Local Curvature Descent: Squeezing More Curvature out of Standard and Polyak GD

Peter Richtárik Simone Maria Giancola Dymitr Lubczyk Robin Yadav

King Abdullah University of Science and Technology

KAUST

Local Curvature

- We revisit the standard convex optimization problem

$$\min_{x\in\mathbb{R}^d} f(x) \qquad f_\star := f(x_\star) \qquad x_\star \in \mathcal{X}_\star$$

- Our aim is to design adaptive matrix-valued step sizes without going the route of fully-fledged second-order methods
- **Our key observation:** For some problems, certain **local curvature information** is available which can be used to obtain powerful matrix step sizes
- **(Convexity and smoothness with local curvature)** We propose a new class of functions based on a given positive semi-definite curvature mapping

$$\mathbf{C} : \mathbb{R}^d \to \mathbb{S}^d_+$$

$$\underbrace{f(y) + \langle \nabla f(y), x-y\rangle + \tfrac{1}{2}\|x-y\|^2_{\mathbf{C}(y)}}_{M^{\text{low}}_{\mathbf{C}}(x;y)} \leq f(x) \qquad (5) \qquad \forall x,y \in \mathbb{R}^d$$

$$f(x) \leq \underbrace{f(y) + \langle \nabla f(y), x-y\rangle + \tfrac{1}{2}\|x-y\|^2_{\mathbf{C}(y)+L_{\mathbf{C}}\cdot\mathbf{I}}}_{M^{\text{up}}_{\mathbf{C}}(x;y)} \qquad (6) \qquad \forall x,y \in \mathbb{R}^d$$

Three New Algorithms

Local Curvature Descent 1 (LCD1)

- Given our assumption, we can derive an analogue to **GD** by minimizing the upper bound

$$\boxed{x_{k+1} = x_k - \left[\mathbf{C}(x_k) + L_{\mathbf{C}}\cdot\mathbf{I}\right]^{-1}\nabla f(x_k)}$$

- **LCD1** is not an adaptive algorithm and reduces to GD when the curvature matrix vanishes

Local Curvature Descent 2 (LCD2)

- Our first adaptive algorithm is an extension of **GD** with Polyak step size enhanced with a local curvature matrix
- We can start by defining a **localization set**:

$$\mathcal{L}_{\mathbf{C}}(y) := \left\{x\in\mathbb{R}^d \;:\; M^{\text{low}}_{\mathbf{C}}(x,y) \leq f_\star\right\}$$

$$\mathcal{X}_\star \subseteq \mathcal{L}_{\mathbf{C}}(y)$$

- The localization set separates $\mathbb{R}^d$ into two regions, one contains the set of minimizers and the other contains the current iterate $y = x_k$
- **LCD2** simply projects the current iterate onto the localization set, bringing us closer to the set of minimizers:

$$\boxed{x_{k+1} = \arg\min_{x\in\mathcal{L}_{\mathbf{C}}(x_k)} \tfrac{1}{2}\|x-x_k\|^2}$$

- The minimization problem has a parametric solution which can be computed using root-finding algorithms

$$x_{k+1} = x_k - \left[\mathbf{C}(x_k) + \beta_k\cdot\mathbf{I}\right]^{-1}\nabla f(x_k)$$

- **LCD2** has a closed-form update step when the curvature matrix is a matrix of rank one, or a multiple of the identity

Local Curvature Descent 3 (LCD3)

- **LCD3** does *not* require executing a subroutine to compute the update step:

$$\boxed{x_{k+1} = \arg\min_{x\in\mathcal{L}_{\mathbf{C}}(x_k)} \tfrac{1}{2}\|x-x_k\|^2_{\mathbf{C}(x_k)}}$$

- We can obtain a closed-form update step by changing the norm in which we project:

$$x_{k+1} = x_k - \left(1 - \sqrt{1 - \frac{2(f(x_k)-f_\star)}{\|\nabla f(x_k)\|^2_{\mathbf{C}^{-1}(x_k)}}}\right)\mathbf{C}^{-1}(x_k)\nabla f(x_k)$$

- We do not provide a convergence theorem for **LCD3** due to the variable nature of the norm (open problem)

Convergence Rates

- **(Convergence of LCD1)** The iterates of **LCD1** satisfy

$$f(x_k) - f_\star \leq \frac{L_{\mathbf{C}}\|x_0 - x_\star\|^2}{2k}$$

- This result extends the reach of classical theorems since it is possible for a function to satisfy our assumption but not be L-smooth
- If a function is L-smooth and convex, then we may obtain improved complexity up to a constant because

$$\inf_{x\in\mathbb{R}^d}\lambda_{\min}(\mathbf{C}(x)) \leq L - L_{\mathbf{C}} \leq \sup_{x\in\mathbb{R}^d}\lambda_{\max}(\mathbf{C}(x)).$$

- **(Convergence of LCD2)** The iterates of **LCD2** *satisfy*

$$\min_{1\leq t\leq k} f(x_t) - f_\star \leq \frac{L_{\mathbf{C}}\|x_0 - x_\star\|^2}{2k}$$

- Our results recover the standard rate of **GD** with constant step size and **GD** with Polyak step size when

$$\mathbf{C}(x) \equiv \mathbf{0}, \qquad L_{\mathbf{C}} = L$$

- **LCD1** and **LCD2** reduce to Newton's method for convex quadratics and converge in one step, which is predicted by our theorems!

Examples

- Using a variable norm induced by the curvature mapping allows us to adjust the tightness of the quadratic lower bound
- Consider the square of the Huber loss, which is *not strongly-convex*

$$h(x) = \begin{cases} \frac{1}{2}x^2 & |x|\leq\delta \\ \delta(|x| - \frac{1}{2}\delta) & |x|>\delta \end{cases} \qquad \mathbf{C}(x) = \begin{cases} x^2 & |x|\leq\delta \\ \delta^2 & |x|>\delta \end{cases} \qquad L_{\mathbf{C}} = 2\delta^2$$

- Squared p-norms also satisfy our assumption when $p \geq 2$, with a diagonal curvature matrix given by:

$$\mathbf{C}(x) = \frac{2}{\|x\|_p^{p-2}}\operatorname{Diag}\left(|x_1|^{p-2},\ldots,|x_d|^{p-2}\right) \qquad \mathbf{C}(x) = 2\nabla f(x)\nabla f(x)^\top$$

- We introduce the class of **absolutely convex functions**

$$\phi(x) \geq |\phi(y) + \langle\nabla\phi(y), x-y\rangle| \quad \forall x,y\in\mathbb{R}^d$$

- When minimizing the sum of squares of absolutely convex functions, the curvature matrix is readily available:

$$x_\star = \arg\min_{x\in\mathbb{R}^d}\left\{f(x) := \frac{1}{n}\sum_{i=1}^n \phi_i^2(x)\right\} \qquad \mathbf{C}(x) = \frac{2}{n}\sum_{i=1}^n \nabla\phi_i(x)\nabla\phi_i(x)^\top$$

- Local curvature calculus is similar to calculus of convex functions, which can be used to derive a variety of additional examples. Example:

If f satisfies (5) and g is convex, then $h := f + g$ also satisfies (5).

Experiments

- We consider logistic regression with regularization, where the curvature matrix is derived from the regularization term
- Increasing the regularization weight improves the performance of **LCD2** over **GD** with Polyak step size
- **LCD2** beats **GD** with Polyak step size on wall clock time even when using a subroutine

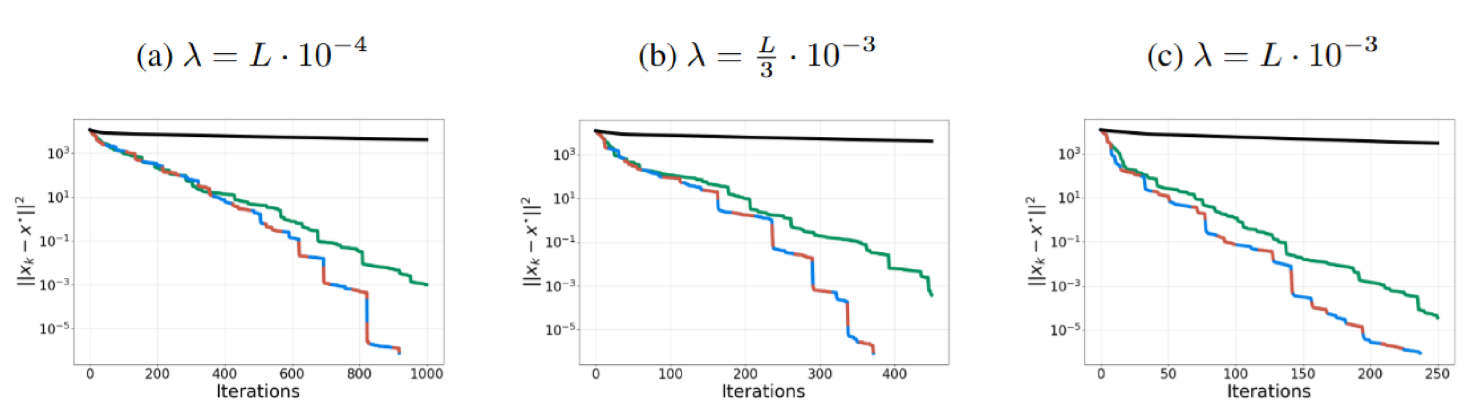

Figure 2: Logistic regression on mushrooms dataset with L_2 regularization.

Figure 4: Logistic regression on mushrooms dataset with L_3 regularization - time convergence.