Online Structured Prediction with Fenchel–Young Losses and Improved Surrogate Regret for Online Multiclass Classification with Logistic Loss

Shinsaku Sakaue SAKAUE@MIST.I.U-TOKYO.AC.JP
The University of Tokyo, Tokyo, Japan

Han Bao BAO@I.KYOTO-U.AC.JP
Kyoto University, Kyoto, Japan

Taira Tsuchiya TSUCHIYA@MIST.I.U-TOKYO.AC.JP
The University of Tokyo, Tokyo, Japan

Taihei Oki OKI@MIST.I.U-TOKYO.AC.JP
The University of Tokyo, Tokyo, Japan

Editors: Shipra Agrawal and Aaron Roth

Abstract

This paper studies online structured prediction with full-information feedback. For online multiclass classification, Van der Hoeven (2020) established *finite* surrogate regret bounds, which are independent of the time horizon, by introducing an elegant *exploit-the-surrogate-gap* framework. However, this framework has been limited to multiclass classification primarily because it relies on a classification-specific procedure for converting estimated scores to outputs. We extend the exploit-the-surrogate-gap framework to online structured prediction with *Fenchel–Young losses*, a large family of surrogate losses that includes the logistic loss for multiclass classification as a special case, obtaining finite surrogate regret bounds in various structured prediction problems. To this end, we propose and analyze *randomized decoding*, which converts estimated scores to general structured outputs. Moreover, by applying our decoding to online multiclass classification with the logistic loss, we obtain a surrogate regret bound of $O(\|\boldsymbol{U}\|_{\mathrm{F}}^2)$, where $\boldsymbol{U}$ is the best offline linear estimator and $\|\cdot\|_{\mathrm{F}}$ denotes the Frobenius norm. This bound is tight up to logarithmic factors and improves the previous bound of $O(d\|\boldsymbol{U}\|_{\mathrm{F}}^2)$ due to Van der Hoeven (2020) by a factor of d, the number of classes.
Keywords: online learning, structured prediction, online multiclass classification

1. Introduction

Many machine learning problems involve predicting outputs in a finite set $\mathcal{Y}$ from input vectors in a vector space $\mathcal{X}$. A typical example is multiclass classification, and other tasks require predicting more complex structured objects, e.g., matchings and trees. Such problems, known as *structured prediction*, are ubiquitous in many applications, including natural language processing and bioinformatics (BakIr et al., 2007). Since working directly on discrete output spaces is often intractable, it is usual to adopt the surrogate loss framework (e.g., Bartlett et al. (2006)). Common examples are the logistic and hinge losses for classification. Blondel et al. (2020) have studied a family of *Fenchel–Young losses*, which subsumes many practical surrogate losses for structured prediction; see Section 2.2 for details.

Structured prediction can be naturally extended to the online learning setting: for $t = 1, \dots, T$, an adversary picks $(\boldsymbol{x}_t, \boldsymbol{y}_t) \in \mathcal{X} \times \mathcal{Y}$ and a learner plays $\widehat{\boldsymbol{y}}_t \in \mathcal{Y}$ given an input $\boldsymbol{x}_t \in \mathcal{X}$. The learner aims to minimize the cumulative target loss $\sum_{t=1}^{T} L(\widehat{\boldsymbol{y}}_t; \boldsymbol{y}_t)$, where $L : \mathcal{Y} \times \mathcal{Y} \to \mathbb{R}_{\geq 0}$ is a target loss function, such as the 0-1 and Hamming losses. This paper focuses on the full-information setting, where the true output $\boldsymbol{y}_t \in \mathcal{Y}$ is available as feedback, while another common setting is the bandit setting, where only $L(\widehat{\boldsymbol{y}}_t; \boldsymbol{y}_t)$ is given. A well-studied special case is online multiclass classification (Fink et al., 2006; Kakade et al., 2008). Let $\mathcal{Y}$ be a set representing d classes, $L(\widehat{\boldsymbol{y}}_t; \boldsymbol{y}_t)$ the standard 0-1 loss, $\boldsymbol{U} : \mathcal{X} \to \mathbb{R}^d$ the best offline linear estimator, and $S : \mathbb{R}^d \times \mathcal{Y} \to \mathbb{R}_{\geq 0}$ a surrogate loss (e.g., logistic or hinge) that measures the discrepancy between an estimated score vector in $\mathbb{R}^d$ and $\boldsymbol{y}_t \in \mathcal{Y}$. In this setting, a reasonable performance metric is the *surrogate regret* $\mathcal{R}_T$ given by

$$\sum_{t=1}^{T} L(\widehat{\boldsymbol{y}}_t; \boldsymbol{y}_t) = \sum_{t=1}^{T} S(\boldsymbol{U}\boldsymbol{x}_t; \boldsymbol{y}_t) + \mathcal{R}_T.$$ [1]

The tenet behind this metric is that while we want to minimize the cumulative target loss, the best we can do in hindsight with the surrogate loss framework is to minimize the cumulative surrogate loss, and the surrogate regret accounts for the extra target loss incurred by actual plays, $\widehat{\boldsymbol{y}}_1, \dots, \widehat{\boldsymbol{y}}_T$.

For online multiclass classification with the logistic, hinge, and smooth hinge surrogate losses, Van der Hoeven (2020) obtained an $O(d\|\boldsymbol{U}\|_{\mathrm{F}}^2)$ surrogate regret bound, where $\|\cdot\|_{\mathrm{F}}$ denotes the Frobenius norm. Notably, the bound is independent of T, or *finite*. The core idea is to exploit the gap between the 0-1 and surrogate losses, which draws inspiration from online classification with abstention (Neu and Zhivotovskiy, 2020). The author also gave an $O(dB\sqrt{T})$ surrogate regret bound for the bandit setting, where B is the ℓ_2-diameter of the domain containing $\boldsymbol{U}$. Later, Van der Hoeven et al. (2021) extended the idea to a more general feedback setting and also obtained lower bounds for the case of the smooth hinge surrogate loss. In the full-information setting, their lower bound is $\Omega(dB^2)$, implying that the $O(d\|\boldsymbol{U}\|_{\mathrm{F}}^2)$ bound for the smooth-hinge case is tight if $\|\boldsymbol{U}\|_{\mathrm{F}} = \Theta(B)$.

However, the finite surrogate regret bound provided by Van der Hoeven (2020) has been limited to online multiclass classification so far. Although the notion of surrogate regret naturally applies to more general structured prediction problems with the surrogate loss framework, the original exploit-the-surrogate-gap technique relies on a classification-specific decoding procedure for converting estimated scores in $\mathbb{R}^d$ to the outputs in $\{1, \dots, d\}$, preventing the extension to structured prediction. Since how to convert scores to structured outputs is non-trivial, it has been unclear when and how we can exploit the surrogate gap to obtain finite surrogate regret bounds in online structured prediction.[2]

We extend the exploit-the-surrogate-gap framework to online structured prediction. Regarding surrogate losses, we consider a class of Fenchel–Young losses generated by *Legendre-type* functions, due to its generality and useful properties (see Section 2.2). The main challenge lies in converting scores to the outputs in structured space $\mathcal{Y}$, for which we propose a *randomized decoding* procedure (Section 3), together with its efficient implementation based on a fast Frank–Wolfe-type algorithm (Section 3.1). Our analysis of randomized decoding (Lemma 4) reveals conditions of the structured output space, target loss, and surrogate loss under which we can obtain finite surrogate regret bounds by offsetting the regret in terms of surrogate losses with the surrogate gap. Consequently, we establish finite surrogate regret bounds that hold in expectation and with high probability (Theorems 7 and 8). Additionally, Theorem 11 shows that our randomized decoding enables *online-to-batch conversion* of surrogate regret bounds to offline guarantees on the target risk.

Although bounding the surrogate regret may seem to become easier by scaling up the surrogate loss relative to the target loss, our analysis of the surrogate regret is indeed sharp regardless of the scale of the surrogate loss. To demonstrate the sharpness, Section 5 addresses online multiclass classification with the logistic loss, the same setting as that of Van der Hoeven (2020). We obtain an $O(\|\boldsymbol{U}\|_{\mathrm{F}}^2)$

1. In statistical learning, the term "surrogate regret" sometimes refers to the excess risk of surrogate losses, but we here use the term in the above sense following Van der Hoeven et al. (2021).
2. Applying Van der Hoeven (2020) naively to $|\mathcal{Y}|$-class classification results in exponentially worse bounds in general.

surrogate regret bound (Theorem 12), which improves the previous bound of $O(d\|\boldsymbol{U}\|_{\mathrm{F}}^2)$ by a factor of d, the number of classes. We also provide an $\Omega(B^2/\ln^2 d)$ lower bound (Theorem 13), implying that our bound is tight up to $\ln d$ factors under $\|\boldsymbol{U}\|_{\mathrm{F}} = \Theta(B)$. These results shed light on an interesting $O(d)$ difference depending on surrogate losses: $O(d\|\boldsymbol{U}\|_{\mathrm{F}}^2)$ is tight for the smooth hinge loss (Van der Hoeven, 2020; Van der Hoeven et al., 2021), while $O(\|\boldsymbol{U}\|_{\mathrm{F}}^2)$ is almost tight for the logistic loss. Our work, grounded in sharp analysis, pushes the boundaries of the exploit-the-surrogate-gap framework and serves as a foundation for obtaining strong guarantees in online structured prediction.

1.1. Additional Related Work

Structured prediction. We here present particularly relevant studies and defer a literature review to Appendix A. Prior to the development of the Fenchel–Young loss framework, Niculae et al. (2018) studied *SparseMAP* inference, which trades off the MAP and marginal inference so that the estimator captures the uncertainty, provides a unique solution, and is tractable. SparseMAP regularizes the output by its squared ℓ_2-norm and solves the resulting problem with optimization algorithms, such as Frank–Wolfe-type algorithms. The fundamental idea of the Fenchel–Young losses is inherited from SparseMAP. To study the relationship between surrogate and target losses, or the *Fisher consistency*, Ciliberto et al. (2016, 2020) introduced target losses of the form $L(\boldsymbol{y}'; \boldsymbol{y}) = \langle \boldsymbol{y}', \boldsymbol{V}\boldsymbol{y}\rangle$ for some $\boldsymbol{V} \in \mathbb{R}^{d\times d}$,[3] termed as the *Structure Encoding Loss Function (SELF)* by subsequent studies. They analyzed the regularized least-square decoder and obtained a *comparison inequality*, a bound on the target excess risk in terms of the surrogate excess risk, for the squared loss. Since SELF encompasses many common target losses,[4] the framework has been oftentimes leveraged by follow-up studies. For example, Blondel (2019) studied the Fisher consistency of Fenchel–Young losses based on projection for a generalized variant of SELF, which includes the 0-1, Hamming, and NDCG losses.

Online multiclass classification. For online binary classification on *linearly separable* data, the classical Perceptron (Rosenblatt, 1958) achieves a finite surrogate regret bound. Van der Hoeven et al. (2021) extended this to multiclass classification, obtaining an $O(B^2)$ surrogate regret bound under the separability assumption, which matches the lower bound of Beygelzimer et al. (2019). By contrast, our $O(\|\boldsymbol{U}\|_{\mathrm{F}}^2)$ surrogate regret bound applies to online multiclass classification with general non-separable data. Another line of work has explored online logistic regression, where the performance is measured by the standard regret of the logistic loss.[5] In this context, Online Newton Step (Hazan et al., 2007) is known as an $O(e^{B/2}\ln T)$-regret algorithm (omitting dimension factors), and obtaining an $O(\mathrm{poly}(B)\ln T)$ regret bound had been a major open problem (McMahan and Streeter, 2012a). Despite a negative answer to the original question by Hazan et al. (2014), a seminal work by Foster et al. (2018) achieved an $O(\ln(BT))$ regret bound (a doubly exponential improvement in B) via *improper learning*, where a learner can use an estimator that is non-linear in $\boldsymbol{x}_t$. While their original algorithm is inefficient, recent studies provide more efficient $O(B\ln T)$-regret improper algorithms (Jézéquel et al., 2021; Agarwal et al., 2022). In contrast to this stream of research, we focus on obtaining finite surrogate regret bounds via proper learning. For a more extensive literature review, we refer the reader to Van der Hoeven (2020).

3. Although we focus on the Euclidean case, Ciliberto et al. (2016, 2020) consider a more general form on Hilbert spaces.
4. Indeed, any target loss on a finite set $\mathcal{Y}$ is written as a SELF with $\boldsymbol{V} \in \mathbb{R}^{|\mathcal{Y}|\times|\mathcal{Y}|}$ such that $V_{\boldsymbol{y}',\boldsymbol{y}} = L(\boldsymbol{y}'; \boldsymbol{y})$, though this representation ignores structural information of $\mathcal{Y}$ and typically causes inefficiency in learning.
5. A bound on the regret also upper bounds the surrogate regret in expectation since $1 - x \le -\log_2 x$ for $x \in (0, 1]$.

2. Preliminaries

Let $\mathbb{R}_{\geq 0}$ be the set of non-negative reals. Let $[n] = \{1, \ldots, n\}$ for any positive integer n. Let $\mathbb{1}_A$ be the 0-1 loss that takes one if A is true and zero otherwise. Let $\|\cdot\|$ be any norm (typically, ℓ_1 or ℓ_2) that satisfies $\kappa\|\boldsymbol{y}\| \geq \|\boldsymbol{y}\|_2$ for some $\kappa > 0$ for any $\boldsymbol{y} \in \mathbb{R}^d$. For a matrix $\boldsymbol{W}$, let $\|\boldsymbol{W}\|_{\mathrm{F}} = \sqrt{\mathrm{tr}(\boldsymbol{W}^\top \boldsymbol{W})}$ be the Frobenius norm. Let $\mathbf{1}$ be the all-ones vector and $\mathbf{e}_i$ the ith standard basis vector, i.e., all zeros except for the ith entry being one. For $\mathcal{C} \subseteq \mathbb{R}^d$, $\mathrm{conv}(\mathcal{C})$ denotes its convex hull, $\mathrm{int}(\mathcal{C})$ its interior, and $I_{\mathcal{C}} : \mathbb{R}^d \to \{0, +\infty\}$ its indicator function, which takes zero if $\boldsymbol{y} \in \mathcal{C}$ and $+\infty$ otherwise. For $\Omega : \mathbb{R}^d \to \mathbb{R} \cup \{+\infty\}$, $\mathrm{dom}(\Omega) := \left\{ \boldsymbol{y} \in \mathbb{R}^d : \Omega(\boldsymbol{y}) < +\infty \right\}$ denotes its effective domain and $\Omega^*(\boldsymbol{\theta}) := \sup\left\{ \langle \boldsymbol{\theta}, \boldsymbol{y} \rangle - \Omega(\boldsymbol{y}) : \boldsymbol{y} \in \mathbb{R}^d \right\}$ its convex conjugate. Let $\triangle^d := \left\{ \boldsymbol{y} \in \mathbb{R}^d_{\geq 0} : \|\boldsymbol{y}\|_1 = 1 \right\}$ be the probability simplex and $\mathsf{H}^{\mathrm{s}}(\boldsymbol{y}) := -\sum_{i=1}^d y_i \ln y_i$ the Shannon entropy of $\boldsymbol{y} \in \triangle^d$.

Let $\Psi : \mathbb{R}^d \to \mathbb{R} \cup \{+\infty\}$ be a strictly convex function differentiable throughout $\mathrm{int}(\mathrm{dom}\,\Psi) \neq \emptyset$. We say Ψ is of *Legendre-type* if $\lim_{i \to \infty} \|\nabla\Psi(\boldsymbol{x}_i)\|_2 = +\infty$ whenever $\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots$ is a sequence in $\mathrm{int}(\mathrm{dom}(\Psi))$ converging to a boundary point of $\mathrm{int}(\mathrm{dom}(\Psi))$ (see, Rockafellar (1970, Section 26)).[6] Also, given a convex set $\mathcal{C} \subseteq \mathrm{dom}(\Psi)$, we say Ψ is λ-strongly convex with respect to $\|\cdot\|$ over $\mathcal{C}$ if $\Psi(\boldsymbol{y}) \geq \Psi(\boldsymbol{y}') + \langle \nabla\Psi(\boldsymbol{y}'), \boldsymbol{y} - \boldsymbol{y}' \rangle + \frac{\lambda}{2}\|\boldsymbol{y} - \boldsymbol{y}'\|^2$ holds for any $\boldsymbol{y} \in \mathcal{C}$ and $\boldsymbol{y}' \in \mathrm{int}(\mathrm{dom}(\Psi)) \cap \mathcal{C}$.

2.1. Problem Setting

Let $\mathcal{X}$ be an input vector space. For consistency with Blondel et al. (2020), we let $\mathcal{Y}$ be the set of outputs embedded into $\mathbb{R}^d$ in the standard manner. For example, we let $\mathcal{Y} = \{\mathbf{e}_1, \ldots, \mathbf{e}_d\}$ in multiclass classification with d classes. We focus on the case where observable feedback comes from $\mathcal{Y}$.

As with online multiclass classification, we consider learning a linear estimator $\boldsymbol{W}$ that maps an input vector $\boldsymbol{x} \in \mathcal{X}$ to a score vector $\boldsymbol{W}\boldsymbol{x} \in \mathbb{R}^d$. The learning proceeds for $t = 1, \ldots, T$. In each tth round, an adversary picks an input vector $\boldsymbol{x}_t \in \mathcal{X}$ and the true output $\boldsymbol{y}_t \in \mathcal{Y}$. The learner receives $\boldsymbol{x}_t$ and computes a score vector $\boldsymbol{\theta}_t = \boldsymbol{W}_t\boldsymbol{x}_t$ with a current estimator $\boldsymbol{W}_t$. The learner then chooses $\widehat{\boldsymbol{y}}_t \in \mathcal{Y}$ based on $\boldsymbol{\theta}_t$, plays it, and incurs a target loss of $L(\widehat{\boldsymbol{y}}_t; \boldsymbol{y}_t)$. The learner receives $\boldsymbol{y}_t$ as feedback and updates $\boldsymbol{W}_t$ to $\boldsymbol{W}_{t+1}$. The goal of the learner is to minimize the cumulative target loss $\sum_{t=1}^T L(\widehat{\boldsymbol{y}}_t; \boldsymbol{y}_t)$. We assume the following conditions on the output space and the target loss.

Assumption 1 *(I) There exists $\nu > 0$ such that $\|\boldsymbol{y} - \boldsymbol{y}'\| \geq \nu$ holds for any $\boldsymbol{y}, \boldsymbol{y}' \in \mathcal{Y}$ with $\boldsymbol{y} \neq \boldsymbol{y}'$, (II) for each $\boldsymbol{y} \in \mathcal{Y}$, the target loss $L(\cdot; \boldsymbol{y})$ is defined on $\mathrm{conv}(\mathcal{Y})$, non-negative, and affine in the first argument,[7] and (III) $L(\boldsymbol{y}'; \boldsymbol{y}) \leq \gamma\|\boldsymbol{y}' - \boldsymbol{y}\|$ holds for some $\gamma > 0$, for any $\boldsymbol{y}' \in \mathrm{conv}(\mathcal{Y})$ and $\boldsymbol{y} \in \mathcal{Y}$.*

These conditions are not restrictive; see Section 2.3 for examples satisfying them. Regarding (I), ν lower bounded in many cases. For instance, if $\|\cdot\|$ is an ℓ_p-norm and $\mathcal{Y} \subseteq \mathbb{Z}^d$, $\nu \geq 1$ holds. In addition, if $\boldsymbol{y}^\top \mathbf{1}$ is constant for all $\boldsymbol{y} \in \mathcal{Y}$, distinct $\boldsymbol{y}, \boldsymbol{y}' \in \mathcal{Y}$ have at least two entries that differ by at least 1 in magnitude, hence $\nu \geq 2^{1/p}$. As for (II), SELFs $L(\boldsymbol{y}'; \boldsymbol{y}) = \langle \boldsymbol{y}', \boldsymbol{V}\boldsymbol{y} \rangle$ are defined on $\mathrm{conv}(\mathcal{Y})$ and affine in $\boldsymbol{y}'$. Moreover, Blondel (2019, Appendix A) provides many target losses expressed as $L(\boldsymbol{y}'; \boldsymbol{y}) = \langle \boldsymbol{y}', \boldsymbol{V}\boldsymbol{y} + \boldsymbol{b} \rangle + c(\boldsymbol{y})$ for $\boldsymbol{y}, \boldsymbol{y}' \in \mathrm{conv}(\mathcal{Y})$ with some $\boldsymbol{V} \in \mathbb{R}^{d \times d}$, $\boldsymbol{b} \in \mathbb{R}^d$ and $c(\boldsymbol{y}) \in \mathbb{R}$, which are also defined on $\mathrm{conv}(\mathcal{Y})$ and affine in $\boldsymbol{y}'$. Condition (III) is typically satisfied by moderate γ values (see Section 2.3). Note that the non-negativity and (III) imply $L(\boldsymbol{y}'; \boldsymbol{y}) = 0$ if $\boldsymbol{y}' = \boldsymbol{y}$.

6. Strictly speaking, this property is *essential smoothness*, which, combined with strict convexity, implies Legendre-type.
7. Condition (II) is assumed for technical convenience, although target losses are inherently defined on $\mathcal{Y}$. This specifically ensures $\mathbb{E}[L(\widetilde{\boldsymbol{y}}; \boldsymbol{y})] = L(\mathbb{E}[\widetilde{\boldsymbol{y}}]; \boldsymbol{y})$ for $\widetilde{\boldsymbol{y}}$ drawn randomly from $\mathcal{Y}$, which we will use in the proof of Lemma 4.

2.2. Fenchel–Young Loss

We adopt the surrogate loss framework considered in Blondel et al. (2020). We define an intermediate *score space* $\mathbb{R}^d$ between $\mathcal{X}$ and $\mathcal{Y}$ and measure the discrepancy between a score vector $\boldsymbol{\theta} \in \mathbb{R}^d$ and the ground truth $\boldsymbol{y} \in \mathcal{Y}$ with a surrogate loss $S : \mathbb{R}^d \times \mathcal{Y} \to \mathbb{R}_{\geq 0}$; here, we suppose $\boldsymbol{\theta}$ to be given by $\boldsymbol{W}_t \boldsymbol{x}_t$ as in Section 2.1. Blondel et al. (2020) provides a general recipe for designing various surrogate losses, called *Fenchel–Young losses*, for structured prediction from regularization functions.

Definition 2 *Let $\Omega : \mathbb{R}^d \to \mathbb{R} \cup \{+\infty\}$ be a regularization function such that $\mathcal{Y} \subseteq \mathrm{dom}(\Omega)$. The Fenchel–Young loss $S_\Omega : \mathrm{dom}(\Omega^*) \times \mathrm{dom}(\Omega) \to \mathbb{R}_{\geq 0}$ generated by Ω is defined as*

$$S_\Omega(\boldsymbol{\theta}; \boldsymbol{y}) \coloneqq \Omega^*(\boldsymbol{\theta}) + \Omega(\boldsymbol{y}) - \langle \boldsymbol{\theta}, \boldsymbol{y} \rangle.$$

By definition, $S_\Omega(\boldsymbol{\theta}; \boldsymbol{y})$ is convex in $\boldsymbol{\theta}$ for any $\boldsymbol{y} \in \mathrm{dom}(\Omega)$. Furthermore, $S_\Omega(\boldsymbol{\theta}; \boldsymbol{y}) \geq 0$ follows from the Fenchel–Young inequality, and $S_\Omega(\boldsymbol{\theta}; \boldsymbol{y}) = 0$ holds if and only if $\boldsymbol{y} \in \partial\Omega^*(\boldsymbol{\theta})$.

We focus on special Fenchel–Young losses studied in Blondel et al. (2020, Section 3.2), which are generated by Ω of the form $\Psi + I_{\mathrm{conv}(\mathcal{Y})}$ (i.e., Ω is the restriction of Ψ to $\mathrm{conv}(\mathcal{Y})$), where Ψ is differentiable, of Legendre-type, and λ-strongly convex w.r.t. $\|\cdot\|$, and satisfies $\mathrm{conv}(\mathcal{Y}) \subseteq \mathrm{dom}(\Psi)$ and $\mathrm{dom}(\Psi^*) = \mathbb{R}^d$. Such Fenchel–Young losses subsume various useful surrogate losses, including the logistic, CRF, and SparseMAP losses, and enjoy the following helpful properties. See Blondel et al. (2020, Propositions 2 and 3) for more details, and also Appendix B for a note on the CRF loss.

Proposition 3 *Let S_Ω be a Fenchel–Young loss generated by $\Omega = \Psi + I_{\mathrm{conv}(\mathcal{Y})}$, where $\Psi : \mathbb{R}^d \to \mathbb{R} \cup \{+\infty\}$ satisfies the above properties. For $\boldsymbol{\theta} \in \mathbb{R}^d$, define the regularized prediction function as*

$$\widehat{\boldsymbol{y}}_\Omega(\boldsymbol{\theta}) \coloneqq \arg\max\{ \langle \boldsymbol{\theta}, \boldsymbol{y} \rangle - \Omega(\boldsymbol{y}) : \boldsymbol{y} \in \mathbb{R}^d \} = \arg\max\{ \langle \boldsymbol{\theta}, \boldsymbol{y} \rangle - \Psi(\boldsymbol{y}) : \boldsymbol{y} \in \mathrm{conv}(\mathcal{Y}) \},$$

where the maximizer is unique. Then, for any $\boldsymbol{y} \in \mathcal{Y}$, $S_\Omega(\boldsymbol{\theta}; \boldsymbol{y})$ is differentiable in $\boldsymbol{\theta}$ and the gradient is the residual, i.e., $\nabla S_\Omega(\boldsymbol{\theta}; \boldsymbol{y}) = \widehat{\boldsymbol{y}}_\Omega(\boldsymbol{\theta}) - \boldsymbol{y}$. Furthermore, $S_\Omega(\boldsymbol{\theta}; \boldsymbol{y}) \geq \frac{\lambda}{2}\|\boldsymbol{y} - \widehat{\boldsymbol{y}}_\Omega(\boldsymbol{\theta})\|^2$ holds.[8]

The last inequality will turn out useful in analyzing our randomized decoding (see Lemma 4). Proposition 3 also implies $\|\nabla S_\Omega(\boldsymbol{\theta}; \boldsymbol{y})\|^2 \leq \frac{2}{\lambda} S_\Omega(\boldsymbol{\theta}; \boldsymbol{y})$, which we will use in the proof of Theorem 7. This type of inequality plays a crucial role in exploiting the surrogate gap, as highlighted in Proposition 6.

2.3. Examples

Below are three typical structured prediction problems and Fenchel–Young losses satisfying the above conditions, and Appendix C gives two more examples; all the five are considered in Blondel (2019, Section 4). More examples of structured outputs are provided in Blondel et al. (2020, Section 7.3).

Multiclass classification. Let $\mathcal{Y} = \{\mathbf{e}_1, \dots, \mathbf{e}_d\}$ and $\|\cdot\|$ be the ℓ_1-norm. Since $\|\mathbf{e}_i - \mathbf{e}_j\|_1 \geq 2$ holds for any distinct $i, j \in [d]$, we have $\nu = 2$. For any $\mathbf{e}_i \in \mathcal{Y}$, the 0-1 loss, $L(\boldsymbol{y}'; \mathbf{e}_i) = \mathbb{1}_{\boldsymbol{y}' \neq \mathbf{e}_i}$, can be extended on $\mathrm{conv}(\mathcal{Y})$ as $L(\boldsymbol{y}'; \mathbf{e}_i) = \langle \boldsymbol{y}', \mathbf{1} - \mathbf{e}_i \rangle$, which is affine in $\boldsymbol{y}'$ and equals $\sum_{j \neq i} y'_j = \frac{1}{2}\left(1 - y'_i + \sum_{j \neq i} y'_j\right) = \frac{1}{2}\|\mathbf{e}_i - \boldsymbol{y}'\|_1$ due to $\sum_{i=1}^n y'_i = 1$, hence $\gamma = \frac{1}{2}$. As detailed in Section 5, the logistic loss can be written as a Fenchel–Young loss generated by an entropic regularizer Ω.[9]

8. Blondel et al. (2020, Proposition 3) shows $S_\Omega(\boldsymbol{\theta}; \boldsymbol{y}) \geq B_\Psi(\boldsymbol{y} \,\|\, \widehat{\boldsymbol{y}}_\Omega(\boldsymbol{\theta}))$, where B_Ψ is the Bregman divergence induced by Ψ, and $B_\Psi(\boldsymbol{y} \,\|\, \widehat{\boldsymbol{y}}_\Omega(\boldsymbol{\theta})) \geq \frac{\lambda}{2}\|\boldsymbol{y} - \widehat{\boldsymbol{y}}_\Omega(\boldsymbol{\theta})\|^2$ follows from the λ-strong convexity of Ψ with respect to $\|\cdot\|$. The same inequality is also used in Blondel (2019, Lemma 3). This is the only part where we need Ψ to be of Legendre-type.

9. The multiclass hinge loss (Crammer and Singer, 2001) is also written as a *cost-sensitive* Fenchel–Young loss (Blondel et al., 2020, Section 3.4). However, this requires the ground truth $\boldsymbol{y}$ when computing a counterpart of $\widehat{\boldsymbol{y}}_\Omega(\boldsymbol{\theta})$, which we need before $\boldsymbol{y}_t$ is revealed (see Algorithm 2). Thus, it seems difficult to apply our approach to the (smooth) hinge loss.

Algorithm 1 Randomized decoding ψ_Ω

Input: $\boldsymbol{\theta} \in \mathbb{R}^d$

1: $\widehat{\boldsymbol{y}}_\Omega(\boldsymbol{\theta}) \leftarrow \arg\max\{ \langle \boldsymbol{\theta}, \boldsymbol{y} \rangle - \Psi(\boldsymbol{y}) : \boldsymbol{y} \in \mathrm{conv}(\mathcal{Y}) \}$

2: $\boldsymbol{y}^* \leftarrow \arg\min\{ \|\boldsymbol{y} - \widehat{\boldsymbol{y}}_\Omega(\boldsymbol{\theta})\| : \boldsymbol{y} \in \mathcal{Y} \}$ (breaking ties arbitrarily)

3: $\Delta^* \leftarrow \|\boldsymbol{y}^* - \widehat{\boldsymbol{y}}_\Omega(\boldsymbol{\theta})\|$ and $p \leftarrow \min\{1, 2\Delta^*/\nu\}$

4: $Z \leftarrow 0$ with probability $1-p$; $Z \leftarrow 1$ with probability p

5: $\widehat{\boldsymbol{y}} \leftarrow \begin{cases} \boldsymbol{y}^* & \text{if } Z = 0 \\ \widetilde{\boldsymbol{y}} & \text{if } Z = 1\text{, where } \widetilde{\boldsymbol{y}} \text{ is randomly drawn from } \mathcal{Y} \text{ so that } \mathbb{E}[\widetilde{\boldsymbol{y}} \mid Z = 1] = \widehat{\boldsymbol{y}}_\Omega(\boldsymbol{\theta}) \end{cases}$

6: **return** $\psi_\Omega(\boldsymbol{\theta}) = \widehat{\boldsymbol{y}}$

Multilabel classification. We consider multilabel classification with $\mathcal{Y} = \{0,1\}^d$. If $\|\cdot\|$ is the ℓ_2-norm, we have $\nu = 1$. A common target loss is the Hamming loss $L(\boldsymbol{y}'; \boldsymbol{y}) = \frac{1}{d}\sum_{i=1}^d \mathbb{1}_{y'_i \neq y_i}$, where the division by d scales the loss to $[0,1]$. For any $\boldsymbol{y} \in \mathcal{Y}$, it is represented on $\mathrm{conv}(\mathcal{Y})$ as $L(\boldsymbol{y}'; \boldsymbol{y}) = \frac{1}{d}(\langle \boldsymbol{y}', \mathbf{1} \rangle + \langle \boldsymbol{y}, \mathbf{1} \rangle - 2\langle \boldsymbol{y}', \boldsymbol{y} \rangle)$, which is affine in $\boldsymbol{y}'$ and satisfies $L(\boldsymbol{y}'; \boldsymbol{y}) = \frac{1}{d}\|\boldsymbol{y}' - \boldsymbol{y}\|_1 \le \frac{1}{\sqrt{d}}\|\boldsymbol{y}' - \boldsymbol{y}\|_2$, hence $\gamma = \frac{1}{\sqrt{d}}$. If we let $\Omega = \frac{1}{2}\|\cdot\|_2^2 + I_{\mathrm{conv}(\mathcal{Y})}$, we have $\lambda = 1$ (i.e., 1-strongly convex), and the resulting Fenchel–Young loss is the SparseMAP loss: $S_\Omega(\boldsymbol{\theta}; \boldsymbol{y}) = \frac{1}{2}\|\boldsymbol{y} - \boldsymbol{\theta}\|_2^2 - \frac{1}{2}\|\widehat{\boldsymbol{y}}_\Omega(\boldsymbol{\theta}) - \boldsymbol{\theta}\|_2^2$.

Ranking. We consider predicting the ranking of n items. Let $d = n^2$ and $\mathcal{Y} \subseteq \{0,1\}^d$ be the set of all $n \times n$ permutation matrices, vectorized into $\{0,1\}^d$. Then, $\mathrm{conv}(\mathcal{Y})$ is the Birkhoff polytope. For $\boldsymbol{y} \in \mathrm{conv}(\mathcal{Y})$, y_{ij} refers to the (i,j) entry of the corresponding matrix. If $\|\cdot\|$ is the ℓ_1-norm, $\|\boldsymbol{y} - \boldsymbol{y}'\|_1 \ge 4$ holds for distinct $\boldsymbol{y}, \boldsymbol{y}' \in \mathcal{Y}$, hence $\nu = 4$. We use a target loss that counts mismatches. Specifically, let $L(\boldsymbol{y}'; \boldsymbol{y}) = \frac{1}{n}\sum_{i=1}^n \mathbb{1}_{y'_{i,j_i} \neq y_{i,j_i}}$ for $\boldsymbol{y}, \boldsymbol{y}' \in \mathcal{Y}$, where $j_i \in [n]$ is a unique index such that $y_{ij_i} = 1$ for each $i \in [n]$ and the division by n scales the loss to $[0,1]$. For any $\boldsymbol{y} \in \mathcal{Y}$, this loss can be represented on $\mathrm{conv}(\mathcal{Y})$ as $L(\boldsymbol{y}'; \boldsymbol{y}) = \frac{1}{n}\langle \boldsymbol{y}', \mathbf{1} - \boldsymbol{y} \rangle$, which is affine in $\boldsymbol{y}'$. Furthermore, it equals $\frac{1}{2n}\sum_{i=1}^n (1 - y'_{ij_i} + \sum_{j \neq j_i} y'_{ij}) = \frac{1}{2n}\|\boldsymbol{y}' - \boldsymbol{y}\|_1$ since $\sum_j y'_{ij} = 1$ holds for each $i \in [n]$, hence $\gamma = \frac{1}{2n}$. Drawing inspiration from celebrated entropic optimal transport (Cuturi, 2013), we consider a Fenchel–Young loss generated by $\Omega = -\frac{1}{\mu}\mathsf{H}^{\mathrm{s}} + I_{\mathrm{conv}(\mathcal{Y})}$, where $\mu > 0$ controls the regularization strength.[10] Since $-\frac{1}{\mu}\mathsf{H}^{\mathrm{s}}$ is $\frac{1}{n\mu}$-strongly convex w.r.t. $\|\cdot\|_1$ over $\mathrm{conv}(\mathcal{Y})$ (Blondel, 2019, Proposition 2), we have $\lambda = \frac{1}{n\mu}$. The resulting $S_\Omega(\boldsymbol{\theta}; \boldsymbol{y})$ is written as $\langle \boldsymbol{\theta}, \widehat{\boldsymbol{y}}_\Omega(\boldsymbol{\theta}) - \boldsymbol{y} \rangle + \frac{1}{\mu}\mathsf{H}^{\mathrm{s}}(\widehat{\boldsymbol{y}}_\Omega(\boldsymbol{\theta}))$, where the first term measures the affinity between $\boldsymbol{\theta}$ and $\boldsymbol{y} \in \mathcal{Y}$, and the second term penalizes the uncertainty of $\widehat{\boldsymbol{y}}_\Omega(\boldsymbol{\theta})$.

3. Randomized Decoding

We present our key technical tool, randomized decoding, for converting a score vector $\boldsymbol{\theta} \in \mathbb{R}^d$ to an output $\widehat{\boldsymbol{y}} \in \mathcal{Y}$. Our randomized decoding (Algorithm 1) returns either $\boldsymbol{y}^* \in \mathcal{Y}$ closest to $\widehat{\boldsymbol{y}}_\Omega(\boldsymbol{\theta}) \in \mathrm{conv}(\mathcal{Y})$ or random $\widetilde{\boldsymbol{y}} \in \mathcal{Y}$ such that $\mathbb{E}[\widetilde{\boldsymbol{y}} \mid Z = 1] = \widehat{\boldsymbol{y}}_\Omega(\boldsymbol{\theta})$, where Ω is a regularization function generating the Fenchel–Young loss S_Ω and Z is the Bernoulli random variable with parameter p, as in Step 4. Intuitively, the closer the regularized prediction $\widehat{\boldsymbol{y}}_\Omega(\boldsymbol{\theta})$ is to $\boldsymbol{y}^*$ (i.e., Δ^* is smaller), the more confident $\boldsymbol{\theta}$ is about $\boldsymbol{y}^*$, and hence the decoding procedure returns $\boldsymbol{y}^*$ with a higher probability; otherwise, $\boldsymbol{\theta}$ is not sufficiently confident about any $\boldsymbol{y} \in \mathcal{Y}$, and hence the decoding procedure more

10. With this choice of Ω, we can efficiently compute the regularized prediction $\widehat{\boldsymbol{y}}_\Omega(\boldsymbol{\theta})$ with the Sinkhorn algorithm.

likely returns a random $\widetilde{\boldsymbol{y}}$. The confidence is quantified by $2\Delta^*/\nu$ (smaller values indicate higher confidence), where ν is the minimum distance between distinct elements in $\mathcal{Y}$, as in Assumption 1.[11]

The following Lemma 4 is our main technical result regarding the randomized decoding. Despite the simplicity of the proof, it plays a crucial role in the subsequent analysis of the surrogate regret.

Lemma 4 *For any $(\boldsymbol{\theta}, \boldsymbol{y}) \in \mathbb{R}^d \times \mathcal{Y}$, the randomized decoding ψ_Ω (Algorithm 1) satisfies*

$$\mathbb{E}[L(\psi_\Omega(\boldsymbol{\theta}); \boldsymbol{y})] \le \frac{4\gamma}{\lambda\nu} S_\Omega(\boldsymbol{\theta}; \boldsymbol{y}).$$

Proof Let $\Delta = \|\boldsymbol{y} - \widehat{\boldsymbol{y}}_\Omega(\boldsymbol{\theta})\|$. Note that $\Delta \ge \Delta^* = \|\boldsymbol{y}^* - \widehat{\boldsymbol{y}}_\Omega(\boldsymbol{\theta})\|$ holds by definition of $\boldsymbol{y}^*$. Since $L(\cdot; \boldsymbol{y})$ is affine as in Assumption 1, we have $\mathbb{E}[L(\widetilde{\boldsymbol{y}}; \boldsymbol{y}) \mid Z = 1] = L(\widehat{\boldsymbol{y}}_\Omega(\boldsymbol{\theta}); \boldsymbol{y})$. Thus, it holds that

$$\begin{aligned}\mathbb{E}[L(\psi_\Omega(\boldsymbol{\theta}); \boldsymbol{y})] &= (1-p)L(\boldsymbol{y}^*; \boldsymbol{y}) + pL(\widehat{\boldsymbol{y}}_\Omega(\boldsymbol{\theta}); \boldsymbol{y}) \\ &= \begin{cases} pL(\widehat{\boldsymbol{y}}_\Omega(\boldsymbol{\theta}); \boldsymbol{y}) & \text{if } \Delta^* \ge \nu/2 \text{ or } \boldsymbol{y}^* = \boldsymbol{y}, \\ (1-p)L(\boldsymbol{y}^*; \boldsymbol{y}) + pL(\widehat{\boldsymbol{y}}_\Omega(\boldsymbol{\theta}); \boldsymbol{y}) & \text{if } \Delta^* < \nu/2 \text{ and } \boldsymbol{y}^* \ne \boldsymbol{y}. \end{cases}\end{aligned}$$

Below, we will prove $\mathbb{E}[L(\psi_\Omega(\boldsymbol{\theta}); \boldsymbol{y})] \le 2\gamma\Delta^2/\nu$; then, the desired bound follows from $\lambda\Delta^2/2 \le S_\Omega(\boldsymbol{\theta}; \boldsymbol{y})$ given in Proposition 3. In the first case, from $p \le 2\Delta^*/\nu \le 2\Delta/\nu$ and $L(\widehat{\boldsymbol{y}}_\Omega(\boldsymbol{\theta}); \boldsymbol{y}) \le \gamma\Delta$ (due to Assumption 1), we obtain $pL(\widehat{\boldsymbol{y}}_\Omega(\boldsymbol{\theta}); \boldsymbol{y}) \le 2\gamma\Delta^2/\nu$. In the second case, by using $p = 2\Delta^*/\nu$, $L(\boldsymbol{y}'; \boldsymbol{y}) \le \gamma\|\boldsymbol{y}' - \boldsymbol{y}\|$ for any $\boldsymbol{y}' \in \mathrm{conv}(\mathcal{Y})$ (Assumption 1), and the triangle inequality, we obtain

$$\begin{aligned}\mathbb{E}[L(\psi_\Omega(\boldsymbol{\theta}); \boldsymbol{y})] &\le (1 - 2\Delta^*/\nu)\gamma\|\boldsymbol{y}^* - \boldsymbol{y}\| + (2\Delta^*/\nu)\gamma\|\widehat{\boldsymbol{y}}_\Omega(\boldsymbol{\theta}) - \boldsymbol{y}\| \\ &\le (1 - 2\Delta^*/\nu)\gamma(\|\boldsymbol{y}^* - \widehat{\boldsymbol{y}}_\Omega(\boldsymbol{\theta})\| + \|\widehat{\boldsymbol{y}}_\Omega(\boldsymbol{\theta}) - \boldsymbol{y}\|) + (2\Delta^*/\nu)\gamma\|\widehat{\boldsymbol{y}}_\Omega(\boldsymbol{\theta}) - \boldsymbol{y}\| \\ &= (1 - 2\Delta^*/\nu)\gamma\Delta^* + \gamma\Delta.\end{aligned}$$

Hence, it suffices to prove $(1 - 2\Delta^*/\nu)\gamma\Delta^* + \gamma\Delta \le 2\gamma\Delta^2/\nu$; by dividing both sides by $\gamma\nu$ and letting $u = \Delta^*/\nu$ and $v = \Delta/\nu$, this can be simplified as $2u^2 + 2v^2 - u - v \ge 0$. From the triangle inequality and $\boldsymbol{y}^* \ne \boldsymbol{y}$, we have $\Delta^* + \Delta \ge \|\boldsymbol{y}^* - \boldsymbol{y}\| \ge \nu$, i.e., $u + v \ge 1$. Also, $\Delta^* < \nu/2$ implies $u < 1/2$. Combining them yields $0 \le u < 1/2 < v$. These imply the desired inequality as follows:

$$2u^2 + 2v^2 - u - v = (u + v - 1)(2u + 2v - 1) + (2v - 1)(1 - 2u) \ge 0.$$

Therefore, we have $\mathbb{E}[L(\psi_\Omega(\boldsymbol{\theta}); \boldsymbol{y})] \le 2\gamma\Delta^2/\nu$ in any case, completing the proof. ∎

As we will see in Proposition 6, given a possibly randomized decoding function $\psi : \mathbb{R}^d \to \mathcal{Y}$, a sufficient condition for achieving finite surrogate regret bounds is the existence of $a \in (0, 1)$ such that $\mathbb{E}[L(\psi(\boldsymbol{\theta}); \boldsymbol{y})] \le (1 - a)S_\Omega(\boldsymbol{\theta}; \boldsymbol{y})$ holds for any $(\boldsymbol{\theta}, \boldsymbol{y}) \in \mathbb{R}^d \times \mathcal{Y}$, which leads to a surrogate regret bound proportional to $1/a$. We call the quantity $aS_\Omega(\boldsymbol{\theta}; \boldsymbol{y})$ the *surrogate gap*.[12] Lemma 4 will ensure that our randomized decoding offers meaningful surrogate gaps.

11. This confidence measure is based on a rationale that $2\Delta^*/\nu < 1$ ensures that $\boldsymbol{y}^*$ is the closest point to $\widehat{\boldsymbol{y}}_\Omega(\boldsymbol{\theta})$ among all $\boldsymbol{y} \in \mathcal{Y}$. Specifically, if $\Delta^* = \|\boldsymbol{y}^* - \widehat{\boldsymbol{y}}_\Omega(\boldsymbol{\theta})\| < \nu/2$ holds, for any $\boldsymbol{y} \in \mathcal{Y} \setminus \{\boldsymbol{y}^*\}$, the triangle inequality implies $\|\boldsymbol{y} - \widehat{\boldsymbol{y}}_\Omega(\boldsymbol{\theta})\| \ge \|\boldsymbol{y} - \boldsymbol{y}^*\| - \|\boldsymbol{y}^* - \widehat{\boldsymbol{y}}_\Omega(\boldsymbol{\theta})\| > \nu - \nu/2 > \|\boldsymbol{y}^* - \widehat{\boldsymbol{y}}_\Omega(\boldsymbol{\theta})\|$. Note that the opposite is not always true.
12. The original definition of the surrogate gap (Van der Hoeven, 2020) slightly differs, but represents a similar quantity.

Necessity of mixing $\boldsymbol{y}^*$ and $\widetilde{\boldsymbol{y}}$. Our randomized decoding is a mixture of two strategies: returning $\boldsymbol{y}^*$ or random $\widetilde{\boldsymbol{y}}$.[13] We explain that either strategy alone does not yield the desired surrogate gap. Let us discuss the deterministic decoding that always returns $\boldsymbol{y}^*$. Consider binary classification with $\mathcal{Y} = \{\mathbf{e}_1, \mathbf{e}_2\}$. Let $\boldsymbol{y} = \mathbf{e}_1$ be the ground truth and $\boldsymbol{\theta} = (\theta_1, \theta_2) = (1, 1 + \ln(2^{1+\varepsilon} - 1))$ for some small $\varepsilon > 0$, which slightly favors $\mathbf{e}_2$ by mistake. Then, the logistic loss is $\log_2(1 + \exp(\theta_2 - \theta_1)) = 1 + \varepsilon$, and the 0-1 loss is 1 since the deterministic decoding converts $\boldsymbol{\theta}$ to $\mathbf{e}_2$. Thus, only a surrogate gap with $a \le \frac{\varepsilon}{1+\varepsilon}$ is left, leading to an arbitrarily large $\Omega(1/\varepsilon)$ surrogate regret bound. By contrast, our randomized decoding applied to this setting yields a surrogate gap with $a = 1 - \ln 2 \in (0, 1)$ (see Theorem 12). Further investigation of this example also suggests that the multiplicative constant of 4 in Lemma 4 cannot be smaller than $2/\ln 2 \approx 2.89$ (see Appendix D). Next, we discuss the strategy that always returns random $\widetilde{\boldsymbol{y}}$. If we do so (i.e., fix p to 1 in the proof of Lemma 4), we only have $\mathbb{E}[L(\psi_\Omega(\boldsymbol{\theta}); \boldsymbol{y})] = L(\widehat{\boldsymbol{y}}_\Omega(\boldsymbol{\theta}); \boldsymbol{y}) \le \gamma\Delta$ by Assumption 1 and $\lambda\Delta^2/2 \le S_\Omega(\boldsymbol{\theta}; \boldsymbol{y})$ by Proposition 3. These do not imply the desired relation, $\mathbb{E}[L(\psi_\Omega(\boldsymbol{\theta}); \boldsymbol{y})] \le (1 - a) S_\Omega(\boldsymbol{\theta}; \boldsymbol{y})$, when $\Delta \ll 1$. (While we have $\mathbb{E}[L(\psi_\Omega(\boldsymbol{\theta}); \boldsymbol{y})] \lesssim \sqrt{S_\Omega(\boldsymbol{\theta}; \boldsymbol{y})}$, this does not enable us to exploit the surrogate gap; see the proof of Proposition 6.) By adjusting the bias toward $\boldsymbol{y}^*$, we can avoid this issue when $\widehat{\boldsymbol{y}}_\Omega(\boldsymbol{\theta})$ is very close to some $\boldsymbol{y}^*$ while moderating the penalty of mistake, $\boldsymbol{y}^* \ne \boldsymbol{y}$.

3.1. Implementation of Randomized Decoding

Algorithm 1 involves computing $\widehat{\boldsymbol{y}}_\Omega(\boldsymbol{\theta})$ and $\boldsymbol{y}^*$, and sampling $\widetilde{\boldsymbol{y}}$. We can obtain $\widehat{\boldsymbol{y}}_\Omega(\boldsymbol{\theta})$ by solving the convex optimization in Step 1, and efficient methods for this problem are extensively discussed in Blondel et al. (2020, Section 8.3); also, we can use a fast Frank–Wolfe-type algorithm of Garber and Wolf (2021) to obtain $\widehat{\boldsymbol{y}}_\Omega(\boldsymbol{\theta})$, as described shortly. Below, we focus on how to obtain $\boldsymbol{y}^*$ and $\widetilde{\boldsymbol{y}}$ first.

In Step 2, we need to find a nearest extreme point $\boldsymbol{y}^* \in \mathcal{Y}$ to $\widehat{\boldsymbol{y}}_\Omega(\boldsymbol{\theta})$ with respect to the distance induced by $\|\cdot\|$. In the case of multiclass classification, we can easily do this by choosing $i \in [d]$ corresponding to the largest entry in $\widehat{\boldsymbol{y}}_\Omega(\boldsymbol{\theta})$ and setting $\boldsymbol{y}^* = \mathbf{e}_i$. More generally, Proposition 5 ensures that if $\mathcal{Y} \subseteq \{0,1\}^d$, which is a common scenario where $\mathrm{conv}(\mathcal{Y})$ constitutes a 0-1 polytope, and $\|\cdot\|$ is an ℓ_p-norm, we can find such a nearest extreme point by solving a linear optimization problem.

Proposition 5 *Let $\mathcal{Y} \subseteq \{0,1\}^d$ and $p \in [1, +\infty)$. For any $\boldsymbol{y}' \in \mathrm{conv}(\mathcal{Y})$, we can find a nearest extreme point $\boldsymbol{y}^* \in \mathcal{Y}$ to $\boldsymbol{y}'$ with respect to $\|\cdot\|_p$, i.e., $\boldsymbol{y}^* \in \arg\min\{\, \|\boldsymbol{y} - \boldsymbol{y}'\|_p : \boldsymbol{y} \in \mathcal{Y} \,\}$, via a single call to a linear optimization oracle that, for any $\boldsymbol{c} \in \mathbb{R}^d$, returns a point in $\arg\min\{\, \langle \boldsymbol{c}, \boldsymbol{y} \rangle : \boldsymbol{y} \in \mathcal{Y} \,\}$.*

Proof We can find a nearest point by minimizing $\|\boldsymbol{y} - \boldsymbol{y}'\|_p^p = \sum_{i=1}^d |y_i - y_i'|^p$ over $\boldsymbol{y} \in \mathcal{Y}$. Since we have $y_i \in \{0,1\}$, we can rewrite each term as $|1 - y_i'|^p y_i + |y_i'|^p (1 - y_i) = (|1 - y_i'|^p - |y_i'|^p) y_i + |y_i'|^p$. Therefore, the problem is equivalent to $\min\left\{\, \sum_{i=1}^d (|1 - y_i'|^p - |y_i'|^p) y_i : \boldsymbol{y} \in \mathcal{Y} \,\right\}$, which we can solve with the linear optimization oracle. ■

Proposition 5 enables efficient computation of $\boldsymbol{y}^*$ for various structures of $\mathrm{conv}(\mathcal{Y})$: the 0-1 hypercube (multilabel classification), the Birkhoff polytope (ranking), and a general matroid polytope. Garber and Wolf (2021, Section 1.2) shows more examples where we can compute nearest extreme points. The 0-1 polytope case is given there only for the ℓ_2-norm, and Proposition 5 extends it to ℓ_p-norms.

We turn to how to sample $\widetilde{\boldsymbol{y}} \in \mathcal{Y}$ such that $\mathbb{E}[\widetilde{\boldsymbol{y}} \mid Z = 1] = \widehat{\boldsymbol{y}}_\Omega(\boldsymbol{\theta})$ in Step 5. This is also easy in multiclass classification: we sample $i \in [d]$ with probability proportional to the ith entry of $\widehat{\boldsymbol{y}}_\Omega(\boldsymbol{\theta})$

13. While Van der Hoeven (2020) uses a similar mixing strategy, a difference lies in the definition of $\widetilde{\boldsymbol{y}}$, which is crucial for shaving the $O(d)$ factor in the case of online multiclass classification. See Section 5 for a detailed discussion.

Algorithm 2 Learning procedure for online structured prediction

Input: Alg with domain $\mathcal{W}$ and decoding function ψ_Ω (Algorithm 1)

Set $\boldsymbol{W}_1$ to the all-zero matrix
for $t = 1, \dots, T$ **do**
  Receive $\boldsymbol{x}_t$ and compute $\boldsymbol{\theta}_t = \boldsymbol{W}_t \boldsymbol{x}_t$
  Play $\widehat{\boldsymbol{y}}_t = \psi_\Omega(\boldsymbol{\theta}_t)$ and observe $\boldsymbol{y}_t$
  Send S_t (or $(\boldsymbol{x}_t, \boldsymbol{y}_t)$) to Alg and get $\boldsymbol{W}_{t+1}$ in return

and set $\widetilde{\boldsymbol{y}} = \mathbf{e}_i$. In general, if we have a convex combination of extreme points of $\mathrm{conv}(\mathcal{Y})$ that equals $\widehat{\boldsymbol{y}}_\Omega(\boldsymbol{\theta})$, we can sample $\widetilde{\boldsymbol{y}}$ by choosing an extreme point with a probability of the corresponding combination coefficient. Such a convex combination can be obtained by applying a Frank–Wolfe-type algorithm to $\min\{ \|\boldsymbol{y} - \widehat{\boldsymbol{y}}_\Omega(\boldsymbol{\theta})\|_2^2 : \boldsymbol{y} \in \mathrm{conv}(\mathcal{Y}) \}$, as considered in Combettes and Pokutta (2023) (or, we may directly compute $\widehat{\boldsymbol{y}}_\Omega(\boldsymbol{\theta})$ with a Frank–Wolfe-type algorithm). In particular, given that we can efficiently compute nearest extreme points as discussed above, we can use the linearly convergent Frank–Wolfe algorithm of Garber and Wolf (2021, Theorem 5), which returns an ε-approximation of $\widehat{\boldsymbol{y}}_\Omega(\boldsymbol{\theta})$ as a convex combination of only $O(M \ln(d/\varepsilon))$ extreme points (typically, $M = O(d^2)$).

4. Surrogate Regret Bounds for Online Structured Prediction

We analyze the surrogate regret for online structured prediction. To simplify the notation, let $L_t(\boldsymbol{y}) = L(\boldsymbol{y}; \boldsymbol{y}_t)$ and $S_t(\boldsymbol{W}) = S_\Omega(\boldsymbol{W}\boldsymbol{x}_t; \boldsymbol{y}_t)$ be the target and surrogate losses in the tth round, respectively. We also use $\mathbb{E}_t$ to represent the expectation taken only with respect to the randomness of the randomized decoding to produce $\widehat{\boldsymbol{y}}_t$ in the tth round, i.e., $\mathbb{E}_t$ is conditioned on $\widehat{\boldsymbol{y}}_1, \dots, \widehat{\boldsymbol{y}}_{t-1}$. The learning procedure is summarized in Algorithm 2. In each tth round, the learner receives $\boldsymbol{x}_t$, computes $\boldsymbol{\theta}_t = \boldsymbol{W}_t \boldsymbol{x}_t$, plays $\widehat{\boldsymbol{y}}_t$ obtained by decoding $\boldsymbol{\theta}_t$, and observe $\boldsymbol{y}_t$. The learner updates $\boldsymbol{W}_t$ using an online convex optimization algorithm, denoted by Alg, with domain $\mathcal{W}$ and loss function S_t. Below, we assume that $\mathcal{W}$ contains the all-zero matrix and set $\boldsymbol{W}_1$ to it for convenience.

As with Van der Hoeven (2020), we here use the online gradient descent (OGD) with a constant learning rate $\eta > 0$ as Alg; we discuss using other online learning methods in Appendix E. This OGD achieves the following regret bound for any $\boldsymbol{U} \in \mathcal{W}$ (Orabona, 2023, Theorem 2.13):

$$\sum_{t=1}^{T} (S_t(\boldsymbol{W}_t) - S_t(\boldsymbol{U})) \leq \frac{\|\boldsymbol{U}\|_{\mathrm{F}}^2}{2\eta} + \frac{\eta}{2} \sum_{t=1}^{T} \|\nabla S_t(\boldsymbol{W}_t)\|_{\mathrm{F}}^2. \tag{1}$$

The next proposition highlights how to obtain a template of finite surrogate regret bounds by learning $\boldsymbol{W}_t$ with this OGD and exploiting the surrogate gap.

Proposition 6 *Assume that there exist constants $a \in (0, 1)$ and $b > 0$ satisfying the following conditions for $t = 1, \dots, T$: (i) $\mathbb{E}_t[L_t(\widehat{\boldsymbol{y}}_t)] \leq (1 - a) S_t(\boldsymbol{W}_t)$ and (ii) $\|\nabla S_t(\boldsymbol{W}_t)\|_{\mathrm{F}}^2 \leq b S_t(\boldsymbol{W}_t)$.*[14] *Let* Alg *be OGD with learning rate $\eta = \frac{2}{b} \min\{\frac{1}{2}, a\}$. Then, it holds that*

$$\sum_{t=1}^{T} \mathbb{E}_t[L_t(\widehat{\boldsymbol{y}}_t)] \leq \sum_{t=1}^{T} S_t(\boldsymbol{U}) + \underbrace{\frac{(1-a) b \|\boldsymbol{U}\|_{\mathrm{F}}^2}{4(1 - \min\{\frac{1}{2}, a\}) \min\{\frac{1}{2}, a\}}}_{\text{(A)}}.$$

14. Surrogate losses satisfying this inequality are said to be *regular* in Van der Hoeven et al. (2021).

Proof By substituting (ii) into (1), we have $\sum_{t=1}^{T}(S_t(\boldsymbol{W}_t) - S_t(\boldsymbol{U})) \le \frac{\|\boldsymbol{U}\|_{\mathrm{F}}^2}{2\eta} + \frac{\eta b}{2}\sum_{t=1}^{T} S_t(\boldsymbol{W}_t)$. Noting $\frac{\eta b}{2} < 1$, we rearrange the terms to obtain a so-called $L^\star$ *bound* (Orabona, 2023, Section 4.2.3), whose right-hand side depends on $\sum_{t=1}^{T} S_t(\boldsymbol{U})$ as follows:

$$\sum_{t=1}^{T}(S_t(\boldsymbol{W}_t) - S_t(\boldsymbol{U})) \le \left(1 - \frac{\eta b}{2}\right)^{-1}\left(\frac{\|\boldsymbol{U}\|_{\mathrm{F}}^2}{2\eta} + \frac{\eta b}{2}\sum_{t=1}^{T} S_t(\boldsymbol{U})\right).$$

Combining this with (i) implies that the surrogate regret, $\sum_{t=1}^{T}\mathbb{E}_t[L_t(\widehat{\boldsymbol{y}}_t)] - \sum_{t=1}^{T} S_t(\boldsymbol{U})$, is at most

$$\begin{aligned}
&(1-a)\sum_{t=1}^{T}(S_t(\boldsymbol{W}_t) - S_t(\boldsymbol{U})) - a\sum_{t=1}^{T} S_t(\boldsymbol{U}) \\
&\le (1-a)\left(1 - \frac{\eta b}{2}\right)^{-1}\frac{\|\boldsymbol{U}\|_{\mathrm{F}}^2}{2\eta} - \left(a - (1-a)\left(1 - \frac{\eta b}{2}\right)^{-1}\frac{\eta b}{2}\right)\sum_{t=1}^{T} S_t(\boldsymbol{U}).
\end{aligned}$$

Since $\frac{\eta b}{2} \le a$ implies $a - (1-a)\left(1 - \frac{\eta b}{2}\right)^{-1}\frac{\eta b}{2} \ge 0$, ignoring the second term does not decrease the right-hand side. Substituting $\eta = \frac{2}{b}\min\left\{\frac{1}{2}, a\right\}$ into the first term yields the desired bound. ∎

The last part in the proof highlights the fundamental idea for achieving a finite surrogate regret bound: offsetting the increase in the regret of OGD, which originates from $\frac{\eta}{2}\sum_{t=1}^{T}\|\nabla S_t(\boldsymbol{W}_t)\|_{\mathrm{F}}^2$ in (1), with the cumulative surrogate gap, $a\sum_{t=1}^{T} S_t(\boldsymbol{U})$, by setting η to a sufficiently small value. This is based on the original idea of exploiting the surrogate gap by Van der Hoeven (2020). The crux of this fundamental idea lies in conditions (i) and (ii) in Proposition 6, which we will verify by using Lemma 4 and the properties of the Fenchel–Young loss in Proposition 3, respectively. Consequently, we obtain the following finite surrogate regret bound in expectation for online structured prediction with Fenchel–Young losses, which is the main result of this paper.

Theorem 7 *Let ψ_Ω be the randomized decoding given in Algorithm 1 and $C > 0$ a constant with $\max_{t\in[T]}\|\boldsymbol{x}_t\|_2 \le C$. If $\lambda > \frac{4\gamma}{\nu}$ holds and* Alg *is OGD with learning rate $\eta = \frac{\lambda}{C^2\kappa^2}\min\left\{\frac{1}{2}, 1 - \frac{4\gamma}{\lambda\nu}\right\}$, for any $\boldsymbol{U} \in \mathcal{W}$, it holds that*

$$\sum_{t=1}^{T}\mathbb{E}_t[L_t(\widehat{\boldsymbol{y}}_t)] \le \sum_{t=1}^{T} S_t(\boldsymbol{U}) + \frac{2\gamma C^2\kappa^2\|\boldsymbol{U}\|_{\mathrm{F}}^2}{\lambda^2\nu\left(1 - \min\left\{\frac{1}{2}, 1 - \frac{4\gamma}{\lambda\nu}\right\}\right)\min\left\{\frac{1}{2}, 1 - \frac{4\gamma}{\lambda\nu}\right\}}.$$

Proof Since Lemma 4 implies $\mathbb{E}_t[L_t(\widehat{\boldsymbol{y}}_t)] \le \frac{4\gamma}{\lambda\nu}S_t(\boldsymbol{W}_t)$, condition (i) in Proposition 6 holds with $a = 1 - \frac{4\gamma}{\lambda\nu} \in (0,1)$. Furthermore, Proposition 3 implies $\nabla S_t(\boldsymbol{W}_t) = (\widehat{\boldsymbol{y}}_\Omega(\boldsymbol{\theta}_t) - \boldsymbol{y}_t)\boldsymbol{x}_t^\top$ and $\|\widehat{\boldsymbol{y}}_\Omega(\boldsymbol{\theta}_t) - \boldsymbol{y}_t\|^2 \le \frac{2}{\lambda}S_t(\boldsymbol{W}_t)$; combining them with $\|\boldsymbol{y}\boldsymbol{x}^\top\|_{\mathrm{F}}^2 = \mathrm{tr}(\boldsymbol{x}^\top\boldsymbol{x}\boldsymbol{y}^\top\boldsymbol{y}) = \|\boldsymbol{x}\|_2^2\|\boldsymbol{y}\|_2^2$, $\|\boldsymbol{x}_t\|_2 \le C$, and $\|\cdot\|_2 \le \kappa\|\cdot\|$ yields

$$\|\nabla S_t(\boldsymbol{W}_t)\|_{\mathrm{F}}^2 = \|\widehat{\boldsymbol{y}}_\Omega(\boldsymbol{\theta}_t) - \boldsymbol{y}_t\|_2^2\|\boldsymbol{x}_t\|_2^2 \le C^2\kappa^2\|\widehat{\boldsymbol{y}}_\Omega(\boldsymbol{\theta}_t) - \boldsymbol{y}_t\|^2 \le \frac{2C^2\kappa^2}{\lambda}S_t(\boldsymbol{W}_t).$$

Thus, condition (ii) with $b = \frac{2C^2\kappa^2}{\lambda}$ holds. Therefore, Proposition 6 provides the desired bound. ∎

It is also worth mentioning that the above OGD is *parameter-free* in the sense that the learning rate η is tuned without any knowledge of $\boldsymbol{U}$ or the size of domain $\mathcal{W}$ (cf. McMahan and Streeter (2012b), Orabona (2013), and Cutkosky and Orabona (2018)).[15] However, the constant learning rate may result in poor empirical performance, particularly when $a = 1 - \frac{4\gamma}{\lambda\nu}$ is very small. Appendix E.2 shows that we can alternatively use a parameter-free algorithm of Cutkosky and Orabona (2018) to achieve a finite surrogate regret bound.

High-probability bound. Similar to Van der Hoeven et al. (2021), we can obtain a finite surrogate regret bound that holds with high probability. Define random variables $Z_t \coloneqq L_t(\widehat{\boldsymbol{y}}_t) - \mathbb{E}_t[L_t(\widehat{\boldsymbol{y}}_t)]$ for $t = 1, \dots, T$, where the randomness comes from the randomized decoding. A crucial step is to ensure that the cumulative deviation $\sum_{t=1}^{T} Z_t$ grows only at the rate of $\sqrt{\sum_{t=1}^{T} S_t(\boldsymbol{U})}$, in which Lemma 4 again turns out to be helpful. Once it is shown, we can obtain a high-probability bound by offsetting the regret of OGD, plus $\sum_{t=1}^{T} Z_t$, with $a\sum_{t=1}^{T} S_t(\boldsymbol{U})$. See Appendix F for the proof.

Theorem 8 *Assume the same condition as Theorem 7 except for the learning rate of OGD, which we here set as $\eta = \frac{a}{b}$. Let D be the diameter of $\mathrm{conv}(\mathcal{Y})$ in terms of $\|\cdot\|$ and $\delta \in (0, 1)$. Then, with probability at least $1 - \delta$, for any $\boldsymbol{U} \in \mathcal{W}$, it holds that*

$$\sum_{t=1}^{T} L_t(\widehat{\boldsymbol{y}}_t) \le \sum_{t=1}^{T} S_t(\boldsymbol{U}) + \left(1 - \frac{4\gamma}{\lambda\nu}\right)^{-1} \left(\frac{8\gamma C^2\kappa^2\|\boldsymbol{U}\|_{\mathrm{F}}^2}{\lambda^2\nu} + \gamma D \ln\frac{1}{\delta}\right).$$

Remark 9 (The case of adaptive adversary) *Theorem 7 remains true even against an adaptive adversary since Lemma 4 ensures that condition (i) in Proposition 6 holds for any adaptive sequence $(\boldsymbol{x}_1, \boldsymbol{y}_1), \dots, (\boldsymbol{x}_T, \boldsymbol{y}_T)$ and the regret bound of OGD in (1) applies to the adaptive case as well. The high-probability bound in Theorem 8 also remains valid in the adaptive case since an additional concentration argument used in the proof is irrelevant to the adversary's type.*

Remark 10 (Asymptotic behavior when $a \to 1$) *The surrogate regret bound (A) in Proposition 6 simplifies to $\frac{b\|\boldsymbol{U}\|_{\mathrm{F}}^2}{4a}$ if $a \le 1/2$ and to $(1-a)b\|\boldsymbol{U}\|_{\mathrm{F}}^2$ if $a > 1/2$, where the latter expression is smaller when $a > 1/2$. Notably, the bound vanishes when $a \to 1$. This property has not been observed in previous studies (Van der Hoeven, 2020; Van der Hoeven et al., 2021), and we have obtained this by taking advantage of the $L^\star$ bound. Note that (i) in Proposition 6 implies that $\mathbb{E}_t[L_t(\widehat{\boldsymbol{y}}_t)]/S_t(\boldsymbol{W}_t)$ goes to zero when $a \to 1$. Therefore, this asymptotic behavior reflects a rationale that the surrogate regret bound should vanish when the target loss scales down relative to the surrogate loss. As in the proof of Theorem 7, $1 - a$ and b are proportional to $1/\lambda$, and hence the surrogate regret bound in Theorem 7 vanishes at the rate of $1/\lambda^2$. The high-probability surrogate regret bound in Theorem 8 also decreases at the rate of $1/\lambda^2$, while the $\gamma D \ln\frac{1}{\delta}$ term persists as it comes from the randomness of the decoding. Here, we can increase or decrease λ by scaling up or down the regularization function Ω generating the Fenchel–Young loss (see also Section 4.1), although increasing λ generally leads to larger $S_t(\boldsymbol{U})$.*

4.1. Application to Specific Problems

Theorems 7 and 8 provide finite surrogate regret bounds for various online structured prediction problems that satisfy Assumption 1 if $\lambda > \frac{4\gamma}{\nu}$ (or $a = 1 - \frac{4\gamma}{\lambda\nu} > 0$) holds, which requires Ω

15. The line of work on parameter-free learning achieves regret bounds that depend almost linearly on the comparator's norm via non-trivial techniques. Compared to this, achieving the surrogate regret bound depending on $\|\boldsymbol{U}\|_{\mathrm{F}}^2$ is easier.

generating the Fenchel–Young loss to be sufficiently strongly convex. Notably, this requirement is automatically satisfied in the case of multiclass classification with the logistic loss (see Section 5). In the multilabel classification example in Section 2.3, we have $\nu = 1$, $\gamma = \frac{1}{\sqrt{d}}$, and $\lambda = 1$; therefore, $\lambda > \frac{4\gamma}{\nu}$ holds if $d > 16$.[16] In general, we can take advantage of the Fenchel–Young loss framework to satisfy $\lambda > \frac{4\gamma}{\nu}$: since we may use any function $\Omega = \Psi + I_{\mathrm{conv}(\mathcal{Y})}$ to generate a Fenchel–Young loss, we can scale up Ψ to satisfy $\lambda > \frac{4\gamma}{\nu}$ if necessary. In the ranking example in Section 2.3, we have $\nu = 4$, $\gamma = \frac{1}{2n}$, and $\lambda = \frac{1}{n\mu}$, where $\mu > 0$ controls the scale of $\Psi = -\frac{1}{\mu}\mathsf{H}^{\mathrm{s}}$. Thus, $\lambda > \frac{4\gamma}{\nu}$ holds if $\mu < 2$. Note that the dependence of λ on γ is inevitable because the surrogate loss encodes no information about the target loss per se. Nonetheless, we can scale λ to satisfy $\lambda > \frac{4\gamma}{\nu}$ as ν is often lower bounded: $\nu \ge 1$ holds if $\mathcal{Y} \subseteq \mathbb{Z}^d$ and $\|\cdot\|$ is an ℓ_p-norm, and $\nu \ge 2^{1/p}$ if $\boldsymbol{y}^\top \mathbf{1}$ is constant.

4.2. Online-to-Batch Conversion

We discuss converting surrogate regret bounds to guarantees for offline structured prediction. In general, surrogate regret bounds may not admit online-to-batch conversion because we cannot apply Jensen's inequality to non-convex target loss. In our case, Lemma 4, which bounds target loss by convex surrogate loss, enables us to sidestep this issue, leading to the following result.

Theorem 11 *Assume the same condition as Theorem 7. If $(\boldsymbol{x}_1, \boldsymbol{y}_1), \dots, (\boldsymbol{x}_T, \boldsymbol{y}_T)$ are drawn i.i.d. from an underlying joint distribution on $\mathcal{X}$ and $\mathcal{Y}$, for any $\boldsymbol{U} \in \mathcal{W}$, it holds that*

$$\mathbb{E}[L(\psi_\Omega(\overline{\boldsymbol{W}}\boldsymbol{x}); \boldsymbol{y})] \le \mathbb{E}[S_\Omega(\boldsymbol{U}\boldsymbol{x}; \boldsymbol{y})] + \frac{1}{T} \cdot \frac{2\gamma C^2 \kappa^2 \|\boldsymbol{U}\|_{\mathrm{F}}^2}{\lambda^2 \nu \left(1 - \min\left\{\frac{1}{2}, 1 - \frac{4\gamma}{\lambda\nu}\right\}\right) \min\left\{\frac{1}{2}, 1 - \frac{4\gamma}{\lambda\nu}\right\}},$$

where $\overline{\boldsymbol{W}} = \frac{1}{T}\sum_{t=1}^T \boldsymbol{W}_t$ is the average of outputs of Alg.

The above bound differs from common excess risk bounds studied in statistical learning, and hence comparing them directly is difficult. However, since (super) fast convergence for structured prediction has been already established by Cabannes et al. (2021), we do not think the bound in Theorem 11 itself is of particular significance. Nonetheless, Theorem 11 offers a better understanding of the relationship between online learning guarantees based on the surrogate regret and statistical learning theory with margin conditions. We present the proof of Theorem 11 in Appendix G.1 and discuss a connection to excess risk bounds in Appendix G.2.

5. Improved Surrogate Regret for Online Multiclass Classification with Logistic Loss

We present an $O(\|\boldsymbol{U}\|_{\mathrm{F}}^2)$ surrogate regret bound for online multiclass classification with the logistic loss by using our general result for structured prediction, thereby improving the $O(d\|\boldsymbol{U}\|_{\mathrm{F}}^2)$ bound of Van der Hoeven (2020). In this section, we let $\mathcal{Y} = \{\mathbf{e}_1, \dots, \mathbf{e}_d\}$ and $\|\cdot\| = \|\cdot\|_1$. We have $\kappa = 1$ since $\|\cdot\|_1 \ge \|\cdot\|_2$. The target loss is the 0-1 loss, $L(\boldsymbol{y}'; \mathbf{e}_i) = \mathbb{1}_{\boldsymbol{y}' \neq \mathbf{e}_i}$. Note that we have $\nu = 2$ and $\gamma = 1/2$, as explained in Section 2.3. We use the same logistic loss as that used by Van der Hoeven (2020). Specifically, for any $\boldsymbol{\theta} \in \mathbb{R}^d$ and $\mathbf{e}_i \in \mathcal{Y}$, we define the logistic loss as

$$S_{\mathrm{logistic}}(\boldsymbol{\theta}; \mathbf{e}_i) := -\log_2 \sigma_i(\boldsymbol{\theta}),$$

16. Note that the target loss is scaled to $[0, 1]$. If not scaled, we need to scale up Ψ to satisfy $\lambda > 4\sqrt{d}$.

where $\sigma_i(\boldsymbol{\theta}) := \frac{\exp(\theta_i)}{\sum_{j=1}^{d}\exp(\theta_j)}$ is the softmax function. This logistic loss is expressed as a Fenchel–Young loss up to a constant factor. For any $\boldsymbol{y} \in \triangle^d$, let Ω be an entropic regularizer given by

$$\Omega(\boldsymbol{y}) = -\mathsf{H}^{\mathrm{s}}(\boldsymbol{y}) + I_{\triangle^d}(\boldsymbol{y}). \tag{2}$$

The Fenchel–Young loss generated by this Ω is $S_\Omega(\boldsymbol{\theta}; \mathrm{e}_i) = -\ln \sigma_i(\boldsymbol{\theta})$ (see Blondel et al. (2020)), hence $S_{\mathrm{logistic}}(\boldsymbol{\theta}; \mathrm{e}_i) = \frac{1}{\ln 2} S_\Omega(\boldsymbol{\theta}; \mathrm{e}_i)$. Moreover, $\widehat{\boldsymbol{y}}_\Omega(\boldsymbol{\theta})$ equals $(\sigma_1(\boldsymbol{\theta}), \dots, \sigma_d(\boldsymbol{\theta}))^\top$, which we can efficiently compute in the randomized decoding (Algorithm 1) without iterative optimization methods.

By applying Algorithm 2 to the above setting, we obtain the following surrogate regret bound.

Theorem 12 *Let ψ_Ω be the randomized decoding given in Algorithm 1 with the entropic regularizer Ω in (2) and $C > 0$ a constant such that $\max_{t\in[T]}\|\boldsymbol{x}_t\|_2 \le C$. If we apply OGD with $\eta = \frac{(1-\ln 2)\ln 2}{C^2}$ to loss functions $S_t(\boldsymbol{W}) = S_{\mathrm{logistic}}(\boldsymbol{W}\boldsymbol{x}_t; \boldsymbol{y}_t)$ ($t = 1, \dots, T$), for any $\boldsymbol{U} \in \mathcal{W}$, Algorithm 2 achieves*

$$\sum_{t=1}^{T} \mathbb{E}_t[\mathbb{1}_{\widehat{\boldsymbol{y}}_t \neq \boldsymbol{y}_t}] \le \sum_{t=1}^{T} S_t(\boldsymbol{U}) + \frac{C^2\|\boldsymbol{U}\|_{\mathrm{F}}^2}{2(1-\ln 2)\ln 2}.$$

Proof The proof resembles that of Theorem 7. From Pinsker's inequality, $-\mathsf{H}^{\mathrm{s}}$ is 1-strongly convex w.r.t. $\|\cdot\|_1$ over $\triangle^d$ (i.e., $\lambda = 1$). Thus, $\mathbb{E}[L(\psi_\Omega(\boldsymbol{\theta}); \boldsymbol{y})] \le \frac{4\gamma}{\lambda\nu} S_\Omega(\boldsymbol{\theta}; \boldsymbol{y}) = \frac{4\cdot 1/2}{1\cdot 2} \ln 2 \cdot S_{\mathrm{logistic}}(\boldsymbol{\theta}; \boldsymbol{y})$ holds due to Lemma 4, leaving a surrogate gap with $a = 1 - \ln 2 \in (0, 1/2)$ in Proposition 6. We also have $\|\nabla S_t(\boldsymbol{W}_t)\|_{\mathrm{F}}^2 \le \frac{2C^2}{\ln 2} S_t(\boldsymbol{W}_t)$ due to Van der Hoeven (2020, Lemma 2), i.e., $b = \frac{2C^2}{\ln 2}$. By setting $\eta = \frac{2a}{b} = \frac{(1-\ln 2)\ln 2}{C^2}$, Proposition 6 implies the desired bound of $\frac{b\|\boldsymbol{U}\|_{\mathrm{F}}^2}{4a} = \frac{C^2\|\boldsymbol{U}\|_{\mathrm{F}}^2}{2(1-\ln 2)\ln 2}$. ∎

Difference from Van der Hoeven (2020) and Van der Hoeven et al. (2021). The main technical difference from the previous studies lies in how to decode $\boldsymbol{\theta} \in \mathbb{R}^d$ to $\boldsymbol{y} \in \mathcal{Y}$. Specifically, when $\boldsymbol{\theta}$ is not confident about any of d classes, their methods increase the likelihood of choosing a class uniformly at random (i.e., uniform exploration with probability $\frac{1}{d}$), which yields a surrogate gap with $a = \frac{1}{d}$ in Proposition 6 (see Van der Hoeven et al. (2021, Lemma 1)), resulting in the extra d factor. Our randomized decoding instead returns random $\widetilde{\boldsymbol{y}} \in \mathcal{Y}$ with $\mathbb{E}[\widetilde{\boldsymbol{y}} \mid Z = 1] = \widehat{\boldsymbol{y}}_\Omega(\boldsymbol{\theta})$, which exploits $\boldsymbol{\theta}$ more aggressively and yields a surrogate gap with $a = 1 - \ln 2$, thus achieving the improved bound of $O(\|\boldsymbol{U}\|_{\mathrm{F}}^2)$.[17] Apart from this, the two decoding procedures have distinct pros and cons: uniform exploration is often extensive for structured spaces,[18] while our randomized decoding enjoys efficient implementations given linear optimization oracles on $\mathcal{Y}$, as discussed in Section 3.1. On the other hand, uniform exploration is applicable to the bandit setting and works with the (smooth) hinge loss. Investigating how to apply our randomized decoding to the bandit setting and how the resulting bound compares with the state-of-the-art (Van der Hoeven, 2020; Agarwal et al., 2022) will be an interesting future direction, whereas applying the randomized decoding to the (smooth) hinge loss seems difficult, as discussed in Footnote 9.

Finally, we give a lower bound showing that Theorem 12 is asymptotically tight up to $\ln d$ factors if $\|\boldsymbol{U}\|_{\mathrm{F}} = \Theta(B)$ holds for the ℓ_2-diameter B of the domain $\mathcal{W}$. The proof, deferred to Appendix H, is inspired by that of Van der Hoeven et al. (2021) for obtaining an $\Omega(dB^2)$ lower bound for the smooth hinge loss, which we modify to deal with the logistic loss.

17. The dependence on C is identical in our bound and that of Van der Hoeven (2020).
18. For example, if $\mathcal{Y}$ consists of perfect matchings of a (possibly incomplete) bipartite graph with n vertices, the current fastest *fully polynomial almost uniform sampler* takes $O(n^7 \ln n)$ time (Jerrum et al., 2004; Bezáková et al., 2008), which is known to be impractical (Newman and Vardi, 2020).

Theorem 13 *Let $d \geq 2$. For $B = \Omega(\ln(dT))$, there exists a sequence $(\boldsymbol{x}_1, \boldsymbol{y}_1), \ldots, (\boldsymbol{x}_T, \boldsymbol{y}_T)$ such that $\|\boldsymbol{x}_t\|_2 = 1$ for $t = 1, \ldots, T$ and any possibly randomized algorithm incurs an $\Omega(B^2/\ln^2 d)$ surrogate regret with respect to the logistic surrogate loss.*

Logistic vs. smooth hinge. As mentioned in Section 1, larger surrogate loss makes it easier to bound the surrogate regret. Thus, considering the (nearly) tight bounds of $O(\|\boldsymbol{U}\|_{\mathrm{F}}^2)$ and $O(d\|\boldsymbol{U}\|_{\mathrm{F}}^2)$ for the logistic and smooth hinge losses, respectively, one may think that the logistic loss is always larger than the smooth hinge loss. However, this is not the case. For example, consider a binary classification setting with $\mathcal{Y} = \{\mathbf{e}_1, \mathbf{e}_2\}$, where $\mathbf{e}_1$ is the ground truth. If an estimator $\boldsymbol{U}$ yields $(\theta_1, \theta_2)^\top = \boldsymbol{U}\boldsymbol{x}_t$ with $\theta_1 - \theta_2 = 0.3$, the logistic and smooth hinge losses take $\log_2(1 + \exp(\theta_2 - \theta_1)) \approx 0.8$ and $\max\{1 - (\theta_1 - \theta_2)^2, 0\} \approx 0.9$, respectively, implying that the logistic loss is not always larger than the smooth hinge loss.

Acknowledgments

We thank anonymous COLT reviewers, who provided truly constructive feedback. In particular, their review comments helped us improve the surrogate regret bounds so that they vanish when $a \to 1$ and brought our attention to the parameter-freeness of the original method by Van der Hoeven (2020). This work was supported by JST ERATO Grant Number JPMJER1903, JST ACT-X Grant Number JPMJAX210E, and JSPS KAKENHI Grant Number JP22K17853.

References

Naman Agarwal, Satyen Kale, and Julian Zimmert. Efficient methods for online multiclass logistic regression. In *Proceedings of the 33rd International Conference on Algorithmic Learning Theory*, volume 167, pages 3–33. PMLR, 2022. Cited on pages 3 and 13.

Gökhan BakIr, Thomas Hofmann, Bernhard Schölkopf, Alexander J. Smola, Ben Taskar, and S. V. N. Vishwanathan. *Predicting Structured Data*. The MIT Press, 2007. Cited on page 1.

Peter L. Bartlett, Michael Collins, Ben Taskar, and David McAllester. Exponentiated gradient algorithms for large-margin structured classification. In *Advances in Neural Information Processing Systems*, volume 17, pages 113–120. Curran Associates, Inc., 2004. Cited on page 19.

Peter L. Bartlett, Michael I. Jordan, and Jon D. McAuliffe. Convexity, classification, and risk bounds. *Journal of the American Statistical Association*, 101(473):138–156, 2006. Cited on page 1.

Quentin Berthet, Mathieu Blondel, Olivier Teboul, Marco Cuturi, Jean-Philippe Vert, and Francis Bach. Learning with differentiable pertubed optimizers. In *Advances in Neural Information Processing Systems*, volume 33, pages 9508–9519. Curran Associates, Inc., 2020. Cited on page 21.

Alina Beygelzimer, David Pal, Balazs Szorenyi, Devanathan Thiruvenkatachari, Chen-Yu Wei, and Chicheng Zhang. Bandit multiclass linear classification: Efficient algorithms for the separable case. In *Proceedings of the 36th International Conference on Machine Learning*, volume 97, pages 624–633. PMLR, 2019. Cited on page 3.

Ivona Bezáková, Daniel Štefankovič, Vijay V. Vazirani, and Eric Vigoda. Accelerating simulated annealing for the permanent and combinatorial counting problems. *SIAM Journal on Computing*, 37(5):1429–1454, 2008. Cited on page 13.

Mathieu Blondel. Structured prediction with projection oracles. In *Advances in Neural Information Processing Systems*, volume 32, pages 12167–12178. Curran Associates, Inc., 2019. Cited on pages 3, 4, 5, 6, 20, and 21.

Mathieu Blondel, André F. T. Martins, and Vlad Niculae. Learning with Fenchel–Young losses. *Journal of Machine Learning Research*, 21(35):1–69, 2020. Cited on pages 1, 4, 5, 8, 13, 19, 20, and 28.

Vivien A. Cabannes, Alessandro Rudi, and Francis Bach. Structured prediction with partial labelling through the infimum loss. In *Proceedings of the 37th International Conference on Machine Learning*, pages 1230–1239. PMLR, 2020. Cited on page 19.

Vivien A. Cabannes, Francis Bach, and Alessandro Rudi. Fast rates for structured prediction. In *Proceedings of the 34th Conference on Learning Theory*, volume 134, pages 823–865. PMLR, 2021. Cited on pages 12, 19, 28, and 29.

Nicolò Cesa-Bianchi and Gábor Lugosi. *Prediction, Learning, and Games*. Cambridge University Press, 2006. Cited on page 26.

Carlo Ciliberto, Lorenzo Rosasco, and Alessandro Rudi. A consistent regularization approach for structured prediction. In *Advances in Neural Information Processing Systems*, volume 29, pages 4412–4420. Curran Associates, Inc., 2016. Cited on pages 3 and 19.

Carlo Ciliberto, Lorenzo Rosasco, and Alessandro Rudi. A general framework for consistent structured prediction with implicit loss embeddings. *Journal of Machine Learning Research*, 21 (98):1–67, 2020. Cited on pages 3 and 19.

Cyrille W. Combettes and Sebastian Pokutta. Revisiting the approximate Carathéodory problem via the Frank–Wolfe algorithm. *Mathematical Programming*, 197(1):191–214, 2023. Cited on page 9.

Koby Crammer and Yoram Singer. On the algorithmic implementation of multiclass kernel-based vector machines. *Journal of Machine Learning Research*, 2:265–292, 2001. Cited on page 5.

Ashok Cutkosky and Francesco Orabona. Black-box reductions for parameter-free online learning in Banach spaces. In *Proceedings of the 31st Conference on Learning Theory*, volume 75, pages 1493–1529. PMLR, 2018. Cited on pages 11 and 23.

Marco Cuturi. Sinkhorn distances: Lightspeed computation of optimal transport. In *Advances in Neural Information Processing Systems*, volume 26, pages 2292–2300. Curran Associates, Inc., 2013. Cited on page 6.

John Duchi, Elad Hazan, and Yoram Singer. Adaptive subgradient methods for online learning and stochastic optimization. *Journal of Machine Learning Research*, 12(61):2121–2159, 2011. Cited on page 22.

Adam N. Elmachtoub and Paul Grigas. Smart "predict, then optimize". *Management Science*, 68(1): 9–26, 2022. Cited on page 21.

Michael Fink, Shai Shalev-Shwartz, Yoram Singer, and Shimon Ullman. Online multiclass learning by interclass hypothesis sharing. In *Proceedings of the 23rd International Conference on Machine Learning*, pages 313–320. ACM, 2006. Cited on page 2.

Jessica Finocchiaro, Rafael Frongillo, and Bo Waggoner. An embedding framework for consistent polyhedral surrogates. In *Advances in Neural Information Processing Systems*, volume 32, pages 10781–10791. Curran Associates, Inc., 2019. Cited on page 19.

Dylan J. Foster, Satyen Kale, Haipeng Luo, Mehryar Mohri, and Karthik Sridharan. Logistic regression: The importance of being improper. In *Proceedings of the 31st Conference on Learning Theory*, volume 75, pages 167–208. PMLR, 2018. Cited on page 3.

Dan Garber and Noam Wolf. Frank–Wolfe with a nearest extreme point oracle. In *Proceedings of the 34th Conference on Learning Theory*, volume 134, pages 2103–2132. PMLR, 2021. Cited on pages 8, 9, and 21.

Elad Hazan, Amit Agarwal, and Satyen Kale. Logarithmic regret algorithms for online convex optimization. *Machine Learning*, 69(2):169–192, 2007. Cited on page 3.

Elad Hazan, Tomer Koren, and Kfir Y. Levy. Logistic regression: Tight bounds for stochastic and online optimization. In *Proceedings of the 27th Conference on Learning Theory*, volume 35, pages 197–209. PMLR, 2014. Cited on page 3.

Yichun Hu, Nathan Kallus, and Xiaojie Mao. Fast rates for contextual linear optimization. *Management science*, 68(6):4236–4245, 2022. Cited on page 21.

Mark Jerrum, Alistair Sinclair, and Eric Vigoda. A polynomial-time approximation algorithm for the permanent of a matrix with nonnegative entries. *Journal of the ACM*, 51(4):671–697, 2004. Cited on page 13.

Rémi Jézéquel, Pierre Gaillard, and Alessandro Rudi. Mixability made efficient: Fast online multiclass logistic regression. In *Advances in Neural Information Processing Systems*, volume 34, pages 23692–23702. Curran Associates, Inc., 2021. Cited on page 3.

Sham M. Kakade, Shai Shalev-Shwartz, and Ambuj Tewari. Efficient bandit algorithms for online multiclass prediction. In *Proceedings of the 25th International Conference on Machine Learning*, pages 440–447. ACM, 2008. Cited on page 2.

Terry Koo, Amir Globerson, Xavier Carreras Pérez, and Michael Collins. Structured prediction models via the matrix-tree theorem. In *Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, pages 141–150. ACL, 2007. Cited on page 19.

John Lafferty, Andrew McCallum, and Fernando C. N. Pereira. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In *Proceedings of the 18th International Conference on Machine Learning*, pages 282–289. Morgan Kaufmann Publishers Inc., 2001. Cited on page 19.

Shaojie Li and Yong Liu. Towards sharper generalization bounds for structured prediction. In *Advances in Neural Information Processing Systems*, volume 34, pages 26844–26857. Curran Associates, Inc., 2021. Cited on page 19.

André F. T. Martins, Noah Smith, Eric Xing, Pedro Aguiar, and Mario Figueiredo. Online learning of structured predictors with multiple kernels. In *Proceedings of the 14th International Conference on Artificial Intelligence and Statistics*, volume 15, pages 507–515. PMLR, 2011. Cited on page 19.

H. Brendan McMahan and Matthew Streeter. Adaptive bound optimization for online convex optimization. In *Proceedings of the 23rd Conference on Learning Theory*, pages 244–256. Omnipress, 2010. Cited on page 22.

H. Brendan McMahan and Matthew Streeter. Open problem: Better bounds for online logistic regression. In *Proceedings of the 25th Annual Conference on Learning Theory*, volume 23, pages 44.1–44.3. PMLR, 2012a. Cited on page 3.

H. Brendan McMahan and Matthew Streeter. No-regret algorithms for unconstrained online convex optimization. In *Advances in Neural Information Processing Systems*, volume 25, pages 2402–2410. Curran Associates, Inc., 2012b. Cited on pages 11, 23, and 24.

Gergely Neu and Nikita Zhivotovskiy. Fast rates for online prediction with abstention. In *Proceedings of the 33rd Conference on Learning Theory*, volume 125, pages 3030–3048. PMLR, 2020. Cited on page 2.

James E. Newman and Moshe Y. Vardi. FPRAS approximation of the matrix permanent in practice. *arXiv:2012.03367*, 2020. Cited on page 13.

Vlad Niculae, André F. T. Martins, Mathieu Blondel, and Claire Cardie. SparseMAP: Differentiable sparse structured inference. In *Proceedings of the 35th International Conference on Machine Learning*, volume 80, pages 3799–3808. PMLR, 2018. Cited on pages 3, 19, and 20.

Alex Nowak-Vila, Francis Bach, and Alessandro Rudi. Sharp analysis of learning with discrete losses. In *Proceedings of the 22nd International Conference on Artificial Intelligence and Statistics*, pages 1920–1929. PMLR, 2019. Cited on page 19.

Alex Nowak-Vila, Francis Bach, and Alessandro Rudi. Consistent structured prediction with max-min margin Markov networks. In *Proceedings of the 37th International Conference on Machine Learning*, pages 7381–7391. PMLR, 2020. Cited on page 19.

Alex Nowak-Vila, Alessandro Rudi, and Francis Bach. On the consistency of max-margin losses. In *Proceedings of the 25th International Conference on Artificial Intelligence and Statistics*, pages 4612–4633. PMLR, 2022. Cited on page 19.

Francesco Orabona. Dimension-free exponentiated gradient. In *Advances in Neural Information Processing Systems*, volume 26, pages 1806–1814. Curran Associates, Inc., 2013. Cited on pages 11, 23, and 24.

Francesco Orabona. A modern introduction to online learning. *arXiv:1912.13213*, 2023. https://arxiv.org/abs/1912.13213v6. Cited on pages 9, 10, 22, and 27.

Anton Osokin, Francis Bach, and Simon Lacoste-Julien. On structured prediction theory with calibrated convex surrogate losses. In *Advances in Neural Information Processing Systems*, volume 30, pages 302–313. Curran Associates, Inc., 2017. Cited on page 19.

Venkata Krishna Pillutla, Vincent Roulet, Sham M. Kakade, and Zaid Harchaoui. A smoother way to train structured prediction models. In *Advances in Neural Information Processing Systems*, volume 31, pages 4766–4778. Curran Associates, Inc., 2018. Cited on page 19.

Ralph Tyrell Rockafellar. *Convex Analysis*. Princeton University Press, 1970. Cited on page 4.

Frank Rosenblatt. The perceptron: a probabilistic model for information storage and organization in the brain. *Psychological Review*, 65(6):386–408, 1958. Cited on page 3.

Anish Thilagar, Rafael Frongillo, Jessica J. Finocchiaro, and Emma Goodwill. Consistent polyhedral surrogates for top-k classification and variants. In *Proceedings of the 39th International Conference on Machine Learning*, pages 21329–21359. PMLR, 2022. Cited on page 19.

Ioannis Tsochantaridis, Thorsten Joachims, Thomas Hofmann, Yasemin Altun, and Yoram Singer. Large margin methods for structured and interdependent output variables. *Journal of Machine Learning Research*, 6(50):1453–1484, 2005. Cited on page 19.

Dirk van der Hoeven. Exploiting the surrogate gap in online multiclass classification. In *Advances in Neural Information Processing Systems*, volume 33, pages 9562–9572. Curran Associates, Inc., 2020. Cited on pages 1, 2, 3, 7, 8, 9, 10, 11, 12, 13, and 14.

Dirk van der Hoeven, Federico Fusco, and Nicolò Cesa-Bianchi. Beyond bandit feedback in online multiclass classification. In *Advances in Neural Information Processing Systems*, volume 34, pages 13280–13291. Curran Associates, Inc., 2021. Cited on pages 2, 3, 9, 11, and 13.

Yutong Wang and Clayton Scott. Weston-Watkins hinge loss and ordered partitions. In *Advances in Neural Information Processing Systems*, volume 33, pages 19873–19883. Curran Associates, Inc., 2020. Cited on page 19.

Appendix A. Literature Review of Structured Prediction

In this literature review, the term "(surrogate) regret" refers to the (surrogate) excess risk.

Earlier studies investigated statistical inference on conditional random fields (Lafferty et al., 2001), max-margin models (Bartlett et al., 2004), and spanning trees (Koo et al., 2007). Tsochantaridis et al. (2005) is a seminal work to provide a general framework based on loss functions for structured prediction problems by extending support vector machines. Independent of the development of SparseMAP (Niculae et al., 2018), Pillutla et al. (2018) proposed a tractable algorithm to optimize the structured hinge loss by introducing a smoothed decoder.

With the SELF framework (Ciliberto et al., 2016, 2020), regret bounds depend on the spectral norm of $\boldsymbol{V}$, which may be exponential in the natural dimension of the output space. To address this issue, Osokin et al. (2017, Appendix E) improved the surrogate regret bound of the quadratic loss for specific target losses such as the block 0-1 and Hamming losses, showing that the dependency of the surrogate regret on the matrix norm can be lifted. Nowak-Vila et al. (2019, Theorem 3.1) systematically extended this result to many multilabel and ranking target losses by obtaining low-rank decomposition of $\boldsymbol{V}$ in SELF for those losses. Moreover, Nowak-Vila et al. (2020) established the surrogate regret bounds beyond the quadratic loss, for max-margin surrogate losses generated by the Fenchel–Young losses; Nowak-Vila et al. (2022) studied necessary conditions of a structured target loss for max-margin losses to be Fisher consistent. Cabannes et al. (2020) elucidated that partial label learning, a type of learning problem with ambiguous structured inputs, can be cast into the framework of the regularized least-square decoder and established the surrogate regret bound for the target loss called the infimum loss, encompassed into SELF. Recently, Cabannes et al. (2021) showed fast convergence rates for the excess risk of SELF under a generalized Tsybakov margin condition; in particular, their result implies exponential convergence in the number of samples under the hard margin condition. Li and Liu (2021) studied generalization bounds for surrogate losses, which imply a fast convergence rate when surrogate losses are smooth.

Apart from the aforementioned studies, another stream of research has studied the consistency of structured prediction problems via *polyhedral losses*. Finocchiaro et al. (2019) and Thilagar et al. (2022) studied the consistency of classification with abstention, top-k prediction, and Lovász hinge loss; Wang and Scott (2020) studied the consistent target loss of the Weston–Watkins hinge loss.

The literature is scarce when it comes to the online setting (except for online multiclass classification discussed in Section 1.1). Martins et al. (2011) studied an online learning approach to structured prediction with multiple kernels based on standard regret bounds for convex surrogate losses.

Appendix B. Note on CRF Loss

We confirm that the *Conditional Random Field (CRF)* loss (Lafferty et al., 2001) satisfies Proposition 3. Consequently, we can treat it as a specific Fenchel-Young loss to obtain the results presented in the main text, similar to the logistic and SparseMAP losses. Although the following discussion is elementary, we include it for completeness.

Below, $\mathcal{Y}$ is equipped with some total order and the components of any $|\mathcal{Y}|$-dimensional vector are arranged in the same order. As detailed in Blondel et al. (2020, Section 7.1), the CRF loss is a Fenchel–Young loss generated by $\Omega(\boldsymbol{y}) = \min\{ -\mathsf{H}^{\mathrm{s}}(\boldsymbol{p}) : \boldsymbol{p} \in \triangle^{|\mathcal{Y}|}, \mathbb{E}_{Y\sim\boldsymbol{p}}[Y] = \boldsymbol{y} \}$. For any $\boldsymbol{\theta} \in \mathbb{R}^d$ and $\boldsymbol{y} \in \mathcal{Y}$, the resulting CRF loss is expressed as $S_\Omega(\boldsymbol{\theta}; \boldsymbol{y}) = -\ln(\exp(\langle \boldsymbol{\theta}, \boldsymbol{y} \rangle)/Z(\boldsymbol{\theta}))$, where $Z(\boldsymbol{\theta}) \coloneqq \sum_{\boldsymbol{y}\in\mathcal{Y}} \exp(\langle \boldsymbol{\theta}, \boldsymbol{y} \rangle)$ denotes the partition function. The regularized prediction is uniquely determined by the marginal inference: $\widehat{\boldsymbol{y}}_\Omega(\boldsymbol{\theta}) = \sum_{\boldsymbol{y}\in\mathcal{Y}} (\exp(\langle \boldsymbol{\theta}, \boldsymbol{y} \rangle)/Z(\boldsymbol{\theta}))\boldsymbol{y}$. Thus, the gradient,

$\nabla S_\Omega(\boldsymbol{\theta}; \boldsymbol{y}) = \widehat{\boldsymbol{y}}_\Omega(\boldsymbol{\theta}) - \boldsymbol{y}$, is also unique. In what follows, we prove $S_\Omega(\boldsymbol{\theta}; \boldsymbol{y}) \geq \frac{\lambda}{2}\|\widehat{\boldsymbol{y}}_\Omega(\boldsymbol{\theta}) - \boldsymbol{y}\|_1^2$ for $\lambda := 1/\max\{\, \|\boldsymbol{y}\|_1^2 : \boldsymbol{y} \in \mathcal{Y} \,\}$ to establish Proposition 3. If $\mathcal{Y} \subseteq \{0,1\}^d$, this λ is at least $1/d^2$.

First, let us observe that for any $\boldsymbol{\theta} \in \mathbb{R}^d$, the CRF loss can be seen as a logistic loss defined on $\mathbb{R}^{|\mathcal{Y}|}$ by associating each $\boldsymbol{y} \in \mathcal{Y}$ with a score $\langle \boldsymbol{\theta}, \boldsymbol{y} \rangle \in \mathbb{R}$. Specifically, let $\boldsymbol{s}(\boldsymbol{\theta}) \in \mathbb{R}^{|\mathcal{Y}|}$ denote a vector whose component corresponding to $\boldsymbol{y} \in \mathcal{Y}$, denoted by $s_{\boldsymbol{y}}(\boldsymbol{\theta})$, equals $\langle \boldsymbol{\theta}, \boldsymbol{y} \rangle$. Then, we have $S_\Omega(\boldsymbol{\theta}; \boldsymbol{y}) = -\ln(\exp(s_{\boldsymbol{y}}(\boldsymbol{\theta}))/\sum_{\boldsymbol{y}' \in \mathcal{Y}} \exp(s_{\boldsymbol{y}'}(\boldsymbol{\theta})))$. By regarding this as the logistic loss of $\boldsymbol{s}(\boldsymbol{\theta})$, the Pinsker's inequality (or Blondel (2019, Proposition 2)) implies

$$S_\Omega(\boldsymbol{\theta}; \boldsymbol{y}) \geq \frac{1}{2}\|\sigma(\boldsymbol{s}(\boldsymbol{\theta})) - \mathbf{e}_{\boldsymbol{y}}\|_1^2, \tag{3}$$

where $\sigma : \mathbb{R}^{|\mathcal{Y}|} \to \triangle^{|\mathcal{Y}|}$ is the softmax function, i.e., $\sigma_{\boldsymbol{y}}(\boldsymbol{s}(\boldsymbol{\theta})) = \exp(s_{\boldsymbol{y}}(\boldsymbol{\theta}))/\sum_{\boldsymbol{y}' \in \mathcal{Y}} \exp(s_{\boldsymbol{y}'}(\boldsymbol{\theta}))$, and $\mathbf{e}_{\boldsymbol{y}} \in \mathbb{R}^{|\mathcal{Y}|}$ is the standard basis vector that has a single one at the component corresponding to $\boldsymbol{y}$.

Next, we show $\|\sigma(\boldsymbol{s}(\boldsymbol{\theta})) - \mathbf{e}_{\boldsymbol{y}}\|_1^2 \geq \lambda\|\widehat{\boldsymbol{y}}_\Omega(\boldsymbol{\theta}) - \boldsymbol{y}\|_1^2$, which combined with (3) yields the desired inequality. To see this, let A be the $d \times |\mathcal{Y}|$ matrix with column vectors $\boldsymbol{y} \in \mathcal{Y}$ aligned horizontally in the same total order. We have $\widehat{\boldsymbol{y}}_\Omega(\boldsymbol{\theta}) = A\sigma(\boldsymbol{s}(\boldsymbol{\theta}))$ and $\boldsymbol{y} = A\mathbf{e}_{\boldsymbol{y}}$, and hence

$$\|\widehat{\boldsymbol{y}}_\Omega(\boldsymbol{\theta}) - \boldsymbol{y}\|_1 = \|A(\sigma(\boldsymbol{s}(\boldsymbol{\theta})) - \mathbf{e}_{\boldsymbol{y}})\|_1 \leq \|A\|_1\|\sigma(\boldsymbol{s}(\boldsymbol{\theta})) - \mathbf{e}_{\boldsymbol{y}}\|_1.$$

Here, $\|A\|_1$ is the operator norm of A between the ℓ_1-normed spaces, i.e., the maximum ℓ_1-norm of the columns of A. Thus, we have $\|A\|_1 = 1/\sqrt{\lambda}$, and hence $\|\sigma(\boldsymbol{s}(\boldsymbol{\theta})) - \mathbf{e}_{\boldsymbol{y}}\|_1^2 \geq \lambda\|\widehat{\boldsymbol{y}}_\Omega(\boldsymbol{\theta}) - \boldsymbol{y}\|_1^2$.

We remark that the marginal inference for computing $\widehat{\boldsymbol{y}}_\Omega(\boldsymbol{\theta})$ is sometimes intractable, which has motivated the development of SparseMAP (Niculae et al., 2018). We refer the reader to Blondel et al. (2020, Section 7.3) for a discussion on the computational complexity.

Appendix C. Additional Applications

We discuss additional applications considered in Blondel (2019). For simplicity, we below let $\|\cdot\|$ be the ℓ_2-norm and S_Ω the SparseMAP loss generated by $\Omega(\boldsymbol{y}) = \frac{1}{2}\|\boldsymbol{y}\|_2^2 + I_{\mathrm{conv}(\mathcal{Y})}(\boldsymbol{y})$, as in the multilabel classification example in Section 2.3; hence, we have $\lambda = 1$. We will confirm the condition of $\lambda > \frac{4\gamma}{\nu}$ in Theorem 7 and discuss the implementation of randomized decoding (Algorithm 1).

C.1. Ranking with Permutahedron

We consider another ranking scenario with different $\mathcal{Y}$. Let $\mathcal{Y} \subseteq \mathbb{Z}^d$ be the set of all points obtained by permuting $(d, d-1, \dots, 1)^\top \in \mathbb{Z}^d$. Then, $\mathrm{conv}(\mathcal{Y})$ is the so-called *permutahedron*. Since $\boldsymbol{y}^\top \mathbf{1}$ is constant for all $\boldsymbol{y} \in \mathcal{Y}$ and $\|\cdot\|$ is the ℓ_2-norm, we have $\nu = \sqrt{2}$. We consider measuring how predicted $\boldsymbol{y}' \in \mathrm{conv}(\mathcal{Y})$ is aligned with the true $\boldsymbol{y} \in \mathcal{Y}$ by $\langle \boldsymbol{y}, \boldsymbol{y} - \boldsymbol{y}' \rangle$, which takes 0 if $\boldsymbol{y}' = \boldsymbol{y}$ and $M := \frac{d(d^2-1)}{6}$ for the least aligned $\boldsymbol{y}'$. Based on this idea, we use a target loss defined by $L(\boldsymbol{y}'; \boldsymbol{y}) = \frac{1}{M}\langle \boldsymbol{y}, \boldsymbol{y} - \boldsymbol{y}' \rangle \in [0,1]$, which is affine in $\boldsymbol{y}'$ and satisfies $L(\boldsymbol{y}'; \boldsymbol{y}) \leq \frac{1}{M}\|\boldsymbol{y}\|_2\|\boldsymbol{y} - \boldsymbol{y}'\|_2 = \gamma\|\boldsymbol{y} - \boldsymbol{y}'\|_2$ for $\gamma = \frac{1}{d-1}\sqrt{\frac{2d+1}{6d(d+1)}}$. Therefore, $\lambda > \frac{4\gamma}{\nu}$ holds for $d \geq 3$.

We provide an efficient implementation of randomized decoding (Algorithm 1) for this setting. To this end, we show that a similar claim to Proposition 5 holds, though $\mathrm{conv}(\mathcal{Y})$ is not a 0-1 polytope.

Proposition 14 *Let $\boldsymbol{y}' \in \mathrm{conv}(\mathcal{Y})$ and π be a permutation on $[d]$ such that $y'_{\pi(1)} \leq \dots \leq y'_{\pi(d)}$. It holds that $\boldsymbol{y}^* := (\pi^{-1}(1), \dots, \pi^{-1}(d))^\top \in \arg\min\{\, \|\boldsymbol{y} - \boldsymbol{y}'\|_p : \boldsymbol{y} \in \mathcal{Y} \,\}$, where $p \in [1, +\infty]$.*

Proof Let $p < \infty$; the case of $p = \infty$ follows by taking the limit $p \to \infty$ in what follows. Suppose that there exists a nearest point $\boldsymbol{y} \in \mathcal{Y}$ to $\boldsymbol{y}'$ such that $y'_i > y'_j$ and $y_i < y_j$ for some $i, j \in [d]$, which implies $y_i - y'_i < y_j - y'_j$. Since $y_i - y'_j$ and $y_j - y'_i$ are contained in $[y_i - y'_i, y_j - y'_j]$, we have $|y_i - y'_i|^p + |y_j - y'_j|^p \geq |y_i - y'_j|^p + |y_j - y'_i|^p$ by the convexity of $x \mapsto |x|^p$. Thus, letting $\breve{\boldsymbol{y}} \in \mathcal{Y}$ be a point obtained from $\boldsymbol{y}$ by swapping the ith and jth components, we obtain

$$\begin{aligned}\|\boldsymbol{y} - \boldsymbol{y}'\|_p^p &= |y_i - y'_i|^p + |y_j - y'_j|^p + \sum_{k \neq i,j} |y_k - y'_k|^p \\ &\geq |\breve{y}_i - y'_i|^p + |\breve{y}_j - y'_j|^p + \sum_{k \neq i,j} |\breve{y}_k - y'_k|^p = \|\breve{\boldsymbol{y}} - \boldsymbol{y}'\|_p^p,\end{aligned}$$

meaning that $\breve{\boldsymbol{y}}$ is also a nearest point to $\boldsymbol{y}'$. Therefore, there must exist a nearest point such that its components have the same order as those of $\boldsymbol{y}'$, and any such points have the same distance to $\boldsymbol{y}'$. Consequently, $\boldsymbol{y}^*$ is a nearest point to $\boldsymbol{y}'$ with respect to $\|\cdot\|_p$ among all $\boldsymbol{y} \in \mathcal{Y}$. ■

Proposition 14 means that we can compute a nearest extreme point $\boldsymbol{y}^* \in \mathcal{Y}$ to any given $\boldsymbol{y}' \in \operatorname{conv}(\mathcal{Y})$ by sorting the components of $\boldsymbol{y}'$ in $O(d \ln d)$ time. Armed with this, we can use the fast Frank–Wolfe algorithm of Garber and Wolf (2021) to compute $\widehat{\boldsymbol{y}}_\Omega(\boldsymbol{\theta})$ and a convex combination of extreme points for sampling $\widetilde{\boldsymbol{y}}$ with $\mathbb{E}[\widetilde{\boldsymbol{y}} \mid Z = 1] = \widehat{\boldsymbol{y}}_\Omega(\boldsymbol{\theta})$, as described in Section 3.1.

C.2. Ordinal Regression

We consider ordinal regression with binary outputs satisfying $y_1 \geq \cdots \geq y_d$. The output space is $\mathcal{Y} = \{\mathbf{0}, \mathbf{e}_1, \mathbf{e}_1 + \mathbf{e}_2, \ldots, \mathbf{e}_1 + \cdots + \mathbf{e}_d\}$. Since $\|\boldsymbol{y} - \boldsymbol{y}'\|_2 \geq 1$ for any distinct $\boldsymbol{y}, \boldsymbol{y}' \in \mathcal{Y}$, $\nu = 1$ holds. The target loss considered in Blondel (2019) is the absolute loss $\|\boldsymbol{y}' - \boldsymbol{y}\|_1$. Scaling it to $[0, 1]$, we can express it on $\operatorname{conv}(\mathcal{Y})$ for any $\boldsymbol{y} \in \mathcal{Y}$ in the same way as the Hamming loss in Section 2.3: $L(\boldsymbol{y}'; \boldsymbol{y}) = \frac{1}{d}(\langle \boldsymbol{y}', \mathbf{1} \rangle + \langle \boldsymbol{y}, \mathbf{1} \rangle - 2\langle \boldsymbol{y}', \boldsymbol{y} \rangle)$, which is affine in $\boldsymbol{y}'$ and upper bounded by $\frac{1}{\sqrt{d}}\|\boldsymbol{y}' - \boldsymbol{y}\|_2$, hence $\gamma = \frac{1}{\sqrt{d}}$. Thus, $\lambda > \frac{4\gamma}{\nu}$ holds for $d > 16$, as with the case of multilabel classification.

Since $\mathcal{Y} \subseteq \{0, 1\}^d$ holds in this case, an efficient implementation of randomized decoding is available, as discussed in Section 3.1.

Other potential applications. Due to the generality of our randomized decoding, we expect that our method is potentially useful for learning tasks involving general linear optimization problems (Elmachtoub and Grigas, 2022; Hu et al., 2022). Application of Fenchel–Young losses to similar tasks is studied in Berthet et al. (2020). We leave the investigation of this direction as future work.

Appendix D. Lower Bound on Constant Factor in Lemma 4

Lemma 4 ensures that for any $(\boldsymbol{\theta}, \boldsymbol{y}) \in \mathbb{R}^d \times \mathcal{Y}$, the randomized decoding ψ_Ω (Algorithm 1) achieves $\mathbb{E}[L(\psi_\Omega(\boldsymbol{\theta}); \boldsymbol{y})] \leq c \cdot \frac{\gamma}{\lambda\nu} S_\Omega(\boldsymbol{\theta}; \boldsymbol{y})$ for $c = 4$. We show that the multiplicative constant c cannot be smaller than $2/\ln 2 \approx 2.89$ in general, suggesting Lemma 4 is nearly tight. To see this, we again use the binary classification example in Section 3. Let $\mathcal{Y} = \{\mathbf{e}_1, \mathbf{e}_2\}$, $\boldsymbol{y} = \mathbf{e}_1$ be the ground truth, and $\boldsymbol{\theta} = (\theta_1, \theta_2) = (1, 1 + \ln(2^{1+\varepsilon} - 1))$ for sufficiently small $\varepsilon > 0$, which slightly favors $\mathbf{e}_2$ by mistake. As explained in Section 2.3, if $\|\cdot\|$ is the ℓ_1-norm, we have $\gamma = 1/2$ and $\nu = 2$. Also, as detailed in Section 5, if Ω is the entropic regularizer given in (2), we have $\lambda = 1$, and the regularized prediction is given by the softmax function, i.e., $\widehat{\boldsymbol{y}}_\Omega(\boldsymbol{\theta}) = (1, \exp(\theta_2 - \theta_1))/(1 + \exp(\theta_2 - \theta_1))$. The resulting Fenchel–Young loss is the base-e logistic loss, hence $S_\Omega(\boldsymbol{\theta}; \mathbf{e}_1) = \ln(1 + \exp(\theta_2 - \theta_1))$. Substituting the above $\boldsymbol{\theta}$ into $\widehat{\boldsymbol{y}}_\Omega(\boldsymbol{\theta})$ and $S_\Omega(\boldsymbol{\theta}; \mathbf{e}_1)$, we obtain

$$\widehat{\boldsymbol{y}}_\Omega(\boldsymbol{\theta}) = \left(\frac{1}{2^{1+\varepsilon}}, 1 - \frac{1}{2^{1+\varepsilon}}\right) \quad \text{and} \quad S_\Omega(\boldsymbol{\theta}; \mathbf{e}_1) = (1+\varepsilon)\ln 2.$$

We then calculate the expected 0-1 loss. Since the closest point to $\widehat{\boldsymbol{y}}_\Omega(\boldsymbol{\theta})$ in $\mathcal{Y}$ is $\boldsymbol{y}^* = \mathbf{e}_2$, it holds that $\Delta^* = \|\mathbf{e}_2 - \widehat{\boldsymbol{y}}_\Omega(\boldsymbol{\theta})\|_1 = 1/2^\varepsilon$, and hence $p = \min\{1, 2\Delta^*/\nu\} = 1/2^\varepsilon$ for $\varepsilon > 0$. Recall that the randomized decoding returns $\boldsymbol{y}^* = \mathbf{e}_2$ with probability $1-p$, or a random $\widetilde{\boldsymbol{y}}$ with probability p. Since $\widetilde{\boldsymbol{y}}$ is drawn from $\mathcal{Y}$ to satisfy $\mathbb{E}[\widetilde{\boldsymbol{y}} \mid Z = 1] = \widehat{\boldsymbol{y}}_\Omega(\boldsymbol{\theta})$, we have $\widetilde{\boldsymbol{y}} = \mathbf{e}_1$ with probability $1/2^{1+\varepsilon}$ and $\mathbf{e}_2$ with probability $1 - 1/2^{1+\varepsilon}$. Therefore, the expected 0-1 loss is

$$\begin{aligned}\mathbb{E}[L(\psi_\Omega(\boldsymbol{\theta}); \boldsymbol{y})] &= (1-p)L(\mathbf{e}_2; \mathbf{e}_1) + p\,\mathbb{E}[L(\widetilde{\boldsymbol{y}}; \mathbf{e}_1) \mid Z = 1] \\ &= \left(1 - \frac{1}{2^\varepsilon}\right)\cdot 1 + \frac{1}{2^\varepsilon}\cdot\left(\frac{1}{2^{1+\varepsilon}}\cdot 0 + \left(1 - \frac{1}{2^{1+\varepsilon}}\right)\cdot 1\right) = 1 - \frac{1}{2^{1+2\varepsilon}}.\end{aligned}$$

To ensure that Lemma 4 holds in the above setting, we need

$$c \geq \frac{\lambda\nu}{\gamma}\cdot\frac{\mathbb{E}[L(\psi_\Omega(\boldsymbol{\theta}); \boldsymbol{y})]}{S_\Omega(\boldsymbol{\theta}; \boldsymbol{y})} = 4\cdot\frac{1 - \frac{1}{2^{1+2\varepsilon}}}{(1+\varepsilon)\ln 2}.$$

The right-hand side converges to $2/\ln 2$ as $\varepsilon \to +0$, implying $c \geq 2/\ln 2 \approx 2.89$.

Appendix E. Bounding Surrogate Regret with Other OCO Algorithms

While we have obtained finite surrogate regret bounds using OGD with the constant learning rate, it may not perform well in practice since it does not utilize information from past rounds. Below, we demonstrate that we can achieve finite surrogate regret bounds with more practical OCO algorithms. Specifically, Appendices E.1 and E.2 discuss using OGD with an adaptive learning rate and a parameter-free OCO algorithm, respectively. In terms of theoretical guarantees, however, the results presented below are weaker than those obtained in the main text by using OGD with the constant learning rate: the result in Appendix E.1 is not parameter-free, and the surrogate regret bound in Appendix E.2 is asymptotically larger by a logarithmic factor and does not vanish when $a \to 1$.

E.1. OGD with Adaptive Learning Rate

Let $a = 1 - \frac{4\gamma}{\lambda\nu}$ and $b = \frac{2C^2\kappa^2}{\lambda}$, as in the proof of Theorem 7. Assume that a domain $\mathcal{W}$ with an ℓ_2-diameter of $B > 0$ is given. We consider using the online gradient descent on $\mathcal{W}$ with an adaptive learning rate as Alg, where the learning rate in the tth round is set to $B/\sqrt{2\sum_{i=1}^T \|\nabla S_i(\boldsymbol{W}_i)\|_\mathrm{F}^2}$ (McMahan and Streeter, 2010; Duchi et al., 2011). Due to $\|\nabla S_t(\boldsymbol{W}_t)\|_\mathrm{F}^2 \leq bS_t(\boldsymbol{W}_t)$, this OGD achieves the following $L^\star$ bound for any $\boldsymbol{U} \in \mathcal{W}$ (Orabona, 2023, Theorem 4.25):

$$\sum_{t=1}^T (S_t(\boldsymbol{W}_t) - S_t(\boldsymbol{U})) \leq 2bB^2 + 2B\sqrt{2b\sum_{t=1}^T S_t(\boldsymbol{U})}. \tag{4}$$

If we learn $\boldsymbol{W}_t$ with this OGD, we can achieve the following expected surrogate regret bound.

Theorem 15 *Let ψ_Ω be the randomized decoding given in Algorithm 1 and $C > 0$ a constant such that $\max_{t\in[T]}\|\boldsymbol{x}_t\|_2 \le C$. If* Alg *satisfies* (4) *and $\lambda > \frac{4\gamma}{\nu}$ holds, for any $\boldsymbol{U} \in \mathcal{W}$, it holds that*

$$\sum_{t=1}^{T} \mathbb{E}_t[L_t(\widehat{\boldsymbol{y}}_t)] \le \sum_{t=1}^{T} S_t(\boldsymbol{U}) + \left(1 - \frac{4\gamma}{\lambda\nu}\right)^{-1} \frac{16\gamma C^2\kappa^2 B^2}{\lambda^2\nu}$$

Proof Since we have $\mathbb{E}_t[L_t(\widehat{\boldsymbol{y}}_t)] \le (1-a)S_t(\boldsymbol{W}_t)$ due to Lemma 4, it holds that

$$\sum_{t=1}^{T} \mathbb{E}_t[L_t(\widehat{\boldsymbol{y}}_t)] - \sum_{t=1}^{T} S_t(\boldsymbol{U}) \le (1-a)\sum_{t=1}^{T}(S_t(\boldsymbol{W}_t) - S_t(\boldsymbol{U})) - a\sum_{t=1}^{T} S_t(\boldsymbol{U}).$$

Substituting (4) into the right-hand side implies that the surrogate regret is at most

$$\begin{aligned} 2(1-a)bB^2 + 2(1-a)B\sqrt{2b\sum_{t=1}^{T} S_t(\boldsymbol{U})} - a\sum_{t=1}^{T} S_t(\boldsymbol{U}) &\le 2(1-a)bB^2 + \frac{2(1-a)^2bB^2}{a} \\ &= \frac{2(1-a)bB^2}{a}, \end{aligned}$$

where the inequality comes from $\sqrt{c_1 x} - c_2 x \le \frac{c_1}{4c_2}$ $(\forall x > 0)$ for $c_1, c_2 > 0$. The desired surrogate regret bound follows from $a = 1 - \frac{4\gamma}{\lambda\nu}$ and $b = \frac{2C^2\kappa^2}{\lambda}$. ■

Similar to the bound in Proposition 6, it vanishes when $a \to 1$. However, as described above, tuning the learning rate requires the knowledge of the domain size, B, hence no longer parameter-free.

E.2. Parameter-Free Algorithm

In the previous section, we have assumed that the ℓ_2-diameter B of the domain $\mathcal{W}$ is known a priori. In practice, however, we rarely know the precise size of $\mathcal{W}$ containing the best estimator $\boldsymbol{U}$ in hindsight. A common workaround is to set B to a sufficiently large value, but this typically results in overly pessimistic regret bounds. *Parameter-free* algorithms (McMahan and Streeter, 2012b; Orabona, 2013; Cutkosky and Orabona, 2018) are designed to avoid this issue by automatically adapting to the norm of $\boldsymbol{U}$. While OGD with the constant learning rate is also parameter-free in our case, parameter-free algorithms studied in this context would be more practical.

We consider using a parameter-free algorithm of Cutkosky and Orabona (2018, Section 3) as Alg instead of OGD. Their algorithm enjoys a regret bound depending on $g_{1:T} \coloneqq \sum_{t=1}^{T}\|\nabla S_t(\boldsymbol{W}_t)\|_{\mathrm{F}}^2$, which is helpful in the subsequent analysis. Specifically, the bound on $\sum_{t=1}^{T}(S_t(\boldsymbol{W}_t) - S_t(\boldsymbol{U}))$ is

$$O\left(\|\boldsymbol{U}\|_{\mathrm{F}} \max\left\{\sqrt{g_{1:T}\ln\left(\frac{\|\boldsymbol{U}\|_{\mathrm{F}}^2 g_{1:T}}{\epsilon^2} + 1\right)}, \ln\frac{\|\boldsymbol{U}\|_{\mathrm{F}} g_{1:T}}{\epsilon}\right\} + \|\boldsymbol{U}\|_{\mathrm{F}}\sqrt{g_{1:T}} + \epsilon\right),$$

where $\epsilon > 0$ is an *initial-wealth* parameter specified by the user. We let $\epsilon = 1$ for simplicity. Then, omitting lower-order terms, the regret bound of Alg reduces to

$$O\left(\|\boldsymbol{U}\|_{\mathrm{F}}\sqrt{g_{1:T}\ln\left(\|\boldsymbol{U}\|_{\mathrm{F}}^2 g_{1:T} + 1\right)}\right). \tag{5}$$

Compared with the standard regret bound of OGD, the dependence on $g_{1:T}$ is worse by a factor of $\sqrt{\ln g_{1:T}}$, which is inevitable in parameter-free learning (McMahan and Streeter, 2012b; Orabona, 2013). This makes the analysis for exploiting the surrogate gap a bit trickier. The next lemma offers a helpful inequality for exploiting the surrogate gap with the parameter-free regret bound (5).

Lemma 16 *Let $a, b, c > 0$. For any $x \geq 0$, it holds that*

$$-ax + \sqrt{bx \ln(cx+1)} \leq \frac{b}{2a}\left(\ln\left(\frac{b}{2a^2}+1\right) + \ln(c+1)\right) + a.$$

Proof Let $f(x) = -ax + \sqrt{bx \ln(cx+1)}$. Since f is continuous and goes to $-\infty$ as $x \to \infty$, it suffices to show that f is bounded as in the lemma statement at any critical point x^*. The derivative of f is $f'(x) = -a + \frac{\sqrt{b}}{2}\frac{\frac{cx}{cx+1}+\ln(cx+1)}{\sqrt{x\ln(cx+1)}}$, and hence any critical point x^* satisfies $\sqrt{bx^* \ln(cx^*+1)} = \frac{b}{2a}\left(\frac{cx^*}{cx^*+1} + \ln(cx^*+1)\right)$. Thus, we have

$$f(x^*) = -ax^* + \frac{b}{2a}\left(\frac{cx^*}{cx^*+1} + \ln(cx^*+1)\right) \leq -ax^* + \frac{b}{2a} + \frac{b}{2a}\ln(cx^*+1). \tag{6}$$

If $c \leq 1$, the right-hand side of (6) is at most $-ax^* + \frac{b}{2a} + \frac{b}{2a}\ln(x^*+1)$, and inspecting its derivative with respect to x^* readily shows that the following inequality holds for any $x^* \geq 0$:

$$-ax^* + \frac{b}{2a} + \frac{b}{2a}\ln(x^*+1) \leq \frac{b}{2a}\max\left\{\ln\left(\frac{b}{2a^2}\right), 0\right\} + a. \tag{7}$$

If $c > 1$, the right-hand side of (6) is at most $-ax^* + \frac{b}{2a} + \frac{b}{2a}\ln(x^*+1) + \frac{b}{2a}\ln c$, and the sum of the first three terms is bounded as in (7). Thus, $f(x^*) \leq \frac{b}{2a}\max\{\ln(\frac{b}{2a^2}), 0\} + \frac{b}{2a}\max\{\ln(c), 0\} + a$ holds in any case. By using $\max\{\ln z, 0\} \leq \ln(z+1)$ for $z > 0$, we obtain the desired bound. ■

Now, we are ready to obtain a finite surrogate regret bound with the parameter-free algorithm.

Theorem 17 *Suppose ψ_Ω and C to be given as in Theorem 7. Fix any $\boldsymbol{U} \in \mathcal{W}$, where $\mathcal{W}$ is some domain. Let* Alg *be the parameter-free algorithm that achieves the regret bound given in (5) without knowing the size of $\mathcal{W}$ (or $\|\boldsymbol{U}\|_{\mathrm{F}}$). If $\lambda > \frac{4\gamma}{\nu}$, it holds that*

$$\sum_{t=1}^{T} \mathbb{E}_t[L_t(\widehat{\boldsymbol{y}}_t)] = \sum_{t=1}^{T} S_t(\boldsymbol{U}) + O\left(\frac{\|\boldsymbol{U}\|_{\mathrm{F}}^2 C^2\kappa^2}{\lambda - 4\gamma/\nu}\ln\left(\frac{\|\boldsymbol{U}\|_{\mathrm{F}}^2 C^2\kappa^2\lambda}{(\lambda - 4\gamma/\nu)^2} + 1\right)\right).$$

Proof Let $a = 1 - \frac{4\gamma}{\lambda\nu} \in (0,1)$ and $x = \sum_{t=1}^{T} S_t(\boldsymbol{W}_t)$. As in the proof of Theorem 7, we have

$$\sum_{t=1}^{T}(\mathbb{E}_t[L_t(\widehat{\boldsymbol{y}}_t)] - S_t(\boldsymbol{W}_t)) \leq -ax \tag{8}$$

and

$$g_{1:T} = \sum_{t=1}^{T}\|\nabla S_t(\boldsymbol{W}_t)\|_{\mathrm{F}}^2 \leq \frac{2C^2\kappa^2}{\lambda}x. \tag{9}$$

Substituting (9) into the regret bound (5) and using (8), we can bound the surrogate regret as follows:

$$\begin{aligned}\sum_{t=1}^{T}\mathbb{E}_t[L_t(\widehat{\boldsymbol{y}}_t)] - \sum_{t=1}^{T} S_t(\boldsymbol{U}) &= \sum_{t=1}^{T}(\mathbb{E}_t[L_t(\widehat{\boldsymbol{y}}_t)] - S_t(\boldsymbol{W}_t)) + \sum_{t=1}^{T}(S_t(\boldsymbol{W}_t) - S_t(\boldsymbol{U})) \\ &\leq -ax + c_{\mathsf{Alg}}\|\boldsymbol{U}\|_{\mathrm{F}}\sqrt{\frac{2C^2\kappa^2}{\lambda}x\ \ln\left(\|\boldsymbol{U}\|_{\mathrm{F}}^2\frac{2C^2\kappa^2}{\lambda}x + 1\right)},\end{aligned}$$

where $c_{\mathsf{Alg}} > 0$ is an absolute constant hidden in the big-O notation in (5). By using Lemma 16 with $a = 1 - \frac{4\gamma}{\lambda\nu}$, $b = c_{\mathsf{Alg}}^2\|\boldsymbol{U}\|_{\mathrm{F}}^2\frac{2C^2\kappa^2}{\lambda}$, and $c = \|\boldsymbol{U}\|_{\mathrm{F}}^2\frac{2C^2\kappa^2}{\lambda}$, we can upper bound the surrogate regret by

$$\begin{aligned}&\frac{b}{2a}\left(\ln\left(\frac{b}{2a^2}+1\right) + \ln(c+1)\right) + a \\ &= O\left(\frac{\|\boldsymbol{U}\|_{\mathrm{F}}^2 C^2\kappa^2}{\lambda - \frac{4\gamma}{\nu}}\left(\ln\left(\frac{\|\boldsymbol{U}\|_{\mathrm{F}}^2 C^2\kappa^2}{\lambda\left(1-\frac{4\gamma}{\lambda\nu}\right)^2}+1\right) + \ln\left(\frac{\|\boldsymbol{U}\|_{\mathrm{F}}^2 C^2\kappa^2}{\lambda}+1\right)\right)\right) \\ &\lesssim O\left(\frac{\|\boldsymbol{U}\|_{\mathrm{F}}^2 C^2\kappa^2}{\lambda - \frac{4\gamma}{\nu}}\ln\left(\frac{\|\boldsymbol{U}\|_{\mathrm{F}}^2 C^2\kappa^2\lambda}{\left(\lambda-\frac{4\gamma}{\nu}\right)^2}+1\right)\right),\end{aligned}$$

where the asymptotic inequality is due to $1 - \frac{4\gamma}{\lambda\nu} < 1$. Thus, we obtain the desired bound. ■

Note that the bound in Theorem 17 depends on $\|\boldsymbol{U}\|_{\mathrm{F}}$ and is free from B, in contrast to the bound in Theorem 15. However, the bound does not vanish when $a = 1 - \frac{4\gamma}{\lambda\nu}$ approaches 1.

Appendix F. Proof of Theorem 8

We prove the high-probability surrogate regret bound. Recall that Z_t for $t = 1, \dots, T$ are random variables defined by $Z_t \coloneqq L_t(\widehat{\boldsymbol{y}}_t) - \mathbb{E}_t[L_t(\widehat{\boldsymbol{y}}_t)]$.
Proof Let $a = 1 - \frac{4\gamma}{\lambda\nu} \in (0,1)$ and $b = \frac{2C^2\kappa^2}{\lambda} > 0$, as in the proof of Theorem 7. We decompose the surrogate regret, $\sum_{t=1}^{T} L_t(\widehat{\boldsymbol{y}}_t) - \sum_{t=1}^{T} S_t(\boldsymbol{U})$, as follows:

$$\sum_{t=1}^{T}(L_t(\widehat{\boldsymbol{y}}_t) - \mathbb{E}_t[L_t(\widehat{\boldsymbol{y}}_t)]) + \sum_{t=1}^{T}(\mathbb{E}_t[L_t(\widehat{\boldsymbol{y}}_t)] - S_t(\boldsymbol{U})). \tag{10}$$

Let $\eta' \coloneqq \frac{\eta b}{2} = \frac{a}{2} < 1$. As in the proof of Proposition 6, the second term in (10) is at most

$$\begin{aligned}&(1-a)\left(1-\frac{\eta b}{2}\right)^{-1}\frac{\|\boldsymbol{U}\|_{\mathrm{F}}^2}{2\eta} - \left(a - (1-a)\left(1-\frac{\eta b}{2}\right)^{-1}\frac{\eta b}{2}\right)\sum_{t=1}^{T} S_t(\boldsymbol{U}) \\ &= \frac{1-a}{1-\eta'}\cdot\frac{b\|\boldsymbol{U}\|_{\mathrm{F}}^2}{4\eta'} - \left(a - \frac{1-a}{1-\eta'}\cdot\eta'\right)\sum_{t=1}^{T} S_t(\boldsymbol{U}).\end{aligned} \tag{11}$$

Below, we derive an upper bound on the first term in (10). Note that the first term equals $\sum_{t=1}^{T} Z_t$. Since Assumption 1 implies $0 \leq L_t(\boldsymbol{y}) \leq \gamma D$ for any $\boldsymbol{y} \in \mathcal{Y}$, we have $|Z_t| \leq \gamma D$, hence $\mathbb{E}_t[Z_t^2] \leq$

$\mathbb{E}_t\left[L_t(\widehat{\boldsymbol{y}}_t)^2\right] \le \gamma D\, \mathbb{E}_t[L_t(\widehat{\boldsymbol{y}}_t)]$. Combining this with Lemma 4 yields $\mathbb{E}_t[Z_t^2] \le \frac{4\gamma^2 D}{\lambda\nu} S_t(\boldsymbol{W}_t) = (1-a)\gamma D S_t(\boldsymbol{W}_t)$. By applying Bernstein's inequality (see, e.g., Cesa-Bianchi and Lugosi (2006, Lemma A.8)) to the bounded martingale difference sequence $Z_1, \dots, Z_T$ with bounded variance, for any $\delta \in (0,1)$, with probability at least $1-\delta$, it holds that

$$\sum_{t=1}^T Z_t \le \sqrt{2(1-a)\gamma D \sum_{t=1}^T S_t(\boldsymbol{W}_t) \ln\frac{1}{\delta}} + \frac{\sqrt{2}}{3}\gamma D \ln\frac{1}{\delta} = \sqrt{2\zeta(1-a)\sum_{t=1}^T S_t(\boldsymbol{W}_t)} + \frac{\sqrt{2}}{3}\zeta, \tag{12}$$

where we let $\zeta = \gamma D \ln\frac{1}{\delta}$ for simplicity. Also, the $L^\star$ bound in the proof of Proposition 6 implies

$$\sum_{t=1}^T S_t(\boldsymbol{W}_t) \le \left(1 - \frac{\eta b}{2}\right)^{-1}\left(\frac{\|\boldsymbol{U}\|_{\mathrm{F}}^2}{2\eta} + \sum_{t=1}^T S_t(\boldsymbol{U})\right) = \frac{1}{1-\eta'} \cdot \frac{b\|\boldsymbol{U}\|_{\mathrm{F}}^2}{4\eta'} + \frac{1}{1-\eta'} \cdot \sum_{t=1}^T S_t(\boldsymbol{U}).$$

Substituting this into (12) and using the subadditivity of $\sqrt{\cdot}$ imply that the first term in (10) is at most

$$\sqrt{\frac{\zeta(1-a)}{1-\eta'} \cdot \frac{b\|\boldsymbol{U}\|_{\mathrm{F}}^2}{2\eta'}} + \sqrt{\frac{2\zeta(1-a)}{1-\eta'} \cdot \sum_{t=1}^T S_t(\boldsymbol{U})} + \frac{\sqrt{2}}{3}\zeta. \tag{13}$$

Therefore, the surrogate regret is bounded from above by the sum of (11) and (13). Since $\eta' < a$, the coefficient of the second term in (11) satisfies $a - \frac{1-a}{1-\eta'} \cdot \eta' > 0$. From $\sqrt{c_1 x} - c_2 x \le \frac{c_1}{4c_2}$ $(\forall x > 0)$ for $c_1, c_2 > 0$, we can offset the middle term in (13) with the second term in (11) as follows:

$$\sqrt{\frac{2\zeta(1-a)}{1-\eta'} \cdot \sum_{t=1}^T S_t(\boldsymbol{U})} - \left(a - \frac{1-a}{1-\eta'} \cdot \eta'\right)\sum_{t=1}^T S_t(\boldsymbol{U}) \le \frac{\zeta}{2} \cdot \frac{1-a}{a-\eta'} = \frac{1-a}{a}\zeta.$$

By putting everything together, the surrogate regret is at most

$$\frac{1-a}{1-\eta'} \cdot \frac{b\|\boldsymbol{U}\|_{\mathrm{F}}^2}{4\eta'} + \sqrt{\frac{\zeta(1-a)}{1-\eta'} \cdot \frac{b\|\boldsymbol{U}\|_{\mathrm{F}}^2}{2\eta'}} + \frac{\sqrt{2}}{3}\zeta + \frac{1-a}{a}\zeta \le \frac{1}{a}\left((1-a)b\|\boldsymbol{U}\|_{\mathrm{F}}^2 + \zeta\right),$$

where the inequality is for simplification; we applied AM–GM to the second term to upper bound it by $\frac{\zeta}{2} + \frac{1-a}{1-\eta'} \cdot \frac{b\|\boldsymbol{U}\|_{\mathrm{F}}^2}{4\eta'}$, used $\frac{1}{1-\eta'} \le 2$ (since $\eta' = \frac{a}{2} < \frac{1}{2}$), and simplified the resulting coefficient of ζ. Substituting $a = 1 - \frac{4\gamma}{\lambda\nu}$, $b = \frac{2C^2\kappa^2}{\lambda}$, and $\zeta = \gamma D \ln\frac{1}{\delta}$ into it completes the proof. ∎

Appendix G. Missing Details of Online-to-Batch Conversion

We prove the offline learning guarantee via online-to-batch conversion. We also discuss its connection to excess risk bounds.

G.1. Proof of Theorem 11

Proof Let $a = 1 - \frac{4\gamma}{\nu\lambda}$ and $b = \frac{2C^2\kappa^2}{\lambda}$, as in the proof of Theorem 7. For any $(\boldsymbol{x}, \boldsymbol{y}) \in \mathcal{X} \times \mathcal{Y}$, Lemma 4 ensures $\mathbb{E}[L(\psi_\Omega(\overline{\boldsymbol{W}}\boldsymbol{x}); \boldsymbol{y})] \le (1-a) S_\Omega(\overline{\boldsymbol{W}}\boldsymbol{x}; \boldsymbol{y})$, where the expectation is about the randomness of ψ_Ω. Let ρ denote the joint distribution on $\mathcal{X} \times \mathcal{Y}$. Since $S_\Omega(\boldsymbol{W}\boldsymbol{x}; \boldsymbol{y})$ is convex in $\boldsymbol{W}$, Jensen's inequality implies that the expected target loss, $\mathbb{E}[L(\psi_\Omega(\overline{\boldsymbol{W}}\boldsymbol{x}); \boldsymbol{y})]$, is bounded as follows:

$$\begin{aligned}\mathop{\mathbb{E}}_{\boldsymbol{x},\boldsymbol{y}\sim\rho}\big[\mathbb{E}\big[L(\psi_\Omega(\overline{\boldsymbol{W}}\boldsymbol{x}); \boldsymbol{y}) \mid \boldsymbol{x}, \boldsymbol{y}\big]\big] &\le (1-a) \mathop{\mathbb{E}}_{\boldsymbol{x},\boldsymbol{y}\sim\rho}\big[S_\Omega(\overline{\boldsymbol{W}}\boldsymbol{x}; \boldsymbol{y})\big] \\ &\le (1-a) \cdot \frac{1}{T}\sum_{t=1}^{T} \mathop{\mathbb{E}}_{\boldsymbol{x},\boldsymbol{y}\sim\rho}[S_\Omega(\boldsymbol{W}_t\boldsymbol{x}; \boldsymbol{y})].\end{aligned}$$

Furthermore, since $\boldsymbol{W}_t$ depends only on $(\boldsymbol{x}_s, \boldsymbol{y}_s)_{s<t}$, the law of total expectation implies $\mathbb{E}[S_t(\boldsymbol{W}_t)] = \mathbb{E}[S_\Omega(\boldsymbol{W}_t\boldsymbol{x}; \boldsymbol{y})]$ (see, e.g., Orabona (2023, Theorem 3.1) for a similar discussion); also, $\mathbb{E}[S_t(\boldsymbol{U})] = \mathbb{E}[S_\Omega(\boldsymbol{U}\boldsymbol{x}; \boldsymbol{y})]$ holds for any fixed $\boldsymbol{U} \in \mathcal{W}$. Therefore, it holds that

$$\begin{aligned}&\mathbb{E}[L(\psi_\Omega(\overline{\boldsymbol{W}}\boldsymbol{x}); \boldsymbol{y})] - \mathbb{E}[S_\Omega(\boldsymbol{U}\boldsymbol{x}; \boldsymbol{y})] \\ &\le (1-a) \cdot \frac{1}{T}\sum_{t=1}^{T} \mathbb{E}[S_\Omega(\boldsymbol{W}_t\boldsymbol{x}; \boldsymbol{y})] - \frac{1}{T}\sum_{t=1}^{T} \mathbb{E}[S_\Omega(\boldsymbol{U}\boldsymbol{x}; \boldsymbol{y})] \\ &= (1-a) \cdot \frac{1}{T}\sum_{t=1}^{T} \mathbb{E}[S_t(\boldsymbol{W}_t)] - \frac{1}{T}\sum_{t=1}^{T} \mathbb{E}[S_t(\boldsymbol{U})] \\ &= \frac{1}{T} \cdot \mathbb{E}\left[(1-a)\sum_{t=1}^{T}(S_t(\boldsymbol{W}_t) - S_t(\boldsymbol{U})) - a\sum_{t=1}^{T} S_t(\boldsymbol{U})\right] \\ &\le \frac{1}{T} \cdot (1-a)\left(1 - \frac{\eta b}{2}\right)^{-1} \frac{\|\boldsymbol{U}\|_{\mathrm{F}}^2}{2\eta}.\end{aligned}$$

where the last inequality follows from the same discussion as that in the proof of Proposition 6. By substituting $a = 1 - \frac{4\gamma}{\nu\lambda}$ and $b = \frac{2C^2\kappa^2}{\lambda}$ into the right-hand side, we obtain the desired bound. ■

G.2. Connection to Excess Risk Bound

We derive fast convergence of the excess risk from the surrogate regret bound in Theorem 11 under separability and realizability assumptions detailed below. The excess risk that we aim to bound is

$$\mathbb{E}[L(\psi_\Omega(\overline{\boldsymbol{W}}\boldsymbol{x}); \boldsymbol{y})] - \mathbb{E}[L(\boldsymbol{y}_0; \boldsymbol{y})],$$

where $\boldsymbol{y}_0$ is the Bayes rule. As Theorem 11 bounds the first term $\mathbb{E}[L(\psi_\Omega(\overline{\boldsymbol{W}}\boldsymbol{x}); \boldsymbol{y})]$ by $\mathbb{E}[S_\Omega(\boldsymbol{U}\boldsymbol{x}; \boldsymbol{y})]$, plus an asymptotically vanishing term of $O(1/T)$, it is sufficient to show

1. that $\boldsymbol{y}_0$ can be decoded from some linear predictor $\boldsymbol{U}_0$ and satisfies $\mathbb{E}[L(\boldsymbol{y}_0; \boldsymbol{y})] = 0$; and
2. that $\mathbb{E}[S_\Omega(\boldsymbol{U}\boldsymbol{x}; \boldsymbol{y})] = 0$ holds for some linear estimator $\boldsymbol{U}$.

Consequently, these combined with Theorem 11 imply the excess risk bound of

$$\mathbb{E}[L(\psi_\Omega(\overline{\boldsymbol{W}}\boldsymbol{x}); \boldsymbol{y})] - \mathbb{E}[L(\boldsymbol{y}_0; \boldsymbol{y})] = O(1/T).$$

Below, we design a Bayes rule (or, the best linear estimator and decoder) and check $\mathbb{E}[L(\boldsymbol{y}_0; \boldsymbol{y})] = 0$, and we confirm $\mathbb{E}[S_\Omega(\boldsymbol{U}\boldsymbol{x}; \boldsymbol{y})] = 0$ by using the SparseMAP surrogate loss S_Ω.

Let us introduce definitions needed in the subsequent discussion. We define the *frontier* (or the decision boundary) in a similar manner to Cabannes et al. (2021) as follows:

$$F \coloneqq \left\{ \boldsymbol{\theta} \in \mathbb{R}^d : |\arg\max_{\boldsymbol{y}\in\mathcal{Y}}\langle \boldsymbol{\theta}, \boldsymbol{y}\rangle| \geq 2 \right\}.$$

We also define the normal cone $N(\boldsymbol{y})$ at $\boldsymbol{y} \in \mathcal{Y}$ and its boundary $E(\boldsymbol{y})$ by

$$N(\boldsymbol{y}) \coloneqq \left\{ \boldsymbol{\theta} \in \mathbb{R}^d : \forall \boldsymbol{y}' \in \mathrm{conv}(\mathcal{Y}), \langle \boldsymbol{\theta}, \boldsymbol{y}' - \boldsymbol{y}\rangle \leq 0 \right\} \quad \text{and}$$
$$E(\boldsymbol{y}) \coloneqq \bigcup_{\boldsymbol{y}'\in\mathcal{Y}\setminus\{\boldsymbol{y}\}} \{ \boldsymbol{\theta} \in N(\boldsymbol{y}) : \langle \boldsymbol{\theta}, \boldsymbol{y}' - \boldsymbol{y}\rangle = 0 \},$$

respectively. One can readily confirm $F = \bigcup_{\boldsymbol{y}\in\mathcal{Y}} E(\boldsymbol{y})$. Let $d(\boldsymbol{\theta}, F) \coloneqq \min\{ \|\boldsymbol{\theta} - \boldsymbol{\theta}'\|_2 : \boldsymbol{\theta}' \in F \}$ denote the distance from $\boldsymbol{\theta}$ to F and D the ℓ_2-diameter of $\mathrm{conv}(\mathcal{Y})$. Below is a fundamental fact.

Lemma 18 *Let $\boldsymbol{y} \in \mathcal{Y}$ and $\boldsymbol{\theta} \in N(\boldsymbol{y})$. If $d(\boldsymbol{\theta}, F) \geq t$ holds for some $t > 0$, we have $\langle \boldsymbol{\theta}, \boldsymbol{y} - \boldsymbol{y}'\rangle \geq t\|\boldsymbol{y} - \boldsymbol{y}'\|_2$ for any $\boldsymbol{y}' \in \mathcal{Y}$.*

Proof The case of $\boldsymbol{y} = \boldsymbol{y}'$ is trivial. Below, we shift $\boldsymbol{y}$ to $\mathbf{0}$ and prove $\langle \boldsymbol{\theta}, \boldsymbol{z}\rangle \leq -t\|\boldsymbol{z}\|_2$ for any $\boldsymbol{z} \in \mathcal{Y} \setminus \{\boldsymbol{y}'\} - \{\boldsymbol{y}\}$; let $N = N(\mathbf{0})$ and $E = E(\mathbf{0})$. Since $\boldsymbol{z} \notin N$, there exists $\boldsymbol{\xi} \in E \setminus \{\mathbf{0}\}$ such that the sum of the angles between $\boldsymbol{\theta}$ and $\boldsymbol{\xi}$ and between $\boldsymbol{\xi}$ and $\boldsymbol{z}$, denoted by $\alpha, \beta \geq 0$, respectively, equals the angle between $\boldsymbol{\theta}$ and $\boldsymbol{z}$ and $\alpha + \beta \leq \pi$ holds. Due to $\boldsymbol{\xi} \in E \subseteq N$, we have $\beta \geq \pi/2$. Furthermore, since $d(\boldsymbol{\theta}, F) \geq t$ implies $d(\boldsymbol{\theta}, E) \geq t$, we have $\sin\alpha \geq t/\|\boldsymbol{\theta}\|_2$. Thus, $\cos(\alpha + \beta) \leq \cos(\alpha + \pi/2) = -\sin\alpha \leq -t/\|\boldsymbol{\theta}\|_2$ holds, hence $\langle \boldsymbol{\theta}, \boldsymbol{z}\rangle = \|\boldsymbol{\theta}\|_2\|\boldsymbol{z}\|_2 \cos(\alpha + \beta) \leq -t\|\boldsymbol{z}\|_2$. ∎

Assumptions. We assume a variant of the margin condition introduced in Cabannes et al. (2021, Assumption 3) and a realizability condition. Specifically, we assume there exists a linear estimator $\boldsymbol{U}_0 : \mathcal{X} \to \mathbb{R}^d$ satisfying the *no-density separation* for some margin $t_0 > 0$, i.e.,

$$\mathbb{P}_{\mathcal{X}}(d(\boldsymbol{U}_0\boldsymbol{x}, F) < t_0) = 0,$$

where $\mathbb{P}_{\mathcal{X}}$ represents the probability with respect to the marginal distribution of ρ on $\mathcal{X}$. In binary classification, this is sometimes called *Massart's noise condition*. We also require the underlying distribution ρ to satisfy a realizability condition that the above $\boldsymbol{U}_0$ satisfies

$$\mathbb{P}(\phi(\boldsymbol{U}_0\boldsymbol{x}) = \boldsymbol{y}) = 1,$$

where $\phi : \mathbb{R}^d \to \mathcal{Y}$ is a decoding function given by $\phi(\boldsymbol{U}_0\boldsymbol{x}) = \arg\max\{ \langle \boldsymbol{U}_0\boldsymbol{x}, \boldsymbol{y}\rangle : \boldsymbol{y} \in \mathcal{Y} \}$ (ties occur with probability zero due to the no-density separation). Then, $\boldsymbol{x} \mapsto \phi(\boldsymbol{U}_0\boldsymbol{x}) \in \mathcal{Y}$ is the Bayes rule that attains the zero target risk, i.e., $\mathbb{E}[L(\boldsymbol{y}_0; \boldsymbol{y})] = \mathbb{E}[L(\phi(\boldsymbol{U}_0\boldsymbol{x}); \boldsymbol{y})] = 0$. Therefore, the first condition is confirmed.

Surrogate loss. To establish the second condition $\mathbb{E}[S_\Omega(\boldsymbol{U}\boldsymbol{x}; \boldsymbol{y})] = 0$, we need a surrogate loss that attains zero for well-separated data. To this end, we employ the SparseMAP loss, which is known to have a unit *structured separation margin* (see Blondel et al. (2020, Section 7.4)). Specifically, for any $(\boldsymbol{\theta}, \boldsymbol{y}) \in \mathbb{R}^d \times \mathcal{Y}$, we have $S_\Omega(\boldsymbol{\theta}; \boldsymbol{y}) = 0$ if

$$\langle \boldsymbol{\theta}, \boldsymbol{y}\rangle \geq \max\left\{ \langle \boldsymbol{\theta}, \boldsymbol{y}'\rangle + \frac{1}{2}\|\boldsymbol{y} - \boldsymbol{y}'\|_2^2 : \boldsymbol{y}' \in \mathcal{Y} \right\}. \tag{14}$$

We show that $\boldsymbol{U} = \frac{D}{2t_0}\boldsymbol{U}_0$ satisfies $\mathbb{E}[S_\Omega(\boldsymbol{U}\boldsymbol{x};\boldsymbol{y})] = 0$. Due to the assumptions and Lemma 18 with $\boldsymbol{U}_0\boldsymbol{x} \in N(\phi(\boldsymbol{U}_0\boldsymbol{x}))$, for $(\boldsymbol{x},\boldsymbol{y})$ drawn from ρ and any $\boldsymbol{y}' \in \mathcal{Y}$, with probability 1, we have

$$\langle \boldsymbol{U}\boldsymbol{x}, \boldsymbol{y}\rangle = \frac{D}{2t_0}\langle \boldsymbol{U}_0\boldsymbol{x}, \phi(\boldsymbol{U}_0\boldsymbol{x})\rangle \geq \frac{D}{2t_0}\langle \boldsymbol{U}_0\boldsymbol{x}, \boldsymbol{y}'\rangle + \frac{D}{2}\|\phi(\boldsymbol{U}_0\boldsymbol{x}) - \boldsymbol{y}'\|_2 \geq \langle \boldsymbol{U}\boldsymbol{x}, \boldsymbol{y}'\rangle + \frac{1}{2}\|\boldsymbol{y} - \boldsymbol{y}'\|_2^2,$$

which implies (14) and thus $\mathbb{E}[S_\Omega(\boldsymbol{U}\boldsymbol{x};\boldsymbol{y})] = 0$ holds.

Remark 19 *As mentioned in Section 4.2, Cabannes et al. (2021) has already achieved an excess risk bound with an exponential rate under weaker assumptions, which is faster than the above $O(1/T)$ rate. Hence, the purpose of the above discussion is to deepen our understanding of the relationship between the surrogate regret and excess risk bounds, rather than to present a novel result.*

Appendix H. Proof of Theorem 13

Proof For simplicity, assume that $M := (B^2 - \ln^2(dT))/\ln^2(2d)$ is a positive integer. We sample true class $i_t \in [d]$ uniformly at random for $t = 1,\dots,M+1$. For $t > M+1$, we set $i_t = i_{M+1}$. Each $\boldsymbol{x}_t$ is a vector of length $M+1$. For $t = 1,\dots,M+1$, we let $\boldsymbol{x}_t = \mathbf{e}_t$, the tth standard basis vector in $\mathbb{R}^{M+1}$. For $t > M+1$, we let $\boldsymbol{x}_t = \mathbf{e}_{M+1}$. We define an offline estimator $\boldsymbol{U}' \in \mathbb{R}^{d\times(M+1)}$ as follows: the tth column of $\boldsymbol{U}'$ is $\ln(2d)\mathbf{e}_{i_t}$ for $t = 1,\dots,M$, and the $(M+1)$th column is $\ln(dT)\mathbf{e}_{M+1}$. Note that $\|\boldsymbol{U}'\|_{\mathrm{F}}^2 = M\ln^2(2d) + \ln^2(dT) = B^2$ always holds. Fix any learner's algorithm. For the first M rounds, the logistic loss of $\boldsymbol{U}'$ is bounded as $-\log_2\frac{2d}{2d+d-1} = \log_2\bigl(1 + \frac{1}{2}\bigl(1 - \frac{1}{d}\bigr)\bigr) \leq \frac{1}{2}\bigl(1 - \frac{1}{d}\bigr)$. Since each $i_t \in [d]$ is sampled uniformly at random, the expected 0-1 loss is $1 - \frac{1}{d}$. Therefore, the expected surrogate regret summed over the first M rounds is at least

$$\sum_{t=1}^{M}\mathbb{E}[\mathbb{1}_{\widehat{\boldsymbol{y}}_t\neq\boldsymbol{y}_t}] - \sum_{t=1}^{M}S_t(\boldsymbol{U}') \geq M\left(1 - \frac{1}{d}\right) - \frac{M}{2}\left(1 - \frac{1}{d}\right) = \frac{M}{2}\left(1 - \frac{1}{d}\right) \geq \frac{M}{4} = \Omega\left(\frac{B^2}{\ln^2 d}\right),$$

where we used $d \geq 2$ and $B = \Omega(\ln(dT))$. As for the remaining $T - M$ rounds, the logistic loss value is at most $\frac{1}{T}\bigl(1 - \frac{1}{d}\bigr)$ by a similar calculus to the above, whereas the expected 0-1 loss is at least $1 - \frac{1}{d}$ since i_{M+1} is sampled uniformly at random. Therefore, the expected surrogate regret over the $T - M$ rounds is non-negative. In total, the expected surrogate regret is $\Omega\Bigl(\frac{B^2}{\ln^2 d}\Bigr)$. ■