Structured Bandit Methods for Optimization

Anthony Nguyen, Raghu Bollapragada, Matt Menickelly and Stefan M. Wild

June 15, 2020

1 Week 1: June 15

1.1 Research Aims

We are targeting unconstrained minimization of functions that have a composite structure involving a zeroth-order oracle $\boldsymbol{F} : \mathbb{R}^n \mapsto \mathbb{R}^p$ and an algebraically available $h : \mathbb{R}^p \mapsto \mathbb{R}$. That is, we want to minimize

$$f(\boldsymbol{x}) = h(\boldsymbol{F}(\boldsymbol{x}); \boldsymbol{x}), \tag{1}$$

where the above notations allows for functions that include algebraic terms that do not involve $\boldsymbol{F}(\boldsymbol{x})$; an example of the latter is $f(\boldsymbol{x}) = \|\boldsymbol{F}(\boldsymbol{x})\|_2^2 + \|\boldsymbol{x}\|_2^2$. We refer to the general functions F_i $(i = 1, \ldots, p)$ as *component functions*. Objective functions of the form eq. (1) are found throughout supervised learning, design optimization, and simulation-based optimization.

We will focus on methods that are based on derivative-free approximations of directional derivatives. Specifically, given a point $\boldsymbol{x}_k \in \mathbb{R}^n$, steplength $\Delta_k > 0$, and vector $\boldsymbol{v} \in \mathbb{R}^n$ (not necessarily of unit length), we would like to use zero-order quantities such as[1] $\boldsymbol{F}(\boldsymbol{x}_k)$ and $\boldsymbol{F}(\boldsymbol{x}_k + \Delta_k \boldsymbol{v}_k)$ in order to minimize eq. (1).

Such information could be used to define an approximation of the directional Jacobian $\boldsymbol{v}^T \nabla_{\boldsymbol{x}} \boldsymbol{F}(\boldsymbol{x}_k)$, which can be expressed componentwise by

$$\delta_{F_i}\left(F_i; \boldsymbol{x}_k; \boldsymbol{v}; \Delta_k\right) := \frac{F_i(\boldsymbol{x}_k + \Delta_k \boldsymbol{v}) - F_i(\boldsymbol{x}_k)}{\Delta_k}, \qquad i = 1, \ldots, p. \tag{2}$$

Although one could have $\boldsymbol{v}$ swipe through the coordinate directions and thus define the entire Jacobian $\nabla_{\boldsymbol{x}} \boldsymbol{F}(\boldsymbol{x}_k)$, here we are interested in having $\boldsymbol{v}$ be a single direction. Randomized strategies could manifest in selection of the direction $\boldsymbol{v}$ and possibly the steplength Δ_k.

Our central question is to understand what can be said about such methods and for what types of functions $(f, h, \boldsymbol{F})$ such methods could present an advantage over similar methods that do not leverage of the composite mapping (i.e., methods that use only the zeroth-order quantities $f(\boldsymbol{x}_k)$ and $f(\boldsymbol{x}_k + \Delta_k \boldsymbol{v}_k)$).

Notes:

- We will initially focus on deterministic functions $\boldsymbol{F}$ so that the only source of randomness is from the algorithm.
- We will assume that $\boldsymbol{F}$ and h are as smooth as needed; however, whenever possible, we would like to not require explicit knowledge of $\boldsymbol{F}$ (e.g., its Lipschitz constants).
- In some cases, it could be beneficial to consider the case when the function h is convex in its primary arguments.
- A cursory background on how the Argonne folks might have thought about derivative-free optimization, bandit/randomized methods, etc. is [**?**]. It would be good to discuss this.

  All references in [**?**] are already in a bib file in this repo.

[1] Central difference and other variants could obviously also fall under such an umbrella.

1.1.1 First Steps

As an initial example, consider the sum of squares

$$f(\boldsymbol{x}) = \sum_{i=1}^{p} F_i(\boldsymbol{x})^2, \tag{3}$$

which has

$$\nabla_{\boldsymbol{x}} f(\boldsymbol{x}) = 2 \sum_{i=1}^{p} F_i(\boldsymbol{x}) \nabla_{\boldsymbol{x}} F_i(\boldsymbol{x}) \tag{4}$$

and

$$\nabla^2_{\boldsymbol{x},\boldsymbol{x}} f(\boldsymbol{x}) = 2 \sum_{i=1}^{p} \left(\nabla_{\boldsymbol{x}} F_i(\boldsymbol{x}) \nabla_{\boldsymbol{x}} F_i(\boldsymbol{x})^T + F_i(\boldsymbol{x}) \nabla^2_{\boldsymbol{x}} F_i(\boldsymbol{x}) \right). \tag{5}$$

It is evident, for example, that a directional derivative estimate could use the zeroth-order information in eq. (4) – namely, the values $\boldsymbol{F}(\boldsymbol{x})$ – by observing that:

$$\boldsymbol{v}^T \nabla_{\boldsymbol{x}} f(\boldsymbol{x}) = 2 \boldsymbol{v}^T \nabla_{\boldsymbol{x}} \boldsymbol{F}(\boldsymbol{x}) \boldsymbol{F}(\boldsymbol{x}). \tag{6}$$

So clearly a directional derivative estimate such as $\delta_{\boldsymbol{F}}(\boldsymbol{F}; \boldsymbol{x}; \boldsymbol{v}; \Delta)$ could be beneficial to use.

But in what ways does such an approximation, which leverages some form of the composite mapping, benefit the overall approximation and its use in an optimization algorithm?

What is the approximation error that results? Does this change current 2-point (and other) bandit methods in some way?

Is there an opportunity to reuse quantities such as $\delta_{\boldsymbol{F}}(\boldsymbol{F}; \boldsymbol{x}; \boldsymbol{v}; \Delta)$ and other readily available zeroth-order information to also start to access higher-order derivatives? When/how does this result in a different step?

1.2 Logistics

All times listed are in US Central time.

1.2.1 Workday

No formal clock in/clock out. You'll get a daily doodle check in. Schedule?

1.2.2 Meetings

We will schedule at least 3 research meetings a week. These times will help us stay connected during a virtual internship. We should address issues as they come up outside of these meeting times; Slack and email are the preferred ways to sync up.

- Mondays 1:00–1:30pm, project-wide: with Matt, Stefan, and Raghu
  Bluejeans: https://bluejeans.com/281217739
- Tuesdays 2:00-2:30 PM with Matt
  Bluejeans: https://bluejeans.com/581344516
- Thursdays 4–4:30pm, with Stefan
  Bluejeans: https://bluejeans.com/114200013
- TBD, with Raghu

There are also a number of other regular events outside of these research-focused meetings, including:

- Mondays 2-3pm, Optimization and Uncertainty coffee: informal updates
  Zoom: https://us02web.zoom.us/j/83111019500?pwd=Qzg5aDlFVUZJZnpLVHdBeERGZ0M0Zz09
- Wednesdays 10:15–11:30am, LANS seminar
  (link sent out via mailing list)
- Thursdays 10-11am, Argonne/CELS student seminar series?
  (link sent out via mailing list)
- Thursdays 1pm, LANS Student mixers
  (links change and are sent out via mailing list)

And more ways to stay connected:

- Wednesdays 2–2:30pm, Computing at Argonne coffee presentations
  (link sent out via mailing list)
- LANS Thirsty Thursday/Happy Hour
  Zoom https://us02web.zoom.us/j/89951606596?pwd=V0lNelRPMVJLREgyTk5iQmd6N0hqZz09
- Fridays 11–11:30am, MCS weekly coffee
  Zoom: https://us02web.zoom.us/j/89384520983?pwd=MTlSM1Y1TUhLY1Brd1BRZjJPS3VDUT09
- Weekdays 2:30–3pm, LANS coffee/tea (best to announce on channel social/student Slack channels)
  Bluejeans: https://argonne.bluejeans.com/121887495

1.2.3 Repository

We will do all of our work in a git repository hosted at https://xgitlab.cels.anl.gov. You can log in there with your Argonne user name once that is established, but it may also be just as easy to set up a new MCS account associated with your university address. To do so, go to the above link and follow the link that says "clicking here."

The actual repo location is
https://xgitlab.cels.anl.gov/wild/structuredbandits20
The default branch is `team`, we'll all push there. I suggest getting in the habit of committing (say once a day), this will also help during our various meetings.

The plan is to make a running log/journal, in this document, of things you're working on, reading through, or have questions on.

1.3 Anthony's Mathematical Journal

1.3.1 Paper 1: Random Gradient-Free Minimization of Convex Functions

The idea is to perform optimization of functions based on zeroth order information.

What the paper proves:

- New complexity bounds for methods of convex optimization on zeroth order information
  - Search directions of schemes based on distributed random Gaussian vectors. Such methods usually needs at most n times more iterations than standard gradient methods, where n is the dimension of the space of variables.
  - This is true for both smooth and non-smooth problems.
- For smooth classes, paper present an accelerated scheme with expected rate of convergence $\mathcal{O}\left(\frac{n^2}{k^2}\right)$, where k is the iteration counter.

- For stochastic optimization, authors propose zeroth order scheme and justify its expected rate of convergence $\mathcal{O}\left(\frac{n}{\sqrt{k}}\right)$
- Authors give some bounds for the rate of convergence of the random gradient-free methods to stionary points of non-convex functions, for both smooth / nonsmooth cases.

Brief Summary of results on Gradient Free Minimization

Notation:

$$D_u f(x)u \equiv g_0(x) := f'(x,u)u \tag{7}$$

where $D_u f(x)$ is the directional derivative of f in the direction of $u \neq 0$.

For finite dimensional vector space E, denote E^* its algebraic dual space. By Riesz theorem, $s \in E^*$ can be identified by $s(x) = \langle s, x\rangle$. Endow the spaces E, E^* with Euclidean norms

$$\|x\| := \langle Bx, x\rangle^{1/2}, x \in E, \quad \|s\|_* := \langle s, B^{-1}s\rangle^{1/2}, s \in E^*$$

where $B = B^* \succ 0$ (Positive Definite, Self-adjoint) is a linear operator from E to E^*. For all $u \in E$, define

$$uu^* : E^* \to E \quad \text{a linear operator, which acts as follows:}$$
$$uu^*(s) = u \cdot \langle s, u\rangle, s \in E^*.$$

Results and definitions:

Given $f \in C^1(E)$ and whose gradient is lipschitz continuous. Then

$$\|f(y) - f(x) - \nabla f(x)(y-x)\| \leq \frac{L_{\nabla f}}{2}\|x-y\|^2, \quad x, y \in E \tag{8}$$

When $f \in C^2(E)$ whose Hessian is Lipschitz continuous. Then the corresponding third order bound is as follows:

$$|f(y) - f(x) - \langle \nabla f(x), y-x\rangle - \frac{1}{2}\langle \nabla^2 f(x)(y-x), y-x\rangle| \leq \frac{L_{\nabla^2 f}}{2}\|x-y\|^3, \quad x, y \in E \tag{9}$$

Definition 1. *We say $f \in C^{1,1}(E)$ is strongly convex with convexity parameter $\tau(f) \geq 0$, if for any $x, y \in E$ we have*

$$f(y) \geq f(x) + \nabla f(x)(y-x) + \frac{\tau(f)}{2}\|x-y\|^2. \tag{10}$$

Definition 2. *Recall that a **subgradient** of a convex function at y is a vector g such that*

$$f(x) \geq f(y) + g^\top(x-y) \quad \text{for all } x \in E \tag{11}$$

Therefore, the set of all subgradient g of f at y is denoted $\partial f(y)$. Mathematically,

$$\partial f(y) := \{g : f(x) \geq f(y) + g^\top(x-y) \text{ for all } x \in E\}$$

Definition 3. *Let $\epsilon \geq 0$. For convex function f, we denote by $\partial f_\epsilon(y)$ its ϵ-**subdifferential** at $y \in E$:*

$$f(x) \geq f(y) - \epsilon + \langle g, x-y\rangle, \quad g \in \partial f_\epsilon(y), x \in E.$$

Observation: When $\epsilon = 0$, $\partial f_\epsilon(y) = \partial f(y)$.

Gaussian Smoothing:

Let $f : E \to \mathbb{R}$ and assume $D_u f$ exists for each $x \in E$. Then using Gaussian smoothing, we have

$$f_\mu(x) = \mathbb{E}_u[f(x + \mu u)] = \frac{1}{\kappa} \int_E f(x + \mu u) e^{-\frac{1}{2}\|u\|^2} du, \tag{12}$$

where

$$\kappa = \int_E e^{-\frac{1}{2}\|u\|^2} du = \frac{(2\pi)^{n/2}}{\sqrt{\det(B)}} \tag{13}$$

is the normalizing constant.

Lemma 1. *f convex $\Rightarrow$ f_μ is convex.*

Lemma 2. *If $f \in C^0(E)$, then $f_\mu \in C^0(E)$ with $L_{f_\mu} \leq L_f$.*

Lemma 3. *If $f \in C^1(E)$ with gradient Lipschitz continuous, then $f_\mu \in C^1(E)$ whose gradient is Lipschitz continuous with $L_{\nabla f_\mu} \leq L_{\nabla f}$.*

Theorem 4. *Let $f \in C^0(E)$ be Lipschitz continuous. Then*

$$|f_\mu(x) - f(x)| \leq \mu L_f \sqrt{n}, x \in E \tag{14}$$

If $f \in C^1(E)$ with gradient Lipschitz continuous, then

$$|f_\mu(x) - f(x)| \leq \frac{\mu^2 n}{2} L_{\nabla f}, x \in E \tag{15}$$

Finally, if $f \in C^2(E)$ and Hessian is Lipschitz continuous, then

$$|f_\mu(x) - f(x) - \frac{\mu^2}{2}\langle \nabla^2 f(x), B^{-1}\rangle| \leq \frac{\mu^3}{3} L_{\nabla^2 f}(n+3)^{3/2}, x \in E \tag{16}$$

Proof. We prove (14). This follows from the triangle inequality and $\frac{1}{\kappa}\int_E \|u\| e^{-\frac{1}{2}\|u\|^2} du \leq \sqrt{n}$ by Lemma 1 in [**?**].

Next, we deduce (15). Since $\int_E u e^{-\frac{1}{2}\|u\|^2} du = 0$, we have

$$f_\mu(x) - f(x) = \frac{1}{\kappa} \int_E [f(x + \mu u) - f(x) - \mu\langle \nabla f(x), u\rangle] e^{-\frac{1}{2}\|u\|^2} du \tag{17}$$

Using (8) and Lemma 1 in '[**?**],

$$|f_\mu(x) - f(x)| \leq \frac{\mu^2 L_{\nabla f}}{2\kappa} \int_E \|u\|^2 e^{-\frac{1}{2}\|u\|^2} du = \frac{\mu^2 L_{\nabla f}}{2} n \tag{18}$$

□

Lemma 5. *Let $f \in C^0(E)$ and Lipschitz continuous with $\mu > 0$. Then $f_\mu \in C^1(E)$ with*

$$L_{\nabla f_\mu} = \frac{\sqrt{n}}{\mu} L_f \tag{19}$$

Theorem 6. *Let f be convex and Lipschitz continuous. Then for $x \in E, \mu \geq 0$, we have*

$$\nabla f_\mu(x) \in \partial_\epsilon f(x), \quad \epsilon = \mu L_f \sqrt{n} \tag{20}$$

Lemma 7. *If $f \in C^1(E)$ with gradient Lipschitz continuous, then*

$$\|\nabla f_\mu(x) - \nabla f(x)\|_* \leq \frac{\mu}{2} L_{\nabla f}(n+3)^{3/2} \tag{21}$$

For $f \in C^2(E)$ with Hessian Lipschitz continuous. Then

$$\|\nabla f_\mu(x) - \nabla f(x)\|_* \leq \frac{\mu^2}{6} L_{\nabla^2 f}(n+4)^2 \tag{22}$$

Lemma 8. *Let $f \in C^1(E)$ with gradient Lipschitz continuous. Then for any $x \in E$, we have*

$$\|\nabla f(x)\|_*^2 \leq 2\|\nabla f_\mu(x)\|_*^2 + \frac{\mu^2}{2} L_{\nabla f}^2 (n+6)^3 \tag{23}$$

Random Gradient-Free Oracles:

Let random vector $u \in E$ have Gaussian distribution with correlation operator B^{-1}. We define the following random gradient free oracles:

- (Forward differencing) Generate random $u \in E$ and return $g_\mu(x) = \frac{f(x+\mu x) - f(x)}{\mu} Bu$
- (Central differencing) Generate random $u \in E$ and return $\hat{g_\mu}(x) = \frac{f(x+\mu u) - f(x-\mu u)}{2\mu} Bu$
- (Generate random $u \in E$ and return $g_0(x) = D_u f(x) Bu$

Fact: $\mathbb{E}_u[g_0(x)] = \nabla f(x) \in \partial f(x)$.

Theorem 9. *If f is differentiable at x, then*

$$\mathbb{E}_u[\|g_0(x)\|_*^2] \leq (n+4)\|\nabla f(x)\|_*^2 \tag{24}$$

Let f be convex. Denote $D(x) = \sup_{x_1, x_2 \in \partial f(x)} \|x_1 - x_2\|$. Then for any $x \in E$, we have

$$\mathbb{E}_u[\|g_0(x)\|_*^2] \leq (n+4)(\|\nabla f(x)\|_*^2 + nD^2(x)) \tag{25}$$

Theorem 10. *Let function f be convex.*

- *If $f \in C^0(E)$ and Lipschitz continuous, then*

$$\mathbb{E}_u[\|g_\mu(x)\|_*^2] \leq L_f^2 (n+4)^2 \tag{26}$$

- *If $f \in C^1(E)$ with gradient Lipschitz continuous, then*

$$\mathbb{E}_u[\|g_\mu(x)\|_*^2] \leq \frac{\mu^2}{2} L_{\nabla f}^2 (n+6)^3 + 2(n+4)\|\nabla f(x)\|_*^2, \tag{27}$$

$$\mathbb{E}_u[\|\hat{g_\mu}(x)\|_*^2] \leq \frac{\mu^2}{8} L_{\nabla f}^2 (n+6)^3 + 2(n+4)\|\nabla f(x)\|_*^2 \tag{28}$$

- *If $f \in C^2(E)$ with Hessian Lipschitz continuous, then*

$$\mathbb{E}_u[\|\hat{g}_\mu(x)\|_*^2] \leq \frac{\mu^4}{18} L_{\nabla^2 f}^2 (n+8)^4 + 2(n+4)\|\nabla f(x)\|_*^2. \tag{29}$$

Lemma 11. *Let $f \in C^1(E)$ with gradient Lipschitz continuous. Then for any $x \in E$, we have*

$$\mathbb{E}_u[\|g_\mu(x)\|_*^2] \leq 4(n+4)\|\nabla f_\mu(x)\|_*^2 + 3\mu^2 L_{\nabla f}^2 (n+4)^3 \tag{30}$$

In later analysis, the author shows

Theorem 12.

$$\nabla f_\mu(x) = \frac{1}{\mu}\mathbb{E}_u[f(x+\mu u)u] = \frac{1}{\mu}\mathbb{E}_u[(f(x+\mu u) - f(x))u] \tag{31}$$

Optimization with access to stochastic function:

Suppose our objective function has the following form:

$$f(x) = \mathbb{E}_\xi[F(x,\xi)] := \int_\Xi F(x,\xi)dP(\xi), \quad x \in Q, \tag{32}$$

where ξ is a random vector with probability distribution $P(\xi), \xi \in \Xi$. We assume $f \in C^0(E)$ and Lipschitz continuous. We define random stochastic gradient-free oracles:

- Generate random $u \in E, \xi \in \Xi$. Return $s_\mu(x) = \frac{F(x+\mu u,\xi)-F(x,\xi)}{\mu} Bu$.
- Generate random $u \in E, \xi \in \Xi$. Return $\hat{s}_\mu(x) = \frac{F(x+\mu u,\xi)-F(x-\mu u,\xi)}{\mu} Bu$.
- Generate random $u \in E, \xi \in \Xi$. Return $s_0(x) = D_x F(x,\xi)[u]Bu$.

1.3.2 Paper 2: An Optimal Algorithm for Bandit and Zero-Order Convex Optimization with Two-Point Feedback

Paper considers the closely related problems of bandit convex optimization with two-point feedback, and zero-order stochastic convex optimization with two function evaluations per round.

- Simple algorithm and analysis which is optimal for convex Lipschitz functions
- This algorithm improves on the algorithm give in paper "Optimal rates for zero-order optimization: the power of two function evaluations." by J. Duchi, M. Jordan, M. Wainwright, and A. Wibisono, which only provides an optimal result for smooth functions.
- Algorithm based on a small but surprisingly powerful modfication of the gradient estimator

Description of the problem of bandit convex optimization: The problem can be defined as a repeated game between a learner and an adversary as follows: At each round t, the adversary picks a convex function f_t on $\mathbb{R}^d$, which is not revealed to the learner. The learner then chooses a point w_t from some known and closed convex set $\mathcal{W} \subseteq \mathbb{R}^d$, and suffers a loss $f_t(w_t)$. As feedback, the learner may choose two points $w_t', w_t'' \in \mathcal{W}$ and receive $f_t(w_t'), f_t(w_t'')$. The learner's goal is to minimize average regret, defined as

$$\frac{1}{T}\sum_{t=1}^{T} f_t(w_t) - \min_{w\in\mathcal{W}} \frac{1}{T}\sum_{t=1}^{T} f_t(w). \tag{33}$$

Paper focuses on obtaining bounds on the expected average regret (expectation of equation (33)).

A closely related and easier setting is zero-order stochastic convex optimization. In this setting, our goal is to approximately solve

$$F(\mathbf{w}) = \min_{\mathbf{w}\in\mathcal{W}} \mathbb{E}_\xi[f(\mathbf{w};\xi)], \tag{34}$$

given limited access to $\{f(\cdot;\xi_t)\}_{t=1}^T$ whre ξ_t are i.i.d. random variables. Specifically, we assume that each $f(\cdot,\xi_t)$ is not directly observed, but rather can be queried at two points.

Brief Summery on Bandit and Zeroth order convex optimization

Fact: Using central differencing over forward differencing due to variance depends on dimension.

1.3.3 Paper 3: Optimization of Convex Functions with Random Pursuit*

- Unconstrained randomized optimization of convex objective functions
- Analyze the Random Pursuit algorithm
  - iteratively computes an approximate solution to the optimization problem by repeated optimization over randomly chosen one-dimensional subspace.
  - This randomized method only uses zeroth order information about the objective function and doesn't need any problem-specific parametrization.
- Prove convergence and give convergence rates for smooth objectives assuming that the one-dimensional optimization can be solved exactly or approximately by an oracle.

Brief Summery on Random Pursuit Paper

Metric of convergence for a zeroth order method: $\mathbb{E}[f(x_N) - f(x^*)]$.

2 Week 2: June 22

2.0.1 Approaches:

Approach # 1: Use Theorem 1 in [**?**]; we want $f = h \circ F$ to have gradient Lipschitz continuous.

A sufficient condition is for F, h and $\nabla F, \nabla h$ to be lipschitz continuous. Observe that $\nabla h(x) = 2x$ is Lipschitz. But in the unconstrained setting $h(x) = \|x\|^2$ need not be Lipschitz.

Approach # 2: Replace F by F_μ and we can estimate the latter by zeroth order oracle calls where $F_\mu := (F_{1,\mu}, \ldots, F_{p,\mu})$; we write $F_{i,\mu} := (F_i)_\mu$ for notational convenience.

By Lemma 3 in [**?**] we can make such replacement assuming $F \in C^1$ with Lipschitz continuous gradient.

$$\begin{aligned}\nabla f(x) &= \nabla h(F(x))\nabla F(x) \\ &= 2F(x)\nabla F(x) \\ &\approx 2F(x)\nabla F_\mu(x) \\ &\approx 2F(x)\left[\frac{F(x+\mu u) - F(x)}{\mu} Bu\right] \end{aligned} \tag{35}$$

We can replace (35) by a central differencing approximator:

$$\nabla f(x) \approx 2F(x)\left[\frac{F(x+\mu u) - F(x-\mu u)}{\mu} Bu\right] \tag{36}$$

Consider approximation $\tilde{f}(x) := h \circ F_\mu(x)$. By the chain rule, $\nabla(h \circ F_\mu)(x) = 2F_\mu(x)\nabla F_\mu(x)$. Want to measure the error

$$\begin{aligned}\|\nabla f(x) - \nabla \tilde{f}(x)\| &\le \|2F(x)\nabla F(x) - 2F(x)\nabla F_\mu(x)\| + \|2F(x)\nabla F_\mu(x) - 2F_\mu(x)\nabla F_\mu(x)\| \\ &\le 2\|F(x)\| \underbrace{\|\nabla F(x) - \nabla F_\mu(x)\|}_{\text{Use Lemma 3}} + \|\nabla F_\mu(x)\| \underbrace{\|F(x) - F_\mu(x)\|}_{\text{Use Theorem 1}} \end{aligned} \tag{37}$$

Obstacle: To bound (37), it is sufficient that F is Lipschitz continuous and F is bounded.

2.0.2 June 24:

Going through details of the gradient-free minimization and thinking of the results in connection to our composite setting.

2.0.3 June 25:

By equation (21), (26) in [**?**], we readily have

$$\mathbb{E}_u[g_\mu(x)] = \nabla f_\mu(x) \tag{38}$$
$$\mathbb{E}_u[\hat{g}_\mu(x)] = \nabla f_\mu(x) \tag{39}$$
$$\mathbb{E}_u[g_0(x)] = \nabla f(x) \tag{40}$$

Something I am thinking about: To exploit some of the results in the Gradient-Free paper, it's enough to assume the objective function $f = h \circ F$ is convex. At this point, h is convex, but the components of F are not necessarily convex. If f is convex and Lipschitz continuous, using Theorem 2 in Gradient-Free paper,

$$\nabla f_\mu(x) \in \partial_\epsilon f(x) \tag{41}$$

Consider when F has one component: To see this

$$\begin{aligned} f_\mu(x) &= \mathbb{E}_u[f(x+\mu u)] \\ &= \frac{1}{\kappa}\int_E h \circ F(x+\mu u) e^{-\frac{1}{2}\|u\|^2} du \\ &= \sum_{i=1}^m \frac{1}{\kappa}\int_E \tilde{h} \circ F_i(x+\mu u) e^{-\frac{1}{2}\|u\|^2} du \\ &= \sum_{i=1}^p \mathbb{E}_u[\tilde{h} \circ F_i(x+\mu u)] \\ &= \sum_{i=1}^p f_{i,\mu}(x) \end{aligned} \tag{42}$$

where $f_{i,\mu}(x) := \mathbb{E}_u[\tilde{h} \circ F_i(x+\mu u)]$ and $\tilde{h}(x) := x^2$.

Here are some papers I can examine mentioned in [**?**].

Nesterov mentions the random optimization approach in [**?**], applied to the minimization problem

$$\min_{x \in \mathbb{R}^n} f(x) \tag{43}$$

by sampling a point y randomly around the current point x (using a Gaussian distirbution) and move to y if $f(y) < f(x)$. The performance of this technique for nonconvex smooth functions was estimated in [**?**] and critized in [**?**].

Task: Look at the Rosenbrock function as a test function to run simulations.

2.0.4 June 26:

An assumption I can impose to ensure that $f = h \circ F$ has gradient continuous lipschitz is the range of F is bounded. Short computation to be added...

3 Week 3: June 29

3.0.1 June 29:

Here's a multi-dimensional Rosenbrock function, a non-convex test function:

$$f(\mathbf{x}) = \sum_{i=1}^{n-1} \left[100(x_{i+1} - x_i^2)^2 + (1 - x_i)^2\right]. \tag{44}$$

Some computation on the Rosenbrock function:

$$D_{e_1} f(\mathbf{x}) = -400x_1(x_2 - x_1^2) - 2(1 - x_1) \tag{45}$$
$$D_{e_n} f(\mathbf{x}) = 200(x_n - x_{n-1}^2) \tag{46}$$
$$D_{e_j} f(\mathbf{x}) = -400(x_{j+1} - x_j^2)x_j - 2(1 - x_j) + 200(x_j - x_{j-1}^2) \quad 1 < j < n. \tag{47}$$

In our setting, $f(\mathbf{x}) = h \circ F(\mathbf{x})$ where $F : \mathbb{R}^n \to \mathbb{R}^{2(n-1)}$ and

$$F(\mathbf{x}) = (F_1(\mathbf{x}), \ldots, F_{n-1}(\mathbf{x}), G_1(\mathbf{x}), \ldots, G_{n-1}(\mathbf{x})) \tag{48}$$

where

$$F_i(\mathbf{x}) = 10(x_{i+1} - x_i^2) \quad 1 \le i \le n - 1 \tag{49}$$
$$G_i(\mathbf{x}) = 1 - x_i \quad 1 \le i \le n - 1 \tag{50}$$

By the chain rule, $\nabla f(\mathbf{x}) = \nabla h(F(\mathbf{x}))\nabla F(\mathbf{x}) = 2F(\mathbf{x})\nabla F(\mathbf{x})$.

$$D_{e_1} F(\mathbf{x}) = -20x_1 e_1 - e_n \tag{51}$$
$$D_{e_n} F(\mathbf{x}) = 10e_{n-1} \tag{52}$$
$$D_{e_i} F(\mathbf{x}) = -20x_i e_i + 10e_{i-1} - e_{(n-1)+i} \quad 1 < i < n. \tag{53}$$

For nonconvex problems, [**?**] considers the problem

$$\min_{x \in E} f(x), \tag{54}$$

where f is nonconvex. We discuss the method (66) in [**?**]. There are two cases:

1. $f \in C^1(E)$ and ∇f is lipschitz.

To get

$$\frac{1}{N+1} \sum_{k=0}^{N} \mathbb{E}_{\mathcal{U}_k}(\|f(x_k)\|_*^2) \le \epsilon^2 \tag{55}$$

where $\mathcal{U}_k = (u_0, \ldots, u_k)$ a random vector whose components are i.i.d., we need to pick $u \le \mathcal{O}\left(\frac{\epsilon}{n^{3/2} L_{\nabla f}}\right)$ and the upper bound for the expected number of steps is $\mathcal{O}\left(\frac{n}{\epsilon^2}\right)$.

2. $f \in C^0(E)$ and f is lipschitz. Observe by lemma 2 in [**?**], $f_\mu \in C^1(E)$ and the gradient is lipschitz continous.

Define $S_N := \sum_{k=0}^{N} h_k$, we have

$$\frac{1}{S_N}\sum_{k=0}^{N} h_k \mathbb{E}_{\mathcal{U}_k}(\|\nabla f_\mu(x_k)\|_*^2) \leq \frac{1}{S_N}\left[(f_\mu(x_0) - f^*) + C(\mu)\sum_{k=0}^{N} h_k^2\right] \tag{56}$$

where

$$C(\mu) := \frac{1}{\mu}\sqrt{n}(n+4)^2 L_f^3 \tag{57}$$

To pick a smooth approximation f_μ of f within an ϵ error threshold, choose $\mu \leq \bar{\mu} = \frac{\epsilon}{\sqrt{n}L_f}$. Choosing $h_k \equiv h, k \geq 0$, the upperbound of (56) becomes

$$\frac{f_{\bar{\mu}}(x_0) - f^*}{(N+1)h} + \frac{h}{\epsilon}n(n+4)^2 L_f^4 \leq \frac{L_f R}{(N+1)h} + \frac{h}{\epsilon}N(n+4)^2 L_f^4 \equiv \rho(h) \tag{58}$$

where R is chosen such that $\|x_0 - x^*\| \leq R$ and $N \geq n$. Minimizing h in the upperbound, we have

$$h^* = \left[\frac{\epsilon R}{n(n+4)^2 L_f^3 (N+1)}\right]^{1/2} \quad , \quad \rho(h^*) = 2\left[\frac{n(n+4)^2 L_f^5 R}{\epsilon(N+1)}\right]^{1/2}.$$

To get the expected squared norm of the gradient of $f_{\bar{\mu}}$ of the order δ, we need

$$\mathcal{O}\left(\frac{n(n+4)^2 L_f^5 R}{\epsilon\delta^2}\right) \tag{59}$$

iterations of the method (66) in [**?**].

Remark: To use the method, we need $f = h \circ F$ to be Lipschitz or have gradient Lipschitz. But in both cases, we see that f is not Lipschitz on the unconstrained setting nor does its gradient is Lipschitz continuous.

By inspection, ∇F is Lipschitz using the computation above. However

$$\begin{aligned}
\|\nabla f(x) - \nabla f(y)\| &= 2\|F(x)\nabla F(x) - F(y)\nabla F(y)\| && (60)\\
&\leq 2\left[\|F(x)\|\|\nabla F(x) - \nabla F(y)\| + \|\nabla F(y)\|\|F(x) - F(y)\|\right] && (61)\\
&\leq 2\left[L_{\nabla F}\|F(x)\|\|x - y\| + \|\nabla F(y)\|\|F(x) - F(y)\|\right] && (62)
\end{aligned}$$

In general, F need not be Lipschitz nor do F and ∇F are required to be bounded.

We need to make some assumptions or try the methods on just the inner function (or something...).

3.0.2 June 30:

Task:

Reading up on some optimization literature to cover some gap in knowledge and for some inspiration / ideas. Such books include Nesterov's Lectures on Convex Optimization, Boyd's Convex optimization, and Nocedal's Numerical Optimization.

Trying to see whether I can adapt our setting to Method $\widehat{\mathcal{RS}_\mu}$ in [**?**].

Currently, I'm inclined to assume that F is bounded and Lipschitz, but this seems restrictive. This is sufficient so that $f = h \circ F$ is Lipschitz continuous (a required condition for the method).

To get f to have gradient lipschitz, we it's sufficient that F is bounded and lipschitz continuous (see (62)).

3.0.3 July 1:

Approach 1: We can use Method $\widehat{\mathcal{RS}_\mu}$ in [**?**] when F is bounded and Lipschitz.

Furthermore, if we want to use $\tilde{f}(x) := h \circ F_\mu(x)$ as an approximation, we assume that F has gradient continuous Lipschitz and using (37) yields the error bound.

Approach 2: We can use Method $\widehat{\mathcal{RS}_\mu}$ in [**?**] but need f to have gradient Lipschitz continuity. The sufficient condition is $F, \nabla F$ are Lipschitz, and F is bounded.

To see why consider the following computation:

$$\begin{aligned}
\|\nabla f(x) - \nabla f(y)\| &= 2\|F(x)\nabla F(x) - F(y)\nabla F(y)\| && (63)\\
&= 2[\|F(x)\nabla F(x) - F(x)\nabla F(y)\| + \|F(x)\nabla F(y) - F(y)\nabla F(y)\|] && (64)\\
&\leq 2[\|F(x)\|\|\nabla F(x) - \nabla F(y)\| + \|\nabla F(y)\|\|F(x) - F(y)\|] && (65)\\
&\leq 2[\|F\|_{L^\infty} L_{\nabla F}\|x - y\| + L_F^2\|x - y\|] && (66)\\
&= \underbrace{[2\|F\|_{L^\infty} L_{\nabla F} + 2L_F^2]}_{L_f}\|x - y\| && (67)
\end{aligned}$$

Remark: Numerical testing? Rosenbrock isn't lipschitz nor is its gradient lipschitz.

3.0.4 July 2:

Test Problem: Consider the least squares problem where $f(x) = \|Ax - b\|^2$ which can be written as $f(x) = h \circ F(x)$ where $F(x) = Ax - b$.

$$\begin{aligned}
\nabla f(x) &= \nabla h(F(x))\nabla F(x)\\
&= 2F(x)^\top \nabla F(x)\\
&= 2F(x)^\top A\\
&= 2(Ax - b)^\top A\\
&= 2(x^\top A^\top A - b^\top A) && (68)\\
\nabla f(x)^\top &= 2(A^\top Ax - A^\top b) && (69)
\end{aligned}$$

Generally, note that f is not Lipschitz. Luckily, ∇f is lipschitz provided that $\|A^\top A\|_2$ is bounded.

$$\begin{aligned}
\|\nabla f(x) - \nabla f(y)\|_2 &= \|2(A^\top Ax - A^\top b) - 2(A^\top Ay - A^\top b)\|\\
&= 2\|A^\top Ax - A^\top Ay\|
\end{aligned}$$

$$\leq 2\|A^\top A\|_2\|x-y\| \tag{70}$$

We assume $A_{ij} \sim \mathcal{N}(0,1)$, then $\|A^\top A\|_2$ is bounded using $\|\cdot\|_2 \leq \|\cdot\|_F$.

Sufficient conditions for Lipschitz continuity	$h = \|x\|^2$	F
f Lipschitz	n/a	F bounded and Lipschitz
∇f Lipschitz	n/a	F bounded, $F, \nabla F$ lipschitz

4 Week 4: July 6

4.0.1 July 6:

Consider two approaches:

Approach # 1:

$$\frac{f(x+\delta e_j) - f(x)}{\delta} \tag{71}$$

applied the the scheme

$$x_{k+1} = x_k - h_k \frac{f(x_k+\mu_k u) - f(x_k)}{\mu_k} u \tag{72}$$

For test problem $f(x) = \|Ax-b\|^2$, we have

$$\frac{f(x+\delta e_j) - f(x)}{\delta} = \frac{1}{\delta}\left[\sum_{i=1}^{p}(a_i^\top x + \delta a_{ij} - b_i)^2 - (a_i^\top x - b_i)^2\right] \tag{73}$$

$$= \frac{1}{\delta}\left[\sum_{i=1}^{p} 2\delta a_{ij}(a_i^\top x - b_i) + \delta^2 a_{ij}^2\right] \tag{74}$$

$$= \sum_{i=1}^{p} 2a_{ij}(a_i^\top x - b_i) + \delta \sum_{i=1}^{p} a_{ij}^2 \tag{75}$$

Approach # 2:

$$\frac{F(x+\delta e_j) - F(x)}{\delta} \tag{76}$$

For $f(x) = \|Ax-b\|^2$, we get

$$\frac{F(x+\delta e_j) - F(x)}{\delta} = A_i \tag{77}$$

$$x_{k+1} = x_k - h_k(2F(x_k))\left\{\frac{F(x_k+\mu_k u) - F(x_k)}{\mu_k}\right\} u \tag{78}$$

$$\nabla f(x) \approx \mathbb{E}_u \left[2[F_1(x), \ldots, F_p(x)] \begin{bmatrix} \frac{F_1(x_k+\mu_k u)-F_1(x_k)}{\mu_k} \\ \vdots \\ \frac{F_p(x_k+\mu_k u)-F_p(x_k)}{\mu_k} \end{bmatrix} u \right] \tag{79}$$

Potential issue:

$$\nabla^2 f(x) = 2 \left(\nabla F(x)^\top \nabla F(x) + \sum_{i=1}^{p} \underbrace{F_i(x)}_{\text{problematic to bound}} \nabla^2 F_i(x) \right) \tag{80}$$

Need to bound $\nabla^2 f(x)$ to get gradient lipschitz of $f(x)$, but the term F_i is problematic to bound.

From the Rosenbrock test function (see June 29), we have

$$G_i(x) = 1 - x_i \tag{81}$$
$$\nabla G_i(x) = -e_i \tag{82}$$
$$\nabla^2 G_i(x) = 0 \tag{83}$$
$$\nabla F_i(x) = -2x_i e_i + 10 e_{i+1} \tag{84}$$
$$\nabla^2 F_i(x) = -2 e_i e_i^\top \tag{85}$$

Since $\|\nabla^2 F_i(x)\| = 2, \|\nabla^2 G_i(x)\| = 0$ for all $x \in \mathbb{R}^n$, G_i, F_i both have gradient lipschitz.

Using the gradient and hessian information,

$$\begin{aligned} \nabla^2 f(x) &= 2 \sum_{i=1}^{n-1} \left(\nabla F_i(x) \nabla F_i(x)^\top + F_i \nabla^2 F_i(x) + \nabla G_i(x) \nabla G_i(x)^\top \right) \\ &= 2 \sum_{i=1}^{n-1} \left((-2x_i e_i + 10 e_{i+1})(-2x_i e_i + 10 e_{i+1})^\top - 20(x_{i+1} - x_i^2) e_i e_i^\top + e_i e_i^\top \right) \\ &= 2 \sum_{i=1}^{n-1} (24 x_i^2 - 20 x_{i+1} + 1) e_i e_i^\top - 20 x_i (e_i e_{i+1}^\top + e_{i+1} e_i^\top) + 100 e_{i+1} e_{i+1}^\top \end{aligned} \tag{86}$$

Simplifying the right hand side of (79),

$$\begin{aligned} \nabla f(x) &\approx 2 F(x) \nabla F_\mu(x) \\ &= 2 \sum_{i=1}^{p} F_i(x) \nabla F_{i,\mu}(x) \end{aligned} \tag{87}$$

Remark: The components of F are convex and Lipschitz, we have convergence given the structure h.

4.0.2 July 7:

Simulations adapted from [**?**] for two test functions $f(x) = \|Ax - b\|^2$ (least squares) and $f(x)$ is the Rosenbrock function. We present three simulation settings for these two test functions:

1. $f_\mu = (h \circ F)_\mu$
2. $2F(x)F_\mu(x)$
3. $2F_\mu(x)F_\mu(x)$

See code for the simulated results. We present the following graphs generated from the python code..... (to be added tomorrow with tuning parameters)

4.0.3 July 8:

Recall $f(x) = h \circ F(x)$ where $h(x) = \|x\|^2$

Scheme 1:

$$x_{k+1} = x_k - h_k \frac{f(x_k + \mu_k u) - f(x_k)}{\mu_k} u \tag{88}$$

Scheme 2:

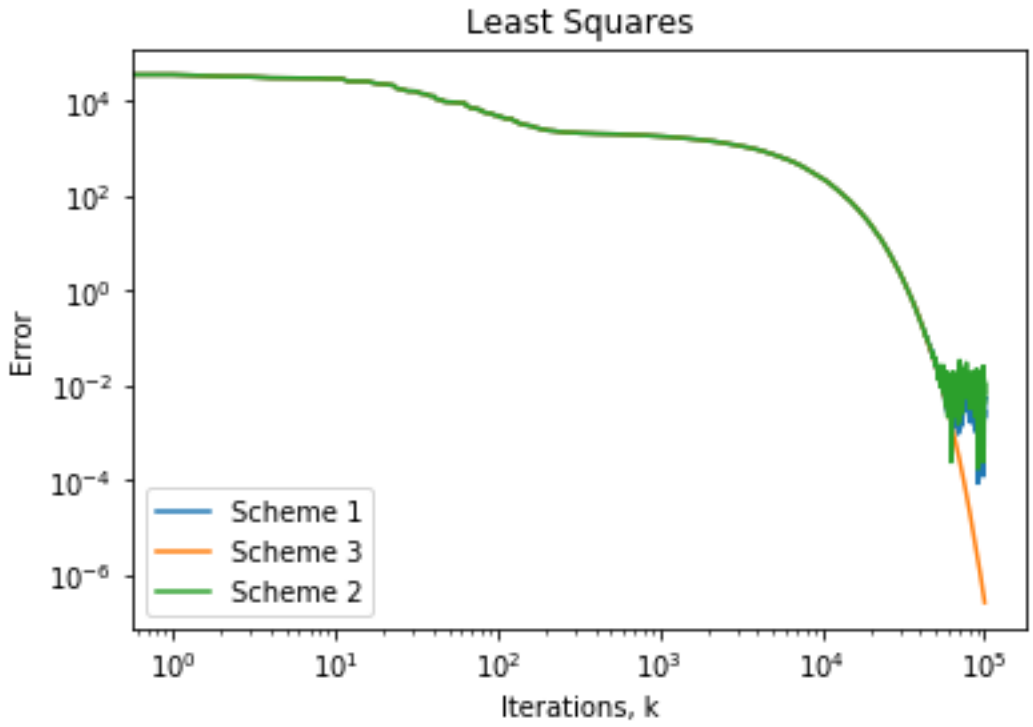
$$x_{k+1} = x_k - h_k 2F(x_k + \mu_k u) \frac{F(x_k + \mu_k u) - F(x_k)}{\mu_k} u \tag{89}$$

Scheme 3:

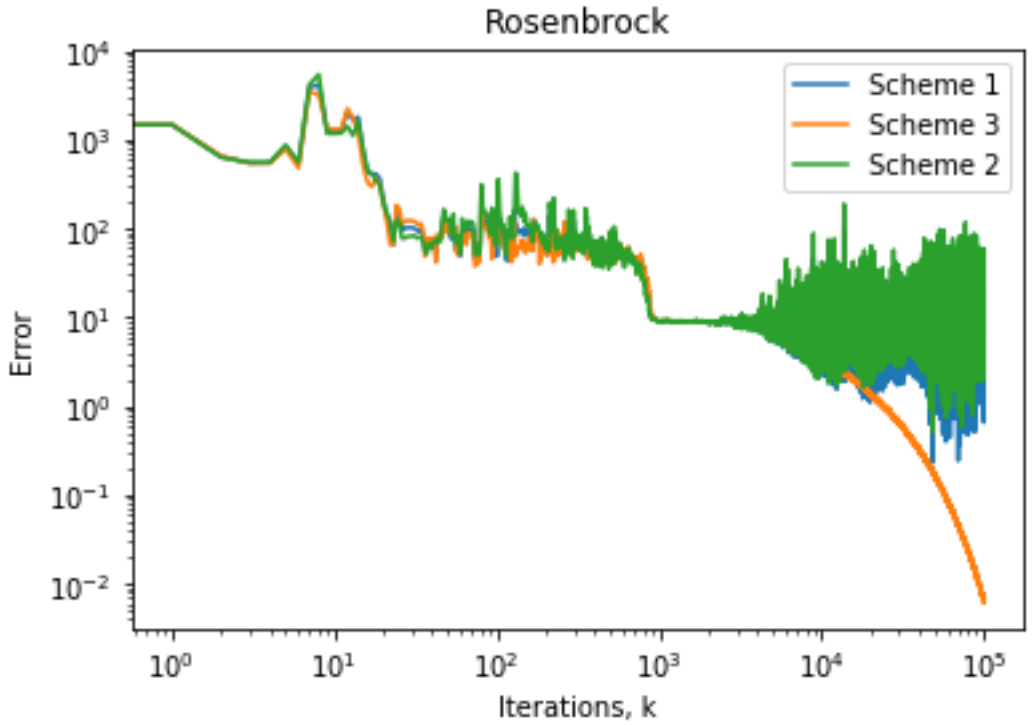
$$x_{k+1} = x_k - h_k 2F(x_k) \frac{F(x_k + \mu_k u) - F(x_k)}{\mu_k} u \tag{90}$$

Remark: For the simulation pertaining to the least squares, we utilize the $\mathcal{RG}_\mu$ in [**?**] and parameters are chosen in Theorem 8 in [**?**]. For the Rosenbrock test cases, we refer the reader to the Nonconvex Problems section in [**?**] and set $L_{\nabla f} = \|\nabla f(x_0)\|_2$ where x_0 is the initial point chosen. We set $n = 4$ and $p = 3$ for all six of our simulations.

Least squares: scheme 1,2,3



Rosenbrock: scheme 1,2,3



In the Rosenbrock simulations, we set $L_{\nabla f} = \|\nabla^2 f(x_0)\|_2$ where x_0 is an initial guess.

4.0.4 July 9:

By Lemma 3 in [**?**], we claim

$$\|g(x) - \nabla f(x)\|_2 \leq 2\|F(x)\|\frac{\mu}{2}(n+3)^{3/2}\sqrt{\sum_{i=1}^{p} L_{\nabla F_i}^2} \tag{91}$$

where $g(x) := 2F(x)\nabla F_\mu(x)$.

Proof. Let the correlation operator $B^{-1} = I$. Using lemma 3 on the components of F,

$$\begin{aligned}
\|g(x) - \nabla f(x)\|^2 &= \|2F(x)\nabla F(x) - 2F(x)\nabla F_\mu(x)\|^2 \\
&\leq 4\|F(x)\|^2\|\nabla F(x) - \nabla F_\mu(x)\|^2 \\
&\leq 4\|F(x)\|^2\sum_{i=1}^{p}\|\nabla F_i(x) - \nabla F_{i,\mu}(x)\|^2 \\
&\leq 4\|F(x)\|^2\sum_{i=1}^{p}\frac{\mu^2}{4}L_{\nabla F_i}^2(n+3)^3 \\
&= 4\|F(x)\|^2\frac{\mu^2}{4}(n+3)^3\sum_{i=1}^{p}L_{\nabla F_i}^2
\end{aligned} \tag{92}$$

□

Assuming the components of F are C^1 with gradient lipschitz, we have

$$\|F(x) - F_\mu(x)\| \leq \frac{\mu^2 n}{2}\sqrt{\sum_{i=1}^{p} L_{\nabla F_i}^2} \tag{93}$$

Proof. Using $F_i \in C^1$ with gradient lipschitz and by Theorem 1 in [**?**]

$$\begin{aligned}
\|F(x) - F_\mu(x)\|^2 &= \sum_{i=1}^{p}\|F_i(x) - F_{\mu,i}(x)\|^2 \\
&\leq \sum_{i=1}^{p}\left(\frac{\mu^2}{2}L_{\nabla F_i}n\right)^2 \\
&= \left(\frac{\mu^2}{2}n\right)^2\sum_{i=1}^{p}L_{\nabla F_i}^2
\end{aligned}$$

□

Recall equation (37), the above two facts, and the components of F are lipschitz together yield

$$\begin{aligned}
\|\nabla f(x) - \nabla(h\circ F_\mu)(x)\| &\leq 2\|F(x)\|\underbrace{\|\nabla F(x) - \nabla F_\mu(x)\|}_{\text{Use Lemma 3}} + \|\nabla F_\mu(x)\|\underbrace{\|F(x) - F_\mu(x)\|}_{\text{Use Theorem 1}} \\
&\leq 2\|F(x)\|\frac{\mu}{2}(n+3)^{3/2}\sqrt{\sum_{i=1}^{p}L_{\nabla F_i}^2} + \sqrt{\sum_{i=1}^{p}L_{\nabla F_i}^2}\frac{\mu^2 n}{2}\sqrt{\sum_{i=1}^{p}L_{\nabla F_i}^2}
\end{aligned}$$

$$= \sqrt{\sum_{i=1}^{p} L_{\nabla F_i}^2} \left(2\|F(x)\| \frac{\mu}{2}(n+3)^{3/2} + \frac{\mu^2 n}{2} \sqrt{\sum_{i=1}^{p} L_{\nabla F_i}^2} \right) \tag{94}$$

Suppose $L_{\nabla F_i} = L_{\nabla F_j}$ for all i, j. Then

$$\|\nabla f(x) - \nabla(h \circ F_\mu)(x)\| \leq \sqrt{p} L_{\nabla F_1} \left(2\|F(x)\| \frac{\mu}{2}(n+3)^{3/2} + \frac{\mu^2 n \sqrt{p}}{2} L_{\nabla F_1} \right) \tag{95}$$

Remark: At the moment, we haven't used any convexity assumptions.

Derivations of the scheme: Scheme 1 in 4.0.3 is taken from equation (54) in [**?**]. To justify scheme 2, observe the following computation:

$$\begin{aligned} \nabla(h \circ F_\mu) &= \nabla h(F_\mu(x)) \nabla F_\mu(x) \\ &= 2F_\mu(x) \nabla F_\mu(x) \\ &= 2\mathbb{E}_u[F(x+\mu u)]\mathbb{E}_u[\vec{g}_\mu(x)] \end{aligned} \tag{96}$$

where

$$\vec{g}_\mu(x) = \begin{bmatrix} \frac{F_1(x+\mu u) - F_1(x)}{\mu} \\ \vdots \\ \frac{F_p(x+\mu u) - F_p(x)}{\mu} \end{bmatrix} Bu \tag{97}$$

$[\vec{g}_\mu(x)]_i = g_{i,\mu}(x)$ with $\mathbb{E}_u[g_{i,\mu}(x)] = \nabla F_{i,\mu}(x)$. In view of (96), we have Scheme 2 as a result.

4.1 July 10

Scheme 3 from 4.0.3 is constructed from Scheme 2 while ignoring any stochasticity from F.

Scheme 3:

$$x_{k+1} = x_k - h_k 2F(x_k)\frac{F(x_k+\mu_k u)-F(x_k)}{\mu_k}u \tag{98}$$

$$g_3 = 2F(x_k)\frac{F(x_k+\mu_k u)-F(x_k)}{\mu_k}u \tag{99}$$

$$E_u[g_3] = 2F(x_k)E_u[\vec{g}_\mu(x)] \tag{100}$$

$$\begin{aligned}\|\nabla f(x) - E_u[g_3])\| &\le 2\|F(x)\|\underbrace{\|\nabla F(x) - \nabla F_\mu(x)\|}_{\text{Use Lemma 3}} \\ &\le 2\|F(x)\|\frac{\mu}{2}(n+3)^{3/2}\sqrt{\sum_{i=1}^p L^2_{\nabla F_i}}\end{aligned} \tag{101}$$

$$E_u[\|g_3 - \nabla f(x)\|] \le E_u[\|g_3 - E_u[g_3\|] + \|\nabla f(x) - E_u[g_3]\| \tag{102}$$

$$\le 2\|F(x_k)\|\left(E_u[\|\vec{g}_\mu(x) - E_u[\vec{g}_\mu(x)]\|] + \frac{\mu}{2}(n+3)^{3/2}\sqrt{\sum_{i=1}^p L^2_{\nabla F_i}}\right) \tag{103}$$

5 Week 5: July 13

5.1 July 13

We consider another test problem (Ridge Regression):

$$f(x) = \|Ax-b\|^2 + \lambda\|x\|^2 \tag{104}$$

where $\lambda > 0$. Observe on average, $\min_x f(x) > 0$. The first and second derivatives:

$$\nabla f(x) = 2(Ax-b)^\top A + 2\lambda x \tag{105}$$

$$\nabla^2 f(x) = 2(A^\top A + \lambda I) \tag{106}$$

In view of (105), the argmin of the ridge regression problem is $x^\star = (A^\top A + \lambda I)^{-1}A^\top b$.

We write $f = h \circ F$ where

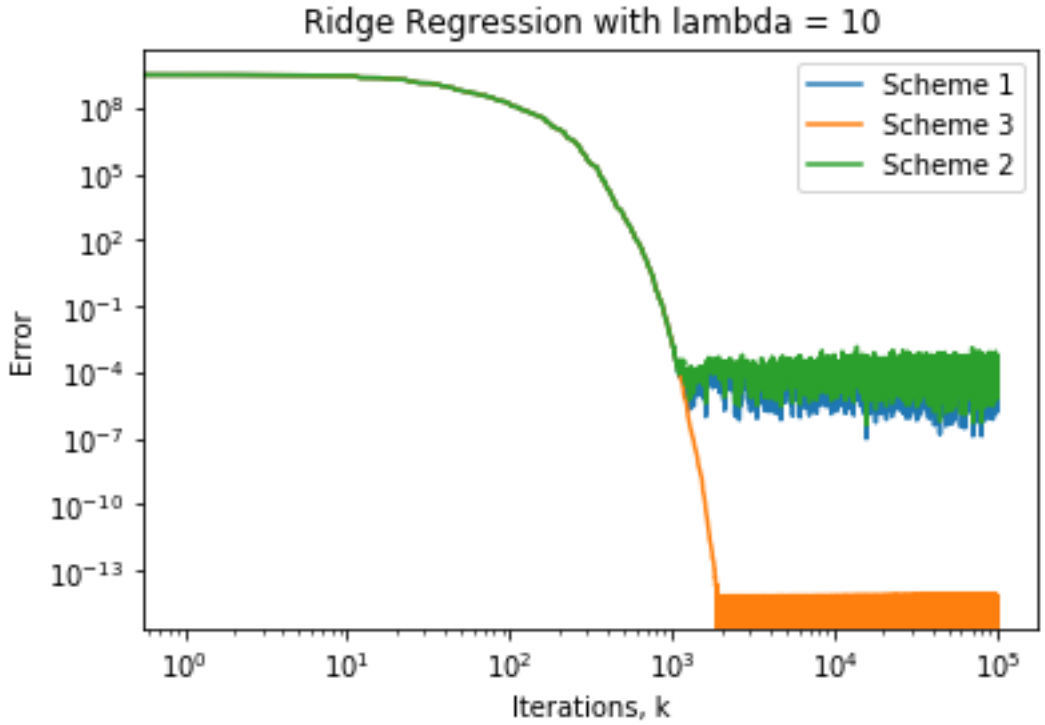
$$F(x) = (F_1(x), \ldots, F_p(x), G_1(x), \ldots, G_n(x)) \tag{107}$$

with

$$F_i(x) = a_i^\top x - b_i \quad \text{for } 1 \le i \le p \tag{108}$$

$$G_j(x) = \lambda^{1/2}x_j \quad \text{for } 1 \le j \le n \tag{109}$$

Here's the simulation for the three schemes (see 4.0.3):



Recall $g_\mu(x) = \frac{f(x+\mu u)-f(x)}{\mu}u$ and $g_{\mu,3}(x) := 2F(x)\frac{F(x+\mu u)-F(x)}{\mu}u$. By the triangle inequality and taking expectation,

$$\begin{aligned}\mathbb{E}_u[\|g_\mu(x) - \nabla f(x)\|] &\leq \mathbb{E}_u[\|g_\mu(x) - \nabla f_\mu(x)\|] + \|\nabla f(x) - \nabla f_u(x)\| \\ &\leq \mathbb{E}_u\|g_\mu(x) - \nabla f_\mu(x)\|] + \frac{\mu}{2}L_{\nabla f}(n+3)^{3/2} \end{aligned} \tag{110}$$

assuming $f \in C^1(E)$ and f has gradient Lipschitz.

Next, we prove the following.

Theorem 13.

$$\|\nabla F_\mu(x) - \nabla F(x)\| \leq \frac{\mu}{2}(n+3)^{3/2}\sqrt{\sum_{i=1}^{p} L_{\nabla F_i}^2} \tag{111}$$

Proof. Assuming the components of F is C^1 and has gradient Lipschitz and by Lemma 3 in [**?**],

$$\begin{aligned}\|\nabla F_\mu(x) - \nabla F(x)\|_2^2 &\leq \|\nabla F_\mu(x) - \nabla F(x)\|_F^2 \\ &= \sum_{i=1}^{p}\|\nabla F_{i,\mu}(x) - \nabla F_i(x)\|_2^2 \\ &\leq \sum_{i=1}^{p}\left(\frac{\mu}{2}L_{\nabla F_i}(n+3)^{3/2}\right)^2 \\ &= \left(\frac{\mu}{2}(n+3)^{3/2}\right)^2\sum_{i=1}^{p}L_{\nabla F_i}^2 \end{aligned} \tag{112}$$

□

Using this fact and assuming that $L_{\nabla F_i} = L_{\nabla F_j}$ for all i, j,

$$\begin{aligned}
\mathbb{E}_u[\|g_{\mu,3}(x) - \nabla f(x)\|] &= \mathbb{E}_u\left[\|2F(x)\left\{\frac{F(x+\mu u)-F(x)}{\mu}u - \nabla F(x)\right\}\|\right] \\
&\leq 2\|F(x)\|\mathbb{E}_u\left[\|\frac{F(x+\mu u)-F(x)}{\mu}u - \nabla F(x)\|\right] \\
&\leq 2\|F(x)\|\mathbb{E}_u\left[\|\frac{F(x+\mu u)-F(x)}{\mu}u - \nabla F_\mu(x)\|\right] + 2\|F(x)\|\|\nabla F_\mu(x) - \nabla F(x)\| \\
&\leq 2\|F(x)\|\mathbb{E}_u\left[\|\frac{F(x+\mu u)-F(x)}{\mu}u - \nabla F_\mu(x)\|\right] + 2\|F(x)\|\frac{\mu}{2}(n+3)^{3/2}\sqrt{p}L_{\nabla F_1}
\end{aligned} \tag{113}$$

Question: How does one deal with $\mathbb{E}_u\|g_\mu(x) - \nabla f_\mu(x)\|$? If such an approach exists, then dealing $\mathbb{E}_u\left[\|\frac{F(x+\mu u)-F(x)}{\mu}u - \nabla F_\mu(x)\|\right]$ can be dealt with similarly.

5.2 July 14

We prove

Theorem 14.

$$\mathbb{E}_u[\|g_\mu(x) - \nabla f(x)\|] \leq \mu L_{\nabla f}(n+6)^{3/2} + 2(\sqrt{n}+3)\|\nabla f(x)\| \tag{114}$$

assuming that $f \in C^1(E)$ with lipschitz gradient and is convex.

Proof. By Theorem 4 in [**?**] and Jensen's inequality, we arrive at

$$\begin{aligned}
\mathbb{E}_u[\|g_\mu(x) - \nabla f(x)\|] &\leq \sqrt{\mathbb{E}_u\|g_\mu(x) - \nabla f(x)\|^2} \\
&\leq \sqrt{2\mathbb{E}_u\{\|g_\mu(x)\|^2 + \|\nabla f(x)\|^2\}} \\
&= \sqrt{2}\sqrt{\mathbb{E}_u\|g_\mu(x)\|^2 + \|\nabla f(x)\|^2} \\
&\leq \sqrt{2}\left(\sqrt{\mathbb{E}_u\|g_\mu\|^2} + \|\nabla f(x)\|\right) \\
&\leq \sqrt{2}\left(\sqrt{\frac{\mu^2}{2}L_{\nabla f}^2(n+6)^3 + 2(n+4)\|\nabla f(x)\|^2} + \|\nabla f(x)\|\right) \\
&\leq \sqrt{2}\left(\sqrt{\frac{\mu^2}{2}L_{\nabla f}^2(n+6)^3} + \sqrt{2(n+4)\|\nabla f(x)\|^2} + \|\nabla f(x)\|\right) \\
&= \sqrt{2}\left(\frac{\mu}{\sqrt{2}}L_{\nabla f}(n+6)^{3/2} + \sqrt{2}\sqrt{n+4}\|\nabla f(x)\| + \|\nabla f(x)\|\right) \\
&\leq \mu L_{\nabla f}(n+6)^{3/2} + 2(\sqrt{n}+3)\|\nabla f(x)\|
\end{aligned} \tag{115}$$

□

By Theorem 2 and recycling the proof, we have

Theorem 15.

$$\mathbb{E}_u[\|\hat{g}_\mu(x) - \nabla f(x)\|] \leq \frac{\mu}{2}L_{\nabla f}(n+6)^{3/2} + 2(\sqrt{n}+3)\|\nabla f(x)\| \tag{116}$$

Assume the components of F are $C^1(E)$ with lipschitz gradient and is convex.

Theorem 16. *We prove*

$$\mathbb{E}_u[\|g_{\mu,3}(x) - \nabla f(x)\|] \leq 2\|F(x)\| \left\{ \sum_{i=1}^{p} \mu L_{\nabla F_i} (n+6)^{3/2} + 2(\sqrt{n}+3)\|\nabla F_i(x)\| \right\} \tag{117}$$

Proof. Using (113) and recycling the proof of the fact (114), we have

$$\begin{aligned}
\mathbb{E}_u \| \frac{F(x+\mu u) - F(x)}{\mu} u - \nabla F(x)\|_2 &\leq \mathbb{E}_u \| \frac{F(x+\mu u) - F(x)}{\mu} u - \nabla F(x)\|_F \\
&= \mathbb{E}_u \sqrt{\sum_{i=1}^{p} \| \frac{F_i(x+\mu u) - F_i(x)}{\mu} u - \nabla F_i(x)\|^2} \\
&\leq \mathbb{E}_u \sum_{i=1}^{p} \| \frac{F_i(x+\mu u) - F_i(x)}{\mu} u - \nabla F_i(x)\| \\
&= \sum_{i=1}^{p} \mathbb{E}_u \| \frac{F_i(x+\mu u) - F_i(x)}{\mu} u - \nabla F_i(x)\| \\
&\leq \sum_{i=1}^{p} \left\{ \mu L_{\nabla F_i} (n+6)^{3/2} + 2(\sqrt{n}+3)\|\nabla F_i(x)\| \right\} \qquad (118)
\end{aligned}$$

Therefore, the conclusion follows. Observe if we assume $L_{\nabla F_i} = L_{\nabla F_j}$ for all i, j,

$$\mathbb{E}_u \|g_{\mu,3}(x) - \nabla f(x)\| \leq 2\|F(x)\| \left\{ p\mu L_{\nabla F_1} (n+6)^{3/2} + 2p(\sqrt{n}+3) \max_{1\leq i\leq p} \|\nabla F_i(x)\| \right\} \tag{119}$$

□

Then using (116), we have

Theorem 17.

$$\mathbb{E}_u \left[\left\| \frac{F(x+\mu u) - F(x-\mu u)}{2\mu} u - \nabla F(x) \right\| \right] \leq \sum_{i=1}^{p} \left\{ \frac{\mu}{2} L_{\nabla F_i} (n+6)^{3/2} + 2(\sqrt{n}+3)\|\nabla F_i(x)\| \right\} \tag{120}$$

Consequently,

$$\mathbb{E}_u[\|\hat{g}_{\mu,3}(x) - \nabla f(x)\|] \leq 2\|F(x)\| \sum_{i=1}^{p} \left\{ \frac{\mu}{2} L_{\nabla F_i} (n+6)^{3/2} + 2(\sqrt{n}+3)\|\nabla F_i(x)\| \right\} \tag{121}$$

Assume that $L_{\nabla F_i} = L_{\nabla F_j}$ for all i, j, we have

$$\mathbb{E}_u \left[\left\| \frac{F(x+\mu u) - F(x-\mu u)}{2\mu} u - \nabla F(x) \right\| \right] \leq \frac{\mu}{2} p L_{\nabla F_1} (n+6)^{3/2} + 2p(\sqrt{n}+3) \max_{1\leq i\leq p} \|\nabla F_i(x)\| \tag{122}$$

Consequently,

$$\mathbb{E}_u[\|\hat{g}_{\mu,3}(x) - \nabla f(x)\|] \leq 2\|F(x)\| \left[\frac{\mu}{2} p L_{\nabla F_1} (n+6)^{3/2} + 2p(\sqrt{n}+3) \max_{1\leq i\leq p} \|\nabla F_i(x)\| \right] \tag{123}$$

5.3 July 15

Lemma 18.

$$\mathbb{E}_u\left[\left\|\frac{F(x+\mu u)-F(x)}{\mu}u-\nabla F(x)\right\|^2\right]\leq\sum_{i=1}^{p}\left(\mu^2L_{\nabla F_i}^2(n+6)^3+4(n+5)\|\nabla F_i(x)\|^2\right)$$

Proof.

$$\begin{aligned}
\mathbb{E}_u\left[\left\|\frac{F(x+\mu u)-F(x)}{\mu}u-\nabla F(x)\right\|^2\right]&\leq\mathbb{E}_u\left[\left\|\frac{F(x+\mu u)-F(x)}{\mu}u-\nabla F(x)\right\|_F^2\right]\\
&=\sum_{i=1}^{p}\mathbb{E}_u\left[\left\|\frac{F_i(x+\mu u)-F_i(x)}{\mu}u-\nabla F_i(x)\right\|^2\right]\\
&\leq 2\sum_{i=1}^{p}\left(\mathbb{E}_u\left[\left\|\frac{F_i(x+\mu u)-F_i(x)}{\mu}u\right\|^2\right]+\|\nabla F_i(x)\|^2\right)\\
&\leq 2\sum_{i=1}^{p}\left(\frac{\mu^2}{2}L_{\nabla F_i}^2(n+6)^3+2(n+4)\|\nabla F_i(x)\|^2+\|\nabla F_i(x)\|^2\right)\\
&\leq 2\sum_{i=1}^{p}\left(\frac{\mu^2}{2}L_{\nabla F_i}^2(n+6)^3+2(n+5)\|\nabla F_i(x)\|^2\right)
\end{aligned}$$

□

Lemma 19.

$$\mathbb{E}_u\left[\left\|\frac{F(x+\mu u)-F(x-\mu u)}{2\mu}u-\nabla F(x)\right\|^2\right]\leq\sum_{i=1}^{p}\left(\frac{\mu^2}{4}L_{\nabla F_i}^2(n+6)^3+4(n+5)\|\nabla F_i(x)\|^2\right)$$

Proof.

$$\begin{aligned}
\mathbb{E}_u\left[\left\|\frac{F(x+\mu u)-F(x-\mu u)}{2\mu}u-\nabla F(x)\right\|^2\right]&\leq\mathbb{E}_u\left[\left\|\frac{F(x+\mu u)-F(x-\mu u)}{2\mu}u-\nabla F(x)\right\|_F^2\right]\\
&=\sum_{i=1}^{p}\mathbb{E}_u\left[\left\|\frac{F_i(x+\mu u)-F_i(x-\mu u)}{2\mu}u-\nabla F_i(x)\right\|^2\right]\\
&\leq 2\sum_{i=1}^{p}\left(\mathbb{E}_u\left[\left\|\frac{F_i(x+\mu u)-F_i(x-\mu u)}{2\mu}u\right\|^2\right]+\|\nabla F_i(x)\|^2\right)\\
&\leq 2\sum_{i=1}^{p}\left(\frac{\mu^2}{8}L_{\nabla F_i}^2(n+6)^3+2(n+4)\|\nabla F_i(x)\|^2+\|\nabla F_i(x)\|^2\right)\\
&\leq 2\sum_{i=1}^{p}\left(\frac{\mu^2}{8}L_{\nabla F_i}^2(n+6)^3+2(n+5)\|\nabla F_i(x)\|^2\right)\\
&=\sum_{i=1}^{p}\left(\frac{\mu^2}{4}L_{\nabla F_i}^2(n+6)^3+4(n+5)\|\nabla F_i(x)\|^2\right)
\end{aligned}$$

□

Define $g_{\mu,2}(x)=2F(x+\mu u)\left[\frac{F(x+\mu u)-F(x)}{\mu}\right]u$.

We have the following fact:

Theorem 20.

$$\mathbb{E}_u \|g_{\mu,2}(x) - \nabla f(x)\| \leq 2L_F\mu \left[n + \sum_{i=1}^{p} \left\{ \mu^2 L_{\nabla F_i}^2 (n+6)^3 + 4(n+5)\|\nabla F_i(x)\|^2 \right\} \right] + 2\|F(x)\| \sum_{i=1}^{p} \left\{ \mu L_{\nabla F_i} (n+6)^{3/2} + 2(\sqrt{n}+3)\|\nabla F_i(x)\| \right\} + 2L_F^2 \mu\sqrt{n} \quad (124)$$

Assuming $L_{\nabla F_i} = L_{\nabla F_j}$ for all i, j, we have

$$\mathbb{E}_u \|g_{\mu,2}(x) - \nabla f(x)\| \leq 2L_F\mu \left[n + p\mu^2 L_{\nabla F_1}^2 (n+6)^3 + 4(n+5)\|\nabla F(x)\|_F^2 \right] + 2\|F(x)\| \left\{ p\mu L_{\nabla F_1} (n+6)^{3/2} + 2p(\sqrt{n}+3) \max_{1\leq i\leq p} \|\nabla F_i(x)\| \right\} + 2L_F^2 \mu\sqrt{n} \quad (125)$$

Proof.

$$\|g_{\mu,2}(x) - \nabla f(x)\| \leq \left\| 2F(x+\mu u) \left[\frac{F(x+\mu u) - F(x)}{\mu} u - \nabla F(x) \right] \right\| + \|[2F(x+\mu u) - 2F(x)]\nabla F(x)\|$$
$$\leq 2\|F(x+\mu u)\| \left\| \frac{F(x+\mu u) - F(x)}{\mu} u - \nabla F(x) \right\| + 2\|\nabla F(x)\|\|F(x+\mu u) - F(x)\|$$

Taking expectation of the above equation and using Jensen's inequality, we have

$$\mathbb{E}_u \|g_{\mu,2}(x) - \nabla f(x)\| \leq 2\mathbb{E}_u \left[\|F(x+\ \mu u) - F(x)\| \left\| \frac{F(x+\mu u) - F(x)}{\mu} u - \nabla F(x) \right\| \right]$$
$$\leq 2\mathbb{E}_u \left[\|F(x)\| \left\| \frac{F(x+\mu u) - F(x)}{\mu} u - \nabla F(x) \right\| \right] + 2L_F^2 \mu\sqrt{n}$$
$$\leq 2L_F\mu\mathbb{E}_u \left[\|u\| \left\| \frac{F(x+\mu u) - F(x)}{\mu} u - \nabla F(x) \right\| \right]$$
$$+ 2\mathbb{E}_u \left[\|F(x)\| \left\| \frac{F(x+\mu u) - F(x)}{\mu} u - \nabla F(x) \right\| \right] + 2L_F^2 \mu\sqrt{n}$$
$$\leq 2L_F\mu\mathbb{E}_u \left[\|u\|^2 + \left\| \frac{F(x+\mu u) - F(x)}{\mu} u - \nabla F(x) \right\|^2 \right]$$
$$+ 2\mathbb{E}_u \left[\|F(x)\| \left\| \frac{F(x+\mu u) - F(x)}{\mu} u - \nabla F(x) \right\| \right] + 2L_F^2 \mu\sqrt{n}$$
$$\leq 2L_F\mu \left[n + \sum_{i=1}^{p} (\mu^2 L_{\nabla F_i}^2 (n+6)^3 + 4(n+5)\|\nabla F_i(x)\|^2) \right]$$
$$+ 2\|F(x)\| \sum_{i=1}^{p} \left\{ \mu L_{\nabla F_i} (n+6)^{3/2} + 2(\sqrt{n}+3)\|\nabla F_i(x)\| \right\} + 2L_F^2 \mu\sqrt{n} \quad (126)$$

Assuming $L_{\nabla F_1} = L_{\nabla F_j}$ for all j, we have

$$\mathbb{E}_u \|g_{\mu,2}(x) - \nabla f(x)\| \leq 2L_F\mu \left[n + p\mu^2 L_{\nabla F_1}^2 (n+6)^3 + 4(n+5)\|\nabla F(x)\|_F^2 \right] + 2\|F(x)\| \left\{ p\mu L_{\nabla F_1} (n+6)^{3/2} + 2p(\sqrt{n}+3) \max_{1\leq i\leq p} \|\nabla F_i(x)\| \right\} + 2L_F^2 \mu\sqrt{n} \quad (127)$$

□

We have the following fact:

Theorem 21.

$$\mathbb{E}_u \|\hat{g}_{\mu,2}(x) - \nabla f(x)\| \leq 2L_F\mu \left[n + \sum_{i=1}^{p} \left\{ \frac{\mu^2}{4} L^2_{\nabla F_i} (n+6)^3 + 4(n+5)\|\nabla F_i(x)\|^2 \right\} \right] + 2\|F(x)\| \sum_{i=1}^{p} \left\{ \frac{\mu}{2} L_{\nabla F_i} (n+6)^{3/2} + 2(\sqrt{n}+3)\|\nabla F_i(x)\| \right\} + 2L_F^2 \mu\sqrt{n} \quad (128)$$

Assuming $L_{\nabla F_i} = L_{\nabla F_j}$ for all i, j, we have

$$\mathbb{E}_u \|\hat{g}_{\mu,2}(x) - \nabla f(x)\| \leq 2L_F\mu \left[n + \frac{p\mu^2}{4} L^2_{\nabla F_1} (n+6)^3 + 4(n+5)\|\nabla F(x)\|_F^2 \right] + 2\|F(x)\| \left[\frac{p\mu}{2} L_{\nabla F_1} (n+6)^{3/2} + 2p(\sqrt{n}+3) \max_{1\leq i\leq p} \|\nabla F_i(x)\| \right] + 2L_F^2 \mu\sqrt{n} \quad (129)$$

Proof.

$$\begin{aligned}
\|\hat{g}_{\mu,2}(x) - \nabla f(x)\| &\leq \left\| 2F(x+\mu u) \left[\frac{F(x+\mu u) - F(x-\mu u)}{2\mu} u - \nabla F(x) \right] \right\| + \|[2F(x+\mu u) - 2F(x)]\nabla F(x)\| \\
&\leq 2\|F(x+\mu u)\| \left\| \frac{F(x+\mu u) - F(x-\mu u)}{2\mu} u - \nabla F(x) \right\| + 2\|\nabla F(x)\|\|F(x+\mu u) - F(x)\| \\
&\leq 2[\|F(x+\mu u) - F(x)\|] \left\| \frac{F(x+\mu u) - F(x-\mu u)}{2\mu} u - \nabla F(x) \right\| \\
&+ 2[\|F(x)\|] \left\| \frac{F(x+\mu u) - F(x-\mu u)}{2\mu} u - \nabla F(x) \right\| + 2L_F^2 \mu\|u\| \\
&\leq 2L_F\mu\|u\| \left\| \frac{F(x+\mu u) - F(x-\mu u)}{2\mu} u - \nabla F(x) \right\| \\
&+ 2\|F(x)\| \left\| \frac{F(x+\mu u) - F(x-\mu u)}{2\mu} u - \nabla F(x) \right\| + 2L_F^2 \mu\|u\|
\end{aligned}$$

Taking the expectation of both sides,

$$\begin{aligned}
\mathbb{E}_u \left[\|\hat{g}_{\mu,2}(x) - \nabla f(x)\| \right] &\leq 2L_F\mu\mathbb{E}_u \left[\|u\| \left\| \frac{F(x+\mu u) - F(x-\mu u)}{2\mu} u - \nabla F(x) \right\| \right] \\
&+ 2\|F(x)\|\mathbb{E}_u \left[\left\| \frac{F(x+\mu u) - F(x-\mu u)}{2\mu} u - \nabla F(x) \right\| \right] + 2L_F^2 \mu\sqrt{n} \\
&\leq 2L_F\mu\mathbb{E}_u\|u\|^2 + 2L_F^2\mu\sqrt{n} + 2L_F\mu\mathbb{E}_u \left[\left\| \frac{F(x+\mu u) - F(x-\mu u)}{2\mu} u - \nabla F(x) \right\|^2 \right] \\
&+ 2\|F(x)\|\mathbb{E}_u \left[\left\| \frac{F(x+\mu u) - F(x-\mu u)}{2\mu} u - \nabla F(x) \right\| \right] \\
&\leq 2L_F^2\mu\sqrt{n} + 2L_F\mu n + 2L_F\mu \sum_{i=1}^{p} \left\{ \frac{\mu^2}{4} L^2_{\nabla F_i} (n+6)^3 + 4(n+5)\|\nabla F_i(x)\|^2 \right\} \\
&+ 2\|F(x)\| \sum_{i=1}^{p} \left\{ \frac{\mu}{2} L_{\nabla F_i} (n+6)^{3/2} + 2(\sqrt{n}+3)\|\nabla F_i(x)\| \right\}
\end{aligned}$$

Assuming $L_{\nabla F_i} = L_{\nabla F_j}$ for all i, j, we have

$$\mathbb{E}_u\left[\|\hat{g}_{\mu,2}(x) - \nabla f(x)\|\right] \leq 2L_F\mu[n + \frac{p\mu^2}{4}L^2_{\nabla F_1}(n+6)^3 + 4(n+5)\|\nabla F(x)\|_F^2] + 2\|F(x)\|\left[\frac{p\mu}{2}L_{\nabla F_1}(n+6)^{3/2} + 2p(\sqrt{n}+3)\max_{1\leq i\leq p}\|\nabla F_i(x)\|\right] + 2L_F^2\mu\sqrt{n}$$

□

Recall the following:

$$g_\mu(x) = \frac{f(x+\mu u) - f(x)}{\mu}u \tag{130}$$

$$\hat{g}_\mu(x) = \frac{f(x+\mu u) - f(x-\mu u)}{2\mu}u \tag{131}$$

$$g_{\mu,2}(x) = 2F(x+\mu u)\left[\frac{F(x+\mu u) - F(x)}{\mu}\right]u \tag{132}$$

$$\hat{g}_{\mu,2}(x) = 2F(x+\mu u)\left[\frac{F(x+\mu u) - F(x-\mu u)}{2\mu}\right]u \tag{133}$$

$$g_{\mu,3}(x) = 2F(x)\left[\frac{F(x+\mu u) - F(x)}{\mu}\right]u \tag{134}$$

$$\hat{g}_{\mu,3}(x) = 2F(x)\left[\frac{F(x+\mu u) - F(x-\mu u)}{2\mu}\right]u \tag{135}$$

Let g be the gradient estimate above. We assume $L_{\nabla F_1} = L_{\nabla F_i}$ for all $1 \leq i \leq p$.

error	$\mathbb{E}_u[\|g - \nabla f(x)\|] \leq$ error
S1 FD	$\mu L_{\nabla f}(n+6)^{3/2} + 2(\sqrt{n}+3)\|\nabla f(x)\|$
S1 CD	$\frac{\mu}{2}L_{\nabla f}(n+6)^{3/2} + 2(\sqrt{n}+3)\|\nabla f(x)\|$
S2 FD	$2L_F\mu[n + p\mu^2 L^2_{\nabla F_1}(n+6)^3 + 4(n+5)\|\nabla F(x)\|_F^2] + 2\|F(x)\|\left\{p\mu L_{\nabla F_1}(n+6)^{3/2} + 2p(\sqrt{n}+3)\max_{1\leq i\leq p}\|\nabla F_i(x)\|\right\} + 2L_F^2\mu\sqrt{n}$
S2 CD	$2L_F\mu\left[n + \frac{p\mu^2}{4}L^2_{\nabla F_1}(n+6)^3 + 4(n+5)\|\nabla F(x)\|_F^2\right] + 2\|F(x)\|\left[\frac{p\mu}{2}L_{\nabla F_1}(n+6)^{3/2} + 2p(\sqrt{n}+3)\max_{1\leq i\leq p}\|\nabla F_i(x)\|\right] + 2L_F^2\mu\sqrt{n}$
S3 FD	$2\|F(x)\|\left\{p\mu L_{\nabla F_1}(n+6)^{3/2} + 2p(\sqrt{n}+3)\max_{1\leq i\leq p}\|\nabla F_i(x)\|\right\}$
S3 CD	$2\|F(x)\|\left\{\frac{\mu}{2}pL_{\nabla F_1}(n+6)^{3/2} + 2p(\sqrt{n}+3)\max_{1\leq i\leq p}\|\nabla F_i(x)\|\right\}$

6 Week 6: July 20

6.1 July 20

For notational brevity, $g_\mu(x) = 2F(x)\left[\frac{F(x+\mu u)-F(x)}{\mu}\right]u$.

We try to mimc the convergence analysis in Theorem 8 in [**?**].

Compute $\mathbb{E}_u\|g_\mu(x)\|^2$.

$$g_\mu(x) = 2\sum_{i=1}^{p} F_i(x)g_{\mu,i}(x)u \tag{136}$$

$$\|g_\mu(x)\|^2 \leq 4\|F(x)\|^2 \left\| \frac{F(x+\mu u)-F(x)}{\mu} u \right\|^2 \tag{137}$$

$$\begin{aligned}
\mathbb{E}_u\|g_\mu(x)\|^2 &\leq 4\|F(x)\|^2 \mathbb{E}_u \left[\left\| \frac{F(x+\mu u)-F(x)}{\mu} u \right\|^2 \right] \\
&\leq 4\|F(x)\|^2 \sum_{i=1}^p \frac{\mu^2}{2} L_{\nabla F_i}^2 (n+6)^3 + 2(n+4)\|\nabla F_i(x)\|^2 \\
&= 4\|F(x)\|^2 \sum_{i=1}^p \frac{\mu^2}{2} L_{\nabla F_i}^2 (n+6)^3 + 2(n+4)\|\nabla F_i(x)\|^2
\end{aligned} \tag{138}$$

Define sequence $x_{k+1} = x_k - hg_\mu(x_k)$ and $r_k = \|x_k - x^*\|$ where x^* is the optimal solution to minimizing $f(x)$.

$$\begin{aligned}
r_{k+1}^2 &= \|x_{k+1} - x^*\|^2 \\
&= \|x_k - hg_\mu(x_k) - x^*\|^2 \\
&= \|x_k - x^* - hg_\mu(x_k)\|^2 \\
&= r_k^2 - 2h\langle g_\mu(x_k), x_k - x^*\rangle + h^2\|g_\mu(x_k)\|^2
\end{aligned} \tag{139}$$

Observe that

$$\mathbb{E}_u g_\mu(x_k) = 2\sum_{i=1}^p F_i(x_k)\nabla F_{\mu,i}(x_k) \tag{140}$$

$$\begin{aligned}
\mathbb{E}_{u_k}(r_{k+1}^2) &\leq r_k^2 - 2h\left\langle \sum_{i=1}^p 2F_i(x_k)\nabla F_{\mu,i}(x_k), x_k - x^* \right\rangle + 2\|F(x_k)\|^2 h^2 \sum_{i=1}^p \frac{\mu^2}{2} L_{\nabla F_i}^2 (n+6)^3 + 2(n+4)\|\nabla F_i(x_k)\|^2 \\
&= r_k^2 - 2h\sum_{i=1}^p \langle 2F_i(x_k)\nabla F_{\mu,i}(x_k), x_k - x^*\rangle + 2\|F(x_k)\|^2 h^2 \sum_{i=1}^p \frac{\mu^2}{2} L_{\nabla F_i}^2 (n+6)^3 + 2(n+4)\|\nabla F_i(x_k)\|^2 \\
&= r_k^2 - 2h\sum_{i=1}^p 2F_i(x_k)\langle \nabla F_{\mu,i}(x_k), x_k - x^*\rangle + 2\|F(x_k)\|^2 h^2 \sum_{i=1}^p \frac{\mu^2}{2} L_{\nabla F_i}^2 (n+6)^3 + 2(n+4)\|\nabla F_i(x_k)\|^2 \\
&\leq r_k^2 - 2h\sum_{i=1}^p 2F_i(x_k)\left[F_i(x_k) - F_{i,\mu}(x^*)\right] + 2\|F(x_k)\|^2 h^2 \sum_{i=1}^p \frac{\mu^2}{2} L_{\nabla F_i}^2 (n+6)^3 + 4(n+4)L_{\nabla F_i}(F_i(x_k) - F_i(x^*))
\end{aligned} \tag{141}$$

Lemma 22. *Assume $F_i, \nabla F_i$ are Lipschitz continuous for all $1 \leq i \leq p$. Then*

$$\|\nabla f(x) - \nabla f(y)\| \leq L_{\nabla f}(x,y)\|x - y\| \quad \forall x, y \in \mathbb{R}^n \tag{142}$$

where $L_{\nabla f}(x,y) = L_{\nabla F}^2 + 2L_F^2 + \|F(x)\|^2$ with

$$L_F = \sqrt{\sum_{i=1}^p L_{F_i}^2} \tag{143}$$

$$L_{\nabla F} = \sqrt{\sum_{i=1}^p L_{\nabla F_i}^2} \tag{144}$$

Proof. Recall

$$\nabla f(x) - \nabla f(y) = 2\sum_{i=1}^{p} \{F_i(x)[\nabla F_i(x) - \nabla F_i(y)] + [F_i(x) - F_i(y)]\nabla F_i(y)\} \tag{145}$$

By hypothesis,

$$\begin{aligned}
\|\nabla f(x) - \nabla f(y)\| &= 2\left\|\sum_{i=1}^{p} F_i(x)[\nabla F_i(x) - \nabla F_i(y)] + [F_i(x) - F_i(y)]\nabla F_i(y)\right\| \\
&\leq 2\sum_{i=1}^{p} \|F_i(x)\|\|\nabla F_i(x) - \nabla F_i(y)\| + \|F_i(x) - F_i(y)\|\|\nabla F_i(y)\| \\
&\leq 2\sum_{i=1}^{p} \|F_i(x)\|L_{\nabla F_i}\|x - y\| + L_{F_i}^2\|x - y\| \\
&= \sum_{i=1}^{p} \left(2L_{\nabla F_i}\|F_i(x)\| + 2L_{F_i}^2\right)\|x - y\| \\
&= \|x - y\|\sum_{i=1}^{p} \left(2L_{\nabla F_i}\|F_i(x)\| + 2L_{F_i}^2\right) \\
&= \|x - y\|\left(\sum_{i=1}^{p} 2L_{\nabla F_i}\|F_i(x)\| + \sum_{i=1}^{p} 2L_{F_i}^2\right) \\
&\leq \|x - y\|\left(\sum_{i=1}^{p} L_{\nabla F_i}^2 + \|F_i(x)\|^2 + 2\sum_{i=1}^{p} L_{F_i}^2\right) \\
&\leq \|x - y\|\left(L_{\nabla F}^2 + \|F(x)\|^2 + 2L_F^2\right) \\
&= \|x - y\|L_{\nabla f}(x, y)
\end{aligned}$$

□

Unless stated otherwise, assume $g_\mu(x) = 2F(x)\left[\frac{F(x+\mu u) - F(x)}{\mu}\right]u$. In our setting,

$$f(x) = \|F(x)\|^2, \tag{146}$$

$$x_{k+1} = x_k - hg_\mu(x_k). \tag{147}$$

Second approach to convergence analysis:

$$\begin{aligned}
f(x_{k+1}) &\leq f(x_k) + \langle \nabla f(x_k), -hg_\mu(x_k)\rangle + \frac{L_{\nabla f}}{2}\|hg_\mu(x_k)\|^2 \\
&= f(x_k) - h\langle \nabla f(x_k), g_\mu(x_k)\rangle + \frac{L_{\nabla f}}{2}h^2\|g_\mu(x_k)\|^2
\end{aligned}$$

By (138),

$$\begin{aligned}
\mathbb{E}_{u_k} f(x_{k+1}) &\leq f(x_k) - h\mathbb{E}_{u_k}\langle \nabla f(x_k), g_\mu(x_k)\rangle + \frac{L_{\nabla f}}{2}h^2\mathbb{E}_{u_k}\|g_\mu(x_k)\|^2 \\
&\leq f(x_k) - h\mathbb{E}_{u_k}\langle \nabla f(x_k), g_\mu(x_k)\rangle + \frac{L_{\nabla f}}{2}h^2\left[2\|F(x_k)\|^2\sum_{i=1}^{p}\frac{\mu^2}{2}L_{\nabla F_i}^2(n+6)^3 + 2(n+4)\|\nabla F_i(x_k)\|^2\right]
\end{aligned}$$

6.2 July 21

Lemma 23. *Assume $F_i, \nabla F_i$ are Lipschitz continuous for all $1 \le i \le p$. Then for all $x, y \in \mathbb{R}^n$,*

$$|f(x) - f(y) - \nabla f(y)(x-y)| \le \frac{1}{2}(L_{\nabla F}^2 + 2L_F^2 + \|F(y)\|^2)\|x-y\|^2 \tag{148}$$

Proof. By the following computation and lemma 22,

$$\begin{aligned}
|f(x) - f(y) - \nabla f(y)(x-y)| &= \left| \int_0^1 [\nabla f(y + t(x-y)) - \nabla f(y)](x-y)dt \right| \\
&\le \int_0^1 \|\nabla f(y + t(x-y)) - \nabla f(y)\|\|x-y\|dt \\
&= \int_0^1 (L_{\nabla F}^2 + 2L_F^2 + \|F(y)\|^2)\|x-y\|^2 t dt \\
&= \frac{1}{2}(L_{\nabla F}^2 + 2L_F^2 + \|F(y)\|^2)\|x-y\|^2
\end{aligned}$$

□

Theorem 24. *Remark: We prove a result analogous to theorem 4 in [**?**], which we state here. Recall that $g_\mu(x) = \frac{f(x+\mu u) - f(x)}{\mu} u$ and $\hat{g}_\mu(x) = \frac{f(x+\mu u) - f(x-\mu u)}{2\mu} u$. Then*

$$\mathbb{E}_u(\|g_\mu(x)\|^2) \le \frac{1}{2}(L_{\nabla F}^2 + 2L_F^2 + \|F(x)\|^2)^2 \mu^2 (n+6)^3 + 2(n+4)\|\nabla f(x)\|^2 \tag{149}$$

$$\mathbb{E}_u(\|\hat{g}_\mu(x)\|^2) \le \frac{\mu^2}{8}(L_{\nabla F}^2 + 2L_F^2 + \|F(x)\|^2)^2 (n+6)^3 + 2(n+4)\|\nabla f(x)\|^2 \tag{150}$$

Proof. By Young's inequality and lemma 23, we have

$$\begin{aligned}
[f(x+\mu u) - f(x)]^2 &= [f(x+\mu u) - f(x) - \langle \nabla f(x), \mu u\rangle + \langle \nabla f(x), \mu u\rangle]^2 \\
&\le 2(f(x+\mu u) - f(x) - \langle \nabla f(x), \mu u\rangle)^2 + 2\mu^2 \langle \nabla f(x), \mu\rangle^2 \\
&\le 2\left(\frac{1}{2}(L_{\nabla F}^2 + 2L_F^2 + \|F(x)\|^2)\mu^2\|u\|^2\right)^2 + 2\mu^2\langle \nabla f(x), u\rangle^2 \\
&= \frac{1}{2}(L_{\nabla F}^2 + 2L_F^2 + \|F(x)\|^2)^2 \mu^4 \|u\|^4 + 2\mu^2 \langle \nabla f(x), u\rangle^2
\end{aligned}$$

Dividing both sides by μ^2, taking expectation, multiplying by $\|u\|^2$, and following the steps in Nesterov's proof 4.2 equation (35), we have

$$\begin{aligned}
\mathbb{E}_u(\|g_\mu(x)\|^2) &\le \frac{1}{2}(L_{\nabla F}^2 + 2L_F^2 + \|F(x)\|^2)^2 \mu^2 \mathbb{E}_u(\|u\|^6) + 2\mathbb{E}_u[\langle \nabla f(x), u\rangle^2 \|u\|^2] \\
&\le \frac{1}{2}(L_{\nabla F}^2 + 2L_F^2 + \|F(x)\|^2)^2 \mu^2 (n+6)^3 + 2(n+4)\|\nabla f(x)\|^2
\end{aligned}$$

For the symmetric oracle $\hat{g}_\mu$, since f is convex, we have

$$\begin{aligned}
f(x+\mu u) - f(x-\mu u) &= [f(x+\mu u) - f(x)] + [f(x) - f(x-\mu u)] \\
&\le [\langle \nabla f(x), \mu u\rangle + \frac{\mu^2}{2}(L_{\nabla F}^2 + 2L_F^2 + \|F(x)\|^2)\|u\|^2] + \mu\langle \nabla f(x), u\rangle
\end{aligned}$$

$$= 2\mu\langle\nabla f(x), u\rangle + \frac{\mu^2}{2}(L_{\nabla F}^2 + 2L_F^2 + \|F(x)\|^2)\|u\|^2$$

Similarly, we claim that $f(x+\mu u) - f(x-\mu u) \geq 2\mu\langle\nabla f(x), u\rangle - \frac{\mu^2}{2}(L_{\nabla F}^2 + 2L_F^2 + \|F(x)\|^2)\|u\|^2$. To see this, we have

$$f(x+\mu u) - f(x) \geq \langle\nabla f(x), \mu u\rangle$$

by convexity. So it suffices to deduce

$$f(x) - f(x-\mu u) \geq \langle\nabla f(x), \mu u\rangle - \frac{\mu^2}{2}(L_{\nabla F}^2 + 2L_F^2 + \|F(x)\|^2)\|u\|^2 \tag{151}$$

By lemma 23, we have

$$|f(x-\mu u) - f(x) - \langle\nabla f(x), -\mu u\rangle| \leq \frac{\mu^2}{2}(L_{\nabla F}^2 + 2L_F^2 + \|F(x)\|^2)\|u\|^2$$

This implies

$$f(x) - f(x-\mu u) - \langle\nabla f(x), \mu u\rangle \geq -\frac{\mu^2}{2}(L_{\nabla F}^2 + 2L_F^2 + \|F(x)\|^2)\|u\|^2$$

This implies (151). Therefore, by Young's inequality and following Nesterov's proof for theorem 4.2 in [**?**], we have

$$\begin{aligned}
\mathbb{E}_u(\|\hat{g}_\mu(x)\|^2) &= \frac{1}{4\mu^2}\mathbb{E}_u([f(x+\mu u) - f(x-\mu u) - 2\mu\langle\nabla f(x), u\rangle + 2\mu\langle\nabla f(x), u\rangle]^2\|u\|^2) \\
&\leq \frac{1}{2\mu^2}\mathbb{E}_u([f(x+\mu u) - f(x-\mu u) - 2\mu\langle\nabla f(x), u\rangle]^2\|u\|^2 + 4\mu^2\langle\nabla f(x), u\rangle^2\|u\|^2) \\
&\leq \frac{1}{2\mu^2}\left[\mathbb{E}_u(\frac{\mu^4}{4}(L_{\nabla F}^2 + 2L_F^2 + \|F(x)\|^2)^2\|u\|^6) + 4\mu^2\mathbb{E}_u\{\langle\nabla f(x), u\rangle^2\|u\|^2\}\right] \\
&= \frac{\mu^2}{8}(L_{\nabla F}^2 + 2L_F^2 + \|F(x)\|^2)^2(n+6)^3 + 2(n+4)\|\nabla f(x)\|^2
\end{aligned}$$

□

6.3 July 22

Theorem 25. *By Theorem 24 and the proof of Theorem 14, we have the following results:*

$$\mathbb{E}_u\|g_\mu(x) - \nabla f(x)\| \leq \mu(L_{\nabla F}^2 + 2L_F^2 + \|F(x)\|^2)(n+6)^{3/2} + 2(\sqrt{n}+3)\|\nabla f(x)\| \tag{152}$$

$$\mathbb{E}_u\|\hat{g}_\mu(x) - \nabla f(x)\| \leq \frac{\mu}{2}(L_{\nabla F}^2 + 2L_F^2 + \|F(x)\|^2)(n+6)^{3/2} + 2(\sqrt{n}+3)\|\nabla f(x)\| \tag{153}$$

Lemma 26.

$$\|\nabla f(x)\| \leq 2\|F(x)\|\sum_{i=1}^{p}\|\nabla F_i(x)\| \tag{154}$$

Proof.

$$\begin{aligned}\|\nabla f(x)\| &= 2\|F(x)\nabla F(x)\| \leq 2\|F(x)\|\|\nabla F(x)\|_F \\ &= 2\|F(x)\|\sqrt{\sum_{i=1}^{p}\|\nabla F_i(x)\|^2} \\ &\leq 2\|F(x)\|\sum_{i=1}^{p}\|\nabla F_i(x)\|\end{aligned}$$

□

Expected Error Sampling

Recall from lemma 22 that $L_F = \sqrt{\sum_{i=1}^{p} L_{F_i}^2}, L_{\nabla F} = \sqrt{\sum_{i=1}^{p} L_{\nabla F_i}^2}$.

By theorem 25 and lemma 26, we have the expected error sampling from scheme 1 with Forward and Central difference are

Scheme 1 Forward Difference:

$$\mathbb{E}_u\|g_\mu(x) - \nabla f(x)\| \leq \mu(L_{\nabla F}^2 + 2L_F^2 + \|F(x)\|^2)(n+6)^{3/2} + 4(\sqrt{n}+3)\|F(x)\|\sum_{i=1}^{p}\|\nabla F_i(x)\| \quad (155)$$

Scheme 1 Central Difference:

$$\mathbb{E}_u\|\hat{g}_\mu(x) - \nabla f(x)\| \leq \frac{\mu}{2}(L_{\nabla F}^2 + 2L_F^2 + \|F(x)\|^2)(n+6)^{3/2} + 4(\sqrt{n}+3)\|F(x)\|\sum_{i=1}^{p}\|\nabla F_i(x)\| \quad (156)$$

Scheme 3 Forward Difference:

$$\mathbb{E}_u\|g_{\mu,3}(x) - \nabla f(x)\| \leq 2\|F(x)\|\sum_{i=1}^{p}\mu L_{\nabla F_i}(n+6)^{3/2} + 4(\sqrt{n}+3)\|F(x)\|\sum_{i=1}^{p}\|\nabla F_i(x)\| \quad (157)$$

Scheme 3 Central Difference:

$$\mathbb{E}_u\|\hat{g}_{\mu,3}(x) - \nabla f(x)\| \leq \|F(x)\|\sum_{i=1}^{p}\mu L_{\nabla F_i}(n+6)^{3/2} + 4(\sqrt{n}+3)\|F(x)\|\sum_{i=1}^{p}\|\nabla F_i(x)\| \quad (158)$$

Scheme 2 Forward Difference:

$$\begin{aligned}\mathbb{E}_u\|g_{\mu,2}(x) - \nabla f(x)\| \leq 2L_F\mu &\left[n + \sum_{i=1}^{p}\left\{\mu^2 L_{\nabla F_i}^2(n+6)^3 + 4(n+5)\|\nabla F_i(x)\|^2\right\}\right] \\ &+ 2\|F(x)\|\sum_{i=1}^{p}\left\{\mu L_{\nabla F_i}(n+6)^{3/2} + 2(\sqrt{n}+3)\|\nabla F_i(x)\|\right\} + 2L_F^2\mu\sqrt{n} \quad (159)\end{aligned}$$

Scheme 2 Central Difference:

$$\begin{aligned}\mathbb{E}_u\|\hat{g}_{\mu,2}(x) - \nabla f(x)\| \leq 2L_F\mu &\left[n + \sum_{i=1}^{p}\left\{\frac{\mu^2}{4} L_{\nabla F_i}^2(n+6)^3 + 4(n+5)\|\nabla F_i(x)\|^2\right\}\right] \\ &+ 2\|F(x)\|\sum_{i=1}^{p}\left\{\frac{\mu}{2} L_{\nabla F_i}(n+6)^{3/2} + 2(\sqrt{n}+3)\|\nabla F_i(x)\|\right\} + 2L_F^2\mu\sqrt{n} \quad (160)\end{aligned}$$

6.4 July 23

Tasks:

1. Work with equation (138); start with the term $\mathbb{E}_u\|g_\mu(x)\|^2$ and work out the algebra rather than proceeding with the first inequality).

2. Incorporate item 1 in the convergence analysis. In addition, think about the two approaches for the convergence analysis. First, the $\mathbb{E}_{u_k} r_{k+1}^2$ and second, $\mathbb{E}_{u_k} f(x_{k+1})$.

6.5 July 24

Lemma 27. *Given $v_1, \ldots, v_n \in \mathbb{R}^n$,*

$$\|v_1 + \cdots + v_n\|^2 \leq n \sum_{i=1}^{n} \|v_i\|^2 \tag{161}$$

Proof. Base case is immediate. By induction hypothesis, we assume equation (161) holds for all $n \geq 1$. For the inductive step,

$$\begin{aligned}
\|v_1 + \cdots + v_{n+1}\|^2 &= \|v_1 + \cdots + v_n\|^2 + \|v_{n+1}\|^2 + 2\langle v_{n+1}, v_1 + \cdots + v_n\rangle \\
&\leq n \sum_{i=1}^{n} \|v_i\|^2 + \|v_{n+1}\|^2 + 2\sum_{i=1}^{n} \langle v_{n+1}, v_i\rangle \\
&\leq n \sum_{i=1}^{n} \|v_i\|^2 + \|v_{n+1}\|^2 + \sum_{i=1}^{n} (\|v_i\|^2 + \|v_{n+1}\|^2) \\
&= (n+1) \sum_{i=1}^{n+1} \|v_i\|^2
\end{aligned}$$

where we used Cauchy Schwarz and Young's inequality. □

For notational convenience, $g_\mu(x) = 2F(x)\left[\frac{F(x+\mu u)-F(x)}{\mu}\right] u = 2\sum_{i=1}^{p} F_i(x)\left[\frac{F_i(x+\mu u)-F_i(x)}{\mu}\right] u$. By lemma 161 and Theorem 4 in [**?**], we have

$$\begin{aligned}
\mathbb{E}_u\|g_\mu(x)\|^2 &\leq 4p \sum_{i=1}^{p} F_i(x)^2 \mathbb{E}_u \left\| \frac{F_i(x+\mu u) - F_i(x)}{\mu} u \right\|^2 \\
&\leq 4p \sum_{i=1}^{p} F_i(x)^2 \left[\frac{\mu^2}{2} L_{\nabla F_i}^2 (n+6)^3 + 2(n+4)\|\nabla F_i(x)\|^2 \right]
\end{aligned} \tag{162}$$

By a computation on July 20 (see section 6.1), recall that $r_k := \|x_k - x^*\|$, $x_{k+1} = x_k - h g_\mu(x_k)$. Taking expectation with respect to u_k and using equation (162), we have

$$\begin{aligned}
\mathbb{E}_{u_k}(r_{k+1}^2) &= r_k^2 - 2h\mathbb{E}_{u_k}\langle g_\mu(x_k), x_k - x^*\rangle + h^2 \mathbb{E}_{u_k}\|g_\mu(x_k)\|^2 \\
&\leq r_k^2 - 2h\mathbb{E}_{u_k}\langle g_\mu(x_k), x_k - x^*\rangle + 4h^2 p \sum_{i=1}^{p} F_i(x_k)^2 \left[\frac{\mu^2}{2} L_{\nabla F_i}^2 (n+6)^3 + 2(n+4)\|\nabla F_i(x_k)\|^2 \right]
\end{aligned}$$

Lemma 28.

$$\mathbb{E}_{u_k}\langle g_\mu(x_k), x_k - x^*\rangle = \sum_{j=1}^{p} 2F_j(x_k)\langle \nabla F_{j,\mu}(x_k), x_k - x^*\rangle$$

Proof.

$$\begin{aligned}
\mathbb{E}_{u_k}\langle g_\mu(x_k), x_k - x^*\rangle &= \sum_{i=1}^{n}\mathbb{E}_{u_k}\sum_{j=1}^{p} 2F_j(x_k)\left[\frac{F_j(x_k+\mu u_k) - F_j(x_k)}{\mu}\right]u_i(x_{k,i} - x_i^*) \\
&= \sum_{i=1}^{n}\sum_{j=1}^{p} 2F_j(x_k)\nabla F_{j,\mu}(x_k)_i(x_{k,i} - x_i^*) \\
&= \sum_{j=1}^{p}\sum_{i=1}^{n} 2F_j(x_k)\nabla F_{j,\mu}(x_k)_i(x_{k,i} - x_i^*) \\
&= \sum_{j=1}^{p} 2F_j(x_k)\langle \nabla F_{j,\mu}(x_k), x_k - x^*\rangle
\end{aligned}$$

Equivalently, we have

$$\mathbb{E}_{u_k}\langle g_\mu(x_k), x_k - x^*\rangle = 2\langle F(x_k), \nabla F_\mu(x_k)(x_k - x^*)\rangle$$

□

Facts: A standard result on upper bounding the gradient squared shows

$$\|\nabla F_i(x_k)\|^2 \le 2L_{\nabla F_i}[F_i(x_k) - F_i(x_i^*)] \tag{163}$$

where x_i^* minmizes F_i. For simplicity, denote $F_i^* := F_i(x_i^*)$.

7 Week 7: July 27

7.1 July 27

Theorem 29.

$$\begin{aligned}
\mathbb{E}_{u_k}[r_{k+1}^2] \le r_k^2 &+ 2h\|F(x_k)\|\|x_k - x^*\|\mu(n+3)^{3/2}L_{\nabla F} + 2h\nabla f(x_k)(x^* - x_k) \\
&+ 4h^2 p\sum_{i=1}^{p} F_i(x_k)^2\left[\frac{\mu^2}{2}L_{\nabla F_i}^2(n+6)^3 + 4(n+4)L_{\nabla F_i}(F_i(x_k) - F_i^*)\right]
\end{aligned}$$

Proof.

$$\begin{aligned}
\mathbb{E}_{u_k}[r_{k+1}^2] &\le r_k^2 + 2h\mathbb{E}_{u_k}\langle g_\mu(x_k), x^* - x_k\rangle + h^2\mathbb{E}_{u_k}\|g_\mu(x_k)\|^2 \\
&= r_k^2 + 2h\langle 2F(x_k), \nabla F_\mu(x_k)(x^* - x_k)\rangle + h^2\mathbb{E}_{u_k}\|g_\mu(x_k)\|^2 \\
&= r_k^2 + 2h\langle 2F(x_k), [\nabla F_\mu(x_k) - \nabla F(x_k)](x^* - x_k)\rangle + 2h\langle 2F(x_k), \nabla F(x_k)(x^* - x_k)\rangle + h^2\mathbb{E}_{u_k}\|g_\mu(x_k)\|^2 \\
&= r_k^2 + 2h\langle 2F(x_k), [\nabla F_\mu(x_k) - \nabla F(x_k)](x^* - x_k)\rangle + 2h\nabla f(x_k)(x^* - x_k) + h^2\mathbb{E}_{u_k}\|g_\mu(x_k)\|^2 \\
&\le r_k^2 + 4h\|F(x_k)\|\|\nabla F_\mu(x_k) - \nabla F(x_k)\|\|x^* - x_k\| + 2h\nabla f(x_k)(x^* - x_k) + h^2\mathbb{E}_{u_k}\|g_\mu(x_k)\|^2 \\
&\le r_k^2 + 2h\|F(x_k)\|\|x_k - x^*\|\mu(n+3)^{3/2}L_{\nabla F} + 2h\nabla f(x_k)(x^* - x_k)
\end{aligned}$$

$$+ 4h^2 p \sum_{i=1}^{p} F_i(x_k)^2 \left[\frac{\mu^2}{2} L_{\nabla F_i}^2 (n+6)^3 + 4(n+4) L_{\nabla F_i} [F_i(x_k) - F_i^*] \right]$$

where these inequalities follow from lemma 28, equation (162), equation (163), and theorem 13. □

Second approach to convergence analysis from Week 6: July 20

$$f(x_{k+1}) \leq f(x_k) - h\langle \nabla f(x_k), g_\mu(x_k) \rangle + \frac{L_{\nabla f}}{2} h^2 \|g_\mu(x_k)\|^2 \tag{164}$$

Taking the expectation of equation (164), we have

$$\begin{aligned}
\mathbb{E}_{u_k} f(x_{k+1}) &\leq f(x_k) - h\mathbb{E}_{u_k} \langle \nabla f(x_k), g_\mu(x_k) \rangle + \frac{L_{\nabla f}}{2} h^2 \mathbb{E}_{u_k} \|g_\mu(x_k)\|^2 \\
&\leq f(x_k) - 2h \nabla f(x_k) F(x_k) \nabla F_\mu(x_k) \\
&+ \frac{4pL_{\nabla f}}{2} h^2 \sum_{i=1}^{p} F_i(x_k)^2 \left[\frac{\mu^2}{2} L_{\nabla F_i}^2 (n+6)^3 + 4(n+4) L_{\nabla F_i} (F_i(x_k) - F_i^*) \right]
\end{aligned}$$

where the inequality follows from a computation analogous to lemma 28 and inequality 162.

Computation based on [**?**]

Lemma 30. *Define* $e(x) = g_\mu(x) - \nabla f(x)$, *where* $g_\mu(x) = 2F(x)\frac{F(x+\mu u) - F(x)}{\mu} u$ *using iterative update* $x_{k+1} = x_k - h g_\mu(x_k)$.

*As done in [***?** *], define* $\mathcal{U}_k = (u_0, \ldots, u_k)$. *Assume* f *is strongly convex with parameter* $\lambda > 0$. *Assuming* $h \leq \frac{1}{L_{\nabla f}}$, *then*

$$\mathbb{E}_{\mathcal{U}_k}[f(x_{k+1}) - f^*] \leq \left[\sum_{\ell=0}^{k} (1 - \lambda h)^{k-\ell} \mathbb{E}_{\mathcal{U}_\ell} \|e(x_\ell)\|^2 \right] + (1 - \lambda h)^{k+1} [f(x_0) - f(x^*)] \tag{165}$$

Proof. Ignore the previous equations. Don't need. □

7.2 July 28

Let $x^+ = x - h g_\mu(x)$. Assume the setting in lemma 30. Then

$$\begin{aligned}
f(x^+) &\leq f(x) + \langle \nabla f(x), x^+ - x \rangle + \frac{L_{\nabla f}}{2} \|x^+ - x\|^2 \\
&= f(x) - h\langle \nabla f(x), g_\mu(x) \rangle + \frac{L_{\nabla f}}{2} h^2 \|g_\mu(x)\|^2 \\
&= f(x) - h \nabla f(x)^\top (\nabla f(x) + e(x)) + \frac{L_{\nabla f}}{2} h^2 \|g_\mu(x)\|^2 \\
&= f(x) - h\|\nabla f(x)\|^2 - h \nabla f(x)^\top e(x) + \frac{L_{\nabla f}}{2} h^2 \|\nabla f(x) + e(x)\|^2 \\
&= f(x) - h\|\nabla f(x)\|^2 - h \nabla f(x)^\top e(x) + \frac{L_{\nabla f}}{2} h^2 \|\nabla f(x)\|^2 + \frac{L_{\nabla f}}{2} h^2 \|e(x)\|^2 + L_{\nabla f} h^2 \nabla f(x)^\top e(x) \\
&= f(x) - h(1 - \frac{hL_{\nabla f}}{2}) \|\nabla f(x)\|^2 + \frac{L_{\nabla f} h^2}{2} \|e(x)\|^2 + h(hL_{\nabla f} - 1) \nabla f(x)^\top e(x)
\end{aligned}$$

$$\leq f(x) - h(1 - \frac{hL_{\nabla f}}{2})\|\nabla f(x)\|^2 + \frac{L_{\nabla f}h^2}{2}\|e(x)\|^2 + h(1 - hL_{\nabla f})\|\nabla f(x)\|\|e(x)\|$$
$$\leq f(x) - h(1 - \frac{hL_{\nabla f}}{2})\|\nabla f(x)\|^2 + \frac{L_{\nabla f}h^2}{2}\|e(x)\|^2 + h(1 - hL_{\nabla f})\left[\frac{\|\nabla f(x)\|^2}{2} + \frac{\|e(x)\|^2}{2}\right]$$

where we utilize $hL_{\nabla f} \leq 1$. Simplifying this expression, we have

$$f(x^+) \leq f(x) - \frac{h}{2}\|\nabla f(x)\|^2 + \frac{h}{2}\|e(x)\|^2$$

Since f is λ-strongly convex, we have

$$\|\nabla f(x)\|^2 \geq 2\lambda(f(x) - f^*)$$

Together, this yields

$$f(x^+) \leq f(x) - h\lambda(f(x) - f^*) + \frac{h}{2}\|e(x)\|^2$$

This implies that

$$f(x^+) - f^* \leq (1 - h\lambda)(f(x) - f^*) + \frac{h}{2}\|e(x)\|^2 \tag{166}$$

Taking the expectation with respect to u, we have

$$\mathbb{E}_u[f(x^+) - f^*] \leq (1 - \lambda h)(f(x) - f^*) + \frac{h}{2}\mathbb{E}_u\|e(x)\|^2 \tag{167}$$

By lemma 18, the above becomes

$$\mathbb{E}_u[f(x^+) - f^*] \leq (1 - \lambda h)(f(x) - f^*) + 2hf(x)\sum_{i=1}^{p}(\mu^2 L_{\nabla F_i}^2(n+6)^3 + 4(n+5)\|\nabla F_i(x)\|^2) \tag{168}$$

Using $\|\nabla F_i(x)\|^2 \leq 2L_{\nabla F_i}[F_i(x) - F_i^*]$, the above becomes

$$\mathbb{E}_u[f(x^+) - f^*] \leq (1 - \lambda h)(f(x) - f^*) + 2hf(x)\sum_{i=1}^{p}(\mu^2 L_{\nabla F_i}^2(n+6)^3 + 8(n+5)L_{\nabla F_i}[F_i(x) - F_i^*]) \tag{169}$$

Using the scheme $x_{k+1} = x_k - hg_\mu(x_k)$ and inserting this in the previous equation, we have

$$\mathbb{E}_{\mathcal{U}_k}[f(x_{k+1}) - f^*] \leq (1 - \lambda h)\mathbb{E}_{\mathcal{U}_{k-1}}[(f(x_k) - f^*)] + 2hf(x_k)\sum_{i=1}^{p}(\mu^2 L_{\nabla F_i}^2(n+6)^3 + 8(n+5)L_{\nabla F_i}[F_i(x_k) - F_i^*])$$
$$\leq (1 - \lambda h)^{k+1}(f(x_0) - f^*) + 2h\sum_{\ell=0}^{k}(1 - \lambda h)^{k-\ell}f(x_\ell)\sum_{i=1}^{p}(\mu^2 L_{\nabla F_i}^2(n+6)^3 + 8(n+5)L_{\nabla F_i}[F_i(x_\ell) - F_i^*] \tag{170}$$

7.3 July 29

By definition of $\mathcal{U}_k$ (see lemma 30) and equation (170), we have

$$\mathbb{E}_{\mathcal{U}_k}[f(x_{k+1}) - f^*] \leq (1-\lambda h)^{k+1}(f(x_0) - f^*) \\ + 2h\sum_{\ell=0}^{k}(1-\lambda h)^{k-\ell}\mathbb{E}_{\mathcal{U}_\ell}\left\{f(x_\ell)\sum_{i=1}^{p}(\mu^2 L_{\nabla F_i}^2(n+6)^3 + 8(n+5)L_{\nabla F_i}[F_i(x_\ell) - F_i^*])\right\} \tag{171}$$

Assumption : Assume the scheme as stated in lemma 30 is so that $F_i(x_k)$ is bounded for each $1 \leq i \leq p$. In other words, for each $1 \leq i \leq p$, there exists $M_i \geq 0$ such that

$$|F_i(x_k)| \leq M_i \quad \text{for all } k \geq 0 \tag{172}$$

Lemma 31. *From assumption 7.3, it follows that*

$$|F_i(x_k) - F_i^*| \leq M_i - F_i^* \tag{173}$$

where defintion of $\mathcal{U}_k$ is defined in see lemma 30.

Lemma 32. *Define* $M := \sqrt{\sum_{i=1}^{p} M_i^2}$. *From assumption 7.3, it follows that*

$$\|F(x_k)\|^2 \leq M^2 \tag{174}$$

Theorem 33. *Recall the definition of* $L_{\nabla F}$ *(see Lemma 22), assuming* $L_{\nabla F_i} = L_{\nabla F_j}$ *for all* $1 \leq i, j \leq p$, *and by assumption 7.3, we have*

$$\mathbb{E}_{\mathcal{U}_k}[f(x_{k+1}) - f^*] \leq (1-\lambda h)^{k+1}[f(x_0) - f^*] \\ + \frac{2M^2}{\lambda}[1-(1-\lambda h)^{k+1}]\left\{p\mu^2 L_{\nabla F}^2(n+6)^3 + 8p(n+5)L_{\nabla F}(M - \min_{1\leq i\leq p} F_i^*)\right\} \tag{175}$$

Proof. The conclusion follows from lemma 31, 32, hypothesis, and the fact

$$\sum_{\ell=0}^{k}(1-\lambda h)^{k-\ell} = \frac{1-(1-\lambda h)^{k+1}}{\lambda h}$$

□

Recall that $e(x) = g_\mu(x) - \nabla f(x)$, we compute $\mathbb{E}_u\|e(x)\|^2$ in two different ways and in two different settings: Forward and central difference. Note that g_μ refers to scheme 3 (see 4.0.3).

Theorem 34.

$$\mathbb{E}_u\|g_\mu(x) - \nabla f(x)\|^2 \leq 8f(x)\sum_{i=1}^{p}\left(\frac{\mu^2}{2}L_{\nabla F_i}^2(n+6)^3 + 2(n+5)\|\nabla F_i(x)\|^2\right) \tag{176}$$

$$\mathbb{E}_u\|\hat{g}_\mu(x) - \nabla f(x)\|^2 \leq 8f(x)\sum_{i=1}^{p}\left(\frac{\mu^2}{8}L_{\nabla F_i}^2(n+6)^3 + 2(n+5)\|\nabla F_i(x)\|^2\right) \tag{177}$$

Proof. By Theorem 4 in [**?**], we have

$$\begin{aligned}
\mathbb{E}_u\|g_\mu(x) - \nabla f(x)\|^2 &\leq 4f(x)\mathbb{E}_u \left\| \frac{F(x+\mu u) - F(x)}{\mu} u - \nabla F(x) \right\|_F^2 \\
&= 4f(x) \sum_{i=1}^{p} \mathbb{E}_u \left\| \frac{F_i(x+\mu u) - F_i(x)}{\mu} u - \nabla F_i(x) \right\|_2^2 \\
&\leq 8f(x) \sum_{i=1}^{p} \left(\mathbb{E}_u \left\| \frac{F_i(x+\mu u) - F_i(x)}{\mu} u \right\|^2 + \|\nabla F_i(x)\|^2 \right) \\
&\leq 8f(x) \sum_{i=1}^{p} \left(\frac{\mu^2}{2} L_{\nabla F_i}^2 (n+6)^3 + 2(n+5)\|\nabla F_i(x)\|^2 \right)
\end{aligned}$$

The proof for the central difference $\hat{g}_\mu$ bound follows from the same argument. □

Theorem 35.

$$\mathbb{E}_u\|g_\mu(x) - \nabla f(x)\|^2 \leq 8p \sum_{i=1}^{p} F_i(x)^2 \left\{ \frac{\mu^2}{2} L_{\nabla F_i}^2 (n+6)^3 + 2(n+5)\|\nabla F_i(x)\|^2 \right\} \tag{178}$$

$$\mathbb{E}_u\|\hat{g}_\mu(x) - \nabla f(x)\|^2 \leq 8p \sum_{i=1}^{p} F_i(x)^2 \left\{ \frac{\mu^2}{8} L_{\nabla F_i}^2 (n+6)^3 + 2(n+5)\|\nabla F_i(x)\|^2 \right\} \tag{179}$$

Proof. By theorem 4 in [**?**] and lemma 161, we have

$$\begin{aligned}
\mathbb{E}_u\|g_\mu(x) - \nabla f(x)\|^2 &= 4\mathbb{E}_u \left\| F(x) \left[\frac{F(x+\mu u) - F(x)}{\mu} u - \nabla F(x) \right] \right\|^2 \\
&= 4\mathbb{E}_u \left\| \sum_{i=1}^{p} F_i(x) \left[\frac{F_i(x+\mu u) - F_i(x)}{\mu} u - \nabla F_i(x) \right] \right\|^2 \\
&\leq 4p \sum_{i=1}^{p} \mathbb{E}_u F_i(x)^2 \left\| \frac{F_i(x+\mu u) - F_i(x)}{\mu} u - \nabla F_i(x) \right\|^2 \\
&\leq 8p \sum_{i=1}^{p} F_i(x)^2 \left\{ \mathbb{E}_u \left\| \frac{F_i(x+\mu u) - F_i(x)}{\mu} u \right\|^2 + \|\nabla F_i(x)\|^2 \right\} \\
&\leq 8p \sum_{i=1}^{p} F_i(x)^2 \left\{ \frac{\mu^2}{2} L_{\nabla F_i}^2 (n+6)^3 + 2(n+5)\|\nabla F_i(x)\|^2 \right\}
\end{aligned}$$

The proof for the central difference $\hat{g}_\mu$ bound follows from the same argument. □

7.4 July 30

Thoughts for today: Still examining the two approaches for convergence rate: 1. $\mathbb{E}[f(x_k) - f^*]$ and 2. $\mathbb{E}\|x_k - x^*\|$.

7.5 July 31

Rough computation for $\mathbb{E}\|x_k - x^*\|$ approach.

$$
\begin{aligned}
\mathbb{E}_{u_k} r_{k+1}^2 &\le r_k^2 + 2h\langle 2F(x_k), \nabla F_\mu(x_k)(x^* - x_k)\rangle + h^2 \mathbb{E}_k \|g_\mu(x_k)\|^2 \\
&= r_k^2 + 2h\langle 2F(x_k), \nabla F(x_k)(x^* - x_k)\rangle + 2h\langle 2F(x_k), (\nabla F_\mu(x_k) - \nabla F(x_k))(x^* - x_k)\rangle + h^2 \mathbb{E}_{u_k} \|g_\mu(x_k)\|^2 \\
&= r_k^2 + 2h\nabla f(x_k)(x_k - x^*) + 2h\langle 2F(x_k), (\nabla F_\mu(x_k) - \nabla F(x_k))(x^* - x_k)\rangle + h^2 \mathbb{E}_{u_k} \|g_\mu(x_k)\|^2 \\
&= r_k^2 + h\|\nabla f(x_k)\|^2 + h r_k^2 + h[4f(x_k) + \|F_\mu(x_k) - \nabla F(x_k)\|^2 r_k^2] + h^2 \mathbb{E}_{u_k} \|g_\mu(x_k)\|^2 \\
&\le r_k^2 + h\|\nabla f(x_k)\|^2 + h r_k^2 + h[4f(x_k) + \frac{\mu^2}{4}(n+3)^3 L_{\nabla F}^2 r_k^2] + h^2 \mathbb{E}_{u_k} \|g_\mu(x_k)\|^2
\end{aligned}
$$

8 Week 8: August 3

8.1 August 3

We revisit the computation of $\mathbb{E}_u \|g_\mu(x)\|^2$. First, we prove a lemma.

Lemma 36. *Assume F has gradient Lipschitz. Then*

$$
\|F(y) - F(x) - \nabla F(x)(y - x)\| \le \frac{L_{\nabla F}}{2}\|y - x\|^2 \tag{180}
$$

Proof.

$$
\begin{aligned}
\|F(y) - F(x) - \nabla F(x)(y - x)\| &= \left\| \int_0^1 [\nabla F(x + t(y - x)) - F(x)](y - x) dt \right\| \\
&\le \int_0^1 \|\nabla F(x + t(y - x)) - \nabla F(x)\| \|x - y\| dt \\
&\le \frac{L_{\nabla F}}{2}\|y - x\|^2
\end{aligned}
$$

□

Theorem 37.

$$
\mathbb{E}_u \|g_\mu(x)\|^2 \le 2\mu^2 f(x) L_{\nabla F}^2 (n+6)^3 + 2(n+4)\|\nabla f(x)\|^2 \tag{181}
$$

Proof. By lemma 36, we have

$$
\begin{aligned}
[2F(x)\{F(x + \mu u) - F(x)\}]^2 &= [2F(x)\{(F(x + \mu u) - F(x) - \mu \nabla F(x)u) + \mu \nabla F(x)u\}]^2 \\
&\le 2(2F(x)\{F(x + \mu u - F(x) - \mu \nabla F(x)u\})^2 + 2(2F(x)\mu \nabla F(x)u)^2 \\
&\le 2(2\|F(x)\| \|F(x + \mu u) - F(x) - \mu \nabla F(x)u\|)^2 + 2\mu^2 \langle \nabla f(x), u\rangle^2 \\
&\le 8f(x)\frac{L_{\nabla F}^2}{4}\mu^4 \|u\|^4 + 2\mu^2 \langle \nabla f(x), u\rangle^2 \\
&= 2f(x) L_{\nabla F}^2 \mu^4 \|u\|^4 + 2\mu^2 \langle \nabla f(x), u\rangle^2
\end{aligned}
$$

By definition of $g_\mu(x)$, and following the proof of Theorem 4 in [**?**], we have

$$
\begin{aligned}
\mathbb{E}\|g_\mu(x)\|^2 &\le 2f(x) L_{\nabla F}^2 \mu^2 \mathbb{E}_u \|u\|^6 + 2\mathbb{E}_u[\langle \nabla f(x), u\rangle^2 \|u\|^2] \\
&\le 2\mu^2 f(x) L_{\nabla F}^2 (n+6)^3 + 2(n+4)\|\nabla f(x)\|^2
\end{aligned}
$$

□

Corollary 1 (Needs Revision. Bounds may not be tight.)**.** *As a result to theorem 37, we have*

$$\mathbb{E}_u\|g_\mu(x) - \nabla f(x)\|^2 \leq 4\mu^2 f(x) L_{\nabla F}^2 (n+6)^3 + 4(n+4.5)\|\nabla f(x)\|^2 \tag{182}$$

Furthermore, if we assume that f has gradient Lipschitz and is convex, we have

$$\mathbb{E}_u\|g_\mu(x) - \nabla f(x)\|^2 \leq 4\mu^2 f(x) L_{\nabla F}^2 (n+6)^3 + 8L_{\nabla f}(n+4.5)(f(x) - f^*) \tag{183}$$

Proof. By theorem 37 and Young's inequality,

$$\begin{aligned}
\mathbb{E}_u\|g_\mu(x) - \nabla f(x)\|^2 &\leq 2\mathbb{E}_u\|g_\mu(x)\|^2 + 2\|\nabla f(x)\|^2 \\
&\leq 4\mu^2 f(x) L_{\nabla F}^2 (n+6)^3 + 4(n+4)\|\nabla f(x)\|^2 + 2\|\nabla f(x)\|^2 \\
&= 4\mu^2 f(x) L_{\nabla F}^2 (n+6)^3 + 4(n+4.5)\|\nabla f(x)\|^2
\end{aligned}$$

Using this, the second equation follows from $\|\nabla f(x)\|^2 \leq 2L_{\nabla f}(f(x) - f^*)$ □

Theorem 38. *Recall the scheme $x_{k+1} = x_k - hg_\mu(x_k)$ where $g_\mu(x) = 2F(x)\frac{F(x+\mu u)-F(x)}{\mu}u$. Pick $x_0 \in \mathbb{R}^n$. Assume that $f^* = 0$ and $\lambda > 4L_{\nabla f}(n+4.5)$ (where $\lambda > 0$ is the strongly-convex parameter of f). Pick*

$$\mu < \sqrt{\frac{\lambda - 4L_{\nabla f}(n+4.5)}{2L_{\nabla F}^2(n+6)^3}} \tag{184}$$

and

$$h < \min\left\{\frac{1}{\lambda}, \frac{1}{L_{\nabla f}}\right\} \tag{185}$$

We have that

$$\mathbb{E}_{\mathcal{U}_k}[f(x_{k+1}) - f^*] \leq (1 + h[2\mu^2 L_{\nabla F}^2 (n+6)^3 + 4L_{\nabla f}(n+4.5) - \lambda])^{k+1}[f(x_0) - f^*] \tag{186}$$

where the definition of $\mathcal{U}_k$ is stated in lemma 30.

Proof. By equation (167), Corollary 1, and hypothesis, we have

$$\begin{aligned}
\mathbb{E}_{u_k}[f(x_{k+1}) - f^*] &\leq (1 - \lambda h)(f(x_k) - f^*) + \frac{h}{2}\mathbb{E}_{u_k}\|g_\mu(x_k) - \nabla f(x_k)\|^2 \\
&\leq (1 - \lambda h)(f(x_k) - f^*) + \frac{h}{2}[4\mu^2 f(x_k) L_{\nabla F}^2 (n+6)^3 + 8L_{\nabla f}(n+4.5)(f(x_k) - f^*)] \\
&= [(1 - \lambda h) + 2h\mu^2 L_{\nabla F}^2 (n+6)^3 + 4hL_{\nabla f}(n+4.5)](f(x_k) - f^*)
\end{aligned}$$

Observe that the quantity $(1 - \lambda h) + 2h\mu^2 L_{\nabla F}^2 (n+6)^3 + 4hL_{\nabla f}(n+4.5) \in (0, 1)$. Therefore by iteratively apply the above inequality and taking the expectation with respect to $\mathcal{U}_k$, we have our conclusion.

Problem: $\lambda > 4L_{\nabla f}(n+4.5)$ can't happen since $L_{\nabla f} \geq \lambda$.

Next step: Try adapt the proof of Theorem 8 in [**?**] using our bounds $\mathbb{E}_u\|g_\mu(x)\|^2$ established in theorem 37.

□

8.2 August 4

We revisit corollary 1 and examine if we can make the bounds (more) tight.

An aspect to make this bound tight is to examine the following step in the proof:

$$\begin{aligned}\mathbb{E}_u\|g_\mu(x)-\nabla f(x)\|^2 &= \mathbb{E}_u\|g_\mu(x)\|^2+\|\nabla f(x)\|^2-2\mathbb{E}_u\langle g_\mu(x),\nabla f(x)\rangle\\ &\le 2\mathbb{E}_u\|g_\mu(x)\|^2+2\|\nabla f(x)\|^2\end{aligned}$$

Before applying Cauchy Schwartz, we have the term

$$\langle g_\mu(x),\nabla f(x)\rangle$$

Recall that $\mathbb{E}_u g_\mu(x)=\nabla f_\mu(x)$ in the unstructured case (setting in [**?**] and $g_\mu(x)=[f(x+\mu u)-f(x)]\frac{u}{\mu}$), we have

$$\begin{aligned}\mathbb{E}_u\langle g_\mu(x),\nabla f(x)\rangle &= \langle\nabla f_\mu(x),\nabla f(x)\rangle = \|\nabla f(x)\|^2+\langle\nabla f_\mu(x)-\nabla f(x),\nabla f(x)\rangle\\ &= \|\nabla f(x)\|^2+\|\nabla f(x)\|\|\nabla f_\mu(x)-\nabla f(x)\|\cos\theta(x)\\ &\le \|\nabla f(x)\|^2+\|\nabla f(x)\|\frac{\mu}{2}L_{\nabla f}(n+3)^{3/2}\end{aligned}\tag{187}$$

where $\theta(x)$ denotes the angle between $\nabla f_\mu(x)-\nabla f(x)$ and $\nabla f(x)$ and used Lemma 3 in [**?**].

Returning to our structured setting with $g_\mu(x)=2F(x)\frac{F(x+\mu u)-F(x)}{\mu}u$.

Observe that $\mathbb{E}_u\left[\frac{F(x+\mu u)-F(x)}{\mu}\right]u_i=\frac{\partial}{\partial x_i}F_\mu(x)$ where u_i is the i-th component of u. Therefore, $\mathbb{E}_u[g_\mu(x)]=2F(x)\nabla F_\mu(x)$. Using this, we arrive at

$$\begin{aligned}\mathbb{E}_u\langle\nabla f(x),g_\mu(x)\rangle &= \langle\nabla f(x),2F(x)\nabla F_\mu(x)\rangle\\ &= \langle\nabla f(x),2F(x)\nabla F(x)\rangle+\langle\nabla f(x),2F(x)[\nabla F_\mu(x)-\nabla F(x)]\rangle+\langle\nabla f(x),2F(x)[\nabla F_\mu(x)-\nabla F(x)]\rangle\\ &= \|\nabla f(x)\|^2+\langle\nabla f(x),2F(x)[\nabla F_\mu(x)-\nabla F(x)]\rangle\\ &= \|\nabla f(x)\|^2+2\|\nabla f(x)\|\|F(x)\|\|\nabla F_\mu(x)-\nabla F(x)\|\cos(\theta(x))\\ &\le \|\nabla f(x)\|^2+2\|\nabla f(x)\|\sqrt{f(x)}\frac{\mu}{2}(n+3)^{3/2}L_{\nabla F}\end{aligned}\tag{188}$$

where $\theta(x)$ is the angle betweeen $\nabla f(x)$ and $2F(x)[\nabla F_\mu(x)-\nabla F(x)]$ and where we used theorem 13 to upperbound the quantity $\|\nabla F_\mu(x)-\nabla F(x)\|$.

Question: Is $\nabla f_\mu(x)=\mathbb{E}_u g_\mu(x)$? Computing, we have $\mathbb{E}_u g_\mu(x)=2F(x)\nabla F_\mu(x)$.

$$\begin{aligned}\mathbb{E}_u\|g_\mu(x)-\nabla f(x)\|^2 &= \mathbb{E}_u\|g_\mu(x)-\mathbb{E}_u g_\mu(x)+\mathbb{E}_u g_\mu(x)-\nabla f(x)\|^2\\ &= \mathbb{E}_u\|g_\mu(x)-\mathbb{E}_u g_\mu(x)\|^2+\mathbb{E}_u\|\mathbb{E}_u g_\mu(x)-\nabla f(x)\|^2+2\mathbb{E}_u\langle g_\mu(x)-\mathbb{E}_u g_\mu(x),\mathbb{E}_u g_\mu(x)-\nabla f(x)\rangle\\ &= \mathbb{E}_u\|g_\mu(x)-2F(x)\nabla F_\mu(x)\|^2+\mathbb{E}_u\|2F(x)[\nabla F_\mu(x)-\nabla f(x)]\|^2\\ &= \mathbb{E}_u\|2F(x)\left[\frac{F(x+\mu u)-F(x)}{\mu}u-\nabla F_\mu(x)\right]\|^2+\mathbb{E}_u\|2F(x)\left[\nabla F_\mu(x)-\nabla F(x)\right]\|^2\\ &\le 4f(x)\mathbb{E}_u\left\|\frac{F(x+\mu u)-F(x)}{\mu}u-\nabla F_\mu(x)\right\|^2+4f(x)\mathbb{E}_u\|\nabla F_\mu(x)-\nabla F(x)\|^2\end{aligned}$$

$$\leq 4f(x)\mathbb{E}_u \left\| \frac{F(x+\mu u)-F(x)}{\mu} u - \nabla F_\mu(x) \right\|^2 + 4f(x)\frac{\mu^2}{4}(n+3)^3 L_{\nabla F}$$
$$= 4f(x) \left[\mathbb{E}_u \left\| \frac{F(x+\mu u)-F(x)}{\mu} u - \nabla F_\mu(x) \right\|^2 + \frac{\mu^2}{4}(n+3)^3 L_{\nabla F} \right]$$

where we used $\mathbb{E}_u[g_\mu(x) - \mathbb{E}_u g_\mu(x)] = 0$ and theorem 13.

8.3 August 5

Following up on the question from August 4 (see 8.2) on whether $\nabla f_\mu(x) = \mathbb{E}_u g_\mu(x)$ where $g_\mu(x) = 2F(x)\frac{F(x+\mu u)-F(x)}{\mu} u$, the answer when F is linear, it is yes. But in general, it isn't true. Here's a counterexample.

Counterexample: Let $h(x) = x^2$ and $F(x) = x^2$. Recall $u \sim \mathcal{N}(0,1)$ and $\mathbb{E}[u^4] = 3$. Then

$$\begin{aligned} f_\mu(x) &= \mathbb{E}_u[f(x+\mu u)] \\ &= \mathbb{E}_u[(x+\mu u)^4] = \mathbb{E}_u \sum_{i=0}^{4} \binom{4}{i} x^i (\mu u)^{4-i} \\ &= 3\mu^4 + 6x^2\mu^2 + x^4 \end{aligned}$$

Then $\nabla f_\mu(x) = 12x\mu^2 + 4x^3$. Next,

$$F_\mu(x) = \mathbb{E}_u[F(x+\mu u)] = \mathbb{E}_u[(x+\mu u)^2] = x^2 + \mu^2$$

Hence, $\nabla F_\mu(x) = 2x$, so $\mathbb{E}_u g_\mu(x) = 2F(x)\nabla F_\mu(x) = 2F(x)\nabla F(x) = \nabla f(x) \neq \nabla f_\mu(x)$.

Now, we show that if F is linear, then $\mathbb{E}_u[g_\mu(x)] = \nabla f_\mu(x)$. In other words, when $F(x) = Ax - b$, $\nabla f_\mu(x) = 2(Ax-b)^\top A$.

Now,

$$\begin{aligned} f_\mu(x) &= \mathbb{E}_u[f(x+\mu u)] \\ &= \mathbb{E}_u[\|(Ax-b) + \mu Au\|^2] \\ &= \mathbb{E}_u\left[\|Ax-b\|^2 + \mu^2\|Au\|^2 + 2\langle Ax-b, \mu Au\rangle\right] \\ &= \mu^2 \mathbb{E}_u\|Au\|^2 + \|Ax-b\|^2 \\ &= \mu^2\langle A, A\rangle + \|Ax-b\|^2 \end{aligned}$$

where $\mathbb{E}_u\|Au\|^2 = \mathbb{E}_u u^\top A^\top Au = \mathbb{E}_u \sum_{i,j}(A^\top A)_{ij} u_i u_j = \sum_{i=1}^n (A^\top A)_{ii} = \text{Trace}(A^\top A) =: \langle A, A\rangle$. Therefore,

$$\nabla f_\mu(x) = 2(Ax-b)^\top A$$

We compute $F_\mu(x)$:

$$F_\mu(x) = \mathbb{E}_u[A(x+\mu u) - b] = Ax - b$$

Therefore, $\nabla F_\mu(x) = A$. Hence, $2F(x)\nabla F_\mu(x) = 2(Ax-b)^\top A$. We conclude from these computation that $\mathbb{E}_u[g_\mu(x)] = \nabla f_\mu(x)$.

Task: Try some computation and observe the bounds stated in corollary 1 using $F(x) = Ax - b$.

8.4 August 6

We compute the first bound in Corollary 1 when $F(x) = Ax - b$. Observe that $\nabla F(x) = A$, so $L_{\nabla F} = 0$. Therefore, the bound from Corollary 1

$$\begin{aligned}\mathbb{E}_u\|g_\mu(x) - \nabla f(x)\|^2 &\leq 4\mu^2 f(x) L_{\nabla F}^2 (n+6)^3 + 4(n+4.5)\|\nabla f(x)\|^2 \\ &= 4(n+4.5)\|\nabla f(x)\|^2\end{aligned}$$

Recall that $g_\mu(x) = 2F(x)\frac{F(x+\mu u)-F(x)}{\mu}u = 2(Ax-b)Auu^\top$ and $\nabla f(x) = 2F(x)\nabla F(x) = 2(Ax-b)^\top A$. Therefore, the bound above becomes

$$\mathbb{E}_u\|g_\mu(x) - \nabla f(x)\|^2 = \mathbb{E}_u\|2(Ax-b)^\top A[uu^\top - I]\|^2 \leq 16(n+4.5)\|(Ax-b)^\top A\|^2$$

Therefore, the above bound is equivalent to

$$\mathbb{E}_u\|(Ax-b)^\top A[uu^\top - I]\|^2 \leq 4(n+4.5)\|(Ax-b)^\top A\|^2 \tag{189}$$

See the python code zerothOrderBound.py which checks if the bound (189) holds or not.

8.5 August 7

Recall from journal entry in subsection 8.3, we asked if $\mathbb{E}_u[g_\mu(x)] = \nabla f_\mu(x)$ where $g_\mu(x) = 2F(x)^\top \left\{\frac{F(x+\mu u)-F(x)}{\mu}\right\} u$ and $f = h \circ F$. We provide a difference computation that answers this question. We claim that

Theorem 39.

$$\mathbb{E}_u\left[\frac{\|F(x+\mu u) - F(x)\|^2}{\mu}u\right] = \nabla f_\mu(x) - \mathbb{E}_u[g_\mu(x)] \tag{190}$$

Proof.

$$\begin{aligned}\mathbb{E}_u\left[\frac{\|F(x+\mu u) - F(x)\|^2}{\mu}u\right] &= \mathbb{E}_u\left[\frac{\|F(x+\mu u)\|^2 - 2\langle F(x+\mu u), F(x)\rangle + \|F(x)\|^2}{\mu}u\right] \\ &= \mathbb{E}_u\left[\frac{\|F(x+\mu u)\|^2 - \|F(x)\|^2 + 2\|F(x)\|^2 - 2\langle F(x+\mu u), F(x)\rangle}{\mu}u\right] \\ &= \mathbb{E}_u\left[\frac{f(x+\mu u) - f(x) + 2F(x)^\top[F(x) - F(x+\mu u)]}{\mu}u\right] \\ &= \mathbb{E}_u\left[\frac{f(x+\mu u) - f(x)}{\mu}u - \frac{2F(x)^\top[F(x+\mu u) - F(x)]}{\mu}u\right] \\ &= \mathbb{E}_u\left[\frac{f(x+\mu u) - f(x)}{\mu}u\right] - \mathbb{E}_u\left[\frac{2F(x)^\top[F(x+\mu u) - F(x)]}{\mu}u\right] \\ &= \nabla f_\mu(x) - \mathbb{E}_u[g_\mu(x)]\end{aligned}$$

□

In view of theorem 39, we can observe that $g_\mu(x)$ is not an unbiased estimate of $\nabla f_\mu(x)$ in general.

Continuing from journal entry on August 4 (see 8.2), we want to examine the quantity

$$\mathbb{E}_u\|g_\mu(x) - \nabla f(x)\|^2 = \mathbb{E}_u\|g_\mu(x)\|^2 + \|\nabla f(x)\|^2 - 2\mathbb{E}_u\langle g_\mu(x), \nabla f(x)\rangle$$

in two settings: Structured and unstructued setting.

In the unstructured case: $g_\mu(x) = \frac{f(x+\mu u)-f(x)}{\mu}u$, we want to upperbound the term $-2\mathbb{E}_u\langle g_\mu(x), \nabla f(x)\rangle$. From equation (187), we have

$$-2\mathbb{E}_u\langle g_\mu(x), \nabla f(x)\rangle = -2\|\nabla f(x)\|^2 - 2\|\nabla f(x)\|\|\nabla f_\mu(x) - \nabla f(x)\|\cos(\theta) \tag{191}$$

where $\theta(x)$ is the angle between $\nabla f(x)$ and $\nabla f_\mu(x) - \nabla f(x)$.

In the structured case: $g_\mu(x) = 2F(x)^\top\left\{\frac{F(x+\mu u)-F(x)}{\mu}\right\}u$, we want to upperbound the term $-2\mathbb{E}_u\langle g_\mu(x), \nabla f(x)\rangle$. From equation (188), we have

$$-2\mathbb{E}_u\langle \nabla f(x), g_\mu(x)\rangle = -2\|\nabla f(x)\|^2 - 4\|\nabla f(x)\|\|F(x)\|\|\nabla F_\mu(x) - \nabla F(x)\|\cos\theta(x) \tag{192}$$

where $\theta(x)$ is the angle between $\nabla f(x)$ and $2F(x)^\top[\nabla F_\mu(x) - \nabla F(x)]$.

9 Week 9: August 10

9.1 August 10

We return to the convergence analysis done in Theorem 8 done in [**?**]. Recall that

$$r_k := \|x_k - x^*\|$$

where $x_{k+1} = x_k - hg_\mu(x_k)$ with

$$g_\mu(x) = 2F(x)^\top\left\{\frac{F(x+\mu u) - F(x)}{\mu}\right\}u$$

Following the proof in Theorem 8,

$$r_{k+1}^2 = r_k^2 - 2h\langle g_\mu(x_k), x_k - x^*\rangle + h^2\|g_\mu(x_k)\|^2 \tag{193}$$

Taking the expectation of (193), theorem 37, and lemma 28, and adding and subtracting ∇f_μ, we have

$$\begin{aligned}
\mathbb{E}_{u_k} r_{k+1}^2 &\le r_k^2 - 2h\mathbb{E}_{u_k}\langle g_\mu(x_k), x_k - x^*\rangle + h^2\mathbb{E}_{u_k}\|g_\mu(x_k)\|^2 \\
&= r_k^2 - 2h\langle 2F(x_k), \nabla F_\mu(x_k)(x_k - x^*)\rangle + h^2\mathbb{E}_{u_k}\|g_\mu(x_k)\|^2 \\
&= r_k^2 - 2h\mathbb{E}_{u_k}[g_\mu(x_k)](x_k - x^*) + h^2\mathbb{E}_{u_k}\|g_\mu(x_k)\|^2 \\
&= r_k^2 - 2h\nabla f_\mu(x_k)(x_k - x^*) - 2h[\mathbb{E}_{u_k}[g_\mu(x_k)] - \nabla f_\mu(x_k)](x_k - x^*) + h^2\mathbb{E}_{u_k}\|g_\mu(x_k)\|^2
\end{aligned}$$

$$
\begin{aligned}
&\leq r_k^2 - 2h(f(x_k) - f_\mu(x^*)) + h^2\left[2\mu^2 f(x_k)L_{\nabla F}^2(n+6)^3 + 2(n+4)\|\nabla f(x_k)\|^2\right] \\
&- 2h[\mathbb{E}_{u_k}[g_\mu(x_k)] - \nabla f_\mu(x_k)](x_k - x^*) \\
&\leq r_k^2 - 2h(f(x_k) - f_\mu(x^*)) + h^2\left[2\mu^2 f(x_k)L_{\nabla F}^2(n+6)^3 + 4(n+4)L_{\nabla f}(f(x_k) - f^*)\right] \\
&- 2h[\mathbb{E}_{u_k}[g_\mu(x_k)] - \nabla f_\mu(x_k)](x_k - x^*) \\
&\leq r_k^2 - 2h(f(x_k) - f^*) + \mu^2 nhL_{\nabla f} + h^2\left[2\mu^2(f(x_k) - f^*)L_{\nabla F}^2(n+6)^3 + 4(n+4)L_{\nabla f}(f(x_k) - f^*)\right] \\
&+ h^2[2\mu^2 f^* L_{\nabla F}^2(n+6)^3] - 2h[\mathbb{E}_{u_k}[g_\mu(x_k)] - \nabla f_\mu(x_k)](x_k - x^*) \qquad (194)
\end{aligned}
$$

where we used that f has gradient Lipschitz and Theorem 1 in [**?**].

Observe that if $g_\mu(x_k)$ is an unbiased estimate of $\nabla f_\mu(x_k)$, the last term in previous equation vanishes. In this case, we should be able to proceed to prove convergence as done in [**?**].

Remark: In the linear least squares, we see convergence.

Question: How do we deal with $g_\mu(x)$ being a biased estimate of $\nabla f_\mu(x)$?

NOTE: Finish convergence analysis in the unbiased setting for convex and strongly convex.

Task 1: look at the quantity $g_\mu(x) - \nabla f_\mu(x)$ and write out and see if I can bound above.

Task 2: SaSSy presentation.

Theorem 40. *For notational convenicence, define* $C(f) := \frac{4f^* L_{\nabla F}^2}{L_{\nabla f}^2}$. *Assume* $g_\mu(x) = 2F(x)^\top\left\{\frac{F(x+\mu u)-F(x)}{\mu}u\right\}$ *is an unbiased estimator of* $\nabla f_\mu(x)$. *Suppose we pick*

$$
h = \frac{1}{4\left(\frac{\mu^2}{2}L_{\nabla F}^2(n+6)^3 + (n+4)L_{\nabla f}\right)}
$$

Then the convergence rate is

$$
\frac{1}{N+1}\sum_{k=0}^{N}(\phi_k - f^*) \leq \frac{4(n+4)L_{\nabla f}\|x_0 - x^*\|^2}{N+1} + \frac{11\mu^2(n+4)^2 L_{\nabla f}\max(C(f),1)}{25} \qquad (195)
$$

where $\phi_k := \mathbb{E}_{\mathcal{U}_{k-1}}(f(x_k))$ *for* $k \geq 1$ *and* $\phi_0 := f(x_0)$ *and definition of* $\mathcal{U}_{k-1}$ *is in lemma 30.*

In addition, if f *is* λ*-strongly convex and* $\delta_\mu := \frac{22\mu^2(n+4)^2 L_{\nabla f}\max(C(f),1)}{25\lambda}$, *then the rate becomes*

$$
\phi_N - f^* \leq \frac{L_{\nabla f}}{2}\left[\delta_\mu + \left(1 - \frac{\lambda}{8(n+4)L_{\nabla f}}\right)^N(\|x_0 - x^*\|^2 - \delta_\mu)\right] \qquad (196)
$$

Proof. By hypothesis, equation (194) becomes

$$
\begin{aligned}
\mathbb{E}_{u_k} r_{k+1}^2 &\leq r_k^2 - 2h(f(x_k) - f^*)\left[1 - h\mu^2 L_{\nabla F}^2(n+6)^3 - 2h(n+4)L_{\nabla f}\right] + \mu^2 nhL_{\nabla f} + 2\mu^2 h^2 f^* L_{\nabla F}^2(n+6)^3 \\
&= r_k^2 - 2h(f(x_k) - f^*)\left[1 - 2h\left\{\frac{\mu^2}{2}L_{\nabla F}^2(n+6)^3 + (n+4)L_{\nabla f}\right\}\right] + \mu^2 nhL_{\nabla f} + 2\mu^2 h^2 f^* L_{\nabla F}^2(n+6)^3
\end{aligned}
$$

$$
\begin{aligned}
&= r_k^2 - \frac{f(x_k) - f^*}{4\left(\frac{\mu^2}{2}L_{\nabla F}^2(n+6)^3 + (n+4)L_{\nabla f}\right)} + \frac{\mu^2 n L_{\nabla f}}{4\left(\frac{\mu^2}{2}L_{\nabla F}^2(n+6)^3 + (n+4)L_{\nabla f}\right)} \\
&+ \frac{\mu^2 f^* L_{\nabla F}^2 (n+6)^3}{8\left(\frac{\mu^2}{2}L_{\nabla F}^2(n+6)^3 + (n+4)L_{\nabla f}\right)^2} \\
&= r_k^2 - \frac{f(x_k) - f^*}{4\left(\frac{\mu^2}{2}L_{\nabla F}^2(n+6)^3 + (n+4)L_{\nabla f}\right)} \\
&+ \frac{\mu^2}{4}\left[\frac{nL_{\nabla f}}{\frac{\mu^2}{2}L_{\nabla F}^2(n+6)^3 + (n+4)L_{\nabla f}} + \frac{f^* L_{\nabla F}^2 (n+6)^3}{2\left(\frac{\mu^2}{2}L_{\nabla F}^2(n+6)^3 + (n+4)L_{\nabla f}\right)^2}\right] \\
&\leq r_k^2 - \frac{f(x_k) - f^*}{4(n+4)L_{\nabla f}} + \frac{\mu^2}{4}\left[\frac{n}{n+4} + \frac{f^* L_{\nabla F}^2 (n+6)^3}{2(n+4)^2 L_{\nabla f}^2}\right] \\
&= r_k^2 - \frac{f(x_k) - f^*}{4(n+4)L_{\nabla f}} + \frac{\mu^2}{4}\left[\frac{n}{n+4} + C(f)\frac{(n+6)^3}{8(n+4)^2}\right] \\
&\leq r_k^2 - \frac{f(x_k) - f^*}{4(n+4)L_{\nabla f}} + \frac{\mu^2 \max(C(f),1)}{4}\left[\frac{n}{n+4} + \frac{(n+6)^3}{8(n+4)^2}\right] \\
&\leq r_k^2 - \frac{f(x_k) - f^*}{4(n+4)L_{\nabla f}} + \frac{11\mu^2(n+4)\max(C(f),1)}{100}
\end{aligned} \tag{197}
$$

where we used

$$
\frac{n}{n+4} + \frac{(n+6)^3}{8(n+4)^2} \leq \frac{11(n+4)}{25} \quad \text{for } n \geq 1
$$

Taking the expectation in $\mathcal{U}_{k-1}$, we obtain

$$
\rho_{k+1} := \mathbb{E}_{\mathcal{U}_k}(r_{k+1}^2) \leq \rho_k - \frac{\phi_k - f^*}{4(n+4)L_{\nabla f}} + \frac{11\mu^2(n+4)\max(C(f),1)}{100}
$$

Summing up these inequalities for $k = 0, \ldots, N$, we have

$$
\begin{aligned}
0 \leq \rho_N &\leq \rho_0 - \sum_{k=0}^{N} \frac{\phi_k - f^*}{4(n+4)L_{\nabla f}} + (N+1)\frac{11\mu^2(n+4)\max(C(f),1)}{100} \\
&= \|x_0 - x^*\|^2 - \sum_{k=0}^{N} \frac{\phi_k - f^*}{4(n+4)L_{\nabla f}} + (N+1)\frac{11\mu^2(n+4)\max(C(f),1)}{100}
\end{aligned}
$$

rearranging yields

$$
\sum_{k=0}^{N} \frac{\phi_k - f^*}{4(n+4)L_{\nabla f}} \leq \|x_0 - x^*\|^2 + (N+1)\frac{11\mu^2(n+4)\max(C(f),1)}{100}
$$

dividing the result by $N+1$, and multiplying across by $4(n+4)L_{\nabla f}$,

$$
\frac{1}{N+1}\sum_{k=0}^{N}(\phi_k - f^*) \leq \frac{4(n+4)L_{\nabla f}\|x_0 - x^*\|^2}{N+1} + \frac{11\mu^2(n+4)^2 L_{\nabla f}\max(C(f),1)}{25}
$$

Assume now that f is λ-strongly convex. As we have seen,

$$\begin{aligned}\mathbb{E}_{u_k} r_{k+1}^2 &\leq r_k^2 - \frac{f(x_k) - f^*}{4(n+4)L_{\nabla f}} + \frac{11\mu^2(n+4)\max(C(f),1)}{100} \\ &\leq \left(1 - \frac{\lambda}{8(n+4)L_{\nabla f}}\right) r_k^2 + \frac{11\mu^2(n+4)\max(C(f),1)}{100}\end{aligned}$$

where we used $-\frac{\lambda}{2} r_k^2 \geq f^* - f(x_k)$ by strong convexity.

Taking the expectation in $\mathcal{U}_{k-1}$, we get

$$\rho_{k+1} \leq \left(1 - \frac{\lambda}{8(n+4)L_{\nabla f}}\right) \rho_k + \frac{11\mu^2(n+4)\max(C(f),1)}{100}$$

this inequality is equivalent to the following one:

$$\rho_{k+1} - \delta_\mu \leq \left(1 - \frac{\lambda}{8(n+4)L_{\nabla f}}\right)(\rho_k - \delta_\mu) \leq \left(1 - \frac{\lambda}{8(n+4)L_{\nabla f}}\right)^{k+1} (\rho_0 - \delta_\mu) \tag{198}$$

Since f has gradient Lipschitz, then $f(x_k) - f^* \leq \frac{L_{\nabla f}}{2} r_k^2$. Taking the expectation with respect to $\mathcal{U}_{k-1}$, we have $\phi_k - f^* \leq \frac{1}{2} L_{\nabla f} \rho_k$. Using this inequality and equation (198), we have

$$\phi_N - f^* \leq \frac{L_{\nabla f}}{2}\left[\delta_\mu + \left(1 - \frac{\lambda}{8(n+4)L_{\nabla f}}\right)^N (\|x_0 - x^*\|^2 - \delta_\mu)\right]$$

□

9.2 August 11

Recall we use the gradient estimator $g_\mu(x) = 2F(x)^\top \left\{\frac{F(x+\mu u) - F(x)}{\mu}\right\} u$ for $\nabla f_\mu(x)$.

Next, we compute ∇f_μ:

$$\begin{aligned}\nabla f_\mu(x) &= \mathbb{E}_u\left[\frac{f(x+\mu u) - f(x)}{\mu} u\right] \\ &= \mathbb{E}_u\left[\frac{\|F(x+\mu u)\|^2 - \|F(x)\|^2}{\mu} u\right] \\ &= \mathbb{E}_u\left[\frac{F(x+\mu u)^\top F(x+\mu u)}{\mu} u\right]\end{aligned}$$

In our meeting yesterday, Raghu mentioned

$$\mathbb{E}_u[g_\mu(x) - \nabla f_\mu(x)] = \frac{1}{\mu}\mathbb{E}_u\left[2\sum_{i=1}^p F_i(x+\mu u)\left[\frac{F_i(x+\mu u) - F_i(x)}{\mu}\right] u\right]$$

I don't see how this computation is correct.

The actual computation goes as follows:

$$\begin{aligned}\mathbb{E}_u\left[g_\mu(x) - \nabla f_\mu(x)\right] &= \mathbb{E}_u\left[\frac{2F(x)^\top F(x+\mu u)u}{\mu}u - \frac{F(x+\mu u)^\top F(x+\mu u)}{\mu}u\right] \\ &= \mathbb{E}_u\left[\sum_{i=1}^{p}\frac{F_i(x+\mu u)[2F_i(x) - F_i(x+\mu u)]}{\mu}u\right]\end{aligned}$$

9.3 August 12

Brain storm about Sassy Slides and thinking about the biased estimator case: $\mathbb{E}_u g_\mu(x) \neq \nabla f_\mu(x)$.

9.4 August 13

Finish up Sassy Slides and ponder more about the biased estimator case.

9.5 August 14

Here's an attempt using Nocedal's approach; not sure if this is useful or not. Nevertheless, I'll document it here.

Recall the update scheme $x_{k+1} = x_k - g_\mu(x_k)$, using bound in theorem 37, equation (192), and equation (167), we have

$$\begin{aligned}\mathbb{E}_{u_k}[f(x_{k+1}) - f^*] &\leq (1-\lambda h)(f(x_k) - f^*) + \frac{h}{2}\mathbb{E}_{u_k}\|g_\mu(x_k) - \nabla f(x_k)\|^2 \\ &= (1-\lambda h)(f(x_k) - f^*) + \frac{h}{2}\left[\mathbb{E}_{u_k}\|g_\mu(x_k)\|^2 - 2\mathbb{E}_{u_k}\langle g_\mu(x_k), \nabla f(x_k)\rangle + \|\nabla f(x_k)\|^2\right] \\ &= (1-\lambda h)(f(x_k) - f^*) \\ &+ \frac{h}{2}\left[\mathbb{E}_{u_k}\|g_\mu(x_k)\|^2 - \|\nabla f(x_k)\|^2 - 4\|\nabla f(x_k)\|\|F(x_k)\|\|\nabla F_\mu(x_k) - \nabla F(x_k)\|\cos\theta(x_k)\right] \\ &\leq (1-\lambda h)(f(x_k) - f^*) \\ &+ \mu^2 h f(x_k) L_{\nabla F}^2 (n+6)^3 + h(n+3.5)\|\nabla f(x_k)\|^2 \\ &- 2h\|\nabla f(x_k)\|\|F(x_k)\|\|\nabla F_\mu(x_k) - \nabla F(x_k)\|\cos\theta(x_k) \qquad (199)\end{aligned}$$

where $\theta(x_k)$ denotes the angle between $\nabla f_\mu(x_k) - \nabla f(x_k), \nabla f(x_k)$.

Thoughts on sufficient condition for $L_{\nabla f}$ to exist:

In "A Derivative-Free Gauss-Newton Method" by Coralia Cartis and Lindon Roberts, the authors mention a sufficient condition for ∇f to be lipschitz continuous (see Assumption 3.1, Remark 3.2, and Lemma 3.3). Not quite sure if this applies in our setting, but it requires the level set $\mathcal{L} := \{x : f(x) \leq f(x_0)\}$ to be bounded.

Task: Transfer slides over to Argonne powerpoint template. May need further revision before I send to Matt/Raghu.

10 Week 10: August 17

10.1 August 17

Return to the computation done on equation (199), we have

$$\begin{aligned}\mathbb{E}_{u_k}[f(x_{k+1}) - f^*] \leq [1 - \lambda h + \mu^2 h(n+6)^3 + 2L_{\nabla f} h(n+3.5)](f(x_k) - f^*) \\ + \mu^2 h(n+6)^3 f^* - 2h\|\nabla f(x_k)\|\|F(x_k)\|\|\nabla F_\mu(x_k) - \nabla F(x_k)\| \cos\theta(x_k) \end{aligned} \tag{200}$$

The term scaled by $(f(x_k) - f^*)$ is greater than 1. This approach used by Nocedal doesn't seem helpful.

Going back to the approach by Nesterov (see (194)), we need to deal with the term

$$[\mathbb{E}_{u_k} g_\mu(x_k) - \nabla f_\mu(x_k)](x_k - x^*)$$

Question: Can we use the identity

$$\nabla f_\mu(x) - \mathbb{E}_u[g_\mu(x)] = \mathbb{E}_u\left[\frac{\|F(x+\mu u) - F(x)\|^2}{\mu} u\right]$$

from Theorem 39 to help advance the convergence analysis proof?

Necessary condition for existence of $L_{\nabla f}$:

Maybe: $L_F, L_{\nabla F}$ exist? We have seen that F, f being bounded is not necessary.

10.2 August 18

10.2.1 Quick Note from Matt

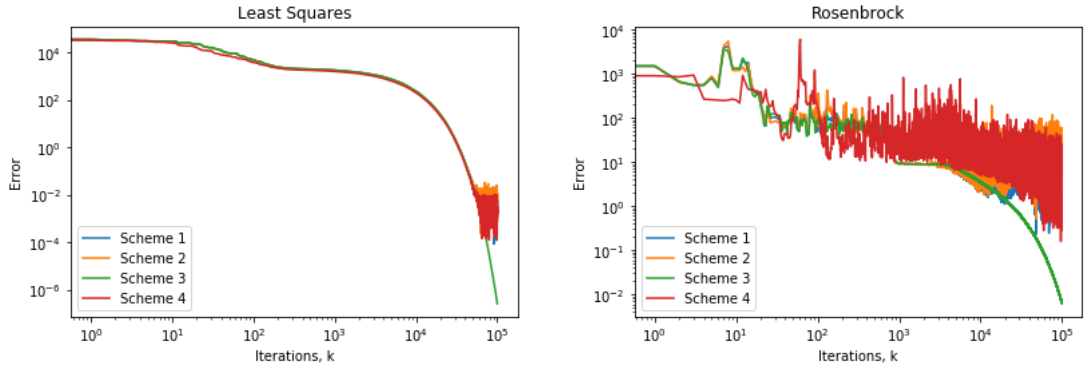
Consider the following (three point!!!) estimator:

$$g_\mu(x) = 2F(x+\mu u_1)^\top \left[\frac{F(x+\mu u_2) - F(x))}{\mu}\right] u_2,$$

where u_1, u_2 are independent samples from whatever (Gaussian) distribution. Then, by independence,

$$\mathbb{E}_{u_1,u_2}[g_\mu(x)] = 2\nabla F_\mu(x) F_\mu(x) = \nabla \sum_{i=1}^{p} F_i^2(x) = \nabla f_\mu(x).$$

Here's a simulation for test problems least squares and Rosenbrock:



We present a corollary to theorem 40 when $f^* = 0$.

Remark: Looks like this method of drawing two independent random vectors (denoted by scheme 4) is not any better than scheme 3.

Corollary 2. *Let $f^* = 0$. Assume $g_\mu(x) = 2F(x)^\top \left\{ \frac{F(x+\mu u)-F(x)}{\mu} u \right\}$ is an unbiased estimator of $\nabla f_\mu(x)$. Suppose we pick*

$$h = \frac{1}{4\left(\frac{\mu^2}{2} L_{\nabla F}^2 (n+6)^3 + (n+4) L_{\nabla f}\right)}$$

Then the convergence rate is

$$\frac{1}{N+1} \sum_{k=0}^{N} (\phi_k - f^*) \leq \frac{4(n+4) L_{\nabla f} \|x_0 - x^*\|^2}{N+1} + \mu^2 L_{\nabla f} n \tag{201}$$

where $\phi_k := \mathbb{E}_{\mathcal{U}_{k-1}}(f(x_k))$ for $k \geq 1$ and $\phi_0 := f(x_0)$ and definition of $\mathcal{U}_{k-1}$ is in lemma 30.

In addition, if f is λ-strongly convex and $\delta_\mu := \frac{2\mu^2 n L_{\nabla f}}{\lambda}$, then the rate becomes

$$\phi_N - f^* \leq \frac{L_{\nabla f}}{2} \left[\delta_\mu + \left(1 - \frac{\lambda}{8(n+4) L_{\nabla f}}\right)^N \left(\|x_0 - x^*\|^2 - \delta_\mu \right) \right] \tag{202}$$

Proof. Using $f^* = 0$ and the proof of theorem 40, we have

$$\mathbb{E}_{u_k} r_{k+1}^2 \leq r_k^2 - \frac{f(x_k) - f^*}{4(n+4) L_{\nabla f}} + \frac{\mu^2}{4} \left[\frac{n}{n+4} \right] \tag{203}$$

The rest follows from the proof of theorem 40. Next, using equation (203) and following the theorem 40 analysis, we have the conclusion. □

10.3 August 19

Recall from equation (194) here:

$$\mathbb{E}_{u_k} r_{k+1}^2 \leq r_k^2 - 2h[f(x_k) - f^*][1 - h\mu^2 L_{\nabla F}^2 (n+6)^3 - 2h(n+4) L_{\nabla f}] + \mu^2 n h L_{\nabla f} + h^2 [2\mu^2 f^* L_{\nabla F}^2 (n+6)^3]$$

$$- 2h[\mathbb{E}_{u_k} g_\mu(x_k) - \nabla f_\mu(x_k)](x_k - x^*)$$

We can proceed in proving the convergence analysis (pretty sure it works...but will finalize details later) when f is strongly convex, $g_\mu(x_k)$ is not necesarily an unbiased estimator of $\nabla f_\mu(x_k)$, F is not necessarily bounded, and using the fact:

$$\begin{aligned}|-2h[\mathbb{E}_{u_k} g_\mu(x_k) - \nabla f_\mu(x_k)](x_k - x^*)| \leq \frac{4hL_F}{\lambda}\mu(n+3)^{3/2}L_{\nabla F}(f(x_k) - f^*) \\ + h\mu\left[\sqrt{f^*}(n+3)^{3/2}L_{\nabla F} + \frac{L_{\nabla f}}{2}(n+3)^{3/2}\right] \\ + 2h(f(x_k) - f^*)\frac{\mu}{\lambda}\left[\sqrt{f^*}(n+3)^{3/2}L_{\nabla F} + \frac{L_{\nabla f}}{2}(n+3)^{3/2}\right]\end{aligned} \tag{204}$$

An additional parameter we need to control in addition to Nesterov's line of reasoning is μ. One remark to add: If we relax strongly convexity of f to convexity, I don't see how we can use Nesterov's approach to prove convergence. Proof to be added...

10.4 August 20

We continue with (204). First, we prove a lemma

Lemma 41. *Let $g_\mu(x) = 2F(x)^\top \left\{\frac{F(x+\mu u) - F(x)}{\mu}\right\} u$. By theorem 13 and lemma 3 in [?], we have*

$$\|\mathbb{E}_u g_\mu(x) - \nabla f_\mu(x)\| \leq \mu(n+3)^{3/2}L_{\nabla F}\|F(x)\| + \frac{\mu}{2}L_{\nabla f}(n+3)^{3/2} \tag{205}$$

Proof.

$$\begin{aligned}\|\mathbb{E}_u g_\mu(x) - \nabla f_\mu(x)\| &= \|2F(x)\nabla F_\mu(x) - 2F(x)\nabla F(x) + \nabla f(x) - \nabla f_\mu(x)\| \\ &\leq \|2F(x)[\nabla F_\mu(x) - \nabla F(x)]\| + \|\nabla f(x) - \nabla f_\mu(x)\| \\ &\leq 2\|F(x)\|\|\nabla F_\mu(x) - \nabla F(x)\| + \|\nabla f(x) - \nabla f_\mu(x)\| \\ &\leq \mu(n+3)^{3/2}L_{\nabla F}\|F(x)\| + \frac{\mu}{2}L_{\nabla f}(n+3)^{3/2}\end{aligned}$$

□

Using lemma 41, we prove equation (204).

Lemma 42. *Suppose F is lipschitz continuous and f is λ-strongly convex. Using lemma 41, we have*

$$\begin{aligned}|-2h[\mathbb{E}_{u_k} g_\mu(x_k) - \nabla f_\mu(x_k)](x_k - x^*)| \leq \frac{4hL_F}{\lambda}\mu(n+3)^{3/2}L_{\nabla F}(f(x_k) - f^*) \\ + h\mu\left[\sqrt{f^*}(n+3)^{3/2}L_{\nabla F} + \frac{L_{\nabla f}}{2}(n+3)^{3/2}\right] \\ + 2h(f(x_k) - f^*)\frac{\mu}{\lambda}\left[\sqrt{f^*}(n+3)^{3/2}L_{\nabla F} + \frac{L_{\nabla f}}{2}(n+3)^{3/2}\right]\end{aligned}$$

Proof. Since F is lipschitz continuous, we have

$$\|F(x)\| \leq \|F(x) - F(x^*)\| + \|F(x^*)\|$$

$$\leq L_F\|x - x^*\| + \|F(x^*)\|$$

By strong convexity of f and Young's inequality, we have

$$\|x_k - x^*\|^2 \leq \frac{2}{\lambda}(f(x_k) - f^*)$$
$$\|x_k - x^*\| \leq \frac{1}{2} + \frac{f(x_k) - f^*}{\lambda}$$

Continuing with the left hand side,

$$\begin{aligned}
|-2h[\mathbb{E}_{u_k} g_\mu(x_k) - \nabla f_\mu(x_k)](x_k - x^*)| &\leq 2h\|\mathbb{E}_{u_k} g_\mu(x_k) - \nabla f_\mu(x_k)\|\|x_k - x^*\| \\
&\leq 2h[\mu(n+3)^{3/2} L_{\nabla F}\|F(x_k)\| + \frac{\mu}{2} L_{\nabla f}(n+3)^{3/2}]\|x_k - x^*\| \\
&\leq 2h[L_F\|x_k - x^*\|\mu(n+3)^{3/2} L_{\nabla F} + \|F(x^*)\|\mu(n+3)^{3/2} L_{\nabla F}]\|x_k - x^*\| \\
&+ 2h[\frac{\mu}{2} L_{\nabla f}(n+3)^{3/2}]\|x_k - x^*\| \\
&= 2hL_F\|x_k - x^*\|^2\mu(n+3)^{3/2} L_{\nabla F} \\
&+ 2h[\|F(x^*)\|\mu(n+3)^{3/2} L_{\nabla F} + \frac{\mu}{2} L_{\nabla f}(n+3)^{3/2}]\|x_k - x^*\| \\
&\leq \frac{4hL_F}{\lambda}(f(x_k) - f^*)\mu(n+3)^{3/2} L_{\nabla F} \\
&+ 2h[\|F(x^*)\|\mu(n+3)^{3/2} L_{\nabla F} + \frac{\mu}{2} L_{\nabla f}(n+3)^{3/2}]\|x_k - x^*\| \\
&\leq \frac{4hL_F}{\lambda}(f(x_k) - f^*)\mu(n+3)^{3/2} L_{\nabla F} \\
&+ h[\|F(x^*)\|\mu(n+3)^{3/2} L_{\nabla F} + \frac{\mu}{2} L_{\nabla f}(n+3)^{3/2}] \\
&+ 2h\left[\frac{f(x_k) - f^*}{\lambda}\right][\|F(x^*)\|\mu(n+3)^{3/2} L_{\nabla F} + \frac{\mu}{2} L_{\nabla f}(n+3)^{3/2}]
\end{aligned}$$

Finally, factoring μ in the second and third term and writing $\sqrt{f^*} = \|F(x^*)\|$, we have our conclusion. □

Task: Using equation (194), we need to pick $h, \mu > 0$ appropriately to advance the convergence analysis.

Continuing with the convergence analysis in equation (194) and using lemma 42, we have

$$\mathbb{E}_{u_k} r_{k+1}^2 \leq r_k^2 - 2h[f(x_k) - f^*][\star_1] + \star_2 \tag{206}$$

where

$$\star_1 := 1 - h\mu^2 L_{\nabla F}^2 (n+6)^3 - 2h(n+4)L_{\nabla f} - \frac{2L_F}{\lambda}\mu(n+3)^{3/2} L_{\nabla F} - \frac{\mu}{\lambda}\left[\sqrt{f^*}(n+3)^{3/2} L_{\nabla F} + \frac{L_{\nabla f}}{2}(n+3)^{3/2}\right]$$
$$\star_2 := h\left[\mu^2 n L_{\nabla f} + \mu\left[\sqrt{f^*}(n+3)^{3/2} L_{\nabla F} + \frac{L_{\nabla f}}{2}(n+3)^{3/2}\right]\right] + h^2[2\mu^2 f^* L_{\nabla F}^2 (n+6)^3]$$

Taking the expectation with respect to $\mathcal{U}_{k-1}$ (see lemma 30),

$$\rho_{k+1} \leq \rho_k - 2h[\phi_k - f^*][\star_1] + \star_2 \tag{207}$$

where ρ_k is defined in [**?**]. Applying this inequality recursively,

$$2h\sum_{k=0}^{N}(\phi_k - f^*)[\star_1] \leq \|x_0 - x^*\|^2 + (N+1)[\star_2] \tag{208}$$

Rearranging yields

$$\frac{1}{N+1}\sum_{k=0}^{N}(\phi_k - f^*) \leq \frac{1}{2h}\left[\frac{\|x_0 - x^*\|^2}{N+1} + [\star_2]\right]\frac{1}{[\star_1]} \tag{209}$$

We show a different convergence analysis result. By strong convexity of f, we have

$$-(f(x_k) - f^*) \leq -\frac{\lambda}{2}\|x_k - x^*\|^2 \tag{210}$$

Using this and equation (206), we have

$$\begin{aligned}\mathbb{E}_{u_k} r_{k+1}^2 &\leq r_k^2 - \lambda h r_k^2[\star_1] + [\star_2] \\ &= (1 - \lambda h[\star_1])r_k^2 + [\star_2]\end{aligned} \tag{211}$$

where we pick $h, \mu > 0$ appropriately so $\star_1 \in [0, 1]$.

Taking the expectation with respect to $\mathcal{U}_{k-1}$ yields

$$\rho_{k+1} \leq (1 - \lambda h[\star_1])\rho_k + \star_2 \tag{212}$$

Using this and picking δ_μ such that $\lambda h[\star_1]\delta_\mu = [\star_2]$ and we have

$$\begin{aligned}\rho_{k+1} - \delta_\mu &\leq (1 - \lambda h[\star_1])(\rho_k - \delta_\mu) \\ &\leq (1 - \lambda h[\star_1])^{k+1}(\rho_0 - \delta_\mu)\end{aligned} \tag{213}$$

Since f has gradient lipschitz, taking expectation and the previous equation, we have

$$\begin{aligned}\phi_N - f^* &\leq \frac{L_{\nabla f}}{2}\rho_N \\ &\leq \frac{L_{\nabla f}}{2}\left[\delta_\mu + (1 - \lambda h[\star_1])^N(\|x_0 - x^*\|^2 - \delta_\mu)\right]\end{aligned} \tag{214}$$

11 Week 11: August 24

11.1 August 24

Unbiased, $f^* = 0$, see theorem 2: Convex:

$$\frac{1}{N+1}\sum_{k=0}^{N}(\phi_k - f^*) \leq \frac{4(n+4)L_{\nabla f}\|x_0 - x^*\|^2}{N+1} + \mu^2 L_{\nabla f} n$$

f is λ-strongly convex:

$$\phi_N - f^* \leq \frac{L_{\nabla f}}{2}\left[\delta_\mu + \left(1 - \frac{\lambda}{8(n+4)L_{\nabla f}}\right)^N (\|x_0 - x^*\|^2 - \delta_\mu)\right]$$

where $\delta_\mu = \frac{2\mu^2 n L_{\nabla f}}{\lambda}$.

Nesterov's convergence rates:

f is convex:

$$\frac{1}{N+1}\sum_{k=0}^{N}(\phi_k - f^*) \leq \frac{4(n+4)L_{\nabla f}\|x_0 - x^*\|^2}{N+1} + \frac{9\mu^2(n+4)^2 L_{\nabla f}}{25}$$

f is λ-strongly convex:

$$\phi_N - f^* \leq \frac{L_{\nabla f}}{2}\left[\delta_\mu + \left(1 - \frac{\lambda}{8(n+4)L_{\nabla f}}\right)^N (\|x_0 - x^*\|^2 - \delta_\mu)\right]$$

where $\delta_\mu = \frac{18\mu^2(n+4)^2 L_{\nabla f}}{25\lambda}$.

Biased estimator g_μ of ∇F_μ:

f is λ-strongly convex:

$$\frac{1}{N+1}\sum_{k=0}^{N}(\phi_k - f^*) \leq \frac{1}{2h}\left[\frac{\|x_0 - x^*\|^2}{N+1} + \star_2\right] + \frac{1}{\star_1}$$

$$\phi_N - f^* \leq \frac{L_{\nabla f}}{2}[\delta_\mu + (1 - \lambda h \star_1)^N (\|x_0 - x^*\|^2 - \delta_\mu)]$$

where $\delta_\mu = \frac{\star_2}{\lambda h \star_1}$

$$\star_1 := 1 - h\mu^2 L_{\nabla F}^2 (n+6)^3 - 2h(n+4)L_{\nabla f} - \frac{2L_F}{\lambda}\mu(n+3)^{3/2}L_{\nabla F} - \frac{\mu}{\lambda}\left[\sqrt{f^*}(n+3)^{3/2}L_{\nabla F} + \frac{L_{\nabla f}}{2}(n+3)^{3/2}\right]$$

$$\star_2 := h\left[\mu^2 n L_{\nabla f} + \mu\left[\sqrt{f^*}(n+3)^{3/2}L_{\nabla F} + \frac{L_{\nabla f}}{2}(n+3)^{3/2}\right]\right] + h^2[2\mu^2 f^* L_{\nabla F}^2 (n+6)^3]$$

11.2 August 25:

On August 5 (see 8.3), under scheme 3, we have shown g_μ is an unbiased estimator of ∇f_μ.

Now, consider the least squares setting where $F(x) = Ax - b$. Here are the Lipschitz constants:

$$L_{\nabla F} = \|A\|_2$$

$$
\begin{aligned}
L_{\nabla F} &= 0 \\
L_{\nabla f} &= 2\|A^\top A\|_2 \\
f^* &= 0
\end{aligned}
$$

where we assumed $b \in \text{range}(A)$. We compute $\star_1, \star_2$, and $\frac{\star_2}{h\star_1}$ using these constants:

$$
\begin{aligned}
\star_1 &= 1 - 4h(n+4)\|A^\top A\|_2 - \frac{\mu}{\lambda}\left[\|A^\top A\|_2(n+3)^{3/2}\right] \\
\star_2 &= h\left[2\mu^2 n\|A^\top A\|_2 + \mu\|A^\top A\|_2(n+3)^{3/2}\right] \\
\frac{\star_2}{h\star_1} &= \frac{2\mu^2 n\|A^\top A\|_2 + \mu\|A^\top A\|_2(n+3)^{3/2}}{1 - 4h(n+4)\|A^\top A\|_2 - \frac{\mu}{\lambda}\|A^\top A\|_2(n+3)^{3/2}}
\end{aligned}
$$

Remark: This bound in equation (214) and the previous equation yields that the error scales like $\mathcal{O}\left(\mu^2 n + \mu n^{3/2}\right)$ which is much worse than what we originally got for the unbiased estimator case where the error scales as $\mathcal{O}(\mu^2 n)$.

11.3 A Note from Matt

Let's analyze things a bit differently for a new perspective on how biased estimators are screwing up our analysis.

I'll use $g_\mu(x;u)$ to denote the "Scheme 3" estimator that we know and love by now. I'm not sure how well this jives with the notation established so far, but I've been interpreting $F(x) : \mathbb{R}^n \to \mathbb{R}^p$ to be defined by the vector $[F_1(x), \ldots, F_p(x)]$ and $F_\mu(x) : \mathbb{R}^n \to \mathbb{R}^p$ to be defined by the vector $[F_{\mu,1}(x), \ldots, F_{\mu,p}(x)]$, where $F_{\mu,i}$ denotes the Gaussian (or whatever distribution) smoothed version of the function $F_i(x)$. We assume that when computing $g_\mu(x;u)$, a single sample u from the distribution is used, i.e., a common u is used in the estimation of each entry of F_μ.

Let's use the classical descent lemma, and unlike Nesterov, let's look at decrease in f, instead of f_μ:

$$
f(x_{k+1}) \leq f(x_k) - h\langle \nabla f(x_k), g_\mu(x_k;u)\rangle + \frac{1}{2}h^2 L_{\nabla f}\|g_\mu(x_k;u)\|^2.
$$

Taking expectations of both sides with respect to u,

$$
\mathbb{E}_u\left[f(x_{k+1})\right] \leq f(x_k) - h\langle \nabla f(x_k), \mathbb{E}_u\left[g_\mu(x_k;u)\right]\rangle + \frac{1}{2}h^2 L_{\nabla f}\mathbb{E}_u\left[\|g_\mu(x_k;u)\|^2\right].
$$

Using Theorem 37,

$$
\mathbb{E}_u\left[f(x_{k+1})\right] \leq f(x_k) - h\langle \nabla f(x_k), \mathbb{E}_u\left[g_\mu(x_k;u)\right]\rangle + \frac{1}{2}h^2 L_{\nabla f}\left(2\mu^2 f(x)L_{\nabla F}^2(n+6)^3 + 2(n+4)\|\nabla f(x)\|^2\right).
$$

Now let's work on that inner product. We seek to *lower bound* the inner product, since we intend to upper bound its negative to get anything useful. Reexpressing (and denoting $\nabla F_\mu, \nabla F$ as the Jacobians of the respective functions in $\mathbb{R}^{p\times n}$).

$$
\langle \nabla f(x_k), \mathbb{E}_u\left[g_\mu(x_k;u)\right]\rangle = \langle \nabla f(x_k), 2\nabla F_\mu(x_k)^\top F(x_k)\rangle = \langle 2\nabla F(x_k)^\top F(x_k), 2\nabla F_\mu(x_k)^\top F(x_k)\rangle =
$$

$$
4F(x_k)^\top\left[\nabla F(x_k)\nabla F_\mu(x_k)^\top\right]F(x_k) = 4F(x_k)^\top\left[\sum_{i=1}^p \nabla F_i(x)\nabla F_{\mu,i}(x_k)^\top\right]F(x_k)
$$

Trivially, then, since the quadratic function is a sum of outer products,

$$\langle \nabla f(x_k), \mathbb{E}_u \left[g_\mu(x_k; u) \right] \rangle \geq 4 \|F(x_k)\|^2 \min_{i=1,\ldots,p} \left[\nabla F_{\mu,i}(x_k)^\top \nabla F_i(x_k) \right]$$

So, the only way this bound is useful is if the minimum eigenvalue is bounded below by zero, i.e., if *all* the inner products between the true gradient and the smoothed gradient are nonnegative! I'm not totally sure how you can guarantee this algorithmically (e.g., by adaptively driving μ) , but **it's an intuitive and reasonable geometric condition**. The other thing that's odd here is that we're getting a term that's effectively $-C\|F(x)\|^2$ for $C > 0$ provided the geometric condition is satisfied, when in SGD methods, one often expects to see $-C\|\nabla f(x)\|^2$. This actually makes the analysis look more like Gauss-Newton analysis. The problem is, however, that our current bound from Theorem 37 is giving us a SGD-type bound on the first-order Taylor remainder.

11.4 Closing Remarks:

In the case $g_\mu(x) = 2F(x)^\top \left\{ \frac{F(x+\mu u) - F(x)}{\mu} \right\} u$ is a biased estimator of $f_\mu(x)$, we didn't see any theoretical improvements in the convergence rate in contrast to Nesterov's algorithms for the unstructured case; we saw a speed up on convergence empirically. My question (which may have been answered) is how common is this?

Things I would like to have done: I am pausing research (in general) until late December 2020 or early January 2021 to secure a full time industry position upon graduation. When research resumes, I would like to incorporate it in my PhD disseration, whatever direction we take it. If an opportunity to continue this research with Stefan exists, I would like to involve my PhD advisor in this (currently we haven't discussed about this summer project). By then, I believe I'll have enough time away from research to be recharged and make some contribution (hopefully a publication). As a reference, I am planning to graduate at the end of July 2021, so hopefully this is realistic.