An Optimal Affine Invariant Smooth Minimization Algorithm.

Alexandre d’Aspremont, CNRS & D.I. ENS.

with Cristóbal Guzmán, Vincent Roulet, Nicolas Boumal & Martin Jaggi.
Support from ERC SIPA.
Introduction

A complexity bound.

\[ O \left( \frac{n \log n}{\epsilon} \right) \]
A complexity bound, if we’re lucky...

\[ O \left( \frac{L \ln \log n}{\epsilon} \right) \]
Introduction

A complexity bound, if we’re lucky. . .

\[ O \left( \frac{L n \log n}{\epsilon} \right) \]

One thing missing: **the data.**
Big gap between worst-case complexity and empirical performance for first-order optimization algorithms.

- Data-driven complexity bounds?
- In particular, quantify the **complexity vs. statistical performance tradeoff**?
Outline

- Affine invariant bounds.
- Renegar’s condition number and compressed sensing.
Solve

\[
\begin{align*}
\text{minimize} & \quad f(x) \\
\text{subject to} & \quad x \in Q, \\
\end{align*}
\]

in \( x \in \mathbb{R}^n \).

- Here, \( f(x) \) is convex, **smooth**.
- Assume \( Q \subset \mathbb{R}^n \) is compact, convex and **simple**.
**Complexity**

**Newton’s method.** At each iteration, take a step in the direction

\[ \Delta x_{nt} = -\nabla^2 f(x)^{-1} \nabla f(x) \]

Assume that

- the function \( f(x) \) is self-concordant, i.e. \( |f'''(x)| \leq 2f''(x)^{3/2} \),
- the set \( Q \) has a self concordant barrier \( g(x) \).

[Nesterov and Nemirovskii, 1994] Newton’s method produces an \( \epsilon \) optimal solution to the barrier problem

\[ \min_x h(x) \overset{\triangle}{=} f(x) + t g(x) \]

for some \( t > 0 \), in at most

\[ \frac{20 - 8\alpha}{\alpha \beta (1 - 2\alpha)^2} (h(x_0) - h^*) + \log_2 \log_2 (1/\epsilon) \] iterations

where \( 0 < \alpha < 0.5 \) and \( 0 < \beta < 1 \) are line search parameters.
**Newton’s method.** Basically

\[
\text{\# Newton iterations} \leq 375 \left( h(x_0) - h^* \right) + 6
\]

- Empirically valid, up to constants.
- **Independent from the dimension** \( n \).
- **Affine invariant.**

In practice, implementation mostly requires **efficient linear algebra.**

- Form the Hessian.
- Solve the Newton (or KKT) system \( \nabla^2 f(x) \Delta x_{nt} = -\nabla f(x) \).
Affine Invariance

Set \( x = Ay \) where \( A \in \mathbb{R}^{n \times n} \) is nonsingular

\[
\begin{align*}
\text{minimize} & \quad f(x) \\
\text{subject to} & \quad x \in Q,
\end{align*}
\]

becomes

\[
\begin{align*}
\text{minimize} & \quad \hat{f}(y) \\
\text{subject to} & \quad y \in \hat{Q},
\end{align*}
\]

in the variable \( y \in \mathbb{R}^n \), where \( \hat{f}(y) \triangleq f(Ay) \) and \( \hat{Q} \triangleq A^{-1}Q \).

- **Identical Newton steps**, with \( \Delta x_{nt} = A\Delta y_{nt} \)
- **Identical complexity bounds** \( 375 (h(x_0) - h^*) + 6 \) since \( h^* = \hat{h}^* \)

Newton’s method is **invariant w.r.t. an affine change of coordinates**. The same is true for its complexity analysis.
Large-Scale Problems

The challenge now is scaling.

- Newton’s method (and derivatives) solve all reasonably large problems.
- Beyond a certain scale, second order information is out of reach.

**Question today:** clean complexity bounds for first order methods?
Franke-Wolfe

**Conditional gradient.** At each iteration, solve

\[
\begin{align*}
\text{minimize} & \quad \langle \nabla f(x_k), u \rangle \\
\text{subject to} & \quad u \in Q
\end{align*}
\]

in \( u \in \mathbb{R}^n \). Define the curvature

\[
C_f \triangleq \sup_{s,x \in \mathcal{M}, \alpha \in [0,1], \ y = x + \alpha (s - x)} \frac{1}{\alpha^2} \left( f(y) - f(x) - \langle y - x, \nabla f(x) \rangle \right).
\]

The Franke-Wolfe algorithm will then produce an \( \epsilon \) solution after

\[
N_{\max} = \frac{4C_f}{\epsilon}
\]

iterations.

- \( C_f \) is affine invariant but the bound is suboptimal in \( \epsilon \) in many cases.
- If \( f(x) \) has a Lipschitz gradient, the lower bound can be as low as \( O \left( \frac{1}{\sqrt{\epsilon}} \right) \).
**Smooth Minimization** algorithm in [Nesterov, 1983] to solve

\[
\begin{align*}
\text{minimize} & \quad f(x) \\
\text{subject to} & \quad x \in Q,
\end{align*}
\]

Original paper was in an Euclidean setting. In the general case . . .

- **Choose a norm** \( \| \cdot \| \). \( \nabla f(x) \) Lipschitz with constant \( L \) w.r.t. \( \| \cdot \| 

\[
f(y) \leq f(x) + \langle \nabla f(x), y - x \rangle + \frac{1}{2}L\|y - x\|^2, \quad x, y \in Q
\]

- **Choose a prox function** \( d(x) \) for the set \( Q \), with

\[
\frac{\sigma}{2}\|x - x_0\|^2 \leq d(x)
\]

for some \( \sigma > 0 \).
Smooth minimization algorithm [Nesterov, 2005]

Input: \( x_0 \), the prox center of the set \( Q \).

1: for \( k = 0, \ldots, N \) do
2: Compute \( \nabla f(x_k) \).
3: Compute \( y_k = \arg\min_{y \in Q} \left\{ \langle \nabla f(x_k), y - x_k \rangle + \frac{1}{2}L\|y - x_k\|^2 \right\} \).
4: Compute \( z_k = \arg\min_{x \in Q} \left\{ \sum_{i=0}^k \alpha_i [f(x_i) + \langle \nabla f(x_i), x - x_i \rangle] + \frac{L}{\sigma}d(x) \right\} \).
5: Set \( x_{k+1} = \tau_k z_k + (1 - \tau_k)y_k \).
6: end for

Output: \( x_N, y_N \in Q \).

Produces an \( \epsilon \)-solution in at most

\[
N_{\text{max}} = \sqrt{\frac{8Ld(x^*)}{\epsilon \sigma}}
\]

iterations. Optimal in \( \epsilon \), but not affine invariant.

Heavily used: TFOCS, NESTA, Structured \( \ell_1 \), \ldots
Choosing norm and prox can have a big impact, beyond the immediate computational cost of computing the prox steps. Consider the following matrix game problem

\[
\begin{align*}
\min_{\{1^T x = 1, x \geq 0\}} & \quad \max_{\{1^T x = 1, x \geq 0\}} x^T A y \\
\end{align*}
\]

- **Euclidean prox.** Pick \( \| \cdot \|_2 \) and \( d(x) = \| x \|_2^2 / 2 \), after regularization, the complexity bound is

\[
N_{\text{max}} = \frac{4\| A \|_2}{N + 1}
\]

- **Entropy prox.** Pick \( \| \cdot \|_1 \) and \( d(x) = \sum_i x_i \log x_i + \log n \), the bound becomes

\[
N_{\text{max}} = \frac{4\sqrt{\log n \log m} \max_{ij} |A_{ij}|}{N + 1}
\]

which can be significantly smaller.

Speedup is roughly \( \sqrt{n} \) when \( A \) is Bernoulli...
Choosing the norm

Invariance means \( \| \cdot \| \) and \( d(x) \) constructed using only \( f \) and the set \( Q \).

Minkovski gauge. Assume \( Q \) is centrally symmetric with non-empty interior.

The Minkowski gauge of \( Q \) is a norm: \( \| x \|_Q \triangleq \inf \{ \lambda \geq 0 : x \in \lambda Q \} \)

**Lemma**

**Affine invariance.** The function \( f(x) \) has Lipschitz continuous gradient with respect to the norm \( \| \cdot \|_Q \) with constant \( L_Q > 0 \), i.e.

\[
f(y) \leq f(x) + \langle \nabla f(x), y - x \rangle + \frac{1}{2}L_Q \| y - x \|_Q^2, \quad x, y \in Q,
\]

if and only if the function \( f(Aw) \) has Lipschitz continuous gradient with respect to the norm \( \| \cdot \|_{A^{-1}Q} \) with the same constant \( L_Q \).

A similar result holds for **strong convexity.** Note that \( \| x \|_Q^* = \| x \|_Q^\circ \).
Choosing the prox.

How do we choose the prox.? Start with two definitions.

**Definition**

**Banach-Mazur distance.** Suppose $\| \cdot \|_X$ and $\| \cdot \|_Y$ are two norms on a space $E$, the **distortion** $d(\| \cdot \|_X, \| \cdot \|_Y)$ is the

smallest product $ab > 0$ such that \[
\frac{1}{b}\|x\|_Y \leq \|x\|_X \leq a\|x\|_Y, \text{ for all } x \in E.
\]

$log(d(\| \cdot \|_X, \| \cdot \|_Y))$ is the Banach-Mazur distance between $X$ and $Y$. 
Choosing the prox.

Regularity constant. Regularity constant of \((E, \| \cdot \|)\), defined in [Juditsky and Nemirovski, 2008] to study large deviations of vector valued martingales.

---

**Definition [Juditsky and Nemirovski, 2008]**

**Regularity constant of a Banach** \((E, \| \cdot \|)\). *The smallest constant \(\Delta > 0\) for which there exists a smooth norm \(p(x)\) such that*

- *The prox \(p(x)^2/2\) has a Lipschitz continuous gradient w.r.t. the norm \(p(x)\), with constant \(\mu\) where \(1 \leq \mu \leq \Delta\),

- *The norm \(p(x)\) satisfies*

\[
\|x\| \leq p(x) \leq \|x\| \left(\frac{\Delta}{\mu}\right)^{1/2}, \quad \text{for all } x \in E
\]

i.e. \(d(p(x), \| \cdot \|) \leq \sqrt{\Delta/\mu}\).
Complexity

Using the algorithm in [Nesterov, 2005] to solve

\[
\begin{align*}
\text{minimize} & \quad f(x) \\
\text{subject to} & \quad x \in Q.
\end{align*}
\]

Proposition [d’Aspremont, Guzman, and Jaggi, 2013]

**Affine invariant complexity bounds.** Suppose \( f(x) \) has a Lipschitz continuous gradient with constant \( L_Q \) with respect to the norm \( \| \cdot \|_Q \) and the space \( (\mathbb{R}^n, \| \cdot \|_Q^* \) \) is \( D_Q \)-regular, then the smooth algorithm in [Nesterov, 2005] will produce an \( \epsilon \) solution in at most

\[
N_{\text{max}} = \sqrt{\frac{4L_QD_Q}{\epsilon}}
\]

iterations. Furthermore, the constants \( L_Q \) and \( D_Q \) are affine invariant.

We can show \( C_f \leq L_QD_Q \), but it is not clear if the bound is attained. . .
Minimizing a smooth convex function over the unit simplex

\[
\begin{align*}
\text{minimize} & \quad f(x) \\
\text{subject to} & \quad 1^T x \leq 1, \ x \geq 0
\end{align*}
\]

in \( x \in \mathbb{R}^n \).

- Choosing \( \| \cdot \|_1 \) as the norm and \( d(x) = \log n + \sum_{i=1}^n x_i \log x_i \) as the prox function, complexity bounded by
  \[
  \sqrt{8 \frac{L_1 \log n}{\epsilon}}
  \]
  (note \( L_1 \) is lowest Lipschitz constant among all \( \ell_p \) norm choices.)

- Symmetrizing the simplex into the \( \ell_1 \) ball. The space \((\mathbb{R}^n, \| \cdot \|_\infty)\) is \( 2 \log n \) regular [Juditsky and Nemirovski, 2008, Ex. 3.2]. The prox function chosen here is \( \| \cdot \|_\alpha^2/2 \), with \( \alpha = 2 \log n/(2 \log n - 1) \) and our complexity bound is
  \[
  \sqrt{16 \frac{L_1 \log n}{\epsilon}}
  \]
In practice

Easy and hard problems.

- The parameter $L_Q$ satisfies

\[ f(y) \leq f(x) + \langle \nabla f(x), y - x \rangle + \frac{1}{2}L_Q\|y - x\|_Q^2, \quad x, y \in Q, \]

On easy problems, $\| \cdot \|$ is large in directions where $\nabla f$ is large, i.e. the sublevel sets of $f(x)$ and $Q$ are aligned.

- For $l_p$ spaces for $p \in [2, \infty]$, the unit balls $B_p$ have low regularity constants,

\[ D_{B_p} \leq \min\{p - 1, 2 \log n\} \]

while $D_{B_1} = n$ (worst case).

- By duality, problems over unit balls $B_q$ for $q \in [1, 2]$ are easier.

- Optimizing over cubes is harder.
Optimality

How good are these bounds?

- Affine invariance does not imply that this complexity bound is tight. . .

- In fact, the worst choice of norm and prox. yields a bound in $\frac{Ld(x^*)}{\sigma}$ that is also affine invariant.

Can we show optimality?
Optimality: upper bounds

Optimizing over $\ell_p$ balls. Focus now on the problem of solving

$$\begin{align*}
\text{minimize} & \quad f(x) \\
\text{subject to} & \quad x \in B_p
\end{align*}$$

in the variable $x \in \mathbb{R}^n$, where $B_p$ is the $\ell_p$ ball. We show that

$$N_{\text{max}} = \sqrt{\frac{4L_p D_p}{\epsilon}}$$

The constants $D_p$ can be computed explicitly (idem for the corresponding norms).

- **When** $p \in [2, \infty]$, we have $D_p = n^{\frac{p-2}{p}}$.

- **When** $p \in [1, 2]$, Juditsky et al. [2009, Ex. 3.2] show

$$D_p = \inf_{2 \leq \rho < \frac{p}{p-1}} (\rho - 1)n^{\rho - \frac{2(p-1)}{p}} \leq \min \left\{ \frac{p}{p-1}, C \log n \right\}$$

where $C > 0$ is an absolute constant.
Optimality: lower bounds

Optimizing over $\ell_p$ balls. In the range $p \in [1, 2]$ the lower bound on risk from Guzmán and Nemirovski [2013] is given by

$$\Omega \left( \frac{L}{T^2 \log[T + 1]} \right)$$

which translates into the following lower bound on iteration complexity

$$\Omega \left( \sqrt{\frac{L}{\epsilon \log n}} \right)$$

Our bound, given by

$$N_{\text{max}} = \sqrt{\frac{4CL \log n}{\epsilon}}$$

where $C > 0$ is an absolute constant, and is thus optimal up to a poly-logarithmic factor.
Optimality: lower bounds

Optimizing over $\ell_p$ balls. In the range $p \in [2, \infty]$ the lower bound on risk from Guzmán and Nemirovski [2013] can be translated to

$$\Omega \left( \sqrt{\frac{ Ln^{1-2/p} }{ \min[p, \log n] \epsilon } } \right).$$

Our bound is then

$$N_{\text{max}} = \sqrt{ \frac{ 4 Ln^{1-2/p} }{ \epsilon } }$$

which is again optimal up to poly-logarithmic factors when $k \sim n$. 
Generalization

- **The Banach space** \((E, \| \cdot \|)\) **is** \((\kappa, r)\) **smooth.** There is \(W(y) : E^* \to \mathbb{R}\) such that \(W(0) = 0\),

\[
W(y) \geq \frac{\|y\|_r^r}{r^*}
\]

and

\[
W(y + z) \leq W(y) + \langle W'(y), z \rangle + \frac{\kappa}{r^*} \|z\|_r^r
\]

- **The function is Hölder smooth**

\[
\|\nabla f(x) - \nabla f(y)\|_{\mu} \leq L \|x - y\|^{\sigma - 1}_{\mu}
\]

The optimal complexity bound, achieved by the algorithm in [Nemirovskii and Nesterov, 1985, Khachiyan et al., 1993], is in this case

\[
O \left( \left( \frac{LR^\sigma}{\epsilon} \right)^{\frac{1}{\mu}} \right), \quad \text{where} \quad \mu = \sigma - 1 + \frac{\sigma(r - 1)}{r}
\]

**Affine invariance:** work in progress...
Outline

- Affine invariant bounds.
- Renegar’s condition number and compressed sensing.
Conic feasibility problems

Alternative **conic linear systems**

\[ Ax = 0, \ x \in C \] \hspace{1cm} (P)

and

\[ -A^T y \in C^* \] \hspace{1cm} (D)

for a given cone \( C \subset \mathbb{R}^p \).
Let $\mathcal{M}^P_{x^*} = \{ A \in \mathbb{R}^{n \times p} : P \text{ is infeasible} \}$, define the **distance to infeasibility**

\[
\rho^P_{x^*}(A) \triangleq \inf_{\Delta A} \{ \| \Delta A \|_2 : A + \Delta A \notin \mathcal{M}^P_{x^*} \}.
\]

**Renegar’s condition number** for problem P with respect to $x^*$ is then defined as the scale-invariant reciprocal of this distance

\[
C^P_{x^*}(A) \triangleq \frac{\| A \|_2}{\rho^P_{x^*}(A)}
\]
■ Renegar’s condition number $C(A)$ and the complexity of solving conic linear systems discussed in [Renegar, 1995, Freund and Vera, 1999b, Epelman and Freund, 2000, Renegar, 2001, Vera et al., 2007, Belloni et al., 2009].

■ In particular, Vera et al. [Vera et al., 2007] link $C(A)$ show that the number of outer barrier method iterations grows as

$$O\left(\sqrt{\nu_C} \log (\nu_C C(A))\right),$$

where $\nu_C$ is the barrier parameter, while the complexity of the linear systems arising at each interior point iteration is controlled by $C(A)^2$. 
Sparse recovery

Sparse recovery problem.

\[
\begin{align*}
\text{minimize} & \quad \|x\| \\
\text{subject to} & \quad \|Ax - y\|_2 \leq \delta \|A\|_2,
\end{align*}
\]

in the variable \(x \in \mathbb{R}^n\).

Define the **conically restricted minimal singular value of** \(A\) as follows

\[
\mu_{x^*}(A) = \inf_{z \in \mathcal{T}(x^*)} \frac{\|Az\|_2}{\|z\|_2}.
\]

where \(\mathcal{T}(x) = \text{cone}\{z : \|x + z\| \leq \|x\|\}\), is the cone of descent directions, then

\[
\|x^* - x_0\|_2 \leq 2 \frac{\delta \|A\|_2}{\mu_{x_0}(A)}.
\]
Theorem [Freund and Vera, 1999a]

**Cone eigenvalues and conditioning.** Distance to feasibility and cone restricted eigenvalues match, i.e. \( \rho^P_{x^*}(A) = \mu_{x^*}(A) \).

Generalizes to a much broader class of recovery problems [Roulet, Boumal, and d’Aspremont, 2015].
Sparse recovery

Classical condition number $\kappa(A)$

Condition number $C_{x_0}(A)$ (lower bound)

Estimation error, L1-Hom., noisy

Exact recovery probability, noiseless

CPU time in lsq solves, L1-Hom., noiseless

#iterations, L1-Hom., noiseless

#iterations, LARS, noiseless

#iterations, TFOCS-BP, noiseless

Alex d'Aspremont Institut des Hautes Études Scientifiques, March. 2016. 33/34
Conclusion

- **Affine invariant** complexity bound for the optimal algorithm [Nesterov, 1983]

\[ N_{\text{max}} = \sqrt{\frac{4L_QD_Q}{\epsilon}} \]

Matches (up to polylog terms) best known lower bounds on \( \ell_p \)-balls.

- Data-driven complexity measure for **sparse recovery problems**, matching statistical performance measures.

Open problems.

- Optimality of product \( L_QD_Q \) in the general case?
- Matches curvature \( C_f \)?
- Best norm choice for non-symmetric sets \( Q \)?
- Systematic, tractable procedure for smoothing \( Q \)?
References


