Differential Equation Modeling For Optimization

Yiping Lu
School of mathematical sciences
Peking University
Beijing, China
luyiping9712@pku.edu.cn

Abstract

Optimization, an area for finding the minimization point of an objective function, is becoming a hot topic. In this technical report, we bridge several famous optimization methods as numerical scheme approximation to different dynamic, such as gradient flow. We discover that there is a second order differential equation behind the accelerated optimization methods. This framework is very wide, can include accelerated mirror descent, gradient flow in Sobolev space and etc. We also extend the acceleration to stochastic gradient and synchronous parallel iteration. We utilize stochastic differential equation and differential equation with time delay to analysis the two different setting. In this report, we will show that adding a stochastic and synchronous parallel term will have the same convergence as the original optimization methods.

1 Introduction

Optimization has becoming a more and more important task in various areas, like signal processing, machine learning and physic simulation. The analysis fo optimization algorithms are always from discrete time scheme, these years analyzing the optimzation methods from a continuous time limit becoming more and more popular. As a simple example to reveal the motivation, the famous gradient descent $x_k = x_{k-1} - \Delta t \nabla f(x_{k-1})$ will converge to the gradient flow $\dot{X} = -\nabla f$. From the continuous dynamic, several converge property can be easily observed. This report will focus on the differential equation limit of the optimization methods. In order to observe the converge rate, Lyapunov analysis is always the core, which means to construct a decreasing energy function.

First using the differential equation to model optimization methods is [1]. Using a Hamilton equation with friction to model momentum methods. [2] also discover the limit when the step size go down to zero is an ODE. Applying Lyapunov analysis, they give a new converge proof. Generalizing [2], [6] give an variational perspective of the first-order acceleration methods and give first-order methods with polynomial convergence of any order. Moreover, [15] gives a stochastic version of the acceleration method.

Stochastic gradient descent is becoming more and more concerned due to the explosive growth number of data. It's impossible to calculated the full gradient of the full

Homepage: www.2prime.cn and about.2prime.cn (2017).

dataset, instead, every time a subset is sampled and descent according to the partial calculated gradient. We will go on the stochastic algorithm in Section3. We will bridge it with MCMC method and using stochastic differential equation to analysis the algorithms. Moreover, in this section, we will discover more about its benefit on machine learning tasks.

In order to accelerating the algorithm further, people want to reduce the communication and waiting time in parallel computing. As a result, asynchronous methods is attracting more attention. The agents calculates its own part and updates the parameter without communication. In section4 we will using time-delay dynamic to characterize this method and find it the same converge rate with the original one which is known as "more iterations per second, same quality".

In short, optimization method can be consider as numerical approximation of continuous gradient flow like dynamic. From continuous dynamic, the behavior of the optimization methods can be easily observed via Lyapunov analysis.

2 Differential Equation Modelling For Momentum Methods

[2] gives a differential modeling for momentum methods. From the differential equation's view, the author claim that the oscillation for Nesterov's method is due to the oscillating trajectory of the dynamic system the due to the fraction term is fast decreasing. Thus, the authors gives a new speed restarting method to ensure the loss will decrease and achieves better results.[12] extend Nesterov's acceleration to mirror descent case and [13] gives a variational generalization via Bregman Divergence. [14] has extend this framework Sobolev gradients for more robust coarse-to-fine PDE-based optimization problems.

2.1 Momentum Methods: Dynamic View

First we derivate the Nesterov momentum methods[19] as a approximation of a second order differential equation. First we rewrite the Nesterov methods as

$$\frac{x_{k+1}-x_k}{\sqrt{s}} = \frac{k-1}{k+2}\frac{x_k - x_{k-1}}{\sqrt{s}} - \sqrt{s}\nabla f(x_k)$$

Introduce the Ansatz $x_k \simeq X(k\sqrt{s})$, the Taylor expansion can be formulated as,

$$\frac{x_{k+1}-x_k}{\sqrt{s}} = \dot{X}(t) + \frac{\sqrt{s}}{2}\ddot{X}(t) + o(\sqrt{s})$$
$$\frac{x_k - x_{k-1}}{\sqrt{s}} = \dot{X}(t) - \frac{\sqrt{s}}{2}\ddot{X}(t) + o(\sqrt{s})$$

Thus we have

$$\dot{X}(t) + \frac{\sqrt{s}}{2}\ddot{X}(t) + o(\sqrt{s}) = \frac{k-1}{k+2}(\dot{X}(t) - \frac{\sqrt{s}}{2}\ddot{X}(t) + o(\sqrt{s})) - \sqrt{s}\nabla f(x_k)$$

By comparing the coefficients of $\sqrt{s}$, we get the Nesterov ODE

$$\ddot{X(t)} + \frac{3}{t}\dot{X} + \nabla f(X) = 0$$

This ODE have the several properties:

- **Time Invariance.** After apply a linear time transform $t' = ct$, the ODE becomes $\ddot{X} + 3/t\dot{X} + \nabla f(X)/c^2 = 0$ For minimizing f/c^2 is equivalent to minimizing f, the ODE is invariant under the time change. It's obvious that the time invariance will hold only if the equation is a **Euler-type Equation**. We will generalize the property about this in the following section.
- **Rotational Invariance.**
- **Initial Asymptotic.** Here we suppose $\lim_{t\to 0} \ddot{X}$ exists and the initial condition $\dot{X}(0) = 0$. Then we have $\lim_{t\to 0} X(t)/t = \frac{X(t)-X(0)}{t} = \ddot{X}(0)$ which means $\ddot{X}(t) \approx -\nabla f(x_0)/4$. So the trajectory at the initial time will have asymptotic behavior

  $$X(t) = -\frac{\nabla f(x_0)t^2}{8} + x_0 + o(t^2)$$

  This asymptotic expansion demonstrates that Nesterov's methods performs slowly in the beginning.

2.2 Oscillating

To give heuristic analysis, we assume the objective is a quadratic function $f(x) = \sum_{i=1}^{n} \lambda_i x_i^2$, the solution of the differential equation is the Bessel function, which has an analytic expansion as

$$J_i(u) = \sum_{m=0}^{\infty} \frac{(-1)^m}{(2m)!!(2m+2)!!} u^{2m+1}$$

For large t, the Bessel function has the following asymptotic form

$$J_1(t) = \sqrt{\frac{2}{\pi t}}(\cos(t - 3\pi/4) + O(1/t))$$

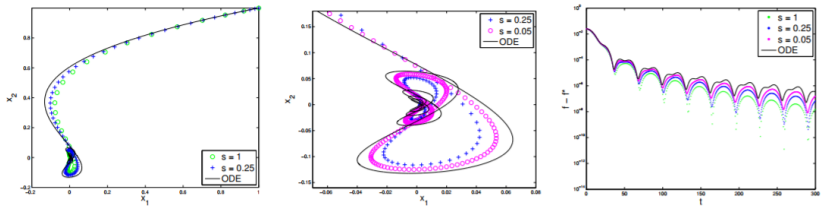
From the above expansion, it is easy to observe the $O(1/k^2)$ converge rate.

Now we will utilize the analytic expansion to show the to true solution of the dynamic

The figure shows that the solution is oscillating. [2] believes that the oscillating is come from the oscillating dynamic, but in the next section, I will give an example that the oscillating comes from the numerical scheme.

2.3 Stiff Equation

First we give an introduction to the stiff equation. *In mathematics, a stiff equation is a differential equation for which certain numerical methods for solving the equation*



(a) Trajectories. (b) Zoomed trajectories. (c) Errors $f - f^\star$.

Figure 1: Minimizing $f = 2 \times 10^{-2} x_1^2 + 5 \times 10^{-3} x_2^2$, starting from $x_0 = (1, 1)$. The black and solid curves correspond to the solution to the ODE.

are numerically unstable, unless the step size is taken to be extremely small. It has proven difficult to formulate a precise definition of stiffness, but the main idea is that the equation includes some terms that can lead to rapid variation in the solution.

When integrating a differential equation numerically, one would expect the requisite step size to be relatively small in a region where the solution curve displays much variation and to be relatively large where the solution curve straightens out to approach a line with slope nearly zero. For some problems this is not the case. Sometimes the step size is forced down to an unacceptably small level in a region where the solution curve is very smooth. The phenomenon being exhibited here is known as stiffness. In some cases we may have two different problems with the same solution, yet problem one is not stiff and problem two is stiff. Clearly the phenomenon cannot be a property of the exact solution, since this is the same for both problems, and must be a property of the differential system itself. It is thus appropriate to speak of stiff systems.(from wiki)

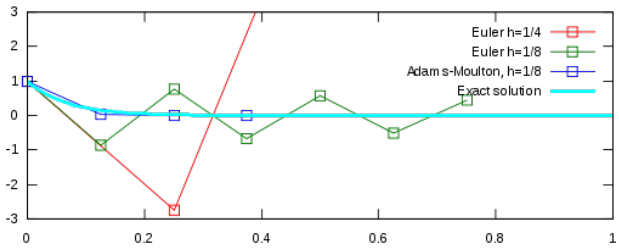


Figure 2: Stiff equation is a set of equations with large condition numbers, the numerical scheme of the dynamic system is oscillating.

In order to justify where the oscillating comes from, we utilize the Lasso problem

$$\min ||Ax - b||_2^2 + \lambda ||x||_1$$

to be the test problem. Instead of decaying the step size, we reparameterize the time variable as

$$\frac{x_{k+1} - x_k}{\sqrt{s}} = \frac{k \cdot rate - 1}{k \cdot rate + 2} \frac{x_k - x_{k-1}}{\sqrt{s}} - \sqrt{s} \nabla f(y_k)$$

Revise the Ansatz $x_k \simeq X(k \cdot rate\sqrt{s})$, the equation will also converge to the Nesterov ODE, like the test in the original paper[2], we also test the high friction

case. In Figure3, it shows that it doesn't appears the oscillating while the step size becomes very small. So it is not surprising that the algorithm becomes faster will the re-parameter rate$= 0.1$

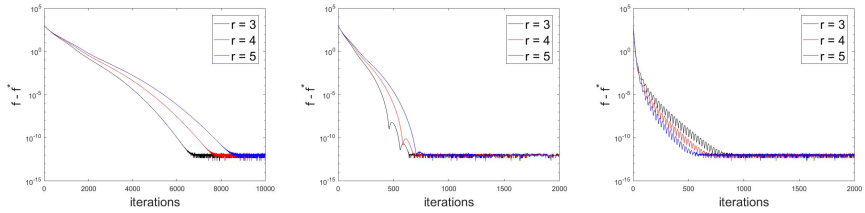


Figure 3: Fig(1) reparameter rate 0.01(2) reparameter rat 0.1 (3) original reparameter rate.

2.4 Second Order Momentum Dynamic

[9] introduced the notation of second order differential equations with asymptotically small dissipation, which can be concluded by

$$\ddot{X}(t) + \alpha(t)\dot{X}(t) + \nabla f(t) = 0$$

First, we will point out that this is a dynamic system not only comes from Nesterov acceleration but also conclude many examples in optimization.

- **Averaged gradient system:** $\dot{z}(s) + \frac{1}{s}\int_0^s g(z(\tau))d\tau = 0$
  After multiplying equation by s and then differentiating, we have
  $$s\ddot{z} + \dot{z} + \nabla f(x) = 0$$
  Instead by $t = 2\sqrt{s}$, we have
  $$\ddot{X} + \frac{1}{t}\dot{X} + \nabla f(X) = 0$$
- **Heavy Ball with Friction system:** If we consider a heavy ball with unit quality and we give a friction which is proportional to the velocity, the dynamic will take $a(t) = \gamma$.
- The fundamental solution of the semi-linear elliptic equation
  $$\Delta u(y) + g(u(y)) = 0$$
  leads to solving the ode
  $$\ddot{x}(r) + \frac{m-1}{r}\dot{x}(r) + g(x(r)) = 0$$
- If we give an approximation of the stochastic gradient system
  $$x^{n+1} = x^n - \epsilon^n g(x^n, \omega^n) + \epsilon^n \eta^n$$
  here $g(x^n, \omega^n)$ means that every calculation of the gradient will bring a noise and η^n is the noise injected by human.
  Now we give an approximation as
  $$x^{n+1} = x^n - \epsilon^n \frac{\sum_{i=1}^n \epsilon_i g(X^i, \omega^{i+1})}{\sum_{i=1}^n \epsilon_i}$$

As the analysis above in averaged gradient system, the dynamic can be consider as

$$\ddot{X}(t) = -\frac{\dot{X}(t) + g(X(t))}{t+\beta}$$

For the second order differential equations with asymptotically small dissipation, we can denote a Lyapunov energy function as

$$E(t) = f(X(t)) + \frac{1}{2}|\dot{X}(t)|$$

From Lyapunov point of view, it is easy to observe

Theorem 2.1. *Let $a : \mathbb{R}_+ \to \mathbb{R}_+$ be a non increasing map such that $\int_0^\infty a(s)ds = \infty$, let $G : H \to \mathbb{R}$ be a coercive function of class C^1 such that ∇G is Lipschitz continuous function on the bounded set of H. Then any solution x to the differential equation $\ddot{X(t)} + a(t)\dot{X(t)} + \nabla f(X) = 0$ satisfies for every $T > 0$,*

$$\liminf_{t\to\infty} \sup_{s\in[t,t+T]} |\dot{x}(s)| = 0$$

As the same prove of applying a subsequence converge in the weak topology, we can also prove the converge of the trajectory at the case of existing a unique minimum. For other case this will no longer be true again. As an example, take the objective function as the constant. The solution can be solved as $x(t) = x(0) + \dot{x}(0)e^{\int_0^s a(s)ds}$. If we need to ensure that solution will converge if and only if

$$\int_0^s a(s)ds < \infty$$

[9] ensures that the trajectory will converge under the assumption $\int_0^s a(s)ds < \infty$.

2.5 Bregman Hamiltonian Variational Framework

In this section, we will generalize the result of the previous section. We will show that, the trajectory of the accelerated optimization ODE will also follows a type of principle of minimum action. Equally, the trajectory will become a solution of a variational problem. From this point of view, optimization algorithms with converge rate of any order polynomial can be created.

First we give the notation about Bregman divergence, a divergence $D[P : Q]$ is a function of its coordinate ξ_P and ξ_Q satisfies certain criteria. We may also express as $D[\xi_P : \xi_Q]$. $D[P : Q]$ is called a divergence when it satisfies the following criteria:

- $D[P : Q] \geq 0$
- $D[P : Q] = 0$ if and only if $P = Q$
- $D[\xi_P : \xi_P + d\xi] = \frac{1}{2}\sum g_{ij}(\xi_p)d\xi_i d\xi_j + O(|d\xi|^3)$

Here $G = (g_{ij})$ is a positive-definite matrix depending on ξ_p and Bregman divergence of convex function ϕ is defined as

$$D_\psi[\xi : \xi_0] = \psi(\xi) - \psi(\xi_0) - \nabla\psi(\xi_0) \cdot (\xi - \xi_0)$$

In [13], they consider the dynamic following under Bregman Lagrangian

$$L(X, Y, V) = e^{\alpha(t)+\gamma(t)}[D_\phi(X + e^{-a(t)}V, X) - e^{b(t)}U(X)]$$

the potential energy U represents the objective function we need to minimize, thus the function f. In the Euclidean case, the Lagrangian is

$$L(X, Y, V) = e^{\alpha(t)+\gamma(t)}[e^{-a(t)}\frac{1}{2}||V||_2^2 - e^{b(t)}U(X)]$$

where can be consider as the kinetic energy adds the potential energy. The trajectory of the optimal solution under the Bregman Lagrangian can be written as

$$\ddot{X}_t + (e^{\alpha_t} - \dot{\alpha}_t)\dot{X}_t + e^{2\alpha_t+\beta_t}[\nabla\psi(X_t + e^{-\alpha_t}\dot{X}_t)]^{-1}\nabla f(X_t) = 0$$

when ψ is strong convex the Hessian matrix $\nabla^2\psi$ is invertible

Theorem 2.2. *If the the parameter satisfies* ***ideal scaling conditions*** *which means*

$$\dot{\beta}_t \leq e^{\alpha_t}$$
$$\dot{\gamma}_t = e^{\alpha_t}$$

Then the converge rate satisfies

$$f(X_t) - f(x^*) \leq O(e^{\beta_t})$$

The prove of the theorem is obvious by utilizing the non-increasing energy functional

$$E_t = D_h(x^*, X_t + e^{-\alpha_t}\dot{X}_t) + e^{\beta_t}(f(X_t) - f(x^*)$$

Time dilation. What surprising is that *A notable property of the Bregman Lagrangian family is that it is closed under time dilation.*([6]) As a simple example, $\dot{X} = -\nabla f(X)$ is an $O(1/k)$ converge rate algorithm[1], then we make a time dilation as $Y(t) = X(t^2)$. Then Y satisfies the dynamic as $\dot{Y}(t) = -\frac{1}{t}\nabla f(Y(t))$, then the converge rate will become $O(1/k^2)$.

Theorem 2.3. *For the re-parameterized curve $Y_t = X_{\tau(t)}$, we only need to adjust the parameter as*

$$\hat{\alpha}_t = \alpha_{\tau(t)} + \log\dot{\tau}(t)$$
$$\hat{\beta}_t = \beta_{\tau(t)}$$
$$\hat{\gamma}_t = \gamma_{\tau(t)}$$

Furthermore, if α, β, γ satisfy the ideal scaling then $\hat{\alpha}, \hat{\beta}, \hat{\gamma}$ also satisfies the ideal scaling.

As the [6] says, *"the time-dilation property means that the entire family of curves for accelerated methods in continuous time corresponds to a single curve in space*

time, which is traveled at different speeds. This suggests that the underlying solution curve has a more fundamental structure that is worth exploring further."

Nesterov's methods' trajectory can be consider as the solution of the variational problem when we take the time parameter as

$$a = \log \frac{k}{t}, b = k \log t + \log \lambda, \gamma = k \log t$$

[6] also provides two ways to discretization, the core process is to rewrite the second-order equation as

$$Z_t = X_t + \frac{t}{p}\dot{X}_t$$
$$\frac{d}{dt}\nabla h(Z_t) = -Cpt^{p-1}\nabla f(X_t)$$

Instead of the naive discretization, [6] also introduce a rate-matching discretization as the Nesterov's constructions of accelerated mirror descent in [12], the algorithm is demonstrated in Algorithm1.

Algorithm 1 A rate-matching discretization for Bregman Lagrangian Variational problem

Output: An estimate of the minima.

1: **for** Not Converge **do**

$$x_{k+1} = \frac{p}{k+p}z_k + \frac{k}{k+p}y_k$$
$$z_k = \arg\min_z \left\{ Cpk^{(p-1)} \langle \nabla f(y_k), z \rangle + \frac{1}{\epsilon} D_h(z, z_{k-1}) \right\}$$

2: **return** x_k

3 Stochastic Case

Stochastic gradient descent is a method that using gradient which is evaluated at a subset of the dataset, which can have faster calculate per iteration. Based on the law of large number, some people assume the noise which is caused by the subsampled calculated gradient is a Gauissan distribution. To make the theorical analysis easier, people begin to consider the algorithm that injecting a Gaussian noise every time, which is called SGLD[23]. Also people find the noise can also have more benefits, like finding a global minima and avoiding overfitting. In this section we will bridge the stochastic method with MCMC and Bayesian methods to show the generalization result of the algorithms.

3.1 Bridge The MCMC

3.1.1 Stochastic As MCMC In Bayesian Models

Here we will utilize the stochastic gradient descent to give an intuition to bridge stochastic optimization methods with Bayesian methods. First we consider the following algorithm called stochastic gradient Langevin methods

Algorithm 2 SGLD(stochastic gradient Langevin dynamic)

Output: An estimate of the minima.

1: **for** Not Converge **do**

$$x_{k+1} = x_k - \eta \nabla \hat{f}(x_k) + \sqrt{\eta} N(0; I)$$

2: **return** x_k

SGLD is a iterative methods which can be consider as an approximation to the Langevin dynamic

$$\partial_t X_t = -\nabla f(X_t) + dW_t$$

for $\eta << \sqrt{\eta}$ when $\eta \to 0$, the noise in the gradient will be cover by the noise from the Bownian Motion. So there is theoretical guarantee that the dynamic will converge. As a common knowledge that the steady state distribution is the well-known Gibbs distribution $\frac{1}{Z} exp(-f(x))$. Sampling a Gibbs distribution is a basic methods in Bayesian setting, for we also need to sample the posterior as $likelihood \times prior$.

The posterior distribution of a set of N data items $X = \{x_i\}$ is: $p(\theta|X) \propto p(\theta) \prod_{i=1}^{N} p(x_i|\theta)$ Now we consider the convergence active when the step size is really small the stochastic gradient Langevin dynamic which can be defined as Every update can be operated as

$$\Delta\theta_t = \frac{\epsilon_t}{2}(\nabla \log p(\theta_t) + \frac{N}{n} \sum_{i=1}^{n} \nabla \log p(x_{ti}|\theta_t)) + \eta_t$$

For step size, we need$\sum_t \epsilon_t = \infty, \sum_t \epsilon_t^2 < \infty$, every η_t is a Gaussian noise with covariance $\Delta_t I_n$

For $Var(\epsilon_t \frac{N}{n} \sum_{i=1}^{n} \nabla \log p(x_{ti}|\theta_t))) << Var(\eta_t)$, [1] proves that the dynamic system will converge to the Langevin dynamic when $\epsilon_t \to 0$. As a well known result in stochastic differential equation, the pdf will follows a Fokker-Planck equation

$$\frac{\partial \pi}{\partial t} = \Delta\pi + \nabla \cdot (\pi \nabla f)$$

To follow the Fokker-Planck equation, there is a well known theoretical result that the dynamic plays a steepest descent of free energy in the W_2 space.

Theorem 3.1. ***(JKO Theorem)[16]*** *Fokker-Planck equation plays the steepest descent in W_2 space on*

$$F(\pi) = \mathbb{E}_\pi[f(x)] - H(\pi)$$

Here W_2 is the space of probability space in L^2, where the metric is defined as

$$W_2(p_1, p_2) = \int |x - y|^2 d\pi(x, y)$$

Here we need $\int_y \pi(x, y)dy = p_1(x), \int_x \pi(x, y)dx = p_2(y)$

The distribution will converge to the Gibbs distribution $\frac{1}{Z}exp(-f(x))$.(which also can be proved to check the detailed balance condition.)

So that the Sgd can be consider as Bayesian sample method, which will bring us following benefits

- Avoid overfit(PAC-Bayesian bound is shown in [17]).
- Robust, which means "flat minima"(sample means we need to consider the probability in the neighborhood.)
- Implicit entropy regularizer.(which will be useful in reinforcement learning)

3.1.2 Accelerating The Dynamic

In this section, we will introduce the method in Section2 to stochastic case. We will utilizing second order stochastic differential equation with time decay friction to accelerate the dynamic.

First let us consider the motion of a particle on the line with friction proportional to the speed driven by the potential V is described by the ODE

$$m\frac{d^2x}{dt^2} = -\gamma\frac{dx}{dt} - \frac{\partial V(x)}{\partial x}$$

Let p denote $m\dot{x}$, then we get the system

$$\dot{x} = \frac{\partial H}{\partial p} - \gamma p$$
$$\dot{p} = -\frac{\partial H}{\partial x}$$

Here H is the total energy of the system $H(x,p) = \frac{p^2}{2m} + V(x)$, thus

$$\frac{d}{dt}H((x(t),p(t)) = -\gamma p^2 < 0$$

If we discretized the Hamilton equation, we can discover that the momentum gradient descent method:

$$x_{n+1} = cx_n + (1-c)x_{n-1} - \Delta t\nabla f(x)$$

is a numerical approximation to the Hamilton equation,

$$(2-c)\dot{X}_{n+1} + \frac{c}{2}\Delta t\ddot{X}_{n+1} = -\nabla f(x)$$

in second order($O(\Delta t^2)$). Moreover, we can discover that the momentum gradient descent method is a special case to the **linear multi-step** method to the original gradient flow $\dot{X} = -\nabla f$. The Hamiltion equation is the modified equation of the numerical scheme. This mention us that even the two method have the same limit can have different behavior if their modified equation is not the same. We want to emphasize that **modified equation may also affect.**

It is also possible to consider the algorithms that approximates the dynamic $\ddot{X} + 3/t\dot{X} + \nabla f + dW_t = 0$. In [15] they consider a accelerated dynamic under the Bregman divergence

$$dZ(t) = -\eta(t)[\nabla f(X(t))dt + \sigma(X(t),t)dB(t)]$$
$$dX(t) = a(t)[\nabla\phi^*(Z(t)/s(t)) - X(t)]dt$$

We consider the energy function defined above

$$L(x,z,t) = r(t)(f(x) - f(x^*)) + s(t)D_{\phi^*}(z(t)/s(t), z^*)$$

Then we have

Theorem 3.2. *Suppose that $a = \eta/r$. Under the case with out the noise term we will have*

$$\frac{d}{dt}L(x(t),z(t),t) \leq (f(x(t)) - f(x^*))(\dot{r}(t) - \eta(t)) + \psi(x^*)\dot{s}(t)$$

Under the stochastic case we have

$$dL(x(t),z(t),t) \leq [(f(x(t))-f(x^*))(\dot{r}(t)-\eta(t))+\psi(x^*)\dot{s}(t)+\frac{nL_{\psi^*}}{2}\cdot\frac{\eta(t)^2\sigma_*^2(t)}{s(t)}]dt+\langle V(t), dB(t)\rangle$$

Here $V(t) = -\eta(t)\sigma(X(t),t)^T(\nabla\psi^*(Z(t)/s(t)) - \nabla\psi^*(z^*))$

As is mentioned, the Hamiltonian System with friction will decrease the energy, we add a brownian motion to give chance to go to a position with higher energy, modify the equation to

$$\dot{x} = \frac{\partial H}{\partial p} - \gamma p + \eta dW_t$$
$$\dot{p} = -\frac{\partial H}{\partial x}$$

It's easy to observe that this method is equal to the Hamiltonian MC, for the equilibrium distribution is also a gibbs distribution. This means **momentum method** will converge to the same equilibrium distribution as the original gradient descent. [20] mentioned that there are three factors that affects the converged distribution(also can understand as noise level)

- The learning rate
- Batch size
- Momentum

[21] using increasing batch size in order to give a simulated annealing and gives the method "don't decay learning rate, increase the batch size". [22] using a approximate control to choose the best learnig rate.

In fact, first oder optimization methods can be separated into two classes:

- Polyak's heavy-ball method(HB)

  $$\omega_{k+1} = \omega_k - \alpha_k\nabla f(\omega_k + \gamma_k(\omega_k - \omega_{k-1})) + \beta_k(\omega_k - \omega_{k-1})$$

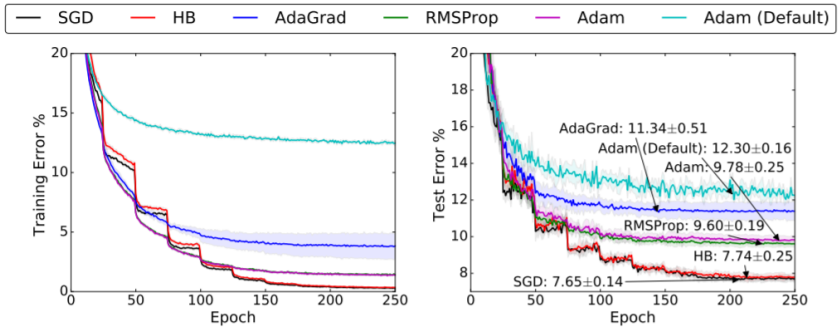
  Example: SG,MSG,NAG

- Adaptive Methods: chose a **local** metric.

  $$\omega_{k+1} = \omega_k - \alpha_k H_k^{-1} \nabla f(\omega_k + \gamma_k(\omega_k - \omega_{k-1})) + \beta_k H_k^{-1} H_{k-1}(\omega_k - \omega_{k-1})$$

  Example: Adagrad, RMSProp, Adam

We have discussed that the Polyak's heavy-ball method will not leads to overfit. [19] find out that the adaptive methods will lead to overfit, results is shown in Figure4.



(a) CIFAR-10 (Train) **(b)** CIFAR-10 (Test)

Figure 4: Adaptive methods leads to overfit.

3.2 Local Entropy

In this section, we will consider stochastic gradient descent's variation in machine learning in order to achieve better test accuracy, also means better generalization. [7] suggested that the flat minima will bring better genrealization in the nonconvex case which is demonstrated at Figure5. This can also been discover in the PAC-Bayesian setting[11]. If $\hat{f}_w$ is the model we select, in the flat minima the loss of $\hat{f}_w$ will similar to the expection of $\hat{f}_{w+v}$, here v is a small gaussion r.v.. And the generalization gap between $\hat{f}_{w+v}$ and f_{w+v} can be guaranteed by the PAC-Bayesian theory that the error is related by $D_{KL}(w + v||N(0, I))$ which is proportion to $||w||_2^2$. Based on the discussion, this section we will focus on finding flat minima.

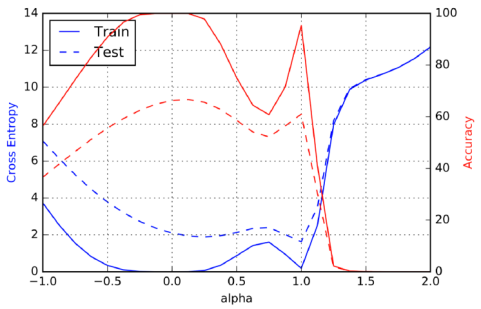


Figure 5: The flater minima will brings better generalization.

[4] gives a method to find the flat minima of the objective function and this section will utilize the differential equation to modeling the algorithms. First we will go on Bayesian setting. For we sample from a Gibbs distribution, it's simple to find that we will have larger probability to lying in a flat area. They modify the gibbs formula

as a new objective in order to reduce the sharp minima, the new objective function is defined as following and we demonstrated the local entropy with different γ in Figure6(a).

Definition 3.1. Local Entropy

$$F(x,\gamma) = \log \int_{x'} \exp(-f(x') - \frac{\gamma}{2}||x - x'||_2^2)dx'$$

[3] point out that the local entropy is the solution to the viscosity Hamilton-Jacobi equation(it is obvious via Hopf-Cole transform)

$$u_t = \frac{1}{2}|\nabla u|^2 + \gamma \Delta u$$

So the optimization problem becomes a multiscale problem, the viscosity Hamilton-Jacobi equation plays the role of fast parameter and the optimzation problem is the slow one. (I'm not an expert on multiscale analysis, I will not go on this part.) Due to the existence of the viscosity, the local minima will change, so what if the objective function can defined as the solution of the Hamilton-Jacobi equation

$$u_t = \frac{1}{2}|\nabla u|^2$$

[3] proved that the local minima will not change the position if using the solution of the Hamilton-Jacobi equation as objective. The solution of $u_t = \frac{1}{2}|\nabla u|^2$ can be given via Hopf-lax lemma, so the update every time can be formulate as

$$x_{k+1} = x_k - \Delta t \nabla \hat{f}(prox_{\lambda f}(x))$$

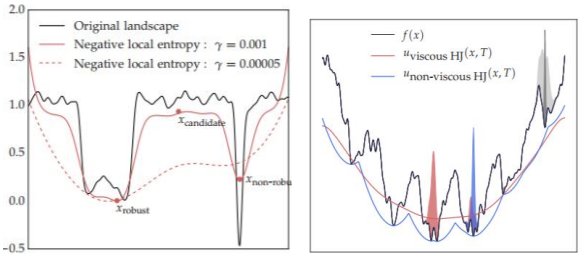


Figure 6: Fig(1) Local entropy concentrates on wide valleys in the energy landscape.(2) Hamilton-Jacobi Regularization Doesn't Change the local minima.

3.3 Stochastic and generalization

As mentioned in the previous section, injecting noise in the optimization algorithm(like SGLD) is equal to the Bayesian methods which is known that avoid overfit. [17] also gives a PAC-Bayesian bound on the algorithm and in this section we will go on the true sgd algorithm.

[18] consider a robustness bound for sgd without Gaussian noise, which is known as "train faster, generalize better". We consider a machine learning problem with the following setting

- model with parameter w; loss function: f; unknown distribution D over examples from some space Z
- a sample $S = (s_1, \cdots, s_n)$ of n examples drawn i.i.d. from D
- population risk
$$R[w] \equiv \mathbb{E}_{s\sim D} f(w; s)$$
- empirical risk
$$R_S[w] \equiv \frac{1}{n}\sum_{i=1}^{n} f(w; s_i)$$

Definition 3.2. A randomized algorithm A is ϵ-uniformly stable if

$$\sup_z \mathbb{E}_A[f(A(S); z) - f(A(S'); z)] \leq \epsilon$$

holds for all samples $S, S' \in Z^n$ that differ in at most one example $\rightsquigarrow$ **uniform stability** implies generalization in expectation

Theorem 3.3. ***generalization from uniform stability*** *Let A be ϵ-uniformly stable. Then,*

$$|\epsilon_{gen}| = |\mathbb{E}_{S,A}[R_S[A(S)] - R[A(S)]]| \leq \epsilon$$

Proof. Denote by $S = (z_1, \cdots, z_n)$ and $S' = (z'_1, \cdots, z'_n)$ two independent random samples and let $S^{(i)} = (z_1, \cdots, z_{i-1}, z'_i, z_{i+1}, \cdots, z_n)$.

$$\begin{aligned}\mathbb{E}_{S,A} R_S[A(S)] &= \mathbb{E}_{S,A}\left[\frac{1}{n}\sum_{i=1}^{n} f(A(S), s_i)\right] \\ &= \mathbb{E}_{S,S',A}\left[\frac{1}{n}\sum_{i=1}^{n} f(A(S^{(i)}), s'_i)\right]\end{aligned}$$

at the same time,

$$\begin{aligned}\mathbb{E}_{S,S',A}\left[\frac{1}{n}\sum_{i=1}^{n} f(A(S^{(i)}), s'_i)\right] &= \mathbb{E}_{S,S',A}\left[\frac{1}{n}\sum_{i=1}^{n} f(A(S), s'_i)\right] + \delta \\ &= \mathbb{E}_{S,A}\left[R[A(S)]\right] + \delta, \quad |\delta| < \epsilon\end{aligned}$$

□

The proof stech is shown following

- Let w and w' be two models trained on two samples S and S' respectively using stochastic gradient method(SGM). S and S' differ in one example.
- Suppose the loss function f is L-lipschitz, we have
$$\mathbb{E}|f(w; z) - f(w'; z)| \leq L\, \mathbb{E}\|w - w'\|$$
- So it suffices to give a bound of $\|w - w'\|$

Assume that f is β-smooth, we have following results for gradient update rule $G_{f,\alpha}(w) = w - \alpha\nabla f(w)$, then we have:

- $G_{f,\alpha}$ is $(1+\alpha\beta)$ expansive
- $G_{f,\alpha}$ is 1 expansive when f is convex and $\alpha < 2/\beta$

Theorem 3.4. *Assume that the loss function $f(\cdot\,; z)$ is β-smooth, convex and L-Lipschitz for every z. Then SGM with step sizes $\alpha_t \leq 2/\beta$ for T steps satisfies uniform stability with*

$$\epsilon_{stab} \leq \frac{2L^2}{n} \sum_{t=1}^{T} \alpha_t$$

Proof. The proof is given following the following sketch:

- Consider the gradient updates $G_1, \cdots, G_T$ and $G'_1, \cdots, G'_T$ induced by running SGM on sample S and S', respectively. Let w_T and w'_T denote the corresponding outputs of SGM and $\delta_T = \|w_T - w'_T\|$.
- With probability $1 - 1/n$ the same example is selected, in this case we can use the 1-expansivity of $G_{T_i} = G'_{T_i}$.
- With probability $1/n$ the selected examples differ. By the triangle inequality,

$$\delta_{t+1} \leq \delta_t + \|G_t(w_t) - w_t\| + \|G'_t(w'_t) - w'_t\| \leq \delta_t + 2L\alpha_t$$

□

4 Asynchronous Parallel Iteration

In this section, we first give a differential equation modeling for the asynchronous gradient descent. [5] propose a dynamic perspective of a asynchronous parallel iteration to the asynchronous gradient descent and we will go futher to the general perturbation case.

Async-parallel update uses a set of agents still perform parallel updates, but synchronization is eliminated or weaked. Hence, each agent continuously applies update, wich reads x from and writes x_i back to the shared memory. k increases whenever any agent completes an update. Formally $x_i^{k+1} = x_i^k - \eta_k \nabla f(x^{k-d_k})$. The lack of synchronization often results in computation with out-of-date information. Figure 7 demonstrated the difference between the sync-parallel update and the async-paralled computing.

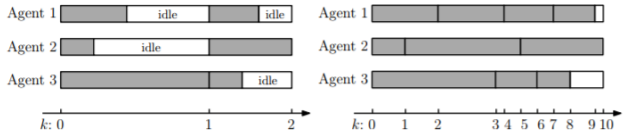


Figure 7: Sync-Parallel computing(left) vs async-paralled computing(right).

4.1 Continuous Time Analysis

First we consider a dynamic of a gradient descent with time delay, let t be time in this section, consider the ODE

$$\dot{x}(t) = -\eta \nabla f(\hat{x}(t))$$

If there is no delay, easily set $\hat{x}(t) = x(t)$, the ODE describe a gradient flow, which monotonically decreases $f(x(t))$ for $\frac{d}{dt}f(x(t)) = \langle \nabla f(x(t)), \dot{x}(t) \rangle = -\frac{1}{\eta}||\dot{x}(t)||_2^2$

Instead, in order to analysis the asynchronous optimization algorithm, we allow delays and impose the bound $c > 0$ on the delays:

$$||\hat{x}(t) - x(t)||_2 \leq \int_{t-c}^{t} ||\dot{x}(s)||_2 ds$$

However We lose monotonicity for the objective function, for

$$\begin{aligned}
\frac{d}{dt}f(x(t)) &= \langle \nabla f(\hat{x}(t)), \dot{x}(t) \rangle + \langle \nabla f(x(t)) - \nabla f(\hat{x}(t)), \dot{x}(t) \rangle \\
&\leq -\frac{1}{\eta}||\dot{x}(t)||_2^2 + L||x(t) - \hat{x}(t)||_2 \cdot ||\dot{x}(t)||_2 \\
&\leq -\frac{1}{2\eta}||\dot{x}(t)||_2^2 + \frac{\eta c L^2}{2} \int_{t-c}^{t} ||\dot{x}(s)||_2^2 ds
\end{aligned}$$

However we can revise the **Energy function** with both f and a weighted total keinetic term, where $\gamma > 0$.

$$\xi(t) = f(x(t)) + \gamma \int_{t-c}^{t} (s - (t-c))||\dot{x}(s)||_2^2 ds \tag{1}$$

$\xi(t)$ has the time derivative

$$\begin{aligned}
\dot{\xi}(t) &= \frac{d}{dt}f(x(t)) + \gamma c||x(t)||_2^2 - \gamma \int_{t-c}^{t} ||\dot{x}(s)||_2^2 ds \\
&\leq -(\frac{1}{\eta} - \gamma)||\dot{x}(t)||_2^2 - (\gamma - \frac{ncL^2}{2}) \int_{t-c}^{t} ||\dot{x}(s)||_2^2 ds
\end{aligned}$$

It is easy to generalize the prove to the discrete algorithm. We can define the Lyapunov function

$$\xi_k := f(x^k) + \frac{L}{2\epsilon} \sum_{i=k-\tau}^{k-1} (i - (k-\tau) + 1)||\Delta^i||_2^2$$

Similarly, we can prove

- $f(x^{k+1}) - f(x^k) \leq \frac{L}{2\epsilon} \sum_{i=k-\tau}^{k-1} ||\Delta^i||_2^2 + [\frac{L(\tau\epsilon+1)}{2} - \frac{L}{\gamma}]||\Delta^k||_2^2$
- $\xi_k - \xi_{k+1} \geq \frac{1}{2}(\frac{1}{\gamma} - \frac{1}{2} - \tau)L \cdot ||\Delta^k||_2^2$

For the converge rate of asynchronuous algorithms can be given straight forward. It is superising that the converge rate is the same as the common gradient descent.

Theorem 4.1. ***Converge Rate Of Asynchronous Gradient Descent.***

$$\lim_k ||\nabla f(x^k)||_2 = 0, \lim_{1\leq i\leq k} ||\nabla f(x^k)||_2 = o(1/\sqrt{k})$$

4.2 Perturbed Dynamic Of Second Order Momentum Dynamic

[10] gives a analysis of perturbed dynamic of inertial dynamics and algorithms with asymptotic vanishing viscosity, where the dynamic can be consider as

$$\ddot{x}(t) + \frac{\alpha}{t}\dot{x}(t) + \nabla\Phi(x(t)) = g(t)$$

The asymptotic behavior can be formulated as following

Theorem 4.2. *Let* Φ *bounded from below and* $\arg\min \Phi \neq \emptyset$

- *If* $\int_{t_0}^{\infty} ||g(t)||dt < +\infty$ *Then*
  - $\sup ||\dot{x}(t)|| < +\infty$
  - $\int_t^{\infty} \frac{1}{\tau}||\dot{x}(\tau)|| < +\infty$
  - $\lim_{t\to+\infty} \Phi(x(t)) = \inf \Phi$
- *If* $\int_{t_0}^{\infty} t||g(t)||dt < +\infty$ *Then*

$$\Phi(x(t)) - \min \Phi = O(1/t^2)$$

The core of the prove is to select two energy function

$$\frac{1}{2}||\dot{x}(t)||^2 + (\Phi(x(t)) - \inf \Phi) + \int_t^T \langle \dot{x}(\tau), g(\tau)\rangle \, d\tau$$

and

$$\frac{2}{\alpha-1}t^2(\Phi(x(t)) - \inf \Phi) + (\alpha-1)||x(t) - x^* + \frac{t}{\alpha-1}\dot{x}(t)||^2 + \int_t^T \tau \left\langle x(\tau) - x^* + \frac{\tau}{\alpha-1}\dot{x}(\tau), g(\tau)\right\rangle$$

There still lack fast converging asynchronous parallel iteration case where we may consider a delay differential system. It is interesting to consider **the asynchronous parallel MC algorithm** and generalization of **asynchronous algorithms**.

5 Discussion

5.1 Futher Work.

In this paper we have analysis the several algorithms:

- Gradient descent and acceleration methods.

- Stochastic gradient descent and acceleration methods.
- Asynchronous parallel gradient descent.

There are still a lot of mystery needs to be studied. We will list some of them in this section.

Asynchronous Algorithms

For asynchronous algorithms, we still have several unknown problems

- Will asynchronous parallel iteration overfit?
- Can accelerated first-order method have an asynchronous version? What is the converge rate?
- Will the accelerated method overfit?
- Analysis the stochastic gradient asynchronous parallel iteration via time-delay stochastic differential equation.

Stiffness of the stochastic algorithm The stability of numerical schemes for stochastic differential equations is still not define. As a result the stiffness of the stochastic dynamic is still not well defined. Will adding a stochastic noise can reduce the stiffness or the noise will make the problem harder? There is also an interesting problem, that how does the second order term $\ddot{X}$ work? We know to have a $n-$th order polynomial decay speed gradient system, we only need to let $\ddot{X}/t^{n-1} = -\nabla f$, this is a simple observation to apply a time transform $\hat{t} = t^n$. So the $\ddot{X}$ here is only to make the problem easier?

A new direction to think the problem. For the second order ODE $\ddot{X} + g(t)\dot{X} = h(X)$, consider the following transform $Y = \frac{X}{f(t)}$, which also means $X = f(t)Y$, and,

$$\ddot{X} = f_{tt}Y + 2f_t\dot{Y} + f\ddot{Y}$$
$$\dot{X} = f_tY + f\dot{Y}$$

If we let $2f_t + g(t)f = 0$, the equation of Y becomes a standard wave equation like equation. The variational perspective in [4] becomes the principle of minimum action. At the same the energy can be easily defined as the Hamiltonian.

References

[1] Boyd, Vandenberghe, Faybusovich. Convex Optimization.

[2] Su W, Boyd S. A differential equation for modeling Nesterov's accelerated gradient method: theory and insights NIPS2015.

[3] Pratik Chaudhari, Adam Oberman, Stanley Osher, Stefano Soatto, and Guillame Carlier, Deep Relaxation: Partial Differential Equations for Optimizing Deep Neural Networks UCLA CAM Report17-20

[4] Chaudhari P, Choromanska A, Soatto S, et al. Entropy-SGD: Biasing Gradient Descent Into Wide Valleys. ICLR2017.

[5] Sun T, Hannah R, Yin W. Asynchronous Coordinate Descent under More Realistic Assumptions. UCLA CAM Report 2017.

[6] A. Wibisono, A. Wilson, and M. I. Jordan. A variational perspective on accelerated methods in optimization. Proceedings of the National Academy of Sciences, 133, E7351-E7358, 2016.

[7] Keskar N S, Mudigere D, Nocedal J, et al. On Large-Batch Training for Deep Learning: Generalization Gap and Sharp Minima. ICLR2017

[8] https://en.wikipedia.org/wiki/Stiff_equation

[9] Cabot A, Engler H, Gadat A E. ON SECOND ORDER DIFFERENTIAL EQUATIONS WITH ASYMPTOTICALLY SMALL DISSIPATION. 2007.

[10] Attouch H, Chbani Z, Peypouquet J, et al. Fast convergence of inertial dynamics and algorithms with asymptotic vanishing viscosity[J]. Mathematical Programming, 2016:1-53.

[11]Neyshabur B, Bhojanapalli S, Mcallester D, et al. Exploring Generalization in Deep Learning. NIPS2017.

[12] Bartlett P L, Bartlett P L, Bartlett P L. Accelerated mirror descent in continuous and discrete time[C]// International Conference on Neural Information Processing Systems. MIT Press, 2015:2845-2853.

[13] A. Wibisono, A. Wilson, and M. I. Jordan. A variational perspective on accelerated methods in optimization. Proceedings of the National Academy of Sciences, 133, E7351-E7358, 2016.

[14] Anthony Yezzi, Ganesh Sundaramoorthi. Accelerated Optimization in the PDE Framework: Formulations for the Active Contour Case arXiv preprint.

[15] Walid Krichene and Peter Bartlett Acceleration and Averaging In Stochastic Descent Dynamics NIPS2017

[16] Jordan R, Otto F, Kinderlehrer D. The Variational Formulation of the Fokker-Planck Equation[J]. Siam Journal on Mathematical Analysis, 1998, 29(1):1-17.

[17] Mou W, Wang L, Zhai X, et al. Generalization Bounds of SGLD for Non-convex Learning: Two Theoretical Viewpoints. arXiv preprint 2017.

[18] Hardt M, Recht B, Singer Y. Train faster, generalize better: Stability of stochastic gradient descent ICML2015.

[19] Nesterov B Y. A method of solving a convex programming problem with convergence rate O($1/k^2$).

[20] Smith S L, Le Q V. A Bayesian Perspective on Generalization and Stochastic Gradient Descent. arXiv preprint 2017.

[21] Smith S L, Kindermans P J, Le Q V. Don't Decay the Learning Rate, Increase the Batch Size. arXiv preprint 2017.

[22] Li Q, Tai C, Weinan E. Stochastic modified equations and adaptive stochastic gradient algorithms. ICML 2017.