[Lectures 13-14]

1 Resilience Beyond TV Distance

We now turn our attention to distances other than the distance $D = \mathsf{TV}$ that we have considered so far. The family of distances we will consider are called *Wasserstein distances*. Given a cost function $c(x, y)$ (which is usually assumed to be a metric), we define the distance $W_c(p, q)$ between two distributions p and q as

$$W_c(p,q) = \inf_{\pi} \mathbb{E}_{x,y\sim\pi}[c(x,y)] \tag{1}$$

$$\text{subject to } \int \pi(x,y)dy = p(x), \int \pi(x,y)dx = q(y). \tag{2}$$

This definition is a bit abstruse so let us unpack it. The decision variable π is called a *coupling* between p and q, and can be thought of as a way of matching points in p with points in q ($\pi(x, y)$ is the amount of mass in $p(x)$ that is matched to $q(y)$). The Wasserstein distance is then the minimum cost coupling (i.e., minimum cost matching) between p and q. Some special cases include:

- $c(x, y) = \mathbb{I}[x \neq y]$. Then W_c is the total variation distance, with the optimal coupling being $\pi(x, x) = \min(p(x), q(x))$ (the off-diagonal $\pi(x, y)$ can be arbitrary as long as the total mass adds up correctly).
- $c(x, y) = \|x - y\|_2$. Then W_c is the *earth-mover distance*—the average amount that we need to move points around to "move" p to q.
- $c(x, y) = \|x - y\|_0$. Then W_c is the average number of coordinates we need to change to move p to q.
- $c(x, y) = \|x - y\|_2^\alpha$, for $\alpha \in [0, 1]$. This is still a metric and interpolates between TV and earthmover distance.

There are a couple key properties of Wasserstein distance we will want to use. The first is that W_c is a metric if c is:

Proposition 1.1. *Suppose that c is a metric. Then W_c is also a metric.*

Proof. TBD □

The second, called *Kantorovich-Rubinstein duality*, provides an alternate definition of W_c distance in terms of functions that are Lipschitz under c, meaning that $|f(x) - f(y)| \leq c(x, y)$.

Theorem 1.2 (Kantorovich-Rubinstein)**.** *Call a function f* Lipschitz in c *if $|f(x) - f(y)| \leq c(x, y)$ for all x, y, and let $\mathcal{L}(c)$ denote the space of such functions. If c is a metric, then we have*

$$W_c(p,q) = \sup_{f\in\mathcal{L}(c)} \mathbb{E}_{x\sim p}[f(x)] - \mathbb{E}_{x\sim q}[f(x)]. \tag{3}$$

As a special case, take $c(x, y) = \mathbb{I}[x \neq y]$ (corresponding to TV distance). Then $f \in \mathcal{L}(c)$ if and only if $|f(x) - f(y)| \leq 1$ for all $x \neq y$. By translating f, we can equivalently take the supremum over all f mapping to $[0, 1]$. This says that

$$\mathsf{TV}(p,q) = \sup_{f:\mathcal{X}\to[0,1]} \mathbb{E}_p[f(x)] - \mathbb{E}_q[f(x)], \tag{4}$$

which recovers the definition of TV in terms of the maximum difference in probability of any event E.

As another special case, take $c(x, y) = \|x - y\|_2$. Then the supremum is over all 1-Lipschitz functions (in the usual sense).

In the next section, we will see how to generalize the definition of resilience to any Wasserstein distance.

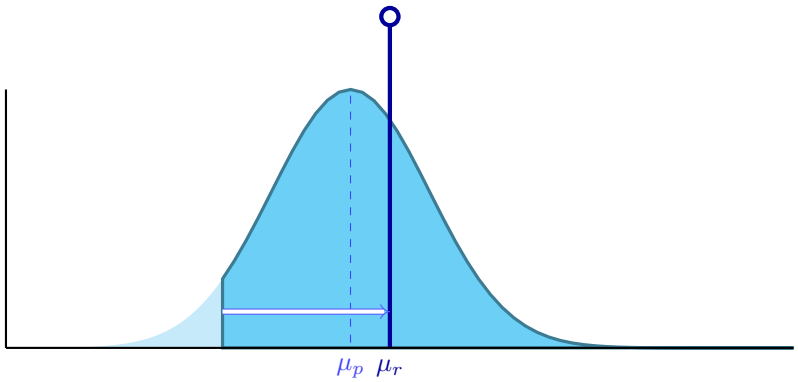
1.1 Resilience for Wasserstein distances

We show how to extend the idea of resilience to Wasserstein distances W_c. Recall that for TV distance, we showed that resilient sets have bounded modulus $\mathfrak{m}$; this crucially relied on the midpoint property that any p_1, p_2 have a midpoint r obtained via *deletions* of p_1 or p_2. In other words, we used the fact that any TV perturbation can be decomposed into a "friendly" operation (deletion) and its opposite (addition). We think of deletion as friendlier than addition, as the latter can move the mean arbitrarily far by adding probability mass at infinity.

To extend this to other Wasserstein distances, we need to identify a similar way of decomposing a Wasserstein perturbation into a friendly perturbation and its inverse. Unfortunately, deletion is closely tied to the TV distance in particular. To get around this, we use the following re-interpretation: *Deletion is equivalent to movement towards the mean under* TV. More precisely:

> $\hat{\mu}$ is a possible mean of an ϵ-deletion of p if and only if some r with mean $\hat{\mu}$ can be obtained from p by moving points *towards* $\hat{\mu}$ with TV distance at most ϵ.

This is more easily seen in the following diagram:



Here we can equivalently either delete the left tail of p or shift all of its mass to μ_r; both yield a modified distribution with the same mean μ_r. Thus we can more generally say that a perturbation is friendly if it only moves probability mass towards the mean. This motivates the following definition:

Definition 1.3 (Friendly perturbation)**.** For a distribution p over $\mathcal{X}$, fix a function $f : \mathcal{X} \to \mathbb{R}$. A distribution r is an ϵ-friendly perturbation of p for f under W_c if there is a coupling π between $X \sim p$ and $Y \sim r$ such that:

- The cost ($\mathbb{E}_\pi[c(X, Y)]$) is at most ϵ.
- All points move towards the mean of r: $f(Y)$ is between $f(X)$ and $\mathbb{E}_r[f(Y)]$ almost surely.

Note that friendliness is defined only in terms of one-dimensional functions $f : \mathcal{X} \to \mathbb{R}$; we will see how to handle higher-dimensional objects later. Intuitively, a friendly perturbation is a distribution r for which there exists a coupling that 'squeezes' p to μ_r.

The key property of deletion in the TV case was the existence of a *midpoint*: for any two distributions that are within ϵ in TV, one can find another distribution that is an ϵ-deletion of both distributions. We would like to show the analogous result for W_c–i.e. that if $W_c(p, q) \leq \epsilon$ then there exists an r that is an ϵ-friendly perturbation of *both* p and q for the function f.

The intuitive reason this is true is that any coupling between two one-dimensional distributions can be separated into two stages: in one stage all the mass only moves towards some point, in the other stage all the mass moves away from that point. This is illustrated in Figure 1.

To formalize this intuitive argument, we need a mild topological property:

Assumption 1.4 (Intermediate value property)**.** *For any x and y and any u with $f(x) < u < f(y)$, there is some z satisfying $f(z) = u$ and* $\max(c(x, z), c(z, y)) \leq c(x, y)$.

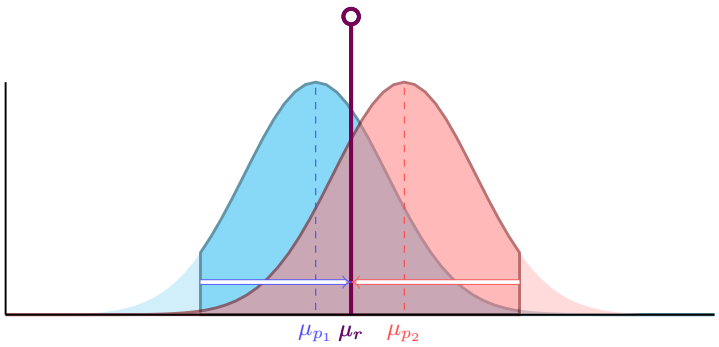


Figure 1: Illustration of midpoint lemma. For any distributions p_1, p_2 that are close under W_c, the coupling between p_1 and p_2 can be split into couplings $\pi_{p_1,r}$, $\pi_{p_2,r}$ such that p_1, p_2 only move towards μ_r under the couplings. We do this by "stopping" the movement from p_1 to p_2 at μ_r.

This holds for any f if $c = \mathbb{I}[x \neq y]$ (TV distance), and for any continuous f if c is a path metric (a metric with "nice" paths between points, which includes the ℓ_2-distance). Under this assumption we can prove the desired midpoint lemma:

Lemma 1.5 (Midpoint lemma for W_c)**.** *Suppose Assumption 1.4 holds. Then for any p_1 and p_2 such that $W_c(p_1, p_2) \leq \epsilon$ and any f, there exists a distribution r that is an ϵ-friendly perturbation of both p_1 and p_2 with respect to f.*

Proof. Given any two points x and y, without loss of generality we assume $f(x) \leq f(y)$. Define

$$s_{xy}(u) = \begin{cases} \min(f(x), f(y)), & u \leq \min(f(x), f(y)) \\ u, & u \in [f(x), f(y)] \\ \max(f(x), f(y)), & u \geq \max(f(x), f(y)). \end{cases} \tag{5}$$

If we imagine u increasing from $-\infty$ to $+\infty$, we can think of s_{xy} as a "slider" that tries to be as close to u as possible while remaining between $f(x)$ and $f(y)$.

By Assumption 1.4, there must exist some point z such that $\max(c(x,z), c(z,y)) \leq c(x,y)$ and $f(z) = s_{xy}(u)$. Call this point $z_{xy}(u)$.

Given a coupling $\pi(x,y)$ from p_1 to p_2, if we map y to $z_{xy}(u)$, we obtain a coupling $\pi_1(x,z)$ to some distribution $r(u)$, which by construction satisfies the squeezing property, except that it squeezes towards u rather than towards the mean $\mu(u) = \mathbb{E}_{X \sim r(u)}[f(X)]$. However, note that $u - \mu(u)$ is a continuous, monotonically non-decreasing function (since $u - s_{xy}(u)$ is non-decreasing) that ranges from $-\infty$ to $+\infty$. It follows that there is a u^* with $\mu(u^*) = u^*$. Then the couplings to $r(u^*)$ squeeze towards its mean $\mu(u^*)$.

Moreover, $\mathbb{E}_{(X,Z)\sim\pi_1}[c(X,Z)] \leq \mathbb{E}_{(X,Y)\sim\pi}[c(X,Y)] = W_c(p_1, p_2)$. The coupling π_1 therefore also has small enough cost, and so is a friendly perturbation. Similarly, the coupling π_2 mapping y to $z_{xy}(u^*)$ satisfies the squeezing property and has small enough cost by the same argument. □

Defining resilience: warm-up. With Lemma 1.5 in hand, we generalize resilience to Wasserstein distances by saying that a distribution is resilient if $\mathbb{E}_r[f(X)]$ is close to $\mathbb{E}_p[f(X)]$ for every η-friendly perturbation r and every function f lying within some appropriate family $\mathcal{F}$. For now, we will focus on second moment estimation under $W_{\|\cdot\|_2}$ (we consider second moment estimation because mean estimation is trivial under $W_{\|\cdot\|_2}$). This corresponds to the loss function

$$L(p, S) = \|\mathbb{E}_{x\sim p}[xx^\top] - S\|. \tag{6}$$

For notational convenience we also typically denote $W_{\|\cdot\|_2}$ as W_1.

For the loss $L(p,S)$, we will take our family $\mathcal{F}$ to be all functions of the form $f_v(x) = \langle x, v\rangle^2$ with $\|v\|_2 = 1$. Thus we define the (ρ, ϵ)-resilient distributions under W_1 as

$$\mathcal{G}_{\text{sec}}^{W_1}(\rho, \epsilon) = \{p \mid |\mathbb{E}_r[\langle x, v\rangle^2] - \mathbb{E}_p[\langle x, v\rangle^2]| \leq \rho \text{ whenever } r \text{ is } \epsilon\text{-friendly under } \langle x, v\rangle^2 \text{ and } \|v\|_2 = 1\}. \tag{7}$$

Note the twist in the definition of $\mathcal{G}_{\text{sec}}^{W_1}$–the allowed r depends on the current choice of v, since friendliness is specific to the function $f_v = \langle x, v\rangle^2$, which is different from deletions in the TV case.

We will first show that $\mathcal{G}_{\text{sec}}^{W_1}$ has small modulus, then derive sufficient moment conditions for p to be (ρ, ϵ)-resilient.

Proposition 1.6. *The set of (ρ, ϵ)-resilient distributions for W_1 has modulus $\mathfrak{m}(\mathcal{G}_{\text{sec}}^{W_1}(\rho, 2\epsilon), \epsilon) \le 2\rho$.*

Proof. For a distribution q, let $S_q = \mathbb{E}_q[xx^\top]$. Suppose that $p_1, p_2 \in \mathcal{G}_{\text{sec}}^{W_1}(\rho, \epsilon)$ and $W_1(p_1, p_2) \le 2\epsilon$. For any v, by Lemma 1.5, there exists an r that is a (2ϵ)-friendly perturbation of both p_1 and p_2 with respect to $\langle x, v\rangle^2$. We conclude that $|\mathbb{E}_{p_i}[\langle x, v\rangle^2] - \mathbb{E}_r[\langle x, v\rangle^2]| \le \rho$ for $i = 1, 2$, and hence $|\mathbb{E}_{p_1}[\langle x, v\rangle^2] - \mathbb{E}_{p_2}[\langle x, v\rangle^2]| \le 2\rho$, which can be written as $|v^\top(S_{p_1} - S_{p_2})v| \le 2\rho$. Taking the sup over $\|v\|_2 = 1$ yields $\|S_{p_1} - S_{p_2}\| \le 2\rho$. Since $L(p_1, \theta^*(p_2)) = \|S_{p_1} - S_{p_2}\|$, this gives the desired modulus bound. □

Sufficient conditions for W_1-resilience. Recall that for mean estimation under TV perturbation, any distribution with bounded ψ-norm was $(\mathcal{O}(\epsilon\psi^{-1}(1/\epsilon)), \epsilon)$-resilient. In particular, bounded covariance distributions were $(\mathcal{O}(\sqrt{\epsilon}), \epsilon)$-resilient. We have an analogous result for W_1-resilience, but with a modified ψ function:

Proposition 1.7. *Let ψ be an Orlicz function, and define $\tilde{\psi}(x) = x\psi(2x)$. Suppose that X (not $X - \mu$) has bounded $\tilde{\psi}$-norm: $\mathbb{E}_p[\tilde{\psi}(|v^\top X|/\sigma)] \le 1$ for all unit vectors v. Also assume that the second moment of p is at most σ^2. Then p is (ρ, ϵ) resilient for $\rho = \max(\sigma\epsilon\psi^{-1}(\frac{2\sigma}{\epsilon}), 4\epsilon^2 + 2\epsilon\sigma)$.*

Let us interpret Proposition 1.7 before giving the proof. Take for instance $\psi(x) = x^2$. Then Proposition 1.7 asks for the 3rd moment to be bounded by $\sigma^3/4$. In that case we have $\rho = \sigma\epsilon\psi^{-1}(2\sigma/\epsilon) = \sqrt{2}\sigma^{3/2}\epsilon^{1/2}$. If the units seem weird, remember that ϵ has units of distance (before it was unitless) and hence $\sigma^{3/2}\epsilon^{1/2}$ has quadratic units, which matches the second moment estimation task.

More generally, taking $\psi(x) = x^k$, we ask for a $(k+1)$st moment bound and get error $\mathcal{O}(\sigma^{1+1/k}\epsilon^{1-1/k})$.

We now turn to proving Proposition 1.7. A helpful auxiliary lemma (here and later) proves a way to use Orlicz norm bounds:

Lemma 1.8. *Let p and q be two distributions over $\mathcal{X}$, $g : \mathcal{X} \to \mathbb{R}$ be any function, c be a non-negative cost function, and ψ be an Orlicz function. Then for any coupling $\pi_{p,q}$ between p and q and any $\sigma > 0$ we have*

$$|\mathbb{E}_{X\sim p}[g(X)] - \mathbb{E}_{Y\sim q}[g(Y)]| \le \sigma\mathbb{E}_{\pi_{p,q}}[c(X,Y)]\psi^{-1}\left(\frac{\mathbb{E}_{\pi_{p,q}}\left[c(X,Y)\psi\left(\frac{|g(X)-g(Y)|}{\sigma c(X,Y)}\right)\right]}{\mathbb{E}_{\pi_{p,q}}[c(X,Y)]}\right). \tag{8}$$

Proof. Note that $|\mathbb{E}_p[g(X)] - \mathbb{E}_q[g(Y)]| = |\mathbb{E}_\pi[g(X) - g(Y)]|$. We weight the coupling π by the cost c to obtain a new probability measure $\pi'(x, y) = c(x, y)\pi(x, y)/\mathbb{E}[c(x, y)]$. We apply Jensen's inequality under π' as follows:

$$\psi\Big(\Big|\frac{\mathbb{E}_\pi[g(X) - g(Y)]}{\sigma\mathbb{E}_\pi[c(X,Y)]}\Big|\Big) = \psi\Big(\Big|\mathbb{E}_\pi\Big[\frac{c(X,Y)}{\mathbb{E}[c(X,Y)]} \cdot \frac{g(X) - g(Y)}{\sigma c(X,Y)}\Big]\Big|\Big) \tag{9}$$

$$= \psi\Big(\Big|\mathbb{E}_{\pi'}\Big[\frac{g(X) - g(Y)}{\sigma c(X,Y)}\Big]\Big|\Big) \tag{10}$$

$$\le \mathbb{E}_{\pi'}\Big[\psi\Big(\frac{|g(X) - g(Y)|}{\sigma c(X,Y)}\Big)\Big] \tag{11}$$

$$= \mathbb{E}_\pi\Big[c(X,Y)\psi\big(\frac{|g(X) - g(Y)|}{\sigma c(X,Y)}\big)\Big]/\mathbb{E}_\pi[c(X,Y)]. \tag{12}$$

Inverting ψ yields the desired result. □

Proof of Proposition 1.7. We apply Lemma 1.8 with $q = r$ an ϵ-friendly perturbation of p under $\langle x, v\rangle^2$, and $g = \langle x, v\rangle^2$; we will also use cost $c'(x, y) = |v^\top(x - y)|$, which satisfies $c'(x, y) \le c(x, y)$. Taking π to be the

ϵ-friendly coupling (under c, not c') between p and r yields

$$|\mathbb{E}_p[\langle x,v\rangle^2] - \mathbb{E}_r[\langle x,v\rangle^2]| \le \sigma\epsilon\psi^{-1}\left(\frac{\mathbb{E}_\pi\left[|\langle x-y,v\rangle|\psi\left(\frac{|\langle x,v\rangle^2 - \langle y,v\rangle^2|}{\sigma|\langle x-y,v\rangle|}\right)\right]}{\epsilon}\right) \tag{13}$$

$$= \sigma\epsilon\psi^{-1}\left(\frac{\mathbb{E}_\pi\left[|\langle x-y,v\rangle|\psi\big(|\langle x,v\rangle + \langle y,v\rangle|/\sigma\big)\right]}{\epsilon}\right). \tag{14}$$

Now we will split into two cases. First, we observe that the worst-case friendly perturbation will either move all of the $\langle x,v\rangle^2$ upwards, or all of the $\langle x,v\rangle^2$ downwards, since otherwise we could take just the upwards part or just the downwards part and perturb the mean further. In other words, we either have (i) $\langle x,v\rangle^2 \ge \langle y,v\rangle^2$ for all $(x,y) \in \mathrm{supp}(\pi)$ with $x \ne y$, or (ii) $\langle x,v\rangle^2 \le \langle y,v\rangle^2$ for all $(x,y) \in \mathrm{supp}(\pi)$ with $x \ne y$. We analyze each case in turn.

Case (i): y moves downwards. In this case we can use the bounds $|\langle x-y,v\rangle| \le 2|\langle x,v\rangle|$ and $|\langle x+y,v\rangle| \le 2|\langle x,v\rangle|$ together with (14) to conclude that

$$|\mathbb{E}_p[\langle x,v\rangle^2] - \mathbb{E}_r[\langle x,v\rangle^2]| \le \sigma\epsilon\psi^{-1}\Big(\mathbb{E}_\pi\Big[2|\langle x,v\rangle|\psi\Big(\frac{2|\langle x,v\rangle|}{\sigma}\Big)\Big]/\epsilon\Big) \tag{15}$$

$$= \sigma\epsilon\psi^{-1}\Big(\mathbb{E}_p\Big[2\sigma\tilde{\psi}\Big(\frac{|\langle x,r\rangle|}{\sigma}\Big)\Big]/\epsilon\Big) \tag{16}$$

$$\le \sigma\epsilon\psi^{-1}(2\sigma/\epsilon), \tag{17}$$

where the final inequality is by bounded Orlicz norm of p.

Case (ii): y moved upwards. In this case by friendliness we have that $|\langle y,v\rangle|^2 \le v^\top S_r v$ whenever $(x,y) \in \mathrm{supp}(\pi)$ and $y \ne x$. Thus

$$|\langle x-y,v\rangle|\psi(|\langle x,v\rangle + \langle y,v\rangle|/\sigma) \le |\langle x-y,v\rangle|\psi(2|\langle y,v\rangle|/\sigma) \le |\langle x-y,v\rangle|\psi(2\sqrt{v^\top S_r v}/\sigma). \tag{18}$$

for all $(x,y) \in \mathrm{supp}(\pi)$. Plugging back into (14) yields

$$|\mathbb{E}_p[\langle x,v\rangle^2] - \mathbb{E}_r[\langle x,v\rangle^2]| \le \sigma\epsilon\psi^{-1}(\mathbb{E}_\pi[|\langle x-y,v\rangle|\psi(2\sqrt{v^\top S_r v}/\sigma)]/\epsilon) \tag{19}$$

$$\le \sigma\epsilon\psi^{-1}(\epsilon\cdot\psi(2\sqrt{v^\top S_r v}/\sigma)/\epsilon) \tag{20}$$

$$= \sigma\epsilon\cdot 2v^\top S_r v/\sigma = 2\epsilon\sqrt{v^\top S_r v}. \tag{21}$$

Here the final inequality is because $\mathbb{E}_\pi[|\langle x-y,v\rangle|] \le \mathbb{E}_\pi[c(x,y)] \le \epsilon$ under the coupling. Comparing the left-hand-side to the final right-hand-side yields $|v^\top S_p v - v^\top S_r v| \le 2\epsilon\sqrt{v^\top S_r v}$. Thus defining $\Delta = |v^\top S_p v - v^\top S_r v|$ and using the fact that $v^\top S_p v \le \sigma^2$, we obtain $\Delta \le 2\epsilon\sqrt{\Delta+\sigma^2}$, which implies (after solving the quadratic) that $\Delta \le 4\epsilon^2 + 2\epsilon\sigma$.

Thus overall we have $|\mathbb{E}_p[\langle x,v\rangle^2] - \mathbb{E}_r[\langle x,v\rangle^2]| \le \max(\sigma\epsilon\psi^{-1}(2\sigma/\epsilon), 4\epsilon^2 + 2\epsilon\sigma)$, as was to be shown. □

1.2 Other Results

Our understanding of robust estimation under W_c distances is still rudimentary. Below are a couple of known results, but many of these may be improved or extended in the near future (perhaps by you!).

The most straightforward extension is from second moment estimation to kth moment estimation. In that case instead of using $\tilde{\psi}(x) = x\psi(2x)$, we use $\tilde{\psi}(x) = x\psi(kx^{k-1})$. Essentially the same proof goes through.

We can also extend to more general loss functions $L(p,\theta)$, as long as L is a convex function of p for fixed θ (this holds e.g. for any $L(p,\theta) = \mathbb{E}_{x\sim p}[\ell(\theta;x)]$, since these loss functions are linear in p and hence also convex). Here the main challenge is defining an appropriate family $\mathcal{F}$ of functions for which to consider friendly perturbations. For second moment estimation our family $\mathcal{F}$ was motivated by the obsevation that $L(p,S) = \sup\{|\mathbb{E}_p[f_v(x)] - \mathbb{E}_q[f_v(x)]| \mid f_v(x) = \langle x,v\rangle^2, \|v\|_2 = 1\}$, but such linear structure need not hold in general. But we can still exploit linear structure by looking at subgradients of the loss. In particular, we can take the Fenchel-Moreau representation

$$L(p,\theta) = \sup_{f\in\mathcal{F}_\theta} \mathbb{E}_{x\sim p}[f(x)] - L^*(f,\theta), \tag{22}$$

which exists for some $\mathcal{F}_\theta$ and L^* whenever $L(p, \theta)$ is convex in p. The family $\mathcal{F}_\theta$ is roughly the family of subgradients of $L(p, \theta)$ as p varies for fixed θ. In this case we obtain conditions $G_\downarrow$ and $G_\uparrow$ as before, asking that

$$\mathbb{E}_r[f(x)] - L^*(f, \theta^*(p)) \le \rho_1 \text{ for all } f \in \mathcal{F}_{\theta^*(p)} \text{ and } \epsilon\text{-friendly } r, \tag{$\downarrow$}$$

and furthermore

$$L(p, \theta) \le \rho_2 \text{ if for every } f \in \mathcal{F}_\theta \text{ there is an } \epsilon\text{-friendly } r \text{ such that } \mathbb{E}_r[f(x)] - L^*(f, \theta) \le \rho_1. \tag{$\uparrow$}$$

Note that for the second condition ($\mathcal{G}_\downarrow$), we allow the perturbation r to depend on the current function f. If r was fixed this would closely match the old definition, but we can only do that for deletions since in general even the set of feasible r depends on f.

Using this, we can (after sufficient algebra) derive sufficient conditions for robust linear regression under W_1, for conditions similar to the hypercontractivity condition from before. This will be a challenge problem on the homework.

Finally, we can define a $\tilde{W}_1$ similar to $\tilde{\mathsf{TV}}$, but our understanding of it is far more rudimentary. In particular, known analyses do not seem to yield the correct finite-sample rates (for instance, the rate of convergence includes an $n^{-1/3}$ term that seems unlikely to actually exist).