Energetic Natural Gradient Descent

Philip S. Thomas PHILIPT@CS.CMU.EDU
Bruno Castro da Silva BSILVA@INF.UFRGS.BR
Christoph Dann CDANN@CS.CMU.EDU
Emma Brunskill EBRUN@CS.CMU.EDU

Abstract

We propose a new class of algorithms for minimizing or maximizing functions of parametric probabilistic models. These new algorithms are natural gradient algorithms that leverage more information than prior methods by using a new metric tensor in place of the commonly used Fisher information matrix. This new metric tensor is derived by computing directions of steepest ascent where the distance between distributions is measured using an approximation of energy distance (as opposed to Kullback-Leibler divergence, which produces the Fisher information matrix), and so we refer to our new ascent direction as the *energetic natural gradient*.

1. Introduction

In this paper we consider a fundamental problem within machine learning research: minimizing or maximizing functions of parametric probabilistic models (also called parametrized probability distributions). This problem lies at the heart of all three branches of machine learning. In unsupervised learning, for instance, it arises in the maximization step of the expectation-maximization algorithm. In supervised learning, nearly every modern algorithm fits a parametric model by minimizing a loss function over candidate models. In reinforcement learning the goal is to find a parametric probabilistic model (a stochastic policy) that maximizes an objective function (expected return).

Although sometimes the parameter vector that minimizes or maximizes a function of a parametric model can be solved for analytically, the use of gradient-based algorithms is still common. When using gradient methods we usually know how the parametric model is parametrized, e.g., that the parameters encode the mean and standard deviation of a normal distribution. *Natural gradient descent* using the *Fisher information matrix* (FIM) is one popular extension of the ordinary gradient descent algorithm that leverages this additional information to improve data efficiency (to speed up convergence).

Often we know more than just how we parametrized the parametric model: we also know what the parametric model is a distribution over. This additional knowledge is not leveraged by gradient descent or natural gradient descent using the FIM. **In this paper we propose a new class of natural gradient algorithms that provide more informed update directions by leveraging knowledge of both how the parametric probabilistic model is parameterized and what it is a distribution over.**

We derive our method by computing directions of steepest ascent of the objective function when the distances between probability distributions are measured using an approximation of energy distance. The resulting class of algorithms are natural gradient algorithms that use a new metric tensor that we call the *energy information matrix* (EIM). We therefore call our method *energetic natural gradient descent*.

2. Setting

Let $\boldsymbol{\theta} \in \mathbb{R}^n$ be the parameters of a *parametric probabilistic model* (PPM), $p(\boldsymbol{\theta})$. That is, for each $\boldsymbol{\theta} \in \mathbb{R}^n$, $p(\boldsymbol{\theta})$ is a probability distribution. Let Ω be the set of outcomes that $p(\boldsymbol{\theta})$ is a distribution over, and let $p(\omega|\boldsymbol{\theta})$ denote the probability (or probability density) of ω under $p(\boldsymbol{\theta})$. Let f be an objective function that takes as input a probability distribution and produces as output a real number. We assume that $f \circ p : \mathbb{R}^n \to \mathbb{R}$ is a continuous function, where $\circ$ denotes function composition, i.e., $(f \circ p)(\boldsymbol{\theta}) \coloneqq f(p(\boldsymbol{\theta}))$. Our goal is to find a (local) minimum of $f \circ p$.

For example, we might wish to find the normal distribution that maximizes the log-likelihood of some observed data.

Proceedings of the 33^{rd} International Conference on Machine Learning, New York, NY, USA, 2016. JMLR: W&CP volume 48.

The research reported here was supported by an ONR Young Investigator award, NSF CAREER grant 1350984 and by the Institute of Education Sciences, U.S. Department of Education, through Grant R305A130215 to Carnegie Mellon University. The opinions expressed are those of the authors and do not represent views of the Institute or the U.S. Department of Education.

In this case we might select $n = 2$ so that $\boldsymbol{\theta} = (\boldsymbol{\theta}_1, \boldsymbol{\theta}_2)$, and let $p(\boldsymbol{\theta})$ denote a normal distribution with mean $\boldsymbol{\theta}_1$ and standard deviation $\boldsymbol{\theta}_2$. We could then define $f(p(\boldsymbol{\theta}))$ to be negative the log-likelihood of the data given the normal distribution $p(\boldsymbol{\theta})$, since minimizing negative the log-likelihood corresponds to maximizing the log-likelihood.

3. Background

3.1. The Space of Probability Distributions

We assume that the space of probability distributions representable by p is a smooth semi-Riemannian manifold. In the remainder of this section we present an intuitive description of what this means.

Consider how we can represent the space of probability distributions that is spanned by p. One way to represent this space is by viewing it as $\mathbb{R}^n$, where each point, $\boldsymbol{\theta} \in \mathbb{R}^n$, corresponds to the probability distribution $p(\boldsymbol{\theta})$. However, this set of probability distributions is more than just $\mathbb{R}^n$—each point has meaning—and so this set has additional structure. Specifically, we can define a notion of distance between different probability distributions.

We do this by defining an inner product, $\langle \cdot, \cdot \rangle_{\boldsymbol{\theta}}$, that describes the topology of the set around the point $\boldsymbol{\theta}$, and which changes smoothly with $\boldsymbol{\theta}$. We define this inner product in terms of a positive semidefinite $n \times n$ matrix, $G(\boldsymbol{\theta})$. Selecting different definitions for $G(\boldsymbol{\theta})$ will correspond to different notions of distance between probability distributions. Specifically, we define: $\langle \mathbf{v}, \mathbf{w} \rangle_{\boldsymbol{\theta}} \coloneqq \mathbf{v}^\intercal G(\boldsymbol{\theta}) \mathbf{w}$, and we call G the *metric tensor*. The vector space resulting from the use of this inner product is called a semi-Riemannian manifold. The distance between two points on the manifold, $\boldsymbol{\theta}$ and $\boldsymbol{\theta} + \Delta$, which are close to each other, will be approximately $\sqrt{\langle \Delta, \Delta \rangle_{\boldsymbol{\theta}}} = \sqrt{\Delta^\intercal G(\boldsymbol{\theta}) \Delta}$. This is only approximately the distance because $G(\boldsymbol{\theta})$ is a local description of distance—it describes how distances should be measured around $\boldsymbol{\theta}$. Measuring the distance between $\boldsymbol{\theta}$ and $\boldsymbol{\theta} + \Delta$ exactly would require considering $G(\boldsymbol{\theta}')$ for $\boldsymbol{\theta}'$ between $\boldsymbol{\theta}$ and $\boldsymbol{\theta} + \Delta$.

Notice that if we define $G(\boldsymbol{\theta}) = I_n$ for all $\boldsymbol{\theta}$, where I_n is the $n \times n$ identity matrix, then we are declaring the space of probability distributions to be Euclidean in our parametrization. That is, we assume that the distance between two probability distributions $p(\boldsymbol{\theta}_1)$ and $p(\boldsymbol{\theta}_2)$ is $\sqrt{(\boldsymbol{\theta}_1 - \boldsymbol{\theta}_2)^\intercal (\boldsymbol{\theta}_1 - \boldsymbol{\theta}_2)}$—the Euclidean distance between $\boldsymbol{\theta}_1$ and $\boldsymbol{\theta}_2$.

3.2. Steepest Descent Methods

One way to find a local minimum of $f \circ p$ is to select some initial probability distribution, q, compute the direction of change to q that most rapidly decreases $f(q)$, and

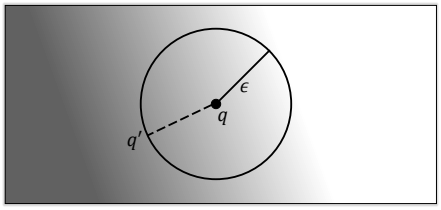


Figure 1: Depiction of the gradient.

then move q a small amount in this direction:

$$q \leftarrow q - \alpha \nabla f(q), \tag{1}$$

where $\alpha \in \mathbb{R}$ is a small positive *step size* and $\nabla f(q)$ is the gradient of f at q—the direction of change to q that most rapidly increases $f(q)$.

Consider in more detail what the gradient—the direction of steepest ascent—really is. Figure 1 depicts the idea behind the gradient. The background denotes the space of all probability distributions, where q is a point (depicted in the middle). The darkness of the background denotes the value of $f(q)$—it is larger for darker regions and smaller for lighter regions. Consider the points on a circle around q with radius ϵ, where ϵ is tiny—infinitesimal. These are all of the points $q + \epsilon\Delta$ where $\|\Delta\| = 1$. Since ϵ is small, and f is continuous, Figure 1 is zoomed in to the point where f appears to be a planar function. Let q' be the point on the circle where f takes the largest value. The direction of the gradient is $q' - q$, and is depicted by the dashed line.

Notice that the direction of the gradient depends on how we measure distances in the space of probability distributions since we require $\|\Delta\| = 1$. This norm, $\|\cdot\|$, encodes our notion of distance in probability space (locally, around q). From the point of view of one notion of distance, a different notion of distance will produce an oblong ellipse rather than a circle. So, different notions of distance can cause the gradient to point in different directions in the space of probability distributions.

In (1) we are performing steepest descent in the space of probability distributions—$\nabla f(q)$ is a direction in the space of probability distributions. Since each $\boldsymbol{\theta}$ corresponds to a probability distribution, we can write this update in terms of $\boldsymbol{\theta}$. This gives our formal definition of steepest descent methods: they compute a sequence, $(\boldsymbol{\theta}_i)_{i=0}^{\infty}$, where $\boldsymbol{\theta}_0$ is chosen arbitrarily and

$$\boldsymbol{\theta}_{i+1} = \boldsymbol{\theta}_i - \alpha_i \widetilde{\nabla}(f \circ p)(\boldsymbol{\theta}_i), \tag{2}$$

where $(\alpha_i)_{i=0}^{\infty}$ is a sequence of step sizes and $\widetilde{\nabla}(f \circ p)(\boldsymbol{\theta})$ is called the *natural gradient* of f at $\boldsymbol{\theta}$. The natural gradient is merely the direction of change to $\boldsymbol{\theta}$ that causes $p(\boldsymbol{\theta})$ to move in the direction of the gradient of f (which is a direction in the space of probability distributions).

In our setting where distances around $p(\boldsymbol{\theta})$ can be measured using the norm $\|\Delta\| = \sqrt{\Delta^\intercal G(\boldsymbol{\theta}) \Delta}$, the natural gradient

has the closed form (Amari, 1998; Thomas, 2014):

$$\widetilde{\nabla}(f \circ p)(\boldsymbol{\theta}_i) \coloneqq G(\boldsymbol{\theta}_i)^+ \frac{\partial (f \circ p)(\boldsymbol{\theta}_i)}{\partial \boldsymbol{\theta}_i},$$

where $G(\boldsymbol{\theta})^+$ denotes the Moore-Penrose pseudoinverse of $G(\boldsymbol{\theta})$ and where here and after we write $\partial(f \circ p)(\boldsymbol{\theta}_i)/\partial \boldsymbol{\theta}_i$ as shorthand for $\frac{\partial (f \circ p)(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}}|_{\boldsymbol{\theta}=\boldsymbol{\theta}_i}$. Also, for the optimization of PPMs, the method of steepest descent is often called *natural gradient descent*.[1]

3.3. Ordinary Gradient Descent

In order to apply the descent algorithm in (2), we must select a G that reflects how we wish to measure distances between probability distributions. One way to do this is to select a G that is simple to compute, but which might not produce a particularly useful notion of distance: let $G(\boldsymbol{\theta}) = I_n$ for all $\boldsymbol{\theta}$. As discussed before, this means that we are defining the distance between two probability distributions, $p(\boldsymbol{\theta}_1)$ and $p(\boldsymbol{\theta}_2)$, to be the Euclidean distance between their parameters: $\sqrt{(\boldsymbol{\theta}_1 - \boldsymbol{\theta}_2)^\intercal (\boldsymbol{\theta}_1 - \boldsymbol{\theta}_2)}$.

Although this metric tensor is trivial to compute, it has a significant drawback: it means that the way that we measure the distance between two probability distributions depends on how we parametrize distributions. Consider what would happen if we constructed two PPMs that span the exact same set of probability distributions. For example, if $n = 2$ and $\boldsymbol{\theta} = (\boldsymbol{\theta}_1, \boldsymbol{\theta}_2)$, we might have $p(\boldsymbol{\theta})$ be the normal distribution with mean $\boldsymbol{\theta}_1$ and *standard deviation* $|\boldsymbol{\theta}_2|$. We can then construct a second parametric model, $q(\boldsymbol{\phi})$, where $\boldsymbol{\phi} = (\boldsymbol{\phi}_1, \boldsymbol{\phi}_2) \in \mathbb{R}^2$ and $q(\boldsymbol{\phi})$ is the normal distribution with mean $\boldsymbol{\phi}_1$ and *variance* $\boldsymbol{\phi}_2$.

Now consider two different probability distributions, both with zero mean, but one with variance equal to one and the other with variance equal to four. These correspond to $p([0, 1]) = q([0, 1])$ and $p([0, 2]) = q([0, 4])$. Under the parametrization used by p, using $G(\boldsymbol{\theta}) = I_n$ means that the distance between $p([0, 1])$ and $p([0, 2])$ is 1.[2] However, under the parameterization used by q, using $G(\boldsymbol{\theta}) = I_n$ means that the distance between the same distributions, $q([0, 1])$ and $q([0, 4])$ is 3. It is straightforward to verify that changes to $\boldsymbol{\theta}_1$ and $\boldsymbol{\phi}_1$ are treated equivalently by the notions of distance induced by p and q, and so one notion of distance is not merely an affine rescaling of the other. So, using this metric tensor, different parametrizations of the PPM induce different notions of distance between probability distributions.

Notice that different notions of distance between probability distributions can result in different directions of steepest ascent. Using $G(\boldsymbol{\theta}) = I_n$ means that the choice of parametrization can change our notion of distance and thus the direction of steepest ascent. Some parametrizations of the PPM might result in a reasonable notion of distance, which in turn results in a reasonable direction of steepest ascent and the sequence of $p(\boldsymbol{\theta}_i)$ produced by the steepest descent method taking a short path to a local minimum. Other parametrizations might result in an absurd notion of distance, thereby producing ascent directions that result in sequences of $p(\boldsymbol{\theta}_i)$ that form a long and curving path to a local minimum.

Using $G(\boldsymbol{\theta}) = I_n$, the closed form equation for the natural gradient becomes particularly simple:

$$\widetilde{\nabla}(f \circ p)(\boldsymbol{\theta}) = \frac{\partial (f \circ p)(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}}.$$

We refer to this direction as the *ordinary gradient*, since it is the gradient that you would get if you used ordinary (non-manifold) calculus to compute the direction of steepest ascent of $f \circ p$ at $\boldsymbol{\theta}$. We also refer to natural gradient descent using this metric tensor as *ordinary gradient descent*. Notice that ordinary gradient descent is a special case of natural gradient descent.

3.4. Fisher Natural Gradient Descent

Rather than allow our choice of parametrization to induce a notion of distance over probability distributions, we can explicitly define a reasonable notion of distance. One popular method for quantifying the similarity of probability distributions is the *Kullback-Leibler divergence* (KLD). The KLD of two probability distributions, $p(\boldsymbol{\theta})$ and $p(\boldsymbol{\theta} + \Delta)$ is given by (for finite Ω):

$$D_{\mathrm{KL}}\left(p(\boldsymbol{\theta}) \| p(\boldsymbol{\theta} + \Delta)\right) \coloneqq \sum_{\omega \in \Omega} p(\omega|\boldsymbol{\theta}) \ln\left(\frac{p(\omega|\boldsymbol{\theta})}{p(\omega|\boldsymbol{\theta} + \Delta)}\right).$$

Pinsker's inequality relates the *square root* of the KLD to a distance metric (Tsybakov, 2009, Lemma 2.5), which suggests that we might treat approximations of the KLD as notions of squared distance (as opposed to distance). So, we will use an approximation, $\frac{1}{2}\Delta^\intercal G(\boldsymbol{\theta})\Delta$, of $D_{\mathrm{KL}}(p(\boldsymbol{\theta})\|p(\boldsymbol{\theta} + \Delta))$ as (half) the squared distance between $p(\boldsymbol{\theta})$ and $p(\boldsymbol{\theta} + \Delta)$.

Let the *Fisher information matrix* (FIM), F, be defined as:

$$F(\boldsymbol{\theta}) \coloneqq \mathbf{E}_{X \sim p(\cdot|\boldsymbol{\theta})}\left[\frac{\partial \ln p(X|\boldsymbol{\theta})}{\partial \boldsymbol{\theta}} \frac{\partial \ln p(X|\boldsymbol{\theta})}{\partial \boldsymbol{\theta}}^\intercal\right],$$

which is guaranteed to be positive semidefinite (although

[1] Technically the natural gradient (Amari, 1998) requires $G(\boldsymbol{\theta})$ to be positive definite for all $\boldsymbol{\theta}$, while the *generalized* natural gradient (Thomas, 2014) only requires $G(\boldsymbol{\theta})$ to be positive semidefinite for all $\boldsymbol{\theta}$. Hereafter we use the generalized natural gradient but refer to it as the natural gradient.

[2] Since $G(\boldsymbol{\theta})$ does not depend on $\boldsymbol{\theta}$, we can compute the distance between two vectors $\mathbf{v}$ and $\mathbf{w}$ as $\sqrt{(\mathbf{v} - \mathbf{w})^\intercal G(\boldsymbol{\theta})(\mathbf{v} - \mathbf{w})}$ (this is not just an approximation).

not necessarily positive definite). In Appendix A we reproduce the known result that $\frac{1}{2}\Delta^\intercal F(\boldsymbol{\theta})\Delta$ is a second order Taylor approximation of $D_{\mathrm{KL}}\left(p(\boldsymbol{\theta})\|p(\boldsymbol{\theta}+\Delta)\right)$, which suggests using $G = F$ (ignoring the scalar, $1/2$, since it corresponds to using a different step size). Using a second order Taylor approximation of KLD is reasonable because $G(\boldsymbol{\theta})$ is only used to define distances locally around $\boldsymbol{\theta}$ (and so higher order terms have little impact). We refer to the natural gradient and natural gradient descent using $G = F$ as *the Fisher natural gradient* and *Fisher natural gradient descent* respectively.

Notice that the FIM depends only on the PPM, not on the objective function, f, which is desirable when f is not known but its gradient is known or can be approximated. Furthermore, it has been shown that the Fisher natural gradient is a *covariant* update direction—its direction in the space of probability distributions does not depend on the parametrization of the PPM. Intuitively, this suggests that the Fisher natural gradient automatically corrects for how a PPM is parametrized. Both of our parametrizations of normal distributions, p and q in earlier examples, will produce the same sequence of probability distributions if used with Fisher natural gradient descent (and an infinitesimal step size, or a step size of fixed length, where length is measured over the semi-Riemannian manifold of probability distributions).

Often the term "natural gradient descent" is used specifically to denote Fisher natural gradient descent. This is likely due to its popularity, which in turn is likely due to the wealth of both theoretical and empirical results it has accrued. For example, it is Fisher efficient (Amari & Douglas, 1998), convergent given only mild assumptions beyond those required by ordinary gradient descent (Thomas, 2014), related to the mirror descent algorithm (Thomas et al., 2013; Raskutti & Mukherjee, 2015), and can be efficiently estimated when using deep neural networks (Desjardins et al., 2015). Empirically it has achieved notable successes for adaptive robotic control (Peters & Schaal, 2008) and stochastic variational inference in topic models (Hoffman et al., 2013).

4. Energetic Natural Gradient Descent

While the Fisher natural gradient corrects for the parametrization of the PPM, it does not depend on f in any way. This is beneficial because often f is not known and can only be sampled, while p (and thus the FIM) is known. However, often some information about f is known. Specifically, usually it is known what set, Ω, called the *sample space*, $p(\boldsymbol{\theta})$ is a distribution over. Often there may be some structure to this space—we may have a distance metric, $d : \Omega \times \Omega \to \mathbb{R}$. We propose a new natural gradient that leverages this additional information.

Consider a simple toy example where we try to find the distribution, $p(\boldsymbol{\theta})$, over daily exercises, $\Omega = \{\omega_1, \omega_2, \omega_3\}$, that minimizes the expected cost of health insurance as predicted by some complicated model, f. We assume that f is unknown, but that we can produce estimates of the gradient of f at any point. Let ω_1 denote a half hour of walking outdoors, ω_2 denote a half hour of walking indoors, and ω_3 denote a half hour of climbing the exterior of tall buildings without safety equipment. Even without knowing more about f, we know that there is some likely structure to the space Ω: ω_1 and ω_2 are quite similar to each other, and quite different from ω_3. For the remainder of this section, let $p(\boldsymbol{\theta}) \coloneqq [\Pr(\omega_1|\boldsymbol{\theta}), \Pr(\omega_2|\boldsymbol{\theta}), \Pr(\omega_3|\boldsymbol{\theta})]^\intercal$.

Consider what happens if we use KLD to measure the squared distance between outcomes for this example. Let $p(\boldsymbol{\theta}_1) = [0.3, 0.3, 0.3]^\intercal$, $p(\boldsymbol{\theta}_2) = [0.2, 0.4, 0.3]^\intercal$, and $p(\boldsymbol{\theta}_3) = [0.2, 0.3, 0.4]^\intercal$. Notice that $D_{\mathrm{KL}}(p(\boldsymbol{\theta}_1)\|p(\boldsymbol{\theta}_2)) = D_{\mathrm{KL}}(p(\boldsymbol{\theta}_1)\|p(\boldsymbol{\theta}_3)) \approx 0.36$. That is, moving probability mass from walking outdoors to walking indoors incurs the same (squared) distance as moving probability mass from walking outdoors to climbing tall buildings without safety equipment. Intuitively, this does not respect what we know about Ω. Because of its use of the KLD, the Fisher natural gradient also ignores this known structure of Ω.

We therefore propose using a notion of distance over probability distributions that, unlike KLD, captures our prior knowledge about the structure of the sample space, Ω. We propose using the *energy distance*, which is sometimes called the *maximum mean discrepancy*. Let $d_{p(\boldsymbol{\theta})}$ be a distance metric over Ω, then the *squared energy distance* between $p(\boldsymbol{\theta}_1)$ and $p(\boldsymbol{\theta}_2)$ is given by:

$$D_{\mathrm{E}}(p(\boldsymbol{\theta}_1), p(\boldsymbol{\theta}_2))^2 \coloneqq \mathbf{E}[2d_{p(\boldsymbol{\theta}_1)}(X, Y) - d_{p(\boldsymbol{\theta}_1)}(X, X') - d_{p(\boldsymbol{\theta}_1)}(Y, Y')],$$

where $X \sim p(\boldsymbol{\theta}_1), X' \sim p(\boldsymbol{\theta}_1), Y \sim p(\boldsymbol{\theta}_2)$, and $Y' \sim p(\boldsymbol{\theta}_2)$. Notice that for generality we let $d_{p(\boldsymbol{\theta})}$ depend on $p(\boldsymbol{\theta})$—this means that the notion of distance between outcomes may depend on the current distribution over outcomes.

As with the KLD, we use a second order Taylor expansion to construct an estimate of the squared energy distance that is in the form that we want. Specifically, we will show that

$$D_{\mathrm{E}}(p(\boldsymbol{\theta}), p(\boldsymbol{\theta}+\Delta))^2 \approx \Delta^\intercal \mathcal{E}(\boldsymbol{\theta})\Delta, \tag{3}$$

where

$$\mathcal{E}(\boldsymbol{\theta}) \coloneqq -\sum_{\omega_1\in\Omega}\sum_{\omega_2\in\Omega} d_{p(\boldsymbol{\theta})}(\omega_1, \omega_2)p(\omega_1|\boldsymbol{\theta})p(\omega_2|\boldsymbol{\theta}) \times \frac{\partial \ln p(\omega_1|\boldsymbol{\theta})}{\partial \boldsymbol{\theta}} \frac{\partial \ln p(\omega_2|\boldsymbol{\theta})}{\partial \boldsymbol{\theta}}^\intercal$$

$$= -\sum_{\omega_1 \in \Omega} \sum_{\omega_2 \in \Omega} d_{p(\boldsymbol{\theta})}(\omega_1, \omega_2) \frac{\partial p(\omega_1|\boldsymbol{\theta})}{\partial \boldsymbol{\theta}} \frac{\partial p(\omega_2|\boldsymbol{\theta})^\intercal}{\partial \boldsymbol{\theta}}$$

$$= -\mathbf{E}_{\substack{X \sim p(\cdot|\boldsymbol{\theta}) \\ Y \sim p(\cdot|\boldsymbol{\theta})}} \left[d_{p(\boldsymbol{\theta})}(X, Y) \frac{\partial \ln p(X|\boldsymbol{\theta})}{\partial \boldsymbol{\theta}} \frac{\partial \ln p(Y|\boldsymbol{\theta})^\intercal}{\partial \boldsymbol{\theta}} \right],$$

where $\times$ denotes scalar multiplication split across multiple lines. We call $\mathcal{E}(\boldsymbol{\theta})$ the *energetic information matrix* (EIM) at $\boldsymbol{\theta}$, because we show later that it is a generalization of the FIM.

First, we show that (3) is a second order Taylor approximation. Although here we skip many algebraic and calculus steps, a step-by-step derivation of this fact is provided in Appendix B. For brevity, let $g_q(\boldsymbol{\theta}) \coloneqq D_{\mathrm{E}}(q, p(\boldsymbol{\theta}))^2$ be the squared energy distance between q and $p(\boldsymbol{\theta})$. The Jacobian and Hessian of g_q at $\boldsymbol{\theta}$ are:

$$\frac{\partial g_q(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}} = 2 \sum_{\omega_1 \in \Omega, \omega_2 \in \Omega} (q(\omega_1) - p(\omega_1|\boldsymbol{\theta}))\, d_{p(\boldsymbol{\theta})}(\omega_1, \omega_2) \frac{\partial p(\omega_2|\boldsymbol{\theta})}{\partial \boldsymbol{\theta}}$$

$$\frac{\partial^2 g_q(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}^2} = 2 \sum_{\omega_1 \in \Omega, \omega_2 \in \Omega} (q(\omega_1) - p(\omega_1|\boldsymbol{\theta}))\, d_{p(\boldsymbol{\theta})}(\omega_1, \omega_2) \frac{\partial^2 p(\omega_2|\boldsymbol{\theta})}{\partial \boldsymbol{\theta}^2}$$

$$- 2 \sum_{\omega_1 \in \Omega, \omega_2 \in \Omega} d_{p(\boldsymbol{\theta})}(\omega_1, \omega_2) \frac{\partial p(\omega_1|\boldsymbol{\theta})}{\partial \boldsymbol{\theta}} \frac{\partial p(\omega_2|\boldsymbol{\theta})^\intercal}{\partial \boldsymbol{\theta}}.$$

A second order Taylor approximation of $g_{p(\boldsymbol{\theta})}(\boldsymbol{\theta} + \Delta)$ around $\boldsymbol{\theta}$ is defined as

$$g_{p(\boldsymbol{\theta})}(\boldsymbol{\theta} + \Delta) \overset{\text{Taylor}_2}{\approx} g_{p(\boldsymbol{\theta})}(\boldsymbol{\theta}) + \Delta^\intercal \frac{\partial g_{p(\boldsymbol{\theta})}(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}} + \frac{1}{2} \Delta^\intercal \frac{\partial^2 g_{p(\boldsymbol{\theta})}(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}^2} \Delta,$$

which, after substituting in our equations for the Jacobian and Hessian and performing algebraic manipulations, is:

$$g_{p(\boldsymbol{\theta})}(\boldsymbol{\theta} + \Delta) \overset{\text{Taylor}_2}{\approx} \Delta^\intercal \mathcal{E}(\boldsymbol{\theta}) \Delta,$$

which by the definition of g is the desired result, (3).

We now have our new natural gradient algorithm, *energetic natural gradient descent*, which is natural gradient descent using $G = \mathcal{E}$. Notice that computing $\mathcal{E}$ requires knowledge of p and d, but not necessarily f. This means that energetic natural gradient descent, like Fisher natural gradient descent, is applicable when f is not known, but where the gradient of f can be estimated.

5. Illustrative Example

In this section we present an example that illustrates the potential benefits of the energetic natural gradient over the Fisher natural gradient and ordinary gradient. Let $\Omega = \{\omega_1, \omega_2, \omega_3\}$, $\boldsymbol{\theta} = (\boldsymbol{\theta}_1, \boldsymbol{\theta}_2)^\intercal \in \mathbb{R}^2$,

$$p(\boldsymbol{\theta}) = \begin{bmatrix} p(\omega_1|\boldsymbol{\theta}) \\ p(\omega_2|\boldsymbol{\theta}) \\ p(\omega_3|\boldsymbol{\theta}) \end{bmatrix} = \begin{bmatrix} \boldsymbol{\theta}_1 \\ \boldsymbol{\theta}_2 \\ 1 - \boldsymbol{\theta}_1 - \boldsymbol{\theta}_2 \end{bmatrix},$$

where $\boldsymbol{\theta}_1 \geq 0$, $\boldsymbol{\theta}_2 \geq 0$, and $\boldsymbol{\theta}_1 + \boldsymbol{\theta}_2 \leq 1$ to ensure that $p(\boldsymbol{\theta})$ is a probability distribution. Let

$$f(p(\boldsymbol{\theta})) \coloneqq -p(\boldsymbol{\theta})^\intercal \begin{bmatrix} 1 & 0.5 & 0 \\ 0.5 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix} p(\boldsymbol{\theta}).$$

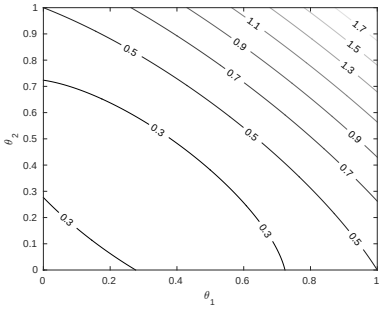
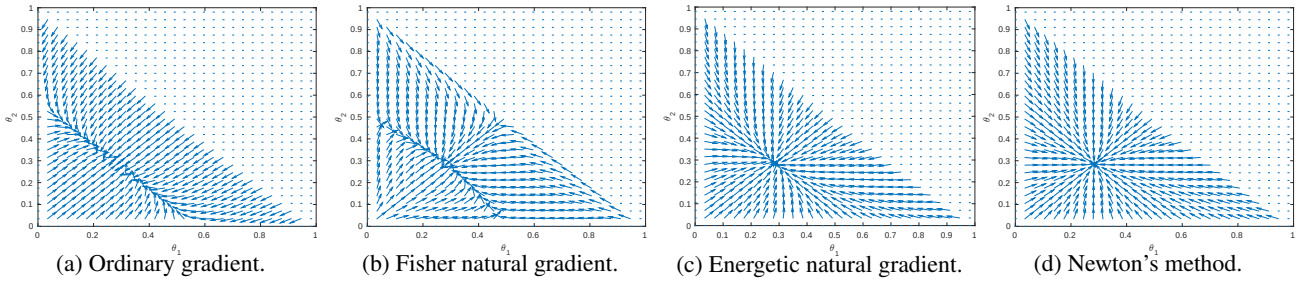


Figure 2: Contour plot of $f \circ p$ for the illustrative example.

Figure 2 depicts $f \circ p$. Consider what the distance metric, $d_{p(\boldsymbol{\theta})}$, over Ω, might be in this case. The matrix in the definition of f suggests a setting where ω_1 and ω_2 are similar, but ω_3 is quite different from ω_1 and ω_2 (like in our health insurance example). We might therefore choose the distance between ω_1 and ω_2 to be some constant in $[0, 1)$, e.g., 0.7, and the distance between ω_3 and either of the other two events to be 1.0. Let $d_{p(\boldsymbol{\theta})}$ be described as a 3×3 matrix where the element in the i^{th} row and j^{th} column corresponds to the distance between ω_i and ω_j, and where $d_{p(\boldsymbol{\theta})}$ is the same for all $\boldsymbol{\theta}$:

$$d_{p(\boldsymbol{\theta})} = \begin{bmatrix} 0 & 0.7 & 1 \\ 0.7 & 0 & 1 \\ 1 & 1 & 0 \end{bmatrix}.$$

Figures 3a, 3b, and 3c show the (negative) ordinary gradient, Fisher natural gradient, and energetic natural gradient of $f \circ p$ given these definitions. Notice that at most points the ordinary gradient does not point towards the global optimum near $\boldsymbol{\theta} = (0.3, 0.3)^\intercal$. This is particularly noticeable when one parameter is near one. The Fisher natural gradient corrects for how p is parametrized, and this helps a little. For example, if either parameter is near one, the natural gradient points much closer to the global optimum than the ordinary gradient. However, the natural gradient does not uniformly point towards the global optimum. The energetic natural gradient leverages our knowledge about how p is parametrized as well as prior knowledge about a reasonable d to improve upon the Fisher natural gradient



(a) Ordinary gradient. (b) Fisher natural gradient. (c) Energetic natural gradient. (d) Newton's method.

Figure 3: Different ascent directions of the illustrative $f \circ p$.

and ordinary gradient—at every point it points close to the global optimum.

Although the energetic natural gradient is an improvement upon the ordinary gradient and Fisher natural gradient in this example, it is not ideal—it does not point exactly towards the global optimum from all points. This happens because we only used intuitive knowledge about f—knowledge about how distances over Ω might be measured. If, however, we knew exactly what f was when selecting $d_{p(\boldsymbol{\theta})}$, then we might have chosen $d_{p(\boldsymbol{\theta})}(\omega_1, \omega_2) = 0.5$, since that causes $\mathcal{E}(\boldsymbol{\theta})$ to be the Hessian of $f \circ p$, and thus the energetic natural gradient to be the update direction of Newton's method, which is depicted in Figure 3d.

This highlights one setting where one should *not* use the energetic natural gradient: when the Hessian of $f \circ p$ is known. In this setting, using Newton's method will generally be preferable to energetic natural gradient descent. However, if the Hessian is not known, then one may still apply energetic natural gradient descent to get improvements over Fisher natural gradient descent as shown by this example. We view Figures 3a through 3d as being a sequence of update directions that leverage (but also require to be available) an increasing amount of knowledge about the optimization problem at hand. To the left we have methods that require little knowledge about f or p, and left of Figure 3a we might place optimization methods like CMA-ES (Hansen, 2006) that require no knowledge of the gradient. On the right we have methods that require more knowledge about f and p, and to the right of Figure 3d we might place analytic solutions to the optimization problem. One argument in favor of the energetic natural gradient is that it leverages knowledge that is typically available when applying the Fisher natural gradient.

6. How to Select d

In the previous sections we derived the energetic natural gradient assuming that a distance metric, d, over outcomes, Ω, is available. An obvious question is: how should d be selected for other problems? This remains an open problem. However, although we have yet to find a satisfactory answer to the question of how to define an optimal d in general, our example from the previous section suggests that even a d chosen based on intuition can provide benefits over the Fisher natural gradient, which ignores distances over events altogether. In Section 8 we provide another example that supports this claim in the context of reinforcement learning.

Moreover, the analysis in the next section shows that d can be chosen to make energetic natural gradients identical to Fisher natural gradients, and so there always exists a d that causes energetic natural gradient descent to perform at least as well as Fisher natural gradient descent.

7. Theoretical Analysis of Energetic Gradient Descent

In the previous sections we gave intuitive motivation for and derived the energetic natural gradient. We then discussed how the distance metrics, d, could be selected. In this section we provide a first theoretical analysis of the EIM, energetic natural gradient, and energetic natural gradient descent.

7.1. Positive definiteness of EIM

For the energetic natural gradient to be an ascent direction, $\mathcal{E}(\boldsymbol{\theta})$ must be positive *definite*. Less restrictively, for the energetic natural gradient to not be a descent direction, $\mathcal{E}(\boldsymbol{\theta})$ must be positive *semidefinite*. The following theorem establishes sufficient conditions for $\mathcal{E}(\boldsymbol{\theta})$ to be positive semidefinite. Specifically, we show that if d is a conditionally negative definite distance, then $\mathcal{E}(\boldsymbol{\theta})$ is positive semidefinite, which in turn implies that the energetic natural gradient will not be a descent direction.[3]

Theorem 1. *If $|\Omega| < \infty$ and $d_{p(\boldsymbol{\theta})}$ is conditionally negative semidefinite, then $\mathcal{E}(\boldsymbol{\theta})$ is positive semidefinite.*

Proof. See Appendix C. □

[3] Recall that when minimizing $f \circ p$, we move $\boldsymbol{\theta}$ in the direction of the *negative* energetic natural gradient, so we want the energetic natural gradient to be an ascent direction (or at least not a descent direction) so that the negative energetic natural gradient is a descent direction (or at least not an ascent direction).

In Appendix D we provide conditions to ensure that a $d_{p(\boldsymbol{\theta})}$ is conditionally negative semidefinite and give an example of a distance metric, $d_{p(\boldsymbol{\theta})}$, that is not conditionally negative semidefinite.

7.2. FIM is a special case of EIM

The following theorem shows that the FIM is a special case of the EIM—there is a choice of d that causes the two to be equivalent. Let $\mathbf{1}_{(\cdot)}$ denote the indicator function.

Theorem 2. *$\mathcal{E}(\boldsymbol{\theta}) = F(\boldsymbol{\theta})$ if $|\Omega| < \infty$ and*

$$d_{p(\boldsymbol{\theta})}(\omega_1, \omega_2) \coloneqq \mathbf{1}_{(\omega_1 \neq \omega_2)} \left(\frac{1}{2p(\omega_1|\boldsymbol{\theta})} + \frac{1}{2p(\omega_2|\boldsymbol{\theta})} \right).$$

Proof. See Appendix E. □

The d that causes FIM and EIM to be equivalent provides insight into why Fisher natural gradient descent tends to converge slowly to distributions with very low probabilities associated with some outcome, ω. Since $1/p(\omega|\boldsymbol{\theta})$ is very large in this case, FIM is inducing a notion of distance over outcomes such that changing $p(\omega|\boldsymbol{\theta})$ incurs a large distance. This will cause the natural gradient to favor update directions that change the probabilities of other events while leaving the probability of ω relatively unchanged, which means that the probability of ω will only slowly approach its ideal value.

7.3. The Energetic Natural Gradient is a Covariant Update Direction

Recall that when using ordinary gradient descent the choice of parametrization impacts the notion of distance over probability distributions. This means that the path that ordinary gradient descent takes through the space of probability distributions depends on the parametrization of the PPM. In general, this is undesirable. We prefer methods that are robust to different parametrizations—methods that ensure that the path through the space of probability distributions does not depend on how we parametrize the PPM. Such methods are called *covariant*. A formal definition of what it means for an update direction to be covariant (there are a few minor technicalities) is provided in Appendix F.

Below we present a theorem which establishes that the energetic natural gradient is a covariant update direction.

Theorem 3. *The energetic natural gradient is a covariant update direction.*

Proof. See Appendix F. □

7.4. Relation to Natural Gradient

The obvious statement that energetic natural gradient descent is a natural gradient descent algorithm has important ramifications. Perhaps most importantly, energetic natural gradient descent therefore inherits the theoretical properties afforded to all natural gradient algorithms. For example, conditions for the convergence of natural gradient algorithms (and thus energetic natural gradient descent) can be found in the work of Thomas (2014, Theorems 2, 3, 4, and 5).

An equivalence between natural gradient descent and mirror descent, a popular smooth, constrained, convex optimization algorithm (Nemirovski & Yudin, 1983; Beck & Teboulle, 2003), was recently established (Raskutti & Mukherjee, 2015). When this equivalence holds (it only holds in certain settings), natural gradient algorithms, including energetic natural gradient descent, inherit the theoretical properties of mirror descent. One important ramification of this is that energetic natural gradient methods can be extended to allow for constraints on the set of parameter vectors, $\boldsymbol{\theta}$, that are allowed (Thomas et al., 2013).

8. Example Application: Reinforcement Learning

In this section we assume that the reader is familiar with reinforcement learning (Sutton & Barto, 1998, RL) and evalute the performance of energetic gradient descent for an RL application. Our experiments provide empirical evidence that an agent optimizing its behavior via energetic natural gradient descent can execute more efficient update steps than one using the ordinary gradient or Fisher natural gradient. In particular, we show that, for a wide range of initial solutions (initial policy parameters), the energetic natural gradient consistently points towards better solutions than the ordinary gradient and Fisher natural gradient.

This presents an interesting question: what experiment would provide evidence that one update *direction* is superior to another? If we create algorithms that try to efficiently estimate each update direction from data and show standard learning curves, then our results would conflate the data-efficiency of a particular algorithm's gradient estimates with the quality of the update direction—i.e., is the energetic natural gradient really a better update direction, or does it just require less data before estimates of it become ascent directions? Similarly, if we use an adaptive step size or a fixed step size, then our algorithms would conflate the quality of an update direction with the compatibility of the selected (adaptive) step size.

Because of these complications, we chose to estimate each update direction as accurately as possible (using large amounts of data) and to use line searches to find the optimal step size for each method. We then compare the performance of the resulting policy after a single update. If we produced standard learning curves, only the first step would

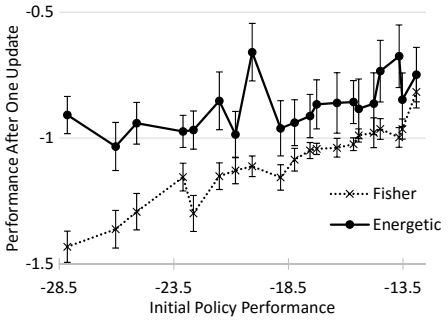


Figure 4: Comparison of updates for mountain car.

provide a meaningful comparison—each gradient method would begin from a different $\boldsymbol{\theta}$ when making its second step and so the quality of $\boldsymbol{\theta}_2$ conflates the quality of two different updates; a method could appear superior over the entire learning process if it produces a better first update but is equivalent thereafter. We therefore compute a set of policies from which each gradient method is run, and we report the performance of the policies that result from using each gradient method for one step. Notice that this tests whether the energetic natural gradient is a better update direction—we do *not* propose or evaluate a specific algorithm for estimating the energetic natural policy gradient efficiently for learning.

We selected a variant of the canonical mountain car domain (Thomas, 2015, Section 4.10.2). Each gradient direction was computed by fixing the agent's policy, generating 20,000 trajectories, and then computing the gradient using REINFORCE, the sample FIM, and the sample EIM. The performance under an optimal policy is zero. Figure 4 shows, on the horizontal axis, the different initial policies from which we execute the different gradient updates; each initial policy was obtained by interpolating between a completely random policy and a policy with intermediate performance. Each point in the x-axis is annotated with the performance of the corresponding initial policy. We plot the performance of the policy produced by a single step of each gradient method.

When computing the energetic natural gradient, we followed the approach taken by Kakade (2002): we treated the policy as a set of distributions—one per state. We then defined the EIM for each state and took the average per-state EIM as our final EIM. We defined the distance metric, d, as the difference between the q-values of two actions at a given state; that is, $d_s(a_1, a_2) = |q(s, a_1) - q(s, a_2)|$. Each point in Figure 4 corresponds to the performance achieved after one update step, averaged over 20 trials and including standard error bars.

From Figure 4 we can see that the energetic natural gradient consistently provides better update directions when evaluated over a wide range of initial policies. In particular, it results in policy performance that is consistently higher than that achieved via the Fisher natural gradient—around 30% higher if applied to a random policy (leftmost point in Figure 4). As the initial policy's performance increases, the gap between energetic natural gradient and Fisher natural gradient decreases, indicating that when we approach a near-optimal solution, there are fewer ways in which it can be significantly improved. However, note that even when this is the case, following the energetic natural gradient still results in consistently higher performance following a policy update. The curve depicting the performance resulting from following the ordinary gradient does not appear in Figure 4 because it never increased past performance -5 when evaluated over the range of possible initial policies.

9. Conclusion

In this paper we introduced a new variant of gradient descent that we call *energetic natural gradient descent*. The energetic natural gradient is a new member of the family of *natural gradient* algorithms (Amari, 1998) that leverages more prior knowledge than the most common natural gradient algorithms, which use the Fisher information matrix. Specifically, the natural gradient using the Fisher information matrix leverages prior knowledge about how a parametric model is parametrized to improve data efficiency relative to ordinary gradient descent. However, it does *not* leverage any knowledge of what the parametric model is a distribution over. By contrast, the energetic natural gradient leverages both information about how the parametric model was parametrized and what the model is a distribution over.

We show that energetic gradient descent has many desirable theoretical properties: **1)** under common conditions, it is always a descent direction, **2)** natural gradients using the Fisher information matrix are a special case of our new more general approach, and **3)** the energetic natural gradient is a covariant update direction—it is not sensitive to how the parametric model was parametrized. Finally, we presented an empirical example where the straightforward application of energetic natural gradients to the problem of policy search in reinforcement learning produces better update directions than the natural gradient using the Fisher information matrix and the ordinary gradient.

Several avenues of future work remain. We have presented a generally applicable optimization algorithm; it remains for this algorithm to be adapted to individual applications. For example, there is a particularly efficient linear-time method for approximating the Fisher natural gradient for policy search in reinforcement learning (Bhatnagar et al., 2009)—does a similarly efficient approximation exist for the energetic natural gradient? Do other definitions of d produce better results for reinforcement learning?

References

Amari, S. Natural gradient works efficiently in learning. *Neural Computation*, 10:251–276, 1998.

Amari, S. and Douglas, S. Why natural gradient? In *Proceedings of the 1998 IEEE International Conference on Acoustics, Speech, and Signal Processing*, volume 2, pp. 1213–1216, 1998.

Beck, A. and Teboulle, M. Mirror descent and nonlinear projected subgradient methods for convex optimization. *Operations Research Letters*, 2003.

Bhatnagar, S., Sutton, R. S., Ghavamzadeh, M., and Lee, M. Natural actor-critic algorithms. *Automatica*, 45(11): 2471–2482, 2009.

Dabney, W. and Thomas, P. S. Natural temporal difference learning. In *Twenty-Eighth AAAI Conference on Artificial Intelligence*, 2014.

Desjardins, G., Simonyan, K., Pascanu, R., et al. Natural neural networks. In *Advances in Neural Information Processing Systems*, pp. 2062–2070, 2015.

Hansen, N. The CMA evolution strategy: a comparing review. In Lozano, J.A., Larranaga, P., Inza, I., and Bengoetxea, E. (eds.), *Towards a new evolutionary computation. Advances on estimation of distribution algorithms*, pp. 75–102. Springer, 2006.

Hoffman, M., Blei, D., Wang, C., and Paisley, J. Stochastic Variational inference. *Journal of Machine Learning Research*, 14:1303–1347, 2013.

Kakade, S. A natural policy gradient. In *Advances in Neural Information Processing Systems*, volume 14, pp. 1531–1538, 2002.

Nemirovski, A. and Yudin, D. *Problem Complexity and Method Efficiency in Optimization*. Wiley, New York, 1983.

Peters, J. and Schaal, S. Natural actor-critic. *Neurocomputing*, 71:1180–1190, 2008.

Rao, C. Radhakrishna. Convexity properties of entropy functions and analysis of diversity. *IMS Lecture Notes - Nonograph Series*, 5:68–77, 1984. URL http://www.jstor.org/stable/10.2307/4355484.

Raskutti, G. and Mukherjee, S. The information geometry of mirror descent. *Information Theory, IEEE Transactions on*, 61(3):1451–1457, 2015.

Schoenberg, I. J. Metric spaces and positive definite functions. *Transactions of the American Mathematical Society*, 44:522–522, 1938. ISSN 0002-9947. doi: 10.1090/S0002-9947-1938-1501980-0.

Sutton, R. S. and Barto, A. G. *Reinforcement Learning: An Introduction*. MIT Press, Cambridge, MA, 1998.

Thomas, P. S. GeNGA: A generalization of natural gradient ascent with positive and negative convergence results. In *Proceedings of the Thirty-First International Conference on Machine Learning*, 2014.

Thomas, P. S. *Safe Reinforcement Learning*. PhD thesis, University of Massachusetts Amherst, 2015.

Thomas, P. S., Dabney, W., Mahadevan, S., and Giguere, S. Projected natural actor-critic. In *Advances in Neural Information Processing Systems 26*, 2013.

Tsybakov, A. B. Introduction to nonparametric estimation. revised and extended from the 2004 french original. translated by vladimir zaiats, 2009.