Statistical Inference Through the Lens of Information Geometry

Tony Wang (twang6@mit.edu)

May 16, 2019

1 Outline

Statistical inference is the process of making conclusions about the world given noisy data about the world. More precisely, the "world" here means a collection of random variables and "data" are just samples of those random variables.

In this paper we attempt to show how information geometry can be a useful perspective from which to study statistical inference. Our paper is organized as follows:

Section 2: We introduce and motivate some problems in statistical inference.

Sections 3-5: We build up some basic theory of information geometry, taking both a local and global approach.

Section 6-7: We use our information geometry theory to tackle the statistical inference problems we introduced earlier.

2 Some Problems in Statistical Inference

2.1 Maximum Entropy Priors

In the Bayesian school of statistical inference, the starting point of any inference task is coming up with a prior – a probability distribution that captures one's pre-existing beliefs about the world. The process of inference then boils down to observing the world and updating ones prior.

As an example imagine we are forecasting the an upcoming national election between two candidates. Our prior for the election outcome is simply the probabilities that we believe each candidate has of winning. As the election draws nearer, we may listen to the candidates give speeches, look at country-wide poll data, and watch news reporters present new facts about the candidates. Each of these new pieces of information have the potential to make us update our prior.

As it turns out, coming up with good priors is an important but difficult task. For example, in the scenario above, how would one quantify one's preexisting beliefs about the candidates? There is no catch all answer to this question, but one commonly used method

is to pick the prior that is consistent with some set of constraints and also has the maximum entropy – in a sense assuming as little as possible. This motivates the following problem:

Problem 2.1. *Let p be a distribution over $\mathcal{X}$ that satisfies*

$$\mathbb{E}_{x\sim p}\left[f_i(x)\right] = \alpha_i,$$

for some finite list of functions $f_i : \mathcal{X} \to \mathbb{R}$ and values $\alpha_i \in \mathbb{R}$.

What is the p that satisfies the above constraints and has the maximum possible entropy?

2.2 Parameter Estimation

A common problem in statistical inference is that of parameter estimation. One classical case of this problem looks as follows.

Problem 2.2. *We have a collection of probability distributions parameterized by a real valued parameter θ. That is to say, for every $\theta \in \mathbb{R}$ we have a probability distribution p_θ over $\mathcal{X}$. We are given a random variable x taking values in $\mathcal{X}$ and we know it is distributed according to p_{θ_0} for some fixed but unknown θ_0. How do figure out what θ_0 from observing x?*

To give a concrete example, x could be a noisy measurement of the speed of light made by a lab instrument. Imagine that based on the properties of our instrument we make the assumption that x is normally distributed with some known variance and unknown mean. A parameter estimation task could be to estimate the mean of x.

3 Setup of Information Geometry

Information geometry is the application of geometrical techniques (particularly those from differential geometry) to information theory. We present some interesting results in information geometry in this section, with the goal of using our results to analyze the problems stated in the previous section.

3.1 Preliminaries: The Space of Probability Distributions

One starting point for geometry is the space in which the objects of interest live in. For example, Euclidean plane geometry can be used to study objects living on a plane, spherical geometry can be used to study objects living on the surface of a sphere, and differential geometry can be used to study objects residing in abstract spaces called smooth manifolds.

In this paper, we will use information geometry to study objects lying within the probability simplex $\Pi_\mathcal{X}$ – the set of all probability distributions over $\mathcal{X}$. We will treat $\Pi_\mathcal{X}$ as the subset of $\mathbb{R}^{|\mathcal{X}|}$ where $\mathbf{x} \in \Pi_\mathcal{X}$ if $\sum_{i=1}^{|\mathcal{X}|} \mathbf{x}_i = 1$ and $\mathbf{x}_i \geq 0$ for all $1 \leq i \leq |\mathcal{X}|$. When it is clear/irrelevant what $\mathcal{X}$ is, we will drop the subscript and just write Π.

To make our analysis simpler, we will take $\mathcal{X}$ to be a finite set and work primarily within the interior of Π, which we denote $\langle\Pi\rangle$. Working with a finite $\mathcal{X}$ lets us avoid issues with convergence and working within $\langle\Pi\rangle$ means our probability distributions are strictly positive, which lets us avoid issues related to dividing by zero. Both of these restrictions can be lifted at the cost of specifying regularity conditions and more careful analysis.

3.2 The KL-Divergence

Having established our space, we spice things up further by introducing a distance-like function on our space: the KL-divergence. Given two probability distributions $p, q \in \mathcal{P}^{\mathcal{X}}$ their KL-divergence is defined as

$$D(p\|q) \triangleq \sum_{x\in\mathcal{X}} p(x) \log \frac{p(x)}{q(x)},$$

where log here and in the rest of the paper is the natural logarithm. Here we take the standard convention and take $0 \log 0 = 0$ and $0 \log \frac{c}{0} = \infty$. Note that by this convention when $p \in \langle \Pi \rangle$ the KL-divergence always evaluates to real number.

A classical result from information theory says that $D(p\|q) \geq 0$ with equality if and only if $p = q$. Thus in some sense D is a measure of distance between two probability distributions. However, there is a troubling flaw to this view: the KL-divergence is not symmetric.

One key insight of information geometry is that despite this asymmetry, many geometric notions related to distance can still be found from the KL-divergence. We explore some of these notions in the following sections.

4 Large Scale Geometry

We begin by taking analyzing some global geometry of the probability simplex. The results and presentation of this section largely follow [1] and [2].

4.1 I-Projections

Using the KL-divergence we can define the following projection operator

Definition 4.1 (Information Projection)**.** *Let $q \in \langle \Pi \rangle$ and let $\mathcal{K} \subset \Pi$ be closed subset. An information projection, or I-projection, of q onto $\mathcal{K}$ is a distribution $p^* \in \mathcal{K}$ that satisfies*

$$D(p^*\|q) = \min_{p\in\mathcal{K}} D(p\|q).$$

Some comments on this definition:

1. The existence of p^* is guaranteed by the fact that $\mathcal{K}$ is closed and D is continuous.
2. For an arbitrary $\mathcal{K}$ the I-projection is not necessarily unique. However, we shall show that it is unique for convex sets – this mirrors the behavior of the projection operator in Euclidean space.

To prove the uniqueness of I-projections onto convex sets we first prove the following inequality:

Theorem 4.2 (Pythagorean Theorem on the Probability Simplex)**.** *Let $q \in \langle \Pi \rangle$, let $\mathcal{C} \subset \Pi$ be a closed and convex set and let p^* be an I-projection of q onto $\mathcal{C}$. Then for any $p \in \mathcal{C}$ we have*

$$D(p\|q) \geq D(p\|p^*) + D(p^*\|q). \tag{1}$$

Proof. Let $p \in \mathcal{C}$ be an arbitrary point. Since $\mathcal{C}$ is convex, the line segment L connecting p^* and p lies within $\mathcal{C}$. Let's parameterize this line segment as a path $\alpha : [0,1] \to L$, where

$$\alpha(t) = p^* + t\,(p - p^*).$$

Since p^* minimizes $D(p\|q)$, we have that

$$\begin{aligned} 0 &\leq \frac{d}{dt} D(\alpha(t)\|q)\Big|_{t=0} \\ &= \sum_{x \in \mathcal{X}} (p(x) - p^*(x)) \log \frac{p^*(x)}{q(x)} \\ &= D(p\|q) - D(p\|p^*) - D(p^*\|q). \end{aligned}$$

Rearranging the first and last lines in the above inequality yields the desired result. □

Using the Pythagorean theorem above we can now prove the uniqueness of I-projection onto convex sets:

Corollary 4.2.1. *Let $q \in \langle \Pi \rangle$, let $\mathcal{C} \subset \Pi$ be a closed and convex set. Then the I-projection p^* of q onto $\mathcal{C}$ is unique.*

Proof. Let $p \in \mathcal{C}$ be a distribution different from p^*. Then $D(p\|p^*) > 0$, which means

$$D(p\|q) \geq D(p\|p^*) + D(p^*\|q) > D(p^*\|q).$$

Hence p^* is the unique projection. □

4.2 Linear Families

We saw in the previous section that enforcing the convexity of the set onto which we I-project yields uniqueness of the projection.

As it turns out, enforcing more conditions on the set onto which we I-project yields even nicer properties. In particular in this section we consider the case of I-projections onto an affine subset of our probability simplex. We make this notion precise via the following definition:

Definition 4.3 (Linear Families)**.** *A nonempty subset $\mathcal{L} \subset \Pi$ is a linear family if $\mathcal{L}$ is of the form*

$$\mathcal{L} = \{p \in \Pi \,|\, \mathbb{E}_p\,[f_i(\mathsf{x})] = \alpha_i,\, \forall\, 1 \leq i \leq K\}\,, \tag{2}$$

where $f_i : \mathcal{X} \to \mathbb{R}$ and $\alpha_i \in \mathbb{R}$ for $1 \leq i \leq K$.

Each constraint $\mathbb{E}_p\,[f_i(\mathsf{x})] = \alpha_i$ is a linear equation in $\mathbb{R}^{|\mathcal{X}|}$, which means a linear family is just the intersection of an affine subspace of $\mathbb{R}^{|\mathcal{X}|}$ with $\mathcal{L} \subset \langle \Pi \rangle$. This is why linear families are named as such.

Linear families have the following important property:

Lemma 4.4. *Let $\mathcal{L}$ be a linear family which has a non-empty intersection with $\langle \Pi \rangle$. Let $q \in \langle \Pi \rangle$. Then p^*, the I-projection of q onto $\mathcal{L}$ lies within $\langle \Pi \rangle$.*

Proof. Since q is in the interior of the simplex, $D(p^*\|q) < \infty$. By the Pythagorean theorem this means $D(p\|p^*) < \infty$ for all $p \in \mathcal{L}$. Since $\mathcal{L}$ has a non-empty intersection with $\langle\Pi\rangle$, this means $D(p\|p^*) < \infty$ for some strictly positive distribution p. Thus p^* cannot lie on the boundary of the simplex and must lie in $\langle\Pi\rangle$. □

We now consider the following question:

Problem 4.5. *Let $\langle\mathcal{L}\rangle = \mathcal{L} \cap \langle\Pi\rangle$. What subset of $\langle\Pi\rangle$ I-projects as some fixed $p_0 \in \langle\mathcal{L}\rangle$ onto $\mathcal{L}$?*

As we shall see, the subset that projects as p_0 onto $\mathcal{L}$ takes the form of an exponential family, which we now define:

Definition 4.6 (Exponential Families)**.** *The exponential family with natural statistics $f_1, \ldots, f_K : \mathcal{X} \to \mathbb{R}$ and base distribution $p_0 \in \langle\Pi\rangle$ is defined as*

$$\mathcal{E}_{p_0,f_1,\ldots,f_K} \triangleq \left\{ p \in \langle\Pi\rangle \;\middle|\; p(\cdot) \propto p_0(\cdot) \exp\left(\sum_{i=1}^{K} \theta_i f_i(\cdot)\right), \text{ for some } (\theta_1, \ldots, \theta_K) \in \mathbb{R}^K \right\}. \quad (3)$$

One important property of exponential families is that they do not depend on their base distribution. This is made precise by the following lemma:

Lemma 4.7. *If $q \in \mathcal{E}_{p,f_1,\ldots,f_K}$ then $\mathcal{E}_{p,f_1,\ldots,f_K} = \mathcal{E}_{q,f_1,\ldots,f_K}$.*

Proof. Left as a exercise for the reader. □

We are now ready to officially give an answer to Problem 4.5 via the following theorem:

Theorem 4.8. *Let $\mathcal{L}$ be a linear family with constraints functions $f_1, \ldots, f_K$ and constraint values $\alpha_1, \ldots, \alpha_K$ and let $\langle\mathcal{L}\rangle$ be nonempty. If $p_0 \in \langle\mathcal{L}\rangle$, the subset of $\langle\Pi\rangle$ which projects as p_0 is precisely the exponential family $\mathcal{E}_{p_0,f_1,\ldots,f_K}$.*

Proof. We begin by showing that $\mathcal{E}_{p_0,f_1,\ldots,f_K}$ projects to p_0.

Let $p \in \mathcal{L}$ and let $q \in \mathcal{E}_{p_0,f_1,\ldots,f_K}$. Then

$$\begin{aligned}
D(p\|q) &= \sum_{x\in\mathcal{X}} p(x) \log \frac{p(x)}{q(x)} \\
&= \sum_{x\in\mathcal{X}} p(x) \log \frac{p(x)}{c \cdot p_0(x) \exp\left(\sum_{i=1}^{K} \theta_i f_i(x)\right)} \\
&= D(p\|p_0) - \log c - \sum_{x\in\mathcal{X}} p(x) \sum_{i=1}^{K} \theta_i f_i(x) \\
&= D(p\|p_0) - \log c - \sum_{i=1}^{K} \theta_i \alpha_i.
\end{aligned}$$

Thus we get that

$$\operatorname*{argmin}_{p\in\mathcal{L}} D(p\|q) = \operatorname*{argmin}_{p\in\mathcal{L}} D(p\|p_0) = p_0.$$

So all of $\mathcal{E}_{p_0,f_1,\ldots,f_K}$ gets I-projected to p_0.

It remains to show that only no other distributions I-project to p_0. To do this we will prove that the exponential families with different base distributions in $\langle\mathcal{L}\rangle$ form a foliation of $\langle\Pi\rangle$, which is to say:

1. They cover all of $\langle\Pi\rangle$.
2. They are all disjoint.
3. They all have the same dimension. This follows immediately from the definition of exponential families, but is not really relevant to our proof.

Since we already shown exponential families project onto their base distributions, proving that exponential families foliate $\langle\Pi\rangle$ suffices to finish our proof.

The disjointedness of exponential families is a consequence of Lemma 4.7.

As for why exponential families cover $\langle\Pi\rangle$, let $q \in \langle\Pi\rangle$ be an arbitrary distribution which projects to p^*. By Lemma 4.4 we have that $p^* \in \langle\mathcal{L}\rangle$. Since p^* is the I-projection and lies on the interior of $\mathcal{L}$, for any distribution $p \in \langle\mathcal{L}\rangle$ we must have

$$\begin{aligned} 0 &= \frac{d}{dt} D(p^* + tp \| q)\Big|_{t=0} \\ &= \sum_{x\in\mathcal{X}} (p(x) - p^*(x)) \log \frac{p^*(x)}{q(x)}. \end{aligned}$$

This means $\log \frac{p^*}{q}$ is orthogonal to $(p - p^*)$ for all $p \in \langle\mathcal{L}\rangle$. But this set of $(p - p^*)$ is just the orthogonal complement to the space spanned by the f_is. Thus $\log \frac{p^*}{q}$ must be spanned by the f_is, meaning we can write

$$q(x) \propto p^*(x) \cdot \exp\left(\sum_{i=1}^{K} \theta_i f_i(x)\right) \tag{4}$$

for some $(\theta_1, \ldots, \theta_K) \in \mathbb{R}^K$. Thus q is in the exponential family with base distribution p^*. Since q was arbitrary element of $\langle\Pi\rangle$ the exponential families cover $\langle\Pi\rangle$. □

5 Local Geometry

Another way of analyzing the geometry of the probability simplex is to take a more local view. To begin, we analyze the local behavior of the KL divergence.

5.1 The Fisher Metric

We note that while the KL-divergence is not symmetric, to leading order it is symmetric. To be precise, we have the following theorem:

Theorem 5.1. *Let $|\mathcal{X}| \geq 2$, let $p \in \langle \Pi \rangle$, and let $\delta : \mathcal{X} \to \mathbb{R}$ be an infinitesimal perturbation such that $\sum_{x \in \mathcal{X}} \delta(x) = 0$. Then*

$$D(p + \delta \| p) = D(p \| p + \delta) = \frac{1}{2} \sum_{x \in \mathcal{X}} \frac{\delta(x)^2}{p(x)} + O(\delta^3). \tag{5}$$

Proof. Taking a Taylor expansion in $\delta(x)$ yields

$$\begin{aligned} D(p \| p + \delta) &= \sum_{x \in \mathcal{X}} p(x) \log \frac{p(x)}{p(x) + \delta} \\ &= \sum_{x \in \mathcal{X}} p(x) \left[-\frac{\delta(x)}{p(x)} + \frac{\delta(x)^2}{2p(x)^2} + O(\delta^3) \right] \\ &= -\sum_{x \in \mathcal{X}} \delta(x) + \frac{1}{2} \sum_{x \in \mathcal{X}} \frac{\delta(x)^2}{p(x)} + O(\delta^3) \\ &= \frac{1}{2} \sum_{x \in \mathcal{X}} \frac{\delta(x)^2}{p(x)} + O(\delta^3). \end{aligned}$$

Doing an analogous expansion for $D(p + \delta \| p)$ yields the desired result. □

Thus in a small region of $\langle \Pi \rangle$, the KL-divergence behaves approximately as a bona fide distance function. Moreover given two distributions p, q from this small region we have

$$D(p \| q) \approx \frac{1}{2} \sum_{x \in \mathcal{X}} \frac{(p(x) - q(x))^2}{p(x)}. \tag{6}$$

We note that the form given in (6) is very similar to that of a Euclidean norm. Since the Euclidean norm is induced via the standard inner product on $\mathbb{R}^n$, this motivates us to define a corresponding local inner product on Π. In differential geometry, such a local inner product is called a metric.

Definition 5.2 (The Fisher Metric)**.** *At a point $p \in \langle \Pi \rangle$, let $g_p : T_p\langle \Pi \rangle \times T_p\langle \Pi \rangle \to \mathbb{R}$ be an inner product where*

$$g_p(\mathbf{u}, \mathbf{v}) \triangleq \sum_{x \in \mathcal{X}} \frac{\mathbf{u}(x)\, \mathbf{v}(x)}{p(x)}. \tag{7}$$

Here $T_p\langle \Pi \rangle$ denotes the tangent space to $\langle \Pi \rangle$ at point p, which consists of vectors in the hyperplane tangent to $\langle \Pi \rangle$ – so vectors in $\mathbb{R}^{|\mathcal{X}|}$ whose components sum to zero.

We emphasize that this inner product varies [1] as move around on $\langle \Pi \rangle$ – this corresponds to the fact that the KL-divergence behaves differently in different portions of the probability simplex.

[1] In fact g actually varies smoothly, a point which we go into detail but actually turns $\langle \Pi \rangle$ into a Riemannian manifold.

5.2 The Gradient

As a refresher, recall the definition of gradient in $\mathbb{R}^n$.

Definition 5.3 (The Gradient in Euclidean Space)**.** *Let $f : \mathbb{R}^n \to \mathbb{R}$ be a differentiable function and let $\mathbf{p} \in \mathbb{R}^n$. The gradient of f at $\mathbf{p}$, written $\nabla f(p)$, is the unique vector in $\mathbb{R}^n$ such that*

$$\langle \nabla f(\mathbf{p}), \mathbf{v} \rangle = \frac{d}{dt} f(\mathbf{p} + t\mathbf{v})\Big|_{t=0} \tag{8}$$

holds for all $\mathbf{v} \in \mathbb{R}^n$. Here $\langle \cdot, \cdot \rangle$ is the standard inner product in $\mathbb{R}^n$.

The local inner product g allows us to construct a local version of the gradient on $\langle \Pi \rangle$:

Definition 5.4 (The Gradient on Π)**.** *Let $f : \langle \Pi \rangle \to \mathbb{R}$ be a differentiable function and let $p \in \langle \Pi \rangle$. Then $\nabla f(p) \in T_p\langle \Pi \rangle$ is the unique tangent vector such that*

$$g_p(\nabla f(p), v) = \frac{d}{dt} f(p + tv)\Big|_{t=0}, \tag{9}$$

for all $v \in T_p\langle \Pi \rangle$ and g_p is the Fisher metric as defined in the previous section.

Note that in the second definition our vectors v live in $T_p\langle \Pi \rangle$ instead of Euclidean space. While we can technically identify $T_p\langle \Pi \rangle$ with $\mathbb{R}^{|\mathcal{X}|}$ via treating Π as a simplex in $\mathbb{R}^{|\mathcal{X}|}$, we choose to write things in terms of $T_p\langle \Pi \rangle$ to emphasize that we are treating $\langle \Pi \rangle$ as a general manifold.

Indeed on a general manifold the tangent spaces at different points can not be identified with one another. For example, the tangent spaces of the sphere in $\mathbb{R}^3$, can be visualized as a tangent plane to the sphere, and these planes changes as we move around the sphere.

One benefit of adopting this more abstract definition is that it hides the complexities of our manifold's global structure and makes it easier to carry out local analysis.

6 Global Information Geometry and Maximum Entropy Priors

Our developments in global information geometry allow us to gain some insight into the form of maximum entropy priors. We explore this idea in this section.

6.1 I-Projections and Maximizing Entropy

I-projections are actually very closely tied to maximizing entropy due to the following identity:

$$D(p\|\text{uniform}) = \log|\mathcal{X}| - H(p). \tag{10}$$

The identity means that the maximum entropy distribution from a set $\mathcal{S}$ is equivalent to the I-projection of the uniform distribution onto $\mathcal{S}$. In particular, this means that the maximum entropy distribution with linear constraints specified by $f_1, \ldots, f_K$ and $\alpha_1, \ldots, \alpha_K$ takes the form

$$p_{\text{maxent}}(\cdot) \propto \exp\left(\sum_{i=1}^{K} \theta_i f_i(\cdot)\right). \tag{11}$$

6.2 An Example of a Max Entropy Distributions

It is a famous result that when one constrains the mean a variance of a distribution over $\mathbb{R}$, the maximum entropy distribution is Gaussian. This is no coincidence. Constrained the mean and variance is introducing constrains $f_1(x) = x$ and $f_2(x) = x^2$ with fixed expectations. Thus from our previous analysis we see that the maximum entropy should take the form (modulo convergence issues resulting from the transition to $\mathcal{X} = \mathbb{R}$)

$$p_{\text{maxent}}(\cdot) \propto \exp\left(ax^2 + bx\right), \tag{12}$$

precisely the form of a Gaussian.

7 Local Information Geometry and Parameter Estimation

In the problem of parameter estimation, it turns out that local information geometry allows us establish an upper bound on how well inference can be performed. We explore this notion in this section.

The presentation of this section is heavily inspired by [3].

7.1 The Gradient and the Cramér-Rao Bound

It turns out that the object which bounds the efficiency of parameter estimation is the gradient of our parameterization function. To be precise, we have the following celebrated result:

Theorem 7.1 (Cramér-Rao Bound)**.** *Using the same notation that we established in Section 2.2, say we have an unbiased*[2] *estimator $\hat{\theta} : \mathcal{X} \to \mathbb{R}$ of our the parameter $\theta : \langle \Pi \rangle \to \mathbb{R}$. Here unbiased means that $\mathbb{E}_{x\sim p}[\hat{\theta}(x)] = \theta(p)$ for all $p \in \langle \Pi \rangle$. Then for any $p \in \langle \mathcal{P}^{\mathcal{X}} \rangle$ (representing the true hidden distribution) it holds that*

$$\text{Var}_p\left(\hat{\theta}\right) \geq \|\nabla\theta(p)\|_p^2, \tag{13}$$

where $\|v\|_p^2 \triangleq g_p(v, v)$.

Proof. Treating $\langle \Pi \rangle$ as a subset of $\mathbb{R}^{|\mathcal{X}|}$, for $x \in \mathcal{X}$, we let $(\nabla\theta)^x$ denote the component of $\nabla\theta$ pointing in the direction of increasing $p(x)$.

We begin by noting that

$$\begin{aligned}
\|\nabla\theta\|_p^2 &= g_p(\nabla\theta(p), \nabla\theta(p)) \\
&= \frac{d}{dt}\,\theta\left(p + t \cdot \nabla\theta(p)\right)\Big|_{t=0}, \\
&= \sum_{x\in\mathcal{X}} \left[\frac{\partial\theta}{\partial p(x)}\right]_p (\nabla\theta(p))^x.
\end{aligned}$$

[2] The condition that $\hat{\theta}$ be unbiased is actually less restrictive than it appears – the proof of the unbiased case can be extended to that of the biased case relatively straightforwardly. Wikipedia offers a fairly simple presentation.

Since $\hat{\theta}$ is unbiased,

$$\left[\frac{\partial \theta}{\partial p(x)}\right]_p = \left[\frac{\partial}{\partial p(x)} \mathbb{E}_{t\sim p}\left[\hat{\theta}(t)\right]\right]_p = \hat{\theta}(x).$$

Thus

$$\|\nabla\theta\|_p^2 = \sum_{x\in\mathcal{X}} \hat{\theta}(x)(\nabla\theta(p))^x.$$

Now since $\nabla\theta(p) \in T_p\langle\Pi\rangle$, we have that $\sum_{x\in\mathcal{X}}(\nabla\theta)^x = 0$. which means

$$\|\nabla\theta\|_p^2 = \sum_{x\in\mathcal{X}} \left(\hat{\theta}(x) - \theta(p)\right)(\nabla\theta)^x.$$

Applying the Cauchy-Schwartz inequality to the right hand side of above equation yields

$$\begin{aligned}\|\nabla\theta\|_p^2 &\leq \left(\sum_{x\in\mathcal{X}} p(x)\left(\hat{\theta}(x) - \theta(p)\right)\right)^{1/2}\left(\sum_{x\in\mathcal{X}} \frac{((\nabla\theta)^x)^2}{p(x)}\right)^{1/2} \\ &= \sqrt{\mathrm{Var}_p\left(\hat{\theta}\right)}\sqrt{\|\nabla\theta(p)\|_p^2}.\end{aligned}$$

The desired result thus follows from a little algebraic manipulation of this final inequality. □

Heuristically some version of this theorem should be true. $\|\nabla\theta\|_p^2$ measures how fast θ changes as you move around in distribution space. For purposes of inference, the farther apart two distributions are the easier we can tell them apart. Thus in a region of fast changing θ a small change in distribution would result in a large change in θ – making estimating θ a difficult task.

7.2 Fixing Dimensionality Issues

In the above proof we assumed that the space of parameterized distributions (the domain of θ) was $\langle\Pi\rangle$. This is a little unrealistic since in practice we would not want to parameterize a high dimensional space Π using a one dimensional parameter θ. There are two ways to resolve this dilemma.

The first resolution is to note that the Cramér-Rao bound (and our proof) works the same when we restrict ourselves to a submanifold (a subset which inherits the metric) of $\langle Pi\rangle$. As a more realistic application, we could use the Cramér-Rao to bound the variance of an estimator for the parameter of a one dimensional exponential family.

The second resolution is to increase the dimension of our parameter space. The following corollary gives the generalization of the Cramér-Rao bound for such a multivariate parameterization.

Corollary 7.1.1 (Multivariate Cramér-Rao Bound)**.** *Say we have an unbiased estimate $\hat{\theta} : \mathcal{X} \to \mathbb{R}^n$ of our the parameter $\theta : \langle\Pi\rangle \to \mathbb{R}^n$. Then for any $p \in \langle\mathcal{P}^{\mathcal{X}}\rangle$ (representing the true hidden distribution) it holds that*

$$\mathrm{Var}_p\left(\hat{\theta}_i\right) \geq \|\nabla\theta_i(p)\|_p^2, \tag{14}$$

$$\mathrm{Var}_p\left(\hat{\theta}_i\right) + \mathrm{Var}_p\left(\hat{\theta}_j\right) + 2\,\mathrm{Cov}_p\left(\hat{\theta}_i, \hat{\theta}_j\right) \geq \|\nabla\theta_i(p)\|_p^2 + \|\nabla\theta_j(p)\|_p^2 + 2\,g_p\left(\nabla\theta_i(p), \nabla\theta_j(p)\right). \tag{15}$$

Proof. The first inequality follows directly from the scalar Cramér-Rao Bound and the second follows from applying the scalar Cramér-Rao Bound to $\theta_i + \theta_j$:

$$\begin{aligned}
&\mathrm{Var}_p\left(\hat{\theta}_i\right) + \mathrm{Var}_p\left(\hat{\theta}_j\right) + 2\,\mathrm{Cov}_p\left(\hat{\theta}_i, \hat{\theta}_j\right) \\
&= \mathrm{Var}_p\left(\hat{\theta}_i + \hat{\theta}_j\right) \\
&\geq \|\nabla\theta_i(p) + \nabla\theta_j(p)\|_p^2 \\
&= \|\nabla\theta_i(p)\|_p^2 + \|\nabla\theta_j(p)\|_p^2 + 2\,g_p\left(\nabla\theta_i(p), \nabla\theta_j(p)\right).
\end{aligned}$$

□

Note that the bound on covariance in the multivariate Cramér-Rao bound depends on the variance of the individual estimators – if the individual estimators are too noisy then we cannot say much about the covariance.

7.3 How Tight is Cramér-Rao?

A natural question to ask is how good the Cramér-Rao bound actually is. As it turns out it is actually fairly tight in the limit of many independent and identically distributed (i.i.d.) samples. To make this point precise we need to extend our Cramér-Rao bound to the i.i.d. multi-sample case:

Corollary 7.1.2 (I.I.D. Multi-Sample Cramér-Rao Bound)**.** *Let $\theta : \langle\Pi_{\mathcal{X}}\rangle \to \mathbb{R}^n$. The i.i.d. multi-sample parameter estimation problem consists of estimating the θ of some $p \in \langle\Pi_{\mathcal{X}}\rangle$ given $x_1, \ldots, x_N$ sampled independently from p*

Now let $\langle\Pi_{\mathcal{X}}\rangle^N \subset \langle\Pi_{\mathcal{X}^N}\rangle$ be the sub-manifold consisting of distributions over $\mathcal{X}^N$ which are i.i.d.. Note that there is a natural identification between $\langle\Pi_{\mathcal{X}}\rangle$ and $\langle\Pi_{\mathcal{X}}\rangle^N$, namely the one where a distribution $p \in \langle\Pi_{\mathcal{X}}\rangle$ is identified with N independent copies of p, which we denote as p^N. To clarify we have $p^N \in \langle\Pi_{\mathcal{X}}\rangle^N$.

This identification also allows us to extend θ to $\langle\Pi_{\mathcal{X}}\rangle^N$ via the relation

$$\theta_N(p^N) = \theta(p).$$

Thus we can re-frame the multi-sample parameter estimation problem as a single parameter estimation problem on the sub-manifold $\langle\Pi_{\mathcal{X}}\rangle^N$ with parameter θ_N.

Thus let $\hat{\theta} : \langle\Pi_{\mathcal{X}}\rangle^N \to \mathbb{R}$ be our unbiased estimator. Then

$$\mathrm{Var}_{p^N}\left(\hat{\theta}\right) \geq \left\|\nabla\theta_N(p^N)\right\|_{p^N}^2 = \frac{\|\nabla\theta(p)\|_p^2}{N}. \tag{16}$$

Proof. The inequality in Equation (16) follows from the single parameter Cramér-Rao bound. It remains the show the equality.

Let $v \in T_p\langle\Pi_{\mathcal{X}}\rangle$. From the definition of the Fisher metric we have that

$$\|v\|_p^2 = \frac{d}{dt} D\left(p \,\|\, p + tv\right)\Big|_{t=0}.$$

Since for $p, q \in \langle \Pi_{\mathcal{X}} \rangle$ we have that $ND(p\|q) = D(p^N\|q^N)$, this means $N\|v\|_p^2 = \left\|v^N\right\|_p^2$. And since a norm uniquely determines an inner product via the polarization identity, we have that

$$Ng_p(u, v) = g_{p^N}\left(u^N, v^N\right)$$

for all $u, v \in T_p\langle \Pi_{\mathcal{X}} \rangle$.

Now we also know that for all $v \in T_p\langle \Pi_{\mathcal{X}} \rangle$ we have

$$g_p\left(\nabla\theta(p), v\right) = \frac{d}{dt}\,\theta\left(p + tv\right)\Big|_{t=0} = \frac{d}{dt}\,\theta_N\left(p^N + tv^N\right)\Big|_{t=0} = g_{p^N}\left(\nabla\theta_N(p^N), v^N\right).$$

Combining these last two facts we get that for all $v \in T_p\langle \Pi_{\mathcal{X}} \rangle$ we have

$$g_{p^N}\left(\nabla\theta_N(p^N), v^N\right) = g_{p^N}\left(\frac{(\nabla\theta(p))^N}{N}, v^N\right).$$

Thus it must be the case that $\nabla\theta_N(p) = \dfrac{(\nabla\theta(p))^N}{N}$. From here we see that

$$\begin{aligned}
\|\nabla\theta(p)\|_p^2 &= g_p\left(\nabla\theta(p), \nabla\theta(p)\right) \\
&= g_{p^N}\left(\nabla\theta_N(p^N), (\nabla\theta(p))^N\right) \\
&= g_{p^N}\left(\nabla\theta_N(p^N), N\nabla\theta_N(p^N)\right) \\
&= N\left\|\nabla\theta_N(p^N)\right\|_{p^N}^2,
\end{aligned}$$

from which our desired equality immediately follows. □

Now back to the question of the tightness of Cramér-Rao. It turns out that with sufficient technical conditions (see Section 7.3 of [6]), the maximum likelihood parameter estimate actually achieves the i.i.d multi-sample Cramér-Rao bound. Thus in this limiting sense Cramér-Rao is tight.

References

[1] 6.437 lecture notes sections 14 and 17, Spring 2019.

[2] Imre Csiszár, Paul C Shields, et al. Information theory and statistics: A tutorial. *Foundations and Trends® in Communications and Information Theory*, 1(4):440–446, 2004.

[3] Anthony D Blaom. A geometer's view of the the cramer-rao bound on estimator variance. *arXiv preprint arXiv:1710.01598*, 2017.

[4] Shun-ichi Amari. *Information geometry and its applications*, volume 194. Springer, 2016.

[5] Thomas M Cover and Joy A Thomas. *Elements of information theory*. John Wiley & Sons, 2012.

[6] Erich Leo Lehmann. *Elements of large-sample theory*. Springer Science & Business Media, 2004.