Dynamical, Symplectic and Stochastic Perspectives on Gradient-Based Optimization

Michael I. Jordan
University of California, Berkeley

August 9, 2018
Computation and Statistics

• A Grand Challenge of our era: tradeoffs between statistical inference and computation
  – most data analysis problems have a time budget
  – and often they’re embedded in a control problem
• Optimization has provided the computational model for this effort (computer science, not so much)
  – it’s provided the algorithms and the insight
• On the other hand, modern large-scale statistics has posed new challenges for optimization
  – millions of variables, millions of terms, sampling issues, nonconvexity, need for confidence intervals, parallel/distributed platforms, etc
For the Journalists in the Audience

• A variety of buzzwords have been used for (aspects of) this general Grand Challenge:
  – machine learning, deep learning, AI, etc

• Most of these buzzwords are poor at capturing the general scope of the challenge
  – many aspects of the problem have little to do with imitating human intelligence, or understanding the brain
  – consider existing and future planetary-scale networked systems for medicine, commerce, transportation, etc
  – a better scoping is “algorithmic decision-making under uncertainty, in large-scale networks and markets”

• This demands a (mathematical) reunification of many separate threads, a return to and updating of the era of Kolmogorov, von Neumann, Wiener, Blackwell, Wald, etc
Computation and Statistics (cont)

• Modern large-scale statistics has posed new challenges for optimization
  – millions of variables, millions of terms, sampling issues, nonconvexity, need for confidence intervals, parallel/distributed platforms, etc

• Current algorithmic focus: what can we do with the following ingredients?
  – gradients
  – stochastics
  – acceleration

• Current theoretical focus: placing lower bounds from statistics and optimization in contact with each other
Outline

• Escaping saddle points efficiently
• Variational, Hamiltonian and symplectic perspectives on Nesterov acceleration
• Acceleration and saddle points
• Acceleration and Langevin diffusions
• Optimization and empirical processes
Part I: How to Escape Saddle Points Efficiently

with Chi Jin, Praneeth Netrapalli, Rong Ge, and Sham Kakade
Nonconvex Optimization and Statistics

- Many interesting statistical models yield nonconvex optimization problems (cf neural networks)
- Bad local minima used to be thought of as the main problem in fitting such models
- But in many convex problems there either are no local optima (provably), or stochastic gradient seems to have no trouble (eventually) finding global optima
- But saddle points abound in these architectures, and they cause the learning curve to flatten out, perhaps (nearly) indefinitely
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):

\[
x_{t+1} = x_t - \eta \nabla f(x_t).
\]
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.
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).
Some Well-Behaved Nonconvex Problems

- PCA, CCA, Matrix Factorization
- Orthogonal Tensor Decomposition (Ge, Huang, Jin, Yang, 2015)
- Complete Dictionary Learning (Sun et al, 2015)
- Phase Retrieval (Sun et al, 2015)
- Matrix Sensing (Bhojanapalli et al, 2016; Park et al, 2016)
- Symmetric Matrix Completion (Ge et al, 2016)
- Matrix Sensing/Completion, Robust PCA (Ge, Jin, Zheng, 2017)

- The problems have no spurious local minima and all saddle points are strict
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

$$\|\nabla f(x)\| \leq \epsilon$$
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.*
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. $x_t \leftarrow x_t + \xi_t, \quad \xi_t$ 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).*
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>GD (Nesterov 1998)</th>
<th>PGD (This Work)</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Guarantees</strong></td>
<td>$\epsilon$-FOSP</td>
<td>$\epsilon$-SOSP</td>
</tr>
<tr>
<td><strong>Iterations</strong></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>
Geometry and Dynamics around Saddle Points

**Challenge:** non-constant Hessian $+ \text{ large step size } \eta = O(1/\ell)$.

Around saddle point, **stuck region** forms a non-flat “pancake” shape.
Geometry and Dynamics around Saddle Points

**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)
Next Questions

• Does acceleration help in escaping saddle points?
• What other kind of stochastic models can we use to escape saddle points?
• How do acceleration and stochastics interact?
Next Questions

• Does acceleration help in escaping saddle points?
• What other kind of stochastic models can we use to escape saddle points?
• How do acceleration and stochastics interact?

• To address these questions we need to understand develop a deeper understanding of acceleration than has been available in the literature to date
Part II: Variational, Hamiltonian and Symplectic Perspectives on Acceleration

with Andre Wibisono, Ashia Wilson and Michael Betancourt
Interplay between Differentiation and Integration

• The 300-yr-old fields: Physics, Statistics
  – cf. Lagrange/Hamilton, Laplace expansions, saddlepoint expansions

• The numerical disciplines
  – e.g., finite elements, Monte Carlo
Interplay between Differentiation and Integration

• The 300-yr-old fields: Physics, Statistics
  – cf. Lagrange/Hamilton, Laplace expansions, saddlepoint expansions

• The numerical disciplines
  – e.g., finite elements, Monte Carlo

• Optimization?
Interplay between Differentiation and Integration

• The 300-yr-old fields: Physics, Statistics
  – cf. Lagrange/Hamilton, Laplace expansions, saddlepoint expansions

• The numerical disciplines
  – e.g., finite elements, Monte Carlo

• Optimization?
  – to date, almost entirely focused on differentiation
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:
  \[
y_{k+1} = x_k - \beta \nabla f(x_k)
y_{k+1} = (1 - \lambda_k)y_{k+1} + \lambda_k y_k
\]
  obtains the (optimal) convergence rate of \(O(1/k^2)\)
The acceleration phenomenon

Two classes of algorithms:

- **Gradient methods**
  - Gradient descent, mirror descent, cubic-regularized Newton’s method (Nesterov and Polyak ’06), etc.
  - Greedy descent methods, relatively well-understood
The acceleration phenomenon

Two classes of algorithms:

- **Gradient methods**
  - Gradient descent, mirror descent, cubic-regularized Newton’s method (Nesterov and Polyak ’06), etc.
  - Greedy descent methods, relatively well-understood

- **Accelerated methods**
  - Nesterov’s accelerated gradient descent, accelerated mirror descent, accelerated cubic-regularized Newton’s method (Nesterov ’08), etc.
  - Important for both theory (optimal rate for first-order methods) and practice (many extensions: FISTA, stochastic setting, etc.)
  - *Not* descent methods, faster than gradient methods, still mysterious
Accelerated methods

- Analysis using Nesterov’s estimate sequence technique
- Common interpretation as “momentum methods” (Euclidean case)
- Many proposed explanations:
  - Chebyshev polynomial (Hardt ’13)
  - Linear coupling (Allen-Zhu, Orecchia ’14)
  - Optimized first-order method (Drori, Teboulle ’14; Kim, Fessler ’15)
  - Geometric shrinking (Bubeck, Lee, Singh ’15)
  - Universal catalyst (Lin, Mairal, Harchaoui ’15)
  - …

But only for strongly convex functions, or first-order methods

**Question:** What is the underlying mechanism that generates acceleration (including for higher-order methods)?
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, Candès ’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
Define the **Bregman Lagrangian**:

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

- Function of position \(x\), velocity \(\dot{x}\), and time \(t\)
- \(D_h(y, x) = h(y) - h(x) - \langle \nabla h(x), y - x \rangle\) is the Bregman divergence
- \(h\) is the convex distance-generating function
- \(f\) is the convex objective function

Ideal scaling conditions:

\[\dot{\beta}_t \leq e^{\alpha t} \quad \dot{\gamma}_t = e^{\alpha t}\]
Define the **Bregman Lagrangian**:

\[
\mathcal{L}(x, \dot{x}, t) = e^{\gamma t - \alpha t} \left( \frac{1}{2} \| \dot{x} \|^2 - e^{2\alpha t + \beta t} f(x) \right)
\]

- Function of position \( x \), velocity \( \dot{x} \), and time \( t \)
- \( D_h(y, x) = h(y) - h(x) - \langle \nabla h(x), y - x \rangle \) is the Bregman divergence
- \( h \) is the convex distance-generating function
- \( f \) is the convex objective function
- \( \alpha_t, \beta_t, \gamma_t \in \mathbb{R} \) are arbitrary smooth functions
- In Euclidean setting, simplifies to damped Lagrangian
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:

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

Optimal curve is characterized by **Euler-Lagrange** equation:

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

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

Optimal curve is characterized by \textbf{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:

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

Theorem

*Theorem* Under ideal scaling, the E-L equation has convergence rate

\[ f(X_t) - f(x^*) \leq O(e^{-\beta t}) \]

**Proof.** Exhibit a Lyapunov function for the dynamics:

\[
\mathcal{E}_t = D_h \left( x^*, X_t + e^{-\alpha t} \dot{X}_t \right) + e^{\beta t} (f(X_t) - f(x^*))
\]

\[
\dot{\mathcal{E}}_t = -e^{\alpha t + \beta t} D_f(x^*, X_t) + (\beta_t - e^{\alpha t}) e^{\beta t} (f(X_t) - f(x^*)) \leq 0
\]

**Note:** Only requires convexity and differentiability of \( f, h \)
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?
Symplectic Integration

- Consider discretizing a system of differential equations obtained from physical principles.
- Solutions of the differential equations generally conserve various quantities (energy, momentum, volumes in phase space).
- Is it possible to find discretizations whose solutions exactly conserve these same quantities?
- Yes!
  - from a long line of research initiated by Jacobi, Hamilton, Poincare’ and others.
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 Hamiltonian formalism:

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

\[
\left( \frac{dq}{dt}, \frac{dv}{dt} \right) \rightarrow \left( \frac{dq}{d\tau}, \frac{dp}{d\tau}, \frac{dt}{d\tau}, \frac{d\mathcal{E}}{d\tau} \right)
\]
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
Problem Setup

**Smooth Assumption:** \( f(\cdot) \) is smooth:

- \( \ell \)-gradient Lipschitz, i.e.
  \[
  \forall x_1, x_2, \| \nabla f(x_1) - \nabla f(x_2) \| \leq \ell \| x_1 - x_2 \|.
  \]

- \( \rho \)-Hessian Lipschitz, i.e.
  \[
  \forall x_1, x_2, \| \nabla^2 f(x_1) - \nabla^2 f(x_2) \| \leq \rho \| x_1 - x_2 \|.
  \]

**Goal:** find **second-order stationary point** (SOSP):
\[
\nabla f(x) = 0, \quad \lambda_{\min}(\nabla^2 f(x)) \geq 0.
\]

Relaxed version: \( \epsilon \)-**second-order stationary point** (\( \epsilon \)-SOSP):
\[
\| \nabla f(x) \| \leq \epsilon, \quad \text{and} \quad \lambda_{\min}(\nabla^2 f(x)) \geq -\sqrt{\rho \epsilon}
\]
Analysis of AGD in the Nonconvex Setting

- **Challenge**: AGD is not a descent algorithm
- **Solution**: Lift the problem to a phase space, and make use of a Hamiltonian
- **Consequence**: AGD is nearly a descent algorithm in the Hamiltonian, with a simple “negative curvature exploitation” (NCE; cf. Carmon et al., 2017) step handling the case when descent isn’t guaranteed
Hamiltonian Perspective on AGD

• AGD is a discretization of the following ODE

\[ \ddot{x} + \tilde{\theta} \dot{x} + \nabla f(x) = 0 \]

• Multiplying by \( \dot{x} \) and integrating from \( t_1 \) to \( t_2 \) gives us

\[ f(x_{t_2}) + \frac{1}{2} \| \dot{x}_{t_2} \|^2 = f(x_{t_1}) + \frac{1}{2} \| \dot{x}_{t_1} \|^2 - \tilde{\theta} \int_{t_1}^{t_2} \| \dot{x}_t \|^2 dt \]

• In convex case, Hamiltonian \( f(x_t) + \frac{1}{2} \| \dot{x}_t \|^2 \) decreases monotonically
Algorithm Perturbed Accelerated Gradient Descent (PAGD)

1. for $t = 0, 1, \ldots$ do
2. if $\|\nabla f(x_t)\| \leq \epsilon$ and no perturbation in last $T$ steps then
3. $x_t \leftarrow x_t + \xi_t, \quad \xi_t$ uniformly $\sim B_0(r)$
4. $y_t \leftarrow x_t + (1 - \theta)v_t$
5. $x_{t+1} \leftarrow y_t - \eta \nabla f(y_t); \quad v_{t+1} \leftarrow x_{t+1} - x_t$
6. if $f(x_t) \leq f(y_t) + \langle \nabla f(y_t), x_t - y_t \rangle - \frac{\gamma}{2} \|x_t - y_t\|^2$ then
7. $x_{t+1} \leftarrow \text{NCE}(x_t, v_t, s); \quad v_{t+1} \leftarrow 0$

- Perturbation (line 2-3);
- Standard AGD (line 4-5);
- Negative Curvature Exploitation (NCE, line 6-7)
  - 1) simple (two steps), 2) auxiliary. [inspired by Carmon et al. 2017]
Hamiltonian Analysis

$f(\cdot)$ between $x_t$ and $x_t + \nu_t$

Not too nonconvex

Too nonconvex (Negative curvature exploitation)

AGD step

$\|\nu_t\|$ large

$\nu_{t+1} = 0$

$f(x_t) + \frac{1}{2\eta} \|\nu_t\|^2$ decreases

$\|\nu_t\|$ small

Move in $\pm \nu_t$ direction

Do an amortized analysis

Enough decrease in a single step
**Convergence Result**

**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 $\varepsilon$-SOSP in iterations:

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

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

**Strongly Convex**

- $\tilde{O}(\ell/\alpha)$
- $\tilde{O}(\sqrt{\ell/\alpha})$

**Nonconvex (SOSP)**

- $\tilde{O}(\Delta_f \cdot \ell/\varepsilon^2)$
- $\tilde{O}(\Delta_f \cdot \ell^{1/2} \rho^{1/4}/\varepsilon^{7/4})$

**Improvement**

- $\ell/\sqrt{\rho \varepsilon}$
- $\sqrt{\kappa}$
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...
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 $n \geq O\left(\frac{d}{\epsilon^2}\right)$ then, $TV(p^{(n)}, p^*) \leq \epsilon$

Durmus & Moulines’16: Guarantees in 2-Wasserstein

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

Cheng and Bartlett’17: Guarantees in KL divergence

If $n \geq O\left(\frac{d}{\epsilon^2}\right)$ 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
Discretization

We can discretize; and at each step evolve according to

\[ d\tilde{x}_t = \tilde{v}_t dt \]
\[ d\tilde{v}_t = -\gamma \tilde{v}_t dt - \lambda \nabla U(\tilde{x}_{t/\delta}) dt + \sqrt{2\gamma\lambda} dB_t \]

we evolve this for time \( \delta \) to get an MCMC algorithm

Notice this is a second-order method. Can we get faster rates?
Quadratic Improvement

Let \( p^{(n)} \) denote the distribution of \((\tilde{x}_{n\delta}, \tilde{\nu}_{n\delta})\). 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 \( W_2(p^{(n)}, p^*) \leq \epsilon \)
Intuition: Smoother Sample Paths

$x_t$ is much smoother for Underdamped Langevin Diffusion, so easier to discretize.

Overdamped Langevin Diffusion  
Underdamped Langevin Diffusion
Beyond Convexity?

So far we assume $U(x)$ is $m$-strongly convex

Goal: Establish rates when $U(x)$ is non-convex

Multiple modes
Strongly Convex Outside a Ball

1. Smooth everywhere
\[ \forall x, y \ |\nabla U(x) - \nabla U(y)|_2 \leq L|x - y|_2 \]

2. Strongly convex outside a ball \( \forall x, y: |x - y|_2 \geq R \)
\[ \langle \nabla U(x) - \nabla U(y), x - y \rangle \geq m|x - y|_2 \]

Cheng, Chatterji, Abbasi-Yakdori, Bartlett, & Jordan ’18:

To get \( W_1(p^{(n)}, p^*) \leq \epsilon \):

- Overdamped MCMC: \( n \geq O\left(\frac{e^{cLR^2d}}{\epsilon^2}\right) \)
- Underdamped MCMC needs: \( n \geq O\left(\frac{e^{cLR^2\sqrt{d}}}{\epsilon}\right) \)
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

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

‘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.
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.
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**.
Algorithm Zero-th order Perturbed SGD (ZPSGD)

1. for $t = 0, 1, \ldots$ do
2. sample $(z^{(1)}_t, \cdots, z^{(m)}_t) \sim \mathcal{N}(0, \sigma^2 I)$
3. $g_t(x_t) \leftarrow \sum_{i=1}^{m} z^{(i)}_t [f(x_t + z^{(i)}_t) - f(x_t)]/(m\sigma^2)$
4. $x_{t+1} \leftarrow x_t - \eta (g_t(x_t) + \xi_t)$, $\xi_t$ uniformly $\sim \mathbb{B}_0(r)$

- Computing stochastic gradient of smoothed function (line 2-3);

$$\tilde{f}_\sigma(x) = \mathbb{E}_{z \sim \mathcal{N}(0, \sigma^2 I)}[f(x + z)]$$
$$\nabla \tilde{f}_\sigma(x) = \mathbb{E}_{z \sim \mathcal{N}(0, \sigma^2 I)}[z(f(x + z) - f(x))]/\sigma^2$$

- Perturbation (line 4).
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 $d$. 
Discussion

• **Data** and **inferential problems** will be everywhere in computer science, and will fundamentally change the field

• Many **conceptual** and **mathematical** challenges arising in taking this effort seriously, in addition to systems challenges and “outreach” challenges

• Facing these challenges will require a rapprochement between **computational thinking** and **inferential thinking**

• This effort is just beginning!
Reference