CS794/CO673: Optimization for Data Science

Lec 08: Metric Gradient

Yaoliang Yu

UNIVERSITY OF WATERLOO | FACULTY OF MATHEMATICS DAVID R. CHERITON SCHOOL OF COMPUTER SCIENCE

October 7, 2022

Problem

Unconstrained minimization:

$$f_\star = \inf_{\mathbf{w}\in\mathbb{R}^d} \; f(w_1, \ldots, w_d)$$

- f: smooth w.r.t. a general norm $\|\cdot\|$ and possibly nonconvex
- For simplicity, no constraints on $\mathbf{w}$

Gradient Compression

Typical problem in ML:

$$f(\mathbf{w}) = \frac{1}{m}\sum_{i=1}^{m} f_i(\mathbf{w}; \mathcal{D}_i)$$

- Each f_i represent a different user/study/processor

$$f'(\mathbf{w}) = \frac{1}{m}\sum_{i=1}^{m} f_i'(\mathbf{w}; \mathcal{D}_i)$$

- For large d, communicating and aggregating the individual gradients are expensive
- Compress the gradients by simply taking its sign?

J. Bernstein et al. (2018). "signSGD: Compressed Optimisation for Non-Convex Problems". In: *Proceedings of the 35th International Conference on Machine Learning*, pp. 560–569; J. Bernstein et al. (2019). "signSGD with Majority Vote is Communication Efficient and Fault Tolerant". In: *International Conference on Learning Representations*.

Definition: Norm

Recall a norm $\|\cdot\|$ satisfies:

- definiteness: $\|\mathbf{x}\| \geq 0$ with 0 attained iff $\mathbf{x} = \mathbf{0}$
- positive homogeneity: $\|\lambda \mathbf{x}\| = |\lambda| \cdot \|\mathbf{x}\|$ for any $\lambda \in \mathbb{R}$
- triangle inequality: $\|\mathbf{x} + \mathbf{z}\| \leq \|\mathbf{x}\| + \|\mathbf{z}\|$

Definition: Dual

The dual norm of a norm $\|\cdot\|$ is

$$\|\mathbf{w}^*\|_\circ := \max_{\|\mathbf{w}\| \leq 1} \langle \mathbf{w}; \mathbf{w}^* \rangle$$

Example:

The dual of the ℓ_p norm $\|\mathbf{w}\|_p := (\sum_j |w_j|^p)^{1/p}$ is ℓ_q norm, where $1/p + 1/q = 1$.

Definition: duality mapping

Let $\mathsf{q} := \frac{1}{2}\|\cdot\|^2$ be "quadratic." We define the duality mapping

$$\mathsf{J} = \partial \mathsf{q} : \mathsf{V} \to \mathsf{V}^*, \quad \mathsf{j} : \mathsf{V} \to \mathsf{V}^*, \mathbf{w} \mapsto \mathsf{j}(\mathbf{w}) \in \mathsf{J}(\mathbf{w}),$$

where j is an arbitrary single-valued selection of J.

$$\langle \mathbf{w}; \mathsf{j}(\mathbf{w})\rangle = \|\mathbf{w}\|^2 = \|\mathsf{j}(\mathbf{w})\|_\circ^2$$

Definition: metric gradient w.r.t. a norm

We define the metric gradient w.r.t. a norm $\|\cdot\|$ as

$$\blacktriangledown f = \mathsf{J}^{-1}(f'), \quad \overline{\nabla} f = \mathsf{j}^{-1}(f') : \mathsf{V} \to \mathsf{V}.$$

Steepest Descent

Another way to recognize the metric gradient is through Kantorovich's steepest descent. Fixing the current iterate $\mathbf{w}_t$, we look for a direction $\mathbf{d}$ such that the univariate function

$$\eta \mapsto h(\eta) := f(\mathbf{w}_t - \eta \mathbf{d})$$

decreases steepest.

Kantorovich (1945) proposed to find the direction $\mathbf{d}$ through the subproblem:

$$\operatorname*{argmin}_{\mathbf{d} \neq \mathbf{0}} \quad \frac{h'(\eta)|_{\eta=0}}{\|\mathbf{d}\|} = \frac{-\langle \mathbf{d}; f'(\mathbf{w}_t) \rangle}{\|\mathbf{d}\|} \quad \Longrightarrow \quad \mathbf{d} = \frac{\overline{\nabla} f(\mathbf{w}_t)}{\| \overline{\nabla} f(\mathbf{w}_t)\|} = \frac{\overline{\nabla} f(\mathbf{w}_t)}{\|\nabla f(\mathbf{w}_t)\|_\circ},$$

which is exactly the normalized metric gradient!

L. V. Kantorovich (1945). "On an effective method of solving extremal problems for quadratic functionals". *Soviet Mathematics Doklady*, vol. 48, no. 7, pp. 595–600.

Algorithm 1: Metric gradient descent for unconstrained smooth minimization

Input: $\mathbf{w}_0$, norm $\|\cdot\|$

```
for t = 0, 1, ... do
    g_t ← ∇̄f(w_t)                         // compute any metric gradient
    if ‖g_t‖ = 0 then
        break
    choose step size η_t > 0
    w_{t+1} ← w_t − η_t g_t               // update
```

Key insight (note the similarity as before):

$$f(\mathbf{w}) \leq f(\mathbf{w}_t) + \langle \mathbf{w} - \mathbf{w}_t; f'(\mathbf{w}_t)\rangle + \frac{1}{2\eta_t}\|\mathbf{w} - \mathbf{w}_t\|^2,$$

i.e. L-smoothness w.r.t. a general norm $\|\cdot\|$.

Apply polar decomposition on the RHS:

$$\min_{\lambda \geq 0} \min_{\|\mathbf{w}-\mathbf{w}_t\|=\lambda} f(\mathbf{w}_t) + \langle \mathbf{w} - \mathbf{w}_t; f'(\mathbf{w}_t) \rangle + \tfrac{1}{2\eta_t}\lambda^2 \quad \equiv \quad \min_{\lambda \geq 0} -\lambda \|f'(\mathbf{w}_t)\|_\circ + \tfrac{1}{2\eta_t}\lambda^2.$$

Thus, $\lambda = \eta_t \|f'(\mathbf{w}_t)\|_\circ$ and

$$\mathbf{w} - \mathbf{w}_t = \lambda \frac{-\nabla f(\mathbf{w}_t)}{\|f'(\mathbf{w}_t)\|_\circ}, \qquad i.e. \qquad \mathbf{w}_{t+1} = \mathbf{w}_t - \eta_t \nabla f(\mathbf{w}_t)$$

Theorem: convergence of metric gradient descent for L-smooth functions

Let $f : \mathbb{R}^d \to \mathbb{R}$ be L-smooth w.r.t. a general norm $\|\cdot\|$ and bounded from below (i.e. $f_\star > -\infty$). If the step size $\eta_t \in [\alpha, \frac{2}{\mathsf{L}} - \beta]$ for some $\alpha, \beta > 0$, then the sequence $\{\mathbf{w}_t\}$ generated satisfies $\overline{\nabla} f(\mathbf{w}_t) \to \mathbf{0}$. Moreover,

$$\min_{0 \le t \le T-1} \|\overline{\nabla} f(\mathbf{w}_t)\| \le \sqrt{\frac{f(\mathbf{w}_0) - f_\star}{\alpha\beta \mathsf{L} T/2}}.$$

- The proof is literally the same as that of gradient descent
- Choosing $\alpha = \beta = \frac{1}{\mathsf{L}}$, the bound reduces to

$$\min_{0 \le t \le T-1} \|\overline{\nabla} f(\mathbf{w}_t)\| \le \sqrt{\frac{2\mathsf{L}[f(\mathbf{w}_0) - f_\star]}{T}}.$$

- Obviously, LHS depends on the norm and so does RHS (through $\mathsf{L} = \mathsf{L}_{\|\cdot\|}$)

ℓ_p norm metric gradient

Let $\mathsf{V} = \mathbb{R}^d$ be equipped with the ℓ_p norm, whose dual is ℓ_q norm with $1/p + 1/q = 1$.

$$\blacktriangledown f(\mathbf{w}) := \left[\mathop{\mathrm{argmax}}_{\|\mathbf{z}\|_p \le \|f'(\mathbf{w})\|_q} \langle \mathbf{z}; f'(\mathbf{w}) \rangle \right] = \|f'(\mathbf{w})\|_q^{1-q/p} \cdot \mathrm{sign}(f'(\mathbf{w})) \cdot |f'(\mathbf{w})|^{q/p}$$

- When $p = q = 2$, we have $\blacktriangledown f = \nabla f$
- When $p = 1, q = \infty$, we have $\blacktriangledown f = \mathrm{conv}\{\nabla_j f \cdot \mathbf{e}_j : |\nabla_j f| = \|\nabla f\|_\infty\}$
- When $p = \infty, q = 1$, we have $\blacktriangledown f = \mathrm{conv}\{\|\nabla f\|_1 \cdot \mathrm{sign}(\nabla f)\}$, $\mathrm{sign}(0) \in [-1, 1]$

metric gradient indeed depends on the norm

Sign gradient descent

Let us equip the input space V (where $\mathbf{w}$ lives) with the ℓ_∞ norm , and the gradient space V^* (where $f'(\mathbf{w})$ lives) with the corresponding dual ℓ_1 norm.

We obtain the so-called sign gradient descent algorithm, where in each iteration we only update with the sign of the gradient:

$$\mathbf{w}_{t+1} = \mathbf{w}_t - \eta_t \|\nabla f(\mathbf{w}_t)\|_1 \cdot \mathrm{sign}(\nabla f(\mathbf{w}_t)),$$

which is particularly appealing in distributed and low-resource devices.

Coordinate gradient descent

Let us equip the input space V (where $\mathbf{w}$ lives) with the ℓ_1 norm , and the gradient space V* (where $f'(\mathbf{w})$ lives) with the corresponding dual ℓ_∞ norm.

We obtain the so-called greedy coordinate gradient descent algorithm, where in each iteration we only take a gradient step along one (block of) coordinate(s):

$$w_{j,t+1} = w_{j,t} - \eta_t \nabla_j f(\mathbf{w}_t), \quad \text{where} \quad |\nabla_j f(\mathbf{w}_t)| = \|\nabla f(\mathbf{w}_t)\|_\infty.$$

- Compute all derivatives to figure out which one is largest
- Most of the computational effort is wasted...

R. V. Southwell (1935). "Stress-Calculation in Frameworks by the Method of "Systematic Relaxation of Constraints". I and II". *Proceedings of the Royal Society of London. Series A, Mathematical and Physical Sciences*, vol. 151, no. 872, pp. 56–95.

Alternatives

An obvious alternative is to update the coordinates cyclically:

$$\textbf{for}\ \ j = 1, \ldots, d$$
$$w_j \leftarrow w_j - \eta \nabla_j f(\mathbf{w})$$

- computing the gradient ∇f vs. computing a single component $\nabla_j f$?
- L-smoothness is w.r.t. different norms!
- Can randomize our choice of the coordinates (Nesterov 2012)
- Might as well go to the extreme:

$$w_j \leftarrow \underset{w}{\operatorname{argmin}}\ f(w_1, \ldots, w_{j-1}, w, w_{j+1}, \ldots, w_d)$$

Y. Nesterov (2012). "Efficiency of Coordinate Descent Methods on Huge-Scale Optimization Problems". *SIAM Journal on Optimization*, vol. 22, no. 2, pp. 341–362.

Definition: metric projection

We define the metric projection w.r.t. an arbitrary norm and a closed set C:

$$\mathrm{P}_C(\mathbf{w}) = \operatorname*{argmin}_{\mathbf{z} \in C} \|\mathbf{w} - \mathbf{z}\|.$$

However, the metric projection may no longer be nonexpansive even when C is convex.

Algorithm 2: Metric projected gradient descent

Input: $\mathbf{w}_0 \in C$, norm $\|\cdot\|$

```
for t = 0, 1, ... do
    g_t ← ∇̄f(w_t)                              // compute any metric gradient
    η_t ← argmin_{η≥0} f(P_C(w_t − η g_t))     // Cauchy's rule
    w_{t+1} ← P_C(w_t − η_t g_t)               // update
```

G. P. McCormick (1969). "Anti-Zig-Zagging by Bending". *Management Science*, vol. 15, no. 5, pp. 315–320.