

Statistical Optimization

Lagrange duality

Zijian Guo

Zhejiang University
Center for Data Science

May 2, 2026

Constrained convex programming

- Convex programs:

$$\begin{aligned} & \text{minimize} && f_0(\theta) \\ & \text{subject to} && f_i(\theta) \leq 0, \quad i = 1, \dots, m, \\ & && h_i(\theta) = 0, \quad i = 1, \dots, p \end{aligned} \tag{1}$$

where $f_0, \dots, f_m$ are convex, and $h_1, \dots, h_p$ are affine.

- Feasible set

$$\Theta = \{\theta \in \mathbb{R}^d : f_i(\theta) \leq 0, i = 1, \dots, m, h_i(\theta) = 0, i = 1, \dots, p\},$$

which is convex.

- Domain

$$\mathcal{D} = \left(\bigcap_{i=0}^m \text{dom}(f_i) \right) \cap \left(\bigcap_{i=1}^p \text{dom}(h_i) \right),$$

often simply taken as $\mathbb{R}^d$.

Outline

Lagrange Duality

KKT Conditions

Applications of KKT

Lagrange duality

Definition. Given optimization problem (1), its *Lagrangian* is the function $L : \mathcal{D} \times \mathbb{R}^m \times \mathbb{R}^p \rightarrow \mathbb{R}$ defined by

$$L(\theta, \lambda, \nu) = f_0(\theta) + \sum_{i=1}^m \lambda_i f_i(\theta) + \sum_{i=1}^p \nu_i h_i(\theta). \quad (2)$$

The λ_i, ν_i are called *Lagrange multipliers*.

The *Lagrange dual function* is the function $g : \mathbb{R}^m \times \mathbb{R}^p \rightarrow \mathbb{R} \cup \{-\infty\}$ defined by

$$g(\lambda, \nu) = \inf_{\theta \in \mathcal{D}} L(\theta, \lambda, \nu). \quad (3)$$

Minimax Formulation of Lagrangian Duality

For any $\theta \in \mathcal{D}$,

$$\sup_{\lambda \geq 0, \nu \in \mathbb{R}^p} L(\theta, \lambda, \nu) = \begin{cases} f_0(\theta), & \theta \text{ feasible,} \\ +\infty, & \theta \text{ infeasible.} \end{cases}$$

Therefore, the primal problem can be written as

$$\min_{\theta \in \Theta} f_0(\theta) = \min_{\theta \in \mathcal{D}} \sup_{\lambda \geq 0, \nu \in \mathbb{R}^p} L(\theta, \lambda, \nu).$$

By exchanging the order of optimization, we also obtain the dual problem

$$\sup_{\lambda \geq 0, \nu \in \mathbb{R}^p} \inf_{\theta \in \mathcal{D}} L(\theta, \lambda, \nu) = \sup_{\lambda \geq 0, \nu \in \mathbb{R}^p} g(\lambda, \nu).$$

Convexity of the primal formulation

For any fixed (λ, ν) with $\lambda \geq 0$,

$$L(\theta, \lambda, \nu) = f_0(\theta) + \sum_{i=1}^m \lambda_i f_i(\theta) + \sum_{i=1}^p \nu_i h_i(\theta)$$

is convex in θ , since $f_0, \dots, f_m$ are convex, h_i are affine, and $\lambda_i \geq 0$. The nonnegative coefficients linear combination of convex and affine functions is still convex.

Therefore,

$$\phi(\theta) := \sup_{\lambda \geq 0, \nu} L(\theta, \lambda, \nu)$$

is also convex in θ , because the pointwise supremum of convex functions is convex.

Hence

$$\min_{\theta \in \mathcal{D}} \phi(\theta)$$

is a convex optimization problem.

Convexity of the dual formulation

For any fixed θ , $L(\theta, \lambda, \nu)$ is affine in (λ, ν) . Hence

$$g(\lambda, \nu) := \inf_{\theta \in \mathcal{D}} L(\theta, \lambda, \nu)$$

is concave in (λ, ν) , because the pointwise infimum of affine functions is concave. Since the dual feasible set

$$\{(\lambda, \nu) : \lambda \geq 0, \nu \in \mathbb{R}^p\}$$

is convex, the dual problem

$$\sup_{\lambda \geq 0, \nu \in \mathbb{R}^p} g(\lambda, \nu)$$

is a concave maximization problem over a convex set which is equal to a convex minimization problem.

Lagrange duality

Theorem (Weak Lagrange duality). Let θ be a feasible solution of the optimization problem (1), meaning that $f_i(\theta) \leq 0$ for $i = 1, \dots, m$ and $h_i(\theta) = 0$ for $i = 1, \dots, p$. Let g be the Lagrange dual function of (1) and $\lambda \in \mathbb{R}^m, \nu \in \mathbb{R}^p$ such that $\lambda \geq 0$. Then

$$g(\lambda, \nu) \leq f_0(\theta).$$

Proof. For any θ belonging to the feasible set,

$$g(\lambda, \nu) \leq L(\theta, \lambda, \nu) = f_0(\theta) + \sum_{i=1}^m \lambda_i f_i(\theta) + \sum_{i=1}^p \nu_i h_i(\theta)$$

$$\underbrace{\sum_{i=1}^m \lambda_i f_i(\theta)}_{\leq 0} + \underbrace{\sum_{i=1}^p \nu_i h_i(\theta)}_{=0} \Rightarrow g(\lambda, \nu) \leq f_0(\theta).$$



Lagrange Dual Problem

Definition. Let g be the Lagrange dual function of the optimization problem (1). Then the Lagrange dual of (1) is the optimization problem

$$\begin{array}{ll}\text{maximize} & g(\lambda, \nu) \\ \text{subject to} & \lambda \geq 0.\end{array}\tag{4}$$

- The equivalent minimization problem

$$\begin{array}{ll}\text{minimize} & -g(\lambda, \nu) \\ \text{subject to} & \lambda \geq 0\end{array}$$

is a convex program, even if (1) is not.

Strong Duality

Theorem. Suppose that the convex program (1) has a feasible solution θ that additionally satisfies $f_i(\theta) < 0$, $i = 1, \dots, m$ (a Slater point). Then

$$\inf_{\theta \in \Theta} f_0(\theta) = \sup_{\lambda \geq 0, \nu \in \mathbb{R}^p} g(\lambda, \nu)$$

for its Lagrange dual (4). Moreover, if this value is finite, it is attained by a feasible solution of the dual (4).

- Strong duality may still hold even without a Slater point, or even when (1) is not convex. Slater's condition simply provides one powerful sufficient condition.

Zero duality gap

Definition (Zero duality gap). Let θ^* be feasible for the primal (1) and (λ^*, ν^*) feasible for the Lagrange dual (4). The primal and dual solutions θ^* and (λ^*, ν^*) are said to have *zero duality gap* if

$$f_0(\theta^*) = g(\lambda^*, \nu^*).$$

Zero duality gap is a certificate for strong duality:

$$\inf_{\theta \in \Theta} f_0(\theta) \leq f_0(\theta^*) = g(\lambda^*, \nu^*) \leq \sup_{\lambda \geq 0, \nu \in \mathbb{R}^p} g(\lambda, \nu).$$

Zero Duality Gap and Its Consequences

If θ^* and (λ^*, ν^*) have zero duality gap, then we have the chain of (in)equalities:

$$\begin{aligned} f_0(\theta^*) &= g(\lambda^*, \nu^*) \\ &= \inf_{\theta \in \mathcal{D}} \left(f_0(\theta) + \sum_{i=1}^m \lambda_i^* f_i(\theta) + \sum_{i=1}^p \nu_i^* h_i(\theta) \right) \\ &\leq f_0(\theta^*) + \underbrace{\sum_{i=1}^m \lambda_i^* f_i(\theta^*)}_{\leq 0} + \underbrace{\sum_{i=1}^p \nu_i^* h_i(\theta^*)}_{=0} \\ &\leq f_0(\theta^*). \end{aligned} \tag{5}$$

All inequalities must be equalities.

Consequences from Zero Duality Gap (KKT)

Lemma (Complementary slackness). If zero duality gap holds, then

$$\lambda_i^* f_i(\theta^*) = 0, \quad i = 1, \dots, m.$$

Lemma (Vanishing Lagrangian gradient). If f_i and h_i are differentiable,

$$\nabla f_0(\theta^*) + \sum_{i=1}^m \lambda_i^* \nabla f_i(\theta^*) + \sum_{i=1}^p \nu_i^* \nabla h_i(\theta^*) = 0.$$

Outline

Lagrange Duality

KKT Conditions

Applications of KKT

KKT Conditions

Primal feasibility: $f_i(\theta^*) \leq 0, \quad i = 1, \dots, m, \quad h_j(\theta^*) = 0, \quad j = 1, \dots, p,$

Dual feasibility: $\lambda_i^* \geq 0, \quad i = 1, \dots, m,$

Complementary slackness: $\lambda_i^* f_i(\theta^*) = 0, \quad i = 1, \dots, m,$

Stationarity:
$$\nabla f_0(\theta^*) + \sum_{i=1}^m \lambda_i^* \nabla f_i(\theta^*) + \sum_{j=1}^p \nu_j^* \nabla h_j(\theta^*) = 0.$$

Global optimality, zero duality gap, and KKT

- **Weak duality** always holds:

$$g(\lambda, \nu) \leq f_0(\theta).$$

- **Strong duality** $\Rightarrow$ **KKT** if the primal and dual optima are attained and all f_i, h_j are differentiable.
- **Global optimality** $\Rightarrow$ **KKT** requires a constraint qualification (e.g. LICQ); convexity is not needed.
- **KKT** $\Rightarrow$ **global minimum + zero duality gap** requires convexity: $f_0, \dots, f_m$ convex and $h_1, \dots, h_p$ affine.
- Under **convexity + strong duality**,

$$\text{global optimality} \iff \text{zero duality gap} \iff \text{KKT}.$$

KKT necessary conditions (zero duality gap)

Theorem. Let θ^* be feasible for the primal problem (1), and let (λ^*, ν^*) be feasible for the dual problem (4). Assume that all f_i and h_i are differentiable. If zero duality gap holds with

$$f_0(\theta^*) = g(\lambda^*, \nu^*),$$

then $(\theta^*, \lambda^*, \nu^*)$ satisfies the KKT conditions:

Primal feasibility: $f_i(\theta^*) \leq 0, \quad i = 1, \dots, m, \quad h_j(\theta^*) = 0, \quad j = 1, \dots, p,$

Dual feasibility: $\lambda_i^* \geq 0, \quad i = 1, \dots, m,$

Complementary slackness: $\lambda_i^* f_i(\theta^*) = 0, \quad i = 1, \dots, m,$

Stationarity:
$$\nabla f_0(\theta^*) + \sum_{i=1}^m \lambda_i^* \nabla f_i(\theta^*) + \sum_{j=1}^p \nu_j^* \nabla h_j(\theta^*) = 0.$$

KKT necessary conditions (local minimum)

Theorem. Assume that all f_i and h_j in (1) are differentiable. Let θ^* be a local minimum of (1) satisfying a constraint qualification (e.g., LICQ: the gradients of the active inequality constraints and the equality constraints at θ^* are linearly independent). Then there exist $\lambda^* \in \mathbb{R}^m$ and $\nu^* \in \mathbb{R}^p$ such that

Primal feasibility: $f_i(\theta^*) \leq 0, \quad i = 1, \dots, m, \quad h_j(\theta^*) = 0, \quad j = 1, \dots, p,$

Dual feasibility: $\lambda_i^* \geq 0, \quad i = 1, \dots, m,$

Complementary slackness: $\lambda_i^* f_i(\theta^*) = 0, \quad i = 1, \dots, m,$

Stationarity:
$$\nabla f_0(\theta^*) + \sum_{i=1}^m \lambda_i^* \nabla f_i(\theta^*) + \sum_{j=1}^p \nu_j^* \nabla h_j(\theta^*) = 0.$$

Proof (Optional)

Proof. Let

$$I(\theta^*) = \{i \in \{1, \dots, m\} : f_i(\theta^*) = 0\}$$

be the active set.

Since θ^* is a local minimum of (1), it is feasible:

$$f_i(\theta^*) \leq 0, \quad h_j(\theta^*) = 0.$$

If there existed $d \in \mathbb{R}^d$ such that

$$\nabla f_0(\theta^*)^\top d < 0, \quad \nabla f_i(\theta^*)^\top d < 0 \quad (i \in I(\theta^*)), \quad \nabla h_j(\theta^*)^\top d = 0 \quad (j = 1, \dots, p),$$

then $\theta^* + td$ would be feasible for all sufficiently small $t > 0$ and

$$f_0(\theta^* + td) < f_0(\theta^*),$$

contradicting the local optimality of θ^* .

Proof (Optional)

Hence no such d exists. By Farkas' lemma (a theorem of alternatives), there exist $\lambda_i^* \geq 0$ for $i \in I(\theta^*)$ and $\nu_j^* \in \mathbb{R}$ such that

$$\nabla f_0(\theta^*) + \sum_{i \in I(\theta^*)} \lambda_i^* \nabla f_i(\theta^*) + \sum_{j=1}^p \nu_j^* \nabla h_j(\theta^*) = 0.$$

Set $\lambda_i^* = 0$ for $i \notin I(\theta^*)$. Then

$$\lambda_i^* \geq 0, \quad \lambda_i^* f_i(\theta^*) = 0, \quad i = 1, \dots, m.$$

Together with feasibility of θ^* , this gives the four KKT conditions.



KKT sufficient conditions

Theorem (KKT sufficient conditions). Let θ^* and (λ^*, ν^*) be feasible solutions of the primal optimization problem (1) and its Lagrange dual (4), respectively.

Further suppose that all f_i and h_i in (1) are differentiable, all f_i are **convex**, and all h_i are **affine**, and that

$$\lambda_i^* f_i(\theta^*) = 0, \quad i = 1, \dots, m, \quad (6)$$

$$\nabla f_0(\theta^*) + \sum_{i=1}^m \lambda_i^* \nabla f_i(\theta^*) + \sum_{i=1}^p \nu_i^* \nabla h_i(\theta^*) = 0. \quad (7)$$

Then θ^* is a **global minimum** of the primal problem (1), and the primal and dual problems have **zero duality gap**.

Proof (1/3)

Proof. Let θ be any feasible point. By convexity of f_0 ,

$$f_0(\theta) \geq f_0(\theta^*) + \nabla f_0(\theta^*)^\top (\theta - \theta^*).$$

Using stationary point,

$$\nabla f_0(\theta^*) = - \sum_{i=1}^m \lambda_i^* \nabla f_i(\theta^*) - \sum_{i=1}^p \nu_i^* \nabla h_i(\theta^*),$$

implying

$$f_0(\theta) - f_0(\theta^*) \geq - \sum_{i=1}^m \lambda_i^* \nabla f_i(\theta^*)^\top (\theta - \theta^*) - \sum_{i=1}^p \nu_i^* \nabla h_i(\theta^*)^\top (\theta - \theta^*).$$

Proof (2/3)

By convexity of each f_i ,

$$f_i(\theta) \geq f_i(\theta^*) + \nabla f_i(\theta^*)^\top (\theta - \theta^*),$$

hence

$$\nabla f_i(\theta^*)^\top (\theta - \theta^*) \leq f_i(\theta) - f_i(\theta^*).$$

Since each h_i is affine,

$$\nabla h_i(\theta^*)^\top (\theta - \theta^*) = h_i(\theta) - h_i(\theta^*).$$

Therefore,

$$f_0(\theta) - f_0(\theta^*) \geq - \sum_{i=1}^m \lambda_i^* (f_i(\theta) - f_i(\theta^*)) - \sum_{i=1}^p \nu_i^* (h_i(\theta) - h_i(\theta^*)).$$

Proof (3/3)

Since θ and θ^* are feasible,

$$f_i(\theta) \leq 0, \quad h_i(\theta) = 0, \quad h_i(\theta^*) = 0.$$

Also, by dual feasibility and complementary slackness,

$$\lambda_i^* \geq 0, \quad \lambda_i^* f_i(\theta^*) = 0.$$

Hence

$$\begin{aligned} f_0(\theta) - f_0(\theta^*) &\geq - \sum_{i=1}^m \lambda_i^* (f_i(\theta) - f_i(\theta^*)) \\ &= - \sum_{i=1}^m \lambda_i^* f_i(\theta) + 0 \geq 0. \end{aligned}$$

Thus $f_0(\theta) \geq f_0(\theta^*)$ for every feasible θ , so θ^* is primal optimal. By weak duality, (λ^*, ν^*) is dual optimal as well.



Strong duality + KKT $\Leftrightarrow$ global optimality

Theorem. Suppose $f_0, f_1, \dots, f_m$ are convex, $h_1, \dots, h_p$ are affine, and strong duality holds. Let θ^* be primal feasible and (λ^*, ν^*) be dual feasible. Then the following are equivalent:

- (A) θ^* is primal optimal and (λ^*, ν^*) is dual optimal,
- (B) $f_0(\theta^*) = g(\lambda^*, \nu^*)$,
- (C) $(\theta^*, \lambda^*, \nu^*)$ satisfies the KKT conditions.

Proof

Proof. We have already proved earlier that

$$(B) \Rightarrow (C), \quad (C) \Rightarrow (A).$$

So it remains only to show

$$(A) \Rightarrow (B).$$

Assume (A) holds, namely: θ^* is primal optimal and (λ^*, ν^*) is dual optimal.
Then

$$f_0(\theta^*) = p^*, \quad g(\lambda^*, \nu^*) = d^*.$$

Since strong duality holds,

$$p^* = d^*.$$

Therefore,

$$f_0(\theta^*) = g(\lambda^*, \nu^*),$$

which is exactly (B) .

Global optimality, zero duality gap, and KKT

- **Weak duality** always holds:

$$g(\lambda, \nu) \leq f_0(\theta).$$

- **Strong duality** $\Rightarrow$ **KKT** if the primal and dual optima are attained and all f_i, h_j are differentiable.
- **Global optimality** $\Rightarrow$ **KKT** requires a constraint qualification (e.g. LICQ); convexity is not needed.
- **KKT** $\Rightarrow$ **global minimum + zero duality gap** requires convexity: $f_0, \dots, f_m$ convex and $h_1, \dots, h_p$ affine.
- Under **convexity + strong duality**,

$$\text{global optimality} \iff \text{zero duality gap} \iff \text{KKT}.$$

Outline

Lagrange Duality

KKT Conditions

Applications of KKT

Example 1: projection onto the ℓ_1 ball

Consider the problem

$$\min_{\theta \in \mathbb{R}^d} \frac{1}{2} \|\theta - \eta\|_2^2 \quad \text{s.t.} \quad \|\theta\|_1 \leq \tau.$$

This is exactly the projection step in projected gradient descent for constrained Lasso:

$$\Pi_{\|\theta\|_1 \leq \tau}(\eta) = \arg \min_{\|\theta\|_1 \leq \tau} \frac{1}{2} \|\theta - \eta\|_2^2.$$

We first reformulate this problem into a smooth constrained optimization problem, and then apply the standard KKT conditions.

Projection keeps sign

We first claim that at the optimum, each coordinate has the same sign as η_i :

$$\theta_i^* \eta_i \geq 0, \quad i = 1, \dots, d.$$

Indeed, changing the sign of θ_i to match η_i does not change $\|\theta\|_1$, but can only decrease $(\theta_i - \eta_i)^2$. So we may write

$$\theta_i^* = \text{sign}(\eta_i) z_i^*, \quad z_i^* \geq 0$$

and

$$|\theta_i^*| = z_i^*, \quad (\theta_i^* - \eta_i)^2 = (z_i^* - |\eta_i|)^2.$$

Hence the projection problem becomes

$$\min_{z \in \mathbb{R}^d} \frac{1}{2} \sum_{i=1}^d (z_i - |\eta_i|)^2 \quad \text{s.t.} \quad \sum_{i=1}^d z_i \leq \tau, \quad z_i \geq 0.$$

KKT conditions for the smooth reformulation

Its Lagrangian is

$$\mathcal{L}(z, \lambda, \nu) = \frac{1}{2} \sum_{i=1}^d (z_i - |\eta_i|)^2 + \lambda \left(\sum_{i=1}^d z_i - \tau \right) - \sum_{i=1}^d \nu_i z_i,$$

where

$$\lambda \geq 0, \quad \nu_i \geq 0.$$

The KKT conditions are:

$$z_i^* - |\eta_i| + \lambda^* - \nu_i^* = 0, \quad i = 1, \dots, d,$$

$$\sum_{i=1}^d z_i^* \leq \tau, \quad z_i^* \geq 0,$$

$$\lambda^* \left(\sum_{i=1}^d z_i^* - \tau \right) = 0, \quad \nu_i^* z_i^* = 0.$$

From the KKT conditions to soft-thresholding

$$z_i^* - |\eta_i| + \lambda^* - \nu_i^* = 0.$$

Case 1: $z_i^* > 0$. Then complementary slackness gives

$$\nu_i^* z_i^* = 0 \implies \nu_i^* = 0.$$

So stationarity becomes

$$z_i^* = |\eta_i| - \lambda^*.$$

Case 2: $z_i^* = 0$. Then stationarity gives

$$-|\eta_i| + \lambda^* - \nu_i^* = 0 \implies \nu_i^* = \lambda^* - |\eta_i|.$$

Since $\nu_i^* \geq 0$, we must have

$$|\eta_i| \leq \lambda^*.$$

Therefore,

$$z_i^* = \max(|\eta_i| - \lambda^*, 0).$$

How do we determine λ^* ?

Since

$$\theta_i^* = \text{sign}(\eta_i) z_i^*,$$

we obtain

$$\theta_i^* = \text{sign}(\eta_i) \max(|\eta_i| - \lambda^*, 0).$$

So the projection has the soft-thresholding form.

To determine λ^* , there are two cases.

Case 1: If $\|\eta\|_1 \leq \tau$, then η is already feasible, so

$$\theta^* = \eta, \quad \lambda^* = 0.$$

How do we determine λ^* ?

Case 2: If $\|\eta\|_1 > \tau$, then the constraint is active, so

$$\sum_{i=1}^d z_i^* = \tau.$$

Using

$$z_i^* = \max(|\eta_i| - \lambda^*, 0),$$

we get

$$\sum_{i=1}^d \max(|\eta_i| - \lambda^*, 0) = \tau.$$

Thus λ^* is the threshold that shrinks the coordinates just enough to make the ℓ_1 norm equal to τ .

Example 2: Simplex projection

We consider the problem

$$\begin{aligned} \min_{\theta \in \mathbb{R}^n} \quad & \frac{1}{2} \|\theta - \eta\|^2 \\ \text{s.t.} \quad & \sum_{i=1}^n \theta_i - 1 = 0, \\ & -\theta_i \leq 0, \quad i = 1, \dots, n. \end{aligned}$$

The Lagrangian is

$$L(\theta, \lambda, \nu) = \frac{1}{2} \|\theta - \eta\|^2 + \nu \left(\sum_{i=1}^n \theta_i - 1 \right) - \sum_{i=1}^n \lambda_i \theta_i.$$

Simplex projection: KKT conditions

The KKT conditions are:

1. Stationarity:

$$\nabla_{\theta_i} L = \theta_i - \eta_i + \nu - \lambda_i = 0, \quad i = 1, \dots, n.$$

2. Primal feasibility:

$$\sum_{i=1}^n \theta_i = 1, \quad \theta_i \geq 0.$$

3. Dual feasibility:

$$\lambda_i \geq 0.$$

4. Complementary slackness:

$$\lambda_i \theta_i = 0.$$

Simplex projection: solving the solution

From stationarity,

$$\theta_i = \eta_i - \nu + \lambda_i.$$

Now use complementary slackness:

$$\theta_i > 0 \implies \lambda_i = 0 \implies \theta_i = \eta_i - \nu.$$

$$\theta_i = 0 \implies \lambda_i = \nu - \eta_i \geq 0 \implies \eta_i \leq \nu.$$

Therefore,

$$\theta_i = (\eta_i - \nu)_+, \quad i = 1, \dots, n.$$

The equality constraint then gives

$$\sum_{i=1}^n (\eta_i - \nu)_+ = 1.$$

Simplex projection: proposition and remark

$$\Delta^n := \left\{ \theta \in \mathbb{R}^n : \theta_i \geq 0, \sum_{i=1}^n \theta_i = 1 \right\}.$$

Proposition. The Euclidean projection of η onto Δ^n is

$$\theta_i^* = (\eta_i - \tau)_+, \quad i = 1, \dots, n,$$

where τ is the unique scalar satisfying

$$\sum_{i=1}^n (\eta_i - \tau)_+ = 1.$$

Remark. The solution is sparse. Complementary slackness tells us $\theta_i = 0$ whenever $\eta_i < \tau$ without any computation.

KL divergence

For probability vectors on Δ^n , a natural non-Euclidean distance-like quantity is KL divergence

$$D_{\text{KL}}(\theta \parallel \eta) := \sum_{i=1}^n \theta_i \log \frac{\theta_i}{\eta_i}.$$

Remark. It can be derived as the Bregman divergence associated with the negative entropy

$$\Phi(\theta) := \sum_{i=1}^n \theta_i \log \theta_i.$$

Intuition. It measures how different two distributions are, and is more suitable than the Euclidean distance when θ represents probabilities.

Softmax map

For any vector $\eta \in \mathbb{R}^n$, define

$$(\text{softmax}(\eta))_i := \frac{e^{\eta_i}}{\sum_{j=1}^n e^{\eta_j}}, \quad i = 1, \dots, n.$$

Basic properties:

- $(\text{softmax}(\eta))_i > 0$ for all i ,
- $\sum_{i=1}^n (\text{softmax}(\eta))_i = 1$,
- so $\text{softmax}(\eta) \in \Delta^n$.

Intuition.

- The largest coordinate of η gets the largest weight. But the other coordinates are not discarded completely.
- If one entry of η is much larger than the others, then $\text{softmax}(\eta)$ is concentrated near that index.
- So it behaves like a *smooth* or *softened* version of $\arg \max_i \eta_i$.

Example 3: Mirror descent on simplex

θ_{t+1} is obtained by the non-Euclidean projection

$$\theta_{t+1} = \arg \min_{\theta \in \Delta^n} \left\{ \langle g_t, \theta \rangle + \frac{1}{\gamma_t} D_{\text{KL}}(\theta \| \theta_t) \right\}, \quad \text{with } g_t := \nabla f(\theta_t)$$

where

$$D_{\text{KL}}(\theta \| \theta_t) = \sum_{i=1}^n \theta_i \log \frac{\theta_i}{(\theta_t)_i}.$$

Equivalent form

We start from the mirror descent update

$$\theta_{t+1} = \arg \min_{\theta \in \Delta^n} \left\{ \langle \mathbf{g}_t, \theta \rangle + \frac{1}{\gamma_t} \sum_{i=1}^n \theta_i \log \frac{\theta_i}{(\theta_t)_i} \right\}.$$

Expand the logarithm and the objective becomes

$$\langle \mathbf{g}_t, \theta \rangle + \frac{1}{\gamma_t} \sum_{i=1}^n \theta_i \log \theta_i - \frac{1}{\gamma_t} \langle \theta, \log \theta_t \rangle.$$

Multiplying by $\gamma_t > 0$ does not change the $\arg \min$, so this is equivalent to

$$\theta_{t+1} = \arg \min_{\theta \in \Delta^n} \left\{ \sum_{i=1}^n \theta_i \log \theta_i - \langle \theta, \eta_t \rangle \right\}.$$

where

$$\eta_t := \log \theta_t - \gamma_t \mathbf{g}_t.$$

Entropy regularization: Lagrangian and stationarity

$$\begin{aligned} \min_{\theta \in \mathbb{R}^n} \quad & \sum_{i=1}^n \theta_i \log \theta_i - \langle \theta, \eta \rangle \\ \text{s.t.} \quad & \sum_{i=1}^n \theta_i - 1 = 0. \end{aligned}$$

The Lagrangian is

$$L(\theta, \nu) = \sum_{i=1}^n \theta_i \log \theta_i - \langle \theta, \eta \rangle + \nu \left(\sum_{i=1}^n \theta_i - 1 \right).$$

The stationarity condition is

$$\nabla_{\theta_i} L = \log \theta_i + 1 - \eta_i + \nu = 0, \quad i = 1, \dots, n.$$

Hence

$$\theta_i = e^{\eta_i - 1 - \nu}, \quad i = 1, \dots, n.$$

Entropy regularization: softmax solution

Using the constraint

$$\sum_{i=1}^n \theta_i = 1,$$

we obtain

$$\theta_i = \frac{e^{\eta_i}}{\sum_{j=1}^n e^{\eta_j}}, \quad i = 1, \dots, n.$$

Therefore,

$$\theta^* = \text{softmax}(\eta).$$

So, under entropy geometry, the closed-form solution is no longer thresholding; it becomes a normalized exponential map.

Mirror descent update on Δ^n

Applying the previous softmax formula with $\eta = \eta_t$, we get

$$\theta_{t+1} = \text{softmax}(\eta_t).$$

Equivalently,

$$(\theta_{t+1})_i = \frac{(\theta_t)_i e^{-\gamma_t (g_t)_i}}{\sum_{j=1}^n (\theta_t)_j e^{-\gamma_t (g_t)_j}}, \quad i = 1, \dots, n.$$

Remark. These three examples illustrate the power of the Lagrangian formulation: it often leads to many closed-form update rules.