Sins of Omission and Commission in Complex Systems[1][2]

Drew Fudenberg[3] and David K. Levine[4]

Abstract

We study how optimal interventions in response to a shock with limited information depend on the complexity of the system. We show that as the complexity of the system grows, the optimal intervention shrinks to zero.

[1]We are grateful to Chris Ackerman, Daniel Clark, Daniel Friedman, George Georgiadis, Giacomo Lanzani, Andy McLennan, John Rust and Larry Samuelson for helpful comments, and to NSF grant SES 1643517 and MIUR (PRIN 2017H5KPLL) for financial support.

[2]This version: 8/22/2020. First version: 08/22/2020

[3]Department of Economics, MIT; drew.fudenberg@gmail.com

[4]Department of Economics and RSCAS, EUI, and Department of Economics, WUSTL; david@dklevine.com

1. Introduction

We consider a decision maker faced with limited information who has the opportunity to intervene in a system that has received a shock. For example, while driving a car, you hear a large bang, the car no longer accelerates properly, and the engine makes loud noises. Should you get out of the car and try to fix it, or simply struggle on? You are not an automobile mechanic, and do not know how engines work, but you are aware of basic facts, for example, that breaking parts of the engine with a hammer will make things worse.

Similar settings abound, such as whether to give aid to an injured person. In settings of aiding others there is a substantial literature arguing that non-intervention is generally thought to be better.[5] However, we expect that that someone's perceived obligation to intervene depends on the amount of information they have relative to the difficulty of the problem. For example, in a health emergency on an airplane a doctor would be expected to intervene when an ordinary passenger would not. By contrast if an elderly person drops a bag of groceries even an ordinary passer-by might be expected to help.

Similar issues arise with respect to economic systems: How much should policy makers intervene in in response to a pandemic? How should rich countries intervene to help a developing country facing a crisis?[6] How should business firms faced with unexpected systems failures or by competition from new products respond?

In static settings economics does not always clearly distinguish between omission and commission, but in dynamic settings such as Stokey (2009) there often is a clear idea of inaction. Here we assume there is a known status quo, so there is a clear distinction between a sin of omission, that is either not intervening or intervening too little, and a sin of comission, intervening but making things worse. Here we argue that utility maximization supports the common intuition that sins of commission are to be avoided in complex systems with limited information.

To make this point we adopt a very simple framework. We study a symmetric quadratic loss function in n-dimensional Euclidean space where the loss is measured by the Euclidean distance from the optimum. Initially the optimum is know to be at zero, but a shock then displaces the optimum. The decision maker knows how large the shock is, that is, how great the loss is, but does not know the direction in which the optimum has moved. The decision maker is restricted to responding in a single dimension, which we interpret as the result of partial ignorance.. The decision maker does, know in which direction the objective function is increasing, - which is akin to knowing not to break parts of the engine with a hammer.

In this setting complexity is measured by the dimension of the system n.

[5] Spranca, Minsk and Baron (1991) and Cushman, Young and Hauser (2006), for example, provide experimental evidence that this is a widely held view and review other evidence and literature.

[6] Easterly (2002) discusses some of the ways such interventions can backfire.

We show that in a one dimensional system the full optimum is obtained, and that as the dimension increases optimal interventions become smaller and the optimized loss grows. The intuition is that with more directions it is more likely that an intervention in a randomly chosen direction will "overshoot" and lead to an additional loss. In the limit as $n \to \infty$ it is optimal not to intervene at all and simply accept the loss. In addition if, following Stokey (2009)there is a small fixed cost of intervention, then small interventions are not worthwhile and no intervention is better. In this sense the more complex is the system, the bigger the shock must be before it is worthwhile to intervene.

The problem of intervention with limited information as we formulate it appears to be new. It is connected to Samuelson (1947)'s LeChatelier principle, which shows that in a setting of certainty, where the location of the optimum is known, the size of intervention is strictly smaller when intervention is possible in a subet of the dimensions instead of all of them. We extend this to the case where the location of the optimum is unknown, and show that this effect increases as the complexity of the system does. Our setup has also some similarities to Ely (2011), who also uses the number of control dimensions as a measure complexity.. Our results are also, in a certain sense, a counterpoint to the "curse of dimensionality" (see, for example, Bellman (1957)) that states that the time it takes to reach the optimum increases exponentially with the dimension of the problem. Here we show that incomplete optimization also does less well as the dimension of the problem increases. We relate this to results about the rate of convergence of gradient descent algorithms.

2. The Model

The state of the system is x in $\Re^n$ where $n \geq 1$ and has the interpretation as the complexity of the system. There is a loss minimizing action $\hat{x}$ and the loss function is quadratic and given by $(1/2)\sum_{i=1}^{n}(x_i - \hat{x}_i)^2$. This function is known. Initially there is a status quo, denoted by x_0, at which the current intervention and the loss minimizing action are both 0, that is, $x_0 = \hat{x}_0 = 0$. A shock then occurs and $\hat{x}$ is chosen randomly according to a constant density on the surface of the sphere of radius $|\hat{x}|$. The decision maker does not observe the new value of $\hat{x}$ but observes size of the loss $L = (1/2)\sum_{i=1}^{n}\hat{x}_i{}^2$ at the status quo, so knows the Euclidean distance $|\hat{x}|$.[7]

The decision maker is restricted to moving in one dimension, which we take to be the x_1 axis. Let d denote the column vector $d = |\hat{x}|(1, 0, 0, \ldots, 0)^T$. The decision maker observes the sign of the directional derivative of L in direction d at the status quo, which is equal to $-\text{sgn}\hat{x}_1$; we denote it σ to lighten notation.[8]

[7] In this formulation the utility from the optimum after the shock is the same as before, which is to say zero. This is a convenient normalization that does not matter for the analysis: if the optimum after the shock is much worse, or better even than before our analysis is unchanged.

[8] We subsequently consider the case where the decision maker observes the magnitude as well as the sign.

Hence, the decision maker chooses a real number $\beta(\sigma)$ and sets $x = \beta(\sigma)d$.

One way to think about one-dimensional adjustment is to imagine the system as a mechanical machine controlled by dials, with each dial corresponding to a coordinate axis. Being restricted to a single dimension corresponds to partial ignorance, in the sense of being aware of only one dial. A decision maker who knew the system better might be aware of more dials/dimensions. For example, in trying to help a person with a severe leg wound a typical person might think that putting pressure on the wound is a good idea, while a doctor might know also that it is possible to use a tourniquet above the wound (a second dial).

The assumption that intervention is only contemplated in a single dimension is a good approximation in many settings with partial ignorance. For example, in the EU debate over pandemic aid the debate focused on a single dimension: the division of aid between grants and loans. A less well known example is that of the Wright brothers. An airplane is a complex system, and marketing one is complicated as well. Faced with competition from Glenn Curtiss airplanes built using the superior aileron technology, the Wrights did not consider redesigning their airplane or change their marketing practices: the only dimension in which they responded was in the filing of patent lawsuits.[9]

It should be clear as well that some tinkering in other dimensions does take place. For example, in the EU pandemic package, the pro-grant countries agreed to decrease the total size of the aid package and increase the rebates of the "Frugal Four." These other dimensions typically are limited to small changes. Subsequently we consider the robustness of our results to the possibility that the decision maker might have some idea of a slightly better direction than d.

The goal of the decision maker is to choose $\beta(\sigma)$ to minimize expected loss. If the choice is $\beta(\sigma)$ the loss is $(1/2)|\hat{x} - \beta(\sigma)d|^2$so that the overall expected loss is this amount integrated with respect to $\hat{x}$ according to a uniform distribution over the n-dimensional sphere.

Preliminary Analysis

The problem is homogeneous, so the optimal solution $\beta^n(\sigma)$ is independent of $|\hat{x}|$, and if the minimum expected loss corresponding to $|\hat{x}| = 1$ is L^n, the corresponding expected loss for general $|\hat{x}|$ is $L^n|\hat{x}|^2$. With this in mind, we normalize $|\hat{x}| = 1$. It is convenient moreover to use coordinates in which $\hat{x} = (1, 0, 0, \ldots, 0)^T$, so that d rather than $\hat{x}$ is random and uniformly distributed over the unit sphere. As the sign of the derivative is observed, the decision maker should move in the direction $r = -\sigma d$ to reduce the loss. Moreover from symmetry the optimal choice for the "bad" hemisphere is opposite that for the "good" hemisphere, that is, for the optimum $\beta(-1) = -\beta(1) \geq 0$. Hence the original problem is equivalent to the decision maker choosing non-negative $\phi \geq 0$ and $x = \phi r$. We use ϕ_n to denote the optimal value of ϕ. Since d is uniform on the sphere and the sign of the derivative reflects the "bad" half-sphere to the

[9] A good account of the sad saga of the Wright brothers can be found in Shulman (2002).

"good" one, r is uniform on the half-sphere defined by $|r| = 1$ and $r_1 \geq 0$. We subsequently study this normalized problem.

In the one dimensional case, $n = 1$, the true optimum lies at $x_1 = 1$ while $r_1 = 1$ so that the optimal choice of ϕ is $\phi = 1$ with corresponding loss $L^1 = 0$. The key point is that in higher dimensions there is a tradeoff. If $r = \hat{x}$ it would be best to choose $\phi = 1$. If r is orthogonal to $\hat{x}$ it would be best to choose $\phi = 0$. Hence the optimal choice of ϕ is a compromise: a large ϕ works well for "good" directions r close to $\hat{x}$ but poorly in "bad" directions far from $\hat{x}$. The intuition we would like to establish is that in higher dimensions there are relatively more "bad" directions so ϕ should be chosen smaller.

3. The Main Result

Theorem 1. *The optimal solution ϕ^n and corresponding loss L^n satisfies $\phi^{n+1} < \phi^n$, $L^{n+1} > L^n$ with $\phi_1 = 1, L^1 = 0$ and $\lim_{n\to\infty} \phi^n = 0$, $\lim_{n\to\infty} L^n = 1/2$.*

This says is that given a shock that without intervention would result in a half unit loss, if the system has high complexity as measured by n it is better to intervene very little with the result that almost the entire loss must be swallowed, while if the system has low complexity a substantial intervention is optimal resulting in a substantial mitigation of the loss.

The case $n = 1$ was proven above. The remainder of the proof follows from several intermediate results which we now prove.

Lemma 1. *The loss function is $(1/2)(1 - 2a^n\phi + \phi^2)$ with $0 < a^n \leq 1$ and $a^1 = 1$.*

Proof. Fix r. Using the fact that $|r| = 1$ we can compute The the loss is $L^n(r, \phi) = (1/2)[(1 - \phi r_1)^2 + \sum_{i=2}^n (\phi r_i)^2] = (1/2)[1 - 2r_1\phi + \sum_{i=1}^n (\phi r_i)^2] = 1/2 - r_1\phi + \phi^2/2$. Hence a^n is the integral of r_1 with respect to r which is distributed uniformly over the unit half-sphere $|r| = 1$ and $r_1 \geq 0$. As r_1 is strictly positive with probability 1 and $r_1 \leq 1$ on the unit half-sphere it follows that $0 < a^n \leq 1$, and the result for $n = 1$ is immediate. □

Corollary 1. *$\phi^n = a^n$ and $L^n = 1/2 - (a^n)^2/2$.*

The theorem will follow if we can show that$a^{n+1} < a^n$ and $\lim_{n\to\infty} a^n = 0$. This we show next.

Lemma 2. *Let $h(n, \theta)$ be the density function*

$$h(n, \theta) = \frac{(\sin\theta)^{n-2}}{\int_0^{\pi/2} (\sin\omega)^{n-2} d\omega}$$

on $[0, \pi/2]$. Then for $n > 1$ we have $a^n = \int_0^{\pi/2} h(n, \theta) \cos\theta d\theta$.

Proof. For $n > 1$ we do the integration by taking $r_1 = \cos\theta$ where $\theta \in [0, \pi/2]$. For given θ the density is given by the surface area S^{n-2} of the $n-2$ dimensional sphere with radius equal to r_1, which is $S^{n-2} = c(n-2)(\sin\theta)^{n-2}$ where $c(n-2)$ is positive, known, and irrelevant. Hence we can compute

$$a^n = \frac{\int_0^{\pi/2} c(n-2)(\sin\theta)^{n-2}\cos\theta d\theta}{\int_0^{\pi/2} c(n-2)(\sin\theta)^{n-2} d\theta}.$$

□

The final step of proving the theorem is then

Lemma 3. *$a^{n+1} < a^n$ and $\lim_{n\to\infty} a^n = 0$.*

Proof. To show $a^{n+1} < a^n$ note that since $\cos\theta$ is strictly decreasing[10] in θ it suffices to prove that $h(n+1,\theta)$ first order stochastically dominates $h(n,\theta)$. Observe that if $\theta' > \theta$ since $\sin\theta$ is strictly increasing we have

$$\frac{h(n+1,\theta')}{h(n+1,\theta)} = \left(\frac{\sin\theta'}{\sin\theta}\right)^{n-1} > \left(\frac{\sin\theta'}{\sin\theta}\right)^{n-2} = \frac{h(n,\theta')}{h(n,\theta)}.$$

Hence as both are density functions it must be that $h(n+1,0) < h(n,0)$ and there is a unique value $\theta \in [0,\pi/2]$ where $h(n+1,\theta) = h(n,\theta)$ which implies stochastic dominance.

To show $\lim_{n\to\infty} a^n = 0$, let $\varepsilon > 0$, and observe that there is $\theta_\varepsilon \in [0,\pi/2]$ such that for all $\theta \in [\theta_\varepsilon, \pi/2,]$, $\cos(\theta) < \varepsilon$. Moreover, since $\sin\theta$ is strictly increasing on $[0,\pi/2]$ for any $M \in (\theta_\varepsilon, \pi/2)$ we have

$$\lim_{n\to\infty} \frac{\int_0^{\theta_\varepsilon} c(n-2)(\sin\theta)^{n-2}d\theta}{\int_0^{\pi/2} c(n-2)(\sin\theta)^{n-2}d\theta} \le \lim_{n\to\infty} \frac{\theta_\varepsilon(\sin\theta_\varepsilon)^{n-2}}{\int_M^{\pi/2}(\sin\theta)^{n-2}d\theta} \le \lim{}_{n\to\infty} \frac{\theta_\varepsilon(\sin\theta_\varepsilon)^{n-2}}{(\pi/2-M)(\sin M)^{n-2}} = 0$$

Since $\cos\theta \le 1$

$$\lim_{n\to\infty} a_n \le \lim_{n\to'\infty} \frac{\int_0^{\theta_\varepsilon} c(n-2)(\sin\theta)^{n-2}d\theta}{\int_0^{\pi/2} c(n-2)(\sin\theta)^{n-2}d\theta} + \frac{\int_{\theta_\epsilon}^{\pi/2} c(n-2((\sin\theta)^{n-2}\varepsilon d\theta}{\int_0^{\pi/2} c(n-2)(\sin\theta)^{n-2}d\theta} \le \varepsilon.$$

Since ϵ was arbitrary it follows that in fact $\lim_{n\to\infty} a^n = 0$. □

More Directions

As we indicated earlier, the decision maker might have some ability to slightly improve the direction r . For example, the decision maker might be limited not just to the coordinate axis, but to directions that lies within an angle $\overline{\theta}$ of the

[10] Note that the only facts we use about $\cos\theta$ is that on $[0,\pi/2]$ it is strictly decreasing from one to zero.

coordinate axis, where $\overline{\theta}$ is relatively small. This corresponds to directions $\tilde{r}$ lying on the unit sphere with $|\tilde{r}'r| \geq \cos\overline{\theta}$. The most information that the decision-maker could have is to know the best direction within that cone. As long as $\overline{\theta} < \pi/2$ this does not effect the monotonicity result which did not depend on the range over which the integrals are taken, but rather on the fact that worse directions are more likely in higher dimensions.

Knowing the best direction in a cone does change the asymptotic results: with $\overline{\theta} = 0$ as in our base model, the optimal $\phi(r)$ on the boundary where $r'\hat{x} = 0$ is $\phi(r) = 0$ with corresponding loss $L(r) = 1$. With $\overline{\theta}>0$ weight is still pushed towards the boundary as n increases, but now when r is orthogonal to $\hat{x}$ the direction of adjustment $\tilde{r}$ now has an angle $\pi/2 - \overline{\theta}/2$ lying closer to the optimum . This implies a boundary optimum of $\phi(\tilde{r}) = \cos(\pi/2 - \overline{\theta}/2)$ with corresponding boundary loss of $1/2 - \cos(\pi/2 - \overline{\theta}/2)^2/2$. In practice, the decision maker is unlikely to know the best direction in the cone, and indeed our results argue it will be more difficult in higher dimensions. Hence these should be viewed as upper bounds on the size of the optimal intervention ϕ and a lower bound on the loss L. The key point is that even for these bounds, if $\overline{\theta}$ is small the asymptotic intervention $\phi(\tilde{r}) = \cos(\pi/2 - \overline{\theta}/2)$ is small, and the loss $1/2 - \cos(\pi/2 - \overline{\theta}/2)^2/2$ is close to $1/2$.

4. Quantitative Information and Gradient Descent

In the base model only qualitative information about the direction of improvement, the sign of the derivative, is available. We now consider the implications of quantitative information and ask what happens when the decision maker observes not only the sign of the derivative but also its absolute value, which is r_1. Now the optimal choice of ϕ may depend upon r_1; we denote it by $\varphi^n(r_1)$ with corresponding expected loss $\lambda_n(r_1)$. Also denote by E^n the expectation with respect to the direction d. If we consider the expected intervention and corresponding expected loss the picture is the same as in the base model.

Theorem 2. *The optimal solution φ^n and corresponding loss λ^n satisfies $E^n\varphi^n = \phi^n$, $L^{n+1} > E^{n+1}\lambda^{n+1} > E^n\lambda^n$ with $\lambda^1 = 0$ and $\lim_{n\to\infty} E^n\lambda^n = 1/2$.*

Proof. From the proof of Lemma 1 the objective function is $1/2 - r_1\phi + \phi^2/2$. Hence $\varphi^n(r_1) = r_1$. It follows that $E^n\varphi^n = \phi^n = E^n r_1$, which is clear since the best response is linear, so it does not matter if we first compute the optimum then take the expectation as here, or first take the expectation and then compute the optimum as in the base model.

Substituting back into the objective function we have $\lambda^n(r_1) = 1/2 - r_1^2/2$ so that $E^n\lambda^n = 1/2 - Er_1^2/2$. Since $L^n = 1/2 - (Er_1)^2/2$ the result $L^{n+1} > E^{n+1}\lambda^{n+1}$ follows from Jensen's inequality. The remaining results follow from Lemma 3 which as observed in footnote 10 holds not only for computing $E^n r_1 = E^n \cos\theta$ but also for computing $E^n r_1^2 = E^n(\cos\theta)^2$. □

The model here with the magnitude of the derivative known is equivalent to a model in which $|\hat{x}|$ (and the length of d) is unknown, but it is possible to do

a line search in the direction r to find the optimum, which will be at $(r_1/|\hat{x}|)\, r$. This provides a connection to the method of gradient descent. To make this connection it is helpful to translate the coordinate system by subtracting the optimum. In these coordinates the optimum $\hat{x}$ is at the origin and the status quo $x_0 = -(1, 0, 0, \ldots, 0)^T$.

We can index a problem of gradient descent by choosing a non-singular symmetric matrix A and a starting point y_0 with the objective function $(1/2)y'A'Ay$. The first step of gradient descent computes the gradient $g = Ay_0$ at y_0, and then conducts a line search to find the optimum y_1 in along that line. It then repeats starting at y_1.

The improvement made at the first step of the gradient descent process can be analyzed using our methods. Let B be a nonsingular linear transformation such that $By_0 = x_0$ and $(AB^{-1})^T AB^{-1} = I$. In the coordinate system $x = By$ the loss function is $(1/2)x^T x$ and the initial condition is x_0. The direction of search is the tranformed gradient $r = -Bg$. Random choice of gradient descent problems, that is (A, y_0, B) then map to random choices of r in our environment. This enables us to translate our results into conclusions about the first step of randomly chosen gradient descent problems. In particular, in the Appendix we show that if the probability distribution over gradient descent problems satisfies a symmetry and monotonicity condition then as the dimension of the problem increases the step size and gain from the first steps grow smaller and in the limit approach zero.

This result about dimensionality and the first step has the same flavor as rate of convergence bound results. There is a standard convergence bound (see, for example, Meza (2010)) for the method of gradient descent. If the matrix A is randomly chosen by independently choosing the diagonal and upper triangle elements from suitable distribution and filling in the rest by symmetry, a result on the distribution of eigenvalues by Wigner (1958) implies, as explained in the Appendix, that the convergence bound deteriorates as the dimension increases.

Finally, we observe that when the decision maker observes the magnitude of the derivative, we can model decision makers who are less ignorant by supposing they can adjust the first m coordinates rather than just the first. We conjecture that similar results hold provided that m/n goes to 0.

5. Conclusion

We have have defined and studied the "optimal intervention problem with limited information." We modeled the complexity of a problem by its dimensionality, and found that as dimension increases it is optimal to make smaller interventions. Our proof follows the intuition that with more directions in which to move there is greater opportunity for mistake, and hence greater need for caution.

We consider two notions of limited information, both corresponding to what the decision maker knows about moving in any arbitrarily-chosen direction. In one formulation, only the sign of the derivative is known, and in the other its

magnitude is known as well. In both cases we assume that the magnitude of the loss from non-intervention is known. In the "magnitude" case if line search is possible we show that this does not matter. More generally, while it may not be the case in practice that the magnitude of the loss is known, in more complex systems this information is likely to be more difficult to come by, which reinforces our results.

It is important to note that our result depends on the decision maker only observing the magnitude of the derivative in a fixed and small number of dimensions. In particular, if the decision maker observes the magnitude of the derivative in as few as n linearly independent directions then the full optimum is achieved.[11]

There is another point worth emphasizing. If $n > 1$ since $\phi^n > 0$ it follows that while the expected loss is less than in the absence of intervention, there is still a positive probability that intervention will make the realized loss worse: Sometimes the intervention will make things worse rather than better.[12]

What do our theoretical results tell us? One case in point is that firms who do not respond to competition from complicated new products are sometimes criticized for doing nothing. However, if a firm lacks the technical expertise to locate the new "optimum," and are limited to adjustments in a single dimension such as price, our results show that it may indeed be optimal to respond only slightly if at all. The responses of WordPerfect to the Windows 3.0 shock in 1990 and of Nokia and Blackberry to the iPhone shock in 2007 all fit into this category. In all three cases, firms did little to respond to the shock, and subsequent events showed that all indeed lacked the technical expertise to build competitive products.

One way to read our theory is that even in time of crisis large interventions are a bad idea. This is not the case. Instead, our model says that large ill-considered interventions are a bad idea. One example is that of the failure of NASDAQ computer systems during the Facebook IPO. Rather than study the system to understand the reason for the failure, programmers were instructed simply to make a large intervention in a single direction, namely to remove a validation check that had caused the system to shut down. The consequences were catastrophic: There was a cascading series of failures and "traders blamed Nasdaq for hundreds of millions of dollars of losses, and the mistake exposed the exchange to litigation, fines, and reputational costs."[13]

[11] This follows from the fact that the problem is quadratic, that the second derivatives are known, and that the distance to the optimum is known.

[12] This can be seen from the fact that if d is orthogonal to $\hat{x}$ the fact that $\phi_n > 0$ implies a strictly greater loss than in the absence of intervention, hence the same must be true for d that are nearly orthogonal to $\hat{x}$.

[13] The story and the quote are from Clearfield and Tilcsik (2018), who also give examples where similar catastrophes were averted because complex malfunctioning systems were not restarted.

References

Bellman, R. E., 1957. *Dynamic programming.* Princeton University Press, p.151.

Clearfield, C. and Tilcsik A. 2018. How to prepare for a crisis you couldn't predict. *Harvard Business Review,* March 16.

Cushman, F., Young, L. and Hauser, M., 2006. The role of conscious reasoning and intuition in moral judgment: Testing three principles of harm. *Psychological science*, 17(12), pp.1082-1089.

Ely, J.C., 2011. Kludged. *American Economic Journal: Microeconomics*, 3(3), pp.210-31.

Easterly, W., 2002. *The elusive quest for growth: economists' adventures and misadventures in the tropics.* MIT press.

Meza, J.C., 2010. Steepest descent. *Wiley Interdisciplinary Reviews: Computational Statistics*, 2(6), pp.719-722.

Samuelson, P.A., 1947. *Foundations of economic analysis.* Harvard University Press.

Shulman, S., 2002. *Unlocking the sky.* Harper-Collins.

Spranca, M., Minsk, E. and Baron, J., 1991. Omission and commission in judgment and choice. *Journal of experimental social psychology*, 27(1), pp.76-105.

Stokey, N. L. 2009. *The Economics of Inaction: Stochastic Control models with fixed costs.* Princeton University Press, 2009.

Wigner, E., 1958. On the Distribution of the Roots of Certain Symmetric Matrices. *Annals of Mathematics* 67, 325-328.

Appendix: Gradient Descent

First Step of Gradient Descent

Recall that we have translated the coordinate system by subtracting the location of the optimum, so that the optimum $\hat{x}$ is at the origin and the status quo $x_0 = -(1, 0, 0, \ldots, 0)^T$. We describe a gradient descent problem as a non-singular symmetric matrix A and a starting point y_0 with the objective function $(1/2)y^T A^T Ay$. We map this into our setting with a nonsingular linear transformation B such that $By_0 = x_0$ and $(AB^{-1})^T AB^{-1} = I$. The direction of gradient descent, that is the negative of the gradient $-g = Ay_0$ maps to $r = BAy_0$.

Observe that

$$-x_0^T r = -x_0^T BAy_0 = -x_0^T BB'x_0 < 0$$

so that r is a direction of decrease for the objective function and lies in the half-sphere defined by x_0 and $\hat{x}$. If we find $\phi_n(\mu)$ and $\lambda_n(\mu)$ in our problem then the optimum in the original problem is given by $\phi_n(\mu)g$ and the loss relative to $|y_0|^2$ is given by $\lambda_n(\mu)$.

Any distribution over (A, y_0, B) induces a distribution of r over our half-sphere. If the induced distribution over r is uniform describe the distribution over (A, y_0, B) *as pseudo-uniform.*Our main theorem then says that in this case as the dimension increases the expected size of the first step of gradient descent and expected gain from that step shrink to zero.

There are in general many pseudo-uniform choices of gradient descent problems. Here is one example: let e be uniform on our half-sphere and $A = \sqrt{2}\left((2/e_1)\left((e-(e_1\hat{x})/2)(e-(e_1\hat{x})/2)^T + (e_1^2/4)I\right)\right)^{1/2}$, $y_0 = A^{-1}x_0$, and $B = A$. Then as $\hat{x} = (1, 0, 0, \ldots, 0)^T$

$$r = 2A^{-1}(A^T)^{-1}\hat{x}$$

$$= \left((2/e_1)\left((e-(e_1\hat{x})/2)(e-(e_1\hat{x})/2)^T + (e_1^2/4)I\right)\right)\hat{x}$$

$$= (2/e_1)\left((e-(e_1\hat{x})/2)(e-(e_1\hat{x})/2)^T\hat{x} + (e_1^2/4)\hat{x}\right)$$

$$= (2/e_1)\left((e-(e_1\hat{x})/2)(e_1 - e_1/2) + (e_1^2/4)\hat{x}\right)$$

$$= (2/e_1)\left((e_1/2)e - (e_1^2\hat{x})/4) + (e_1^2/4)\hat{x}\right) = e$$

is also uniform on our half-sphere so this distribution is indeed pseudo uniform. Note that since $B = A$ we can easily extend this to a measure with full support over A. Take $y_0 = A^{-1}x_0$. The mapping $r = BAy_0 = A^{-1}x_0$ induces equivalence classes of matrices A with two matrices being equivalent the give rise the the same value of r . Regardless of the distribution over matrices conditional on equivalent class, the example given provides a probability distribution over equivalence classes such that we get pseudo uniformity.

Note pseudo uniformity can be relaxed substantially. In proving Theorem 1 we conditioned on the angle θ. For given θ we used the fact that the distribution of directions r was symmetric. Let us say that a distribution over (A, y_0, B) is *symmetric* if this is the case. We view this as a fairly neutral assumption, that the gradient descent problem does not favor any particular direction. Second we used the fact that the distribution over angles $g_n(\theta)$ is uniform on $[0, \pi/2]$. However: the proof only requires that it be strictly positive and independent of n. Moreover, it is clear that if as we increase n the distribution for $n+1$ weakly stochastically dominates that for n then this enhances the results given. If this is the case we say that the ensemble of distributions over (A, y_0, B) is *weakly monotone.* We conclude that symmetry plus weak monotonicity is sufficent for our results.

Wigner's Theorem and the Condition Number

A standard convergence bound (see, for example, Meza (2010)) shows that the rate of convergence declines as the ratio of the largest to smallest eigenvalue

of $A'A$ grows. If the matrix A is randomly chosen by independently choosing the diagonal and upper triangle elements from a fixed and standardized distribution with moments of all orders, and filling in the rest by symmetry Wigner (1958)'s semi-circle law says that the fraction of normalized eigenvalues $\lambda_j/n^{1/2}$ of A that lie in an interval $[\underline{\lambda}, \overline{\lambda}] \subseteq [-1, 1]$ converges in probability to

$$(2/\pi) \int_{\underline{\lambda}}^{\overline{\lambda}} \sqrt{(1-\lambda^2)} d\lambda.$$

Graphically on $[-1, 1]$ the function $\sqrt{(1-\lambda^2)}$ is a semi-circle, hence the name. The implication for the ratio of the largest to smallest eigenvalue of AA' is clear: for any ϵ with high probability for large enough n there must be eigenvalues (and indeed quite a few of them) in $[0, 2\epsilon]$ and in $[1/2, 1/2 + 2\epsilon]$ so that the ratio is at least $1/\epsilon$. Hence, asymptotically, the convergence bound has no bite.