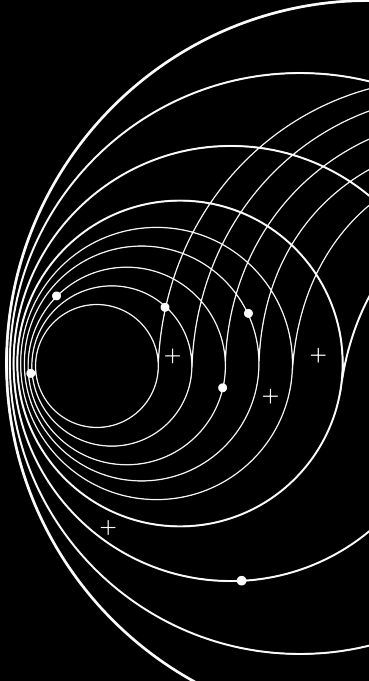
Yandex Research

Nesterov Finds GRAAL: Optimal and Adaptative Gradient Method for Convex Optimization

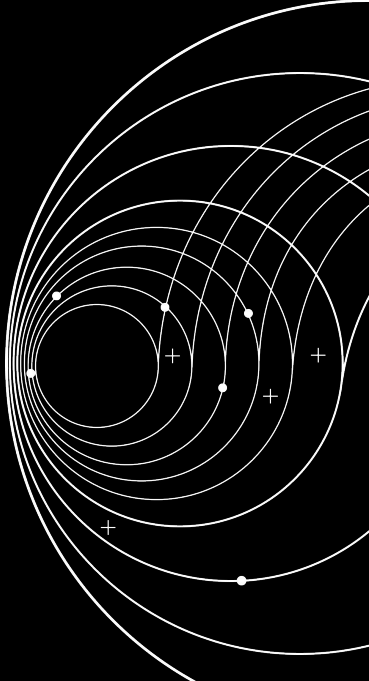
Dmitry Kovalev
Yandex Research

May 13, 2026



Yandex Research

Introduction



Optimization

Optimization Problem

$$\min_{x \in \mathbb{R}^d} f(x)$$

Optimization

Optimization Problem

$$\min_{x \in \mathbb{R}^d} f(x)$$

- $f(x)\colon \mathbb{R}^d \to \mathbb{R}$ – differentiable objective function

Optimization

Optimization Problem

$$\min_{x \in \mathbb{R}^d} f(x)$$

- $f(x) \colon \mathbb{R}^d \to \mathbb{R}$ – differentiable objective function
- $x^* \in \mathbb{R}^d$ – solution, $f^* = f(x^*)$

Optimization

Optimization Problem

$$\min_{x \in \mathbb{R}^d} f(x)$$

- $f(x) \colon \mathbb{R}^d \to \mathbb{R}$ – differentiable objective function
- $x^* \in \mathbb{R}^d$ – solution, $f^* = f(x^*)$

Example: Empirical Risk Minimization

$$f(x) = \frac{1}{N} \sum_{i=1}^{N} \ell(b_i, h(a_i; x))$$

Optimization

Optimization Problem

$$\min_{x \in \mathbb{R}^d} f(x)$$

- $f(x) \colon \mathbb{R}^d \to \mathbb{R}$ – differentiable objective function
- $x^* \in \mathbb{R}^d$ – solution, $f^* = f(x^*)$

Example: Empirical Risk Minimization

$$f(x) = \frac{1}{N} \sum_{i=1}^{N} \ell(b_i, h(a_i; x))$$

- i-th data sample: $a_i \in \mathcal{A}$ – feature vector, $b_i \in \mathcal{B}$ – target variable

Optimization

Optimization Problem

$$\min_{x \in \mathbb{R}^d} f(x)$$

- $f(x) \colon \mathbb{R}^d \to \mathbb{R}$ – differentiable objective function
- $x^* \in \mathbb{R}^d$ – solution, $f^* = f(x^*)$

Example: Empirical Risk Minimization

$$f(x) = \frac{1}{N} \sum_{i=1}^{N} \ell(b_i, h(a_i; x))$$

- i-th data sample: $a_i \in \mathcal{A}$ – feature vector, $b_i \in \mathcal{B}$ – target variable
- $h(a; x) \colon \mathcal{A} \times \mathbb{R}^d \to \mathcal{B}$ – ML model, parameterized by vector $x \in \mathbb{R}^d$

Optimization

Optimization Problem

$$\min_{x \in \mathbb{R}^d} f(x)$$

- $f(x) \colon \mathbb{R}^d \to \mathbb{R}$ – differentiable objective function
- $x^* \in \mathbb{R}^d$ – solution, $f^* = f(x^*)$

Example: Empirical Risk Minimization

$$f(x) = \frac{1}{N} \sum_{i=1}^{N} \ell(b_i, h(a_i; x))$$

- i-th data sample: $a_i \in \mathcal{A}$ – feature vector, $b_i \in \mathcal{B}$ – target variable
- $h(a; x) \colon \mathcal{A} \times \mathbb{R}^d \to \mathcal{B}$ – ML model, parameterized by vector $x \in \mathbb{R}^d$
- $\ell(b, b') \colon \mathcal{B} \times \mathcal{B} \to \mathbb{R}$ – loss function

Optimization

Optimization Problem

$$\min_{x \in \mathbb{R}^d} f(x)$$

- $f(x)\colon \mathbb{R}^d \to \mathbb{R}$ – differentiable objective function
- $x^* \in \mathbb{R}^d$ – solution, $f^* = f(x^*)$

Example: Empirical Risk Minimization

$$f(x) = \frac{1}{N} \sum_{i=1}^{N} \ell(b_i, h(a_i; x))$$

- i-th data sample: $a_i \in \mathcal{A}$ – feature vector, $b_i \in \mathcal{B}$ – target variable
- $h(a; x)\colon \mathcal{A} \times \mathbb{R}^d \to \mathcal{B}$ – ML model, parameterized by vector $x \in \mathbb{R}^d$
- $\ell(b, b')\colon \mathcal{B} \times \mathcal{B} \to \mathbb{R}$ – loss function

First-Order Optimization Methods

Optimization

Optimization Problem

$$\min_{x \in \mathbb{R}^d} f(x)$$

- $f(x)\colon \mathbb{R}^d \to \mathbb{R}$ – differentiable objective function
- $x^* \in \mathbb{R}^d$ – solution, $f^* = f(x^*)$

Example: Empirical Risk Minimization

$$f(x) = \frac{1}{N} \sum_{i=1}^{N} \ell(b_i, h(a_i; x))$$

- i-th data sample: $a_i \in \mathcal{A}$ – feature vector, $b_i \in \mathcal{B}$ – target variable
- $h(a; x)\colon \mathcal{A} \times \mathbb{R}^d \to \mathcal{B}$ – ML model, parameterized by vector $x \in \mathbb{R}^d$
- $\ell(b, b')\colon \mathcal{B} \times \mathcal{B} \to \mathbb{R}$ – loss function

First-Order Optimization Methods

- Perform iterative updates using $f(x_k)$ and $\nabla f(x_k)$

Optimization

Optimization Problem

$$\min_{x \in \mathbb{R}^d} f(x)$$

- $f(x)\colon \mathbb{R}^d \to \mathbb{R}$ – differentiable objective function
- $x^* \in \mathbb{R}^d$ – solution, $f^* = f(x^*)$

Example: Empirical Risk Minimization

$$f(x) = \frac{1}{N} \sum_{i=1}^{N} \ell(b_i, h(a_i; x))$$

- i-th data sample: $a_i \in \mathcal{A}$ – feature vector, $b_i \in \mathcal{B}$ – target variable
- $h(a; x)\colon \mathcal{A} \times \mathbb{R}^d \to \mathcal{B}$ – ML model, parameterized by vector $x \in \mathbb{R}^d$
- $\ell(b, b')\colon \mathcal{B} \times \mathcal{B} \to \mathbb{R}$ – loss function

First-Order Optimization Methods

- Perform iterative updates using $f(x_k)$ and $\nabla f(x_k)$
- Simplest example – GD: $x_{k+1} = x_k - \eta \nabla f(x_k), \eta > 0$ – stepsize

Convex Optimization

Optimization Problem

$$\min_{x \in \mathbb{R}^d} f(x)$$

Convex Optimization

Optimization Problem

$$\min_{x \in \mathbb{R}^d} f(x)$$

- $f(x)$ – convex function

Convex Optimization

Optimization Problem

$$\min_{x \in \mathbb{R}^d} f(x)$$

- $f(x)$ – convex function

Why Convexity?

Convex Optimization

Optimization Problem

$$\min_{x \in \mathbb{R}^d} f(x)$$

- $f(x)$ – convex function

Why Convexity?

- Inspiration for efficient optimization methods: GD with momentum, AdaGrad/Adam, etc.

Convex Optimization

Optimization Problem

$$\min_{x \in \mathbb{R}^d} f(x)$$

- $f(x)$ – convex function

Why Convexity?

- Inspiration for efficient optimization methods: GD with momentum, AdaGrad/Adam, etc.
- Deep neural networks may adhere to convexity or its relaxations (star/quasar convexity) in practice (Kleinberg et al., 2018; Zhou et al., 2019)

Convex Optimization

Optimization Problem

$$\min_{x \in \mathbb{R}^d} f(x)$$

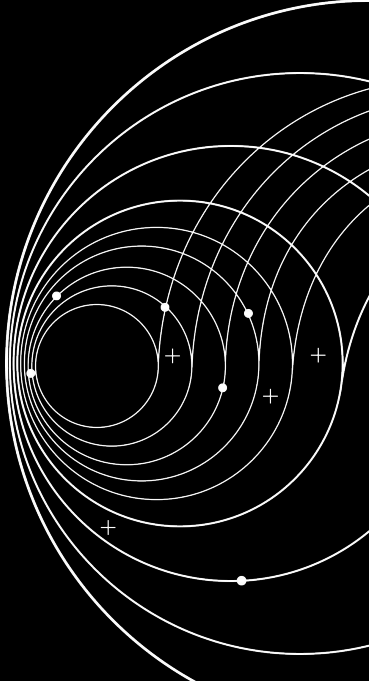
- $f(x)$ – convex function

Why Convexity?

- Inspiration for efficient optimization methods: GD with momentum, AdaGrad/Adam, etc.
- Deep neural networks may adhere to convexity or its relaxations (star/quasar convexity) in practice (Kleinberg et al., 2018; Zhou et al., 2019)
- It is not possible to obtain meaningful convergence guarantees with respect to the objective function gap, $f(x) - f^* \leq \epsilon$, in the general non-convex case; only first-order stationarity, $\|\nabla f(x)\| \leq \epsilon$, can be guaranteed

Yandex Research

Adaptive Methods



Non-Adaptative Methods

Gradient Descent

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

Non-Adaptative Methods

Gradient Descent

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

- Slow $\mathcal{O}(1/k)$ rate for L-smooth convex functions

Non-Adaptative Methods

Gradient Descent

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

- Slow $\mathcal{O}(1/k)$ rate for L-smooth convex functions
- Improved $\mathcal{O}(1/k^2)$ rate if Nesterov acceleration is used (Nesterov, 1983)

Non-Adaptative Methods

Gradient Descent

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

- Slow $\mathcal{O}(1/k)$ rate for L-smooth convex functions
- Improved $\mathcal{O}(1/k^2)$ rate if Nesterov acceleration is used (Nesterov, 1983)

Issues with GD and Accelerated GD: **requires tuning stepsize** η_k

Non-Adaptative Methods

Gradient Descent

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

- Slow $\mathcal{O}(1/k)$ rate for L-smooth convex functions
- Improved $\mathcal{O}(1/k^2)$ rate if Nesterov acceleration is used (Nesterov, 1983)

Issues with GD and Accelerated GD: **requires tuning stepsize** η_k

- Globally with hyperparameter tuning

Non-Adaptative Methods

Gradient Descent

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

- Slow $\mathcal{O}(1/k)$ rate for L-smooth convex functions
- Improved $\mathcal{O}(1/k^2)$ rate if Nesterov acceleration is used (Nesterov, 1983)

Issues with GD and Accelerated GD: **requires tuning stepsize** η_k

- Globally with hyperparameter tuning
- At each iteration with line-search

AdaGrad-type Methods

AdaGrad (scalar-stepsize vaiant)

(Duchi et al., 2011)

$$\eta_k = \eta \cdot \left[\sum_{i=0}^{k} \|\nabla f(x_i)\|^2\right]^{-1/2}$$

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

AdaGrad-type Methods

AdaGrad (scalar-stepsize vaiant)

(Duchi et al., 2011)

$$\eta_k = \eta \cdot \left[\sum_{i=0}^{k} \|\nabla f(x_i)\|^2\right]^{-1/2}$$
$$x_{k+1} = x_k - \eta_k \nabla f(x_k)$$

Significance of AdaGrad

AdaGrad was a baseline in the development of

AdaGrad-type Methods

AdaGrad (scalar-stepsize vaiant)

(Duchi et al., 2011)

$$\eta_k = \eta \cdot \left[\sum_{i=0}^{k} \|\nabla f(x_i)\|^2\right]^{-1/2}$$
$$x_{k+1} = x_k - \eta_k \nabla f(x_k)$$

Significance of AdaGrad

AdaGrad was a baseline in the development of

- RMSProp (Tieleman, 2012)

AdaGrad-type Methods

AdaGrad (scalar-stepsize vaiant)

(Duchi et al., 2011)

$$\eta_k = \eta \cdot \left[\sum_{i=0}^{k} \|\nabla f(x_i)\|^2\right]^{-1/2}$$
$$x_{k+1} = x_k - \eta_k \nabla f(x_k)$$

Significance of AdaGrad

AdaGrad was a baseline in the development of

- RMSProp (Tieleman, 2012)
- Adam (Kingma and Ba, 2014)

AdaGrad-type Methods

AdaGrad (scalar-stepsize vaiant)

(Duchi et al., 2011)

$$\eta_k = \eta \cdot \left[\sum_{i=0}^{k} \|\nabla f(x_i)\|^2\right]^{-1/2}$$
$$x_{k+1} = x_k - \eta_k \nabla f(x_k)$$

Significance of AdaGrad

AdaGrad was a baseline in the development of

- RMSProp (Tieleman, 2012)
- Adam (Kingma and Ba, 2014)

Universality of AdaGrad

(Orabona, 2023)

AdaGrad-type Methods

AdaGrad (scalar-stepsize vaiant)

(Duchi et al., 2011)

$$\eta_k = \eta \cdot \left[\sum_{i=0}^{k} \|\nabla f(x_i)\|^2\right]^{-1/2}$$
$$x_{k+1} = x_k - \eta_k \nabla f(x_k)$$

Significance of AdaGrad

AdaGrad was a baseline in the development of

- RMSProp (Tieleman, 2012)
- Adam (Kingma and Ba, 2014)

Universality of AdaGrad

(Orabona, 2023)

- Adapts to different levels of smoothness, i.e., ν-Hölder smoothness with $\nu \in [0, 1]$

AdaGrad-type Methods

AdaGrad (scalar-stepsize vaiant)

(Duchi et al., 2011)

$$\eta_k = \eta \cdot \left[\sum_{i=0}^{k} \|\nabla f(x_i)\|^2\right]^{-1/2}$$
$$x_{k+1} = x_k - \eta_k \nabla f(x_k)$$

Significance of AdaGrad

AdaGrad was a baseline in the development of

- RMSProp (Tieleman, 2012)
- Adam (Kingma and Ba, 2014)

Universality of AdaGrad

(Orabona, 2023)

- Adapts to different levels of smoothness, i.e., ν-Hölder smoothness with $\nu \in [0, 1]$
- With a single choice of the stepsize $\eta \propto \|x_0 - x^*\|$

AdaGrad-type Methods

AdaGrad (scalar-stepsize vaiant)

(Duchi et al., 2011)

$$\eta_k = \eta \cdot \left[\sum_{i=0}^{k} \|\nabla f(x_i)\|^2\right]^{-1/2}$$
$$x_{k+1} = x_k - \eta_k \nabla f(x_k)$$

Significance of AdaGrad

AdaGrad was a baseline in the development of

- RMSProp (Tieleman, 2012)
- Adam (Kingma and Ba, 2014)

Issues with AdaGrad:

Universality of AdaGrad

(Orabona, 2023)

- Adapts to different levels of smoothness, i.e., ν-Hölder smoothness with $\nu \in [0, 1]$
- With a single choice of the stepsize $\eta \propto \|x_0 - x^*\|$

AdaGrad-type Methods

AdaGrad (scalar-stepsize vaiant)

(Duchi et al., 2011)

$$\eta_k = \eta \cdot \left[\sum_{i=0}^{k} \|\nabla f(x_i)\|^2\right]^{-1/2}$$
$$x_{k+1} = x_k - \eta_k \nabla f(x_k)$$

Significance of AdaGrad

AdaGrad was a baseline in the development of

- RMSProp (Tieleman, 2012)
- Adam (Kingma and Ba, 2014)

Issues with AdaGrad:

- Stepsize η_k is always non-increasing

Universality of AdaGrad

(Orabona, 2023)

- Adapts to different levels of smoothness, i.e., ν-Hölder smoothness with $\nu \in [0, 1]$
- With a single choice of the stepsize $\eta \propto \|x_0 - x^*\|$

AdaGrad-type Methods

AdaGrad (scalar-stepsize vaiant)

(Duchi et al., 2011)

$$\eta_k = \eta \cdot \left[\sum_{i=0}^{k} \|\nabla f(x_i)\|^2\right]^{-1/2}$$
$$x_{k+1} = x_k - \eta_k \nabla f(x_k)$$

Significance of AdaGrad

AdaGrad was a baseline in the development of

- RMSProp (Tieleman, 2012)
- Adam (Kingma and Ba, 2014)

Issues with AdaGrad:

- Stepsize η_k is always non-increasing
- **Cannot truly adapt to the local curvature of** $f(x)$

Universality of AdaGrad

(Orabona, 2023)

- Adapts to different levels of smoothness, i.e., ν-Hölder smoothness with $\nu \in [0, 1]$
- With a single choice of the stepsize $\eta \propto \|x_0 - x^*\|$

Algorithms with Local Curvature Estimators

Local Curvature Estimator

$$\Lambda(x, x') = \frac{\|x - x'\|}{\|\nabla f(x) - \nabla f(x')\|}$$

Algorithms with Local Curvature Estimators

Local Curvature Estimator

$$\Lambda(x, x') = \frac{\|x - x'\|}{\|\nabla f(x) - \nabla f(x')\|}$$

- local estimator of the inverse gradient Lipschitz constant

Algorithms with Local Curvature Estimators

GRAAL (Malitsky, 2020)

$$\eta_k = \min\left\{(1+\gamma)\eta_{k-1}, \frac{\nu\Lambda^2(x_k, x_{k-1})}{\eta_{k-2}}\right\}$$

$$x_{k+1} = x_k - \eta_k \nabla f(\hat{x}_k), \quad \hat{x}_k = x_k + \theta(x_k - x_{k-1})$$

Local Curvature Estimator

$$\Lambda(x, x') = \frac{\|x - x'\|}{\|\nabla f(x) - \nabla f(x')\|}$$

- local estimator of the inverse gradient Lipschitz constant

Algorithms with Local Curvature Estimators

GRAAL (Malitsky, 2020)

$$\eta_k = \min\left\{(1+\gamma)\eta_{k-1}, \frac{\nu\Lambda^2(x_k, x_{k-1})}{\eta_{k-2}}\right\}$$
$$x_{k+1} = x_k - \eta_k \nabla f(\hat{x}_k), \quad \hat{x}_k = x_k + \theta(x_k - x_{k-1})$$

AdGD (Malitsky and Mishchenko, 2020)

$$\eta_k = \min\left\{\sqrt{1 + \frac{\eta_{k-1}}{\eta_{k-2}}}\eta_{k-1}, \nu\Lambda(x_k, x_{k-1})\right\}$$
$$x_{k+1} = x_k - \eta_k \nabla f(x_k)$$

Local Curvature Estimator

$$\Lambda(x, x') = \frac{\|x - x'\|}{\|\nabla f(x) - \nabla f(x')\|}$$

- local estimator of the inverse gradient Lipschitz constant

Algorithms with Local Curvature Estimators

GRAAL (Malitsky, 2020)

$$\eta_k = \min\left\{(1+\gamma)\eta_{k-1}, \frac{\nu\Lambda^2(x_k, x_{k-1})}{\eta_{k-2}}\right\}$$
$$x_{k+1} = x_k - \eta_k \nabla f(\hat{x}_k), \quad \hat{x}_k = x_k + \theta(x_k - x_{k-1})$$

AdGD (Malitsky and Mishchenko, 2020)

$$\eta_k = \min\left\{\sqrt{1 + \frac{\eta_{k-1}}{\eta_{k-2}}}\eta_{k-1}, \nu\Lambda(x_k, x_{k-1})\right\}$$
$$x_{k+1} = x_k - \eta_k \nabla f(x_k)$$

Local Curvature Estimator

$$\Lambda(x, x') = \frac{\|x - x'\|}{\|\nabla f(x) - \nabla f(x')\|}$$

- local estimator of the inverse gradient Lipschitz constant

Algorithms with Local Curvature Estimators

GRAAL (Malitsky, 2020)

$$\eta_k = \min\left\{(1+\gamma)\eta_{k-1}, \frac{\nu\Lambda^2(x_k, x_{k-1})}{\eta_{k-2}}\right\}$$
$$x_{k+1} = x_k - \eta_k \nabla f(\hat{x}_k), \quad \hat{x}_k = x_k + \theta(x_k - x_{k-1})$$

AdGD (Malitsky and Mishchenko, 2020)

$$\eta_k = \min\left\{\sqrt{1 + \frac{\eta_{k-1}}{\eta_{k-2}}}\eta_{k-1}, \nu\Lambda(x_k, x_{k-1})\right\}$$
$$x_{k+1} = x_k - \eta_k \nabla f(x_k)$$

Issues with GRAAL and AdGD:

Local Curvature Estimator

$$\Lambda(x, x') = \frac{\|x - x'\|}{\|\nabla f(x) - \nabla f(x')\|}$$

- local estimator of the inverse gradient Lipschitz constant

Algorithms with Local Curvature Estimators

GRAAL (Malitsky, 2020)

$$\eta_k = \min\left\{(1+\gamma)\eta_{k-1}, \frac{\nu\Lambda^2(x_k, x_{k-1})}{\eta_{k-2}}\right\}$$
$$x_{k+1} = x_k - \eta_k \nabla f(\hat{x}_k), \quad \hat{x}_k = x_k + \theta(x_k - x_{k-1})$$

AdGD (Malitsky and Mishchenko, 2020)

$$\eta_k = \min\left\{\sqrt{1 + \frac{\eta_{k-1}}{\eta_{k-2}}}\eta_{k-1}, \nu\Lambda(x_k, x_{k-1})\right\}$$
$$x_{k+1} = x_k - \eta_k \nabla f(x_k)$$

Issues with GRAAL and AdGD:

- **No Nesterov acceleration and $\mathcal{O}(1/k^2)$ rate**

Local Curvature Estimator

$$\Lambda(x, x') = \frac{\|x - x'\|}{\|\nabla f(x) - \nabla f(x')\|}$$

- local estimator of the inverse gradient Lipschitz constant

Algorithms with Local Curvature Estimators

GRAAL (Malitsky, 2020)

$$\eta_k = \min\left\{(1+\gamma)\eta_{k-1}, \frac{\nu\Lambda^2(x_k, x_{k-1})}{\eta_{k-2}}\right\}$$
$$x_{k+1} = x_k - \eta_k \nabla f(\hat{x}_k), \quad \hat{x}_k = x_k + \theta(x_k - x_{k-1})$$

AdGD (Malitsky and Mishchenko, 2020)

$$\eta_k = \min\left\{\sqrt{1 + \frac{\eta_{k-1}}{\eta_{k-2}}}\eta_{k-1}, \nu\Lambda(x_k, x_{k-1})\right\}$$
$$x_{k+1} = x_k - \eta_k \nabla f(x_k)$$

Issues with GRAAL and AdGD:

- **No Nesterov acceleration and $\mathcal{O}(1/k^2)$ rate**
- Slow $\mathcal{O}(1/k)$ rate for L-smooth convex functions

Local Curvature Estimator

$$\Lambda(x, x') = \frac{\|x - x'\|}{\|\nabla f(x) - \nabla f(x')\|}$$

- local estimator of the inverse gradient Lipschitz constant

Accelerated Algorithms with Local Curvature Estimators

Accelerated Algorithms with Local Curvature Estimators

AC-FGM (Li and Lan, 2025)

$$
\begin{aligned}
\eta_k &= \min\left\{ \frac{(k+1)}{k} \cdot \eta_k, \frac{k}{8} \cdot \Lambda(\bar{x}_{k-1}, \bar{x}_k) \right\} \\
\hat{x}_{k+1} &= x_k - \eta_k \nabla f(\bar{x}_k) \\
x_{k+1} &= \tau x_k + (1-\tau)\hat{x}_{k+1} \\
\bar{x}_{k+1} &= \alpha_{k+1}\hat{x}_{k+1} + (1-\alpha_{k+1})\bar{x}_k
\end{aligned}
$$

Accelerated Algorithms with Local Curvature Estimators

AC-FGM (Li and Lan, 2025)

$$\eta_k = \min\left\{\frac{(k+1)}{k}\cdot\eta_k, \frac{k}{8}\cdot\Lambda(\bar{x}_{k-1},\bar{x}_k)\right\}$$
$$\hat{x}_{k+1} = x_k - \eta_k\nabla f(\bar{x}_k)$$
$$x_{k+1} = \tau x_k + (1-\tau)\hat{x}_{k+1}$$
$$\bar{x}_{k+1} = \alpha_{k+1}\hat{x}_{k+1} + (1-\alpha_{k+1})\bar{x}_k$$

Local Curvature Estimator

$$\Lambda(x,x') = \frac{2\mathrm{D}_f(x;x')}{\|\nabla f(x) - \nabla f(x')\|}$$

Accelerated Algorithms with Local Curvature Estimators

AC-FGM (Li and Lan, 2025)

$$\eta_k = \min\left\{\frac{(k+1)}{k}\cdot\eta_k, \frac{k}{8}\cdot\Lambda(\bar{x}_{k-1},\bar{x}_k)\right\}$$
$$\hat{x}_{k+1} = x_k - \eta_k\nabla f(\bar{x}_k)$$
$$x_{k+1} = \tau x_k + (1-\tau)\hat{x}_{k+1}$$
$$\bar{x}_{k+1} = \alpha_{k+1}\hat{x}_{k+1} + (1-\alpha_{k+1})\bar{x}_k$$

Local Curvature Estimator

$$\Lambda(x,x') = \frac{2\mathrm{D}_f(x;x')}{\|\nabla f(x) - \nabla f(x')\|}$$

- local estimator of the inverse gradient Lipschitz constant for convex functions

Accelerated Algorithms with Local Curvature Estimators

AC-FGM (Li and Lan, 2025)

$$\eta_k = \min\left\{\frac{(k+1)}{k}\cdot\eta_k, \frac{k}{8}\cdot\Lambda(\bar{x}_{k-1},\bar{x}_k)\right\}$$
$$\hat{x}_{k+1} = x_k - \eta_k\nabla f(\bar{x}_k)$$
$$x_{k+1} = \tau x_k + (1-\tau)\hat{x}_{k+1}$$
$$\bar{x}_{k+1} = \alpha_{k+1}\hat{x}_{k+1} + (1-\alpha_{k+1})\bar{x}_k$$

Local Curvature Estimator

$$\Lambda(x,x') = \frac{2\mathrm{D}_f(x;x')}{\|\nabla f(x) - \nabla f(x')\|}$$

- local estimator of the inverse gradient Lipschitz constant for convex functions

Issues with AC-FGM:

Accelerated Algorithms with Local Curvature Estimators

AC-FGM (Li and Lan, 2025)

$$\eta_k = \min\left\{\frac{(k+1)}{k}\cdot\eta_k, \frac{k}{8}\cdot\Lambda(\bar{x}_{k-1},\bar{x}_k)\right\}$$
$$\hat{x}_{k+1} = x_k - \eta_k\nabla f(\bar{x}_k)$$
$$x_{k+1} = \tau x_k + (1-\tau)\hat{x}_{k+1}$$
$$\bar{x}_{k+1} = \alpha_{k+1}\hat{x}_{k+1} + (1-\alpha_{k+1})\bar{x}_k$$

Local Curvature Estimator

$$\Lambda(x,x') = \frac{2\mathrm{D}_f(x;x')}{\|\nabla f(x) - \nabla f(x')\|}$$

- local estimator of the inverse gradient Lipschitz constant for convex functions

Issues with AC-FGM:

- **Allows only sub-geometric growth of stepsize:**

$$\eta_k \le k\eta_0$$

Accelerated Algorithms with Local Curvature Estimators

AC-FGM (Li and Lan, 2025)

$$\eta_k = \min\left\{\frac{(k+1)}{k}\cdot\eta_k, \frac{k}{8}\cdot\Lambda(\bar{x}_{k-1},\bar{x}_k)\right\}$$
$$\hat{x}_{k+1} = x_k - \eta_k\nabla f(\bar{x}_k)$$
$$x_{k+1} = \tau x_k + (1-\tau)\hat{x}_{k+1}$$
$$\bar{x}_{k+1} = \alpha_{k+1}\hat{x}_{k+1} + (1-\alpha_{k+1})\bar{x}_k$$

Local Curvature Estimator

$$\Lambda(x,x') = \frac{2\mathrm{D}_f(x;x')}{\|\nabla f(x) - \nabla f(x')\|}$$

- local estimator of the inverse gradient Lipschitz constant for convex functions

Issues with AC-FGM:

- **Allows only sub-geometric growth of stepsize:**

$$\eta_k \le k\eta_0$$

- GRAAL allows geometric growth of stepsize:

$$\eta_k \le (1+\gamma)^k\eta_0$$

GRAAL Stepsize

$$\eta_k = \min\left\{(1+\gamma)\eta_{k-1}, \nu\Lambda_k^2/\eta_{k-2}\right\}$$

Accelerated Algorithms with Local Curvature Estimators

AC-FGM (Li and Lan, 2025)

$$\eta_k = \min\left\{\frac{(k+1)}{k}\cdot\eta_k, \frac{k}{8}\cdot\Lambda(\bar{x}_{k-1},\bar{x}_k)\right\}$$
$$\hat{x}_{k+1} = x_k - \eta_k\nabla f(\bar{x}_k)$$
$$x_{k+1} = \tau x_k + (1-\tau)\hat{x}_{k+1}$$
$$\bar{x}_{k+1} = \alpha_{k+1}\hat{x}_{k+1} + (1-\alpha_{k+1})\bar{x}_k$$

Local Curvature Estimator

$$\Lambda(x,x') = \frac{2\mathrm{D}_f(x;x')}{\|\nabla f(x) - \nabla f(x')\|}$$

- local estimator of the inverse gradient Lipschitz constant for convex functions

Issues with AC-FGM:

- **Allows only sub-geometric growth of stepsize:**

$$\eta_k \le k\eta_0$$

- GRAAL allows geometric growth of stepsize:

$$\eta_k \le (1+\gamma)^k\eta_0$$

- **Sub-geometric adaptation speed is insufficient**

GRAAL Stepsize

$$\eta_k = \min\left\{(1+\gamma)\eta_{k-1}, \nu\Lambda_k^2/\eta_{k-2}\right\}$$

Accelerated Algorithms with Local Curvature Estimators

AC-FGM (Li and Lan, 2025)

$$\eta_k = \min\left\{\frac{(k+1)}{k}\cdot\eta_k, \frac{k}{8}\cdot\Lambda(\bar{x}_{k-1},\bar{x}_k)\right\}$$
$$\hat{x}_{k+1} = x_k - \eta_k\nabla f(\bar{x}_k)$$
$$x_{k+1} = \tau x_k + (1-\tau)\hat{x}_{k+1}$$
$$\bar{x}_{k+1} = \alpha_{k+1}\hat{x}_{k+1} + (1-\alpha_{k+1})\bar{x}_k$$

Local Curvature Estimator

$$\Lambda(x,x') = \frac{2\mathrm{D}_f(x;x')}{\|\nabla f(x) - \nabla f(x')\|}$$

- local estimator of the inverse gradient Lipschitz constant for convex functions

Issues with AC-FGM:

- **Allows only sub-geometric growth of stepsize:**

$$\eta_k \le k\eta_0$$

- GRAAL allows geometric growth of stepsize:

$$\eta_k \le (1+\gamma)^k\eta_0$$

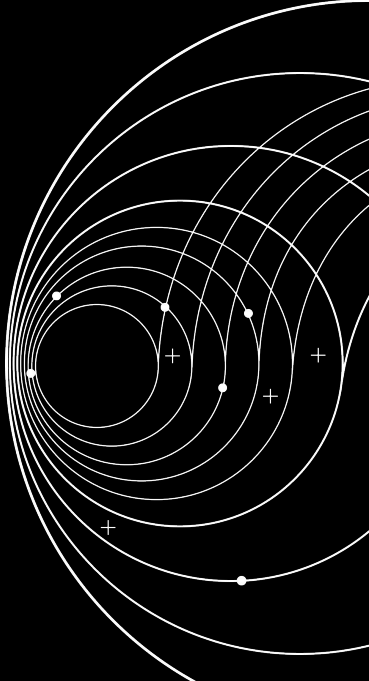
- **Sub-geometric adaptation speed is insufficient**
- **AdaNAG (Suh and Ma, 2025) has similar issues**

GRAAL Stepsize

$$\eta_k = \min\left\{(1+\gamma)\eta_{k-1}, \nu\Lambda_k^2/\eta_{k-2}\right\}$$

Yandex Research

GRAAL with Nesterov Acceleration



GRAAL Extrapolation and Nesterov Acceleration

Local Curvature Estimator

$$\Lambda(x, x') = \frac{2\mathrm{D}_f(x; x')}{\|\nabla f(x) - \nabla f(x')\|}$$

GRAAL Extrapolation and Nesterov Acceleration

GRAAL (Malitsky, 2020)

Local Curvature Estimator

$$\Lambda(x, x') = \frac{2\mathrm{D}_f(x; x')}{\|\nabla f(x) - \nabla f(x')\|}$$

GRAAL Extrapolation and Nesterov Acceleration

GRAAL (Malitsky, 2020)

- gradient step: $x_{k+1} = x_k - \eta_k \nabla f(\hat{x}_k)$

Local Curvature Estimator

$$\Lambda(x, x') = \frac{2\mathrm{D}_f(x; x')}{\|\nabla f(x) - \nabla f(x')\|}$$

GRAAL Extrapolation and Nesterov Acceleration

GRAAL (Malitsky, 2020)

- gradient step: $x_{k+1} = x_k - \eta_k \nabla f(\hat{x}_k)$
- extrapolation: $\hat{x}_{k+1} = x_{k+1} + \theta(x_{k+1} - x_k)$

Local Curvature Estimator

$$\Lambda(x, x') = \frac{2\mathrm{D}_f(x; x')}{\|\nabla f(x) - \nabla f(x')\|}$$

GRAAL Extrapolation and Nesterov Acceleration

GRAAL (Malitsky, 2020)

- gradient step: $x_{k+1} = x_k - \eta_k \nabla f(\hat{x}_k)$
- extrapolation: $\hat{x}_{k+1} = x_{k+1} + \theta(x_{k+1} - x_k)$

Acceleration Framework (Borodich et al., 2025)

Local Curvature Estimator

$$\Lambda(x, x') = \frac{2\mathrm{D}_f(x; x')}{\|\nabla f(x) - \nabla f(x')\|}$$

GRAAL Extrapolation and Nesterov Acceleration

GRAAL (Malitsky, 2020)

- gradient step: $x_{k+1} = x_k - \eta_k \nabla f(\hat{x}_k)$
- extrapolation: $\hat{x}_{k+1} = x_{k+1} + \theta(x_{k+1} - x_k)$

Acceleration Framework (Borodich et al., 2025)

- Replace objective function $f(x)$ with

$$f_k(x) = \alpha_k^{-1} f(\alpha_k x + (1 - \alpha_k)\bar{x}_k)$$

where $\bar{x}_k \in \mathbb{R}^d$ and $\alpha_k \in (0, 1]$

Local Curvature Estimator

$$\Lambda(x, x') = \frac{2\mathrm{D}_f(x; x')}{\|\nabla f(x) - \nabla f(x')\|}$$

GRAAL Extrapolation and Nesterov Acceleration

GRAAL (Malitsky, 2020)

- gradient step: $x_{k+1} = x_k - \eta_k \nabla f(\hat{x}_k)$
- extrapolation: $\hat{x}_{k+1} = x_{k+1} + \theta(x_{k+1} - x_k)$

Acceleration Framework (Borodich et al., 2025)

- Replace objective function $f(x)$ with

$$f_k(x) = \alpha_k^{-1} f(\alpha_k x + (1 - \alpha_k)\bar{x}_k)$$

  where $\bar{x}_k \in \mathbb{R}^d$ and $\alpha_k \in (0, 1]$
- Choose $\bar{x}_{k+1} = \alpha_k \hat{x}_k + (1 - \alpha_k)\bar{x}_k$

Local Curvature Estimator

$$\Lambda(x, x') = \frac{2\mathrm{D}_f(x; x')}{\|\nabla f(x) - \nabla f(x')\|}$$

GRAAL Extrapolation and Nesterov Acceleration

GRAAL (Malitsky, 2020)

- gradient step: $x_{k+1} = x_k - \eta_k \nabla f(\hat{x}_k)$
- extrapolation: $\hat{x}_{k+1} = x_{k+1} + \theta(x_{k+1} - x_k)$

Local Curvature Estimator

$$\Lambda(x, x') = \frac{2\mathrm{D}_f(x; x')}{\|\nabla f(x) - \nabla f(x')\|}$$

Acceleration Framework (Borodich et al., 2025)

- Replace objective function $f(x)$ with

  $$f_k(x) = \alpha_k^{-1} f(\alpha_k x + (1 - \alpha_k)\bar{x}_k)$$

  where $\bar{x}_k \in \mathbb{R}^d$ and $\alpha_k \in (0, 1]$
- Choose $\bar{x}_{k+1} = \alpha_k \hat{x}_k + (1 - \alpha_k)\bar{x}_k$

GRAAL + Acceleration

$$\bar{x}_{k+1} = \alpha_k \hat{x}_k + (1 - \alpha_k)\bar{x}_k, \qquad \hat{x}_{k+1} = x_{k+1} + \theta(x_{k+1} - x_k), \qquad x_{k+1} = x_k - \eta_k \nabla f(\bar{x}_{k+1})$$

Problem: How to Choose η_k and α_k?

Local Curvature Estimator

$$\Lambda(x, x') = \frac{2\mathrm{D}_f(x; x')}{\|\nabla f(x) - \nabla f(x')\|}$$

GRAAL + Acceleration

$$\bar{x}_{k+1} = \alpha_k \hat{x}_k + (1 - \alpha_k)\bar{x}_k$$
$$\hat{x}_{k+1} = x_{k+1} + \theta(x_{k+1} - x_k)$$
$$x_{k+1} = x_k - \eta_k \nabla f(\bar{x}_{k+1})$$

Problem: How to Choose η_k and α_k?

Restrictions on η_k and α_k (from analysis)

Local Curvature Estimator

$$\Lambda(x, x') = \frac{2\mathrm{D}_f(x; x')}{\|\nabla f(x) - \nabla f(x')\|}$$

GRAAL + Acceleration

$$\begin{aligned}
\bar{x}_{k+1} &= \alpha_k \hat{x}_k + (1 - \alpha_k)\bar{x}_k \\
\hat{x}_{k+1} &= x_{k+1} + \theta(x_{k+1} - x_k) \\
x_{k+1} &= x_k - \eta_k \nabla f(\bar{x}_{k+1})
\end{aligned}$$

Problem: How to Choose η_k and α_k?

Restrictions on η_k and α_k (from analysis)

- Upper bound on the stepsize η_k:

$$\eta_k \leq \frac{\nu(1-\alpha_k)}{\alpha_k} \cdot \Lambda(\bar{x}_k, \bar{x}_{k+1})$$

Local Curvature Estimator

$$\Lambda(x, x') = \frac{2\mathrm{D}_f(x; x')}{\|\nabla f(x) - \nabla f(x')\|}$$

GRAAL + Acceleration

$$\bar{x}_{k+1} = \alpha_k \hat{x}_k + (1-\alpha_k)\bar{x}_k$$
$$\hat{x}_{k+1} = x_{k+1} + \theta(x_{k+1} - x_k)$$
$$x_{k+1} = x_k - \eta_k \nabla f(\bar{x}_{k+1})$$

Problem: How to Choose η_k and α_k?

Restrictions on η_k and α_k (from analysis)

- Upper bound on the stepsize η_k:

$$\eta_k \leq \frac{\nu(1-\alpha_k)}{\alpha_k} \cdot \Lambda(\bar{x}_k, \bar{x}_{k+1})$$

- Parameters η_k and α_k must satisfy the following:

$$\frac{\eta_k}{\alpha_k} \leq \frac{\eta_{k-1}}{\alpha_{k-1}} + \eta_k$$

Local Curvature Estimator

$$\Lambda(x, x') = \frac{2\mathrm{D}_f(x; x')}{\|\nabla f(x) - \nabla f(x')\|}$$

GRAAL + Acceleration

$$\bar{x}_{k+1} = \alpha_k \hat{x}_k + (1-\alpha_k)\bar{x}_k$$
$$\hat{x}_{k+1} = x_{k+1} + \theta(x_{k+1} - x_k)$$
$$x_{k+1} = x_k - \eta_k \nabla f(\bar{x}_{k+1})$$

Problem: How to Choose η_k and α_k?

Restrictions on η_k and α_k (from analysis)

- Upper bound on the stepsize η_k:

$$\eta_k \leq \frac{\nu(1-\alpha_k)}{\alpha_k} \cdot \Lambda(\bar{x}_k, \bar{x}_{k+1})$$

- Parameters η_k and α_k must satisfy the following:

$$\frac{\eta_k}{\alpha_k} \leq \frac{\eta_{k-1}}{\alpha_{k-1}} + \eta_k$$

Issues with parameters η_k and α_k:

Local Curvature Estimator

$$\Lambda(x, x') = \frac{2\mathrm{D}_f(x; x')}{\|\nabla f(x) - \nabla f(x')\|}$$

GRAAL + Acceleration

$$\bar{x}_{k+1} = \alpha_k \hat{x}_k + (1-\alpha_k)\bar{x}_k$$
$$\hat{x}_{k+1} = x_{k+1} + \theta(x_{k+1} - x_k)$$
$$x_{k+1} = x_k - \eta_k \nabla f(\bar{x}_{k+1})$$

Problem: How to Choose η_k and α_k?

Restrictions on η_k and α_k (from analysis)

- Upper bound on the stepsize η_k:

$$\eta_k \leq \frac{\nu(1-\alpha_k)}{\alpha_k} \cdot \Lambda(\bar{x}_k, \bar{x}_{k+1})$$

- Parameters η_k and α_k must satisfy the following:

$$\frac{\eta_k}{\alpha_k} \leq \frac{\eta_{k-1}}{\alpha_{k-1}} + \eta_k$$

Issues with parameters η_k and α_k:

- **Not possible to choose α_k adaptively**
  cyclic dependency: $\eta_k \leftarrow \bar{x}_{k+1} \leftarrow \alpha_k \leftarrow \eta_k$

Local Curvature Estimator

$$\Lambda(x, x') = \frac{2\mathrm{D}_f(x; x')}{\|\nabla f(x) - \nabla f(x')\|}$$

GRAAL + Acceleration

$$\bar{x}_{k+1} = \alpha_k \hat{x}_k + (1-\alpha_k)\bar{x}_k$$
$$\hat{x}_{k+1} = x_{k+1} + \theta(x_{k+1} - x_k)$$
$$x_{k+1} = x_k - \eta_k \nabla f(\bar{x}_{k+1})$$

Problem: How to Choose η_k and α_k?

Restrictions on η_k and α_k (from analysis)

- Upper bound on the stepsize η_k:

$$\eta_k \leq \frac{\nu(1-\alpha_k)}{\alpha_k} \cdot \Lambda(\bar{x}_k, \bar{x}_{k+1})$$

- Parameters η_k and α_k must satisfy the following:

$$\frac{\eta_k}{\alpha_k} \leq \frac{\eta_{k-1}}{\alpha_{k-1}} + \eta_k$$

Issues with parameters η_k and α_k:

- **Not possible to choose α_k adaptively**
  cyclic dependency: $\eta_k \leftarrow \bar{x}_{k+1} \leftarrow \alpha_k \leftarrow \eta_k$

Local Curvature Estimator

$$\Lambda(x, x') = \frac{2\mathrm{D}_f(x; x')}{\|\nabla f(x) - \nabla f(x')\|}$$

GRAAL + Acceleration

$$\bar{x}_{k+1} = \alpha_k \hat{x}_k + (1-\alpha_k)\bar{x}_k$$
$$\hat{x}_{k+1} = x_{k+1} + \theta(x_{k+1} - x_k)$$
$$x_{k+1} = x_k - \eta_k \nabla f(\bar{x}_{k+1})$$

Problem: How to Choose η_k and α_k?

Restrictions on η_k and α_k (from analysis)

- Upper bound on the stepsize η_k:

$$\eta_k \leq \frac{\nu(1-\alpha_k)}{\alpha_k} \cdot \Lambda(\bar{x}_k, \bar{x}_{k+1})$$

- Parameters η_k and α_k must satisfy the following:

$$\frac{\eta_k}{\alpha_k} \leq \frac{\eta_{k-1}}{\alpha_{k-1}} + \eta_k$$

Issues with parameters η_k and α_k:

- **Not possible to choose α_k adaptively**
  cyclic dependency: $\eta_k \leftarrow \bar{x}_{k+1} \leftarrow \alpha_k \leftarrow \eta_k$

Local Curvature Estimator

$$\Lambda(x, x') = \frac{2\mathrm{D}_f(x; x')}{\|\nabla f(x) - \nabla f(x')\|}$$

GRAAL + Acceleration

$$\bar{x}_{k+1} = \alpha_k \hat{x}_k + (1-\alpha_k)\bar{x}_k$$
$$\hat{x}_{k+1} = x_{k+1} + \theta(x_{k+1} - x_k)$$
$$x_{k+1} = x_k - \eta_k \nabla f(\bar{x}_{k+1})$$

Problem: How to Choose η_k and α_k?

Restrictions on η_k and α_k (from analysis)

- Upper bound on the stepsize η_k:

$$\eta_k \leq \frac{\nu(1-\alpha_k)}{\alpha_k} \cdot \Lambda(\bar{x}_k, \bar{x}_{k+1})$$

- Parameters η_k and α_k must satisfy the following:

$$\frac{\eta_k}{\alpha_k} \leq \frac{\eta_{k-1}}{\alpha_{k-1}} + \eta_k$$

Issues with parameters η_k and α_k:

- **Not possible to choose α_k adaptively**
  cyclic dependency: $\eta_k \leftarrow \bar{x}_{k+1} \leftarrow \alpha_k \leftarrow \eta_k$

Local Curvature Estimator

$$\Lambda(x, x') = \frac{2\mathrm{D}_f(x; x')}{\|\nabla f(x) - \nabla f(x')\|}$$

GRAAL + Acceleration

$$\bar{x}_{k+1} = \alpha_k \hat{x}_k + (1-\alpha_k)\bar{x}_k$$
$$\hat{x}_{k+1} = x_{k+1} + \theta(x_{k+1} - x_k)$$
$$x_{k+1} = x_k - \eta_k \nabla f(\bar{x}_{k+1})$$

Problem: How to Choose η_k and α_k?

Restrictions on η_k and α_k (from analysis)

- Upper bound on the stepsize η_k:

$$\eta_k \leq \frac{\nu(1-\alpha_k)}{\alpha_k} \cdot \Lambda(\bar{x}_k, \bar{x}_{k+1})$$

- Parameters η_k and α_k must satisfy the following:

$$\frac{\eta_k}{\alpha_k} \leq \frac{\eta_{k-1}}{\alpha_{k-1}} + \eta_k$$

Issues with parameters η_k and α_k:

- **Not possible to choose α_k adaptively**
  cyclic dependency: $\eta_k \leftarrow \bar{x}_{k+1} \leftarrow \alpha_k \leftarrow \eta_k$
- Li and Lan (2025) choose $\alpha_k = 2/(k+2)$ in AC-FGM
  non-adaptive, implies sub-geometric growth of η_k

Local Curvature Estimator

$$\Lambda(x, x') = \frac{2\mathrm{D}_f(x; x')}{\|\nabla f(x) - \nabla f(x')\|}$$

GRAAL + Acceleration

$$\bar{x}_{k+1} = \alpha_k \hat{x}_k + (1-\alpha_k)\bar{x}_k$$
$$\hat{x}_{k+1} = x_{k+1} + \theta(x_{k+1} - x_k)$$
$$x_{k+1} = x_k - \eta_k \nabla f(\bar{x}_{k+1})$$

Solution: Additional Coupling Step

Local Curvature Estimator

$$\Lambda(x, x') = \frac{2\mathrm{D}_f(x; x')}{\|\nabla f(x) - \nabla f(x')\|}$$

GRAAL + Acceleration

$$\bar{x}_{k+1} = \alpha_k \hat{x}_k + (1 - \alpha_k)\bar{x}_k$$
$$\hat{x}_{k+1} = x_{k+1} + \theta(x_{k+1} - x_k)$$
$$x_{k+1} = x_k - \eta_k \nabla f(\bar{x}_{k+1})$$

Solution: Additional Coupling Step

Additional Coupling Step

Local Curvature Estimator

$$\Lambda(x, x') = \frac{2\mathrm{D}_f(x; x')}{\|\nabla f(x) - \nabla f(x')\|}$$

GRAAL + Acceleration

$$\bar{x}_{k+1} = \alpha_k \hat{x}_k + (1 - \alpha_k) x_k$$
$$\hat{x}_{k+1} = x_{k+1} + \theta(x_{k+1} - x_k)$$
$$x_{k+1} = x_k - \eta_k \nabla f(\bar{x}_{k+1})$$

Solution: Additional Coupling Step

Additional Coupling Step

- Update $\bar{x}_{k+1}$ as follows:

$$\bar{x}_{k+1} = \beta_k \tilde{x}_k + (1 - \beta_k)\bar{x}_k$$
$$\tilde{x}_k = \alpha_k \hat{x}_k + (1 - \alpha_k)\bar{x}_k$$

where $\alpha_k, \beta_k \in (0, 1]$

Local Curvature Estimator

$$\Lambda(x, x') = \frac{2\mathrm{D}_f(x; x')}{\|\nabla f(x) - \nabla f(x')\|}$$

GRAAL + Acceleration

$$\bar{x}_{k+1} = \alpha_k \hat{x}_k + (1 - \alpha_k)\bar{x}_k$$
$$\hat{x}_{k+1} = x_{k+1} + \theta(x_{k+1} - x_k)$$
$$x_{k+1} = x_k - \eta_k \nabla f(\bar{x}_{k+1})$$

Solution: Additional Coupling Step

Additional Coupling Step

- Update $\bar{x}_{k+1}$ as follows:

$$\bar{x}_{k+1} = \beta_k \tilde{x}_k + (1 - \beta_k)\bar{x}_k$$
$$\tilde{x}_k = \alpha_k \hat{x}_k + (1 - \alpha_k)\bar{x}_k$$

  where $\alpha_k, \beta_k \in (0, 1]$

- replace $\nabla f(\bar{x}_{k+1}) \to \nabla f(\tilde{x}_k)$

Local Curvature Estimator

$$\Lambda(x, x') = \frac{2\mathrm{D}_f(x; x')}{\|\nabla f(x) - \nabla f(x')\|}$$

GRAAL + Acceleration

$$\bar{x}_{k+1} = \alpha_k \hat{x}_k + (1 - \alpha_k)\bar{x}_k$$
$$\hat{x}_{k+1} = x_{k+1} + \theta(x_{k+1} - x_k)$$
$$x_{k+1} = x_k - \eta_k \nabla f(\bar{x}_{k+1})$$

Solution: Additional Coupling Step

Additional Coupling Step

- Update $\bar{x}_{k+1}$ as follows:

$$\bar{x}_{k+1} = \beta_k \tilde{x}_k + (1 - \beta_k)\bar{x}_k$$
$$\tilde{x}_k = \alpha_k \hat{x}_k + (1 - \alpha_k)\bar{x}_k$$

  where $\alpha_k, \beta_k \in (0, 1]$
- replace $\nabla f(\bar{x}_{k+1}) \to \nabla f(\tilde{x}_k)$

Local Curvature Estimator

$$\Lambda(x, x') = \frac{2\mathrm{D}_f(x; x')}{\|\nabla f(x) - \nabla f(x')\|}$$

Accelerated GRAAL

$$\tilde{x}_k = \alpha_k \hat{x}_k + (1 - \alpha_k)\bar{x}_k$$
$$\bar{x}_{k+1} = \beta_k \tilde{x}_k + (1 - \beta_k)\bar{x}_k$$
$$\hat{x}_{k+1} = x_{k+1} + \theta(x_{k+1} - x_k)$$
$$x_{k+1} = x_k - \eta_k \nabla f(\tilde{x}_k)$$

Solution: Additional Coupling Step

Choosing α_k, β_k, and η_k (from analysis)

Local Curvature Estimator

$$\Lambda(x, x') = \frac{2\mathrm{D}_f(x; x')}{\|\nabla f(x) - \nabla f(x')\|}$$

Accelerated GRAAL

$$\begin{aligned}
\tilde{x}_k &= \alpha_k \hat{x}_k + (1 - \alpha_k)\bar{x}_k \\
\bar{x}_{k+1} &= \beta_k \tilde{x}_k + (1 - \beta_k)\bar{x}_k \\
\hat{x}_{k+1} &= x_{k+1} + \theta(x_{k+1} - x_k) \\
x_{k+1} &= x_k - \eta_k \nabla f(\tilde{x}_k)
\end{aligned}$$

Solution: Additional Coupling Step

Choosing α_k, β_k, and η_k (from analysis)

- Parameters α_k, β_k, and η_k satisfy the following:

$$\frac{\eta_k}{\alpha_k \beta_k} \leq \frac{\eta_{k-1}}{\alpha_{k-1}\beta_{k-1}} + \eta_k \leq \cdots \leq H_k = \sum_{i=0}^{k} \eta_i$$

Local Curvature Estimator

$$\Lambda(x, x') = \frac{2\mathrm{D}_f(x; x')}{\|\nabla f(x) - \nabla f(x')\|}$$

Accelerated GRAAL

$$\begin{aligned}
\tilde{x}_k &= \alpha_k \hat{x}_k + (1 - \alpha_k)\bar{x}_k \\
\bar{x}_{k+1} &= \beta_k \tilde{x}_k + (1 - \beta_k)\bar{x}_k \\
\hat{x}_{k+1} &= x_{k+1} + \theta(x_{k+1} - x_k) \\
x_{k+1} &= x_k - \eta_k \nabla f(\tilde{x}_k)
\end{aligned}$$

Solution: Additional Coupling Step

Choosing α_k, β_k, and η_k (from analysis)

- Parameters α_k, β_k, and η_k satisfy the following:

$$\frac{\eta_k}{\alpha_k \beta_k} \leq \frac{\eta_{k-1}}{\alpha_{k-1}\beta_{k-1}} + \eta_k \leq \cdots \leq H_k = \sum_{i=0}^{k} \eta_i$$

- Hence, choose $\beta_k = \eta_k / (\alpha_k H_k)$

Local Curvature Estimator

$$\Lambda(x, x') = \frac{2\mathrm{D}_f(x; x')}{\|\nabla f(x) - \nabla f(x')\|}$$

Accelerated GRAAL

$$\begin{aligned}
\tilde{x}_k &= \alpha_k \hat{x}_k + (1 - \alpha_k)\bar{x}_k \\
\bar{x}_{k+1} &= \beta_k \tilde{x}_k + (1 - \beta_k)\bar{x}_k \\
\hat{x}_{k+1} &= x_{k+1} + \theta(x_{k+1} - x_k) \\
x_{k+1} &= x_k - \eta_k \nabla f(\tilde{x}_k)
\end{aligned}$$

Solution: Additional Coupling Step

Choosing α_k, β_k, and η_k (from analysis)

- Parameters α_k, β_k, and η_k satisfy the following:

$$\frac{\eta_k}{\alpha_k \beta_k} \leq \frac{\eta_{k-1}}{\alpha_{k-1}\beta_{k-1}} + \eta_k \leq \cdots \leq H_k = \sum_{i=0}^{k} \eta_i$$

- Hence, choose $\beta_k = \eta_k / (\alpha_k H_k)$
- How to choose $\alpha_k \in (0, 1]$ and ensure $\beta_k \in (0, 1]$?

Local Curvature Estimator

$$\Lambda(x, x') = \frac{2\mathrm{D}_f(x; x')}{\|\nabla f(x) - \nabla f(x')\|}$$

Accelerated GRAAL

$$\begin{aligned}
\tilde{x}_k &= \alpha_k \hat{x}_k + (1 - \alpha_k)\bar{x}_k \\
\bar{x}_{k+1} &= \beta_k \tilde{x}_k + (1 - \beta_k)\bar{x}_k \\
\hat{x}_{k+1} &= x_{k+1} + \theta(x_{k+1} - x_k) \\
x_{k+1} &= x_k - \eta_k \nabla f(\tilde{x}_k)
\end{aligned}$$

Solution: Additional Coupling Step

Choosing α_k, β_k, and η_k (from analysis)

- Parameters α_k, β_k, and η_k satisfy the following:

$$\frac{\eta_k}{\alpha_k \beta_k} \leq \frac{\eta_{k-1}}{\alpha_{k-1}\beta_{k-1}} + \eta_k \leq \cdots \leq H_k = \sum_{i=0}^{k} \eta_i$$

- Hence, choose $\beta_k = \eta_k / (\alpha_k H_k)$
- How to choose $\alpha_k \in (0, 1]$ and ensure $\beta_k \in (0, 1]$?
- Restrictions on stepsize η_k:

$$\eta_k \leq (1+\gamma)\eta_{k-1}$$

$$\eta_k \leq \frac{\nu H_{k-2}}{\eta_{k-2}} \cdot \min\{\Lambda(\bar{x}_k, \tilde{x}_k), \Lambda(\bar{x}_k, \tilde{x}_{k-1})\}$$

Local Curvature Estimator

$$\Lambda(x, x') = \frac{2\mathrm{D}_f(x; x')}{\|\nabla f(x) - \nabla f(x')\|}$$

Accelerated GRAAL

$$\begin{aligned}
\tilde{x}_k &= \alpha_k \hat{x}_k + (1-\alpha_k)\bar{x}_k \\
\bar{x}_{k+1} &= \beta_k \tilde{x}_k + (1-\beta_k)\bar{x}_k \\
\hat{x}_{k+1} &= x_{k+1} + \theta(x_{k+1} - x_k) \\
x_{k+1} &= x_k - \eta_k \nabla f(\tilde{x}_k)
\end{aligned}$$

Solution: Additional Coupling Step

Choosing α_k, β_k, and η_k (from analysis)

- Parameters α_k, β_k, and η_k satisfy the following:

$$\frac{\eta_k}{\alpha_k\beta_k} \le \frac{\eta_{k-1}}{\alpha_{k-1}\beta_{k-1}} + \eta_k \le \cdots \le H_k = \sum_{i=0}^{k} \eta_i$$

- Hence, choose $\beta_k = \eta_k/(\alpha_k H_k)$
- How to choose $\alpha_k \in (0,1]$ and ensure $\beta_k \in (0,1]$?
- Restrictions on stepsize η_k:

$$\eta_k \le (1+\gamma)\eta_{k-1}$$

$$\eta_k \le \frac{\nu H_{k-2}}{\eta_{k-2}} \cdot \min\{\Lambda(\bar{x}_k, \tilde{x}_k), \Lambda(\bar{x}_k, \tilde{x}_{k-1})\}$$

- Choose $\alpha_k \in (0,1]$ adaptively as follows:

$$\alpha_k = \frac{(1+\gamma)\eta_{k-1}}{H_{k-1} + (1+\gamma)\eta_{k-1}} \quad \Rightarrow \quad \beta_k \in (0,1]$$

Local Curvature Estimator

$$\Lambda(x, x') = \frac{2\mathrm{D}_f(x; x')}{\|\nabla f(x) - \nabla f(x')\|}$$

Accelerated GRAAL

$$\tilde{x}_k = \alpha_k \hat{x}_k + (1-\alpha_k)\bar{x}_k$$
$$\bar{x}_{k+1} = \beta_k \tilde{x}_k + (1-\beta_k)\bar{x}_k$$
$$\hat{x}_{k+1} = x_{k+1} + \theta(x_{k+1} - x_k)$$
$$x_{k+1} = x_k - \eta_k \nabla f(\tilde{x}_k)$$

Solution: Additional Coupling Step

Choosing α_k, β_k, and η_k (from analysis)

- Parameters α_k, β_k, and η_k satisfy the following:

$$\frac{\eta_k}{\alpha_k\beta_k} \le \frac{\eta_{k-1}}{\alpha_{k-1}\beta_{k-1}} + \eta_k \le \cdots \le H_k = \sum_{i=0}^{k} \eta_i$$

- Hence, choose $\beta_k = \eta_k/(\alpha_k H_k)$
- How to choose $\alpha_k \in (0,1]$ and ensure $\beta_k \in (0,1]$?
- Restrictions on stepsize η_k:

$$\eta_k \le (1+\gamma)\eta_{k-1}$$

$$\eta_k \le \frac{\nu H_{k-2}}{\eta_{k-2}} \cdot \min\{\Lambda(\bar{x}_k, \tilde{x}_k), \Lambda(\bar{x}_k, \tilde{x}_{k-1})\}$$

- Choose $\alpha_k \in (0,1]$ adaptively as follows:

$$\alpha_k = \frac{(1+\gamma)\eta_{k-1}}{H_{k-1} + (1+\gamma)\eta_{k-1}} \quad \Rightarrow \quad \beta_k \in (0,1]$$

- **Implementable: no cyclic dependencies**

Local Curvature Estimator

$$\Lambda(x, x') = \frac{2\mathrm{D}_f(x; x')}{\|\nabla f(x) - \nabla f(x')\|}$$

Accelerated GRAAL

$$\begin{aligned}
\tilde{x}_k &= \alpha_k \hat{x}_k + (1-\alpha_k)\bar{x}_k \\
\bar{x}_{k+1} &= \beta_k \tilde{x}_k + (1-\beta_k)\bar{x}_k \\
\hat{x}_{k+1} &= x_{k+1} + \theta(x_{k+1} - x_k) \\
x_{k+1} &= x_k - \eta_k \nabla f(\tilde{x}_k)
\end{aligned}$$

Accelerated GRAAL: Convergence Analysis

Accelerated GRAAL Stepsize

$$\eta_k = \min\left\{(1+\gamma)\eta_{k-1}, \frac{\nu H_{k-2}\Lambda_k}{\eta_{k-2}}\right\}$$

$$H_k = \sum_{i=0}^{k}\eta_i, \quad \Lambda_k = \min\{\Lambda(\bar{x}_k, \tilde{x}_k), \Lambda(\bar{x}_k, \tilde{x}_{k-1})\}$$

Local Curvature Estimator

$$\Lambda(x, x') = \frac{2\mathrm{D}_f(x; x')}{\|\nabla f(x) - \nabla f(x')\|}$$

Accelerated GRAAL

$$\begin{aligned}
\tilde{x}_k &= \alpha_k \hat{x}_k + (1-\alpha_k)\bar{x}_k \\
\bar{x}_{k+1} &= \beta_k \tilde{x}_k + (1-\beta_k)\bar{x}_k \\
\hat{x}_{k+1} &= x_{k+1} + \theta(x_{k+1} - x_k) \\
x_{k+1} &= x_k - \eta_k \nabla f(\tilde{x}_k)
\end{aligned}$$

Accelerated GRAAL: Convergence Analysis

Accelerated GRAAL Stepsize

$$\eta_k = \min\left\{(1+\gamma)\eta_{k-1}, \frac{\nu H_{k-2}\Lambda_k}{\eta_{k-2}}\right\}$$

$$H_k = \sum_{i=0}^{k}\eta_i, \quad \Lambda_k = \min\{\Lambda(\bar{x}_k, \tilde{x}_k), \Lambda(\bar{x}_k, \tilde{x}_{k-1})\}$$

Theorem 1 (Borodich and Kovalev, 2025)

There exist universal constants $\theta, \gamma, \nu > 0$ such that

$$\begin{aligned}\tfrac{1}{2}\|x_K - x\|^2 + H_{K-1}(f(\bar{x}_K) - f(x)) \\ \leq \tfrac{1}{2}\|x_0 - x\|^2 + \tfrac{(1+\gamma\theta)}{2}\eta_0^2\|\nabla f(x_0)\|^2\end{aligned}$$

Local Curvature Estimator

$$\Lambda(x, x') = \frac{2\mathrm{D}_f(x; x')}{\|\nabla f(x) - \nabla f(x')\|}$$

Accelerated GRAAL

$$\begin{aligned}\tilde{x}_k &= \alpha_k\hat{x}_k + (1-\alpha_k)\bar{x}_k \\ \bar{x}_{k+1} &= \beta_k\tilde{x}_k + (1-\beta_k)\bar{x}_k \\ \hat{x}_{k+1} &= x_{k+1} + \theta(x_{k+1} - x_k) \\ x_{k+1} &= x_k - \eta_k\nabla f(\tilde{x}_k)\end{aligned}$$

Accelerated GRAAL: Convergence Analysis

Accelerated GRAAL Stepsize

$$\eta_k = \min\left\{(1+\gamma)\eta_{k-1}, \frac{\nu H_{k-2}\Lambda_k}{\eta_{k-2}}\right\}$$

$$H_k = \sum_{i=0}^{k}\eta_i, \quad \Lambda_k = \min\{\Lambda(\bar{x}_k, \tilde{x}_k), \Lambda(\bar{x}_k, \tilde{x}_{k-1})\}$$

Theorem 1 (Borodich and Kovalev, 2025)

There exist universal constants $\theta, \gamma, \nu > 0$ such that

$$\begin{aligned}\tfrac{1}{2}\|x_K - x\|^2 + H_{K-1}(f(\bar{x}_K) - f(x)) \\ \leq \tfrac{1}{2}\|x_0 - x\|^2 + \tfrac{(1+\gamma\theta)}{2}\eta_0^2\|\nabla f(x_0)\|^2\end{aligned}$$

- **Allows geometric growth of the stepsize:**

$$\eta_k \leq (1+\gamma)^k \eta_0$$

Local Curvature Estimator

$$\Lambda(x, x') = \frac{2\mathrm{D}_f(x; x')}{\|\nabla f(x) - \nabla f(x')\|}$$

Accelerated GRAAL

$$\begin{aligned}\tilde{x}_k &= \alpha_k\hat{x}_k + (1-\alpha_k)\bar{x}_k \\ \bar{x}_{k+1} &= \beta_k\tilde{x}_k + (1-\beta_k)\bar{x}_k \\ \hat{x}_{k+1} &= x_{k+1} + \theta(x_{k+1} - x_k) \\ x_{k+1} &= x_k - \eta_k\nabla f(\tilde{x}_k)\end{aligned}$$

Accelerated GRAAL: Convergence Analysis

Accelerated GRAAL Stepsize

$$\eta_k = \min\left\{(1+\gamma)\eta_{k-1}, \frac{\nu H_{k-2}\Lambda_k}{\eta_{k-2}}\right\}$$

$$H_k = \sum_{i=0}^{k}\eta_i, \quad \Lambda_k = \min\{\Lambda(\bar{x}_k, \tilde{x}_k), \Lambda(\bar{x}_k, \tilde{x}_{k-1})\}$$

Theorem 1 (Borodich and Kovalev, 2025)

There exist universal constants $\theta, \gamma, \nu > 0$ such that

$$\begin{aligned}\tfrac{1}{2}\|x_K - x\|^2 + H_{K-1}(f(\bar{x}_K) - f(x)) \\ \leq \tfrac{1}{2}\|x_0 - x\|^2 + \tfrac{(1+\gamma\theta)}{2}\eta_0^2\|\nabla f(x_0)\|^2\end{aligned}$$

- **Allows geometric growth of the stepsize:**

$$\eta_k \leq (1+\gamma)^k \eta_0$$

- **Accelerated stepsize:** $\eta_k^2 \lesssim \Lambda \cdot \sum_{i=0}^{k}\eta_i$

Local Curvature Estimator

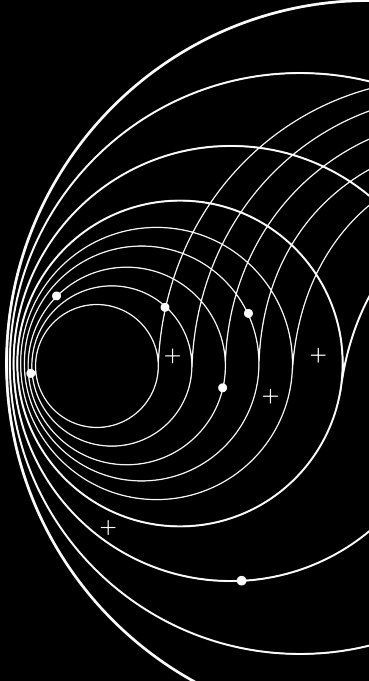
$$\Lambda(x, x') = \frac{2\mathrm{D}_f(x; x')}{\|\nabla f(x) - \nabla f(x')\|}$$

Accelerated GRAAL

$$\begin{aligned}\tilde{x}_k &= \alpha_k\hat{x}_k + (1-\alpha_k)\bar{x}_k \\ \bar{x}_{k+1} &= \beta_k\tilde{x}_k + (1-\beta_k)\bar{x}_k \\ \hat{x}_{k+1} &= x_{k+1} + \theta(x_{k+1} - x_k) \\ x_{k+1} &= x_k - \eta_k\nabla f(\tilde{x}_k)\end{aligned}$$

Yandex Research

Convergence Analysis for (L_0, L_1)-Smooth Functions



(L_0, L_1)-Smooth Functions

Local Curvature Estimator

$$\Lambda(x, x') = \frac{2\mathrm{D}_f(x; x')}{\|\nabla f(x) - \nabla f(x')\|}$$

Accel. GRAAL Parameters

$$\alpha_k = \frac{(1+\gamma)\eta_{k-1}}{H_{k-1}+(1+\gamma)\eta_{k-1}}$$

$$\Lambda_k = \min\{\Lambda(\bar{x}_k, \tilde{x}_k), \Lambda(\bar{x}_k, \tilde{x}_{k-1})\}$$

$$\eta_k = \min\left\{(1+\gamma)\eta_{k-1}, \frac{\nu H_{k-2}\Lambda_k}{\eta_{k-2}}\right\}$$

$$H_k = \sum_{i=0}^{k}\eta_i, \quad \beta_k = \frac{\eta_k}{\alpha_k H_k}$$

(L_0, L_1)-Smooth Functions

(L_0, L_1)-Smoothness (Zhang et al., 2019)

$$\|\nabla^2 f(x)\| \leq L_0 + L_1 \|\nabla f(x)\|$$

Local Curvature Estimator

$$\Lambda(x, x') = \frac{2\mathrm{D}_f(x; x')}{\|\nabla f(x) - \nabla f(x')\|}$$

Accel. GRAAL Parameters

$$\alpha_k = \frac{(1+\gamma)\eta_{k-1}}{H_{k-1} + (1+\gamma)\eta_{k-1}}$$

$$\Lambda_k = \min\{\Lambda(\bar{x}_k, \tilde{x}_k), \Lambda(\bar{x}_k, \tilde{x}_{k-1})\}$$

$$\eta_k = \min\left\{(1+\gamma)\eta_{k-1}, \frac{\nu H_{k-2}\Lambda_k}{\eta_{k-2}}\right\}$$

$$H_k = \sum_{i=0}^{k} \eta_i, \quad \beta_k = \frac{\eta_k}{\alpha_k H_k}$$

(L_0, L_1)-Smooth Functions

(L_0, L_1)-Smoothness (Zhang et al., 2019)

$$\|\nabla^2 f(x)\| \leq L_0 + L_1 \|\nabla f(x)\|$$

- Covers L-smooth functions ($L_0 = L, L_1 = 0$)

Local Curvature Estimator

$$\Lambda(x, x') = \frac{2\mathrm{D}_f(x; x')}{\|\nabla f(x) - \nabla f(x')\|}$$

Accel. GRAAL Parameters

$$\alpha_k = \frac{(1+\gamma)\eta_{k-1}}{H_{k-1}+(1+\gamma)\eta_{k-1}}$$

$$\Lambda_k = \min\{\Lambda(\bar{x}_k, \tilde{x}_k), \Lambda(\bar{x}_k, \tilde{x}_{k-1})\}$$

$$\eta_k = \min\left\{(1+\gamma)\eta_{k-1}, \frac{\nu H_{k-2}\Lambda_k}{\eta_{k-2}}\right\}$$

$$H_k = \sum_{i=0}^{k} \eta_i, \quad \beta_k = \frac{\eta_k}{\alpha_k H_k}$$

(L_0, L_1)-Smooth Functions

(L_0, L_1)-Smoothness (Zhang et al., 2019)

$$\|\nabla^2 f(x)\| \leq L_0 + L_1 \|\nabla f(x)\|$$

- Covers L-smooth functions ($L_0 = L, L_1 = 0$)
- More realistic than L-smoothness

Local Curvature Estimator

$$\Lambda(x, x') = \frac{2\mathrm{D}_f(x; x')}{\|\nabla f(x) - \nabla f(x')\|}$$

Accel. GRAAL Parameters

$$\alpha_k = \frac{(1+\gamma)\eta_{k-1}}{H_{k-1}+(1+\gamma)\eta_{k-1}}$$

$$\Lambda_k = \min\{\Lambda(\bar{x}_k, \tilde{x}_k), \Lambda(\bar{x}_k, \tilde{x}_{k-1})\}$$

$$\eta_k = \min\left\{(1+\gamma)\eta_{k-1}, \frac{\nu H_{k-2}\Lambda_k}{\eta_{k-2}}\right\}$$

$$H_k = \sum_{i=0}^{k} \eta_i, \quad \beta_k = \frac{\eta_k}{\alpha_k H_k}$$

(L_0, L_1)-Smooth Functions

(L_0, L_1)-Smoothness (Zhang et al., 2019)

$$\|\nabla^2 f(x)\| \leq L_0 + L_1 \|\nabla f(x)\|$$

- Covers L-smooth functions ($L_0 = L, L_1 = 0$)
- More realistic than L-smoothness

Examples

Local Curvature Estimator

$$\Lambda(x, x') = \frac{2\mathrm{D}_f(x; x')}{\|\nabla f(x) - \nabla f(x')\|}$$

Accel. GRAAL Parameters

$$\alpha_k = \frac{(1+\gamma)\eta_{k-1}}{H_{k-1}+(1+\gamma)\eta_{k-1}}$$

$$\Lambda_k = \min\{\Lambda(\bar{x}_k, \tilde{x}_k), \Lambda(\bar{x}_k, \tilde{x}_{k-1})\}$$

$$\eta_k = \min\left\{(1+\gamma)\eta_{k-1}, \frac{\nu H_{k-2}\Lambda_k}{\eta_{k-2}}\right\}$$

$$H_k = \sum_{i=0}^{k} \eta_i, \quad \beta_k = \frac{\eta_k}{\alpha_k H_k}$$

(L_0, L_1)-Smooth Functions

(L_0, L_1)-Smoothness (Zhang et al., 2019)

$$\|\nabla^2 f(x)\| \leq L_0 + L_1 \|\nabla f(x)\|$$

- Covers L-smooth functions ($L_0 = L, L_1 = 0$)
- More realistic than L-smoothness

Examples

- $x \mapsto \|x\|^p, \ p \geq 2$

Local Curvature Estimator

$$\Lambda(x, x') = \frac{2\mathrm{D}_f(x; x')}{\|\nabla f(x) - \nabla f(x')\|}$$

Accel. GRAAL Parameters

$$\alpha_k = \frac{(1+\gamma)\eta_{k-1}}{H_{k-1}+(1+\gamma)\eta_{k-1}}$$

$$\Lambda_k = \min\{\Lambda(\bar{x}_k, \tilde{x}_k), \Lambda(\bar{x}_k, \tilde{x}_{k-1})\}$$

$$\eta_k = \min\left\{(1+\gamma)\eta_{k-1}, \frac{\nu H_{k-2}\Lambda_k}{\eta_{k-2}}\right\}$$

$$H_k = \sum_{i=0}^{k} \eta_i, \quad \beta_k = \frac{\eta_k}{\alpha_k H_k}$$

(L_0, L_1)-Smooth Functions

(L_0, L_1)-Smoothness (Zhang et al., 2019)

$$\|\nabla^2 f(x)\| \leq L_0 + L_1 \|\nabla f(x)\|$$

- Covers L-smooth functions ($L_0 = L, L_1 = 0$)
- More realistic than L-smoothness

Examples

- $x \mapsto \|x\|^p, \ p \geq 2$
- $x \mapsto \exp(\langle a, x \rangle), \ a \in \mathbb{R}^d$

Local Curvature Estimator

$$\Lambda(x, x') = \frac{2\mathrm{D}_f(x; x')}{\|\nabla f(x) - \nabla f(x')\|}$$

Accel. GRAAL Parameters

$$\alpha_k = \frac{(1+\gamma)\eta_{k-1}}{H_{k-1} + (1+\gamma)\eta_{k-1}}$$

$$\Lambda_k = \min\{\Lambda(\bar{x}_k, \tilde{x}_k), \Lambda(\bar{x}_k, \tilde{x}_{k-1})\}$$

$$\eta_k = \min\left\{(1+\gamma)\eta_{k-1}, \frac{\nu H_{k-2} \Lambda_k}{\eta_{k-2}}\right\}$$

$$H_k = \sum_{i=0}^{k} \eta_i, \quad \beta_k = \frac{\eta_k}{\alpha_k H_k}$$

(L_0, L_1)-Smooth Functions

(L_0, L_1)-Smoothness (Zhang et al., 2019)

$$\|\nabla^2 f(x)\| \leq L_0 + L_1 \|\nabla f(x)\|$$

- Covers L-smooth functions ($L_0 = L, L_1 = 0$)
- More realistic than L-smoothness

Examples

- $x \mapsto \|x\|^p, \ p \geq 2$
- $x \mapsto \exp(\langle a, x \rangle), \ a \in \mathbb{R}^d$
- $x \mapsto \ln(1 + \exp(\langle a, x \rangle)), \ a \in \mathbb{R}^d$

Local Curvature Estimator

$$\Lambda(x, x') = \frac{2\mathrm{D}_f(x; x')}{\|\nabla f(x) - \nabla f(x')\|}$$

Accel. GRAAL Parameters

$$\alpha_k = \frac{(1+\gamma)\eta_{k-1}}{H_{k-1} + (1+\gamma)\eta_{k-1}}$$

$$\Lambda_k = \min\{\Lambda(\bar{x}_k, \tilde{x}_k), \Lambda(\bar{x}_k, \tilde{x}_{k-1})\}$$

$$\eta_k = \min\left\{(1+\gamma)\eta_{k-1}, \frac{\nu H_{k-2} \Lambda_k}{\eta_{k-2}}\right\}$$

$$H_k = \sum_{i=0}^{k} \eta_i, \quad \beta_k = \frac{\eta_k}{\alpha_k H_k}$$

Key Lemmas

Lemma 7 (Borodich and Kovalev, 2025)

At least one of the following options holds:

Local Curvature Estimator

$$\Lambda(x, x') = \frac{2\mathrm{D}_f(x; x')}{\|\nabla f(x) - \nabla f(x')\|}$$

Accel. GRAAL Parameters

$$\alpha_k = \frac{(1+\gamma)\eta_{k-1}}{H_{k-1}+(1+\gamma)\eta_{k-1}}$$

$$\Lambda_k = \min\{\Lambda(\bar{x}_k, \tilde{x}_k), \Lambda(\bar{x}_k, \tilde{x}_{k-1})\}$$

$$\eta_k = \min\left\{(1+\gamma)\eta_{k-1}, \frac{\nu H_{k-2}\Lambda_k}{\eta_{k-2}}\right\}$$

$$H_k = \sum_{i=0}^{k}\eta_i, \quad \beta_k = \frac{\eta_k}{\alpha_k H_k}$$

Key Lemmas

Lemma 7 (Borodich and Kovalev, 2025)

At least one of the following options holds:

- $\Lambda_k \geq \text{const}/L_0$

Local Curvature Estimator

$$\Lambda(x, x') = \frac{2\mathrm{D}_f(x; x')}{\|\nabla f(x) - \nabla f(x')\|}$$

Accel. GRAAL Parameters

$$\alpha_k = \frac{(1+\gamma)\eta_{k-1}}{H_{k-1}+(1+\gamma)\eta_{k-1}}$$

$$\Lambda_k = \min\{\Lambda(\bar{x}_k, \tilde{x}_k), \Lambda(\bar{x}_k, \tilde{x}_{k-1})\}$$

$$\eta_k = \min\left\{(1+\gamma)\eta_{k-1}, \frac{\nu H_{k-2}\Lambda_k}{\eta_{k-2}}\right\}$$

$$H_k = \sum_{i=0}^{k}\eta_i, \quad \beta_k = \frac{\eta_k}{\alpha_k H_k}$$

Key Lemmas

Lemma 7 (Borodich and Kovalev, 2025)

At least one of the following options holds:

- $\Lambda_k \geq \text{const}/L_0$
- $\Lambda_k \geq \text{const}/\max\{L_1\|\nabla f(\tilde{x}_k)\|, L_1\|\nabla f(\tilde{x}_{k-1})\|\}$

Local Curvature Estimator

$$\Lambda(x, x') = \frac{2\mathrm{D}_f(x; x')}{\|\nabla f(x) - \nabla f(x')\|}$$

Accel. GRAAL Parameters

$$\alpha_k = \frac{(1+\gamma)\eta_{k-1}}{H_{k-1}+(1+\gamma)\eta_{k-1}}$$

$$\Lambda_k = \min\{\Lambda(\bar{x}_k, \tilde{x}_k), \Lambda(\bar{x}_k, \tilde{x}_{k-1})\}$$

$$\eta_k = \min\left\{(1+\gamma)\eta_{k-1}, \frac{\nu H_{k-2}\Lambda_k}{\eta_{k-2}}\right\}$$

$$H_k = \sum_{i=0}^{k}\eta_i, \quad \beta_k = \frac{\eta_k}{\alpha_k H_k}$$

Key Lemmas

Lemma 7 (Borodich and Kovalev, 2025)

At least one of the following options holds:

- $\Lambda_k \geq \text{const}/L_0$
- $\Lambda_k \geq \text{const}/\max\{L_1\|\nabla f(\tilde{x}_k)\|, L_1\|\nabla f(\tilde{x}_{k-1})\|\}$
- $\Lambda_k \geq \text{const}/\max\{L_1^2 \mathrm{D}_f(\bar{x}_k, \tilde{x}_k), L_1^2 \mathrm{D}_f(\bar{x}_{k-1}, \tilde{x}_{k-1})\}$

Local Curvature Estimator

$$\Lambda(x, x') = \frac{2\mathrm{D}_f(x; x')}{\|\nabla f(x) - \nabla f(x')\|}$$

Accel. GRAAL Parameters

$$\alpha_k = \frac{(1+\gamma)\eta_{k-1}}{H_{k-1}+(1+\gamma)\eta_{k-1}}$$

$$\Lambda_k = \min\{\Lambda(\bar{x}_k, \tilde{x}_k), \Lambda(\bar{x}_k, \tilde{x}_{k-1})\}$$

$$\eta_k = \min\left\{(1+\gamma)\eta_{k-1}, \frac{\nu H_{k-2}\Lambda_k}{\eta_{k-2}}\right\}$$

$$H_k = \sum_{i=0}^{k}\eta_i, \quad \beta_k = \frac{\eta_k}{\alpha_k H_k}$$

Key Lemmas

Lemma 7 (Borodich and Kovalev, 2025)

At least one of the following options holds:

- $\Lambda_k \geq \text{const}/L_0$
- $\Lambda_k \geq \text{const}/\max\{L_1\|\nabla f(\tilde{x}_k)\|, L_1\|\nabla f(\tilde{x}_{k-1})\|\}$
- $\Lambda_k \geq \text{const}/\max\{L_1^2 \mathrm{D}_f(\bar{x}_k, \tilde{x}_k), L_1^2 \mathrm{D}_f(\bar{x}_{k-1}, \tilde{x}_{k-1})\}$

Lemma 5 (Borodich and Kovalev, 2025)

$$\sum_{i=1}^{K}(\eta_i \mathrm{D}_f(\bar{x}_i, \tilde{x}_i) + \eta_i^2\|\nabla f(x_i)\|^2) \leq \mathcal{D}^2$$

Local Curvature Estimator

$$\Lambda(x, x') = \frac{2\mathrm{D}_f(x; x')}{\|\nabla f(x) - \nabla f(x')\|}$$

Accel. GRAAL Parameters

$$\alpha_k = \frac{(1+\gamma)\eta_{k-1}}{H_{k-1}+(1+\gamma)\eta_{k-1}}$$

$$\Lambda_k = \min\{\Lambda(\bar{x}_k, \tilde{x}_k), \Lambda(\bar{x}_k, \tilde{x}_{k-1})\}$$

$$\eta_k = \min\left\{(1+\gamma)\eta_{k-1}, \frac{\nu H_{k-2}\Lambda_k}{\eta_{k-2}}\right\}$$

$$H_k = \sum_{i=0}^{k}\eta_i, \quad \beta_k = \frac{\eta_k}{\alpha_k H_k}$$

Key Lemmas

Lemma 7 (Borodich and Kovalev, 2025)

At least one of the following options holds:

- $\Lambda_k \geq \text{const}/L_0$
- $\Lambda_k \geq \text{const}/\max\{L_1\|\nabla f(\tilde{x}_k)\|, L_1\|\nabla f(\tilde{x}_{k-1})\|\}$
- $\Lambda_k \geq \text{const}/\max\{L_1^2 \mathrm{D}_f(\bar{x}_k, \tilde{x}_k), L_1^2 \mathrm{D}_f(\bar{x}_{k-1}, \tilde{x}_{k-1})\}$

Lemma 5 (Borodich and Kovalev, 2025)

$$\sum_{i=1}^{K}(\eta_i \mathrm{D}_f(\bar{x}_i, \tilde{x}_i) + \eta_i^2\|\nabla f(x_i)\|^2) \leq \mathcal{D}^2$$

- Lemmas 5 and 7 allow to upper-bound the number of iterations where $\Lambda_k < \text{const}/L_0$

Local Curvature Estimator

$$\Lambda(x, x') = \frac{2\mathrm{D}_f(x; x')}{\|\nabla f(x) - \nabla f(x')\|}$$

Accel. GRAAL Parameters

$$\alpha_k = \frac{(1+\gamma)\eta_{k-1}}{H_{k-1}+(1+\gamma)\eta_{k-1}}$$

$$\Lambda_k = \min\{\Lambda(\bar{x}_k, \tilde{x}_k), \Lambda(\bar{x}_k, \tilde{x}_{k-1})\}$$

$$\eta_k = \min\left\{(1+\gamma)\eta_{k-1}, \frac{\nu H_{k-2}\Lambda_k}{\eta_{k-2}}\right\}$$

$$H_k = \sum_{i=0}^{k}\eta_i, \quad \beta_k = \frac{\eta_k}{\alpha_k H_k}$$

Key Lemmas

Lemma 7 (Borodich and Kovalev, 2025)

At least one of the following options holds:

- $\Lambda_k \geq \text{const}/L_0$
- $\Lambda_k \geq \text{const}/\max\{L_1\|\nabla f(\tilde{x}_k)\|, L_1\|\nabla f(\tilde{x}_{k-1})\|\}$
- $\Lambda_k \geq \text{const}/\max\{L_1^2 \mathrm{D}_f(\bar{x}_k, \tilde{x}_k), L_1^2 \mathrm{D}_f(\bar{x}_{k-1}, \tilde{x}_{k-1})\}$

Lemma 5 (Borodich and Kovalev, 2025)

$$\sum_{i=1}^{K}(\eta_i \mathrm{D}_f(\bar{x}_i, \tilde{x}_i) + \eta_i^2\|\nabla f(x_i)\|^2) \leq \mathcal{D}^2$$

- Lemmas 5 and 7 allow to upper-bound the number of iterations where $\Lambda_k < \text{const}/L_0$
- This allows to lower-bound H_k

Local Curvature Estimator

$$\Lambda(x, x') = \frac{2\mathrm{D}_f(x; x')}{\|\nabla f(x) - \nabla f(x')\|}$$

Accel. GRAAL Parameters

$$\alpha_k = \frac{(1+\gamma)\eta_{k-1}}{H_{k-1}+(1+\gamma)\eta_{k-1}}$$

$$\Lambda_k = \min\{\Lambda(\bar{x}_k, \tilde{x}_k), \Lambda(\bar{x}_k, \tilde{x}_{k-1})\}$$

$$\eta_k = \min\left\{(1+\gamma)\eta_{k-1}, \frac{\nu H_{k-2}\Lambda_k}{\eta_{k-2}}\right\}$$

$$H_k = \sum_{i=0}^{k}\eta_i, \quad \beta_k = \frac{\eta_k}{\alpha_k H_k}$$

Main Result

Theorem 3 (Borodich and Kovalev, 2025)

Let $\eta_0 L_0 \exp(L_1\|x_0 - x^*\|) \leq 1$. Then $\mathcal{D} = \mathcal{O}(\|x_0 - x^*\|)$ and the following inequality holds:

$$H_K \geq \frac{c^2}{L_0}\Big[K - \big(1 + L_1^3\mathcal{D}^3\big) - \big(1 + L_1^2\mathcal{D}^2\big)\ln\Big[\tfrac{1}{\eta_0 L_0}\Big]\Big]^2$$

Local Curvature Estimator

$$\Lambda(x, x') = \frac{2\mathrm{D}_f(x; x')}{\|\nabla f(x) - \nabla f(x')\|}$$

Accel. GRAAL Parameters

$$\alpha_k = \tfrac{(1+\gamma)\eta_{k-1}}{H_{k-1}+(1+\gamma)\eta_{k-1}}$$
$$\Lambda_k = \min\{\Lambda(\bar{x}_k, \tilde{x}_k), \Lambda(\bar{x}_k, \tilde{x}_{k-1})\}$$
$$\eta_k = \min\Big\{(1+\gamma)\eta_{k-1}, \tfrac{\nu H_{k-2}\Lambda_k}{\eta_{k-2}}\Big\}$$
$$H_k = \textstyle\sum_{i=0}^{k}\eta_i, \quad \beta_k = \tfrac{\eta_k}{\alpha_k H_k}$$

Main Result

Theorem 3 (Borodich and Kovalev, 2025)

Let $\eta_0 L_0 \exp(L_1\|x_0 - x^*\|) \leq 1$. Then $\mathcal{D} = \mathcal{O}(\|x_0 - x^*\|)$ and the following inequality holds:

$$H_K \geq \frac{c^2}{L_0}\Big[K - \big(1 + L_1^3\mathcal{D}^3\big) - \big(1 + L_1^2\mathcal{D}^2\big)\ln\Big[\tfrac{1}{\eta_0 L_0}\Big]\Big]^2$$

Corollary 3 (Borodich and Kovalev, 2025)

To reach precision $f(\bar{x}_K) - f^* \leq \epsilon$, the following number of iterations is sufficient:

$$K = \mathcal{O}\Big[1 + \sqrt{L_0\mathcal{D}^2/\epsilon} + L_1^3\mathcal{D}^3 + \big(1 + L_1^2\mathcal{D}^2\big)\ln\Big[\tfrac{1}{\eta_0 L_0}\Big]\Big]$$

Local Curvature Estimator

$$\Lambda(x, x') = \frac{2\mathrm{D}_f(x; x')}{\|\nabla f(x) - \nabla f(x')\|}$$

Accel. GRAAL Parameters

$$\alpha_k = \tfrac{(1+\gamma)\eta_{k-1}}{H_{k-1}+(1+\gamma)\eta_{k-1}}$$
$$\Lambda_k = \min\{\Lambda(\bar{x}_k, \tilde{x}_k), \Lambda(\bar{x}_k, \tilde{x}_{k-1})\}$$
$$\eta_k = \min\Big\{(1+\gamma)\eta_{k-1}, \tfrac{\nu H_{k-2}\Lambda_k}{\eta_{k-2}}\Big\}$$
$$H_k = \textstyle\sum_{i=0}^{k}\eta_i, \quad \beta_k = \tfrac{\eta_k}{\alpha_k H_k}$$

Comparison with AC-FGM and Ada-NAG

Corollary 3 (Borodich and Kovalev, 2025)

Let $\eta_0 L_0 \exp(L_1 \mathcal{D}) \leq 1$. Then $\mathcal{D} = \mathcal{O}(\|x_0 - x^*\|)$, and Accelerated GRAAL has the following iteration complexity:

$$K = \mathcal{O}\left[1 + \sqrt{L_0 \mathcal{D}^2/\epsilon} + L_1^3 \mathcal{D}^3 + \left(1 + L_1^2 \mathcal{D}^2\right) \ln\left[\frac{1}{\eta_0 L_0}\right]\right]$$

Comparison with AC-FGM and Ada-NAG

Corollary 3 (Borodich and Kovalev, 2025)

Let $\eta_0 L_0 \exp(L_1 \mathcal{D}) \leq 1$. Then $\mathcal{D} = \mathcal{O}(\|x_0 - x^*\|)$, and Accelerated GRAAL has the following iteration complexity:

$$K = \mathcal{O}\left[1 + \sqrt{L_0 \mathcal{D}^2/\epsilon} + L_1^3 \mathcal{D}^3 + \left(1 + L_1^2 \mathcal{D}^2\right) \ln\left[\frac{1}{\eta_0 L_0}\right]\right]$$

Comparison with AC-FGM and AdaNAG:

Comparison with AC-FGM and Ada-NAG

Corollary 3 (Borodich and Kovalev, 2025)

Let $\eta_0 L_0 \exp(L_1 \mathcal{D}) \leq 1$. Then $\mathcal{D} = \mathcal{O}(\|x_0 - x^*\|)$, and Accelerated GRAAL has the following iteration complexity:

$$K = \mathcal{O}\left[1 + \sqrt{L_0 \mathcal{D}^2/\epsilon} + L_1^3 \mathcal{D}^3 + \left(1 + L_1^2 \mathcal{D}^2\right) \ln\left[\frac{1}{\eta_0 L_0}\right]\right]$$

Comparison with AC-FGM and AdaNAG:

- We can ensure $\eta_0 L_0 \exp(L_1 \|x_0 - x^*\|) \leq 1$ by choosing a very small η_0 at the cost of logarithmic factors

Comparison with AC-FGM and Ada-NAG

Corollary 3 (Borodich and Kovalev, 2025)

Let $\eta_0 L_0 \exp(L_1 \mathcal{D}) \leq 1$. Then $\mathcal{D} = \mathcal{O}(\|x_0 - x^*\|)$, and Accelerated GRAAL has the following iteration complexity:

$$K = \mathcal{O}\left[1 + \sqrt{L_0 \mathcal{D}^2/\epsilon} + L_1^3 \mathcal{D}^3 + \left(1 + L_1^2 \mathcal{D}^2\right) \ln\left[\frac{1}{\eta_0 L_0}\right]\right]$$

Comparison with AC-FGM and AdaNAG:

- We can ensure $\eta_0 L_0 \exp(L_1 \|x_0 - x^*\|) \leq 1$ by choosing a very small η_0 at the cost of logarithmic factors
- Li and Lan (2025) perform a line search at the first iteration of AC-FGM instead

Comparison with AC-FGM and Ada-NAG

Corollary 3 (Borodich and Kovalev, 2025)

Let $\eta_0 L_0 \exp(L_1 \mathcal{D}) \leq 1$. Then $\mathcal{D} = \mathcal{O}(\|x_0 - x^*\|)$, and Accelerated GRAAL has the following iteration complexity:

$$K = \mathcal{O}\left[1 + \sqrt{L_0 \mathcal{D}^2/\epsilon} + L_1^3 \mathcal{D}^3 + \left(1 + L_1^2 \mathcal{D}^2\right) \ln\left[\frac{1}{\eta_0 L_0}\right]\right]$$

Comparison with AC-FGM and AdaNAG:

- We can ensure $\eta_0 L_0 \exp(L_1 \|x_0 - x^*\|) \leq 1$ by choosing a very small η_0 at the cost of logarithmic factors
- Li and Lan (2025) perform a line search at the first iteration of AC-FGM instead
- No results under (L_0, L_1)-smoothness for AC-FGM (Li and Lan, 2025) and AdaNAG (Suh and Ma, 2025)

Comparison with AC-FGM and Ada-NAG

Corollary 3 (Borodich and Kovalev, 2025)

Let $\eta_0 L_0 \exp(L_1 \mathcal{D}) \leq 1$. Then $\mathcal{D} = \mathcal{O}(\|x_0 - x^*\|)$, and Accelerated GRAAL has the following iteration complexity:

$$K = \mathcal{O}\left[1 + \sqrt{L_0 \mathcal{D}^2/\epsilon} + L_1^3 \mathcal{D}^3 + \left(1 + L_1^2 \mathcal{D}^2\right) \ln\left[\frac{1}{\eta_0 L_0}\right]\right]$$

Comparison with AC-FGM and AdaNAG:

- We can ensure $\eta_0 L_0 \exp(L_1 \|x_0 - x^*\|) \leq 1$ by choosing a very small η_0 at the cost of logarithmic factors
- Li and Lan (2025) perform a line search at the first iteration of AC-FGM instead
- No results under (L_0, L_1)-smoothness for AC-FGM (Li and Lan, 2025) and AdaNAG (Suh and Ma, 2025)
- Adaptation at a geometric rate is crucial, as Λ_k may be exponentially small, $\Lambda_k \geq \exp(-3 L_1 \mathcal{D})$ (Borodich and Kovalev, 2025, Lemma 6)

Comparison with Algorithms for (L_0, L_1)-smooth Functions

Algorithm	Complexity	Remarks

Comparison with Algorithms for (L_0, L_1)-smooth Functions

Algorithm	Complexity	Remarks
Similar Triangles Method (Gasnikov and Nesterov, 2016) (Gorbunov et al., 2024)	$\sqrt{\frac{L_0\mathcal{D}^2}{\epsilon}} \times \sqrt{1 + L_1\mathcal{D}\exp(L_1\mathcal{D})}$	✘ Non-adaptive ✘ Not optimal

Comparison with Algorithms for (L_0, L_1)-smooth Functions

Algorithm	Complexity	Remarks
Similar Triangles Method (Gasnikov and Nesterov, 2016) (Gorbunov et al., 2024)	$\sqrt{\frac{L_0\mathcal{D}^2}{\epsilon}} \times \sqrt{1 + L_1\mathcal{D}\exp(L_1\mathcal{D})}$	✘ Non-adaptive ✘ Not optimal
AdGD (Malitsky and Mishchenko, 2020) (Gorbunov et al., 2024)	$\frac{L_0\mathcal{D}^2}{\epsilon} + (L_1\mathcal{D})^6$	✔ Adaptive ✘ Non-accelerated

Comparison with Algorithms for (L_0, L_1)-smooth Functions

Algorithm	Complexity	Remarks
Similar Triangles Method (Gasnikov and Nesterov, 2016) (Gorbunov et al., 2024)	$\sqrt{\frac{L_0\mathcal{D}^2}{\epsilon}} \times \sqrt{1 + L_1\mathcal{D}\exp(L_1\mathcal{D})}$	✘ Non-adaptive ✘ Not optimal
AdGD (Malitsky and Mishchenko, 2020) (Gorbunov et al., 2024)	$\frac{L_0\mathcal{D}^2}{\epsilon} + (L_1\mathcal{D})^6$	✔ Adaptive ✘ Non-accelerated
AGMsDR (Vankov et al., 2024)	$\sqrt{\frac{L_0\mathcal{D}^2}{\epsilon}} + (L_1\mathcal{D})^{5/3}$	✘ Non-adaptive ✔ Optimal

Comparison with Algorithms for (L_0, L_1)-smooth Functions

Algorithm	Complexity	Remarks
Similar Triangles Method (Gasnikov and Nesterov, 2016) (Gorbunov et al., 2024)	$\sqrt{\frac{L_0\mathcal{D}^2}{\epsilon}} \times \sqrt{1 + L_1\mathcal{D}\exp(L_1\mathcal{D})}$	✘ Non-adaptive ✘ Not optimal
AdGD (Malitsky and Mishchenko, 2020) (Gorbunov et al., 2024)	$\frac{L_0\mathcal{D}^2}{\epsilon} + (L_1\mathcal{D})^6$	✔ Adaptive ✘ Non-accelerated
AGMsDR (Vankov et al., 2024)	$\sqrt{\frac{L_0\mathcal{D}^2}{\epsilon}} + (L_1\mathcal{D})^{5/3}$	✘ Non-adaptive ✔ Optimal
AGD (Tyurin, 2025)	$\sqrt{\frac{L_0\mathcal{D}^2}{\epsilon}} + (L_1\mathcal{D})^2$	✘ Non-adaptive ✔ Optimal

Comparison with Algorithms for (L_0, L_1)-smooth Functions

Algorithm	Complexity	Remarks
Similar Triangles Method (Gasnikov and Nesterov, 2016) (Gorbunov et al., 2024)	$\sqrt{\frac{L_0\mathcal{D}^2}{\epsilon}} \times \sqrt{1 + L_1\mathcal{D}\exp(L_1\mathcal{D})}$	✘ Non-adaptive ✘ Not optimal
AdGD (Malitsky and Mishchenko, 2020) (Gorbunov et al., 2024)	$\frac{L_0\mathcal{D}^2}{\epsilon} + (L_1\mathcal{D})^6$	✔ Adaptive ✘ Non-accelerated
AGMsDR (Vankov et al., 2024)	$\sqrt{\frac{L_0\mathcal{D}^2}{\epsilon}} + (L_1\mathcal{D})^{5/3}$	✘ Non-adaptive ✔ Optimal
AGD (Tyurin, 2025)	$\sqrt{\frac{L_0\mathcal{D}^2}{\epsilon}} + (L_1\mathcal{D})^2$	✘ Non-adaptive ✔ Optimal
Accelerated GRAAL (Borodich and Kovalev, 2025)	$\sqrt{\frac{L_0\mathcal{D}^2}{\epsilon}} + (L_1\mathcal{D})^3$	✔ Adaptive ✔ Optimal

References I

Borodich, E., Gasnikov, A., and Kovalev, D. (2025). On linear convergence in smooth convex-concave bilinearly-coupled saddle-point optimization: Lower bounds and optimal algorithms. *Accepted to International Conference on Machine Learning*.

Borodich, E. and Kovalev, D. (2025). Nesterov finds graal: Optimal and adaptive gradient method for convex optimization. *arXiv preprint arXiv:2507.09823*.

Duchi, J., Hazan, E., and Singer, Y. (2011). Adaptive subgradient methods for online learning and stochastic optimization. *Journal of machine learning research*, 12(7).

Gasnikov, A. and Nesterov, Y. (2016). Universal fast gradient method for stochastic composit optimization problems. *arXiv preprint arXiv:1604.05275*.

Gorbunov, E., Tupitsa, N., Choudhury, S., Aliev, A., Richtárik, P., Horváth, S., and Takáč, M. (2024). Methods for convex (l_0, l_1)-smooth optimization: Clipping, acceleration, and adaptivity. *arXiv preprint arXiv:2409.14989*.

Kingma, D. P. and Ba, J. (2014). Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

Kleinberg, B., Li, Y., and Yuan, Y. (2018). An alternative view: When does sgd escape local minima? In *International conference on machine learning*, pages 2698–2707. PMLR.

References II

Li, T. and Lan, G. (2025). A simple uniformly optimal method without line search for convex optimization: T. li, g. lan. *Mathematical Programming*, pages 1–38.

Malitsky, Y. (2020). Golden ratio algorithms for variational inequalities. *Mathematical Programming*, 184(1):383–410.

Malitsky, Y. and Mishchenko, K. (2020). Adaptive gradient descent without descent. In *International Conference on Machine Learning*, pages 6702–6712. PMLR.

Nesterov, Y. (1983). A method for unconstrained convex minimization problem with the rate of convergence $O(1/k^2)$. *Dokl. Akad. Nauk. SSSR*, 269(3):543.

Orabona, F. (2023). Normalized gradients for all. *arXiv preprint arXiv:2308.05621*.

Suh, J. J. and Ma, S. (2025). An adaptive and parameter-free nesterov's accelerated gradient method for convex optimization. *arXiv preprint arXiv:2505.11670*.

Tieleman, T. (2012). Lecture 6.5-rmsprop: Divide the gradient by a running average of its recent magnitude. *COURSERA: Neural networks for machine learning*, 4(2):26.

Tyurin, A. (2025). Near-optimal convergence of accelerated gradient methods under generalized and (l_0, l_1)-smoothness. *arXiv preprint arXiv:2508.06884*.

Vankov, D., Rodomanov, A., Nedich, A., Sankar, L., and Stich, S. U. (2024). Optimizing (l_0, l_1)-smooth functions by gradient methods. *arXiv preprint arXiv:2410.10800*.

References III

Zhang, J., He, T., Sra, S., and Jadbabaie, A. (2019). Why gradient clipping accelerates training: A theoretical justification for adaptivity. *arXiv preprint arXiv:1905.11881*.

Zhou, Y., Yang, J., Zhang, H., Liang, Y., and Tarokh, V. (2019). Sgd converges to global minimum in deep learning via star-convex path. *arXiv preprint arXiv:1901.00451*.