Cubuk. "Adversarial Examples Are a Natural Consequence of Test Error in Noise". In: (2019). arXiv: 1901.10513 [cs.LG].

[For21] V. Fortuin. "Priors in Bayesian Deep Learning: A Review". In: (2021). arXiv: 2105.06868 [stat.ML].

[For+95] J. Forbes, T. Huang, K. Kanazawa, and S. Russell. "The BATmobile: Towards a Bayesian Automated Taxi". In: *IJCAI*. 1995.

[Fot+14] N. Foti, J. Xu, D. Laird, and E. Fox. "Stochastic variational inference for hidden Markov models". In: *NIPS*. 2014, pp. 3599–3607.

[Fox+08] E. Fox, E. Sudderth, M. Jordan, and A. Willsky. "An HDP-HMM for Systems with State Persistence". In: *ICML*. 2008.

[Fox09] E. B. Fox. "Bayesian nonparametric learning of complex dynamical phenomena". PhD thesis. Massachusetts Institute of Technology, 2009.

[FP08] N Friel and A. N. Pettitt. "Marginal Likelihood Estimation via Power Posteriors". In: *J. of Royal Stat. Soc. Series B* 70.3 (2008), pp. 589–607.

[FP69] D. C. Fraser and J. E. Potter. "The optimum linear smoother as a combination of two optimum linear filters". In: *IEEE Trans. on Automatical Control* (1969), pp. 387–390.

[FPD09] P. Frazier, W. Powell, and S. Dayanik. "The knowledge-gradient policy for correlated normal beliefs". In: *INFORMS J. on Computing* 21.4 (2009), pp. 599–613.

[Fra08] A. Fraser. *Hidden Markov Models and Dynamical Systems*. SIAM Press, 2008.

[Fra+16] M. Fraccaro, S. K. Sønderby, U. Paquet, and O. Winther. "Sequential Neural Models with Stochastic Layers". In: *NIPS*. 2016.

[Fra18] P. I. Frazier. "Bayesian Optimization". In: *Recent Advances in Optimization and Modeling of Contemporary Problems*. INFORMS TutORials in Operations Research. INFORMS, 2018, pp. 255–278.

[Fre14] A. A. Freitas. "Comprehensible classification models: a position paper". In: *ACM SIGKDD explorations newsletter* 15.1 (2014), pp. 1–10.

[Fre+22] M. Freitag, D. Grangier, Q. Tan, and B. Liang. "High Quality Rather than High Model Probability: Minimum Bayes Risk Decoding with Neural Metrics". In: *TACL*. 2022.

[Fre98] B. Frey. *Graphical Models for Machine Learning and Digital Communication*. MIT Press, 1998.

[Fre99] R. M. French. "Catastrophic forgetting in connectionist networks". In: *Trends in Cognitive Science* (1999).

[Fri09] K. Friston. "The free-energy principle: a rough guide to the brain?" en. In: *Trends Cogn. Sci.* 13.7 (2009), pp. 293–301.

[Fro+13] A. Frome, G. Corrado, J. Shlens, S. Bengio, J. Dean, M. Ranzato, and T. Mikolov. "Devise: A deep visual-semantic embedding model". In: (2013).

[Fro+21] R. Frostig, M. Johnson, D. Maclaurin, A. Paszke, and A. Radul. "Decomposing reverse-mode automatic differentiation". In: *LAFI workshop at POPL 2021*. 2021.

[FS07] S. Fruhwirth-Schnatter. *Finite Mixture and Markov Switching Models*. Springer, 2007.

[FSF10] S. Fruhwirth-Schnatter and R. Fruhwirth. "Data Augmentation and MCMC for Binary and Multinomial Logit Models". In: *Statistical Modelling and Regression Structures*. Ed. by T. Kneib and G. Tutz. Springer, 2010, pp. 111–132.

[FST98] S. Fine, Y. Singer, and N. Tishby. "The Hierarchical Hidden Markov Model: Analysis and Applications". In: *Machine Learning* 32 (1998), p. 41.

[FT05] M. Fashing and C. Tomasi. "Mean shift is a bound optimization". en. In: *IEEE Trans. Pattern Anal. Mach. Intell.* 27.3 (2005), pp. 471–474.

[FT19] A. Finke and A. H. Thiery. "On the relationship between variational inference and adaptive importance sampling". In: (2019). arXiv: 1907.10477 [stat.ML].

[FT74] J. H. Friedman and J. W. Tukey. "A Projection Pursuit Algorithm for Exploratory Data Analysis". In: *IEEE Trans. Comput.* C-23.9 (1974), pp. 881–890.

[Fu15] M. Fu, ed. *Handbook of Simulation Optimization*. 1st ed. Springer-Verlag New York, 2015.

[Fu+17] Z. Fu, X. Tan, N. Peng, D. Zhao, and R. Yan. "Style transfer in text: Exploration and evaluation". In: *arXiv preprint arXiv:1711.06861* (2017).

[Fu+19] H. Fu, C. Li, X. Liu, J. Gao, A. Celikyilmaz, and L. Carin. "Cyclical Annealing Schedule: A Simple Approach to Mitigating KL Vanishing". In: *NAACL*. 2019.

[Fu+20] J. Fu, A. Kumar, O. Nachum, G. Tucker, and S. Levine. *D4RL: Datasets for Deep Data-Driven Reinforcement Learning*. arXiv:2004.07219. 2020.

[Fuj05] S. Fujishige. *Submodular functions and optimization*. Vol. 58. Elsevier Science, 2005.

[Ful+20] I. R. Fulcher, I. Shpitser, S. Marealle, and E. J. Tchetgen Tchetgen. "Robust inference on population indirect causal effects: the generalized front door criterion". In: *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 82.1 (2020), pp. 199–214. eprint: https://rss.onlinelibrary.wiley.com/doi/pdf/10.1111/rssb.12345.

[FV15] V. Feldman and J. Vondrák. "Tight Bounds on Low-Degree Spectral Concentration of Submodular and XOS Functions". In: *2015 IEEE 56th Annual Symposium on Foundations of Computer Science*. 2015, pp. 923–942.

[FV16] V. Feldman and J. Vondrák. "Optimal bounds on approximation of submodular and XOS functions by juntas". In: *SIAM Journal on Computing* 45.3 (2016), pp. 1129–1170.

[FV17] R. C. Fong and A. Vedaldi. "Interpretable explanations of black boxes by meaningful perturbation". In: *Proceedings of the IEEE international conference on computer vision*. 2017, pp. 3429–3437.

[FW12] N. Friel and J. Wyse. "Estimating the evidence – a review". In: *Stat. Neerl.* 66.3 (2012), pp. 288–308.

[FW21] R. Friedman and Y. Weiss. "Posterior Sampling for Image Restoration using Explicit Patch Priors". In: (2021). arXiv: 2104.09895 [cs.CV].

[FWW21] M. Finzi, M. Welling, and A. G. Wilson. "A Practical Method for Constructing Equivariant Multilayer Perceptrons for Arbitrary Matrix Groups". In: *ICML*. 2021.

[Gaj+19] A. Gajewski, J. Clune, K. O. Stanley, and J. Lehman. "Evolvability ES: Scalable and Direct Optimization of Evolvability". In: *Proc. of the Conf. on Genetic and Evolutionary Computation*. 2019.

[Gan07] M Ganapathiraju. "Application of language technologies in biology: Feature extraction and modeling for transmembrane helix prediction". en. PhD thesis. 2007.

[Gan+16a] Y Ganin, E Ustinova, H Ajakan, P Germain, and others. "Domain-adversarial training of neural networks". In: *JMLR* (2016).

[Gan+16b] Y. Ganin, E. Ustinova, H. Ajakan, P. Germain, H. Larochelle, F. Laviolette, M. Marchand, and V. Lempitsky. "Domain-adversarial training of neural networks". In: *The journal of machine learning research* 17.1 (2016), pp. 2096–2030.

[Gao+18] R. Gao, J. Xie, S.-C. Zhu, and Y. N. Wu. "Learning Grid-like Units with Vector Representation of Self-Position and Matrix Representation of Self-Motion". In: *arXiv preprint arXiv:1810.05597* (2018).

[Gao+20] R. Gao, E. Nijkamp, D. P. Kingma, Z. Xu, A. M. Dai, and Y. N. Wu. "Flow contrastive estimation of energy-based models". In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2020, pp. 7518–7528.

[Gär03] T. Gärtner. "A Survey of Kernels for Structured Data". In: *SIGKDD Explor. Newsl.* 5.1 (2003), pp. 49–58.

[GAR16] R. B. Grosse, S. Ancha, and D. M. Roy. "Measuring the reliability of MCMC inference with bidirectional Monte Carlo". In: *NIPS*. 2016.

[Gar+18a] J. Gardner, G. Pleiss, K. Q. Weinberger, D. Bindel, and A. G. Wilson. "GPyTorch: Blackbox Matrix-Matrix Gaussian Process Inference with GPU Acceleration". In: *NIPS*. Ed. by S Bengio, H Wallach, H Larochelle, K Grauman, N Cesa-Bianchi, and R Garnett. Curran Associates, Inc., 2018, pp. 7576–7586.

[Gar+18b] J. R. Gardner, G. Pleiss, R. Wu, K. Q. Weinberger, and A. G. Wilson. "Product Kernel Interpolation for Scalable Gaussian Processes". In: *AISTATS*. 2018.

[Gar+18c] T. Garipov, P. Izmailov, D. Podoprikhin, D. Vetrov, and A. G. Wilson. "Loss Surfaces, Mode Connectivity, and Fast Ensembling of DNNs". In: *NIPS*. 2018.

[Gar+18d] M. Garnelo, D. Rosenbaum, C. Maddison, T. Ramalho, D. Saxton, M. Shanahan, Y. W. Teh, D. Rezende, and S. M. A. Eslami. "Conditional Neural Processes". In: *ICML*. Ed. by J. Dy and A. Krause. Vol. 80. Proceedings of Machine Learning Research. PMLR, 2018, pp. 1704–1713.

[Gar+18e] M. Garnelo, J. Schwarz, D. Rosenbaum, F. Viola, D. J. Rezende, S. M. Ali Eslami, and Y. W. Teh. "Neural Processes". In: *ICML workshop on Theoretical Foundations and Applications of Deep Generative Models*. 2018.

[Gar+19] S. Garg, V. Perot, N. Limtiaco, A. Taly, E. H. Chi, and A. Beutel. "Counterfactual Fairness in Text Classification through Robustness". In: *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society*. AIES '19. Association for Computing Machinery, 2019, 219–226.

[Gar22] R. Garnett. *Bayesian Optimization*. in preparation. Cambridge University Press, 2022.

[Gas+19] J. Gasthaus, K. Benidis, Y. Wang, S. S. Rangapuram, D. Salinas, V. Flunkert, and T. Januschowski. "Probabilistic Forecasting with Spline Quantile Function RNNs". In: *ICML*. Vol. 89. Proceedings of Machine Learning Research. PMLR, 2019, pp. 1901–1910.

[GAZ19] A. Ghorbani, A. Abid, and J. Zou. "Interpretation of neural networks is fragile". In: *Proceedings of the AAAI Conference on Artificial Intelligence*. Vol. 33. 01. 2019, pp. 3681–3688.

[GB00] Z. Ghahramani and M. Beal. "Variational inference for Bayesian mixtures of factor analysers". In: *NIPS-12*. 2000.

[GB09] A. Guillory and J. Bilmes. "Label Selection on Graphs". In: *NIPS*. Vancouver, Canada, 2009.

[GB10] X. Glorot and Y. Bengio. "Understanding the difficulty of training deep feedforward neural networks". In: *AISTATS*. 2010, pp. 249–256.

[GB11] A. Guillory and J. Bilmes. "Active Semi-Supervised Learning using Submodular Functions". In: *UAI*. Barcelona, Spain: AUAI, 2011.

[GB+18] R. Gómez-Bombarelli, J. N. Wei, D. Duvenaud, J. M. Hernández-Lobato, B. Sánchez-Lengeling, D. Sheberla, J. Aguilera-Iparraguirre, T. D. Hirzel, R. P. Adams, and A. Aspuru-Guzik. "Automatic Chemical Design Using a Data-Driven Continuous Representation of Molecules". en. In: *American Chemical Society Central Science* 4.2 (2018), pp. 268–276.

[GBB11] X. Glorot, A. Bordes, and Y. Bengio. "Deep Sparse Rectifer Neural Networks". In: *AISTATS*. 2011.

[GBC16] I. Goodfellow, Y. Bengio, and A. Courville. *Deep Learning*. http://www.deeplearningbook.org. MIT Press, 2016.

[GBJ18] R. Giordano, T. Broderick, and M. I. Jordan. "Covariances, Robustness, and Variational Bayes". In: *JMLR* 19.51 (2018), pp. 1–49.

[GBP18] C. Gurau, A. Bewley, and I. Posner. "Dropout Distillation for Efficiently Estimating Model Confidence". 2018.

[GC11] M. Girolami and B. Calderhead. "Riemann manifold Langevin and Hamiltonian Monte Carlo methods". In: *J. of Royal Stat. Soc. Series B* 73.2 (2011), pp. 123–214.

[GC90] R. P. Goldman and E. Charniak. "Dynamic Construction of Belief Networks". In: *UAI*. 1990.

[GCW19] M. Gerber, N. Chopin, and N. Whiteley. "Negative association, ordering and convergence of resampling methods". In: *Ann. Stat.* 47.4 (2019), pp. 2236–2260.

[GD20] T. Geffner and J. Domke. "A Rule for Gradient Estimator Selection, with an Application to Variational Inference". In: *AISTATS*. Ed. by S. Chiappa and R. Calandra. Vol. 108. Proceedings of Machine Learning Research. PMLR, 2020, pp. 1803–1812.

[GDFY16] S. Ghosh, F. M. Delle Fave, and J. Yedidia. "Assumed Density Filtering Methods for Learning Bayesian Neural Networks". In: *AAAI*. 2016.

[Geb+21] T. Gebru, J. Morgenstern, B. Vecchione, J. W. Vaughan, H. Wallach, H. D. Iii, and K. Crawford. "Datasheets for datasets". In: *Communications of the ACM* 64.12 (2021), pp. 86–92.

[Ged+20] D. Gedon, N. Wahlström, T. B. Schön, and L. Ljung. "Deep State Space Models for Nonlinear System Identification". In: (Mar. 2020). arXiv: 2003.14162 [eess.SY].

[Gei+20a] R. Geirhos, J.-H. Jacobsen, C. Michaelis, R. Zemel, W. Brendel, M. Bethge, and F. A. Wichmann. "Shortcut Learning in Deep Neural Networks". In: *arXiv preprint arXiv:2004.07780* (2020).

[Gei+20b] R. Geirhos, J. Jacobsen, C. Michaelis, R. S. Zemel, W. Brendel, M. Bethge, and F. A. Wichmann. "Shortcut Learning in Deep Neural Networks". In: *CoRR* abs/2004.07780 (2020). arXiv: 2004.07780.

[Gel+04] A. Gelman, J. Carlin, H. Stern, and D. Rubin. *Bayesian data analysis*. 2nd edition. Chapman and Hall, 2004.

[Gel06] A. Gelman. "Prior distributions for variance parameters in hierarchical models (comment on article by Browne and Draper)". en. In: *Bayesian Anal.* 1.3 (2006), pp. 515–534.

[Gel+08] A. Gelman, A. Jakulin, M. G. Pittau, and Y.-S. Su. "A weakly informative default prior distribution for logistic and other regression models". en. In: *The Annals of Applied Statistics* 2.4 (2008), pp. 1360–1383.

[Gel+14a] A. Gelman, J. B. Carlin, H. S. Stern, D. B. Dunson, A. Vehtari, and D. B. Rubin. *Bayesian Data Analysis, Third Edition*. Third edition. Chapman and Hall/CRC, 2014.

[Gel+14b] A. Gelman, A. Vehtari, P. Jylänki, C. Robert, N. Chopin, and J. P. Cunningham. "Expectation propagation as a way of life". In: (2014). arXiv: 1412.4869 [stat.CO].

[Gel+20] A. Gelman, A. Vehtari, D. Simpson, C. C. Margossian, B. Carpenter, Y. Yao, L. Kennedy, J. Gabry, P.-C. Bürkner, and M. Modrák. "Bayesian Workflow". In: (2020). arXiv: 2011.01808 [stat.ME].

[Gel+22] A. Gelman, J. Hill, B. Goodrich, J. Gabry, D. Simpson, and A. Vehtari. *Applied Regression and Multilevel Models*. To appear. 2022.

[Gel74] A. Gelb. *Applied Optimal Estimation*. en. The MIT Press, May 1974.

[Gel90] M. Gelbrich. "On a formula for the L2 Wasserstein metric between measures on Euclidean and Hilbert spaces". In: *Mathematische Nachrichten* 147.1 (1990), pp. 185–203.

[Gen+19] A. Genevay, L. Chizat, F. Bach, M. Cuturi, and G. Peyré. "Sample complexity of sinkhorn divergences". In: *The 22nd International Conference on Artificial Intelligence and Statistics*. PMLR. 2019, pp. 1574–1583.

[Geo+17] D. George et al. "A generative vision model that trains with high data efficiency and breaks text-based CAPTCHAs". In: *Science* 358.6368 (2017), eaag2612. eprint: https://www.science.org/doi/pdf/10.1126/science.aag2612.

[Geo+18] T. George, C. Laurent, X. Bouthillier, N. Ballas, and P. Vincent. "Fast Approximate Natural Gradient Descent in a Kronecker Factored Eigenbasis". In: *NIPS*. Curran Associates, Inc., 2018, pp. 9550–9560.

[Geo88] H.-O. Georgii. *Gibbs Measures and Phase Transitions*. en. Walter De Gruyter Inc, 1988.

[Ger+15] M. Germain, K. Gregor, I. Murray, and H. Larochelle. "MADE: Masked Autoencoder for Distribution Estimation". In: *ICML*. Ed. by F. Bach and D. Blei. Vol. 37. Proceedings of Machine Learning Research. PMLR, 2015, pp. 881–889.

[Gér19] A. Géron. *Hands-On Machine Learning with Scikit-Learn and TensorFlow: Concepts, Tools, and Techniques for Building Intelligent Systems (2nd edition)*. en. O'Reilly Media, Incorporated, 2019.

[Ger19] S. J. Gershman. "What does the free energy principle tell us about the brain?" In: *Neurons, Behavior, Data Analysis, and Theory* (2019).

[GEY19] Y. Geifman and R. El-Yaniv. "SelectiveNet: A Deep Neural Network with an Integrated Reject Option". In: *ICML*. 2019.

[Gey92] C. Geyer. "Practical Markov chain Monte Carlo". In: *Statistical Science* 7 (1992), pp. 473–483.

[GF00] E. George and D. Foster. "Calibration and empirical Bayes variable selection". In: *Biometrika* 87.4 (2000), pp. 731–747.

[GF09] I. E. Givoni and B. J. Frey. "A Binary Variable Model for Affinity Propagation". In: *Neural Computation* 21.6 (2009), pp. 1589–1600.

[GF+15] Á. F. García-Fernández, L. Svensson, M. R. Morelande, and S. Särkkä. "Posterior Linearization Filter: Principles and Implementation Using Sigma Points". In: *IEEE Trans. Signal Process.* 63.20 (Oct. 2015), pp. 5561–5573.

[GF17] B. Goodman and S. Flaxman. "European Union regulations on algorithmic decision-making and a "right to explanation"". In: *AI magazine* 38.3 (2017), pp. 50–57.

[GFSS17] Á. F. García-Fernández, L. Svensson, and S. Särkkä. "Iterated Posterior Linearization Smoother". In: *IEEE Trans. Automat. Contr.* 62.4 (2017), pp. 2056–2063.

[GFTS19] Á. F. García-Fernández, F. Tronarp, and S. Särkkä. "Gaussian Process Classification Using Posterior Linearization". In: *IEEE Signal Process. Lett.* 26.5 (2019), pp. 735–739.

[GFWO20] T. Galy-Fajou, F. Wenzel, and M. Opper. "Automated Augmented Conjugate Inference for Non-conjugate Gaussian Process Models". In: *AISTATS*. 2020.

[GG11] T. L. Griffiths and Z. Ghahramani. "The Indian Buffet Process: An Introduction and Review." In: *JMLR* 12.4 (2011).

[GG14] S. J. Gershman and N. D. Goodman. "Amortized Inference in Probabilistic Reasoning". In: *36th Annual Conference of the Cognitive Science Society*. 2014.

[GG16] Y. Gal and Z. Ghahramani. "Dropout as a Bayesian Approximation: Representing Model Uncertainty in Deep Learning". In: *ICML*. 2016.

[GG84] S. Geman and D. Geman. "Stochastic Relaxation, Gibbs Distributions, and the Bayesian Restoration of Images". In: *IEEE PAMI* 6.6 (1984).

[GGG15] M. Gygli, H. Grabner, and L. Gool. "Video summarization by learning submodular mixtures of objectives". In: *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (2015), pp. 3090–3098.

[GGS21] M. Grcic, I. Grubisic, and S. Segvic. "DenseFlow: Official implementation of Densely connected normalizing flows". en. In: *NIPS*. 2021.

[GH07] A. Gelman and J. Hill. *Data analysis using regression and multilevel/ hierarchical models*. Cambridge, 2007.

[GH10] M. Gutmann and A. Hyvärinen. "Noise-contrastive estimation: A new estimation principle for unnormalized statistical models". In: *Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics*. 2010, pp. 297–304.

[GH12] M. Gutmann and J.-i. Hirayama. "Bregman divergence as general framework to estimate unnormalized statistical models". In: *arXiv:1202.3727* (2012).

[GH96a] Z. Ghahramani and G. Hinton. *Parameter estimation for linear dynamical systems*. Tech. rep. CRG-TR-96-2. Dept. Comp. Sci., Univ. Toronto, 1996.

[GH96b] Z. Ghahramani and G. Hinton. *The EM Algorithm for Mixtures of Factor Analyzers*. Tech. rep. Dept. of Comp. Sci., Uni. Toronto, 1996.

[GH98] Z. Ghahramani and G. Hinton. "Variational learning for switching state-space models". In: *Neural Computation* 12.4 (1998), pp. 963–996.

[Gha+15] M. Ghavamzadeh, S. Mannor, J. Pineau, and A. Tamar. "Bayesian Reinforcement Learning: A Survey". In: *Foundations and Trends in ML* (2015).

[Gha+21] A. Ghandeharioun, B. Kim, C.-L. Li, B. Jou, B. Eoff, and R. W. Picard. *DISSECT: Disentangled Simultaneous Explanations via Concept Traversals*. 2021. arXiv: 2105.15164 [cs.LG].

[GHC20] C. Geng, S.-J. Huang, and S. Chen. "Recent Advances in Open Set Recognition: A Survey". In: *IEEE PAMI* (2020).

[GHC21] S. Gould, R. Hartley, and D. J. Campbell. "Deep Declarative Networks". en. In: *IEEE PAMI* PP (2021).

[GHK17] Y. Gal, J. Hron, and A. Kendall. "Concrete Dropout". In: (2017). arXiv: 1705.07832 [stat.ML].

[Gho+19] A. Ghorbani, J. Wexler, J. Zou, and B. Kim. *Towards Automatic Concept-based Explanations*. 2019. arXiv: 1902.03129 [stat.ML].

[Gho+21] A. Ghosh, A. Honoré, D. Liu, G. E. Henter, and S. Chatterjee. "Normalizing Flow based Hidden Markov Models for Classification of Speech Phones with Explainability". In: (July 2021). arXiv: 2107.00730 [cs.LG].

[GHV14] A. Gelman, J. Hwang, and A. Vehtari. "Understanding predictive information criteria for Bayesian models". In: *Statistics and Computing* 24.6 (2014), pp. 997–1016.

[GHV20] A. Gelman, J. Hill, and A. Vehtari. *Regression and Other Stories*. Cambridge, 2020.

[GIG17] Y. Gal, R. Islam, and Z. Ghahramani. "Deep bayesian active learning with image data". In: *International Conference on Machine Learning*. PMLR. 2017, pp. 1183–1192.

[Gil+18a] J. Gilmer, R. P. Adams, I. Goodfellow, D. Andersen, and G. E. Dahl. "Motivating the rules of the game for adversarial example research". In: *arXiv preprint arXiv:1807.06732* (2018).

[Gil+18b] J. Gilmer, L. Metz, F. Faghri, S. S. Schoenholz, M. Raghu, M. Wattenberg, and I. Goodfellow. "Adversarial spheres". In: *arXiv preprint arXiv:1801.02774* (2018).

[Gil88] J. R. Gilbert. "Some nested dissection order is nearly optimal". In: *Inf. Process. Lett.* 26.6 (1988), pp. 325–328.

[Gir+14] R. Girshick, J. Donahue, T. Darrell, and J. Malik. "Rich feature hierarchies for accurate object detection and semantic segmentation". In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2014, pp. 580–587.

[Gir+15] R. Girshick, F. Iandola, T. Darrell, and J. Malik. "Deformable Part Models are Convolutional Neural Networks". In: *CVPR*. 2015.

[Gir+21] L. Girin, S. Leglaive, X. Bie, J. Diard, T. Hueber, and X. Alameda-Pineda. "Dynamical Variational Autoencoders: A Comprehensive Review". In: *Foundations and Trends® in Machine Learning* 15.1-2 (2021), pp. 1–175.

[Git89] J. Gittins. *Multi-armed Bandit Allocation Indices*. Wiley, 1989.

[GJ97] Z. Ghahramani and M. Jordan. "Factorial Hidden Markov Models". In: *Machine Learning* 29 (1997), pp. 245–273.

[GK19] L. Graesser and W. L. Keng. *Foundations of Deep Reinforcement Learning: Theory and Practice in Python*. en. 1 edition. Addison-Wesley Professional, 2019.

[GKS05] C. Guestrin, A. Krause, and A. P. Singh. "Near-optimal sensor placements in gaussian processes". In: *Proceedings of the 22nd international conference on Machine learning*. 2005, pp. 265–272.

[GL10] K. Gregor and Y. LeCun. "Learning fast approximations of sparse coding". In: *ICML*. 2010, pp. 399–406.

[GL97] F. Glover and M. Laguna. Kluwer Academic Publishers, 1997.

[Gla03] P. Glasserman. *Monte Carlo Methods in Financial Engineering*. 1st ed. Stochastic Modelling and Applied Probability. Springer-Verlag New York, 2003.

[Gle02] S. Glennan. "Rethinking mechanistic explanation". In: *Philosophy of science* 69.S3 (2002), S342–S353.

[GLM15] J. Ghosh, Y. Li, and R. Mitra. "On the Use of Cauchy Prior Distributions for Bayesian Logistic Regression". In: (2015). arXiv: 1507.07170 [stat.ME].

[GLP21] I. Gulrajani and D. Lopez-Paz. "In Search of Lost Domain Generalization". In: *ICLR*. 2021.

[GLS81] M. Grötschel, L. Lovász, and A. Schrijver. "The ellipsoid method and its consequences in combinatorial optimization". In: *Combinatorica* 1.2 (1981), pp. 169–197.

[GM12] G. Gordon and J. McNulty. *Matroids: a geometric introduction*. Cambridge University Press, 2012.

[GM15] J. Gorham and L. Mackey. "Measuring sample quality with Stein's method". In: *Advances in Neural Information Processing Systems*. 2015, pp. 226–234.

[GM16] R. Grosse and J. Martens. "A Kronecker-factored approximate Fisher matrix for convolution layers". In: *ICML*. 2016.

[GM17] P. L. Green and S Maskell. "Estimating the parameters of dynamical systems from Big Data using Sequential Monte Carlo samplers". In: *Mech. Syst. Signal Process.* 93 (2017), pp. 379–396.

[GM98] A. Gelman and X.-L. Meng. "Simulating normalizing constants: from importance sampling to bridge sampling to path sampling". In: *Statisical Science* 13 (1998), pp. 163–185.

[GMAR16] Y. Gal, R. T. Mc Allister, and C. E. Rasmussen. "Improving PILCO with Bayesian Neural Network Dynamics Models". In: *ICML workshop on Data-efficient machine learning*. 2016.

[GMH20] M. Gorinova, D. Moore, and M. Hoffman. "Automatic Reparameterisation of Probabilistic Programs". In: *ICML*. Vol. 119. Proceedings of Machine Learning Research. PMLR, 2020, pp. 3648–3657.

[GNM19] D. Greenberg, M. Nonnenmacher, and J. Macke. "Automatic Posterior Transformation for Likelihood-Free Inference". In: *ICML*. 2019.

[Goe+09] M. X. Goemans, N. J. Harvey, S. Iwata, and V. Mirrokni. "Approximating submodular functions everywhere". In: *Proceedings of the twentieth annual ACM-SIAM symposium on Discrete algorithms*. SIAM. 2009, pp. 535–544.

[Gol+04] J. Goldberger, S. T. Roweis, G. E. Hinton, and R. Salakhutdinov. "Neighbourhood Components Analysis". In: *NIPS*. 2004.

[Gol+17] N. Gold, M. G. Frasch, C. Herry, B. S. Richardson, and X. Wang. "A Doubly Stochastic Change Point Detection Algorithm for Noisy Biological Signals". en. In: *Front. Physiol.* (2017), p. 106088.

[Gol17] Y. Goldberg. *Neural Network Methods for Natural Language Processing*. Morgan & Claypool Publishers, 2017.

[Gol+19] Z. Goldfeld, E. van den Berg, K. H. Greenewald, I. Melnyk, N. Nguyen, B. Kingsbury, and Y. Polyanskiy. "Estimating Information Flow in Deep Neural Networks". In: *ICML*. 2019.

[Gol89] D. E. Goldberg. *Genetic Algorithms in Search, Optimization and Machine Learning*. 1st. Addison-Wesley Longman Publishing Co., Inc., 1989.

[Gom+17] A. N. Gomez, M. Ren, R. Urtasun, and R. B. Grosse. "The Reversible Residual Network: Backpropagation Without Storing Activations". In: *NIPS*. 2017.

[Gon+11] J. Gonzalez, Y. Low, A. Gretton, and C. Guestrin. "Parallel gibbs sampling: From colored fields to thin junction trees". In: *AISTATS*. 2011, pp. 324–332.

[Gon+14] B. Gong, W.-L. Chao, K. Grauman, and F. Sha. "Diverse sequential subset selection for supervised video summarization". In: *Advances in neural information processing systems* 27 (2014), pp. 2069–2077.

[Gon+20] P. J. Goncalves et al. "Training deep neural density estimators to identify mechanistic models of neural dynamics". In: *Elife* 9 (2020).

[Goo+14] I. J. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio. "Generative Adversarial Networks". In: *NIPS*. 2014.

[Goo16] I. Goodfellow. "NIPS 2016 Tutorial: Generative Adversarial Networks". In: *NIPS Tutorial*. 2016.

[Goo85] I. Good. "Weight of evidence: A brief survey". In: *Bayesian statistics* 2 (1985), pp. 249–270.

[Gor+14] A. D. Gordon, T. A. Henzinger, A. V. Nori, and S. K. Rajamani. "Probabilistic programming". In: *Intl. Conf. on Software Engineering*. 2014.

[Gor+19] J. Gordon, J. Bronskill, M. Bauer, S. Nowozin, and R. E. Turner. "Meta-Learning Probabilistic Inference For Prediction". In: *ICLR*. 2019.

[Gor93] N. Gordon. "Novel Approach to Nonlinear/Non-Gaussian Bayesian State Estimation". In: *IEE Proceedings (F)* 140.2 (1993), pp. 107–113.

[Gor95] G. J. Gordon. "Stable Function Approximation in Dynamic Programming". In: *ICML*. 1995, pp. 261–268.

[Gou+96] C. Gourieroux, M. Gourieroux, A. Monfort, and D. A. Monfort. *Simulation-based econometric methods*. Oxford university press, 1996.

[Goy+19] Y. Goyal, Z. Wu, J. Ernst, D. Batra, D. Parikh, and S. Lee. *Counterfactual Visual Explanations*. 2019. arXiv: 1904.07451 [cs.LG].

[Goy+22] S. Goyal, M. Sun, A. Raghunathan, and Z. Kolter. "Test-Time Adaptation via Conjugate Pseudo-labels". In: (July 2022). arXiv: 2207.09640 [cs.LG].

[GPS89] D. Greig, B. Porteous, and A. Seheult. "Exact maximum a posteriori estimation for binary images". In: *J. of Royal Stat. Soc. Series B* 51.2 (1989), pp. 271–279.

[GR06a] M Girolami and S Rogers. "Variational Bayesian Multinomial Probit Regression with Gaussian Process Priors". In: *Neural Comput.* 18.8 (2006), pp. 1790–1817.

[GR06b] M. Girolami and S. Rogers. "Variational Bayesian multinomial probit regression with Gaussian process priors". In: *Neural Comptuation* 18.8 (2006), pp. 1790 –1817.

[GR07a] T. Gneiting and A. E. Raftery. "Strictly Proper Scoring Rules, Prediction, and Estimation". In: *JASA* 102.477 (2007), pp. 359–378.

[GR07b] T. Gneiting and A. E. Raftery. "Strictly proper scoring rules, prediction, and estimation". In: *Journal of the American statistical Association* 102.477 (2007), pp. 359–378.

[Gra+10] T. Graepel, J. Quinonero-Candela, T. Borchert, and R. Herbrich. "Web-Scale Bayesian Click-Through Rate Prediction for Sponsored Search Advertising in Microsoft's Bing Search Engine". In: *ICML*. 2010.

[Gra11] A. Graves. "Practical Variational Inference for Neural Networks". In: *NIPS*. 2011.

[Gra+18] W. Grathwohl, R. T. Q. Chen, J. Bettencourt, I. Sutskever, and D. Duvenaud. "FFJORD: Free-form Continuous Dynamics for Scalable Reversible Generative Models". In: (2018). arXiv: 1810.01367 [cs.LG].

[Gra+20a] W. Grathwohl, J. Kelly, M. Hashemi, M. Norouzi, K. Swersky, and D. Duvenaud. "No MCMC for me: Amortized sampling for fast and stable training of energy-based models". In: *arXiv preprint arXiv:2010.04230* (2020).

[Gra+20b] W. Grathwohl, K.-C. Wang, J.-H. Jacobsen, D. Duvenaud, M. Norouzi, and K. Swersky. "Your classifier is secretly an energy based model and you should treat it like one". In: *ICLR*. 2020.

[Gra+20c] W. Grathwohl, K.-C. Wang, J.-H. Jacobsen, D. Duvenaud, and R. Zemel. "Cutting out the Middle-Man: Training and Evaluating Energy-Based Models without Sampling". In: *arXiv preprint arXiv:2002.05616* (2020).

[Gre03] P. Green. "Tutorial on trans-dimensional MCMC". In: *Highly Structured Stochastic Systems*. Ed. by P. Green, N. Hjort, and S. Richardson. OUP, 2003.

[Gre+12] A. Gretton, K. M. Borgwardt, M. J. Rasch, B. Schölkopf, and A. Smola. "A Kernel Two-Sample Test". In: *JMLR* 13.Mar (2012), pp. 723–773.

[Gre+14] K. Gregor, I. Danihelka, A. Mnih, C. Blundell, and D. Wierstra. "Deep AutoRegressive Networks". In: *ICML*. 2014.

[Gre20] F. Greenlee. *Transformer VAE*. 2020.

[Gre+22] P. L. Green, R. E. Moore, R. J. Jackson, J. Li, and S. Maskell. "Increasing the efficiency of Sequential Monte Carlo samplers through the use of approximately optimal L-kernels". In: *Mech. Syst. Signal Process.* 162 (2022).

[Gre98] P. Green. "Reversible Jump Markov Chain Monte Carlo computation and Bayesian model determination". In: *Biometrika* 82 (1998), pp. 711–732.

[Gri+04] T. Griffiths, M. Steyvers, D. Blei, and J. Tenenbaum. "Integrating Topics and Syntax". In: *NIPS*. 2004.

[Gri20] T. L. Griffiths. "Understanding Human Intelligence through Human Limitations". en. In: *Trends Cogn. Sci.* 24.11 (2020), pp. 873–883.

[Gri+20] J.-B. Grill, F. Strub, F. Altché, C. Tallec, P. H. Richemond, E. Buchatskaya, C. Doersch, B. A. Pires, Z. D. Guo, M. G. Azar, et al. "Bootstrap your own latent: A new approach to self-supervised learning". In: *arXiv preprint arXiv:2006.07733* (2020).

[GRS96] W. Gilks, S. Richardson, and D. Spiegelhalter. *Markov Chain Monte Carlo in Practice*. Chapman and Hall, 1996.

[GS04] T. Griffiths and M. Steyvers. "Finding scientific topics". In: *PNAS* 101 (2004), pp. 5228–5235.

[GS08] Y Guo and D Schuurmans. "Efficient global optimization for exponential family PCA and low-rank matrix factorization". In: *2008 46th Annual Allerton Conference on Communication, Control, and Computing*. 2008, pp. 1100–1107.

[GS13] A. Gelman and C. R. Shalizi. "Philosophy and the practice of Bayesian statistics". en. In: *Br. J. Math. Stat. Psychol.* 66.1 (Feb. 2013), pp. 8–38.

[GS15] R. B. Grosse and R. Salakhutdinov. "Scaling Up Natural Gradient by Sparsely Factorizing the Inverse Fisher Matrix". In: *ICML*. 2015.

[GS90] A. Gelfand and A. Smith. "Sampling-based approaches to calculating marginal densities". In: *JASA* 85 (1990), pp. 385–409.

[GS92] G. Grimmett and D. Stirzaker. *Probability and Random Processes*. Oxford, 1992.

[GSA14] M. A. Gelbart, J. Snoek, and R. P. Adams. "Bayesian Optimization with Unknown Constraints". In: *UAI*. 2014.

[GSD13] A. Guez, D. Silver, and P. Dayan. "Scalable and Efficient Bayes-Adaptive Reinforcement Learning Based on Monte-Carlo Tree Search". In: *JAIR* 48 (2013), pp. 841–883.

[GSJ19] A. Gretton, D. Sutherland, and W. Jitkrittum. *NIPS tutorial on interpretable comparison of distributions and models*. 2019.

[GSK18] S. Gidaris, P. Singh, and N. Komodakis. "Unsupervised Representation Learning by Predicting Image Rotations". In: *International Conference on Learning Representations*. 2018.

[GSM18] U. Garciarena, R. Santana, and A. Mendiburu. "Expanding Variational Autoencoders for Learning and Exploiting Latent Representations in Search Distributions". In: *Proc. of the Conf. on Genetic and Evolutionary Computation*. 2018, pp. 849–856.

[GSS15] I. J. Goodfellow, J. Shlens, and C. Szegedy. "Explaining and Harnessing Adversarial Examples". In: *ICLR*. 2015.

[GSZ21] P. Grünwald, T. Steinke, and L. Zakynthinou. "PAC-Bayes, MAC-Bayes and Conditional Mutual Information: Fast rate bounds that handle general VC classes". In: *COLT*. 2021.

[GT86] J. R. Gilbert and R. E. Tarjan. "The analysis of a nested dissection algorithm". In: *Numer. Math.* 50.4 (1986), pp. 377–404.

[Gue19] B. Guedj. "A primer on PAC-Bayesian learning". In: *arXiv preprint arXiv:1901.05353* (2019).

[Gul+17] I. Gulrajani, F. Ahmed, M. Arjovsky, V. Dumoulin, and A. C. Courville. "Improved training of wasserstein gans". In: *NIPS*. 2017, pp. 5767–5777.

[Gul+20] C. Gulcehre et al. *RL Unplugged: Benchmarks for Offline Reinforcement Learning*. arXiv:2006.13888. 2020.

[Gum54] E. J. Gumbel. *Statistical theory of extreme values and some practical applications;: A series of lectures (United States. National Bureau of Standards. Applied mathematics series)*. en. 1st edition. U.S. Govt. Print. Office, 1954.

[Guo09] Y. Guo. "Supervised exponential family principal component analysis via convex optimization". In: *NIPS*. 2009.

[Guo+17] C. Guo, G. Pleiss, Y. Sun, and K. Q. Weinberger. "On Calibration of Modern Neural Networks". In: *ICML*. 2017.

[Gup+16] M. Gupta, A. Cotter, J. Pfeifer, K. Voevodski, K. Canini, A. Mangylov, W. Moczydlowski, and A. van Esbroeck. "Monotonic Calibrated Interpolated Look-Up Tables". In: *Journal of Machine Learning Research* 17.109 (2016), pp. 1–47.

[Gur+18] S. Gururangan, S. Swayamdipta, O. Levy, R. Schwartz, S. R. Bowman, and N. A. Smith. "Annotation Artifacts in Natural Language Inference Data". In: *CoRR* abs/1803.02324 (2018). arXiv: 1803.02324.

[Gus01] M. Gustafsson. "A probabilistic derivation of the partial least-squares algorithm". In: *Journal of Chemical Information and Modeling* 41 (2001), pp. 288–294.

[Gut+14] M. U. Gutmann, R. Dutta, S. Kaski, and J. Corander. "Statistical inference of intractable generative models via classification". In: *arXiv preprint arXiv:1407.4981* (2014).

[Gut22] M. U. Gutmann. "Pen and Paper Exercises in Machine Learning". In: (June 2022). arXiv: 2206.13446 [cs.LG].

[GV17] S. Ghosal and A. van der Vaart. *Fundamentals of Nonparametric Bayesian Inference*. en. 1st ed. Cambridge University Press, 2017.

[GW08] A. Griewank and A. Walther. *Evaluating Derivatives: Principles and Techniques of Algorithmic Differentiation*. Second. Society for Industrial and Applied Mathematics, 2008.

[GW92] W. Gilks and P. Wild. "Adaptive rejection sampling for Gibbs sampling". In: *Applied Statistics* 41 (1992), pp. 337–348.

[GXG18] H. Ge, K. Xu, and Z. Ghahramani. "Turing: a language for flexible probabilistic inference". In: *AISTATS*. 2018, pp. 1682–1690.

[GZG19] S. K. S. Ghasemipour, R. S. Zemel, and S. Gu. "A Divergence Minimization Perspective on Imitation Learning Methods". In: *CORL*. 2019, pp. 1259–1277.

[HA21] R. J. Hyndman and G. Athanasopoulos. *Forecasting: Principles and Practice*. en. 3rd ed. Otexts, 2021.

[Haa+17] T. Haarnoja, H. Tang, P. Abbeel, and S. Levine. "Reinforcement learning with deep energy-based policies". In: *Proceedings of the 34th International Conference on Machine Learning-Volume 70*. 2017, pp. 1352–1361.

[Haa+18a] T. Haarnoja, A. Zhou, P. Abbeel, and S. Levine. "Soft Actor-Critic: Off-Policy Maximum Entropy Deep Reinforcement Learning with a Stochastic Actor". In: *International Conference on Machine Learning*. 2018, pp. 1861–1870.

[Haa+18b] T. Haarnoja, A. Zhou, P. Abbeel, and S. Levine. "Soft Actor-Critic: Off-Policy Maximum Entropy Deep Reinforcement Learning with a Stochastic Actor". In: *ICML*. 2018.

[Haa+18c] T. Haarnoja et al. "Soft Actor-Critic Algorithms and Applications". In: (2018). arXiv: 1812.05905 [cs.LG].

[HAB17] J. H. Huggins, R. P. Adams, and T. Broderick. "PASS-GLM: polynomial approximate sufficient statistics for scalable Bayesian GLM inference". In: *NIPS*. 2017.

[Had+20] R. Hadsell, D. Rao, A. A. Rusu, and R. Pascanu. "Embracing Change: Continual Learning in Deep Neural Networks". en. In: *Trends Cogn. Sci.* 24.12 (2020), pp. 1028–1040.

[HAE16] M. Huh, P. Agrawal, and A. A. Efros. "What makes ImageNet good for transfer learning?" In: *arXiv preprint arXiv:1608.08614* (2016).

[Haf18] D. Hafner. *Building Variational Auto-Encoders in TensorFlow*. Blog post. 2018.

[Haf+19] D. Hafner, T. Lillicrap, I. Fischer, R. Villegas, D. Ha, H. Lee, and J. Davidson. "Learning Latent Dynamics for Planning from Pixels". In: *ICML*. 2019.

[Haf+20] D. Hafner, T. Lillicrap, J. Ba, and M. Norouzi. "Dream to Control: Learning Behaviors by Latent Imagination". In: *ICLR*. 2020.

[Hag+17] M. Hagen, M. Potthast, M. Gohsen, A. Rathgeber, and B. Stein. "A Large-Scale Query Spelling Correction Corpus". In: *Proceedings of the 40th International ACM SIGIR Conference on Research and Development in Information Retrieval*. SIGIR '17. ACM, 2017, pp. 1261–1264.

[Háj08] A. Hájek. "Dutch Book Arguments". In: *The Oxford Handbook of Rational and Social Choice*. Ed. by P. Anand, P. Pattanaik, and C. Puppe. Oxford University Press, 2008.

[Haj88] B. Hajek. "Cooling Schedules for Optimal Annealing". In: *Math. Oper. Res.* 13.2 (1988), pp. 311–329.

[Häl+21] H. Hälvä, S. L. Corff, L. Leh'ericy, J. So, Y. Zhu, E. Gassiat, and A. Hyvärinen. "Disentangling Identifiable Features from Noisy Data with Structured Nonlinear ICA". In: *NeurIPS*. 2021.

[Ham90] J. Hamilton. "Analysis of time series subject to changes in regime". In: *J. Econometrics* 45 (1990), pp. 39–70.

[Han16] N. Hansen. "The CMA Evolution Strategy: A Tutorial". In: (2016). arXiv: 1604.00772 [cs.LG].

[Han+20] K. Han et al. "A Survey on Vision Transformer". In: (2020). arXiv: 2012.12556 [cs.CV].

[Han80] T. S. Han. "Multiple mutual informations and multiple interactions in frequency data". In: *Information and Control* 46.1 (1980), pp. 26–45.

[Har+17] J. Hartford, G. Lewis, K. Leyton-Brown, and M. Taddy. "Deep IV: A flexible approach for counterfactual prediction". In: *Proceedings of the 34th International Conference on Machine Learning-Volume 70*. JMLR. org. 2017, pp. 1414–1423.

[Har18] K. Hartnett. "To Build Truly Intelligent Machines, Teach Them Cause and Effect". In: *Quanta Magazine* (2018).

[Har+22] W. Harvey, S. Naderiparizi, V. Masrani, C. Weilbach, and F. Wood. "Flexible Diffusion Modeling of Long Videos". In: (May 2022). arXiv: 2205.11495 [cs.CV].

[Har90] A. C. Harvey. *Forecasting, Structural Time Series Models, and the Kalman Filter*. Cambridge Univerity Press, 1990.

[Has10] H. van Hasselt. "Double Q-learning". In: *NIPS*. Ed. by J. D. Lafferty, C. K. I. Williams, J Shawe-Taylor, R. S. Zemel, and A Culotta. Curran Associates, Inc., 2010, pp. 2613–2621.

[Has70] W. Hastings. "Monte Carlo Sampling Methods Using Markov Chains and Their Applications". In: *Biometrika* 57.1 (1970), pp. 97–109.

[Hau+10] J. R. Hauser, O. Toubia, T. Evgeniou, R. Befurt, and D. Dzyabura. "Disjunctions of conjunctions, cognitive simplicity, and consideration sets". In: *Journal of Marketing Research* 47.3 (2010), pp. 485–496.

[Hau+11] M. Hauschild, M. Pelikan, M. Hauschild, and M. Pelikan. "An introduction and survey of estimation of distribution algorithms". In: *Swarm and Evolutionary Computation*. 2011.

[Haw71] A. G. Hawkes. "Point spectra of some mutually exciting point processes". In: *Journal of the Royal Statistical Society: Series B (Methodological)* 33.3 (1971), pp. 438–443.

[HBB10] M. Hoffman, D. Blei, and F. Bach. "Online learning for latent Dirichlet allocation". In: *NIPS*. 2010.

[HBW19] E. Hoogeboom, R. van den Berg, and M. Welling. "Emerging Convolutions for Generative Normalizing Flows". In: *ICML*. 2019.

[HC93] G. Hinton and D. V. Camp. "Keeping Neural Networks Simple by Minimizing the Description Length of the Weights". In: *in Proc. of the 6th Ann. ACM Conf. on Computational Learning Theory*. ACM Press, 1993, pp. 5–13.

[HCG20] S. Huang, Y. Cao, and R. Grosse. "Evaluating Lossy Compression Rates of Deep Generative Models". In: *ICML*. 2020.

[HCL06] R. Hadsell, S. Chopra, and Y. LeCun. "Dimensionality Reduction by Learning an Invariant Mapping". In: *2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'06)* 2 (2006), pp. 1735–1742.

[HD19] D. Hendrycks and T. Dietterich. "Benchmarking Neural Network Robustness to Common Corruptions and Perturbations". In: *ICLR*. 2019.

[HDL17] D. Ha, A. M. Dai, and Q. V. Le. "HyperNetworks". In: *ICLR*. 2017.

[He+16a] K. He, X. Zhang, S. Ren, and J. Sun. "Deep Residual Learning for Image Recognition". In: *CVPR*. 2016.

[HE16a] J. Ho and S. Ermon. "Generative adversarial imitation learning". In: *Proceedings of the 30th International Conference on Neural Information Processing Systems*. 2016, pp. 4572–4580.

[He+16b] K. He, X. Zhang, S. Ren, and J. Sun. "Identity Mappings in Deep Residual Networks". In: *ECCV*. 2016.

[HE16b] J. Ho and S. Ermon. "Generative Adversarial Imitation Learning". In: *NIPS*. 2016, pp. 4565–4573.

[HE18] D. Ha and D. Eck. "A Neural Representation of Sketch Drawings". In: *ICLR*. 2018.

[He+19] J. He, D. Spokoyny, G. Neubig, and T. Berg-Kirkpatrick. "Lagging Inference Networks and Posterior Collapse in Variational Autoencoders". In: *ICLR*. 2019.

[He+20] K. He, H. Fan, Y. Wu, S. Xie, and R. Girshick. "Momentum contrast for unsupervised visual representation learning". In: *CVPR*. 2020, pp. 9729–9738.

[He+21] K. He, X. Chen, S. Xie, Y. Li, P. Dollár, and R. Girshick. *Masked Autoencoders Are Scalable Vision Learners*. 2021. arXiv: 2111.06377 [cs.CV].

[He+22] K. He, X. Chen, S. Xie, Y. Li, P. Dollár, and R. Girshick. "Masked autoencoders are scalable vision learners". In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2022, pp. 16000–16009.

[Heg06] P. Heggernes. "Minimal triangulations of graphs: A survey". In: *Discrete Math.* 306.3 (2006), pp. 297–317.

[Hel17] J. Helske. "KFAS: Exponential Family State Space Models in R". In: *J. Stat. Softw.* (2017).

[Hen+15] J. Hensman, A. Matthews, M. Filippone, and Z. Ghahramani. "MCMC for Variationally Sparse Gaussian Processes". In: *NIPS*. 2015, pp. 1648–1656.

[Hen+16] L. A. Hendricks, Z. Akata, M. Rohrbach, J. Donahue, B. Schiele, and T. Darrell. "Generating visual explanations". In: *European conference on computer vision*. Springer. 2016, pp. 3–19.

[Hen+18] G. E. Henter, J. Lorenzo-Trueba, X. Wang, and J. Yamagishi. "Deep Encoder-Decoder Models for Unsupervised Learning of Controllable Speech Synthesis". In: (2018). arXiv: 1807.11470 [eess.AS].

[Hen+19a] O. J. Henaff, A. Razavi, C. Doersch, S. M. Ali Eslami, and A. van den Oord. "Data-Efficient Image Recognition with Contrastive Predictive Coding". In: *arXiv [cs.CV]* (2019).

[Hen+19b] D. Hendrycks, S. Basart, M. Mazeika, A. Zou, J. Kwon, M. Mostajabi, J. Steinhardt, and D. Song. "Scaling Out-of-Distribution Detection for Real-World Settings". In: (2019). arXiv: 1911.11132 [cs.CV].

[Hen+20] D. Hendrycks*, N. Mu*, E. D. Cubuk, B. Zoph, J. Gilmer, and B. Lakshminarayanan. "AugMix: A Simple Data Processing Method to Improve Robustness and Uncertainty". In: *ICLR*. 2020.

[Hen+21] C. Henning, M. R. Cervera, F. D'Angelo, J. von Oswald, R. Traber, B. Ehret, S. Kobayashi, B. F. Grewe, and J. Sacramento. "Posterior Meta-Replay for Continual Learning". In: *NIPS*. 2021.

[Hes00] T. Heskes. "On "Natural" Learning and Pruning in Multilayered Perceptrons". In: *Neural Comput.* 12.4 (2000), pp. 881–901.

[Hes+18] M. Hessel, J. Modayil, H. van Hasselt, T. Schaul, G. Ostrovski, W. Dabney, D. Horgan, B. Piot, M. Azar, and D. Silver. "Rainbow: Combining Improvements in Deep Reinforcement Learning". In: *AAAI*. 2018.

[Heu+17a] M. Heusel, H. Ramsauer, T. Unterthiner, B. Nessler, and S. Hochreiter. "GANs Trained by a Two Time-Scale Update Rule Converge to a Local Nash Equilibrium". In: *NIPS*. 2017.

[Heu+17b] M. Heusel, H. Ramsauer, T. Unterthiner, B. Nessler, and S. Hochreiter. "Gans trained by a two time-scale update rule converge to a local nash equilibrium". In: *Advances in neural information processing systems*. 2017, pp. 6626–6637.

[HFL13] J. Hensman, N. Fusi, and N. D. Lawrence. "Gaussian Processes for Big Data". In: *UAI*. 2013.

[HFM17] D. W. Hogg and D. Foreman-Mackey. "Data analysis recipes: Using Markov Chain Monte Carlo". In: (2017). arXiv: 1710.06068 [astro-ph.IM].

[HG12] D. I. Hastie and P. J. Green. "Model Choice using Reversible Jump Markov Chain Monte Carlo". In: *Statistica Neerlandica* 66 (2012), pp. 309–338.

[HG14] M. D. Hoffman and A. Gelman. "The No-U-Turn Sampler: Adaptively Setting Path Lengths in Hamiltonian Monte Carlo". In: *JMLR* 15 (2014), pp. 1593–1623.

[HG16] D. Hendrycks and K. Gimpel. "Gaussian Error Linear Units (GELUs)". In: *arXiv [cs.LG]* (2016).

[HGMG18] J. Hron, A. G. de G. Matthews, and Z. Ghahramani. "Variational Bayesian dropout: pitfalls and fixes". In: *ICML*. 2018.

[HGS16] H. van Hasselt, A. Guez, and D. Silver. "Deep Reinforcement Learning with Double Q-Learning". In: *AAAI*. AAAI'16. AAAI Press, 2016, pp. 2094–2100.

[HH06] C. Holmes and L. Held. "Bayesian auxiliary variable models for binary and multinomial regression". In: *Bayesian Analysis* 1.1 (2006), pp. 145–168.

[HHC12] J. Hu, P. Hu, and H. S. Chang. "A Stochastic Approximation Framework for a Class of Randomized Optimization Algorithms". In: *IEEE Trans. Automatic Control* 57.1 (2012).

[HHH09] A. Hyvarinen, J. Hurri, and P. Hoyer. *Natural Image Statistics: a probabilistic approach to early computational vision*. Springer, 2009.

[HHK19] M. Haußmann, F. A. Hamprecht, and M. Kandemir. "Sampling-Free Variational Inference of Bayesian Neural Networks by Variance Backpropagation". In: *UAI*. 2019.

[HHLB11] F. Hutter, H. H. Hoos, and K. Leyton-Brown. "Sequential Model-Based Optimization for General Algorithm Configuration". In: *Intl. Conf. on Learning and Intelligent Optimization (LION)*. 2011, pp. 507–523.

[HHLMF18] M. Havasi, J. M. Hernández-Lobato, and J. J. Murillo-Fuentes. "Inference in Deep Gaussian Processes using Stochastic Gradient Hamiltonian Monte Carlo". In: (2018). arXiv: 1806.05490 [stat.ML].

[Hig+17a] I. Higgins, L. Matthey, A. Pal, C. Burgess, X. Glorot, M. Botvinick, S. Mohamed, and A. Lerchner. "beta-VAE: Learning Basic Visual Concepts with a Constrained Variational Framework". In: *ICLR*. 2017.

[Hig+17b] I. Higgins, L. Matthey, A. Pal, C. P. Burgess, X. Glorot, M. M. Botvinick, S. Mohamed, and A. Lerchner. "beta-VAE: Learning Basic Visual Concepts with a Constrained Variational Framework". In: *ICLR*. 2017.

[Hin02] G. E. Hinton. "Training products of experts by minimizing contrastive divergence". en. In: *Neural Computation* 14.8 (2002), pp. 1771–1800.

[Hin10] G. Hinton. *A Practical Guide to Training Restricted Boltzmann Machines*. Tech. rep. U. Toronto, 2010.

[Hin14] G. Hinton. *Lecture 6e on neural networks (RMSprop: Divide the gradient by a running average of its recent magnitude)*. 2014.

[Hin+95] G. E. Hinton, P Dayan, B. J. Frey, and R. M. Neal. "The "wake-sleep" algorithm for unsupervised neural networks". en. In: *Science* 268.5214 (1995), pp. 1158–1161.

[HIY19] K. Hayashi, M. Imaizumi, and Y. Yoshida. "On Random Subsampling of Gaussian Process Regression: A Graphon-Based Analysis". In: (2019). arXiv: 1901.09541 [stat.ML].

[HJ12] T. Hazan and T. Jaakkola. "On the Partition Function and Random Maximum A-Posteriori Perturbations". In: *ICML*. June 2012.

[HJ20] J. Huang and N. Jiang. "From Importance Sampling to Doubly Robust Policy Gradient". In: *ICML*. 2020.

[HJA20] J. Ho, A. Jain, and P. Abbeel. "Denoising Diffusion Probabilistic Models". In: *NIPS*. 2020.

[Hje+18] R. D. Hjelm, A. Fedorov, S. Lavoie-Marchildon, K. Grewal, P. Bachman, A. Trischler, and Y. Bengio. "Learning deep representations by mutual information estimation and maximization". In: *International Conference on Learning Representations*. 2018.

[Hjo+10] N. Hjort, C. Holmes, P. Muller, and S. Walker, eds. *Bayesian Nonparametrics*. Cambridge, 2010.

[HJT18] M. D. Hoffman, M. J. Johnson, and D. Tran. "Autoconj: Recognizing and Exploiting Conjugacy Without a Domain-Specific Language". In: *NIPS*. 2018.

[HKO22] P. Hennig, H. Kersting, and M. Osborne. *Probabilistic Numerics: Computation as Machine Learning*. 2022.

[HKP91] J. Hertz, A. Krogh, and R. G. Palmer. *An Introduction to the Theory of Neural Comptuation*. Addison-Wesley, 1991.

[HKZ12] D. Hsu, S. Kakade, and T. Zhang. "A spectral algorithm for learning hidden Markov models". In: *J. of Computer and System Sciences* 78.5 (2012), pp. 1460–1480.

[HL04] D. R. Hunter and K. Lange. "A Tutorial on MM Algorithms". In: *The American Statistician* 58 (2004), pp. 30–37.

[HL+16a] J. Hernandez-Lobato, Y. Li, M. Rowland, T. Bui, D. Hernandez-Lobato, and R. Turner. "Black-Box Alpha Divergence Minimization". en. In: *ICML*. 2016, pp. 1511–1520.

[HL+16b] M. Hernandez-Lobato, M. A. Gelbart, R. P. Adams, M. W. Hoffman, and Z. Ghahramani. "A General Framework for Constrained Bayesian Optimization using Information-based Search". In: *JMLR* (2016).

[HL20] X. Hu and J. Lei. "A Distribution-Free Test of Covariate Shift Using Conformal Prediction". In: (2020). arXiv: 2010.07147 [stat.ME].

[HLA15a] J. M. Hernández-Lobato and R. P. Adams. "Probabilistic Backpropagation for Scalable Learning of Bayesian Neural Networks". In: *ICML*. 2015.

[HLA15b] J. M. Hernández-Lobato and R. P. Adams. "Probabilistic Backpropagation for Scalable Learning of Bayesian Neural Networks". In: *ICML*. 2015.

[HLA19] J. Ho, E. Lohn, and P. Abbeel. "Compression with Flows via Local Bits-Back Coding". In: *NeurIPS*. 2019.

[HLC19] Y.-P. Hsieh, C. Liu, and V. Cevher. "Finding mixed nash equilibria of generative adversarial networks". In: *International Conference on Machine Learning*. 2019, pp. 2810–2819.

[HLHG14] J. Hernandez-Lobato, M. W. Hoffman, and Z. Ghahramani. "Predictive entropy search for efficient global optimization of black-box functions". In: *NIPS*. 2014.

[HLR16] K. Hofmann, L. Li, and F. Radlinski. "Online Evaluation for Information Retrieval". In: *Foundations and Trends in Information Retrieval* 10.1 (2016), pp. 1–117.

[HLRW14] J. R. Hershey, J. Le Roux, and F. Weninger. "Deep Unfolding: Model-Based Inspiration of Novel Deep Architectures". In: (2014). arXiv: 1409.2574 [cs.LG].

[HLS03] R. Herbrich, N. D. Lawrence, and M. Seeger. "Fast Sparse Gaussian Process Methods: The Informative Vector Machine". In: *NIPS*. MIT Press, 2003, pp. 625–632.

[HM81] R. Howard and J. Matheson. "Influence diagrams". In: *Readings on the Principles and Applications of Decision Analysis, volume II*. Ed. by R. Howard and J. Matheson. Strategic Decisions Group, 1981.

[HMD18] J. C. Higuera, D Meger, and G Dudek. "Synthesizing Neural Network Controllers with Probabilistic Model-Based Reinforcement Learning". In: *IROS*. 2018, pp. 2538–2544.

[HMD19] D. Hendrycks, M. Mazeika, and T. Dietterich. "Deep Anomaly Detection with Outlier Exposure". In: *ICLR*. 2019.

[HMK04] D. Heckerman, C. Meek, and D. Koller. *Probabilistic Models for Relational Data*. Tech. rep. MSR-TR-2004-30. Microsoft Research, 2004.

[HNBK18] J. He, G. Neubig, and T. Berg-Kirkpatrick. "Unsupervised Learning of Syntactic Structure with Invertible Neural Projections". In: *EMNLP*. 2018.

[HNP09] A. Halevy, P. Norvig, and F. Pereira. "The unreasonable effectiveness of data". In: *IEEE Intelligent Systems* 24.2 (2009), pp. 8–12.

[HO00] A. Hyvarinen and E. Oja. "Independent component analysis: algorithms and applications". In: *Neural Networks* 13 (2000), pp. 411–430.

[Ho+21] J. Ho, C. Saharia, W. Chan, D. J. Fleet, M. Norouzi, and T. Salimans. "Cascaded Diffusion Models for High Fidelity Image Generation". In: (May 2021). arXiv: 2106.15282 [cs.CV].

[Ho+22] J. Ho, T. Salimans, A. Gritsenko, W. Chan, M. Norouzi, and D. J. Fleet. "Video Diffusion Models". In: (Apr. 2022). arXiv: 2204.03458 [cs.CV].

[HO48] C. G. Hempel and P. Oppenheim. "Studies in the Logic of Explanation". In: *Philosophy of science* 15.2 (1948), pp. 135–175.

[Hob69] A. Hobson. "A new theorem of information theory". In: *Journal of Statistical Physics* 1.3 (1969), pp. 383–391.

[Hoe12] J. M. ver Hoef. "Who invented the delta method?" In: *The American Statistician* 66.2 (2012), pp. 124–127.

[Hoe+21] T. Hoefler, D. Alistarh, T. Ben-Nun, N. Dryden, and A. Peste. "Sparsity in Deep Learning: Pruning and growth for efficient inference and training in neural networks". In: (2021). arXiv: 2102.00554 [cs.LG].

[Hof09] P. D. Hoff. *A First Course in Bayesian Statistical Methods*. Springer, 2009.

[Hof+13] M. D. Hoffman, D. M. Blei, C. Wang, and J. Paisley. "Stochastic Variational Inference". In: *JMLR* 14 (2013), pp. 1303–1347.

[Hof17] M. D. Hoffman. "Learning Deep Latent Gaussian Models with Markov Chain Monte Carlo". In: *ICML*. 2017, pp. 1510–1519.

[Hof+18] J. Hoffman, E. Tzeng, T. Park, J.-Y. Zhu, P. Isola, K. Saenko, A. Efros, and T. Darrell. "Cycada: Cycle-consistent adversarial domain adaptation". In: *International conference on machine learning*. PMLR. 2018, pp. 1989–1998.

[Hof+19] M. Hoffman, P. Sountsov, J. V. Dillon, I. Langmore, D. Tran, and S. Vasudevan. "NeuTra-lizing Bad Geometry in Hamiltonian Monte Carlo Using Neural Transport". In: (2019). arXiv: 1903.03704 [stat.CO].

[Hof99] T. Hofmann. "Probabilistic latent semantic indexing". In: *Research and Development in Information Retrieval* (1999), pp. 50–57.

[Hoh+20] F. Hohman, M. Conlen, J. Heer, and D. H. P. Chau. "Communicating with interactive articles". In: *Distill* 5.9 (2020), e28.

[Hol86] P. W. Holland. "Statistics and Causal Inference". In: *JASA* 81.396 (1986), pp. 945–960.

[Hol92] J. H. Holland. *Adaptation in Natural and Artificial Systems*. https://mitpress.mit.edu/books/adaptation-natural-and-artificial-systems. Accessed: 2017-11-26. 1992.

[Hon+10] A. Honkela, T. Raiko, M. Kuusela, M. Tornio, and J. Karhunen. "Approximate Riemannian Conjugate Gradient Learning for Fixed-Form Variational Bayes". In: *JMLR* 11.Nov (2010), pp. 3235–3268.

[Hop82] J. J. Hopfield. "Neural networks and physical systems with emergent collective computational abilities". In: *PNAS* 79.8 (1982), 2554–2558.

[Hor+05] E. Horvitz, J. Apacible, R. Sarin, and L. Liao. "Prediction, Expectation, and Surprise: Methods, Designs, and Study of a Deployed Traffic Forecasting Service". In: *UAI*. 2005.

[Hor61] P Horst. "Generalized canonical correlations and their applications to experimental data". en. In: *J. Clin. Psychol.* 17 (1961), pp. 331–347.

[Hos+20a] T. Hospedales, A. Antoniou, P. Micaelli, and A. Storkey. "Meta-Learning in Neural Networks: A Survey". In: (2020). arXiv: 2004.05439 [cs.LG].

[Hos+20b] R. Hostettler, F. Tronarp, Á. F. García-Fernández, and S. Särkkä. "Importance Densities for Particle Filtering Using Iterated Conditional Expectations". In: *IEEE Signal Process. Lett.* 27 (2020), pp. 211–215.

[HOT06a] G. Hinton, S. Osindero, and Y. Teh. "A fast learning algorithm for deep belief nets". In: *Neural Computation* 18 (2006), pp. 1527–1554.

[HOT06b] G. E. Hinton, S. Osindero, and Y. W. Teh. "A Fast Learning Algorithm for Deep Belief Nets". In: *Neural Computation* 18 (2006), pp. 1527–1554.

[Hot36] H. Hotelling. "Relations Between Two Sets of Variates". In: *Biometrika* 28.3/4 (1936), pp. 321–377.

[Hou+12] N. Houlsby, F. Huszar, Z. Ghahramani, and J. M. Hernández-lobato. "Collaborative Gaussian Processes for Preference Learning". In: *NIPS*. 2012, pp. 2096–2104.

[Hou+19] N. Houlsby, A. Giurgiu, S. Jastrzebski, B. Morrone, Q. De Laroussilhe, A. Gesmundo, M. Attariyan, and S. Gelly. "Parameter-efficient transfer learning for NLP". In: *International Conference on Machine Learning*. PMLR. 2019, pp. 2790–2799.

[HOW11] P. Hall, J. T. Ormerod, and M. P. Wand. "Theory of Gaussian Variational Approximation for a Generalised Linear Mixed Model". In: *Statistica Sinica* 21 (2011), pp. 269–389.

[HP10] J. Hacker and P. Pierson. *Winner-Take-All Politics: How Washington Made the Rich Richer — and Turned Its Back on the Middle Class*. Simon & Schuster, 2010.

[HPR19] C. Herzog né Hoffmann, E. Petersen, and P. Rostalski. "Iterative Approximate Nonlinear Inference via Gaussian Message Passing on Factor Graphs". In: *IEEE Control Systems Letters* 3.4 (2019), pp. 978–983.

[HR17] C. Hoffmann and P. Rostalski. "Linear Optimal Control on Factor Graphs — A Message Passing Perspective". In: *Intl. Federation of Automatic Control* 50.1 (2017), pp. 6314–6319.

[HR20a] M. Hernan and J. Robins. *Causal Inference: What If*. CRC Press, 2020.

[HR20b] M. Hernán and J. Robins. *Causal Inference: What If*. Boca Raton: Chapman & Hall/CRC., 2020.

[HS05] H. Hoos and T. Stutzle. *Stochastic local search: Foundations and applications*. Morgan Kauffman, 2005.

[HS06a] G. Hinton and R. Salakhutdinov. "Reducing the dimensionality of data with neural networks". In: *Science* 313.5786 (2006), pp. 504–507.

[HS06b] G. E. Hinton and R. R. Salakhutdinov. "Reducing the dimensionality of data with neural networks". In: *science* 313.5786 (2006), pp. 504–507.

[HS09] M. Heaton and J. Scott. *Bayesian computation and the linear model*. Tech. rep. Duke, 2009.

[HS12] P. Hennig and C. Schuler. "Entropy search for information-efficient global optimization". In: *JMLR* 13 (2012), pp. 1809–1837.

[HS13] J. Y. Hsu and D. S. Small. "Calibrating Sensitivity Analyses to Observed Covariates in Observational Studies". In: *Biometrics* 69.4 (2013), pp. 803–811. eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/biom.12101.

[HS18] D. Ha and J. Schmidhuber. "World Models". In: *NIPS*. 2018.

[HS21] J. Ho and T. Salimans. "Classifier-Free Diffusion Guidance". In: *NIPS Workshop on Deep Generative Models and Downstream Applications*. 2021.

[HS88] D. S. Hochbaum and D. B. Shmoys. "A polynomial approximation scheme for scheduling on uniform processors: Using the dual approximation approach". In: *SICOMP*. 1988.

[HS97] S Hochreiter and J Schmidhuber. "Flat minima". en. In: *Neural Comput.* 9.1 (1997), pp. 1–42.

[HSDK12] C. Hillar, J. Sohl-Dickstein, and K. Koepsell. *Efficient and Optimal Binary Hopfield Associative Memory Storage Using Minimum Probability Flow*. Tech. rep. 2012. arXiv: 1204.2916.

[HSG06] J. D. Hol, T. B. Schon, and F. Gustafsson. "On Resampling Algorithms for Particle Filters". In: *IEEE Nonlinear Statistical Signal Processing Workshop*. 2006, pp. 79–82.

[HSGF21] S. Hassan, S. Sarkka, and A. F. Garcia-Fernandez. "Temporal Parallelization of Inference in Hidden Markov Models". In: *IEEE Trans. Signal Processing* 69 (2021), pp. 4875–4887.

[Hsu+18] Y.-C. Hsu, Y.-C. Liu, A. Ramasamy, and Z. Kira. "Re-evaluating Continual Learning Scenarios: A Categorization and Case for Strong Baselines". In: *NIPS Continual Learning Workshop*. 2018.

[HT01] G. E. Hinton and Y. Teh. "Discovering multiple constraints that are frequently approximately satisfied". In: *UAI*. 2001.

[HT09] H. Hoefling and R. Tibshirani. "Estimation of Sparse Binary Pairwise Markov Networks using Pseudo-likelihoods". In: *JMLR* 10 (2009).

[HT15] J. H. Huggins and J. B. Tenenbaum. "Risk and regret of hierarchical Bayesian learners". In: *ICML*. 2015.

[HT17] T. J. Hastie and R. J. Tibshirani. *Generalized additive models*. Routledge, 2017.

[HTF09] T. Hastie, R. Tibshirani, and J. Friedman. *The Elements of Statistical Learning*. 2nd edition. Springer, 2009.

[HTW15] T. Hastie, R. Tibshirani, and M. Wainwright. *Statistical Learning with Sparsity: The Lasso and Generalizations*. CRC Press, 2015.

[Hu+00] M. Hu, C. Ingram, M.Sirski, C. Pal, S. Swamy, and C. Patten. *A Hierarchical HMM Implementation for Vertebrate Gene Splice Site Prediction*. Tech. rep. Dept. Computer Science, Univ. Waterloo, 2000.

[Hu+12] J. Hu, Y. Wang, E. Zhou, M. C. Fu, and S. I. Marcus. "A Survey of Some Model-Based Methods for Global Optimization". en. In: *Optimization, Control, and Applications of Stochastic Systems*. Systems & Control: Foundations & Applications. Birkhäuser, Boston, 2012, pp. 157–179.

[Hu+17] W. Hu, C. J. Li, L. Li, and J.-G. Liu. "On the diffusion approximation of nonconvex stochastic gradient descent". In: (2017). arXiv: 1705.07562 [stat.ML].

[Hu+18] W. Hu, G. Niu, I. Sato, and M. Sugiyama. "Does Distributionally Robust Supervised Learning Give Robust Classifiers?" In: *ICML*. 2018.

[Hua+17a] G. Huang, Y. Li, G. Pleiss, Z. Liu, J. Hopcroft, and K. Weinberger. "Snapshot ensembles: train 1, get M for free". In: *ICLR*. 2017.

[Hua+17b] Y Huang, Y Zhang, N Li, Z Wu, and J. A. Chambers. "A Novel Robust Student's t-Based Kalman Filter". In: *IEEE Trans. Aerosp. Electron. Syst.* 53.3 (2017), pp. 1545–1554.

[Hua+18a] C.-Z. A. Huang, A. Vaswani, J. Uszkoreit, N. Shazeer, I. Simon, C. Hawthorne, A. M. Dai, M. D. Hoffman, M. Dinculescu, and D. Eck. "Music Transformer". In: (2018). arXiv: 1809.04281 [cs.LG].

[Hua+18b] C.-W. Huang, D. Krueger, A. Lacoste, and A. Courville. "Neural Autoregressive Flows". In: *ICML*. 2018.

[Hua+19] Y. Huang, Y. Zhang, Y. Zhao, and J. A. Chambers. "A Novel Robust Gaussian–Student's t Mixture Distribution Based Kalman Filter". In: *IEEE Trans. Signal Process.* 67.13 (2019), pp. 3606–3620.

[Hug+19] J. H. Huggins, T. Campbell, M. Kasprzak, and T. Broderick. "Scalable Gaussian Process Inference with Finite-data Mean and Variance Guarantees". In: *AISTATS*. 2019.

[Hug+20] J. Huggins, M. Kasprzak, T. Campbell, and T. Broderick. "Validated Variational Inference via

Practical Posterior Error Bounds". In: *AISTATS*. Ed. by S. Chiappa and R. Calandra. Vol. 108. Proceedings of Machine Learning Research. PMLR, 2020, pp. 1792–1802.

[Hus17a] F. Huszár. *Is Maximum Likelihood Useful for Representation Learning?* 2017.

[Hus17b] F. Huszár. "Variational inference using implicit distributions". In: *arXiv preprint arXiv:1702.08235* (2017).

[Hut89] M. F. Hutchinson. "A stochastic estimator of the trace of the influence matrix for Laplacian smoothing splines". In: *Communications in Statistics-Simulation and Computation* 18.3 (1989), pp. 1059–1076.

[HVD14] G. Hinton, O. Vinyals, and J. Dean. "Distilling the Knowledge in a Neural Network". In: *NIPS DL workshop*. 2014.

[HW13] A. Huang and M. P. Wand. "Simple Marginally Noninformative Prior Distributions for Covariance Matrices". en. In: *Bayesian Analysis* 8.2 (2013), pp. 439–452.

[HXW17] H. He, B. Xin, and D. Wipf. "From Bayesian Sparsity to Gated Recurrent Nets". In: *NIPS*. 2017.

[HY01] M. Hansen and B. Yu. "Model selection and the principle of minimum description length". In: *JASA* (2001).

[Hyv05] A. Hyvärinen. "Estimation of non-normalized statistical models by score matching". In: *JMLR* 6.Apr (2005), pp. 695–709.

[Hyv07a] A. Hyvarinen. "Connections between score matching, contrastive divergence, and pseudolikelihood for continuous-valued variables". In: *IEEE Transactions on neural networks* 18.5 (2007), pp. 1529–1531.

[Hyv07b] A. Hyvärinen. "Some extensions of score matching". In: *Computational statistics & data analysis* 51.5 (2007), pp. 2499–2512.

[IB12] R. Iyer and J. Bilmes. "Algorithms for Approximate Minimization of the Difference Between Submodular Functions, with Applications". In: *Uncertainty in Artificial Intelligence (UAI)*. Catalina Island, USA: AUAI, 2012.

[IB13] R. Iyer and J. Bilmes. "Submodular Optimization with Submodular Cover and Submodular Knapsack Constraints". In: *Neural Information Processing Society (NeurIPS, formerly NIPS)*. Lake Tahoe, CA, 2013.

[IB15] R. K. Iyer and J. A. Bilmes. "Polyhedral aspects of Submodularity, Convexity and Concavity". In: *Arxiv, CoRR* abs/1506.07329 (2015).

[IB98] M. Isard and A. Blake. "CONDENSATION - conditional density propagation for visual tracking". In: *Intl. J. of Computer Vision* 29.1 (1998), pp. 5–18.

[IBK06] E. L. Ionides, C Bretó, and A. A. King. "Inference for nonlinear dynamical systems". en. In: *PNAS* 103.49 (2006), pp. 18438–18443.

[IFF00] S. Iwata, L. Fleischer, and S. Fujishige. "A combinatorial strongly polynomial algorithm for minimizing submodular functions". In: *Journal of the ACM* (2000).

[IFW05] A. T. Ihler, J. W. Fischer III, and A. S. Willsky. "Loopy Belief Propagation: Convergence and Effects of Message Errors". In: *JMLR* 6 (2005), pp. 905–936.

[IJ01] H. Ishwaran and L. F. James. "Gibbs sampling methods for stick-breaking priors". In: *Journal of the American Statistical Association* 96.453 (2001), pp. 161–173.

[IJB13] R. Iyer, S. Jegelka, and J. Bilmes. "Curvature and Optimal Algorithms for Learning and Minimizing Submodular Functions". In: *NIPS*. Lake Tahoe, CA, 2013.

[IKB21] A. Immer, M. Korzepa, and M. Bauer. "Improving predictions of Bayesian neural nets via local linearization". In: *AISTATS*. Ed. by A. Banerjee and K. Fukumizu. Vol. 130. Proceedings of Machine Learning Research. PMLR, 2021, pp. 703–711.

[IM17] J. Ingraham and D. Marks. "Bayesian Sparsity for Intractable Undirected Models". In: *ICML*. 2017.

[Imb03] G. Imbens. "Sensitivity to Exogeneity Assumptions in Program Evaluation". In: *The American Economic Review* (2003).

[Imb19] G. W. Imbens. "Potential Outcome and Directed Acyclic Graph Approaches to Causality: Relevance for Empirical Practice in Economics". In: (2019). arXiv: 1907.07271 [stat.ME].

[Imm+21] A. Immer, M. Bauer, V. Fortuin, G. Rätsch, and M. E. Khan. "Scalable Marginal Likelihood Estimation for Model Selection in Deep Learning". In: *ICML*. 2021.

[IN09] S. Iwata and K. Nagano. "Submodular function minimization under covering constraints". In: *Proceedings of the 50th Annual IEEE Symposium on Foundations of Computer Science (FOCS)*. 2009, pp. 671–680.

[Ing20] M. Ingram. *Deterministic ADVI in JAX (blog post)*. https://martiningram.github.io/deterministic-advi/. 2020.

[INK18] P. Izmailov, A. Novikov, and D. Kropotov. "Scalable Gaussian Processes with Billions of Inducing Inputs via Tensor Train Decomposition". In: *ICML*. 2018.

[Inn20] M. Innes. "Sense & Sensitivities: The Path to General-Purpose Algorithmic Differentiation". In: *Proceedings of Machine Learning and Systems*. Ed. by I. Dhillon, D. Papailiopoulos, and V. Sze. Vol. 2. 2020, pp. 58–69.

[IO09] S. Iwata and J. B. Orlin. "A simple combinatorial algorithm for submodular function minimization". In: *Proceedings of the twentieth annual ACM-SIAM symposium on Discrete algorithms*. SIAM. 2009, pp. 1230–1237.

[IR00] D. R. Insua and F. Ruggeri. *Robust Bayesian Analysis*. Springer, 2000.

[IR15] G. Imbens and D. Rubin. *Causal Inference in Statistics, Social and Biomedical Sciences: An Introduction*. Cambridge University Press, 2015.

[IS15] S. Ioffe and C. Szegedy. "Batch Normalization: Accelerating Deep Network Training by Reducing Internal Covariate Shift". In: *ICML*. 2015, pp. 448–456.

[Isa03] M. Isard. "PAMPAS: Real-Valued Graphical Models for Computer Vision". In: *CVPR*. Vol. 1. 2003, p. 613.

[Isl+19] R. Islam, R. Seraj, S. Y. Arnob, and D. Precup. "Doubly Robust Off-Policy Actor-Critic Algorithms for Reinforcement Learning". In: *NeurIPS Workshop on Safety and Robustness in Decision Making*. 2019.

[Iso+17] P. Isola, J.-Y. Zhu, T. Zhou, and A. A. Efros. "Image-to-Image Translation with Conditional Adversarial Networks". In: *CVPR*. 2017.

[IX00] K Ito and K Xiong. "Gaussian filters for nonlinear filtering problems". In: *IEEE Trans. Automat. Contr.* 45.5 (2000), pp. 910–927.

[Iye+21] R. Iyer, N. Khargonkar, J. Bilmes, and H. Asnani. "Generalized Submodular Information Measures: Theoretical Properties, Examples, Optimization Algorithms, and Applications". In: *IEEE Transactions on Information Theory* (2021).

[Iza+15] H. Izadinia, B. C. Russell, A. Farhadi, M. D. Hoffman, and A. Hertzmann. "Deep classifiers from image tags in the wild". In: *Proceedings of the 2015 Workshop on Community-Organized Multimodal Mining: Opportunities for Novel Solutions*. 2015, pp. 13–18.

[Izm+18] P. Izmailov, D. Podoprikhin, T. Garipov, D. Vetrov, and A. G. Wilson. "Averaging Weights Leads to Wider Optima and Better Generalization". In: *UAI*. 2018.

[Izm+19] P. Izmailov, W. J. Maddox, P. Kirichenko, T. Garipov, D. Vetrov, and A. G. Wilson. "Subspace Inference for Bayesian Deep Learning". In: *UAI*. 2019.

[Izm+21a] P. Izmailov, P. Nicholson, S. Lotfi, and A. G. Wilson. "Dangers of Bayesian Model Averaging under Covariate Shift". In: *NIPS*. 2021.

[Izm+21b] P. Izmailov, S. Vikram, M. D. Hoffman, and A. G. Wilson. "What Are Bayesian Neural Network Posteriors Really Like?" In: *ICML*. 2021.

[Jaa01] T. Jaakkola. "Tutorial on variational approximation methods". In: *Advanced mean field methods*. Ed. by M. Opper and D. Saad. MIT Press, 2001.

[Jac+21] M. Jacobs, M. F. Pradier, T. H. McCoy, R. H. Perlis, F. Doshi-Velez, and K. Z. Gajos. "How machine-learning recommendations influence clinician treatment selections: the example of antidepressant selection". In: *Translational psychiatry* 11.1 (2021), pp. 1–9.

[Jad+17] M. Jaderberg, V. Mnih, W. M. Czarnecki, T. Schaul, J. Z. Leibo, D. Silver, and K. Kavukcuoglu. "Reinforcement Learning with Unsupervised Auxiliary Tasks". In: *ICLR*. 2017.

[Jak21] K. Jakkala. "Deep Gaussian Processes: A Survey". In: (2021). arXiv: 2106.12135 [cs.LG].

[Jan+17] P. A. Jang, A. Loeb, M. Davidow, and A. G. Wilson. "Scalable Levy Process Priors for Spectral Kernel Learning". In: *Advances in Neural Information Processing Systems*. 2017.

[Jan18] E. Jang. *Normalizing Flows Tutorial*. 2018.

[Jan+19] M. Janner, J. Fu, M. Zhang, and S. Levine. "When to Trust Your Model: Model-Based Policy Optimization". In: *NIPS*. 2019.

[Jas] Jason Antic and Jeremy Howard and Uri Manor. *Decrappification, DeOldification, and Super Resolution (Blog post)*.

[Jay03] E. T. Jaynes. *Probability theory: the logic of science*. Cambridge university press, 2003.

[Jay+20] S. M. Jayakumar, W. M. Czarnecki, J. Menick, J. Schwarz, J. Rae, S. Osindero, Y. W. Teh, T. Harley, and R. Pascanu. "Multiplicative Interactions and Where to Find Them". In: *ICLR*. 2020.

[JB03] A. Jakulin and I. Bratko. "Analyzing Attribute Dependencies". In: *Proc. 7th European Conf. on Principles and Practice of Knowledge Discovery in Databases*. 2003.

[JB16] S. Jegelka and J. Bilmes. "Graph cuts with interacting edge weights: examples, approximations, and algorithms". In: *Mathematical Programming* (2016), pp. 1–42.

[JB65] C. Jacobi and C. W. Borchardt. "De investigando ordine systematis aequationum differentialium vulgarium cujuscunque." In: *Journal für die reine und angewandte Mathematik* 1865.64 (1865), pp. 297–320.

[JBB09] D. Jian, A. Barthels, and M. Beetz. "Adaptive Markov logic networks: Learning statistical relational models with dynamic parameters". In: *9th European Conf. on AI*. 2009, 937–942.

[Jef04] R. Jeffrey. *Subjective Probability: The Real Thing*. Cambridge, 2004.

[Jen+17] R. Jenatton, C. Archambeau, J. González, and M. Seeger. "Bayesian Optimization with Tree-structured Dependencies". en. In: *ICML*. 2017, pp. 1655–1664.

[Jeu+19] O. Jeunen, D. Mykhaylov, D. Rohde, F. Vasile, A. Gilotte, and M. Bompaire. "Learning from Bandit Feedback: An Overview of the State-of-the-art". In: (2019). arXiv: 1909.08471 [cs.IR].

[JG20] A. Jacovi and Y. Goldberg. "Towards faithfully interpretable NLP systems: How should we define and evaluate faithfulness?" In: *arXiv preprint arXiv:2004.03685* (2020).

[JG21] A. Jacovi and Y. Goldberg. "Aligning faithful interpretations with their social attribution". In: *Transactions of the Association for Computational Linguistics* 9 (2021), pp. 294–310.

[JGH18] A. Jacot, F. Gabriel, and C. Hongler. "Neural Tangent Kernel: Convergence and Generalization in Neural Networks". In: *NIPS*. 2018.

[JGP17] E. Jang, S. Gu, and B. Poole. "Categorical Reparameterization with Gumbel-Softmax". In: *ICLR*. 2017.

[Jia+19] R. Jia, A. Raghunathan, K. Göksel, and P. Liang. "Certified Robustness to Adversarial Word Substitutions". In: *EMNLP*. 2019.

[Jia+21] C. Jia, Y. Yang, Y. Xia, Y.-T. Chen, Z. Parekh, H. Pham, Q. V. Le, Y. Sung, Z. Li, and T. Duerig. *Scaling Up Visual and Vision-Language Representation Learning With Noisy Text Supervision*. 2021. arXiv: 2102.05918 [cs.CV].

[Jia21] H. Jiang. "Minimizing convex functions with integral minimizers". In: *Proceedings of the 2021 ACM-SIAM Symposium on Discrete Algorithms (SODA)*. SIAM. 2021, pp. 976–985.

[Jih+12] Jihong Min, J Kim, Seunghak Shin, and I. S. Kweon. "Efficient Data-Driven MCMC sampling for vision-based 6D SLAM". In: *ICRA*. 2012, pp. 3025–3032.

[Jin11] Y. Jin. "Surrogate-assisted evolutionary computation: Recent advances and future challenges". In: *Swarm and Evolutionary Computation* 1.2 (2011), pp. 61–70.

[Jit+16] W. Jitkrittum, Z. Szabó, K. P. Chwialkowski, and A. Gretton. "Interpretable Distribution Features with Maximum Testing Power". In: *NIPS*. Curran Associates, Inc., 2016, pp. 181–189.

[JJ00] T. S. Jaakkola and M. I. Jordan. "Bayesian parameter estimation via variational methods". In: *Statistics and Computing* 10 (2000), pp. 25–37.

[JKG18] H. Jiang, B. Kim, and M. Y. Guan. "To Trust Or Not To Trust A Classifier". In: *NIPS*. 2018.

[JKJ12] K. Jiang, B. Kulis, and M. Jordan. "Small-variance asymptotics for exponential family Dirichlet process mixture models". In: *Advances in Neural Information Processing Systems* 25 (2012), pp. 3158–3166.

[JKK95] C. S. Jensen, A. Kong, and U. Kjaerulff. "Blocking-Gibbs Sampling in Very Large Probabilistic Expert Systems". In: *Intl. J. Human-Computer Studies* (1995), pp. 647–666.

[JL15a] V. Jalali and D. Leake. "CBR meets big data: A case study of large-scale adaptation rule generation". In: *International Conference on Case-Based Reasoning*. Springer. 2015, pp. 181–196.

[JL15b] V. Jalali and D. B. Leake. "Enhancing case-based regression with automatically-generated ensembles of adaptations". In: *Journal of Intelligent Information Systems* 46 (2015), pp. 237–258.

[JL16] N. Jiang and L. Li. "Doubly Robust Off-policy Evaluation for Reinforcement Learning". In: *ICML*. 2016, pp. 652–661.

[JM00] D. Jurafsky and J. H. Martin. *Speech and language processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition*. Prentice-Hall, 2000.

[JM08] D. Jurafsky and J. H. Martin. *Speech and language processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition*. 2nd edition. Prentice-Hall, 2008.

[JM18] A. Jolicoeur-Martineau. "The relativistic discriminator: a key element missing from standard GAN". In: *arXiv preprint arXiv:1807.00734* (2018).

[JM70] D. H. Jacobson and D. Q. Mayne. *Differential Dynamic Programming*. Elsevier Press, 1970.

[JMW06] J. K. Johnson, D. M. Malioutov, and A. S. Willsky. "Walk-sum interpretation and analysis of Gaussian belief propagation". In: *NIPS*. 2006, pp. 579–586.

[JNJ20] C. Jin, P. Netrapalli, and M. I. Jordan. "What is local optimality in nonconvex-nonconcave minimax optimization?" In: *Proceedings of the 34th International Conference on Machine Learning-Volume 73*. 2020.

[JO18] M. Jankowiak and F. Obermeyer. "Pathwise Derivatives Beyond the Reparameterization Trick". In: *ICML*. 2018.

[JOA10] T. Jaksch, R. Ortner, and P. Auer. "Near-optimal Regret Bounds for Reinforcement Learning". In: *JMLR* 11 (2010), pp. 1563–1600.

[JOA17] P. E. Jacob, J. O'Leary, and Y. F. Atchadé. "Unbiased Markov Chain Monte Carlo with couplings". In: *arXiv preprint arXiv:1708.03625* (2017).

[Joh12] M. J. Johnson. "A Simple Explanation of A Spectral Algorithm for Learning Hidden Markov Models". In: (2012). arXiv: 1204.2477 [stat.ME].

[Jon01] D. R. Jones. "A Taxonomy of Global Optimization Methods Based on Response Surfaces". In: *J. Global Optimiz.* 21.4 (2001), pp. 345–383.

[Jor07] M. I. Jordan. *An Introduction to Probabilistic Graphical Models*. In preparation. 2007.

[Jor11] M. I. Jordan. "The era of Big Data". In: *ISBA Bulletin*. Vol. 18. 2011, pp. 1–3.

[Jor+98] M. I. Jordan, Z. Ghahramani, T. S. Jaakkola, and L. K. Saul. "An introduction to variational methods for graphical models". In: *Learning in Graphical Models*. Ed. by M. Jordan. MIT Press, 1998.

[Jos+17] A. Joshi, S. Ghosh, M. Betke, S. Sclaroff, and H. Pfister. "Personalizing gesture recognition using hierarchical Bayesian neural networks". In: *CVPR*. Honolulu, HI: IEEE, July 2017.

[Jos20] C. Joshi. *Transformers are Graph Neural Networks*. Tech. rep. 2020.

[Jos+20] M. Joshi, D. Chen, Y. Liu, D. S. Weld, L. Zettlemoyer, and O. Levy. "Spanbert: Improving pre-training by representing and predicting spans". In: *Transactions of the Association for Computational Linguistics* 8 (2020), pp. 64–77.

[Jos+22] L. V. Jospin, W. Buntine, F. Boussaid, H. Laga, and M. Bennamoun. "Hands-on Bayesian Neural Networks – a Tutorial for Deep Learning Users". In: (2022).

[Jou+16] A. Joulin, L. Van Der Maaten, A. Jabri, and N. Vasilache. "Learning visual features from large weakly supervised data". In: *European Conference on Computer Vision*. Springer. 2016, pp. 67–84.

[JP95] R. Jirousek and S. Preucil. "On the effective implementation of the iterative proportional fitting procedure". In: *Computational Statistics & Data Analysis* 19 (1995), pp. 177–189.

[JS93] M. Jerrum and A. Sinclair. "Polynomial-time approximation algorithms for the Ising model". In: *SIAM J. on Computing* 22 (1993), pp. 1087–1116.

[JS96] M. Jerrum and A. Sinclair. "The Markov chain Monte Carlo method: an approach to approximate counting and integration". In: *Approximation Algorithms for NP-hard problems*. Ed. by D. S. Hochbaum. PWS Publishing, 1996.

[JSY19] P. Jaini, K. A. Selby, and Y. Yu. "Sum-of-Squares Polynomial Flow". In: *ICML*. 2019, pp. 3009–3018.

[JU97] S. Julier and J. Uhlmann. "A New Extension of the Kalman Filter to Nonlinear Systems". In: *Proc. of AeroSense: The 11th Intl. Symp. on Aerospace/Defence Sensing, Simulation and Controls*. 1997.

[JUDW00] S Julier, J Uhlmann, and H. F. Durrant-Whyte. "A new method for the nonlinear transformation of means and covariances in filters and estimators". In: *IEEE Trans. Automat. Contr.* 45.3 (Mar. 2000), pp. 477–482.

[JW14] M. Johnson and A. Willsky. "Stochastic Variational Inference for Bayesian Time Series Models". en. In: *ICML*. 2014, pp. 1854–1862.

[JW19] S. Jain and B. C. Wallace. "Attention is not explanation". In: *arXiv preprint arXiv:1902.10186* (2019).

[Kaa12] Kaare Brandt Petersen and Michael Syskind Pedersen. *The Matrix Cookbook*. 2012.

[KAG19] A. Kirsch, J. van Amersfoort, and Y. Gal. "BatchBALD: Efficient and Diverse Batch Acquisition for Deep Bayesian Active Learning". In: *NIPS*. 2019.

[KAH19] F. H. Kingma, P. Abbeel, and J. Ho. "Bit-Swap: Recursive Bits-Back Coding for Lossless Compression with Hierarchical Latent Variables". In: *ICML*. 2019.

[Kai58] H. Kaiser. "The varimax criterion for analytic rotation in factor analysis". In: *Psychometrika* 23.3 (1958).

[Kak02] S. M. Kakade. "A Natural Policy Gradient". In: *NIPS*. 2002, pp. 1531–1538.

[Kal06] O. Kallenberg. *Foundations of modern probability*. Springer Science & Business Media, 2006.

[Kal+11] M. Kalakrishnan, S. Chitta, E. A. Theodorou, P. Pastor, and S. Schaal. "STOMP: Stochastic Trajectory Optimization for Motion Planning". In: *ICRA*. 2011, pp. 4569–4574.

[Kal+18a] D. Kalashnikov et al. "QT-Opt: Scalable Deep Reinforcement Learning for Vision-Based Robotic Manipulation". In: *CORL*. 2018.

[Kal+18b] N. Kalchbrenner, E. Elsen, K. Simonyan, S. Noury, N. Casagrande, E. Lockhart, F. Stimberg, A. Oord, S. Dieleman, and K. Kavukcuoglu. "Efficient neural audio synthesis". In: *International Conference on Machine Learning*. PMLR. 2018, pp. 2410–2419.

[Kam16] E. Kamar. "Directions in Hybrid Intelligence: Complementing AI Systems with Human Intelligence." In: *IJCAI*. 2016, pp. 4070–4073.

[Kam+22] S. Kamthe, S. Takao, S. Mohamed, and M. P. Deisenroth. "Iterative state estimation in non-linear dynamical systems using approximate expectation propagation". In: *Trans. on Machine Learning Research* (2022).

[Kan+20] T. Kaneko, H. Kameoka, K. Tanaka, and N. Hojo. "CycleGAN-VC3: Examining and Improving CycleGAN-VCs for Mel-spectrogram Conversion". In: *Interspeech conference proceedings* (2020).

[Kan42] L. Kantorovich. "On the transfer of masses (in Russian)". In: *Doklady Akademii Nauk* 37.2 (1942), pp. 227–229.

[Kap+22] S. Kapoor, W. J. Maddox, P. Izmailov, and A. G. Wilson. "On Uncertainty, Tempering, and Data Augmentation in Bayesian Classification". In: *arXiv preprint arXiv:2203.16481* (2022).

[Kar+18] T. Karras, T. Aila, S. Laine, and J. Lehtinen. "Progressive Growing of GANs for Improved Quality, Stability, and Variation". In: *ICLR*. 2018.

[Kar+20a] A.-H. Karimi, G. Barthe, B. Balle, and I. Valera. *Model-Agnostic Counterfactual Explanations for Consequential Decisions*. 2020. arXiv: 1905.11190 [cs.LG].

[Kar+20b] A.-H. Karimi, G. Barthe, B. Schölkopf, and I. Valera. "A survey of algorithmic recourse: definitions, formulations, solutions, and prospects". In: *arXiv preprint arXiv:2010.04050* (2020).

[Kar+20c] T. Karras, S. Laine, M. Aittala, J. Hellsten, J. Lehtinen, and T. Aila. "Analyzing and improving the image quality of stylegan". In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2020, pp. 8110–8119.

[Kar+21] T. Karras, M. Aittala, S. Laine, E. Härkönen, J. Hellsten, J. Lehtinen, and T. Aila. "Alias-Free Generative Adversarial Networks". In: *arXiv preprint arXiv:2106.12423* (2021).

[Kat05] T. Katayama. *Subspace Methods for Systems Identification*. Springer Verlag, 2005.

[Kat+06] H. G. Katzgraber, S. Trebst, D. A. Huse, and M. Troyer. "Feedback-optimized parallel tempering Monte Carlo". In: *Journal of Statistical Mechanics: Theory and Experiment* 2006.03 (2006), P03018.

[Kat+17] G. Katz, C. Barrett, D. L. Dill, K. Julian, and M. J. Kochenderfer. "Reluplex: An efficient SMT solver for verifying deep neural networks". In: *International Conference on Computer Aided Verification*. Springer. 2017, pp. 97–117.

[Kat+19] N. Kato, H. Osone, K. Oomori, C. W. Ooi, and Y. Ochiai. "GANs-Based Clothes Design: Pattern Maker Is All You Need to Design Clothing". In: *Proceedings of the 10th Augmented Human International Conference 2019*. Association for Computing Machinery, 2019.

[Kau+19] V. Kaushal, R. Iyer, S. Kothawade, R. Mahadev, K. Doctor, and G. Ramakrishnan. "Learning from less data: A unified data subset selection and active learning framework for computer vision". In: *2019 IEEE Winter Conference on Applications of Computer Vision (WACV)*. IEEE. 2019, pp. 1289–1299.

[Kau+20] H. Kaur, H. Nori, S. Jenkins, R. Caruana, H. Wallach, and J. Wortman Vaughan. "Interpreting interpretability: understanding data scientists' use of interpretability tools for machine learning". In: *Proceedings of the 2020 CHI conference on human factors in computing systems*. 2020, pp. 1–14.

[Kau+21] D. Kaushik, A. Setlur, E. Hovy, and Z. C. Lipton. "Explaining The Efficacy of Counterfactually Augmented Data". In: *ICLR*. 2021.

[KB00] H. J. Kushner and A. S. Budhiraja. "A nonlinear filtering algorithm based on an approximation of the conditional distribution". In: *IEEE Trans. Automat. Contr.* 45.3 (2000), pp. 580–585.

[KB14a] D. P. Kingma and J. Ba. "Adam: A method for stochastic optimization". In: *arXiv preprint arXiv:1412.6980* (2014).

[KB14b] K. Kirchhoff and J. Bilmes. "Submodularity for Data Selection in Machine Translation". In: *Empirical Methods in Natural Language Processing (EMNLP)*. 2014.

[KB15] D. Kingma and J. Ba. "Adam: A Method for Stochastic Optimization". In: *ICLR*. 2015.

[KB16] T. Kim and Y. Bengio. "Deep directed generative models with energy-based probability estimation". In: *arXiv preprint arXiv:1606.03439* (2016).

[KBH19] F. Kunstner, L. Balles, and P. Hennig. "Limitations of the Empirical Fisher Approximation". In: (2019). arXiv: 1905.12558 [cs.LG].

[KCC20] V. Kumar, A. Choudhary, and E. Cho. "Data Augmentation using Pre-trained Transformer Models". In: *Proceedings of the 2nd Workshop on Lifelong Learning for Spoken Language Systems*. Suzhou, China: Association for Computational Linguistics, Dec. 2020, pp. 18–26.

[KD18a] S. Kamthe and M. P. Deisenroth. "Data-Efficient Reinforcement Learning with Probabilistic Model Predictive Control". In: *AISTATS*. 2018.

[KD18b] D. P. Kingma and P. Dhariwal. "Glow: Generative Flow with Invertible 1×1 Convolutions". In: *NIPS*. 2018.

[Ke+19a] L. Ke, M. Barnes, W. Sun, G. Lee, S. Choudhury, and S. Srinivasa. "Imitation Learning as f-Divergence Minimization". In: *arXiv preprint arXiv:1905.12888* (2019).

[Ke+19b] L. Ke, S. Choudhury, M. Barnes, W. Sun, G. Lee, and S. Srinivasa. *Imitation Learning as f-Divergence Minimization*. arXiv:1905.12888. 2019.

[Ke+21] A. Ke, W. Ellsworth, O. Banerjee, A. Y. Ng, and P. Rajpurkar. "CheXtransfer: performance and parameter efficiency of ImageNet models for chest X-Ray interpretation". In: *Proceedings of the Conference on Health, Inference, and Learning*. 2021, pp. 116–124.

[Kei06] F. C. Keil. "Explanation and understanding". In: *Annu. Rev. Psychol.* 57 (2006), pp. 227–254.

[Kel+20] J. Kelly, J. Bettencourt, M. J. Johnson, and D. Duvenaud. "Learning Differential Equations that are Easy to Solve". In: *Neural Information Processing Systems*. 2020.

[Kem+06] C. Kemp, J. Tenenbaum, T. Y. T. Griffiths and, and N. Ueda. "Learning systems of concepts with an infinite relational model". In: *AAAI*. 2006.

[Kem+10] C. Kemp, J. Tenenbaum, S. Niyogi, and T. Griffiths. "A probabilistic model of theory formation". In: *Cognition* 114 (2010), pp. 165–196.

[Ken16] E. H. Kennedy. "Semiparametric theory and empirical processes in causal inference". In: *Statistical causal inferences and their applications in public health research*. ICSA Book Ser. Stat. Springer, [Cham], 2016, pp. 141–167.

[Ken17] E. H. Kennedy. *Semiparametric theory*. 2017. arXiv: 1709.06418 [stat.ME].

[Ken20] E. H. Kennedy. *Optimal doubly robust estimation of heterogeneous causal effects*. 2020. arXiv: 2004.14497 [math.ST].

[Kes+17] N. S. Keskar, D. Mudigere, J. Nocedal, M. Smelyanskiy, and P. T. P. Tang. "On Large-Batch Training for Deep Learning: Generalization Gap and Sharp Minima". In: *ICLR*. 2017.

[KF09a] D. Koller and N. Friedman. *Probabilistic Graphical Models: Principles and Techniques*. MIT Press, 2009.

[KF09b] D. Krishnan and R. Fergus. "Fast Image Deconvolution using Hyper-Laplacian Priors". In: *NIPS*. 2009, pp. 1033–1041.

[KFL01] F. Kschischang, B. Frey, and H.-A. Loeliger. "Factor Graphs and the Sum-Product Algorithm". In: *IEEE Trans Info. Theory* (2001).

[KG05] A. Krause and C. Guestrin. "Near-optimal Nonmyopic Value of Information in Graphical Models". In: *Proc. of the 21st Annual Conf. on Uncertainty in Artificial Intelligence (UAI 2005)*. AUAI Press, 2005, pp. 324–331.

[KG17] A. Kendall and Y. Gal. "What Uncertainties Do We Need in Bayesian Deep Learning for Computer Vision?" In: *NIPS*. Curran Associates, Inc., 2017, pp. 5574–5584.

[KGO12] H. J. Kappen, V. Gómez, and M. Opper. "Optimal control as a graphical model inference problem". In: *Mach. Learn.* 87.2 (2012), pp. 159–182.

[KGW11] M. Kalli, J. E. Griffin, and S. G. Walker. "Slice sampling mixture models". In: *Statistics and computing* 21.1 (2011), pp. 93–105.

[KH22] L. Kurscheidt and M. Hein. "Lost in Translation: Modern Image Classifiers still degrade even under simple Translations". In: *ICML 2022 Shift Happens Workshop*. July 2022.

[Kha+10] M. E. Khan, B. Marlin, G. Bouchard, and K. P. Murphy. "Variational bounds for mixed-data factor analysis". In: *NIPS*. 2010.

[Kha+18] M. E. Khan, D. Nielsen, V. Tangkaratt, W. Lin, Y. Gal, and A. Srivastava. "Fast and Scalable Bayesian Deep Learning by Weight-Perturbation in Adam". In: *ICML*. 2018.

[Kha20] M. E. Khan. *Deep learning with Bayesian principles*. NeurIPS tutorial. 2020.

[Kha+21] S. Khan, M. Naseer, M. Hayat, S. W. Zamir, F. S. Khan, and M. Shah. "Transformers in Vision: A Survey". In: *ACM Computing Surveys December* (2021).

[Khe+20] I. Khemakhem, R. Monti, D. Kingma, and A. Hyvarinen. "ICE-BeeM: Identifiable Conditional Energy-Based Deep Models Based on Nonlinear ICA". In: *Advances in Neural Information Processing Systems*. Ed. by H. Larochelle, M. Ranzato, R. Hadsell, M. Balcan, and H. Lin. Vol. 33. Curran Associates, Inc., 2020, pp. 12768–12778.

[KHH20] A. Kristiadi, M. Hein, and P. Hennig. "Being Bayesian, Even Just a Bit, Fixes Overconfidence in ReLU Networks". In: *ICML*. 2020.

[KHL20] D. Kaushik, E. Hovy, and Z. C. Lipton. *Learning the Difference that Makes a Difference with Counterfactually-Augmented Data*. 2020. arXiv: 1909.12434 [cs.CL].

[Kho+20] P. Khosla, P. Teterwak, C. Wang, A. Sarna, Y. Tian, P. Isola, A. Maschinot, C. Liu, and D. Krishnan. "Supervised Contrastive Learning". In: *ArXiv* abs/2004.11362 (2020).

[Kil+20] K. Killamsetty, D. Sivasubramanian, G. Ramakrishnan, and R. Iyer. "GLISTER: Generalization based Data Subset Selection for Efficient and Robust Learning". In: *arXiv preprint arXiv:2012.10630* (2020).

[Kim+18a] B. Kim, M. Wattenberg, J. Gilmer, C. Cai, J. Wexler, F. Viegas, and R. Sayres. "Interpretability Beyond Feature Attribution: Quantitative Testing with Concept Activation Vectors (TCAV)". In: *ICML*. 2018.

[Kim+18b] B. Kim, M. Wattenberg, J. Gilmer, C. Cai, J. Wexler, F. Viegas, et al. "Interpretability beyond feature attribution: Quantitative testing with concept activation vectors (tcav)". In: *International conference on machine learning*. PMLR. 2018, pp. 2668–2677.

[Kim+18c] Y. Kim, S. Wiseman, A. C. Miller, D. Sontag, and A. M. Rush. "Semi-Amortized Variational Autoencoders". In: *ICML*. 2018.

[Kim+19] S. Kim, S.-G. Lee, J. Song, J. Kim, and S. Yoon. "FloWaveNet : A Generative Flow for Raw Audio". In: *Proceedings of the 36th International Conference on Machine Learning*. 2019, pp. 3370–3378.

[Kin+14a] D. P. Kingma, D. J. Rezende, S. Mohamed, and M. Welling. "Semi-Supervised Learning with Deep Generative Models". In: *NIPS*. 2014.

[Kin+14b] D. P. Kingma, D. J. Rezende, S. Mohamed, and M. Welling. *Semi-Supervised Learning with Deep Generative Models*. 2014. arXiv: 1406.5298 [cs.LG].

[Kin+16] D. P. Kingma, T. Salimans, R. Jozefowicz, X. Chen, I. Sutskever, and M. Welling. "Improved Variational Inference with Inverse Autoregressive Flow". In: *NIPS*. 2016.

[Kin+19] P.-J. Kindermans, S. Hooker, J. Adebayo, M. Alber, K. T. Schütt, S. Dähne, D. Erhan, and B. Kim. "The (un) reliability of saliency methods". In: *Explainable AI: Interpreting, Explaining and Visualizing Deep Learning*. Springer, 2019, pp. 267–280.

[Kin+21] D. P. Kingma, T. Salimans, B. Poole, and J. Ho. "Variational Diffusion Models". In: *NIPS*. July 2021.

[Kir+17] J. Kirkpatrick et al. "Overcoming catastrophic forgetting in neural networks". en. In: *PNAS* 114.13 (2017), pp. 3521–3526.

[Kir+21] A. Kirsch, J. M. J. van Amersfoort, P. H. S. Torr, and Y. Gal. "On pitfalls in OoD detection: Predictive entropy considered harmful". In: *ICML Workshop on Uncertainty in Deep Learning*. 2021.

[Kit04] G. Kitagawa. "The two-filter formula for smoothing and an implementation of the Gaussian-sum smoother". In: *Annals of the Institute of Statistical Mathematics* 46.4 (2004), pp. 605–623.

[KIW20] P. Kirichenko, P. Izmailov, and A. G. Wilson. "Why Normalizing Flows Fail to Detect Out-of-Distribution Data". In: (2020). arXiv: 2006.08545 [stat.ML].

[KJ12a] J. Z. Kolter and T. S. Jaakkola. "Approximate Inference in Additive Factorial HMMs with Application to Energy Disaggregation". In: *AISTATS*. 2012.

[KJ12b] B. Kulis and M. I. Jordan. "Revisiting K-Means: New Algorithms via Bayesian Nonparametrics". In: *Proceedings of the 29th International Coference on International Conference on Machine Learning*. ICML'12. Omnipress, 2012, 1131–1138.

[Kja90] U. Kjaerulff. *Triangulation of graphs – algorithms giving small total state space*. Tech. rep. R-90-09. Dept. of Math. and Comp. Sci., Aalborg Univ., Denmark, 1990.

[Kja92] U. Kjaerulff. "Optimal decomposition of probabilistic networks by simulated annealing". In: *Statistics and Computing*. Vol. 2. 1992, pp. 7–17.

[KJD18] J. Knoblauch, J. Jewson, and T. Damoulas. "Doubly Robust Bayesian Inference for Non-Stationary Streaming Data with β-Divergences". In: *NIPS*. 2018.

[KJD19] J. Knoblauch, J. Jewson, and T. Damoulas. "Generalized Variational Inference: Three arguments for deriving new Posteriors". In: (2019). arXiv: 1904.02063 [stat.ML].

[KJD21] J. Knoblauch, J. Jewson, and T. Damoulas. "An Optimization-centric View on Bayes' Rule: Reviewing and Generalizing Variational Inference". In: *JMLR* (2021).

[KJM19] N. M. Kriege, F. D. Johansson, and C. Morris. "A Survey on Graph Kernels". In: (2019). arXiv: 1903.11835 [cs.LG].

[KJV83] S. Kirkpatrick, C. G. Jr., and M. Vecchi. "Optimization by simulated annealing". In: *Science* 220 (1983), pp. 671–680.

[KK11] P. Krähenbühl and V. Koltun. "Efficient Inference in Fully Connected CRFs with Gaussian Edge Potentials". In: *NIPS*. 2011.

[KKG22] A. Kirsch, J. Kossen, and Y. Gal. "Marginal and Joint Cross-Entropies & Predictives for Online Bayesian Inference, Active Learning, and Active Sampling". In: (May 2022). arXiv: 2205.08766 [cs.LG].

[KKH20] I. Khemakhem, D. P. Kingma, and A. Hyvärinen. "Variational Autoencoders and Nonlinear ICA: A Unifying Framework". In: *AISTATS*. 2020.

[KKL20] N. Kitaev, L. Kaiser, and A. Levskaya. "Reformer: The Efficient Transformer". In: *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net, 2020.

[KKR95] K. Kanazawa, D. Koller, and S. Russell. "Stochastic Simulation Algorithms for Dynamic Probabilistic Networks". In: *UAI*. 1995.

[KKS20] F. Kunstner, R. Kumar, and M. Schmidt. "Homeomorphic-Invariance of EM: Non-Asymptotic Convergence in KL Divergence for Exponential Families via Mirror Descent". In: (2020). arXiv: 2011.01170 [cs.LG].

[KKT03] D. Kempe, J. Kleinberg, and É. Tardos. "Maximizing the spread of influence through a social network". In: *Proceedings of the ninth ACM SIGKDD international conference on Knowledge discovery and data mining*. 2003, pp. 137–146.

[KL02] S. Kakade and J. Langford. "Approximately Optimal Approximate Reinforcement Learning". In: *ICML*. ICML '02. Morgan Kaufmann Publishers Inc., 2002, pp. 267–274.

[KL09] H. Kawakatsu and A. Largey. "EM algorithms for ordered probit models with endogenous regressors". In: *The Econometrics Journal* 12.1 (2009), pp. 164–186.

[KL10] D. P. Kingma and Y. LeCun. "Regularized estimation of image statistics by score matching". In: *Advances in neural information processing systems*. 2010, pp. 1126–1134.

[KL17a] M. E. Khan and W. Lin. "Conjugate-Computation Variational Inference : Converting Variational Inference in Non-Conjugate Models to Inferences in Conjugate Models". In: *AISTATS*. 2017.

[KL17b] P. W. Koh and P. Liang. "Understanding black-box predictions via influence functions". In: *International Conference on Machine Learning*. PMLR. 2017, pp. 1885–1894.

[KL21] W. M. Kouw and M. Loog. "A review of domain adaptation without target labels". en. In: *IEEE Trans. Pattern Anal. Mach. Intell.* (2021).

[Kla+06] M. Klaas, M. Briers, N. de Freitas, A. Doucet, S. Maskell, and D. Lang. "Fast Particle Smoothing: If I Had a Million Particles". In: *ICML*. 2006.

[KLA19] T. Karras, S. Laine, and T. Aila. "A Style-Based Generator Architecture for Generative Adversarial Networks". In: *CVPR*. 2019.

[Kle17] G. A. Klein. *Sources of power: How people make decisions*. MIT press, 2017.

[Kle18] J. Kleinberg. "Inherent Trade-Offs in Algorithmic Fairness". In: *ACM International Conference on Measurement and Modeling of Computer Systems*. SIGMETRICS '18. Irvine, CA, USA: Association for Computing Machinery, June 2018, p. 40.

[KLM19] A. Kumar, P. Liang, and T. Ma. "Verified Uncertainty Calibration". In: *NIPS*. 2019.

[KM00] J. Kwon and K. Murphy. *Modeling Freeway Traffic with Coupled HMMs*. Tech. rep. Univ. California, Berkeley, 2000.

[KM08] U. Kjaerulff and A. Madsen. *Bayesian Networks and Influence Diagrams: A Guide to Construction and Analysis*. Springer, 2008.

[KM22] J. Kuan and J. Mueller. "Back to the Basics: Revisiting Out-of-Distribution Detection Baselines". In: *ICML PODS Workshop*. July 2022.

[KMB08] N. Kriegeskorte, M. Mur, and P. Bandettini. "Representational similarity analysis - connecting the branches of systems neuroscience". en. In: *Front. Syst. Neurosci.* 2 (Nov. 2008), p. 4.

[KML20] A. Kumar, T. Ma, and P. Liang. "Understanding Self-Training for Gradual Domain Adaptation". In: *ICML*. Ed. by H. D. Iii and A. Singh. Vol. 119. Proceedings of Machine Learning Research. PMLR, 2020, pp. 5468–5479.

[KMR16] J. Kleinberg, S. Mullainathan, and M. Raghavan. "Inherent trade-offs in the fair determination of risk scores". In: *arXiv preprint arXiv:1609.05807* (2016).

[KMY04] D. Kersten, P. Mamassian, and A. Yuille. "Object perception as Bayesian inference". en. In: *Annu. Rev. Psychol.* 55 (2004), pp. 271–304.

[KN09] J. Z. Kolter and A. Y. Ng. "Near-Bayesian Exploration in Polynomial Time". In: *ICML*. 2009.

[KN95] R. Kneser and H. Ney. "Improved backing-off for n-gram language modeling". In: *ICASSP*. Vol. 1. 1995, pp. 181–184.

[KN98] C.-J. Kim and C. Nelson. *State-Space Models with Regime-Switching: Classical and Gibbs-Sampling Approaches with Applications*. MIT Press, 1998.

[KNP11] K. Kersting, S. Natarajan, and D. Poole. *Statistical Relational AI: Logic, Probability and Computation*. Tech. rep. UBC, 2011.

[KNT20] I. Kostrikov, O. Nachum, and J. Tompson. "Imitation Learning via Off-Policy Distribution Matching". In: *ICLR*. 2020.

[Koe05] R. Koenker. *Quantile Regression*. en. Cambridge University Press, 2005.

[Koh+20a] P. W. Koh, T. Nguyen, Y. S. Tang, S. Mussmann, E. Pierson, B. Kim, and P. Liang. "Concept bottleneck models". In: *International Conference on Machine Learning*. PMLR. 2020, pp. 5338–5348.

[Koh+20b] P. W. Koh et al. "WILDS: A Benchmark of in-the-Wild Distribution Shifts". In: (Dec. 2020). arXiv: 2012.07421 [cs.LG].

[Kol+20] A. Kolesnikov, L. Beyer, X. Zhai, J. Puigcerver, J. Yung, S. Gelly, and N. Houlsby. "Big transfer (BiT): General visual representation learning". In: *ECCV*. Springer. 2020, pp. 491–507.

[Kon+21] Z. Kong, W. Ping, J. Huang, K. Zhao, and B. Catanzaro. "DiffWave: A Versatile Diffusion Model for Audio Synthesis". In: *ICLR*. 2021.

[Koo03] G. Koop. *Bayesian econometrics*. Wiley, 2003.

[Kor+15] A. Korattikara, V. Rathod, K. Murphy, and M. Welling. "Bayesian Dark Knowledge". In: *NIPS*. 2015.

[Kor+19] S. Kornblith, M. Norouzi, H. Lee, and G. Hinton. "Similarity of neural network representations revisited". In: *arXiv preprint arXiv:1905.00414* (2019).

[Kor+20] A. Korotin, V. Egiazarian, A. Asadulaev, A. Safin, and E. Burnaev. "Wasserstein-2 Generative Networks". In: *International Conference on Learning Representations*. 2020.

[Kor+21] S. Kornblith, T. Chen, H. Lee, and M. Norouzi. "Why do better loss functions lead to less transferable features?" In: *Advances in Neural Information Processing Systems*. Vol. 34. 2021.

[Kot+17] L. Kotthoff, C. Thornton, H. H. Hoos, F. Hutter, and K. Leyton-Brown. "Auto-WEKA 2.0: Automatic model selection and hyperparameter optimization in WEKA". In: *JMLR* 18.25 (2017), pp. 1–5.

[Kot+22] S. Kothawade, V. Kaushal, G. Ramakrishnan, J. Bilmes, and R. Iyer. "PRISM: A Rich Class of Parameterized Submodular Information Measures for Guided Subset Selection". In: *Proceedings of the AAAI Conference on Artificial Intelligence*. 2022.

[Koy+10] S. Koyama, L. C. Pérez-Bolde, C. R. Shalizi, and R. E. Kass. "Approximate Methods for State-Space Models". en. In: *JASA* 105.489 (2010), pp. 170–180.

[Koz92] J. R. Koza. *Genetic Programming*. https://mitpress.mit.edu/books/genetic-programming. Accessed: 2017-11-26. 1992.

[KP20] A. Kumar and B. Poole. "On Implicit Regularization in β-VAEs". In: *ICML*. 2020.

[KPB19] I. Kobyzev, S. Prince, and M. A. Brubaker. "Normalizing Flows: An Introduction and Review of Current Methods". In: (2019). arXiv: 1908.09257 [stat.ML].

[KPHL17] M. J. Kusner, B. Paige, and J. M. Hernández-Lobato. "Grammar Variational Autoencoder". In: *ICML*. 2017.

[KPT21] J. S. Kim, G. Plumb, and A. Talwalkar. "Sanity Simulations for Saliency Methods". In: *arXiv preprint arXiv:2105.06506* (2021).

[KR21a] M. Khan and H. Rue. "The Bayesian Learning Rule". In: (2021).

[KR21b] S. Kong and D. Ramanan. "OpenGAN: Open-Set Recognition via Open Data Generation". In: *ICCV*. 2021.

[Kra+08] A. Krause, H Brendan McMahan, C. Guestrin, and A. Gupta. "Robust Submodular Observation Selection". In: *JMLR* 9 (2008), pp. 2761–2801.

[Kri+05] B. Krishnapuram, L. Carin, M. Figueiredo, and A. Hartemink. "Learning sparse Bayesian classifiers: multi-class formulation, fast algorithms, and generalization bounds". In: *IEEE Transaction on Pattern Analysis and Machine Intelligence* (2005).

[Kri19] N. Kriegeskorte. *What's the best measure of representational dissimilarity?* 2019.

[KRL08] K. Kavukcuoglu, M. Ranzato, and Y. LeCun. "Fast Inference in Sparse Coding Algorithms with Applications to Object Recognition". In: *NIPS workshop on optimization in machine learning*. 2008.

[KRS14] B. Kim, C. Rudin, and J. A. Shah. "The bayesian case model: A generative approach for case-based reasoning and prototype classification". In: *Advances in neural information processing systems*. 2014, pp. 1952–1960.

[Kru13] J. K. Kruschke. "Bayesian estimation supersedes the t test". In: *J. Experimental Psychology: General* 142.2 (2013), pp. 573–603.

[KS06] L. Kocsis and C. Szepesvári. "Bandit Based Monte-Carlo Planning". In: *ECML*. 2006, pp. 282–293.

[KS07] J. D. Y. Kang and J. L. Schafer. "Demystifying double robustness: a comparison of alternative strategies for estimating a population mean from incomplete data". In: *Statist. Sci.* 22.4 (2007), pp. 523–539.

[KS15] H. Kaya and A. A. Salah. "Adaptive Mixtures of Factor Analyzers". In: (2015). arXiv: 1507.02801 [stat.ML].

[KSB21] B. Kompa, J. Snoek, and A. Beam. "Empirical Frequentist Coverage of Deep Learning Uncertainty Quantification Procedures". In: *Entropy* 23.12 (2021).

[KSC98] S. Kim, N. Shephard, and S. Chib. "Stochastic Volatility: Likelihood Inference and Comparison with ARCH Models". In: *Review of Economic Studies* 65.3 (1998), pp. 361–393.

[KSH12a] A. Krizhevsky, I. Sutskever, and G. Hinton. "Imagenet classification with deep convolutional neural networks". In: *NIPS*. 2012.

[KSH12b] A. Krizhevsky, I. Sutskever, and G. E. Hinton. "Imagenet classification with deep convolutional neural networks". In: *Advances in neural information processing systems*. 2012, pp. 1097–1105.

[KSL19] S. Kornblith, J. Shlens, and Q. V. Le. "Do better ImageNet models transfer better?" In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2019, pp. 2661–2671.

[KSL21] S. Kim, Q. Song, and F. Liang. "Stochastic gradient Langevin dynamics with adaptive drifts". In: *J. Stat. Comput. Simul.* (2021), pp. 1–19.

[KSN99] N. L. Kleinman, J. C. Spall, and D. Q. Naiman. "Simulation-Based Optimization with Stochastic Approximation Using Common Random Numbers". In: *Manage. Sci.* 45.11 (1999), pp. 1570–1578.

[KSS17] R. G. Krishnan, U. Shalit, and D. Sontag. "Structured Inference Networks for Nonlinear State Space Models". In: *AAAI*. 2017.

[KT11] A. Kulesza and B. Taskar. "k-DPPs: Fixed-size determinantal point processes". In: *ICML*. 2011.

[KT+12] A. Kulesza, B. Taskar, et al. "Determinantal Point Processes for Machine Learning". In: *Foundations and Trends in Machine Learning* 5.2–3 (2012), pp. 123–286.

[KTB11] D. P. Kroese, T. Taimre, and Z. I. Botev. *Handbook of Monte Carlo Methods*. en. 1 edition. Wiley, 2011.

[KTX20] R. Kohavi, D. Tang, and Y. Xu. *Trustworthy Online Controlled Experiments: A Practical Guide to A/B Testing*. en. 1st ed. Cambridge University Press, 2020.

[KU21] S. Khan and J. Ugander. *Adaptive normalization for IPW estimation*. 2021. arXiv: 2106.07695 [stat.ME].

[Kua+09] P. Kuan, G. Pan, J. A. Thomson, R. Stewart, and S. Keles. *A hierarchical semi-Markov model for detecting enrichment with application to ChIP-Seq experiments*. Tech. rep. U. Wisconsin, 2009.

[Kub04] M. Kubale. *Graph colorings*. Vol. 352. American Mathematical Society, 2004.

[Kuc+16] A. Kucukelbir, D. Tran, R. Ranganath, A. Gelman, and D. M. Blei. "Automatic Differentiation Variational Inference". In: *JMLR* (2016).

[Kuh55] H. W. Kuhn. "The Hungarian method for the assignment problem". In: *Naval Research Logistics Quarterly* 2 (1955), pp. 83–97.

[Kul+13] T. Kulesza, S. Stumpf, M. Burnett, S. Yang, I. Kwan, and W.-K. Wong. "Too much, too little, or just right? Ways explanations impact end users' mental models". In: *2013 IEEE Symposium on visual languages and human centric computing*. IEEE. 2013, pp. 3–10.

[Kul+19] M. Kull, M. Perello-Nieto, M. Kängsepp, T. S. Filho, H. Song, and P. Flach. "Beyond temperature scaling: Obtaining well-calibrated multiclass probabilities with Dirichlet calibration". In: *NIPS*. 2019.

[Kum+19a] A. Kumar, J. Fu, M. Soh, G. Tucker, and S. Levine. "Stabilizing Off-Policy Q-Learning via Bootstrapping Error Reduction". In: *NeurIPS*. 2019, pp. 11761–11771.

[Kum+19b] M. Kumar, M. Babaeizadeh, D. Erhan, C. Finn, S. Levine, L. Dinh, and D. P. Kingma. "VideoFlow: A flow-based generative model for video". In: *ICML Workshop on Invertible Neural Networks and Normalizing Flows* (2019).

[Kum+19c] R. Kumar, S. Ozair, A. Goyal, A. Courville, and Y. Bengio. "Maximum entropy generators for energy-based models". In: *arXiv preprint arXiv:1901.08508* (2019).

[Kün+19] S. R. Künzel, J. S. Sekhon, P. J. Bickel, and B. Yu. "Metalearners for estimating heterogeneous treatment effects using machine learning". In: *Proceedings of the National Academy of Sciences* 116.10 (2019), pp. 4156–4165. eprint: https://www.pnas.org/content/116/10/4156.full.pdf.

[Kun20] J. Kuntz. "Markov chains revisited". In: (Jan. 2020). arXiv: 2001.02183 [math.PR].

[Kur+19] T. Kurutach, I. Clavera, Y. Duan, A. Tamar, and P. Abbeel. "Model-Ensemble Trust-Region Policy Optimization". In: *ICLR*. 2019.

[Kur+20] R. Kurle, B. Cseke, A. Klushyn, P. van der Smagt, and S. Günnemann. "Continual Learning with Bayesian Neural Networks for Non-Stationary Data". In: *ICLR*. 2020.

[Kus+18] M. J. Kusner, J. R. Loftus, C. Russell, and R. Silva. *Counterfactual Fairness*. 2018. arXiv: 1703.06856 [stat.ML].

[Kus64] H. J. Kushner. "A New Method of Locating the Maximum Point of an Arbitrary Multipeak Curve in the Presence of Noise". In: *J. Basic Eng* 86.1 (1964), pp. 97–106.

[KVK10] A. Klami, S. Virtanen, and S. Kaski. "Bayesian exponential family projections for coupled data sources". In: *UAI*. 2010.

[KW14] D. P. Kingma and M. Welling. "Auto-encoding variational Bayes". In: *ICLR*. 2014.

[KW18] A. S. I. Kim and M. P. Wand. "On expectation propagation for generalised, linear and mixed models". In: *Aust. N. Z. J. Stat.* 60.1 (2018), pp. 75–102.

[KW19a] D. P. Kingma and M. Welling. "An Introduction to Variational Autoencoders". In: *Foundations and Trends in Machine Learning* 12.4 (2019), pp. 307–392.

[KW19b] M. J. Kochenderfer and T. A. Wheeler. *Algorithms for Optimization*. en. The MIT Press, 2019.

[KW70] G. S. Kimeldorf and G. Wahba. "A Correspondence Between Bayesian Estimation on Stochastic Processes and Smoothing by Splines". en. In: *Ann. Math. Stat.* 41.2 (1970), pp. 495–502.

[KW96] R. E. Kass and L. Wasserman. "The Selection of Prior Distributions by Formal Rules". In: *JASA* 91.435 (1996), pp. 1343–1370.

[KWV06] K. Kurihara, M. Welling, and N. Vlassis. "Accelerated variational DP mixture models". In: *NIPS*. 2006.

[KWW22] M. J. Kochenderfer, T. A. Wheeler, and K. Wray. *Algorithms for Decision Making*. The MIT Press, 2022.

[KY94] J. J. Kosowsky and A. L. Yuille. "The invisible hand algorithm: Solving the assignment problem with statistical physics". In: *Neural networks* 7.3 (1994), pp. 477–490.

[Kyn+19] T. Kynkäänniemi, T. Karras, S. Laine, J. Lehtinen, and T. Aila. "Improved Precision and Recall

Metric for Assessing Generative Models". In: *NeurIPS*. 2019.

[KZ02] V. Kolmogorov and R. Zabih. "What energy functions can be minimized via graph cuts?" In: *Computer Vision—ECCV 2002* (2002), pp. 185–208.

[KZB19] A. Kolesnikov, X. Zhai, and L. Beyer. "Revisiting self-supervised visual representation learning". In: *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 2019, pp. 1920–1929.

[LA87] P. J. M. Laarhoven and E. H. L. Aarts, eds. *Simulated Annealing: Theory and Applications*. Kluwer Academic Publishers, 1987.

[Lag+19] I. Lage, E. Chen, J. He, M. Narayanan, B. Kim, S. J. Gershman, and F. Doshi-Velez. "Human evaluation of models built for interpretability". In: *Proceedings of the AAAI Conference on Human Computation and Crowdsourcing*. Vol. 7. 1. 2019, pp. 59–67.

[Lak+17] B. M. Lake, T. D. Ullman, J. B. Tenenbaum, and S. J. Gershman. "Building Machines That Learn and Think Like People". en. In: *Behav. Brain Sci.* (2017), pp. 1–101.

[Lal+21] A. Lal, M. W. Lockhart, Y. Xu, and Z. Zu. "How Much Should We Trust Instrumental Variable Estimates in Political Science? Practical Advice based on Over 60 Replicated Studies". In: (2021).

[Lam92] D. Lambert. "Zero-Inflated Poisson Regression, with an Application to Defects in Manufacturing". In: *Technometrics* 34.1 (1992), pp. 1–14.

[Lan95a] K. Lange. "A gradient algorithm locally equivalent to the em algorithm". en. In: *J. of Royal Stat. Soc. Series B* 57.2 (July 1995), pp. 425–437.

[Lan95b] K. Lange. "A QUASI-NEWTON ACCELERATION OF THE EM ALGORITHM". In: *Statistica Sinica* 5.1 (1995), pp. 1–18.

[Lao+20] J. Lao, C. Suter, I. Langmore, C. Chimisov, A. Saxena, P. Sountsov, D. Moore, R. A. Saurous, M. D. Hoffman, and J. V. Dillon. "tfp.mcmc: Modern Markov Chain Monte Carlo Tools Built for Modern Hardware". In: *PROBPROG*. 2020.

[Lar+16] A. B. L. Larsen, S. K. Sonderby, H. Larochelle, and O. Winther. "Autoencoding beyond pixels using a learned similarity metric". In: *International conference on machine learning*. PMLR. 2016, pp. 1558–1566.

[Lar+21] B. W. Larsen, S. Fort, N. Becker, and S. Ganguli. "How many degrees of freedom do we need to train deep networks: a loss landscape perspective". In: (2021). arXiv: 2107.05802 [cs.LG].

[Las08] K. B. Laskey. "MEBN: A language for first-order Bayesian knowledge bases". In: *Artif. Intell.* 172.2 (2008), pp. 140–178.

[Lau92] S. L. Lauritzen. "Propagation of probabilities, means and variances in mixed graphical association models". In: *JASA* 87.420 (1992), pp. 1098–1108.

[Lau95] S. L. Lauritzen. "The EM algorithm for graphical association models with missing data". In: *Computational Statistics and Data Analysis* 19 (1995), pp. 191–201.

[Law05] N. D. Lawrence. "Probabilistic non-linear principal component analysis with Gaussian process latent variable models". In: *JMLR* 6 (2005), pp. 1783–1816.

[Law+22] D. Lawson, A. Raventós, A. Warrington, and S. Linderman. "SIXO: Smoothing Inference with Twisted Objectives". In: (June 2022). arXiv: 2206.05952 [cs.LG].

[LB09] H. Lin and J. A. Bilmes. "How to Select a Good Training-data Subset for Transcription: Submodular Active Selection for Sequences". In: *Proc. Annual Conference of the International Speech Communication Association (INTERSPEECH)*. Brighton, UK, 2009.

[LB10a] H. Lin and J. Bilmes. "Multi-document Summarization via Budgeted Maximization of Submodular Functions". In: *North American chapter of the Association for Computational Linguistics/Human Language Technology Conference (NAACL/HLT-2010)*. Los Angeles, CA, 2010.

[LB10b] H. Lin and J. A. Bilmes. "An Application of the Submodular Principal Partition to Training Data Subset Selection". In: *Neural Information Processing Society (NeurIPS, formerly NIPS) Workshop*. NeurIPS (formerly NIPS) Workshop on Discrete Optimization in Machine Learning: Submodularity, Sparsity & Polyhedra (DISCML). Vancouver, Canada, 2010.

[LB11] H. Lin and J. Bilmes. "A Class of Submodular Functions for Document Summarization". In: *The 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies (ACL/HLT-2011)*. (long paper). Portland, OR, 2011.

[LB12] H. Lin and J. Bilmes. "Learning Mixtures of Submodular Shells with Application to Document Summarization". In: *Uncertainty in Artificial Intelligence (UAI)*. Catalina Island, USA: AUAI, 2012.

[LB19] A. Levray and V. Belle. "Learning Tractable Probabilistic Models in Open Worlds". In: (2019). arXiv: 1901.05847 [cs.LG].

[LBB20] Y. Liu, P.-L. Bacon, and E. Brunskill. "Understanding the Curse of Horizon in Off-Policy Evaluation via Conditional Importance Sampling". In: *ICML*. 2020.

[LBH15] Y. LeCun, Y. Bengio, and G. Hinton. "Deep learning". en. In: *Nature* 521.7553 (May 2015), pp. 436–444.

[LBH22] X. Lu, A. Boukouvalas, and J. Hensman. "Additive Gaussian Processes Revisited". In: *ICML*. June 2022.

[LBL16] H. Lakkaraju, S. H. Bach, and J. Leskovec. "Interpretable decision sets: A joint framework for description and prediction". In: *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*. 2016, pp. 1675–1684.

[LBM06] J. A. Lasserre, C. M. Bishop, and T. P. Minka. "Principled Hybrids of Generative and Discriminative Models". In: *CVPR*. Vol. 1. June 2006, pp. 87–94.

[LBS01] T. Lefebvre, H. Bruyninckx, and J. D. Schutter. "Comment on "A New Method for the Nonlinear Transformation of Means and Covariances in Filters and Estimators"". In: *IEEE Trans. on Automatic Control* 47.8 (2001), pp. 1406–1409.

[LBS17] C. Lakshminarayanan, S. Bhatnagar, and C. Szepesvári. "A Linearly Relaxed Approximate Linear Program for Markov Decision Processes". In: *IEEE Transactions on Automatic Control* 63.4 (2017), pp. 1185–1191.

[LBW17] T. A. Le, A. G. Baydin, and F. Wood. "Inference Compilation and Universal Probabilistic Programming". In: *AISTATS*. 2017.

[LC02] J. Langford and R. Caruana. "(Not) bounding the true error". In: *NIPS*. 2002.

[LCG12] Y. Lou, R. Caruana, and J. Gehrke. "Intelligible models for classification and regression". In: *Proceedings of the 18th ACM SIGKDD international conference on Knowledge discovery and data mining*. 2012, pp. 150–158.

[LCR21] T. Lesort, M. Caccia, and I. Rish. "Understanding Continual Learning Settings with Data Distribution Drift Analysis". In: (2021). arXiv: 2104.01678 [cs.LG].

[LDZ11] Y Li, H Duan, and C. X. Zhai. "Cloudspeller: Spelling correction for search queries by using a unified hidden markov model with web-scale resources". In: *SIGIR*. 2011.

[LDZ12] Y. Li, H. Duan, and C. Zhai. "A Generalized Hidden Markov Model with Discriminative Training for Query Spelling Correction". In: *SIGIR*. 2012, pp. 611–620.

[Le+18] T. A. Le, M. Igl, T. Rainforth, T. Jin, and F. Wood. "Auto-Encoding Sequential Monte Carlo". In: *ICLR*. 2018.

[L'E18] P. L'Ecuyer. "Randomized Quasi-Monte Carlo: An Introduction for Practitioners". In: *Monte Carlo and Quasi-Monte Carlo Methods*. Springer International Publishing, 2018, pp. 29–52.

[Le+19] T. A. Le, A. R. Kosiorek, N Siddharth, Y. W. Teh, and F. Wood. "Revisiting Reweighted Wake-Sleep for Models with Stochastic Control Flow". In: *UAI*. 2019.

[LeC+89] Y. LeCun, B. Boser, J. S. Denker, D. Henderson, R. E. Howard, W. Hubbard, and L. D. Jackel. "Backpropagation Applied to Handwritten Zip Code Recognition". In: *Neural Computation* 1.4 (1989), pp. 541–551.

[LeC+98] Y. LeCun, L. Bottou, G. B. Orr, and K.-R. Müller. "Efficient BackProp". en. In: *Neural Networks: Tricks of the Trade*. Lecture Notes in Computer Science. Springer, Berlin, Heidelberg, 1998, pp. 9–50.

[Lee06] J. de Leeuw. "Principal Component Analysis of Binary Data by Iterated Singular Value Decomposition". In: *Comput. Stat. Data Anal.* 50.1 (2006), pp. 21–39.

[Lee+09] H. Lee, R. B. Grosse, R. Ranganath, and A. Ng. "Convolutional deep belief networks for scalable unsupervised learning of hierarchical representations". In: *ICML '09*. 2009.

[Lee+10] J. Lee, V. Mirrokni, V. Nagarajan, and M. Sviridenko. "Maximizing Nonmonotone Submodular Functions under Matroid or Knapsack Constraints". In: *SIAM Journal on Discrete Mathematics* 23.4 (2010), pp. 2053–2078.

[Lee+18] J. Lee, Y. Bahri, R. Novak, S. S. Schoenholz, J. Pennington, and J. Sohl-Dickstein. "Deep Neural Networks as Gaussian Processes". In: *ICLR*. 2018.

[Lei+18] J. Lei, M. G'Sell, A. Rinaldo, R. J. Tibshirani, and L. Wasserman. "Distribution-Free Predictive Inference For Regression". In: *JASA* (2018).

[Lei+20] F. Leibfried, V. Dutordoir, S. T. John, and N. Durrande. "A Tutorial on Sparse Gaussian Processes and Variational Inference". In: (2020). arXiv: 2012.13962 [cs.LG].

[Lem09] C. Lemieux. *Monte Carlo and Quasi-Monte Carlo Sampling*. Springer, New York, NY, 2009.

[Léo14] C. Léonard. "A survey of the Schrödinger problem and some of its connections with optimal transport". In: *Discrete & Continuous Dynamical Systems* 34.4 (2014), p. 1533.

[Let+15a] B. Letham, C. Rudin, T. H. McCormick, and D. Madigan. "Interpretable classifiers using rules and Bayesian analysis: Building a better stroke prediction model". In: *The Annals of Applied Statistics* 9.3 (2015).

[Let+15b] B. Letham, C. Rudin, T. H. McCormick, and D. Madigan. "Interpretable classifiers using rules and bayesian analysis: Building a better stroke prediction model". In: *The Annals of Applied Statistics* 9.3 (2015), pp. 1350–1371.

[Lev18] S. Levine. "Reinforcement Learning and Control as Probabilistic Inference: Tutorial and Review". In: (2018). arXiv: 1805.00909 [cs.LG].

[Lev+20] S. Levine, A. Kumar, G. Tucker, and J. Fu. *Offline Reinforcement Learning: Tutorial, Review, and Perspectives on Open Problems*. arXiv:2005.01643. 2020.

[LF+21] L. Le Folgoc, V. Baltatzis, S. Desai, A. Devaraj, S. Ellis, O. E. Martinez Manzanera, A. Nair, H. Qiu, J. Schnabel, and B. Glocker. "Is MC Dropout Bayesian?" In: (2021). arXiv: 2110.04286 [cs.LG].

[LFG21] F. Lin, X. Fang, and Z. Gao. "Distributionally Robust Optimization: A review on theory and applications". In: *Numer. Algebra Control Optim.* 12.1 (2021), pp. 159–212.

[LG94] D. D. Lewis and W. A. Gale. "A sequential algorithm for training text classifiers". In: *SIGIR94*. Springer. 1994, pp. 3–12.

[LGMT11] F. Le Gland, V. Monbet, and V.-D. Tran. "Large Sample Asymptotics for the Ensemble Kalman Filter". In: *Oxford Handbook of Nonlinear Filtering*. Ed. by D Crisan And. 2011.

[LHF17] R. M. Levy, A. Haldane, and W. F. Flynn. "Potts Hamiltonian models of protein co-variation, free energy landscapes, and evolutionary fitness". en. In: *Curr. Opin. Struct. Biol.* 43 (2017), pp. 55–62.

[LHLT15] Y. Li, J. M. Hernandez-Lobato, and R. E. Turner. "Stochastic Expectation Propagation". In: *NIPS*. 2015.

[LHR20] A. Lucic, H. Haned, and M. de Rijke. "Why does my model fail? contrastive local explanations for retail forecasting". In: *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*. 2020, pp. 90–98.

[Li+10] L. Li, W. Chu, J. Langford, and R. E. Schapire. "A contextual-bandit approach to personalized news article recommendation". In: *WWW*. 2010.

[Li+16] C. Li, C. Chen, D. Carlson, and L. Carin. "Preconditioned Stochastic Gradient Langevin Dynamics for Deep Neural Networks". In: *AAAI*. 2016.

[Li+17a] A. Li, A. Jabri, A. Joulin, and L. van der Maaten. "Learning visual n-grams from web data". In: *Proceedings of the IEEE International Conference on Computer Vision*. 2017, pp. 4183–4192.

[Li+17b] C.-L. Li, W.-C. Chang, Y. Cheng, Y. Yang, and B. Poczos. "Mmd gan: Towards deeper understanding of moment matching network". In: *Advances in Neural Information Processing Systems*. 2017, pp. 2203–2213.

[Li+17c] L. Li, K. Jamieson, G. De Salvo, A. Rostamizadeh, and A. Talwalkar. "Hyperband: bandit-based configuration evaluation for hyperparameter optimization". In: *ICLR*. 2017.

[Li+17d] O. Li, H. Liu, C. Chen, and C. Rudin. *Deep Learning for Case-Based Reasoning through Prototypes: A Neural Network that Explains Its Predictions*. 2017. arXiv: 1710.04806 [cs.AI].

[Li+17e] T.-C. Li, J.-Y. Su, W. Liu, and J. M. Corchado. "Approximate Gaussian conjugacy: parametric recursive filtering under nonlinearity, multimodality, uncertainty, and constraint, and beyond". In: *Frontiers of Information Technology & Electronic Engineering* 18.12 (2017), pp. 1913–1939.

[Li18] Y. Li. "Deep Reinforcement Learning". In: (2018). arXiv: 1810.06339 [cs.LG].

[Li+18a] C. Li, H. Farkhoor, R. Liu, and J. Yosinski. "Measuring the Intrinsic Dimension of Objective Landscapes". In: *ICLR*. 2018.

[Li+18b] C. Li, H. Farkhoor, R. Liu, and J. Yosinski. "Measuring the Intrinsic Dimension of Objective Landscapes". In: *ICLR*. 2018.

[Li+18c] X. Li, C. Li, J. Chi, J. Ouyang, and W. Wang. "Black-box Expectation Propagation for Bayesian Models". In: *ICDM*. Proceedings. Society for Industrial and Applied Mathematics, 2018, pp. 603–611.

[Li+19] J. Li, S. Qu, X. Li, J. Szurley, J. Z. Kolter, and F. Metze. "Adversarial Music: Real world Audio Adversary against Wake-word Detection System". In: *NIPS*. Curran Associates, Inc., 2019, pp. 11908–11918.

[Li+20] C. Li, X. Gao, Y. Li, B. Peng, X. Li, Y. Zhang, and J. Gao. "Optimus: Organizing Sentences via Pre-trained Modeling of a Latent Space". In: *EMNLP*. 2020.

[Li+21] Y. Li, R. Pogodin, D. J. Sutherland, and A. Gretton. "Self-Supervised Learning with Kernel Dependence Maximization". In: *NeurIPS*. 2021.

[Lia+07] L. Liao, D. J. Patterson, D. Fox, and H. Kautz. "Learning and Inferring Transportation Routines". In: *Artificial Intelligence* 171.5 (2007), pp. 311–331.

[Lia+08] F. Liang, R. Paulo, G. Molina, M. Clyde, and J. Berger. "Mixtures of g-priors for Bayesian Variable Selection". In: *JASA* 103.481 (2008), pp. 410–423.

[Lia+19] V. Liao, R. Ballamy, M. Muller, and H. Candello. "Human-AI Collaboration: Towards Socially-Guided Machine Learning". In: *CHI Workshop on Human-Centered Machine Learning Perspectives*. 2019.

[Lil+16] T. P. Lillicrap, J. J. Hunt, A. Pritzel, N. Heess, T. Erez, Y. Tassa, D. Silver, and D. Wierstra. "Continuous control with deep reinforcement learning". In: *ICLR*. 2016.

[Lin13a] W. Lin. "Agnostic notes on regression adjustments to experimental data: Reexamining Freedman's critique". In: *The Annals of Applied Statistics* 7.1 (2013), pp. 295–318.

[Lin13b] D. A. Linzer. "Dynamic Bayesian Forecasting of Presidential Elections in the States". In: *JASA* 108.501 (2013), pp. 124–134.

[Lin+17a] K. Lin, D. Li, X. He, Z. Zhang, and M.-T. Sun. "Adversarial ranking for language generation". In: *Advances in Neural Information Processing Systems*. 2017, pp. 3155–3165.

[Lin+17b] T.-Y. Lin, P. Dollár, R. Girshick, K. He, B. Hariharan, and S. Belongie. "Feature pyramid networks for object detection". In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2017, pp. 2117–2125.

[Lin+20] H. Lin, H. Chen, T. Zhang, C. Laroche, and K. Choromanski. "Demystifying Orthogonal Monte Carlo and Beyond". In: (2020). arXiv: 2005.13590 [cs.LG].

[Lin+21a] T. Lin, Y. Wang, X. Liu, and X. Qiu. "A Survey of Transformers". In: (2021). arXiv: 2106.04554 [cs.LG].

[Lin+21b] A. Lindholm, N. Wahlström, F. Lindsten, and T. B. Schön. *Machine Learning - A First Course for Engineers and Scientists*. 2021.

[Lin+21c] J. Lindqvist, S. Särkkä, Á. F. García-Fernández, M. Raitoharju, and L. Svensson. "Posterior linearisation smoothing with robust iterations". In: (Dec. 2021). arXiv: 2112.03969 [math.OC].

[Lin+22] Z. Lin, J. Shi, D. Pathak, and D. Ramanan. "The CLEAR Benchmark: Continual LEArning on Real-World Imagery". In: *NIPS Datasets and Benchmarks Track*. Jan. 2022.

[Lin56] D. Lindley. "On a measure of the information provided by an experiment". In: *The Annals of Math. Stat.* (1956), 986–1005.

[Lin88a] B. Lindsay. "Composite Likelihood Methods". In: *Contemporary Mathematics* 80.1 (1988), pp. 221–239.

[Lin88b] R Linsker. "Self-organization in a perceptual network". In: *Computer* 21.3 (1988), pp. 105–117.

[Lin88c] R. Linsker. "Self-organization in a perceptual network". In: *Computer* 21.3 (1988), pp. 105–117.

[Lin92] L.-J. Lin. "Self-Improving Reactive Agents Based on Reinforcement Learning, Planning and Teaching". In: *Mach. Learn.* 8.3-4 (1992), pp. 293–321.

[Lip18] Z. C. Lipton. "The Mythos of Model Interpretability: In machine learning, the concept of interpretability is both important and slippery." In: *Queue* 16.3 (2018), pp. 31–57.

[Liu01] J. Liu. *Monte Carlo Strategies in Scientific Computation*. Springer, 2001.

[Liu+15] Z. Liu, P. Luo, X. Wang, and X. Tang. "Deep Learning Face Attributes in the Wild". In: *ICCV*. 2015.

[Liu+18a] L. Liu, X. Liu, C.-J. Hsieh, and D. Tao. "Stochastic Second-order Methods for Non-convex Optimization with Inexact Hessian and Gradient". In: (2018). arXiv: 1809.09853 [math.OC].

[Liu+18b] Q. Liu, L. Li, Z. Tang, and D. Zhou. "Breaking the Curse of Horizon: Infinite-horizon Off-policy Estimation". In: *NeurIPS*. Curran Associates Inc., 2018, pp. 5361–5371.

[Liu+19a] H. Liu, Y.-S. Ong, Z. Yu, J. Cai, and X. Shen. "Scalable Gaussian Process Classification with Additive Noise for Various Likelihoods". In: (2019). arXiv: 1909.06541 [stat.ML].

[Liu+19b] R. Liu, J. Regier, N. Tripuraneni, M. I. Jordan, and J. McAuliffe. "Rao-Blackwellized Stochastic Gradients for Discrete Distributions". In: *ICML*. 2019.

[Liu+19c] Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer, and V. Stoyanov. "Roberta: A robustly optimized bert pretraining approach". In: *arXiv preprint arXiv:1907.11692* (2019).

[Liu+20a] F. Liu, W. Xu, J. Lu, G. Zhang, A. Gretton, and D. J. Sutherland. "Learning Deep Kernels for Non-Parametric Two-Sample Tests". In: *ICML*. 2020.

[Liu+20b] F. Liu, W. Xu, J. Lu, G. Zhang, A. Gretton, and D. J. Sutherland. "Learning deep kernels for non-parametric two-sample tests". In: *International Conference on Machine Learning*. PMLR. 2020, pp. 6316–6326.

[Liu+20c] H. Liu, Y.-S. Ong, X. Shen, and J. Cai. "When Gaussian Process Meets Big Data: A Review of Scalable GPs". In: *IEEE Transactions on Neural Networks and Learning Systems* 31.1 (2020).

[Liu+20d] J. Z. Liu, Z. Lin, S. Padhy, D. Tran, T. Bedrax-Weiss, and B. Lakshminarayanan. "Simple and Principled Uncertainty Estimation with Deterministic Deep Learning via Distance Awareness". In: *NIPS*. 2020.

[Liu+21a] P. Liu, W. Yuan, J. Fu, Z. Jiang, H. Hayashi, and G. Neubig. "Pre-train, Prompt, and Predict: A Systematic Survey of Prompting Methods in Natural Language Processing". In: (2021). arXiv: 2107.13586 [cs.CL].

[Liu+21b] W. Liu, X. Wang, J. D. Owens, and Y. Li. "Energy-based Out-of-distribution Detection". In: *NIPS*. 2021.

[Liu+22] Z. Liu, H. Mao, C.-Y. Wu, C. Feichtenhofer, T. Darrell, and S. Xie. "A ConvNet for the 2020s". In: (2022). arXiv: 2201.03545 [cs.CV].

[LJ08] P. Liang and M. I. Jordan. "An Asymptotic Analysis of Generative, Discriminative, and Pseudo-likelihood Estimators". In: *International Conference on Machine Learning (ICML)*. 2008.

[Lju87] L. Ljung. *System Identificiation: Theory for the User*. Prentice Hall, 1987.

[LJY19] S. Liu, Y. Jiang, and T. Yu. "Modelling RNA-Seq data with a zero-inflated mixture Poisson linear model". en. In: *Genet. Epidemiol.* 43.7 (2019), pp. 786–799.

[LK07] P. Liang and D. Klein. *Structured Bayesian Nonparametric Models with Variational Inference*. ACL Tutorial. 2007.

[LK09] P. Liang and D. Klein. "Online EM for Unsupervised Models". In: *NAACL*. 2009.

[LKJ09] D. Lewandowski, D. Kurowicka, and H. Joe. "Generating random correlation matrices based on vines and extended onion method". In: *J. Multivar. Anal.* 100.9 (2009), pp. 1989–2001.

[LL02] P. Larranaga and J. A. Lozano. *Estimation of Distribution Algorithms: A New Tool for Evolutionary Computation*. Kluwer Academic Publishers, 2002.

[LL17] S. M. Lundberg and S.-I. Lee. "A unified approach to interpreting model predictions". In: *NIPS*. 2017, pp. 4765–4774.

[LLC20] M. Locher, K. B. Laskey, and P. C. G. Costa. "Design patterns for modeling first-order expressive Bayesian networks". In: *Knowl. Eng. Rev.* 35 (2020).

[LLJ16] Q. Liu, J. Lee, and M. Jordan. "A kernelized Stein discrepancy for goodness-of-fit tests". In: *International conference on machine learning*. 2016, pp. 276–284.

[LLN06] B. Lehmann, D. Lehmann, and N. Nisan. "Combinatorial auctions with decreasing marginal utilities". In: *Games and Economic Behavior* 55.2 (2006), pp. 270–296.

[Llo+14] J. R. Lloyd, D. Duvenaud, R. Grosse, J. B. Tenenbaum, and Z. Ghahramani. "Automatic Construction and Natural-Language Description of Nonparametric Regression Models". In: *AAAI*. 2014.

[LM03] T. S. Lee and D. Mumford. "Hierarchical Bayesian inference in the visual cortex." In: *Journal of the Optical Society of America. A, Optics, image science, and vision* 20 7 (2003), pp. 1434–48.

[LM11] H. Larochelle and I. Murray. "The neural autoregressive distribution estimator". In: *AISTATS*. Vol. 15. 2011, pp. 29–37.

[LM20] M. L. Leavitt and A. Morcos. "Towards falsifiable interpretability research". In: *arXiv preprint arXiv:2010.12016* (2020).

[LMP01] J. Lafferty, A. McCallum, and F. Pereira. "Conditional Random Fields: Probabilistic Models for Segmenting and Labeling Sequence Data". In: *ICML*. 2001.

[LMR15] F. Lavancier, J. Møller, and E. Rubak. "Determinantal point process models and statistical inference". In: *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 77.4 (2015), pp. 853–877.

[LMS16] B. Leimkuhler, C. Matthews, and G. Stoltz. "The computation of averages from equilibrium and nonequilibrium Langevin molecular dynamics". In: *IMA J. Numer. Anal.* 36.1 (2016), pp. 13–79.

[LN01] S. Lauritzen and D. Nilsson. "Representing and solving decision problems with limited information". In: *Management Science* 47 (2001), pp. 1238–1251.

[LN19] H. Lin and V. Ng. "Abstractive summarization: A survey of the state of the art". In: *Proceedings of the AAAI Conference on Artificial Intelligence*. Vol. 33. 01. 2019, pp. 9815–9822.

[LN96] o. Lee and J. A. Nelder. "Hierarchical Generalized Linear Models". In: *J. of Royal Stat. Soc. Series B* 58.4 (1996), pp. 619–678.

[Loc+18] F. Locatello, S. Bauer, M. Lucic, G. Rätsch, S. Gelly, B. Schölkopf, and O. Bachem. "Challenging Common Assumptions in the Unsupervised Learning of Disentangled Representations". In: (2018). arXiv: 1811.12359 [cs.LG].

[Loc+20a] F. Locatello, S. Bauer, M. Lucic, G. Raetsch, S. Gelly, B. Schölkopf, and O. Bachem. "A Sober Look at the Unsupervised Learning of Disentangled Representations and their Evaluation". In: *Journal of Machine Learning Research* 21.209 (2020), pp. 1–62.

[Loc+20b] F. Locatello, B. Poole, G. Rätsch, B. Schölkopf, O. Bachem, and M. Tschannen. "Weakly-supervised disentanglement without compromises". In: *International Conference on Machine Learning*. PMLR. 2020, pp. 6348–6359.

[Lod+02] H. Lodhi, C. Saunders, J. Shawe-Taylor, N. Cristianini, and C. Watkins. "Text classification using string kernels". en. In: *J. Mach. Learn. Res.* (2002).

[Loe04] H Loeliger. "An introduction to factor graphs". In: *IEEE Signal Process. Magazine* 21.1 (2004), pp. 28–41.

[Loe+07] H Loeliger, J Dauwels, J Hu, S Korl, L Ping, and F. R. Kschischang. "The Factor Graph Approach to Model-Based Signal Processing". In: *Proc. IEEE* 95.6 (2007), pp. 1295–1322.

[Loe+16] H.-A. Loeliger, L. Bruderer, H. Malmberg, F. Wadehn, and N. Zalmai. "On Sparsity by NUV-EM, Gaussian Message Passing, and Kalman Smoothing". In: *ITA Workshop*. 2016.

[Loi+21] N. Loizou, S. Vaswani, I. Laradji, and S. Lacoste-Julien. "Stochastic Polyak step-size for SGD:

An adaptive learning rate for fast convergence". In: *AISTATS*. 2021.

[Lon+18] M. Long, Z. Cao, J. Wang, and M. I. Jordan. "Conditional adversarial domain adaptation". In: *Neural Information Processing Systems* (2018).

[Lot+22] S. Lotfi, P. Izmailov, G. Benton, M. Goldblum, and A. G. Wilson. "Bayesian Model Selection, the Marginal Likelihood, and Generalization". In: *ICML*. 2022.

[Lov+20] T. Lovett, M. Briers, M. Charalambides, R. Jersakova, J. Lomax, and C. Holmes. "Inferring proximity from Bluetooth Low Energy RSSI with Unscented Kalman Smoothers". In: (2020). arXiv: 2007.05057 [eess.SP].

[Lov83] L. Lovász. "Submodular functions and convexity". In: *Mathematical programming the state of the art*. Springer, 1983, pp. 235–257.

[LÖW21] T. van de Laar, A. Özçelikkale, and H. Wymeersch. "Application of the Free Energy Principle to Estimation and Control". In: *IEEE Trans. Signal Process.* 69 (2021), pp. 4234–4244.

[LP01] U. Lerner and R. Parr. "Inference in Hybrid Networks: Theoretical Limits and Practical Algorithms". In: *UAI*. 2001.

[LP03] M. G. Lagoudakis and R. Parr. "Least-Squares Policy Iteration". In: *JMLR* 4 (2003), pp. 1107–1149.

[LP06] N. Lartillot and H. Philippe. "Computing Bayes factors using thermodynamic integration". en. In: *Systematic Biology* 55.2 (2006), pp. 195–207.

[LPB17] B. Lakshminarayanan, A. Pritzel, and C. Blundell. "Simple and Scalable Predictive Uncertainty Estimation using Deep Ensembles". In: *NIPS*. 2017.

[LPO17] D. Lopez-Paz and M. Oquab. "Revisiting classifier two-sample tests". In: *International Conference on Learning Representations*. 2017.

[LPR17] D. Lopez-Paz and M. Ranzato. "Gradient Episodic Memory for Continual Learning". In: *NIPS*. 2017.

[LR19] F. Lattimore and D. Rohde. "Replacing the do-calculus with Bayes rule". In: (2019). arXiv: 1906.07125 [stat.ML].

[LR85] T. L. Lai and H. Robbins. "Asymptotically efficient adaptive allocation rules". en. In: *Adv. Appl. Math.* (1985).

[LR87] R. J. Little and D. B. Rubin. *Statistical Analysis with Missing Data*. Wiley and Son, 1987.

[LR95] C. Liu and D. Rubin. "ML Estimation of the T distribution using EM and its extensions, ECM and ECME". In: *Statistica Sinica* 5 (1995), pp. 19–39.

[LRC19] Y. Li, B. I. P. Rubinstein, and T. Cohn. "Truth Inference at Scale: A Bayesian Model for Adjudicating Highly Redundant Crowd Annotations". In: *WWW*. 2019.

[LS01] D. Lee and S. Seung. "Algorithms for non-negative matrix factorization". In: *NIPS*. 2001.

[LS19] T. Lattimore and C. Szepesvari. *Bandit Algorithms*. Cambridge, 2019.

[LS79] P. W. Lewis and G. S. Shedler. "Simulation of nonhomogeneous Poisson processes by thinning". In: *Naval research logistics quarterly* 26.3 (1979), pp. 403–413.

[LS99] D. D. Lee and H. S. Seung. "Learning the parts of objects by non-negative matrix factorization". In: *Nature* 401.6755 (1999), pp. 788–791.

[LST15] B. M. Lake, R. Salakhutdinov, and J. B. Tenenbaum. "Human-level concept learning through probabilistic program induction". In: *Science* 350 (2015), pp. 1332 –1338.

[LST21] N. Loo, S. Swaroop, and R. E. Turner. "Generalized Variational Continual Learning". In: *ICLR*. 2021.

[LST90] J. K. Lenstra, D. B. Shmoys, and É. Tardos. "Approximation algorithms for scheduling unrelated parallel machines". In: *Mathematical programming*. 1990.

[LSV09] J. Lee, M. Sviridenko, and J. Vondrák. "Submodular maximization over multiple matroids via generalized exchange properties". In: *Approximation, Randomization, and Combinatorial Optimization. Algorithms and Techniques* (2009), pp. 244–257.

[LSW15] Y. T. Lee, A. Sidford, and S. C.-w. Wong. "A faster cutting plane method and its implications for combinatorial and convex optimization". In: *2015 IEEE 56th Annual Symposium on Foundations of Computer Science*. IEEE. 2015, pp. 1049–1065.

[LSZ15] Y. Li, K. Swersky, and R. Zemel. "Generative Moment Matching Networks". In: *ICML*. 2015.

[LT16] M.-Y. Liu and O. Tuzel. "Coupled Generative Adversarial Networks". In: *NIPS*. 2016, pp. 469–477.

[LTW15] Q. Li, C. Tai, and E Weinan. "Stochastic modified equations and adaptive stochastic gradient algorithms". In: *ICML*. 2015.

[Lu+21a] X. Lu, I. Osband, B. Van Roy, and Z. Wen. "Evaluating Probabilistic Inference in Deep Learning: Beyond Marginal Predictions". In: (2021). arXiv: 2107.09224 [cs.LG].

[Lu+21b] X. Lu, B. Van Roy, V. Dwaracherla, M. Ibrahimi, I. Osband, and Z. Wen. "Reinforcement Learning, Bit by Bit". In: (Mar. 2021). arXiv: 2103.04047 [cs.LG].

[Lu+22] X. Lu, I. Osband, B. Van Roy, and Z. Wen. "From Predictions to Decisions: The Importance of Joint Predictive Distributions". In: (2022). arXiv: 2107.09224 [cs.LG].

[Luc+18] A Lucas, M Iliadis, R Molina, and A. K. Katsaggelos. "Using Deep Neural Networks for Inverse Problems in Imaging: Beyond Analytical Methods". In: *IEEE Signal Process. Mag.* 35.1 (2018), pp. 20–36.

[Luc+19] J. Lucas, G. Tucker, R. Grosse, and M. Norouzi. "Don't blame the ELBO! A linear VAE perspective on posterior collapse". In: *NIPS*. 2019.

[Luh58] H. P. Luhn. "The automatic creation of literature abstracts". In: *IBM Journal of research and development* 2.2 (1958), pp. 159–165.

[Lun+20] S. M. Lundberg, G. Erion, H. Chen, A. DeGrave, J. M. Prutkin, B. Nair, R. Katz, J. Himmelfarb, N. Bansal, and S.-I. Lee. "From local explanations to global understanding with explainable AI for trees". In: *Nature machine intelligence* 2.1 (2020), pp. 56–67.

[Luo+19] Y. Luo, H. Xu, Y. Li, Y. Tian, T. Darrell, and T. Ma. "Algorithmic Framework for Model-based Deep Reinforcement Learning with Theoretical Guarantees". In: *ICLR*. 2019.

[Lut16] J Luttinen. "BayesPy: variational Bayesian inference in Python". In: *JMLR* (2016).

[LV06] F. Liese and I. Vajda. "On divergences and informations in statistics and information theory". In: *IEEE Transactions on Information Theory* 52.10 (2006), pp. 4394–4412.

[LW04] H. Lopes and M. West. "Bayesian model assessment in factor analysis". In: *Statisica Sinica* 14 (2004), pp. 41–67.

[LW16] C. Louizos and M. Welling. "Structured and Efficient Variational Deep Learning with Matrix Gaussian Posteriors". In: *ICML*. 2016.

[LW17] C. Louizos and M. Welling. "Multiplicative Normalizing Flows for Variational Bayesian Neural Networks". In: *Proceedings of the 34th International Conference on Machine Learning*. 2017, pp. 2218–2227.

[LWS18] Z. C. Lipton, Y.-X. Wang, and A. Smola. "Detecting and Correcting for Label Shift with Black Box Predictors". In: *ICML*. 2018.

[LY17] J. H. Lim and J. C. Ye. "Geometric gan". In: *arXiv preprint arXiv:1705.02894* (2017).

[Lyu11] S. Lyu. "Unifying non-maximum likelihood learning objectives with minimum KL contraction". In: *Advances in Neural Information Processing Systems*. 2011, pp. 64–72.

[Lyu12] S. Lyu. "Interpretation and generalization of score matching". In: *arXiv preprint arXiv:1205.2629* (2012).

[LZ20] B. Lim and S. Zohren. "Time Series Forecasting With Deep Learning: A Survey". In: (2020). arXiv: 2004.13408 [stat.ML].

[MA10] I. Murray and R. P. Adams. "Slice sampling covariance hyperparameters of latent Gaussian models". In: *NIPS*. 2010, pp. 1732–1740.

[Maa+16] L. Maaløe, C. K. Sønderby, S. K. Sønderby, and O. Winther. "Auxiliary Deep Generative Models". In: *ICML*. 2016.

[Maa+19] L. Maaløe, M. Fraccaro, V. Liévin, and O. Winther. "BIVA: A Very Deep Hierarchy of Latent Variables for Generative Modeling". In: *NIPS*. 2019.

[Mac03] D. MacKay. *Information Theory, Inference, and Learning Algorithms*. Cambridge University Press, 2003.

[Mac+11] J. H. Macke, L. Büsing, J. P. Cunningham, B. M. Y. Ece, K. V. Shenoy, and M. Sahani. "Empirical models of spiking in neural populations". In: *NIPS*. 2011.

[Mac+15] D. Maclaurin, D. Duvenaud, M. Johnson, and R. P. Adams. *Autograd: Reverse-mode differentiation of native Python*. 2015.

[Mac+19] D. Maclaurin, A. Radul, M. J. Johnson, and D. Vytiniotis. "Dex: array programming with typed indices". In: *NeurIPS workshop: Program Transformations for Machine Learning* (2019).

[Mac75] O. Macchi. "The coincidence approach to stochastic point processes". In: *Advances in Applied Probability* 7.1 (1975), pp. 83–122.

[Mac92a] D. MacKay. "The evidence framework applied to classification networks". In: *Neural Comptutation* 4.5 (1992), pp. 720–736.

[Mac92b] D. J. C. MacKay. "A Practical Bayesian Framework for Backpropagation Networks". In: *Neural Comput.* 4.3 (1992), pp. 448–472.

[Mac95] D. MacKay. "Probable networks and plausible predictions — a review of practical Bayesian methods for supervised neural networks". In: *Network: Computation in Neural Systems* 6.3 (1995), pp. 469–505.

[Mac98] D. MacKay. "Introduction to Gaussian Processes". In: *Neural Networks and Machine Learning*. Ed. by C. Bishop. 1998.

[Mac99a] S. N. MacEachern. "Dependent nonparametric processes". In: *ASA proceedings of the section on Bayesian statistical science*. Vol. 1. Alexandria, Virginia. Virginia: American Statistical Association; 1999. 1999, pp. 50–55.

[Mac99b] D. MacKay. "Comparision of approximate methods for handling hyperparameters". In: *Neural Computation* 11.5 (1999), pp. 1035–1068.

[Mad+17] C. J. Maddison, D. Lawson, G. Tucker, N. Heess, M. Norouzi, A. Mnih, A. Doucet, and Y. W. Teh. "Filtering Variational Objectives". In: *NIPS*. 2017.

[Mad+18] A. Madry, A. Makelov, L. Schmidt, D. Tsipras, and A. Vladu. "Towards Deep Learning Models Resistant to Adversarial Attacks". In: *ICLR*. 2018.

[Mad+19] W. J. Maddox, P. Izmailov, T. Garipov, D. P. Vetrov, and A. G. Wilson. "A Simple Baseline for Bayesian Uncertainty in Deep Learning". In: *NIPS*. Curran Associates, Inc., 2019, pp. 13153–13164.

[MAD20] W. R. Morningstar, A. A. Alemi, and J. V. Dillon. "PAC^m-Bayes: Narrowing the Empirical Risk Gap in the Misspecified Bayesian Regime". In: (2020). arXiv: 2010.09629 [cs.LG].

[Mah07] R. P. S. Mahler. *Statistical Multisource-Multitarget Information Fusion*. Artech House, Inc., 2007.

[Mah13] R Mahler. "Statistics 102 for Multisource-Multitarget Detection and Tracking". In: *IEEE J. Sel. Top. Signal Process.* 7.3 (2013), pp. 376–389.

[Mah+18] D. Mahajan, R. Girshick, V. Ramanathan, K. He, M. Paluri, Y. Li, A. Bharambe, and L. van der Maaten. "Exploring the limits of weakly supervised pretraining". In: *Proceedings of the European Conference on Computer Vision (ECCV)*. 2018, pp. 181–196.

[Mah+19] N. Maheswaranathan, A. H. Williams, M. D. Golub, S. Ganguli, and D. Sussillo. "Universality and individuality in neural dynamics across large populations of recurrent networks". In: *Advances in Neural Information Processing Systems* 2019 (2019), p. 15629.

[Mai+10] J. Mairal, F. Bach, J. Ponce, and G. Sapiro. "Online learning for matrix factorization and sparse coding". In: *JMLR* 11 (2010), pp. 19–60.

[Mai13] J. Mairal. "Stochastic Majorization-minimization Algorithms for Large-scale Optimization". In: *NIPS*. 2013, pp. 2283–2291.

[Mai15] J Mairal. "Incremental Majorization-Minimization Optimization with Application to Large-Scale Machine Learning". In: *SIAM J. Optim.* 25.2 (2015), pp. 829–855.

[Mai+22] Z. Mai, R. Li, J. Jeong, D. Quispe, H. Kim, and S. Sanner. "Online continual learning in image classification: An empirical survey". In: *Neurocomputing* 469 (2022), pp. 28–51.

[Mak+15a] A. Makhzani, J. Shlens, N. Jaitly, I. Goodfellow, and B. Frey. "Adversarial Autoencoders". In: (2015). arXiv: 1511.05644 [cs.LG].

[Mak+15b] A. Makhzani, J. Shlens, N. Jaitly, I. Goodfellow, and B. Frey. "Adversarial autoencoders". In: *arXiv preprint arXiv:1511.05644* (2015).

[Mak+20] A. Makkuva, A. Taghvaei, S. Oh, and J. Lee. "Optimal transport mapping via input convex neural networks". In: *International Conference on Machine Learning*. PMLR. 2020, pp. 6672–6681.

[Mal+17] D. M. Malioutov, K. R. Varshney, A. Emad, and S. Dash. "Learning interpretable classification rules with boolean compressed sensing". In: *Transparent Data Mining for Big and Small Data*. Springer, 2017, pp. 95–121.

[Man] B. Mann. *How many times should you shuffle a deck of cards?* Tech. rep. Dartmouth.

[Man+07] V. Mansinghka, D. Roy, R. Rifkin, and J. Tenenbaum. "AClass: An online algorithm for generative classification". In: *AISTATS*. 2007.

[Man+19] D. J. Mankowitz, N. Levine, R. Jeong, Y. Shi, J. Kay, A. Abdolmaleki, J. T. Springenberg, T. Mann, T. Hester, and M. Riedmiller. "Robust Reinforcement Learning for Continuous Control with Model Misspecification". In: (2019). arXiv: 1906.07516 [cs.LG].

[Man22] V. Manokhin. "Machine Learning for Probabilistic Prediction". PhD thesis. Royal Holloway, University of London, June 2022.

[Man90] C. F. Manski. "Nonparametric Bounds on Treatment Effects". In: *The American Economic Review* 80.2 (1990), pp. 319–323.

[Mao+17] X. Mao, Q. Li, H. Xie, R. Y. K. Lau, Z. Wang, and S. P. Smolley. "Least Squares Generative Adversarial Networks". In: *ICCV*. 2017.

[Mar03] B. Marlin. "Modeling User Rating Profiles for Collaborative Filtering". In: *NIPS*. 2003.

[Mar+06] A. Margolin, I. Nemenman, K. Basso, C. Wiggins, G. Stolovitzky, and R. F. abd A. Califano. "ARACNE: An Algorithm for the Reconstruction of Gene Regulatory Networks in a Mammalian Cellular Context". In: *BMC Bionformatics* 7 (2006).

[Mar08] B. Marlin. "Missing Data Problems in Machine Learning". PhD thesis. U. Toronto, 2008.

[Mar+10] B. M. Marlin, K. Swersky, B. Chen, and N. de Freitas. "Inductive Principles for Restricted Boltzmann Machine Learning". In: *AISTATS*. 2010.

[Mar10a] J Martens. "Deep learning via Hessian-free optimization". In: *ICML*. 2010.

[Mar10b] J. Martens. "Learning the Linear Dynamical System with ASOS". In: *ICML*. ICML'10. Omnipress, 2010, pp. 743–750.

[Mar16] J. Martens. "Second-order optimization for neural networks". PhD thesis. Toronto, 2016.

[Mar18] O. Martin. *Bayesian analysis with Python*. Packt, 2018.

[Mar20] J. Martens. "New insights and perspectives on the natural gradient method". In: *JMLR* (2020).

[Mas+20] M. Masana, X. Liu, B. Twardowski, M. Menta, A. D. Bagdanov, and J. van de Weijer. "Class-incremental learning: survey and performance evaluation on image classification". In: (2020). arXiv: 2010.15277 [cs.LG].

[Mat14] C. Mattfeld. "Implementing spectral methods for hidden Markov models with real-valued emissions". MA thesis. ETH Zurich, 2014.

[Mat+16] A. Matthews, J. Hensman, R. Turner, and Z. Ghahramani. "On Sparse Variational Methods and the Kullback-Leibler Divergence between Stochastic Processes". en. In: *AISTATS*. 2016, pp. 231–239.

[Mav16] Mavrogonatou, L And Vyshemirsky, "Sequential Importance Sampling for Online Bayesian Changepoint Detection". In: *22nd International Conference on Computational Statistics*. 2016.

[May+19] A. May, J. Zhang, T. Dao, and C. Ré. "On the downstream performance of compressed word embeddings". In: *Advances in neural information processing systems* 32 (2019), p. 11782.

[May79] P. Maybeck. *Stochastic models, estimation, and control*. Academic Press, 1979.

[Maz+22] P. Mazzaglia, T. Verbelen, O. Çatal, and B. Dhoedt. "The Free Energy Principle for Perception and Action: A Deep Learning Perspective". en. In: *Entropy* 24.2 (2022).

[MAZA18] X. B. P. amd Marcin Andrychowicz, W. Zaremba, and P. Abbeel. "Sim-to-Real Transfer of Robotic Control with Dynamics Randomization". In: *ICRA*. 2018, pp. 1–8.

[MB16] Y. Miao and P. Blunsom. "Language as a Latent Variable: Discrete Generative Models for Sentence Compression". In: *EMNLP*. 2016.

[MB18] A. Mensch and M. Blondel. "Differentiable Dynamic Programming for Structured Prediction and Attention". In: *ICML*. 2018.

[MB21] M. Y. Michelis and Q. Becker. "On Linear Interpolation in the Latent Space of Deep Generative Models". In: *ICLR Workshop on Geometrical and Topological Representation Learning*. 2021.

[MB88] T. Mitchell and J. Beauchamp. "Bayesian Variable Selection in Linear Regression". In: *JASA* 83 (1988), pp. 1023–1036.

[MBJ06] J. McAuliffe, D. Blei, and M. Jordan. "Nonparametric empirical Bayes for the Dirichlet process mixture model". In: *Statistics and Computing* 16.1 (2006), pp. 5–14.

[MBJ20] T. M. Moerland, J. Broekens, and C. M. Jonker. "Model-based Reinforcement Learning: A Survey". In: (2020). arXiv: 2006.16712 [cs.LG].

[MBK20] X. Meng, R. Bachmann, and M. E. Khan. "Training Binary Neural Networks using the Bayesian Learning Rule". In: *ICML*. 2020.

[MBL20] B. Mirzasoleiman, J. Bilmes, and J. Leskovec. "Coresets for data-efficient training of machine learning models". In: *International Conference on Machine Learning*. PMLR. 2020, pp. 6950–6960.

[MBW20] W. J. Maddox, G. Benton, and A. G. Wilson. "Rethinking Parameter Counting in Deep Models: Effective Dimensionality Revisited". In: *arXiv preprint arXiv:2003.02139* (2020).

[MC03] P. Moscato and C. Cotta. "A Gentle Introduction to Memetic Algorithms". en. In: *Handbook of Metaheuristics*. International Series in Operations Research & Management Science. Springer, Boston, MA, 2003, pp. 105–144.

[MC19] P. Moreno Comellas. "Vision as inverse graphics for detailed scene understanding". en. PhD thesis. 2019.

[McA99] D. A. McAllester. "PAC-Bayesian model averaging". In: *Proceedings of the twelfth annual conference on computational learning theory*. 1999.

[McC03] A. McCray. "An upper level ontology for the biomedical domain". In: *Comparative and Functional Genomics* 4 (2003), pp. 80–84.

[McE20] R. McElreath. *Statistical Rethinking: A Bayesian Course with Examples in R and Stan (2nd edition)*. en. Chapman and Hall/CRC, 2020.

[McG54] W. McGill. "Multivariate information transmission". In: *Psychometrika* 19 (1954), pp. 97–116.

[McM+13] H. B. McMahan, G. Holt, D Sculley, M. Young, D. Ebner, J. Grady, L. Nie, T. Phillips, E. Davydov, D. Golovin, et al. "Ad click prediction: a view from the trenches". In: *KDD*. 2013, pp. 1222–1230.

[Md+19] C. de Masson d'Autume, M. Rosca, J. Rae, and S. Mohamed. "Training language GANs from Scratch". In: (2019). arXiv: 1905.09922 [cs.CL].

[MD97] X. L. Meng and D. van Dyk. "The EM algorithm — an old folk song sung to a fast new tune (with Discussion)". In: *J. Royal Stat. Soc. B* 59 (1997), pp. 511–567.

[MDA15] D. Maclaurin, D. Duvenaud, and R. P. Adams. "Gradient-based Hyperparameter Optimization through Reversible Learning". In: *ICML*. 2015.

[MDM19] S. Mahloujifar, D. I. Diochnos, and M. Mahmoody. "The curse of concentration in robust learning: Evasion and poisoning attacks from concentration of measure". In: *Proceedings of the AAAI Conference on Artificial Intelligence*. Vol. 33. 2019, pp. 4536–4543.

[MDR94] S. Muggleton and L. De Raedt. "Inductive logic programming: Theory and methods". In: *The Journal of Logic Programming* 19 (1994), pp. 629–679.

[ME17] H. Mei and J. M. Eisner. "The neural hawkes process: A neurally self-modulating multivariate point process". In: *Advances in Neural Information Processing Systems*. 2017, pp. 6754–6764.

[Med+21] M. A. Medina, J. L. M. Olea, C. Rush, and A. Velez. "On the Robustness to Misspecification of α-Posteriors and Their Variational Approximations". In: (2021). arXiv: 2104.08324 [stat.ML].

[Mee+18] J.-W. van de Meent, B. Paige, H. Yang, and F. Wood. *An introduction to probabilistic programming*. Foundations and Trends in Machine Learning, 2018.

[Mei18a] N. Meinshausen. "Causality from a distributional robustness point of view". In: *IEEE Data Science Workshop (DSW)*. 2018, pp. 6–10.

[Mei18b] N. Meinshausen. "CAUSALITY FROM A DISTRIBUTIONAL ROBUSTNESS POINT OF VIEW". In: *2018 IEEE Data Science Workshop (DSW)*. 2018, pp. 6–10.

[Mer] *Definition of interpret*. 2022. URL: https://www.merriam-webster.com/dictionary/interpret.

[Mer+00] R. van der Merwe, A. Doucet, N. de Freitas, and E. Wan. "The Unscented Particle Filter". In: *NIPS-13*. 2000.

[Met16] C. Metz. *In Two Moves, AlphaGo and Lee Sedol Redefined the Future*. 2016. URL: https://www.wired.com/2016/03/two-moves-alphago-lee-sedol-redefined-future/ (visited on 01/07/2022).

[Met+16] L. Metz, B. Poole, D. Pfau, and J. Sohl-Dickstein. "Unrolled Generative Adversarial Networks". In: (2016).

[Met+17] L. Metz, J. Ibarz, N. Jaitly, and J. Davidson. "Discrete Sequential Prediction of Continuous Actions for Deep RL". In: (2017). arXiv: 1705.05035 [cs.LG].

[Met+53] N. Metropolis, A. Rosenbluth, M. Rosenbluth, A. Teller, and E. Teller. "Equation of state calculations by fast computing machines". In: *J. of Chemical Physics* 21 (1953), pp. 1087–1092.

[Mey+18] F. Meyer, T. Kropfreiter, J. Williams, R. Lau, F. Hlawatsch, P. Braca, and M. Win. "Message passing algorithms for scalable multitarget tracking". In: *Proc. IEEE* 106.2 (2018).

[Mey+21] R. A. Meyer, C. Musco, C. Musco, and D. P. Woodruff. "Hutch++: Optimal Stochastic Trace Estimation". In: *SIAM Symposium on Simplicity in Algorithms (SOSA21)*. 2021.

[Mey22] S. Meyn. *Control Systems and Reinforcement Learning*. Cambridge, 2022.

[MFP00] A. McCallum, D. Freitag, and F. Pereira. "Maximum Entropy Markov Models for Information Extraction and Segmentation". In: *ICML*. 2000.

[MFR20] G. M. Martin, D. T. Frazier, and C. P. Robert. "Computing Bayes: Bayesian Computation from 1763 to the 21st Century". In: (2020). arXiv: 2004.06425 [stat.CO].

[MG15] J. Martens and R. Grosse. "Optimizing Neural Networks with Kronecker-factored Approximate Curvature". In: *ICML*. 2015.

[MGM06] I. Murray, Z. Ghahramani, and D. J. C. MacKay. "MCMC for doubly-intractable distributions". In: *Proceedings of the 22nd Annual Conference on Uncertainty in Artificial Intelligence (UAI-06)*. AUAI Press, 2006, pp. 359–366.

[MGN18a] L. Mescheder, A. Geiger, and S. Nowozin. "Which Training Methods for GANs do actually Converge?" In: *ICML*. 2018.

[MGN18b] L. Mescheder, A. Geiger, and S. Nowozin. "Which training methods for GANs do actually converge?" In: *International conference on machine learning*. PMLR. 2018, pp. 3481–3490.

[MGR18] H. Mania, A. Guy, and B. Recht. "Simple random search of static linear policies is competitive for reinforcement learning". In: *NIPS*. Ed. by S Bengio, H Wallach, H Larochelle, K Grauman, N Cesa-Bianchi, and R Garnett. Curran Associates, Inc., 2018, pp. 1800–1809.

[MH12] R. Mazumder and T. Hastie. *The Graphical Lasso: New Insights and Alternatives*. Tech. rep. Stanford Dept. Statistics, 2012.

[MH14] J. W. Miller and M. T. Harrison. "Inconsistency of Pitman-Yor process mixtures for the number of components". In: *JMLR* 15.1 (2014), pp. 3333–3370.

[MH20] I. Mordatch and J. Hamrick. *ICML tutorial on model-based methods in reinforcement learning*. https://sites.google.com/corp/view/mbrl-tutorial. 2020.

[MHB17] S. Mandt, M. D. Hoffman, and D. M. Blei. "Stochastic Gradient Descent As Approximate Bayesian Inference". In: *JMLR* 18.1 (2017), pp. 4873–4907.

[MHH14] F. Meyer, O. Hlinka, and F. Hlawatsch. "Sigma point belief propagation". In: *IEEE Signal Processing Letters* 21.2 (2014), pp. 145–149.

[MHN13] A. L. Maas, A. Y. Hannun, and A. Y. Ng. "Rectifier Nonlinearities Improve Neural Network Acoustic Models". In: *ICML*. Vol. 28. 2013.

[Mik+13] T. Mikolov, I. Sutskever, K. Chen, G. S. Corrado, and J. Dean. "Distributed representations of words and phrases and their compositionality". In: *NIPS*. 2013, pp. 3111–3119.

[Mil+05] B. Milch, B. Marthi, S. Russell, D. Sontag, D. Ong, and A. Kolobov. "BLOG: Probabilistic Models with Unknown Objects". In: *IJCAI*. 2005.

[Mil19] T. Miller. "Explanation in artificial intelligence: Insights from the social sciences". In: *Artificial intelligence* 267 (2019), pp. 1–38.

[Mil+20] B. Millidge, A. Tschantz, A. K. Seth, and C. L. Buckley. "On the Relationship Between Active Inference and Control as Inference". In: *International Workshop on Active Inference*. 2020.

[Mil+21] J. P. Miller, R. Taori, A. Raghunathan, S. Sagawa, P. W. Koh, V. Shankar, P. Liang, Y. Carmon, and L. Schmidt. "Accuracy on the Line: on the Strong Correlation Between Out-of-Distribution and In-Distribution Generalization". In: *ICML*. Ed. by M. Meila and T. Zhang. Vol. 139. Proceedings of Machine Learning Research. PMLR, 2021, pp. 7721–7735.

[Min00a] T. Minka. *Bayesian linear regression*. Tech. rep. MIT, 2000.

[Min00b] T. Minka. *Bayesian model averaging is not model combination*. Tech. rep. MIT Media Lab, 2000.

[Min00c] T. Minka. *Estimating a Dirichlet distribution*. Tech. rep. MIT, 2000.

[Min01a] T. Minka. "A family of algorithms for approximate Bayesian inference". PhD thesis. MIT, 2001.

[Min01b] T. Minka. "Expectation Propagation for approximate Bayesian inference". In: *UAI*. 2001.

[Min04] T. Minka. *Power EP*. Tech. rep. MSR-TR-2004-149. 2004.

[Min05] T. Minka. *Divergence measures and message passing*. Tech. rep. MSR Cambridge, 2005.

[Min+18] T. Minka, J. Winn, J. Guiver, Y. Zaykov, D. Fabian, and J. Bronskill. *Infer.NET 0.3*. Microsoft Research Cambridge. 2018.

[Min78] M. Minoux. "Accelerated greedy algorithms for maximizing submodular set functions". In: *Optimization Techniques*. Ed. by J. Stoer. Vol. 7. Lecture Notes in Control and Information Sciences. 10.1007/BFb0006528. Springer Berlin / Heidelberg, 1978, pp. 234–243.

[Mis+18] A. Mishkin, F. Kunstner, D. Nielsen, M. Schmidt, and M. E. Khan. "SLANG: Fast Structured Covariance Approximations for Bayesian Deep Learning with Natural Gradient". In: *NIPS*. Curran Associates, Inc., 2018, pp. 6245–6255.

[Mit+19] M. Mitchell, S. Wu, A. Zaldivar, P. Barnes, L. Vasserman, B. Hutchinson, E. Spitzer, I. D. Raji, and T. Gebru. "Model cards for model reporting". In: *Proceedings of the conference on fairness, accountability, and transparency*. 2019, pp. 220–229.

[Mit+20] J. Mitrovic, B. McWilliams, J. Walker, L. Buesing, and C. Blundell. *Representation Learning via Invariant Causal Mechanisms*. 2020. arXiv: 2010.07922 [cs.LG].

[Mit97] T. Mitchell. *Machine Learning*. McGraw Hill, 1997.

[Miy+18a] T. Miyato, T. Kataoka, M. Koyama, and Y. Yoshida. "Spectral Normalization for Generative Adversarial Networks". In: *ICLR*. 2018.

[Miy+18b] T. Miyato, T. Kataoka, M. Koyama, and Y. Yoshida. "Spectral Normalization for Generative Adversarial Networks". In: *ICLR*. 2018.

[Miy+18c] T. Miyato, T. Kataoka, M. Koyama, and Y. Yoshida. "Spectral Normalization for Generative Adversarial Networks". In: *International Conference on Learning Representations*. 2018.

[MJ97] M. Meila and M. Jordan. *Triangulation by continuous embedding*. Tech. rep. 1605. MIT AI Lab, 1997.

[MK05] J. Mooij and H. Kappen. "Sufficient conditions for convergence of loopy belief propagation". In: *UAI*. 2005.

[MK07] G. J. McLachlan and T. Krishnan. *The EM Algorithm and Extensions (Second Edition)*. Wiley, 2007.

[MK18] T. Miyato and M. Koyama. "cGANs with Projection Discriminator". In: *International Conference on Learning Representations*. 2018.

[MK19] J. Menick and N. Kalchbrenner. "Generating high fidelity images with subscale pixel networks and multidimensional upscaling". In: *ICLR*. 2019.

[MK97] G. J. McLachlan and T. Krishnan. *The EM Algorithm and Extensions*. Wiley, 1997.

[MKG21] W. J. Ma, K. Kording, and D. Goldreich. *Bayesian models of perception and action*. MIT Press, 2021.

[MKH19] R. Müller, S. Kornblith, and G. E. Hinton. "When does label smoothing help?" In: *NIPS*. 2019, pp. 4694–4703.

[MKL11] O. Martin, R. Kumar, and J. Lao. *Bayesian Modeling and Computation in Python*. CRC Press, 2011.

[MKL21] O. A. Martin, R. Kumar, and J. Lao. *Bayesian Modeling and Computation in Python*. CRC Press, 2021.

[MKS21] K. Murphy, A. Kumar, and S. Serghiou. "Risk score learning for COVID-19 contact tracing apps". In: *Machine Learning for Healthcare*. 2021.

[ML02] T. Minka and J. Lafferty. "Expectation-propagation for the Generative Aspect Model". In: *UAI*. Morgan Kaufmann Publishers Inc., 2002, pp. 352–359.

[ML16] S. Mohamed and B. Lakshminarayanan. "Learning in Implicit Generative Models". In: (2016). arXiv: 1610.03483 [stat.ML].

[MLN19] P. Michel, O. Levy, and G. Neubig. "Are Sixteen Heads Really Better than One?" In: *NIPS*. 2019.

[MLW19] V. Masrani, T. A. Le, and F. Wood. "The Thermodynamic Variational Objective". In: *NIPS*. Curran Associates, Inc., 2019, pp. 11521–11530.

[MM01] T. K. Marks and J. R. Movellan. *Diffusion networks, products of experts, and factor analysis*. Tech. rep. University of California San Diego, 2001.

[MM90] D. Q. Mayne and H Michalska. "Receding horizon control of nonlinear systems". In: *IEEE Trans. Automat. Contr.* 35.7 (1990), pp. 814–824.

[MMC98] R. J. McEliece, D. J. C. MacKay, and J. F. Cheng. "Turbo decoding as an instance of Pearl's 'belief propagation' algorithm". In: *IEEE J. on Selected Areas in Comm.* 16.2 (1998), pp. 140–152.

[MMP13] L. Malagò, M. Matteucci, and G. Pistoni. "Natural gradient, fitness modelling and model selection: A unifying perspective". In: *IEEE Congress on Evolutionary Computation*. 2013.

[MMP87] J. Marroquin, S. Mitter, and T. Poggio. "Probabilistic solution of ill-posed problems in computational vision". In: *JASA* 82.297 (1987), pp. 76–89.

[MMT17] C. J. Maddison, A. Mnih, and Y. W. Teh. "The Concrete Distribution: A Continuous Relaxation of Discrete Random Variables". In: *ICLR*. 2017.

[MN89] P. McCullagh and J. Nelder. *Generalized linear models*. 2nd edition. Chapman and Hall, 1989.

[MNG17a] L. Mescheder, S. Nowozin, and A. Geiger. "Adversarial variational bayes: Unifying variational autoencoders and generative adversarial networks". In: *International Conference on Machine Learning*. PMLR. 2017, pp. 2391–2400.

[MNG17b] L. Mescheder, S. Nowozin, and A. Geiger. "The numerics of gans". In: *Advances in Neural Information Processing Systems*. 2017, pp. 1825–1835.

[Mni+15] V. Mnih et al. "Human-level control through deep reinforcement learning". In: *Nature* 518.7540 (2015), pp. 529–533.

[Mni+16] V. Mnih, A. P. Badia, M. Mirza, A. Graves, T. P. Lillicrap, T. Harley, D. Silver, and K. Kavukcuoglu. "Asynchronous Methods for Deep Reinforcement Learning". In: *ICML*. 2016.

[MO14] M. Mirza and S. Osindero. "Conditional generative adversarial nets". In: *arXiv preprint arXiv:1411.1784* (2014).

[Mob16] H. Mobahi. "Closed Form for Some Gaussian Convolutions". In: (2016). arXiv: 1602.05610 [math.CA].

[Moc+96] J. Mockus, W. Eddy, A. Mockus, L. Mockus, and G. Reklaitis. *Bayesian Heuristic Approach to Discrete and Global Optimization: Algorithms, Visualization, Software, and Applications*. Kluwer, 1996.

[Moh+19] S. Mohamed, M. Rosca, M. Figurnov, and A. Mnih. "Monte Carlo Gradient Estimation in Machine Learning". In: (2019). arXiv: 1906.10652 [stat.ML].

[Mon81] G. Monge. "Mémoire sur la théorie des déblais et des remblais". In: *Histoire de l'Académie Royale des Sciences* (1781), pp. 666–704.

[Mor+11] F. Morcos, A. Pagnani, B. Lunt, A. Bertolino, D. S. Marks, C. Sander, R. Zecchina, J. N. Onuchic, T. Hwa, and M. Weigt. "Direct-coupling analysis of residue coevolution captures native contacts across many protein families". en. In: *Proc. Natl. Acad. Sci. U. S. A.* 108.49 (2011), E1293–301.

[Mor+21a] W. Morningstar, C. Ham, A. Gallagher, B. Lakshminarayanan, A. Alemi, and J. Dillon. "Density of States Estimation for Out of Distribution Detection". In: *AISTATS*. Ed. by A. Banerjee and K. Fukumizu. Vol. 130. Proceedings of Machine Learning Research. PMLR, 2021, pp. 3232–3240.

[Mor+21b] W. Morningstar, S. Vikram, C. Ham, A. Gallagher, and J. Dillon. "Automatic Differentiation Variational Inference with Mixtures". In: *AISTATS*. Ed. by A. Banerjee and K. Fukumizu. Vol. 130. Proceedings of Machine Learning Research. PMLR, 2021, pp. 3250–3258.

[Mor63] T. Morimoto. "Markov Processes and the H-Theorem". In: *J. Phys. Soc. Jpn.* 18.3 (1963), pp. 328–331.

[MOT15] A. Mordvintsev, C. Olah, and M. Tyka. *Inceptionism: Going Deeper into Neural Networks*. https://ai.googleblog.com/2015/06/inceptionism-going-deeper-into-neural.html. Accessed: NA-NA-NA. 2015.

[Mov08] J. R. Movellan. "A minimum velocity approach to learning". In: *unpublished draft, Jan* (2008).

[MP01] K. Murphy and M. Paskin. "Linear time inference in hierarchical HMMs". In: *NIPS*. 2001.

[MP21] D. Mazza and M. Pagani. "Automatic Differentiation in PCF". In: *Proc. ACM Program. Lang.* 5.POPL (2021).

[MP95] D. MacKay and L. Peto. "A hierarchical dirichlet language model". In: *Natural Language Engineering* 1.3 (1995), pp. 289–307.

[MPS18] O. Mangoubi, N. S. Pillai, and A. Smith. "Does Hamiltonian Monte Carlo mix faster than a random walk on multimodal densities?" In: (2018). arXiv: 1808.03230 [math.PR].

[MR09] A. Melkumyan and F. Ramos. "A Sparse Covariance Function for Exact Gaussian Process Inference in Large Datasets". In: *IJCAI*. 2009, pp. 1936–1942.

[MR10] B. Milch and S. Russell. "Extending Bayesian Networks to the Open-Universe Case". In: *Heuristics, Probability and Causality. A Tribute to Judea Pearl*. Ed. by R. Dechter, H. Geffner, and J. Y. Halper. College Publications, 2010.

[MRB18] A. S. Morcos, M. Raghu, and S. Bengio. "Insights on representational similarity in neural networks with canonical correlation". In: *Advances in neural information processing systems* (2018).

[MRW19] B. Mittelstadt, C. Russell, and S. Wachter. "Explaining explanations in AI". In: *Proceedings of the conference on fairness, accountability, and transparency*. 2019, pp. 279–288.

[MS96] V Matveev V and R Shrock. "Complex-temperature singularities in Potts models on the square lattice". en. In: *Phys. Rev. E Stat. Phys. Plasmas Fluids Relat. Interdiscip. Topics* 54.6 (1996), pp. 6174–6185.

[MS99] C. Manning and H. Schuetze. *Foundations of statistical natural language processing*. MIT Press, 1999.

[MSA18] S. Makridakis, E. Spiliotis, and V. Assimakopoulos. "The M4 Competition: Results, findings, conclusion and way forward". In: *Int. J. Forecast.* 34.4 (2018), pp. 802–808.

[MT12] A. Mnih and Y. W. Teh. "A fast and simple algorithm for training neural probabilistic language models". In: *ICML*. 2012, pp. 419–426.

[MTM14] C. J. Maddison, D. Tarlow, and T. Minka. "A* Sampling". In: *NIPS*. 2014.

[MTS22] Y. Ma, D. Tsao, and H.-Y. Shum. "On the Principles of Parsimony and Self-Consistency for the Emergence of Intelligence". In: (July 2022). arXiv: 2207.04630 [cs.AI].

[Mua+17] K. Muandet, K. Fukumizu, B. Sriperumbudur, and B. Schölkopf. "Kernel Mean Embedding of Distributions: A Review and Beyond". In: *Foundations and Trends* 10.1-2 (2017), pp. 1–141.

[Mua+20] K. Muandet, A. Mehrjou, S. K. Lee, and A. Raj. *Dual Instrumental Variable Regression*. 2020.

Author: Kevin P. Murphy. (C) MIT Press. CC-BY-NC-ND license

[Muk+18] S. Mukherjee, D. Shankar, A. Ghosh, N. Tathawadekar, P. Kompalli, S. Sarawagi, and K. Chaudhury. "ARMDN: Associative and Recurrent Mixture Density Networks for eRetail Demand Forecasting". In: (2018). arXiv: 1803.03800 [cs.LG].

[Mül+19a] T. Müller, B. McWilliams, F. Rousselle, M. Gross, and J. Novák. "Neural Importance Sampling". In: *SIGGRAPH*. 2019.

[Mül+19b] T. Müller, B. McWilliams, F. Rousselle, M. Gross, and J. Novák. "Neural importance sampling". In: *ACM Transactions on Graphics* 38.5 (2019), p. 145.

[Mun14] R. Munos. "From Bandits to Monte-Carlo Tree Search: The Optimistic Principle Applied to Optimization and Planning". In: *Foundations and Trends in Machine Learning* 7.1 (2014), pp. 1–129.

[Mun+16] R. Munos, T. Stepleton, A. Harutyunyan, and M. G. Bellemare. "Safe and Efficient Off-Policy Reinforcement Learning". In: *NIPS*. 2016, pp. 1046–1054.

[Mun57] J. Munkres. "Algorithms for the assignment and transportation problems". In: *Journal of the society for industrial and applied mathematics* 5.1 (1957), pp. 32–38.

[Mur00] K. Murphy. "Bayesian Map Learning in Dynamic Environments". In: *NIPS*. Vol. 12. 2000.

[Mur02] K. Murphy. "Dynamic Bayesian Networks: Representation, Inference and Learning". PhD thesis. Dept. Computer Science, UC Berkeley, 2002.

[Mur+19] W. J. Murdoch, C. Singh, K. Kumbier, R. Abbasi-Asl, and B. Yu. "Definitions, methods, and applications in interpretable machine learning". In: *Proceedings of the National Academy of Sciences* 116.44 (2019), pp. 22071–22080.

[Mur22] K. P. Murphy. *Probabilistic Machine Learning: An introduction*. MIT Press, 2022.

[MW07] J. Møller and R. P. Waagepetersen. "Modern statistics for spatial point processes". In: *Scandinavian Journal of Statistics* 34.4 (2007), pp. 643–684.

[MW15] S. Morgan and C. Winship. *Counterfactuals and Causal Inference*. 2nd. Cambridge University Press, 2015.

[MWJ99] K. Murphy, Y. Weiss, and M. Jordan. "Loopy Belief Propagation for Approximate Inference: an Empirical Study". In: *UAI*. 1999.

[MYM18] J. Marino, Y. Yue, and S. Mandt. "Iterative Amortized Inference". In: *ICML*. 2018.

[Nac+17] O. Nachum, M. Norouzi, K. Xu, and D. Schuurmans. "Bridging the Gap Between Value and Policy Based Reinforcement Learning". In: *NIPS*. 2017, pp. 2772–2782.

[Nac+19a] O. Nachum, Y. Chow, B. Dai, and L. Li. "DualDICE: Behavior-agnostic Estimation of Discounted Stationary Distribution Corrections". In: *NeurIPS*. 2019, pp. 2315–2325.

[Nac+19b] O. Nachum, B. Dai, I. Kostrikov, Y. Chow, L. Li, and D. Schuurmans. *Algae: Policy Gradient from Arbitrary Experience*. CoRR abs/1912.02074. 2019.

[Nad+19] S. Naderi, K. He, R. Aghajani, S. Sclaroff, and P. Felzenszwalb. "Generalized Majorization-Minimization". In: *ICML*. 2019.

[Nae+18] C. A. Naesseth, S. W. Linderman, R. Ranganath, and D. M. Blei. "Variational Sequential Monte Carlo". In: *AISTATS*. 2018.

[Nal18] E. T. Nalisnick. "On Priors for Bayesian Neural Networks". PhD thesis. UC Irvine, 2018.

[Nal+19a] E. Nalisnick, A. Matsukawa, Y. W. Teh, D. Gorur, and B. Lakshminarayanan. "Do Deep Generative Models Know What They Don't Know?" In: *ICLR*. 2019.

[Nal+19b] E. Nalisnick, A. Matsukawa, Y. W. Teh, D. Gorur, and B. Lakshminarayanan. "Hybrid Models with Deep and Invertible Features". In: *ICML*. 2019, pp. 4723–4732.

[Nau04] J. Naudts. "Estimators, escort probabilities and ϕ-exponential families in statistical physics". In: *J. of Inequalities in Pure and Applied Mathematics* 5.4 (2004).

[NB05] M. Narasimhan and J. Bilmes. "A Submodular-Supermodular Procedure with Applications to Discriminative Structure Learning". In: *Uncertainty in Artificial Intelligence: Proceedings of the Twentieth Conference (UAI-2004)*. Edinburgh, Scotland: Morgan Kaufmann Publishers, 2005.

[NB06] M. Narasimhan and J. Bilmes. *Learning Graphical Models over partial k-trees*. Tech. rep. UWEETR-2006-0001. https://vannevar.ece.uw.edu/techsite/papers/refer/UWEETR-2006-0001.html. University of Washington, Department of Electrical Engineering, 2006.

[NBS18] B. Neyshabur, S. Bhojanapalli, and N. Srebro. "A PAC-Bayesian Approach to Spectrally-Normalized Margin Bounds for Neural Networks". In: *ICLR*. 2018.

[NCH15] M. Naeini, G. Cooper, and M. Hauskrecht. "Obtaining well calibrated probabilities using Bayesian binning". In: *AAAI*. 2015.

[NCL20] T. Nguyen, Z. Chen, and J. Lee. "Dataset Meta-Learning from Kernel Ridge-Regression". In: *International Conference on Learning Representations*. 2020.

[NCT16a] S. Nowozin, B. Cseke, and R. Tomioka. "f-GAN: Training Generative Neural Samplers using Variational Divergence Minimization". In: *NIPS*. Ed. by D. D. Lee, M Sugiyama, U. V. Luxburg, I Guyon, and R Garnett. Curran Associates, Inc., 2016, pp. 271–279.

[NCT16b] S. Nowozin, B. Cseke, and R. Tomioka. "f-gan: Training generative neural samplers using variational divergence minimization". In: *NIPS*. 2016, pp. 271–279.

[NCT16c] S. Nowozin, B. Cseke, and R. Tomioka. "f-gan: Training generative neural samplers using variational divergence minimization". In: *Advances in neural information processing systems*. 2016, pp. 271–279.

[ND20] O. Nachum and B. Dai. "Reinforcement Learning via Fenchel-Rockafellar Duality". In: (2020). arXiv: 2001.01866 [cs.LG].

[ND21] A. Nichol and P. Dhariwal. "Improved Denoising Diffusion Probabilistic Models". In: *ICML*. Feb. 2021.

[NDL20] A. Nishimura, D. Dunson, and J. Lu. "Discontinuous Hamiltonian Monte Carlo for discrete parameters and discontinuous likelihoods". In: *Biometrika* (2020).

[Nea00] R. Neal. "Markov Chain Sampling Methods for Dirichlet Process Mixture Models". In: *JCGS* 9.2 (2000), pp. 249–265.

[Nea01] R. M. Neal. "Annealed Importance Sampling". In: *Statistics and Computing* 11 (2001), pp. 125–139.

[Nea03] R. Neal. "Slice sampling". In: *Annals of Statistics* 31.3 (2003), pp. 7–5–767.

[Nea+08] R. Neal et al. "Computing likelihood functions for high-energy physics experiments when distributions are defined by simulators with nuisance parameters". In: (2008).

[Nea10] R. Neal. "MCMC using Hamiltonian Dynamics". In: *Handbook of Markov Chain Monte Carlo*. Ed. by S. Brooks, A. Gelman, G. Jones, and X.-L. Meng. Chapman & Hall, 2010.

[Nea12] R. C. Neath. "On Convergence Properties of the Monte Carlo EM Algorithm". In: *arXiv [math.ST]* (2012).

[Nea20] B. Neal. *Introduction to Causal Inference from a Machine Learning Perspective*. 2020.

[Nea92] R. Neal. "Connectionist learning of belief networks". In: *Artificial Intelligence* 56 (1992), pp. 71–113.

[Nea93] R. M. Neal. *Probabilistic Inference using Markov Chain Monte Carlo Methods*. Tech. rep. CRG-TR-93-1. 144pp. Dept. of Computer Science, University of Toronto, 1993.

[Nea96] R. Neal. *Bayesian learning for neural networks*. Springer, 1996.

[Nef+02] A. Nefian, L. Liang, X. Pi, X. Liu, and K. Murphy. "Dynamic Bayesian Networks for Audio-Visual Speech Recognition". In: *J. Applied Signal Processing* (2002).

[Neg+21] J. Negrea, J. Yang, H. Feng, D. M. Roy, and J. H. Huggins. "Statistical inference with stochastic gradient algorithms". 2021.

[Nei+18] D. Neil, J. Briody, A. Lacoste, A. Sim, P. Creed, and A. Saffari. "Interpretable graph convolutional neural networks for inference on noisy knowledge graphs". In: *arXiv preprint arXiv:1812.00279* (2018).

[Nel21] Nelson Elhage and Neel Nanda and Catherine Olsson and Tom Henighan and Nicholas Joseph and Ben Mann and Amanda Askell and Yuntao Bai and Anna Chen and Tom Conerly and Nova DasSarma and Dawn Drain and Deep Ganguli and Zac Hatfield-Dodds and Danny Hernandez and Andy Jones and Jackson Kernion and Liane Lovitt and Kamal Ndousse and Dario Amodei and Tom Brown and Jack Clark and Jared Kaplan and Sam McCandlish and Chris Olah. *A Mathematical Framework for Transformer Circuits*. Tech. rep. Anthropic, 2021.

[Neu11] G. Neumann. "Variational Inference for Policy Search in Changing Situations". In: *ICML*. 2011, pp. 817–824.

[Ney+17] B. Neyshabur, S. Bhojanapalli, D. McAllester, and N. Srebro. "Exploring generalization in deep learning". In: *NIPS*. 2017.

[NF16] M. Noroozi and P. Favaro. "Unsupervised learning of visual representations by solving jigsaw puzzles". In: *European conference on computer vision*. Springer. 2016, pp. 69–84.

[NG01] D. Nilsson and J. Goldberger. "Sequentially finding the N-Best List in Hidden Markov Models". In: *IJCAI*. 2001, pp. 1280–1285.

[Ng+11] A. Ng et al. "Sparse autoencoder". In: *CS294A Lecture notes* 72.2011 (2011), pp. 1–19.

[Ngi+11] J. Ngiam, Z. Chen, P. W. Koh, and A. Y. Ng. "Learning deep energy models". In: *Proceedings of the 28th international conference on machine learning (ICML-11)*. 2011, pp. 1105–1112.

[Ngi+18] J. Ngiam, D. Peng, V. Vasudevan, S. Kornblith, Q. V. Le, and R. Pang. "Domain adaptive transfer learning with specialist models". In: *arXiv preprint arXiv:1811.07056* (2018).

[Ngu+16] A. Nguyen, A. Dosovitskiy, J. Yosinski, T. Brox, and J. Clune. *Synthesizing the preferred inputs for neurons in neural networks via deep generator networks*. 2016. arXiv: 1605.09304 [cs.NE].

[Ngu+18] C. V. Nguyen, Y. Li, T. D. Bui, and R. E. Turner. "Variational Continual Learning". In: *ICLR*. 2018.

[Ngu+19] T. T. Nguyen, Q. V. H. Nguyen, D. T. Nguyen, D. T. Nguyen, Thien Huynh-The, S. Nahavandi, T. T. Nguyen, Q.-V. Pham, and C. M. Nguyen. "Deep Learning for Deepfakes Creation and Detection: A Survey". In: (2019). arXiv: 1909.11573 [cs.CV].

[Ngu+21] T. Nguyen, R. Novak, L. Xiao, and J. Lee. "Dataset Distillation with Infinitely Wide Convolutional Networks". In: *Advances in Neural Information Processing Systems*. Ed. by A. Beygelzimer, Y. Dauphin, P. Liang, and J. W. Vaughan. 2021.

[NH98a] R. M. Neal and G. E. Hinton. "A new view of the EM algorithm that justifies incremental and other variants". In: *Learning in Graphical Models*. Ed. by M. Jordan. MIT Press, 1998.

[NH98b] R. M. Neal and G. E. Hinton. "A View of the EM Algorithm that Justifies Incremental, Sparse, and other Variants". In: *Learning in Graphical Models*. Ed. by M. I. Jordan. Springer Netherlands, 1998, pp. 355–368.

[NHLS19] E. Nalisnick, J. M. Hernández-Lobato, and P. Smyth. "Dropout as a Structured Shrinkage Prior". In: *ICML*. 2019.

[NHR99] A. Ng, D. Harada, and S. Russell. "Policy invariance under reward transformations: Theory and application to reward shaping". In: *ICML*. 1999.

[NI92] H. Nagamochi and T. Ibaraki. "Computing edge-connectivity of multigraphs and capacitated graphs". In: *SIAM J. Discrete Math.* 5 (1992), pp. 54–66.

[Nic+21] A. Nichol, P. Dhariwal, P. Ramesh Aditya and> Shyam, P. Mishkin, B. McGrew, I. Sutskever, and M. Chen. "GLIDE: Towards Photorealistic Image Generation and Editing with Text-Guided Diffusion Models". In: (2021). arXiv: 2112.10741 [cs.CV].

[Nig+00] K. Nigam, A. K. Mccallum, S. Thrun, and T. Mitchell. "Text Classification from Labeled and Unlabeled Documents using EM". In: *MLJ* 39.2 (May 2000), pp. 103–134.

[Nij+19] E. Nijkamp, M. Hill, S.-C. Zhu, and Y. N. Wu. "Learning non-convergent non-persistent short-run MCMC toward energy-based model". In: *NIPS*. 2019, pp. 5232–5242.

[Nix+19] J. Nixon, M. Dusenberry, L. Zhang, G. Jerfel, and D. Tran. "Measuring Calibration in Deep Learning". In: (2019). arXiv: 1904.01685 [cs.LG].

[NJ00] A. Y. Ng and M. Jordan. "PEGASUS: A policy search method for large MDPs and POMDPs". In: *UAI*. 2000.

[NJB05] M. Narasimhan, N. Jojic, and J. A. Bilmes. "Q-clustering". In: *Advances in Neural Information Processing Systems* 18 (2005), pp. 979–986.

[NK17] V. Nagarajan and J. Z. Kolter. "Gradient descent GAN optimization is locally stable". In: *Ad-*

vances in neural information processing systems*. 2017, pp. 5585–5595.

[NKI10] K. Nagano, Y. Kawahara, and S. Iwata. "Minimum average cost clustering". In: *Advances in Neural Information Processing Systems* 23 (2010), pp. 1759–1767.

[NLS15] C. A. Naesseth, F. Lindsten, and T. B. Schön. "Nested Sequential Monte Carlo Methods". In: *ICML*. 2015.

[NLS19] C. A. Naesseth, F. Lindsten, and T. B. Schön. "Elements of Sequential Monte Carlo". In: *Foundations and Trends in Machine Learning* (2019).

[NM12] A. Nenkova and K. McKeown. "A survey of text summarization techniques". In: *Mining text data*. Springer, 2012, pp. 43–76.

[NMC05] A. Niculescu-Mizil and R. Caruana. "Predicting Good Probabilities with Supervised Learning". In: *ICML*. 2005.

[NNP19] W. Nie, N. Narodytska, and A. Patel. "RelGAN: Relational Generative Adversarial Networks for Text Generation". In: *International Conference on Learning Representations*. 2019.

[Noc+21] L. Noci, K. Roth, G. Bachmann, S. Nowozin, and T. Hofmann. "Disentangling the Roles of Curation, Data-Augmentation and the Prior in the Cold Posterior Effect". In: *NIPS*. 2021.

[Noé+19] F. Noé, S. Olsson, J. Köhler, and H. Wu. "Boltzmann generators: Sampling equilibrium states of many-body systems with deep learning". In: *Science* 365 (2019).

[Nou+02] M. N. Nounou, B. R. Bakshi, P. K. Goel, and X. Shen. "Process modeling by Bayesian latent variable regression". In: *Am. Inst. Chemical Engineers Journal* 48.8 (2002), pp. 1775–1793.

[Nov+19] R. Novak, L. Xiao, J. Lee, Y. Bahri, G. Yang, J. Hron, D. A. Abolafia, J. Pennington, and J. Sohl-Dickstein. "Bayesian Deep Convolutional Networks with Many Channels are Gaussian Processes". In: *ICLR*. 2019.

[NP03] W. K. Newey and J. L. Powell. "Instrumental variable estimation of nonparametric models". In: *Econometrica* 71.5 (2003), pp. 1565–1578.

[NR00a] A. Ng and S. Russell. "Algorithms for inverse reinforcement learning". In: *ICML*. 2000.

[NR00b] A. Y. Ng and S. Russell. "Algorithms for Inverse Reinforcement Learning". In: *in Proc. 17th International Conf. on Machine Learning*. Citeseer. 2000.

[NR94] M. Newton and A. Raftery. "Approximate Bayesian Inference with the Weighted Likelihood Bootstrap". In: *J. of Royal Stat. Soc. Series B* 56.1 (1994), pp. 3–48.

[NS01] K. Nowicki and T. A. B. Snijders. "Estimation and Prediction for Stochastic Blockstructures". In: *Journal of the American Statistical Association* 96.455 (2001), pp. 1077–1087.

[NS17] E. Nalisnick and P. Smyth. "Variational Reference Priors". In: *ICLR Workshop*. 2017.

[NS18] E. Nalisnick and P. Smyth. "Learning Priors for Invariance". In: *AISTATS*. 2018.

[NW06] J. Nocedal and S. Wright. *Numerical Optimization*. Springer, 2006.

[NW20] X Nie and S Wager. "Quasi-oracle estimation of heterogeneous treatment effects". In: *Biometrika* 108.2 (Sept. 2020), pp. 299–319. eprint: https://academic.oup.com/biomet/article-pdf/108/2/299/37938939/asaa076.pdf.

[NWF78] G. Nemhauser, L. Wolsey, and M. Fisher. "An analysis of approximations for maximizing submodular set functions—I". In: *Mathematical Programming* 14.1 (1978), pp. 265–294.

[NWJ09] X. Nguyen, M. J. Wainwright, and M. I. Jordan. "On Surrogate Loss Functions and f-Divergences". In: *Ann. Stat.* 37.2 (2009), pp. 876–904.

[NWJ+09] X. Nguyen, M. J. Wainwright, M. I. Jordan, et al. "On surrogate loss functions and f-divergences". In: *The Annals of Statistics* 37.2 (2009), pp. 876–904.

[NWJ10a] X Nguyen, M. J. Wainwright, and M. I. Jordan. "Estimating Divergence Functionals and the Likelihood Ratio by Convex Risk Minimization". In: *IEEE Trans. Inf. Theory* 56.11 (2010), pp. 5847–5861.

[NWJ10b] X. Nguyen, M. J. Wainwright, and M. I. Jordan. "Estimating divergence functionals and the likelihood ratio by convex risk minimization". In: *IEEE Transactions on Information Theory* 56.11 (2010), pp. 5847–5861.

[NYC15] A. Nguyen, J. Yosinski, and J. Clune. "Deep Neural Networks are Easily Fooled: High Confidence Predictions for Unrecognizable Images". In: *CVPR*. 2015.

[OA09] M. Opper and C. Archambeau. "The variational Gaussian approximation revisited". en. In: *Neural Comput.* 21.3 (2009), pp. 786–792.

[OAC18] I. Osband, J. Aslanides, and A. Cassirer. "Randomized prior functions for deep reinforcement learning". In: *NIPS*. 2018.

[Obe+19] F. Obermeyer, E. Bingham, M. Jankowiak, J. Chiu, N. Pradhan, A. Rush, and N. Goodman. "Tensor Variable Elimination for Plated Factor Graphs". In: *ICML*. 2019.

[OCM21] L. A. Ortega, R. Cabañas, and A. R. Masegosa. "Diversity and Generalization in Neural Network Ensembles". In: (2021). arXiv: 2110.13786 [cs.LG].

[ODK96] M. Ostendorf, V. Digalakis, and O. Kimball. "From HMMs to segment models: a unified view of stochastic modeling for speech recognition". In: *IEEE Trans. on Speech and Audio Processing* 4.5 (1996), pp. 360–378.

[OED21] J. Ortiz, T. Evans, and A. J. Davison. "A visual introduction to Gaussian Belief Propagation". In: *arXiv preprint arXiv:2107.02308* (2021).

[OF96] B. A. Olshausen and D. J. Field. "Emergence of simple cell receptive field properties by learning a sparse code for natural images". In: *Nature* 381 (1996), pp. 607–609.

[O'H78] A. O'Hagan. "Curve Fitting and Optimal Design for Prediction". In: *J. of Royal Stat. Soc. Series B* 40 (1978), pp. 1–42.

[OKK16] A. Van den Oord, N. Kalchbrenner, and K. Kavukcuoglu. "Pixel Recurrent Neural Networks". In: *ICML*. 2016.

[Oll+17] Y. Ollivier, L. Arnold, A. Auger, and N. Hansen. "Information-Geometric Optimization Algorithms: A Unifying Picture via Invariance Principles". In: *JMLR* 18 (2017), pp. 1–65.

[Oll18] Y. Ollivier. "Online natural gradient as a Kalman filter". en. In: *Electron. J. Stat.* 12.2 (2018), pp. 2930–2961.

[OLV18a] A. van den Oord, Y. Li, and O. Vinyals. "Representation Learning with Contrastive Predictive Coding". In: (2018). arXiv: 1807.03748 [cs.LG].

[OLV18b] A. v. d. Oord, Y. Li, and O. Vinyals. "Representation learning with contrastive predictive coding". In: *arXiv preprint arXiv:1807.03748* (2018).

[OM96] P. V. Overschee and B. D. Moor. *Subspace Identification for Linear Systems: Theory, Implementation, Applications*. Kluwer Academic Publishers, 1996.

[OMS17] C. Olah, A. Mordvintsev, and L. Schubert. "Feature Visualization". In: *Distill* 2.11 (2017).

[O'N09] B. O'Neill. "Exchangeability, Correlation, and Bayes' Effect". In: *Int. Stat. Rev.* 77.2 (2009), pp. 241–250.

[ONS18] V. M.-H. Ong, D. J. Nott, and M. S. Smith. "Gaussian Variational Approximation With a Factor Covariance Structure". In: *J. Comput. Graph. Stat.* 27.3 (2018), pp. 465–478.

[oor+16] A. Van den oord, S. Dieleman, H. Zen, K. Simonyan, O. Vinyals, A. Graves, N. Kalchbrenner, A. Senior, and K. Kavukcuoglu. "WaveNet: A Generative Model for Raw Audio". In: (2016). arXiv: 1609.03499 [cs.SD].

[Oor+16] A. van den Oord, N. Kalchbrenner, O. Vinyals, L. Espeholt, A. Graves, and K. Kavukcuoglu. "Conditional Image Generation with PixelCNN Decoders". In: (2016). arXiv: 1606.05328 [cs.CV].

[Oor+18] A. van den Oord et al. "Parallel WaveNet: Fast High-Fidelity Speech Synthesis". In: *ICML*. Ed. by J. Dy and A. Krause. Vol. 80. Proceedings of Machine Learning Research. PMLR, 2018, pp. 3918–3926.

[Oor+19] A. van den Oord, B. Poole, O. Vinyals, and A. Razavi. "Fixing Posterior Collapse with delta-VAEs". In: *ICLR*. 2019.

[OOS17] A. Odena, C. Olah, and J. Shlens. "Conditional image synthesis with auxiliary classifier gans". In: *International conference on machine learning*. 2017, pp. 2642–2651.

[Opt88] Optimal information processing and Bayes' theorem. In: *The American Statistician* 42.4 (1988), pp. 278–280.

[Oqu+14] M. Oquab, L. Bottou, I. Laptev, and J. Sivic. "Learning and transferring mid-level image representations using convolutional neural networks". In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2014, pp. 1717–1724.

[OR20] A. Owen and D. Rudolf. "A strong law of large numbers for scrambled net integration". In: (2020). arXiv: 2002.07859 [math.NA].

[Ore+20] B. N. Oreshkin, D. Carpov, N. Chapados, and Y. Bengio. "N-BEATS: Neural basis expansion analysis for interpretable time series forecasting". In: *ICLR*. 2020.

[ORW21] S. W. Ober, C. E. Rasmussen, and M. van der Wilk. "The Promises and Pitfalls of Deep Kernel Learning". In: *ICML*. Feb. 2021.

[Osa+18] T. Osa, J. Pajarinen, G. Neumann, J. A. Bagnell, P. Abbeel, and J. Peters. "An Algorithmic Perspective on Imitation Learning". In: *Foundations and Trends in Robotics* 7.1–2 (2018), pp. 1–179.

[Osa+19a] K. Osawa, S. Swaroop, A. Jain, R. Eschenhagen, R. E. Turner, R. Yokota, and M. E. Khan. "Practical Deep Learning with Bayesian Principles". In: *NIPS*. 2019.

[Osa+19b] K. Osawa, Y. Tsuji, Y. Ueno, A. Naruse, R. Yokota, and S. Matsuoka. "Large-Scale Distributed Second-Order Optimization Using Kronecker-Factored Approximate Curvature for Deep Convolutional Neural Networks". In: *CVPR*. 2019.

[Osb16] I. Osband. "Risk versus Uncertainty in Deep Learning: Bayes, Bootstrap and the Dangers of Dropout". In: *NIPS workshop on Bayesian deep learning*. 2016.

[Osb+21] I. Osband, Z. Wen, S. M. Asghari, V. Dwaracherla, B. Hao, M. Ibrahimi, D. Lawson, X. Lu, B. O'Donoghue, and B. Van Roy. "The Neural Testbed: Evaluating Predictive Distributions". In: (2021). arXiv: 2110.04629 [cs.LG].

[Ose11] I. V. Oseledets. "Tensor-Train Decomposition". In: *SIAM J. Sci. Comput.* 33.5 (2011), pp. 2295–2317.

[OT05] A. B. Owen and S. D. Tribble. "A quasi-Monte Carlo Metropolis algorithm". en. In: *PNAS* 102.25 (2005), pp. 8844–8849.

[Ova+19] Y. Ovadia, E. Fertig, J. Ren, Z. Nado, D Sculley, S. Nowozin, J. V. Dillon, B. Lakshminarayanan, and J. Snoek. "Can You Trust Your Model's Uncertainty? Evaluating Predictive Uncertainty Under Dataset Shift". In: *NIPS*. 2019.

[OVK17] A. van den Oord, O. Vinyals, and K. Kavukcuoglu. "Neural Discrete Representation Learning". In: *NIPS*. 2017.

[Owe13] A. B. Owen. *Monte Carlo theory, methods and examples*. 2013.

[Owe17] A. B. Owen. "A randomized Halton algorithm in R". In: *arXiv [stat.CO]* (2017).

[Oxl11] J. Oxley. *Matroid Theory: Second Edition*. Oxford University Press, 2011.

[Pac+14] J. Pacheco, S. Zuffi, M. Black, and E. Sudderth. "Preserving Modes and Messages via Diverse Particle Selection". en. In: *ICML*. 2014, pp. 1152–1160.

[Pai] *Explainable AI in Practice Falls Short of Transparency Goals*. https://partnershiponai.org/xai-in-practice/. Accessed: 2021-11-23.

[Pai+14] T. L. Paine, P. Khorrami, W. Han, and T. S. Huang. "An analysis of unsupervised pre-training in light of recent advances". In: *arXiv preprint arXiv:1412.6597* (2014).

[Pan+10] L. Paninski, Y. Ahmadian, D. G. Ferreira, S. Koyama, K. Rahnama Rad, M. Vidne, J. Vogelstein, and W. Wu. "A new look at state-space models for neural data". en. In: *J. Comput. Neurosci.* 29.1-2 (2010), pp. 107–126.

[Pan+21] G. Pang, C. Shen, L. Cao, and A. Van Den Hengel. "Deep Learning for Anomaly Detection: A Review". In: *ACM Comput. Surv.* 54.2 (2021), pp. 1–38.

[Pap+17] N. Papernot, P. McDaniel, I. Goodfellow, S. Jha, Z Berkay Celik, and A. Swami. "Practical Black-Box Attacks against Deep Learning Systems using Adversarial Examples". In: *ACM Asia Conference on Computer and Communications Security*. 2017.

[Pap+19] G. Papamakarios, E. Nalisnick, D. J. Rezende, S. Mohamed, and B. Lakshminarayanan. "Normalizing Flows for Probabilistic Modeling and Inference". In: (2019). arXiv: 1912.02762 [stat.ML].

[Par+19] S Parameswaran, C Deledalle, L Denis, and T. Q. Nguyen. "Accelerating GMM-Based Patch Priors for Image Restoration: Three Ingredients for a 100× Speed-Up". In: *IEEE Trans. Image Process.* 28.2 (2019), pp. 687–698.

[Par81] G. Parisi. "Correlation functions and computer simulations". In: *Nuclear Physics B* 180.3 (1981), pp. 378–384.

[Pas+02] H. Pasula, B. Marthi, B. Milch, S. Russell, and I. Shpitser. "Identity Uncertainty and Citation Matching". In: *NIPS*. 2002.

[Pas+21a] A. Paszke, D. Johnson, D. Duvenaud, D. Vytiniotis, A. Radul, M. Johnson, J. Ragan-Kelley, and D. Maclaurin. "Getting to the Point: Index Sets and Parallelism-Preserving Autodiff for Pointful Array Programming". In: *Proc. ACM Program. Lang.* 5.ICFP (2021).

[Pas+21b] A. Paszke, M. J. Johnson, R. Frostig, and D. Maclaurin. "Parallelism-Preserving Automatic Differentiation for Second-Order Array Languages". In: *Proceedings of the 9th ACM SIGPLAN International Workshop on Functional High-Performance and Numerical Computing*. FHPNC 2021. Association for Computing Machinery, 2021, 13–23.

[Pat+16] D. Pathak, P. Krahenbuhl, J. Donahue, T. Darrell, and A. A. Efros. "Context encoders: Feature learning by inpainting". In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2016, pp. 2536–2544.

[PB14] R. Pascanu and Y. Bengio. "Revisiting Natural Gradient for Deep Networks". In: *ICLR*. 2014.

[PBM16a] J. Peters, P. Bühlmann, and N. Meinshausen. "Causal inference by using invariant prediction: identification and confidence intervals". In: *Journal of the Royal Statistical Society. Series B (Statistical Methodology)* 78.5 (2016), pp. 947–1012.

[PBM16b] J. Peters, P. Bühlmann, and N. Meinshausen. "Causal inference using invariant prediction: identification and confidence intervals". In: *J. of Royal Stat. Soc. Series B* 78.5 (2016), pp. 947–1012.

[PC08] T. Park and G. Casella. "The Bayesian Lasso". In: *JASA* 103.482 (2008), pp. 681–686.

[PC09] J. Paisley and L. Carin. "Nonparametric Factor Analysis with Beta Process Priors". In: *ICML*. 2009.

[PC12] N. Pinto and D. D. Cox. "High-throughput-derived biologically-inspired features for unconstrained face recognition". In: *Image Vis. Comput.* 30.3 (2012), pp. 159–168.

[PD03] J. D. Park and A. Darwiche. "A Differential Semantics for Jointree Algorithms". In: *NIPS*. MIT Press, 2003, pp. 801–808.

[PD11] H. Poon and P. Domingos. "Sum-Product Networks: A New Deep Architecture". In: *UAI*. Java code at http://alchemy.cs.washington.edu/spn/. Short intro at http://lessoned.blogspot.com/2011/10/intro-to-sum-product-networks.html. 2011.

[PdC20] F.-P. Paty, A. d'Aspremont, and M. Cuturi. "Regularity as Regularization: Smooth and Strongly Convex Brenier Potentials in Optimal Transport". In: *Proceedings of the Twenty Third International Conference on Artificial Intelligence and Statistics*. Ed. by S. Chiappa and R. Calandra. Vol. 108. Proceedings of Machine Learning Research. PMLR, 2020, pp. 1222–1232.

[PDL+12] M. Parry, A. P. Dawid, S. Lauritzen, et al. "Proper local scoring rules". In: *The Annals of Statistics* 40.1 (2012), pp. 561–592.

[PE16] V. Papyan and M. Elad. "Multi-Scale Patch-Based Image Restoration". en. In: *IEEE Trans. Image Process.* 25.1 (2016), pp. 249–261.

[Pea09a] J. Pearl. *Causality*. 2nd. Cambridge University Press, 2009.

[Pea09b] J. Pearl. *Causality: Models, Reasoning and Inference (Second Edition)*. Cambridge Univ. Press, 2009.

[Pea09c] J. Pearl. "Causal inference in statistics: An overview". In: *Stat. Surv.* 3.0 (2009), pp. 96–146.

[Pea12] J. Pearl. "The Causal Foundations of Structural Equation Modeling". In: *Handbook of structural equation modeling*. Ed. by R. H. Hoyle. Vol. 68. 2012.

[Pea19] J. Pearl. "The Seven Tools of Causal Inference, with Reflections on Machine Learning". In: *Comm. of the ACM* 62.3 (2019), pp. 54–60.

[Pea36] K. Pearson. "Method of moments and method of maximum likelihood". In: *Biometrika* 28.1/2 (1936), pp. 34–59.

[Pea84] J. Pearl. *Heuristics: Intelligent Search Strategies for Computer Problem Solving*. Addison-Wesley Longman Publishing Co., Inc., 1984.

[Pea88] J. Pearl. *Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference*. Morgan Kaufmann, 1988.

[Pea94] B. A. Pearlmutter. "Fast Exact Multiplication by the Hessian". In: *Neural Comput.* 6.1 (1994), pp. 147–160.

[Peh+20] R. Peharz, S. Lang, A. Vergari, K. Stelzner, A. Molina, M. Trapp, G. Van den Broeck, K. Kersting, and Z. Ghahramani. "Einsum Networks: Fast and Scalable Learning of Tractable Probabilistic Circuits". In: (2020). arXiv: 2004.06231 [cs.LG].

[Pel05] M. Pelikan. *Hierarchical Bayesian Optimization Algorithm: Toward a New Generation of Evolutionary Algorithms*. en. Softcover reprint of hardcover 1st ed. 2005 edition. Springer, 2005.

[Pen13] J. Pena. "Reading dependencies from covariance graphs". In: *Intl. J. of Approximate Reasoning* 54.1 (2013).

[Per+18] E. Perez, F. Strub, H. de Vries, V. Dumoulin, and A. Courville. "FiLM: Visual Reasoning with a General Conditioning Layer". In: *AAAI*. 2018.

[Per+21] V. Perrone, M. Donini, M. B. Zafar, R. Schmucker, K. Kenthapadi, and C. Archambeau. "Fair Bayesian Optimization". In: *AAAI/ACM Conference on AI, Ethics, and Society (AIES '21)*. 2021.

[Pes+21] H. Pesonen et al. "ABC of the Future". In: (2021). arXiv: 2112.12841 [stat.AP].

[Pey20] G. Peyre. "Course notes on Optimization for Machine Learning". 2020.

[Pez+21] M. Pezeshki, S.-O. Kaba, Y. Bengio, A. Courville, D. Precup, and G. Lajoie. "Gradient Starvation: A Learning Proclivity in Neural Networks". In: *NIPS*. 2021.

[PF03] G. V. Puskorius and L. A. Feldkamp. "Parameter-based Kalman filter training: Theory and implementation". In: *Kalman Filtering and Neural Networks*. John Wiley & Sons, Inc., 2003, pp. 23–67.

[PF91] G. V. Puskorius and L. A. Feldkamp. "Decoupled extended Kalman filter training of feedforward layered networks". In: *International Joint Conference on Neural Networks*. Vol. i. 1991, 771–777 vol.1.

[PFW21] R. Prado, M. Ferreira, and M. West. *Time Series: Modelling, Computation and Inference (2nd ed)*. CRC Press, 2021.

[PG98] M. Popescu and P. D. Gader. "Image content retrieval from image databases using feature integration by Choquet integral". In: *Storage and Retrieval for Image and Video Databases VII*. Ed. by M. M. Yeung, B.-L. Yeo, and C. A. Bouman. Vol. 3656. International Society for Optics and Photonics. SPIE, 1998, pp. 552 –560.

[PGCP00] M Pelikan, D. E. Goldberg, and E Cantú-Paz. "Linkage problem, distribution estimation, and Bayesian networks". en. In: *Evol. Comput.* 8.3 (2000), pp. 311–340.

[PGJ16] J. Pearl, M. Glymour, and N. Jewell. *Causal inference in statistics: a primer*. Wiley, 2016.

[PHL12] M. Pelikan, M. Hausschild, and F. Lobo. *Introduction to estimation of distribution algorithms*. Tech. rep. U. Missouri, 2012.

[PHR18] E. Petersen, C. Hoffmann, and P. Rostalski. "On Approximate Nonlinear Gaussian Message Passing On Factor Graphs". In: *IEEE Statistical Signal Processing Workshop (SSP)*. 2018.

[Phu+18] M. Phuong, M. Welling, N. Kushman, R. Tomioka, and S. Nowozin. "The Mutual Autoencoder: Controlling Information in Latent Code Representations". In: *Arxiv* (2018).

[Pir+13] M. Pirotta, M. Restelli, A. Pecorino, and D. Calandriello. "Safe Policy Iteration". In: *ICML*. 3. 2013, pp. 307–317.

[PJD21] Y. Petetin, Y. Janati, and F. Desbouvries. "Structured Variational Bayesian Inference for Gaussian State-Space Models With Regime Switching". In: *IEEE Signal Process. Lett.* 28 (2021), pp. 1953–1957.

[PJS17] J. Peters, D. Janzing, and B. Schölkopf. *Elements of Causal Inference: Foundations and Learning Algorithms (Adaptive Computation and Machine Learning series)*. The MIT Press, 2017.

[PKP21] A. Plaat, W. Kosters, and M. Preuss. "High-Accuracy Model-Based Reinforcement Learning, a Survey". In: (2021). arXiv: 2107.08241 [cs.LG].

[PL03] M. A. Paskin and G. D. Lawrence. *Junction Tree Algorithms for Solving Sparse Linear Systems*. Tech. rep. UCB/CSD-03-1271. UC Berkeley, 2003.

[Pla00] J. Platt. "Probabilities for SV machines". In: *Advances in Large Margin Classifiers*. Ed. by A. Smola, P. Bartlett, B. Schoelkopf, and D. Schuurmans. MIT Press, 2000.

[Pla18] E. Plaut. "From Principal Subspaces to Principal Components with Linear Autoencoders". In: *ArXiv* abs/1804.10253 (2018).

[Ple+18] G. Pleiss, J. R. Gardner, K. Q. Weinberger, and A. G. Wilson. "Constant-Time Predictive Distributions for Gaussian Processes". In: *International Conference on Machine Learning*. 2018.

[Plu+20] G. Plumb, M. Al-Shedivat, Á. A. Cabrera, A. Perer, E. Xing, and A. Talwalkar. "Regularizing black-box models for improved interpretability". In: *Advances in Neural Information Processing Systems* 33 (2020).

[PM18a] N. Papernot and P. McDaniel. *Deep k-Nearest Neighbors: Towards Confident, Interpretable and Robust Deep Learning*. 2018. arXiv: 1803.04765 [cs.LG].

[PM18b] J. Pearl and D. Mackenzie. *The book of why: the new science of cause and effect*. 2018.

[PMT18] G. Plumb, D. Molitor, and A. Talwalkar. "Supervised Local Modeling for Interpretability". In: *CoRR* abs/1807.02910 (2018). arXiv: 1807.02910.

[Pol+19] A. A. Pol, V. Berger, G. Cerminara, C. Germain, and M. Pierini. "Anomaly Detection With Conditional Variational Autoencoders". In: *IEEE International Conference on Machine Learning and Applications*. 2019.

[Pom89] D. Pomerleau. "ALVINN: An Autonomous Land Vehicle in a Neural Network". In: *NIPS*. 1989, pp. 305–313.

[Poo+12] D. Poole, D. Buchman, S. Natarajan, and K. Kersting. "Aggregation and Population Growth: The Relational Logistic Regression and Markov Logic Cases". In: *Statistical Relational AI workshop*. 2012.

[Poo+19a] B. Poole, S. Ozair, A. van den Oord, A. A. Alemi, and G. Tucker. "On Variational Bounds of Mutual Information". In: *ICML*. 2019.

[Poo+19b] B. Poole, S. Ozair, A. van den Oord, A. A. Alemi, and G. Tucker. "On variational lower bounds of mutual information". In: *ICML*. 2019.

[Pou04] M. Pourahmadi. *Simultaneous Modelling of Covariance Matrices: GLM, Bayesian and Nonparametric Perspectives*. Tech. rep. Northern Illinois University, 2004.

[Poy+20] R. Poyiadzi, K. Sokol, R. Santos-Rodriguez, T. De Bie, and P. Flach. "FACE: Feasible and actionable counterfactual explanations". In: *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*. 2020, pp. 344–350.

[PPC09] G. Petris, S. Petrone, and P. Campagnoli. *Dynamic linear models with R*. Springer, 2009.

[PPG91] C. S. Pomerleau, O. F. Pomerleau, and A. W. Garcia. "Biobehavioral research on nicotine use in women". In: *British Journal of Addiction* 86.5 (1991), pp. 527–531.

[PPM17] G. Papamakarios, T. Pavlakou, and I. Murray. "Masked Autoregressive Flow for Density Estimation". In: *NIPS*. 2017.

[PPS18] T. Pierrot, N. Perrin, and O. Sigaud. "First-order and second-order variants of the gradient descent in a unified framework". In: (2018). arXiv: 1810.08102 [cs.LG].

[PR03] O. Papaspiliopoulos and G. O. Roberts. "Non-Centered Parameterisations for Hierarchical Models and Data Augmentation". In: *Bayesian Statistics* 7 (2003), pp. 307–326.

[Pra+18] S. Prabhumoye, Y. Tsvetkov, R. Salakhutdinov, and A. W. Black. "Style Transfer Through Back-Translation". In: *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. 2018, pp. 866–876.

[Pre05] S. J. Press. *Applied multivariate analysis, using Bayesian and frequentist methods of inference*. Second edition. Dover, 2005.

[Pre+17a] V. Premachandran, D. Tarlow, A. L. Yuille, and D. Batra. "Empirical Minimum Bayes Risk Prediction". en. In: *IEEE PAMI* 39.1 (Jan. 2017), pp. 75–86.

Author: Kevin P. Murphy. (C) MIT Press. CC-BY-NC-ND license

[Pre+17b] O. Press, A. Bar, B. Bogin, J. Berant, and L. Wolf. "Language generation with recurrent generative adversarial networks without pre-training". In: *arXiv preprint arXiv:1706.01399* (2017).

[Pre+88] W. Press, W. Vetterling, S. Teukolosky, and B. Flannery. *Numerical Recipes in C: The Art of Scientific Computing*. Second. Cambridge University Press, 1988.

[PRG17] M. Probst, F. Rothlauf, and J. Grahl. "Scalability of using Restricted Boltzmann Machines for combinatorial optimization". In: *Eur. J. Oper. Res.* 256.2 (2017), pp. 368–383.

[Pri58] R. Price. "A useful theorem for nonlinear devices having Gaussian inputs". In: *IRE Trans. Info. Theory* 4.2 (1958), pp. 69–72.

[PS07] J. Peters and S. Schaal. "Reinforcement Learning by Reward-Weighted Regression for Operational Space Control". In: *ICML*. 2007, pp. 745–750.

[PS08a] B. A. Pearlmutter and J. M. Siskind. "Reverse-Mode AD in a Functional Framework: Lambda the Ultimate Backpropagator". In: *ACM Trans. Program. Lang. Syst.* 30.2 (2008).

[PS08b] J. Peters and S. Schaal. "Reinforcement Learning of Motor Skills with Policy Gradients". In: *Neural Networks* 21.4 (2008), pp. 682–697.

[PS12] N. G. Polson and J. G. Scott. "On the Half-Cauchy Prior for a Global Scale Parameter". en. In: *Bayesian Anal.* 7.4 (2012), pp. 887–902.

[PS17] N. G. Polson and V. Sokolov. "Deep Learning: A Bayesian Perspective". en. In: *Bayesian Anal.* 12.4 (2017), pp. 1275–1304.

[PSCP06] M. Pelikan, K. Sastry, and E. Cantú-Paz. *Scalable Optimization via Probabilistic Modeling: From Algorithms to Applications (Studies in Computational Intelligence)*. Springer-Verlag New York, Inc., 2006.

[PSD00] J. K. Pritchard, M Stephens, and P Donnelly. "Inference of population structure using multilocus genotype data". In: *Genetics* 155.2 (2000), pp. 945–959.

[PSDG14] B. Poole, J. Sohl-Dickstein, and S. Ganguli. "Analyzing noise in autoencoders and deep networks". In: *arXiv preprint arXiv:1406.1831* (2014).

[PSM19] G. Papamakarios, D. Sterratt, and I. Murray. "Sequential Neural Likelihood: Fast Likelihood-free Inference with Autoregressive Flows". In: *AISTATS*. 2019.

[PSS00] D. Precup, R. S. Sutton, and S. P. Singh. "Eligibility Traces for Off-Policy Policy Evaluation". In: *ICML*. ICML '00. Morgan Kaufmann Publishers Inc., 2000, pp. 759–766.

[PT13] S. Patterson and Y. W. Teh. "Stochastic Gradient Riemannian Langevin Dynamics on the Probability Simplex". In: *NIPS*. 2013.

[PT87] C. Papadimitriou and J. Tsitsiklis. "The complexity of Markov decision processes". In: *Mathematics of Operations Research* 12.3 (1987), pp. 441–450.

[PT94] P. Paatero and U. Tapper. "Positive Matrix Factorization: A Non-negative Factor Model with Optimal Utilization of Error Estimates of Data Values". In: *Environmetrics* 5 (1994), pp. 111–126.

[PTD20] A. Prabhu, P. H. S. Torr, and P. K. Dokania. "GDumb: A simple approach that questions our progress in continual learning". In: *ECCV*. Lecture notes in computer science. Springer International Publishing, 2020, pp. 524–540.

[Put94] M. L. Puterman. *Markov Decision Processes: Discrete Stochastic Dynamic Programming*. Wiley, 1994.

[PVC19] R. Prenger, R. Valle, and B. Catanzaro. "WaveGLOW: A flow-based generative network for speech synthesis". In: *Proceedings of the 2019 IEEE International Conference on Acoustics, Speech and Signal Processing*. IEEE. 2019, pp. 3617–3621.

[PW05] S. Parise and M. Welling. "Learning in Markov Random Fields: An Empirical Study". In: *Joint Statistical Meeting*. 2005.

[PY10] G. Papandreou and A. L. Yuille. "Gaussian sampling by local perturbations". In: *NIPS*. 2010.

[PY11] G Papandreou and A. L. Yuille. "Perturb-and-MAP random fields: Using discrete optimization to learn and sample from energy models". In: *ICCV*. Nov. 2011, pp. 193–200.

[PY14] G. Papandreou and A. Yuille. "Perturb-and-MAP Random Fields: Reducing Random Sampling to Optimization, with Applications in Computer Vision". In: *Advanced Structured Prediction*. Ed. by S. Nowozin, P. Gehler, J. Jancsary, C. Lampert. MIT Press, 2014.

[PY97] J. Pitman and M. Yor. "The two-parameter Poisson-Dirichlet distribution derived from a stable subordinator". In: *The Annals of Probability* (1997), pp. 855–900.

[QC+06] J. Quiñonero-Candela, C. E. Rasmussen, F. Sinz, O. Bousquet, and B. Schölkopf. "Evaluating Predictive Uncertainty Challenge". In: *Machine Learning Challenges. Evaluating Predictive Uncertainty, Visual Object Classification, and Recognising Tectual Entailment*. Lecture Notes in Computer Science. Springer Berlin Heidelberg, 2006, pp. 1–27.

[QC+08] J. Quinonero-Candela, M. Sugiyama, A. Schwaighofer, and N. D. Lawrence, eds. *Dataset Shift in Machine Learning*. en. The MIT Press, 2008.

[QCR05] J. Quinonero-Candela and C. Rasmussen. "A unifying view of sparse approximate Gaussian process regression". In: *JMLR* 6.3 (2005), pp. 1939–1959.

[Qin+20] C. Qin, Y. Wu, J. T. Springenberg, A. Brock, J. Donahue, T. P. Lillicrap, and P. Kohli. "Training Generative Adversarial Networks by Solving Ordinary Differential Equations". In: *arXiv preprint arXiv:2010.15040* (2020).

[Qu+21] H. Qu, H. Rahmani, L. Xu, B. Williams, and J. Liu. "Recent Advances of Continual Learning in Computer Vision: An Overview". In: (2021). arXiv: 2109.11369 [cs.CV].

[Qua+07] A. Quattoni, S. Wang, L.-P. Morency, M. Collins, and T. Darrell. "Hidden conditional random fields". In: *IEEE PAMI* 29.10 (2007), pp. 1848–1852.

[Que98] M. Queyranne. "Minimizing symmetric submodular functions". In: *Math. Programming* 82 (1998), pp. 3–12.

[QZW19] Y. Qiu, L. Zhang, and X. Wang. "Unbiased Contrastive Divergence Algorithm for Training Energy-Based Latent Variable Models". In: *ICLR*. 2019.

[RA13] O. Rippel and R. P. Adams. "High-dimensional probability estimation with deep density models". In: *ArXiv Preprint arXiv:1302.5125* (2013).

[Rab89] L. R. Rabiner. "A Tutorial on Hidden Markov Models and Selected Applications in Speech Recognition". In: *Proc. of the IEEE* 77.2 (1989), pp. 257–286.

[Rad+18] A. Radford, K. Narasimhan, T. Salimans, and I. Sutskever. *Improving Language Understanding by Generative Pre-Training*. Tech. rep. OpenAI, 2018.

[Rad+19] A. Radford, J. Wu, R. Child, D. Luan, D. Amodei, and I. Sutskever. *Language Models are Unsupervised Multitask Learners*. Tech. rep. OpenAI, 2019.

[Rad+21] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, et al. "Learning transferable visual models from natural language supervision". In: *arXiv preprint arXiv:2103.00020* (2021).

[Raf+20a] C. Raffel, N. Shazeer, A. Roberts, K. Lee, S. Narang, M. Matena, Y. Zhou, W. Li, and P. J. Liu. "Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer". In: *JMLR* (2020).

[Raf+20b] C. Raffel, N. M. Shazeer, A. Roberts, K. Lee, S. Narang, M. Matena, Y. Zhou, W. Li, and P. J. Liu. "Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer". In: *ArXiv* abs/1910.10683 (2020).

[RAG04] B. Ristic, S. Arulampalam, and N. Gordon. *Beyond the Kalman Filter: Particle Filters for Tracking Applications*. Artech House Radar Library, 2004.

[Rag+17] M. Raghu, J. Gilmer, J. Yosinski, and J. Sohl-Dickstein. "Svcca: Singular vector canonical correlation analysis for deep learning dynamics and interpretability". In: *Advances in Neural Information Processing Systems*. 2017, pp. 6076–6085.

[Rag+19] M. Raghu, C. Zhang, J. Kleinberg, and S. Bengio. "Transfusion: Understanding transfer learning for medical imaging". In: *NIPS*. 2019, pp. 3347–3357.

[Rag+21] M. Raghu, T. Unterthiner, S. Kornblith, C. Zhang, and A. Dosovitskiy. "Do Vision Transformers See Like Convolutional Neural Networks?" In: *NIPS*. 2021.

[Rai+18a] T. Rainforth, A. R. Kosiorek, T. A. Le, C. J. Maddison, M. Igl, F. Wood, and Y. W. Teh. "Tighter Variational Bounds are Not Necessarily Better". In: *ICML*. 2018.

[Rai+18b] M. Raitoharju, L. Svensson, Á. F. García-Fernández, and R. Piché. "Damped Posterior Linearization Filter". In: *IEEE Signal Process. Lett.* 25.4 (2018).

[Rai+20] T. Rainforth, A. Golinski, F. Wood, and S. Zaidi. "Target–Aware Bayesian Inference: How to Beat Optimal Conventional Estimators". In: *JMLR* 21.88 (2020), pp. 1–54.

[Rai68] H. Raiffa. *Decision Analysis*. Addison Wesley, 1968.

[Rak+08] A. Rakotomamonjy, F. Bach, S. Canu, and Y. Grandvalet. "SimpleMKL". In: *JMLR* 9 (2008), pp. 2491–2521.

[Ram+21a] A. Ramesh, M. Pavlov, G. Goh, S. Gray, C. Voss, A. Radford, M. Chen, and I. Sutskever. "Zero-Shot Text-to-Image Generation". In: (2021). arXiv: 2102.12092 [cs.CV].

[Ram+21b] A. Ramesh, M. Pavlov, G. Goh, S. Gray, C. Voss, A. Radford, M. Chen, and I. Sutskever. "Zero-shot text-to-image generation". In: *International Conference on Machine Learning*. PMLR. 2021, pp. 8821–8831.

[Ram+22] A. Ramesh, P. Dhariwal, A. Nichol, C. Chu, and M. Chen. "Hierarchical Text-Conditional Image Generation with CLIP Latents". In: (Apr. 2022). arXiv: 2204.06125 [cs.CV].

[Ran+06] M. Ranzato, C. S. Poultney, S. Chopra, and Y. LeCun. "Efficient Learning of Sparse Representations with an Energy-Based Model". In: *NIPS*. 2006.

[Ran16] R. Ranganath. "Hierarchical Variational Models". In: *ICML*. 2016.

[Ran+18] S. S. Rangapuram, M. W. Seeger, J. Gasthaus, L. Stella, Y. Wang, and T. Januschowski. "Deep State Space Models for Time Series Forecasting". In: *NIPS*. Curran Associates, Inc., 2018, pp. 7796–7805.

[Rao10] A. V. Rao. "A Survey of Numerical Methods for Optimal Control". In: *Adv. Astronaut. Sci.* 135.1 (2010).

[Rao99] R. P. Rao. "An optimal estimation approach to visual perception and learning". en. In: *Vision Res.* 39.11 (1999), pp. 1963–1989.

[Ras00] C. Rasmussen. "The Infinite Gaussian Mixture Model". In: *NIPS*. 2000.

[Ras+15] A. Rasmus, H. Valpola, M. Honkala, M. Berglund, and T. Raiko. *Semi-Supervised Learning with Ladder Networks*. 2015. arXiv: 1507.02672 [cs.NE].

[Rat+09] M. Rattray, O. Stegle, K. Sharp, and J. Winn. "Inference algorithms and learning theory for Bayesian sparse factor analysis". In: *Proc. Intl. Workshop on Statistical-Mechanical Informatics*. 2009.

[Rav+18] S. Ravuri, S. Mohamed, M. Rosca, and O. Vinyals. "Learning Implicit Generative Models with the Method of Learned Moments". In: *International Conference on Machine Learning*. 2018, pp. 4314–4323.

[RB16] G. P. Rigby BR. "The Efficacy of Equine-Assisted Activities and Therapies on Improving Physical Function." In: *J Altern Complement Med.* (2016).

[RBB18a] H. Ritter, A. Botev, and D. Barber. "A Scalable Laplace Approximation for Neural Networks". In: *ICLR*. 2018.

[RBB18b] H. Ritter, A. Botev, and D. Barber. "Online Structured Laplace Approximations for Overcoming Catastrophic Forgetting". In: *NIPS*. Curran Associates, Inc., 2018, pp. 3738–3748.

[RBS16] L. J. Ratliff, S. A. Burden, and S. S. Sastry. "On the characterization of local Nash equilibria in continuous games". In: *IEEE transactions on automatic control* 61.8 (2016), pp. 2301–2307.

[RBS84] J. Ramsay, J. ten Berge, and G. Styan. "Matrix correlation". In: *Psychometrika* 49.3 (1984), pp. 403–423.

[RC04] C. Robert and G. Casella. *Monte Carlo Statisical Methods*. 2nd edition. Springer, 2004.

[RC+18] M. Rojas-Carulla, B. Schölkopf, R. Turner, and J. Peters. "Invariant Models for Causal Transfer Learning". In: *Journal of Machine Learning Research* 19.36 (2018), pp. 1–34.

[RD06] M. Richardson and P. Domingos. "Markov logic networks". In: *Machine Learning* 62 (2006), pp. 107–136.

[RDV18] A. S. Ross and F. Doshi-Velez. "Improving the adversarial robustness and interpretability of deep neural networks by regularizing their input gradients".

In: *Thirty-second AAAI conference on artificial intelligence*. 2018.

[RE76] P Robert and Y Escoufier. "A unifying tool for linear multivariate statistical methods: The RV- coefficient". In: *J. R. Stat. Soc. Ser. C Appl. Stat.* 25.3 (1976), p. 257.

[Rea+19] E. Real, A. Aggarwal, Y. Huang, and Q. V. Le. "Regularized Evolution for Image Classifier Architecture Search". In: *AAAI*. 2019.

[Rec19] B. Recht. "A Tour of Reinforcement Learning: The View from Continuous Control". In: *Annual Review of Control, Robotics, and Autonomous Systems* 2 (2019), pp. 253–279.

[Ree+17] S. Reed, A. van den Oord, N. Kalchbrenner, S. G. Colmenarejo, Z. Wang, D. Belov, and N. de Freitas. "Parallel Multiscale Autoregressive Density Estimation". In: (2017). arXiv: 1703.03664 [cs.CV].

[Rei+10] J. Reisinger, A. Waters, B. Silverthorn, and R. Mooney. "Spherical topic models". In: *ICML*. 2010.

[Rei13] S. Reich. "A Nonparametric Ensemble Transform Method for Bayesian Inference". In: *SIAM J. Sci. Comput.* 35.4 (2013), A2013–A2024.

[Rei16] P. C. Reiss. "Just How Sensitive are Instrumental Variable Estimates?" In: *Foundations and Trends in Accounting* 10.2-4 (2016).

[Rei22] Reinforcement Learning:Theory and Algorithms. *Alekh Agarwal; Nan Jiang; Sham M. Kakade; Wen Sun*. 2022.

[Rei+22] P. Reizinger, L. Gresele, J. Brady, J. von Kügelgen, D. Zietlow, B. Scholkopf, G. Martius, W. Brendel, and M. Besserve. "Embrace the Gap: VAEs Perform Independent Mechanism Analysis". In: 2022.

[Ren+19] J. Ren, P. J. Liu, E. Fertig, J. Snoek, R. Poplin, M. A. DePristo, J. V. Dillon, and B. Lakshminarayanan. "Likelihood Ratios for Out-of-Distribution Detection". In: *NIPS*. 2019.

[Rén61] A. Rényi. "On Measures of Entropy and Information". en. In: *Proceedings of the Fourth Berkeley Symposium on Mathematical Statistics and Probability, Volume 1: Contributions to the Theory of Statistics*. The Regents of the University of California, 1961.

[RFB15] O. Ronneberger, P. Fischer, and T. Brox. "U-Net: Convolutional Networks for Biomedical Image Segmentation". In: *MICCAI (Intl. Conf. on Medical Image Computing and Computer Assisted Interventions)*. 2015.

[RG17] M Roth and F Gustafsson. "Computation and visualization of posterior densities in scalar nonlinear and non-Gaussian Bayesian filtering and smoothing problems". In: *ICASSP*. 2017, pp. 4686–4690.

[RGB11] S. Ross, G. J. Gordon, and D. Bagnell. "A Reduction of Imitation Learning and Structured Prediction to No-Regret Online Learning". In: *AISTATS*. 2011, pp. 627–635.

[RGB14] R. Ranganath, S. Gerrish, and D. M. Blei. "Black Box Variational Inference". In: *AISTATS*. 2014.

[RGL19] S. Rabanser, S. Günnemann, and Z. C. Lipton. "Failing Loudly: An Empirical Study of Methods for Detecting Dataset Shift". In: *NIPS*. 2019.

[RH05] H. Rue and L. Held. *Gaussian Markov Random Fields: Theory and Applications*. Vol. 104. Monographs on Statistics and Applied Probability. London: Chapman & Hall, 2005.

[RHDV17] A. S. Ross, M. C. Hughes, and F. Doshi-Velez. "Right for the right reasons: Training differentiable models by constraining their explanations". In: *IJCAI* (2017).

[RHG16] D. Ritchie, P. Horsfall, and N. D. Goodman. "Deep Amortized Inference for Probabilistic Programs". In: (2016). arXiv: 1610.05735 [cs.AI].

[RHS22] S. Rissanen, M. Heinonen, and A. Solin. "Generative Modelling With Inverse Heat Dissipation". In: (June 2022). arXiv: 2206.13397 [cs.CV].

[RHW86a] D. Rumelhart, G. Hinton, and R. Williams. "Learning internal representations by error propagation". In: *Parallel Distributed Processing: Explorations in the Microstructure of Cognition*. Ed. by D. Rumelhart, J. McClelland, and the PDD Research Group. MIT Press, 1986.

[RHW86b] D. Rumelhart, G. E. Hinton, and R. J. Williams. "Learning representations by back-propagating errors". In: *Nature* 323 (1986), pp. 533–536.

[Ric03] T. Richardson. "Markov properties for acyclic directed mixed graphs". In: *Scandinavian J. of Statistics* 30 (2003), pp. 145–157.

[Ric95] J. Rice. *Mathematical statistics and data analysis*. 2nd edition. Duxbury, 1995.

[Rif+11] S. Rifai, P. Vincent, X. Muller, X. Glorot, and Y. Bengio. "Contractive auto-encoders: Explicit invariance during feature extraction". In: *Icml*. 2011.

[Rip05] B. D. Ripley. *Spatial statistics*. Vol. 575. John Wiley & Sons, 2005.

[Rip77] B. D. Ripley. "Modelling spatial patterns". In: *Journal of the Royal Statistical Society: Series B (Methodological)* 39.2 (1977), pp. 172–192.

[Ris+08] I. Rish, G. Grabarnik, G. Cecchi, F. Pereira, and G. Gordon. "Closed-form supervised dimensionality reduction with generalized linear models". In: *ICML*. 2008.

[Riv87] R. L. Rivest. "Learning decision lists". In: *Machine learning* 2.3 (1987), pp. 229–246.

[RK04] R. Rubinstein and D. Kroese. *The Cross-Entropy Method: A Unified Approach to Combinatorial Optimization, Monte-Carlo Simulation, and Machine Learning*. Springer-Verlag, 2004.

[RL17] S. Ravi and H. Larochelle. "Optimization as a Model for Few-Shot Learning". In: *ICLR*. 2017.

[RM15] D. J. Rezende and S. Mohamed. "Variational Inference with Normalizing Flows". In: *ICML*. 2015.

[RMB08] N. L. Roux, P.-A. Manzagol, and Y. Bengio. "Topmoumoute Online Natural Gradient Algorithm". In: *NIPS*. 2008, pp. 849–856.

[RMC09] H. Rue, S. Martino, and N. Chopin. "Approximate Bayesian Inference for Latent Gaussian Models Using Integrated Nested Laplace Approximations". In: *J. of Royal Stat. Soc. Series B* 71 (2009), pp. 319–392.

[RMC15] A. Radford, L. Metz, and S. Chintala. "Unsupervised Representation Learning with Deep Convolutional Generative Adversarial Networks". In: *arXiv* (2015).

[RMC16a] A. Radford, L. Metz, and S. Chintala. "Unsupervised Representation Learning with Deep Convolutional Generative Adversarial Networks". In: *ICLR*. 2016.

[RMC16b] A. Radford, L. Metz, and S. Chintala. "Unsupervised Representation Learning with Deep Convolutional Generative Adversarial Networks". In: *CoRR* abs/1511.06434 (2016).

[RMK21] G. Roeder, L. Metz, and D. P. Kingma. "On Linear Identifiability of Learned Representations". In: *ArXiv* abs/2007.00810 (2021).

[RMW14a] D. Rezende, S. Mohamed, and D. Wierstra. "Stochastic backpropagation and approximate inference in deep generative models". In: *ICML*. 2014.

[RMW14b] D. J. Rezende, S. Mohamed, and D. Wierstra. "Stochastic Backpropagation and Approximate Inference in Deep Generative Models". In: *ICML*. Ed. by E. P. Xing and T. Jebara. Vol. 32. Proceedings of Machine Learning Research. PMLR, 2014, pp. 1278–1286.

[RN02] S. Russell and P. Norvig. *Artificial Intelligence: A Modern Approach*. 2nd edition. Prentice Hall, 2002.

[RN10] S. Russell and P. Norvig. *Artificial Intelligence: A Modern Approach*. 3rd edition. Prentice Hall, 2010.

[RN19] S. Russell and P. Norvig. *Artificial Intelligence: A Modern Approach*. 4th edition. Prentice Hall, 2019.

[RN94] G. A. Rummery and M Niranjan. *On-Line Q-Learning Using Connectionist Systems*. Tech. rep. Cambridge Univ. Engineering Dept., 1994.

[RN95] S. Russell and P. Norvig. *Artificial Intelligence: A Modern Approach*. Prentice Hall, 1995.

[RNA22] L. Regenwetter, A. H. Nobari, and F. Ahmed. "Deep Generative Models in Engineering Design: A Review". In: *J. Mech. Des.* (2022).

[Rob07] C. P. Robert. *The Bayesian Choice: From Decision-Theoretic Foundations to Computational Implementation*. en. 2nd edition. Springer Verlag, New York, 2007.

[Rob+13] S Roberts, M Osborne, M Ebden, S Reece, N Gibson, and S Aigrain. "Gaussian processes for time-series modelling". en. In: *Philos. Trans. A Math. Phys. Eng. Sci.* 371.1984 (2013), p. 20110550.

[Rob+18] C. P. Robert, V. Elvira, N. Tawn, and C. Wu. "Accelerating MCMC Algorithms". In: (2018). arXiv: 1804.02719 [stat.CO].

[Rob+21] J. Robinson, C.-Y. Chuang, S. Sra, and S. Jegelka. "Contrastive Learning with Hard Negative Samples". In: *ArXiv* abs/2010.04592 (2021).

[Rob63] L. G. Roberts. "Machine Perception of Three-Dimensional Solids". In: *Outstanding Dissertations in the Computer Sciences*. 1963.

[Rob86] J. Robins. "A new approach to causal inference in mortality studies with a sustained exposure period—application to control of the healthy worker survivor effect". In: *Mathematical Modelling* 7.9 (1986), pp. 1393–1512.

[Rob95a] C. Robert. "Simulation of truncated normal distributions". In: *Statistics and computing* 5 (1995), pp. 121–125.

[Rob95b] A. Robins. "Catastrophic Forgetting, Rehearsal and Pseudorehearsal". In: *Conn. Sci.* 7.2 (1995), pp. 123–146.

[Rod14] J. Rodu. "Spectral estimation of hidden Markov models". PhD thesis. U. Penn, 2014.

[RÖG13] M. Roth, E. Özkan, and F. Gustafsson. "A Student's t filter for heavy tailed process and measurement noise". In: *ICASSP*. 2013, pp. 5770–5774.

[Roh21] D. Rohde. "Causal Inference, is just Inference: A beautifully simple idea that not everyone accepts". In: *I (Still) Can't Believe It's Not Better! NeurIPS 2021 Workshop*. 2021.

[Ros10] P. Rosenbaum. *Design of Observational Studies*. 2010.

[Ros+21] M. Rosca, Y. Wu, B. Dherin, and D. G. Barrett. "Discretization Drift in Two-Player Games". In: (2021).

[Ros+22] C. Rosato, L. Devlin, V. Beraud, P. Horridge, T. B. Schön, and S. Maskell. "Efficient Learning of the Parameters of Non-Linear Models Using Differentiable Resampling in Particle Filters". In: *IEEE Trans. Signal Process.* 70 (2022), pp. 3676–3692.

[Ros22] C. Ross. *AI gone astray: How subtle shifts in patient data send popular algorithms reeling, undermining patient safety*. en. https://www.statnews.com/2022/02/28/sepsis-hospital-algorithms-data-shift/. Accessed: 2022-3-2. 2022.

[Rot+17] M. Roth, G. Hendeby, C. Fritsche, and F. Gustafsson. "The Ensemble Kalman filter: a signal processing perspective". In: *EURASIP J. Adv. Signal Processing* 2017.1 (2017), p. 56.

[Rot+18] W. Roth, R. Peharz, S. Tschiatschek, and F. Pernkopf. "Hybrid generative-discriminative training of Gaussian mixture models". In: *Pattern Recognit. Lett.* 112 (Sept. 2018), pp. 131–137.

[Rot96] D. Roth. "On the hardness of approximate reasoning". In: *Artificial Intelligence* 82.1-2 (1996), pp. 273–302.

[ROV19] A. Razavi, A. van den Oord, and O. Vinyals. "Generating diverse high resolution images with VA-VAE-2". In: *NIPS*. 2019.

[Row97] S. Roweis. "EM algorithms for PCA and SPCA". In: *NIPS*. 1997.

[Roy+21] N. Roy et al. "From Machine Learning to Robotics: Challenges and Opportunities for Embodied Intelligence". In: (Oct. 2021). arXiv: 2110.15245 [cs.RO].

[RPC19] Y. Romano, E. Patterson, and E. J. Candès. "Conformalized Quantile Regression". In: *NIPS*. 2019.

[RPH21] A. Robey, G. J. Pappas, and H. Hassani. *Model-Based Domain Generalization*. 2021. arXiv: 2102.11436 [stat.ML].

[RR01a] A. Rao and K. Rose. "Deterministically Annealed Design of Hidden Markov Model Speech Recognizers". In: *IEEE Trans. on Speech and Audio Proc.* 9.2 (2001), pp. 111–126.

[RR01b] G. Roberts and J. Rosenthal. "Optimal scaling for various Metropolis-Hastings algorithms". In: *Statistical Science* 16 (2001), pp. 351–367.

[RR08] A. Rahimi and B. Recht. "Random Features for Large-Scale Kernel Machines". In: *NIPS*. Curran Associates, Inc., 2008, pp. 1177–1184.

[RR09] A. Rahimi and B. Recht. "Weighted Sums of Random Kitchen Sinks: Replacing minimization with randomization in learning". In: *NIPS*. Curran Associates, Inc., 2009, pp. 1313–1320.

[RR11] T. S. Richardson and J. M. Robins. "Single World Intervention Graphs: A Primer". In: *Second UAI workshop on causal structure learning*. 2011.

[RR13] T. S. Richardson and J. M. Robins. "Single World Intervention Graphs (SWIGs): A Unification of the Counterfactual and Graphical Approaches to Causality". 2013.

[RR14] D. Russo and B. V. Roy. "Learning to Optimize via Posterior Sampling". In: *Math. Oper. Res.* 39.4 (2014), pp. 1221–1243.

[RR83] P. R. Rosenbaum and D. B. Rubin. "Assessing Sensitivity to an Unobserved Binary Covariate in an Observational Study with Binary Outcome". In: *Journal of the Royal Statistical Society. Series B (Methodological)* 45.2 (1983), pp. 212–218.

[RRR21] E. Rosenfeld, P. Ravikumar, and A. Risteski. "The Risks of Invariant Risk Minimization". In: *ICML*. 2021.

[RRS00] J. M. Robins, A. Rotnitzky, and D. O. Scharfstein. "Sensitivity analysis for selection bias and unmeasured confounding in missing data and causal inference models". In: *Statistical models in epidemiology, the environment, and clinical trials*. Springer, 2000, pp. 1–94.

[RS07] M. Raphan and E. P. Simoncelli. "Learning to be Bayesian without supervision". In: *Advances in neural information processing systems*. 2007, pp. 1145–1152.

[RS11] M. Raphan and E. P. Simoncelli. "Least squares estimation without priors or supervision". In: *Neural computation* 23.2 (2011), pp. 374–420.

[RS20] A. Rotnitzky and E. Smucler. "Efficient Adjustment Sets for Population Average Causal Treatment Effect Estimation in Graphical Models." In: *J. Mach. Learn. Res.* 21 (2020), pp. 188–1.

[RS97a] G. O. Roberts and S. K. Sahu. "Updating Schemes, Correlation Structure, Blocking and Parameterization for the Gibbs Sampler". In: *J. of Royal Stat. Soc. Series B* 59.2 (1997), pp. 291–317.

[RS97b] G. O. Roberts and S. K. Sahu. "Updating schemes, correlation structure, blocking and parameterization for the Gibbs sampler". In: *J. of Royal Stat. Soc. Series B* 59.2 (1997), pp. 291–317.

[RSC20] Y. Romano, M. Sesia, and E. J. Candès. "Classification with Valid and Adaptive Coverage". In: *NIPS*. 2020.

[RSG16a] M. T. Ribeiro, S. Singh, and C. Guestrin. "" Why should i trust you?" Explaining the predictions of any classifier". In: *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*. 2016, pp. 1135–1144.

[RSG16b] M. T. Ribeiro, S. Singh, and C. Guestrin. "Model-agnostic interpretability of machine learning". In: *arXiv preprint arXiv:1606.05386* (2016).

[RSG17] S. Rabanser, O. Shchur, and S. Günnemann. "Introduction to Tensor Decompositions and their Applications in Machine Learning". In: (2017). arXiv: 1711.10781 [stat.ML].

[RT16] S. Reid and R. Tibshirani. "Sparse regression and marginal testing using cluster prototypes". In: *Biostatistics* 17.2 (2016), pp. 364–376.

[RT82] D. B. Rubin and D. T. Thayer. "EM algorithms for ML factor analysis". In: *Psychometrika* 47.1 (1982), pp. 69–76.

[RT96] G. O. Roberts and R. L. Tweedie. "Exponential convergence of Langevin distributions and their discrete approximations". en. In: *Bernoulli* 2.4 (1996), pp. 341–363.

[RTS18] C. Riquelme, G. Tucker, and J. Snoek. "Deep Bayesian Bandits Showdown: An Empirical Comparison of Bayesian Deep Networks for Thompson Sampling". In: *ICLR*. 2018.

[RTS65] H. E. Rauch, F. Tung, and C. T. Striebel. "Maximum likelihood estimates of linear dynamic systems". In: *AIAA Journal* 3.8 (1965), pp. 1445–1450.

[Rub+20] Y. Rubanova, D. Dohan, K. Swersky, and K. Murphy. "Amortized Bayesian Optimization over Discrete Spaces". In: *UAI*. 2020.

[Rub74] D. B. Rubin. "Estimating causal effects of treatments in randomized and nonrandomized studies". In: *J. Educ. Psychol.* 66.5 (1974), pp. 688–701.

[Rub84] D. B. Rubin. "Bayesianly Justifiable and Relevant Frequency Calculations for the Applied Statistician". In: *Ann. Stat.* 12.4 (1984), pp. 1151–1172.

[Rub97] R. Y. Rubinstein. "Optimization of computer simulation models with rare events". In: *Eur. J. Oper. Res.* 99.1 (1997), pp. 89–112.

[Rud19] C. Rudin. *Stop Explaining Black Box Machine Learning Models for High Stakes Decisions and Use Interpretable Models Instead*. 2019. arXiv: 1811.10154 [stat.ML].

[Ruf+21] L. Ruff, J. R. Kauffmann, R. A. Vandermeulen, G. Montavon, W. Samek, M. Kloft, T. G. Dietterich, and K.-R. Müller. "A Unifying Review of Deep and Shallow Anomaly Detection". In: *Proc. IEEE* 109.5 (2021), pp. 756–795.

[Rus15] S. Russell. "Unifying Logic and Probability". In: *Commun. ACM* 58.7 (2015), pp. 88–97.

[Rus+16] A. A. Rusu, N. C. Rabinowitz, G. Desjardins, H. Soyer, J. Kirkpatrick, K. Kavukcuoglu, R. Pascanu, and R. Hadsell. "Progressive Neural Networks". In: (2016). arXiv: 1606.04671 [cs.LG].

[Rus+18] D. J. Russo, B. Van Roy, A. Kazerouni, I. Osband, and Z. Wen. "A Tutorial on Thompson Sampling". In: *Foundations and Trends in Machine Learning* 11.1 (2018), pp. 1–96.

[Rus+95] S. Russell, J. Binder, D. Koller, and K. Kanazawa. "Local learning in probabilistic networks with hidden variables". In: *IJCAI*. 1995.

[RV19] S. Ravuri and O. Vinyals. "Classification accuracy score for conditional generative models". In: *Advances in Neural Information Processing Systems*. 2019, pp. 12268–12279.

[RW06] C. E. Rasmussen and C. K. I. Williams. *Gaussian Processes for Machine Learning*. MIT Press, 2006.

[RW11] M. D. Reid and R. C. Williamson. "Information, Divergence and Risk for Binary Experiments". In: *Journal of Machine Learning Research* 12.3 (2011).

[RW15] D. Rosenbaum and Y. Weiss. "The Return of the Gating Network: Combining Generative Models and Discriminative Training in Natural Image Priors". In: *NIPS*. 2015, pp. 2665–2673.

[RW18] E. Richardson and Y. Weiss. "On GANs and GMMs". In: *NIPS*. 2018.

[RWD17] G. Roeder, Y. Wu, and D. Duvenaud. "Sticking the Landing: An Asymptotically Zero-Variance Gradient Estimator for Variational Inference". In: *NIPS*. 2017.

[RY21] D. Roberts and S. Yaida. *The Principles of Deep Learning Theory: An Effective Theory Approach to Understanding Neural Network*. 2021.

[Ryc+19] B. Rychalska, D. Basaj, A. Gosiewska, and P. Biecek. "Models in the Wild: On Corruption Robustness of Neural NLP Systems". In: *International Conference on Neural Information Processing (ICONIP)*. Springer International Publishing, 2019, pp. 235–247.

[Ryu+20] M. Ryu, Y. Chow, R. Anderson, C. Tjandraatmadja, and C. Boutilier. "CAQL: Continuous Action Q-Learning". In: *ICLR*. 2020.

[RZL17] P. Ramachandran, B. Zoph, and Q. V. Le. "Searching for Activation Functions". In: (2017). arXiv: 1710.05941 [cs.NE].

[SA19] F. Schafer and A. Anandkumar. "Competitive gradient descent". In: *NIPS*. 2019, pp. 7625–7635.

[Sac+05] K. Sachs, O. Perez, D. Pe'er, D. Lauffenburger, and G. Nolan. "Causal Protein-Signaling Networks Derived from Multiparameter Single-Cell Data". In: *Science* 308 (2005).

[SAC17] J. Schulman, P. Abbeel, and X. Chen. *Equivalence Between Policy Gradients and Soft Q-Learning*. arXiv:1704.06440. 2017.

[Sag+20] S. Sagawa, P. W. Koh, T. B. Hashimoto, and P. Liang. "Distributionally Robust Neural Networks for Group Shifts: On the Importance of Regularization for Worst-Case Generalization". In: *ICLR*. 2020.

[Sah+21] C. Saharia, W. Chan, H. Chang, C. A. Lee, J. Ho, T. Salimans, D. J. Fleet, and M. Norouzi. "Palette: Image-to-Image Diffusion Models". In: (Nov. 2021). arXiv: 2111.05826 [cs.CV].

[Sah+22] C. Saharia et al. "Photorealistic Text-to-Image Diffusion Models with Deep Language Understanding". In: (May 2022). arXiv: 2205.11487 [cs.CV].

[Sai+20] M. Saito, S. Saito, M. Koyama, and S. Kobayashi. "Train Sparsely, Generate Densely: Memory-Efficient Unsupervised Training of High-Resolution Temporal GAN". In: *International Journal of Computer Vision* 128 (2020), pp. 2586–2606.

[Saj+18] M. S. Sajjadi, O. Bachem, M. Lucic, O. Bousquet, and S. Gelly. "Assessing generative models via precision and recall". In: *Proceedings of the 32nd International Conference on Neural Information Processing Systems*. 2018, pp. 5234–5243.

[Sal16] T. Salimans. "A Structured Variational Auto-encoder for Learning Deep Hierarchies of Sparse Features". In: (2016). arXiv: 1602.08734 [stat.ML].

[Sal+16] T. Salimans, I. Goodfellow, W. Zaremba, V. Cheung, A. Radford, and X. Chen. "Improved Techniques for Training GANs". In: (2016). arXiv: 1606.03498 [cs.LG].

[Sal+17a] M. Salehi, A. Karbasi, D. Scheinost, and R. T. Constable. "A Submodular Approach to Create Individualized Parcellations of the Human Brain". In: *Medical Image Computing and Computer Assisted Intervention - MICCAI 2017*. Ed. by M. Descoteaux, L. Maier-Hein, A. Franz, P. Jannin, D. L. Collins, and S. Duchesne. Cham: Springer International Publishing, 2017, pp. 478–485.

[Sal+17b] T. Salimans, J. Ho, X. Chen, and I. Sutskever. "Evolution Strategies as a Scalable Alternative to Reinforcement Learning". In: (2017). arXiv: 1703.03864 [stat.ML].

[Sal+17c] T. Salimans, A. Karpathy, X. Chen, and D. P. Kingma. "PixelCNN++: Improving the PixelCNN with Discretized Logistic Mixture Likelihood and Other Modifications". In: *ICLR*. 2017.

[Sal+19a] D. Salinas, M. Bohlke-Schneider, L. Callot, R. Medico, and J. Gasthaus. "High-Dimensional Multivariate Forecasting with Low-Rank Gaussian Copula Processes". In: *NIPS*. 2019.

[Sal+19b] D. Salinas, V. Flunkert, J. Gasthaus, and T. Januschowski. "DeepAR: Probabilistic forecasting with autoregressive recurrent networks". In: *International Journal of Forecasting* (2019).

[Sal+20] H. Salman, A. Ilyas, L. Engstrom, A. Kapoor, and A. Madry. "Do Adversarially Robust ImageNet Models Transfer Better?" In: *arXiv preprint arXiv:2007.08489* (2020).

[Sal+21] M. Salehi, H. Mirzaei, D. Hendrycks, Y. Li, M. H. Rohban, and M. Sabokrou. "A Unified Survey on Anomaly, Novelty, Open-Set, and Out-of-Distribution Detection: Solutions and Future Challenges". In: (2021). arXiv: 2110.14051 [cs.CV].

[Sam68] F. Sampson. "A Novitiate in a Period of Change: An Experimental and Case Study of Social Relationships". PhD thesis. Cornell, 1968.

[Sam74] P. A. Samuelson. "Complementarity: An essay on the 40th anniversary of the Hicks-Allen revolution in demand theory". In: *Journal of Economic literature* 12.4 (1974), pp. 1255–1289.

[San17] R. Santana. "Gray-box optimization and factorized distribution algorithms: where two worlds collide". In: (2017). arXiv: 1707.03093 [cs.NE].

[San+17] A. Santoro, D. Raposo, D. G. Barrett, M. Malinowski, R. Pascanu, P. Battaglia, and T. Lillicrap. "A simple neural network module for relational reasoning". In: *Advances in neural information processing systems*. 2017, pp. 4967–4976.

[Sar08] S. Sarkka. "Unscented Rauch–Tung–Striebel Smoother". In: *IEEE Trans. Automat. Contr.* 53.3 (Apr. 2008), pp. 845–849.

[Sar13] S. Sarkka. *Bayesian Filtering and Smoothing*. Cambridge University Press, 2013.

[Sar18] H. Sarin. "Playing a game of GANstruction". In: *The Gradient* (2018).

[Say+19] R. Sayres, S. Xu, T Saensuksopa, M. Le, and D. R. Webster. "Assistance from a deep learning system improves diabetic retinopathy assessment in optometrists". In: *Investigative Ophthalmology & Visual Science* 60.9 (2019), pp. 1433–1433.

[SB01] A. J. Smola and P. L. Bartlett. "Sparse Greedy Gaussian Process Regression". In: *NIPS*. Ed. by T. K. Leen, T. G. Dietterich, and V Tresp. MIT Press, 2001, pp. 619–625.

[SB12] J. Staines and D. Barber. "Variational Optimization". In: (2012). arXiv: 1212.4507 [stat.ML].

[SB13] J Staines and D Barber. "Optimization by Variational Bounding". In: *European Symposium on ANNs*. elen.ucl.ac.be, 2013.

[SB18] R. Sutton and A. Barto. *Reinforcement learning: an introduction (2nd edn)*. MIT Press, 2018.

[SBG07] S. Siddiqi, B. Boots, and G. Gordon. "A constraint generation approach to learning stable linear dynamical systems". In: *NIPS*. 2007.

[SBP17] Y Sun, P Babu, and D. P. Palomar. "Majorization-Minimization Algorithms in Signal Processing, Communications, and Machine Learning". In: *IEEE Trans. Signal Process.* 65.3 (2017), pp. 794–816.

Author: Kevin P. Murphy. (C) MIT Press. CC-BY-NC-ND license

[SC13] C. Schäfer and N. Chopin. "Sequential Monte Carlo on large binary sampling spaces". In: *Stat. Comput.* 23.2 (2013), pp. 163–184.

[SC86] R. Smith and P. Cheeseman. "On the Representation and Estimation of Spatial Uncertainty". In: *Intl. J. Robotics Research* 5.4 (1986), pp. 56–68.

[SC90] R. Schwarz and Y. Chow. "The n-best algorithm: an efficient and exact procedure for finding the n most likely hypotheses". In: *ICASSP*. 1990.

[Sca21] S. Scardapane. *Lecture 8: Beyond single-task supervised learning*. 2021.

[Sch00] A. Schrijver. "A combinatorial algorithm minimizing submodular functions in strongly polynomial time". In: *Journal of Combinatorial Theory, Series B* 80.2 (2000), pp. 346–355.

[Sch02] N. N. Schraudolph. "Fast Curvature Matrix-Vector Products for Second-Order Gradient Descent". In: *Neural Computation* 14 (2002).

[Sch04] A. Schrijver. *Combinatorial Optimization*. Springer, 2004.

[Sch+12a] B. Schoelkopf, D. Janzing, J. Peters, E. Sgouritsa, K. Zhang, and J. Mooij. "On Causal and Anticausal Learning". In: *ICML*. 2012.

[Sch+12b] B. Schölkopf, D. Janzing, J. Peters, E. Sgouritsa, K. Zhang, and J. Mooij. "On causal and anticausal learning". In: *Proceedings of the 29th International Coference on International Conference on Machine Learning*. 2012, pp. 459–466.

[Sch14] J. Schmidhuber. *Deep Learning in Neural Networks: An Overview*. Tech. rep. 2014.

[Sch+15a] J. Schulman, N. Heess, T. Weber, and P. Abbeel. "Gradient Estimation Using Stochastic Computation Graphs". In: *NIPS*. 2015.

[Sch+15b] J. Schulman, S. Levine, P. Moritz, M. I. Jordan, and P. Abbeel. "Trust Region Policy Optimization". In: *ICML*. 2015.

[Sch+16a] T. Schaul, J. Quan, I. Antonoglou, and D. Silver. "Prioritized Experience Replay". In: *ICLR*. 2016.

[Sch+16b] J. Schulman, P. Moritz, S. Levine, M. Jordan, and P. Abbeel. "High-Dimensional Continuous Control Using Generalized Advantage Estimation". In: *ICLR*. 2016.

[Sch+17] J. Schulman, F. Wolski, P. Dhariwal, A. Radford, and O. Klimov. "Proximal Policy Optimization Algorithms". In: (2017). arXiv: 1707.06347 [cs.LG].

[Sch+18] J. Schwarz, J. Luketina, W. M. Czarnecki, A. Grabska-Barwinska, Y. W. Teh, R. Pascanu, and R. Hadsell. "Progress & Compress: A scalable framework for continual learning". In: *ICML*. 2018.

[Sch19] B. Schölkopf. "Causality for Machine Learning". In: (2019). arXiv: 1911.10500 [cs.LG].

[Sch20] J. Schmidhuber. *Planning & Reinforcement Learning with Recurrent World Models and Artificial Curiosity*. 2020.

[Sch+20] J. Schrittwieser et al. "Mastering Atari, Go, Chess and Shogi by Planning with a Learned Model". In: *Nature* (2020).

[Sch+21a] D. O. Scharfstein, R. Nabi, E. H. Kennedy, M.-Y. Huang, M. Bonvini, and M. Smid. *Semiparametric Sensitivity Analysis: Unmeasured Confounding In Observational Studies*. 2021. arXiv: 2104.08300 [stat.ME].

[Sch+21b] B. Schölkopf, F. Locatello, S. Bauer, N. R. Ke, N. Kalchbrenner, A. Goyal, and Y. Bengio. "Toward Causal Representation Learning". In: *Proc. IEEE* 109.5 (2021), pp. 612–634.

[Sch+21c] B. Schölkopf, F. Locatello, S. Bauer, N. R. Ke, N. Kalchbrenner, A. Goyal, and Y. Bengio. "Towards Causal Representation Learning". In: *CoRR* abs/2102.11107 (2021). arXiv: 2102.11107.

[Sch78] G. Schwarz. "Estimating the dimension of a model". In: *Annals of Statistics* 6.2 (1978), pp. 461–464.

[Sco02] S Scott. "Bayesian methods for hidden Markov models: Recursive computing in the 21st century." In: *JASA* (2002).

[Sco09] S. Scott. "Data augmentation, frequentist estimation, and the Bayesian analysis of multinomial logit models". In: *Statistical Papers* (2009).

[Sco10] S. Scott. "A modern Bayesian look at the multi-armed bandit". In: *Applied Stochastic Models in Business and Industry* 26 (2010), pp. 639–658.

[SCPD22] R. Sanchez-Cauce, I. Paris, and F. J. Diez. "Sum-Product Networks: A Survey". en. In: *IEEE PAMI* 44.7 (July 2022), pp. 3821–3839.

[SCS19] A. Subbaswamy, B. Chen, and S. Saria. *A Universal Hierarchy of Shift-Stable Distributions and the Tradeoff Between Stability and Performance*. 2019. arXiv: 1905.11374 [stat.ML].

[SCS22] A. Subbaswamy, B. Chen, and S. Saria. "A unifying causal framework for analyzing dataset shift-stable learning algorithms". en. In: *Journal of Causal Inference* 10.1 (Jan. 2022), pp. 64–89.

[SD12] J. Sohl-Dickstein. "The Natural Gradient by Analogy to Signal Whitening, and Recipes and Tricks for its Use". In: (2012). arXiv: 1205.1828 [cs.LG].

[SD+15a] J. Sohl-Dickstein, E. Weiss, N. Maheswaranathan, and S. Ganguli. "Deep Unsupervised Learning using Nonequilibrium Thermodynamics". In: *ICML*. 2015, pp. 2256–2265.

[SD+15b] J. Sohl-Dickstein, E. A. Weiss, N. Maheswaranathan, and S. Ganguli. "Deep Unsupervised Learning using Nonequilibrium Thermodynamics". In: *ICML*. 2015.

[SDBD11] J. Sohl-Dickstein, P. Battaglino, and M. R. DeWeese. "Minimum probability flow learning". In: *Proceedings of the 28th International Conference on International Conference on Machine Learning*. 2011, pp. 905–912.

[SE19] Y. Song and S. Ermon. "Generative Modeling by Estimating Gradients of the Data Distribution". In: *NIPS*. 2019, pp. 11895–11907.

[SE20a] J. Song and S. Ermon. "Multi-label Contrastive Predictive Coding". In: *NIPS*. 2020.

[SE20b] Y. Song and S. Ermon. "Improved Techniques for Training Score-Based Generative Models". In: *NIPS*. 2020.

[See+17] M. Seeger, S. Rangapuram, Y. Wang, D. Salinas, J. Gasthaus, T. Januschowski, and V. Flunkert. "Approximate Bayesian Inference in Linear State Space Models for Intermittent Demand Forecasting at Scale". In: (2017). arXiv: 1709.07638 [stat.ML].

[Sej20] T. J. Sejnowski. "The unreasonable effectiveness of deep learning in artificial intelligence". en. In: *PNAS* 117.48 (Dec. 2020), pp. 30033–30038.

[Sel+17] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra. "Grad-cam: Visual explanations from deep networks via gradient-based localization". In: *Proceedings of the IEEE international conference on computer vision*. 2017, pp. 618–626.

[Sel+19] A. D. Selbst, D. Boyd, S. A. Friedler, S. Venkatasubramanian, and J. Vertesi. "Fairness and Abstraction in Sociotechnical Systems". In: *Proceedings of the Conference on Fairness, Accountability, and Transparency*. FAT* '19. Atlanta, GA, USA: Association for Computing Machinery, 2019, 59–68.

[Sen+08] P. Sen, G. Namata, M. Bilgic, L. Getoor, B. Galligher, and T. Eliassi-Rad. "Collective Classification in Network Data". en. In: *AI Magazine* 29.3 (2008), pp. 93–93.

[Ser+20] J. Serrà, D. Álvarez, V. Gómez, O. Slizovskaia, J. F. Núñez, and J. Luque. "Input complexity and out-of-distribution detection with likelihood-based generative models". In: *ICLR*. 2020.

[Set12] B. Settles. "Active learning". In: *Synthesis Lectures on Artificial Intelligence and Machine Learning* 6 (2012), 1–114.

[Set94] J. Sethuraman. "A constructive definition of Dirichlet priors". In: *Statistica Sinica* (1994), pp. 639–650.

[SF08] Z. Svitkina and L. Fleischer. "Submodular approximation: Sampling-based algorithms and lower bounds". In: *FOCS*. 2008.

[SF20] K. Sokol and P. Flach. "Explainability fact sheets: a framework for systematic assessment of explainable approaches". In: *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*. 2020, pp. 56–67.

[SFB18] S. A. Sisson, Y. Fan, and M. A. Beaumont. "Overview of ABC". In: *Handbook of approximate Bayesian computation*. Chapman and Hall/CRC, 2018, pp. 3–54.

[SG05] E. Snelson and Z. Ghahramani. "Compact Approximations to Bayesian Predictive Distributions". In: *ICML*. 2005.

[SG06a] E. Snelson and Z. Ghahramani. "Sparse Gaussian processes using pseudo-inputs". In: *NIPS*. 2006.

[SG06b] E. Snelson and Z. Ghahramani. "Sparse Gaussian Processes using Pseudo-inputs". In: *Advances in Neural Information Processing Systems*. Ed. by Y. Weiss, B. Schölkopf, and J. Platt. Vol. 18. MIT Press, 2006.

[SG07] M. Steyvers and T. Griffiths. "Probabilistic topic models". In: *Latent Semantic Analysis: A Road to Meaning*. Ed. by T. Landauer, D McNamara, S. Dennis, and W. Kintsch. Laurence Erlbaum, 2007.

[SG09] R. Silva and Z. Ghahramani. "The Hidden Life of Latent Variables: Bayesian Learning with Mixed Graph Models". In: *JMLR* 10 (2009), pp. 1187–1238.

[SGF21] S. Särkkä and Á. F. García-Fernández. "Temporal Parallelization of Bayesian Filters and Smoothers". In: *IEEE Trans. Automat. Contr.* 66.1 (2021).

[SGS16] A. Sharghi, B. Gong, and M. Shah. "Query-focused extractive video summarization". In: *European Conference on Computer Vision*. Springer. 2016, pp. 3–19.

[SH07] R. Salakhutdinov and G. Hinton. "Using Deep Belief Nets to Learn Covariance Kernels for Gaussian Processes". In: *NIPS*. 2007.

[SH09] R. Salakhutdinov and G. Hinton. "Deep Boltzmann Machines". In: *AISTATS*. Vol. 5. 2009, pp. 448–455.

[SH10] R. Salakhutdinov and G. Hinton. "Replicated Softmax: an Undirected Topic Model". In: *NIPS*. 2010.

[SH22] T. Salimans and J. Ho. "Progressive Distillation for Fast Sampling of Diffusion Models". In: *ICLR*. Feb. 2022.

[SHA15] M. A. Skoglund, G. Hendeby, and D. Axehill. "Extended Kalman filter modifications based on an optimization view point". In: *2015 18th International Conference on Information Fusion (Fusion)*. July 2015, pp. 1856–1861.

[Sha+16] B Shahriari, K Swersky, Z. Wang, R. P. Adams, and N de Freitas. "Taking the Human Out of the Loop: A Review of Bayesian Optimization". In: *Proc. IEEE* 104.1 (2016), pp. 148–175.

[Sha16] L. S. Shapley. *17. A value for n-person games*. Princeton University Press, 2016.

[Sha+16] M. Sharif, S. Bhagavatula, L. Bauer, and M. K. Reiter. "Accessorize to a Crime: Real and Stealthy Attacks on State-of-the-Art Face Recognition". In: *Proceedings of the 2016 ACM SIGSAC Conference on Computer and Communications Security*. ACM, 2016, pp. 1528–1540.

[Sha+19] A. Shaikhha, A. Fitzgibbon, D. Vytiniotis, and S. Peyton Jones. "Efficient differentiable programming in a functional array-processing language". In: *Proceedings of the ACM on Programming Languages* 3.ICFP (2019), pp. 1–30.

[Sha+20] H. Shah, K. Tamuly, A. Raghunathan, P. Jain, and P. Netrapalli. "The Pitfalls of Simplicity Bias in Neural Networks". In: *NIPS*. 2020.

[Sha22] C. Shalizi. *Advanced Data Analysis from an Elementary Point of View*. Cambridge University Press, 2022.

[Sha48] C. Shannon. "A mathematical theory of communication". In: *Bell Systems Tech. Journal* 27 (1948), pp. 379–423.

[Sha98] R. Shachter. "Bayes-Ball: The Rational Pastime (for determining Irrelevance and Requisite Information in Belief Networks and Influence Diagrams)". In: *UAI*. 1998.

[She+11] C. Shen, X. Li, L. Li, and M. C. Were. "Sensitivity analysis for causal inference using inverse probability weighting". In: *Biometrical Journal* 53.5 (2011), pp. 822–837. eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/bimj.201100042.

[She+17] T. Shen, T. Lei, R. Barzilay, and T. Jaakkola. "Style transfer from non-parallel text by cross-alignment". In: *Advances in neural information processing systems* 30 (2017), pp. 6830–6841.

[She+20] T. Shen, J. Mueller, R. Barzilay, and T. Jaakkola. "Educating Text Autoencoders: Latent Representation Guidance via Denoising". In: *ICML*. 2020.

[She+21] Z. Shen, J. Liu, Y. He, X. Zhang, R. Xu, H. Yu, and P. Cui. "Towards Out-Of-Distribution Generalization: A Survey". In: (2021). arXiv: 2108.13624 [cs.LG].

[SHF15] R. Steorts, R. Hall, and S. Fienberg. "A Bayesian Approach to Graphical Record Linkage and De-duplication". In: *JASA* (2015).

[Shi00] B. Shipley. *Cause and Correlation in Biology: A User's Guide to Path Analysis, Structural Equations and Causal Inference*. Cambridge, 2000.

[Shi+21] C. Shi, D. Sridhar, V. Misra, and D. M. Blei. "On the Assumptions of Synthetic Control Methods". In: (2021). arXiv: 2112.05671 [stat.ME].

[SHM14] D. Soudry, I. Hubara, and R. Meir. "Expectation backpropagation: Parameter-free training of multilayer neural networks with continuous or discrete weights". In: *NIPS*. 2014.

[Shr+16] A. Shrikumar, P. Greenside, A. Shcherbina, and A. Kundaje. "Not just a black box: Learning important features through propagating activation differences". In: *arXiv preprint arXiv:1605.01713* (2016).

[SHS] D. Stutz, M. Hein, and B. Schiele. "Confidence-calibrated adversarial training: Generalizing to unseen attacks". In: ().

[SHS01] B. Schölkopf, R. Herbrich, and A. J. Smola. "A Generalized Representer Theorem". In: *COLT*. COLT '01/EuroCOLT '01. Springer-Verlag, 2001, pp. 416–426.

[Shu+19a] K. Shu, L. Cui, S. Wang, D. Lee, and H. Liu. "defend: Explainable fake news detection". In: *Proceedings of the 25th ACM SIGKDD international conference on knowledge discovery & data mining*. 2019, pp. 395–405.

[Shu+19b] R. Shu, Y. Chen, A. Kumar, S. Ermon, and B. Poole. "Weakly supervised disentanglement with guarantees". In: *arXiv preprint arXiv:1910.09772* (2019).

[SI00] M. Sato and S. Ishii. "On-line EM algorithm for the normalized Gaussian network". In: *Neural Computation* 12 (2000), pp. 407–432.

[Sil+14] D. Silver, G. Lever, N. Heess, T. Degris, D. Wierstra, and M. Riedmiller. "Deterministic Policy Gradient Algorithms". In: *ICML*. ICML'14. JMLR.org, 2014, pp. I–387–I–395.

[Sil+16] D. Silver et al. "Mastering the game of Go with deep neural networks and tree search". en. In: *Nature* 529.7587 (2016), pp. 484–489.

[Sil18] D. Silver. *Lecture 9L Exploration and Exploitation*. 2018.

[Sil+18] D. Silver et al. "A general reinforcement learning algorithm that masters chess, shogi, and Go through self-play". en. In: *Science* 362.6419 (2018), pp. 1140–1144.

[Sil85] B. W. Silverman. "Some Aspects of the Spline Smoothing Approach to Non-Parametric Regression Curve Fitting". In: *J. R. Stat. Soc. Series B Stat. Methodol.* 47.1 (1985), pp. 1–52.

[Sim06] D. Simon. *Optimal State Estimation: Kalman, H Infinity, and Nonlinear Approaches*. Wiley, 2006.

[Sin+00] S. Singh, T. Jaakkola, M. L. Littman, and C. Szepesvári. "Convergence Results for Single-Step On-PolicyReinforcement-Learning Algorithms". In: *MLJ* 38.3 (2000), pp. 287–308.

[Sin67] R. Sinkhorn. "Diagonal Equivalence to Matrices with Prescribed Row and Column Sums". In: *The American Mathematical Monthly* 74.4 (1967), pp. 402–405.

[SJ08] S. Shirdhonkar and D. W. Jacobs. "Approximate earth mover's distance in linear time". In: *2008 IEEE Conference on Computer Vision and Pattern Recognition*. IEEE. 2008, pp. 1–8.

[SJ15] A. Swaminathan and T. Joachims. "Batch Learning from Logged Bandit Feedback through Counterfactual Risk Minimization". In: *JMLR* 16.1 (2015), pp. 1731–1755.

[SJ95] L. Saul and M. Jordan. "Exploiting tractable substructures in intractable networks". In: *NIPS*. Vol. 8. 1995.

[SJ99] L. Saul and M. Jordan. "Mixed memory Markov models: Decomposing complex stochastic processes as mixture of simpler ones". In: *Machine Learning* 37.1 (1999), pp. 75–87.

[SJJ96] L. Saul, T. Jaakkola, and M. Jordan. "Mean Field Theory for Sigmoid Belief Networks". In: *JAIR* 4 (1996), pp. 61–76.

[SJR04] S. Singh, M. James, and M. Rudary. "Predictive state representations: A new theory for modeling dynamical systems". In: *UAI*. 2004.

[SK19] C. Shorten and T. M. Khoshgoftaar. "A survey on Image Data Augmentation for Deep Learning". en. In: *Journal of Big Data* 6.1 (2019), pp. 1–48.

[SK20] S. Singh and S. Krishnan. "Filter Response Normalization Layer: Eliminating Batch Dependence in the Training of Deep Neural Networks". In: *CVPR*. 2020.

[SK89] R. Shachter and C. R. Kenley. "Gaussian Influence Diagrams". In: *Managment Science* 35.5 (1989), pp. 527–550.

[Ski89] J. Skilling. "The eigenvalues of mega-dimensional matrices". In: *Maximum Entropy and Bayesian Methods*. Springer, 1989, pp. 455–466.

[SKM18] S. Schwöbel, S. Kiebel, and D. Marković. "Active Inference, Belief Propagation, and the Bethe Approximation". en. In: *Neural Comput.* 30.9 (2018), pp. 2530–2567.

[SKM21] M. Shanahan, C. Kaplanis, and J. Mitrović. "Encoders and Ensembles for Task-Free Continual Learning". In: (2021). arXiv: 2105.13327 [cs.LG].

[SKP15] F. Schroff, D. Kalenichenko, and J. Philbin. "Facenet: A unified embedding for face recognition and clustering". In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2015, pp. 815–823.

[SKTF18] H. Shao, A. Kumar, and P Thomas Fletcher. "The Riemannian Geometry of Deep Generative Models". In: *CVPR*. 2018, pp. 315–323.

[SKW15] T. Salimans, D. Kingma, and M. Welling. "Markov Chain Monte Carlo and Variational Inference: Bridging the Gap". In: *ICML*. 2015, pp. 1218–1226.

[SL08] A. L. Strehl and M. L. Littman. "An Analysis of Model-based Interval Estimation for Markov Decision Processes". In: *J. of Comp. and Sys. Sci.* 74.8 (2008), pp. 1309–1331.

[SL18] S. L. Smith and Q. V. Le. "A Bayesian Perspective on Generalization and Stochastic Gradient Descent". In: *ICLR*. 2018.

[SL90] D. J. Spiegelhalter and S. L. Lauritzen. "Sequential updating of conditional probabilities on directed graphical structures". In: *Networks* 20 (1990).

[Sla+20] D. Slack, S. Hilgard, E. Jia, S. Singh, and H. Lakkaraju. "Fooling lime and shap: Adversarial attacks on post hoc explanation methods". In: *Proceed-*

ings of the AAAI/ACM Conference on AI, Ethics, and Society. 2020, pp. 180–186.

[SLG17] A. Sharghi, J. S. Laurel, and B. Gong. "Query-focused video summarization: Dataset, evaluation, and a memory network based approach". In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2017, pp. 4788–4797.

[Sli19] A. Slivkins. "Introduction to Multi-Armed Bandits". In: *Foundations and Trends in Machine Learning* (2019).

[SLL09] A. L. Strehl, L. Li, and M. L. Littman. "Reinforcement Learning in Finite MDPs: PAC Analysis". In: *JMLR* 10 (2009), pp. 2413–2444.

[SLM92] B. Selman, H. Levesque, and D. Mitchell. "A New Method for Solving Hard Satisfiability Problems". In: *Proceedings of the Tenth National Conference on Artificial Intelligence*. AAAI'92. AAAI Press, 1992, pp. 440–446.

[SLW19] M. Sadinle, J. Lei, and L. Wasserman. "Least Ambiguous Set-Valued Classifiers With Bounded Error Levels". In: *JASA* 114.525 (2019), pp. 223–234.

[SM07] C. Sutton and A. McCallum. "Improved Dynamic Schedules for Belief Propagation". In: *UAI*. 2007.

[SM12] Y Saika and K Morimoto. "Generalized MAP estimation via parameter scheduling and maximizer of the posterior marginal estimate for image reconstruction using multiple halftone images". In: *12th International Conference on Control, Automation and Systems*. 2012, pp. 1285–1289.

[SMB10] H. Schulz, A. Müller, and S. Behnke. "Investigating convergence of restricted Boltzmann machine learning". In: *NIPS 2010 Workshop on Deep Learning and Unsupervised Feature Learning*. Vol. 1. 2. 2010, pp. 6–1.

[SME21] J. Song, C. Meng, and S. Ermon. "Denoising Diffusion Implicit Models". In: *ICLR*. 2021.

[SMH07] R. R. Salakhutdinov, A. Mnih, and G. E. Hinton. "Restricted Boltzmann machines for collaborative filtering". In: *ICML*. Vol. 24. 2007, pp. 791–798.

[Smi+00] G. Smith, J. F. G. de Freitas, T. Robinson, and M. Niranjan. "Speech Modelling Using Subspace and EM Techniques". In: *NIPS*. MIT Press, 2000, pp. 796–802.

[Smi+06] V. Smith, J. Yu, T. Smulders, A. Hartemink, and E. Jarvis. "Computational Inference of Neural Information Flow Networks". In: *PLOS Computational Biology* 2 (2006), pp. 1436–1439.

[Smi11] N. Smith. *Linguistic structure prediction*. Morgan Claypool, 2011.

[Smi+17] D. Smilkov, N. Thorat, B. Kim, F. Viégas, and M. Wattenberg. *SmoothGrad: removing noise by adding noise*. 2017. arXiv: 1706.03825 [cs.LG].

[Smo86] P. Smolensky. "Information processing in dynamical systems: foundations of harmony theory". In: *Parallel Distributed Processing: Explorations in the Microstructure of Cognition. Volume 1*. Ed. by D. Rumehart and J. McClelland. McGraw-Hill, 1986.

[SMS17] M. Saito, E. Matsumoto, and S. Saito. "Temporal Generative Adversarial Nets with Singular Value Clipping". In: *ICCV*. 2017.

[SMT18] M. R. U. Saputra, A. Markham, and N. Trigoni. "Visual SLAM and Structure from Motion in Dynamic Environments: A Survey". In: *ACM Computing Surveys* 51.2 (2018), pp. 1–36.

[Smy20] S. Smyl. "A hybrid method of exponential smoothing and recurrent neural networks for time series forecasting". In: *Int. J. Forecast.* 36.1 (2020), pp. 75–85.

[Sn+16] C. K. Sø nderby, T. Raiko, L. Maaløe, S. R. K. Sø nderby, and O. Winther. "Ladder Variational Autoencoders". In: *NIPS*. Curran Associates, Inc., 2016, pp. 3738–3746.

[SNM16] M. Suzuki, K. Nakayama, and Y. Matsuo. "Joint Multimodal Learning with Deep Generative Models". In: (2016). arXiv: 1611.01891 [stat.ML].

[SOB12] R. Snyder, J. K. Ord, and A. Beaumont. "Forecasting the intermittent demand for slow-moving inventories: A modelling approach". In: *Int. J. Forecast.* 28.2 (2012), pp. 485–496.

[Soh16] K. Sohn. "Improved deep metric learning with multi-class n-pair loss objective". In: *Advances in Neural Information Processing Systems*. 2016, pp. 1857–1865.

[Søn+16] C. Sønderby, T. Raiko, L. Maaløe, S. Sønderby, and O. Winther. "How to Train Deep Variational Autoencoders and Probabilistic Ladder Networks". In: *ICML*. 2016.

[Son+19] Y. Song, S. Garg, J. Shi, and S. Ermon. "Sliced Score Matching: A Scalable Approach to Density and Score Estimation". In: *Proceedings of the Thirty-Fifth Conference on Uncertainty in Artificial Intelligence, UAI 2019, Tel Aviv, Israel, July 22-25, 2019*. 2019, p. 204.

[Son+21] Y. Song, J. Sohl-Dickstein, D. P. Kingma, A. Kumar, S. Ermon, and B. Poole. "Score-Based Generative Modeling through Stochastic Differential Equations". In: *ICLR*. 2021.

[Son98] E. D. Sontag. *Mathematical Control Theory: Deterministic Finite Dimensional Systems*. 2nd. Vol. 6. Texts in Applied Mathematics. Springer, 1998.

[SOS92] H. Seung, M. Opper, and H. Sompolinsky. "Query by committee". In: *5th Annual Workshop on Computational Learning Theory*. 1992, 287–294.

[SPD92] S. Shah, F. Palmieri, and M. Datum. "Optimal filtering algorithms for fast learning in feedforward neural networks". In: *Neural Netw.* 5.5 (1992), pp. 779–787.

[Spi71] M. Spivak. *Calculus On Manifolds: A Modern Approach To Classical Theorems Of Advanced Calculus*. Westview Press; 5th edition, 1971.

[SPL20] M. B. Sariyildiz, J. Perez, and D. Larlus. "Learning visual representations with caption annotations". In: *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part VIII 16*. Springer. 2020, pp. 153–170.

[Spr+14] J. T. Springenberg, A. Dosovitskiy, T. Brox, and M. Riedmiller. "Striving for simplicity: The all convolutional net". In: *arXiv preprint arXiv:1412.6806* (2014).

[Spr+16] J. T. Springenberg, A. Klein, S. Falkner, and F. Hutter. "Bayesian Optimization with Robust Bayesian Neural Networks". In: *NIPS*. 2016, pp. 4141–4149.

[SPW18] M. H. S. Segler, M. Preuss, and M. P. Waller. "Planning chemical syntheses with deep neural networks and symbolic AI". en. In: *Nature* 555.7698 (2018), pp. 604–610.

[SPZ09] P. Schniter, L. C. Potter, and J. Ziniel. "Fast Bayesian Matching Pursuit: Model Uncertainty and

Parameter Estimation for Sparse Linear Models". In: *IEEE Trans. on Signal Processing* (2009).

[SR+14] A. Sharif Razavian, H. Azizpour, J. Sullivan, and S. Carlsson. "CNN features off-the-shelf: an astounding baseline for recognition". In: *Proceedings of the IEEE conference on computer vision and pattern recognition workshops*. 2014, pp. 806–813.

[SRG03] R. Salakhutdinov, S. T. Roweis, and Z. Ghahramani. "Optimization with EM and Expectation-Conjugate-Gradient". In: *ICML*. 2003.

[Sri+09] B. K. Sriperumbudur, K. Fukumizu, A. Gretton, B. Schölkopf, and G. R. G. Lanckriet. "On integral probability metrics, φ-divergences and binary classification". In: (2009). arXiv: 0901.2698 [cs.IT].

[Sri+10] N. Srinivas, A. Krause, S. Kakade, and M. Seeger. "Gaussian Process Optimization in the Bandit Setting: No Regret and Experimental Design". In: *ICML*. 2010, pp. 1015–1022.

[Sri+14a] N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov. "Dropout: A Simple Way to Prevent Neural Networks from Over tting". In: *JMLR* (2014).

[Sri+14b] N. Srivastava, G. E. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov. "Dropout: a simple way to prevent neural networks from overfitting". In: *J. Mach. Learn. Res.* 15 (2014), pp. 1929–1958.

[Sri+17] A. Srivastava, L. Valkov, C. Russell, M. U. Gutmann, and C. Sutton. "Veegan: Reducing mode collapse in gans using implicit variational learning". In: *Proceedings of the 31st International Conference on Neural Information Processing Systems*. 2017, pp. 3310–3320.

[SRS10] P. Schnitzspan, S. Roth, and B. Schiele. "Automatic discovery of meaningful object parts with latent CRFs". In: *CVPR*. 2010.

[SS17a] A. Srivastava and C. Sutton. "Autoencoding Variational Inference For Topic Models". In: *ICLR*. 2017.

[SS17b] A. Srivastava and C. Sutton. "Autoencoding Variational Inference For Topic Models". In: *ICLR*. 2017.

[SS18a] O. Sener and S. Savarese. "Active Learning for Convolutional Neural Networks: A Core-Set Approach". In: *International Conference on Learning Representations*. 2018.

[SS18b] A. Subbaswamy and S. Saria. "Counterfactual Normalization: Proactively Addressing Dataset Shift and Improving Reliability Using Causal Mechanisms". In: *Proceedings of the 34th Conference on Uncertainty in Artificial Intelligence (UAI), 2018*. 2018.

[SS19] S. Sarkka and A. Solin. *Applied stochastic differential equations*. en. Cambridge University Press, 2019.

[SS20a] S. Sarkka and L. Svensson. "Levenberg-Marquardt and Line-Search Extended Kalman Smoothers". In: *ICASSP*. Barcelona, Spain: IEEE, May 2020.

[SS20b] K. E. Smith and A. O. Smith. "Conditional GAN for timeseries generation". In: *arXiv preprint arXiv:2006.16477* (2020).

[SS21] I. Sucholutsky and M. Schonlau. "Soft-Label Dataset Distillation and Text Dataset Distillation". In: *2021 International Joint Conference on Neural Networks (IJCNN)*. 2021, pp. 1–8.

[SS82] R. H. Shumway and D. S. Stoffer. "An approach to time series smoothing and forecasting using the em algorithm". en. In: *J. Time Ser. Anal.* 3.4 (July 1982), pp. 253–264.

[SSA14] K. Swersky, J. Snoek, and R. P. Adams. "Freeze-Thaw Bayesian Optimization". In: (2014). arXiv: 1406.3896 [stat.ML].

[SSA18] K. Shmelkov, C. Schmid, and K. Alahari. "How good is my GAN?" In: *Proceedings of the European Conference on Computer Vision (ECCV)*. 2018, pp. 213–229.

[SSB17] S. Semeniuta, A. Severyn, and E. Barth. "A Hybrid Convolutional Variational Autoencoder for Text Generation". In: (2017). arXiv: 1702 . 02390 [cs.CL].

[SSE18] Y. Song, J. Song, and S. Ermon. "Accelerating Natural Gradient with Higher-Order Invariance". In: *ICML*. 2018.

[SSF16] M. W. Seeger, D. Salinas, and V. Flunkert. "Bayesian Intermittent Demand Forecasting for Large Inventories". In: *NIPS*. 2016, pp. 4646–4654.

[SSG18a] S. Semeniuta, A. Severyn, and S. Gelly. "On Accurate Evaluation of GANs for Language Generation". In: *arXiv preprint arXiv:1806.04936* (2018).

[SSG18b] S. Semeniuta, A. Severyn, and S. Gelly. "On Accurate Evaluation of GANs for Language Generation". In: (2018). arXiv: 1806.04936 [cs.CL].

[SSG19] R. Singh, M. Sahani, and A. Gretton. "Kernel Instrumental Variable Regression". In: *Advances in Neural Information Processing Systems*. 2019, pp. 4593–4605.

[SSH13] S. Sarkka, A. Solin, and J. Hartikainen. "Spatio-Temporal Learning via Infinite-Dimensional Bayesian Filtering and Smoothing: A look at Gaussian process regression through Kalman filtering". In: *IEEE Signal Processing Magazine* (2013).

[SSJ12] R. Sipos, P. Shivaswamy, and T. Joachims. "Large-margin learning of submodular summarization models". In: *Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics*. 2012, pp. 224–233.

[SSK12] M. Sugiyama, T. Suzuki, and T. Kanamori. *Density Ratio Estimation in Machine Learning*. en. Cambridge University Press, 2012.

[SSM18] S. Santurkar, L. Schmidt, and A. Madry. "A classification-based study of covariate shift in gan distributions". In: *International Conference on Machine Learning*. PMLR. 2018, pp. 4480–4489.

[SSZ17] J. Snell, K. Swersky, and R. Zemel. "Prototypical networks for few-shot learning". In: *NIPS*. 2017, pp. 4077–4087.

[Sta] *Scientific Explanation*. https : / / plato . stanford . edu / entries / scientific - explanation / #ConcOpenIssuFutuDire. Accessed: 2021-11-23.

[Sta07] K. O. Stanley. "Compositional pattern producing networks: A novel abstraction of development". In: *Genet. Program. Evolvable Mach.* 8.2 (2007), pp. 131–162.

[Sta+17] N. Stallard, F. Miller, S. Day, S. W. Hee, J. Madan, S. Zohar, and M. Posch. "Determination of the optimal sample size for a clinical trial accounting for the population size". en. In: *Biom. J.* 59.4 (2017), pp. 609–625.

[Sta+19] K. O. Stanley, J. Clune, J. Lehman, and R. Miikkulainen. "Designing neural networks through

neuroevolution". In: *Nature Machine Intelligence* 1.1 (2019).

[Ste81] C. M. Stein. "Estimation of the mean of a multivariate normal distribution". In: *The annals of Statistics* (1981), pp. 1135–1151.

[Sto09] A. J. Storkey. "When Training and Test Sets are Different: Characterising Learning Transfer". In: *Dataset Shift in Machine Learning*. 2009.

[Sto17] J. Stoehr. "A review on statistical inference methods for discrete Markov random fields". In: (2017). arXiv: 1704.03331 [stat.ME].

[STR10] Y. Saatchi, R. Turner, and C. E. Rasmussen. "Gaussian Process Change Point Models". In: *ICML*. unknown, 2010, pp. 927–934.

[Str+17] H. Strobelt, S. Gehrmann, H. Pfister, and A. M. Rush. "Lstmvis: A tool for visual analysis of hidden state dynamics in recurrent neural networks". In: *IEEE transactions on visualization and computer graphics* 24.1 (2017), pp. 667–676.

[Str19] M. Streeter. "Bayes Optimal Early Stopping Policies for Black-Box Optimization". In: (2019). arXiv: 1902.08285 [cs.LG].

[Stu+22] D. Stutz, Krishnamurthy, Dvijotham, A. T. Cemgil, and A. Doucet. "Learning Optimal Conformal Classifiers". In: *ICLR*. 2022.

[STY17] M. Sundararajan, A. Taly, and Q. Yan. *Axiomatic Attribution for Deep Networks*. 2017. arXiv: 1703.01365 [cs.LG].

[Suc+20] F. P. Such, A. Rawal, J. Lehman, K. Stanley, and J. Clune. "Generative teaching networks: Accelerating neural architecture search by learning to generate synthetic training data". In: *International Conference on Machine Learning*. PMLR. 2020, pp. 9206–9216.

[Sud+03] E. Sudderth, A. Ihler, W. Freeman, and A. Willsky. "Nonparametric Belief Propagation". In: *CVPR*. 2003.

[Sud06] E. Sudderth. "Graphical Models for Visual Object Recognition and Tracking". PhD thesis. MIT, 2006.

[Sud+10] E. Sudderth, A. Ihler, M. Isard, W. Freeman, and A. Willsky. "Nonparametric Belief Propagation". In: *Comm. of the ACM* 53.10 (2010).

[Sug+13] M. Sugiyama, T. Kanamori, T. Suzuki, M. C. du Plessis, S. Liu, and I. Takeuchi. "Density-difference estimation". en. In: *Neural Comput.* 25.10 (2013), pp. 2734–2775.

[Sun+09] L. Sun, S. Ji, S. Yu, and J. Ye. "On the Equivalence Between Canonical Correlation Analysis and Orthonormalized Partial Least Squares". In: *IJCAI*. 2009.

[Sun+17] C. Sun, A. Shrivastava, S. Singh, and A. Gupta. "Revisiting unreasonable effectiveness of data in deep learning era". In: *Proceedings of the IEEE international conference on computer vision*. 2017, pp. 843–852.

[Sun+18] S. Sun, G. Zhang, C. Wang, W. Zeng, J. Li, and R. Grosse. "Differentiable Compositional Kernel Learning for Gaussian Processes". In: *ICML*. 2018.

[Sun+19a] S. Sun, G. Zhang, J. Shi, and R. Grosse. "Functional variational bayesian neural networks". In: *arXiv preprint arXiv:1903.05779* (2019).

[Sun+19b] S. Sun, Z. Cao, H. Zhu, and J. Zhao. "A Survey of Optimization Methods from a Machine Learning Perspective". In: (2019). arXiv: 1906.06821 [cs.LG].

[Sun+19c] M. Sundararajan, J. Xu, A. Taly, R. Sayres, and A. Najmi. "Exploring Principled Visualizations for Deep Network Attributions." In: *IUI Workshops*. Vol. 4. 2019.

[Sun+20] Y. Sun, X. Wang, Z. Liu, J. Miller, A. Efros, and M. Hardt. "Test-Time Training with Self-Supervision for Generalization under Distribution Shifts". In: *ICML*. Vol. 119. Proceedings of Machine Learning Research. PMLR, 2020, pp. 9229–9248.

[Sun+22] T. Sun, M. Segu, J. Postels, Y. Wang, L. Van Gool, B. Schiele, F. Tombari, and F. Yu. "SHIFT: A Synthetic Driving Dataset for Continuous Multi-Task Domain Adaptation". In: *CVPR*. June 2022.

[Sut+17] D. J. Sutherland, H.-Y. Tung, H. Strathmann, S. De, A. Ramdas, A. Smola, and A. Gretton. "Generative Models and Model Criticism via Optimized Maximum Mean Discrepancy". In: *ICLR*. 2017.

[Sut19] R. Sutton. *The Bitter Lesson*. 2019.

[Sut88] R. Sutton. "Learning to predict by the methods of temporal differences". In: *Machine Learning* 3.1 (1988), pp. 9–44.

[Sut90] R. S. Sutton. "Integrated Architectures for Learning, Planning, and Reacting Based on Approximating Dynamic Programming". In: *ICML*. Ed. by B. Porter and R. Mooney. Morgan Kaufmann, 1990, pp. 216–224.

[Sut96] R. S. Sutton. "Generalization in Reinforcement Learning: Successful Examples Using Sparse Coarse Coding". In: *NIPS*. Ed. by D. S. Touretzky, M. C. Mozer, and M. E. Hasselmo. MIT Press, 1996, pp. 1038–1044.

[Sut+99] R. Sutton, D. McAllester, S. Singh, and Y. Mansour. "Policy Gradient Methods for Reinforcement Learning with Function Approximation". In: *NIPS*. 1999.

[SV08] G. Shafer and V. Vovk. "A Tutorial on Conformal Prediction". In: *JMLR* 9.Mar (2008), pp. 371–421.

[SV98] M. Studenty and J. Vejnarova. "The multiinformation function as a tool for measuring stochastic dependence". In: *Learning in graphical models*. Ed. by M. Jordan. MIT Press, 1998, pp. 261–297.

[SVE04] A. Smola, S. V. N. Vishwanathan, and E. Eskin. "Laplace Propagation". In: *NIPS*. MIT Press, 2004, pp. 441–448.

[Svi04] M. Sviridenko. "A note on maximizing a submodular set function subject to a knapsack constraint". In: *Operations Research Letters* 32.1 (2004), pp. 41–43.

[SVK19] J. Su, D. V. Vargas, and S. Kouichi. "One pixel attack for fooling deep neural networks". In: *IEEE Trans. Evol. Comput.* 23.5 (2019).

[SVZ13] K. Simonyan, A. Vedaldi, and A. Zisserman. "Deep inside convolutional networks: Visualising image classification models and saliency maps". In: *arXiv preprint arXiv:1312.6034* (2013).

[SW06] J. E. Smith and R. L. Winkler. "The Optimizer's Curse: Skepticism and Postdecision Surprise in Decision Analysis". In: *Manage. Sci.* 52.3 (2006), pp. 311–322.

[SW13] G. J. Sussman and J. Wisdom. *Functional Differential Geometry*. Functional Differential Geometry. MIT Press, 2013.

[SW20] V. G. Satorras and M. Welling. "Neural Enhanced Belief Propagation on Factor Graphs". In: (2020). arXiv: 2003.01998 [cs.LG].

[SW87a] M. Shewry and H. Wynn. "Maximum entropy sampling". In: *J. Applied Statistics* 14 (1987), 165–170.

[SW87b] R. Swendsen and J.-S. Wang. "Nonuniversal critical dynamics in Monte Carlo simulations". In: *Physical Review Letters* 58 (1987), pp. 86–88.

[SW89] S. Singhal and L. Wu. "Training Multilayer Perceptrons with the Extended Kalman Algorithm". In: *NIPS*. Vol. 1. 1989.

[Swe+10] K. Swersky, B. Chen, B. Marlin, and N. de Freitas. "A Tutorial on Stochastic Approximation Algorithms for Training Restricted Boltzmann Machines and Deep Belief Nets". In: *Information Theory and Applications (ITA) Workshop*. 2010.

[Swe+13] K. Swersky, D. Duvenaud, J. Snoek, F. Hutter, and M. A. Osborne. "Raiders of the Lost Architecture: Kernels for Bayesian Optimization in Conditional Parameter Spaces". In: *NIPS BayesOpt workshop*. 2013.

[SWL03] M. Seeger, C. K. I. Williams, and N. D. Lawrence. "Fast Forward Selection to Speed Up Sparse Gaussian Process Regression". In: *AISTATS*. 2003.

[SWW08] E. Sudderth, M. Wainwright, and A. Willsky. "Loop series and Bethe variational bounds in attractive graphical models". In: *NIPS* (2008).

[SYD19] R. Sen, H.-F. Yu, and I. Dhillon. "Think Globally, Act Locally: A Deep Neural Network Approach to High-Dimensional Time Series Forecasting". In: *NIPS*. 2019.

[SZ22] R. Shwartz-Ziv. "Information Flow in Deep Neural Networks". PhD thesis. 2022.

[SZ+22] R. Shwartz-Ziv, M. Goldblum, H. Souri, S. Kapoor, C. Zhu, Y. LeCun, and A. G. Wilson. "Pre-Train Your Loss: Easy Bayesian Transfer Learning with Informative Priors". In: (May 2022). arXiv: 2205.10279 [cs.LG].

[Sze10] C. Szepesvari. *Algorithms for Reinforcement Learning*. Morgan Claypool, 2010.

[Sze+14] C. Szegedy, W. Zaremba, I. Sutskever, J. Bruna, D. Erhan, I. Goodfellow, and R. Fergus. "Intriguing properties of neural networks". In: *ICLR*. 2014.

[Sze+15] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich. "Going Deeper with Convolutions". In: *CVPR*. 2015.

[TAH20] D. Teney, E. Abbasnejad, and A. van den Hengel. "Learning What Makes a Difference from Counterfactual Examples and Gradient Supervision". In: *CoRR* abs/2004.09034 (2020). arXiv: 2004.09034.

[Tan+16] X. Tan, S. A. Naqvi, K. A. Heller, and V. A. Rao. "Content-based Modeling of Reciprocal Relationships using Hawkes and Gaussian Processes." In: *Proceedings of the Thirty-Second Conference on Uncertainty in Artificial Intelligence*. 2016.

[TB16] P. S. Thomas and E. Brunskill. "Data-Efficient Off-Policy Policy Evaluation for Reinforcement Learning". In: *ICML*. 2016, pp. 2139–2148.

[TB22] A. Tiulpin and M. B. Blaschko. "Greedy Bayesian Posterior Approximation with Deep Ensembles". In: *Trans. on Machine Learning Research* (2022).

[TB99] M. Tipping and C. Bishop. "Probabilistic principal component analysis". In: *J. of Royal Stat. Soc. Series B* 21.3 (1999), pp. 611–622.

[TBA19] N. Tremblay, S. Barthelmé, and P.-O. Amblard. "Determinantal Point Processes for Coresets." In: *J. Mach. Learn. Res.* 20 (2019), pp. 168–1.

[TBB19] J. Townsend, T. Bird, and D. Barber. "Practical Lossless Compression with Latent Variables using Bits Back Coding". In: *ICLR*. 2019.

[TBF06] S. Thrun, W. Burgard, and D. Fox. *Probabilistic Robotics*. MIT Press, 2006.

[TBS13] R. Turner, S. Bottone, and C. Stanek. "Online variational approximations to non-exponential family change point models: With application to radar tracking". In: *NIPS*. 2013.

[TCG21] Y. Tian, X. Chen, and S. Ganguli. "Understanding self-supervised learning dynamics without contrastive pairs". In: *International Conference on Machine Learning*. PMLR. 2021, pp. 10268–10278.

[Teh06] Y. W. Teh. "A hierarchical Bayesian language model based on Pitman-Yor processes". In: *Proc. of the Assoc. for Computational Linguistics*. 2006, 985=992.

[Teh+06a] Y.-W. Teh, M. Jordan, M. Beal, and D. Blei. "Hierarchical Dirichlet processes". In: *JASA* 101.476 (2006), pp. 1566–1581.

[Teh+06b] Y. W. Teh, M. I. Jordan, M. J. Beal, and D. M. Blei. "Hierarchical Dirichlet Processes". In: *JASA* 101.476 (2006), pp. 1566–1581.

[Teh+20] N. Tehrani, N. S. Arora, Y. L. Li, K. D. Shah, D. Noursi, M. Tingley, N. Torabi, S. Masouleh, E. Lippert, and E. Meijer. "Bean Machine: A Declarative Probabilistic Programming Language For Efficient Programmable Inference". In: *Proceedings of the 10th International Conference on Probabilistic Graphical Models*. Ed. by M. Jaeger and T. D. Nielsen. Vol. 138. Proceedings of Machine Learning Research. PMLR, 2020, pp. 485–496.

[Ten+20] I. Tenney, J. Wexler, J. Bastings, T. Bolukbasi, A. Coenen, S. Gehrmann, E. Jiang, M. Pushkarna, C. Radebaugh, E. Reif, et al. "The language interpretability tool: Extensible, interactive visualizations and analysis for NLP models". In: *arXiv preprint arXiv:2008.05122* (2020).

[TF03] M. Tipping and A. Faul. "Fast marginal likelihood maximisation for sparse Bayesian models". In: *AI/Stats*. 2003.

[TG09] Y. W. Teh and D. Gorur. "Indian buffet processes with power-law behavior". In: *NIPS*. 2009, pp. 1838–1846.

[TG18] L. Tu and K. Gimpel. "Learning Approximate Inference Networks for Structured Prediction". In: *ICLR*. 2018.

[TGFS18] F. Tronarp, Á. F. García-Fernández, and S. Särkkä. "Iterative Filtering and Smoothing in Nonlinear and Non-Gaussian Systems Using Conditional Moments". In: *IEEE Signal Process. Lett.* 25.3 (2018), pp. 408–412.

[TGK03] B. Taskar, C. Guestrin, and D. Koller. "Max-Margin Markov Networks". In: *NIPS*. 2003.

[TH09] T. Tieleman and G. Hinton. "Using Fast Weights to Improve Persistent Contrastive Divergence". In: *ICML*. 2009, pp. 1033–1040.

[Tho+19] V. Thomas, F. Pedregosa, B. van Merriënboer, P.-A. Mangazol, Y. Bengio, and N. Le Roux.

"Information matrices and generalization". In: (2019). arXiv: 1906.07774 [cs.LG].

[Tho33] W. R. Thompson. "On the Likelihood that One Unknown Probability Exceeds Another in View of the Evidence of Two Samples". In: *Biometrika* 25.3/4 (1933), pp. 285–294.

[Thr+04] S. Thrun, M. Montemerlo, D. Koller, B. Wegbreit, J. Nieto, and E. Nebot. "FastSLAM: An efficient solution to the simultaneous localization and mapping problem with unknown data association". In: *JMLR* 2004 (2004).

[Thr98] S. Thrun. "Lifelong learning algorithms". In: *Learning to learn*. Ed. by S. Thrun and L. Pratt. Kluwer, 1998, pp. 181–209.

[Thu+21] S. Thulasidasan, S. Thapa, S. Dhaubhadel, G. Chennupati, T. Bhattacharya, and J. Bilmes. "An Effective Baseline for Robustness to Distributional Shift". In: (2021). arXiv: 2105.07107 [cs.LG].

[Tia+20] Y. Tian, C. Sun, B. Poole, D. Krishnan, C. Schmid, and P. Isola. "What makes for good views for contrastive learning". In: *ArXiv* abs/2005.10243 (2020).

[Tib96] R. Tibshirani. "Regression shrinkage and selection via the lasso". In: *J. Royal. Statist. Soc B* 58.1 (1996), pp. 267–288.

[Tie08] T. Tieleman. "Training restricted Boltzmann machines using approximations to the likelihood gradient". In: *ICML*. ACM New York, NY, USA. 2008, pp. 1064–1071.

[Tip01] M. Tipping. "Sparse Bayesian learning and the relevance vector machine". In: *JMLR* 1 (2001), pp. 211–244.

[Tip98] M. Tipping. "Probabilistic visualization of high-dimensional binary data". In: *NIPS*. 1998.

[Tit09] M. K. Titsias. "Variational Learning of Inducing Variables in Sparse Gaussian Processes". In: *AISTATS*. 2009.

[TK86] L. Tierney and J. Kadane. "Accurate approximations for posterior moments and marginal densities". In: *JASA* 81.393 (1986).

[TKI20] Y. Tian, D. Krishnan, and P. Isola. "Contrastive Multiview Coding". In: *ECCV*. 2020.

[TKW08] Y. W. Teh, K. Kurihara, and M. Welling. "Collapsed variational inference for HDP". In: *Advances in neural information processing systems*. 2008, pp. 1481–1488.

[TL05] E. Todorov and W. Li. "A Generalized Iterative LQG Method for Locally-optimal Feedback Control of Constrained Nonlinear Stochastic Systems". In: *ACC*. 2005, pp. 300–306.

[TL18a] S. J. Taylor and B. Letham. "Forecasting at scale". en. In: *The American Statistician* 72.1 (2018), pp. 37–45.

[TL18b] G. Tucker and D. Lawson. "Doubly Reparameterized Gradient Estimators for Monte Carlo Objectives". In: *1st Symposium on Advances in Approximate Bayesian Inference*. 2018.

[TL19] M. Tan and Q. Le. "Efficientnet: Rethinking model scaling for convolutional neural networks". In: *International Conference on Machine Learning*. PMLR. 2019, pp. 6105–6114.

[TLG14] M. Titsias and M. Lázaro-Gredilla. "Doubly Stochastic Variational Bayes for non-Conjugate Inference". In: *ICML*. 2014, pp. 1971–1979.

[TMD12] A. Talhouk, K. Murphy, and A. Doucet. "Efficient Bayesian Inference for Multivariate Probit Models with Sparse Inverse Correlation Matrices". In: *J. Comp. Graph. Statist.* 21.3 (2012), pp. 739–757.

[TMK04] G. Theocharous, K. Murphy, and L. Kaelbling. "Representing hierarchical POMDPs as DBNs for multi-scale robot localization". In: *ICRA*. 2004.

[TN13] L. S. L. Tan and D. J. Nott. "Variational Inference for Generalized Linear Mixed Models Using Partially Noncentered Parametrizations". In: *Stat. Sci.* (2013).

[TN18] L. S. L. Tan and D. J. Nott. "Gaussian variational approximation with sparse precision matrices". In: *Stat. Comput.* 28.2 (2018), pp. 259–275.

[TND21] M.-N. Tran, T.-N. Nguyen, and V.-H. Dao. "A practical tutorial on Variational Bayes". In: (2021). arXiv: 2103.01327 [stat.CO].

[TOB16] L. Theis, A. van den Oord, and M. Bethge. "A note on the evaluation of generative models". In: *ICLR*. 2016.

[Tol22] S. Toledo. "UltimateKalman: Flexible Kalman Filtering and Smoothing Using Orthogonal Transformations". In: (July 2022). arXiv: 2207 . 13526 [math.NA].

[Tom+20] R. Tomsett, D. Harborne, S. Chakraborty, P. Gurram, and A. Preece. "Sanity checks for saliency metrics". In: *Proceedings of the AAAI conference on artificial intelligence*. Vol. 34. 04. 2020, pp. 6021–6029.

[Tom22] J. M. Tomczak. *Deep Generative Modeling*. en. 1st ed. Springer, 2022.

[Tou09] M. Toussaint. "Robot Rrajectory Optimization using Approximate Inference". In: *ICML*. 2009, pp. 1049–1056.

[Tou14] M. Toussaint. *Bandits, Global Optimization, Active Learning, and Bayesian RL – understanding the common ground*. Autonomous Learning Summer School. 2014.

[Tou+19] J Toubeau, J Bottieau, F Vallée, and Z De Grève. "Deep Learning-Based Multivariate Probabilistic Forecasting for Short-Term Scheduling in Power Markets". In: *IEEE Trans. Power Syst.* 34.2 (2019), pp. 1203–1215.

[TOV18] C. Truong, L. Oudre, and N. Vayatis. "Selective review of offline change point detection methods". In: (2018). arXiv: 1801.00718 [cs.CE].

[TP97] S. Thrun and L. Pratt, eds. *Learning to learn*. Kluwer, 1997.

[TPB00] N. Tishby, F. C. Pereira, and W. Bialek. "The information bottleneck method". In: *ArXiv* physics/0004057 (2000).

[TPB99] N. Tishby, F. Pereira, and W. Biale. "The Information Bottleneck method". In: *The 37th annual Allerton Conf. on Communication, Control, and Computing*. 1999, pp. 368–377.

[TR19] M. K. Titsias and F. Ruiz. "Unbiased Implicit Variational Inference". In: *AISTATS*. Ed. by K. Chaudhuri and M. Sugiyama. Vol. 89. Proceedings of Machine Learning Research. PMLR, 2019, pp. 167–176.

[TR97] J. Tsitsiklis and B. V. Roy. "An analysis of temporal-difference learning with function approximation". In: *IEEE Trans. on Automatic Control* 42.5 (1997), pp. 674–690.

Author: Kevin P. Murphy. (C) MIT Press. CC-BY-NC-ND license

[Tra+19] D. Tran, K. Vafa, K. K. Agrawal, L. Dinh, and B. Poole. "Discrete Flows: Invertible Generative Models of Discrete Data". In: *Advances in Neural Information Processing Systems*. 2019.

[Tra+20a] L. Tran, B. S. Veeling, K. Roth, J. Swiatkowski, J. V. Dillon, J. Snoek, S. Mandt, T. Salimans, S. Nowozin, and R. Jenatton. "Hydra: Preserving Ensemble Diversity for Model Distillation". In: (2020). arXiv: 2001.04694 [cs.LG].

[Tra+20b] M.-N. Tran, N Nguyen, D Nott, and R Kohn. "Bayesian Deep Net GLM and GLMM". In: *J. Comput. Graph. Stat.* 29.1 (2020), pp. 97–113.

[TRB16] D. Tran, R. Ranganath, and D. M. Blei. "The Variational Gaussian Process". In: *ICLR*. 2016.

[Tri21] K. Triantafyllopoulos. *Bayesian Inference of State Space Models: Kalman Filtering and Beyond*. en. 1st ed. Springer, 2021.

[TS06] M. Toussaint and A. Storkey. "Probabilistic inference for solving discrete and continuous state Markov Decision Processes". In: *ICML*. 2006, pp. 945–952.

[Tsa+18] Y.-H. Tsai, W.-C. Hung, S. Schulter, K. Sohn, M.-H. Yang, and M. Chandraker. "Learning to adapt structured output space for semantic segmentation". In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2018, pp. 7472–7481.

[Tsa+19] Y.-H. H. Tsai, S. Bai, M. Yamada, L.-P. Morency, and R. Salakhutdinov. "Transformer Dissection: An Unified Understanding for Transformer's Attention via the Lens of Kernel". In: *EMNLP*. Association for Computational Linguistics, 2019, pp. 4344–4353.

[Tsa+21] Y.-H. H. Tsai, S. Bai, L.-P. Morency, and R. Salakhutdinov. "A Note on Connecting Barlow Twins with Negative-Sample-Free Contrastive Learning". In: *ArXiv* abs/2104.13712 (2021).

[Tsa88] C. Tsallis. "Possible generalization of Boltzmann-Gibbs statistics". In: *J. of Statistical Physics* 52 (1988), pp. 479–487.

[Tsc+14] S. Tschiatschek, R. Iyer, H. Wei, and J. Bilmes. "Learning Mixtures of Submodular Functions for Image Collection Summarization". In: *NIPS*. Montreal, Canada, 2014.

[Tsc+19] M. Tschannen, J. Djolonga, P. K. Rubenstein, S. Gelly, and M. Lucic. "On Mutual Information Maximization for Representation Learning". In: *arXiv preprint arXiv:1907.13625* (2019).

[Tsi+17] P. A. Tsividis, T. Pouncy, J. L. Xu, J. B. Tenenbaum, and S. J. Gershman. "Human Learning in Atari". en. In: *AAAI Spring Symposium Series*. 2017.

[Tso+05] I. Tsochantaridis, T. Joachims, T. Hofmann, and Y. Altun. "Large Margin Methods for Structured and Interdependent Output Variables". In: *JMLR* 6 (2005), pp. 1453–1484.

[TT13] E. G. Tabak and C. V. Turner. "A family of nonparametric density estimation algorithms". In: *Communications on Pure and Applied Mathematics* 66.2 (2013), pp. 145–164.

[TT17] B. Trippe and R. Turner. "Overpruning in Variational Bayesian Neural Networks". In: *NIPS Workshop on Advances in Approximate Bayesian Inference*. 2017.

[Tuc+19] G. Tucker, D. Lawson, B. Dai, and R. Ranganath. "Revisiting auxiliary latent variables in generative models". In: (2019).

[TUI17] T. Taketomi, H. Uchiyama, and S. Ikeda. "Visual SLAM algorithms: a survey from 2010 to 2016". en. In: *IPSJ Transactions on Computer Vision and Applications* 9.1 (2017), p. 16.

[Tul+18] S. Tulyakov, M.-Y. Liu, X. Yang, and J. Kautz. "Mocogan: Decomposing motion and content for video generation". In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2018, pp. 1526–1535.

[Tur+08] R. Turner, P. Berkes, M. Sahani, and D. Mackay. *Counterexamples to variational free energy compactness folk theorems*. Tech. rep. U. Cambridge, 2008.

[TVE10] E. G. Tabak and E. Vanden-Eijnden. "Density estimation by dual ascent of the log-likelihood". In: *Communications in Mathematical Sciences* 8.1 (2010), pp. 217–233.

[TW16] J. M. Tomczak and M. Welling. "Improving variational auto-encoders using Householder flow". In: *NeurIPS Workshop on Bayesian Deep Learning* (2016).

[TX00] J. B. Tenenbaum and F. Xu. "Word learning as Bayesian inference". In: *Proc. 22nd Annual Conf. of the Cognitive Science Society*. 2000.

[TZ02] Z. Tu and S. Zhu. "Image Segmentation by Data-Driven Markov Chain Monte Carlo". In: *IEEE PAMI* 24.5 (2002), pp. 657–673.

[Tze+17] E. Tzeng, J. Hoffman, K. Saenko, and T. Darrell. "Adversarial discriminative domain adaptation". In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2017, pp. 7167–7176.

[UCS17] S. Ubaru, J. Chen, and Y. Saad. "Fast Estimation of $tr(f(A))$ via Stochastic Lanczos Quadrature". In: *SIAM J. Matrix Anal. Appl.* 38.4 (2017), pp. 1075–1099.

[Ude+16] M. Udell, C. Horn, R. Zadeh, and S. Boyd. "Generalized Low Rank Models". In: *Foundations and Trends in Machine Learning* 9.1 (2016), pp. 1–118.

[UHJ20] M. Uehara, J. Huang, and N. Jiang. "Minimax Weight and Q-Function Learning for Off-Policy Evaluation". In: *ICML*. 2020.

[UML13] B. Uria, I. Murray, and H. Larochelle. "RNADE: The real-valued neural autoregressive density-estimator". In: *NIPS*. 2013.

[UML14] B. Uria, I. Murray, and H. Larochelle. "A Deep and Tractable Density Estimator". In: *ICML*. 2014.

[UN98] N. Ueda and R. Nakano. "Deterministic annealing EM algorithm". In: *Neural Networks* 11 (1998), pp. 271–282.

[UR16] B. Ustun and C. Rudin. "Supersparse linear integer models for optimized medical scoring systems". In: *Machine Learning* 102.3 (2016), pp. 349–391.

[Uri+16] B. Uria, M.-A. Côté, K. Gregor, I. Murray, and H. Larochelle. "Neural Autoregressive Distribution Estimation". In: *JMLR* (2016).

[UTR14] B. Ustun, S. Tracà, and C. Rudin. *Supersparse Linear Integer Models for Interpretable Classification*. 2014. arXiv: 1306.6677 [stat.ML].

[Uur+17] V. Uurtio, J. M. Monteiro, J. Kandola, J. Shawe-Taylor, D. Fernandez-Reyes, and J. Rousu. "A Tutorial on Canonical Correlation Methods". In: *ACM Computing Surveys* (2017).

[UVL16] D. Ulyanov, A. Vedaldi, and V. Lempitsky. "Instance Normalization: The Missing Ingredient for Fast Stylization". In: (2016). arXiv: 1607.08022 [cs.CV].

[UVL18] D. Ulyanov, A. Vedaldi, and V. Lempitsky. "Deep Image Prior". In: *CVPR*. 2018.

[Vaa00] A. W. Van der Vaart. *Asymptotic statistics*. Vol. 3. Cambridge university press, 2000.

[Val00] H. Valpola. "Bayesian Ensemble Learning for Nonlinear Factor Analysis". PhD thesis. Helsinki University of Technology, 2000.

[Van10] J. Vanhatalo. "Speeding up the inference in Gaussian process models". PhD thesis. Helsinki Univ. Technology, 2010.

[van+18] H. van Hasselt, Y. Doron, F. Strub, M. Hessel, N. Sonnerat, and J. Modayil. *Deep Reinforcement Learning and the Deadly Triad*. arXiv:1812.02648. 2018.

[Vas+17a] A. B. Vasudevan, M. Gygli, A. Volokitin, and L. Van Gool. "Query-adaptive video summarization via quality-aware relevance estimation". In: *Proceedings of the 25th ACM international conference on Multimedia*. 2017, pp. 582–590.

[Vas+17b] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin. "Attention is all you need". In: *NIPS*. 2017, pp. 5998–6008.

[Vas+17c] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin. "Attention Is All You Need". In: *NIPS*. 2017.

[Vaz+22] S. Vaze, K. Han, A. Vedaldi, and A. Zisserman. "Open-Set Recognition: A Good Closed-Set Classifier is All You Need". In: *ICLR*. 2022.

[VBW15] S. S. Villar, J. Bowden, and J. Wason. "Multi-armed Bandit Models for the Optimal Design of Clinical Trials: Benefits and Challenges". en. In: *Stat. Sci.* 30.2 (2015), pp. 199–215.

[VDMW03] R. Van Der Merwe and E. Wan. "Sigma-Point Kalman Filters for probabilistic inference in dynamic state-space models". In: *Workshop on Advances in ML*. 2003.

[Ved+18] R. Vedantam, I. Fischer, J. Huang, and K. Murphy. "Generative Models of Visually Grounded Imagination". In: *ICLR*. 2018.

[Veh+19] A. Vehtari, A. Gelman, D. Simpson, B. Carpenter, and P.-C. Bürkner. "Rank-normalization, folding, and localization: An improved $\widehat{R}$ for assessing convergence of MCMC". In: (2019). arXiv: 1903.08008 [stat.CO].

[Vei+21] V. Veitch, A. D'Amour, S. Yadlowsky, and J. Eisenstein. "Counterfactual Invariance to Spurious Correlations: Why and How to Pass Stress Tests". In: *Advances in Neural Information Processing Systems*. 2021.

[Vel+17] P. Veličković, G. Cucurull, A. Casanova, A. Romero, P. Lio, and Y. Bengio. "Graph attention networks". In: *arXiv preprint arXiv:1710.10903* (2017).

[Ver18] R. Vershynin. *High-Dimensional Probability: An Introduction with Applications in Data Science*. en. 1 edition. Cambridge University Press, 2018.

[Ver+19] A. Vergari, A. Molina, R. Peharz, Z. Ghahramani, K. Kersting, and I. Valera. "Automatic Bayesian Density Analysis". In: *AAAI*. 2019.

[VF+80] B. C. Van Fraassen et al. *The scientific image*. Oxford University Press, 1980.

[VGG17] A. Vehtari, A. Gelman, and J. Gabry. "Practical Bayesian model evaluation using leave-one-out cross-validation and WAIC". In: *Stat. Comput.* 27.5 (2017), pp. 1413–1432.

[VGS05] V. Vovk, A. Gammerman, and G. Shafer. *Algorithmic Learning in a Random World*. en. 2005th ed. Springer, 2005.

[Vid99] P. Vidoni. "Exponential Family State Space Models Based on a Conjugate Latent Process". In: *J. R. Stat. Soc. Series B Stat. Methodol.* 61.1 (1999), pp. 213–221.

[Vil08] C. Villani. *Optimal Transport: Old and New*. Grundlehren der mathematischen Wissenschaften. Springer Berlin Heidelberg, 2008.

[Vil+19] R. Villegas, A. Pathak, H. Kannan, D. Erhan, Q. V. Le, and H. Lee. "High Fidelity Video Prediction with Large Stochastic Recurrent Neural Networks". In: *NIPS*. 2019.

[Vin+08] P. Vincent, H. Larochelle, Y. Bengio, and P.-A. Manzagol. "Extracting and composing robust features with denoising autoencoders". In: *Proceedings of the 25th international conference on Machine learning*. 2008, pp. 1096–1103.

[Vin+10a] P. Vincent, H. Larochelle, I. Lajoie, Y. Bengio, and P.-A. Manzagol. "Stacked denoising autoencoders: Learning useful representations in a deep network with a local denoising criterion". In: *Journal of machine learning research* 11.Dec (2010), pp. 3371–3408.

[Vin+10b] M. Vinyals, J. Cerquides, J. Rodriguez-Aguilar, and A. Farinelli. "Worst-case bounds on the quality of max-product fixed-points". In: *NIPS*. 2010.

[Vin11] P. Vincent. "A connection between score matching and denoising autoencoders". In: *Neural computation* 23.7 (2011), pp. 1661–1674.

[Vir10] S. Virtanen. "Bayesian exponential family projections". MA thesis. Aalto University, 2010.

[Vis+06] S. V. N. Vishwanathan, N. N. Schraudolph, M. W. Schmidt, and K. P. Murphy. "Accelerated training of conditional random fields with stochastic gradient methods". In: *ICML*. ACM Press, 2006.

[Vis+10] S. V. N. Vishwanathan, N. N. Schraudolph, R. Kondor, and K. M. Borgward. "Graph Kernels". In: *JMLR* 11 (2010), pp. 1201–1242.

[Vit67] A. Viterbi. "Error bounds for convolutional codes and an asymptotically optimal decoding algorithm". In: *IEEE Trans. on Information Theory* 13.2 (1967), pp. 260–269.

[VJMP22] V. Voleti, A. Jolicoeur-Martineau, and C. Pal. "MCVD: Masked Conditional Video Diffusion for Prediction, Generation, and Interpolation". In: (May 2022). arXiv: 2205.09853 [cs.CV].

[VJV09] J. Vanhatalo, P. Jylänki, and A. Vehtari. "Gaussian process regression with Student-t likelihood". In: *NIPS*. Vol. 22. 2009.

[VK20a] A. Vahdat and J. Kautz. "NVAE: A Deep Hierarchical Variational Autoencoder". In: *NIPS*. 2020.

[VK20b] A. Vahdat and J. Kautz. "Nvae: A deep hierarchical variational autoencoder". In: *arXiv preprint arXiv:2007.03898* (2020).

[VLT21] G. M. van de Ven, Z. Li, and A. S. Tolias. "Class-Incremental Learning with Generative Classi-

fiers". In: *CVPR workshop on Continual Learning in Computer Vision (CLVision)*. 2021.

[Vo+15] B.-N. Vo, M. Mallick, Y. Bar-Shalom, S. Coraluppi, R. Osborne, R. Mahler, B. t Vo, and J. Webster. *Multitarget tracking*. John Wiley and Sons, 2015.

[Von13] P. Vontobel. "The Bethe permanent of a non-negative matrix". In: *IEEE Trans. Info. Theory* 59.3 (2013).

[Vov13] V. Vovk. "Kernel Ridge Regression". In: *Empirical Inference: Festschrift in Honor of Vladimir N. Vapnik*. Ed. by B. Schölkopf, Z. Luo, and V. Vovk. Springer Berlin Heidelberg, 2013, pp. 105–116.

[VPV10] J. Vanhatalo, V. Pietilainen, and A. Vehtari. "Approximate inference for disease mapping with sparse Gaussian processes". In: *Statistics in Medicine* 29.15 (2010), pp. 1580–1607.

[vR11] M. van der Laan and S. Rose. *Targeted Learning: Causal Inference for Observational and Experimental Data*. Jan. 2011.

[VR18] S. Verma and J. Rubin. "Fairness Definitions Explained". In: *2018 IEEE/ACM International Workshop on Software Fairness (FairWare)*. May 2018, pp. 1–7.

[VRF11] C. Varin, N. Reid, and D. Firth. "An overview of composite likelihood methods". In: *Stat. Sin.* 21.1 (2011), pp. 5–42.

[VT11] S. I. VanderWeele TJ. "A new criterion for confounder selection". In: *Biometrics* (2011).

[VT18] G. M. van de Ven and A. S. Tolias. "Three scenarios for continual learning". In: *NeurIPS Continual Learning workshop*. 2018.

[Vyt+19] D. Vytiniotis, D. Belov, R. Wei, G. Plotkin, and M. Abadi. "The differentiable curry". In: *NeurIPS 2019 Workshop Program Transformations*. 2019.

[VZ20] V. Veitch and A. Zaveri. "Sense and Sensitivity Analysis: Simple Post-Hoc Analysis of Bias Due to Unobserved Confounding". In: *Advances in Neural Information Processing Systems*. Ed. by H. Larochelle, M. Ranzato, R. Hadsell, M. F. Balcan, and H. Lin. Vol. 33. Curran Associates, Inc., 2020, pp. 10999–11009.

[WA13] A. Wilson and R. Adams. "Gaussian process kernels for pattern discovery and extrapolation". In: *International conference on machine learning*. 2013, pp. 1067–1075.

[Wal+09] H. Wallach, I. Murray, R. Salakhutdinov, and D. Mimno. "Evaluation Methods for Topic Models". In: *ICML*. 2009.

[Wal+20] M. Walmsley et al. "Galaxy Zoo: probabilistic morphology through Bayesian CNNs and active learning". In: *Monthly Notices Royal Astronomial Society* 491.2 (2020), pp. 1554–1574.

[Wan+16] Z. Wang, T. Schaul, M. Hessel, H. van Hasselt, M. Lanctot, and N. de Freitas. "Dueling Network Architectures for Deep Reinforcement Learning". In: *ICML*. 2016.

[Wan17] M. P. Wand. "Fast Approximate Inference for Arbitrarily Large Semiparametric Regression Models via Message Passing". In: *JASA* 112.517 (2017), pp. 137–168.

[Wan+17a] T. Wang, C. Rudin, F. Doshi-Velez, Y. Liu, E. Klampfl, and P. MacNeille. "A bayesian framework for learning rule sets for interpretable classification". In: *The Journal of Machine Learning Research* 18.1 (2017), pp. 2357–2393.

[Wan+17b] X. Wang, T. Li, S. Sun, and J. M. Corchado. "A Survey of Recent Advances in Particle Filters and Remaining Challenges for Multitarget Tracking". en. In: *Sensors* 17.12 (2017).

[Wan+17c] Y. Wang et al. "Tacotron: Towards End-to-End Speech Synthesis". In: *Interspeech*. 2017.

[Wan+18] T.-C. Wang, M.-Y. Liu, J.-Y. Zhu, G. Liu, A. Tao, J. Kautz, and B. Catanzaro. "Video-to-video synthesis". In: *Proceedings of the 32nd International Conference on Neural Information Processing Systems*. 2018, pp. 1152–1164.

[Wan+19a] K. Wang, G. Pleiss, J. Gardner, S. Tyree, K. Q. Weinberger, and A. G. Wilson. "Exact Gaussian Processes on a Million Data Points". In: *NIPS*. 2019, pp. 14622–14632.

[Wan+19b] S. Wang, W. Bai, C. Lavania, and J. Bilmes. "Fixing Mini-batch Sequences with Hierarchical Robust Partitioning". In: *Proceedings of the Twenty-Second International Conference on Artificial Intelligence and Statistics*. Ed. by K. Chaudhuri and M. Sugiyama. Vol. 89. Proceedings of Machine Learning Research. PMLR, 2019, pp. 3352–3361.

[Wan+19c] Y. Wang, A. Smola, D. C. Maddix, J. Gasthaus, D. Foster, and T. Januschowski. "Deep Factors for Forecasting". In: *ICML*. 2019.

[Wan+20a] D. Wang, E. Shelhamer, S. Liu, B. Olshausen, and T. Darrell. "Tent: Fully Test-Time Adaptation by Entropy Minimization". In: *ICLR*. 2020.

[Wan+20b] H. Wang, X. Wu, Z. Huang, and E. P. Xing. "High-frequency component helps explain the generalization of convolutional neural networks". In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2020, pp. 8684–8694.

[Wan+20c] T. Wang, J.-Y. Zhu, A. Torralba, and A. A. Efros. *Dataset Distillation*. 2020. arXiv: 1811.10959 [cs.LG].

[Wan+20d] Z. Wang, S. Cheng, L. Yueru, J. Zhu, and B. Zhang. "A Wasserstein Minimum Velocity Approach to Learning Unnormalized Models". In: *Proceedings of the Twenty Third International Conference on Artificial Intelligence and Statistics*. Ed. by S. Chiappa and R. Calandra. Vol. 108. Proceedings of Machine Learning Research. PMLR, 2020, pp. 3728–3738.

[Wan+21] J. Wang, C. Lan, C. Liu, Y. Ouyang, W. Zeng, and T. Qin. "Generalizing to Unseen Domains: A Survey on Domain Generalization". In: *IJCAI*. 2021.

[Wan+22] H. Wang, Y. Yang, D. Pati, and A. Bhattacharya. "Structured Variational Inference in Bayesian State-Space Models". In: (2022).

[Was06] L. Wasserman. *All of Nonparametric Statistics*. Springer, 2006.

[Wat10] S. Watanabe. "Asymptotic Equivalence of Bayes Cross Validation and Widely Applicable Information Criterion in Singular Learning Theory". In: *JMLR* 11 (2010), pp. 3571–3594.

[Wat13] S. Watanabe. "A Widely Applicable Bayesian Information Criterion". In: *JMLR* 14 (2013), pp. 867–897.

[Wat60] S. Watanabe. "Information theoretical analysis of multivariate correlation". In: *IBM J. of Research and Development* 4 (1960), pp. 66–82.

[WB05] J. Winn and C. Bishop. "Variational Message Passing". In: *JMLR* 6 (2005), pp. 661–694.

[WB12] C. Wang and D. M. Blei. "Truncation-free online variational inference for Bayesian nonparametric models". In: *Advances in neural information processing systems*. 2012, pp. 413–421.

[WB20] T. Wang and J. Ba. "Exploring Model-based Planning with Policy Networks". In: *ICLR*. 2020.

[WBC21] Y. Wang, D. M. Blei, and J. P. Cunningham. "Posterior Collapse and Latent Variable Non-identifiability". In: *NIPS*. 2021.

[WCS08] M. Welling, C. Chemudugunta, and N. Sutter. "Deterministic Latent Variable Models and their Pitfalls". In: *ICDM*. 2008.

[WD92] C. Watkins and P. Dayan. "Q-learning". In: *Machine Learning* 8.3 (1992), pp. 279–292.

[WDN15] A. G. Wilson, C. Dann, and H. Nickisch. "Thoughts on Massively Scalable Gaussian Processes". In: *arXiv preprint arXiv:1511.01870* (2015). https://arxiv.org/abs/1511.01870.

[Web+17] T. Weber et al. "Imagination-Augmented Agents for Deep Reinforcement Learning". In: *NIPS*. 2017.

[Wei00] Y. Weiss. "Correctness of local probability propagation in graphical models with loops". In: *Neural Computation* 12 (2000), pp. 1–41.

[Wei+13] K. Wei, Y. Liu, K. Kirchhoff, and J. Bilmes. "Using Document Summarization Techniques for Speech Data Subset Selection." In: *HLT-NAACL*. 2013, pp. 721–726.

[Wei+14] K. Wei, Y. Liu, K. Kirchhoff, and J. Bilmes. "Unsupervised Submodular Subset Selection for Speech Data". In: *Proc. IEEE Intl. Conf. on Acoustics, Speech, and Signal Processing*. Florence, Italy, 2014.

[Wei+15a] K. Wei, R. Iyer, S. Wang, W. Bai, and J. Bilmes. "How to Intelligently Distribute Training Data to Multiple Compute Nodes: Distributed Machine Learning via Submodular Partitioning". In: *Neural Information Processing Society (NeurIPS, formerly NIPS) Workshop*. LearningSys Workshop, http://learningsys.org. Montreal, Canada, 2015.

[Wei+15b] K. Wei, R. Iyer, S. Wang, W. Bai, and J. Bilmes. "Mixed Robust/Average Submodular Partitioning: Fast Algorithms, Guarantees, and Applications". In: *Neural Information Processing Society (NeurIPS, formerly NIPS)*. Montreal, Canada, 2015.

[Wel11] M. Welling. "Bayesian Learning via Stochastic Gradient Langevin Dynamics". In: *ICML*. 2011.

[Wen+17] R. Wen, K. Torkkola, B. Narayanaswamy, and D. Madeka. "A Multi-Horizon Quantile Recurrent Forecaster". In: *NIPS Time Series Workshop*. 2017.

[Wen+19a] L. Wenliang, D. Sutherland, H. Strathmann, and A. Gretton. "Learning deep kernels for exponential family densities". In: *International Conference on Machine Learning*. 2019, pp. 6737–6746.

[Wen+19b] F. Wenzel, T. Galy-Fajou, C. Donner, M. Kloft, and M. Opper. "Efficient Gaussian Process Classification Using Polya-Gamma Data Augmentation". In: *AAAI*. 2019.

[Wen+20a] C. Wendler, A. Amrollahi, B. Seifert, A. Krause, and M. Püschel. "Learning set functions that are sparse in non-orthogonal Fourier bases". In: *arXiv preprint arXiv:2010.00439* (2020).

[Wen+20b] F. Wenzel, K. Roth, B. S. Veeling, J. Świątkowski, L. Tran, S. Mandt, J. Snoek, T. Salimans, R. Jenatton, and S. Nowozin. "How Good is the Bayes Posterior in Deep Neural Networks Really?" In: *ICML*. 2020.

[Wen+20c] F. Wenzel, J. Snoek, D. Tran, and R. Jenatton. "Hyperparameter Ensembles for Robustness and Uncertainty Quantification". In: *NIPS*. 2020.

[Wen21] L. Weng. "What are diffusion models?" In: *lilianweng.github.io/lil-log* (2021).

[Wes03] M. West. "Bayesian Factor Regression Models in the "Large p, Small n" Paradigm". In: *Bayesian Statistics 7* (2003).

[Wes87] M. West. "On scale mixtures of normal distributions". In: *Biometrika* 74 (1987), pp. 646–648.

[WF01a] Y. Weiss and W. T. Freeman. "Correctness of belief propagation in Gaussian graphical models of arbitrary topology". In: *Neural Computation* 13.10 (2001), pp. 2173–2200.

[WF01b] Y. Weiss and W. T. Freeman. "On the optimality of solutions of the max-product belief propagation algorithm in arbitrary graphs". In: *IEEE Trans. Information Theory, Special Issue on Codes on Graphs and Iterative Algorithms* 47.2 (2001), pp. 723–735.

[WF14] Z. Wang and N. de Freitas. "Theoretical Analysis of Bayesian Optimisation with Unknown Gaussian Process Hyper-Parameters". In: (2014). arXiv: 1406.7758 [stat.ML].

[WF16] Z. Wang and N. de Freitas. "Theoretical Analysis of Bayesian Optimisation with Unknown Gaussian Process Hyper-Parameters". In: *BayesOpt Workshop*. 2016.

[WG18] M. Wu and N. Goodman. "Multimodal Generative Models for Scalable Weakly-Supervised Learning". In: *NIPS*. 2018.

[WGY21] G. Weiss, Y. Goldberg, and E. Yahav. "Thinking Like Transformers". In: *ICML*. 2021.

[WH02] M. Welling and G. E. Hinton. "A new learning algorithm for mean field Boltzmann machines". In: *International Conference on Artificial Neural Networks*. Springer. 2002, pp. 351–357.

[WH18] Y. Wu and K. He. "Group Normalization". In: *ECCV*. 2018.

[WH97] M. West and J. Harrison. *Bayesian forecasting and dynamic models*. Springer, 1997.

[WHD18] J. T. Wilson, F. Hutter, and M. P. Deisenroth. "Maximizing acquisition functions for Bayesian optimization". In: *NIPS*. 2018.

[Whi16] T. White. "Sampling Generative Networks". In: *arXiv* (2016).

[Whi88] P Whittle. "Restless bandits: activity allocation in a changing world". In: *J. Appl. Probab.* 25.A (1988), pp. 287–298.

[WHT19] Y. Wang, H. He, and X. Tan. "Truly Proximal Policy Optimization". In: *UAI*. 2019.

[WI20] A. G. Wilson and P. Izmailov. "Bayesian Deep Learning and a Probabilistic Perspective of Generalization". In: *NIPS*. 2020.

[WIB15] K. Wei, R. Iyer, and J. Bilmes. "Submodularity in Data Subset Selection and Active Learning". In: *Proceedings of the 32nd international conference on Machine learning*. Lille, France, 2015.

[Wie+14] D Wierstra, T Schaul, J Peters, and J Schmidhuber. "Natural Evolution Strategies". In: *JMLR* 15.1 (2014), pp. 949–980.

[Wik21] Wikipedia contributors. *CliffsNotes — Wikipedia, The Free Encyclopedia*. [Online; accessed 29-December-2021]. 2021.

[Wil+14] A. G. Wilson, E. Gilboa, A. Nehorai, and J. P. Cunningham. "Fast kernel learning for multidimensional pattern extrapolation". In: *Advances in Neural Information Processing Systems*. 2014, pp. 3626–3634.

[Wil14] A. G. Wilson. "Covariance kernels for fast automatic pattern discovery and extrapolation with Gaussian processes". PhD thesis. University of Cambridge, 2014.

[Wil+16] A. G. Wilson, Z. Hu, R. Salakhutdinov, and E. P. Xing. "Deep Kernel Learning". en. In: *AISTATS*. 2016, pp. 370–378.

[Wil+17] A. G. Wills, J. Hendriks, C. Renton, and B. Ninness. "A Bayesian Filtering Algorithm for Gaussian Mixture Models". In: (2017). arXiv: 1705.05495 [stat.ML].

[Wil20] A. G. Wilson. "The Case for Bayesian Deep Learning". In: (2020). arXiv: 2001.10995 [cs.LG].

[Wil+20a] J. T. Wilson, V. Borovitskiy, A. Terenin, P. Mostowsky, and M. P. Deisenroth. "Efficiently Sampling Functions from Gaussian Process Posteriors". In: *ICML*. 2020.

[Wil+20b] A. B. Wiltschko, B. Sanchez-Lengeling, B. Lee, E. Reif, J. Wei, K. J. McCloskey, L. Colwell, W. Qian, and Y. Wang. "Evaluating Attribution for Graph Neural Networks". In: (2020).

[Wil69] A. G. Wilson. "The use of entropy maximising models, in the theory of trip distribution, mode split and route split". In: *Journal of transport economics and policy* (1969), pp. 108–126.

[Wil92] R. J. Williams. "Simple statistical gradient-following algorithms for connectionist reinforcement learning". In: *MLJ* 8.3-4 (1992), pp. 229–256.

[Wil98] C. Williams. "Computation with infinite networks". In: *Neural Computation* 10.5 (1998), pp. 1203–1216.

[Win] J. Winn. *VIBES*.

[Win+19] J. Winn, C. Bishop, T. Diethe, J. Guiver, and Y. Zaykov. *Model-based Machine Learning*. 2019.

[WIP20] J. Watson, A. Imohiosen, and J. Peters. "Active Inference or Control as Inference? A Unifying View". In: *International Workshop on Active Inference*. 2020.

[Wit] *DEEPFAKES: PREPARE NOW (PERSPECTIVES FROM BRAZIL)*. https://lab.witness.org/brazil-deepfakes-prepare-now/. Accessed: 2021-08-18.

[Wiy+19] R. R. Wiyatno, A. Xu, O. Dia, and A. de Berker. "Adversarial Examples in Modern Machine Learning: A Review". In: (2019). arXiv: 1911.05268 [cs.LG].

[WJ08] M. J. Wainwright and M. I. Jordan. "Graphical models, exponential families, and variational inference". In: *Foundations and Trends in Machine Learning* 1–2 (2008), pp. 1–305.

[WJ21] Y. Wang and M. I. Jordan. *Desiderata for Representation Learning: A Causal Perspective*. 2021. arXiv: 2109.03795 [stat.ML].

[WJW03] M. J. Wainwright, T. S. Jaakkola, and A. S. Willsky. "Tree-based reparameterization framework for analysis of sum-product and related algorithms". In: *IEEE Trans. on Information Theory* 49.5 (2003), pp. 1120–1146.

[WK18] E. Wong and Z. Kolter. "Provable defenses against adversarial examples via the convex outer adversarial polytope". In: *International Conference on Machine Learning*. PMLR. 2018, pp. 5286–5295.

[WK96] G. Widmer and M. Kubat. "Learning in the presence of concept drift and hidden contexts". In: *Mach. Learn.* 23.1 (1996), pp. 69–101.

[WKS21] V. Wild, M. Kanagawa, and D. Sejdinovic. "Connections and Equivalences between the Nyström Method and Sparse Variational Gaussian Processes". In: (2021). arXiv: 2106.01121 [stat.ML].

[WL14a] N. Whiteley and A. Lee. "Twisted particle filters". en. In: *Annals of Statistics* 42.1 (Feb. 2014), pp. 115–141.

[WL14b] J. L. Williams and R. A. Lau. "Approximate evaluation of marginal association probabilities with belief propagation". In: *IEEE Trans. Aerosp. Electron. Syst.* 50.4 (2014).

[WL19] A. Wehenkel and G. Louppe. "Unconstrained Monotonic Neural Networks". In: *NIPS*. 2019.

[WLL16] W. Wang, H. Lee, and K. Livescu. "Deep Variational Canonical Correlation Analysis". In: *arXiv* (2016).

[WLZ19] D. Widmann, F. Lindsten, and D. Zachariah. "Calibration tests in multi-class classification: A unifying framework". In: *NIPS*. Curran Associates, Inc., 2019, pp. 12236–12246.

[WM12] K. Wakabayashi and T. Miura. "Forward-Backward Activation Algorithm for Hierarchical Hidden Markov Models". In: *NIPS*. 2012.

[WMR17] S. Wachter, B. Mittelstadt, and C. Russell. "Counterfactual explanations without opening the black box: Automated decisions and the GDPR". In: *Harv. JL & Tech.* 31 (2017), p. 841.

[WMR18] S. Wachter, B. Mittelstadt, and C. Russell. *Counterfactual Explanations without Opening the Black Box: Automated Decisions and the GDPR*. 2018. arXiv: 1711.00399 [cs.AI].

[WN01] E. A. Wan and A. T. Nelson. "Dual EKF Methods". In: *Kalman Filtering and Neural Networks*. Ed. by S. Haykin. Wiley, 2001.

[WN07] D. Wipf and S. Nagarajan. "A new view of automatic relevancy determination". In: *NIPS*. 2007.

[WN10] D. Wipf and S. Nagarajan. "Iterative Reweighted ℓ_1 and ℓ_2 Methods for Finding Sparse Solutions". In: *J. of Selected Topics in Signal Processing (Special Issue on Compressive Sensing)* 4.2 (2010).

[WN15] A. G. Wilson and H. Nickisch. "Kernel Interpolation for Scalable Structured Gaussian Processes (KISS-GP)". In: *ICML*. ICML'15. JMLR.org, 2015, pp. 1775–1784.

[WN18] C. K. I. Williams and C. Nash. "Autoencoders and Probabilistic Inference with Missing Data: An Exact Solution for The Factor Analysis Case". In: (2018). arXiv: 1801.03851 [cs.LG].

[WO12] M. Wiering and M. van Otterlo, eds. *Reinforcement learning: State-of-the-art*. Springer, 2012.

[Woł+21] M. Wołczyk, M. Zając, R. Pascanu, Ł. Kuciński, and P. Miłoś. "Continual World: A Robotic Benchmark For Continual Reinforcement Learning". In: *NIPS*. 2021.

[Wol76] P. Wolfe. "Finding the nearest point in a polytope". In: *Mathematical Programming* 11 (1976), pp. 128–149.

[Wol92] D. Wolpert. "Stacked Generalization". In: *Neural Networks* 5.2 (1992), pp. 241–259.

[Woo+09] F. Wood, C. Archambeau, J. Gasthaus, L. James, and Y. W. Teh. "A Stochastic Memoizer for Sequence Data". In: *ICML*. 2009.

[Woo+11] F. Wood, J. Gasthaus, C. Archambeau, L. James, and Y. W. Teh. "The sequence memoizer". In: *Comm. of the ACM* 54.2 (2011), pp. 91–98.

[Woo+19] B. Woodworth, S. Gunasekar, P. Savarese, E. Moroshko, I. Golan, J. Lee, D. Soudry, and N. Srebro. "Kernel and Rich Regimes in Overparametrized Models". In: (2019). arXiv: 1906.05827 [cs.LG].

[WP18] H. Wang and H. Poon. "Deep Probabilistic Logic: A Unifying Framework for Indirect Supervision". In: *EMNLP*. 2018.

[WP19] S. Wiegreffe and Y. Pinter. "Attention is not not explanation". In: *arXiv preprint arXiv:1908.04626* (2019).

[WR15] F. Wang and C. Rudin. "Falling Rule Lists". In: *Proceedings of the Eighteenth International Conference on Artificial Intelligence and Statistics*. Ed. by G. Lebanon and S. V. N. Vishwanathan. Vol. 38. Proceedings of Machine Learning Research. San Diego, California, USA: PMLR, 2015, pp. 1013–1022.

[WRN10] D. Wipf, B. Rao, and S. Nagarajan. "Latent Variable Bayesian Models for Promoting Sparsity". In: *IEEE Transactions on Information Theory* (2010).

[WRZH04] M. Welling, M. Rosen-Zvi, and G. Hinton. "Exponential family harmoniums with an application to information retrieval". In: *NIPS-14*. 2004.

[WS01] C. K. I. Williams and M. Seeger. "Using the Nyström Method to Speed Up Kernel Machines". In: *NIPS*. Ed. by T. K. Leen, T. G. Dietterich, and V Tresp. MIT Press, 2001, pp. 682–688.

[WS05] M. Welling and C. Sutton. "Learning in Markov Random Fields with Contrastive Free Energies". In: *Tenth International Workshop on Artificial Intelligence and Statistics (AISTATS)*. 2005.

[WS93] D. Wolpert and C. Strauss. "What Bayes has to say about the evidence procedure". In: *Proc. Workshop on Maximum Entropy and Bayesian methods*. 1993.

[WSG21] C. Wang, S. Sun, and R. Grosse. "Beyond Marginal Uncertainty: How Accurately can Bayesian Regression Models Estimate Posterior Predictive Correlations?" In: *AISTATS*. Ed. by A. Banerjee and K. Fukumizu. Vol. 130. Proceedings of Machine Learning Research. PMLR, 2021, pp. 2476–2484.

[WSN00] B. Williams, T. Santner, and W. Notz. "Sequential design of computer experiments to minimize integrated response functions". In: *Statistica Sinica* 10 (2000), pp. 1133–1152.

[WSS21] W. J. Wilkinson, S. Särkkä, and A. Solin. "Bayes-Newton Methods for Approximate Bayesian Inference with PSD Guarantees". In: (2021). arXiv: 2111.01721 [stat.ML].

[WT01] M. Welling and Y.-W. Teh. "Belief Optimization for Binary Networks: a Stable Alternative to Loopy Belief Propagation". In: *UAI*. 2001.

[WT19] R. Wen and K. Torkkola. "Deep Generative Quantile-Copula Models for Probabilistic Forecasting". In: *ICML*. 2019.

[WT90] G. Wei and M. Tanner. "A Monte Carlo implementation of the EM algorithm and the poor man's data augmentation algorithms". In: *JASA* 85.411 (1990), pp. 699–704.

[WTB20] Y. Wen, D. Tran, and J. Ba. "BatchEnsemble: an Alternative Approach to Efficient Ensemble and Lifelong Learning". In: *ICLR*. 2020.

[WTN19] Y. Wu, G. Tucker, and O. Nachum. *Behavior Regularized Offline Reinforcement Learning*. arXiv:1911.11361. 2019.

[Wu+06] Y Wu, D Hu, M Wu, and X Hu. "A Numerical-Integration Perspective on Gaussian Filters". In: *IEEE Trans. Signal Process.* 54.8 (2006), pp. 2910–2921.

[Wu+17] Y. Wu, E. Mansimov, S. Liao, R. Grosse, and J. Ba. "Scalable trust-region method for deep reinforcement learning using Kronecker-factored approximation". In: *NIPS*. 2017.

[Wu+19a] A. Wu, S. Nowozin, E. Meeds, R. E. Turner, J. M. Hernández-Lobato, and A. L. Gaunt. "Fixing Variational Bayes: Deterministic Variational Inference for Bayesian Neural Networks". In: *ICLR*. 2019.

[Wu+19b] M. Wu, S. Parbhoo, M. Hughes, R. Kindle, L. Celi, M. Zazzi, V. Roth, and F. Doshi-Velez. "Regional tree regularization for interpretability in black box models". In: *AAAI* (2019).

[Wu+21] T. Wu, M. T. Ribeiro, J. Heer, and D. S. Weld. "Polyjuice: Generating Counterfactuals for Explaining, Evaluating, and Improving Models". In: *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics, 2021.

[Wüt+16] M. Wüthrich, S. Trimpe, C. Garcia Cifuentes, D. Kappler, and S. Schaal. "A new perspective and extension of the Gaussian Filter". en. In: *The International Journal of Robotics Research* 35.14 (2016), pp. 1731–1749.

[WW12] Y. Wu and D. P. Wipf. "Dual-Space Analysis of the Sparse Linear Model". In: *NIPS*. 2012.

[WY02] D. Wilkinson and S. Yeung. "Conditional simulation from highly structured Gaussian systems with application to blocking-MCMC for the Bayesian analysis of very large linear models". In: *Statistics and Computing* 12 (2002), pp. 287–300.

[WYG14] J. Wen, C.-N. Yu, and R. Greiner. "Robust Learning under Uncertain Test Distributions: Relating Covariate Shift to Model Misspecification". In: *ICML*. Vol. 32. Proceedings of Machine Learning Research. PMLR, 2014, pp. 631–639.

[WZ19] J. Wei and K. Zou. "EDA: Easy Data Augmentation Techniques for Boosting Performance on Text Classification Tasks". In: *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*. Hong Kong, China: Association for Computational Linguistics, Nov. 2019, pp. 6382–6388.

[WZR20] S. Wu, H. R. Zhang, and C. Ré. "Understanding and Improving Information Transfer in Multi-Task Learning". In: *International Conference on Learning Representations*. 2020.

[XC19] Z. Xia and A. Chakrabarti. "Training Image Estimators without Image Ground-Truth". In: *NIPS*. 2019.

[XD18] J. Xu and G. Durrett. "Spherical Latent Spaces for Stable Variational Autoencoders". In: *EMNLP*. 2018.

[Xia+21] K. Xiao, L. Engstrom, A. Ilyas, and A. Madry. "Noise or Signal: The Role of Image Backgrounds in Object Recognition". In: *ICLR*. 2021.

[Xie+16] J. Xie, Y. Lu, S.-C. Zhu, and Y. N. Wu. "A Theory of Generative ConvNet". In: *ICML*. 2016.

[Xie+18] J. Xie, Y. Lu, R. Gao, and Y. N. Wu. "Cooperative Learning of Energy-Based Model and Latent Variable Model via MCMC Teaching." In: *AAAI*. Vol. 1. 6. 2018, p. 7.

[Xie+22] S. M. Xie, A. Raghunathan, P. Liang, and T. Ma. "An Explanation of In-context Learning as Implicit Bayesian Inference". In: *ICLR*. 2022.

[XJ96] L. Xu and M. I. Jordan. "On Convergence Properties of the EM Algorithm for Gaussian Mixtures". In: *Neural Computation* 8 (1996), pp. 129–151.

[XKV22] Z. Xiao, K. Kreis, and A. Vahdat. "Tackling the Generative Learning Trilemma with Denoising Diffusion GANs". In: *ICLR*. 2022.

[Xu+06] Z. Xu, V. Tresp, K. Yu, and H.-P. Kriegel. "Infinite hidden relational models". In: *UAI*. 2006.

[Xu+07] Z. Xu, V. Tresp, S. Yu, K. Yu, and H.-P. Kriegel. "Fast Inference in Infinite Hidden Relational Models". In: *Workshop on Mining and Learning with Graphs*. 2007.

[Xu+15] J. Xu, L. Mukherjee, Y. Li, J. Warner, J. M. Rehg, and V. Singh. "Gaze-enabled egocentric video summarization via constrained submodular maximization". In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2015, pp. 2235–2244.

[Xu18] J. Xu. "Distance-based Protein Folding Powered by Deep Learning". In: (2018). arXiv: 1811.03481 [q-bio.BM].

[Xu+19] M. Xu, M. Quiroz, R. Kohn, and S. A. Sisson. "Variance reduction properties of the reparameterization trick". In: *AISTATS*. Ed. by K. Chaudhuri and M. Sugiyama. Vol. 89. Proceedings of Machine Learning Research. PMLR, 2019, pp. 2711–2720.

[Yad+18] S. Yadlowsky, H. Namkoong, S. Basu, J. Duchi, and L. Tian. "Bounds on the conditional and average treatment effect with unobserved confounding factors". In: *arXiv e-prints*, arXiv:1808.09521 (Aug. 2018), arXiv:1808.09521. arXiv: 1808.09521 [stat.ME].

[Yad+21] S. Yadlowsky, S. Fleming, N. Shah, E. Brunskill, and S. Wager. *Evaluating Treatment Prioritization Rules via Rank-Weighted Average Treatment Effects*. 2021. arXiv: 2111.07966 [stat.ME].

[Yan+17] S Yang, L Xie, X Chen, X Lou, X Zhu, D Huang, and H Li. "Statistical parametric speech synthesis using generative adversarial networks under a multi-task learning framework". In: *IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*. 2017, pp. 685–691.

[Yan19] G. Yang. "Scaling Limits of Wide Neural Networks with Weight Sharing: Gaussian Process Behavior, Gradient Independence, and Neural Tangent Kernel Derivation". In: (2019). arXiv: 1902.04760 [cs.NE].

[Yan+21] J. Yang, K. Zhou, Y. Li, and Z. Liu. "Generalized OOD Detection: A Survey". In: (2021).

[Yan74] H Yanai. "Unification of various techniques of multivariate analysis by means of generalized coefficient of determination (GCD)". In: *J. Behaviormetrics* 1.1 (1974), pp. 45–54.

[Yan81] M. Yannakakis. "Computing the minimum fill-in is NP-complete". In: *SIAM J. Alg. Discrete Methods* 2 (1981), pp. 77–79.

[Yao+18a] Y. Yao, A. Vehtari, D. Simpson, and A. Gelman. "Using Stacking to Average Bayesian Predictive Distributions (with Discussion)". en. In: *Bayesian Analysis* 13.3 (2018), pp. 917–1007.

[Yao+18b] Y. Yao, A. Vehtari, D. Simpson, and A. Gelman. "Yes, but Did It Work?: Evaluating Variational Inference". In: *ICML*. Vol. 80. Proceedings of Machine Learning Research. PMLR, 2018, pp. 5581–5590.

[YBM20] Y. Yang, R. Bamler, and S. Mandt. *Improving Inference for Neural Image Compression*. 2020. arXiv: 2006.04240 [eess.IV].

[YBW15] F. Yang, S. Balakrishnan, and M. J. Wainwright. "Statistical and Computational Guarantees for the Baum-Welch Algorithm". In: (2015). arXiv: 1512.08269 [stat.ML].

[Yed11] J. S. Yedidia. "Message-Passing Algorithms for Inference and Optimization". In: *J. Stat. Phys.* 145.4 (2011), pp. 860–890.

[Yeh+18] C.-K. Yeh, J. S. Kim, I. E. H. Yen, and P. Ravikumar. *Representer Point Selection for Explaining Deep Neural Networks*. 2018. arXiv: 1811.09720 [cs.LG].

[Yeh+19a] C.-K. Yeh, C.-Y. Hsieh, A. Suggala, D. I. Inouye, and P. K. Ravikumar. "On the (in) fidelity and sensitivity of explanations". In: *Advances in Neural Information Processing Systems* 32 (2019), pp. 10967–10978.

[Yeh+19b] C.-K. Yeh, B. Kim, S. O. Arik, C.-L. Li, T. Pfister, and P. Ravikumar. "On completeness-aware concept-based explanations in deep neural networks". In: *arXiv preprint arXiv:1910.07969* (2019).

[Yeu+17] S. Yeung, A. Kannan, Y. Dauphin, and L. Fei-Fei. "Tackling Over-pruning in Variational Autoencoders". In: *ICML Workshop on "Principled Approaches to Deep Learning"*. 2017.

[Yeu91a] R. W. Yeung. "A new outlook on Shannon's information measures". In: *IEEE Trans. Inf. Theory* 37.3 (1991), pp. 466–474.

[Yeu91b] R. W. Yeung. "A new outlook on Shannon's information measures". In: *IEEE Trans. on Information Theory* 37 (1991), pp. 466–474.

[YFW00] J. Yedidia, W. T. Freeman, and Y. Weiss. "Generalized Belief Propagation". In: *NIPS*. 2000.

[YH21] G. Yang and E. J. Hu. "Feature Learning in Infinite-Width Neural Networks". In: *ICML*. 2021.

[Yin+19a] D. Yin, R. G. Lopes, J. Shlens, E. D. Cubuk, and J. Gilmer. "A Fourier Perspective on Model Robustness in Computer Vision". In: *NIPS*. 2019.

[Yin+19b] P. Yin, J. Lyu, S. Zhang, S. Osher, Y. Qi, and J. Xin. "Understanding Straight-Through Estimator in Training Activation Quantized Neural Nets". In: *ICLR*. 2019.

[Yin+19c] R. Ying, D. Bourgeois, J. You, M. Zitnik, and J. Leskovec. "Gnnexplainer: Generating explanations for graph neural networks". In: *Advances in neural information processing systems* 32 (2019), p. 9240.

[Yin+20] D. Yin, M. Farajtabar, A. Li, N. Levine, and A. Mott. "Optimization and Generalization of Regularization-Based Continual Learning: a Loss Approximation Viewpoint". In: (2020). arXiv: 2006.10974 [cs.LG].

[YK06] A. Yuille and D. Kersten. "Vision as Bayesian inference: analysis by synthesis?" en. In: *Trends Cogn. Sci.* 10.7 (2006), pp. 301–308.

[YK19] M. Yang and B. Kim. *Benchmarking Attribution Methods with Relative Feature Importance*. 2019. arXiv: 1907.09701 [cs.LG].

[YMT22] Y. Yang, S. Mandt, and L. Theis. "An Introduction to Neural Data Compression". In: (2022). arXiv: 2202.06533 [cs.LG].

[Yoo+18] K. Yoon, R. Liao, Y. Xiong, L. Zhang, E. Fetaya, R. Urtasun, R. Zemel, and X. Pitkow. "Inference in Probabilistic Graphical Models by Graph Neural Networks". In: *ICLR Workshop*. 2018.

[You19] A. Young. "Consistency without inference: Instrumental variables in practical application". In: (2019).

[You89] L. Younes. "Parameter estimation for imperfectly observed Gibbsian fields". In: *Probab. Theory and Related Fields* 82 (1989), pp. 625–645.

[You99] L. Younes. "On the convergence of Markovian stochastic algorithms with rapidly decreasing ergodicity rates". In: *Stochastics: An International Journal of Probability and Stochastic Processes* 65.3-4 (1999), pp. 177–228.

[Yu+06] S. Yu, K. Yu, V. Tresp, K. H-P., and M. Wu. "Supervised probabilistic principal component analysis". In: *KDD*. 2006.

[Yu10] S.-Z. Yu. "Hidden Semi-Markov Models". In: *Artificial Intelligence J.* 174.2 (2010).

[Yu+16] F. X. X. Yu, A. T. Suresh, K. M. Choromanski, D. N. Holtmann-Rice, and S. Kumar. "Orthogonal Random Features". In: *NIPS*. Curran Associates, Inc., 2016, pp. 1975–1983.

[Yu+17] L. Yu, W. Zhang, J. Wang, and Y. Yu. "Seqgan: Sequence generative adversarial nets with policy gradient". In: *Thirty-first AAAI conference on artificial intelligence*. 2017.

[Yu+18] Y. Yu et al. "Dynamic Control Flow in Large-Scale Machine Learning". In: *Proceedings of the Thirteenth EuroSys Conference*. EuroSys '18. Association for Computing Machinery, 2018.

[Yu+20] L. Yu, Y. Song, J. Song, and S. Ermon. "Training Deep Energy-Based Models with f-Divergence Minimization". In: *arXiv preprint arXiv:2003.03463* (2020).

[Yu+21] J. Yu, X. Li, J. Y. Koh, H. Zhang, R. Pang, J. Qin, A. Ku, Y. Xu, J. Baldridge, and Y. Wu. "Vector-quantized Image Modeling with Improved VQGAN". In: (2021). arXiv: 2110.04627 [cs.CV].

[Yu+22] J. Yu et al. "Scaling Autoregressive Models for Content-Rich Text-to-Image Generation". 2022.

[Yua+19] X. Yuan, P. He, Q. Zhu, and X. Li. "Adversarial Examples: Attacks and Defenses for Deep Learning". en. In: *IEEE Trans. Neural Networks and Learning Systems* 30.9 (2019), pp. 2805–2824.

[Yui01] A. Yuille. "CCCP algorithms to minimze the Bethe and Kikuchi free energies: convergent alternatives to belief propagation". In: *Neural Computation* 14 (2001), pp. 1691–1722.

[YW04] C. Yanover and Y. Weiss. "Finding the M Most Probable Configurations in Arbitrary Graphical Models". In: *NIPS*. 2004.

[YWX17] J.-g. Yao, X. Wan, and J. Xiao. "Recent advances in document summarization". In: *Knowledge and Information Systems* 53.2 (2017), pp. 297–336.

[YZ19] G. Yaroslavtsev and S. Zhou. "Approximate F_2-Sketching of Valuation Functions". In: *Approximation, Randomization, and Combinatorial Optimization. Algorithms and Techniques (APPROX-/RANDOM 2019)*. Ed. by D. Achlioptas and L. A. Végh. Vol. 145. Leibniz International Proceedings in Informatics (LIPIcs). Dagstuhl, Germany: Schloss Dagstuhl–Leibniz-Zentrum fuer Informatik, 2019, 69:1–69:21.

[YZ22] Y. Yang and P. Zhai. "Click-through rate prediction in online advertising: A literature review". In: *Inf. Process. Manag.* 59.2 (2022), p. 102853.

[ZA12] J. Zou and R. Adams. "Priors for Diversity in Generative Latent Variable Models". In: *NIPS*. 2012.

[Zaf+22] M. Zaffran, A. Dieuleveut, O. Féron, Y. Goude, and J. Josse. "Adaptive Conformal Predictions for Time Series". In: (2022). arXiv: 2202.07282 [stat.ML].

[Zai+20] S. Zaidi, A. Zela, T. Elsken, C. Holmes, F. Hutter, and Y. W. Teh. "Neural Ensemble Search for Performant and Calibrated Predictions". In: (2020). arXiv: 2006.08573 [cs.LG].

[Zan21] N. Zanichelli. *IAML Distill Blog: Transformers in Vision*. 2021.

[ZB18] T. Zhou and J. Bilmes. "Minimax curriculum learning: Machine teaching with desirable difficulties and scheduled diversity". In: *International Conference on Learning Representations*. 2018.

[ZB21] B. Zhao and H. Bilen. "Dataset Condensation with Differentiable Siamese Augmentation". In: *Proceedings of the 38th International Conference on Machine Learning, ICML 2021, 18-24 July 2021, Virtual Event*. Ed. by M. Meila and T. Z. 0001. Vol. 139. Proceedings of Machine Learning Research. PMLR, 2021, pp. 12674–12685.

[Zbo+21] J. Zbontar, L. Jing, I. Misra, Y. LeCun, and S. Deny. "Barlow twins: Self-supervised learning via redundancy reduction". In: *arXiv preprint arXiv:2103.03230* (2021).

[ZDK15] J. Zhang, J. Djolonga, and A. Krause. "Higher-order inference for multi-class log-supermodular models". In: *Proceedings of the IEEE International Conference on Computer Vision*. 2015, pp. 1859–1867.

[ZE01a] B. Zadrozny and C. Elkan. "Obtaining calibrated probability estimaets from decision trees and naive Bayesian classifiers". In: *ICML*. 2001.

[ZE01b] B. Zadrozny and C. Elkan. "Transforming classifier scores into accurate multiclass probability estimates". In: *KDD*. 2001.

[Zec+18a] J. R. Zech, M. A. Badgeley, M. Liu, A. B. Costa, J. J. Titano, and E. K. Oermann. "Variable generalization performance of a deep learning model to detect pneumonia in chest radiographs: A cross-sectional study". en. In: *PLoS Med.* 15.11 (Nov. 2018), e1002683.

[Zec+18b] J. R. Zech, M. A. Badgeley, M. Liu, A. B. Costa, J. J. Titano, and E. K. Oermann. "Variable generalization performance of a deep learning model to detect pneumonia in chest radiographs: a cross-sectional study". In: *PLoS medicine* 15.11 (2018), e1002683.

Author: Kevin P. Murphy. (C) MIT Press. CC-BY-NC-ND license

[Zel76] A. Zellner. "Bayesian and non-Bayesian analysis of the regression model with multivariate Student-t error terms". In: *JASA* 71.354 (1976), pp. 400–405.

[Zel86] A. Zellner. "On assessing prior distributions and Bayesian regression analysis with g-prior distributions". In: *Bayesian inference and decision techniques, Studies of Bayesian and Econometrics and Statistics volume 6*. North Holland, 1986.

[Zen+18] C. Zeno, I. Golan, E. Hoffer, and D. Soudry. "Task Agnostic Continual Learning Using Online Variational Bayes". In: (2018). arXiv: 1803.10123 [stat.ML].

[Zen+21] C. Zeno, I. Golan, E. Hoffer, and D. Soudry. "Task-Agnostic Continual Learning Using Online Variational Bayes With Fixed-Point Updates". en. In: *Neural Comput.* 33.11 (2021), pp. 3139–3177.

[Zer+19] J. Zerilli, A. Knott, J. Maclaurin, and C. Gavaghan. "Transparency in algorithmic and human decision-making: is there a double standard?" In: *Philosophy & Technology* 32.4 (2019), pp. 661–683.

[ZF14] M. D. Zeiler and R. Fergus. "Visualizing and understanding convolutional networks". In: *European conference on computer vision*. Springer. 2014, pp. 818–833.

[ZFV20] G. Zeni, M. Fontana, and S. Vantini. "Conformal Prediction: a Unified Review of Theory and New Challenges". In: (2020). arXiv: 2005.07972 [cs.LG].

[ZG21] D. Zou and Q. Gu. "On the Convergence of Hamiltonian Monte Carlo with Stochastic Gradients". In: *ICML*. Ed. by M. Meila and T. Zhang. Vol. 139. Proceedings of Machine Learning Research. PMLR, 2021, pp. 13012–13022.

[ZGR21] L. Zhang, M. Goldstein, and R. Ranganath. "Understanding Failures in Out-of-Distribution Detection with Deep Generative Models". In: *ICML*. Ed. by M. Meila and T. Zhang. Vol. 139. Proceedings of Machine Learning Research. PMLR, 2021, pp. 12427–12436.

[Zha+13a] K. Zhang, B. Schölkopf, K. Muandet, and Z. Wang. "Domain Adaptation under Target and Conditional Shift". In: *Proceedings of the 30th International Conference on Machine Learning*. 2013, pp. 819–827.

[Zha+13b] K. Zhang, B. Scholkopf, K. Muandet, and Z. Wang. "Domain Adaptation under Target and Conditional Shift". In: *ICML*. Vol. 28. 2013.

[Zha+16] S. Zhai, Y. Cheng, R. Feris, and Z. Zhang. "Generative adversarial networks as variational training of energy based models". In: *arXiv preprint arXiv:1611.01799* (2016).

[Zha+17] C. Zhang, S. Bengio, M. Hardt, B. Recht, and O. Vinyals. "Understanding deep learning requires rethinking generalization". In: *ICLR*. 2017.

[Zha+18] G. Zhang, S. Sun, D. Duvenaud, and R. Grosse. "Noisy Natural Gradient as Variational Inference". In: *ICML*. 2018.

[Zha+19a] X. Zhai, J. Puigcerver, A. Kolesnikov, P. Ruyssen, C. Riquelme, M. Lucic, J. Djolonga, A. S. Pinto, M. Neumann, A. Dosovitskiy, et al. "A large-scale study of representation learning with the visual task adaptation benchmark". In: *arXiv preprint arXiv:1910.04867* (2019).

[Zha+19b] C. Zhang, J. Butepage, H. Kjellstrom, and S. Mandt. "Advances in Variational Inference". In: *IEEE PAMI* (2019), pp. 2008–2026.

[Zha+19c] H. Zhang, I. Goodfellow, D. Metaxas, and A. Odena. "Self-attention generative adversarial networks". In: *International conference on machine learning*. PMLR. 2019, pp. 7354–7363.

[Zha+19d] L. Zhao, K. Korovina, W. Si, and. M. Cheung. "Approximate inference with Graph Neural Networks". 2019.

[Zha+20a] A. Zhang, Z. Lipton, M. Li, and A. Smola. *Dive into deep learning*. 2020.

[Zha+20b] H. Zhang, A. Li, J. Guo, and Y. Guo. "Hybrid models for open set recognition". In: *European Conference on Computer Vision*. Springer. 2020, pp. 102–117.

[Zha+20c] R. Zhang, B. Dai, L. Li, and D. Schuurmans. "GenDICE: Generalized Offline Estimation of Stationary Values". In: *ICLR*. 2020.

[Zha+20d] R. Zhang, C. Li, J. Zhang, C. Chen, and A. G. Wilson. "Cyclical stochastic gradient MCMC for Bayesian deep learning". In: *ICLR*. 2020.

[Zha+20e] X. Zhang, Y. Li, Z. Zhang, and Z.-L. Zhang. "f-GAIL: Learning f-Divergence for Generative Adversarial Imitation Learning". In: *Neural Information Processing Systems* (2020).

[Zha+20f] Y. Zhang, X. Chen, Y. Yang, A. Ramamurthy, B. Li, Y. Qi, and L. Song. "Efficient Probabilistic Logic Reasoning with Graph Neural Networks". In: *ICLR*. 2020.

[Zha+21a] X. Zhai, A. Kolesnikov, N. Houlsby, and L. Beyer. "Scaling vision transformers". In: *arXiv preprint arXiv:2106.04560* (2021).

[Zha+21b] H. Zhang, J. Y. Koh, J. Baldridge, H. Lee, and Y. Yang. "Cross-Modal Contrastive Learning for Text-to-Image Generation". In: *CVPR*. 2021.

[Zhe+15] S. Zheng, S. Jayasumana, B. Romera-Paredes, V. Vineet, Z. Su, D. Du, C. Huang, and P. Torr. "Conditional Random Fields as Recurrent Neural Networks". In: *ICCV*. 2015.

[Zho+19a] S. Zhou, M. Gordon, R. Krishna, A. Narcomey, L. F. Fei-Fei, and M. Bernstein. "HYPE: A Benchmark for Human eYe Perceptual Evaluation of Generative Models". In: *NIPS*. Curran Associates, Inc., 2019, pp. 3444–3456.

[Zho+19b] S. Zhou, M. L. Gordon, R. Krishna, A. Narcomey, L. Fei-Fei, and M. S. Bernstein. "HYPE: a benchmark for human eye perceptual evaluation of generative models". In: *Proceedings of the 33rd International Conference on Neural Information Processing Systems*. 2019, pp. 3449–3461.

[Zho20] G. Zhou. "Mixed Hamiltonian Monte Carlo for Mixed Discrete and Continuous Variable". In: (2020). arXiv: 1909.04852 [stat.CO].

[ZHT06] H. Zou, T. Hastie, and R. Tibshirani. "Sparse principal component analysis". In: *JCGS* 15.2 (2006), pp. 262–286.

[Zhu+17] J.-Y. Zhu, T. Park, P. Isola, and A. A. Efros. "Unpaired Image-to-Image Translation using Cycle-Consistent Adversarial Networks". In: *ICCV*. 2017.

[Zhu+18] X. Zhu, A. Singla, S. Zilles, and A. N. Rafferty. "An overview of machine teaching". In: *arXiv preprint arXiv:1801.05927* (2018).

[Zhu+21] F. Zhuang, Z. Qi, K. Duan, D. Xi, Y. Zhu, H. Zhu, H. Xiong, and Q. He. "A Comprehensive Survey on Transfer Learning". In: *Proc. IEEE* 109.1 (2021).

[Zie+08] B. D. Ziebart, A. L. Maas, J. A. Bagnell, and A. K. Dey. "Maximum Entropy Inverse Reinforcement Learning". In: *AAAI*. 2008, pp. 1433–1438.

[ZIE16] R. Zhang, P. Isola, and A. A. Efros. "Colorful image colorization". In: *European conference on computer vision*. Springer. 2016, pp. 649–666.

[ZIE17] R. Zhang, P. Isola, and A. A. Efros. "Split-brain autoencoders: Unsupervised learning by cross-channel prediction". In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2017, pp. 1058–1067.

[Zin+20] L. Zintgraf, K. Shiarlis, M. Igl, S. Schulze, Y. Gal, K. Hofmann, and S. Whiteson. "VariBAD: A Very Good Method for Bayes-Adaptive Deep RL via Meta-Learning". In: *ICLR*. 2020.

[Ziy+19] L. Ziyin, Z. Wang, P. P. Liang, R. Salakhutdinov, L.-P. Morency, and M. Ueda. "Deep gamblers: Learning to abstain with portfolio theory". In: *NIPS*. June 2019.

[ZL21] A. Zhou and S. Levine. "Training on Test Data with Bayesian Adaptation for Covariate Shift". In: *NIPS*. 2021.

[ZLF21] M. Zhang, S. Levine, and C. Finn. "MEMO: Test Time Robustness via Adaptation and Augmentation". In: (2021). arXiv: 2110.09506 [cs.LG].

[ZLG20] R. Zivan, O. Lev, and R. Galiki. "Beyond Trees: Analysis and Convergence of Belief Propagation in Graphs with Multiple Cycles". In: *AAAI*. 2020.

[ZMB21] B. Zhao, K. R. Mopuri, and H. Bilen. "Dataset Condensation with Gradient Matching". In: *International Conference on Learning Representations*. 2021.

[ZMG19] G. Zhang, J. Martens, and R. B. Grosse. "Fast Convergence of Natural Gradient Descent for Over-Parameterized Neural Networks". In: *NIPS*. 2019, pp. 8082–8093.

[ZML16] J. J. Zhao, M. Mathieu, and Y. LeCun. "Energy-based Generative Adversarial Network". In: (2016).

[Zob09] O. Zobay. "Mean field inference for the Dirichlet process mixture model". In: *Electronic J. of Statistics* 3 (2009), pp. 507–545.

[Zoe07] O. Zoeter. "Bayesian generalized linear models in a terabyte world". In: *Proc. 5th International Symposium on image and Signal Processing and Analysis*. 2007.

[Zon+18] B. Zong, Q. Song, M. R. Min, W. Cheng, C. Lumezanu, D. Cho, and H. Chen. "Deep Autoencoding Gaussian Mixture Model for Unsupervised Anomaly Detection". In: *ICLR*. 2018.

[ZP00] G. Zweig and M. Padmanabhan. "Exact alpha-beta computation in logarithmic space with application to map word graph construction". In: *ICSLP*. 2000.

[ZP96] N. Zhang and D. Poole. "Exploiting causal independence in Bayesian network inference". In: *JAIR* (1996), pp. 301–328.

[ZR19a] Z. Ziegler and A. Rush. "Latent Normalizing Flows for Discrete Sequences". In: *Proceedings of the 36th International Conference on Machine Learning*. 2019, pp. 7673–7682.

[ZR19b] Z. M. Ziegler and A. M. Rush. "Latent Normalizing Flows for Discrete Sequences". In: *ICML*. 2019.

[ZSB19] Q. Zhao, D. S. Small, and B. B. Bhattacharya. "Sensitivity analysis for inverse probability weighting estimators via the percentile bootstrap". In: *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 81.4 (2019), pp. 735–761. eprint: https://rss.onlinelibrary.wiley.com/doi/pdf/10.1111/rssb.12327.

[ZSE19] S. Zhao, J. Song, and S. Ermon. "InfoVAE: Information Maximizing Variational Autoencoders". In: *AAAI*. 2019.

[ZW11] D Zoran and Y Weiss. "From learning models of natural image patches to whole image restoration". In: *ICCV*. 2011.

[ZW12] D. Zoran and Y. Weiss. "Natural Images, Gaussian Mixtures and Dead Leaves". In: *NIPS*. 2012, pp. 1736–1744.

[ZWM97] C. S. Zhu, N. Y. Wu, and D. Mumford. "Minimax Entropy Principle and Its Application to Texture Modeling". In: *Neural Computation* 9.8 (1997).

[ZY08] J.-H. Zhao and P. L. H. Yu. "Fast ML Estimation for the Mixture of Factor Analyzers via an ECM Algorithm". In: *IEEE. Trans. on Neural Networks* 19.11 (2008).

[ZY21] Y. Zhang and Q. Yang. "A Survey on Multi-Task Learning". In: *IEEE Trans. Knowl. Data Eng.* (2021).

[ZY97] Z. Zhang and R. W. Yeung. "A non-Shannon-type conditional inequality of information quantities". In: *IEEE Transactions on Information Theory* 43.6 (1997), pp. 1982–1986.

[ZY98] Z. Zhang and R. W. Yeung. "On characterization of entropy function via information inequalities". In: *IEEE Transactions on Information Theory* 44.4 (1998), pp. 1440–1452.

Author: Kevin P. Murphy. (C) MIT Press. CC-BY-NC-ND license