Mathematical Institute

Master's Thesis

Statistical Science for the Life and Behavioural Sciences

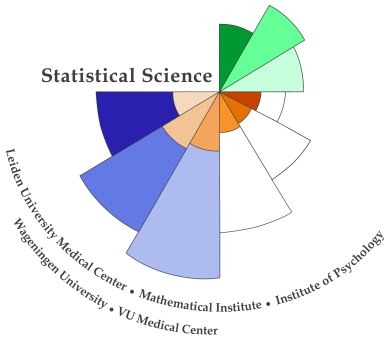
Is Mirror Descent a special case of Exponential Weights?

Author:
Dirk van der Hoeven

Supervisor:
Dr. Tim van Erven
Universiteit Leiden

August 2016

Universiteit Leiden
The Netherlands



Abstract

Online Convex Optimization is a setting in which a forecaster is to sequentially predict outcomes. In this thesis we focus on two algorithms in the Online Convex Optimization setting, namely Mirror Descent and Exponential Weights. Exponential Weights is usually seen as a special case of Mirror Descent. However, we developed an interpretation of Exponential Weights that sees Mirror Descent as the mean of Exponential Weights, and thus a special case of Exponential Weights. Specifically, different priors for Exponential Weights lead to different Mirror Descent algorithms. The link between Exponential Weights and Mirror Descent hinges on the link between cumulant generating functions, related to the prior in Exponential Weights, and Legendre functions, related to the update step in Mirror Descent.

Contents

Chapter 1

Introduction

Imagine you are a gambler that predicts outcomes of football games coming season. Say you know ten experts that also predict the outcome of football games and you want to take their advice. However, you do not know how much to listen to whom, but you do know how bad or how good the expert predictions were so far. You want minimize your regret at the end of the season for listening to the experts, since you want to make as much money gambling as possible. This setting roughly describes the *Prediction with Expert Advice* setting, which is a special case of the *Online Convex Optimization* setting .

Online Convex Optimization is a sequential prediction setting that proceeds in rounds in which a forecaster is to predict an unknown sequence of elements. In each round the forecaster suffers a convex loss, which accumulates over rounds. An example of a problem that can be modeled in the Online Convex Optimization setting is given by Hazan (2015). Consider your email service, which has a spam filter. During the day the spam filter has to classify emails as spam or valid. The spam filter may represent each email as a vector $\boldsymbol{x} \in \{0,1\}^d$, where the number of dimensions d is the number of words in the dictionary. The elements of the vector are all zero unless a word corresponding to an element in the vector occurs in the email. The spam filter learns a filter, for example a vector $\boldsymbol{a} \in \mathbb{R}^d$. An email $\boldsymbol{x}$ is now classified by the sign of the inner product between $\boldsymbol{a}$ and $\boldsymbol{x}$: $\hat{y} = \text{sign}\langle \boldsymbol{x}, \boldsymbol{a} \rangle$, where $\langle \boldsymbol{x}, \boldsymbol{a} \rangle$ denotes the inner product between vectors $\boldsymbol{x}$ and $\boldsymbol{a}$ and with $\hat{y} = 1$ meaning spam and $\hat{y} = -1$ meaning valid. Assuming we know the true label $y \in \{-1, 1\}$ after classification the spam filter now receives a loss, which we restrict to convex loss functions, for example the square loss: $\hat{\ell}(\hat{y}, y) = (\hat{y} - y)^2$. At the end of the day we can now measure something called *regret*, which embodies how sorry one feels for choosing a given filter as opposed the best filter. The goal of algorithms in the Online Convex Optimization setting is to minimize regret.

In this thesis we focus on three such algorithms, namely the *Exponential Weights*, *Mirror Descent* and *Online Gradient Descent* algorithms. In the literature, the Online Gradient Descent and Exponential Weights algorithms are known as special cases of the Mirror Descent algorithm. However, a recent observation by Koolen (2016) changed the understanding of the relationship between Exponential Weights and Online Gradient Descent: Online Gradient Descent may also be viewed as a special case of the Exponential Weights algorithm. This raises the question of whether other Mirror Descent type algorithms are also special cases of Exponential Weights. Mirror Descent is a class of algorithms that maps parameters to a different "mirror" space by so-called Legendre functions and performs the optimization in the mirror space. Exponential Weights may be initialized with different choices of priors. In this thesis we show that the Mirror descent algorithm can be seen as the mean of the Exponential Weights algorithm. We show that the prior for Exponential Weights is related to the Legendre function in Mirror Descent: when a Legendre function is a cumulant generating function of an exponential family, a member of this exponential family is the prior for Exponential Weights and vice versa. This provides unification and understanding of a large class of algorithms. Furthermore, this interpretation of the Exponential Weights algorithm greatly reduces its computational complexity, which makes it applicable in the Online Convex Optimization setting as opposed to before this interpretation.

Outline As for the organization of the thesis: in chapter 2 we give a formal introduction to the Online Convex Optimization and Prediction with Expert Advice settings to detail the Online Gradient Descent, Mirror Descent and Exponential Weigths algorithm. In chapter 3 the proof of Koolen (2016) is replicated. In chapter 4 the main result of this thesis is presented in Theorem 3: we prove that, under one condition, the MD algorithm is a special case of the EW algorithm. In chapter 5 we explore when a Legendre function is a cumulant generating function, the condition for Theorem 3. Furthermore, two new and constructive sufficient conditions for which the relationship between MD and EW holds are given in Theorems 7 and 8. Finally, in chapter 6 some examples of the relationship between the EW and MD algorithms are given.

Chapter 2

Online Convex Optimization

The analysis of many efficient online learning tools has been influenced by convex optimization tools. Most efficient algorithms can be analyzed based on the following model, which summarizes the Online Convex Optimization setting (OCO) setting (Shalev-Shwartz, 2011):

Input: A convex set $S \subset \mathbb{R}^d$
For $t = 1, 2, \ldots, T$
    predict a vector $\boldsymbol{w}_t \in S$
    receive a convex loss function $f_t : S \to \mathbb{R}$
    suffer loss $f_t(\boldsymbol{w})$

For example, as in the spam filter above, say we receive outcomes $y_t \in \{-1, 1\}$ and predict a weight vector (filter) $\boldsymbol{w}_t$ that classifies incoming emails $\boldsymbol{x}_t$ at time t, $\hat{y}_t = \text{sign}\langle \boldsymbol{x}_t, \boldsymbol{w}_t \rangle$. For simplicity we may use $f_t(\boldsymbol{w}) = (\hat{y} - y_t)^2$ to evaluate our performance in a given round, which is convex. At the end of the day, we hope that we chose the best filter possible, which would have minimized the loss function.

This chapter will describe some algorithms working in this model. We focus on analyzing the regret, which is measured with respect to a reference weight vector $\overline{\boldsymbol{w}} \in S$ (Shalev-Shwartz, 2011):

$$\mathcal{R}_T(\overline{\boldsymbol{w}}) = \sum_{t=1}^{T} f_t(\boldsymbol{w}_t) - \sum_{t=1}^{T} f_t(\overline{\boldsymbol{w}}). \tag{2.1}$$

An example of a reference weight vector $\overline{\boldsymbol{w}}$ would be the minimizer of the cumulative loss, $\sum_{t=1}^{T} f_t(\boldsymbol{w})$. We may interpret regret by how sorry a forecaster is for choosing $\boldsymbol{w}_1, \ldots, \boldsymbol{w}_T$ instead of choosing $\overline{\boldsymbol{w}}$. The goal is to find an algorithm

that has regret that grows sub-linear with T with respect to any reference weights. Under appropriate conditions this is achievable by the Online Gradient Descent (OGD) algorithm (Zinkevich, 2003). Without loss of generality we assume that S is centered around 0. The OGD algorithm initializes $\boldsymbol{w}_t$ as a zero vector $\mathbf{0}$, then updates these weights with the following update rule:

$$\boldsymbol{w}_{t+1} = \boldsymbol{w}_t - \eta \nabla f_t(\boldsymbol{w}_t), \tag{2.2}$$

with learning rate parameter $\eta > 0$, which controls how fast the algorithm learns, and where $\nabla f_t(\boldsymbol{w}_t)$ denotes the gradient of f_t at $\boldsymbol{w}_t$ (hence online *gradient* descent). Intuitively, the OGD algorithm moves $\boldsymbol{w}_t$ in the direction of the minimum of f_t, but not by too much because it wants to remember the effect of $f_1, f_2, \ldots, f_{t-1}$. The OGD algorithm guarantees regret bounded by (Shalev-Shwartz, 2011):

$$\mathcal{R}_T(\overline{\boldsymbol{w}}) \leq \frac{||\overline{\boldsymbol{w}}||_2^2}{2\eta} + \frac{\eta}{2} \sum_{t=1}^{T} ||\nabla f_t(\boldsymbol{w}_t)||_2^2. \tag{2.3}$$

If $||\overline{\boldsymbol{w}}||_2 \leq B$, $||\nabla f_t(\boldsymbol{w}_t)||_2 \leq L$ and $\eta = \frac{B}{L\sqrt{2T}}$ we obtain:

$$\mathcal{R}_T(\overline{\boldsymbol{w}}) \leq BL\sqrt{2T}, \tag{2.4}$$

which grows sub-linearly with T.

A generalization of OGD is the *Mirror Descent* (MD) algorithm. It maps the weights to a different mirror space with the gradient of a function of type Legendre before updating the weights and maps them back with the inverse function of the gradient after the update. A Legendre function has several desirable properties, which will be discussed in chapter 4. Let $\phi = \nabla F^*$ be the gradient of Legendre function F^*. We use F^* instead of F for notational purposes, which will become apparent in section 4.2 when convex conjugates are discussed. The update rule of the MD algorithm is the following (Shalev-Shwartz, 2011):

$$\boldsymbol{w}_{t+1} = \phi^{-1}\Big(\phi(\boldsymbol{w}_t) - \eta \nabla f_i(\boldsymbol{w}_t)\Big), \tag{2.5}$$

where ϕ^{-1} is the inverse function of ϕ. For each different choice of ϕ this leads to a different algorithm with different properties. Choosing ϕ to be the identity function leads to the OGD algorithm. Setting $\phi(\boldsymbol{w}) = \log \boldsymbol{w} + 1, \boldsymbol{w} \in \triangle$, where $\triangle$ denotes the probability simplex and log the natural logarithm (henceforth we shall denote the natural logarithm by log), leads to the *Exponential Weights* (EW) algorithm, which is commonly used in the *Prediction with Expert Advice* setting to be discussed in section 2.1.

2.1 Prediction with Expert Advice

Prediction with Expert Advice (PwEA) is a setting in online prediction where, in each round t, K experts predict an outcome y_t. In each round the forecaster has access to the predictions $\hat{y}_t^k$ of the experts. On the basis of these expert predictions the forecaster forms his own predictions by choosing a probability distribution p_t on the experts. The forecaster's loss becomes $\hat{\ell}_t = \mathop{\mathbb{E}}_{k \sim p_t}[\ell_t^k]$, where $\ell_t^k = \ell(y_t, \hat{y}_t^k)$ is the loss of expert k at time t with respect to outcome y_t and prediction $\hat{y}_t^k$. Loss $\hat{\ell}_t$ can be motivated in several ways:

1. If the forecaster randomly chooses an expert $k \sim p_t$ then this is the expected loss.
2. If ℓ is linear in the second argument and the forecaster predicts $\hat{y}_t = \mathbb{E}_{p_t}[\hat{y}_k^t]$, the mean of the expert predictions, then $\hat{\ell}_t$ is the forecaster's loss.
3. If ℓ is convex in the second argument and the forecaster predicts $\hat{y}_t = \mathbb{E}_{p_t}[\hat{y}_t^k]$ then $\hat{\ell}_t$ is an upper bound on the forecaster's loss.

Note that the PwEA setting is a special case of the OCO setting in which weight vector $p_t \in S$ is a probability distribution over the experts. In other words, the role of $\boldsymbol{w}_t$ in the OCO setting is played by p_t in the PwEA setting. The cumulative loss is minimized by a distribution $\overline{p}$ that is a point mass on the single expert that has the lowest cumulative loss. As with online convex optimization the goal is to achieve a total loss after T rounds that is not much worse than the total loss of the best expert, measured by regret:

$$\mathcal{R}_T(k) = \sum_{t=1}^{T} \hat{\ell}_t - \sum_{t=1}^{T} \ell_t^k. \tag{2.6}$$

As in the online convex optimization setting it is possible to achieve a regret that grows sublinearly with the number of rounds (Shalev-Shwartz, 2011):

$$\mathcal{R}_T(k) \leq \sqrt{\frac{T \log(K)}{2}} \qquad \forall k. \tag{2.7}$$

The most common algorithm in the PwEA setting to achieve this regret bound is the *Exponential Weights* (EW) algorithm (Shalev-Shwartz, 2011):

$$p_{t+1}(k) = \frac{\pi(k)\ \exp(-\eta \sum_{i=1}^{t} \ell_i^k)}{\sum_{k=1}^{K} \pi(k)\ \exp(-\eta \sum_{i=1}^{t} \ell_i^k)}, \tag{2.8}$$

where π is a prior distribution on the experts, usually chosen to be the uniform distribution over the experts such that $\pi(k) = \frac{1}{K}$, and $\eta > 0$ a parameter of the algorithm. $\sum_{k=1}^{K} \pi(k) \exp(-\eta \sum_{i=1}^{t} \ell_i^k)$ is a normalization factor. This algorithm gives high weights to experts with small losses. If $\ell_t^k \in [0, 1]$ then the EW algorithm achieves the following regret bound with a uniform prior (Shalev-Shwartz, 2011):

$$R_T(k) \leq \frac{\log(K)}{\eta} + \frac{\eta T}{8} \qquad \forall k, \tag{2.9}$$

which is a trade-off between the number of experts and rounds, regulated by η. The optimal η is found by $\eta = \sqrt{\frac{\log(K)}{T/8}}$, which leads to (2.7).

Chapter 3

Online Gradient Descent as a special case of Exponential Weights

In this chapter we elaborate on the proof given by Koolen (2016) that shows that OGD is a special case of EW. A non-standard interpretation of the EW algorithm arises if we apply EW with a continuous set of experts and a non-uniform prior π. We now use p_t as the probability distribution over a continuous set of experts parametrized by $\boldsymbol{z} \in \mathbb{R}^d$ on time point t and use the mean of p_t as weights $\boldsymbol{w}_t$. As shown by Koolen (2016), with a normal prior the EW algorithm becomes the Online Gradient Descent algorithm.

We proceed in the OCO setting. In each round t the experts $\boldsymbol{z}$ receive a loss $\ell_t^{\boldsymbol{z}} = \langle \boldsymbol{z}, \boldsymbol{g}_t \rangle$. This loss function is motivated by the following. Let f_t be a convex loss function and let $\tilde{f}_t(\boldsymbol{w}) = \langle \boldsymbol{w}, \boldsymbol{g}_t \rangle$, with $\boldsymbol{g}_t = \nabla f_t(\boldsymbol{w}_t)$. Our regret is then bounded by:

$$\mathcal{R}_t(\overline{\boldsymbol{w}}) = \sum_{i=1}^{t} f_i(\boldsymbol{w}_t) - f_i(\overline{\boldsymbol{w}}) \leq \sum_{i=1}^{t} \tilde{f}_i(\boldsymbol{w}) - \tilde{f}_i(\overline{\boldsymbol{w}}). \tag{3.1}$$

The forecaster's loss $\hat{\ell}_t$ is:

$$\begin{aligned} \hat{\ell}_t &= \mathbb{E}_{\boldsymbol{z} \sim p_t}[\ell_t^{\boldsymbol{z}}] \\ &= \mathbb{E}_{\boldsymbol{z} \sim p_t}[\langle \boldsymbol{z}, \boldsymbol{g}_t \rangle] \\ &= \langle \mathbb{E}_{\boldsymbol{z} \sim p_t}[\boldsymbol{z}], \boldsymbol{g}_t \rangle \\ &= \langle \boldsymbol{w}_t, \boldsymbol{g}_t \rangle. \end{aligned} \tag{3.2}$$

Thus, if we use $\boldsymbol{w}_t = \mathbb{E}_{p_t}[\boldsymbol{z}]$ in the OCO setting the losses in OCO and PwEA are equal. In each round we update probability density p_t with the following:

$$p_{t+1}(\boldsymbol{z}) = \frac{\pi(\boldsymbol{z}) \exp(-\eta \sum_{i=1}^{t} \ell_i^{\boldsymbol{z}})}{\int_{\mathbb{R}^d} \pi(\boldsymbol{z}) \exp(-\eta \sum_{i=1}^{t}, \ell_i^{\boldsymbol{z}}) d\boldsymbol{z}}, \tag{3.3}$$

in which our initial choice of the probability density p_t is represented by prior π. The above leads to the following Theorem, which appears to have been published only as a blog post by Koolen (2016):

Theorem 1. *Let p_{t+1} be the exponential weight distribution at time $t+1$ and $\boldsymbol{w}_{t+1} \in S = \mathbb{R}^d$ be the mean of p_{t+1}. If we choose $\pi(\boldsymbol{z}) = N(\boldsymbol{0}, \sigma^2 \boldsymbol{I})$ and $\ell_i^{\boldsymbol{z}} = \langle \boldsymbol{z}, \boldsymbol{g}_i \rangle$ the Exponential Weights algorithm yields a multivariate normal distribution $p_{t+1} = N(\boldsymbol{w}_{t+1}, \sigma^2 \boldsymbol{I})$ with mean*

$$\begin{aligned} \mathop{\mathbb{E}}_{\boldsymbol{z} \sim p_{t+1}}[\boldsymbol{z}] &= \boldsymbol{w}_{t+1} \\ &= \sigma^2 \eta \sum_{i=1}^{t} \boldsymbol{g}_i \\ &= \boldsymbol{w}_t - \sigma^2 \eta \boldsymbol{g}_t, \end{aligned} \tag{3.4}$$

which are the weights of the Online Gradient Descent algorithm with learning rate $\sigma^2 \eta$.

Proof. The proof is given by plugging in the multivariate normal density function for $\pi = N(\boldsymbol{0}, \sigma^2 \boldsymbol{I})$ and working out the algebra. The multivariate normal density with mean $\boldsymbol{w}$ and covariance matrix $\boldsymbol{\Sigma}$ is defined as:

$$N(\boldsymbol{w}, \boldsymbol{\Sigma}) = (2\pi)^{-\frac{d}{2}} \det(\boldsymbol{\Sigma})^{-\frac{1}{2}} \exp(-\frac{1}{2}(\boldsymbol{z} - \boldsymbol{w})^T \boldsymbol{\Sigma}^{-1} (\boldsymbol{z} - \boldsymbol{w})). \tag{3.5}$$

We start by computing the numerator of equation 3.3:

$$\begin{aligned} \pi(\boldsymbol{z}) \times e^{-\eta \sum_{i=1}^{t} \ell_i^{\boldsymbol{z}}} &= (2\pi)^{-\frac{d}{2}} \det(\sigma^2 \boldsymbol{I})^{-\frac{1}{2}} \exp(\frac{-\langle \boldsymbol{z}, \boldsymbol{z} \rangle - 2\sigma^2 \eta \langle \boldsymbol{z}, \sum_{i=1}^{t} \boldsymbol{g}_i \rangle}{2\sigma^2}) \\ &= (2\pi)^{-\frac{d}{2}} \sigma^{-d} \exp(-\frac{\langle \boldsymbol{z} + \sigma^2 \eta \sum_{i=1}^{t} \boldsymbol{g}_i, \boldsymbol{z} + \sigma^2 \eta \sum_{i=1}^{t} \boldsymbol{g}_i \rangle}{2\sigma^2}) \\ &\quad \exp(\frac{\sigma^2 \langle \eta \sum_{i=1}^{t} \boldsymbol{g}_i, \eta \sum_{t=1}^{i} \boldsymbol{g}_i \rangle}{2}). \end{aligned} \tag{3.6}$$

which is, up to normalization, the density function for the multivariate normal distribution $N(-\eta\sigma^2 \sum_{i=1}^{t} \boldsymbol{g}_i,\ \sigma^2 \boldsymbol{I})$. The mean of this distribution is equal to the updating rule for the gradient descent, with learning rate $\eta^* = \sigma^2\eta$.

□

Hence, by Theorem 1 the OGD algorithm is a special case of the EW algorithm on a continuous set of experts.

Chapter 4

Mirror Descent is a special case of Exponential Weights

In Theorem 1 we showed that OGD is a special case of EW. This raises the question of whether other MD type algorithms are also special cases of the EW algorithm. In this chapter it is proven that, under some conditions, the MD algorithm can be seen as a special case of the EW algorithm. As in Theorem 1 viewing the MD update step (2.5) as the mean of the EW probability distribution yields the relationship between EW and MD. To identify the relation between MD and EW in general, we will need to introduce the following three concepts: the *Follow the Regularized Leader* algorithm, *exponential families* of distributions, and the *minimum relative entropy principle*.

In section 4.1 we introduce the Follow the Regularized Leader algorithm. The Follow the Regularized Leader algorithm is used to see the algorithms as similar minimization problems with different regularization functions. With the Follow the Regularized Leader representation of MD and EW the only difference between MD and EW is the regularization function used. The EW algorithm has the *Kullback Leibler divergence* as regularization function, which is a measure of divergence between two distributions. The MD algorithm has the *Bregman divergence* as regularization function, which can be interpreted as a measure of how convex a function is. Section 4.2 introduces exponential families, which are sets of distributions that can be written in a certain form. This form of exponential families is used in section 4.3 to show a crucial result: the KL divergence between two members of an exponential family reduces to the *Bregman divergence*. Section 4.4 presents a technical result from the minimum relative entropy principle. This result is used to prove that it poses no restriction to minimize the Follow the Regularized Leader representation of EW over exponential families only. Combined

these ideas show that MD is a special case of EW. We continue by introducing the Follow the Regularized Leader algorithm.

4.1 Follow the Regularized Leader

The Follow The Leader (FTL) algorithm and its counterpart the Follow The Regularized Leader (FTRL) algorithms both work in the OCO setting. The FTL algorithm does what its name suggests: it follows the weight vector that had the smallest loss of the preceding rounds:

$$\boldsymbol{w}_{t+1} = \arg\min_{\boldsymbol{w} \in S} \left\{ \sum_{i=1}^{t} f_i(\boldsymbol{w}) \right\}. \tag{4.1}$$

However, the FTL algorithm may lead to a regret that grows linearly with the number of rounds (Shalev-Shwartz, 2011). To overcome this issue the FTRL algorithm was introduced, in which a regularization function is appended to (4.1):

$$\boldsymbol{w}_{t+1} = \arg\min_{\boldsymbol{w} \in S} \left\{ \sum_{i=1}^{t} f_i(\boldsymbol{w}) + R(\boldsymbol{w}) \right\}, \tag{4.2}$$

where $R : S \to \mathbb{R}$ is the regularization function. Different choices for the regularization function lead to different algorithms, of which three are given in the following.

Lemma 1 (Shalev-Shwartz (2011))**.** *Let $S = \mathbb{R}^d$, $\boldsymbol{w}_1 = 0$, and $f_t(\boldsymbol{w}) = \langle \boldsymbol{w}, \boldsymbol{g}_t \rangle$. In the OCO setting Follow The Regularized Leader with regularization $R(\boldsymbol{w}) = \frac{1}{2\eta}||\boldsymbol{w}_t||_2^2$ yields the Gradient Descent algorithm:*

$$\begin{aligned} \boldsymbol{w}_{t+1} &= \arg\min_{\boldsymbol{w}} \left\{ \sum_{i=1}^{t} f_i(\boldsymbol{w}) + \frac{1}{2\eta}||\boldsymbol{w}||_2^2 \right\} \\ &= \boldsymbol{w}_t - \eta \nabla f_t(\boldsymbol{w}_t) \end{aligned} \tag{4.3}$$

The Exponential Weights algorithm can also be seen as a FTRL algorithm. To show this we first introduce the Kullback Leibler (KL) divergence (Kullback and Leibler, 1951). The KL divergence between two discrete distributions $p, \pi \in \triangle$,

where $\triangle$ is the probability simplex, is defined as:

$$\begin{aligned} KL(p||\pi) &= \sum_{k=1}^{K} p(k) \log \frac{p(k)}{\pi(k)} \\ &= \mathop{\mathbb{E}}_{k\sim p}[\log p(k) - \log \pi(k)]. \end{aligned} \tag{4.4}$$

For continuous distributions the summation in (4.4) becomes an integral. The KL divergence (also known as relative entropy) is a measure of difference between two distributions. It is not a distance, as it does not obey the triangle inequality nor does $KL(p||\pi) = KL(\pi||p)$ hold in general.

Lemma 2 (Shalev-Shwartz (2011))**.** *In the PwEA setting Follow The Regularized Leader with regularization $R(p) = \frac{1}{\eta} KL(p||\pi)$ yields the Exponential Weights algorithm:*

$$\begin{aligned} p_{t+1} &= \arg\min_{p\in\triangle} \left\{ \sum_{i=1}^{t} \mathbb{E}_{k\sim p}[\ell_t^k] + \frac{1}{\eta} KL(p||\pi) \right\} \\ &= k \mapsto \frac{\pi(k) \exp(-\eta \sum_{i=1}^{t} \ell_i^k)}{\sum_{k=1}^{K} \pi(k) \exp(-\eta \sum_{i=1}^{t} \ell_i^k)}. \end{aligned} \tag{4.5}$$

The mirror descent algorithm is a FTRL algorithm if we use the Bregman divergence (Bregman, 1967) in combination with learning parameter η as the regularization function. Bregman divergences are generated by Legendre functions, which are introduced in the following. A function $F^* : S \to \mathbb{R}$ is called Legendre if it obeys the following (Cesa-Bianchi and Lugosi, 2006, Chapter 11):

1. $S \subset \mathbb{R}^d$ is nonempty and its interior is convex.
2. F^* is strict convex with continuous first partial derivatives throughout the interior of S.
3. if $x_1, x_2, \ldots, x_n \in S$ is a sequence converging to a boundary point in S, then $||\nabla F^*(x_n)|| \to \infty$ as $n \to \infty$.

The Bregman divergence generated by Legendre function F^* for vectors $\boldsymbol{x}, \boldsymbol{y} \in S$ is defined as:

$$B_{F^*}(\boldsymbol{x}||\boldsymbol{y}) = F^*(\boldsymbol{x}) - F^*(\boldsymbol{y}) - (\boldsymbol{x} - \boldsymbol{y}) \nabla F^*(\boldsymbol{y}). \tag{4.6}$$

Note that the Bregman divergence is not a distance in general, as it does not satisfy the triangle inequality nor is it symmetric in its arguments. The Bregman

divergence may be interpreted as the difference between $F^*(\boldsymbol{x})$ and the tangent of F^* at $\boldsymbol{y}$. In other words, the Bregman divergence is a measure of how convex F^* is.

Lemma 3 (Shalev-Shwartz (2011))**.** *Let $S = \mathbb{R}^d$ and $f_t(\boldsymbol{w}) = \langle \boldsymbol{w}, \boldsymbol{g}_t \rangle$. In the OCO setting Follow The Regularized Leader with regularization $R(\boldsymbol{w}) = \frac{1}{\eta} B_{F^*}(\boldsymbol{w}||\boldsymbol{w}_0)$, where $\boldsymbol{w}_0$ denotes the starting weights for the algorithm, yields the Mirror Descent algorithm:*

$$\begin{aligned} \boldsymbol{w}_{t+1} &= \arg\min_{\boldsymbol{w}} \left\{ \sum_{i=1}^{t} f_i(\boldsymbol{w}) + \frac{1}{\eta} B_{F^*}(\boldsymbol{w}||\boldsymbol{w}_0) \right\} \\ &= \phi^{-1}\Big(\phi(\boldsymbol{w}_t) - \eta \nabla f_i(\boldsymbol{w}_t)\Big), \end{aligned} \tag{4.7}$$

with $\phi = \nabla F^*$.

As before, both the EW and GD algorithm are usually seen as special cases of the MD algorithm. If we set F^* to the negative entropy $F^*(p) = \sum_k p(k) \log p(k)$ and the second argument to some arbitrary distribution π we obtain the KL divergence with respect to π. In turn, this yields the FTRL representation of the Exponential Weights algorithm. If we set F^* to half the squared L_2 norm ($F^*(\boldsymbol{w}) = \frac{1}{2}||\boldsymbol{w}||_2^2$) and set the second argument of the Bregman Divergence to a zero vector we obtain the FTRL representation of the Gradient Descent algorithm.

4.2 Exponential families

In order to show that the KL divergence between two members of the same exponential family reduces to the Bregman divergence we first need to introduce exponential families. Exponential families are sets of distributions that share important properties and can be written in a certain form. The multivariate normal and many other common distributions such as the multinomial, gamma and Poisson distributions are exponential families. The following will introduce exponential families and some of their properties.

An exponential family can be written in the following form:

$$p_{\boldsymbol{\theta}}(\boldsymbol{z}) = e^{\langle \boldsymbol{\theta}, T(\boldsymbol{z}) \rangle - F(\boldsymbol{\theta})} K(\boldsymbol{z}). \tag{4.8}$$

Here, $\boldsymbol{\theta}$ is called the natural parameter vector coming from the non-empty, convex, open parameter space Θ. We use $T(\boldsymbol{z})$ to denote the sufficient statistic of the distribution, which completely summarizes the data to recover the density function.

Furthermore, we require $T(\boldsymbol{z})$ to be minimal, which is to say that the components of $T(\boldsymbol{z})$ are affinely independent, i.e., $\nexists$ a non-zero $\boldsymbol{a} \in \mathbb{R}^d$ such that $\langle \boldsymbol{a}, T(\boldsymbol{z}) \rangle = b$ (a constant). Furthermore, $K(\boldsymbol{z})$ is called the carrier measure and $F(\boldsymbol{\theta})$ is known as the cumulant generating function. Cumulant generating function F is a function of Legendre type (Barndorff-Nielsen, 1978, Theorems 8.2, 9.1, and 9.3). We may interpret F as the logarithm of the normalization factor:

$$\begin{aligned} \int_{\mathbb{R}^d} e^{\langle \boldsymbol{\theta}, T(\boldsymbol{z}) \rangle - F(\boldsymbol{\theta})} K(\boldsymbol{z}) d\boldsymbol{z} = 1 \\ \Leftrightarrow F(\boldsymbol{\theta}) = \log \int_{\mathbb{R}^d} e^{\langle \boldsymbol{\theta}, T(\boldsymbol{z}) \rangle} K(\boldsymbol{z}) d\boldsymbol{z}. \end{aligned} \tag{4.9}$$

The expectation of a member of an exponential family and its natural parameters have a one to one relationship (Barndorff-Nielsen, 1978, Theorem 8.1), which can be shown using conjugate functions. The expectation of a member of an exponential family is computed by (Barndorff-Nielsen, 1978, Theorem 8.1): $\boldsymbol{\mu} = \mathop{\mathbb{E}}_{\boldsymbol{z} \sim p_{\boldsymbol{\theta}}} [T(\boldsymbol{z})] = \nabla F(\boldsymbol{\theta})$. The conjugate function of F, F^*, is given by:

$$F^*(\boldsymbol{\mu}) = \sup_{\boldsymbol{\theta}} \{ \langle \boldsymbol{\theta}, \boldsymbol{\mu} \rangle - F(\boldsymbol{\theta}) \}, \tag{4.10}$$

where sup denotes the supremum and $\boldsymbol{\mu}$ represents the expectation vector of $p_{\boldsymbol{\theta}}$. Since F is a Legendre function, F^* is also a Legendre function (Cesa-Bianchi and Lugosi, 2006, chapter 11). The conjugate function of F^* is again F (Cesa-Bianchi and Lugosi, 2006, chapter 11). Note that $\langle \boldsymbol{\theta}, \boldsymbol{\mu} \rangle \leq F(\boldsymbol{\theta}) + F^*(\boldsymbol{\mu})$, with equality if and only if $\boldsymbol{\mu} = \nabla F(\boldsymbol{\theta})$.

In section 4.1 it was mentioned that F^* was used instead of F to generate Bregman divergences for notational purposes. This is because we actually use the convex conjugate of F to generate the Bregman divergence. We do this to exploit the relationship between the KL divergence for two members of the same exponential family and the Bregman divergence, which is detailed in Theorem 2 below.

4.3 KL divergence is Bregman divergence

The above definition of exponential families leads to the following crucial connection between the Kullback-Leibler divergence between two members of the same exponential family and the Bregman divergence.

Theorem 2 (Banerjee et al. (2005); Nielsen and Nock (2010))**.** *The KL divergence between two members of the same exponential family, $p_{\boldsymbol{\theta}}$ and $\pi_{\boldsymbol{\theta}}$, with cumulant generating function F can be expressed by the Bregman divergence between their natural parameters, $\boldsymbol{\theta}_p$ and $\boldsymbol{\theta}_\pi$, or their expectation parameters, $\boldsymbol{\mu}_p$ and $\boldsymbol{\mu}_\pi$. The first Bregman divergence is generated by the cumulant generating function F and the secend Bregman divergence is generated by the convex conjugate of the cumulant generating function F^*:*

$$KL(p_{\boldsymbol{\theta}}||\pi_{\boldsymbol{\theta}}) = B_F(\boldsymbol{\theta}_\pi||\boldsymbol{\theta}_p) = B_{F^*}(\boldsymbol{\mu}_p||\boldsymbol{\mu}_\pi). \tag{4.11}$$

The proof follows from computing the KL divergence between two members of the same exponential family:

Proof.

$$\begin{aligned}
KL(p_{\boldsymbol{\theta}}||\pi_{\boldsymbol{\theta}}) &= \mathop{\mathbb{E}}_{z\sim p_{\boldsymbol{\theta}}}\Big[\log p_{\boldsymbol{\theta}}(\boldsymbol{z}) - \log \pi_{\boldsymbol{\theta}}(\boldsymbol{z})]\Big] \\
&= \mathop{\mathbb{E}}_{z\sim p_{\boldsymbol{\theta}}}\Big[\langle \boldsymbol{\theta}_p, T(\boldsymbol{z})\rangle - F(\boldsymbol{\theta}_p) + \log(K(\boldsymbol{z})) - \langle \boldsymbol{\theta}_\pi, T(\boldsymbol{z})\rangle + F(\boldsymbol{\theta}_\pi) - \log(K(\boldsymbol{z}))\Big] \\
&= \langle \boldsymbol{\theta}_p - \boldsymbol{\theta}_\pi, \mathop{\mathbb{E}}_{z\sim p_{\boldsymbol{\theta}}}[T(\boldsymbol{z})]\rangle - F(\boldsymbol{\theta}_p) + F(\boldsymbol{\theta}_\pi) \\
&= F(\boldsymbol{\theta}_\pi) - F(\boldsymbol{\theta}_p) - \langle \boldsymbol{\theta}_\pi - \boldsymbol{\theta}_p, \nabla F(\boldsymbol{\theta}_p)\rangle \\
&= B_F(\boldsymbol{\theta}_\pi||\boldsymbol{\theta}_p).
\end{aligned} \tag{4.12}$$

To show that $B_F(\boldsymbol{\theta}_\pi||\boldsymbol{\theta}_p) = B_{F^*}(\boldsymbol{\mu}_p||\boldsymbol{\mu}_\pi)$ we use the convex conjugate of F:

$$\begin{aligned}
B_F(\boldsymbol{\theta}_\pi||\boldsymbol{\theta}_p) &= F(\boldsymbol{\theta}_\pi) - F(\boldsymbol{\theta}_p) - \langle \boldsymbol{\theta}_\pi - \boldsymbol{\theta}_p, \nabla F(\boldsymbol{\theta}_p)\rangle \\
&= \langle \boldsymbol{\theta}_\pi, \boldsymbol{\mu}_\pi\rangle - F^*(\boldsymbol{\mu}_\pi)\big) - \big(\langle \boldsymbol{\theta}_p, \boldsymbol{\mu}_p\rangle - F^*(\boldsymbol{\mu}_p)\big) - \langle \boldsymbol{\theta}_\pi - \boldsymbol{\theta}_p, \boldsymbol{\mu}_p\rangle \\
&= F^*(\boldsymbol{\mu}_p) - F^*(\boldsymbol{\mu}_\pi) + \langle \boldsymbol{\theta}_\pi, \boldsymbol{\mu}_\pi\rangle - \langle \boldsymbol{\theta}_p, \boldsymbol{\mu}_p\rangle - \langle \boldsymbol{\theta}_\pi, \boldsymbol{\mu}_p\rangle + \langle \boldsymbol{\theta}_p, \boldsymbol{\mu}_p\rangle \\
&= F^*(\boldsymbol{\mu}_p) - F^*(\boldsymbol{\mu}_\pi) - \langle \boldsymbol{\mu}_p - \boldsymbol{\mu}_\pi, \boldsymbol{\theta}_\pi\rangle \\
&= F^*(\boldsymbol{\mu}_p) - F^*(\boldsymbol{\mu}_\pi) - \langle \boldsymbol{\mu}_p - \boldsymbol{\mu}_\pi, \nabla F^*(\boldsymbol{\mu}_\pi)\rangle \\
&= B_{F^*}(\boldsymbol{\mu}_p||\boldsymbol{\mu}_\pi)
\end{aligned} \tag{4.13}$$

□

4.4 Minimum relative entropy principle

To apply Theorem 2 to Exponential Weights we need to show that a member of an exponential family reaches the minimum in the FTRL representation of EW. We utilize a technical result from the literature on the minimum relative entropy principle (Jaynes, 1957; Grünwald, 2007, Chapter 19). Say we have to make an initial guess about the weights for the experts with some distribution π and then learn that $\mathbb{E}[\boldsymbol{z}] = \boldsymbol{\mu}$. The minimum relative entropy principle tells us to choose the distribution p with $\mathop{\mathbb{E}}_{z\sim p}[\boldsymbol{z}] = \boldsymbol{\mu}$ that is closest in KL divergence to π:

$$p_{mre} = \arg\min_{p\in\mathcal{P}_\mu} KL(p||\pi), \tag{4.14}$$

where $\mathcal{P}_\mu$ is defined as:

$$\mathcal{P}_\mu = \{p : \mathop{\mathbb{E}}_{z\sim p}[\boldsymbol{z}] = \boldsymbol{\mu}\}. \tag{4.15}$$

Let $\mathcal{E}_\pi = \{p : e^{\langle \boldsymbol{z},\boldsymbol{\theta}\rangle - F(\boldsymbol{\theta})}\pi(\boldsymbol{z})\}$ be an exponential family with cumulant generating function F, sufficient statistic $\boldsymbol{z}$, and carrier $\pi(\boldsymbol{z})$. It can be shown that if $p_\theta \in \mathcal{E}_\pi$ exists such that $\mathbb{E}_{p_\theta}[\boldsymbol{z}] = \boldsymbol{\mu}$ then $p_{mre} = p_\theta$.

Summarizing, we obtain the following Lemma.

Lemma 4. *For any $\boldsymbol{\mu}$, the minimum in*

$$\arg\min_{p\in\mathcal{P}_\mu} KL(p||\pi), \tag{4.16}$$

is achieved by $p_\theta \in \mathcal{E}_\pi$ such that $\mathbb{E}_{p_\theta}[\boldsymbol{z}] = \boldsymbol{\mu}$, provided such a p_θ exists.

4.5 MD is a special case of EW

In this section the main result of the Thesis is presented. We exploit the relationship between the KL divergence and Bregman divergence to show that the Mirror Descent algorithm is a special case of the Exponential weights algorithm.

Theorem 3. *Let p_{t+1} be the Exponential Weights distribution at time $t+1$ with prior π, let the loss of expert $\boldsymbol{z}$ be $\ell_t^z = \langle \boldsymbol{z}, \boldsymbol{g}_t \rangle$, let the forecasters loss be $\hat{\ell}_t = \mathop{\mathbb{E}}_{\boldsymbol{z}\sim p_{t+1}}[\langle \boldsymbol{z}, \boldsymbol{g}_t \rangle]$, and let F be the cumulant generating function of $\mathcal{E}_\pi$. Let Mirror Descent be used with Bregman divergence B_{F^*} generated by F^*, the convex conjugate of cumulant generating function F. Then the Mirror Descent algorithm is the mean of the Exponential Weights algorithm:*

$$\mathop{\mathbb{E}}_{\boldsymbol{z}\sim p_{t+1}}[\boldsymbol{z}] = \boldsymbol{w}_{t+1} = \phi^{-1}\Big(\phi(\boldsymbol{w}_t) - \eta \nabla f_i(\boldsymbol{w}_t)\Big), \tag{4.17}$$

where $\phi = \nabla F^*$.

Proof. We utilize the FTRL representation of both the EW algorithm and MD algorithm. Let $\boldsymbol{\theta}_\pi$ denote the natural parameter of π and let $\boldsymbol{w}_\pi$ denote the expectation of π. Recall the FTRL representation of the Exponential Weights algorithm:

$$p_{t+1} = \arg\min_{p\in\mathcal{E}_\pi}\left\{ \sum_{i=1}^{t} \mathop{\mathbb{E}}_{\boldsymbol{z}\sim p}[\langle \boldsymbol{z}, \boldsymbol{g}_i \rangle] + \frac{1}{\eta} KL(p||\pi) \right\}. \tag{4.18}$$

Note that p is restricted to $\mathcal{E}_\pi$ in (4.18). Say that we are minimizing over all distributions. The distribution that attains the minimum in (4.18) has mean $\mathop{\mathbb{E}}_{\boldsymbol{z}\sim p_{t+1}}[\boldsymbol{z}] = \boldsymbol{w}_{t+1}$. Because of this we could restrict p to distributions with mean $\boldsymbol{w}_{t+1}$. If we restrict p to distributions with mean $\boldsymbol{w}_{t+1}$ we are minimizing the KL divergence plus a constant $\sum_{i=1}^{t}\langle \boldsymbol{w}_{t+1}, \boldsymbol{g}_i \rangle$, which we may ignore. Hence, we are minimizing the KL divergence for distributions with a given mean, to which we can apply Lemma 4.

For any $p \in \mathcal{E}_\pi$ the expression inside the curly brackets in (4.18) reduces to:

$$\begin{aligned} \sum_{i=1}^{t} \mathop{\mathbb{E}}_{\boldsymbol{z}\sim p}[\langle \boldsymbol{z}, \boldsymbol{g}_t \rangle] + \frac{1}{\eta} KL(p||\pi) &= \sum_{i=1}^{t} \langle \boldsymbol{w}, \boldsymbol{g}_i \rangle + \frac{1}{\eta} B_F(\boldsymbol{\theta}_\pi||\boldsymbol{\theta}) \\ &= \sum_{i=1}^{t} \langle \boldsymbol{w}, \boldsymbol{g}_i \rangle + \frac{1}{\eta} B_{F^*}(\boldsymbol{w}||\boldsymbol{w}_\pi), \end{aligned} \tag{4.19}$$

where $\boldsymbol{w}$ is the expectation and $\boldsymbol{\theta}$ is the natural parameter of some $p \in \mathcal{E}_\pi$. Both steps in (4.19) follow from Theorem 2. The EW distribution p_{t+1} has expectation parameter $\boldsymbol{w}$ that minimizes the expression in (4.19). Hence, to find $\mathop{\mathbb{E}}_{\boldsymbol{z}\sim p_{t+1}}[\boldsymbol{z}]$ we find

the $\boldsymbol{w}$ that minimizes the expression in (4.19):

$$\underset{\boldsymbol{z}\sim p_{t+1}}{\mathbb{E}}[\boldsymbol{z}] = \arg\min_{\boldsymbol{w}} \left\{ \sum_{i=1}^{t} \langle \boldsymbol{w}, \boldsymbol{g}_i \rangle + \frac{1}{\eta} B_{F^*}(\boldsymbol{w} || \boldsymbol{w}_\pi) \right\}, \tag{4.20}$$

which is the FTRL representation of the Mirror Descent algorithm.

□

For example, we may use the exponential family representation of the multivariate normal distribution with the identity matrix as covariance matrix:

$$\begin{aligned} \boldsymbol{\theta} &= \boldsymbol{w} \\ F(\boldsymbol{\theta}) &= \frac{1}{2} ||\boldsymbol{\theta}||_2^2 \\ F^*(\boldsymbol{w}) &= \frac{1}{2} ||\boldsymbol{w}||_2^2 \\ T(\boldsymbol{z}) &= \boldsymbol{z} \\ K(\boldsymbol{z}) &= (2\pi)^{-\frac{d}{2}} e^{-\frac{1}{2}\langle \boldsymbol{z}, \boldsymbol{z} \rangle}. \end{aligned} \tag{4.21}$$

Now we use Theorem 3 and obtain:

$$\underset{\boldsymbol{z}\sim p_{t+1}}{\mathbb{E}}[\boldsymbol{z}] = \boldsymbol{w}_t - \eta \boldsymbol{g}_t, \tag{4.22}$$

which is the update step of the Gradient Descent algorithm.

Chapter 5

When is a function a cumulant generating function?

In this chapter the condition in Theorem 3 is explored. The condition is that the dual of function F^* from which the Bregman divergence B_{F^*} is generated is the cumulant generating function of an exponential family. Previous studies by Banerjee et al. (2005) relate cumulant generating functions to exponentially convex functions (Akhiezer, 1965; Ehm et al., 2003). A function is called exponentially convex if it is positive definite. In turn, this gives the result that every positive semi-definite function is a cumulant generating function. However, the positive semi-definiteness of a function is a technical condition that is difficult to interpret and also non-constructive: the fact that a corresponding exponential family exists does not tell us what that family is. It gives little insight into the carrier measure, or in our case prior that corresponds to an exponentially convex function. To find a simpler, more constructive condition for when a function is a cumulant generating function we explore the relation between cumulant generating functions, *moment generating functions*, *characteristic functions*, *Laplace transforms*, and *Fourier transforms*.

In the following we introduce exponentially convex functions, Fourier tranforms, as well as two generating functions: the moment generating function (MGF) and the characteristic function (CF). Exponentially convex functions are introduced in section 5.1 and are used to provide a necessary and sufficient condition for a function to be a cumulant generating function. To obtain a probabilistic interpretation of exponentially convex functions and to obtain the link with Fourier transformations we introduce generating functions in section 5.2. Fourier transformations are used to gain access to inversion formulas to find the carrier (or prior). These inversion formulas are introduced in section 5.3 as well as a *saddle point approx-*

imation to the inversion formulas. Furthermore, in section 5.3 the saddle point approximation is used to develop Theorem 7, in which we present a new sufficient condition for a function to be a cumulant generating function. In section 5.4 the bijection between exponentially convex functions and Fourier transformations of non-negative functions due to Ehm et al. (2003) is formally stated. In combination with a Lemma from Wendland (2004) and the derivation of the saddle point approximation this bijection leads to a second new sufficient condition for a function to be a cumulant generating function in Theorem 8.

5.1 Exponentially convex functions

Banerjee et al. (2005) gives a bijection between exponentially convex functions and cumulant generating functions. A function $\psi : \Theta \to \mathbb{R}_{++}, \Theta \in \mathbb{R}^d$, where $\mathbb{R}_{++}$ denotes the positive reals, is called exponentially convex if

$$\sum_{i,j=1}^{n} \psi(\boldsymbol{\theta}_i - \boldsymbol{\theta}_j) c_i, \overline{c_j} \geq 0, \tag{5.1}$$

for any set $\{\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_n\}$ with $\boldsymbol{\theta}_i + \boldsymbol{\theta}_j \in \Theta$, $\forall i, j, \{c_1, \ldots c_n\} \in \mathbb{C}$, where $\mathbb{C}$ denotes the complex plane and $\overline{c_j}$ denotes the complex conjugate of c_j.

The crucial property of exponentially convex functions we require is the following. If and only if a function ψ is exponentially convex there exists a unique, bounded, non-negative measure K such that ψ can be represented as (Devinatz et al., 1955):

$$\psi(\boldsymbol{\theta}) = \int_{\mathbb{R}^d} e^{\langle \boldsymbol{x}, \boldsymbol{\theta} \rangle} K(\boldsymbol{x}) d\boldsymbol{x}. \tag{5.2}$$

If one compares the definition of a cumulant generating function (see equation (4.9)) with (5.2) then it is tempting to say that a cumulant generating function F is the logarithm of an exponentially convex function ψ: $F(\boldsymbol{\theta}) = \log \psi(\boldsymbol{\theta})$. Banerjee et al. (2005) does exactly that and formally derives a bijection between cumulant generating functions and exponentially convex functions.

Theorem 4 (Banerjee et al. (2005))**.** *Let $\psi : \Theta \to \mathbb{R}_{++}$ be an exponentially convex function such that Θ is open and $F(\boldsymbol{\theta}) = \log \psi(\boldsymbol{\theta})$ is strictly convex. Then F is the cumulant generating function of an exponential family. Conversely, if $F(\boldsymbol{\theta})$ is the cumulant generating function of an exponential family, then $\psi(\boldsymbol{\theta}) = \exp(F(\boldsymbol{\theta}))$ is an exponentially convex function.*

5.2 Generating functions

To gain a probabilistic interpretation of exponentially convex functions and find a link with Fourier transforms two generating functions are introduced: moment generating and characteristic functions. The link with Fourier transforms gives us access to inversion integrals, which will be used in the derivation of two sufficient conditions for a function to be a cumulant generating function. The moment generating function of distribution p is a function $M_p : \mathbb{R}^d \to \mathbb{R}$ which is defined as the Laplace transform of p (Billingsley, 2008, Section 21):

$$\begin{aligned} M_p(\boldsymbol{\theta}) &= \int_{\mathbb{R}^d} e^{\langle \boldsymbol{\theta}, \boldsymbol{x} \rangle} p(\boldsymbol{x}) d\boldsymbol{x} \\ &= \mathop{\mathbb{E}}_{x \sim p} [e^{\langle \boldsymbol{\theta}, \boldsymbol{x} \rangle}], \end{aligned} \tag{5.3}$$

with $\boldsymbol{\theta} \in \mathbb{R}^d$.

To relate moment generating functions to exponential families consider the following. Let F be the cumulant generating function of an exponential family with carrier $K(\boldsymbol{x})$ and sufficient statistic $T(\boldsymbol{x}) = \boldsymbol{x}$. If $F(\mathbf{0}) = 0$, then since $K(\boldsymbol{x})$ is always positive, $K(\boldsymbol{x})$ is a probability distribution:

$$1 = e^{F(\mathbf{0})} = \int_{\mathbb{R}^d} e^{\langle \boldsymbol{x}, \mathbf{0} \rangle} K(\boldsymbol{x}) d\boldsymbol{x} = \int_{\mathbb{R}^d} K(\boldsymbol{x}) d\boldsymbol{x}. \tag{5.4}$$

This opens up the relationship with moment generating functions. The cumulant generating function F of the exponential family with sufficient statistic $\boldsymbol{x}$ and carrier p is the logarithm of the moment generating function of p (Jørgensen and Labouriau, 2012): $\log M_p(\boldsymbol{\theta}) = F(\boldsymbol{\theta})$. The domain of M_p and F is $\Theta = \{\boldsymbol{\theta} \in \mathbb{R}^d : M_p(\boldsymbol{\theta}) < \infty\}$. Note that 5.3 gives a relation with exponentially convex functions. Since $p(\boldsymbol{x})$ is non-negative and bounded M_p is an exponentially convex function (see equation (5.2)).

The characteristic function ρ_p of any random variable always exists and uniquely determines its distribution function (Shiryaev, 1996, Chapter 12). The characteristic function of p is defined as the Fourier transform of p:

$$\rho_p(\boldsymbol{\theta}) = \int_{\mathbb{R}^d} e^{i\langle \boldsymbol{\theta}, \boldsymbol{x} \rangle} p(\boldsymbol{x}) d\boldsymbol{x}, \tag{5.5}$$

where i is the imaginary number. There is a relationship between moment generating and characteristic functions: $\rho_p(\boldsymbol{\theta}) = M_p(i\boldsymbol{\theta})$ (Jørgensen and Labouriau, 2012). A necessary and sufficient condition for a function to be a characteristic function is given by the Bochner-Khinchin Theorem:

Theorem 5 (Bochner (1933))**.** *Let $\rho(\boldsymbol{s})$ be continuous, $\boldsymbol{s} \in S = \mathbb{R}^d$, with $\rho(\mathbf{0}) = 1$. A necessary and sufficient condition that ρ is a characteristic function is that it is positive semi-definite, i.e. that for any set $\{\boldsymbol{s}_1, \ldots, \boldsymbol{s}_n\} \subseteq S$ with $\boldsymbol{s}_i + \boldsymbol{s}_j \in S$, $\forall i, j$, and $c \in \mathbb{C}$*

$$\sum_{i,j=1}^{n} \rho(s_i - s_j) c_i \overline{c_j} \geq 0. \tag{5.6}$$

The Bochner-Khinchin Theorem tells us that if and only if a function is positive semi-definite then it is a characteristic function. Dropping the condition that $\rho(\mathbf{0}) = 1$ gives us a necessary and sufficient condition for ρ to be the Fourier transform of a finite non-negative measure on $\mathbb{R}^d$. This is the same condition as for a function to be called exponentially convex (Akhiezer, 1965; Ehm et al., 2003). Some properties of positive semi-definite functions are given by the following Lemma:

Lemma 5 (Shiryaev (1996), Chapter 12; Wendland (2004), Theorem 6.2)**.** *Let ρ_p be a positive semi-definite function, then ρ_p has the following properties:*

1. $|\rho_p(\boldsymbol{\theta})| \leq \rho_p(0)$.
2. *ρ_p is uniformly continuous for $\boldsymbol{\theta} \in \mathbb{R}^d$.*
3. $\rho_p(\boldsymbol{\theta}) = \overline{\rho_p(-\boldsymbol{\theta})}$, *where $\overline{\rho_p(-\boldsymbol{\theta})}$ denotes the complex conjugate of $\rho_p(-\boldsymbol{\theta})$.*
4. *$\rho_p(\boldsymbol{\theta})$ is real valued if and only if p is symmetric.*
5. $\rho_p(\mathbf{0}) \geq 0$.

The properties in Lemma 5 give necessary conditions for both positive semi-definite and characteristic functions: if any of the above does not apply to a function then it is not positive semi-definite.

5.3 Inversion of generating functions

In order to derive new sufficient conditions for a function to be a cumulant generating function we utilize inversion integrals of characteristic and moment generating

functions. To find a distribution given a moment generating M_p or characteristic function ρ_p we may employ the following inversion formula (Daniels, 1954):

$$\begin{aligned} p(\boldsymbol{x}) &= \frac{1}{(2\pi)^d} \int_{\mathbb{R}^d} \rho_p(\boldsymbol{\theta}) e^{-i\langle \boldsymbol{\theta}, \boldsymbol{x} \rangle} \, d\boldsymbol{\theta} \\ &= \frac{1}{(2\pi)^d} \int_{\mathbb{R}^d} M_p(i\boldsymbol{\theta}) e^{-i\langle \boldsymbol{\theta}, \boldsymbol{x} \rangle} \, d\boldsymbol{\theta}. \end{aligned} \tag{5.7}$$

An alternative, equivalent inversion integral to find the distribution is given by the following (Daniels, 1954):

$$p(\boldsymbol{x}) = \frac{1}{i(2\pi)^d} \int_{\boldsymbol{\gamma} - \mathbf{i}\infty}^{\boldsymbol{\gamma} + \mathbf{i}\infty} e^{F(T) - \langle T, \boldsymbol{x} \rangle} \, dT, \tag{5.8}$$

where $F = \log(M_p)$, $T = \boldsymbol{\theta} + i\boldsymbol{y}, T \in \mathbb{C}^d$, and $\boldsymbol{\gamma} \in \Theta$.

In order to derive a new sufficient condition for a function to be a cumulant generating function we make use of saddle point approximations. Saddle point approximations are used in asymptotics to find accurate approximations to integrals like (5.8). To derive a new sufficient statistic we make use of a saddle point approximation due to Daniels (1954), who used it to approximate the density function p_n of the mean of n i.i.d. variables.

Theorem 6 (Reid (1988))**.** *Let $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n$ be independent, identically distributed random vectors from a density p on $\mathbb{R}^d$ and let $F(\boldsymbol{\theta})$ the cumulant generating function of p. The saddle point expansion of density of the mean $\overline{\boldsymbol{x}} = \frac{1}{n} \sum_{i=1}^n \boldsymbol{x}_i$ is given by:*

$$p_n(\overline{\boldsymbol{x}}) = (2\pi)^{-\frac{d}{2}} \left(\frac{n}{\det(\nabla^2 F(T_0))} \right)^{\frac{1}{2}} e^{n(F(T_0) - \langle T_0, \overline{\boldsymbol{x}} \rangle)} (1 + R_n), \tag{5.9}$$

where R_n is the remainder term, $T_0 \in \mathbb{R}^d$ is the saddle point, and $\nabla^2 F(T_0)$ is the Hessian of F. The right hand side of (5.9), excluding the $1 + R_n$ factor, is called the saddle point approximation $g(\overline{\boldsymbol{x}})$. The saddle point T_0 is found when the following holds true:

$$\nabla F(T_0) = \overline{\boldsymbol{x}}. \tag{5.10}$$

The saddle point exists if the convex conjugate of F exists (Reid, 1988). This leads to a different expression for $g(\overline{\boldsymbol{x}})$ given by (McCullagh, 1987, Chapter 6):

$$g(\overline{\boldsymbol{x}}) = (2\pi)^{-\frac{d}{2}}(n\det(\nabla^2 F^*(\overline{\boldsymbol{x}})))^{\frac{1}{2}}e^{-nF^*(\overline{\boldsymbol{x}}))} \tag{5.11}$$

The saddle point approximation can also be applied to approximate the carrier measure of a member of the exponential family (Reid, 1988). Let F be a cumulant generating function. The saddle point approximation of the carrier is given by (5.11). The saddle point approximation $\widetilde{p}$ to said exponential family for $n = 1$ is now given by:

$$\widetilde{p}(\overline{\boldsymbol{x}}) = e^{\langle \boldsymbol{\theta}, \overline{\boldsymbol{x}}\rangle - F(\boldsymbol{\theta})} g(\overline{\boldsymbol{x}}). \tag{5.12}$$

It is not guaranteed that (5.12) integrates to 1. In fact, it may integrate to some function $b : \Theta \to \mathbb{R}_{++}$:

$$\begin{aligned} &\int_{\mathbb{R}^d} e^{\langle \boldsymbol{\theta}, \overline{\boldsymbol{x}}\rangle - F(\boldsymbol{\theta})} g(\overline{\boldsymbol{x}}) d\overline{\boldsymbol{x}} = b(\boldsymbol{\theta}) \\ &\Leftrightarrow \int_{\mathbb{R}^d} e^{\langle \boldsymbol{\theta}, \overline{\boldsymbol{x}}\rangle} g(\overline{\boldsymbol{x}}) d\overline{\boldsymbol{x}} = b(\boldsymbol{\theta}) e^{F(\boldsymbol{\theta})}. \end{aligned} \tag{5.13}$$

However, if (5.12) integrates to a constant then $e^{F(\boldsymbol{\theta})}$ is positive semi-definite, which is shown in the following Theorem.

Theorem 7. *Legendre functions F for which*

$$\int_{\mathbb{R}^d} e^{\langle \boldsymbol{\theta}, \overline{\boldsymbol{x}}\rangle - F(\boldsymbol{\theta})} (2\pi)^{-\frac{d}{2}} (\det(\nabla^2 F^*(\overline{\boldsymbol{x}})))^{\frac{1}{2}} e^{-F^*(\overline{\boldsymbol{x}}))} d\overline{\boldsymbol{x}} \tag{5.14}$$

integrates to a constant b independent of $\boldsymbol{\theta}$ are cumulant generating functions of exponential families with carrier $\frac{1}{b}g(\overline{x}) = \frac{1}{b}(\det(\nabla^2 F^*(\overline{\boldsymbol{x}})))^{\frac{1}{2}} e^{-F^*(\overline{\boldsymbol{x}}))}$.

Proof. Let $g(\overline{\boldsymbol{x}})$ (as constructed in (5.11)) denote the saddle point approximation to the carrier measure for $n = 1$. We construct a distribution with (5.12). Now, using (5.13):

$$\begin{aligned} \sum_{i,j=1}^{n} b e^{F(\boldsymbol{\theta}_i + \boldsymbol{\theta}_j)} \boldsymbol{c}_i \overline{\boldsymbol{c}_j} &= \sum_{i,j=1}^{n} \boldsymbol{c}_i \overline{\boldsymbol{c}_j} \int_{\mathbb{R}^d} e^{\langle \boldsymbol{\theta}_i + \boldsymbol{\theta}_j, \overline{\boldsymbol{x}}\rangle} g(\overline{\boldsymbol{x}}) d\overline{\boldsymbol{x}} \\ &= \int_{\mathbb{R}^d} \sum_{i=1}^{n} \boldsymbol{c}_i e^{\langle \boldsymbol{\theta}_i, \overline{\boldsymbol{x}}\rangle} \sum_{j=1}^{n} \overline{\boldsymbol{c}_j} e^{\langle \boldsymbol{\theta}_j, \overline{\boldsymbol{x}}\rangle} g(\overline{\boldsymbol{x}}) d\overline{\boldsymbol{x}} \\ &= \int_{\mathbb{R}^d} \left[\sum_{i=1}^{n} \boldsymbol{c}_i e^{\langle \boldsymbol{\theta}_i, \overline{\boldsymbol{x}}\rangle} \right]^2 g(\overline{\boldsymbol{x}}) d\overline{\boldsymbol{x}} \\ &\geq 0, \end{aligned} \tag{5.15}$$

where the last equation holds because $g(\overline{\boldsymbol{x}})$ is positive by construction and we may move the summation past the integration sign due to Tonelli's theorem (Tonelli, 1909). Now, noting that b is a positive constant means that e^F is positive semi-definite. Furthermore, F may be represented as:

$$F(\boldsymbol{\theta}) = \log \int_{\mathbb{R}^d} e^{\langle \boldsymbol{\theta}, \overline{\boldsymbol{x}} \rangle} \frac{1}{b} g(\overline{\boldsymbol{x}}) d\overline{\boldsymbol{x}}, \tag{5.16}$$

which concludes the proof. □

Examples of Legendre functions $F : \Theta \to \mathbb{R}$ that satisfy the condition of Theorem 7 are given by Blæsild and Jensen (1985): $F(\boldsymbol{\theta}) = \frac{1}{2}||\boldsymbol{\theta}||_2^2, \boldsymbol{\theta} \in \mathbb{R}^d$, the cumulant generating function of the normal distribution, and $F(\boldsymbol{\theta}) = \sum_{i=1}^d \Gamma(\boldsymbol{\theta}_i + 1) - (\boldsymbol{\theta}_i + 1)\log(-k), \boldsymbol{\theta} \in (-1, \infty)^d$, k is fixed, and Γ denotes the gamma function; the cumulant generating function of the gamma distribution. Another example is given by $F(\boldsymbol{\theta}) = \sum_{i=1}^d -\log(-\boldsymbol{\theta}_i)$ with $\boldsymbol{\theta} \in \mathbb{R}_{--}^d$, where $\mathbb{R}_{--}$ denotes the negative real numbers. The proof of this example is given in chapter 6.

5.4 Bijection exponentially convex functions and Fourier transforms

This section will formally state the relation between positive definite functions and exponentially convex functions in order to derive a new sufficient condition for functions to be cumulant generating functions. Furthermore, we give a sufficient condition for a function to be positive definite due to Wendland (2004). Ehm et al. (2003) gives a bijection between exponentially convex and positive semi-definite functions. If ψ is an exponentially convex function then it is analytic and $\psi(i\boldsymbol{\theta})$ is the Fourier transform of a non-negative Borel measure. A function is analytic if a function can be represented as a power series that converges everywhere on its domain:

$$\psi(\boldsymbol{\theta}) = \sum_{i=1}^{\infty} a_n \boldsymbol{\theta}^n. \tag{5.17}$$

Note that all elementary functions are analytic (Parks and Krantz, 1992). Any products, sums, or compositions of analytic functions are also analytic. Examples of analytic functions are exponentials, polynomials, trigonometric functions, and the natural logarithm. The inversion integral in (5.7) combined with the bijection between exponentially convex functions and Fourier transforms yields a route to find if a function is positive definite, a stricter condition than positive

semi-definite. For positive definite functions the equation in Theorem 5 becomes a strict inequality. We now give a sufficient condition for a function to be positive definite.

Lemma 6 (Wendland (2004), Theorem 6.11)**.** *Let $\rho : \Theta \to \mathbb{C}$ be a bounded continuous function where $\int |\rho(\boldsymbol{\theta})| d\boldsymbol{\theta} < \infty$. Then ρ is a positive definite function if and only if*

$$\frac{1}{(2\pi)^d} \int_{\mathbb{R}^d} \rho(\boldsymbol{\theta}) e^{-i\langle \boldsymbol{\theta}, \boldsymbol{x} \rangle} d\boldsymbol{\theta} > 0. \tag{5.18}$$

That is, if and only if the inverse Fourier transform of ρ is non-negative and not identically equal to zero then ρ is a positive definite function.

5.5 Application to Legendre functions

To apply the above theory to exponents of Legendre functions we combine the ideas from Ehm et al. (2003), Wendland (2004), and Daniels (1954) to derive a new sufficient condition for a function to be positive semi-definite. Specifically, we use the bijection between exponentially convex functions and positive semi-definite functions to be able to use the Fourier inversion integral in (5.7). Then we use the derivation of the saddle point approximation by Daniels (1954) to show that the inverse Fourier transorm is always positive, which leads to the results that the exponent of the original function is positive definite by Lemma 6. First we require a brief introduction in contour integration. We omit proofs of the following, instead we prefer an informal, visual approach which gives some intuition on the ideas that are used by Daniels (1954).

Say we want to solve an integral like $\frac{1}{2\pi i} \int_{\gamma - i\infty}^{\gamma + i\infty} M(z) dz$. This integral is a contour integral. The contour looks like the image in Figure 5.1. Just as with regular integration we may split the integral in several parts. For instance, the sum of the integral from A to B plus the integral from B to A is equal to the integral over the contour. For analytic functions we may distort the contour while the value for the integral remains the same. Since moment generating functions are analytic this distortion is allowed.

Now, lets say the function $M(z)$, with $z \in \mathbb{C}$ generates the surface plot in Figure 5.2 and say we want to find the value of the integral $\frac{1}{2\pi i} \int_{\gamma - i\infty}^{\gamma + i\infty} M(z) dz$. The saddle point is located at the center of Figure 5.2 (with some imagination one can see a horse saddle here). The idea is to capture as much of the value of integral as possible in a small as possible part of the contour. This is done by the *method of*

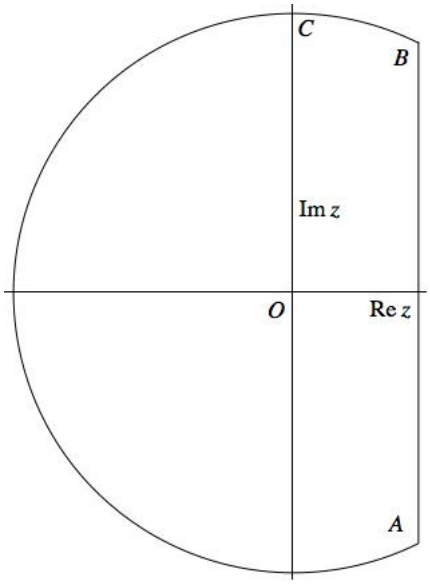


Figure 5.1: *Example of a contour of integration.*

steepest descent, not to be confused with Online Gradient Descent. We arrange the contour such that most of the contour is where $M(z)$ is small, which is where the value of the integral is also small. We shall call this part of the integral $D1$. Then we go over the saddle point by the steepest possible route to capture as much of the integral in a small part of the contour as possible, which we will call $D2$. The result of all this is that $D2$ is around the saddle point and is larger than $D1$ in the absolute sense.

We continue by combining the ideas of Ehm et al. (2003), Daniels (1954), Wendland (2004) and derive a, to our knowledge, new sufficient condition for a function to be a cumulant generating function.

Theorem 8. *Let F be an analytic function of Legendre type with*

(a) $e^{F(i\boldsymbol{\theta})} = \overline{e^{F(-i\boldsymbol{\theta})}}$.

(b) $\int_{\mathbb{R}^d} |e^{F(i\boldsymbol{\theta})}| d\boldsymbol{\theta} < \infty$

(c) $|e^{F(i\boldsymbol{\theta}}| \leq B, B \in \mathbb{R}$

Then e^F is an exponentially convex function and F is a cumulant generating function.

Proof. If F is analytic then we may uniquely extend it to the complex plane.

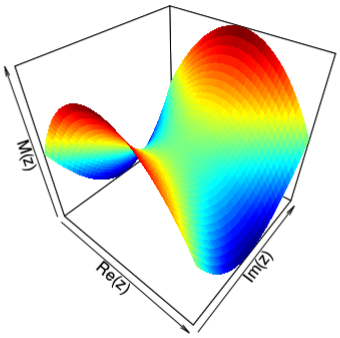


Figure 5.2: *Surface plot of $M(z), z \in \mathbb{C}$. The saddle point is located at the center of the plot.*

If $e^{F(i\boldsymbol{\theta})}$ satisfies condition (c) it is bounded. Combined with condition (b) this makes Lemma 6 applicable. Conditions (b) and (c) tell us that $e^{F(i\boldsymbol{\theta})}$ can be represented as the Fourier transformation of some not necessary positive function $\hat{f}$: $e^{F(i\boldsymbol{\theta})} = \int_{-\infty}^{\infty} \hat{f}(x) e^{i\langle \boldsymbol{x}, \boldsymbol{\theta} \rangle} d\boldsymbol{x}$ (Wendland, 2004, chapter 6). If the inverse Fourier transform of $e^{F(i\boldsymbol{\theta})}$ is positive then $e^{F(i\boldsymbol{\theta})}$ is positive definite by Lemma 6. To guarantee that the inverse Fourier transform of $e^{F(i\boldsymbol{\theta})}$ is real we require condition (a). Let p be the inverse Fourier transform of $e^{F(i\boldsymbol{\theta})}$, then:

$$
\begin{aligned}
p(\boldsymbol{x}) &= \frac{1}{(2\pi)^d} \int_{-\infty}^{\infty} e^{F(i\boldsymbol{\theta})} e^{-i\langle \boldsymbol{\theta}, \boldsymbol{x} \rangle} \, d\boldsymbol{\theta} \\
&= \frac{1}{(2\pi)^d} \int_{0}^{\infty} e^{F(i\boldsymbol{\theta})} e^{-i\langle \boldsymbol{\theta}, \boldsymbol{x} \rangle} \, d\boldsymbol{\theta} + \frac{1}{(2\pi)^d} \int_{0}^{\infty} e^{F(-i\boldsymbol{\theta})} e^{i\langle \boldsymbol{\theta}, \boldsymbol{x} \rangle} \, d\boldsymbol{\theta} \\
&= \frac{1}{(2\pi)^d} \int_{0}^{\infty} e^{F(i\boldsymbol{\theta})} e^{-i\langle \boldsymbol{\theta}, \boldsymbol{x} \rangle} \, d\boldsymbol{\theta} + \frac{1}{(2\pi)^d} \int_{0}^{\infty} \overline{e^{F(i\boldsymbol{\theta})} e^{-i\langle \boldsymbol{\theta}, \boldsymbol{x} \rangle}} \, d\boldsymbol{\theta} \\
&= \frac{1}{(2\pi)^d} \int_{0}^{\infty} e^{F(i\boldsymbol{\theta})} e^{-i\langle \boldsymbol{\theta}, \boldsymbol{x} \rangle} \, d\boldsymbol{\theta} + \frac{1}{(2\pi)^d} \overline{\int_{0}^{\infty} e^{F(i\boldsymbol{\theta})} e^{-i\langle \boldsymbol{\theta}, \boldsymbol{x} \rangle} \, d\boldsymbol{\theta}} \\
&= 2\Re\left(\frac{1}{(2\pi)^d} \int_{0}^{\infty} e^{F(i\boldsymbol{\theta})} e^{-i\langle \boldsymbol{\theta}, \boldsymbol{x} \rangle} \, d\boldsymbol{\theta}) \right),
\end{aligned}
\tag{5.19}
$$

where $\Re(z)$ is the real part of z. To prove that the inverse Fourier transform is positive everywhere we use the derivation of the saddle point approximation of inverse Fourier transforms due to Daniels (1954).

For $T = T_0 + i\boldsymbol{y}$, where T_0 is the saddle point, the Fourier inversion integral is equal to (Daniels, 1954):

$$p(\boldsymbol{x}) = \frac{1}{(2\pi)^d} \int_{-\infty}^{\infty} e^{F(T_0+i\boldsymbol{y})-\langle (T_0+i\boldsymbol{y}),\boldsymbol{x}\rangle} d\boldsymbol{y}. \tag{5.20}$$

On any admissible line parallel to the imaginary axis the integral attains its maximum modulus only where the line crosses the real axis. For on the line $T = \boldsymbol{\tau} + i\boldsymbol{y}$ (Daniels, 1954):

$$\begin{aligned} |e^{F(T)-\langle T,\boldsymbol{x}\rangle}| &= e^{-\langle \boldsymbol{\tau},\boldsymbol{x}\rangle} \left| \int_{-\infty}^{\infty} e^{\langle T,\boldsymbol{x}\rangle} \hat{f}(\boldsymbol{x}) d\boldsymbol{x} \right| \\ &\leq e^{-\langle \boldsymbol{\tau},\boldsymbol{x}\rangle + F(\boldsymbol{\tau})}. \end{aligned} \tag{5.21}$$

Furthermore, by the Riemann-Lebesgue Lemma, as $\boldsymbol{y} \to \infty$, $e^{F(\boldsymbol{\tau}+i\boldsymbol{y})} = O(\frac{1}{|\boldsymbol{y}|})$, which tells us that the integral cannot approach the modulus of the integral arbitrarily as $\boldsymbol{y} \to \infty$ (Daniels, 1954). The contour of integration is deformed in the complex plane with the curve of steepest descent approach. We focus on two parts of the integral, the integral on the curve of the steepest descent, which we will call $D2$, and the remainder of the integral, which we will call $D1$. On the steepest descent curve, $F(T) - \langle T, \boldsymbol{x}\rangle$ is real and $e^{F(T)-\langle T,\boldsymbol{x}\rangle}$ decreases steadily on each side of T_0 (Daniels, 1954). Since the integral contains most of its value in the neighborhood of the saddle point, which is on the real axis, we are guaranteed that $D2 > |D1|$ (Daniels, 1954). For Legendre functions the saddle point T_0 exists and the integral that appears in (5.20) is positive and real: the sum of $D1$ and $D2$ must be real by condition (a), $D2 > |D1|$, and since $D2 > 0$, $D2 + D1 > 0$. Hence, the Fourier inversion of $e^{F(i\boldsymbol{\theta})}$ is positive, which means $e^{F(i\boldsymbol{\theta})}$ is positive definite by Lemma 6, $e^{F(\boldsymbol{\theta})}$ is exponentially convex, and F is a cumulant generating function. □

Chapter 6

Examples of update steps

In this chapter some examples of update steps for different prior distributions are given. First we show that p_{t+1}, the update step for the EW algorithm, is an exponential family with carrier π. Next we start with some Legendre functions and give give some examples of prior distributions. Finally, some examples of prior distributions and related means are given.

Let π be a member of an exponential family with natural parameter $\boldsymbol{\theta}_\pi$. For $\ell_i^z = \langle \boldsymbol{z}, \boldsymbol{g}_i \rangle$ the update step for the EW algorithm is:

$$p_{t+1}(\boldsymbol{z}) = \frac{\pi(\boldsymbol{z}) \exp(-\eta \sum_{i=1}^t \ell_i^{\boldsymbol{z}})}{\int_{\mathbb{R}^d} \pi(\boldsymbol{z}) \exp(-\eta \sum_{i=1}^t \ell_i^{\boldsymbol{z}}) d\boldsymbol{z}}, \tag{6.1}$$

Note that the update step is equal to exponentially tilting π (see Escher (1932)). Let $\boldsymbol{\theta}_\pi$ be the natural parameter of π, then:

$$\begin{aligned} p_{t+1}(\boldsymbol{z}) &= \frac{\pi(\boldsymbol{z}) \exp(-\eta \sum_{i=1}^t \langle \boldsymbol{z}, \boldsymbol{g}_i \rangle)}{\int_{\mathbb{R}^d} \pi(\boldsymbol{z}) \exp(-\eta \sum_{i=1}^t, \langle \boldsymbol{z}, \boldsymbol{g}_i \rangle) d\boldsymbol{z}} \\ &= \frac{\exp(-F_\pi(\boldsymbol{\theta}_\pi) + \langle \boldsymbol{\theta}_\pi, \boldsymbol{z} \rangle - \eta \sum_{i=1}^t \langle \boldsymbol{z}, \boldsymbol{g}_i \rangle K_\pi(\boldsymbol{z})}{\exp(F_\pi(\boldsymbol{\theta}_\pi - \eta \sum_{i=1}^t \boldsymbol{g}_i) - F_\pi(\boldsymbol{\theta}_\pi))} \\ &= \exp(-F_\pi(\boldsymbol{\theta}_{p_{t+1}}) + \langle \boldsymbol{\theta}_\pi - \eta \sum_{i=1}^t \boldsymbol{g}_i, \boldsymbol{z} \rangle) K_\pi(\boldsymbol{z}) \\ &= \exp(-F_\pi(\boldsymbol{\theta}_{p_{t+1}}) + \langle \boldsymbol{\theta}_{p_{t+1}}, \boldsymbol{z} \rangle) K_\pi(\boldsymbol{z}). \end{aligned} \tag{6.2}$$

Hence, p_{t+1} is a member of the exponential family with cumulant generating function $F_\pi(\boldsymbol{\theta}_{p_{t+1}})$, natural parameter $\boldsymbol{\theta}_{p_{t+1}} = \boldsymbol{\theta}_\pi - \eta \sum_{i=1}^t \boldsymbol{g}_i$, sufficient statistic $\boldsymbol{z}$,

and carrier measure $K_{p_{t+1}}(\boldsymbol{z}) = K_\pi(\boldsymbol{z})$. The mean of this distribution can now be found with $\boldsymbol{\mu}_{p_{t+1}} = \nabla F_{p_{t+1}}(\boldsymbol{\theta}_{p_{t+1}})$. Note that the cumulant generating function and the natural statistic are the only parts of π we require to find the weights for subsequent rounds. The computational complexity of updating p is linear in the number of rounds, which makes EW applicable in large scale problems.

6.1 From Legendre function to prior

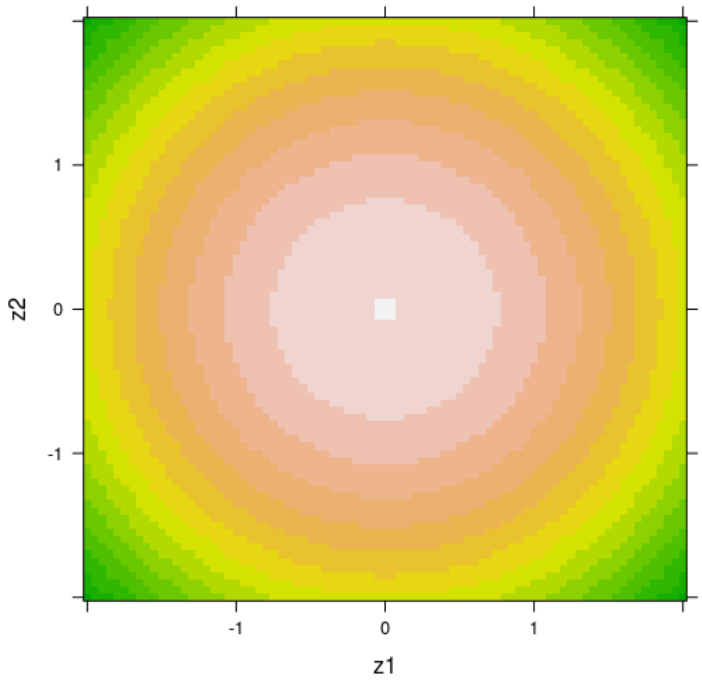


Figure 6.1: *Bivariate standard normal contour plot on* $\boldsymbol{z} \in [-2, 2]^2$. *The scale goes from white (high) to green (low).*

In this section we give some application of the theory developed in chapter 5 to find a prior for a given Legendre function. We start with a trivial example, namely $F(\boldsymbol{\theta}) = \frac{1}{2}||\boldsymbol{\theta}||_2^2$. For this example we will apply Theorem 8. Since F is a polynomial it is analytic and we may extend it to the complex plane. We have $e^{F(i\boldsymbol{\theta})} = e^{-\frac{1}{2}||\boldsymbol{\theta}||_2^2}$ after which condition (a) from Theorem 8 is trivial. As for condition (b): $\int |e^{-\frac{1}{2}||\boldsymbol{\theta}||_2^2}| d\boldsymbol{\theta} < \infty$. Condition (c) is simply noting that $|e^{-\frac{1}{2}||\boldsymbol{\theta}||_2^2}| \leq |e^{-\frac{1}{2}||\mathbf{0}||_2^2}|$. Hence, we may conclude that $\frac{1}{2}||\boldsymbol{\theta}||_2^2$ is a cumulant generating function. The carrier can be found explicitly in this case. Omitting the derivation, the

Fourier inversion integral yields $(2\pi)^{-\frac{d}{2}} e^{-\frac{1}{2}||\boldsymbol{z}||_2^2}$, the carrier for the standard normal distribution. The carrier is also the prior for the EW algorithm in this case. The prior is visualized for $\boldsymbol{z} \in [-2, 2]^2$ in Figure 6.1. We see that when $\boldsymbol{z}$ is further from the origin the prior density diminishes.

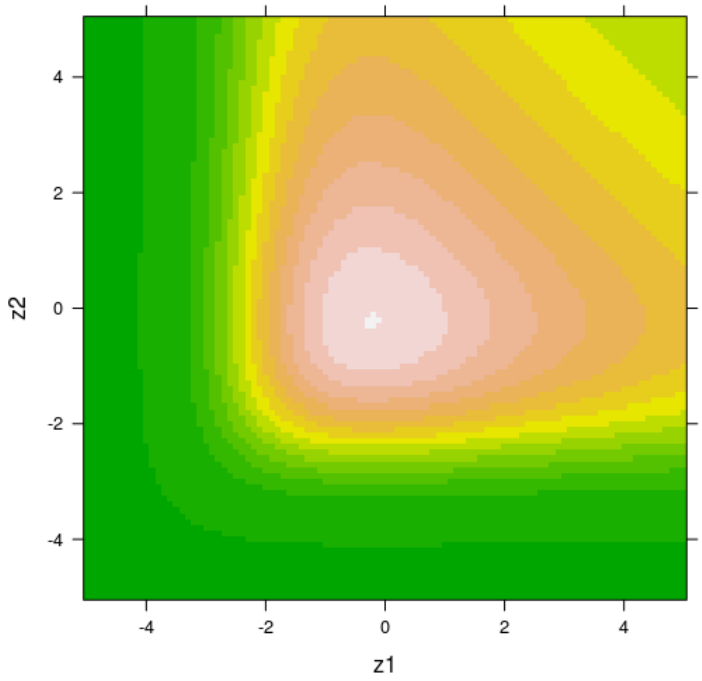


Figure 6.2: *Numerical evaluation of* (5.8) *for* $F(\boldsymbol{\theta}) = \sum_{i=1}^{2} \boldsymbol{\theta}_i \log \boldsymbol{\theta}_i, \boldsymbol{z} \in [-5, 5]$. *The scale goes from white (high) to green (low).*

A less trivial application of Theorem 8 is given by $F(\boldsymbol{\theta}) = \sum_{i=1}^{d} \boldsymbol{\theta}_i \log \boldsymbol{\theta}_i$, $\boldsymbol{\theta} \in \mathbb{R}_{++}^d$, where $\mathbb{R}_{++}^d$ denotes the positive reals. Note that when $\boldsymbol{\theta}$ is restricted to the probability simplex this is the negative entropy. Since F is a product of elementary functions it is analytic. We continue for $d = 1$ for simplicity. As for condition (a): $\overline{e^{-i\theta \log -i\theta}} = \overline{e^{-\frac{\theta\pi}{2} - i\theta \log \theta}} = e^{-\frac{\theta\pi}{2} + i\theta \log \theta} = e^{i\theta \log i\theta}$. As for condition (b): $|e^{i\theta \log i\theta}| = |(i\theta)^{(i\theta)}| = e^{-\theta \arg(i\theta)}$, where $\arg(ix)$ is the arg function, which gives the angle between the positive real axis and the line between its argument and the origin. The arg function for purely imaginary numbers iy is $\text{sign}(y)\frac{\pi}{2}$. We have that $\int |e^{-\theta \arg(i\theta)}| d\boldsymbol{\theta} < \infty$. Condition (c) is also satisfied, $|e^{-\theta \arg(i\theta)}| \leq |e^{-\mathbf{0} \arg(i\mathbf{0})}|$, and thus F is a cumulant generating function. Finding the carrier for F is problematic since neither inversion integral is easy to solve for this function and Theorem 7 does not work. However, we may still numerically evaluate the integral in (5.8). For $\boldsymbol{z} \in [-10 : 10]^2$ the integral in (5.8) yields Figure 6.2. Software used for inversion

was R (R Core Team, 2016) with the package "Inversion of Laplace-Transformed Functions" (Barry, 2015). We see an asymmetrical contour plot, which agrees with 4. in Lemma 5: $e^{F(i\theta)}$ is real if and only if the prior is symmetric. We see that if we move $\boldsymbol{z}$ away from the origin in the direction of $(-\infty, -\infty)$ the density quickly diminishes. As opposed to the prior for $F(\boldsymbol{\theta}) = \frac{1}{2}||\boldsymbol{\theta}||^2$ we see that the density does not diminish as quickly in the other directions.

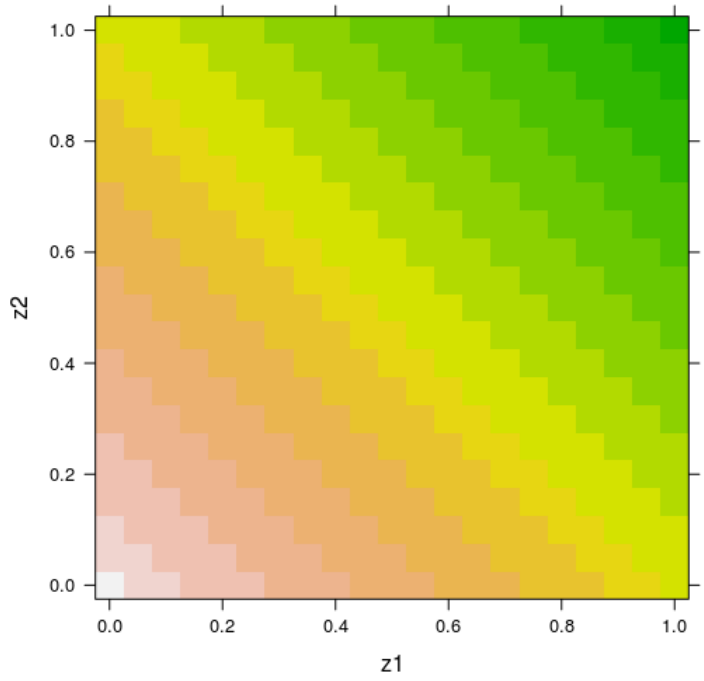


Figure 6.3: *Contour plot for the exponential family associated with* $F(\boldsymbol{\theta}) = \sum_{i=1}^{d} -\log(-\boldsymbol{\theta}_i)$, $\boldsymbol{\theta} = -1$, *and* $\boldsymbol{z} \in [0, 1]$. *The scale goes from white (high) to green (low).*

An application of Theorem 7 is given by $F(\boldsymbol{\theta}) = \sum_{i=1}^{d} -\log(-\boldsymbol{\theta}_i)$ with $\boldsymbol{\theta} \in \mathbb{R}^d_{--}$. The convex conjugate of F is $F^*(\overline{\boldsymbol{x}}) = \sum_{i=1}^{d} -\log \overline{\boldsymbol{x}_i}$ and the Hessian of the dual is $\frac{1}{\overline{\boldsymbol{x}}^2}$. We obtain $g(\overline{\boldsymbol{x}}) = (2\pi)^{-\frac{d}{2}} \frac{1}{\overline{\boldsymbol{x}}} e^{\log \overline{\boldsymbol{x}}} = (2\pi)^{-\frac{d}{2}}$. Now, $\int_{\overline{\boldsymbol{x}}} e^{\langle \overline{\boldsymbol{x}}, \boldsymbol{\theta} \rangle + \sum_{i=1}^{d} \log -\boldsymbol{\theta}} g(\overline{\boldsymbol{x}}) d\overline{\boldsymbol{x}} = (2\pi)^{\frac{d}{2}}$, which shows that F is a cumulant generating function and the carrier is 1. Note that this carrier not directly gives us a prior for EW, since this is not a density function. However, we may use any member of the exponential family associated with F, which is the exponential distribution, as a prior. For $\boldsymbol{\theta} = -1$, $\overline{\boldsymbol{x}} \in [0, 3]^2$ this prior produces the contour plot in Figure 6.4. We see that as we move away from the origin to (∞, ∞) the density diminishes orthogonal to the line from the origin to (∞, ∞).

6.2 From prior to update step

Prior distribution	$F(\boldsymbol{\theta})$	$\boldsymbol{\mu}_{p_{t+1}}$
Multivariate Gaussian with identity covariance	$\frac{1}{2}\|\|\boldsymbol{\theta}\|\|_2^2$	$\boldsymbol{\theta}_\pi - \eta \sum_{i=1}^t \boldsymbol{g}_i$
Poisson (d = 1)	$\exp(\theta)$	$\exp(\theta_\pi - \eta \sum_{i=1}^t g_i)$
Binomial (d = 1)	$N \log(1 + \exp(\theta))$	$\frac{N \exp(\theta_\pi - \eta \sum_{i=1}^t g_i)}{\exp(\theta_\pi - \eta \sum_{i=1}^t g_i)+1}$
Exponential (d = 1)	$-\log(-\theta)$	$\frac{1}{-\theta_\pi + \eta \sum_{i=1}^t g_i}$
Chi squared (d = 1)	$\log \Gamma(\theta + 1) + (\theta + 1) \log 2$	$\Psi_0(\theta_\pi - \eta \sum_{i=1}^t g_i + 1) + \log 2$
Centered laplacian (d = 1)	$\log(-\frac{2}{\boldsymbol{\theta}})$	$\frac{1}{-\theta_\pi + \eta \sum_{i=1}^t g_i}$
Negative entropy	$\sum_{i=1}^d \boldsymbol{\theta}_i \log \boldsymbol{\theta}_i$	$\log(\boldsymbol{\theta}_\pi - \eta \sum_{i=1}^t g_i) + 1$
Negative binomial (d = 1) with number of failures r	$-r \log(1 - e^{-\boldsymbol{\theta}})$	$-re^{-\boldsymbol{\theta}_\pi + \eta \sum_{i=1}^t g_i} - 1$

Table 6.1: *Prior distributions, their cumulant generating functions, and updated means. As for notiation:* Γ *denotes the gamma function and* Ψ_0 *denotes the polygamma function of order 0.*

Examples of prior distributions, their cumulant generating functions, and updated means can be found in Table 6.1. Most distributions are given in one dimension, but the cumulant generating functions are easily extended to multiple dimensions. To extend the cumulant generating functions we simply evaluate F at every element of the natural parameter vector and sum the results. The mean becomes a vector that is the derivative of F evaluated at the elements of the natural parameter vector. Figure 1 shows the effect on the mean as $\eta \sum_{i=1}^t g_i$ changes. For example, we see that the for the standard multivariate normal the mean just linearly grows as we get closer to the minimum of the loss functions. A different pattern occurs for the Poisson: the mean grows exponentially as the loss goes to the minimum. The other distributions have upper bounded means, which tells us that no single element in $\boldsymbol{z}$ can grow too important for subsequent rounds.

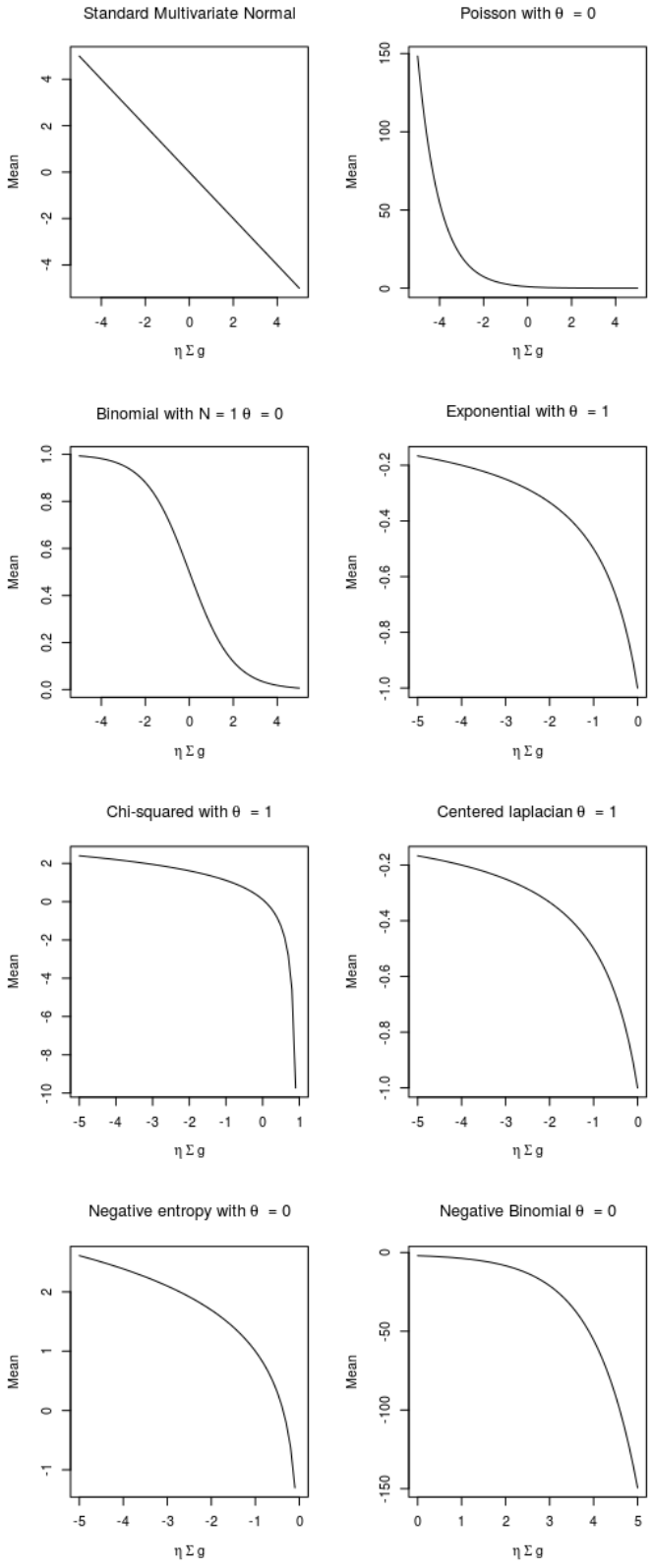


Figure 6.4: *Changes in mean as the sum of* $\eta \sum_{i=1}^{t} g_i$ *changes for the distributions in Table 6.1 .*

Chapter 7

Conclusion and Future work

In Theorem 3 we have shown that the Mirror Descent algorithm, under a single condition, can be seen as the mean of the Exponential Weights algorithm. Our analysis gives a large class of algorithms in the Online Convex Optimization setting a new interpretation: the update step is updating a distribution and taking the mean of said distribution as the weights for the coming round. Furthermore, two relatively simple and constructive sufficient conditions to see whether a Legendre function is a cumulant generating function were given in Theorems 7 and 8. These sufficient conditions are constructive in that they provide methods to find the corresponding prior distribution of a Legendre function. We have provided simple (and fast) update steps for the EW algorithm. Only the natural parameter of the prior distribution needs to be updated. This gives the EW algorithm a computational complexity that is linear in the number of rounds which makes EW scalable and approriate for large scale machine learning tasks. The application and implementation of the EW algorithm is significantly simplified by this development.

As for future work, this interpretation of EW provides a means to export extensions developed in the EW world to the MD world. For example: recently Koolen and Van Erven (2015) developed a means to learn the learning rate η, which was regarded as a fixed constant in this thesis. Or one could image a problem in which the experts are not independent of each other. A prior distribution with corresponding means that models this dependence probably reduces the regret. An example of such a distribution would be a multivariate normal with covariance other than the identity matrix. However, to learn the covariance matrix the algorithm probably needs some adjustments. The relationship between cumulant generating functions and Legendre functions can probably be further developed. Grünwald and Dawid (2004) section 7.4.1 gives an alternative representation of the cumulant generating

function. Here, the cumulant generating function is represented as the convex dual of the Kullback Leibler divergence. It is easy to show that every (convex dual of) a Legendre function can be seen as a generalized entropy (Dawid, 1998). However, it is difficult to prove that this generalized entropy is always a Kullback Leibler divergence. Reid et al. (2014) conjecture that, up to a constant, every generalized entropy related to a Legendre Function has the same associated loss function. If this were the case then the convex dual of every Legendre function, would be a Kullback Leibler divergence. Nevertheless, it is only a conjecture. Future work will have to prove this true or false.

Bibliography

Akhiezer, N. (1965). The classical moment problem and some related questions in analysis. 1965. *Hafner, New York*.

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

Barndorff-Nielsen, O. (1978). *Information and exponential families in statistical theory*. John Wiley & Sons.

Barry, C. (2015). *invLT: Inversion of Laplace-Transformed Functions*. R package version 0.2.1.

Billingsley, P. (2008). *Probability and measure*. John Wiley & Sons.

Blæsild, P. and Jensen, J. L. (1985). Saddlepoint formulas for reproductive exponential models. *Scandinavian journal of statistics*, pages 193–202.

Bochner, S. (1933). Monotone funktionen, stieltjessche integrale und harmonische analyse. *Mathematische Annalen*, 108(1):378–410.

Bregman, L. M. (1967). The relaxation method of 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(3):200–217.

Cesa-Bianchi, N. and Lugosi, G. (2006). *Prediction, learning, and games*. Cambridge university press.

Daniels, H. E. (1954). Saddlepoint approximations in statistics. *The Annals of Mathematical Statistics*, pages 631–650.

Dawid, A. P. (1998). Coherent measures of discrepancy, uncertainty and dependence, with applications to Bayesian predictive experimental design. *Department of Statistical Science, University College London. http://www. ucl. ac. uk/Stats/research/abs94. html, Tech. Rep*, 139.

Devinatz, A. et al. (1955). The representation of functions as a laplace-stieltjes integrals. *Duke Mathematical Journal*, 22(2):185–191.

Ehm, W., Genton, M. G., Gneiting, T., et al. (2003). Stationary covariances associated with exponentially convex functions. *Bernoulli*, 9(4):607–615.

Escher, F. (1932). On the probability function in the collective theory of risk. *Skandinavisk. Aktuarietidskrift*, 15:175–195.

Grünwald, P. D. (2007). *The minimum description length principle*. MIT press.

Grünwald, P. D. and Dawid, A. P. (2004). Game theory, maximum entropy, minimum discrepancy and robust bayesian decision theory. *Annals of Statistics*, pages 1367–1433.

Hazan, E. (2015). Introduction to online convex optimization. *Book draft*.

Jaynes, E. T. (1957). Information theory and statistical mechanics 1. *The Physical Review*, 106:620–630.

Jørgensen, B. and Labouriau, R. (2012). Exponential families and theoretical inference.

Koolen, W. (2016). Gradient descent as exponential weights. *Blog February 21: `http://blog.wouterkoolen.info/GDasEW/post.html/`*.

Koolen, W. M. and Van Erven, T. (2015). Second-order quantile methods for experts and combinatorial games. In *Proceedings of The 28th Conference on Learning Theory*, pages 1155–1175.

Kullback, S. and Leibler, R. A. (1951). On information and sufficiency. *The annals of mathematical statistics*, 22(1):79–86.

McCullagh, P. (1987). *Tensor methods in statistics*, volume 161. Chapman and Hall London.

Nielsen, F. and Nock, R. (2010). Entropies and cross-entropies of exponential families. In *Image Processing (ICIP), 2010 17th IEEE International Conference on*, pages 3621–3624. IEEE.

Parks, H. R. and Krantz, S. (1992). *A primer of real analytic functions*. Birkhäuser Verlag.

R Core Team (2016). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

Reid, M. D., Frongillo, R. M., Williamson, R. C., and Mehta, N. (2014). Generalized mixability via entropic duality. *arXiv preprint arXiv:1406.6130*.

Reid, N. (1988). Saddlepoint methods and statistical inference. *Statistical Science*, pages 213–227.

Rockafellar, R. T. (1970). *Convex analysis*. Princeton university press.

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

Shiryaev, A. N. (1996). *Probability, volume 95 of Graduate texts in mathematics*. Springer-Verlag, New York,.

Tonelli, L. (1909). Sullintegrazione per parti. *Atti della Accademia Nazionale dei Lincei*, 5(18):246–253.

Wendland, H. (2004). *Scattered data approximation*, volume 17. Cambridge university press.

Zinkevich, M. (2003). Online convex programming and generalized infinitesimal gradient ascent. *Proceedings of the Twenty International Conference on Machine Learning (ICML'03), (T. Fawcett and N. Mishra, eds.), pp. 928-936, AAAI Press*.