CPSC 661: Sampling Algorithms in Machine Learning March 10, 2021

Lecture 11

Lecturer: Andre Wibisono *Scribe: Linyong Nan*

1 Outline

Today's lecture is about the optimization and dynamics. It covers three main parts:

- Optimization as a universal modeling approach to problems in many disciplines
- Optimization from dynamics perspective
- Hierarchy of convex functions for optimization

2 Optimization

We start by introducing optimization as a universal modeling principle for problems/tasks in many disciplines. Specifically, we give examples to show how we can use it to describe and achieve goals in the following fields:

- **General disciplines**: Different disciplines have different quantities that they want to optimize. For examples, *performance* and *cost* in engineering, *utility* and *reward* in economics, *food* and *reproduction* in biology, and *happiness* (possibly) for psychology.
- **Computer science**: There are lots of problems involve searching for an optimal solution (e.g. finding the best cut in a graph, certain elements in a set), and greedy algorithm for instance, provides a reasonable guide for local optimal moves.
- **Machine learning**: Learning from data can be described as optimization of objective function which encodes the goal. Depending on the properties we desire for the solution of the problem, we can design different objective functions. For example, in a binary classification problem, if we want to minimize error of classification, we can add some regularization or to force a maximum margin solution with SVM; if we want to induce sparsity in the solution, we can add L1 regularization to the objective function. It's also worth noting that there exists many challenges: comparing to classical models that have statistical properties that lead to convex objective functions, in modern application we usually deal with sophisticated methods like neural networks or variational inference that gives rise to non-convex objectives that are more difficult to conduct optimization, especially assuming the setting of learning from

large-scale, high-dimensional and noisy data. However, some objectives might have *hidden convexity* that we can exploit via parameterization or manifold.

- **Stochastic systems**: simulated annealing can be used for optimization by sampling from a distribution and slowly decreasing the temperature/noise.

- **Physics**:

  Consider the Newton's second law of motion: Force = mass × acceleration

  $$m\ddot{X}_t = -\nabla U(X_t)$$

  where the force is described as the negative gradient of an objective function $U(X_t)$. This law conserves energy (Hamiltonian):

  $$\mathcal{H} = \frac{m}{2}||\dot{X}_t||^2 + U(X_t)$$

  where $U(X_t)$ represents the potential energy. For example, consider a harmonic oscillator that conform to $m\ddot{X}_t = -X_t$ given that $U(X) = \frac{1}{2}||X||^2$, this system can be viewed as a Hamiltonian flow:

  $$\dot{X}_t = V_t$$
  $$\dot{V}_t = -\frac{1}{m}X_t$$

  Now we describe the Newton's second law of motion as an optimization of the **Principle of least action**, which states that curve of the motion minimizes action

  $$\mathcal{A} = \int_{t_0}^{t_1} \mathcal{L}(X_t, \dot{X}_t)\ dt$$

  where $\mathcal{L}(X_t, \dot{X}_t) = \frac{m}{2}||\dot{X}_t||^2 - U(X_t)$ is the *Lagrangian*.

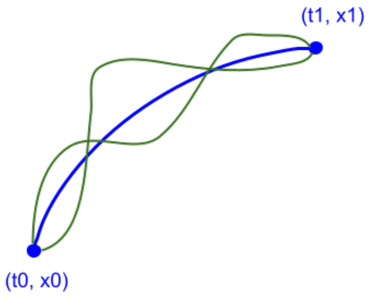

Figure 1: Illustration of optimization over curves

Considering two points in spacetime $(t_0, x_0), (t_1, x_1)$ where $X_{t_0} = x_0, X_{t_1} = x_1$ (shown in Figure 1), the curve/motion that followed the Newton's second law is the optimal curve (in blue) given by the principle of least action. Therefore we see that comparing to Newton's law that is local in time, optimization viewpoint of the motion problem is a more general perspective in the spacetime and the space of all motion. In addition, this formulation of motion also captures intrinsic geometry, it has covariant representation and it governs all physics, including electromagnetism, relativity and quantum physics.

- **Randomness**: Even random processes like Random walk or Brownian motion can be considered as maximizing *entropy*, which is a measure of randomness. Entropy increases along Brownian motion, and we know that Brownian motion (or heat flow) is gradient flow for maximizing entropy. This formulation applies to problems of continuous and discrete space.

Now we provide a general description of the optimization problem:

Given a space $\mathcal{X}$ and an objective function $f : \mathcal{X} \to \mathbb{R}$

Want to find minimizer

$$x^* = \underset{x \in \mathcal{X}}{\text{argmin}}\ f(x)$$

In general, we cannot find x^* exactly. Instead we find $\tilde{x}$ such that:

$$f(\tilde{x}) - f(x^*) \leq \epsilon \text{ or } d(\tilde{x}, x^*) \leq \epsilon \text{ for } \epsilon > 0$$

In the worst case, finding a solution to this problem can be NP-hard (exponential time). However, if we assume some structures (such as convexity), then we can solve efficiently. Note that for this lecture $\mathcal{X} = \mathbb{R}^n$, but it also works for manifolds with different assumptions.

3 Dynamics

We see dynamics as a system modeling approach with which we can solve optimization problem in continuous time.

Definition 1 (Dynamics)**.** *A* ***dynamics*** *on $\mathbb{R}^n$ is determined by a vector field $\phi : \mathbb{R}^n \to \mathbb{R}^n$*

From any $X_0 \in \mathbb{R}^n$, generate a flow $(X_t)_{t \leq 0}$ following:

$$\dot{X}_t = \phi(X_t)$$

For small dt, we have $X_{t+dt} = X_t + \phi(X_t)\ dt$, so that we can apply the following chain rule given some function f:

$$\frac{d}{dt} f(X_t) = \langle \nabla f(X_t), \dot{X}_t \rangle = \langle \nabla f(X_t), \phi(X_t) \rangle$$

We introduce two different dynamics for solving optimization problems:

- **Gradient flow**:

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

- **Heavy ball/accelerated gradient flow**:

$$\ddot{X}_t + \gamma \dot{X}_t + \nabla f(X_t) = 0$$

For today's lecture, we mainly explored using gradient flow for optimization.

3.1 Gradient Flow

Gradient flow is also called First-order in time dynamics, and it is the analog of the **greedy** method for optimization in the continuous time setting, as it is the solution to the following problem:

$$-\nabla f(X_t) = \underset{v \in \mathbb{R}^n}{\operatorname{argmin}} \{ \langle \nabla f(x), v \rangle + \frac{1}{2} ||v||^2 \}$$

where the second term is a regularization term that constrains the size of v. Gradient flow can also be viewed as a **descent method** since by chain rule we have:

$$\frac{d}{dt} f(X_t) = \langle \nabla f(X_t), \dot{X}_t \rangle = -||\nabla f(X_t)||^2 \leq 0$$

3.1.1 Example

We consider an example when f is a quadratic function.

Assume $\mathcal{X} = \mathbb{R}^n$, and let $f(x) = \frac{1}{2} x^T A x$ for some $A \succeq 0$: A is a positive semi-definite matrix (also symmatric). Note that in general, any $A \in \mathbb{R}^{n \times n}$ can be written as:

$$A = A_{sym} + A_{ant}$$

where $A_{sym} = A_{sym}^T$ and $A_{ant} = -A_{ant}^T$, and that:

$$f(x) = \frac{1}{2} x^T A x = \frac{1}{2} x^T (A_{sym} + A_{ant}) x = \frac{1}{2} x^T A_{sym} x, \text{ since } x^T A_{ant} x = 0$$

From given we have the gradient flow:

$$\dot{X}_t = -A X_t$$

with the following solution:

$$X_t = e^{-At} X_0$$

in which e^{-At} is the matrix exponential taking the form of:

$$e^{At} = I + tA + \frac{t^2}{2!} A^2 + \frac{t^3}{3!} A^3 + \cdots$$

We can easily show this for $n = 1$, for which we have $X_t \in \mathbb{R}, A \geq 0$, and then:

$$\begin{aligned} \dot{X}_t = -AX_t \Leftrightarrow \frac{d}{dt}\log X_t = \frac{\dot{X}_t}{X_t} = -A \\ \Rightarrow \log X_t = \log X_0 - At \\ \Rightarrow X_t = X_0 \cdot e^{-At} \end{aligned}$$

If A has eigenvalues $\lambda_1 \geq \cdots \geq \lambda_m > 0 = \lambda_{m+1} = \cdots = \lambda_n$, then

$$||X_t - x^*||^2 \leq e^{-2\lambda_m t}||X_0 - x^*||^2$$

where x^* is the projection of X_0 to the kernel of A.

4 Hierarchy of Convex Functions

We are interested in knowing and quantifying the conditions that allow gradient flow of a function to have a reasonable convergence guarantee. To achieve this, we need to first construct the hierarchy of convex functions.

Definition 2. *Let $f : \mathbb{R}^n \to \mathbb{R}$ be differentiable and $x^* = \operatorname{argmin}_{x\in\mathbb{R}^n} f(x)$.*

1. *f is α-**strongly convex** if:*

$$\begin{aligned} (\nabla f(x) - \nabla f(y))^T(x-y) \geq \alpha||x-y||^2 \qquad (1) \\ \Leftrightarrow \nabla^2 f(x) \succeq \alpha I \\ \Leftrightarrow \forall v \in \mathbb{R}^n : v^T\nabla^2 f(x)v \geq v^T(\alpha I)v = \alpha||v||^2 \end{aligned}$$

   where $\alpha > 0$ and $\nabla^2 f(x)$ is the hessian and is lower bounded by α times the identity matrix. When $\alpha = 0$, f is weakly convex.

2. *f is α-**gradient dominated** if:*

$$||\nabla f(x)||^2 \geq 2\alpha(f(x) - f(x^*)) \qquad (2)$$

   (also known as Polyak-Łojaciewicz inequality)

3. *f is α-**sufficient growth** if:*

$$f(x) - f(x^*) \geq \frac{\alpha}{2}||x - x^*||^2 \qquad (3)$$

Then we have the following theorem:

Theorem 1. *Given the above definition, we have:*

$$(1) \Rightarrow (2) \Rightarrow (3)$$

4.1 Example

We revisit the example in which f is a quadratic function. Let $f(x) = \frac{1}{2}x^T Ax$ for some $A \succeq 0$.

If A has eigenvalues $\lambda_1 \geq \cdots \geq \lambda_m > 0 = \lambda_{m+1} = \cdots = \lambda_n$, then:

1. f is **strongly convex** with $\alpha = \lambda_n = 0$ (since $\alpha_{sc} = \lambda_{min}(\nabla^2 f(x)) = \lambda_{min}(A)$)
2. f is **gradient dominated** with $\alpha = \lambda_m > 0$
3. f is **sufficient growth** with $\alpha = \lambda_m > 0$

4.2 Convergence Rates

From Theorem 1 we are able to derive the following result about convergence rates.

Theorem 2. *Given Definition 2, we have:*

1. *If f is α-**strongly convex**, then gradient flow has exponential contraction:*

$$\textit{For } \dot{X}_t = -\nabla f(X_t), \dot{Y}_t = -\nabla f(Y_t)$$

$$||X_t - Y_t||^2 \leq e^{-2\alpha t}||X_0 - Y_0||^2$$

2. *If f is α-**gradient dominated**, then along gradient flow:*

$$f(X_t) - f(x^*) \leq e^{-2\alpha t}(f(X_0) - f(x^*))$$

3. *If f is convex and has α-**sufficient growth**, along gradient flow:*

$$||X_t - x^*||^2 \leq e^{-\alpha t}||X_0 - x^*||^2$$

Proof. 1. Consider $\dot{X}_t = -\nabla f(X_t)$, $\dot{Y}_t = -\nabla f(Y_t)$, then we have (by strong convexity):

$$\frac{d}{dt}||X_t - Y_t||^2 = 2\langle X_t - Y_t, \dot{X}_t - \dot{Y}_t\rangle = -2\langle X_t - Y_t, \nabla f(X_t) - \nabla f(Y_t)\rangle \leq -2\alpha||X_t - Y_t||^2$$

Let $U_t = ||X_t - Y_t||^2 \geq 0$, then we have (the following routine is also called *Grönwall's inequality*):

$$\dot{U}_t = \frac{d}{dt}U_t \leq -2\alpha U_t$$

$$\Leftrightarrow \quad \frac{d}{dt}\log U_t = \frac{\dot{U}_t}{U_t} \leq -2\alpha$$

$$\Rightarrow \quad \log U_t - \log U_0 \leq -2\alpha t$$

$$\Leftrightarrow \quad U_t = ||X_t - Y_t||^2 \leq U_0 \cdot e^{-2\alpha t} = ||X_0 - Y_0||^2 \cdot e^{-2\alpha t}$$

2. Compute:

$$\frac{d}{dt}(f(X_t) - f(x^*)) = \langle \nabla f(X_t), \dot{X}_t \rangle = -||\nabla f(X_t)||^2 \text{ since } \dot{X}_t = -\nabla f(X_t)$$
$$\leq -2\alpha(f(X_t) - f(x^*)) \text{ by gradient dominance}$$

Following the routine of Grönwall's inequality, we have:

$$f(X_t) - f(x^*) \leq e^{-2\alpha t}(f(X_0) - f(x^*))$$

3. Compute:

$$\frac{d}{dt}||X_t - x^*||^2 = 2\langle X_t - x^*, \dot{X}_t \rangle = -2\langle X_t - x^*, \nabla f(X_t) \rangle$$
$$\leq -2(f(X_t) - f(x^*)) \text{ by convexity of } f$$
$$\leq -\alpha ||X_t - x^*||^2$$

Again by Grönwall's inequality we have:

$$||X_t - x^*||^2 \leq e^{-\alpha t}||X_0 - x^*||^2$$

□

References

[1] Boyd and Vandenberghe, *Convex Optimization*, Cambridge University Press, 2004

[2] Nesterov, *Lectures on Convex Optimization*, Springer, 2004

[3] Nocedal and Wright, *Numerical Optimization*, Springer, 2006

[4] Vishnoi, *Algorithms for Convex Optimization*, 2021

[5] Recht, Optimization, *Big Data Bootcamp at Simons Institute*, 2013

[6] Boyd, *Convex Optimization*, Lecture videos for EE 364A at Stanford, 2008