# Transportation method, convex Lipschitz concentration

Recall Donsker-Varadhan: For any $Q \ll P$,

$$D_{KL}(Q\|P) = \sup_{U : \mathbb{E}_P e^U < \infty} \mathbb{E}_Q U - \log \mathbb{E}_P e^U.$$

The following is its dual form:

Lemma (Gibbs variational principle): For any r.v. $U$ on $(\Omega, P)$,

$$\log \mathbb{E}_P e^U = \sup_{Q \ll P} \mathbb{E}_Q U - D_{KL}(Q\|P).$$

Proof: Suppose $U \le M$ a.s. for a constant $M < \infty$. By Donsker-Varadhan, $D_{KL}(Q\|P) \ge \mathbb{E}_Q U - \log \mathbb{E}_P e^U$, so $\log \mathbb{E}_P e^U \ge \sup_{Q \ll P} \mathbb{E}_Q U - D_{KL}(Q\|P)$. Equality holds when $\frac{dQ}{dP} = \frac{e^U}{\mathbb{E}_P e^U}$. For unbounded $U$, apply this to $\min(U, M)$ and then take sup over $M$. $\blacksquare$

Thm: Let $Z$ be a r.v. on $(\Omega, P)$, $\phi : [0,\infty) \to [0,\infty)$ convex, differentiable w/ $\phi(0) = \phi'(0) = 0$. The following are equivalent:

(a) $\log \mathbb{E} e^{\lambda(Z - \mathbb{E}Z)} \le \phi(\lambda) \quad \forall \lambda \ge 0.$

(b) For any $Q \ll P$, $\mathbb{E}_Q Z - \mathbb{E}_P Z \le \phi^{*-1}(D_{KL}(Q\|P))$

where $\phi^*(t) = \sup_{\lambda \geq 0} \lambda t - \phi(\lambda)$.

In particular, $\log \mathbb{E} e^{\lambda(Z - \mathbb{E}Z)} \leq \frac{\lambda^2 \sigma^2}{2}$ if and only if

$$\forall Q \ll P, \quad \mathbb{E}_Q Z - \mathbb{E}_P Z \leq \sqrt{2\sigma^2 D_{KL}(Q \| P)}$$

Proof: By Gibbs variational principle,

$$\log \mathbb{E}_P e^{\lambda(Z - \mathbb{E}_P Z)} = \sup_{Q \ll P} \lambda(\mathbb{E}_Q Z - \mathbb{E}_P Z) - D_{KL}(Q \| P).$$

Then $(a) \Leftrightarrow \forall \lambda \geq 0, Q \ll P, \quad \lambda(\mathbb{E}_Q Z - \mathbb{E}_P Z) - D_{KL}(Q \| P) \leq \phi(\lambda)$

$$\Leftrightarrow \forall Q \ll P, \quad D_{KL}(Q \| P) \geq \sup_{\lambda \geq 0} \lambda(\mathbb{E}_Q Z - \mathbb{E}_P Z) - \phi(\lambda)$$

$$= \phi^*(\mathbb{E}_Q Z - \mathbb{E}_P Z)$$

$\Leftrightarrow (b)$. ($\phi^*(t)$ is increasing on $[0, \infty)$, $0$ on $(-\infty, 0]$) $\blacksquare$

Thus: Let $(X_1, \ldots, X_n) \in \mathcal{X}^n$ w/ joint law $P$, and $Z = f(X_1, \ldots, X_n)$. Then

$$\mathbb{E} e^{\lambda(Z - \mathbb{E}Z)} \leq e^{\frac{\lambda^2 \sigma^2}{2}} \quad \forall \lambda \geq 0$$

iff, for any $(Y_1, \ldots, Y_n) \in \mathcal{X}^n$ w/ joint law $Q$,

$$\mathbb{E} f(Y_1, \ldots, Y_n) - \mathbb{E} f(X_1, \ldots, X_n) \leq \sqrt{2\sigma^2 D_{KL}(Q \| P)} \quad (**)$$

Goal: Bound left side of $(**)$ by <u>coupling</u> $P$ and $Q$, and using bounded differences / Lipschitz properties of $f$.

## Bounded differences revisited

Thm (McDiarmid): Let $X_1, \ldots, X_n$ be independent, $Z = f(X_1, \ldots, X_n)$,

$$\|D_i f\|_\infty = \sup_{x_1, \ldots, x_n, x_i'} |f(x_1, \ldots, x_n) - f(x_1, \ldots, x_i', \ldots, x_n)|.$$

Then $Z - \mathbb{E}Z$ is $\frac{1}{4}\sum_{i=1}^n \|D_i f\|_\infty^2$-subgaussian.

[We proved this in Lecture 1 using the martingale method.]

Under any coupling of $(X_1, \ldots, X_n) \sim P$ w/ $(Y_1, \ldots, Y_n) \sim Q$,

$$\mathbb{E}\left[f(Y_1, \ldots, Y_n) - f(X_1, \ldots, X_n)\right] \le \mathbb{E}\left[\sum_{i=1}^n \|D_i f\|_\infty \cdot \mathbb{1}\{X_i \ne Y_i\}\right] \le \sum_{i=1}^n \|D_i f\|_\infty \cdot \mathbb{P}[X_i \ne Y_i]$$

Lemma (Pinsker's inequality): Let $Q \ll P$, and

$$TV(P, Q) = \sup_{A \subseteq \Omega} |P(A) - Q(A)| = \min_{(X,Y) \sim \text{couplings}(P,Q)} \mathbb{P}[X \ne Y].$$

Then $TV(P,Q)^2 \le \frac{1}{2} D_{KL}(Q \| P)$.

Proof: Let $Z = \mathbb{1}_A(\omega)$ or $-\mathbb{1}_A(\omega)$ for any $A \subseteq \Omega$. Since $Z$ is bounded, by Hoeffding's Lemma,

$$\log \mathbb{E}_P e^{\lambda(Z - \mathbb{E}_P Z)} \le \frac{\lambda^2}{8} \quad \forall \lambda \ge 0.$$

Then by above theorem, $\mathbb{E}_Q Z - \mathbb{E}_P(Z) \le \sqrt{\frac{1}{2} D_{KL}(Q \| P)}$.

So $|Q(A) - P(A)| \le \sqrt{\frac{1}{2} D_{KL}(Q \| P)}$ for any $A \subseteq \Omega$. ∎

Thm (Marton): Let $P = \bigotimes_{i=1}^n P_i$ be a product measure on $\mathbb{R}^n$ s.t., for some $w: \mathbb{R}^2 \to [0,\infty)$, convex $\phi: [0,\infty) \to [0,\infty)$, and all $i \in \{1,\dots,n\}$ and $Q_i \ll P_i$,

$$\min_{(X_i, Y_i) \sim \text{couplings}(P_i, Q_i)} \phi(\mathbb{E}\, w(X_i, Y_i)) \leq D_{KL}(Q_i \| P_i).$$

Then for all $Q \ll P$,

$$\min_{(X,Y) \sim \text{couplings}(P,Q)} \sum_{i=1}^n \phi(\mathbb{E}\, w(X_i, Y_i)) \leq D_{KL}(Q \| P).$$

Proof: Induct on $n$. Base case $n=1$ is the given condition.

Suppose result holds for $n-1$, and consider $Q \ll P = \bigotimes_{i=1}^n P_i$.

For $Y = (Y_1, \dots, Y_n) \sim Q$, let

$$Q^{(n-1)} = \text{marginal law of } (Y_1, \dots, Y_{n-1})$$

$$Q_n(\cdot \mid Y_1, \dots, Y_{n-1}) = \text{conditional law of } Y_n.$$

By chain rule for KL-divergence:

$$D_{KL}(Q \| P) = D_{KL}\left(Q^{(n-1)} \,\Big\|\, \bigotimes_{i=1}^{n-1} P_i\right) + \mathbb{E}\, D_{KL}(Q_n(\cdot \mid Y_1, \dots, Y_{n-1}) \| P_n)$$

Consider the coupling $(X, Y)$ of $(P, Q)$ s.t.

- $(X_1, \ldots, X_{n-1}), (Y_1, \ldots, Y_{n-1})$ is coupling of $\bigotimes_{i=1}^{n-1} P_i$ and $Q^{(n-1)}$ that minimizes $\sum_{i=1}^{n-1} \phi(\mathbb{E}\, w(X_i, Y_i))$.
- $X_n, Y_n \mid Y_1, \ldots, Y_{n-1}$ is coupling of $P_n$ and $Q_n(\cdot \mid Y_1, \ldots, Y_{n-1})$ that minimizes $\phi(\mathbb{E}[w(X_n, Y_n) \mid Y_1, \ldots, Y_{n-1}])$

$$\Rightarrow D_{KL}(Q \| P) \geq \underbrace{\sum_{i=1}^{n-1} \phi(\mathbb{E}\, w(X_i, Y_i))}_{\text{induction hypothesis}} + \mathbb{E}\, \underbrace{\phi(\mathbb{E}[w(X_n, Y_n) \mid Y_1, \ldots, Y_{n-1}])}_{\text{base case } n=1}$$

(Jensen's)
$$\geq \sum_{i=1}^{n-1} \phi(\mathbb{E}\, w(X_i, Y_i)) + \phi(\mathbb{E}\, w(X_n, Y_n))$$

completing the induction. ∎

Proof of Thm: Recall, for any coupling $(X, Y)$ of $(P, Q)$,

$$\mathbb{E} f(Y_1, \ldots, Y_n) - \mathbb{E} f(X_1, \ldots, X_n) \leq \sum_{i=1}^{n} \|D_i f\|_\infty \cdot \mathbb{P}[X_i \neq Y_i]$$

$$\leq \left(\sum_{i=1}^{n} \|D_i f\|_\infty^2\right)^{1/2} \left(\sum_{i=1}^{n} \mathbb{P}[X_i \neq Y_i]^2\right)^{1/2}$$

By Pinsker's inequality: $\forall Q_i \ll P_i$, $\exists$ coupling $(X_i, Y_i)$ of $(P_i, Q_i)$ s.t.

$$\mathbb{E}[\mathbb{1}\{X_i \neq Y_i\}]^2 = \mathbb{P}[X_i \neq Y_i]^2 \leq \frac{1}{2} D_{KL}(\mathbb{Q}_i \| \mathbb{P}_i).$$

Then by Marton's tensorization theorem ($w(x,y) = \mathbb{1}\{x \neq y\}$, $\phi(x) = x^2$),

$\exists$ coupling $(X,Y)$ of $(\mathbb{P},\mathbb{Q})$ s.t. $\sum_{i=1}^n \mathbb{P}[X_i \neq Y_i]^2 \leq \frac{1}{2} D_{KL}(\mathbb{Q}\|\mathbb{P})$.

$$\Rightarrow \mathbb{E}f(Y_1,\ldots,Y_n) - \mathbb{E}f(X_1,\ldots,X_n) \leq \sqrt{\frac{1}{2}\sum_{i=1}^n \|D_i f\|_\infty^2 \cdot D_{KL}(\mathbb{Q}\|\mathbb{P})}$$

Let $Z = f(X_1,\ldots,X_n)$. By transportation method, $\mathbb{E}e^{\lambda(Z-\mathbb{E}Z)} \leq e^{\frac{\lambda^2\sigma^2}{2}}$ $\forall \lambda \geq 0$

for $\sigma^2 = \frac{1}{4}\sum_{i=1}^n \|D_i f\|_\infty^2$. For $\lambda \leq 0$, apply to $-f$. $\blacksquare$

## Gaussian concentration revisited

**Thm** (Tsirelson-Ibragimov-Sudakov): Let $X_1,\ldots,X_n \overset{iid}{\sim} N(0,1)$,

$f: \mathbb{R}^n \to \mathbb{R}$ $L$-Lipschitz, $Z = f(X_1,\ldots,X_n)$. Then $Z - \mathbb{E}Z$ is

$L^2$-subgaussian.

**Lemma** (Stein): Let $Z \sim N(0,1)$, $\phi(z) = \frac{1}{\sqrt{2\pi}} e^{-\frac{z^2}{2}}$, and $f: \mathbb{R} \to \mathbb{R}$

be absolutely continuous s.t. $\mathbb{E}|Zf(Z)| < \infty$ and $\lim_{|z|\to\infty} f(z)\phi(z) = 0$.

Then $\mathbb{E}\, Zf(Z) = \mathbb{E}f'(Z)$.

Proof: Applying $\phi'(z) = -z\phi(z)$ and integration-by-parts,

$$\int_a^b f'(z)\phi(z)\,dz = f(b)\phi(b) - f(a)\phi(a) - \int_a^b f(z)\phi'(z)\,dz$$
$$= f(b)\phi(b) - f(a)\phi(a) + \int_a^b z f(z)\phi(z)\,dz.$$

Take $a \to -\infty$, $b \to \infty$ on both sides. $\blacksquare$

Lemma ($T_2$-inequality): Let $P = \mathcal{N}(0,1)$ and $Q \ll P$. Then

$$\min_{(X,Y)\sim\text{couplings}(P,Q)} \mathbb{E}(X-Y)^2 \le 2 D_{KL}(Q \| P).$$

[The left side is the squared Wasserstein-2 distance $W_2(P,Q)^2$.]

Proof: Let $P, Q$ be the CDFs and $\phi, q$ the densities.

Suppose first $\varepsilon \le \frac{q(x)}{\phi(x)} \le \frac{1}{\varepsilon}$ $\forall x \in \mathbb{R}$, so $Q$ strictly increasing.

The optimal coupling is $X \sim \mathcal{N}(0,1)$, $Y = T(X) := Q^{-1}(P(X))$.

By chain rule, $T'(x) = \dfrac{\phi(x)}{q(Q^{-1}(P(x)))} = \dfrac{\phi(x)}{q(T(x))}$

$$\Rightarrow D_{KL}(Q\|P) = \mathbb{E} \log \frac{q(Y)}{\phi(Y)}$$
$$= \mathbb{E} \log \frac{\phi(X)}{\phi(Y)\,T'(X)} = \mathbb{E}\left[-\frac{X^2}{2} + \frac{Y^2}{2} - \log T'(X)\right].$$
$$\ge \mathbb{E}\left[-\frac{X^2}{2} + \frac{Y^2}{2} + 1 - T'(X)\right].$$

$$Q(t) = \int_{-\infty}^{t} q(y)\,dy \le \frac{1}{\varepsilon}\int_{-\infty}^{t} \phi(y)\,dy$$

$$= \frac{1}{\varepsilon} P(t) \le P\left(\frac{t}{2}\right) \text{ for all large } t.$$

$\Rightarrow T(x) = Q^{-1}(P(x)) \le 2x$ for all large $x$

Similarly $T(-x) \ge -2x$ for all large $x$. So

- $\lim\limits_{|x|\to\infty} T(x)\phi(x) = 0$
- $\mathbb{E}|X \cdot T(X)| = \mathbb{E}|XY| < \infty$.

By Stein's Lemma: $D_{KL}(Q\|P) \ge \mathbb{E}\left[-\frac{X^2}{2} + \frac{Y^2}{2} + 1 - XY\right] = \frac{\mathbb{E}(X-Y)^2}{2}$.

For general $q(x)$: Let $w^\varepsilon(x) = \max\left(\varepsilon, \min\left(\frac{1}{\varepsilon}, \frac{q(x)}{\phi(x)}\right)\right)$,

$$q^\varepsilon(x) = \frac{w^\varepsilon(x)\cdot\phi(x)}{Z_\varepsilon}, \quad Z_\varepsilon = \int w^\varepsilon(x)\,\phi(x)\,dx,$$

$$Y^\varepsilon \sim Q^\varepsilon \text{ w/ density } q^\varepsilon$$

Note $w^\varepsilon(x) \le \varepsilon + \frac{q(x)}{\phi(x)}$. Thus, as $\varepsilon \to 0$, by dominated convergence:

- $Z^\varepsilon \to 1$, $q^\varepsilon(x) \to q(x)$, $Y^\varepsilon \xrightarrow{L} Y$
- $D_{KL}(Q^\varepsilon\|P) = \mathbb{E}\left[\frac{q^\varepsilon(X)}{\phi(X)} \log \frac{q^\varepsilon(X)}{\phi(X)}\right] \to D_{KL}(Q\|P)$.

So $\mathbb{E}(X-Y)^2 \le \liminf\limits_{\varepsilon\to 0} \mathbb{E}(X-Y^\varepsilon)^2 \le \lim\limits_{\varepsilon\to 0} 2D_{KL}(Q^\varepsilon\|P) = 2D_{KL}(Q\|P)$.

Cor: Let $P \sim \mathcal{N}(0, I)$ in $\mathbb{R}^n$ and $Q \ll P$. Then

$$\min_{(X,Y)\sim \text{couplings}(P,Q)} \mathbb{E}\|X-Y\|_2^2 \le 2 D_{KL}(Q\|P).$$

Proof: Apply Marton's tensorization w/ $w(x,y) = (x-y)^2$; $\phi(x) = x$.

Proof of Thm: For any $Q \ll P$, $\exists$ coupling $(X,Y)$ of $(P,Q)$ s.t.

$$\mathbb{E}f(Y) - \mathbb{E}f(X) \le L \cdot \mathbb{E}\|X-Y\|_2 \le \sqrt{2L^2 D_{KL}(Q\|P)}.$$

Apply this with both $f$ and $-f$. $\blacksquare$

<u>Convex Lipschitz concentration</u>

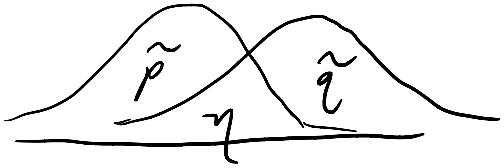
Thm (Talagrand): Let $X_1, \ldots, X_n$ be independent, $X_i \in [0,1]$. If $f: \mathbb{R}^n \to \mathbb{R}$ is $L$-Lipschitz and convex, $Z = f(X_1, \ldots, X_n)$, then $Z - \mathbb{E}Z$ is $L^2$-subgaussian.

Lemma: For any $Q \ll P$,

$$\min_{(X,Y)\sim \text{couplings}(P,Q)} \mathbb{E}\left[\mathbb{P}[X \ne Y \mid X]^2\right] + \mathbb{E}\left[\mathbb{P}[X \ne Y \mid Y]^2\right] \le 2 D_{KL}(Q\|P).$$

Proof: Let $p, q$ be the densities w.r.t. an underlying measure $\mu$.

Let $\eta(x) = \min(p(x), q(x))$, $\tilde{p}(x) = p(x) - \eta(x)$, $\tilde{q}(x) = q(x) - \eta(x)$.




Consider the coupling $(X,Y)$:

(1) w/ probability $\int \eta \, d\mu$, let $X = Y \sim \frac{\eta}{\int \eta \, d\mu}$

(2) w/ probability $1 - \int \eta \, d\mu = \int \tilde{p} \, d\mu = \int \tilde{q} \, d\mu$,

let $X \sim \frac{\tilde{p}}{\int \tilde{p} \, d\mu}$, $Y \sim \frac{\tilde{q}}{\int \tilde{q} \, d\mu}$ independently.

Then
$$\mathbb{P}[X \neq Y \mid X = x] = \frac{\tilde{p}(x)}{p(x)} = 1 - \frac{\eta(x)}{p(x)} = \left(1 - \frac{q(x)}{p(x)}\right)_+,$$
$$\mathbb{P}[X \neq Y \mid Y = y] = \frac{\tilde{q}(y)}{q(y)} = 1 - \frac{\eta(y)}{q(y)} = \left(1 - \frac{p(y)}{q(y)}\right)_+,$$
optimal for each fixed $x, y$.

$$\Rightarrow \min_{(X,Y) \sim \text{couplings}(P,Q)} \mathbb{E}\left[\mathbb{P}[X \neq Y \mid X]^2 + \mathbb{P}[X \neq Y \mid Y]^2\right]$$
$$= \mathbb{E}_P\left[\left(1 - \frac{q}{p}\right)_+^2\right] + \mathbb{E}_Q\left[\left(1 - \frac{p}{q}\right)_+^2\right]$$
$$= \mathbb{E}_P\left[\left(1 - \frac{q}{p}\right)_+^2 + \frac{q}{p}\left(1 - \frac{p}{q}\right)_+^2\right]$$
$$\leq \mathbb{E}_P\left[2\left(\frac{q}{p}\log\frac{q}{p} - \frac{q}{p} + 1\right)\right] \quad \text{by } (1-x)_+^2 + x\left(1 - \frac{1}{x}\right)_+^2 \leq 2(x\log x - x + 1)$$
$$= 2 D_{KL}(Q \| P).$$

$\blacksquare$

Thm (Marton): Let $P = \bigotimes_{i=1}^n P_i$ be a product measure on $\mathbb{R}^n$ s.t., for some $w: \mathbb{R}^2 \to [0,\infty)$, convex $\phi: [0,\infty) \to [0,\infty)$, and all $i \in \{1,\dots,n\}$ and $Q_i \ll P_i$,

$$\min_{(X_i, Y_i) \sim \text{couplings}(P_i, Q_i)} \mathbb{E}\left[\phi\left(\mathbb{E}[w(X_i, Y_i) \mid X_i]\right) + \phi\left(\mathbb{E}[w(X_i, Y_i) \mid Y_i]\right)\right] \le D_{KL}(Q_i \| P_i)$$

Then for all $Q \ll P$,

$$\min_{(X,Y) \sim \text{couplings}(P,Q)} \sum_{i=1}^n \mathbb{E}\left[\phi\left(\mathbb{E}[w(X_i, Y_i) \mid X]\right) + \phi\left(\mathbb{E}[w(X_i, Y_i) \mid Y]\right)\right] \le D_{KL}(Q \| P)$$

Proof: Induction on $n$, analogous to preceding tensorization theorem.

Proof of Thm: By convexity of $f$,

$$f(y) - f(x) \le \nabla f(y)^T (y - x)$$

$$= \sum_{i=1}^n \partial_i f(y) \cdot \underbrace{(y_i - x_i)}_{\in [0,1]} \le \sum_{i=1}^n |\partial_i f(y)| \cdot \mathbb{1}\{x_i \ne y_i\}.$$

For any $Q \ll P$, $\exists$ coupling of $(X_1, \dots, X_n) \sim P$ and $(Y_1, \dots, Y_n) \sim Q$ s.t.

$$\mathbb{E} f(Y_1, \dots, Y_n) - \mathbb{E} f(X_1, \dots, X_n) \le \sum_{i=1}^n \mathbb{E}\left[|\partial_i f(Y)| \cdot \mathbb{1}\{X_i \ne Y_i\}\right]$$

$$= \sum_{i=1}^n \mathbb{E}\left[|\partial_i f(Y)| \cdot \mathbb{E}\left[\mathbb{1}\{X_i \ne Y_i\} \mid Y\right]\right]$$

$$\leq \left(\mathbb{E}\sum_{i=1}^n |\partial_i f(Y)|^2\right)^{1/2}\left(\mathbb{E}\sum_{i=1}^n \mathbb{P}[X_i \neq Y_i | Y]^2\right)^{1/2}$$

$$\leq L\cdot\left(\mathbb{E}\sum_{i=1}^n \mathbb{P}[X_i \neq Y_i | Y]^2\right)^{1/2}$$

$$\leq L\cdot\sqrt{2D_{KL}(Q\|P)} \quad \text{by Marton's tensorization.}$$

Similarly, $-f(y)+f(x) \leq \nabla f(x)^T(x-y)$, so

$$\mathbb{E}[-f(Y_1,\dots,Y_n)]-\mathbb{E}[-f(X_1,\dots,X_n)] \leq \sum_{i=1}^n \mathbb{E}\left[|\partial_i f(X)|\cdot \mathbb{1}\{X_i \neq Y_i\}\right]$$

$$\leq L\cdot\left(\mathbb{E}\sum_{i=1}^n \mathbb{P}[X_i\neq Y_i|X]^2\right)^{1/2} \leq L\cdot\sqrt{2D_{KL}(Q\|P)}$$

Apply transportation method to both $f$ and $-f$. ∎

Example: Let $X\in\mathbb{R}^{m\times n}$ have independent entries, $X_{ij}\in[a,b]$.

Recall from Lecture 4 that $\sigma_{max}(X)$ is a convex, $(b-a)$-Lipschitz function of $Z_{ij} = \frac{X_{ij}-a}{b-a}\in[0,1]$

$$\Rightarrow \mathbb{P}\left[|\sigma_{max}(X)-\mathbb{E}\sigma_{max}(X)|\geq t\right] \leq 2e^{-\frac{t^2}{2(b-a)^2}}$$