Matrix Algebra and Optimization for Statistics and Machine Learning

Yiyuan She

Department of Statistics, Florida State University

- Proximal methods and linearization

Soft-thresholding

- Recall the soft-thresholding for solving the lasso

$$\begin{aligned}\Theta(y;\lambda) &:= 1_{|y|>\lambda}(y - \mathrm{sgn}(y)\lambda) \\ &= \arg\min_{\beta} \frac{1}{2}(y-\beta)^2 + \lambda|\beta|\end{aligned}$$

- Similarly we showed that singular-value soft thresholding $\Theta^\sigma(Y;\lambda)$ solves $\min_B \frac{1}{2}\|Y - B\|_F^2 + \lambda\|B\|_*$
- These proximity operators can effectively handle statistical learning problems of form $\min l(\beta) + P(\beta)$

Proximity operators

- Given a closed proper convex function P (i.e., its epigraph $\{(x,t) \in \mathbb{R}^n \times \mathbb{R} : P(x) \leq t\}$ is a nonempty closed convex set), define

$$\mathrm{prox}_P(y) = \arg\min_x \frac{1}{2}(y-x)^2 + P(x)$$

- Projection operators $\arg\min_{x \in P} \frac{1}{2}(y-x)^2$ are special cases, since we can introduce an indicator function $\iota_P(x) = 0$ if $x \in P$, and $+\infty$ otherwise

Examples

- $P = \frac{\lambda}{2}x^2$ leads to proportional (ridge) scaling

$$\mathrm{prox}_P(y) = \frac{y}{1+\lambda}$$

- $P = \iota_{Ax=b}$:

$$\mathrm{prox}_P(y) = A^+b + \mathbf{P}_A^\perp y = A^+b + (I - A^+A)y$$

- $P = \iota_{\{L \leq x \leq U\}}$ (e.g. $\iota_{\{x \geq 0\}}$): truncation
- $P = \iota_{\|x\|_2 \leq 1}$: $\mathrm{prox}_P(y) = \begin{cases} y^\circ := y/\|y\|_2, & \text{if } \|y\|_2 \geq 1 \\ y, & \text{o/w} \end{cases}$

- For convex P, prox is well-defined (due to s-convexity)
  - Q: Do we really need convexity to define prox_P?
- From $0 \in x - y + \partial P(x)$, we can write it in the resolvent form (which corresponds a unique solution):

$$\text{prox}_P = (I + \partial P)^{-1}$$

- Convex f: $x^\star \in \arg\min f(x) \Leftrightarrow x^\star = \text{prox}_{f/\rho}(x^\star)$ (fixed point), due to $x^\star = \arg\min_x \frac{\rho}{2}\|x - x^-\|_2^2 + f(x)|_{x^- = x^\star}$,
  - Proximal point algorithm: $x^{t+1} = \text{prox}_{f/\rho}(x^t)$ $(\rho > 0)$

Some properties

- $f(\begin{bmatrix} x \\ y \end{bmatrix}) = g(x) + h(y)$, $\mathrm{prox}_f(\begin{bmatrix} x \\ y \end{bmatrix}) = \begin{bmatrix} \mathrm{prox}_g(x) \\ \mathrm{prox}_h(y) \end{bmatrix}$
- $f(x) = g(ax + b)$, $\mathrm{prox}_f(x) = \frac{1}{a}\mathrm{prox}_{a^2 g}(ax + b) - b)$
- $f(x) = \lambda g(x/\lambda)$, $\mathrm{prox}_f(x) = \lambda \mathrm{prox}_{g/\lambda}(x/\lambda)$. (So from $(\lambda f)^*(\cdot) = \lambda f^*(\cdot/\lambda)$, $\mathrm{prox}_{(\lambda f)^*}(x) = \lambda \mathrm{prox}_{f^*/\lambda}(x/\lambda)$.)

Moreau decomposition

- In the convex setting,

$$\mathrm{prox}_f + \mathrm{prox}_{f^*} = Id$$

- With a scaling $\lambda > 0$, $\mathrm{prox}_{\lambda f}(x) + \lambda \mathrm{prox}_{f^*/\lambda}(x/\lambda) = x$
- This is because $u = \mathrm{prox}_f(x) \Leftrightarrow x - u \in \partial f(u) \overset{\text{conjugate}}{\Longleftrightarrow} u \in \partial f^*(x-u) \Leftrightarrow x - u = \mathrm{prox}_{f^*}(x)$
- Formally, $(Id - (Id + \partial f)^{-1})^{-1} - Id = (\partial f)^{-1} = \partial f^*$
- This generalizes the subspace decomposition $(\mathbf{P}_A, \mathbf{P}_A^{\perp})$

Examples

- $P(x) = \lambda\|x\|$, $P^*(y) = \iota_{\|y\|_* \le \lambda}$, and so
$$\operatorname{prox}_P(x) = x - \mathbf{P}_{\|x\|_* \le \lambda}(x),$$
where projection may facilitate the calculation of prox
- In general, $\operatorname{prox}_{\lambda S_C}(x) = x - \lambda \mathbf{P}_C(x/\lambda)$, where S_C is the support function of C (i.e., $S_C = \iota_C^*$).
- $P(x) = x_{[1]} + \cdots + x_{[k]}$, $P^*(y) = \iota_{0 \preceq y \preceq 1, 1^T y = k}$
- $P = \|\cdot\|_2$: from the projection on $\|\cdot\|_2 \le 1$,
$$\operatorname{prox}_{\lambda P}(x) = \vec{\Theta}_{\text{soft}}(x; \lambda) = \Theta_{\text{soft}}(\|x\|_2; \lambda) x^\circ \quad \left(0 \cdot \frac{0}{0} := 0\right)$$

Extension to thresholding

- In practice the penalties (or losses) of interest are often nonconvex. We consider a nonconvex extension of prox
- A threshold function is a real-valued function $\Theta(t;\lambda)$ defined for $-\infty < t < \infty$ and $0 \leq \lambda < \infty$ such that (i) $\Theta(-t;\lambda) = -\Theta(t;\lambda)$; (ii) $\Theta(t;\lambda) \leq \Theta(t';\lambda)$ for $t \leq t'$; (iii) $\lim_{t\to\infty} \Theta(t;\lambda) = \infty$; (iv) $0 \leq \Theta(t;\lambda) \leq t$ for $t \geq 0$.
- Given any Θ, $\vec{\Theta}$ is defined for any vector $a \in \mathbb{R}^m$ such that $\vec{\Theta}(a;\lambda) = a\Theta(\|a\|_2;\lambda)/\| a\|_2$ for $a \neq 0$ and 0 o/w

$\Theta \to P$

- A sparsity-inducing penalty should result in some kind of thresholding rule (*many-to-one*)
- Given an arbitrary thresholding Θ, let P be any function associated with Θ through

$$P(t;\lambda) - P(0;\lambda) = P_\Theta(t;\lambda) + q(t;\lambda),$$

$$P_\Theta(t;\lambda) = \int_0^{|t|} [\sup\{s : \Theta(s;\lambda) \le u\} - u]\, \mathrm{d}u$$

  for some nonnegative $q(\theta;\lambda)$ satisfying $q\{\Theta(\cdot;\lambda)\} = 0$
- When Θ has discontinuities, there are infinitely many q

- Then, $\hat{\beta} = \vec{\Theta}(y; \lambda)$ is a globally optimal solution to (S09, 12)

$$\min_{\beta \in \mathbb{R}^n} \frac{1}{2}\|y - \beta\|_2^2 + P(\|\beta_j\|_2; \lambda)$$

- A componentwise version: $\Theta(y; \lambda)$ solves $\sum P(\beta_j; \lambda)$
- The solution is not unique when Θ had discontinuities
- Examples: ridge-scaling $\to \ell_2$, soft $\to \ell_1$, elastic net; SCAD, MCP, ℓ_r $(0 < r < 1)$, capped ℓ_1 (nonconvex)

- A particular instance is the hard-thresholding

$$\Theta_H(t;\lambda) = t 1_{|t|\geq\lambda},$$

which induces

$$P_H(t;\lambda) = (-\frac{t^2}{2} + \lambda|t|)1_{|t|<\lambda} + \frac{\lambda^2}{2}1_{|t|\geq\lambda},$$
$$P_0(t;\lambda) = \frac{\lambda^2}{2}1_{t\neq 0}$$

- The 1st uses $q \equiv 0$. The 2nd: $q = \frac{(|t|-\lambda)^2}{2}1_{0<|t|<\lambda}$
- Notice the nonconvexity and many-to-one mapping

Generalized Moreau for robust estimation

- Standard robustification: OLS minimizes $\|y - X\beta\|_2^2$ or solves $X^T(X\beta - y) = 0$ (assume $n > p$ for now)
- Use a robust *loss*: $\min \sum_i \rho(y_i - X_i^T\beta)$
  - ρ: Huber's loss or a bounded nonconvex loss
- Use a ψ-*function*: $X^T\psi(X\beta - y) = 0$
  - ψ: Huber's ψ or a redescending ψ
- Modern challenges: theory, tuning, computation, etc.
- We give an additive robustification scheme

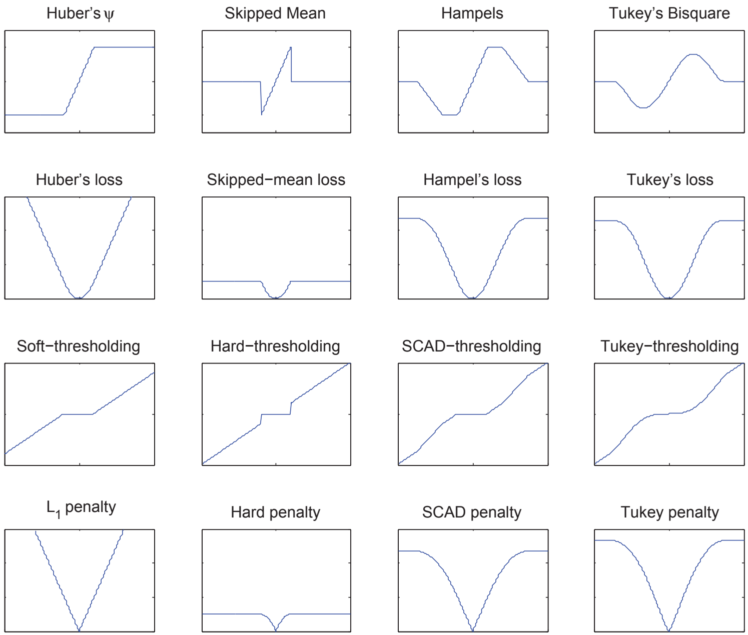
M-estimators & nonconvex penalized regression

- Let Θ be **any** thresholding rule which induces P
- Then given any coordinate minimum point $(\hat{\beta}, \hat{\gamma})$ of

$$\frac{1}{2}\|y - X\beta - \gamma\|_2^2 + \sum_{i=1}^{n} P(\gamma_i; \lambda_i),$$

$\hat{\beta}$ is necessarily an M-estimate associated with ψ (S & Owen 11), as long as (Θ, ψ) satisfies

$$\Theta + \psi = Id.$$


Huber's ψ
Skipped Mean
Hampels
Tukey's Bisquare
Huber's loss
Skipped−mean loss
Hampel's loss
Tukey's loss
Soft−thresholding
Hard−thresholding
SCAD−thresholding
Tukey−thresholding
L_1 penalty
Hard penalty
SCAD penalty
Tukey penalty

An **identity** on the (generalized) Moreau envelop

- The additive robust scheme goes beyond $n > p$, since (S & Chen 17)

$$\frac{1}{2}\{r-\Theta(r;\lambda)\}^2+P_\Theta\{\Theta(r;\lambda);\lambda\} = \int_0^{|r|} \psi(t;\lambda)\,\mathrm{d}t,\ \forall r \in \mathbb{R},$$

- So the equivalence holds much more generally, with β subject to an arbitrary constraint or penalty, and regardless of the number of responses and predictors

- Given any convex P, let Θ (ψ) be its (dual) proximity
- Let $M_P(r) = \frac{1}{2}\{r - \Theta(r;\lambda)\}^2 + P\{\Theta(r;\lambda);\lambda\}$, the Moreau envelope of P (with $1/\rho = 1$). Then

$$M_P(r) = \int_0^r \psi(t;\lambda)\,\mathrm{d}t + M_P(0)$$

- This is because $\psi = \mathrm{prox}_{P*} = \nabla M_P$
  - $M_P(y) = \inf_x \frac{1}{2}\|y - x\|_2^2 + P(x) \Rightarrow \nabla M_P(y) = y - \mathrm{prox}_P(y) \Rightarrow \mathrm{prox}_{P*}(y) = \nabla M_P(y)$

Proximal gradient method

- Proximity can help us design optimization algorithms
- Consider $\min_\beta l(\beta) + P(\beta)$, where prox_P is accessible
- Recall the gradient update: $\beta^{t+1} = \beta^t - \alpha_t \nabla l(\beta^t)$. Due to the existence of P, we add a proximity step:

$$\beta^{t+1} = \text{prox}_{(1/\rho_t)P}(\beta^t - \frac{1}{\rho_t}\nabla l(\beta^t)),$$

where $\rho^t > 0$

Linearization

- PGD is an outcome of **linearizing** the loss (only)

$$g_\rho(\beta, \beta^-) = l(\beta^-) + \langle \nabla l(\beta^-), \beta - \beta^- \rangle + \frac{\rho}{2}\|\beta - \beta^-\|_2^2 + P(\beta)$$

- In fact, $\beta^{t+1} = \arg\min_\beta g_{\rho_t}(\beta, \beta^t)$ which is equivalent to

$$\arg\min \frac{1}{2}\|\beta - (\beta^t - \frac{1}{\rho_t}\nabla l(\beta^t))\|_2^2 + \frac{1}{\rho_t} P(\beta)$$

- Power of linearization: a general loss l is now reduced to $\|\cdot\|_2^2$ (without any design) in g-optimization

Stepsize

- If ∇l is Lip(L), choosing $\rho_t \geq L$ guarantees

$$f(\beta^{t+1}) \leq g_{\rho_t}(\beta^{t+1}, \beta^t) \leq g_{\rho_t}(\beta^t, \beta^t) = f(\beta^t)$$

  where the 2nd inequality gives sufficient decrease
- In general, run a line search on ρ to meet the criterion

$$f(\beta^{t+1}(\rho)) \leq g_\rho(\beta^{t+1}(\rho), \beta^t)$$

- The analysis is similar to that of gradient descent

Example: lasso

- Lasso: $l(\beta) = \|y - X\beta\|_2^2/2$, $P(\beta) = \lambda\|\beta\|_1$
- Proximal gradient results in iterative soft-thresholding

$$\begin{aligned}\beta^{t+1} &= \arg\min \|\beta - (\beta^t - \frac{1}{\rho}(X^T X\beta^t - X^T y))\|_2^2 + \frac{\lambda}{\rho}\|\beta\|_1 \\ &= \Theta_{\text{soft}}(\beta^t - \frac{1}{\rho}(X^T X\beta^t - X^T y); \frac{\lambda}{\rho})\end{aligned}$$

  where $\rho = \|X\|_2^2$
- The linearization removes the design matrix here

- It actually uses the subgradient of the next iterate: $\beta^{t+1} = \beta^t - \nabla l(\beta^t)/\rho - (\lambda/\rho)\widetilde{\mathrm{sgn}}(\beta^{t+1})$
- To speed its convergence, apply the 1st acceleration:

$$\gamma^{(t)} = \beta^{(t)} + \theta_t(\theta_{t-1}^{-1} - 1)(\beta^{(t)} - \beta^{(t-1)}),$$
$$\beta^{(t+1)} = \Theta_{\mathrm{soft}}(\gamma^t - \frac{1}{\rho}(X^T X\gamma^t - X^T y); \frac{\lambda}{\rho})$$

where $\theta_0 = 1$, $\theta_{t+1} = (\sqrt{\theta_t^4 + 4\theta_t^2} - \theta_t^2)/2$

- This algorithm shares similarity with the CD lasso
- PGD: $\mathcal{O}(1/T)$, APG: $\mathcal{O}(1/T^2)$. CD: exact min
- The technique also applies to nonconvex losses and/or nonconvex penalties like SCAD, MCP, ℓ_r $(r \geq 0)$ (S 09)
- Note that the linearization step can always be accelerated (S & Wang 17)

Example: classification with feature clustering

- The problem can be formulated as (S 10)

$$\min -\langle y, X\beta\rangle + \langle 1, b(X\beta)\rangle + \lambda \sum_{j \neq j'} w_{j,j'} |\beta_j - \beta_{j'}|$$

- Here, $b(t) = \log(1 + \exp(t))$. We can introduce a sparse matrix $T \in \mathbb{R}^{\frac{p(p-1)}{2} \times p}$ to denote the pairwise differences.
- **Linearization**: $g(\beta, \beta^-) = l(\beta^-) + \langle X^T(b'(X\beta) - y),$ $\beta - \beta^-\rangle + \lambda \|T\beta\|_1 + \frac{\rho}{2}\|\beta - \beta^-\|_2^2$, $\rho \geq \|\nabla^2 l\|_2 = \|X\|_2^2/4$

- So with the help of linearization, it suffices to solving

$$\frac{1}{2}\|z-\beta\|_2^2+\frac{\lambda}{\rho}\|T\beta\|_1$$

  But $\text{prox}_{\|T\cdot\|_1}$ does not have a closed form as T is 'tall'
- Introduce $\gamma = T\beta$ to decouple: $\|z-\beta\|_2^2+\|\gamma\|_1$.
- We can derive a dual algorithm or a primal-dual one or ADMM. An example based on PGD is given as follows.

- Let $L(\beta, \gamma, \nu) = \|z - \beta\|_2^2/2 + \lambda' \|\gamma\|_1 + \langle \nu, T\beta - \gamma \rangle$
- With $\beta^o(\nu) = z - T^T\nu$, we just need to solve

$$\max_{\nu} g(\nu) = \|z\|_2^2/2 - \|T^T\nu - z\|_2^2/2 - (\lambda' \|\cdot\|_1)^*(\nu)$$

- Applying proximal gradient on the dual leads to

$$\begin{aligned} \nu^+ &= \mathrm{prox}_{(\lambda'\|\cdot\|_1)^*}(\nu - \varrho(TT^T\nu - Tz)) \\ &= \nu - \varrho(TT^T\nu - Tz) - \Theta(\nu - \varrho(TT^T\nu - Tz); \lambda') \end{aligned}$$

- A universal stepsize: $\varrho = 1/\|T\|_2^2$. **APG** can be used.

- Equivalently, we can write the algorithm as

$$\begin{aligned}\beta^+ &= z - T^T\nu \\ \gamma^+ &= \Theta(\nu + \varrho T\beta; \lambda')/\varrho = \Theta(T\beta + \nu/\varrho; \lambda'/\varrho) \\ \nu^+ &= \nu + \varrho(T\beta^+ - \gamma^+)\end{aligned}$$

- γ^+ is not the same one by minimizing L; interestingly, $\gamma^+ = \arg\min_\gamma \|z - \beta\|_2^2/2 + \lambda'\|\gamma\|_1 + \langle \nu, T\beta - \gamma\rangle + (\varrho/2)\|T\beta - \gamma\|_2^2$ (augmented Lagrangian)

- An alternative reparametrization via γ, $H \triangleq T^{+}$:

$$\min -\langle y, XH\gamma\rangle + \langle 1, b(XH\gamma)\rangle + \lambda\|\gamma\|_1 \text{ s.t. } TH\gamma = \gamma$$

- The linearization wrt $b(XH\cdot)$ reduces the problem to

$$\min_{\gamma} \frac{1}{2}\|z' - \gamma\|_2^2 + \lambda'\|\gamma\|_1 \text{ s.t. } P_T^{\perp}\gamma = 0$$

- Even though the penalty and the constraint are convex, it is not easy to get the optimal solution
  - Alternating prox/proj does not work in general!

Dykstra's projections

- Recall the dual problem for $\min \|y-\beta\|_2^2/2 + P_1(\beta) + P_2(\beta)$ (with μ, ν introduced for $\beta = \beta_1, \beta = \beta_2$)

$$\min_{\mu,\nu} \|y-\mu-\nu\|_2^2/2 + P_1^*(\mu) + P_2^*(\nu)$$

  where $\beta^o(\mu,\nu) = y - \mu - \nu$.
- Now apply BCD + Moreau decomposition:

$$\mu^+ = \mathrm{prox}_{P_1^*}(y-\nu) = y - \nu - \mathrm{prox}_{P_1}(y-\nu)$$
$$\nu^+ = \mathrm{prox}_{P_2^*}(y-\mu^+) = y - \mu^+ - \mathrm{prox}_{P_2}(y-\mu^+)$$

- Let $\beta = y - \mu - \nu$, $\beta^+ = y - \mu^+ - \nu$, $\beta^{++} = y - \mu^+ - \nu^+$.
- Then $\beta^+ = \text{prox}_{P_1}(y - \nu) = \text{prox}_{P_1}(\beta + \mu)$, $\beta^{++} = \text{prox}_{P_2}(\beta^+ + \nu)$ or

$$\begin{cases} \beta^+ = \text{prox}_{P_1}(\beta + \mu) \\ \mu^+ = \beta + \mu - \beta^+ \\ \beta^{++} = \text{prox}_{P_2}(\beta^+ + \nu) \\ \nu^+ = \beta^+ + \nu - \beta^{++} \end{cases}$$

Example: matrix completion

- The problem (noiseless version) is often defined by

$$\min_X \|X\|_* \quad \text{s.t. } X_{ij} = M_{ij}, \forall (i,j) \in \Omega$$

- In general, the problem is given by $\min \|X\|_*$ s.t. $\mathcal{A}(X) = b$, where the $\mathcal{A}$ is a **linear** mapping.
- To apply PGD, switch to an ℓ_2-regularized version and conduct successive optimization with $\lambda \to +\infty$:

$$\min \lambda \|X\|_* + \frac{1}{2}\|X\|_F^2 \quad \text{s.t. } \mathcal{A}(X) = b$$

- $g(Z) = \inf_X \lambda\|X\|_* + \|X\|_F^2/2 + \langle Z, \mathcal{A}(X) - b\rangle$
- Recall $\partial g(Z) = \mathbf{conv}\{\mathcal{A}(X^\star(Z)) - b\}$. Due to the s-convexity of $L(\cdot, Z)$, $g(\cdot)$ must be differentiable.
- Primal **proximity** + dual **ascent**:

$$\begin{cases} X^+ & = \Theta^\sigma(-\mathcal{A}^*(Z)); \lambda) \\ Z^+ & = Z + \alpha(\mathcal{A}(X^+) - b) \end{cases}$$

where $\langle Z, \mathcal{A}(X)\rangle = \langle \mathcal{A}^*(Z), X\rangle$(or $(\,\mathrm{vec}Z)^T A\,\mathrm{vec}X = (A^T\,\mathrm{vec}Z)^T\,\mathrm{vec}X$) and $\alpha = 1/\|\mathcal{A}\|_2^2$ (say)
- [*Augmented* Lagrangian: $+(\rho/2)\|\mathcal{A}(X) - b\|_F^2$]