# Adversarial risk via optimal transport and optimal couplings (Supplementary)

## Proofs from Section 3

**Theorem 1.** *Consider the binary classification setup with $\mathcal{Y} = \{0,1\}$, where the input $x \in \mathcal{X}$ is drawn with equal probability from two distributions $p_0$ (for label 0) and $p_1$ (for label 0). We consider a set of binary classifiers of the form $\mathbb{1}\{x \in A\}$, where $A \subseteq \mathcal{X}$ is a topologically closed set. That is, the classifier corresponding to $A$ assigns the label 1 for all $x \in A$ and the label 0 for all $x \notin A$. Consider the $0-1$ loss function $\ell((x,y),A) = \mathbb{1}\{x \in A, y=0\} + \mathbb{1}\{x \notin A, y=1\}$. The adversarial risk with the data perturbing adversary is given by*

$$R_\epsilon^* = \frac{1}{2}\left[1 - D_\epsilon(p_0,p_1)\right]. \tag{1}$$

*Proof.* Let $A \subseteq \mathcal{X}$ be a closed set such that the classifier declares 1 on $A$ and 0 on $A^c$. The robust risk over the hypothesis class of closed sets is given by

$$\begin{aligned} R_\epsilon^* &= \min_{A \subseteq \mathcal{X}} \frac{1}{2}\left(p_0(A^\epsilon) + p_1\left((A^c)^\epsilon\right)\right) \\ &= \frac{1}{2}\left(1 - \sup_{A \text{ closed}} \left\{p_1\left(A^{-\epsilon}\right) - p_0(A^\epsilon)\right\}\right), \end{aligned}$$

where we define $A^{-\epsilon} := ((A^c)^\epsilon)^c$.

Strassen's theorem is a special case for the Kantorovich duality in the case of a $0-1$ loss. The statement provided below is as in Villani [3, Corollary 1.28]:

**Lemma 0.1.** *Let the input $X$ be drawn from a Polish space $\mathcal{X}$. Let $\Pi(p_0,p_1)$ be the set of all probability measures on $\mathcal{X} \times \mathcal{X}$ with marginals $p_0$ and $p_1$. Then for $\epsilon \geq 0$ and $A \subseteq \mathcal{X}$,*

$$\inf_{\pi \in \Pi(p_0,p_1)} \pi[d(x,x') > \epsilon] = \sup_{A \quad \textit{closed}} \{p_0(A) - p_1(A^\epsilon)\}.$$

For any two distributions, Strassen's theorem states that

$$D_\epsilon(p_0,p_1) = \sup_A \left\{p_0(A) - p_1(A^{2\epsilon})\right\}.$$

To prove the equality $R^*_\epsilon = \frac{1}{2}[1 - D_\epsilon(p_0, p_1)]$, notice that it is enough to prove that for measures $\mu$ and $\nu$,

$$\sup_A \mu(A^{-\epsilon}) - \nu(A^\epsilon) = \sup_A \mu(A) - \nu(A^{2\epsilon}). \tag{2}$$

We may assume that the set achieving the supremum on the right hand side in (2) is a closed set for the following reason: For any measurable set $A \subseteq \mathcal{X}$, $A^\epsilon$ is closed, and $A \subseteq \overline{A}$, where $\overline{A}$ is the closure of $A$ in $\mathcal{X}$. Hence, $\mu(\overline{A}) - \nu(\overline{A}^{2\epsilon}) \geq \mu(A) - \nu(A^{2\epsilon})$. Similarly, we may assume that the set achieving the supremum on the left hand side in (2) is also a closed set. Hence, it suffices to prove (2) for the case when the supremum on both sides is taken over closed sets $A$, instead of the larger class of measurable sets.

In the following lemma, we show that $A^\epsilon$ and $A^{-\epsilon}$ are also closed sets if $A$ is a closed set.

**Lemma 0.2.** *If $A$ is closed, then $A^\epsilon$ is also closed. Similarly, if $A$ is closed, then $A^{-\epsilon}$ is also closed.*

*Proof.* Let $A$ be a closed set and let $B$ be the closed ball of radius $\epsilon$. Let $\{z_i\}_{i\geq 1}$ be a sequence of points in $A^\epsilon$ converging to a limit $z$. We shall show that $z \in A^\epsilon$ as well. Note that every $z_i$ admits an expression $z_i = a_i + b_i$, where $a_i \in A$ and $b_i \in B$. Since $B$ is a compact set, there exists a subsequence among the $\{b_i\}$ sequence that converges to $b^* \in B$. Call the subsequence $\{\tilde{b}_i\}_{i\geq 1}$ such that $\tilde{b}_i \to b^*$ and $|\tilde{b}_i - b^*| < \delta$ for some small $\delta > 0$. Since $z = \tilde{a}_i + \tilde{b}_i = \tilde{a}_i + \tilde{b}_i - b^* + b^*$. Thus, we see that $\tilde{a}_i$ is also bounded within a ball of radius $\delta$ around $z - b^*$, and so there exists a convergence subsequence within the $\{\tilde{a}_i\}$ sequence. Let that subsequence converge to $a^*$. We must have $a^* \in A$ and $b^* \in B$ since $A$ and $B$ are closed. This means $z = a^* + b^*$ must lies in $A^\epsilon$, which shows that $A^\epsilon$ is closed.

Recall that $A^{-\epsilon} = ((A^c)^\epsilon)^c$. Since $A^c$ is an open set, it is enough to show that $C^\epsilon$ is open if $C$ is open. Let $z \in C^\epsilon$. We know that $z = c + b$ for some $c \in C$ and $b \in B$. Consider a small open ball of radius $\delta$ around $c$, called $N_\delta(c)$ that lies entirely in $C$. This is possible since $C$ is assumed to be open. Now observe that $N_\delta(z) \subseteq C^\epsilon$, since $N_\delta(z) = N_\delta(c) + b$. This shows that every point $z \in C^\epsilon$ admits a small ball around it that is contained in $C^\epsilon$, or equivalently, $C^\epsilon$ is open. This completes the proof. □

Next, we show the order in which a set is thickened by $\epsilon$ and thinned by $\epsilon$ affects the size of the original set in opposing ways:

**Lemma 0.3.** *Let $A$ be a closed set. Then $(A^{-\epsilon})^\epsilon \subseteq A$ and $A \subseteq (A^\epsilon)^{-\epsilon}$.*

*Proof.* Notice that a point $x \in A^{-\epsilon}$ if and only if $N_\epsilon(x)$ (which is the ball of radius $\epsilon$ centered at $x$) lies entirely in $A$. If this were not the case, then we could find a $y \in A^c$ such that $d(x,y) \leq \epsilon$, and so $x \in (A^c)^\epsilon$, which implies $x \notin ((A^c)^\epsilon)^c = A^{-\epsilon}$. This observation implies that $(A^{-\epsilon})^\epsilon \subseteq A$.

Similarly, a point $x \in (A^\epsilon)^{-\epsilon}$ if and only if $N_\epsilon(x) \in A^\epsilon$. By definition of $A^\epsilon$, every point $x \in A$ satisfies $N_\epsilon(x) \in A^\epsilon$. Thus, if $x \in A$ then $x \in (A^\epsilon)^{-\epsilon}$. Equivalently, $A \subseteq (A^\epsilon)^{-\epsilon}$. □

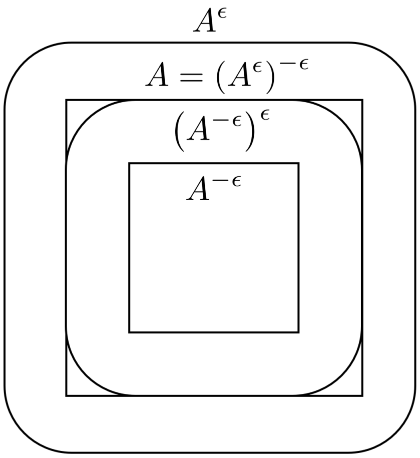

Figure 1: Illustration of $A, A^\epsilon, A^{-\epsilon}, (A^\epsilon)^{-\epsilon}$, and $(A^{-\epsilon})^\epsilon$ for a square in $(\mathbb{R}^2, \|\cdot\|_2)$. Observe that $(A^{-\epsilon})^\epsilon \subseteq A$ and $A \subseteq (A^\epsilon)^{-\epsilon}$.

Figure 1 illustrates the above lemma when $A$ is a square in $\mathbb{R}^2$ with the Euclidean distance metric. Using the above lemmas, we now establish the desired equality from (2): We have the sequence of inequalities

$$\begin{aligned}\sup_{A \text{ closed}} \mu(A) - \nu(A^{2\epsilon}) &\overset{(a)}{\geq} \sup_{A \text{ closed}} \mu(A^{-\epsilon}) - \nu((A^{-\epsilon})^{2\epsilon}) \\ &\overset{(b)}{\geq} \sup_{A \text{ closed}} \mu(A^{-\epsilon}) - \nu(A^\epsilon).\end{aligned}$$

Here, $(a)$ follows because $A^{-\epsilon}$ is contained in the set of all closed sets by Lemma 0.2. Inequality $(b)$ follows using Lemma 0.3, since

$$(A^{-\epsilon})^{2\epsilon} = [(A^{-\epsilon})^\epsilon]^\epsilon \subseteq A^\epsilon,$$

and so $\nu((A^{-\epsilon})^{2\epsilon}) \leq \nu(A^\epsilon)$.

For the other direction, notice that

$$\begin{aligned}\sup_{A \text{ closed}} \mu(A^{-\epsilon}) - \nu(A^{\epsilon}) &\overset{(a)}{\geq} \sup_{A \text{ closed}} \mu((A^{\epsilon})^{-\epsilon}) - \nu((A^{\epsilon})^{\epsilon}) \\ &\overset{(b)}{\geq} \sup_{A \text{ closed}} \mu(A) - \nu(A^{2\epsilon}).\end{aligned}$$

Here, $(a)$ follows because $A^{\epsilon}$ is a closed set according to Lemma 0.2. To see $(b)$, first note that $(A^{\epsilon})^{\epsilon} = A^{2\epsilon}$, and so $\nu((A^{\epsilon})^{\epsilon}) = \nu(A^{2\epsilon})$. Moreover, Lemma 0.3 states that

$$A \subseteq (A^{\epsilon})^{-\epsilon},$$

and so $\mu(A) \leq \mu((A^{\epsilon})^{-\epsilon})$. This completes the proof. □

**Comparison with Bhagoji et al.[1]:** We note that a similar result was obtained recently in Bhagoji et al.[1]. While the duality in their proof was established for a larger hypothesis class of measurable sets $A$, our proof relies on Strassen's duality theorem and properties of closed sets. Using closed sets and directly using Strassen's theorem allows us to considerably simplify the technical details as compared with Bhagoji et al.[1].

**Corollary 0.1.** *Under the setup considered in Theorem 1, we have the following bound for $p \geq 1$:*

$$R^*_{\epsilon} \geq \frac{1}{2}\left[1 - \left(\frac{W_p(p_0, p_1)}{2\epsilon}\right)^p\right]. \tag{3}$$

*Proof.* From Theorem 1, we have

$$R^*_{\epsilon} = \frac{1}{2}\left[1 - \inf_{\pi \in \Pi(\mu,\nu)} \mathbb{E}_{(x,x')\sim\pi}[\mathbb{1}\{d(x,x') > 2\epsilon\}]\right].$$

For $p \geq 1$ and any $\pi \in \Pi(\mu,\nu)$, we have the following:

$$\begin{aligned}\mathbb{E}_{(x,x')\sim\pi}[\mathbb{1}\{d(x,x') > 2\epsilon\}] &= \mathbb{E}_{(x,x')\sim\pi}[\mathbb{1}\{d(x,x')^p > (2\epsilon)^p\}] \\ &\leq \mathbb{E}_{(x,x')\sim\pi}\left[\left(\frac{d(x,x')}{2\epsilon}\right)^p\right],\end{aligned}$$

where the last inequality follows from Markov's inequality. Therefore,

$$R^*_{\epsilon} = \frac{1}{2}\left[1 - \inf_{\pi \in \Pi(\mu,\nu)} \mathbb{E}_{(x,x')\sim\pi}\left[\left(\frac{d(x,x')}{2\epsilon}\right)^p\right]\right] = \frac{1}{2}\left[1 - \left(\frac{W_p(p_0, p_1)}{2\epsilon}\right)^p\right].$$

□

## Proofs from Section 4

**Theorem 2.** *Let $\mu$ and $\nu$ be finite positive measures on $\mathbb{R}^d$ that are absolutely continuous with respect to the Lebesgue measure and have bounded supports. Then $D_\epsilon(\mu,\nu) = 0$ if and only if $W_\infty(\mu,\nu) \leq 2\epsilon$. Here, $W_\infty(\mu,\nu) = \lim_{p\to\infty} W_p(\mu,\nu)$.*

*Proof.* An alternate description of $W_\infty(\mu,\nu)$ as per Givens and Shortt [2] is

$$W_\infty(\mu,\nu) = \inf\{\delta > 0 \ : \ \mu(A) \leq \nu(A^\delta) \text{ for all measurable } A\}.$$

Naturally, if $W_\infty(\mu,\nu) \leq 2\epsilon$, then $\mu(A) \leq \nu(A^{2\epsilon})$ for all closed sets $A$. Hence,

$$D_\epsilon(\mu,\nu) = \sup_{A \text{ closed}} \mu(A) - \nu(A^{2\epsilon}) \leq 0.$$

Since $D_\epsilon(\mu,\nu) \geq 0$, we conclude that $D_\epsilon(\mu,\nu) = 0$.

For the reverse direction, suppose that $D_\epsilon(\mu,\nu) = 0$. This means there exists a sequence of couplings $\{\pi\}_{i\geq 1}$ such that $\mathbb{E}_{\pi_i} c_\epsilon(x,x') \to 0$ where $\pi_i \in \Pi(\mu,\nu)$. Equivalently, the probability $\alpha_i := \mathbb{P}_{\pi_i}(\|x-x'\| > 2\epsilon) \to 0$ as $i \to \infty$. For any fixed $p \geq 1$, we have the inequality

$$\begin{aligned} W_p(\mu,\nu) &\leq \left(\mathbb{E}_{\pi_i}\|x-x'\|^p\right)^{1/p} \\ &\leq \left((2\epsilon)^p(1-\alpha_i) + C^p\alpha_i\right)^{1/p}, \end{aligned}$$

where $C = \sup\|x-x'\|$ which is a constant since $\mu$ and $\nu$ are assumed to have bounded supports. Pick $i_0$ large enough so that $(2\epsilon)^p(1-\alpha_{i_0}) > C^p\alpha_{i_0}$. Now letting $p \to \infty$ and calculating the limits, we conclude that $W_\infty(\mu,\nu) \leq 2\epsilon$. □

**Theorem 3.** *Let $\mu$ and $\nu$ be finite positive measures on $\mathbb{R}$ that are absolutely continuous with respect to the Lebesgue measure with Radon-Nikodyn derivatives $f(\cdot)$ and $g(\cdot)$, respectively. The cumulative distribution function (cdf) of $\mu$ is defined as $F(x) = \mu((-\infty,x])$, and for $t \in [0,1]$, the inverse cdf (or quantile function) is defined as $F^{-1}(t) = \inf\{x \in \mathbb{R} \ : \ F(x) \geq t\}$. The cdf $G(\cdot)$ and inverse cdf $G^{-1}(\cdot)$ are defined analogously. Suppose that $\mu(\mathbb{R}) = \nu(\mathbb{R}) = U$. Then $D_\epsilon(\mu,\nu) = 0$ if and only if $\|F^{-1} - G^{-1}\|_\infty \leq 2\epsilon$.*

*Proof.* Consider the monotone transport map from $\mu$ to $\nu$ given by $T(x) = G^{-1}(F(x))$ for $x \in \mathbb{R}$ [3]. We shall show that this map satisfies $|T(x)-x| \leq 2\epsilon$ for all $x \in \mathbb{R}$, and so the optimal transport cost $D_\epsilon$ must be 0. To see this, note that

$$\begin{aligned} T(x) - x &= G^{-1}(F(x)) - x \\ &\leq F^{-1}(F(x)) + 2\epsilon - x \\ &= 2\epsilon, \end{aligned}$$

where the last equality is in the $\mu$-almost sure sense. A similar argument shows $x - T(x) \leq 2\epsilon$, and thus $|T(x) - x| \leq 2\epsilon$.

For the converse, suppose that there exists a $t_0 \in (0, 1)$ such that $G^{-1}(t_0) - F^{-1}(t_0) > 2\epsilon$. Equivalently, $G^{-1}(t_0) > F^{-1}(t_0) + 2\epsilon$. Applying the $G$ function on both sides,

$$t_0 > G(F^{-1}(t_0) + 2\epsilon).$$

Consider the set $\tilde{A} = (-\infty, F^{-1}(t_0)]$. For this set, notice that

$$\nu(\tilde{A}^{2\epsilon}) = \nu((-\infty, F^{-1}(t_0) + 2\epsilon]) = G(F^{-1}(t_0) + 2\epsilon).$$

Thus, we have

$$\begin{aligned} D_\epsilon(\mu, \nu) &= \sup_A \mu(A) - \nu(A^{2\epsilon}) \\ &\geq \mu(\tilde{A}) - \nu(\tilde{A}^{2\epsilon}) \\ &= t_0 - G(F^{-1}(t_0) + 2\epsilon) \\ &> 0. \end{aligned}$$

A similar argument may also be made for the case when $F^{-1}(t_0) - G^{-1}(t_0) > 2\epsilon$. □

**Corollary 0.2.** *Let $\mu$ and $\nu$ be as in Theorem 3. Suppose that for every $x \in \mathbb{R}$, we have $F(x) \geq G(x)$ and $F(x) \leq G(x + 2\epsilon)$. Then $D_\epsilon(\mu, \nu) = 0$.*

*Proof.* Applying the $G^{-1}$ function to both sides of both inequalities, we arrive at

$$T(x) \leq x, \quad \text{and} \quad T(x) \geq x + 2\epsilon.$$

This gives $|T(x) - x| \leq 2\epsilon$ for all $x$, which concludes the proof. □

Before we examine the case of Gaussian distributions with identical means but different variances, we define notions of transport for measures with unequal masses.

**Definition 1.** [Optimal transport cost for general measures] Let $\mu$ and $\nu$ be as in Theorem 3. Suppose that $\mu(\mathbb{R}) = U$ and $\nu(\mathbb{R}) = V$ and $U \leq V$. Let $\nu'$ be a measure on $\mathbb{R}$ with Radon-Nikodyn derivative $g'$ such that $\nu'(\mathbb{R}) = U$. We say $\nu' \subseteq \nu$, or $\nu'$ is contained in $\nu$, if $g(x) \geq g'(x)$ $\nu$-almost surely. Then the optimal transport cost $D_\epsilon(\mu, \nu)$ is defined as

$$D_\epsilon(\mu, \nu) = \inf_{\nu' \subseteq \nu} D_\epsilon(\mu, \nu').$$

Note that the amount of mass being moved is $\min(U, V) = U$.

**Lemma 0.4.** *Let $\mu$ and $\nu$ be as in Theorem 3. Assume that $\mu(\mathbb{R}) = U$ and $\nu(\mathbb{R}) = V$. Suppose the following conditions hold:*

1. *The support of $g$ is $[a, +\infty)$ and the support of $f$ is $[a + 2\epsilon, +\infty) =: [a', +\infty)$.*
2. *For all $x \in \mathbb{R}$, we have $g(x) \leq f(x + 2\epsilon)$.*

*Then $D_\epsilon(\mu, \nu) = 0$. A similar result holds if the supports of $g$ and $f$ are $(-\infty, -a]$ and $(-\infty, -a - 2\epsilon]$, and $f(-x - 2\epsilon) \geq g(-x)$.*

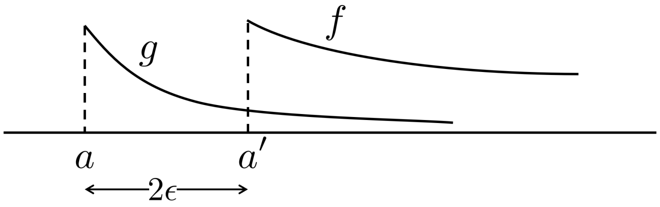

Figure 2: Figure illustrating the conditions in Lemma 0.4.

*Proof.* Consider the transport map $T(x) = x + 2\epsilon$ applied to $\nu$. This map has the effect of "translating" the measure $\nu$ by $2\epsilon$ to the right. Call this translated measure $\eta$. Since $f(x) \geq g(x - 2\epsilon)$, it is immediate that $\eta \subseteq \mu$. Moreover, the transport cost is $D_\epsilon(\nu, \eta) = 0$. This shows that $D_\epsilon(\mu, \nu) = 0$. □

**Lemma 0.5.** *Let $\mu$ and $\nu$ be as in Theorem 3. Assume that $\mu(\mathbb{R}) = \nu(\mathbb{R}) = U$. Suppose the following conditions hold (see Figure 3 for an illustration):*

1. *Let $a, b \in \mathbb{R}$ be such that $b - a > 2\epsilon$. The support of $f$ is $[a, b]$ and the support of $g$ is $[a', b] := [a + 2\epsilon, b]$.*

2. *There exists* $t \in [a, b]$ *such that* $f(x) \geq g(x)$ *for* $x \in [a, t)$*, and* $f(x) \leq g(x)$ *for* $x \in (t, b]$*.*

3. *Let* $\tilde{g}(x) = g(x+2\epsilon)$*. Note that* $\tilde{g}$ *is supported on* $[a, b-2\epsilon]$*. There exists* $\tilde{t} \in [a, b-2\epsilon]$ *such that* $f(x) \leq \tilde{g}(x)$ *for* $x \in [a, \tilde{t})$*, and* $f(x) \geq \tilde{g}(x)$ *for* $x \in (\tilde{t}, b-2\epsilon]$*.*

*Then* $D_\epsilon(\mu, \nu) = 0$*. A mirror image of this result also holds:* $D_\epsilon(\mu, \nu) = 0$ *when the support of* $f$ *is* $[b, c+2\epsilon]$*, that of* $g$ *is* $[b, c]$*, and* $f(x) \leq g(x)$ *for* $x \in [b, t)$ *and* $f(x) \geq g(x)$ *for* $x \in [t, c+2\epsilon]$*; and for* $\tilde{g}(x) = g(x+2\epsilon)$ *we have* $f(x) \geq \tilde{g}(x)$ *for* $x \in [b+2\epsilon, \tilde{t})$ *and* $f(x) \leq g(x)$ *for* $x \in [\tilde{t}, c+2\epsilon]$ *.*

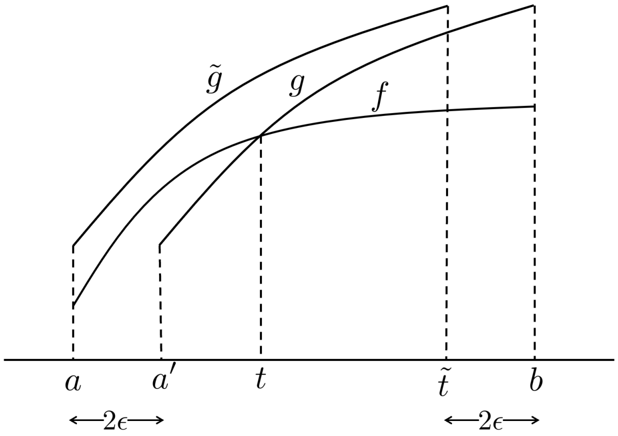
$\tilde{g}$
$g$
$f$
$a$
$a'$
$t$
$\tilde{t}$
$b$
$\longleftarrow 2\epsilon \longrightarrow$
$\longleftarrow 2\epsilon \longrightarrow$

Figure 3: Figure illustrating the conditions in Lemma 0.5.

*Proof.* We first prove $F(x) \geq G(x)$. To see this, consider $H(x) = F(x) - G(x)$. Since the derivative of $H$ is $f - g$, it must be that $H$ is increasing from $[a, t)$ and decreasing from $[t, b]$. Also, we have $H(a) = H(b) = 0$, and so the function $H$ must be non-negative in $[a, b]$. Equivalently, we must have $F(x) \geq G(x)$ for $x \in \mathbb{R}$. We now prove $F(x) \leq G(x + 2\epsilon)$. Consider $\tilde{H}(x) = F(x) - \tilde{G}(x)$. By condition (3), the derivative of this function is negative from $[a, \tilde{t}]$ and positive from $[\tilde{t}, b]$. Thus, the function $\tilde{H}$ decreases on the interval $[a, \tilde{t})$ and increases on the interval $[\tilde{t}, b]$. Note that since $\tilde{H}(a) = \tilde{H}(b) = 0$, the function $\tilde{H}$ must be non-positive in the interval $[a, b]$. Thus, we have $F(x) \leq G(x + 2\epsilon)$. Applying Corollary 0.2 concludes the proof. □

Our next lemma is specific to Gaussian pdfs:

**Lemma 0.6.** *Let $f$ and $g$ be Gaussian pdfs corresponding to $\mathcal{N}(\mu_1, \sigma_1^2)$ and $\mathcal{N}(\mu_2, \sigma_2^2)$, respectively. Assume $\sigma_1^2 > \sigma_2^2$. Then the equation $f(x) - g(x) = 0$ has exactly two solutions $s_1 < \mu_2 < s_2$.*

*Proof.* By scaling and translating, we may set $\mu_2 = 0$ and $\sigma_2^2 = 1$. Solving $f(x) - g(x) = 0$ is equivalent to solving the quadratic equation

$$\frac{x^2}{2} - \frac{(x-\mu_1)^2}{2\sigma_1^2} = \log \sigma_1.$$

Simplifying, we wish to solve

$$x^2(\sigma_1^2 - 1) + 2\mu_1 x - (\mu_1^2 + 2\sigma_1^2 \log \sigma_1) = 0.$$

Since $\sigma_1 > 1$, the above quadratic has two distinct roots: one negative and one positive. This proves the claim. □

We shall call the two points where $f$ and $g$ intersect as the left and right intersection points.

**Theorem 4.** *Let $\mu$ and $\nu$ be the Gaussian measures $\mathcal{N}(0, \sigma_1^2)$ and $\mathcal{N}(0, \sigma_2^2)$, respectively. Assume $\sigma_1^2 > \sigma_2^2$ without loss of generality. Let $m > 0$ be such that $f(m+\epsilon) = g(m-\epsilon)$. Let $A = (-\infty, -m] \cup [m, +\infty)$. Then the optimal transport cost between $\mu$ and $\nu$ is given by*

$$\begin{aligned} D_\epsilon(\mu, \nu) &= \mu(A^{-\epsilon}) - \nu(A^\epsilon) \\ &= 2Q\left(\frac{m+\epsilon}{\sigma_1}\right) - 2Q\left(\frac{m-\epsilon}{\sigma_2}\right). \end{aligned}$$

*The corresponding robust risk is*

$$R_\epsilon^* = \frac{1 - \mu(A^{-\epsilon}) + \nu(A^\epsilon)}{2}.$$

*Moreover, if $\mu$ corresponds to hypothesis 1, the optimal robust classifier decides 1 on the set $A$.*

*Proof.* We shall propose a map that transports $\mu$ to $\nu$. (See Figure 4 for an illustration.) Consider $r \in (0, m-\epsilon)$ whose value will be provided later. First, we partition $\mathbb{R}$ into the five regions for $\mu$ and $\nu$, as shown in Table 1. For $\mu$, these partitions are $(-\infty, -m-\epsilon]$, $(-m-\epsilon, -r]$, $(-r, +r)$, $[r, m+\epsilon)$, and $[m+\epsilon, \infty)$. Let $\mu$ restricted to these intervals be $\mu_{--}$, $\mu_-$, $\mu_0$, $\mu_+$, and $\mu_{++}$, respectively. The measure $\nu$ is also partitioned five ways, but the intervals used in this case are slightly modified to be $(-\infty, -m+\epsilon]$,

| | |
|---|---|
| $\mu_{--}$ | $(-\infty, -m-\epsilon]$ |
| $\mu_{-}$ | $(-m-\epsilon, -r]$ |
| $\mu_0$ | $(-r, +r)$ |
| $\mu_{+}$ | $[r, m+\epsilon)$ |
| $\mu_{++}$ | $[m+\epsilon, \infty)$ |

| | |
|---|---|
| $\nu_{--}$ | $(-\infty, -m+\epsilon]$ |
| $\nu_{-}$ | $(-m+\epsilon, -r]$ |
| $\nu_0$ | $(-r, +r)$ |
| $\nu_{+}$ | $[r, m-\epsilon)$ |
| $\nu_{++}$ | $[m-\epsilon, \infty)$ |

Table 1: The real line is partitioned into five regions for $\mu$ and $\nu$, as shown in the table.

$(-m+\epsilon, -r]$, $(-r, r)$, $[r, m-\epsilon)$, and $[m-\epsilon, +\infty)$. Call $\nu$ restricted to these intervals $\nu_{--}$, $\nu_{-}$, $\nu_0$, $\nu_{+}$, and $\nu_{++}$, respectively.

The transport plan from $\mu$ to $\nu$ will consist of five maps transporting $\mu_{--} \to \nu_{--}$, $\mu_{-} \to \nu_{-}$, $\mu_0 \to \nu_0$, $\mu_{+} \to \nu_{+}$, and $\mu_{++} \to \nu_{++}$. In each case, we plan to show that $D_\epsilon(\mu_*, \nu_*) = 0$, where $*$ ranges over all possible subscripts. Note that these measures do not necessarily have identical masses, and thus by Definition 1, we are transporting a quantity of mass equal to the minimum mass among the two measures. For this reason, even though the transport cost is $D_\epsilon(\mu_*, \nu_*) = 0$, it does not mean $D_\epsilon(\mu, \nu) = 0$.

Consider $\mu_{++}$ and $\nu_{++}$. We have $f(m+\epsilon) = g(m-\epsilon)$ by the choice of $m$. We argue that for any $t \geq 0$, we must have $f(m+\epsilon+t) \geq g(m-\epsilon+t)$. This is because any two Gaussian pdfs can intersect in at most two points. By Lemma 0.6, the $\epsilon$-shifted Gaussian pdfs $f(x+\epsilon)$ and $g(x-\epsilon)$ have $m$ as their right intersection point, and there are no additional points of intersection to the right of $m$. Since the tail of $f$ is heavier, it means that $f(m+\epsilon+t) \geq g(m-\epsilon+t)$ for all $t \geq 0$. By Lemma 0.4, we can now conclude $D_\epsilon(\mu_{++}, \nu_{++}) = 0$. A similar argument also shows $D_\epsilon(\mu_{--}, \nu_{--}) = 0$.

Before we consider $\mu_{-}$ and $\nu_{-}$, we first define $r$ as follows: Pick $r > 0$ such that $\mu([-m-\epsilon, -r)) = \nu([-m+\epsilon, -r))$. To see that such an $r$ must exist, consider the functions $a(t) := \mu([-m-\epsilon, t))$ and $b(t) := \nu([-m+\epsilon, t))$ as $t$ ranges over $(-m+\epsilon, 0)$. When $t = -m+\epsilon$, we have $a(t) > b(t) = 0$. When $t = 0$, we have $a(t) = 1/2 - \mu_{--}(\mathbb{R}) < b(t) = 1/2 - \nu_{--}(\mathbb{R})$. Thus, there must exist a $t_0 \in (-m+\epsilon, 0)$ such that $a(t_0) = b(t_0)$. Pick the smallest (i.e., the leftmost) such $t_0$, and set $-r = t_0$. Call $f(\cdot)$ restricted to $[-m-\epsilon, -r)$ and $g(\cdot)$ restricted to $[-m+\epsilon, -r)$ as $f_{-}$ and $g_{-}$, respectively, and their corresponding cdfs $F_{-}$ and $G_{-}$, respectively. We claim that $\mu_{-}$ and $\nu_{-}$ satisfy all three conditions from Lemma 0.5. Since the supports of $f_{-}$ and $g_{-}$ are $[-m-\epsilon, -r)$ and $[-m+\epsilon, -r)$, condition (1) is immediately verified. To check condition (2), we break up the interval $[-m-\epsilon, -r)$ into two parts: $[-m-\epsilon, -s)$ and $[-s, -r)$, where $s$ is such that $f(-s) = g(-s)$. Observe that $f_- \geq g_-$ on $[-m-\epsilon, -s)$, whereas $f_- \leq g_-$ on $[-s, -r)$. This shows that condition (2) is satisfied. We have $g_-(-m+\epsilon) = f_-(-m-\epsilon)$. Again, using Lemma 0.6 the $2\epsilon$-shifted Gaussian pdf $f(x-2\epsilon)$ and $g(x)$ have $-m+\epsilon$ as their left intersection point, and the right intersection point is to the right of 0. Thus, we have $f(x-2\epsilon) \leq g(x)$ for all $x \in [-m+\epsilon, 0] \supseteq [-m+\epsilon, r)$. Using this domination, we conclude that $f_- \leq \tilde{g}_-$ in the interval $[-m-\epsilon, -r-2\epsilon)$ and $f_- \geq g_- = 0$ in the interval $(-r-2\epsilon, -r]$, and so condition (3) is satisfied. Applying Lemma 0.5, we conclude $D_\epsilon(\mu_-, \nu_-) = 0$. An essentially identical argument may be used to show $D_\epsilon(\mu_+, \nu_+) = 0$. The minor difference being that $r$ is chosen to satisfy $\mu([r, m+\epsilon)) = \nu([r, m-\epsilon))$, and the mirror image of Lemma 0.5 is applied.

Finally, consider the interval $(-r, +r)$. In this interval, $f(x) \leq g(x)$ for every point. Hence, a transport map from $\mu_0$ to $\nu_0$ is obtained by simply considering the identity function. Any remaining mass in $\mu$ is moved to $\nu$ arbitrarily, incurring a cost of at most 1 per unit mass. The total cost of transport is then upper-bounded by

$$\begin{aligned}
D_\epsilon(\mu, \nu) &\leq 1 - [\min(\mu_{--}, \nu_{--}) + \min(\mu_-, \nu_-) + \min(\mu_0, \nu_0) \\
&\qquad + \min(\mu_+, \nu_+) + \min(\mu_{++}, \nu_{++})] \\
&= 1 - [\nu_{--} + \mu_- + \mu_0 + \mu_+ + \nu_{++}] \\
&= 1 - \mu([-m-\epsilon, m+\epsilon]) - 2\nu([m-\epsilon, \infty)) \\
&= \mu(A^{-\epsilon}) - \nu(A^\epsilon) \\
&= 2Q\left(\frac{m+\epsilon}{\sigma_1}\right) - 2Q\left(\frac{m-\epsilon}{\sigma_2}\right).
\end{aligned}$$

where for brevity we have denoted $\mu_*(\mathbb{R})$ as $\mu_*$. However, we also have

$$D_\epsilon(\mu, \nu) \geq \mu(A^{-\epsilon}) - \nu(A^\epsilon).$$

The lower and upper bounds match and this concludes the proof. The robust risk $R_\epsilon^*$ is given by Theorem 1. The robust risk of the classifier that decides 1 on the set $A$ is easily seen to be $R_\epsilon^*$. □

**Theorem 5.** *Let $\mu$ and $\nu$ be Gaussian measures $\mathcal{N}(\mu_1, \sigma_1^2)$ and $\mathcal{N}(\mu_2, \sigma_2^2)$ respectively. Assume $\sigma_1^2 > \sigma_2^2$ without loss of generality. Let $m_1, m_2 > 0$ be such that $f(-m_1-\epsilon) = g(-m_1+\epsilon)$ and $f(m_2+\epsilon) = g(m_2-\epsilon)$. Let $A = (-\infty, -m_1] \cup [m_2, \infty)$. Then the optimal transport cost between $\mu$ and $\nu$ is given by*

$$D_\epsilon(\mu, \nu) = \mu(A^{-\epsilon}) - \nu(A^\epsilon).$$

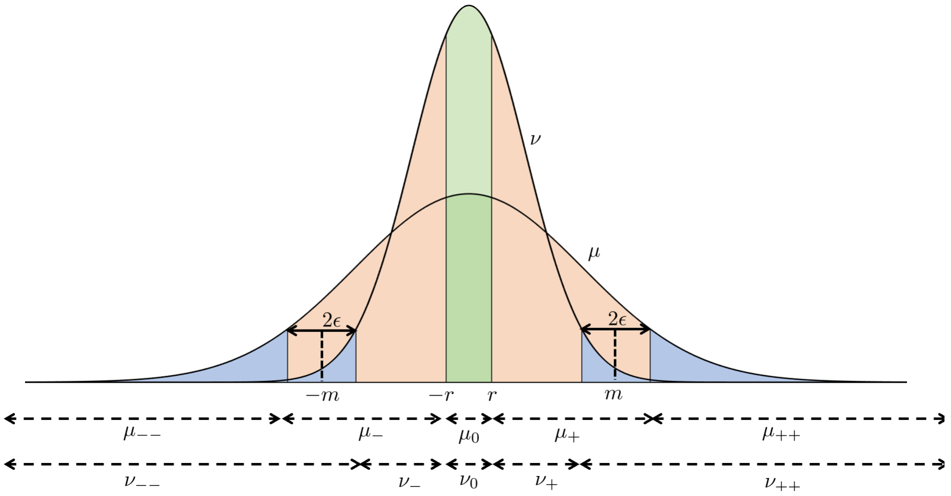

Figure 4: Optimal transport coupling for centered Gaussian distributions $\mu$ and $\nu$. As in the proof of Theorem 4, we divide the real line into five regions. The transport plan from $\mu$ to $\nu$ consists of five maps transporting $\mu_{--} \to \nu_{--}$ (blue regions to the left), $\mu_- \to \nu_-$ (orange regions to the left), $\mu_0 \to \nu_0$ (green regions in the middle), $\mu_+ \to \nu_+$ (orange regions to the right), and $\mu_{++} \to \nu_{++}$ (blue regions to the right).

*Consequently, the robust risk is given by*

$$R^*_\epsilon = \frac{1}{2}(1 - \mu(A^{-\epsilon}) + \nu(A^\epsilon)).$$

*If $\mu$ corresponds to hypothesis 1, the optimal robust classifier decides 1 on the set $A$.*

*Proof.* As in the proof of Theorem 4, we shall divide the real line into five regions as shown in Table 2 where we define $r_1$ and $r_2$ shortly. Using an identical strategy as in Theorem 4, we conclude $D_\epsilon(\mu_{--}, \nu_{--}) = D_\epsilon(mu_{++}, \nu_{++}) = 0$. Define $r_1$ as the leftmost point where $\mu([-m_1 - \epsilon, r_1)) = \nu([-m_1 + \epsilon, r_1))$. Similarly, define $r_2$ to be the rightmost point such that $\mu([r_2, m_2 + \epsilon)) = \nu([r_2, m_2 - \epsilon))$. We shall now prove $D_\epsilon(\mu_-, \nu_-) = 0$ by using Lemma 0.5. Verifying conditions (1) and (2) is exactly as in that of Theorem 4. The novel component of this proof is verifying condition (3), since the domination used in the proof of Theorem 4 does not work in this case due to the asymmetry. Consider the pdfs $f_-(x)$ and $g_-(x + 2\epsilon)$. These two pdfs, being restrictions of Gaussian pdfs to suitable intervals, may only intersect in at most two points. One of these points of intersection

| | |
|---|---|
| $\mu_{--}$ | $(-\infty, -m_1 - \epsilon]$ |
| $\mu_{-}$ | $(-m_1 - \epsilon, -r_1]$ |
| $\mu_0$ | $(-r_1, +r_2)$ |
| $\mu_{+}$ | $[r_2, m_2 + \epsilon)$ |
| $\mu_{++}$ | $[m_2 + \epsilon, \infty)$ |

| | |
|---|---|
| $\nu_{--}$ | $(-\infty, -m_1 + \epsilon]$ |
| $\nu_{-}$ | $(-m_1 + \epsilon, -r_1]$ |
| $\nu_0$ | $(-r_1, +r_2)$ |
| $\nu_{+}$ | $[r_2, m_2 - \epsilon)$ |
| $\nu_{++}$ | $[m_2 - \epsilon, \infty)$ |

Table 2: The real line is partitioned into five regions for $\mu$ and $\nu$ as shown in the table.

is $-m_1 - \epsilon$ by the choice of $m_1$, so there can be at most one other point of intersection in the interval $[-m_1 - \epsilon, -r_1 - 2\epsilon]$. Note that there may be no point of intersection in this interval. However, the key observation is that in both cases, condition (3) continues to be satisfied. To see this, suppose that there is a point of interaction $\tilde{t}$. In this case, $f_- \leq \tilde{g}_-$ in $[-m_1 - \epsilon, \tilde{t})$, and $f_- \geq g_-$ in $(\tilde{t}, -r_1]$. If there is no point of intersection, then $f_- \leq \tilde{g}_-$ in $[-m_1 - \epsilon, -r_1 - 2\epsilon)$, and $f_- \geq g_- = 0$ in $(-r_1 - 2\epsilon, -r_1]$. This verifies condition (3). Using Lemma 0.5, we conclude $D_\epsilon(\mu_-, \nu_-) = 0$. An identical approach gives $D_\epsilon(\mu_+, \nu_+) = 0$. Since $f(x) \leq g(x)$ for all points in the interval $(-r_1, r_2)$, the identity map may be used to conclude $D_\epsilon(\mu_0, \nu_0) = 0$.

Any remaining mass in $\mu$ is moved to $\nu$ arbitrarily, incurring a cost of at most 1 per unit mass. The total cost of transport is then upper-bounded by

$$\begin{aligned}
D_\epsilon(\mu, \nu) &\leq 1 - [\min(\mu_{--}, \nu_{--}) + \min(\mu_-, \nu_-) + \min(\mu_0, \nu_0) \\
&\qquad + \min(\mu_+, \nu_+) + \min(\mu_{++}, \nu_{++})] \\
&= 1 - [\nu_{--} + \mu_- + \mu_0 + \mu_+ + \nu_{++}] \\
&= 1 - \mu([-m_1 - \epsilon, m_2 + \epsilon]) - \nu((-\infty, -m_1 + \epsilon)) - \nu([m_2 - \epsilon, \infty)) \\
&= \mu(A^{-\epsilon}) - \nu(A^\epsilon),
\end{aligned}$$

where for brevity we have denoted $\mu_*(\mathbb{R})$ as $\mu_*$. The rest of the proof is identical to that of Theorem 4. □

**Theorem 6** (Uniform distributions)**.** *Let $\mu$ and $\nu$ be uniform measures on closed intervals $I$ and $J$ respectively. Without loss of generality, we assume $|I| \leq |J|$. Then the optimal robust risk is $\nu(I^{2\epsilon})$ and the optimal classifier is given by $A = I^\epsilon$.*

*Proof.* Like in the proof for Theorem 4, we prove Theorem 6 by partitioning the real line into several regions for $\mu$ and $\nu$, and transporting mass between these regions. Figure 5 shows the optimal coupling for the case when $I^{2\epsilon} \subseteq J$.

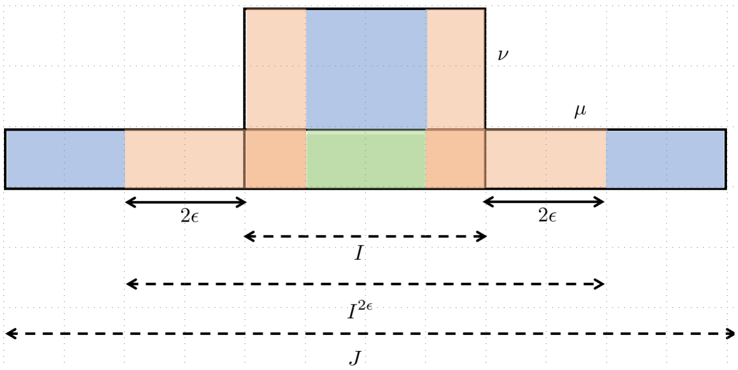

Figure 5: Optimal coupling for two uniform distributions. The region shaded in green is kept in place (at no cost). The two regions shaded in orange are transported monotonically from either side at a cost not exceeding $2\epsilon$ per unit mass. The remaining region in blue is moved at the cost of 1 per unit mass.

We first prove a lower bound. Choose the set $A = I$, we have that

$$D_\epsilon(\mu, \nu) \geq \mu(A) - \nu(A^{2\epsilon}) = 1 - \nu(I^{2\epsilon}).$$

To establish the upper bound, we need to find a coupling that transports $\mu$ to $\nu$ such that the cost of transportation is bounded above by $1 - \nu(I^{2\epsilon})$. Let $I = [i_1, i_2]$ and $J = [j_1, j_2]$. As shown in Figure 5, we pick $t_1 \in [i_1, i_2]$ such that $\nu([i_1 - 2\epsilon, t_1]) = \mu([i_1, t_1])$. Similarly, pick $t_2 \in [t_1, i_2]$ such that $\mu([t_2, i_2]) = \nu([t_2, i_2 + 2\epsilon])$.

We now present a plan to transport $\nu$ to $\mu$. This plan consists of four mini-plans:

1. First, transport the mass $\nu([i_1 - 2\epsilon, t_1])$ to $\mu([i_1, t_1])$ using a monotone transport map.

2. Then transport the mass from $\nu([t_2, i_2 + 2\epsilon])$ to $\mu([t_2, i_2])$ using a monotone transport map.

3. Keep any mass in $[t_1, t_2]$ in its place.

4. Move any remaining mass in $\nu$ arbitrarily to the necessary places in $\mu$.

The key point to note is that in maps (1) and (2), the total distance moved by every unit of mass is at most $2\epsilon$. The proof of this part is along similar lines to that of Theorem 5. Thus, the transport cost in steps (1) and (2) is 0. Naturally, the transport cost in (3) is 0. This means that all the mass in the interval $[i_1 - 2\epsilon, i_2 + 2\epsilon]$ can be transported into $\mu$ for zero cost. The total cost of transportation is therefore at most $1 - \nu([i_1 - 2\epsilon, i_2 + 2\epsilon])$, which matches our lower bound. It is easily checked that the error attained by the proposed classifier also matches the bound, which completes the proof. □

## References

[1] A. N. Bhagoji, D. Cullina, and P. Mittal. Lower bounds on adversarial robustness from optimal transport. *Conference on Neural Information Processing Systems*, 2019.

[2] C. R. Givens and R. M. Shortt. A class of Wasserstein metrics for probability distributions. *The Michigan Mathematical Journal*, 31(2):231–240, 1984.

[3] C. Villani. *Topics in Optimal Transportation.* American Mathematical Soc., 2003.