Sih-Ray Mao

Linear Coupling: An Ultimate Unification of Gradient and Mirror Descent - Zenyuan Allen-Zhu & Lorenzo Orrecchia

Last modified: 2025-10-06

1 Summary

Allen-Zhu and Orecchia [AO14] proposed an algorithm, Accelerated Gradient Method, for smooth convex optimization with a convergence rate of $O(1/T^2)$, which matches that of Nesterov's accelerated gradient descent [Nes05]. This is achieved by a technique called *linear coupling* that combines gradient descent and mirror descent. Compared to Nesterov's accelerated gradient method, the analysis of Accelerated Gradient Method is more intuitive and straightforward. Moreover, Accelerated Gradient Method works in more general settings than Nesterov's algorithm.

2 Preliminaries

This paper considers smooth convex optimization problems. A smooth convex optimization problem is of the following form:

$$\min_{x \in \mathcal{X}} f(x),$$

where $\mathcal{X} \subseteq \mathbb{R}^n$ is a closed convex set and $f : \mathbb{R}^n \to \mathbb{R}$ is a convex differentiable function that is L-smooth with respect to a norm $\|\cdot\|$ on $\mathcal{X}$.

Definition 2.1 (Smoothness)**.** *A differentiable function $f : \mathbb{R}^n \to \mathbb{R}$ is said to be **L-smooth** with respect to a norm $\|\cdot\|$ on $\mathcal{X} \subseteq \mathbb{R}^n$ if we have*

$$\|\nabla f(x) - \nabla f(y)\|_* \leq L\|x - y\|, \quad \forall x, y \in \mathcal{X},$$

*where $\|\cdot\|_*$ denotes the dual norm of $\|\cdot\|$. The constant $L > 0$ is called the **smoothness parameter**.*

If $f : \mathbb{R}^n \to \mathbb{R}$ is an L-smooth function with respect to $\|\cdot\|$ on $\mathcal{X} \subseteq \mathbb{R}^n$, then we have

$$f(y) \leq f(x) + \langle \nabla f(x), y - x\rangle + \frac{L}{2}\|y - x\|^2, \quad \forall x, y \in \mathcal{X}. \tag{2.1}$$

Namely, we have a quadratic upper bound of f at each point in $\mathcal{X}$.

Definition 2.2 (Strong convexity)**.** *A differentiable function $f : \mathbb{R}^n \to \mathbb{R}$ is said to be **μ-strongly convex** with respect to a norm $\|\cdot\|$ on $\mathcal{X} \subseteq \mathbb{R}^n$ if we have*

$$\|f(x) - f(y)\|_* \geq \mu\|x - y\|, \quad \forall x, y \in \mathcal{X},$$

*where $\|\cdot\|_*$ denotes the dual norm of $\|\cdot\|$. The constant $\mu > 0$ is called the **strong convexity parameter**.*

If $f : \mathbb{R}^n \to \mathbb{R}$ is μ-strongly convex with respect to $\|\cdot\|$ on $\mathcal{X} \subseteq \mathbb{R}^n$, then we can show that

$$f(y) \geq f(x) + \langle \nabla f(x), y - x\rangle + \frac{\mu}{2}\|y - x\|^2, \quad \forall x, y \in \mathcal{X}. \tag{2.2}$$

In words, we can construct a quadratic lower bound of f at each point in $\mathcal{X}$.

Definition 2.3 (Bregman divergence)**.** *Let $h : \mathbb{R}^n \to \mathbb{R}$ be a differentiable function. The **Bregman divergence** with respect to to h, denoted by $D_h : \mathrm{dom}(h) \times \mathrm{dom}(\nabla h) \to \mathbb{R}$, is defined as*

$$D_h(y \,\|\, x) \coloneqq h(y) - h(x) - \langle \nabla h(x), y - x\rangle, \quad \forall x \in \mathrm{dom}(\nabla h), y \in \mathrm{dom}(h).$$

Let $h : \mathbb{R}^n \to \mathbb{R}$ be 1-strongly convex with respect to $\|\cdot\|$ on $\mathcal{X} \subseteq \mathbb{R}^n$. Then, from Equation (2.2), we have

$$D_h(y \,\|\, x) = h(y) - h(x) - \langle \nabla h(x), y - x\rangle \geq \frac{1}{2}\|y - x\|^2,$$

for all $x \in \mathrm{int}(\mathcal{X})$ and $y \in \mathcal{X}$.

Lemma 2.4 (Three-point equality). *Let $\mathcal{X} \subseteq \mathbb{R}^n$ be a closed convex set and let $h : \mathcal{X} \to \mathbb{R}$ be differentiable. Then, we have*

$$D_h(x \,\|\, y) + D_h(y \,\|\, z) = D_h(x \,\|\, z) + \langle \nabla h(z) - \nabla h(y), x - y \rangle \,,$$

for all $x \in \mathcal{X}$ and $y, z \in \mathrm{int}(\mathcal{X})$.

Theorem 2.5 (Optimality condition). *Let $\mathcal{X} \subseteq \mathbb{R}^n$ be a closed convex set and let $f : \mathbb{R}^n \to \mathbb{R}$ be convex differentiable on $\mathcal{X}$. Then, we have $x^\star \in \arg\min_{x \in \mathcal{X}} f(x)$ if and only if*

$$\langle \nabla f(x^\star), x - x^\star \rangle \geq 0, \quad \forall x \in \mathcal{X}.$$

Definition 2.6 (Fenchel conjugate). *Let $f : \mathbb{R}^n \to [-\infty, +\infty]$ be an extended real-valued function. The* ***Fenchel conjugate*** *of f, $f^* : \mathbb{R}^n \to [-\infty, +\infty]$, is defined as*

$$f^*(\theta) \coloneqq \sup_{x \in \mathbb{R}^n} \{\langle \theta, x \rangle - f(x)\}, \quad \forall \theta \in \mathbb{R}^n.$$

Recall that the Fenchel conjugate of the halved square norm $f(x) = \|x\|^2/2$ is given by $f^*(\theta) = \|\theta\|_*^2/2$, where $\|\cdot\|_*$ denotes the dual norm of $\|\cdot\|$.

3 Gradient Descent, Mirror Descent and Nesterov's Accelerated Gradient Method

3.1 Gradient Descent

Let $\mathcal{X} \subseteq \mathbb{R}^n$ be a closed convex set, and let $f : \mathbb{R}^n \to \mathbb{R}$ be a convex differentiable function that is L-smooth with respect to $\|\cdot\|$ on $\mathcal{X}$. In each round $t = 1, \ldots, T$, Gradient Descent iterates as

$$x_{t+1} \leftarrow \arg\min_{y \in \mathcal{X}} \left\{ \langle \nabla f(x_t), y - x_t \rangle + \frac{L}{2} \|y - x_t\|^2 \right\}, \tag{3.1}$$

where we choose the initial point x_1 as any feasible point in the constraint set $\mathcal{X}$. The updating rule of Gradient Descent can be interpreted as follows. Since f is L-smooth with respect to $\|\cdot\|$ on $\mathcal{X}$, by Equation (2.1), we can write

$$f(x_t) - f(y) \geq \langle \nabla f(x_t), x_t - y \rangle - \frac{L}{2} \|y - x_t\|^2, \quad \forall y \in \mathcal{X}. \tag{3.2}$$

Thus, from Equation (3.1) and Equation (3.2), we can see that Gradient Descent chooses the next iterate x_{t+1} as the point that maximizes the "objective progress." We formally define the notion of gradient step and objective progress in the following.

Definition 3.1 (Gradient step, objective progress). *Let $\mathcal{X} \subseteq \mathbb{R}^n$ be a closed convex set, and let $f : \mathbb{R}^n \to \mathbb{R}$ be a convex differentiable function that is L-smooth with respect to $\|\cdot\|$ on $\mathcal{X}$. The* ***gradient step*** *with respect to f (with step size $1/L$) is defined as*

$$\mathsf{Grad}_f(x) \coloneqq \arg\min_{y \in \mathcal{X}} \left\{ \langle \nabla f(x), y - x \rangle + \frac{L}{2} \|y - x\|^2 \right\}, \quad \forall x \in \mathcal{X}.$$

The ***objective progress*** *with respect to f is defined as*

$$\mathsf{Prog}_f(x) \coloneqq \max_{y \in \mathcal{X}} \left\{ \langle \nabla f(x), x - y \rangle - \frac{L}{2} \|x - y\|^2 \right\}, \quad \forall x \in \mathcal{X}.$$

Observe that we immediately have

$$\mathsf{Prog}_f(x) = \max_{y \in \mathcal{X}} \left\{ \langle \nabla f(x), y - x \rangle - \frac{L}{2} \|x - y\|^2 \right\} \geq \langle \nabla f(x), x - x \rangle - \frac{L}{2} \|x - x\|^2 = 0. \tag{3.3}$$

Lemma 3.2 (Gradient descent lemma). *Let $\mathcal{X} \subseteq \mathbb{R}^n$ be a closed convex set and let $x \in \mathcal{X}$. Let $f : \mathbb{R}^n \to \mathbb{R}$ be a convex differentiable function that is L-smooth with respect to $\|\cdot\|$ on $\mathcal{X}$. If $\tilde{x} \coloneqq \mathsf{Grad}_f(x)$, then we have*

$$f(\tilde{x}) \leq f(x) - \mathsf{Prog}_f(x). \tag{3.4}$$

Moreover, if $\mathcal{X} = \mathbb{R}^n$ (i.e., the unconstrained case), we have

$$f(\tilde{x}) \leq f(x) - \frac{\|\nabla f(x)\|_*^2}{2L}. \tag{3.5}$$

Proof. By the definition of gradient step and objective progress (Definition 3.1), we can write

$$\langle \nabla f(x), \tilde{x} - x \rangle + \frac{L}{2}\|\tilde{x} - x\|^2 = \min_{y \in \mathcal{X}} \left\{ \langle \nabla f(x), y - x \rangle + \frac{L}{2}\|y - x\|^2 \right\} = -\mathsf{Prog}_f(x).$$

Then, by the smoothness of f, we have

$$f(\tilde{x}) \le f(x) + \langle \nabla f(x), \tilde{x} - x \rangle + \frac{L}{2}\|\tilde{x} - x\|^2 \le f(x) - \mathsf{Prog}_f(x).$$

This proves Equation (3.4). Now, let us consider the unconstrained case. We can write

$$\begin{aligned} \mathsf{Prog}_f(x) &= \max_{y \in \mathbb{R}^n} \left\{ \langle \nabla f(x), x - y \rangle - \frac{L}{2}\|x - y\|^2 \right\} \\ &= L \cdot \max_{y \in \mathbb{R}^n} \left\{ \left\langle \frac{\nabla f(x)}{L}, x - y \right\rangle - \frac{\|x - y\|^2}{2} \right\} \\ &= \frac{\|\nabla f(x)\|_*^2}{2L}, \end{aligned}$$

where the last equality follows from the fact that the Fenchel conjugate of $\|\cdot\|^2/2$ is $\|\cdot\|_*^2/2$. This proves Equation (3.5). □

With the definition of gradient step (Definition 3.1), we can rewrite the updating rule of Gradient Descent as $x_{t+1} \leftarrow \mathsf{Grad}_f(x_t)$. Moreover, from Lemma 3.2, we know that $f(x_t) - f(x_{t+1}) \ge \mathsf{Prog}_f(x_t)$. Together with the fact that $\mathsf{Prog}_f(x) \ge 0$ for all $x \in \mathcal{X}$ (Equation (3.3)), if $\{x_t\}_{t=1}^T$ is the iterates of Gradient Descent, then we know that $\{f(x_t)\}_{t=1}^T$ is a monotonically decreasing sequence. In other words, if the objective function is smooth, then Gradient Descent is an optimization algorithm that decreases the function value in each iteration.

Remark 3.3. For the unconstrained case, the objective progress of Gradient Descent is lower bounded by

$$f(x_t) - f(x_{t+1}) \ge \frac{\|\nabla f(x_t)\|_*^2}{2L}.$$

Therefore, Gradient Descent makes larger progress when the gradient norm $\|\nabla f(x_t)\|_*$ is large.

Theorem 3.4 (Gradient descent guarantee)**.** *Let $f : \mathbb{R}^n \to \mathbb{R}$ be a convex differentiable function that is L-smooth with respect to $\|\cdot\|$, and let $x_1 \in \mathbb{R}^n$ be an arbitrary initial point. Then,* Gradient Descent *achieves*

$$f(x_T) - f(x^\star) \le \frac{2LR^2}{T},$$

where $x^\star$ is a minimizer of f and $R = \max_{f(x) \le f(x_1)} \|x - x^\star\|$.

Proof. From the gradient descent lemma (Lemma 3.2), we have

$$f(x_{t+1}) - f(x_t) \le -\frac{\|\nabla f(x_t)\|_*^2}{2L}, \tag{3.6}$$

for $t = 1, \ldots, T-1$. By the convexity of f and the Hölder's inequality, we can write

$$f(x_t) - f(x^\star) \le \langle \nabla f(x_t), x_t - x^\star \rangle \le \|\nabla f(x_t)\|_* \|x_t - x^\star\| \le R\|\nabla f(x_t)\|_*, \tag{3.7}$$

for $t = 1, \ldots, T$, where the last inequality follows from the definition of R. For each $t = 1, \ldots, T$, define Δ_t as the gap between $f(x_t)$ and $f(x^\star)$, that is, $\Delta_t := f(x_t) - f(x^\star)$. Putting Equation (3.6) and Equation (3.7) together yields

$$\Delta_t^2 \le 2LR^2(\Delta_t - \Delta_{t+1}). \tag{3.8}$$

By dividing $\Delta_t \Delta_{t+1} > 0$ from both sides of Equation (3.8), we get

$$1 \le \frac{\Delta_t}{\Delta_{t+1}} \le 2LR^2\left(\frac{1}{\Delta_{t+1}} - \frac{1}{\Delta_t}\right). \tag{3.9}$$

Summing the Equation (3.9) over $t = 1, \ldots, T-1$ yields

$$\frac{1}{\Delta_T} - \frac{1}{\Delta_1} = \sum_{t=1}^{T-1}\left(\frac{1}{\Delta_{t+1}} - \frac{1}{\Delta_t}\right) \ge \sum_{t=1}^{T}\frac{1}{2LR^2} = \frac{T-1}{2LR^2}. \tag{3.10}$$

We proceed by upper bounding the gap Δ_1. By the smoothness of f (Equation (2.1)) and the optimality condition (Theorem 2.5), we have

$$\Delta_1 = f(x_1) - f(x^\star) \leq \langle \nabla f(x^\star), x_1 - x^\star \rangle + \frac{L}{2}\|x_1 - x^\star\|^2 \leq \frac{LR^2}{2}. \tag{3.11}$$

Putting Equation (3.10) and Equation (3.11) together gives

$$\frac{1}{\Delta_T} \geq \frac{T-1}{2LR^2} + \frac{1}{\Delta_1} \geq \frac{T+3}{2LR^2} \geq \frac{T}{2LR^2}.$$

Rearranging, we obtain

$$f(x_T) - f(x^\star) \leq \frac{2LR^2}{T}.$$

□

In other words, Gradient Descent has a convergence rate of $O(1/T)$.

Remark 3.5. This is essentially Corollary 2.1.2 in *Lectures on Convex Optimization.*

Question 3.6. Can we prove the existence of the constant R in Theorem 3.4? Moreover, is it possible to obtain a convergence guarantee of Gradient Descent for the constrained cases?

3.2 Mirror Descent

Let $\mathcal{X} \subseteq \mathbb{R}^n$ be a closed convex set. Let $f : \mathbb{R}^n \to \mathbb{R}$ be convex differentiable (not necessarily smooth) on $\mathcal{X}$, and let $h : \mathbb{R}^n \to \mathbb{R}$ be 1-strongly convex with respect to $\|\cdot\|$ on $\mathcal{X}$. Let $\eta > 0$ be the fixed step size. In each round $t = 1, \ldots, T$, Mirror Descent with step size η iterates as

$$x_{t+1} \leftarrow \operatorname*{arg\,min}_{y \in \mathcal{X}} \left\{ \langle \nabla f(x_t), y - x_t \rangle + \frac{1}{\eta} D_h(y \,\|\, x_t) \right\},$$

where we choose the initial point x_1 is chosen as any feasible point in the constraint set $\mathcal{X}$. The convergence analysis of Mirror Descent relies heavily on lower bounding of the optimal value $f(x^\star)$ as follows. By the convexity of f, we can write

$$f(x^\star) \geq f(x_t) + \langle \nabla f(x_t), x^\star - x_t \rangle .$$

We further lower bound the inner product $\langle \nabla f(x_t), x^\star - x_t \rangle$ by decomposing it into

$$\langle \nabla f(x_t), x^\star - x_t \rangle = \langle \nabla f(x_t), x^\star - x_{t+1} \rangle + \langle \nabla f(x_t), x_{t+1} - x_t \rangle .$$

Lemma 3.7. *Let $\mathcal{X} \subseteq \mathbb{R}^n$ be a closed convex set, let $f : \mathbb{R}^n \to \mathbb{R}$ be convex differentiable on $\mathcal{X}$, and let $h : \mathbb{R}^n \to \mathbb{R}$ be 1-strongly convex with respect to $\|\cdot\|$ on $\mathcal{X}$. Set*

$$x_{t+1} \leftarrow \operatorname*{arg\,min}_{y \in \mathcal{X}} \left\{ \langle \nabla f(x_t), y - x_t \rangle + \frac{1}{\eta} D_h(y \,\|\, x_t) \right\},$$

where $\eta > 0$ is the fixed step size. Then, we have

$$\begin{aligned} f(x_t) - f(u) &\leq \langle \nabla f(x_t), x_t - u \rangle && (3.12) \\ &\leq \frac{\eta}{2}\|\nabla f(x_t)\|_*^2 + \frac{1}{\eta}(D_h(u \,\|\, x_t) - D_h(u \,\|\, x_{t+1})), && (3.13) \end{aligned}$$

for all $u \in \mathcal{X}$.

Proof. Let $u \in \mathcal{X}$ be fixed. Equation (3.12) follows immediately from the convexity of f. We prove Equation (3.13) by decomposing the inner product $\langle \nabla f(x_t), x_t - u \rangle$. We have

$$\begin{aligned} \langle \nabla f(x_t), x_t - u \rangle &= \langle \nabla f(x_t), x_t - x_{t+1} \rangle + \langle \nabla f(x_t), x_{t+1} - u \rangle \\ &\leq \langle \nabla f(x_t), x_t - x_{t+1} \rangle + \frac{1}{\eta} \langle \nabla h(x_{t+1}) - \nabla h(x_t), u - x_{t+1} \rangle \\ &= \langle \nabla f(x_t), x_t - x_{t+1} \rangle + \frac{1}{\eta}(D_h(u \,\|\, x_t) - D_h(u \,\|\, x_{t+1}) - D_h(x_{t+1} \,\|\, x_t)) \\ &= \|\nabla f(x_t)\|_* \|x_t - x_{t+1}\| - \frac{1}{\eta} D_h(x_{t+1} \,\|\, x_t) + \frac{1}{\eta}(D_h(u \,\|\, x_t) - D_h(u \,\|\, x_{t+1})) \\ &\leq \|\nabla f(x_t)\|_* \|x_t - x_{t+1}\| - \frac{1}{2\eta}\|x_{t+1} - x_t\|^2 + \frac{1}{\eta}(D_h(u \,\|\, x_t) - D_h(u \,\|\, x_{t+1})) \\ &\leq \frac{\eta}{2}\|\nabla f(x_t)\|_*^2 + \frac{1}{\eta}(D_h(u \,\|\, x_t) - D_h(u \,\|\, x_{t+1})), \end{aligned}$$

where the first inequality follows from the optimality condition (Theorem 2.5), the second equality follows from the three-point equality (Lemma 2.4), the third equality follows from Hölder's inequality, the second inequality follows from the strong convexity of h, and the last inequality follows by optimizing a quadratic function. This completes the proof. □

Remark 3.8. Observe that Equation (3.13) is tighter when the gradient norm $\|\nabla f(x_t)\|_*$ is small. This result is complementary to Remark 3.3.

Theorem 3.9 (Mirror descent guarantee)**.** *Let $\mathcal{X} \subseteq \mathbb{R}^n$ be a closed convex set, let $f : \mathbb{R}^n \to \mathbb{R}$ be convex differentiable on $\mathcal{X}$, and let $h : \mathbb{R}^n \to \mathbb{R}$ be 1-strongly convex with respect to $\|\cdot\|$ on $\mathcal{X}$. Suppose that $\|\nabla f(x_t)\|_* \leq G$ for $t = 1, \ldots, T$ for some $G > 0$, and that $D_h(u \,\|\, x_1) \leq D$ for all $u \in \mathcal{X}$ for some $D > 0$. Then,* Mirror Descent *with step size $\eta \coloneqq \sqrt{2D/(G^2 T)}$ achieves*

$$f(\bar{x}) - f(x^\star) \leq \sqrt{\frac{2DG^2}{T}},$$

where $\bar{x} \coloneqq \sum_{t=1}^{T} x_t/T$ is the averages of the x_t's and $x^\star$ is a minimizer of f over $\mathcal{X}$.

Proof. Let $u \in \mathcal{X}$ be arbitrary. Summing up Equation (3.13) over $t = 1, \ldots, T$ yields

$$\begin{aligned}
\sum_{t=1}^{T}(f(x_t) - f(u)) &\leq \frac{\eta}{2}\sum_{t=1}^{T}\|\nabla f(x_t)\|_*^2 + \frac{1}{\eta}\sum_{t=1}^{T}(D_h(u \,\|\, x_t) - D_h(u \,\|\, x_{t+1})) \\
&= \frac{\eta}{2}\sum_{t=1}^{T}\|\nabla f(x_t)\|_*^2 + \frac{1}{\eta}D_h(u \,\|\, x_1) - \frac{1}{\eta}D_h(u \,\|\, x_{T+1}) \\
&\leq \frac{\eta}{2}\sum_{t=1}^{T}\|\nabla f(x_t)\|_*^2 + \frac{1}{\eta}D_h(u \,\|\, x_1), \\
&\leq \frac{\eta T G^2}{2} + \frac{D}{\eta},
\end{aligned} \tag{3.14}$$

where the second inequality follows from the fact that $D_h(u \,\|\, x_{T+1}) \geq 0$, and the last inequality follows from the bounded domain and bounded gradient assumptions. Let $\bar{x} \coloneqq \sum_{t=1}^{T} x_t/T$ be the average of the x_t's. By the convexity of f and Equation (3.14), we can write

$$T(f(\bar{x}) - f(u)) \leq \sum_{t=1}^{T}(f(x_t) - f(u)) \leq \frac{\eta T G^2}{2} + \frac{D}{\eta}. \tag{3.15}$$

Let $x^\star$ be a minimizer of f over $\mathcal{X}$. Substitute $u = x^\star$ into Equation (3.15), we obtain

$$f(\bar{x}) - f(x^\star) \leq \frac{\eta G^2}{2} + \frac{D}{\eta T}. \tag{3.16}$$

Substitute $\eta \coloneqq \sqrt{2D/(G^2 T)} > 0$ into Equation (3.16), we get

$$f(\bar{x}) - f(x^\star) \leq \sqrt{\frac{2DG^2}{T}}.$$

□

In other words, Mirror Descent has a convergence rate of $O(1/\sqrt{T})$.

3.3 Nesterov's Accelerated Gradient Method

Nesterov [Nes05] proposed a first-order method for smooth convex optimization problems with a convergence rate $O(1/T^2)$. In fact, this convergence rate is optimal. However, the convergence analysis of Nesterov's accelerated gradient method is not intuitive.

4 Fixed Step Accelerated Gradient Method

In this section, we consider the Accelerated Gradient Method in the unconstrained setting with a fixed step size. We call this version the Fixed Step Accelerated Gradient Method.

Let $f : \mathbb{R}^n \to \mathbb{R}$ be a convex differentiable function that is L-smooth with respect to $\|\cdot\|$ for some $L > 0$. In each iteration, Fixed Step Accelerated Gradient Method performs both a gradient step and a mirror step, while ensuring that the two steps are *linearly coupled.* As mentioned earlier in Remark 3.3 and Remark 3.8, the objective progress is larger when $\|\nabla f(x_t)\|_*$ is large, and the lower bound on $f(x^\star)$ is tighter when $\|\nabla f(x_t)\|_*$ is small. Fixed Step Accelerated Gradient Method balances this tradeoff by introducing a *coupling rate* τ. The Fixed Step Accelerated Gradient Method algorithm is as follows.

Algorithm 1 Fixed Step Accelerated Gradient Method

```
Inputs: A convex differentiable function $f : \mathbb{R}^n \to \mathbb{R}$ that is $L$-smooth with respect to $\|\cdot\|$, an 1-strongly convex function $h : \mathbb{R}^n \to \mathbb{R}$ with respect to $\|\cdot\|$, a coupling rate $\tau \in (0,1)$, a fixed step size $\eta > 0$, and an initial point $x_1 \in \mathbb{R}^n$.
Set $y_1 \leftarrow x_1$.
Set $z_1 \leftarrow x_1$.
for $t = 1, \ldots, T$ do
    Set $x_{t+1} \leftarrow \tau z_t + (1-\tau) y_t$.
    Performs a gradient step
        $y_{t+1} \leftarrow \arg\min_{y \in \mathbb{R}^n} \left\{ \langle \nabla f(x_{t+1}), y - x_{t+1} \rangle + \frac{L}{2} \|y - x_{t+1}\|^2 \right\}.$
    Performs a mirror step
        $z_{t+1} \leftarrow \arg\min_{z \in \mathbb{R}^n} \left\{ \langle \nabla f(x_{t+1}), z - z_t \rangle + \frac{1}{\eta} D_h(z \,\|\, z_t) \right\}.$
end for
```

Lemma 4.1. *Let $f : \mathbb{R}^n \to \mathbb{R}$ be a convex differentiable function that is L-smooth with respect to $\|\cdot\|$, and let $h : \mathbb{R}^n \to \mathbb{R}$ be 1-strongly convex with respect to $\|\cdot\|$. Let $\eta > 0$ be the fixed step size. Then,* Fixed Step Accelerated Gradient Method *achieves*

$$\begin{aligned}\langle \nabla f(x_{t+1}), z_t - u \rangle &\le \frac{\eta}{2} \|\nabla f(x_{t+1})\|_*^2 + \frac{1}{\eta}(D_h(u \,\|\, z_t) - D_h(u \,\|\, z_{t+1})) \qquad (4.1)\\ &\le \eta L (f(x_{t+1}) - f(y_{t+1})) + \frac{1}{\eta}(D_h(u \,\|\, z_t) - D_h(u \,\|\, z_{t+1})), \qquad (4.2)\end{aligned}$$

for all $u \in \mathbb{R}^n$.

Proof. Equation (4.1) follows from Lemma 3.7, and Equation (4.2) follows from Lemma 3.2. □

Lemma 4.2 (Unconstrained coupling lemma)**.** *Let $f : \mathbb{R}^n \to \mathbb{R}$ be a convex differentiable function that is L-smooth with respect to $\|\cdot\|$, and let $h : \mathbb{R}^n \to \mathbb{R}$ be 1-strongly convex with respect to $\|\cdot\|$. Let $\eta > 0$ be the fixed step size and let $\tau \in (0,1)$ such that $(1-\tau)/\tau = \eta L$. Then,* Fixed Step Accelerated Gradient Method *achieves*

$$\langle \nabla f(x_{t+1}), x_{t+1} - u \rangle \le \eta L (f(y_t) - f(y_{t+1})) + \frac{1}{\eta}(D_h(u \,\|\, z_t) - D_h(u \,\|\, z_{t+1})), \qquad (4.3)$$

for all $u \in \mathbb{R}^n$.

Proof. Let $u \in \mathbb{R}^n$ be fixed. Since we set $x_{t+1} = \tau z_t + (1-\tau) y_t$ and $(1-\tau)/\tau = \eta L$, we can write

$$x_{t+1} - z_t = \frac{1-\tau}{\tau}(y_t - x_{t+1}) = \eta L (y_t - x_{t+1}). \qquad (4.4)$$

Thus, we have

$$\begin{aligned}\langle \nabla f(x_{t+1}), x_{t+1} - u \rangle &= \langle \nabla f(x_{t+1}), x_{t+1} - z_t \rangle + \langle \nabla f(x_{t+1}), z_t - u \rangle\\ &= \eta L \langle \nabla f(x_{t+1}), y_t - x_{t+1} \rangle + \langle \nabla f(x_{t+1}), z_t - u \rangle\\ &\le \eta L (f(y_t) - f(x_{t+1})) + \langle \nabla f(x_{t+1}), z_t - u \rangle\\ &\le \eta L (f(y_t) - f(y_{t+1})) + \frac{1}{\eta}(D_h(u \,\|\, z_t) - D_h(u \,\|\, z_{t+1})),\end{aligned}$$

where the second equality follows from Equation (4.4), the first inequality follows from the convexity of f, and the last inequality follows from Lemma 4.1. □

Observe that the coupling rate τ is chosen precisely to balance the objective decrease $f(x_{t+1}) - f(y_{t+1})$ and the potential objective increase $f(y_t) - f(y_{t+1})$ in the proof of Lemma 4.2. Since $(1-\tau)/\tau = \eta L$, we have $\tau = 1/(\eta L + 1)$ and thus τ is indeed a real in the open interval $(0, 1)$. Therefore, this choice of coupling rate is feasible as long as we have the knowledge of the smoothness parameter L.

Theorem 4.3 (Fixed step accelerated gradient method guarantee)**.** *Let $f : \mathbb{R}^n \to \mathbb{R}$ be a convex differentiable function that is L-smooth with respect to $\|\cdot\|$, and let $h : \mathbb{R}^n \to \mathbb{R}$ be 1-strongly convex with respect to $\|\cdot\|$. Let $x^\star$ be a minimizer of f. Suppose that $f(x_1) - f(x^\star) \le d$ and that $D_h(u \,\|\, x_1) \le D$ for all $u \in \mathbb{R}^n$ for some $D > 0$. Set $\eta := \sqrt{D/(Ld)}$ and $\tau \in (0, 1)$ such that $(1-\tau)/\tau = \eta L$. Then,* Fixed Step Accelerated Gradient Method *achieves*

$$f(\bar{x}) - f(x^\star) \le \frac{2\sqrt{DLd}}{T},$$

where $\bar{x} := \sum_{t=1}^{T} x_{t+1}/T$ is the average of the iterates.

Proof. Let $u \in \mathbb{R}^n$ be arbitrary. Summing up Equation (4.3) over $t = 1, \dots, T$ yields

$$\begin{aligned}
\sum_{t=1}^{T} \langle \nabla f(x_{t+1}), x_{t+1} - u \rangle &\le \eta L \sum_{t=1}^{T} (f(y_t) - f(y_{t+1})) + \frac{1}{\eta} \sum_{t=1}^{T} (D_h(u \,\|\, z_t) - D_h(u \,\|\, z_{t+1})) \\
&= \eta L (f(y_1) - f(y_{T+1})) + \frac{1}{\eta} D_h(u \,\|\, z_1) - \frac{1}{\eta} D_h(u \,\|\, z_{T+1}) \\
&\le \eta L (f(y_1) - f(y_{T+1})) + \frac{1}{\eta} D_h(u \,\|\, z_1) \\
&\le \eta L d + \frac{D}{\eta},
\end{aligned}$$

where the second inequality follows from the fact that $D_h(u \,\|\, z_{T+1}) \ge 0$, and the last inequality follows from the assumption that $f(x_1) - f(x^\star) \le d$ and $D_h(u \,\|\, x_1) \le D$. Let $\bar{x} := \sum_{t=1}^{T} x_{t+1}/T$ be the average of the iterates. By the convexity of f, we can write

$$T(f(\bar{x}) - f(u)) \le \sum_{t=1}^{T} (f(x_{t+1}) - f(u)) \tag{4.5}$$

Substitute $u = x^\star$ and $\eta = \sqrt{D/(Ld)}$ into Equation (4.5), we obtain

$$f(\bar{x}) - f(x^\star) \le \frac{2\sqrt{DLd}}{T}.$$

□

Corollary 4.4. *Let $f : \mathbb{R}^n \to \mathbb{R}$ be a convex differentiable function that is L-smooth with respect to $\|\cdot\|$, and let $h : \mathbb{R}^n \to \mathbb{R}$ be 1-strongly convex with respect to $\|\cdot\|$. Let $x^\star$ be a minimizer of f. Suppose that $f(x_1) - f(x^\star) \le d$ and that $D_h(u \,\|\, x_1) \le D$ for all $u \in \mathbb{R}^n$ for some $D > 0$. Set $\eta = \sqrt{D/(Ld)}$ and $\tau \in (0, 1)$ such that $(1-\tau)/\tau = \eta L$. After $T = \lceil 4\sqrt{DL/d} \rceil$ iterations,* Fixed Step Accelerated Gradient Method *achieves*

$$f(\bar{x}) - f(x^\star) \le \frac{d}{2},$$

where $\bar{x} := \sum_{t=1}^{T} x_{t+1}/T$ is the average of the iterates.

Proof. From Theorem 4.3, we know that after T iterations, Fixed Step Accelerated Gradient Method guarantees

$$f(\bar{x}) - f(x^\star) \le \frac{2\sqrt{DLd}}{T},$$

where $\bar{x} := \sum_{t=1}^{T} x_{t+1}/T$. Note that $2\sqrt{DLd}/T \le d/2$ when $T \ge 4\sqrt{DL/d}$. Thus, after $T = \lceil 4\sqrt{DL/d} \rceil$ iterations, we have

$$f(\bar{x}) - f(x^\star) \le \frac{d}{2}.$$

□

We can show that Fixed Step Accelerated Gradient Method converges to an ε-approximate solution in $\Omega(\sqrt{DL/\varepsilon})$ iterations. Let $x_1 \in \mathbb{R}^n$ and suppose that $f(x_1) - f(x^\star) \le d$. From Corollary 4.4, we know that

$$f(\bar{x}) - f(x^\star) \le \frac{d}{2},$$

after $T = \lceil 4\sqrt{DL/d}\rceil$ iterations, where $\bar{x} := \sum_{t=1}^{T} x_{t+1}/T$ is the average of the x_t's. Then, we could restart the algorithm with $u_1 := \bar{x}$ as the initial point. Again, from Corollary 4.4, after $T = \lceil 4\sqrt{2DL/d}\rceil$ iterations, we have

$$f(\bar{u}) - f(x^\star) \leq \frac{d}{4},$$

where $\bar{u} := \sum_{t=1}^{T} u_{t+1}/T$ is the average of the u_t's. Continuing in this fashion, we can see that Fixed Step Accelerated Gradient Method converges to an ε-approximate solution in

$$\Omega\left(\sqrt{\frac{DL}{\varepsilon}} + \sqrt{\frac{DL}{2\varepsilon}} + \dots\right) = \Omega\left(\sqrt{\frac{DL}{\varepsilon}}\right)$$

iterations. Equivalently, Fixed Step Accelerated Gradient Method has a convergence rate of $O(1/T^2)$, matching that of Nesterov's accelerated gradient method [Nes05].

We point out a few drawbacks of the Fixed Step Accelerated Gradient Method. In the proof of Theorem 4.3, the step size η is determined with the knowledge of the parameters D, d and L, where $D = \max_{u \in \mathbb{R}^n} D_h(u \,\|\, x_1)$, d is an upper bound on the difference between the $f(x_1)$ and $f(x^\star)$, and L is the smoothness parameter. Moreover, in order to obtain an ε-approximate solution, we need to restart the algorithm. Nonetheless, we will show that the general version of Fixed Step Accelerated Gradient Method overcomes these flaws.

5 Accelerated Gradient Method

In this section, we extend Fixed Step Accelerated Gradient Method to the unconstrained case with time-varying step sizes. The resulting algorithm is referred to as Accelerated Gradient Method. The main theorem, Theorem 5.3, shows that Accelerated Gradient Method match the $O(1/T^2)$ convergence rate of Nesterov's accelerated gradient method.

Algorithm 2 Accelerated Gradient Method

Inputs: A closed convex set $\mathcal{X} \subseteq \mathbb{R}^n$, a convex differentiable function $f : \mathcal{X} \to \mathbb{R}$ that is L-smooth with respect to $\|\cdot\|$, an 1-strongly convex function $h : \mathcal{X} \to \mathbb{R}$ with respect to $\|\cdot\|$, and an initial point $x_1 \in \mathcal{X}$.

```
Set $y_1 \leftarrow x_1$.
Set $z_1 \leftarrow x_1$.
for $t = 1, \dots, T$ do
    Set $\eta_t \leftarrow {}^{t+2}\!/_{2L}$.
    Set $\tau_t \leftarrow {}^{1}\!/_{\eta_t L}$.
    Set $x_{t+1} \leftarrow \tau_t z_t + (1 - \tau_t) y_t$.
    Performs a gradient step
        $$y_{t+1} \leftarrow \operatorname*{arg\,min}_{y \in \mathcal{X}} \left\{ \langle \nabla f(x_{t+1}), y - x_{t+1} \rangle + \frac{L}{2}\|y - x_{t+1}\|^2 \right\}.$$
    Performs a mirror step
        $$z_{t+1} \leftarrow \operatorname*{arg\,min}_{z \in \mathcal{X}} \left\{ \langle \nabla f(x_{t+1}), z - z_t \rangle + \frac{1}{\eta_t} D_h(z \,\|\, z_t) \right\}.$$
end for
```

Observe that the step size sequence $\{\eta_t\}_{t=1}^{T}$ is nondecreasing. Since $\tau_t = 1/(\eta_t L) = 2/(t+2)$, we know that the coupling rate sequence $\{\tau_t\}_{t=1}^{T}$ is a nonincreasing sequence in $(0, 1)$.

Lemma 5.1. *Let $\mathcal{X} \subseteq \mathbb{R}^n$ be a closed convex set, let $f : \mathbb{R}^n \to \mathbb{R}$ be a convex differentiable function that is L-smooth with respect to $\|\cdot\|$ on $\mathcal{X}$, and let $h : \mathbb{R}^n \to \mathbb{R}$ be 1-strongly convex with respect to $\|\cdot\|$ on $\mathcal{X}$. Then,* Accelerated Gradient Method *achieves*

$$\begin{aligned}
\langle \nabla f(x_{t+1}), z_t - u \rangle &\leq \eta_t L \mathsf{Prog}_f(x_{t+1}) + \frac{1}{\eta_t}(D_h(u \,\|\, z_t) - D_h(u \,\|\, z_{t+1})) \qquad (5.1)\\
&\leq \eta_t L(f(x_{t+1}) - f(y_{t+1})) + \frac{1}{\eta_t}(D_h(u \,\|\, z_t) - D_h(u \,\|\, z_{t+1})), \qquad (5.2)
\end{aligned}$$

for all $u \in \mathcal{X}$.

Proof. Let $u \in \mathcal{X}$. We can write

$$\begin{aligned}
\langle \nabla f(x_{t+1}), z_t - u\rangle &= \langle \nabla f(x_{t+1}), z_t - z_{t+1}\rangle + \langle \nabla f(x_{t+1}), z_{t+1} - u\rangle \\
&\leq \langle \nabla f(x_{t+1}), z_t - z_{t+1}\rangle + \frac{1}{\eta_t}\langle \nabla h(z_{t+1}) - \nabla h(z_t), u - z_{t+1}\rangle \\
&= \langle \nabla f(x_{t+1}), z_t - z_{t+1}\rangle + \frac{1}{\eta_t}(D_h(u \,\|\, z_t) - D_h(u \,\|\, z_{t+1}) - D_h(z_{t+1} \,\|\, z_t)) \\
&\leq \langle \nabla f(x_{t+1}), z_t - z_{t+1}\rangle - \frac{1}{2\eta_t}\|z_{t+1} - z_t\|^2 + \frac{1}{\eta_t}(D_h(u \,\|\, z_t) - D_h(u \,\|\, z_{t+1})), \qquad (5.3)
\end{aligned}$$

where the first inequality follows from the optimality condition (Theorem 2.5), the second equality follows from the three-point equality (Lemma 2.4), and the last inequality follows from the strong convexity of h. Let $v := \tau_t z_{t+1} + (1-\tau_t)y_t \in \mathcal{X}$. We have

$$x_{t+1} - v = \tau_t z_t + (1-\tau_t)y_t - \tau_t z_{t+1} - (1-\tau_t)y_t = \tau_t(z_t - z_{t+1}). \tag{5.4}$$

By Equation (5.4), we can write

$$\begin{aligned}
\langle f(x_{t+1}), z_t - z_{t+1}\rangle - \frac{1}{2\eta_t}\|z_{t+1} - z_t\|^2 &= \frac{1}{\tau_t}\langle \nabla f(x_{t+1}), x_{t+1} - v\rangle - \frac{1}{2\eta_t\tau_t^2}\|x_{t+1} - v\|^2 \\
&= \eta_t L\langle \nabla f(x_{t+1}), x_{t+1} - v\rangle - \frac{\eta_t L^2}{2}\|x_{t+1} - v\|^2 \\
&= \eta_t L(\langle \nabla f(x_{t+1}), x_{t+1} - v\rangle - \frac{L}{2}\|x_{t+1} - v\|^2) \\
&\leq \eta_t L \mathsf{Prog}_f(x_{t+1}), \qquad (5.5)
\end{aligned}$$

where the second equality follows from the fact that $\tau_t = 1/(\eta_t L)$, and the last inequality follows from the definition of objective progress (Definition 3.1). By combining Equation (5.3) and Equation (5.5), we prove Equation (5.1). Equation (5.2) follows immediately from the gradient descent lemma (Lemma 3.2). □

Lemma 5.2 (Constrained coupling lemma)**.** *Let $\mathcal{X} \subseteq \mathbb{R}^n$ be a closed convex set, let $f : \mathbb{R}^n \to \mathbb{R}$ be a convex differentiable function that is L-smooth with respect to $\|\cdot\|$ on $\mathcal{X}$, and let $h : \mathcal{X} \to \mathbb{R}$ be 1-strongly convex with respect to $\|\cdot\|$ on $\mathcal{X}$. Then,* Accelerated Gradient Method *achieves*

$$(\eta_t^2 L)f(y_{t+1}) - (\eta_t^2 L - \eta_t)f(y_t) + D_h(u \,\|\, z_{t+1}) - D_h(u \,\|\, z_t) \leq \eta_t f(u), \tag{5.6}$$

for all $u \in \mathcal{X}$.

Proof. Let $u \in \mathcal{X}$. Since we set $x_{t+1} \leftarrow \tau_t z_t + (1-\tau_t)y_t$ and $\tau_t \leftarrow 1/(\eta_t L)$ in Accelerated Gradient Method, we have

$$x_{t+1} - z_t = \frac{1-\tau_t}{\tau_t}(y_t - x_{t+1}) = (\eta_t L - 1)(y_t - x_{t+1}). \tag{5.7}$$

We can write

$$\begin{aligned}
f(x_{t+1}) - f(u) &\leq \langle \nabla f(x_{t+1}), x_{t+1} - u\rangle \\
&= \langle \nabla f(x_{t+1}), x_{t+1} - z_t\rangle + \langle \nabla f(x_{t+1}), z_t - u\rangle \\
&= (\eta_t L - 1)\langle \nabla f(x_{t+1}), y_t - x_{t+1}\rangle + \langle \nabla f(x_{t+1}), z_t - u\rangle \\
&\leq (\eta_t L - 1)(f(y_t) - f(x_{t+1})) + \langle \nabla f(x_{t+1}), z_t - u\rangle \\
&= f(x_{t+1}) + (\eta_t L - 1)f(y_t) - \eta_t L f(y_{t+1}) + \frac{1}{\eta_t}(D_h(u \,\|\, z_t) - D_h(u \,\|\, z_{t+1})), \qquad (5.8)
\end{aligned}$$

where the first inequality follows from the convexity of f, the second equality follows from Equation (5.7), the second inequality again follows from the convexity of f, and the last equality follows from Lemma 5.1. By multiplying both sides of Equation (5.8) by η_t and rearranging, we arrive at Equation (5.6). □

Theorem 5.3 (Accelerated gradient method guarantee)**.** *Let $\mathcal{X} \subseteq \mathbb{R}^n$ be a closed convex set, let $f : \mathbb{R}^n \to \mathbb{R}$ be a convex differentiable function that is L-smooth with respect to $\|\cdot\|$ on $\mathcal{X}$, and let $h : \mathbb{R}^n \to \mathbb{R}$ be 1-strongly convex with respect to $\|\cdot\|$ on $\mathcal{X}$. Let $x^\star$ be a minimizer of f over $\mathcal{X}$. Suppose that $D_h(u \,\|\, x_1) \leq D$ for all $u \in \mathcal{X}$ for some $D > 0$. Then, the* Accelerated Gradient Method *outputs $y_{T+1} \in \mathcal{X}$ such that*

$$f(y_{T+1}) - f(x^\star) \leq \frac{4DL + 4(f(y_1) - f(x^\star))}{T(T+4)}.$$

Proof. Since we set $\eta_t \leftarrow t + 2/(2L)$, we have

$$(\eta_t^2 L - \eta_t) - \eta_{t-1}^2 L = \left(\frac{(t+2)^2}{4L} - \frac{t+2}{2L}\right) - \frac{(t+1)^2}{4L} = -\frac{1}{4L}. \tag{5.9}$$

By Equation (5.9), we can rewrite Equation (5.6) as

$$(\eta_t^2 L) f(y_{t+1}) - (\eta_{t-1}^2 L) f(y_t) + \frac{1}{4L} f(y_t) + D_h(u \,\|\, z_{t+1}) - D_h(u \,\|\, z_t) \leq \eta_t f(u), \tag{5.10}$$

for all $u \in \mathcal{X}$. Summing up the Equation (5.10) over $t = 1, \ldots, T$ yields

$$\begin{aligned}
\sum_{t=1}^{T} \eta_t f(u) &\geq \sum_{t=1}^{T} ((\eta_t^2 L) f(y_{t+1}) - (\eta_{t-1}^2 L) f(y_t)) + \frac{1}{4L} \sum_{t=1}^{T} f(y_t) + \sum_{t=1}^{T} (D_h(u \,\|\, z_{t+1}) - D_h(u \,\|\, z_t)) \\
&= (\eta_T^2 L) f(y_{T+1}) - (\eta_0^2 L) f(y_1) + \frac{1}{4L} \sum_{t=1}^{T} f(y_t) + D_h(u \,\|\, z_{T+1}) - D_h(u \,\|\, z_1) \\
&\geq (\eta_T^2 L) f(y_{T+1}) - (\eta_0^2 L) f(y_1) + \frac{T}{4L} f(x^\star) + D_h(u \,\|\, z_{T+1}) - D_h(u \,\|\, z_1) \\
&\geq (\eta_T^2 L) f(y_{T+1}) - (\eta_0^2 L) f(y_1) + \frac{T}{4L} f(x^\star) - D,
\end{aligned} \tag{5.11}$$

where the second inequality follows from the fact that $f(x_t) \geq f(x^\star)$ for $t = 1, \ldots, T$, and the last inequality follows from the fact that $D_h(u \,\|\, z_{T+1}) \geq 0$ and the assumption that $D_h(u \,\|\, x_1) \leq D$ for all $u \in \mathcal{X}$. Rearrange Equation (5.11), we get

$$(\eta_T^2 L) f(y_{T+1}) - \sum_{t=1}^{T} \eta_t f(u) \leq (\eta_0^2 L) f(y_1) - \frac{T}{4L} f(x^\star) + D. \tag{5.12}$$

Substitute $\eta_t = (t+2)/(2L)$ and $u = x^\star$ into the Equation (5.12), we have

$$\frac{(T+2)^2}{4L} f(y_{T+1}) - \frac{T(T+5)}{4L} f(x^\star) \leq \frac{f(y_1)}{L} - \frac{T}{4L} f(x^\star) + D \tag{5.13}$$

We can further simplify Equation (5.13) into

$$\frac{T(T+4)}{4L} (f(y_{T+1}) - f(x^\star)) \leq \frac{f(y_1) - f(y_{T+1})}{L} + D \leq \frac{f(y_1) - f(x^\star)}{L} + D, \tag{5.14}$$

where the last inequality follows from the fact that $f(y_{T+1}) \geq f(x^\star)$. By dividing $T(T+4)/(4L)$ from both sides of Equation (5.14), we obtain

$$f(y_{T+1}) - f(x^\star) \leq \frac{4LD + 4(f(y_1) - f(x^\star))}{T(T+4)}.$$

□

In other words, Accelerated Gradient Method has a convergence rate $O(1/T^2)$.

Remark 5.4. The bound in Theorem 5.3 is slightly different than the one given in the paper. However, the convergence rates are the same.

Question 5.5. In Theorem 5.3, we have the term $f(y_1 - f(x^\star))$ in our upper bound. Is it possible to further bound $f(y_1) - f(x^\star)$?

Question 5.6. What are the differences between Allen-Zhu and Orecchia's accelerated gradient method and Nesterov's accelerated gradient method?

6 Conclusion

This paper consider smooth convex optimization problems. For smooth objectives, gradient descent achieves a $O(1/\sqrt{T})$ convergence rate while mirror descent achieves a $O(1/\sqrt{T})$ convergence rate. The optimal convergence rate for such problems is $O(1/T^2)$, which is achieved by Nesterov's accelerated gradient method. Allen-Zhu and Orrecchia proposed a technique called linear coupling that unifies gradient descent and mirror descent. The resulting algorithm Accelerated Gradient Method enjoys the same convergence rate as Nesterov's accelerated gradient method while having a more intuitive convergence analysis than Nesterov's method.

References

[1] Zeyuan Allen-Zhu and Lorenzo Orecchia. “Linear Coupling: An Ultimate Unification of Gradient and Mirror Descent”. In: *arXiv preprint arXiv:1407.1537* (2014).

[2] Yu Nesterov. “Smooth Minimization of Non-smooth Functions”. In: *Mathematical programming* 103 (2005), pp. 127–152.

[3] Yurii Nesterov. *Lectures on Convex Optimization.* Vol. 137. Springer, 2018.