MA 574 Reading Project: Linear Coupling, An Ultimate Unification of Gradient and Mirror Descent

Zitao Song

1 Introduction

First-order methods play a central role in large-scale machine learning. Throughout the semester, MA 574 at Purdue University covers a broad range of first-order methods, starting from simple gradient descent algorithms on smooth functions, covering different splitting methods and dual methods for non-smooth functions, and ending with first-order methods on the Riemannian manifold. From my own perspective, modern optimization techniques emphasize leveraging the intrinsic geometric structure of the data and deriving faster algorithms for the given problem. That is where mirror descent and Riemannian optimization algorithms come into the game by solving optimization problems on non-Euclidean spaces. Furthermore, we have also witnessed the widely used linear interpolation techniques in the optimization communities, i.e., the linear interpolation between previous updates and the reflection operators gives us the general Douglas-Rachford splitting and the linear interpolation between uniform sampling weight and Lipchitz-biased sampling weight generates Importance Sampling based Stochastic Gradient Descent algorithms [2]. By linear interpolation, we hope we can derive the convergence result that is at least as good as the convergence result before doing the interpolation.

In this reading project, we target the black-box constraint optimization problem $\min_Q f(x)$ where f is convex and smooth, Q is the constraint set of the problem. We talk about another linear interpolation algorithm which is between gradient descent and mirror descent, named after Linear Coupling [1]. Since gradient descent has a nice descent property for each iteration (primal progress on function values) and mirror descent tackles the dual problem by constructing stronger lower bounds to the optimum through minimizing the maximized regret of each iteration, i.e., $f(x_k)-f(x) \leq \langle \nabla f(x_k), x_k - x\rangle \leq M$. The author of Linear Coupling observes that the performances of gradient and mirror descent are complementary so that faster algorithms can be designed by linearly interpolating the two. Interestingly, the paper has shown the convergence result of Nesterov's accelerated gradient methods, i.e., $\Omega(\frac{1}{\sqrt{\epsilon}})$ could be reconstructed using a linear coupling, which gives a cleaner interpretation than the 'algebraic miracle' in Nesterov's original proofs.

In this short report, we will first derive the decent lemma and convergence result for gradient descent, which is slightly different from what we've taught in the lecture and in the paper's proof in Section (2). Next, in Section (3) and (5), we derive the lemma for the regret upper bound in each iteration for both mirror descent and linear coupling, which will finally build up the convergence result for both of them by telescoping the regret upper bound. In Section (6), we include a toy numerical experiment of the least square problem to test the efficiency of the proposed linear coupling problem.

2 Gradient Descent Convergence Result

Definition 2.1 (Gradient Descent Update)**.** *The gradient descent step with step length η can be described as*

$$x_{k+1} = Grad_\eta(x_t) \quad where \quad x_{k+1} \leftarrow \arg\min_{y\in\mathbb{R}^n} \Big\{ \langle \nabla f(x_k), y - x_k\rangle + \frac{1}{2\eta}||y-x_k||^2 \Big\}, \tag{1}$$

the step takes the familiar additive form $x_{k+1} = x_k - \eta \nabla f(x_k)$.

Lemma 2.2 (Descent Lemma for GD)**.** *Under the assumption of an L-smooth function f, i.e., $||\nabla f(x) - \nabla f(y)|| \leq L||x-y||$ for every x, y, we have a global quadratic upper bound on the function*

around a query point x $\forall y$:

$$f(y) \leq f(x) + \langle \nabla f(x), y - x \rangle + \frac{L}{2}||y - x||^2, \tag{2}$$

furthermore, under the gradient descent update (2.1) with step size η,i.e., $x_{k+1} = x_k - \eta \nabla f(x)$, the magnitude for each iteration update is at least

$$f(x_k) - f(x_{k+1}) \geq (\eta - \frac{L}{2}\eta^2)||\nabla f(x)||_*^2, \tag{3}$$

when $\eta \in (0, \frac{2}{L})$, $f(x_k) - f(x_{k+1})$ is non-negative. When $\eta = \frac{1}{L}$, we have $f(x_k) - f(x_{k+1}) \geq \frac{1}{2L}||\nabla f(x)||_*^2$.

Proof. The proof for Equation (2) is a direct application of the Multivariate Quadratic Taylor's Theorem. To prove Equation (3), we just need to replace y and x by x_{k+1} and x_k in Equation (2), and we get

$$\begin{aligned} f(x_{k+1}) &\leq f(x_k) + \langle \nabla f(x_k), x_{k+1} - x_k \rangle + \frac{L}{2}||x_{k+1} - x_k||^2 \\ &\leq f(x_k) + \langle \nabla f(x_k), -\eta \nabla f(x_k) \rangle + \frac{L}{2}|| - \eta \nabla f(x_k)||_*^2 \\ &\leq f(x_k) + (\frac{L}{2}\eta^2 - \eta)||\nabla f(x)||_*^2 \end{aligned}$$

□

Lemma 2.3. *For x_{k+1} and x_k in gradient descent method (2.1), i.e., $x_{k+1} = Grad_\eta(x_k)$, we have*

$$\forall x \quad \eta \langle x_k - x_*, \nabla f(x_k) \rangle = \frac{1}{2}||x_k - x_*||^2 - \frac{1}{2}||x_{k+1} - x_*||^2 + \frac{\eta^2}{2}||\nabla f(x_k)||_*^2 \tag{4}$$

Proof.

$$\begin{aligned} &||x_{k+1} - x_*||^2 - ||x_k - x_*||^2 \\ =&||x_k - \eta \nabla f(x_k) - x_*||^2 - ||x_k - x_*||^2 \\ =&||x_k - x_*|| - 2\eta \langle x_k - x_*, \nabla f(x_k) \rangle + \eta^2 ||\nabla f(x_k)||_*^2 - ||x_k - x_*||^2 \\ =& - 2\eta \langle x_k - x_*, \nabla f(x_k) \rangle + \eta^2 ||\nabla f(x_k)||_*^2, \end{aligned}$$

Rearranging equations we have,

$$\eta \langle x_k - x_*, \nabla f(x_k) \rangle = \frac{1}{2}||x_k - x_*||^2 - \frac{1}{2}||x_{k+1} - x_*||^* + \frac{\eta^2}{2}||\nabla f(x_k)||_*^2$$

□

Theorem 2.4 (Convergence Result for GD)**.** *Assume $f(x)$ is convex and L-smooth and assume $f(x)$ has a global minimizer $f(x) \geq f(x_*)$ $\forall x$. Then gradient descent method (2.1) starting with x_0 and taking constant step size $\frac{1}{L}$ in every iteration produces $x_1, \ldots, x_T$ such that, the following holds:*

$$f(x_T) - f(x_*) \leq \frac{1}{T}\sum_{k=0}^{T-1} f(x_{k+1}) - \frac{1}{T}\sum_{k=0}^{T-1} f(x_*) \leq \frac{L}{2T}||x_0 - x_*||^2, \tag{5}$$

Intuitively, to reach an accuracy of ϵ, we need $\Omega(\frac{||x_0 - x_*||^2 L}{2\epsilon})$ iterations.

Proof. Here we give a proof different from the proof in [1] and the lecture notes.

$$
\begin{aligned}
f(x_{k+1}) &\le f(x_k) - \frac{1}{2L}||\nabla f(x_k)||_*^2 && \text{apply Descent Lemma (2.2)}\\
f(x_{k+1}) - f(x_*) &\le f(x_k) - f(x_*) - \frac{1}{2L}||\nabla f(x_k)||_*^2 &&\\
f(x_{k+1}) - f(x_*) &\le \langle \nabla f(x_k), x_k - x_*\rangle - \frac{1}{2L}||\nabla f(x_k)||_*^2 && \text{apply convexity of } f(x)\\
f(x_{k+1}) - f(x_*) &\le \frac{L}{2}||x_k - x_*||^2 - \frac{L}{2}||x_{k+1} - x_*||^* && \text{apply Lemma (2.3)}\\
\sum_{k=0}^{T-1} f(x_{k+1}) - \sum_{k=0}^{T-1} f(x_*) &\le \frac{L}{2}||x_0 - x_*||^2 && \text{telescoping}\\
f(x_T) - f(x_*) &\le \frac{L}{2T}||x_0 - x_*||^2 && f(x_T) \le \sum_{k=0}^{T} f(x_{k+1}).
\end{aligned}
$$

□

3 Mirror Descent Convergence Result

Definition 3.1 (Bregman divergence)**.** *The Bregman divergence from x to y with respect to a 1-strongly convex function h defines a distance based on how the function differs from its linear approximation:*

$$D_h(y||x) := h(y) - h(x) - \langle \nabla h(x), y - x\rangle \tag{6}$$

Definition 3.2 (Mirror Descent Update)**.** *The mirror descent update with step length α can be described as*

$$x_{k+1} = Mirr_\alpha(x_k; \partial f(x)) \quad \textit{where} \quad Mirr_\alpha(x_k;\xi) \leftarrow \arg\min_{y\in Q}\Big\{\langle \xi, y - x_k\rangle + \frac{1}{\alpha}D_h(y||x_k)\Big\}. \tag{7}$$

the step takes the familiar additive form $\nabla h(x_{k+1}) = \nabla h(x_k) - \alpha\partial f(x_k)$.

In Lemma (2.3), we use the gradient descent update to obtain the relation between the one step regret $\langle \nabla f(x_k), x_k - x\rangle$, the ℓ_2 difference $\frac{1}{2}||x_k - x||^2 - \frac{1}{2}||x_{k+1} - x||^2$, and the gradient norm $||\nabla f(x)||_*^2$. This encourages us to investigate whether there exists a similar relationship between them when using the mirror descent update.

Lemma 3.3. *For x_{k+1} and x_k in mirror descent method, i.e, $x_{k+1} = Mirr_\alpha(x_k)$ (3.2), we have*

$$\forall x \in Q \quad \alpha\langle \partial f(x_k), x_k - x\rangle \le D_h(x||x_k) - D_h(x||x_{k+1}) + \frac{\alpha^2}{2}||\partial f(x_k)||_*^2. \tag{8}$$

Proof.

$$
\begin{aligned}
&D_h(x||x_k) - D_h(x||x_{k+1})\\
=&h(x) - h(x_k) - \langle \nabla h(x_k), x - x_k\rangle - h(x) + h(x_{k+1}) + \langle \nabla h(x_{k+1}), x - x_{k+1}\rangle\\
=&h(x_{k+1}) - h(x_k) + \langle \nabla h(x_{k+1}), x - x_{k+1}\rangle + \langle \nabla h(x_k), x_k - x\rangle\\
=&D_h(x_{k+1}||x_k) + \langle \nabla h(x_k), x_{k+1} - x_k\rangle + \langle \nabla h(x_{k+1}), x - x_{k+1}\rangle + \langle \nabla h(x_k), x_k - x\rangle\\
=&D_h(x_{k+1}||x_k) + \langle \nabla h(x_k) - \nabla h(x_{k+1}), x_{k+1} - x\rangle
\end{aligned}
$$

Since mirror decent update give us $\nabla h(x_{k+1}) = \nabla h(x_k) - \alpha\partial f(x_k)$, we have:

$$=D_h(x_{k+1}||x_k) + \langle \alpha\partial f(x_k), x_{k+1} - x\rangle,$$

Since h is 1-strongly convex, we have $D_h(x_{k+1}||x_k) \ge \frac{1}{2}||x_{k+1} - x_k||^2$

$$
\begin{aligned}
&\ge \frac{1}{2}||x_{k+1} - x_k||^2 + \langle \alpha\partial f(x_k), x_{k+1} - x\rangle\\
&= \frac{1}{2}||x_{k+1} - x_k||^2 + \langle \alpha\partial f(x_k), x_{k+1} - x_k\rangle + \langle \alpha\partial f(x_k), x_k - x\rangle
\end{aligned}
$$

Since $\frac{1}{2}||\alpha\partial f(x_k)+x_{k+1}-x_k||^2 \geq 0$, we know $\frac{1}{2}||x_{k+1}-x_k||^2+\alpha\langle\partial f(x_k), x_{k+1}-x_k\rangle \geq -\frac{\alpha^2}{2}||\partial f(x_k)||^2$

$$\geq -\frac{\alpha^2}{2}||\partial f(x_k)||^2 + \langle\alpha\partial f(x_k), x_k - x\rangle$$

Finally, by rearranging the inequality, we obtain

$$\alpha\langle\partial f(x_k), x_k - x\rangle \leq D_h(x||x_k) - D_h(x||x_{k+1}) + \frac{\alpha^2}{2}||\partial f(x_k)||_*^2.$$

□

Unlike gradient descent, which has a descent Lemma to ensure stepwise improvement, mirror descent is not guaranteed to decrease the function value monotonically, we only have Lemma (3.3) to prove the convergence.

Theorem 3.4 (Mirror Descent Regret Bound)**.** *Assume $f(x)$ is convex and $||\partial f(x)||_*$ is bounded by G. Assume $f(x)$ takes minimizer at x_*. Then mirror descent algorithm starting with x_0 and taking constant step size α in every iteration produces $x_1, \ldots, x_{T-1}$ such that for any $x_* \in Q$:*

$$f(\bar{x}) - f(x_*) \leq \frac{1}{T}\sum_{k=0}^{T-1} f(x_k) - \frac{1}{T}\sum_{k=0}^{T-1} f(x_*) \leq \frac{1}{\alpha T} D_h(x_*||x_0) + \frac{\alpha}{2T}\sum_{k=0}^{T-1}||\partial f(x_k)||_*^2, \tag{9}$$

Let Θ be any upper bound on $D_h(x_*||x_0)$ and take step length $\alpha = \frac{\sqrt{2\Theta}}{G\cdot\sqrt{T}}$, we have $f(\bar{x}) - f(x_*) \leq \frac{\sqrt{2\Theta}\cdot G}{\sqrt{T}}$ and we need $\Omega(\frac{2\Theta\cdot G^2}{\epsilon^2})$ iterations to reach the accuracy of ϵ.

Proof. By convexity, we have $f(x_*) \geq f(x_k) + \langle\partial f(x_k), x_* - x_k\rangle$, so

$$\begin{aligned}
f(x_k) - f(x_*) &\leq \langle\partial f(x_k), x_k - x_*\rangle \\
&\leq \frac{1}{\alpha}D_h(x_*||x_k) - \frac{1}{\alpha}D_h(x_*||x_{k+1}) + \frac{\alpha}{2}||\partial f(x_k)||_*^2 && \text{apply Lemma (3.3)} \\
\sum_{k=0}^{T-1} f(x_k) - \sum_{k=0}^{T-1} f(x_*) &\leq \frac{1}{\alpha}D_h(x_*||x_0) - \frac{1}{\alpha}D_h(x_*||x_T) + \frac{\alpha}{2}\sum_{k=0}^{T-1}||\partial f(x_k)||_*^2 && \text{telescoping} \\
f(\bar{x}) - f(x_*) &\leq \frac{1}{\alpha T}D_h(x_*||x_0) + \frac{\alpha}{2T}\sum_{k=0}^{T-1}||\partial f(x_k)||_*^2 && \text{Jensen Inequality} \\
&\leq \frac{\Theta}{\alpha T} + \frac{\alpha G^2}{2}
\end{aligned}$$

take $\alpha = \frac{\sqrt{2\Theta}}{G\cdot\sqrt{T}}$ will balance the residuals

$$\leq \frac{\sqrt{2\Theta}\cdot G}{\sqrt{T}}$$

□

4 Connection to Accelerate Gradient Descent

Nesterov Accelerated gradient descent [3] for L-smooth functions gives a $\Omega(\sqrt{\frac{L}{\epsilon}})$ bound on iterations. Although accelerated gradient methods have been widely applied, they are often regarded as "analytical tricks" because their convergence analyses are somewhat complicated and lack of intuitions. Thus, in paper [1], the author provides a simple, alternative, but complete version of the accelerated gradient method by constructing two sequences of updates: one sequence of gradient-descent updates and one sequence of mirror-descent updates.

By descent lemma (2.2) and Theorem (3.4), we know a large gradient norm will have a large function value decrease when doing gradient descent while have a slower convergence speed when doing mirror descent update. Suppose $f(x)$ is unconstrained and L-smooth, Zhu et al. [1] attempts to define a

cut-off value K for $||\nabla f(x)||_2$, the norm of the observed gradients, such that if $||\nabla f(x)||_2$ is always $\geq K$, we perform T gradient decent steps; otherwise we perform T mirror-descent steps.

Considering gradient-descent steps always decrease the objective while mirror descent steps may sometimes increase the objective, canceling the effect of the gradient descent, we cannot simply alternative between doing the gradient descent steps and mirror descent steps. For this reason, Zhu et al. [1] choose a linear combination of gradient decent steps and mirror descent steps, i.e., $x_{k+1} \leftarrow \tau z_k + (1-\tau)y_k$ in the k-th iteration and use the same gradient $\nabla f(x_{k+1})$ to continue the gradient and mirror steps of the next iteration. Whenever τ is carefully chosen (like the cut-off value K), the decent descent sequences provide a coupled bound on the error guarantee. Surprisingly, the recovered error bound is the same as accelerated gradient methods.

5 Linear Coupling between Gradient Descent and Mirror Descent

We now formalize how Zhu et al. [1] derive the coupled error bound for both constant step lengths and variable step lengths for Linear Coupling. Here we focus on the convergence of unconstraint settings for both constant step lengths and variable step lengths. The constraint case for variable step size can be similarly derived by switching to the constraint gradient descent update, i.e., the descent lemma in gradient descent changed to $f(Grad(x)) \leq f(x) - Prog(x)$ where $Prog(x) := -\min_{y\in Q}\{\frac{L}{2}||y-x||^2 + \langle\nabla f(x), y-x\rangle\} \geq 0$.

5.1 Constant Step Lengths

Definition 5.1. *The linearly coupled update with a fixed step length can be described as*

$$x_{k+1} \leftarrow \tau z_k + (1-\tau)y_k, \tag{10}$$

where $y_{k+1} = Grad_\eta(x_{k+1})$, $z_{k+1} = Mirr_\alpha(z_k; \nabla f(x_{k+1}))$, *and* τ *is the parameter controlling the coupling rate.*

For the linearly coupled update for x_k, we can have the similar regret bound as (2.3) and (3.3) in gradient descent update and mirror descent update.

Lemma 5.2 (Coupling)**.** *Letting* $\tau \in (0,1)$ *satisfy that* $\frac{1-\tau}{\tau} = \alpha L$, *we have that*

$$\forall x \in Q = \mathbb{R}^n, \alpha\langle\nabla f(x_{k+1}), x_{k+1} - x\rangle \leq \alpha L^2\big(f(y_k) - f(y_{k+1})\big) + (D_h(x||z_k) - D_h(x||z_{k+1})) \tag{11}$$

Proof. Since $z_{k+1} := Mirr_\alpha(z_k; \nabla f(x_{k+1}))$, we can plug into (3.3), we have

$$\begin{aligned}\alpha\langle\nabla f(x_{k+1}), z_k - x\rangle &\leq \frac{\alpha^2}{2}||\nabla f(x_{k+1})||_*^2 + D_h(x||z_k) - D_h(x||z_{k+1}) \\ &\leq \alpha^2 L(f(x_{k+1}) - f(y_{k+1})) + D_h(x||z_k) - D_h(x||z_{k+1}) \quad \text{descent lemma on } y_{k+1}\end{aligned}$$

Since $f(x_{k+1}) - f(y_{k+1})$ does not telescope,, we need

$$\begin{aligned}&\alpha\langle\nabla f(x_{k+1}), x_{k+1} - x\rangle - \alpha\langle\nabla f(x_{k+1}), z_k - x\rangle \\ =&\alpha\langle\nabla f(x_{k+1}), x_{k+1} - z_k\rangle \\ =&\frac{1-\tau}{\tau}\alpha\langle\nabla f(x_{k+1}), y_k - x_{k+1}\rangle && \tau(x_{k+1} - z_k) = (1-\tau)(y_k - x_{k+1}) \\ \leq&\frac{1-\tau}{\tau}\alpha(f(y_k) - f(x_{k+1})) && \text{convexity of } f(x)\end{aligned}$$

Summing the above two inequalities, when $\frac{1-\tau}{\tau} = \alpha L$, we have

$$\alpha\langle\nabla f(x_{k+1}), x_{k+1} - x\rangle \leq \alpha L^2\big(f(y_k) - f(y_{k+1})\big) + (D_h(x||z_k) - D_h(x||z_{k+1})) \tag{12}$$

□

Finally, we can telescope inequality in Lemma (5.2) for $k = 0, 1, \ldots, T-1$ and get similar results in Theorem (2.4) and Theorem (3.4), i.e.,

$$\alpha T(f(\bar{x}) - f(x^*)) \leq \sum_{k=0}^{T-1} \alpha \langle \nabla f(x_k), x_k - x^* \rangle \leq \alpha^2 L(f(y_0) - f(y_T)) + D_h(x^*||x_0) - D_h(x^*||x_T) \quad (13)$$

Suppose the initial point of error is at most d, i.e., $f(y_0) - f(x^*) \leq d$ and suppose $D_h(x^*||x_0) \leq \Theta$, then we have $f(\bar{x}) - f(x^*) \leq \frac{1}{T}(\alpha L d + \frac{\Theta}{\alpha})$. Choosing $\alpha = \sqrt{\frac{\Theta}{Ld}}$ to be the value that balances the two terms, we obtain $f(\bar{x}) - f(x^*) \leq \frac{2\sqrt{L\Theta d}}{T}$. We can observe that, the bound for constant step lengths is still not satisfying (We still need $\tilde{\Omega}(\frac{1}{\epsilon})$ iterations) and Zhu et al. [1] propose to restart this entire procedure a few numbers of times and obtain ϵ-approximate solution in $O(\frac{1}{\sqrt{\epsilon}})$ iterations.

5.2 Variable Step Lengths

Definition 5.3. *The linearly coupled update with a variable step length can be described as*

$$x_{k+1} \leftarrow \tau_k z_k + (1 - \tau_k) y_k, \quad (14)$$

where $y_{k+1} = Grad_\eta(x_{k+1})$, $z_{k+1} = Mirr_{\alpha_k}(z_k; \nabla f(x_{k+1}))$, and τ is the parameter controlling the coupling rate.

Lemma 5.4 (Coupling)**.** *If $\tau_k = \frac{1}{\alpha_{k+1} L}$, we have that*

$$\begin{aligned} \forall x \in Q = \mathbb{R}^n, \alpha_{k+1}\langle \nabla f(x_{k+1}), x_{k+1} - x \rangle \leq & (\alpha_{k+1}^2 L - \alpha_{k+1}) f(y_k) - (\alpha_{k+1}^2 L) f(y_{k+1}) + \\ & \alpha_{k+1} f(x_{k+1}) + (D_h(x||z_k) - D_h(x||z_{k+1})). \end{aligned} \quad (15)$$

Proof. Under the assumption of $Q = \mathbb{R}^n$, we can get a proof similar to Lemma (5.2) by replacing τ_k and α_k. Since $z_{k+1} := Mirr_{\alpha_{k+1}}(z_k; \nabla f(x_{k+1}))$, we can plug into (3.3), we have

$$\begin{aligned} \alpha_{k+1}\langle \nabla f(x_{k+1}), z_k - x \rangle &\leq \frac{\alpha_{k+1}^2}{2} ||\nabla f(x_{k+1})||_*^2 + D_h(x||z_k) - D_h(x||z_{k+1}) \\ &\leq \alpha_{k+1}^2 L(f(x_{k+1}) - f(y_{k+1})) + D_h(x||z_k) - D_h(x||z_{k+1}) \quad \text{descent lemma on } y_{k+1} \end{aligned}$$

Since $f(x_{k+1}) - f(y_{k+1})$ does not telescope,, we need

$$\begin{aligned} & \alpha_{k+1}\langle \nabla f(x_{k+1}), x_{k+1} - x \rangle - \alpha_{k+1}\langle \nabla f(x_{k+1}), z_k - x \rangle \\ =& \alpha_{k+1}\langle \nabla f(x_{k+1}), x_{k+1} - z_k \rangle \\ =& \frac{1 - \tau_k}{\tau_k} \alpha_{k+1}\langle \nabla f(x_{k+1}), y_k - x_{k+1} \rangle && \tau_k(x_{k+1} - z_k) = (1 - \tau_k)(y_k - x_{k+1}) \\ \leq & \frac{1 - \tau_k}{\tau_k} \alpha_{k+1}(f(y_k) - f(x_{k+1})) && \text{convexity of } f(x) \\ =& (\alpha_{k+1} L - 1)\alpha_{k+1}(f(y_k) - f(x_{k+1})) && \tau_k = \frac{1}{\alpha_{k+1} L} \end{aligned}$$

Summing up the above two inequalities, we obtain

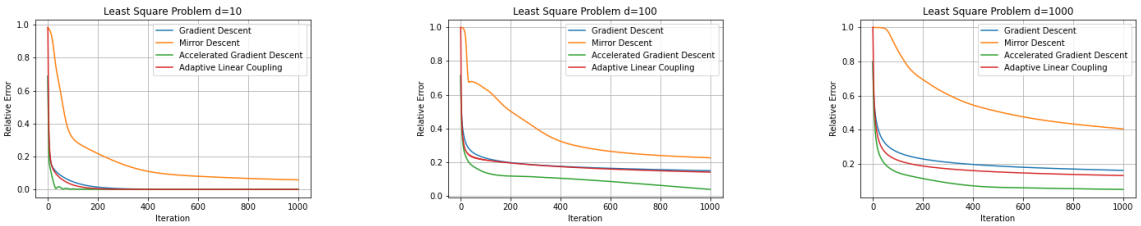
$$\begin{aligned} \alpha_{k+1}\langle \nabla f(x_{k+1}), x_{k+1} - x \rangle \leq & (\alpha_{k+1}^2 L - \alpha_{k+1}) f(y_k) - (\alpha_{k+1}^2 L) f(y_{k+1}) + \\ & \alpha_{k+1} f(x_{k+1}) + (D_h(x||z_k) - D_h(x||z_{k+1})). \end{aligned}$$

□

Theorem 5.5. *Assume $f(x)$ is convex and L-smooth, $h(x)$ is 1-strongly convex, then linear coupling update [AGM] y_T satisfying $f(y_T) - f(x^*) \leq 4\Theta L / T^2$, where Θ is any upper bound on $D_h(x^*||x_0)$.*

Proof. since $f(x)$ is convex, we have $f(x_{k+1}) - f(x) \leq \langle \nabla f(x_{k+1}), x_{k+1} - x \rangle$, we can rewrite Lemma (5.4) as

$$\begin{aligned} \alpha_{k+1}(f(x_{k+1}) - f(x)) &\leq (\alpha_{k+1}^2 L - \alpha_{k+1}) f(y_k) - (\alpha_{k+1}^2 L) f(y_{k+1}) + \alpha_{k+1} f(x_{k+1}) + (D_h(x||z_k) - D_h(x||z_{k+1})) \\ -\alpha_{k+1} f(x) &\leq (\alpha_{k+1}^2 L - \alpha_{k+1}) f(y_k) - (\alpha_{k+1}^2 L) f(y_{k+1}) + (D_h(x||z_k) - D_h(x||z_{k+1})) \\ \alpha_{k+1} f(x) &\geq (\alpha_{k+1}^2 L) f(y_{k+1}) - (\alpha_{k+1}^2 L - \alpha_{k+1}) f(y_k) + (D_h(x||z_{k+1}) - D_h(x||z_k)) \end{aligned}$$



(a) $d = 10$ (b) $d = 100$ (c) $d = 1000$

Figure 1: Solving Least Square by different optimization algorithms

In order to telescope, we need to set the sequence of α_k so that $\alpha_k L \approx \alpha_{k+1}^2 L - \alpha_{k+1}$ as well as $\tau_k = \frac{1}{\alpha_{k+1} L}$. In Zhu et al. [1], $\alpha_k = \frac{k+1}{2L}$ so that $\alpha_k^2 L = \alpha_{k+1}^2 L - \alpha_{K+1} + \frac{1}{4L}$, summing up the above inequalities for $k = 0, 1, \ldots, T-1$, we obtain

$$\sum_{k=1}^{T} \alpha_k f(x) \geq \alpha_T^2 L f(y_T) - (\alpha_0^2 L - \alpha_0) f(y_0) + \sum_{k=1}^{T-1} \frac{1}{4L} f(y_k) + D_h(x||z_T) - D_h(x||z_0)$$

assume the minimize x^* exists, we known $\sum_{k=1}^{T} \alpha_k = \frac{T(T+3)}{4L}$, $f(y_k) \geq f(x^*)$, $D_h(x^*||z_T) \geq 0$, $(\alpha_0^2 L - \alpha_0) f(y_0) < 0$ and $D_h(x||z_0) \leq \Theta$ we obtain

$$\left(\frac{T(T+3)}{4L} - \frac{T-1}{4L} \right) f(x^*) + \Theta \geq \frac{(T+1)^2}{4L^2} L f(y_T) \tag{16}$$

after simplification, we get $f(y_T) \leq f(x^*) + \frac{4\Theta L}{(T+1)^2}$ □

6 Experiments

We perform the experiments of the proposed linear coupling algorithm on the toy least square problem:

$$\min \frac{1}{2} ||\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta}||^2 \tag{17}$$

we create a random symmetric matrix $\boldsymbol{X} \in \mathbb{R}^{n \times n}$, ground truth $\boldsymbol{\beta}$ on the computer and calculate the $\boldsymbol{Y}$ through $\boldsymbol{X}\boldsymbol{\beta}$.

In our experiment, we compare Linear Coupling with gradient descent, mirror descent, and accelerated gradient method. For mirror descent, we choose the distance-generating function to be the negative entropy function, i.e., $h(x) = \sum_{i=1}^{n} x_i \log x_i$. We present the results in Figure 1. We observe mirror descent comes to be the slowest algorithm. The adaptive linear coupling algorithm is slightly faster than gradient descent, but it indeed is not as fast as the accelerate gradient descent. We conjecture this is because we down-scaled the step size for mirror descent by a factor c during implementation. We found if we use the original $(i+2)/2L$ stepsize for mirror descent, the mirror descent update will explode as i increases and the algorithm will fail to converge. Thus the down-scaled step size may lead to a degraded performance for Linear Coupling.

References

[1] Zeyuan Allen-Zhu and Lorenzo Orecchia. Linear coupling: An ultimate unification of gradient and mirror descent. *arXiv preprint arXiv:1407.1537*, 2014.

[2] Deanna Needell, Rachel Ward, and Nati Srebro. Stochastic gradient descent, weighted sampling, and the randomized kaczmarz algorithm. *Advances in neural information processing systems*, 27, 2014.

[3] Yurii Nesterov. A method for solving the convex programming problem with convergence rate o (1/k2). In *Dokl akad nauk Sssr*, volume 269, page 543, 1983.