Section 8
Convex Optimisation 2

Lagrangian

Consider a general optimization problem

$$\begin{aligned} &\min_{\boldsymbol{x}} \; f\left(\boldsymbol{x}\right) \\ \text{subject to } &h_i\left(\boldsymbol{x}\right) \leq 0,\ i = 1, \cdots, m, \\ &\ell_j\left(\boldsymbol{x}\right) = 0,\ j = 1, \cdots, r. \end{aligned}$$

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

Definition 8.1 (Lagrangian)

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

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

Lagrange Dual Function

Definition 8.2 (Lagrange Dual Function)

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

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

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

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

$$g(\boldsymbol{u}, \boldsymbol{v}) = \min_{\boldsymbol{x} \in \mathbb{R}^n} L(\boldsymbol{x}, \boldsymbol{u}, \boldsymbol{v}) \leq \min_{\boldsymbol{x} \in \mathcal{X}} L(\boldsymbol{x}, \boldsymbol{u}, \boldsymbol{v}) \leq f(\boldsymbol{x}). \tag{13}$$

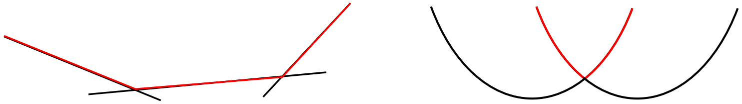
Concavity of Lagrange Dual Function

Lemma 8.3

The Lagrange dual function
$$g(\boldsymbol{u},\boldsymbol{v}) = \min_{\boldsymbol{x}\in\mathbb{R}^n} L(\boldsymbol{x},\boldsymbol{u},\boldsymbol{v}) = \min_{\boldsymbol{x}\in\mathbb{R}^n} f(\boldsymbol{x}) + \sum_{i=1}^m u_i h_i(\boldsymbol{x}) + \sum_{j=1}^r v_j \ell_j(\boldsymbol{x})$$
is concave in $(\boldsymbol{u},\boldsymbol{v})$.

Lemma 8.4

- *Let* $f_\alpha(\boldsymbol{x})$ *be concave functions. Then* $g(\boldsymbol{x}) = \inf_\alpha \; f_\alpha(\boldsymbol{x})$ *is concave.*
- *Let* $f_\alpha(\boldsymbol{x})$ *be convex functions. Then* $g(\boldsymbol{x}) = \sup_\alpha \; f_\alpha(\boldsymbol{x})$ *is convex.*



Proofs

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

$$\begin{aligned} g\left(\lambda \boldsymbol{x} + (1-\lambda)\, \boldsymbol{y}\right) &= \inf_{\alpha}\; f_{\alpha}\left(\lambda \boldsymbol{x} + (1-\lambda)\, \boldsymbol{y}\right) \\ &\geq \inf_{\alpha}\; \lambda f_{\alpha}(\boldsymbol{x}) + (1-\lambda)\, f_{\alpha}(\boldsymbol{y}) \\ &\geq \lambda \inf_{\alpha}\; f_{\alpha}(\boldsymbol{x}) + (1-\lambda) \inf_{\alpha}\; f_{\alpha}(\boldsymbol{y}) . \end{aligned}$$

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

Lagrange Dual Problem

Given the primal problem

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

Its Lagrange dual problem is

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

Weak and Strong Duality

Weak duality: the dual optimal value $g^\star$ satisfies

$$f^\star \geq g^\star.$$

This is a direct consequence of (13).
Strong duality is referred to as the case that

$$f^\star = g^\star.$$

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

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

then strong duality holds. (Proof is omitted.)

Karush-Kuhn-Tucker conditions

Given the optimization problem

$$\begin{aligned} &\text{minimize } f(\boldsymbol{x}) \\ &\text{subject to } h_i(\boldsymbol{x}) \leq 0,\ i = 1, \cdots, m, \\ &\qquad\qquad \ell_j(\boldsymbol{x}) = 0,\ i = 1, \cdots, r. \end{aligned}$$

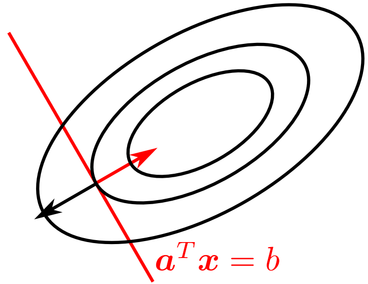
The Karush-Kuhn-Tucker (KKT) conditions are:

- $\mathbf{0} \in \partial f(\boldsymbol{x}) + \sum_{i=1}^m u_i \partial h_i(\boldsymbol{x}) + \sum_{j=1}^r v_j \partial \ell_j(\boldsymbol{x})$. (stationarity)
- $u_i h_i(\boldsymbol{x}) = 0$, $\forall i$. (complementary slackness)
- $h_i(\boldsymbol{x}) \leq 0$, $\ell_j(\boldsymbol{x}) = 0$, $\forall i$, $\forall j$. (primal feasibility)
- $u_i \geq 0$, $\forall i$. (dual feasibility)

KKT conditions are

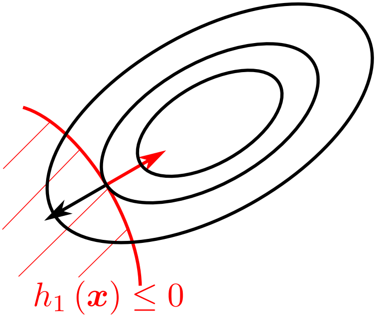
- Always sufficient.
- Necessary under strong duality.

Geometric Intuition: Equality Constraints



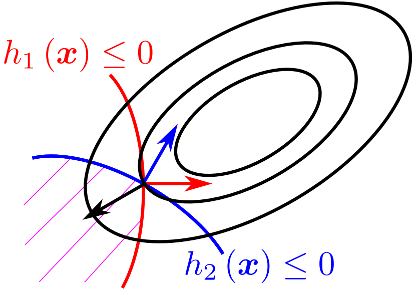
$\partial f(\boldsymbol{x})$ is a linear combination of $\partial \ell_j(\boldsymbol{x})$'s.

Geometric Intuition: One Inequality Constraint



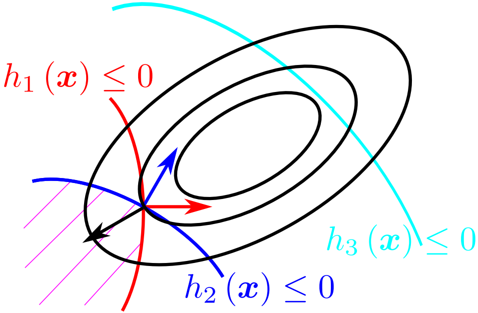
$$\partial f(\boldsymbol{x}) + u_1 \partial h_1(\boldsymbol{x}) = \mathbf{0}.$$
$$h_1(\boldsymbol{x}) = 0.$$

Geometric Intuition: Inequality Constraints



$$\partial f(\boldsymbol{x}) + \sum_{i=1}^{2} u_i \partial h_i(\boldsymbol{x}) = \mathbf{0}.$$
$$h_1(\boldsymbol{x}) = 0,\ h_2(\boldsymbol{x}) = 0.$$

Geometric Intuition: Inequality Constraints



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

Sufficiency

If $\boldsymbol{x}^\star, \boldsymbol{u}^\star, \boldsymbol{v}^\star$ satisfy the KKT conditions, then $\boldsymbol{x}^\star$ and $\boldsymbol{u}^\star, \boldsymbol{v}^\star$ are primal and dual solutions.

If $\boldsymbol{x}^\star, \boldsymbol{u}^\star, \boldsymbol{v}^\star$ satisfy the KKT conditions, then

$$\begin{aligned} g\left(\boldsymbol{u}^\star, \boldsymbol{v}^\star\right) &= f\left(\boldsymbol{x}^\star\right) + \sum_{i=1}^{m} u_i^\star h_i\left(\boldsymbol{x}^\star\right) + \sum_{j=1}^{r} v_i^\star \ell_i\left(\boldsymbol{x}^\star\right) \\ &= f\left(\boldsymbol{x}^\star\right), \end{aligned}$$

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

Necessity

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

Due to the strong duality, one has

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

In other words, all the inequalities are actually equalities.

Quadratic Programming with Equality Constraints

Let $\boldsymbol{Q} \succeq 0$.

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

By KKT conditions, $\boldsymbol{x}$ is the minimizer if and only if

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

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

The optimal $\boldsymbol{x}^\star$ can be obtained by solving the linear inverse problem.

Water Filling

$$\min_{\boldsymbol{x}} \; -\sum_{i=1}^{n} \log\left(\alpha_i + x_i\right) \text{ subject to } \boldsymbol{x} \geq \mathbf{0},\; \mathbf{1}^T \boldsymbol{x} = 1.$$

By KKT conditions,

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

Eliminate $\boldsymbol{u}$. The first two conditions become

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

Therefore, the solution:

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

where v is chosen such that

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

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

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

Its Lagrangian is

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

The dual function

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

The dual problem

$$\max \; g(\boldsymbol{v})$$

The Dual Ascent Method (2)

.

Dual ascent method

$$\boldsymbol{x}^{k+1} = \underset{\boldsymbol{x}}{\operatorname{argmin}}\, L\left(\boldsymbol{x}, \boldsymbol{v}^k\right)$$

$$\boldsymbol{v}^{k+1} = \boldsymbol{v}^k + \alpha^k \nabla_{\boldsymbol{v}} L\left(\boldsymbol{x}^{k+1}, \boldsymbol{v}\right) = \boldsymbol{v}^k + \alpha^k \left(\boldsymbol{A}\boldsymbol{x}^{k+1} - \boldsymbol{b}\right)$$

With appropriate chosen α^k, $g\left(\boldsymbol{v}^{k+1}\right) > g\left(\boldsymbol{v}^k\right)$ 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:

$$\min f(\boldsymbol{x}) + \frac{\rho}{2} \|\boldsymbol{A}\boldsymbol{x} - \boldsymbol{b}\|_2^2$$
$$\text{subject to } \boldsymbol{A}\boldsymbol{x} = \boldsymbol{b}$$

Lagrangian:

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

Method of multipliers:

$$\boldsymbol{x}^{k+1} = \underset{\boldsymbol{x}}{\operatorname{argmin}}\, L_\rho\left(\boldsymbol{x}, \boldsymbol{v}^k\right)$$
$$\boldsymbol{v}^{k+1} = \boldsymbol{v}^k + \rho\left(\boldsymbol{A}\boldsymbol{x}^{k+1} - \boldsymbol{b}\right)$$

Note the fixed step size ρ.

The Method of Multipliers: Step Size ρ

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

$$\boldsymbol{A}\boldsymbol{x}^* - \boldsymbol{b} = \boldsymbol{0}, \quad \nabla f\left(\boldsymbol{x}^*\right) + \boldsymbol{A}^T \boldsymbol{v}^* = \boldsymbol{0}.$$

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

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

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

ADMM (1)

ADMM solves problems in the form

$$\min f(\boldsymbol{x}) + g(\boldsymbol{z})$$
$$\text{subject to } \boldsymbol{A}\boldsymbol{x} + \boldsymbol{B}\boldsymbol{z} = \boldsymbol{c}$$

The optimal value of this problem is denoted by

$$p^\star = \inf \left\{ f(\boldsymbol{x}) + g(\boldsymbol{z}) \,|\, \boldsymbol{A}\boldsymbol{x} + \boldsymbol{B}\boldsymbol{z} = \boldsymbol{c} \right\}.$$

The augmented Lagrangian:

$$L_\rho(\boldsymbol{x}, \boldsymbol{z}, \boldsymbol{v}) = f(\boldsymbol{x}) + g(\boldsymbol{z}) + \boldsymbol{v}^T (\boldsymbol{A}\boldsymbol{x} + \boldsymbol{B}\boldsymbol{z} - \boldsymbol{c})$$
$$+ \frac{\rho}{2} \|\boldsymbol{A}\boldsymbol{x} + \boldsymbol{B}\boldsymbol{z} - \boldsymbol{c}\|_2^2$$

where $\rho > 0$.

ADMM (2)

ADMM is an iterative algorithm with iterations:

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

ADMM is different from the method of multipliers which has iterations

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

Scaled Form

Define the *primal residual*

$$\boldsymbol{r} = \boldsymbol{A}\boldsymbol{x} + \boldsymbol{B}\boldsymbol{z} - \boldsymbol{c}.$$

Define the *scaled dual variable* $\boldsymbol{u} = (1/\rho)\,\boldsymbol{v}$, then

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

The *scaled form of ADMM*:

$$\begin{aligned} \boldsymbol{x}^{k+1} &= \operatorname*{argmin}_{\boldsymbol{x}}\ f\left(\boldsymbol{x}\right) + \frac{\rho}{2}\left\|\boldsymbol{A}\boldsymbol{x} + \boldsymbol{B}\boldsymbol{z}^k - \boldsymbol{c} + \boldsymbol{u}^k\right\|_2^2 \\ \boldsymbol{z}^{k+1} &= \operatorname*{argmin}_{\boldsymbol{z}}\ g\left(\boldsymbol{z}\right) + \frac{\rho}{2}\left\|\boldsymbol{A}\boldsymbol{x}^{k+1} + \boldsymbol{B}\boldsymbol{z} - \boldsymbol{c} + \boldsymbol{u}^k\right\|_2^2 \\ \boldsymbol{u}^{k+1} &= \boldsymbol{u}^k + \boldsymbol{A}\boldsymbol{x}^{k+1} + \boldsymbol{B}\boldsymbol{z}^{k+1} - \boldsymbol{c}. \end{aligned}$$

Example 1

Lasso problem:

$$\min \ \frac{1}{2} \left\| \boldsymbol{y} - \boldsymbol{A}\boldsymbol{x} \right\|_2^2 + \lambda \left\| \boldsymbol{x} \right\|_1 .$$

ADMM version:

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

ADMM iterations:

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

Example 2

Constrained Lasso problem:

$$\min \frac{1}{2} \|\boldsymbol{y} - \boldsymbol{A}\boldsymbol{x}\|_2^2 + \lambda \|\boldsymbol{x}\|_1$$
$$\text{subject to } \boldsymbol{B}\boldsymbol{x} \leq \boldsymbol{c}$$

ADMM version:

$$\min \frac{1}{2} \|\boldsymbol{y} - \boldsymbol{A}\boldsymbol{x}\|_2^2 + g(\boldsymbol{z})$$
$$\text{subject to } \begin{bmatrix} \boldsymbol{I} \\ \boldsymbol{B} \end{bmatrix} \boldsymbol{x} + \begin{bmatrix} -\boldsymbol{I} & \\ & \boldsymbol{I} \end{bmatrix} \begin{bmatrix} \boldsymbol{z}_1 \\ \boldsymbol{z}_2 \end{bmatrix} = \begin{bmatrix} \boldsymbol{0} \\ \boldsymbol{c} \end{bmatrix},$$

where

$$g(\boldsymbol{z}) = \lambda \|\boldsymbol{z}_1\|_1 + \mathbb{1}_{\geq 0}(\boldsymbol{z}_2)$$
$$\mathbb{1}_{\geq 0}(z) = \begin{cases} \infty & \text{if } z < 0 \\ 0 & \text{if } z \geq 0 \end{cases}.$$

Example 2 - Continued

ADMM Iterations:

$$
\begin{aligned}
\boldsymbol{x}^{k+1} &= \underset{\boldsymbol{x}}{\operatorname{argmin}} \frac{1}{2}\|\boldsymbol{y}-\boldsymbol{A}\boldsymbol{x}\|_2^2 + \frac{\rho}{2}\left\|\boldsymbol{B}'\boldsymbol{x}+\boldsymbol{D}'\boldsymbol{z}^k-\boldsymbol{c}'+\boldsymbol{u}^k\right\|_2^2 \\
\boldsymbol{z}^{k+1} &= \underset{\boldsymbol{z}_1,\boldsymbol{z}_2}{\operatorname{argmin}} \lambda\|\boldsymbol{z}\|_1 + \frac{\rho}{2}\left\|\boldsymbol{x}^{k+1}-\boldsymbol{z}_1+\boldsymbol{u}_1^k\right\|_2^2 \\
&\qquad\qquad \mathbb{1}_{\geq 0}(\boldsymbol{z}_2) + \frac{\rho}{2}\left\|\boldsymbol{B}\boldsymbol{x}^{k+1}+\boldsymbol{z}_2-\boldsymbol{c}+\boldsymbol{u}_2^k\right\|_2^2 \\
\boldsymbol{u}^{k+1} &= \boldsymbol{u}^k+\boldsymbol{B}'\boldsymbol{x}^{k+1}+\boldsymbol{D}'\boldsymbol{z}^{k+1}-\boldsymbol{c}'.
\end{aligned}
$$

Each step is easy to compute.

Optimality Conditions (1)

The necessary and sufficient conditions for optimality are primal feasibility

$$\boldsymbol{A}\boldsymbol{x}^\star + \boldsymbol{B}\boldsymbol{z}^\star - \boldsymbol{c} = \boldsymbol{0}. \tag{15}$$

and dual feasibility

$$\boldsymbol{0} \in \partial f\left(\boldsymbol{x}^\star\right) + \boldsymbol{A}^T\boldsymbol{v}^\star \tag{16}$$

$$\boldsymbol{0} \in \partial g\left(\boldsymbol{z}^\star\right) + \boldsymbol{B}^T\boldsymbol{v}^\star. \tag{17}$$

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

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

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

Optimality Conditions (2)

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

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

Or equivalently

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

The *dual residual* is defined as

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

Convergence of ADMM

Under mild conditions, ADMM converges:

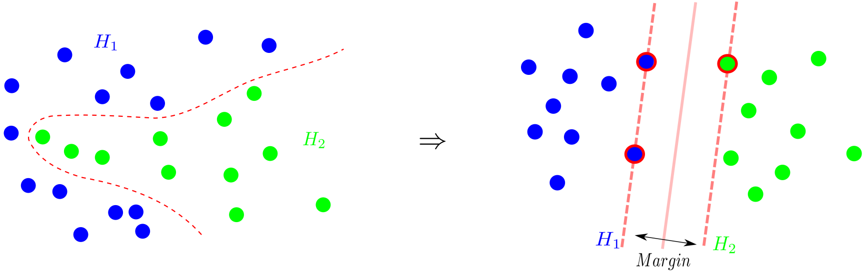
- Primal residual convergence: $\boldsymbol{r}^k \to \mathbf{0}$ as $k \to \infty$.
- Objective convergence: $f\left(\boldsymbol{x}^k\right) + g\left(\boldsymbol{z}^k\right) \to \boldsymbol{p}^\star$ as $k \to \infty$.
- Dual variable convergence: $\boldsymbol{v}^k \to \boldsymbol{v}^\star$ as $k \to \infty$.

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

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

Section 10
Support Vector Machine

Idea of SVM

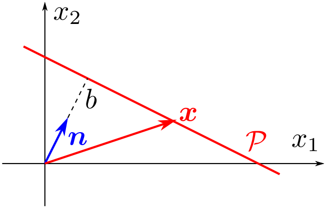


A Hyperplane

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

$$\mathcal{P} = \left\{ \boldsymbol{x} : \quad \boldsymbol{n}^T \boldsymbol{x} = b \right\}.$$

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



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

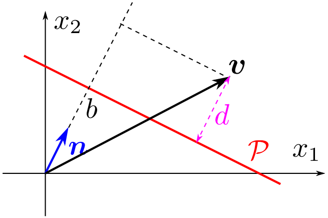
- If $\|\boldsymbol{n}\|_2 \neq 1$, then

$$\mathcal{P} = \left\{ \boldsymbol{x} : \quad \boldsymbol{n}^T \boldsymbol{x} = b \right\} = \left\{ \boldsymbol{x} : \quad \boldsymbol{n}^T \boldsymbol{x} / \|\boldsymbol{n}\|_2 = b / \|\boldsymbol{n}\|_2 \right\}.$$

Distance to a Hyperplane

Define a hyperplane $\mathcal{P} = \left\{ \boldsymbol{x} : \quad \boldsymbol{n}^T \boldsymbol{x} = b \right\}$ where $\|\boldsymbol{n}\|_2 = 1$.
Let $\boldsymbol{v}$ be an arbitrary point.
The distance between $\boldsymbol{v}$ and $\mathcal{P}$ is given by

$$d = d\left(\boldsymbol{v}, \mathcal{P}\right) = \left| \boldsymbol{n}^T \boldsymbol{v} - b \right|. \tag{18}$$



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

$$d = \left| \frac{\boldsymbol{n}^T}{\|\boldsymbol{n}\|_2} \boldsymbol{v} - \frac{b}{\|\boldsymbol{n}\|_2} \right| = \frac{\left| \boldsymbol{n}^T \boldsymbol{v} - b \right|}{\|\boldsymbol{n}\|_2}. \tag{19}$$

SVM: Separate Points from Two Different Classes

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

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

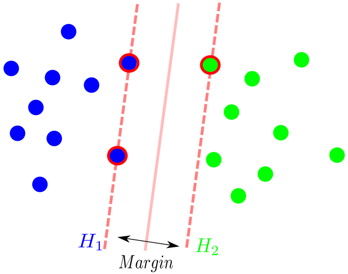
or equivalently

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

In other words, find a hyperplane $\left\{ \boldsymbol{x} : \; \boldsymbol{\beta}^T \boldsymbol{x} - b \right\}$ s.t.

- Distance from one class to the hyperplane is $1 / \left\| \boldsymbol{\beta} \right\|_2$.
- Distance between the two classes (along direction $\boldsymbol{\beta}$) is $2 / \left\| \boldsymbol{\beta} \right\|_2$.

SVM: Best Separation



SVM: a convex optimization problem:

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

Lagrange Dual Problem of SVM

Lagrangian of the SVM primal optimization problem:

$$L = \frac{1}{2} \left\| \boldsymbol{\beta} \right\|^2 + \sum_i \lambda_i \left(1 - y_i \left(\boldsymbol{\beta}^T \boldsymbol{x}_i + b \right) \right), \tag{21}$$

where $\lambda_i \geq 0$.

Lagrange Dual Problem

$$\max_{\boldsymbol{\lambda}} \quad \underbrace{\min_{\boldsymbol{\beta},b} \; L}_{\text{Lagrange dual function } L_D}$$

The Dual Function

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

$$\frac{\partial L}{\partial \boldsymbol{\beta}} = \boldsymbol{\beta} - \sum_i \lambda_i y_i \boldsymbol{x}_i = 0 \;\Rightarrow\; \boldsymbol{\beta} = \sum_i \lambda_i y_i \boldsymbol{x}_i. \tag{22}$$

$$\frac{\partial L}{\partial b} = \sum_i \lambda_i y_i = 0. \tag{23}$$

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

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

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

The Dual Problem

The dual problem becomes:

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

The KKT Condition

$$
\begin{aligned}
\frac{\partial L}{\partial \boldsymbol{\beta}} &= \boldsymbol{\beta} - \sum_i \lambda_i y_i \boldsymbol{x}_i = 0, && (26) \\
\frac{\partial L}{\partial b} &= \sum_i \lambda_i y_i = 0, && (27) \\
1 - y_i \left(\boldsymbol{\beta}^T \boldsymbol{x}_i + b \right) &\leq 0, && (28) \\
\lambda_i &\geq 0, && (29) \\
\lambda_i \left(1 - y_i \left(\boldsymbol{\beta}^T \boldsymbol{x}_i + b \right) \right) &= 0. && (30)
\end{aligned}
$$

SVM Classifier: Support Vectors

Condition (30) implies

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

Hence from (26),

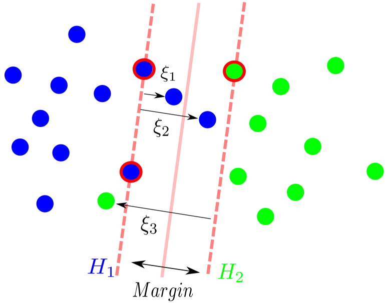
$$\boldsymbol{\beta} = \sum_{i \in \mathcal{I}} \lambda_i y_i \boldsymbol{x}_i, \quad \mathcal{I} = \left\{ i : \ y_i \left(\boldsymbol{\beta}^T \boldsymbol{x}_i + b\right) = 1 \quad (\text{or } \lambda_i \neq 0) \right\}.$$

For a new test data $\boldsymbol{x}^{\text{new}}$,

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

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

SVM for Overlapping Classes



Primal Problem for Overlapping Classes

The constraints:

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

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

SVM Primal Problem:

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

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

Dual Function

The Lagrangian

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

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

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

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

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

The Dual Problem

The dual problem:

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

where $K_{i,j} = y_i \boldsymbol{x}_i^T \boldsymbol{x}_j y_j$.
The only difference is that now λ_i's are upper bounded by C.

Again, only boundary points are involved.

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

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

The General Case

- Two classes $\Rightarrow$ multiple classes
  - Regression
- Data space $\Rightarrow$ feature space
  Define a kernel function $\varphi : \mathbb{R}^n \to \mathcal{H}$ and work on the space of $\varphi(\boldsymbol{x}_i)$.

In SVM, what really matters is $\boldsymbol{x}_i^T \boldsymbol{x}_j$.
In the general case (kernel method), what matters is

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

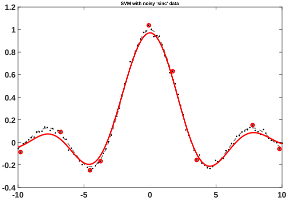
Example of nonlinear features:

- $\kappa(\boldsymbol{x}, \boldsymbol{y}) = \left(\boldsymbol{x}^T \boldsymbol{y}\right)^2 = (x_1 y_1 + x_2 y_2)^2 = x_1^2 y_2^2 + 2 x_1 y_1 x_2 y_2 + x_2^2 y_2^2$.
  - $\kappa(\boldsymbol{x}, \boldsymbol{y}) = \varphi^T(\boldsymbol{x})\, \varphi(\boldsymbol{y})$ with $\varphi(\boldsymbol{x}) = \left[x_1^2, \sqrt{2} x_1 x_2, x_2^2\right]^T$.
- $\kappa(\boldsymbol{x}, \boldsymbol{y}) = \exp\left(- \|\boldsymbol{x} - \boldsymbol{y}\|_2^2 / 2\sigma^2\right)$.
  - $\varphi(\boldsymbol{x})$ has infinite dimension.

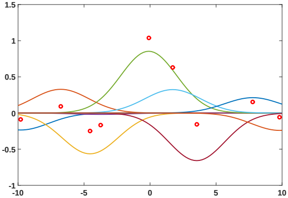
SVM for Regression

Regression problem: find $\boldsymbol{\beta}$ and b s.t.

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



=



The Primal Optimization Problem

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

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

The constraints are equivalent to

$$\begin{aligned} y_i - \boldsymbol{\beta}^T \varphi\left(\boldsymbol{x}_i\right) - b \le \epsilon, \\ \boldsymbol{\beta}^T \varphi\left(\boldsymbol{x}_i\right) + b - y_i \le \epsilon. \end{aligned}$$

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

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

Lagrangian

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

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

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

The Dual Problem

The objective function of the dual problem

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

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

The optimizatoin constraints are

$$\sum \left(\lambda_i - \lambda_i^\star\right) = 0, \\ 0 \leq \lambda_i, \lambda_i^\star \leq C.$$

KKT Condition and Support Vectors

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

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

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

The Standard Form

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

$$\min_{\boldsymbol{\gamma}} \quad \frac{1}{2}\boldsymbol{\gamma}^T \boldsymbol{Q} \boldsymbol{\gamma} + \boldsymbol{v}^T \boldsymbol{\gamma},$$

$$\text{subject to} \quad 0 \leq \gamma_i \leq C, \quad \sum_{i=1}^{n} \gamma_i - \sum_{i=n+1}^{2n} \gamma_i = 0.$$

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

For a new data point $\boldsymbol{x}^{\text{new}}$, the regression is

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

Section 11
Gaussian Distribution

Gaussian Random Vectors

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

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

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

$$\text{PDF} \underset{\text{Inverse Fourier Transform}}{\overset{\text{Fourier Transform}}{\rightleftharpoons}} \text{Characteristic function } \mathrm{E}\left[e^{i\langle \boldsymbol{\lambda}, \boldsymbol{X}\rangle}\right].$$

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

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

It is well defined even when $\boldsymbol{\Sigma}$ is not invertible.

Affine Transformation

Lemma 11.1

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

$$\boldsymbol{A}\boldsymbol{X} + \boldsymbol{b} \sim \mathcal{N}\left(\boldsymbol{A}\boldsymbol{\mu} + \boldsymbol{b}, \boldsymbol{A}\boldsymbol{\Sigma}\boldsymbol{A}^T\right).$$

Proof:

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

Gaussian Conditioning Lemma

Lemma 11.2

Let $\boldsymbol{X} \sim \mathcal{N}(\mathbf{0}, \boldsymbol{\Sigma})$.
Let $\boldsymbol{X}_{\mathcal{A}}$ *and* $\boldsymbol{X}_{\mathcal{B}}$ *be two subvectors of* $\boldsymbol{X}$, *i.e.,* $\boldsymbol{X} = \left[\boldsymbol{X}_{\mathcal{A}}^T, \boldsymbol{X}_{\mathcal{B}}^T\right]^T$.
Let $\boldsymbol{K} := \boldsymbol{\Sigma}^{-1} = \begin{bmatrix} \boldsymbol{K}_{\mathcal{A}\mathcal{A}} & \boldsymbol{K}_{\mathcal{A}\mathcal{B}} \\ \boldsymbol{K}_{\mathcal{B}\mathcal{A}} & \boldsymbol{K}_{\mathcal{B}\mathcal{B}} \end{bmatrix}$ *be the precision matrix.*
Then $\boldsymbol{X}_{\mathcal{A}} | \boldsymbol{X}_{\mathcal{B}} \sim P_{\boldsymbol{X}_{\mathcal{A}}|\boldsymbol{X}_{\mathcal{B}}} = \mathcal{N}\left(-\boldsymbol{K}_{\mathcal{A}\mathcal{A}}^{-1}\boldsymbol{K}_{\mathcal{A}\mathcal{B}}\boldsymbol{X}_{\mathcal{B}}, \boldsymbol{K}_{\mathcal{A}\mathcal{A}}^{-1}\right)$. *In other words,*

$$\boldsymbol{X}_{\mathcal{A}} = -\boldsymbol{K}_{\mathcal{A}\mathcal{A}}^{-1}\boldsymbol{K}_{\mathcal{A}\mathcal{B}}\boldsymbol{X}_{\mathcal{B}} + \boldsymbol{\epsilon},$$

where $\boldsymbol{\epsilon} \sim \mathcal{N}\left(\mathbf{0}, \boldsymbol{K}_{\mathcal{A}\mathcal{A}}^{-1}\right)$ *is independent of* $\boldsymbol{X}_{\mathcal{B}}$.

Remark: $\boldsymbol{K}_{\mathcal{A}\mathcal{A}}^{-1} \neq \boldsymbol{\Sigma}_{\mathcal{A}\mathcal{A}}$.

Matrix Identities

- Block matrix inverse (BMI)

$$\begin{bmatrix} \boldsymbol{A} & \boldsymbol{B} \\ \boldsymbol{C} & \boldsymbol{D} \end{bmatrix}^{-1} =$$
$$\begin{bmatrix} \left(\boldsymbol{A}-\boldsymbol{B}\boldsymbol{D}^{-1}\boldsymbol{C}\right)^{-1} & -\boldsymbol{A}^{-1}\boldsymbol{B}\left(\boldsymbol{D}-\boldsymbol{C}\boldsymbol{A}^{-1}\boldsymbol{B}\right)^{-1} \\ -\boldsymbol{D}^{-1}\boldsymbol{C}\left(\boldsymbol{A}-\boldsymbol{B}\boldsymbol{D}^{-1}\boldsymbol{C}\right)^{-1} & \left(\boldsymbol{D}-\boldsymbol{C}\boldsymbol{A}^{-1}\boldsymbol{B}\right)^{-1} \end{bmatrix}. \tag{31}$$

- Woodbury matrix identity (WMI)

$$\left(\boldsymbol{A}+\boldsymbol{U}\boldsymbol{C}\boldsymbol{V}\right)^{-1} = \boldsymbol{A}^{-1} - \boldsymbol{A}^{-1}\boldsymbol{U}\left(\boldsymbol{C}^{-1}+\boldsymbol{V}\boldsymbol{A}^{-1}\boldsymbol{U}\right)^{-1}\boldsymbol{V}\boldsymbol{A}^{-1}. \tag{32}$$

Proof of Gaussian Conditioning Lemma

By Bayes rule, $p\left(\boldsymbol{x}_{\mathcal{A}}|\boldsymbol{x}_{\mathcal{B}}\right)=p\left(\boldsymbol{x}_{\mathcal{A}},\boldsymbol{x}_{\mathcal{B}}\right)/p\left(\boldsymbol{x}_{\mathcal{B}}\right)$. Then

$$\begin{aligned}\ln p\left(\boldsymbol{x}_{\mathcal{A}}|\boldsymbol{x}_{\mathcal{B}}\right)&=\ln p\left(\boldsymbol{x}_{\mathcal{A}},\boldsymbol{x}_{\mathcal{B}}\right)-\ln p\left(\boldsymbol{x}_{\mathcal{B}}\right)\\&=c-\frac{1}{2}\boldsymbol{x}_{\mathcal{A}}^{T}\boldsymbol{K}_{\mathcal{A}\mathcal{A}}\boldsymbol{x}_{\mathcal{A}}-\boldsymbol{x}_{\mathcal{A}}^{T}\boldsymbol{K}_{\mathcal{A}\mathcal{B}}\boldsymbol{x}_{\mathcal{B}}-\frac{1}{2}\boldsymbol{x}_{\mathcal{B}}^{T}\left(\boldsymbol{K}_{\mathcal{B}\mathcal{B}}-\boldsymbol{\Sigma}_{\mathcal{B}\mathcal{B}}^{-1}\right)\boldsymbol{x}_{\mathcal{B}},\end{aligned}$$

where c is a constant. By (31),

$$\boldsymbol{\Sigma}_{\mathcal{B}\mathcal{B}}^{-1}=\boldsymbol{K}_{\mathcal{B}B}-\boldsymbol{K}_{\mathcal{B}\mathcal{A}}\boldsymbol{K}_{\mathcal{A}\mathcal{A}}^{-1}\boldsymbol{K}_{\mathcal{A}\mathcal{B}}.$$

One has

$$\ln p\left(\boldsymbol{x}_{\mathcal{A}}|\boldsymbol{x}_{\mathcal{B}}\right)=c-\frac{1}{2}\left(\boldsymbol{x}_{\mathcal{A}}+\boldsymbol{K}_{\mathcal{A}\mathcal{A}}^{-1}\boldsymbol{K}_{\mathcal{A}\mathcal{B}}\boldsymbol{x}_{\mathcal{B}}\right)^{T}\boldsymbol{K}_{\mathcal{A}\mathcal{A}}\left(\boldsymbol{x}_{\mathcal{A}}+\boldsymbol{K}_{\mathcal{A}\mathcal{A}}^{-1}\boldsymbol{K}_{\mathcal{A}\mathcal{B}}\boldsymbol{x}_{\mathcal{B}}\right).$$

That is, $\boldsymbol{X}_{\mathcal{A}}|\boldsymbol{X}_{\mathcal{B}}\sim\mathcal{N}\left(-\boldsymbol{K}_{\mathcal{A}\mathcal{A}}^{-1}\boldsymbol{K}_{\mathcal{A}\mathcal{B}}\boldsymbol{X}_{\mathcal{B}},\boldsymbol{K}_{\mathcal{A}\mathcal{A}}^{-1}\right)$.

A Signal Processing Application

The problem:
Given

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

where $\boldsymbol{X} \sim \mathcal{N}(\mathbf{0}, \boldsymbol{\Sigma}_x)$ and $\boldsymbol{W} \sim \mathcal{N}(\mathbf{0}, \boldsymbol{\Sigma}_w)$.

Given observation $\boldsymbol{y}$, want to find $\hat{\boldsymbol{x}} = f(\boldsymbol{y})$ s.t. the mean squared error $\mathrm{E}\left[\|\boldsymbol{x} - \hat{\boldsymbol{x}}\|_2^2\right]$ is minimized (MMSE solution).

Fact: The general MMSE solution is given by

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

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

Finding the MMSE Solution

1. $\boldsymbol{Y} = \boldsymbol{A}\boldsymbol{X} + \boldsymbol{W}$ is Gaussian distributed $\mathcal{N}\left(\mathbf{0}, \boldsymbol{A}\boldsymbol{\Sigma}_x\boldsymbol{A}^T + \boldsymbol{\Sigma}_w\right)$.
2. $$\left[\begin{array}{c} \boldsymbol{X} \\ \boldsymbol{Y} \end{array}\right] \sim \mathcal{N}\left(\mathbf{0}, \underbrace{\left[\begin{array}{cc} \boldsymbol{\Sigma}_x & \boldsymbol{\Sigma}_x\boldsymbol{A}^T \\ \boldsymbol{A}\boldsymbol{\Sigma}_x & \boldsymbol{A}\boldsymbol{\Sigma}_x\boldsymbol{A}^T + \boldsymbol{\Sigma}_w \end{array}\right]}_{\boldsymbol{\Sigma}}\right).$$
3. Find the precision matrix from $\boldsymbol{\Sigma}$:
$$\boldsymbol{K} = \left[\begin{array}{cc} \boldsymbol{\Sigma}_x^{-1} + \boldsymbol{A}^T\boldsymbol{\Sigma}_w^{-1}\boldsymbol{A} & -\boldsymbol{A}^T\boldsymbol{\Sigma}_w^{-1} \\ -\boldsymbol{\Sigma}_w^{-T}\boldsymbol{A} & \text{sth} \end{array}\right]$$
4. $\boldsymbol{X}|\boldsymbol{Y} \sim \mathcal{N}\left(-\boldsymbol{K}_{\mathcal{A}\mathcal{A}}^{-1}\boldsymbol{K}_{\mathcal{A}\mathcal{B}}\boldsymbol{Y}, \boldsymbol{K}_{\mathcal{A}\mathcal{A}}^{-1}\right)$ by Gaussian Conditioning Lemma.

   We use the conditional mean as the estimate $\hat{\boldsymbol{x}}$:
$$\hat{\boldsymbol{x}} = \left(\boldsymbol{\Sigma}_x^{-1} + \boldsymbol{A}^T\boldsymbol{\Sigma}_w^{-1}\boldsymbol{A}\right)^{-1}\boldsymbol{A}^T\boldsymbol{\Sigma}_w^{-1}\boldsymbol{y}. \tag{33}$$
$$\boldsymbol{\Sigma}_{X|Y} = \boldsymbol{K}_{\mathcal{A}\mathcal{A}}^{-1} = \left(\boldsymbol{\Sigma}_x^{-1} + \boldsymbol{A}^T\boldsymbol{\Sigma}_w^{-1}\boldsymbol{A}\right)^{-1}. \tag{34}$$

Calculation of The $\boldsymbol{K}$ Matrix

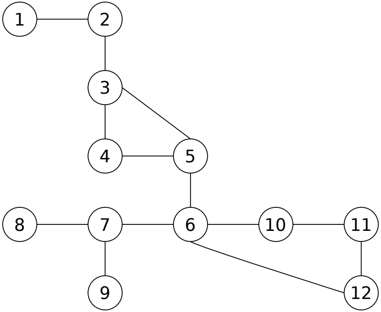
$$
\begin{aligned}
\boldsymbol{K}_{\mathcal{A}\mathcal{A}} &\overset{\text{BMI(31)}}{=} \left(\boldsymbol{\Sigma}_x - \boldsymbol{\Sigma}_x \boldsymbol{A}^T \left(\boldsymbol{A}\boldsymbol{\Sigma}_x\boldsymbol{A}^T + \boldsymbol{\Sigma}_w\right)^{-1} \boldsymbol{A}\boldsymbol{\Sigma}_x\right)^{-1} \\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

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

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

Section 12
Gaussian Graphic Model

Motivation: Gaussian Graphic Model



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

Correlation and Conditional Independence

Sneeze — Catch Cold — Weather Change

Observation: "Weather Change" and "Sneeze" are correlated.

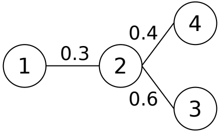
- "Weather Change" and "Catch Cold" are highly correlated.
- "Catch Cold" and "Sneeze" are highly correlated.

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

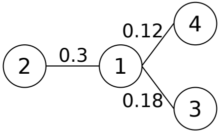
- Given that "Catch Cold" is false, "Sneeze" is likely to be false, independent of whether "Weather Change" is true or not.
- Given that "Catch Cold" is true, "Sneeze" is likely to be true, independent of whether "Weather Change" is true or not.

Other Examples

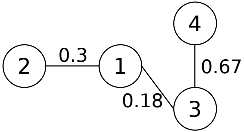
Suppose that $\rho(X_1, X_2) = 0.3$, $\rho(X_1, X_3) = 0.18$, and $\rho(X_1, X_4) = 0.12$.
Suppose that on one day, $X_2 \uparrow 0.2$, $X_3 \downarrow 0.1$, and $X_4 \downarrow 0.5$.
Find the expected change of X_1.



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

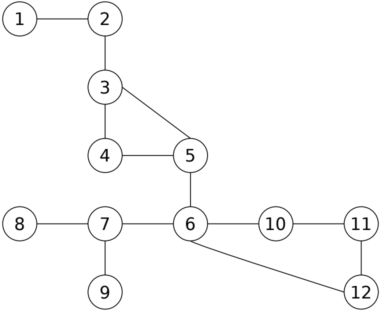


$$\begin{aligned}\mathrm{E}[\Delta X_1] &= 0.2 \times 0.3 - 0.1 \times 0.18 - 0.5 \times 0.12 \\ &= -0.018.\end{aligned}$$



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

Nondirected Graphical Model



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

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

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 $\boldsymbol{K}$ be the precision matrix of $\boldsymbol{X}$.

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 $\epsilon_a \sim \mathcal{N}\left(0, K_{aa}^{-1}\right)$ 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

$$\operatorname{cor}\left(X_a, X_b | \boldsymbol{X}_{\mathcal{C}}\right) = -\frac{K_{ab}}{\sqrt{K_{aa}K_{bb}}}.$$

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

$$\operatorname{cov}\left(\boldsymbol{X}_{\{a,b\}} | \boldsymbol{X}_{\mathcal{C}}\right) = \begin{bmatrix} K_{aa} & K_{ab} \\ K_{ba} & K_{bb} \end{bmatrix}^{-1} = \frac{1}{K_{aa}K_{bb} - K_{ab}^2} \begin{bmatrix} K_{bb} & -K_{ba} \\ -K_{ab} & K_{aa} \end{bmatrix}.$$

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

$$\frac{1}{m} \sum \left(\boldsymbol{X} - \bar{\boldsymbol{X}} \right) \left(\boldsymbol{X} - \bar{\boldsymbol{X}} \right)^T \to \boldsymbol{\Sigma} \tag{35}$$

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

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

Assumption: $\boldsymbol{K}$ is sparse.

Estimation via Regression (1)

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

$$\mathrm{E}\left[X_a | X_b : \ b \neq a\right] = \sum_b \theta_{ba} X_b.$$

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

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

Or in matrix format

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

where $\Theta = \{\boldsymbol{\Theta} : \ \mathrm{diag}\left(\boldsymbol{\Theta}\right) = \mathbf{0}\}$.

Estimation via Regression (2)

The objective function can be rewritten as

$$\begin{aligned}
\mathrm{E}\left[\left\|\boldsymbol{X}-\boldsymbol{\Theta}^T\boldsymbol{X}\right\|_2^2\right] &\approx \frac{1}{m}\sum\left(\boldsymbol{x}-\boldsymbol{\Theta}^T\boldsymbol{x}\right)^T\left(\boldsymbol{x}-\boldsymbol{\Theta}^T\boldsymbol{x}\right) \\
&= \frac{1}{m}\left\|\begin{bmatrix}\boldsymbol{x}_{(1)}^T \\ \vdots \\ \boldsymbol{x}_{(m)}^T\end{bmatrix}-\begin{bmatrix}\boldsymbol{x}_{(1)}^T \\ \vdots \\ \boldsymbol{x}_{(m)}^T\end{bmatrix}\boldsymbol{\Theta}\right\|_F^2 \\
&= \frac{1}{m}\left\|\boldsymbol{X}-\boldsymbol{X}\boldsymbol{\Theta}\right\|_F^2.
\end{aligned}$$

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

Estimation via Regression (3)

The overall optimization problem:

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

Or

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