Solving linear feasibility in linear number of steps with self-concordant Perceptron.

Adrien Chan-Hon-Tong

Solving linear feasibility in linear number of steps with self-concordant Perceptron.

Adrien CHAN-HON-TONG

October 12, 2020

Abstract

This paper offers an algorithms which solves any linear feasibility instance with a number of step which is at most linear in the binary size of the instance. This algorithm combines Perceptron and self concordant functions (log barrier) features. In particular, algorithm step are damped Newton step mainly consisting in solving a set of linear equations.

Using common conversion, this algorithm could be used to solve linear programming, but, this leads to a slightly higher time complexity than state of the art. Inversely, times complexity for linear separability is state of the art, and, binary complexity is also mastered thank to a specific rounding process.

Claim

This paper claims the following statements

- **Definition 1:** Linear feasibility or linear separability is the task of computing $x \in \mathbb{Q}^N$ such that $Ax > \mathbf{0}$ any given matrix $A \in \mathbb{Z}^{M \times N}$ but assuming such x exists.

  For complexity analysis, let M be the number of constrains, and N be the number of variables from the original problem. Let B be the maximal binary size of any entry of A, and, L the total binary size ($L \leq MNB$ but mainly $L \approx MNB$).

- **Remark:** This linear separability task is not *real-life friendly* because it is hard to have the assumption that a solution exists without having solving the problem. Yet, this problem is significant for complexity point of view (in particular it is still equivalent to linear programming).

  Let also stress that there is no *approximation* in this task: the output should be x such that $Ax > \mathbf{0}$ or equivalently $Ax \geq \mathbf{1}$. So this algorithm is not comparable with common solvers (e.g. [4, 1, 5, 10, 11]) for which $Ax \geq \mathbf{1}$ may not be true for some rows even when an exact solution exists (it may be for a tiny set of rows but anyway it is not exact solution).

- **Lemma 1 (common knowledge):** there exists a trivial normalization process which convert $A \in \mathbb{Z}^{M\times N}$ into $\mathcal{A} \in \mathbb{Z}^{4M\times N+2}$ such that rows of $\mathcal{A}$ can be normalized on $\mathbb{Q}$ (i.e. $\forall m, \sqrt{\mathcal{A}_m \mathcal{A}_m^T} \in \mathbb{Z}$), and, such that binary sizes of A and $\mathcal{A}$ are almost the same, and, such that $\exists \chi \in \mathbb{R}^N \ / \ A\chi > \mathbf{0} \Leftrightarrow \exists \omega \in \mathbb{R}^{N+2} \ / \ \mathcal{A}\omega > \mathbf{0}$ (link between χ and ω being trivial).

- **Lemma 2 (common knowledge):** If $\exists \chi \in \mathbb{R}^N \ / \ A\chi > \mathbf{0}$ then $\Omega_A = \min\limits_{\omega \ / \ \forall m, A_m\omega \geq \mathbf{1}} \omega^T\omega$ is well defined, and, $\Omega = \widetilde{O}(NB)$.

  Interestingly, when one could apply lemma 2 after lemma 1, and, as normalizing A does not change Ω_A, this allows to state that it does not restrict generality to consider that A is normalized on $\mathbb{Q}$ when solving a linear separability instance $Ax > \mathbf{0}$ (binary size should be understand as binary size on $\mathbb{Z}$ before normalization).

- **definition 2:** $\forall v \in \mathbb{Q}^M, A \in \mathbb{Q}^{M\times N}, F_A(v) = \frac{v^T A A^T v}{2} - \sum\limits_{m=1}^{M} \log(v_m)$

- **Theorem 1:** let A be a normalized instance of linear feasibility (so Ω_A is well defined), then,

  - $F_A(\frac{1}{M}\mathbf{1}) = \widetilde{O}(M)$
  - $F_A(v)$ is lower bounded
  - let write F_A^* for the minimum, then $F_A^* = \widetilde{O}(M \log(\Omega_A))$
  - and for all v, $F_A(v) - F_A^* \leq \frac{1}{2M\Omega_A+2} \Rightarrow AA^T v > \mathbf{0}$

  So minimizing sufficiently F_A allows to find $x = A^T v$ such that $Ax > \mathbf{0}$ i.e. allows to solve the linear separability instance.

- **Corollary 1 (application of common knowledge):** As F_A is self concordant (in context of linear separability), damped Newton descent from $\frac{1}{M}\mathbf{1}$ allows to compute v such that $F_A(v) - F_A^* \leq \frac{1}{2M\Omega_A+2}$ in less than $\widetilde{O}(M \log(\Omega_A))$.

  So, solving linear separability instance A (by computing x such that $Ax > \mathbf{0}$) can be done with at most $\widetilde{O}(MNB) = \widetilde{O}(L)$ damped Newton step - each step mainly consisting in solving a $M \times M$ linear system.

- **Corollary 2:** When representing variable as fraction of two infinite-size integer, using a specific rounding process allows to use an common denominator, plus, bounding the binary size of all numerators after each steps of the algorithm. The bound being $\widetilde{O}(M \log(\Omega_A))$, it results that the algorithm is polynomial times, with strongly polynomial steps having binary complexity $\widetilde{O}(LM^\gamma)$ with M^γ the number of integer multiplication required to solve a $M \times M$ linear system. So, the global algorithm has time complexity $\widetilde{O}(L)$ which is state of the art for linear separability and binary time complexity $\widetilde{O}(L^2 M^\gamma)$.

- **Corollary 3 (common knowledge):** using common conversion process from linear separability to linear programming, it holds that this algorithm can be used to solve general linear programming (without any assumption on input contrary to linear separability).

  These conversion leads to an algorithm solving linear programming in $\widetilde{O}(ML)$ damped Newton steps.

  This complexity is slightly higher than state of the art which is $\widetilde{O}(\sqrt{N}L)$ for central path log barrier [8, 7] and even less for [3, 2, 6] (but higher than just $\widetilde{O}(L)$).

  Yet, this algorithm is better for solving linear separability than to shape the instance into a linear programming one and to apply state of the art. Indeed, central path log barrier solves linear programming in $\widetilde{O}(\sqrt{N}L)$ steps but it does not solve linear feasibility faster than $\widetilde{O}(\sqrt{N}L)$, while our algorithm is specific for this last task with a $\widetilde{O}(L)$ time complexity.

Common knowledge is presented in appendix for completeness, but, is composed of trivial and/or classical results. Conversions (appendix A) and bound on Ω_A (appendix C) are based on Cramer rules and Hadamard bound. Self concordant theory (appendix D) is briefly recalled but could be found for example in [7]. Only contribution is theorem 1 which is the encoding of a linear separability into a self concordant function which share feature with Perceptron [9] (presented in appendix B for completeness).

Let stress that this is not classical log barrier optimization because if the goal was to minimize $v^T AA^T v$ under constraint $v \geq \mathbf{0}$, then, the result would have just trivially been 0. Here the goal is a common minimization of $v^T AA^T v$ while keeping $v > \mathbf{0}$ which forces $AA^T v$ not to be negative. Let stress that as a result $v > \mathbf{0}$ at the end of the algorithm, but, this is just a side effect - the goal is just to produce $x = A^T v$ such that $Ax > \mathbf{0}$.

So this is not log barrier applied to convex optimization but self concordant Perceptron (i.e. encoding of the Perceptron into a self concordant function).

Normally, all notations of the paper are standards. The index $_A$ of F_A and Ω_A would be omitted when no ambiguity is possible.

Proof

As it does not restrict generality (see lemma 1 and 2), let assume that $A \in \mathbb{Q}^{M \times N}$ is a normalized linear separability instance i.e. rows are normalized and exists x such that $Ax \geq \mathbf{1}$, and so, $\Omega = \min_{\omega \ / \ \forall m, A_m \omega \geq \mathbf{1}} \omega^T \omega$ is well defined. Then,

$$F(v) = \frac{v^T AA^T v}{2} - \sum_{m=1}^{M} \log(v_m)$$

Existence of a minimum

First, as A is normalized $A_i A_j^T \leq 1$ due to Cauchy, so it trivially holds that $F(\frac{1}{M}\mathbf{1}) = \frac{1}{M^2}\sum_{i,j} A_i A_j^T + \sum_{m=1} \log(M) \leq M\log(M) + 1 = \widetilde{O}(M)$.

then, $v^T A\omega = (A^T v)^T \omega \leq \sqrt{v^T AA^T v \times \omega^T \omega}$. But, by definition $A\omega \geq \mathbf{1}$, so $\forall v \geq \mathbf{0}$, $v^T \mathbf{1} \leq v^T A\omega \leq \sqrt{v^T AA^T v \times \Omega}$. So, $\forall v \geq \mathbf{0}$, $\frac{(v^T \mathbf{1})^2}{\Omega} \leq v^T AA^T v$, and, $(v^T\mathbf{1})^2 > v^T v$ $(v \geq \mathbf{0})$. So $\forall v \geq \mathbf{0}$, $\frac{v^T v}{\Omega} \leq v^T AA^T v$. This is the same idea than in Perceptron, if v is large then $||x|| = ||A^T v||$ can not be too low.

Let introduce $f(t) = \frac{t^2}{2\Omega} - \log(t)$, from previous inequality it stands that $F(v) \geq \sum_m f(v_m)$.

Now, f is a single variable function which goes to infinity when t goes to 0 ($t^2 \to 0$ but $-\log(t) \to \infty$) or to infinity (t^2 growths faster than $\log(t)$). So, f has a minimum and so F too (more precisely, as f goes to ∞ on 0 or ∞, one could define a compact on which $F(v) \leq F(v_{start})$, but on this compact it has a minimum which is also the minimum of the function as the value if higher outside the compact). Let call them f^* and F^*.

Importantly, if there is no x such that $Ax \geq \mathbf{1}$ then, there exists y such that $Ay = \mathbf{0}$ and $y > \mathbf{0}$, and, F is not bounded (can go to $-\infty$). But, the existence of x (assumption of linear separability) forces the existence of F^* (thank to Cauchy because $\frac{v^T v}{\Omega} \leq v^T AA^T v$ for positive v).

As f, F are smooth the minimums are characterized by a null derivative or gradient. $f'(t) = \frac{t}{\Omega} - \frac{1}{t}$, so, $f'(\sqrt{\Omega}) = 0$, so $f^* = f(\sqrt{\Omega}) = \frac{1}{2} - \frac{1}{2}\log(\Omega) \geq -\log(\Omega)$. Thus, the minimum of F verifies $F^* \geq Mf^* \geq -M\log(\Omega)$.

So F is bounded and $F(v_{start}) - F^* \leq \widetilde{O}(M\log(\Omega))$ for $v_{start} = \frac{1}{M}\mathbf{1}$.

Normalization, linearization and lemmas

Independently, let remark that $\theta(t) = F(tv) = \frac{v^T AA^T v}{2} t^2 - \mathbf{1}^T \log(v) - M\log(t)$ is minimal when $v^T AA^T v = M$. So for any w, one could build a $v = \mu w$ such that $v^T AA^T v = M$ and $F(v) \leq F(w)$. In other words, it stands that $F\left(\sqrt{\frac{M}{v^T AA^T v}} v\right) \leq F(v)$.

So, let consider $v \geq \mathbf{0}$ such that $v^T AA^T v = M$. As, $v^T AA^T v \geq \frac{(\mathbf{1}v)^2}{\Omega}$, no v_m could be higher than $\sqrt{M\Omega}$ i.e. $\mathbf{0} \leq v \leq \sqrt{M\Omega}\mathbf{1}$.

Let also remark that $F(v+w) = \frac{v^T AA^T v}{2} + \frac{w^T AA^T w}{2} + w^T AA^T v - \mathbf{1}^T \log(v) - \mathbf{1}^T \log(1 + \frac{w}{v}) = F(v) + \frac{w^T AA^T w}{2} + w^T AA^T v - \mathbf{1}^T \log(1 + \frac{w}{v})$

Finally, let consider the following lemmas from basic analysis:

1. $\varphi(t) = \frac{1}{2}\alpha t^2 - \log(1+t) \leq \frac{1}{2}(\alpha+1)t^2 - t = \psi(t)$ for $t \geq 0$
2. $\psi(\frac{1}{\alpha+1}) \leq -\frac{1}{2\alpha+2}$
3. $\varphi(\frac{1}{\alpha+1}) \leq -\frac{1}{2\alpha+2}$ i.e. $\forall \alpha \geq 0$, $\frac{1}{2}\frac{\alpha}{(\alpha+1)^2} - \log(1 + \frac{1}{\alpha+1}) \leq -\frac{1}{2\alpha+2}$

First, $\psi'(t) - \phi'(t) = (\alpha+1)t - 1 - \alpha t + \frac{1}{1+t} = t - 1 + \frac{1}{1+t} = \frac{t^2}{1+t} > 0$, so $\psi(t) - \phi(t)$ always increase. But, $\psi(0) = \phi(0) = 0$ so $\psi(t) \geq \phi(t)$ for $t \geq 0$. Second, $\psi(\frac{1}{\alpha+1}) = \frac{1}{2}(\alpha+1)\frac{1}{(\alpha+1)^2} - \frac{1}{\alpha+1} = -\frac{1}{2\alpha+2}$. Three is just 1+2.

Convergence

Now, if $A_k A^T v \leq 0$ and $v^T A A^T v = M$, let introduce $w = v + \frac{vk}{v_k^2+1}\mathbf{1}_k$: this idea is again Perceptron based, one could increase v and decrease $||A^T v||$ (precisely F here) in the same times. Then $F(w) = F(v + \frac{vk}{v_k^2+1}\mathbf{1}_k) = F(v) + \frac{A_k A_k^T}{2}(\frac{vk}{v_k^2+1})^2 + A_k A^T v \times \frac{vk}{v_k^2+1} - \log(1 + \frac{1}{v_k^2+1})$. But, $A_k A^T v \leq 0$ and $A_k A_k^T = 1$, so $F(w) \leq F(v) + \frac{1}{2}(\frac{vk}{v_k^2+1})^2 - \log(1 + \frac{1}{v_k^2+1})$.

And, from lemmas just above, $F(w) \leq F(v) - \frac{1}{2v_k^2+2}$. But, $v_k \leq \sqrt{M\Omega}$, so, $F(w) \leq F(v) - \frac{1}{2M\Omega+2}$ which is impossible if $F(v) - F^* < \frac{1}{2M\Omega+2}$. So, $\forall v > \mathbf{0}$ such that $v^T A A^T v = M$, $F(v) - F^* \leq \frac{1}{2M\Omega+2} \Rightarrow AA^T v > \mathbf{0}$.

Finally, $\forall v > \mathbf{0}$, $F(v) - F^* \leq \frac{1}{2M\Omega+2} \Rightarrow F(\sqrt{\frac{M}{v^T AA^T v}}v) - F^* \leq \frac{1}{2M\Omega+2}$

$\Rightarrow \sqrt{\frac{M}{v^T AA^T v}} AA^T v > \mathbf{0} \Rightarrow AA^T v > \mathbf{0}$. This finishes the proof of the main theorem.

Proof of corollaries

Self concordance

Despite that all this paper relies on self concordance theory, there is nothing to prove to apply results from [7] recalled in appendix D to F (AA^T is positive because $v^T AA^T v = ||A^T v||_2^2 \geq 0$ and F has a minimum see theorem 1).

So results recalled in in appendix D directly leads to state that damped Newton descent starting from any v_{start} builds v such that $F(v) - F^* \leq \frac{1}{2Mx^T x+2}$ (i.e. a solution of the linear feasibility) in less than $\widetilde{O}(F(v_{start}) - F^* + \log\log(2Mx^T x))$ steps.

But as proven in theorem 1, $F(v_{start}) - F^* = M\log(\Omega)$ for $v_{start} = \frac{1}{M}\mathbf{1}$. So, damped Newton descent builds a solution of the linear feasibility problem in less than $\widetilde{O}(M\log(\Omega) + \log\log(M\Omega))$ steps. Precisely, almost all the optimization is done in the so called phase 1 which may last $M\log(\Omega)$ because the phase 2 only requires $\log\log(M\Omega)$ steps i.e. it is negligible.

Using Cramer rules and Hadarmard bound (see appendix C), one can then proof that $\Omega \leq N 2^{2NB\log(N)}$ with 2^B being a bound on each entry of $A \in \mathbb{Z}^{M\times N}$ before normalization - i.e. maximal binary size of all entries. This finally gives a $\widetilde{O}(NMB)$ number of steps, which becomes $\widetilde{O}(L)$ by reintroducing the total binary size. So, the complexity is established:

The offered algorithm solves linear feasibility in $\widetilde{O}(L)$ damped Newton steps (L being the total binary size of the input matrix A).

Binary complexity

For practical floating point implementation, the previous claim that complexity is $\widetilde{O}(L)$ Newton steps is sufficient. Yet, floating point implementation is impossible with too large numbers (and may lead to numerical issue even with not too large numbers).

So, from theoretical point of view, it is required to have also a bound on the binary complexity of each step when representing rational number as fraction of two infinite size integers - let recall that naive Gaussian elimination is exponential when not taking care about intermediary values.

For establishing binary complexity of each step, one should use approximation instead of square root and round the number after each damped Newton step without breaking the convergence. This is possible as presented bellow.

Computing λ is impossible on $\mathbb{Q}$, but, as F is convex, it holds that $F(v - \theta(\nabla_v^2 F)^{-1}(\nabla_v F)) \leq \frac{1}{2}(F(v) + F(v - \frac{1}{1+\lambda(v)}(\nabla_v^2 F)^{-1}(\nabla_v F)))$ if $\frac{1}{2}\frac{1}{1+\lambda(v)} \leq \theta \leq \frac{1}{1+\lambda(v)}$. So approximating λ by a factor 2 approximation is sufficient and possible on $\mathbb{Q}$ (only twice number of steps are required compared to infeasible exact computation version). Then, with the same idea, normalizing $v^T AA^T v = M$ is not possible exactly, but, $\frac{M}{2} \leq v^T AA^T v \leq 2M$ is possible and still guarantees that $v \leq \widetilde{O}(M\Omega)$.

Importantly, finding μ such that $\mu \leq \sqrt{\rho} \leq \mu+1$ is in $\log(\rho)$ (with bisection) which is too much here. But, finding μ such that $\mu \leq \sqrt{\rho} \leq 2\mu$ is in $\log(\log(\rho))$ because bisection can be done on power. So, computing approximation of λ or normalization should be done using this fast bisection, and this way, binary complexity of this operation will be $\widetilde{O}(B)$ and not $O(B^2)$ (currently there are dedicated algorithms to compute square root, but, here, there are not even required - this may allows to find an 2 approximation of the optimal value instead of a fixed $\frac{1}{\lambda+1}$ for the same cost).

Finally, flooring variable could be dramatic due to the log but ceiling is acceptable. So at each step, one should ceil v: $v_{round} = int(Q*v+1)/Q$ (after Newton and normalization). So, one has $F(v_{round}) \leq F(v) + 2|v - v_{round}|\mathbf{1}^T AA^T v$ (logs are not affected by ceiling) so $F(v_{round}) \leq F(v) + \frac{2}{Q}\mathbf{1}^T AA^T v$ but as v is bounded by $O(M\Omega)$ and A is normalized this leads to $F(v_{round}) \leq F(v) + O(\frac{M^2\Omega}{Q})$. By setting Q such that this remainder is - at most - half the minimal decreasing during a damped Newton descent ($\frac{1}{4} - \log(\frac{5}{4})$ in phase 1), one could guarantee that the only twice the number of damped Newton steps are required to reach a solution while keeping a mastered infinite integer representation (with a common denominator Q having a binary size limited by $\widetilde{O}(\log(M\Omega))$ and numerators with binary size also bounded by $\widetilde{O}(\log(M\Omega))$ as $v \leq O(M\Omega)$.

So, binary size of all numbers can be round to $\widetilde{O}(NB)$ at the end of each damped Newton step. Within a step, binary size will not exceed $\widetilde{O}(NMB)$ (because there is $O(M)$ multiplications). This leads to an algorithm without any numerical issue, and, steps being the resolution of a $M \times M$ linear system with entries with size $\widetilde{O}(NMB)$.

So, finally, for **floating point implementation** (i.e. with possible numeri-

cal instabilities and not arbitrary large numbers), the algorithm converges after solving less than $\widetilde{O}(M\Omega)$ linear systems of size $M \times M$.

And, using **exact infinite integer representation**, the algorithm converges after solving less than $\widetilde{O}(L)$ linear systems of size $M \times M$ with fraction coded by two integers with maximal binary value being $\widetilde{O}(L)$ using the offered rounding strategy. This leads to a $\widetilde{O}(N^{\omega}L^2)$ binary complexity (assuming fast multiplication of number of binary size B is done in $\widetilde{O}(B)$ binary operations and solving a $M \times M$ linear system requires $\widetilde{O}(N^{\omega})$ multiplications) for $A \in \mathbb{Z}^{M \times N}$ with maximal entry bounded (in absolute value) by 2^B and total binary size $L \approx NMB$.

References

[1] Chih-Chung Chang and Chih-Jen Lin. Libsvm: A library for support vector machines. *ACM transactions on intelligent systems and technology (TIST)*, 2011.

[2] Sergei Chubanov. A polynomial projection algorithm for linear feasibility problems. *Mathematical Programming*, 2015.

[3] Michael B Cohen, Yin Tat Lee, and Zhao Song. Solving linear programs in the current matrix multiplication time. In *Proceedings of the 51st annual ACM SIGACT symposium on theory of computing*, 2019.

[4] Rong-En Fan, Kai-Wei Chang, Cho-Jui Hsieh, Xiang-Rui Wang, and Chih-Jen Lin. Liblinear: A library for large linear classification. *Journal of machine learning research*, 2008.

[5] Thorsten Joachims. Svmlight: Support vector machine. *SVM-Light Support Vector Machine http://svmlight. joachims. org/, University of Dortmund*, 1999.

[6] Yin Tat Lee and Aaron Sidford. Path finding methods for linear programming: Solving linear programs in o (vrank) iterations and faster algorithms for maximum flow. In *2014 IEEE 55th Annual Symposium on Foundations of Computer Science*. IEEE, 2014.

[7] Arkadi Nemirovski. Interior point polynomial time methods in convex programming. *Lecture notes*, 2004.

[8] Yurii Nesterov and Arkadii Nemirovskii. *Interior-point polynomial algorithms in convex programming*. Siam, 1994.

[9] Frank Rosenblatt. The perceptron: a probabilistic model for information storage and organization in the brain. *Psychological review*, 1958.

[10] Shai Shalev-Shwartz, Yoram Singer, Nathan Srebro, and Andrew Cotter. Pegasos: Primal estimated sub-gradient solver for svm. *Mathematical programming*, 2011.

[11] Shai Shalev-Shwartz and Nathan Srebro. Svm optimization: inverse dependence on training set size. In *Proceedings of the 25th international conference on Machine learning*, 2008.

Appendix A: Equivalence on linear programming

This paper provides an algorithm $\mathbf{algo}_0$ such that for any A such that $\exists x \ / \ Ax \geq \mathbf{1}$, the algorithm returns v such that $AA^Tv > \mathbf{0}$ in $\widetilde{O}(L)$ Newton (v is positive but this does not matter).

- Trivially, if $AA^Tv > \mathbf{0}$ then $A(A^Tv) > \mathbf{0}$. So $\mathbf{algo}_1(A) = A^T\mathbf{algo}_0(A)$ returns χ such that $A\chi > \mathbf{0}$ (when exists) with the same complexity.
- Let consider any A, b such that $\exists x \ / \ Ax > b$, finding such x is equivalent to find a pair x, t such that $Ax - t \times b > \mathbf{0}$, $t > 0$ because $\frac{x}{t}$ is a solution. Let $\mathcal{A}$ the matrix with A plus $\mathbf{1}$ as additional column and $(\mathbf{0} \ \ 1)$ as additional row. Thus, $\mathbf{algo}_2(A, b)$ can be implemented by computing $(x \ \ t) = \mathbf{algo}_1(\mathcal{A})$ and returning $\frac{x}{t}$ (complexity still unchanged)
- Let any A and b such that $\exists x / Ax \geq b$, finding such x is equivalent to find a pair x, t such that $Ax + t\mathbf{1} \times > b$, $0 < t < \frac{1}{\max\limits_{\mathcal{S} \subset A} Det(\mathcal{S})}$. Indeed, from such pair, one could project x on $\{z \ / \ Az \geq b\}$ while minimizing t, this leads to $\hat{x}, \hat{t}$ with $\hat{t} \leq t < \frac{1}{\max\limits_{\mathcal{S} \subset A} Det(\mathcal{S})}$ but $\hat{x}, \hat{t}$ is a vertex of A. So Cramer rule applies, and so there exists $\mathcal{S}$ such that $\hat{t} = \frac{Det(\mathcal{S}_{partiel})}{Det(\mathcal{S})}$ but seeing constraint on $\hat{t}$ it forces $\hat{t} = 0$ i.e. $A\hat{x} \geq b$. So, $\mathbf{algo}_3(A, b)$ can be implemented by
  - computing $(x \ \ t) = \mathbf{algo}_2(\mathcal{A}, \beta)$ with $\mathcal{A}$ being A plus $\mathbf{1}$ as additional column, and $(\mathbf{0} \ \ 1)$, $(\mathbf{0} \ \ -1)$ as additional rows and $\beta = b$ plus 0 and $-\frac{1}{\max\limits_{\mathcal{S} \subset A} Det(\mathcal{S})}$
  - projecting $(x \ \ t)$ on $\{(z \ \ \tau) \ / \ Az + \tau\mathbf{1} \geq b\}$ while minimizing t
  - returning x

  Importantly, if highest number in A, b is bounded by 2^B, then, maximal determinant is lower than $2^{MB + M\log(M)}$ (Hadamard bound), so, the complexity of $\mathbf{algo}_3$ is just increased by a factor M compared to the complexity of $\mathbf{algo}_2$.
- Let any A, b - **without assumption** - solving $Ax \geq b$ (or producing a certificate) is equivalent to solve $\min\limits_{z \ / Az + t\mathbf{1} \geq b, t \geq 0} t$ (there is a solution if the minimum is 0). Yet, this last linear program is structurally feasible ($x = 0$ and a sufficiently large t) and bounded because $t \geq 0$. Thus, primal dual theory gives a system $A_{primal-dual}(x \ \ y) \geq b_{primal-dual}$ whose solution contains solution of the linear program $\min\limits_{z \ / Az + t\mathbf{1} \geq b, t \geq 0} t$

and thus of x such that $Ax \geq b$ or a certificate that no solution exists. So a possible implementation of $\mathbf{algo}_4(A,b)$ is to compute $(x \quad y) = \mathbf{algo}_3(A_{primal-dual}, b_{primal-dual})$ returning only x or the certificate.

Importantly, the number of variables-constraints is scaled two folds but from theoretical point of view, it does not change the complexity. Let stress again that from $\mathbf{algo}_4$ there is no assumption on the input: for any A, b $\mathbf{algo}_4(A,b)$ either returns x such that $Ax \geq b$, or, a certificate that no such x exists.

- Finally, for any A, b, c without any assumption solving $\min_{x \ / Ax \geq b} c^T x$ can be done with 2 $\mathbf{algo}_4$ calls and one $\mathbf{algo}_3$ call:
  - one to know if the problem is feasible i.e. $\mathbf{algo}_4(A,b)$
  - one on the dual to known if it is bounded $\mathbf{algo}_4(A_{dual}, b_{dual})$
  - and one call to $\mathbf{algo}_3$ on the primal dual to get the optimal solution (if one passes the two first step than the problem is feasible and bounded so the primal dual has a solution).

  Again, from theoretical point of view, the complexity does not change: it only does 3 calls on instances only scaled 2 times. At the end, it returns the optimal solution or a certificate that the problem is not feasible or not bounded. Let stress that the opposite way is trivial $\mathbf{algo}_5(AA^T, \mathbf{1}, \mathbf{0})$ is a correct implementation of $\mathbf{algo}_1(A)$ for any A.

This recall the common knowledge that the ability to solve linear separation $v \ / \ AA^T v > 0$ (assuming a solution exists) allows to solve linear programming in general just adding a factor M.

Yet, an important point is that an algorithm solving linear programming in $\widetilde{O}(\sqrt{N}L)$ (e.g. central path log barrier) will solve linear feasibility in $\widetilde{O}(\sqrt{N}L)$ if just applied to the linear feasibility instance shaped into a linear programming one. Inversely, this paper offers and algorithm which directly tackles linear separability with complexity $\widetilde{O}(L)$ i.e. better than central path log barrier for linear separability but worse for linear programming. The question of the extension of the Perceptron log barrier.

The same holds for normalization: let A such that $\exists x \ / \ Ax \geq \mathbf{1}$, then let consider $\mathcal{A}$ which has 4 times more constraint and 2 additional variables (so complexity does not change) built as follow: for all constraint A_n, $\mathcal{A}$ gets the 4 constraints $2A_n :: \pm 1 :: \pm 2A_n A_n^T$.

The point is that each such new constraints has a norm $4A_n A_n^T + 1 + 4(A_n A_n^T)^2 = (2A_n A_n^T + 1)^2$. So, each new constraint can be normalized on $\mathbb{Q}$ as the root of the norm is an integer (while binary size is only scaled twice). Importantly, $(x \quad 0 \quad 0)$ is a solution for $\mathcal{A}$, and inversely, any solution χ for $\mathcal{A}$ is a solution for A when removing the last two coordinates (if not zeros, it always make one from the 4 constraints worse).

Let stress that normalizing is mainly for understanding - the offered algorithm works with the raw matrix but proof is a bit harder.

Appendix B: Perceptron

The goal of the Perceptron algorithm [9] is to find x such that $Ax > \mathbf{0}$ (assuming such x exists). As, this does not restrict generality (see Appendix A), one could assume that A has normalized rows and could consider $\Omega_A = \min_{\omega \ / \ A\omega \geq \mathbf{1}} \omega^T \omega$. Let write ω_A for the solution of this SVM problem.

The key idea of Perceptron is to look for x as $x = A^T v$ with positive v, and, to remark that either $Ax > \mathbf{0}$ or it is possible both to increase v and decrease $x^T x$. But then, a Cauchy inequality linking $v^T v$ (which increases) and $x^T x$ (which decreases) would forces this process to converge.

Precisely, one could consider the algorithm $x_{t+1} = x_t + A_{i_t}^T$ with $A_{i_t} x_t \leq 0$ (possible if the algorithm has not reached $Ax > \mathbf{0}$) with $x_0 = \mathbf{0}$.

Let recall the Cauchy inequality: $\forall u, v, u^T v \leq \sqrt{u^T u \times v^T v}$. Then, $(\omega_A^T x_t)^2 \leq x_t^T x_t \Omega_A$.

But $x_{t+1}^T x_{t+1} = x_t^T x_t + 2A_{i_t} x_t + A_{i_t}^T A_{i_t} \leq x_t^T x_t + 1$ because $2A_{i_t} x_t \leq 0$ and $A_{i_t}^T A_{i_t} = 1$.

So, $x_{t+1}^T x_{t+1} \leq t$.

While $\omega_A^T x_{t+1} = \omega_A^T x_t + \omega_A^T A_{i_t}^T \geq \omega_A^T x_t + 1$ because $A\omega_A \geq \mathbf{1}$.

So, $\omega_A^T x_t \geq t$.

So $(\omega_A^T x_t)^2 \leq x_t^T x_t \Omega_A$ becomes:

$$t^2 \leq (\omega_A^T x_t)^2 \leq x_t^T x_t \Omega_A \leq \Omega_A \times t$$

But, t^2 can not be lower than $\Omega_A \times t$ for $t \geq \Omega_A$, so the algorithm converges in less than Ω_A steps and returns some $x = A^T v$ with $v \geq \mathbf{0}$.

Now, Ω_A can be large: classical tricks combining Cramer rules and Hadarmard bound allow to state that if the maximal value of all entries of $A \in \mathbf{Z}^{M \times N}$ (before normalization) is 2^B, then, $\Omega_A \leq N 2^{2BN \log(N)}$ (see appendix C for completeness), so this algorithm is exponential because it may require $\widetilde{O}(N 2^{2BN \log(N))}$ steps. Inversely, the self concordant Perceptron is *just* an updated version of this Perceptron such that the number of steps goes from Ω_A to $M \log(\Omega_A) = \widetilde{O}(NMB) = \widetilde{O}(L)$.

Let stress that Perceptron steps are much faster than Newton ones: Perceptron step is just checking if $Ax_t > \mathbf{0}$ and if not $x_{t+1} = x_t + A_{i_t}^T$, so binary step complexity is just $\widetilde{O}(NMB)$ while damped Newton steps require to solve a $M \times M$ linear system. More generally, as *step* may have different complexity, it is important to distinguish *binary complexity* (total number of binary operations - unbiased) and *complexity* (number of steps - a step being a strongly polynomial times sub algorithm - like solving a linear system - biased).

Appendix C: SVM

Assuming $\exists x \in \mathbb{Q}^N \ / Ax \geq \mathbf{1}$ with $A \in \mathbb{Q}^{M \times N}$ a normalized matrix, then, there exists a solution to the SVM problem $\Omega_A = \min_{\omega \ / \ A\omega \geq \mathbf{1}} \omega^T \omega$, and, Perceptron will converge in $\widetilde{O}(\Omega_A)$ steps while the self concordant Perceptron in $\widetilde{O}(M \log(\Omega_A))$

Now, this value Ω_A should be characterized. Currently, it is not trivial to show directly that this Ω_A is small. But the classical trick is to consider the following theoretical problem (that nobody is going to solve) $\min_{\chi\ /A\chi\geq\mathbf{1}} \Upsilon((\chi\ \ A\chi-\mathbf{1})$ where $\Upsilon(p) > \Upsilon(q)$ if and only if $\upsilon(p) > \upsilon(q)$ or $\upsilon(p) = \upsilon(q)$ and $||p||_1 > ||q||_1$ and where $\upsilon(p)$ is the lexicographic rank of not-zero index vector p i.e. $\upsilon(p) > \upsilon(q)$ if and only if $\exists i\ /(\forall j < i,\ p_j = 0 \Leftrightarrow q_i = 0) \wedge p_i \neq 0 \wedge q_i = 0$. The solution of this problem is unique (because if there is 2, one could explore the linear combination of the two either to saturated some $A_j\chi \geq 1$ or to make 0 some χ_i). And, this solution is naturally a vertex because it maximizes the size of a set of equality.

Thus, the solution of this theoretical problem is also the only solution of a set of equality $A_I\chi = \mathbf{1},\ \chi_J = \mathbf{0}$ with some set I, J. Thus, this system is full rank, and one could extract a square not singular matrix $\mathcal{A}$ from it, and, χ is the only solution of $\mathcal{A}x = \mathcal{B}$ with $\mathcal{A}$ being a sub matrix of A (plus some Dirac vector) and $\mathcal{B}$ being 0-1 vectors. So Cramer rules hold, thus, each component of χ is a fraction of sub determinants of A. But, Hadamard bound (again) gives that if maximal entry of this matrix is 2^B, then $\chi^T\chi \leq 2^{2NB\log(N)}$.

So, there exists χ such that $A\chi \geq \mathbf{1}$ and such that $\log(\Omega_A) \leq \log(\chi^T\chi) = \widetilde{O}(NB)$.

Appendix D: Self concordant theory

This appendix only reports few results from [7] which could be considered for completeness.

A smooth single variable function ϕ is self concordant if and only if $|\phi'''(t)| \leq 2(\phi''(t))^{\frac{3}{2}}$. And, a smooth multi variable function Φ is self concordant if each ray $t \to \Phi(at + b)$ is self concordant (i.e. for all vectors a, b).

Mainly, self concordant functions are quadratic, linear, constant and $-\log$ ones. For example, the function $x \to x^TQx + q^Tx - \log(\alpha x - \beta)$ is a self concordant function for any vector q, α, bias β and any matrix Q as soon as Q is positive (not necessarily strictly positive, $\mathbf{0}$ is possible).

Now, if G is a smooth self concordant function, let introduce λ and the damped Newton descent:

$$\lambda_G(x) = \sqrt{(\nabla_x G)^T(\nabla_x^2 G)^{-1}(\nabla_x G)}$$

$$x = x - \frac{1}{1 + \lambda_G(x)}(\nabla_x^2 G)^{-1}(\nabla_x G)$$

The important property of Newton descent on self concordant functions is that: **if** G has a minimum G^*, then, damped Newton descent starting from any x_{start} allows to build x such that $G(x) - G^* \leq \varepsilon$ in less than $\widetilde{O}(G(x_{start}) - G^* + \log\log(\frac{1}{\varepsilon}))$ damped Newton steps.

Precisely, [7] describes this optimization in two phases:

- while $\lambda(x) \geq \frac{1}{4}$, each damped Newton step decreases G of at least $\frac{1}{4} - \log(\frac{5}{4})$, this is the so called phase 1.
- as soon as $\lambda(x) \leq \frac{1}{4}$, then, the two next statements hold

  $$G(x) - G^* \leq \frac{\lambda_G(x)^2}{2(1-\lambda_G(x))} \quad ; \quad \lambda(x - \frac{1}{1+\lambda_G(x)}(\nabla_x^2 G)^{-1}(\nabla_x G)) \leq 2\lambda(x)^2$$

  this is the so called phase 2 (with $G(x) - G^*$ and/or $\lambda_G(x)$ having a quadratic convergence).

Classically, damped Newton descent is considered in the phase 2 while it is considered in phase 1 in this paper (phase 2 being negligible as performed in $\widetilde{O}(\log\log(M\Omega_A))$.