Notes on Invariant Risk Minimization

Karl Stratos

1 Objective

Assume some loss function $l : \mathcal{Y} \times \mathcal{Y} \to [0, \infty)$. Let $e \in \mathcal{E}$ denote an "environment" that has its own distribution $p_{X^e Y^e}$ over input-output pairs and is associated with a risk $R^e(f) = \mathbf{E}\left[l(Y^e, f(X^e)\right]$. Our goal is to learn f that minimizes the out-of-domain risk $R^{\text{OOD}}(f) = \max_{e \in \mathcal{E}} R^e(f)$ given sampling access to $\mathcal{E}_{\text{tr}} \subset \mathcal{E}$. A naive approach is to reduce the problem into empirical risk minimization (ERM) by "pooling" samples from all environments and minimizing $\sum_{e \in \mathcal{E}_{\text{tr}}} R^e(f)$, but this can still lead to large R^{OOD}.

Instead, suppose we can transform the input so that the output is distributed the same across all environments conditioning on the transformation of the input. Let's stick with regression for simplicity and let $\mathcal{Y} = \mathbb{R}^d$ and $l(y, y') = ||y - y'||^2$. If we have a transformation $\Phi : \mathcal{X} \to \mathcal{H}$ such that for any $x \in \mathcal{X}$,

$$\mathbf{E}[Y^e | \Phi(X^e) = \Phi(x)] = \mathbf{E}[Y^{e'} | \Phi(X^{e'}) = \Phi(x)] \qquad \forall e, e' \in \mathcal{E} \tag{1}$$

then any predictor $g : \mathcal{H} \to \mathcal{Y}$ that is optimal for one environment is optimal for all environments under Φ. Invariant risk minimization (IRM) [1] aims to learn such a transformation by the following nested objective

$$\Phi^*, g^* = \underset{\substack{\Phi: \mathcal{X} \to \mathcal{H} \\ g \in \arg\min_{g': \mathcal{H} \to \mathcal{Y}} R^e(g' \circ \Phi)\ \forall e \in \mathcal{E}_{\text{tr}}}}{\arg\min} \sum_{e \in \mathcal{E}_{\text{tr}}} R^e(g \circ \Phi) \tag{2}$$

where for every considered representation we only consider a predictor g that is *simultaneously optimal* for all environments. The point is that if Φ induces a simultaneously optimal predictor g, then it must satisfy the stability property (1) across $\mathcal{E}_{\text{tr}}$ because $g(x) = \mathbf{E}[Y^e | \Phi(X^e) = \Phi(x)]$ for every $e \in \mathcal{E}_{\text{tr}}$ (see Lemma B.2). Under certain diversity conditions on $\mathcal{E}_{\text{tr}}$, invariance across $\mathcal{E}_{\text{tr}}$ implies invariance across $\mathcal{E}$.

Optimization. A soft version of (2) is to minimize $\sum_{e \in \mathcal{E}_{\text{tr}}} R^e(g \circ \Phi) + \lambda D(g, \Phi, e)$ over Φ and g where λ is a hyperparamater and $D(g, \Phi, e)$ measures the suboptimality of g for minimizing the risk R^e under Φ. One way to measure the suboptimality is the magnitude of the gradient of $R^e(g \circ \Phi)$ wrt g since it should be zero at optimum. Finally, noting that in the overparameterized case we can fix g to be an arbitrary simple function, say element-wise multiplication (denoted $\cdot$) by the scalar $w = 1$, and focus on learning just Φ, we have the practical objective (IRMv1):

$$\Phi^* = \underset{\Phi: \mathcal{X} \to \mathbb{R}^d}{\arg\min} \sum_{e \in \mathcal{E}_{\text{tr}}} R^e(\Phi) + \lambda \left(\nabla_w R^e(w \cdot \Phi)\big|_{w=1}\right)^2 \tag{3}$$

A more verbose justification is provided in Appendix A.

Empirical objective. Let $(x_1^e, y_1^e) \dots (x_{N^e}^e, y_{N^e}^e) \sim p_{X^e Y^e}$ denote N^e iid samples from each training environment $e \in \mathcal{E}_{\text{tr}}$. Let (J^e, K^e) denote a balanced partition of the indices $\{1 \dots N^e\}$. An empirical estimate of IRMv1 is given by

$$\widehat{J}(\Phi) = \sum_{e \in \mathcal{E}_{\text{tr}}} \left(\frac{1}{N^e} \sum_{i=1}^{N^e} l(y_i^e, \Phi(x_i^e)) + \lambda \left(\frac{1}{|J^e|} \sum_{j \in J^e} \nabla_w l(y_j^e, w \cdot \Phi(x_j^e)) \big|_{w=1} \right) \times \left(\frac{1}{|K^e|} \sum_{k \in K^e} \nabla_w l(y_k^e, w \cdot \Phi(x_k^e)) \big|_{w=1} \right) \right)$$

and we can update $\Phi \leftarrow \textbf{step}(\Phi, \nabla_\Phi \widehat{J}(\Phi), \eta)$. Modern autodiff software like PyTorch can build a computation graph that calculate the gradient of $l(y, w \cdot \Phi(x))$ with respect to w evaluated at $w = 1$ (a scalar).

Relation to other objectives. IRM seems to have an upper hand over related objectives in generalization capabilities. In the example in Section A.1, any form of vanilla **ERM** will result in infinite R^{OOD} (because of nonzero w_2^+) whereas IRM can in principle achieve finite R^{OOD}. More generally, ERM breaks when the assumption that all samples are generated from the *same* distribution breaks. In adversarial domain adaptation (**ADA**), one learns Φ such that the distribution of $\Phi(X^e)$ is the same across $e \in \mathcal{E}$. But this is quite different from the *conditional* stability in (1) and suffers poor generalization when $p_{Y^e|X^e}$ is different across environments. In a conditional adaptation of ADA (**C-ADA**), one learns Φ such that the conditional distribution of $\Phi(X^e)$ given Y^e is the same across $e \in \mathcal{E}$. This is the same objective as IRM only if the distribution of Y^e is the same across environments; the paper gives an example where C-ADA fails.

References

[1] Arjovsky, M., Bottou, L., Gulrajani, I., and Lopez-Paz, D. (2019). Invariant risk minimization. *arXiv preprint arXiv:1907.02893*.

A More Verbose Justification of IRMv1

The paper [1] uses specific linear settings to motivate (i) why using the squared norm of the gradient is a good measure of suboptimality and (ii) why it's okay to fix g to be an arbitrary linear function.

A.1 Definition of Suboptimality

Suppose each environment $e \in \mathcal{E}$ is associated with variance $\sigma_e^2 \in (0, \sigma_{\max}^2]$ and defines $(X^e, Y^e) \in \mathbb{R}^2 \times \mathbb{R}$ by

$$\begin{aligned} x_1^e &\sim \mathcal{N}(0, \sigma_e^2) & y^e &= x_1^e + \epsilon_1^e \\ \epsilon_1^e &\sim \mathcal{N}(0, \sigma_e^2) & x_2^e &= y^e + \epsilon_2^e \\ \epsilon_2^e &\sim \mathcal{N}(0, 1) \end{aligned}$$

This is really a one-dimensional problem since Y^e is "caused" by X_1^e but not by X_2^e. If we know that fact, we would regress *only* using X_1^e for some training environment e and obtain $w_1^* = \arg\min_{w_1} \mathbf{E}[(Y - w_1 X_1^e)^2] = 1$. This optimal predictor $f(x^e) = x_1^e$ achieves $R^{\text{OOD}}(f) = \max_e \sigma_e^2$. In general, we don't know this fact and regress using both input variables: $(w_1^+, w_2^+) = \arg\min_{w_1, w_2} \mathbf{E}[(Y - w_1 X_1^e - w_2 X_2^e)^2] = \left(\frac{1}{1+\sigma_e^2}, \frac{\sigma_e^2}{1+\sigma_e^2}\right)$. This "wrong" optimal predictor suffers infinite R^{OOD}. Now suppose $\Phi(x) = x_1 + c x_2$ which defines the optimal regressor $w_\Phi^e = \mathbf{E}\left[\Phi(X^e)\Phi(X^e)^\top\right]^{-1} \mathbf{E}\left[\Phi(X^e) Y^e\right] = \left(\frac{1}{1+\sigma_e^2}, \frac{\sigma_e^2}{c(1+\sigma_e^2)}\right)$. We might consider the squared error $||w_\Phi^e - w||^2$ to be a natural metric of suboptimality, but for $w = (1, 0)$ the error is equal to $\left(\frac{\sigma_e^2}{1+\sigma_e^2}\right)^2 \left(1 + \frac{1}{c^2}\right)$ which goes to infinity as $c \to 0$. The proposed solution is to undo the inversion in the squared error and define

$$D(w, \Phi, e) = \left|\left|\mathbf{E}\left[\Phi(X^e) Y^e\right] - \mathbf{E}\left[\Phi(X^e)\Phi(X^e)^\top\right] w\right|\right|^2$$

Then $D((1,0), \Phi, e) = c^2 \sigma_e^4$ which goes to zero as $c \to 0$ as we wanted. We "lift" this definition by noting that $D(w, \Phi, e) = ||\nabla_w R^e(w \circ \Phi)||^2$.

A.2 Fixing the Predictor

Let $(X^e, Y^e) \in \mathbb{R}^d \times \mathbb{R}$ and $\Phi \in \mathbb{R}^{p \times d}, w \in \mathbb{R}^p$. Note that the final model can be written as a d-dimensional vector $v = \Phi^\top w$. Assume that $R^e : \mathbb{R}^d \to [0, \infty)$ is convex. The following lemma is easy to verify.

Lemma A.1. *A sufficient and necessary condition for $v \in \mathbb{R}^d$ to be a feasible point in IRM (2) (i.e., $v = \Phi^\top w$ for some $\Phi \in \mathbb{R}^{p \times d}, w \in \mathbb{R}^p$ where $w \in \arg\min_{w'} R^e(w' \circ \Phi)$ for all $e \in \mathcal{E}_{\text{tr}}$) is $v^\top \nabla_v R^e(v) = 0$ for all $e \in \mathcal{E}_{\text{tr}}$.*

The proof of the lemma implies the following corollary.

Corollary A.2. *Pick any feasible point $v \in \mathbb{R}^d$ in IRM (2). It admits a rank-p decomposition $v = \Phi^\top w$ for some $\Phi \in \mathbb{R}^{p \times d}, w \in \mathbb{R}^p$ where $p = \text{nullity}\left([\nabla_v R^e(v)]_{e \in \mathcal{E}_{\text{tr}}}\right) \geq 1$.*

If $p = 1$ in the corollary, then we can restrict our search to $\Phi \in \mathbb{R}^d$ by fixing $w = 1$. We maintain this argument even when $p > 1$. We also generalize the argument to a multi-dimensional output $Y^e \in \mathbb{R}^{d'}$ by fixing $w = 1$ to be an element-wise multiplication.

B Risk Minimization

We assume an unknown joint distribution p_{XY} over input-output pairs $(x, y) \in \mathcal{X} \times \mathcal{Y}$ and learn a mapping $f : \mathcal{X} \to \mathcal{Y}$ by minimizing the expected value of a loss function $l : \mathcal{Y} \times \mathcal{Y} \to [0, \infty)$. This objective is called the "risk" associated with f and denoted by

$$R(f) := \mathbf{E}\left[l(Y, f(X))\right]$$

The minimal possible risk is called the **Bayes risk**:

$$R^* := \min_{f:\mathcal{X}\to\mathcal{Y}} R(f)$$

Any mapping f^* that achieves the Bayes risk (i.e., $R(f^*) = R^*$) is called **Bayes optimal** or simply optimal.

Lemma B.1 (Classification)**.** *Assume $\mathcal{Y}$ is discrete. Choose the zero-one loss $l(y, y') = [[y \neq y']]$. Then $f : \mathcal{X} \to \mathcal{Y}$ is an optimal classifier iff for all $x \in \mathcal{X}$*

$$f(x) \in \arg\max_{y\in\mathcal{Y}} \; p_{Y|X}(y|x)$$

Proof. For all $x \in \mathcal{X}$,

$$\begin{aligned}
\Pr\left(Y = f(X)|X = x\right) &= \sum_{a\in\mathcal{Y}} \Pr\left(Y = a|X = x, f(X) = a\right) \Pr\left(f(X) = a|X = x\right) \\
&= \sum_{a\in\mathcal{Y}} p_{Y|X}(a|x)\left[[f(x) = a]\right] \\
&= p_{Y|X}(f(x)|x)
\end{aligned}$$

and thus

$$\Pr\left(Y = f(X)\right) = \sum_{x\in\mathcal{X}} p_X(x) p_{Y|X}(f(x)|x)$$

This is maximized (and thus $R(f)$ is minimized) iff $f(x) \in \arg\max_{y\in\mathcal{Y}} \; p_{Y|X}(y|x)$ for all $x \in \mathcal{X}$. □

Lemma B.2 (Regression)**.** *Assume $\mathcal{Y} = \mathbb{R}^d$. Choose the squared loss $l(y, y') = ||y - y'||^2$. Then $f : \mathcal{X} \to \mathbb{R}^d$ is an optimal regressor iff for all $x \in \mathcal{X}$*

$$f(x) = \mathbf{E}\left[Y|X = x\right]$$

Proof. Let $f : \mathcal{X} \to \mathbb{R}^d$ be any mapping. For any $x \in \mathcal{X}$,

$$\left(Y - \mathbf{E}\left[Y|X = x\right]\right)^\top \left(f(X) - \mathbf{E}\left[Y|X = x\right]\right)$$

is zero in expectation with respect to $p_{Y|X=x}$. Thus we can decompose the squared loss by the Pythagorean theorem:

$$R(f) = \mathbf{E}\left[\left|\left|Y - \mathbf{E}\left[Y|X\right]\right|\right|^2\right] + \mathbf{E}\left[\left|\left|f(X) - \mathbf{E}\left[Y|X\right]\right|\right|^2\right]$$

The first term is independent of f (i.e., intrinsic risk). The second term is minimized to zero iff $f(x) = \mathbf{E}\left[Y|X = x\right]$ for all $x \in \mathcal{X}$. □