Journal of Machine Learning Research 26 (2025) 1-68 Submitted 2/25; Revised 7/25; Published 11/25

Gradient Equilibrium in Online Learning: Theory and Applications

Anastasios N. Angelopoulos ANGELOPOULOS@BERKELEY.EDU
University of California, Berkeley

Michael I. Jordan JORDAN@CS.BERKELEY.EDU
University of California, Berkeley
Inria, Paris

Ryan J. Tibshirani RYANTIBS@BERKELEY.EDU
University of California, Berkeley

Editor: Zhengyuan Zhou

Abstract

We present a new perspective on online learning that we refer to as *gradient equilibrium*: a sequence of iterates achieves gradient equilibrium if the average of gradients of losses along the sequence converges to zero. In general, this condition is not implied by, nor implies, sublinear regret. It turns out that gradient equilibrium is achievable by standard online learning methods such as gradient descent and mirror descent with *constant* step sizes (rather than decaying step sizes, as is usually required for no regret). Further, as we show through examples, gradient equilibrium translates into an interpretable and meaningful property in online prediction problems spanning regression, classification, quantile estimation, and others. Notably, we show that the gradient equilibrium framework can be used to develop a debiasing scheme for black-box predictions under arbitrary distribution shift, based on simple post hoc online descent updates. We also show that post hoc gradient updates can be used to calibrate predicted quantiles under distribution shift, and that the framework leads to unbiased Elo scores for pairwise preference prediction.

Keywords: online learning, gradient equilibrium, distribution shift, calibration, debiasing

1. Introduction

Online learning is a powerful paradigm for the analysis of sequential data, with applications in diverse areas such as forecasting, calibration, and control. Distinct from classical methods in statistics for analyzing time series and stochastic processes, the online learning framework makes no stochastic or generative assumptions about the sequence of data points. That is, the guarantees from typical online learning analyses apply to an individual sequence of data points that can be entirely arbitrary. As such, online learning offers algorithmic and analytic strategies to cope with arbitrary distribution drift and adversarial behavior—issues which are central in real deployments of machine learning systems.

Most online learning analyses focus on *regret* as a key metric—given an algorithm that is presented with one data point at a time, and must produce predictions accordingly, how small is the total loss relative to an oracle who can wait until the end, seeing the entire batch of data, and outputting the single best prediction? Although regret provides a theoretical

handle on the challenging problem of optimizing sums of losses with respect to arbitrary sequences, it is not always apparent how regret aligns with various goals of learning. It is often the case that certain properties of a predictive algorithm are of interest, and not just its loss relative to a fixed reference; for example, we might seek to understand its bias, coverage, or other statistical properties. While extensions of regret may be able to target such behaviors, they can become complex.

In the current paper, we offer a simple, alternative perspective on online learning which directly targets various properties that we might want an online algorithm to exhibit. Such properties tend to have a statistical flavor, but in the spirit of online learning, we achieve them without making any stochastic assumptions. For example, our framework can guarantee a set of predictions is unbiased over a completely arbitrary sequence of data points; here, bias refers to the difference between the average prediction and average response over the sequence, not in expectation. We refer to our online learning framework as *gradient equilibrium*.

Next we introduce the basic idea of gradient equilibrium, and we provide three examples of applications that highlight the simplicity and scalability of gradient equilibrium. In Section 1.2, we return to a discussion of regret, demonstrating that our framework is not the same as regret—sublinear regret does not subsume gradient equilibrium nor vice versa. In Section 2, we provide comprehensive theory on gradient equilibrium in online gradient descent. Section 3 extends the analysis to incorporate regularization. Although our main focus is constant step sizes, in Section 4 we treat the case of arbitrary step sizes. Finally, in Section 5, we give several examples of applications, and in Section 6, our conclusions.

1.1 Gradient equilibrium

To fix notation, let ℓ_t, $t = 1, 2, 3, \ldots$ be a sequence of loss functions, with each $\ell_t : \mathbb{R}^d \to (-\infty, \infty]$ assumed to be finite and subdifferentiable, in a generalized sense made precise below, on a common domain $D \subseteq \mathbb{R}^d$. Each loss ℓ_t will typically depend on a data point $(x_t, y_t) \in \mathcal{X} \times \mathcal{Y}$, e.g., we could have $\ell_t(\theta) = \frac{1}{2}(y_t - x_t^T\theta)^2$ with $\mathcal{X} = \mathbb{R}^d$ and $\mathcal{Y} = \mathbb{R}$, though we suppress this dependence for notational simplicity.

Let θ_t, $t = 1, 2, 3, \ldots$ denote a sequence of iterates produced by some online optimization scheme applied to ℓ_t, $t = 1, 2, 3, \ldots$, such as gradient descent. Next we define our main condition of study.

Definition 1 *A sequence of iterates $\theta_t \in D \subseteq \mathbb{R}^d$, $t = 1, 2, 3, \ldots$ satisfies* gradient equilibrium *with respect to a sequence of loss functions ℓ_t, $t = 1, 2, 3, \ldots$ provided that*

$$\frac{1}{T}\sum_{t=1}^{T} g_t(\theta_t) \to 0, \quad \text{as } T \to \infty, \tag{1}$$

where each $g_t(\theta_t)$ is a subgradient of ℓ_t at θ_t.

In this definition and in general, we use $g_t(\theta) \in \mathbb{R}^d$ to denote a generalized subgradient of ℓ_t at $\theta \in D$ as defined in Appendix D. For our purposes in what follows, the precise definition of a generalized subgradient is not crucial, and the most important point to convey is that this generalization fluidly encapsulates both subgradients of convex functions, and gradients of differentiable (and possibly nonconvex) functions. We will often call $g_t(\theta)$ a "gradient," for simplicity, even though it technically denotes a generalized subgradient.

Gradient equilibrium, simply put, says that the average of gradients along the iterate sequence tends to zero. This condition turns out to be interpretable in various problem settings: for example, for squared losses, it translates into a sequential notion of unbiasedness; for quantile losses, it reduces to one-sided coverage; for generalized linear model (GLM) losses, it becomes uncorrelatedness of the GLM residuals and the covariates. Generally, one can understand these conditions as sequential analogs of the first-order optimality condition in M-estimation problems. Table 1 gives a summary. In Section 2.3, we will examine these and other examples in more detail, also including a derivation of the equilibrium conditions.

Loss	Equilibrium condition	Interpretation
$\ell_t(\theta) = \frac{1}{2}(y_t - \theta)^2$	$\frac{1}{T}\sum_{t=1}^T \theta_t \asymp \frac{1}{T}\sum_{t=1}^T y_t$	θ_t is unbiased for y_t
$\ell_t(\theta) = \rho_\tau(y_t - \theta)$	$\frac{1}{T}\sum_{t=1}^T 1\{y_t \le \theta_t\} \asymp \tau$	θ_t lies above y_t with frequency τ
$\ell_t(\theta) = -y_t x_t^\mathsf{T}\theta + \psi(x_t^\mathsf{T}\theta)$	$\frac{1}{T}\sum_{t=1}^T (\psi'(x_t^\mathsf{T}\theta_t) - y_t)x_t \asymp 0$	GLM residuals from θ_t are uncorrelated with features

Table 1: *Examples of loss functions and gradient equilibrium conditions. For sequences a_T and b_T, we write $a_T \asymp b_T$ to mean $a_T - b_T \to 0$ as $T \to \infty$. In the second row, ρ_τ denotes the quantile loss at level $\tau \in [0,1]$, i.e., $\rho_\tau(u) = \tau|u|$ for $u \ge 0$ and $\rho_\tau(u) = (1-\tau)|u|$ for $u < 0$. In the third row, ψ is the cumulant generating function for the GLM (e.g., $\psi(u) = \frac{1}{2}u^2$ for linear regression, $\psi(u) = \log(1+e^u)$ for logistic regression, and $\psi(u) = e^u$ for Poisson regression), and ψ' is its derivative.*

To further motivate our study, we briefly review a few applications of gradient equilibrium conditions in machine learning. An important application is the debiasing of a black-box prediction model under arbitrary distribution shift, a problem of significant interest across a wide range of machine learning deployments. To give a flavor of how this works, and how simple it is, suppose that $f_t(x_t)$ is a prediction of a response y_t at t. We form an adjusted prediction via

$$\tilde{f}_t(x_t) = f_t(x_t) + \theta_t, \quad \text{where} \quad \theta_t = \theta_{t-1} + \eta(y_{t-1} - f_{t-1}(x_{t-1}) - \theta_{t-1}).$$

Here $\eta > 0$ is a step size (i.e., learning rate). The parameter θ_t can be seen as the result of an online gradient descent update applied to the loss $\ell_{t-1}(\theta) = \frac{1}{2}(y_{t-1} - f_{t-1}(x_{t-1}) - \theta)^2$. Gradient equilibrium translates into

$$\frac{1}{T}\sum_{t=1}^T \tilde{f}_t(x_t) - \frac{1}{T}\sum_{t=1}^T y_t \to 0, \quad \text{as } T \to \infty,$$

the unbiasedness of the adjusted predictions $\tilde{f}_t(x_t)$, $t = 1, 2, 3, \ldots$ along the sequence. The same idea can be applied to carry out debiasing based on features, which may or may not be a part of the feature x_t used by the original predictor f_t. For example, we can achieve multigroup (i.e., groupwise) debiasing guarantees by setting ℓ_{t-1} to be the loss which regresses

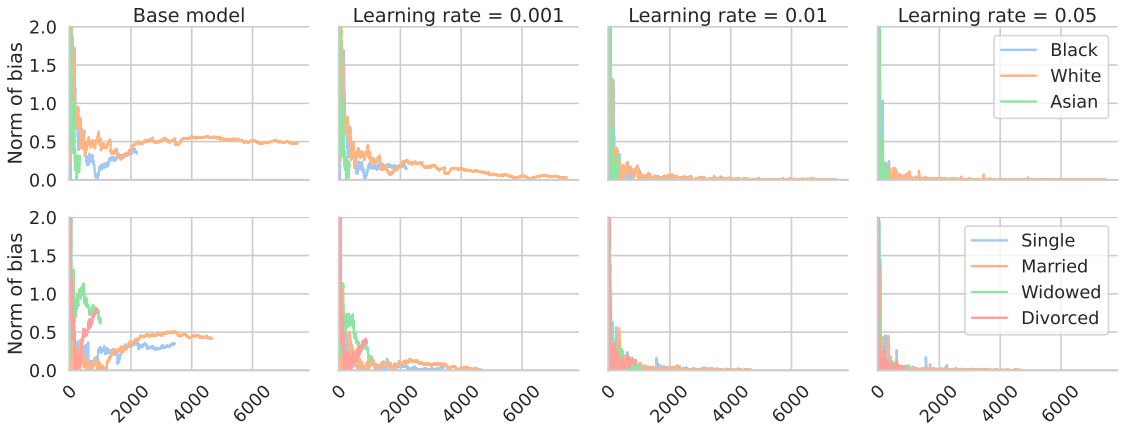


Figure 1: *Multigroup debiasing results on the MIMIC dataset, on predicting length-of-stay of patients in a hospital system in Boston, Massachusetts. We train an XGBoost model on a large number of features from this dataset and run our multigroup debiasing procedure with respect to ethnicity (top row) and marital status (bottom row), with each column showing a different learning rate. Gradient equilibrium (third row of Table 1 where the features are group indicators) for this problem says that we achieve zero bias for each ethnicity and marital status, in the long run.*

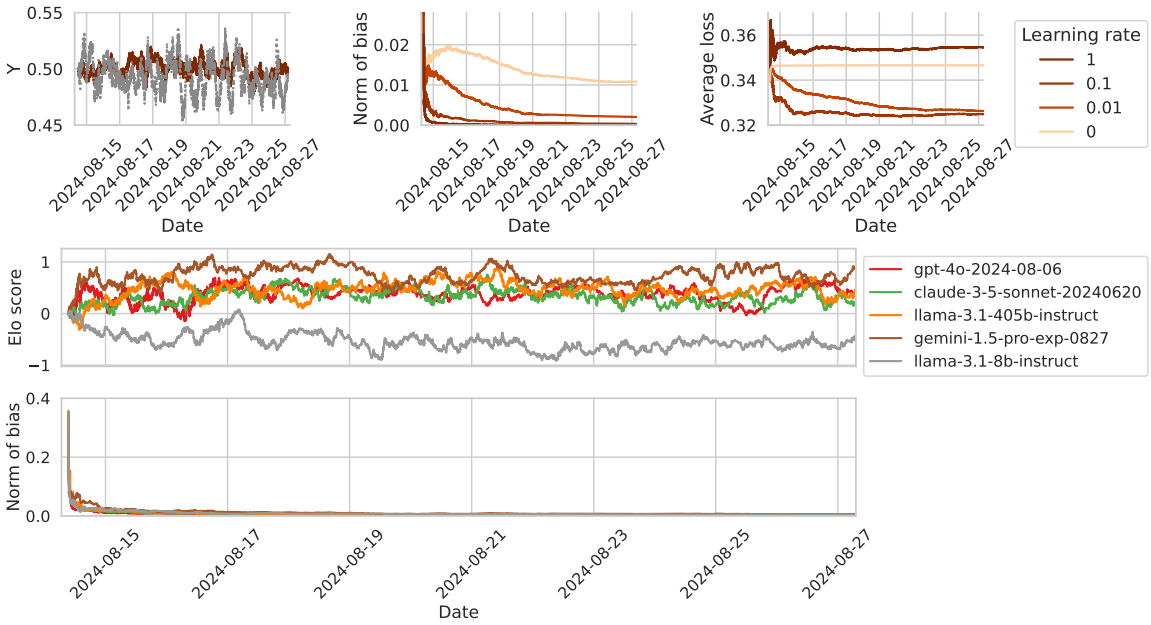


Figure 2: *Results on the Chatbot Arena dataset, comprised of predictions of human preferences between pairs of large language models (LLMs). The top row shows statistics of the predictions over time, with different learning rates. The left plot shows a rolling average of the predicted win rate along with a rolling average of the ground-truth labels in gray over time. The middle plot gives the absolute bias of the predictions over the sequence. On the right is the average binary cross-entropy loss of predictions. The bottom two plots show the sequence of Elo scores, and the bias of the Elo win-rate predictions on a per-model level. Gradient equilibrium for this problem says that we achieve zero bias per model, in the long run.*

the residual $y_{t-1} - f_{t-1}(x_{t-1})$ onto a feature that indicates group membership. Figure 1 showcases multigroup debiasing on a medical length-of-stay prediction dataset. This setting is studied formally in Section 2.3 and empirically in Sections 5.2 and 5.3.

A second example we highlight is the learning of Elo scores for pairwise preference prediction. Elo scores are a widely-used rating system for competitive two-player games, like chess. Given M competitors, we first initialize a coefficient vector of all zeros, $\theta_1 = 0 \in \mathbb{R}^d$. During a sequence $t = 1, 2, 3, \ldots$ of battles between two distinct competitors $a_t, b_t \in \{1, \ldots, M\}$, we define $y_t = 1$ if b_t wins and $y_t = 0$ otherwise. After the battle at each t, we perform the following update (abbreviating $a = a_t$ and $b = b_t$):

$$\theta_{ta} = \theta_{t-1,a} - \eta(y_t - p_t)$$
$$\theta_{tb} = \theta_{t-1,b} - \eta(p_t - y_t).$$

Here $\eta > 0$ is a learning rate, and $p_t = \sigma(\theta_{t-1,b} - \theta_{t-1,a})$, where $\sigma(x) = e^x/(1 + e^x)$ is the sigmoid function. All other coefficients are carried forward, $\theta_{tm} = \theta_{t-1,m}$ for all $m \neq a, b$. Gradient equilibrum translates into an unbiasedness guarantee for this algorithm: namely, for each player $m \in \{1, \ldots, M\}$,

$$\frac{1}{|I_m|} \sum_{t \in I_m} p_t^m - \frac{1}{|I_m|} \sum_{t \in I_m} y_t^m \to 0, \quad \text{as } T \to \infty, \text{ where } I_m = \{t \leq T : a_t = m \text{ or } b_t = m\},$$

where p_t^m, y_y^m are the win-rate prediction and win indicator, respectively, for m (equal to p_t and y_t if $b_t = m$, or $1 - p_t$ and $1 - y_t$ if $a_t = m$). That is, for every player, the win-rate predictions are unbiased in the long run, over the sequence. Figure 2 showcases this capability in the domain of pairwise human preference evaluations of large language models (LLMs). We return to a more detailed discussion of this problem in Section 5.5.

A third and last application is quantile calibration: we can use simple post hoc online gradient updates (analogous to those described above for debiasing) to adjust predicted quantiles, so that they achieve finite-sample coverage guarantees under arbitrary distribution shift. This was already studied in Gibbs and Candès (2021) and Angelopoulos et al. (2023), albeit not under the framework of gradient equilibrium. We show formally how this line of work relates to gradient equilibrium in Section 2.3, and revisit it through empirical examples in Section 5.4.

1.2 Comparison to regret

Regret (including its various generalizations) is the most common metric used to analyze algorithms in the literature on online learning. See the related work discussion in Section 1.3. The *regret* associated with the sequence θ_t, $t = 1, 2, 3, \ldots$ at iteration T is defined by:

$$\text{Regret}_T = \sum_{t=1}^{T} \ell_t(\theta_t) - \inf_{\theta} \sum_{t=1}^{T} \ell_t(\theta).$$

This measures the difference in loss incurred by the iterate sequence to the loss of the best fixed parameter choice "in hindsight." In this paper, we make use of the following definition.

Definition 2 *A sequence of iterates* θ_t, $t = 1, 2, 3, \ldots$ *satisfies* no regret *with respect to a sequence of loss functions* ℓ_t, $t = 1, 2, 3, \ldots$ *provided that*

$$\frac{\text{Regret}_T}{T} = \frac{1}{T}\sum_{t=1}^{T} \ell_t(\theta_t) - \inf_{\theta} \frac{1}{T}\sum_{t=1}^{T} \ell_t(\theta) \to 0, \quad \text{as } T \to \infty. \tag{2}$$

The condition in (2) is usually called "sublinear regret" in the online learning literature, as much of this literature is focused on quantifying the precise growth rate of Regret_T as a function of T for certain online algorithms under certain conditions on the sequence ℓ_t (e.g., $\sqrt{T}$ for gradient descent on convex functions, or $\log T$ for strongly convex functions, and so on). We refer to condition (2) as "no regret" for simplicity.

How do regret and gradient equilibrium compare? To glean some insight, Figure 3 displays the result of a simple experiment with squared losses, $\ell_t(\theta) = \frac{1}{2}(y_t - \theta)^2$, $t = 1, 2, 3, \ldots$, where the data y_t, $t = 1, 2, 3, \ldots$ are i.i.d. standard Gaussian samples. We use gradient descent, initialized at $\theta_1 = 0$, to define the sequence θ_t, $t = 2, 3, \ldots$. We use a constant step size, $\eta = 0.2$, which is not typical for online optimization, and does not imply a no-regret guarantee in general; for strongly convex functions like the squared losses in the current setting, a decaying step size schedule such as $\eta_t = 1/(\mu t)$ would be typical (where μ is the strong convexity constant), since this would lead to a sharp no-regret guarantee.

The iterate sequence from gradient descent here does not appear to have no regret, as we can see in the left panel of Figure 3 that Regret_T/T does not vanish as $T \to \infty$. However, something interesting happens to the magnitude of the average bias, $|\frac{1}{T}\sum_{t=1}^{T}(\theta_t - y_t)|$, plotted in the right panel of the figure: this is driven to zero as T grows. What this shows, in fact, is that along the gradient descent sequence the average of the gradients $\frac{1}{T}\sum_{t=1}^{T} g_t(\theta_t)$ can be made small even if the regret is large.

It is not hard to see that, in general, gradient equilibrium (1) does not imply no regret (2). For example, the iterates θ_t, $t = 1, 2, 3, \ldots$ could "bounce around" the minimizer of $\bar{\ell}_T = \frac{1}{T}\sum_{t=1}^{T} \ell_t$ in such a way that the average gradient tends to zero but the average loss is not minimal; just as in Figure 4b below. Further, if $\bar{\ell}_T$ is nonconvex, then it could have stationary points (with zero gradient) that are strict local minima or strict saddle points,

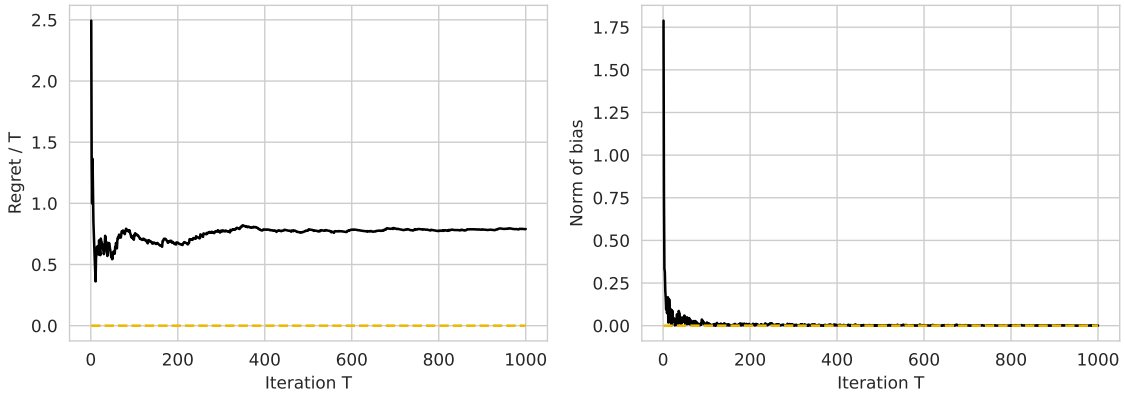


Figure 3: *Regret and bias for gradient descent on squared losses, with constant step sizes.*

and θ_t could be bouncing around (or even converging to) such points, and in doing so it could satisfy gradient equilibrium without attaining anywhere close to minimal average loss. Conversely, no regret does not imply gradient equilibrium in general, as we study next through examples.

1.2.1 Two illustrative examples

To help elucidate the relationship between no regret (NR) and gradient equilibrium (GEQ), we work through two illustrative examples.

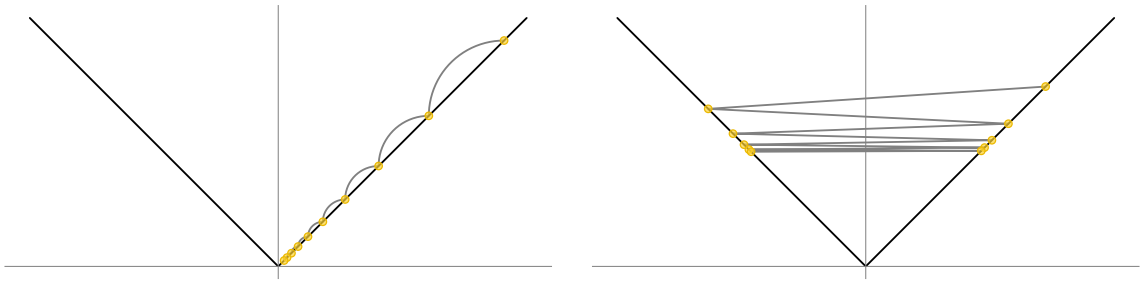
Example 1: absolute value losses. For the first example, let $\ell_t(\theta) = |\theta|$, for all $t = 1, 2, 3, \ldots$. To see that NR does not imply GEQ, consider any sequence such that $\theta_t > 0$ for all t, and $\theta_t \to 0$. Then we have $\text{Regret}_T/T \to 0$ but $\frac{1}{T}\sum_{t=1}^T g_t(\theta_t) = 1$ for all T, thus GEQ does not hold. This is depicted in Figure 4a.

For the opposite direction, to see that GEQ does not imply NR, consider any sequence with $\theta_t > 0$ for odd t, and $\theta_t < 0$ for even t. (This behavior of "bouncing around" the minimum was discussed earlier, and the current setting provides a concrete picture.) Then $\frac{1}{T}|\sum_{t=1}^T g_t(\theta_t)| \leq \frac{1}{T} \to 0$, but as long as $|\theta_t|$ remains bounded away from zero, NR does not hold. This is depicted in Figure 4b.

Example 2: squared losses. After the last example, illustrated in Figure 4a, it is natural to ask whether this kind of behavior, where iterates can satisfy NR but not GEQ, is limited to settings where the gradients are nonsmooth, as in the absolute value loss. The answer is "no": as the next example shows, a sequence of iterates can still satisfy NR for smooth and even strongly convex loss functions, without attaining GEQ. Let $\ell_t(\theta) = \frac{1}{2}(y_t - \theta)^2$, for $t = 1, 2, 3, \ldots$, where

$$(y_1, y_2, y_3, \ldots) = (\underbrace{a, a, \ldots, a}_{n \text{ times}}, \underbrace{b, b, \ldots, b}_{m \text{ times}}, \underbrace{a, a, \ldots, a}_{n \text{ times}}, \underbrace{b, b, \ldots, b}_{m \text{ times}}, \ldots),$$

that is, a repeating pattern of a for n elements, b for m elements, a for n elements, and so on. Suppose that $a < 0 < b$ and $n > m$ are such that $na + mb = 0$. Suppose that T is a



(a) NR $\not\Rightarrow$ GEQ (b) GEQ $\not\Rightarrow$ NR

Figure 4: *Two examples with $\ell_t(\theta) = |\theta|$ which show that neither NR nor GEQ necessarily implies the other. In each panel, the iterates start at the upper right-most point, and the thin gray lines are simply as visual aid to demonstrate the order of the sequence.*

multiple of $n+m$. Then

$$\inf_\theta \frac{1}{T}\sum_{t=1}^{T}(y_t-\theta)^2 = \frac{na^2+mb^2}{n+m} = \alpha a^2+\beta b^2,$$

where $\alpha = n/(n+m)$ and $\beta = m/(n+m)$. Meanwhile, define

$$(\theta_1,\theta_2,\theta_3,\dots) = (\underbrace{u,u,\dots,u}_{n \text{ times}},\underbrace{v,v,\dots,v}_{m \text{ times}},\underbrace{u,u,\dots,u}_{n \text{ times}},\underbrace{v,v,\dots,v}_{n \text{ times}},\dots),$$

that is, a repeating pattern of u for n elements, v for m elements, u for n elements, and so on. The basic idea is to choose u, v in such a way that average loss of the sequence θ_t equals $\alpha a^2+\beta b^2$, but the average gradient remains positive and bounded away from zero (possible due to the square growth of the loss functions versus the linear growth of their gradients). To this end, we set $u=a$ and then choose $v>b$ by equating $\alpha a^2+\beta b^2$ and $\frac{1}{T}\sum_{t=1}^{T}(y_t-\theta_t)^2 = \beta(b-v)^2$, which results in $v = b+\sqrt{\alpha a^2/\beta + b^2}$. This makes the average gradient:

$$\frac{1}{T}\sum_{t=1}^{T}\theta_t = \alpha a+\beta v = \sqrt{\alpha\beta a^2+\beta^2 b^2}.$$

Hence, we have constructed a sequence with zero regret but with average gradient, or equivalently average bias, equal to $\sqrt{\alpha\beta a^2+\beta^2 b^2}$, for arbitrarily large T.

Zooming out from this particular construction, our next result shows that for *any* data sequence y_t with nontrivial sample variance, we can always find an iterate sequence θ_t that has zero regret and average bias bounded away from zero. The proof is very simple, and deferred until Appendix A.1.

Proposition 3 *For any sequence y_t, $t=1,\dots,T$, denote its sample mean and variance by $\bar{y}_T = \frac{1}{T}\sum_{t=1}^{T} y_t$ and $s_T^2 = \frac{1}{T}\sum_{t=1}^{T}(y_t-\bar{y}_T)^2$. There exists a sequence θ_t, $t=1,\dots,T$ such that*

$$\frac{1}{T}\sum_{t=1}^{T}(y_t-\theta_t)^2 = s_T^2, \quad \text{and} \quad \left|\frac{1}{T}\sum_{t=1}^{T}\theta_t - \frac{1}{T}\sum_{t=1}^{T}y_t\right| \geq s_T.$$

1.2.2 The broader picture

What did we learn from the last two examples? The main takeaway:

> *No regret does not imply gradient equilibrium, even for smooth and strongly convex functions.*

This is an important point, and—together with the converse fact, that gradient equilibrium does not imply no regret, even with strong assumptions on the loss functions in question—supports the study of gradient equilibrium as a standalone property of interest in online learning, beyond no regret.

A careful look back at the last subsection actually suggests a second takeaway:

> *Iterate convergence does not imply gradient equilibrium if the gradients are nonsmooth.*

Here by *iterate convergence* we mean convergence of the iterate sequence to a minimizer of the average loss:

$$\inf\left\{\|\theta_T - \theta_T^\star\|_2 : \theta_T^\star \text{ minimizes } \bar{\ell}_T = \frac{1}{T}\sum_{t=1}^{T} \ell_t\right\} \to 0, \quad \text{as } T \to \infty. \tag{3}$$

This is a very strong condition, especially in the online setting, and as such it is not typically studied in the online optimization literature. The point of raising it here is not to suggest that it is interesting, but instead to emphasize that even under a strong condition on the iterates such as (3), the GEQ property (1) is not guaranteed to hold if the gradients are nonsmooth, as the example in Figure 4a demonstrates.[1]

The discussion of regret-type properties in relation to gradient equilibrium is continued in Appendix C, where we introduce a condition we call *no move regret*, which can be viewed as a conceptual stepping stone between NR and GEQ.

1.3 Related work

Online learning studies algorithms that process data sequentially, aiming to maximize a cumulative measure of performance while making no assumptions about putative mechanisms that generated the sequence. We begin by overviewing the historical roots of the adversarial sequence model, which underlie both the classical regret-based perspective on online learning and our gradient equilibrium perspective.

1.3.1 Historical roots

The study of algorithms for analyzing arbitrary sequences under no statistical assumptions arose historically in separate threads in various fields over several decades. One important thread was game theory. Blackwell, inspired by von Neumann's minimax theorem for two-player, zero-sum games, asked whether a similar result could be developed for vector-valued payoffs (Blackwell, 1956). After realizing that no such theorem could be obtained for single-move games, he asked instead what could be achieved in a repeated, non-zero-sum game, where the e evaluation of performance was expressed in terms of a time-averaged vector-valued payoff. Specifically, one player was conceived of as trying to force a time-averaged payoff to approach a given set, while the competing player was viewed as trying to prevent the time-averaged payoff from approaching that set. Blackwell derived conditions under which such "approachability" can be achieved, and developed a simple randomized algorithm that a player could employ to ensure approachability, even in the face of an adversarial sequence of actions from the opposing player. This paper developed time-averaged performance measures, adversarial sequences, and vector-valued payoffs, all elements which have formed the basis of various further developments in game theory and (subsequently) online learning. Concurrent work by Hannan, on repeated games, further refined the notion of a time-averaged performance measure, by introducing the "regret" of a player via a comparison to an oracle who could see the entire sequence of play (Hannan, 1957). Recent work has established

1. It is not hard to see that some level of nonsmoothness is needed for iterate convergence to fail to imply gradient equilibrium. If the loss functions ℓ_t are smooth enough such that, say, $\nabla\bar{\ell}_T$ is L-Lipschitz (where the constant $L > 0$ does not depend on T), then we have $\|\nabla\bar{\ell}_T(\theta_T)\|_2 = \|\nabla\bar{\ell}_T(\theta_T) - \nabla\bar{\ell}_T(\theta_T^\star)\|_2 \le L\|\theta_T - \theta_T^\star\|_2 \to 0$.

the equivalence of the approachability framework and the regret-minimizing framework (Abernethy et al., 2011).

A different thread came from information theory and finance, as methods for investment decisions and compression were developed that eschewed traditional stochastic assumptions, and ideas such as hedging and loss minimization began to take shape (Kelly, 1956; Ziv and Lempel, 1978; Cover, 1991).

Online learning initially developed independently of this earlier literature; it was focused on computational issues and on specific gradient-based and multiplicative algorithms for prediction and decision-making (Vovk, 1990; Littlestone and Warmuth, 1994; Cesa-Bianchi et al., 1997). The performance measures were of a statistical flavor, with the canonical examples being squared error and zero-one loss, but the focus was on theoretical guarantees on predicting labels or selecting models that were obtained without making classical statistical assumptions. Given this stance, connections to the earlier literature on arbitrary, possibly adversarial data soon became apparent. In fact, when Cesa-Bianchi and Lugosi wrote the first comprehensive book on online learning (Cesa-Bianchi and Lugosi, 2006), connections to Blackwell, Cover, Hannan, and Kelly were explicit and well-developed.

An additional crucial ingredient in the melting pot of online learning was optimization theory, specifically the branch of optimization theory where the focus was iterative, gradient-based algorithms (Kiefer and Wolfowitz, 1952; Nemirovski and Yudin, 1985). An important unifying step was made by Zinkevich, who considered optimization problems involving sums of convex functions, doing so within the regret perspective of online learning (Zinkevich, 2003). This opened the door to many further connections. Indeed, it turned out that the algorithms studied in gradient-based optimization theory—most notably, the mirror descent algorithm of Nemirovski and Yudin—not only exhibit favorable convergence in optimization-theoretic analysis but also have favorable regret guarantees in an online-learning analysis. Thus, recent textbooks have emphasized the powerful connections between optimization and online learning (Hazan, 2016; Orabona, 2019). The story has been enriched by regularization; the two main algorithmic success stories in online learning—mirror descent and follow the regularized leader (McMahan et al., 2013)—both make essential use of regularized updates, which are needed to handle the adversarial nature of the problem.

Further connections emerged from statistics and economics in the form of calibration (Foster and Vohra, 1998). Calibration can be viewed as a weak but useful desideratum for a learning algorithm. A well-known example is weather prediction: when a meteorologist says "the chance of rain today is 70%," how do we evaluate such a prediction given that it either rains or it doesn't? A standard calibration method would define a grid on the interval $(0, 1)$, consider the subset of days in which the prediction falls in a particular bin, and ask that the empirical proportion of rainy days for that subset is close to the number at the center of the bin (say 0.7). It is relatively straightforward to calibrate predictions in batch mode, at the end of a sequence of predictions, but it is more useful in many problems to calibrate as the data arrive. Moreover, the goal is generally to calibrate for the particular sequence at hand. This motivates studying calibration within an online, regret-based framework. An example of such a framework was presented by Foster (Foster, 1999), based on a connection to approachability.

1.3.2 Recent developments

A closely related line of work to this manuscript is the literature on online conformal prediction, introduced in Gibbs and Candès (2021). Inspired by the literature on conformal prediction (Vovk et al., 2005), in online conformal we seek to construct prediction sets which satisfy a coverage guarantee on average over the sequence: $\frac{1}{T}\sum_{t=1}^{T} 1\{y_t \in C_t\} \to 1-\alpha$, as $T \to \infty$, for a coverage level $1-\alpha$, where y_t, $t = 1, 2, 3, \ldots$. It is well-known that the adaptive conformal inference algorithm of Gibbs and Candès (2021) is an instance of online gradient descent with respect to the quantile loss; many papers have built on this to derive new results on coverage (Zaffran et al., 2022; Feldman et al., 2023; Bhatnagar et al., 2023; Angelopoulos et al., 2023; Lekeufack et al., 2024), regret (Chen et al., 2023; Bhatnagar et al., 2023; Ge et al., 2024; Gibbs and Candès, 2023; Podkopaev et al., 2024; Zhang et al., 2024), and iterate convergence (Angelopoulos et al., 2024; Zecchin and Simeone, 2024).

In a way, our paper builds on the online conformal prediction literature and extends these ideas to general optimization problems. As we will show in Section 2.3.1, online conformal prediction is equivalent to finding a sequence of quantiles that satisfy gradient equilibrium with respect to the quantile loss. However, our paper takes this idea much farther, defining gradient equilibrium as a general condition, analyzing it in general, and deriving its implications for other families of losses beyond the quantile loss.

Another closely related line of work is that on multivalid conformal (Gupta et al., 2021; Bastani et al., 2022; Jung et al., 2023; Deng et al., 2023; Gibbs et al., 2023; Blot et al., 2024; Noarov et al., 2023), and especially its online variant as described in Bastani et al. (2022). In this line of work, it is shown that conformal prediction is an intercept-only quantile regression problem, and that simultaneous coverage guarantees over overlapping groups can be obtained by running a quantile regression on the vector of group indicators, i.e., (possibly overlapping) subsets of the domain $\mathcal{X}$. The literature on multivalidity was inspired by multicalibration (Hébert-Johnson et al., 2018) and multiaccuracy (Kim et al., 2019), which provide technical tools towards mitigating systematic biases in learning systems and have strong connections to the literature on fairness in machine learning (Barocas et al., 2023). The guarantees given by multivalid conformal algorithms parallel the multigroup debiasing guarantees in Sections 2.3.4, 3.2, and 3.3, and the multiaccuracy guarantees in Section 3.4. We view our work is complementary: in some sense, it is broader in scope (since we study gradient equilibrium in general), while our guarantees may be less developed and coarser. We also focus on applying online gradient descent as our main workhorse algorithm, rather than creating bespoke iterative procedures.

Farther afield, we remark that gradient equilibrium can be seen—viewing the time-average in (1) as an expectation—as an online variant of an estimating equation. See Godambe (1991) and Qin and Lawless (1994) for classical references. In a similar vein, the gradient equilibrium conditions which we examine in this paper are often interpretable as online analogs of standard first-order optimality guarantees in M-estimation. The main difference is that we are averaging over a sequence in a nonstochastic adversarial setting, whereas M-estimation and estimating equations are typically studied in a stochastic setting. Revisiting Table 1 will give the flavor of this parallel; recall, the squared loss leads to unbiasedness, the quantile loss leads to coverage, and so on.

Finally, in modern optimization-based online learning analyses, we note that it is common to analyze the convergence of the average gradient norm (for example, Theorem 2.1 of Ghadimi and Lan (2013) or Proposition 2.6 of Hazan et al. (2017)). Although this problem seems relevant at first glance, there is an important difference: these references study the *average norm of the gradient*, while we study the *norm of the average gradient*. These are very different objects. In gradient equilibrium, the individual gradients need not converge to zero. This allows our theory, as we will see in the next section, to handle a much broader class of functions than typical analyses.

1.3.3 Stochastic approximation

Lastly, closely related to our study of gradient equilibrium (and online optimization in general) is the topic of stochastic optimization. This field of course has an enormous literature of its own, and we will not be able to review it in any detail; it is still worth noting some connections. The seminal work of Robbins and Monro (Robbins and Monro, 1951)—which gave birth to the idea of stochastic gradient descent—does not require bona fide gradients to be used in the iterations. Robbins and Monro were concerned with stochastic approximation of the roots of nonlinear equations, a setting which encompasses stochastic optimization, but is genuinely broader in scope. Similarly, our analysis of online gradient descent does not require bona fide gradients; rather, the analysis applies to general vector fields that satisfy the conditions that we impose.

The connections run deeper—celebrated work on iterate averaging in stochastic approximation (Ruppert, 1988; Polyak and Juditsky, 1992) actually leverages gradient equilibrium (although not by this name and using a different analysis than ours) to prove a central limit theorem for the average iterate in certain settings. Though we adopt an online perspective in this work, it would be interesting to explore the stochastic setting in the future.

2. Gradient descent

We now study how to achieve gradient equilibrium via *online gradient descent*, which, given an initial point $\theta_1 \in \mathbb{R}^d$, produces iterates according to:

$$\theta_{t+1} = \theta_t - \eta_t g_t(\theta_t), \quad t = 1, 2, 3, \ldots, \tag{4}$$

for a sequence $\eta_t > 0$, $t = 1, 2, 3, \ldots$ of step sizes. Technically, this is the online subgradient method, since, recall, we use $g_t(\theta_t)$ to denote a generalized subgradient of ℓ_t at θ_t. For simplicity, we refer to the algorithm in (4) as gradient descent (GD), and $g_t(\theta_t)$ as a gradient.

In what follows, we assume that each loss ℓ_t is finite and subdifferentiable on all of $\mathbb{R}^d$. Importantly, we do not assume convexity. We also focus on constant step sizes, $\eta_t = \eta$, for all t. Later, in Section 4, we will see that the main conclusions discovered here for fixed step sizes all have analogs for decaying step sizes.

2.1 Bounded or slowly growing iterates

We begin by deriving a simple but useful bound on the average gradient.

Proposition 4 *Consider gradient descent (4), with arbitrary initialization $\theta_1 \in \mathbb{R}^d$, and constant step sizes $\eta_t = \eta > 0$, for all t. The average gradient satisfies*

$$\frac{1}{T}\sum_{t=1}^{T} g_t(\theta_t) = \frac{\theta_1 - \theta_{T+1}}{\eta T}, \tag{5}$$

and therefore

$$\left\| \frac{1}{T}\sum_{t=1}^{T} g_t(\theta_t) \right\|_2 \leq \frac{\|\theta_1\|_2 + \|\theta_{T+1}\|_2}{\eta T}. \tag{6}$$

Proof Rewrite the iteration (4), with $\eta_t = \eta$, as $\theta_t - \theta_{t+1} = \eta g_t(\theta_t)$. Adding this up over $t = 1, \ldots, T$, the left-hand side telescopes, yielding

$$\theta_1 - \theta_{T+1} = \eta \sum_{t=1}^{T} g_t(\theta_t).$$

Dividing both sides by ηT proves (5). The bound in (6) follows by simply taking the norm of both sides, and then applying the triangle inequality. ■

Proposition 4 is entirely elementary, yet the end result reveals an important property of gradient descent with constant step sizes: this algorithm achieves gradient equilibrium whenever the iterates remain bounded, since, if $\sup_{t\geq 1} \|\theta_t\|_2 \leq b$ for a constant $b > 0$ (not depending on T), then by (6),

$$\left\| \frac{1}{T}\sum_{t=1}^{T} g_t(\theta_t) \right\|_2 \leq \frac{2b}{\eta T} \to 0, \quad \text{as } T \to \infty.$$

It is worth noting that the bound b does not need to be known in order to run GD in the first place. It is also worth reiterating the contrast to regret: in general, GD with constant step sizes does not deliver a no-regret guarantee (as the example in Figure 3 illustrates).

It turns out that boundedness of the GD iterates arises naturally with some loss functions, and we will discuss this shortly. Meanwhile, a key realization is that boundedness is not actually necessary for gradient equilibrium: from (6), we see that we only need $\|\theta_{T+1}\|_2$ to grow sublinearly in T. We use the term *slowly growing* to refer to an iterate sequence which satisfies this sublinearity condition, writing it as $\|\theta_t\|_2 = o(t)$. Notice that from (5), gradient equilibrium itself implies $\|\theta_t - \theta_1\|_2$, hence $\|\theta_t\|_2$, is slowly growing. The next result summarizes these observations, for completeness.

Proposition 5 *For gradient descent (4) with constant step sizes, gradient equilibrium (1) holds if and only if the iterates are slowly growing.*

While sublinearity of the iterate norm $\|\theta_t\|_2$ clearly arises from (5) and (6) as a necessary and sufficient condition for gradient equilibrium, it would be dissatisfying for the analysis to end here. This is because we have not yet elucidated which problem settings (which loss sequences) lead to sublinear iterate growth if we apply GD. This is the topic of the analysis in the next subsection.

2.2 Restorative loss functions

For a parameter $h \geq 0$, and nonnegative function $\phi : \mathbb{R}^d \to \mathbb{R}_+$, we will say that a loss function $\ell : \mathbb{R}^d \to \mathbb{R}$ admits a (h, ϕ)-*restorative negative gradient field* provided that

$$g(\theta)^\mathsf{T}\theta \geq \phi(\theta), \quad \text{for all } \|\theta\|_2 > h, \text{ and all generalized subgradients } g(\theta) \text{ of } \ell \text{ at } \theta. \tag{7}$$

Informally, this condition says that if θ is far enough away from the origin (a distance of at least h), then the negative gradient field of ℓ is antialigned with directions of growth, because $-g(\theta)^\mathsf{T}\theta \leq -\phi(\theta) \leq 0$. This field hence imparts something of a restorative "force," which stops the parameter from "escaping to infinity" too quickly. The parameter h, which we call the *horizon*, is the distance at which this force activates; ϕ, which we call the *curvature function*, controls the strength of the restorative force. From here on, we will refer to a loss ℓ as being (h, ϕ)-restorative, with the understanding that this is really a property of its negative gradient field. See Figure 5 for an illustration.

In what follows, we examine the growth of GD iterates for what we call a *restorative loss sequence*. This is a sequence ℓ_t, $t = 1, 2, 3, \ldots$, such that

- each ℓ_t is (h_t, ϕ_t)-restorative, for some horizon h_t and curvature function ϕ_t;
- the horizon sequence is sublinear, $h_t = o(t)$, and nondecreasing.

It turns out that the univariate and multivariate cases are fundamentally different, in terms of the required curvature in the negative gradient field. We therefore separate our investigation into cases. The proofs of all results below are elementary, but deferred until the appendix to preserve the flow of ideas.

2.2.1 Univariate losses with weak curvature

Our first result studies restorative univariate loss sequences with zero curvature. Note that in the univariate case, $d = 1$, the restorative condition (7) with $\phi = 0$ reduces to

$$\text{sign}(g(\theta)) = \text{sign}(\theta), \quad \text{for all } |\theta| > h, \text{ and all generalized subgradients } g(\theta) \text{ of } \ell \text{ at } \theta. \tag{8}$$

In other words, the negative gradient field points back toward the origin beyond a horizon h. This property, along with a Lipschitz assumption on the loss, ensures strong control of univariate GD iterates even under the weakest possible curvature assumption, with $\phi = 0$. The proof of the next result is in Appendix A.2.

Proposition 6 *Let $d = 1$ and assume that each ℓ_t is L-Lipschitz, and $(h_t, 0)$-restorative. Assume also that h_t is nondecreasing. Then gradient descent (4), under arbitrary initialization $\theta_1 \in \mathbb{R}$, and constant step sizes $\eta_t = \eta > 0$, for all t, satisfies*

$$|\theta_{T+1}| \leq \max\{|\theta_1|, h_T\} + \eta L. \tag{9}$$

In particular, if h_t is sublinear, then the iterates are slowly growing. By (6), they satisfy gradient equilibrium:

$$\left|\frac{1}{T}\sum_{t=1}^{T} g_t(\theta_t)\right| \leq \frac{2|\theta_1|}{\eta T} + \frac{L}{T} + \frac{h_T}{\eta T} \to 0, \quad \text{as } T \to \infty. \tag{10}$$

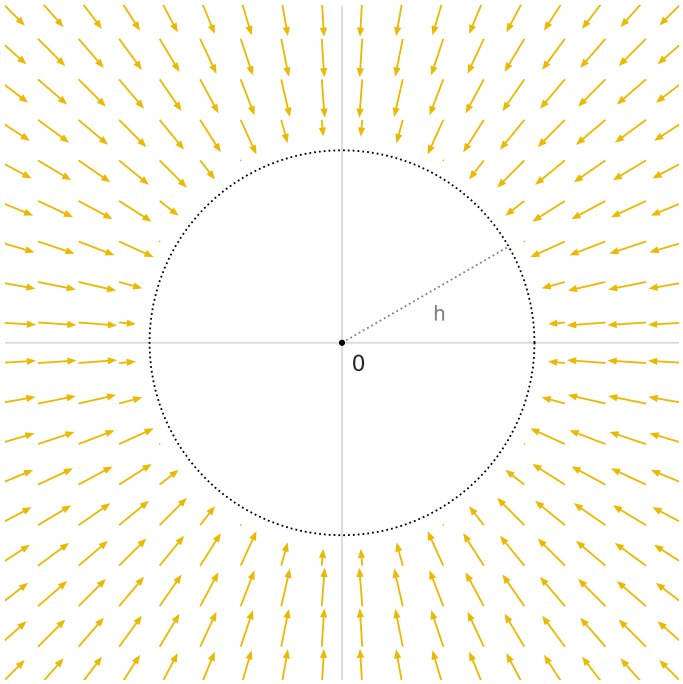
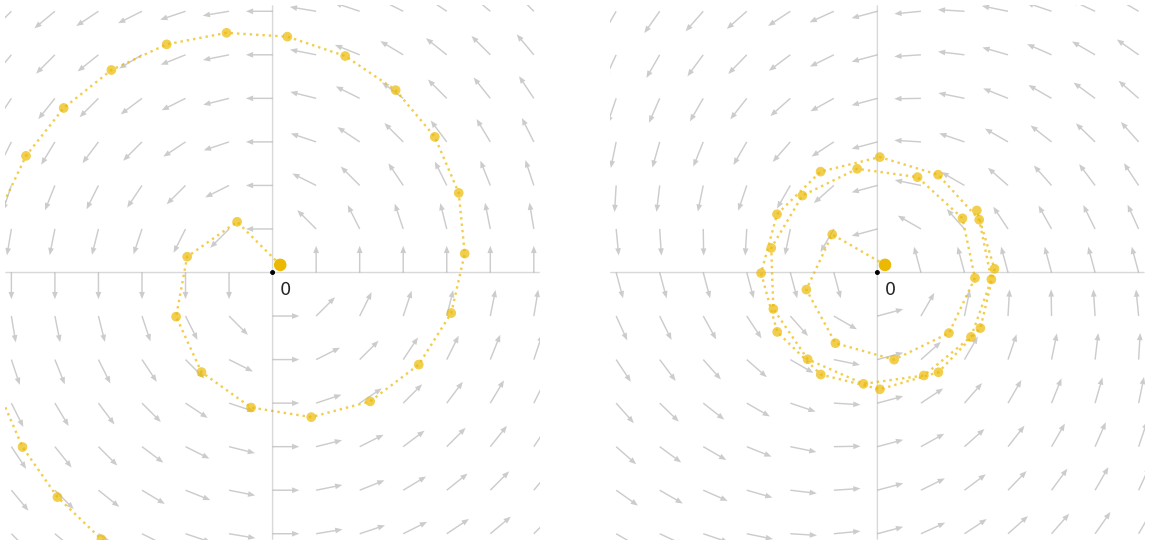


Figure 5: *Illustration of a restorative field, where each gold arrow represents the negative gradient of ℓ_t at a particular point. Note that if we take this point to be θ_t, then the gradient descent update would move θ_{t+1} in the direction of arrow. The gold arrows need to point inwards outside of a radius h; within the radius, the field can be arbitrary, so it is not drawn.*



(a) Zero curvature (b) Positive curvature

Figure 6: *Two example gradient descent trajectories, in gold, resulting from different negative gradient fields. In the left plot, the field has zero curvature, pointing at a 90° angle from each line to the origin, causing the gradient descent iterates to spiral outwards. In the right plot, the field has a positive curvature of 15°, and this causes the iterates to remain bounded.*

When the horizon sequence is constant, $h_t = h$ for all t, the result in (9) shows that the iterates remain bounded, and (10) shows that the average gradient diminishes at the rate $1/T$. We note that this effectively generalizes results from Gibbs and Candès (2021) and Angelopoulos et al. (2023) on quantile losses, which, under a boundedness assumption on the data, satisfy the property (8) for a global constant $h > 0$. We return to this setting in Section 2.3.1.

2.2.2 Multivariate losses with weak curvature

When $d \geq 2$, the story is quite different under zero curvature in a restorative negative gradient field, as the next result shows. Its proof is given in Appendix A.3.

Proposition 7 *Assume that each ℓ_t is L-Lipschitz and $(h_t, 0)$-restorative. Then gradient descent (4), with arbitrary initialization $\theta_1 \in \mathbb{R}^d$, and constant step sizes $\eta_t = \eta > 0$, for all t, satisfies*

$$\|\theta_{T+1}\|_2 \leq \sqrt{\|\theta_1\|_2^2 + \eta^2 L^2 T + 2\eta L \sum_{t=1}^{T} h_t}. \tag{11}$$

In particular, if h_t is sublinear and nondecreasing, then the iterates are slowly growing. By (6), they satisfy gradient equilibrium:

$$\left\| \frac{1}{T} \sum_{t=1}^{T} g_t(\theta_t) \right\|_2 \leq \frac{2\|\theta_1\|_2}{\eta T} + \sqrt{\frac{L^2}{T} + \frac{2Lh_T}{\eta T}} \to 0, \quad \text{as } T \to \infty. \tag{12}$$

Note that Proposition 7 can never guarantee bounded iterates. The upper bound (11) on $\|\theta_{T+1}\|_2$ grows at best at the rate $\sqrt{T}$, and hence the bound on the average gradient in (12) grows at best at the rate $1/\sqrt{T}$. These rates are achieved by a constant horizon sequence, $h_t = h$ for all t.

Meanwhile, it is not difficult to argue that Proposition 7 cannot be improved in terms of its dependence on T when $d \geq 2$, under the stated assumptions. Suppose that $g_t(\theta_t)^\mathsf{T}\theta_t = 0$ and $\|g_t(\theta_t)\|_2 = L$, for all t (this can be satisfied when $d \geq 2$, but not when $d = 1$). Then we have

$$\begin{aligned}
\|\theta_{T+1}\|_2^2 &= \|\theta_T\|_2^2 + \eta^2 \|g_T(\theta_T)\|_2^2 - 2\eta g_T(\theta_T)^\mathsf{T}\theta_T \\
&= \|\theta_T\|_2^2 + \eta^2 L^2 \\
&= \|\theta_1\|_2^2 + \eta^2 L^2 T,
\end{aligned}$$

thus $\|\theta_{T+1}\|_2$ grows precisely at the rate $\sqrt{T}$. Figure 6a gives an illustration. Hence we see that, in order to maintain bounded iterates in multiple dimensions, we need to strengthen the curvature assumption.

2.2.3 Multivariate losses with strong curvature

Our next result considers restorative loss sequences under a positive but constant curvature assumption. Its proof is given in Appendix A.4.

Proposition 8 *Assume that each ℓ_t is L-Lipschitz and (h_t, ϕ_t)-restorative, where the curvature is a sufficiently large positive constant:*

$$\phi_t(\theta) \geq \frac{\eta L^2}{2}, \quad \text{for all } \|\theta\|_2 > h_t. \tag{13}$$

Assume also that h_t is nondecreasing. Then gradient descent (4), with arbitrary initialization $\theta_1 \in \mathbb{R}^d$, and constant step sizes $\eta_t = \eta > 0$, for all t, satisfies

$$\|\theta_{T+1}\|_2 \leq \max\{\|\theta_1\|_2, h_T\} + \eta L. \tag{14}$$

Thus if h_t is sublinear, then the iterates are slowly growing. By (6), they satisfy gradient equilibrium:

$$\left\| \frac{1}{T} \sum_{t=1}^{T} g_t(\theta_t) \right\|_2 \leq \frac{2\|\theta_1\|_2}{\eta T} + \frac{L}{T} + \frac{h_T}{\eta T} \to 0, \quad \text{as } T \to \infty. \tag{15}$$

We note that Proposition 8 produces bounds of the same form as in Proposition 6, but it does so under the minimum positive curvature assumption (13). For a constant horizon sequence, $h_t = h$ for all t, the upper bound on $\|\theta_{T+1}\|_2$ in (14) is of constant order, and the average gradient bound (15) scales at the rate $1/T$. Figure 6b gives an illustration.

Our next result weakens the Lipschitz condition on the loss, while assuming even stronger curvature. Its proof is given in Appendix A.5.

Proposition 9 *Assume that each ℓ_t is L_t-Lipschitz on $\{\theta \in \mathbb{R}^d : \|\theta\|_2 \leq h_t\}$, equivalently,*

$$\|g_t(\theta)\|_2 \leq L_t, \quad \text{for all } \|\theta\|_2 \leq h_t, \tag{16}$$

and assume that ℓ_t is (h_t, ϕ_t)-restorative, where the curvature is a sufficiently large quadratic in the gradient:

$$\phi_t(\theta) \geq \frac{\eta}{2}\|g_t(\theta)\|_2^2, \quad \text{for all } \|\theta\|_2 > h_t. \tag{17}$$

Assume also that h_t, L_t are nondecreasing. Then gradient descent (4), with arbitrary initialization $\theta_1 \in \mathbb{R}^d$, and constant step sizes $\eta_t = \eta > 0$, for all t, satisfies

$$\|\theta_{T+1}\|_2 \leq \max\{\|\theta_1\|_2, h_T\} + \eta L_T. \tag{18}$$

Thus if h_t, L_t are sublinear, then the iterates are slowly growing. By (6), they satisfy gradient equilibrium:

$$\left\| \frac{1}{T} \sum_{t=1}^{T} g_t(\theta_t) \right\|_2 \leq \frac{2\|\theta_1\|_2}{\eta T} + \frac{L_T}{T} + \frac{h_T}{\eta T} \to 0, \quad \text{as } T \to \infty. \tag{19}$$

Like Proposition 8, this result gives bounded iterates (18), and the average gradient (19) vanishes at the rate $1/T$, for a constant horizon sequence, $h_t = h$ for all t. The difference is that we have only assumed local Lipschitzness (16), along with stronger curvature (17) in the negative gradient field.

2.3 Examples of restorative losses

We work through several examples of restorative losses.

2.3.1 Quantile losses

First, consider $\ell_t(\theta) = \rho_\tau(y_t - \theta)$, where ρ_τ denotes the quantile loss at level $\tau \in [0,1]$, i.e., $\rho_\tau(u) = \tau|u|$ for $u \geq 0$ and $(1-\tau)|u|$ for $u < 0$. We will show that this loss exhibits the restorative condition (8) with zero curvature. For $\theta \neq y_t$, the loss ℓ_t is differentiable at θ with gradient

$$g_t(\theta) = \begin{cases} -\tau & \text{if } \theta < y_t, \\ 1-\tau & \text{if } \theta > y_t. \end{cases}$$

We can see that (8) is met, provided that we choose a horizon $h \geq |y_t|$. Note also that ℓ_t is Lipschitz with constant $L = \max\{\tau, 1-\tau\} \leq 1$. The next result summarizes the conclusion for quantile loss. Its proof is in Appendix A.6.

Corollary 10 *Let $\ell_t(\theta) = \rho_\tau(y_t - \theta)$. Then ℓ_t satisfies the $(h_t, 0)$-restorative condition (8) for any $h_t \geq |y_t|$. Hence supposing each $|y_t| \leq b_t$, where b_t is sublinear and nondecreasing, we can set $h_t = b_t$, and by (10) (and assuming we take $g_t(\theta_t) = -\tau$ whenever $\theta_t = y_t$),*

$$\left| \frac{1}{T} \sum_{t=1}^{T} 1\{y_t \leq \theta_t\} - \tau \right| \leq \frac{2|\theta_1| + \eta + b_T}{\eta T}. \tag{20}$$

The result in Corollary 10 reproduces an online conformal prediction result from Angelopoulos et al. (2023). These authors consider a setting in which each y_t is replaced by a score, denoted $s_t = s(x_t, y_t)$, where s is a score function, x_t is a feature, and y_t is a response. The learned parameter θ_t is used to build a prediction set via

$$C_t(x_t) = \{y : s(x_t, y) \leq \theta_t\},$$

for the (unseen) response y_t at t. Because $y_t \in C_t(x_t) \iff s_t \leq \theta_t$, note that (20) applied to the scores s_t, $t = 1, 2, 3, \ldots$ (which we assume are bounded in magnitude by b) translates into the following guarantee:

$$\left| \frac{1}{T} \sum_{t=1}^{T} 1\{y_t \in C_t(x_t)\} - \tau \right| \leq \frac{\eta + b}{\eta T}, \tag{21}$$

where we set $\theta_1 = 0$ for simplicity. This problem setting will be studied empirically in Section 5.4.

2.3.2 Squared losses

Moving on to squared loss, $\ell_t(\theta) = \frac{1}{2}(y_t - \theta)^2$, we will show that this loss satisfies the restorative property (7) with quadratic curvature (17), the strongest kind of curvature. For such ℓ_t, this condition becomes

$$\theta(\theta - y_t) \geq \frac{\eta}{2}(y_t - \theta)^2, \quad \text{for } |\theta| > h_t.$$

Roughly speaking, we can always satisfy this condition for large enough $|\theta|$, and $\eta \leq 2$: if (say) $\theta_t = 2y_t > 0$, then the left-hand side is $2y_t \cdot y_t = 2y_t^2$ but the right-hand side is only $(\eta/2)y_t^2 \leq y_t^2$. The next result makes this idea precise. Its proof is given in Appendix A.7.

Corollary 11 *Let $\ell_t(\theta) = \frac{1}{2}(y_t - \theta)^2$. Fixing any $\delta \in (0,1)$, ℓ_t satisfies the (h_t, ϕ_t)-restorative condition (7) with quadratic curvature (17), for any horizon $h_t \geq |y_t|/\delta$ and step size $\eta \leq 2(1-\delta)/(1+\delta)^2$. Furthermore, suppose each $|y_t| \leq b_t$. Then the local Lipschitz condition (16) is met with $L_t = h_t + b_t$. If b_t is sublinear and nondecreasing, then for any $\delta \in (0,1)$, we can choose $\eta \leq 2(1-\delta)/(1+\delta)^2$ and $h_t = b_t/\delta$, and by (19),*

$$\left| \frac{1}{T} \sum_{t=1}^{T} \theta_t - \frac{1}{T} \sum_{t=1}^{T} y_t \right| \leq \frac{2|\theta_1| + b_T \eta (1 + 1/\delta) + b_T/\delta}{\eta T}. \tag{22}$$

Corollary 11 leads to guarantees for a scheme where we use GD to debias the output of a prediction model. In this setting, each y_t is replaced by a residual $y_t - f_t(x_t)$, where f_t is a black-box predictor, x_t is a feature, and y_t is a response. We use the parameter θ_t (learned by gradient descent) to augment each prediction, via $f_t(x_t) + \theta_t$. Note that (22) applied to the residuals $y_t - f_t(x_t)$, $t = 1, 2, 3, \ldots$ (which we assume are bounded in magnitude by b) translates into the following guarantee:

$$\left| \frac{1}{T} \sum_{t=1}^{T} (f_t(x_t) + \theta_t) - \frac{1}{T} \sum_{t=1}^{T} y_t \right| \leq \frac{b\eta(1 + 1/\delta) + b/\delta}{\eta T}, \tag{23}$$

where we set $\theta_1 = 0$ for simplicity. This problem setting will be studied empirically in Section 5.2.

2.3.3 Logistic losses

Next we study a generalized logistic loss, $\ell_t(\theta) = -y_t\theta + (b-a)\log(1+e^\theta) + a\theta$, for arbitrary values $a < b$. This generalization accommodates responses y_t lying in $[a, b]$ (for the usual logistic loss, we would set $a = 0$ and $b = 1$). We will show that this loss satisfies the restorative condition (8) with zero curvature. Note that ℓ_t has a gradient at θ of

$$g_t(\theta) = -y_t + (b-a)\frac{e^\theta}{1+e^\theta} + a.$$

We can see that ℓ_t is Lipschitz with constant $L = b - a$, and yet, it is clear that condition (8) cannot be met when $y_t = a$ or $y_t = b$. On the other hand, if y_t is bounded away from a and b, then (8) can be satisfied, as the next proposition shows. Its proof is in Appendix A.8.

Corollary 12 *Let $\ell_t(\theta) = -y_t\theta + (b-a)\log(1+e^\theta) + a\theta$, where $a < b$. Then ℓ_t satisfies the $(h_t, 0)$-restorative condition (8) for any*

$$h_t \geq \max\left\{ \log\frac{y_t - a}{b - y_t}, \log\frac{b - y_t}{y_t - a} \right\}.$$

(Note that h_t is infinite when $y_t = a$ or $y_t = b$.) Supposing each $y_t \in [a + \epsilon_t, b - \epsilon_t]$, where $\epsilon_t \in (0, (b-a)/2)$, is subexponentially vanishing in the sense that $1/\epsilon_t = \exp(o(t))$, and is nonincreasing, we can thus set

$$h_t = \log \frac{b-a}{2\epsilon_t}.$$

This will be sublinear and nondecreasing, and by (10),

$$\left| \frac{1}{T} \sum_{t=1}^{T} \left((b-a) \frac{e^{\theta_t}}{1+e^{\theta_t}} + a \right) - \frac{1}{T} \sum_{t=1}^{T} y_t \right| \leq \frac{2|\theta_1| + (b-a)\eta + \log((b-a)/(2\epsilon_t))}{\eta T}. \tag{24}$$

Analogous to the case of squared loss, Corollary 12 leads to guarantees for a scheme where we use GD to debias the output of a probabilistic classifer. In this setting, we replace y_t in the corollary by $y_t - p_t(x_t)$, for a black-box probabilistic classifer p_t, feature x_t, and binary response $y_t \in \{0, 1\}$. The learned parameter θ_t is used to augment the prediction via $p_t(x_t) + 2e^{\theta_t}/(1+e^{\theta_t}) - 1$. Notice now that (24) applied to the residuals $y_t - p_t(x_t)$, $t = 1, 2, 3, \ldots$ (we take $a = -1$ and $b = 1$, and assume the residuals lie in $[-1+\epsilon, 1-\epsilon]$, which can be accomplished by restricting $p_t(x_t)$ to lie in $[\epsilon, 1-\epsilon]$) translates into the following guarantee:

$$\left| \frac{1}{T} \sum_{t=1}^{T} \left(p_t(x_t) + \frac{2e^{\theta_t}}{1+e^{\theta_t}} - 1 \right) - \frac{1}{T} \sum_{t=1}^{T} y_t \right| \leq \frac{2\eta + \log(1/\epsilon)}{\eta T}, \tag{25}$$

where we set $\theta_1 = 0$ for simplicity. This problem setting will be studied empirically in Section 5.2.

2.3.4 GLMs with orthogonal features

Lastly, we study a generalized linear model (GLM) loss, $\ell_t(\theta) = -y_t x_t^\mathsf{T} \theta + \psi(x_t^\mathsf{T} \theta)$, where x_t is a feature, y_t is a response, and ψ is the cumulant generating function in the underlying exponential family, e.g., $\psi(u) = \frac{1}{2}u^2$ for linear regression and $\psi(u) = (b-a)\log(1+e^u) + au$ for generalized logistic regression. The GD updates (4) for a GLM, assuming as usual constant step sizes $\eta_t = \eta > 0$, for all t, are

$$\theta_{t+1} = \theta_t + \eta x_t (y_t - \psi'(x_t^\mathsf{T} \theta_t)), \quad t = 1, 2, 3, \ldots, \tag{26}$$

where ψ' denotes the derivative of ψ. Something interesting occurs in (26) when the features are orthogonal. Let u_j, $j = 1, \ldots, d$ be an orthonormal basis for $\mathbb{R}^d$, and suppose that

$$x_t \in \{u_j\}_{j=1}^d, \quad t = 1, 2, 3, \ldots,$$

i.e., the features only take values in this orthonormal set. We can then reparametrize the GD iterates by

$$\theta_t = \sum_{j=1}^{d} \vartheta_{tj} u_j, \quad t = 1, 2, 3, \ldots,$$

where $\vartheta_{tj} = u_j^\mathsf{T} \theta_t$, for $j = 1, \ldots, d$. Taking an inner product on each side of (26) with x_t gives

$$\vartheta_{t+1,j_t} = \vartheta_{t,j_t} + \eta(y_t - \psi'(\vartheta_{t,j_t})), \quad t = 1, 2, 3, \ldots, \tag{27}$$

where $j_t \in \{1, \dots, d\}$ is the index such that $x_t = u_{j_t}$. Thus, gradient descent for GLM losses in the current setting actually *decouples* into d separate GD processes, one per coordinate in the new parametrization.

These processes are truly decoupled in the sense that each coordinate ϑ_{tj} is only updated when $x_t = u_j$, and not at any other t. Hence to analyze GD in the current setting, we can simply analyze the GD iterations for each coordinate separately, effectively reducing the analysis to that of univariate GLMs, which was given in Corollaries 11 and 12. The next result provides details, and its proof is given in Appendix A.9.

Corollary 13 *Let $\ell_t(\theta) = -y_t x_t^\mathsf{T}\theta + \psi(x_t^\mathsf{T}\theta)$. Let u_j, $j = 1, \dots, d$ be an orthonormal basis $\mathbb{R}^d$, and suppose that each $x_t \in \{u_j\}_{j=1}^d$. Denote $I_j(T) = \{t \le T : j_t = j\}$, and $T_j = |I_j(T)|$, for $j = 1, \dots, d$.*

(a) If $\psi(u) = \frac{1}{2}u^2$ (linear regression), and $|y_t| \le b_t$, where b_t is sublinear and nondecreasing, then for any $\delta \in (0,1)$, and $\eta \le 2(1-\delta)/(1+\delta)^2$, we have for each $j = 1, \dots, d$,

$$\left|\frac{1}{T_j}\sum_{t\in I_j(T)} \vartheta_{tj} - \frac{1}{T_j}\sum_{t\in I_j(T)} y_t\right| \le \frac{2|\vartheta_{1j}| + b_T\eta(1+1/\delta) + b_T/\delta}{\eta T_j}. \tag{28}$$

(b) If $\psi(u) = (b-a)\log(1+e^u) + au$ (generalized logistic), and $y_t \in [a+\epsilon_t, b-\epsilon_t]$, where $\epsilon_t \in (0, (b-a)/2)$, is subexponentially vanishing in the sense that $1/\epsilon_t = \exp(o(t))$, and is nonincreasing, then we have for each $j = 1, \dots, d$,

$$\left|\frac{1}{T_j}\sum_{t\in I_j(T)} \left((b-a)\frac{e^{\vartheta_{tj}}}{1+e^{\vartheta_{tj}}} + a\right) - \frac{1}{T_j}\sum_{t\in I_j(T)} y_t\right| \le \frac{2|\vartheta_{1j}| + (b-a)\eta + \log((b-a)/(2\epsilon_t))}{\eta T_j}. \tag{29}$$

An important special case of Corollary 13 is when the features are indicators of group membership. This leads to guarantees precisely as in (23) and (25) for each group, which we call *multigroup debiasing*. Given a black-box predictor f_t, feature x_t, and response y_t, we can implement the GD iterations (26) with $\psi(u) = \frac{1}{2}u^2$, the residual $y_t - f_t(x_t)$ in place of y_t, and a group indicator vector $z_t \in \mathbb{R}^d$ in place of x_t. This has $z_{tj} = 1$ if observation t belongs to group j and 0 otherwise. Assuming that the groups are disjoint (nonoverlapping), the desired orthogonality condition is met, and (28) implies (assuming also that each residual $y_t - f_t(x_t)$ is bounded in magnitude by b):

$$\left|\frac{1}{T_j}\sum_{t\in I_j(T)} (f_t(x_t) + z_t^\mathsf{T}\theta_t) - \frac{1}{T_j}\sum_{t\in I_j(T)} y_t\right| \le \frac{b\eta(1+1/\delta) + b/\delta}{\eta T_j}, \tag{30}$$

for each $j = 1, \dots, d$, where we set $\theta_1 = 0$ for simplicity. Similarly, given a black-box probabilistic classifier p_t, applying (26) with $\psi(u) = (b-a)\log(1+e^u) + au$, the residual $y_t - p_t(x_t)$ in place of y_t, and z_t in place of x_t, implies (taking $a = -1$ and $b = 1$, and assuming each residual $y_t - p_t(x_t)$ lies in $[-1+\epsilon, 1-\epsilon]$):

$$\left|\frac{1}{T_j}\sum_{t\in I_j(T)} \left(p_t(x_t) + \frac{2e^{z_t^\mathsf{T}\theta}}{1+e^{z_t^\mathsf{T}\theta}} - 1\right) - \frac{1}{T_j}\sum_{t\in I_j(T)} y_t\right| \le \frac{2\eta + \log(1/\epsilon)}{\eta T_j}, \tag{31}$$

for each $j = 1, \dots, d$, where again we set $\theta_1 = 0$ for simplicity. The multigroup debiasing problem setting will be studied empirically in Section 5.3.

2.4 Connections to traditional conditions

As a final part of this section on gradient descent theory, we investigate connections between the restorative conditions used in the propositions and more traditional conditions used in optimization and adjacent fields. Copied here for convenience, the key assumptions are as follows:

$$\text{Proposition 7}: \quad g_t(\theta)^\mathsf{T}\theta \geq 0, \qquad \text{for all } \|\theta\|_2 > h_t, \tag{32}$$

$$\text{Proposition 8}: \quad g_t(\theta)^\mathsf{T}\theta \geq \tfrac{\eta L^2}{2}, \qquad \text{for all } \|\theta\|_2 > h_t, \tag{33}$$

$$\text{Proposition 9}: \quad g_t(\theta)^\mathsf{T}\theta \geq \tfrac{\eta}{2}\|g_t(\theta)\|_2^2, \qquad \text{for all } \|\theta\|_2 > h_t, \tag{34}$$

These are reminiscent of the following conditions arising in convex analysis and monotone operator theory:

$$\text{Monotonicity}: \quad (g_t(\theta) - g_t(\theta'))^\mathsf{T}(\theta - \theta') \geq 0, \qquad \text{for all } \theta, \theta', \tag{35}$$

$$\text{Strong monotonicity}: \quad (g_t(\theta) - g_t(\theta'))^\mathsf{T}(\theta - \theta') \geq \alpha\|\theta - \theta'\|_2^2, \qquad \text{for all } \theta, \theta', \tag{36}$$

$$\text{Co-coercivity}: \quad (g_t(\theta) - g_t(\theta'))^\mathsf{T}(\theta - \theta') \geq \mu\|g_t(\theta) - g_t(\theta')\|_2^2, \qquad \text{for all } \theta, \theta', \tag{37}$$

where $\alpha, \mu > 0$ are constants. A standard result in convex analysis is that, for differentiable ℓ_t, monotonicity of the gradient (35) is equivalent to convexity of ℓ_t, whereas strong monotonicity (36) is equivalent to strong convexity of ℓ_t. Another important result is the following: for convex ℓ_t, co-coercivity of the gradient (37) is equivalent to Lipschitzness of g_t (Baillon and Haddad, 1977). This result is often referred to as the *Baillon-Haddad theorem* in the literature on monotone operator theory (Bauschke and Combettes, 2009).

Immediately, we can see that conditions (32)–(34) only assume lower bounds on gradient inner products when $\|\theta\|_2 > h_t$, and not on the whole space, as in (35)–(37). Thus in no way do Propositions 7–9 require, e.g., convexity or strong convexity of ℓ_t. On the other hand, the curvature conditions (32)–(34) can be seen as weaker, *restricted* versions of (35)–(37), where "restricted" refers to the fact that the conditions are only assumed to hold when $\|\theta\|_2 > h_t$. The next proposition, whose proof is given in Appendix A.10, makes this connection precise.

Proposition 14 *Assume the restricted α_t-strongly monotone condition:*

$$(g_t(\theta) - g_t(0))^\mathsf{T}\theta \geq \alpha_t\|\theta\|_2^2, \quad \textit{for all } \|\theta\|_2 > h_t, \textit{ and all generalized subgradients } g_t(\theta). \tag{38}$$

Note that this condition is met when ℓ_t is α_t-strongly convex; but (38) is much weaker than strong convexity in general. Let $b_t \geq \|g_t(0)\|_2$. Then the following holds.

(a) For any $h_t \geq b_t/\alpha_t$, ℓ_t satisfies the restorative condition with zero curvature (32).

(b) For any $h_t \geq b_t/\alpha_t + \sqrt{\eta/2} \cdot L/\alpha_t$, ℓ_t satisfies the restorative condition with positive curvature (33).

(c) Let us additionally assume restricted β_t-co-coercivity and restricted β_t-smoothness conditions: that for all $\|\theta\|_2 > h_t$, and all generalized subgradients $g_t(\theta)$,

$$\beta_t(g_t(\theta) - g_t(0))^\mathsf{T}\theta \geq \|g_t(\theta) - g_t(0)\|_2^2 \quad \textit{and} \tag{39}$$

$$\|g_t(\theta) - g_t(0)\|_2 \leq \beta_t\|\theta\|_2. \tag{40}$$

Note that these conditions are simultaneously met when ℓ_t is differentiable and convex with β_t-Lipschitz gradient; but (39) and (40) are much weaker conditions in general. Then for any $\eta \leq 2/\beta_t$, and for any

$$h_t \geq \frac{b_t(1+\eta\beta_t+\sqrt{\eta/2})}{\alpha_t(1-\eta\beta_t/2)},$$

ℓ_t satisfies the restorative condition with quadratic curvature (34).

We reiterate that the conditions used in Proposition 14 are implied by standard conditions in optimization: (38) is implied by strong convexity and (39), (40) are implied by Lipschitzness of the gradient; however, the conditions in the proposition are weaker in general.

Remark 15 *As a sanity check (and investigation into the sharpness of the results in Proposition 14 in terms of their dependence on problem parameters, such as b_t and η), we can ask what the proposition says about squared loss, $\ell_t(\theta) = \frac{1}{2}(y_t-\theta)^2$. First, note that the gradient at the origin is $g_t(0) = y_t$, so we can take b_t to be a bound on $|y_t|$. Second, it is easy to check that ℓ_t is strongly convex with parameter $\alpha_t = 1$, and has a Lipschitz gradient with parameter $\beta_t = 1$. Thus we can apply part (c) of Proposition 14, which says that the restorative condition with quadratic curvature holds when $\eta \leq 2$ and*

$$h_t \geq \frac{b_t(1+\eta+\sqrt{\eta/2})}{1-\eta/2}.$$

Fix any $\delta \in (0,1)$; we can take $\eta \leq 2(1-\delta)/(1+\delta)^2$, and calculate $1-\eta/2 = \delta(3+\delta)/(1+\delta)^2 \geq \delta/(1+\delta)$, which means $1/(1-\eta/2) \leq (1+\delta)/\delta$. Using this along with the simple upper bounds $\delta \leq 1$ and $\eta \leq 2$, we see that a sufficient condition for the above display is

$$h_t \geq \frac{b_t(1+\eta+\sqrt{\eta/2})(1+\delta)}{\delta} = \frac{8b_t}{\delta}.$$

In other words, Proposition 14 part (c) implies that if $h_t \geq 8b_t/\delta$, then ℓ_t satisfies the restorative condition with quadratic curvature. We can see that this is less sharp than the direct analysis for squared loss from Corollary 11 by a constant factor of 8, but admits the same precise dependence on b_t and η (via δ).

Remark 16 *Meanwhile, for quantile and logistic losses the results in Proposition 14 cannot be applied. This is because neither loss satisfies the restricted strong monotonicity property in (38). Nevertheless, as we saw via direct analyses in Corollaries 10 and 12, useful results can still be derived in these cases, which emphasizes the fact that restricted strong monotonicity (or strong convexity, as a simpler sufficient condition) is certainly not the only avenue for establishing restorativeness of the negative gradient field.*

3. Regularization

We consider regularization applied to the loss sequence ℓ_t, $t = 1, 2, 3, \ldots$. At a high level, we will show that proper regularization can be used to modify the loss sequence so that

it satisfies restorative properties (when the original sequence does not), and therefore is amenable to applying gradient descent in order to achieve gradient equilibrium. Importantly, we also show that this translates back to a statement about the original equilibrium condition of interest (with respect to the original sequence), since regularization perturbs each gradient by a controlled amount.

To fix notation, let r be a regularizer, assumed to be finite and subdifferentiable on all of $\mathbb{R}^d$. As in our analysis of gradient descent, in the last section, we assume the same of each loss ℓ_t. Define the regularized loss sequence $\tilde{\ell}_t = \ell_t + r$, $t = 1, 2, 3, \ldots$. Two common choices of regularizers are:

- $r(\theta) = \lambda\|\theta\|_1$, often called a *lasso* penalty (Tibshirani, 1996); and
- $r(\theta) = \lambda\|\theta\|_2^2$, often called a *ridge* penalty (Hoerl and Kennard, 1970).

In either case $\lambda \geq 0$ is a tuning parameter which controls the strength of regularization. We first examine gradient equilibrium for a general regularized sequence, and then we analyze the lasso and ridge cases.

3.1 Gradient equilibrium revisited

We investigate gradient equilibrium for $\tilde{\ell}_t$, $t = 1, 2, 3, \ldots$, and relate it back to the original sequence. A key fact is that subgradients of $\tilde{\ell}_t$ can be related to those of ℓ_t: our generalized notion of subgradients allows us to take, as a subgradient $\tilde{g}_t(\theta)$ of $\tilde{\ell}_t$ at θ,

$$\tilde{g}_t(\theta) = g_t(\theta) + g_r(\theta), \tag{41}$$

where $g_t(\theta)$ is a subgradient of ℓ_t at θ, and $g_r(\theta)$ denotes a subgradient of r at θ. (Appendix D gives a proof of this fact.) Thus, for a sequence of subgradients, $\tilde{g}_t(\theta)$, $t = 1, 2, 3 \ldots$, which are chosen to satisfy (41),

$$\frac{1}{T}\sum_{t=1}^{T} \tilde{g}_t(\theta_t) = \frac{1}{T}\sum_{t=1}^{T} g_t(\theta_t) + \frac{1}{T}\sum_{t=1}^{T} g_r(\theta_t).$$

Rearranging, and using the triangle inequality, we get

$$\left\|\frac{1}{T}\sum_{t=1}^{T} g_t(\theta_t)\right\|_2 \leq \left\|\frac{1}{T}\sum_{t=1}^{T} \tilde{g}_t(\theta_t)\right\|_2 + \frac{1}{T}\sum_{t=1}^{T} \|g_r(\theta_t)\|_2. \tag{42}$$

If we apply gradient descent to $\tilde{\ell}_t$, $t = 1, 2, 3, \ldots$ then first term can be made small provided that r endows the regularized sequence with the appropriate restorative properties. The second term can be made small by a judicious choice of the amount of regularization (which is controlled by λ in the lasso and ridge penalties). The next two subsections follow this general template.

3.2 Logistic losses with lasso penalties

We consider generalized logistic losses with arbitrary features, $\ell_t(\theta) = -y_t x_t^\mathsf{T}\theta + (b-a)\log(1 + e^{x_t^\mathsf{T}\theta}) + a x_t^\mathsf{T}\theta$, where x_t is a feature, and y_t is a response lying in $[a, b]$. For arbitrary $x_t \in \mathbb{R}^d$,

this loss need not satisfy the restorative condition (7) for any $\phi(\theta) \geq 0$, even when y_t is bounded away from the endpoints a, b of the given range. This is because

$$g(\theta)^\mathsf{T}\theta = \bigg((b-a)\frac{e^{x_t^\mathsf{T}\theta}}{1+e^{x_t^\mathsf{T}\theta}} + a - y_t\bigg)x_t^\mathsf{T}\theta, \tag{43}$$

and the leading factor $(b-a)e^{x_t^\mathsf{T}\theta}/(1+e^{x_t^\mathsf{T}\theta}) + a - y_t$ can be negative when $x_t^\mathsf{T}\theta$ is small and positive.

An important realization is that the gradient inner product in (43) cannot be arbitrarily small. Next we give a finite lower bound on its infimum. The proof is given in Appendix A.11.

Lemma 17 *Let $g(u) = (b-a)e^u/(1+e^u) + a - y$, where $y \in [a,b]$, for $a < b$. Then*

$$\inf_u \, g(u)^\mathsf{T}u \geq -0.279(b-a).$$

The finite infimum in Lemma 17 suggests that an appropriate level of ℓ_1 regularization can "overwhelm" a possibly negative gradient inner product in order to ensure suitable curvature, and thus restorativeness, for the regularized loss. The proof of the next result is given in Appendix A.12.

Proposition 18 *Let $\tilde{\ell}_t(\theta) = -y_t x_t^\mathsf{T}\theta + (b-a)\log(1+e^{x_t^\mathsf{T}\theta}) + a x_t^\mathsf{T}\theta + \lambda\|\theta\|_1$, for arbitrary $\lambda > 0$, and assume $\|x_t\|_2 \leq c$, and $y \in [a,b]$. Then $\tilde{\ell}_t$ is Lipschitz continuous with parameter $L = (b-a)c + \lambda\sqrt{d}$, and satisfies the (h_t, ϕ_t)-restorative condition (7) with positive curvature (13) for any*

$$h_t \geq \frac{0.279(b-a) + \eta(c(b-a)+\lambda\sqrt{d})^2/2}{\lambda}.$$

Taking h_t to match this lower bound, we have from (15) that gradient descent (4), with $\tilde{g}_t$ in place of g_t and constant step sizes $\eta_t = \eta > 0$, for all t, satisfies

$$\bigg\|\frac{1}{T}\sum_{t=1}^T \bigg((b-a)\frac{e^{x_t^\mathsf{T}\theta_t}}{1+e^{x_t^\mathsf{T}\theta_t}} + a - y_t\bigg)x_t\bigg\|_2 \leq \frac{C_1(\lambda)}{\eta T} + \frac{C_2(\lambda)}{\lambda\eta T} + \lambda\sqrt{d}, \tag{44}$$

where $C_1(\lambda) = 2\|\theta_1\|_2 + (c(b-a)+\lambda\sqrt{d})/\eta$, and $C_2(\lambda) = 0.279(b-a) + \eta(c(b-a)+\lambda\sqrt{d})^2/2$.

Remark 19 *In this proposition, note we do not need y_t to be bounded away from a, b, the endpoints of its range (as we did in the unregularized, univariate logistic theory in Corollary 12, or the theory for orthogonal features in Corollary 13). Furthermore, a lasso penalty is chosen (instead of, say, a ridge penalty) here so that the regularized loss retains the Lipschitz property of the original logistic loss, which allows us to apply the positive curvature result from Proposition 8 (and allows the step size $\eta > 0$ to be arbitrary).*

Proposition 18 can be used to derive guarantees on a post-processing scheme for a black-box probabilistic classifier, where we use gradient descent to decorrelate the errors with any features of interest. Precisely, let $p_t(x_t)$ be a probabilistic prediction of a binary

response $y_t \in \{0,1\}$, based on a feature x_t. Replace y_t in the proposition by $y_t - p_t(x_t)$, and replace x_t by another feature $z_t \in \mathbb{R}^d$. To adjust predictions, we use:

$$p_t(x_t) + \frac{2e^{z_t^\mathsf{T}\theta_t}}{1+e^{z_t^\mathsf{T}\theta_t}} - 1, \quad t = 1,2,3,\ldots, \tag{45}$$

where θ_t, $t = 1,2,3,\ldots$ is obtained by gradient descent on the lasso-penalized logistic loss sequence (and we set $a = -1$ and $b = 1$ in the language of the proposition). If each $\|z_t\|_2 \leq c$, then the result in (44) says:

$$\left\| \frac{1}{T}\sum_{t=1}^{T}\left(p_t(x_t) + \frac{2e^{z_t^\mathsf{T}\theta_t}}{1+e^{z_t^\mathsf{T}\theta_t}} - 1 - y_t\right) z_t \right\|_2 \leq \frac{2c+\lambda\sqrt{d}}{T} + \frac{1.116 + \eta(2c+\lambda\sqrt{d})^2}{2\lambda\eta T} + \lambda\sqrt{d}, \tag{46}$$

where we set $\theta_1 = 0$ for simplicity. The bound in (46) can be made small by taking λ to be small and T to be large. For example, taking $\lambda = 1/\sqrt{T}$ to approximately balance the second and third terms gives

$$\left\| \frac{1}{T}\sum_{t=1}^{T}\left(p_t(x_t) + \frac{2e^{z_t^\mathsf{T}\theta_t}}{1+e^{z_t^\mathsf{T}\theta_t}} - 1 - y_t\right) z_t \right\|_2 \leq O\left(\frac{1}{\sqrt{T}}\right),$$

where we have simplified the bound to highlight the dependence on T alone.

As an example, we can apply this post-processing strategy in a setting in which z_t indicates membership with respect to a set of overlapping groups. In this setting (with overlap), the features will not be orthogonal and so the previous theory (Corollary 13 part (b)) does not apply. In Sections 5.3 and 5.5, we study empirical examples with overlapping groups and regularization. We must emphasize, however, that the post-processing strategy described above accommodates a much broader set of applications, as it allows the user to specify the secondary feature set z_t arbitrarily. We return to this point in Section 3.4.

3.3 Squared losses with ridge penalties

We now consider squared losses with arbitrary features, $\ell_t(\theta) = \frac{1}{2}(y_t - x_t^\mathsf{T}\theta)^2$, where x_t is a feature, and y_t is a response with $|y_t| \leq b_t$. For arbitrary $x_t \in \mathbb{R}^d$, this loss need not satisfy the restorative property (7) for any $\phi(\theta) \geq 0$. This is because

$$g(\theta)^\mathsf{T}\theta = (x_t^T\theta - y_t)x_t^\mathsf{T}\theta, \tag{47}$$

and the leading factor $x_t^T\theta - y_t$ can be negative when $x_t^\mathsf{T}\theta$ is small and positive.

Similar to the earlier logistic setting, a key realization is that the gradient inner product (47) cannot be arbitrarily small, and is subject to a finite lower bound. The proof of the next result is in Appendix A.13.

Lemma 20 *Let $g(u) = u - y$, where $|y| \leq b$. Then for arbitrary $a_1 > a_2$,*

$$\inf_u \left\{ a_1 g(u)^\mathsf{T} u - a_2(u-y)^2 \right\} \geq -\frac{(a_1 - 2a_2)^2 b^2}{4(a_1 - a_2)} - a_2 b^2,$$

Again, in a similar vein to our earlier analysis of logistic regression, the finite infimum in Lemma 20 means that we can use an appropriate level of regularization to "overwhelm" a potentially negative gradient inner product, thus ensuring proper curvature and restorativeness for the regularized loss. (The generalized form of the criterion, involving $a_1, a_2 \geq 0$, will allow us to verify quadratic curvature.) The proof of the next result is given in Appendix A.14.

Proposition 21 *Let $\tilde{\ell}_t(\theta) = \frac{1}{2}(y_t - x_t^\top \theta)^2 + \frac{\lambda}{2}\|\theta\|_2^2$, for arbitrary $\lambda > 0$, and assume $|y_t| \leq b_t$ and $\|x_t\|_2 \leq c_t$ for sublinear and nondecreasing sequences b_t, c_t. Then $\tilde{\ell}_t$ is Lipschitz continuous on $\{\theta \in \mathbb{R}^d : \|\theta\|_2 \leq h_t\}$ with parameter $L_t = (c_t^2 + \lambda)h_t + b_t c_t$, and satisfies the (h_t, ϕ_t)-restorative condition (7) with quadratic curvature (17) for any $\eta < 1/(\lambda + c_t^2/2)$, and*

$$h_t^2 \geq \frac{C_{0t}(\lambda)}{\lambda},$$

where $C_{0t}(\lambda) = ((1 - \lambda\eta - \eta c_t^2)^2 b_t^2/(4(1 - \lambda\eta - \eta c_t^2/2)) + \eta c_t^2 b_t^2/2)/(1 - \lambda\eta/2)$. Taking h_t to match this lower bound, we have from (19) that gradient descent (4), with $\tilde{g}_t$ in place of g_t and constant step sizes $\eta_t = \eta > 0$, for all t, satisfies

$$\left\| \frac{1}{T} \sum_{t=1}^{T} (x_t^T \theta_t - y_t) x_t \right\|_2 \leq \frac{C_{1T}(\lambda)}{\eta T} + \frac{C_{2T}(\lambda)}{\sqrt{\lambda}\eta T} + C_{3T}(\lambda), \tag{48}$$

where $C_{1T}(\lambda) = 2\|\theta_1\|_2 + \eta b_T c_T$, $C_{2T}(\lambda) = \sqrt{C_{0T}(\lambda)}(\eta(c_T^2 + \lambda) + 1)$, and $C_{3T}(\lambda) = \sqrt{\lambda} C_{2T}(\lambda) + \lambda(\|\theta_1\|_2 + \eta b_T c_T)$.

Remark 22 *The use of a ridge penalty (instead of, say, a lasso penalty) is important here since it leads to quadratic curvature, which allows us to apply the result from Proposition 9. This type of curvature is needed because the squared loss is not Lipschitz, only locally Lipschitz, so other results (in particular, Proposition 8) cannot be applied. Moreover, we note that relying on strong convexity alone to establish restorativeness (as in Proposition 14) would be too blunt to deliver the result in (48). Deriving this requires a more specialized analysis, which relies on the fact that the gradient inner product (47) for squared loss has a finite infimum. It is possible that a similar analysis could be extended to cover all GLMs, but we do not pursue this.*

Proposition 21 can be used to derive guarantees on a scheme which post-processes a black-box prediction model, using gradient descent to decorrelate the errors with any given features of interest. Let $f_t(x_t)$ be the prediction of a response y_t, based on a feature x_t. Replace y_t in the proposition by the residual $y_t - f_t(x_t)$, and x_t by a secondary feature $z_t \in \mathbb{R}^d$. We adjust predictions using:

$$f_t(x_t) + \theta_t, \quad t = 1, 2, 3, \ldots, \tag{49}$$

where θ_t, $t = 1, 2, 3, \ldots$ is obtained by gradient descent on the ridge-penalized squared loss sequence. If each $|y_t - f_t(x_t)| \leq b$ and $\|z_t\|_2 \leq c$, then the result in (48) says:

$$\left\| \frac{1}{T} \sum_{t=1}^{T} (f_t(x_t) + z_t^T \theta_t - y_t) z_t \right\|_2 \leq \frac{bc}{T} + \frac{\sqrt{C(\lambda)}}{\sqrt{\lambda}\eta T} + \sqrt{\lambda C(\lambda)} + \lambda\eta bc, \tag{50}$$

where $C(\lambda) = [((1-\lambda\eta-\eta c^2)^2 b^2/(4(1-\lambda\eta-\eta c^2/2)) + \eta c^2 b^2/2)/(1-\lambda\eta/2)] \cdot (\eta(c^2+\lambda)+1)$, and we set $\theta_1 = 0$ for simplicity. The bound in (50) can be made small by taking λ to be small and T to be large. For example, taking $\lambda = 1/T$ to approximately balance the third and fourth terms gives

$$\left\| \frac{1}{T} \sum_{t=1}^{T} (f_t(x_t) + z_t^T \theta_t - y_t) z_t \right\|_2 \leq O\left(\frac{1}{\sqrt{T}}\right),$$

where we have simplified the bound to highlight the dependence on T.

As in the logistic setting, we note that we can apply such a post-processing scheme to achieve groupwise debiased predictions in a setting with overlapping groups. This is important as it expands our current set of available tools—the features will not be orthogonal (due to overlap), so the previous theory (Corollary 13 part (a)) cannot be applied. However, the result in (50) is considerably broader, because the secondary feature z_t can be specified arbitrarily. Next, we discuss implications for multiaccuracy.

3.4 Implications for multiaccuracy

We show how the decorrelation results (46), (50) can be used to derive an online method with multiaccuracy guarantees. Let $\mathcal{F}$ be a linear, finite-dimensional class of real-valued functions on the space of features. We can write any function $F \in \mathcal{F}$ as

$$F(x) = \sum_{j=1}^{d} \alpha_j b_j(x),$$

for coefficients $\alpha_j \in \mathbb{R}$, $j = 1, \ldots, d$ and basis functions b_j, $j = 1, \ldots, d$. Abbreviate the basis expansion at x by $F(x) = b(x)^\mathsf{T}\alpha$. With $z_t = b(x_t)$ for $t = 1, 2, 3, \ldots$, consider the logistic post-processing scheme following Proposition 18, where we adjust the predictions from a black-box probabilistic classifier p_t as in (45). Observe

$$\begin{aligned}\left| \frac{1}{T} \sum_{t=1}^{T} \left(p_t(x_t) + \frac{2e^{z_t^\mathsf{T}\theta_t}}{1+e^{z_t^\mathsf{T}\theta_t}} - 1 - y_t \right) F(x_t) \right| &= \left| \frac{1}{T} \sum_{t=1}^{T} \left(p_t(x_t) + \frac{2e^{z_t^\mathsf{T}\theta_t}}{1+e^{z_t^\mathsf{T}\theta_t}} - 1 - y_t \right) z_t^\mathsf{T}\alpha \right| \\ &\leq \left\| \frac{1}{T} \sum_{t=1}^{T} \left(p_t(x_t) + \frac{2e^{z_t^\mathsf{T}\theta_t}}{1+e^{z_t^\mathsf{T}\theta_t}} - 1 - y_t \right) z_t \right\|_2 \|\alpha\|_2,\end{aligned}$$

where the first line uses $F(x_t) = z_t^\mathsf{T}\alpha$ for each t, and the second line uses Cauchy-Schwarz. Thus, introducing the norm $\|F\| = \|\alpha\|_2$ on $\mathcal{F}$, and assuming $\|b(x)\|_2 \leq c$ for all x, the result in (46) implies

$$\begin{aligned}\sup_{F \in \mathcal{F}, \|F\| \leq 1} &\left| \frac{1}{T} \sum_{t=1}^{T} \left(p_t(x_t) + \frac{2e^{z_t^\mathsf{T}\theta_t}}{1+e^{z_t^\mathsf{T}\theta_t}} - 1 - y_t \right) F(x_t) \right| \\ &\leq \frac{2c + \lambda\sqrt{d}}{T} + \frac{1.116 + \eta(2c+\lambda\sqrt{d})^2}{2\lambda\eta T} + \lambda\sqrt{d}.\end{aligned} \tag{51}$$

Similarly, if we apply the squared loss post-processing scheme following Proposition 21, where we adjust the predictions from a black box-model f_t as in (49), then the result in (50) implies

$$\sup_{F \in \mathcal{F},\, \|F\| \leq 1} \left| \frac{1}{T} \sum_{t=1}^{T} (f_t(x_t) + z_t^T \theta_t - y_t) F(x_t) \right| \leq \frac{bc}{T} + \frac{\sqrt{C(\lambda)}}{\sqrt{\lambda} \eta T} + \sqrt{\lambda C(\lambda)} + \lambda \eta bc, \tag{52}$$

where as before $C(\lambda) = [((1 - \lambda\eta - \eta c^2)^2 b^2 / (4(1 - \lambda\eta - \eta c^2/2)) + \eta c^2 b^2/2)/(1 - \lambda\eta/2)] \cdot (\eta(c^2 + \lambda) + 1)$.

Each of (51), (52) are online multiaccuracy bounds. They apply to a relatively restricted function class $\mathcal{F}$ (linear and finite-dimensional), but they make no assumptions about the data or the original predictors p_t or f_t whatsoever (other than boundedness), i.e., they apply deterministically. The post-processing scheme based on gradient descent is also very simple to implement practically. See Section 1.3 for a discussion of related work in the context of multiaccuracy.

4. Arbitrary step sizes

We extend our theory to the case of decaying step sizes. As in the last two sections, we assume that each loss ℓ_t is finite and subdifferentiable on all of $\mathbb{R}^d$. Core to our extension is a modification of Proposition 4.

Proposition 23 *Consider gradient descent (4), with arbitrary initialization $\theta_1 \in \mathbb{R}^d$, and abitrary positive step sizes $\eta_t > 0$, $t = 1, 2, 3, \ldots$. For convenience, let $\eta_0^{-1} = 0$, and define*

$$\Delta_t = \eta_t^{-1} - \eta_{t-1}^{-1}, \quad t = 1, 2, 3, \ldots,$$

The average gradient satisfies

$$\frac{1}{T} \sum_{t=1}^{T} g_t(\theta_t) = \frac{1}{T} \sum_{t=1}^{T} (\theta_t - \theta_{T+1}) \Delta_t, \tag{53}$$

and therefore

$$\left\| \frac{1}{T} \sum_{t=1}^{T} g_t(\theta_t) \right\|_2 \leq \frac{2}{T} \Big(\max_{t \leq T+1} \|\theta_t\|_2 \Big) \sum_{t=1}^{T} |\Delta_t|. \tag{54}$$

When the step sizes η_t, $t = 1, 2, 3, \ldots$ are nonincreasing,

$$\left\| \frac{1}{T} \sum_{t=1}^{T} g_t(\theta_t) \right\|_2 \leq \frac{2}{\eta_T T} \Big(\max_{t \leq T+1} \|\theta_t\|_2 \Big). \tag{55}$$

Proof Rewrite the iteration (4) as $\theta_t - \theta_{t+1} = \eta_t g_t(\theta_t)$. Adding this up over $t = s, \ldots, T$, the left-hand side telescopes, yielding

$$\theta_s - \theta_{T+1} = \sum_{t=s}^{T} \eta_t g_t(\theta_t)$$

Next, for all t, we have

$$\sum_{s=1}^{t} \Delta_s = \eta_t^{-1}.$$

This allows us to write

$$\begin{aligned}
\frac{1}{T}\sum_{t=1}^{T} g_t(\theta_t) &= \frac{1}{T}\sum_{t=1}^{T}\sum_{s=1}^{t} \Delta_s \eta_t g_t(\theta_t) \\
&= \frac{1}{T}\sum_{s=1}^{T} \Delta_s \sum_{t=s}^{T} \eta_t g_t(\theta_t) \\
&= \frac{1}{T}\sum_{s=1}^{T} \Delta_s (\theta_s - \theta_{T+1}).
\end{aligned}$$

This proves (53). Furthermore, writing $\Delta = (\Delta_1, \ldots, \Delta_T) \in \mathbb{R}^d$,

$$\begin{aligned}
\left\| \frac{1}{T}\sum_{t=1}^{T} g_t(\theta_t) \right\|_2 &\leq \frac{1}{T}\sum_{s=1}^{T} |\Delta_s| \|\theta_s - \theta_{T+1}\|_2 \\
&\leq \frac{1}{T}\|\Delta\|_1 \Big(\max_{s \leq T} \|\theta_s - \theta_{T+1}\|_2 \Big) \\
&\leq \frac{2}{T}\|\Delta\|_1 \Big(\max_{s \leq T+1} \|\theta_s\|_2 \Big),
\end{aligned}$$

which proves (54). Lastly, when the step sizes are nonincreasing,

$$\|\Delta\|_1 = \eta_1^{-1} + \sum_{t=1}^{T} (\eta_t^{-1} - \eta_{t-1}^{-1}) = \eta_T^{-1},$$

which verifies (55), and completes the proof. ■

Observe that the key term in either of (54), (55) is the maximum iterate norm $\max_{t \leq T+1} \|\theta_t\|_2$. The next result summarizes when these bounds lead to gradient equilibrium.

Proposition 24 *For gradient descent (4) with arbitrary positive sizes, gradient equilibrium (1) holds if the iterates satisfy* $\max_{t \leq T+1} \|\theta_t\|_2 = o(T / \sum_{t=1}^{T} |\Delta_t|)$. *For nonincreasing step sizes, this condition simplifies to* $\max_{t \leq T+1} \|\theta_t\|_2 = o(\eta_T T)$. *When the step sizes decrease slowly enough in order for* $\eta_t t$, $t = 1, 2, 3, \ldots$ *to be nonincreasing, it further simplifies to* $\|\theta_{t+1}\|_2 = o(\eta_t t)$.

Fortunately, the restorative theory established in Section 2.2, which provides control over iterate norms, carries over to arbitrary step sizes. The next result extends Propositions 6–9. Its proof is basically identical to that of these propositions, and is hence omitted.

Proposition 25 *Consider gradient descent (4), with arbitrary initialization* $\theta_1 \in \mathbb{R}^d$, *and abitrary positive step sizes* $\eta_t > 0$, $t = 1, 2, 3, \ldots$. *Abbreviate* $N_t = \max_{s \leq t} \eta_s$, *and let* $c > 0$ *and* $\alpha \in (0, 1)$ *be arbitrary.*

(a) If $d = 1$, and each ℓ_t is L-Lipschitz and $(h_t, 0)$-restorative, for nondecreasing h_t, then

$$|\theta_{T+1}| \le \max\{|\theta_1|, h_T\} + N_T L. \tag{56}$$

Thus if we set each $\eta_t = ct^{-\alpha}$, and $h_t = o(t^{1-\alpha})$, then (55) implies gradient equilibrium:

$$\left|\frac{1}{T}\sum_{t=1}^{T} g_t(\theta_t)\right| \le \frac{2|\theta_1|/c}{T^{1-\alpha}} + \frac{2(L + h_T/c)}{T^{1-\alpha}} \to 0, \quad \text{as } T \to \infty. \tag{57}$$

(b) For general d, if each ℓ_t is L-Lipschitz and $(h_t, 0)$-restorative, then

$$\|\theta_{T+1}\|_2 \le \sqrt{\|\theta_1\|_2^2 + \sum_{t=1}^{T}(\eta_t^2 L^2 + 2\eta_t h_t L)}. \tag{58}$$

Thus if we set each $\eta_t = ct^{-\alpha}$, and $h_t = o(t^{1-\alpha})$, then (55) implies gradient equilibrium:

$$\left\|\frac{1}{T}\sum_{t=1}^{T} g_t(\theta_t)\right\|_2 \le \frac{2\|\theta_1\|_2/c}{T^{1-\alpha}} + \sqrt{\sum_{t=1}^{T}\frac{t^{-2\alpha}L^2 + 2t^{-\alpha}h_t L/c}{T^{2(1-\alpha)}}} \to 0, \quad \text{as } T \to \infty. \tag{59}$$

(c) If each ℓ_t is L-Lipschitz and (h_t, ϕ_t)-restorative with positive curvature: $\phi_t(\theta) \ge \eta_t L^2/2$, for nondecreasing h_t, then

$$\|\theta_{T+1}\|_2 \le \max\{\|\theta_1\|_2, h_T\} + N_T L. \tag{60}$$

Thus if we set each $\eta_t = ct^{-\alpha}$, and $h_t = o(t^{1-\alpha})$, then (55) implies gradient equilibrium:

$$\left\|\frac{1}{T}\sum_{t=1}^{T} g_t(\theta_t)\right\|_2 \le \frac{2\|\theta_1\|_2/c}{T^{1-\alpha}} + \frac{2(L + h_T/c)}{T^{1-\alpha}} \to 0, \quad \text{as } T \to \infty. \tag{61}$$

(d) If each ℓ_t is L_t-Lipschitz on the set $\{\theta \in \mathbb{R}^d : \|\theta\|_2 \le h_t\}$, and also (h_t, ϕ_t)-restorative with quadratic curvature: $\phi_t(\theta) \ge \eta_t \|g_t(\theta)\|_2^2/2$, for nondecreasing h_t, L_t, then

$$\|\theta_{T+1}\|_2 \le \max\{\|\theta_1\|_2, h_T\} + N_T L_T. \tag{62}$$

Thus if we set each $\eta_t = ct^{-\alpha}$, and $h_t = o(t^{1-\alpha})$, $L_t = o(t^{1-\alpha})$, then (55) implies gradient equilibrium:

$$\left\|\frac{1}{T}\sum_{t=1}^{T} g_t(\theta_t)\right\|_2 \le \frac{2\|\theta_1\|_2/c}{T^{1-\alpha}} + \frac{2(L_T + h_T/c)}{T^{1-\alpha}} \to 0, \quad \text{as } T \to \infty. \tag{63}$$

Remark 26 *In the gradient equilibrium results (57), (59), (61), (63) in the proposition, we have focused on decaying step sizes of the form $\eta_t = ct^{-\alpha}$, in order to make the conclusions salient. However, we note that the iterate bounds (56), (58), (60), (62) combined with (54) will lead to gradient equilibrium guarantees for a broader class of step size schedules, including adaptive ones.*

Remark 27 *Another reason to focus on decaying step sizes is that it allows us to fluidly draw in standard, parallel results on regret. For example, by Theorem 3.1 in Hazan (2016), we know that online projected gradient descent for L-Lipschitz losses ℓ_t, $t = 1, 2, 3, \ldots$, with step sizes $\eta = D/(L\sqrt{t})$, $t = 1, 2, 3, \ldots$, where D is the diameter of the constraint set, leads to an average regret bound*

$$\frac{\text{Regret}_T}{T} \leq \frac{3}{2}\frac{DL}{\sqrt{T}} \to 0, \quad \text{as } T \to \infty.$$

Our theory does not require bounded constraint sets, but it does extend to constraints and projections; see Appendix B. The topline conclusion is that gradient equilibrium and no regret are not at odds, and gradient descent with decaying step sizes can achieve them both simultaneously. A secondary conclusion is that our restorative theory may be itself helpful for regret analysis in unbounded domains, by virtue of its control over the iterate norms in gradient descent.

5. Experiments

In this section, we present a set of worked examples. These work together to showcase the generality of the gradient equilibrium framework laid out in previous sections. Generally, our examples will follow a common structure: we define a base model (which may be a pre-trained or trained online), and a parameter sequence θ_t, $t = 1, 2, 3, \ldots$ that we will learn online. Then, we define a loss function over these parameters and state a resulting gradient equilibrium guarantee. Finally, we examine a given dataset and present application-specific plots and interpretations to showcase the utility of the method. Code to reproduce all of our experiments is available at: `https://github.com/aangelopoulos/gradient-equilibrium`.

Henceforth we will denote by f_t, $t = 1, 2, 3, \ldots$ a sequence of predictions of a ground-truth label sequence y_t, $t = 1, 2, 3, \ldots$. Also, z_t, $t = 1, 2, 3, \ldots$ denote features of interest, which are not necessarily the same as the ones used to make the predictions. They could be, for example, societally-meaningful covariates such as age, race, sex, etc. Also, we note that in many of our experiments, we test the following ad hoc settings of the learning rate: $\{0, 0.001, 0.01, 0.05\}$. While there is a substantial literature on choosing learning rates in optimization, we do not pursue such methods here.

5.1 Datasets

COMPAS dataset. The Correctional Offender Management Profiling for Alternative Sanctions (COMPAS) Recidivism and Racial Bias dataset is a widely-used benchmark for evaluating algorithmic fairness and bias in predictive models. It contains information about individuals arrested in Broward County, Florida, and it includes features such as criminal history, demographics, and a recidivism risk score, which is calculated by the COMPAS algorithm for the purpose of making pretrial bail release decisions. One of the key challenges in this dataset is the presence of significant racial disparities in the COMPAS risk scores, which has led to concerns about the fairness and bias of the COMPAS predictions.

Figure 7 plots the racial distribution of the COMPAS data, along with the predictive bias. The predicted recidivism rate is obtained by taking the COMPAS recidivism score, which lies on an integer scale from 1 to 10, and dividing it by 10. We preprocess the dataset

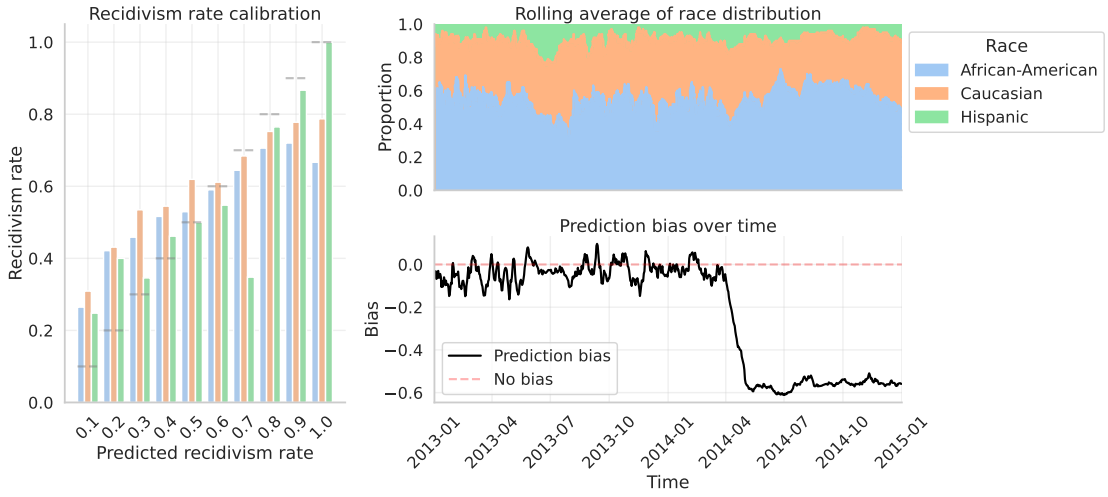


Figure 7: *Statistics of the COMPAS dataset. On the left-hand side is a calibration plot showing the predicted recidivism rate on the horizontal axis and the true recidivism rate—conditionally on the predicted one—on the vertical axis. The COMPAS algorithm tends to overpredict recidivism for African-Americans as the predicted recidivism rate grows, as compared to Caucasians and Hispanics. On the top right, we show a rolling average of the race distribution over time, with window size 100 (in terms of individuals screened). On the bottom right, we show a rolling average of the prediction bias in time, again with window size 100. A positive bias indicates an overprediction of recidivism. Thus, in general, the algorithm is underpredicting recidivism. The drastic drop in bias towards the end is an artifact of the dataset—the distribution changes drastically so that almost all individuals screened are recidivist.*

to remove rare groups, and only consider the Caucasian, African-American, and Hispanic subgroups. Together, these groups make up over 90% of the dataset.

HelpSteer2 dataset. The HelpSteer2 dataset is an open-source dataset specifically designed for training and fine-tuning reward models in alignment tasks, particularly for language models. This dataset contains around 10,000 response pairs for training and 1,000 for validation, each annotated for helpfulness, correctness, coherence, complexity, and verbosity on a Likert scale. Figure 8 provides two examples from this dataset. We post-train a Gemma2-2B (Gemma Team et al., 2024) model to predict helpfulness on the training set, and use the validation set for the purposes of our experiments. The goal is to debias the Gemma2-2B reward model using this data, in an online fashion.

Chatbot Arena dataset. Chatbot Arena (Chiang et al., 2024) is an open-source dataset of conversations between humans and anonymized pairs of large language models (LLMs). At the end of the conversation, the human votes for which response they prefer, encoded as a binary preference. This binary preference data is used to perform a Bradley-Terry regression (Bradley and Terry, 1952), whose coefficients estimate each AI model's relative "strength" or

Prompt:	Think of 5 fairly uncommon words but don't tell me what they are. Ask me for the definition of any one of these words. Wait for my answer.
Response:	Sure, here are five fairly uncommon words: 1. Ephemeral - lasting only a short time; short-lived. 2. Esoteric - intended for or understood by only a small group of people with a specialized knowledge or interest. 3. Elusive - difficult to find or capture; elusive. 4. Perennial - lasting or existing for a long or indefinite time. 5. Serendipitous - happening by chance; accidental.
Helpfulness:	0/4

Prompt:	config static ip address in ubuntu
Response:	To configure a static IP address in Ubuntu, follow these steps: 1. Open the Network Manager 2. Click the "Edit" button to modify. 3. Go to the "IPv4 Settings" tab. 4. Select "Manual". 5. Click "Add" to add a new IP address. 6. Enter the static IP address, netmask, and gateway. 7. Click "Save" to apply the changes.
Helpfulness:	4/4

Figure 8: *Two examples from the HelpSteer2 dataset. We display a prompt, a response, and the helpfulness as judged by an expert human.*

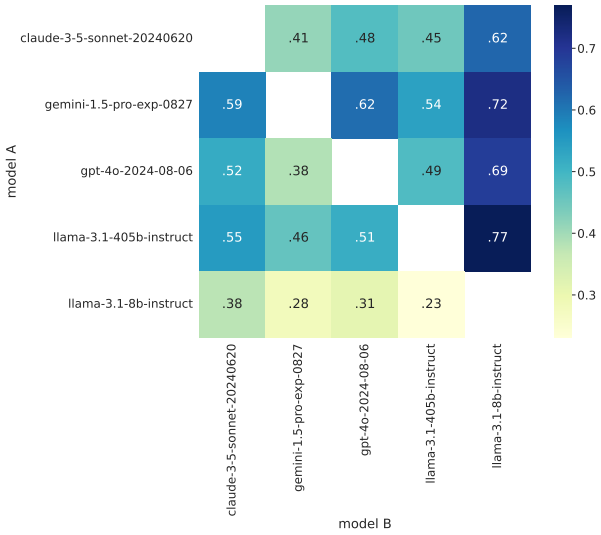


Figure 9: *Chatbot Arena win matrix. We display the win rates between five major models.*

capability. This allows for a leaderboard to be created which reflects how models perform against each other in direct comparisons, and predicts future preferences.

Bradley-Terry regression can be reframed as a particular form of logistic regression with binary features. Therefore, our online gradient descent theory immediately applies, and we can get a guarantee on an online variant of this regression. The online form of Bradley-Terry regression produces what are also known as Elo scores (Elo, 1967). Unlike the previous examples, here we are not debiasing a model post hoc, we are instead giving a guarantee on the Elo score, as it is normally computed. Being composed of human ratings, we note that this data set may possess some adversarial elements, more so than others considered in this section.

MIMIC dataset. The Medical Information Mart for Intensive Care (MIMIC-III) database is a publicly available, de-identified clinical dataset encompassing health information from over 40,000 patients admitted to the intensive care units of Beth Israel Deaconess Medical Center between 2001 and 2012. It includes data types such as demographics, vital signs, lab results, medications, caregiver notes, and mortality data. Our goal is length-of-stay

estimation: predicting a patient's length of stay in the hospital from their covariates at the time of admission. We use a gradient boosting tree as our base model, retrained periodically.

5.2 Simple debiasing

We consider the task of debiasing a sequence of predictions, so their average is approximately equal to the average of responses. The structure differs slightly when handling regression and classification problems.

Regression. In the regression setting, given real-valued predictions f_t, $t = 1, 2, 3, \ldots$, the goal is to produce a sequence of real-valued parameters θ_t, $t = 1, 2, 3, \ldots$, with each θ_t depending only on past data, such that (using $a_T \asymp b_T$ to mean $a_T - b_T \to 0$ as $T \to \infty$, as before):

$$\frac{1}{T}\sum_{t=1}^{T}(f_t + \theta_t) \asymp \frac{1}{T}\sum_{t=1}^{T} y_t. \tag{64}$$

In other words, the adjusted predictions $f_t + \theta_t$ should be unbiased for y_t, in the long run along the sequence. As discussed in Section 2.3.2, online gradient descent with the squared loss, $\ell_t(\theta) = \frac{1}{2}(y_t - f_t - \theta)^2$, achieves this goal under mild conditions. See Algorithm 1 below for pseudocode.

Algorithm 1 Simple debiasing (regression)

Input: Predictions $f_1, \ldots, f_T$; responses $y_1, \ldots, y_T$; learning rate $\eta > 0$

Initialize $\theta_1 = 0$

for $t = 1, \ldots, T$ **do**

  Compute the gradient: $g_t = f_t + \theta_t - y_t$

  Update the parameter: $\theta_{t+1} = \theta_t - \eta g_t$

Output: Adjusted predictions $f_1 + \theta_1, \ldots, f_T + \theta_T$

Figure 10 shows the results of this method on the HelpSteer2 dataset. As expected, gradient equilibrium holds, and thus the predictions are unbiased in the long run. Note also that the loss is not strongly impacted by the choice of learning rate.

Classification. In the classification setting, given $[0, 1]$-valued probabilistic predictions p_t, $t = 1, 2, 3, \ldots$ of binary labels, the goal is to produce a sequence of real-valued parameters θ_t, $t = 1, 2, 3, \ldots$, where each θ_t depends only on past data, such that:

$$\frac{1}{T}\sum_{t=1}^{T}\left(p_t + \frac{2e^{\theta_t}}{1 + e^{\theta_t}} - 1\right) \asymp \frac{1}{T}\sum_{t=1}^{T} y_t. \tag{65}$$

That is, the adjusted predictions $p_t + 2e^{\theta_t}/(1 + e^{\theta_t}) - 1$ should be unbiased for y_t, in the long run. As covered in Section 2.3.3, online gradient descent on the generalized logistic loss, $\ell_t(\theta) = -(y_t - p_t)\theta + 2\log(1 + e^{\theta}) - \theta$, achieves this goal under mild conditions. See Algorithm 2 below for pseudocode.

Algorithm 2 Simple debiasing (classification)

Input: Probabilistic predictions $p_t, \ldots, p_T$; binary labels $y_1, \ldots, y_T$; learning rate $\eta > 0$

1: Initialize $\theta_1 = 0$
2: **for** $t = 1, \ldots, T$ **do**
3: Compute the gradient: $g_t = p_y - y_t + 2\frac{e^{\theta_t}}{1+e^{\theta_t}} - 1$
4: Update the parameter: $\theta_{t+1} = \theta_t - \eta g_t$

Output: Adjusted predictions $p_1 + \frac{2e^{\theta_1}}{1+e^{\theta_1}} - 1, \ldots, p_T + \frac{2e^{\theta_T}}{1+e^{\theta_T}} - 1$

Figure 11 displays the results of this method on the COMPAS dataset. As expected, gradient equilibrium holds, thus the predictions are long-run unbiased. Again the loss is not significantly impacted by the choice of learning rate. We repeat this experiment for several choices of a decaying learning rate in Figure 12. The larger decay rates lead to slower convergence to gradient equilibrium (compared to one another) and smooth out the predictions.

5.3 Multigroup debiasing

We now turn to the debiasing of a predictor simultaneously over multiple, possibly overlapping groups. Let $z_t \in \{0,1\}^d$ be a vector of group indicators, that is, every element of z_t corresponds to some attribute of the data point observed at t, such as ethnicity, sex, marital status, and so on. Our goal is to achieve unbiasedness simultaneously for each of these attributes. As before, we will have slightly different approches to regression and classification.

Regression. In the regression setting, given real-valued predictions f_t, $t = 1, 2, 3, \ldots$ and group indicator vectors $z_t \in \{0,1\}^d$, $t = 1, 2, 3, \ldots$, the goal is produce a sequence of parameter vectors $\theta_t \in \mathbb{R}^d$, $t = 1, 2, 3, \ldots$, where each θ_t depends only on past data, such that:

$$\frac{1}{|I_j|}\sum_{t \in I_j}(f_t + z_t^\mathsf{T}\theta_t) \asymp \frac{1}{|I_j|}\sum_{t \in I_j} y_t, \quad \text{where } I_j = \{t \leq T : z_{tj} = 1\}, \tag{66}$$

for each $j = 1, \ldots, d$ such that $|I_j|$ grows linearly in T. Observe that (66) is the same as (64), but with $z_t^\mathsf{T}\theta_t$ in place of θ_t, and the sum being taken only over indexes where a particular group indicator is active. That is, the adjusted predictions $f_t + z_t^\mathsf{T}\theta_t$ should be long-run unbiased for y_t, for each j appearing often enough. As discussed in Section 2.3.4, when the groups are nonoverlapping, online gradient descent on the squared loss, $\ell_t(\theta) = \frac{1}{2}(y_t - f_t - z_t^\mathsf{T}\theta)^2$, achieves this goal under mild conditions; as discussed in Section 3.3, for arbitrary groups, a ridge-regularized version can be used. See Algorithm 3 below (for the unregularized version).

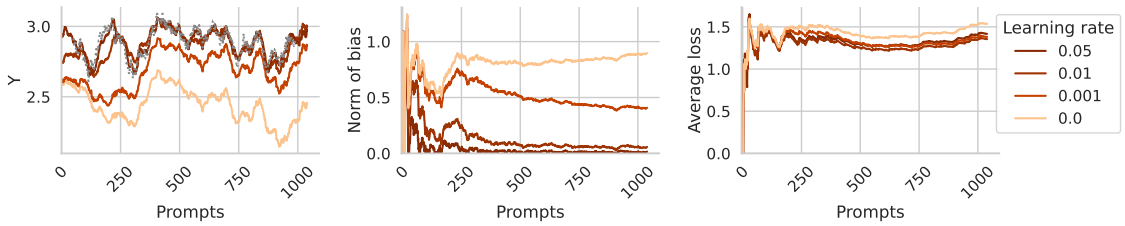


Figure 10: *Simple debiasing results on the HelpSteer2 dataset. On the left, we display a rolling average of the response as a gray dotted line along with the predictions in different colors, corresponding to the learning rate. In the middle, we show the norm of the bias of the adjusted predictor. It diminishes more quickly for larger learning rates. On the right, we plot the average loss of the adjusted predictor for different learning rates.*

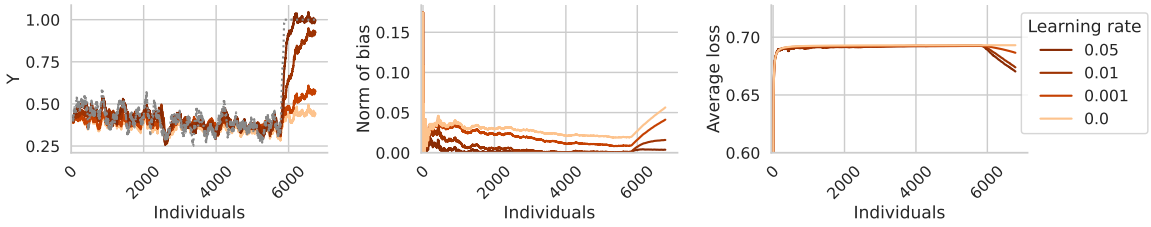


Figure 11: *Simple debiasing results on the COMPAS dataset. Same layout as Figure 10.*

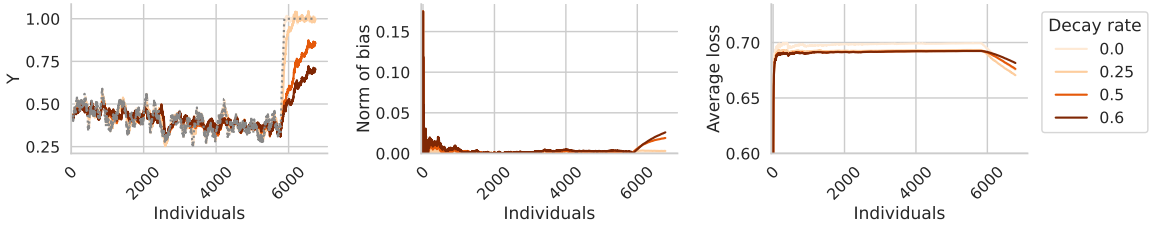


Figure 12: *Simple debiasing results on the COMPAS dataset, now with decaying learning rates.*

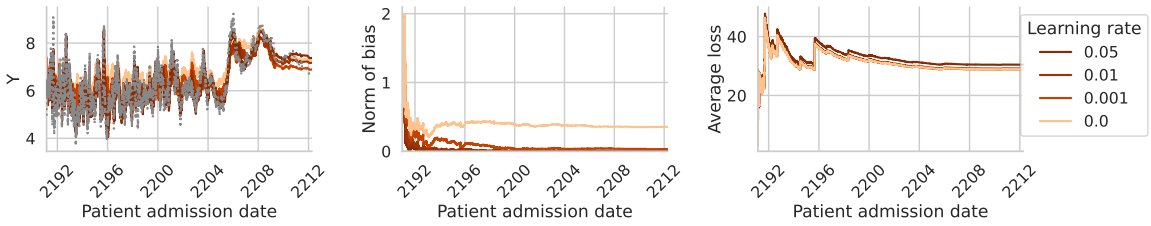


Figure 13: *Multigroup debiasing results on the MIMIC dataset. Same layout as Figure 10.*

Algorithm 3 Multigroup debiasing (regression)

Input: Predictions $f_1, \ldots, f_T$; group indicators $z_1, \ldots, z_T$; responses $y_1, \ldots, y_T$; learning rate $\eta > 0$

Initialize $\theta_1 = 0$
for $t = 1, \ldots, T$ **do**
  Compute the gradient: $g_t = z_t(f_t + z_t^\mathsf{T}\theta_t - y_t)$
  Update the parameter: $\theta_{t+1} = \theta_t - \eta g_t$

Output: Adjusted predictions $f_1 + z_1^\mathsf{T}\theta_1, \ldots, f_T + z_T^\mathsf{T}\theta_T$

Figure 13 shows marginal results of this procedure on the MIMIC dataset. Notice that the marginal bias also goes to zero, and the average loss remains essentially the same regardless of the learning rate. Figure 1 (in the introduction) displays the bias stratified by group. It can be seen that as compared to the base model, applying our method drives the group-stratified bias to zero within all groups. (It is also interesting to note that ethnicity and marital status form overlapping groups, yet in this case, gradient equilibrium is achieved even without regularization.)

Classification. In the classification setting, given $[0,1]$-valued predictions f_t, $t = 1, 2, 3, \ldots$ of binary labels, and group indicator vectors $z_t \in \{0,1\}^d$, $t = 1, 2, 3, \ldots$, the goal is produce a sequence of parameter vectors $\theta_t \in \mathbb{R}^d$, $t = 1, 2, 3, \ldots$, where each θ_t depends only on past data, such that:

$$\frac{1}{|I_j|}\sum_{t \in I_j}\left(p_t + \frac{2e^{z_t^\mathsf{T}\theta_t}}{1 + e^{z_t^\mathsf{T}\theta_t}} - 1\right) \asymp \frac{1}{|I_j|}\sum_{t \in I_j} y_t, \quad \text{where } I_j = \{t \le T : z_{tj} = 1\}, \tag{67}$$

for each $j = 1, \ldots, d$ such that $|I_j|$ grows linearly in T. Note (67) is the same as (65), but with $z_t^\mathsf{T}\theta_t$ in place of θ_t, and the sum being taken only over indexes where a particular group indicator is active. As before, the adjusted predictions $p_t + 2e^{\theta_t}/(1 + e^{\theta_t}) - 1$ should be unbiased for y_t should be long-run unbiased for y_t, for each j appearing often enough. As established in Section 2.3.4, when the groups are nonoverlapping, online gradient descent on the generalized logistic loss, $\ell_t(\theta) = -(y_t - p_t)z_t^\mathsf{T}\theta + 2\log(1 + e^{z_t^T\theta}) - z_t^\mathsf{T}\theta$, achieves this goal under mild conditions. As shown in Section 3.2, for arbitrary groups, a lasso-regularized version can be used. See Algorithm 4 below (for the unregularized version).

Algorithm 4 Multigroup debiasing (classification)

Input: Probabilistic predictions $p_1, \ldots, p_T$; group indicators $z_1, \ldots, z_T$; binary labels $y_1, \ldots, y_T$;
learning rate $\eta > 0$

Initialize $\theta_1 = 0$
for $t = 1, \ldots, T$ **do**
  Compute the gradient: $g_t = z_t\big(p_y - y_t + 2\frac{e^{\theta_t}}{1+e^{\theta_t}} - 1\big)$
  Update the parameter: $\theta_{t+1} = \theta_t - \eta g_t$

Output: Adjusted predictions $p_1 + \frac{2e^{z_1^\mathsf{T}\theta_1}}{1+e^{z_1^\mathsf{T}\theta_1}} - 1, \ldots, p_T + \frac{2e^{z_T^\mathsf{T}\theta_T}}{1+e^{z_T^\mathsf{T}\theta_T}} - 1$

Figures 14 and 15 show results of this procedure on the COMPAS dataset. Like previously, the gradient equilibrium conditions manifest, leading to zero long-run bias per group (improving on the COMPAS base model considerably). Meanwhile, the loss does not change much as a function of learning rate.

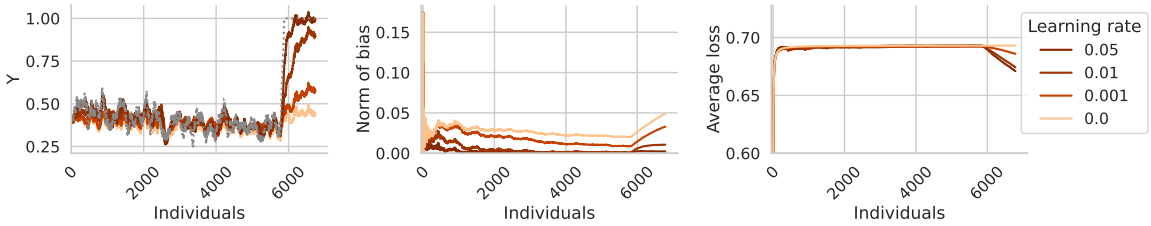


Figure 14: *Multigroup debiasing results on the COMPAS dataset. Same layout as Figure 10.*

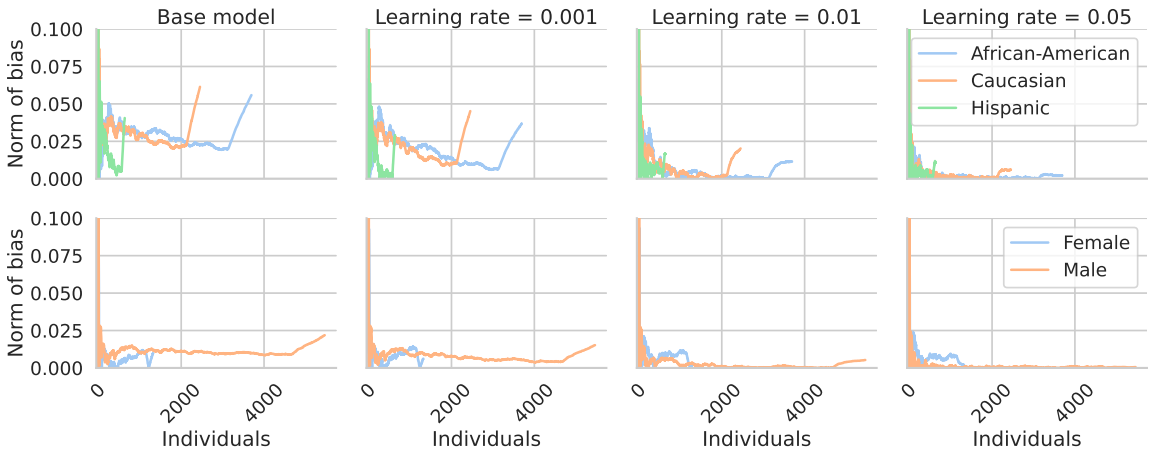


Figure 15: *Multigroup debiasing results on the COMPAS dataset. Same layout as Figure 1.*

5.4 Quantile tracking and ensembling

Another application of our framework is to choose the weights of an ensemble. Unlike the subsections above, this one does not involve debiasing. Instead, we focus on sequential quantile estimation, first recalling how to provide coverage using online gradient updates, then showing how to form an ensemble for more robust performance. In the setting of online conformal (Gibbs and Candès, 2021), we focus on quantile tracking as in (Angelopoulos et al., 2023). Given real-valued predictions f_t, $t = 1, 2, 3, \ldots$ and responses y_t, $t = 1, 2, 3, \ldots$, the goal is to construct real-valued parameters θ_t, $t = 1, 2, 3, \ldots$, where each θ_t depends only on past data, such that:

$$\frac{1}{T}\sum_{t=1}^{T} 1\{y_t \leq f_t + \theta_t\} \asymp \tau. \tag{68}$$

In other words, $f_t + \theta_t$ lies above y_t with frequency τ. We refer to this property as "coverage," insofar as it represents the ability to contain y_t with a one-sided interval. As discussed in

Section 2.3.1, online gradient descent with the quantile loss, $\ell_t(\theta) = \rho_\tau(y_t - f_t - \theta)$, achieves this goal under mild conditions. To be clear, this result is not original to the current paper, and is well-known in the online conformal literature. Also, these online gradient descent updates are sometimes referred to as quantile tracking.

As with nearly all online learning methods, practical deployments of quantile tracking require a choice of learning rate η. In some problems, the best choice of η may actually be nonstationary (time-varying). Initial efforts have been made to set the learning rate via mixture-of-experts schemes, as in (Zaffran et al., 2022; Gibbs and Candès, 2023). Here, we propose a different strategy: learning a mixture of quantile tracking iterates via online mirror descent. To fix notation, consider K learning rates, $\nu_1, \ldots, \nu_K$, and let θ_t^k denote the quantile tracking iterate under step size ν_k, at iteration t. We produce a mixed parameter estimate

$$\theta_t = w_t^\mathsf{T}(\theta_t^1, \ldots \theta_t^K),$$

where $w_t \in \mathbb{R}^K$ is a weight vector living in the standard K-dimensional probability simplex. Furthermore, we learn w_t $t = 1, 2, 3, \ldots$ by online mirror descent on the sequence of losses $\ell_t(w) = \rho_\tau(y_t - f_t - w^\mathsf{T}(\theta_t^1, \ldots \theta_t^K))$, $t = 1, 2, 3, \ldots$. Algorithm 5 gives pseudocode. Mirror descent is formally studied in Appendix B, where it is shown that it satisfies a form of gradient equilibrium. How this might relate to a rigorous coverage statement, as in (68) but at the ensemble level, is left to future work.

Algorithm 5 Quantile ensembling

Input: Predictions $f_1, \ldots, f_T$; responses $y_1, \ldots, y_T$; quantile level τ; base learning rates $\nu_1, \ldots, \nu_K$;
ensemble learning rate ν

Initialize w_1 to be uniform on the simplex
Initialize $\theta_1^k = 0$, $k = 1, \ldots, K$
Initialize $\theta_1 = 0$
for $t = 1, \ldots, T$ **do**
  Compute signed errors: $\sigma_t^k = 1\{y_t \leq f_t + \theta_t^k\} - \tau$, $k = 1, \ldots, K$ and $\epsilon_t = 1\{y_t \leq f_t + \theta_t\} - \tau$
  Compute quantile updates: $\theta_{t+1}^k = \theta_t^k - \nu_k \sigma_t^k$, $k = 1, \ldots, K$
  Compute ensemble update:

$$z_{t+1,k} = w_{tk} \exp(-\eta \theta_t^k \epsilon_t), \quad k = 1, \ldots, K$$
$$w_{t+1} = z_{t+1} / \|z_{t+1}\|_1$$

  Compute ensemble prediction: $\theta_{t+1} = w_{t+1}^\mathsf{T}(\theta_{t+1}^1, \ldots, \theta_{t+1}^K)$

Output: Adjusted quantile predictions $f_1 + \theta_1, \ldots, f_T + \theta_T$

In Figure 16, we present results of this procedure for predicting quantiles of the length-of-stay variable on the MIMIC dataset. In the middle row, we see that the coverage gap diminishes quickly (as expected by the gradient equilibrium guarantees) for quantile tracking as we increase the learning rate, but this comes at the expense of increasing the quantile loss. The top row shows that the ensemble achieves coverage faster than many of the individual

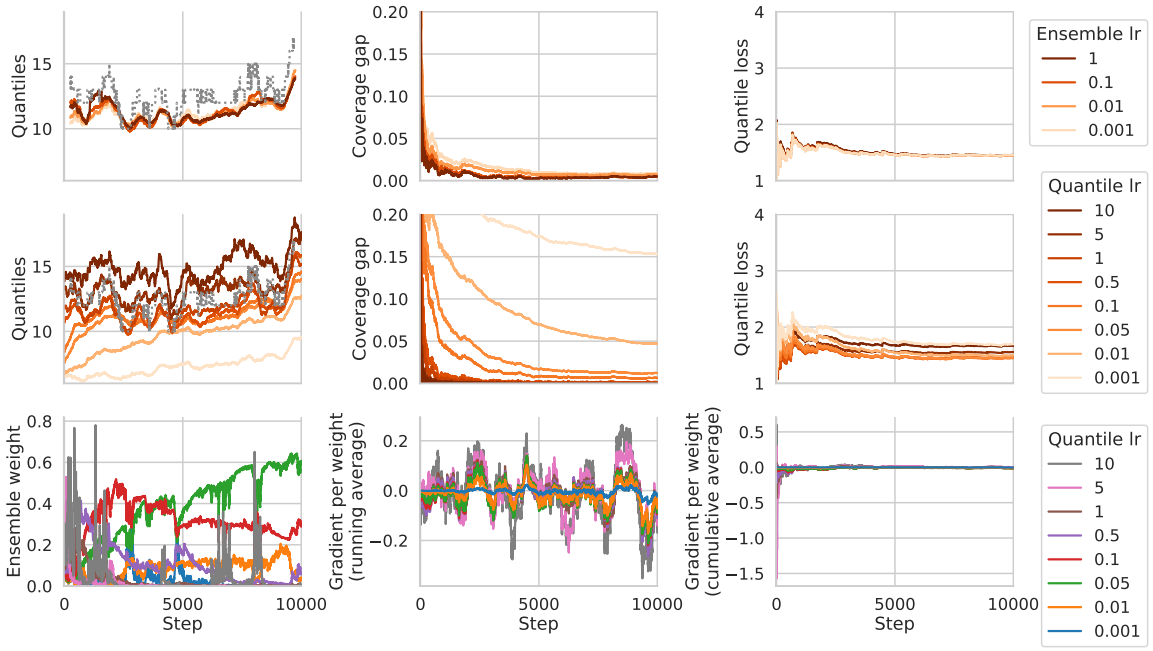


Figure 16: *Quantile ensembling results on the MIMIC dataset. We show results of our procedure learning to balance (via weights) different choices of learning rates, in quantile tracking. The top row follows the format of Figure 10. The middle row also follows this format, but this time with respect to different quantile tracker learning rates as opposed to different ensemble learning rates. The last row shows the behavior and properties of the weights for the quantile learning rates, along the sequence.*

quantile trackers, while matching essentially the best individual quantile loss. This behavior is almost invariant to the ensemble learning rate. Finally, the third row reveals how the ensemble upweights or downweights individual quantile trackers over the sequence.

5.5 Pairwise preference prediction

As a last example, we use our gradient equilibrium framework to examine the well-known Elo rating system (Elo, 1967), often used in chess and other two-player competitions, as previewed in Figure 2 in the introduction. We use the Chatbot Arena dataset for this example, where we have access to M models and a sequence of T battles between pairs of models $a_t, b_t \in \{1, \ldots, M\}$, and we define $y_t = 1$ when a human rater votes for the response of model b_t over a_t, and $y_t = 0$ otherwise.

The Elo rating system assigns a strength to each model at each t, encapsulated by a parameter $\theta_t \in \mathbb{R}^M$. The Elo updates can be recast as online gradient descent with respect to the logistic loss function on a certain set of features. In particular, define a feature vector $z_t \in \mathbb{R}^M$ with $z_{tj} = -1$ if $j = a_t$, $z_{tj} = 1$ if $j = b_t$, and $z_{tj} = 0$ otherwise. Then the Elo updates are

$$\theta_{t+1} = \theta_t - \eta z_t(\sigma(z_t^\mathsf{T}\theta) - y_t), \quad t = 1, 2, 3, \ldots,$$

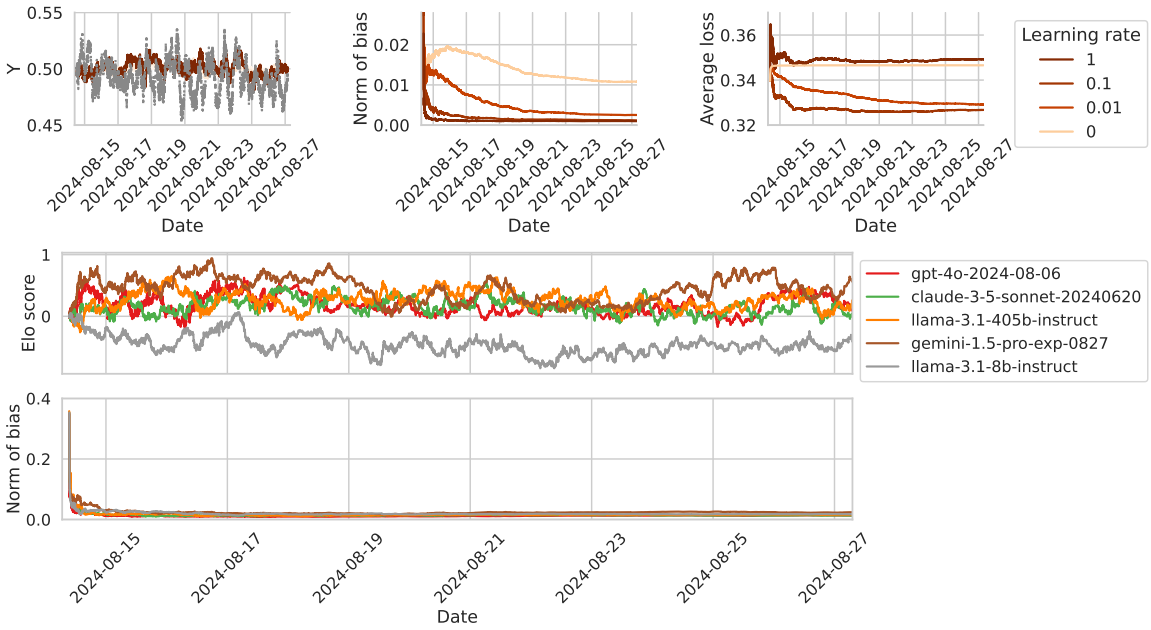


Figure 17: *Results on Chatbot Arena with regularization. Same format as Figure 2.*

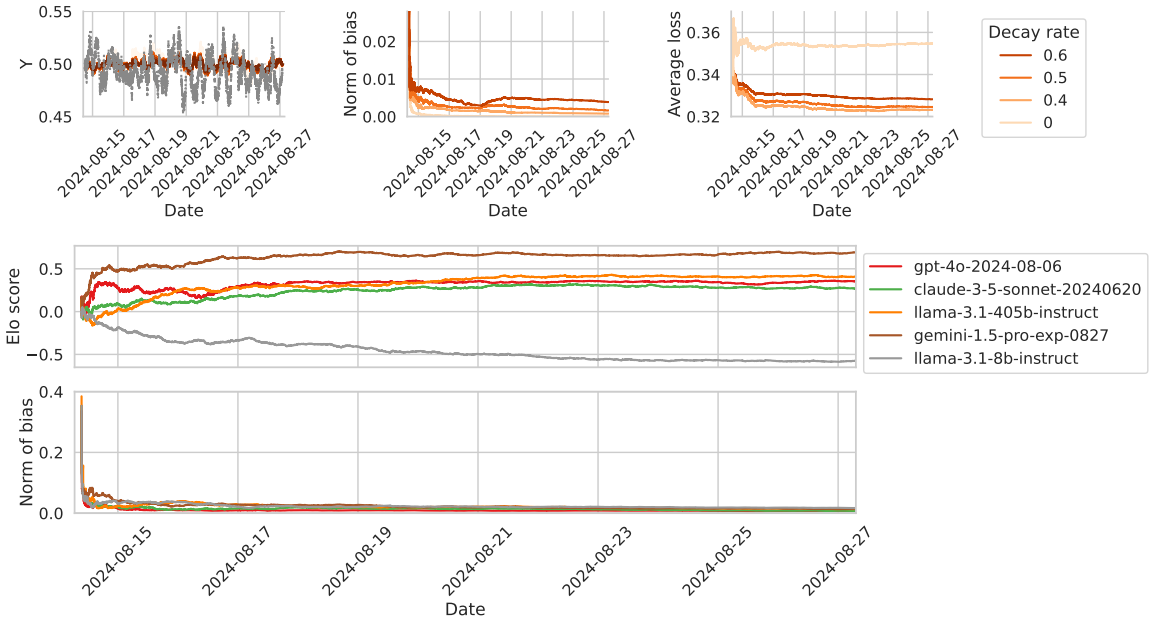


Figure 18: *Results on Chatbot Arena with decaying learning rates. Same format as Figure 2. For the bottom plots, we set a learning rate decay of* $\alpha = 0.5$.

where $\sigma(x) = e^x/(1+e^x)$ is the sigmoid function. Gradient equilibrium for this problem translates into the statement that

$$\frac{1}{|I_m|}\sum_{t\in I_m} p_t^m - \frac{1}{|I_m|}\sum_{t\in I_m} y_t^m \to 0, \quad \text{as } T\to\infty, \text{ where } I_m = \{t \le T : a_t = m \text{ or } b_t = m\},$$

for each model m such that $|I_m|$ grows linearly in T, where $p_t^m = p_t$ and $y_t^m = y_t$ if $b_t = m$, or $p_t^m = 1 - p_t$ and $y_t^m = 1 - y_t$ if $a_t = m$. This guarantee says that the Elo score gives unbiased win-rate predictions for all players appearing often enough, even in an adversarial online setting.

Recall, our most basic theory in Proposition 4 says that the gradient equilibrium holds whenever the Elo iterates remain bounded or slowly growing. As shown in Section 3.2, this can be achieved by adding in lasso regularization to the updates, but it is possible a more refined analysis would show that the original updates themselves lead to bounded or slowly growing iterates in general.

Figure 2 (in the introduction) shows results of the Elo procedure on Chatbot Arena, confirming that the gradient equilibrium guarantee can hold in practice, without any modification to the updates. However, this plot also underscores the importance of choosing a good learning rate for these sorts of problems; the largest learning rate, which reaches equilibrium the fastest, also has the highest loss. Meanwhile, the second largest learning rate present a much more favorable tradeoff. In Figures 17 and 18, we run two further experiments, with regularization and decaying learning rates, respectively. Regularization (a lasso penalty with $\lambda = 0.001$) leads to a slight smoothing of the sequence of predictions, but it comes at a cost: the bias does not go to zero. Instead, it is driven to a small positive value. With decaying learning rates we now get much smoother Elo scores, and the bias is still driven to zero, albeit more slowly than with constant learning rates.

6. Discussion

This paper has introduced gradient equilibrium as a property of interest for online learning algorithms in the standard adversarial sequence model. Online gradient descent with constant step sizes (Section 2) will lead to gradient equilibrium whenever its iterates remain bounded or grow slowly, and iterates grow slowly under a condition on the loss sequence that we refer to as the restorative property. If this property does not hold, then a careful use of regularization (Section 3) can supply it and minimally perturb the gradient equilibrium condition. Decaying step sizes (Section 4) will lead to gradient equilibrium at a slower rate, but at the same time accommodates other objectives, such as no regret. The analysis can be generalized to handle constraints (Appendix B), and to cover online proximal mirror descent, a class of methods which includes mirror descent, projected gradient descent, and proximal gradient descent. The study of gradient equilibrium has practical implications for machine learning (Section 5): gradient equilibrium can help us debias black-box predictions, calibrate quantiles, learn ensemble weights, prove guarantees on the Elo score, and more.

There are a number of possible directions for further exploration that are related to gradient equilibrium. In general, gradient descent with a large step size rapidly converges to gradient equilibrium, but in the process may produce a parameter sequence θ_t, $t = 1, 2, 3, \ldots$ that is volatile (i.e., nonsmooth in t). From the point of view of correcting predictions, as in

the debiasing application (or other ones considered in this paper), this may not be desirable. One might thus ask if we can achieve gradient equilibrium quickly, while maintaining a smooth parameter sequence. If we measure smoothness using the path-length metric: $\frac{1}{T}\sum_{t=1}^{T-1} \|\theta_t - \theta_{t+1}\|_2$, which is common in dynamic regret analyses, then a simple calculation reveals that with gradient descent, to balance gradient equilibrium and path-length we must take decaying (not constant) step sizes. Still, gradient descent may not be the best tool for balancing these two objectives, and other online learning methods may balance them better. This would be an interesting topic for future work, as would investigations into lower bounds (optimality) relating to gradient equilibrium, more generally.

Other directions that might be interesting to study from the perspective of gradient equilibrium involve building connections to calibration and control theory. For example, it is possible that various probabilistic calibration conditions can be expressed as gradient equilibria corresponding to particular losses, or that the techniques in our paper can otherwise be used to recover calibration guarantees. On the control front, a key point worth making is that our theory does not actually depend on g_t being a gradient (or subgradient), and simply analyzes its action as a map from vectors to vectors; recall the use of the restorative negative gradient field condition. This means that gradient descent updates could be applied and analyzed well outside of a setting motivated by loss minimization, for example, when we seek to hit a fixed point (on average) of some operator, which may be useful for applications in control or other fields.

Acknowledgments

We are grateful to Sivaraman Balakrishnan, Francis Bach, Emmanuel Candès, Isaac Gibbs, Jonathan Taylor, and Eric Zhao for various helpful discussions, and Isaac and Eric for helping us to identify related work. We thank Evan Frick for help setting up the HelpSteer2 experiment, and we thank Hyunsuk Kim for finding and fixing a bug in an earlier version of the quantile ensembling experiment. Lastly, we thank the associate editor and reviewers of this work for their thoughtful feedback.

MIJ was funded by the European Union, ERC-2022-SYG-OCEAN-101071601. Views and opinions expressed are however those of the author(s) only and do not necessarily reflect those of the European Union or the European Research Council Executive Agency. Neither the European Union nor the granting authority can be held responsible for them. MIJ was also funded by the Chair "Markets and Learning," supported by Air Liquide, BNP PARIBAS ASSET MANAGEMENT Europe, EDF, Orange and SNCF, sponsors of the Inria Foundation. RJT was funded by the Office of Naval Research, ONR grant N00014-20-1-2787.

Appendix A. Proofs

A.1 Proof of Proposition 3

Define $\theta_t = y_t + s_T$, $t = 1, \ldots, T$. Then

$$\frac{1}{T}\sum_{t=1}^{T}(y_t - \theta_t)^2 = s_T^2,$$

and

$$\frac{1}{T}\sum_{t=1}^{T}\theta_t = \bar{y}_T + s_T,$$

which proves the claim.

A.2 Proof of Proposition 6

For convenience, redefine each $h_t = \max\{|\theta_1|, h_t\}$, and let $h_0 = |\theta_1|$. We will use induction to prove (9). The base case, for $T = 0$, is trivial. Assume that the result is true up through T. We break up the argument for $T + 1$ into cases. If $|\theta_T| \le h_T$, then note that by the triangle inequality

$$\begin{aligned} |\theta_{T+1}| &\le |\theta_T| + \eta|g_T(\theta_T)| \\ &\le h_T + \eta L, \end{aligned}$$

where the second line uses L-Lipschitzness of ℓ_t. If instead $|\theta_{T+1}| > h_T$, then we further divide this case into two subcases. If $\theta_{T+1} > h_T$, then by (8), we have $\theta_{T+1} \le \theta_T$, and moreover, $\theta_{T+1} \ge h_T - \eta g_T(\theta_T) \ge -\eta L$, using L-Lipschitzness (and $h_T \ge 0$). Thus, putting these bounds together, we have $\theta_{T+1} \in [-\eta L, \theta_T]$. In the case $\theta_{T+1} < -h_T$, similar arguments lead to $\theta_{T+1} \in [\theta_T, \eta L]$. Thus, when $|\theta_{T+1}| > h_T$, we have shown

$$\theta_{T+1} \in \big[\min\{\theta_T, -\eta L\}, \max\{\theta_T, \eta L\}\big].$$

By the inductive hypothesis, this gives $|\theta_{T+1}| \le h_{T-1} + \eta L \le h_T + \eta L$, using the nondecreasing property of h_T. This completes the inductive step and proves (9). The average gradient bound (10) now follows from (6), and then using the simple inequality $\max\{a, b\} \le a + b$ to write the bound more cleanly.

A.3 Proof of Proposition 7

Taking the squared norm on both sides of the update (4) and expanding yields

$$\begin{aligned} \|\theta_{T+1}\|_2^2 &= \|\theta_T\|_2^2 + \eta^2\|g_T(\theta_T)\|_2^2 - 2\eta g_T(\theta_T)^\mathsf{T}\theta_T \\ &\le \|\theta_T\|_2^2 + \eta^2 L^2 - 2\eta g_T(\theta_T)^\mathsf{T}\theta_T \\ &\le \|\theta_1\|_2^2 + \eta^2 L^2 T - 2\eta\sum_{t=1}^{T} g_t(\theta_t)^\mathsf{T}\theta_t, \end{aligned} \tag{69}$$

where the second line uses L-Lipschitzness of the loss (or equivalently, $\|g_t\|_2 \leq L$), and the third unravels the iteration over $t = 1, \ldots, T$. Next we show each summand on last line above is bounded by

$$-2\eta g_t(\theta_t)^\mathsf{T}\theta_t \leq 2\eta L h_t. \tag{70}$$

We split the argument into cases. If $\|\theta_t\|_2 \leq h_t$, then the upper bound (70) follows from the Cauchy-Schwarz inequality and L-Lipschitzness. If $\|\theta_t\|_2 > h_t$, then the restorative condition (7) provides the stronger upper bound $-\ell_t(\theta_t)^\mathsf{T}\theta_t \leq 0$. This completes the proof of (70), and combining this with (69), then taking a square root, proves (11). The second result (12) follows from $\sum_{t=1}^T h_T \leq T h_T$, as h_T is nondecreasing, applying (6), and then using $\sqrt{a+b} \leq \sqrt{a} + \sqrt{b}$ to write the bound more cleanly.

A.4 Proof of Proposition 8

We use induction, as in the proof of Proposition 6. As in that proof, redefine each $h_t = \max\{\|\theta_1\|_2, h_t\}$ and let $h_0 = \|\theta_1\|_2$. The base case, for $T = 0$, is trivial. Assume that the result is true up through T. We break up the argument for $T+1$ into cases. If $\|\theta_T\|_2 \leq h_T$, then note that by the triangle inequality

$$\begin{aligned}\|\theta_{T+1}\|_2 &\leq \|\theta_T\|_2 + \eta\|g_T(\theta_T)\|_2 \\ &\leq h_T + \eta L,\end{aligned}$$

where the second line uses L-Lipschitzness of ℓ_T. If instead $\|\theta_T\|_2 > h_T$, then

$$\begin{aligned}\|\theta_{T+1}\|_2^2 &= \|\theta_T\|_2^2 + \eta^2\|g_T(\theta_T)\|_2^2 - 2\eta g_T(\theta_T)^\mathsf{T}\theta_T \\ &\leq \|\theta_T\|_2^2 + \eta^2 L^2 - 2\eta g_T(\theta_T)^\mathsf{T}\theta_T \\ &\leq (h_{T-1} + \eta L)^2 + \eta^2 L^2 - 2\eta\phi_T(\theta) \\ &\leq (h_T + \eta L)^2,\end{aligned}$$

where the second line uses L-Lipschitzness of ℓ_T, the third uses the inductive hypothesis and the restorative condition (7) on ℓ_T, and the fourth uses the nondecreasing property of h_T and condition (13) on $\phi_T(\theta)$ (which implies that $\eta^2 L - 2\eta\phi_T(\theta) \leq 0$). Taking a square root proves (14). The second result (15) is implied by (6), and using $\max\{a, b\} \leq a + b$ to write the bound more cleanly.

A.5 Proof of Proposition 9

This proof is similar to that of Proposition 9, but it differs slightly in the inductive argument for $T+1$. We again divide into two cases. If $\|\theta_T\|_2 \leq h_T$, then the triangle inequality and *local* L_T-Lipschitzness of ℓ_T, as in (16), implies that $\|\theta_{T+1}\|_2 \leq h_T + \eta L_T$. If instead $\|\theta_T\|_2 > h_T$, then expanding the squared norm of the gradient update, and using the inductive hypothesis and nondecreasingness of h_T, L_T, gives

$$\|\theta_{T+1}\|_2^2 \leq (h_T + \eta L_T)^2 + \eta^2\|g_T(\theta_T)\|_2^2 - 2\eta g_T(\theta_T)^\mathsf{T}\theta_T.$$

The restorative condition (7) with quadratic curvature (17) implies

$$\eta^2\|g_T(\theta_T)\|_2^2 - 2\eta g_T(\theta_T)^\mathsf{T}\theta_T \leq \eta^2\|g_T(\theta_T)\|_2^2 - 2\eta\phi_T(\theta) \leq 0,$$

and thus $\|\theta_{T+1}\|_2 \leq h_T + \eta L_T$, completing the proof of (18).

A.6 Proof of Corollary 10

Observe

$$\begin{aligned}\frac{1}{T}\sum_{t=1}^{T} g_t(\theta_t) &= \frac{1}{T}\Big(-\tau \sum_{t=1}^{T} 1\{y_t > \theta_t\} + (1-\tau)\sum_{t=1}^{T} 1\{y_t \le \theta_t\}\Big) \\ &= \frac{1}{T}\sum_{t=1}^{T} 1\{y_t \le \theta_t\} - \tau,\end{aligned}$$

where the second line uses the assumption that $g_t(\theta_t) = -\tau$ whenever $\theta_t = y_t$. The result follows directly from (10).

A.7 Proof of Corollary 11

Fix any $|\theta| > h_t$. Let $\mu = \eta/2$, and $b_t = |y_t|$. As highlighted in the display before the proposition, we seek to prove that

$$\theta(\theta - y_t) \ge \mu(y_t - \theta)^2.$$

It suffices to show that

$$|\theta|(|\theta| - b_t) \ge \mu(b_t + |\theta|)^2,$$

or

$$\mu \le \frac{|\theta|(|\theta| - b_t)}{(b_t + |\theta|)^2} = \underbrace{\frac{|\theta|^2}{(b_t + |\theta|)^2}}_{f_+(\theta)} - \underbrace{\frac{b_t|\theta|}{(b_t + |\theta|)^2}}_{f_-(\theta)}.$$

Now we will minimize $f_+(\theta)$ and maximize $f_-(\theta)$ over $|\theta| \ge h_t$. Clearly, the minimum of $f_+(\theta)$ occurs at the boundary of this range, $|\theta| = h_t$, and the minimum value is $h_t^2/(b_t + h_t)^2$. For $f_-(\theta)$, rewrite this as

$$f_-(\theta) = \frac{b_t}{(b_t/\sqrt{|\theta|} + \sqrt{|\theta|})^2}.$$

Observe that its unconstrained maximum is obtained by globally minimizing the denominator, which occurs at $|\theta| = b_t$. As $f_-(\theta)$ is monotone increasing in $|\theta|$ to the right of b_t (and also $h_t > b_t$), its maximum over the range $|\theta| \ge h_t$ again occurs at $|\theta| = h_t$, and the maximum value is therefore $b_t h_t/(b_t + h_t)^2$. Putting the last two results together (on the extrema of f_+ and f_-), we see that it suffices to have

$$\mu \le \frac{h_t(h_t - b_t)}{(b_t + h_t)^2}.$$

Using $b_t \le \delta h_t$, the right-hand side above is lower bounded by $h_t^2(1-\delta)/(\delta h_t + h_t)^2 = (1-\delta)/(1+\delta)^2$, which completes the proof of the first part of the proposition. For the second part, we simply note that the gradient of the squared loss is bounded by $|\theta_t - y_t| \le h_t + b_t$ for $|\theta| \le h_t$ and $|y_t| \le b_t$. The rest is a direct application of (19).

A.8 Proof of Corollary 12

We must verify that

$$(b-a)\frac{e^\theta}{1+e^\theta} + a \leq y_t, \quad \text{for all } \theta < -h,$$
$$(b-a)\frac{e^\theta}{1+e^\theta} + a \geq y_t, \quad \text{for all } \theta > h.$$

Note that the function $\theta \mapsto (b-a)e^\theta/(1+e^\theta) + a$ is an increasing bijection from $\mathbb{R}$ to $[a,b]$, whose inverse is $u \mapsto \log((u-a)/(b-u))$, so the first line above holds for $-h = \log((y_t-a)/(b-y_t))$, while the second line holds for $h = \log((y_t-a)/(b-y_t))$. Taking the max of these values two proves the first claim in the corollary. The remaining claims simply follow from noting $y_t \in [a+\epsilon_t, b-\epsilon_t]$, where $\epsilon_t \in (0,(b-a)/2)$, implies

$$\max\left\{\log\frac{y_t-a}{b-y_t}, \log\frac{b-y_t}{y_t-a}\right\} \leq \log\frac{b-a}{2\epsilon_t},$$

and the rest is a direct application of (10).

A.9 Proof of Corollary 13

Recalling (27), we can rewrite this as parallel streams of GD processes, one for each $j = 1,\ldots,d$:

$$\vartheta_{i_j(t+1),j} = \vartheta_{i_j(t),j} + \eta(y_{i_j(t)} - \psi'(\vartheta_{i_j(t),j})), \quad t = 1,2,3,\ldots,$$

where $i_j(t)$ is the index at which j^{th} feature is visited for the t^{th} time, in the sense that $|I_j(i_j(t))| = t$. We can then apply Corollaries 11 or 12 to the above GD iterations (which are the standard GD iterations studied in these corollaries, but we have simply reparametrized the index from $t = 1,2,3,\ldots$ to $i_j(t)$, $t = 1,2,3,\ldots$) to establish the desired results.

A.10 Proof of Proposition 14

We will prove the result in parts (a) and (b) separately from that in part (c).

Parts (a) and (b). Fix any $\|\theta\|_2 > h_t$. Rewrite the restricted α_t-strong monotonicity condition (38) as

$$g_t(\theta)^\mathsf{T}\theta \geq g_t(0)^\mathsf{T}\theta + \alpha_t\|\theta\|_2^2. \tag{71}$$

It suffices to prove that

$$g_t(0)^\mathsf{T}\theta + \alpha_t\|\theta\|_2^2 \geq \kappa,$$

where we will set $\kappa = 0$ for part (a) and $\kappa = \eta L^2/2$ for part (b). By the Cauchy-Schwarz inequality and the assumed bound on the gradient at the origin, $g_t(0)^\mathsf{T}\theta \geq -\|g_t(0)\|_2\|\theta\|_2 \geq -b_t\|\theta\|_2$, so it suffices to prove

$$\alpha_t\|\theta\|_2^2 - b_t\|\theta\|_2 - \kappa \geq 0.$$

We can treat this as a univariate quadratic inequality $q(x) \geq 0$ in $x = \|\theta\|_2$. The quadratic q has roots

$$x_- = \frac{b_t - \sqrt{b_t^2+4\kappa}}{2\alpha_t}, \quad \text{and} \quad x_+ = \frac{b_t + \sqrt{b_t^2+4\kappa}}{2\alpha_t} \leq \frac{b_t+\sqrt{\kappa}}{\alpha_t},$$

where the upper bound uses the simple inequality $\sqrt{a+b} \leq \sqrt{a} + \sqrt{b}$. As q is convex, it will be nonnegative to the right of its larger root x_+. Thus we can set h_t to be larger or equal to the right-hand side in the above display, which proves parts (a) and (b), after plugging in for κ appropriately.

Part (c). Fix any $\|\theta\|_2 > h_t$. Expand the right-hand side of the restricted β_t-co-coercivity condition (39) and rearrange to yield

$$\|g_t(\theta)\|_2^2 \leq \beta_t (g_t(\theta) - g_t(0))^\mathsf{T} \theta - \|g_t(0)\|_2^2 + 2 g_t(\theta)^\mathsf{T} g_t(0).$$

It suffices to prove that

$$g_t(\theta)^\mathsf{T} \theta \geq \frac{\eta}{2} \Big(\beta_t (g_t(\theta) - g_t(0))^\mathsf{T} \theta - \|g_t(0)\|_2^2 + 2 g_t(\theta)^\mathsf{T} g_t(0) \Big),$$

or equivalently

$$\bigg(1 - \frac{\eta \beta_t}{2}\bigg) g_t(\theta)^\mathsf{T} \theta + \frac{\eta \beta_t}{2} g_t(0)^\mathsf{T} \theta + \frac{\eta}{2} \|g_t(0)\|_2^2 - \eta g_t(\theta)^\mathsf{T} g_t(0) \geq 0.$$

Recall (71) from the restricted α_t-strong monotonicity condition. Assuming $\eta \leq 2/\beta_t$ so that $1 - \eta\beta_t/2 \geq 0$, we can multiply both sides of (71) by $1 - \eta\beta_t/2$ to show that the above display is implied by

$$g_t(0)^\mathsf{T} \theta + \alpha_t \bigg(1 - \frac{\eta \beta_t}{2}\bigg) \|\theta\|_2^2 + \frac{\eta}{2} \|g_t(0)\|_2^2 - \eta g_t(\theta)^\mathsf{T} g_t(0) \geq 0,$$

or equivalently,

$$g_t(0)^\mathsf{T} \theta + \alpha_t \bigg(1 - \frac{\eta \beta_t}{2}\bigg) \|\theta\|_2^2 - \frac{\eta}{2} \|g_t(0)\|_2^2 + \eta (g_t(0) - g_t(\theta))^\mathsf{T} g_t(0) \geq 0.$$

By the Cauchy-Schwarz inequality and the assumed bound on the gradient at the origin, it suffices to prove

$$\alpha_t \bigg(1 - \frac{\eta \beta_t}{2}\bigg) \|\theta\|_2^2 - b_t \|\theta\|_2 - \frac{\eta b_t^2}{2} + \eta (g_t(0) - g_t(\theta))^\mathsf{T} g_t(0) \geq 0.$$

For the last term, we use Cauchy-Schwarz, the restricted β_t-Lipschitz condition on the gradient (40), and the gradient bound at the origin, to lower bound $(g_t(0) - g_t(\theta))^\mathsf{T} g_t(0) \geq -\|g_t(0) - g_t(\theta)\|_2 \|g_t(0)\|_2 \geq -\beta_t b_t \|\theta\|_2$. Thus it suffices to prove

$$\alpha_t \bigg(1 - \frac{\eta \beta_t}{2}\bigg) \|\theta\|_2^2 - b_t (1 + \eta\beta_t) \|\theta\|_2 - \frac{\eta b_t^2}{2} \geq 0.$$

As before, we can treat this as a univariate convex quadratic inequality $q(x) \geq 0$ in $x = \|\theta\|_2$, and the larger of its two roots is

$$x_+ = \frac{b_t(1 + \eta\beta_t) + \sqrt{(b_t(1+\eta\beta_t))^2 + 2\eta b_t^2}}{2\alpha_t(1 - \eta\beta_t/2)} \leq \frac{b_t(1+\eta\beta_t) + \sqrt{\eta/2} \cdot b_t}{\alpha_t(1 - \eta\beta_t/2)},$$

Hence we can set h_t to be larger or equal to the right-hand side above, which completes the proof of part (c).

A.11 Proof of Lemma 17

Consider the standard logistic case with $a = 0$ and $b = 1$. Let $f(u; y) = (e^u/(1 + e^u) - y)u$. For any given u, this is linear in y, and hence the minimum over y must occur at an endpoint of the range $[0, 1]$. Therefore

$$f(u; y) \geq \min\{f(u; 0), f(u; 1)\}.$$

Each of $f(\cdot; 0), f(\cdot; 1)$ are smooth quasi-convex functions. Indeed $f(u; 0) = f(-u; 1)$, and thus it suffices to minimize, say, $f(\cdot; 1)$. This can be done by finding the root of its derivative, i.e., solving the transcendental equation $u = 1 + e^{-u}$. Applying the bisection method gives

$$\inf_u f(u; 1) = -0.2784645... \geq -0.279.$$

The result for general $a < b$ follows similarly.

A.12 Proof of Proposition 18

To verify that $\tilde{\ell}_t$ is Lipschitz, we compute its subgradient

$$\tilde{g}_t(\theta) = \left((b-a)\frac{e^{x_t^\mathsf{T}\theta}}{1+e^{x_t^\mathsf{T}\theta}} + a - y_t\right)x_t + \lambda s,$$

where $s \in \partial\|\theta\|_1$. Thus we see that

$$\|\tilde{g}_t(\theta)\|_2 \leq \left|(b-a)\frac{e^{x_t^\mathsf{T}\theta}}{1+e^{x_t^\mathsf{T}\theta}} + a - y_t\right|\|x_t\|_2 + \lambda\|s\|_2 \leq c(b-a) + \lambda\sqrt{d},$$

which proves the claim. Next, to check that $\tilde{\ell}_t$ satisfies the restorative property, we compute

$$\tilde{g}_t(\theta)^\mathsf{T}\theta = \left((b-a)\frac{e^{x_t^\mathsf{T}\theta}}{1+e^{x_t^\mathsf{T}\theta}} + a - y_t\right)x_t^\mathsf{T}\theta + \lambda\|\theta\|_1.$$

By Lemma 17, the first term above can be globally lower bounded:

$$\tilde{g}_t(\theta)^\mathsf{T}\theta \geq -0.279(b-a) + \lambda\|\theta\|_1.$$

To verify the restorative property with positive curvature, it suffices to show that for $\|\theta\|_2 \geq h_t$,

$$-0.279(b-a) + \lambda\|\theta\|_1 \geq \frac{\eta(c(b-a) + \lambda\sqrt{d})^2}{2}.$$

Since $\|\theta\|_1 \geq \|\theta\|_2$, it suffices to have

$$-0.279(b-a) + \lambda h_t \geq \frac{\eta(c(b-a) + \lambda\sqrt{d})^2}{2},$$

and rearranging gives the desired lower bound on h_t. Finally, setting h_t to this lower bound, we can apply (15) to yield

$$\left\|\frac{1}{T}\sum_{t=1}^{T}\left[\left((b-a)\frac{e^{x_t^\mathsf{T}\theta_t}}{1+e^{x_t^\mathsf{T}\theta_t}} + a - y_t\right)x_t + \lambda s_t\right]\right\|_2 \leq \frac{2\|\theta_1\|_2}{\eta T} + \frac{L}{T} + \frac{0.279(b-a) + \eta L^2/2}{\lambda\eta T},$$

where $s_t \in \partial\|\theta_t\|_1$. To verify (44) we use (42) and $\|s_t\|_2 \le \sqrt{d}$ to yield

$$\left\|\frac{1}{T}\sum_{t=1}^{T}\left((b-a)\frac{e^{x_t^\mathsf{T}\theta_t}}{1+e^{x_t^\mathsf{T}\theta_t}}+a-y_t\right)x_t\right\|_2 \le \frac{2\|\theta_1\|_2}{\eta T}+\frac{L}{T}+\frac{0.279(b-a)+\eta L^2/2}{\lambda\eta T}+\lambda\sqrt{d}.$$

This completes the proof.

A.13 Proof of Lemma 20

Let $f(u;y) = a_1(u-y)u - a_2(u-y)^2$. Observe that

$$f(u;y) = (a_1-a_2)u^2 - (a_1-2a_2)yu - a_2y^2,$$

which is a convex quadratic in u, provided that $a_1 > a_2$. Its minimizer is

$$\bar{u} = \frac{(a_1-2a_2)y}{2(a_1-a_2)},$$

and its minimum is

$$f(\bar{u};y) = -\frac{(a_1-2a_2)^2y^2}{4(a_1-a_2)} - a_2y^2,$$

and using $|y| \le b$ to lower bound this further gives the claimed result.

A.14 Proof of Proposition 21

To verify that $\tilde{\ell}_t$ is locally Lipschitz, we compute its gradient

$$\tilde{g}_t(\theta) = (x_t^T\theta - y_t)x_t + \lambda\theta.$$

Thus we see that for $\|\theta\|_2 \le h_t$,

$$\|\tilde{g}_t(\theta)\|_2 \le |x_t^T\theta - y_t|\|x_t\|_2 + \lambda\|\theta\|_2 \le (c_th_t+b_t)c_t + \lambda h_t,$$

which proves the claim. Next, to check that $\tilde{\ell}_t$ satisfies the restorative property, we compute

$$\tilde{g}_t(\theta)^\mathsf{T}\theta = (x_t^T\theta - y_t)x_t^\mathsf{T}\theta + \lambda\|\theta\|_2^2,$$

and

$$\frac{\eta}{2}\|\tilde{g}_t(\theta)\|_2^2 = \frac{\eta}{2}(x_t^T\theta-y_t)^2\|x_t\|_2^2 + \lambda\eta(x_t^T\theta-y_t)x_t^\mathsf{T}\theta + \frac{\lambda^2\eta}{2}\|\theta\|_2^2.$$

To verify the restorative property with quadratic curvature it suffices to show that for $\|\theta\|_2 \ge h_t$,

$$\underbrace{(1-\lambda\eta)(x_t^T\theta-y_t)x_t^\mathsf{T}\theta - \frac{\eta c_t^2}{2}(x_t^T\theta-y_t)^2}_{f(x_t^\mathsf{T}\theta;\,y_t)} + \lambda\left(1-\frac{\lambda\eta}{2}\right)\|\theta\|_2^2 \ge 0,$$

where $f(u;y) = (1-\lambda\eta)(u-y)u - (\eta c_t^2/2)(u-y)^2$. By Lemma 20, under the assumption $\eta < 1/(\lambda + c_t^2/2)$, we can globally lower bound f (first term), and we can also simply lower bound $\|\theta\|_2$ by h_t (second term), thus it suffices to have

$$-\frac{(1-\lambda\eta-\eta c_t^2)^2b_t^2}{4(1-\lambda\eta-\eta c_t^2/2)} - \frac{\eta c_t^2b_t^2}{2} + \lambda\left(1-\frac{\lambda\eta}{2}\right)h_t^2 \ge 0.$$

Rearranging gives the desired lower bound on h_t. Setting h_t to this lower bound, we can apply (19) to yield

$$\left\| \frac{1}{T} \sum_{t=1}^{T} \left[(x_t^T \theta_t - y_t) x_t + \lambda \theta_t \right] \right\|_2 \leq \frac{2\|\theta_1\|_2}{\eta T} + \frac{b_T c_T}{T} + \frac{\sqrt{C_{0T}(\lambda)}(\eta(c_T^2 + \lambda) + 1)}{\sqrt{\lambda}\eta T}.$$

where $C_{0T}(\lambda)$ is as in the proposition statement. Now we use (42) to yield

$$\begin{aligned}
&\left\| \frac{1}{T} \sum_{t=1}^{T} (x_t^T \theta_t - y_t) x_t \right\|_2 \\
\leq &\frac{2\|\theta_1\|_2}{\eta T} + \frac{b_T c_T}{T} + \frac{\sqrt{C_{0T}(\lambda)}(\eta(c_T^2 + \lambda) + 1)}{\sqrt{\lambda}\eta T} + \frac{\lambda}{T} \sum_{t=1}^{T} \|\theta_t\|_2 \\
\leq &\frac{2\|\theta_1\|_2}{\eta T} + \frac{b_T c_T}{T} + \frac{\sqrt{C_{0T}(\lambda)}(\eta(c_T^2 + \lambda) + 1)}{\sqrt{\lambda}\eta T} + \lambda(\max\{\|\theta_1\|_2, h_T\} + \eta L_T),
\end{aligned}$$

where the last line uses (18). Applying the upper bound

$$\lambda(\max\{\|\theta_1\|_2, h_T\} + \eta L_T) \leq \lambda(\|\theta_1\|_2 + \eta b_T c_T) + \sqrt{\lambda C_{0T}(\lambda)}(\eta(c_T^2 + \lambda) + 1),$$

leads to (48), and completes the proof.

Appendix B. Constraints

We extend the perspective on gradient equilibrium from the main text by considering losses ℓ_t, $t = 1, 2, 3, \ldots$ subject to a constraint $\theta \in C \subseteq \mathbb{R}^d$. In this section, we establish theory for online projected gradient descent and online mirror descent analogous that in Section 2 for online gradient descent. Our first step is to revisit and interpret the gradient equilibrium condition under constraints.

B.1 Constrained equilibrium condition

Let $\tilde{\ell}_t = \ell_t + I_C$, $t = 1, 2, 3, \ldots$, where I_C is the characteristic function of $C \subseteq \mathbb{R}^d$,

$$I_C(\theta) = \begin{cases} 0 & \text{if } \theta \in C, \\ \infty & \text{if } \theta \notin C. \end{cases}$$

Assume that each ℓ_t is finite and subdifferentiable on C. As noted in the regularization section (and verified in Appendix D), our notion of subgradients allows us to take, as a subgradient $\tilde{g}_t(\theta)$ of $\tilde{\ell}_t$ at θ,

$$\tilde{g}_t(\theta) = g_t(\theta) + g_{I_C}(\theta), \tag{72}$$

where as usual $g_t(\theta)$ is a subgradient of ℓ_t at θ, and $g_{I_C}(\theta)$ denotes a subgradient of I_C at θ. Subgradients of I_C at θ are elements of the normal cone to C at θ, a fact which will be leveraged shortly. For a sequence of subgradients $\tilde{g}_t(\theta)$, $t = 1, 2, 3 \ldots$ which are chosen to satisfy (72), we have

$$\frac{1}{T} \sum_{t=1}^{T} \tilde{g}_t(\theta_t) \asymp \frac{1}{T} \sum_{t=1}^{T} g_t(\theta_t) + \frac{1}{T} \sum_{t=1}^{T} g_{I_C}(\theta_t),$$

and therefore (using $a_T \asymp b_T$ to mean $a_T - b_T \to 0$ as $T \to \infty$, as before)

$$\frac{1}{T}\sum_{t=1}^{T} \tilde{g}_t(\theta_t) \asymp 0 \iff \frac{1}{T}\sum_{t=1}^{T} g_t(\theta_t) \asymp -\frac{1}{T}\sum_{t=1}^{T} g_{I_C}(\theta_t). \tag{73}$$

That is, the gradient equilibrium condition for the constrained losses can be recast as a modified equilibrium condition for the original losses.

We can further develop the right-hand side in (73), to help elucidate this condition. As mentioned above, subgradients $g_{I_C}(\theta)$ of the characteristic function I_C at θ are elements of the normal cone $N_C(\theta)$ to the set C at θ. The consists of vectors $v \in \mathbb{R}^d$ such that $v^\mathsf{T}\delta \leq o(\|\delta\|_2)$, for all directions $\delta \in \mathbb{R}^d$ such that $\theta + \delta \in C$. Appendix D gives a more formal characterization. The second condition in (73) can hence be rewritten as

$$\frac{1}{T}\sum_{t=1}^{T} g_t(\theta_t) \asymp \frac{1}{T}\sum_{t=1}^{T} v_t, \quad \text{where } v_t \in -N_C(\theta_t),\ t = 1, 2, 3, \ldots. \tag{74}$$

From this, we can form the following interpretation. If $g_t(\theta_t) = v_t$ for an individual t, then by definition of the normal cone, $g_t(\theta_t)^\mathsf{T}\delta \geq o(\|\delta\|_2)$, for all directions $\delta \in \mathbb{R}^d$ such that $\theta_t + \delta \in C$. Intuitively, any infinitesimal move away from θ_t which remains feasible (remains within C) cannot decrease ℓ_t, according to its first-order Taylor expansion. In a similar vein, we can view (74) as saying that the average gradient $\frac{1}{T}\sum_{t=1}^{T} g_t(\theta_t)$ will have nonnegative inner product with any infinitesimal direction of "average feasibility" (by this we mean an average of directions which maintain feasibility), as T grows. This can be seen as a sequential analog of the standard first-order condition for optimality in constrained problems.

Remark 28 *When θ lies in the interior of C, the normal cone is trivially $N_C(\theta) = \{0\}$. Only when θ lies on the boundary ∂C of C is the normal cone nontrivial. This allows us to rewrite (74) once more as*

$$\frac{1}{T}\sum_{t=1}^{T} g_t(\theta_t) \asymp \frac{1}{T}\sum_{t:\theta_t \in \partial C} v_t, \quad \textit{where } v_t \in -N_C(\theta_t),\ t = 1, 2, 3, \ldots. \tag{75}$$

Thus if the number t for which $\theta_t \in \partial C$ grows sublinearly, $|\{t \leq T : \theta_t \in \partial C\}| = o(T)$, and v_t, $t = 1, 2, 3, \ldots$ are uniformly bounded, then (75) reduces to the usual gradient equilibrium condition (1). Unfortunately, this observation is of no consequence to us in the theory that follows, as we analyze projected descent methods, which produce iterates that generically lie on the boundary at all iterations. However, it might be useful for analyzing alternative (interior point) methods.

B.2 Examples of constrained equilibrium

We work through examples of the equilibrium condition (74) for constrained losses.

B.2.1 Linear equalities

Consider a set of linear equality constraints, $C = \{\theta \in \mathbb{R}^d : A\theta = b\}$, for an arbitrary matrix $A \in \mathbb{R}^{k \times d}$ and vector $b \in \mathbb{R}^k$. Fix $\theta \in C$. Since C is convex, the normal cone $N_C(\theta)$ has a

more explicit form, recalling (102): it consists of all vectors $v \in \mathbb{R}^d$ such that $v^\mathsf{T}(z-\theta) \leq 0$, for all $z \in C$. Note that for our given C,

$$z \in C \iff \delta = z - \theta \in \text{null}(A),$$

where $\text{null}(A)$ is the null space of A. Also, if $v^\mathsf{T}\delta \leq 0$ for all $\delta \in \text{null}(A)$, then we must have $v^\mathsf{T}\delta = 0$ for all $\delta \in \text{null}(A)$. In other words, we have shown that the normal cone contains all vectors orthogonal to $\text{null}(A)$. Hence, if $U \in \mathbb{R}^{d \times p}$ is a matrix whose columns form a basis for $\text{null}(A)$, then $U^\mathsf{T} v = 0$ for all $v \in N_C(\theta)$. We can then multiply by U^T on both sides of the gradient equilibrium condition (74), which gives

$$\frac{1}{T}\sum_{t=1}^{T} U^\mathsf{T} g_t(\theta_t) \asymp 0. \tag{76}$$

This is in fact equivalent to the usual (unconstrained) gradient equilibrium condition had we reparametrized our losses to satisfy the linear equality constraints. To see this, define $f_t(\beta) = \ell_t(U\beta + \theta_0)$, where $A\theta_0 = b$. The usual gradient equilibrium condition for f_t, $t = 1, 2, 3, \ldots$, invoking the chain rule, is

$$\frac{1}{T}\sum_{t=1}^{T} U^\mathsf{T} g_t(U\beta_t + \theta_0) \asymp 0. \tag{77}$$

(To check that the chain rule still holds for our generalized definition of subgradients, see Theorem 10.6 and Exercise 10.7 of Rockafellar and Wets (2009).) Observe that (77) reduces to (76) once we identify $\theta_t = U\beta_t + \theta_0$.

B.2.2 Simplex

Consider the standard d-dimensional probability simplex, $C = \{\theta \in \mathbb{R}^d : \sum_{i=1} \theta_i = 1,\ \theta_i \geq 0,\ i = 1, \ldots, d\}$. Let $\theta \in \partial C$, and let F be the smallest face of C containing θ. This is the unique face such that θ lies in its relative interior. Write $\text{aff}(F) = L + a$ for the affine span of F, where $L \subseteq \mathbb{R}^d$ is a linear subspace and $a \in \mathbb{R}^d$ is an offset. Again by convexity of C, the normal cone has an explicit form, recalling (102): it consists of all vectors $v \in \mathbb{R}^d$ such that $v^\mathsf{T}(z-\theta) \leq 0$, for all $z \in C$. Now consider all z lying in the face $F \subseteq C$, which are close to θ:

$$\delta = z - \theta \in L \cap B_\epsilon(\theta),$$

for a ball $B_\epsilon(\theta)$ around θ, and some small radius $\epsilon > 0$. Then, in order to have $v^\mathsf{T}\delta \leq 0$ for all $\delta \in L \cap B_\epsilon(\theta)$, we must have $v^\mathsf{T}\delta = 0$ for all $\delta \in L$. In other words, we have shown that the normal cone contains contains all vectors orthogonal to L.

Note that the linear part L of a face F of the standard probability simplex generally takes the form

$$L = \left\{ v \in \mathbb{R}^d : \sum_{i \in S} v_i = 0,\ v_i = 0,\ i \notin S \right\},$$

for some subset $S \subseteq \{1, \ldots, d\}$ of coordinates. Hence, the space of vectors vectors orthogonal to L, written as $L^\perp$, is spanned by 1_S (vector with 1s for elements in S and 0 otherwise), and e_i (vector with 1 in the i^{th} element and 0 otherwise), for $i \notin S$.

Now, assume that F contains v_t in its relative interior, for all $t = 1, 2, 3, \ldots$. (This assumption could be relaxed to F containing all but an $o(T)$ number of elements, with a boundedness assumption on those left out.) Taking an inner product with the aforementioned basis for $L^\perp$, on both sides of the gradient equilibrium condition (74), gives

$$\frac{1}{T} \sum_{t=1}^{T} \sum_{i \in S} [g_t(\theta_t)]_i \asymp 0, \tag{78}$$

$$\frac{1}{T} \sum_{t=1}^{T} [g_t(\theta_t)]_i \asymp 0, \quad i \notin S. \tag{79}$$

That is, the constrained equilibrium condition for the simplex in this case implies the usual notion of gradient equilibrium (79) for a subset of the coordinates, and an aggregate equilibrium condition (78) for the rest.

B.3 Proximal mirror descent

We turn to analyzing *online proximal mirror descent*. As we will see, this algorithm generalizes both online mirror descent and online projected gradient descent. Let $r : \mathbb{R}^d \to (-\infty, \infty]$ be a convex function, finite and subdifferentiable on a set $D \in \mathbb{R}^d$. Assume that each loss ℓ_t is also finite and subdifferentiable on D (but not necessarily convex). Finally, let $\Phi : \mathbb{R}^d \to \mathbb{R}$ be a strictly convex and differentiable function. Given an initial point $\theta_1 \in D$, online proximal mirror descent produces iterates according to:

$$\left.\begin{aligned} \nabla\Phi(z_{t+1}) &= \nabla\Phi(\theta_t) - \eta_t g_t(\theta_t) \\ \theta_{t+1} &= \operatorname*{argmin}_{\theta} \left\{ D_\Phi(\theta, z_{t+1}) + \eta_t r(\theta) \right\} \end{aligned}\right\} \quad t = 1, 2, 3, \ldots, \tag{80}$$

where $D_\Phi(\theta, z) = \Phi(\theta) - \Phi(z) - \nabla\Phi(z)^\mathsf{T}(\theta - z)$ is the Bregman divergence with respect to Φ. The quantity $\nabla\Phi$ in this setting is often referred to as the *mirror map*. Henceforth, in referring to (80), we will drop the qualifier "online" and simply call this proximal mirror descent. Similarly, we will drop the qualifier "online" when discussing all algorithms in this section.

Proximal mirror descent (80) has many notable special cases:

- when $r = I_C$, the characteristic function of a set C, it reduces to mirror descent;
- when $\nabla\Phi = \mathrm{Id}$ (the identity map), it reduces to proximal gradient descent;
- when $r = I_C$ and $\nabla\Phi = \mathrm{Id}$, it reduces to projected gradient descent;
- when $r = 0$ and $\nabla\Phi = \mathrm{Id}$, it reduces to gradient descent.

In what follows, our main interest will be in the cases where $r = I_C$, i.e., mirror descent and projected gradient descent, as we are concerned with studying gradient equilibrium under constraints. But since it is fluid to work with the framework offered by proximal mirror descent, we will begin our analysis in greater generality. Throughout, we let $\tilde{\ell}_t = \ell_t + r$ denote the modified loss and write $\tilde{g}_t(\theta)$ for its subgradient at θ. Recall that we can take

this to be of the form $\tilde{g}_t(\theta) = g_t(\theta) + g_r(\theta)$, where g_r is a subgradient of r at θ. We focus on gradient equilibrium for the regularized loss sequence,

$$\frac{1}{T}\sum_{t=1}^{T}\tilde{g}_t(\theta_t) \asymp 0 \iff \frac{1}{T}\sum_{t=1}^{T} g_t(\theta_t) \asymp -\frac{1}{T}\sum_{t=1}^{T} g_r(\theta_t). \tag{81}$$

Next we derive a generalization of the basic result in Proposition 4 for proximal mirror descent.

Proposition 29 *Consider proximal mirror descent (80), with arbitrary initialization $\theta_1 \in D$, and constant step sizes $\eta_t = \eta > 0$, for all t. Let $g_r(\theta_{t+1}) = (\nabla\Phi(z_{t+1}) - \nabla\Phi(\theta_{t+1}))/\eta$, which is indeed a subgradient of r at θ_{t+1}, for each t, and let $g_r(\theta_1)$ be an arbitrary subgradient of r at θ_1. Then*

$$\frac{1}{T}\sum_{t=1}^{T}\tilde{g}_t(\theta_t) = \frac{\nabla\Phi(\theta_1) + \eta g_r(\theta_1) - \nabla\Phi(\theta_{T+1}) - \eta g_r(\theta_{T+1})}{\eta T}, \tag{82}$$

and therefore

$$\left\|\frac{1}{T}\sum_{t=1}^{T}\tilde{g}_t(\theta_t)\right\|_2 \leq \frac{\|\nabla\Phi(\theta_1) + \eta g_r(\theta_1)\|_2 + \|\nabla\Phi(\theta_{T+1}) + \eta g_r(\theta_{t+1})\|_2}{\eta T}. \tag{83}$$

Proof Consider first the second step in (80), with $\eta_t = \eta$. By the subgradient optimality condition,

$$0 \in \nabla_\theta D(\theta_{t+1}, z_{t+1}) + \eta\partial r(\theta_{t+1}) \iff 0 \in \nabla\Phi(\theta_{t+1}) - \nabla\Phi(z_{t+1}) + \eta\partial r(\theta_{t+1}),$$

where we have used convexity of Φ, r to decompose the subgradients. This verifies the claim that $g_r(\theta_{t+1}) = (\nabla\Phi(z_{t+1}) - \nabla\Phi(\theta_{t+1}))/\eta$ is a subgradient of r at θ_{t+1}. Notice, by the first step in (80), with $\eta_t = \eta$,

$$\begin{aligned}
\nabla\Phi(\theta_{t+1}) &= \nabla\Phi(z_{t+1}) + \nabla\Phi(\theta_{t+1}) - \nabla\Phi(z_{t+1}) \\
&= \nabla\Phi(\theta_t) - \eta g_t(\theta_t) + \nabla\Phi(\theta_{t+1}) - \nabla\Phi(z_{t+1}) \\
&= \nabla\Phi(\theta_t) - \eta g_t(\theta_t) - \eta g_r(\theta_{t+1}).
\end{aligned}$$

Now rewrite the last line as $\nabla\Phi(\theta_t) - \nabla\Phi(\theta_{t+1}) = \eta g_t(\theta_t) + \eta g_r(\theta_{t+1})$. Adding this up over $t = 1, \ldots, T$, the left-hand side telescopes, yielding

$$\nabla\Phi(\theta_1) - \nabla\Phi(\theta_{T+1}) = \eta\sum_{t=1}^{T}(g_t(\theta_t) + g_r(\theta_{t+1})).$$

Identifying $\tilde{g}_t(\theta_t) = g_t(\theta_t) + g_r(\theta_t)$, rearranging, and dividing both sides by ηT proves (82). The bound (83) follows by taking the norm of both sides, and then applying the triangle inequality. ■

The simple bound in (83) reveals an important property of proximal mirror descent (and thus all special cases thereof) with constant step sizes. The next result summarizes.

Proposition 30 *For proximal mirror descent (80) with constant step sizes, gradient equilibrium (81) holds if the mirror map $\nabla\Phi$ delivers slowly growing iterates, $\|\nabla\Phi(\theta_t)\|_2 = o(t)$, and the regularizer r delivers slowly growing residuals $\|g_r(\theta_t)\|_2 = \|\nabla\Phi(z_t) - \nabla\Phi(\theta_t)\|_2/\eta = o(t)$.*

Below, we specialize to $r = I_C$ for a convex set C, and we investigate consequences of the basic theory developed above for both bounded and unbounded constraint sets.

B.3.1 Bounded constraints, general $\nabla\Phi$

The analysis for bounded C is particularly simple, and it does not require any of the restorative theory from Section 2.2. If $\nabla\Phi$ is continuous, then it has a finite maximum on the bounded set C. Combined with a mild (local) Lipschitz condition on ℓ_t, this gives the desired boundedness of the right-hand side in (82).

Proposition 31 *Let $C \subseteq \mathbb{R}^d$ be a bounded convex set. Assume that $\nabla\Phi$ continuous on C, and denote*

$$M = \sup_{\theta \in C} \|\nabla\Phi(\theta)\|_2 < \infty.$$

Assume also that each ℓ_t is L_t-Lipschitz on C, where L_t is sublinear. Then mirror descent (80) with $r = I_C$, arbitrary initialization $\theta_1 \in C$, and constant step sizes $\eta_t = \eta > 0$, for all t, satisfies gradient equilibrium:

$$\left\| \frac{1}{T} \sum_{t=1}^{T} \tilde{g}_t(\theta_t) \right\|_2 \leq \frac{\|\nabla\Phi(\theta_1) + \eta g_r(\theta_1)\|_2}{\eta T} + \frac{M}{\eta T} + \frac{L_T}{T} \to 0, \quad \text{as } T \to \infty. \tag{84}$$

Proof Return to (83), and rewrite the right-hand side as

$$\begin{aligned} \frac{1}{T} \sum_{t=1}^{T} \tilde{g}_t(\theta_t) &\leq \frac{\|\nabla\Phi(\theta_1) + \eta g_r(\theta_1)\|_2 + \|\nabla\Phi(z_{T+1})\|_2}{\eta T} \\ &= \frac{\|\nabla\Phi(\theta_1) + \eta g_r(\theta_1)\|_2 + \|\nabla\Phi(\theta_T) - \eta g_T(\theta_T)\|_2}{\eta T}. \end{aligned}$$

The result (84) follows by the using triangle inequality, $\|\nabla\Phi(\theta_T)\|_2 \leq M$, and $\|g_T(\theta_T)\|_2 \leq L_T$. ■

B.3.2 Unbounded constraints, $\nabla\Phi = \mathrm{Id}$

For unbounded C, fortunately, we can leverage much of the restorative theory from Section 2.2. We assume, without loss of generality, that C contains the origin. This tied to the fact that the restorative condition (7) is based around the origin. (If $0 \notin C$, then we can translate the parameter space to make it so.) We further make the restriction that $\nabla\Phi = \mathrm{Id}$, the identity map, and thus focus on projected gradient descent.

Note that $g_r(\theta_t) \in \partial r(\theta_t) = \partial I_C(\theta_t) = N_C(\theta_t)$, the normal cone to C at θ_t. As C is convex, the normal cone has a special structure (102), which implies $g_r(\theta_t)^\mathsf{T}\theta_t \geq g_r(\theta_t)^\mathsf{T} z$ for all $z \in C$, and so, taking $z = 0$,

$$g_r(\theta_t)^\mathsf{T}\theta_t \geq 0.$$

This simple fact will be critical in porting over the restorative theory to projected gradient descent descent with an unbounded set C. It implies $\|\theta_t\|_2^2 + \eta^2\|g_r(\theta_t)\|_2^2 \leq \|\theta_t + \eta g_r(\theta_t)\|_2^2$, and in particular,

$$\|\theta_t\|_2 \leq \|\theta_t + \eta g_r(\theta_t)\|_2. \tag{85}$$

We are ready to state our result for projected gradient descent, which generalizes Propositions 6–9.

Proposition 32 *Let $C \subseteq \mathbb{R}^d$ be a convex set containing the origin, and consider projected gradient descent (80) with $r = I_C$ and $\nabla\Phi = \mathrm{Id}$, arbitrary initialization $\theta_1 \in C$, and constant step sizes $\eta_t = \eta > 0$, for all t.*

(a) If $d = 1$, and each ℓ_t is L-Lipschitz and $(h_t, 0)$-restorative, for nondecreasing h_t, then

$$|\theta_{T+1} + \eta g_r(\theta_{T+1})| \leq \max\{|\theta_1 + \eta g_r(\theta_1)|,\, h_T\} + \eta L. \tag{86}$$

Thus if h_t is sublinear, then (83) implies gradient equilibrium:

$$\left|\frac{1}{T}\sum_{t=1}^{T} \tilde{g}_t(\theta_t)\right| \leq \frac{2|\theta_1 + \eta g_r(\theta_1)|}{\eta T} + \frac{L}{T} + \frac{h_T}{\eta T} \to 0, \quad \text{as } T \to \infty. \tag{87}$$

(b) For general d, if each ℓ_t is L-Lipschitz and $(h_t, 0)$-restorative, then

$$\|\theta_{T+1} + \eta g_r(\theta_{T+1})\|_2 \leq \sqrt{\|\theta_1 + \eta g_r(\theta_1)\|_2^2 + \eta^2 L^2 T + 2\eta L \sum_{t=1}^{T} h_t}. \tag{88}$$

Thus if h_t is sublinear and nondecreasing, then (83) implies gradient equilibrium:

$$\left\|\frac{1}{T}\sum_{t=1}^{T} \tilde{g}_t(\theta_t)\right\|_2 \leq \frac{2\|\theta_1 + \eta g_r(\theta_1)\|_2}{\eta T} + \sqrt{\frac{L^2}{T} + \frac{2Lh_T}{\eta T}} \to 0, \quad \text{as } T \to \infty. \tag{89}$$

(c) If each ℓ_t is L-Lipschitz and (h_t, ϕ_t)-restorative with positive curvature (13), for nondecreasing h_t, then

$$\|\theta_{T+1} + \eta g_r(\theta_{T+1})\|_2 \leq \max\{\|\theta_1 + \eta g_r(\theta_1)\|_2,\, h_T\} + \eta L. \tag{90}$$

Thus if h_t is sublinear, then (83) implies gradient equilibrium:

$$\left\|\frac{1}{T}\sum_{t=1}^{T} \tilde{g}_t(\theta_t)\right\|_2 \leq \frac{2\|\theta_1 + \eta g_r(\theta_1)\|_2}{\eta T} + \frac{L}{T} + \frac{h_T}{\eta T} \to 0, \quad \text{as } T \to \infty. \tag{91}$$

(d) If each ℓ_t is L_t-Lipschitz on the set $\{\theta \in \mathbb{R}^d : \|\theta\|_2 \leq h_t\}$, and also (h_t, ϕ_t)-restorative with quadratic curvature (17), for nondecreasing h_t, L_t, then

$$\|\theta_{T+1} + \eta g_r(\theta_{T+1})\|_2 \leq \max\{\|\theta_1 + \eta g_r(\theta_1)\|_2,\, h_T\} + \eta L_T. \tag{92}$$

Thus if h_t, L_t are sublinear, then (83) implies gradient equilibrium:

$$\left\|\frac{1}{T}\sum_{t=1}^{T} \tilde{g}_t(\theta_t)\right\|_2 \leq \frac{2\|\theta_1 + \eta g_r(\theta_1)\|_2}{\eta T} + \frac{L_T}{T} + \frac{h_T}{\eta T} \to 0, \quad \text{as } T \to \infty. \tag{93}$$

Proof It will be helpful to rewrite the projected gradient descent updates using the regularizer subgradient $g_r(\theta_{t+1}) = (z_{t+1} - \theta_{t+1})/\eta$, introduced in Proposition 29. Note that we can rewrite (80) (with $\nabla\Phi = \mathrm{Id}$) as

$$\theta_{t+1} + \eta g_r(\theta_{t+1}) = \theta_t - \eta g_t(\theta_t) \quad t = 1, 2, 3, \ldots. \tag{94}$$

We now prove parts (a)–(d) separately.

Part (a). We follow the proof of Proposition 6 in Appendix A.2. For convenience, redefine $h_t = \max\{|\theta_1 + \eta g_r(\theta_1)|, h_t\}$, and let $h_0 = |\theta_1 + \eta g_r(\theta_1)|$. We will use induction to prove (86). The base case, for $T = 0$, is trivial. Assume that the result is true up through T. We break up the argument for $T + 1$ into cases. If $|\theta_T| \le h_T$, then note that by (94) and the triangle inequality

$$\begin{aligned} |\theta_{T+1} + \eta g_r(\theta_{T+1})| &\le |\theta_T| + \eta |g_T(\theta_T)| \\ &\le h_T + \eta L, \end{aligned}$$

where the second line uses L-Lipschitzness of ℓ_t. If instead $|\theta_{T+1}| > h_T$, then the same argument (applying one-sided inequalities) as in Appendix A.2 leads to

$$\theta_{T+1} + \eta g_r(\theta_{t+1}) \in \big[\min\{\theta_T, -\eta L\}, \max\{\theta_T, \eta L\}\big].$$

By (85) and the inductive hypothesis, $|\theta_{T+1} + \eta g_r(\theta_{t+1})| \le h_{T-1} + \eta L \le h_T + \eta L$, using the nondecreasing property of h_T. This completes the inductive step and proves (86). The average gradient bound (87) follows from (83), and then using the simple inequality $\max\{a, b\} \le a + b$ to write the bound more cleanly.

Part (b). We follow the proof of Proposition 7 in Appendix A.3. Taking the squared norm on both sides of the update (94) and expanding yields

$$\begin{aligned} \|\theta_{T+1} + \eta g_r(\theta_{T+1})\|_2^2 &= \|\theta_T\|_2^2 + \eta^2 \|g_T(\theta_T)\|_2^2 - 2\eta g_T(\theta_T)^\mathsf{T} \theta_T \\ &\le \|\theta_T\|_2^2 + \eta^2 L^2 - 2\eta g_T(\theta_T)^\mathsf{T} \theta_T \\ &\le \|\theta_1\|_2^2 + \eta^2 L^2 T - 2\eta \sum_{t=1}^T g_t(\theta_t)^\mathsf{T} \theta_t. \end{aligned}$$

Here the second line uses L-Lipschitzness, while the third unravels the iteration over $t = 1, \ldots, T$, using (85) at each step. The same argument as in Appendix A.3 shows that each $-2\eta g_t(\theta_t)^\mathsf{T} \theta_t \le 2\eta L h_t$. Plugging this in, and taking a square root, proves (88). The second result (12) follows from bounding $\sum_{t=1}^T h_T \le T h_T$, as h_T is nondecreasing, applying (6), and then using $\sqrt{a+b} \le \sqrt{a} + \sqrt{b}$ to write the bound more cleanly.

Part (c). We follow the proof of Proposition 8 in Appendix A.4. Redefine $h_t = \max\{\|\theta_1 + \eta g_r(\theta_1)\|_2, h_t\}$, and let $h_0 = \|\theta_1 + \eta g_r(\theta_1)\|_2$. We use induction to prove (90). The base case, for $T = 0$, is trivial. Assume that the result is true up through T. We break up the argument for $T + 1$ into cases. If $\|\theta_T\|_2 \le h_T$, then note that by (94) and the triangle inequality

$$\begin{aligned} \|\theta_{T+1} + \eta_t g_r(\theta_{T+1})\|_2 &\le \|\theta_T\|_2 + \eta \|g_T(\theta_T)\|_2 \\ &\le h_T + \eta L, \end{aligned}$$

where the second line uses L-Lipschitzness of ℓ_T. If instead $\|\theta_T\|_2 > h_T$, then

$$\begin{aligned}\|\theta_{T+1} + \eta g_r(\theta_{T+1})\|_2^2 &= \|\theta_T\|_2^2 + \eta^2\|g_T(\theta_T)\|_2^2 - 2\eta g_T(\theta_T)^\mathsf{T}\theta_T \\ &\leq \|\theta_T\|_2^2 + \eta^2 L^2 - 2\eta g_T(\theta_T)^\mathsf{T}\theta_T \\ &\leq (h_{T-1} + \eta L)^2 + \eta^2 L^2 - 2\eta g_T(\theta_T)^\mathsf{T}\theta_T \\ &\leq (h_{T-1} + \eta L)^2 + \eta^2 L^2 - 2\eta\phi_T(\theta) \\ &\leq (h_T + \eta L)^2.\end{aligned}$$

Here the second line uses L-Lipschitzness of ℓ_T, the third uses (85) and the inductive hypothesis, the fourth uses the restorative condition (7), and the last uses the nondecreasing property of h_T and positive curvature (13). Taking a square root proves (90). The second result (91) follows from (83), and using $\max\{a, b\} \leq a + b$ to write the bound more cleanly.

Part (d). This follows a similar proof to part (c) (and also to that of Proposition 9 in Appendix A.5), but differs slightly in the inductive argument for $T + 1$. As before, we divide into two cases. If $\|\theta_T\|_2 \leq h_T$, then the triangle inequality and local L_T-Lipschitzness of ℓ_T implies $\|\theta_{T+1}\|_2 \leq h_T + \eta L_T$. If instead $\|\theta_T\|_2 > h_T$, then expanding the squared norm of the gradient update, using (85) together with the inductive hypothesis, and nondecreasingness of h_T, L_T, gives

$$\|\theta_{T+1}\|_2^2 \leq (h_T + \eta L_T)^2 + \eta^2\|g_T(\theta_T)\|_2^2 - 2\eta g_T(\theta_T)^\mathsf{T}\theta_T.$$

The restorative condition (7) with quadratic curvature (17) implies

$$\eta^2\|g_T(\theta_T)\|_2^2 - 2\eta g_T(\theta_T)^\mathsf{T}\theta_T \leq \eta^2\|g_T(\theta_T)\|_2^2 - 2\eta\phi_T(\theta) \leq 0,$$

and thus $\|\theta_{T+1}\|_2 \leq h_T + \eta L_T$, completing the proof of (92). ■

Remark 33 *All uses of the restorative property, throughout the proof, were limited to examining gradient inner products of the form $g_t(\theta_t)^\mathsf{T}\theta_t$ for $\theta_t \in C$ (by the nature of projected gradient descent). That is, instead of (7), we only really require the following C-restricted restorative condition:*

$$g(\theta)^\mathsf{T}\theta \geq \phi(\theta), \quad \text{for all } \|\theta\|_2 > h,\ \theta \in C, \text{ and all generalized subgradients } g(\theta) \text{ of } \ell \text{ at } \theta. \tag{95}$$

Appendix C. No move regret

Consider the following definition of a regret-type property.

Definition 34 *A sequence of iterates θ_t, $t = 1, 2, 3, \ldots$ is satisfies* no move regret *(NMR) with respect to a sequence of loss functions ℓ_t, $t = 1, 2, 3, \ldots$ provided that for any bounded set $D \subseteq \mathbb{R}^d$,*

$$\liminf_{T\to\infty} \inf_{\delta\in D} \frac{1}{T}\left(\sum_{t=1}^{T} \ell_t(\theta_t + \delta) - \sum_{t=1}^{T} \ell_t(\theta_t)\right) \geq 0. \tag{96}$$

This concept serves as somewhat of a bridge between gradient equilibrium (GEQ) and no regret (NR).

C.1 Relationship to GEQ

For convex losses, it is straightforward to show that GEQ implies NMR.

Proposition 35 *Assume each ℓ_t, $t = 1, 2, 3, \ldots$ is convex. If the iterates θ_t, $t = 1, 2, 3, \ldots$ satisfy gradient equilibrium (1), then they satisfy no move regret (96).*

Proof Fix any bounded set D, and $\delta \in D$. Since ℓ_t is convex, any subgradient $g_t(\theta_t)$ at θ_t must satisfy

$$\ell_t(\theta_t + \delta) - \ell_t(\theta_t) \geq g_t(\theta_t)^\mathsf{T} \delta.$$

Average this over $t = 1, \ldots, T$ and using linearity gives

$$\frac{1}{T}\left(\sum_{t=1}^T \ell_t(\theta_t + \delta) - \sum_{t=1}^T \ell_t(\theta_t)\right) \geq \left(\frac{1}{T}\sum_{t=1}^T g_t(\theta_t)\right)^\mathsf{T} \delta.$$

Letting b denote the bound on vectors in D, by Cauchy-Schwarz,

$$\inf_{\delta \in D} \frac{1}{T}\left(\sum_{t=1}^T \ell_t(\theta_t + \delta) - \sum_{t=1}^T \ell_t(\theta_t)\right) \geq \left(\frac{1}{T}\sum_{t=1}^T g_t(\theta_t)\right)^\mathsf{T} \delta \geq -b\left\|\frac{1}{T}\sum_{t=1}^T g_t(\theta_t)\right\|_2.$$

Taking the limit infimum as $T \to \infty$ proves the claim. ■

On the other hand, for smooth losses, it is likewise straightforward to show that NMR implies GEQ.

Proposition 36 *Assume each ℓ_t, $t = 1, 2, 3, \ldots$ is β-smooth (differentiable with β-Lipschitz gradient). If the iterates θ_t, $t = 1, 2, 3, \ldots$ satisfy no move regret (96), then they satisfy gradient equilibrium (1).*

Proof Fix any δ. By a well-known fact for β-smooth functions,

$$\ell_t(\theta_t + \delta) - \ell_t(\theta_t) \leq \nabla \ell_t(\theta_t)^\mathsf{T} \delta + \frac{\beta}{2}\|\delta\|_2^2.$$

Averaging this over $t = 1, \ldots, T$ gives

$$\frac{1}{T}\left(\sum_{t=1}^T \ell_t(\theta_t + \delta) - \sum_{t=1}^T \ell_t(\theta_t)\right) \leq \left(\frac{1}{T}\sum_{t=1}^T \nabla \ell_t(\theta_t)\right)^\mathsf{T} \delta + \frac{\beta}{2}\|\delta\|_2^2.$$

Letting $G_T = \frac{1}{T}\sum_{t=1}^T \nabla \ell_t(\theta_t)$, the right-hand side above is minimized at $\delta = -G_T/\beta$, and it has a minimum value of $-\|G_T\|_2^2/(2\beta)$. Hence, if $G_t \not\to 0$ as $T \to \infty$, then we would have a violation of NMR. This proves the claim via the contrapositive. ■

In the above, we assumed convexity and smoothness for simplicity. It is possible that connections could be drawn in greater generality.

C.2 Relationship to NR

For convex, smooth losses, we know from the above that NMR and NR cannot be equivalent (Propositions 35 and 36 imply that NMR is equivlaent to GEQ, whereas the examples and discussion in Section 1.2 show that GEQ is not equivalent to NR). That said, the NMR condition in (96) can be recast in a way to more closely resemble NR in (2). Simply rearranging terms, and using the fact that $-\liminf_{T\to\infty} a_T = \limsup_{T\to\infty} -a_T$ for any sequence a_T, observe that NMR can be rewritten as

$$\limsup_{T\to\infty} \left(\frac{1}{T}\sum_{t=1}^{T} \ell_t(\theta_t) - \inf_{\delta\in D} \frac{1}{T}\sum_{t=1}^{T} \ell_t(\theta_t+\delta) \right) \leq 0, \tag{97}$$

for any bounded set $D \subseteq \mathbb{R}^d$. Whereas NR (2) compares the average witnessed loss to the best possible fixed parameter θ in hindsight, we find that NMR (97) compares the average witnessed loss to the best possible fixed "move" δ away from the given parameter sequence θ_t, $t = 1, 2, 3, \ldots$. As explained above, a comparison against the best fixed parameter (as in NR) is neither strong nor weaker than a comparison against the best fixed move (as in NMR), in general. These are simply different comparisons.

C.3 Interpretation for GLMs

We now explore NMR in the context of GLMs, to help understand and interpret the comparison to NR (and GEQ). Recall, a GLM loss is of the form

$$\ell_t(\theta) = -y_t x_t^\mathsf{T}\theta + \psi(x_t^\mathsf{T}\theta),$$

where x_t is a feature, y_t is a response, and ψ is the cumulant generating function in the underlying exponential family, e.g., $\psi(u) = \frac{1}{2}u^2$ for linear regression, $\psi(u) = (b-a)\log(1+e^u) + au$ for generalized logistic regression, or $\psi(u) = e^u$ for Poisson regression. Given a sequence of such GLM losses, NR (2) means that solving the "batch" GLM optimization problem

$$\inf_{\theta} \frac{1}{T}\sum_{t=1}^{T} \Big(-y_t x_t^\mathsf{T}\theta + \psi(x_t^\mathsf{T}\theta) \Big)$$

cannot asymptotically improve on the average witnessed loss $\frac{1}{T}\sum_{t=1}^{T}(-y_t x_t^\mathsf{T}\theta + \psi(x_t^\mathsf{T}\theta))$. On the other hand, NMR (97) means that the solving "posthoc" GLM optimization

$$\inf_{\delta\in D} \frac{1}{T}\sum_{t=1}^{T} \Big(-y_t x_t^\mathsf{T}(\theta_t+\delta) + \psi(x_t^\mathsf{T}(\theta_t+\delta)) \Big)$$

cannot asymptotically improve on the average witnessed loss, for any bounded set D. This is posthoc in the sense that we are trying to adjust the witnessed linear predictor $x_t^\mathsf{T}\theta_t$ by the amount $\eta_t = x_t^\mathsf{T}\delta$, at each t. As the above optimization cannot improve on the average witnessed loss, we know that the infimum is obtained at $\delta = 0$, in the asymptotic sense.

Meanwhile, GEQ has its own interpretation for GLMs, transparently different from NMR, but equivalent to NMR when ψ is convex (all GLMs) and twice differentiable with bounded second derivative ψ'' (linear and logistic regression, or Poisson regression on a bounded

parameter space), and the data x_t, y_t, $t = 1, 2, 3, \ldots$ is bounded, according to the results in Propositions 35 and 36. Observe that GEQ (1) in the current setting is

$$\frac{1}{T}\sum_{t=1}^{T}(y_t - \psi'(x_t^\mathsf{T}\theta_t))x_t \asymp 0,$$

which means that

$$\frac{1}{T}\sum_{t=1}^{T}(y_t - \psi'(x_t^\mathsf{T}\theta_t) - \psi'(x_t^\mathsf{T}\theta))x_t \asymp 0$$

is asymptotically solved when the fitted value is $\mu_t = \psi'(x_t^\mathsf{T}\theta) = 0$, at each t. As the above is the first-order optimality condtion for a GLM with residuals $y_t - \psi'(x_t^\mathsf{T}\theta_t)$, $t = 1, 2, 3, \ldots$ in place of the original responses, this says that the "batch" GLM optimization on the residuals

$$\inf_\theta \frac{1}{T}\sum_{t=1}^{T}\Big(- (y_t - \psi'(x_t^\mathsf{T}\theta_t))x_t^\mathsf{T}\theta + \psi(x_t^\mathsf{T}\theta)\Big)$$

will result in null fitted values $\mu_t = \psi'(x_t^\mathsf{T}\theta) = 0$, at each t, in the asymptotic sense (i.e., for linear regression, the infimum will be asymptotically obtained at $\theta = 0$, and for logistic, it will be asymptotically obtained by sending $\theta \to -\infty$ coordinatewise).

Appendix D. Generalized subgradients

A generalized subgradient of ℓ at $\theta \in D$, denoted $g(\theta)$, can be taken to be any vector $v \in \mathbb{R}^d$ satisfying

$$\liminf_{\substack{z\to\theta \\ z\neq\theta}} \frac{\ell(z) - \ell(\theta) - v^\mathsf{T}(z-\theta)}{\|z-\theta\|_2} \geq 0. \tag{98}$$

We note that Rockafellar and Wets (2009) call this a "regular" subgradient, in their Definition 8.3. Two important special cases to highlight are as follows:

- if ℓ is convex, then $g(\theta)$ reduces to a classical subgradient of ℓ at θ, written $g(\theta) \in \partial\ell(\theta)$;
- if ℓ is differentiable at θ, then $g(\theta)$ reduces (uniquely) to the gradient of ℓ at θ, written $g(\theta) = \nabla\ell(\theta)$.

See Exercise 8.8(a) and Proposition 8.12 of Rockafellar and Wets (2009) for proofs of these facts. In general, when a subgradient exists at θ, the function ℓ is said to be subdifferentiable at this point.

Note that we may abbreviate (98) as

$$\ell(z) \geq \ell(\theta) + v^\mathsf{T}(z-\theta) + o(\|z-\theta\|_2), \quad \text{for all } z. \tag{99}$$

This allows us to easily verify the following fact. If $\ell = f_1 + f_2$, where each f_1, f_2 are subdifferentiable at θ with subgradients v_1, v_2 respectively, then ℓ is subdifferentiable at θ with subgradent $v_1 + v_2$. To check this, simply note that by adding together

$$f_1(z) \geq f_1(\theta) + v_1^\mathsf{T}(z-\theta) + o(\|z-\theta\|_2),$$
$$f_2(z) \geq f_2(\theta) + v_2^\mathsf{T}(z-\theta) + o(\|z-\theta\|_2),$$

we get

$$(f_1 + f_2)(z) \geq (f_1 + f_2)(\theta) + (v_1 + v_2)^\mathsf{T}(z - \theta) + o(\|z - \theta\|_2).$$

Another useful fact concerns the characteristic function I_C of a set C,

$$I_C(\theta) = \begin{cases} 0 & \text{if } \theta \in C, \\ \infty & \text{if } \theta \notin C, \end{cases}$$

whose subgradients are intimately connected to the geometry of C. It is not hard to see that subgradients of I_C are vectors $v \in \mathbb{R}^d$ satisfying

$$\liminf_{\substack{z \xrightarrow{C} \theta \\ z \neq \theta}} \frac{v^\mathsf{T}(z - \theta)}{\|z - \theta\|_2} \leq 0. \tag{100}$$

Here $z \xrightarrow{C} \theta$ means that z converges to θ from within C. A vector $v \in \mathbb{R}^d$ that satisfies the above condition is called a normal vector to C a θ. The set of all such vectors form what we call the normal cone, denoted $N_C(\theta)$. We note that Rockafellar and Wets (2009) call such vectors "regular" normal vectors. Similar to the notation for subgradients, it is often convenient to abbreviate (100) as

$$v^\mathsf{T}(z - \theta) \leq o(\|z - \theta\|_2), \quad \text{for all } z \in C. \tag{101}$$

The form (101) is used to describe normal vectors in Appendix B. An important special case to highlight: if C is convex, then $N_C(\theta)$ reduces to the usual definition of the normal cone from convex geometry,

$$N_C(\theta) = \{v \in \mathbb{R}^d : v^\mathsf{T}(z - \theta) \leq 0, \text{ for all } z \in C\}. \tag{102}$$

See Theorem 6.9 of Rockafellar and Wets (2009) for a proof of this fact.

References

Jacob Abernethy, Peter L. Bartlett, and Elad Hazan. Blackwell approachability and low-regret learning are equivalent. In *Proceedings of the Annual Conference on Learning Theory*, 2011.

Anastasios N. Angelopoulos, Emmanuel J. Candès, and Ryan J. Tibshirani. Conformal PID control for time series prediction. In *Advances in Neural Information Processing Systems*, 2023.

Anastasios N. Angelopoulos, Rina Foygel Barber, and Stephen Bates. Online conformal prediction with decaying step sizes. In *Proceedings of the International Conference on Machine Learning*, 2024.

Jean-Bernard Baillon and Georges Haddad. Quelques propriétés des opérateurs angle-bornés et n-cycliquement monotones. *Israel Journal of Mathematics*, 26:137–150, 1977.

Solon Barocas, Moritz Hardt, and Arvind Narayanan. *Fairness and Machine Learning: Limitations and Opportunities*. MIT Press, 2023.

Osbert Bastani, Varun Gupta, Christopher Jung, Georgy Noarov, Ramya Ramalingam, and Aaron Roth. Practical adversarial multivalid conformal prediction. In *Advances in Neural Information Processing Systems*, 2022.

Heinz H. Bauschke and Patrick L. Combettes. The Baillon-Haddad theorem revisited. *Journal of Convex Analysis*, 17(3):781–787, 2009.

Aadyot Bhatnagar, Huan Wang, Caiming Xiong, and Yu Bai. Improved online conformal prediction via strongly adaptive online learning. In *Proceedings of the International Conference on Machine Learning*, 2023.

David Blackwell. An analog of the minimax theorem for vector payoffs. *Pacific Journal of Mathematics*, 6(1):1–8, 1956.

Vincent Blot, Anastasios N. Angelopoulos, Michael I. Jordan, and Nicolas J. B. Brunel. Automatically adaptive conformal risk control. arXiv preprint arXiv:2406.17819, 2024.

Ralph A. Bradley and Milton E. Terry. Rank analysis of incomplete block designs: I. The method of paired comparisons. *Biometrika*, 39(3/4):324–345, 1952.

Nicoló Cesa-Bianchi and Gabor Lugosi. *Prediction, Learning, and Games*. Cambridge University Press, 2006.

Nicolò Cesa-Bianchi, Yoav Freund, David Haussler, David P. Helmbold, Robert E. Schapire, and Manfred K. Warmuth. How to use expert advice. *Journal of the ACM*, 44:427–485, 1997.

Jiadong Chen, Yang Luo, Xiuqi Huang, Fuxin Jiang, Yangguang Shi, Tieying Zhang, and Xiaofeng Gao. IPOC: An adaptive interval prediction model based on online chasing and conformal inference for large-scale systems. In *Proceedings of Conference on Knowledge Discovery and Data Mining*, 2023.

Wei-Lin Chiang, Lianmin Zheng, Ying Sheng, Anastasios Nikolas Angelopoulos, Tianle Li, Dacheng Li, Hao Zhang, Banghua Zhu, Michael Jordan, Joseph E. Gonzalez, and Ion Stoica. Chatbot Arena: An open platform for evaluating LLMs by human preference. arXiv preprint arXiv: 2403.04132, 2024.

Thomas Cover. Universal portfolios. *Mathematical Finance*, 1:1–29, 1991.

Zhun Deng, Cynthia Dwork, and Linjun Zhang. HappyMap: A generalized multicalibration method. In *Proceedings of the Innovations in Theoretical Computer Science Conference*, 2023.

Arpad E. Elo. The proposed USCF rating system, its development, theory, and applications. *Chess Life*, 22(8):242–247, 1967.

Shai Feldman, Liran Ringel, Stephen Bates, and Yaniv Romano. Achieving risk control in online learning settings. *Transactions on Machine Learning Research*, 2023. ISSN 2835-8856.

Dean Foster. A proof of calibration via Blackwell's approachability theorem. *Games and Economic Behavior*, 29:73–78, 1999.

Dean Foster and Rakesh Vohra. Asymptotic calibration. *Biometrika*, 85:379–390, 1998.

Haosen Ge, Hamsa Bastani, and Osbert Bastani. Stochastic online conformal prediction with semi-bandit feedback. arXiv preprint arXiv:2405.13268, 2024.

Gemma Team, Thomas Mesnard, Cassidy Hardin, Robert Dadashi, Surya Bhupatiraju, Shreya Pathak, Laurent Sifre, Morgane Rivière, Mihir Sanjay Kale, Juliette Love, et al. Gemma: Open models based on Gemini research and technology. arXiv preprint arXiv:2403.08295, 2024.

Saeed Ghadimi and Guanghui Lan. Stochastic first-and zeroth-order methods for nonconvex stochastic programming. *SIAM Journal on Optimization*, 23(4):2341–2368, 2013.

Isaac Gibbs and Emmanuel J. Candès. Adaptive conformal inference under distribution shift. In *Advances in Neural Information Processing Systems*, 2021.

Isaac Gibbs and Emmanuel J. Candès. Conformal inference for online prediction with arbitrary distribution shifts. *Journal of Machine Learning Research*, 25(162):1–36, 2023.

Isaac Gibbs, John J. Cherian, and Emmanuel J. Candès. Conformal prediction with conditional guarantees. arXiv preprint arXiv:2305.12616, 2023.

Vidyadhar P. Godambe. *Estimating functions*. Oxford University Press, 1991.

Varun Gupta, Christopher Jung, Georgy Noarov, Mallesh M Pai, and Aaron Roth. Online multivalid learning: Means, moments, and prediction intervals. arXiv preprint arXiv:2101.01739, 2021.

James Hannan. Approximation to Bayes risk in repeated play. *Contributions to the Theory of Games*, 3:97–139, 1957.

Elad Hazan. *Introduction to Online Optimization*. Cambridge University Press, 2016.

Elad Hazan, Karan Singh, and Cyril Zhang. Efficient regret minimization in non-convex games. In *Proceedings of the International Conference on Machine Learning*, 2017.

Ursula Hébert-Johnson, Michael P. Kim, Omer Reingold, and Guy Rothblum. Multicalibration: Calibration for the (computationally-identifiable) masses. In *Proceedings of the International Conference on Machine Learning*, 2018.

Arthur Hoerl and Robert Kennard. Ridge regression: Biased estimation for nonorthogonal problems. *Technometrics*, 12(1):55–67, 1970.

Christopher Jung, Georgy Noarov, Ramya Ramalingam, and Aaron Roth. Batch multivalid conformal prediction. In *Proceedings of the International Conference on Learning Representations*, 2023.

John L. Kelly. A new interpretation of information rate. *Bell System Technical Journal*, 35: 917–926, 1956.

Jack Kiefer and Jacob Wolfowitz. Stochastic estimation of the maximum of a regression function. *Annals of Mathematical Statistics*, 23:462–466, 1952.

Michael P. Kim, Amirata Ghorbani, and James Zou. Multiaccuracy: Black-box post-processing for fairness in classification. In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, 2019.

Jordan Lekeufack, Anastasios N. Angelopoulos, Andrea Bajcsy, Michael I. Jordan, and Jitendra Malik. Conformal decision theory: Safe autonomous decisions from imperfect predictions. In *Proceedings of the IEEE International Conference on Robotics and Automation*. IEEE, 2024.

Nick Littlestone and Manfred Warmuth. The weighted majority algorithm. *Information and Computation*, 108(2):212–261, 1994.

H. Brendan McMahan, Gary Holt, David Sculley, Michael Young, Dietmar Ebner, Julian Grady, Lan Nie, Todd Phillips, Eugene Davydov, Daniel Golovin, Sharat Chikkerur, Dan Liu, Martin Wattenberg, Arnar Mar Hrafnkelsson, Tom Boulos, and Jeremy Kubica. Ad click prediction: A view from the trenches. In *Proceedings of the Conference on Knowledge Discovery and Data Mining*, 2013.

Arkadi S. Nemirovski and David Yudin. *Problem Complexity and Method Efficiency in Optimization*. John Wiley & Sons, 1985.

Georgy Noarov, Ramya Ramalingam, Aaron Roth, and Stephan Xie. High-dimensional unbiased prediction for sequential decision making. In *NeurIPS Workshop on Optimization for Machine Learning*, 2023.

Francesco Orabona. A modern introduction to online learning. arXiv preprint arXiv: 1901.01608, 2019.

Aleksandr Podkopaev, Darren Xu, and Kuang-chih Lee. Adaptive conformal inference by betting. arXiv preprint arXiv:2412.19318, 2024.

Boris T. Polyak and Anatoly B. Juditsky. Acceleration of stochastic approximation by averaging. *SIAM Journal on Control and Optimization*, 30(4):838–855, 1992.

Jin Qin and Jerry Lawless. Empirical likelihood and general estimating equations. *Annals of Statistics*, 22(1):300–325, 1994.

Herbert Robbins and Sutton Monro. A stochastic approximation method. *Annals of Mathematical Statistics*, 22(3):400–407, 1951.

R. Tyrrell Rockafellar and Roger J-B Wets. *Variational Analysis*. Springer, 2009. third printing.

David Ruppert. Efficient estimations from a slowly convergent Robbins-Monro process. Technical Report, Department of Operations Research and Industrial Engineering, Cornell University, 1988.

Robert Tibshirani. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society: Series B*, 58(1):267–288, 1996.

Vladimir Vovk. Aggregating strategies. In *Proceedings of the Annual Conference Computational Learning Theory*, 1990.

Vladimir Vovk, Alex Gammerman, and Glenn Shafer. *Algorithmic Learning in a Random World*. Springer, 2005.

Margaux Zaffran, Olivier Féron, Yannig Goude, Julie Josse, and Aymeric Dieuleveut. Adaptive conformal predictions for time series. In *Proceedings of the International Conference on Machine Learning*, 2022.

Matteo Zecchin and Osvaldo Simeone. Localized adaptive risk control. arXiv preprint arXiv:2405.07976, 2024.

Zhiyu Zhang, Zhou Lu, and Heng Yang. The importance of being Bayesian in online conformal prediction. In *NeurIPS Workshop on Bayesian Decision-Making and Uncertainty*, 2024.

Martin Zinkevich. Online convex programming and generalized infinitesimal gradient ascent. In *Proceedings of the International Conference on Machine Learning*, 2003.

Jacob Ziv and Abraham Lempel. Compression of individual sequences via variable-rate coding. *IEEE Transactions on Information Theory*, 24:530–537, 1978.