On the Theory of Gradient-Based Learning: A View from Continuous Time

Michael I. Jordan
University of California, Berkeley

May 17, 2019
Machine Learning (aka, AI) Successes

- First Generation (‘90-’00): the backend
  - e.g., fraud detection, search, supply-chain management
- Second Generation (‘00-’10): the human side
  - e.g., recommendation systems, commerce, social media
- Third Generation (‘10-now): pattern recognition
  - e.g., speech recognition, computer vision, translation
- Fourth Generation (emerging): decisions and markets
  - not just one agent making a decision or sequence of decisions
  - rather, a huge interconnected web of data, agents, decisions
  - many new challenges!
Algorithmic and Theoretical Progress

• Nonconvex optimization
  – avoidance of saddle points
  – rates that have dimension dependence
  – acceleration, dynamical systems and lower bounds
  – statistical guarantees from optimization guarantees

• Computationally-efficient sampling
  – nonconvex functions
  – nonreversible MCMC
  – links to optimization

• Market design
  – approach to saddle points
  – recommendations and two-way markets
Sampling vs. Optimization: The Tortoise and the Hare

- Folk knowledge: Sampling is slow, while optimization is fast
  - but sampling provides *inferences*, while optimization only provides *point estimates*
- But there hasn’t been a clear theoretical analysis that establishes this folk knowledge as true
Sampling vs. Optimization: The Tortoise and the Hare

• Folk knowledge: Sampling is slow, while optimization is fast
  – but sampling provides *inferences*, while optimization only provides *point estimates*

• But there hasn’t been a clear theoretical analysis that establishes this folk knowledge as true

• Is it really true?
Sampling vs. Optimization: The Tortoise and the Hare

- Folk knowledge: Sampling is slow, while optimization is fast
  - but sampling provides inferences, while optimization only provides point estimates
- But there hasn’t been a clear theoretical analysis that establishes this folk knowledge as true
- Is it really true?
- Define the *mixing time*:

  $$\tau(\epsilon, p^0) = \min\{k \mid \|p^k - p^*\|_{TV} \leq \epsilon\}$$

- We’ll study the Unadjusted Langevin Algorithm (ULA) and the Metropolis-Adjusted Langevin Algorithm (MALA)
Theorem. For \( p^* \propto e^{-U} \), we assume that \( U \) is \( m \)-strongly convex outside of a region of radius \( R \) and \( L \)-smooth. Let \( \kappa = L/m \) denote the condition number of \( U \). Let \( p^0 = \mathcal{N}(0, \frac{1}{L} I) \) and let \( \epsilon \in (0, 1) \). Then ULA satisfies

\[
\tau_{ULA}(\epsilon, p^0) \leq \mathcal{O} \left( e^{32LR^2} \kappa^2 \frac{d}{\epsilon^2} \ln \left( \frac{d}{\epsilon^2} \right) \right).
\]

For MALA,

\[
\tau_{MALA}(\epsilon, p^0) \leq \mathcal{O} \left( e^{16LR^2} \kappa^{1.5} \left( d \ln \kappa + \ln \left( \frac{1}{\epsilon} \right) \right)^{3/2} d^{1/2} \right).
\]
**Optimization**

**Theorem.** For any radius $R > 0$, Lipschitz and strong convexity constants $L \geq 2m > 0$, probability $0 < p \leq 1$, there exists an objective function $U(x)$ where $x \in \mathbb{R}^d$ and $U$ is $L$-Lipschitz smooth and $m$-strongly convex for $\|x\|_2 > 2R$, such that for any optimization algorithm that inputs $\{U(x), \nabla U(x), \ldots, \nabla^n U(x)\}$, for some $n$, at least

$$K \geq \mathcal{O}\left(p \cdot \left( LR^2 / \epsilon \right)^{d/2}\right)$$

steps are required for $\epsilon \leq \mathcal{O}(LR^2)$ so that $P(|U(x_K) - U(x^*)| < \epsilon) \geq p$. 
Part I: How to Escape Saddle Points Efficiently

with Chi Jin, Praneeth Netrapalli, Rong Ge, and Sham Kakade
The Importance of Saddle Points

- How to escape?
  - need to have a negative eigenvalue that’s strictly negative
- How to escape efficiently?
  - in high dimensions how do we find the direction of escape?
  - should we expect exponential complexity in dimension?
A Few Facts

- Gradient descent will asymptotically avoid saddle points (Lee, Simchowitz, Jordan & Recht, 2017)
- Gradient descent can take exponential time to escape saddle points (Du, Jin, Lee, Jordan, & Singh, 2017)
- Stochastic gradient descent can escape saddle points in polynomial time (Ge, Huang, Jin & Yuan, 2015)
  - but that’s still not an explanation for its practical success
- Can we prove a stronger theorem?
Optimization

Consider problem:

\[
\min_{x \in \mathbb{R}^d} f(x)
\]

Gradient Descent (GD):

\[
x_{t+1} = x_t - \eta \nabla f(x_t).
\]

Convex: converges to global minimum; dimension-free iterations.
Convergence to FOSP

Function $f(\cdot)$ is $\ell$-smooth (or gradient Lipschitz)

$$\forall x_1, x_2, \|\nabla f(x_1) - \nabla f(x_2)\| \leq \ell \|x_1 - x_2\|.$$ 

Point $x$ is an $\epsilon$-first-order stationary point ($\epsilon$-FOSP) if

$$\|\nabla f(x)\| \leq \epsilon$$

**Theorem** [GD Converges to FOSP (Nesterov, 1998)]

For $\ell$-smooth function, GD with $\eta = 1/\ell$ finds $\epsilon$-FOSP in iterations:

$$\frac{2\ell(f(x_0) - f^*)}{\epsilon^2}$$

*Number of iterations is dimension free.*
Nonconvex Optimization

**Non-convex**: converges to Stationary Point (SP) $\nabla f(x) = 0$.

SP: local min / local max / saddle points

Many applications: no spurious local min (see full list later).
Definitions and Algorithm

Function $f(\cdot)$ is $\rho$-Hessian Lipschitz if

$$\forall x_1, x_2, \| \nabla^2 f(x_1) - \nabla^2 f(x_2) \| \leq \rho \| x_1 - x_2 \|.$$ 

Point $x$ is an $\epsilon$-second-order stationary point ($\epsilon$-SOSP) if

$$\| \nabla f(x) \| \leq \epsilon, \quad \text{and} \quad \lambda_{\min}(\nabla^2 f(x)) \geq -\sqrt{\rho \epsilon}.$$
Definitions and Algorithm

Function $f(\cdot)$ is $\rho$-Hessian Lipschitz if

$$\forall x_1, x_2, \|\nabla^2 f(x_1) - \nabla^2 f(x_2)\| \leq \rho \|x_1 - x_2\|.$$ 

Point $x$ is an $\epsilon$-second-order stationary point ($\epsilon$-SOSP) if

$$\|\nabla f(x)\| \leq \epsilon, \quad \text{and} \quad \lambda_{\min}(\nabla^2 f(x)) \geq -\sqrt{\rho \epsilon}$$

Algorithm Perturbed Gradient Descent (PGD)

1. for $t = 0, 1, \ldots$ do
2. if perturbation condition holds then
3. $x_t \leftarrow x_t + \xi_t, \quad \xi_t \text{ uniformly } \sim B_0(r)$
4. $x_{t+1} \leftarrow x_t - \eta \nabla f(x_t)$

Adds perturbation when $\|\nabla f(x_t)\| \leq \epsilon$; no more than once per $T$ steps.
Main Result

**Theorem** [PGD Converges to SOSP]

For \( \ell \)-smooth and \( \rho \)-Hessian Lipschitz function \( f \), PGD with \( \eta = O(1/\ell) \) and proper choice of \( r, T \) w.h.p. finds \( \epsilon \)-SOSP in iterations:

\[
\tilde{O}\left(\frac{\ell(f(x_0) - f^*)}{\epsilon^2}\right)
\]

*Dimension dependence in iteration is \( \log^4(d) \) (almost dimension free).

<table>
<thead>
<tr>
<th>Assumptions</th>
<th>( \ell )-grad-Lip</th>
<th>( \ell )-grad-Lip + ( \rho )-Hessian-Lip</th>
</tr>
</thead>
<tbody>
<tr>
<td>Guarantees</td>
<td>( \epsilon )-FOSP</td>
<td>( \epsilon )-SOSP</td>
</tr>
<tr>
<td>Iterations</td>
<td>( 2\ell(f(x_0) - f^*)/\epsilon^2 )</td>
<td>( \tilde{O}(\ell(f(x_0) - f^*)/\epsilon^2) )</td>
</tr>
</tbody>
</table>
**Challenge:** non-constant Hessian + large step size $\eta = O(1/\ell)$.

Around saddle point, **stuck region** forms a non-flat “pancake” shape.

**Key Observation:** although we don’t know its shape, we know it's thin! (Based on an analysis of two nearly coupled sequences)
How Fast Can We Go?

• Important role of lower bounds (Nemirovski & Yudin)
  – strip away inessential aspects of the problem to reveal fundamentals

• The acceleration phenomenon (Nesterov)
  – achieve the lower bounds
  – second-order dynamics
  – a conceptual mystery

• Our perspective: it’s essential to go to continuous time
  – the notion of ”acceleration” requires a continuum topology to support it
Part II: Variational, Hamiltonian and Symplectic Perspectives on Acceleration

with Andre Wibisono, Ashia Wilson and Michael Betancourt
Accelerated gradient descent

Setting: Unconstrained convex optimization

\[
\min_{x \in \mathbb{R}^d} f(x)
\]

- Classical gradient descent:

\[
x_{k+1} = x_k - \beta \nabla f(x_k)
\]

obtains a convergence rate of \(O(1/k)\)
Accelerated gradient descent

**Setting:** Unconstrained convex optimization

\[
\min_{x \in \mathbb{R}^d} f(x)
\]

- Classical gradient descent:

\[
x_{k+1} = x_k - \beta \nabla f(x_k)
\]

obtains a convergence rate of \(O(1/k)\)

- Accelerated gradient descent:

\[
y_{k+1} = x_k - \beta \nabla f(x_k)
\]

\[
x_{k+1} = (1 - \lambda_k) y_{k+1} + \lambda_k y_k
\]

obtains the (optimal) convergence rate of \(O(1/k^2)\)
Accelerated methods: Continuous time perspective

- Gradient descent is discretization of gradient flow

\[ \dot{X}_t = -\nabla f(X_t) \]

(and mirror descent is discretization of natural gradient flow)
Accelerated methods: Continuous time perspective

- Gradient descent is discretization of gradient flow

\[ \dot{X}_t = -\nabla f(X_t) \]

(and mirror descent is discretization of natural gradient flow)

- Su, Boyd, Candes ’14: Continuous time limit of accelerated gradient descent is a second-order ODE

\[ \ddot{X}_t + \frac{3}{t} \dot{X}_t + \nabla f(X_t) = 0 \]
Accelerated methods: Continuous time perspective

- Gradient descent is discretization of gradient flow
  \[ \dot{X}_t = -\nabla f(X_t) \]
  (and mirror descent is discretization of natural gradient flow)
- Su, Boyd, Candes ’14: Continuous time limit of accelerated gradient descent is a second-order ODE
  \[ \ddot{X}_t + \frac{3}{t} \dot{X}_t + \nabla f(X_t) = 0 \]

- These ODEs are obtained by taking continuous time limits. Is there a deeper generative mechanism?

Our work: A general variational approach to acceleration
A systematic discretization methodology
Bregman Lagrangian

\[ \mathcal{L}(x, \dot{x}, t) = e^{\gamma t + \alpha t} \left( D_h(x + e^{-\alpha t} \dot{x}, x) - e^{\beta t} f(x) \right) \]

Variational problem over curves:

\[ \min_X \int \mathcal{L}(X_t, \dot{X}_t, t) \, dt \]

Optimal curve is characterized by Euler-Lagrange equation:

\[ \frac{d}{dt} \left\{ \frac{\partial \mathcal{L}}{\partial \dot{X}}(X_t, \dot{X}_t, t) \right\} = \frac{\partial \mathcal{L}}{\partial X}(X_t, \dot{X}_t, t) \]

E-L equation for Bregman Lagrangian under ideal scaling:

\[ \dddot{X}_t + (e^{\alpha t} - \dot{\alpha}_t) \ddot{X}_t + e^{2\alpha t + \beta t} \left[ \nabla^2 h(X_t + e^{-\alpha t} \dot{X}_t) \right]^{-1} \nabla f(X_t) = 0 \]
Mysteries

- **Why** can’t we discretize the dynamics when we are using exponentially fast clocks?
- **What** happens when we arrive at a clock speed that we can discretize?
- **How** do we discretize once it’s possible?
Towards A Symplectic Perspective

- We’ve discussed discretization of Lagrangian-based dynamics
- Discretization of Lagrangian dynamics is often fragile and requires small step sizes
- We can build more robust solutions by taking a Legendre transform and considering a \textit{Hamiltonian} formalism:

\[
L(q, \nu, t) \rightarrow H(q, p, t, \mathcal{E})
\]

\[
\begin{pmatrix}
\frac{dq}{dt} & \frac{d\nu}{dt}
\end{pmatrix}
\rightarrow
\begin{pmatrix}
\frac{dq}{d\tau} & \frac{dp}{d\tau} & \frac{dt}{d\tau} & \frac{d\mathcal{E}}{d\tau}
\end{pmatrix}
\]
Symplectic Integration of Bregman Hamiltonian
Symplectic vs Nesterov

$p = 2, N = 2, C = 0.0625, \varepsilon = 0.1$
Symplectic vs Nesterov

\[p = 2, \; N = 2, \; C = 0.0625, \; \varepsilon = 0.25\]
Part III: Acceleration and Saddle Points

with Chi Jin and Praneeth Netrapalli
Hamiltonian Analysis

\[ f(\cdot) \text{ between } x_t \text{ and } x_t + v_t \]

\( f(x_t) + \frac{1}{2\eta} \|v_t\|^2 \text{ decreases} \)

\( \|v_t\| \text{ large} \)

\( v_{t+1} = 0 \)

\( \|v_t\| \text{ small} \)

Move in \( \pm v_t \) direction

Not too nonconvex

(Too nonconvex

(Negative curvature exploitation)

Do an amortized analysis

Enough decrease in a single step
PAGD Converges to SOSP Faster (Jin et al. 2017)

For $\ell$-gradient Lipschitz and $\rho$-Hessian Lipschitz function $f$, PAGD with proper choice of $\eta, \theta, r, T, \gamma, s$ w.h.p. finds $\epsilon$-SOSP in iterations:

$$\tilde{O}\left(\frac{\ell^{1/2} \rho^{1/4} (f(x_0) - f^*)}{\epsilon^{7/4}}\right)$$

### Comparison Table

<table>
<thead>
<tr>
<th></th>
<th>Strongly Convex</th>
<th>Nonconvex (SOSP)</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Assumptions</strong></td>
<td>$\ell$-grad-Lip &amp; $\alpha$-str-convex</td>
<td>$\ell$-grad-Lip &amp; $\rho$-Hessian-Lip</td>
</tr>
<tr>
<td>(Perturbed) GD</td>
<td>$\tilde{O}(\ell/\alpha)$</td>
<td>$\tilde{O}(\Delta_f \cdot \ell/\epsilon^2)$</td>
</tr>
<tr>
<td>(Perturbed) AGD</td>
<td>$\tilde{O}(\sqrt{\ell/\alpha})$</td>
<td>$\tilde{O}(\Delta_f \cdot \ell^{1/2} \rho^{1/4} / \epsilon^{7/4})$</td>
</tr>
<tr>
<td><strong>Condition $\kappa$</strong></td>
<td>$\ell/\alpha$</td>
<td>$\ell/\sqrt{\rho \epsilon}$</td>
</tr>
<tr>
<td><strong>Improvement</strong></td>
<td>$\sqrt{\kappa}$</td>
<td>$\sqrt{\kappa}$</td>
</tr>
</tbody>
</table>

**Convergence Result**

Impact of $\kappa^{14/14}$ Michael Jordan AGD Escape Saddle Points Faster than GD
Part IV: Acceleration and Stochastics

with Xiang Cheng, Niladri Chatterji and Peter Bartlett
Acceleration and Stochastics

• Can we accelerate diffusions?
• There have been negative results…
• …but they’ve focused on classical overdamped diffusions
Acceleration and Stochastics

• Can we accelerate diffusions?
• There have been negative results…
• …but they’ve focused on classical overdamped diffusions
• Inspired by our work on acceleration, can we accelerate underdamped diffusions?
Overdamped Langevin MCMC

Described by the Stochastic Differential Equation (SDE):
\[ dx_t = -\nabla U(x_t) dt + \sqrt{2} dB_t \]
where \( U(x) : \mathbb{R}^d \to \mathbb{R} \) and \( B_t \) is standard Brownian motion. The stationary distribution is \( p^*(x) \propto \exp(U(x)) \)

Corresponding Markov Chain Monte Carlo Algorithm (MCMC):
\[ \tilde{x}_{(k+1)\delta} = \tilde{x}_{k\delta} - \nabla U(\tilde{x}_{k\delta}) + \sqrt{2\delta} \xi_k \]
where \( \delta \) is the step-size and \( \xi_k \sim N(0, I_{d \times d}) \)
Guarantees under Convexity

Assuming $U(x)$ is $L$-smooth and $m$-strongly convex:

Dalalyan’14: Guarantees in Total Variation

If $n \geq 0 \left( \frac{d}{\epsilon^2} \right)$ then, $TV(p^{(n)}, p^*) \leq \epsilon$

Durmus & Moulines’16: Guarantees in 2-Wasserstein

If $n \geq 0 \left( \frac{d}{\epsilon^2} \right)$ then, $W_2(p^{(n)}, p^*) \leq \epsilon$

Cheng and Bartlett’17: Guarantees in KL divergence

If $n \geq 0 \left( \frac{d}{\epsilon^2} \right)$ then, $KL(p^{(n)}, p^*) \leq \epsilon$
**Underdamped Langevin Diffusion**

Described by the *second-order* equation:

\[
\begin{align*}
    dx_t &= v_t \, dt \\
    dv_t &= -\gamma v_t \, dt + \lambda \nabla U(x_t) \, dt + \sqrt{2\gamma\lambda} \, dB_t
\end{align*}
\]

The stationary distribution is \( p^*(x, v) \propto \exp \left( -U(x) - \frac{|v|^2}{2\lambda} \right) \)

Intuitively, \( x_t \) is the position and \( v_t \) is the velocity

\( \nabla U(x_t) \) is the force and \( \gamma \) is the drag coefficient
Quadratic Improvement

Let $p^{(n)}$ denote the distribution of $(\tilde{x}_{n\delta}, \tilde{v}_{n\delta})$. Assume $U(x)$ is strongly convex

Cheng, Chatterji, Bartlett, Jordan ’17:
If $n \geq O\left(\frac{\sqrt{d}}{\epsilon}\right)$ then $W_2(p^{(n)}, p^*) \leq \epsilon$

Compare with Durmus & Moulines ’16 (Overdamped)
If $n \geq O\left(\frac{d}{\epsilon^2}\right)$ then $W_2(p^{(n)}, p^*) \leq \epsilon$
Proof Idea: Reflection Coupling

Tricky to prove continuous-time process contracts. Consider two processes,

\[ dx_t = -\nabla U(x_t)dt + \sqrt{2} \, dB_t^x \]
\[ dy_t = -\nabla U(y_t)dt + \sqrt{2} \, dB_t^y \]

where \( x_0 \sim p_0 \) and \( y_0 \sim p^* \). Couple these through Brownian motion

\[ dB_t^y = \left[ I_{d \times d} - \frac{2 \cdot (x_t - y_t)(x_t - y_t)^T}{|x_t - y_t|_2^2} \right] dB_t^x \]

“reflection along line separating the two processes”
Reduction to One Dimension

By Itô’s Lemma we can monitor the evolution of the separation distance

\[ d|x_t - y_t|_2 = -\left(\frac{x_t - y_t}{|x_t - y_t|_2}, \nabla U(x_t) - \nabla U(y_t)\right) dt + 2\sqrt{2} dB^1_t \]

‘Drift’

’1-d random walk’

Two cases are possible

1. If \( |x_t - y_t|_2 \leq R \) then we have strong convexity; the drift helps.
2. If \( |x_t - y_t|_2 \geq R \) then the drift hurts us, but Brownian motion helps stick*

Rates not exponential in \( d \) as we have a 1-d random walk

*Under a clever choice of Lyapunov function.
Part VI: Acceleration and Sampling

With Yi-An Ma, Niladri Chatterji, and Xiang Cheng
Acceleration of SDEs

• The underdamped Langevin stochastic differential equation is Nesterov acceleration on the manifold of probability distributions, with respect to the KL divergence (Ma, et al., to appear)
Part V: Population Risk and Empirical Risk

with Chi Jin and Lydia Liu
Population Risk vs Empirical Risk

Well-behaved population risk \Rightarrow rough empirical risk

- Even when $R$ is smooth, $\hat{R}_n$ can be non-smooth and may even have many additional local minima (ReLU deep networks).
- Typically $\|R - \hat{R}_n\|_{\infty} \leq O(1/\sqrt{n})$ by empirical process results.

Can we find local min of $R$ given only access to the function value $\hat{R}_n$?
Our Contribution

Our answer: **Yes!** Our SGD approach finds $\epsilon-$SOSP of $F$ if $\nu \leq \epsilon^{1.5}/d$, which is optimal among all polynomial queries algorithms.

Complete characterization of error $\nu$ vs accuracy $\epsilon$ and dimension $d$. 
Part VII: Market Design Meets Gradient-Based Learning

with Lydia Liu, Horia Mania and Eric Mazumdar
Two Examples of Current Projects

• How to find saddle points in high dimensions?
  – not just any saddle points; we want to find the Nash equilibria (and only the Nash equilibria)

• Competitive bandits and two-way markets
  – how to find the “best action” when supervised training data is not available, when other agents are also searching for best actions, and when there is conflict (e.g., scarcity)
Executive Summary

• ML (AI) has come of age
• But it is far from being a solid engineering discipline that can yield robust, scalable solutions to modern data-analytic problems
• There are many hard problems involving uncertainty, inference, decision-making, robustness and scale that are far from being solved
  – not to mention economic, social and legal issues