

Section 8

Convex Optimisation 2

Lagrangian

Consider a general optimization problem

$$\begin{aligned} \min_{\mathbf{x}} \quad & f(\mathbf{x}) \\ \text{subject to} \quad & h_i(\mathbf{x}) \leq 0, \quad i = 1, \dots, m, \\ & \ell_j(\mathbf{x}) = 0, \quad j = 1, \dots, r. \end{aligned}$$

The objective function f needs not to be convex. Of course we pay special attention to the convex case.

Definition 8.1 (Lagrangian)

$$L(\mathbf{x}, \mathbf{u}, \mathbf{v}) = f(\mathbf{x}) + \sum_{i=1}^m u_i h_i(\mathbf{x}) + \sum_{j=1}^r v_j \ell_j(\mathbf{x}).$$

Here $\mathbf{u} \in \mathbb{R}^m$, $\mathbf{v} \in \mathbb{R}^r$, and $\mathbf{u} \geq \mathbf{0}$.

*all elements
are positive*

Lagrange Dual Function

Definition 8.2 (Lagrange Dual Function)

$$g(\mathbf{u}, \mathbf{v}) := \min_{\mathbf{x} \in \mathbb{R}^n} L(\mathbf{x}, \mathbf{u}, \mathbf{v}) = \min_{\mathbf{x} \in \mathbb{R}^n} f(\mathbf{x}) + \sum_{i=1}^m u_i h_i(\mathbf{x}) + \sum_{j=1}^r v_j \ell_j(\mathbf{x}).$$

- For every feasible \mathbf{x} ($\mathbf{x} \in \mathcal{X}$), $L(\mathbf{x}, \mathbf{u}, \mathbf{v}) \leq f(\mathbf{x})$

$$L(\mathbf{x}, \mathbf{u}, \mathbf{v}) = f(\mathbf{x}) + \underbrace{\sum_{i=1}^m u_i h_i(\mathbf{x})}_{\leq 0} + \underbrace{\sum_{j=1}^r v_j \ell_j(\mathbf{x})}_{=0}.$$

- Let \mathcal{X} denote the primal feasible set.

$$g(\mathbf{u}, \mathbf{v}) = \min_{\mathbf{x} \in \mathbb{R}^n} L(\mathbf{x}, \mathbf{u}, \mathbf{v}) \overset{\text{subset}}{\leq} \min_{\mathbf{x} \in \mathcal{X}} L(\mathbf{x}, \mathbf{u}, \mathbf{v}) \overset{\text{definition}}{\leq} f(\mathbf{x}). \quad (13)$$

maximize

Concavity of Lagrange Dual Function

hyperplane: both convex and concave

$$\begin{aligned} f(u, v) &= \min_x L(x, u, v) \\ &= \min_x (f(x) + \sum u_i h_i(x) + \sum v_j \ell_j(x)) \end{aligned}$$

Lemma 8.3

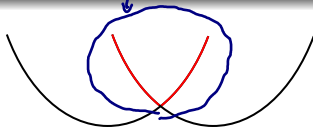
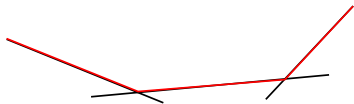
The Lagrange dual function

$$g(u, v) = \min_{x \in \mathbb{R}^n} L(x, u, v) = \min_{x \in \mathbb{R}^n} f(x) + \sum_{i=1}^m u_i h_i(x) + \sum_{j=1}^r v_j \ell_j(x)$$

is concave in (u, v) .

Lemma 8.4

- ▶ Let $f_\alpha(x)$ be concave functions. Then $g(x) = \inf_\alpha f_\alpha(x)$ is concave.
- ▶ Let $f_\alpha(x)$ be convex functions. Then $g(x) = \sup_\alpha f_\alpha(x)$ is convex.



Proofs

Proof of Lemma 8.4: For any $\lambda \in [0, 1]$,

$$\begin{aligned} g(\lambda \mathbf{x} + (1 - \lambda) \mathbf{y}) &= \inf_{\alpha} f_{\alpha}(\lambda \mathbf{x} + (1 - \lambda) \mathbf{y}) \\ &\stackrel{\text{concave}}{\geq} \inf_{\alpha} [\lambda f_{\alpha}(\mathbf{x}) + (1 - \lambda) f_{\alpha}(\mathbf{y})] \stackrel{\text{total inf}}{\geq} \lambda \inf_{\alpha} f_{\alpha}(\mathbf{x}) + (1 - \lambda) \inf_{\alpha} f_{\alpha}(\mathbf{y}) \stackrel{\text{sum of individual inf}}{\geq} \lambda g(\mathbf{x}) + (1 - \lambda) g(\mathbf{y}) \end{aligned}$$

Proof of Lemma 8.3: For any given \mathbf{x} , $L(\mathbf{x}, \mathbf{u}, \mathbf{v})$ is linear in (\mathbf{u}, \mathbf{v}) , and hence concave in (\mathbf{u}, \mathbf{v}) . The minimum of concave functions is concave based on Lemma 8.4.

$$g(\mathbf{u}, \mathbf{v}) = \min_{\mathbf{x} \in \mathbb{R}^n} L(\mathbf{x}, \mathbf{u}, \mathbf{v})$$

Lagrange Dual Problem

Given the primal problem

$$\begin{aligned} & \min_{\mathbf{x}} f(\mathbf{x}) \\ & \text{subject to } h_i(\mathbf{x}) \leq 0, \quad i = 1, \dots, m, \\ & \quad \quad \ell_j(\mathbf{x}) = 0, \quad j = 1, \dots, r. \end{aligned}$$

Its Lagrange dual problem is

$$\max_{\mathbf{u} \in \mathbb{R}^m, \mathbf{v} \in \mathbb{R}^r} g(\mathbf{u}, \mathbf{v}), \quad \text{subject to } \mathbf{u} \geq \mathbf{0}.$$

Weak and Strong Duality

$$g^* = \max_{u,v} g(u,v) \leq \min_x f(x) = f^*$$

Weak duality: the dual optimal value g^* satisfies

$$f^* \geq g^*.$$

This is a direct consequence of (13).

Strong duality is referred to as the case that

$$f^* = g^*.$$

Slater's condition: if the primal is a ^①convex problem (i.e., f and g_i 's are convex and ℓ_j 's are affine), and there exists ^②at least one strictly feasible $\mathbf{x} \in \mathbb{R}^n$ satisfying

$$h_i(\mathbf{x}) < 0, \forall i \in [m], \text{ and } \ell_j(\mathbf{x}) = 0, \forall j \in [r],$$

then **strong duality holds**. (Proof is omitted.)

Karush-Kuhn-Tucker conditions

Suppose only equality constraints.

$$1) 0 = \nabla f(x) + \sum \lambda_i \nabla g_i$$

$$2) g_i = 0$$

Given the optimization problem

$$\min_x \frac{1}{2} x^T A x + b^T x \quad \text{s.t. } Cx = 0$$

$$L(x, \nu) = \frac{1}{2} x^T A x + b^T x + \nu^T Cx$$

minimize $f(x)$

$$\frac{dL}{dx} = Ax + b + C^T \nu = 0$$

$$\begin{bmatrix} A & C^T \\ C & 0 \end{bmatrix} \begin{bmatrix} x \\ \nu \end{bmatrix} = \begin{bmatrix} -b \\ 0 \end{bmatrix}$$

subject to $h_i(x) \leq 0, i = 1, \dots, m,$
 $\ell_j(x) = 0, i = 1, \dots, r.$

$$\Rightarrow \begin{bmatrix} x \\ \nu \end{bmatrix} = \begin{bmatrix} A & C^T \\ C & 0 \end{bmatrix}^+ \begin{bmatrix} -b \\ 0 \end{bmatrix}$$

The Karush-Kuhn-Tucker (KKT) conditions are:

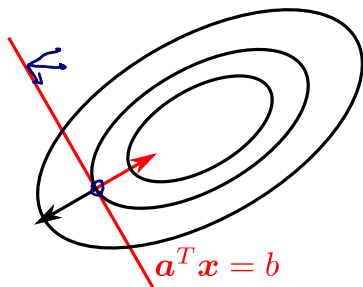
→ global optimal

- ▶ $0 \in \nabla f(x) + \sum_{i=1}^m u_i \nabla h_i(x) + \sum_{j=1}^r v_j \nabla \ell_j(x)$. (stationarity)
 (unconstrained subgradient includes 0)
 (linear combination of ∇ is zero)
- ▶ $u_i h_i(x) = 0, \forall i.$ (complementary slackness)
- ▶ $h_i(x) \leq 0, \ell_j(x) = 0, \forall i, \forall j.$ (primal feasibility)
- ▶ $u_i \geq 0, \forall i.$ (dual feasibility)

KKT conditions are

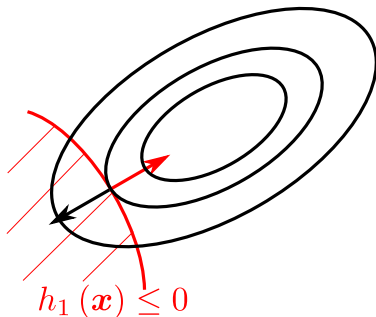
- ▶ Always sufficient.
- ▶ Necessary under strong duality.

Geometric Intuition: Equality Constraints



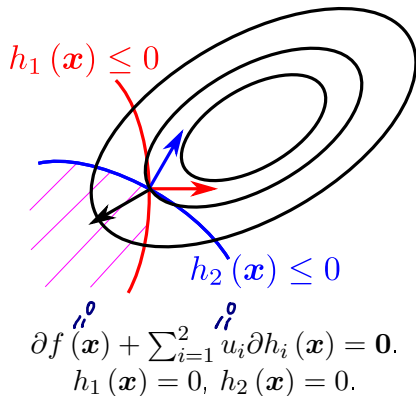
$\partial f(x)$ is a linear combination of $\partial \ell_j(x)$'s.

Geometric Intuition: One Inequality Constraint

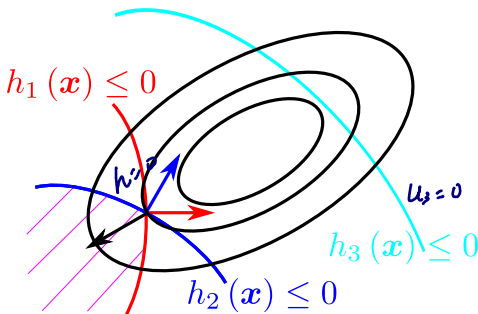


$$\begin{aligned}\partial f(\mathbf{x}) + u_1 \partial h_1(\mathbf{x}) &= \mathbf{0}. \\ h_1(\mathbf{x}) &= 0.\end{aligned}$$

Geometric Intuition: Inequality Constraints



Geometric Intuition: Inequality Constraints



$$\begin{aligned}\partial f(\mathbf{x}) + \sum_{i=1}^3 u_i \partial h_i(\mathbf{x}) &= \mathbf{0}. \\ h_1(\mathbf{x}) &= 0, \quad h_2(\mathbf{x}) = 0, \\ h_3(\mathbf{x}) &< 0 \text{ but } u_3 = 0 \text{ so that } u_3 h_3(\mathbf{x}) = 0.\end{aligned}$$

Sufficiency $g(u, v) = \min_{x \in \mathbb{R}^n} f(x) + \sum_i u_i h_i(x) + \sum_j v_j l_j(x)$

If x^*, u^*, v^* satisfy the KKT conditions, then x^* and u^*, v^* are primal and dual solutions.

If x^*, u^*, v^* satisfy the KKT conditions, then

$$g(u^*, v^*) \stackrel{\substack{\text{definition} \\ \text{stationary}}}{=} f(x^*) + \sum_{i=1}^m u_i^* h_i(x^*) + \sum_{j=1}^r v_j^* l_j(x^*)$$

$$\stackrel{\substack{\text{complementary} \\ \text{slackness}}}{=} f(x^*),$$

where the first equality follows from stationarity, and the second follows from complementary slackness. This equality suggests the duality gap is zero. Hence, x^*, u^* and v^* are primal and dual optimal.

Necessity

KKT holds \Leftrightarrow zero duality gap

Suppose that the strong duality holds and that \mathbf{x}^* and $\mathbf{u}^*, \mathbf{v}^*$ are primal and dual solutions. Then $\mathbf{x}^*, \mathbf{u}^*, \mathbf{v}^*$ satisfy the KKT conditions.

Due to the strong duality, one has

$$\begin{aligned} f(\mathbf{x}^*) &= g(\mathbf{u}^*, \mathbf{v}^*) \\ &= \min_{\mathbf{x} \in \mathbb{R}^n} f(\mathbf{x}) + \sum_{i=1}^m u_i^* h_i(\mathbf{x}) + \sum_{j=1}^r v_j^* \ell_j(\mathbf{x}) \\ &\leq f(\mathbf{x}^*) + \sum_{i=1}^m u_i^* h_i(\mathbf{x}^*) + \sum_{j=1}^r v_j^* \ell_j(\mathbf{x}^*) \\ &\leq f(\mathbf{x}^*). \end{aligned}$$

In other words, all the inequalities are actually equalities.

Quadratic Programming with Equality Constraints

$$f(x) = \frac{1}{2}x^T Q x + c^T x$$

$$L(v, x) = \frac{1}{2}x^T Q x + c^T x + v^T A x$$

Let $Q \succeq 0$.

$$\frac{\partial L}{\partial v} = 0 \Rightarrow A x = 0$$

$$\frac{\partial L}{\partial x} = 0 \Rightarrow Q x + c + A^T v = 0$$

$$\min_x \frac{1}{2}x^T Q x + c^T x \text{ subject to } A x = 0.$$

By KKT conditions, x is the minimizer if and only if

$$\begin{bmatrix} Q & A^T \\ A & 0 \end{bmatrix} \begin{bmatrix} x \\ u \end{bmatrix} = \begin{bmatrix} -c \\ 0 \end{bmatrix},$$

where the first set of linear equations come from the stationarity and the second set follows from the primal feasibility.

The optimal x^* can be obtained by solving the linear inverse problem.

Water Filling

$$L(u, v, x) = -\sum_i \log(\alpha_i + x_i) - \sum_i u_i x_i + v (\sum_i x_i - 1) = 0$$

$$\begin{cases} \frac{\partial L}{\partial x_i} = -\frac{1}{\alpha_i + x_i} - u_i + v = 0 \\ u_i x_i = 0 \\ x_i \geq 0, \mathbf{1}^T x = 1, u \geq 0 \end{cases} \Rightarrow \begin{cases} v \geq \frac{1}{\alpha_i + x_i} \\ (v - \frac{1}{\alpha_i + x_i}) x_i = 0 \end{cases}$$

$$\Rightarrow x_i = 0 \text{ or } v = \frac{1}{\alpha_i + x_i}$$

$$\min_x -\sum_{i=1} \log(\alpha_i + x_i) \text{ subject to } x \geq 0, \mathbf{1}^T x = 1.$$

$$v \alpha_i + v x_i = 1$$

$$x_i = \frac{1 - v \alpha_i}{v} = \frac{1}{v} - \alpha_i$$

By KKT conditions,

- ▶ $-1/(\alpha_i + x_i) - u_i + v = 0, \forall i$
- ▶ $u_i x_i = 0, \forall i$
- ▶ $x \geq 0, \mathbf{1}^T x = 1, u \geq 0.$

$$\Rightarrow \begin{cases} x_i = (\frac{1}{v} - \alpha_i)^+ \\ \sum_i x_i = 1 \end{cases}$$

Eliminate u . The first two conditions become

$$1/(\alpha_i + x_i) \leq v, \text{ and } x_i (v - 1/(\alpha_i + x_i)) = 0, \forall i.$$

Therefore, the solution:

$$x_i = \max(0, 1/v - \alpha_i)$$

where v is chosen such that

$$\sum_{i=1}^n \max(0, 1/v - \alpha_i) = 1.$$

Lasso Dual (1)

Lasso Primal:

$$\min_x \frac{1}{2} \|y - Ax\|_2^2 + \lambda \|x\|_1$$

low dim. unconstrained

This is equivalent to

$$\min_{x, y_r} \frac{1}{2} \|y_r\|_2^2 + \lambda \|x\|_1 \text{ s.t. } y_r = y - Ax$$

high dim. constrained

The Lagrangian:

$$L(x, y_r, u) = \frac{1}{2} \|y_r\|_2^2 + \lambda \|x\|_1 + \langle u, y - Ax - y_r \rangle$$

Lasso Dual (2)

$$L(x, y_r, u) = \frac{1}{2} \|y_r\|_2^2 + \lambda \|x\|_1 + \underbrace{\langle u, y_r - Ax \rangle}_{\text{const}} = \underbrace{\langle u, y_r \rangle}_{\text{const}} - \underbrace{\langle A^T u, x \rangle}_{\text{const}} - \underbrace{u^T y_r}_{\text{const}}$$

1.

$$\frac{dL}{dy_r} = y_r - u = 0 \Rightarrow L_1 = -\frac{1}{2} \|u\|_2^2$$

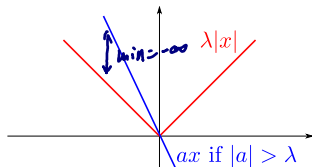
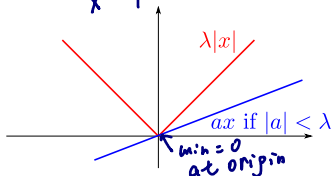
$$\min_{y_r} \frac{1}{2} \|y_r\|_2^2 - u^T y_r = -\frac{1}{2} \|u\|_2^2$$

The minimum is achieved when $y_r = u$.

2.

$$\min_x \lambda \|x\|_1 - \langle A^T u, x \rangle = \begin{cases} 0 & \text{if } \|A^T u\|_\infty \leq \lambda \\ -\infty & \text{otherwise} \end{cases}$$

$$= \min_x \sum_i (\lambda |x_i| - v_i x_i)$$



Lasso Dual (3)

Dual function:

$$-\frac{1}{2}\|\mathbf{u}\|_2^2 + \mathbf{u}^T \mathbf{y} \quad \text{s.t.} \quad \|\mathbf{A}^T \mathbf{u}\|_\infty \leq \lambda$$

L. const. L2=0

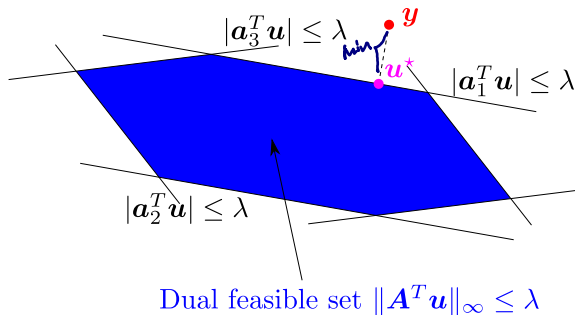
or equivalently

$$-\frac{1}{2}\|\mathbf{u} - \mathbf{y}\|_2^2 + \frac{1}{2}\|\mathbf{y}\|_2^2 \quad \text{s.t.} \quad \|\mathbf{A}^T \mathbf{u}\|_\infty \leq \lambda$$

Lasso dual problem: $\max_{\mathbf{u}} -\frac{1}{2}\|\mathbf{u}-\mathbf{y}\|_2^2 \quad \text{s.t.} \quad \|\mathbf{A}^T \mathbf{u}\|_\infty \leq \lambda$

$$\min_{\mathbf{u}} \|\mathbf{u} - \mathbf{y}\|_2^2 \quad \text{s.t.} \quad \|\mathbf{A}^T \mathbf{u}\|_\infty \leq \lambda$$

Lasso Dual (4)



Section 9

Alternating Direction Method of Multipliers (ADMM)

Boyd, Stephen, Neal Parikh, Eric Chu, Borja Peleato, and Jonathan Eckstein. "Distributed optimization and statistical learning via the alternating direction method of multipliers." *Foundations and Trends® in Machine learning* 3, no. 1 (2011): 1-122.

Dual Ascent Method (1)

Consider the convex optimization problem

$$\begin{aligned} \min \quad & f(\mathbf{x}) \\ \text{subject to} \quad & \mathbf{Ax} = \mathbf{b} \end{aligned} \tag{14}$$

Its Lagrangian is

$$L(\mathbf{x}, \mathbf{v}) = f(\mathbf{x}) + \mathbf{v}^T (\mathbf{Ax} - \mathbf{b})$$

The dual function

$$g(\mathbf{v}) = \inf_{\mathbf{x}} L(\mathbf{x}, \mathbf{v})$$

The dual problem

$$\max \quad g(\mathbf{v})$$

The Dual Ascent Method (2)

Dual ascent method

$$x^{k+1} = \operatorname{argmin}_x L(x, v^k) \quad \text{fix } v, \text{ minimise } x$$

$$v^{k+1} = v^k + \underbrace{\alpha^k}_{\text{step}} \nabla_v L(x^{k+1}, v) = v^k + \alpha^k (Ax^{k+1} - b) \quad \text{update } v \text{ by gradient ascent}$$

With appropriate chosen α^k , $g(v^{k+1}) > g(v^k)$ and dual ascent method converges under some assumptions.

However, the required assumptions do not hold in many applications.

Augmented Lagrangian and the Method of Multipliers

Problem:

$$\min f(x) + \underbrace{\frac{\rho}{2} \|Ax - b\|_2^2}_{\text{extra term: only influence the speed of convergence.}}$$

subject to $Ax = b$

Lagrangian:

$$\underline{L_\rho(x, v) = f(x) + v^T (Ax - b) + \frac{\rho}{2} \|Ax - b\|_2^2}$$

Method of multipliers:



$$x^{k+1} = \underset{x}{\operatorname{argmin}} L_\rho(x, v^k)$$

$$v^{k+1} = v^k + \rho (Ax^{k+1} - b)$$

Note the fixed step size ρ .

The Method of Multipliers: Step Size ρ

$$L_\rho(\mathbf{x}, \mathbf{v}) = f(\mathbf{x}) + \mathbf{v}^T (\mathbf{Ax} - \mathbf{b}) + \frac{\rho}{2} \|\mathbf{Ax} - \mathbf{b}\|_2^2$$

The optimality conditions for (14) are primal and dual feasibility

$$\underbrace{\mathbf{Ax}^* - \mathbf{b}}_{\text{constraint}} = \mathbf{0}, \quad \nabla f(\underbrace{\mathbf{x}^*}_{\text{KKT}}) + \mathbf{A}^T \mathbf{v}^* = \mathbf{0}.$$

As \mathbf{x}^{k+1} minimizes $L_\rho(\mathbf{x}, \mathbf{v}^k)$, it holds that

$$\begin{aligned} \mathbf{0} &= \nabla_{\mathbf{x}} L_\rho(\mathbf{x}^{k+1}, \mathbf{v}^k) \\ &= \nabla_{\mathbf{x}} f(\mathbf{x}^{k+1}) + \mathbf{A}^T \left(\mathbf{v}^k + \underbrace{\rho(\mathbf{Ax}^{k+1} - \mathbf{b})}_{\mathbf{v}^{k+1}} \right) \\ &= \nabla_{\mathbf{x}} f(\mathbf{x}^{k+1}) + \mathbf{A}^T \mathbf{v}^{k+1}. \end{aligned}$$

Using ρ as step size, $(\mathbf{x}^{k+1}, \mathbf{v}^{k+1})$ is dual feasible.

ADMM (1)

ADMM solves problems in the form

$$\begin{aligned} & \min f(\mathbf{x}) + g(\mathbf{z}) \\ & \text{subject to } \mathbf{Ax} + \mathbf{Bz} = \mathbf{c} \end{aligned}$$

The optimal value of this problem is denoted by

$$p^* = \inf \{f(\mathbf{x}) + g(\mathbf{z}) \mid \mathbf{Ax} + \mathbf{Bz} = \mathbf{c}\}.$$

The augmented Lagrangian:

$$\begin{aligned} L_\rho(\mathbf{x}, \mathbf{z}, \mathbf{v}) = & f(\mathbf{x}) + g(\mathbf{z}) + \underbrace{\mathbf{v}^T (\mathbf{Ax} + \mathbf{Bz} - \mathbf{c})}_{\text{combine}} \\ & + \underbrace{\frac{\rho}{2} \|\mathbf{Ax} + \mathbf{Bz} - \mathbf{c}\|_2^2}_{\text{combine}} \end{aligned}$$

where $\rho > 0$.

ADMM (2)

ADMM is an iterative algorithm with iterations:

$$\begin{aligned} \mathbf{x}^{k+1} &= \underset{\mathbf{x}}{\operatorname{argmin}} L_{\rho}(\mathbf{x}, \mathbf{z}^k, \mathbf{v}^k) \\ \mathbf{z}^{k+1} &= \underset{\mathbf{z}}{\operatorname{argmin}} L_{\rho}(\mathbf{x}^{k+1}, \mathbf{z}, \mathbf{v}^k) \\ \mathbf{v}^{k+1} &= \mathbf{v}^k + \rho (\mathbf{A}\mathbf{x}^{k+1} + \mathbf{B}\mathbf{z}^{k+1} - \mathbf{c}). \end{aligned}$$

gradient ascent

ADMM is different from the method of multipliers which has iterations

$$\begin{aligned} \underbrace{(\mathbf{x}^{k+1}, \mathbf{z}^{k+1})}_{\substack{\text{joint} \\ \text{minimization}}} &= \underset{\mathbf{x}, \mathbf{z}}{\operatorname{argmin}} L_{\rho}(\mathbf{x}, \mathbf{z}, \mathbf{v}^k) \\ &\quad \mathbf{v}^{k+1} = \mathbf{v}^k + \rho (\mathbf{A}\mathbf{x}^{k+1} + \mathbf{B}\mathbf{z}^{k+1} - \mathbf{c}). \end{aligned}$$

Augmented Lagrangian.
 Scaled Form $\mathcal{L}_\rho(\mathbf{x}, \mathbf{z}, \mathbf{v}) = f(\mathbf{x}) + g(\mathbf{z}) + \mathbf{v}^T(\mathbf{Ax} + \mathbf{Bz} - \mathbf{c}) + \frac{\rho}{2} \|\mathbf{Ax} + \mathbf{Bz} - \mathbf{c}\|_2^2$
 Define the primal residual $= f(\mathbf{x}) + g(\mathbf{z}) + \mathbf{v}^T \mathbf{r} + \frac{\rho}{2} \|\mathbf{r}\|_2^2$

$$\mathbf{r} = \mathbf{Ax} + \mathbf{Bz} - \mathbf{c}.$$

Define the scaled dual variable $\mathbf{u} = \mathbf{v}/\rho$, then

$$\begin{aligned} \mathbf{v}^T \mathbf{r} + \frac{\rho}{2} \|\mathbf{r}\|_2^2 &= \frac{\rho}{2} \left\| \mathbf{r} + \frac{1}{\rho} \mathbf{v} \right\|_2^2 - \frac{1}{2\rho} \|\mathbf{v}\|_2^2 \\ &= \frac{\rho}{2} \|\mathbf{r} + \mathbf{u}\|_2^2 - \frac{\rho}{2} \|\mathbf{u}\|_2^2. \end{aligned}$$

The scaled form of ADMM:

$$\begin{aligned} \triangle \mathbf{x}^{k+1} &= \underset{\mathbf{x}}{\operatorname{argmin}} f(\mathbf{x}) + \frac{\rho}{2} \left\| \mathbf{Ax} + \mathbf{Bz}^k - \mathbf{c} + \mathbf{u}^k \right\|_2^2 \\ \mathbf{z}^{k+1} &= \underset{\mathbf{z}}{\operatorname{argmin}} g(\mathbf{z}) + \frac{\rho}{2} \left\| \mathbf{Ax}^{k+1} + \mathbf{Bz} - \mathbf{c} + \mathbf{u}^k \right\|_2^2 \\ \mathbf{u}^{k+1} &= \mathbf{u}^k + \mathbf{Ax}^{k+1} + \mathbf{Bz}^{k+1} - \mathbf{c}. \end{aligned}$$

Example 1

Lasso problem:

$$\min \frac{1}{2} \|\mathbf{y} - \mathbf{A}\mathbf{x}\|_2^2 + \lambda \underbrace{\|\mathbf{x}\|_1}_{\text{non-differentiable}}.$$

ADMM version:

$$\begin{aligned} \min \quad & \frac{1}{2} \|\mathbf{y} - \mathbf{A}\mathbf{x}\|_2^2 + \lambda \|\mathbf{z}\|_1 \\ \text{subject to} \quad & \mathbf{x} = \mathbf{z} \end{aligned}$$

ADMM iterations:

$$\begin{aligned} \mathbf{x}^{k+1} &= \underset{\mathbf{x}}{\operatorname{argmin}} \frac{1}{2} \|\mathbf{y} - \mathbf{A}\mathbf{x}\|_2^2 + \frac{\rho}{2} \left\| \mathbf{x} - \mathbf{z}^k + \mathbf{u}^k \right\|_2^2 \\ \mathbf{z}^{k+1} &= \underset{\mathbf{z}}{\operatorname{argmin}} \lambda \|\mathbf{z}\|_1 + \frac{\rho}{2} \left\| \mathbf{x}^{k+1} - \mathbf{z} + \mathbf{u}^k \right\|_2^2 \\ \mathbf{u}^{k+1} &= \mathbf{u}^k + \left(\mathbf{x}^{k+1} - \mathbf{z}^{k+1} \right). \end{aligned}$$

Example 2

Constrained Lasso problem:

$$\min \frac{1}{2} \|y - Ax\|_2^2 + \lambda \|x\|_1$$

$$\text{subject to } Bx \leq c$$

\Downarrow

convex set \rightarrow indicator function

ADMM version:

$$\frac{1}{2} \|y - Ax\|_2^2 + \lambda \|x\|_1 + \mathbb{1}_{Bx \leq c}(x)$$

$$\min \frac{1}{2} \|y - Ax\|_2^2 + g(z)$$

$$\text{subject to } \begin{bmatrix} I \\ B \end{bmatrix} x + \begin{bmatrix} -I & I \end{bmatrix} \begin{bmatrix} z_1 \\ z_2 \end{bmatrix} = \begin{bmatrix} 0 \\ c \end{bmatrix},$$

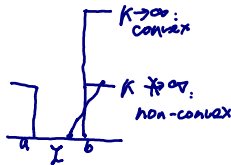
where

$$\begin{cases} Bx + z_2 = c \\ z_2 \geq 0 \end{cases}$$

$$g(z) = \lambda \|z_1\|_1 + \mathbb{1}_{\geq 0}(z_2)$$

$$\mathbb{1}_{\geq 0}(z) = \begin{cases} \infty & \text{if } z < 0 \\ 0 & \text{if } z \geq 0 \end{cases}.$$

indicator function — convex




Example 2 - Continued

ADMM Iterations:

$$\mathbf{x}^{k+1} = \underset{\mathbf{x}}{\operatorname{argmin}} \frac{1}{2} \|\mathbf{y} - \mathbf{A}\mathbf{x}\|_2^2 + \frac{\rho}{2} \left\| \mathbf{B}'\mathbf{x} + \mathbf{D}'\mathbf{z}^k - \mathbf{c}' + \mathbf{u}^k \right\|_2^2$$

$$\mathbf{z}^{k+1} = \underset{\mathbf{z}_1, \mathbf{z}_2}{\operatorname{argmin}} \underbrace{\lambda \|\mathbf{z}_1\|_1 + \frac{\rho}{2} \left\| \mathbf{x}^{k+1} - \mathbf{z}_1 + \mathbf{u}_1^k \right\|_2^2}_{\text{soft thresholding}} + \underbrace{\mathbb{1}_{\geq 0}(\mathbf{z}_2) + \frac{\rho}{2} \left\| \mathbf{B}\mathbf{x}^{k+1} + \mathbf{z}_2 - \mathbf{c} + \mathbf{u}_2^k \right\|_2^2}_{\text{projection}}$$

$$\mathbf{u}^{k+1} = \mathbf{u}^k + \mathbf{B}'\mathbf{x}^{k+1} + \mathbf{D}'\mathbf{z}^{k+1} - \mathbf{c}'.$$

$\mathbf{z}^k: \begin{cases} y_i > 0 \\ 0 < y_i < 0 \end{cases}$


Each step is easy to compute.

Optimality Conditions (1)

The necessary and sufficient conditions for optimality are primal feasibility

$$\mathbf{A}\mathbf{x}^* + \mathbf{B}\mathbf{z}^* - \mathbf{c} = \mathbf{0}. \quad (15)$$

and dual feasibility

$$\mathbf{0} \in \partial f(\mathbf{x}^*) + \mathbf{A}^T \mathbf{v}^* \quad (16)$$

$$\mathbf{0} \in \partial g(\mathbf{z}^*) + \mathbf{B}^T \mathbf{v}^*. \quad (17)$$

It turns out that \mathbf{z}^{k+1} and \mathbf{v}^{k+1} always satisfy (17):

$$\begin{aligned} \mathbf{0} &\in \partial g(\mathbf{z}^{k+1}) + \mathbf{B}^T \mathbf{v}^k + \rho \mathbf{B}^T (\mathbf{A}\mathbf{x}^{k+1} + \mathbf{B}\mathbf{z}^{k+1} - \mathbf{c}) \\ &= \partial g(\mathbf{z}^{k+1}) + \mathbf{B}^T \mathbf{v}^{k+1}. \end{aligned}$$

The situation about \mathbf{x}^{k+1} is different.

Optimality Conditions (2)

By definition \mathbf{x}^{k+1} minimizes $L_\rho(\mathbf{x}, \mathbf{z}^k, \mathbf{v}^k)$. It holds that

$$\begin{aligned} \mathbf{0} &\in \partial f(\mathbf{x}^{k+1}) + \mathbf{A}^T \mathbf{v}^k + \rho \mathbf{A}^T (\mathbf{A} \mathbf{x}^{k+1} + \mathbf{B} \mathbf{z}^k - \mathbf{c}) \\ &= \partial f(\mathbf{x}^{k+1}) + \mathbf{A}^T (\mathbf{v}^k + \rho \mathbf{r}^{k+1} + \rho \mathbf{B} (\mathbf{z}^k - \mathbf{z}^{k+1})) \\ &= \partial f(\mathbf{x}^{k+1}) + \mathbf{A}^T \mathbf{v}^{k+1} + \rho \mathbf{A}^T \mathbf{B} (\mathbf{z}^k - \mathbf{z}^{k+1}). \end{aligned}$$

Or equivalently

$$\rho \mathbf{A}^T \mathbf{B} (\mathbf{z}^{k+1} - \mathbf{z}^k) \in \partial f(\mathbf{x}^{k+1}) + \mathbf{A}^T \mathbf{v}^{k+1}.$$

The *dual residual* is defined as

$$\mathbf{s}^{k+1} = \rho \mathbf{A}^T \mathbf{B} (\mathbf{z}^{k+1} - \mathbf{z}^k).$$

Convergence of ADMM

Under mild conditions, ADMM converges:

- ▶ Primal residual convergence: $\mathbf{r}^k \rightarrow \mathbf{0}$ as $k \rightarrow \infty$.
- ▶ Objective convergence: $f(\mathbf{x}^k) + g(\mathbf{z}^k) \rightarrow \mathbf{p}^*$ as $k \rightarrow \infty$.
- ▶ Dual variable convergence: $\mathbf{v}^k \rightarrow \mathbf{v}^*$ as $k \rightarrow \infty$.

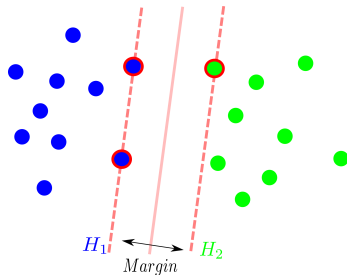
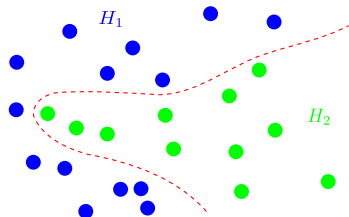
In practice, a reasonable criterion of terminating ADMM iterations is that the primal and dual residuals are small, i.e.,

$$\left\| \mathbf{r}^k \right\|_2 \leq \epsilon^{\text{pri}}, \quad \left\| \mathbf{s}^k \right\|_2 \leq \epsilon^{\text{dual}}.$$

Section 10

Support Vector Machine

Idea of SVM

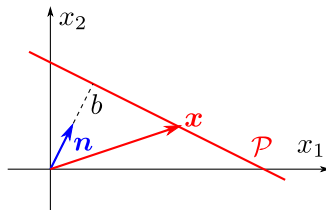


A Hyperplane

A hyperplane in \mathbb{R}^n can be defined using its normal vector $\mathbf{n} \in \mathbb{R}^n$:

$$\mathcal{P} = \{\mathbf{x} : \mathbf{n}^T \mathbf{x} = b\}.$$

- Usually we assume $\|\mathbf{n}\|_2 = 1$.



The projection $\|\text{Proj}(\mathbf{x}, \text{span}(\mathbf{n}))\|_2 = b$.

- If $\|\mathbf{n}\|_2 \neq 1$, then

$$\mathcal{P} = \{\mathbf{x} : \mathbf{n}^T \mathbf{x} = b\} = \{\mathbf{x} : \mathbf{n}^T \mathbf{x} / \|\mathbf{n}\|_2 = b / \|\mathbf{n}\|_2\}.$$

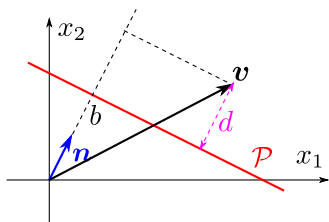
Distance to a Hyperplane

Define a hyperplane $\mathcal{P} = \{\mathbf{x} : \mathbf{n}^T \mathbf{x} = b\}$ where $\|\mathbf{n}\|_2 = 1$.

Let \mathbf{v} be an arbitrary point.

The distance between \mathbf{v} and \mathcal{P} is given by

$$d = d(\mathbf{v}, \mathcal{P}) = |\mathbf{n}^T \mathbf{v} - b|. \quad (18)$$



When $\|\mathbf{n}\|_2 \neq 1$,

$$d = \left| \frac{\mathbf{n}^T}{\|\mathbf{n}\|_2} \mathbf{v} - \frac{b}{\|\mathbf{n}\|_2} \right| = \frac{|\mathbf{n}^T \mathbf{v} - b|}{\|\mathbf{n}\|_2}. \quad (19)$$

SVM: Basics & KKT Conditions

SVM: Separate Points from Two Different Classes

Given training dataset $\{\mathbf{x}_i, y_i\}$ where the labels $y_i \in \{-1, 1\}$, want to find β and b s.t.

$$\beta^T \mathbf{x}_i + b \geq +1 \quad \text{for } y_i = +1, \quad y_i (\beta^T \mathbf{x}_i + b) \geq 1$$

$$\beta^T \mathbf{x}_i + b \leq -1 \quad \text{for } y_i = -1. \quad y_i (\beta^T \mathbf{x}_i + b) \geq 1$$

or equivalently

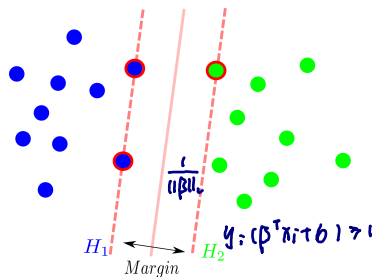
$$y_i (\beta^T \mathbf{x}_i + b) - 1 \geq 0, \quad \forall i.$$

$$\text{dist}(\mathbf{x}_i, \mathcal{P}) = \frac{|\beta^T \mathbf{x}_i + b|}{\|\beta\|_2} = \frac{1}{\|\beta\|_2}$$

In other words, find a hyperplane $\{\mathbf{x} : \beta^T \mathbf{x} + b\}^{\perp}$ s.t.

- ▶ Distance from one class to the hyperplane is $1/\|\beta\|_2$.
- ▶ Distance between the two classes (along direction β) is $2/\|\beta\|_2$.

SVM: Best Separation



SVM: a convex optimization problem:

$$\begin{aligned} \min_{\beta, b} \quad & \frac{1}{2} \|\beta\|_2^2, \\ \text{subject to} \quad & 1 - y_i (\beta^T x_i + b) \leq 0. \end{aligned} \tag{20}$$

Lagrange Dual Problem of SVM

Lagrangian of the SVM primal optimization problem:

$$L = \frac{1}{2} \|\beta\|^2 + \sum_i \lambda_i (1 - y_i (\beta^T x_i + b)), \quad (21)$$

where $\lambda_i \geq 0$.

$$\frac{\partial L}{\partial \beta} = \beta - \sum_i \lambda_i y_i x_i = 0$$

$$\frac{\partial L}{\partial b} = \sum_i \lambda_i y_i = 0$$

Lagrange Dual Problem $\Rightarrow L = \frac{1}{2} \sum_i \sum_j \lambda_i \lambda_j y_i y_j x_i^T x_j + \sum_i \lambda_i - \sum_i \lambda_i y_i (\sum_j \lambda_j y_j x_j^T x_i)$

$$= \sum_i \lambda_i - \frac{1}{2} \sum_i \sum_j \lambda_i \lambda_j y_i y_j x_i^T x_j$$

$$\max_{\lambda} \quad \min_{\beta, b} L$$

Lagrange dual function L_D

The Dual Function

To solve $\min_{\beta, b} L$, set $\partial L / \partial \beta$ and $\partial L / \partial b$ to zero:

$$\frac{\partial L}{\partial \beta} = \beta - \sum_i \lambda_i y_i \mathbf{x}_i = 0 \Rightarrow \beta = \sum_i \lambda_i y_i \mathbf{x}_i. \quad (22)$$

$$\frac{\partial L}{\partial b} = \sum_i \lambda_i y_i = 0. \quad (23)$$

Substitute (22) and (23) into the Lagrangian (21). It holds that

$$\begin{aligned} L_D &= \sum \lambda_i - \frac{1}{2} \|\beta\|_2^2 = \sum_i \lambda_i - \frac{1}{2} \sum_{i,j} \lambda_i \lambda_j y_i y_j \mathbf{x}_i^T \mathbf{x}_j \\ &= -\frac{1}{2} \boldsymbol{\lambda}^T \mathbf{K} \boldsymbol{\lambda} + \mathbf{1}^T \boldsymbol{\lambda}, \end{aligned} \quad (24)$$

where $K_{i,j} = y_i \mathbf{x}_i^T \mathbf{x}_j y_j$.

The Dual Problem

The dual problem becomes:

$$\begin{aligned} \max_{\boldsymbol{\lambda}} \quad & -\frac{1}{2}\boldsymbol{\lambda}^T \mathbf{K} \boldsymbol{\lambda} + \mathbf{1}^T \boldsymbol{\lambda}, \\ \text{subject to} \quad & \lambda_i \geq 0, \quad \forall i, \\ & \sum_i \lambda_i y_i = 0. \end{aligned} \tag{25}$$

The KKT Condition

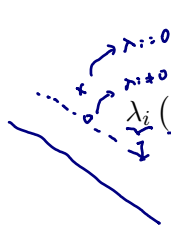
$$\frac{\partial L}{\partial \beta} = \beta - \sum_i \lambda_i y_i \mathbf{x}_i = 0, \quad (26)$$

$$\frac{\partial L}{\partial b} = \sum_i \lambda_i y_i = 0, \quad (27)$$

$$1 - y_i (\beta^T \mathbf{x}_i + b) \leq 0, \quad (28)$$

$$\lambda_i \geq 0, \quad (29)$$

$$\lambda_i (1 - y_i (\beta^T \mathbf{x}_i + b)) = 0. \quad (30)$$



SVM Classifier: Support Vectors

Condition (30) implies

$$\begin{cases} \text{if } \lambda_i \neq 0 & \text{then } 1 = y_i (\beta^T \mathbf{x}_i + b), \\ \text{if } 1 \neq y_i (\beta^T \mathbf{x}_i + b) & \text{then } \lambda_i = 0. \end{cases}$$

boundary points
inner points

Hence from (26),

$$\beta = \sum_{i \in \mathcal{I}} \lambda_i y_i \mathbf{x}_i, \quad \mathcal{I} = \{i : y_i (\beta^T \mathbf{x}_i + b) = 1 \quad (\text{or } \lambda_i \neq 0)\}.$$

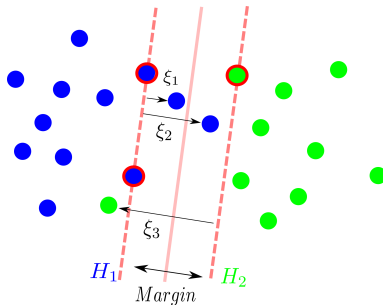
For a new test data \mathbf{x}^{new} ,

$$y^{\text{new}} = \text{sign} \left(\sum_{i \in \mathcal{I}} \lambda_i y_i \mathbf{x}_i^T \mathbf{x}^{\text{new}} + b \right).$$

The classifier only uses the boundary points (**sparsity!**).

SVM: Overlapping Classes

SVM for Overlapping Classes



Primal Problem for Overlapping Classes

The constraints:

$$\begin{aligned}\beta^T \mathbf{x}_i + b &\geq +1 - \xi_i & \text{for } y_i = +1, \\ \beta^T \mathbf{x}_i + b &\leq -1 + \xi_i & \text{for } y_i = -1,\end{aligned}$$

where $\xi_i \geq 0, \forall i$.

SVM Primal Problem:

$$\begin{aligned}\min_{\beta, b, \xi} \quad & \frac{1}{2} \|\beta\|_2^2 + C \left(\sum_i \xi_i \right)^k \\ \text{subject to} \quad & 1 - \xi_i - y_i (\beta^T \mathbf{x}_i + b) \leq 0, \\ & -\xi_i \leq 0, \quad \forall i,\end{aligned}$$

where $C > 0$ is a constant and k is a positive integer. Usually $k = 1$.

Dual Function

The Lagrangian

$$L = \frac{1}{2} \|\boldsymbol{\beta}\|_2^2 + C \sum \xi_i + \sum \lambda_i (1 - \xi_i - y_i (\boldsymbol{\beta}^T \mathbf{x}_i - b)) - \sum u_i \xi_i,$$

where $\lambda_i \geq 0$, $u_i \geq 0$ are Lagrange multipliers.

The dual function

$$L_D = \min_{\boldsymbol{\beta}, b, \boldsymbol{\xi}} L.$$

To find the dual function

$$\frac{dL}{d\boldsymbol{\beta}} = 0 \quad \Rightarrow \quad \boldsymbol{\beta} = \sum \lambda_i y_i \mathbf{x}_i.$$

$$\frac{dL}{db} = 0 \quad \Rightarrow \quad \sum \lambda_i y_i = 0.$$

$$\frac{dL}{d\xi} = 0 \quad \Rightarrow \quad C - \lambda_i - u_i = 0 \quad \Rightarrow \quad \lambda_i = C - u_i \leq C.$$

The Dual Problem

The dual problem:

$$\begin{aligned} \max_{\boldsymbol{\lambda}} \quad & \sum \lambda_i - \frac{1}{2} \|\boldsymbol{\beta}\|_2^2 = -\frac{1}{2} \boldsymbol{\lambda}^T \mathbf{K} \boldsymbol{\lambda} + \mathbf{1}^T \boldsymbol{\lambda} \\ \text{subject to} \quad & 0 \leq \lambda_i \leq C, \\ & \sum \lambda_i y_i = 0, \end{aligned}$$

where $K_{i,j} = y_i \mathbf{x}_i^T \mathbf{x}_j y_j$.

The only difference is that now λ_i 's are upper bounded by C .

Again, only **boundary points** are involved.

$$\boldsymbol{\beta} = \sum_{i \in \mathcal{I}} \lambda_i y_i \mathbf{x}_i, \quad \mathcal{I} = \{i : \lambda_i \neq 0\},$$

which comes from the KKT condition $\lambda_i (1 - \xi_i - y_i (\boldsymbol{\beta}^T \mathbf{x}_i + b)) = 0$.

SVM: General Cases

The General Case

- ▶ Two classes \Rightarrow multiple classes

- ▶ Regression

- ▶ Data space \Rightarrow feature space

Define a **kernel** function $\varphi : \mathbb{R}^n \rightarrow \mathcal{H}$ and work on the space of $\varphi(\mathbf{x}_i)$.

In SVM, what really matters is $\mathbf{x}_i^T \mathbf{x}_j$.

In the general case (**kernel method**), what matters is

$$\kappa(\mathbf{x}_i, \mathbf{x}_j) = \varphi^T(\mathbf{x}_i) \varphi(\mathbf{x}_j).$$

Example of nonlinear features:

- ▶ $\kappa(\mathbf{x}, \mathbf{y}) = (\mathbf{x}^T \mathbf{y})^2 = (x_1 y_1 + x_2 y_2)^2 = x_1^2 y_2^2 + 2x_1 y_1 x_2 y_2 + x_2^2 y_2^2$.

- ▶ $\kappa(\mathbf{x}, \mathbf{y}) = \varphi^T(\mathbf{x}) \varphi(\mathbf{y})$ with $\varphi(\mathbf{x}) = [x_1^2, \sqrt{2}x_1 x_2, x_2^2]^T$.

- ▶ $\kappa(\mathbf{x}, \mathbf{y}) = \exp\left(-\|\mathbf{x} - \mathbf{y}\|_2^2 / 2\sigma^2\right)$. *Gaussian kernel*

- ▶ $\varphi(\mathbf{x})$ has infinite dimension.

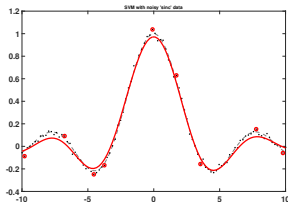
SVM for Regression

Regression problem: find β and b s.t.

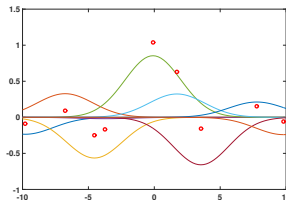
regression: label $y_i = 1$

$$\begin{aligned} y_i &= f(x_i) = \beta^T \varphi(x_i) + b \\ &= \sum_j \lambda'_j \varphi^T(x_j) \varphi(x_i) + b \\ &= \sum_j \lambda'_j \kappa(x_i, x_j) + b. \end{aligned}$$

$\varphi \rightarrow \kappa \rightarrow$ Gaussian function



=



The Primal Optimization Problem

Let $\epsilon > 0$ be the error tolerance. Then one has

$$\begin{aligned} \min \quad & \frac{1}{2} \|\boldsymbol{\beta}\|_2^2 \\ \text{subject to} \quad & |y_i - \boldsymbol{\beta}^T \varphi(\mathbf{x}_i) - b| \leq \epsilon. \end{aligned}$$

The constraints are equivalent to

$$\begin{aligned} y_i - \boldsymbol{\beta}^T \varphi(\mathbf{x}_i) - b &\leq \epsilon, \\ \boldsymbol{\beta}^T \varphi(\mathbf{x}_i) + b - y_i &\leq \epsilon. \end{aligned}$$

Now if we allow additional noise, represented by $\xi_i \geq 0$ and $\xi_i^* \geq 0$. Then

$$\begin{aligned} \min \quad & \frac{1}{2} \|\boldsymbol{\beta}\|_2^2 + C \sum (\xi_i + \xi_i^*) \\ \text{subject to} \quad & y_i - \boldsymbol{\beta}^T \varphi(\mathbf{x}_i) - b \leq \epsilon + \xi_i, \\ & \boldsymbol{\beta}^T \varphi(\mathbf{x}_i) + b - y_i \leq \epsilon + \xi_i^*, \\ & -\xi_i \leq 0, \quad -\xi_i^* \leq 0. \end{aligned}$$

Lagrangian

$$\begin{aligned} L = & \frac{1}{2} \|\boldsymbol{\beta}\|_2^2 + C \sum (\xi_i + \xi_i^*) - \sum_i \left(u_i \xi_i + \sum u_i^* \xi_i^* \right) \\ & + \lambda_i \left(y_i - \boldsymbol{\beta}^T \boldsymbol{\varphi}(\mathbf{x}_i) - b - \epsilon - \xi_i \right) \\ & + \lambda_i^* \left(\boldsymbol{\beta}^T \boldsymbol{\varphi}(\mathbf{x}_i) + b - y_i - \epsilon - \xi_i^* \right), \end{aligned}$$

where $u_i, u_i^*, \xi_i, \xi_i^* \geq 0$ are Lagrange multiplier. To minimize L ,

$$\begin{aligned} dL/d\boldsymbol{\beta} = \mathbf{0} & \Rightarrow \boldsymbol{\beta} = \sum_i (\lambda_i - \lambda_i^*) \boldsymbol{\varphi}(\boldsymbol{\xi}_i), \\ dL/db = \mathbf{0} & \Rightarrow \sum \lambda_i = \sum \lambda_i^*, \\ dL/d\xi_i = 0, dL/d\xi_i^* = 0 & \Rightarrow \lambda_i \leq C, \lambda_i^* \leq C. \end{aligned}$$

The Dual Problem

The **objective function** of the dual problem

$$L_D = -\epsilon \sum (\lambda_i + \lambda_i^*) + y_i \sum (\lambda_i - \lambda_i^*) \\ - \underbrace{\frac{1}{2} \sum_{i,j} (\lambda_i - \lambda_i^*) (\lambda_j - \lambda_j^*) \kappa(\mathbf{x}_i, \mathbf{x}_j)}_{\|\boldsymbol{\beta}\|_2^2},$$

where $\kappa(\mathbf{x}_i, \mathbf{x}_j) = \varphi^T(\mathbf{x}_i) \varphi(\mathbf{x}_j)$.

The optimizatoin **constraints** are

$$\sum (\lambda_i - \lambda_i^*) = 0, \\ 0 \leq \lambda_i, \lambda_i^* \leq C.$$

KKT Condition and Support Vectors

Part of the KKT condition is that $\forall i$,

$$\begin{cases} \lambda_i (y_i - \beta^T \varphi(\mathbf{x}_i) - b - \epsilon - \xi_i) = 0, \\ \lambda_i^* (\beta^T \varphi(\mathbf{x}_i) + b - y_i - \epsilon - \xi_i^*) = 0. \end{cases}$$

- ▶ Interior points: $|y_i - \beta^T \varphi(\mathbf{x}_i) - b| < \epsilon + \xi_i$.
 - ▶ Both λ_i and λ_i^* are zero.
- ▶ Boundary points: $|y_i - \beta^T \varphi(\mathbf{x}_i) - b| = \epsilon + \xi_i$.
 - ▶ One of λ_i and λ_i^* is zero.
 - ▶ $\lambda_i \neq \lambda_i^*$.

The Standard Form

Let $\gamma_i = \lambda_i$ and $\gamma_{i+n} = \lambda_i^*$ (Merge $\boldsymbol{\lambda}$ and $\boldsymbol{\lambda}^*$ into a single vector).
The dual problem becomes

$$\begin{aligned} \min_{\boldsymbol{\gamma}} \quad & \frac{1}{2} \boldsymbol{\gamma}^T \mathbf{Q} \boldsymbol{\gamma} + \mathbf{v}^T \boldsymbol{\gamma}, \\ \text{subject to} \quad & 0 \leq \gamma_i \leq C, \quad \sum_{i=1}^n \gamma_i - \sum_{i=n+1}^{2n} \gamma_i = 0. \end{aligned}$$

The **boundary points** are given by $\mathcal{I} = \{i : \gamma_i - \gamma_{i+n} \neq 0\}$.

For a new data point \mathbf{x}^{new} , the regression is

$$f(\mathbf{x}^{\text{new}}) = \sum_i (\gamma_i - \gamma_{i+n}) \kappa(\mathbf{x}_i, \mathbf{x}^{\text{new}}) + b.$$

Section 11

Gaussian Distribution

Gaussian Random Vectors

A random vector $\mathbf{X} \in \mathbb{R}^n$ is Gaussian distributed $\mathbf{X} \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ if its pdf is given by

$$p(\mathbf{x}) = |2\pi\boldsymbol{\Sigma}|^{-\frac{1}{2}} \exp\left(-\frac{1}{2}(\mathbf{x} - \boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1}(\mathbf{x} - \boldsymbol{\mu})\right),$$

where $\boldsymbol{\mu} \in \mathbb{R}^n$ and $\boldsymbol{\Sigma} \in \mathcal{S}_+^n$ (the set of $n \times n$ symmetric positive semidefinite matrices).

Here, we have assumed that $\boldsymbol{\Sigma}$ is invertible (of full rank).

Gaussian Random Vectors: Characteristic Function

PDF $\xrightleftharpoons[\text{Inverse Fourier Transform}]{\text{Fourier Transform}}$ Characteristic function $E[e^{i\langle \lambda, \mathbf{X} \rangle}]$.

$$\begin{aligned}
 f(\lambda) &= E[e^{i\langle \lambda, \mathbf{X} \rangle}] = \frac{1}{|\Sigma|^{1/2}} \int e^{i\langle \lambda, \mathbf{x} \rangle} e^{-\frac{1}{2} \mathbf{x}^T \Sigma^{-1} \mathbf{x}} d\mathbf{x} \\
 &= \frac{1}{|\Sigma|^{1/2}} \int e^{-\frac{1}{2} \mathbf{x}^T \Sigma^{-1} \mathbf{x} + i\langle \lambda, \mathbf{x} \rangle} d\mathbf{x} \\
 &= \frac{1}{|\Sigma|^{1/2}} \int e^{-\frac{1}{2} (\mathbf{x} + i\Sigma \lambda)^T \Sigma^{-1} (\mathbf{x} + i\Sigma \lambda)} e^{-\frac{1}{2} \lambda^T \Sigma \lambda} d\mathbf{x}
 \end{aligned}$$

$\mathbf{X} \sim \mathcal{N}(\mu, \Sigma)$ if

$$E[e^{i\langle \lambda, \mathbf{X} \rangle}] = \exp\left(i\langle \lambda, \mu \rangle - \frac{1}{2} \lambda^T \Sigma \lambda\right).$$

It is well defined even when Σ is not invertible.

Affine Transformation

Lemma 11.1

Let $\mathbf{X} \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$. Then for any $\mathbf{A} \in \mathbb{R}^{m \times n}$ and $\mathbf{b} \in \mathbb{R}^m$,

$$\mathbf{AX} + \mathbf{b} \sim \mathcal{N}(\mathbf{A}\boldsymbol{\mu} + \mathbf{b}, \mathbf{A}\boldsymbol{\Sigma}\mathbf{A}^T).$$

Proof:

$$\begin{aligned} \mathbb{E} \left[e^{i\langle \boldsymbol{\lambda}, \mathbf{AX} + \mathbf{b} \rangle} \right] &= \mathbb{E} \left[e^{i\langle \mathbf{A}^T \boldsymbol{\lambda}, \mathbf{X} \rangle + i\langle \boldsymbol{\lambda}, \mathbf{b} \rangle} \right] \\ &= \exp \left(i\langle \mathbf{A}^T \boldsymbol{\lambda}, \boldsymbol{\mu} \rangle - \frac{1}{2} (\mathbf{A}^T \boldsymbol{\lambda})^T \boldsymbol{\Sigma} (\mathbf{A}^T \boldsymbol{\lambda}) \right) e^{i\langle \boldsymbol{\lambda}, \mathbf{b} \rangle} \\ &= \exp \left(i\langle \boldsymbol{\lambda}, \mathbf{A}^T \boldsymbol{\mu} + \mathbf{b} \rangle - \frac{1}{2} \boldsymbol{\lambda}^T (\mathbf{A}\boldsymbol{\Sigma}\mathbf{A}^T) \boldsymbol{\lambda} \right). \end{aligned}$$

Gaussian Conditioning Lemma

Lemma 11.2

Let $\mathbf{X} \sim \mathcal{N}(\mathbf{0}, \Sigma)$.

Let \mathbf{X}_A and \mathbf{X}_B be two subvectors of \mathbf{X} , i.e., $\mathbf{X} = [\mathbf{X}_A^T, \mathbf{X}_B^T]^T$.

Let $\mathbf{K} := \Sigma^{-1} = \begin{bmatrix} \mathbf{K}_{AA} & \mathbf{K}_{AB} \\ \mathbf{K}_{BA} & \mathbf{K}_{BB} \end{bmatrix}$ be the precision matrix.

Then $\mathbf{X}_A | \mathbf{X}_B \sim P_{\mathbf{X}_A | \mathbf{X}_B} = \mathcal{N}(-\mathbf{K}_{AA}^{-1} \mathbf{K}_{AB} \mathbf{X}_B, \mathbf{K}_{AA}^{-1})$. In other words,

$$\mathbf{X}_A = -\mathbf{K}_{AA}^{-1} \mathbf{K}_{AB} \mathbf{X}_B + \epsilon,$$

where $\epsilon \sim \mathcal{N}(\mathbf{0}, \mathbf{K}_{AA}^{-1})$ is independent of \mathbf{X}_B .

Remark: $\mathbf{K}_{AA}^{-1} \neq \Sigma_{AA}$.

Matrix Identities

► Block matrix inverse (BMI)

$$\begin{bmatrix} \mathbf{A} & \mathbf{B} \\ \mathbf{C} & \mathbf{D} \end{bmatrix}^{-1} = \begin{bmatrix} (\mathbf{A} - \mathbf{B}\mathbf{D}^{-1}\mathbf{C})^{-1} & -\mathbf{A}^{-1}\mathbf{B}(\mathbf{D} - \mathbf{C}\mathbf{A}^{-1}\mathbf{B})^{-1} \\ -\mathbf{D}^{-1}\mathbf{C}(\mathbf{A} - \mathbf{B}\mathbf{D}^{-1}\mathbf{C})^{-1} & (\mathbf{D} - \mathbf{C}\mathbf{A}^{-1}\mathbf{B})^{-1} \end{bmatrix}. \quad (31)$$

► Woodbury matrix identity (WMI)

$$(\mathbf{A} + \mathbf{UCV})^{-1} = \mathbf{A}^{-1} - \mathbf{A}^{-1}\mathbf{U}(\mathbf{C}^{-1} + \mathbf{VA}^{-1}\mathbf{U})^{-1}\mathbf{VA}^{-1}. \quad (32)$$

Proof of Gaussian Conditioning Lemma

By Bayes rule, $p(\mathbf{x}_A|\mathbf{x}_B) = p(\mathbf{x}_A, \mathbf{x}_B) / p(\mathbf{x}_B)$. Then

$$\begin{aligned}\ln p(\mathbf{x}_A|\mathbf{x}_B) &= \ln p(\mathbf{x}_A, \mathbf{x}_B) - \ln p(\mathbf{x}_B) \\ &= c - \frac{1}{2} \mathbf{x}_A^T \mathbf{K}_{AA} \mathbf{x}_A - \mathbf{x}_A^T \mathbf{K}_{AB} \mathbf{x}_B - \frac{1}{2} \mathbf{x}_B^T (\mathbf{K}_{BB} - \Sigma_{BB}^{-1}) \mathbf{x}_B,\end{aligned}$$

where c is a constant. By (31),

$$\Sigma_{BB}^{-1} = \mathbf{K}_{BB} - \mathbf{K}_{BA} \mathbf{K}_{AA}^{-1} \mathbf{K}_{AB}.$$

One has

$$\ln p(\mathbf{x}_A|\mathbf{x}_B) = c - \frac{1}{2} (\mathbf{x}_A + \mathbf{K}_{AA}^{-1} \mathbf{K}_{AB} \mathbf{x}_B)^T \mathbf{K}_{AA} (\mathbf{x}_A + \mathbf{K}_{AA}^{-1} \mathbf{K}_{AB} \mathbf{x}_B).$$

That is, $\mathbf{X}_A|\mathbf{X}_B \sim \mathcal{N}(-\mathbf{K}_{AA}^{-1} \mathbf{K}_{AB} \mathbf{X}_B, \mathbf{K}_{AA}^{-1})$.

A Signal Processing Application

The problem:

Given

$$\mathbf{Y} = \mathbf{A}\mathbf{X} + \mathbf{W},$$

where $\mathbf{X} \sim \mathcal{N}(\mathbf{0}, \Sigma_x)$ and $\mathbf{W} \sim \mathcal{N}(\mathbf{0}, \Sigma_w)$. $\mathcal{P}_{X|Y}(x|y) \sim \mathcal{N}(\mu_x, \Sigma_x)$

Given observation \mathbf{y} , want to find $\hat{\mathbf{x}} = f(\mathbf{y})$ s.t. the mean squared error $E[\|\mathbf{x} - \hat{\mathbf{x}}\|_2^2]$ is minimized (MMSE solution). $\begin{pmatrix} \mathbf{x} \\ \mathbf{y} \end{pmatrix} \sim \mathcal{N} \begin{pmatrix} \mathbf{0} \\ \mathbf{0} \end{pmatrix}, \begin{bmatrix} \Sigma_x & \Sigma_x \mathbf{A}^T \\ \mathbf{A} \Sigma_x \mathbf{A}^T + \Sigma_w & \Sigma_x \mathbf{A}^T \end{bmatrix}$

Fact: The general MMSE solution is given by

$$\hat{\mathbf{x}} = E[\mathbf{X}|\mathbf{Y} = \mathbf{y}] = \int \mathbf{x} \cdot p_{\mathbf{X}|\mathbf{Y}}(\mathbf{x}|\mathbf{y}) d\mathbf{x}.$$

Hence for Gaussian random variables, Gaussian conditioning lemma can be used.

Finding the MMSE Solution

1. $\mathbf{Y} = \mathbf{A}\mathbf{X} + \mathbf{W}$ is Gaussian distributed $\mathcal{N}(\mathbf{0}, \mathbf{A}\Sigma_x\mathbf{A}^T + \Sigma_w)$.

2.

$$\begin{bmatrix} \mathbf{X} \\ \mathbf{Y} \end{bmatrix} \sim \mathcal{N}\left(\mathbf{0}, \underbrace{\begin{bmatrix} \Sigma_x & \Sigma_x\mathbf{A}^T \\ \mathbf{A}\Sigma_x & \mathbf{A}\Sigma_x\mathbf{A}^T + \Sigma_w \end{bmatrix}}_{\Sigma}\right).$$

3. Find the precision matrix from Σ : $\mathbf{K} = \Sigma^{-1}$

$$\mathbf{K} = \begin{bmatrix} \Sigma_x^{-1} + \mathbf{A}^T \Sigma_w^{-1} \mathbf{A} & -\mathbf{A}^T \Sigma_w^{-1} \\ -\Sigma_w^{-T} \mathbf{A} & \text{sth} \end{bmatrix}$$

4. $\mathbf{X}|\mathbf{Y} \sim \mathcal{N}(-\mathbf{K}_{\mathcal{A}\mathcal{A}}^{-1}\mathbf{K}_{\mathcal{A}\mathcal{B}}\mathbf{Y}, \mathbf{K}_{\mathcal{A}\mathcal{A}}^{-1})$ by Gaussian Conditioning Lemma.

We use the conditional mean as the estimate $\hat{\mathbf{x}}$:

$$\hat{\mathbf{x}} = (\Sigma_x^{-1} + \mathbf{A}^T \Sigma_w^{-1} \mathbf{A})^{-1} \mathbf{A}^T \Sigma_w^{-1} \mathbf{y}. \quad (33)$$

$$\Sigma_{X|Y} = \mathbf{K}_{\mathcal{A}\mathcal{A}}^{-1} = (\Sigma_x^{-1} + \mathbf{A}^T \Sigma_w^{-1} \mathbf{A})^{-1}. \quad (34)$$

Calculation of The \mathbf{K} Matrix

$$\begin{aligned}\mathbf{K}_{\mathcal{AA}} &\stackrel{\text{BMI(31)}}{=} \left(\boldsymbol{\Sigma}_x - \boldsymbol{\Sigma}_x \mathbf{A}^T (\mathbf{A} \boldsymbol{\Sigma}_x \mathbf{A}^T + \boldsymbol{\Sigma}_w)^{-1} \mathbf{A} \boldsymbol{\Sigma}_x \right)^{-1} \\ &\stackrel{\text{WMI(32)}}{=} \left((\boldsymbol{\Sigma}_x^{-1} + \mathbf{A}^T \boldsymbol{\Sigma}_w^{-1} \mathbf{A})^{-1} \right)^{-1} \\ &= \boldsymbol{\Sigma}_x^{-1} + \mathbf{A}^T \boldsymbol{\Sigma}_w^{-1} \mathbf{A}.\end{aligned}$$

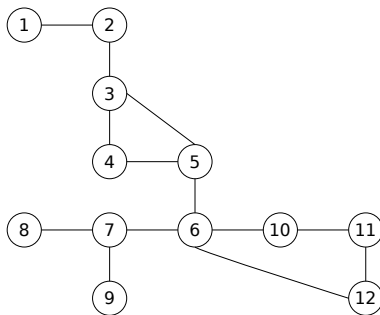
$$\begin{aligned}\mathbf{K}_{\mathcal{AB}} &\stackrel{\text{BMI(31)}}{=} -\boldsymbol{\Sigma}_x^{-1} (\boldsymbol{\Sigma}_x \mathbf{A}^T) (\mathbf{A} \boldsymbol{\Sigma}_x \mathbf{A}^T + \boldsymbol{\Sigma}_w - \mathbf{A} \boldsymbol{\Sigma}_x \boldsymbol{\Sigma}_x^{-1} \boldsymbol{\Sigma}_x \mathbf{A}^T)^{-1} \\ &= -\mathbf{A}^T \boldsymbol{\Sigma}_w^{-1}.\end{aligned}$$

Hence $\boldsymbol{\Sigma}_{X|Y} = (\boldsymbol{\Sigma}_x^{-1} + \mathbf{A}^T \boldsymbol{\Sigma}_w^{-1} \mathbf{A})^{-1}$ and $\mathbf{L} = \boldsymbol{\Sigma}_{X|Y} \mathbf{A}^T \boldsymbol{\Sigma}_w^{-1}$.

Section 12

Gaussian Graphic Model

Motivation: Gaussian Graphic Model



Encoding the **conditional dependencies** between n random variables X_1, \dots, X_n by a graph.

Correlation and Conditional Independence

*correlated
independent given middle condition*

Sneeze — Catch Cold — Weather Change

Observation: “Weather Change” and “Sneeze” are correlated.

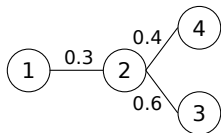
- ▶ “Weather Change” and “Catch Cold” are highly correlated.
- ▶ “Catch Cold” and “Sneeze” are highly correlated.

However, given the status of “Catch Cold”, “Weather Change” and “Sneeze” are independent.

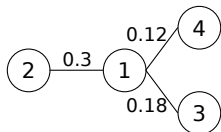
- ▶ Given that “Catch Cold” is false, “Sneeze” is likely to be false, independent of whether “Weather Change” is true or not.
- ▶ Given that “Catch Cold” is true, “Sneeze” is likely to be true, independent of whether “Weather Change” is true or not.

Other Examples

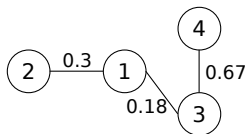
Suppose that $\rho(X_1, X_2) = 0.3$, $\rho(X_1, X_3) = 0.18$, and $\rho(X_1, X_4) = 0.12$. Suppose that on one day, $X_2 \uparrow 0.2$, $X_3 \downarrow 0.1$, and $X_4 \downarrow 0.5$. Find the expected change of X_1 .



$$E[\Delta X_1] = 0.2 \times 0.3 = 0.06.$$

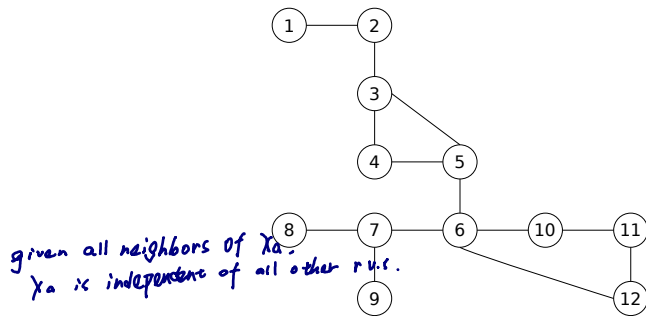


$$\begin{aligned} E[\Delta X_1] &= 0.2 \times 0.3 - 0.1 \times 0.18 - 0.5 \times 0.12 \\ &= -0.018. \end{aligned}$$



$$\begin{aligned} E[\Delta X_1] &= 0.2 \times 0.3 - 0.1 \times 0.18 \\ &= 0.042. \end{aligned}$$

Nondirected Graphical Model



The distribution of the Gaussian random vector $\mathbf{X} = [X_1, \dots, X_n]^T$ is a graphic model according to the graph g if

for all a : $X_a \perp \{X_b : b \notin \text{ne}(a), b \neq a\}$ given $\{X_c : c \in \text{ne}(a)\}$.

Or, given X_c 's, $c \in \text{ne}(a)$, X_a and X_b 's are independent for all b not in the neighborhood.

Consequence of Gaussian Conditioning

Recall the Gaussian conditioning lemma (Lemma 11.2).
Let \mathbf{K} be the precision matrix of \mathbf{X} .

Corollary 12.1

For any $a \in [n]$,

$$X_a = - \sum_{b: b \neq a} \frac{K_{ab}}{K_{aa}} X_b + \epsilon_a,$$

where $\epsilon_a \sim \mathcal{N}(0, K_{aa}^{-1})$ is independent of $\{X_b : b \neq a\}$.

Proof: Apply Lemma 11.2 with $A = \{a\}$ and $B = [n] \setminus \{a\} = A^c$.

Remark: Find the neighboring points.

Conditional Correlation

Corollary 12.2

$$\text{cor}(X_a, X_b | \mathbf{X}_C) = -\frac{K_{ab}}{\sqrt{K_{aa}K_{bb}}}.$$

Proof: From Gaussian Conditioning (Lemma 11.2), it holds that

$$\text{cov}(\mathbf{X}_{\{a,b\}} | \mathbf{X}_C) = \begin{bmatrix} K_{aa} & K_{ab} \\ K_{ba} & K_{bb} \end{bmatrix}^{-1} = \frac{1}{K_{aa}K_{bb} - K_{ab}^2} \begin{bmatrix} K_{bb} & -K_{ba} \\ -K_{ab} & K_{aa} \end{bmatrix}.$$

Plug this formula into the definition of conditional correlation. Corollary 12.2 is proved.

Remark: Find the correlation between neighboring points.

Estimate the Precision Matrix

From the definition $\mathbf{K} = \mathbf{\Sigma}^{-1}$, the computation seems straightforward. However, the commonly used fact

$$\frac{1}{m} \sum (\mathbf{X} - \bar{\mathbf{X}}) (\mathbf{X} - \bar{\mathbf{X}})^T \rightarrow \mathbf{\Sigma} \quad (35)$$

is based on the assumption that n is fixed and $m \rightarrow \infty$.

In reality, we may not have sufficient data m . Hence (35) may not be applicable.

Assumption: \mathbf{K} is sparse.

Estimation via Regression (1)

Define the matrix Θ by $\theta_{ab} = -K_{ab}/K_{bb}$ for $b \neq a$ and $\theta_{aa} = 0$. Then Corollary 12.1 implies

$$\underset{\text{expectation}}{\mathbb{E}[X_a | X_b : b \neq a]} = \sum_b \theta_{ba} X_b.$$

Hence we need to find θ_{ba} 's ($b \neq a$) to minimize

$$\mathbb{E} \left[\left(X_a - \sum_b \theta_{ba} X_b \right)^2 \right].$$

Or in matrix format

$$\hat{\Theta} = \arg \min_{\Theta \in \Theta} \mathbb{E} \left[\| \mathbf{X} - \Theta^T \mathbf{X} \|_2^2 \right],$$

where $\Theta = \{ \Theta : \text{diag}(\Theta) = \mathbf{0} \}$.

Estimation via Regression (2)

The objective function can be rewritten as

$$\begin{aligned} \overset{\text{expectation}}{\mathbb{E}} \left[\|\mathbf{X} - \Theta^T \mathbf{X}\|_2^2 \right] &\approx \overset{\text{average of training points}}{\frac{1}{m}} \sum (x - \Theta^T x)^T (x - \Theta^T x) \\ &= \frac{1}{m} \left\| \begin{bmatrix} \mathbf{x}_{(1)}^T \\ \vdots \\ \mathbf{x}_{(m)}^T \end{bmatrix} - \begin{bmatrix} \mathbf{x}_{(1)}^T \\ \vdots \\ \mathbf{x}_{(m)}^T \end{bmatrix} \Theta \right\|_F^2 \\ &= \frac{1}{m} \|\mathbf{X} - \mathbf{X}\Theta\|_F^2. \end{aligned}$$

Note that the \mathbf{X} on this slide is the data matrix and the \mathbf{X} on previous slides are random vectors.

Estimation via Regression (3)

The overall optimization problem:

$$\min_{\Theta \in \Theta} \frac{1}{m} \|\mathbf{X} - \mathbf{X}\Theta\|_F^2 + \lambda \sum_{a \neq b} |\theta_{ab}|,$$

Or

$$\min_{\Theta \in \Theta} \frac{1}{m} \|\mathbf{X} - \mathbf{X}\Theta\|_F^2 + \lambda \sum_{a < b} \sqrt{\theta_{ab}^2 + \theta_{ba}^2}.$$