A nonconvex loss function with identical critical values

Mastane Achab

Deep Gambit Limited, Masdar City, Abu Dhabi, UAE

UAE MathDay, University of Sharjah, 12 April 2025

Gradient-based optimization

- Task: minimize loss function $f : \mathbb{R}^d \to \mathbb{R}$
- Continuous gradient flow ODE: $\frac{dx}{dt}(t) = -\nabla f(x(t))$
- Discretization step-size $\lambda > 0$ a.k.a. learning rate
- Gradient descent (explicit Euler): $x_{t+1} = x_t - \lambda \nabla f(x_t)$
- Proximal point algorithm a.k.a. PPA (implicit Euler): $x_{t+1} = x_t - \lambda \nabla f(x_{t+1})$
- Convergence to a minimizer of f under various assumptions (smoothness, convexity, Polyak-Lojasiewicz condition)

Proximal operator

- assume f is convex $\Rightarrow$ its gradient is monotone: $\forall x, x'$,

$$\langle \nabla f(x) - \nabla f(x'), x - x' \rangle \geq 0$$

- PPA update $x_{t+1} = x_t - \lambda \nabla f(x_{t+1})$
- $\iff x_{t+1} = \mathrm{prox}_{\lambda f}(x_t) = \mathrm{argmin}_{z \in \mathbb{R}^d} f(z) + \frac{1}{2\lambda}\|z - x_t\|^2$
- $\mathrm{Fix}(\mathrm{prox}_{\lambda f}) = \mathrm{Crit}(f) = \{x : \nabla f(x) = 0\}$
- The proximal operator is firmly nonexpansive: $\forall x, x'$,

$$\langle \mathrm{prox}_{\lambda f}(x) - \mathrm{prox}_{\lambda f}(x'), x - x' \rangle \geq \|\mathrm{prox}_{\lambda f}(x) - \mathrm{prox}_{\lambda f}(x')\|^2$$

- Hence $\mathrm{prox}_{\lambda f} \circ \cdots \circ \mathrm{prox}_{\lambda f}(x_0) \to x^* = \mathrm{prox}_{\lambda f}(x^*)$ if $\mathrm{Fix}(\mathrm{prox}_{\lambda f}) \neq \emptyset$

Mirror descent (MD)

- probability simplex
  $\Delta_K = \{p = (p_1, \ldots, p_K) \in (0,1)^K : p_1 + \cdots + p_K = 1\}$
- Task: minimize a function $f : \Delta_K \to \mathbb{R}$
- mirror map ∇h with negative entropy $h(p) = \sum_k p_k \log(p_k)$
- MD update: $\nabla h(p^{t+1}) = \nabla h(p^t) - \lambda \nabla f(p^t)$
- $\iff p^{t+1} = \mathrm{argmin}_{q \in \Delta_K} \nabla f(p^t)^\intercal (q - p^t) + \frac{1}{\lambda} KL(q \| p^t)$
- Note proximity term here is KL, not Euclidean

Baryconvex optimization (A., 2024)

- assume $f = (f_1, \dots, f_K)$ with every f_k convex
- generalized proximal operator:

$$(x^{t+1}, p^{t+1}) = \mathrm{prox}_{\lambda f}(x^t, p^t) =$$

$$\arg\operatorname{minimax}_{(z,q)\in\mathbb{R}^d\times\Delta_K} q^\intercal f(z) + \frac{1}{2\lambda}\|z - x^t\|^2 - \frac{1}{\lambda} KL(q\|p^t)$$

- joint update of the pair (x^t, p^t)

Extended properties

- the operator $(x, p) \mapsto \begin{pmatrix} \mathrm{Jac}_f(x)^\intercal p \\ -f(x) \end{pmatrix}$ is monotone
- our generalized prox is Bregman firmly nonexpansive w.r.t. Euclidean-KL product geometry
- Bregman geometry associated to $(x, p) \mapsto \frac{1}{2}\|x\|^2 + h(p)$

Fixed points as critical points

- Denote $F(x, \xi) = \sigma(\xi)^\intercal f(x)$ (nonconvex) with σ the softmax function
- Then for $\mathrm{Crit}(F) = \{(x, \xi) : \nabla F(x, \xi) = 0\}$ we have:

$$\mathrm{Fix}(\mathrm{prox}_{\lambda f}) = \{(x, \sigma(\xi)) : (x, \xi) \in \mathrm{Crit}(F)\}$$

- At most one critical value i.e. $\mathrm{Crit}(F) \neq \emptyset \implies F(\mathrm{Crit}(F))$ is a singleton
- $\to$ generalizes the fact that a convex function has at most one critical value (global minimum)

Thank you!