On Stochastic Mirror Descent: Convergence Analysis and Adaptive Variants

Ryan D'Orazio [1 2] **Nicolas Loizou** [1 2] **Issam Laradji** [3 4] **Ioannis Mitliagkas** [1 2 5]

Abstract

We investigate the convergence of stochastic mirror descent in both relatively smooth and smooth convex optimization. In relatively smooth convex optimization we provide new convergence guarantees for stochastic mirror descent (SMD) with a constant stepsize. For smooth convex optimization we propose a new adaptive stepsize scheme – the mirror stochastic Polyak stepsize (mSPS). Notably, our convergence results in both settings do not make bounded gradient assumptions or bounded variance assumptions, and we show convergence to a neighborhood that vanishes under interpolation. We complement our results with experiments across various supervised learning tasks and different instances of SMD, demonstrating the effectiveness of mSPS. Code is available at `https://github.com/IssamLaradji/mirror-sps`.

1. Introduction

We address the problem of finite-sum optimization,

$$\min_{x\in\mathcal{X}} \left[f(x) = \frac{1}{n}\sum_{i=1}^{n} f_i(x) \right], \qquad (1)$$

where $\mathcal{X} \subseteq \mathbb{R}^d$ is a closed convex set. A common iterative approach to solve (1) when $\mathcal{X} = \mathbb{R}^d$ is stochastic gradient descent (SGD), iterates are updated in the negative direction of a gradient computed from a single $i \in \{1, \cdots, n\}$. When the problem is constrained, $\mathcal{X} \subset \mathbb{R}^d$, one may employ projected methods such as stochastic projected gradient descent (SPGD). However, the convergence guarantees of both SGD and SPGD depend on values measured by the Euclidean norm. If the Euclidean structure is not naturally suited to the problem, *e.g.* the function is smooth with respect to a norm that is not the Euclidean norm, then SGD and SPGD can suffer a worse dependence on the dimension d of the space Bubeck 2014. A powerful generalization of SGD and SPGD is stochastic mirror descent (SMD) (Nemirovsky and Yudin 1983; Beck and Teboulle 2003), permitting better convergence guarantees by matching the geometry of the problem.

A typical analysis of mirror descent—and first-order methods in general—usually involves the notion of smoothness with respect to some norm $||\cdot||$, which is often used in selecting the appropriate instance of mirror descent. However, a recent trend is to study non-euclidean methods like mirror descent with the more general assumption of relative smoothness (Birnbaum, Devanur, and Xiao 2011; Bauschke, Bolte, and Teboulle 2017; Lu, Freund, and Nesterov 2018). In contrast to deterministic methods, stochastic methods under relative smoothness have received less attention.

Our contributions are summarized as follows; **novel analysis of SMD**, our analysis includes new convergence guarantees in both the relatively smooth and smooth setting. In both cases our analysis does not make bounded gradient or bounded variance assumptions, instead we leverage the finite optimal objective difference (Loizou et al. 2021),

$$\sigma^2 := f(x_*) - \mathbb{E}\left[f_i^*\right] < \infty, \qquad (2)$$

where $x_* = \min_{x\in\mathcal{X}} f(x)$ and $f_i^* := \inf_{x\in\mathbb{R}^d} f_i(x)$. **Novel adaptive SMD** for smooth optimization, we propose a natural extension of the recent adaptive stochastic Polyak stepsize (SPS) to mirror descent and provide convergence guarantees. **Over-parametrized models and interpolation**, as a corollary of our theoretical results we obtain fast convergence of both constant and adaptive stepsize SMD under the interpolation setting.

2. Background

We denote vectors within the feasible set as $x \in \mathcal{X} \subseteq \mathbb{R}^d$, where $\mathbb{R}$ is the set of real numbers. We denote a minimum of (1) as $x_* \in \mathcal{X}$ and assume it exists. We also use the subscript to denote time, after t time steps we denote the average of the iterates as $\bar{x}_t = {}^1\!/\!_t \sum_{s=1}^{t} x_s$. With a slight abuse of notation we may also refer to the ith coordinate of x as x_i, $x = (x_1, \cdots, x_d)$. Whether the subscript refers to time or the coordinate is clear from context. We denote $||\cdot||_2$ as the Euclidean norm and $||\cdot||$ as any arbitrary norm with

[1] DIRO, Université de Montréal [2] MILA [3] Element AI [4] McGill University [5] Canada CIFAR AI Chair. Correspondence to: Ryan D'Orazio <ryan.dorazio@mila.quebec>.

Proceedings of the 38th International Conference on Machine Learning, PMLR 139, 2021. Copyright 2021 by the author(s).

corresponding dual norm $||x||_* = \sup_y\{\langle x, y\rangle : ||y|| \leq 1\}$.

For a differentiable function ψ, we define the difference between $\psi(x)$ and the first order approximation of it at y as the Bregman divergence $B_\psi(x; y)$, formally defined below.

Definition 1 (Bregman Divergence)**.** *Let $\psi : \mathcal{D} \to \mathbb{R}$ be a continuously differentiable on $\operatorname{int}\mathcal{D}$. Then the Bregman divergence with respect to ψ is $B_\psi : \mathcal{D} \times \operatorname{int}\mathcal{D} \to \mathbb{R}$, defined as $B_\psi(x; y) = \psi(x) - \psi(y) - \langle\nabla\psi(y), x - y\rangle$.*

A continously differentiable function f is convex on a set $\mathcal{X}$ if $B_f(x; y) \geq 0$ for any $x, y \in \mathcal{X}$. Similarly, a function f is L smooth with respect to a norm $||\cdot||$ if $B_f(x; y) \leq \frac{L}{2}||x - y||^2$, and is μ strongly convex if $\frac{\mu}{2}||x - y||^2 \leq B_f(x; y)$.

We will also refer to the generalization of smoothness and strong convexity – relative smoothness and relative strong convexity (Bauschke, Bolte, and Teboulle 2017; Lu, Freund, and Nesterov 2018). For any $x, y \in \mathcal{X}$, a differentiable function $f : \mathcal{X} \to \mathbb{R}$ is L smooth relative to ψ if $B_f(x; y) \leq LB_\psi(x; y)$, and is μ strongly convex relative to ψ if $\mu B_\psi(x; y) \leq B_f(x; y)$.

2.1. Mirror descent

To solve problem (1) we consider the general stochastic mirror descent update

$$x_{t+1} = \arg\min_{x\in\mathcal{X}} \langle\nabla f_i(x_t), x\rangle + \frac{1}{\eta_t}B_\psi(x; x_t). \quad (3)$$

In the non-smooth or deterministic setting $\nabla f_i(x_t)$ may be replaced by a subgradient or the full gradient respectively. To make the updates well defined, all we require is that $x_{t+1} \in \operatorname{int}\mathcal{D}$ in update (3).

Assumption 1. *Let $\mathcal{X} \subseteq \mathcal{D}$, then for any g, $x_{t+1} = \arg\min_{x\in\mathcal{X}} \langle g, x\rangle + \frac{1}{\eta_t}B_\psi(x; x_t) \in \operatorname{int}\mathcal{D}$.*

For example the following assumption by Orabona (2019) would be sufficient to guarantee Assumption 1.

Assumption 2. *Let $\psi : \mathcal{D} \to \mathbb{R}$ be a strictly convex function such that $\mathcal{X} \subseteq \mathcal{D}$, we require either one of the following to hold: $\lim_{x\to\partial\mathcal{X}} ||\nabla\psi(x)||_2 = +\infty$ or $\mathcal{X} \subseteq \operatorname{int}\mathcal{D}$.*

We note that other assumptions can be made to guarantee $x_{t+1} \in \operatorname{int}\mathcal{D}$.

The following standard one step mirror descent lemma will be used often, and we include the proof in the appendix. All other omitted proofs are deferred to the appendix.

Lemma 1. *Let B_ψ be the Bregman divergence with respect to a convex function $\psi : \mathcal{D} \to \mathbb{R}$ and assume assumption 1 holds. Let $x_{t+1} = \arg\min_{x\in\mathcal{X}} \langle g_t, x\rangle + \frac{1}{\eta_t}B_\psi(x; x_t)$. Then for any $x_* \in \mathcal{X}$*
$B_\psi(x_*; x_{t+1}) \leq B_\psi(x_*; x_t) - \eta_t\langle g_t, x_t - x_*\rangle - B_\psi(x_{t+1}; x_t) + \eta_t\langle g_t, x_t - x_{t+1}\rangle$.
Furthermore if ψ is μ_ψ strongly convex over $\mathcal{X}$ then $B_\psi(x_*; x_{t+1}) \leq B_\psi(x_*; x_t) - \eta_t\langle g_t, x_t - x_*\rangle + \frac{\eta_t^2}{2\mu_\psi}||g_t||_*^2$.

3. Constant and Polyak stepsize for mirror descent

In this section we provide background on constant stepsize selection for mirror descent and introduce our extensions of the classic Polyak stepsize and the recent stochastic Polyak stepsize (SPS) to mirror descent.

When a function is L smooth with respect to the Euclidean norm, a commonly prescribed stepsize for gradient descent is $\eta = {}^1\!/_L$, allowing for convergence in many settings (Bubeck 2014). Similarly if a function is L relatively smooth with respect to a function ψ, then the prescribed stepsize for mirror descent using ψ is $\eta = {}^1\!/_L$ (Birnbaum, Devanur, and Xiao 2011; Lu, Freund, and Nesterov 2018).

An alternative method to selecting a stepsize, as suggested by Polyak (1987), is to take η_t by minizing an upper bound on $||x_{t+1} - x_*||_2^2$. From Lemma 1, if we take $\psi = \frac{1}{2}||\cdot||_2^2$ and assume $g_t \in \partial f(x_t)$ is a subgradient for some convex function then we recover a well known inequality for projected subgradient descent: $\frac{1}{2}||x_* - x_{t+1}||_2^2 \leq \frac{1}{2}||x_* - x_t||_2^2 - \eta_t(f(x_t) - f(x_*)) + \frac{\eta_t^2}{2}||g_t||_2^2$.

Selecting η_t that minimizes this bound yields Polyak's stepsize, $\eta_t = {}^{(f(x_t)-f(x_*))}\!/_{||g_t||_2^2}$ (Polyak 1987; Beck 2017). Following in a similar fashion, we propose a generalization of Polyak's stepsize for mirror descent. If ψ is μ_ψ strongly convex[1] with respect to the norm $||\cdot||$ then we can minimize the right hand side of equation (1) to arrive at the mirror Polyak stepsize $\eta_t = {}^{\mu_\psi(f(x_t)-f(x_*))}\!/_{||g_t||_*^2}$.

Despite the well-known connection between projected subgradient descent and mirror descent, this generalization of Polyak's stepsize is absent from the literature. For completeness, we include analysis of the non-smooth case in the appendix. As expected, mirror descent with the mirror Polyak stepsize maintains the benefits of mirror descent, however, it is impractical – knowledge of $f(x_*)$ and the use of an exact gradient or subgradient is required.

In the stochastic setting Loizou et al. (2021) propose the more practical stochastic polyak stepsize (SPS), $\eta_t = {}^{(f_i(x_t)-f_i^*)}\!/_{c||\nabla f_i(x_t)||_2^2}$, and the bounded variant SPS$_{\max}$, $\eta_t = \min\{{}^{(f_i(x_t)-f_i^*)}\!/_{c||\nabla f_i(x_t)||_2^2}, \eta_b\}$. Where f_i^* is commonly known in many machine learning applications, and c is a scaling parameter that depends on the class of functions being optimized. Similar to the non-smooth case we can propose a generalization of SPS and SPS$_{\max}$, the mirror

[1] Note that without loss of generality we could assume ψ to be 1-strongly convex and scale ψ by ${}^1\!/_{\mu_\psi}$, however, this would not change the stepsize. Any scaling of ψ inversely scales the stepsize.

stochastic Polyak stepsize (mSPS) and the bounded variant mSPS_{max},

$$\text{mSPS} : \eta_t = \frac{\mu_\psi(f_i(x_t) - f_i^*)}{c\,||\nabla f_i(x_t)||_*^2}, \tag{4}$$

$$\text{mSPS}_{\text{max}} : \eta_t = \min\left\{\frac{\mu_\psi(f_i(x_t) - f_i^*)}{c\,||\nabla f_i(x_t)||_*^2}, \eta_b\right\}. \tag{5}$$

3.1. Self-bounding property of mSPS

An important property of SPS and mSPS is its self-bounding property for when f_i is L_i smooth and μ_i strongly convex with respect to a norm $||\cdot||$,

$$\frac{\mu_\psi}{2cL_i} \leq \eta_t = \frac{\mu_\psi(f_i(x_t) - f_i^*)}{c\,||\nabla f_i(x_t)||_*^2} \leq \frac{\mu_\psi}{2c\mu_i}. \tag{6}$$

We will often make use of the lower bound, also known as the self-bounding property of smooth functions (Srebro, Sridharan, and Tewari 2010), and we provide a complete proof in the appendix. A proof of the upper bound can be found for example in Orabona (2019)[Corollary 7.6].

3.2. Related work on adaptive stepsizes

Adaptive stepsizes for first order methods have a long history. In the context of mirror descent the method of accumulating past gradients or subgradients to set a stepsize, $\eta_t \propto 1/\sqrt{\sum_{s=1}^t ||g_s||_*^2}$, can be traced back to online learning (Auer, Cesa-Bianchi, and Gentile 2002; Streeter and McMahan 2010). More recently, a similar strategy has been used to develop adaptive coordinate-wise stepsizes such as ADAGRAD (*i.e.*, variable metric methods) (McMahan and Streeter 2010; J. Duchi, Hazan, and Singer 2011). Unfortunately, all the existing mirror descent methods with the aforementioned stepsize scheme require a bounded constraint when guarantees are provided using regret bounds; when the problem is unconstrained Orabona and Pál (2018) prove a $\Omega(t)$ worst case lower bound for the regret. However, it may still be possible to show meaningful results without relying on regret bounds. For example, see Li and Orabona (2019) for convergence results in the case of unconstrained SGD. Furthermore, in the stochastic case bounded gradient assumptions are made. In contrast our methods employ a completely different stepsize selection strategy and we make a very weak assumption on the noise.

4. Constant stepsize in relatively smooth optimization

In this section we provide new convergence results for SMD with constant stepsize under relatively smooth optimization. For an appropriately selected stepsize, we have that SMD enjoys a linear rate of convergence to a neighborhood of the minimum x_*.

Theorem 1. *Assume ψ satisfies assumption 1 . Furthermore assume f to be μ strongly convex relative to ψ, and f_i to be L-smooth relative to ψ. Then SMD with stepsize $\eta \leq \frac{1}{L}$ guarantees $\mathbb{E}\left[B_\psi(x_*; x_{t+1})\right] \leq (1-\mu\eta)^t B_\psi(x_*; x_1) + \frac{\sigma^2}{\mu}$.*

Under interpolation we have $\sigma^2 = 0$, and SMD will converge to the true solution. If ψ is strongly convex then Theorem 1 provides a linear rate on the expected distance $||x_{t+1} - x_*||^2$ for some norm $||\cdot||$.

Similar to Theorem 1 we can show convergence of a quantity to a neighborhood when only assuming f_i to be L smooth relative to ψ, where f or f_i need not be convex.

Theorem 2. *Assume ψ satisfies assumption 1. Furthermore assume f_i to L-smooth relative to ψ. Then SMD with stepsize $\eta \leq \frac{1}{L}$ guarantees $\mathbb{E}\left[\frac{1}{t}\sum_{s=1}^t B_f(x_*; x_s)\right] \leq \frac{B_\psi(x_*;x_1)}{\eta t} + \sigma^2$.*

The above guarantee also shows a result for the "best" iterate, $\mathbb{E}\left[\min_{1\leq s\leq t} B_f(x_*; x_s)\right]$, to a neighborhood. If f is strictly convex then this implies at least one iterate x_s is converging to a neighborhood of x_* on expectation.

In the constant stepsize and relatively smooth regime, Hanzely and Richtárik (2018) and Dragomir, Even, and Hendrikx (2021) provide convergence guarantees for SMD under different assumptions and to different neighborhoods. Hanzely and Richtárik (2018) make an assumption akin to bounded variance. Dragomir, Even, and Hendrikx (2021) consider a more restrictive version of mirror descent with a smaller stepsize, and assume $\nabla f(x_*) = 0$, but show convergence to a smaller neighborhood.

5. Convergence of mirror SPS

In this section we present our convergence results for SMD with mSPS_{max} when f_i are L_i smooth with respect to some norm and with varying assumptions. First, we consider the case when f is strongly convex relative to ψ, a common assumption when analysing mirror descent under strong convexity (Hazan and Kale 2014). Then we present rates under convexity and smoothness but without relatively strong convexity.

5.1. Smooth and strong convexity

With strong convexity of ψ and f being relatively strongly convex with respect to ψ we can show a linear rate of convergence to a neighborhood.

Theorem 3. *Assume f_i are convex and L_i smooth with respect to the norm $||\cdot||$. Furthermore, assume that f is*

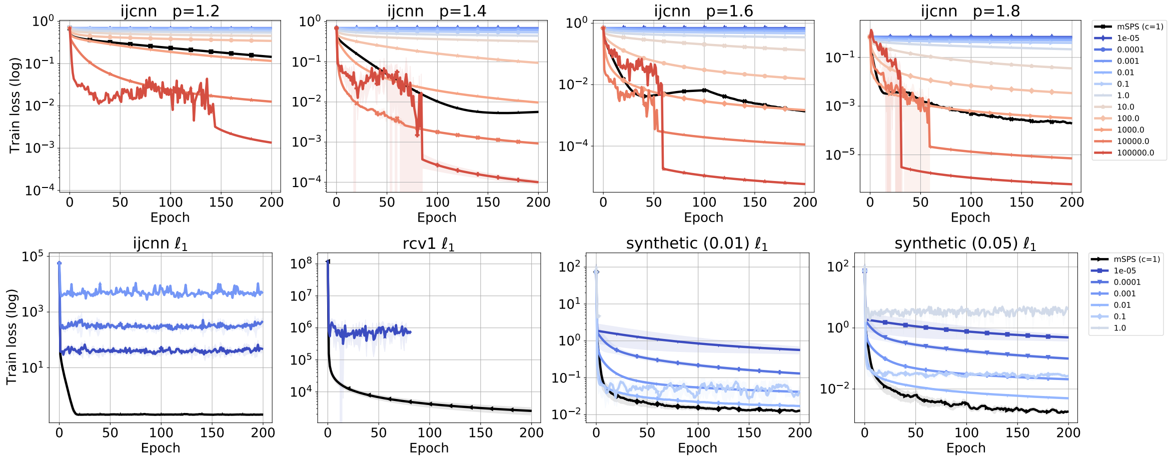

Figure 1. Comparison between mSPS with $c = 1$ and constant step-sizes on convex binary-classification problem with no constraints (row1), and with ℓ_1 constraints (row2).

μ strongly convex relative to ψ, where ψ is μ_ψ strongly convex with respect to the norm $||\cdot||$ and assumption 1 holds. Then SMD with mSPS$_{max}$ and $c \geq \frac{1}{2}$ guarantees $\mathbb{E}\left[B_\psi(x_*; x_{t+1})\right] \leq (1-\mu\alpha)^t B_\psi(x_*; x_1) + \frac{\eta_b \sigma^2}{\alpha\mu}$. Where $\alpha := \min\{^{\mu_\psi}/_{2cL_{\max}}, \eta_b\}$ and $L_{\max} = \max_i \{L_i\}_{i=1}^n$.

Theorem 3 generalizes the existing SPS$_{\max}$ results for SGD (Loizou et al. 2021)[Theorem 3.1] in fact we show that it also holds for SPGD.

5.2. Smooth and convex

Without f being relatively strongly convex we can attain convergence results on the average function value.

Theorem 4. *Assume f_i are convex and L_i smooth with respect to a norm $||\cdot||$, assumption 1 holds, and ψ is μ_ψ strongly convex with respect to the norm $||\cdot||$. Then mirror descent with mSPS$_{max}$ and $c \geq 1$ guarantees $\mathbb{E}\left[f(\bar{x}_t) - f(x_*)\right] \leq \frac{2B_\psi(x_*; x_1)}{\alpha t} + \frac{2\eta_b\sigma^2}{\alpha}$. Where $\alpha := \min\{^{\mu_\psi}/_{2cL_{\max}}, \eta_b\}$ and $L_{\max} = \max_i \{L_i\}_{i=1}^n$.*

6. Experiments

We test the performance of mSPS on different supervised learning domains with convex losses and with different instances of mirror descent and use mSPS with $c = 1$. Although in theory the bounded stepsize mSPS$_{\max}$ is required in absence of interpolation, in practice we observe mSPS converges.

We consider 2 series of experiments.[2] First, in Section 6.1 we consider unconstrained convex problems with mSPS and different p-norm algorithms, $\psi(x) = ||x||_p^2$. Second, in Section 6.2 we solve a convex problem with a ℓ_1 constraint using mSPS and the exponentiated gradient algorithm (EG).

6.1. Mirror descent across p-norms

We consider a convex binary-classification problems using radial basis function (RBF) kernels without regularization. We experiment on the ijcnn dataset obtained from LIBSVM (Chang and Lin 2011) which does not satisfy interpolation. However we show in the Appendix results on the mushroom dataset which satisfies interpolation. For these experiments we set $\psi(x) = ||x||_p^2$ and compare across $p \in \{1.2, 1.4, 1.6, 1.8\}$ between mSPS and the standard constant stepsize method. The first row of Figure 1 shows the training loss for the different optimizers with a softmax loss. We make the following observations: (i) mSPS performs reasonably well across different values of p and outperforms most stepsizes of SMD. (ii) mSPS performs well on ijcnn even in absence of inerpolation ($\sigma^2 > 0$).

6.2. Exponentiated gradient with ℓ_1 constraint

To test the effectiveness of mSPS with EG we consider the ijcnn and rcv1 datasets (Chang and Lin 2011) with logitistic regression where parameters are constrained to the ℓ_1 ball, $\mathcal{X} = \{x : ||x||_1 \leq \lambda\}$. To solve this problem with EG, we employ the trick of reducing an ℓ_1 ball constraint to a simplex constraint (M. Zinkevich 2003; Schuurmans and M. A. Zinkevich 2016).

For these experiments we test our optimizers on rcv1 and ijcnn and two synthetic datasets and report their results in row 2 of Figure 1. Like in the previous experiments, mSPS is significantly faster than most constant stepsizes in some cases outperforms the best tuned SMD. Note that the constant stepsizes that don't appear have diverged.

[2] We provide additional details and more experiments in the appendix.

References

Auer, Peter, Nicolò Cesa-Bianchi, and Claudio Gentile (2002). "Adaptive and Self-Confident On-Line Learning Algorithms". In: *Journal of Computer and System Sciences* 64.1, pp. 48–75. ISSN: 0022-0000. DOI: https://doi.org/10.1006/jcss.2001.1795. URL: https://www.sciencedirect.com/science/article/pii/S0022000001917957.

Bauschke, Heinz H, Jérôme Bolte, and Marc Teboulle (2017). "A descent lemma beyond Lipschitz gradient continuity: first-order methods revisited and applications". In: *Mathematics of Operations Research* 42.2, pp. 330–348.

Beck, Amir (2017). *First-Order Methods in Optimization*. Philadelphia, PA: Society for Industrial and Applied Mathematics. DOI: 10.1137/1.9781611974997. eprint: https://epubs.siam.org/doi/pdf/10.1137/1.9781611974997. URL: https://epubs.siam.org/doi/abs/10.1137/1.9781611974997.

Beck, Amir and Marc Teboulle (2003). "Mirror descent and nonlinear projected subgradient methods for convex optimization". In: *Operations Research Letters* 31.3, pp. 167–175. ISSN: 0167-6377. DOI: https://doi.org/10.1016/S0167-6377(02)00231-6. URL: https://www.sciencedirect.com/science/article/pii/S0167637702002316.

Birnbaum, Benjamin, Nikhil R Devanur, and Lin Xiao (2011). "Distributed algorithms via gradient descent for Fisher markets". In: *Proceedings of the 12th ACM conference on Electronic commerce*, pp. 127–136.

Bubeck, Sébastien (2014). "Convex optimization: Algorithms and complexity". In: *arXiv preprint arXiv:1405.4980*.

Chang, Chih-Chung and Chih-Jen Lin (2011). "LIBSVM: A library for support vector machines". In: *ACM Transactions on Intelligent Systems and Technology*. Software available at http://www.csie.ntu.edu.tw/~cjlin/libsvm.

Collins, Michael, Amir Globerson, Terry Koo, Xavier Carreras Pérez, and Peter Bartlett (2008). "Exponentiated gradient algorithms for conditional random fields and max-margin markov networks". In: *Journal of Machine Learning Research* 9, pp. 1775–1822.

Dragomir, Radu-Alexandru, Mathieu Even, and Hadrien Hendrikx (2021). "Fast Stochastic Bregman Gradient Methods: Sharp Analysis and Variance Reduction". In: *arXiv preprint arXiv:2104.09813*.

Duchi, John, Elad Hazan, and Yoram Singer (2011). "Adaptive subgradient methods for online learning and stochastic optimization." In: *Journal of machine learning research* 12.7.

Duchi, John C (2018). "Introductory lectures on stochastic optimization". In: *The mathematics of data* 25, p. 99.

Hanzely, Filip and Peter Richtárik (2018). "Fastest rates for stochastic mirror descent methods". In: *arXiv preprint arXiv:1803.07374*.

Hazan, Elad and Satyen Kale (2014). "Beyond the regret minimization barrier: optimal algorithms for stochastic strongly-convex optimization". In: *The Journal of Machine Learning Research* 15.1, pp. 2489–2512.

He, Kaiming, Xiangyu Zhang, Shaoqing Ren, and Jian Sun (2016). "Deep residual learning for image recognition". In: *CVPR*.

Li, Xiaoyu and Francesco Orabona (16–18 Apr 2019). "On the Convergence of Stochastic Gradient Descent with Adaptive Stepsizes". In: *Proceedings of the Twenty-Second International Conference on Artificial Intelligence and Statistics*. Ed. by Kamalika Chaudhuri and Masashi Sugiyama. Vol. 89. Proceedings of Machine Learning Research. PMLR, pp. 983–992. URL: http://proceedings.mlr.press/v89/li19c.html.

Loizou, Nicolas, Sharan Vaswani, Issam Hadj Laradji, and Simon Lacoste-Julien (2021). "Stochastic polyak step-size for sgd: An adaptive learning rate for fast convergence". In: *International Conference on Artificial Intelligence and Statistics*. PMLR, pp. 1306–1314.

Lu, Haihao, Robert M Freund, and Yurii Nesterov (2018). "Relatively smooth convex optimization by first-order methods, and applications". In: *SIAM Journal on Optimization* 28.1, pp. 333–354.

McMahan, H Brendan and Matthew Streeter (2010). "Adaptive bound optimization for online convex optimization". In: *COLT*.

Nemirovsky, A.S. and D.B. Yudin (1983). *Problem Complexity and Method Efficiency in Optimization*. A Wiley-Interscience publication. Wiley. ISBN: 9780471103455. URL: https://books.google.ca/books?id=6ULvAAAAMAAJ.

Orabona, Francesco (2019). "A modern introduction to online learning". In: *arXiv preprint arXiv:1912.13213*.

Orabona, Francesco and Dávid Pál (2018). "Scale-free online learning". In: *Theoretical Computer Science* 716. Special Issue on ALT 2015, pp. 50–69. ISSN: 0304-3975. DOI: https://doi.org/10.1016/j.tcs.2017.11.021. URL: https://www.sciencedirect.com/science/article/pii/S0304397517308514.

Polyak, Boris (1987). *Introduction to optimization*. New York : Optimization Software, Publications Division.

Schuurmans, Dale and Martin A Zinkevich (2016). "Deep learning games". In: *Advances in Neural Information Processing Systems*, pp. 1678–1686.

Srebro, Nathan, Karthik Sridharan, and Ambuj Tewari (2010). "Optimistic rates for learning with a smooth loss". In: *arXiv preprint arXiv:1009.3896*.

Streeter, Matthew and H Brendan McMahan (2010). “Less regret via online conditioning”. In: *arXiv preprint arXiv:1002.4862*.

Vaswani, Sharan, Aaron Mishkin, Issam Laradji, Mark Schmidt, Gauthier Gidel, and Simon Lacoste-Julien (2019). “Painless stochastic gradient: Interpolation, line-search, and convergence rates”. In: *Advances in Neural Information Processing Systems*, pp. 3727–3740.

Zinkevich, Martin (2003). “Online convex programming and generalized infinitesimal gradient ascent”. In: *Proceedings of the 20th international conference on machine learning (icml-03)*, pp. 928–936.

Appendices

Table of Contents

A. Mirror descent lemmas

Lemma 2 (Three Point Property (Bubeck 2014; Orabona 2019))**.** *Let B_ψ be the Bregman divergence with respect to $\psi : \mathcal{D} \to \mathbb{R}$. Then for any three points $x, y \in \operatorname{int} \mathcal{D}$, and $z \in \mathcal{D}$, the following holds*

$$B_\psi(z; x) + B_\psi(x; y) - B_\psi(z; y) = \langle \nabla\psi(y) - \nabla\psi(x), z - x\rangle.$$

A.1. Proof of Lemma 1

Lemma 1. *Let B_ψ be the Bregman divergence with respect to a convex function $\psi : \mathcal{D} \to \mathbb{R}$ and assume assumption 1 holds. Let $x_{t+1} = \arg\min_{x\in\mathcal{X}} \langle g_t, x\rangle + \frac{1}{\eta_t} B_\psi(x; x_t)$. Then for any $x_* \in \mathcal{X}$*

$$B_\psi(x_*; x_{t+1}) \leq B_\psi(x_*; x_t) - \eta_t\langle g_t, x_t - x_*\rangle - B_\psi(x_{t+1}; x_t) + \eta_t\langle g_t, x_t - x_{t+1}\rangle. \tag{7}$$

Furthermore if ψ is μ_ψ strongly convex over $\mathcal{X}$ then

$$B_\psi(x_*; x_{t+1}) \leq B_\psi(x_*; x_t) - \eta_t\langle g_t, x_t - x_*\rangle + \frac{\eta_t^2}{2\mu_\psi} \|g_t\|_*^2 . \tag{8}$$

Proof. The proof follows closely to the one presented in Orabona (2019)[Lemma 6.7]. First observe that x_{t+1} statisfies the first order optimality condition

$$\langle \eta_t g_t + \nabla\psi(x_{t+1}) - \nabla\psi(x_t), x_* - x_{t+1}\rangle \geq 0,$$

since $\nabla_x B_\psi(x; x_t) = \nabla\psi(x) - \nabla\psi(x_t)$.

We start by examining the inner product $\langle \eta_t g_t, x_t - x_*\rangle$ and adding subtracting quantities to make the first order optimality condition appear.

$$\begin{aligned}\langle \eta_t g_t, x_t - x_*\rangle &= \langle \eta_t g_t + \nabla\psi(x_{t+1}) - \nabla\psi(x_t), x_{t+1} - x_*\rangle + \langle \nabla\psi(x_{t+1}) - \nabla\psi(x_t), x_* - x_{t+1}\rangle + \langle \eta_t g_t, x_t - x_{t+1}\rangle \\ &\leq \langle \nabla\psi(x_{t+1}) - \nabla\psi(x_t), x_* - x_{t+1}\rangle + \langle \eta_t g_t, x_t - x_{t+1}\rangle \text{(first order optimality)} \\ &= B_\psi(x_*; x_t) - B_\psi(x_*; x_{t+1}) - B_\psi(x_{t+1}; x_t) + \langle \eta_t g_t, x_t - x_{t+1}\rangle \text{ (three point property).}\end{aligned}$$

Rearranging gives the first result. Note at this point we only require ψ to be convex and ψ to be differentiable at x_t and x_{t+1}, which is guaranteed by assumption 1. To obtain the second result, observe

$$\begin{aligned}\langle \eta_t g_t, x_t - x_*\rangle &\leq B_\psi(x_*; x_t) - B_\psi(x_*; x_{t+1}) - B_\psi(x_{t+1}; x_t) + \langle \eta_t g_t, x_t - x_{t+1}\rangle \text{ (from above)} \\ &\leq B_\psi(x_*; x_t) - B_\psi(x_*; x_{t+1}) - \frac{\mu_\psi}{2}\left|\left|x_{t+1} - x_t\right|\right|^2 + \langle \eta_t g_t, x_t - x_{t+1}\rangle \text{ (strong convexity)} \\ &\leq B_\psi(x_*; x_t) - B_\psi(x_*; x_{t+1}) + \frac{\eta_t^2}{2\mu_\psi}\left|\left|g_t\right|\right|_*^2 \text{ (Fenchel-Young inequality).}\end{aligned}$$

Rearranging gives the second result. □

B. Non-smooth analysis of mirror SPS for Lipschitz functions

As we have already mentioned in the main paper, the Polyak step-size is used extensively in the literature of projected subgradient descent for solving non-smooth optimization problems. However to the best of our knowledge there is no efficient generalization of this step-size for the more general mirror descent update.

Theorem 5 (Non-smooth deterministic)**.** *Assume f is convex with bounded subgradients, $\left|\left|\partial f(x_t)\right|\right|_* \leq G$. Let ψ be μ_ψ strongly convex with respect to the norm $||\cdot||$, and assume that Assumption 1 holds. Then mirror descent with stepsize $\eta_t = \frac{\mu_\psi(f(x_t) - f(x_*))}{||\partial f(x_t)||_*^2}$ satisfies,*

$$f\left(\bar{x}_t\right) - f(x_*) \leq G\sqrt{\frac{\frac{2}{\mu_\psi}B_\psi(x_*; x_1)}{t}},$$

where $\bar{x}_t = \frac{1}{t}\sum_{s=1}^t x_s$. The same result holds for the best iterate $f(x_t^*) = \min_s\{f(x_s)\}_{1\leq s\leq t}$.

Proof. Let g_t be a subgradient of f at x_t used to compute η_t. Then by Lemma 1 we have

$$\begin{aligned}B_\psi(x_*; x_{t+1}) &\leq B_\psi(x_*; x_t) - \eta_t\langle g_t, x_t - x_*\rangle + \frac{\eta_t^2}{2\mu_\psi}\left|\left|g_t\right|\right|_*^2 \\ &\leq B_\psi(x_*; x_t) - \eta_t(f(x_t) - f(x_*)) + \frac{\eta_t^2}{2\mu_\psi}\left|\left|g_t\right|\right|_*^2 \text{ (by convexity)} \\ &= B_\psi(x_*; x_t) - \frac{\mu_\psi\left(f(x_t) - f(x_*)\right)^2}{\left|\left|g_t\right|\right|_*^2} + \frac{\mu_\psi\left(f(x_t) - f(x_*)\right)^2}{2\left|\left|g_t\right|\right|_*^2} \text{ (by definition of } \eta_t\text{)} \\ &= B_\psi(x_*; x_t) - \frac{\mu_\psi\left(f(x_t) - f(x_*)\right)^2}{2\left|\left|g_t\right|\right|_*^2}.\end{aligned}$$

Rearranging and summing across time we have

$$\sum_{s=1}^t \frac{\mu_\psi\left(f(x_s) - f(x_*)\right)^2}{2\left|\left|g_s\right|\right|_*^2} \leq B_\psi(x_*; x_1) - B_\psi(x_*; x_{t+1}) \leq B_\psi(x_*; x_1).$$

Applying the upper bound $||g_s||_* \leq G$ and taking the square root gives,

$$\sum_{s=1}^{t} (f(x_s) - f(x_*))^2 \leq G\sqrt{\frac{2B_\psi(x_*;x_1)}{\mu_\psi}}.$$

The result then follows by the convexity of f and concavity of the square root function,

$$f(\bar{x}_t) - f(x_*) \leq \frac{1}{t}\sum_{s=1}^{t}(f(x_s) - f(x_*)) = \frac{1}{t}\sum_{s=1}^{t}\sqrt{(f(x_s) - f(x_*))^2} \leq \sqrt{\frac{1}{t}\sum_{s=1}^{t}(f(x_s) - f(x_*))^2}$$
$$\leq G\sqrt{\frac{2B_\psi(x_*;x_1)}{t\mu_\psi}}.$$

To obtain the best iterate result notice that $f(x_t^*) - f(x_*)) \leq \frac{1}{t}\sum_{s=1}^{t}(f(x_s) - f(x_*))$. □

C. Proof of mSPS lower bound in section 3

The lower bound of mSPS (6) when f_i is L smooth, restated below, is vital to our analysis,

$$\frac{\mu_\psi}{2cL} \leq \eta_t = \frac{\mu_\psi(f_i(x_t) - f_i^*)}{c\,||\nabla f(x_t)||_*^2}.$$

Notice the above inequality is equivalent to

$$\frac{1}{2L} \leq \frac{(f_i(x_t) - f_i^*)}{||\nabla f(x_t)||_*^2}.$$

The first inequality is attained by multiplying both sides by μ_ψ/c. We provide a detailed proof below.

Lemma 3. *If $f : \mathbb{R}^n \to \mathbb{R}$ is L-smooth with respect to a norm $||\cdot||$ then*

$$\frac{||\nabla f(x)||_*^2}{2L} \leq f(x) - \inf_{y\in\mathbb{R}^n} f(y).$$

Rearranging and defining $f^* = \inf_{y\in\mathbb{R}^n} f(y)$ gives

$$\frac{1}{2L} \leq \frac{f(x) - f^*}{||\nabla f(x)||_*^2}.$$

Proof. Since f is L-smooth we have

$$f(y) \leq f(x) + \langle\nabla f(x), y - x\rangle + \frac{L}{2}||x - y||^2 \quad \forall x, y \in \mathbb{R}^n.$$

Therefore we have the following upper bound on $\inf_y f(y)$.

$$\begin{aligned}
\inf_y f(y) &\leq \min_y \left\{ f(x) + \langle \nabla f(x), y - x\rangle + \frac{L}{2} ||x-y||^2 \right\} \\
&= \min_{r\geq 0, ||z||\leq 1} \left\{ f(x) + r\langle \nabla f(x), z\rangle + \frac{L}{2} r^2 ||z||^2 \right\} \\
&\leq \min_{r\geq 0, ||z||\leq 1} \left\{ f(x) + r\langle \nabla f(x), z\rangle + \frac{L}{2} r^2 \right\} \\
&= f(x) + \min_{r\geq 0} \left\{ \min_{||z||\leq 1} \{ r\langle \nabla f(x), z\rangle \} + \frac{L}{2} r^2 \right\} \\
&= f(x) + \min_{r\geq 0} \left\{ -r \max_{||z||\leq 1} \{ \langle \nabla f(x), -z\rangle \} + \frac{L}{2} r^2 \right\} \\
&= f(x) + \min_{r\geq 0} \left\{ -r ||\nabla f(x)||_* + \frac{L}{2} r^2 \right\} \text{ by the definition of } ||\cdot||_* \\
&\overset{(r=||\nabla f(x)||_*/L)}{=} f(x) - \frac{||\nabla f(x)||_*^2}{L} + \frac{||\nabla f(x)||_*^2}{2L}
\end{aligned}$$

Simplifying and rearranging gives the result. □

D. Proofs for section 4

In this section we provide proofs of our main results in the relative smooth setting. For convenience we denote the expectation over index $i \in \{1, \cdots, n\}$ conditional on knowing x_t as $\mathbb{E}_t[\cdot]$.

First we provide the following lemma which allows us to bound the last two terms in (Lemma 1). This result can be seen as a generalization of Lemma 2 in Collins et al. (2008), where the exponentiated gradient algorithm is studied under the relative smoothness assumption.

Lemma 4. *Suppose f is L smooth relative to ψ. Then if $\eta \leq \frac{1}{L}$ we have*

$$-B_\psi(x_{t+1}; x_t) + \eta\langle \nabla f(x_t), x_t - x_{t+1}\rangle \leq \eta(f(x_t) - f(x_{t+1})).$$

Proof. Since f is L smooth relative to ψ it is also $\frac{1}{\eta}$ smooth relative to ψ (because $L \leq \frac{1}{\eta}$ and ψ is convex). Therefore,

$$\begin{aligned}
B_f(x_{t+1}; x_t) &\leq \frac{1}{\eta} B_\psi(x_{t+1}; x_t) \\
\implies -B_\psi(x_{t+1}; x_t) + \eta B_f(x_{t+1}; x_t) &\leq 0.
\end{aligned}$$

Now we examine the inner product $\eta\langle \nabla f(x_t), x_t - x_{t+1}\rangle$,

$$\begin{aligned}
\eta\langle \nabla f(x_t), x_t - x_{t+1}\rangle &= \eta\left(f(x_{t+1}) - f(x_t) - \langle \nabla f(x_t), x_{t+1} - x_t\rangle + f(x_t) - f(x_{t+1})\right) \\
&= \eta\left(B_f(x_{t+1}; x_t) + f(x_t) - f(x_{t+1})\right).
\end{aligned}$$

Therefore, we have the following

$$\begin{aligned}
-B_\psi(x_{t+1}; x_t) + \eta\langle \nabla f(x_t), x_t - x_{t+1}\rangle &= -B_\psi(x_{t+1}; x_t) + \eta B_f(x_{t+1}; x_t) + \eta(f(x_t) - f(x_{t+1}) \\
&\leq \eta(f(x_t) - f(x_{t+1}).
\end{aligned}$$

□

D.1. Proof of Theorem 1

Theorem 1. *Assume ψ satisfies assumption 1 and is strictly convex. Furthermore assume f to be μ strongly convex relative to ψ, and f_i to be L-smooth relative to ψ. Then stochastic mirror decent with stepsize $\eta \leq \frac{1}{L}$ guarantees*

$$\mathbb{E}\left[B_\psi(x_*; x_{t+1})\right] \leq (1 - \mu\eta)^t B_\psi(x_*; x_1) + \frac{\sigma^2}{\mu}.$$

Proof. From Lemma 1 (before applying strong convexity but assuming convexity of ψ) we have

$$\begin{aligned} B_\psi(x_*;x_{t+1}) &\le B_\psi(x_*;x_t) - \eta\langle\nabla f_i(x_t), x_t - x_*\rangle - B_\psi(x_{t+1};x_t) + \eta\langle\nabla f_i(x_t), x_t - x_{t+1}\rangle \\ &\le B_\psi(x_*;x_t) - \eta\langle\nabla f_i(x_t), x_t - x_*\rangle + \eta(f_i(x_t) - f_i(x_{t+1})) \text{ (by Lemma 4)} \\ &\le B_\psi(x_*;x_t) - \eta\langle\nabla f_i(x_t), x_t - x_*\rangle + \eta(f_i(x_t) - f_i^*) \text{ (by definition of } f_i^*) \\ &= B_\psi(x_*;x_t) - \eta\langle\nabla f_i(x_t), x_t - x_*\rangle + \eta(f_i(x_t) - f_i(x_*)) + \eta(f_i(x_*) - f_i^*). \end{aligned}$$

By taking an expectation conditioning on x_t we obtain,

$$\begin{aligned} \mathbb{E}_t\left[B_\psi(x_*;x_{t+1})\right] &\le B_\psi(x_*;x_t) - \eta\langle\nabla f(x_t), x_t - x_*\rangle + \eta(f(x_t) - f(x_*)) + \eta(f(x_*) - \mathbb{E}_t\left[f_i^*\right]) \\ &= B_\psi(x_*;x_t) - \eta\underbrace{\left(f(x_*) - f(x_t) - \langle\nabla f(x_t), x_* - x_t\rangle\right)}_{B_f(x_*;x_t)} + \eta(f(x_*) - \mathbb{E}_t\left[f_i^*\right]) \qquad (9) \\ &\le B_\psi(x_*;x_t)(1-\mu\eta) + \eta(f(x_*) - \mathbb{E}_t\left[f_i^*\right]) \text{ (by relative strongly convexity of } f). \end{aligned}$$

Now by the tower property of expectations and applying the definition of σ^2,

$$\mathbb{E}\left[B_\psi(x_*;x_{t+1})\right] \le \mathbb{E}\left[B_\psi(x_*;x_t)\right](1-\mu\eta) + \eta\sigma^2.$$

Iterating the inequality gives,

$$\begin{aligned} \mathbb{E}\left[B_\psi(x_*;x_{t+1})\right] &\le B_\psi(x_*;x_1)(1-\mu\eta)^t + \sum_{s=0}^{t-1}\eta\sigma^2(1-\mu\eta)^s \\ &\le B_\psi(x_*;x_1)(1-\mu\eta)^t + \frac{\sigma^2}{\mu}. \end{aligned}$$

Where the last inequality follows by $\sum_{s=0}^{t-1}(1-\mu\eta)^s \le \sum_{s=0}^{\infty}(1-\mu\eta)^s = {}^1\!/_{\mu\eta}$. □

D.2. Proof of Theorem 2

Theorem 2. *Assume ψ satisfies assumption 1. Furthermore assume f_i to L-smooth relative to ψ. Then stochastic mirror decent with stepsize $\eta \le \frac{1}{L}$ guarantees*

$$\mathbb{E}\left[\frac{1}{t}\sum_{s=1}^{t} B_f(x_*;x_s)\right] \le \frac{B_\psi(x_*;x_1)}{\eta t} + \sigma^2.$$

Proof. Note that in the proof of Theorem 1 relative strong convexity is not used to attain the inequality (9). Therefore we have,

$$\mathbb{E}_t\left[B_\psi(x_*;x_{t+1})\right] \le B_\psi(x_*;x_t) - \eta B_f(x_*;x_t) + \eta(f(x_*) - \mathbb{E}_t\left[f_i^*\right]).$$

After applying the tower property, definition of σ^2, and rearranging, we have

$$\eta\mathbb{E}\left[B_f(x_*;x_t)\right] \le \mathbb{E}\left[B_\psi(x_*;x_t)\right] - \mathbb{E}\left[B_\psi(x_*;x_{t+1})\right] + \eta\sigma^2.$$

Summing across time and dividing by ηt gives the result. □

E. Proofs for section 5

In this section we provide proofs of our main results in the smooth setting. For convenience we denote the expectation over index $i \in \{1, \cdots, n\}$ conditional on knowing x_t as $\mathbb{E}_t\left[\cdot\right]$.

Notice that by definition of mSPS$_{\text{max}}$ we have the following upper bound

$$\eta_t \le \frac{\mu_\psi(f_i(x_t) - f_i^*)}{c\,||\nabla f_i(x_t)||_*^2}.$$

Muliplying both sides of the inequality with $\eta_t||\nabla f_i(x_t)||_*^2/\mu_\psi$ gives the following useful inequality,

$$\frac{\eta_t^2\,||\nabla f_i(x_t)||_*^2}{\mu_\psi} \leq \frac{\eta_t(f_i(x_t)-f_i^*)}{c}. \tag{10}$$

The inequality holds with equality for mSPS.

E.1. Proof of Theorem 3

Theorem 3. *Assume f_i are convex and L_i smooth with respect to the norm $||\cdot||$. Furthermore, assume that f is μ strongly convex relative to ψ, where ψ is μ_ψ strongly convex with respect to the norm $||\cdot||$ and assumption 1 holds. Then stochastic mirror descent with mSPS_{max} and $c \geq \frac{1}{2}$ guarantees*

$$\mathbb{E}\left[B_\psi(x_*;x_{t+1})\right] \leq (1-\mu\alpha)^t B_\psi(x_*;x_1) + \frac{\eta_b\sigma^2}{\alpha\mu}.$$

Where $\alpha := \min\{\mu_\psi/2cL_{\max}, \eta_b\}$ and $L_{\max} = \max_i\{L_i\}_{i=1}^n$.

Proof.

$$\begin{aligned}
B_\psi(x_*;x_{t+1}) &\leq B_\psi(x_*;x_t) - \eta_t\langle\nabla f_i(x_t), x_t - x_*\rangle + \frac{\eta_t^2}{2\mu_\psi}\,||\nabla f_i(x_t)||_*^2 \\
&\overset{(10)}{\leq} B_\psi(x_*;x_t) - \eta_t\langle\nabla f_i(x_t), x_t - x_*\rangle + \eta_t\frac{(f_i(x_t)-f_i^*)}{2c} \\
&\overset{(c\geq 1/2)}{\leq} B_\psi(x_*;x_t) - \eta_t\langle\nabla f_i(x_t), x_t - x_*\rangle + \eta_t(f_i(x_t)-f_i^*) \\
&= B_\psi(x_*;x_t) - \eta_t\langle\nabla f_i(x_t), x_t - x_*\rangle + \eta_t(f_i(x_t) - f_i(x_*) + f_i(x_*) - f_i^*) \\
&= B_\psi(x_*;x_t) - \underbrace{\eta_t\left(f_i(x_*) - f_i(x_t) - \langle\nabla f_i(x_t), x_* - x_t\rangle\right)}_{\geq 0} + \eta_t(f_i(x_*) - f_i^*) \\
&\overset{(6)}{\leq} B_\psi(x_*;x_t) - \min\left\{\frac{\mu_\psi}{2cL_i}, \eta_b\right\}\left(f_i(x_*) - f_i(x_t) - \langle\nabla f_i(x_t), x_* - x_t\rangle\right) + \eta_b(f_i(x_*) - f_i^*)
\end{aligned}$$

Taking an expectation over i condition on x_t gives

$$\begin{aligned}
\mathbb{E}_t\left[B_\psi(x_*;x_{t+1})\right] &\leq B_\psi(x_*;x_t) - \min\left\{\frac{\mu_\psi}{2cL_i}, \eta_b\right\}\left(f(x_*) - f(x_t) - \langle\nabla f(x_t), x_* - x_t\rangle\right) + \eta_b\mathbb{E}_t\left[(f_i(x_*) - f_i^*)\right] \\
&\leq B_\psi(x_*;x_t)\left(1 - \mu\min\left\{\frac{\mu_\psi}{2cL_{\max}}, \eta_b\right\}\right) + \eta_b\mathbb{E}_t\left[(f_i(x_*) - f_i^*)\right] \text{ (by relative strong convexity of } f) \\
&= B_\psi(x_*;x_t)\left(1-\mu\alpha\right) + \eta_b\mathbb{E}_t\left[(f_i(x_*) - f_i^*)\right].
\end{aligned}$$

Now by the tower property of expectations and applying the definition of σ^2,

$$\mathbb{E}\left[B_\psi(x_*;x_{t+1})\right] \leq \mathbb{E}\left[B_\psi(x_*;x_t)\right](1-\mu\alpha) + \eta_b\sigma^2.$$

Iterating the inequality gives,

$$\begin{aligned}
\mathbb{E}\left[B_\psi(x_*;x_{t+1})\right] &\leq B_\psi(x_*;x_1)(1-\mu\alpha)^t + \sum_{s=0}^{t-1}\eta_b\sigma^2(1-\mu\alpha)^s \\
&\leq B_\psi(x_*;x_1)(1-\mu\alpha)^t + \frac{\eta_b\sigma^2}{\alpha\mu}.
\end{aligned}$$

Where the last inequality follows by $\sum_{s=0}^{t-1}(1-\mu\alpha)^s \leq \sum_{s=0}^{\infty}(1-\mu\alpha)^s = 1/\mu\alpha$. □

E.2. Proof of Theorem 4

Theorem 4. *Assume f_i are convex and L_i smooth with respect to a norm $||\cdot||$, assumption 1 holds, and ψ is μ_ψ strongly convex with respect to the norm $||\cdot||$. Then stochastic mirror descent with mSPS$_{max}$ and $c \geq 1$ guarantees*

$$\mathbb{E}\left[f(\bar{x}_t) - f(x_*)\right] \leq \frac{2B_\psi(x_*;x_1)}{\alpha t} + \frac{2\eta_b\sigma^2}{\alpha}.$$

Where $\alpha := \min\{\mu_\psi/2cL_{\max}, \eta_b\}$ and $L_{\max} = \max_i\{L_i\}_{i=1}^n$.

Proof. We begin with Lemma 1,

$$\begin{aligned}
B_\psi(x_*;x_{t+1}) &\leq B_\psi(x_*;x_t) - \eta_t\langle\nabla f_i(x_y), x_t - x_*\rangle + \frac{\eta_t^2}{2\mu_\psi}\left|\left|\nabla f_i(x_t)\right|\right|_*^2 \\
&\leq B_\psi(x_*;x_t) - \eta_t\left(f_i(x_t) - f_i(x_*)\right) + \frac{\eta_t^2}{2\mu_\psi}\left|\left|\nabla f_i(x_t)\right|\right|_*^2 \text{ by convexity} \\
&\overset{(10)}{\leq} B_\psi(x_*;x_t) - \eta_t\left(f_i(x_t) - f_i(x_*)\right) + \frac{\eta_t(f_i(x_t) - f_i^*)}{2c} \\
&\overset{(c\geq1)}{\leq} B_\psi(x_*;x_t) - \eta_t\left(f_i(x_t) - f_i(x_*)\right) + \frac{\eta_t(f_i(x_t) - f_i^*)}{2} \\
&= B_\psi(x_*;x_t) - \eta_t\left(f_i(x_t) - f_i^* + f_i^* - f_i(x_*)\right) + \frac{\eta_t(f_i(x_t) - f_i^*)}{2} \\
&= B_\psi(x_*;x_t) - \eta_t\left(1 - \frac{1}{2}\right)\left(f_i(x_t) - f_i^*\right) + \eta_t(f_i(x_*) - f_i^*) \\
&= B_\psi(x_*;x_t) - \frac{\eta_t}{2}\underbrace{\left(f_i(x_t) - f_i^*\right)}_{\geq 0} + \eta_t(f_i(x_*) - f_i^*) \\
&\overset{(6)}{\leq} B_\psi(x_*;x_t) - \frac{\alpha}{2}\left(f_i(x_t) - f_i^*\right) + \eta_b(f_i(x_*) - f_i^*) \\
&= B_\psi(x_*;x_t) - \frac{\alpha}{2}\left(f_i(x_t) - f_i(x_*)\right) - \frac{\alpha}{2}\left(f_i(x_*) - f_i^*\right) + \eta_b(f_i(x_*) - f_i^*) \\
&\leq B_\psi(x_*;x_t) - \frac{\alpha}{2}\left(f_i(x_t) - f_i(x_*)\right) + \eta_b(f_i(x_*) - f_i^*)
\end{aligned}$$

Recall from (6) that we have

$$\alpha = \min\left\{\frac{\mu_\psi}{2cL_{\max}}, \eta_b\right\} \leq \eta_t \leq \eta_b.$$

By a simple rearrangement we have

$$\frac{\alpha}{2}\left(f_i(x_t) - f_i^*\right) \leq B_\psi(x_*;x_t) - B_\psi(x_*;x_{t+1}) + \eta_b(f_i(x_*) - f_i^*).$$

Taking an expectation on both sides, dividing by α, and applying the definition of σ^2 yields

$$\mathbb{E}\left[f(x_t) - f(x_*)\right] \leq \frac{2}{\alpha}\left(\mathbb{E}\left[B_\psi(x_*;x_t)\right] - \mathbb{E}\left[B_\psi(x_*;x_{t+1})\right]\right) + \frac{2\eta_b}{\alpha}\sigma^2.$$

Summing across time, applying convexity of f, and dividing by t gives

$$\mathbb{E}\left[f(\bar{x}_t) - f(x_*)\right] \leq \frac{1}{t}\sum_{s=1}^{t}\mathbb{E}\left[f(x_s) - f(x_*)\right] \leq \frac{2B_\psi(x_*;x_1)}{\alpha t} + \frac{2\eta_b\sigma^2}{\alpha}.$$

□

E.2.1. Constant stepsize corollary

In this section we present the constant stepsize corollary for Theorem 4. If $\eta_b \leq \mu_\psi/2L_{\max}$ then mSPS$_{\text{max}}$ with $c = 1$ is a constant stepsize because of the lower bound (6), $\eta_t = \eta_b$, and we have that $\eta_b = \alpha$. Therefore plugging in these values into Theorem 4 gives the following corollary.

Corollary 8. *Assume f_i are convex and L_i smooth with respect to a norm $||\cdot||$, assumption 1 holds, and ψ is μ_ψ strongly convex with respect to the norm $||\cdot||$. Then stochastic mirror descent with $\eta \leq \mu_\psi/2L_{\max}$ guarantees*

$$\mathbb{E}\left[f(\bar{x}_t) - f(x_*)\right] \leq \frac{2B_\psi(x_*; x_1)}{\eta t} + 2\sigma^2.$$

Where $L_{\max} = \max_i\{L_i\}_{i=1}^n$.

F. Experiment details

In this section we provide details for our experiments including the updates for different mirror descent algorithms. Note that in all our experiments we have $f_i^* = 0$, and for the constant stepsize we sweep over $\{10^{-5}, 10^{-4}, 10^{-3}, 10^{-2}, 10^{-1}, 1, 10^1, 10^2, 10^3, 10^4, 10^5\}$.

F.1. Compute resources

We ran around a thousand experiments using an internal cluster, where each experiment uses a single NVIDIA Tesla P100 GPU, 40GB of RAM, and 4 CPUs. Some experiments like the synthetic ones took only few minutes to complete, while the deep learning experiments like CIFAR10 took about 12 hours.

F.2. Mirror descent across p-norms

We select $\psi(x) = \frac{1}{2}||x||_p^2$ and $\mathcal{X} = \mathbb{R}^d$ for $1 < p \leq 2$. We have in this case that ψ is $\mu_\psi = (p-1)$ strongly convex with respect to the norm $||\cdot||_p$ with dual norm $||\cdot||_q$ where q is such that $1/p + 1/q = 1$ (Orabona 2019). Therefore, mSPS$_{\text{max}}$ with $c = 1$ is

$$\eta_t = \min\left\{\frac{(p-1)(f_i(x_t) - f_i^*)}{||\nabla f_i(x_t)||_q^2}, \eta_b\right\},$$

and similarly for mSPS.

The closed form update for mirror descent in this case is given by the following coordinate wise updates (J. C. Duchi 2018): let $\phi^p : \mathbb{R}^d \to \mathbb{R}^d$ with component functions $\phi_i^p(x) = (||x||_p)^{2-p}\,\text{sign}(x_i)|x_i|^{p-1}$, then the mirror descent update with stepsize η_t is

$$x_{t+1} = \phi^q(\phi^p(x_t) - \eta_t \nabla f_i(x_t)).$$

F.3. Exponentiated gradient with ℓ_1 constraint

We consider the case of supervised learning with constraint set $\mathcal{X} = \{x : ||x||_1 \leq \lambda\}$. To consider the exponentiated gradient algorithm we equivalently write the set $\mathcal{X}$ as a convex hull of its corners, $\mathcal{X} = \{\Lambda x : x \in \Delta_{2d}\}$ where Δ_{2d} is the $2d$-dimensional probability simplex and Λ is a matrix with $2d$ columns and d rows,

$$\Lambda = \begin{bmatrix} \lambda & -\lambda & 0 & 0 & \cdots & 0 & 0 \\ 0 & 0 & \lambda & -\lambda & \cdots & 0 & 0 \\ 0 & 0 & 0 & 0 & \cdots & 0 & 0 \\ \vdots & \vdots & \vdots & \vdots & \cdots & \vdots & \vdots \\ 0 & 0 & 0 & 0 & \cdots & \lambda & -\lambda \end{bmatrix}.$$

Therefore we can use the exponentiated algorithm with constraint set Δ_{2d} by selecting $\psi(x) = \sum_{i=1}^{2d} x_i \log(x_i)$. In this case ψ is $\mu_\psi = 1$ strongly convex on Δ_{2d} with respect to the norm $||\cdot||_1$. Since the dual norm $||\cdot||_* = ||\cdot||_\infty$ we have that mSPS$_{\text{max}}$ with $c = 1$ is

$$\eta_t = \min\left\{\frac{(f_i(x_t) - f_i^*)}{||\nabla f_i(x_t)||_\infty^2}, \eta_b\right\},$$

and similarly for mSPS.

The mirror descent update then can be written in two steps (Bubeck 2014),

$$y_{t+1} = x_t \odot \exp(-\eta_t \nabla f_i(x_t))$$
$$x_{t+1} = \frac{y_{t+1}}{||y_{t+1}||_1}.$$

Where $\odot$ and $\exp$ are component wise multiplication and component wise exponentiation respectively.

F.4. Additional Results across p-norms

We observe in Figure 2 that mSPS outperforms a large grid of step-sizes for most values of p. Note that we used the mushrooms dataset with the kernel bandwidth selected in Vaswani et al. (2019) which satisfies interpolation.

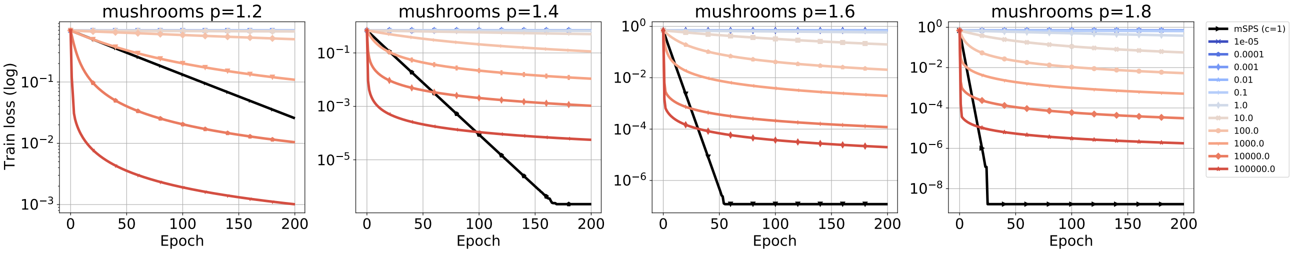

Figure 2. Comparison between mSPS with $c = 1$ and constant step-sizes on convex binary-classification problem on the mushroom dataset.

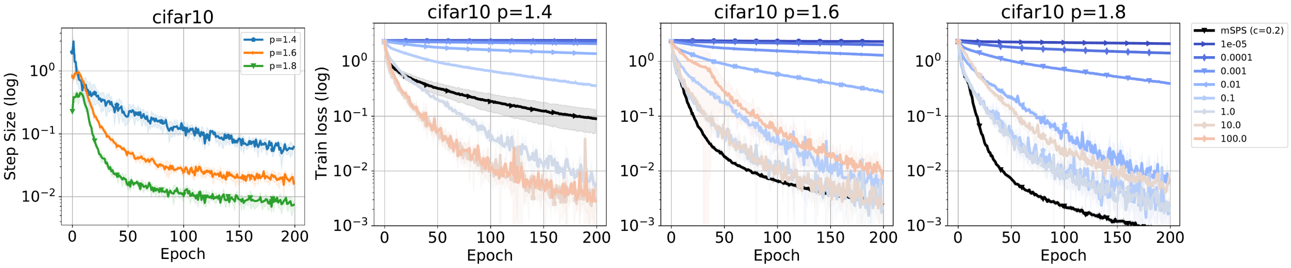

Figure 3. Comparison between mSPS with $c = 0.2$ and constant step-sizes on non-convex multiclass classification with deep networks. The leftmost plot shows the step-size evolution for different p values.

For mutliclass-classification with deep networks, we considered the p-norm algorithms for the CIFAR10 dataset and we set $c = 0.2$, as recommended by Loizou et al. (2021). CIFAR10 has 10 classes and we used the standard training set consisting of 50k examples and a test set of 10k. As in the kernel experiments, we evaluated the optimizers using the softmax loss for different values of p. We used the experimental setup proposed in Loizou et al. (2021) and used a batch-size of 128 for all methods and datasets. We used the standard image-classification architecture ResNet-34 (He et al. 2016). As in the other experiments, each optimizer was run with five different random seeds in the final experiment. The optimizers were run until the performance of most methods saturated; 200 epochs for the models on the CIFAR10 dataset.

From Figure 3, we observe that: (i) mSPS with $c = 0.2$ constantly converges to a good solution much faster when compared to most constant stepsizes. (ii) The gap between the performance of mSPS and constant stepsize increases as p decreases suggesting that, like in the convex setting, our method is robust to different values of p. Note that we used smoothing for computing the SPS step-size which was recommended in Loizou et al. (2021) for the deep learning experiments.