High-Dimensional Statistics

Or How I Learned to Stop Worrying and Love Geometry

Collin Cademartori

May 2, 2020

What Makes a Problem High Dimensional?

- Suppose we make n measurements of p variables
- Ex: We ask $n = 200$ poll respondants $p = 5$ questions
- Classical low-dimensional asymptotics: p fixed, $n \to \infty$
- Like getting a larger sample size for the same poll
- High-dimensional asymptotics: $p/n \to C > 0$
- I.e. the number of variables grows with the number of observations
- Asymptotics helps us understand if our estimators converge
- "Do we get the right answer with infinite data"
- In the finite data setting, high-dimensionality means $p > n$
- Or even just n not much larger than p!

High Dimensional Problems: Complex Data

- Why is this an interesting question?
- Sometimes our data is naturally high-dimensional
- Human genome data - 3 billion base pairs!
- Stress testing financial institutions - many macro variables
- Choosing interactions, n predictors, $2^n - n$ interactions!

High Dimensional Problems: Complex Models

- Sometimes we design models to have more parameters than data
- Recall the Bayesian hierarchical model for 8 schools
- For each school, we get one test score measurement
- Each school has a separate (long-run) average score parameter
- We then assume these come from a common population
- ...which itself has some mean and variance
- So that's 10 parameters and 8 data points
- Problem persists even if we add more schools

You Take the High Road...

- What makes high dimensional problems more difficult?
- Large-sample limits are more complex (asymptotic theory)
- Optimization problems may lack unique solutions (frequentist inference)
- Sampling in many dimensions runs into curse of dimensionality (Bayesian inference)
- We will focus on the latter two issues

A Brief Digression on Identifiability

- You may have heard “correlation does not imply causation”
- Central problem of science: many competing explanations for phenomena
- Need to design experiments that can distinguish explanations
- A similar thing happens in frequentist inference
- The phenomena are the observed data
- Explanations are values for parameters along with a model
- Model identifiable if we can recover true parameters from data

More Formally...

- We can make that idea more precise
- A statistical model is just a probability distribution
- In frequentist stats, a sampling distribution:

$$p(x_1, \ldots, x_n \mid \theta_1, \ldots, \theta_p)$$

- Recall maximum likelihood inference:

$$\hat{\theta} = \arg\max_{\theta} p(x_1, \ldots, x_n \mid \theta)$$

- What if

$$p\left(x \mid \theta^1\right) = \max_{\theta} p(x \mid \theta) = p\left(x \mid \theta^2\right)$$

- There is no unique maximum likelihood estimate!

Identifiability in High Dimensions

- A model is unidentifiable if data can't distinguish between parameters
- This phenomenon is common in high dimensions
- Why? Let's think about a deterministic example
- Estimate the coefficients of a degree k monic polynomial
- Need to infer the k coefficients
- I give you $k-1$ zeros z_i of the polynomial
- Define
$$q(x) = (x - z_1)(x - z_2) \cdots (x - z_{k-1})$$
- This is a degree $k-1$ polynomial with these roots
- $(x-z)q(x)$ is degree k and has all z_i as roots for any z
- Infinitely many solutions to the problem!

Unidentified = Unemployed?

- When there are multiple solutions, often infinitely many
- In some sense our sample size is just too small
- Imagine trying to infer 10 coefficients with just 3 samples
- We clearly would need more data!
- So why should we expect to be able to solve high-dimensional problems?
- In fact, these problems are not solvable in general

Sparsity - Hide and Seek in High Dimensions

- But many cases that arise in practice are solvable!
- Imagine regressing disease variables on all genome base pairs
- We often expect only a reltively small number of pairs are relavant
- But we may have very little idea which are important
- If we could identify them in advance, we could do low-dimensional regression
- We can rephrase that idea in terms of coefficients
- We expect most of the (true) coefficients in our high-dimensional model to be 0
- This kind of assumption is called a sparsity assumption

How to Find Sparse Solutions

- Let θ be the (p-dimensional) vector of parameters
- Define $\|\theta\|_0$ to be the number of nonzero elements
- Suppose c is an upper bound on the # of nonzero elements
- Then we can replace the maximum likelihood problem with a sparse version:

$$\arg\max_{\theta} p(x \mid \theta)$$
$$\text{such that } \|\theta\|_0 \leq c$$

- This problem is computationally too hard! There are

$$\binom{p}{p-c} = \frac{p!}{(p-c)!c!}$$

ways to choose $p - c$ coefficients to be 0, essentially have to search them all

How to Find Sparse Solutions and be Smart About It

- Note the previous optimization problem did solve the identifiability problem
- But it was too hard to compute the solution
- Can we find a more computationally tractable solution?
- Turns out the answer is yes!
- Let $\|\theta\|_1 = \sum_{i=1}^{p} |\theta_i|$ (this is called the L_1 norm)
- Consider the optimization problem

$$\arg\max_{\theta} p(x \mid \theta)$$

$$\text{such that } \|\theta\|_1 \leq c$$

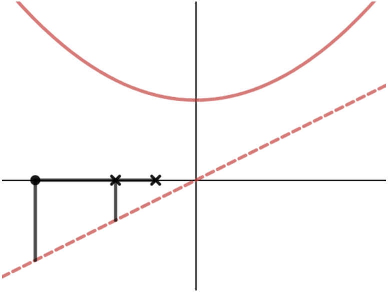
  Computationally tractable and has sparse solutions!
- The solution to this problem is called the LASSO estimator

Why is the LASSO Computable?

- Need to understand basic principles of optimization
- I give you arbitrary $f(x)$ and ask for $\min_x f(x)$
- Can you find this? What assumptions do we need on $f(x)$?
- Very hard in general, but what if f is differentiable?
- Can look for solution to $f'(x) = 0$
- May not have a unique solution
- Even if solution is unique, may not have closed form solution!
- We may not be able to "jump" straight to maximum
- But if we start at x, can we move closer to the maximum?

Optimization in One Picture

- Derivative negative $\rightarrow$ move in positive direction
- And vice versa
- Derivative larger $\rightarrow$ take bigger steps
- So if we start at x
- Move to $x_+ = x - f'(x)$
- And repeat until $f'(x) < \epsilon$
- In practice need to do $x_+ = x - \gamma f'(x)$ for $\gamma < 0$
- If $f'(x)$ is too large, we will overstep the minimum
- This is gradient descent



Gradient Descent Works

- Gradient descent effective on many differentiable functions
- Can prove it works well on very nice functions
- Previous function was bowl-shaped or convex
- In formal terms, this means that $f''(x) \geq 0$ everywhere
- For gradient descent to converge, need to avoid over-stepping
- Can assume derivative is Lipschitz continuous

$$|f'(x) - f'(y)| \leq L|x - y|$$

- Ensures tangent slopes don't vary too rapidly
- Claim: If f is twice-differentiable, convex, and with Lipschitz derivative, gradient descent converges

Proof That Gradient Descent Works

- First observe that

$$|f''(x)| = \lim_{h\to 0} \frac{|f'(x+h) - f'(x)|}{|h|} \leq \lim_{h\to 0} \frac{L|x+h-x|}{|h|} = L$$

- Next we Taylor expand f around our current point x:

$$\begin{aligned} f(y) &= f(x) + f'(x)(y-x) + \int_x^y f''(z)(y-x)dz \\ &\leq f(x) + f'(x)(y-x) + L\int_x^y (y-z)dz \\ &= f(x) + f'(x)(y-x) + \frac{L}{2}(y-x)^2 \end{aligned}$$

Proof Continued

- We got a quadratic upper bound for f:

$$f(y) \leq f(x) + f'(x)(y-x) + \frac{L}{2}(y-x)^2$$

- Next we plug in our gradient descent step $y = x_+ = x - \gamma f'(x)$:

$$\begin{aligned} f(x_+) &\leq f(x) + f'(x)(x + \gamma f'(x) - x) + \frac{L}{2}(x - \gamma f'(x) - x)^2 \\ &= f(x) + \gamma f'(x)^2 + \frac{L}{2}\gamma^2 f'(x)^2 \\ &= f(x) + \gamma \left(\frac{L}{2}\gamma - 1 \right) f'(x)^2 \end{aligned}$$

- If we take $\gamma = \frac{1}{L}$, then the above becomes

$$f(x) - \frac{1}{2L} f'(x)^2$$

I Promise This is the End of the Proof

- We got that
$$f(x_+) \leq f(x) - \frac{1}{2L} f'(x)^2$$
- This is less than $f(x)$ if $f'(x) \neq 0$
- But $f'(x) = 0$ only if we were already at the minimum
- So we successfully move toward minimum at each step
- Many functions of interest not differentiable everywhere
- This strategy can be generalized to those situations

Back to the LASSO

- So why is the LASSO computable?
- It is the solution to "constrained" problem

$$\arg\max_{\theta} p(x \mid \theta)$$
$$\|\theta\|_1 \leq c$$

- $p(x \mid \theta)$ not convex in θ
- But we can often replace $p(x \mid \theta)$ with a convex function for which $\arg\min_\theta f(\theta) = \arg\max_\theta p(x \mid \theta)$
- Let $f(\theta)$ be this new function, then we solve

$$\arg\min_{\theta} f(\theta)$$
$$\|\theta\|_1 \leq c$$

Computing the LASSO

- We can replace our constrained problem with a "penalized" problem:
$$\arg\min_{\theta} f(\theta) + \lambda\|\theta\|_1$$
- λ is a constant which is related to c
- $f(\theta)$ and $\|\theta\|_1$ are both convex
- $\|\theta\|_1$ is not differentiable
- So can't use gradient descent, but can use more general descent methods
- Today we have specialized algorithms that compute LASSO orders of magnitude faster than descent optimizers
- Compare to $f(\theta) + \lambda\|\theta\|_0$
- $\|\theta\|_0$ isn't even continuous! Can't use descent methods

Did Someone Say Sparsity?

- So we can compute the LASSO...
- The point was to find estimates of θ that were sparse
- It turns out LASSO solution is sparse almost always!
- Suppose that

$$p(x \mid \theta) = \frac{1}{2\pi} \exp\left(-\frac{1}{2}\left((x^1 - \theta^1)^2 + (x^2 - \theta^2)^2\right)\right)$$

  is the (2-dimensional) normal distribution
- Recall that max likelihood is equivalent to least squares:

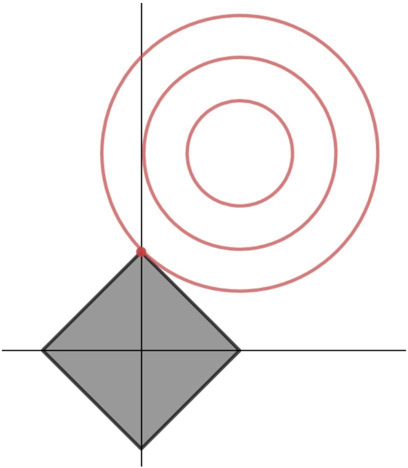
$$\arg \min_{(\theta^1, \theta^2)} \sum_{i=1}^{n} \left[(x_i^1 - \theta^1)^2 + (x_i^2 - \theta^2)^2\right] = \arg\min_{\theta} \sum_{i=1}^{n} d\left(x_i, \theta\right)^2$$

Sparsity in One Picture

- Then we want to solve

$$\arg\min_{\theta} \sum_{i=1}^{n} d\left(x_i, \theta\right)^2$$
$$\text{such that } \|\theta\|_1 \leq c$$

- The black region is the set allowed by the constraint
- Red circles are level curves of function to minimize
- Want to be on the inner-most valid level curve
- This occurs at a corner
- The corners are all sparse!



Recap of High-Dimensionality in Parameter Estimation

- When $p > n$, our models become unidentifiable
- We don't have enough data to estimate effects of all p variables
- But if we assume most have no effect...
- We can estimate which have no effect and then estimate the effects for remaining variables
- The LASSO is a computationally tractable way to get sparse solutions to high-dimensional linear regression problems

High-Dimensionality in Bayesian Statistics

- Recall that in Bayesian statistics we don't construct parameter estimators
- We want to compute a posterior distribution $p(\theta \mid x)$
- This is a probability distribution on parameters
- Tells us which parameters could have generated the data
- Instead of optimizing a parameter, we want to sample from the whole posterior distribution
- This is also very hard in high dimensions! Why?
- The curse of dimensionality!

The Curse of Dimensionality

- Many problems tend to get exponentially harder as the dimension increases
- This is due to an interesting fact about high-dimensional geometry
- Suppose I give you two circles with radii r and $r - \epsilon$
- What proportion of the outer circles area is taken up by the inner circle?

$$\frac{\pi(r-\epsilon)^2}{\pi r^2} = \left(1 - \frac{\epsilon}{r}\right)^2$$

- If $r = 1$ and $\epsilon = 0.1$, then this proportion is 0.81
- In other words, most of the outer circle

The Curse of Dimensionality

- What happens when the dimension increases?
- Now I give you d-dimensional spheres with radii r and $r - \epsilon$
- The ratio of the inner volume to the outer volume becomes

$$\left(1 - \frac{\epsilon}{r}\right)^d$$

  Again if $r = 1$ and $\epsilon = 0.1$, this is 0.9^d
- For $d = 10$, this is 0.35
- For $d = 50$, this is 0.005!
- As d increases, the proportion of the volume of a d-sphere that is contained near its boundary increases exponentially in d

The Goal of Sampling

- Why does this pose a problem for generating samples from the posterior?
- The answer depends on what we want to do with the samples
- Recall that we usually want some summaries of the posterior distribution
- If $\phi(\theta)$ is a function, we may be interested in

$$\int \phi(\theta) p(\theta \mid x) d\theta$$

- This is the average of ϕ over this distribution
- If $\phi(\theta) = \theta$, this is just the mean of the distribution
- Similarly, we can estimate the variance, or percentiles, or other functions

Estimating Posterior Averages

- Integrals of the form $\int \phi(\theta)p(\theta \mid x)d\theta$ may be impossible to solve exactly
- But if $(\theta^1, \dots, \theta^S)$ are samples from $p(\theta \mid x)$, then

$$\frac{1}{S}\sum_{i=1}^{S}\phi(\theta^i) \approx \int \phi(\theta)p(\theta \mid x)d\theta$$

  by the law of large numbers.
- But not all samples are equally good for computing these averages
- To see why, think about Riemann sums
- We can approximate the integral by a sum of rectangular regions

The Typical Set

- What is the volume of such a rectangular region?
- Take the base of the region to be some $B \subset \mathbb{R}^d$
- Let θ^* be the center point, so the height is $\phi(\theta^*)p(\theta^* \mid x)$
- Then the volume is $\mathrm{Vol}(B) \times \phi(\theta^*)p(\theta^* \mid x)$
- For now we will neglect the $\phi(\theta^*)$ factor
- We would like to be agnostic to choice of ϕ
- Then a region contributes significantly to the integral if $\mathrm{Vol}(B)$ and $p(\theta^* \mid x)$ are both large
- Those regions for which this is true make up the “typical set” of $p(\theta \mid x)$

The Typical Set

- All regions outside the typical set contribute negligibly to the integral
- Want to focus on generating samples from the typical set
- What is the typical set of a normal distribution?

$$p(\theta) = \left(\frac{1}{2\pi}\right)^{n/2} \exp\left(-\frac{1}{2}\sum_{i=1}^{d}\theta_i^2\right)$$

- In regions far from the origin, $p(\theta)$ is exponentially small
- What about near the origin?

The Curse of Dimensionality Returns

- Near the origin, $p(\theta)$ is maximal
- But in high dimensions, the proportion of total volume around the origin is vanishingly small!
- The only area with high volume and high probability is in a spherical shell
- As the dimension increases, the diameter of the typical set decreases
- This is called concentration of measure

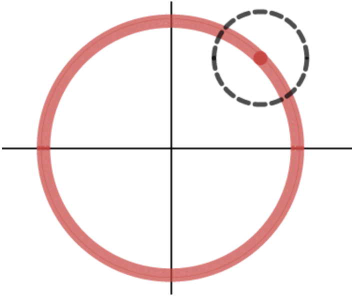
Metropolis Sampler

- To see why this poses a problem for sampling, have to understand sampling algorithms
- The Metropolis algorithm is a common sampling algorithm with nice properties
- Also easy to describe!
- Start at a sample point θ_0
- Take a random step by sampling θ^* from normal$(\theta_0, 1)$ distribution
- If $p(\theta^*)/p(\theta_0) > 1$, set $\theta_1 = \theta^*$ and repeat
- If $p(\theta^*)/p(\theta_0) \leq 1$, set $\theta_1 = \theta^*$ with probability $p(\theta^*)/p(\theta_0)$, otherwise stay at θ_0 and repeat

Metropolis Sampler

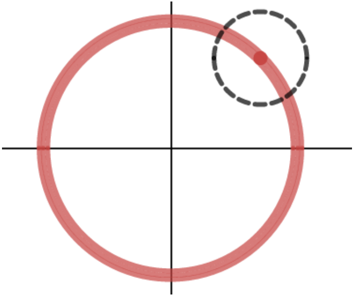
- The Metropolis sampler tries to move toward regions of higher probability
- Still sometimes randomly steps toward lower probability areas
- Thus it tries to explore adequately while staying away from useless low-probability regions
- But it fails horribly in high dimensions. Why?
- We can visualize the problem in two dimensions

The Problem with Metropolis



- Suppose the typical set is the red circular region
- The starting point is the red dot in the typical set
- Metropolis will take a random step from this point
- But the direction of the step is uniformly distributed
- Thus any point on the black circle is equally likely

The Problem with Metropolis



- Thus almost every step will be off of the circle
- Steps on the outer side will almost always reject
- Steps on the inner side will lead to a very low volume region
- The sampler will thus be very slow and over sample regions of little relevance to estimating to the integrals of interest

The Solution: Follow the Typical Set

- It is possible to design samplers that are smarter
- When generating the next step, we can use the geometry of the distribution
- We can follow the curvature of the typical set
- This involves deep connections to differential geometry and Hamiltonian dynamics
- These are the methods used by state-of-the-art samplers like that available in Stan