Gradient Descent using Duality Structures

Thomas Flynn tflynn@gradcenter.cuny.edu
Department of Computer Science
Graduate Center of CUNY
New York, NY 10016, USA

Abstract

In most applications of gradient-based optimization to complex problems the choice of step size is based on trial-and-error and other heuristics. A case when it is easy to choose the step sizes is when the function has a Lipschitz continuous gradient. Many functions of interest do not appear at first sight to have this property, but often it can be established with the right choice of underlying metric. We find a simple recipe for choosing step sizes when a function has a Lipschitz gradient with respect to any Finsler structure that verifies an exponential bound. When relevant problem structure can be encoded in the metric to yield a significantly tighter bound while keeping optimization tractable, this may lead to rigorous and efficient algorithms. Our general result can be applied to yield an optimization algorithm with non-asymptotic performance guarantees for batch optimization of multilayer neural networks.

1 Introduction

The past decade has witnessed significant advances in the application of neural networks to computer vision problems, such as representation learning [10, 8], image classification [4, 6], scene labeling [2], and multimodal processing [11]. All of these works achieve their goals through gradient based optimization, using carefully tuned heuristics to determine the step size taken at each iteration. Using a more rigorous approach to gradient descent in these problems can improve the practice of machine learning, for instance by avoiding the time consuming process of manually tuning algorithms.

Here we consider a generalization of Euclidean gradient descent where instead of requiring a global bound on the norm of the second derivative of the function of interest, we require that the second derivative be bounded with respect to a Finsler structure. We find that if the Finsler structure obeys certain exponential bounds, then exact solution of the corresponding line search problems yields a convergence guarantee. If the Finsler structures themselves are not too complicated, these sub-problems can be easily solved and the procedure becomes practical. When we apply this to machine learning classification tasks, the result is a full-batch gradient descent method for minimizing the empirical error. Our main result says the function is guaranteed to decrease on every iteration, and also provides a bound on the number of iterations needed to reach a point with an arbitrarily small gradient, measured with respect to the local norm determined by the Finsler structure. We then show that the algorithm, and its associated performance guarantee, is applicable to multilayer neural networks, by constructing a Finsler structure which reflects the hierarchical structure of the network. Numerical experiments on standard data sets suggest the resulting step sizes are not too conservative. All proofs may be found in the full version of the paper [3].

2 Finsler Gradient Descent

Let $W = \mathbb{R}^n$ be the parameter space. We begin by defining Finsler duality structures.

Definition 2.1. Let $\|\cdot\|_w$ be an assignment of a norm on $\mathbb{R}^n$ to each point of W. The notation $\|u\|_w$ refers to the norm of the vector u at the parameter w. We say that $\|\cdot\|_w$ is a *Finsler structure* if the map $(w, u) \mapsto \|u\|_w$ is continuous on $W \times \mathbb{R}^n$.

The Finsler structure induces a norm on the dual $\mathcal{L}(\mathbb{R}^n, \mathbb{R})$ at each point w; if $\ell \in \mathcal{L}(\mathbb{R}^n, \mathbb{R})$ then

$$\|\ell\|_w = \sup_{\|u\|_w = 1} \ell(u). \tag{1}$$

OPTML 2017: 10th NIPS Workshop on *Optimization for Machine Learning* (NIPS 2017).

Definition 2.2. A *duality structure* is an assignment of a duality map to each $w \in W$. The notation $\rho(\ell)_w$ refers to the value of the duality map at w applied to the functional ℓ. That is, a duality structure is a function $\rho : W \times \mathcal{L}(\mathbb{R}^n, \mathbb{R}) \to \mathbb{R}^n$ satisfying, for all $(w, \ell) \in W \times \mathcal{L}(\mathbb{R}^n, \mathbb{R})$, the two properties $\|\rho(\ell)_w\|_w = 1$ and $\ell(\rho(\ell)_w) = \|\ell\|_w$.

We introduce a growth condition on the Finsler structure.

Assumption 2.3. *There is an $\eta \geq 0$ so that for all $(w, \ell) \in W \times \mathcal{L}(\mathbb{R}^n, \mathbb{R})$ and $\lambda \in \mathbb{R}$, if $u = \rho(\ell)_w$, then $\|u\|_{w+\lambda u} \leq \|u\|_w \exp(\eta|\lambda|)$.*

Let $f : W \to \mathbb{R}$ be the function to be optimized. We introduce the notation Δw to refer to $\rho(\frac{\partial f}{\partial w}(w))_w$, that is, the duality map at w applied to the linear functional $\frac{\partial f}{\partial w}(w)$.

Assumption 2.4. *The function f is twice differentiable, bounded from below with $f \geq f^*$, and there is an $L \geq 0$ such that, for all $w \in W$ and $\lambda \geq 0$, $|\frac{\partial^2 f}{\partial w^2}(w - \lambda \Delta w)[\Delta w, \Delta w]| \leq L \|\Delta w\|^2_{w - \lambda \Delta w}$.*

We now describe the algorithm and convergence guarantee. We obtain convergence of the gradients in terms of the local-norms $\|\cdot\|_{w(n)}$; this is a common criteria for gradient convergence in the manifold setting, and can be compared with Theorem 4 of [1] or, in the stochastic case, Theorem 2 of [12].

Theorem 2.5. *Let Assumptions 2.3 and 2.4 hold. Starting from $w(0) \in W$, define $w(n)$ as*

$$w(n+1) = w(n) - \epsilon(n)\Delta(n) \tag{2}$$

where $\Delta(n) = \rho\left(\frac{\partial f}{\partial w}(w(n))\right)_{w(n)}$ and

$$\epsilon(n) = \arg\min_{\epsilon} \left[-\epsilon \|\tfrac{\partial f}{\partial w}(w(n))\|_{w(n)} + L \int_0^{\epsilon} \int_0^{u} \|\Delta(n)\|^2_{w(n) - \lambda \Delta(n)} d\lambda du \right]. \tag{3}$$

Then $\|\frac{\partial f}{\partial w}(w(n))\|_{w(n)} \to 0$. Furthermore, the following non-asymptotic performance guarantee holds. Defining the function

$$g(x, \eta, L) = \begin{cases} \frac{1}{2\eta}\left[\log(1 + \frac{2\eta}{L}x)(x + \frac{L}{2\eta}) - x\right] & \text{if } \eta > 0, \\ \frac{x^2}{2L} & \text{if } \eta = 0, \end{cases}$$

then $\min_{0 \leq i \leq n-1} \|\frac{\partial f}{\partial w}(w(i))\|_{w(i)} \leq \epsilon$ when $n \geq \frac{1}{g(\epsilon, \eta, L)} (f(w_0) - f^*)$.

3 Application to Neural Networks with Multiple Layers

In any application of the methodology there are three tasks. First, one must define the Finsler and duality structures for the space, and check that the exponential bounds hold. Secondly, one must verify the Lipschitz-like condition on the gradient. This determines the search directions. Finally, one must devise a solution to the resulting optimization problems, in order to obtain the step sizes.

Let the input to a neural network be of dimensionality n_0, and let $n_1, \ldots, n_K$ specify the number of nodes in each of $K-1$ non-input layers. For $k = 1, \ldots, K$ define $W_k = \mathbb{R}^{n_k \times n_{k-1}}$ to be the space of $n_k \times n_{k-1}$ matrices; a matrix in W_k specifies weights from nodes in layer $k-1$ to nodes in layer k. The overall parameter space is then $W = W_1 \times \ldots \times W_{K-1}$. Let us denote the 2-norm by $\|\cdot\|_2$. For an input $y \in \mathbb{R}^{n_0}$, the output of the network is $x^K(w; y) \in \mathbb{R}^{n_K}$ where $x^0(w; y) = y$ and for $1 \leq l \leq K$,

$$x_i^k(w; y) = \sigma\left(\sum_{j=1}^{n_{k-1}} w_{k,i,j} x_j^{k-1}(w; y) \right), \quad i = 1, 2, \ldots, n_k.$$

Given m input/output pairs $(y_1, t_1), (y_2, t_2), \ldots, (y_m, t_m)$, where $(y_n, t_n) \in \mathbb{R}^{n_0} \times \mathbb{R}^{n_K}$, we seek to minimize the *empirical error*

$$f(w) = \frac{1}{m} \sum_{i=1}^{m} \|x^K(w; y_i) - t_i\|_2^2. \tag{4}$$

Our assumptions on the nonlinearity σ, the inputs y_i, and the targets t_i, are as follows:

Assumption 3.1. $\|\sigma\|_\infty \leq 1$, $\|\sigma'\|_\infty < \infty$, $\|\sigma''\|_\infty < \infty$ *(Nonlinearity bounds) and for $i = 1, 2, \ldots, m$, $\|y_i\|_\infty \leq 1$ and $\|t_i\|_\infty \leq 1$ (Input/target bounds).*

We define the Finsler structure on W:

1. Each space $\mathbb{R}^{n_i}$ has the norm $\|\cdot\|_\infty$.

2. The spaces $W_1, \ldots, W_K$ have the norm induced by $\|\cdot\|_\infty$, which is the maximum-absolute-row-sum norm: for an $r \times c$ matrix, $\|m\|_\infty = \max_{1\le i\le r} \sum_{j=1}^{c} |m_{i,j}|$.

3. The Finsler structure is then defined as

$$\|(\delta w_1, \ldots, \delta w_K)\|_w = p_1(w)\|\delta w_1\|_\infty + \ldots + p_K(w)\|\delta w_K\|_\infty \tag{5}$$

where the functions p_i are defined as follows. Let $r_0 = 1$ and for $n > 0$ define $r_n(z_1, \ldots, z_n) = \|\sigma'\|_\infty^n \prod_{i=1}^n z_i$. Then define q_n recursively, with $q_0 = 0$, $q_1(z_1) = \|\sigma''\|_\infty z_1^2$, and for $n > 1$,

$$q_n(z_1, \ldots, z_n) = \|\sigma''\|_\infty z_n^2 \|\sigma'\|_\infty^{2(n-1)} \prod_{i=1}^{n-1} z_i^2 + \|\sigma'\|_\infty z_n q_{n-1}(z_1, \ldots, z_{n-1})$$

Define $s_0, \ldots, s_{K-1}$ as

$$s_i(z_1, \ldots, z_i) = n_K \|\sigma'\|_\infty^2 r_i^2(z_1, \ldots, z_i) + 2n_K \|\sigma'\|_\infty^2 q_i(z_1, \ldots, z_i) + 2n_K \|\sigma''\|_\infty r_i(z_1, \ldots, z_i)$$

Finally, the $p_1, \ldots, p_K$ are $p_i(w) = \sqrt{s_{K-i}(\|w_{i+1}\|_\infty, \ldots, \|w_K\|_\infty) + 1}$. (6)

For example, in a network with one hidden layer, the two polynomials p_1, p_2 are

$$p_1(w) = \sqrt{(n_2\|\sigma'\|_\infty^4 + 2n_2\|\sigma'\|_\infty^2\|\sigma''\|_\infty)\|w_2\|_\infty^2 + 2n_2\|\sigma'\|_\infty\|\sigma''\|_\infty\|w_2\|_\infty + 1} \tag{7}$$

$$p_2(w) = \sqrt{n_2(\|\sigma'\|_\infty^2 + 2\|\sigma''\|_\infty) + 1}. \tag{8}$$

To obtain the duality structure, first we derive a duality map for matrices with the norm $\|\cdot\|_\infty$, and then use a standard construction for product spaces. The first part is summarized in the following.

Proposition 3.2. *Let $\ell \in \mathcal{L}(\mathbb{R}^{r\times c}, \mathbb{R})$ be defined on the space of matrices with the norm $\|\cdot\|_\infty$. Then*

$$\|\ell\|_\infty = \textstyle\sum_{i=1}^r \max_{1\le j\le c} |\ell_{i,j}| \tag{9}$$

and one duality map is ρ_∞, which sends ℓ to a matrix that 'picks out' a maximum in each row:

$$\rho(\ell)_\infty = m \text{ where } m_{i,j} = \begin{cases} \operatorname{sgn}(\ell_{i,j}) \text{ if } j = \arg\max_k |w_{i,k}|, \\ 0 \text{ otherwise.} \end{cases} \tag{10}$$

In the next result we construct a duality map for a product space from duality maps on the components.

Proposition 3.3. *If $X_1, \ldots, X_n$ are normed spaces, carrying duality maps $\rho_{X_1}, \ldots, \rho_{X_n}$ respectively, and the product $Z = X_1 \times \ldots \times X_n$ has norm $\|(x_1, \ldots, x_n)\|_Z = p_1\|x_1\|_{X_1} + \ldots + p_n\|x_n\|_{X_n}$, for some positive coefficients $p_1, \ldots, p_n$, then the dual norm for Z is $\|(\ell_1, \ldots, \ell_n)\|_Z = \max\left\{\frac{1}{p_1}\|\ell_1\|_{X_1}, \ldots, \frac{1}{p_n}\|\ell_n\|_{X_n}\right\}$ and a duality map for Z is given by $\rho((\ell_1, \ldots, \ell_n))_Z = \left(0, \ldots, \frac{1}{p_{i^*}}\rho(\ell_{i^*})_{X_{i^*}}, \ldots, 0\right)$ where $i^* = \arg\max_i \left\{\frac{1}{p_i}\|\ell_i\|_{X_i}\right\}$.*

Based on this, we define the Finsler duality structure on W:

1. Each space $W_1, \ldots, W_K$ has the duality map $\rho(\cdot)_\infty$, defined according to (10).

2. The duality map at each point w is defined according to Proposition 3.3:

$$\rho((\ell_1, \ldots, \ell_K))_w = \left(0, \ldots, \tfrac{1}{p_{i^*}(w)}\rho(\ell_{i^*})_\infty, \ldots, 0\right) \text{ where } i^* = \arg\max_i \left\{\tfrac{1}{p_i(w)}\|\ell_i\|_\infty\right\} \tag{11}$$

We can now set up the line search problems at each step and determine their solution. Let $w \in W$ and let $\Delta w = \rho(\frac{\partial f}{\partial w}(w))_w$. Set $i^* = \arg\max_i \left\{\frac{1}{p_i(w)}\|\frac{\partial f}{\partial w_i}(w)\|_\infty\right\}$. Then $\arg\min_\epsilon \left[-\epsilon\|\frac{\partial f}{\partial w}(w)\|_w + \int_0^\epsilon \int_0^u \|\Delta w\|_{w-\lambda\Delta w}^2 d\lambda du\right] = \frac{1}{p_{i^*}(w)}\|\frac{\partial f}{\partial w_{i^*}}(w)\|_\infty$.

We now arrive at the convergence result for batch training of multilayer networks:

Proposition 3.4. *Let f be defined as in (4), let Assumption 3.1 hold, and endow W with the Finsler structure (5) and duality structure (11). Then Assumption 2.3 is satisfied with $\eta = 0$, Assumption 2.4 is satisfied with $L = 1$, and the sequence $w(n)$ defined by Eqn. 2 is guaranteed to satisfy the conclusion of Theorem 2.5. In particular, $\min_{0\le i\le n-1}\|\frac{\partial f}{\partial w}(w(i))\|_{w(i)} \le \epsilon$ when $n \ge \frac{2L}{\epsilon^2} f(w_0)$.*

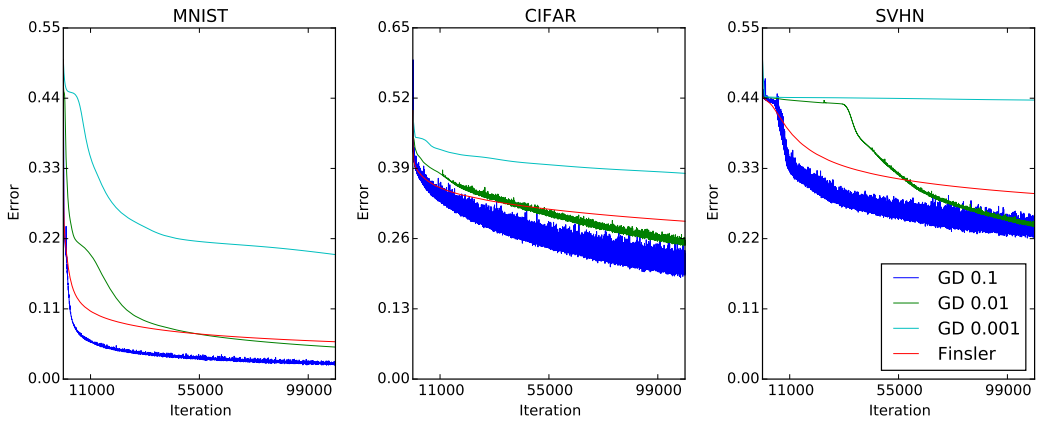

Figure 1: A comparison of Finsler gradient descent with normal Euclidean gradient descent with constant step sizes. The curves indicate the empirical error (Eqn. 4).

4 Numerical Experiment

We considered minimization of the empirical error in the MNIST [7], SVHN [9], and CIFAR-10 [5] classification tasks. See [3] for architecture and initialization details. In all cases the objective f was the average squared error over the first 10,000 training examples.

Algorithm 1: Finsler gradient descent for a network with one hidden layer

for $n = 0, 1, \ldots$ **do**

Compute $\frac{\partial f}{\partial w_1}(w(n))$, $\frac{\partial f}{\partial w_2}(w(n))$ via back-propagation.

Compute $\frac{1}{p_1(w(n))}\|\frac{\partial f}{\partial w_1}(w(n))\|_\infty$ and $\frac{1}{p_2(w(n))}\|\frac{\partial f}{\partial w_2}(w(n))\|_\infty$ via eqns. (7, 8, 9).

Compute the matrices $\rho(\frac{\partial f}{\partial w_1}(w(n)))_\infty$ and $\rho(\frac{\partial f}{\partial w_2}(w(n)))_\infty$ via equation (10).

if $\frac{1}{p_2(w(n))}\|\frac{\partial f}{\partial w_2}(w(n))\|_\infty > \frac{1}{p_1(w(n))}\|\frac{\partial f}{\partial w_1}(w(n))\|_\infty$ **then**

$$w(n+1) = \left(w_1(n),\ w_2(n) - \tfrac{1}{p_2(w(n))}\|\tfrac{\partial f}{\partial w_2}(w(n))\|_\infty \rho(\tfrac{\partial f}{\partial w_2}(w(n)))_\infty\right)$$

else

$$w(n+1) = \left(w_1(n) - \tfrac{1}{p_1(w(n))}\|\tfrac{\partial f}{\partial w_1}(w(n))\|_\infty \rho(\tfrac{\partial f}{\partial w_1}(w(n)))_\infty,\ w_2(n)\right)$$

end

end

We compared Finsler gradient descent (Algorithm 1) with vanilla Euclidean gradient descent (GD). For each of the three problems we ran four algorithms: GD with a constant step size of $\epsilon = 0.1$, 0.01, and 0.001, and the Finsler gradient descent. The results are shown in Figure 1. In all cases we see that the step size has a big effect on the behavior of the algorithm. Small step sizes like $\epsilon = 0.001$ lead to very slow optimization. The function decreases much faster with larger ϵ, but this leads to oscillations. The Finsler based optimization procedure, on the other hand, always produces a smooth decrease in the empirical error (guaranteed by Proposition 3.4), with at most a moderate slow down compared to the large step size algorithms. Note that GD requires extra time to tune the step size.

5 Discussion

In this work we presented an approach to neural network optimization which involves computing step sizes and search directions with the help of a pair of geometric structures: a Finsler structure and a duality structure. Numerical results suggest the promise of the approach. A reason for this may be that our framework is better able to integrate problem structure as compared to naive Euclidean gradient descent. The Finsler structure uses a good deal of problem information, such as the hierarchical structure of the network, bounds on various derivatives, and bounds on the input.

References

[1] N. Boumal, P.-A. Absil, and C. Cartis. Global rates of convergence for nonconvex optimization on manifolds. *arXiv preprint arXiv:1605.08101*, 2016.

[2] C. Farabet, C. Couprie, L. Najman, and Y. LeCun. Learning hierarchical features for scene labeling. *IEEE transactions on pattern analysis and machine intelligence*, 35(8):1915–1929, 2013.

[3] T. Flynn. Gradient Descent using Duality Structures. *ArXiv e-prints*.

[4] G. E. Hinton and R. R. Salakhutdinov. Reducing the dimensionality of data with neural networks. *Science*, 313(5786):504–507, 2006.

[5] A. Krizhevsky. Learning multiple layers of features from tiny images. Technical report, University of Toronto, 2009.

[6] A. Krizhevsky, I. Sutskever, and G. Hinton. Imagenet classification with deep convolutional neural networks. In *Advances in Neural Information Processing Systems 25*, pages 1106–1114, 2012.

[7] Y. LeCun, C. Cortes, and C. J. Burges. The MNIST database of handwritten digits. http://yann.lecun.com/exdb/mnist/.

[8] H. Lee, R. Grosse, R. Ranganath, and A. Y. Ng. Convolutional deep belief networks for scalable unsupervised learning of hierarchical representations. In *Proceedings of the 26th annual international conference on machine learning*, pages 609–616. ACM, 2009.

[9] Y. Netzer, T. Wang, A. Coates, A. Bissacco, B. Wu, and A. Y. Ng. Reading digits in natural images with unsupervised feature learning. In *NIPS Workshop on Deep Learning and Unsupervised Feature Learning*, 2011.

[10] R. Salakhutdinov and G. E. Hinton. Learning a nonlinear embedding by preserving class neighbourhood structure. In *International Conference on Artificial Intelligence and Statistics*, pages 412–419, 2007.

[11] N. Srivastava and R. Salakhutdinov. Multimodal learning with deep boltzmann machines. *The Journal of Machine Learning Research*, 15(1):2949–2980, 2014.

[12] H. Zhang, S. J. Reddi, and S. Sra. Riemannian svrg: Fast stochastic optimization on riemannian manifolds. In D. D. Lee, M. Sugiyama, U. V. Luxburg, I. Guyon, and R. Garnett, editors, *Advances in Neural Information Processing Systems 29*, pages 4592–4600. Curran Associates, Inc., 2016.