Math of ML : Exercises 6 *

October 27, 2025

Exercise 1 (Conjugate kernels for step and ReLU activations)**.** *Consider the two-layer neural network*

$$f(x; w, \eta) = \frac{1}{\sqrt{m}} \sum_{j=1}^{m} \eta_j \sigma(w_j^\top x),$$

where (w, η) *are the weights of the network and* σ *is the activation function.*

Suppose that the weights $w = (w_1, \dots, w_j)$ *are sampled independently from the multivariate standard normal distribution* $w_j \sim \mathcal{N}(0, I_d)$*, where* I_d *is the* $d \times d$ *identity matrix.*

When the weights w *are held fixed and only the weights* η *are trained, training the neural network* $f(x; w\eta)$ *corresponds to fitting a kernel method with the feature map*

$$\phi(x)_j = \frac{1}{\sqrt{m}} \sigma(w_j^\top x)$$

with the corresponding kernel $\widehat{k}_m(x, x') = \frac{1}{m} \sum_{j=1}^{m} \sigma(w_j^\top x)\sigma(w_j^\top x')$*. By the law of large numbers, we have*

$$\lim_{m \to \infty} \widehat{k}_m(x, x') = k(x, x') = \mathbf{E}_{w \sim \mathcal{N}(0, I_d)} \left[\sigma(w^\top x)\sigma(w^\top x') \right].$$

The kernel k *is called the* conjugate kernel*. Fix any* x, x' *and let* $\theta = \arccos(x^\top x'/(\|x\|_2 \|x'\|_2))$*. Prove the following closed-form expressions for* k*.*

1. *For the step activation function* $\sigma(x) = 1$ *if* $x \geq 0$ *and* $\sigma(x) = 0$ *if* $x < 0$ *we have*

$$k(x, x') = \frac{1}{2\pi}(\pi - \theta).$$

2. $(\star)$ *For the ReLU activation function* $\sigma(x) = \max\{0, x\}$ *we have*

$$k(x, x') = \frac{\|x\|_2 \|x'\|_2}{2\pi} \left(\sin\theta + (\pi - \theta)\cos\theta \right). \tag{1}$$

Hint: *use the fact that for any orthogonal matrix* P *we have* $k(x, x') = k(Px, Px')$*.*

Solution 1. *Let* $H = \operatorname{span}(x, x')$*. Let* $u_1 = x/\|x\|_2$*. If* $\dim(H) = 2$*, take* $u_2 \in H$ *to be any orthonormal vector to* u_1*. Extend the set of vectors* $\{u_1, u_2\}$ *(or* $\{u_1\}$ *if* $\dim(H) = 1$*) to an orthonormal basis of* $\mathbb{R}^d$ *denoted by* $\{u_1, u_2, \dots, u_d\}$*. Let* $P \in \mathbb{R}^d$ *be a matrix with the* i*-th row equal to* u_i*. Then* P *is an orthogonal matrix and hence, we have*

$$k(x, x') = \mathbf{E}_{w \sim \mathcal{N}(0, I_d)} \left[\sigma(w^\top x)\sigma(w^\top x') \right]$$

*Lénaïc Chizat EPFL lenaic.chizat@epfl.ch

$$
\begin{aligned}
&= \mathbf{E}_{w\sim\mathcal{N}(0,I_d)}\left[\sigma((Pw)^\top(Px))\sigma((Pw)^\top(Px'))\right] \\
&= \mathbf{E}_{w\sim\mathcal{N}(0,I_d)}\left[\sigma((Pw)^\top(Px))\sigma((Pw)^\top(Px'))\right] \\
&= \mathbf{E}_{w\sim\mathcal{N}(0,I_d)}\left[\sigma((w)^\top(Px))\sigma((w)^\top(Px'))\right] \\
&= k(Px, Px'),
\end{aligned}
$$

where the penultimate step follows because the standard Gaussian measure is invariant under orthogonal transformations. By our construction of the change of basis matrix P we have $Px = (a_1, 0, 0, \ldots, 0)^\mathsf{T} \in \mathbb{R}^d$ and $Px' = (b_1, b_2, 0, \ldots, 0)^\mathsf{T} \in \mathbb{R}^d$. Hence, we have

$$
\begin{aligned}
k(x, x') &= \int_{\mathbb{R}^d} \frac{1}{(2\pi)^{d/2}} \exp(-\|w\|_2^2/2)\sigma(a_1 w_1)\sigma(b_1 w_1 + b_2 w_2) dw_1 dw_2 \ldots dw_d \\
&= \frac{1}{2\pi}\int_{\mathbb{R}^2} \exp(-(w_1^2 + w_2^2)/2)\sigma(a_1 w_1)\sigma(b_1 w_1 + b_2 w_2) dw_1 dw_2.
\end{aligned}
$$

Now observe that

$$
a_1 = \|x\|_2, b_1 = \|x'\|\cos\theta, b_2 = \|x'\|\sin\theta.
$$

Hence, we have

$$
k(x, x') = \frac{1}{2\pi}\int_{\mathbb{R}^2} \exp(-(w_1^2 + w_2^2)/2)\sigma(\|x\|_2 w_1)\sigma(\|x'\|_2(\cos(\theta)w_1 + \sin(\theta)w_2)) dw_1 dw_2 \quad (2)
$$

In what follows, we assume that $\sin\theta \neq 0$. If $\sin\theta = 0$, then the vectors x and x' are co-linear, and this case can be considered separately. Under this assumption, we can introduce change of variables

$$
u = w_1, v = \cos(\theta)w_1 + \sin(\theta)w_2 \quad (3)
$$

so that $(w_1, w_2) = T(u, v) = (u, v/\sin(\theta) - u\cos(\theta)/\sin(\theta))$. In particular, the absolute value of the determinant of the Jacobian is equal to

$$
|\det J| = \left|\det \begin{pmatrix} 1 & 0 \\ -\frac{\cos(\theta)}{\sin(\theta)} & \frac{1}{\sin(\theta)} \end{pmatrix}\right| = \frac{1}{\sin(\theta)},
$$

where we removed the absolute value in the last line, since by definition of θ we have $\theta \in [-\pi, \pi]$.

1. *Let σ_{step} be the step activation function. Then, the integral in (2) becomes*

$$
\begin{aligned}
&k(x, x') \\
&= \frac{1}{2\pi}\int_{\mathbb{R}^2} \exp(-(w_1^2 + w_2^2)/2)\sigma_{step}(\|x\|_2 w_1)\sigma_{step}(\|x'\|_2(\cos(\theta)w_1 + \sin(\theta)w_2)) dw_1 dw_2 \\
&= \frac{1}{2\pi}\int_{\mathbb{R}^2} \exp(-(w_1^2 + w_2^2)/2)\sigma_{step}(w_1)\sigma_{step}(\cos(\theta)w_1 + \sin(\theta)w_2) dw_1 dw_2
\end{aligned}
$$

Hence, with the change of variables (3) we have

$$
\begin{aligned}
&k(x, x') \\
&= \frac{1}{2\pi}\frac{1}{\sin(\theta)}\int_{\mathbb{R}^2} \exp\left(-\frac{u^2 + v^2 - 2uv\cos(\theta)}{2\sin^2(\theta)}\right)\sigma_{step}(u)\sigma_{step}(v) du dv \\
&= \frac{1}{2\pi}\frac{1}{\sin(\theta)}\int_0^\infty\int_0^\infty \exp\left(-\frac{u^2 + v^2 - 2uv\cos(\theta)}{2\sin^2(\theta)}\right) du dv. \quad (4)
\end{aligned}
$$

We now use the substitution[1] $v = us, dv = u\,ds$ *for* $s > 0$. *We then have*

$$
\begin{aligned}
&k(x,x') \\
&= \frac{1}{2\pi}\frac{1}{\sin(\theta)}\int_0^\infty\int_0^\infty \exp\left(-\frac{u^2(1+s^2-2s\cos(\theta))}{2\sin^2(\theta)}\right) u\,du\,ds. \\
&= \frac{1}{2\pi}\frac{1}{\sin(\theta)}\int_0^\infty \frac{\sqrt{2\pi\sin^2(\theta)}}{\sqrt{1+s^2-2s\cos(\theta)}}\int_0^\infty u\frac{\sqrt{1+s^2-2s\cos(\theta)}}{\sqrt{2\pi\sin^2(\theta)}}\exp\left(-\frac{u^2(1+s^2-2s\cos(\theta))}{2\sin^2(\theta)}\right) du\,ds. \\
&= \frac{1}{2\pi}\frac{1}{\sin(\theta)}\int_0^\infty \frac{\sqrt{2\pi\sin^2(\theta)}}{\sqrt{1+s^2-2s\cos(\theta)}}\left[\frac{1}{2}\mathbf{E}_{Z\sim\mathcal{N}\left(0,\frac{\sin^2(\theta)}{1+s^2-2s\cos(\theta)}\right)}\left[|Z|\right]\right] ds. \\
&= \frac{1}{2\pi}\frac{1}{\sin(\theta)}\int_0^\infty \frac{\sqrt{2\pi\sin^2(\theta)}}{\sqrt{1+s^2-2s\cos(\theta)}}\left[\frac{1}{2}\sqrt{\frac{\sin^2(\theta)}{1+s^2-2s\cos(\theta)}}\mathbf{E}_{Z\sim\mathcal{N}(0,1)}\left[|Z|\right]\right] ds. \\
&= \frac{1}{2\pi}\frac{1}{\sin(\theta)}\int_0^\infty \frac{\sqrt{2\pi\sin^2(\theta)}}{\sqrt{1+s^2-2s\cos(\theta)}}\left[\frac{1}{\sqrt{2\pi}}\sqrt{\frac{\sin^2(\theta)}{1+s^2-2s\cos(\theta)}}\right] ds. \\
&= \frac{1}{2\pi}\frac{1}{\sin(\theta)}\int_0^\infty \frac{\sin^2(\theta)}{1+s^2-2s\cos(\theta)} ds. \\
&= \sin(\theta)\frac{1}{2\pi}\left[\frac{\arctan\left(\frac{s-\cos(\theta)}{\sqrt{1-\cos^2(\theta)}}\right)}{\sqrt{1-\cos^2(\theta)}}\right]_{s=0}^{s=\infty} \\
&= \sin(\theta)\frac{1}{2\pi}\left[\frac{\arctan\left(\frac{s-\cos(\theta)}{\sin(\theta)}\right)}{\sin(\theta)}\right]_{s=0}^{s=\infty} \\
&= \frac{1}{2\pi}\left[\arctan\left(\frac{s-\cos(\theta)}{\sin(\theta)}\right)\right]_{s=0}^{s=\infty} \\
&= \frac{1}{2\pi}\left[\frac{\pi}{2} - \arctan\left(\frac{\cos(\theta)}{\sin(\theta)}\right)\right] \\
&= \frac{1}{2\pi}\left[\frac{\pi}{2} - \left(\theta - \frac{\pi}{2}\right)\right] \\
&= \frac{1}{2\pi}\left[\pi - \theta\right].
\end{aligned}
$$

2. *Computations for the ReLU activation function are more involved. See, for example, the appendix in the below-cited reference.*

   Observe that $\sigma_{ReLU}(x) = \sigma_{step}(x)x$. *Hence, repeating the same steps used to compute the conjugate kernel for the step activation up to* (4) *we have*

$$
\begin{aligned}
&k(x,x') \\
&= \frac{\|x\|_2\|x'\|_2}{2\pi}\frac{1}{\sin(\theta)}\int_0^\infty\int_0^\infty \exp\left(-\frac{u^2+v^2-2uv\cos(\theta)}{2\sin^2(\theta)}\right) uv\,du\,dv. \qquad (5)
\end{aligned}
$$

   Observe that

$$\sin^2(\theta)\exp\left(-\frac{v^2}{2\sin^2(\theta)}\right)$$

[1] This substitution to evaluate Gaussian integrals dates back at least to Laplace https://en.wikipedia.org/wiki/Gaussian_integral#By_Cartesian_coordinates.

$$= \int_0^\infty -\frac{d}{du}\sin^2(\theta)\exp\left(-\frac{u^2+v^2-2uv\cos(\theta)}{2\sin^2(\theta)}\right)du$$
$$= \int_0^\infty u - v\cos(\theta)\exp\left(-\frac{u^2+v^2-2uv\cos(\theta)}{2\sin^2(\theta)}\right)du$$
$$= \int_0^\infty u\exp\left(-\frac{u^2+v^2-2uv\cos(\theta)}{2\sin^2(\theta)}\right)du - v\cos(\theta)\int_0^\infty \exp\left(-\frac{u^2+v^2-2uv\cos(\theta)}{2\sin^2(\theta)}\right)du.$$

In particular, we have

$$\int_0^\infty\int_0^\infty \exp\left(-\frac{u^2+v^2-2uv\cos(\theta)}{2\sin^2(\theta)}\right)uvdudv$$
$$= \int_0^\infty v\left[\int_0^\infty u\exp\left(-\frac{u^2+v^2-2uv\cos(\theta)}{2\sin^2(\theta)}\right)du\right]dv$$
$$= \int_0^\infty v\left[\sin^2(\theta)\exp\left(-\frac{v^2}{2\sin^2(\theta)}\right) + v\cos(\theta)\int_0^\infty \exp\left(-\frac{u^2+v^2-2uv\cos(\theta)}{2\sin^2(\theta)}\right)du\right]dv$$

$$= \sin^2(\theta)\int_0^\infty v\exp\left(-\frac{v^2}{2\sin^2(\theta)}\right)dv + \cos(\theta)\int_0^\infty\int_0^\infty v^2\exp\left(-\frac{u^2+v^2-2uv\cos(\theta)}{2\sin^2(\theta)}\right)dudv$$

$$= \sin^4(\theta) + \cos(\theta)\int_0^\infty\int_0^\infty v^2\exp\left(-\frac{u^2+v^2-2uv\cos(\theta)}{2\sin^2(\theta)}\right)dudv.$$

Plugging the above identity into (5) yields

$$k(x,x') = \frac{\|x\|_2\|x'\|_2}{2\pi}\left(\sin^3(\theta) + \frac{\cos(\theta)}{\sin(\theta)}\int_0^\infty\int_0^\infty v^2\exp\left(-\frac{u^2+v^2-2uv\cos(\theta)}{2\sin^2(\theta)}\right)dudv\right) \quad (6)$$

It thus remains to evaluate the double integral in the above display equation. With the change of variables $v = us, dv = uds$ as in the proof for the step activation we obtain

$$\int_0^\infty\int_0^\infty v^2\exp\left(-\frac{u^2+v^2-2uv\cos(\theta)}{2\sin^2(\theta)}\right)dudv$$
$$= \int_0^\infty s^2\left[\int_0^\infty u^3\exp\left(-\frac{u^2(1+s^2-2s\cos(\theta))}{2\sin^2(\theta)}\right)du\right]ds$$
$$= \int_0^\infty s^2\frac{\sqrt{2\pi}\sin(\theta)}{\sqrt{1+s^2-2s\cos(\theta)}}\left[\int_0^\infty \frac{\sqrt{1+s^2-2s\cos(\theta)}}{\sqrt{2\pi}\sin(\theta)}u^3\exp\left(-\frac{u^2(1+s^2-2s\cos(\theta))}{2\sin^2(\theta)}\right)du\right]ds$$

$$= \int_0^\infty s^2\frac{\sqrt{2\pi}\sin(\theta)}{\sqrt{1+s^2-2s\cos(\theta)}}\left[\frac{1}{2}\mathbf{E}_{Z\sim\mathcal{N}(0,\frac{\sin^2(\theta)}{1+s^2-2s\cos(\theta)})}\left[|Z|^3\right]\right]ds$$
$$= \int_0^\infty s^2\frac{\sqrt{2\pi}\sin(\theta)}{\sqrt{1+s^2-2s\cos(\theta)}}\left[\frac{1}{2}\cdot 2\sqrt{\frac{2}{\pi}}\cdot\left(\frac{\sin^2(\theta)}{1+s^2-2s\cos(\theta)}\right)^{3/2}\right]ds$$
$$= 2\sin^4(\theta)\int_0^\infty\left(\frac{s}{1+s^2-2s\cos(\theta)}\right)^2 ds$$
$$= 2\sin^4(\theta)\left[\frac{-\frac{\pi/2}{\sin(\theta)}}{-2\sin^2(\theta)} - \frac{\cos(\theta) - \left(\frac{\tan^{-1}(-\cot(\theta))}{\sin(\theta)}\right)}{-2\sin^2(\theta)}\right]$$

$$
\begin{aligned}
&= -\sin^2(\theta)\left[-\frac{\pi/2}{\sin(\theta)} - \cos(\theta) + \left(\frac{\tan^{-1}(-\cot(\theta))}{\sin(\theta)}\right)\right] \\
&= -\sin^2(\theta)\left[-\frac{\pi/2}{\sin(\theta)} - \cos(\theta) + \left(\frac{\theta - \frac{\pi}{2}}{\sin(\theta)}\right)\right] \\
&= -\sin(\theta)\left[-\pi/2 - \cos(\theta)\sin(\theta) + \left(\theta - \frac{\pi}{2}\right)\right] \\
&= \sin(\theta)\,(\pi - \theta) + \sin^2(\theta)\cos(\theta).
\end{aligned}
$$

Plugging in the above identity into (6) yields

$$
\begin{aligned}
&k(x, x') \\
&= \frac{\|x\|_2\|x'\|_2}{2\pi}\left(\sin^3(\theta) + \cos(\theta)\,(\pi - \theta) + \sin(\theta)\cos^2(\theta)\right) \\
&= \frac{\|x\|_2\|x'\|_2}{2\pi}\left(\sin^3(\theta) + \cos(\theta)\,(\pi - \theta) + \sin(\theta)(1 - \sin^2(\theta))\right) \\
&= \frac{\|x\|_2\|x'\|_2}{2\pi}\left(\sin(\theta) + \cos(\theta)\,(\pi - \theta)\right).
\end{aligned}
$$

Our proof is complete.

Remark 1. *Exercise 1 is based on the paper by Cho and Saul [2009]. See the appendix of the above cited paper for computations for a more general family of activation functions of which the step and ReLU activations are special cases.*

Exercise 2 (Two-layer neural networks as kernel methods)**.** *Consider the setting of Exercise 1. Consider sampling input vectors* $(x, 1) \in \mathbb{R}^2$*, where* $x \sim \text{Uniform}([0, 1])$*. Conditionally on the input vector* $(x, 1)$*, generate a response variable* $y = f^*(x) + \mathcal{N}(0, 0.1)$ *for the choices* $f^*(x) = x, f^*(x) = 1$ *and* $f^*(x) = \sin(2\pi x)$*. For each choice of* f^**, generate a dataset* $(x_i, y_i)_{i=1}^n$ *of size* $n = 100$*. Fit the kernel ridge regression estimator* $\widehat{f}_\infty$ *for the conjugate kernel of ReLU network (1). For* $m = 5, 10, 100, 1000$*, fit the kernel ridge regression estimator* $\widehat{f}_m$ *using the kernel* $\widehat{k}_m$ *(cf. Exercise 1) with ReLU activation. Plot the learned functions* $\widehat{f}_m$ *(for* $m = 5, 10, 100, 1000, \infty$*) and the function* f^* *used to generate the data.*

The next exercise is not about kernels nor neural networks. It introduces the notion of duality, which plays a central role in convex optimization, and will be useful later in the course.

Exercise 3 (KKT Conditions)**.** *Consider the convex optimization problem:*

$$
\begin{aligned}
&\min_{x \in \mathbb{R}^d} f(x) \\
&\text{subject to} \quad f_i(x) \leq 0 \text{ for } i = 1, \ldots, m,
\end{aligned}
\tag{7}
$$

where the functions $f, f_1, \ldots, f_m : \mathbb{R}^d \to \mathbb{R}$ *are convex and differentiable.*

1. *The Lagrangian* $L : \mathbb{R}^d \times \mathbb{R}^m \to \mathbb{R}$ *is defined by* $L(x, \lambda) = f(x) + \sum_{i=1}^m \lambda_i f_i(x)$*. Prove that the problem (7) can be reformulated as*

$$
\inf_{x \in \mathbb{R}^d} \sup_{\lambda \geq 0} L(x, \lambda), \tag{8}
$$

where the inequality $\lambda \geq 0$ *is to be interpreted coordinate-wise. We will refer to the problem (8) as the primal problem.*

2. *Denote the optimal (minimum) value attained by the primal minimization problem (8) by $p^* \in \mathbb{R} \cup \{\pm\infty\}$. We now introduce the dual problem, the maximization problem in $\lambda \geq 0$ defined by*

$$\sup_{\lambda \geq 0} \inf_{x \in \mathbb{R}^d} L(x, \lambda). \tag{9}$$

Denote the optimal (maximum) value attained by the dual maximization problem (9) by $d^* \in \mathbb{R} \cup \{\pm\infty\}$. Prove that $d^* \leq p^*$ (weak duality; $p^* - d^*$ is called the duality gap).

3. *Prove that $x^* \in \mathbb{R}^d$ is optimal for (8), $\lambda^* \in \mathbb{R}^m$ is optimal for (9), and $p^* = d^*$ (strong duality holds)* if and only if *the pair (x^*, λ^*) satisfies the Karush-Kuhn-Tucker (KKT) conditions:*

   (a) *(Stationarity) $\nabla f(x^*) + \sum_{i=1}^m \lambda_i^* \nabla f_i(x^*) = 0$.*

   (b) *(Primal feasibility) $f_i(x^*) \leq 0$ for $i = 1, \ldots, m$.*

   (c) *(Dual feasibility) $\lambda_i^* \geq 0$ for $i = 1, \ldots, m$.*

   (d) *(Complementary slackness) $\lambda_i^* f_i(x_i^*) = 0$ for $i = 1, \ldots, m$.*

   Remark: Strong duality can be ensured under so-called constraint qualification conditions. One such condition, called Slater's condition, *is the existence of $x \in \mathbb{R}^d$ such that $f_i(x) < 0$ for $i = 1, \ldots, m$ (i.e., the existence of a feasible solution to (7) such that all constraints are satisfied with strict inequalities).*

Solution 2. 1. *If $x \in \mathbb{R}^d$ is not feasible for the problem (7), then there exists some i such that $f_i(x) > 0$. It follows that $\sup_{\lambda \geq 0} L(x, \lambda) = \infty$. On the other hand, if $x \in \mathbb{R}^d$ is feasible for the problem (7), then $\sup_{\lambda \geq 0} L(x, \lambda) = L(x, 0) = f(x)$, which is equal to the objective value of the minimization problem (7).*

2. *The* $\sup \inf \leq \inf \sup$ *inequality*

$$\sup_{y \in \mathcal{Y}} \inf_{x \in \mathcal{X}} f(x, y) \leq \inf_{x \in \mathcal{X}} \sup_{y \in \mathcal{Y}} f(x, y)$$

holds without any assumptions for any function $f : \mathcal{X} \times \mathcal{Y} \to \mathbb{R}$. Indeed, for any $x^* \in \mathcal{X}, y^* \in \mathcal{Y}$ we have

$$\inf_{x \in \mathcal{X}} f(x, y^*) \leq f(x^*, y^*) \leq \sup_{y \in \mathcal{Y}} f(x^*, y).$$

Since the choice of (x^*, y^*) was arbitrary, it follows that

$$\sup_{y \in \mathcal{Y}} \inf_{x \in \mathcal{X}} f(x, y) \leq \inf_{x \in \mathcal{X}} \sup_{y \in \mathcal{Y}} f(x, y),$$

which is what we wanted to show.

3. *Suppose that the pair (x^*, y^*) satisfies the KKT conditions. Then, we have*

$$\begin{aligned}
&\sup_{\lambda \geq 0} L(x^*, \lambda) \\
&= f(x^*) && \text{(primal feasibility)} \\
&= f(x^*) + \sum_{i=1}^m \lambda_i^* f_i(x_i^*) && \text{(complementary slackness)} \\
&= \inf_{x \in \mathbb{R}^d} L(x, \lambda^*) && \text{(stationarity)}.
\end{aligned}$$

Consequently,

$$\sup_{\lambda \geq 0} L(x^*, \lambda) = \inf_{x \in \mathbb{R}^d} L(x, \lambda^*) \leq \inf_{x \in \mathbb{R}^d} \sup_{\lambda \geq 0} L(x, \lambda)$$

so x^* is a minimizer of (8); in particular $p^* = \sup_{\lambda \geq 0} L(x^*, \lambda)$. Likewise, $\lambda^* \geq 0$ (dual feasibility) and

$$\inf_{x \in \mathbb{R}^d} L(x, \lambda^*) = \sup_{\lambda \geq 0} L(x^*, \lambda) \geq \sup_{\lambda \geq 0} \inf_{x \in \mathbb{R}^d} L(x, \lambda)$$

so λ^* is a maximizer of (9), and in particular $d^* = \inf_{x \in \mathbb{R}^d} L(x, \lambda^*)$ and so $d^* = p^*$. This completes the proof of the "if" implication.

Conversely, suppose that x^* is primal-optimal, λ^* is dual-optimal, and the duality gap is zero. We will show that (x^*, λ^*) satisfies the KKT conditions. First, by the assumption that x^* is optimal for the primal problem and λ^* is optimal for the dual problem, it follows that x^* and λ^* satisfy the primal and dual feasibility conditions, respectively. It remains to prove the stationarity and complementary slackness conditions. We have

$$f(x^*) = \sup_{\lambda \geq 0} L(x^*, \lambda) \geq L(x^*, \lambda^*) \geq \inf_{x \in \mathbb{R}^d} L(x, \lambda^*).$$

By the zero duality gap assumption, all the inequalities above are actually equalities. Hence, we have

$$f(x^*) = L(x^*, \lambda^*),$$

from which the complementary slackness condition follows. Finally, the stationarity condition follows from the first-order optimality conditions for

$$L(x^*, \lambda^*) = \inf_{x \in \mathbb{R}^d} L(x, \lambda^*).$$

Our proof is complete.

References

Youngmin Cho and Lawrence Saul. Kernel methods for deep learning. *Advances in neural information processing systems*, 22, 2009.