Tracking the Best Regressor*

Mark Herbster and **Manfred K. Warmuth**
Department of Computer Science
University of California at Santa Cruz
Applied Sciences Building
Santa Cruz, CA 95064 (USA)
mark@cs.ucsc.edu, manfred@cs.ucsc.edu

Abstract

In most of the on-line learning research the total on-line loss of the algorithm is compared to the total loss of the best off-line predictor $\boldsymbol{u}$ from a comparison class of predictors. We call such bounds static bounds. The interesting feature of these bounds is that they hold for an arbitrary sequence of examples. Recently some work has been done where the comparison vector $\boldsymbol{u}_t$ at each trial t is allowed to change with time, and the total on-line loss of the algorithm is compared to the sum of the losses of $\boldsymbol{u}_t$ at each trial plus the total "cost" for shifting to successive comparison vectors. This is to model situations in which the examples change over time and different predictors from the comparison class are best for different segments of the sequence of examples. We call such bounds shifting bounds. Shifting bounds still hold for arbitrary sequences of examples and also for arbitrary partitions. The algorithm does not know the off-line partition and the sequence of predictors that its performance is compared against.

Naturally shifting bounds are much harder to prove. The only known bounds are for the case when the comparison class consists of a finite sets of experts or boolean disjunctions. In this paper we develop the methodology for lifting known static bounds to the shifting case. In particular we obtain bounds when the comparison class consists of linear neurons (linear combinations of experts). Our essential technique consists of the following. At the end of each trial we *project* the hypothesis of the static algorithm into a suitably chosen convex region. This keeps the hypothesis of the algorithm well-behaved and the static bounds can be converted to shifting bounds so that the cost for shifting remains reasonable.

*The authors were supported by the NSF grant CCR-9700201.

COLT 98 Madison WI USA

1 INTRODUCTION

Consider the following by now standard on-line learning model which is a generalization of a model introduced by Littlestone [Lit89, Lit88, LW94]. Learning proceeds in trials $t = 1, 2, \ldots, \ell$. The algorithm maintains a parameter vector (hypothesis), denoted by $\boldsymbol{w}_t \in \Re^n$. In each trial the algorithm receives an *instance* $\mathbf{x}_t$. It then produces some action or a prediction for $\mathbf{x}_t$ based on $\boldsymbol{w}_t$. Finally the algorithm receives an *outcome* y_t and incurs a loss describing how well the hypothesis $\boldsymbol{w}_t$ "performed" on the *example* $(\mathbf{x}_t, y_t)$. This loss is some non-negative function $L(\boldsymbol{w}_t, (\mathbf{x}_t, y_t))$.

For example, the algorithm might predict with $\sigma(\boldsymbol{w}_t \cdot \mathbf{x}_t)$, where σ is the logistic function, and the loss might be the square loss, i.e. $\frac{1}{2}(\sigma(\boldsymbol{w}_t \cdot \mathbf{x}_t) - y_t)^2$. Thus in this example $\boldsymbol{w}$ represents a hypothesis $h_{\boldsymbol{w}}(\boldsymbol{x}) = \sigma(\boldsymbol{w} \cdot \boldsymbol{x})$. By choosing a different function σ, the hypotheses class of the algorithm changes. We abbreviate the loss in trial t by $L_t(\boldsymbol{w}_t)$ and the total loss of the algorithm A on a sequence $S = \langle(\boldsymbol{x}_1, y_1), \ldots, (\boldsymbol{x}_\ell, y_\ell)\rangle$ of examples by $L(A, S) = \sum_{t=1}^{\ell} L_t(\boldsymbol{w}_t)$.

In the methodology of worst-case loss bounds the total loss of the algorithm is expressed as a function of the total loss of any member in a comparison class of predictors, which is usually a subset of the hypotheses class. Such a predictor $\boldsymbol{u}$ also incurs a loss $L_t(\boldsymbol{u})$ in each trial and the total loss of $\boldsymbol{u}$ on the entire sequence is abbreviated as $L(\boldsymbol{u}, S) = \sum_{t=1}^{\ell} L_t(\boldsymbol{u})$. In the simplest case the bounds have the following form

$$L(A, S) \leq cL(\boldsymbol{u}, S) + c\,\mathrm{size}(\boldsymbol{u}). \tag{1}$$

Here c is a small constant, $\boldsymbol{u}$ is any vector in the comparison class, and $\mathrm{size}(\boldsymbol{u})$ is a measurement of the size of $\boldsymbol{u}$. We call such bounds *static* bounds, because the comparison vector $\boldsymbol{u}$ is any member from the comparison class but it does not change with time. Surprisingly, such bounds are achievable even when there are no probabilistic assumptions made on the sequence of examples [Lit88, Vov90, HKW97, CBLW96, KW94, Byl97, HKW95]. In this paper we allow the comparison vector $\boldsymbol{u}$ to shift with time. For a sequence S of examples of length ℓ and a schedule of predictors $\langle \boldsymbol{u}_1, \ldots, \boldsymbol{u}_\ell \rangle$ from the comparison class, we seek an upper bound of the form

$$L(A, S) \leq cL(\langle \boldsymbol{u}_1, \ldots, \boldsymbol{u}_\ell \rangle, S) + c\,\mathrm{size}(\langle \boldsymbol{u}_1, \ldots, \boldsymbol{u}_\ell \rangle). \tag{2}$$

Here $L(\langle \boldsymbol{u}_1, \ldots, \boldsymbol{u}_\ell \rangle, S)$ is the loss of the schedule of predictors on S, i.e. $L(\langle \boldsymbol{u}_1, \ldots, \boldsymbol{u}_\ell \rangle, S) = \sum_{t=1}^{\ell} L_t(\boldsymbol{u}_t)$, and, intuitively, size$(\langle \boldsymbol{u}_1, \ldots, \boldsymbol{u}_\ell \rangle)$ measures the amount of shifting that occurs in the schedule. We call such bounds *shifting* bounds. In this paper our bounds use the following measure of size$(\langle \boldsymbol{u}_1, \ldots, \boldsymbol{u}_\ell \rangle)$:

$$\|\langle \boldsymbol{u}_1, \ldots, \boldsymbol{u}_\ell \rangle\|_p = \sum_{t=1}^{\ell-1} \|\boldsymbol{u}_t - \boldsymbol{u}_{t+1}\|_p + \|\boldsymbol{u}_\ell\|_p,$$

where $\|\cdot\|_p$ is a p-norm. This is in accord with previous work on worst case loss bounds, where the static bound also grows with a p-norm of the comparison vector. The new shifting bounds presented in this paper build on previous work of the authors [HW98] where the loss of the algorithm was compared against the loss of the best shifting expert (see also [Vov97]) or the best shifting disjunction [AW98]. The work on shifting experts has been applied to predicting disk idle times [HLS96] as well as load balancing problems [BB97]. In this paper we discuss general methods of lifting static bounds to the more complicated shifting case.

We focus on a class of algorithms that is characterized by a *link function* f. Such functions are arbitrary mappings for which the domain is some open set in $\Re^n$, the range is a subset of $\Re^n$, and furthermore the Jacobian of f is strictly positive definite on the whole domain. The update for a given link function f is as follows:

$$\boldsymbol{w}_{t+1} = f^{-1}\left(f(\boldsymbol{w}_t) - \eta \nabla L_t(\boldsymbol{w}_t)\right). \tag{3}$$

Here, f^{-1} denotes the inverse of f and η is a non-negative learning rate. The general form of these updates was introduced in [GLS97] for linear threshold functions and in [JW98, KW97b] for regression problems. We refer to these updates as *generalized gradient descent algorithms*. For the sake of simplicity we focus on only two link functions in this paper, the identity function and the logarithm function. For the identity link $\mathrm{id}(\boldsymbol{w}) = \boldsymbol{w}$ the domain is all of $\Re^n$ and the above update becomes the standard Gradient Descent Algorithm (GD). For the logarithm link function the domain is $(0, \infty)^n$ and $\ln(\boldsymbol{w})$ is defined as the component-wise application of the ln function, i.e. $\ln(\boldsymbol{w}) = (\ln w_1, \ldots, \ln w_n)$. The generalized gradient descent algorithm for the ln link is the Un-normalized Exponentiated Gradient (EGU) Algorithm [KW97a] (see Figure 1).

The key tool for analyzing the update associated with a link function f is a divergence function introduced by Bregman [Bre67, CL81, Csi91]. These divergences were previously used in convex programming in connection with generalized maximum entropy methods. In the context of online learning various versions of these divergence functions were first used in [KW97b, GLS97, JW98]). At this point we introduce the notion of divergence without listing a number of technical conditions specified in the Appendix. Given a strictly convex function $F : E \to \Re$, where $E \subset \Re^n$ is a closed convex set and ri E denotes the relative interior of E, then the Bregman divergence $D_F : E \times \text{ri } E \to [0, \infty)$ is defined as

$$D_F(\boldsymbol{u}, \boldsymbol{w}) = F(\boldsymbol{u}) - F(\boldsymbol{w}) + (\boldsymbol{w} - \boldsymbol{u}, \nabla F(\boldsymbol{w})). \tag{4}$$

Here $(\boldsymbol{u}, \boldsymbol{w})$ denotes the usual dot product in $\Re^n$.

The link function f is simply the gradient of F, i.e. $f = \nabla F$. Note that adding the constant term to F does not change the corresponding link function. Also, when a linear term is added to F, the link function changes by an additive constant. However, adding a linear term to F does not change the divergence D_F and the corresponding update (3).

The convex function $sq(\boldsymbol{w}) = \frac{1}{2}\sum_{i=1}^{n} {w_i}^2$ corresponds to the identity link, and the divergence for this choice of F is the squared Euclidean distance, denoted as $D_{\mathrm{sq}}(\boldsymbol{u}, \boldsymbol{w}) = \sum_{i=1}^{n} \frac{1}{2}(u_i - w_i)^2$. Similarly, the convex function $ne(\boldsymbol{w}) = \sum_{i=1}^{n} w_i \ln w_i - w_i$ (essentially the negative entropy) gives rise to the logarithm link and the unnormalized relative entropy as the divergence function (see Figure 1): $D_{\mathrm{ne}} = \sum_{i=1}^{n} (u_i \ln \frac{u_i}{w_i} + w_i - u_i)$.

The divergence functions D_F are used as potential functions for proving static loss bounds for the corresponding generalized gradient descent algorithms[CBLW96, KW97a, HKW95, Byl97, KW97b]. As in the previous work in convex programming, we use projections based on these divergences.

We now outline how worst-case loss bounds are obtained in the static case. At the center of all the static proofs [KW94, HKW95, Byl97, KW97b] for the generalized gradient descent algorithms lies the following type of inequality:

$$a(\eta)\, L_t(\boldsymbol{w}_t) - b(\eta)\, L_t(\boldsymbol{u}) \leq D_F(\boldsymbol{u}, \boldsymbol{w}_t) - D_F(\boldsymbol{u}, \boldsymbol{w}_{t+1}). \tag{5}$$

Here a and b are non-negative functions that depend on the learning rate η and upper bounds on the norms of the instances. Note that $D_F(\boldsymbol{u}, \boldsymbol{w}_t) - D_F(\boldsymbol{u}, \boldsymbol{w}_{t+1})$ may be seen as the progress towards the off-line weight vector $\boldsymbol{u}$. Since this inequality holds for each trial, we can sum Equation (5) over trials. The progresses towards $\boldsymbol{u}$ that appear on the right-hand sides form a telescoping sum and only the first and last terms survive:

$$a(\eta) L(A, S) - b(\eta) L(\langle \boldsymbol{u}_1, \ldots, \boldsymbol{u}_\ell \rangle, S) \leq D_F(\boldsymbol{u}, \boldsymbol{w}_1) - D_F(\boldsymbol{u}, \boldsymbol{w}_{\ell+1}). \tag{6}$$

The last term $D_F(\boldsymbol{u}, \boldsymbol{w}_{\ell+1})$ can be dropped, since the divergence is non-negative and we are forming an upper bound. The first term $D_F(\boldsymbol{u}, \boldsymbol{w}_1)$ plays the role of the size measure of $\boldsymbol{u}$. Rearranging the above and picking a good choice for η gives bounds of the form (1).

The new approach (which allows us to prove bounds for the shifting case) starts with picking a suitable convex region Γ. After the usual update the hypothesis vector is projected into this convex region. This gives us two benefits: first, we can show that the generalized gradient descent algorithms may be modified to maintain convex constraints on the hypothesis vector. Second, we show that shifting bounds may readily be obtained. The convex region plays a key role in bounding the size of the schedule of predictors.

Definition and properties of the Projection Update

The projection of a point x onto a closed convex set Γ is simply the point in Γ closest to x w.r.t. the divergence D_F.

Definition 1.1. *The projection of a point $\boldsymbol{w}$ w.r.t. divergence $D_F : E \times ri\ E \to [0, \infty)$ where $E \subseteq \Re^n$ onto a closed convex set Γ is defined by:*

$$P_{(\Gamma, D_F)}(\boldsymbol{w}) = argmin_{\boldsymbol{u} \in \Gamma \cap E} D_F(\boldsymbol{u}, \boldsymbol{w}). \tag{7}$$

In the Appendix we show that the projection exists and is unique.

Recall the standard Pythagorean Theorem for the divergence associated with the identity link:

$$\|\boldsymbol{u}-\boldsymbol{w}\|^2 = \|\boldsymbol{u}-\boldsymbol{p}\|^2 + \|\boldsymbol{p}-\boldsymbol{w}\|^2, \text{ when } \boldsymbol{u}-\boldsymbol{p} \perp \boldsymbol{p}-\boldsymbol{w}.$$

The orthogonality condition can be rewritten as $\boldsymbol{p}$ being a projection of $\boldsymbol{w}$ onto a plane that contains $\boldsymbol{u}$. Here projections are w.r.t. the (squared) Euclidean distance. Surprisingly, the Pythagorean Theorem holds for projections w.r.t. D_F divergences.

Theorem 1.2. (Generalized Pythagoras) *Given a divergence $D_F : E \times ri\, E \to [0,\infty)$, a closed convex set Γ such that $\Gamma \bigcap ri\, E \neq \emptyset$, and points $\boldsymbol{w} \in ri\, E$ and $\boldsymbol{u} \in \Gamma$, then*

$$D_F(\boldsymbol{u},\boldsymbol{w}) \geq D_F(\boldsymbol{u}, P_{(\Gamma,D_F)}(\boldsymbol{w})) + D_F(P_{(\Gamma,D_F)}(\boldsymbol{w}),\boldsymbol{w}). \tag{8}$$

In the special case when Γ is a hyperplane then the above becomes an equality.

Note that this inequality is opposite to the triangle inequality that holds for a (distance) metric. This is the reason that we call D_F a divergence instead of a distance.

The above theorem has been proven many times under a variety of assumptions, see for instance [Bre67, Csi91, JB90]. For the sake of completeness we provide a streamlined proof in the Appendix. The below corollary of the Generalized Pythagorean Theorem will be used repeatedly in the analyses of our algorithms.

Corollary 1.3. *Given a divergence $D_F : E \times ri\, E \to [0,\infty)$, a closed convex set Γ such that $\Gamma \bigcap ri\, E \neq \emptyset$, and points $\boldsymbol{w} \in ri\, E$ and $\boldsymbol{u} \in \Gamma$, then*

$$D_F(\boldsymbol{u},\boldsymbol{w}) - D_F(\boldsymbol{u}, P_{(\Gamma,D_F)}(\boldsymbol{w})) \geq 0 \tag{9}$$

Using the above definition of projection we now introduce our modification of the generalized gradient descent algorithms:

$$\begin{aligned} \boldsymbol{w}_t^m &= f^{-1}\left(f(\boldsymbol{w}_t) - \eta \nabla L_t(\boldsymbol{w}_t)\right) && \text{gen.grad.desc.upd.} \\ \boldsymbol{w}_{t+1} &= P_{(\Gamma,D_F)}(\boldsymbol{w}_t^m) && \text{proj.upd.} \end{aligned} \tag{10}$$

Since we now have two updates, we will refer to the intermediate weight vector following the generalized gradient update as $\boldsymbol{w}_t^m$. We call this algorithm the Constrained Generalized Gradient Descent Algorithm. We use C-GD(Γ) and C-EGU(Γ) for the example algorithms of this paper.

The constraint sets Γ_{gd} and Γ_{egu} chosen for to obtain shifting bounds for GD and EGU, respectively, are intimately related to the static loss bounds proven for the original algorithms. In [KW97a] it is observed that the bounds of GD depend on the 2-norm of both the instances $\mathbf{x}_t$ and the predictor $\boldsymbol{u}$, while in EGU (as well as EG) the bounds depend on the ∞-norm of the instances $\mathbf{x}_t$ and the 1-norm of the predictor $\boldsymbol{u}$. The constraint set is associated with the predictor and the norm dual to the predictor's norm. Thus the constraint set Γ_{gd} is the level set of the two-norm, and the constraint set Γ_{egu} is a shifted level set of the ∞-norm. Our worst-case loss bounds for C-GD(Γ_{gd}) thus depend on $\|\langle \boldsymbol{u}_1,\ldots,\boldsymbol{u}_\ell\rangle\|_2$ and $\|\mathbf{x}_t\|_2$, and the bounds for C-EGU(Γ_{egu}) depend on $\|\langle \boldsymbol{u}_1,\ldots,\boldsymbol{u}_\ell\rangle\|_1$ of the predictor schedules and $\|\mathbf{x}_t\|_\infty$.

Calculating the projection $P_{(\Gamma,D_F)}(\boldsymbol{w})$ may be computationally expensive. For our example algorithms the projections $P_{(\Gamma_{\text{gd}},D_{\text{sq}})}(\boldsymbol{w})$ and $P_{(\Gamma_{\text{egu}},D_{\text{ne}})}(\boldsymbol{w})$ are easily computable (see Figure 1) in $O(n)$ time. In particular, the projection update $P_{(\Gamma_{\text{egu}},D_{\text{ne}})}(\boldsymbol{w})$ corresponds to simple weight clipping. Similar updates were used in the Fixed-share Algorithm [HW98], WML [LW94] and the algorithms for tracking disjunctions [AW98].

An overview of proof techniques based on projections

We have three results. First, we show that if the predictor $\boldsymbol{u}$ from the comparison class lies in the constraint set, then the previous static loss bounds hold for the constraint algorithm. Second, we prove a shifting bound on a predictor schedule $\langle \boldsymbol{u}_1,\ldots,\boldsymbol{u}_\ell\rangle$ whose predictors must all lie in Γ. These results follow almost immediately from properties of projections and the static bounds. Third (more tricky), we show for C-EGU(Γ_{egu}) that the shifting bound may be extended from predictors in the constraint set $\Gamma_{\text{egu}} = [\frac{\gamma}{n},\frac{n}{\gamma}]^n$ to predictors in the larger set $[0,\frac{n}{\gamma}]^n$. Interestingly, in this case the shifting loss bound obtained via the projection update of this paper is similar to the loss bound of the Fixed-share Algorithm[HW98].

For the first part recall that $\boldsymbol{w}_t$ is the weight vector at the start of a trial, and $\boldsymbol{w}_t^m$ the weight vector between the two updates. With this notation, Inequality (5) becomes

$$a(\eta)L_t(\boldsymbol{w}_t) - b(\eta)L_t(\boldsymbol{u}_t) \leq D_F(\boldsymbol{u},\boldsymbol{w}_t) - D_F(\boldsymbol{u},\boldsymbol{w}_t^m). \tag{15}$$

Note that this inequality no longer telescopes. By Corollary 1.3 we have

$$0 \leq D_F(\boldsymbol{u},\boldsymbol{w}_t^m) - D_F(\boldsymbol{u},\boldsymbol{w}_{t+1}),$$

provided that $\boldsymbol{u}$ lies in the constraint set. The sum of the two inequalities is again a telescoping inequality with the same functions a and b:

$$a(\eta)L_t(\boldsymbol{w}_t) - b(\eta)L_t(\boldsymbol{u}_t) \leq D_F(\boldsymbol{u},\boldsymbol{w}_t) - D_F(\boldsymbol{u},\boldsymbol{w}_{t+1}). \tag{16}$$

Thus the constraint algorithms have the same static bounds, provided that the comparison vector $\boldsymbol{u}$ lies in the constraint set.

For the second part we assume that all the predictors of predictor schedule $\langle \boldsymbol{u}_1,\ldots,\boldsymbol{u}_\ell\rangle$ lie in the constraint set. Thus Equation (16) holds for each trial:

$$a(\eta)L_t(\boldsymbol{w}_t) - b(\eta)L_t(\boldsymbol{u}_t) \leq D_F(\boldsymbol{u}_t,\boldsymbol{w}_t) - D_F(\boldsymbol{u}_t,\boldsymbol{w}_{t+1}). \tag{17}$$

This again does not telescope. We fix this by adding

$$G_F(\boldsymbol{u}_t,\boldsymbol{u}_{t+1}) := D_F(\boldsymbol{u}_t,\boldsymbol{w}_{t+1}) - D_F(\boldsymbol{u}_{t+1},\boldsymbol{w}_{t+1}). \tag{18}$$

Intuitively, $G_F(\boldsymbol{u}_t,\boldsymbol{u}_{t+1})$ measures the cost for shifting from $\boldsymbol{u}_t$ to $\boldsymbol{u}_{t+1}$. We will return to a discussion of this cost later.

Convex Function $F(\boldsymbol{w})$	$\mathrm{ne}(\boldsymbol{w}) = \sum_{i=1}^{n} w_i \ln w_i - w_i$, $\boldsymbol{w} \in [0,\infty)^n$	$\mathrm{sq}(\boldsymbol{w}) = \frac{1}{2}\sum_{i=1}^{n} {w_i}^2$, $\boldsymbol{w} \in \Re^n$
Link Function $f(\boldsymbol{w})$	$\ln(\boldsymbol{w}) = (\ln w_1, \ldots, w_n)$, $\boldsymbol{w} \in (0,\infty)^n$	$\mathrm{id}(\boldsymbol{w}) = \boldsymbol{w}$, $\boldsymbol{w} \in \Re^n$
$D_F(\boldsymbol{u}, \boldsymbol{w})$	$D_{\mathrm{ne}}(\boldsymbol{u}, \boldsymbol{w}) = \sum_{i=1}^{n} (u_i \ln \frac{u_i}{w_i} + w_i - u_i)$	$D_{\mathrm{sq}}(\boldsymbol{u}, \boldsymbol{w}) = \sum_{i=1}^{n} \frac{1}{2}(u_i - w_i)^2$
Gen. Grad. Desc. upd.	$w^m_{t+1,i} = w_{t,i} e^{-\eta \nabla_{t,i}}$ (11)	$w^m_{t+1,i} = w_{t,i} - \nabla_{t,i}$ (12)
Constraint Set Γ	$\Gamma_{\mathrm{egu}} = [\frac{\gamma}{n}, \frac{n}{\gamma}]^n$	$\Gamma_{\mathrm{gd}} = \{\boldsymbol{u} : \|\boldsymbol{u}\|_2 \leq \gamma\}$
Projection Update $\boldsymbol{w}_{t+1} = P_{(\Gamma, D_F)}(\boldsymbol{w}^m_t)$	$\forall i : 1, \ldots, n : w_{t+1,i} = \begin{cases} w^m_{t,i} & w^m_{t,i} \in [\frac{\gamma}{n}, \frac{n}{\gamma}] \\ \frac{\gamma}{n} & w^m_{t,i} < \frac{\gamma}{n} \\ \frac{n}{\gamma} & w^m_{t,i} > \frac{n}{\gamma} \end{cases}$ (13)	$\boldsymbol{w}_{t+1} = \begin{cases} \boldsymbol{w}^m_t & \boldsymbol{w}^m_t \in \Gamma_{\mathrm{gd}} \\ \frac{\gamma \boldsymbol{w}^m_t}{\|\boldsymbol{w}^m_t\|_2} & \boldsymbol{w}^m_t \notin \Gamma_{\mathrm{gd}} \end{cases}$ (14)

Figure 1: Two constrained generalized gradient descent algorithms C-EGU(Γ_{egu}) and C-GD(Γ_{gd}), based on the Unnormalized Exponentiated Gradient and the Gradient Descent. Here $\nabla_{t,i}$ is shorthand for $\frac{\partial L_t(\boldsymbol{w}_t)}{\partial w_{t,i}}$. In the case of the square loss, $\nabla_{t,i} = 2x_{t,i}(\boldsymbol{w}_t \cdot \mathbf{x}_t - y_t)$.

By adding the Equation (17) and (18) and by summing over all trials we get the following:

$$L(A,S) \leq \frac{b(\eta)}{a(\eta)} L(\langle \boldsymbol{u}_1, \ldots, \boldsymbol{u}_\ell \rangle, S)[\frac{1}{a(\eta)}[D_F(\boldsymbol{u}_1, \boldsymbol{w}_1) - D_F(\boldsymbol{u}_{\ell+1}, \boldsymbol{w}_{\ell+1}) - \sum_{t=1}^{\ell} G_F(\boldsymbol{u}_t, \boldsymbol{u}_{t+1})]. \quad (19)$$

This would be a good bound if $G_F(\boldsymbol{u}_t, \boldsymbol{u}_{t+1})$ was lower bounded. Unfortunately this is not true for arbitrary $\boldsymbol{w}_{t+1}$. For example, for the squared Euclidean distance,

$$G_{\mathrm{sq}}(\boldsymbol{u}_t, \boldsymbol{u}_{t+1}) = \|\boldsymbol{u}_t\|_2^2 - \|\boldsymbol{u}_{t+1}\|_2^2 - 2\boldsymbol{w}_{t+1} \cdot (\boldsymbol{u}_t - \boldsymbol{u}_{t+1}). \quad (20)$$

Observe that unless we have a bound on some norm of $\boldsymbol{w}_{t+1}$, the cost $G_{\mathrm{sq}}(\boldsymbol{u}_t, \boldsymbol{u}_{t+1})$ cannot be lower bounded solely in terms of $\boldsymbol{u}_t$ and $\boldsymbol{u}_{t+1}$. However, $\boldsymbol{w}_{t+1} \in \Gamma_{\mathrm{gd}}$ and thus by our choice of the constraint set Γ_{gd}, $\|\boldsymbol{w}_{t+1}\|_2 \leq \gamma$. Through an application of Hölder's inequality, we have

$$G_{\mathrm{sq}}(\boldsymbol{u}_t, \boldsymbol{u}_{t+1}) \geq \|\boldsymbol{u}_t\|_2^2 - \|\boldsymbol{u}_{t+1}\|_2^2 - 2\gamma \|\boldsymbol{u}_t - \boldsymbol{u}_{t+1}\|_2, \quad (21)$$

and plugging into Equation (19) we obtain a reasonable upper bound of the form (2):

$$L(\text{C-GD}(\Gamma_{\mathrm{gd}}), S) \leq \frac{b(\eta)}{a(\eta)} L(\langle \boldsymbol{u}_1, \ldots, \boldsymbol{u}_\ell \rangle, S) + \frac{1}{a(\eta)}[D_{\mathrm{sq}}(\boldsymbol{u}_1, \boldsymbol{w}_1) - D_{\mathrm{sq}}(\boldsymbol{u}_{\ell+1}, \boldsymbol{w}_{\ell+1}) + \|\boldsymbol{u}_{\ell+1}\|_2^2 - \|\boldsymbol{u}_1\|_2^2 + 2\gamma \sum_{t=1}^{\ell} \|\boldsymbol{u}_t - \boldsymbol{u}_{t+1}\|_2].$$

Part two for the divergence function D_{ne} is similar. Again the constraint set is used to obtain a lower bound of the G function: $Gegu(\boldsymbol{u}_t, \boldsymbol{u}_{t+1}) \geq -H(\boldsymbol{u}_t) + H(\boldsymbol{u}_{t+1}) - \|\boldsymbol{u}_t\|_1 + \|\boldsymbol{u}_{t+1}\|_1 - \ln(\frac{n}{\gamma})\|\boldsymbol{u}_t - \boldsymbol{u}_{t+1}\|_1$, where H denotes the entropy function. By plugging into (19) we obtain

$$L(\text{C-EGU}(\Gamma_{\mathrm{egu}}), S) \leq \frac{b(\eta)}{a(\eta)} L(\langle \boldsymbol{u}_1, \ldots, \boldsymbol{u}_\ell \rangle, S) + \frac{1}{a(\eta)}[D_{\mathrm{ne}}(\boldsymbol{u}_1, \boldsymbol{w}_1) - D_{\mathrm{ne}}(\boldsymbol{u}_{\ell+1}, \boldsymbol{w}_{\ell+1}) + H(\boldsymbol{u}_1) - H(\boldsymbol{u}_{\ell+1}) + \|\boldsymbol{u}_0\|_1 - \|\boldsymbol{u}_{\ell+1}\|_1 + \ln(\frac{n}{\gamma}) \sum_{t=1}^{\ell} \|\boldsymbol{u}_t - \boldsymbol{u}_{t+1}\|_1]. \quad (22)$$

The third part is more involved and is covered in detail in the rest of the paper. We only give a rough outline in this section. The analysis applies directly to EGU and extends the comparison class from $\Gamma_{\mathrm{egu}} = [\frac{\gamma}{n}, \frac{n}{\gamma}]^n$ to $[0, \frac{n}{\gamma}]^n$. Since the comparison vectors do not lie in the constraint set we cannot use the Pythagorean Theorem. Instead, we directly lower bound $G_{\mathrm{ne}}(\boldsymbol{u}_t, \boldsymbol{u}_{t+1})$. This lower bound will contain a constant term $-\gamma$, having the effect that our loss bounds for the extended concept class grow with the sequence length. Careful tunings of γ ameliorate this effect.

2 SHIFTING C-EGU(Γ_{egu}) FOR PREDICTORS IN $[0, \frac{n}{\gamma}]^n$

We begin by proving a lower bound for $G_{\mathrm{ne}}(\boldsymbol{u}_t, \boldsymbol{u}_{t+1})$ as defined in Equation (18) when $\boldsymbol{u}_t$ and $\boldsymbol{u}_{t+1}$ lie in $[0, \frac{n}{\gamma}]^n$ instead of $\Gamma_{\mathrm{egu}} = [\frac{\gamma}{n}, \frac{n}{\gamma}]^n$

Lemma 2.1. *Let $\boldsymbol{w} \in [0,\infty)^n$, and let $\boldsymbol{u}_t, \boldsymbol{u}_{t+1} \in [0, \frac{n}{\gamma}]^n$. Then*

$$G_{ne}(\boldsymbol{u}_t, \boldsymbol{u}_{t+1}) \geq -H(\boldsymbol{u}_t) + H(\boldsymbol{u}_{t+1}) - \|\boldsymbol{u}_t\|_1 + \|\boldsymbol{u}_{t+1}\|_1 - \ln(\frac{n}{\gamma})\|\boldsymbol{u}_t - \boldsymbol{u}_{t+1}\|_1 - \gamma. \quad (23)$$

Proof. Note that the divergence D_{ne} is a simple sum over coordinates, and the constraint set Γ_{egu} is the same in each

coordinate. By a one-dimensional argument one can show that

$$P_{(\Gamma_{\mathrm{egu}},D_{\mathrm{ne}})}(\boldsymbol{w}) = P_{([\frac{\gamma}{n},\infty)^n,D_{\mathrm{ne}})}(P_{([0,\frac{n}{\gamma}]^n,D_{\mathrm{ne}})}(\boldsymbol{w})).$$

Let $\boldsymbol{w}' = P_{([0,\frac{n}{\gamma}]^n,D_{\mathrm{ne}})}(\boldsymbol{w})$, and let $\boldsymbol{w}'' = P_{([\frac{\gamma}{n},\infty)^n,D_{\mathrm{ne}})}(\boldsymbol{w}')$. The sum of the following three inequalities gives Inequality (23), thus proving the lemma:

$$D_{\mathrm{ne}}(\boldsymbol{u}_t,\boldsymbol{w}) - D_{\mathrm{ne}}(\boldsymbol{u}_t,\boldsymbol{w}') \geq 0 \tag{24}$$

$$D_{\mathrm{ne}}(\boldsymbol{u}_t,\boldsymbol{w}') - D_{\mathrm{ne}}(\boldsymbol{u}_t,\boldsymbol{w}'') \geq -\gamma \tag{25}$$

$$D_{\mathrm{ne}}(\boldsymbol{u}_t,\boldsymbol{w}'') - D_{\mathrm{ne}}(\boldsymbol{u}_{t+1},\boldsymbol{w}'') \geq -H(\boldsymbol{u}_t) + H(\boldsymbol{u}_{t+1})$$
$$- \|\boldsymbol{u}_t\|_1 + \|\boldsymbol{u}_{t+1}\|_1 - \ln(\frac{n}{\gamma})\|\boldsymbol{u}_t - \boldsymbol{u}_{t+1}\|_1. \tag{26}$$

Inequality (24) holds by Corollary 1.3 when all $\boldsymbol{u}_t$ lie in the constraint set $[0,\frac{n}{\gamma}]^n$. Recall that the latter convex set was used as the constraint set for defining $\boldsymbol{w}'$ as a projection.

For Inequality (25) we expand the definition of D_{ne}:

$$D_{\mathrm{ne}}(\boldsymbol{u}_t,\boldsymbol{w}') - D_{\mathrm{ne}}(\boldsymbol{u}_t,\boldsymbol{w}'') = \sum_{i=1}^{n}(w'_i - w''_i) + \sum_{i=1}^{n} u_{t,i}\ln\frac{w''_i}{w'_i}$$

From the definition of $\boldsymbol{w}'$ and $\boldsymbol{w}''$ it follows that $w'_i - w''_i \geq -\frac{\gamma}{n}$ and $w''_i - w'_i \geq 0$. Thus the first sum is lower bounded by $-\gamma$ and the second sum is lower bounded by zero, giving us Inequality (25).

For the proof of Inequality (26) we expand D_{ne}, apply Hölder's inequality and the fact that $w''_i \in [\frac{\gamma}{n},\frac{n}{\gamma}]$:

$$\begin{aligned} &D_{\mathrm{ne}}(\boldsymbol{u}_t,\boldsymbol{w}'') - D_{\mathrm{ne}}(\boldsymbol{u}_{t+1},\boldsymbol{w}'') \\ &\quad = -H(\boldsymbol{u}_t) + H(\boldsymbol{u}_{t+1}) - \|\boldsymbol{u}_t\|_1 + \|\boldsymbol{u}_{t+1}\|_1 + \\ &\sum_{i=1}^{n}\left[(u_{t,i} - u_{t+1,i})\ln\frac{1}{w''_i}\right] \\ &\quad \geq -H(\boldsymbol{u}_t) + H(\boldsymbol{u}_{t+1}) - \|\boldsymbol{u}_t\|_1 + \|\boldsymbol{u}_{t+1}\|_1 - \\ &\qquad \ln(\frac{n}{\gamma})\|\boldsymbol{u}_t - \boldsymbol{u}_{t+1}\|_1. \end{aligned}$$

□

By plugging the lower bound of the lemma into Equation (19) we obtain an upper bound which is the same as Equation (22), except for the additional term $\frac{\gamma\ell}{a(\eta)}$:

For all predictor schedules $\langle \boldsymbol{u}_1,\ldots,\boldsymbol{u}_\ell\rangle \in \{[0,\frac{n}{\gamma}]^n\}^\ell$,

$$L(\text{C-EGU}(\Gamma_{\mathrm{egu}}),S) \leq \frac{b(\eta)}{a(\eta)}L(\langle \boldsymbol{u}_1,\ldots,\boldsymbol{u}_\ell\rangle,S)$$
$$+ \frac{1}{a(\eta)}[D_{\mathrm{ne}}(\boldsymbol{u}_1,\boldsymbol{w}_1) - D_{\mathrm{ne}}(\boldsymbol{u}_{\ell+1},\boldsymbol{w}_{\ell+1}) + H(\boldsymbol{u}_1) -$$
$$H(\boldsymbol{u}_{\ell+1}) + \|\boldsymbol{u}_0\|_1 - \|\boldsymbol{u}_{\ell+1}\|_1 + \ln(\frac{n}{\gamma})\sum_{t=1}^{\ell}\|\boldsymbol{u}_t - \boldsymbol{u}_{t+1}\|_1 + \gamma\ell].$$

Finally, we give a tuning of γ for the purpose of minimizing the above . The tuning involves two parameters, $\hat{\ell}$ and $\hat{U}$, where $\hat{\ell}$ is an upper bound on the length of the trial sequence and $\hat{U}$ is an approximation to $\|\boldsymbol{u}_1\|_1 + \|\langle \boldsymbol{u}_1,\ldots,\boldsymbol{u}_\ell\rangle\|_1$. We set $\gamma = \frac{\hat{U}}{\hat{U}+\hat{\ell}}$ and $\boldsymbol{w}_1 = (\frac{\gamma}{n},\ldots,\frac{\gamma}{n})$. Then for all predictor schedules $\langle \boldsymbol{u}_1,\ldots,\boldsymbol{u}_\ell\rangle \in \{[0,\frac{n}{\gamma}]^n\}^\ell$, provided that $\hat{U} \leq \hat{\ell}$, we have the bound

$$L(\text{C-EGU}(\Gamma_{\mathrm{egu}}),S) \leq \frac{b(\eta)}{a(\eta)}L(\langle \boldsymbol{u}_1,\ldots,\boldsymbol{u}_\ell\rangle,S) +$$
$$\frac{1}{a(\eta)}([\|\boldsymbol{u}_1\|_1 + \|\langle \boldsymbol{u}_1,\ldots,\boldsymbol{u}_\ell\rangle\|_1](\ln n + \ln(\frac{\hat{\ell}}{\hat{U}}) + \ln 2) + \hat{U}).$$

3 OPEN PROBLEMS

We feel that we are developing important methods based on projections that can be used for proving worst-case loss bounds when the predictor from the comparison class is allowed to shift. These methods apply to such algorithms as the WM Algorithm[LW94], the Aggregating Algorithm[Vov95], the Hedge Algorithm[FS97], and various exponentiated gradient algorithms [KW97a, HKW95, Byl97, KW97b], as well as Winnow [Lit88]. Just the static bounds for the constraint algorithms are already interesting in their own right. A number of detailed applications will be included in the full paper.

There were two algorithms for the shifting expert setting. The Fixed-share Algorithm, which essentially lower bounded the weights, and the Variable-share Algorithm, which lets the weights of the algorithm go to zero. So far we have only developed the methodology for the updates corresponding to the Fixed-share Algorithm (i.e. we use clipping to define the convex region Γ_{egu} associated with the exponentiated gradient algorithms). For the alternate bounds (in progress) we want to let the weights go close to zero and only pull back the weights when the master incurs large loss. Thus we need to use a dynamic convex region that we project into at end of each trial.

A nice application for the projection update may be made to prediction portfolios on-line[Cov91, HSSW96]. In this case linear inequality constraints may be used to express relations that must be maintained between the instruments of the portfolio.

Acknowledgments We would like to thank Claudio Gentile, Andrew Klingler, Debra Lewis and Harold Widom for valuable discussions.

References

[AW98] P. Auer and M. K. Warmuth. Tracking the best disjunction. *Journal of Machine Learning*, 1998. Special issue on concept drift.

[BB97] A. Blum and C. Burch. On-line learning and the metrical task system. In *Proc. 10th Annu. Workshop on Comput. Learning Theory*. ACM Press, New York, NY, 1997.

[Bre67] L.M. Bregman. 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 Physics*, 7:200–217, 1967.

[Byl97] T. Bylander. The binary exponentiated gradient algorithm for learning linear functions. In *Proc. 8th Annu. Conf. on Comput. Learning Theory*, pages 184–192. ACM, 1997.

[CBLW96] N. Cesa-Bianchi, P. Long, and M.K. Warmuth. Worst-case quadratic loss bounds for on-line prediction of linear functions by gradient descent. *IEEE Transactions on Neural Networks*, 7(2):604–619, May 1996.

[CL81] Y. Censor and A. Lent. An iterative row-action method for interval convex programming. *Journal of Optimization Theory and Applications*, 34(3):321–353, July 1981.

[Cov91] T. M. Cover. Universal portfolios. *Mathematical Finance*, 1(1):1–29, 1991.

[Csi91] Imre Csiszar. Why least squares and maximum entropy? An axiomatic approach for linear inverse problems. *The Annals of Statistics*, 19(4):2032–2066, 1991.

[FS97] Yoav Freund and Robert E. Schapire. A decision-theoretic generalization of on-line learning and an application to boosting. *Journal of Computer and System Sciences*, 55(1):119–139, August 1997.

[GLS97] A. J. Grove, N. Littlestone, and D. Schuurmans. General convergence results for linear discriminant updates. In *Proc. 8th Annu. Conf. on Comput. Learning Theory*, pages 171–183. ACM, 1997.

[HKW95] D. P. Helmbold, J. Kivinen, and M. K. Warmuth. Worst-case loss bounds for sigmoided linear neurons. In *Proc. 1995 Neural Information Processing Conference*, pages 309–315. MIT Press, Cambridge, MA, November 1995.

[HKW97] D. Haussler, J. Kivinen, and M. K. Warmuth. Tight worst-case loss bounds for predicting with expert advice. IEEE Transactions on Information Theory, 1997. To appear.

[HLS96] D.P. Helmbold, D.D.E. Long, and B. Sherrod. A dynamic disk spin-down technique for mobile computing. In *Proceedings of the Second Annual ACM International Conference on Mobile Computing and Networking*. ACM/IEEE, November 1996.

[HSSW96] D. Helmbold, R. E. Schapire, Y. Singer, and M. K. Warmuth. On-line portfolio selection using multiplicative updates. In *Proc. 13th International Conference on Machine Learning*, pages 243–251. Morgan Kaufmann, San Francisco, July 1996.

[HW98] M. Herbster and M. K. Warmuth. Tracking the best expert. *Journal of Machine Learning*, 1998. Special issue on concept drift.

[JB90] L. Jones and C. Byrne. General entropy criteria for inverse problems, with applications to data compression, pattern classification and cluster analysis. *IEEE Transactions on Information Theory*, 36(1):23–30, 1990.

[JW98] A. Jagota and M. K. Warmuth. Continuous and discrete time nonlinear gradient descent: relative loss bounds and convergence. In R. Greiner E. Boros, editor, *Electronic Proceedings of Fifth International Symposium on Artificial Intelligence and Mathematics*, pages – Electronic,http://rutcor.rutgers.edu/~amai, 1998.

[KW94] J. Kivinen and M. Warmuth. Using experts for predicting continuous outcomes. In *Computational Learning Theory: Eurocolt '93*, volume New Series Number 53 of *The Institute of Mathematics and its Applications Conference Series*, pages 109–120, Oxford, 1994. Oxford University Press.

[KW97a] J. Kivinen and M. K. Warmuth. Additive versus exponentiated gradient updates for linear prediction. *Information and Computation*, 132(1):1–64, January 1997.

[KW97b] J. Kivinen and M. K. Warmuth. Relative loss bounds for multidimensional regression problems. In *Advances in Neural Information Processing Systems, 11*, Cambridge, MA, 1997. MIT Press. To appear.

[Lit88] N. Littlestone. Learning when irrelevant attributes abound: A new linear-threshold algorithm. *Machine Learning*, 2:285–318, 1988.

[Lit89] N. Littlestone. *Mistake Bounds and Logarithmic Linear-threshold Learning Algorithms*. PhD thesis, Technical Report UCSC-CRL-89-11, University of California Santa Cruz, 1989.

[LW94] N. Littlestone and M. K. Warmuth. The weighted majority algorithm. *Information and Computation*, 108(2):212–261, 1994.

[Roc70] R. Rockafellar. *Convex Analysis*. Princeton University Press, 1970.

[Rud91] W. Rudin. *Functional Analysis*. McGraw-Hill, 1991.

[Vov90] V. Vovk. Aggregating strategies. In *Proc. 3rd Annu. Workshop on Comput. Learning Theory*, pages 371–383. Morgan Kaufmann, 1990.

[Vov95] V. G. Vovk. A game of prediction with expert advice. In *Proc. 8th Annu. Conf. on Comput. Learning Theory*, pages 51–60. ACM Press, New York, NY, 1995.

[Vov97] V. Vovk. Derandomizing stochastic prediction strategies. In *Proc. 10th Annu. Workshop on Comput. Learning Theory*. ACM Press, New York, NY, 1997.

APPENDIX: DIVERGENCES AND THE GENERALIZED PYTHAGOREAN THEOREM

We begin by giving a more detailed definition of Bregman divergences. Here *ri* and *bd* denote the relative interior and relative boundary of a set [Roc70].

Definition A.1. *Let F be a function from a closed convex set $E \subset \Re^n$ into $\Re$ such that the following three conditions hold:*

(C1) *F is strictly convex.*

(C2) *F is differentiable.*

(C3) $\forall \boldsymbol{r} \in bd\, E, \forall \boldsymbol{s}, \boldsymbol{t} \in ri\, E : (\nabla F(\boldsymbol{r}) - \nabla F(\boldsymbol{s}), \boldsymbol{t} - \boldsymbol{r}) < 0.$

Then the Bregman divergence $D_F : E \times ri\, E \to [0, \infty)$ is defined as

$$D_F(\boldsymbol{u}, \boldsymbol{w}) = F(\boldsymbol{u}) - F(\boldsymbol{w}) + (\boldsymbol{w} - \boldsymbol{u}, \nabla F(\boldsymbol{w})). \quad (27)$$

The conditions that F is strictly convex and F is differentiable guarantee that $D_F(\boldsymbol{u}, \boldsymbol{w}) \geq 0$ and $\boldsymbol{u} = \boldsymbol{w} \Leftrightarrow D_F(\boldsymbol{u}, \boldsymbol{w}) = 0$.

The technical Condition (C3) guarantees that any projection (Definition 1.1) w.r.t. to the divergence will not lie on the boundary of E. Note that the inner product in Condition (C3) is the directional derivative of $D_F(\cdot, \boldsymbol{s})$ at the point $\boldsymbol{r}$ in the direction $\boldsymbol{t} - \boldsymbol{r}$. The direction $\boldsymbol{t} - \boldsymbol{r}$ is the direction from $\boldsymbol{r}$ (any point in bd E) to $\boldsymbol{t}$ (any point in ri E).

The divergence $D_F(\boldsymbol{u}, \boldsymbol{w})$ is strictly convex in $\boldsymbol{u}$ since it is the sum of a strictly convex function and a linear function. Similarly, $D_F(\boldsymbol{u}, \boldsymbol{w})$ is continuous in $\boldsymbol{u}$. We define the open and closed balls relative to a point and a divergence as $B_\epsilon(\boldsymbol{w}) = \{\boldsymbol{v} : D_F(\boldsymbol{v}, \boldsymbol{w}) < \epsilon\}$ and $\overline{B}_\epsilon(\boldsymbol{w}) = \{\boldsymbol{v} : D_F(\boldsymbol{v}, \boldsymbol{w}) \leq \epsilon\}$. Since $D_F(\boldsymbol{u}, \boldsymbol{w})$ is continuous in $\boldsymbol{u}$, the closure $B_\epsilon(\boldsymbol{w})$ is equal to $\overline{B}_\epsilon(\boldsymbol{w})$. It is straightforward to check that these balls are strictly convex sets since $D_F(., \boldsymbol{w})$ is strictly convex. In the full paper we will also show the following:

Proposition A.2. *For any point $\boldsymbol{w} \in ri\, E$, $\overline{B}_\epsilon(\boldsymbol{w})$ is bounded.*

The uniqueness of the projection essentially follows from the boundedness and strict convexity of the balls, and the convexity of the constraint set.

Proposition A.3. *Given $\boldsymbol{w} \in ri\, E$ and a closed convex set Γ such that $\Gamma \bigcap ri\, E \neq \emptyset$, then $P_{(\Gamma, D_F)}(\boldsymbol{w})$ exists and is unique.*

Proof. Without loss of generality, assume $\Gamma \subseteq E$. We first show existence. Let $A = D_F(\Gamma, \boldsymbol{w})$. Since all points in A are lower bounded by 0, $a = \inf A$ exists. We now prove a lies in A, which proves the existence of a projection. Let $\delta > 0$. The ball $\overline{B}_{a+\delta}(\boldsymbol{w})$ is closed and bounded. Therefore the intersection $B = \overline{B}_{a+\delta}(\boldsymbol{w}) \bigcap \Gamma$ is closed and bounded. Since we are in $\Re^n$, this implies that B is compact. Let $C = D_F(B, \boldsymbol{w})$. Since B is compact and $D_F(\cdot, \boldsymbol{w})$ is continuous, it follows that C is compact. Since C is compact, $\inf C$ must lie in C. It follows that a projection exists and all projections lie in $D = \{\boldsymbol{v} : D_F(\boldsymbol{v}, \boldsymbol{w}) = \inf C\} \bigcap \Gamma$. We now show that the set D has a single element, and hence the projection is unique. Let $\boldsymbol{s}, \boldsymbol{t}$ be distinct elements of D and let $\boldsymbol{u} = \frac{\boldsymbol{s}+\boldsymbol{t}}{2}$. Then $\boldsymbol{u} \in \Gamma$, since Γ is convex. Since $D_F(\cdot, \boldsymbol{w})$ is strictly convex, $D_F(\boldsymbol{u}, \boldsymbol{w}) < \frac{D_F(\boldsymbol{s}, \boldsymbol{w})}{2} + \frac{D_F(\boldsymbol{t}, \boldsymbol{w})}{2} = \inf C$, which is a contradiction. Hence $P_{(\Gamma, D_F)}(\boldsymbol{w})$ is unique. □

We prove a sequence of three propositions. The first and third comprise Theorem 1.2, the generalization of the Pythagorean Theorem. The following proposition treats the special case of this theorem when Γ is a hyperplane (second part of Theorem 1.2).

Proposition A.4. *Given a divergence $D_F : E \times ri\, E \to [0, \infty)$, a hyperplane $H = \{\boldsymbol{v} : (\boldsymbol{x}, \boldsymbol{v}) = y\}$ such that $H \bigcap ri\, E \neq \emptyset$, and points $\boldsymbol{w} \in ri\, E$ and $\boldsymbol{u} \in H$, then*

$$D_F(\boldsymbol{u}, \boldsymbol{w}) = D_F(\boldsymbol{u}, P_{(H, D_F)}(\boldsymbol{w})) + D_F(P_{(H, D_F)}(\boldsymbol{w}), \boldsymbol{w}). \quad (28)$$

Proof. If $\boldsymbol{w} \in H$, then $P_{(H, D_F)}(\boldsymbol{w}) = \boldsymbol{w}$ and the proposition trivially holds. Assume $\boldsymbol{w} \notin H$. By expanding the D_F (cf. Definition A.1) we get the following equivalent form of Equation (28):

$$(\boldsymbol{u} - P_{(H, D_F)}(\boldsymbol{w}), \nabla F(\boldsymbol{w}) - \nabla F(P_{(H, D_F)}(\boldsymbol{w}))) = 0. \quad (29)$$

We now proceed to prove an implicit form of the projection. Computing the projection is a convex programming problem, i.e. the computation of $\text{argmin}_{\boldsymbol{v}} D_F(\boldsymbol{v}, \boldsymbol{w})$ subject to the constraint that $\boldsymbol{v}$ lies in the convex set $E \bigcap H$. We may ignore the constraint $\boldsymbol{v} \in E$, since technical Condition (C3) ensures that $\boldsymbol{v}$ is in ri E. We introduce a Lagrange multiplier λ for the equality constraint that $\boldsymbol{v} \in H$:

$$L(\lambda, \boldsymbol{v}) = D_F(\boldsymbol{v}, \boldsymbol{w}) + \lambda((\boldsymbol{x}, \boldsymbol{v}) - y). \quad (30)$$

Thus for the projection the following equation holds:

$$\begin{aligned} 0 &= \nabla_{\boldsymbol{v}} L(\lambda, \boldsymbol{v}) \mid_{\boldsymbol{v} = P_{(H, D_F)}(\boldsymbol{w})} \\ &= \nabla F(P_{(H, D_F)}(\boldsymbol{w})) - \nabla F(\boldsymbol{w}) + \lambda \boldsymbol{x}. \end{aligned}$$

Rewriting the above equation, we have

$$\nabla F(\boldsymbol{w}) - \nabla F(P_{(H, D_F)}(\boldsymbol{w})) = \lambda \boldsymbol{x}, \quad (31)$$

and the left-hand-side of (29) becomes $\lambda(\boldsymbol{u} - P_{(H, D_F)}(\boldsymbol{w}), \boldsymbol{x})$. Since $(\boldsymbol{x}, P_{(H, D_F)}(\boldsymbol{w})) = y$ and $(\boldsymbol{x}, \boldsymbol{u}) = y$, this inner product is zero. □

We next show the generalized Pythagorean Theorem for projections onto halfspaces.

Proposition A.5. *Given a divergence $D_F : E \times ri\, E \to [0, \infty)$, a closed halfspace $\mathcal{H} = \{\boldsymbol{v} : (\boldsymbol{x}, \boldsymbol{v}) \leq y\}$ such that $\mathcal{H} \bigcap ri\, E \neq \emptyset$, and points $\boldsymbol{w} \in ri\, E$ and $\boldsymbol{u} \in \mathcal{H}$, then*

$$D_F(\boldsymbol{u}, \boldsymbol{w}) \geq D_F(\boldsymbol{u}, P_{(\mathcal{H}, D_F)}(\boldsymbol{w})) + D_F(P_{(\mathcal{H}, D_F)}(\boldsymbol{w}), \boldsymbol{w}). \quad (32)$$

Proof. The case $\boldsymbol{w} \in \mathcal{H}$ is trivially true. Assume $\boldsymbol{w} \notin \mathcal{H}$. Equation (32) has the following equivalent form:

$$(\boldsymbol{u} - P_{(\mathcal{H}, D_F)}(\boldsymbol{w}), \nabla F(\boldsymbol{w}) - \nabla F(P_{(\mathcal{H}, D_F)}(\boldsymbol{w}))) \leq 0. \quad (33)$$

The projection of $\boldsymbol{w}$ clearly lies on the boundary of the halfspace, i.e. the hyperplane $H = \{\boldsymbol{v} : (\boldsymbol{x}, \boldsymbol{v}) = y\}$. Thus $P_{(\mathcal{H}, D_F)}(\boldsymbol{w}) = P_{(H, D_F)}(\boldsymbol{w})$. Finding the projection is a convex programming problem subject to an inequality constraint. Thus the parameter λ in (30) is now a Kuhn-Tucker coefficient and Equation (31) again holds. This lets us rewrite the left-hand-side of (33) as $\lambda(\boldsymbol{u} - P_{(H, D_F)}(\boldsymbol{w}), \boldsymbol{x})$. We complete the proof by showing that this expression is at most zero. First note that the projection lies on the boundary H of $\mathcal{H}$ and thus the Kuhn-Tucker coefficient is positive. Finally, $(\boldsymbol{x}, P_{(H, D_F)}(\boldsymbol{w})) = y$ and $(\boldsymbol{x}, \boldsymbol{u}) \leq y$, so we have that

$$\lambda(\boldsymbol{u} - P_{(H, D_F)}(\boldsymbol{w}), \boldsymbol{x}) \leq 0.$$

□

We now prove the first part of the generalized Pythagorean Theorem:

Proposition A.6. *Given a divergence $D_F : E \times ri\ E \to [0, \infty)$, a closed convex set Γ such that $\Gamma \bigcap ri\ E \neq \emptyset$, and points $\boldsymbol{w} \in ri\ E$ and $\boldsymbol{u} \in \Gamma$, then*

$$D_F(\boldsymbol{u}, \boldsymbol{w}) \geq D_F(\boldsymbol{u}, P_{(\Gamma, D_F)}(\boldsymbol{w})) + D_F(P_{(\Gamma, D_F)}(\boldsymbol{w}), \boldsymbol{w}). \tag{34}$$

Proof. The case $\boldsymbol{w} \in \Gamma$ is again trivially true. Assume $\boldsymbol{w} \notin \Gamma$. Let $\epsilon = D_F(P_{(\Gamma, D_F)}(\boldsymbol{w}), \boldsymbol{w})$. Consider the balls $B_\epsilon(\boldsymbol{w})$ and $\overline{B}_\epsilon(\boldsymbol{w})$. Note that $B_\epsilon(\boldsymbol{w}) \bigcap \Gamma = \emptyset$ and $\overline{B}_\epsilon(\boldsymbol{w}) \bigcap \Gamma = P_{(\Gamma, D_F)}(\boldsymbol{w})$. Let $\mathcal{H}$ be a halfspace that contains Γ but does not intersect with $B_\epsilon(\boldsymbol{w})$. The existence of such a halfspace is implied by the Hahn-Banach separation Theorem ([Rud91, p59]). (This theorem says that for any disjoint open and closed sets there exists a halfspace containing the closed set such that this halfspace is disjoint with the open set.)

We now claim that $P_{(\Gamma, D_F)}(\boldsymbol{w}) = P_{(\mathcal{H}, D_F)}(\boldsymbol{w})$. Provided this claim is true, the current proposition clearly follows from the previous one. Since $\mathcal{H} \supseteq \Gamma$,

$$D_F(P_{(\mathcal{H}, D_F)}(\boldsymbol{w}), \boldsymbol{w}) \leq D_F(P_{(\Gamma, D_F)}(\boldsymbol{w}), \boldsymbol{w}) = \epsilon.$$

On the other hand $<$ is impossible because $B_\epsilon(\boldsymbol{w})$ and $\mathcal{H}$ are disjoint. Thus the above inequality must be an equality and claim follows from the uniqueness of projections. □