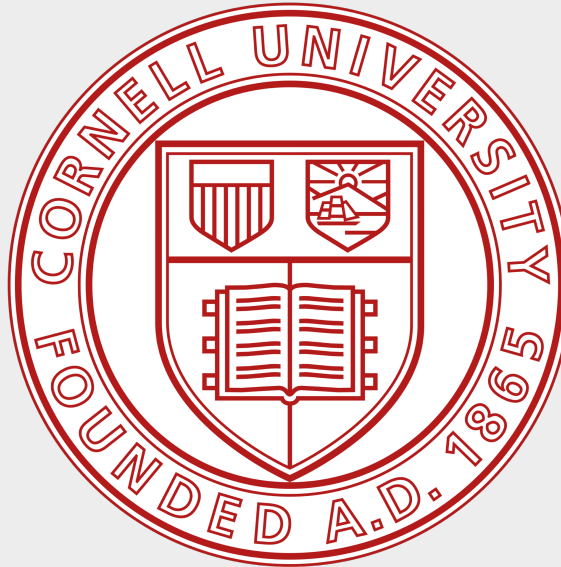


 HELMHOLTZ CENTER FOR INFORMATION SECURITY	Efficient Gradient Methods for Functions with Super-Quadratic Curvature			
	Anton Rodomanov ¹ Nikita Doikov ² Sebastian Stich ¹			
	¹ CISPA Helmholtz Center for Information Security, Germany ² Cornell University, USA			

Introduction
We consider the optimization problem $f^* := \min_{x \in \mathbb{R}^n} f(x), \tag{P}$ where f is differentiable. Assume (P) has a solution x^* and denote $R := \ x_0 - x^*\ ,$ where $x_0 \in \mathbb{R}^n$ is a given starting point, $\ \cdot\ $ is the standard Euclidean norm.
■ We show that, in usual Gradient Descent (GD) and Fast Gradient Method, we can assume convexity and smoothness on the ball $B(x^*, R)$ instead of globally: $f(y) \geq f(x) + \langle \nabla f(x), y - x \rangle, \quad \forall x, y \in B(x^*, R), \tag{A1}$$\ \nabla f(x) - \nabla f(y)\ \leq L_*(R)\ x - y\ , \quad \forall x, y \in B(x^*, R), \tag{A2}$ and all classic convergence results still hold with global smoothness constant L replaced by $L_*(R)$.
■ One application is the minimization of an RGH function (defined later). We show that GD is as efficient as the Bregman Method (BM) with a properly chosen prox-function, while being much simpler and more practical.

Gradient Descent (GD)
Gradient step: $T_M(x) := x - \frac{1}{M}\nabla f(x)$.
Basic method with constant step size: $x_{k+1} = T_M(x_k), \quad M \equiv L_*(R).$
In this method, $x_k \in B(x^*, R)$ for all $k \geq 0$.
Efficiency: At most the following number of steps to get $f(x_{k+1}) - f^* \leq \epsilon$: $N_{\text{GD}}(\epsilon) = \frac{L_*(R)R^2}{\epsilon}.$

Gradient Descent with Line Search (GD-LS)
Step sizes can be selected automatically, no need to know $L_*(R)$:
Choose $\hat{M}_0 > 0$. for $k = 0, 1, 2, \dots$ do Find the smallest integer $i_k \geq 0$ such that, for $M_k = 2^{i_k}\hat{M}_k, \quad x_{k+1} = T_{M_k}(x_k),$ we have $f(x_{k+1}) \leq f(x_k) + \langle \nabla f(x_k), x_{k+1} - x_k \rangle + \frac{M_k}{2}\ x_{k+1} - x_k\ ^2$. Set $\hat{M}_{k+1} = \frac{1}{2}M_k$.
This method also ensures $x_k \in B(x^*, R)$ for all $k \geq 0$.
Efficiency: At most the following number of oracle queries: $N_{\text{GD-LS}}(\epsilon) = \frac{M_*(R)R^2}{\epsilon} + \log_2 \frac{4M_*(R)}{\hat{M}_0},$ where $M_*(R) = \max\{2L_*(R), \hat{M}_0\}$. In particular, if $\hat{M}_0 \leq 2L_*(R)$, then $N_{\text{GD-LS}}(\epsilon) \simeq N_{\text{GD}}(\epsilon)$, up to an additive logarithmic factor $\log \frac{L_*(R)}{\hat{M}_0}$.

Fast Gradient Method (FGM)
$v_0 = x_0, \quad A_0 = 0,$$M_k a_{k+1}^2 = A_k + a_{k+1}, \quad A_{k+1} = A_k + a_{k+1},$$y_k = \frac{A_k x_k + a_{k+1} v_k}{A_{k+1}}, \quad x_{k+1} = T_{M_k}(y_k), \quad v_{k+1} = v_k - a_{k+1} \nabla f(y_k),$
where M_k ensures $f(x_{k+1}) \leq f(y_k) + \langle \nabla f(y_k), x_{k+1} - y_k \rangle + \frac{M_k}{2}\ x_{k+1} - y_k\ ^2$. (*)
Basic version: $M_k \equiv L_*(R)$. Then, (*) is satisfied, $x_k, y_k, v_k \in B(x^*, R)$, and $N_{\text{FGM}}(\epsilon) = \sqrt{\frac{2L_*(R)R^2}{\epsilon}}.$
Can also use line search (similar to GD-LS). Number of oracle queries: $N_{\text{FGM-LS}}(\epsilon) = 4\sqrt{\frac{2M_*(R)R^2}{\epsilon}} + \log_2 \frac{2M_*(R)}{\hat{M}_0}.$

Main Technical Result
Main result: If $x \in B(x^*, R)$ and $M \geq L_*(R)$, then $T_M(x) \in B(x^*, R)$.
Key lemma: Let f be convex and $L_{\bar{x}}(r)$ -smooth on $B(\bar{x}, r)$. Then, $f(\bar{x}) - f(x) - \langle \nabla f(x), \bar{x} - x \rangle \geq \frac{1}{2L_{\bar{x}}(r)}\ \nabla f(\bar{x}) - \nabla f(x)\ ^2, \quad \forall x \in B(\bar{x}, r).$
Extends a well-known result for globally convex and Lipschitz-smooth function.

Functions with Radially Growing Hessian (RGH)
RGH function: twice differentiable f is called RGH around $\bar{x} \in \mathbb{R}^n$ if $\ \nabla^2 f(x)\ \leq L_{\bar{x}}(\ x - \bar{x}\), \quad \forall x \in \mathbb{R}^n,$ where $L_{\bar{x}}(\cdot)$ is a non-decreasing function.
Basic examples: ■ $f(\tau) = \tau ^p, p > 2 \implies L_0(r) = p(p-1)r^{p-2}$. ■ $f(\tau) = e^\tau \implies L_0(r) = e^r$. ■ $f(\tau) = e^{\tau^2} \implies L_0(r) = 2(1+2r^2)e^{r^2}$. ■ Radial function $f(x) = \psi(\ x - \bar{x}\) \implies L_{\bar{x}}(r) = \psi''(r)$, assuming $\psi'(0) = 0$ and $\psi''(\cdot)$ is non-negative and non-decreasing. In particular: ■ $f(x) = \ x - \bar{x}\ ^p, p > 2 \implies L_{\bar{x}}(r) = p(p-1)r^{p-2}$. ■ $f(x) = e^{\ x - \bar{x}\ ^2} \implies L_{\bar{x}}(r) = 2(1+2r^2)e^{r^2}$.
Basic calculus rules: ■ $f(x) = cg(x), c \in \mathbb{R} \implies L_{\bar{x}}(r) = c L_{g,\bar{x}}(r)$. ■ $f(x) = f_1(x) + f_2(x) \implies L_{\bar{x}}(r) = L_{f_1,\bar{x}}(r) + L_{f_2,\bar{x}}(r)$. ■ $f(x) = g(Ax + b) \implies L_{\bar{x}}(r) = \alpha^2 L_{g,\bar{y}}(\alpha r + \Delta)$, where $\alpha := \ A\$ and $\Delta := \ A\bar{x} + b - \bar{y}\$.
More interesting examples: ■ $f(x) = \frac{1}{2}\ A_1 x - b_1\ ^2 + \frac{1}{p}\ A_2 x - b_2\ ^p, p > 2$: $\implies L_{\bar{x}}(r) = \ A_1\ ^2 + (p-1)\ A_2\ ^2(\ A_2\ r + \ A_2\bar{x} - b_2\)^{p-2}.$ ■ $f(x) = \sum_{i=1}^m e^{\langle a_i, x \rangle + b_i} \implies L_{\bar{x}}(r) = \sum_{i=1}^m \alpha_i^2 e^{\alpha_i r + \Delta_i}$, where $\alpha_i = \ a_i\$ and $\Delta_i = \langle a_i, \bar{x} \rangle + b_i$.

Application: Minimizing RGH Function
Consider problem (P), where f is a convex RGH function around x_0 : $\ \nabla^2 f(x)\ \leq L(\ x - x_0\), \quad \forall x \in \mathbb{R}^n.$ According to our new theory, the complexity of GD on this problem is $N_{\text{GD}}(\epsilon) = \frac{L_*(R)R^2}{\epsilon} \simeq \frac{L(R)R^2}{\epsilon}$ since $L_*(R) \leq L(2R)$ and $L(R) \leq L_*(2R)$.

Comparison: Bregman Method
Relative smoothness [BBT17; LFN18]: $\nabla^2 f(x) \preceq L_d \nabla^2 d(x), \forall x \in \mathbb{R}^n$. Bregman distance: $\beta_d(x, y) := d(y) - d(x) - \langle \nabla d(x), y - x \rangle$. Bregman Method (BM): $x_{k+1} = \arg\min_{x \in \mathbb{R}^n} \{\langle \nabla f(x_k), x \rangle + L_d \beta_d(x_k, x)\}$. Prox-function: $d(x) = \psi(\ x - x_0\), \quad \psi(r) = \int_0^r \tau L(\tau) d\tau.$ ■ The above $\psi(\cdot)$ is the best possible modulus (in terms of the resulting complexity estimate) ensuring the relative smoothness condition and satisfying the normalization conditions $L_d = 1, d(x_0) = 0, \nabla d(x_0) = 0$. ■ For $L(r) = \alpha_0 + \alpha_p r^p$, we get $\psi(r) = \frac{\alpha_0}{2}r^2 + \frac{\alpha_p}{p+2}r^{p+2}$, which is an existing recommendation from [LFN18].
Implementing BM: Find r_{k+1} from the non-linear equation $L(r)r = \ L(r_k)(x_k - x_0) - \nabla f(x_k)\ ,$ and then compute $x_{k+1} = x_k - \frac{1}{L(r_{k+1})}\nabla f(x_k) - \left[1 - \frac{L(r_k)}{L(r_{k+1})}\right](x_k - x_0).$
Complexity: $N_{\text{BM}}(\epsilon) = \frac{L_d R_d^2}{2\epsilon} = \frac{d(x^*)}{\epsilon} = \frac{\psi(R)}{\epsilon} \simeq \frac{L(R)R^2}{\epsilon},$ where $\frac{1}{2}R_d^2 := \beta_d(x_0, x^*)$ and $\frac{3}{8}L(\frac{R}{2})R^2 \leq \psi(R) \leq \frac{1}{2}L(R)R^2$.
Comparison with GD: ■ BM has essentially the same efficiency guarantee as GD. ■ However, GD is simpler: no need to solve non-linear equations; needs only a scalar $L_*(R)$ in contrast to an entire function $L(\cdot)$ in BM. ■ Knowing $L(\cdot)$ is hard: it depends on the function's structure and usually involves expensive-to-compute matrix norms. ■ Knowing $L_*(R)$ is even more difficult, but we do not need to (line search). For BM, line search is possible, but we cannot prove anything “good”. ■ GD can be accelerated $\implies$ FGM. For BM, this is an open question.

References
[BBT17] Heinz H. Bauschke, Jérôme Bolte, and Marc Teboulle. “A Descent Lemma Beyond Lipschitz Gradient Continuity: First-Order Methods Revisited and Applications”. In: <i>Mathematics of Operations Research</i> 42.2 (2017), pp. 330–348.
[LFN18] Haihao Lu, Robert M. Freund, and Yurii Nesterov. “Relatively Smooth Convex Optimization by First-Order Methods, and Applications”. In: <i>SIAM Journal on Optimization</i> 28.1 (2018), pp. 333–354.