The information geometry of mirror descent

Garvesh Raskutti[1] Sayan Mukherjee[2,3]

[1] Department of Statistics, University of Wisconsin-Madison
[2] Departments of Statistical Science, Computer Science, and Mathematics, Duke University
[3] Institute for Genome Sciences & Policy, Duke University

Abstract

Information geometry applies concepts in differential geometry to probability and statistics and is especially useful for parameter estimation in exponential families where parameters are known to lie on a Riemannian manifold. Connections between the geometric properties of the induced manifold and statistical properties of the estimation problem are well-established through the Cramér-Rao lower bound. However developing first-order methods that scale to larger problems has been less of a focus in the information geometry community. The best known algorithm that incorporates manifold structure is the *natural* gradient descent algorithm introduced by Amari, which is a second-order method. On the other hand, stochastic approximation methods have led to the development of first-order methods for optimizing noisy objective functions. The classical Robbins-Monro and Keifer-Wolfowitz algorithms are the basis for many stochastic approaximation algorithms. A recent generalization of the Robbins-Monro algorithm known as mirror descent, developed by Nemirovski and Yudin is a first order method that induces non-Euclidean geometries. However current analysis of mirror descent does not precisely characterize the induced non-Euclidean geometry nor does it consider performance in terms of statistical relative efficiency. In this paper, we prove that mirror descent induced by Bregman divergences is equivalent to the *natural* gradient descent algorithm on the *dual* Riemannian manifold. Using this equivalence between natural gradient descent and mirror descent, it follows that (1) mirror descent is the steepest descent direction along the Riemannian manifold of the exponential family; (2) mirror descent with log-likelihood loss applied to parameter estimation in exponential families asymptotically achieves the classical Cramér-Rao lower bound and (3) natural gradient descent for manifolds corresponding to exponential families can be implemented as a first-order method through mirror descent.

1 Introduction

Information geometry, which marries concepts from differential geometry with statistical decision theory was born out of the pioneering work of Rao [18]. Along with Cramér, Rao proved the classical Cramér-Rao bound [12, 18] and showed that the Fisher information matrix for parametric families induces a Riemannian manifold. Detailed characterizations of the differential geometric properties for various statistical manifolds are provided in [5]. Work by Amari and other authors in recent years has provided further signifcant contributions to information geometry by making links to information theory and incorporating ideas from optimization and online learning (see e.g. [2, 3, 4]).

In particular Amari proposed *natural* gradient descent [2] as an online learning algorithm that incorporates manifold structure. The algoirthm generalizes online gradient descent by multiplying the gradient update by the inverse of the Riemannian metric. For regular parametric families, the Riemannian metric corresponds to the inverse of the Fisher information matrix. Importantly, Amari proves that using the natural gradient descent step asymptotically satisfies the Cramér-Rao lower bound for parameter estimation in exponential families. However, the natural gradient descent algorithm is a second-order method since it requires inversion of the Hessian or Fisher-information matrix correpsonding to the Riemannian manifold. When the number of parameters is large, second-order methods are often a significant computational burden.

On the other hand, there has recently been a strong focus on developing first-order methods to estimate noisy objective functions in the machine learning community. The idea of stochastic updates or stochastic optimization procedures has a long history in optimizing statistical objectives going back to Robbins and Monro [20] and Kiefer and Wolfowitz [16]. The widely used online gradient descent algorithm is simply a special case of the Robbins-Monro algorithm. More recently, Nemirovski and Yudin developed a generalization of online or stochastic gradient descent known as *mirror* descent [17]. The mirror descent algorithm allows non-Euclidean geometries to be induced via the choice of a proximity function (see e.g. [23]). Importantly mirror descent is a first-order method and has subseqently received significant attention (see e.g. [1, 8, 13, 23]). However, to the best of our knowledge, the mirror descent literature has generally not addressed two important issues: (i) developing a precise characterization of the non-Euclidean geometry induced by mirror descent and (ii) statistical analysis with respect to a model that considers variance and efficiency properties of the mirror descent update.

In this paper we consider mirror descent updates where the proximity function is a *Bregman divergence* [10] or equivalently, the Kullback-Leibler divergence for an exponential family. Bregman divergences are differentiable functions on $\mathbb{R}^p \times \mathbb{R}^p$ indexed by strictly convex differentiable functions G. Bregman divergences induce Riemannian manifolds and for each Bregman divergence, there exists a *dual* Bregman divergence inducing a dual Riemannain manifold (see Amari and Cichocki [3] for details). Using this connection, we prove that mirror descent with Bregman divergences is equivalent to natural gradient descent [2] along the *dual* Riemannian manifold. An immediate consequence of this equivalence between mirror descent with Bregman divergences and natural gradient descent allows us to make three novel statements about mirror descent. Firstly, the mirror descent step is the direction of steepest descent in the dual Riemannian manifold corresponding to the Bregman diveregence. Secondly, using the one-to-one correspondence between Bregman divergences and exponential families (see e.g. [6, 7]), mirror descent applied to parameter estimation in the exponential family corresponding to the Bregman divergence asymptotically achieves the Cramér-Rao lower bound [12, 18]. Hence we address both of the stated issues for mirror descent when the proximity function is a Bregman divergence. Thirdly, the equivalence proves that mirror descent with Bregman divergences is a first-order implementation of natural gradient descent along the dual Riemannian manifold.

We also discuss connections between mirror descent, natural gradient descent and other online algorithms that directly impose a Riemannian structure. The standard gradient descent step on a Riemannian manifold requires applying the exponential map to the gradient update (see e.g. [9]). Since computation of the exponential map is challenging in general, approximations to the exponential map are used. We prove that natural gradient descent (and consequently mirror descent) is equivalent to standard Riemannian gradient descent with the exponential map approximated by its first-order Taylor approximation. Hence mirror descent and natural gradient descent can be motivated as first-order approximations to the computationally intensive Riemannian gradient descent step.

The remainder of the paper is organized as follows: In Section 2 we introduce the basic concepts including mirror descent, Bregman divergences, convex duality, Riemannian manifolds and natural gradient descent. Section 3 presents the main result on the equivalence between mirror descent and natural gradient descent as well as consequences for statistical estimation and connections to other online algorithms for Riemannian manifolds. The discussion and conclusion is presented in Section 4.

2 Background and preliminaries

We begin by introducing the notation and problem setup for online convex optimization as defined in Shalev-Shwartz [22]. Let $\{f_t\}_{t=1}^{\infty}$ denote a sequence of convex differentiable cost functions $f_t : \Theta \to \mathbb{R}$ where $\Theta \subset \mathbb{R}^p$ is a convex set. The online learning problem is to predict a sequence of vectors $\{\theta_t\}_{t=1}^{\infty}$ and incur a loss $f_t(\theta_t)$ at each iterate t. The sum of losses $\sum_{t=1}^{T} f_t(\theta_t)$ is referred to as the *regret* at T and the goal is to construct a sequence with the smallest possible regret. There is a large body of work on providing bounds on regret for various online algorithms (see e.g. [1, 13, 23]), however the main focus of this paper is understanding the geometry of the mirror descent algorithm in the context of a generative statistical model by proving an equivalence to natural gradient descent.

2.1 Mirror descent with Bregman divergences and convex duality

The most common approach to construct a sequence $\{\theta_t\}_{t=1}^{\infty}$ is based on online or stochastic gradient descent. The online gradient descent update is:

$$\theta_{t+1} = \theta_t - \alpha_t \nabla f_t(\theta_t), \tag{1}$$

where $(\alpha_t)_{t=0}^{\infty}$ denotes a sequence of step-sizes. Note that the online gradient descent step can alternatively be expressed as:

$$\theta_{t+1} = \arg\min_{\theta \in \Theta} \left\{ \langle \theta, \nabla f_t(\theta_t) \rangle + \frac{1}{2\alpha_t} \|\theta - \theta_t\|_2^2 \right\}.$$

By re-expressing the stochastic gradient step in this way, Nemirovski and Yudin [17] introduced a generalization of gradient descent as follows: Denote the *proximity* function $\Psi : \mathbb{R}^p \times \mathbb{R}^p \to \mathbb{R}^+$, strictly convex in the first argument, then define the *mirror* descent step as:

$$\theta_{t+1} = \arg\min_{\theta \in \Theta} \left\{ \langle \theta, \nabla f_t(\theta_t) \rangle + \Psi(\theta, \theta_t) \right\}. \tag{2}$$

Setting $\Psi(\theta, \theta') = \frac{1}{2}\|\theta - \theta'\|_2^2$ yields the standard gradient descent update, hence (2) is a generalization of online gradient descent.

A standard choice for the proximity function Ψ is the so-called *Bregman divergence* since they correspond to the Kullback-Leibler divergence for an exponential family. In particular, let $G : \Theta \to \mathbb{R}$ denote a strictly convex twice-differentiable function, the divergence introduced by Bregman [10] $B_G : \Theta \times \Theta \to \mathbb{R}^+$ is:

$$B_G(\theta, \theta') = G(\theta) - G(\theta') - \langle \nabla G(\theta'), \theta - \theta' \rangle.$$

Bregman divergences are widely used in statistical inference, optimization, machine learning, and information geometry (see e.g. [3, 7]). Letting $\Psi(\cdot, \cdot) = B_G(\cdot, \cdot)$, the mirror descent step defined is:

$$\theta_{t+1} = \arg\min_{\theta} \left\{ \langle \theta, \nabla f_t(\theta_t) \rangle + \frac{1}{\alpha_t} B_G(\theta, \theta_t) \right\}. \tag{3}$$

Examples of G and the induced Bregman divergences are listed in Table 1. For a more extensive list, see e.g. [7]. There is a one-to-one correspondence between Bregman divergences and exponential families [7] which we exploit later when we discuss estimation in exponential families.

The concept of convex conjugate functions is central to the main result in the paper. The convex conjugate function of G is defined to be:

$$H(\eta) := \sup_{\theta \in \Theta} \left\{ \langle \theta, \eta \rangle - G(\theta) \right\}.$$

$G(\theta)$	$B_G(\theta, \theta')$
$\frac{1}{2}\|\theta\|_2^2$	$\frac{1}{2}\|\theta - \theta'\|_2^2$
$\exp(\theta)$	$\exp(\theta) - \exp(\theta') - \langle \exp(\theta'), \theta - \theta' \rangle$
$\log(1 + \exp(\theta))$	$\log\left(\frac{1+e^{\theta}}{1+e^{\theta'}}\right) - \langle \frac{e^{\theta'}}{1+e^{\theta'}}, \theta - \theta' \rangle$

Table 1: Bregman divergence examples

$G(\theta)$	$H(\eta)$	$B_H(\eta, \eta')$
$\frac{1}{2}\|\theta\|_2^2$	$\frac{1}{2}\|\eta\|_2^2$	$\frac{1}{2}\|\eta - \eta'\|_2^2$
$\exp(\theta)$	$\langle \eta, \log \eta \rangle - \eta$	$\eta \log \frac{\eta}{\eta'}$
$\log(1 + \exp(\theta))$	$\eta \log \eta + (1 - \eta) \log(1 - \eta)$	$(1 - \eta) \log\left(\frac{1-\eta}{1-\eta'}\right) + \eta \log \frac{\eta}{\eta'}$

Table 2: Dual Bregman divergence examples

If G is lower semi-continuous, G is the convex conjugate of H, implying a dual relationship between G and H. Further, since G is strictly convex and twice differentiable, so is H. Note also that if $g = \nabla G$ and $h = \nabla H$, $g = h^{-1}$. For additional properties and motivation for the convex conjugate function, see Rockafeller [21].

Let $\eta = g(\theta) \in \Phi$ be the point at which the supremum for the dual function is attained represent the *dual* co-ordinate system to θ. The dual Bregman divergence $B_H : \Phi \times \Phi \to \mathbb{R}^+$ is:

$$B_H(\eta, \eta') = H(\eta) - H(\eta') - \langle \nabla H(\eta'), \eta - \eta' \rangle.$$

Using the dual co-ordinate relationship, it is straightforward to show that $B_H(\eta, \eta') = B_G(h(\eta'), h(\eta))$ and $B_G(\theta, \theta') = B_H(g(\theta'), g(\theta))$. Dual functions and Bregman divergences for examples in Table 1 are presented in Table 2. For more examples see Banerjee et al. [7].

2.2 Riemannian manifolds and natural gradient descent

Let $(\mathcal{M}, \mathcal{H})$ be a p-dimensional Riemannian manifold with metric tensor $\mathcal{H} = (h_{jk})$ and $\mathcal{M} \subset \mathbb{R}^p$. For a thorough introduction to Riemannian manifolds, see doCarmo [11].

Now, define a sequence of functions $\{\tilde{f}_t\}_{t=0}^{\infty}$ on the Riemannian manifold $\tilde{f}_t : \mathcal{M} \to \mathbb{R}$. The *natural* gradient descent step introduced in Amari [2] is:

$$\eta_{t+1} = \eta_t - \alpha_t \mathcal{H}^{-1}(\eta_t) \nabla \tilde{f}_t(\theta_t), \tag{4}$$

where $\mathcal{H}^{-1}$ is the inverse of the Riemannian mentric $\mathcal{H} = (h_{jk})$. Theorem 1 in [2] proves that the natural gradient algorithm steps in the direction of steepest descent along the Riemannian manifold $(\mathcal{M}, \mathcal{H})$. Hence the name natural gradient descent. The choice of notation η both to parameterize elements of the Riemannian manifold $\mathcal{M}$ and the dual co-ordinates $\eta = g(\theta)$ is intended to pre-empt the link between natural gradient descent and mirror descent.

2.3 Bregman divergences and Riemannian manifolds

Now we explain how every Bregman divergence and its dual induces a pair of Riemannian manifolds as described in Amari and Cichocki [3]. For the Bregman divergence $B_G : \Theta \times \Theta \to \mathbb{R}^+$ induced by the convex function G, define the Riemannian metric on Θ, $\mathcal{G} = \nabla^2 G$ (i.e. the Hessian matrix). Since G is a strictly convex twice differentiable function, $\nabla^2 G(\theta)$ is a positive definite matrix for all $\theta \in \Theta$. Hence $B_G(\cdot, \cdot)$ induces the Riemannian manifold $(\Theta, \nabla^2 G)$. Now let Φ be the image of Θ under the continuous map $g = \nabla G$. $B_H : \Phi \times \Phi \to \mathbb{R}^+$ induces a Riemannian manifold $(\Phi, \mathcal{H})$, where $\mathcal{H} = \nabla^2 H$. Let $(\Theta, \nabla^2 G)$ denote the *primal* Riemannian manifold and $(\Phi, \nabla^2 H)$ denote the *dual* Riemannian manifold.

3 Equivalence of algorithms

In this section we present our main result, the equivalence of mirror descent and natural gradient descent. We also discuss consequences and implications.

Theorem 1. *The mirror descent step* (3) *with Bregman divergence defined by G applied to the sequence of functions $(f_t)_{t=0}^{\infty}$ in the space Θ is equivalent to the natural gradient step* (4) *along the dual Riemannian manifold $(\Phi, \nabla^2 H)$.*

The proof follows by stating mirror descent in the dual Riemannian manifold and simple applications of the chain rule.

Proof. Recall that the mirror descent update is:

$$\theta_{t+1} = \arg\min_{\theta} \left\{ \langle \theta, \nabla f_t(\theta_t) \rangle + \frac{1}{\alpha_t} B_G(\theta, \theta_t) \right\}.$$

Finding the minimum by differentiation yields the step:

$$g(\theta_{t+1}) = g(\theta_t) - \alpha_t \nabla_\theta f_t(\theta_t),$$

where $g = \nabla G$. In terms of the dual variable $\eta = g(\theta)$ and noting that $\theta = h(\eta) = \nabla H(\eta)$,

$$\eta_{t+1} = \eta_t - \alpha_t \nabla_\theta f_t(h(\eta_t)).$$

Applying the chain rule to $\nabla_\eta f_t(h(\eta)) = \nabla_\eta h(\eta) \nabla_\theta f_t(h(\eta))$ implies that

$$\nabla_\theta f_t(h(\eta_t)) = [\nabla_\eta h(\eta_t)]^{-1} \nabla_\eta f_t(h(\eta_t)).$$

Therefore

$$\eta_{t+1} = \eta_t - \alpha_t [\nabla^2 H(\eta_t)]^{-1} \nabla_\eta f_t(h(\eta_t)),$$

which corresponds to the natural gadient descent step. This completes the proof. □

In the subsequent sections, we discuss how this connection directly yields optimal efficiency results for mirror descent and discuss connections to other online algorithm on Riemannian manifolds.

$G(\theta)$	$B_G(\theta, \theta')$	Family
$\frac{1}{2}\|\theta\|_2^2$	$\frac{1}{2}\|\theta - \theta'\|_2^2$	Gaussian
$\exp(\theta)$	$\exp(\theta) - \exp(\theta') - \langle \exp(\theta'), \theta - \theta' \rangle$	Poisson
$\log(1 + \exp(\theta))$	$\log\left(\frac{1+e^\theta}{1+e^{\theta'}}\right) - \langle \frac{e^{\theta'}}{1+e^{\theta'}}, \theta - \theta' \rangle$	Bernoulli

Table 3: Exponential famililes and Bregman divergences

3.1 Efficient parameter estimation in exponential families

In this section we exploit the connection between mirror descent and natural gradient descent to study the efficiency of mirror descent from a statistical perspective. Prior work on the statistical theory of mirror descent has largely focussed on regret analysis and we are not aware of analysis on statistical efficiency. We will see that Fisher efficiency [14, 15, 19] which is an optimality criterian on the covariance of a parameter estimate is an immediate consequence of the equivalence between mirror descent and natural gradient descent.

The statistical problem we consider is parameter estimation in exponential families. Consider a *natural parameter* exponential family with density:

$$p(y \mid \theta) = h(y) \exp(\langle \theta, y \rangle - G(\theta)),$$

where $\theta \in \mathbb{R}^p$ and $G : \mathbb{R}^p \to \mathbb{R}$ is a strictly convex differentiable function. The probability density function can be re-expressed in terms of the Bregman divergence $B_G(\cdot, \cdot)$ as follows:

$$p(y \mid \theta) = \tilde{h}(y) \exp(-B_G(\theta, h(y))),$$

where recall that $h = \nabla H$ and H is the conjugate dual function of G. The distribution can be expressed in terms of the *mean parameter* $\eta = g(\theta)$ and the dual Bregman divergence $B_H(\cdot, \cdot)$:

$$p(y \mid \eta) = \tilde{h}(y) \exp(-B_H(y, \eta)).$$

There is a one-to-one correspondence between exponential families and Bregman divergence [6, 7]. Table 3 displays the exponential families corresponding to the Breman divergences in Table 2.

Consider the mirror descent update for the natural parameter θ with proximty function $B_G(\cdot, \cdot)$ when the function to be minimized is the standard log loss:

$$f_t(\theta; y_t) = -\log p(y_t \mid \theta) = B_G(\theta, h(y_t)).$$

Then the mirror descent step is:

$$\theta_{t+1} = \arg\min_\theta \left\{ \langle \theta, \nabla_\theta B_G(\theta, h(y_t))|_{\theta=\theta_t} \rangle + \frac{1}{\alpha_t} B_G(\theta, \theta_t) \rangle \right\}. \tag{5}$$

Now if we consider the natural gradient descent step for the mean parameter η, the function to be minimized is again the standard log-loss in the η co-ordinates:

$$\tilde{f}_t(\eta; y_t) = -\log p(y_t \mid \eta) = B_H(y_t, \eta).$$

Using Theorem 1 (or by showing it directly), the natrual gradient step is:

$$\eta_{t+1} = \eta_t - \alpha_t [\nabla^2 H]^{-1} \nabla B_H(y_t, \eta_t). \tag{6}$$

A parallel argument holds if the mirror descent step was expressed in terms of the mean parameter and the natural gradient step in terms of the natural parameter.

Now we use Theorem 2 in Amari [2] to prove that mirror descent yields an asymptotically Fisher efficient for η. The Cramér-Rao theorem states that any unbiased estimator based on T independent samples $y_1, y_2, ..., y_T$ of η, which we denote by $\widehat{\eta}_T$ satisfies the following lower bound:

$$\mathbb{E}[(\widehat{\eta}_T - \eta)(\widehat{\eta}_T - \eta)^T] \succeq \frac{1}{T}\nabla^2 H,$$

where $\succeq$ refers to the standard matrix inequality. A sequence of estimators $(\widehat{\eta}_t)_{t=1}^{\infty}$ is asymptotically Fisher efficient if:

$$\lim_{T\to\infty} T\mathbb{E}[(\widehat{\eta}_T - \eta)(\widehat{\eta}_T - \eta)^T] \to \nabla^2 H.$$

Now by using a result from Amari [2] for natural gradient descent, we prove that mirror descent is Fisher efficient, thus achieving the same asymptotic efficiency as any batch-mode method. The following corollary is a direct consequence of Theorem 2 in Amari [2].

Corollary 1. *The mirror descent step applied to the log loss* (5) *with step-sizes* $\alpha_t = \frac{1}{t}$ *asymptotically achieves the Cramér-Rao lower bound.*

For a more detailed discussion on the statistical properties of natural gradient see Amari [2]. Here we have illustrated how the equivalence between mirror descent with Bregman divergences and natural gradient descent gives second-order optimality properties of mirror descent.

3.2 Connection to other online methods on Riemannian manifolds

In this section, we discuss connections between natural gradient descent, mirror descent and online algorithms developed in Bonnabel [9] that directly use Riemannian manifold structures. To define the online gradient descent step for general Riemannian manifolds used in Bonnabel [9], we need to define the exponential map and differentiation in curved spaces.

The *exponential map* at a point $\eta \in \mathcal{M}$ is a map $\exp_\eta : T_\eta\mathcal{M} \to \mathcal{M}$ where $T_\eta\mathcal{M}$ is the tangent space at each point $\eta \in \mathcal{M}$ (see e.g. [11]). Consider the geodesic curve $\gamma : [0,1] \to \mathcal{M}$, with $\gamma(0) = \eta$ and $\dot{\gamma}(0) = v$, where $v \in T_\eta\mathcal{M}$ then $\exp_\eta(v) = \gamma(1)$.

Now we define differentiation. Let $f : \mathcal{M} \to \mathbb{R}$ be a differentiable function on $\mathcal{M}$. The gradient vector field $\bigtriangledown_{\mathcal{M}} f$ takes the form $\bigtriangledown_{\mathcal{M}} f(\eta) = \bigtriangledown_v (f(\exp_\eta(v)))|_{v=0}$ noting that $f(\exp_\eta(v))$ is a smooth function on $T_\eta\mathcal{M}$.

For the sequence of functions $\{f_t\}_{t=0}^{\infty}$ where $f_t : \mathcal{M} \to \mathbb{R}$ the online gradient descent step analyzed in Bonnabel [9] is:

$$\eta_{t+1} = \exp_{\eta_t}(-\alpha_t \nabla_{\mathcal{M}} f_t(\eta_t)). \tag{7}$$

The key reason why the update (7) is the standard gradient descent step instead of the natural gradient descent step introduced by Amari is that η_{t+1} is always guaranteed to lie on the manifold $\mathcal{M}$ for (7), but not for the natural gradient descent step. Unfortunately, the exponential map is extremely difficult to evaluate in general since it is the solution of a system of second-order differential equations [11].

Consequently a standard strategy is to use a computable *retraction* $R_\eta : T_\eta\mathcal{M} \to \mathbb{R}^p$ of the exponential map which yields the approximate gradient descent step:

$$\eta_{t+1} = R_{\eta_t}(-\alpha_t \nabla_{\mathcal{M}} f_t(\eta_t)). \tag{8}$$

The retraction $R_\eta(v) = \eta + v$ corresponds to the first-order Taylor approximation of the exponential map and yields the natural gradient descent step in Amari [2]. Therefore as pointed out in Bonnabel [9], natural gradient descent can be cast as an approximation to gradient descent for Riemannian manifolds. Consequently mirror descent can be viewed as an easily computable first-order approximation to gradient descent for any Riemannian manifold induced by a Bregman divergence.

4 Discussion

In this paper we prove that mirror descent with proximity function Ψ equal to a Bregman divergence is equivalent to the natural gradint descent algorithm along the dual Riemannian manifold. Based on this equivalence, we use results developed by Amari [2] to conclude that mirror descent is the direction of steepest in the corresponding Riemannian space and for parameter estimation in exponential families with the associated Bregman divergence, mirror descent achieves the Cramér-Rao lower bound. Furthermore, this connection proves that the natural gradient step can be implemented as a first-order method using mirror descent which has computational gains for larger datasets.

Following on from this connection, there are a number of interesting and open directions. Firstly, one of the important issues for any online learning algorithm is choice of step-size. Using the connection between mirror descent and natural gradient, it would be interesting to determine whether adaptive choices of step-sizes proposed in Amari [2] that exploit the Riemannian structure can improve performance of mirror descent. It would also be useful to determine a precise characterization of the geometry of mirror descent for other proximity functions such as ℓ_p-norms and explore links online algorithms such as projected gradient descent.

Acknowledgements

GR was partially supported by the NSF under Grant DMS-1127914 to the Statistical and Applied Mathematical Sciences Institute. SM was supported by grants: NIH (Systems Biology): 5P50-GM081883, AFOSR: FA9550-10-1-0436, and NSF CCF-1049290.

References

[1] J. Abernethy, A. Agarwal, P. Bartlett, and A. Rakhlin. A stochastic view of optimal regret through minimax duality. In *Proceedings of COLT*, 2009.

[2] S. Amari. Natural gradient works efficiently in learning. *Neural Computation*, 10(2):251–276, February 1998.

[3] S. Amari and A. Cichocki. Information geometry of divergence functions. *Bulletin of the Polish Academy of Sciences: Technical Sciences*, 58(1):183–195, March 2010.

[4] S. Amari and H. Nagaoka. *Methods of information geometry*. AMS, Providence, RI, 2000.

[5] S.-I Amari, O. E. Barndoff-Nielsen, R. E. Kass, S. L. Lauritzen, and C. R. Rao. *Differential geometry in statistical inference*. IMS Lecture Notes - Monograph Series, Hayward, CA, 1987.

[6] K. S. Azoury and M. K. Warmuth. Relative loss bounds for on-line density estimation with the exponential family of dsitributions. *Machine Learning*, 43(3):211–246, 2001.

[7] A. Banerjee, S. Merugu, I. S. Dhillon, and J. Ghosh. Clustering with bregman divergences. *Journal of Machine Learning Research*, 6:1705–1749, 2005.

[8] A. Beck and M. Taboulle. Mirror descent and nonlinear projected subgradient methods for convex optimization. *Operations Research Letters*, 31:167–175, 2003.

[9] S. Bonnabel. Stochastic gradient descent on Riemannian manifiolds. Technical report, Mines Paris Tech, November 2011.

[10] L. M. Bregman. The relaxation method for finding the common point of convex sets and its application to the solution of problems in convex programming. *USSR Computational Mathematics and Mathematical Physics*, 7:191–204, 1967.

[11] M. P. Do Carmo. *Riemannian Geometry*. Springer Series in Statistics. Birkhauser, Boston, MA, 1992.

[12] H. Cramér. *Mathematical Methods of Statistics*. Princeton University Press, 1946.

[13] J. Duchi, S. Shalev-Shwartz, Y. Singer, and A. Tewari. Composite objective mirror descent. In *Conference on Learning Theory (COLT)*, 2010.

[14] B. Efron. Defining the Curvature of a Statistical Problem (with Applications to Second Order Efficiency). *Annals of Statistics*, 3(6):1189–1242, 1975.

[15] R.A. Fisher. Theory of statistical estimation. *Mathematical Proceedings of the Cambridge Philosophical Society*, 22:700–725, 1925.

[16] J. Kiefer and J. Wolfowitz. Stochastic estimation of the maximum of a regression function. *Annals of Mathematical Statistics*, 23:452–466, 1952.

[17] A. Nemirovski and D. Yudin. *Problem Complexity and Method Efficiency in Optimization*. Wiley, 1983.

[18] C. R. Rao. Information and accuracy obtainable in the estimation of statistical parameters. *Bulletin Calcutta Math. Soc.*, 37:81–91, 1945.

[19] C.R. Rao. Asymptotic efficiency and limiting information. In *Fourth Berkeley Symposium on Mathematical Statistics and Probability*, volume 1, pages 531–546. 1961.

[20] H. Robbins and S. Monro. A stochastic optimization method. *Annals of Mathematical Statistics*, 22:400–407, 1951.

[21] R. T. Rockafeller. *Convex Analysis*. Princeton University Press, Princeton, 1970.

[22] S. Shalev-Shwartz. Online learning and online convex optimization. *Foundations and Trends in Machine Learning*, 4(2):107–194, 2011.

[23] N. Srebro, K. Sridharan, and A. Tewari. On the universality of online mirror descent. In *Proceedings of the NIPS Conference*, Vancouver, Canada, 2011.