Gradient Projection Iterative Sketch for large-scale constrained Least-squares (Supplementary materials)

Junqi Tang [1] **Mohammad Golbabaee** [1] **Mike E. Davies** [1]

1. Supplementary materials

1.1. the proof for Theorem 1

Proof. At first we denote the underlying cost function of GPIS as $f_t(x)$:

for $t = 0$, we have the cost function of the classical sketch (CS):

$$f_t(x) := \frac{1}{2}\|Sy - SAx\|_2^2, \tag{1}$$

for $t = 1, 2, ..., N$ we have the the cost function of Iterative Hessian Sketch (IHS):

$$f_t(x) = \frac{1}{2}\|S^{t+1}A(x - x^t)\|_2^2 - mx^T A^T(y - Ax^t), \tag{2}$$

and then we denote the optimal solution of f_t constrained to set $\mathcal{K}$ as $x_\star^t$ and $\|r_{i+1}^t\|_2 = \|x_{i+1}^t - x_\star^t\|_2$ have:

$$\|r_{i+1}^t\|_2 = \|x_{i+1}^t - x_\star^t\|_2 = \|\mathcal{P}_\mathcal{K}(x_i^t - \eta\nabla f(x_i)) - x_\star^t\|_2 \tag{3}$$

then we denote cone $\mathcal{C}_t$ to be the smallest close cone at $x_\star^t$ containing the set $\mathcal{K} - x_\star^t$, again because of the distance preservation of translation by Lemma 6.3 of (Oymak et al., 2015), we have:

$$\begin{aligned}\|r_{i+1}^t\|_2 &= \|\mathcal{P}_{\mathcal{K}-x_\star^t}(x_i^t - \eta\nabla f(x_i) - x_\star^t)\|_2 \\ &= \sup_{v\in\mathcal{C}_t\cap\mathcal{B}^d}\left\{v^T(x_i - x_\star^t - \mu\nabla f(x_i))\right\},\end{aligned} \tag{4}$$

then because of the optimality condition on the constrained LS solution $x_\star^t$, we have:

$$\begin{aligned}\|r_{i+1}^t\|_2 &= \sup_{v\in\mathcal{C}_t\cap\mathcal{B}^d}\left\{v^T(x_i - x_\star^t - \eta\nabla f(x_i))\right\} \\ &\le \sup_{v\in\mathcal{C}_t\cap\mathcal{B}^d}\left\{v^T(x_i - x_\star^t - \eta\nabla f(x_i)) + \eta v^T\nabla f(x_\star^t)\right\} \\ &= \sup_{v\in\mathcal{C}_t\cap\mathcal{B}^d}\left\{v^T(x_i - x_\star^t) - \eta v^T(\nabla f(x_i) - \nabla f(x_\star^t))\right\} \\ &= \sup_{v\in\mathcal{C}_t\cap\mathcal{B}^d}\left\{v^T(I - \eta A^T S^T SA)r_i^t\right\} \\ &\le \sup_{u,v\in\mathcal{C}_t\cap\mathcal{B}^d}\left\{v^T(I - \eta A^T S^T SA)u\right\}\|r_i^t\|_2 \\ &\le \sup_{u,v\in\mathcal{B}^d}\left\{v^T(I - \eta A^T S^T SA)u\right\}\|r_i^t\|_2,\end{aligned} \tag{5}$$

We denote:

$$\alpha_t = \sup_{u,v\in\mathcal{B}^d} v^T(I - \eta A^T S^T SA)u, \tag{6}$$

then by recursive subsitution we have:

$$\|r_{i+1}^t\|_2 \le \alpha_t^i\|r_0^t\|_2, \tag{7}$$

and suppose we run GPIHS inner loop k_t time, we have:

$$\|r_{k_t+1}^t\|_2 \le \{\alpha_t\}^{k_t}\|r_0^t\|_2, \tag{8}$$

and we transfer it in terms of A-norm:

$$\|r_{k_t+1}^t\|_A \le \{\alpha_t\}^{k_t}\sqrt{\frac{L}{\mu}}\|r_0^t\|_A. \tag{9}$$

From the main theorems of the Classical sketch (Pilanci & Wainwright, 2015) and Iterative Hessian Sketch (Pilanci & Wainwright, 2016) we have following relationships:

$$\|x_\star^0 - x^\star\|_A \le 2\rho_0\|Ax^\star - y\|_2 = 2\rho_0\|e\|_2, \tag{10}$$

and,

$$\|x_\star^t - x^\star\|_A \le \rho_t\|x_0^t - x^\star\|_A. \tag{11}$$

Then by triangle inequality we have:

$$\|x_0^1 - x^\star\|_A \le \|x_0^1 - x_\star^0\|_A + 2\rho_0\|e\|_2, \tag{12}$$

and,

$$\|x_0^{t+1} - x^\star\|_A \le \|x_0^{t+1} - x_\star^t\|_A + \rho_t\|x_0^t - x^\star\|_A. \tag{13}$$

[1]Institute of Digital Communications, the University of Edinburgh, EH9 3JL, UK. Correspondence to: Junqi Tang <J.Tang@ed.ac.uk>.

Supplementary materials. Proceedings of the 34^{th} International Conference on Machine Learning, Sydney, Australia, 2017. JMLR: W&CP.

Then for $t = 0$ we can have:

$$\begin{aligned}\|x_0^1 - x^\star\|_A &\leq \|x_0^1 - x_\star^0\|_A + 2\rho_0\|e\|_2 \\ &\leq \{\alpha_t\}^{k_t}\sqrt{\frac{L}{\mu}}\|x_0^0 - x_\star^0\|_A + 2\rho_0\|e\|_2,\end{aligned} \tag{14}$$

for $t = 1, 2, ..., N$ we have:

$$\begin{aligned}&\|x_0^t - x^\star\|_A \\ &\leq \|x_0^t - x_\star^{t-1}\|_A + \rho_t\|x_0^{t-1} - x^\star\|_A \\ &\leq \{\alpha_t\}^{k_t}\sqrt{\frac{L}{\mu}}\|x_0^{t-1} - x_\star^{t-1}\|_A \\ &+ \rho_t\|x_0^{t-1} - x^\star\|_A \\ &\leq \left\{\{\alpha_t\}^{k_t}\left((1+\rho_t)\sqrt{\frac{L}{\mu}}\right) + \rho_t\right\}\|x_0^{t-1} - x^\star\|_A,\end{aligned} \tag{15}$$

The last inequality holds because:

$$\begin{aligned}\|x_0^{t-1} - x_{f_{N-1}}^\star\|_A &\leq \|x_0^{t-1} - x^\star\|_A + \|x_\star^{t-1} - x^\star\|_A \\ &\leq \{1+\rho_t\}\|x_0^{t-1} - x^\star\|_A,\end{aligned} \tag{16}$$

Then we denote:

$$\rho_t^\star = \{\alpha_t\}^{k_t}\left((1+\rho_t)\sqrt{\frac{L}{\mu}}\right) + \rho_t \tag{17}$$

and do recursive substitution we can have:

$$\|x_0^t - x^\star\|_A \leq \left\{\prod_{t=1}^{N}\rho_t^\star\right\}\|x_0^1 - x^\star\|_A. \tag{18}$$

hence we finish the proof of Theorem 1. □

1.2. The proofs for Theorem 2 and 3

Proof. From the theory of the Classical sketch and Iterative Hessian Sketch we have following relationships:

$$\|x_\star^0 - x^\star\|_A \leq 2\rho_0\|Ax^\star - y\|_2 = 2\rho_0\|e\|_2, \tag{19}$$

and,

$$\|x_\star^t - x^\star\|_A \leq \rho_t\|x_0^t - x^\star\|_A. \tag{20}$$

Then by triangle inequality we have:

$$\|x_0^1 - x^\star\|_A \leq \|x_0^1 - x_\star^0\|_A + 2\rho_0\|e\|_2, \tag{21}$$

and,

$$\|x_0^{t+1} - x^\star\|_A \leq \|x_0^{t+1} - x_\star^t\|_A + \rho_t\|x_0^t - x^\star\|_A. \tag{22}$$

The remaining task of this proof is just bound the term $\|x_0^{t+1} - x_\star^t\|_A$ for both GPIS and Acc-GPIS algorithm and then chain it. For all the sketched objective function $f_t(x)$, $t = 0, 1, ..., N$, and any pair of vectors $x, x' \in \mathcal{K}$ we have:

$$f_t(x) - f_t(x') - < \nabla f_t(x'), x - x' >= \|S^tA(x - x')\|_2^2 \tag{23}$$

If we set $x' = x_\star^t$, by first order optimality condition we immediately have:

$$\begin{aligned}&f_t(x) - f_t(x_\star^t) \geq \|S^tA(x - x_\star^t)\|_2^2 \\ = \quad &\|S^t\frac{A(x - x_\star^t)}{\|A(x - x_\star^t)\|_2}\|A(x - x_\star^t)\|_2\|_2^2 \\ \geq \quad &\left\{\inf_{v\in range(A)\cap\mathbb{S}^{n-1}}\|S^tv\|_2^2\right\}\|x - x_\star^t\|_A^2,\end{aligned} \tag{24}$$

so we have:

$$\|x - x_\star^t\|_A \leq \frac{\sqrt{f_t(x) - f_t(x_\star^t)}}{\inf_{v\in range(A)\cap\mathbb{S}^{n-1}}\|S^tv\|_2}, \tag{25}$$

From the convergence theory in (Beck & Teboulle, 2009) which the authors in their Remark 2.1 have stated to hold for convex constrained sets, for GPIS inner iterates we have:

$$f_t(x_k) - f_t(x_\star^t) \leq \frac{\beta LR\sup_{v\in range(A)\cap\mathbb{S}^{n-1}}\|S^tv\|_2^2}{2k}, \tag{26}$$

and for Acc-GPIS inner loop we have:

$$f_t(x_k) - f_t(x_\star^t) \leq \frac{2\beta LR\sup_{v\in range(A)\cap\mathbb{S}^{n-1}}\|S^tv\|_2^2}{(k+1)^2}, \tag{27}$$

hence for GPIS:

$$\|x_0^{t+1} - x_\star^t\|_A \leq \sqrt{\frac{\beta L\sigma_t R}{2k}}, \tag{28}$$

for Acc-GPIS,

$$\|x_0^{t+1} - x_\star^t\|_A \leq \sqrt{\frac{2\beta L\sigma_t R}{(k+1)^2}}, \tag{29}$$

Then by simply towering the inequalities we shall obtain the desired results in Theorem 2 and 3. □

1.3. The proofs for quantitative bounds of α_t, ρ_t and σ_t for Gaussian sketches

To prove the results in Proposition 1, 2 and 3 we need the following concentration lemmas as pillars:

Lemma 1. *For any $g \in \mathcal{R}^d$, we have:*

$$\sup_{v\in\mathcal{C}\cap\mathcal{B}^d} v^Tg = \max\left\{0, \sup_{u\in\mathcal{C}\cap\mathbb{S}^{d-1}} u^Tg\right\} \tag{30}$$

Proof. By the definition of cone projection operator we have:

$$\sup_{v\in\mathcal{C}\cap\mathcal{B}^d} v^T g = \|\mathcal{P}_{\mathcal{C}}(g)\|_2 \geq 0 \tag{31}$$

if $\sup_{v\in\mathcal{C}\cap\mathcal{B}^d} v^T g > 0$:

$$\sup_{v\in\mathcal{C}\cap\mathcal{B}^d} v^T g = \sup_{v\in\mathcal{C}\cap\mathcal{B}^d} \|v\|_2 \frac{v^T g}{\|v\|_2} \leq \sup_{u\in\mathcal{C}\cap\mathbb{S}^{d-1}} u^T g, \tag{32}$$

and meanwhile since $\mathcal{C}\cap\mathbb{S}^{d-1} \in \mathcal{C}\cap\mathcal{B}^d$ we have:

$$\sup_{v\in\mathcal{C}\cap\mathcal{B}^d} v^T g \geq \sup_{u\in\mathcal{C}\cap\mathbb{S}^{d-1}} u^T g, \tag{33}$$

hence we have:

$$\sup_{v\in\mathcal{C}\cap\mathcal{B}^d} v^T g = \sup_{u\in\mathcal{C}\cap\mathbb{S}^{d-1}} u^T g, \tag{34}$$

□

Lemma 2. *If* $\sup_{u,v\in\mathcal{C}\cap\mathcal{B}^d} v^T M u > 0$*, we have:*

$$\sup_{u,v\in\mathcal{C}\cap\mathcal{B}^d} v^T M u = \sup_{u,v\in\mathcal{C}\cap\mathbb{S}^{d-1}} v^T M u \tag{35}$$

Proof. Since $u, v \in \mathcal{C}\cap\mathcal{B}^d$, $\|u\|_2$ and $\|v\|_2$ are both less than or equal to 1, we can have the following upper bound:

$$\begin{aligned}\sup_{u,v\in\mathcal{C}\cap\mathcal{B}^d} v^T M u &= \sup_{u,v\in\mathcal{C}\cap\mathcal{B}^d} (\frac{v^T M u}{\|v\|_2\|u\|_2})\|v\|_2\|u\|_2 \\ &\leq \sup_{u,v\in\mathcal{C}\cap\mathbb{S}^{d-1}} v^T M u,\end{aligned}$$

and meanwhile since $\mathcal{C}\cap\mathbb{S}^{d-1} \in \mathcal{C}\cap\mathcal{B}^d$ we have:

$$\sup_{u,v\in\mathcal{C}\cap\mathcal{B}^d} v^T M u \geq \sup_{u,v\in\mathcal{C}\cap\mathbb{S}^{d-1}} v^T M u, \tag{36}$$

hence we have:

$$\sup_{u,v\in\mathcal{C}\cap\mathcal{B}^d} v^T M u = \sup_{u,v\in\mathcal{C}\cap\mathbb{S}^{d-1}} v^T M u \tag{37}$$

□

Lemma 3. *If the entries of the sketching matrix* S *is i.i.d drawn from Normal distribution and* $v \in \mathcal{C}$*, we have:*

$$\|SAv\|_2 \geq \sqrt{\mu}(b_m - \mathcal{W} - \theta)\|v\|_2, \tag{38}$$

$$\|SAv\|_2 \leq \sqrt{L}(b_m + \mathcal{W} + \theta)\|v\|_2, \tag{39}$$

with probability at least $1 - e^{-\frac{\theta^2}{2}}$*.* ($b_m = \sqrt{2}\frac{\Gamma(\frac{m+1}{2})}{\Gamma(\frac{m}{2})} \approx \sqrt{m}$, $\mathcal{W} := \mathcal{W}(A\mathcal{C}\cap\mathbb{S}^{n-1})$)

Proof. This Lemma follows the result of the simplified form of the Gordon's Lemma [Lemma 6.7](Oymak et al., 2015):

$$\begin{aligned}\|SAv\|_2 &\geq (b_m - \mathcal{W}(A\mathcal{C}\cap\mathbb{S}^{n-1}) - \theta)\|Av\|_2 \\ &\geq \sqrt{\mu}(b_m - \mathcal{W}(A\mathcal{C}\cap\mathbb{S}^{n-1}) - \theta)\|v\|_2\end{aligned}$$

$$\begin{aligned}\|SAv\|_2 &\leq (b_m + \mathcal{W}(A\mathcal{C}\cap\mathbb{S}^{n-1}) + \theta)\|Av\|_2 \\ &\leq \sqrt{L}(b_m + \mathcal{W}(A\mathcal{C}\cap\mathbb{S}^{n-1}) + \theta)\|v\|_2\end{aligned}$$

□

1.3.1. THE PROOF FOR PROPOSITION 1

Proof. Let's mark out the feasible region of the step-size η:

$$\begin{aligned}&\alpha(\eta, S^t A) \\ =\ &\sup_{u,v\in\mathcal{B}^d} v^T(I - \eta A^T S^T S A)u \\ \geq\ &\sup_{v\in\mathcal{B}^d} v^T(I - \eta A^T S^T S A)v \\ =\ &\sup_{v\in\mathcal{B}^d} (\|v\|_2^2 - \eta\|SAv\|_2^2) \\ \geq\ &\sup_{v\in\mathcal{B}^d} ((1 - \eta L(b_m + \sqrt{d} + \theta - \epsilon)^2)\|v\|_2^2),\end{aligned}$$

so if we choose a step size $\eta \leq \frac{1}{L(b_m+\sqrt{d}+\theta)^2}$ we can ensure that with probability $1 - e^{-\frac{(\theta-\epsilon)^2}{2}} (\epsilon > 0)$ we have $\alpha(\eta, S^t A) > 0$ and the Lemma 2 become applicable:

$$\begin{aligned}&\alpha(\eta, S^t A) \\ =\ &\sup_{u,v\in\mathcal{B}^d} v^T(I - \eta A^T S^T S A)u \\ =\ &\sup_{u,v\in\mathbb{S}^{d-1}} v^T(I - \eta A^T S^T S A)u \\ =\ &\sup_{u,v\in\mathbb{S}^{d-1}} \frac{1}{4}[(u+v)^T(I - \eta A^T S^T S A)(u+v) \\ &-(u-v)^T(I - \eta A^T S^T S A)(u-v)] \\ =\ &\sup_{u,v\in\mathbb{S}^{d-1}} \frac{1}{4}[\|u+v\|_2^2 - \eta\|SA(u+v)\|_2^2 \\ &-\|u-v\|_2^2 + \eta\|SA(u-v)\|_2^2] \\ \leq\ &\sup_{u,v\in\mathbb{S}^{d-1}} \frac{1}{4}[(1 - \eta\mu(b_m - \sqrt{d} - \theta)^2)\|u+v\|_2^2 \\ &+(\eta L(b_m + \sqrt{d} + \theta)^2 - 1)\|u-v\|_2^2]\end{aligned}$$

The last line of inquality holds with probability at least $1 - 2e^{-\frac{\theta^2}{2}}$ according to Lemma 3. Then since we have set $\eta \leq \frac{1}{L(b_m+\sqrt{d}+\theta+\epsilon)^2}$, and meanwhile notice the fact that $\|u + v\|_2^2 \leq 4$ we have:

$$\begin{aligned}&\alpha(\eta, S^t A) \\ \leq\ &\sup_{u,v\in\mathbb{S}^{d-1}} \frac{1}{4}(1 - \eta\mu(b_m - \sqrt{d} - \theta)^2\|u+v\|_2^2 \\ \leq\ &(1 - \eta\mu(b_m - \sqrt{d} - \theta)^2)\end{aligned}$$

If we chose $\eta = \frac{1}{L(b_m+\sqrt{d}+\theta)^2}$ we have:

$$\alpha(\eta, S^t A) \leq \left(1 - \frac{\mu}{L}\frac{(b_m - \sqrt{d} - \theta)^2}{(b_m + \sqrt{d} + \theta)^2}\right), \tag{40}$$

Then let $\epsilon \to 0$, we shall get the result shown in Proposition 1. □

1.3.2. The proof for Proposition 2

Proof. Recall that ρ_t is defined as:

$$\rho(S^t, A) = \frac{\sup_{v \in A\mathcal{C}\cap\mathbb{S}^{n-1}} v^T(\frac{1}{m}S^{t^T}S^t - I)z}{\inf_{v\in A\mathcal{C}\cap\mathbb{S}^{n-1}} \frac{1}{m}\|S^t v\|_2^2}, \tag{41}$$

we start by lower-bounding the denominator, by simplified Gordon's lemma [Lemma 6.7](Oymak et al., 2015) we directly have:

$$\inf_{v\in A\mathcal{C}\cap\mathbb{S}^{n-1}} \frac{1}{m}\|Sv\|_2^2 \geq \frac{(b_m - \mathcal{W} - \theta)^2}{m}, \tag{42}$$

with probability at least $(1 - e^{-\frac{\theta^2}{2}})$.Then we move to the upper bound for the numerator:

$$\begin{aligned} &v^T\left(\frac{S^{t^T}S^t}{m} - I\right)z \\ &= \frac{1}{4}\{(v+z)^T(\frac{S^{t^T}S^t}{m} - I)(v+z) \\ &- (v-z)^T(\frac{S^{t^T}S^t}{m} - I)(v-z)\} \\ &= \frac{1}{4}\{\frac{1}{m}\|S^t(v+z)\|_2 - \|v+z\|_2 \\ &+ \|v-z\|_2 - \frac{1}{m}\|S^t(v-z)\|_2\}, \end{aligned} \tag{43}$$

and,

$$\begin{aligned} \mathcal{W}(A\mathcal{C}\cap\mathbb{S}^{n-1} - z) &= \mathbb{E}_g(\sup_{v\in A\mathcal{C}\cap\mathbb{S}^{n-1}} g^T(v-z)) \\ &= \mathbb{E}_g(g^T z + \sup_{v\in A\mathcal{C}\cap\mathbb{S}^{n-1}} v^T g) \\ &= \mathcal{W}(A\mathcal{C}\cap\mathbb{S}^{n-1}) \end{aligned} \tag{44}$$

hence we have the following by [Lemma 6.8](Oymak et al., 2015):

$$\begin{aligned} &v^T\left(\frac{S^{t^T}S^t}{m} - I\right)z \\ &\leq \frac{1}{4}\left\{\frac{1}{m}(b_m\|v+z\|_2 + \mathcal{W} + \theta)^2 - \|v+z\|_2^2\right\} \\ &+ \frac{1}{4}\left\{\frac{1}{m}(b_m\|v-z\|_2 + \mathcal{W} + \theta)^2 - \|v-z\|_2^2\right\} \\ &= \frac{1}{4}\left\{(\frac{b_m^2}{m} - 1)\|v+z\|_2^2 + \frac{2b_m(\mathcal{W}+\theta)}{m}\|v+z\|_2\right\} \\ &+ \frac{1}{4}\left\{(1 - \frac{b_m^2}{m})\|v-z\|_2^2 + \frac{2b_m(\mathcal{W}+\theta)}{m}\|v-z\|_2\right\}, \end{aligned} \tag{45}$$

with probability at least $(1 - 8e^{-\frac{\theta^2}{8}})$. Note that $\|v+z\|_2 + \|v-z\|_2 \leq 2\sqrt{2}$ and $\|v+z\|_2^2 + \|v-z\|_2^2 \leq 4$, we have:

$$\begin{aligned} &v^T\left(\frac{S^{t^T}S^t}{m} - I\right)z \\ &\leq \frac{2b_m(\mathcal{W}+\theta)}{m}\frac{\|v+z\|_2 + \|v-z\|_2}{4} + |\frac{b_m^2}{m} - 1| \\ &\leq \frac{\sqrt{2}b_m(\mathcal{W}+\theta)}{m} + |\frac{b_m^2}{m} - 1| \end{aligned} \tag{46}$$

thus finishes the proof. □

1.3.3. The proof for Proposition 3

Proof. Recall that σ_t is defined as:

$$\sigma(S^t, A) = \frac{\sup_{v\in range(A)\cap\mathbb{S}^{n-1}} \|S^t v\|_2^2}{\inf_{v\in range(A)\cap\mathbb{S}^{n-1}} \|S^t v\|_2^2}, \tag{47}$$

by simply apply again the Gordon's lemma [Lemma 6.7](Oymak et al., 2015), with $\mathcal{W}(A\mathbb{S}^{d-1}) \leq \sqrt{d}$, we with obtain the upper bound on the numerator:

$$\sup_{v\in range(A)\cap\mathbb{S}^{n-1}} \|S^t v\|_2^2 \leq (b_m + \sqrt{d} + \theta)^2, \tag{48}$$

and the lower bound:

$$\inf_{v\in range(A)\cap\mathbb{S}^{n-1}} \|S^t v\|_2^2 \geq (b_m - \sqrt{d} - \theta)^2, \tag{49}$$

both with probability at least $1 - e^{-\frac{\theta^2}{2}}$. □

1.4. Details of the implementation of algorithms and numerical experiments

For our GPIS and Acc-GPIS algorithms, we have several key points of implemenations:

- **Count sketch**

  As described in the main text.

- **Line search**

  We implement the line-search scheme given by (Nesterov, 2007) and is described by Algorithm 3 for GPIS and Acc-GPIS in our experiments with parameters $\gamma_u = 2$, and $\gamma_d = 2$. Such choice of line-search parameters simply means: when even we find the condition $f_t(\mathcal{P}_{\mathcal{K}}(x_i - \eta\nabla f_t(x_i))) \leq m_L$ does not hold, we shrink the step size by a factor of 2; and then at the beginning of each iteration, we increase the step size chosen at previous iteration by a factor of 2, then do backtracking again. Hence our methods are able to ensure we use an aggressive step size safely in each iteration. This is an important advantage of the sketched gradient method since we observe that for stochastic gradient such as SAGA a heuristic backtracking method similar to Algorithm 3 may work but it will demand a very small γ_d (tends to 1) otherwise SAGA may go unstable, and an aggressive choice like our $\gamma_d = 2$ is unacceptable for SAGA. (Hence we suspect that SAGA is unlikely to be able to benefit computational gains from line-search as our method does.)

- **Gradient restart for Acc-GPIS**

  (O'Donoghue & Candes, 2015) has proposed two heuristic adaptive restart schemes - *gradient restart* and *function restart* for the accelerated gradient methods and have shown significant improvements without the need of the knowledge of the functional parameters μ and L. Such restart methods are directly applicable for the Acc-GPIS by nature due to its sketched deterministic iterations. Here we choose the *gradient restart* since it achieves comparable performance in practice as *function restart* but cost only $\mathcal{O}(d)$ operations.

1.4.1. Procedure to generate synthetic data sets

The procedure we used to generate a constrained least-square problem sized n by 100 with approximately s-sparse solution and a condition number κ strictly follows:

1) Generate a random matrix A sized n by 100 with i.i.d entries drawn from $\mathcal{N}(0, 1)$.

2) Calculate A's SVD: $A = U\Sigma V^T$ and replace the singular values $diag(\Sigma)_i$ by a sequence:

$$diag(\Sigma)_i = \frac{diag(\Sigma)_{i-1}}{\kappa^{\frac{1}{d}}} \tag{50}$$

3) Generate the "ground truth" vector x_{gt} sized 100 by 1 randomly with only s non-zero entries in a orthongonal transformed domain Φ, and calculate the l_1 norm of it ($r = \|\Phi x_{gt}\|_1$). Hence the constrained set can be described as $\mathcal{K} = \{x : \|\Phi x\|_1 \leq r\}$.

4) Generate a random error vector w with i.i.d entries such that $\frac{\|Ax_{gt}\|_2}{\|w\|_2} = 10$.

5) Set $y = Ax_{gt} + w$

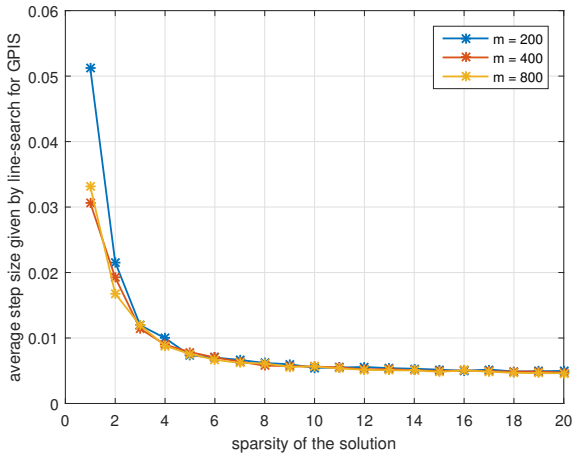

Figure 1. Experimental results on the average choices of GPIS's step sizes given by line-search scheme (Nesterov, 2013)

Table 1. Synthetic data set for step size experiment

DATA SET	SIZE	S	Φ
SYN4	(20000, 100)	-	I

1.4.2. Extra experiment for step size choice

We explore the step size choices the GPIS algorithm produce through using the line-search scheme with respect to different sparsity level of the solution. The result we shown is the average of 50 random trials.

The result of the step-size simulation demonstrates that the step sizes chosen on average by the line-search scheme for the GPIS algorithm is actually related with the sparsity of the ground truth x_{gt}: at a regime when the x_{gt} is sparse enough, the step size one can achieve goes up rapidly w.r.t the sparsity. While in our Proposition 2 we revealed that the outerloop of GPIS/Acc-GPIS can benefit from the constrained set, and here surprisingly we also find out numerically that the inner loop's can also benefit from the constrained set by aggressively choosing the large step sizes. Such a result echos the analysis of the PGD algorithm on constrained Least-squares with a Gaussian map A (Oymak et al., 2015). Further experiments and theoretical analysis of such greedy step sizes for sketched gradients and full

gradients on general maps is of great interest and will go beyond the state of the art analysis for convex optimization.

References

Beck, Amir and Teboulle, Marc. A fast iterative shrinkage-thresholding algorithm for linear inverse problems. *SIAM Journal on Imaging Sciences*, 2(1):183–202, 2009. doi: 10.1137/080716542. URL http://dx.doi.org/10.1137/080716542.

Nesterov, Yurii. Gradient methods for minimizing composite objective function. Technical report, UCL, 2007.

Nesterov, Yurii. Gradient methods for minimizing composite functions. *Mathematical Programming*, 140(1): 125–161, 2013.

O'Donoghue, Brendan and Candes, Emmanuel. Adaptive restart for accelerated gradient schemes. *Foundations of computational mathematics*, 15(3):715–732, 2015.

Oymak, Samet, Recht, Benjamin, and Soltanolkotabi, Mahdi. Sharp time–data tradeoffs for linear inverse problems. *arXiv preprint arXiv:1507.04793*, 2015.

Pilanci, Mert and Wainwright, Martin J. Randomized sketches of convex programs with sharp guarantees. *Information Theory, IEEE Transactions on*, 61(9):5096–5115, 2015.

Pilanci, Mert and Wainwright, Martin J. Iterative hessian sketch: Fast and accurate solution approximation for constrained least-squares. *Journal of Machine Learning Research*, 17(53):1–38, 2016.