Stochastic linear optimization never overfits with quadratically-bounded losses on general data

Matus Telgarsky <mjt@illinois.edu>

Editors: Po-Ling Loh and Maxim Raginsky

Abstract

This work provides test error bounds for iterative fixed point methods on linear predictors — specifically, stochastic and batch mirror descent (MD), and stochastic temporal difference learning (TD) — with two core contributions: (a) a single proof technique which gives high probability guarantees despite the absence of projections, regularization, or any equivalents, even when optima have large or infinite norm, for quadratically-bounded losses (e.g., providing unified treatment of squared and logistic losses); (b) locally-adapted rates which depend not on global problem structure (such as conditions numbers and maximum margins), but rather on properties of low norm predictors which may suffer some small excess test error. The proof technique is an elementary and versatile coupling argument, and is demonstrated here in the following settings: stochastic MD under realizability; stochastic MD for general Markov data; batch MD for general IID data; stochastic MD on heavy-tailed data (still without projections); stochastic TD on approximately mixing Markov chains (all prior stochastic TD bounds are in expectation).

1. Introduction

This work studies iterative fixed point methods resembling (stochastic) gradient descent, specifically

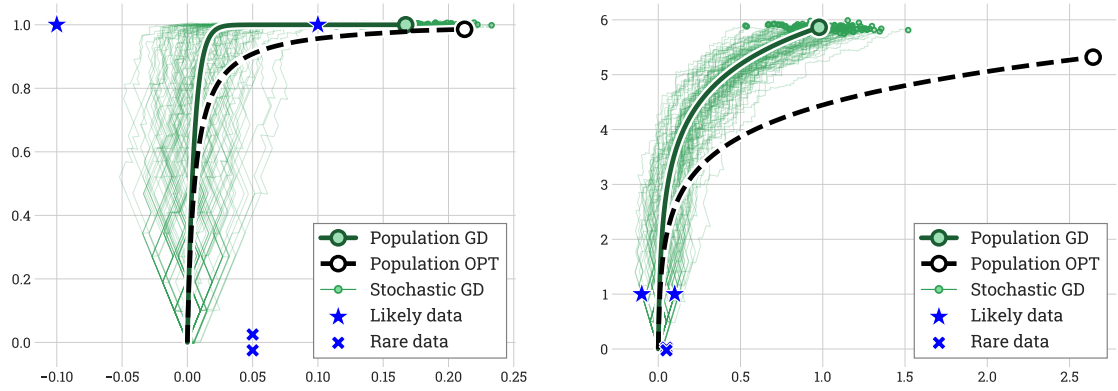
$$\text{gradient descent (GD)}, \quad w_{i+1} := w_i - \eta g_{i+1}, \tag{1}$$

$$\text{mirror descent (MD)}, \quad w_{i+1} := \arg\min \left\{ \langle \eta g_{i+1}, w \rangle + D_\psi(w, w_i) \ : \ w \in S \right\}, \tag{2}$$

$$\text{temporal difference learning (TD)}, \quad w_{i+1} := w_i - \eta G_{i+1}(w_i), \tag{3}$$

where g_{i+1} is stochastic or batch gradient, G_{i+1} is a superficially similar affine mapping related to the Bellman error in reinforcement learning, and D_ψ is a Bregman divergence. (Details will come in Section 1.1.)

The goal here is to control the excess risk of these procedures with high probability, meaning to ensure that $\frac{1}{t}\sum_{i<t} \mathcal{R}(w_i) - \mathcal{R}(w_{\text{ref}})$ is small, where the risk $\mathcal{R}$ in the simplest setting is $\mathcal{R}(w) = \text{E}x_{,y}\ell(y, w^\mathsf{T}x)$ with ℓ a *quadratically-bounded loss*, which roughly speaking means $\|\partial_w \ell(y, w^\mathsf{T}x)\|_*$ can be related to $\|w - w_{\text{ref}}\|$ (cf. Section 1.1), and w_{ref} is some good (but not necessarily optimal) *reference solution*. So far, this is not too outlandish, however the focus of the work is (a) a single proof technique for all methods and settings without projections (e.g., $S = \mathbb{R}^d$ in MD in eq. (2)), constraints, or regularization, despite the possibility of all minimizers being at infinity, as is often the case in practice, and (b) rates which depend not on global structure, but rather on the behavior of reasonably good but low norm predictors. In more detail, these two contributions and their relationship with prior work is as follows.



(*a*) Squared loss, step size $\eta = 0.1$.

(*b*) Logistic loss, step size $\eta = 1$.

Figure 1: 100 parallel runs of 400 SGD iterations on the same data distribution in $\mathbb{R}^2$ using the squared loss (cf. Figure 1(*a*)) and the logistic loss (cf. Figure 1(*b*)). The distribution consists of two "likely" upper data points (sampled with probability 90%), and two "rare" lower data points (sampled with probability 10%), all with common label $+1$. Three types of trajectory are depicted: "population OPT" is the curve of optimal constrained solutions $\{\arg\min_{\|w\|\leq B} \mathcal{R}(w) : B \geq 0\}$, "population GD" is GD applied to the population risk $\mathcal{R}$, and "stochastic GD" uses a fresh sample for each update. In all cases, the methods procrastinate convergence towards their asymptotic destinations, respectively the minimum norm and maximum margin solutions, and instead spend a good deal of time heading upwards, specifically towards low norm, low risk solutions. Analyzing this early trend is a goal of the present work, which is achieved through appropriate choices of w_{ref}, as detailed in the illustrative examples in Sections 2 and 3.

1. **Single coupling-based proof technique.** The core contribution is a single proof technique which can handle MD (which generalizes GD) and TD, obtaining excess risk rates with high probability without projections, constraints, regularization, or equivalents. Despite the extensive history and development of these methods throughout machine learning and optimization (Robbins and Monro, 1951; Nemirovski and Yudin, 1983; Bottou, 2010; Kingma and Ba, 2014), prior work either requires projections, regularization, and constraints (Rakhlin et al., 2012; Harvey et al., 2019), or makes noise and comparator assumptions which effectively necessitate bounded iterates (Li and Orabona, 2020), or it provides bounds only in expectation (Hardt et al., 2016), or is tailored to specific data and loss settings, for instance exponentially-tailed losses and linearly separable data (Soudry et al., 2017; Ji and Telgarsky, 2018b; Shamir, 2021), to mention a few. By contrast, the present work not only handles all such cases, it does so with an elementary and unified coupling-based proof technique for any *quadratically-bounded loss*, or more generally fixed point mappings with quadratic growth, such as the TD update, which has no prior high probability analysis (even with projections). This lack of projections is relevant in contemporary usage, since deep learning typically has minimal or nonexistent regularization (Neyshabur et al., 2014; Zhang et al., 2017).

2. **Locally-adapted rates.** Even if there is some natural constraint or regularization in effect within the optimization procedure, the optimal solution may be unsatisfactory: e.g., it may simply be very large, and competing with it could require a large number of samples. On the other hand, the present proofs and rates only rely upon properties of reasonable reference solutions, which may fail to be optimal, but instead have much lower norm.

   As an illustration, consider Figure 1, which runs stochastic GD on a single set of points with either the squared loss $(y, \hat{y}) \mapsto (y - \hat{y})^2/2$ (cf. Figure 1(*a*)), or the logistic loss $(y, \hat{y}) \mapsto \ln(1+\exp(-y\hat{y}))$ (cf. Figure 1(*b*)). Many trajectories of stochastic GD are plotted along with GD run directly on the population risk $\mathcal{R}$ (labeled "population GD"), as well as the curve of optimal constrained solutions $\{\arg\min_{\|w\|\leq B} \mathcal{R}(w) : B \geq 0\}$ (labeled "population OPT"). All trajectories take a long time to rotate towards their asymptotic targets (respectively the minimum norm and max margin solutions); their early behavior is better characterized by rather different low norm but higher risk comparators.

In detail, the concrete contributions and organization of this work are as follows.

1. **Theorem 5: stochastic MD with realizable IID data.** This first guarantee is for stochastic MD (mirror descent with stochastic gradients, as detailed in Section 1.1) on IID *realizable* data; as discussed in Section 2, realizability is encoded as the existence of w_{ref} with *population* risk roughly $\mathcal{O}(1/t)$. Sections 1.1 and 2.1 will discuss this condition, and Section 2 will provide an outline of the general coupling-based proof technique used throughout this work. This realizable setting not only generalizes margin-based analyses for exponentially-tailed losses on linearly-separable data (Ji and Telgarsky, 2018b; Shamir, 2021), it extends them to control the convex risk and not just misclassification risk, to settings with only approximate linear separability, and lastly handles realizable *regression* settings, where $1/t$ high probability rates seem missing in the literature (even with projection).

2. **Theorem 8 and Theorem 9: stochastic MD and stochastic TD on Markovian data.** Dropping the realizability assumption, Section 3 analyzes stochastic MD with not just IID but Markovian data, and uses the same proof technique to handle the standard TD approximate fixed point method used extensively in reinforcement learning (Sutton, 1988). For mirror descent, the closest prior work used projections (Duchi et al., 2012). For TD, there appear to be no prior high probability bounds, even with projections; the closest prior projection-free analysis is in expectation only (Hu et al., 2022), and most prior works make use of not only projections, but also full rank and mixing assumptions (Bhandari et al., 2018). For MD, the squared loss is considered as an illustrative example, where stochastic GD is shown to adapt to local structure in the strong sense of competing with singular value thresholding.

3. **Theorem 10 and Theorem 11: heavy-tailed and batch data.** As a brief auxiliary investigation to demonstrate the proof technique, rates are given in Section 4 for MD on batch and heavy-tailed data. Once again, prior work either requires projections, or violates one of the other goals (e.g., being custom-tailored to exponential-tailed losses (Zhang and Yu, 2005; Telgarsky, 2013; Soudry et al., 2017)).

Rounding out the organization, this introduction concludes with notation and setting in Section 1.1, and the work itself concludes with further related work and open problems in Section 5.

1.1. Notation

Loss functions. In order to handle regression and classification settings simultaneously, each loss $\ell : \mathbb{R} \times \mathbb{R} \to \mathbb{R}_{\geq 0}$ will have an auxiliary scalar function $\widetilde{\ell}$ with exactly one of two forms: either ℓ is a *classification (margin) loss* $\ell(y, \hat{y}) = \widetilde{\ell}(\operatorname{sgn}(y)\hat{y})$, where $\operatorname{sgn}(y) := 2\mathbb{1}[y \geq 0] - 1 \in \{-1, +1\}$, or ℓ is a *regression (distance) loss* $\ell(y, \hat{y}) = \widetilde{\ell}(y - \hat{y})$. Subgradients of ℓ will always be in the second argument, and always exist since $\widetilde{\ell}$ is always convex in this work. The core loss property, *quadratic boundedness*, is defined as follows.

Assumption 1.1 *A loss ℓ is* (C_1, C_2)-quadratically-bounded *(for nonnegative C_1, C_2) if*

$$|\ell'(y, \hat{y})| \leq C_1 + C_2 \left(|y| + |\hat{y}|\right), \qquad \forall y, \hat{y}.$$

This property is quite pessimistic in the sense that stronger variants can be satisfied for all standard losses. Even so, its worst-case nature demonstrates the utility of the core proof technique (which doesn't explode even for this formulation), and captures standard losses via the following lemma. (Throughout this work, $\|\partial f(w)\| := \sup\{\|g\| : g \in \partial f(w)\}$, and ℓ' means $\partial \ell$.)

Lemma 1 *If ℓ is α-Lipschitz (i.e., $\sup_z |\partial\widetilde{\ell}(z)| \leq \alpha$), then ℓ is $(\alpha, 0)$-quadratically-bounded. If ℓ is β-smooth (i.e., $|\widetilde{\ell}'(z) - \widetilde{\ell}'(\hat{z})| \leq \beta|z - \hat{z}|$ $\forall z, \hat{z}$), then ℓ is $(|\partial\ell(0)|, \beta)$-quadratically-bounded.*

A second crucial property is *self-boundedness*, which is used in the realizable rates of Section 2.

Definition 2 *A loss function ℓ is* ρ-self-bounding *if $\widetilde{\ell}$ satisfies $\widetilde{\ell}'(z)^2 \leq 2\rho\widetilde{\ell}(z)$ for all $z \in \mathbb{R}$.*

Notably, the two primary losses in machine learning, the logistic and squared losses, are both self-bounding and quadratically-bounded.

Lemma 3 *The squared loss $\ell(y, \hat{y}) := \frac{1}{2}(y - \hat{y})^2$ is 1-smooth, 1-self-bounding, and $(0, 1)$-quadratically-bounded, whereas the logistic loss $\ell(y, \hat{y}) := \ln(1 + \exp(-y\hat{y}))$ is $(1/4)$-smooth, 1-Lipschitz, $(1/2)$-self-bounding, and $(1, 0)$-quadratically-bounded.*

A few remarks on self-bounding are in order. Firstly, the definition has appeared before (Zhang, 2004), however in a generalized form and with a calculation for the logistic loss which implies it is 0-self-bounding under the present definition; that the logistic loss is 1-self-bounding was first observed in (Telgarsky, 2013), and is crucial for obtaining $1/t$ rates under realizability. Secondly, it may seem that self-bounding is simply a reformulation of smoothness, but firstly it is satisfied for certain nonsmooth losses (in the sense of bounded second derivatives), such as the exponential loss, and secondly replacing self-boundedness with smoothness breaks the current proofs.

Probabilities, expectations, and Markov chains. When data arrives IID, then $\mathrm{E}\mathrm{X}$ and $\mathrm{P}\mathrm{R}$ will respectively denote expectations and probabilities. Correspondingly, the risk $\mathcal{R}(w)$ is defined by $\mathrm{E}\mathrm{X}_{x,y} = \ell(y, w^{\mathsf{T}}x)$. In either case, whenever data $((x_i, y_i))_{i=1}^t$ and a loss ℓ are available, define $\ell_i(v) := \ell(y_i, x_i^{\mathsf{T}}v)$, though $\ell_{x,y}(v) := \ell(y, x^{\mathsf{T}}v)$ is also used, whereby $\mathcal{R}(w) = \mathrm{E}\mathrm{X}_{x,y}\ell_{x,y}(w)$.

With Markov chains and stochastic processes, $\mathrm{E}\mathrm{X}_{\leq i}$ will be used to condition on $\mathcal{F}_{\leq i}$, the σ-algebra of all information up through time i. It will not be necessary for the stationary processes here to be exactly Markovian (or possess a precise stationary distribution); instead, inspired by the *Ergodic Mirror Descent* analysis by Duchi et al. (2012), the stationarity assumption here will be approximate.

Definition 4 *Let $(x_i)_{i\geq 0}$ be samples from a stochastic process, and let P_i^t denote the conditional distribution of x_t conditioned on time $i < t$. For any $\epsilon \geq 0$, a triple (π, τ, ϵ) is an* approximate stationarity witness *if*

$$\sup_{t\in\mathbb{Z}_{\geq 0}} \mathrm{TV}(P_t^{t+\tau}, \pi) \leq \epsilon.$$

(For a similar condition in prior work, see (Duchi et al., 2012, Assumption C).)

Note that if $(x_i)_{i\geq 0}$ are IID, then we can choose $(\pi, \tau, \epsilon) = (P_1^1, 1, 0)$, and the corresponding stochastic MD bounds in Section 3 exhibit no degradation in the IID case. For a broad variety of Markov chains, for any $\epsilon > 0$ we can establish $\tau = \mathcal{O}(\ln(1/\epsilon))$, with hidden constants uniform in ϵ (Meyn and Tweedie, 2012). We will always bake in $\epsilon = 1/\sqrt{t}$, which suggests $\tau = \mathcal{O}(\ln(t))$.

For risk minimization over Markov data as in Theorem 8, the risk will refer to $\mathcal{R}(w) := \mathrm{E}_{x,y\sim\pi}\ell_{x,y}(w)$, where π is an approximate stationary distribution provided by Definition 4.

Mirror descent. Mirror descent is a powerful generalization of gradient descent, which operates as follows. Given a differentiable 1-strongly-convex *mirror map* ψ and the corresponding Bregman divergence $D_\psi(w, v) := \psi(w) - [\psi(v) + \langle\nabla\psi(v), w - v\rangle]$, and a sequence of objective functions $(f_i)_{i\leq t}$, mirror descent chooses a new iterate w_{i+1} from the old iterate w_i and a step size $\eta \geq 0$ and subgradient $g_{i+1} \in \partial f_{i+1}(w_i)$ and a closed convex constraint set S (with $S = \mathbb{R}^d$ allowed) via eq. (3), repeated here verbatim for convenience as

$$w_{i+1} := \arg\min_{w\in S} \left(\langle\eta g_{i+1}, w\rangle + D_\psi(w, w_i)\right).$$

In the present work, typically $f_{i+1} = \ell_{i+1}$, but in Theorem 11 it will be the full batch empirical risk. As before, we will use the notation $\|\partial f_{i+1}(w)\|_* = \sup\{\|g\|_* : g \in \partial f_{i+1}(w)\}$, and the particular choice of subgradient will never matter. The vector space for iterates will be $\mathbb{R}^d$ mainly for sake of presentation, however none of the bounds have any dependence on dimension, and a future version may simply use a separable Hilbert space. Norms without any subscript are simply general norms (i.e., not necessarily Euclidean). Batch and stochastic gradient descent can be written as mirror descent via $\Psi(v) := \|v\|_2^2/2$; for more information on mirror descent, there are many excellent texts (Duchi et al., 2012; Bubeck, 2015; Nemirovski and Yudin, 1983).

TD is only used in Theorem 9, and its presentation is deferred to Section 3.1.

The comparator w_{ref}. The bounds will rely not on global minimizers, but rather on merely good comparators w_{ref}. These comparators will either satisfy $\mathcal{R}(w_{\mathrm{ref}}) \leq D_\psi(w_{\mathrm{ref}}, w_0)/t$ in the realizable case (cf. Theorem 5), or $\mathcal{R}(w_{\mathrm{ref}}) \leq D_\psi(w_{\mathrm{ref}}, w_0)/\sqrt{t} + \inf_v \mathcal{R}(v)$ in the general case (cf. Theorem 8, Theorem 10, Theorem 11). Here are a few sanity checks on this arguably awkward definition (which has w_{ref} on both sides). Firstly, if a minimizer w_{ref} exists (meaning $\mathcal{R}(w_{\mathrm{ref}}) = \inf_v \mathcal{R}(v)$), then immediately $\mathcal{R}(w_{\mathrm{ref}}) \leq D_\psi(w_{\mathrm{ref}}, w_0)/t^\alpha + \inf_v \mathcal{R}(v)$ for all $\alpha \geq 0$ and all t. Secondly, if there exists a w_{ref} satisfying $\mathcal{R}(w_{\mathrm{ref}}) \leq D_\psi(w_{\mathrm{ref}}, w_0)/\sqrt{t} + \inf_v \mathcal{R}(v)$, then we can also use the *oracle solution* $\hat{w} := \arg\min_u \left[D_\psi(u, w_0)/\sqrt{t} + \mathcal{R}(u)\right]$, since the nonnegativity of D_ψ and existence of w_{ref} imply

$$\mathcal{R}(\hat{w}) \leq \mathcal{R}(\hat{w}) + \frac{D_\psi(\hat{w}, w_0)}{\sqrt{t}} = \inf_u \left[\mathcal{R}(u) + \frac{D_\psi(u, w_0)}{\sqrt{t}}\right] \leq \frac{D_\psi(w_{\mathrm{ref}}, w_0)}{\sqrt{t/4}} + \inf_v \mathcal{R}(v),$$

meaning we can use $\hat{w}$ at an earlier time $t/4$. In general, the admissible choices for w_{ref} depend on ℓ, the data distribution, and on t (i.e., different w_{ref} are relevant at different times, as desired); detailed discussions are given for the logistic and squared losses in Section 2 and Section 3.

2. Realizable case, illustrative examples, and proof scheme

The first bound is for self- and quadratically-bounded losses, and requires *realizability*: as discussed at the end of Section 1.1, this corresponds to the existence of w_{ref} with $\mathcal{R}(w_{\text{ref}}) \leq D_\psi(w_{\text{ref}}, w_0)/t$, which will be discussed momentarily for the logistic loss.

Theorem 5 *Suppose ℓ is convex, (C_1, C_2)-quadratically-bounded, and ρ-self-bounding. Let t be given, and suppose $((x_i, y_i))_{i \leq t}$ are IID samples with $\max\{\|x_i\|_*, |y_i|\} \leq 1$ almost surely. Let reference solution w_{ref} and initial point w_0 be given, and suppose w_{ref} satisfies $\mathcal{R}(w_{\text{ref}}) \leq \rho D_\psi(w_{\text{ref}}, w_0)/t$, and let C_4 be given so that $\max_{j<t} |\ell_{j+1}(w_{\text{ref}})| \leq C_4$ almost surely. Then with probability at least $1 - 2t\delta$, every $i \leq t$ satisfies*

$$\frac{8}{3i\eta} D_\psi(w_{\text{ref}}, w_i) + \frac{1}{i}\sum_{j<i} \mathcal{R}(w_j) \leq \frac{2B_w^2\left(1 + C_1 + C_2(1 + \|w_{\text{ref}}\|) + C_4\right)}{i\eta} + \frac{4}{\eta}\mathcal{R}(w_{\text{ref}}),$$

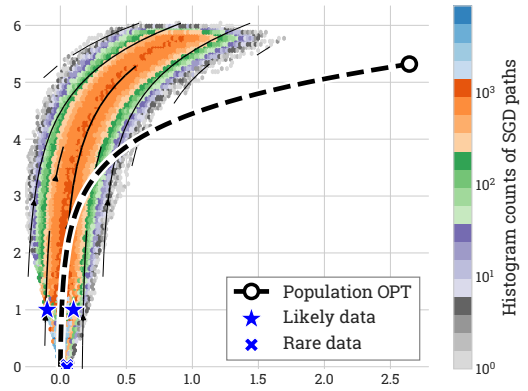
where $B_w := \max\left\{1, 4\sqrt{D_\psi(w_{\text{ref}}, w_0)}, \sqrt{(64C_4/\rho)\ln(1/\delta)}\right\}$ and $\eta \leq 1/(2\rho)$.

All bounds in this work will have roughly the form of Theorem 5, which can be summarized as follows. As stated in the introduction, the bound has a regret-style average risk on the left hand side, and the risk of the comparator w_{ref} in the right hand side. Unusual elements are the control for all times $i < t$, the left hand side term $D_\psi(w_{\text{ref}}, w_i)$, and the coefficient exceeding one on $\mathcal{R}(w_{\text{ref}})$. The control for all times $i < t$ is an artifact of the proof scheme, and will be discussed below. The left hand side term $D_\psi(w_{\text{ref}}, w_i)$ is crucial to the operation of the proof; it is an *implicit bias* which prevents iterates from growing too large. This term is dropped in all standard presentations of mirror descent (Bubeck, 2015; Duchi et al., 2012; Nemirovski and Yudin, 1983), but was exploited in the original perceptron convergence proof (Novikoff, 1962). The large coefficient on $\mathcal{R}(w_{\text{ref}})$ is a consequence of realizability, and can not be removed with the current proof scheme.

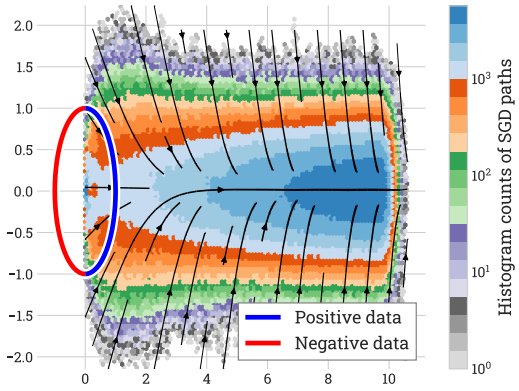
2.1. Illustrative example: the logistic loss and approximate separability

To investigate Theorem 5 more closely, consider the logistic loss $\ell(y, \hat{y}) = \ln(1 + \exp(-y\hat{y}))$ on classification data. In typical implicit regularization works, the logistic loss is treated as having an exponential tail, and inducing convergence to maximum margin directions (Zhang and Yu, 2005; Telgarsky, 2013; Soudry et al., 2017; Shamir, 2021). As in Figure 1(*b*), it can take a while for the maximum margin asymptotics to kick in; for example, the risk rate for SGD in (Ji and Telgarsky, 2018b, Theorem 1.1) is $\mathcal{O}(\ln(t)/(t\gamma^2))$, where the population margin γ can be taken to mean $\Pr[\bar{u}^\intercal xy \geq \gamma] = 1$ for some unit vector $\bar{u}$. Returning to Figure 1(*b*), the global margin γ is very small, but the initial dynamics are governed by the much larger margin of the likely points oriented vertically. As a first step towards formalizing this and connecting back to Theorem 5, consider the following *approximate realizability/separability* characterization, and its consequences on the choice of w_{ref}.

Proposition 6 *Let t be given. Suppose $\ell(y, \hat{y}) = \ln(1 + \exp(-\operatorname{sgn}(y)\hat{y}))$ is the logistic loss, the data satisfies $\|x\|_2 \leq 1$ and $y \in \{\pm 1\}$ almost surely, and that there exists a unit vector $u_t \in \mathbb{R}^d$ and a scalar $\gamma_t > 0$ so that $\Pr[u_t^\intercal xy \geq \gamma_t] \geq 1 - 1/t$. Then the reference solution $w_{\text{ref}} := u_t \ln(t)/\gamma_t$ satisfies $\mathcal{R}(w_{\text{ref}}) \leq (2 + \ln(t)/\gamma_t)/t$.*



(a) Another visualization of Figure 1(b).



(b) Spherical data from Proposition 7.

Figure 2: These plots show SGD on the logistic loss, but rather than depicting many trajectories as a filigree (as in Figure 1), they are histogrammed into hexagonal bins, with additional black lines showing the vector field. The color scheme gradations are logarithmic, and seem to exhibit an exponential concentration around the GD path, a phenomenon not established in this work. Figure 2(a) shows this visualization technique on the same data from Figure 1, whereas Figure 2(b) shows the spherical 0-margin data from Proposition 7.

Applying Theorem 5 with this w_{ref} and with step size $\eta = 1$ gives, with probability at least $1-\delta$,

$$\frac{8}{3t}\|w_t - w_{\text{ref}}\|^2 + \frac{1}{t}\sum_{i<t}\mathcal{R}(w_i) \leq \mathcal{O}\left(\frac{\ln(t)^2\sqrt{\ln(t/\delta)}}{t\gamma_t^2}\right).$$

As a sanity check, if the global maximum margin $\gamma \leq \gamma_t$ is used, then this bound matches the bound from prior work mentioned before (Ji and Telgarsky, 2018b, Theorem 1.1), and also matches standard margin-based generalization bounds (Schapire et al., 1997). A key difference is that γ_t is not a hard margin, but allows some fraction of margin violations. In particular, returning to Figure 1(b), if we choose $t = 10$, then we can choose γ_t to be the large "likely data" margin, and our convergence rate scales with this $1/\gamma_t^2$, rather than the much smaller "rare data" margin.

Rather than relying on an opaque γ_t, here is another example where γ_t may be calculated. Consider Figure 2(b), where there is a perfect but 0-margin classifier (thus breaking standard bounds), and the marginal distribution of x along this perfect classifier is uniform. In this setting, the optimal predictor of length r is unique and achieves risk only $1/r$, in contrast to the standard margin setting (roughly as in Proposition 6, where one hopes for a predictor of length $\ln(r)/\gamma$ for risk $1/r$).

Proposition 7 *Let dimension $d \geq 2$ be given, let μ_0 denote the uniform probability density on the sphere $\mathcal{S}_{d-1} := \{x \in \mathbb{R}^d : \|x\| = 1\}$, and let μ denote a reweighting of μ_0 along the axis e_1 so that every orthogonal slice has equal density, meaning $\mathrm{d}\mu(x) = p(x_1)\,\mathrm{d}\mu_0(x)$ for some p, whereby $\Pr_\mu[\{x \in \mathcal{S}_{d-1} : -1 \leq a \leq x_1 \leq b \leq 1\}] = (b-a)/2$, and suppose $\Pr[y = 1|x] = \mathbb{1}[x_1 \geq 0]$. Then for any norm $r > 0$, the vector $u_r := re_1$ is the unique minimizer of $\mathcal{R}$ with norm r, and*

moreover if $r \geq 1$ *then*

$$\left| \mathcal{R}(u_r) - \frac{\pi^2}{12r} \right| \leq \frac{2}{\exp(r)}.$$

Now consider applying the (approximately) realizable analysis in Theorem 5 to this setting. Choosing $\eta = 1$ as suggested there, and $w_{\text{ref}} = e_1/t^{1/3}$ as suggested by Proposition 7 after optimizing terms, then $\max\{\|w_{\text{ref}}\|, \mathcal{R}(w_{\text{ref}})\} = \mathcal{O}(1/t^{1/3})$ and $\mathcal{R}(w_{\text{ref}}) = \mathcal{O}(\|w_{\text{ref}}\|^2/t)$ as required by the realizability conditions in Theorem 5. Thus, with probability at least $1 - \delta$,

$$\frac{8}{3t}\|w_{\text{ref}} - w_t\|^2 + \frac{1}{t}\sum_{i<t} \mathcal{R}(w_i) \leq \mathcal{O}\left(\frac{\ln(t/\delta)}{t^{1/3}}\right).$$

There does not appear to be any prior work analyzing these infinitessimally-separable scenarios; furthermore, such examples necessitated the realizability formulation in Theorem 5.

2.2. Proof sketch: the core coupling-based argument

This subsection provides the basic form of the coupling-based argument used within all proofs in this work. Rather than handling the realizable setting of Theorem 5 directly, it is stated for the simpler setting of stochastic gradient descent with IID data and a step size $\eta = \mathcal{O}(1/\sqrt{t})$, with a few remarks afterwards then adjusting it for Theorem 5.

The proof scheme consists of the following three steps.

1. **Coupling unconstrained iterates $(w_i)_{i<t}$ with constrained iterates $(v_i)_{i<t}$.** Because $(w_i)_{i<t}$ are unconstrained, it is unclear how to apply standard concentration inequalities to them. Instead, define *projected* iterates $(v_i)_{i<t}$ which are coupled to $(w_i)_{i<t}$ in the following strong sense: $v_0 = w_0$, and thereafter, v_{i+1} is defined using the *same* randomness as w_{i+1}, meaning

$$\begin{aligned} w_{i+1} &:= w_i - \eta \partial_w \ell(y_{i+1}, x_{i+1}^{\mathsf{T}} w_i), \\ v_{i+1} &:= \Pi_S \left(v_i - \eta \partial_v \ell(y_{i+1}, x_{i+1}^{\mathsf{T}} v_i) \right), \end{aligned}$$

   where the constraint set $S := \left\{ v \in \mathbb{R}^d : \|v - w_{\text{ref}}\| \leq B_w \right\}$ has a few key choices. Firstly, it projects onto a ball around the desired comparator w_{ref}; algorithmically, this would require clairvoyantly re-running the algorithm with knowledge of w_{ref}, but here it is only used as a mathematical construct. Since $(v_i)_{i\leq t}$ explicitly depends on w_{ref}, relating w_i to v_i will in turn relate w_i to w_{ref}. A description of the radius B_w will come shortly.

2. **Implicitly-biased MD analysis of $(v_i)_{i\leq t}$.** Because $(v_i)_{i\leq t}$ are constrained to a small ball around w_{ref}, we can easily apply MD and concentration guarantees and expect all quantities to scale with properties of w_{ref}. Concretely, following the standard MD proof scheme specialized to GD via $\Psi(w) = \frac{1}{2}\|w\|_2^2$ (whereby $D_\psi(w_{\text{ref}}, w) = \frac{1}{2}\|w_{\text{ref}} - w\|_2^2$), and writing $h_{j+1} := \partial_v \ell(y_{j+1}, x_{j+1}^{\mathsf{T}} v_j)$ for the stochastic gradient at time $j+1$ for v_j,

$$\|v_{j+1} - w_{\text{ref}}\|_2^2 \leq \|v_j - w_{\text{ref}}\|_2^2 + 2\eta \left[\ell_{j+1}(w_{\text{ref}}) - \ell_j(v_j)\right] + \eta^2 \|h_{j+1}\|^2,$$

   which after recursing and rearranging (alternatively applying $\sum_{j<i}$ to both sides) gives

$$\|v_i - w_{\text{ref}}\|_2^2 \leq \|v_0 - w_{\text{ref}}\|_2^2 + 2\eta \sum_{j<i} \left[\ell_{j+1}(w_{\text{ref}}) - \ell_{j+1}(v_j) + \eta^2 \|h_{j+1}\|^2\right].$$

Applying Azuma's inequality, with probability at least $1-\delta$,

$$\|v_i - w_{\text{ref}}\|_2^2 \le \|v_0 - w_{\text{ref}}\|_2^2 + 2\eta \sum_{j<i} [\mathcal{R}(w_{\text{ref}}) - \mathcal{R}(v_j)] + \eta\Big[\text{deviations} + \eta \sum_{j<i} \|h_{j+1}\|^2\Big]. \tag{4}$$

What is the magnitude of the error term on the second line? Azuma's inequality scales with the range of the relevant random variables, and thus if the loss has quadratic growth, we can expect the entire second line to be $\mathcal{O}(B_w^2)$, which deserves quite a bit more discussion.

This quantity $\mathcal{O}(B_w^2)$ (and the choice of quadratically-bounded losses) is crucial. The left hand side of the bound has $\|v_i - w_{\text{ref}}\|_2^2$, which is at most B_w^2 by the choice of S. As such, if the $\mathcal{O}(B_w^2)$ in the right hand side has a leading constant less than 1, then *the projection operation is never invoked*, and we should be able to show $w_i = v_i$. In fact, this observation was the starting point of this work, and "quadratically-bounded loss" is merely a reverse-engineered concept to make it go through. Moreover, it is clear that none of this would be possible if the left hand term $\|v_i - w_{\text{ref}}\|_2^2$ were deleted, as is standard in MD.

3. **Proving $(w_i)_{i\le t} = (v_i)_{i\le t}$ via induction.** We are now in position to complete the proof.

   Let E denote the failure event for the earlier regret guarantee in eq. (4), which rules out certain wild trajectories for $(v_i)_{i\le t}$. The underlying sample space for this event is $((x_i, y_i))_{i\le t}$, and therefore this event also controls the behavior of $(w_i)_{i\le t}$; in fact, this proof will show that ruling out E deletes not just the wild trajectories of $(v_i)_{i\le t}$, but also that one may interpret these projected iterates as mere proxies to get a handle on the wild trajectories of $(w_i)_{i\le t}$. This proof technique is then a truncation argument, as is standard throughout probability theory.

   In detail, consider w_{i+1}, and suppose the inductive hypothesis $(w_j)_{j\le i} = (v_j)_{j\le i}$. Writing out a *deterministic* gradient descent inequality for w_{i+1} (cf. Lemma 17) and then invoking the inductive hypothesis to transplant $(v_j)_{j\le i}$, gives (under event E)

$$\begin{aligned}
\|w_{i+1} - w_{\text{ref}}\|_2^2 &= \|w_0 - w_{\text{ref}}\|_2^2 + 2\eta \sum_{j\le i} [\ell_{j+1}(w_{\text{ref}}) - \ell_{j+1}(w_j)] + \eta^2 \sum_{j\le i} \|\partial \ell_{j+1}(w_j)\|^2 \\
&= \|v_0 - w_{\text{ref}}\|_2^2 + 2\eta \sum_{j\le i} [\ell_{j+1}(w_{\text{ref}}) - \ell_{j+1}(v_j)] + \eta^2 \sum_{j\le i} \|\partial \ell_{j+1}(v_j)\|^2 \\
&\le \|v_0 - w_{\text{ref}}\|_2^2 + 2\eta \sum_{j\le i} [\mathcal{R}(w_{\text{ref}}) - \mathcal{R}(v_j)] + i\eta^2 \mathcal{O}(B_w^2).
\end{aligned}$$

   With some tuning of η and B_w, the final term $i\eta^2\mathcal{O}(B_w^2)$ is in fact strictly less than B_w^2, which suffices to imply projections are never invoked, and $w_{i+1} = v_{i+1}$. This argument is repeated for every iteration $i \le t$, so in fact there are t different failure events $(E_i)_{i\le t}$, and unioning them together gives the final statement.

The preceding proof was for GD not MD, but the standard MD proof scheme is identical Lemma 17, even with the left hand implicit bias term added in.

Handling the realizable case has a few important differences. The first is that the squared gradient term $\|\partial \ell_{j+1}(v_j)\|_*^2$ is swallowed into the loss term via the definition of ρ-self-bounding. Moreover, to obtain a rate $1/t$ not $1/\sqrt{t}$, Freedman's inequality is used rather than Azuma's inequality,

which needs the conditional variances to be small (which invokes realizability). Lastly, to allow for a simple step size, two separate concentration inequalities are applied: one to control norms, and another to control risks; using just one concentration inequality would give similar rates but require a messy step size as in Theorem 8.

3. General MD analysis, illustrative examples, and TD analysis

The section exhibits slower rates, meaning $1/\sqrt{t}$ rather than the $1/t$ in Theorem 5, but allows an important generalization: the data need not be IID, but instead is approximately Markovian (cf. Definition 4), and need not be realizable.

The first bound here is for stochastic MD. As discussed in Section 1.1, the condition on w_{ref} is now $\mathcal{R}(w_{\text{ref}}) \leq D_\psi(w_{\text{ref}}, w_0)/\sqrt{t} + \inf_v \mathcal{R}(v)$.

Theorem 8 *Suppose ℓ is convex and (C_1, C_2)-quadratically-bounded. Let t be given, and suppose $((x_i, y_i))_{i\leq t}$ are drawn from a stochastic process with approximate stationarity witness $(\pi, \tau, 1/\sqrt{t})$ with $\max\{\|x_i\|_*, |y_i|\} \leq 1$ almost surely. Let reference solution w_{ref} and initial point w_0 be given, and suppose w_{ref} satisfies $\mathcal{R}(w_{\text{ref}}) \leq D_\psi(w_{\text{ref}}, w_0)/\sqrt{t} + \inf_v \mathcal{R}(v)$. Then with probability at least $1 - t\tau\delta$, every $i \leq t$ satisfies*

$$\frac{1}{i\eta} D_\psi(w_{\text{ref}}, w_i) + \frac{1}{i}\sum_{j<i} \mathcal{R}(w_j) \leq \frac{B_w^2}{8i\eta} + \mathcal{R}(w_{\text{ref}}),$$

where $B_w = \max\left\{1, C_2\|w_{\text{ref}}\|, 4\sqrt{D_\psi(w_{\text{ref}}, w_0)}\right\}$ and $\eta \leq \frac{1}{4096\max\{1,C_1,C_2\}\sqrt{t\tau\ln(1/\delta)}}$.

Illustrative examples of Theorem 8 will be provided shortly in Section 3.2. The form of the bound is similar to Theorem 5, but has a rate $\mathcal{O}(1/\sqrt{t})$ after expanding η. Unlike Theorem 5, the step size is messy; this seems necessary with the current proof technique, which seems to have no recourse but to use η to swallow some terms; the realizable analysis was able to avoid this thanks to applying two concentration inequalities, the first of which relied heavily on realizability.

The proof of Theorem 8 follows the sketch in Section 2.2 exactly, with two exceptions. The first is that GD is replaced by MD, following a nearly standard analysis with an added non-standard implicit bias term $D_\psi(w_{\text{ref}}, w_i)$ (cf. Lemma 17). The second difference is that Azuma's inequality is replaced with a Markov chain concentration inequality, which itself uses a standard technique of treating the data as τ interleaved sequences of nearly-IID data, and applying Azuma's inequality to each. This concentration inequality is detailed in Lemma 14, but is abstracted from a proof due to Duchi et al. (2012).

3.1. TD analysis

The second Markovian guarantee on TD. It is not necessary to be familiar with any RL concepts to make sense of this theorem, and in fact it can be stated as a fixed point property, but here is some brief background. The sequence $(x_i)_{i\geq 0}$ with $x_i \in \mathbb{R}^d$ is interpreted as combined state/action vectors, and instead of labels there are scalar rewards $(r_i)_{i\geq 1}$, whose conditional distribution is fully determined by the preceding state/action vector, meaning $r_{i+1}|\mathcal{F}_{\leq i} = r_{i+1}|x_i$. The stochastic TD update is

$$w_{i+1} := w_i - \eta G_{i+1}(w_i), \quad \text{where } G_{i+1}(v) = x\left(\left\langle x_i - \gamma x'_{i+1}, v\right\rangle - r_{i+1}\right), \tag{5}$$

where the *discount factor* $\gamma \in (0,1)$ is fixed throughout.

In prior work, this method is only studied in expectation, often with a variety of boundedness/projection and full rank conditions (Zou et al., 2019), or further stationarity and sampling conditions (Bhandari and Russo, 2019). Some recent work has aimed to reduce these assumptions, but still was only able to achieve bounds in expectation (Hu et al., 2022). Meanwhile, invoking essentially the same proof as for Theorem 8 leads to a high probability guarantee; the only real difference is that the deterministic MD analysis (from Lemma 17) is replaced with a similar deterministic TD analysis (from Lemma 18), even though TD is not in any sense a gradient method.

Theorem 9 *Let a stochastic process $((x_i, r_i))_{i\geq 0}$ be given, where $(x_i)_i \geq 0$ form a Markov chain and $\max\{\|x_i\|, |r_i|\} \leq 1$ almost surely, and define auxiliary random variables $\zeta_{i+1} = (x_i, x_{i+1}, r_{i+1})$, and let $(\pi, \tau, 1/\sqrt{t})$ denote an approximate stationarity witness for $(\zeta_i)_{i\geq 1}$. Let reference solution w_{ref} be given with $\|\mathrm{E}_{\zeta\sim\pi} G_\zeta(w_{\mathrm{ref}})\| \leq \|w_{\mathrm{ref}} - w_0\|^2/\sqrt{t}$, where $G_\zeta(w_{\mathrm{ref}}) := x\left(\langle x - \gamma x', w_{\mathrm{ref}}\rangle - r\right)$ for $\zeta = (x, x', r)$. Then, with probability at least $1 - t\tau\delta$,*

$$\|w_t - w_{\mathrm{ref}}\|^2 + \eta(1-\gamma)^2 \sum_{i<t} \mathrm{E}_{x\sim\pi} \langle x, w_i - w_{\mathrm{ref}}\rangle^2 \leq B_w^2 + \frac{t\eta B_w}{512} \|\mathrm{E}_{\zeta\sim\pi} G_\zeta(w_{\mathrm{ref}})\|,$$

where $B_w = \max\{1, 4\|w_{\mathrm{ref}}\|, 4\|w_0 - w_{\mathrm{ref}}\|\}$ and $\eta \leq \frac{1}{1024\sqrt{t\tau\ln(1/\delta)}}$.

Notably, as a parallel to the approximate optimality of w_{ref} in Theorems 5 and 8, the reference solution in Theorem 9 need only be an *approximate* fixed point: $\|\mathrm{E}_{\zeta\sim\pi} G_\zeta(w_{\mathrm{ref}})\| = \mathcal{O}(1/\sqrt{t})$.

3.2. Illustrative examples

Squared loss. First consider the squared loss $\ell(y, \hat{y}) := (y - \hat{y})^2/2$. Typically $\mathcal{R}$ in this setting is treated as strongly convex (perhaps along a subspace), and SGD converges at a rate $1/t$. Unfortunately, this rate also scales with $1/\sigma_{\min}^2$, the inverse of the smallest positive eigenvalue of the population covariance, which has no reason to be large or stable.

The goal of the present work is to eschew global dependencies, and depend on local properties; this is also exhibited in Figure 1(*a*), where both stochastic GD as well as GD on $\mathcal{R}$ spend a long time pointing *away* from the minimum norm solution of $\mathcal{R}$ they eventually converge to.

As a concrete construction of w_{ref}, consider the case of *singular value thresholding*: rather than seeking out the population solution $\bar{w} := [\mathrm{E}xx^\intercal]^+ [\mathrm{E}xy]$, where the "+" denotes a pseudoinverse, consider $\bar{w}_k := [\mathrm{E}xx^\intercal]_k^+ [\mathrm{E}xy]$, where k truncates the spectrum of $\mathrm{E}xx^\intercal$ to have only the k largest eigenvalues. Correspondingly, suppose $w_0 = 0$, and choose t small enough so that

$$\mathcal{R}(\bar{w}_k) \leq \frac{\|\bar{w}_k\|^2}{2\sqrt{t}} + \inf_v \mathcal{R}(v) = \frac{D_\psi(\bar{w}_k, w_0)}{\sqrt{t}} + \mathcal{R}(\bar{w});$$

whenever this holds, this $\bar{w}_k$ and t can be plugged in to Theorem 8, giving a stochastic GD guarantee which not only competes with $\mathcal{R}(\bar{w}_k)$, but moreover the constants in the rate scale with $\|\bar{w}_k\|^2$, which is on the order $1/\sigma_k^2$, rather than being on the order $1/\sigma_{\min}^2$ as with the minimum norm least squares solution $\bar{w}$. This gives some explanation of the behavior of Figure 1(*a*): early in training, the path is closer to low norm solutions such as the singular value thresholded solution $\bar{w}_k$. Said another way, stochastic GD competes with $\bar{w}_k$ *for every* k without any specialized algorithm!

Geometric medians. As another example, suppose the regression problem $\ell(y, \hat{y}) := \|y - \hat{y}\|$, which is Lipschitz, nonsmooth, and unbounded. Applying stochastic GD to this problem leads to a pleasing algorithm especially in the univariate case: explicitly, to find the median of a stream of data points, it simply compares its estimate w_i to the new test point y_{i+1}, and adjust by $\pm\eta$ depending on whether it is to the right or left. Standard analyses of this method would require projections or regularization (and some outer doubling loop to guess the radius), but Theorem 8 can handle a direct the projection-free stochastic GD.

4. Final examples: batch data and heavy-tailed data

This final set of results will relax two conditions which may have seemed necessary to the proof scheme in Section 2.2: data may be heavy-tailed, and data may be handled as a batch.

4.1. Heavy-tailed data

All preceding sections required bounded data: $\max\{\|x\|_*, |y|\} \leq 1$ almost surely. Instead, the following bound is similar to the non-realizable setting of Theorem 8, except the data is IID, and may have two types of heavy tail.

Theorem 10 *Suppose ℓ is convex and (C_1, C_2)-quadratically-bounded. Let t be given, and suppose $((x_i, y_i))_{i\leq t}$ are drawn IID with auxiliary random variables $Z_i := \max\{1, \|x\|_*^4, |y|^4\}$ satisfying one of the following two tail behaviors with a corresponding constant C.*

1. ***Subgaussian tails.*** *Each Z_i is subgaussian with variance proxy σ^2, and define $C := \mathrm{E}\mathrm{x} Z_1 + 2\sigma\sqrt{\ln(1/\delta)/t}$.*
2. ***Polynomial tails.*** *Defining a moment bound $M := \max\{p/e,\ \sup_{2\leq r\leq p} \mathrm{E}\mathrm{x}|Z_i - \mathrm{E}\mathrm{x} Z_i|^r\}$ for each Z_i for some power p satisfying $8|p$, define $C := \mathrm{E}\mathrm{x} Z_1 + 2M\left(\frac{2}{\delta}\right)^{1/p}/\sqrt{t}$.*

Let reference solution w_{ref} and initial point w_0 be given, and suppose w_{ref} satisfies $\mathcal{R}(w_{\text{ref}}) \leq D_\psi(w_{\text{ref}}, w_0)/\sqrt{t} + \inf_v \mathcal{R}(v)$. Then with probability at least $1 - 2t\delta$, every $i \leq t$ satisfies

$$\frac{1}{i\eta} D_\psi(w_{\text{ref}}, w_i) + \frac{1}{i}\sum_{j<i} \mathcal{R}(w_j) \leq \frac{B_w^2}{8i\eta} + \mathcal{R}(w_{\text{ref}}),$$

where $B_w = \max\left\{1, C_2\|w_{\text{ref}}\|, 4\sqrt{D_\psi(w_{\text{ref}}, w_0)}\right\}$ and $\eta \leq \frac{1}{4096\max\{1,C_1,C_2\}\sqrt{t(1+C)\ln(1/\delta)}}$.

A few remarks are in order. Firstly, with polynomial tails, the bound is not what is generally called a high probability bound, as the familiar $\ln(1/\delta)$ is replaced with $(1/\delta)^{1/p}$. This has a particularly bad interaction with union bounds: in fact, since the proof technique utilized a union bound over all t iterations, this term should in fact be interpreted as $(t/\delta)^{1/p}$. Expanding the corresponding term C in the final bound, the rate becomes $\max\{t^{-1/2}, t^{-1+1/(2p)}\}$, which is reasonable, though the dependence on $1/\delta$ is still unpleasant.

A second remark is on the proof. The mirror descent guarantee in the case of general data with smooth losses ends up having terms of the form $\sum_{i<t}\max\{\|x\|_*^4, |y|^4\}$ appear in a few places, which were simply upper bounded by t in the earlier general analyses. The step size η is then made large to swallow these terms (i.e., the C above appears within η), but there are still two issues:

firstly, C must be controlled, and secondly, as this quantity is random, we can not simply invoke Azuma's inequality with a random range. To solve the first problem, there exist a variety of heavy tail concentration inequalities, as detailed in the proofs in the appendix. For the second problem, we use a very nice variant of Azuma's inequality which allows the ranges to not be specified up front (van Handel, 2016, Problem 3.11).

It appears guarantees of this type have not appeared before; the most similar analyses consider specialized scenarios and moreover modify the descent method, for instance by using minibatches with specially-tuned batch sizes to exhibit strong convexity structure assumed to hold over the population (Zhu et al., 2022), or by gradient clipping and similar procedures (Gorbunov et al., 2020; Davis and Drusvyatskiy, 2020; Nazin et al., 2019).

4.2. Batch data

All the bounds in this work so far have worked with stochastic data, arriving one point at a time; correspondingly, the concentration inequalities seemed to rely upon martingale structure. The final bound shows that this is not necessary: data can be handled in a batch, and the concentration inequality can simply be a generalization bound. This method replaces the stochastic gradient g_{j+1} with a full batch gradient:

$$g_{i+1} := \partial_w \widehat{\mathcal{R}}(w_i), \qquad \text{where } \widehat{\mathcal{R}}(w_i) := \frac{1}{n}\sum_{k=1}^{n} \ell(y_k, x_k^{\mathsf{T}} w_i). \tag{6}$$

Theorem 11 *Suppose ℓ is convex and (C_1, C_2)-quadratically-bounded. Suppose $((x_i, y_i))_{i\le n}$ are drawn IID with $\max\{\|x_i\|_*, |y_i|\}$ almost surely, and that $(w_i)_{i\le t}$ are given by batch MD with $t \le n$, where batch gradients are given eq. (6). Let reference solution w_{ref} and initial point w_0 be given, and suppose w_{ref} satisfies $\mathcal{R}(w_{\text{ref}}) \le D_\psi(w_{\text{ref}}, w_0)/\sqrt{t} + \inf_v \mathcal{R}(v)$. Then with probability at least $1 - 4\delta$, every $i \le t$ satisfies*

$$\frac{1}{i\eta} D_\psi(w_{\text{ref}}, w_i) + \frac{1}{i}\sum_{j<i} \mathcal{R}(w_j) \le \frac{B_w^2}{8i\eta} + \mathcal{R}(w_{\text{ref}}),$$

where $B_w = \max\left\{1, C_2\|w_{\text{ref}}\|, 4\sqrt{D_\psi(w_{\text{ref}}, w_0)}\right\}$ and $\eta \le \frac{1}{4096\max\{1, C_1, C_2\}\sqrt{t\ln(1/\delta)}}$.

The proof is basically the same as Theorem 8, except the Markov chain concentration inequality is replaced with a generalization bound for linear predictors. As a technical note, the generalization bound is directly on a ball defined by D_ψ, rather than the corresponding norm, as happens in the other proofs; this is thanks to a clean Rademacher bound on Bregman balls due to Kakade et al. (2008, Theorem 3 and Example (4)).

Although the main focus of this work is on methods which process one example at a time, general unconstrained batch guarantees as above similarly do not seem to have appeared in the literature; as with the stochastic analyses, the closest prior work has rates depending on structural properties of the loss and training data (Soudry et al., 2017; Ji and Telgarsky, 2018b).

5. Further related work and open problems

Implicit regularization. The extensive literature on implicit bias/regularization was a major source of techniques and inspiration for the present work. These works typically show convergence to a

specific (limiting) good solution over the training set; this has been shown for coordinate descent (Zhang and Yu, 2005; Telgarsky, 2013). gradient descent (Soudry et al., 2017; Ji and Telgarsky, 2018b), deep linear networks (Ji and Telgarsky, 2018a; Arora et al., 2019), ReLU networks (Lyu and Li, 2019; Chizat and Bach, 2020; Ji and Telgarsky, 2020), mirror descent (Gunasekar et al., 2018), and many others. One point of contrast is that the focus in the preceding works is on structure of the training set, not the structure of the population distribution as is used here. Moreover, the relaxed criterion here (where there is no effort to prove $w_t \to w_{\text{ref}}$ in any topology) allows treatment of previously difficult cases, such as the spherical zero-margin data in Section 2.1.

Is there a more refined comparison between the approaches? Is there a stronger convergence property over the distribution than the ones here? Figure 1 suggests that SGD concentrates along the path of population GD; is there an easy way to prove this, perhaps via implicit regularization techniques or the techniques in the present work?

That said, there is an increasing body of work which takes the view of mirror descent here, namely of not dropping the term $D_\psi(w_{\text{ref}}, w_t)$ and treating it as crucial (Ji and Telgarsky, 2018b; Shamir, 2021; Vaskevicius et al., 2020). There is also work on the unbounded *online* setting (Cutkosky and Orabona, 2018), but the guarantees there contain gradient norms and other terms which in the present statistical work can be ensured small.

Deep networks. Deep learning is a natural target for the techniques presented here. Unfortunately, the proofs heavily rely upon convexity. Is there some adjustment that can be made to hold for general deep networks, meaning those far outside the initial linearization regime?

Concentration-based proof technique. Is there a way to prove these results without needing a coupled sequence $(v_i)_{i \le t}$? For instance, is there a powerful concentration inequality which can directly establish concentration along the population GD path?

SGD vs GD. Many recent works aim to exhibit and study cases where SGD behaves *differently* from GD, with an eye towards giving further justification to the extensive use of SGD in practice (Wu et al., 2020); in this sense, the present work is a bit unambitious. Is there some way to use the coupling-based proof technique here — perhaps by coupling with a very different path — to establish other behaviors of SGD?

Acknowledgments

The author thanks Daniel Hsu, Ziwei Ji, Francesco Orabona, Jeroen Rombouts, and Danny Son for valuable discussions, as well as the COLT 2022 reviewers for many helpful comments, specifically regarding readability. The author thanks the NSF for support under grant IIS-1750051.

References

Alekh Agarwal, Daniel Hsu, Satyen Kale, John Langford, Lihong Li, and Robert Schapire. Taming the monster: A fast and simple algorithm for contextual bandits. In *International Conference on Machine Learning*, pages 1638–1646. PMLR, 2014.

Sanjeev Arora, Nadav Cohen, Wei Hu, and Yuping Luo. Implicit regularization in deep matrix factorization. *Advances in Neural Information Processing Systems*, 32, 2019.

Jalaj Bhandari and Daniel Russo. Global optimality guarantees for policy gradient methods. *arXiv preprint arXiv:1906.01786*, 2019.

Jalaj Bhandari, Daniel Russo, and Raghav Singal. A finite time analysis of temporal difference learning with linear function approximation. *arXiv preprint arXiv:1806.02450*, 2018.

Avrim Blum, John Hopcroft, and Ravindran Kannan. Foundations of data science, 2017. URL https://www.cs.cornell.edu/jeh/book.pdf.

Léon Bottou. Large-scale machine learning with stochastic gradient descent. In *Proceedings of COMPSTAT'2010*, pages 177–186. Springer, 2010.

Sébastien Bubeck. Convex optimization: Algorithms and complexity. *Foundations and Trends in Machine Learning*, 2015.

Lenaic Chizat and Francis Bach. Implicit bias of gradient descent for wide two-layer neural networks trained with the logistic loss. *arXiv preprint arXiv:2002.04486*, 2020.

Ashok Cutkosky and Francesco Orabona. Black-box reductions for parameter-free online learning in banach spaces. In *COLT*, 2018.

Damek Davis and Dmitriy Drusvyatskiy. High probability guarantees for stochastic convex optimization. In *COLT*, 2020.

John C Duchi, Alekh Agarwal, Mikael Johansson, and Michael I Jordan. Ergodic mirror descent. *SIAM Journal on Optimization*, 22(4):1549–1578, 2012.

Eduard Gorbunov, Marina Danilova, and Alexander Gasnikov. Stochastic optimization with heavy-tailed noise via accelerated gradient clipping. In *NeurIPS*, 2020.

Suriya Gunasekar, Jason Lee, Daniel Soudry, and Nathan Srebro. Characterizing implicit bias in terms of optimization geometry. *arXiv preprint arXiv:1802.08246*, 2018.

Moritz Hardt, Ben Recht, and Yoram Singer. Train faster, generalize better: Stability of stochastic gradient descent. In *International conference on machine learning*, pages 1225–1234. PMLR, 2016.

Nicholas J. A. Harvey, Christopher Liaw, Yaniv Plan, and Sikander Randhawa. Tight analyses for non-smooth stochastic gradient descent. In *COLT*, 2019.

Jean-Baptiste Hiriart-Urruty and Claude Lemaréchal. *Fundamentals of Convex Analysis*. Springer Publishing Company, Incorporated, 2001.

Yuzheng Hu, Ziwei Ji, and Matus Telgarsky. Actor-critic is implicitly biased towards high entropy optimal policies. In *ICLR*, 2022.

Ziwei Ji and Matus Telgarsky. Gradient descent aligns the layers of deep linear networks. *arXiv preprint arXiv:1810.02032*, 2018a.

Ziwei Ji and Matus Telgarsky. Risk and parameter convergence of logistic regression. *arXiv preprint arXiv:1803.07300v2*, 2018b.

Ziwei Ji and Matus Telgarsky. Directional convergence and alignment in deep learning. *arXiv preprint arXiv:2006.06657*, 2020.

Sham M Kakade, Karthik Sridharan, and Ambuj Tewari. On the complexity of linear prediction: Risk bounds, margin bounds, and regularization. *Advances in neural information processing systems*, 21, 2008.

Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

Xiaoyu Li and Francesco Orabona. A high probability analysis of adaptive sgd with momentum. *arXiv preprint arXiv:2007.14294*, 2020.

Kaifeng Lyu and Jian Li. Gradient descent maximizes the margin of homogeneous neural networks. *arXiv preprint arXiv:1906.05890*, 2019.

Sean P Meyn and Richard L Tweedie. *Markov chains and stochastic stability*. Springer Science & Business Media, 2012.

Alexander V Nazin, Arkadi S Nemirovsky, Alexandre B Tsybakov, and Anatoli B Juditsky. Algorithms of robust stochastic optimization based on mirror descent method. *Automation and Remote Control*, 80(9):1607–1627, 2019.

A. S. Nemirovski and D. B. Yudin. *Problem complexity and method efficiency in optimization*. John Wiley & Sons, 1983.

Behnam Neyshabur, Ryota Tomioka, and Nathan Srebro. In search of the real inductive bias: On the role of implicit regularization in deep learning. `arXiv:1412.6614 [cs.LG]`, 2014.

Albert B.J. Novikoff. On convergence proofs on perceptrons. *In Proceedings of the Symposium on the Mathematical Theory of Automata*, 12:615–622, 1962.

Alexander Rakhlin, Ohad Shamir, and Karthik Sridharan. Making gradient descent optimal for strongly convex stochastic optimization. In *ICML*. Citeseer, 2012.

H. Robbins and S. Monro. A stochastic approximation method. *Annals of Mathematical Statistics*, 22:400–407, 1951.

Robert E. Schapire, Yoav Freund, Peter Bartlett, and Wee Sun Lee. Boosting the margin: A new explanation for the effectiveness of voting methods. In *ICML*, pages 322–330, 1997.

Shai Shalev-Shwartz and Shai Ben-David. *Understanding Machine Learning: From Theory to Algorithms*. Cambridge University Press, 2014.

Ohad Shamir. Gradient methods never overfit on separable data. *Journal of Machine Learning Research*, 22(85):1–20, 2021.

Daniel Soudry, Elad Hoffer, Mor Shpigel Nacson, Suriya Gunasekar, and Nathan Srebro. The implicit bias of gradient descent on separable data. *arXiv preprint arXiv:1710.10345*, 2017.

Richard S Sutton. Learning to predict by the methods of temporal differences. *Machine learning*, 3 (1):9–44, 1988.

Terence Tao. 254a notes 1: Concentration of measure, Jan 2010. URL http://terrytao.wordpress.com/2010/01/03/254a-notes-1-concentration-of-measure/.

Matus Telgarsky. Margins, shrinkage, and boosting. In *ICML*, 2013.

Matus Telgarsky and Sanjoy Dasgupta. Moment-based uniform deviation bounds for k-means and friends. In *NIPS*, 2013.

Ramon van Handel. Apc 550 lecture notes: Probability in high dimensions, Dec 2016. URL https://web.math.princeton.edu/~rvan/APC550.pdf.

Tomas Vaskevicius, Varun Kanade, and Patrick Rebeschini. The statistical complexity of early-stopped mirror descent. *Advances in Neural Information Processing Systems*, 33:253–264, 2020.

Cèdric Villani. *Optimal Transport: Old and New*. Springer Science & Business Media, 2008.

Wikipedia contributors. Polylogarithm — Wikipedia, the free encyclopedia. https://en.wikipedia.org/w/index.php?title=Polylogarithm&oldid=1057121753, 2021. [Online; accessed 9-February-2022].

Jingfeng Wu, Difan Zou, Vladimir Braverman, and Quanquan Gu. Direction matters: On the implicit bias of stochastic gradient descent with moderate learning rate. *arXiv preprint arXiv:2011.02538*, 2020.

Chiyuan Zhang, Samy Bengio, Moritz Hardt, Benjamin Recht, and Oriol Vinyals. Understanding deep learning requires rethinking generalization. *ICLR*, 2017.

Tong Zhang. Solving large scale linear prediction problems using stochastic gradient descent algorithms. In *ICML*, 2004.

Tong Zhang and Bin Yu. Boosting with early stopping: Convergence and consistency. *The Annals of Statistics*, 33:1538–1579, 2005.

Wanrong Zhu, Zhipeng Lou, and Wei Biao Wu. Beyond sub-gaussian noises: Sharp concentration analysis for stochastic gradient descent. *Journal of Machine Learning Research*, 23(46):1–22, 2022.

Shaofeng Zou, Tengyu Xu, and Yingbin Liang. Finite-sample analysis for SARSA with linear function approximation. *Advances in neural information processing systems*, 32, 2019.

Appendix A. Technical preliminaries

This first appendix proves basic properties of the loss functions considered, then proves a variety of concentration inequalities, and lastly provides the proofs for the examples in Section 2.1.

A.1. Losses

First, the proof of Lemma 1, that Lipschitzness or smoothness suffice for quadratic-boundedness.

Proof (of Lemma 1) For both assumptions, it is easiest to consider classification and regression losses separately. If ℓ is α-Lipschitz, then if it is a classification loss $|\partial\ell(y,\hat{y})| = |\partial\tilde{\ell}(\text{sgn}(y)\hat{y})| \leq \alpha$, whereas for regression $|\partial\ell(y,\hat{y})| = |\partial\tilde{\ell}(y-\hat{y})| \leq \alpha$. For the case of a smooth loss, with classification

$$|\partial\ell(y,\hat{y})| \leq |\partial\tilde{\ell}(\text{sgn}(y)\hat{y}) - \partial\tilde{\ell}(0)| + |\partial\tilde{\ell}(0)| \leq \beta|\hat{y}| + |\partial\tilde{\ell}(0)|,$$

and for regression

$$|\partial\ell(y,\hat{y})| \leq |\partial\tilde{\ell}(y-\hat{y}) - \partial\tilde{\ell}(0)| + |\partial\tilde{\ell}(0)| \leq \beta|y-\hat{y}| + |\partial\tilde{\ell}(0)|.$$

■

Next, the special properties of the logistic and squared losses.

Proof (of Lemma 3) This proof is split into the two losses.

1. For the squared loss, $\widetilde{\ell}''(z) = 1$ and $\widetilde{\ell}'(z)^2 = z^2 = 2\widetilde{\ell}(z)$, implying 1-smoothness and 1-self-boundedness. For $(0,1)$-quadratic-boundedness, it suffices to apply Lemma 1 and note that $\partial\widetilde{\ell}(0) = 0$.

2. For the logistic loss, a standard calculations reveal $(1/4)$-smoothness via $\widetilde{\ell}''(z) \leq \widetilde{\ell}''(0) = 1/4$ and 1-Lipschitz via $|\widetilde{\ell}'(z)| \leq 1$. The $(1/2)$-self-bounding property was stated in (Telgarsky, 2013). For $(1,0)$-quadratically-bounding, it suffices to Lemma 1; it is nicer to use the Lipschitz bound since it shrinks some of the bounds throughout this work.

■

Lastly, a few key consequences of the definition of quadratic-boundedness, which is used in all proofs (except for TD).

Lemma 12 *If ℓ is (C_1, C_2)-quadratically-bounded, then for any (x,y) with $B_x := \max\{\|x\|_*, |y|\}$, and any u, v,*

$$\begin{aligned}\|\partial\ell_{x,y}(u)\|_* &\leq B_x\left[C_1 + C_2 B_x\left(1 + \|u\|\right)\right],\\ \|\ell_{x,y}(u) - \ell_{x,y}(v)\| &\leq B_x\|u-v\|\left[C_1 + C_2 B_x\left(1 + \|u\|\right)\right].\end{aligned}$$

In particular, given any reference point w_{ref} and set $S := \{u \in \mathbb{R}^d : \|u - w_{\text{ref}}\| \leq B_0\}$ with $B_0 \geq 1$, then every $u, v \in S$ satisfies

$$\begin{aligned}\|\partial\ell_{x,y}(u)\|_* &\leq B_x\left[C_1 + 2C_2 B_0 + C_2\|w_{\text{ref}}\|\right],\\ \|\ell_{x,y}(u) - \ell_{x,y}(v)\| &\leq B_x\|u-v\|\left[C_1 + 2C_2 B_0 + C_2\|w_{\text{ref}}\|\right].\end{aligned}$$

Proof For the first inequality, by the chain rule,

$$\|\partial_u\ell(y, x^{\mathsf{T}}u)\|_* = \|x\ell'(y, x^{\mathsf{T}}u)\|_* \leq B_x\left|C_1 + C_2\left(|y| + \left|x^{\mathsf{T}}u\right|\right)\right| \leq B_x\left[C_1 + C_2 B_x\left(1 + \|u\|\right)\right].$$

For the second inequality, using a version of the fundamental theorem of calculus for subdifferentials (Hiriart-Urruty and Lemaréchal, 2001, Theorem D.2.3.4),

$$\begin{aligned}
|\ell_{x,y}(u) - \ell_{x,y}(v)| &= \left| \int_0^1 \langle \partial \ell_{x,y}(v + t(u-v)), u - v \rangle \, \mathrm{d}t \right| \\
&\leq \int_0^1 \|\partial \ell_{x,y}(v + t(u-v))\|_* \|u - v\| \, \mathrm{d}t \\
&\leq B_x \|u - v\| \int_0^1 \left[C_1 + C_2 B_x (1 + \|v + t(u-v)\|) \right] \mathrm{d}t \\
&\leq B_x \|u - v\| \left[C_1 + C_2 B_x (1 + \|v\| + \|u - v\|/2) \right].
\end{aligned}$$

For the bounds with w_{ref} and B_0, invoking the previous two bounds with $u \in S$ and $v := w_{\text{ref}}$ gives

$$\begin{aligned}
\|\partial_u \ell(y, x^\mathsf{T} y)\|_* &\leq B_x \left[C_1 + C_2 B_x \left(1 + \|u - w_{\text{ref}}\| + \|w_{\text{ref}}\| \right) \right] \\
&\leq B_x \left[C_1 + 2 C_2 B_x B_0 + C_2 B_x \|w_{\text{ref}}\| \right], \\
|\ell_{x,y}(u) - \ell_{x,y}(v)| &\leq B_x \|u - v\| \left[C_1 + C_2 B_x (B_0 + \|w_{\text{ref}}\| + B_0/2) \right] \\
&\leq B_x \|u - v\| \left[C_1 + 2 C_2 B_x B_0 + 2 C_2 B_x \|w_{\text{ref}}\| \right],
\end{aligned}$$

as desired. ■

A.2. Concentration inequalities

The first concentration inequality is a tiny bit of algebra on top of a convenient reformulation (with elementary proof) of Freedman's inequality due to Agarwal et al. (2014). This concentration inequality will be used to get $1/t$ rates in realizable settings.

Lemma 13 *(See also Agarwal et al., 2014, Lemma 9) Let nonnegative random variables $(X_1, \ldots, X_t)$ be given with $|X_i| \leq B$. Then, for any $c \geq 4$, with probability at least $1 - \delta$,*

$$\sum_{i=1}^t \left[X_i - \mathrm{E}_{<i} X_i \right] \leq \frac{1}{c} \sum_{i=1}^t \mathrm{E}_{<i} |X_i| + cB \ln\left(\frac{1}{\delta}\right).$$

Proof For each i, define $Y_i := X_i - \mathrm{E}_{<i} X_i$, whereby $\mathrm{E}_{<i} Y_i = 0$, and $|Y_i| \leq 4(e-2)B \leq c(e-2)B$, and

$$\mathrm{E}_{<i} Y_i^2 = \mathrm{E}_{<i} \left(X_i - \mathrm{E}_{<i} X_i \right)^2 = \mathrm{E}_{<i} X_i^2 - \left(\mathrm{E}_{<i} X_i \right)^2 \leq \mathrm{E}_{<i} X_i^2 \leq B \mathrm{E}_{<i} |X_i| .$$

As such, by a version of Freedman's inequality (Agarwal et al., 2014, Lemma 9), with probability at least $1 - \delta$,

$$\sum_{i=1}^t Y_i \leq \frac{1}{cB} \sum_{i=1}^t \mathrm{E}_{<i} Y_i^2 + (e-2) cB \ln(1/\delta) \leq \frac{1}{c} \sum_{i=1}^t \mathrm{E}_{<i} |X_i| + cB \ln(1/\delta).$$

■

Next comes a concentration inequality for Markov chains mentioned in the body. The proof is based on a very nice one due to Duchi et al. (2012, Proposition 1), though re-organized and fully decoupled from mirror descent; e.g., this same bound will be used in the TD proofs. That said, all the core ideas are from (Duchi et al., 2012, Proposition 1).

Lemma 14 *(See also Duchi et al., 2012, Proposition 1) Let $\epsilon \geq 0$ be given along with approximate stationarity witness (π, τ, ϵ) on a stochastic process $(x_i)_{i \leq t}$. Let f be given with $|f(x; w)| \leq B_f$ almost surely, and suppose $|f(x_{i+1}; w_i) - f(x_{i+1}; w_{i-\tau+1})| \leq B_i$ almost surely for all i. With probability at least $1 - \tau\delta$,*

$$\sum_{i<t} \left[f(x_{i+1}; w_i) - \mathrm{E}_{x \sim \pi} f(x; w_i) \right] \leq 2B_f \left(2\tau - 2 + t\epsilon + \sqrt{t\tau \ln(1/\delta)} \right) + \sum_{i=\tau-1}^{t-1} B_i.$$

A notable characteristic of this bound (also present in the version due to Duchi et al. (2012)) is that in the IID setting with $\tau_{\mathrm{TV}}(\epsilon) = 1$ and $\epsilon \approx 0$, the bound is exactly what one would expect from Azuma's inequality, with no excess.

Proof The key idea of the proof is to introduce gaps of length τ wherever x_i and w_i interact, making them approximately independent. To accomplish this, following a proof idea due to Duchi et al. (2012), the summation over time will be replaced with τ interleaved summations, where w_i only interacts with $x_{i+\tau}$, meaning enough time has been inserted to ensure mixing.

Before proceeding with the bulk of the proof, let's dispense with the nuisance case $t < 2\tau$. If $t = 0$, the bound is direct, and if $t > 0$, by the definition of B_f, almost surely

$$\sum_{i<t} \left[f(x_{i+1}; w_i) - \mathrm{E}_{x \sim \pi} f(x; w_i) \right] \leq 2tB_f \leq 2B_f(2\tau - 1) \leq 2B_f(2\tau - 2 + \sqrt{t\tau \ln(1/\delta)}).$$

which is upper bounded by the final desired quantity and completes the proof in the case $t < 2\tau$. For the remainder of the proof, suppose $t \geq 2\tau$.

Due essentially to boundary conditions, a bit of additional notation and care are needed, especially to ensure that the bound loses nothing when $\tau = 1$ (e.g., the IID case). Let $\rho \geq \tau - 1$ denote the smallest time so that $\tau | (t - \rho)$ and let $n := (t - \rho)/\tau$; this ρ is the number of iterates that will be thrown out so that the overall sum can be split into n interleaved sums. Since $t \geq 2\tau$, then ρ always exists and satisfies $\tau - 1 \leq \rho \leq 2(\tau - 1)$ (if $t - (\tau - 1)$ is not a multiple of τ, then there must be a multiple of τ within $\{t - 2(\tau - 1), \ldots, t - \tau\}$), and $t/(2\tau) \leq n \leq t/\tau$.

To start, defining $B_\tau := \sum_{i=\tau-1}^{t-1} B_i$ for convenience,

$$\begin{aligned}
\sum_{i<t} f(x_{i+1}; w_i) &= \sum_{i<\tau-1} f(x_{i+1}; w_i) \\
&\quad + \sum_{i=\tau-1}^{t-\rho+\tau-2} \left(f(x_{i+1}; w_i) - f(x_{i+1}; w_{i-\tau+1}) + f(x_{i+1}; w_{i-\tau+1})\right) \\
&\quad + \sum_{i=t-\rho+\tau-1}^{t-1} f(x_{i+1}; w_i) \\
&\le \rho B_f + B_\tau + \sum_{i=0}^{t-\rho-1} f(x_{i+\tau}; w_i) \\
&\le \rho B_f + B_\tau + \sum_{i=0}^{t-\rho-1} \mathrm{E}_{\mathrm{X}_i} f(x_{i+\tau}; w_i) + \sum_{i=0}^{t-\rho-1} \left(f(x_{i+\tau}; w_i) - \mathrm{E}_{\mathrm{X}_i} f(x_{i+\tau}; w_i)\right) \\
&\le 2\rho B_f + B_\tau + \sum_{i<\tau} \mathrm{E}_{x\sim\pi} f(x; w_i) \\
&\quad + \sum_{i=0}^{t-\rho-1} \left(\mathrm{E}_{\mathrm{X}_i} f(x_{i+\tau}; w_i) - \mathrm{E}_{x\sim\pi} f(x; w_i)\right) \\
&\quad + \sum_{i=0}^{t-\rho-1} \left(f(x_{i+\tau}; w_i) - \mathrm{E}_{\mathrm{X}_i} f(x_{i+\tau}; w_i)\right). \qquad (7)
\end{aligned}$$

The rest of the proof will handle these last two summations, the first via the definition of τ and the mixing properties of the stochastic process, and the second via concentration inequalities and the aforementioned interleaved sum technique.

For the first summation, fix any i, and let ξ_i be a coupling between the distribution of $x_{i+\tau}$ conditioned on $\mathcal{F}_i$, and the stationary distribution π. By the coupling characterization of total variation (Villani, 2008, Equation 6.11), and since $\mathrm{TV}(P_i^{i+\tau}, \pi) \le \epsilon$ by the definition of $\tau = \tau_{\mathrm{TV}}(\epsilon)$,

$$\begin{aligned}
\mathrm{E}_{\mathrm{X}_i} f(x_{i+\tau}; w_i) - \mathrm{E}_{x\sim\pi} f(x; w_i) &= \mathrm{E}_{(y,x)\sim\xi_i} \left(f(y; w_i) - f(x; w_i)\right) \\
&\le 2B_f \mathrm{Pr}_{(y,x)\sim\xi_i}[x \ne y] \\
&\le 2B_f \epsilon. \qquad (8)
\end{aligned}$$

Summing these errors adds a term $2\epsilon(t-\rho)B_f$ to eq. (7).

The final summation in eq. (7) will utilize the aforementioned technique of splitting the sum into τ interleaved sums. Concretely,

$$\sum_{i=0}^{t-\rho-1} \left(f(x_{i+\tau}; w_i) - \mathrm{E}_{\mathrm{X}_i} f(x_{i+\tau}; w_i)\right) = \sum_{j=0}^{\tau-1} \left[\sum_{k=0}^{n-1} \left(f(x_{k\tau+j+\tau}; w_{k\tau+j}) - \mathrm{E}_{\mathrm{X}_i} f(x_{k\tau+j+\tau}; w_{k\tau+j})\right)\right],$$

and all that remains is to apply a concentration inequality to the τ inner summations, and collect terms. By τ applications of Azuma's inequality, with probability at least $1 - \tau\delta$, simultaneously for

all $j \in \{0, \dots, \tau - 1\}$,

$$\sum_{k=0}^{n-1} \left(f(x_{k\tau+j+\tau}; w_{k\tau+j}) - \mathrm{Ex}_i f(x_{k\tau+j+\tau}; w_{k\tau+j}) \right) \leq B_f \sqrt{2n \ln(1/\delta)}.$$

Combining all this with eq. (7), the total variation bound in eq. (8), and simplifying to replace ρ and n with their respective bounds gives

$$\begin{aligned} \sum_{i<t} \left[f(x_{i+1}; w_i) - \mathrm{Ex}_{x\sim\pi} f(x; w_i) \right] &\leq 2(\rho + \epsilon(t-\rho))B_f + B_\tau + B_f \tau \sqrt{2n \ln(1/\delta)} \\ &\leq 2(2\tau - 2 + t\epsilon)B_f + B_\tau + B_f \sqrt{2t\tau \ln(1/\delta)}, \end{aligned}$$

which completes the proof. ■

The last collection of concentration inequalities needed are for heavy-tailed losses. There appear to be many such inequalities, for example here is one due to Blum et al. (2017).

Lemma 15 *(See also Blum et al., 2017) Let IID random variables $(Z_i, \dots, Z_t)$ be given with variance at most σ^2, and suppose there exists $m \leq t\sigma^2/2$ with $\left| \mathrm{Ex} \left(\|y_i x_i\|^2 - \mathrm{Ex}\|y_i x_i\|^2 \right)^r \right| \leq \sigma^2 r!$ for $r \in \{3, \dots, m\}$. If $\delta \geq 1/(n\sigma^2)^{s/2}$, then with probability at least $1 - \delta$,*

$$\left| \sum_i \left(Z_i - \mathrm{Ex} Z_i^2 \right) \right| \leq \frac{\sigma\sqrt{2sn}}{\delta^{1/s}}.$$

Proof This version performs minor algebra to repackage the original. ■

The conditions on the preceding are a little complicated, so here is a potentially worse bound with many fewer conditions to check. It is presented in (Telgarsky and Dasgupta, 2013), but follows a proof scheme due to Tao (2010, Equation 7), making adjustments to drop boundedness assumptions.

Lemma 16 *(Appears in (Telgarsky and Dasgupta, 2013) following a proof scheme due to Tao (2010, Equation 7).) Let IID random variables $(Z_i, \dots, Z_t)$ be given. Suppose p is even, and define $M := \max\{p/e, \sup_{2\leq r\leq p} \mathrm{Ex} |Z_i - \mathrm{Ex} Z_i|^r\}$. Then with probability at least $1 - \delta$,*

$$\left| \sum_{i=1}^t (Z_i - \mathrm{Ex} Z_i) \right| \leq 2M\sqrt{t} \left(\frac{2}{\delta} \right)^{1/p}.$$

Proof This is the same as (Telgarsky and Dasgupta, 2013, Lemma A.3), except the requirement $M \geq p/e$ simplifies the bound, in particular the requirement $t \geq p/(Me)$ is now simply $t \geq 1$. ■

A.3. Analysis of examples in Section 2.1

To conclude this section of technical preliminaries, here are full proofs corresponding to the illustrative examples presented in Section 2.1. First is the simple margin-like bound.

Proof (of Proposition 6) First note that if $u^\mathsf{T} xy \geq \gamma$, then $w_{\text{ref}}^\mathsf{T} xy \geq \ln t$; otherwise, $w_{\text{ref}}^\mathsf{T} xy \geq -\|w_{\text{ref}}\| \geq -\ln(t)/\gamma$ almost surely. Using the provided choice of w_{ref} and the elementary upper bounds $\ell(y,\hat{y}) = \ln(1+\exp(-y\hat{y})) \leq \exp(-y\hat{y})$ when $y\hat{y} \geq \gamma$ and $\ell(y,\hat{y}) \leq 1 - y\hat{y}$ when $y\hat{y} \leq 0$ gives

$$\begin{aligned}
\mathrm{E}_{x,y}\ell(y, x^\mathsf{T} w_{\text{ref}}) &\leq \widetilde{\ell}\left(\ln(t)\right) \mathrm{Pr}[u^\mathsf{T} xy \geq \gamma] + \widetilde{\ell}\left(-\ln(t)/\gamma\right) \mathrm{Pr}[u^\mathsf{T} xy < \gamma] \\
&\leq \exp(-\ln(t)) + \left(1 + \frac{\ln t}{\gamma}\right)\frac{1}{t} \\
&\leq \frac{1}{t}\left(2 + \frac{\ln(t)}{\gamma}\right).
\end{aligned}$$

■

Next is the characterization of the optimal path corresponding to the sphere data depicted in Figure 2(*b*).

Proof (of Proposition 7) The proof first establishes that $u_r := re_1$ (for $r > 0$) is the unique risk minimizer with norm r, which will be established via symmetry argument. Consider any other solution v with $\|v\| = r$. To avoid dealing with labels, let μ_1 denote the density on the points $\{z \in \mathcal{S}_{d-1} : z_1 \geq 0\}$ obtain by sampling (x,y) and then multiplying to obtain $z := xy$. For any z in the support of μ_1, consider the behavior of v on z and $z' := 2z_1 e_1 - z$, which reflects z around $z_1 e_1$. With probability 1, $z \neq e_1$, and thus $z \neq z'$, and by strict convexity

$$\begin{aligned}
\frac{1}{2}\ln(1+\exp(-v^\mathsf{T} z)) + \frac{1}{2}\ln(1+\exp(-v^\mathsf{T} z')) &> \ln(1+\exp(-v^\mathsf{T}(z+z')/2)) \\
&= \ln(1+\exp(-v^\mathsf{T} e_1 z_1)) \\
&\geq \ln(1+\exp(-u_r^\mathsf{T} z)).
\end{aligned}$$

Since z and $2z_1 e_1 - z$ have equal probability density, letting μ_2 denote the probability density obtained from μ_1 by conditioning on $z_2 > 0$,

$$\begin{aligned}
\mathcal{R}(v) &= \int \ln(1+\exp(-v^\mathsf{T} z))\,\mathrm{d}\mu_1(z) \\
&= \int \ln(1+\exp(-v^\mathsf{T} z))\,\mathrm{d}\mu_1(z) \\
&= \frac{1}{2}\int \left(\ln(1+\exp(-v^\mathsf{T} z)) + \ln(1+\exp(-v^\mathsf{T}(2z_1 e_1 - z)))\right)\mathrm{d}\mu_2(z) \\
&> \int \ln(1+\exp(-u_r^\mathsf{T} z))\,\mathrm{d}\mu_2(z) \\
&= \frac{1}{2}\int \left(\ln(1+\exp(-u_r^\mathsf{T}(2z_1 e_1 - z))) + \ln(1+\exp(-u_r^\mathsf{T} z))\right)\mathrm{d}\mu_2(z) \\
&= \int \ln(1+\exp(-u_r^\mathsf{T} z))\,\mathrm{d}\mu_1(z) \\
&= \mathcal{R}(u_r),
\end{aligned}$$

establishing that u_r is the unique optimal solution of norm r.

Next, still letting ρ denote the uniform measure on the hemisphere with combined variable z, to compute $\mathcal{R}(u_r)$ exactly, recall the *dilogarithm function* $\mathrm{Li}_2(z)$ (Wikipedia contributors, 2021), which satisfies

$$\mathrm{Li}_2(z) := \sum_{k\geq 1} \frac{z^k}{k^2}, \qquad \frac{\mathrm{d}}{\mathrm{d}z}\mathrm{Li}_2(z) = -\ln(1-z).$$

Then

$$\begin{aligned}
\mathcal{R}(u_r) = \int \ln(1+\exp(-u_r^{\mathsf{T}} z))\,\mathrm{d}\rho(z) &= \int_0^1 \ln(1+\exp(-rs))\,\mathrm{d}s \\
&= -\frac{1}{r}\,\mathrm{Li}(-\exp(-rs))|_0^1 \\
&= \frac{1}{r}\left(-\mathrm{Li}(-\exp(-r)) + \frac{\pi^2}{12}\right).
\end{aligned}$$

To control this further, $\mathrm{Li}(-\exp(-r))$ can be written

$$\begin{aligned}
\mathrm{Li}(-\exp(-r)) &= \sum_{k\geq 1} \frac{(-\exp(-r))^k}{k} \\
&= \int_r^\infty \sum_{k\geq 1} \exp(-kr)) \\
&= \sum_{k\geq 1} \int_r^\infty \exp(-kr)) \\
&= \sum_{k\geq 1} \frac{\exp(-kr)}{k},
\end{aligned}$$

which is at least $\exp(-r)$, implying

$$\mathcal{R}(u_r) \leq \frac{1}{r}\left(-\exp(-r) + \frac{\pi^2}{12}\right).$$

For the lower bound, if $r \geq 1$, then

$$\begin{aligned}
\sum_{k\geq 1} \frac{\exp(-kr)}{k} &\leq \exp(-r) + \frac{1}{2}\sum_{k\geq 2} \exp(-kr) \\
&= \exp(-r) + \frac{1}{2}\frac{\exp(-2r)}{1-\exp(-r)} \\
&= \exp(-r)\left(1 + \frac{1}{2(\exp(r)-1)}\right) \\
&\leq 2\exp(-r),
\end{aligned}$$

which completes the proof. ■

Appendix B. Analysis of mirror descent (MD)

This section collects all proofs related to mirror descent.

B.1. Deterministic mirror descent analysis

While the core deterministic proof for mirror descent is completely standard, the standard presentation always omits the term $D_\psi(w_{\text{ref}}, w_t)$, which as mentioned is the basis for this entire work. As such, the standard guarantee is re-proved with this term included; the proof itself has otherwise nothing new over other versions, and most closely follows one due to Duchi et al. (2012).

Lemma 17 *Let closed convex S be given, let $w_0 \in S$ and $w_{\text{ref}} \in S$ be given, and define $w_{i+1} := \arg\min_{w \in S} \left(\langle \eta g_{i+1}, w \rangle + D_\psi(w, w_i) \right)$ with a corresponding given sequence of functions $(f_i)_{i=1}^t$ and arbitrary subgradients $g_{i+1} \in \partial f_{i+1}(w_i)$ (whereby $\|g_{i+1}\|_* \leq \|\partial f_{i+1}(w_i)\|_*$), where $D_\psi(w, v) := \psi(w) - [\psi(v) + \langle \nabla\psi(v), w - v\rangle]$ and ψ is 1-strongly-convex with respect to some norm $\|\cdot\|$. Then, for all $i \leq t$,*

$$D_\psi(w_{\text{ref}}, w_t) \leq D_\psi(w_{\text{ref}}, w_0) + \eta \sum_{i<t} \langle \partial f_{i+1}(w_i), w_{\text{ref}} - w_i \rangle + \sum_{i<t} \frac{\eta^2}{2} \|\partial f_{i+1}(w_i)\|_*^2,$$

and it holds for every $i < t$ that $\|w_{i+1} - w_i\| \leq \eta \|g_{i+1}\|_*$. If additionally each f_i is convex, then

$$D_\psi(w_{\text{ref}}, w_t) \leq D_\psi(w_{\text{ref}}, w_0) + \eta \sum_{i<t} [f_{i+1}(w_{\text{ref}}) - f_{i+1}(w_i)] + \sum_{i<t} \frac{\eta^2}{2} \|\partial f_{i+1}(w_i)\|_*^2.$$

Proof Fix any iteration i. By the first-order conditions on the choice of w_{i+1}, for any $v \in S$,

$$\langle \eta g_{i+1} + \nabla\psi(w_{i+1}) - \nabla\psi(w_i), v - w_{i+1} \rangle \geq 0.$$

Instantiating this first-order condition with $v = w_i$ rearranges to give

$$\eta \|g_{i+1}\|_* \|w_i - w_{i+1}\| \geq \langle \eta g_{i+1}, w_i - w_{i+1} \rangle \geq \langle \nabla\psi(w_{i+1}) - \nabla\psi(w_i), w_{i+1} - w_i \rangle \geq \|w_{i+1} - w_i\|^2,$$

which implies $\|w_{i+1} - w_i\| \leq \eta \|g_{i+1}\|_* \leq \eta \|\partial f_{i+1}(w_i)\|_*$ from the statement. On the other hand, instantiating the first-order condition with $v = w_{\text{ref}}$ gives

$$\begin{aligned} \langle \nabla\psi(w_i) - \nabla\psi(w_{i+1}), w_{\text{ref}} - w_{i+1} \rangle &\leq \langle \eta g_{i+1}, w_{\text{ref}} - w_{i+1} \rangle \\ &\leq \eta \left[\langle g_{i+1}, w_{\text{ref}} - w_i \rangle + \langle g_{i+1}, w_i - w_{i+1} \rangle \right], \end{aligned}$$

which combines with the definition of D_ψ to give

$$\begin{aligned} D_\psi(w_{\text{ref}}, w_{i+1}) - D_\psi(w_{\text{ref}}, w_i) &= \psi(w_i) - \psi(w_{i+1}) - \langle \nabla\psi(w_{i+1}), w_{\text{ref}} - w_{i+1} \rangle + \langle \nabla\psi(w_i), w_{\text{ref}} - w_i \rangle \\ &= -D_\psi(w_{i+1}, w_i) + \langle \nabla\psi(w_i) - \nabla\psi(w_{i+1}), w_{\text{ref}} - w_{i+1} \rangle \\ &\leq \eta \langle g_{i+1}, w_{\text{ref}} - w_i \rangle + \eta \left[\langle g_{i+1}, w_i - w_{i+1} \rangle - \frac{1}{\eta} D_\psi(w_{i+1}, w_i) \right]. \end{aligned}$$

To simplify this further, by the Fenchel-Young inequality and strong convexity of D_ψ,

$$\begin{aligned} \langle \partial f_{i+1}(w_i), w_i - w_{i+1} \rangle - \frac{1}{\eta} D_\psi(w_{i+1}, w_i) &\leq \frac{\eta}{2} \|\partial f_{i+1}(w_i)\|_*^2 + \frac{1}{2\eta} \|w_i - w_{i+1}\|^2 - \frac{1}{\eta} D_\psi(w_{i+1}, w_i) \\ &\leq \frac{\eta}{2} \|\partial f_{i+1}(w_i)\|_*^2. \end{aligned}$$

which after summing for all $i < t$, telescoping and rearranging gives

$$D_\psi(w_{\text{ref}}, w_t) \leq D_\psi(w_{\text{ref}}, w_0) + \eta \sum_{i<t} \langle g_{i+1}, w_{\text{ref}} - w_i \rangle + \eta \sum_{i<t} \frac{\eta}{2} \|\partial f_{i+1}(w_i)\|_*^2.$$

Lastly, the version of the claim with convexity follows from $\langle g_{i+1}, w_{\text{ref}} - w_i \rangle \leq f_{i+1}(w_{\text{ref}}) - f_{i+1}(w_i)$. ∎

B.2. The realizable case: Theorem 5

This proof makes use of the variant of Freedman's inequality restated in Lemma 13.

Proof (of Theorem 5) As in the other proofs, let $(v_i)_{i\leq t}$ be coupled iterates projected onto the ball $S := \{v \in \mathbb{R}^d : \|v - w_{\text{ref}}\| \leq B_w\}$, where $v_0 = w_0$ and common random data $((x_i, y_i))_{i=1}^t$ are used. The proof will apply concentration inequalities on the common sample space, and then show that $(w_i)_{i\leq t} = (v_i)_{i\leq t}$ and that both share good risk and norm guarantees.

Unlike the other proofs, this realizable setting will apply two separate concentration inequalities in order to allow cleaner step sizes. Concretely, the first concentration inequality will be on $\sum_{j<i} \ell_{j+1}(w_{\text{ref}})$ alone. Since $|\ell_{j+1}(w_{\text{ref}})| \leq C_4$ almost surely by assumption, then by t applications of Lemma 13 with constant $c = 4$ gives, with probability at least $1 - t\delta$, simultaneously for all $i \leq t$,

$$\eta \sum_{j<i} \ell_{j+1}(w_{\text{ref}}) \leq \frac{5\eta i}{4} \mathcal{R}(w_{\text{ref}}) + 4\eta C_4 \ln(1/\delta) = \frac{5\eta i}{4} \mathcal{R}(w_{\text{ref}}) + \frac{B_w^2}{16}.$$

The second concentration inequality will as usual involve both $\ell_{j+1}(z)$ and $\ell_{j+1}(v_j)$. To start, again using the constant C_4 but also Theorem 12, and defining $C_5 := (C_4 + (C_1 + 2C_2 B_w + C_2\|w_{\text{ref}}\|)B_w)/2$ for convenience, for any $j < t$ and any $u \in S$,

$$|2\ell_{j+1}(w_{\text{ref}}) - \ell_{j+1}(v)| \leq |\ell_{j+1}(w_{\text{ref}})| + |\ell_{j+1}(w_{\text{ref}}) - \ell_{j+1}(v)| \leq 2C_5.$$

Combining this bound with with another t applications of Lemma 13 with constant $c = 4$, with probability at least $1 - t\delta$, simultaneously for all $i \leq t$,

$$\sum_{j<i} [\ell_{j+1}(w_{\text{ref}}) - (1/2)\ell_{j+1}(v_j)] \leq \sum_{j<i} [(5/4)\mathcal{R}(w_{\text{ref}}) - (3/8)\mathcal{R}(v_j)] + 4C_5 \ln(1/\delta).$$

For the remainder of the proof, discard the combined failure event from the preceding bounds, which together removes $2t\delta$ probability mass.

The first concentration alone will now be used to prove $w_i = v_i \in S$ by induction; the base case $w_0 = v_0 \in S$ is direct, thus consider some $i > 0$. By the concentration inequality on $\sum_{j<i} \ell_{j+1}(w_{\text{ref}})$, the fact that $\ell_{j+1}(w_j) \geq 0$, and using additionally $\mathcal{R}(w_{\text{ref}}) \leq \rho D_\psi(w_{\text{ref}}, w_0)/t$, and lastly the definition of ρ-self-bounding, the deterministic mirror descent guarantee from Lemma 17

becomes

$$\begin{aligned} D_\psi(w_{\text{ref}}, w_i) &\le D_\psi(w_{\text{ref}}, w_0) + \eta \sum_{j<i} [\ell_{j+1}(w_{\text{ref}}) - \ell_{j+1}(w_j)] + \sum_{j<i} \frac{\eta^2}{2} \ell'_{j+1}(w_j)^2 \|x_{j+1} y_{j+1}\|_*^2 \\ &\le \frac{B_w^2}{4} + \eta \sum_{j<i} [\ell_{j+1}(w_{\text{ref}}) - (1-\eta\rho)\ell_{j+1}(w_j)] \qquad (9) \\ &\le \frac{B_w^2}{4} + \eta \sum_{j<i} \ell_{j+1}(w_{\text{ref}}) \\ &\le \frac{B_w^2}{4} + \frac{5i\eta}{4} \mathcal{R}(w_{\text{ref}}) + \frac{B_w^2}{16} \\ &\le \frac{B_w^2}{4} + \frac{5 D_\psi(w_{\text{ref}}, w_0)}{8} + \frac{B_w^2}{16} \le \frac{3B_w^2}{8}, \end{aligned}$$

which establishes the desired norm control since $D_\psi(w_{\text{ref}}, w_i) \ge \|w_i - w_{\text{ref}}\|^2/2$, but also since $v_{i-1} = w_{i-1}$, then the construction of v_i will not invoke the constraint and $v_i = w_i$.

For the risk control, for any $i \le t$, by the second concentration inequality above, using $(v_j)_{j<i} = (w_j)_{j<i}$ and continuing with the deterministic mirror descent bound from eq. (9),

$$\begin{aligned} D_\psi(w_{\text{ref}}, w_i) &\le \frac{B_w^2}{4} + \eta \sum_{j<i} [\ell_{j+1}(w_{\text{ref}}) - (1/2)\ell_{j+1}(w_j)] \\ &= \frac{B_w^2}{4} + \eta \sum_{j<i} [\ell_{j+1}(w_{\text{ref}}) - (1/2)\ell_{j+1}(v_j)] \\ &\le \frac{B_w^2}{4} + \eta \sum_{j<i} [(5/4)\mathcal{R}(w_{\text{ref}}) - (3/8)\mathcal{R}(v_j)] + 4\eta C_5 \ln(1/\delta), \end{aligned}$$

which after expanding the choice of C_5 gives the desired bound. ■

B.3. The non-realizable, Markov case: Theorem 8

This proof makes use of the Markov chain concentration inequality in Lemma 14.

Proof (of Theorem 8) Let $(v_i)_{i\le t}$ denote the coupled projected mirror descent iterates using projection ball $S := \{v \in \mathbb{R}^d : \|v - w_{\text{ref}}\| \le B_w\}$, which are coupled in the strong sense that that $v_0 = w_0$, and thereafter $(v_i)_{i\le t}$ and $(w_i)_{i\le t}$ use the exact same data sequence $((x_i, y_i))_{i=1}^t$. As in the general proof scheme, the first step is to apply a concentration inequality on this shared data $((x_i, y_i))_{i=1}^t$, and then show that $(v_i)_{i\le t} = (w_i)_{i\le t}$ by induction.

Before starting on the general proof scheme, to simplify the interaction with the loss conditions, define $C_3 := C_1 + 2C_2 B_w + B_w$ for convenience, and note by Theorem 12 and since $\max\{\|x\|_*, |y|\} \le 1$, for any $u \in S$,

$$\sup_{j<t} \|\partial \ell_{j+1}(u)\|_* \le C_3, \qquad \sup_{j<t} \|\ell_{j+1}(u) - \ell_{j+1}(v)\| \le C_3 \|u - v\|. \qquad (10)$$

Lastly, $C_3 \le 4B_w \max\{1, C_1, C_2\}$.

The concentration inequality will be the general stochastic process bound in Lemma 14, applied to $\eta \sum_{j<i} [\ell_{j+1}(w_{\text{ref}}) - \ell_{j+1}(v_j)]$ for all $i \leq t$, which requires almost sure bounds on two quantities. The first is a uniform control on individual differences within this summation, which thanks to eq. (10) is simply

$$\sup_{j<i} |\ell_{j+1}(w_{\text{ref}}) - \ell_{j+1}(v_j)| \leq C_3 B_w \qquad \text{a.s..}$$

The second almost sure bound is on a similar difference but on iterates which are τ apart, which follows by combining eq. (10) with the per-iteration guarantee from Lemma 17:

$$|\ell_{i+\tau+1}(v_{i+\tau}) - \ell_{i+\tau+1}(v_i)| \leq C_3 \|v_{i+\tau} - v_i\| \leq C_3 \sum_{j=i}^{i+\tau-1} \eta \|\nabla \ell_{j+1}(v_j)\|_* \leq \eta\tau C_3^2.$$

As such, union bounding t applications of Lemma 14 with probability at least $1 - t\tau\delta$, simultaneously for all $i \leq t$,

$$\eta \sum_{j<i} [\ell_{j+1}(w_{\text{ref}}) - \ell_{j+1}(v_j) - \mathcal{R}(w_{\text{ref}}) + \mathcal{R}(v_j)] \leq 2\eta C_3 B_w \left[2\tau + i\epsilon + \sqrt{i\tau \ln(1/\delta)}\right] + i\eta^2 \tau C_3^2.$$

If $i \geq 2\tau$, then this bound simplifies to

$$\begin{aligned}
\eta \sum_{j<i} [\ell_{j+1}(w_{\text{ref}}) - \ell_{j+1}(v_j) - \mathcal{R}(w_{\text{ref}}) + \mathcal{R}(v_j)] &\leq 2\eta C_3 B_w \left[\sqrt{i\tau} + \sqrt{i} + \sqrt{i\tau \ln(1/\delta)}\right] + i\tau\eta^2 C_3^2 \\
&\leq \frac{6B_w^2 + B_w^2}{1024} \\
&\leq \frac{B_w^2}{128}.
\end{aligned}$$

On the other hand, if $i < 2\tau$, then forgoing Lemma 14 entirely and using the almost sure bounds on the left hand side directly,

$$\eta \sum_{j<i} [\ell_{j+1}(w_{\text{ref}}) - \ell_{j+1}(v_j) - \mathcal{R}_\pi(w_{\text{ref}}) + \mathcal{R}_\pi(v_j)] \leq 2\eta i C_3 B_w \leq 2\eta\sqrt{2i\tau} C_3 B_w \leq \frac{B_w^2}{128}.$$

The remainder of the proof discards the common $t\tau\delta$ failure probability on the underlying sample space $((x_i, y_i))_{i_1}^t$ which is shared by $(v_i)_{i\leq t}$ and $(w_i)_{i\leq t}$.

The proof now proceeds by induction, establishing $v_i = w_i$ for $i \leq t$ and the corresponding risk bound. The base case $w_0 = v_0$ is by definition of the coupling, thus consider the construction of some w_i with $i > 0$; by the deterministic mirror descent guarantee in Lemma 17, the inductive

hypothesis $(v_i)_{j<i} = (w_i)_{j<i}$, and the above concentration inequality on $(v_j)_{j<i}$,

$$\begin{aligned}
D_\psi(w_{\text{ref}}, w_i) &\le D_\psi(w_{\text{ref}}, w_0) + \eta \sum_{j<i} [\ell_{j+1}(w_{\text{ref}}) - \ell_{j+1}(w_j)] + \sum_{j<i} \frac{\eta^2}{2} \|\nabla \ell_{j+1}(w_j)\|_*^2 \\
&= D_\psi(w_{\text{ref}}, w_0) + \eta \sum_{j<i} [\ell_{j+1}(w_{\text{ref}}) - \ell_{j+1}(v_j)] + \sum_{j<i} \frac{\eta^2}{2} \|\nabla \ell_{j+1}(v_j)\|_*^2 \\
&\le \frac{B_w^2}{16} + \eta \sum_{j<i} [\mathcal{R}(w_{\text{ref}}) - \mathcal{R}(v_j)] + \frac{B_w^2}{128} + \frac{i\eta^2 C_3^2}{2} \\
&\le \eta \sum_{j<i} [\mathcal{R}(w_{\text{ref}}) - \mathcal{R}(v_j)] + \frac{B_w^2}{8},
\end{aligned}$$

which establishes the risk guarantee on $(w_j)_{j<i}$ after substituting $(v_j)_{j<i} = (w_j)_{j<i}$ back in. To verify $w_i = v_i \in S$, it suffices to establish $\|w_i - w_{\text{ref}}\| < B_w$, which means v_i will not encounter its projection and $w_i = v_i$; to this end, combining the preceding with the fact $\mathcal{R}(w_{\text{ref}}) \le \frac{D_\psi(w_{\text{ref}}, w_0)}{\sqrt{t}} + \inf_v \mathcal{R}(v) \le \frac{B_w^2}{16\sqrt{t}} + \inf_v \mathcal{R}(v)$, then

$$\frac{1}{2}\|w_i - w_{\text{ref}}\|^2 \le D_\psi(w_{\text{ref}}, w_i) \le \frac{B_w^2}{8} + \eta \sum_{j<i} [\mathcal{R}(w_{\text{ref}}) - \mathcal{R}(w_j)] \le \frac{B_w^2}{8} + \eta \sum_{j<i} \frac{B_w^2}{16\sqrt{t}} \le \frac{B_w^2}{4}$$

as desired. ∎

B.4. Heavy-tailed data: Theorem 10

This proof makes use of the heavy-tail concentration inequalities at the end of Appendix A.2.

Proof (of Theorem 10) This proof will follow the usual scheme — applying concentration to projected iterates $(v_i)_{i\le t}$ with $v_0 = w_0$ which are coupled to $(w_i)_{i\le t}$ and satisfy $v_i \in S$, and then separately apply concentration to $(v_i)_{i\le t}$ and derive $v_i = w_i \in S$ and a risk bound on both — but will additionally apply concentration to $\|x_i y_i\|_*^2$. To this end and to simplify a few terms, define $C_3 := C_1 + C_2 B_w + B_w$, whereby Theorem 12 grants, for any $j < t$ and any $v \in S$,

$$\|\partial \ell_{j+1}(v)\| \le Z_{j+1} C_3, \qquad |\ell_{j+1}(w_{\text{ref}}) - \ell_{j+1}(v)| \le Z_{j+1}^2 C_3 B_w. \tag{11}$$

To this end, the first step is to use one of the two assumptions to control $\sum_{j=1}^i \max\{1, \|x_j\|_*^4, |y_j|^4\}$.

1. **(Subgaussian tails.)** Union bounding over $2t$ standard subgaussian bounds (van Handel, 2016), simultaneously for every $i \le t$,

$$\sum_{j=1}^i Z_j \le t\text{E}\text{x}Z_1 + 2\sigma\sqrt{t\ln(1/\delta)} =: tC_7.$$

2. **(Polynomial tails.)** Union bounding over $2t$ applications of Theorem 16, simultaneously for every $i \le t$,

$$\sum_{j=1}^i Z_j \le t\text{E}\text{x}Z_1 + 2M\sqrt{t}\left(\frac{2}{\delta}\right)^{1/p} =: tC_8.$$

The rest of the proof will simply use C to denote either C_7 or C_8, and the final bounds will be obtained by using the appropriate setting to expand the definition of C.

Next comes the concentration inequality on $\sum_{j<i}[\ell_{j+1}(w_{\text{ref}}) - \ell_{j+1}(v_j)]$ for all $i \le t$. It will not be possibly to apply Azuma's inequality directly, since the increments do not have a uniform control; instead, a very nice extension of Azuma's inequality, presented by (van Handel, 2016, Problem 3.11), will allow us to use the varying increments which were controlled with high probability above. In particular, combining the above moment bounds with the loss bounds from eq. (11) gives

$$\sum_{j<i} |\ell_{j+1}(w_{\text{ref}}) - \ell_{j+1}(v_j)|^2 \le \sum_{j<i} Z_{j+1}^4 C_3^2 B_w^2 \le 16 t C C_3^2 B_w^2.$$

As such, applying the variant of Azuma's inequality from (van Handel, 2016, Problem 3.11) to each $i \le t$ and union bounding, and using the earlier control on the data norms to remove the "and" case from the bound in (van Handel, 2016, Problem 3.11), then with probability at least $1 - t\delta$, simultaneously for every $i \le t$,

$$\begin{aligned}\sum_{j<i} [\ell_{j+1}(w_{\text{ref}}) - \ell_{j+1}(v_j) - \mathcal{R}(w_{\text{ref}}) + \mathcal{R}(v_j)] &\le \sqrt{\frac{16 t C C_3^2 B_w^2 \ln(1/\delta)}{2}} \\ &\le 4 C_3 B_w \sqrt{tC \ln(1/\delta)}.\end{aligned}$$

This completes the expanded concentration part of the proof technique.

The induction part now proceeds as usual. The base case has $w_0 = v_0 \in S$ by the initial conditions, thus consider $i > 0$. By the deterministic mirror descent guarantee in Lemma 17 and since $(w_j)_{j<i} = (v_j)_{j<i}$, and also controlling the gradient norm via eq. (11),

$$\begin{aligned} D_\psi(w_{\text{ref}}, w_i) &\le D_\psi(w_{\text{ref}}, w_0) + \eta \sum_{j<i} [\ell_{j+1}(w_{\text{ref}}) - \ell_{j+1}(w_j)] + \sum_{j<i} \frac{\eta^2}{2} \|\partial \ell_{j+1}(w_j)\|_*^2 \\ &\le D_\psi(w_{\text{ref}}, w_0) + \eta \sum_{j<i} [\ell_{j+1}(w_{\text{ref}}) - \ell_{j+1}(v_j)] + \frac{\eta^2 C_3^2}{2} \sum_{j<i} Z_{j+1}^2 \\ &\le \frac{B_w^2}{16} + \eta \sum_{j<i} [\mathcal{R}(w_{\text{ref}}) - \mathcal{R}(v_j)] + 4\eta C_3 B_w \sqrt{tC \ln(1/\delta)} + \eta^2 C_3^2 tC \\ &\le \frac{B_w^2}{8} + \eta \sum_{j<i} [\mathcal{R}(w_{\text{ref}}) - \mathcal{R}(v_j)], \end{aligned}$$

which establishes the risk guarantee for $(w_j)_{j<i}$ after substituting $(v_j)_{j<i} = (w_j)_{j<i}$ back in. To see that the projection is not invoked and in fact $v_i = w_i$, then using the bound $\mathcal{R}(w_{\text{ref}}) \le D_\psi(w_{\text{ref}}, w_0)/\sqrt{t} + \inf_v \mathcal{R}(v)$, the preceding simplifies further to give

$$\frac{1}{2}\|w_{\text{ref}} - w_i\|^2 \le D_\psi(w_{\text{ref}}, w_i) \le \frac{B_w^2}{8} + \eta \sum_{j<i} [\mathcal{R}(w_{\text{ref}}) - \mathcal{R}(v_j)] \le \frac{B_w^2}{4},$$

meaning the projection set is not exceeded, and $v_i = w_i$. ∎

B.5. Batch data: Theorem 11

This batch data proof uses a generalization bound over Bregman balls to handle concentration (Kakade et al., 2008, Theorem 3 and Example (4)).

Proof (of Theorem 11) Let $(v_i)_{i\leq t}$ denote full batch projected mirror descent iterates onto a constraint set S_ψ using initial condition $v_0 = w_0$ and the same sample as $(w_i)_{i\leq t}$. Unlike other invocations, the constraint set here is defined in terms of the Bregman divergence:

$$S_\psi := \left\{ v \in \mathbb{R}^d : D_\psi(w_{\text{ref}}, v) \leq B_w^2/2 \right\}.$$

Due to strong convexity, if follows that $v \in S_\psi$ satisfies $\|v - w_{\text{ref}}\|^2/2 \leq D_\psi(w_{\text{ref}}, v) \leq B_w^2/2$; the reason for this altered constraint set is for the application of an appropriate concentration inequality. As usual, this will be the first step, and then the risk control will come second.

The concentration inequality will in fact be a generalization bound, which allows for data re-use across iterations, unlike the martingale concentration inequalities in the other sections. It is still necessary to bound the Lipschitz constant and range of the predictors; by Theorem 12, defining $C_3 := C_1 + 2C_2B_w + B_w$ for convenience, for all $v \in S_\psi$,

$$\begin{aligned}
|\ell_j(y, x^\mathsf{T} v) - \ell_j(y, x^\mathsf{T} w_{\text{ref}})| &\leq C_3 B_w, \\
\left|\ell'(y, x^\mathsf{T} v)\right| &\leq C_3, \\
\left\|\partial \widehat{\mathcal{R}}(v_j)\right\|_* &\leq \frac{1}{n}\sum_{i<n} \|\partial \ell_i(v_j)\|_* \leq C_3.
\end{aligned}$$

Combining these bounds with the fundamental theorem of Rademacher complexity and its Lipschitz composition lemma (Shalev-Shwartz and Ben-David, 2014), as well as a Rademacher complexity bound for predictor sets of the form S_ψ (Kakade et al., 2008, Theorem 3 and Example (4)), with probability at least $1 - 4\delta$, simultaneously for every $v \in S_\psi$,

$$\begin{aligned}
\left|\widehat{\mathcal{R}}(w_{\text{ref}}) - \widehat{\mathcal{R}}(v) - \mathcal{R}(w_{\text{ref}}) + \mathcal{R}(v)\right| &\leq \text{Rad}\left(\{(\ell_1(v), \ldots, \ell_n(v) : v \in S_\psi)\}\right) + 6C_3B_w\sqrt{\frac{\ln(1/\delta)}{2n}} \\
&\leq C_3\text{Rad}\left(\{(x_1^\mathsf{T} v, \ldots, x_n^\mathsf{T} v) : v \in S_\psi)\}\right) + 6C_3B_w\sqrt{\frac{\ln(1/\delta)}{2n}} \\
&\leq \frac{C_3B_w}{\sqrt{n}}\left(1 + 6\sqrt{\ln(1/\delta)}\right).
\end{aligned}$$

The proof now proceeds by induction. The base case $w_0 = v_0 \in S_\psi$ is direct, thus consider $i > 0$. By the deterministic mirror descent guarantee applied to w_i but now with batch gradients

$\nabla\widehat{\mathcal{R}}(w_i)$, and using $(w_j)_{j<i} = (v_j)_{j<i}$, and lastly using $t \leq n$,

$$\begin{aligned}
D_\psi(w_{\text{ref}}, w_i) &\leq D_\psi(w_{\text{ref}}, w_0) + \eta \sum_{j<i} \left[\widehat{\mathcal{R}}(w_{\text{ref}}) - \widehat{\mathcal{R}}(w_j)\right] + \eta^2 \sum_{j<i} \left\|\nabla\widehat{\mathcal{R}}(w_j)\right\|_*^2 \\
&\leq \frac{B_w^2}{16} + \eta \sum_{j<i} \left[\widehat{\mathcal{R}}(w_{\text{ref}}) - \widehat{\mathcal{R}}(v_j)\right] + \eta^2 \sum_{j<i} \left\|\nabla\widehat{\mathcal{R}}(v_j)\right\|_*^2 \\
&\leq \frac{B_w^2}{16} + \eta \sum_{j<i} \left[\mathcal{R}(w_{\text{ref}}) - \mathcal{R}(v_j)\right] + \eta C_3 B_w \sqrt{t}\left(1 + 6\sqrt{\ln(1/\delta)}\right) + t\eta^2 C_3^2 \\
&\leq \frac{B_w^2}{8} + \eta \sum_{j<i} \left[\mathcal{R}(w_{\text{ref}}) - \mathcal{R}(v_j)\right],
\end{aligned}$$

which establishes the risk guarantee on $(w_j)_{j<i}$ after using $(v_j)_{j<i} = (w_j)_{j<i}$. For $w_i = v_i \in S_\psi$, continuing from the preceding inequality but additionally making use of $\mathcal{R}(w_{\text{ref}}) \leq D_\psi(w_{\text{ref}}, w_0)/\sqrt{t} + \inf_v \mathcal{R}(v)$,

$$\begin{aligned}
D_\psi(w_{\text{ref}}, w_i) &\leq \frac{B_w^2}{8} + \eta \sum_{j<i} \left[\mathcal{R}(w_{\text{ref}}) - \mathcal{R}(v_j)\right], \\
&\leq \frac{B_w^2}{8} + \eta \sum_{j<i} \frac{D_\psi(w_{\text{ref}}, w_0)}{\sqrt{t}} \\
&\leq \frac{B_w^2}{4},
\end{aligned}$$

as desired. ■

Appendix C. Analysis of Temporal Difference learning (TD)

As with the proof schemes for mirror descent, there is both a deterministic part (provided for mirror descent in Lemma 17), and a random part (using fact:conc:markov for concentration of Markov chains).

C.1. Deterministic TD analysis

Even though TD is not a gradient-based method, the analysis here follows the same expand-the-square plan as described for gradient descent in Section 2.2, which is also the idea behind the mirror descent bound in Lemma 17.

Lemma 18 *Let $S \subseteq \mathbb{R}^d$ denote an arbitrary closed convex constraint set, let $w_{\text{ref}} \in S$ be arbitrary and $w_0 \in S$, and given any vectors $(x_i)_{i\geq 0}$ with $\|x_i\| \leq 1$ and scalars $(r_i)_{i\geq 1}$ with $|r_i| \leq 1$, consider the corresponding projected TD iterates $w_{i+1} := \Pi_S\left(w_i - \eta_{i+1} G_{i+1}(w_i)\right)$ (where $G_{i+1}(\cdot)$ is defined in eq. (5)). Then, for any t,*

$$\begin{aligned}
\|w_t - w_{\text{ref}}\|^2 \leq \|w_0 - w_{\text{ref}}\|^2 + \eta \sum_{i<t} \Big[&- \langle x_i, w_i - w_{\text{ref}}\rangle^2 + \langle \gamma x_{i+1}, w_i - w_{\text{ref}}\rangle^2 \\
&- 2\langle G_{i+1}(w_{\text{ref}}), w_i - w_{\text{ref}}\rangle + 4\eta \|G_{i+1}(w_{\text{ref}})\|^2\Big].
\end{aligned}$$

Proof Proceeding just as in mirror descent, first note for any i that

$$\begin{aligned}\|w_{i+1} - w_{\text{ref}}\|^2 &= \|\Pi_S(w_i - \eta_{i+1}G_{i+1}(w_i)) - w_{\text{ref}}\|^2 \\ &\leq \|w_i - \eta_{i+1}G_{i+1}(w_i) - w_{\text{ref}}\|^2 \\ &= \|w_i - w_{\text{ref}}\|^2 - 2\eta_{i+1}\left\langle G_{i+1}(w_i), w_i - w_{\text{ref}}\right\rangle + \eta_i^2\|G_{i+1}(w_i)\|^2.\end{aligned}$$

To simplify the two latter terms, since $G_{i+1}(w_i) - G_{i+1}(w_{\text{ref}}) = x_i\left\langle x_i - \gamma x_{i+1}, w_i - w_{\text{ref}}\right\rangle$, note firstly that

$$\begin{aligned}-\left\langle G_{i+1}(w_i), w_i - w_{\text{ref}}\right\rangle &= -\left\langle G_{i+1}(w_i) - G_{i+1}(w_{\text{ref}}), w_i - w_{\text{ref}}\right\rangle - \left\langle G_{i+1}(w_{\text{ref}}), w_i - w_{\text{ref}}\right\rangle \\ &= -\left\langle x_i, w_i - w_{\text{ref}}\right\rangle\left\langle x_i - \gamma x_{i+1}, w_i - w_{\text{ref}}\right\rangle - \left\langle G_{i+1}(w_{\text{ref}}), w_i - w_{\text{ref}}\right\rangle \\ &= -\left\langle x_i, w_i - w_{\text{ref}}\right\rangle^2 + \left\langle x_i, w_i - w_{\text{ref}}\right\rangle\left\langle \gamma x_{i+1}, w_i - w_{\text{ref}}\right\rangle - \left\langle G_{i+1}(w_{\text{ref}}), w_i - w_{\text{ref}}\right\rangle,\end{aligned}$$

and secondly

$$\begin{aligned}\frac{1}{2}\|G_{i+1}(w_i)\|^2 &\leq \|G_{i+1}(w_i) - G_{i+1}(w_{\text{ref}})\|^2 + \|G_{i+1}(w_{\text{ref}})\|^2 \\ &= \|x_i\|^2\left\langle x_i - \gamma x_{i+1}, w_i - w_{\text{ref}}\right\rangle^2 + 2\|G_{i+1}(w_{\text{ref}})\|^2 \\ &\leq \left\langle x_i - \gamma x_{i+1}, w_i - w_{\text{ref}}\right\rangle^2 + 2\|G_{i+1}(w_{\text{ref}})\|^2 \\ &= \left\langle x_i, w_i - w_{\text{ref}}\right\rangle^2 - 2\left\langle x_i, w_i - w_{\text{ref}}\right\rangle\left\langle \gamma x_{i+1}, w_i - w_{\text{ref}}\right\rangle \\ &\qquad + \left\langle \gamma x_{i+1}, w_i - w_{\text{ref}}\right\rangle^2 + 2\|G_{i+1}(w_{\text{ref}})\|^2,\end{aligned}$$

which together with $2\left\langle x_i, w_i - w_{\text{ref}}\right\rangle\left\langle \gamma x_{i+1}, w_i - w_{\text{ref}}\right\rangle \leq \left\langle x_i, w_i - w_{\text{ref}}\right\rangle^2 + \left\langle \gamma x_{i+1}, w_i - w_{\text{ref}}\right\rangle^2$ give

$$\begin{aligned}&-2\left\langle G_{i+1}(w_i), w_i - w_{\text{ref}}\right\rangle + \eta\|G_{i+1}(w_i)\|^2 \\ &\quad\leq -2(1-\eta)\left\langle x_i, w_i - w_{\text{ref}}\right\rangle^2 + 2(1-2\eta)\left\langle x_i, w_i - w_{\text{ref}}\right\rangle\left\langle \gamma x_{i+1}, w_i - w_{\text{ref}}\right\rangle \\ &\qquad + 2\eta\left\langle \gamma x_{i+1}, w_i - w_{\text{ref}}\right\rangle^2 - 2\left\langle G_{i+1}(w_{\text{ref}}), w_i - w_{\text{ref}}\right\rangle + 4\eta\|G_{i+1}(w_{\text{ref}})\|^2 \\ &\quad\leq -\left\langle x_i, w_i - w_{\text{ref}}\right\rangle^2 + \left\langle \gamma x_{i+1}, w_i - w_{\text{ref}}\right\rangle^2 \\ &\qquad - 2\left\langle G_{i+1}(w_{\text{ref}}), w_i - w_{\text{ref}}\right\rangle + 4\eta\|G_{i+1}(w_{\text{ref}})\|^2.\end{aligned}$$

Combining all the inequalities so far gives

$$\begin{aligned}\|w_{i+1} - w_{\text{ref}}\|^2 - \|w_i - w_{\text{ref}}\|^2 \leq &-\eta\left\langle x_i, w_i - w_{\text{ref}}\right\rangle^2 + \eta\left\langle \gamma x_{i+1}, w_i - w_{\text{ref}}\right\rangle^2 \\ &- 2\eta\left\langle G_{i+1}(w_{\text{ref}}), w_i - w_{\text{ref}}\right\rangle + 4\eta^2\|G_{i+1}(w_{\text{ref}})\|^2,\end{aligned}$$

while applying $\sum_{i<t}$ to both sides and rearranging gives the final overall inequality. ■

C.2. Stochastic TD analysis

As mentioned in the body, the underlying Markov chain is $(x_i)_{i\leq t}$: the distribution of x_{i+1} is wholly determined by x_i. Moreover, there are additional scalars $(r_i)_{i=1}^n$, where the distribution of r_{i+1} is

also fully specified given x_i, but meanwhile r_{i+1} says nothing about x_{i+1}. Confusing matters, the TD update makes use of a triple $\zeta = (x_i, x_{i+1}, r_{i+1})$; in particular, define

$$G_\zeta(v) = x\left(\langle x - \gamma x', v\rangle - r\right) \quad \text{and} \quad G_{i+1}(v) = x\left(\langle x_i - \gamma x'_{i+1}, v\rangle - r_{i+1}\right).$$

The underlying mixing assumptions will be placed on $(\zeta_i)_{i\le t}$, not on $(x_i)_{i\le t}$; this is simply to avoid messiness arising from the use of multiple indexes. This is all smoothed over with the approximate stationarity of Definition 4. That said, there is a place in the proof where the Markov structure on $(x_i)_{i\le t}$ is needed (and explicitly mentioned).

Proof (of Theorem 9) Following the standard proof structure, let $(v_i)_{i\le t}$ denote projected TD iterates constrained to lie within $S := \{v \in \mathbb{R}^d : \|v - w_{\text{ref}}\| \le B_w\}$. The $(v_i)_{i\le t}$ are coupled to $(w_i)_{i\le t}$ in the strong sense used throughout this work: $v_0 := w_0$, and thereafter both are updated using the exact same random data sequence $(\zeta_i)_{i\le t}$. The next step of the proof will be to apply concentration inequalities to control the underlying sample space $(\zeta_i)_{i\le t}$, and then use these to show that in fact $w_i = v_i \in S$ via the deterministic TD guarantees in Lemma 18.

The concentration inequality will be the one for stochastic processes in Lemma 14, which first requires a variety of uniform bounds. For convenience, define $j := i - \tau + 1$ and a mapping f as

$$f(\zeta; v) := -\langle x, v - w_{\text{ref}}\rangle^2 + \langle \gamma x', v - w_{\text{ref}}\rangle^2 - 2\langle G_\zeta(w_{\text{ref}}), v - w_{\text{ref}}\rangle + 4\eta\|G_\zeta(w_{\text{ref}})\|^2;$$

the uniform controls will be on this mapping f, which corresponds to the f in Lemma 14, and is the stochastic term in lemma 18. The first step is to bound $|f|$, which is direct: using $\eta \le 1$ and $\gamma \le 1$,

$$\begin{aligned}
|f(\zeta; v)| &\le \|x\|^2\|v - w_{\text{ref}}\|^2 + \gamma^2\|x'\|^2\|v - w_{\text{ref}}\|^2 \\
&\qquad + \|G_\zeta(w_{\text{ref}})\|^2 + \|v - w_{\text{ref}}\|^2 + 4\eta\|x\|^2\left(\langle x - \gamma x', w_{\text{ref}}\rangle - r\right)^2 \\
&\le B_w^2 + \gamma^2 B_w^2 + B_w^2 + (1 + 4\eta)(2 + (2 + 2\gamma^2)\|w_{\text{ref}}\|^2) \\
&\le 30(B_w^2 + \|w_{\text{ref}}\|^2) \\
&\le 31 B_w^2 =: B_f,
\end{aligned}$$

To bound $|f(\zeta_{i+1}; v_i) - f(\zeta_{i+1}; v_j)|$, first note that

$$\begin{aligned}
\|v_i - v_j\| &= \left\|\sum_{k=j}^{i-1} \eta\left(G_{k+1}(v_k) - G_{k+1}(w_{\text{ref}}) + G_{k+1}(w_{\text{ref}})\right)\right\| \\
&= \sum_{k=j}^{i-1} \eta\left\|x_k\langle x_k - \gamma x_{k+1}, v_k - w_{\text{ref}}\rangle + x_k\left(\langle x_k - \gamma x_{k+1}, w_{\text{ref}}\rangle - r_{k+1}\right)\right\| \\
&\le 2\eta(\tau - 1)(1 + B_w + \|w_{\text{ref}}\|) \\
&\le 6\eta B_w =: B_G,
\end{aligned}$$

whereby

$$\begin{aligned}
|f(\zeta_{i+1}; v_i) - f(\zeta_{i+1}; v_j)| &= \Big| - \langle x_i, v_i - v_j + v_j - w_{\text{ref}}\rangle^2 + \langle x_i, v_j - w_{\text{ref}}\rangle^2 \\
&\quad + \langle \gamma x_{i+1}, v_i - v_j + v_j - w_{\text{ref}}\rangle^2 - \langle \gamma x_{i+1}, v_j - w_{\text{ref}}\rangle^2 \\
&\quad + 2\langle G_\zeta(w_{\text{ref}}), v_j - v_i\rangle \Big| \\
&= \Big| - \langle x_i, v_i - v_j\rangle^2 - 2\langle x_i, v_i - v_j\rangle \langle x_i, v_j - w_{\text{ref}}\rangle \\
&\quad + \gamma^2 \langle x_{i+1}, v_i - v_j\rangle^2 + 2\gamma^2 \langle x_{i+1}, v_i - v_j\rangle \langle x_{i+1}, v_j - w_{\text{ref}}\rangle \\
&\quad + 2\langle x_i, v_j - v_i\rangle \langle x_i - \gamma x_{i+1}, w_{\text{ref}}\rangle - 2r_{i+1}\langle x_i, v_j - v_i\rangle \Big| \\
&\le B_G^2 + 2B_w B_G + \gamma^2 B_G^2 + 2\gamma^2 B_G B_w + 2(1+\gamma) B_G \|w_{\text{ref}}\| + 2B_G \\
&\le B_G\left(2 + 4\|w_{\text{ref}}\| + 4B_w\right) + 2B_G^2 \\
&\le 42\eta B_w^2 + 36\eta^2 B_w^2 \\
&\le 80\eta B_w^2 =: B_i.
\end{aligned}$$

Applying Lemma 14 to each $i \le t$ and union bounding, then with probability at least $1 - t\delta$, simultaneously for every $i \le t$,

$$\begin{aligned}
\eta \sum_{j<i} \left[f(\zeta_{j+1}; v_j) - \text{E}_{\zeta\sim\pi} f(\zeta; v_j)\right] &\le 2\eta B_f\left(2\tau - 2 + t\epsilon + \sqrt{t\tau \ln(1/\delta)}\right) + \eta \sum_{i<t} B_i \\
&\le \eta B_w^2\left(62(2\tau - 2 + t\epsilon) + 80t\eta + \sqrt{2t\tau\ln(1/\delta)}\right) \\
&\le \frac{B_w^2}{1024}(124 + 80 + 2) \le \frac{B_w^2}{4}.
\end{aligned}$$

Henceforth, condition away the failure event for the preceding inequalities.

What remains is to inductively invoke the deterministic TD guarantee from Lemma 18 to bound the error of $(w_i)_{i\le t}$ and simultaneously show $w_i = v_i \in S$ for all i. The base case

Throw out the preceding failure event; the remainder of the proof proceeds by induction, establishing that the iterate sequence never exits S, The proof now proceeds by induction; the claim holds automatically for v_0, since $v_0 = w_0 \in S$ by construction, thus consider w_i for some $i > 0$. Invoking the deterministic TD guarantee from Lemma 18 to w_i, together with the inductive hypothesis $(w_j)_{j<i} = (v_j)_{j<i}$, the earlier concentration inequalities, and lastly making the single appeal

to the Markov property on $(x_i)_{i\leq t}$ to obtain $\mathrm{E}_{\mathrm{X}_{\zeta_{i+1}}} x_{i+1} = \mathrm{E}_{\mathrm{X}_{\zeta_{i+1}}} x_i$, note

$$\begin{aligned}
\|\tilde{w}_i - w_{\text{ref}}\|^2 &\leq \|w_0 - w_{\text{ref}}\|^2 + \eta \sum_{j<i} \Big[- \langle x_j, w_j - w_{\text{ref}}\rangle^2 + \langle \gamma x_{j+1}, w_j - w_{\text{ref}}\rangle^2 \\
&\qquad - 2\langle G_j(w_{\text{ref}}), w_j - w_{\text{ref}}\rangle + 4\eta \|G_j(w_{\text{ref}})\|^2 \Big] \\
&\leq \|w_0 - w_{\text{ref}}\|^2 + \eta \sum_{j<i} \Big[- \langle x_j, v_j - w_{\text{ref}}\rangle^2 + \langle \gamma x_{j+1}, v_j - w_{\text{ref}}\rangle^2 \\
&\qquad - 2\langle G_j(w_{\text{ref}}), v_j - w_{\text{ref}}\rangle + 4\eta \|G_j(w_{\text{ref}})\|^2 \Big] \\
&\leq \|w_0 - w_{\text{ref}}\|^2 + \frac{B_w^2}{4} + \eta \sum_{j<i} \mathrm{E}_{\mathrm{X}_{\zeta\sim\pi}} \Big[- \langle x, v_j - w_{\text{ref}}\rangle^2 + \langle \gamma x, v_j - w_{\text{ref}}\rangle^2 \\
&\qquad - 2\langle G_\zeta(w_{\text{ref}}), v_j - w_{\text{ref}}\rangle + 4\eta(2 + 2(1+\gamma^2)\|w_{\text{ref}}\|^2) \Big] \\
&\leq \|w_0 - w_{\text{ref}}\|^2 + \frac{B_w^2}{2} \\
&\qquad + \eta \sum_{j<i} \mathrm{E}_{\mathrm{X}_{\zeta\sim\pi}} \Big[- (1-\gamma^2) \langle x, w_j - w_{\text{ref}}\rangle^2 - 2\langle G_\zeta(w_{\text{ref}}), w_j - w_{\text{ref}}\rangle \Big], \quad (12)
\end{aligned}$$

which rearranges to give the desired TD error bound. To control the norms from here, since $\|\mathrm{E}_{\mathrm{X}_{\zeta\sim\pi}} G_\zeta(w_{\text{ref}})\| \leq \frac{\|w_{\text{ref}} - w_0\|^2}{\sqrt{t}}$, and moreover since $-(1-\gamma^2)\langle x, v_i - w_{\text{ref}}\rangle^2$ is negative and can be dropped, then eq. (12) implies

$$\begin{aligned}
\|w_i - w_{\text{ref}}\|^2 &\leq \|v_0 - w_{\text{ref}}\|^2 + \frac{B_w^2}{2} + 2\eta \sum_{j<i} \|G_\pi(w_{\text{ref}})\| \, \|v_j - w_{\text{ref}}\| \\
&\leq \frac{B_w^2}{16} + \frac{B_w^2}{2} + \frac{B_w}{8} \\
&< B_w^2,
\end{aligned}$$

meaning the new unconstrained iterate w_i satisfies $w_i \in S$, whereby v_i will also not encounter the constraint since $v_{i-1} = w_{i-1}$ and the update is the same, and thus $w_i = v_i \in S$. ■