CSE 599: Interplay between Convex Optimization and Geometry **Winter 2018**

Lecture 7: Gradient Mapping and First-Order Methods

Lecturer: Yin Tat Lee

Disclaimer: *Please tell me any mistake you noticed.*

The goal of this and next lecture is to cover, in one shot, basic first-order methods including gradient descent, mirror descent, accelerated gradient descent, coordinate descent, stochastic gradient method, accelerated coordinate descent and accelerated stochastic gradient method. More importantly, we want to make the proof as intuitive as possible. Therefore, many proofs here are less-traditional and more restrictive than what it is needed. To be clear, all proofs here are known and modified from [41, 42, 43, 44]. In particular, our proof is based on the linear coupling by [41] with the idea that one can accelerate first-order methods in general [43].

For simplicity, we consider the Euclidean space here only. We will discuss different space in later lectures.

7.1 Gradient Mapping

The key idea to avoid re-deriving accelerated gradient for coordinate descent and stochastic gradient separately is to develop a general way to accelerate an optimization method. To this end, we need to first develop an optimization method to solve problem of the form

$$\min_x \phi(x) \stackrel{\text{def}}{=} f(x) + h(x)$$

where $f(x)$ is strongly convex and smooth and $h(x)$ is convex. We assume that we access the function g and h differently via:

1. Let $\mathcal{T}_g$ be the cost of computing $\nabla f(x)$.
2. Let $\mathcal{T}_{h,\lambda}$ be the cost of minimizing $h(x) + \frac{\lambda}{2}\|x-c\|_2^2$ exactly.

The idea is to move whatever we can optimize in ϕ to h and hopefully this makes the remaining part of ϕ, f, as smooth and strongly convex as possible. However, in that case, we cannot assume the regularity of h too much. In fact, we do not even assume h to be differentiable. To handle this issue, we need to define an approximate derivative of h that we can compute.

Definition 7.1.1. We define the gradient step

$$p_{x,\gamma} = \text{argmin}_y f(x) + \nabla f(x)^\top (y-x) + \frac{\gamma}{2}\|y-x\|^2 + h(y)$$

and the gradient mapping

$$g_{x,\gamma} = \gamma(x - p_{x,\gamma}).$$

If f is β-smooth, we simply use p_x to denote $p_{x,\beta}$ and g_x to denote $g_{x,\beta}$.

Note that if $h = 0$, then $p_{x,\gamma} = x - \frac{1}{\gamma}\nabla f(x)$ and $g_{x,\gamma} = \nabla f(x)$. In general, if $\phi \in \mathcal{C}^2$, then we have that

$$p_{x,\gamma} = x - \frac{1}{\gamma}\nabla\phi(x) + O(\frac{1}{\gamma^2}).$$

Therefore, we have that $g_{x,\gamma} = \nabla\phi(x) + O(\frac{1}{\gamma})$. Hence, the gradient mapping is an approximation of the gradient of ϕ that is computable in time $\mathcal{T}_g + \mathcal{T}_{h,\gamma}$.

The key lemma we use here is that ϕ satisfies a lower bound defining using $g_{x,\gamma}$. Ideally, we would love to get a lower bound as follows:

$$\phi(z) \geq \phi(x) + g_{x,\gamma}^\top (z - x) + \frac{\alpha}{2}\|z - x\|_2^2 .$$

But it is WRONG. If that was true for all z, then we would have $g_{x,\gamma} = \nabla\phi(x)$. However, if $\phi \in \mathcal{C}^2$ is α strongly convex, then we have

$$\begin{aligned}\phi(z) &\geq \phi(x) + \nabla\phi(x)^\top (z - x) + \frac{\alpha}{2}\|z - x\|_2^2 \\ &\geq \phi(x - \frac{1}{\gamma}\nabla\phi(x)) + \frac{1}{2\gamma}\|\nabla\phi(x)\|^2 + \nabla\phi(x)^\top (z - x) + \frac{\alpha}{2}\|z - x\|_2^2 . \end{aligned} \tag{7.1}$$

It turns out that this is true and is exactly what we need for proving gradient descent, mirror descent and accelerated gradient descent.

Theorem 7.1.2. *Given $\phi = f + h$. Suppose that f is α strongly convex and β smooth. Then, for any $\gamma \geq \beta$ and any z, we have that*

$$\phi(z) \geq \phi(p_{x,\gamma}) + g_{x,\gamma}^\top (z - x) + \frac{1}{2\gamma}\|g_{x,\gamma}\|_2^2 + \frac{\alpha}{2}\|z - x\|_2^2 .$$

Proof. To prove this, naturally, we want to follow the calculation in (7.1) and this corresponds to the following:

$$\begin{aligned}\phi(z) - \frac{\alpha}{2}\|z - x\|_2^2 &\geq f(x) + \nabla f(x)^\top (z - x) + h(z) \\ &= f(x) + \nabla f(x)^\top (p_{x,\gamma} - x) + \nabla f(x)^\top (z - p_{x,\gamma}) + h(z) \\ &\geq f(p_{x,\gamma}) - \frac{\gamma}{2}\|p_{x,\gamma} - x\|^2 + \nabla f(x)^\top (z - p_{x,\gamma}) + h(z) \\ &= \phi(p_{x,\gamma}) - \frac{1}{2\gamma}\|g_{x,\gamma}\|^2 + \nabla f(x)^\top (z - p_{x,\gamma}) + h(z) - h(p_{x,\gamma})\end{aligned}$$

where we used f is $\gamma \geq \beta$ smooth. Unfortunately, this is what we needed. Comparing this to what we need, it suffices to prove that

$$-\frac{1}{2\gamma}\|g_{x,\gamma}\|^2 + \nabla f(x)^\top (z - p_{x,\gamma}) + h(z) - h(p_{x,\gamma}) \geq g_{x,\gamma}^\top (z - x) + \frac{1}{2\gamma}\|g_{x,\gamma}\|_2^2 .$$

Let $\overline{f}(y) = f(x) + \nabla f(x)^\top (y - x) + \frac{\gamma}{2}\|y - x\|_2^2$, then this is same as

$$\nabla\overline{f}(p_{x,\gamma})^\top (z - p_{x,\gamma}) + h(z) - h(p_{x,z}) \geq 0 \tag{7.2}$$

Let $p_t = p_{x,\gamma} + t(z - p_{x,\gamma})$. Using that p_0 is the minimizer of $\overline{f} + h$, we have that

$$\overline{f}(p_0) + h(p_0) \leq \overline{f}(p_t) + h(p_t) \leq \overline{f}(p_0) + \nabla\overline{f}(p_0)^\top (p_t - p_0) + \frac{\gamma}{2}\|p_t - p_0\|^2 + h(p_t).$$

Hence, we have that

$$\begin{aligned}0 &\leq \nabla\overline{f}(p_0)^\top (p_t - p_0) + \frac{\gamma}{2}\|p_t - p_0\|^2 + h(p_t) - h(p_0) \\ &= t \cdot \left(\nabla\overline{f}(p_0)^\top (z - p_0) + h(z) - h(p_0)\right) + \frac{\gamma t^2}{2}\|z - p_0\|^2 .\end{aligned}$$

Taking $t \to 0^+$, we have (7.2). □

The next lemma shows that $\|g_{x,\gamma}\|_2 \le 2G$ if ϕ is G-Lipschitz.

Lemma 7.1.3. *If ϕ is G-Lipschitz, then $\|g_{x,\gamma}\|_2 \le 2G$ for all x and $\gamma \ge \beta$.*

Proof. By the definition of gradient mapping (namely, $p_{x,\gamma}$ is the minimizer of a function), we have that

$$f(x) + \nabla f(x)^\top (p_{x,\gamma} - x) + \frac{\gamma}{2}\|p_{x,\gamma} - x\|^2 + h(p_{x,\gamma}) \le f(x) + h(x).$$

Using $h(p_{x,\gamma}) \ge h(x) + \nabla h(x)^\top (p_{x,\gamma} - x)$, we have that

$$\begin{aligned} 0 &\ge \nabla\phi(x)^\top (p_{x,\gamma} - x) + \frac{\gamma}{2}\|p_{x,\gamma} - x\|^2 \\ &\ge -G\|p_{x,\gamma} - x\|_2 + \frac{\gamma}{2}\|p_{x,\gamma} - x\|^2. \end{aligned}$$

Hence, we have that $\|p_{x,\gamma} - x\|_2 \le \frac{2}{\gamma}G$ and hence $\|p_{x,\gamma}\|_2 \le 2G$. □

7.2 Gradient Descent

Putting $z = x$ and $\gamma = \beta$ in Theorem 7.1.2, we have that

$$\phi(p_x) \le \phi(x) - \frac{1}{2\beta}\|g_x\|_2^2. \tag{7.1}$$

This shows that each step of the gradient step decreases the function value by $\frac{1}{2\beta}\|g_x\|_2^2$. Therefore, if the gradient is large, then we decrease the function value by a lot. On the other hand, it is easy to show that if the gradient is small and domain is bounded, we are close to the optimal. Combining these two facts, we have the following theorem:

Theorem 7.2.1 (Gradient Descent Convergence)**.** *Suppose that f is β smooth. Consider the algorithm $x_{k+1} \leftarrow p_{x_k}$, we have that*

$$\phi(x_T) - \phi(x^*) \le \frac{2\beta R^2}{T+3} \quad \textit{with} \quad R = \max_{\phi(x)\le\phi(x_1)} \|x - x^*\|_2$$

for any minimizer x^* of ϕ.

Remark. Note that $\phi(x_0) - \phi(x^*) \le \frac{\beta R^2}{2}$. Therefore, gradient descent decreases the error by roughly a $\frac{1}{T}$ factor after T steps.

Proof. Recall from (7.1) that

$$\phi(x_{k+1}) \le \phi(x_k) - \frac{1}{2\beta}\|g_{x_k}\|_2^2.$$

On the other hand, Theorem 7.1.2 (with $z = x^*$ and $\gamma = \beta$)

$$\phi(x_{k+1}) - \phi(x^*) \le g_{x_k}^\top (x_k - x^*) \le \|g_{x_k}\|_2 \cdot \|x_k - x^*\|_2 \le R \cdot \|\nabla f(x_k)\|_2$$

where we used that $\phi(x_k)$ is decreasing. Let $\varepsilon_k = \phi(x_k) - \phi(x^*)$. Then, we have

$$\varepsilon_k - \varepsilon_{k+1} \ge \frac{1}{2\beta}\|g_{x_k}\|_2^2 \ge \frac{\varepsilon_{k+1}^2}{2\beta R^2}.$$

Note that $\varepsilon_1 = \phi(x_1) - \phi(x^*) \le \phi(x^*) + \frac{\beta}{2}\|x_1 - x^*\|^2 \le \frac{\beta}{2}R^2$. Using this initial condition, one can solve the recurrence and get

$$\varepsilon_k \le \frac{2\beta R^2}{k+3}.$$

□

7.3 Mirror Descent

7.3.1 Intuition

To give the intuition for the formula of mirror descent, we consider the simple case $h = 0$. It is like the opposite of gradient descent. Instead of optimizing over the upper bound, it optimizes over the lower bound of ϕ. Recall that a function ϕ is α strongly convex if

$$f(y) \geq f(x) + \nabla f(x)^\top (y - x) + \frac{\alpha}{2} \|y - x\|_2^2$$

Suppose that we have queried points $x_1, \cdots, x_k$. Then, a convex combination of the lower bound is given by

$$f(y) \geq \frac{1}{k} \sum_{i=1}^{k} \left(f(x_i) + \nabla f(x_i)^\top (y - x_i) + \frac{\alpha}{2} \|y - x_i\|_2^2 \right) \tag{7.1}$$

Mirror descent chooses

$$x_{k+1} = \text{argmin}_x \sum_{i=1}^{k} \left(\nabla f(x_i)^\top (y - x_i) + \frac{\alpha}{2} \|y - x_i\|_2^2 \right). \tag{7.2}$$

Since x_k is the minimizer of the quadratic function $\sum_{i=1}^{k-1} \left(\nabla f(x_i)^\top (y - x_i) + \frac{\alpha}{2} \|y - x_i\|_2^2 \right)$, we have that

$$\sum_{i=1}^{k-1} \left(\nabla f(x_i)^\top (y - x_i) + \frac{\alpha}{2} \|y - x_i\|_2^2 \right) = C_k + \frac{\alpha(k-1)}{2} \|y - x_k\|^2 .$$

Putting it into (7.2) gives

$$\begin{aligned} x_{k+1} &= \text{argmin}_x \frac{\alpha(k-1)}{2} \|y - x_k\|_2^2 + \nabla f(x_k)^\top (y - x_k) + \frac{\alpha}{2} \|y - x_k\|_2^2 \\ &= \text{argmin}_x \nabla f(x_k)^\top (y - x_k) + \frac{\alpha k}{2} \|y - x_k\|_2^2 . \end{aligned}$$

This shows that another way to write down mirror descent in Euclidean space is

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

This is basically gradient descent with a decreasing step size. This is special for Euclidean space. For other norms, the mirror descent is different from gradient descent.

7.3.2 Analysis

For the general case with $h \neq 0$, we simply define the mirror descent as

$$x_{k+1} = x_k - \frac{1}{\alpha k} g_{x_k}. \tag{7.3}$$

To analyze the mirror descent, we use Theorem 7.1.2 and the convexity of ϕ to get that

$$\phi(\frac{1}{k} \sum_{i=1}^{k} y_{x_i}) - \phi(x^*) \leq \frac{1}{k} \sum_{i=1}^{k} (\phi(y_{x_i}) - \phi(x^*)) \leq \frac{1}{k} \sum_{i=1}^{k} g_{x_i}^\top (x_i - x^*). \tag{7.4}$$

Therefore, it suffices to upper bound $g_{x_i}^\top (x_i - x^*)$. The following lemma shows that if $g_{x_i}^\top (x_i - x^*)$ is large, then either the gradient is large or the distance to optimum moves a lot. It turns out this holds for any vector g, not necessarily an approximate gradient.

Lemma 7.3.1 (Mirror Descent Lemma)**.** *Let $p = x - \eta g$. Then, we have that*

$$g^\top (x-u) \le \frac{\eta}{2}\|g\|_2^2 + \frac{1}{2\eta}\left(\|x-u\|_2^2 - \|p-u\|_2^2\right)$$

for any u.

Proof. Note that

$$\begin{aligned}\|p-u\|^2 &= \|x-u-\eta g\|^2 \\ &= \|x-u\|^2 - 2\eta \cdot g^\top (x-u) + \eta^2 \|g\|^2 .\end{aligned}$$

□

Using this, we can prove the convergence of mirror descent.

Theorem 7.3.2 (Mirror Descent Convergence)**.** *Suppose that f is α-strongly convex and ϕ is G-Lipschitz. Consider the algorithm $x_{k+1} \leftarrow x_k - \frac{1}{\alpha k} g_{x_k}$, we have*

$$\phi\left(\frac{1}{T}\sum_{k=1}^{T} p_{x_k}\right) - \min_x \phi(x) \le O(\log T)\cdot \frac{G^2}{\alpha T}.$$

Remark. Note that $\phi(x_0) - \min_x \phi(x) \le \frac{G^2}{2\alpha}$. Therefore, mirror descent decreases the error by roughly a $\frac{1}{T}$ factor after T steps.

Proof. Theorem 7.1.2 shows that

$$\phi(p_{x_k}) - \phi(x^*) \le g_{x_k}^\top (x_i - x^*) - \frac{\alpha}{2}\|x_k - x^*\|_2^2 .$$

Using Lemma 7.3.1, we have that

$$\begin{aligned}\phi(p_{x_k}) - \phi(x^*) &\le \frac{1}{2\alpha k}\|g_{x_k}\|_2^2 + \frac{\alpha k}{2}\left(\|x_k - x^*\|_2^2 - \|x_{k+1} - x^*\|_2^2\right) - \frac{\alpha}{2}\|x_k - x^*\|^2 \\ &= \frac{1}{2\alpha k}\|g_{x_k}\|_2^2 + \frac{\alpha(k-1)}{2}\|x_k - x^*\|_2^2 - \frac{\alpha k}{2}\|x_{k+1} - x^*\|_2^2 .\end{aligned}$$

Summing both sides and using the Lipschitz constant (Lemma 7.1.3), we have

$$\begin{aligned}\phi\left(\frac{1}{T}\sum_{k=1}^{T} p_{x_k}\right) - \min_x \phi(x) &\le \frac{1}{T}\sum_{k=1}^{T}(f(p_{x_k}) - f(x^*)) \\ &\le \frac{1}{T}\sum_{k=1}^{T}\frac{4G^2}{2\alpha k} + \frac{\alpha(1-1)}{2}\|x_1 - x^*\|_2^2 - \frac{\alpha T}{2}\|x_{T+1} - x^*\|_2^2 \\ &= O(\log T)\cdot\frac{G^2}{\alpha T}.\end{aligned}$$

□

Exercise 7.3.3. Improve the bound in Theorem 7.3.2 to $O(\frac{G^2}{\alpha T})$ by optimizing the step sizes in (7.3).

7.4 Accelerated Gradient Descent

Recall (7.1) shows that if the gradient is large, gradient descent makes a large progress. On the other hand, if the gradient is small, (7.4) shows that mirror descent makes a large progress. Therefore, it is natural to combine two approaches.

- Starting from $x_1 = y_1 = z_1$,
- For $k = 1, \cdots, T-1$
  - Define $x_{k+1} = \tau z_k + (1-\tau) y_k$.
  - Perform a gradient step $y_{k+1} = x_{k+1} - \frac{1}{\beta} g_{x_{k+1}}$.
  - Perform a mirror step $z_{k+1} = z_k - \eta g_{x_{k+1}}$.

where τ and η are some parameters to be chosen. Note that if $\tau = 1$, the algorithm is simply mirror descent and if $\tau = 0$, the algorithm is gradient descent.

Theorem 7.4.1. *Setting $\frac{1-\tau}{\tau} = \eta\beta$ and $\eta = \frac{1}{\sqrt{\alpha\beta}}$, we have that*

$$\phi(\frac{1}{T}\sum_{k=1}^{T} p_{x_{k+1}}) - \phi(x^*) \le \frac{2}{T}\sqrt{\frac{\beta}{\alpha}}\left(\phi(x_1) - \phi(x^*)\right).$$

In particular, if we restart the algorithm every $4\sqrt{\frac{\beta}{\alpha}}$ iterations, we can find x such that

$$\phi(x) - \phi(x^*) \le 2(1 - \sqrt{\frac{\alpha}{\beta}})^{\Omega(T)}\left(\phi(x_1) - \phi(x^*)\right)$$

in T steps. Furthermore, each step takes $\mathcal{T}_f + \mathcal{T}_{h,\beta}$

Proof. Lemma 7.3.1 and (7.1) showed that

$$g_{x_{k+1}}^\top (z_k - x^*) \le \frac{\eta}{2}\left\|g_{x_{k+1}}\right\|_2^2 + \frac{1}{2\eta}\left(\|z_k - x^*\|_2^2 - \|z_{k+1} - x^*\|_2^2\right) \tag{7.1}$$

This shows that if the mirror descent has large error $g_{x_{k+1}}^\top (z_k - x^*)$, then the gradient descent makes a large progress ($\frac{\eta}{2}\left\|g_{x_{k+1}}\right\|_2^2$).

To make the left-hand side usable, note that $x_{k+1} = z_k + \frac{1-\tau}{\tau}\cdot(y_k - x_{k+1})$ and hence

$$\begin{aligned} g_{x_{k+1}}^\top (x_{k+1} - x^*) &= g_{x_{k+1}}^\top (z_k - x^*) + \frac{1-\tau}{\tau}\cdot g_{x_{k+1}}^\top (y_k - x_{k+1}) \\ &\le g_{x_{k+1}}^\top (z_k - x^*) + \frac{1-\tau}{\tau}(\phi(y_k) - \phi(y_{k+1}) - \frac{1}{2\beta}\left\|g_{x_{k+1}}\right\|_2^2) \\ &\le \frac{\eta}{2}\left\|g_{x_{k+1}}\right\|_2^2 + \frac{1-\tau}{\tau}(\phi(y_k) - \phi(y_{k+1}) - \frac{1}{2\beta}\left\|g_{x_{k+1}}\right\|_2^2) + \frac{1}{2\eta}\left(\|z_k - x^*\|_2^2 - \|z_{k+1} - x^*\|_2^2\right). \end{aligned}$$

where we used Theorem 7.1.2 in the middle and (7.1) at the end.

Now, we set $\frac{1-\tau}{\tau} = \eta\beta$ and get

$$g_{x_{k+1}}^\top (x_{k+1} - x^*) \le \eta\beta(\phi(y_k) - \phi(y_{k+1})) + \frac{1}{2\eta}\left(\|z_k - x^*\|_2^2 - \|z_{k+1} - x^*\|_2^2\right).$$

Taking a sum on both side and let $\overline{x} = \frac{1}{T}\sum_{k=1}^{T} y_{x_{k+1}}$, we have that

$$\begin{aligned}\eta T \cdot (\phi(\overline{x}) - \phi(x^*)) &\le \eta T \sum_{k=1}^{T} (\phi(y_{x_{k+1}}) - \phi(x^*)) \\ &\le \eta T \sum_{k=1}^{T} g_{x_{k+1}}^{\top}(x_{k+1} - x^*) \\ &\le \eta^2 \beta(\phi(y_1) - \phi(y_{T+1})) + \frac{1}{2}\left\|z_1 - x^*\right\|_2^2.\end{aligned}$$

Hence, we have that

$$\begin{aligned}\phi(\overline{x}) - \phi(x^*) &\le \frac{\eta\beta}{T}(\phi(x_1) - \phi(x^*)) + \frac{1}{2\eta T}\left\|x_1 - x^*\right\|_2^2 \\ &\le \left(\frac{\eta\beta}{T} + \frac{1}{2\eta T\alpha}\right)(\phi(x_1) - \phi(x^*)).\end{aligned}$$

The conclusion follows from our setting of η. □

7.5 (Accelerated) Coordinate Descent

If some of the coordinates are more important than other, it makes sense to update the important coordinates more often than others.

Lemma 7.5.1. *Given a convex function ℓ. Suppose that $\frac{\partial^2}{\partial x_i^2}\ell(x) \le \beta_i$ for all x and let $B = \sum \beta_i$. If we sample coordinate i with probability $p_i = \frac{\beta_i}{B}$, then,*

$$\mathbb{E}_i \ell(x - \frac{1}{\beta_i}\frac{\partial}{\partial x_i}\ell(x)e_i) \le \ell(x) - \frac{1}{2B}\left\|\nabla\ell(x)\right\|_2^2.$$

Proof. Note that the function $\zeta(t) = \ell(x + te_i)$ is β_i smooth. Hence, we have that

$$\ell(x - \frac{1}{\beta_i}\frac{\partial}{\partial x_i}\ell(x)e_i) \le \ell(x) - \frac{1}{2\beta_i}\frac{\partial}{\partial x_i}\ell(x)^2.$$

Since we sample coordinate i with probability $p_i = \frac{\beta_i}{B}$, we have that

$$\begin{aligned}\mathbb{E}\ell(x - \frac{1}{\beta_i}\frac{\partial}{\partial x_i}\ell(x)e_i) &= \ell(x) - \sum_i \frac{\beta_i}{B}\frac{1}{2\beta_i}\frac{\partial}{\partial x_i}\ell(x)^2 \\ &= \ell(x) - \frac{1}{2B}\sum_i \frac{\partial}{\partial x_i}\ell(x)^2 \\ &= \ell(x) - \frac{1}{2B}\left\|\nabla\ell(x)\right\|_2^2.\end{aligned}$$

□

By the same proof as Theorem 7.2.1 but replacing (7.1) by Lemma 7.5.1, we have the following

Theorem 7.5.2 (Coordinate Descent Convergence)**.** *Given a convex function ℓ. Suppose that $\frac{\partial^2}{\partial x_i^2}\ell(x) \le \beta_i$ for all x and let $B = \sum \beta_i$. Consider the algorithm $x^{(k+1)} \leftarrow x^{(k)} - \frac{1}{\beta_i}\frac{\partial}{\partial x_i}\ell(x^{(k)})e_i$, we have that*

$$\mathbb{E}\ell(x^{(T)}) - \ell(x^*) \le \frac{2BR^2}{T+3} \quad \textit{with} \quad R = \max_{\ell(x)\le\ell(x^{(1)})}\left\|x - x^*\right\|_2$$

for any minimizer x^* of ℓ. If ℓ is α strongly convex, then we also have

$$\mathbb{E}\ell(x^{(T)}) - \ell(x^*) \leq 2(1 - \frac{B}{\alpha})^{\Omega(T)}(\ell(x^{(0)}) - \ell(x^*)).$$

Proof. The first part follows by the same proof in Theorem 7.2.1. The second part follows from

$$R^2 \leq \frac{1}{\alpha}(\ell(x^{(0)}) - \ell(x^*)).$$

□

Now, the really fun part is here. Consider the function

$$\phi(x) = f(x) + h(x) \quad \text{with} \quad f(x) = \frac{\alpha}{2}\|x\|_2^2 \text{ and } h(x) = \ell(x) - \frac{\alpha}{2}\|x\|_2^2.$$

Since f is $\alpha + \frac{B}{n}$ smooth (YES!, I know this is also α smooth) and α strongly convex and since h is convex, we apply Theorem 7.4.1 and get an algorithm that takes $O^*(\sqrt{\frac{B}{n\alpha}})$ steps. Note that each step involves $\mathcal{T}_f + \mathcal{T}_{h,\alpha+\frac{B}{n}}$. Obviously, $\mathcal{T}_f = 0$. Next, note that $\mathcal{T}_{h,\alpha+\frac{B}{n}}$ involves solving a problem of the form

$$\begin{aligned} y_x &= \text{argmin}_y(\frac{\alpha}{2} + \frac{B}{2n})\|y - x\|^2 + (\ell(y) - \frac{\alpha}{2}\|x\|^2) \\ &= \text{argmin}_y \ell(y) - \alpha y^\top x + \frac{B}{2n}\|y - x\|^2. \end{aligned}$$

Now, we can apply Theorem 7.5.2 to solve this problem. It takes

$$O^*(\frac{B + (B/n)\cdot n}{B/n}) = O^*(n) \text{ coordinate steps.}$$

Therefore, in total it takes

$$O^*(\sqrt{\frac{B}{n\alpha}}) \cdot O^*(n) = O^*(\sqrt{\frac{Bn}{\alpha}}) \text{ coordinate steps.}$$

Hence, we have the following theorem

Theorem 7.5.3 (Accelerated Coordinate Descent Convergence)**.** *Given an α strongly-convex function ℓ. Suppose that $\frac{\partial^2}{\partial x_i^2}\ell(x) \leq \beta_i$ for all x and let $B = \sum \beta_i$. We can minimize ℓ in $O^*(\sqrt{\frac{Bn}{\alpha}})$ coordinate steps.*

References

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

[42] Roy Frostig, Rong Ge, Sham Kakade, and Aaron Sidford. Un-regularizing: approximate proximal point and faster stochastic algorithms for empirical risk minimization. In *International Conference on Machine Learning*, pages 2540–2548, 2015.

[43] Hongzhou Lin, Julien Mairal, and Zaid Harchaoui. A universal catalyst for first-order optimization. In *Advances in Neural Information Processing Systems*, pages 3384–3392, 2015.

[44] Yu Nesterov. Introductory lectures on convex programming volume i: Basic course. *Lecture notes*, 1998.