Efficient Primal-Dual Algorithms for Large-Scale Multiclass Classification

Dmitry Babichev, Dmitrii M. Ostrovskii, Francis Bach

Efficient Primal-Dual Algorithms for Large-Scale Multiclass Classification

Dmitry Babichev [*,†] Dmitrii M. Ostrovskii [*,†] Francis Bach [†]

{dmitry.babichev, dmitrii.ostrovskii, francis.bach}@inria.fr

Abstract

We develop efficient algorithms to train ℓ_1-regularized linear classifiers with large dimensionality d of the feature space, number of classes k, and sample size n. Our focus is on a special class of losses that includes, in particular, the multiclass hinge and logistic losses. Our approach combines several ideas: (i) passing to the equivalent saddle-point problem with a quasi-bilinear objective; (ii) applying stochastic mirror descent with a proper choice of geometry which guarantees a favorable accuracy bound; (iii) devising non-uniform sampling schemes to approximate the matrix products. In particular, for the multiclass hinge loss we propose a *sublinear* algorithm with iterations performed in $O(d+n+k)$ arithmetic operations.

1 Introduction

We study optimization problems arising in multiclass linear classification with a large number of classes and features. Formally, consider a dataset of n pairs (x_i, y_i), $i \in [n] := \{1, ..., n\}$, where $x_i \in \mathbb{R}^d$ is the feature vector of the i-th example, and $y_i \in \{e_1, ..., e_k\}$ is the label vector encoding one of k possible classes; here $e_1, ..., e_k$ are the standard basis vectors in $\mathbb{R}^k$. Given such data, our goal is to find a linear classifier that minimizes the ℓ_1-regularized empirical risk. We thus consider a minimization problem of the form

$$\min_{U \in \mathbb{R}^{d\times k}} \frac{1}{n}\sum_{i=1}^{n} \ell(U^\top x_i, y_i) + \lambda \|U\|_1. \tag{1}$$

Here, $U \in \mathbb{R}^{d\times k}$ is the matrix whose columns specify the parameter vectors for each of the k classes; $\ell(U^\top x, y)$, with $\ell : \mathbb{R}^k \times \Delta_k \to \mathbb{R}$ and $\Delta_k \subset \mathbb{R}^k$ being the unit probability simplex, is the *loss* corresponding to the margins $U^\top x \in \mathbb{R}^k$ assigned to x when its class is encoded by y; finally, the regularization term $\lambda\|U\|_1$, $\lambda \geqslant 0$, uses the elementwise ℓ_1-norm $\|U\|_1 = \sum_{i=1}^d \sum_{j=1}^k |U_{ij}|$. Apart from inducing sparsity of features and classes (Bühlmann and Van De Geer, 2011), this choice of regularization is crucial from the algorithmic perspective, as will be explained in Sec. 3.

Our focus is on the so-called *Fenchel-Young* losses, introduced by Blondel et al. (2018), which can be expressed as

$$\ell(U^\top x, y) = \max_{v \in \Delta_k} \left\{ -\mathfrak{f}(v, y) + (v - y)^\top U^\top x \right\}, \tag{2}$$

[*] Equal contribution.

[†] SIERRA Project-Team, INRIA and École Normale Supérieure, PSL Research University, Paris, France.

where Δ_k is the probability simplex in $\mathbb{R}^k$, and the function $\mathfrak{f}(\cdot, y) : \Delta_k \to \mathbb{R}$ is convex and "simple" (i.e., quasi-separable in v), which implies that maximization in (2) can be performed in running time $O(k)$. In particular, this allows us to address two commonly used multiclass losses:

- The multiclass *logistic (or softmax) loss*

$$\log\left(\sum_{l=1}^{k}\exp(U_l^\top x)\right) - y^\top U^\top x, \tag{3}$$

  where U_l is the l-th column of U so that $U_l^\top x$ is the l-th element of $U^\top x$. This loss corresponds to (2) with the negative entropy term $\mathfrak{f}(v, y) = \sum_{l=1}^{k} v_l \log v_l$, which is independent of y.

- The multiclass *hinge loss*, given by

$$\max_{l\in[k]}\left\{\mathbb{1}[e_l \neq y] + U_l^\top x\right\} - y^\top U^\top x, \tag{4}$$

  and used in multiclass support vector machines (SVM). This loss reduces to (2) by setting $\mathfrak{f}(v, y) = v^\top y - 1$ (refer to Appendix A for additional details).

Arranging the feature vectors into $X \in \mathbb{R}^{n\times d}$, and the class labels into $Y \in \mathbb{R}^{n\times k}$, and using the Fenchel-type representation (2) of the loss, we can recast the initial problem (1) as the following convex-concave saddle-point problem:

$$\min_{U\in\mathbb{R}^{d\times k}}\max_{V\in\mathcal{V}} \; -\mathcal{F}(V, Y) + \frac{1}{n}\mathrm{tr}\left[(V - Y)^\top XU\right] + \lambda\|U\|_1, \tag{5}$$

$$\text{with} \;\; \mathcal{F}(V, Y) := \frac{1}{n}\sum_{i=1}^{n}\mathfrak{f}(v_i, y_i), \tag{6}$$

where $v_i, y_i \in \Delta_k$ are the i-th rows of V and Y, and define the Cartesian product of probability simplices

$$\mathcal{V} := \Delta_k^{\otimes n} \subset \mathbb{R}^{n\times k}, \tag{7}$$

the set comprised of all right-stochastic matrices in $\mathbb{R}^{n\times k}$. Taking into account the Fenchel-type representation (2), this reduction is quite natural. Indeed, while the objective in (1) can be non-smooth, the essential part of the objective in (5),

$$\Phi(U, V - Y) := \frac{1}{n}\mathrm{tr}\left[(V - Y)^\top XU\right], \tag{8}$$

is not only smooth but *bilinear* in U and $V - Y$. On the other hand, the presence of the dual constraints, as given by (7), does not seem problematic since $\mathcal{V}$ allows for a computationally cheap projection oracle. Finally, in the saddle-point formulation we can control the duality gap which provides an *accuracy certificate* for the initial problem, see, e.g., Nemirovski et al. (2010); Ostrovskii and Harchaoui (2018).

In this work, we propose efficient algorithms for solving (1) via the associated saddle-point problem (5), built upon the vector-field formulation of stochastic mirror descent (SMD), a well-known general optimization method, see, e.g., Juditsky and Nemirovski (2011a) and references therein. Mirror descent, as well as its cousin Mirror Prox (Juditsky and Nemirovski, 2011b), allows to solve convex-concave saddle-point problems (CCSPPs) whenever the first-order information about the objective is available, and the primal and dual feasible sets have simple

structures that allow for easily computable prox-mappings. While these algorithms are poorly adapted for obtaining high-accuracy solutions, this is not a limitation in the context of empirical risk minimization where the ultimate goal is to minimize the *expected risk*, and there is often a natural level of statistical accuracy, going beyond which is unnecessary, see, e.g., Mokhtari et al. (2016). On the other hand, mirror descent is especially well-suited to quasi-bilinear CCSPPs of the form (5).

First, it uses the Bregman divergence, rather than the standard Euclidean distance, as a proximity measure, which allows to adjust to the specific geometry associated to $\mathcal{V}$ and $\|\cdot\|_1$.

Second, it retains its favorable convergence guarantee when the exact partial gradients

$$\nabla_U[\Phi(U, V-Y)] = \frac{1}{n}XU, \quad \nabla_V[\Phi(U, V-Y)] = \frac{1}{n}X^\top(V-Y),$$

cf. (8), are replaced with their unbiased estimates. This is especially important: recall that the exact computation of the matrix products XU, $X^\top(V-Y)$ requires $O(dnk)$ arithmetic operations (a.o.'s) while the subsequent proximal mappings can usually be done in linear time in the combined size of the variables, i.e., $O(dk+nk)$; thus, computation of the matrix products becomes the main bottleneck. On the other hand, these matrix products can be approximated via *randomized subsampling* of the elements of U, $V-Y$, and X. While a similar approach has already been explored by Juditsky and Nemirovski (2011b) in the case of bilinear CCSPPs with vector variables arising in sparse recovery, its extension to problems of the type (5) is non-trivial. In fact, for a sampling scheme to be deemed "good", it clearly has to satisfy two concurrent requirements.

(a) On one hand, one must control the stochastic variability of the estimates in the chosen sampling scheme. Ideally, the additional term due to sampling should not be much larger than the term already present in the accuracy bound for deterministic mirror descent.

(b) On the other hand, the estimates must be computationally cheap. The immediate goal is $O(dk+nk)$ per iteration, i.e., the cost of the proximal mapping given the full gradients. However, one might want to go beyond that, to $O(d+n+k)$ per iteration, obtaining a *sublinear* algorithm with complexity of an iteration much smaller than the combined size of the variables.

Devising a sampling scheme that satisfies both these requirements simultaneously is a delicate task. To solve it, one should carefully exploit the specific geometric structure of (5) associated to $\mathcal{V}$, the norm $\|\cdot\|_1$, and the term $\mathcal{F}(V,Y)$.

Contributions and Outline. We propose two sampling schemes with various levels of "aggressiveness" satisfying the above requirements, study their statistical properties, and analyze the numerical complexity of stochastic mirror descent (SMD) equipped with them. In particular, we show that SMD with appropriately balanced entropy-type potentials (see Sec. 2) has nearly the same complexity estimate, in terms of the number of iterations to guarantee a given duality gap, as its deterministic counterpart with exact gradients, while at the same time enjoying a drastically improved cost of iterations, with improvement depending on the scheme.

- The *partial sampling scheme* (see Sec. 3.1) is applicable to any problem of the form (5). In it, we sample a single *row* of U and $V-Y$ at a time, with probabilities minimizing the expected squared norm of the gradient estimate at the current iterate. This leads to the cost $O(dn+dk+nk)$ of iterations.

- In the *full sampling scheme* (see Sec. 3.2), sampling of the rows of U and $V - Y$ is augmented with column sampling. Applying it to the multiclass hinge loss (4), we construct a sublinear algorithm (see Sec. 4) with iteration cost of $O(d + n + k)$ a.o.'s, and a one-time additional cost of $O(dn + nk)$ a.o.'s (starting from the zero primal solution allows to remove the $O(dk)$ term).

We conclude the paper with numerical experiments that illustrate our approach (see Sec. 5).

Related Work. Instead of passing to the saddle-point problem (5) we can solve the original problem (1) directly. For deterministic first-order algorithms, this leads to $O(dnk)$ numerical complexity of one iteration, and for stochastic algorithms that sample one example at a time, the complexity is $O(dk)$, similarly to our approach with partial sampling. In particular, algorithms such as SAG (Schmidt et al., 2017), SVRG (Johnson and Zhang, 2013), SDCA (Shalev-Shwartz and Zhang, 2013), SAGA (Defazio et al., 2014) use variance reduction techniques to obtain accelerated convergence rates in terms of the number of iterations, but all have $O(dk)$ or $O(dk + nk)$ runtime. Such variance reduction techniques have been extended to saddle-point problems (Palaniappan and Bach, 2016; Shi et al., 2017) but with the same overall complexity (our full sampling scheme could also be adapted to them); still, none of these algorithms are sublinear.

Regarding sublinear algorithms, several results can be found in the literature for the biclass setting. The case of bilinear CCSPPs was first considered by Grigoriadis and Khachiyan (1995), then by Juditsky and Nemirovski (2011b), and by Xiao et al. (2017). Hazan et al. (2011) proposed a sublinear algorithm for biclass SVM, and Garber and Hazan (2011, 2016) addressed semidefinite programs; more general results were given by Clarkson et al. (2012). We reuse some of the tools (mirror descent and multiplicative updates) considered in this literature. However, none of these approaches can be easily extended to the multiclass setting without an extra $O(k)$ factor in the cost of iterations.

2 Mirror Descent for Quasi-Bilinear CCSPPs

Preliminary Reductions. We focus on a CCSPP given by (5)–(8), and make the following assumption:

Assumption 1. *We assume that the $\|\cdot\|_1$-radius of an optimal solution U^* to (5) is known:*

$$\|U^*\|_1 = R_*.$$

Remark 2.1. *The accuracy bounds presented later on depend on R_*, and are preserved when R_* becomes an upper bound on $\|U^*\|_1$. Since $U = 0$ is feasible, we always have $R_* \leqslant 1/\lambda$ when the loss is non-negative, but this bound is usually loose since λ typically decreases with n. Alternatively, we can solve a series of constrained problems, starting with a small radius, and increasing it by a constant factor until the obtained solution leaves the boundary of the feasible set.*

With Assumption 1, we can put (5) in the constrained form:

$$\min_{\|U\|_1 \leq R_*} \max_{V \in \mathcal{V}} -\mathcal{F}(V, Y) + \Phi(U, V - Y) + \lambda \|U\|_1, \tag{9}$$

and then reduce (9) to a simplex-constrained problem. Indeed, let $\mathbb{1}_{2d \times k} \in \mathbb{R}^{2d \times k}$ be the all-ones matrix, and define

$$\mathcal{U} := \big\{U \in \mathbb{R}^{2d \times k} : U_{ij} \geqslant 0,\ \operatorname{tr}[\mathbb{1}_{2d \times k}^\top U] \leqslant R_*\big\}, \tag{10}$$

i.e., the "solid" simplex in $\mathbb{R}^{2d\times k}$ (note that $\mathrm{tr}[\mathbb{1}_{2d\times k}^\top U] = \|U\|_1$ on $\mathcal{U}$). Consider now the following CCSPP:

$$\min_{U\in\mathcal{U}}\max_{V\in\mathcal{V}} -\mathcal{F}(V,Y) + \widehat{\Phi}(U, V-Y) + \lambda\,\mathrm{tr}[\mathbb{1}_{2d\times k}^\top U], \tag{11}$$

where $\mathcal{F}(V,Y)$ and $\mathcal{V}$ are given by (6)–(7), $\mathcal{U}$ by (10), and

$$\widehat{\Phi}(U, V-Y) := \frac{1}{n}\mathrm{tr}\left[(V-Y)^\top \widehat{X} U\right],\ \widehat{X} := \left[X, -X\right], \tag{12}$$

using the "Matlab notation" for matrix concatenation (i.e., $\widehat{X} \in \mathbb{R}^{n\times 2d}$). One can verify that (11) is equivalent to (9), and hence to (5), in the following sense: any ε-accurate (in terms of the primal accuracy or duality gap, see below) solution (U,V) to (11) with $U=[U_1;U_2]$ results in the ε-accurate solution (U_1-U_2,V) to (9). This reduction is motivated by the fact that mirror descent with an entropy-type potential on $\mathcal{U}$ reduces to multiplicative updates, which is crucial in the sublinear algorithm presented in Sec. 4.

Background on CCSPPs. The accuracy of a candidate solution $(\bar{U},\bar{V})$ to a CCSPP

$$\min_{U\in\mathcal{U}}\max_{V\in\mathcal{V}} f(U,V)$$

with compact sets $\mathcal{U},\mathcal{V}$ can be quantified via the *duality gap*

$$\mathtt{Gap}(\bar{U},\bar{V}) := \max_{V\in\mathcal{V}} f(\bar{U},V) - \min_{U\in\mathcal{U}} f(U,\bar{V}). \tag{13}$$

Under certain conditions which in particular hold for (11), see Sion (1958), this CCSPP possesses an optimal solution $W^* = (U^*,V^*)$, called a *saddle point*, for which it holds $f(U^*,V^*) = \max_{V\in\mathcal{V}} f(U^*,V) = \min_{U\in\mathcal{U}} f(U,V^*)$. That is, U^* (resp. V^*) is optimal in the primal problem of minimizing $f_{\mathrm{prim}}(U) := \max_{V\in\mathcal{V}} f(U,V)$ (resp. the dual problem of maximizing $f_{\mathrm{dual}}(V) := \min_{U\in\mathcal{U}} f(U,\bar{V})$). Thus, the duality gap bounds from above the primal accuracy – in our case, the accuracy of solving the initial problem (1).

2.1 Choice of the Geometry

When applied to CCSPPs, the geometry of mirror descent is specified by the choice of the pair of the primal and dual norms $\|\cdot\|_{\mathscr{U}}$, $\|\cdot\|_{\mathscr{V}}$ in which the optimal solution is likely to be small, and the *potentials*, or *distance-generating functions* in the terminology of Nemirovski (2004), that must satisfy some compatibility properties with respect to these norms.

Norms. It is natural to use $\|\cdot\|_1$ as the primal norm $\|\cdot\|_{\mathscr{U}}$. On the other hand, due to (7), it is reasonable to choose the norm $\|\cdot\|_{\mathscr{V}}$ within the family of mixed $\ell_p\times\ell_q$ norms

$$\|V\|_{p\times q} := \left(\sum_i \|V(i,:)\|_q^p\right)^{1/p},\quad p,q\in[1,\infty], \tag{14}$$

i.e., the ℓ_p-norm of the vector of ℓ_q-norms of the rows, with $q=1$. The question is how to choose p. The naive choice $p=\infty$, corresponding to the direct-product structure of $\mathcal{V}$, does not allow for a compatible potential (see Juditsky and Nemirovski (2011a)). The remedy is to replace $p=\infty$ with $p=2$, leading to the choice $\|\cdot\|_{\mathscr{V}} = \|\cdot\|_{2\times 1}$. In fact, this choice can also be motivated from the black-box model perspective (Nemirovsky and Yudin, 1983).

Remark 2.2. *Other choices of the regularization norm and the norm $\|\cdot\|_{\mathscr{V}}$ are explored in Appendix E. As it turns out, the choice described here is the only one in a broad class of those using the mixed $\ell_p\times\ell_q$ norms, for which one can achieve the goal stated in Section 1, that is, have both a favorable accuracy guarantee and an efficient (sublinear) algorithmic implementation.*

Partial Potentials. A potential $\phi_{\mathcal{U}} : \mathcal{U} \to \mathbb{R} \cup \{\infty\}$ is called *compatible* with the norm $\|\cdot\|_{\mathscr{U}}$ in the sense of Nemirovski (2004) when it admits a continous selection of subgradients in the relative interior of $\mathcal{U}$, and is 1-strongly convex on $\mathcal{U}$ with respect to $\|\cdot\|_{\mathscr{U}}$. In order to obtain favorable convergence guarantees, the primal and dual potentials $\phi_{\mathcal{U}}$, $\phi_{\mathcal{V}}$ must satisfy two conditions. First, they must be compatible with the chosen norms; second, *the potential differences*, defined by

$$\begin{aligned} \Omega_{\mathcal{U}} &:= \max_{U \in \mathcal{U}} \phi_{\mathcal{U}}(U) - \min_{U \in \mathcal{U}} \phi_{\mathcal{U}}(U), \\ \Omega_{\mathcal{V}} &:= \max_{V \in \mathcal{V}} \phi_{\mathcal{V}}(V) - \min_{V \in \mathcal{V}} \phi_{\mathcal{V}}(V), \end{aligned} \tag{15}$$

must be upper-bounded, up to logarithmic factors in the problem dimension, with the squared radii of $\mathcal{U}, \mathcal{V}$ in the corresponding norms (Nemirovsky and Yudin, 1983). We now specify the potentials that satisfy these requirements.

The natural choice for the dual potential $\phi_{\mathcal{V}}(\cdot)$, reflecting the product structure of $\mathcal{V}$, is the sum of negative entropies (Beck and Teboulle, 2003):

$$\phi_{\mathcal{V}}(V) = \sum_{i=1}^{n} \sum_{l=1}^{k} V_{il} \log(V_{il}). \tag{16}$$

Its compatibility with $\|\cdot\|_{2\times 1}$ follows from Pinsker's inequality (Kemperman, 1969) applied rowwise to V. On the other hand, we have

$$\Omega_{\mathcal{V}} = n \log k, \tag{17}$$

whereas the squared $\|\cdot\|_{2\times 1}$-norm of any feasible solution V to (9) is precisely n. Thus, (16) is a valid potential on $\mathcal{V}$.

Regarding the choice of the potential on $\mathcal{U}$, consider first the unit "solid" simplex, i.e., the set (10) with $R_* = 1$. On this set, one can define the *unnormalized negative entropy*

$$\mathcal{H}(U) = \sum_{i=1}^{d} \sum_{j=1}^{k} \{U_{ij} \log U_{ij} - U_{ij}\}. \tag{18}$$

Clearly, $\mathcal{H}(\cdot)$ is continuously differentiable in the interior of its domain, and one can show that it is 1-strongly convex on it (see Yu (2013)). Due to Assumption 1, we can consider

$$\phi_{\mathcal{U}}(U) := R_*^2 \cdot \mathcal{H}(U/R_*), \tag{19}$$

which is then a compatible potential on $\mathcal{U}$ that satisfies

$$\Omega_{\mathcal{U}} = R_*^2 \log(2dk), \tag{20}$$

and thus is a valid potential on $\mathcal{U}$. Note that for our choice of the potentials, the corresponding Bregman divergences are expressed in terms of the Kullback-Leibler divergence, and thus lead to multiplicative updates. This circumstance is crucial for the sublinear algorithm considered in Section 4.

2.2 Composite Saddle-Point Mirror Descent

We use the composite variant of saddle-point mirror descent applicable for quasi-bilinear CC-SPPs, see Juditsky and Nemirovski (2011b); Ostrovskii and Harchaoui (2018). Introducing $W = (U, V) \in \mathcal{W}$ $[:= \mathcal{U} \times \mathcal{V}]$, the algorithm can be summarized as follows. First, one constructs the *joint potential* $\phi(W)$ on $\mathcal{W}$ by reweighting $\phi_{\mathcal{U}}(U)$ and $\phi_{\mathcal{V}}(V)$:

$$\phi(W) = \frac{\phi_{\mathcal{U}}(U)}{2\Omega_{\mathcal{U}}} + \frac{\phi_{\mathcal{V}}(V)}{2\Omega_{\mathcal{V}}}. \tag{21}$$

Such reweighting, possible in our case due to Assumption 1, allows to improve the accuracy bound, replacing the factor $\Omega_{\mathcal{U}} + \Omega_{\mathcal{V}}$ with $\sqrt{\Omega_{\mathcal{U}}\Omega_{\mathcal{V}}}$ (cf. Theorem 2.1 below). Then, initializing with $W^0 = \min_{W\in\mathcal{W}} \phi_{\mathcal{W}}(W)$, one iterates

$$W^{t+1} = \operatorname*{argmin}_{W\in\mathcal{W}} h(W) + \langle G(W^t), W\rangle + \frac{D_\phi(W, W^t)}{\gamma_t}, \tag{22}$$

where $\{\gamma_t\}$ is the sequence of stepsizes, $h(W) = \mathcal{F}(V, Y) + \lambda \operatorname{tr}[\mathbb{1}_{2d\times k}^\top U]$ is the combined "simple" term (cf. (6)),

$$G(W) = \left[\frac{1}{n}X^\top(V - Y), -\frac{1}{n}XU\right] \tag{23}$$

is the vector field of the partial gradients of $\widehat{\Phi}(U, V - Y)$, cf. (12), and $D(\cdot,\cdot)$ is the Bregman divergence[1] linked to $\phi(\cdot)$:

$$D(W, W') = \phi(W) - \phi(W') - \left\langle \nabla\phi(W'), W - W'\right\rangle.$$

Note that in the case of (11), this amounts to the initialization

$$V^0 = \frac{1}{k}\mathbb{1}_{n\times k}, \quad U^0 = \frac{R_*}{2dk}\mathbb{1}_{2d\times k}, \tag{Init}$$

and iterations (separable in U and V) of the form

$$\begin{aligned}&(U^{t+1}, V^{t+1})\\ &= \operatorname*{argmin}_{U\in\mathcal{U}, V\in\mathcal{V}} \left\{\lambda \operatorname{tr}[\mathbb{1}_{2d\times k}^\top U] + \widehat{\Phi}(U, V^t - Y) + \frac{D_{\mathcal{U}}(U, U^t)}{2\gamma_t\Omega_{\mathcal{U}}} + \mathcal{F}(V, Y) - \widehat{\Phi}(U^t, V) + \frac{D_{\mathcal{V}}(V, V^t)}{2\gamma_t\Omega_{\mathcal{V}}}\right\},\end{aligned} \tag{MD}$$

where $D_{\mathcal{U}}, D_{\mathcal{V}}$ are the Bregman divergences for $\phi_{\mathcal{U}}, \phi_{\mathcal{V}}$.

Complexity of Iterations. One iteration in (MD) has running time $O(dnk)$, and is dominated by the computation of the matrix products XU and $X^\top(V - Y)$; once they are known, the proximal step only requires $O(dk + nk)$ a.o.'s. Indeed, given $S^t = \frac{1}{n}\widehat{X}^\top(V^t - Y)$, the primal update with the choice (19) of $\phi_{\mathcal{U}}$ can be expressed in closed form (refer to Lemma C.1 in Appendix C for the derivation):

$$\begin{aligned}U_{il}^{t+1} &= U_{il}^t\, e^{-2\gamma_t S_{il}^t R_* L} \min\left\{e^{-2\gamma_t\lambda R_* L}, R_*/M\right\},\\ \text{where } L &:= \log(2dk), \quad \text{and } M := \sum_{i=1}^{2d}\sum_{l=1}^{k} U_{il}^t \cdot e^{-2\gamma_t S_{il}^t R_* L}.\end{aligned} \tag{24}$$

On the other hand, our ability to perform the dual updates depends on the form of $\mathfrak{f}(v, y)$ in the representation (2). In the general case, due to $\mathfrak{f}(v, y)$ being separable in v, we can reduce the dual update to $O(nk)$ *one-dimensional* optimization problems. This can be done by passing to the Lagrangian dual problem (which is separable), minimizing the Lagrangian for the given value of multiplier by solving $O(nk)$ one-dimensional problems, and finding the optimal Lagrange multiplier via root search. Moreover, for the multiclass hinge loss (4) we have the closed-form updates:

$$V_{il}^{t+1} = \frac{V_{il}^t \exp\left(2\gamma_t[\widehat{X}U^t - Y]_{il}\log(k)\right)}{\sum_{l=1}^k V_{il}^t \exp\left(2\gamma_t[\widehat{X}U^t - Y]_{il}\log(k)\right)}. \tag{25}$$

[1] We ignore subtleties related to the correct definition of the domain of $D(W,\cdot)$ (see, e.g., (Beck and Teboulle, 2003)); nonetheless, the subsequent algorithms are correctly defined.

Convergence Rate. The convergence rate of mirror descent for CCSPPs of the form (11) depends on the quantity

$$\mathcal{L}_{\mathscr{U},\mathscr{V}} := \frac{1}{n} \sup_{\|U\|_{\mathscr{U}} \leqslant 1} \|\widehat{X}U\|_{\mathscr{V}^*}, \tag{26}$$

where $\|\cdot\|_{\mathscr{V}^*}$ is the dual norm to $\|\cdot\|_{\mathscr{V}}$. Thus, $\mathcal{L}_{\mathscr{U},\mathscr{V}}$ is the $(\mathscr{U}, \mathscr{V}^*)$-subordinate norm of the linear mapping $U \mapsto \frac{1}{n}\widehat{X}U$. For the norms chosen in Section 2.1, $\mathcal{L}_{\mathscr{U},\mathscr{V}}$ is expressed as a mixed norm (14) (see Appendix for the proof):

Proposition 2.1. *For $\|\cdot\|_{\mathscr{U}} = \|\cdot\|_1$ and $\|\cdot\|_{\mathscr{V}} = \|\cdot\|_{2\times 1}$, one has*

$$\mathcal{L}_{\mathscr{U},\mathscr{V}} = \frac{1}{n}\|X^\top\|_{\infty\times 2}. \tag{27}$$

We obtain the following convergence guarantee for our variant of mirror descent applied to CCSPP (11). For simplicity, we consider constant stepsize and simple averaging; empirically we observe similar results for the time-varying stepsize $\gamma_t \propto 1/\sqrt{t+1}$.

Theorem 2.1. *Let $(\bar{U}^T, \bar{V}^T) = \frac{1}{T}\sum_{t=0}^{T-1}(U^t, V^t)$ be the average of the first T iterates of mirror descent (**MD**) with initialization (**Init**) and stepsize $\gamma_t \equiv 1/(\mathcal{L}_{\mathscr{U},\mathscr{V}}\sqrt{5T\Omega_{\mathcal{U}}\Omega_{\mathcal{V}}})$, with $\mathcal{L}_{\mathscr{U},\mathscr{V}}$, $\Omega_{\mathcal{U}}$, $\Omega_{\mathcal{V}}$ given by (27), (20), (17). Then the duality gap can be bounded as*

$$\begin{aligned}\mathtt{Gap}(\bar{U}^T, \bar{V}^T) &\leqslant \frac{2\sqrt{5}\mathcal{L}_{\mathscr{U},\mathscr{V}}\sqrt{\Omega_{\mathcal{U}}\Omega_{\mathcal{V}}}}{\sqrt{T}} + \frac{\mathtt{r}}{T} \\ &\leqslant \frac{2\sqrt{5}\|X^\top\|_{\infty\times 2}}{\sqrt{n}}\frac{\log(2dk)R_*}{\sqrt{T}} + \frac{\mathtt{r}}{T}.\end{aligned} \tag{28}$$

Here $\mathtt{r} = 0$ for the multiclass hinge loss (4) and $\mathtt{r} = \max_{y\in\Delta_k}\left\{\mathfrak{f}(\frac{\mathbb{1}_k}{k}, y) - \min_{v\in\Delta_k}\mathfrak{f}(v, y)\right\}$ in the general case; in particular, $\mathtt{r} = O(\log(d))$ for the softmax loss (3).

Proof. The first bound follows from the general result for quasi-bilinear CCSPPs, see Theorem B.1 in Appendix.[2] Its combination with (17), (20), and (27) results in (28). □

Remark 2.3. *Note that the j-th column $X_j \in \mathbb{R}^n$ of X represents the empirical distribution of the feature φ_j. Hence, when the data is i.i.d., we have*

$$\frac{\|X^\top\|_{\infty\times 2}}{\sqrt{n}} = \max_{j\in[d]}\sqrt{\frac{\|X_j\|_2^2}{n}} \xrightarrow[n\to\infty]{a.s.} \max_{j\in[d]}\left(\mathbb{E}_{\text{data}}[\varphi_j^2]\right)^{1/2}, \tag{29}$$

where the expectation $\mathbb{E}_{\text{data}}$ is over the data distribution. In other words, $\|X^\top\|_{\infty\times 2}/\sqrt{n}$ has a finite limit, converging to the largest L_2-norm of a feature. In the non-asymptotic regime, $\|X^\top\|_{\infty\times 2}/\sqrt{n}$ is the largest empirical *L_2-norm of a feature, and can be controlled if the features are bounded, or sufficiently light-tailed, via standard concentration inequalities. In particular, $\|X^\top\|_{\infty\times 2}/\sqrt{n} \leqslant B$ if the features are uniformly bounded with B, and for the Gaussian features $\varphi_j \sim \mathcal{N}(0, \sigma_j^2)$ with $\sigma_j \leq B$ we have, with probability at least $1-\delta$,*

$$\frac{\|X^\top\|_{\infty\times 2}}{\sqrt{n}} \leqslant C\sigma\left(1 + \sqrt{\frac{\log(d/\delta)}{n}}\right), \tag{30}$$

for some constant C, see Laurent and Massart (2000, Lem. 1).

[2] Note that the results of Duchi et al. (2010) nor those of Nesterov and Nemirovski (2013) cannot be readily applied in our setup.

Remark 2.4. *It is known that the convergence rate can be improved to $O(1/T)$ for the composite version of Mirror Prox, but this improvement is not preserved in the stochastic setting. On the other hand, this allows to emulate the "mini-batching" technique, by sampling the matrix products repeatedly (or in parallel), and controlling the variability of the averaged gradient estimates via Nemirovski's inequalities in the vein of Juditsky and Nemirovski (2011b, Sec. 2.5.1).*

3 Sampling Schemes

Recall that the bottleneck of mirror descent iterations (**MD**) is computing the matrix products $\widehat{X}U^t, \widehat{X}^\top(V^t - Y)$ which requires $O(dnk)$ a.o.'s. Inspired by Juditsky and Nemirovski (2011b), we propose sampling schemes that produce unbiased estimates ξ_{U^t} and $\eta_{V^t,Y}$ of $\widehat{X}U^t$ and $\widehat{X}^\top(V^t - Y)$ with reduced complexity of computation, and use them to approximate the true partial gradients, arriving at the following variant of stochastic mirror descent:

$$\begin{aligned}&(U^{t+1}, V^{t+1})\\ &= \operatorname*{argmin}_{U\in\mathcal{U}, V\in\mathcal{V}} \Big\{\lambda \operatorname{tr}[\mathbb{1}_{2d\times k}^\top U] + \frac{1}{n}\operatorname{tr}\left[\eta_{V^t,Y} U\right] + \frac{D_{\mathcal{U}}(U, U^t)}{2\gamma_t \Omega_{\mathcal{U}}}\\ &\qquad + \mathcal{F}(V, Y) - \frac{1}{n}\operatorname{tr}\left[\xi_{U^t}^\top V\right] + \frac{D_{\mathcal{V}}(V, V^t)}{2\gamma_t \Omega_{\mathcal{V}}}\Big\}.\end{aligned} \tag{SMD}$$

Since the gradients are now replaced with their unbiased estimates, the accuracy bound gets augmented with an extra term that reflects the variability of these estimates. This extra term is known to be $O(\sqrt{(\Omega_{\mathcal{U}}\sigma_{\mathcal{V}}^2 + \Omega_{\mathcal{V}}\sigma_{\mathcal{U}}^2)/T})$, where $\sigma_{\mathcal{U}}^2$ and $\sigma_{\mathcal{V}}^2$ are "variance proxies" – the natural analogues of the variances of ξ_U and $\eta_{V,Y}$ for the chosen norms:

$$\begin{aligned}\sigma_{\mathcal{U}}^2 &:= \frac{1}{n^2}\sup_{U\in\mathcal{U}} \mathbb{E}\Big[\big\|\widehat{X}U - \xi_U\big\|_{\mathscr{Y}^*}^2\Big],\\ \sigma_{\mathcal{V}}^2 &:= \frac{1}{n^2}\sup_{(V,Y)\in\mathcal{V}\times\mathcal{V}} \mathbb{E}\Big[\big\|\widehat{X}^\top(V - Y) - \eta_{V,Y}\big\|_{\mathscr{U}^*}^2\Big].\end{aligned} \tag{31}$$

We consider two sampling schemes for ξ_U and $\eta_{V,Y}$: *partial sampling* where the estimates are obtained by sampling the rows of U and $V - Y$, and *full sampling*, where one subsequently samples their columns. In both cases, we derive the data-dependent sampling disrtibutions with near-optimal variance proxies. Our finding is that for the mirror descent geometry chosen in Sec. 2, and under mild assumptions on the data distribution, application of both schemes with the found distributions results in essentially the same convergence rate as in the deterministic case.

3.1 Partial Sampling Scheme

In the *partial sampling scheme*, we choose a pair of distributions $p = (p_1, ..., p_{2d}) \in \Delta_{2d}$ and $q = (q_1, ..., q_n) \in \Delta_n$, and draw one row of U and $V - Y$ (i.e., a feature and a tratining example) at a time according to p and q. In other words, we produce the estimates

$$\xi_U(p) = \widehat{X}\frac{e_i e_i^\top}{p_i}U, \quad \eta_{V,Y}(q) = \widehat{X}^\top\frac{e_j e_j^\top}{q_j}(V - Y), \tag{Part-SS}$$

where $e_i \in \Delta_{2d}$ and $e_j \in \Delta_n$ are standard basis vectors, and $i \in [2d], j \in [n]$ are drawn from p, q correspondingly; clearly, this gives unbiased estimates. The challenge is to choose the distributions p, q. In the Euclidean case, i.e., when $\|\cdot\|_{\mathscr{U}^*}, \|\cdot\|_{\mathscr{Y}^*}$ are Frobenius norms, one can

explicitly minimize the variances of the resulting estimates, and it is equivalent to minimizing the second moments $\mathbb{E}\left[\|\xi_U(p)\|_{\mathrm{F}}^2\right], \mathbb{E}\left[\|\eta_{V,Y}(q)\|_{\mathrm{F}}^2\right]$. In general, this is not the case. Next we show that for our mixed norms, the problem of minimizing the second moment proxies, i.e., finding

$$\begin{aligned} p^* = p^*(\widehat{X}, U) &\in \operatorname*{Argmin}_{p\in\Delta_{2d}} \mathbb{E}\left[\|\xi_U(p)\|_{\mathscr{Y}^*}^2\right], \\ q^* = q^*(\widehat{X}, V, Y) &\in \operatorname*{Argmin}_{q\in\Delta_n} \mathbb{E}\left[\|\eta_{V,Y}(q)\|_{\mathscr{U}^*}^2\right], \end{aligned} \tag{32}$$

can be solved explicitly, due to the matrices in the right-hand side of (**Part-SS**) being one-rank. The variance proxies (cf. (31)) can then be bounded via the triangle inequality.

Proposition 3.1. *When $\|\cdot\|_{\mathscr{U}} = \|\cdot\|_1$ and $\|\cdot\|_{\mathscr{Y}} = \|\cdot\|_{2\times 1}$, the optimal distributions p^* and q^*, cf. (32), are given by*

$$\begin{aligned} p_i^* &\propto \|\widehat{X}(:,i)\|_2 \cdot \|U(i,:)\|_\infty, \\ q_j^* &\propto \|\widehat{X}(j,:)\|_\infty \cdot \|V(j,:) - Y(j,:)\|_\infty, \end{aligned} \tag{33}$$

where $A(i,:)$ and $A(:,j)$ are the i-th row and j-th column of A. Moreover, the corresponding variance proxies satisfy

$$\begin{aligned} \sigma_{\mathcal{U}}^2(p^*) &\leqslant \frac{4R_*^2\|X^\top\|_{\infty\times 2}^2}{n^2}, \\ \sigma_{\mathcal{V}}^2(q^*) &\leqslant \frac{8\|X^\top\|_{\infty\times 2}^2}{n} + \frac{8\|X\|_{1\times\infty}^2}{n^2}. \end{aligned} \tag{34}$$

See Appendix D.2 for the proof of an extended result for the general mixed norms (14). Combined with the general result for composite saddle-point stochastic mirror descent (Theorem B.2 in Appendix), Proposition 3.1 implies the following result:

Theorem 3.1. *Let $(\bar{U}^T, \bar{V}^T) = \frac{1}{T}\sum_{t=0}^{T-1}(U^t, V^t)$ be the average of T iterates of stochastic mirror descent (**SMD**) initialized with (**Init**), equipped with sampling scheme (**Part-SS**) with distributions (33), and with stepsize*

$$\gamma_t \equiv \frac{1}{\sqrt{2T}} \min\left\{\frac{1}{\mathcal{L}_{\mathscr{U},\mathscr{Y}}\sqrt{5\Omega_{\mathcal{U}}\Omega_{\mathcal{V}}}}, \frac{1}{\sqrt{\Omega_{\mathcal{U}}\bar{\sigma}_{\mathcal{V}}^2 + \Omega_{\mathcal{V}}\bar{\sigma}_{\mathcal{U}}^2}}\right\}, \tag{35}$$

with $\mathcal{L}_{\mathscr{U},\mathscr{Y}}, \Omega_{\mathcal{U}}, \Omega_{\mathcal{V}}$ given by (27), (17), (20), and the upper bounds $\bar{\sigma}_{\mathcal{U}}^2, \bar{\sigma}_{\mathcal{V}}^2$ on the variance proxies given by (34). Then

$$\begin{aligned} \mathbb{E}[\mathtt{Gap}(\bar{U}^T, \bar{V}^T)] &\leqslant \frac{2\sqrt{10}\mathcal{L}_{\mathscr{U},\mathscr{Y}}\sqrt{\Omega_{\mathcal{U}}\Omega_{\mathcal{V}}}}{\sqrt{T}} + \frac{2\sqrt{2}\sqrt{\Omega_{\mathcal{U}}\bar{\sigma}_{\mathcal{V}}^2 + \Omega_{\mathcal{V}}\bar{\sigma}_{\mathcal{U}}^2}}{\sqrt{T}} + \frac{\mathtt{r}}{T} \\ &\leqslant \left(\frac{16.2\,\|X^\top\|_{\infty\times 2}}{\sqrt{n}} + \frac{8\|X\|_{1\times\infty}}{n}\right)\frac{\log(2dk)R_*}{\sqrt{T}} + \frac{\mathtt{r}}{T}, \end{aligned} \tag{36}$$

where the expectation is over the randomness of the algorithm, and $\mathtt{r}$ is the same as in Theorem 2.1.

Remark 3.1. *Comparing (36) with (28), we see that the partial sampling (**Part-SS**) does not deteriorate the convergence rate as long as the extra term $\|X\|_{1\times\infty}/n$ does not dominate $\|X^\top\|_{\infty\times 2}/\sqrt{n}$. When the data is* light-tailed, *the two terms are comparable. Indeed, $\|X^\top\|_{\infty\times 2}/\sqrt{n}$ a.s. converges to $\max_{j\in[d]}\left(\mathbb{E}_{\mathrm{data}}[\varphi_j^2]\right)^{1/2}$, where φ_j's are the features (cf. (29)), and the term $\|X\|_{1\times\infty}/n$ clearly converges to $\mathbb{E}_{\mathrm{data}}\max_{j\in[d]}|\varphi_j|$. When φ_j's are subgaussian, we have*

$$\mathbb{E}_{\mathrm{data}}[\max_{j\in[d]}|\varphi_j|] \leqslant \widetilde{O}_d(1)\max_{j\in[d]}(\mathbb{E}_{\mathrm{data}}[\varphi_j^2])^{1/2},$$

where $\widetilde{O}_d(1)$ *is a log-factor in* d*. Similar conclusions hold in finite sample: both terms admit the same bound in terms of the uniform bound on the features (cf. Remark 2.3), and when* $\varphi_j \sim \mathcal{N}(0, \sigma_j^2)$ *with* $\sigma_j \leqslant \sigma$ *for any* $j \in [d]$*, we have*

$$\|X\|_{1\times\infty}/n \leqslant C\sigma\sqrt{\log(dn/\delta)},$$

w.p. $1-\delta$*, with a similar bound for* $\|X^\top\|_{\infty\times 2}/\sqrt{n}$*, cf. (30).*

Complexity. $O(dn)$ a.o.'s are needed once to compute the row and column norms of $\widehat{X}$ in (33). Producing $\xi_U(p^*), \eta_{V,Y}(q^*)$ costs $O(dk+nk)$ a.o.'s, including the computation of the distibutions (33); given them, the proximal step has the same complexity as discussed in Sec. 2.

3.2 Full Sampling Scheme

In the *full sampling scheme*, sampling of the rows of U and $V-Y$ is augmented with a subsequent column sampling:

$$\begin{aligned}\xi_U(p,P) &= \widehat{X}\frac{e_i e_i^\top}{p_i}U\frac{e_l e_l^\top}{P_{il}},\\ \eta_{V,Y}(q,Q) &= \widehat{X}^\top\frac{e_j e_j^\top}{q_j}(V-Y)\frac{e_l e_l^\top}{Q_{jl}},\end{aligned} \qquad \textbf{(Full-SS)}$$

where $i \in [2d]$ and $j \in [n]$ are drawn from distributions $p \in \Delta_{2d}, q \in \Delta_n$ as before, and the rows of the matrices $P \in \Delta_k^{\otimes 2d}$ and $Q \in \Delta_k^{\otimes n}$ specify the conditional sampling distribution of the class $l \in [k]$ given i and j. Unbiasedness of these estimates is easy to verify. Next we derive the optimal sampling distributions and bound their variance proxies (refer to Appendix D.3 for the proof).

Proposition 3.2. *Let* $\|\cdot\|_{\mathscr{U}} = \|\cdot\|_{1\times 1}$*,* $\|\cdot\|_{\mathscr{V}} = \|\cdot\|_{2\times 1}$*. The optimal solutions* $(p*, P^*)$*,* (q^*, Q^*) *to*

$$\min_{p\in\Delta_{2d}, P\in\Delta_k^{\otimes 2d}} \mathbb{E}\|\xi_U(p,P)\|_{\mathscr{V}*}^2, \qquad \min_{q\in\Delta_n, Q\in\Delta_k^{\otimes n}} \mathbb{E}\|\eta_{V,Y}(q,Q)\|_{\mathscr{U}*}^2$$

are unique and given by

$$\begin{aligned} p_i^* &\propto \|\widehat{X}(:,i)\|_2\,\|U(i,:)\|_1, & P_{il}^* &\propto U_{il};\\ q_j^* &\propto \|\widehat{X}(j,:)\|_\infty\,\|V(j,:)-Y(j,:)\|_1, & Q_{jl}^* &\propto |V_{jl}-Y_{jl}|.\end{aligned} \qquad (37)$$

The respective variance proxies still admit the bounds (34).

Corollary 3.1. *The second bound of (36) in Theorem 3.1 remains true when we replace* (**Part-SS**) *with the sampling scheme* (**Full-SS**) *with distributions given by (37).*

4 Sublinear Algorithm for Multiclass ℓ_1-SVM

For the hinge loss (4), we provide a *sublinear* implementation of the bundle (**SMD**) + (**Full-SS**) with sampling distributions (37).

Lazy Updates. Note that although the estimates $\xi_U, \eta_{V,Y}$ produced in (**Full-SS**) are sparse (each contains a single non-zero column), the updates in (**SMD**), which can be expressed as (24)–(25) with $\xi_U, \eta_{V,Y}$ instead of the corresponding matrix products, are *dense*, and implementing them naively costs $O(dk+nk)$ a.o.'s. Fortunately, these updates have a special form: all elements in each row of U^t and V^t are simply rescaled with the same factor – except for at most two elements corresponding to a single non-zero element of $\eta_{V^t,Y}$ and at most two non-zero elements of $\xi_{U^t} - Y$ in this row. To exploit this fact, we perform "lazy" updates: instead of explicitly computing the actual iterates (U^t, V^t), we maintain the quadruple $(\widetilde{U}, \alpha, \widetilde{V}, \beta)$, where $\widetilde{U}, \widetilde{V}$ have the same dimensions as U, V, while $\alpha \in \mathbb{R}^{2d}$ and $\beta \in \mathbb{R}^n$ are the "scaling vectors", so that at any iteration t it holds

$$U^t(i,:) = \widetilde{U}(i,:) \cdot \alpha(i), \quad V^t(j,:) = \widetilde{V}(j,:) \cdot \beta(j) \tag{38}$$

for any row of U^t and V^t. Initializing with $(\widetilde{U}, \widetilde{V}) = (U, V)$, $\alpha = \mathbb{1}_{2d}$, $\beta = \mathbb{1}_n$, we can update the whole quadruple, while maintaining (38), by updating at most two elements in each row of $\widetilde{U}$ and $\widetilde{V}$, and encapsulating the overall scaling of rows in α and β. Clearly, this update requires only $O(d+n)$ operations once $\xi_{U^t}, \eta_{V^t,Y}$ have been drawn.

Sampling. Computing the distributions p^*, q^* from (37) requires the knowledge of $\|\widehat{X}(:,i)\|_2$ and $\|\widehat{X}(j,:)\|_\infty$ which can be precomputed in $O(dn)$ a.o.'s, and maintaining $O(d+n)$ norms π_i, ρ_j of the rows of U^t and $V^t - Y$ that can maintained in $O(1)$ a.o.'s each using (38). Thus, p^* and q^* can be updated in $O(d+n)$. Once it is done, we can sample $i^t \sim p^*$ and $j^t \sim q^*$, and then sample the class from P^* and Q^*, cf. (37), by computing only the i^t-th row of P^* and the j^t-th row of Q^*, both in $O(k)$ a.o.'s. Thus, the total complexity of producing $\xi_{U^t}, \eta_{V^t,Y}$ is $O(d+n+k)$.

Tracking the Averages. Similar "lazy" updates can be performed for the running averages of the iterates. Omitting the details, this requires $O(d+n)$ a.o.'s per iteration, plus post-processing of $O(dk+nk)$ a.o.'s.

The above ideas are implemented in Algorithm 1 whose correctness is formally shown in Appendix F (see also Sec. G for an additional discussion). Its close inspection shows the iteration cost of $O(d+n+k)$ a.o.'s, plus $O(dn+dk+nk)$ a.o.'s for pre/post-processing, and the memory complexiy of $O(dn+dk+nk)$. Moreover, the term $O(dk)$, which dominates in high-dimensional and highly multiclass problems, can be removed if one exploits sparsity of the corresponding primal solution to the ℓ_1-constrained problem (9), and outputs it directly, bypassing the explicit storage of $\widetilde{U}$ (see Appendix F for details). Note that when $n = O(\min(d,k))$, the resulting algorithm enters the sublinear regime after as few as $O(n)$ iterations.

5 Experiments

Sublinear Runtime. To illustrate the sublinear iteration cost of Algorithm 1, we consider the following experiment. Fixing $n = d = k$, we generate X with i.i.d. standard Gaussian entries, take U^o to be the identity matrix (thus very sparse), and generate the labels by $\operatorname{argmax}_{l\in[k]} x_j U^o_l + \frac{1}{\sqrt{d}}\mathcal{N}(0, I_d)$, where x_j's are the rows of X, and U^o_l's are the columns of U^o. This is repeated 10 times with $n = d = k$ increasing by a constant factor κ; each time we run Algorithm 1 for a fixed (large) number of iterations to dominate the cost of pre/post-processing, with $R_* = \|U^o\|_1$ and $\lambda = 10^{-3}$, and measure its runtime. We observe (see Tab. 1) that the runtime is proportional to κ, as expected.

Synthetic Data Experiment. We compare Algorithm 1 with two competitors: $\|\cdot\|_1$-composite stochastic subgradient method (SSM) for the primal problem (1), in which one uniformly samples one training example at a time (Shalev-Shwartz et al., 2011), leading to $O(dk)$ iteration cost; deterministic saddle-point Mirror Prox (MP) with geometry chosen as in Algorithm 1, for which we have $O(dnk)$ cost of iterations but $O(1/T)$ convergence in terms of the number of iterations. We generate data as in the previous experiment, fixing $n = d = k = 10^3$. The randomized algorithms are run 10 times for $T \in \{10^{m/2}, m = 1, ..., 12\}$ iterations with constant stepsize (we use stepsize (35) in Algorithm 1, choose the one recommended in Theorem 2.1 for MP, and use the theoretical stepsize for SSM, explicitly computing the variance of subgradients and the Lipschitz constant). Each time we compute the duality gap and the primal accuracy, and measure the runtime (see Fig. 1). We see that Algorithm 1 outperforms SSM, which might be the combined effect of sublinearity and our choice of geometry. It also outmatches MP up to

Algorithm 1 Sublinear Multiclass ℓ_1-Regularized SVM

Input: $X \in \mathbb{R}^{n\times d}$, $y \in [k]^{\otimes n}$, λ, R_*, $T \geqslant 1$, $\{\gamma_t\}_{t=0}^{T-1}$

Obtain $Y \in \Delta_k^{\otimes n}$ from the labels y; $\quad \widehat{X} \equiv [X, -X]$

$\alpha \leftarrow \mathbb{1}_{2d}$; $\quad \widetilde{U} \leftarrow \frac{R_* \mathbb{1}_{2d\times k}}{2dk}$; $\quad \beta \leftarrow \mathbb{1}_n$; $\quad \widetilde{V} \leftarrow \frac{\mathbb{1}_{n\times k}}{k}$

for $\imath = 1$ **to** $2d$ **do**

  $\sigma(\imath) \equiv \|\widehat{X}(:, \imath)\|_2$; $\quad \pi(\imath) \leftarrow \|\widetilde{U}(\imath, :)\|_1$

end for

for $\jmath = 1$ **to** n **do**

  $\tau(\jmath) \equiv \|\widehat{X}(\jmath, :)\|_\infty$; $\quad \rho(\jmath) \leftarrow \|\widetilde{V}(\jmath, :) - Y(\jmath, :)\|_1$

end for

 # Initialize machinery to track the cumulative sums

$U_\Sigma \leftarrow 0_{2d\times k}$; $\quad V_\Sigma \leftarrow 0_{n\times k}$ *# Cumulative sums*

$A \leftarrow 0_{2d}$; $\quad B \leftarrow 0_n$; $\quad A_{\text{pr}} \leftarrow 0_{2d\times k}$; $\quad B_{\text{pr}} \leftarrow 0_{n\times k}$

for $t = 0$ **to** $T - 1$ **do** *# (**SMD**) iterations*

  Draw $\boldsymbol{\jmath} \sim \tau \circ \rho$ *# $\circ$ is the elementwise product*

  Draw $\boldsymbol{l} \sim |\widetilde{V}(\boldsymbol{\jmath}, :) \cdot \beta_{\boldsymbol{\jmath}} - Y(\boldsymbol{\jmath}, :)|$

  $[U_\Sigma, A_{\text{pr}}, A] \leftarrow \textsc{TrackPrimal}(\widetilde{U}, U_\Sigma, A_{\text{pr}}, A, \alpha, \boldsymbol{l})$

  *# The only non-zero column of $\eta_{V^t,Y}$, cf. (**Full-SS**):*

  $\eta \leftarrow \widehat{X}(\boldsymbol{\jmath}, :) \cdot \frac{\sum_{\jmath=1}^n \tau(\jmath)\cdot\rho(\jmath)\cdot\text{sign}[\beta_\jmath \cdot V(\jmath,\boldsymbol{l}) - Y(\jmath,\boldsymbol{l})]}{\tau(\boldsymbol{\jmath})}$

  $[\widetilde{U}, \alpha, \pi] \leftarrow \textsc{UpdatePrimal}(\widetilde{U}, \alpha, \pi, \eta, \boldsymbol{l}, \gamma_t, \lambda, R_*)$

  Draw $\boldsymbol{\imath} \sim \sigma \circ \pi$

  Draw $\boldsymbol{\ell} \sim \widetilde{U}(\boldsymbol{\imath}, :)$

  $[V_\Sigma, B_{\text{pr}}, B] \leftarrow \textsc{TrackDual}(\widetilde{V}, V_\Sigma, B_{\text{pr}}, B, \beta, \boldsymbol{\ell}, y)$

  *# The only non-zero column of ξ_{U^t}, cf. (**Full-SS**):*

  $\xi \leftarrow \widehat{X}(:, \boldsymbol{\imath}) \cdot \frac{\sum_{\imath=1}^{2d} \sigma(\imath)\cdot\pi(\imath)}{\sigma(\boldsymbol{\imath})}$

  $[\widetilde{V}, \beta, \rho] \leftarrow \textsc{UpdateDual}(\widetilde{V}, Y, \beta, \rho, \xi, \boldsymbol{\ell}, y, \gamma_t)$

end for

for $l = 1$ **to** k **do** *# Postprocessing of cumulative sums*

  $U_\Sigma(:, l) \leftarrow U_\Sigma(:, l) + \widetilde{U}(:, l) \circ (\alpha + A - A_{\text{pr}}(:, l))$

  $V_\Sigma(:, l) \leftarrow V_\Sigma(:, l) + \widetilde{V}(:, l) \circ (\beta + B - B_{\text{pr}}(:, l))$

end for

Output: $\frac{1}{T+1} U_\Sigma, \frac{1}{T+1} V_\Sigma$ *# Averages $(\bar{U}^{T+1}, \bar{V}^{T+1})$*

Procedure 1 UpdatePrimal

Input: $\widetilde{U} \in \mathbb{R}^{2d \times k}$, $\alpha, \pi, \eta \in \mathbb{R}^{2d}$, $l \in [k]$, γ, λ, R_*

$L \equiv \log(2dk)$
for $i = 1$ **to** $2d$ **do**
  $\mu_i = \pi_i - \alpha_i \cdot \widetilde{U}(i,l) \cdot (1 - e^{-2\gamma L R_* \eta_i / n})$
end for
$M = \sum_{i=1}^{2d} \mu_i$
$\nu = \min\{e^{-2\gamma L R_* \lambda}, R_*/M\}$
for $i = 1$ **to** $2d$ **do**
  $\widetilde{U}(i,l) \leftarrow \widetilde{U}(i,l) \cdot e^{-2\gamma L R_* \eta_i / n}$
  $\alpha_i^+ = \nu \cdot \alpha_i$
  $\pi_i^+ = \nu \cdot \mu_i$
end for

Output: $\widetilde{U}, \alpha^+, \pi^+$

Procedure 2 UpdateDual

Input: $\widetilde{V}, Y \in \mathbb{R}^{n \times k}$, $\beta, \rho, \xi \in \mathbb{R}^n$, $\ell \in [k]$, $y \in [k]^{\otimes n}$, γ

$\theta = e^{-2\gamma \log(k)}$
for $j = 1$ **to** n **do**
  $\omega_j = e^{2\gamma \log(k) \xi_j}$
  $\varepsilon_j = e^{-2\gamma \log(k) Y(j,\ell)}$
  $\chi_j = 1 - \beta_j \cdot \widetilde{V}(j,\ell) \cdot (1 - \omega_j \cdot \varepsilon_j)$
  if $\ell \neq y_j$ **then** *# not the actual class of j drawn*
    $\chi_j \leftarrow \chi_j - \beta_j \cdot \widetilde{V}(j, y_j) \cdot (1 - \theta)$
  end if
  $\beta_j^+ = \beta_j / \chi_j$
  $\widetilde{V}(j,\ell) \leftarrow \widetilde{V}(j,\ell) \cdot \omega_j \cdot \varepsilon_j$
  $\widetilde{V}(j, y_j) \leftarrow \widetilde{V}(j, y_j) \cdot \omega_j \cdot \theta$
  $\rho_j^+ = 2 - 2\beta_j^+ \cdot \widetilde{V}(j, y_j)$
end for

Output: $\widetilde{V}, \beta^+, \rho^+$

Procedure 3 TrackPrimal

Input: $\widetilde{U}, U_\Sigma, A_{\mathrm{pr}} \in \mathbb{R}^{2d \times k}$, $A, \alpha \in \mathbb{R}^{2d}$, $l \in [k]$

for $i = 1$ **to** $2d$ **do**
  $U_\Sigma(i,l) \leftarrow U_\Sigma(i,l) + \widetilde{U}(i,l) \cdot (A_i + \alpha_i - A_{\mathrm{pr}}(i,l))$
  $A_{\mathrm{pr}}(i,l) \leftarrow A_i + \alpha_i$
  $A_i \leftarrow A_i + \alpha_i$
end for

Output: U_Σ, A_{pr}, A

$n = d = k$	400	800	1600	3200	6400
$T = 10^4$	1.17	2.07	4.27	7.55	15.56
$T = 2 \cdot 10^4$	2.47	4.27	8.74	14.65	30.77

Table 1: Runtime (in seconds) of Algorithm 1 in a synthetic data experiment.

Procedure 4 TRACKDUAL

Input: $\widetilde{V}, V_\Sigma, B_{\rm pr} \in \mathbb{R}^{n\times k}$, $B, \beta \in \mathbb{R}^n$, $\ell \in [k]$, $y \in [k]^{\otimes n}$

for $j = 1$ **to** n **do**
  for $l \in \{\ell, y_j\}$ **do** *# $\{\ell, y_j\}$ has 1 or 2 elements*
    $V_\Sigma(j,l) \leftarrow V_\Sigma(j,l) + \widetilde{V}(j,l) \cdot (B_j + \beta_j - B_{\rm pr}(j,l))$
    $B_{\rm pr}(j,l) \leftarrow B_j + \beta_j$
  end for
  $B_j \leftarrow B_j + \beta_j$
end for

Output: V_Σ, $B_{\rm pr}$, B

high accuracy due to the sublinear effect (MP eventually "wins" because of its $O(1/T)$ rate).[3]

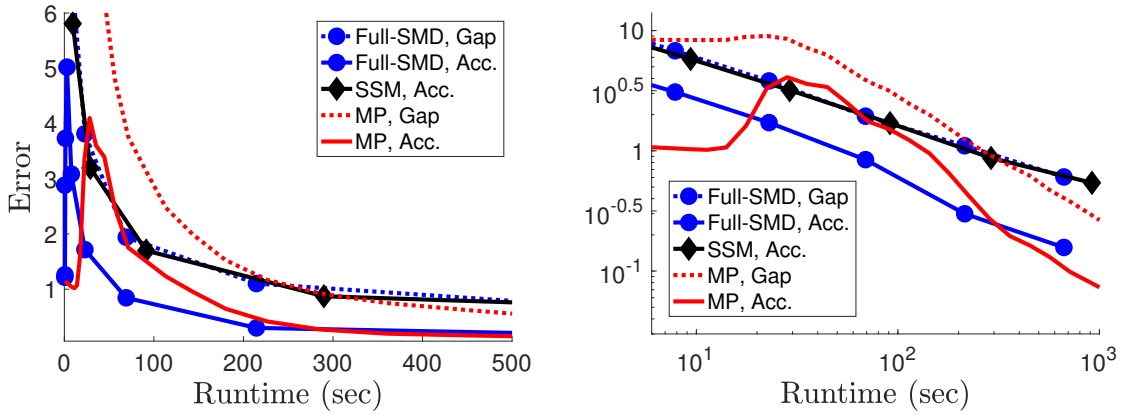

Figure 1: Primal accuracy and duality gap (when available) for Algorithm 1, stochastic subgradient method (SSM), and Mirror Prox (MP) with exact gradients, on a synthetic data benchmark, in the natural scale (left) and the log-log scale (right).

6 Conclusion and Perspectives

We proposed efficient algorithms based on stochastic mirror descent with entropy-type potentials, that allows to train ℓ_1-regularized multiclass linear classifiers in the case when the loss admits an explicit Fenchel-type representation, by reducing the finite-sum minimization problem to its saddle-point equivalent. In particular, in the case of the multiclass hinge loss we were able to construct a sublinear algorithm with the cost $O(d+n+k)$ of iterations, which was possible due to the multiplicative form of the updates. We envision the following directions for future work.

- It would be interesting to investigate whether out approach can also yield sublinear algorithms for other Fenchel-Young losses, in particular for the multiclass logistic model (3), which is widely used in Natural Language Processing (NLP) problems, where d, n and k are on order of millions or even billions (Chelba et al., 2013; Partalas et al., 2015).
- It would be useful to implement more flexible stepsizes, including the online stepsize search in the vein of Juditsky and Nemirovski (2011a), and to conduct larger scale experiments, including those on real data.

[3] The codes of our experiments are available online at `https://github.com/flykiller/sublinear-svm`.

- According to our result in Theorem 3.1 and Corollary 3.1, the loss of accuracy due to sampling is negligible when the data is light-tailed, but otherwise, the additional error due to sampling might be significant (see Remark 3.1). On the other hand, in the case of (non-composite) bilinear problems with vector variables, Juditsky and Nemirovski (2011b, Sec. 2.5.2.3) propose a technique of transforming the problem to an equivalent one, for which the loss of accuracy due to sampling is *always* tolerable. Extending their technique to our situation is non-trivial, and could be a worthwhile direction for future research.

Acknowledgments

DB and FB acknowledge support from the European Research Council (grant SEQUOIA 724063). DB was partly funded from the European Unions H2020 Framework Programme (H2020-MSCA-ITN-2014) under grant agreement №642685 MacSeNet. DO was supported by the ERCIM Alain Bensoussan Fellowship. We would like to thank Anatoli Juditsky for interesting discussions related to this work.

References

A. Beck and M. Teboulle. Mirror descent and nonlinear projected subgradient methods for convex optimization. *Operations Research Letters*, 31(3):167–175, 2003.

M. Blondel, A. F. T. Martins, and V. Niculae. Learning classifiers with Fenchel-Young losses: Generalized entropies, margins, and algorithms. *arXiv preprint arXiv:1805.09717*, 2018.

P. Bühlmann and S. Van De Geer. *Statistics for high-dimensional data: methods, theory and applications*. Springer Science & Business Media, 2011.

C. Chelba, T. Mikolov, M. Schuster, Q. Ge, T. Brants, P. Koehn, and T. Robinson. One billion word benchmark for measuring progress in statistical language modeling. *arXiv preprint arXiv:1312.3005*, 2013.

K. L. Clarkson, E. Hazan, and D. P. Woodruff. Sublinear optimization for machine learning. *Journal of the ACM (JACM)*, 59(5):23, 2012.

A. Defazio, F. Bach, and S. Lacoste-Julien. SAGA: A fast incremental gradient method with support for non-strongly convex composite objectives. In *Advances in neural information processing systems*, pages 1646–1654, 2014.

J. C. Duchi, S. Shalev-Shwartz, Y. Singer, and A. Tewari. Composite objective mirror descent. In *COLT*, pages 14–26, 2010.

D. Garber and E. Hazan. Approximating semidefinite programs in sublinear time. In *Advances in Neural Information Processing Systems*, pages 1080–1088, 2011.

D. Garber and E. Hazan. Sublinear time algorithms for approximate semidefinite programming. *Mathematical Programming*, 158(1-2):329–361, 2016.

M. D. Grigoriadis and L. G. Khachiyan. A sublinear-time randomized approximation algorithm for matrix games. *Operations Research Letters*, 18(2):53–58, 1995.

E. Hazan, T. Koren, and N. Srebro. Beating SGD: Learning SVMs in sublinear time. In *Advances in Neural Information Processing Systems*, pages 1233–1241, 2011.

R. Johnson and T. Zhang. Accelerating stochastic gradient descent using predictive variance reduction. In *Advances in Neural Information Processing Systems*, pages 315–323, 2013.

A. Juditsky and A. Nemirovski. First-order methods for nonsmooth convex large-scale optimization, I: General purpose methods. *Optimization for Machine Learning*, pages 121–148, 2011a.

A. Juditsky and A. Nemirovski. First order methods for nonsmooth convex large-scale optimization, ii: utilizing problems structure. *Optimization for Machine Learning*, pages 149–183, 2011b.

J. H. B. Kemperman. On the optimum rate of transmitting information. In *Probability and information theory*, pages 126–169. Springer, 1969.

B. Laurent and P. Massart. Adaptive estimation of a quadratic functional by model selection. *The Annals of Statistics*, 28(5):1302–1338, 2000.

A. Mokhtari, H. Daneshmand, A. Lucchi, T. Hofmann, and A. Ribeiro. Adaptive Newton method for empirical risk minimization to statistical accuracy. In *Advances in Neural Information Processing Systems*, pages 4062–4070, 2016.

A. Nemirovski. Prox-method with rate of convergence $O(1/t)$ for variational inequalities with Lipschitz continuous monotone operators and smooth convex-concave saddle point problems. *SIAM Journal on Optimization*, 15(1):229–251, 2004.

A. Nemirovski, S. Onn, and U. G. Rothblum. Accuracy certificates for computational problems with convex structure. *Mathematics of Operations Research*, 35(1):52–78, 2010.

A. Nemirovsky and D. Yudin. *Problem complexity and method efficiency in optimization.* Chichester, 1983.

Y. Nesterov and A. Nemirovski. On first-order algorithms for l_1/nuclear norm minimization. *Acta Numerica*, 22:509–575, 2013.

D. Ostrovskii and Z. Harchaoui. Efficient first-order algorithms for adaptive signal denoising. In *Proceedings of the 35th ICML conference*, volume 80, pages 3946–3955, 2018.

B. Palaniappan and F. Bach. Stochastic variance reduction methods for saddle-point problems. In *Advances in Neural Information Processing Systems*, pages 1416–1424, 2016.

I. Partalas, A. Kosmopoulos, N. Baskiotis, T. Artieres, G. Paliouras, E. Gaussier, I. Androutsopoulos, M.-R. Amini, and P. Galinari. LSHTC: A benchmark for large-scale text classification. *arXiv preprint arXiv:1503.08581*, 2015.

M. Schmidt, N. Le Roux, and F. Bach. Minimizing finite sums with the stochastic average gradient. *Mathematical Programming*, 162(1-2):83–112, 2017.

S. Shalev-Shwartz and S. Ben-David. *Understanding machine learning: From theory to algorithms*. Cambridge university press, 2014.

S. Shalev-Shwartz and T. Zhang. Stochastic dual coordinate ascent methods for regularized loss minimization. *Journal of Machine Learning Research*, 14(Feb):567–599, 2013.

S. Shalev-Shwartz, Y. Singer, N. Srebro, and A. Cotter. Pegasos: Primal estimated sub-gradient solver for SVM. *Mathematical programming*, 127(1):3–30, 2011.

Z. Shi, X. Zhang, and Y. Yu. Bregman divergence for stochastic variance reduction: saddle-point and adversarial prediction. In *Advances in Neural Information Processing Systems*, pages 6031–6041, 2017.

M. Sion. On general minimax theorems. *Pacific Journal of Mathematics*, 8(1):171–176, 1958.

S. Sra. Fast projections onto mixed-norm balls with applications. *Data Mining and Knowledge Discovery*, 25(2):358–377, 2012.

L. Xiao, A. W. Yu, Q. Lin, and W. Chen. DSCOVR: Randomized primal-dual block coordinate algorithms for asynchronous distributed optimization. *arXiv preprint arXiv:1710.05080*, 2017.

Y.-L. Yu. The strong convexity of von Neumann's entropy. *Unpublished note*, June 2013. URL `http://www.cs.cmu.edu/~yaoliang/mynotes/sc.pdf`.

A Motivation for Multiclass Hinge Loss

We justify the multiclass extension (4) of the hinge loss due to (Shalev-Shwartz and Ben-David, 2014). In the binary case, the hinge loss is

$$\frac{1}{n}\sum_{i=1}^{n}\left[\max(0, 1-\widetilde{y}_i u^\top x_i)\right],$$

where $u \in \mathbb{R}^d$ and $\widetilde{y}_i \in \{-1, 1\}$. Introducing $y = e_{\widetilde{y}} \in \{e_{-1}, e_1\}$ where e_j is the j-th standard basis vector (the dimensions of space are symbolically indexed in $\{-1, 1\}$), and putting $u_1 = -u_{-1} = \frac{u}{2}$, we can rewrite the loss as

$$\max(0, 1-\widetilde{y}u^\top x) = \max_{k\in\{1,-1\}}\left\{\mathbb{1}\{e_k \neq y\} + u_k^\top x - u_{\widetilde{y}}^\top x\right\}.$$

The advantage of this reformulation is that we can naturally pass to the multiclass case, by replacing the set $\{-1, 1\}$ with $\{1, ..., K\}$ and introducing $u_1, ..., u_K \in \mathbb{R}^d$ without any restrictions:

$$\begin{aligned}\max_{k\in\{1,\dots,K\}}\left\{\mathbb{1}\{e_k \neq y\} + u_k^\top x - u_{\widetilde{y}}^\top x\right\} &= \max_{v\in\{e_1,\dots,e_K\}}\left\{\mathbb{1}\{v \neq y\} + \sum_{l=1}^{K}(v[l]-y[l])u_l^\top x\right\}\\ &= \max_{v\in\{e_1,\dots,e_K\}}\left\{\mathbb{1}\{v \neq y\} + (v-y)^\top U^\top x\right\} =: \ell(U,(x,y)),\end{aligned}$$

where $a[l]$ denotes the l-th element of a column-vector a, and $U \in \mathbb{R}^{d\times K}$ has u_l as its l-th column. Finally, we can rewrite $\ell(U,(x,y))$ as follows:

$$\ell(U,(x,y)) = \max_{v\in\Delta_K}\left\{1 - v^\top y + (v-y)^\top U^\top x\right\}.$$

This is because we maximize an affine function of v, and $1 - v^\top y = \mathbb{1}\{v \neq y\}$ at the vertices. Thus, we obtain the Fechel dual representation of the multiclass hinge loss. Adding the regularization term $\|\cdot\|_{\mathscr{U}}$, we also arrive at the saddle-point problem

$$\min_{U\in\mathbb{R}^{d\times k}}\max_{V\in\Delta_k^{\otimes n}} 1 - \frac{1}{n}\mathrm{tr}[V^\top Y] + \frac{1}{n}\mathrm{tr}\left[(V-Y)^\top XU\right] + \lambda\|U\|_{\mathscr{U}}.$$

B General Accuracy Bounds for Composite Saddle-Point Mirror Descent

Deterministic Case. Here we provide general accuracy bounds which are instantiated in Theorems 2.1 and 3.1. Below we outline the general setting that encompasses, in particular, the case of (11) solved via (**MD**) with initialization (**Init**).

- We consider a convex-concave saddle-point problem

$$\min_{U\in\mathcal{U}}\max_{V\in\mathcal{V}} f(U,V)$$

with a composite objective

$$f(U,V) = \Phi(U, V-Y) + \Upsilon(U) - \mathcal{F}(V),$$

where

$$\Phi(U,V) = \frac{1}{n} V^\top X U$$

is a bilinear function, and $\Upsilon(U), \mathcal{F}(V)$ are convex "simple" terms. Moreover, we assume that the primal feasible set $\mathcal{U}$ belongs to the $\|\cdot\|_{\mathscr{U}}$-norm ball with radius R_*, the dual constraint set $\mathcal{V}$ belongs to the $\|\cdot\|_{\mathscr{V}}$-norm ball with radius $R_{\mathcal{V}}$, and $\|Y\|_{\mathscr{V}} \leqslant R_{\mathcal{V}}$.[4] To simplify the results, we make the assumption (satisfied in all known to us situations):

$$\Omega_{\mathcal{U}} \geqslant R_*^2, \quad \Omega_{\mathcal{V}} \geqslant R_{\mathcal{V}}^2. \tag{39}$$

- Recall that the vector field of partial gradients of $\Psi(U,V) := \Phi(U, V-Y)$ is

$$\begin{aligned} G(W) :&= (\nabla_U \Psi(U,V), -\nabla_V \Psi(U,V)) \\ &= \frac{1}{n}(X^\top (V-Y), -XU) \end{aligned} \tag{40}$$

- Given the partial proximal setups $(\|\cdot\|_{\mathscr{U}}, \phi_{\mathcal{U}}(\cdot))$ and $(\|\cdot\|_{\mathscr{V}}, \phi_{\mathcal{V}}(\cdot))$, we run Composite Mirror Descent (22) on the vector field $G(W)$ with the joint penalty term

$$h(W) = \Upsilon(U) + \mathcal{F}(V),$$

the "balanced" joint potential given by (21), and stepsizes γ_t.

We now provide the convergence analysis of Mirror Descent, extending the argument of Lem. 1 in (Duchi et al., 2010) to composite saddle-point optimization.

Theorem B.1. *In the above setting, let* $(\bar{U}^T, \bar{V}^T) = \frac{1}{T}\sum_{t=0}^{T-1}(U^t, V^t)$ *be the average of the first* T *iterates of the composite Mirror Descent* (22) *with constant stepsize*

$$\gamma_t \equiv \frac{1}{\mathcal{L}_{\mathscr{U},\mathscr{V}}\sqrt{5T\Omega_{\mathcal{U}}\Omega_{\mathcal{V}}}},$$

where

$$\mathcal{L}_{\mathscr{U},\mathscr{V}} := \frac{1}{n}\sup_{\|U\|_{\mathscr{U}}\leq 1} \|XU\|_{\mathscr{V}^*}.$$

Then we have the following guarantee for the duality gap of $f(\cdot)$*:*

$$\texttt{Gap}(\bar{U}^T, \bar{V}^T) \leqslant \frac{2\sqrt{5}\mathcal{L}_{\mathscr{U},\mathscr{V}}\sqrt{\Omega_{\mathcal{U}}\Omega_{\mathcal{V}}}}{\sqrt{T}} + \frac{\Upsilon(U^0) - \min_{U\in\mathcal{U}}\Upsilon(U)}{T} + \frac{\mathcal{F}(V^0) - \min_{V\in\mathcal{V}}\mathcal{F}(V)}{T}.$$

Moreover, if one of the functions $\Upsilon(U)$, $\mathcal{F}(V)$ *is affine, the corresponding* $O(1/T)$ *error term vanishes from the bound.*

Proof. **1°**. We begin by introducing the norm for $W = (U,V)$:

$$\|W\|_{\mathscr{W}} = \sqrt{\frac{\|U\|_{\mathscr{U}}^2}{2\Omega_{\mathcal{U}}} + \frac{\|V\|_{\mathscr{V}}^2}{2\Omega_{\mathcal{V}}}}, \tag{41}$$

[4] Note that the linear term $\frac{1}{n}Y^\top XU$ can be absorbed into the simple term $\Upsilon(U)$, which will slightly improve the bound in Theorem B.1. However, this improvement is impossible in the stochastic version of the algorithm where we sample the linear form $Y^\top X$ but not the gradient of $\Upsilon(U)$.

and its dual norm defined for $G = (G_U, G_V)$ with $G_U \in \mathbb{R}^{d\times k}$ and $G_V \in \mathbb{R}^{n\times k}$:

$$\|G\|_{\mathscr{W}^*} = \sqrt{2\Omega_{\mathcal{U}}\|G_U\|^2_{\mathscr{U}^*} + 2\Omega_{\mathcal{V}}\|G_V\|^2_{\mathscr{V}^*}}, \tag{42}$$

where $\|\cdot\|_{\mathscr{U}^*}$ and $\|\cdot\|_{\mathscr{V}^*}$ are the dual norms for $\|\cdot\|_{\mathscr{U}}$ and $\|\cdot\|_{\mathscr{V}}$ correspondingly. We now make a few observations. First, the joint potential $\phi_{\mathcal{W}}(W)$ given by (21) is 1-strongly convex with respect to the norm $\|\cdot\|_{\mathscr{W}}$. Second, we can compute the potential difference corresponding to $\phi_{\mathcal{W}}$:

$$\Omega_{\mathcal{W}} := \max_{W\in\mathcal{W}} \phi_{\mathcal{W}}(W) - \min_{W\in\mathcal{W}} \phi_{\mathcal{W}}(W) = 1 \tag{43}$$

Finally, by (40) and (26) we have

$$\max_{W\in\mathcal{W}} \|G_U(W)\|_{\mathscr{U}^*} \leqslant 2\mathcal{L}_{\mathscr{U},\mathscr{V}} R_{\mathcal{V}}, \quad \max_{W\in\mathcal{W}} \|G_V(W)\|_{\mathscr{V}^*} \leqslant \mathcal{L}_{\mathscr{U},\mathscr{V}} R_*,$$

combining which with (39) we bound the $\|\cdot\|_{\mathscr{W}^*}$-norm of $G(W)$ on $\mathcal{W}$:

$$\max_{W\in\mathcal{W}} \|G(W)\|_{\mathscr{W}^*} \leqslant \sqrt{10}\mathcal{L}_{\mathscr{U},\mathscr{V}}\sqrt{\Omega_{\mathcal{U}}\Omega_{\mathcal{V}}}. \tag{44}$$

2°. We now follow the convergence analysis of composite Mirror Descent, see (Duchi et al., 2010), extending it to convex-concave objectives. By the convexity properties of $\Psi(U,V) = \Phi(U, V-Y)$, for any $(\bar{U},\bar{V}) \in \mathcal{W}$ and $(U,V)\in\mathcal{W}$ it holds

$$\begin{aligned}\Psi(\bar{U},V) - \Psi(U,\bar{V}) &= \Psi(\bar{U},V) - \Psi(\bar{U},\bar{V}) + \Psi(\bar{U},\bar{V}) - \Psi(U,\bar{V})\\ &\leqslant \left\langle \nabla_U\Psi(\bar{U},\bar{V}), \bar{U}-U\right\rangle - \left\langle \nabla_V\Psi(\bar{U},\bar{V}), \bar{V}-V\right\rangle\\ &= \left\langle G(\bar{W}), \bar{W}-W\right\rangle,\end{aligned}$$

Let $W^t = (U^t, V^t)$ be the t-th iterate of (22) for $t \geq 1$. By convexity of $\Upsilon(U)$ and $\mathcal{F}(V)$, and denoting $h(W) = \Upsilon(U) + \mathcal{F}(V)$, we have, for any $W = [U,V]$, that

$$\begin{aligned}\Psi(U^{t-1},V) - \Psi(U,V^{t-1}) + h(W^t) - h(W)&\\ \leqslant \left\langle G(W^{t-1}), W^{t-1}-W\right\rangle + \left\langle \partial h(W^t), W^t - W\right\rangle&.\end{aligned} \tag{45}$$

Let us now bound the right-hand side. Note that the first-order optimality condition for (22) (denoting $\phi(\cdot) := \phi_{\mathcal{W}}(\cdot)$ the joint potential) writes[5]

$$\left\langle \gamma_t[G(W^{t-1}) + \partial h(W^t)] + \nabla\phi(W^t) - \nabla\phi(W^{t-1}), W^t - W\right\rangle \leqslant 0. \tag{46}$$

Combining this with (45), we get

$$\begin{aligned}\gamma_t[\Psi(U^{t-1},V) - \Psi(U,V^{t-1}) + h(W^t) - h(W)] \leqslant \left\langle \nabla\phi(W^{t-1}) - \nabla\phi(W^t), W^t - W\right\rangle&\\ + \gamma_t\left\langle G(W^{t-1}), W^{t-1} - W^t\right\rangle&.\end{aligned} \tag{47}$$

By the well-known identity,

$$\left\langle \nabla\phi(W^{t-1}) - \nabla\phi(W^t), W^t - W\right\rangle = D_\phi(W, W^{t-1}) - D_\phi(W, W^t) - D_\phi(W^t, W^{t-1}), \tag{48}$$

see, e.g., Beck and Teboulle (2003, Lemma 4.1). On the other hand, by the Fenchel-Young inequality we have

$$\begin{aligned}\gamma_t\left\langle G(W^{t-1}), W^{t-1} - W^t\right\rangle &\leqslant \frac{\gamma_t^2\|G(W^{t-1})\|^2_{\mathscr{W}^*}}{2} + \frac{\|W^{t-1}-W^t\|^2_{\mathscr{W}}}{2}\\ &\leqslant 5\gamma_t^2\mathcal{L}^2_{\mathscr{U},\mathscr{V}}\Omega_{\mathcal{U}}\Omega_{\mathcal{V}} + D_\phi(W^t, W^{t-1}),\end{aligned} \tag{49}$$

where we used (44) and 1-strong convexity of $\phi(\cdot)$ with respect to $\|\cdot\|_{\mathscr{W}}$. Thus, we obtain

$$\gamma_t[\Psi(U^{t-1},V)-\Psi(U,V^{t-1})+h(W^t)-h(W)] \leq D_\phi(W,W^{t-1})-D_\phi(W,W^t)+5\gamma_t^2\mathcal{L}^2_{\mathscr{U},\mathscr{V}}\Omega_{\mathcal{U}}\Omega_{\mathcal{V}}. \tag{50}$$

3°. Now, assuming the constant stepsize, by the convexity properties of $\Psi(\cdot,\cdot)$ and $h(\cdot)$ we obtain

$$\begin{aligned} f(\bar{U}^T,V)-f(U,\bar{V}^T) &= \Psi(\bar{U}^T,V)-\Psi(U,\bar{V}^T)+h(\bar{W}^T)-h(W) \\ &\leqslant \frac{1}{T}\sum_{t=1}^T \Psi(U^{t-1},V)-\Psi(U,V^{t-1})+h(W^{t-1})-h(W) \\ &\leqslant \frac{1}{T}\left(h(W^0)-h(W^T)+\frac{D_\phi(W,W^0)}{\gamma}+5T\gamma\mathcal{L}^2_{\mathscr{U},\mathscr{V}}\Omega_{\mathcal{U}}\Omega_{\mathcal{V}}\right) \\ &\leqslant \frac{1}{T}\left(h(W^0)-\min_{W\in\mathcal{W}}h(W)+\frac{1}{\gamma}+5T\gamma\mathcal{L}^2_{\mathscr{U},\mathscr{V}}\Omega_{\mathcal{U}}\Omega_{\mathcal{V}}\right). \end{aligned} \tag{51}$$

where for the third line we substituted (50), simplified the telescoping sum, and used that $D(W,W^T)\geqslant 0$, and in the last line we used $D(W,W^0)\leqslant\Omega_{\mathcal{W}}\leqslant 1$, cf. (43). The choice

$$\gamma=\frac{1}{\mathcal{L}_{\mathscr{U},\mathscr{V}}\sqrt{5T\Omega_{\mathcal{U}}\Omega_{\mathcal{V}}}},$$

results in the accuracy bound from the premise of the theorem:

$$\texttt{Gap}(\bar{U}^T,\bar{V}^T)\leqslant\frac{2\sqrt{5}\mathcal{L}_{\mathscr{U},\mathscr{V}}\sqrt{\Omega_{\mathcal{U}}\Omega_{\mathcal{V}}}}{\sqrt{T}}+\frac{h(W^0)-\min_{W\in\mathcal{W}}h(W)}{T}.$$

Finally, assume that one of the terms $\Upsilon(U)$, $\mathcal{F}(V)$ is affine – w.l.o.g. let it be $\Upsilon(U)$. Then, since $\nabla\Upsilon(U)$ is constant, $\partial h(W^t)=(\nabla\Upsilon(U^t),\partial\mathcal{F}(V^t))$ in (46) can be replaced with $(\nabla\Upsilon(U^{t-1}),\partial\mathcal{F}(V^t))$. Then in (50) we can replace $h(W^t)-h(W^0)$ with $\Upsilon(U^{t-1})-\Upsilon(U)+\mathcal{F}(V^t)-\mathcal{F}(V)$, implying that the term $h(W^0)-h(W^T)$ in the right-hand side of (51) gets replaced with $\mathcal{F}(V^0)-\mathcal{F}(V^t)$. The claim is proved. □

Stochastic Mirror Descent. We now consider the stochastic setting that allows to encompass (**SMD**). Stochastic Mirror Descent is given by

$$\begin{aligned} W^0 &= \min_{W\in\mathcal{W}}\phi_{\mathcal{W}}(W); \\ W^t &= \arg\min_{W\in\mathcal{W}}\left\{h(W)+\langle\Xi(W^{t-1}),W\rangle+\frac{1}{\gamma_t}D_{\phi_{\mathcal{W}}}(W,W^{t-1})\right\},\ t\geq 1, \end{aligned} \tag{52}$$

where

$$\Xi(W):=\frac{1}{n}(\eta_{V,Y},-\xi_U)$$

is the unbiased estimate of the first-order oracle $G(W)=\frac{1}{n}(X^\top(V-Y),-XU)$. Let us introduce the corresponding variance proxies (refer to the preamble of Section 3 for the discussion):

$$\sigma^2_{\mathcal{U}}=\frac{1}{n^2}\sup_{U\in\mathcal{U}}\mathbb{E}\left[\|XU-\xi_U\|^2_{\mathscr{V}^*}\right],\quad \sigma^2_{\mathcal{V}}=\frac{1}{n^2}\sup_{(V,Y)\in\mathcal{V}\times\mathcal{V}}\mathbb{E}\left[\left\|X^\top(V-Y)-\eta_{V,Y}\right\|^2_{\mathscr{U}^*}\right]. \tag{53}$$

We assume that the noises $G(W^{t-1})-\Xi(W^{t-1})$ are independent along the iterations of (52). In this setting, we prove the following generalization of Theorem B.1:

[5]Note that $\phi(W)$ is continuously differentiable in the interior of $\mathcal{W}$, and $\nabla\phi$ diverges on the boundary of $\mathcal{W}$, then the iterates are guaranteed to stay in the interior of $\mathcal{W}$ (Beck and Teboulle, 2003).

Theorem B.2. *Let $(\bar{U}^T, \bar{V}^T) = \frac{1}{T}\sum_{t=0}^{T-1}(U^t, V^t)$ be the average of the first T iterates of stochastic composite mirror descent (52) with constant stepsize*

$$\gamma_t \equiv \frac{1}{\sqrt{T}} \min\left\{ \frac{1}{\sqrt{10}\mathcal{L}_{\mathscr{U},\mathscr{V}}\sqrt{\Omega_{\mathcal{U}}\Omega_{\mathcal{V}}}}, \frac{1}{\sqrt{2}\sqrt{\Omega_{\mathcal{U}}\bar{\sigma}_{\mathcal{V}}^2 + \Omega_{\mathcal{V}}\bar{\sigma}_{\mathcal{U}}^2}} \right\},$$

where $\mathcal{L}_{\mathscr{U},\mathscr{V}}, \Omega_{\mathcal{U}}, \Omega_{\mathcal{V}}$ are the same as in Theorem B.2, and $\bar{\sigma}_{\mathcal{U}}^2, \bar{\sigma}_{\mathcal{V}}^2$ are the upper bounds for $\sigma_{\mathcal{U}}^2, \sigma_{\mathcal{V}}^2$, cf. (53). Then it holds

$$\begin{aligned}\mathbb{E}[\mathtt{Gap}(\bar{U}^T, \bar{V}^T)] \leqslant{}& \frac{2\sqrt{10}\mathcal{L}_{\mathscr{U},\mathscr{V}}\sqrt{\Omega_{\mathcal{U}}\Omega_{\mathcal{V}}}}{\sqrt{T}} + \frac{2\sqrt{2}\sqrt{\Omega_{\mathcal{U}}\bar{\sigma}_{\mathcal{V}}^2 + \Omega_{\mathcal{V}}\bar{\sigma}_{\mathcal{U}}^2}}{\sqrt{T}} \\ &+ \frac{\Upsilon(U^0) - \min_{U\in\mathcal{U}}\Upsilon(U)}{T} + \frac{\mathcal{F}(V^0) - \min_{V\in\mathcal{V}}\mathcal{F}(V)}{T},\end{aligned}$$

where $\mathbb{E}[\cdot]$ is the expectation over the randomness in (52). Moreover, if one of the functions $\Upsilon(U)$, $\mathcal{F}(V)$ is affine, the corresponding $O(1/T)$ term can be discarded.

Proof. The proof closely follows that of Theorem B.1. First, $\mathbf{1}^o$ remains unchanged. Then, in the first-order condition (46) one must replace $G(W^{t-1})$ with $\Xi(W^{t-1})$, which results in replacing (47) with

$$\begin{aligned}&\gamma_t[\Psi(U^{t-1}, V) - \Psi(U, V^{t-1}) + h(W^t) - h(W)] \\ &\quad\leqslant \left\langle \nabla\phi(W^{t-1}) - \nabla\phi(W^t), W^t - W\right\rangle \\ &\quad+ \gamma_t\left\langle \Xi(W^{t-1}), W^{t-1} - W^t\right\rangle \\ &\quad+ \gamma_t\left\langle G(W^{t-1}) - \Xi(W^{t-1}), W^{t-1} - W\right\rangle,\end{aligned}$$

where the last term has zero mean. The term $\gamma_t\left\langle \Xi(W^{t-1}), W^{t-1} - W^t\right\rangle$ can be bounded using Young's inequality, and 1-strong convexity of $\phi(\cdot)$, cf. (49):

$$\begin{aligned}&\gamma_t\left\langle \Xi(W^{t-1}), W^{t-1} - W^t\right\rangle \\ &\quad\leqslant \frac{\gamma_t^2\|\Xi(W^{t-1})\|_{\mathscr{W}^*}^2}{2} + \frac{\|W^{t-1} - W^t\|_{\mathscr{W}}^2}{2} \\ &\quad\leqslant \gamma_t^2\|G(W^{t-1})\|_{\mathscr{W}^*}^2 + \gamma_t^2\|\Xi(W^{t-1}) - G(W^{t-1})\|_{\mathscr{W}^*}^2 + D_\phi(W^t, W^{t-1}).\end{aligned}$$

Combining (44), (42), and (53), this implies

$$\gamma_t\left\langle \Xi(W^{t-1}), W^{t-1} - W^t\right\rangle \leqslant 10\gamma_t^2\mathcal{L}_{\mathscr{U},\mathscr{V}}^2\Omega_{\mathcal{U}}\Omega_{\mathcal{V}} + 2\gamma_t^2\left(\Omega_{\mathcal{U}}\bar{\sigma}_{\mathcal{V}}^2 + \Omega_{\mathcal{V}}\bar{\sigma}_{\mathcal{U}}^2\right) + D_\phi(W^t, W^{t-1}).$$

Using (48) and (43), this results in

$$\begin{aligned}\mathbb{E}[\mathtt{Gap}(\bar{U}^T, \bar{V}^T)] &= \mathbb{E}\left[\max_{(U,V)\in\mathcal{W}}\left\{f(\bar{U}^T, V) - f(U, \bar{V}^T)\right\}\right] \\ &\leqslant \frac{1}{T}\left[h(W^0) - \min_{W\in\mathcal{W}} h(W) + \frac{1}{\gamma} + 2T\gamma\left(5\mathcal{L}_{\mathscr{U},\mathscr{V}}^2\Omega_{\mathcal{U}}\Omega_{\mathcal{V}} + \Omega_{\mathcal{U}}\bar{\sigma}_{\mathcal{V}}^2 + \Omega_{\mathcal{V}}\bar{\sigma}_{\mathcal{U}}^2\right)\right];\end{aligned}$$

note that maximization on the left is *under* the expectation (and not vice versa) because the right hand side is independent from $W = (U, V)$. Choosing γ to balance the terms, we arrive at the desired bound. Finally, improvement in the case of affine $\Upsilon(U)$, $\mathcal{F}(V)$ is obtained in the same way as in Theorem B.2. □

C Auxiliary Lemmas

Lemma C.1. *Let* $X^0 \in \mathbb{R}^n_+$ *and* $X^1 = \arg\min_{\substack{\|X\|_1 \leqslant R \\ X \in \mathbb{R}^n_+}} \left\{ C_1\|X\|_1 + \langle S, X\rangle + C_2 \sum_{i=1}^n X_i \log \frac{X_i}{X_i^0} \right\}$. *Then,*

$$X_i^1 = \rho \cdot \frac{X_i^0 \cdot \exp(-S_i/C_2)}{M},$$

where $M = \sum_{j=1}^n X_j^0 \cdot \exp(-S_j/C_2)$ *and* $\rho = \min(M \cdot e^{-\frac{C_1+C_2}{C_2}}, R)$.

Proof. Clearly, we have

$$X^1 = \arg\min_{r \leqslant R} \min_{\substack{\|X\|_1 = r \\ X \in \mathbb{R}^n_+}} \left\{ C_1 r + \langle S, X\rangle + C_2 \sum_{i=1}^n X_i \log \frac{X_i}{X_i^0} \right\}.$$

Let us first do the internal minimization. By simple algebra, the first-order optimality condition for the Lagrangian dual problem (with constraint $\|X\|_1 = R$) amounts to

$$S_i + C_2 + C_2 \log X_i^1 - C_2 \log X_i^0 + \kappa = 0,$$

where κ is Lagrange multiplier, and $\sum_{i=1}^n X_i^1 = r$. Equivalently,

$$X_i^1 = X_i^0 \cdot \exp\left(- \left(\frac{\kappa + C_2 + S_i}{C_2} \right) \right),$$

that is,

$$X_i^1 = r \cdot \frac{X_i^0 \cdot \exp(-S_i/C_2)}{\sum_j X_j^0 \cdot \exp(-S_j/C_2)}.$$

Denoting $D_j = \exp(-S_j/C_2)$ and $M = \sum_j X_j^0 \cdot D_j$ and substituting for X_1 in the external minimization problem, we arrive at

$$\rho = \arg\min_{r \leqslant R} \left\{ C_1 r + r \sum_i \frac{X_i^0 D_i S_i}{M} + C_2 \cdot r \sum_i \frac{X_i^0 D_i}{M} \cdot \log \left[\frac{r \cdot D_i}{M} \right] \right\}.$$

One can easily verify that the counterpart of this minimization problem with $R = \infty$ has a unique stationary point $r^* = M \cdot e^{-\frac{C_1+C_2}{C_2}} > 0$. As the minimized function is convex, the minimum is attained at the point $\rho = \min(r^*, R)$. □

Lemma C.2. *Given* $X \in \mathbb{R}^{n\times d}$ *and mixed norms (cf. (14))* $\|\cdot\|_{p_U^1 \times p_U^2}$ *on* $\mathbb{R}^{d\times k}$ *and* $\|\cdot\|_{q_V^1 \times q_V^2}$ *on* $\mathbb{R}^{n\times k}$ *with* $p_U^2 \leqslant q_V^2$, *one has*

$$\sup_{\|U\|_{p_U^1 \times p_U^2} \leqslant 1} \left\{ \sum_{i=1}^d \|X(:,i)\|_{q_V^1} \cdot \|U(i,:)\|_{q_V^2} \right\} = \|X^\top\|_{q_U^1 \times q_V^1},$$

where q_U^1 *is the conjugate of* p_U^1, *i.e.,* $1/p_U^1 + 1/q_U^1 = 1$.

Proof. First assume $p_U^2 = q_V^2$. Let $a_i = \|X(:,i)\|_{q_V^1}$, $u_i = \|U(i,:)\|_{q_V^2}$, $1 \leqslant i \leqslant d$. Then,

$$\sup_{\|U\|_{p_U^1 \times p_U^2} \leqslant 1} \left\{ \sum_{i=1}^d \|X(:,i)\|_{q_V^1} \cdot \|U(i,:)\|_{q_V^2} \right\} = \sup_{\|u\|_{p_U^1} \leqslant 1} \sum_{i=1}^d a_i u_i = \|a\|_{q_U^1} = \|X^\top\|_{q_U^1 \times q_V^1}.$$

Now let $p_U^2 < q_V^2$. Then, for any $i \leq d$ one has $\|U(i,:)\|_{q_V^2} < \|U(i,:)\|_{p_U^2}$ unless $U(i,:)$ has a single non-zero element, in which case $\|U(i,:)\|_{q_V^2} = \|U(i,:)\|_{p_U^2}$. Hence, the supremum must be attained on such U, for which the previous argument applies. □

Lemma C.3. *In the setting of Lemma C.2, for any $q_U^1 \geqslant 1$ and $q_U^2 \geqslant 1$ one has:*

$$\sup_{\|V\|_{\infty \times 1} \leqslant 1} \left\{ \sum_{i=1}^n \|X^\top(:,i)\|_{q_U^1} \cdot \|V(i,:)\|_{q_U^2} \right\} = \|X\|_{1 \times q_U^1}.$$

Proof. The claim follows by instatiating Lemma C.2. □

D Deferred Proofs

D.1 Proof of Proposition 2.1

By (26), and verifying that the dual norm to $\|\cdot\|_{2\times 1}$ is $\|\cdot\|_{2\times\infty}$, we have

$$\mathcal{L}_{\mathscr{U},\mathscr{V}} = \sup_{\|U\|_{1\times 1} \leqslant 1} \|XU\|_{2\times\infty}.$$

The maximization over the unit ball $\|U\|_{1\times 1} \leqslant 1$ can be replaced with that over its extremal points, which are the matrices U that have zeroes in all positions except for one in which there is 1. Let (i,j) be this position, then for every such U we have:

$$\|XU\|_{2\times\infty} = \sqrt{\sum_{l=1}^n \sup_j |X(l,:)U(:,j)|^2} = \sqrt{\sum_{l=1}^n |X(l,i)|^2} = \sqrt{\sum_{l=1}^n |X^\top(i,l)|^2}.$$

As a result,

$$\sup_{\|U\|_{1\times 1} \leqslant 1} \|XU\|_{2\times\infty} = \sup_{1 \leq i \leq k} \sqrt{\sum_{l=1}^n |X^\top(i,l)|^2} = \left\|X^\top\right\|_{\infty\times 2}. \quad \square$$

D.2 Proof of Proposition 3.1

We prove an extended result that holds when $\|\cdot\|_{\mathscr{U}}$ and $\|\cdot\|_{\mathscr{V}}$ are more general mixed $(\ell_p \times \ell_q)$-norms, cf. (14).

Proposition D.1. *Let $\|\cdot\|_{\mathscr{U}} = \|\cdot\|_{p_U^1 \times p_U^2}$ on $\mathbb{R}^{2d\times k}$, and $\|\cdot\|_{\mathscr{V}} = \|\cdot\|_{p_V^1 \times p_V^2}$ on $\mathbb{R}^{n\times k}$. Then, optimal solutions $p^* = p^*(\widehat{X}, U)$ and $q^* = q^*(\widehat{X}, V, Y)$ to (32) are given by*

$$p_i^* = \frac{\|\widehat{X}(:,i)\|_{q_V^1} \cdot \|U(i,:)\|_{q_V^2}}{\sum_{\imath=1}^{2d} \|\widehat{X}(:,\imath)\|_{q_V^1} \cdot \|U(\imath,:)\|_{q_V^2}}, \quad q_j^* = \frac{\|\widehat{X}(j,:)\|_{q_U^1} \cdot \|V(j,:) - Y(j,:)\|_{q_U^2}}{\sum_{\jmath=1}^{n} \|\widehat{X}(\jmath,:)\|_{q_U^1} \cdot \|V(\jmath,:) - Y(\jmath,:)\|_{q_U^2}}.$$

Moreover, we can bound their respective variance proxies (cf. (31)): introducing

$$\mathcal{L}^2_{\mathscr{U},\mathscr{V}} = \frac{1}{n^2} \sup_{\|U\|_{\mathscr{U}} \leq 1} \|\widehat{X}U\|^2_{\mathscr{V}^*},$$

we have, as long as $p_U^2 \leqslant q_V^2$,

$$\sigma_{\mathcal{U}}^2(p^*) \leqslant 2R_*^2\mathcal{L}_{\mathscr{U},\mathscr{V}}^2 + \frac{2}{n^2}R_*^2\|\widehat{X}^\top\|_{q_U^1\times q_V^1}^2,$$

and, as long as $p_V^1 \geqslant 2$,

$$\sigma_{\mathcal{V}}^2(q^*) \leqslant 8n\mathcal{L}_{\mathscr{U},\mathscr{V}}^2 + \frac{8}{n^2}\|\widehat{X}\|_{1\times q_U^1}^2.$$

Proof. Note that the dual norms to $\|\cdot\|_{p_U^1\times p_U^2}$ and $\|\cdot\|_{p_V^1\times p_V^2}$ are given by $\|\cdot\|_{q_U^1\times q_U^2}$ and $\|\cdot\|_{q_V^1\times q_V^2}$ correspondingly, see, e.g., (Sra, 2012).

1^o. For $\mathbb{E}\left[\|\xi_U(p)\|_{\mathscr{V}^*}^2\right]$ we have:

$$\mathbb{E}\left[\|\xi_U(p)\|_{\mathscr{V}^*}^2\right] = \sum_{i=1}^{2d} p_i \left\|\widehat{X}\frac{e_i e_i^\top}{p_i}U\right\|_{q_V^1\times q_V^2}^2 = \sum_{i=1}^{2d}\frac{1}{p_i}\left\|\widehat{X}(:,i)\cdot U(i,:)\right\|_{q_V^1\times q_V^2}^2$$
$$= \sum_{i=1}^{2d}\frac{1}{p_i}\|\widehat{X}(:,i)\|_{q_V^1}^2\cdot\|U(i,:)\|_{q_V^2}^2,$$

where the last transition can be verified directly. The right-hand side can be easily minimized on Δ_{2d} explicitly, which results in

$$p_i^* = \frac{\|\widehat{X}(:,i)\|_{q_V^1}\cdot\|U(i,:)\|_{q_V^2}}{\sum_{\imath=1}^{2d}\|\widehat{X}(:,\imath)\|_{q_V^1}\cdot\|U(\imath,:)\|_{q_V^2}}$$

and

$$\mathbb{E}\left[\|\xi_U(p^*)\|_{\mathscr{V}^*}^2\right] = \left[\sum_{i=1}^{2d}\|\widehat{X}(:,i)\|_{q_V^1}\cdot\|U(i,:)\|_{q_V^2}\right]^2.$$

Now we can bound $\sigma_{\mathcal{U}}^2(p^*)$ via the triangle inequality:

$$\sigma_{\mathcal{U}}^2(p^*) \leqslant \frac{2}{n^2}\sup_{U\in\mathcal{U}}\|\widehat{X}U\|_{\mathscr{V}^*}^2 + \frac{2}{n^2}\sup_{U\in\mathcal{U}}\mathbb{E}\left[\|\xi_U(p^*)\|_{\mathscr{V}^*}^2\right]$$
$$= 2R_*^2\mathcal{L}_{\mathscr{U},\mathscr{V}}^2 + \frac{2}{n^2}\sup_{U\in\mathcal{U}}\left[\sum_{i=1}^{2d}\|\widehat{X}(:,i)\|_{q_V^1}\cdot\|U(i,:)\|_{q_V^2}\right]^2$$
$$= 2R_*^2\mathcal{L}_{\mathscr{U},\mathscr{V}}^2 + \frac{2}{n^2}\sup_{\|U\|_{\mathscr{U}}\leq R_*}\left[\sum_{i=1}^{2d}\|\widehat{X}(:,i)\|_{q_V^1}\cdot\|U(i,:)\|_{q_V^2}\right]^2$$
$$= 2R_*^2\mathcal{L}_{\mathscr{U},\mathscr{V}}^2 + \frac{2}{n^2}R_*^2\|\widehat{X}^\top\|_{q_U^1\times q_V^1}^2,$$

where we used Lemma C.2 (see Appendix C) in the last transition.

2^o. We now deal with $\mathbb{E}\left[\|\eta_{V,Y}(q)\|_{\mathscr{U}^*}^2\right]$. As previously, we can explicitly compute

$$q_j^* = \frac{\|\widehat{X}(j,:)\|_{q_U^1}\cdot\|V(j,:)-Y(j,:)\|_{q_U^2}}{\sum_{\jmath=1}^{n}\|\widehat{X}(\jmath,:)\|_{q_U^1}\cdot\|V(\jmath,:)-Y(\jmath,:)\|_{q_U^2}}$$

and

$$\mathbb{E}\left[\|\eta_{V,Y}(q^*)\|_{\mathscr{U}^*}^2\right] = \left[\sum_{j=1}^{n}\|\widehat{X}(j,:)\|_{q_U^1}\cdot\|V(j,:)-Y(j,:)\|_{q_U^2}\right]^2.$$

Thus, by the triangle inequality,

$$\begin{aligned}
\sigma^2_{\mathcal{V}}(q_*) &\leqslant \frac{2}{n^2}\sup_{(V,Y)\in\mathcal{V}\times\mathcal{V}}\|\widehat{X}^\top(V-Y)\|^2_{\mathscr{U}^*} + \frac{2}{n^2}\sup_{(V,Y)\in\mathcal{V}\times\mathcal{V}}\mathbb{E}\left[\|\eta_{V,Y}(q^*)\|^2_{\mathscr{U}^*}\right] \\
&\leqslant \frac{2}{n^2}\sup_{\|V\|_{\infty\times 1}\leqslant 2}\|\widehat{X}^\top V\|^2_{\mathscr{U}^*} + \frac{2}{n^2}\sup_{\|V\|_{\infty\times 1}\leqslant 2}\left[\sum_{j=1}^n \|\widehat{X}(j,:)\|_{q_U^1}\cdot\|V(j,:)\|_{q_U^2}\right]^2 \\
&= \frac{8}{n}\sup_{\|V\|_{2\times 1}\leqslant 1}\|\widehat{X}^\top V\|^2_{\mathscr{U}^*} + \frac{8}{n^2}\|\widehat{X}\|^2_{1\times q_U^1} \\
&\leqslant \frac{8}{n}\sup_{\|V\|_{p_V^1\times p_V^2}\leqslant 1}\|\widehat{X}^\top V\|^2_{\mathscr{U}^*} + \frac{8}{n^2}\|\widehat{X}\|^2_{1\times q_U^1} \\
&= 8n\mathcal{L}^2_{\mathscr{U},\mathscr{V}} + \frac{8}{n^2}\|\widehat{X}\|^2_{1\times q_U^1}.
\end{aligned}$$

Here in the second line we used that the Minkowski sum $\Delta_k + (-\Delta_k)$ belongs to the ℓ_1-ball with radius 2 (whence $\mathcal{V}+(-\mathcal{V})$ belongs to the $(\ell_\infty\times\ell_1)$-ball with radius 2); in the third line we used Lemma C.3 (see Appendix C) and the relation on $\mathbb{R}^{n\times k}$:

$$\|\cdot\|_{2\times 1}\leqslant\sqrt{n}\|\cdot\|_{\infty\times 1};$$

lastly, we used that $p_V^1\geq 2$ and that $\|\cdot\|_{p_V^1\times p_V^2}$ is non-increasing in $p_V^1, p_V^2\geq 1$. □

Proof of Proposition 3.1. We instantiate Proposition D.1 with $\|\cdot\|_{\mathscr{U}} = \|\cdot\|_{1\times 1}$ and $\|\cdot\|_{\mathscr{V}} = \|\cdot\|_{2\times 1}$, and observe, using Proposition 2.1, that for $\widehat{X} = [X, -X]\in\mathbb{R}^{n\times 2d}$ it holds

$$\mathcal{L}_{\mathscr{U},\mathscr{V}} = \frac{1}{n}\|\widehat{X}^\top\|_{\infty\times 2} = \frac{1}{n}\|X^\top\|_{\infty\times 2}$$

and

$$\|\widehat{X}\|_{1\times\infty} = \|X\|_{1\times\infty}.$$

□

D.3 Proof of Proposition 3.2

We have

$$\begin{aligned}
\min_{\substack{p\in\Delta_{2d},\\ P\in(\Delta_k^\top)^{\otimes 2d}}}\mathbb{E}\|\xi_U(p,P)\|^2_{2\times\infty} &= \min_{\substack{p\in\Delta_{2d},\\ P\in(\Delta_k^\top)^{\otimes 2d}}}\sum_{i=1}^{2d}\frac{1}{p_i}\|\widehat{X}(:,i)\|_2^2\cdot\left[\sum_{l=1}^k\frac{1}{P_{il}}\cdot|U(i,l)|^2\right] \\
&= \min_{p\in\Delta_{2d}}\sum_{i=1}^{2d}\frac{1}{p_i}\|\widehat{X}(:,i)\|_2^2\cdot\|U(i,:)\|_1^2,
\end{aligned}$$

where we carried out the internal minimization explicitly, obtaining

$$P^*_{il} = \frac{|U_{il}|}{\|U(i,:)\|_1}.$$

Optimization in p gives:

$$p_i^* = \frac{\|\widehat{X}(:,i)\|_2\cdot\|U(i,:)\|_1}{\sum_{i=1}^{2d}\|\widehat{X}(:,i)\|_2\cdot\|U(i,:)\|_1}, \quad \mathbb{E}\|\xi_U(p^*,P^*)\|^2_{2\times\infty} = \sum_{i=1}^{2d}\|\widehat{X}(:,i)\|_2\cdot\|U(i,:)\|_1.$$

Defining

$$\mathcal{L}_{\mathscr{U},\mathscr{V}} = \frac{1}{n} \sup_{\|U\|_{\mathscr{U}} \leqslant 1} \|XU\|_{\mathscr{V}^*}$$

and proceeding as in the proof of Proposition D.1, we get

$$\begin{aligned}
\sigma^2_{\mathcal{U}}(p^*, P^*) &\leqslant \frac{2}{n^2} \sup_{U \in \mathcal{U}} \|\widehat{X}U\|^2_{2\times\infty} + \frac{2}{n^2} \sup_{U \in \mathcal{U}} \mathbb{E}\left[\|\xi_U(p^*, P^*)\|^2_{2\times\infty}\right] \\
&= 2R_*^2 \mathcal{L}^2_{\mathscr{U},\mathscr{V}} + \frac{2}{n^2} \sup_{U \in \mathcal{U}} \left[\sum_{i=1}^{2d} \|\widehat{X}(:,i)\|_2 \cdot \|U(i,:)\|_1\right]^2 \\
&= 2R_*^2 \mathcal{L}^2_{\mathscr{U},\mathscr{V}} + \frac{2R_*^2}{n^2} \sup_{\|U\|_{1\times 1} \leqslant 1} \left[\sum_{i=1}^{2d} \|\widehat{X}(:,i)\|_2 \cdot \|U(i,:)\|_1\right]^2 \\
&= 2R_*^2 \mathcal{L}^2_{\mathscr{U},\mathscr{V}} + \frac{2}{n^2} R_*^2 \|\widehat{X}^\top\|^2_{2\times 1} \\
&= \frac{4}{n^2} R_*^2 \|X^\top\|^2_{2\times 1},
\end{aligned}$$

where in the last two transitions we used Lemma C.2 and Proposition 2.1 (note that $\|\widehat{X}^\top\|_{2\times 1} = \|X^\top\|_{2\times 1}$). Note that the last transition requires that $\|\cdot\|_{\mathscr{U}}$ has ℓ_1-geometry in the classes – otherwise, Lemma C.2 cannot be applied.

To obtain (q^*, Q^*) we proceed in a similar way:

$$\begin{aligned}
\min_{\substack{q\in\Delta_n,\\ Q\in(\Delta_k^\top)^{\otimes n}}} \mathbb{E}\|\eta_{V,Y}(q,Q)\|^2_{\infty\times\infty} &= \min_{\substack{q\in\Delta_n,\\ Q\in(\Delta_k^\top)^{\otimes n}}} \sum_{j=1}^n \frac{1}{q_j} \|\widehat{X}(j,:)\|^2_\infty \cdot \left[\sum_{l=1}^k \frac{1}{Q_{jl}} \cdot |V(j,l) - Y(j,l)|^2\right] \\
&= \min_{q\in\Delta_n} \sum_{j=1}^n \frac{1}{q_j} \|\widehat{X}(j,:)\|^2_\infty \cdot \|V(j,:) - Y(j,:)\|^2_1,
\end{aligned}$$

which results in

$$q_j^* = \frac{\|\widehat{X}(j,:)\|_\infty \cdot \|V(j,:) - Y(j,:)\|_1}{\sum_{j=1}^n \|\widehat{X}(j,:)\|_\infty \cdot \|V(j,:) - Y(j,:)\|_1} \qquad Q^*_{jl} = \frac{|V_{jl} - Y_{jl}|}{\|V(j,:) - Y(j,:)\|_1}.$$

The corresponding variance proxy can then be bounded in the same way as in the proof of Proposition D.1. □

E Discussion of Alternative Geometries

Here we consider alternative choices of the proximal geometry in mirror descent applied to the saddle-point formulation of the CCSPP (1), possibly with other choices of regularization than the entrywise ℓ_1-norm. The goal is to show that the geometry chosen in Sec. 2 is the only one for which we can obtain favorable accuracy guarantees for stochastic mirror descent (**SMD**).

Given the structure of the primal and dual feasible sets, it is reasonable to consider general mixed norms of the type (14):

$$\|\cdot\|_{\mathscr{U}} = \|\cdot\|_{p_U^1 \times p_U^2}, \quad \|\cdot\|_{\mathscr{V}} = \|\cdot\|_{p_V^1 \times p_V^2},$$

where $p_U^{1,2}, p_V^{1,2} \geq 1$ (in the case of $\|\cdot\|_{\mathscr{U}}$, we also assume the same norm for regularization). Note that their dual norms can be easily computed: the dual norm of $\|\cdot\|_{p^1\times p^2}$ is $\|\cdot\|_{q^1\times q^2}$,

where $q^{1,2}$ are the corresponding conjugates to $p^{1,2}$, i.e., $1/p^i+1/q^i=1$ (see, e.g., Lemma 3 in Sra (2012)). Moreover, it makes sense to fix $p_V^1=2$ for the reasons discussed in Section 2.1. This leaves us with the obvious choices $p_V^2\in\{1,2\}$, $p_U^2\in\{1,2\}$ which corresponds to the sparsity-inducing or the standard Euclidean geometry of the *classes* in the dual/primal; $p_U^1\in\{1,2\}$ which corresponds to the sparsity-inducing or Euclidean geometry of the *features*. Finally, the choice $p_U^1=2$ (i.e., the Euclidean geometry in the features) can also be excluded: its combination with $p_V^1=2$ is known to lead to the large variance term in the *biclass* case.[6] This leaves us with the possibilities

$$p_U^2, p_V^2 \in \{1,2\}\times\{1,2\}. \tag{54}$$

In all these cases, the quantity $\mathcal{L}_{\mathcal{U},\mathcal{V}}$ defined in (26) can be controlled by extending Proposition 2.1:

Proposition E.1. *For any $\alpha\geqslant 1$ and $\beta\geqslant 1$ such that $\beta\geqslant\alpha$ it holds:*

$$\mathcal{L}_{\mathcal{U},\mathcal{V}} := \frac{\|X\|_{1\times\alpha,\,2\times\beta}}{n} = \frac{\|X^\top\|_{\infty\times 2}}{n}.$$

The proof of this proposition follows the steps in the proof of Proposition 2.1, and is omitted.

Finally, the corresponding partial potentials could be constructed by combining the Euclidean and an entropy-type potential in a way similar to the one described in Sec. 2.1 for the dual variable; alternatively, one could use the power potential of Nesterov and Nemirovski (2013) that results in the same rates up to a constant factor.

Using Proposition E.1, we can also compute the potential differences for the four remaining setups (54). The results are shown in Table 2. Up to logarithmic factors, we have equivalent results for all four geometries, with the radius R_* evaluated in the corresponding norm $\|\cdot\|_{1\times 2}$ or $\|\cdot\|_{1\times 1}=\|\cdot\|_1$.

		Norm for $V\in\mathbb{R}^{n\times k}$	
		2×1	2×2
Norm for $U\in\mathbb{R}^{d\times k}$	1×2	$\Omega_{\mathcal{U}}=\|U^*\|_{1\times 2}^2\log d$ $\Omega_{\mathcal{V}}=n\log k$	$\Omega_{\mathcal{U}}=\|U^*\|_{1\times 2}^2\log d$ $\Omega_{\mathcal{V}}=n$
	1×1	$\Omega_{\mathcal{U}}=\|U^*\|_1^2\log(dk)$ $\Omega_{\mathcal{V}}=n\log k$	$\Omega_{\mathcal{U}}=\|U^*\|_1^2\log(dk)$ $\Omega_{\mathcal{V}}=n$

Table 2: Comparison of the potential differences for the norms corresponding to (54).

As a result, for the deterministic Mirror Descent (with balanced potentials) we obtain the accuracy bound (cf. (28)):

$$\texttt{Gap}(\bar U^T,\bar V^T)\leqslant\frac{O(1)\mathcal{L}_{\mathcal{U},\mathcal{V}}\sqrt{\Omega_{\mathcal{U}}\Omega_{\mathcal{V}}}}{\sqrt{T}}\leqslant\frac{\widetilde{O}_{d,k}(1)R_*}{\sqrt{T}}\frac{\|X^\top\|_{\infty\times 2}}{\sqrt{n}}+\frac{\mathtt{r}}{T}.$$

in all four cases, where $\widetilde{O}_{d,k}(1)$ is a logarithmic factor in d and k, and $R_*=\|U^*\|_{1\times 2}$ or $R_*=\|U^*\|_1$ depending on $p_U^2\in\{1,2\}$. In other words, the deterministic accuracy bound of Theorem 2.1 is essentially preserved for all four geometries in (54). On the other hand, using

[6] Note that in the biclass case, our variance estimate for the partial sampling scheme (cf. Theorem 3.1) reduces to those in (Juditsky and Nemirovski, 2011b, Section 2.5.2.3). They consider the cases of ℓ_1/ℓ_1 and ℓ_1/ℓ_2 geometries for the primal/dual, and omit the case of ℓ_2/ℓ_2-geometry, in which the sampling variance "explodes".

Proposition D.1, we obtain that in the case of (**Part-SS**), the extra part of the accuracy bound due to sampling (cf. (36)) is also essentially preserved:

$$\mathbb{E}[\mathtt{Gap}(\bar{U}^T, \bar{V}^T)] \leqslant \frac{\widetilde{O}_{d,k}(1)R_*}{\sqrt{T}}\left(\frac{\|X^\top\|_{\infty\times 2}}{\sqrt{n}} + \frac{\|X\|_{1\times\infty}}{n}\right) + \frac{\mathtt{r}}{T}.$$

However, if we consider *full sampling*, the situation changes: in the case $p_U^2 = 2$ the variance bound that holds for (**Part-SS**) is not preserved for (**Full-SS**). This is because our argument to control the variance of the full sampling scheme always requires that $p_U^2 \leqslant 1$ (see the proof of Proposition 3.2 in Appendix D.3 for details; note that for q_V^2 we do not have such a restriction since the variance proxy $\sigma_{\mathcal{V}}^2$ is controlled on the *set* $\mathcal{V}$ given by (7) that has $\ell_\infty \times \ell_1$-type geometry regardless of the norm $\|\cdot\|_{\mathscr{V}}$. This leaves us with the final choice between the $\|\cdot\|_{2\times 1}$ and $\|\cdot\|_{2\times 2}$ norm in the dual, as we have to use the elementwise $\|\cdot\|_1$-norm in the primal. Both choices result in essentially the same accuracy bound (note that this choice only influences the algorithm but not the saddle-point problem itself). We have focused on the $\|\cdot\|_{2\times 1}$ norm because of the algorithmic considerations: with this norm, we have multiplicative updates in the case of the multiclass hinge loss, which allows for a sublinear algorithm presented in Section 4.

F Correctness of Subroutines in Algorithm 1

In this section, we recall the subroutines used in Algorithm 1 – those for performing the lazy updates and tracking the running averages – and demonstrate their correctness.

Procedure 1 UpdatePrimal

Input: $\widetilde{U} \in \mathbb{R}^{2d\times k}$, $\alpha, \pi, \eta \in \mathbb{R}^{2d}$, $l \in [k]$, γ, λ, R_*

$L \equiv \log(2dk)$
for $i = 1$ **to** $2d$ **do**
  $\mu_i = \pi_i - \alpha_i \cdot \widetilde{U}(i,l) \cdot (1 - e^{-2\gamma L R_* \eta_i/n})$
end for
$M = \sum_{i=1}^{2d} \mu_i$
$\nu = \min\{e^{-2\gamma L R_* \lambda}, R_*/M\}$
for $i = 1$ **to** $2d$ **do**
  $\widetilde{U}(i,l) \leftarrow \widetilde{U}(i,l) \cdot e^{-2\gamma L R_* \eta_i/n}$
  $\alpha_i^+ = \nu \cdot \alpha_i$
  $\pi_i^+ = \nu \cdot \mu_i$
end for

Output: $\widetilde{U}, \alpha^+, \pi^+$

Primal Updates (Procedure 1). To demonstrate the correctness of Procedure 1, we prove the following result:

Lemma F.1. *Suppose that at t-iteration of Algorithm 1, Procedure 1 was fed with $\widetilde{U} = \widetilde{U}^t, \alpha = \alpha^t, \pi = \pi^t, \eta = \eta^t, l = l^t$ for which one had*

$$\widetilde{U}^t(:,l) \circ \alpha^t = U^t(:,l), \quad \forall l \in [k], \tag{55}$$

*where U^t is the t-th primal iterate of (**SMD**) equipped with (**Full-SS**) with the optimal sampling distributions (37), and η^t was the only non-zero column $\eta_{V^t,Y}(:,l^t)$ of $\eta_{V^t,Y}$. Moreover, suppose also that*

$$\pi^t(\imath) = \|U^t(\imath,:)\|_1 = \sum_{l\in[k]} U^t(\imath,l), \quad \imath \in [2d], \tag{56}$$

were the correct norms at the t-th step. Then Procedure 1 will output $\widetilde{U}^{t+t}, \alpha^{t+1}, \pi^{t+1}$ *such that*

$$\widetilde{U}^{t+1}(:,l) \circ \alpha^{t+1} = U^{t+1}(:,l), \ \forall l \in [k]$$

and

$$\pi^{t+1}(\imath) = \sum_{l \in [k]} U^{t+1}(\imath, l), \ \imath \in [2d].$$

Proof. Recall that the matrix $\eta^t = \eta_{V^t,Y}$ produced in (**Full-SS**) has a single non-zero column $\eta^t = \eta^t(:, l^t)$, and according to (**SMD**), the primal update $U^t \to U^{t+1}$ writes as (cf. (24)):

$$U_{il}^{t+1} = U_{il}^t \cdot e^{-2\gamma_t R_* L \eta_{V^t,Y}(i,l)/n} \cdot \min\left\{e^{-2\gamma_t R_* L\lambda}, R_*/M\right\},$$

where

$$L := \log(2dk), \quad M_t := \sum_{i=1}^{2d}\sum_{l=1}^{k} U_{il}^t \cdot e^{-2\gamma_t R_* L \eta_{V^t,Y}(i,l)/n},$$

and $\eta_{V^t,Y}$ has a single non-zero column $\eta^t = \eta^t_{V^t,Y}(:, l^t)$. This can be rewritten as

$$U_{il}^{t+1} = \begin{cases} U_{il}^t \cdot \nu \cdot q_i^t, & l = l^t, \\ U_{il}^t \cdot \nu, & l \neq l^t, \end{cases} \tag{57}$$

where

$$\begin{aligned} q_i^t &= e^{-2\gamma_t L R_* \eta_i^t/n}, \\ \nu &= \min\{e^{-2\gamma_t L R_* \lambda}, R_*/M\}, \\ M &= \sum_{i \in [2d]} U_{il^t}^t \cdot q_i^t + \sum_{i \in [2d]} \sum_{l \in [k] \setminus \{l^t\}} U_{il}^t. \end{aligned}$$

Thus, M and ν can be expressed via $\pi^t(i) = \sum_{l \in [k]} U^t(i,l)$, cf. (56):

$$M = \sum_{i \in [2d]} \pi_i^t - \underbrace{\alpha_i^t \widetilde{U}_{i,l^t}^t}_{U_{i,l^t}^t}(1 - q_i^t), \tag{58}$$

where we used the premise (55). Now we can see that lazy updates of $\widetilde{U}$ can be expressed as

$$\begin{aligned} \alpha_i^{t+1} &= \nu \cdot \alpha_i^t, \\ \widetilde{U}_{i,l^t}^{t+1} &= \widetilde{U}_{i,l^t}^t \cdot q_i^t, \end{aligned} \tag{59}$$

and the updates for the norms π^{t+1} as

$$\pi_i^{t+1} = \nu^t \left[\pi_i^t + \alpha_i^t \widetilde{U}_{i,l^t}^t (q_i^t - 1)\right] \tag{60}$$

One can immediately verify that this is exactly the update produced in the call of Procedure 1 in line 15 of Algorithm 1. ∎

Procedure 2 UpdateDual

Input: $\widetilde{V}, Y \in \mathbb{R}^{n\times k}$, $\beta, \rho, \xi \in \mathbb{R}^n$, $\ell \in [k]$, $y \in [k]^{\otimes n}$, γ

$\theta = e^{-2\gamma \log(k)}$
for $j = 1$ **to** n **do**
  $\omega_j = e^{2\gamma \log(k)\xi_j}$
  $\varepsilon_j = e^{-2\gamma \log(k) Y(j,\ell)}$
  $\chi_j = 1 - \beta_j \cdot \widetilde{V}(j,\ell) \cdot (1 - \omega_j \cdot \varepsilon_j)$
  if $\ell \neq y_j$ **then** # *not the actual class of j drawn*
    $\chi_j \leftarrow \chi_j - \beta_j \cdot \widetilde{V}(j, y_j) \cdot (1 - \theta)$
  end if
  $\beta_j^+ = \beta_j / \chi_j$
  $\widetilde{V}(j,\ell) \leftarrow \widetilde{V}(j,\ell) \cdot \omega_j \cdot \varepsilon_j$
  $\widetilde{V}(j, y_j) \leftarrow \widetilde{V}(j, y_j) \cdot \omega_j \cdot \theta$
  $\rho_j^+ = 2 - 2\beta_j^+ \cdot \widetilde{V}(j, y_j)$
end for

Output: $\widetilde{V}, \beta^+, \rho^+$

Dual Updates (Procedure 2). To demonstrate the correctness of Procedure 2, we prove the following lemma.

Lemma F.2. *Suppose that at t-iteration of Algorithm 1, Procedure 2 was fed with $\widetilde{V} = \widetilde{V}^t, \beta = \beta^t, \rho = \rho^t, \xi = \xi^t, \ell = \ell^t$, for which one had*

$$\widetilde{V}^t(:,l) \circ \beta^t = V^t(:,l), \quad \forall l \in [k], \tag{61}$$

*where V^t is the t-th dual iterate of (**SMD**) equipped with (**Full-SS**) with the optimal sampling distributions (37), and ξ^t was the only non-zero column $\xi_{U^t}(:,\ell^t)$ of ξ_{U^t}. Moreover, suppose also that*

$$\rho^t(\jmath) = \|V^t(\jmath,:) - Y(\jmath,:)\|_1, \quad \jmath \in [n] \tag{62}$$

were the correct norms at the t-th step. Then Procedure 2 will output $\widetilde{V}^{t+t}, \beta^{t+1}, \rho^{t+1}$ such that

$$\widetilde{V}^{t+1}(:,l) \circ \beta^{t+1} = V^{t+1}(:,l), \ \forall l \in [k]$$

and

$$\rho^{t+1}(\jmath) = \|V^{t+1}(\jmath,:) - Y(\jmath,:)\|_1, \quad \jmath \in [n].$$

Proof. Recall that the random matrix ξ_{U^t} has a single non-zero column $\xi^t := \xi_{U^t}(:,\ell^t)$, and according to (**SMD**), the update $V^t \to V^{t+1}$ writes as (cf. (25)):

$$V_{jl}^{t+1} = V_{jl}^t \cdot \frac{\exp[2\gamma_t \log(k) \cdot (\xi_{U^t}(j,l) - Y(j,l))]}{\sum_{\ell=1}^k V_{j\ell}^t \cdot \exp[2\gamma_t \log(k) \cdot (\xi_{U^t}(j,\ell) - Y(j,\ell))]}. \tag{63}$$

Note that all elements of the matrix $\xi_{U^t} - Y$ in each row j have value 1, except for at most two elements in the columns ℓ^t and y_j, where y_j is the actual label of the j-th training example, that is, the only $l \in [k]$ for which $Y(j,l) = 1$. Recall also that $\sum_{\ell\in[k]} V^t(j,\ell) = 1$ for any $j \in [n]$. Thus, introducing

$$\omega_j = e^{2\gamma_t \log(k)\xi_j^t}, \quad \varepsilon_j = e^{-2\gamma_t \log(k) Y(j,l)}$$

as defined in Procedure 2, we can express the denominator in (63) as

$$\chi_j = \begin{cases} 1 - \underbrace{\beta_j^t \cdot \widetilde{V}_{j,\ell^t}^t}_{V_{j,\ell^t}^t} \cdot (1 - \omega_j \cdot \varepsilon_j), & \text{if } \ell^t = y_j, \\ 1 - \underbrace{\beta_j^t \cdot \widetilde{V}_{j,\ell^t}^t}_{V_{j,\ell^t}^t} \cdot (1 - \omega_j \cdot \varepsilon_j) - \underbrace{\beta_j^t \cdot \widetilde{V}_{j,y_j}^t}_{V_{j,y_j}^t} \cdot (1 - e^{-2\gamma_t \log(k)}), & \text{if } \ell^t \neq y_j, \end{cases} \tag{64}$$

where we used the premise (61). One can verify that this corresponds to the value of χ_j produced by line 8 of Procedure 2. Then, examining the numerator in (63), we can verify that lines 9–11 guarantee that

$$\widetilde{V}_{j,l}^{t+1} \cdot \beta_j^{t+1} = V_{j,l}^{t+1}, \ \forall l \in [k]$$

holds for the updated values. To verify the second invariant, we combining this result with the premise (62). This gives

$$\rho_j^{t+1} = 2 - V_{j,y_j}^{t+1} = 2 - 2\beta_j^{t+1} \cdot \widetilde{V}_{j,y_j}^{t+1}, \tag{65}$$

which indeed corresponds to the update in line 12 of the procedure. □

Correctness of Tracking the Cumulative Sums. We only consider the primal variables (Procedure 3 and line 24 of Algorithm 1); the complimentary case can be treated analogously. Note that due to the previous two lemmas, at any iteration t of Algorithm 1 Procedure 3 is fed with $l = l^t$, $\widetilde{U} = \widetilde{U}^t$, $\alpha = \alpha^t$ for which it holds $\widetilde{U}^t \alpha^t = U^t$. Now, assume that all previous input values $A^\tau, \tau \leq t$, of variable A, and the current inputs $A_{\text{pr}}^t, U_\Sigma^t$ of variables A_{pr}, U_Σ, satisfy the following:

$$A^\tau = \sum_{s=0}^{\tau-1} \alpha^s, \quad \forall \tau \leq t, \tag{66}$$

$$A_{\text{pr}}^t(i,l) = A_i^{\tau^t(i,l)}, \tag{67}$$

$$U_\Sigma^t(i,l) = \sum_{s=0}^{\tau^t(i,l)} U^s(i,l), \tag{68}$$

where $0 \leqslant \tau^t(i,l) \leqslant t-1$ is the latest moment s, **strictly before** t, when the sampled $l^s \in [k]$ coincided with the given l:

$$\tau^t(i,l) = \operatorname*{argmax}_{s \leqslant t-1} \{s : l^s = l\}. \tag{69}$$

Let us show that this invariant will be preserved aftet the call of Procedure 3 – in other words, that (66)–(69) hold for $t+1$, i.e., for the ouput values $U_\Sigma^{t+1}, A_{\text{pr}}^{t+1}, A^{t+1}$ (note that the variables $U_\Sigma, A_{\text{pr}}, A^{t+1}$ only changed within Procedure 3, so their output values are also the input values at the next iteration).

Proof. Indeed, it is clear that (66) will be preserved (cf. line 4 of Procedure 3). To verify (67), note that $A_{\text{pr}}(i,l)$ only gets updated when $l = l^t$ (cf. line 3), and in this case we will have $\tau^{t+1}(i,l) = t$, and otherwise $\tau^{t+1}(i,l) = \tau^t(i,l)$, cf. (69).

Thus, it only remains to verify the validity of (68) after the update. To this end, note that by (69) we know that the value $\widetilde{U}^s(i,l)$ of the variable $\widetilde{U}(i,l)$ remained constant for $\tau^t(i,l) \leq s < t$, and it will not change after the call at t-th iteration unless $l^t = l$, that is, unless $\tau^{t+1}(i,l) = t$. This is exactly when line (2) is invoked, and it ensures (68) for $t+1$. □

Finally, invoking (66)–(69) at $t = T$, we see that line 24 results in the correct final value $\sum_{t=0}^T U^t$ of the cumulative sum U_Σ. Thus, the correctness of Algorithm 1 is verified.

G Additional Remarks on Algorithm 1

Removing the $O(dk)$ Complexity Term. In fact, the extra term $O(dk)$ in the runtime and memory complexities of Algorithm 1 can be easily avoided. To see this, recall that when solving the simplex-constrained CCSPP (11), we are foremost interested in solving the ℓ_1-constrained CCSPP (9), and an ε-accurate solution $U = [U_1; U_2] \in \mathbb{R}^{2d\times k}$ to (11) yields an ε-accurate solution $\widehat{U} = U_1 - U_2 \in \mathbb{R}^{d\times k}$ to (9). Recall that we initialize Algorithm 1 with $\widetilde{U}^0 = \mathbb{1}_{2d\times k}$ and $\alpha^0 = \mathbb{1}_{2d}$, which corresponds to $\widehat{U}^0 = 0_{2d\times k}$. Moreover, at any iteration we change a single entry of $\widetilde{U}^t$, and scale the whole scaling vector $\alpha^0 = \mathbb{1}_{2d}$ by a constant (in fact, all entries of α^t are always equal to each other; we omitted this fact in the main text to simplify the presentation, since the entries of β generally have different values). Hence, the final candidate solution $\widehat{U}^T = [\bar{U}_1^T - \bar{U}_2^T]$ to the ℓ_1-constrained problem will actually have at most $O(dT)$ non-zero entries that correspond to the entries of $\widetilde{U}$ changed in the course of the algorithm. To exploit this, we can modify Algorithm 1 as follows:

- Instead of explicitly initializing and storing the whole matrices $\widetilde{U}$, U_Σ, and A_{pr}, we can hard-code the "default" value $\widetilde{U}(i,l) = 1$ (cf. line 2 of Algorithm 1), and use a bit mask to flag the entries $\widetilde{U}(i,l)$ that have already been changed at least once. This mask can be stored as a list `Changed` of pairs (i,l), i.e. in a sparse form.
- When post-processing the cumulative sum U_Σ (see line (24) of Algorithm 1), instead of post-processing all entries of U_Σ, we can only process those in the list `Changed`, and ignore the remaining ones, since the corresponding to them entries in $\widehat{U}^T$ (a candidate solution to (9)) will have zero values. We can then directly output $\widehat{U}^T$ in a sparse form.

It is clear that such modification of Algorithm 1 results in the replacement of the $O(dk)$ term in runtime complexity with $O(dT)$ (which is always an improvement since $O(d)$ a.o.'s are done anyway in each iteration); moreover, the memory complexity changes from $O(dn + nk + dk)$ to $O(dn + nk + d\min(T,k))$.

Infeasibility of the Noisy Dual Iterates. Note that when we generate an estimate of the primal gradient $X^T(V^t - Y)$ according to (**Full-SS**) or (**Part-SS**), we also obtain an unbiased estimate of the dual iterate V^t, and vice versa. In the setup with vector variables, Juditsky and Nemirovski (2011b) propose to average such noisy iterates instead of the acutal iterates (U^t, V^t) as we do in (**SMD**). Averaging of the noisy iterates is easier to implement since they are sparse (one does not need to track the cumulative sums), and one could show similar guarantees for the primal accuracy of their running averages. However, in the case of the dual variable its noisy counterpart is infeasible (Juditsky and Nemirovski, 2011b, Sec. 2.5.1); as a result, one loses the guarantee for the duality gap. Hence, we prefer to track the averages of the actual iterates of (**SMD**) as we do in Algorithm 1.