

在线倾斜经验风险最小化的好处： 异常检测和鲁棒回归的案例研究

Yiğit E. Yıldırım, Samet Demir, Zafer Doğan

MLIP Research Group, KUIS AI Center & Department of EEE, Koç University
İstanbul, Turkey

ABSTRACT

经验风险最小化 (ERM) 是监督学习的基础框架，但主要优化平均情况下的性能，经常忽视公平性和鲁棒性的考虑。倾斜经验风险最小化 (TERM) 通过引入一个指数倾斜超参数 t 来平衡平均情况下的准确性与最坏情况下的公平性和鲁棒性。然而，在在线或流数据设置中，数据一次一个样本地到达，经典 TERM 目标退化为标准 ERM，失去倾斜敏感度。我们通过提出一种移除经典目标中的对数的在线 TERM 形式来解决这一局限，保留倾斜效果而无需额外的计算或内存开销。该形式允许由 t 控制的连续权衡，平滑地在 ERM ($t \rightarrow 0$)、公平性强调 ($t > 0$) 和对异常值的鲁棒性 ($t < 0$) 之间插值。我们在两个代表性的流任务上进行了在线 TERM 的经验验证：对抗性异常值下的稳健线性回归和二元分类中少数类检测。我们的结果显示，负倾斜有效抑制了异常值的影响，而正倾斜则在精度影响最小的情况下提高了召回率，所有这些都以每个样本的计算成本相当于 ERM 的方式实现。因此，在高效单一样本学习环境中，在线 TERM 恢复了经典 TERM 的完整鲁棒性-公平性光谱。

Index Terms— 倾斜经验风险最小化，在线学习，异常值检测，鲁棒性，公平性。

We acknowledge that this work is supported partially by TÜBİTAK under project 124E063 in ARDEB 1001 program. Y.E.Y. is supported by the TÜBİTAK project. S.D. is supported by an AI Fellowship provided by KUIS AI Research Center and a PhD Scholarship (BİDEB 2211) from TÜBİTAK. The corresponding author is Zafer Doğan (zdogan@ku.edu.tr).

1. 介绍

经验风险最小化 (ERM) 是大多数现代监督学习算法的基础原理。[1]。ERM 目标通过最小化 N 个样本的平均损失来选择模型参数 θ ：

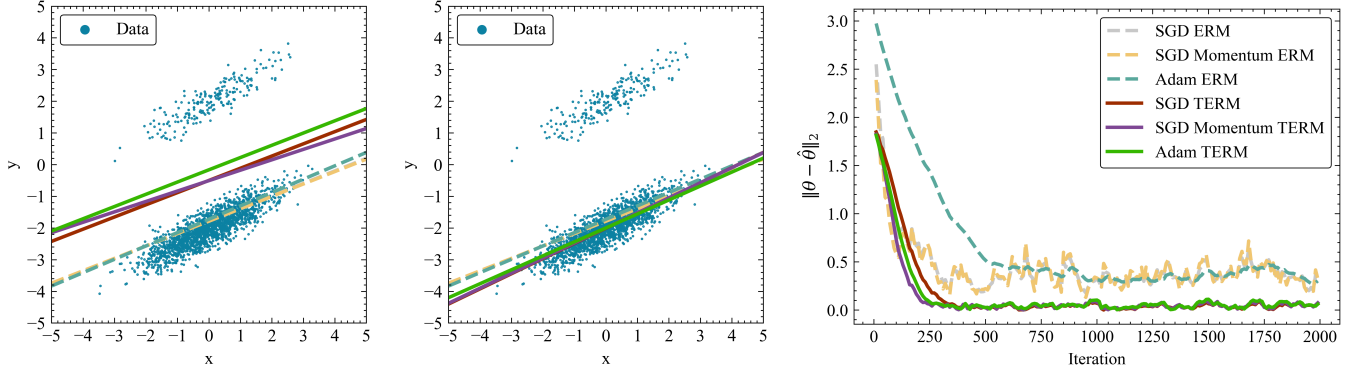
$$R(\theta) := \frac{1}{N} \sum_{i=1}^N \ell_i(\theta), \quad (1)$$

其中， ℓ_i 表示第 i 个样本的损失。

ERM 享有优雅的统计保证，并且在训练数据忠实代表测试分布时提供了强大的预测性能。然而，由于设计原因，ERM 侧重于均值性能。具体来说，每个示例都平等地贡献给目标函数，无论其是否具有代表性、非典型或甚至错误标记。这种均匀加权使得 ERM 在存在异常值、噪声数据 [2, 3] 或类别不平衡 [4, 5] 的情况下变得脆弱，这些都是实际学习场景中常见的问题。大量的研究工作已经解决了标准 ERM [6, 7, 8] 的这些缺点。研究人员提出了公平感知的目标函数以确保子组之间的平衡训练 [6, 7]，并且改进了稳健性公式以提高测试性能 [8]。在寻找 ERM 替代方案的过程中，已在不同设置中探索了各种倾斜技术，包括重要性采样 [9]、序列决策制定 [10, 11] 和大偏差理论 [12]。倾斜经验风险最小化 (TERM) 通过一个倾斜超参数 [13, 14, 15] 扩展了 ERM 框架。TERM 不是最小化平均损失，而是最小化一个倾斜的损失：

$$\bar{R}(t; \theta) := \frac{1}{t} \log \left(\frac{1}{N} \sum_{i=1}^N e^{t\ell_i(\theta)} \right), \quad (2)$$

其中 $t \in \mathbb{R}$ 是倾斜超参数。当 $t \rightarrow 0$ 时，目标函数简化为普通的 ERM。对于 $t > 0$ ，指数加权放大了较大



(a) 在线 TERM 与 $t = +0.5$: 拟合线强调大损失样本。 (b) 在线 TERM 与 $t = -0.5$: 拟合线减小离群值的影响。 (c) 预测误差与 $t < 0$: 到真实内点权重的欧氏距离，每 10 次迭代记录一次。

Fig. 1: 我们通过一个带有异常值的噪声线性回归示例来突出说明普通 ERM 和在线 t -倾斜目标之间的差异。(a) 和 (b) 在合成样本流上比较了这两种方法，其中 90% 个点遵循内点线 $y = 0.52x - 2$ ，其余 10% 是向上平移四单位的平行异常值， $y = 0.52x + 2$ 。每个观测值都受到独立同分布的高斯噪声影响，标准差为 0.3。使用平方损失，并以每次一个样本更新模型——不进行批量处理——使用各种优化器。学习率设置为 ERM 的是 10^{-2} ，TERM 使用 $t < 0$ 是 10^{-2} ，TERM 使用 $t > 0$ 是 10^{-4} 。SGD Momentum ERM/TERM 使用动量参数 0.3，Adam ERM/TERM 的超参数设置为 $\beta_1 = 0.9, \beta_2 = 0.999, \epsilon = 10^{-8}, \bar{\epsilon} = 0$ 。这些速率的选择是为了平衡稳定性并揭示长期行为。(c) 跟踪当前参数向量与真实内点参数 $\theta^* = (0.52, -2)$ 之间的欧几里得距离 $\|\theta - \theta^*\|_2$ ——每十个更新计算一次。

的个别损失，平滑地插值到最大损失目标 [16, 17]，从而促进公平性。对于 $t < 0$ ，加权抑制了大的损失，使得目标对损坏或对抗样本 [18] 具有鲁棒性。因此，倾斜超参数 t 提供了一个在公平性和鲁棒性之间连续的权衡范围。

TERM 通常是批学习制定的，在参数更新之前会聚合一组 $N > 1$ 样本的损失。然而，在在线设置中——即学习者每次迭代处理一个样本（即 $N = 1$ ）——TERM 目标函数 (2) 退化为普通的 ERM，倾斜超参数 t 实际上消失。这导致了 TERM 在公平性和鲁棒性方面的优势丧失。

为了解决这一限制，我们通过引入一个简单的修改来提出一种在线 TERM 公式，在线环境中保留倾斜超参数。然后，我们系统地展示了这种新公式的实际好处。具体来说，我们在两个示例问题设置上进行研究：鲁棒性线性回归和二元分类中的异常值检测。在这两种情况下，数据生成都包括受控的异常样本。通过在正负范围内变化倾斜超参数 t ，我们量化了在线 TERM 如何权衡稳健性和公平性。这些控制实验表

明，单个超参数 t 可以平滑地插值平均损失、公平性和鲁棒性。因此，我们实证显示，在线 TERM 是处理涉及公平性和鲁棒性问题的实际工具。

2. 在线倾斜经验风险最小化

在本节中，我们推导了一个保持倾斜超参数的 TERM 在线版本。为此，我们首先考察倾斜目标函数 (2) 的梯度：

$$\nabla_{\theta} \bar{R}(t; \theta) = \sum_i \frac{e^{t\ell_i(\theta)}}{\sum_j e^{t\ell_j(\theta)}} \nabla_{\theta} \ell_i(\theta), \quad (3)$$

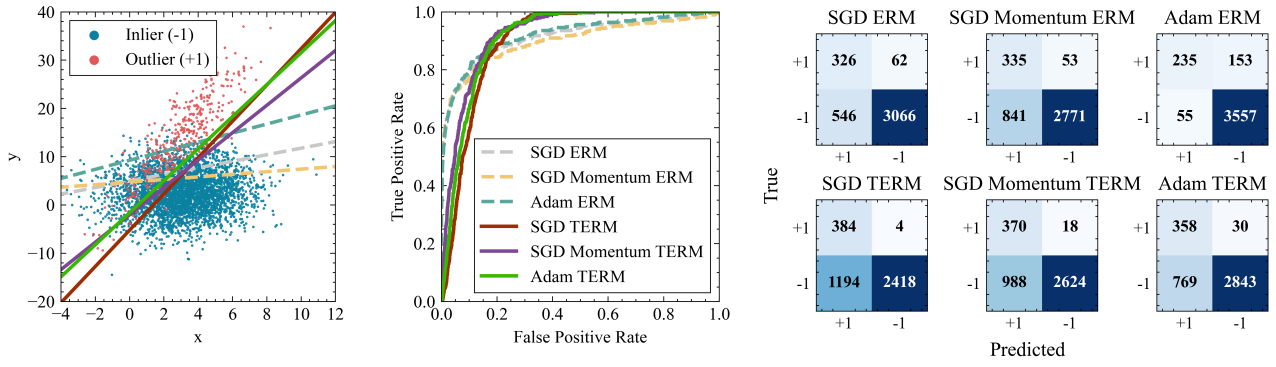
其中每个样本对梯度的影响与其归一化权重 $e^{t\ell_i(\theta)} / \sum_j e^{t\ell_j(\theta)}$ 成正比。

为了在在线设置中保留这种加权机制，其中 $N = 1$ ，我们从 (2) 中去掉对数，并提出以下替代倾斜目标：

$$\tilde{R}(t; \theta) := \frac{1}{t} e^{t\ell_i(\theta)}. \quad (4)$$

该目标产生梯度

$$\nabla_{\theta} \tilde{R}(t; \theta) = e^{t\ell_i(\theta)} \nabla_{\theta} \ell_i(\theta), \quad (5)$$



(a) 使用在线 TERM 方法并通过倾斜度 $t = 0.2$ 获得的决策边界。 (b) 使用在线 TERM 获得的 ROC 曲线采用了倾斜 $t = 0.2$ 。 (c) 使用在线 TERM 获得的混淆矩阵，倾斜度为 $t = 0.2$ 。

Fig. 2: 我们通过一个包含内点和外点的玩具二分类任务来研究正倾斜下的在线 TERM 的行为。在每次迭代中，学习器观察一个独立从两个线性模型的混合中抽取的单一样本 (x, y) ，并添加了高斯噪声。以概率 0.9，内点样本是根据 $y = 0.2x + 2 + \varepsilon$, $\varepsilon \sim \mathcal{N}(0, \sigma^2)$ ，抽取的，并被赋予标签 -1 。以概率 0.1，外点样本是根据 $y = 4x + 2 + \varepsilon$, $\varepsilon \sim \mathcal{N}(0, \sigma^2)$ ，抽取的，并被赋予标签 $+1$ 。噪声标准差设置为 $\sigma = 4.0$ ，以模拟高方差扰动。学习器保持一个线性评分函数的形式 $s_{\theta}(x, y) = \theta^{\top}(x, y, 1)$ ，其中 $\theta \in \mathbb{R}^3$ ，并通过 4000 次迭代使用普通 SGD、带有动量的 SGD 或 Adam 优化器以及平方损失来更新参数。SGD ERM、SGD Momentum ERM 和 Adam ERM 变体分别使用学习率 10^{-3} 、 10^{-3} 和 2×10^{-3} 。相应的 TERM 变体在降低的学习率下进行训练——SGD TERM 和 SGD Momentum TERM 为 2×10^{-4} ，Adam TERM 为 8×10^{-4} ——以抵消由正倾斜引起的梯度爆炸风险。SGD Momentum ERM/TERM 使用动量值 0.3，而 Adam ERM/TERM 的超参数设置为 $\beta_1 = 0.9, \beta_2 = 0.999, \epsilon = 10^{-8}, \bar{\epsilon} = 0$ 。

其中指数因子 $e^{t\ell_i(\theta)}$ 在 $t > 0$ 时放大了大损失样本, 在 $t < 0$ 时则减弱它们。这种指数加权重新赋予了控制 TERM 内在公平性和鲁棒性折衷的能力。

然而, 与批量 TERM 不同的是, 当 t 为正时, (5) 中的指数因子可能会爆炸, 因此需要仔细选择学习率。在本文的其余部分中, 我们将展示在线 TERM 公式 (4) 在鲁棒回归和异常值检测任务中的优势。

2.1. 稳健的线性回归与在线 TERM 名词

在本小节中, 我们展示了使用在线 TERM 的鲁棒线性回归的仿真结果。具体来说, 我们考虑一个线性回归问题, 在这个问题中 90% 的数据是正常值, 剩下的 10% 是异常值。在这种情况下, ERM 无法保持对异常值的鲁棒性, 而当倾斜超参数 t 选择得当时, TERM 可以有效地处理它们。

图 1 展示了我们在这鲁棒回归问题上的结果。在图 1a 中, 倾斜超参数设置为 $t = 0.5$ 。尽管它们的更新规则有所不同, 各种优化器——SGD、带有动量的 SGD 和 Adam——表现出质地上相似的行为: 它们产生的回归线被内点和离群点样本之间所牵引。这是因为指数权重 $e^{t\ell_i(\theta)}$ 对于大损失迅速增长, 导致回归线受到离群点的严重影响。

相反, 图 1b 显示了 $t = -0.5$ 的结果。在这里, 负倾斜抑制了大的损失, 导致 TERM 优化器将异常值的权重分配为可忽略不计, 从而保持鲁棒性。拟合线在不同优化器之间的相似性突出了两个关键观察结果: (i) 负倾斜恢复了原始批次 TERM 目标通常提供的鲁棒性, 以及 (ii) 这种鲁棒性在不同的优化器选择中是一致的。这些发现表明, 在存在异常值的情况下, 在线 TERM 与 $t < 0$ 对稳健线性回归是有益的。

与基于 ERM 的优化器相比, 具有负倾斜 ($t < 0$) 的 TERM 优化器能够更好地拟合真实值, 这一点在图 1b 和图 1c 中显而易见。特别是, 图 1c 显示尽管所有优化器都从相同的随机初始化开始, 但在处理相同的数据流时, 它们的参数轨迹会随着时间的推移逐渐发散。具有负倾斜的 TERM 优化器的权重误差在最初的几百次迭代内更稳定地收敛到零, 然后趋于稳定。这种改进的一致性来自于因子 $e^{t\ell_i(\theta)}$, 它显著减少了梯度大小, 从而限制了振荡和突然更新。相比之下, ERM 优化器

($t = 0$) 表现出更多的震荡行为和较差的收敛性。

重要的是, 一旦 TERM 优化器以负倾斜接近真实参数, 它们在整个训练期间都会保持稳定, 没有振荡的迹象。这些定量证据补充了拟合回归线的可视化解释, 并支持结论: 适度的负倾斜可靠地降低异常值的重要性, 加速收敛并维持参数稳定性, 无论使用哪种优化器。

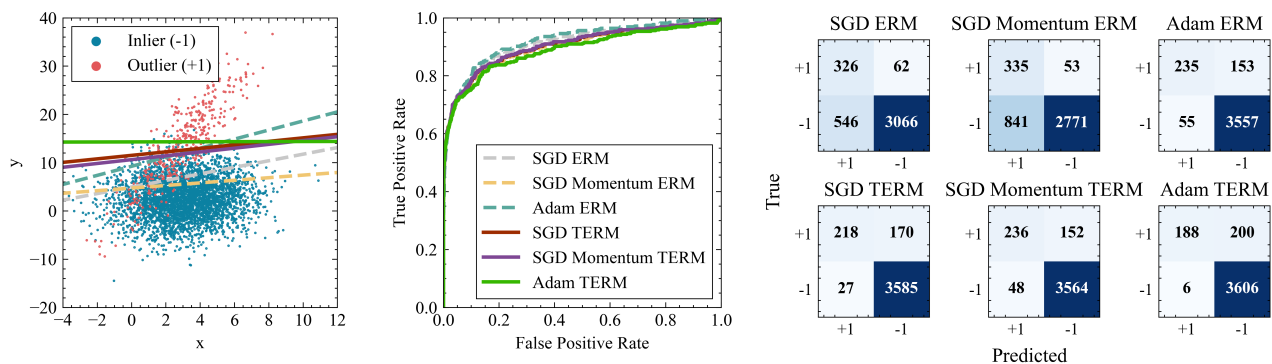
总结而言, 在线 t 倾斜更新具有负倾角提供了一个轻量级、无需调参的鲁棒估计器替代方案。它保留了二次损失函数的简洁性, 并且按顺序处理数据点。如图 1c 所示, 其中权重误差曲线减小, 它迅速收敛到真实的内群线, 而 ERM 变体则受到离群值的不利影响。

2.2. 异常值检测在二元分类中的在线 TERM 检测

我们接下来将在线 t 倾斜更新应用于一个流式二元分类任务, 旨在检测离群点的同时最小化这些罕见点的误分类。

在此设定下, 大损失的示例正好对应我们要识别的异常值。设置 $t = 0.2$ 加剧了它们的影响 (参见图 2), 因为对于异常值而言, 指数权重可能比典型的内点高出几个数量级。因此, 图 2a 中的决策边界适应以正确分类异常值。这种效果在图 2c 所示的混淆矩阵中很明显, 该矩阵显示了真正例 (正确识别的异常值) 增加和假阴性 (未识别的异常值) 减少的情况, 尽管代价是误报 (错误分类的内点) 增多。图 2b 进一步支持这一点, TERM 优化器的 ROC 曲线高于 ERM 基准, 表明超过某些阈值后召回率有所提高。当可以容忍假正例时, TERM 优化器因此能有效地识别几乎所有异常值, 最小化假阴性。

相反, 设置 $t = -0.2$ (见图 3) 在平方误差较大时减小梯度的大小, 实际上指示优化器降低异常值的影响。如图 3a 所示, 这产生了一个几乎水平的决策边界, 与内点分布紧密对齐。图 3c 中的混淆矩阵确认了假阳性显著减少和真阴性增加, 尽管代价是大约遗漏了一半的异常值。相应地, 图 3b 中的 ROC 曲线表明具有负倾斜的 TERM 优化器在低假正率区域优于其 ERM 对应物, 在保持低假阳数的同时实现更高的真正率。这使得它们更可取, 当假阳性成本很高时, 强调了对鲁棒性的偏好而非敏感性。



(a) 使用在线 TERM 方法并倾斜 $t = -0.2$ 获得的决策边界。 (b) 使用在线 TERM 获得的 ROC 曲线采用倾斜 $t = -0.2$ 。 (c) 使用在线 TERM 获得的混淆矩阵，倾斜 $t = -0.2$ 。

Fig. 3: 我们用带有离群值的在线 TERM 在负倾斜下的玩具二分类示例进行说明 ($t < 0$)。设置与图 2 相同，除了这里解释的细节。在每次迭代中，学习器接收一个从先前描述的被污染线性模型中抽取的单个 (x, y) 对。SGD TERM、SGD Momentum TERM 和 Adam TERM 优化器分别使用学习率 2×10^{-3} 、 2×10^{-3} 和 8×10^{-3} 。这些增加的学习率补偿了由负倾斜超参数引起的梯度幅度减小，该超参数强烈降低了大残差的影响。这种调整确保了有效的参数更新，尽管离群值样本的影响被抑制。

这些发现与第 2.1 节中的回归实验观察结果一致。倾斜参数 t 的符号作为离群值检测（公平性）和离群值抵抗（鲁棒性）之间的可调权衡。正向倾斜放大了离群值的损失，因此在稀有点对任务至关重要的情况下表现出色。负向倾斜则降低了这些损失的重要性，从而产生一个更保守的分类器，能更好地抵御极端噪声。在所有三种 TERM 优化器中观察到的一致定性行为表明，这种效果源自倾斜加权机制本身，而不是任何特定的优化算法。在线场景中数据依次到达时，这一简单的修改恢复了 TERM 的灵活性，使从业者能够通过一个超参数平滑地导航鲁棒性和敏感度之间的光谱。

3. 讨论

在实践中，倾斜符号的选择应该反映高损失样本在其学习目标中的作用，这解释了为什么两个玩具问题中最佳方向会有所不同。对于在线鲁棒线性回归，目标是恢复主导内群分布的真实斜率和截距，同时将偶尔出现的极端点视为干扰。在这种情况下，大的残差是干扰信号；远离当前拟合的点很可能是离群值或噪声，因此它们的梯度影响应该减小。负倾斜通过指

数级地降低大损失的权重来实现这一点，有效地剪辑了梯度并使估计器能够专注于大多数内群样本。实证上，这引导参数向量稳定地朝向真实系数，并在收敛后保持稳定性。

相反，在面向异常值检测的流式二元分类任务中，罕见点构成了正类，并且是主要的目标识别对象。大的损失对应于误分类的异常值，需要对决策边界进行激进的更新以在未来迭代中更好地捕捉这些点。正倾斜放大了这些损失，显著增加了梯度贡献并将分离超平面拉向异常值。相比之下，负倾斜抑制了这一关键信息，导致一个保守的分类器很少标记异常值。因此，在希望对外部异常值具有鲁棒性（如回归）的情况下，负倾斜是可取的；而在需要提高敏感性和检测异常值（如分类）时，正倾斜更为适用。这说明相同的倾斜加权机制可以根据任务的不同服务于不同的统计目标。

展望未来，进一步研究的一个关键方向是开发基于任务特征或数据属性自动选择适当倾斜参数的原则性方法。由于倾斜的最佳符号和幅度在很大程度上取决于高损失示例在学习目标中的作用——即强调还是抑制它们——自动化调整策略将大大增强在线 TERM 的实用性能。这些方法可能会利用自适应方案、元学习或数据驱动的启发式方法，在训练过程中动态调整

倾斜, 确保稳健且敏感的性能而无需人工干预。

总而言之, online TERM 中的倾斜参数 t 所提供的灵活性为在鲁棒性和敏感性之间的权衡提供了强大的工具, 但要实现其全部潜力, 则需要针对不同的学习场景进一步提高自动化超参数选择的能力。

4. 结论

我们介绍了一种在线版本的倾斜经验风险最小化 (TERM) 框架, 在严格的流式处理设置中仅一次处理一个样本时, 该框架保持了鲁棒性和公平性之间的可调折衷。我们的方法利用指数加权方案来维持倾斜敏感性而不增加额外的记忆或计算负担, 克服了批量 TERM 在 $N = 1$ 处的局限性。

通过在含有异常值的合成流回归和二元分类任务上的广泛实验, 我们展示了倾斜超参数 t 的符号和大小能够平滑地插值标准 ERM 行为 ($t \rightarrow 0$)、对异常值的鲁棒性 ($t < 0$) 以及公平敏感学习 ($t > 0$)。值得注意的是, 负倾斜显著提高了稳健回归中的收敛稳定性和准确性, 而正倾斜则增强了分类中异常检测的召回率。

这些发现突显了在线 TERM 作为一种实用且理论基础扎实的优化方法, 能够无缝适应多样化的鲁棒性和公平性要求, 并且只需最小程度的调优。未来的工作将研究自适应方案, 用于自动选择针对特定任务和数据分布定制的倾斜参数, 进一步扩大 TERM 在实际流应用程序中的适用性。

5. REFERENCES

- [1] Ian Goodfellow, Yoshua Bengio, and Aaron Courville, Deep Learning, MIT Press, 2016.
- [2] Lu Jiang, Zhengyuan Zhou, Thomas Leung, Li-Jia Li, and Li Fei-Fei, “MentorNet: Learning data-driven curriculum for very deep neural networks on corrupted labels,” in International Conference on Machine Learning, 2018.
- [3] Ashish Khetan, Zachary C Lipton, and Animashree Anandkumar, “Learning from noisy singly-labeled data,” in International Conference on Learning Representations, 2018.
- [4] Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollar, “Focal loss for dense object detection,” in IEEE International Conference on Computer Vision, 2017.
- [5] Tomasz Malisiewicz, Abhinav Gupta, and Alexei A. Efros, “Ensemble of exemplar-svms for object detection and beyond,” in IEEE International Conference on Computer Vision, 2011.
- [6] Tatsunori Hashimoto, Megha Srivastava, Hongseok Namkoong, and Percy Liang, “Fairness without demographics in repeated loss minimization,” in International Conference on Machine Learning, 2018.
- [7] Samira Samadi, Uthaipon Tantipongpipat, Jamie H Morgenstern, Mohit Singh, and Santosh Vempala, “The price of fair pca: One extra dimension,” in Advances in Neural Information Processing Systems, 2018.
- [8] John Duchi and Hongseok Namkoong, “Variance-based regularization with convex objectives,” Journal of Machine Learning Research, vol. 20, no. 68, pp. 1–55, 2019.
- [9] M.J. Wainwright, T.S. Jaakkola, and A.S. Willsky, “A new class of upper bounds on the log partition function,” IEEE Transactions on Information Theory, vol. 51, no. 7, pp. 2313–2335, 2005.
- [10] Ronald A. Howard and James E. Matheson, “Risk-sensitive markov decision processes,” Management Science, vol. 18, no. 7, pp. 356 – 369, Mar. 1972.
- [11] David Nass, Boris Belousov, and Jan Peters, “Entropic risk measure in policy search,” in

IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), 2019.

- [12] Ahmad Beirami, Robert Calderbank, Mark M. Christiansen, Ken R. Duffy, and Muriel Médard, “A characterization of guesswork on swiftly tilting curves,” *IEEE Transactions on Information Theory*, vol. 65, no. 5, pp. 2850–2871, 2019.
- [13] Tian Li, Ahmad Beirami, Maziar Sanjabi, and Virginia Smith, “Tilted empirical risk minimization,” in *International Conference on Learning Representations*, 2021.
- [14] Tian Li, Ahmad Beirami, Maziar Sanjabi, and Virginia Smith, “On tilted losses in machine learning: Theory and applications,” *Journal of Machine Learning Research*, vol. 24, no. 142, pp. 1–79, 2023.
- [15] Gholamali Aminian, Amir R Asadi, Tian Li, Ahmad Beirami, Gesine Reinert, and Samuel N Cohen, “Generalization error of the tilted empirical risk,” in *International Conference on Machine Learning*, 2025.
- [16] E. Y. Pee and J. O. Royset, “On solving large-scale finite minimax problems using exponential smoothing,” *Journal of Optimization Theory and Applications*, vol. 148, no. 2, pp. 390 – 421, Oct. 2010.
- [17] Barry W. Kort and Dimitri P. Bertsekas, “A new penalty function method for constrained minimization,” in *IEEE Conference on Decision and Control and 11th Symposium on Adaptive Processes*, 1972, pp. 162–166.
- [18] Xueqin Wang, Yunlu Jiang, Mian Huang, and Heping Zhang and, “Robust variable selection with exponential squared loss,” *Journal of the American Statistical Association*, vol. 108, no. 502, pp. 632–643, 2013.