An Optimal Affine Invariant Smooth Minimization Algorithm.

Alexandre d'Aspremont, *CNRS & École Polytechnique*.

Joint work with Martin Jaggi. Support from ERC SIPA.

A Basic Convex Problem

Solve

$$\begin{array}{ll} \text{minimize} & f(x) \\ \text{subject to} & x \in Q, \end{array}$$

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

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

Assume that

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

[Nesterov and Nemirovskii, 1994] Newton's method produces an ϵ optimal solution to the barrier problem

$$\min_x h(x) \triangleq f(x) + t\, g(x)$$

for some $t > 0$, in at most

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

where $0 < \alpha < 0.5$ and $0 < \beta < 1$ are line search parameters.

Complexity

Newton's method. Basically

$$\#\text{ Newton iterations} \quad \leq \quad 375\,(h(x_0) - h^*) + 6$$

- Empirically valid, up to constants.
- **Independent from the dimension $\mathbf{n}$.**
- **Affine invariant.**

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

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

Affine Invariance

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

$$\begin{array}{ll}\text{minimize} & f(x)\\ \text{subject to} & x \in Q,\end{array} \quad \textbf{becomes} \quad \begin{array}{ll}\text{minimize} & \hat{f}(y)\\ \text{subject to} & y \in \hat{Q},\end{array}$$

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 $\Delta x_{\mathrm{nt}} = A\Delta y_{\mathrm{nt}}$
- **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.**

"Real men/women solve optimization problems with terabytes of data."

(Michael Jordan, Paris, 2013.)

- 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?

Frank-Wolfe

Conditional gradient. At each iteration, solve

$$\begin{array}{ll} \text{minimize} & \langle \nabla f(x_k), u \rangle \\ \text{subject to} & u \in Q \end{array}$$

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

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

The Franke-Wolfe algorithm will then produce an ϵ solution after

$$N_{\max} = \frac{4C_f}{\epsilon}$$

iterations.

- **C_f is affine invariant** but the bound is **suboptimal in ϵ.**
- If $f(x)$ has a Lipschitz gradient, the lower bound is $O\left(\frac{1}{\sqrt{\epsilon}}\right)$.

Optimal First-Order Methods

Smooth Minimization algorithm in [Nesterov, 1983] to solve

$$\begin{array}{ll} \text{minimize} & f(x) \\ \text{subject to} & x \in Q, \end{array}$$

- **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 $\sigma > 0$.

Optimal First-Order Methods

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 $y_k = \mathrm{argmin}_{y \in Q} \left\{ \langle \nabla f(x_k), y - x_k \rangle + \frac{1}{2} L \|y - x_k\|^2 \right\}$.

4: Compute $z_k = \mathrm{argmin}_{x \in Q} \left\{ \sum_{i=0}^{k} \alpha_i [f(x_i) + \langle \nabla f(x_i), x - x_i \rangle] + \frac{L}{\sigma} d(x) \right\}$.

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 ϵ-solution in at most

$$\sqrt{\frac{8L}{\epsilon} \frac{d(x^\star)}{\sigma}}$$

iterations. **Optimal in ϵ, but not affine invariant.**

Heavily used: TFOCS, NESTA, Structured ℓ_1, . . .

Optimal First-Order Methods

Choosing norm and prox can have a big impact. Consider the following matrix game problem

$$\min_{\{\mathbf{1}^T x = 1, x \geq 0\}} \max_{\{\mathbf{1}^T x = 1, x \geq 0\}} x^T A y$$

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

$$N_{\max} = \frac{4\|A\|_2}{N+1}$$

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

$$N_{\max} = \frac{4\sqrt{\log n \log m}\, \max_{ij} |A_{ij}|}{N+1}$$

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*

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

$\log(d(\|\cdot\|_X, \|\cdot\|_Y))$ 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 $(E, \|.\|)$. *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 μ 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. $d(p(x), \|.\|) \leq \sqrt{\Delta/\mu}$.

Complexity

Using the algorithm in [Nesterov, 2005] to solve

$$\begin{array}{ll} \text{minimize} & f(x) \\ \text{subject to} & x \in Q. \end{array}$$

Proposition [d'Aspremont 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 ϵ solution in at most*

$$N_{\max} = \sqrt{\frac{4L_Q D_Q}{\epsilon}}$$

iterations. Furthermore, the constants L_Q and D_Q are ***affine invariant.***

We can show $C_f \leq L_Q D_Q$, but it is not clear if the bound is attained. . .

Complexity

A few more facts about L_Q and D_Q...

Suppose we **scale $Q \rightarrow \alpha Q$,** with $\alpha > 0$,

- the Lipschitz constant $L_{\alpha Q}$ satisfies $\alpha^2 L_Q \leq L_{\alpha Q}$.
- the smoothness term D_Q remains unchanged.
- Given our choice of norm (hence L_Q), $L_Q D_Q$ is the best possible bound.

Also, from [Juditsky and Nemirovski, 2008], in the dual space

- The regularity constant decreases on a subspace F, i.e. $D_{Q \cap F} \leq D_Q$.
- From D regular spaces $(E_i, \|\cdot\|)$, we can construct a $2D+2$ regular product space $E \times \ldots \times E_m$.

Complexity, ℓ_1 example

Minimizing a **smooth convex function over the unit simplex**

$$\begin{array}{ll} \text{minimize} & f(x) \\ \text{subject to} & \mathbf{1}^T x \leq 1,\ x \geq 0 \end{array}$$

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 ℓ_p norm choices.)

- Symmetrizing the simplex into the ℓ_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 ∇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.

Conclusion

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

$$N_{\max} = \sqrt{\frac{4 L_Q D_Q}{\epsilon}}$$

- Matches best known bounds on key examples.

Open problems.

- Prove optimality of product $L_Q D_Q$. Matches curvature C_f?
- Symmetrize non-symmetric sets Q.
- Systematic, tractable procedure for smoothing Q.

*

References

Alexandre d'Aspremont and Martin Jaggi. An optimal affine invariant smooth minimization algorithm. *arXiv preprint arXiv:1301.0465*, 2013.

A. Juditsky and A.S. Nemirovski. Large deviations of vector-valued martingales in 2-smooth normed spaces. *arXiv preprint arXiv:0809.0813*, 2008.

Y. Nesterov. A method of solving a convex programming problem with convergence rate $O(1/k^2)$. *Soviet Mathematics Doklady*, 27(2): 372–376, 1983.

Y. Nesterov. Smooth minimization of non-smooth functions. *Mathematical Programming*, 103(1):127–152, 2005.

Y. Nesterov and A. Nemirovskii. *Interior-point polynomial algorithms in convex programming*. Society for Industrial and Applied Mathematics, Philadelphia, 1994.