clean the report, share it with them, adding results of using different scaling mechanisms, using gradient descent directly onf_N starting with p_0.

- well-defined algorithm: algorithm can run and exit in finite time. Has nothing to do with convergence or algorithms.
- Be exact in formulating theorems or similar arguments: 1) make sure all the constants/parameters in theorems are at least mentioned where they are defined in the statement of theorems. 2) be concise. A theorem or similar arguments only involve assumptions and results. Any further results that can be inferred from the theorem should not be included in the statement of it.
- think it deeply: for example, the tradeoff of Assumption.2 on C and p and other things, whether some counterintuitive things will happen. Also, compare results across theorems.
- For statements involving probability, the probability should be viewed as a parameter and the statements are still be: exists/for all ..., then ...

1 Mokhtari, Ozdaglar, and Jadbabaie 2019

The problem is to minimize empirical risk via an incremental sampling technique.

Notations:

Notation	Meaning
$w \in \mathbb{R}^n$	model weight, optimization variable
$w_p \in \mathbb{R}^n$	model weight, output where $w_p \approx \operatorname{argmin} f_p(w)$
$w_p^t \in \mathbb{R}^n$	model weight in the tth subiteration for $\{\min_w f_p(w)\}$
$x_\xi \in \mathbb{R}^d$	a data point following distribution ξ
$F(w; x_\xi) \in \mathbb{R}$	loss for data x_ξ with weight w
$f_N(w) \in \mathbb{R}$	$f_N(w) = \frac{1}{N}\sum_{i=1}^N F(w; x_{\xi_i})$, empirical risk
$f(w) \in \mathbb{R}$	$f(w) = \mathbb{E}_\xi[F(w; x_\xi)]$, statistical risk
$\mathbb{B}_n(r) \subseteq \mathbb{R}^n$	$\mathbb{B}_n(r) = \{w \in \mathbb{R}^n \mid \|w\|_2 \leq r\}$

Definitions for symmetric matrix $A \in \mathbb{S}^{n\times n}$ with only real eigenvalues:

Definitions for a function $g : \mathbb{R}^n \to \mathbb{R}$

Goal: given N datapoints $\{x_{\xi_i}\}_{i=1}^N$, solve

$$\min_{w\in\mathbb{R}^n} f_N(w)$$

Definition	Meaning
$\lambda_i(A)$	ith largest eigenvalues of A
$\|A\|_{op}$	$\|A\|_{op} = \max\{\lambda_1, -\lambda_n\}$, operator norm

Definition	Meaning
(α, β)-strong-Morse	$\forall w$ s.t. $\|\nabla g(w)\|_2 \leq \alpha$, it holds: $\|\lambda_i(\nabla^2 g(w))\| \geq \beta, \forall i \in \{1, \cdots, n\}$

via incremental sampling.

Assumption 1 (Smoothness of gradient and Hessian). There exists constants M and L such that $\forall x \in \{x_{\xi_i}\}_{i=1}^N$ and $\forall (w_1, w_2) \in \mathbb{R}^n \times \mathbb{R}^n$, we have

$$\|\nabla_w F(w_1, x) - \nabla_w F(w_2, x)\|_2 \leq M\|w_1 - w_2\|_2, \text{ and}$$
$$\|\nabla_w^2 F(w_1, x) - \nabla_w^2 F(w_2, x)\|_{op} \leq L\|w_1 - w_2\|_2.$$

Assumption 2. There exist a $(r, \tau) \in (0, \infty) \times (0, \infty)$, such that for all $\delta \in (0, 1)$ there exist a constant C depends on δ and if the sample size $p \geq Cn\sqrt{\log n}$, the following holds with probability $(1 - \delta)$:

$$\sup_{w \in \mathbb{B}^n(r)} \|\nabla f_p(w) - \nabla f(w)\|_2 \leq \tau \sqrt{\frac{Cn \log p}{p}},$$
$$\sup_{w \in \mathbb{B}^n(r)} \|\nabla^2 f_p(w) - \nabla^2 f(w)\|_{op} \leq \tau^2 \sqrt{\frac{Cn \log p}{p}}.$$

Proposition 3. Suppose Assumption.2 holds. In addition, let (r, τ) be defined by Assumption.2, and for any $\delta \in (0, 1)$, let C be defined by Assumption.2. And suppose $p \geq Cn\sqrt{\log n}$. Then, for all $w_p \in \mathbb{B}^n(r)$ satisfying

$$\|\nabla f_p(w_p)\|_2 \leq \tau \sqrt{\frac{Cn \log p}{p}},$$

if the sample set of f_{2p} contains that of f_p, we have:

$$\underbrace{\|\nabla f_{2p}(w_p) - \nabla f_p(w_p)\|_2 \leq \tau \sqrt{\frac{Cn \log p}{p}}}_{\text{holds with probability } (1-\delta)}$$

and

$$\underbrace{\|\nabla f_{2p}(w_p)\|_2 \leq 2\tau \sqrt{\frac{Cn \log p}{p}}}_{\text{holds with probability } (1-\delta)^2}, \quad \underbrace{\|\nabla f(w_p)\|_2 \leq 2\tau \sqrt{\frac{Cn \log p}{p}}}_{\text{holds with probability } (1-\delta)}.$$

See Appendix B for proof.

Assumption 4 (Conditions(Topology) of problems). There exists the same $r \in (0, \infty)$ as defined in Assumption.2, such that in the ball $\mathbb{B}^n(r)$, the f is (α, β)-strongly-Morse.

Now, for the simplicity of the argument, we let the probability parameter $\delta \in (0, 1)$ be a fixed constant throughout the report. Moreover, for any assumptions or propositions that utilize Assumption.2 and use notation (n, r, τ, δ, C), we let the (n, r, τ, δ, C) to be the same constant as defined in Assumption.2.

Proposition 5. Under Assumption.2 in which (n, r, τ, δ, C) is defined and under Assumption.4 in which (α, β) is defined, when the sample size p satisfies

$$p \geq Cn\sqrt{\log n} \text{ and } \frac{p}{\log p} \geq \max\{\frac{Cn\tau^2}{\alpha^2}, \frac{Cn\tau^4}{\beta^2}\},$$

the f_p is (α_p, β_p)-strongly-Morse with at least probability $(1 - \delta)$ where

$$\alpha_p = \alpha - \epsilon(p) = \alpha - \tau\sqrt{\frac{Cn\log p}{p}} \geq 0,$$
$$\beta_p = \beta - \tau\epsilon(p) = \beta - \tau^2\sqrt{\frac{Cn\log p}{p}} \geq 0.$$

See proofs in Appendix C.

Let p_0 be the initial sample size. Let the condition for the initial approximate solution w_{p_0} be:

$$\|\nabla f_{p_0}(w_{p_0})\|_2 \leq \tau\sqrt{\frac{Cn\log p_0}{p_0}} := \epsilon(p_0) \text{ and } \nabla^2 f_{p_0}(w_{p_0}) \succeq 0. \tag{1}$$

Now, their algorithm processes as follows:

Algorithm 1 Mokhtari, Ozdaglar, and Jadbabaie 2019

Input: initial sample size p_0, initial approximate solution w_{p_0} satisfying (1), constant (C, τ), dimension n, data set $\{x_{\xi_1}, \cdots, x_{\xi_N}\}$, iterator $k = 0$.

Output: $w^* \in \text{argmin}\{f_N(w)\}$.

1: **for** $k = 1 \ldots, \lceil \log_2 \frac{N}{p_0} \rceil$ **do**

2: Let $p_k = \min\{2p_{k-1}, N\}$ and adding $(p_k - p_{k-1})$ extra samples to the current samples.

3: Use a subproblem solver and take $w_{p_{k-1}}$ as the starting point, solve $\min\{f_{p_k}(w)\}$ approximately with the solution w_{p_k} satisfying

$$\|\nabla f_{p_k}(w_{p_k})\|_2 \leq \tau \sqrt{\frac{Cn \log p_k}{p_k}} := \epsilon(p_k).$$

4: Set $k = k + 1$.

5: **end for**

6: Output $w^* = w_{k-1}$.

2 Analysis

2.1 Consequence of choosing p_0 and w_{p_0} satisfying (1)

The choice of the p_0 is to ensure certain conditions are satisfied, and two of them are

(1). f_{p_0} is a $(\alpha_{p_0}, \beta_{p_0}) \in \mathbb{R}_+ \times \mathbb{R}_+$ strong morse function,

(2). when $\|\nabla f_{p_0}(w_{p_0})\|_2 \leq \epsilon(p_0)$, it holds $\|\nabla f_{p_0}(w_{p_0})\|_2 \leq \alpha_{p_0}$.

If w_{p_0} satisfies $\|\nabla f_{p_0}(w_{p_0})\|_2 \leq \alpha_{p_0}$, then by strong morse function, all eigenvalues satisfy $\max_{i=1,\cdots,n}\{|\lambda_i(\nabla^2 f_{p_0}(w_{p_0}))|\} \geq \beta_{p_0}$. Combining with (1) that $\nabla^2 f_{p_0}(w_{p_0}) \succeq 0$ we have $\lambda_i(\nabla^2 f_{p_0}(w_{p_0})) \geq \beta_{p_0}$ for all $i = 1, \cdots, n$.

Let us take the doubling sample size as an example. The initial p_0 is to satisfy

$$p_0 \geq Cn\sqrt{\log n} \text{ and } \frac{p_0}{\log p_0} \geq \max\{\frac{9Cn\tau^2}{\alpha^2}, \frac{4Cn\tau^4}{\beta^2}\}.$$

Then by Prop.5 we have that f_{p_0} is a $(\alpha_{p_0}, \beta_{p_0})$ strong morse function where

$$\alpha_{p_0} = \alpha - \tau\sqrt{\frac{Cn\log p_0}{p_0}} \geq \frac{2}{3}\alpha,$$
$$\beta_{p_0} = \beta - \tau^2\sqrt{\frac{Cn\log p_0}{p_0}} \geq \frac{1}{2}\beta.$$

Moreover we have $\epsilon(p_0) = \tau\sqrt{\frac{Cn\log p_0}{p_0}} \leq \frac{1}{3}\alpha$ and when the solution w_{p_0} satisfies $\|\nabla f_{p_0}(w_{p_0})\|_2 \leq \epsilon(p_0) \leq \alpha_{p_0}$, by the definition of strong morse we have $|\lambda_i(\nabla^2 f_{p_0}(w_{p_0}))| \geq \beta_{p_0}$ for all i. And by the requirement of w_{p_0} that $\nabla^2 f_{p_0}(w_{p_0}) \succeq 0$ we have $\nabla^2 f_{p_0}(w_{p_0}) \succeq \beta_{p_0} I$.

2.2 Other ways of choosing samples

$\mathbb{Q}$: Does the analysis depend on the way how samples are added?
$\mathbb{A}$: I don't think it depends on how samples are added, as long as we are adding samples. The whole analysis relies on what they call the 'uniform convergence theorem', similar to Assump.2 in this report. It holds as long as the number of samples is larger than the number, without a specific distribution of choosing the sample data.

However, if the set of active samples is randomly chosen at each time, instead of being added by other samples, then there are some differences.

As far as I am concerned, the ways of choosing samples will only affect the bound for $\|\nabla f_{p_{k+1}}(w_{p_k})\|_2$. Define $\mathcal{E}(p_{k+1}, p_k) \in \mathbb{R}_+$ to be the bound for $\|\nabla f_{p_{k+1}}(w_{p_k})\|_2$, which is

$$\|\nabla f_{p_{k+1}}(w_{p_k})\|_2 \leq \mathcal{E}(p_{k+1}, p_k).$$

then

(1). when $p_{k+1} = (1+a)p_k$, extra samples are added to the current sample set, and the gradient $\|\nabla f_{p_k}(w_{p_k})\|_2 \leq \epsilon(p_k)$, we have

$$\begin{aligned}\mathcal{E}(p_{k+1}, p_k) &:= \tau\frac{p_{k+1}-p_k}{p_{k+1}}\sqrt{Cn}\left(\sqrt{\frac{\log(p_{k+1}-p_k)}{p_{k+1}-p_k}} + \sqrt{\frac{\log p_k}{p_k}}\right) \\ &= \underbrace{\left(\frac{a+\sqrt{a\left(1+\log_{p_k} a\right)}}{1+a}\right)}_{:=c(a,p_k)}\epsilon(p_k).\end{aligned}$$

And that $\epsilon(p_k) = \tau\sqrt{Cn}\sqrt{\frac{\log p_k}{p_k}} = \mathcal{E}(2p_k, p_k)$.

Since the value p_k in $c(a, p_k)$ is different for different subproblems, I strengthen it by assuming $p_k \geq p_0 \geq 2$ and $c(a, p_k) \leq c(a, 2)$. For simplicity of notation, since value $c(a, 2)$ depends only on a, we use the notation of $c(a)$ instead.

(2). when $p_{k+1} = (1+a)p_k$, the p_{k+1} samples are randomly chosen, and the gradient $\|\nabla f_{p_k}(w_{p_k})\|_2 \leq \epsilon(p_k)$, we have

$$\begin{aligned}\mathcal{E}(p_{k+1}, p_k) &:= \epsilon(p_{k+1}) + \epsilon(p_k) \\ &= \underbrace{\left(\frac{a+1+\sqrt{(a+1)\left(1+\log_{p_k}(a+1)\right)}}{1+a}\right)}_{:=c'(a,p_k)} \epsilon(p_k).\end{aligned}$$

And we can see that

$$c'(a, p_k) = \frac{a+2}{a+1} \cdot c(a+1, p_k) > \max\{c(a+1, p_k), c(a, p_k)\},$$

where $c'(a, p_k) > c(a, p_k)$ is by comparing their formulas.

2.3 The complexity of computing w_{p_0}

The

2.4 Propositions

Proposition 6 (Subproblems are locally strongly convex). Under Assumption.2 in which the constant (n, r, τ, δ, C) is defined and under Assumption.4 in which the constant (α, β) is defined, if the initial sample size p_0 satisfies:

$$p_0 \geq Cn\sqrt{\log n} \text{ and } \frac{p_0}{\log p_0} \geq \max\{\frac{9Cn\tau^2}{\alpha^2}, \frac{4Cn\tau^4}{\beta^2}\}.$$

Then with probability

$$(1-\delta)^{1+5\left(\lceil \log_2 \frac{N}{p_0} \rceil - 1\right)},$$

the following holds for all k:

$$\|\nabla f_{p_{k+1}}(w_{p_k})\|_2 \leq \alpha_{p_{k+1}}, \nabla^2 f_{p_{k+1}}(w_{p_k}) \succeq \beta_{p_{k+1}} I \tag{2}$$

See Appendix D for proof.

Proposition 7 (Complexity using gradient descent)*.* Under assumptions and statements in Prop.6 in which the constant $(n, r, \tau, \delta, C, \alpha, \beta, p_0)$ is defined. In addition, under Assumption.1 in which the constant (L, M) is defined, if all the subproblems in line 5 of Alg.0 are solved by gradient descent:

$$w_{p_k}^{t+1} = w_{p_k}^{t} - \eta_k \nabla f_{p_k}(w_{p_k}^{t}), \text{ where } \eta_k = \min\{\frac{1}{M}, \frac{\beta_{p_k}}{2L\epsilon(p_k)}\},$$

then with probability

$$(1-\delta)^{1+5\left(\lceil \log_2 \frac{N}{p_0} \rceil - 1\right)},$$

the total number of individual gradient evaluations of all subproblems from w_{p_0} to w_N is bounded by

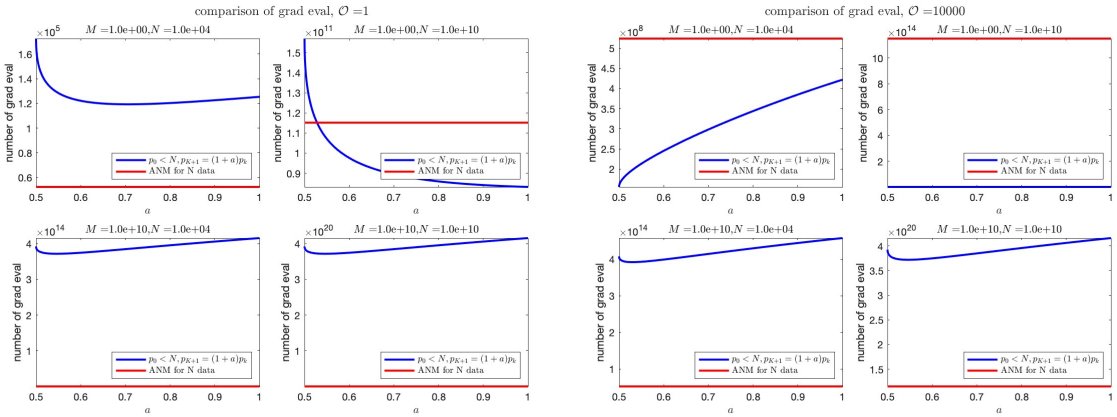
$$\max\left\{\frac{6M}{\beta}\log 2, \frac{12L\alpha}{\beta^2}\log 2\right\} N.$$

See Appen.E for proof.

Table 1: Comparison of different rates. Here we let $p_0 \geq Cn\sqrt{\log n}$ to ensure the requirements of Assumption.2 holds. We also let $p_0 \geq 2$ for the ease of analysis. In addition, the number of added samples can not be less than one, which requires a to satisfy $p_{k+1} - p_k = ap_k \geq ap_0 \geq 1$. Moreover, we define $\sigma := \min\{\sqrt{\alpha}, \beta\}$ and $\sigma_{p_0} := \min\{\sqrt{\alpha_{p_0}}, \beta_{p_0}\}$.

choice of $\{p_k\}$	$p_0 = N$	$p_{k+1} = 2p_k$	$p_{k+1} = (1+a)p_k$	$p_1 = N$
requirements for p_0	$\frac{\log p_0}{p_0} \leq \min\left\{\frac{\alpha^2}{9\tau^2 Cn}, \frac{\beta^2}{4\tau^4 Cn}\right\}$	$\frac{\log p_0}{p_0} \leq \min\left\{\frac{\alpha^2}{9\tau^2 Cn}, \frac{\beta^2}{4\tau^4 Cn}\right\}$	$\frac{\log p_0}{p_0} \leq \min\left\{\frac{\alpha^2}{(2+c(a))^2\tau^2 Cn}, \frac{\beta^2}{(1+c(a))^2\tau^4 Cn}\right\}$	$\frac{\log p_0}{p_0} \leq \min\left\{\frac{25\alpha^2}{324\tau^2 Cn}, \frac{25\beta^2}{169\tau^4 Cn}\right\}$
consequence of choice of p_0	$\sigma_{p_0} \geq \frac{1}{2}\sigma$	$\sigma_{p_0} \geq \frac{1}{2}\sigma$	$\sigma_{p_0} \geq \frac{c(a)}{1+c(a)}\sigma$, and $a \geq 1/p_0$	$\sigma_{p_0} \geq \frac{8}{13}\sigma$
#grad eval of ANM to compute w_{p_0}	$\mathcal{O}\left(N\frac{M^{1/2}L^2\log\sigma_{p_0}^{-2}}{\sigma_{p_0}^{7/2}} + N\frac{M^{1/2}}{\sigma_{p_0}^{1/2}}\log\frac{\sqrt{N}}{\sigma_{p_0}}\right)$	$\mathcal{O}\left(p_0\frac{M^{1/2}L^2\log\sigma_{p_0}^{-2}}{\sigma_{p_0}^{7/2}} + p_0\frac{M^{1/2}}{\sigma_{p_0}^{1/2}}\log\frac{\sqrt{p_0}}{\sigma_{p_0}}\right)$	$\mathcal{O}\left(p_0\frac{M^{1/2}L^2\log\sigma_{p_0}^{-2}}{\sigma_{p_0}^{7/2}} + p_0\frac{M^{1/2}}{\sigma_{p_0}^{1/2}}\log\frac{\sqrt{p_0}}{\sigma_{p_0}}\right)$	$\mathcal{O}\left(p_0\frac{M^{1/2}L^2\log\sigma_{p_0}^{-2}}{\sigma_{p_0}^{7/2}} + p_0\frac{M^{1/2}}{\sigma_{p_0}^{1/2}}\log\frac{\sqrt{p_0}}{\sigma_{p_0}}\right)$
#grad evals of GD (excludes computation of w_{p_0})	not applicable	$N \cdot \max\left\{\frac{6M}{\beta}, \frac{12L\alpha}{\beta^2}\right\} \cdot \log 2$	$N \cdot \max\{\frac{2(1+c(a))M}{c(a)\beta}, \frac{2(1+c(a))^2 L\alpha}{c(a)^2\beta^2}\} \cdot \log\left((1+c(a))\sqrt{(1+a)}\right)$	$N \cdot \max\left\{\frac{13M}{4\beta}, \frac{169L\alpha}{32\beta^2}\right\} \cdot \log\left(\min\left\{\frac{13\alpha}{18}, \frac{\beta}{\tau}\right\}\sqrt{\frac{N}{\log N}}\right)$
#iterations of GD (excludes computation of w_{p_0})	not applicable	$\log_2\frac{N}{p_0} \cdot \max\left\{\frac{6M}{\beta}, \frac{12L\alpha}{\beta^2}\right\} \cdot \log 2$	$\log_{1+a}\frac{N}{p_0} \cdot \max\{\frac{2(1+c(a))M}{c(a)\beta}, \frac{2(1+c(a))^2 L\alpha}{c(a)^2\beta^2}\} \cdot \log\left((1+c(a))\sqrt{(1+a)}\right)$	$\max\left\{\frac{13M}{4\beta}, \frac{169L\alpha}{32\beta^2}\right\} \cdot \log\left(\min\left\{\frac{13\alpha}{18}, \frac{\beta}{\tau}\right\}\sqrt{\frac{N}{\log N}}\right)$
note	the requirement for p_0 is only by σ_{p_0}	not applicable	not applicable	$N \geq p_0$ gives: $\min\left\{\frac{13\alpha}{18}, \frac{\beta}{\tau}\right\}\sqrt{\frac{N}{\log N}} \geq \frac{13}{5}$

The result of total gradient evaluations includes the computation of w_{p_0}.



(a) $\mathcal{O} = 1$. (b) $\mathcal{O} = 10^4$.

Figure 1: Total number of gradient evaluations when N and M are different. For both (a) and (b), the left column chooses samples $N = 10^4$ and the right choose $N = 10^{10}$. The first row chose $M = 1$ and the second row chose $M = 10^{10}$. The parameters $(\alpha, \beta, L, \tau, C)$ are all one and $n = 10$. The initial sample size p_0 is chosen such that $\frac{\log p_0}{p_0} \leq \frac{1}{p_0^{1/2}} = \min\{\cdots\}$ as in the second row of the table.

How does (a, p_0, N) relates to each other:

$$a \text{ is given} \to p_0 \text{ computed} \to N \geq p_0.$$

Since p_0 depends on a and not depend on (M, N, c_o), we draw the plot of p_0 below with $(\alpha, \beta, L, \tau, C)$ are all one and $n = 10$:

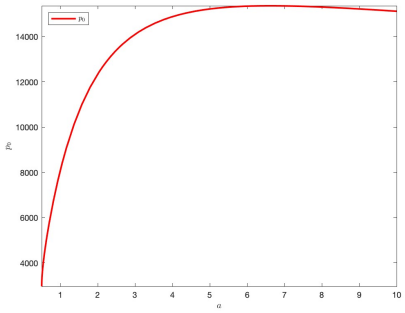
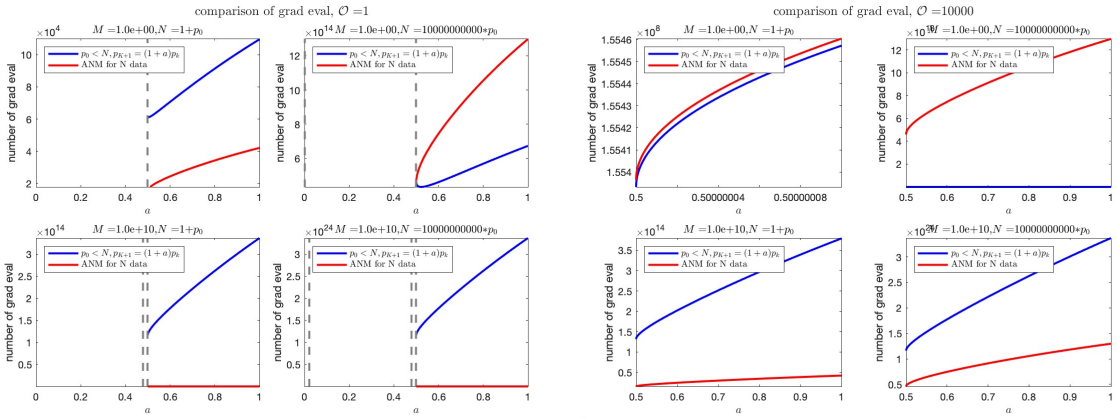


Figure 2: Required initial sample size p_0. The parameters $(\alpha, \beta, L, \tau, C)$ are all one and $n = 10$. The initial sample size p_0 is chosen such that $\frac{\log p_0}{p_0} \leq \frac{1}{p_0^{1/2}} = \min\{\cdots\}$ as in the second row of the table.

Moreover, when $(\alpha, \beta, L, \tau, C)$ are all one, $\sigma_N \geq \sigma_0 \geq 1$ and $n = 10$, suppose the $\mathcal{O}(a) = c_o \cdot a$, we have the total number of gradient evaluations:

$$\text{ANM: } \frac{1}{2} c_o M^{\frac{1}{2}} N \log N,$$
$$\text{incremental sampling: } \frac{1}{2} c_o M^{\frac{1}{2}} p_0 \log p_0 + (constant(a)) MN.$$

We draw a plot that shares the same parameter settings as the previous one with the only difference in N. For the left column in (a) and (b), we choose sample size $N = 1 + p_0$, and for the right column choose it to be $N = 10^{10} \cdot p_0$.



(a) $\mathcal{O} = 1$. (b) $\mathcal{O} = 10^4$.

Figure 3: Total number of gradient evaluations when N and M are different. For both (a) and (b), the left column chooses samples $N = 1 + p_0$ and the right choose $N = 10^{10} \cdot p_0$. The first row chose $M = 1$ and the second row chose $M = 10^{10}$. The parameters $(\alpha, \beta, L, \tau, C)$ are all one and $n = 10$. The initial sample size p_0 is chosen such that $\frac{\log p_0}{p_0} \leq \frac{1}{p_0^{1/2}} = \min\{\cdots\}$ as in the second row of the table.

A Assumptions for Sample Average

Assumption 8. Bounded noise on gradient and Hessian:

- When w is within a ball, the gradient of $F(w, x_\xi)$ is τ^2-sub-Gaussian:
  For all $w \in \mathbb{B}_n(r)$ and for all $y \in \mathbb{R}^n$,

$$\mathbb{E}\left[\exp\left\{\langle y, \nabla_w F(w, x_\xi) - \mathbb{E}_\xi\left[\nabla_w F(w, x_\xi)\right]\rangle\right\}\right] \leq \exp\left[\frac{\tau^2\|y\|_2^2}{2}\right].$$

- When w is within a ball, the Hessain of $F(w, x_\xi)$ is τ^2-sub-exponential:
  For all $w \in \mathbb{B}_n(r) = \{w \in \mathbb{R}^n | \|w\|_2 \leq r\}$ and for all $y \in \mathbb{B}^n(1)$, (namely $\|y\|_2 \leq 1$),

$$\mathbb{E}\left[\exp\left\{\frac{1}{\tau^2}\langle y, \nabla_w^2 F(w, x_\xi)y - \mathbb{E}_\xi\left[\nabla_w^2 F(w, x_\xi)\right] y\rangle\right\}\right] \leq 2.$$

These assumptions are the same as the Assumption 1 & 2 in Mei, Bai, and Montanari 2018.

Assumption 9 (Smoothness of gradient and Hessian). There exists constants $M = \tau^2 n^{c_h}$ and $L = M = \tau^3 n^{c_h}$ where c_h is a dimensionless constant, such that $\forall x \in \mathbb{R}^d$ and $\forall (w_1, w_2) \in \mathbb{R}^n \times \mathbb{R}^n$, we have

$$\|\nabla_w F(w_1, x) - \nabla_w F(w_2, x)\|_2 \leq M\|w_1 - w_2\|, \text{ and}$$
$$\|\nabla_w^2 F(w_1, x) - \nabla_w^2 F(w_2, x)\|_{op} \leq L\|w_1 - w_2\|,$$

where for $A \in \mathbb{R}^{n\times n}$ with eigenvalues $\{\lambda_1, \cdots, \lambda_n\}$ ranked in nonincreasing order, the $\|A\|_{op} = \max\{\lambda_1, -\lambda_n\}$.

This assumption is a stronger version of Assumption 3 in Mei, Bai, and Montanari 2018 in both the gradient and Hessian. The original assumption is:

Assumption 10. Original assumption:

- The Hessian of statistical risk $f(w)$ is bounded at one point:
  There exists a w^* such that $\|\nabla^2 f(w^*)\|_{op} \leq M \leq \tau^2 n^{c_h}$.

- The Hessian is Lipschitz in expectation:
  $\mathbb{E}_\xi\left[\sup_{\forall w_1 \neq w_2 \in \mathbb{R}^n}\left\{\frac{\|\nabla_w^2 F(w_1, x_\xi) - \nabla_w^2 F(w_2, x_\xi)\|_{op}}{\|w_1 - w_2\|_2}\right\}\right] \leq L \leq \tau^3 n^{c_h}$.

Proposition 11. Under Assumption (8,9), for any $\delta \in (0,1)$, there exists a constant C depends on ($\underbrace{r, \tau,}_{\text{by assumptions}}$ δ) such that when the sample size $p \geq Cn\sqrt{\log n}$, each of the following holds with at least probability $1-\delta$:

$$\sup_{w\in\mathbb{B}^n(r)} \|\nabla f_p(w) - \nabla f(w)\|_2 \leq \tau\sqrt{\frac{Cn\log p}{p}},$$

$$\sup_{w\in\mathbb{B}^n(r)} \|\nabla^2 f_p(w) - \nabla^2 f(w)\|_{op} \leq \tau^2\sqrt{\frac{Cn\log p}{p}}.$$

B Proof for Prop.3

Proof. We first show the result for $\|\nabla f_{2p}(w_p)\|_2$. Note that for the notation f_p, the subscript for f only indicates the number of samples instead of the set of samples. Only in this proposition, we allow the subscript of f being a set, representing the average loss in that set, i.e. if S is a sample set, then $f_S(w) := \frac{1}{|S|}\sum_{\xi_i\in S} F(w; x_{\xi_i})$.

Let the set of samples for f_p and f_{2p} be S_p and S_{2p}, where $S_p \subset S_{2p}$ by Algorithm.0. By definition of $f_{S_{2p}}$, we have

$$\begin{aligned}
f_{S_{2p}}(w) &= \frac{1}{2p}\sum_{i\in S_{2p}} F(w; x_{\xi_i}) \\
&= \left(\frac{1}{2p}\sum_{i\in S_p} F(w; x_{\xi_i})\right) + \left(\frac{1}{2p}\sum_{i\in S_{2p}/S_p} F(w; x_{\xi_i})\right) \\
&= \frac{1}{2}\left(\frac{1}{p}\sum_{i\in S_p} F(w; x_{\xi_i})\right) + \frac{1}{2}\left(\frac{1}{p}\sum_{i\in S_{2p}/S_p} F(w; x_{\xi_i})\right) \\
&= \frac{1}{2}f_{S_p}(w) + \frac{1}{2}f_{S_{2p}/S_p}(w).
\end{aligned}$$

Hence

$$\nabla f_{S_{2p}}(w) = \frac{1}{2}\nabla f_{S_p}(w) + \frac{1}{2}\nabla f_{S_{2p}/S_p}(w),$$

and

$$\begin{aligned}
\nabla f_{S_{2p}}(w) - \nabla f_{S_p}(w) &= -\frac{1}{2}\nabla f_{S_p}(w) + \frac{1}{2}\nabla f_{S_{2p}/S_p}(w) \\
&= \frac{1}{2}\left(\nabla f(w) - \nabla f_{S_p}(w)\right) + \frac{1}{2}\left(\nabla f_{S_{2p}/S_p}(w) - \nabla f(w)\right).
\end{aligned} \tag{3}$$

By the Assumption.2, for any $w_p \in \mathbb{B}^n(r)$, each of the following holds with probability $(1-\delta)$:

$$\begin{aligned} \|\nabla f_{S_p}(w_p) - \nabla f(w_p)\|_2 &\leq \tau\sqrt{\frac{Cn\log p}{p}}, \\ \|\nabla f_{S_{2p}/S_p}(w_p) - \nabla f(w_p)\|_2 &\leq \tau\sqrt{\frac{Cn\log p}{p}}. \end{aligned} \tag{4}$$

Since each of the above inequalities is independent, both of them hold with probability $(1-\delta)^2$.

Combining with the (3) and using triangular inequality, we have that with probability $(1-\delta)^2$:

$$\begin{aligned} &\|\nabla f_{S_{2p}}(w_p) - \nabla f_{S_p}(w_p)\|_2 \\ &\leq \frac{1}{2}\left\|\nabla f(w_p) - \nabla f_{S_p}(w_p)\right\|_2 + \frac{1}{2}\left\|\nabla f_{S_{2p}/S_p}(w_p) - \nabla f(w_p)\right\|_2 \\ &\leq \tau\sqrt{\frac{Cn\log p}{p}}. \end{aligned}$$

Hence on condition that the stopping condition for $\{\min f_{S_p}(w)\}$ is satisfied, i.e. $\|\nabla f_{S_p}(w_p)\|_2 \leq \tau\sqrt{\frac{Cn\log p}{p}}$, we have

$$\|\nabla f_{S_{2p}}(w_p)\|_2 \leq 2\tau\sqrt{\frac{Cn\log p}{p}} \text{ with probability } (1-\delta)^2.$$

Or equivalently, let $\mathbb{P}$ represent probability, we have

$$\mathbb{P}\left(\|\nabla f_{S_{2p}}(w_p)\|_2 \leq 2\tau\sqrt{\frac{Cn\log p}{p}}\middle|\|\nabla f_{S_p}(w_p)\|_2 \leq \tau\sqrt{\frac{Cn\log p}{p}}\right) \geq (1-\delta)^2.$$

And the same goes for the Hessian.

For the bound on $\|\nabla f(w_p)\|_2$, we have $\|\nabla f(w_p)\|_2 \leq \|\nabla f(w_p) - \nabla f_{S_p}(w_p)\|_2 + \|\nabla f_{S_p}(w_p)\|_2$ by triangle inequality. Moreoever, condition on $\|\nabla f_{S_p}(w_p)\|_2 \leq \tau\sqrt{\frac{Cn\log p}{p}}$, by (4) we have that with probability $(1-\delta)$:

$$\begin{aligned} \|\nabla f(w_p)\|_2 &\leq \|\nabla f(w_p) - \nabla f_{S_p}(w_p)\|_2 + \|\nabla f_{S_p}(w_p)\|_2 \\ &\leq 2\tau\sqrt{\frac{Cn\log p}{p}}. \end{aligned}$$

And the similar goes for $\nabla f_{S_{2p}}(w_{2p})$. □

C Assumptions on Problem Objectives

Assumption 12. The $f(w)$ is $(\alpha, \beta) \in \mathbb{R}_+ \times \mathbb{R}_+$ strongly morse function: for all w such that $\|\nabla f(w)\|_2 \leq \alpha$, the Hessian's eigenvalues satisfy $|\lambda_i(\nabla^2 f(w))| \geq \beta$ for all $i = 1, \cdots, n$.

Proposition 13 (Conditions(Topology) of subproblems). If assumptions in Proposition 4 and Assumption 5 hold ($f(w)$ is (α, β)-strongly-morse), then for sample size

$$p \geq Cn\sqrt{\log n} \text{ and } \frac{p}{\log p} \geq \max\{\frac{Cn\tau^2}{\alpha^2}, \frac{Cn\tau^4}{\beta^2}\},$$

the function $f_p(w)$ is (α_p, β_p)-strongly-morse with at least probability $(1-\delta)$ where

$$\alpha_p = \alpha - \epsilon(p) = \alpha - \tau\sqrt{\frac{Cn\log p}{p}} \geq 0,$$

$$\beta_p = \beta - \tau\epsilon(p) = \beta - \tau^2\sqrt{\frac{Cn\log p}{p}} \geq 0.$$

D Proof for Prop.6

Proposition 14 (Subproblems are locally strongly convex). Under Assumption.2 in which the constant (n, r, τ, δ, C) is defined and under Assumption.4 in which the constant (α, β) is defined, if the initial sample size p_0 satisfies:

$$p_0 \geq Cn\sqrt{\log n} \text{ and } \frac{p_0}{\log p_0} \geq \max\{\frac{9Cn\tau^2}{\alpha^2}, \frac{4Cn\tau^4}{\beta^2}\}. \tag{5}$$

Then with probability

$$(1-\delta)^{1+5\left(\lceil \log_2 \frac{N}{p_0} \rceil - 1\right)},$$

the following holds for all k:

$$\|\nabla f_{p_{k+1}}(w_{p_k})\|_2 \leq \alpha_{p_{k+1}}, \nabla^2 f_{p_{k+1}}(w_{p_k}) \succeq \beta_{p_{k+1}} I \tag{6}$$

Proof. By the fact that the function value $\frac{\log p}{p}$ decreasing when p increases, and for all $k \in \{0, \cdots, \lceil \log_2 \frac{N}{p_0} \rceil - 1\}$ it holds $p_{k+1} \geq p_k \geq p_0$, we have:

$$\tau\sqrt{\frac{Cn\log p_k}{p_k}} \leq \tau\sqrt{\frac{Cn\log p_0}{p_0}}.$$

Moreover, by the selection of p_0 in (5), we have

$$\tau\sqrt{\frac{Cn\log p_k}{p_k}} \le \tau\sqrt{\frac{Cn\log p_0}{p_0}} \le \frac{\alpha}{3} \;\Rightarrow\; 3\tau\sqrt{\frac{Cn\log p_k}{p_k}} \le \alpha,$$

which gives

$$2\tau\sqrt{\frac{Cn\log p_k}{p_k}} \le \alpha - \tau\sqrt{\frac{Cn\log p_k}{p_k}} \le \alpha - \tau\sqrt{\frac{Cn\log p_{k+1}}{p_{k+1}}} = \alpha_{p_{k+1}}.$$

As a result, for all $0 \le k \le \lceil \log_2 \frac{N}{p_0} \rceil - 1$ that satisfies $\|\nabla f_{p_k}(w_{p_k})\|_2 \le \epsilon(p_k)$, with probability $(1-\delta)^2$, it holds

$$\|\nabla f_{p_{k+1}}(w_{p_k})\|_2 \le 2\tau\sqrt{\frac{Cn\log p_k}{p_k}} \le \alpha_{p_{k+1}}.$$

Hence in total, the probability of the gradient condition holds for all k is:

$$(1-\delta)^{2\left(\lceil \log_2 \frac{N}{p_0} \rceil - 1\right)}.$$

Next, we prove the condition of Hessian by induction. Let the subproblem counter k be the induction variable. When $k = 0$, by the choice of w_{p_0} satisfying (5), we have

$$\|\nabla f_{p_0}(w_{p_0})\|_2 \le \tau\sqrt{\frac{Cn\log p_0}{p_0}} \le \alpha_{p_0},\ \textit{and}\ \lambda_i(\nabla^2 f_{p_0}(w_{p_0})) > 0,\ \forall i = 1, \cdots, n.$$

Since with probability $(1-\delta)$, the f_{p_0} is $(\alpha_{p_0}, \beta_{p_0})$ strong morse function, i.e. all $|\lambda_i(\nabla^2 f_{p_0}(w_{p_0}))| \ge \beta_{p_0}$. Combining with nonnegative eigenvalues, we have

$$\{\lambda_i(\nabla^2 f_{p_0}(w_{p_0})) \ge \beta_{p_0}, \forall i\} \text{ holds with probability } (1-\delta).$$

In summary, with probability $(1-\delta)$, both conditions for ∇f_{p_0} and $\nabla^2 f_{p_0}$ hold.

Suppose at kth subproblem, the inexact solution w_p satisfies (6) and $\nabla^2 f_{p_k}(w_{p_k}) \succeq \beta_{p_k} I$, then for the $(k+1)$th subproblem, by the first part of the proof, we have:

$$\|\nabla f_{p_{k+1}}(w_{p_k})\|_2 \le \alpha_{p_{k+1}} \text{ holds with probability } (1-\delta)^2.$$

Moreover, with probability $(1-\delta)$, the $f_{p_{k+1}}$ is $(\alpha_{p_{k+1}}, \beta_{p_{k+1}})$ strong morse, i.e. $|\lambda_i(\nabla^2 f_{p_{k+1}}(w_{p_k}))| \ge \beta_{p_{k+1}}$ for all i from one to n. Then for all $v \in \mathbb{R}^n/\{0\}$, by the definition of operator norm we have:

$$\begin{aligned} v^T\nabla^2 f_{p_{k+1}}(w_{p_k})v &= v^T\nabla^2 f_{p_k}(w_{p_k})v + v^T\left(\nabla^2 f_{p_{k+1}}(w_{p_k}) - \nabla^2 f_{p_k}(w_{p_k})\right)v \\ &\ge \beta_{p_k}\|v\|_2^2 - \|\nabla^2 f_{p_{k+1}}(w_{p_k}) - \nabla^2 f_{p_k}(w_{p_k})\|_{op}\|v\|_2^2. \end{aligned}$$

By Proposition.3, with probability $(1-\delta)^2$ it holds $\|\nabla^2 f_{p_{k+1}}(w_{p_k}) - \nabla^2 f_{p_k}(w_{p_k})\|_{op} \leq \tau^2\sqrt{\frac{Cn\log p_k}{p_k}}$, we have that with probability $(1-\delta)^2$:

$$\begin{aligned} v^T\nabla^2 f_{p_{k+1}}(w_{p_k})v &\geq \left(\beta_{p_k} - \tau^2\sqrt{\frac{Cn\log p_k}{p_k}}\right)\|v\|_2^2 \\ &= (\beta - 2\tau^2\sqrt{\frac{Cn\log p_k}{p_k}})\|v\|_2^2. \end{aligned}$$

And by the choice of p_0, we have for all $p_k \geq p_0$,

$$\beta - 2\tau^2\sqrt{\frac{Cn\log p_k}{p_k}} \geq 0.$$

Hence $v^T\nabla^2 f_{p_{k+1}}(w_{p_k})v \geq 0$ for all nonzero v. By that $f_{p_{k+1}}$ is strong morse, the $\nabla^2 f_{p_{k+1}}(w_{p_k})$ is P.D. and w_{p_k} is within a neighborhood of a strict local minimal of $f_{p_{k+1}}$.

In summary, the conditions of both gradient and Hessian for $f_{p_{k+1}}$ hold with probability $(1-\delta)^5$. In all, the probability for the conditions to hold for all k is around

$$(1-\delta)^{1+5\left(\lceil\log_2\frac{N}{p_0}\rceil - 1\right)}.$$

□

E Proof for Prop.7

Proof. We first bound the complexity for the subproblem $\{\min_w f_{p_k}(w)\}$. By the Taylor's theorem and the L-Lipschtiz Hessian, we have:

$$\begin{aligned} \|\nabla f_{p_k}(w_{p_k}^{t+1})\|_2 &= \|\nabla f_{p_k}(w_{p_k}^t - \eta_t\nabla f_{p_k}(w_{p_k}^t))\|_2 \\ &\leq \left\|\nabla f_{p_k}(w_{p_k}^t) - \eta_t\nabla^2 f_{p_k}(w_{p_k}^t)\nabla f_{p_k}(w_p^t)\right\|_2 + \frac{L}{2}\eta_t^2\|\nabla f_{p_k}(w_{p_k}^t)\|_2^2 \\ &\leq \left\|I - \eta_t\nabla^2 f_p(w_p^t)\right\|_2\|\nabla f_{p_k}(w_{p_k}^t)\|_2 + \frac{L}{2}\eta_t^2\|\nabla f_{p_k}(w_{p_k}^t)\|_2^2. \end{aligned}$$

Since $\nabla^2 f_{p_k}$ is at least β_{p_k} strongly convex with probability, and that ∇f_{p_k} is M Lipschtiz, we have

$$\beta_{p_k} I \preceq \nabla^2 f_{p_k} \preceq MI.$$

Also, since $\eta_t = \min\{\frac{1}{M}, \frac{\beta_{p_k}}{2L\epsilon(p_{k-1})}\} \leq \frac{1}{M}$, we have

$$\begin{aligned}\left\|I - \eta_t \nabla^2 f_p(w_p^t)\right\|_2 &= \max_i\{|1 - \eta_t \lambda_i(\nabla^2 f_{p_k}(w_p^t))|\} \\ &= \max_i\{1 - \eta_t \lambda_i(\nabla^2 f_{p_k}(w_p^t))\} \\ &\leq 1 - \eta_t \beta_{p_k}.\end{aligned}$$

And

$$\begin{aligned}\|\nabla f_{p_k}(w_{p_k}^{t+1})\|_2 &\leq \left\|I - \eta_t \nabla^2 f_{p_k}(w_{p_k}^t)\right\|_2 \|\nabla f_{p_k}(w_{p_k}^t)\|_2 + \frac{L}{2}\eta_t^2 \|\nabla f_{p_k}(w_{p_k}^t)\|_2^2 \\ &\leq (1 - \eta_t \beta_{p_k})\|\nabla f_{p_k}(w_{p_k}^t)\|_2 + \frac{L}{2}\eta_t^2 \|\nabla f_{p_k}(w_{p_k}^t)\|_2^2 \\ &= \left(1 - \eta_t \beta_{p_k} + \frac{L}{2}\eta_t^2 \|\nabla f_{p_k}(w_{p_k}^t)\|_2\right) \|\nabla f_{p_k}(w_{p_k}^t)\|_2.\end{aligned}$$

Combining with the fact that $\|\nabla f_{p_k}(w_{p_k}^t)\|_2 \leq \|\nabla f_{p_k}(w_{p_k}^0)\|_2 \leq 2\epsilon(p_{k-1})$, we have

$$\|\nabla f_{p_k}(w_{p_k}^{t+1})\|_2 \leq \left(1 - \eta_t \beta_{p_k} + L\eta_t^2 \epsilon(p_{k-1})\right) \|\nabla f_{p_k}(w_{p_k}^t)\|_2.$$

Since $\eta_t = \min\{\frac{1}{M}, \frac{\beta_{p_k}}{2L\epsilon(p_{k-1})}\}$ and combining with the 'warm start' $w_{p_k}^0 = w_{p_{k-1}}$, we have

$$\begin{aligned}\|\nabla f_{p_k}(w_{p_k}^{t+1})\|_2 &\leq \left(1 - \frac{1}{2}\eta_t \beta_{p_k}\right) \|\nabla f_p(w_{p_k}^t)\|_2 \\ &= \left(1 - \frac{\beta_{p_k}}{2}\min\{\frac{1}{M}, \frac{\beta_{p_k}}{2L\epsilon(p_{k-1})}\}\right) \|\nabla f_{p_k}(w_{p_k}^t)\|_2 \\ &\leq \left(1 - \frac{\beta_{p_k}}{2}\min\{\frac{1}{M}, \frac{\beta_{p_k}}{2L\epsilon(p_{k-1})}\}\right)^{k+1} \|\nabla f_{p_k}(w_{p_k}^0)\|_2 \\ &= \left(1 - \min\{\frac{\beta_{p_k}}{2M}, \frac{\beta_{p_k}^2}{4L\epsilon(p_{k-1})}\}\right)^{k+1} \|\nabla f_{p_k}(w_{p_{k-1}})\|_2.\end{aligned}$$

As a result, the number of iterations $T(p_k)$ to exit the subproblem $\{\min_w f_{p_k}\}$ is upper bounded by

$$T(p_k) \leq \log_{\left(1 - \min\{\frac{\beta_{p_k}}{2M}, \frac{\beta_{p_k}^2}{4L\epsilon(p_{k-1})}\}\right)} \frac{\epsilon(p_k)}{\|\nabla f_{p_k}(w_{p_{k-1}})\|_2} = \frac{-\log\left(\frac{\epsilon(p_k)}{\|\nabla f_{p_k}(w_{p_{k-1}})\|_2}\right)}{-\log\left(1 - \min\{\frac{\beta_{p_k}}{2M}, \frac{\beta_{p_k}^2}{4L\epsilon(p_{k-1})}\}\right)}.$$

When $p_{k-1} \geq 2$ and $p_k = 2p_{k-1}$, we have

$$
\begin{aligned}
-\log\left(\frac{\epsilon(p_k)}{\|\nabla f_{p_k}(w_{p_{k-1}})\|_2}\right) &= \log\left(\frac{\|\nabla f_{p_k}(w_{p_{k-1}})\|_2}{\epsilon(p_k)}\right) \leq \log\frac{2\epsilon_{p_{k-1}}}{\epsilon_{p_k}} \\
&= \log\left(2\sqrt{\frac{2\log p_{k-1}}{\log 2p_{k-1}}}\right) \leq \frac{3}{2}\log 2.
\end{aligned}
$$

Here, the last inequality is by the fact that $\frac{\log x}{\log 2x} = \frac{\log x}{\log 2 + \log x} \leq 1$ for all $x > 1$.

For the dominator, by the fact that $\log(1+x) \leq x$ for all $x > -1$, we have

$$
-\log\left(1 - \min\{\frac{\beta_{p_k}}{2M}, \frac{\beta_{p_k}^2}{4L\epsilon(p_{k-1})}\}\right) \geq \min\{\frac{\beta_{p_k}}{2M}, \frac{\beta_{p_k}^2}{4L\epsilon(p_{k-1})}\}.
$$

Combining the above two inequalities with the bound for k, we have

$$
T(p_k) \leq \max\left\{\frac{3M}{\beta_{p_k}}\log 2, \frac{6L\epsilon(p_{k-1})}{\beta_{p_k}^2}\log 2\right\}.
$$

Since by the choice of p_0, we have that $2\epsilon_{p_{k-1}} \leq \alpha_{p_{k-1}} \leq \alpha$, and $\beta_{p_k} \geq \frac{1}{2}\beta$. Hence the above bound can be relaxed into a constant bound

$$
T(p_k) \leq \max\left\{\frac{6M}{\beta}\log 2, \frac{12L\alpha}{\beta^2}\log 2\right\}.
$$

As a result, the total number of gradient evaluations from w_{p_0} to w_N is at most

$$
\begin{aligned}
\sum_{\forall k} \max\left\{\frac{6M}{\beta}\log 2, \frac{12L\alpha}{\beta^2}\log 2\right\} p_k &= \max\left\{\frac{6M}{\beta}\log 2, \frac{12L\alpha}{\beta^2}\log 2\right\} \sum_{\forall p} p \\
&\leq \max\left\{\frac{6M}{\beta}\log 2, \frac{12L\alpha}{\beta^2}\log 2\right\} \sum_{i=0}^{\log_2 N} 2^i.
\end{aligned}
$$

By the summation of the geometric series, we have $\sum_{i=0}^{\log_2 N} 2^i = \frac{2^{\log_2 N}-1}{2-1} \leq N$, the above result is upper bounded by

$$
\max\left\{\frac{6M}{\beta}\log 2, \frac{12L\alpha}{\beta^2}\log 2\right\} N. \tag{7}
$$

□

Since this proposition builds on the foundation of Prop.6 and does not introduce any extra probabilities when analyzing the complexity. Hence the result of this proposition holds with the same probability as in Prop.6, i.e. the (7) holds with probability $(1-\delta)^{1+5\left(\lceil\log_2 \frac{N}{p_0}\rceil - 1\right)}$.

F $p_{k+1} = (1+a)p_k, a > 0$

Suppose the same assumptions in Prop.7 in which $(n, r, \tau, C, \alpha, \beta, p_0)$ is defined. And w_{p_0} satisfy (1):

$$\|\nabla f_{p_0}(w_{p_0})\|_2 \leq \tau\sqrt{\frac{Cn\log p_0}{p_0}} := \epsilon(p_0) \text{ and } \nabla^2 f_{p_0}(w_{p_0}) \succ 0.$$

For any $(p_{k+1}, p_k) \in \mathbb{Z}_+ \times \mathbb{Z}_+$ where $p_{k+1} > p_k$, define $\mathcal{E}$ to be:

$$\mathcal{E}(p_{k+1}, p_k) := \tau\frac{p_{k+1}-p_k}{p_{k+1}}\sqrt{Cn}\left(\sqrt{\frac{\log(p_{k+1}-p_k)}{p_{k+1}-p_k}} + \sqrt{\frac{\log p_k}{p_k}}\right).$$

Then when $p_{k+1} = (1+a)p_k$, we have

$$\mathcal{E}\left((1+a)p_k, p_k\right) = \frac{a}{1+a}\tau\sqrt{Cn}\left(\sqrt{\frac{\log(ap_k)}{ap_k}} + \sqrt{\frac{\log p_k}{p_k}}\right),$$

and it is easy to see that when a is fixed, the function value $\mathcal{E}\left((1+a)p_k, p_k\right)$ decreases when p_k increasing. Moreover, we have the relatonship bettween ϵ and $\mathcal{E}$ being that $\epsilon(p_k) = \tau\sqrt{Cn}\sqrt{\frac{\log p_k}{p_k}} = \mathcal{E}(2p_k, p_k)$.

By universal convergence and triangle inequality, we have that with probability $(1-\delta)^2$, for all $w \in \mathbb{B}^n(r)$ it holds

$$\begin{aligned}
&\|\nabla f_{p_{k+1}}(w) - \nabla f_{p_k}(w)\|_2 \leq \mathcal{E}(p_{k+1}, p_k),\\
&\|\nabla^2 f_{p_{k+1}}(w) - \nabla^2 f_{p_k}(w)\|_{op} \leq \tau\mathcal{E}(p_{k+1}, p_k).
\end{aligned}$$

Hence at any iteration $(k+1)$, on condition that $\|\nabla f_{p_k}(w_{p_k})\|_2 \leq \epsilon(p_k)$, with probability $(1-\delta)^2$ the initial gradient is bounded by

$$\|\nabla f_{p_{k+1}}(w_{p_k})\|_2 \leq \epsilon(p_k) + \mathcal{E}(p_{k+1}, p_k) = \epsilon(p_k) + \mathcal{E}((1+a)p_k, p_k).$$

Note that when $p_{k+1} = (1+a)p_k$, both ϵ and $\mathcal{E}$ are decreasing when p_k increases, while $\alpha_k = \alpha - \epsilon(p_k)$ increases when p_k increases. Hence the worst case happens when $k = 0$, the initial subproblem. Once $\epsilon(p_0) + \mathcal{E}((1+a)p_0, p_0) \leq \alpha_{p_0} \leq \alpha_{p_1}$ holds, then for all k, the gradient $\|\nabla f_{p_{k+1}}(w_{p_k})\|_2 \leq \alpha_{k+1}$. As a result, the first condition is

$$\epsilon(p_0) + \mathcal{E}((1+a)p_0, p_0) \leq \alpha - \epsilon(p_0). \tag{8}$$

Next is for strong convexity. To ensure strongly convex for $f_{k+1}(w)$ at w_{p_k}, on the condition that $\nabla^2 f_{p_k} \succeq \beta_{p_k} I$, we have that for all $v \in \mathbb{R}^n/\{0\}$, it holds:

$$\begin{aligned} v^T \nabla^2 f_{p_{k+1}}(w_{p_k}) v &= v^T \nabla^2 f_{p_k}(w_{p_k}) v + v^T \left(\nabla^2 f_{p_{k+1}}(w_{p_k}) - \nabla^2 f_{p_k}(w_{p_k}) \right) v \\ &\geq \beta_{p_k} \|v\|_2^2 - \|\nabla^2 f_{p_{k+1}}(w_{p_k}) - \nabla^2 f_{p_k}(w_{p_k})\|_{op} \|v\|_2^2. \end{aligned}$$

By the universal convergence theorem, with probability $(1-\delta)^2$ it holds

$$\begin{aligned} v^T \nabla^2 f_{p_{k+1}}(w_{p_k}) v &\geq \left(\beta_{p_k} - \tau \mathcal{E}(p_{k+1}, p_k)\right) \|v\|_2^2 \\ &= \underbrace{\left(\beta_{p_k} - \tau \mathcal{E}((1+a)p_k, p_k)\right) \|v\|_2^2}_{(i)}. \end{aligned}$$

Note that (i) increases when p_k increases, hence for all subproblem counter k we have:

$$\beta_{p_k} - \tau \mathcal{E}((1+a)p_k, p_k) \geq \beta_{p_0} - \tau \mathcal{E}((1+a)p_0, p_0).$$

And requiring p_0 satisfying

$$\beta_{p_0} - \tau \mathcal{E}((1+a)p_0, p_0) = \beta - \tau \epsilon(p_0) - \tau \mathcal{E}((1+a)p_0, p_0) \geq 0 \tag{9}$$

can ensures $(i) \geq 0$ for all k. On condition that $\|\nabla f_{p_{k+1}}\|_2 \leq \alpha_{p_{k+1}}$, and combing with the fact that with probability of $(1-\delta)$ the function $f_{p_{k+1}}$ is $(\alpha_{p_{k+1}}, \beta_{p_{k+1}})$ strongly Morse hence it is $\beta_{p_{k+1}}$ strongly convex at w_{p_k}.

In summary, when p_0 satisfies (8), then for any k such $\|\nabla f_{p_k}(w_{p_k})\|_2 \leq \epsilon(p_k)$, with probability $(1-\delta)^2$ it satisfies $\|\nabla f_{p_{k+1}}\|_2 \leq \alpha_{p_{k+1}}$. In addition, when p_0 satisfies (9), with probability of $(1-\delta)^3$ the point w_{p_k} is within a $\beta_{p_{k+1}}$ strongly convex region of $f_{p_{k+1}}$.

For simplicity, we use $\mathcal{E}(p_k)$ to represent $\mathcal{E}((1+a)p_k, p_k)$ since the latter one only depends on p_k when a is fixed. Similar to Prop.7 and note that $(\mathcal{E}(p_{k-1} + \epsilon(p_{k-1})) \leq \alpha_{p_k} \leq \alpha$, we have the bound

$$\begin{aligned} \|\nabla f_{p_k}(w_{p_k}^{t+1})\|_2 &\leq \left(1 - \min\{\frac{\beta_{p_k}}{2M}, \frac{\beta_{p_k}^2}{2L(\mathcal{E}(p_{k-1} + \epsilon(p_{k-1}))}\}\right)^{k+1} \|\nabla f_{p_k}(w_{p_k}^0)\|_2 \\ &\leq \left(1 - \min\{\frac{\beta_{p_k}}{2M}, \frac{\beta_{p_k}^2}{2L\alpha)}\}\right)^{k+1} \|\nabla f_{p_k}(w_{p_k}^0)\|_2. \end{aligned}$$

And iteration complexity for k is

$$\begin{aligned}
T(p_k) &\leq \log_{\left(1-\min\{\frac{\beta_{p_k}}{2M},\frac{\beta_{p_k}^2}{2L\alpha)}\}\right)} \frac{\epsilon(p_k)}{\|\nabla f_{p_k}(w_{p_k}^0)\|_2} = \frac{-\log\left(\frac{\epsilon(p_k)}{\|\nabla f_{p_k}(w_{p_k}^0)\|_2}\right)}{-\log\left(1-\min\{\frac{\beta_{p_k}}{2M},\frac{\beta_{p_k}^2}{2L\alpha}\}\right)} \\
&= \frac{\log\left(\frac{\|\nabla f_{p_k}(w_{p_k}^0)\|_2}{\epsilon(p_k)}\right)}{-\log\left(1-\min\{\frac{\beta_{p_k}}{2M},\frac{\beta_{p_k}^2}{2L\alpha)}\}\right)} \leq \frac{\log\left(\frac{\epsilon(p_{k-1})+\mathcal{E}(p_{k-1})}{\epsilon(p_k)}\right)}{\min\{\frac{\beta_{p_k}}{2M},\frac{\beta_{p_k}^2}{2L\alpha}\}} \\
&= \max\{\frac{2M}{\beta_{p_k}},\frac{2L\alpha)}{\beta_{p_k}^2}\}\log\left(\frac{\epsilon(p_{k-1})+\mathcal{E}(p_{k-1})}{\epsilon(p_k)}\right).
\end{aligned}$$

Next, we show that $T(p_k)$ can be upper bounded by a constant independent of p_k. To achieve this, we first bound the value $\mathcal{E}((1+a)p,p)$ for $(a,p)\in(0,\infty)\times\mathbb{Z}_+$. We have

$$\begin{aligned}
\frac{1}{\tau\sqrt{Cn}}\mathcal{E}\left((1+a)p,p\right) &= \frac{a}{1+a}\left(\sqrt{\frac{\log(ap)}{ap}}+\sqrt{\frac{\log p}{p}}\right) \\
&= \frac{a}{1+a}\left(\underbrace{\sqrt{\frac{\log(ap)}{a\log p}}}_{(i)}+1\right)\sqrt{\frac{\log p}{p}}.
\end{aligned}$$

And $(i)=\sqrt{\frac{\log a+\log p}{a\log p}}\xrightarrow{p\to\infty}\sqrt{\frac{1}{a}}$. In addition, when $p\geq 2$,

$$\begin{aligned}
(i) &= \sqrt{\frac{\log a+\log p}{a\log p}} = \sqrt{\frac{1}{a}+\frac{\log a}{a}\frac{1}{\log p}} \\
&= \sqrt{\frac{1}{a}+\frac{1}{a}\log_p a} \leq \sqrt{\frac{1}{a}\left(1+\log_2 a\right)}.
\end{aligned}$$

Hence

$$\mathcal{E}\left((1+a)p,p\right) \leq \underbrace{\frac{a+\sqrt{a+a\log_2 a}}{1+a}}_{:=c(a)}\tau\sqrt{\frac{Cn\log p}{p}} = c(a)\epsilon(p).$$

Combining previous bounds gives us

$$\frac{\epsilon(p_{k-1})+\mathcal{E}(p_{k-1})}{\epsilon(p_k)} \le (1+c(a))\frac{\epsilon(p_{k-1})}{\epsilon(p_k)} = (1+c(a))\sqrt{\frac{p_k \log p_{k-1}}{p_{k-1}\log p_k}}$$
$$= (1+c(a))\sqrt{\frac{(1+a)\log p_{k-1}}{\log((1+a)p_{k-1})}} \le (1+c(a))\sqrt{(1+a)}.$$

Moreover, the two conditions for initial p_0 can be relaxed to be

$$(1+c(a))\epsilon(p_0) \le \alpha - \epsilon(p_0) \text{ and}$$
$$\beta - \tau(1+c(a))\epsilon(p_0) \ge 0.$$

The second one gives a bound for β_{p_k} by $\beta_{p_k} \ge \beta_{p_0} = \beta - \tau\epsilon(p_0) \ge \frac{c(a)}{1+c(a)}\beta$. Combining all previous bounds, we have

$$T(p_k) \le \max\{\frac{2(1+c(a))M}{c(a)\beta}, \frac{2(1+c(a))^2 L\alpha}{c(a)^2\beta^2}\}\log\left((1+c(a))\sqrt{(1+a)}\right).$$

Moreover,

$$\sum_{k=1}^{\log_{1+a} N}(1+a)^k = a\frac{N-1}{a}.....$$

Hence the total number of individual gradient evaluations is bounded by

$$\max\{\frac{2(1+c(a))M}{c(a)\beta}, \frac{2(1+c(a))^2 L\alpha}{c(a)^2\beta^2}\}\log\left((1+c(a))\sqrt{(1+a)}\right)N.$$

G One shot sampling complexity.

Suppose the same assumptions in Prop.7 in which $(n, r, \tau, C, \alpha, \beta, p_0)$ is defined. And w_{p_0} satisfy (1):

$$\|\nabla f_{p_0}(w_{p_0})\|_2 \le \tau\sqrt{\frac{Cn\log p_0}{p_0}} := \epsilon(p_0) \text{ and } \nabla^2 f_{p_0}(w_{p_0}) \succ 0.$$

We have that with probability $(1-\delta)^2$, for all $w \in \mathbb{B}^n(r)$ that

$$\|\nabla f_{p_0}(w) - \nabla f_N(w)\|_2 \le \tau\frac{N-p_0}{N}\sqrt{Cn}\left(\sqrt{\frac{\log(N-p_0)}{N-P_0}} + \sqrt{\frac{\log p_0}{p_0}}\right) := \mathcal{E}(N, p_0),$$
$$\|\nabla^2 f_{p_0}(w) - \nabla^2 f_N(w)\|_{op} \le \tau\mathcal{E}(N, p_0).$$

Hence with probability $(1-\delta)^2$ it holds

$$\|f_N(w_{p_0})\|_2 \leq \mathcal{E}(N, p_0) + \epsilon(p_0).$$

As a result, viewing N as fixed and solving the following inequality for p_0:

$$\mathcal{E}(N, p_0) + \epsilon(p_0) \leq \alpha_N \leq \alpha_{p_0} = \alpha - \epsilon(p_0) \tag{10}$$

gives the first conditions for p_0.

Next is for strong convexity. Similar to the derivation in Appendix.F, another condition for p_0 is given by

$$\beta_{p_0} - \tau\mathcal{E}(N, p_0) = \beta - \tau\epsilon(p_0) - \tau\mathcal{E}(N, p_0) \geq 0. \tag{11}$$

And with another probability of $(1-\delta)$ that f_N is (α_N, β_N) strongly Morse hence β_N strongly convex at w_{p_0}. Recall that in Appendix.F, when $p_{k+1} = (1+a)p_k$, we have $\mathcal{E}(p_{k+1}, p_k) \leq c(a)\epsilon(p_k)$ where $c(a) = \frac{a+\sqrt{a+a\log_2 a}}{1+a}$. As long as $a \geq \frac{1}{2}$ then c is well defined and we have

$$\max_{a \geq \frac{1}{2}} |c(a)| \leq \frac{8}{5}.$$

Let $p_k = p_0$ and $p_1 = N = (1 + \frac{N-p_0}{p_0})p_0$, we have

$$\mathcal{E}(N, p_0) \leq c(\frac{N-p_0}{p_0})\epsilon(p_0) \leq \frac{8}{5}\epsilon(p_0).$$

Replace the relaxed bound for $\mathcal{E}$ in (10,11), we have explicit and more strict conditions for p_0:

$$\begin{cases} \frac{8}{5}\epsilon(p_0) + \epsilon(p_0) \leq \alpha - \epsilon(p_0) \\ \beta - \tau\epsilon(p_0) - \frac{8}{5}\tau\epsilon(p_0) \geq 0. \end{cases} \Rightarrow \epsilon(p_0) = \tau\sqrt{\frac{Cn\log p_0}{p_0}} \leq \min\left\{\frac{5\alpha}{18}, \frac{5\beta}{13\tau}\right\}. \tag{12}$$

In summary, when p_0 satisfies (10,11), with probability of $(1-\delta)^5$ the point w_{p_0} is within a β_N strongly convex region of f_N.

In terms of complexity, following a similar gradient descent

$$w_N^{t+1} = w_N^t - \eta_1 \nabla f_N(w_N^t), \text{ where } \eta_1 = \min\{\frac{1}{M}, \frac{\beta_N}{L\alpha_N}\},$$

and similar arguments in Appendix.F, we have the number of iterations is bounded by

$$T(N) \leq \log_{\left(1-\min\{\frac{\beta_N}{2M}, \frac{\beta_N^2}{2L\alpha_N}\}\right)} \frac{\epsilon(N)}{\|\nabla f_N(w_{p_0})\|_2} = \frac{-\log\left(\frac{\epsilon(N)}{\|\nabla f_N(w_{p_0})\|_2}\right)}{-\log\left(1-\min\{\frac{\beta_N}{2M}, \frac{\beta_N^2}{2L\alpha_N}\}\right)}.$$

Similarly, by the termination condition and universal convergence theorem, we have $\|\nabla f_N(w_{p_0})\|_2 \leq \epsilon(p_0) + \mathcal{E}(N, p_0) \leq \frac{13}{5}\epsilon(p_0)$, we have

$$\begin{aligned} -\log\left(\frac{\epsilon(N)}{\|\nabla f_N(w_{p_0})\|_2}\right) &= \log\left(\frac{\|\nabla f_N(w_{p_0})\|_2}{\epsilon(N)}\right) \leq \log\frac{13\epsilon(p_0)}{5\epsilon(N)} \\ &= \log\left(\min\left\{\frac{13\alpha}{18}, \frac{\beta}{\tau}\right\}\sqrt{\frac{N}{\log N}}\right). \end{aligned}$$

For the dominator, by the fact that $\log(1+x) \leq x$ for all $x > -1$ and $\alpha_N \leq \alpha$, we have

$$-\log\left(1 - \min\{\frac{\beta_N}{2M}, \frac{\beta_N^2}{2L\alpha_N}\}\right) \geq \min\{\frac{\beta_N}{2M}, \frac{\beta_N^2}{2L\alpha_N}\} \geq \min\{\frac{\beta_N}{2M}, \frac{\beta_N^2}{2L\alpha}\}.$$

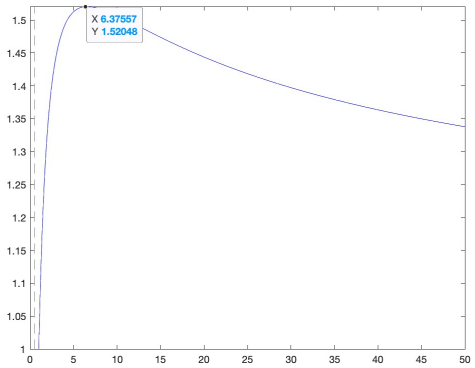
Moreover, we have

$$\beta_N \geq \beta_{p_0} = \beta - \tau\epsilon(p_0) \geq \frac{8}{13}\beta.$$

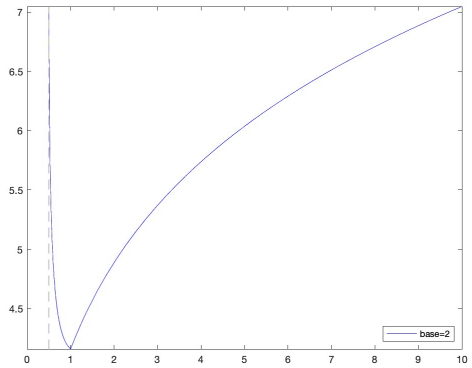
Combining the above three bounds, we have the total number of individual gradient evaluations is bounded by

$$T(N) \leq N\log\left(\min\left\{\frac{13\alpha}{18}, \frac{\beta}{\tau}\right\}\sqrt{\frac{N}{\log N}}\right) \cdot \max\left\{\frac{13M}{4\beta}, \frac{169L\alpha}{32\beta^2}\right\}.$$

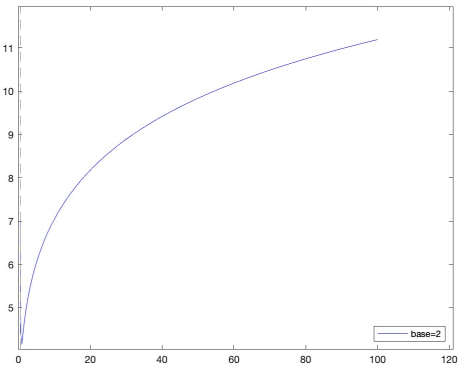
H extra



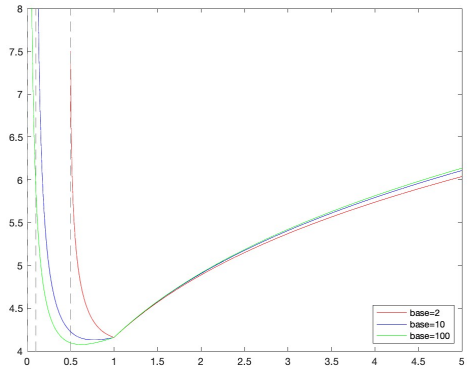
(a) function c.



(b) function g



(c) function g



(d) function g

Figure 4: figures for c and g, where c is defined previously and $g(a) = \max\{\frac{2(1+c(a))M}{c(a)\beta}, \frac{2(1+c(a))^2 L\alpha}{c(a)^2\beta^2}\} \cdot \log\left((1+c(a))\sqrt{(1+a)}\right)$ where $(L, M, \alpha, \beta) = (1, 1, 0.5, 1)$. The base of g: g depends on c, the function c has a term $\log_{p_k} a$ which is strengthened by $\log_b a$if we assume $p \geq b$.

I 1.

Optimization problem

$$\min_x f(x) = \mathbb{E}_\xi[F(x;\xi)]$$
$$s.t.\ c(x) = \mathbb{E}_\xi[C(x;\xi)] = 0$$

with

Notation	Meaning
x	in $\mathbb{R}^n$, optimization variable
ξ	distribution of data
$F(x;\xi)$	in $\mathbb{R}$, loss for data ξ with variable x
$C(x;\xi)$	in $\mathbb{R}^m$, constraints value for ξ and x
$f_N(x)$	$f_N(x) = \frac{1}{N}\sum_{i=1}^N F(x;\xi_i)$
$c_N(x)$	$c_N(x) = \frac{1}{N}\sum_{i=1}^N C(x;\xi_i)$
$A_N(x)$	in $\mathbb{R}^{m\times n}$, Jacobian of $C_N(x)$

Following Oztoprak, Byrd, and Nocedal 2023, the modified SQP with bounded noise, take

Notation	Meaning
H_k	$H_k = \beta_k I, (\beta_k > 0)$, Hessian in SQP objective
y_k	$y_k = (A_N(x_k)A_N(x_k)^T)^{-1}A_N(x_k)\nabla f_N(x_k)$
$P_N(x)$	$P_N = I - A_N^T(A_N A_N^T)^{-1}A_N$, projection matrix.
τ	in $(0,1)$, fixed constant
$\pi_{k,N}$	in $\mathbb{R}$, penalty parameter, $\pi_{k,N} = \begin{cases} \pi_{k-1,N}\ if\ \frac{\|y_k\|_\infty}{1-\tau} \le \pi_{k-1,N} \\ \frac{2}{1-\tau}\|y_k\|_\infty,\ else. \end{cases}$
Δ_N^M	change from sample N to M. E.g. $\Delta_N^M f(x_k) = f_M(x_k) - f_N(x_k)$

Optimality criterion at x_k with N samples:

$$\psi_N(x_k) := \frac{1}{\beta_k}\|P_N(x_k)\nabla f_N(x_k)\|_2^2 + \tau\pi_{k,N}\|c_N(x_k)\|_1.$$

The algorithm changes the sample set to M after x_k, then

$$\psi_M(x_k) := \frac{1}{\beta_k}\|P_M(x_k)\nabla f_M(x_k)\|_2^2 + \tau\pi_{k,M}\|c_M(x_k)\|_1.$$

Changes in $P(x_k)\nabla f(x_k)$:

$$\begin{aligned}\Delta_N^M\{P(x_k)\nabla f(x_k)\} = &\left[\Delta_N^M P(x_k)\right]\nabla f_N(x_k)\\ &+ P_N(x_k)\left[\Delta_N^M \nabla f(x_k)\right]\\ &+ \left[\Delta_N^M P(x_k)\right]\left[\Delta_N^M \nabla f(x_k)\right].\end{aligned}$$

Changes in $\pi_k\|c(x_k)\|_1$:

$$\begin{aligned}\Delta_N^M\{\pi_k\|c(x_k)\|_1\} = &\left[\Delta_N^M \pi_k(x_k)\right]\|c_N(x_k)\|_1\\ &+ \pi_{k,N}(x_k)\left[\Delta_N^M\|c(x_k)\|_1\right]\\ &+ \left[\Delta_N^M \pi_k(x_k)\right]\left[\Delta_N^M\|c(x_k)\|_1\right].\end{aligned}$$

Question:

- What kinds of assumptions on constraints shall we make, i.e. bounded variance? Do the objective and constraints follow the same distribution ξ?
- Should the objective and constraints take the same samples?
- Are the samples containing previous ones or just take new sampling?
- Methods for SQP complexity?
- Sample complexity. Need more specific assumptions?

J Adaptive sampling for stochastic objective and deterministic constraint

- Xie et al. 2024: Composite objective – proximal gradient type method – sample size is nondecreasing but is resampled – linear convergence for strongly convex objective, sublinear(1/k) convergence for convex objective. (stochastics: only assumes on expectation of gradient, uses inequalities on general probabilistic and expresses the rest of things in terms of expectation. In practice replace expectation with average.)
- Bollapragada et al. 2023: Augmented Lagrangian type method – samples are resampled – quadratic convergence for convex objective. – complexity $\epsilon^{-3-\delta}$ (stochastics: assumptions on first/second momentum. No explicit assumptions on function form).

- Beiser et al. 2023: Augmented Lagrangian type – samples are resampled – (stochastics: assumptions on first/second momentum. No explicit assumptions on function form)

- Berahas, Bollapragada, and Zhou 2022: SQP – only requires gradient estimate more and more accurate – complexity worst at ϵ^{-2} (stochastics: assumptions on first and second momentum of gradient, also assumptions on quantities of exact and inexact terms)

How does linear/quadratic convergence link with complexity result?

What their paper is, especially on assumptions on functions and solving subproblems and expanding working sample size.

What other papers are, especially on assumptions such as projected gradient test, and norm test?

What assumptions should we make?

References

Beiser, Florian et al. (2023). "Adaptive sampling strategies for risk-averse stochastic optimization with constraints". In: *IMA Journal of Numerical Analysis* 43.6, pp. 3729–3765.

Berahas, Albert S, Raghu Bollapragada, and Baoyu Zhou (2022). "An adaptive sampling sequential quadratic programming method for equality constrained stochastic optimization". In: *arXiv preprint arXiv:2206.00712*.

Bollapragada, Raghu et al. (2023). "An adaptive sampling augmented Lagrangian method for stochastic optimization with deterministic constraints". In: *Computers & Mathematics with Applications* 149, pp. 239–258.

Mei, Song, Yu Bai, and Andrea Montanari (2018). "The landscape of empirical risk for nonconvex losses". In: *The Annals of Statistics* 46.6A, pp. 2747–2774.

Mokhtari, Aryan, Asuman Ozdaglar, and Ali Jadbabaie (2019). "Efficient nonconvex empirical risk minimization via adaptive sample size methods". In: *The 22nd International Conference on Artificial Intelligence and Statistics*. PMLR, pp. 2485–2494.

Oztoprak, Figen, Richard Byrd, and Jorge Nocedal (2023). "Constrained optimization in the presence of noise". In: *SIAM Journal on Optimization* 33.3, pp. 2118–2136.

Xie, Yuchen et al. (2024). "Constrained and composite optimization via adaptive sampling methods". In: *IMA Journal of Numerical Analysis* 44.2, pp. 680–709.