Optimization for Data Science

Lec 09: Alternating Minimization

Yaoliang Yu

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

Problem

Composite minimization:

$$f_\star = \inf_{\mathbf{w}\in\mathbb{R}^d} f(\mathbf{w}), \quad \text{where} \quad f(\mathbf{w}) = f_0(\mathbf{w}) + \sum_{j=1}^{d} f_j(w_j)$$

- $f_0 : \mathbb{R}^d \to \mathbb{R}$ smooth
- $f_j : \mathbb{R} \to \mathbb{R}$ can be nonsmooth, but they are separable
- More generally, each w_j can be a block of variables
- With $f_j(w_j) = \iota_{C_j}(w_j)$, we reduce to the constrained problem:

$$\min_{\mathbf{w}\in C_1\times C_2\cdots\times C_d} f_0(\mathbf{w})$$

Convex Function Estimation

Least-squares regression:

$$y = f(\mathbf{x}) + \epsilon, \qquad \min_f \; \hat{\mathbb{E}}[y - f(\mathbf{x})]^2$$

- Can assume f is linear: $f(\mathbf{x}) = \langle \mathbf{x}, \mathbf{w} \rangle$ and solve for $\mathbf{w}$
- Can assume f is convex and solve for f directly!

Example: Univariate convex function estimation, primal

Let $d = 1$ and assume w.l.o.g. that $x_1 > x_2 > \cdots > x_n$. Let $z_i = f(x_i)$.

$$\min_{\mathbf{z} \in \mathbb{R}^n} \; \tfrac{1}{2}\|\mathbf{z} - \mathbf{y}\|_2^2, \quad \text{s.t.} \quad \tfrac{z_i - z_{i+1}}{x_i - x_{i+1}} \geq \tfrac{z_{i+1} - z_{i+2}}{x_{i+1} - x_{i+2}}$$

Example: Univariate convex function estimation, dual

Lagragian with dual variable $\boldsymbol{\lambda} \geq \mathbf{0}$:

$$\min_{\mathbf{z}} \max_{\boldsymbol{\lambda} \geq \mathbf{0}} \ \tfrac{1}{2}\|\mathbf{z} - \mathbf{y}\|_2^2 + \langle B\mathbf{z}, \boldsymbol{\lambda} \rangle$$

Setting $\mathbf{z} = \mathbf{y} - B^\top \boldsymbol{\lambda}$ we obtain the dual problem:

$$\min_{\boldsymbol{\lambda} \geq \mathbf{0}} \ \tfrac{1}{2}\boldsymbol{\lambda}^\top A \boldsymbol{\lambda} + \boldsymbol{\lambda}^\top \mathbf{b}, \quad \text{where} \quad A = BB^\top, \mathbf{b} = -B\mathbf{y}$$

- How to solve the primal problem?
- How to solve the dual problem?

C. Hildreth. "Point Estimates of Ordinates of Concave Functions". *Journal of the American Statistical Association*, vol. 49, no. 267 (1954), pp. 598–619.

Algorithm 1: Alternating Minimization (AltMin)

Input: $\mathbf{w} \in \operatorname{dom} f$

```
for t = 1, 2, ... do
    choose coordinate j                                  // cyclic, randomized or greedy
    w_j ← argmin_z f(w_1, ..., w_{j-1}, z, w_{j+1}, ..., w_d)
      // argmin_z f_0(w_1, ..., w_{j-1}, z, w_{j+1}, ..., w_d) + f_j(z), univariate problem!
```

- Can replace each exact minimization with simply a (proximal) gradient (or descent) step
- Can replace Gauss-Seidel update with a Jacobi update for parallelism
- Appealing in practice due to simplicity, flexibility (could be derivative-free), convenience (could be step size free), lightweight (minimum storage) and surprising efficiency

A Nice Univariate Result

Theorem: constrained univariate convex minimization

For any univariate convex function f and convex interval $C = [a, b]$, we have

$$\mathrm{P}_C\left(\operatorname*{argmin}_{w \in \mathbb{R}} f(w)\right) \subseteq \operatorname*{argmin}_{w \in C} f(w),$$

where $\mathrm{P}_C(w) = \mathrm{P}_{[a,b]}(w) = (a \vee w) \wedge b$ is the closest point in C to w.

Why Separability?

$$f(\mathbf{w}) = f_0(\mathbf{w}) + \sum_j f_j(w_j)$$

- What happens if $f_0 \equiv 0$, i.e. f is separable?
- What happens if the domain of f is not separable?

$$\min_{w+z=0} \; w^2 + z^2$$

The Difficulty for Nonsmooth f_0

Example:

Consider the strongly convex function

$$\min_{w,z} \underbrace{w \vee z}_{f_0} + \epsilon[(w-2)^2 + (z-2)^2],$$

where $\epsilon > 0$ is arbitrary. Due to symmetry, it is clear that

$$w_\star = z_\star = 2 - \frac{1}{2\epsilon}.$$

However, if we start with $w_* = z_* = 2$, then alternating minimization immediately gets stuck!

The Difficulty for Nonconvex f

$$\inf_{x,y,z} \; -xy-yz-zx+(x-1)_+^2+(-x-1)_+^2+(y-1)_+^2+(-y-1)_+^2+(z-1)_+^2+(-z-1)_+^2,$$

- Continuously differentiable and convex in each coordinate
- Taking $x = y = z$ yields

$$-3x^2 + 3(x-1)_+^2 + 3(-x-1)_+^2 = \begin{cases} -6x+3, & \text{if } x \geq 1 \\ -3x^2, & \text{if } x \in [-1,1] \\ 6x+3, & \text{if } x \leq -1 \end{cases}.$$

- Stationary points exist at $xyz = 0, x+y+z = 0, x, y, z \in \{0, \pm 2\}$

M. J. D. Powell. "On search directions for minimization algorithms". *Mathematical Programming*, vol. 4 (1973), pp. 193–201.

- Fixing y and z we obtain:

$$\begin{cases} -x(y+z)+(x-1)^2, & \text{if } x \geq 1 \\ -x(y+z), & \text{if } x \in [-1,1], \\ -x(y+z)+(x+1)^2, & \text{if } x \leq -1 \end{cases} \quad \text{with } x_* = \operatorname{sign}(y+z)+\tfrac{1}{2}(y+z)$$

- Start with $(-1-\epsilon, 1+\frac{1}{2}\epsilon, -1-\frac{1}{4}\epsilon)$, in two passes we obtain

$$(-1-\epsilon,1+\tfrac{1}{2}\epsilon,-1-\tfrac{1}{4}\epsilon)\to(1+\tfrac{1}{8}\epsilon,1+\tfrac{1}{2}\epsilon,-1-\tfrac{1}{4}\epsilon)\to(1+\tfrac{1}{8}\epsilon,-1-\tfrac{1}{16}\epsilon,-1-\tfrac{1}{4}\epsilon)\to$$
$$\to(1+\tfrac{1}{8}\epsilon,-1-\tfrac{1}{16}\epsilon,1+\tfrac{1}{32}\epsilon)\to(-1-\tfrac{1}{64}\epsilon,-1-\tfrac{1}{16}\epsilon,1+\tfrac{1}{32}\epsilon)\to(-1-\tfrac{1}{64}\epsilon,1+\tfrac{1}{128}\epsilon,1+\tfrac{1}{32}\epsilon)\to$$
$$\to(-1-\tfrac{1}{64}\epsilon,1+\tfrac{1}{128}\epsilon,-1-\tfrac{1}{256}\epsilon),$$

i.e. reducing ϵ by a factor of 64.

- AltMin cycles around the 6 limit points:

$$(-1,1,-1)\to(1,1,-1)\to(1,-1,-1)\to(1,-1,1)\to(-1,-1,1)\to(-1,1,1)\to(-1,1,-1),$$

neither of which is optimal or stationary.

What Does AltMin Try to Find?

Algorithm 2: Alternating Minimization (AltMin)

Input: $\mathbf{w} \in \mathrm{dom}\, f$

for $t = 1, 2, \ldots$ **do**

  choose coordinate j `// cyclic, randomized or greedy`

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

  `//` $\operatorname*{argmin}_z f_0(w_1, \ldots, w_{j-1}, z, w_{j+1}, \ldots, w_d) + f_j(z)$`, univariate problem!`

- Call $\mathbf{w}$ a (Nash) equilibrium of f if

$$\forall j,\ \ w_j \in \operatorname*{argmin}_z f(w_1, \ldots, w_{j-1}, z, w_{j+1}, \ldots, w_d).$$

- AltMin, if converges at all, converges to a Nash equilibrium?
- A Nash equilibrium may not be a minimizer, or even a stationary point of f!

Theorem: Convergence of AltMin for two blocks

Let $d = 2$ and consider any function $f(\mathbf{x}, \mathbf{y})$ that is separately continuous in its product domain. Assume AltMin is well-defined. Then, any limit point (if any) of $\{\mathbf{w}_t\}$ is an equilibrium.

Example: Nash equilibrium $\neq$ minimizer

Consider the strongly convex function

$$\min_{w,z} \; w \vee z + \epsilon[(w-2)^2 + (z-2)^2],$$

where $\epsilon > 0$ is arbitrary. Due to symmetry, it is clear that

$$w_\star = z_\star = 2 - \tfrac{1}{2\epsilon}.$$

However, if we start with $w_* = z_* = 2$, then AltMin immediately gets stuck!

L. Grippof and M. Sciandrone. "On the convergence of the block nonlinear Gauss–Seidel method under convex constraints". *Operations Research Letters*, vol. 26, no. 3 (2000), pp. 127–136.

Theorem: Convergence of AltMin for any number of blocks

Let $f(\mathbf{w}) = f_0(\mathbf{w}) + \sum_j f_j(w_j)$ be convex and continuous on the sublevel set $[\![f \leq f(\mathbf{w}_0)]\!]$ which we assume to be compact. Assume f_0 is smooth and choose the cyclic rule. Then, any limit point of AltMin is an equilibrium.

Theorem: Convergence of AltMin under uniqueness

Let f be continuous on the sublevel set $[\![f \leq f(\mathbf{w}_0)]\!]$ which we assume to be compact. Assume $\mathrm{dom}\, f$ to be separable and choose the cyclic rule. If for all but one j and any $\mathbf{w}$, the function $z \mapsto f(w_1, \ldots, w_{j-1}, z, w_{j+1}, \ldots, w_d)$ is attained at a unique minimizer, then any limit point of AltMin is an equilibrium.

$$\left[\min_{\mathbf{z}} \min_{\mathbf{w}} f(\mathbf{w}) + \tfrac{1}{2\eta}\|\mathbf{z} - \mathbf{w}\|_2^2\right] = \min_{\mathbf{z}} \mathrm{M}_f^\eta(\mathbf{z})$$

P. Tseng. "Convergence of a Block Coordinate Descent Method for Nondifferentiable Minimization". *Journal of Optimization Theory and Applications*, vol. 109 (2001), pp. 475–494.

Example: The shooting algorithm for lasso

Recall the lasso problem for sparse estimation:

$$\min_{\mathbf{w}\in\mathbb{R}^d} \frac{1}{2n}\|X\mathbf{w}-\mathbf{y}\|_2^2 + \lambda\|\mathbf{w}\|_1,$$

Any limit point of AltMin is a *bona fide* minimizer!

To update the j-th coordinate, we need to solve the subproblem:

$$\min_{w} \frac{1}{2n} \|\mathbf{x}_{:j}(w - w_j) + \mathbf{r}\|_2^2 + \lambda|w|, \quad \text{where} \quad \mathbf{r} := X\mathbf{w}_t - \mathbf{y},$$

(Univariate) soft-shrinkage operator in closed-form.

After updating $w_j \leftarrow w_j^+$, we update $\mathbf{r} \leftarrow \mathbf{r} - \mathbf{x}_{:j}w_j + \mathbf{x}_{:j}w_j^+$.

Complexity on par with gradient algorithms: $O(nd)$ for a full sweep.

W. J. Fu. "Penalized Regressions: The Bridge versus the Lasso". *Journal of Computational and Graphical Statistics*, vol. 7, no. 3 (1998), pp. 397–416.

Example: Sparse precision matrix estimation

Let $S := \frac{1}{n}\sum_{i=1}^{n}(\mathbf{x}_i - \boldsymbol{\mu})(\mathbf{x}_i - \boldsymbol{\mu})^\top$ be the sample covariance matrix. Consider

$$\begin{aligned}
\hat{\Sigma}^{-1} &:= \underset{X \succ \mathbf{0}}{\operatorname{argmax}} \; \log\det X - \operatorname{tr}(SX) - \lambda \|X\|_1 \\
&= \max_{X \succ \mathbf{0}} \min_{\|U\|_\infty \le \lambda} \log\det X - \operatorname{tr}(SX) - \operatorname{tr}(UX) \\
&\equiv \min_{\|U\|_\infty \le \lambda} \; -\log\det(S+U), \quad \text{where} \quad X = (S+U)^{-1} \\
&\equiv \max_{\|W-S\|_\infty \le \lambda} \log\det W
\end{aligned}$$

O. Banerjee, L. E. Ghaoui, and A. d'Aspremont. "Model Selection Through Sparse Maximum Likelihood Estimation for Multivariate Gaussian or Binary Data". *Journal of Machine Learning Research*, vol. 9 (2008), pp. 485–516.

- Diagonal $W_{jj} = S_{jj} + \lambda$ due to monotonicity of $\log\det$
- Sweep j-th column (and row) while fixing everything else:

$$\mathbf{w}_j = \underset{\|\mathbf{w}-\mathbf{s}_j\|_\infty \le \lambda}{\operatorname{argmin}} \ \mathbf{w}^\top W_{\setminus j,\setminus j}^{-1} \mathbf{w}$$

- Dual problem is:

$$\min_{\mathbf{z}} \ \mathbf{z}^\top W_{\setminus j,\setminus j} \mathbf{z} - \mathbf{s}_j^\top \mathbf{z} + \lambda \|\mathbf{z}\|_1$$

- This is just the Lasso problem!
- When is $\mathbf{w}_j \equiv \mathbf{0}$, i.e. sparse column/row in the precision matrix?