



**Gradient Flow. Accelerated gradient flow.**

**Daniil Merkulov**

Optimization methods. MIPT

## Gradient Flow

## Gradient Flow intuition

- Antigradient  $-\nabla f(x)$  indicates the direction of steepest descent at the point  $x$ .

## Gradient Flow intuition

- Antigradient  $-\nabla f(x)$  indicates the direction of steepest descent at the point  $x$ .
- Note also, that the antigradient solves the problem of minimization the Taylor linear approximation of the function on the Euclidian ball

$$\begin{aligned} & \min_{\delta x \in \mathbb{R}^n} \nabla f(x_0)^\top \delta x \\ \text{s.t. } & \delta x^\top \delta x = \varepsilon^2 \end{aligned}$$

## Gradient Flow intuition

- Antigradient  $-\nabla f(x)$  indicates the direction of steepest descent at the point  $x$ .
- Note also, that the antigradient solves the problem of minimization the Taylor linear approximation of the function on the Euclidian ball

$$\begin{aligned} \min_{\delta x \in \mathbb{R}^n} \quad & \nabla f(x_0)^\top \delta x \\ \text{s.t.} \quad & \delta x^\top \delta x = \varepsilon^2 \end{aligned}$$

- The gradient descent is the most classical iterative algorithm to minimize differentiable functions. It comes with a plenty of forms: steepest, stochastic, pre-conditioned, conjugate, proximal, projected, accelerated, etc.

## Gradient Flow intuition

- Antigradient  $-\nabla f(x)$  indicates the direction of steepest descent at the point  $x$ .
- Note also, that the antigradient solves the problem of minimization the Taylor linear approximation of the function on the Euclidian ball

$$\begin{aligned} \min_{\delta x \in \mathbb{R}^n} \quad & \nabla f(x_0)^\top \delta x \\ \text{s.t.} \quad & \delta x^\top \delta x = \varepsilon^2 \end{aligned}$$

- The gradient descent is the most classical iterative algorithm to minimize differentiable functions. It comes with a plenty of forms: steepest, stochastic, pre-conditioned, conjugate, proximal, projected, accelerated, etc.

$$x_{k+1} = x_k - \alpha_k \nabla f(x_k)$$

## Gradient Flow intuition

- Antigradient  $-\nabla f(x)$  indicates the direction of steepest descent at the point  $x$ .
- Note also, that the antigradient solves the problem of minimization the Taylor linear approximation of the function on the Euclidian ball

$$\begin{aligned} \min_{\delta x \in \mathbb{R}^n} \quad & \nabla f(x_0)^\top \delta x \\ \text{s.t.} \quad & \delta x^\top \delta x = \varepsilon^2 \end{aligned}$$

- The gradient descent is the most classical iterative algorithm to minimize differentiable functions. It comes with a plenty of forms: steepest, stochastic, pre-conditioned, conjugate, proximal, projected, accelerated, etc.

$$\begin{aligned} x_{k+1} &= x_k - \alpha_k \nabla f(x_k) \\ x_{k+1} - x_k &= -\alpha_k \nabla f(x_k) \end{aligned}$$

## Gradient Flow intuition

- Antigradient  $-\nabla f(x)$  indicates the direction of steepest descent at the point  $x$ .
- Note also, that the antigradient solves the problem of minimization the Taylor linear approximation of the function on the Euclidian ball

$$\begin{aligned} \min_{\delta x \in \mathbb{R}^n} \quad & \nabla f(x_0)^\top \delta x \\ \text{s.t.} \quad & \delta x^\top \delta x = \varepsilon^2 \end{aligned}$$

- The gradient descent is the most classical iterative algorithm to minimize differentiable functions. It comes with a plenty of forms: steepest, stochastic, pre-conditioned, conjugate, proximal, projected, accelerated, etc.

$$\begin{aligned} x_{k+1} &= x_k - \alpha_k \nabla f(x_k) \\ x_{k+1} - x_k &= -\alpha_k \nabla f(x_k) \\ \frac{x_{k+1} - x_k}{\alpha_k} &= -\nabla f(x_k) \end{aligned}$$



## Gradient Flow intuition

- Antigradient  $-\nabla f(x)$  indicates the direction of steepest descent at the point  $x$ .
- Note also, that the antigradient solves the problem of minimization the Taylor linear approximation of the function on the Euclidian ball

$$\begin{aligned} \min_{\delta x \in \mathbb{R}^n} \nabla f(x_0)^\top \delta x \\ \text{s.t. } \delta x^\top \delta x = \varepsilon^2 \end{aligned}$$

- The gradient descent is the most classical iterative algorithm to minimize differentiable functions. It comes with a plenty of forms: steepest, stochastic, pre-conditioned, conjugate, proximal, projected, accelerated, etc.

$$\begin{aligned} x_{k+1} &= x_k - \alpha_k \nabla f(x_k) \\ x_{k+1} - x_k &= -\alpha_k \nabla f(x_k) \\ \frac{x_{k+1} - x_k}{\alpha_k} &= -\nabla f(x_k) \end{aligned}$$

- The gradient flow is essentially the limit of gradient descent when the step-size  $\alpha_k$  tends to zero

## Gradient Flow intuition

- Antigradient  $-\nabla f(x)$  indicates the direction of steepest descent at the point  $x$ .
- Note also, that the antigradient solves the problem of minimization the Taylor linear approximation of the function on the Euclidian ball

$$\begin{aligned} \min_{\delta x \in \mathbb{R}^n} \nabla f(x_0)^\top \delta x \\ \text{s.t. } \delta x^\top \delta x = \varepsilon^2 \end{aligned}$$

- The gradient descent is the most classical iterative algorithm to minimize differentiable functions. It comes with a plenty of forms: steepest, stochastic, pre-conditioned, conjugate, proximal, projected, accelerated, etc.

$$\begin{aligned} x_{k+1} &= x_k - \alpha_k \nabla f(x_k) \\ x_{k+1} - x_k &= -\alpha_k \nabla f(x_k) \\ \frac{x_{k+1} - x_k}{\alpha_k} &= -\nabla f(x_k) \end{aligned}$$

- The gradient flow is essentially the limit of gradient descent when the step-size  $\alpha_k$  tends to zero

## Gradient Flow intuition

- Antigradient  $-\nabla f(x)$  indicates the direction of steepest descent at the point  $x$ .
- Note also, that the antigradient solves the problem of minimization the Taylor linear approximation of the function on the Euclidian ball

$$\begin{aligned} \min_{\delta x \in \mathbb{R}^n} \quad & \nabla f(x_0)^\top \delta x \\ \text{s.t.} \quad & \delta x^\top \delta x = \varepsilon^2 \end{aligned}$$

- The gradient descent is the most classical iterative algorithm to minimize differentiable functions. It comes with a plenty of forms: steepest, stochastic, pre-conditioned, conjugate, proximal, projected, accelerated, etc.

$$\begin{aligned} x_{k+1} &= x_k - \alpha_k \nabla f(x_k) \\ x_{k+1} - x_k &= -\alpha_k \nabla f(x_k) \\ \frac{x_{k+1} - x_k}{\alpha_k} &= -\nabla f(x_k) \end{aligned}$$

- The gradient flow is essentially the limit of gradient descent when the step-size  $\alpha_k$  tends to zero



$$\frac{dx}{dt} = -\nabla f(x)$$

# Gradient Flow

$k = 100$



- **Simplified analyses.** The gradient flow has no step-size, so all the traditional annoying issues regarding the choice of step-size, with line-search, constant, decreasing or with a weird schedule are unnecessary.

Рис. 1:  Source

# Gradient Flow

$k = 100$

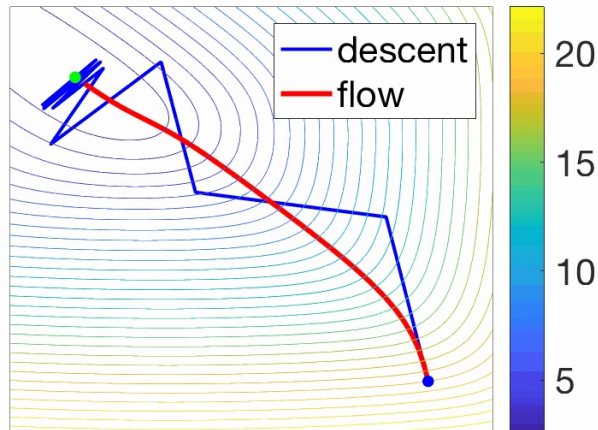


- **Simplified analyses.** The gradient flow has no step-size, so all the traditional annoying issues regarding the choice of step-size, with line-search, constant, decreasing or with a weird schedule are unnecessary.
- **Analytical solution in some cases.** For example, one can consider quadratic problem with linear gradient, which will form a linear ODE with known exact formula.

Рис. 1: Source

# Gradient Flow

$k = 100$



- **Simplified analyses.** The gradient flow has no step-size, so all the traditional annoying issues regarding the choice of step-size, with line-search, constant, decreasing or with a weird schedule are unnecessary.
- **Analytical solution in some cases.** For example, one can consider quadratic problem with linear gradient, which will form a linear ODE with known exact formula.
- **Different discretization leads to different methods.** We will see, that the continuous-time object is pretty rich in terms of the variety of produced algorithms. Therefore, it is interesting to study optimization from this perspective.

Рис. 1:  Source

# Gradient Flow discretization

Consider Gradient Flow ODE:

$$\frac{dx}{dt} = -\nabla f(x)$$

Explicit Euler discretization:

# Gradient Flow discretization

Consider Gradient Flow ODE:

$$\frac{dx}{dt} = -\nabla f(x)$$

Explicit Euler discretization:

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

Leads to ordinary Gradient Descent method



# Gradient Flow discretization

Consider Gradient Flow ODE:

$$\frac{dx}{dt} = -\nabla f(x)$$

Explicit Euler discretization:

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

Leads to ordinary Gradient Descent method

$$\boxed{x_{k+1} = x_k - \alpha \nabla f(x_k)} \quad (\text{GD})$$

# Gradient Flow discretization

Consider Gradient Flow ODE:

$$\frac{dx}{dt} = -\nabla f(x)$$

Explicit Euler discretization:

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

Leads to ordinary Gradient Descent method

$$\boxed{x_{k+1} = x_k - \alpha \nabla f(x_k)}$$

(GD)

Implicit Euler discretization:

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

# Gradient Flow discretization

Consider Gradient Flow ODE:

$$\frac{dx}{dt} = -\nabla f(x)$$

Explicit Euler discretization:

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

Leads to ordinary Gradient Descent method

$$x_{k+1} = x_k - \alpha \nabla f(x_k)$$

(GD)

Implicit Euler discretization:

$$\begin{aligned} \frac{x_{k+1} - x_k}{\alpha} &= -\nabla f(x_{k+1}) \\ \frac{x_{k+1} - x_k}{\alpha} + \nabla f(x_{k+1}) &= 0 \end{aligned}$$

# Gradient Flow discretization

Consider Gradient Flow ODE:

$$\frac{dx}{dt} = -\nabla f(x)$$

Explicit Euler discretization:

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

Leads to ordinary Gradient Descent method

$$\boxed{x_{k+1} = x_k - \alpha \nabla f(x_k)}$$

Implicit Euler discretization:

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

$$\frac{x_{k+1} - x_k}{\alpha} + \nabla f(x_{k+1}) = 0$$

$$\frac{x - x_k}{\alpha} + \nabla f(x) \Big|_{x=x_{k+1}} = 0$$

(GD)

# Gradient Flow discretization

Consider Gradient Flow ODE:

$$\frac{dx}{dt} = -\nabla f(x)$$

Explicit Euler discretization:

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

Leads to ordinary Gradient Descent method

$$\boxed{x_{k+1} = x_k - \alpha \nabla f(x_k)}$$

(GD)

Implicit Euler discretization:

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

$$\frac{x_{k+1} - x_k}{\alpha} + \nabla f(x_{k+1}) = 0$$

$$\frac{x - x_k}{\alpha} + \nabla f(x) \Big|_{x=x_{k+1}} = 0$$

$$\nabla \left[ \frac{1}{2\alpha} \|x - x_k\|_2^2 + f(x) \right] \Big|_{x=x_{k+1}} = 0$$

# Gradient Flow discretization

Consider Gradient Flow ODE:

$$\frac{dx}{dt} = -\nabla f(x)$$

Explicit Euler discretization:

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

Leads to ordinary Gradient Descent method

$$\boxed{x_{k+1} = x_k - \alpha \nabla f(x_k)}$$

(GD)

Implicit Euler discretization:

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

$$\frac{x_{k+1} - x_k}{\alpha} + \nabla f(x_{k+1}) = 0$$

$$\frac{x - x_k}{\alpha} + \nabla f(x) \Big|_{x=x_{k+1}} = 0$$

$$\nabla \left[ \frac{1}{2\alpha} \|x - x_k\|_2^2 + f(x) \right] \Big|_{x=x_{k+1}} = 0$$

$$x_{k+1} = \arg \min_{x \in \mathbb{R}^n} \left[ f(x) + \frac{1}{2\alpha} \|x - x_k\|_2^2 \right]$$

# Gradient Flow discretization

Consider Gradient Flow ODE:

$$\frac{dx}{dt} = -\nabla f(x)$$

Explicit Euler discretization:

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

Leads to ordinary Gradient Descent method

$$\boxed{x_{k+1} = x_k - \alpha \nabla f(x_k)}$$

(GD)

Implicit Euler discretization:

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

$$\frac{x_{k+1} - x_k}{\alpha} + \nabla f(x_{k+1}) = 0$$

$$\frac{x - x_k}{\alpha} + \nabla f(x) \Big|_{x=x_{k+1}} = 0$$

$$\nabla \left[ \frac{1}{2\alpha} \|x - x_k\|_2^2 + f(x) \right] \Big|_{x=x_{k+1}} = 0$$

$$x_{k+1} = \arg \min_{x \in \mathbb{R}^n} \left[ f(x) + \frac{1}{2\alpha} \|x - x_k\|_2^2 \right]$$

# Gradient Flow discretization

Consider Gradient Flow ODE:

$$\frac{dx}{dt} = -\nabla f(x)$$

Explicit Euler discretization:

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

Leads to ordinary Gradient Descent method

$$\boxed{x_{k+1} = x_k - \alpha \nabla f(x_k)}$$

(GD)

Implicit Euler discretization:

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

$$\frac{x_{k+1} - x_k}{\alpha} + \nabla f(x_{k+1}) = 0$$

$$\frac{x - x_k}{\alpha} + \nabla f(x) \Big|_{x=x_{k+1}} = 0$$

$$\nabla \left[ \frac{1}{2\alpha} \|x - x_k\|_2^2 + f(x) \right] \Big|_{x=x_{k+1}} = 0$$

$$x_{k+1} = \arg \min_{x \in \mathbb{R}^n} \left[ f(x) + \frac{1}{2\alpha} \|x - x_k\|_2^2 \right]$$

$$\boxed{x_{k+1} = \text{prox}_{\alpha f}(x_k)}$$

(PPM)



## Convergence analysis. Convex case.

1. Simplest proof of monotonic decrease of GF:

$$\frac{d}{dt}f(x(t)) = \nabla f(x(t))^\top \frac{dx(t)}{dt} = -\|\nabla f(x(t))\|_2^2 \leq 0.$$

If  $f$  is bounded from below, then  $f(x(t))$  will always converge as a non-increasing function which is bounded from below. It is straightforward, that GF converges to the stationary point, where  $\nabla f = 0$  (potentially including minima, maxima and saddle points).

## Convergence analysis. Convex case.

1. Simplest proof of monotonic decrease of GF:

$$\frac{d}{dt}f(x(t)) = \nabla f(x(t))^\top \frac{dx(t)}{dt} = -\|\nabla f(x(t))\|_2^2 \leq 0.$$

If  $f$  is bounded from below, then  $f(x(t))$  will always converge as a non-increasing function which is bounded from below. It is straightforward, that GF converges to the stationary point, where  $\nabla f = 0$  (potentially including minima, maxima and saddle points).

2. If we additionally have convexity:

$$f(x) \geq f(y) + \nabla f(y)^\top (x - y) \quad \Rightarrow \quad \nabla f(y)^\top (x - y) \leq f(x) - f(y)$$

## Convergence analysis. Convex case.

1. Simplest proof of monotonic decrease of GF:

$$\frac{d}{dt}f(x(t)) = \nabla f(x(t))^\top \frac{dx(t)}{dt} = -\|\nabla f(x(t))\|_2^2 \leq 0.$$

If  $f$  is bounded from below, then  $f(x(t))$  will always converge as a non-increasing function which is bounded from below. It is straightforward, that GF converges to the stationary point, where  $\nabla f = 0$  (potentially including minima, maxima and saddle points).

2. If we additionally have convexity:

$$f(x) \geq f(y) + \nabla f(y)^\top (x - y) \quad \Rightarrow \quad \nabla f(y)^\top (x - y) \leq f(x) - f(y)$$

3. Finally, using convexity:

$$\frac{d}{dt}[\|x(t) - x^*\|^2] = -2(x(t) - x^*)^\top \nabla f(x(t)) \leq -2[f(x(t)) - f^*]$$

## Convergence analysis. Convex case.

1. Simplest proof of monotonic decrease of GF:

$$\frac{d}{dt}f(x(t)) = \nabla f(x(t))^\top \frac{dx(t)}{dt} = -\|\nabla f(x(t))\|_2^2 \leq 0.$$

If  $f$  is bounded from below, then  $f(x(t))$  will always converge as a non-increasing function which is bounded from below. It is straightforward, that GF converges to the stationary point, where  $\nabla f = 0$  (potentially including minima, maxima and saddle points).

2. If we additionally have convexity:

$$f(x) \geq f(y) + \nabla f(y)^\top (x - y) \quad \Rightarrow \quad \nabla f(y)^\top (x - y) \leq f(x) - f(y)$$

3. Finally, using convexity:

$$\frac{d}{dt}[\|x(t) - x^*\|^2] = -2(x(t) - x^*)^\top \nabla f(x(t)) \leq -2[f(x(t)) - f^*]$$

4. Leading to, by integrating from 0 to  $t$ , and using the monotonicity of  $f(x(t))$ :

$$f(x(t)) - f^* \leq \frac{1}{t} \int_0^t [f(x(u)) - f^*] du \leq \frac{1}{2t} \|x(0) - x^*\|^2 - \frac{1}{2t} \|x(t) - x^*\|^2 \leq \frac{1}{2t} \|x(0) - x^*\|^2.$$

## Convergence analysis. Convex case.

1. Simplest proof of monotonic decrease of GF:

$$\frac{d}{dt} f(x(t)) = \nabla f(x(t))^\top \frac{dx(t)}{dt} = -\|\nabla f(x(t))\|_2^2 \leq 0.$$

If  $f$  is bounded from below, then  $f(x(t))$  will always converge as a non-increasing function which is bounded from below. It is straightforward, that GF converges to the stationary point, where  $\nabla f = 0$  (potentially including minima, maxima and saddle points).

2. If we additionally have convexity:

$$f(x) \geq f(y) + \nabla f(y)^\top (x - y) \quad \Rightarrow \quad \nabla f(y)^\top (x - y) \leq f(x) - f(y)$$

3. Finally, using convexity:

$$\frac{d}{dt} [\|x(t) - x^*\|^2] = -2(x(t) - x^*)^\top \nabla f(x(t)) \leq -2[f(x(t)) - f^*]$$

4. Leading to, by integrating from 0 to  $t$ , and using the monotonicity of  $f(x(t))$ :

$$f(x(t)) - f^* \leq \frac{1}{t} \int_0^t [f(x(u)) - f^*] du \leq \frac{1}{2t} \|x(0) - x^*\|^2 - \frac{1}{2t} \|x(t) - x^*\|^2 \leq \frac{1}{2t} \|x(0) - x^*\|^2.$$

## Convergence analysis. Convex case.

1. Simplest proof of monotonic decrease of GF:

$$\frac{d}{dt} f(x(t)) = \nabla f(x(t))^\top \frac{dx(t)}{dt} = -\|\nabla f(x(t))\|_2^2 \leq 0.$$

If  $f$  is bounded from below, then  $f(x(t))$  will always converge as a non-increasing function which is bounded from below. It is straightforward, that GF converges to the stationary point, where  $\nabla f = 0$  (potentially including minima, maxima and saddle points).

2. If we additionally have convexity:

$$f(x) \geq f(y) + \nabla f(y)^\top (x - y) \quad \Rightarrow \quad \nabla f(y)^\top (x - y) \leq f(x) - f(y)$$

3. Finally, using convexity:

$$\frac{d}{dt} [\|x(t) - x^*\|^2] = -2(x(t) - x^*)^\top \nabla f(x(t)) \leq -2[f(x(t)) - f^*]$$

4. Leading to, by integrating from 0 to  $t$ , and using the monotonicity of  $f(x(t))$ :

$$f(x(t)) - f^* \leq \frac{1}{t} \int_0^t [f(x(u)) - f^*] du \leq \frac{1}{2t} \|x(0) - x^*\|^2 - \frac{1}{2t} \|x(t) - x^*\|^2 \leq \frac{1}{2t} \|x(0) - x^*\|^2.$$

We recover the usual rates in  $\mathcal{O}\left(\frac{1}{k}\right)$ , with  $t = \alpha k$ .

## Convergence analysis. PL case.

1. The analysis is straightforward. Suppose, the function satisfies PL-condition:

$$\|\nabla f(x)\|^2 \geq 2\mu(f(x) - f^*) \quad \forall x$$

## Convergence analysis. PL case.

1. The analysis is straightforward. Suppose, the function satisfies PL-condition:

$$\|\nabla f(x)\|^2 \geq 2\mu(f(x) - f^*) \quad \forall x$$

2. Then

$$\frac{d}{dt}[f(x(t)) - f(x^*)] = \nabla f(x(t))^\top \dot{x}(t) = -\|\nabla f(x(t))\|_2^2 \leq -2\mu[f(x(t)) - f^*]$$



## Convergence analysis. PL case.

1. The analysis is straightforward. Suppose, the function satisfies PL-condition:

$$\|\nabla f(x)\|^2 \geq 2\mu(f(x) - f^*) \quad \forall x$$

2. Then

$$\frac{d}{dt}[f(x(t)) - f(x^*)] = \nabla f(x(t))^\top \dot{x}(t) = -\|\nabla f(x(t))\|_2^2 \leq -2\mu[f(x(t)) - f^*]$$

3. Finally,

$$f(x(t)) - f^* \leq \exp(-2\mu t)[f(x(0)) - f^*],$$

## Accelerated Gradient Flow

# Accelerated Gradient Flow

Remember one of the forms of Nesterov Accelerated Gradient

$$\begin{aligned}x_{k+1} &= y_k - \alpha \nabla f(y_k) \\ y_k &= x_k + \frac{k-1}{k+2}(x_k - x_{k-1})\end{aligned}$$

The corresponding <sup>1</sup> ODE is:

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

---

<sup>1</sup>A Differential Equation for Modeling Nesterov's Accelerated Gradient Method: Theory and Insights, Weijie Su, Stephen Boyd, Emmanuel J. Candes

# Accelerated Gradient Flow

Define the *energy*

$$E(t) = t^2(f(X(t)) - f^*) + 2\left\|X(t) - x^* + \frac{t}{2}\dot{X}(t)\right\|^2.$$

A direct differentiation using the ODE yields  $\dot{E}(t) \leq 0$  for all  $t > 0$ ; hence  $E(t)$  is non-increasing. Because the second term is non-negative we obtain the *convergence theorem*

$$\boxed{f(X(t)) - f^* \leq \frac{2\|x_0 - x^*\|^2}{t^2}}. \quad (\text{AGF-rate})$$

Thus AGF enjoys the same  $\mathcal{O}(1/t^2)$  rate that discrete NAG achieves in  $\mathcal{O}(1/k^2)$  iterations. A similar argument with a *restarted* ODE gives an exponential rate for  $\mu$ -strongly convex  $f$ .

## Stochastic Gradient Flow

# Stochastic Gradient Flow

How to model stochasticity in the continuous process? A simple idea would be:  $\frac{dx}{dt} = -\nabla f(x) + \xi$  with variety of options for  $\xi$ , for example  $\xi \sim \mathcal{N}(0, \sigma^2) \sim \sigma^2 \mathcal{N}(0, 1)$ .

Therefore, one can write down Stochastic Differential Equation (SDE) for analysis:

$$dx(t) = -\nabla f(x(t)) dt + \sigma dW(t)$$

Here  $W(t)$  is called Wiener process. It is interesting, that one could analyze the convergence of the stochastic process above in two possible ways:

- Watching the trajectories of  $x(t)$

# Stochastic Gradient Flow

How to model stochasticity in the continuous process? A simple idea would be:  $\frac{dx}{dt} = -\nabla f(x) + \xi$  with variety of options for  $\xi$ , for example  $\xi \sim \mathcal{N}(0, \sigma^2) \sim \sigma^2 \mathcal{N}(0, 1)$ .

Therefore, one can write down Stochastic Differential Equation (SDE) for analysis:

$$dx(t) = -\nabla f(x(t)) dt + \sigma dW(t)$$

Here  $W(t)$  is called Wiener process. It is interesting, that one could analyze the convergence of the stochastic process above in two possible ways:

- Watching the trajectories of  $x(t)$
- Watching the evolution of distribution density function of  $\rho(t)$

# Stochastic Gradient Flow

How to model stochasticity in the continuous process? A simple idea would be:  $\frac{dx}{dt} = -\nabla f(x) + \xi$  with variety of options for  $\xi$ , for example  $\xi \sim \mathcal{N}(0, \sigma^2) \sim \sigma^2 \mathcal{N}(0, 1)$ .

Therefore, one can write down Stochastic Differential Equation (SDE) for analysis:

$$dx(t) = -\nabla f(x(t)) dt + \sigma dW(t)$$

Here  $W(t)$  is called Wiener process. It is interesting, that one could analyze the convergence of the stochastic process above in two possible ways:

- Watching the trajectories of  $x(t)$
- Watching the evolution of distribution density function of  $\rho(t)$



# Stochastic Gradient Flow

How to model stochasticity in the continuous process? A simple idea would be:  $\frac{dx}{dt} = -\nabla f(x) + \xi$  with variety of options for  $\xi$ , for example  $\xi \sim \mathcal{N}(0, \sigma^2) \sim \sigma^2 \mathcal{N}(0, 1)$ .

Therefore, one can write down Stochastic Differential Equation (SDE) for analysis:

$$dx(t) = -\nabla f(x(t)) dt + \sigma dW(t)$$

Here  $W(t)$  is called Wiener process. It is interesting, that one could analyze the convergence of the stochastic process above in two possible ways:

- Watching the trajectories of  $x(t)$
- Watching the evolution of distribution density function of  $\rho(t)$

! Fokker-Planck equation

$$\frac{\partial \rho}{\partial t} = \nabla (\rho(t) \nabla f) + \frac{\sigma^2}{2} \Delta \rho(t)$$

# Sources

- Francis Bach blog

# Sources

- Francis Bach blog
- Off convex Path blog

# Sources

- Francis Bach blog
- Off convex Path blog
- Stochastic gradient algorithms from ODE splitting perspective

# Sources

- Francis Bach blog
- Off convex Path blog
- Stochastic gradient algorithms from ODE splitting perspective
- NAG-GS: Semi-Implicit, Accelerated and Robust Stochastic Optimizer

# Sources

- Francis Bach blog
- Off convex Path blog
- Stochastic gradient algorithms from ODE splitting perspective
- NAG-GS: Semi-Implicit, Accelerated and Robust Stochastic Optimizer
- Introduction to Gradient Flows in the 2-Wasserstein Space

# Sources

- Francis Bach blog
- Off convex Path blog
- Stochastic gradient algorithms from ODE splitting perspective
- NAG-GS: Semi-Implicit, Accelerated and Robust Stochastic Optimizer
- Introduction to Gradient Flows in the 2-Wasserstein Space
- Stochastic Modified Equations and Dynamics of Stochastic Gradient Algorithms I: Mathematical Foundations

# Sources

- Francis Bach blog
- Off convex Path blog
- Stochastic gradient algorithms from ODE splitting perspective
- NAG-GS: Semi-Implicit, Accelerated and Robust Stochastic Optimizer
- Introduction to Gradient Flows in the 2-Wasserstein Space
- Stochastic Modified Equations and Dynamics of Stochastic Gradient Algorithms I: Mathematical Foundations
- Understanding Optimization in Deep Learning with Central Flows