# Natural Gradients, Mahalanobis Distances, and Distances between Distributions

Mathias Winther Madsen
mathias@micropsi-industries.com

June 29, 2016

This document defines and discusses the concept of natural gradients.

A gradient is always a vector pointing in the direction of maximal output change per input change, but a natural gradient does so under some "natural" measure of distance in the input space. This note discusses how to compute such directions when the metric on the input space deviates substantially from the conventional Euclidean distance.

## Contents

# 1 Natural Gradients

This introductory section rehearses the concept of directional derivatives and introduces the problem of finding directions of maximum growth per distance.

## 1.1 Directional Derivatives in Euclidean Space

Suppose we would like to know how fast the function $f : \mathbb{R}^n \to \mathbb{R}$ grows at some point $\mathbf{x}$ when we move in the direction of a vector $\mathbf{u}$.

If the function is differentiable at $\mathbf{x}$, its local behavior is well described by the first-order approximation

$$f(\mathbf{x} + \mathbf{u}) \approx f(\mathbf{x}) + \mathbf{u}^T \, \nabla f(\mathbf{x}).$$

Since the dot product $\mathbf{u}^T \, \nabla f(\mathbf{x})$ measures how parallel $\mathbf{u}$ and $\nabla f(\mathbf{x})$ are, this approximation essentially postulates that the only way to achieve growth in $f$ is to follow the gradient.

Using this first-order expansion, we can find the directional derivative of $f$ in an arbitrary direction $\mathbf{u}$ to be

$$\begin{aligned}
\frac{\partial}{\partial \mathbf{u}} f(\mathbf{x}) &= \lim_{\varepsilon \to 0} \frac{f(\mathbf{x} + \varepsilon \mathbf{u}) - f(\mathbf{x})}{\|\varepsilon \mathbf{u}\|} \\
&= \lim_{\varepsilon \to 0} \frac{\varepsilon \mathbf{u}^T \, \nabla f(\mathbf{x})}{\varepsilon \, \|\mathbf{u}\|} \\
&= \frac{\mathbf{u}^T}{\|\mathbf{u}\|} \, \nabla f(\mathbf{x}).
\end{aligned}$$

Again, changes in the direction of $\mathbf{u}$ thus matter to the extent that $\mathbf{u}$ is parallel to the gradient: when they are perfectly parallel, the function grows by an increment of $\|\nabla f(\mathbf{x})\|$ every time we change $\mathbf{x}$ by a normalized increment of $\mathbf{u} / \, \|\mathbf{u}\|$, and this is the fastest growth we can attain.

## 1.2 Directional Derivatives under Weird Distances

Suppose now that we impose some exotic distance measure on the domain of $f$, in the form of a norm $\| \cdot \|_{\mathbf{\Sigma}}$.

In such a context, we might want to know what direction we could walk in in order to trade a $\| \cdot \|_{\mathbf{\Sigma}}$-small change in $\mathbf{x}$ for a huge change in $f(\mathbf{x})$. In other words, we might be interested in maximizing the distance-sensitive directional derivative

$$\left( \frac{\partial}{\partial_{\mathbf{\Sigma}} \mathbf{u}} \right) f(\mathbf{x}) = \lim_{\varepsilon \to 0} \frac{\varepsilon \mathbf{u}^T \, \nabla f(\mathbf{x})}{\varepsilon \, \|\mathbf{u}\|_{\mathbf{\Sigma}}} = \frac{\mathbf{u}^T}{\|\mathbf{u}\|_{\mathbf{\Sigma}}} \, \nabla f(\mathbf{x}).$$

This is once again a dot product of a unit vector with a gradient, but now “unit vector” means a vector with $\| \cdot \|_{\mathbf{\Sigma}}$-norm 1, as opposed to a Euclidean norm of one. Whereas the usual derivative measures the amount of $f$ we get per

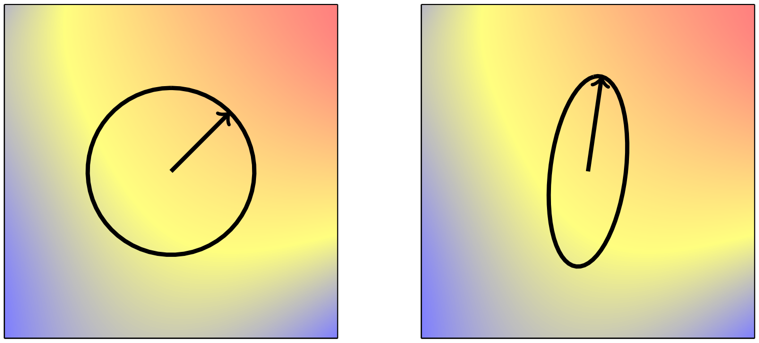

Figure 1: When you are allowed to step a distance of $\|\Delta\mathbf{x}\| = \varepsilon$ in any direction, the gradient points in the direction that gives you the largest increase $\Delta f(\mathbf{x})$ per unit step. However, if you redefine the meaning of a unit step, the optimal direction may be a different one.

Euclidean unit of $\mathbf{u}$, this modified derivative measures the amount of $f$ we get per $\|\cdot\|_{\mathbf{\Sigma}}$-unit $\mathbf{u}$.

In order to find the $\mathbf{u}$ for which this relative gain is maximized, we need to solve the constrained maximization problem

$$\max_{\mathbf{u}} \mathbf{u}^T \, \nabla f(\mathbf{x}) \quad s.t. \quad \|\mathbf{u}\|_{\mathbf{\Sigma}}^2 = 1.$$

By the Lagrange method, we find that the solution to this problem satisfies

$$\nabla_u \left( \mathbf{u}^T \, \nabla_{\mathbf{x}} f(\mathbf{x}) \, - \, \lambda \, \|\mathbf{u}\|_{\mathbf{\Sigma}}^2 \right) = 0$$

for some stretching factor $\lambda$. The solution vector $\mathbf{u}$ thus satisfies

$$\nabla_{\mathbf{x}} f(\mathbf{x}) = \lambda \, \nabla_{\mathbf{u}} \, \|\mathbf{u}\|_{\mathbf{\Sigma}}^2 \, .$$

The solution to this problem is the **natural gradient** of $f$ with respect to the norm $\|\cdot\|_{\mathbf{\Sigma}}$. At the moment, however, we cannot reduce this expression any further.

## 2 Mahalanobis Distances

In this section, I introduce a large class of vector norms which is well-suited to the describe the local behavior of many distance metrics. I then discuss how to compute natural gradients under this class of norms.

### 2.1 Paper Distances vs. Vector Norms

Suppose you take a pencil and piece of graph paper and draw the vector

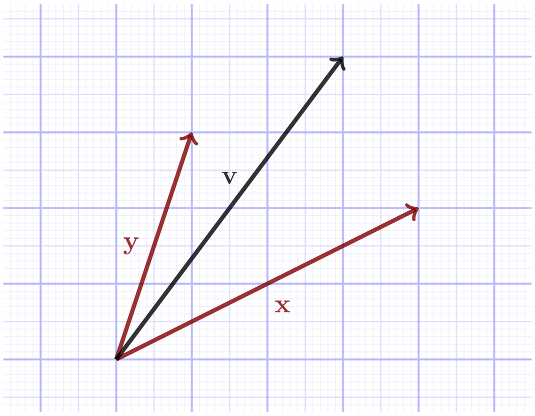

Figure 2: The graph paper example discussed in the text.

$$\mathbf{v} = \begin{pmatrix} 3\,\text{cm} \\ 4\,\text{cm} \end{pmatrix}.$$

In Euclidean terms, the squared norm of this vector is $\|\mathbf{v}\|^2 = 25\,\text{cm}^2$, so the actual pencil line is 5 cm long.

Suppose now that I draw a strange coordinate system on your paper (cf. Fig. 2), using the unit vectors

$$\mathbf{x} = \begin{pmatrix} 4\,\text{cm} \\ 2\,\text{cm} \end{pmatrix}, \quad \mathbf{y} = \begin{pmatrix} 1\,\text{cm} \\ 3\,\text{cm} \end{pmatrix}.$$

In this coordinate system, an abstract vector

$$\mathbf{w} = \begin{pmatrix} 3 \\ 2 \end{pmatrix}$$

should be drawn at the paper coordinates

$$\mathbf{Mw} = \begin{pmatrix} 4\,\text{cm} & 1\,\text{cm} \\ 2\,\text{cm} & 3\,\text{cm} \end{pmatrix} \begin{pmatrix} 3 \\ 2 \end{pmatrix} = \begin{pmatrix} 14\,\text{cm} \\ 12\,\text{cm} \end{pmatrix},$$

whereas your paper vector $\mathbf{v}$ corresponds to the abstract coordinates

$$\mathbf{M}^{-1}\mathbf{v} = \begin{pmatrix} 0.3\,\text{cm}^{-1} & -0.1\,\text{cm}^{-1} \\ -0.2\,\text{cm}^{-1} & 0.4\,\text{cm}^{-1} \end{pmatrix} \begin{pmatrix} 3\,\text{cm} \\ 4\,\text{cm} \end{pmatrix} = \begin{pmatrix} 0.5 \\ 1.0 \end{pmatrix}.$$

Relative to my strange coordinate system, the squared norm of your vector is thus

$$\left\|\mathbf{M}^{-1}\mathbf{v}\right\|^2 = 0.50^2 + 1.00^2 = 1.25.$$

In this coordinate system, which had quite long unit vectors, distances are thus substantially smaller than the Euclidean paper distances.

## 2.2 Mahalanobis Distance

This squared norm of a back-translated vector $\mathbf{M}^{-1}\mathbf{v}$ can also be computed as a dot product,

$$\begin{aligned}\left\|\mathbf{M}^{-1}\mathbf{v}\right\|^2 &= (\mathbf{M}^{-1}\mathbf{v})^T(\mathbf{M}^{-1}\mathbf{v}) \\ &= \mathbf{v}^T(\mathbf{M}^{-1})^T\mathbf{M}^{-1}\mathbf{v} \\ &= \mathbf{v}^T(\mathbf{M}\mathbf{M}^T)^{-1}\mathbf{v}.\end{aligned}$$

The distance measure defined by this vector norm is called the **Mahalanobis distance** associated with the covariance matrix $\mathbf{\Sigma} = \mathbf{M}\mathbf{M}^T$:

$$\|\mathbf{v}\|_{\mathbf{\Sigma}}^2 = \mathbf{v}^T\mathbf{\Sigma}^{-1}\mathbf{v}.$$

Since the ordinary Euclidean norm satisfies $\|\mathbf{v}\|^2 = \mathbf{v}^T\mathbf{v}$, this is a special case of the Mahalanobis distance, arising when $\mathbf{\Sigma} = \mathbf{M} = \mathbf{I}$.

## 2.3 The Gradient of the Mahalanobis Distance

Let's compute the gradient of the Mahalanobis norm with respect to the vector being measured. This gradient is comprised of a list of derivatives of the form

$$\frac{\partial}{\partial\mathbf{v}_n}\|\mathbf{v}\|_{\mathbf{\Sigma}}^2 = \frac{\partial}{\partial\mathbf{v}_n}\mathbf{v}^T\mathbf{\Sigma}^{-1}\mathbf{v} = \frac{\partial}{\partial\mathbf{v}_n}\sum_i\sum_j \mathbf{v}_i\mathbf{\Sigma}_{ij}^{-1}\mathbf{v}_j.$$

If we think about this double sum in terms of a big square table, we find that the vector entry $\mathbf{v}_n$ shows up in one row and one column of this table. Correcting for the double counting of the overlap of the two we thus find that

$$\sum_i\sum_j \mathbf{\Sigma}_{ij}^{-1}\mathbf{v}_i\mathbf{v}_j = \underbrace{\sum_i \mathbf{\Sigma}_{in}^{-1}\mathbf{v}_i\mathbf{v}_n}_{\text{column } n} + \underbrace{\sum_j \mathbf{\Sigma}_{nj}^{-1}\mathbf{v}_n\mathbf{v}_j}_{\text{row } n} - \underbrace{\mathbf{\Sigma}_{nn}^{-1}\mathbf{v}_n\mathbf{v}_n}_{\text{cell } (n,n)}.$$

However, $\mathbf{\Sigma}$ and therefore $\mathbf{\Sigma}^{-1}$ are symmetric matrices, so the derivative of this double sum can be simplified to

$$\frac{\partial}{\partial\mathbf{v}_n}\left(\sum_i \mathbf{\Sigma}_{in}^{-1}\mathbf{v}_i\mathbf{v}_n + \sum_j \mathbf{\Sigma}_{nj}^{-1}\mathbf{v}_n\mathbf{v}_j - \mathbf{\Sigma}_{nn}^{-1}\mathbf{v}_n\mathbf{v}_n\right) = 2\sum_j \mathbf{\Sigma}_{nj}^{-1}\mathbf{v}_i.$$

This is the dot product of $\mathbf{v}$ and the $n$th column of $\mathbf{\Sigma}^{-1}$, times two. Since the result of a matrix multiplication is a vector of such dot products,

$$\nabla_{\mathbf{v}}\|\mathbf{v}\|_{\mathbf{\Sigma}}^2 = 2\,\mathbf{\Sigma}^{-1}\mathbf{v}.$$

In particular, when $\mathbf{\Sigma} = \mathbf{I}$, the gradient is twice the vector itself.

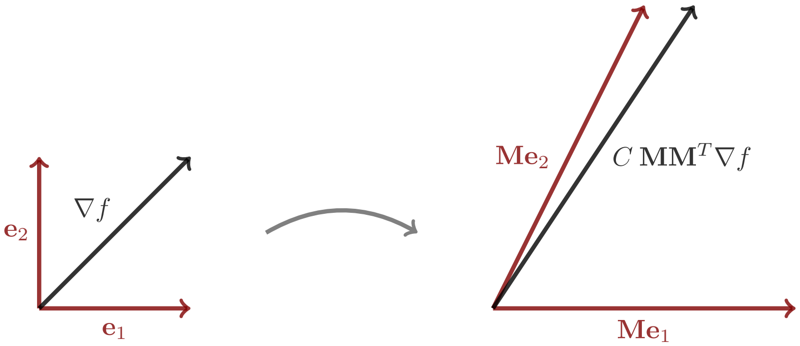

Figure 3: The relationship between the directions of maximal growth per distance, using either the warped or unwarped distance measure. $C$ is a scaling constant which can be chosen such that $C\,\mathbf{M}\mathbf{M}^T\nabla f$ is a unit vector.

### 2.4 Mahalanobis-Natural Gradients

Consider again the problem of determining the natural gradient of a function $f$ with respect to the Mahalanobis norm $\|\cdot\|_{\boldsymbol{\Sigma}}$ on the input space. We have already seen that the natural gradient $\mathbf{u}$ satisfies

$$\nabla_{\mathbf{x}} f(\mathbf{x}) = \lambda \nabla_{\mathbf{u}} \|\mathbf{u}\|_{\boldsymbol{\Sigma}}^2 ,$$

and the gradient of the Mahalanobis distance is $\nabla_{\mathbf{u}} \|\mathbf{u}\|_{\boldsymbol{\Sigma}}^2 = 2\boldsymbol{\Sigma}^{-1}\mathbf{u}$. It follows that the natural gradient under Mahalanobis distances satisfies

$$\nabla_{\mathbf{x}} f(\mathbf{x}) = 2\lambda\boldsymbol{\Sigma}^{-1}\mathbf{u},$$

which is to say that

$$\mathbf{u} = \frac{1}{2\lambda}\boldsymbol{\Sigma}\nabla_{\mathbf{x}} f(\mathbf{x}) \propto \boldsymbol{\Sigma}\nabla_{\mathbf{x}} f(\mathbf{x}).$$

Here, the proportionality constant is there to ensure that this solution is a Mahalanobis-unit vector.

Note also that there is something slightly counterintuitive about this result: even though the Euclidean and the warped input spaces are related by a linear transformation $\mathbf{M}$, and $\nabla f$ is the direction of fastest growth in the warped space $\mathbf{M}^{-1}\mathbb{R}^n$, the direction of maximal growth per Mahalanobis-unit is not $\mathbf{M}\nabla f$. Rather, it is

$$\boldsymbol{\Sigma}\nabla f = \mathbf{M}\mathbf{M}^T\nabla f.$$

This is, roughly speaking, a consequence of the fact that we are looking for the maximum of a linear function on a ellipse. This effect will hopefully be more clear in the following example.

## 2.5 An Explicit Example

Suppose that the unit vectors of a strange coordinate system are given as the columns of the matrix

$$\mathbf{M} = \begin{pmatrix} 2 & 1 \\ 0 & 2 \end{pmatrix}.$$

The covariance matrix associated with this linear transformation is

$$\boldsymbol{\Sigma} = \mathbf{M}\mathbf{M}^T = \begin{pmatrix} 2 & 1 \\ 0 & 2 \end{pmatrix} \begin{pmatrix} 2 & 0 \\ 1 & 2 \end{pmatrix} = \begin{pmatrix} 5 & 2 \\ 2 & 4 \end{pmatrix}.$$

This implies that

$$\boldsymbol{\Sigma}^{-1} = \frac{1}{16} \begin{pmatrix} 4 & -2 \\ -2 & 5 \end{pmatrix}.$$

The squared Mahalanobis-norm $\|\mathbf{v}\|^2_{\boldsymbol{\Sigma}} = \mathbf{v}^T\boldsymbol{\Sigma}^{-1}\mathbf{v}$ of a generic vector $\mathbf{v} = (x, y)^T$ is therefore

$$\|\mathbf{v}\|^2_{\boldsymbol{\Sigma}} = \frac{1}{16} \begin{pmatrix} x \\ y \end{pmatrix}^T \begin{pmatrix} 4 & -2 \\ -2 & 5 \end{pmatrix} \begin{pmatrix} x \\ y \end{pmatrix} = \frac{1}{16} \left(4x^2 - 4xy + 5y^2\right).$$

Ignoring multiplicative constants, the gradient of this norm with respect to the input vector is equal to

$$\nabla_{\mathbf{v}} \|\mathbf{v}\|^2_{\boldsymbol{\Sigma}} \propto \begin{pmatrix} 4x - 2y \\ 5y - 2x \end{pmatrix}.$$

Let's now find the gradient as well as the natural gradient of the function

$$f \begin{pmatrix} x \\ y \end{pmatrix} = 3y.$$

Clearly, the ordinary gradient of this function is

$$\nabla f \begin{pmatrix} x \\ y \end{pmatrix} = \begin{pmatrix} 0 \\ 3 \end{pmatrix}$$

at all points of evaluation. According to the computation in the previous section, the natural gradient of $f$ is parallel to

$$\boldsymbol{\Sigma}\nabla f = \begin{pmatrix} 5 & 2 \\ 2 & 4 \end{pmatrix} \begin{pmatrix} 0 \\ 3 \end{pmatrix} = \begin{pmatrix} 6 \\ 12 \end{pmatrix} \propto \begin{pmatrix} 1 \\ 2 \end{pmatrix}.$$

This vector too is the same at any point of evaluation.

This result can also be confirmed by a direct computation. Since $f$ only grows in the $y$-direction, we get the most $f$ for our money when we move to the highest point on the $\| \cdot \|_{\mathbf{\Sigma}}$-circle around the point of evaluation. Since a $\| \cdot \|_{\mathbf{\Sigma}}$-circle around the origin is in fact an ellipse,

$$4x^2 \,-\, 4xy \,+\, 5y^2 \;=\; 16,$$

we can write $y$ as a function of $x$ as follows:

$$y(x) \;=\; \frac{2}{5}\left(2\sqrt{5 - x^2} + x\right).$$

The maximum of this function is $y(1) \,=\, 2$, confirming the natural gradient of

$$\mathbf{\Sigma}\nabla f \;\propto\; \begin{pmatrix} 1 \\ 2 \end{pmatrix}.$$

This vector is also, in fact, a Mahalanobis-unit vector, since

$$\frac{1}{16}\left(4 \cdot 1^2 \,-\, 4 \cdot 1 \cdot 2 + 5 \cdot 2^2\right) \;=\; 1.$$

# 3 Kullback-Leibler Divergence

In this final section, I discuss how to apply the results about natural gradients to distances between distributions. In order to help intuitions along, I start in the one-dimensional case and then scale up.

## 3.1 Definition and Taylor Expansion

The Kullback-Leibler divergence from on probability density to another is the expectation

$$KL(p \,||\, q) \;=\; E_p\left[\ln\frac{p}{q}\right] \;=\; \int p(\omega)\,\ln\frac{p(\omega)}{q(\omega)}\,d\omega.$$

This statistic measures the speed with which evidence accumulates in favor of $p$ when we are trying to distinguish between $p$ and $q$ in a stochastic environment governed by $p$.

The Taylor expansion of the function

$$\ln\frac{p}{q} \;=\; -\ln\left(1 + \frac{q-p}{p}\right)$$

around the expansion point $q \,=\, p$ is

$$\ln\frac{p}{q} \approx \left(\frac{p-q}{p}\right) + \frac{1}{2}\left(\frac{p-q}{p}\right)^2 + \frac{1}{3}\left(\frac{p-q}{p}\right)^3 + \frac{1}{4}\left(\frac{p-q}{p}\right)^4 + \cdots$$

This provides an approximation of the log-likelihood ratio $\ln p(\omega) \,-\, \ln q(\omega)$ at any specific sample point $\omega$. We can use this approximation to estimate the

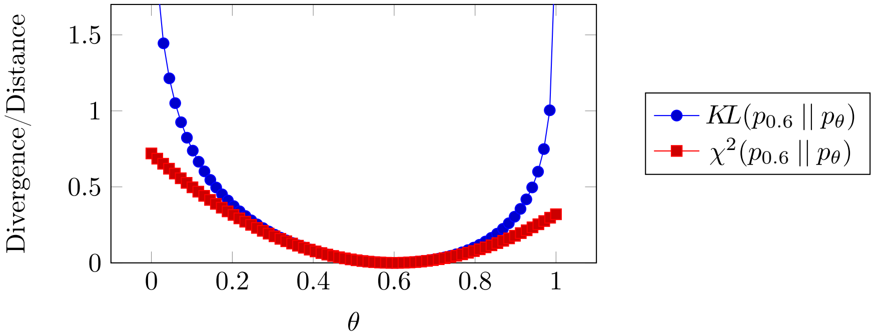

Figure 4: The distance between two coin-flipping distributions, one with parameter $\theta_0 = 0.6$, and one with a free parameter $\theta$. For $\theta \approx \theta_0$, the Kullback-Leibler divergence is approximated well by the $\chi^2$ distance.

Kullback-Leibler divergence between two closely related distibutions by taking the expectation of this Taylor series:

$$E_p\left[\ln\frac{p}{q}\right] \approx E_p\left[\left(\frac{p-q}{p}\right)\right] + \frac{1}{2}\,E_p\left[\left(\frac{p-q}{p}\right)^2\right] + \frac{1}{3}\,E_p\left[\left(\frac{p-q}{p}\right)^3\right] + \cdots$$

The first of these moments is zero, however, since

$$E_p\left[\frac{p-q}{p}\right] = \int\left(\frac{p-q}{p}\right)p = \int p - \int q = 1 - 1 = 0.$$

Truncating the expansion after the second term, we thus find that

$$KL(p\,||\,q) \approx \frac{1}{2}\,E_p\left[\left(\frac{p-q}{p}\right)^2\right].$$

In expectation, the log-likelihood ratio between the probability distributions with density functions $p$ and $q$ is thus close to (half) the $\chi^2$-distance between them. This also happens to be equal to the $L_2$ distance between the two probability density functions.

## 3.2 Fisher Information and KL Divergence

Let's continue to work with this $\chi^2$ distance, now assuming that $p = p_{\theta_0}$ and $q = p_\theta$ are two members of a parametrized family of densities. A linear extrapolation around $\theta = \theta_0$ then shows us that

$$p_{\theta_0}(\omega) - p_\theta(\omega) \approx \delta\frac{\partial}{\partial\theta}p_{\theta_0}(\omega).$$

This implies that

$$\frac{p_{\theta_0} - p_\theta}{p_{\theta_0}} = \frac{\delta\frac{\partial}{\partial\theta}p_\theta}{p_{\theta_0}} = \delta\frac{\partial}{\partial\theta}\ln p_\theta = \delta S,$$

where $S$ is the **Fisher score** of the density $p_\theta$ with respect to $\theta$ (at some sample point $\omega$ which is here suppressed from the notation). Since the **Fisher information** is the mean square of the score, $F_\theta = E[S^2]$, the Kullback-Leibler divergence is approximated well by

$$KL(p_\theta \,||\, p_{\theta_0}) \;=\; \frac{1}{2}\, E_{\theta_0}\left[(\delta S)^2\right] \;=\; \frac{\delta^2}{2}\, F_\theta.$$

In a context where we are able to compute the sample values of $S(\omega)^2$ but not its expectation, we can thus fall back on the empirical approximation

$$KL(p_\theta \,||\, p_{\theta_0}) \;\approx\; \frac{\delta^2}{2}\, \hat{E}_{\theta_0}\left[S^2\right].$$

In the following, we will see higher-dimensional versions of this result, when $\theta$ is a vector of parameters rather than just a scalar.

## 3.3 High-Dimensional Taylor Expansions

The Hessian matrix of a function $f : \mathbb{R}^n \to \mathbb{R}$ is the matrix that contains its second-order partial derivatives,

$$\mathbf{H}(\mathbf{x}) \;=\; \left[\frac{\partial^2}{\partial x_i \partial x_j} f(\mathbf{x})\right]_{ij}.$$

This matrix can be used to compute the square term in the Taylor expansion of $f$, with the gradient being responsible for the first term:

$$f(\mathbf{x} + \mathbf{u}) \;\approx\; f(\mathbf{x}) + \mathbf{u}^T \nabla f(\mathbf{x}) + \frac{1}{2}\mathbf{u}^T \mathbf{H}(\mathbf{x})\mathbf{u}.$$

As a special case of this Taylor approximation, consider the log-likelihood ratio

$$r(\theta) \;=\; \ln \frac{p_\theta(\omega)}{p_{\theta_0}(\omega)},$$

developed around the expansion point $\theta = \theta_0$. This function has the second-order Taylor approximation

$$r(\theta) \;\approx\; r(\theta_0) + \theta^T\, \nabla_\theta r(\theta_0) + \frac{1}{2}\theta^T \mathbf{H}(\theta_0)\theta,$$

where $\mathbf{H}$ is the Hessian matrix of $r$ with respect to the (high-dimensional) parameter vector $\theta$.

If we can find the expectation of this function at some point $\theta = \theta_1$, we will have an approximation of the Kullback-Leibler divergence between the two high-dimensional densities $p_{\theta_0}$ and $p_{\theta_1}$. In the following, I will compute these expectations one by one.

### 3.4 Taylor Terms of the High-Dimensional KL Divergence

We will now compute the expected value of the value, gradient, and Hessian matrix of a twice differentiable log-likelihood function:

1. The **constant term** $r(\theta_0)$ vanishes at all sample points $\omega$:

$$\ln \frac{p_{\theta_0}}{p_{\theta_0}} \;=\; \ln 1 \;=\; 0.$$

2. In order to evaluate the **first-order term** $\theta^T\, \nabla_\theta r(\theta_0)$, we first need to compute the gradient components

$$\frac{\partial}{\partial\theta_i} \ln \frac{p_\theta}{p_{\theta_0}} \;=\; \frac{\partial}{\partial\theta_i} \ln p_\theta \;-\; \frac{\partial}{\partial\theta_i} \ln p_{\theta_0} \;=\; \frac{1}{p_\theta}\frac{\partial}{\partial\theta_i} p_\theta,$$

which are the relative growth rates of the likelihood along each unit vector. Taking expectations, all of these gradient components vanish (at least in cases where we can interchange of integration and differentiation):

$$E_\theta\left[\frac{1}{L(\theta)}\frac{\partial}{\partial\theta_i} L(\theta)\right] \;=\; \int \frac{1}{p_\theta}\left(\frac{\partial}{\partial\theta_i} p_\theta\right) p_\theta \;=\; \frac{\partial}{\partial\theta_i}\int p_\theta \;=\; \frac{\partial}{\partial\theta_i} 1 \;=\; 0.$$

3. The **Hessian matrix** of $r(\theta)$ contains the second-order derivatives

$$\frac{\partial^2}{\partial\theta_i\partial\theta_j} \ln \frac{p_\theta}{p_{\theta_0}} \;=\; \frac{\partial^2}{\partial\theta_i\partial\theta_j}\left(\ln p_\theta \;-\; \ln p_{\theta_0}\right) \;=\; \frac{\partial^2}{\partial\theta_i\partial\theta_j} \ln p_\theta.$$

By the chain and product rules of calculus, these are equal to

$$\frac{\partial}{\partial\theta_i}\left(\frac{1}{p_\theta}\frac{\partial}{\partial\theta_j} p_\theta\right) \;=\; \frac{1}{p_\theta}\frac{\partial^2}{\partial\theta_i\partial\theta_j} p_\theta \;-\; \frac{1}{p_\theta^2}\left(\frac{\partial}{\partial\theta_j} p_\theta\right)\left(\frac{\partial}{\partial\theta_j} p_\theta\right).$$

By the same kind of interchange argument as we saw in the case of the gradient, the first of these terms vanish in expectation. The second term, however, evaluates to

$$\begin{aligned} E_\theta\left[-\frac{1}{p_\theta^2}\left(\frac{\partial}{\partial\theta_j} p_\theta\right)\left(\frac{\partial}{\partial\theta_j} p_\theta\right)\right] \;&=\; -E_\theta\left[\frac{1}{p_\theta^2}\left(\frac{\partial}{\partial\theta_j} p_\theta\right)\left(\frac{\partial}{\partial\theta_j} p_\theta\right)\right] \\ &=\; -E_\theta\left[\left(\frac{\partial}{\partial\theta_j} \ln p_\theta\right)\left(\frac{\partial}{\partial\theta_j} \ln p_\theta\right)\right]. \end{aligned}$$

By using the notation $\mathbf{S} \;=\; \nabla \ln p_\theta$ for the **Fisher score vector**, we can restate this result in the more compact form

$$E_\theta\left[\mathbf{H}\right] \;=\; -E_\theta\left[\mathbf{S}\mathbf{S}^T\right].$$

The matrix $\mathbf{F}_\theta \;=\; E_\theta\left[\mathbf{S}\mathbf{S}^T\right]$ is also called the **Fisher information matrix**. This last of these three results thus states that we can read the expected values of second-order partial derivatives off the Fisher information matrix:

$$\mathbf{F}_\theta \;=\; -E_\theta[\mathbf{H}].$$

This is perhaps somewhat surprising, since the Fisher information matrix can be computed using only first-order partial derivatives.

### 3.5 A Quadractic Approximation to the KL Divergence

Putting these three results together, we find that the Kullback-Leibler divergence (i.e., the expected log-likelihood ratio), can be approximated by the quadractic function

$$\begin{aligned} KL(p_{\theta_0} \,||\, p_\theta) &\approx -\frac{1}{2}(\theta - \theta_0)^T E\left[\mathbf{H}(\theta)\right](\theta - \theta_0) \\ &= \frac{1}{2}(\theta - \theta_0)^T \mathbf{F}_\theta (\theta - \theta_0). \end{aligned}$$

This statistic is relatively easy to compute, since the elements of the Fisher information matrix can be estimated empirically from data $\omega \sim p_\theta$.

Another important consequence of this derivation is that it shows us that the Kullback-Leibler divergence behaves approximately like a Mahalanobis distance when we don't look too far away from the current point of evaluation. This also means that we can apply our results about natural gradients to Kullback-Leibler divergences, at least as long as we only take small steps.

Suppose for instance that we have found that steps in the direction of

$$\nabla_\theta E_\theta[R]$$

yields the greatest amount of reward per Euclidean unit step. It then follows that the vector

$$\mathbf{F}_\theta^{-1} \nabla_\theta E_\theta[R]$$

points in the direction of the greatest gain in reward per KL-unit step.

### 3.6 An explicit Example

Suppose $p_{\mu,\tau}$ is the family of Gaussian densities. Under this parametric assumption, the score of an observation $X = x$ is the vector

$$\mathbf{S} = \begin{pmatrix} \tau(x - \mu) \\ \frac{1}{2\tau} - \frac{1}{2}(x - \mu)^2 \end{pmatrix}.$$

The Fisher information matrix corresponding to this score is then

$$\begin{aligned} \mathbf{F}_{\mu,\tau} &= E_{\mu,\tau}\left[\mathbf{S}\mathbf{S}^T\right] \\ &= E_{\mu,\tau}\left[\begin{pmatrix} \tau^2(X - \mu)^2 & \frac{1}{2}(X - \mu) - \frac{\tau}{2}(X - \mu)^3 \\ \frac{1}{2}(X - \mu) - \frac{\tau}{2}(X - \mu)^3 & \frac{1}{4\tau^2} + \frac{1}{4}(X - \mu)^4 - \frac{1}{2\tau}(X - \mu)^2 \end{pmatrix}\right]. \end{aligned}$$

Suppose further that we have currently set our parameters to the values $(\mu_0, \tau_0) = (0, 1)$. Evaluated at this parameter vector, the Fisher information matrix takes the form

$$\mathbf{F}_{\mu_0,\tau_0} = E_{\mu_0,\tau_0}\left[\begin{pmatrix} X^2 & \frac{1}{2}X - X^3 \\ \frac{1}{2}X - X^3 & \frac{1}{4} + \frac{1}{4}X^4 - \frac{1}{2}X^2 \end{pmatrix}\right] = \begin{pmatrix} 1 & 0 \\ 0 & \frac{1}{2} \end{pmatrix}$$

(where I have used standard formulas for computing the moments of a Gaussian). Since the inverse of this matrix is

$$\mathbf{F}_{\mu_0,\tau_0}^{-1} = \begin{pmatrix} 1 & 0 \\ 0 & 2 \end{pmatrix},$$

a parameter gradient pointing in the direction

$$\nabla f = \begin{pmatrix} 1 \\ 1 \end{pmatrix}$$

gives rise to a natural gradient of

$$\mathbf{F}_{\mu_0,\tau_0}^{-1} \nabla f = \begin{pmatrix} 1 \\ 2 \end{pmatrix}.$$

When the distance between two parameter vectors is measured in terms of the divergence between the two corresponding distributions, we should thus spend less of our energy changing the mean of the distribution, and more of it changing its precision.