AIE6003: Optimization for Machine Learning

Xiaopeng Li
CUHK(SZ)
SAI

April 4, 2026

Outline

Fenchel Duality

Dual First-Order Methods

ALM & ADMM

Fenchel Conjugate

Definition: for $f : \mathbb{R}^n \to \mathbb{R} \cup \{+\infty\}$, the Fenchel conjugate is

$$f^*(p) = \sup_{x \in \mathbb{R}^n} \{p^\top x - f(x)\}.$$

Geometric interpretation: negative y-intercept of the tightest affine lower bound of f with slope p.

Examples:

- $f(x) = \frac{1}{2}\|x\|^2$. Since the maximizer is $x = p$,

$$f^*(p) = \sup_x \left\{p^\top x - \tfrac{1}{2}\|x\|^2\right\} = \tfrac{1}{2}\|p\|^2$$

- $f(x) = \|x\|$.

$$f^*(p) = \begin{cases} 0 & \|p\|_* \le 1 \\ +\infty & \text{otherwise} \end{cases} = \text{indicator of dual-norm ball.}$$

- $f(x) = \delta_\Omega(x)$.

$$f^*(p) = \sup_{x \in \Omega} p^\top x = \text{support function of } \Omega$$

Basic Properties

Convexity: f^* is convex, even if f is not convex.

Closedness: f^* is lower semi-continuous.

Fenchel–Young: for any x and p,

$$f(x) + f^*(p) \geq p^\top x.$$

Equality: $f(x) + f^*(p) = p^\top x$ if and only if x achieves the supremum in $f^*(p) = \sup_z \{p^\top z - f(z)\}$, i.e., x is a maximizer.

Biconjugate Theorem

Biconjugate: apply conjugation twice:

$$f^{**}(x) = \sup_{p \in \mathbb{R}^n} \{p^\top x - f^*(p)\}.$$

We always have $f^{**} \leq f$ by Fenchel–Young.

Biconjugate Theorem

If f is proper, closed, and convex, then $f^{**} = f$.

Meaning: a closed convex function is fully characterized by its conjugate.

If f is not convex, then f^{**} is the convex envelope of f, the largest convex function that lower bounds f.

Conjugate and Subdifferential

Conjugate Subgradient Theorem

For f proper, closed, and convex, the following are equivalent:

$$p \in \partial f(x) \iff x \in \partial f^*(p) \iff f(x) + f^*(p) = p^\top x.$$

Proof sketch:

- $p \in \partial f(x)$ means $f(z) \geq f(x) + p^\top(z - x) \ \forall\, z$, i.e., $f^*(p) = p^\top x - f(x)$.
- This is exactly $f(x) + f^*(p) = p^\top x$.
- By biconjugate, the same argument with f^* gives $x \in \partial f^*(p)$.

Consequence: if f is differentiable, then $p = \nabla f(x)$ iff $x = \nabla f^*(p)$. Conjugation is the inverse of the gradient map.

Smoothness & Strong Convexity Duality

Smoothness & Strong Convexity

Let f be proper, closed, and convex. Then

- f is $\frac{1}{\gamma}$-smooth $\iff$ f^* is γ-strongly convex.
- f is γ-strongly convex $\iff$ f^* is $\frac{1}{\gamma}$-smooth.

Setup: let $p_i = \nabla f(x_i)$, so $x_i = \nabla f^*(p_i)$ and $f^*(p_i) = p_i^\top x_i - f(x_i)$.

Proof: it suffices to show

- f is $\frac{1}{\gamma}$-smooth $\implies$ f^* is γ-strongly convex.
- f is γ-strongly convex $\implies$ f^* is $\frac{1}{\gamma}$-smooth.

If so then the converse immediately follows from biconjugate theorem.

Proof

Claim: f is $\frac{1}{\gamma}$-smooth $\implies$ f^* is γ-strongly convex.

Proof: by $\frac{1}{\gamma}$-smoothness, f has a quadratic upper bound:

$$f(x) \leq f(x_1) + p_1^\top (x - x_1) + \frac{1}{2\gamma}\|x - x_1\|^2.$$

So $f^*(p_2) \geq \sup_x \{p_2^\top x - f(x_1) - p_1^\top (x - x_1) - \frac{1}{2\gamma}\|x - x_1\|^2\}$.

Maximizing over x yields

$$\begin{aligned} f^*(p_2) &\geq p_2^\top x_1 - f(x_1) + \gamma\|p_2 - p_1\|^2 - \frac{\gamma}{2}\|p_2 - p_1\|^2 \\ &= \underbrace{p_1^\top x_1 - f(x_1)}_{f^*(p_1)} + \underbrace{(p_2 - p_1)^\top x_1}_{\nabla f^*(p_1)^\top (p_2 - p_1)} + \frac{\gamma}{2}\|p_2 - p_1\|^2. \end{aligned}$$

Proof

Claim: f is γ-strongly convex $\implies$ f^* is $\frac{1}{\gamma}$-smooth.

Proof: By γ-strong convexity of f:

$$\begin{aligned} f^*(p_2) &= p_2^\top x_2 - f(x_2) \\ &\leq p_2^\top x_2 - f(x_1) - p_1^\top (x_2 - x_1) - \frac{\gamma}{2}\|x_2 - x_1\|^2 \\ &= f^*(p_1) + (p_2 - p_1)^\top x_1 + (p_2 - p_1)^\top (x_2 - x_1) - \frac{\gamma}{2}\|x_2 - x_1\|^2. \end{aligned}$$

Apply Young's inequality,

$$(p_2 - p_1)^\top (x_2 - x_1) \leq \frac{1}{2\gamma}\|p_2 - p_1\|^2 + \frac{\gamma}{2}\|x_2 - x_1\|^2.$$

Notice that $(p_2 - p_1)^\top x_1 = \nabla f^*(p_1)^\top (p_2 - p_1)$, so

$$f^*(p_2) \leq f^*(p_1) + \nabla f^*(p_1)^\top (p_2 - p_1) + \frac{1}{2\gamma}\|p_2 - p_1\|^2.$$

Fenchel Duality Theorem

Setup: consider $\min_{x\in\mathbb{R}^n} \left[f(x)+g(x)\right]$ with f, g convex.

Trick: insert $p^\top x - p^\top x = 0$, by definition of Fenchel conjugate,

$$\begin{aligned} f(x)+g(x) &= (f(x)-p^\top x)+(g(x)+p^\top x) \\ &\geq -f^*(p)+(-g^*(-p)), \end{aligned}$$

Weak duality: for any p, $\min_x \left[f(x)+g(x)\right] \geq -f^*(p)-g^*(-p)$.

Fenchel Duality

If f, g are proper, closed, convex, and $\mathrm{dom}(f)\cap\mathrm{dom}(g)\neq\emptyset$, then

$$\min_{x\in\mathbb{R}^n} \left[f(x)+g(x)\right] = \max_{p\in\mathbb{R}^n} \left[-f^*(p)-g^*(-p)\right].$$

Outline

Fenchel Duality

Dual First-Order Methods

ALM & ADMM

Putting Conjugates to Work

How to build algorithms for constrained optimization using

- Fenchel conjugate f^* and its properties;
- relation between smoothness of f and strong convexity of f^*;
- subdifferential: $p \in \partial f(x) \iff x \in \partial f^*(p)$.

Recall: the equality-constrained problem and its dual:

$$\underbrace{\min_x \; f(x) \;\text{ s.t. }\; Ax = b}_{\text{primal}} \quad \Longleftrightarrow \quad \underbrace{\max_u \; g(u)}_{\text{dual}} .$$

Questions:

- Can we express $g(u)$ in terms of f^*?
- Can we compute $\nabla g(u)$ without knowing f^* explicitly?
- How to use duality to do convergence analysis?

Dual Function via Conjugate

Lagrangian: $\mathcal{L}(x, u) = f(x) + u^\top (Ax - b)$.

Dual function: minimize over x:

$$\begin{aligned} g(u) &= \min_x \{f(x) + u^\top Ax - u^\top b\} \\ &= \min_x \{f(x) + (A^\top u)^\top x\} - u^\top b \\ &= -\max_x \{(-A^\top u)^\top x - f(x)\} - u^\top b \\ &= -f^*(-A^\top u) - u^\top b. \end{aligned}$$

Dual problem:

$$\max_u \ g(u) = \max_u \{-f^*(-A^\top u) - u^\top b\}.$$

Observation: the dual is expressed entirely through f^*.

Subgradient of the Dual

Question: can we compute $\partial g(u)$ without knowing f^* explicitly?

Let $x(u) \in \arg\min_x \{f(x) + u^\top Ax\}$. Then:

$$\partial g(u) = Ax(u) - b.$$

How to see this?

- the optimality condition implies $-A^\top u \in \partial f(x(u))$.
- conjugate subgradient theorem implies $x(u) \in \partial f^*(-A^\top u)$.
- $\partial g(u) = A\,\partial f^*(-A^\top u) - b = Ax(u) - b$.

Interpretation: ∂g at u is the constraint violation $Ax(u) - b$.

- $Ax(u) = b$: stationary point of the dual (KKT point).
- $Ax(u) \neq b$: the violation tells us which direction to move u.

Dual Subgradient Method

Problem: $\max_u \; g(u)$, with $\partial g(u) = Ax(u) - b$.

Algorithm: start with $u^{(0)}$. For $k = 1, 2, \ldots$:

$$\begin{cases} x^{(k)} \in \arg\min_x \left\{ f(x) + {u^{(k-1)}}^\top Ax \right\} & \text{(primal update)} \\ u^{(k)} = u^{(k-1)} + t_k \left(Ax^{(k)} - b \right) & \text{(dual ascent)} \end{cases}$$

Step sizes t_k chosen by standard rules, e.g., $t_k = c/\sqrt{k}$.

Interpretation:

- Primal update: find x satisfying KKT stationarity for current u.
- Dual ascent: increase u in the direction of constraint violation.

Dual Gradient Ascent

If f is strictly convex, then f^* is differentiable, and $x(u) = \arg\min_x \{f(x) + u^\top Ax\}$ is unique.

The subgradient becomes a gradient: $\nabla g(u) = Ax(u) - b$.

Dual gradient ascent: start with $u^{(0)}$. For $k = 1, 2, \ldots$:

$$\begin{cases} x^{(k)} = \arg\min_x \left\{f(x) + {u^{(k-1)}}^\top Ax\right\} \\ u^{(k)} = u^{(k-1)} + t_k \left(Ax^{(k)} - b\right) \end{cases}$$

Since g is now differentiable, we can apply

- gradient ascent with optimal step sizes;
- accelerated gradient methods;
- proximal gradient methods, if the dual has additional constraints.

Convergence of Dual Gradient Ascent

Recall: f is m-strongly convex $\iff$ f^* is $\frac{1}{m}$-smooth.

Since $g(u) = -f^*(-A^\top u) - u^\top b$, by the chain rule:

$$\nabla g(u) = A\nabla f^*(-A^\top u) - b.$$

Thus, g is $\frac{\|A\|^2}{m}$-smooth.

Convergence rates:

- f is m-strongly convex: step size $t_k = \frac{m}{\|A\|^2}$, rate $O(1/k)$.
- f is m-strongly convex and ∇f is L-Lipschitz: g is both smooth and strongly concave, linear rate.

Primal vs Dual: Swapped Roles

	Assumption on f	**Rate**
Primal GD	ℓ-smooth	$O(1/k)$
Dual GD	m-strongly convex	$O(1/k)$
Primal GD	ℓ-smooth & m-strongly convex	linear
Dual GD	ℓ-smooth & m-strongly convex	linear

Observation: duality swap the roles of smoothness and strong convexity.

Tradeoff:

- Primal GD needs smoothness of f.
- Dual GD needs strong convexity of f (harder to satisfy in practice).
- But dual GD handles equality constraints naturally, while primal GD needs projection.

Outline

Fenchel Duality

Dual First-Order Methods

ALM & ADMM

Limitations of Dual Ascent

Recall: dual gradient ascent $O(1/k)$ rate requires f to be strongly convex, but many problems are not strongly convex, e.g., Lasso, hinge loss SVM, linear programs, etc.

Idea: modify the primal to force strong convexity:

$$\min_x \ f(x) + \frac{\rho}{2}\|Ax - b\|^2 \quad \text{s.t.} \quad Ax = b,$$

where $\rho > 0$ is a parameter.

Reason: for any feasible x, the penalty $\frac{\rho}{2}\|Ax - b\|^2 = 0$. So the modified problem has the same solutions as the original.

Benefit: when A has full column rank, $f(x) + \frac{\rho}{2}\|Ax - b\|^2$ is strongly convex even if f is not.

Augmented Lagrangian Method (ALM)

Augmented Lagrangian: for $\rho > 0$,

$$\mathcal{L}_\rho(x, u) = f(x) + u^\top (Ax - b) + \frac{\rho}{2}\|Ax - b\|^2.$$

Augmented dual function: $g_\rho(u) := \min_x \mathcal{L}_\rho(x, u)$.

Method of Multipliers: start with $u^{(0)}$. For $k = 1, 2, \ldots$:

$$\begin{cases} x^{(k)} = \arg\min_x \ \mathcal{L}_\rho(x, u^{(k-1)}) \\ u^{(k)} = u^{(k-1)} + \rho\,(Ax^{(k)} - b) \end{cases}$$

Compared to dual ascent:

- Same dual update structure, but with step size $t_k = \rho$.
- The x-update minimizes the augmented Lagrangian.
- Convergence is guaranteed without strong convexity of f.

ALM as Proximal Algorithm

Claim: let $h = -g$ (convex). ALM is the proximal point method on h:

$$u^{(k)} = \mathrm{prox}_{\rho h}(u^{(k-1)}) = \arg\min_u \left\{ -g(u) + \frac{1}{2\rho}\|u - u^{(k-1)}\|^2 \right\}.$$

Proof: the ALM x-update gives

$$0 \in \partial f(x^{(k)}) + A^\top(u^{(k-1)} + \rho(Ax^{(k)} - b)) = \partial f(x^{(k)}) + A^\top u^{(k)}.$$

So $x^{(k)} \in \arg\min_x \{f(x) + {u^{(k)}}^\top Ax\}$, meaning $Ax^{(k)} - b \in \partial g(u^{(k)})$.

Substituting $Ax^{(k)} - b = \frac{1}{\rho}(u^{(k)} - u^{(k-1)})$:

$$0 \in -\partial g(u^{(k)}) + \frac{1}{\rho}(u^{(k)} - u^{(k-1)}) = \partial h(u^{(k)}) + \frac{1}{\rho}(u^{(k)} - u^{(k-1)}),$$

which is exactly the optimality condition for $u^{(k)} = \mathrm{prox}_{\rho h}(u^{(k-1)})$.

Convergence Proof

Use optimality $\frac{1}{\rho}(u^{(k-1)} - u^{(k)}) \in \partial h(u^{(k)})$ and convexity of h,

$$h(u^*) \geq h(u^{(k)}) + \frac{1}{\rho}(u^{(k-1)} - u^{(k)})^\top (u^* - u^{(k)}).$$

Rearranging: $h(u^{(k)}) - h(u^*) \leq \frac{1}{\rho}(u^{(k-1)} - u^{(k)})^\top (u^{(k)} - u^*)$.

Use the identity $a^\top b = \frac{1}{2}(\|a\|^2 + \|b\|^2 - \|a-b\|^2)$:

$$h(u^{(k)}) - h(u^*) \leq \frac{1}{2\rho}(\|u^{(k-1)} - u^*\|^2 - \|u^{(k)} - u^*\|^2 - \|u^{(k)} - u^{(k-1)}\|^2).$$

Consequences:

- $\|u^{(k)} - u^*\|$ is non-increasing.
- Telescope over $k = 1, \ldots, K$:

$$\min_{1 \leq k \leq K} (h(u^{(k)}) - h(u^*)) \leq \frac{\|u^{(0)} - u^*\|^2}{2\rho K} = O(1/K).$$

- Feasibility $Ax^{(k)} - b \to 0$ because

$$\sum_k \|Ax^{(k)} - b\|^2 = \frac{1}{\rho^2} \sum_k \|u^{(k)} - u^{(k-1)}\|^2 < \infty$$

Limitation: Not Decomposable

Consider block-separable $f(x) = \sum_{i=1}^{B} f_i(x_i)$ with $A = [A_1, \ldots, A_B]$.

Dual ascent decomposes:

$$\min_x \{f(x) + u^\top A x\} = \min_{x_1,\ldots,x_B} \sum_{i=1}^{B} (f_i(x_i) + u^\top A_i x_i)$$

$$= \sum_{i=1}^{B} \underbrace{\min_{x_i} \{f_i(x_i) + u^\top A_i x_i\}}_{\text{independent in } x_i}.$$

ALM does not decompose:

$$\|Ax - b\|^2 = \left\| \sum_{i=1}^{B} A_i x_i - b \right\|^2 = \sum_i \|A_i x_i\|^2 + 2 \sum_{i<j} (A_i x_i)^\top (A_j x_j) + \cdots$$

The cross terms $(A_i x_i)^\top (A_j x_j)$ couple different blocks.

Dilemma: dual ascent decomposes but needs strong convexity; ALM converges without it but loses decomposability. Can we have both?

ADMM

ALM loses decomposability because it jointly minimizes all variables.

Idea: instead of joint minimization, alternate between blocks.

Setup: split the problem into two groups of variables:

$$\min_{x,z} \; f(x) + g(z) \quad \text{s.t.} \quad Ax + Bz = c.$$

Even if the original problem is not in this form, we can often reformulate:

$$\boxed{\min_{\beta} \; \tfrac{1}{2}\|y - X\beta\|^2 + \lambda\|\beta\|_1} \iff \boxed{\begin{array}{l} \min\limits_{\beta,\alpha} \; \tfrac{1}{2}\|y - X\beta\|^2 + \lambda\|\alpha\|_1 \\ \text{s.t.} \;\; \beta = \alpha. \end{array}}$$

ADMM: minimize over x with z fixing, then over z with x fixing, then update u. This alternating structure allows decomposition.

ADMM Algorithm

Augmented Lagrangian for the split problem:

$$\mathcal{L}_\rho(x, z, u) = f(x) + g(z) + u^\top (Ax + Bz - c) + \frac{\rho}{2}\|Ax + Bz - c\|^2.$$

Alternating **D**irection **M**ethod of **M**ultipliers (ADMM) splits this into alternating minimizations. For $k = 1, 2, \ldots,$

$$\begin{cases} x^{(k)} = \arg\min_x \ \mathcal{L}_\rho(x,\, z^{(k-1)},\, u^{(k-1)}) & (x\text{-update, } z \text{ fixed}) \\ z^{(k)} = \arg\min_z \ \mathcal{L}_\rho(x^{(k)},\, z,\, u^{(k-1)}) & (z\text{-update, } x \text{ fixed}) \\ u^{(k)} = u^{(k-1)} + \rho\,(Ax^{(k)} + Bz^{(k)} - c) & (\text{dual update}) \end{cases}$$

Differences from ALM:

- Each subproblem involves only x or z, not both.
- The z-update uses the freshly computed $x^{(k)}$ (Gauss–Seidel).
- $Ax^{(k)} + Bz^{(k)} - c \neq \nabla g_\rho(u^{(k-1)})$.

Convergence

ADMM Convergence Theorem

If f and g are closed and convex, and a solution exists, then $\forall\, \rho > 0$,

1. Residual: $Ax^{(k)} + Bz^{(k)} - c \to 0$.
2. Objective: $f(x^{(k)}) + g(z^{(k)}) \to f^* + g^*$.
3. Dual: $u^{(k)} \to u^*$.

While any $\rho > 0$ converges, the speed depends heavily on ρ:

- ρ large: fast feasibility, but subproblems are dominated by the penalty, insensitive to the objective.
- ρ small: subproblems are easy, but slow feasibility convergence.

Heuristics: monitor primal and dual residuals at each step:

$$r^{(k)} = Ax^{(k)} + Bz^{(k)} - c, \quad s^{(k)} = \rho A^\top B(z^{(k)} - z^{(k-1)}).$$

Increase ρ if $\|r^{(k)}\| \gg \|s^{(k)}\|$; decrease ρ if $\|s^{(k)}\| \gg \|r^{(k)}\|$.

Convergence

Primal residual: $r^{(k)} = Ax^{(k)} + Bz^{(k)} - c$. Measures constraint violation.

Dual residual: measures the gap from KKT stationarity.
The x-update gives

$$0 \in \partial f(x^{(k)}) + A^\top (u^{(k-1)} + \rho(Ax^{(k)} + Bz^{(k-1)} - c)).$$

Substituting $u^{(k)} = u^{(k-1)} + \rho(Ax^{(k)} + Bz^{(k)} - c)$:

$$0 \in \partial f(x^{(k)}) + A^\top u^{(k)} + \rho A^\top B(z^{(k-1)} - z^{(k)}).$$

Let $s^{(k)} = \rho A^\top B(z^{(k)} - z^{(k-1)})$. If $s^{(k)} \to 0$, stationarity holds.

Together: $r^{(k)} \to 0$ and $s^{(k)} \to 0 \implies$ KKT satisfied in the limit.

Rate: $\|r^{(k)}\|^2 + \|s^{(k)}\|^2 = O(1/k)$.

Caveat: $(x^{(k)}, z^{(k)})$ may not converge without additional assumptions.

Stopping criterion: $\|r^{(k)}\| \le \varepsilon_{\text{pri}}$ and $\|s^{(k)}\| \le \varepsilon_{\text{dual}}$.

Scaled Form

Motivation: we can simplify ADMM by completing the square.

Scaled dual variable: let $w = u/\rho$, then

$$\begin{aligned} & u^T(Ax + Bz - c) + \frac{\rho}{2}\|Ax + Bz - c\|^2 \\ &= \frac{\rho}{2}(2w^T(Ax + Bz - c) + \|Ax + Bz - c\|^2) \\ &= \frac{\rho}{2}\|Ax + Bz - c + w\|^2 - \frac{\rho}{2}\|w\|^2. \end{aligned}$$

Scaled ADMM: for $k = 1, 2, \ldots$:

$$\begin{cases} x^{(k)} = \arg\min_x \ f(x) + \frac{\rho}{2}\|Ax + Bz^{(k-1)} - c + w^{(k-1)}\|^2 \\ z^{(k)} = \arg\min_z \ g(z) + \frac{\rho}{2}\|Ax^{(k)} + Bz - c + w^{(k-1)}\|^2 \\ w^{(k)} = w^{(k-1)} + Ax^{(k)} + Bz^{(k)} - c \end{cases}$$

Connection to Proximal Operators

Special case: $A = I$, $B = -I$, $c = 0$.

The scaled ADMM updates simplify to:

$$\begin{cases} x^{(k)} = \arg\min_x \ f(x) + \frac{\rho}{2}\|x - (z^{(k-1)} - w^{(k-1)})\|^2 \\ z^{(k)} = \arg\min_z \ g(z) + \frac{\rho}{2}\|z - (x^{(k)} + w^{(k-1)})\|^2 \\ w^{(k)} = w^{(k-1)} + x^{(k)} - z^{(k)} \end{cases}$$

Notice: each update is a proximal operator!

$$\begin{cases} x^{(k)} = \mathrm{prox}_{f/\rho}\big(z^{(k-1)} - w^{(k-1)}\big) \\ z^{(k)} = \mathrm{prox}_{g/\rho}\big(x^{(k)} + w^{(k-1)}\big) \\ w^{(k)} = w^{(k-1)} + x^{(k)} - z^{(k)} \end{cases}$$

This is the Douglas–Rachford splitting method.

Example: Lasso

Lasso: $\min_\beta \ \frac{1}{2}\|y - X\beta\|^2 + \lambda\|\beta\|_1$.

Splitting: introduce a dummy variable $\alpha = \beta$ to separate two terms:

$$\min_{\beta,\alpha} \ \frac{1}{2}\|y - X\beta\|^2 + \lambda\|\alpha\|_1$$

$$\text{s.t.} \ \ \beta = \alpha.$$

Here $f(\beta) = \frac{1}{2}\|y - X\beta\|^2$, $g(\alpha) = \lambda\|\alpha\|_1$, $A = I$, $B = -I$, $c = 0$.

Scaled ADMM:

$$\begin{cases} \beta^{(k)} = \arg\min_\beta \ \frac{1}{2}\|y - X\beta\|^2 + \frac{\rho}{2}\|\beta - \alpha^{(k-1)} + w^{(k-1)}\|^2 \\ \alpha^{(k)} = \arg\min_\alpha \ \lambda\|\alpha\|_1 + \frac{\rho}{2}\|\beta^{(k)} - \alpha + w^{(k-1)}\|^2 \\ w^{(k)} = w^{(k-1)} + \beta^{(k)} - \alpha^{(k)} \end{cases}$$

Both subproblems have closed-form solutions.

Example: Lasso

β-**update:** solving a ridge regression,

$$(X^\top X + \rho I)\,\beta^{(k)} = X^\top y + \rho(\alpha^{(k-1)} - w^{(k-1)}).$$

Solution: $\beta^{(k)} = (X^\top X + \rho I)^{-1}(X^\top y + \rho(\alpha^{(k-1)} - w^{(k-1)}))$.

Note $X^\top X + \rho I$ is always invertible and can be precomputed once.

α-**update:** proximal operator of $\frac{\lambda}{\rho}\|\cdot\|_1$:

$$\alpha^{(k)} = \mathrm{prox}_{\lambda/\rho\,\|\cdot\|_1}(\beta^{(k)} + w^{(k-1)}) = S_{\lambda/\rho}(\beta^{(k)} + w^{(k-1)}),$$

where $S_t(x)_j = \mathrm{sign}(x_j)\max(|x_j| - t, 0)$ is soft-thresholding.

Interpretation: each ADMM iteration does ridge regression (smooth fit) followed by soft-thresholding (sparsify).

Example: Group Lasso

Group Lasso: variables are partitioned into G groups $\beta = (\beta_1, \ldots, \beta_G)$:

$$\min_{\beta} \ \frac{1}{2}\|y - X\beta\|^2 + \lambda \sum_{g=1}^{G} c_g \|\beta_g\|_2.$$

The ℓ_2 norm on each group encourages entire groups to be zero.

Splitting: same as Lasso, introduce a dummy variable $\alpha = \beta$:

$$\begin{aligned} \min_{\beta,\alpha} \ & \tfrac{1}{2}\|y - X\beta\|^2 + \lambda \sum_{g=1}^{G} c_g \|\alpha_g\|_2 \\ \text{s.t.} \ & \beta = \alpha. \end{aligned}$$

β**-update:** identical to Lasso, ridge regression:

$$\beta^{(k)} = (X^\top X + \rho I)^{-1}(X^\top y + \rho(\alpha^{(k-1)} - w^{(k-1)})).$$

α**-update:** group-wise block soft-thresholding:

$$\alpha_g^{(k)} = \left(1 - \frac{\lambda c_g/\rho}{\|\beta_g^{(k)} + w_g^{(k-1)}\|_2}\right)_+ (\beta_g^{(k)} + w_g^{(k-1)}).$$

Consensus ADMM

Problem: $\min_x \sum_{i=1}^{B} f_i(x)$, all agents share the same variable x.

Splitting: give each agent a local copy x_i, enforce consensus:

$$\min_{x_1,\ldots,x_B,\,z} \sum_{i=1}^{B} f_i(x_i) \quad \text{s.t.} \quad x_i = z, \ \ \forall\, i.$$

Scaled ADMM updates:

$$\begin{cases} x_i^{(k)} = \arg\min_{x_i} \ f_i(x_i) + \frac{\rho}{2}\|x_i - z^{(k-1)} + w_i^{(k-1)}\|^2, \quad i = 1,\ldots,B \\ z^{(k)} = \frac{1}{B}\sum_{i=1}^{B}\left(x_i^{(k)} + w_i^{(k-1)}\right) \\ w_i^{(k)} = w_i^{(k-1)} + x_i^{(k)} - z^{(k)}, \quad i = 1,\ldots,B \end{cases}$$

Communication pattern:

- Each agent i solves a local subproblem independently.
- Server averages local solutions to form consensus $z^{(k)}$.
- Applicable in federated learning or distributed learning paradigm.

Wrap-up and Next Week

Week 10: Key takeaways

- Fenchel conjugate, properties, and Fenchel duality.
- Dual subgradient method, convergence.
- ALM, proximal viewpoint, convergence, limitation.
- ADMM, convergence, scaled-form, examples, consensus ADMM.

Thank you for your attendance!

Questions?

Next week (Week 11): low-rank optimization