MIT OpenCourseWare
http://ocw.mit.edu

18.175 Theory of Probability
Fall 2008

For information about citing these materials or our Terms of Use, visit: http://ocw.mit.edu/terms.

Section 21

Prekopa-Leindler inequality, entropy and concentration.

In this section we will make several connections between the Kantorovich-Rubinstein theorem and other classical objects. Let us start with the following classical inequality.

Theorem 51 *(Prekopa-Leindler) Consider nonnegative integrable functions $w, u, v : \mathbb{R}^n \to [0,\infty)$ such that for some $\lambda \in (0,1)$,*

$$w(\lambda x + (1-\lambda)y) \geq u(x)^{\lambda} v(y)^{1-\lambda} \quad \textit{for all} \quad x, y \in \mathbb{R}^n.$$

Then,

$$\int w dx \geq \Bigl(\int u dx\Bigr)^{\lambda} \Bigl(\int v dx\Bigr)^{1-\lambda}.$$

Proof. The proof will proceed by induction on n. Let us first show the induction step. Suppose the statement holds for n and we would like to show it for $n+1$. By assumption, for any $x, y \in \mathbb{R}^n$ and $a, b \in \mathbb{R}$

$$w(\lambda x + (1-\lambda)y, \lambda a + (1-\lambda)b) \geq u(x,a)^{\lambda} v(y,b)^{1-\lambda}.$$

Let us fix a and b and consider functions

$$w_1(x) = w(x, \lambda a + (1-\lambda)b), \;\; u_1(x) = u(x,a), \;\; v_1(x) = v(x,b)$$

on $\mathbb{R}^n$ that satisfy

$$w_1(\lambda x + (1-\lambda)y) \geq u_1(x)^{\lambda} v_1(y)^{1-\lambda}.$$

By induction assumption,

$$\int_{\mathbb{R}^n} w_1 dx \geq \Bigl(\int_{\mathbb{R}^n} u_1 dx\Bigr)^{\lambda} \Bigl(\int_{\mathbb{R}^n} v_1 dx\Bigr)^{1-\lambda}.$$

These integrals still depend on a and b and we can define

$$w_2(\lambda a + (1-\lambda)b) = \int_{\mathbb{R}^n} w_1 dx = \int_{\mathbb{R}^n} w(x, \lambda a + (1-\lambda)b) dx$$

and, similarly,

$$u_2(a) = \int_{\mathbb{R}^n} u_1(x,a) dx, \;\; v_2(b) = \int_{\mathbb{R}^n} v_1(x,b) dx$$

so that

$$w_2(\lambda a + (1-\lambda)b) \geq u_2(a)^{\lambda} v_2(b)^{1-\lambda}.$$

These functions are defined on $\mathbb{R}$ and, by induction assumption,

$$\int_{\mathbb{R}} w_2 ds \geq \Bigl(\int_{\mathbb{R}} u_2 ds\Bigr)^{\lambda}\Bigl(\int_{\mathbb{R}} v_2 ds\Bigr)^{1-\lambda} \Longrightarrow \int_{\mathbb{R}^{n+1}} w dz \geq \Bigl(\int_{\mathbb{R}^{n+1}} u dz\Bigr)^{\lambda}\Bigl(\int_{\mathbb{R}^{n+1}} v dz\Bigr)^{1-\lambda},$$

which finishes the proof of the induction step. It remains to prove the case $n = 1$. Let us show two different proofs.

1. One approach is based on the Brunn-Minkowski inequality on the real line which says that, if γ is the Lebesgue measure and A, B are Borel sets on $\mathbb{R}$, then

$$\gamma(\lambda A + (1-\lambda)B) \geq \lambda\gamma(A) + (1-\lambda)\gamma(B),$$

where $A+B$ is the set addition, i.e. $A+B = \{a+b : a \in A, b \in B\}$. We can also assume that $u, v, w : \mathbb{R} \to [0,1]$ because the inequality is homogeneous to scaling. We have

$$\{w \geq a\} \supseteq \lambda\{u \geq a\} + (1-\lambda)\{v \geq a\}$$

because if $u(x) \geq a$ and $v(y) \geq a$ then, by assumption,

$$w(\lambda x + (1-\lambda)y) \geq u(x)^{\lambda} v(y)^{1-\lambda} \geq a^{\lambda} a^{1-\lambda} = a.$$

The Brunn-Minkowski inequality implies that

$$\gamma(w \geq a) \geq \lambda\gamma(u \geq a) + (1-\lambda)\gamma(v \geq a).$$

Finally,

$$\begin{aligned}
\int_{\mathbb{R}} w(z)dz &= \int_{\mathbb{R}}\int_0^1 I(x \leq w(z))dxdz = \int_0^1 \gamma(w \geq x)dx \\
&\geq \lambda\int_0^1 \gamma(u \geq x)dx + (1-\lambda)\int_0^1 \gamma(v \geq x)dx \\
&= \lambda\int_{\mathbb{R}} u(z)dz + (1-\lambda)\int_{\mathbb{R}} v(z)dz \geq \Bigl(\int_{\mathbb{R}} u(z)dz\Bigr)^{\lambda}\Bigl(\int_{\mathbb{R}} v(z)dz\Bigr)^{1-\lambda}.
\end{aligned}$$

2. Another approach is based on the transportation of measure. We can assume that $\int u = \int v = 1$ by rescaling

$$u \to \frac{u}{\int u},\ \ v \to \frac{v}{\int v},\ \ w \to \frac{w}{(\int u)^{\lambda}(\int v)^{1-\lambda}}.$$

Then we need to show that $\int w \geq 1$. Without loss of generality, let us assume that $u, v \geq 0$ are smooth and strictly positive, since one can easily reduce to this case. Define $x(t), y(t)$ for $0 \leq t \leq 1$ by

$$\int_{-\infty}^{x(t)} u(s)ds = t,\ \ \int_{-\infty}^{y(t)} v(s)ds = t.$$

Then

$$u(x(t))x'(t) = 1,\ \ u(y(t))y'(t) = 1$$

and the derivatives $x'(t), y'(t) > 0$. Define $z(t) = \lambda x(t) + (1-\lambda)y(t)$. Then

$$\int_{-\infty}^{+\infty} w(s)ds = \int_0^1 w(z(s))dz(s) = \int_0^1 w(\lambda x(s) + (1-\lambda)y(s))z'(s)ds.$$

By arithmetic-geometric mean inequality

$$z'(s) = \lambda x'(s) + (1-\lambda)y'(s) \geq (x'(s))^{\lambda}(y'(s))^{1-\lambda}$$

and, by assumption,

$$w(\lambda x(s)+(1-\lambda)y(s)) \geq u(x(s))^{\lambda}v(y(s))^{1-\lambda}.$$

Therefore,

$$\int w(s)ds \geq \int_0^1 \Big(u(x(s))x'(s)\Big)^{\lambda}\Big(v(y(s))y'(s)\Big)^{1-\lambda}ds = \int_0^1 1ds = 1.$$

This finishes the proof of theorem.

□

Entropy and the Kullback-Leibler divergence. Consider a probability measure $\mathbb{P}$ on $\mathbb{R}^n$ and a nonnegative measurable function $u : \mathbb{R}^n \to [0,\infty)$.

Definition (Entropy) *We define the* entropy *of* u *with respect to* $\mathbb{P}$ *by*

$$\mathbf{Ent}_{\mathbb{P}}(u) = \int u\log u d\mathbb{P} - \int u d\mathbb{P}\cdot\log\int u d\mathbb{P}.$$

One can give a different representation of entropy by

$$\mathbf{Ent}_{\mathbb{P}}(u) = \sup\Big\{\int uv d\mathbb{P} : \int e^v d\mathbb{P} \leq 1\Big\}. \tag{21.0.1}$$

Indeed, if we consider a convex set $V = \{v : \int e^v d\mathbb{P} \leq 1\}$ then the above supremum is obviously a solution of the following saddle point problem:

$$L(v,\lambda) = \int uv d\mathbb{P} - \lambda\Big(\int e^v d\mathbb{P} - 1\Big) \to \sup_v \inf_{\lambda\geq 0}.$$

The functional L is linear in λ and concave in v. Therefore, by the minimax theorem, a saddle point solution exists and $\sup\inf = \inf\sup$. The integral

$$\int uv d\mathbb{P} - \lambda\int e^v d\mathbb{P} = \int (uv - \lambda e^v) d\mathbb{P}$$

can be maximized pointwise by taking v such that $u = \lambda e^v$. Then

$$L(v,\lambda) = \int u\log\frac{u}{\lambda}d\mathbb{P} - \int u d\mathbb{P} + \lambda$$

and maximizing over λ gives $\lambda = \int u$ and $v = \log(u/\int u)$. This proves (21.0.1). Suppose now that a law $\mathbb{Q}$ is absolutely continuous with respect to $\mathbb{P}$ and denote its Radon-Nikodym derivative by

$$u = \frac{d\mathbb{Q}}{d\mathbb{P}}. \tag{21.0.2}$$

Definition (Kullback-Leibler divergence) *The quantity*

$$D(\mathbb{Q}||\mathbb{P}) := \int \log u \, d\mathbb{Q} = \int \log\frac{d\mathbb{Q}}{d\mathbb{P}}\, d\mathbb{Q}$$

is called the Kullback-Leibler divergence *between* $\mathbb{P}$ *and* $\mathbb{Q}$.

Clearly, $D(\mathbb{Q}||\mathbb{P}) = \mathbf{Ent}_{\mathbb{P}}(u)$, since

$$\mathbf{Ent}_{\mathbb{P}}(u) = \int \log\frac{d\mathbb{Q}}{d\mathbb{P}}\cdot\frac{d\mathbb{Q}}{d\mathbb{P}}\,d\mathbb{P} - \int\frac{d\mathbb{Q}}{d\mathbb{P}}\,d\mathbb{P}\cdot\log\int\frac{d\mathbb{Q}}{d\mathbb{P}}\,d\mathbb{P} = \int\log\frac{d\mathbb{Q}}{d\mathbb{P}}\,d\mathbb{Q}.$$

The variational characterization (21.0.1) implies that

$$\text{if} \quad \int e^v d\mathbb{P} \leq 1 \quad \text{then} \quad \int v d\mathbb{Q} = \int uv d\mathbb{P} \leq D(Q||P). \tag{21.0.3}$$

Transportation inequality for log-concave measures. Suppose that a probability distribution $\mathbb{P}$ on $\mathbb{R}^n$ has the Lebesgue density $e^{-V(x)}$ where $V(x)$ is strictly convex in the following sense:

$$tV(x) + (1-t)V(y) - V(tx + (1-t)y) \geq C_p(1 - t + \mathbf{o}(1-t))|x-y|^p \tag{21.0.4}$$

as $t \to 1$ for some $p \geq 2$ and $C_p > 0$.

Example. One example of the distribution that satisfies (21.0.4) is the non-degenerate normal distribution $N(0, C)$ that corresponds to

$$V(x) = \frac{1}{2}(C^{-1}x, x) + \text{const}$$

for some covariance matrix C, $\det C \neq 0$. If we denote $A = C^{-1}/2$ then

$$\begin{aligned} &t(Ax, x) + (1-t)(Ay, y) - (A(tx + (1-t)y), (tx + (1-t)y)) \\ &= t(1-t)(A(x-y), (x-y)) \geq \frac{1}{2\lambda_{\max}(C)} t(1-t)|x-y|^2, \end{aligned} \tag{21.0.5}$$

where $\lambda_{\max}(C)$ is the largest eigenvalue of C. Thus, (21.0.4) holds with $p = 2$ and $C_p = 1/(2\lambda_{\max}(C))$. $\square$

Let us prove the following useful inequality for the Wasserstein distance.

Theorem 52 *If $\mathbb{P}$ satisfies (21.0.4) and $\mathbb{Q}$ is absolutely continuous w.r.t. $\mathbb{P}$ then*

$$W_p(\mathbb{Q}, \mathbb{P})^p \leq \frac{1}{C_p} D(\mathbb{Q} \| \mathbb{P}).$$

Proof. Take functions $f, g \in C(\mathbb{R}^n)$ such that

$$f(x) + g(y) \leq \frac{1}{t(1-t)} C_p(1 - t + \mathbf{o}(1-t))|x-y|^p.$$

Then, by (21.0.4),

$$f(x) + g(y) \leq \frac{1}{t(1-t)}\Big(tV(x) + (1-t)V(y) - V(tx + (1-t)y)\Big)$$

and

$$t(1-t)f(x) - tV(x) + t(1-t)g(y) - (1-t)V(y) \leq -V(tx + (1-t)y).$$

This implies that

$$w(tx + (1-t)y) \geq u(x)^t v(y)^{1-t}$$

for

$$u(x) = e^{(1-t)f(x) - V(x)},\ v(y) = e^{tg(y) - V(y)} \ \text{ and } \ w(z) = e^{-V(z)}.$$

By the Prekopa-Leindler inequality,

$$\Big(\int e^{(1-t)f(x) - V(x)} dx\Big)^t \Big(\int e^{tg(x) - V(x)} dx\Big)^{1-t} \leq \int e^{-V(x)} dx$$

and since e^{-V} is the density of $\mathbb{P}$ we get

$$\Big(\int e^{(1-t)f} d\mathbb{P}\Big)^t \Big(\int e^{tg} d\mathbb{P}\Big)^{1-t} \leq 1 \ \text{ and } \ \Big(\int e^{(1-t)f} d\mathbb{P}\Big)^{\frac{1}{1-t}} \Big(\int e^{tg} d\mathbb{P}\Big)^{\frac{1}{t}} \leq 1.$$

It is a simple calculus exercise to show that

$$\lim_{s \to 0} \Big(\int e^{sf} d\mathbb{P}\Big)^{\frac{1}{s}} = e^{\int f d\mathbb{P}},$$

and, therefore, letting $t \to 1$ proves that

$$\text{if} \quad f(x) + g(y) \le C_p |x-y|^p \quad \text{then} \quad \int e^g d\mathbb{P} \cdot e^{\int f d\mathbb{P}} \le 1.$$

If we denote $v = g + \int f d\mathbb{P}$ then the last inequality is $\int e^v d\mathbb{P} \le 1$ and (21.0.3) implies that

$$\int v d\mathbb{Q} = \int f d\mathbb{P} + \int g d\mathbb{Q} \le D(\mathbb{Q}||\mathbb{P}).$$

Finally, using the Kantorovich-Rubinstein theorem, (20.0.2), we get

$$\begin{aligned} W_p(\mathbb{Q},\mathbb{P})^p &= \frac{1}{C_p} \inf\Big\{ \int C_p |x-y|^p d\mu(x,y) : \mu \in M(\mathbb{P},\mathbb{Q}) \Big\} \\ &= \frac{1}{C_p} \sup \Big\{ \int f d\mathbb{P} + \int g d\mathbb{Q} : f(x) + g(y) \le C_p |x-y|^p \Big\} \le \frac{1}{C_p} D(\mathbb{Q}||\mathbb{P}) \end{aligned}$$

and this finishes the proof.

□

Concentration of Gaussian measure. Applying this result to the example before Theorem 52 gives that for the non-degenerate Gaussian distribution $\mathbb{P} = N(0, C)$,

$$W_2(\mathbb{P},\mathbb{Q}) \le \sqrt{2\lambda_{\max}(C) D(\mathbb{Q}||\mathbb{P})}. \tag{21.0.6}$$

Given a measurable set $A \subseteq \mathbb{R}^n$ with $\mathbb{P}(A) > 0$, define the conditional distribution $\mathbb{P}_A$ by

$$\mathbb{P}_A(C) = \frac{\mathbb{P}(CA)}{\mathbb{P}(A)}.$$

Then, obviously, the Radon-Nikodym derivative

$$\frac{d\mathbb{P}_A}{d\mathbb{P}} = \frac{1}{\mathbb{P}(A)} \mathrm{I}_A$$

and the Kullback-Leibler divergence

$$D(\mathbb{P}_A||\mathbb{P}) = \int_A \log \frac{1}{\mathbb{P}(A)} d\mathbb{P}_A = \log \frac{1}{\mathbb{P}(A)}.$$

Since W_2 is a metric, for any two Borel sets A and B

$$W_2(\mathbb{P}_A, \mathbb{P}_B) \le W_2(\mathbb{P}_A,\mathbb{P}) + W_2(\mathbb{P}_B,\mathbb{P}) \le \sqrt{2\lambda_{\max}(C)} \Big(\sqrt{\log \frac{1}{\mathbb{P}(A)}} + \sqrt{\log \frac{1}{\mathbb{P}(B)}} \Big).$$

Suppose that the sets A and B are apart from each other by a distance t, i.e. $d(A, B) \ge t > 0$. Then any two points in the support of measures $\mathbb{P}_A$ and $\mathbb{P}_B$ are at a distance at least t from each other and the transportation distance $W_2(\mathbb{P}_A, \mathbb{P}_B) \ge t$. Therefore,

$$t \le W_2(\mathbb{P}_A,\mathbb{P}_B) \le \sqrt{2\lambda_{\max}(C)} \Big(\sqrt{\log \frac{1}{\mathbb{P}(A)}} + \sqrt{\log \frac{1}{\mathbb{P}(B)}} \Big) \le \sqrt{4\lambda_{\max}(C) \log \frac{1}{\mathbb{P}(A)\mathbb{P}(B)}}.$$

Therefore,

$$\mathbb{P}(B) \le \frac{1}{\mathbb{P}(A)} \exp\Big(-\frac{t^2}{4\lambda_{\max}(C)} \Big).$$

In particular, if $B = \{x : d(x, A) \ge t\}$ then

$$\mathbb{P}\big(d(x, A) \ge t\big) \le \frac{1}{\mathbb{P}(A)} \exp\Big(-\frac{t^2}{4\lambda_{\max}(C)} \Big).$$

If the set A is not too small, e.g. $\mathbb{P}(A) \geq 1/2$, this implies that

$$\mathbb{P}\big(d(x,A) \geq t\big) \leq 2\exp\Big(-\frac{t^2}{4\lambda_{\max}(C)}\Big).$$

This shows that the Gaussian measure is exponentially concentrated near any "large enough" set. The constant 1/4 in the exponent is not optimal and can be replaced by 1/2; this is just an example of application of the above ideas. The optimal result is the famous Gaussian isoperimetry,

$$\text{if } \mathbb{P}(A) = \mathbb{P}(B) \text{ for some half-space } B \text{ then } \mathbb{P}(A^t) \geq \mathbb{P}(B^t).$$

Gaussian concentration via the Prekopa-Leindler inequality. If we denote $c = 1/\lambda_{\max}(C)$ then setting $t = 1/2$ in (21.0.5),

$$V(x) + V(y) - 2V\Big(\frac{x+y}{2}\Big) \geq \frac{c}{4}|x-y|^2.$$

Given a function f on $\mathbb{R}^n$ let us define its *infimum-convolution* by

$$g(y) = \inf_x \Big(f(x) + \frac{c}{4}|x-y|^2\Big).$$

Then, for all x and y,

$$g(y) - f(x) \leq \frac{c}{4}|x-y|^2 \leq V(x) + V(y) - 2V\Big(\frac{x+y}{2}\Big). \tag{21.0.7}$$

If we define

$$u(x) = e^{-f(x)-V(x)},\ \ v(y) = e^{g(y)-V(y)},\ \ w(z) = e^{-V(z)}$$

then (21.0.7) implies that

$$w\Big(\frac{x+y}{2}\Big) \geq u(x)^{1/2}v(y)^{1/2}.$$

The Prekopa-Leindler inequality with $\lambda = 1/2$ implies that

$$\int e^g d\mathbb{P} \int e^{-f} d\mathbb{P} \leq 1. \tag{21.0.8}$$

Given a measurable set A, let f be equal to 0 on A and $+\infty$ on the complement of A. Then

$$g(y) = \frac{c}{4}d(x,A)^2$$

and (21.0.8) implies

$$\int \exp\Big(\frac{c}{4}d(x,A)^2\Big)d\mathbb{P}(x) \leq \frac{1}{\mathbb{P}(A)}.$$

By Chebyshev's inequality,

$$\mathbb{P}\big(d(x,A) \geq t\big) \leq \frac{1}{\mathbb{P}(A)}\exp\Big(-\frac{ct^2}{4}\Big) = \frac{1}{\mathbb{P}(A)}\exp\Big(-\frac{t^2}{4\lambda_{\max}(C)}\Big).$$

□

Trivial metric and total variation.

Definition A total variation *distance between probability measure* $\mathbb{P}$ *and* $\mathbb{Q}$ *on a measurable space* $(S,\mathcal{B})$ *is defined by*

$$\mathrm{TV}(\mathbb{P},\mathbb{Q}) = \sup_{A\in\mathcal{B}} |\mathbb{P}(A) - \mathbb{Q}(A)|.$$

Using the Hahn-Jordan decomposition, we can represent a signed measure $\mu = \mathbb{P} - \mathbb{Q}$ as $\mu = \mu^+ - \mu^-$ such that for some set $D \in \mathcal{B}$ and for any set $E \in \mathcal{B}$,

$$\mu^+(E) = \mu(ED) \geq 0 \text{ and } \mu^-(E) = -\mu(ED^c) \geq 0.$$

Therefore, for any $A \in \mathcal{B}$,

$$\mathbb{P}(A) - \mathbb{Q}(A) = \mu^+(A) - \mu^-(A) = \mu^+(AD) - \mu^-(AD^c)$$

which makes it obvious that

$$\sup_{A \in \mathcal{B}} |\mathbb{P}(A) - \mathbb{Q}(A)| = \mu^+(D).$$

Let us describe some connections of the total variation distance to the Kullback-Leibler divergence and the Kantorovich-Rubinstein theorem. Let us start with the following simple observation.

Lemma 43 *If f is a measurable function on S such that $|f| \le 1$ and $\int f d\mathbb{P} = 0$ then for any $\lambda \in \mathbb{R}$,*

$$\int e^{\lambda f} d\mathbb{P} \le e^{\lambda^2/2}.$$

Proof. Since $(1+f)/2, (1-f)/2 \in [0,1]$ and

$$\lambda f = \frac{1+f}{2}\lambda + \frac{1-f}{2}(-\lambda),$$

by convexity of e^x we get

$$e^{\lambda f} \le \frac{1+f}{2}e^{\lambda} + \frac{1-f}{2}e^{-\lambda} = \mathrm{ch}(\lambda) + f\mathrm{sh}(\lambda).$$

Therefore,

$$\int e^{\lambda f} d\mathbb{P} \le \mathrm{ch}(\lambda) \le e^{\lambda^2/2},$$

where the last inequality is easy to see by Taylor's expansion.

□

Let us now consider a *trivial metric* on S given by

$$d(x,y) = I(x \ne y). \tag{21.0.9}$$

Then a 1-Lipschitz function f w.r.t. d, $\|f\|_{\mathrm{L}} \le 1$, is defined by the condition that for all $x, y \in S$,

$$|f(x) - f(y)| \le 1. \tag{21.0.10}$$

Formally, the Kantorovich-Rubinstein theorem in this case would state that

$$\begin{aligned} W(\mathbb{P},\mathbb{Q}) &:= \inf\Bigl\{\int I(x \ne y) d\mu(x,y) : \mu \in M(\mathbb{P},\mathbb{Q})\Bigr\} \\ &= \sup\Bigl\{\Bigl|\int f d\mathbb{Q} - \int f d\mathbb{P}\Bigr| : \|f\|_{\mathrm{L}} \le 1\Bigr\} =: \gamma(\mathbb{P},\mathbb{Q}). \end{aligned}$$

However, since any uncountable set S is not separable w.r.t. a trivial metric d, we can not apply the Kantorovich-Rubinstein theorem directly. In this case one can use the Hahn-Jordan decomposition to show that γ coincides with the total variation distance,

$$\gamma(\mathbb{P},\mathbb{Q}) = \mathrm{TV}(\mathbb{P},\mathbb{Q})$$

and it is easy to construct a measure $\mu \in M(\mathbb{P},\mathbb{Q})$ explicitly that witnesses the above equality. We leave this as an exercise. Thus, for the trivial metric d,

$$W(\mathbb{P},\mathbb{Q}) = \gamma(\mathbb{P},\mathbb{Q}) = \mathrm{TV}(\mathbb{P},\mathbb{Q}).$$

We have the following analogue of the KL divergence bound.

Theorem 53 *If $\mathbb{Q}$ is absolutely continuous w.r.t. $\mathbb{P}$ then*

$$\mathrm{TV}(\mathbb{P},\mathbb{Q}) \le \sqrt{2D(\mathbb{Q}||\mathbb{P})}.$$

Proof. Take f such that (21.0.10) holds. If we define $g(x) = f(x) - \int f d\mathbb{P}$ then, clearly, $|g| \le 1$ and $\int g d\mathbb{P} = 0$. The above lemma implies that for any $\lambda \in \mathbb{R}$,

$$\int e^{\lambda f - \lambda \int f d\mathbb{P} - \lambda^2/2} d\mathbb{P} \le 1.$$

The variational characterization of entropy (21.0.3) implies that

$$\lambda \int f d\mathbb{Q} - \lambda \int f d\mathbb{P} - \lambda^2/2 \le D(\mathbb{Q}||\mathbb{P})$$

and for $\lambda > 0$ we get

$$\int f d\mathbb{Q} - \int f d\mathbb{P} \le \frac{\lambda}{2} + \frac{1}{\lambda} D(\mathbb{Q}||\mathbb{P}).$$

Minimizing the right hand side over $\lambda > 0$, we get

$$\int f d\mathbb{Q} - \int f d\mathbb{P} \le \sqrt{2D(\mathbb{Q}||\mathbb{P})}.$$

Applying this to f and $-f$ yields the result.

□