CS794/CO673: Optimization for Data Science

Lec 23: Prox-Linear

Yaoliang Yu

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

December 2, 2022

Problem

Composite minimization:

$$\min_{\mathbf{w}\in\mathbb{R}^d} f(\mathbf{w}), \quad \text{where} \quad f(\mathbf{w}) = \varphi(\mathbf{s}(\mathbf{w}))$$

- $\mathrm{s} : \mathbb{R}^d \to \mathbb{R}^p$ is a sufficiently smooth vector-valued function
- $\varphi : \mathbb{R}^p \to \mathbb{R}$ is possibly nonsmooth

Gauss-Newton

- Given $\mathbf{w}_t$, we linearize the inner function $\mathbf{s}$ and proceed to minimize the outer function φ:

$$\mathbf{w}_{t+1} = \underset{\mathbf{w}}{\operatorname{argmin}} \ \varphi\big(\mathbf{s}(\mathbf{w}_t) + \mathbf{s}'(\mathbf{w}_t)(\mathbf{w} - \mathbf{w}_t)\big)$$

- It may happen that $f(\mathbf{w}_{t+1}) > f(\mathbf{w}_t)$, since our linearization only holds locally around $\mathbf{w}_t$ while there is no guarantee that $\mathbf{w}_{t+1}$ will remain close to $\mathbf{w}_t$

Example: Nonlinear least squares

Often we need to find a solution to some nonlinear equation, i.e. $\mathbf{s}(\mathbf{w}) = \mathbf{0}$. Operationally, it is preferred to solve the nonlinear least-squares reformulation:

$$\min_{\mathbf{w}} \ \tfrac{1}{2}\|\mathbf{s}(\mathbf{w})\|_2^2, \quad \text{where} \quad \varphi = \tfrac{1}{2}\|\cdot\|_2^2.$$

- Directly solving the above problem may be challenging
- Reduce it to a sequence of linear least squares problems:

$$\mathbf{w}_{t+1} = \operatorname*{argmin}_{\mathbf{w}} \ \tfrac{1}{2}\|\mathbf{s}(\mathbf{w}_t) + \mathbf{s}'(\mathbf{w}_t)(\mathbf{w} - \mathbf{w}_t)\|_2^2$$

- Typically worsens the condition number
- Taking square root we arrive at an equivalent reformulation:

$$\min_{\mathbf{w}} \ \|\mathbf{s}(\mathbf{w})\|_2, \quad \text{where} \quad \varphi = \|\cdot\|_2.$$

Prox-linear

$$\mathbf{w}_{t+1} = \operatorname*{argmin}_{\mathbf{w}} \underbrace{\varphi\big(\mathbf{s}(\mathbf{w}_t) + \mathbf{s}'(\mathbf{w}_t)(\mathbf{w} - \mathbf{w}_t)\big)}_{\tilde{f}_t(\mathbf{w}) = \tilde{f}(\mathbf{w};\mathbf{w}_t)} + \tfrac{1}{2\eta_t}\|\mathbf{w} - \mathbf{w}_t\|_2^2, \ \ i.e., \ \ \mathbf{w}_{t+1} = \mathrm{P}_{\tilde{f}_t}^{\eta_t}(\mathbf{w}_t)$$

- Prox-linear adds regularization to the Gauss-Newton algorithm
- Could also turn the implicit regularization into an explicit constraint, resulting in the so-called trust region methods
- When the outer function φ is convex, the regularized problem is strongly convex, while the original function $f = \varphi \circ \mathbf{s}$ may not even be convex
- Can show that the increment $\|\mathbf{w}_{t+1} - \mathbf{w}_t\|_2$ is (continuous) increasing w.r.t. η_t while $\|\mathbf{w}_{t+1} - \mathbf{w}_t\|_2/\eta_t$ is (continuous) decreasing w.r.t. η_t

Making Sense of Prox-linear

- For sufficiently small η_t, $\mathbf{w}_{t+1}$ will remain close to $\mathbf{w}_t$ so that decreasing the surrogate function $\tilde{f}$ leads to decrease in the original function f as well:

$$\frac{\mathrm{d} f(\mathbf{w}_{t+1})}{\mathrm{d}\eta_t} = f'(\mathbf{w}_{t+1})\frac{\mathrm{d}\mathbf{w}_{t+1}}{\mathrm{d}\eta_t} = f'(\mathbf{w}_{t+1})[-(\mathrm{Id} + \eta_t \tilde{f}_t''(\mathbf{w}_{t+1}))^{-1}\tilde{f}_t'(\mathbf{w}_{t+1})],$$

  where we differentiated the optimality condition of $\mathbf{w}_{t+1}$ w.r.t. η_t in the last step:

$$\eta_t \tilde{f}_t'(\mathbf{w}_{t+1}) + \mathbf{w}_{t+1} - \mathbf{w}_t = 0.$$

  Noting that $\mathbf{w}_{t+1} \to \mathbf{w}_t$ if $\eta_t \downarrow 0$, under mild continuity assumptions (e.g. φ and s are sufficiently smooth or convex), we have

$$\frac{\mathrm{d} f(\mathbf{w}_{t+1})}{\mathrm{d}\eta_t}\Big|_{\eta_t=0} = -\|f'(\mathbf{w}_t)\|_2^2 < 0$$

- If $\mathbf{w}_{t+1} = \mathbf{w}_t = \mathbf{w}$, then clearly $\mathbf{w}$ is a stationary point of $\tilde{f}_t$ and hence of f

The Generality of Composition

- Let $\tilde{\mathrm{s}}(\mathbf{w}) = (\mathbf{s}(\mathbf{w}), \mathbf{w})$ and $\tilde{\varphi}(\mathbf{z}, \mathbf{w}) = \varphi(\mathbf{z}) + r(\mathbf{w})$. Show that

  $$\tilde{\varphi}(\tilde{\mathbf{s}}(\mathbf{w})) = \varphi(\mathbf{s}(\mathbf{w})) + r(\mathbf{w}),$$

  and the Gauss-Newton update for the left-hand side reduces to:

  $$\mathbf{w}_{t+1} = \operatorname*{argmin}_{\mathbf{w}} \ \varphi\big(\mathbf{s}(\mathbf{w}_t) + \mathbf{s}'(\mathbf{w}_t)(\mathbf{w} - \mathbf{w}_t)\big) + r(\mathbf{w})$$

- Find s and φ so that the Gauss-Newton update for $\varphi \circ \mathrm{s}$ reduces to the generalized conditional gradient update for $\ell + r$.
- Find s and φ so that the prox-linear update for $\varphi \circ \mathrm{s}$ reduces to the gradient update for $\ell + r$.
- Find s and φ so that the prox-linear update for $\varphi \circ \mathrm{s}$ reduces to the proximal gradient update for $\ell + r$, with a forward step for ℓ and a backward step for r.

Properties of Prox-linear

$$\mathbf{w}_{t+1} = \underset{\mathbf{w}}{\operatorname{argmin}} \; \underbrace{\varphi\big(\mathbf{s}(\mathbf{w}_t) + \mathbf{s}'(\mathbf{w}_t)(\mathbf{w} - \mathbf{w}_t)\big)}_{\tilde{f}_t(\mathbf{w}) = \tilde{f}(\mathbf{w};\mathbf{w}_t)} + \tfrac{1}{2\eta_t}\|\mathbf{w} - \mathbf{w}_t\|_2^2, \;\; i.e., \;\; \mathbf{w}_{t+1} = \mathrm{P}_{\tilde{f}_t}^{\eta_t}(\mathbf{w}_t)$$

- Gauss-Newton is affine equivariant while prox-linear is not
- Prox-linear is an interpolation between Gauss-Newton and gradient descent
  - $\eta \to \infty$: reduces to Gauss-Newton
  - $\eta \to 0$: reduces to gradient descent (upon normalization)
- Convergence can be proved as before (Nesterov, Drusvyatskiy and Lewis)
- Quadratic variants (Bolte et al. 2020):

$$\mathbf{w}_{t+1} = \underset{\mathbf{w}}{\operatorname{argmin}} \, \varphi\Big(\mathbf{s}(\mathbf{w}_t) + \mathbf{s}'(\mathbf{w}_t)(\mathbf{w} - \mathbf{w}_t) + \tfrac{1}{2\eta_t}\|\mathbf{w} - \mathbf{w}_t\|^2\Big),$$

Y. Nesterov. "Modified Gauss-Newton scheme with worst case guarantees for global performance". *Optimization Methods and Software*, vol. 22, no. 3 (2007), pp. 469–483, D. Drusvyatskiy and A. S. Lewis. "Error Bounds, Quadratic Growth, and Linear Convergence of Proximal Methods". *Mathematics of Operations Research*, vol. 43, no. 3 (2018), pp. 919–948, J. Bolte et al. "The multiproximal linearization method for convex composite problems". *Mathematical Programming*, vol. 182 (2020), pp. 1–36.

Stochas Updates and Variance Reduction

$$\min_{\mathbf{w}} \; \mathbb{E}_{\boldsymbol{\xi}}[\varphi(\mathbf{s}(\mathbf{w}, \boldsymbol{\xi}))]$$

- When linearize s, can use the same stochastic idea as in SGD
- If the expectation is over a finite dataset, can apply variance reduction

D. Drusvyatskiy and C. Paquette. "Efficiency of minimizing compositions of convex functions and smooth maps". *Mathematical Programming*, vol. 178 (2019), pp. 503–558, J. Zhang and L. Xiao. "Stochastic variance-reduced prox-linear algorithms for nonconvex composite optimization". *Mathematical Programming* (2021).

Quasi-Newton Method

$$\mathbf{w}_{t+1} = \mathbf{w}_t - \eta_t \cdot H_t \cdot f'(\mathbf{w}_t)$$

- H_t is some approximation of the inverse Hessian: $H_t \approx [f''(\mathbf{w}_t)]^{-1}$
- Can approximate Hessian using $O(d)$ evals of gradient:

$$\frac{f'(\mathbf{w}_t + \alpha \mathbf{e}_j) - f'(\mathbf{w}_t)}{\alpha}, \; j = 1, \ldots, d$$

- Let $\mathbf{h}_t = f'(\mathbf{w}_{t+1}) - f'(\mathbf{w}_t)$ and $\mathbf{p}_t = \eta_t H_t f'(\mathbf{w}_t) = \mathbf{w}_{t+1} - \mathbf{w}_t$
- Use previous gradients to directly approximate Hessian inverse:

$$H_{t+1} = \underset{H}{\operatorname{argmin}} \; \|H - H_t\| \quad \text{s.t.} \quad H\mathbf{h}_t = \mathbf{p}_t$$

- Davidon-Fletcher-Powell:

$$H_{t+1} = H_t - \frac{H_t \mathbf{h}_t \mathbf{h}_t^\top H_t}{\mathbf{h}_t^\top H_t \mathbf{h}_t} + \frac{\mathbf{p}_t \mathbf{p}_t^\top}{\mathbf{p}_t^\top \mathbf{h}_t}$$

- Broyden-Fletcher-Goldfarb-Shanno (BFGS):

$$H_{t+1} = (I - \frac{\mathbf{p}_t \mathbf{h}_t^\top}{\mathbf{h}_t^\top \mathbf{p}_t}) H_t (I - \frac{\mathbf{h}_t \mathbf{p}_t^\top}{\mathbf{h}_t^\top \mathbf{p}_t}) + \frac{\mathbf{p}_t \mathbf{p}_t^\top}{\mathbf{p}_t^\top \mathbf{h}_t}$$