Loss Functions Modulate the Optimal Bias-Variance Trade-off

Adam Bear (adambear@fas.harvard.edu)
Department of Psychology, Harvard University
Cambridge, MA 02143 USA

Fiery Cushman (cushman@fas.harvard.edu)
Department of Psychology, Harvard University
Cambridge, MA 02143 USA

Abstract

Prediction problems vary in the extent to which accuracy is rewarded and inaccuracy is penalized—i.e., in their loss functions. Here, we focus on a particular feature of loss functions that controls how much large errors are penalized relative to how much precise correctness is rewarded: convexity. We show that prediction problems with convex loss functions (i.e., those in which large errors are particularly harmful) favor simpler models that tend to be biased, but exhibit low variability. Conversely, problems with concave loss functions (in which precise correctness is particularly rewarded) favor more complex models that are less biased, but exhibit higher variability. We discuss how this relationship between the bias-variance trade-off and the shape of the loss function may help explain features of human psychology, such as dual-process psychology and fast versus slow learning strategies, and inform statistical inference.

Keywords: judgment; decision-making; dual-process theory; statistics

Introduction

"That's the difference between us, Allison. You wanna lose small; I wanna win big." - Harvey Specter, *Suits*

You're coaching a team in the "Bias-Variance Darts" championship. In Round 1 of the competition the high-scoring regions of the target are large, so anything close to the center of the board tends to win a lot of points. In Round 2 the high-scoring region is small, so unless you hit very close to the center, you won't get any points at all. You've got two players on your team. Lefty's darts always hit in the same small area just a few inches to the left of the center. Loosey's darts fall all over the place, in a wide scatter centered on the middle of the board. Which player should play which round?

As Figure 1 shows, there is a pretty clear answer to this problem. Lefty should play Round 1: because he's only a few inches off he'll usually be "close enough" on a dart board where close enough counts, and he's never too far off. (Loosey is a bad choice because she sometimes throws so far off that she gets very bad scores.) Meanwhile, Loosey should play round 2: although she'll often get no points, at least it is *possible* for her to hit the bulls eye and (occasionally) score points. Since Lefty is almost always off by a few inches, chances are far too small that he'll ever score a point at all.

In other words, before choosing a dart throwing strategy, we must first ask the following question: How close is "close enough" to count?

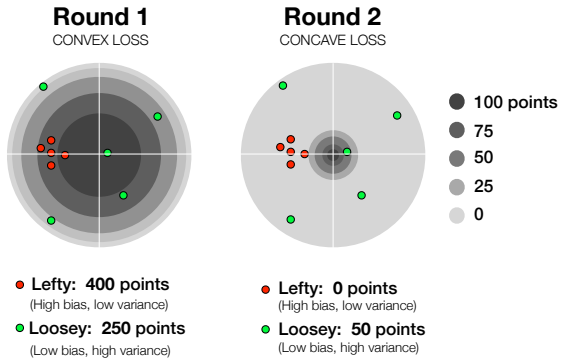

Figure 1: Suppose that a prediction problem is like a game of darts, and an estimation strategy is like a dart-thrower. A high-bias/low-variance strategy generates predictions (like Lefty's throws) with systematic errors that are small but consistent. A low-bias/high-variance strategy generates predictions (like Loosey's throws) with unsystematic errors that may be small or large. Now suppose that we apply a loss function to these errors, specifying how costly they are to a decision-maker (analogous to the points earned on the darts board). We find that when loss is convex—that is, one's goal is to avoid being very wrong—this tends to favor a high-bias/low-variance strategy. Conversely, when loss is concave—that is, one's goal is to be precisely right—this tends to favor a low-bias/high variance strategy. In other words, the structure of the loss function dictates the optimal bias-variance trade-off.

Bias-variance Trade-off

Most of life isn't a dart-throwing competition, but the trade-off between the two strategies illustrated above is analogous to a ubiquitous problem in learning, prediction and decision-making. When making predictions based on a limited sample of data, we face a dilemma about the optimal model to use to make these predictions. Models that are good at finding true regularities are also more likely to overfit noise in the training data. Conversely, models that are less responsive to noise are more likely to miss out on meaningful regularities (underfit).

This dilemma between underfitting and overfitting is often referred to as the "bias-variance trade-off" (Geman, Bienenstock, & Doursat, 1992). A high-bias model systematically misses the target of prediction, on average. However, because

they are simpler, models with more bias tend to be more robust to fluctuations in the training data they receive. They are systematically off-target, but consistent in how much they're off-target. In this sense they are like Lefty, whose throws are biased on average (to the left), but are highly consistent in their degree of error (slight—rarely egregious). In contrast, the predictions of a high-variance model are highly variable, based on the particular training data the model receives. This variability can lead to large unsystematic errors, even if bias (that is, *systematic* error) is low. In this sense they are like Loosey, whose throws are unbiased (i.e., the spatial average is the true center of the board), but exhibit large and inconsistent errors (i.e., often fall far from the center). Typically, there's some optimal level of model complexity that balances the relative magnitudes of systematic and unsystematic prediction errors.

Bias-variance tradeoffs are widely explored in statistics and machine learning. For example, in a regression problem in which a model is trying to uncover some relationship between known features and some real-valued output, there is a trade-off between fitting a complex model that might include many features or complex functional relationships (e.g., higher-order polynomials) and fitting a sparser model. The average prediction of the complex model is more likely to be close to the target, but individual predictions fluctuate more in response to noise in the training data. This overfitting problem has inspired techniques, like ridge and Lasso regression, that penalize complexity.

More generally, an ideal model must balance how much it pools information across times, categories, or contexts. In a dynamic system like the stock market, older information will generally be less relevant than newer information, but biasing predictions with this older information may usefully reduce the risk of detecting false patterns in recent data. In nested data structures, in which observations are stratified by groups or individuals, it can be useful to use information about one group to learn about another, even if this biases inference. Of course, too much of this sort of pooling may cause the model to miss out on important differences between groups.

Bias-variance trade-offs in human psychology

Just like artificial systems, humans routinely encounter the bias-variance trade-off. For instance, when making decisions, we face the choice between using simple heuristics or relying on complex models. Should we buy the most popular model of car or consider all the variables that could determine which specific model would be best for us? Should we take the route that has usually been fastest in the past or consider all the variables that might affect traffic patterns today? Should we simply hire employees who come from our *alma mater* or carefully consider all the information in their dossiers?

One dimension of this problem has been exhaustively explored elsewhere, and we set it aside here: heuristics are usually cognitively "cheap", while reasoning over complex models can sometimes be cognitively demanding (Kahneman, 2011; Dolan & Dayan, 2013). This is true, but it is not our concern.

Rather, even setting aside computational demands, the bias-variance trade-off may even help explain how—and when—simple heuristics can sometimes outperform more complex reasoning strategies (Brighton & Gigerenzer, 2015). In other words, simple "biased" models may not just be more efficient; they can actually be more accurate. Gigerenzer and colleagues (1999; 2007) have documented several real-world cases in which a trivial strategy, like investing one's money equally among several candidate funds, outperforms complicated estimation procedures, like the mean-variance analysis used in Modern Portfolio Theory. These biased heuristics dominate other methods particularly when data is limited, which may happen for one of two reasons. First, data is simply sparse in nature, and even if we wanted to acquire it, we couldn't. Second, data is, in principle, available, but difficult to generate. This latter problem is particularly prevalent in human decision-making, which often relies on costly sampling-based algorithms (Stewart, Chater, & Brown, 2006; Vul, Goodman, Griffiths, & Tenenbaum, 2014). If samples from a generative model take time to retrieve, the decider must act with limited information.

We take up a key question: under what circumstances should simpler models (e.g., estimating the value of a car according to the single feature of its market share) outperform more complex models? As we describe below, an important part of the answer is the structure of the costs associated with error. If small errors are benign but large errors become catastrophic (as in Dart Board 1), a simple model (high-bias, low variance) is typically favored. When such a model is employed, small errors will be frequent, but they cost little; large errors will be rare. On the other hand, if exact precision is rewarded and all errors (large or small) are roughly equal in value (as in Dart Board 2), a complex model (low-bias, high variance) is typically favored. Like playing the lottery, employing such a model makes it possible to at least occasionally reap the rewards of an accurate prediction. As we describe in the general discussion, this relationship provides new insight into when we should expect humans to rely on heuristic-like versus deliberative forms of cognition. Our next goal is to develop these ideas more precisely.

Convexity of Loss

Many methods in machine learning have been developed to find the optimal balance between bias and variance. A classic paper provides a general formulation of the bias-variance trade-off for all symmetric loss functions (James, 2003), including those that we will discuss. Here, we aim to extend this work—and apply it to cognitive science—by noting how a particular structural feature of loss functions that arise in decision-making influences the optimal allocation of bias versus variance. As the introductory example about darts illustrates, the ideal amount of systematic vs. unsystematic error that a model should strive for will depend on the shape of the reward function. The optimization problem in the simple

darts example can be characterized as a maximization of the probability of hitting any of the five reward regions multiplied by their point values. This is a very simple payoff scheme. In the real world, however, payoffs can be arbitrarily complex. Thus, we turn our attention to a property that we call "convexity".

(Although, strictly speaking, functions are defined categorically as either convex or not, we treat convexity as a graded property—similar to its use in some branches of finance. In particular, for all the functions we will discuss except one, we can simply characterize convexity as the magnitude of the function's second derivative for values of $x > 0$. Because these functions are symmetric, we need only consider positive values.)

Keeping with the convention in statistics and machine learning of characterizing payoffs in terms of *loss functions* to be minimized, we consider several continuous, symmetric functions that vary in their curvature. Convex loss functions, such as the commonly used quadratic (L2) loss, penalize errors at an accelerating rate. For example, supposing that Dart Board 1 (convex loss) is 10cm in radius, the payoff difference between being 1cm off and 4cm off is small, whereas the difference between being 6cm off and 9cm off is large. Concave functions (like Board 2), on the other hand, have the opposite property: errors near the target matter a lot, while errors far from the target matter relatively less.

In the real world, most loss functions are neither completely convex nor completely concave. For example, if a wealthy investor makes a risky leveraged bet on a speculative stock, his losses may be initially convex if the stock begins to fall and he loses more and more of his money at an accelerating rate. At some point, though, the investor will go broke, and thus, further losses will not hurt him as much. In other words, the loss function has a sigmoidal shape, which is initially convex and then becomes concave.

Nevertheless, for simplicity, we focus mostly on loss functions that are globally convex or concave. These functions can be used to approximate locally optimal model behavior. Roughly speaking, in contexts with convex losses it is essential to *avoid* major errors, as the costs of major errors outweigh the gains from precise accuracy. In the limit, a loss function of this sort turns into a version of the darts game in which you only need to hit the dartboard to win the game. With these types of payoffs, you want to avoid missing the board, and therefore, limiting variance is crucial.

In contexts with concave losses it pays to focus on precision, as even minor deviations from the target matter a lot. In the limit, the payoffs look similar to a version of the darts game in which you need to hit the center of the board exactly in order to get reward. In these cases, a model that is consistently biased away from that target will incur heavy losses. Meanwhile, occasional large unsystematic errors will not be much costlier than small errors. Thus, unsystematic variance may be relatively less harmful than systematic bias. In other words, it is better to miss badly often but hit exactly sometimes than to always miss by a small amount.

Simulation Results

To evaluate the trade-off between bias and variance, we consider a simplified setup in which prediction error is Gaussian, and the target y is deterministic. Let $\hat{f}$ denote a given model's prediction. Then we suppose that

$$\hat{f} - y \sim \mathcal{N}(\beta, \sigma^2).$$

In this Gaussian formulation of prediction error, β denotes the bias, and σ^2 denotes the variance. Since we stipulate that y is deterministic, there is no irreducible error, and the combination of β and σ fully determines prediction error.

We consider a few types of loss functions, which vary proportional to the absolute magnitude of error. Let $x = \hat{f} - y$. We consider the following functional forms:

$$L(x;n) = |x|^n \quad \text{(Power)}$$

$$L(x;c) = c \log |x| \quad \text{(Log)}$$

$$L(x;c) = e^{c|x|} \quad \text{(Exp)}$$

$$L(x;c) = -e^{-cx^2} \quad \text{(Mass)}$$

The shapes of these functions, for select parameters, are shown in Figure 2.

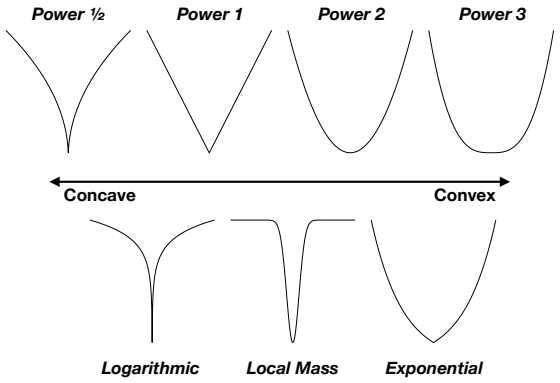

Figure 2: Example loss functions.

We focus primarily on the set of power functions, as the exponent n in these functions most directly modulates convexity. Loss is convex when $n > 1$ and concave when $n < 1$ (i.e., the second derivative is positive and negative for $x > 0$, respectively). We also consider exponential and logarithmic functions, which are well-known convex and concave functions, respectively. Finally, we consider the local mass function (Brainard & Freeman, 1997), which is akin to an inverted bell curve. This function has the interesting property that it rewards some degree of precision near the target, but is also

almost completely insensitive to the relative magnitudes of large errors. In mathematical terms, its second derivative is positive for small values of $x > 0$ (indicating convexity), but negative for larger values of x (indicating concavity). It perhaps maps most closely to many real-world contexts in which reasonable—though not perfect—precision near the target is important, but losses are capped for larger errors (Vul et al., 2014).

Now we consider how expected loss varies as a function of β and σ. We use a Reimann approximation on the bounded interval [-50, 50] to calculate the expected loss:

$$\int_{-\infty}^{\infty} L(x) f(x|\beta, \sigma^2) dx,$$

where $f(x|\beta, \sigma^2)$ is the normal density. Standardized expected losses are plotted in contour plots, where color indicates the magnitude of loss, as well as line plots, where these losses are plotted as a function of σ (square root of variance) for different β values (Figures 3 and 4).

The green line in each of the contour plots displays the optimal combination of β and σ that satisfy the constraint that $\beta + \sigma = C$, where C is a constant 'cost' that is varied. This line, which runs perpendicular to the contour lines, shows the relative benefits of minimizing bias versus variance. A line with a slope that's shallower than $\sigma = \beta$ indicates that one unit of bias is less costly than one unit of square-rooted variance, and vice versa for lines steeper than $\sigma = \beta$. (Note that the y-axis of the contour plots starts at $\sigma = 1$ while the x-axis starts at $\beta = 0$.)

Results from a set of power losses with different exponents (n) are shown in Figure 3. As can be seen by the green lines, the relative harm from β vs. σ decreases with n. For $n = 2$, the green line is simply the identity line, $\beta = \sigma$, which is implied by the well-known bias-variance decomposition for squared error, in which one unit of squared bias produces as much expected loss as one unit of variance (James, Witten, Hastie, & Tibshirani, 2013). However, the more convex loss function with $n = 3$ is relatively more harmed by (square-root of) variance, compared to bias. The opposite is true for $n < 2$.

Interestingly, when the loss function is concave ($n = .5$), adding prediction variance can actually *reduce* expected loss—if bias is high. This can be seen in the contour lines for high levels of loss, which bend back along the x-axis. It can also be seen in the non-monotonicity of the bluer (high β) lines in the line plot. This is in keeping with previous analysis of the bias-variance trade-off for general symmetric loss functions (James, 2003) and has also been noted for 0/1-loss functions used for classification (Friedman, 1997).

We see a similar trend in the other loss functions shown in Figure 4. (We set the parameters c for each function in such a way to make the plots as readable as possible, but the key qualitative results do not depend on particular values.) The logarithmic and local mass functions favor variance over bias, while the highly convex exponential function greatly favors bias. And for the former two functions, we again see non-

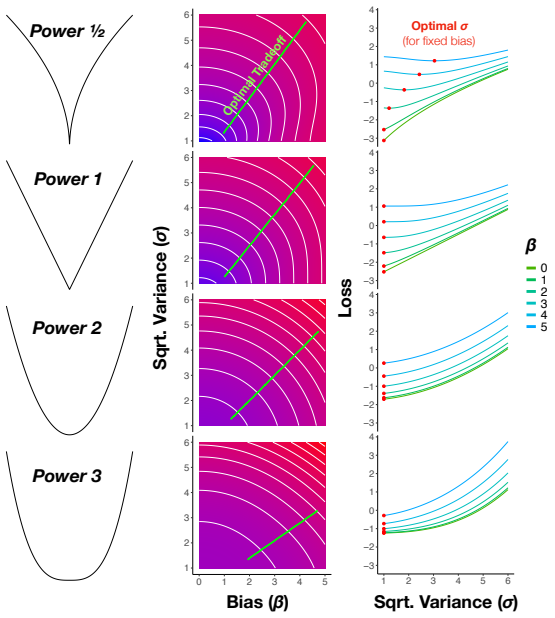

Figure 3: Standardized expected loss for power loss functions (exponents of .5, 1, 2, and 3). Middle: Contour plot with color indicating magnitude of expected loss (blue = low, red = high). Green lines show the combination of β and σ that minimize expected loss given the constraint that $\beta + \sigma = C$, where C is a constant. Right: Line plot with different lines for β, varying as a function of σ. Red dots indicate optimal σ values for each β.

monotonicity in the influence of σ on loss for high β values: when bias is high, variance can actually reduce loss.

In short, we find that the structure of the loss function importantly dictates the optimal balance between bias and variance. Loss functions with accelerating costs of errors favor simple, biased strategies over less biased, but noisy ones. Loss functions with decelerating costs of errors favor unsystematic noise, since even small systematic deviations from the correct answer incur large costs.

Discussion

The analysis presented here uncovers an important relationship between a problem's objective function and optimal model complexity. Because of the bias-variance trade-off, modelers are forced to navigate between underfitting data with an overly biased model and overfitting data with an excessively sensitive model. But the costs of underfitting and overfitting are not fixed in real environments. Although statisticians often use mathematically convenient loss functions like quadratic or absolute error, people face a wide array of loss functions. Sometimes, approximate solutions are acceptable, and agents must primarily avoid major errors. Other times, extreme rewards can only be achieved (or extreme costs can only be avoided) by being precisely correct. Our results suggest that relatively simpler models are favored in the

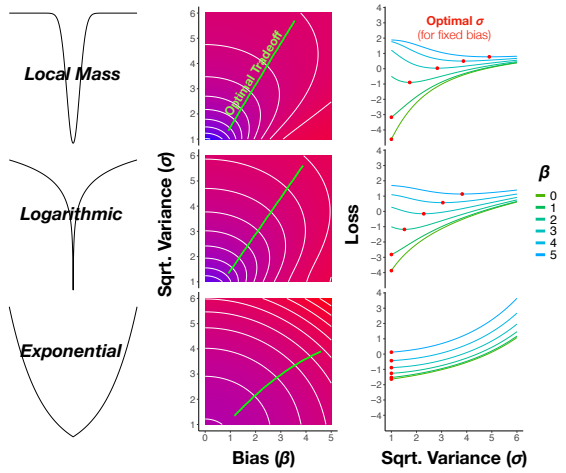

Figure 4: Standardized expected loss for other (non-power) loss functions, with $c = .5$ for local mass loss, $c = 1$ for log loss, and $c = .2$ for exponential loss. Middle: Contour plot with color indicating magnitude of expected loss (blue = low, red = high). Green lines show the combination of β and σ that minimize expected loss given the constraint that $\beta + \sigma = C$. Right: Line plot with different lines for β, varying as a function of σ. Red dots indicate optimal σ values for each β.

former kinds of cases and relatively more complicated models in the latter kinds. This relationship bears on key issues in statistics, machine learning, and cognitive science. Here, we focus on its implications for psychological research.

Application to human cognition

Much work in psychology has emphasized two broad types of cognitive processing: one type that is fast and automatic, and another type that is slow and deliberative (Evans, 2008; Kahneman, 2011; Sloman, 1996; Dolan & Dayan, 2013). A key question is how humans decide which type of processing to deploy for any given task. Typically this is presumed to involve a trade-off between better decisions (given by deliberation) and faster or less cognitively taxing decisions (given by automatic processes) (Kool, Gershman, & Cushman, 2017; Daw, Niv, & Dayan, 2005; Lee, Shimojo, & O'Doherty, 2014; Keramati, Dezfouli, & Piray, 2011; Shenhav, Botvinick, & Cohen, 2013). Yet some work illustrates contexts in which simple heuristics can in fact generate better decisions than complex cognitive strategies *even in the absence of cognitive constraints* because of the bias-variance trade-off (Brighton & Gigerenzer, 2015).

In order to view this issue in its most general form, it helps to conceive of human decision-making as a series of problems that are never identical, but often similar. Take, for example, your choice of what food to eat today. You have made many choices like this before, but the circumstances were not identical. One decision strategy is to choose the thing that you have enjoyed the most in the past. This kind of strategy—relying on historical average rewards in similar past circumstances—is characteristic of habitual (or "model-free") decision-making. It is a canonical example of automatic cognition. It is likely to be a biased decision strategy because the "average" situation you faced in the past almost certainly deviates systematically from the exact situation you face today. Thus, the optimal decision averaging across past episodes is unlikely to be the exact optimal decision today. But it is likely to be a low variance decision strategy because you are averaging over a very large number of similar past episodes, and what was best on average is unlikely to be *far* from best today.

Alternatively, rather than simply averaging over all past episodes of eating various foods, you could estimate a high-dimensional function relating specific features of any given day to the optimal food for that day—features such as whether you're hungry, whether you're sick, how much money you have, etc. For example, this might take the form of inferring a generative causal model of the reward structure of your environment. This kind of strategy is characteristic of planning (or "model based" decision making). It is a canonical example of controlled, deliberative cognition. The resulting model may reduce systematic bias, reflecting the fact that it attempts to derive the value of foods in this particular, unique circumstance (rather than on average). But it will likely introduce unsystematic errors generated by over-fitting a high-dimensional model to the limited training set of your past data.

In sum, two forms of value-guided decision-making widely regarded as canonical examples of an automatic process (habit, or "model-free" choice) and a deliberative process (planning, or "model-based" choice) may occupy distinct locations on a bias-variance trade-off. If so, then, according to our results, decision problems characterized by concave loss ("close doesn't count") should favor model-based deliberation while decision problems characterized by convex loss ("close is good enough") should favor relatively automatic, model-free processes.

Similarly, one can make decisions by copying the behavior of others on the assumption that their choices are adaptive (Deutsch & Gerard, 1955; Richerson & Boyd, 2008). This strategy introduces bias because other people do not face precisely the same decision that you do, and the "average" optimal action for them is unlikely to be precisely the same as the optimal action for you. But it may reduce variance because you are able to average over a large population of similarly situated people.

One can also make decisions by relying on an innate instinct shaped by natural selection. Natural selection favors instincts that worked well over many generations of ancestors who faced similar, but not identical, problems to the ones we face today. These decision strategies occupy a position of even greater bias than habit learning, since they average not only across diverse situations but also across diverse individuals. According to our results, policies derived from adaptive processes such as cultural or biological selection might be

favored in decision problems characterized by extreme convex loss, as compared to policies derived not only from complex model-based reasoning, but potentially also from simple model-free learning.

In other words, heuristics (including instinct, conformity, or habit) may dominate when it is more essential to avoid big mistakes than minimize small ones. In contrast, heuristics should be less useful, or even harmful, in contexts in which precision is essential—for example, in competitive winner-takes-all dynamics. In these cases, even an approach that is less accurate on average can be favored because it has the potential to be perfectly accurate. Such approaches might involve, for instance, sampling from a complex generative model. (Here, decision error could arise either from overfitting in the model (Gaissmaier, Wilke, Scheibehenne, McCanney, & Barrett, 2016) or from the cognitive constraints on the number of samples that can be drawn (Stewart et al., 2006; Vul et al., 2014).)

Future work could explore whether people modulate their prediction strategies based on the nature of the payoffs they face. There are several broad ways in which this might happen. At one extreme, people might be highly sensitive to specific loss functions of particular problems and, in turn, flexibly modulate their prediction strategies to optimally navigate the bias-variance trade-off as they move from task to task. For example, if people are asked to make online predictions with real-time reward feedback, the nature of this feedback might influence the ways in which people make their predictions.

At the other extreme, to the extent that the distribution of resources or risks in one's environment is stable across intergenerational timescales, there might emerge a culturally ingrained tendency towards working with either simpler or more complex models—one that is not tailored to particular contexts, but rather applies across the diverse contexts encountered in a lifespan.

Other possibilities fall between these extremes. Even if we do not continuously monitor the shape of payoffs from moment to moment, we nevertheless approach different broad classes of problems with prediction strategies tailored to their typical objective functions. For problems with salient left-tail risks, such as death or extreme harm, people may adopt a more "high-bias" mindset than the "high-variance" mindset that they would adopt for problems with salient right-tail benefits, such as great wealth or fame. These general strategies could be learned individually, or inherited biologically or culturally.

Application to statistics and machine learning

Apart from its psychological applications, the current work may help inform the ways in which researchers select statistical models based on their complexity. Of course, in many contexts, there is little cost to optimizing out-of-sample prediction using techniques like cross-validation. But when this is not possible, it is important to keep in mind that relatively more complex models may be most useful when the researcher's goal is precision, whereas relatively simpler models may be most useful when the researcher's goal is to make consistent, safe predictions. Broadly speaking, if a researcher is focused on exploring new patterns in the data that are worth following up on, large mistaken inferences should be relatively harmless, and thus a more complex model may be advised. On the other hand, it may be much riskier to employ such a model when large errors are very costly. Indeed, we hope to apply these ideas to medical contexts in which large errors can result in serious injury or even death.

Finally, it is worth noting that the present work is limited in its focus on Gaussian prediction error and neglect of other irreducible factors besides bias and variance that contribute to loss. Moreover, the simple approach we pursue here is quite abstract and may not generalize to all applications of the bias-variance dilemma. Future work should explore how much the loss function impacts specific modeling choices, such as the penalization terms in ridge or Lasso regression or the depth of decision trees. Nevertheless, we hope that the analysis presented here calls attention to an important relationship between how one makes predictions and how close to the truth one needs them to be.

Acknowledgements

This research was supported by grant N00014-19-1-2025 from the Office of Naval Research.

References

Brainard, D. H., & Freeman, W. T. (1997). Bayesian color constancy. *JOSA A*, *14*(7), 1393–1411.

Brighton, H., & Gigerenzer, G. (2015). The bias bias. *Journal of Business Research*, *68*(8), 1772–1784.

Daw, N. D., Niv, Y., & Dayan, P. (2005). Uncertainty-based competition between prefrontal and dorsolateral striatal systems for behavioral control. *Nature Neuroscience*, *8*(12), 1704–1711.

Deutsch, M., & Gerard, H. B. (1955). A study of normative and informational social influences upon individual judgment. *The Journal of Abnormal and Social Psychology*, *51*(3), 629.

Dolan, R. J., & Dayan, P. (2013). Goals and habits in the brain. *Neuron*, *80*(2), 312–325.

Evans, J. S. B. (2008). Dual-processing accounts of reasoning, judgment, and social cognition. *Annual Review of Psychology*, *59*, 255–278.

Friedman, J. H. (1997). On bias, variance, 0/1—loss, and the curse-of-dimensionality. *Data Mining and Knowledge Discovery*, *1*(1), 55–77.

Gaissmaier, W., Wilke, A., Scheibehenne, B., McCanney, P., & Barrett, H. C. (2016). Betting on illusory patterns: Probability matching in habitual gamblers. *Journal of Gambling Studies*, *32*(1), 143–156.

Geman, S., Bienenstock, E., & Doursat, R. (1992). Neural networks and the bias/variance dilemma. *Neural Computation*, *4*(1), 1–58.

Gigerenzer, G. (2007). *Gut feelings: The intelligence of the unconscious*. Penguin.

Gigerenzer, G., & Todd, P. M. (1999). *Simple heuristics that make us smart*. Oxford University Press.

James, G. (2003). Variance and bias for general loss functions. *Machine Learning*, *51*(2), 115–135.

James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013). *An introduction to statistical learning* (Vol. 112). Springer.

Kahneman, D. (2011). *Thinking, fast and slow*. Macmillan.

Keramati, M., Dezfouli, A., & Piray, P. (2011). Speed/accuracy trade-off between the habitual and the goal-directed processes. *PLoS Computational Biology*, *7*(5).

Kool, W., Gershman, S. J., & Cushman, F. A. (2017). Cost-benefit arbitration between multiple reinforcement-learning systems. *Psychological Science*, *28*(9), 1321–1333.

Lee, S. W., Shimojo, S., & O'Doherty, J. P. (2014). Neural computations underlying arbitration between model-based and model-free learning. *Neuron*, *81*(3), 687–699.

Richerson, P. J., & Boyd, R. (2008). *Not by genes alone: How culture transformed human evolution*. University of Chicago Press.

Shenhav, A., Botvinick, M. M., & Cohen, J. D. (2013). The expected value of control: An integrative theory of anterior cingulate cortex function. *Neuron*, *79*(2), 217–240.

Sloman, S. A. (1996). The empirical case for two systems of reasoning. *Psychological Bulletin*, *119*(1), 3–22.

Stewart, N., Chater, N., & Brown, G. D. (2006). Decision by sampling. *Cognitive Psychology*, *53*(1), 1–26.

Vul, E., Goodman, N., Griffiths, T. L., & Tenenbaum, J. B. (2014). One and done? optimal decisions from very few samples. *Cognitive Science*, *38*(4), 599–637.