12 Minimax

Goal

Minimax formulation, weak and strong duality, saddle point, robust optimization, minimax theorem, alternating, oscillation, Uzawa's algorithm, gradient-descent-ascent

Alert 12.1: Convention

Gray boxes are not required hence can be omitted for unenthusiastic readers.

This note is likely to be updated again soon.

Definition 12.2: Minimax problem

In this and the following few lectures we are interested in solving the minimax problem:

$$\mathfrak{p}_\star = \inf_{\mathbf{w}\in\mathbb{W}} \sup_{\mathbf{z}\in\mathbb{W}} f(\mathbf{w},\mathbf{z}), \tag{12.1}$$

where $\mathbb{W} \subseteq \mathbb{R}^p$, $\mathbb{Z} \subseteq \mathbb{R}^d$ and $f : \mathbb{W} \times \mathbb{Z} \to \mathbb{R}$ is a (block) bivariate function. Equivalently, introducing the upper and lower envelope functions

$$\overline{f}(\mathbf{w}) := \sup_{\mathbf{z}\in\mathbb{Z}} f(\mathbf{w},\mathbf{z}), \qquad \underline{f}(\mathbf{z}) := \inf_{\mathbf{w}\in\mathbb{W}} f(\mathbf{w},\mathbf{z}),$$

we may rewrite the minimax problem as the familiar minimization problem:

$$\mathfrak{p}_\star = \inf_{\mathbf{w}\in\mathbb{W}} \overline{f}(\mathbf{w}) \tag{12.2}$$

and the closely related "twin" (or dual) maximin problem:

$$\mathfrak{d}^\star = \left[\sup_{\mathbf{z}\in\mathbb{Z}} \inf_{\mathbf{w}\in\mathbb{W}} f(\mathbf{w},\mathbf{z})\right] = \sup_{\mathbf{z}\in\mathbb{Z}} \underline{f}(\mathbf{z}), \tag{12.3}$$

where the ordering of the inf and sup has been switched. Note that even for a smooth function f the envelopes $\underline{f}$ and $\overline{f}$ may still be nonsmooth so the equivalent problem (12.2) usually amounts to minimizing a nonsmooth function (and similarly for (12.3)).

For later use, let us define the two optimal sets:

$$\mathbb{W}_\star := \operatorname*{argmin}_{\mathbf{w}\in\mathbb{W}} \overline{f}(\mathbf{w}), \qquad \mathbb{Z}^\star := \operatorname*{argmax}_{\mathbf{z}\in\mathbb{Z}} \underline{f}(\mathbf{z}). \tag{12.4}$$

For $\mathbf{w} \in \mathbb{W}$ and $\mathbf{z} \in \mathbb{Z}$ we also define the sets

$$\mathbb{Z}^{\mathbf{w}} := \mathbb{Z}(\mathbf{w}) := \operatorname*{argmax}_{\mathbf{z}\in\mathbb{Z}} f(\mathbf{w},\mathbf{z}), \qquad \mathbb{W}_{\mathbf{z}} := \mathbb{W}(\mathbf{z}) := \operatorname*{argmin}_{\mathbf{w}\in\mathbb{W}} f(\mathbf{w},\mathbf{z}). \tag{12.5}$$

Theorem 12.3: Weak duality

Weak duality, i.e. $\mathfrak{p}_\star \geq \mathfrak{d}^\star$, *always holds.* ■

When equality holds we say strong duality holds.

Definition 12.4: Saddle point in minimax problems

We call the pair $(\mathbf{w}_\star, \mathbf{z}^\star) \in \mathbb{W} \times \mathbb{Z}$ a saddle point of $f(\mathbf{w}, \mathbf{z})$ over $\mathbb{W} \times \mathbb{Z}$ if

$$\forall \mathbf{w} \in \mathbb{W},\ \forall \mathbf{z} \in \mathbb{Z}, \quad f(\mathbf{w}_\star, \mathbf{z}) \leq f(\mathbf{w}_\star, \mathbf{z}^\star) \leq f(\mathbf{w}, \mathbf{z}^\star). \tag{12.6}$$

In other words,

- fixing $\mathbf{w}_\star$, $\mathbf{z}^\star \in \mathrm{argmax}_{\mathbf{z} \in \mathbb{Z}} f(\mathbf{w}_\star, \mathbf{z})$, as can be seen from the left inequality in (12.6);
- fixing $\mathbf{z}^\star$, $\mathbf{w}_\star \in \mathrm{argmin}_{\mathbf{w} \in \mathbb{W}} f(\mathbf{w}, \mathbf{z}^\star)$, as can be seen from the right inequality in (12.6).

We will study algorithms that find saddle points, i.e. solve the primal problem (12.2) and the dual problem (12.3) *simultaneously.*

Alert 12.5: This saddle point is not that saddle point!

The name saddle point is also used to refer to points where the gradient ∇f vanishes but the Hessian $\nabla^2 f$ is indefinite. Do not confuse it with the saddle point in our minimax setting (although the two are actually related).

Theorem 12.6: Strong duality and saddle points

The following are true:

- *If there exists a saddle point, then strong duality holds and $\mathbb{W}_\star \times \mathbb{Z}^\star$ is the set of* all *saddle points.*
- *If both $\mathbb{W}_\star$ and $\mathbb{Z}^\star$ are nonempty, then strong duality holds iff there exists a saddle point.*
- *If strong duality holds, then $(\mathbf{w}_\star, \mathbf{z}^\star)$ is a saddle point iff $\mathbf{w}_\star \in \mathbb{W}_\star$ and $\mathbf{z}^\star \in \mathbb{Z}^\star$.*

Proof: Indeed, when an (arbitrary) saddle point $(\mathbf{w}_\star, \mathbf{z}^\star)$ exists, we have

$$\mathfrak{p}_\star := \left[\inf_{\mathbf{w} \in \mathbb{W}} \overline{f}(\mathbf{w})\right] \leq \overline{f}(\mathbf{w}_\star) \overset{(12.6)}{=} f(\mathbf{w}_\star, \mathbf{z}^\star) \overset{(12.6)}{=} \underline{f}(\mathbf{z}^\star) \leq \left[\sup_{\mathbf{z} \in \mathbb{Z}} \underline{f}(\mathbf{z})\right] =: \mathfrak{d}^\star,$$

hence equality, i.e. strong duality, follows from (12.3). Since we have in fact equality throughout above, it follows that $\mathbf{w}_\star \in \mathbb{W}_\star$ and $\mathbf{z}^\star \in \mathbb{Z}^\star$.

Conversely, for any $\mathbf{w}_\star \in \mathbb{W}_\star$ and $\mathbf{z}^\star \in \mathbb{Z}^\star$, we have

$$\mathfrak{p}_\star := \left[\inf_{\mathbf{w} \in \mathbb{W}} \overline{f}(\mathbf{w})\right] = \overline{f}(\mathbf{w}_\star) := \left[\sup_{\mathbf{z} \in \mathbb{Z}} f(\mathbf{w}_\star, \mathbf{z})\right] \geq f(\mathbf{w}_\star, \mathbf{z}^\star) \geq \left[\inf_{\mathbf{w} \in \mathbb{W}} f(\mathbf{w}, \mathbf{z}^\star)\right] =: \underline{f}(\mathbf{z}^\star) = \left[\sup_{\mathbf{z} \in \mathbb{Z}} \underline{f}(\mathbf{z})\right] =: \mathfrak{d}^\star.$$

Thanks to strong duality, we have in fact equality throughout above. Thus, $(\mathbf{w}_\star, \mathbf{z}^\star)$ is a saddle point. ■

We point out that strong duality may still hold even when there is no saddle point, due to non-attainment of the infimum or supremum (i.e. $\mathbb{W}_\star = \emptyset$ and/or $\mathbb{Z}^\star = \emptyset$).

Alert 12.7: Stability

Let $(\mathbf{w}_\star, \mathbf{z}^\star)$ be a saddle point of f over $\mathbb{W} \times \mathbb{Z}$. Clearly, from the definitions (12.6) and (12.5) we have

$$\mathbb{W}_\star \subseteq \mathbb{W}(\mathbf{z}^\star), \qquad \mathbb{Z}^\star \subseteq \mathbb{Z}(\mathbf{w}_\star), \tag{12.7}$$

where the containment may be strict. We call the saddle point $(\mathbf{w}_\star, \mathbf{z}^\star)$ stable if equality holds in (12.7).

On the other hand, if both $(\mathbf{w}_\star, \mathbf{z}^\star)$ and $(\mathbf{u}_\star, \mathbf{v}^\star)$ are saddle points, then so are $(\mathbf{w}_\star, \mathbf{v}^\star)$ and $(\mathbf{u}_\star, \mathbf{z}^\star)$.

Example 12.8: Nonsmoothness arising from minimax

Consider the trivial nonsmooth minimization problem

$$0 = \mathfrak{p}_\star = \min_w \overline{f}(w), \quad \text{where} \quad \overline{f}(w) = |w|,$$

which can be rewritten as an equivalent smooth minimax problem

$$\min_w \max_{|z|\le 1} wz,$$

where the function $f(w, z) = wz$ clearly is smooth and convex in w and concave in z. The corresponding maximin problem is

$$0 = \mathfrak{d}^\star = \max_{|z|\le 1} \underline{f}(z), \quad \text{where} \quad \underline{f}(z) = \inf_w wz = \begin{cases} -\infty, & \text{if } z \neq 0 \\ 0, & \text{if } z = 0 \end{cases}.$$

Since $\mathfrak{p}_\star = \mathfrak{d}^\star = 0$, strong duality holds. Clearly, $\mathbb{W}_\star = \{0\}$ while $\mathbb{Z}^\star = \{0\}$ so we have a unique saddle point, which is not stable. Indeed, $\mathbb{W}(0) = \mathbb{R} \supsetneq \mathbb{W}_\star$ while $\mathbb{Z}(0) = [-1, 1] \supsetneq \mathbb{Z}^\star$.

Example 12.9: Fenchel-Rockafellar duality

More generally, we may derive the Fenchel-Rockafellar duality from minimax formulations:

$$\begin{aligned}\left[\inf_{\mathbf{w}} \; g(A\mathbf{w}) + h(\mathbf{w})\right] = \left[\inf_{\mathbf{w}} \sup_{\mathbf{z}} \; \underbrace{\langle A\mathbf{w}; \mathbf{z}\rangle - g^*(\mathbf{z}) + h(\mathbf{w})}_{f(\mathbf{w},\mathbf{z})}\right] &\ge \left[\sup_{\mathbf{z}} \inf_{\mathbf{w}} \langle A\mathbf{w}; \mathbf{z}\rangle - g^*(\mathbf{z}) + h(\mathbf{w})\right] \\ &= -\inf_{\mathbf{z}} \sup_{\mathbf{w}} \langle \mathbf{w}; -A^\top \mathbf{z}\rangle + g^*(\mathbf{z}) - h(\mathbf{w}) \\ &= -\inf_{\mathbf{z}} g^*(\mathbf{z}) + h^*(-A^\top \mathbf{z})\end{aligned}$$

where the function f is convex in $\mathbf{w}$ and concave in $\mathbf{z}$, provided that h and g are both convex. Conditions for strong duality include:

- $\mathbf{0} \in \operatorname{core}(\operatorname{dom} g - A \operatorname{dom} h)$, i.e. for any $\mathbf{d}$ there exists some $\lambda = \lambda(\mathbf{d}) > 0$ such that for any $t \in [0, \lambda]$, there exists $\mathbf{w} \in \operatorname{dom} h$ so that $A\mathbf{w} + t\mathbf{d} \in \operatorname{dom} g$.
- $A \operatorname{dom} h \cap \operatorname{cont}(g) \neq \emptyset$, where $\operatorname{cont}(g)$ is the set of points at which g is continuous.

Example 12.10: Robust optimization (Ben-Tal et al. 2009)

Real datasets are noisy and sometimes contain even gross (human) errors. It is thus natural to learn models that are robust against worst-case perturbations:

$$\inf_{\mathbf{w}} \; \mathop{\mathbb{E}}_{(\mathbf{x},y)\sim\mathcal{D}} \left[\sup_{\|\mathbf{z}\|\le\epsilon} \ell(y, \langle \mathbf{x} + \mathbf{z}; \mathbf{w}\rangle)\right], \quad \text{or equivalently} \quad \inf_{\mathbf{w}} \sup_{\|\mathbf{z}(\cdot)\|\le\epsilon} \mathop{\mathbb{E}}_{(\mathbf{x},y)\sim\mathcal{D}} \ell(y, \langle \mathbf{x} + \mathbf{z}(\mathbf{x}, y); \mathbf{w}\rangle).$$

In adversarial ML, we may interpret the minimizer as a defender that tries to learn a good model $\mathbf{w}$ and the maximizer as an attacker that tries to construct a difficult dataset through perturbations $\mathbf{z}$. The dual problem

$$\sup_{\|\mathbf{z}(\cdot)\|\leq\epsilon} \inf_{\mathbf{w}} \mathop{\mathbb{E}}_{(\mathbf{x},y)\sim\mathcal{D}} \ell(y, \langle \mathbf{x}+\mathbf{z}(\mathbf{x},y); \mathbf{w}\rangle)$$

represent the opposite scenario where the attacker acts first while the defender responds.

More generally, one may consider perturbing the distribution $\mathcal{D}$ under some metric dist:

$$\inf_{\mathbf{w}} \sup_{\mathrm{dist}(\tilde{\mathcal{D}},\mathcal{D})\leq\epsilon} \mathop{\mathbb{E}}_{(\mathbf{x},y)\sim\tilde{\mathcal{D}}} \ell(y, \langle \mathbf{x}; \mathbf{w}\rangle) \qquad \geq \qquad \sup_{\mathrm{dist}(\tilde{\mathcal{D}},\mathcal{D})\leq\epsilon} \inf_{\mathbf{w}} \mathop{\mathbb{E}}_{(\mathbf{x},y)\sim\tilde{\mathcal{D}}} \ell(y, \langle \mathbf{x}; \mathbf{w}\rangle),$$

which is known as distributionally robust optimization.

Ben-Tal, A., L. E. Ghaoui, and A. Nemirovski (2009). “Robust Optimization”. Princeton University Press.

Exercise 12.11: Lasso revisited

Let us consider the familiar (square root) linear regression problem:

$$\inf_{\mathbf{w}} \|X\mathbf{w}-\mathbf{y}\|_2, \quad \text{where} \quad X = [\mathbf{x}_1, \ldots, \mathbf{x}_n]^\top.$$

Now suppose we perturb each *feature*, i.e., columns in X, independently, arriving at the robust linear regression problem:

$$\inf_{\mathbf{w}} \sup_{\forall j, \|\mathbf{z}_j\|_2\leq\lambda} \|(X+Z)\mathbf{w}-\mathbf{y}\|_2,$$

where the perturbation matrix $Z = [\mathbf{z}_1, \ldots, \mathbf{z}_d]$. Prove that robust linear regression is exactly equivalent to (square-root) Lasso (note the absence of the square on the ℓ_2 norm):

$$\inf_{\mathbf{w}\in\mathbb{R}^d} \|X\mathbf{w}-\mathbf{y}\|_2 + \lambda\|\mathbf{w}\|_1,$$

where recall that $\|\mathbf{w}\|_1 = \sum_j |w_j|$.

Exercise 12.12: Robust empirical risk minimization

Consider the familiar empirical risk minimization with a loss $\ell : \mathbb{R}\times\mathbb{R}\to\mathbb{R}$ that is convex in the second input, i.e. the function $\ell_y(\cdot) := \ell(y, \cdot)$ is convex. Prove that if we perturb each data instance instead, then

$$\sup_{\|\mathbf{z}\|\leq\epsilon} \ell(y, \langle \mathbf{x}+\mathbf{z}; \mathbf{w}\rangle) = \max\begin{cases} \sup_{y^*\geq 0} y^*(\langle \mathbf{x}; \mathbf{w}\rangle + \epsilon\|\mathbf{w}\|_\circ) - \ell_y^*(y^*) \\ \sup_{y^*\leq 0} y^*(\langle \mathbf{x}; \mathbf{w}\rangle - \epsilon\|\mathbf{w}\|_\circ) - \ell_y^*(y^*) \end{cases}.$$

If ℓ_y is decreasing (as in SVM, logistic regression, etc., for a positive instance), then

$$\sup_{\|\boldsymbol{\delta}\|\leq\epsilon} \ell(y, \langle \mathbf{x}+\boldsymbol{\delta}; \mathbf{w}\rangle) = \sup_{y^*\leq 0} y^*(\langle \mathbf{x}; \mathbf{w}\rangle - \epsilon\|\mathbf{w}\|_\circ) - \ell_y^*(y^*) = \ell(y, \langle \mathbf{x}; \mathbf{w}\rangle - \epsilon\|\mathbf{w}\|_\circ),$$

which agrees with a direct calculation. Obviously, the result for a negative instance is the other case.

Note however that the robust loss we derived on the right-hand side may no longer be convex in $\mathbf{w}$.

Alert 12.13: Convex games

In most of our results below, for simplicity we assume $\mathbb{W}$ and $\mathbb{Z}$ to be closed convex, and f to be smooth and convex in $\mathbf{w}$ while concave in $\mathbf{z}$. There is significant interest in extending the algorithms and analyses to the nonconvex setting, see e.g. Zhang et al. (2020) and the references therein.

Under the above convexity assumptions, we recognize that the upper envelope $\overline{f}(\mathbf{w})$ is convex and the lower envelope $\underline{f}(\mathbf{z})$ is concave, hence both the primal and dual problems in (12.4) are convex programs, and the saddle point set $\mathbb{W}_\star \times \mathbb{Z}^\star$ is always convex.

Zhang, G., P. Poupart, and Y. Yu (2020). "Optimality and Stability in Non-convex Smooth Games".

Definition 12.14: Quasiconvexity

We call a function f quasiconvex if all its sublevel sets are convex, i.e. $\llbracket f \leq t \rrbracket$ is convex for all $t \in \mathbb{R}$. Or equivalently, if for all $\mathbf{x}, \mathbf{y}$ and $\lambda \in [0,1]$,

$$f(\lambda \mathbf{x} + (1-\lambda)\mathbf{y}) \leq \max\{f(\mathbf{x}), f(\mathbf{y})\}.$$

Obviously, a function is quasiconcave iff its negation is quasiconvex. In contrast, recall that a function f is convex if its epigraph $\operatorname{epi} f = \{(\mathbf{x}, t) : f(\mathbf{x}) \leq t\}$ is convex. Clearly, any convex function is quasiconvex but the converse may fail.

We point out that quasiconvexity, unlike convexity, in general is not preserved under summation!

Theorem 12.15: Minimax theorem (in topological vector spaces, TVS)

Let $f(\mathbf{w}, \mathbf{z}) : \mathbb{W} \times \mathbb{Z} \to \mathbb{R}$ be a real-valued function, where $\mathbb{W}$ and $\mathbb{Z}$ are convex subsets of TVS $\mathcal{W}$ and $\mathcal{Z}$, respectively. Suppose

- *$f(\mathbf{w}, \cdot) : \mathbb{Z} \to \mathbb{R}$ is semicontinuous (on line segments) and quasi-concave on $\mathbb{Z}$ for each $\mathbf{w} \in \mathbb{W}$;*
- *$f(\cdot, \mathbf{z}) : \mathbb{W} \to \mathbb{R}$ is l.s.c. and quasi-convex on $\mathbb{W}$ for each $\mathbf{z} \in \mathbb{Z}$;*
- *For some finite $F \subseteq \mathbb{Z}$, $\max_{\mathbf{z} \in F} f(\cdot, \mathbf{z})$ is inf-compact, i.e. $\bigcap_{\mathbf{z} \in F} \{\mathbf{w} \in \mathbb{W} : f(\mathbf{w}, \mathbf{z}) \leq \alpha\}$ is compact for all $\alpha \in \mathbb{R}$;*

then strong duality holds and the minimum of the primal problem is attained:

$$\min_{\mathbf{w} \in \mathbb{W}} \sup_{\mathbf{z} \in \mathbb{Z}} f(\mathbf{w}, \mathbf{z}) = \sup_{\mathbf{z} \in \mathbb{Z}} \inf_{\mathbf{w} \in \mathbb{W}} f(\mathbf{w}, \mathbf{z}).$$

A similar statement holds by swapping the role of $\mathbf{w}$ and $\mathbf{z}$.

Proof: The proof here is based on Komiya (1988). We may assume w.l.o.g. $f(\cdot, \mathbf{z}_c)$ is inf-compact for some $\mathbf{z}_c \in \mathbb{Z}$. Note that in this case the left infimum over $\mathbb{W}$ is attained (while the right infimum is also attained if $f(\cdot, \mathbf{z})$ is inf-compact for all $\mathbf{z}$).

Let $\alpha < \beta < \inf_{\mathbf{w} \in \mathbb{W}} \sup_{\mathbf{z} \in \mathbb{Z}} f(\mathbf{w}, \mathbf{z})$, we need only prove $\alpha < \sup_{\mathbf{z} \in \mathbb{Z}} \inf_{\mathbf{w} \in \mathbb{W}} f(\mathbf{w}, \mathbf{z})$, i.e., there exists some $\mathbf{z}^* \in \mathbb{Z}$ such that $\alpha < \inf_{\mathbf{w} \in \mathbb{W}} f(\mathbf{w}, \mathbf{z}^*)$. For each $\mathbf{z} \in \mathbb{Z}$, define

$$\mathbb{W}_{\mathbf{z}}(t) := \{\mathbf{w} \in \mathbb{W} : f(\mathbf{w}, \mathbf{z}) \leq t\},$$

which, by assumption, is closed and convex. Clearly, $\bigcap_{\mathbf{z} \in \mathbb{Z}} \mathbb{W}_{\mathbf{z}}(\beta) = \emptyset$. Since $\mathbb{W}_{\mathbf{z}_c}(\beta)$ is compact, there exist finitely many $\mathbf{z}_1, \ldots, \mathbf{z}_n \in \mathbb{Z}$ such that $\bigcap_{i=1}^n \mathbb{W}_{\mathbf{z}_i}(\beta) = \emptyset$, that is, $\alpha < \inf_{\mathbf{w} \in \mathbb{W}} \max_{1 \leq i \leq n} f(\mathbf{w}, \mathbf{z}_i)$. We want to prove the existence of $\mathbf{z}^* \in \mathbb{Z}$ such that $\alpha < \inf_{\mathbf{w} \in \mathbb{W}} f(\mathbf{w}, \mathbf{z}^*)$. The result clearly holds if $n = 1$ (simply take $\mathbf{z}^* = \mathbf{z}_1$). Suppose the result holds for $n = k - 1$, and

Claim: If $\alpha < \inf_{\mathbf{w} \in \mathbb{W}} f(\mathbf{w}, \mathbf{z}_1) \vee f(\mathbf{w}, \mathbf{z}_2)$ for any $\mathbf{z}_1, \mathbf{z}_2 \in \mathbb{Z}$, then $\alpha < \inf_{\mathbf{w} \in \mathbb{Z}} f(\mathbf{w}, \mathbf{z}^*)$ for some $\mathbf{z}^* \in \mathbb{Z}$.

Now for $n = k$, $\alpha < \beta < \inf_{\mathbf{w} \in \mathbb{W}} \max_{1 \leq i \leq k} f(\mathbf{w}, \mathbf{z}_i)$ implies $\beta < \inf_{\mathbf{w} \in \mathbb{W}_{\mathbf{z}_k}(\beta)} \max_{1 \leq i \leq k-1} f(\mathbf{w}, \mathbf{z}_i)$. By the induction hypothesis, we have $\beta < \inf_{\mathbf{w} \in \mathbb{W}_{\mathbf{z}_k}(\beta)} f(\mathbf{w}, \mathbf{z}^\star)$ for some $\mathbf{z}^\star \in \mathbb{Z}$. Hence, $\alpha < \beta \leq \inf_{\mathbf{w} \in \mathbb{W}} f(\mathbf{w}, \mathbf{z}^\star) \vee f(\mathbf{w}, \mathbf{z}_k)$. Using the claim we have for some $\mathbf{z}^* \in \mathbb{Z}$ that $\alpha < \inf_{\mathbf{w} \in \mathbb{W}} f(\mathbf{w}, \mathbf{z}^*)$.

To prove the claim, suppose for the sake of contradiction that $\alpha \geq \inf_{\mathbf{w} \in \mathbb{W}} f(\mathbf{w}, \mathbf{z})$ for all $\mathbf{z} \in \mathbb{Z}$. Choose some β such that $\alpha < \beta < \inf_{\mathbf{w} \in \mathbb{W}} f(\mathbf{w}, \mathbf{z}_1) \vee f(\mathbf{w}, \mathbf{z}_2)$. For any $\mathbf{u}$ on the line segment $[\mathbf{z}_1, \mathbf{z}_2]$, define

$\mathbb{W}_{\mathbf{u}}(t)$ as before, which is closed, convex, and nonempty for all $t > \alpha$. Note that $\mathbb{W}_{\mathbf{z}_1}(\beta) \cap \mathbb{W}_{\mathbf{z}_2}(\beta) = \emptyset$. Since $f(\mathbf{w}, \cdot)$ is quasi-concave, we have $\mathbb{W}_{\mathbf{u}}(\beta) \subseteq \mathbb{W}_{\mathbf{z}_1}(\beta) \cup \mathbb{W}_{\mathbf{z}_2}(\beta)$. Since $\mathbb{W}_{\mathbf{u}}(\beta)$ is convex hence connected, we have either $\mathbb{W}_{\mathbf{u}}(t) \subseteq \mathbb{W}_{\mathbf{u}}(\beta) \subseteq \mathbb{W}_{\mathbf{z}_1}(\beta)$ or $\mathbb{W}_{\mathbf{u}}(t) \subseteq \mathbb{W}_{\mathbf{u}}(\beta) \subseteq \mathbb{W}_{\mathbf{z}_2}(\beta)$, where we fix arbitrarily $t \in]\alpha, \beta[$. Thus, we can partition the line segment $[\mathbf{z}_1, \mathbf{z}_2]$ into two disjoint sets I and J where for $\mathbf{u} \in I$, say $\mathbb{W}_{\mathbf{u}}(t) \subseteq \mathbb{W}_{\mathbf{z}_1}(\beta)$, and for $\mathbf{u} \in J$, $\mathbb{W}_{\mathbf{u}}(t) \subseteq \mathbb{W}_{\mathbf{z}_2}(\beta)$. Let $I \ni \mathbf{u}_n \to \mathbf{u} \in [\mathbf{z}_1, \mathbf{z}_2]$. For any $\mathbf{w} \in \mathbb{W}_{\mathbf{u}}(t)$, we have $f(\mathbf{w}, \mathbf{u}) \leq t < \beta$, hence by semicontinuity, there exists some n such that $f(\mathbf{w}, \mathbf{u}_n) < \beta$, i.e., $\mathbf{w} \in \mathbb{W}_{\mathbf{u}_n}(\beta)$. Since $\mathbb{W}_{\mathbf{z}_1}(\beta) \supseteq \mathbb{W}_{\mathbf{u}_n}(t) \subseteq \mathbb{W}_{\mathbf{u}_n}(\beta)$, $\mathbf{w} \in \mathbb{W}_{\mathbf{u}_n}(\beta) \subseteq \mathbb{W}_{\mathbf{z}_1}(\beta)$. Since $\mathbf{w} \in \mathbb{W}_{\mathbf{u}}(t)$ is arbitrary, $\mathbf{u} \in I$, that is, I is closed. Similarly, we can prove J is also closed, which is impossible since $I \cup J = [\mathbf{z}_1, \mathbf{z}_2]$ and $I \cap J = \emptyset$. ■

The inf-compactness assumption is satisfied if (needless to say, a similar result holds for $\mathbf{z}$):

- $\mathbb{W}$ is compact, which is the usual assumption; or
- $\mathbb{W}$ is closed and f is inf-bounded in $\mathbf{z}$, in particular if f is strongly convex in $\mathbf{z}$.

Komiya, H. (1988). "Elementary proof for Sion's Minimax Theorem". *Kodai Mathematical Journal*, vol. 11, no. 1, pp. 5–7.

History 12.16: Minimax theorem

We briefly mention some history behind the development of the minimax theorem. The first nontrivial result is due to von Neumann (1928), where the function f is bilinear and the sets X and Y are simplices in finite dimensional spaces. Note that von Neumann's result was published in 1928, followed by his celebrated game theory book in 1944. The next improvement is due to Kneser (1952), where X and Y are convex sets with X compact and f bilinear and u.s.c. in $\mathbf{x}$ for all $\mathbf{y} \in \mathsf{Y}$. Further refines appeared in Fan (1953), Nikaidô (1954), and cultivated in Sion (1958) which amounts to a compact convex X (or Y) in Theorem 12.15. Wu (1959) made another significant extension by completely removing the vector space structure, which has since been further developed and refined by Tuy (1974) and König (1992).

von Neumann, J. (1928). "Zur Theorie der Gesellschaftsspiele". *Mathematische Annalen*, vol. 100. Translation in Contributions in the Theory of Games IV., pp. 295–320.

Kneser, H. (1952). "Sur un théoréme fondamental de la théorie des jeux". *Comptes rendus de l'Académie des sciences*, vol. 234, no. 1, pp. 2418–2420.

Fan, K. (1953). "Minimax Theorems". *Proceedings of the National Academy of Sciences*, vol. 39, pp. 42–47.

Nikaidô, H. (1954). "On von Neumann's Minimax Theorem". *Pacific Journal of Mathematics*, vol. 4, no. 1, pp. 65–72.

Sion, M. (1958). "On General Minimax Theorems". *Pacific Journal of Mathematics*, vol. 8, no. 1, pp. 171–176.

Wu, W.-T. (1959). "A remark on the fundamental theorem in the theory of games". *Science Record*, vol. 5, pp. 229–233.

Tuy, H. (1974). "On a general minimax theorem". *Soviet Mathematics*, vol. 15, pp. 1689–1693.

König, H. (1992). "A general minimax theorem based on connectedness". *Archiv der Mathematik*, vol. 59, no. 1, pp. 55–64. Archiv der Mathematik, vol.64, 139–143, 1995.

Example 12.17: Lagrangian duality and Slater's condition

Recall that for the generic minimization problem

$$\inf_{\mathbf{w}} \ h(\mathbf{w}) \quad \text{s.t.} \quad \mathbf{g}(\mathbf{w}) \leq 0$$

we may construct the Lagrangian which implicitly removes the functional constraints:

$$\inf_{\mathbf{w}} \sup_{\mathbf{z} \geq 0} \ \underbrace{h(\mathbf{w}) + \langle \mathbf{g}(\mathbf{w}), \mathbf{z} \rangle}_{f(\mathbf{w}, \mathbf{z})}.$$

If h and $\mathbf{g}$ are both (closed) convex, then f is (closed) convex in $\mathbf{w}$ and linear (hence concave) in $\mathbf{z}$. Under Slater's condition, i.e., there exists some $\mathbf{w}_0 \in \operatorname{dom} h$ such that $\mathbf{g}(\mathbf{z}_0) < 0$, we know f is sup-compact in $\mathbf{z}$, i.e., the set

$$\{\mathbf{z} \geq \mathbf{0} : f(\mathbf{w}_0, \mathbf{z}) \geq \alpha\} = \{\mathbf{z} \geq \mathbf{0} : \langle \mathbf{g}(\mathbf{w}_0), \mathbf{z} \rangle \geq \alpha - h(\mathbf{w}_0)\} \subseteq \{\mathbf{0} \leq \mathbf{z} \leq \tfrac{\alpha - h(\mathbf{w}_0)}{\mathbf{g}(\mathbf{w}_0)}\}$$

is compact. Applying the minimax Theorem 12.15 (with $\mathbf{w}$ and $\mathbf{z}$ switched) we obtain strong duality:

$$\inf_{\mathbf{w}} \sup_{\mathbf{z} \geq 0} \underbrace{h(\mathbf{w}) + \langle \mathbf{g}(\mathbf{w}), \mathbf{z} \rangle}_{f(\mathbf{w},\mathbf{z})} \quad = \quad \max_{\mathbf{z} \geq 0} \inf_{\mathbf{w}} \underbrace{h(\mathbf{w}) + \langle \mathbf{g}(\mathbf{w}), \mathbf{z} \rangle}_{f(\mathbf{w},\mathbf{z})}.$$

We caution again that for a given dual solution $\mathbf{z}^\star$, $\mathbb{W}(\mathbf{z}^\star) \supseteq \mathbb{W}_\star$, whereas equality holds if (say) h is strictly convex, in which case the primal solution is unique.

Algorithm 12.18: Alternating, may not work!!!

The saddle point in Definition 12.4 resembles the notion of alternating minimizer (see Definition 13.12) so strikingly that a similar alternating algorithm is inevitable:

```
Algorithm: Alternating Minimax
  Input: $(\mathbf{w}_0, \mathbf{z}_0) \in \mathbb{W} \times \mathbb{Z} \cap \operatorname{dom} f$
  for $t = 0, 1, 2, \ldots$ do
      $\mathbf{w}_{t+1} \leftarrow \operatorname{argmin}_{\mathbf{w} \in \mathbb{W}} \; f(\mathbf{w}, \mathbf{z}_t)$
      $\mathbf{z}_{t+1} \leftarrow \operatorname{argmax}_{\mathbf{z} \in \mathbb{Z}} \; f(\mathbf{w}_{t+1}, \mathbf{z})$        // or $\mathbf{z}_{t+1} \leftarrow \operatorname{argmax}_{\mathbf{z} \in \mathbb{Z}} \; f(\mathbf{w}_t, \mathbf{z})$
```

When f is convex in $\mathbf{w}$ and concave in $\mathbf{z}$, each step is a convex program and can be solved with any sensible algorithm (e.g. projected gradient Algorithm 3.15).

Alert 12.19: Alternating does not work!

Consider the trivial minimax problem:

$$\min_{w \in [-1,1]} \max_{z \in [-1,1]} \quad wz.$$

It is easy to see that strong duality holds (Theorem 12.15) and

$$\overline{f}(w) = |w|, \qquad \underline{f}(z) = -|z|,$$

so that we have a unique saddle point $(w_\star, z^\star) = (0, 0)$, which is not stable: $\mathbb{W}(0) = [-1, 1] \supsetneq \mathbb{W}_\star = \{0\}$ and similarly $\mathbb{Z}(0) = [-1, 1] \supsetneq \mathbb{Z}^\star = \{0\}$. Applying the alternating Algorithm 12.18 with any $z_0 \neq 0$ we obtain

$$z_0 \neq 0 \implies w_1 = z_1 = -\operatorname{sign}(z_0) \implies w_2 = z_2 = \operatorname{sign}(z_0) \implies w_3 = z_3 = -\operatorname{sign}(z_0) \implies \cdots,$$

which oscillates between $w = z = -1$ and $w = z = 1$ hence never converges to the unique saddle point!

Alert 12.20: Alternating does not work?

Consider the modified minimax problem:

$$\min_{w \in [-1,1]} \max_{z \in [-1,1]} \quad z \exp(w).$$

It is easy to see that strong duality holds (Theorem 12.15) and

$$\overline{f}(w) = \exp(w), \qquad \underline{f}(z) = z \exp(-\operatorname{sign}(z)),$$

so that we have a unique saddle point $(w_\star, z^\star) = (-1, 1)$ which is now stable. Applying the alternating

Algorithm 12.18 with any z_0 we obtain

$$w_1 = -\operatorname{sign}(z_0), z_1 = 1 \implies w_2 = -1, z_2 = 1 \implies w_3 = -1, z_3 = 1 \implies \cdots,$$

which converges to the unique saddle point in two iterations!

Algorithm 12.21: Uzawa's algorithm (Uzawa 1958)

By interpretting the minimax problem (12.1) as a nonsmooth minimization problem (12.2) we can simply apply the subgradient Algorithm 5.14, as long as we can compute a subgradient of $\overline{f}(\mathbf{w})$. This last missing piece was supplied by Danskin (1967) (for convex functions) and by Dem'yanov (1969) (for differentiable functions), and leads to the following algorithm of Uzawa (1958):

```
Algorithm: Uzawa's algorithm for minimax
   Input: $(\mathbf{w}_0, \mathbf{z}_0) \in \mathbb{W} \times \mathbb{Z} \cap \operatorname{dom} f$
1  for $t = 0, 1, \ldots$ do
2      $\mathbf{z}_t = \operatorname{argmax}_{\mathbf{z} \in \mathbb{Z}} f(\mathbf{w}_t, \mathbf{z})$        // solve inner maximization exactly
3      compute subgradient $\mathbf{g}_t = \partial_{\mathbf{w}} f(\mathbf{w}_t, \mathbf{z}_t)$        // treating $\mathbf{z}_t$ as constant
4      choose step size $\eta_t$        // see Algorithm 5.14
5      optional: $\mathbf{g}_t \leftarrow \mathbf{g}_t / \|\mathbf{g}_t\|$        // normalization
6      $\mathbf{w}_{t+1} = \mathrm{P}_{\mathbb{W}}[\mathbf{w}_t - \eta_t \mathbf{g}_t]$        // subgrad on outer minimization
```

Uzawa's algorithm can be seen as an approximation of the alternating Algorithm 12.18, where instead of finding the exact minimizer in $\mathbf{w}$, we simply perform a gradient descent step. The downside of Uzawa's algorithm is that we still have to solve the inner maximization problem exactly in line 2, which seems quite wasteful: $\mathbf{w}_t$ is going to change in the next iteration anyways so maybe a crude, inexact maximizer in $\mathbf{z}$, such as a gradient ascent step, suffices?

Uzawa, H. (1958). "Iterative methods for concave programming". In: *Studies in linear and non-linear programming*. Ed. by K. J. Arrow, L. Hurwicz, and H. Uzawa. Standford University Press, pp. 154–165.
Danskin, J. M. (1967). "The theory of max-min and its application to weapons allocation problems". Springer.
Dem'yanov, V. F. (1969). "On the minimax problem". *Soviet Mathematics Doklady*, vol. 187, no. 2, pp. 255–258.

Algorithm 12.22: Gradient descent ascent (GDA)

Indeed, we may simply replace the exact inner maximization step in Uzawa's Algorithm 12.21 with a single gradient ascent step. This idea can be traced back to (at least) Brown and Neumann (1950) and Arrow and Hurwicz (1958), who studied the continuous analogue.

```
Algorithm: Gradient descent ascent for minimax
Input: $(\mathbf{w}_0, \mathbf{z}_0) \in \operatorname{dom} f \cap \mathbb{W} \times \mathbb{Z}$
for $t = 0, 1, \ldots$ do
    choose step size $\eta_t > 0$
    $\mathbf{w}_{t+1} = \mathrm{P}_{\mathbb{W}}[\mathbf{w}_t - \eta_t \partial_{\mathbf{w}} f(\mathbf{w}_t, \mathbf{z}_t)]$        // GD on minimization
    $\mathbf{z}_{t+1} = \mathrm{P}_{\mathbb{Z}}[\mathbf{z}_t - \eta_t \partial_{\mathbf{z}} - f(\mathbf{w}_t, \mathbf{z}_t)]$        // GA on maximization
```

Variations of Algorithm 12.22 include (but are not limited to):

- use different step sizes on $\mathbf{w}$ and $\mathbf{z}$;
- use $\mathbf{w}_{t+1}$ in the update on $\mathbf{z}$ (or vice versa);
- use stochastic gradients in both steps (more on this later);
- after every update in $\mathbf{w}$, perform k updates in $\mathbf{z}$ (or vice versa).

Brown, G. W. and J. v. Neumann (1950). "Solutions of Games by Differential Equations". In: *Contributions to the Theory of Games I*. Ed. by H. W. Kuhn and A. W. Tucker. Princeton University Press, pp. 73–79.

Arrow, K. J. and L. Hurwicz (1958). "Gradient method for concave programming I: Local results". In: *Studies in linear and non-linear programming*. Ed. by K. J. Arrow, L. Hurwicz, and H. Uzawa. Standford University Press, pp. 117–126.

History 12.23: Arrow and Hurwicz

Both Arrow and Hurwicz won the Nobel prize in economics.

Example 12.24: Vanilla GDA may never converge for any step size

Let us consider again the simple problem:

$$\min_{w\in[-1,1]} \max_{z\in[-1,1]} wz \quad \equiv \quad \max_{z\in[-1,1]} \min_{w\in[-1,1]} wz,$$

which, as we showed before, has a unique (non-stable) saddle-point at $(w_\star, z^\star) = (0, 0)$.

If we run vanilla (projected) GDA with step size $\eta_t \geq 0$, then

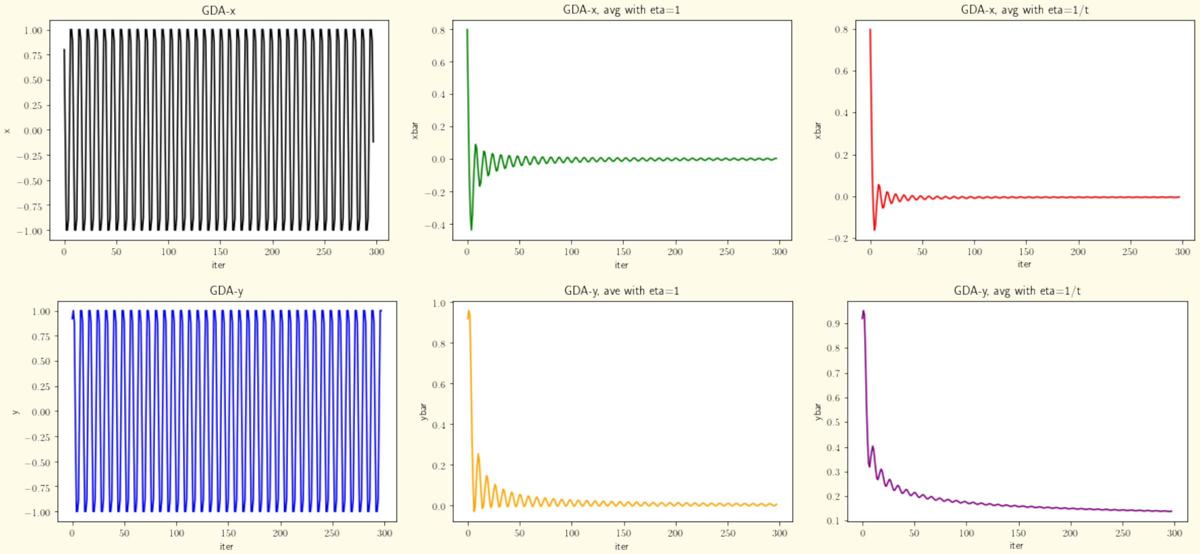
$$\begin{aligned} w_{t+1} &= [w_t - \eta_t z_t]_{-1}^{1} \\ z_{t+1} &= [z_t + \eta_t w_t]_{-1}^{1}, \end{aligned}$$

where $[t]_{-1}^{1} := (t \wedge 1) \vee (-1)$ is the projection of t onto the interval $[-1, 1]$. Thus, we have

$$w_{t+1}^2 + z_{t+1}^2 \geq 1 \wedge [(w_t - \eta_t z_t)^2 + (z_t + \eta_t w_t)^2] = 1 \wedge [(1 + \eta_t^2)(w_t^2 + z_t^2)] \geq 1 \wedge (w_t^2 + z_t^2).$$

Therefore, if we do *not* initialize at the saddle point $(w_\star, z^\star) = (0, 0)$, then the norm of (w_t, z_t) will always be lower bounded by $1 \wedge \|\binom{w_0}{z_0}\| > 0 = \|(w_\star, z^\star)\|$. In other words, (w_t, z_t) will not converge to $(w_\star, z^\star)$.

Indeed, the left plots below verify this result. Interestingly, with averaging (i.e. Line 6-7 of Algorithm 12.22), we recover convergence in both the middle and right plots (with different η).



Remark 12.25: Convergence of GDA

Convergence of GDA, under the assumption of stability of the saddle point set $\mathbb{W}_\star \times \mathbb{Z}^\star$, was first proved by Gol'shtein (1972) and later refined by Maistroskii (1976), which required for instance merely strict convexity in $\mathbf{w}$ (hence one-sided stability).

Gol'shtein, E. G. (1972). "A generalized gradient method for finding saddlepoints". *Ekonomika i matematicheskie metody*, vol. 8, no. 4, pp. 569–579.

Maistroskii, D. (1976). "Gradient methods for finding saddle points". *Ekonomika i matematicheskie metody*, vol. 12, no. 5, pp. 917–929.