Non-smooth hyper-parameter learning

Clarice Poon

March 31, 2022

Consider

$$\min_y h(y), \quad \text{where} \quad h(y) = f(y, x(y)),$$
$$\text{s.t.} \quad x(y) \in \operatorname{argmin}_x g(y, x).$$

In general, we have

$$\partial_x g(y, x(y)) = 0,$$

and by the implicit function theorem, provided that $\partial_x^2 g(y, x(y))$ is invertible, $y \mapsto x(y)$ is differentiable with

$$x'(y) = -\partial_{xx} g(y, x(y))^{-1} \partial_{yx} g(y, x(y)).$$

One can then evaluate the gradient of h as

$$\nabla h(y) = \partial_y f(y, x(y)) - \partial_x f(y, x(y))^\top \partial_{xx} g(y, x(y))^{-1} \partial_{yx} g(y, x(y)). \quad (1)$$

1 Quasi-Newton?

One approach is to find p_k such taht

$$\partial_{yx} g(y, x(y)) \approx \partial_{yy} g(y, x(y)) p_k$$

and evaluate $\nabla h(y) \approx \partial_y f(y, x(y)) - \partial_x f(y, x(y))^\top p_k$.

Can we do a quasi-Newton approach? We know that

$$\partial_x g(y, x(y)) = 0$$

Let $x_k \triangleq x(y_k)$. Then

$$\begin{aligned}\partial_x g(y_k, x(y_{k-1})) &= \partial_x g(y_k, x_{k-1}) - \partial_x g(y_k, x_k) \\ &= \partial_{xx} g(y_k, x_k)(x_{k-1} - x_k) + o(\|x_{k-1} - x_k\|)\end{aligned}$$

Suppose we find B_k such that it minimises

$$\min_B \|B \partial_x g(y_k, x(y_{k-1})) - (x_{k-1} - x_k)\|$$

and treat it as an approximation to $\partial_{xx} g(y_k, x(y_k))^{-1}$. The idea is to compute

$$\nabla h(y_k) \approx \partial_y f(y_k, x(y_k)) - \partial_x f(y_k, x(y_k))^\top B_k \partial_{yx} g(y_k, x(y_k)).$$

One possible update of B_k is as $\tau_k \mathrm{Id} + u_k u_k^\top$. Define $s_k = x_{k-1} - x_k$ and $z_k = \partial_x g(y_k, x(y_{k-1}))$. We want to find diagonal + rank-1 matrix B to minimise

$$\min_B \|B z_k - s_k\|$$

i) Define $\tau_k = \langle s_k,\, z_k\rangle / \|z_k\|^2$ and project onto $[\tau_{\min}, \tau_{\max}]$. Note that before projection, $\tau_k = \mathrm{argmin}_\tau \|\tau z_k - s_k\|$.

ii) Let $B_0 \triangleq \gamma \tau_k \mathrm{Id}$ where $\gamma \in (0, 1)$.

iii) If $\langle s_k - B_0 z_k,\, z_k\rangle \leqslant 10^{-8} \|z_k\| \|s - B_0 z_k\|$ the $U_k = 0$. Else:

$$U_k = \frac{(s_k - B_0 z_k)(s_k - B_0 z_k)^\top}{\langle s_k - B_0 z_k,\, z_k\rangle}.$$

iv) Let $B_k = B_0 + U_k$.

Note that for step iii) the choice of U_k is precisely finding $U_k = uu^\top$ such that

$$B_0 z_k + u\langle u,\, z_k\rangle - s_k = 0.$$

[ToDo:

1. **If we repeatedly updated B_k with rank-1 matrices, show that B_k converges to $\partial_{xx} g(y_*, x(y_*))^{-1}$.**

2. **Suppose that $f = g$ and at iteration k, we have an approximate solution $\hat{x}_k \approx x(y_k)$. Possible ways of computing $\nabla h(y_k)$ include**

   i) $p_1 = \partial_y f(y_k, \hat{x}_k))$

   ii) $p_2 = \partial_y f(y_k, \hat{x}_k)) - \partial_x f(y_k, \hat{x}_k)^\top B_k \partial_{yx} f(y_k, \hat{x}_k)$

   iii) $p_3 = \partial_y f(y_k, \hat{x}_k) + \partial_x f(y_k, \hat{x}_k)^\top \partial_y \hat{x}_k$ **where we obtain** $\partial_y \hat{x}_k$ **via autodiff.**

   The first and 3rd option have been analysed recently (Ablin et al). For the second approach, can we bound the difference between taking approximation B_k and the true Hessian?

]

1.1 When the outer and inner problems are the same

Consider

$$\min_y h(y), \quad \text{where} \quad h(y) = \min_x f(x,y).$$

By differentiating $\partial_x f(x(y),y) = 0$, we obtain for $x = x(y)$,

$$\partial_x^2 f(x,y) x'(y) = -\partial_x \partial_y f(x,y)$$

and

$$\nabla h(x) = \partial_y f(x,y) - \partial_{xy} f(x,y)^\top \partial_x^2 f(x,y)^{-1} \partial_x f(x,y)$$

If $x(y) = \operatorname{argmin}_x f(x,y)$ is computed exactly, then $\nabla h(y) = \partial_y f(x,y)$. The question is what happens when $x(y)$ is only approximated by $\hat{x}$. In this case, one can use the approximation

$$\hat{p} = \partial_y f(\hat{x},y) - \partial_{xy} f(\hat{x},y)^\top \partial_x^2 f(\hat{x},y)^{-1} \partial_x f(\hat{x},y)$$

How effective is the quasi-Newton approximation to $\partial_x^2 f(\hat{x},y)^{-1}$?

Examples:

(i) Lasso $f(x,y) \triangleq \frac{1}{2}\|x\|^2 + \frac{1}{2}\|y\|^2 + \frac{1}{2\lambda} L(xy)$

$$\partial_y f = y + \frac{1}{\lambda} x \odot \nabla L(xy), \quad \partial_x f = x + \frac{1}{\lambda} y \odot \nabla L(xy)$$

$$\partial_{xy} f = \frac{1}{\lambda} \big(\operatorname{diag}(\nabla L(xy)) + \operatorname{diag}(y) \nabla^2 L(xy) \operatorname{diag}(x)\big)$$

and

$$\partial_{xx} f = \mathrm{Id} + \frac{1}{\lambda} \operatorname{diag}(y) \nabla^2 L(xy) \operatorname{diag}(y)$$

2 A differentiable approach to nonsmooth bilevel programming

One example is where g is nonsmooth is when y correspond to a hyperparamter λ and x is the lasso regression coefficients:

$$f(\lambda,\beta) \triangleq \|A_{\text{test}}\beta - y\|^2 \quad \text{and} \quad g(\lambda,\beta) \triangleq \frac{1}{2}\|A_{\text{train}}\beta - b\|_2^2 + \lambda\|\beta\|_1.$$

The difficulty is in this case is that since g is non-smooth, the formula (1) cannot be used. One alternative is to consider instead

$$f(\lambda,(u,v)) \triangleq \|A_{\text{test}} uv - y\|^2 \quad \text{and} \quad g(\lambda,(u,v)) \triangleq \|A_{\text{train}} uv - b\|_2^2 + \lambda\|u\|^2/2 + \lambda\|v\|^2/2.$$

The advantage with this approach is that g is a smooth function and one can show that the Hessian of g is invertible when $\beta \triangleq u(\lambda) \odot v(\lambda)$ is a nondegenerate solution, that is,

$$\max_{i \notin \mathrm{Supp}(\beta)} |A_{\text{train}}^\top (A_{\text{train}}\beta - b)|_i < 1.$$

Things to do

- Check properties of the Hessian of g.
- Acceleration using support pruning.
- Can we handle regularisers such ash $\|L\beta\|_1$ where L is a (possibly singular) linear operator?

For

$$\min_\lambda f(\lambda, \beta)$$

where $\beta \in \operatorname{argmin}_\beta \lambda\|L\beta\|_1 + \frac{1}{2}\|A\beta - y\|^2$, consider instead

$$\min_\lambda h(\lambda) \triangleq f(\lambda, \beta(v(\lambda), \lambda))$$

where

$$v(\lambda) \in \operatorname{argmin}_v \psi(v, \lambda) \triangleq \min_\beta \max_\alpha -\frac{1}{2\lambda}\|\alpha v\|^2 + \frac{\lambda}{2}\|v\|^2 + \frac{1}{2}\|A\beta - y\|^2 + \langle L\beta,\, \alpha\rangle$$

and

$$\beta(v, \lambda) \in \operatorname{argmin}_\beta \max_\alpha \frac{1}{2\lambda}\|\alpha v\|^2 + \frac{\lambda}{2}\|v\|^2 + \frac{1}{2}\|A\beta - y\|^2 + \langle L\beta,\, \alpha\rangle$$

We certainly have that ψ is differentiable and can compute $\nabla\psi$. To compute $\nabla h(\lambda)$, we need

$$\partial_\lambda f + \partial_\beta f[\partial_v \beta \partial_\lambda v + \partial_\lambda \beta]$$

3 The square root lasso

The square root lasso is

$$\min_{\beta \in \mathbb{R}^n} \|X\beta - y\|_2 + \lambda\|\beta\|_1.$$

One interesting aspect of this is that when $y = X\beta_0 + w$, the minimiser β satisfies

$$\|\beta - \beta_0\| \lesssim \|w\|$$

for some constant λ. This is remarkable since, for the Lasso, to achieve this kind of error bound, one would require that $\lambda \sim \|w\|$ and some knowledge of the noise level is required.

One remark is that the square root lasso is equivalent to

$$\min_{\sigma > 0} \min_\beta \frac{1}{2\sigma}\|X\beta - y\|_2^2 + \frac{\sigma}{2} + \lambda\|\beta\|_1,$$

and we can therefore write this in the bilevel formulation with

$$f(\sigma, \beta) = \frac{1}{2\sigma}\|X\beta - y\|_2^2 + \frac{\sigma}{2} + \lambda\|\beta\|_1$$

and

$$g(\sigma, \beta) = \frac{1}{2\sigma}\|X\beta - y\|_2^2 + \lambda\|\beta\|_1.$$

One question I have is what happens when we consider

$$\begin{aligned} f(\sigma, \beta) &= \frac{1}{2\sigma}\|A_{\text{test}}\beta - y\|_2^2 + \frac{\varepsilon\sigma}{2} + \iota_{\sigma>0} \\ g(\sigma, \beta) &= \frac{1}{2\sigma}\|A_{\text{train}}\beta - y\|_2^2 + \|\beta\|_1 \end{aligned}$$

This is precisely the hyperparameter learning framework but with added regularisation on the parameter σ. Note that the outer problem can be written as an unconstrained smooth problem as follows: Let $z = \sqrt{\sigma}$ and $v = A_{\text{test}}\beta - y$, then

$$\min_{z\in\mathbb{R}} f(z, \beta(z^2)), \quad \text{where} \quad \begin{aligned} f(z, \beta) &\triangleq \min_{zv = A_{\text{test}}\beta - y} \frac{1}{2}\|v\|^2 + \frac{\varepsilon}{2}z^2, \\ \beta(z^2) &\triangleq \operatorname{argmin}_\beta g(z^2, \beta) \end{aligned}$$

Notice that the minimisation problem in f is convex wrt v, so by taking the convex dual,

$$f(z, \beta) = \max_{\alpha\in\mathbb{R}^m} -\frac{\varepsilon}{2}z^2\|\alpha\|^2 + \frac{\varepsilon}{2}z^2 + \langle\alpha, -A_{\text{test}}\beta + y\rangle$$

The maximiser α is unique (as the problem is strongly concave) and

$$\partial_\beta f = -A_{\text{test}}^\top\alpha \quad \text{and} \quad \partial_z f = z\|\alpha\|^2.$$

Numerically, we can certainly handle this, the question is whether this kind of regularisation is interesting in practice.

Let $F(\sigma) = f(\sigma, \beta(\sigma))$. Let's look at the optimality conditions

$$\begin{aligned} \partial_\sigma f &= \frac{-1}{\sigma^2}\|A_{\text{test}}\beta - y_{\text{test}}\|^2 + \frac{\varepsilon}{2} \\ \partial_\beta f &= \frac{1}{\sigma}A_{\text{test}}^\top(A_{\text{test}}\beta - y_{\text{test}}) \end{aligned}$$

Also, $\beta = \beta(\sigma)$ satisfies

$$A_{\text{train}}^\top A_{\text{train}}\beta = A_{\text{train}}^\top y - \sigma\,\text{sign}(\beta)$$

In general, $\sigma \mapsto \beta(\sigma)$ is differentiable almost everywhere with gradient

$$\beta'(\sigma) = -(A_{\text{train}}^\top A_{\text{train}})_{J,J}^{-1}\,\text{sign}(\beta).$$

where $J = \text{Supp}(\beta)$. So, when $F'(\sigma) = 0$, we have

$$\frac{-1}{\sigma^2}\|A_{\text{test}}\beta - y_{\text{test}}\|^2 - \frac{1}{\sigma}\langle(A_{\text{test}}\beta - y_{\text{test}}),\, A_{\text{test}}(A_{\text{train}}^\top A_{\text{train}})_{J,J}^{-1}\,\text{sign}(\beta)\rangle + \frac{\varepsilon}{2} = 0$$

which implies $\lambda = 1/\sigma$ satisfies, for $C \triangleq \langle (A_{\text{test}}\beta - y_{\text{test}}),\ A_{\text{test}}(A_{\text{train}}^\top A_{\text{train}})^{-1}_{J,J} \operatorname{sign}(\beta)\rangle$,

$$\lambda = \frac{-C + \sqrt{C^2 + 2\varepsilon \|A_{\text{test}}\beta - y_{\text{test}}\|^2}}{2\|A_{\text{test}}\beta - y_{\text{test}}\|^2}$$

NB: For the standard problem where $f(\sigma, \beta) = \frac{1}{2}\|A_{\text{test}}\beta - y\|^2$, then

$$F'(\sigma) = -\langle (A_{\text{test}}\beta - y_{\text{test}}),\ A_{\text{test}}(A_{\text{train}}^\top A_{\text{train}})^{-1}_{J,J} \operatorname{sign}(\beta)\rangle.$$

Suppose $A_{\text{test}} = A_{\text{train}}$, then this says that

$$F'(\sigma) = \sigma \langle (A_{\text{train}}^\top A_{\text{train}})^{-1}_{J,J} \operatorname{sign}(\beta),\ \operatorname{sign}(\beta)\rangle > 0$$

which means that we optimise to $\sigma = 0$ as expected.