An information geometric perspective on active learning

Chen-Hsiang Yeang

Artificial Intelligence Lab, MIT, Cambridge, MA 02139, USA
chyeang@ai.mit.edu

Abstract. The Fisher information matrix plays a very important role in both active learning and information geometry. In a special case of active learning (nonlinear regression with Gaussian noise), the inverse of the Fisher information matrix – the dispersion matrix of parameters – induces a variety of criteria for optimal experiment design. In information geometry, the Fisher information matrix defines the metric tensor on model manifolds. In this paper, I explore the intrinsic relations of these two fields. The conditional distributions which belong to exponential families are known to be dually flat. Moreover, the author proves for a certain type of conditional models, the embedding curvature in terms of true parameters also vanishes. The expected Riemannian distance between current parameters and the next update is proposed to be the loss function for active learning. Examples of nonlinear and logistic regressions are given in order to elucidate this active learning scheme.

1 Introduction

Active learning is a subcategory of machine learning. The learner seeks new examples from a specific region of input space instead of passively taking the examples generated by an unknown oracle. It is crucial when the effort of acquiring output information is much more demanding than collecting the input data. When the objective is to learn the parameters of an unknown distribution, a data point $(\mathbf{x}, \mathbf{y})$ contains input variables $\mathbf{x}$ and output variables $\mathbf{y}$. Actively choosing $\mathbf{x}$ distorts the natural distribution of $p(\mathbf{x}, \mathbf{y})$ but generates no bias on the conditional distribution $p(\mathbf{y}|\mathbf{x})$. Therefore, parameters of $p(\mathbf{y}|\mathbf{x})$ can be estimated without bias.

One of the most well-known active learning schemes for parameter estimation is optimal experiment design ([7]). Suppose $y \sim N(\theta \cdot \mathbf{f}(\mathbf{x}), \sigma^2)$. Define the *Fisher information matrix* of n inputs $(\mathbf{x_1}, \cdots, \mathbf{x_n})$ as $M = \sum_{t=1}^{n} \mathbf{f}(\mathbf{x_t}) \cdot \mathbf{f}(\mathbf{x_t})^T$, and $Y = \sum_{t=1}^{n} y_t \mathbf{f}(\mathbf{x_t})$. Then the maximum likelihood estimator of θ is the linear estimator $\hat{\theta} = M^{-1}Y$, and the dispersion matrix of estimated parameters $\hat{\theta}$ is $V(\hat{\theta}) = E\{(\hat{\theta} - \theta)(\hat{\theta} - \theta)^T\} = M^{-1}$. The dispersion matrix measures the deviation of the estimated parameters from the true parameters. A variety of loss functions based on the dispersion matrix are proposed. An optimal experiment design scheme selects an input configuration $(\mathbf{x_1}, \cdots, \mathbf{x_n})$ which minimizes the loss function.

The Fisher information matrix of a probability density function $p(\mathbf{x};\zeta)$ is

$$g_{ij}(\zeta) = E_{\mathbf{x}}\{\frac{\partial}{\partial\zeta_i}\log p(\mathbf{x};\zeta)\frac{\partial}{\partial\zeta_j}\log p(\mathbf{x};\zeta)\}. \tag{1}$$

where ζ is the parameter of the probability density function. This quantity constitutes the foundation of information geometry. A statistical model can be viewed as a manifold S imbedded in the high dimensional Euclidean space with the coordinate system ζ. The metric tensor on S in terms of coordinates ζ is the Fisher information matrix (equation 1). The metric tensor defines the *distance* of distributions on the manifold. A manifold in which the metric tensor is defined is called a Riemannian manifold.

The presence of Fisher information matrix in both contexts is no coincidence. The inverse Fisher information matrix asymptotically tends to the dispersion matrix according to Cramer-Rao theorem ([7]). This matrix characterizes the deviation of the estimated parameters from the real one. Similarly, the metric tensor characterizes the "distance" between two distributions on the model manifold. Therefore, it is strongly motivated to devise a loss function for active learning from a geometric perspective.

Previous works on information geometry focus on applying the notion of projection on various problems (for instance, [6], [3], and [1]). On the other hand, previous works on active learning are aimed at extending the scheme beyond parameter estimation (for example, [12]) or adopting a Bayesian approach to learning (for example, [11]). The contribution of this paper is to treat active learning under the framework of information geometry. I define the loss function as the expected Riemannian distance on the manifold between the current and the new estimated parameters. This loss function is closely related to the Kullback-Leibler divergence for exponential families.

The structure of this paper is organized as follows. Section 2 introduces basic notions of information geometry and sets the form of input-output models in terms of information geometry. Section 3 states a sufficient condition for vanishing embedding curvature in terms of true parameters on the model manifold. Section 4 introduces the loss function and scheme for active learning. Section 5 illustrates the active learning by examples of nonlinear and logistic regressions. Section 6 summarizes the works in this paper and points out future works.

2 Information geometry of input-output models

In this paper, I focus on a special type of active learning problem: estimating the parameters of a conditional probability density $p(\mathbf{y}|\mathbf{x},\theta)$. θ is the true parameter to be learned. The learner is free to choose input $\mathbf{x}$. The output is generated according to the conditional density $p(\mathbf{y}|\mathbf{x},\theta)$ (for convenience I write $p(\mathbf{y}|\mathbf{x},\theta) = p(\mathbf{y};\mathbf{x},\theta)$). $\mathbf{y},\mathbf{x},\theta$ are all vectors.

The conditional density in an exponential family can be written in the following form:

$$p(\mathbf{y};\mathbf{x},\theta) = exp\{\mathbf{T}(\mathbf{y})\cdot\mathbf{F}(\mathbf{x},\theta) - \psi(\mathbf{x},\theta) + k(\mathbf{y})\}, \tag{2}$$

where $\mathbf{T}(\mathbf{y})$ is the sufficient statistic of $\mathbf{y}$, $\mathbf{F}(\mathbf{x},\theta)=\zeta$ is the natural parameter and $\psi(\mathbf{x},\theta)$ is the partition function of $p(\mathbf{y};\mathbf{x},\theta)$, $k(\mathbf{y})$ is the Jacobi matrix of transforming from $\mathbf{y}$ to the sufficient statistic $\mathbf{T}(\mathbf{y})$. The natural parameter is a function of both inputs $\mathbf{x}$ and true parameters θ. We can view $p(\mathbf{y};\mathbf{x},\theta)$ as a curved exponential family where its natural parameters $\zeta=\mathbf{F}(\mathbf{x},\theta)$ are characterized by inputs $\mathbf{x}$ and true parameters θ.

In order to study the properties on the model manifold some notions of differential geometry need to be introduced. The metric tensor $g_{ij}(\zeta)$ has the same form of equation 1 (substitute $p(\mathbf{x},\zeta)$ with $p(\mathbf{y};\zeta)$). It defines the Riemannian distance between $p(\mathbf{y};\zeta)$ and an infinitesimally close density function $p(\mathbf{y};\zeta+\mathbf{d}\zeta)$ on the manifold S:

$$ds^2 = g_{ij}(\zeta)d\zeta^i d\zeta^j,$$

where the Einstein's summation of index convention is used in tensorial equations: summing over the indices which appear more than once in the formula. The value of the metric tensor changes when different coordinate systems are used, but the square distance ds^2 is invariant under coordinate transformations.

Another important quantity in information geometry is connection. The connection Γ^i_{jk} is a three-index quantity which defines the correspondence between vectors in different tangent spaces of the manifold ([5]). Notice this is a coordinate system-specific quantity. The *actual rate of change* of a vector field X on the manifold is the change of X along some coordinate curve, plus the change of the coordinate curve itself:

$$\frac{DX^i}{d\zeta^j} = \frac{\partial X^i(\zeta)}{\partial \zeta^j} + \Gamma^i_{jk}X^k(\zeta).$$

This is called the covariant derivative of X. The *geodesic* on a manifold is a curve whose rate of change of its tangent vectors along the curve is zero. It corresponds to the notion of a straight line in a Euclidean space. Setting the covariant derivative of $\dot{\zeta}(t)$ equal to zero, the equation of a geodesic curve $\zeta(t)$ is

$$\frac{d^2\zeta^i(t)}{dt^2} + \Gamma^i_{jk}\dot{\zeta}^j(t)\dot{\zeta}^k(t) = 0. \tag{3}$$

Notice when $\Gamma^i_{jk}=0$ equation 3 reduces to an ordinary second order differential equation, and the solution becomes

$$\zeta(t) = \zeta_0 + t(\zeta_1 - \zeta_0).$$

This corresponds to a straight line in an Euclidean space. Therefore, a manifold is flat when there exists some coordinate system which makes the connection vanish.

The α-connection of an statistical manifold is defined by Amari ([5]):

$$\Gamma^{(\alpha)}_{ijk} = E_{\mathbf{y}}\{\partial_i\partial_j\ell(\mathbf{y};\zeta)\partial_k\ell(\mathbf{y};\zeta)\} + \tfrac{1-\alpha}{2}E_{\mathbf{y}}\{\partial_i\ell(\mathbf{y};\zeta)\partial_j\ell(\mathbf{y};\zeta)\partial_k\ell(\mathbf{y};\zeta)\}, \tag{4}$$

where $\ell(.) = \log p(.)$, $\partial_i = \frac{\partial}{\partial \zeta_i}$, and $\Gamma_{ijk} = \Gamma_{ij}^{m} g_{mk}$, g_{mk} is the inverse matrix of g^{mk}. When $\alpha = 1$, then $\Gamma_{ijk}^{(\alpha)}(\zeta) = 0$ for an exponential family $p(\mathbf{y}; \zeta)$, thus $\zeta(t) = \zeta_0 + t(\zeta_1 - \zeta_0)$ is a geodesic on the manifold. The manifold is called e-flat in this case. Similarly, we can express $\Gamma_{ijk}^{(\alpha)}$ in terms of expectation parameters η, where $\eta_i = E\{T(x)_i\}$. $\Gamma_{ijk}^{(\alpha)}(\eta) = 0$ for mixture family distributions when $\alpha = -1$. This is called m-flat ([5]).

The dual flatness theorem was proved by Amari ([4]): a statistical manifold is e-flat if and only if it is m-flat. Therefore, both exponential and mixture families are dually flat manifolds. This theorem is important because it allow us to treat an exponential family manifold as a Euclidean space. However, it does not guarantee that the connection of the manifold vanishes for every coordinate system. The connection under a particular coordinate system is called the *embedding curvature* of the manifold. The embedding curvature of some coordinate system may not vanish when the manifold is both e-flat and m-flat.

The manifold of multiple data points is the product space of manifolds of individual data points. For ordinary exponential families, let $\mathbf{X} = (\mathbf{x_1}, \cdots, \mathbf{x_n})$ be n samples drawn from the same distribution, then the joint probability $p(\mathbf{X})$ forms a manifold $S_n = S \times \cdots \times S$ which is imbedded in an $n \times m$ dimensional Euclidean space. Here we are interested in the conditional densities of input-output models. Let $\mathbf{X} = (\mathbf{x_1}, \cdots, \mathbf{x_n}) = (\mathbf{a_1}, \cdots, \mathbf{a_n})$ be n fixed inputs and $\mathbf{Y} = (\mathbf{y_1}, \cdots, \mathbf{y_n})$ be their responses. The joint density of the n samples is

$$p(\mathbf{y_1}, \cdots, \mathbf{y_n}, \mathbf{x_1}, \cdots, \mathbf{x_1}, \cdots, \mathbf{x_n}; \theta) = \prod_{t=1}^{n} p(\mathbf{y_t}|\mathbf{x_t}) p(\mathbf{x_t})$$
$$= \prod_{t=1}^{n} p(\mathbf{y_t}|\mathbf{x_t} = \mathbf{a_t}) p(\mathbf{x_t} = \mathbf{a_t}) \quad = \exp\{\sum_{t=1}^{n} \mathbf{T}(\mathbf{y_t}) \cdot \mathbf{F}(\mathbf{x_t}, \theta) - \psi(\mathbf{x_t}, \theta) + k(\mathbf{y_t}) + \log p(\mathbf{x_t} = \mathbf{a_t})\}.$$

Each fixed input $\mathbf{x_t} = \mathbf{a_t}$ induces a manifold $S(\mathbf{a_t})$ and the product of conditional densities forms a submanifold M of the product manifold $S_n(\mathbf{a}) = \prod_{t=1}^{n} S(\mathbf{a_t})$. Each $S(\mathbf{a_t})$ is imbedded in an r-dimensional Euclidean space because it is parameterized by θ. However, M is also parameterized by θ, thus it lives in a much more compact space than $S_n(\mathbf{a})$.

3 Embedding curvature on the true parameter coordinate systems

The dual flatness theorem affirms the flatness of an exponential family manifold. Since M is a submanifold of an exponential family manifold, it is also flat. However, the α-connection of an exponential family vanishes only under the natural parameter coordinate systems; namely,

$$\zeta = (\zeta_\mathbf{1}, \cdots, \zeta_\mathbf{n}) = (\mathbf{F}(\mathbf{x_1}, \theta), \cdots, \mathbf{F}(\mathbf{x_n}, \theta)) = \mathbf{F}(\mathbf{X}, \theta).$$

When we use the coordinate system of true parameters $\theta = (\theta_1, \cdots, \theta_r)$, the α-connection does not guarantee to vanish. This means the curve

$$\theta(t) = \theta_0 + (\theta_1 - \theta_0)t$$

is no longer a geodesic. Knowing the geodesics yields the advantages of evaluating distances efficiently. Therefore, it is important to understand the condition when the embedding curvature vanishes.

Theorem 1 Let $p(\mathbf{y};\mathbf{x},\theta)$ be the conditional probability density in equation 2. $\zeta = \mathbf{F}(\mathbf{x},\theta)$ has the following form:

$$F(\mathbf{x},\theta) = (\theta_1 f_1(\mathbf{x}), \cdots, \theta_r f_r(\mathbf{x})). \tag{5}$$

where r is the dimension of y's sufficient statistic. Let

$$B_a^i = \delta_a^i \frac{\partial(\theta_i f_i(\mathbf{x}))}{\partial \theta_a}, \tag{6}$$

be the Jacobian from θ to ζ. Then both metric tensor and α-connection are invariant under coordinate transformation:

$$g_{ab}(\mathbf{x},\theta) = B_a^i B_b^j g_{ij}(\zeta), \tag{7}$$

$$\Gamma_{abc}^{(\alpha)}(\mathbf{x},\theta) = B_a^i B_b^j B_c^k \Gamma_{ijk}^{(\alpha)}(\zeta). \tag{8}$$

i, j, k are indices of natural parameter ζ components and a, b, c are indices of true parameter θ components.

Proof

The first statement holds for any coordinate transformation (see [5]).

From the definition of α-connection,

$$\Gamma_{abc}^{(\alpha)}(\mathbf{x},\theta) = E_{\mathbf{y}}\{\partial_a\partial_b\ell\partial_c\ell\} + \frac{1-\alpha}{2} E_{\mathbf{y}}\{\partial_a\ell\partial_b\ell\partial_c\ell\}$$

Since $\partial_a\ell = B_a^i\partial_i\ell$, the second term conforms with equation 8 after coordinate transformation. Apply coordinate transformation to the first term.

$$\begin{aligned}\partial_a\partial_b\ell &= \partial_a(B_b^j\partial_j\ell)\\ &= \partial_a(\delta_b^j \tfrac{\partial(\theta_j f_j(\mathbf{x}))}{\partial\theta_b}\partial_j\ell)\\ &= B_b^j\partial_a\partial_j\ell + \delta_b^j\partial_a(\tfrac{\partial\theta_j f_j(\mathbf{x})}{\partial\theta_b})\partial_j\ell\\ &= B_b^j\partial_a\partial_j\ell + \partial_a(f_b(\mathbf{x}))\partial_b\ell.\end{aligned} \tag{9}$$

Since $f_b(\mathbf{x})$ is a constant for θ, the second term of equation 9 vanishes. Thus both terms of the α-connection follow the form of equation 8. The theorem holds. Q.E.D.

Theorem 1 preserves the α-connection form of the manifold under coordinate transformation $\theta = F^{-1}(x,\zeta)$. This theorem holds under a specific type of input-output model (equation 5). Under this model each component of the natural parameter is decoupled into the effect of input $f_i(\mathbf{x})$ and the effect of parameter θ_i, and natural parameters ζ and true parameters θ have the same dimension. While the transformation is linear in true parameters, it does not need to be linear in inputs. Since the connection on natural parameters $\mathit{\Gamma}^{(\alpha)}_{ijk}(\zeta) = 0$, the connection on true parameters $\mathit{\Gamma}^{(\alpha)}_{abc}(x,\theta)$ also vanishes. Therefore, the curve $\theta(t) = \theta_0 + t(\theta_1 - \theta_0)$ is a geodesic on the manifold. This property allows us to evaluate the Riemannian distance on the manifold efficiently.

4 Active learning schemes on manifolds

With the notion of information geometry, active learning procedures can be viewed from a geometric perspective. The goal is to find the true parameter θ^* of the underlying process. The learner selects inputs of the samples. Here the conventional *myopic learning* scheme is used ([13]). Under this scheme the learner chooses an input that optimizes the immediate loss based on current estimate of θ. Procedures of an active learning scheme are as follows:

1. Start with a collection of random samples D_0. Repeat the following steps.
2. Find the maximum likelihood estimate $\hat{\theta}(D_n)$ based on D_n.
   The problem of maximum likelihood parameter estimation can be viewed as the projection from the data points to the model manifold along the m-geodesic ([6] and [3]). The manifold of θ changes with inputs $(\mathbf{x_1}, \cdots, \mathbf{x_n})$ (denoted as $\mathbf{M}(\theta; \mathbf{x_1}, \cdots, \mathbf{x_n})$).
3. For each candidate input for the next step $\mathbf{x_{n+1}}$, evaluate some expected loss function $E_x\{E_y\{\mathcal{L}(\hat{\theta}(D_n), \hat{\theta}(D_n, (\mathbf{x}, \mathbf{y})))\}\}$.
   This routine is usually the most time-consuming part of active learning. Suppose a new input-output pair $(\mathbf{x_{n+1}}, \mathbf{y_{n+1}})$ is given, then one can compute the ML estimate $\hat{\theta}(D_n, (\mathbf{x}, \mathbf{y}))$. A loss function $\mathcal{L}(\hat{\theta}(D_n), \hat{\theta}(D_n, (\mathbf{x_{n+1}}, \mathbf{y_{n+1}})))$ is constructed to capture the *deviation* between the new estimated parameter and the original one. Since $\mathbf{y_{n+1}}$ is generated from the unknown process, the resultant $\hat{\theta}(D_n, (\mathbf{x_{n+1}}, \mathbf{y_{n+1}}))$ is a random variable. The expected loss function is evaluated under the current estimate of the distribution of $\mathbf{y_{n+1}}$ and $\mathbf{x}$:

   $$\begin{aligned}&E_{\mathbf{x}}\{E_{\mathbf{y_{n+1}}}\{\mathcal{L}(\hat{\theta}(D_n), \hat{\theta}(D_n, (\mathbf{x_{n+1}}, \mathbf{y_{n+1}})))\}\} = \\ &\int q(\mathbf{x}) p(\mathbf{y_{n+1}}; \mathbf{x_{n+1}}, \hat{\theta}(D_n)) \mathcal{L}(\hat{\theta}(D_n), \hat{\theta}(D_n, (\mathbf{x_{n+1}}, \mathbf{y_{n+1}}))) d\mathbf{y_{n+1}} d\mathbf{x}.\end{aligned}$$

4. Find the input $\hat{\mathbf{x}}_{n+1}$ which minimizes the expected loss function $E_{\mathbf{x}}\{E_{\mathbf{y}}\{\mathcal{L}(\hat{\theta}(D_n), \hat{\theta}(D_n, (\mathbf{x_{n+1}}, \mathbf{y})))\}\}$.
   Ideally, the expected loss function of all inputs should be evaluated. Since it may not have simple analytic forms, a sampling strategy is usually adopted.
5. Generate a sample by querying the output with input $\hat{\mathbf{x}}$. Incorporate this sample into D_n to form D_{n+1}.

The crux of this scheme is the choice of the loss function. Various loss functions related to the dispersion matrix $V(\hat{\theta})$ (the inverse of Fisher information matrix) are proposed ([7]): for example, the determinant of V (D-optimal), the trace of V, or the maximum of any $\psi V \psi^T$ among normalized vector ψ (min-max optimal). While these loss functions might capture the *dispersion* of $\hat{\theta}$ evaluated by adding new possible samples, they do not explicitly bear geometrical interpretations on the manifold. One sensible choice of the loss function is the Riemmanian (square) distance between the conditional density according to current estimate $p(\mathbf{y};\mathbf{x},\theta(\hat{\mathbf{D}}_\mathbf{n}))$ and the new estimate by incorporating $(\mathbf{x_{n+1}},\mathbf{y_{n+1}})$ $(p(\mathbf{y};\mathbf{x},\hat{\theta}(\mathbf{D_n},(\mathbf{x_{n+1}},\mathbf{y_{n+1}}))))$ on the manifold $M(\mathbf{x_1},\cdots,\mathbf{x_{n+1}})$ of conditional distributions. The Riemannian distance between two points p_0 and p_1 on a Riemannian manifold is the square length of the geodesic $C(t)$ connecting them:

$$D(\theta_0,\theta_1)=(\int_0^1\sqrt{g_{ij}(C(t))\frac{dC^i(t)}{dt}\frac{dC^j(t)}{dt}}dt)^2, \tag{10}$$

where $C(t)$ is parameterized such that $C(0)=p_0$ and $C(1)=p_1$. This is in general a non-trivial task. On a dually flat manifold of statistical models, however, the Kullback-Leibler divergence is usually treated as a (quasi) distance metric. Amari ([5]) has proved the KL divergence between two infinitesimally close distributions is half of their Riemannian distance:

$$D_{KL}(p(x;\theta)\|p(x;\theta+d\theta))=\frac{1}{2}g_{ij}(\theta)d\theta^i d\theta^j.$$

Moreover, it is also known that the KL divergence is the Riemannian distance under Levii-Civita connection ([4]). The computation of this distance uses different geodesic paths when traversing in opposite directions. From a point P to another point R it firstly projects P to a point Q along an m-geodesic then connects Q and R via an e-geodesic. Conversely from R to P it firstly finds the projection Q' of R along the m-geodesic then connects Q' and P by an e-geodesic. This property makes the KL divergence asymmetric. Here we are interested in the distance between two distributions on the model manifold of curved exponential families. Therefore the Riemannian distance under the connection in terms of true parameters θ is more appropriate. In most conditions when evaluating the Riemannian distance on the manifold is cumbersome, the KL divergence is a reasonable substitute for the distance between distributions.

The Riemannian distance between the current estimator and the new estimator is a random variable because the next output value has not sampled yet. Therefore the true loss function is the expected Riemannian distance over potential output values. There are two possible ways of evaluating the expected loss. A local expectation fixes previous data $D_n=\{(\mathbf{x_1},\mathbf{y_1}),\cdots,(\mathbf{x_n},\mathbf{y_n})\}$ and varies the next output $\mathbf{y_{n+1}}$ according to $\hat{\theta}(D_n)$ and $\mathbf{x_{n+1}}$ when performing parameter estimation at the next step:

$$E_{\tilde{\mathbf{y}}_\mathbf{n+1}\sim p(\mathbf{y};\mathbf{x_{n+1}},\hat{\theta}(\mathbf{D_n}))}\{D(p(\mathbf{y};\mathbf{x},\hat{\theta}(\mathbf{D_n}))\|p(\mathbf{y};\mathbf{x},\hat{\theta}(\mathbf{D_n},(\mathbf{x_{n+1}},\tilde{\mathbf{y}}_\mathbf{n+1}))))\}. \tag{11}$$

A global expectation varies all the output values up to step $n+1$ when performing parameter estimation at the next step:

$$E_{\tilde{\mathbf{y}}_\mathbf{1},\cdots,\tilde{\mathbf{y}}_\mathbf{n+1}\sim p(\mathbf{y};\mathbf{x_1},\cdots,\mathbf{x_{n+1}},\hat{\theta}(\mathbf{D_n}))}\{D(p(\mathbf{y};\mathbf{x},\hat{\theta}(\mathbf{D_n}))\|p(\mathbf{y};\mathbf{x},\hat{\theta}((\mathbf{x_1},\tilde{\mathbf{y}}_\mathbf{1}),\cdots,(\mathbf{x_{n+1}},\tilde{\mathbf{y}}_\mathbf{n+1}))))\}. \tag{12}$$

In both scenarios, the output values are assumed to be generated from the distribution with parameters $\hat{\theta}(D_n)$.

While the local expectation is much easier to compute, it has an inherent problem. The myopic nature of the learning procedure makes it minimizes (in expectation) the distance between estimated parameter and previous parameters. Since all previous input-output pairs are fixed, the only possibility to make the estimated parameter at the next step differ from the current estimate is the fluctuation of y_{n+1}. However, as the number of data points grow, a single data point becomes immaterial. Empirically I found the local expectation scenario ends up sampling the same input over and over after a few steps. On the contrary, this problem is less serious in global expectation since we "pretend" to estimate the parameter at the next step using the regenerated samples.

The expected distance in equation 12 is a function of input $\mathbf{x}$. This is undesirable because we can enlarge or reduce the distance between two conditional densities by varying the input values even if their parameters are fixed. This discrepancy is due to the fact that $D(p(\mathbf{y};\mathbf{x},\theta_{\mathbf{1}})\|p(\mathbf{y};\mathbf{x},\theta_{\mathbf{2}}))$ is the Riemannian distance on the manifold $M(\mathbf{x})$ which depends on the input values. To resolve this problem the true loss function is the expected loss in equation 12 over input values.

$$\mathcal{L}(\theta_{\mathbf{1}},\theta_{\mathbf{2}}) = E_{\mathbf{x}\sim q(\mathbf{x})}\{D(p(\mathbf{y};\mathbf{x},\theta_{\mathbf{1}})\|p(\mathbf{y};\mathbf{x},\theta_{\mathbf{2}}))\},$$

where $q(\mathbf{x})$ is the empirical distribution of input $\mathbf{x}$. The sampling procedure in an active learning scheme distorts the input distribution, thus $q(\mathbf{x})$ can only be obtained either from the observations independent of the active sampling or be arbitrarily determined (for instance, by setting it uniformly distributed).

To sum up the active learning scheme can be expressed as the following optimization equation:

$$\hat{\mathbf{x}}_{\mathbf{n+1}} = \arg\min_{\mathbf{x}_{\mathbf{n+1}}} E_{\mathbf{x}\sim q(\mathbf{x})}\{E_{\tilde{\mathbf{Y}}\sim p(\mathbf{Y};\mathbf{X},\hat{\theta}(\mathbf{D_n}))}\{D(p(\mathbf{y};\mathbf{x},\hat{\theta}(\mathbf{D_n}))\|p(\mathbf{y};\mathbf{x},\hat{\theta}(\tilde{\mathbf{D}}_{\mathbf{n+1}})))\}\}, \tag{13}$$

where $\mathbf{X} = (\mathbf{x_1},\cdots,\mathbf{x_{n+1}})$, $\tilde{\mathbf{Y}} = (\tilde{\mathbf{y}}_{\mathbf{1}},\cdots,\tilde{\mathbf{y}}_{\mathbf{n+1}})$, and $\tilde{D}_{n+1} = \{(\mathbf{x_1},\tilde{\mathbf{y}}_{\mathbf{1}}),\cdots,(\mathbf{x_{n+1}},\tilde{\mathbf{y}}_{\mathbf{n+1}})\}$.

5 Examples

In this section I use two examples to illustrate the new active learning scheme.

5.1 Nonlinear regression

The first example is nonlinear regression with Gaussian noise. Assume the scalar output variable y is a nonlinear function of a vector input variable x plus a Gaussian noise e:

$$y = \sum_{i=1}^{r} \theta_i f_i(\mathbf{x}) + e,$$

where $e \sim N(0,\sigma^2)$ and σ is known. f_is comprise basis functions used to model y, for instance, polynomials. Here the index notation is a little abused such that tensor indices (superscript and subscript), example indices (subscript) and iteration steps (subscript) are mixed.

The distribution of y is Gaussian parametrized by θ and $\mathbf{x}$:

$$p(y;\mathbf{x},\theta) = \frac{1}{\sqrt{2\pi\sigma^2}}\exp\{\frac{-1}{2\sigma^2}(y-\theta\cdot\mathbf{f}(x))^2\}.$$

To save space I write the sufficient statistics and the parameters in vector forms. The sufficient statistic, natural parameter, and partition function of y are

$$T(y) = (y, y^2)$$

$$\zeta = (\frac{\theta\cdot\mathbf{f}(x)}{\sigma^2}, \frac{-1}{2\sigma^2}).$$

$$\psi(\zeta) = \frac{-1}{4}\zeta_1^2\zeta_2^{-1} - \frac{1}{2}\log(-\zeta_2) + \frac{1}{2}\log\pi.$$

The Fisher information matrix in terms of ζ is

$$g_{ij}(\zeta) = \begin{pmatrix} \frac{-1}{2\zeta_2} & \frac{\zeta_1}{2\zeta_2^2} \\ \frac{\zeta_1}{2\zeta_2^2} & \frac{1-\zeta_1^2}{2\zeta_2^3} \end{pmatrix}.$$

By applying theorem 1, the Fisher information matrix in terms of θ is

$$g_{ij}(\mathbf{x},\theta) = \frac{1}{\sigma^2}f_i(\mathbf{x})f_j(\mathbf{x}),$$

which is a pure effect of $\mathbf{x}$. Hence the metric tensor is a constant when the inputs are fixed. The differential Riemannian distance is

$$ds^2 = g_{ij}d\theta^i d\theta^j = \frac{1}{\sigma^2}f_i(\mathbf{x})f_j(\mathbf{x})d\theta^i d\theta^j.$$

Plugging it into equation 10, the (square) curve length along the geodesic $\theta(t) = \theta_0 + t(\theta_1 - \theta_0)$ becomes

$$D^2(\theta_0,\theta_1) = \frac{1}{\sigma^2}f_i(\mathbf{x})f_j(\mathbf{x})(\theta_1^i-\theta_0^i)(\theta_1^j-\theta_0^j) = \frac{1}{\sigma^2}\mathbf{f^T}(\mathbf{x})(\theta_\mathbf{1}-\theta_\mathbf{0})(\theta_\mathbf{1}-\theta_\mathbf{0})^\mathbf{T}\mathbf{f}(\mathbf{x}),$$

where the second equation is written in the matrix form. It can be easily verified that this is the KL divergence of conditional Gaussian distributions.

Let

$$M_n = \sum_{t=1}^{n} f(\mathbf{x_t})f^T(\mathbf{x_t}), Y_n = \sum_{t=1}^{n} y_t f(\mathbf{x_t})$$

be obtained from input-output pairs D_n up to step n. The maximum likelihood estimator is

$$\hat{\theta}_n = \hat{\theta}(D_n) = M_n^{-1}Y_n.$$

Assume $\tilde{D}_{n+1} = \{(\mathbf{x_1}, \tilde{y}_1), \cdots, (\mathbf{x_{n+1}}, \tilde{y}_{n+1})\}$ are the *virtual* data resampled from the distribution with parameter $\hat{\theta}_n$. Then the maximum likelihood estimator at the next step $\hat{\theta}_{n+1} = \hat{\theta}(\tilde{D}_{n+1})$ satisfies the following conditions ([7]):

$$\begin{aligned} &E\{\hat{\theta}_{n+1}\} = \hat{\theta}_n. \\ &V\{\hat{\theta}_{n+1}\} = \sigma^2 M_{n+1}^{-1}. \end{aligned}$$

The expected Riemannian distance between $\hat{\theta}_n$ and $\hat{\theta}_{n+1}$ thus becomes

$$\begin{aligned} E_{\tilde{y}_1,\cdots,\tilde{y}_{n+1}}\{D(p(y;\mathbf{x},\hat{\theta}_n)\|p(y;\mathbf{x},\hat{\theta}_{n+1}))\} &= \tfrac{1}{2\sigma^2} E\{f^T(\mathbf{x})(\hat{\theta}_{n+1}-\hat{\theta}_n)(\hat{\theta}_{n+1}-\hat{\theta}_n)^T f(\mathbf{x})\} \\ &= \tfrac{1}{2\sigma^2} f^T(\mathbf{x}) V(\hat{\theta}_{n+1}) f(\mathbf{x}) \\ &= \tfrac{1}{2} f^T(\mathbf{x}) M_{n+1}^{-1} f(\mathbf{x}). \end{aligned}$$

The learning criteria becomes

$$\mathbf{\hat{x}_{n+1}} = \arg\min_{\mathbf{x_{n+1}}} \int q(\mathbf{x}) f^T(\mathbf{x})(M_n + f(\mathbf{x_{n+1}}) f^T(\mathbf{x_{n+1}}))^{-1} f(\mathbf{x}) d\mathbf{x}. \tag{14}$$

Figure 1 shows the experiment results on the regression of the function $y = \theta^0 + \theta^1 sin(\frac{\pi}{6}x) + \theta^2 sin(\frac{\pi}{4}x^2) + \theta^3 sin(\frac{\pi}{3}x^3) + e$, where θ =[15 -13 -3 1]^T and $\sigma = 4$. The average square error of the estimated parameters at each iteration over 500 experiments is plotted. The initial dataset D_0 contains 5 random samples, and the learning curve of the 5 initial samples is not plotted. The results clearly indicate active learning schemes outperform passive learning when a small number of samples are allowed to draw from the unknown distribution. As the size of the data grows, their difference tends to decrease. I also compare the difference in terms of the loss function in active learning. The Riemannian distance loss function (the solid curve) performs slightly better than the trace of the dispersion matrix (the dash-dot curve), although the difference is not as significant as the difference between active and passive learning schemes.

5.2 Logistic regression

Logistic regression is a standard distribution of modeling the influence of continuous inputs on discrete outputs. For simplicity here I only discuss the case of binary variables. Suppose y is a binary random variable which is affected by continuous input variables $\mathbf{x}$. The conditional probability mass function of y can be expressed as

$$p(y;\mathbf{x},\theta) = exp\{\mathbf{f^T}(\mathbf{x}) \cdot \theta\delta(y = \mathit{1}) - \log(1 + e^{\mathbf{f^T}(\mathbf{x})\cdot\theta})\}.$$

By treating $\zeta = \mathbf{f^T}(\mathbf{x}) \cdot \theta$ as the natural parameter of the exponential family, the metric tensor in terms of the true parameters θ can be obtained by coordinate transformation:

$$g_{ab}(\theta,\mathbf{x}) = \frac{\partial\zeta}{\partial\theta^a}\frac{\partial\zeta}{\partial\theta^b} g_{\mathit{11}}(\zeta) = f_a(\mathbf{x}) f_b(\mathbf{x}) \frac{\partial^2\psi(\zeta)}{\partial\zeta^2} = f_a(\mathbf{x}) f_b(\mathbf{x}) \frac{e^{\zeta(\mathbf{x},\theta)}}{(1 + e^{\zeta(\mathbf{x},\theta)})^2}.$$

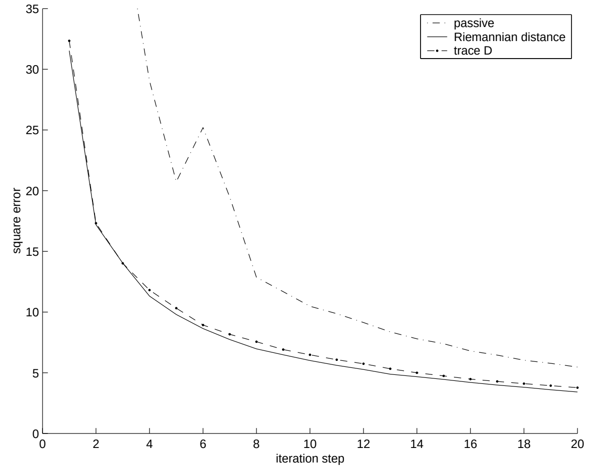

Fig. 1. Active and passive learnings on nonlinear regression

Notice $\frac{e^{\zeta(\mathbf{x},\theta)}}{(1+e^{\zeta(\mathbf{x},\theta)})^2}$ is a symmetric function of θ. The Riemannian distance between two parameters θ_0 and θ_1 can be computed from equation 10:

$$\begin{aligned} D^2(\theta_0,\theta_1) &= (\textstyle\int_0^1 \sqrt{g_{ij}(\theta(t),x)(\theta_1-\theta_0)^i(\theta_1-\theta_0)^j}dt)^2 \\ &= (\textstyle\int_0^1 \sqrt{\frac{e^{\zeta(\theta(t),\mathbf{x})}}{(1+e^{\zeta(\theta(t),\mathbf{x})})^2}[\mathbf{f^T}(\mathbf{x})(\theta_1-\theta_0)(\theta_1-\theta_0)^T\mathbf{f}(\mathbf{x})]}dt)^2 \\ &= (\textstyle\int_0^1 \sqrt{\frac{e^{\zeta(\theta(t),\mathbf{x})}}{(1+e^{\zeta(\theta(t),\mathbf{x})})^2}}dt)^2[\mathbf{f^T}(\mathbf{x})(\theta_1-\theta_0)(\theta_1-\theta_0)^T\mathbf{f}(\mathbf{x})] \\ &= (\tfrac{2}{a}[arctan(e^{\frac{a+b}{2}})-arctan(e^{\frac{b}{2}})])^2[\mathbf{f^T}(\mathbf{x})(\theta_1-\theta_0)(\theta_1-\theta_0)^T\mathbf{f}(\mathbf{x})], \end{aligned}$$

where $a=\sum_{i=1}^r \mathbf{f}(\mathbf{x})_\mathbf{i}(\theta_1^i-\theta_0^i)$ and $b=\sum_{i=1}^r \mathbf{f}(\mathbf{x})_\mathbf{i}\theta_0^i$. The key for simplification is because the metric tensor is a symmetric function of $\theta(t)$, hence t does not appear in individual components. If y takes more than two values, then the square distance has a complicated form.

Although the distance function is considerably simplified, the active learning of logistic regression is still cumbersome. Unlike nonlinear regression with Gaussian noise, there is no analytic solution for maximum likelihood estimators in logistic regression. It is usually obtained by numerical or approximation methods such as gradient descent, Newton's method or variational methods ([9]). Therefore, the expectation of the square distance between current estimate and the next estimate over possible outputs can only be computed by approximation or sampling. Due to the lack of time the numerical experiment for logistic regression is left for future works.

6 Conclusion

In this paper I propose an active learning scheme from the perspective of information geometry. The deviation between two distributions is measured by the

Riemannian distance on the model manifold. The model manifold of exponential families is dually flat. Moreover, for the distributions whose log densities are linear in terms of parameters, the embedding curvature of their manifolds in terms of the true coordinate systems also vanishes. The active learning loss function is the expected Riemannian distance over the input and the output data (equation 13). This scheme is illustrated by two examples: nonlinear regression and logistic regression.

There are abundant future works to be pursued. The active learning scheme is computationally intensive. More efficient algorithms for evaluating expected loss function need to be developed. Secondly, the Bayesian approach for parameter estimation is not yet incorporated into the framework. Moreover, for the model manifolds which are not dually flat, the KL divergence is no longer proportional to the Riemannian distance. How to evaluate the Riemannian distance efficient on a curved manifold needs to be studied.

Acknowledgement

The author would like to thank professor Tommi Jaakkola for the discussion about information geometry and Jason Rennie for reviewing and commenting on the paper.

References

1. Amari, S.I. (2001). *Information geometry of hierarchy of probability distributions*, *IEEE transactions on information theory*, *47:50*, 1701-1711.
2. Amari, S.I. (1996). *Information geometry of neural networks – a new Bayesian duality theory*, *International conference on neural information processing.*
3. Amari, S.I. (1995). *Information geometry of the EM and em algorithms for neural networks. Neural networks*, *9*, 1379-1408.
4. Amari, S.I. (1985). *Differential geometrical methods in statistics*, Springer Lecture Notes in Statistics, 28, Springer.
5. Amari, S.I. (1982). *Differential geometry of curved exponential families – curvatures and information loss. Annals of statistics*, *10:2*, 357-385.
6. Csiszár, I. and Tusnády, G. (1984). *Information geometry and alternating minimization procedures*, *Statistics & Decisions, Supplement Issue*, *1*, 205-237.
7. Fedorov, V.V. (1972). *Theory of optimal experiments.* New York: Academic Press.
8. MacKay, D.J.C. (1992). *Information-based objective functions for active data selection. Neural computation*, *4*, 589-603.
9. Minka, T.P. (2001). *Algorithms for maximum-likelihood logistic regression. CMU Statistics Technical Report 758.*
10. Sokolnikoff, I.S. (1964). *Tensor analysis.* New York:John Wiley & Sons.
11. Sung, K.K. and Niyogi, P. (1995). *Active learning for function approximation. Advances in neural information processing systems*, *7*, 593-600.
12. Tong, S. and Koller, D. (2001). *Active learning for structure in Bayesian networks. International joint conference on artificial intelligence.*
13. Tong, S. and Koller, D. (2000). *Active learning for parameter estimation in Bayesian networks. Advances in neural information processing systems*, *13*, 647-653.