

Algorithm

Recall

$$F(x) = (1/d^6) \max\{d^5 \|Ax\|_\infty - 1, \max_{i \in [N]} v_i^T x - \gamma_{v_i}\}$$

is the objective. Let $a_i \sim \mathcal{H}_d, i = 1, \dots, \lfloor \frac{d}{2} \rfloor$ denote the rows of A .

Let $\bar{x} = x - \text{Proj}(x)$, where $\text{Proj}(x)$ is projection of x to $\text{Ker}(A)$.

Then we have that for all $a \in \mathcal{H}_d$,

$$\begin{aligned} (I - \frac{aa^T}{d})x &= (I - \frac{aa^T}{d})(\text{Proj}(x) + \bar{x}) = \text{Proj}(x) - \frac{aa^T}{d}\text{Proj}(x) + (I - \frac{aa^T}{d})\bar{x} \\ &= \text{Proj}(x) + (I - \frac{aa^T}{d})\bar{x} \end{aligned}$$

The last equality holds because $\text{Proj}(x)$ is orthogonal to the rows of A . Therefore

$$\prod_{i=1}^m (I - \frac{a_{k_i} a_{k_i}^T}{d}) x_0 = \text{Proj}(x_0) + \prod_{i=1}^m (I - \frac{a_{k_i} a_{k_i}^T}{d}) \bar{x}_0$$

The intuition is that every step the following algorithm proceeds equals to multiplying $I - \frac{a_{k_i} a_{k_i}^T}{d}$ to x for some a_{k_i} , which equals to multiplying $I - \frac{a_{k_i} a_{k_i}^T}{d}$ to \bar{x} and leave $\text{Proj}(x)$ unchanged.

Algorithm 1 The projection algorithm

Input: x
 $(u, f) \leftarrow \text{Query}(x)$ $\triangleright u$ and f are gradient and function value respectively.
while $\|u\|_2 > d^{-1}$ **do** $\triangleright \|u\|_2 > d^{-1}$ indicates that $u = d^{-1}a_i$ for some i .
 $x \leftarrow x - \langle x, u \rangle / \|u\|_2$
 $(v, f) \leftarrow \text{Query}(x)$
end while
Return x

Convergence analysis

For any x , $\max_{i \leq \lfloor \frac{d}{2} \rfloor} |a_i^T x| = \max_{i \leq \lfloor \frac{d}{2} \rfloor} |a_i^T \bar{x}|$ holds. And

$$\max_{i \leq \lfloor \frac{d}{2} \rfloor} |a_i^T \bar{x}| = \|A\bar{x}\|_\infty \geq \frac{\|A\bar{x}\|_2}{\sqrt{d}} \geq \frac{\sigma_{\lfloor \frac{d}{2} \rfloor} |\bar{x}|}{\sqrt{d}}$$

Where $\sigma_{\lfloor \frac{d}{2} \rfloor}$ is the smallest singular value of A . (Actually it can be the smallest nonzero singular value with more analysis.) With the following theorem from Terence Tao's matrix book:

Theorem 2.7.1 (Lower bound). *Let $M = (\xi_{ij})_{1 \leq i \leq p; 1 \leq j \leq n}$ be an $n \times p$ Bernoulli matrix, where $1p \leq (1 - \delta)n$ for some $\delta > 0$ (independent of n). Then with exponentially high probability (i.e. $1 - O(e^{-cn})$ for some $c > 0$), one has $\sigma_p(M) \geq c\sqrt{n}$, where $c > 0$ depends only on δ .*

Take $\delta = \frac{1}{2}$, we have that $\sigma_{\lfloor \frac{d}{2} \rfloor}$ is greater than $C\sqrt{d}$ for some constant C with high probability. Therefore we have that

$$|\frac{a_{k_i} a_{k_i}^T}{d} \bar{x}| = \frac{|a_{k_i}| |a_{k_i}^T \bar{x}|}{d} = \frac{|a_{k_i}|}{d} \max_{i \leq \lfloor \frac{d}{2} \rfloor} |a_i^T \bar{x}| \geq \frac{|a_{k_i}|}{d} \frac{\sigma_{\lfloor \frac{d}{2} \rfloor} |\bar{x}|}{\sqrt{d}} = \sigma_{\lfloor \frac{d}{2} \rfloor} \frac{|\bar{x}|}{d} \geq \frac{C}{\sqrt{d}} |\bar{x}|$$

The last inequality holds w.h.p. Therefore

$$|(I - \frac{a_{k_i} a_{k_i}^T}{d}) \bar{x}| \leq \sqrt{1 - \frac{C^2}{d}} |\bar{x}|$$

And the iterations it takes to convergent given ϵ and $|x| = 1$ is

$$\log_{\sqrt{1 - \frac{C^2}{d}}} \epsilon = O(d \log(\frac{1}{\epsilon}))$$

Let $R(v_i) := \{x | F(x) = v_i^T x - i\gamma\}$ be the "Realm" of v_i . However, note that once it reaches $\bigcup_{i=1}^N R(v_i)$, the projection algorithm cannot proceed anymore.

Though an arbitrarily close projection into $\text{Ker}(A)$ is not achievable, the output of this algorithm is already close enough for some further applications. The following paragraph is cited from a document that I have recently been working on, which attempts to view all $v_i, i = 1 \dots N$ in $O(N)$ iterations. Once we can see all gradients efficiently under constrained memory, we can optimize $F(x)$ efficiently.

Example of usage from a recent work

"Let $f_v(x) = \langle v, x \rangle - \gamma_v$, and

$$F(x) = \max \left\{ \max_{i=1 \dots N} f_{v_i}(x), d^5 \|Ax\|_\infty - 1 \right\}, x \in \mathbb{B}$$

where $A \in \mathbb{R}^{p \times d}$. Let l_F and u_F denote the lower and upper bound respectively, of $\max_{i=1 \dots N} f_{v_i}(x)$. Note that in the special case of Marsden et al. $f_{v_i}(x) = \langle v_i, x \rangle - i\gamma$, and $l_F = -O(1/\sqrt{N})$. Further, consider

$$\begin{aligned} U &= \{x | \|Ax\|_\infty \leq d^{-5}(1 + u_F)\} \\ L &= \{x | \|Ax\|_\infty \leq d^{-5}(1 + l_F)\} \end{aligned}$$

Recall that $R(v_{k_1}, v_{k_2}, \dots, v_{k_n}) := \{x | F(x) = v_{k_1}^T x - k_1 \gamma = \dots = v_{k_n}^T x - k_n \gamma\}$. Write $R(V)$ for the short hand of $R(v_1, \dots, v_N)$, then it's easy to see that $L \subseteq R(V) \subseteq U$. Construct the refined projection algorithm (RP) as follows:

$$RP(x) = P(P(x) / \|P(x)\|) \cdot \|P(x)\| ,$$

where $P(x)$ is the result of running the projection algorithm starting from x . It is easy to see that for any vector $x \in \mathbb{B}$, $P(P(x) / \|P(x)\|) \in R(V) \subseteq U$, and

$$\|A(RP(x))\|_\infty \leq d^{-5}(1 + u_F) \|P(x)\| \leq d^{-5}(1 + u_F) \|x\|$$

(This result might be further improved, but it seems enough for now.)”