Project description **Nicholas M. Boffi**

1 Introduction and background

Optimization problems are ubiquitous. Diverse applications in mathematics, science, and engineering all rely on optimization, such as training statistical models for image recognition tasks, inverse materials design problems, and control of nonlinear dynamical systems. The analysis and design of computationally efficient optimization algorithms is essential for powering these common use-cases across applied mathematics.

The goal of this proposal is to apply variational mechanics to analyze fast first-order algorithms for constrained optimization and for adaptive control. Interest in first-order algorithms has exploded in recent years due to the proliferation of high-dimensional statistical models [1]. "Accelerated" algorithms that achieve optimal convergence rates such as Nesterov's celebrated method [2] have been designed for unconstrained optimization [3]. However, our understanding of first-order methods for constrained problems is comparatively limited: for example, fundamental limits on convergence speed are still unknown. Constrained methods are critical for solving high-dimensional statistical problems, where constraints are used to impose sparsity or low-rank structure. Research in adaptive control has also taken off in recent years due to promising applications in robotics and as a model of reinforcement learning [4]. Variational mechanics provides a common methodology for deriving both optimization and adaptive control algorithms.

A significant difficulty in designing fast first-order methods for constrained optimization and adaptive control is the non-intuitive structure of accelerated methods. While technically sound, their convergence proofs are algebraic, and infamously exploit obtuse inequalities at key points. An appealing alternative is to use variational methods, which provide a systematic approach to derive continuous time limits of accelerated optimization algorithms [5]. These limits can then be carefully discretized to obtain an implementation with guarantees that match those of the continuous equation. Symplectic integration has been particularly useful in providing such rate-matching discretizations [6, 7]. While this general approach has proved successful in unconstrained optimization, adding constraints poses some significant mathematical challenges. Similarly, the discretization of continuous adaptive control algorithms derived through variational procedures has not been studied, which is essential for understanding if accelerated convergence is possible at all in adaptive control.

2 Previous work

Continuous time limits are appealing in the design of fast constrained optimization algorithms because they often provide intuitive interpretations of accelerated unconstrained methods [8]. Let $x \in \mathbb{R}^n$ be a vector of parameters, $f : \mathbb{R}^n \to \mathbb{R}$ be a convex loss function, and $\alpha_t : \mathbb{R}_{\geq 0} \to \mathbb{R}$, $\beta_t : \mathbb{R}_{\geq 0} \to \mathbb{R}$, $\gamma_t : \mathbb{R}_{\geq 0} \to \mathbb{R}$ be time-dependent *scaling factors*. Many algorithms can be derived through variational mechanics from a single object known as the *Bregman Lagrangian*

$$\mathcal{L} = e^{\gamma_t - \alpha_t} \left\| \dot{x} \right\|_2^2 - e^{\alpha_t + \beta_t + \gamma_t} f(x), \tag{1}$$

which reveals an underlying geometric principle for acceleration [5]. Accelerated algorithms follow a single optimal curve in parameter space determined by the loss function (Figure 1), and the discrete implementation dictates how many steps it takes to traverse the curve. The norm $\left\| \cdot \right\|_2$ in (1) can be replaced with a more general distance measure known as a Bregman divergence, which is restricted to the Euclidean norm here for simplicity.

In continuous time, convergence proofs are derived through analysis techniques classically applied to nonlinear dynamical systems such as Lyapunov theory [9] and contraction analysis [10]. Algorithm design is completed by discretizing the continuous dynamics, so that the existence of (1) reduces deriving a discrete implementation to numerical analysis of its Euler Lagrange equations.

Through past research in nonlinear dynamical systems, control theory, and computational physics, the PI is well equipped to tackle problems with these two diverse skillsets.

In his thesis work, the PI applied (1) to the setting of adaptive control [11], and a new class of Nesterov-like adaptive control algorithms was derived and proven to be globally convergent through Lyapunov theory. An analogy to mirror descent in optimization [12] was used to design a second new class of adaptive control methods that exploit non-Euclidean geometry. The implicit regularization of these algorithms was categorized by analogy to their optimization counterparts [13], and they were subsequently applied to several interdisciplinary problems such as sparse identification of a chemical reaction network, sparse identification of a Hamiltonian consistent with a physical system, and control of a dynamical system with control primitives.

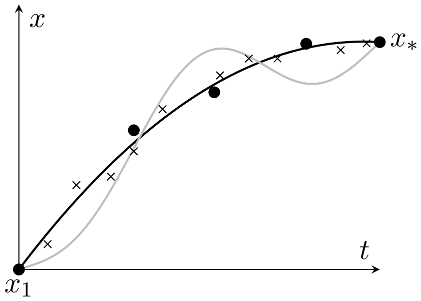

Figure 1: The curve which minimizes (1) is shown in black, and a suboptimal curve is shown in gray. All algorithms generated by (1) follow the black curve. A trajectory from Nesterov's accelerated method [2] is illustrated in circles, and a trajectory from a slower, suboptimal, method is in crosses. Nesterov's method takes larger steps and optimizes faster.

In another work, the PI used contraction analysis to analyze distributed first-order stochastic optimization algorithms from a continuous time lens [14]. New convergence theorems relevant for training high-dimensional statistical models were proven, and novel insight was gained into the behavior of distributed stochastic algorithms through synchronization theory. Synchronization of the distributed optimizers induced by communication was shown to reduce their individual noise levels, and predictions of the theory were verified through numerical simulation on image recognition tasks.

In numerical analysis and computational physics, the PI developed in his thesis work a three-dimensional projection algorithm for simulating the deformation of hard amorphous materials based on a mapping to computational fluid dynamics [15]. The algorithm was also extended to more general domains and boundary conditions using a coordinate transformation methodology [16]. As a Fulbright Scholar, the PI developed a finite difference approach for solving the Hartree-Fock equations from condensed matter physics in real space [17].

3 Manifold constrained optimization

Present approaches for the development of accelerated manifold-constrained optimization algorithms are limited in scope, because the algebraic inequalities used to prove acceleration in Euclidean space often break down in the manifold setting. Remarkably, the variational formulation afforded by (1) provides a simple method to handle manifold constraints by analogy to holonomic constraints in physics. Consider the equality constrained minimization problem

$$\min_x f(x) \quad \text{subject to } g(x) = 0 \tag{2}$$

with $g : \mathbb{R}^n \to \mathbb{R}^m$. If the constraint Jacobian has rank m globally, the surface $g(x) = 0$ has the structure of a differentiable manifold $\mathcal{M}$ and admits local coordinates q (Figure 2). Let $q^\perp$ denote local coordinates orthogonal to $\mathcal{M}$, and define a potential function $U_N(q, q^\perp) = f(x(q, q^\perp)) + N \left\|q^\perp\right\|_2^2$, which imposes a quadratic penalty for constraint violation.

Consider the motion of a particle $x_N(t)$ described by (1), but replace the loss $f(x)$ with the potential $U_N(q, q^\perp)$. Then the limit $x_\infty(t) = \lim_{N\to\infty} x_N(t)$ exists and is a solution to the Euler-Lagrange equations $\frac{d}{dt}\frac{\partial \mathcal{L}_*}{\partial \dot{q}} = \frac{\partial \mathcal{L}_*}{\partial q}$ where $\mathcal{L}_*$ is the Lagrangian with $q^\perp = \dot{q}^\perp = 0$ [18]. Manipulation of these equations leads to the master dynamics

$$\ddot{q}_i + (\dot{\gamma}_t - \dot{\alpha}_t)\, \dot{q}_i + \sum_{b,e} \Gamma^i_{b,e} \dot{q}^b \dot{q}^e + e^{\beta_t + 2\alpha_t} \mathsf{grad}(f)_i = 0, \tag{3}$$

with $\Gamma^i_{b,e}$ the Christoffel symbols for the metric on $\mathcal{M}$ and where $\mathsf{grad}(\cdot)$ denotes the Riemannian gradient in local coordinates. The solution curve $x(q(t))$ to (3) lies entirely on $\mathcal{M}$ by construction. Moreover, (3) has striking theoretical appeal. It takes a similar form to the master equation generated by (1), but has three key differences: it is written entirely in local coordinates on the manifold, an additional term dictated by the manifold geometry appears, and the gradient is replaced by the Riemannian gradient on the manifold.

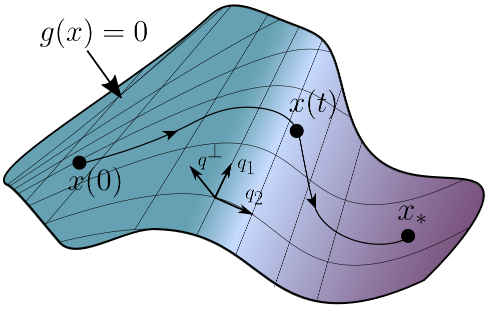

Figure 2: Originally developed for constrained physical problems, Lagrangian mechanics leads to a generative procedure for deriving fast optimization algorithms on manifolds.

A convergence rate of the master equation (3) will be proven through explicit construction of a Lyapunov function by analogy to the unconstrained case [5]. The unconstrained Lyapunov function will be modified to account for the geometry-dependent term in (3), and alternative assumptions on f may be required for convergence, such as geodesic convexity on $\mathcal{M}$ [19].

In the unconstrained setting, a Lyapunov argument shows that the convergence rate of the master dynamics is dictated by the scaling factors, and their choice sets the underlying algorithm. Therefore, once a Lyapunov function has been constructed for the manifold constrained setting, the landscape of possible algorithms generated by (3) can be easily categorized through choice of scaling factors. In particular, picking the scaling factors to correspond to common convergence rates such as $\mathcal{O}(1/t), \mathcal{O}(1/t^2)$, and $\mathcal{O}(e^{-\lambda t})$ for $\lambda > 0$ will provide explicit upper bounds on convergence rates for Riemannian optimization.

To derive implementations, two discretization techniques are proposed inspired by classical mechanics and optimization, respectively. The Hamiltonian form of (3) will be discretized via symplectic integration [7], while a linear coupling scheme [20] will be applied directly to (3). Convergence rates for the discrete iterations will be obtained through discrete time Lyapunov theory by analogy to the unconstrained case [21].

(3) may be expensive to implement if the manifold is high-dimensional or if calculation of $x(q)$ is computationally intensive. The first and third term can be written without reference to coordinates as $\nabla \dot{x}$ where ∇ denotes the Levi-Cevita connection. Guided by this observation, convergence and discretization of the coordinate agnostic dynamics $\nabla \dot{x} + (\dot{\gamma}_t - \dot{\alpha}_t)\,\dot{x} + e^{\beta_t + 2\alpha_t}\mathsf{grad} f(x) = 0$ will be studied through identical Lyapunov approaches, where x is a point on $\mathcal{M}$.

4 Inequality constrained optimization

Consider the inequality constrained problem with $g_i : \mathbb{R}^n \to \mathbb{R}$ and all g_i convex,

$$\min_x \; f(x) \quad \text{subject to } g_i(x) < 0, \quad i = 1, \ldots, m. \tag{4}$$

Many statistical problems such as compressed sensing [22] and matrix completion [23] employ inequality constraints. While some accelerated methods for this setting are known [24], there is no systematic generative procedure to derive accelerated methods, and no geometric understanding as provided by (1) for unconstrained optimization. For (4), the modified Bregman Lagrangian

$$\mathcal{L} = e^{\gamma_t - \alpha_t} \|\dot{x}\|_2^2 - e^{\alpha_t + \beta_t + \gamma_t} \left(f(x) + e^{\delta_t} p(x) \right), \tag{5}$$

is proposed, which contains a new scaling factor $\delta_t : \mathbb{R}_{\geq 0} \to \mathbb{R}$ and a function $p(x) : \mathbb{R}^n \to \mathbb{R}$. The penalty term $p(x)$ is chosen to impose inequality constraints, while the scale factor δ_t provides a degree of freedom to relax or increase the penalty with time. To understand the role of geometric effects in fast optimization, such as specular reflection off the constraint boundary, two penalty models (Figure 3) will be studied corresponding to hard and soft constraints. The Lyapunov function for the unconstrained setting [5] will be modified to account for the new penalty term to study convergence rates in both cases.

For soft constraints, transient constraint violation will be tolerated throughout optimization, but the structure of the loss function will be preserved inside the feasible region. Performance of polynomial penalties $p(x) = \sum_i \max(0, g_i(x)^k)$ for k an integer greater than one will be categorized. Physically, this scenario corresponds to a heavy ball rolling with friction up the side of a potential. Here, δ_t will be chosen as an increasing function of time to ensure the constraints are asymptotically satisfied.

Hard constraints can be implemented using self-concordant barriers, such as the logarithmic barrier $p(x) = -\sum_i \log(-g_i(x))$ [25]. Here, p diverges on the boundary of the constraint set, but is nonzero on the interior. This ensures that constraints will not be violated throughout optimization, and physically corresponds to a smooth infinite potential at the boundary of the constraint region. δ_t will be chosen as a decreasing function of time to restore the loss surface inside the feasible region asympotically and ensure convergence to a minimizer of f.

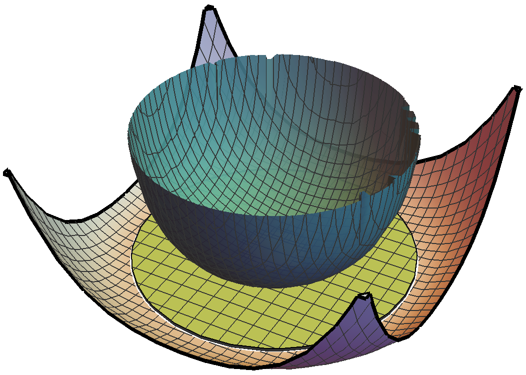

Figure 3: The geometry of hard (blue) and soft (orange) constraints over the disk (yellow).

5 Accelerated adaptive control

Adaptive control theory is the study of concurrent learning and control of dynamical systems. Consider a nonlinear dynamical system $\dot{x} = f(x,t) - Y(x,t)a + u(x,t)$, with $x \in \mathbb{R}^n$ the state, $a \in \mathbb{R}^p$ an unknown vector of parameters, $f : \mathbb{R}^n \times \mathbb{R}_{\geq 0}$ a known nominal dynamics, $Y : \mathbb{R}^n \times \mathbb{R}_{\geq 0} \to \mathbb{R}^{n \times p}$ a known regressor, and $u : \mathbb{R}^n \times \mathbb{R}_{\geq 0} \to \mathbb{R}^n$ the input to the system. Adaptive control uses the input $u = Y(x,t)\hat{a}(t) + u_d(x,t)$ where u_d controls the system towards a desired trajectory $x_d(t)$ in the absence of the unknown $Y(x,t)a$. A continuous optimization algorithm $\frac{d}{dt}\hat{a}$ must then be specified to maintain stability and ensure that $x(t) \to x_d(t)$.

A variational approach for adaptive control was recently developed by the PI via the Bregman Lagrangian [11]. The Euler-Lagrange equations were shown to be stable, convergent, and qualitatively similar to Nesterov's method. However, naive discretizations with standard methods, such as forward Euler and Runge-Kutta approaches, show nearly identical trajectories to classic algorithms (Figure 4). This parallels the optimization setting, where discretization must be performed carefully to match the rate of the continuous dynamics, and raises a fundamental question: *do there exist accelerated algorithms for adaptive control?*

Unlike in optimization, where (1) defines an optimal curve and the discrete iteration dictates the rate at which that curve is traversed, a timescale is fixed *by the system* in adaptive control. Moreover, because the system and adaptation algorithm are coupled in feedback, an optimal curve cannot be specified through parameter space alone. For these geometric reasons, accelerated convergence may be forbidden in adaptive control.

To answer this question, convergence rates will be analyzed through Lyapunov theory for classic gradient-like adaptive control algorithms and new Nesterov-like algorithms derived via (1). Convergence in adaptive control is generally guaranteed asymptotically, because an arbitrary $x_d(t)$ may change suddenly at large time, which forbids finite time guarantees. As a resolution, the study of convergence to *constant* desired trajectories is proposed, where it will be possible to perform a non-asymptotic analysis. In addition to characterizing convergence rates, this analysis will elucidate the dependence of convergence guarantees on relevant problem parameters such as the ambient dimension of the system and the number of unknowns, which can then be used to design mirror descent algorithms with improved dependencies.

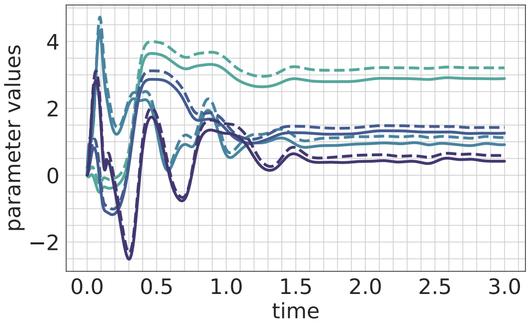

Figure 4: Four parameter trajectories for gradient descent-like (solid) and Nesterov-like (dashed) adaptive control methods, integrated with a fourth-order Runge-Kutta scheme, for control of a double pendulum. The parameter trajectories for the two methods lie directly on top of each other and are shifted for visual clarity.

If convergence rates match for the two classes of methods in continuous time, one possibility is that Nesterov-like algorithms may allow for less frequent updates, akin to larger steps in optimization. By analogy to effective discretization schemes in optimization, maximal timesteps will be probed for symplectic methods and linear coupling schemes, both analytically and numerically. To decouple the system timestep from the algorithm timestep for analytical treatment, the system will be assumed continuous while the parameter updates will be left as a discrete iteration.

6 Significance and broader impacts

The mathematical questions described in this proposal are of both fundamental and applied interest in the areas of statistics, optimization, and control. Because the proposed variational procedures may lead to new algorithms with diverse applications, the PI will develop and release freely available open source software with efficient implementations. The PI's experience contributing to the open source code PARSEC [26, 17] will be helpful in this area.

The PI will pursue research opportunities for motivated undergraduates interested in optimization and dynamical systems, and will be proactive about offering projects to underrepresented minorities and women. One project suitable for an undergraduate with an advanced applied mathematics background would be to contribute to implementations of the manifold constrained algorithms derived via (3), which could later be released as an open source package for Riemannian optimization. A second, more applied, project could be to implement discretizations of Nesterov-like adaptive control laws on real robotic hardware and compare their performance to pre-existing approaches.

7 Sponsoring scientist, institution, and career development

The sponsoring scientist has deep expertise in optimization and statistics, and has made foundational contributions to both fields. His experience in these areas complements the PI's strengths in nonlinear dynamical systems, numerical analysis, and adaptive control. The sponsoring institution, The University of California at Berkeley, has many excellent researchers in statistics, optimization, and control such as B. Recht, P. Bartlett, and M. Wainwright, all of whom frequently collaborate with the sponsor. They will serve as invaluable sources of support and collaboration for the PI during the fellowship.

The PI will benefit greatly from working with the research sponsor to pursue the proposed questions in optimization and adaptive control, as well as from the lively and collaborative environment at Berkeley. The PI's academic career goals will be supported by providing him with the opportunity and resources to pursue compelling and independent research at the intersection of dynamical systems, optimization, and statistics.

References

[1] Léon Bottou. Large-scale machine learning with stochastic gradient descent. In Saporta G. Lechevallier Y., editor, *Proceedings of COMPSTAT 2010*, pages 177–186. Physica-Verlag HD, 2010.

[2] Yurii Nesterov. A Method for Solving a Convex Programming Problem with Convergence Rate $O(1/k^2)$. *Soviet Mathematics Doklady*, 26:367–372, 1983.

[3] Arkadii Nemirovskii and David Yudin. *Problem Complexity and Method Efficiency in Optimization.* John Wiley, 1983.

[4] Yasin Abbasi-Yadkori and Csaba Szepesvári. Regret bounds for the adaptive control of linear quadratic systems. volume 19 of *Proceedings of Machine Learning Research*, pages 1–26, Budapest, Hungary, 09–11 Jun 2011. JMLR Workshop and Conference Proceedings.

[5] Andre Wibisono, Ashia C. Wilson, and Michael I. Jordan. A variational perspective on accelerated methods in optimization. *Proceedings of the National Academy of Sciences*, 113(47), 2016.

[6] Michael Betancourt, Michael I. Jordan, and Ashia C. Wilson. On symplectic optimization. *arXiv:1802.03653*, 2018.

[7] Guilherme Franca, Michael I. Jordan, and René Vidal. On dissipative symplectic integration with applications to gradient-based optimization. *arXiv:2004.06840*, 2020.

[8] Weijie Su, Stephen Boyd, and Emmanuel J. Candès. A differential equation for modeling nesterov's accelerated gradient method: Theory and insights. *Journal of Machine Learning Research*, 17(153):1–43, 2016.

[9] A. M. Lyapunov. *The general problem of the stability of motion (in Russian).* PhD thesis, University of Kharkov, 1892.

[10] Winfried Lohmiller and Jean-Jacques E. Slotine. On contraction analysis for non-linear systems. *Automatica*, 34(6):683–696, 1998.

[11] Nicholas M. Boffi and Jean-Jacques E. Slotine. Implicit regularization and momentum algorithms in nonlinear adaptive control and prediction. *Neural Computation (accepted)*, Sept 2020.

[12] Amir Beck and Marc Teboulle. Mirror descent and nonlinear projected subgradient methods for convex optimization. *Operations Research Letters*, 31(3):167 – 175, 2003.

[13] Suriya Gunasekar, Jason Lee, Daniel Soudry, and Nathan Srebro. Characterizing implicit bias in terms of optimization geometry. volume 80 of *Proceedings of Machine Learning Research*, pages 1832–1841, Stockholmsmässan, Stockholm Sweden, 10–15 Jul 2018. PMLR.

[14] Nicholas M. Boffi and Jean-Jacques E. Slotine. A continuous-time analysis of distributed stochastic gradient. *Neural Computation*, 32(1):36–96, 2020.

[15] Nicholas M. Boffi and Chris H. Rycroft. Parallel three-dimensional simulations of quasi-static elastoplastic solids. *Computer Physics Communications*, 257:107254, 2020.

[16] Nicholas M. Boffi and Chris H. Rycroft. Coordinate transformation methodology for simulating quasistatic elastoplastic solids. *Phys. Rev. E*, 101:053304, 2020.

[17] Nicholas M. Boffi, Manish Jain, and Amir Natan. Efficient computation of the Hartree–Fock exchange in real-space with projection operators. *Journal of Chemical Theory and Computation*, 12(8):3614–3622, 2016.

[18] Vladimir I. Arnold. *Mathematical Methods of Classical Mechanics*. Springer-Verlag, second edition, 1989.

[19] T. Rapcsák. Geodesic convexity in nonlinear optimization. *Journal of Optimization Theory and Applications*, 69(1):169–183, 1991.

[20] Zeyuan Allen-Zhu and Lorenzo Orecchia. Linear coupling: An ultimate unification of gradient and mirror descent. volume 8 of *Innovations in Theoretical Computer Science (ITCS 2017)*, page 3:1–3:2, 2017.

[21] Ashia C. Wilson. *A Lyapunov Analysis of Momentum Methods in Optimization*. PhD thesis, University of California, Berkeley, 2018.

[22] E. J. Candes, J. Romberg, and T. Tao. Robust uncertainty principles: exact signal reconstruction from highly incomplete frequency information. *IEEE Transactions on Information Theory*, 52(2):489–509, 2006.

[23] Emmanuel J. Candès and Benjamin Recht. Exact matrix completion via convex optimization. *Foundations of Computational Mathematics*, 9(6):717, Apr 2009.

[24] Amir Beck and Marc Teboulle. A fast iterative shrinkage-thresholding algorithm for linear inverse problems. *SIAM J. Img. Sci.*, 2(1):183–202, 2009.

[25] Yurii Nesterov. *Introductory Lectures on Convex Optimization*. 1st edition, 2004.

[26] http://real-space.org/.