Introduction to Gradient Flow

Fatiha BARRADE, Fatima NOUTFI

Keywords. *Flow Gradient ; Steepest descent curves , metric spaces , Minimization*

Abstract

Gradient flow, a fundamental concept in mathematics, reveals the dynamic evolution of systems by minimizing functions. This paper explores gradient flow from its basic principles in Euclidean spaces to more intricate variants, discussing its significance in optimization problems and dynamic system analysis. Through variants such as convex and semi-convex gradient flows, we delve into convergence behaviors and solution uniqueness. Moreover, we analyze the role of gradient flow in understanding model trainability, particularly in least squares loss scenarios, offering insights into model fitting capabilities and convergence trajectories in machine learning applications.

1 Introduction

Gradient flow, a foundational concept in mathematics, reveals the dynamic evolution of systems. Originating in linear spaces, it seamlessly extends to more general metric spaces, showcasing adaptability across diverse mathematical contexts.

Operating by minimizing functions, gradient flow guides systems along the steepest descent path, as dictated by the negative gradient of a relevant function. Its versatility extends beyond theoretical realms, finding practical applications in non-smooth analysis, elucidating partial differential equations (PDEs), and optimizing functions, particularly in the landscape of machine learning.

This theory's broad applicability spans traditional mathematical structures and effortlessly adapts to general metric spaces. In essence, gradient flow provides a clear lens for understanding dynamic system evolution, bridging classical mathematics with contemporary applications, notably in fields like machine learning.

2 What is Gradient Flow ?

In the context of comprehending gradient flows within general metric spaces, it is beneficial to start by examining the simplest scenario :the Euclidean space $\mathbb{R}^n$. Although the underlying principles remain applicable in any arbitrary Hilbert space, for the sake of clarity, our discussion will be confined to the finite-dimensional case.

Let $F : \mathbb{R}^n \to \mathbb{R}$ be a sufficiently smooth function, and let $x_0 \in \mathbb{R}^n$ be an initial point. In the Euclidean space, a gradient flow is conceptualized as a trajectory $x(t)$, commencing at $t = 0$ with the starting point x_0. At each time instance, this trajectory dynamically chooses the direction that most effectively minimizes the function F, thereby inducing a descent in the function's value. Mathematically, the solution to this dynamic process is encapsulated by the Cauchy Problem:

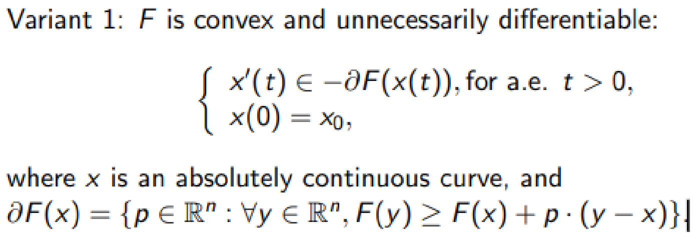
$$\begin{cases} x'(t) = -\nabla F(x(t)) & \text{for } t > 0, \\ x(0) = x_0. \end{cases}$$

This foundational concept of a gradient flow in Euclidean space establishes the groundwork for a more extensive exploration of such flows within more intricate metric spaces.

3 Gradient Flow Variants

3.1 Variant 1 : F is convex and unnecessarily differentiable:

The First Variant of the gradient flow is a mathematical formulation describing the evolution of a system over time, particularly in the context of optimization problems. It involves a convex function F that might not necessarily be differentiable. This variant highlights the dynamic nature of the system's evolution, emphasizing that it is driven by the subdifferential (∂F) of the convex function. The formulation sets up conditions for the initial conditions and the evolution of the system to ensure a unique and well-behaved solution.
Explanation: Now, let's break down the components of the definition:



Variant 1: F is convex and unnecessarily differentiable:

$$\begin{cases} x'(t) \in -\partial F(x(t)), \text{for a.e. } t > 0, \\ x(0) = x_0, \end{cases}$$

where x is an absolutely continuous curve, and
$\partial F(x) = \{p \in \mathbb{R}^n : \forall y \in \mathbb{R}^n, F(y) \geq F(x) + p \cdot (y - x)\}$

Figure 1: Definition

1. **Convex Function (F):**
   A function $F : \mathbb{R}^n \to \mathbb{R}$ is convex if, for any two points x and y in its domain and for any λ in the interval [0, 1], the following inequality holds:

   $$F(\lambda x + (1 - \lambda)y) \leq \lambda F(x) + (1 - \lambda)F(y)$$

2. **Set-Valued Gradient ($\partial F(x)$):**

   The subdifferential or set-valued gradient $\partial F(x)$ at a point x in the domain of F is defined as:

   $$\partial F(x) = \{p \in \mathbb{R}^n : F(y) \geq F(x) + p \cdot (y - x) \ \forall \ y \in \mathbb{R}^n\}$$

3. **Differential Inclusion Equation:**
   The dynamics of the system are described by the differential inclusion equation:

   $$x'(t) \in -\partial F(x(t)), \text{ for almost every } t > 0$$

   This equation expresses that the derivative of the curve $x(t)$ at each point in time is within the negation of the subdifferential of F at $x(t)$.
4. **Initial Condition:**
   The system starts at a given initial condition $x(0) = x_0$, where x_0 is an element in the domain of F.
5. **Absolutely Continuous Curve:**
   The curve $x(t)$ is absolutely continuous, indicating that it is a function with derivatives almost everywhere, allowing for the differentiation of $x(t)$ in the differential inclusion equation.
6. **Uniqueness Theorem:**
   Any two solutions x_1 and x_2 of the differential inclusion problem with different initial conditions satisfy the Lipschitz-like condition:

   $$|x_1(t) - x_2(t)| \leq |x_1(0) - x_2(0)|$$

   This theorem ensures the uniqueness of solutions, indicating that different initial conditions lead to solutions that remain close to each other over time.

3.2 Variant 2: Semi-convex Gradient Flow

Variant 2 of the gradient flow introduces a specific type of convexity known as λ-convexity, where the convex function F is modified by subtracting a quadratic term. This modification is designed to capture the semi-convex nature of F, and the resulting system is described by a differential inclusion equation. In this context, the curve $x(t)$ evolves over time, with the initial conditions and the λ-convex property influencing the system's behavior. The accompanying theorem provides insights into the convergence behavior of solutions, offering a decay estimate for the distance between solutions with different initial conditions. This variant combines the principles of convex analysis with a quadratic penalization term, contributing to a nuanced understanding of gradient flows in optimization problems.

1. **Definition (λ-Convex Function):**
   F is λ-convex ($\lambda \in \mathbb{R}$) if $F(x) - \frac{\lambda}{2}|x|^2$ is convex.
2. **Differential Inclusion Equation:**
   The dynamics of the system are described by the differential inclusion equation:

   $$x'(t) \in -\partial F(x(t)), \text{ for almost every } t > 0$$

   where $x'(t)$ represents the derivative of $x(t)$ with respect to time.
3. **Initial Condition:**
   The system starts at a given initial condition $x(0) = x_0$, where x_0 is an element in the domain of F.
4. **Absolutely Continuous Curve:**

   The curve $x(t)$ is absolutely continuous, indicating that it is a function with derivatives almost everywhere.
5. **Set-Valued Gradient ($\partial F(x)$):**
   The subdifferential or set-valued gradient $\partial F(x)$ at a point x in the domain of F is defined as:

   $$\partial F(x) = \{p \in \mathbb{R}^n : \forall y \in \mathbb{R}^n, F(y) \geq F(x) + p \cdot (y - x) + \frac{\lambda}{2}|y - x|^2\}$$

6. **Theorem:**
   Any two solutions x_1 and x_2 of the above problem with different initial conditions satisfy:

   $$|x_1(t) - x_2(t)| \leq e^{-\lambda t}|x_1(0) - x_2(0)|$$

   This theorem provides a decay estimate on the distance between solutions over time.

4 Understanding the Role of Gradient Flow

Gradient flow is fundamental in understanding how models learn and adapt during training, particularly in the context of least squares loss. By expanding the least squares loss function, we can derive equations that describe how model parameters evolve over time. These equations, such as the differential equation below, represent the change in the residual vector, which captures the differences between model predictions and actual values:

$$\frac{d}{dt}\left[\mathbf{f}[\mathbf{X},\boldsymbol{\phi}]-\mathbf{y}\right]=-\left(\frac{\partial\mathbf{f}[\mathbf{X},\boldsymbol{\phi}]}{\partial\boldsymbol{\phi}}\frac{\partial\mathbf{f}[\mathbf{X},\boldsymbol{\phi}]}{\partial\boldsymbol{\phi}}^T\right)\left(\mathbf{f}[\mathbf{X},\boldsymbol{\phi}]-\mathbf{y}\right)$$

This equation illustrates how the residual vector evolves over time. The product of the gradients in the exponential term determines the rate of change, leading to an exponential decay of the residual:

$$\mathbf{f}[\mathbf{X},\boldsymbol{\phi}_t]-\mathbf{y}=\exp\left[-\frac{\partial\mathbf{f}[\mathbf{X},\boldsymbol{\phi}_t]}{\partial\boldsymbol{\phi}}\frac{\partial\mathbf{f}[\mathbf{X},\boldsymbol{\phi}_t]}{\partial\boldsymbol{\phi}}^T\cdot t\right]\left(\mathbf{f}[\mathbf{X},\boldsymbol{\phi}_0]-\mathbf{y}\right)$$

This exponential decay provides a closed-form solution for the evolution of the function at training points:

$$\mathbf{f}[\mathbf{X},\boldsymbol{\phi}_t]=\mathbf{y}+\exp\left[-\frac{\partial\mathbf{f}[\mathbf{X},\boldsymbol{\phi}_t]}{\partial\boldsymbol{\phi}}\frac{\partial\mathbf{f}[\mathbf{X},\boldsymbol{\phi}_t]}{\partial\boldsymbol{\phi}}^T\cdot t\right]\left(\mathbf{f}[\mathbf{X},\boldsymbol{\phi}_0]-\mathbf{y}\right)$$

In the context of a linear model, the derivative does not depend on the current parameters. Substituting these expressions into the equation for function evolution yields:

$$\mathbf{f}[\mathbf{X},\boldsymbol{\phi}_t]=\mathbf{X}^T\boldsymbol{\phi}_t,\quad \frac{\partial\mathbf{f}[\mathbf{X},\boldsymbol{\phi}_t]}{\partial\boldsymbol{\phi}_t}=\mathbf{X}^T$$

This equation demonstrates how the model predictions evolve over time, considering the initial parameters and the data matrix. Examining the equation for function evolution, we can draw conclusions about the model's trainability based on the exponential term. If the data matrix is full rank, implying a sufficient number of data points relative to parameters, the exponential term converges to zero as time approaches infinity. This indicates that the model fits the prediction exactly. Conversely, if the data matrix is not full rank, the model's ability to fit the prediction diminishes, and residual errors depend on the null space of the matrix.

This aligns with our expectations. The data matrix $\mathbf{X}$ has size $(D+1) \times I$ where D is the input data dimension. Hence, the matrix $\mathbf{X}^T\mathbf{X}$ will be full rank if $I < D+1$, i.e., if the number of data points is less than or equal to the number of parameters. For the case of fitting a line, there are $D+1=2$ parameters, and we can fit $I=1$ or $I=2$ points exactly. However, a line cannot pass exactly through $I > 2$ data points in general position.

Conclusion

In conclusion, gradient flow stands as a foundational concept that underpins dynamic system evolution and optimization processes. Through its versatile applicability in various mathematical contexts, from Euclidean spaces to more general metric spaces, gradient flow provides a unified framework for understanding system dynamics. Variants such as convex and semi-convex gradient flows offer deeper insights into convergence behaviors and solution uniqueness, enriching our understanding of optimization problems. Moreover, in the context of machine learning, gradient flow plays a crucial role in understanding model trainability and convergence speed, particularly evident in least squares loss scenarios. By unraveling the exponential decay of residual vectors, we can assess model fitting capabilities and convergence trajectories, guiding the design and optimization of machine learning models. Overall, gradient flow serves as a cornerstone in mathematical analysis and optimization theory, offering a powerful framework for understanding dynamic system behavior and optimizing complex functions.

References

[1] Santambrogio, F. (2017). Euclidean, metric, and Wasserstein gradient flows: an overview. *Lectures in Mathematics, ETH Zürich.* Retrieved from Springerlink.com.

[2] Ambrosio, L., Gigli, N., & Savaré, G. (2008). *Gradient Flows in Metric Spaces and in the Space of Probability Measures.* Birkhäuser Verlag, Basel – Boston – Berlin. ISBN 3-7643-2428-7.

[3] S. Prince, *Gradient Flow.* Borealis AI, 01/2, `https://www.borealisai.com/research-blogs/gradient-flow/`