Mirrorless Mirror Descent: A Natural Derivation of Mirror Descent

Suriya Gunasekar
Microsoft Research

Blake Woodworth
TTIC

Nathan Srebro
TTIC

Abstract

We present a primal only derivation of Mirror Descent as a "partial" discretization of gradient flow on a Riemannian manifold where the metric tensor is the Hessian of the Mirror Descent potential. We contrast this discretization to Natural Gradient Descent, which is obtained by a "full" forward Euler discretization. This view helps shed light on the relationship between the methods and allows generalizing Mirror Descent to general Riemannian geometries, even when the metric tensor is *not* a Hessian, and thus there is no "dual."

1 Introduction

Mirror Descent (Nemirovsky and Yudin, 1983; Beck and Teboulle, 2003) is an important template first-order optimization method for optimizing w.r.t. a geometry specified by a strongly convex potential function. It enjoys rigorous guarantees, and its stochastic and online variants are even optimal for certain learning settings (Srebro et al., 2011). As its name implies, Mirror Descent was derived, and is typically described, in terms of performing gradient steps in the *dual space* using a *mirror map*: in each iteration, one maps the iterate to the dual space through a link function, performs an update there, and then mirrors the updates back to the primal space. Understanding Mirror Descent in this way requires explicitly discussing the dual space or the link function.

In this paper we derive a direct "primal" understanding of Mirror Descent, and in order to do so, turn to Riemannian Gradient Flow. The infinitesimal limit of Mirror Descent, where the stepsize is taken to zero, corresponds to a Riemannian Gradient Flow on a manifold with a metric tensor that is given by the Hessian of the potential function used by Mirror Descent (see Section 2.2). The standard forward Euler discretization of this Riemannian Gradient Flow gives rise to the Natural Gradient Descent algorithm Amari (1998). Our main observation is that a "partial" discretization of the flow, where we discretize the optimization objective but not the metric tensor specifying the geometry, gives rise precisely to Mirror Descent (see Section 2.2). This view allows us to understand how Mirror Descent is, in a sense, more "faithful" to the geometry compared to Natural Gradient Descent.

The relationship we reveal between Mirror Descent and Natural Gradient Descent is different from, and complementary to, the relationship discussed by Raskutti and Mukherjee (2015)—while their work showed how Mirror Descent and Natural Gradient Descent are *dual* to each other, in the sense that Mirror Descent is equivalent to Natural Gradient Descent in the dual space, we avoid the duality altogether. We work *only* in the primal space, derive Mirror Descent directly, without considering the dual or link functions, and show how both methods are *different discretizations* of the *same* flow (more details in Section 2.5).

As a consequence, our derivation of Mirror Descent allows us to conceptually generalize Mirror Descent to any Riemannian manifold, including situations where metric tensor is *not* specified by the Hessian of any potential, and so there is no dual, no link function and no Bregman divergence.

1.1 Background: Mirror Descent

Consider optimizing a smooth objective $F : \mathcal{W} \to \mathbb{R}$ over a closed convex set $\mathcal{W} \subseteq \mathbb{R}^d$, $\min_{\mathbf{w}\in\mathcal{W}} F(\mathbf{w})$. We will focus on unconstrained optimization, *i.e.,* $\mathcal{W} = \mathbb{R}^d$.

Mirror Descent is a template first-order optimization algorithm specified by a strictly convex potential function $\psi : \mathcal{W} \to \mathbb{R}$. Mirror Descent was developed as a generalization of gradient descent to non-Euclidean geometries, where the local geometry is specified by the Bregman divergences w.r.t. ψ given by $D_\psi(\mathbf{w}, \mathbf{w}') = \psi(\mathbf{w}) - \psi(\mathbf{w}') - \langle \nabla\psi(\mathbf{w}'), \mathbf{w} - \mathbf{w}' \rangle$. The

Proceedings of the 24th International Conference on Artificial Intelligence and Statistics (AISTATS) 2021, San Diego, California, USA. PMLR: Volume 130.

iterative updates of Mirror Descent with stepsize η are defined as:

$$\mathbf{w}_{\text{MD}}^{(k+1)} = \operatorname*{argmin}_{\mathbf{w}\in\mathcal{W}} \eta\langle \mathbf{w}, \nabla F(\mathbf{w}_{\text{MD}}^{(k)})\rangle + D_\psi(\mathbf{w}, \mathbf{w}_{\text{MD}}^{(k)}). \quad (1)$$

For unconstrained optimization, the updates in (1) are equivalently given by:

$$\nabla\psi(\mathbf{w}_{\text{MD}}^{(k+1)}) = \nabla\psi(\mathbf{w}_{\text{MD}}^{(k)}) - \eta\nabla F(\mathbf{w}_{\text{MD}}^{(k)}), \quad (2)$$

where $\nabla\psi$ is called the *link function* and provides a mapping between the primal optimization space $\mathbf{w} \in \mathcal{W}$ and the dual space of gradients. Mirror Descent thus performs gradient updates in the dual "mirror".

1.2 Background: Riemannian Gradient Flow

Let $(\mathcal{W}, H)$ denote a Riemannian manifold over $\mathcal{W} = \mathbb{R}^d$ equipped with a metric tensor $H(\mathbf{w})$ at each point $\mathbf{w} \in \mathcal{W}$. The metric tensor $H(\mathbf{w}) : T_{\mathcal{W}}(\mathbf{w}) \times T_{\mathcal{W}}(\mathbf{w}) \to \mathbb{R}$ denotes a smoothly varying *local* inner product on the tangent space at $\mathbf{w}$. Intuitively, the tangent space $T_{\mathcal{W}}(\mathbf{w})$ is the vector space of all infinitesimal directions $\mathrm{d}\mathbf{w}$ that we can move in while following a smooth path on $\mathcal{W}$ (for more detailed exposition see, e.g. Do Carmo, 2016). For manifolds over $\mathbb{R}^d$, we can take the tangent space as $T_{\mathcal{W}}(\mathbf{w}) = \mathbb{R}^d$ and the metric tensors can be identified with positive definite matrices $H(\mathbf{w}) \in \mathcal{S}_{++}^d$ that define local distances at $\mathbf{w}$ as $d(\mathbf{w}, \mathbf{w} + \mathrm{d}\mathbf{w}) = \sqrt{\mathrm{d}\mathbf{w}^\top H(\mathbf{w})\mathrm{d}\mathbf{w}}$ for infinitesimal $\mathrm{d}\mathbf{w}$.

The Riemannian Gradient Flow dynamics $\mathbf{w}(t)$ for the optimization problem $\min_{\mathbf{w}\in\mathcal{W}} F(\mathbf{w})$ with initialization $\mathbf{w}(0) = \mathbf{w}_{\text{init}}$ are obtained by seeking an infinitesimal change in $\mathbf{w}(t)$ that would lead to the best improvement in objective value, while controlling the length of the change in terms of the manifold geometry, that is,

$$\mathbf{w}(t + \mathrm{d}t) = \operatorname*{argmin}_{\mathbf{w}} \; F(\mathbf{w})\mathrm{d}t + \frac{1}{2}d(\mathbf{w}, \mathbf{w}(t))^2. \quad (3)$$

For infinitesimal $\mathrm{d}t$, using $\mathrm{d}\mathbf{w}(t) = \mathbf{w}(t + \mathrm{d}t) - \mathbf{w}(t)$, we can replace $F(\mathbf{w})$ and $d(\mathbf{w}, \mathbf{w}(t))$ with their first order approximations[1] $F(\mathbf{w}(t)) + \langle \mathrm{d}\mathbf{w}, \nabla F(\mathbf{w}(t))\rangle$, and $d(\mathbf{w}, \mathbf{w}(t)) = \sqrt{\mathrm{d}\mathbf{w}^\top H(\mathbf{w}(t))\mathrm{d}\mathbf{w}}$:

$$\mathrm{d}\mathbf{w}(t) = \operatorname*{argmin}_{\mathrm{d}\mathbf{w}} \langle \mathrm{d}\mathbf{w}, \nabla F(\mathbf{w}(t))\rangle \mathrm{d}t + \frac{1}{2}\mathrm{d}\mathbf{w}^\top H(\mathbf{w}(t))\mathrm{d}\mathbf{w}. \quad (4)$$

Solving for $\mathrm{d}\mathbf{w}$, we obtain:

$$\dot{\mathbf{w}}(t) = -H(\mathbf{w}(t))^{-1}\nabla F(\mathbf{w}(t)), \quad \text{(GF)}$$

where here and throughout we denote $\dot{\mathbf{w}} = \frac{\mathrm{d}\mathbf{w}}{\mathrm{d}t}$.

[1] We use $\langle ., . \rangle$ to denote the canonical inner product in $\mathbb{R}^d$ and ∇ denotes the gradient operator such that $\langle \nabla F(\mathbf{w}), \mathrm{d}\mathbf{w}\rangle = F(\mathbf{w} + \mathrm{d}\mathbf{w}) - F(\mathbf{w})$ for all infinitesimal $\mathrm{d}\mathbf{w}$

We refer to the path specified by (GF) and initial condition $\mathbf{w}(0) = \mathbf{w}_{\text{init}}$ as *Riemannian Gradient Flow* or sometimes simply as *gradient flow*.

Examples of Riemannian metrics and corresponding gradient flow that arise in learning and related areas:

1. The standard Euclidean geometry is recovered with $H(\mathbf{w}) = I$. In this case (GF) reduces to the standard gradient flow $\dot{\mathbf{w}} = -\nabla F(\mathbf{w})$. When $H(\mathbf{w}) = H$ is fixed to some other positive definite H, we get the pre-conditioned gradient flow $\dot{\mathbf{w}} = -H^{-1}\nabla F(\mathbf{w})$, which can also be thought of as the gradient flow dynamics on a reparametrization $\widetilde{\mathbf{w}} = H^{1/2}\mathbf{w}$, i.e. with respect to geometry specified by a linear distortion.
2. For any strongly convex potential function ψ over $\mathcal{W}$, the Hessian $\nabla^2\psi$ defines a non-Euclidean metric tensor. Examples include squared ℓ_p norms $\psi(\mathbf{w}) = \|\mathbf{w}\|_p^2$ for $1 < p \le 2$ ($p = 2$ again recovers the standard Euclidean geometry) and a particularly important example is the simplex, endowed with an entropy potential $\psi(\mathbf{w}) = -\sum_i \mathbf{w}_i \log \mathbf{w}_i$.
3. Information geometry (Amari, 2012) is concerned with a manifold of probability distributions, e.g. in a parametric family $\{p(x;\theta) : \theta \in \Theta \subseteq \mathbb{R}^d\}$, typically endowed with the metric derived from the KL-divergence. In our notation, we would consider this as defining a Riemannian metric structure over the manifold parameters $\mathcal{W} = \Theta$, with a metric tensor given by the Fisher information matrix $H(\theta) = \mathcal{I}(\theta) = \mathbb{E}_x[-\nabla_\theta \log(p(x;\theta))\nabla_\theta \log(p(x;\theta))^\top \mid \theta]$. Such a geometry can also be obtained by considering the entropy as a potential function and taking its Hessian.

2 Discretizing Riemannian Gradient Flow

The Riemannian Gradient Flow is a continuous object defined in terms of a differential equation (GF). To utilize it algorithmically, we consider discretizations of the flow.

2.1 Natural Gradient Descent

Natural Gradient Descent is obtained as the forward Euler discretization with stepsize η of the gradient flow (GF):

$$\mathbf{w}_{\text{NGD}}^{(k+1)} = \mathbf{w}_{\text{NGD}}^{(k)} - \eta H(\mathbf{w}_{\text{NGD}}^{(k)})^{-1}\nabla F(\mathbf{w}_{\text{NGD}}^{(k)}), \quad \text{where } \mathbf{w}_{\text{NGD}}^{(0)} = \mathbf{w}_{\text{init}}. \quad (5)$$

These Natural Gradient Descent updates were suggested and popularized by Amari (1998), particularly

in the context of *information geometry*, where the metric tensor is given by the Fisher information matrix of some family of distributions.

An equivalent way to view the updates (5) is by discretizing the right-hand-side of the differential equation (GF) as follows:

$$\dot{\mathbf{w}}(t) = -H\left(\mathbf{w}(\lfloor t \rfloor_\eta)\right)^{-1} \nabla F\left(\mathbf{w}(\lfloor t \rfloor_\eta)\right), \quad \text{(NGD)}$$

where $\lfloor t \rfloor_\eta := \eta \lfloor t/\eta \rfloor$ denotes a discretization at scale η, i.e. the largest $t' < t$ such that t' is an integer multiple of η.

The differential equation (NGD) specifies a piecewise linear solution $\mathbf{w}(t)$ that interpolates between the Natural Gradient Descent iterates. In particular, the Natural Gradient Descent iterates in (5) are given by $\mathbf{w}^{(k)}_{\text{NGD}} = \mathbf{w}(\eta k)$ where $\mathbf{w}(t)$ is the solution of (NGD) with the initial condition $\mathbf{w}(0) = \mathbf{w}_{\text{init}}$, and we could have alternatively defined the forward Euler discretization in this way.

2.2 Mirror Descent

In (NGD) we fully discretized the Riemannian Gradient Flow (GF). Now consider an alternate, partial, forward discretization of (GF), where we discretize the gradient $\nabla F(\mathbf{w})$, but not the local metric $H(\mathbf{w})$:

$$\dot{\mathbf{w}}(t) = -H(\mathbf{w}(t))^{-1} \nabla F(\mathbf{w}(\lfloor t \rfloor_\eta)). \quad \text{(MD)}$$

The resulting solution $\mathbf{w}(t)$ is piecewise smooth. We will again consider the sequence of iterates at discrete points $t = \eta k$:

$$\mathbf{w}^{(k)} := \mathbf{w}(\eta k). \quad (6)$$

Our main result is that if $H(\mathbf{w}) = \nabla^2 \psi(\mathbf{w})$, then the updates (6) from the solution of (MD) are precisely the Mirror Descent updates in (2) with potential ψ.

Theorem 1. *Let $\psi : \mathcal{W} \to \mathbb{R}$ be strictly convex and twice differentiable and let $H(\mathbf{w}) = \nabla^2 \psi(\mathbf{w})$ be invertible everywhere. Consider the updates $\mathbf{w}^{(k)} = \mathbf{w}(k\eta)$ obtained from the solution of (MD) with initial condition $\mathbf{w}(0)$ and stepsize η. Then $\mathbf{w}^{(k)}$ are the same as the Mirror Descent updates $\mathbf{w}^{(k)}_{MD}$ in (2) obtained with potential ψ and the same initialization and stepsize.*

Proof. Consider the Mirror Descent iterates with step size η for link function $\nabla\psi$ from (2)

$$\nabla\psi(\mathbf{w}^{(k+1)}_{\text{MD}}) = \nabla\psi(\mathbf{w}^{(k)}_{\text{MD}}) - \eta \nabla F(\mathbf{w}^{(k)}_{\text{MD}}).$$

Define a Mirror Descent path $\hat{\mathbf{w}}(t)$ by linearly interpolating in the dual space as follows:

$$\forall k, \forall t \in [k\eta, (k+1)\eta) :$$
$$\nabla\psi(\hat{\mathbf{w}}(t)) = \nabla\psi(\mathbf{w}^{(k)}_{\text{MD}}) - (t - k\eta)\nabla F(\mathbf{w}^{(k)}_{\text{MD}}).$$

One can easily check that $\mathbf{w}^{(k)}_{\text{MD}} = \hat{\mathbf{w}}(\eta k)$. The above equation describing a piecewise smooth path $\hat{\mathbf{w}}(t)$ equivalently corresponds to, $\frac{\mathrm{d}\nabla\psi(\hat{\mathbf{w}}(t))}{\mathrm{d}t} = -\nabla F\big(\hat{\mathbf{w}}(\lfloor t \rfloor_\eta)\big)$. Using the chain rule, we see that $\hat{\mathbf{w}}(t)$ follows the discretization path in (MD):

$$\begin{aligned}\dot{\hat{\mathbf{w}}}(t) &= -\nabla^2\psi(\hat{\mathbf{w}}(t))^{-1}\nabla F\big(\hat{\mathbf{w}}(\lfloor t \rfloor_\eta)\big) \\ &= -H(\hat{\mathbf{w}}(t))^{-1}\nabla F\big(\hat{\mathbf{w}}(\lfloor t \rfloor_\eta)\big).\end{aligned}$$

This completes the proof of the theorem. □

2.3 A More Faithful Discretization?

Comparing the two discretizations (NGD) and (MD) allows us to understand the relationship between Natural Gradient updates (5) and Mirror Descent updates (2). Although both updates have the same infinitesimal limit, as previously discussed by, e.g. Gunasekar et al. (2018), they differ in how the discretization is done with finite stepsizes: while Natural Gradient Descent corresponds to discretizing both the objective *and* the geometry, Mirror Descent involves discretizing only the objective (accessed via $\nabla F(\mathbf{w}(t))$), but not the geometry (specified by $H(\mathbf{w}(t))$). In this sense, Mirror Descent is a "more accurate" discretization, being more faithful to the geometry of the search space.

This view of Mirror Descent also allows us to contrast the computational aspects of implementing the two algorithms. While both the algorithms are first order methods, which only require gradient access to the objective (at discrete iterates, $\nabla F(\mathbf{w}^{(k)})$, Natural Gradient Descent can be implemented if we can compute the inverse Hessian of the metric tensor (*i.e.,* we need to either obtain and invert the metric tensor, or have direct access to its inverse). But at least for a traditional implementation of the Mirror Descent updates in (2), we need (a) the metric tensor $H(\mathbf{w})$ to be a Hessian map, i.e. the differential equation $H(\mathbf{w}) = \nabla^2\psi(\mathbf{w})$ should have a solution, and (b) we need to be able to efficiently calculate the link $\nabla\psi$ and inverse link $\nabla\psi^{-1}$ functions. More generally, one needs some way of solving the ordinary differential equation (MD) to implement Mirror Descent.

2.4 When Does a Potential Exist?

In Theorem 1 we established that for any smooth strictly convex potential $\psi(\mathbf{w})$ with everywhere invertible Hessian, we can define a metric tensor $H(\mathbf{w}) = \nabla^2\psi(\mathbf{w})$ so that Mirror Descent is obtained as a discretization of the Riemannian Gradient Flow (GF). One might ask whether this connection with Mirror Descent holds for *any* Riemannian Gradient Flow.

The Riemannian Gradient Flow (GF), and hence also

the discretization (MD), can be defined for any smooth, invertible Riemannian metric tensor $H : \mathbb{R}^d \to \mathcal{S}^d_{++}$. But the classic Mirror Descent updates (2) are only defined in terms of a potential function ψ. To relate Riemannian Gradient Flow w.r.t. some metric tensor $H(\mathbf{w})$ to Mirror Descent, we need to identify such a potential function, i.e. a function $\psi : \mathcal{W} \to \mathbb{R}$ s.t. $\nabla^2\psi = H$. That is, we need there to be a solution to the partial differential equation $\nabla^2\psi = H$, or in other words for H to be a Hessian map.

When is a metric tensor a Hessian map? Or, when is there a solution to $\nabla^2\psi = H$? By Poincaré's lemma, the rows of the metric tensor, $H_i : \mathbb{R}^d \to \mathbb{R}^d$ are gradients (*i.e.*, $\forall i, H_i = \nabla\phi_i$ for some $\phi_i : \mathbb{R}^d \to \mathbb{R}$) if and only if they satisfy the following symmetry condition:

$$\forall_{\mathbf{w}\in\mathcal{W}}\,\forall_{i,j,k\in\{1,\ldots,n\}} \quad \frac{\partial H_{i,j}(\mathbf{w})}{\partial \mathbf{w}_k} = \frac{\partial H_{i,k}(\mathbf{w})}{\partial \mathbf{w}_j}. \tag{7}$$

Thus, (7) is equivalent to H being a Jacobian of some vector-valued function $\phi = [\phi_1, \phi_2, \ldots, \phi_d] : \mathbb{R}^d \to \mathbb{R}^d$. Further, since H is symmetric (by definition), ϕ also satisfies the same symmetry condition $\frac{\partial \phi_i(\mathbf{w})}{\partial \mathbf{w}_j} = H_{i,j}(\mathbf{w}) = H_{j,i}(\mathbf{w}) = \frac{\partial \phi_j(\mathbf{w})}{\partial \mathbf{w}_i}$, and hence in turn is a gradient field (*i.e.*, $\phi = \nabla\psi$ for some ψ). Therefore, (7) is equivalent to $\nabla^2\psi = H$ having a solution, and since $H(\mathbf{w})$ is positive definite (in order to be a valid metric tensor), ψ must be strictly convex, as we would desire of a potential function. Hence, we can conclude that the discretization of Riemannian Gradient Flow in (MD) corresponds to classic Mirror Descent (2) for some strictly convex potential ψ if and only if (7) holds. The requirement in (7) is non-trivial and does not hold in general, for instance, a seemingly simple metric tensor $H(\mathbf{w}) = I + \mathbf{w}\mathbf{w}^\top$ fails to satisfy (7) and therefore is not a Hessian map[2].

[2]This metric tensor $H(\mathbf{w}) = I + \mathbf{w}\mathbf{w}^\top$ can arise by considering the d-dimensional manifold embedded in $\mathbb{R}^{d+1}$ which consists of points $\begin{bmatrix}\mathbf{w}\\ \frac{1}{2}\|\mathbf{w}\|^2\end{bmatrix}$ with local distances induced by the Euclidean geometry on $\mathbb{R}^{d+1}$. The distance between $\mathbf{w}$ and $\mathbf{w}+\mathrm{d}\mathbf{w}$ is then given by

$$\begin{aligned} d(\mathbf{w}, \mathbf{w}+\mathrm{d}\mathbf{w}) &= \left\| \begin{bmatrix}\mathbf{w}\\ \frac{1}{2}\|\mathbf{w}\|^2\end{bmatrix} - \begin{bmatrix}\mathbf{w}+\mathrm{d}\mathbf{w}\\ \frac{1}{2}\|\mathbf{w}+\mathrm{d}\mathbf{w}\|^2\end{bmatrix} \right\| \\ &= \sqrt{\mathrm{d}\mathbf{w}^\top(I+\mathbf{w}\mathbf{w}^\top)\mathrm{d}\mathbf{w} + \frac{1}{4}\|\mathrm{d}\mathbf{w}\|^4 - \langle \mathbf{w}, \mathrm{d}\mathbf{w}\rangle \|\mathrm{d}\mathbf{w}\|^2} \end{aligned}$$

For infinitesimal $\mathrm{d}\mathbf{w}$, this means $d(\mathbf{w}, \mathbf{w}+\mathrm{d}\mathbf{w}) = \sqrt{\mathrm{d}\mathbf{w}^\top(I+\mathbf{w}\mathbf{w}^\top)\mathrm{d}\mathbf{w}}$, and indeed the metric tensor is described by $H(\mathbf{w}) = I + \mathbf{w}\mathbf{w}^\top$. That this is not a Hessian map can be seen by simply calculating $\frac{\partial}{\partial \mathbf{w}_1} H_{1,2}(\mathbf{w}) = \frac{\partial}{\partial \mathbf{w}_1}\mathbf{w}_1\mathbf{w}_2 = \mathbf{w}_2 \neq \frac{\partial}{\partial \mathbf{w}_2} H_{1,1}(\mathbf{w}) = \frac{\partial}{\partial \mathbf{w}_2} 1 + \mathbf{w}_1^2 = 0$.

2.5 Contrast With Prior Derivations

We emphasize that our derivation of Mirror Descent as a partial discretization of Riemannian Gradient Flow is rather different from, and complementary to, a previous relationship pointed out between Natural Gradient Descent and Mirror Descent by Raskutti and Mukherjee (2015). Their derivation *does* rely on duality and existence of a potential, and thus a link function. Raskutti and Mukherjee showed that Mirror Descent over $\mathbf{w} \in \mathcal{W}$ corresponds to Natural Gradient Descent *in the dual*, that is after a change of parametrization given by the link function $\widetilde{\mathbf{w}} = \nabla\psi(\mathbf{w})$.

In terms of the discretized differential equations (NGD) and (MD), the above relationship can be stated as follows: the path of the Natural Gradient Descent discretization (NGD) is piecewise linear in the primal space, i.e. $\mathbf{w}(t)$ is piecewise linear, while the path of the Mirror Descent discretization (MD) is piecewise linear in the *dual space*, i.e. $\nabla\psi(\mathbf{w}(t))$ is piecewise linear, and consequently curved in the primal space. But this view does not explain why Mirror Descent might be preferable when we are interested in the primal geometry. In contrast, here we focus only on the (primal) Riemannian geometry, do not use a link function nor the dual, and highlight why Mirror Descent is more faithful to this primal geometry.

Raskutti and Mukherjee's dual view is also captured by another popular way of developing Mirror Descent as a discretization of a differential equation: when the metric tensor is a Hessian map and $H(\mathbf{w}) = \nabla^2\psi(\mathbf{w})$, then the partial differential equation (GF) is equivalent to the following[3] (Nemirovsky and Yudin, 1983; Warmuth and Jagota, 1997; Raginsky and Bouvrie, 2012):

$$\frac{\mathrm{d}}{\mathrm{d}t}\nabla\psi(\mathbf{w}(t)) = -\nabla F(\mathbf{w}(t)). \tag{8}$$

A forward Euler discretization of the differential equation (8) yields the Mirror Descent updates in (2). This can be viewed as using standard (full discretization) forward Euler, corresponding to piecewise linear updates and Natural Gradient Descent, but on the *dual* variables, i.e. discretizing $\frac{\mathrm{d}\widetilde{\mathbf{w}}}{\mathrm{d}t}$ where $\widetilde{\mathbf{w}} = \nabla\psi(\mathbf{w})$.

Viewing Mirror Descent as a forward Euler discretization of (8), or as dual to Natural Gradient Descent, as in previous derivations and discussions of Mirror Descent still depends on having a link function $\nabla\psi(\mathbf{w})$ such that the metric tensor is a Hessian map $H = \nabla^2\psi$. One might ask whether we could perform such derivations relying on a change-of-variables "link" function

[3]To see this, apply the chain rule to the left hand side of (8) to get $\nabla^2\psi(\mathbf{w}(t))\dot{\mathbf{w}}(t) = -\nabla F(\mathbf{w}(t))$ and then multiply both sides by $\nabla^2\psi(\mathbf{w}(t))^{-1} = H(\mathbf{w}(t))^{-1}$ to get (GF)

even if the metric tensor H is not a Hessian map. In other words, could we have a function $\phi : \mathbb{R}^d \to \mathbb{R}^d$ such that $\phi(\mathbf{w}(t))$ is piecewise linear under the Mirror Descent dynamics (MD), in which case we could derive Mirror Descent as Natural Gradient Descent on $\phi(\mathbf{w}(t))$ or as the forward Euler discretization

$$\frac{\mathrm{d}}{\mathrm{d}t}\phi(\mathbf{w}(t)) = -\nabla F(\mathbf{w}(\lfloor t \rfloor_\eta)). \tag{9}$$

That is, when does there exists $\phi : \mathbb{R}^d \to \mathbb{R}^d$ such that (9) is equivalent to (MD) for any smooth objective $F(\mathbf{w})$? Applying the chain rule to the left hand side of (9), this would require that $\nabla\phi = H$, which further implies that H is a Hessian map[4]. Therefore, if the metric tensor is *not* a Hessian map, there is no analogue to the link function, and we cannot obtain the discretization in (MD) as Natural Gradient Descent after some change of variables, nor as a forward Euler discretization of a differential equation similar to (8).

An distinguishing feature of our novel derivation of Mirror Descent that clearly differentiates it from all prior derivations, is that it does *not* require a potential, dual or link function, and so it does *not* rely on the metric tensor being a Hessian map. This is exemplified by the fact that, unlike any prior derivation, it allows us to conceptually generalize Mirror Descent to metric tensors that are *not* Hessian maps.

3 Potential-free Mirror Descent

As we emphasized, a significant difference between our derivation and previous, or "classical", derivations of Mirror Descent, is that our derivations did not involve, or even rely on the existence of potential function—that is, it did not rely on the metric tensor being a Hessian map. If the metric tensor is *not* a Hessian map, we cannot define the link function nor Bregman divergence, and the standard Mirror Descent updates (1)–(2) are not defined, nor are any prior derivations of Mirror Descent that we are aware of. Nevertheless our equivalent primal-only derivation of Mirror Descent (6) *does* allow us to generalize Mirror Descent as a first order optimization procedure to *any* metric tensor, even if it is not a Hessian map—we simply use (6) as the definition of Mirror Descent.

To be more precise, we can define $\mathbf{w}_{\mathrm{MD}}^{(k)}$ iteratively as follows: given $\mathbf{w}_{\mathrm{MD}}^{(k)}$ and the gradient $g^{(k)} = \nabla F(\mathbf{w}_{\mathrm{MD}}^{(k)})$, consider the path defined by the differential equation

$$\dot{\mathbf{w}}(t) = -H(\mathbf{w}(t))^{-1} g^{(k)} \text{ with } \mathbf{w}(0) = \mathbf{w}_{\mathrm{MD}}^{(k)}, \quad \text{and let } \mathbf{w}_{\mathrm{MD}}^{(k+1)} = \mathbf{w}(\eta). \tag{10}$$

For a general metric tensor H, the above updates requires computing the solution of a differential equation at each step, which may or may not be efficiently computable (just as the standard Mirror Descent updates may or may not be efficiently computable depending on the link function). Nevertheless, it is important to note that the differential equation (10) depends on the objective F only through a single gradient $g^{(k)} = \nabla F(\mathbf{w}_{\mathrm{MD}}^{(k)})$. That is, the only required access to the objective in order to implement the method is a single gradient access per iteration—the rest is just computation in terms of the pre-specified geometry (similar to computing the link and inverse link in standard Mirror Descent). The updates (10) thus define a valid first order optimization method, and independent of the tractability of solving the differential equation, could be of interest in studying optimization with first order oracle access under general geometries.

We also show in Appendix B that when the eigenvalues of $H(\mathbf{w})$ are bounded from above and below, and when the objective F is smooth and strongly convex with respect to L2, then the updates (10) guarantee linear convergence to a minimizer of F, even when the metric tensor is not a Hessian. While this result is limited to the L2 geometry on $\mathcal{W}$, it at least suggests that the algorithm can be expected to succeed without relying on H being a Hessian map.

4 Importance of the Parametrization

Our development in Section 2 relied not only on a Riemannian manifold $(\mathcal{W}, H)$, but on a specific parametrization (or "chart") for the manifold, or in our presentation, on identifying the manifold $\mathcal{W}$, and its tangent space, with $\mathbb{R}^d$. Let us consider now the effect of a change of parametrization (i.e. on using a different chart).

Consider a change of parameters $\widetilde{\mathbf{w}} = \phi(\mathbf{w})$ for some smooth invertible ϕ with invertible Jacobian $\nabla\phi$, that specifies an isometric Riemannian manifold $(\widetilde{\mathcal{W}}, \widetilde{H})$, i.e., such that $d_H(\mathbf{w}, \mathbf{w}+\mathrm{d}\mathbf{w}) = d_{\widetilde{H}}(\phi(\mathbf{w}), \phi(\mathbf{w}+\mathrm{d}\mathbf{w}))$ for all infinitesimal $\mathrm{d}\mathbf{w}$. The metric tensor $\widetilde{H}(\widetilde{\mathbf{w}})$ for the isometric manifold is given by

$$\widetilde{H}(\widetilde{\mathbf{w}}) = \nabla\phi^{-1}(\widetilde{\mathbf{w}})^\top H(\phi^{-1}(\widetilde{\mathbf{w}}))\nabla\phi^{-1}(\widetilde{\mathbf{w}}), \tag{11}$$

where recall that the Jacobian of the inverse is the inverse Jacobian, $\nabla\phi(\mathbf{w})^{-1} = \nabla\phi^{-1}(\widetilde{\mathbf{w}})$. This can

[4] Applying the chain rule on (9) and substituting (MD) we get $\nabla\phi(\mathbf{w}(t))H(\mathbf{w}(t))^{-1}\nabla F(\mathbf{w}(\lfloor t \rfloor_\eta)) = \nabla F(\mathbf{w}(\lfloor t \rfloor_\eta))$. If this holds for any objective $F(\mathbf{w})$, it must be that $\nabla\phi H^{-1} = I$. But $\nabla\phi = H$ indicates that H is a Jacobian map, which for symmetric H further implies (7) since both sides evaluate to $\partial_{j,k}\phi_i$.

also be thought of as using a different chart for the manifold (in our case, a global chart, since the manifold is isomorphic to $\mathbb{R}^d$).

In understanding methods operating on a manifold, it is important to separate what is intrinsic to the manifold and its geometry, and what aspects of the method are affected by the parametrization, especially since one might desire "intrinsic" methods that depend only on the manifold and its geometry, but not on the parametrization. We therefore ask how changing the parametrization affects our development. In particular, does the Mirror Descent discretization, and with it the Mirror Descent updates change with parameterization?

Consider minimizing $F(\mathbf{w})$, which after the reparametrization we denote as $\widetilde{F}(\widetilde{\mathbf{w}}) = F(\phi^{-1}(\widetilde{\mathbf{w}}))$. The Riemannian Gradient Flow on $\widetilde{\mathbf{w}}$ is

$$\dot{\widetilde{\mathbf{w}}}(t) = -\widetilde{H}(\widetilde{\mathbf{w}}(t))^{-1}\nabla\widetilde{F}(\widetilde{\mathbf{w}}(t)), \tag{12}$$

where note that $\nabla\widetilde{F}(\widetilde{\mathbf{w}}) = \nabla\phi^{-1}(\widetilde{\mathbf{w}})^\top \nabla F(\phi^{-1}(\widetilde{\mathbf{w}}))$.

Since our initial development of the Riemannian Gradient Flow in eq. (3) was independent of the parametrization, it should be the case that the solution $\widetilde{\mathbf{w}}(t)$ of (12), i.e. as gradient flow in $(\widetilde{\mathcal{W}}, \widetilde{H})$, is equivalent to gradient flow in $(\mathcal{W}, H)$, i.e. $\widetilde{\mathbf{w}}(t) = \phi(\mathbf{w}(t))$ where $\mathbf{w}(t)$ is the solution of (GF). It is however, insightful to verify this directly: to do so, let us take the solution $\mathbf{w}(t)$ of (GF), define $\widetilde{\mathbf{w}}(t) \doteq \phi(\mathbf{w}(t))$, and check whether it is in-fact a solution to (12). Starting from the left hand side of (12), we have:

$$\begin{aligned}\dot{\widetilde{\mathbf{w}}} &= \nabla\phi(\mathbf{w})\dot{\mathbf{w}} = -\nabla\phi(\mathbf{w})H(\mathbf{w})^{-1}\nabla F(\mathbf{w}) \\ &= -\widetilde{H}(\widetilde{\mathbf{w}})^{-1}\nabla\widetilde{F}(\widetilde{\mathbf{w}}),\end{aligned} \tag{13}$$

thus verifying that $\phi(\mathbf{w}(t))$ indeed satisfies (12).

Do the same arguments hold also for the Mirror Descent discretization (MD)? Taking the solution $\mathbf{w}(t)$ of (MD) and setting $\widetilde{\mathbf{w}} \doteq \phi(\mathbf{w})$, we can follow the same derivation as above, except now the metric tensor H and gradient ∇F are calculated at different points, $\mathbf{w}(t)$ and $\mathbf{w}(\lfloor t \rfloor_\eta)$, respectively.

$$\begin{aligned}\dot{\widetilde{\mathbf{w}}} &= -\nabla\phi(\mathbf{w})H(\mathbf{w})^{-1}\nabla F(\mathbf{w}(\lfloor t \rfloor_\eta)) \\ &= -\widetilde{H}(\widetilde{\mathbf{w}})^{-1}\Big(\nabla\phi(\mathbf{w}(\lfloor t \rfloor_\eta))^{-1}\nabla\phi(\mathbf{w})\Big)^\top \nabla\widetilde{F}(\widetilde{\mathbf{w}}(\lfloor t \rfloor_\eta)).\end{aligned} \tag{14}$$

We can see why the Mirror Descent discretization, and hence also the Mirror Descent iterates are *not* invariant to changes in parametrization: if $\nabla\phi(\mathbf{w})$ is fixed, i.e., the reparametrization is affine, we have $\nabla\phi(\mathbf{w}(\lfloor t \rfloor_\eta))^{-1}\nabla\phi(\mathbf{w}) = I$ and (14) shows that $\widetilde{\mathbf{w}} = \phi(\mathbf{w})$ satisfies the Mirror Descent discretized differential equation w.r.t. $\widetilde{H}(\widetilde{\mathbf{w}})$. But more generally, the discretization would be affected by the "alignment" of the Jacobians along the solution path. We note that, for essentially the same reason, NGD is not generally invariant to reparametrization either.

A related question is how a reparametrization affects whether the metric tensor is a Hessian map. Indeed, for a particular parametrization (i.e. chart), the existence of a potential function ψ such that $H = \nabla^2\psi$ depends on whether H satisfies (7), and it may well be the case that $H(\mathbf{w})$ is not a Hessian map but $\widetilde{H}(\widetilde{\mathbf{w}})$ is, or visa versa (in fact, in general if $\phi(\mathbf{w})$ is non-affine we cannot expect both $H(\mathbf{w})$ and $\widetilde{H}(\widetilde{\mathbf{w}})$ to be Hessian maps). Does every Riemannian manifold have a reparametrization (i.e. chart) where the metric tensor is a Hessian map, i.e. which corresponds to "classical" Mirror Descent? Amari and Armstrong (2014) showed that while all Riemannian manifolds isomorphic to $\mathbb{R}^2$ admit a parametrization for which the metric tensor is a Hessian map, this is not true in higher dimensions; even a manifold isomorphic to $\mathbb{R}^3$ might not admit any parametrization with a Hessian metric tensor.

We see then how our potential-free derivation of Section 3 can indeed be much more general than the traditional view of Mirror Descent which applies only when the metric tensor is a Hessian map and a potential function exists: for many Riemannian manifolds, there is no parametrization with a Hessian metric tensor, and so it is not possible to define Mirror Descent updates classically such that the Riemannian gradient flow is obtained as their limit. Yet, the approach of Section 3 always allows us to do so. Furthermore, even for manifolds for which there exists a parametrization where the metric tensor is the Hessian of some potential function, our approach allows considering discretizations in other isometric parametrization.

Finally, in light of our characterization of Mirror Descent, several readers have suspected that Mirror Descent might be equivalent to Riemannian Gradient Descent (see Absil et al., 2009) using steps that follow geodesics on the manifold (i.e. using the exponential map retraction). However, the Riemannian Gradient and geodesics are intrinsic, whereas we have shown that Mirror Descent is not.

5 Summary

In this paper we presented a "primal" derivation of Mirror Descent, based on a discretization of the Riemannian Gradient Flow, and showed how it can be useful for understanding, thinking about, and potentially analyzing Mirror Descent, Natural Gradient Descent, and Riemannian Gradient Flow. We also showed how this

view suggests an generalization of (Mirrorless) Mirror Descent to any Riemannian geometry. It is important to identify interesting and useful examples of metric tensors H that are not Hessian maps for which this Mirrorless Mirror Descent perspective can lead to new algorithms and analysis.

Acknowledgements We thank André Neves for a helpful discussion about Riemannian geometry and for pointing out Amari and Armstrong (2014). This work was supported by NSF-RI 1764032 and was done, in part, while SG and NS were visiting the Simons Institute for the Theory of Computing. BW is supported by a Google Research PhD fellowship.

Broader Impact

Mirror Descent and Natural Gradient Descent are important and popular optimization approaches, both theoretically and practically, and both play central roles in machine learning. Aside from a direct role as a method for minimizing a given optimization objective, Mirror Descent, and ideas derived from it, such as the role of the potential function, also play a central role in online learning (e.g. Shalev-Shwartz, 2012, for a survey), and throughout learning theory. Obtaining a better understanding of Mirror Descent, and its relationship with Natural Gradient Descent, has thus been an ongoing endeavour in the optimization and machine learning communities, with past work, e.g. that of Raskutti and Mukherjee (2015), being influential in guiding the community's thinking about these methods. There is also been much interest in the community lately in understanding and re-deriving optimization methods as discretizations of continuous solutions to differential equations (e.g. Wibisono et al., 2016). Obtaining a novel, and very different, derivation of Mirror Descent as such a discretization can thus be very impactful in guiding our thinking about it, and in devising novel insights and methods based on it.

Our novel view could be particularly impactful since unlike all prior derivations of Mirror Descent, our approach does *not* rely on a dual and is thus valid much more broadly, and allows generalizing Mirror Descent to many more settings (as discussed in Section 3).

Cross-Disciplinary Impact and Impact in Education Beyond the possible practical implications, our primal-only view also has pedagogical implications, as it can allows for an arguably more direct derivation of Mirror Descent that might be easier to understand intuitively, especially by an audience not familiar with duality, link functions and Bregman divergences. As such, it can open up understanding of this method to a wider audience. In fact, the derivation was initially derived in order to explain Mirror Descent to physicists, and several colleagues already adopted it in the classroom.

A Stochastic Discretization

In this bonus appendix, we briefly discuss how yet another discretization of Riemannian Gradient Flow captures Stochastic Mirror Descent (Nemirovsky and Yudin, 1983), and could be useful in studying optimization versus statistical issues in training.

We have so far discussed exact, or *batch* Mirror Descent, but a popular variant is *Stochastic* Mirror Descent, where at each iteration we update based on an unbiased estimator $g^{(k)}$ of the gradient $\nabla F(\mathbf{w}^{(k)})$, i.e. such that $\mathbb{E}g^{(k)} = \nabla F(\mathbf{w}^{(k)})$ as,

$$\hat{\mathbf{w}}_{\mathrm{MD}}^{(k+1)} = \operatorname*{argmin}_{\mathbf{w}\in\mathcal{W}} \eta\langle g^{(k)}, \mathbf{w}\rangle + D_\psi(\mathbf{w}, \hat{\mathbf{w}}_{\mathrm{MD}}^{(k)}). \quad (15)$$

Consider stochastic objective of the form:

$$F(\mathbf{w}) = \mathbb{E}_z f(\mathbf{w}, z). \quad (16)$$

We can derive Stochastic Mirror Descent from the following stochastic discretization of Riemannian Gradient Flow (GF):

$$\dot{\mathbf{w}}(t) = -H(\mathbf{w}(t))^{-1}\nabla f(\mathbf{w}(\lfloor t\rfloor_\eta), z_{\lfloor t\rfloor_\nu}) \quad (17)$$

where z_t are sampled i.i.d., and we used two different resolutions, η and ν, to control the discretization.

Setting $\nu = \eta$ and taking $\hat{\mathbf{w}}_{\mathrm{MD}}^{(k)} = \mathbf{w}(\eta k)$ we recover "single example" Stochastic Mirror Descent, i.e. where at each iteration we use a gradient estimator $g^{(k)} = \nabla f(\mathbf{w}^{(k)}, z)$ based on a single i.i.d. example. But varying ν relative to η also allows us to obtain other variants.

Taking $\nu < \eta$, e.g. $\eta = b\cdot\nu$ for $b > 1$, we recover Mini-Batch Stochastic Mirror Descent, where at each iteration we use a gradient estimator obtained by averaging across b i.i.d. examples. To see this, note that solving (17) as in Theorem 1 we have that for $i = 0,\dots,b-1$, $\nabla\psi(\mathbf{w}(k\eta + (i+1)\nu)) = \nabla\psi(\mathbf{w}(k\eta + i\nu)) - \nu\nabla f(\mathbf{w}(k\eta), z_{k\eta+i\nu})$ and so $\nabla\psi(\mathbf{w}((k+1)\eta)) = \nabla\psi(\mathbf{w}(k\eta + b\nu)) = \nabla\psi(\mathbf{w}(k\eta)) - \eta\frac{1}{b}\sum_i \nabla f(\mathbf{w}(k\eta), z_{k\eta+i\nu})$.

At an extreme, as $\nu \to 0$, the solution of (17) converges to the solution of the Mirror Descent discretization (MD) and we recover the exact Mirror Descent updates on the population objective.

It is also interesting to consider $\eta < \nu$, in particular when $\eta \to 0$ while $\nu > 0$ is fixed. This corresponds

to optimization using stochastic (infinitesimal) gradient flow, where over a time T we use T/ν samples. Studying how close the discretization (17) remains to the population Riemannian Gradient Flow (GF), in terms of η and ν, could allow us to tease apart the optimization complexity and sample complexity of learning (minimizing the population objective).

B Convergence of Mirrorless Mirror Descent

Theorem 2. *Let the metric tensor $0 \prec \alpha I \preceq H(\mathbf{w}) \preceq \beta I$ for all $\mathbf{w}$ and let F be γ-smooth and λ-strongly convex with respect to L2. Then the updates (10) with constant stepsize $\eta = \frac{\alpha^2}{\gamma\beta}$ will converge at a rate*

$$F(\mathbf{w}_{MD}^{(K)}) - F^* \leq \left(F(\mathbf{w}_{MD}^{(0)}) - F^*\right) \exp\left(-\frac{\lambda\alpha^2 K}{\gamma\beta^2}\right)$$

Proof. Throughout this proof, we will use $\|\cdot\|$ exclusively to denote the L2 norm. We begin by observing that

$$\begin{aligned}\mathbf{w}_{\mathrm{MD}}^{(k+1)} - \mathbf{w}_{\mathrm{MD}}^{(k)} &= -\int_{\eta k}^{\eta(k+1)} H(\mathbf{w}(t))^{-1}\nabla F(\mathbf{w}_{\mathrm{MD}}^{(k)})dt \\ &= -\hat{H}_k \nabla F(\mathbf{w}_{\mathrm{MD}}^{(k)})\end{aligned}$$

for some matrix $\eta\beta^{-1}I \preceq \hat{H}_k \preceq \eta\alpha^{-1}I$. Furthermore, by the γ-smoothness of F

$$\begin{aligned}&F(\mathbf{w}_{\mathrm{MD}}^{(k+1)}) - F^* \\ &\leq F(\mathbf{w}_{\mathrm{MD}}^{(k)}) - F^* + \langle \nabla F(\mathbf{w}_{\mathrm{MD}}^{(k)}), \mathbf{w}_{\mathrm{MD}}^{(k+1)} - \mathbf{w}_{\mathrm{MD}}^{(k)}\rangle \\ &\qquad + \frac{\gamma}{2}\left\|\mathbf{w}_{\mathrm{MD}}^{(k+1)} - \mathbf{w}_{\mathrm{MD}}^{(k)}\right\|^2 \\ &= F(\mathbf{w}_{\mathrm{MD}}^{(k)}) - F^* - \langle \nabla F(\mathbf{w}_{\mathrm{MD}}^{(k)}), \hat{H}_k \nabla F(\mathbf{w}_{\mathrm{MD}}^{(k)})\rangle \\ &\qquad + \frac{\gamma}{2}\left\|\hat{H}_k \nabla F(\mathbf{w}_{\mathrm{MD}}^{(k)})\right\|^2 \\ &\leq F(\mathbf{w}_{\mathrm{MD}}^{(k)}) - F^* - \eta\beta^{-1}\left\|\nabla F(\mathbf{w}_{\mathrm{MD}}^{(k)})\right\|^2 \\ &\qquad + \frac{\eta^2\gamma}{2\alpha^2}\left\|\nabla F(\mathbf{w}_{\mathrm{MD}}^{(k)})\right\|^2 \\ &= F(\mathbf{w}_{\mathrm{MD}}^{(k)}) - F^* - \frac{\alpha^2}{2\gamma\beta^2}\left\|\nabla F(\mathbf{w}_{\mathrm{MD}}^{(k)})\right\|^2\end{aligned}$$

where we used that $\eta = \frac{\alpha^2}{\gamma\beta}$ for the final equality. Finally, we note that by the λ-strong convexity of F, for any $\mathbf{w}$

$$\|\nabla F(\mathbf{w})\|^2 \geq 2\lambda\left(F(\mathbf{w}) - F^*\right)$$

We conclude that

$$F(\mathbf{w}_{\mathrm{MD}}^{(k+1)}) - F^* \leq \left(1 - \frac{\lambda\alpha^2}{\gamma\beta^2}\right)\left(F(\mathbf{w}_{\mathrm{MD}}^{(k)}) - F^*\right)$$

Unrolling this recursion yields the stated bound. □

References

P-A Absil, Robert Mahony, and Rodolphe Sepulchre. *Optimization algorithms on matrix manifolds*. Princeton University Press, 2009.

Shun-Ichi Amari. Natural gradient works efficiently in learning. *Neural computation*, 1998.

Shun-ichi Amari. *Differential-geometrical methods in statistics*. Springer Science & Business Media, 2012.

Shun-ichi Amari and John Armstrong. Curvature of hessian manifolds. *Differential Geometry and its Applications*, pages 1–12, 2014.

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

Manfredo P Do Carmo. *Differential Geometry of Curves and Surfaces: Revised and Updated Second Edition*. Courier Dover Publications, 2016.

Suriya Gunasekar, Jason Lee, Daniel Soudry, and Nathan Srebro. Characterizing implicit bias in terms of optimization geometry. In *International Conference on Machine Learning*, pages 1832–1841, 2018.

Anatoli Juditsky, Joon Kwon, and Éric Moulines. Unifying mirror descent and dual averaging. *arXiv preprint arXiv:1910.13742*, 2019.

Arkadii Semenovich Nemirovsky and David Borisovich Yudin. Problem complexity and method efficiency in optimization. 1983.

Maxim Raginsky and Jake Bouvrie. Continuous-time stochastic mirror descent on a network: Variance reduction, consensus, convergence. In *IEEE Conference on Decision and Control (CDC)*. IEEE, 2012.

Garvesh Raskutti and Sayan Mukherjee. The information geometry of mirror descent. *IEEE Transactions on Information Theory*, 2015.

Shai Shalev-Shwartz. Online learning and online convex optimization. *Foundations and Trends® in Machine Learning*, pages 107–194, 2012.

Nati Srebro, Karthik Sridharan, and Ambuj Tewari. On the universality of online mirror descent. In *Advances in neural information processing systems*, pages 2645–2653, 2011.

Manfred K Warmuth and Arun K Jagota. Continuous and discrete-time nonlinear gradient descent: Relative loss bounds and convergence. In *International Symposium on Artificial Intelligence and Mathematics*, 1997.

Andre Wibisono, Ashia C Wilson, and Michael I Jordan. A variational perspective on accelerated methods in optimization. *Proceedings of the National Academy of Sciences*, pages E7351–E7358, 2016.