A Supplementary Material

B Proofs

Before we prove Theorem 2, let us first show a useful result which justifies switching from Tikhonov to Ivanov regularization and vice versa. Note that this result does not hold in general but relies on the tightness of the bound on the regularizing constraint, i.e. Eq. (2).

Proposition 1 *Let $D \subset \mathbb{R}^d$ be a convex set, be $f, g : D \to \mathbb{R}$ convex functions. Consider the convex optimization tasks*

$$\min \boldsymbol{x} \in D \quad f(\boldsymbol{x}) + \sigma g(\boldsymbol{x}), \tag{1a}$$

$$\min_{\boldsymbol{x} \in D : g(\boldsymbol{x}) \leq \tau} \quad f(\boldsymbol{x}). \tag{1b}$$

Assume that some contraint qualification holds in (1b)*, which gives rise to strong duality, e.g. that Slater's condition is satisfied. Furthermore assume that the constraint cannot be removed without changing the optimal solution, i.e.*

$$\inf_{\boldsymbol{x} \in D} f(\boldsymbol{x}) \ < \inf_{\boldsymbol{x} \in D : g(\boldsymbol{x}) \leq \tau} f(\boldsymbol{x}). \tag{2}$$

Then we have that for each $\sigma > 0$ there exists a $\tau > 0$, and vice versa, such that OP (1a) *is equivalent to OP* (1b)*, i.e. each optimal solution of the one is an optimal solution of the other, and vice versa.*

Proof.
(a). Let be $\sigma > 0$, be $\boldsymbol{x}^*$ optimal in (1a). We have to show that there exists a $\tau > 0$ such that $\boldsymbol{x}^*$ is optimal in (1b). We set $\tau = g(\boldsymbol{x}^*)$. Suppose $\boldsymbol{x}^*$ is not optimal in (1b), i.e. it exists $\tilde{\boldsymbol{x}} \in D : g(\tilde{\boldsymbol{x}}) \leq \tau$ such that $f(\tilde{\boldsymbol{x}}) < f(\boldsymbol{x}^*)$. Then we have

$$f(\tilde{\boldsymbol{x}}) + \sigma g(\tilde{\boldsymbol{x}}) < f(\boldsymbol{x}^*) + \sigma\tau,$$

and by $\tau = g(\boldsymbol{x}^*)$:

$$f(\tilde{\boldsymbol{x}}) + \sigma g(\tilde{\boldsymbol{x}}) < f(\boldsymbol{x}^*) + \sigma g(\boldsymbol{x}^*).$$

This contradics the optimality of $\boldsymbol{x}^*$ in (1a), and hence shows that $\boldsymbol{x}^*$ is optimal in (1b), which was to be shown.
(b). Vice versa, let be $\tau > 0$, be $\boldsymbol{x}^*$ optimal in (1b). The Lagrangian of (1b) is given by

$$\mathcal{L}(\sigma) = f(\boldsymbol{x}) + \sigma \left(g(\boldsymbol{x}) - \tau \right), \quad \sigma \geq 0.$$

By strong duality $\boldsymbol{x}^*$ is optimal in the sattle point problem

$$\sigma^* := \operatorname*{argmax}_{\sigma \geq 0} \ \min_{\boldsymbol{x} \in D} \quad f(\boldsymbol{x}) + \sigma \left(g(\boldsymbol{x}) - \tau \right),$$

and by the strong max-min property (cf. [1], p.238) we may exchange the order of maximization and minimization. Hence $\boldsymbol{x}^*$ is optimal in

$$\min_{\boldsymbol{x} \in D} \quad f(\boldsymbol{x}) + \sigma^* \left(g(\boldsymbol{x}) - \tau \right). \tag{3}$$

Removing the constant term $-\sigma^*\tau$, and setting $\sigma = \sigma^*$, we have that $\boldsymbol{x}^*$ is optimal in (1a), which was to be shown. Moreover by (2) we have that

$$\boldsymbol{x}^* \neq \operatorname*{argmin}_{\boldsymbol{x} \in D} f(\boldsymbol{x}),$$

and hence we see from Eq. (3) that $\sigma^* > 0$, which completes the proof of the proposition. □

We are now ready to prove Theorem 2.

Proof. Let be $(\tilde{C}, \mu) > 0$. By Prop. 1 we have that (4) is equivalent to

$$\min_{\boldsymbol{w}, b, \boldsymbol{\theta}} \quad \tilde{C} \sum_{i=1}^{n} V \left(\sum_{m=1}^{M} \boldsymbol{w}_m^\top \psi_m(\boldsymbol{x}) + b, \ y_i \right) + \frac{1}{2} \sum_{m=1}^{M} \frac{||\boldsymbol{w}_m||_2^2}{\theta_m}$$
$$\text{s.t.} \quad ||\boldsymbol{\theta}||_p^p \leq \tau,$$

for some $\tau > 0$. Consider the optimal solution $(\boldsymbol{w}^\star, b^\star, \boldsymbol{\theta}^\star)$ corresponding to a given parametrization $(\tilde{C}, \tau)$. For any $\lambda > 0$, the bijective transformation $(\tilde{C}, \tau) \mapsto (\lambda^{-1/p} C, \lambda\tau)$ will yield $(\boldsymbol{w}^\star, b^\star, \lambda^{1/p}\boldsymbol{\theta}^\star)$ as optimal solution. Hence it is equivalent: it represents the same classification function. Applying the transformation with $\lambda := 1/\tau$ and renaming the variable $C = \tilde{C}\tau^{\frac{1}{p}}$ yields (5), which was to be shown. □

References

[1] S. Boyd and L. Vandenberghe. *Convex Optimization*. Cambrigde University Press, Cambridge, UK, 2004.