JMLR: Workshop and Conference Proceedings vol 40:1–3, 2015

Open Problem: The Oracle Complexity of Smooth Convex Optimization in Nonstandard Settings

Cristóbal Guzmán CGUZMAN@GATECH.EDU
ISyE, Georgia Institute of Technology
755 Ferst Drive NW, Atlanta GA, 30332-0205

Abstract

First-order convex minimization algorithms are currently the methods of choice for large-scale sparse – and more generally parsimonious – regression models. We pose the question on the limits of performance of black-box oriented methods for convex minimization in *non-standard settings*, where the regularity of the objective is measured in a norm not necessarily induced by the feasible domain. This question is studied for ℓ_p/ℓ_q-settings, and their matrix analogues (Schatten norms), where we find surprising gaps on lower bounds compared to state of the art methods. We propose a conjecture on the optimal convergence rates for these settings, for which a positive answer would lead to significant improvements on minimization algorithms for parsimonious regression models.

Let $(\mathbf{E}, \|\cdot\|)$ be a finite-dimensional normed space. Given parameters, $1 < \kappa \leq 2$, and $L > 0$, we consider the class $\mathcal{F}_{\|\cdot\|}(\kappa, L)$ of convex functions that are (κ, L)-smooth w.r.t. norm $\|\cdot\|$. One such function $f : \mathbf{E} \to \mathbb{R}$ satisfies that

$$\|\nabla f(y) - \nabla f(x)\|_* \leq L\|x - y\|^{\kappa-1} \quad \forall x, y \in \mathbf{E}.$$

Notice that the case $\kappa \to 1$ corresponds essentially to nonsmooth (Lipschitz continuous) convex functions, and $\kappa = 2$ corresponds to smooth (with Lipschitz continuous gradients) convex functions.

Given a convex body $X \subseteq \mathbf{E}$, we are interested on the complexity of the problem class $\mathcal{P} = (\mathcal{F}_{\|\cdot\|}(\kappa, L), X)$, comprised of optimization problems with objective $f \in \mathcal{F}_{\|\cdot\|}(\kappa, L)$

$$\text{Opt}(f, X) = \min\{f(x) : \ x \in X\}. \qquad (P_{f,X})$$

We study a black-box oracle model where most algorithms based on subgradient computations can be implemented[1]. Here, an algorithm is allowed to perform queries $x \in \mathbf{E}$, and for any such query the oracle returns $\mathcal{O}_f(x)$ (e.g., for first-order methods, $\mathcal{O}_f(x) = \nabla f(x)$). The only assumption on the oracle is *locality*: For any $x \in \mathbf{E}$ and $f, g \in \mathcal{F}$ such that there exists a neighborhood $B(x, \delta)$ where $f \equiv g$, then $\mathcal{O}_f(x) = \mathcal{O}_g(x)$.

Given $T > 0$, we consider an algorithm $\mathcal{A}$ whose output $x^T(\mathcal{A}, f)$ is only determined by T (adaptive) oracle queries. We define the accuracy of algorithm $\mathcal{A}$ on an instance f as $\varepsilon(\mathcal{A}, f) := f(x^T(\mathcal{A}, f)) - \text{Opt}(f, X)$, if $x^T(\mathcal{A}, f) \in X$, otherwise $\varepsilon(\mathcal{A}, f) = \infty$. We characterize optimal

[1] Notable exceptions are methods exploiting explicit saddle-point description, e.g., the smoothing technique by Nesterov (2005). Note however that such algorithms do not give improvement in the smooth case.

algorithms in terms of the *minimax risk*,

$$\mathrm{Risk}(\mathcal{P}) = \inf_{\mathcal{A}} \sup_{f\in\mathcal{F}} \varepsilon(\mathcal{A}, f),$$

where the infimum is taken among all algorithms $\mathcal{A}$ performing only T queries. From now on, we restrict our attention to the large-scale regime, where $T \leq n$.

Until very recently, tools for studying the oracle complexity of convex optimization beyond the nonsmooth case were scarce. The only cases where there were nontrivial lower bounds were given by convex quadratic minimization over a feasible domain given by an Euclidean ball. In Guzmán and Nemirovski (2015) and Guzmán (2015) we provide new techniques and lower bounds for oracle complexity of convex optimization.

Our results have interesting consequences for the ℓ_p/ℓ_q-setting, $1 \leq p, q \leq \infty$, where $X = R\,B_p^n$ is the ℓ_p^n-ball of radius R, and $(\mathbf{E}, \|\cdot\|) = (\mathbb{R}^n, \|\cdot\|_q)$. In short, the obtained risk lower bounds, and the ratio w.r.t. fastest known methods, $\mathcal{R}(T)$, can be seen in Table 1.

Range q	Range p	Risk Lower bound	$\mathcal{R}(T)$
$1 \leq q < 2$	$p < q$	$\Omega\left(\frac{1}{[\ln n]^{\kappa-1}} \frac{LR^\kappa}{T^{\kappa[\frac{3}{2}+\frac{1}{p}-\frac{1}{q}]-1}}\right)$	$\tilde{O}\left(T^{\kappa[\frac{1}{p}-\frac{1}{q}]}\right)$
	$p \geq q$	$\Omega\left(\frac{n^{\kappa(\frac{1}{q}-\frac{1}{p})}}{[\ln n]^{\kappa-1}} \frac{LR^\kappa}{T^{\frac{3\kappa}{2}-1}}\right)$	$\tilde{O}(1)$
$2 \leq q \leq \infty$	$p < q$	$\Omega\left(\frac{1}{[\min\{q, \ln n\}]^{\kappa-1}} \frac{LR^\kappa}{T^{\kappa[1+\frac{1}{p}]-1}}\right)$	$\tilde{O}\left(T^{\kappa[\frac{1}{p}-\frac{1}{q}]}\right)$
	$p \geq q$	$\Omega\left(\frac{n^{\kappa(\frac{1}{q}-\frac{1}{p})}}{[\min\{q, \ln n\}]^{\kappa-1}} \frac{LR^\kappa}{T^{\kappa[1+\frac{1}{q}]-1}}\right)$	$\tilde{O}(1)$

Table 1: Minimax risk lower bounds for (κ, L)-smooth convex optimization in the ℓ_p/ℓ_q-setting.

Similarly we can study the p/q Schatten norm setting[2], where $(\mathbf{E}, \|\cdot\|) = (\mathbb{R}^{n\times n}, \|\cdot\|_{\mathsf{Sch}_q})$, and $X = \{x \in \mathbb{R}^{n\times n} : \ \|x\|_{\mathsf{Sch}_p} \leq R\}$. We refer the reader to Guzmán (2015) for further discussions.

The reader may observe that when $p \geq q$ lower bounds are nearly-tight. However, when $q > p$ we observe substantial gaps in the complexity, which worsen as the smoothness parameter κ grows. Best existing algorithms defining the ratio $\mathcal{R}(T)$ are versions of Nesterov's method Nemirovskii and Nesterov (1985) for the standard ℓ_q/ℓ_q-setting, which do not exploit the geometry of the much smaller ℓ_p^n-ball when $p < q$. This observation, together with the gaps above, lead us to the following

[2] Recall that the Schatten p-norm of a matrix $x \in \mathbb{R}^{n\times n}$ is given by $\|x\|_{\mathsf{Sch}_p} := \left[\sum_i \sigma_i(x)^p\right]^{1/p}$, where $\sigma_1(x), \ldots, \sigma_n(x)$ are the singular values of X.

Open Problem 1 *What is the large-scale minimax risk of minimization within the class $\mathcal{F}_{\|\cdot\|_q}(\kappa, L)$ over the unit ball of ℓ_p^n in the black-box oracle model? Can the rate obtained by Nesterov's method be significantly improved?*

We finish this discussion by showing the importance of closing these gaps for certain classes of regression problems. Our main motivation is the study of linear regression models, where we search for a linear predictor within a norm-bounded set X, e.g., $X \subseteq B_p^n$; and the performance of a predictor is measured by a loss function arising from random samples $(a_1, b_1), \ldots, (a_m, b_m) \in B_{q*}^n \times [-1, 1]$. Thus the empirical risk minimization problem we obtain, $\min\{\frac{1}{m}\sum_{j=1}^m (a_j'x - b_j)^2 : \|x\|_p \leq 1\}$, fits within the ℓ_p/ℓ_q-setting, with $\kappa = 2$ and $L = R = 1$.

Perhaps the most important application of the regression model above is the case of compressed sensing, where $p = 1$ and $q = 2$ (we can also consider the matrix analog of nuclear norm minimization for low-rank matrix recovery). In this case, Nesterov's method gives a rate of convergence $O(1/T^2)$, whereas our lower bound is $\tilde{\Omega}(1/T^3)$. Our conjecture says that the optimal convergence rate here is better than $O(1/T^2)$, although to the best of our knowledge, results on sublinear algorithms beyond this rate are nonexisting[3]. In this sense, we believe surpassing the $O(1/T^2)$ rate is indeed an extremely challenging problem.

Finally, we believe any progress on designing faster algorithms in this setting, will not only have an impact on compressed sensing and low-rank matrix recovery, but more broadly in convex minimization methods for parsimonious regression models, and possibly to the stochastic and online settings.

Acknowledgments

I would like to thank Alexandre d'Aspremont and Arkadi Nemirovski for valuable discussions leading to this open problem.

References

A. Agarwal, S. Negahban, and M. Wainwright. Fast global convergence of gradient methods for high-dimensional statistical recovery. *The Annals of Statistics*, 40(5):2452–2482, 10 2012.

C. Guzmán. *Information, Complexity and Structure in Convex Optimization*. PhD thesis, Georgia Institute of Technology, May 2015.

C. Guzmán and A. Nemirovski. On lower complexity bounds for large-scale smooth convex optimization. *Journal of Complexity*, 31(1):1 – 14, 2015.

A. Nemirovskii and Y. Nesterov. Optimal methods of smooth convex optimization *(in Russian)* . *Zh. Vychisl. Mat. i Mat. Fiz.*, 25(3):356–369, 1985.

Y. Nesterov. Smooth Minimization of Non-Smooth Functions. *Mathematical Programming*, 103 (1):127–152, 2005.

[3] Notice that in the case the objective satisfies a *restricted strong convexity* property then projected gradient descent converges linearly up to the statistical error Agarwal et al. (2012). Our open problem considers models where we do not have such nice properties.