# Model Agreement via Anchoring

**Eric Eaton** EEATON@SEAS.UPENN.EDU
*University of Pennsylvania*

**Surbhi Goel** SURBHIG@SEAS.UPENN.EDU
*University of Pennsylvania*

**Marcel Hussing** MHUSSING@SEAS.UPENN.EDU
*University of Pennsylvania*

**Michael Kearns** MKEARNS@SEAS.UPENN.EDU
*University of Pennsylvania*

**Aaron Roth** AAROTH@SEAS.UPENN.EDU
*University of Pennsylvania*

**Sikata Bela Sengupta** SIKATA@SEAS.UPENN.EDU
*University of Pennsylvania*

**Jessica Sorrell** JESS@JHU.EDU
*Johns Hopkins University*

**Editors:** Steve Hanneke and Tor Lattimore

## Abstract

Numerous lines of aim to control *model disagreement* — the extent to which two machine learning models disagree in their predictions. We adopt a simple and standard notion of model disagreement in real-valued prediction problems, namely the expected squared difference in predictions between two models trained on independent samples, without any coordination of the training processes. We would like to be able to drive disagreement to zero with some natural parameter(s) of the training procedure using analyses that can be applied to existing training methodologies.

We develop a simple general technique for proving bounds on independent model disagreement based on *anchoring* to the average of two models within the analysis. We then apply this technique to prove disagreement bounds for four commonly used machine learning algorithms: (1) stacked aggregation over an arbitrary model class (where disagreement is driven to 0 with the number of models $k$ being stacked) (2) gradient boosting (where disagreement is driven to 0 with the number of iterations $k$) (3) neural network training with architecture search (where disagreement is driven to 0 with the size $n$ of the architecture being optimized over) and (4) regression tree training over all regression trees of fixed depth (where disagreement is driven to 0 with the depth $d$ of the tree architecture). For clarity, we work out our initial bounds in the setting of one-dimensional regression with squared error loss — but then show that all of our results generalize to multi-dimensional regression with any strongly convex loss.

**Keywords:** model agreement, stacking, gradient boosting, neural networks, regression trees

## 1. Introduction

Two predictive models $f_1, f_2 : \mathcal{X} \to \mathbb{R}$, trained on data sampled from the same distribution $\mathcal{D}$, might frequently *disagree* in the sense that on a typical test example $x \sim \mathcal{D}$, $f_1(x)$ and $f_2(x)$ take very different values. In fact, this can happen even when the two models are trained on the same dataset, if the model class is not convex and the training process is stochastic. This kind of model *disagreement*, sometimes known as model or predictive multiplicity (Marx et al., 2020; Black et al., 2022; Roth and Tolbert, 2025) or the *Rashomon effect* (Breiman, 2001), is a concern for many different reasons. Pragmatically, predictions are used to inform downstream actions, and two models that make different predictions produce ambiguity about which is the best action to take when we can only take one. This has led to a literature on how two predictive models (or a predictive model and a human) can engage in short test-time interactions so as to "agree" on a single prediction or action that is more accurate than either model could have made alone (Aumann, 1976; Aaronson, 2005; Donahue et al., 2022; Frongillo et al., 2023; Peng et al., 2025; Collina et al., 2025, 2026). In industrial applications, this same phenomenon is known as model or predictive *churn*; there is a large body of work that aims to reduce it, because churn for predictions in ways that do not produce accuracy improvements can needlessly disrupt downstream pipelines built around an initial model (Milani Fard et al., 2016; Bahri and Jiang, 2021; Hidey et al., 2022; Watson-Daniels et al., 2024). The phenomenon of predictive multiplicity has led to concern about the potential arbitrariness of decisions informed by statistical models, and hence the procedural fairness of using such models in high-stakes settings (Marx et al., 2020; Black et al., 2022; Watson-Daniels et al., 2024). The same phenomenon is what underlies the desire for *replicability* of machine learning algorithms, which has recently attracted widespread study (Impagliazzo et al., 2022; Bun et al., 2023; Eaton et al., 2023; Kalavasis et al., 2024b,a; Karbasi et al., 2023; Diakonikolas et al., 2025; Eaton et al., 2026).

In this paper we ask when training on independent samples from a common distribution results in models that approximately agree on most inputs. Unlike the (model) agreement literature (Aumann, 1976; Aaronson, 2005; Collina et al., 2025) we want approximate agreement "out of the box", without the need for any test-time interaction or coordination. And unlike the literature on replicability (Impagliazzo et al., 2022; Bun et al., 2023; Eaton et al., 2023; Karbasi et al., 2023), we do not want our analyses to apply only to custom-designed (and often impractical) algorithms: we want methods for analyzing existing families of practical learning algorithms. We continue a discussion of additional related work in Appendix A.

### 1.1. Our Results

Our notion of approximate model agreement is that the expected squared difference between two models $f_1$ and $f_2$ should be small: $D(f_1, f_2) := \mathbb{E}_{x\sim P}[(f_1(x) - f_2(x))^2] \leq \varepsilon$. Our goal is to show that for broad classes of model training methods, this disagreement level $\varepsilon$ can be driven to 0 with some tunable parameter of the method. We aim for high agreement in this sense via independent training, i.e., without the need for any interaction or coordination between the learners beyond the fact that they are sampling data from a common distribution. We give an abstract recipe for establishing guarantees like this based on a "midpoint anchoring argument" and then give four applications of the recipe: (1) to the popular ensembling technique of "stacking", (2) to gradient boosting and similar methods that iteratively build up linear combinations over a class of base models, (3) to neural network training with architecture search, and (4) to regression tree training over all regression trees of bounded depth. For clarity, we first establish all of our guarantees for

models that solve a one dimensional regression problem to optimize for squared loss, but we then show how our results generalize to multi-dimensional strongly convex loss functions. We include our results on these generalizations in Section 6 and Appendix E.

### 1.1.1. THE MIDPOINT ANCHORING METHOD

Our core technique is built around a simple "midpoint identity" for squared loss. For any two predictors $f_1, f_2 : \mathcal{X} \to \mathbb{R}$, let $\bar{f}(x) := \frac{1}{2}(f_1(x) + f_2(x))$ denote the (hypothetical) model corresponding to their average. Then

$$D(f_1, f_2) \;=\; 2\Big(\mathrm{MSE}(f_1) + \mathrm{MSE}(f_2) - 2\,\mathrm{MSE}(\bar{f})\Big).$$

This reduces proving independent disagreement bounds (the goal of this paper) to bounding the *error gap* between the constituent models $f_1$ and $f_2$ and their average. If $\bar{f}$ lies in the same hypothesis class $\mathcal{H}$ as $f_1$ and $f_2$, then this error gap can be bounded by any convergence analysis that establishes that $\mathrm{MSE}(f)$ will approach error optimality within $\mathcal{H}$. More frequently, for non-convex classes, $\bar{f}$ will not be representable within the same class of functions as $f_1$ and $f_2$ — but for many natural concept classes, the average of two models trained within some class of models parameterized by a measure of complexity (size, depth) will be representable within a class that is "not much larger". This will give us stability guarantees in terms of the "local learning curve" of this complexity parameter, which because of error boundedness and monotonicity must tend to zero at values of the complexity parameter that can be bounded independently of the instance.

All of our stability bounds are "agnostic" in the sense that they hold without any distributional or realizability assumptions. In other words, our bounds will always follow from the ability to *optimize* within given model classes, without needing to assume that the model class is able to represent the relationship between the features and the labels to any non-trivial degree.

It is instructive at the outset to compare the midpoint anchoring method to a more naive methods for establishing agreement bounds. Any pair of models $f_1$ and $f_2$ that both have almost perfect accuracy in the sense that $\mathrm{MSE}(f_1), \mathrm{MSE}(f_2) \leq \varepsilon$ must also satisfy $D(f_1, f_2) \leq O(\varepsilon)$. This follows by anchoring on hypothetical perfect predictions $f^*(x) = y$. Of course, such bounds will rarely apply because very few settings are compatible with near perfect prediction. The benefit of our more general midpoint anchoring method is that it will allow us to argue for independent model agreement without needing to make any realizability assumptions — high accuracy is not needed for high agreement, as if $f_1$ and $f_2$ have high error, so might the average model $\bar{f}$.

### 1.1.2. APPLICATIONS: ENSEMBLING, BOOSTING, NEURAL NETS, AND REGRESSION TREES

We choose our four applications below to show the various ways in which we can apply our method in settings that are progressively more challenging. First, as a warm-up, we study stacked aggregation, which ensembles independently trained models. We show how the midpoint anchoring method can recover strong agreement results as a function of the *local error curve*.

Next, we study gradient boosting. Gradient boosting, like stacking, learns a linear combination of base models, but unlike stacking, does not rely on independently trained models. The models in gradient boosting are found by adaptively and iteratively solving a "weak learning" problem. As our midpoint anchoring method does not rely on model independence, we are still able to use it to recover strong agreement bounds tending to 0 at a rate of $O(1/k)$, where $k$ is the number of iterations of gradient boosting.

The constituent models used in gradient boosting can be arbitrary and non-convex (e.g., depth 5 regression trees), but the aggregation method is still linear and is implicitly approximating a (infinite dimensional) convex optimization problem — minimizing mean squared error amongst linear models in the span of the set of weak learner models. One might wonder if the kind of agreement bounds we are able to prove are implicitly relying on this convexity. In our third and fourth applications, we see that the answer is no. We study error minimization over arbitrary ReLU neural networks of size $n$ (implying architecture search) as well as arbitrary regression trees of depth $d$. These are highly non-convex optimization landscapes. Thus approximate error minimizers can generally be very far from agreement in parameter space. Nevertheless, we are able to apply our midpoint anchoring method to show strong bounds on agreement that can be driven to $0$ as a function of the size of the neural network $n$ in the first case and the depth of the regression tree $d$ in the second case, recovering agreement in prediction space *despite* arbitrary disagreement in parameter space.

### 1.2. Interpreting Local Learning Curve Stability

Our results for stacking, neural network training, and regression tree training all have the form of local learning curve stability bounds: $D(f_1, f_2) \leq 4(R(\mathcal{F}_n) - R(\mathcal{F}_{2n}))$— where $R(\mathcal{F}_k)$ refers to the optimal error amongst models with "complexity" $k$ (parameterizing the number of models being ensembled, network size, and depth in the cases of stacking, neural networks, and regression trees respectively). These kinds of bounds are actionable and well aligned with optimizing for accuracy. They are actionable because (with enough data) it is possible to empirically plot the local learning curve by training with different parameter values, and picking $n$ such that the curve is locally flat — $R(\mathcal{F}_n) \approx R(\mathcal{F}_{2n})$. This is aligned with the goal of optimizing for accuracy since if we could substantially improve accuracy by locally increasing the complexity of the model, then in the high data regime, we should. It is also descriptive in the sense that if we assume that most deployed models are not "leaving money on the table" in the sense of being able to substantially improve accuracy by locally increasing complexity, then we should expect stability amongst deployed models. Because the error sequence $\{R(\mathcal{F}_n)\}_n$ is bounded from above and below and monotonically decreasing in $n$, the local learning curve is also guaranteed to "flatten out" to a value $\alpha$ for a value of $n$ that is independent of the problem complexity, and at most $2^{1/\alpha}$. However, in practice we expect the local learning curve to flatten out even more gracefully. Empirical studies of neural scaling laws (Kaplan et al., 2020; Hoffmann et al., 2022) have consistently found that across a wide variety of domains, the optimal error $R(\mathcal{F}_n)$ decreases as a power law in model complexity: $R(\mathcal{F}_n) \approx R^* + cn^{-\gamma}$ for some constants $c > 0$, $\gamma > 0$, and irreducible error $R^*$. Under such a power law, the *gap* in the local learning curve becomes: $R(\mathcal{F}_n) - R(\mathcal{F}_{2n}) = c\,(n^{-\gamma} - (2n)^{-\gamma}) = c(1 - 2^{-\gamma})n^{-\gamma} = O(n^{-\gamma})$. That is, the local learning curve gap shrinks polynomially in model complexity, which by our results implies that independent model disagreement $D(f_1, f_2)$ decreases at the same rate. Crucially, this does not require low absolute error rather only that the marginal benefit of increasing complexity diminishes. The exponent $\gamma$ varies by domain (typically $0.05$–$0.5$ for large-scale neural networks), but is reliably positive. Our results provide theoretical grounding for empirical observations that larger models exhibit greater prediction-level consistency across independent training runs (Bhojanapalli et al., 2021; Jordan, 2024), and may help explain the surprisingly high levels of agreement observed empirically across independently trained large language models (Gorecki and Hardt, 2025).

## 2. Preliminaries and Midpoint Anchoring Lemmas

We consider a setting in which we train two models on independently drawn datasets. Let $\mathcal{X} \subseteq \mathbb{R}^d$ be the data domain and $\mathcal{Y} \subseteq \mathbb{R}$ be the label domain. We assume access to datasets $S = ((x_i, y_i))_{i=0}^{n-1}$ that are independently drawn from a joint distribution $P$ on $\mathcal{X} \times \mathcal{Y}$. Note that unless otherwise stated, all expectations will be with respect to $x, y \sim P$ or where appropriate just the marginal over $x$. A model is then defined as a function mapping $f : \mathcal{X} \mapsto \mathcal{Y}$. We define the norm $\|f\| := \big(\mathbb{E}[f(x)^2]\big)^{1/2}$. With this we define the mean squared error objective and the corresponding population risk

$$\mathrm{MSE}(f) = \mathbb{E}[(y - f(x))^2], \quad R(\mathcal{F}) := \inf_{f \in \mathcal{F}} \mathrm{MSE}(f).$$

We next define the disagreement between two models as their expected squared difference.

**Definition 1 (Disagreement)** *For any two functions $f_1 : \mathcal{X} \mapsto \mathcal{Y}, f_2 : \mathcal{X} \mapsto \mathcal{Y}$, we define the expected disagreement between them as*

$$D(f_1, f_2) := \mathbb{E}[(f_1(x) - f_2(x))^2] \ .$$

We are now ready to state and prove a simple identity that will form the backbone of our analyses. It relates the *disagreement* between two models to the degree to which their errors could be improved by averaging the models.

**Lemma 2 (Midpoint identity for squared loss)** *For any two functions $f_1 : \mathcal{X} \mapsto \mathcal{Y}$ and $f_2 : \mathcal{X} \mapsto \mathcal{Y}$, let $\bar{f}(x) := \frac{1}{2}(f_1(x) + f_2(x))$. Then*

$$D(f_1, f_2) \ = \ 2\Big(\mathrm{MSE}(f_1) + \mathrm{MSE}(f_2) - 2\,\mathrm{MSE}(\bar{f})\Big).$$

**Proof** Let $r_i(x) := f_i(x) - y$ for $i \in \{1, 2\}$. Then $\bar{f}(x) - y = \frac{1}{2}(r_1(x) + r_2(x))$ and $f_1(x) - f_2(x) = r_1(x) - r_2(x)$. Expanding squares and using linearity of expectation gives

$$\mathbb{E}[(r_1 - r_2)^2] = \mathbb{E}[r_1^2] + \mathbb{E}[r_2^2] - 2\mathbb{E}[r_1 r_2].$$

On the other hand,

$$\mathbb{E}\Big[\Big(\tfrac{1}{2}(r_1 + r_2)\Big)^2\Big] = \tfrac{1}{4}\mathbb{E}[(r_1 + r_2)^2] = \tfrac{1}{4}\mathbb{E}[r_1^2 + r_2^2 + 2r_1 r_2] = \tfrac{1}{4}\mathbb{E}[r_1^2] + \tfrac{1}{4}\mathbb{E}[r_2^2] + \tfrac{1}{2}\mathbb{E}[r_1 r_2].$$

Therefore,

$$\begin{aligned} 2\Big(\mathbb{E}[r_1^2] + \mathbb{E}[r_2^2] - 2\mathbb{E}\big[\big(\tfrac{1}{2}(r_1 + r_2)\big)^2\big]\Big) &= 2\Big(\mathbb{E}[r_1^2] + \mathbb{E}[r_2^2] - 2\Big(\tfrac{1}{4}\mathbb{E}[r_1^2] + \tfrac{1}{4}\mathbb{E}[r_2^2] + \tfrac{1}{2}\mathbb{E}[r_1 r_2]\Big)\Big) \\ &= 2\big(\tfrac{1}{2}\mathbb{E}[r_1^2] + \tfrac{1}{2}\mathbb{E}[r_2^2] - \mathbb{E}[r_1 r_2]\big) \\ &= \mathbb{E}[r_1^2] + \mathbb{E}[r_2^2] - 2\mathbb{E}[r_1 r_2] = \mathbb{E}[(r_1 - r_2)^2]. \end{aligned}$$

Substituting back $\mathbb{E}[r_i^2] = \mathrm{MSE}(f_i)$ and $\mathbb{E}\big[\big(\frac{1}{2}(r_1 + r_2)\big)^2\big] = \mathrm{MSE}(\bar{f})$ yields the claim. ■

A useful corollary of this identity is that we can upper bound the disagreement between two models by the degree to which they are sub-optimal relative to the *best model* in any family that contains their average.

**Corollary 3 (Disagreement via the midpoint anchor)** *For any two functions $f_1, f_2 : \mathcal{X} \to \mathcal{Y}$, let $\bar{f}(x) := \frac{1}{2}(f_1(x) + f_2(x))$. If $\bar{f} \in \mathcal{H}$ for some class of predictors $\mathcal{H}$, then*

$$D(f_1, f_2) \;\le\; 2\big(\mathrm{MSE}(f_1) - R(\mathcal{H})\big) + 2\big(\mathrm{MSE}(f_2) - R(\mathcal{H})\big).$$

**Proof** By Lemma 2, we have

$$D(f_1, f_2) \;=\; 2\Big(\mathrm{MSE}(f_1) + \mathrm{MSE}(f_2) - 2\,\mathrm{MSE}(\bar{f})\Big).$$

If $\bar{f} \in \mathcal{H}$ then $\mathrm{MSE}(\bar{f}) \ge R(\mathcal{H})$, so substituting yields the claim. ■

If the model class from which $f_1$ and $f_2$ were trained contains their average, then we can relate the disagreement between $f_1$ and $f_2$ to the sub-optimality of the loss of $f_1$ and $f_2$ to the global optimum within the class in which they were trained. However, non-convex model classes will not satisfy this closure-under-averaging property. To analyze these classes it is useful to consider local learning-curve bounds with respect to a hierarchy of model classes, such that each level $\mathcal{F}_{2n}$ in the hierarchy is expressive enough to represent the average of any pair of models in $\mathcal{F}_n$. We will see that this property is satisfied by neural networks (where $n$ parametrizes the number of internal nodes) and regression trees (where $n$ parametrizes the depth).

**Lemma 4 (Local learning-curve bound from midpoint closure)** *Let $(\mathcal{F}_n)_{n\ge1}$ be a nested sequence of predictor classes and assume that for every $n$ and every $f_1, f_2 \in \mathcal{F}_n$, the midpoint predictor $\bar{f} := \frac{1}{2}(f_1 + f_2)$ lies in $\mathcal{F}_{2n}$. Fix $n \ge 1$ and suppose $f_1, f_2 \in \mathcal{F}_n$ satisfy $\mathrm{MSE}(f_i) \le R(\mathcal{F}_n) + \varepsilon$ for $i \in \{1, 2\}$. Then*

$$D(f_1, f_2) \;\le\; 4\big(R(\mathcal{F}_n) - R(\mathcal{F}_{2n}) + \varepsilon\big).$$

**Proof** By midpoint closure we have $\bar{f} \in \mathcal{F}_{2n}$, so Lemma 3 with $\mathcal{H} = \mathcal{F}_{2n}$ gives

$$D(f_1, f_2) \le 2\big(\mathrm{MSE}(f_1) - R(\mathcal{F}_{2n})\big) + 2\big(\mathrm{MSE}(f_2) - R(\mathcal{F}_{2n})\big).$$

Using $\mathrm{MSE}(f_i) \le R(\mathcal{F}_n) + \varepsilon$ for both $i$ yields the claim. ■

In the following sections, we apply Lemma 3 and Lemma 4 by verifying that the midpoint predictor lies in an appropriate hypothesis class.

## 3. Warmup Application: Stacking

Proofs in this section will are deferred to Appendix B.

Stacking is an ensembling method which first trains $k$ independent base models in some arbitrary fashion and then uses linear regression over these base models to combine their predictions.

Let $Q$ be a probability distribution on models of the form $g : \mathcal{X} \to \mathbb{R}$. Concretely, $Q$ could represent the law of a base predictor obtained by training a fixed learning algorithm $M$ on a *random shard* of the training sample of size $n/k$, with a fresh i.i.d. draw of examples and fresh algorithmic randomness; independent draws from $Q$ correspond to training $M$ on independent shards. We remark in passing that other intepretations of $Q$ also make sense. For example, perhaps all parties share the same training set (because e.g. it is the training set for a standard benchmark dataset like ImageNet). Then there is no need to have different models be trained on different shards, and $Q$ can

**Algorithm 1** Ensembling via Stacking

**Input:** $M : G \to \mathcal{H}$ black-box learning algorithm, $D \sim P^n$ dataset of size $n$, number of shards $k$

Randomly split $D$ into $k$ disjoint shards $G_i$ each of size $|G_i| = \left\lfloor \frac{n}{k} \right\rfloor$

**for** $i \in [k]$ **do**
  $g_i \leftarrow M(G_i)$
**end**
$f \leftarrow \mathrm{OLS}(g_1, \ldots, g_k)$ **return** $f$

represent only the randomness of the training procedure, which might re-use samples in arbitrary ways. We will analyze the population least squares predictor over the span of these base models. That is, we sample $k$ models $G = \{g_1, ..., g_k\} \sim Q^k$ and define $V(G)$ to be the linear span of the sampled models in $G$. We will consider the predictor

$$\arg \min_{f \in V(G)} \mathrm{MSE}(f).$$

Note that this is just a finite dimensional least squares problem, so a minimizer exists, and multiset multiplicities do not affect the span $V(G)$. For $t \in \mathbb{N}$, let $R_t$ denote the random variable $R(G)$ when $G = \{g_1, \ldots, g_t\}$ with $g_1, \ldots, g_t \overset{\text{i.i.d.}}{\sim} Q$, and write $\bar{R}_t := \mathbb{E}_{\{g_1,\ldots,g_t\}\sim Q^t}[R_t]$. We will use the shorthand $\bar{R}_t := \mathbb{E}_G[R_t]$

### 3.1. An Agreement Upper Bound

We instantiate our agreement upper bound for Stacking using the midpoint anchoring lemma. In this case we compare $f_1$ and $f_2$ to the risk $R(G^*)$ where $G^* := G \cup G'$ is the union of the base models used in training $f_1$ and $f_2$. Here $f_1$ is the MSE minimizer over the set of base models $G = \{g_1, ..., g_k\}$ and $f_2$ is the MSE minimizer over the set of base models $G' = \{g'_1, ..., g'_k\}$. We know that $V(G), V(G') \subseteq V(G \cup G')$, and that the midpoint predictor $\frac{1}{2}(f_1 + f_2)$ lies in $V(G \cup G')$. This, together with the fact that the set of $2k$ models in $G \cup G'$ is exchangeable lets us prove the following agreement bound:

**Theorem 5 (Agreement for Stacked Aggregation)** *Let* $G = \{g_1, \ldots, g_k\} \overset{\text{i.i.d.}}{\sim} Q^k$ *and* $G' = \{g'_1, \ldots, g'_k\} \overset{\text{i.i.d.}}{\sim} Q^k$ *be independent. Define* $f_1, f_2$ *as follows:*

$$f_1 = \arg \min_{f \in V(G)} \mathrm{MSE}(f), \quad f_2 = \arg \min_{f \in V(G')} \mathrm{MSE}(f)$$

*Then we have that*

$$\mathbb{E}_{f_1,f_2}\big[D(f_1, f_2)\big] \;\le\; 4\big(\bar{R}_k - \bar{R}_{2k}\big).$$

Note that Theorem 5 depends on the slope of the *local learning curve* at $k$: $(\bar{R}_k - \bar{R}_{2k})$. This is a strength; dependence on the *global* learning curve $(\bar{R}_k - R_\infty)$ would be significantly weaker. To see this, note that if $Q$ contained only a single "good model" with arbitrarily small weight, the global learning curve could fail to flatten out for arbitrarily large $k$. On the other hand, simply by monotonicity, for any value of $\alpha$, if labels are bounded in (say) $[0, 1]$ then there must be a value of $k \le 2^{1/\alpha}$ such that $(\bar{R}_k - \bar{R}_{2k}) \le \alpha$ (as error can drop by $\alpha$ at most $1/\alpha$ times before contradicting the non-negativity of squared error). While this depends exponentially on $\alpha$, it is independent of the dimensionality or complexity of the instance, in contrast to bounds depending on the global learning curve.

### 3.2. Stacking Lower Bound

Theorem 5 gives an upper bound with constant $4$. We now show that this factor cannot be improved in general: for every fixed $k$ and every $\varepsilon > 0$, there exists a data distribution $P$ and a distribution $Q$ over base models such that two independent stacking runs have disagreement at least $(4 - \varepsilon)$ times the gap $\bar{R}_k - \bar{R}_{2k}$.

**Theorem 6 (Near-tightness of the factor 4)** *Fix an integer $k \geq 1$. For every $\varepsilon > 0$, there exists a data distribution $P$ and a distribution $Q$ over base models such that if $G, G' \overset{\text{i.i.d.}}{\sim} Q^k$ are independent $k$–tuples and*

$$f_1 = \arg\min_{f \in V(G)} \mathrm{MSE}(f), \qquad f_2 = \arg\min_{f \in V(G')} \mathrm{MSE}(f),$$

*then*

$$\mathbb{E}_{f_1, f_2}\big[D(f_1, f_2)\big] \;\geq\; (4 - \varepsilon)\,\big(\bar{R}_k - \bar{R}_{2k}\big).$$

## 4. Gradient Boosting

In this section we apply our midpoint anchoring argument to *gradient boosting*, an algorithm that iteratively builds up an ensemble model by repeatedly chooses a weak learning model $g \in \mathcal{C}$ that correlates with the residual of our current ensemble model and then adds $g$ to it. Unlike stacking, the models that make up two independently trained ensembles $f_1$ and $f_2$ are *not* exchangeable, since the weak learners are not selected independently, but rather *adaptively* in a path dependent way. Nevertheless, we show that we can apply midpoint anchoring to drive disagreement to $0$ at a $1/k$ rate (where $k$ is the number of iterations of gradient boosting). Here we abstract away finite sample issues by modeling our weak learning algorithm in the style of an SQ oracle (Kearns, 1998) — i.e. rather than obtaining the $g \in \mathcal{C}$ which exactly maximizes covariance with the residuals of our current model, it can return any $g \in \mathcal{C}$ that is an $\epsilon$-approximate maximizer. This models e.g. solving an ERM problem over any sample that is sufficient for $\varepsilon$-approximate uniform convergence over $\mathcal{C}$. Proofs in this section are deferred to Appendix C.

We assume for simplicity that our weak learning class $\mathcal{C}$ satisfies the following mild regularity conditions (which are enforceable under a boundedness assumption discussed below): Symmetry ($g \in \mathcal{C} \Rightarrow -g \in \mathcal{C}$), normalization ($\|g\| \leq 1$ for all $g \in \mathcal{C}$) and non-degeneracy ($0 \notin \mathcal{C}$). We will use the normalization condition with respect to the Atomic and Euclidean Norm, which can be enforced by dividing the original functions (unnormalized) by the maximum of its Atomic norm, Euclidean norm, and 1 when the these norms are bounded. We will assume that there exists $B < \infty$ such that $||g|| \leq B \quad \forall g \in C$. For simplicity, we shall state all bounds and analyses after rescaling the atom class such that $B = 1$. This rescaling is of the weak-learner directions rather than of the loss function itself, which is only valid under the boundedness assumption above. This rescaling preserves the linear span $V(C)$ and therefore preserves $R(V(C))$. However, it changes the scale of the induced atomic norm. Therefore, quantities like $\tau^*$ should be interpreted with respect to the normalized atomic class. Note in this section, for the sake of clarity, we will use the standard inner product $\langle f, g \rangle = f^T g$. When we list $||f||$ it will still corresponding to the norm we defined in the Preliminaries of $(\mathbb{E}[f(x)^2])^{1/2}$. When needed, we will explicitly mention the expectations we are computing. We model weak-learning via an $\varepsilon$-approximate SQ-style oracle: at iteration $t$, the oracle returns any $g_t \in \mathcal{C}$ such that

$$\mathbb{E}[\langle r_{t-1}(x), g_t(x) \rangle] \;\geq\; \sup_{g \in \mathcal{C}} \mathbb{E}[\langle r_{t-1}(x), g(x) \rangle] \;-\; \varepsilon_t, \quad r_{t-1} := y - f_{t-1}.$$

**Algorithm 2** Gradient Boosting

**Input:** SQ-oracle for weak learner class $\mathcal{C}$

$f_0 \equiv 0$, $G_0 = \emptyset$
**for** $t \in [k]$ **do**
  $r_{t-1} := y - f_{t-1}$
  Choose $g_t \in \mathcal{C}$ with $\mathbb{E}[\langle r_{t-1}(x), g_t(x)\rangle] \geq \sup_{g\in\mathcal{C}} \mathbb{E}[\langle r_{t-1}(x), g(x)\rangle] - \varepsilon_t$. (SQ-oracle)
  $\alpha_t := \arg\min_{\alpha\in\mathbb{R}} \mathbb{E}[(r_{t-1}(x) - \alpha g_t(x))^2] = \mathbb{E}[\langle r_{t-1}(x), g_t(x)\rangle]/\|g_t\|^2$
  $f_t := f_{t-1} + \alpha_t g_t$
  set $G_t := G_{t-1} \cup \{g_t\}$
**end**
**return** $f_k$ and $G := G_k$

Algorithm 2 provides the details of how to use this oracle within the Gradient Boosting procedure. We will be interested in comparing the MSE of the gradient boosting iterates with the risk of the best minimizer in the weak learner class $R\big(V(\mathcal{C})\big) := \inf_{f\in V(\mathcal{C})} \mathrm{MSE}(f)$. We will bound the disagreement of two independently trained models $f_1$ and $f_2$ by anchoring to the best model $f^*$ in the span of the weak learner class $\mathcal{C}$, and then apply our anchoring lemma from Section 2. Since anchoring bounds disagreement in terms of each model's error gap to $f^*$, it remains to upper bound that gap. We do so below, starting by bounding the single-step error improvement of gradient boosting.

**Lemma 7 (Single Iterate Progress)** *With $\alpha_t = \arg\min_{\alpha\in\mathbb{R}} \|r_{t-1} - \alpha g_t\|^2$ and $\|g_t\| \leq 1$,*

$$\mathrm{MSE}(f_{t-1}) - \mathrm{MSE}(f_t) \;\geq\; \mathbb{E}[\langle r_{t-1}(x), g_t(x)\rangle]^2.$$

**Proof** Note that $\mathrm{MSE}(f_{t-1}) - \mathrm{MSE}(f_t) = ||r_{t-1}||^2 - ||r_t||^2 = ||r_{t-1}||^2 - ||r_{t-1} - \alpha_t g_t||^2$ By exact line search,

$$\begin{aligned}
\|r_{t-1} - \alpha_t g_t\|^2 &= \min_{\alpha} \|r_{t-1} - \alpha g_t\|^2 \\
&= \min_{\alpha}(||r_{t-1}||^2 - 2\alpha\mathbb{E}[\langle r_{t-1}(x), g_t(x)\rangle] + \alpha^2||g_t||^2) \\
&= ||r_{t-1}||^2 - 2\mathbb{E}[\langle r_{t-1}(x), g_t(x)\rangle]^2/\|g_t\|^2 + \mathbb{E}[\langle r_{t-1}(x), g_t(x)\rangle]^2/\|g_t\|^2 \\
&= ||r_{t-1}||^2 - \mathbb{E}[\langle r_{t-1}(x), g_t(x)\rangle]^2/\|g_t\|^2
\end{aligned}$$

Therefore, we have that $\mathrm{MSE}(f_{t-1}) - \mathrm{MSE}(f_t) = \mathbb{E}[\langle r_{t-1}(x), g_t(x)\rangle]^2/\|g_t\|^2$. Using $\|g_t\| \leq 1$ gives the stated bound. ■

Now, we define the radius $\tau > 0$ with the corresponding convex hull $\mathcal{K}_\tau := \tau\mathrm{conv}(\mathcal{C})$. Let $f^* \in V(\mathcal{C})$ be the population least-squares minimizer over the span of the weak learning class. Define the corresponding atomic norm radius $\tau^* := \|f^*\|_{\mathcal{A}}$, where the atomic norm induced by $\mathcal{C}$ is

$$\|f\|_{\mathcal{A}} \;:=\; \inf\Big\{ \sum_{j=1}^{k} |\alpha_j| :\; f = \lim_{k\to\infty} \sum_{j=1}^{k} \alpha_j g_j,\; g_j \in \mathcal{C}, \sum_{j=1}^{k} |\alpha_j| \leq \infty \Big\}.$$

That is, $\tau^*$ corresponds to the smallest total weight needed to represent $f^*$ within the weak-learner class. Throughout the remainder of the paper, $|\cdot|_A$ and $\tau^*$ should be interpreted with respect to the

normalized weak learner class. To interpret bounds with respect to the unnormalized weak learner class, $\tau^*_{\texttt{unnormalized}} = B\tau^*$. We have now related the MSE gap between the models of two runs in terms of the square of the max correlation of the residuals of the earlier model with a model in the weak learner class. Next, we will lower bound the largest possible correlation between the residuals of a model $f$ and a function in the weak learner class in terms of the difference between the MSE of the current model $f$ and the error of the best model in the span of the weak learners, scaled by the atomic norm of $f^*$.

**Lemma 8 (Correlation Lower Bound w.r.t. Weak Learning Anchor Gap)** *For any $f$, writing $M(f) := \sup_{g\in\mathcal{C}} |\mathbb{E}[\langle y - f, g\rangle]|$, we have*

$$M(f) \;\geq\; \frac{\mathrm{MSE}(f) - R\big(V(\mathcal{C})\big)}{2\,\tau^*}.$$

We have lower bounded the maximum residual–model correlation over the weak learner class by a quantity depending on the gap between the current model's error and the best error in the weak-learner span. We now relate the per-step error gap to that best error via a recurrence.

**Proposition 9 (Gap Recurrence Toward $R\big(V(\mathcal{C})\big)$)** *Let $E_t := \mathrm{MSE}(f_t) - R\big(V(\mathcal{C})\big)$. We will use the shorthand $u_+^2 = (\max\{u, 0\})^2$. Then, for $t \geq 1$,*

$$E_{t-1} - E_t \;\geq\; \Big(\tfrac{E_{t-1}}{2\tau^*} - \varepsilon_t\Big)_+^2.$$

**Proof** By Lemma 7, $\mathrm{MSE}(f_{t-1}) - \mathrm{MSE}(f_t) \geq \mathbb{E}[\langle r_{t-1}(x), g_t(x)\rangle]^2$. The oracle gives $\mathbb{E}[\langle r_{t-1}(x), g_t(x)\rangle] \geq M(f_{t-1}) - \varepsilon_t$. Hence $\mathbb{E}[\langle r_{t-1}, g_t\rangle]^2 \geq (M(f_{t-1}) - \varepsilon_t)_+^2$. Finally, Lemma 8 gives $M(f_{t-1}) \geq E_{t-1}/(2\tau^*)$, yielding the claim. ■

Finally, we can use the recurrence relation to bound the difference between the MSE of the model at iteration $t$ and the MSE of the best model in the span of the weak learner class–we can see that the first term is inversely proportional to $t$ and depends on the atomic norm of the best model in span of the weak learner class. It also includes a term that depends on the SQ-oracle error at every iteration.

**Theorem 10 (Weak Learning Anchor Gap Upper Bound)** *For all $t \geq 1$,*

$$\mathrm{MSE}(f_t) - R\big(V(\mathcal{C})\big) \;\leq\; \frac{8\,(\tau^*)^2}{t} \;+\; \sum_{s=1}^{t} \varepsilon_s^2.$$

Recall that the displayed bounds are displayed with respect to the normalized units. Since the SQ errors are also defined with respect to the normalized problem, to interpret the disagreement upper bounds with respect to the unnormalized class, one can multiply these bounds by a factor of $B^2$ when $B \geq 1$. We can now use the anchoring lemmas from Section 2 to relate two independent stagewise runs.

**Theorem 11 (Gradient Boosting Agreement Bound)** *Let $f_1$ and $f_2$ be two independent gradient boosting runs (using the same weak learning class $\mathcal{C}$ and number of iterations $k$) driven by $\{\varepsilon_t\}$ and $\{\varepsilon'_t\}$ respectively. Let $f^* \in V(\mathcal{C})$ denote the population least-squares predictor over $V(\mathcal{C})$. Then*

$$D(f_1, f_2) \;\leq\; 2\big(\mathrm{MSE}(f_1) - R\big(V(\mathcal{C})\big)\big) \;+\; 2\big(\mathrm{MSE}(f_2) - R\big(V(\mathcal{C})\big)\big).$$

*Consequently, using Theorem 10, for all $k \geq 1$,*

$$D(f_1, f_2) \;\leq\; \frac{32\,(\tau^*)^2}{k} \;+\; 2\Big(\sum_{t=1}^{k} \varepsilon_t^2 \;+\; \sum_{t=1}^{k} \varepsilon_t'^2\Big).$$

**Proof** Let $\bar{f} := \frac{1}{2}(f_1 + f_2)$. Since each gradient boosting run outputs a predictor in $V(\mathcal{C})$, we have $\bar{f} \in V(\mathcal{C})$. Applying Lemma 3 with $\mathcal{H} = V(\mathcal{C})$ gives

$$D(f_1, f_2) \;\leq\; 2\big(\mathrm{MSE}(f_1) - R(V(\mathcal{C}))\big) \;+\; 2\big(\mathrm{MSE}(f_2) - R(V(\mathcal{C}))\big).$$

Applying Theorem 10 to both runs yields the stated bound. ■

Thus we have shown that gradient boosting yields independent agreement tending to 0 at a rate of $O(1/k)$, where $k$ is the number of iterations. This bound also depends on $\tau^*$, which is a problem-dependent constant. In Section 6 we analyze a variant of gradient boosting based on the Frank Wolfe algorithm (for more general loss functions) that always produces a predictor that has norm at most $\tau$, where $\tau$ is a user defined parameter. We give a variant of this analysis in which we anchor to the best model in the span of the weak learner class that also has norm at most $\tau$. This removes any dependence on $\tau^*$, and obtain similar rates depending only on $\tau$ — replacing the problem dependent constant with a user defined parameter that trades of agreement with accuracy as desired.

## 5. Neural Networks, Regression Trees, and Other Classes Satisfying Hierarchical Midpoint Closure

Next, we show that certain function classes including ReLU neural networks and regression trees admit strong agreement bounds under approximate population loss minimization. These function classes may be highly non-convex, meaning that approximate loss minimizers may be very far in parameter space—or even incomparable in the sense that they may be of different architectures. Nevertheless, by anchoring on the midpoint predictor $\bar{f}(x) = \frac{1}{2}(f_1(x) + f_2(x))$ and using that the relevant model classes are closed under averaging we show that they must be close in *prediction space*. We will use Lemma 4 from Section 2. To apply it, we need midpoint closure of the form $\bar{f} \in \mathcal{F}_{2n}$ whenever $f_1, f_2 \in \mathcal{F}_n$. The form of our theorems will be identical for any class satisfying this kind of "hierarchical midpoint closure". Proofs from this section are deferred to Appendix D.

### 5.1. Application to Neural Networks

We work with feed-forward ReLU networks. Let $\sigma(t) := \max\{0, t\}$ denote the ReLU activation. For $n \geq 0$, let $\mathrm{NN}_n$ denote the class of functions $f : \mathcal{X} \to \mathcal{Y}$ computable by a finite directed acyclic graph in which each internal (non-input, non-output) node computes $\sigma(\langle w, u\rangle + b)$ for some affine function of its inputs, and the output node computes an affine combination of the values at the input coordinates and internal nodes. First, we demonstrate midpoint closure for this class.

**Lemma 12 (Neural-network midpoint closure)** *For every $n \geq 0$ and every $f_1, f_2 \in \mathrm{NN}_n$, the midpoint predictor $\bar{f} := \frac{1}{2}(f_1 + f_2)$ lies in $\mathrm{NN}_{2n}$.*

**Proof** Fix realizations of $f_1$ and $f_2$ as ReLU networks with at most $n$ internal nodes each. Construct a new network by taking a disjoint copy of the internal computation graph for each of $f_1$ and $f_2$, and wiring both copies to the same input $x$. This yields a single feed-forward network that computes both $f_1(x)$ and $f_2(x)$ in parallel, using at most $2n$ internal ReLU nodes.

Define the output node to return the affine combination $\frac{1}{2}f_1(x) + \frac{1}{2}f_2(x)$. This adds no new internal nodes, so the resulting network computes $\bar{f}$ and has size at most $2n$, i.e., $\bar{f} \in \mathrm{NN}_{2n}$. ■

**Corollary 13 (Neural-network agreement)** *Fix $n \geq 1$ and $\varepsilon > 0$. If $f_1, f_2 \in \mathrm{NN}_n$ satisfy $\mathrm{MSE}(f_i) \leq R(\mathrm{NN}_n) + \varepsilon$ for $i \in \{1, 2\}$, then*

$$D(f_1, f_2) \leq 4\big(R(\mathrm{NN}_n) - R(\mathrm{NN}_{2n}) + \varepsilon\big).$$

**Proof** Apply Lemma 4 with $\mathcal{F}_n = \mathrm{NN}_n$ and use Lemma 12. ■

Observe that this is exactly the same form of local learning curve guarantee that we got for Stacking in Theorem 5. In particular, as loss is bounded and optimal loss is monotonically decreasing in network size, for any value of $\alpha$, there must be a value of $n \leq 2^{1/\alpha}$ such that $R(\mathrm{NN}_n) - R(\mathrm{NN}_{2n}) \leq \alpha$ (as error can drop by $\alpha$ at most $1/\alpha$ times before contradicting the non-negativity of squared error). For such a value of $n$, we have $D(f_1, f_2) \leq 4(\alpha + \epsilon)$. As with stacking, this bound is completely independent of the complexity of the instance and does not require that "global optimality" can be obtained by a small neural network (i.e. it requires only flatness of the local loss curve, which can always be guaranteed at modest values of $n$, not the global loss curve, which cannot). This kind of "learning curve" bound for neural networks is reminiscent of the argument used by Błasiok et al. (2024) to show that "most sizes" of ReLU networks are approximately multicalibrated with respect to all neural network architectures of bounded size.

### 5.2. Application to Regression Trees

We observe that the same arguments apply almost verbatim to regression trees. We work with axis-aligned regression trees. A depth-$d$ tree is a rooted binary tree in which every internal node is labeled by a coordinate $j \in [d]$ and a threshold $t \in \mathbb{R}$, and routes an input $x \in \mathcal{X} \subseteq \mathbb{R}^d$ to the left or right child depending on whether $x_j \leq t$ or $x_j > t$. Each leaf is labeled by a constant prediction value in $[0, 1]$. The predictor computed by the tree is the leaf value reached by $x$. We write $\mathsf{Tree}_d$ for the class of such predictors of depth at most $d$.

**Lemma 14 (Regression-tree midpoint closure)** *For every $d \geq 0$ and every $f_1, f_2 \in \mathsf{Tree}_d$, the midpoint predictor $\bar{f} := \frac{1}{2}(f_1 + f_2)$ lies in $\mathsf{Tree}_{2d}$.*

We now get an immediate corollary:

**Corollary 15 (Regression tree agreement)** *Fix $d \geq 1$ and $\varepsilon > 0$. If $f_1, f_2 \in \mathsf{Tree}_d$ satisfy $\mathrm{MSE}(f_i) \leq R(\mathsf{Tree}_d) + \varepsilon$ for $i \in \{1, 2\}$, then*

$$D(f_1, f_2) \leq 4\big(R(\mathsf{Tree}_d) - R(\mathsf{Tree}_{2d}) + \varepsilon\big).$$

**Proof** Apply Lemma 4 with $\mathcal{F}_d = \mathsf{Tree}_d$ and use Lemma 14. ■

Again, this is a local learning curve agreement guarantee of exactly the same form as our theorem for Stacking (Theorem 5) and our theorem for neural network training (Corollary 13). An immediate implication is that for any value of $\alpha$ that there is a value $d \leq 2^{1/\alpha}$ (i.e. independent of the complexity of the instance) guaranteeing that for that value of $d$, $D(f_1, f_2) \leq 4(\alpha + \epsilon)$.

**Algorithm 3** Multi-Dimensional Frank–Wolfe

**Input:** SQ-oracle for weak learner class $\mathcal{C}$, budget $\tau > 0$

$f_0 \equiv 0$, $G_0 = \emptyset$ **for** $t \in [k]$ **do**

Choose $s_t \in \mathcal{C}$ such that $\mathbb{E}[\langle -\nabla_p \mathcal{L}(y, f_{t-1}(x)), s_t(x)\rangle] \geq \max_{s\in\mathcal{C}} \mathbb{E}[\langle -\nabla_p \mathcal{L}(y, f_{t-1}(x)), s(x)\rangle] - \varepsilon_t$

Choose $g_t \in \mathcal{K}_\tau$ such that $g_t = \tau s_t / \|s_t\|_{\mathcal{A}}$

$\alpha_t = \frac{2}{t+1}$

$f_t := f_{t-1} + \alpha_t(g_t - f_{t-1})$; $G_t := G_{t-1} \cup \{g_t\}$

**end**

**return** $f_k$ and $G := G_k$

## 6. Generalizations

In Appendix E we generalize all of our results to multi-dimensional convex loss functions. In Appendix F, we generalize our results to cross-entropy loss. Here we focus on one of those generalizations, which also corresponds to a new algorithm. Our gradient boosting result in Section 4 gave disagreement bounds that diminished at a $O(1/k)$ rate with the number of iterations $k$, but that also depended on a problem-dependent constant $\tau^*$. Below we state our theorem for a generalized Frank-Wolfe based gradient boosting algorithm (details in Appendix E) which maintains as its hypothesis a linear combination of weak learners of norm at most $\tau$, where $\tau$ is a user-specified parameter. Our disagreement bound now scales with the user-defined parameter $\tau$ rather than $\tau^*$, and hence is universally bounded. We state the theorem below in the more general setting studied in Appendix E.

**Theorem 16** *(FW Gradient Boosting Agreement Bound) Fix any $\mathcal{L}$ that is $L$-smooth and $\mu$-strongly convex. Let $f_1, f_2$ be the output of any two runs of Algorithm 3 parameterized with the same $\tau, k, \mathcal{C}$ such that the sequence of SQ oracle errors are $\{\varepsilon_t, \varepsilon'_t\}_{t\in[k]}$ respectively. Let $f^* = \arg\min_{f\in\mathcal{K}_\tau} R(f)$. Then, we have that*

$$D(f_1, f_2) \leq \frac{64L\tau^2}{\mu(k+1)} + \frac{8\tau}{\mu(k+1)}\left(\sum_{j=1}^{k}\varepsilon_j + \sum_{j=1}^{k}\varepsilon'_j\right)$$

## Acknowledgments

EE and MH were partially supported by DARPA grant #HR00112420305. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views, position, or policy of DARPA or the US Government. AR supported in part by the Simons Collaboration on Algorithmic Fairness and the NSF EnCORE Tripods Institute. SS supported in part by an NSF Graduate Research Fellowship.

## References

Scott Aaronson. The complexity of agreement. In *Proceedings of the thirty-seventh annual ACM symposium on Theory of computing*, pages 634–643, 2005.

Samuel Ainsworth, Jonathan Hayase, and Siddhartha Srinivasa. Git re-basin: Merging models modulo permutation symmetries. In *The Eleventh International Conference on Learning Representations*, 2023.

Noga Alon, Roi Livni, Maryanthe Malliaris, and Shay Moran. Private pac learning implies finite littlestone dimension. In *Proceedings of the 51st Annual ACM SIGACT Symposium on Theory of Computing*, pages 852–860, 2019.

Robert J Aumann. Agreeing to disagree. *The Annals of Statistics*, 4(6):1236–1239, 1976.

Christina Baek, Yiding Jiang, Aditi Raghunathan, and J. Zico Kolter. Agreement-on-the-line: Predicting the performance of neural networks under distribution shift. In *Advances in Neural Information Processing Systems*, volume 35, 2022.

Dara Bahri and Heinrich Jiang. Locally adaptive label smoothing improves predictive churn. In *Proceedings of the 38th International Conference on Machine Learning*, 2021.

Yamini Bansal, Preetum Nakkiran, and Boaz Barak. Revisiting model stitching to compare neural representations. In *Advances in Neural Information Processing Systems*, 2021.

Srinadh Bhojanapalli, Kimberly Wilber, Andreas Veit, Ankit Singh Rawat, Seungyeon Kim, Aditya Menon, and Sanjiv Kumar. On the reproducibility of neural network predictions. *arXiv preprint arXiv:2102.03349*, 2021.

Emily Black, Manish Raghavan, and Solon Barocas. Model multiplicity: Opportunities, concerns, and solutions. In *Proceedings of the 2022 ACM conference on fairness, accountability, and transparency*, pages 850–863, 2022.

Jarosław Błasiok, Parikshit Gopalan, Lunjia Hu, Adam Tauman Kalai, and Preetum Nakkiran. Loss minimization yields multicalibration for large neural networks. In *15th Innovations in Theoretical Computer Science Conference (ITCS 2024)*, volume 287, pages 17–1. Schloss Dagstuhl–Leibniz-Zentrum f"ur Informatik, 2024.

Olivier Bousquet and André Elisseeff. Stability and generalization. *Journal of machine learning research*, 2(Mar):499–526, 2002.

Leo Breiman. Statistical modeling: The two cultures (with comments and a rejoinder by the author). *Statistical science*, 16(3):199–231, 2001.

Mark Bun, Roi Livni, and Shay Moran. An equivalence between private classification and online prediction. In *2020 IEEE 61st Annual Symposium on Foundations of Computer Science (FOCS)*, pages 389–402. IEEE, 2020.

Mark Bun, Marco Gaboardi, Max Hopkins, Russell Impagliazzo, Rex Lei, Toniann Pitassi, Satchit Sivakumar, and Jessica Sorrell. Stability is stable: Connections between replicability, privacy, and adaptive generalization. In *Proceedings of the 55th Annual ACM Symposium on Theory of Computing*, pages 520–527, 2023.

Andrea Caponnetto and Alexander Rakhlin. Stability properties of empirical risk minimization over donsker classes. *Journal of Machine Learning Research*, 7(138):2565–2583, 2006.

Zachary Charles and Dimitris Papailiopoulos. Stability and generalization of learning algorithms that converge to global optima. In *International conference on machine learning*, pages 745–754. PMLR, 2018.

Natalie Collina, Surbhi Goel, Varun Gupta, and Aaron Roth. Tractable agreement protocols. In *Proceedings of the 57th Annual ACM Symposium on Theory of Computing*, pages 1532–1543, 2025.

Natalie Collina, Ira Globus-Harris, Surbhi Goel, Varun Gupta, Aaron Roth, and Mirah Shi. Collaborative prediction: Tractable information aggregation via agreement. In *Proceedings of the ACM-SIAM Symposium on Discrete Algorithms*, 2026.

Rachel Cummings, Katrina Ligett, Kobbi Nissim, Aaron Roth, and Zhiwei Steven Wu. Adaptive learning with robust generalization guarantees. In *Conference on Learning Theory*, pages 772–814. PMLR, 2016.

Ilias Diakonikolas, Jingyi Gao, Daniel Kane, Sihan Liu, and Christopher Ye. Replicable distribution testing. *arXiv preprint arXiv:2507.02814*, 2025.

Kate Donahue, Alexandra Chouldechova, and Krishnaram Kenthapadi. Human-algorithm collaboration: Achieving complementarity and avoiding unfairness. In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*, pages 1639–1656, 2022.

Felix Draxler, Kambis Veschgini, Manfred Salmhofer, and Fred Hamprecht. Essentially no barriers in neural network energy landscape. In *Proceedings of the 35th International Conference on Machine Learning*, 2018.

Cynthia Dwork and Aaron Roth. The algorithmic foundations of differential privacy. *Foundations and trends® in theoretical computer science*, 9(3–4):211–407, 2014.

Cynthia Dwork, Frank McSherry, Kobbi Nissim, and Adam Smith. Calibrating noise to sensitivity in private data analysis. In *Theory of cryptography conference*, pages 265–284. Springer, 2006.

Cynthia Dwork, Vitaly Feldman, Moritz Hardt, Toniann Pitassi, Omer Reingold, and Aaron Leon Roth. Preserving statistical validity in adaptive data analysis. In *Proceedings of the forty-seventh annual ACM symposium on Theory of computing*, pages 117–126, 2015.

Eric Eaton, Marcel Hussing, Michael Kearns, and Jessica Sorrell. Replicable reinforcement learning. *Advances in Neural Information Processing Systems*, 36:15172–15185, 2023.

Eric Eaton, Marcel Hussing, Michael Kearns, Aaron Roth, Sikata Bela Sengupta, and Jessica Sorrell. Replicable reinforcement learning with linear function approximation. In *The Fourteenth International Conference on Learning Representations*, 2026.

Rahim Entezari, Hanie Sedghi, Olga Saukh, and Behnam Neyshabur. The role of permutation invariance in linear mode connectivity of neural networks. In *International Conference on Learning Representations*, 2022.

Rafael Frongillo, Eric Neyman, and Bo Waggoner. Agreement implies accuracy for substitutable signals. In *Proceedings of the 24th ACM Conference on Economics and Computation*, pages 702–733, 2023.

Timur Garipov, Pavel Izmailov, Dmitrii Podoprikhin, Dmitry P Vetrov, and Andrew G Wilson. Loss surfaces, mode connectivity, and fast ensembling of dnns. *Advances in neural information processing systems*, 31, 2018.

John D Geanakoplos and Heraklis M Polemarchakis. We can't disagree forever. *Journal of Economic theory*, 28(1):192–200, 1982.

Mila Gorecki and Moritz Hardt. Monoculture or multiplicity: Which is it? In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025.

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

Christopher Hidey, Fei Liu, and Rahul Goel. Reducing model churn: Stable re-training of conversational agents. In Oliver Lemon, Dilek Hakkani-Tur, Junyi Jessy Li, Arash Ashrafzadeh, Daniel Hernández Garcia, Malihe Alikhani, David Vandyke, and Ondřej Dušek, editors, *Proceedings of the 23rd Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pages 14–25, Edinburgh, UK, September 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.sigdial-1.2. URL https://aclanthology.org/2022.sigdial-1.2/.

Jordan Hoffmann, Sebastian Borgeaud, Arthur Mensch, Elena Buchatskaya, Trevor Cai, Eliza Rutherford, Diego de Las Casas, Lisa Anne Hendricks, Johannes Welbl, Aidan Clark, et al. Training compute-optimal large language models. In *Proceedings of the 36th International Conference on Neural Information Processing Systems*, pages 30016–30030, 2022.

Max Hopkins, Russell Impagliazzo, and Christopher Ye. Approximate replicability in learning. *arXiv preprint arXiv:2510.20200*, 2025.

Russell Impagliazzo, Rex Lei, Toniann Pitassi, and Jessica Sorrell. Reproducibility in learning. In *Proceedings of the 54th annual ACM SIGACT symposium on theory of computing*, pages 818–831, 2022.

Arthur Jacot, Franck Gabriel, and Clement Hongler. Neural tangent kernel: Convergence and generalization in neural networks. In *Advances in Neural Information Processing Systems*, 2018.

Yiding Jiang, Vaishnavh Nagarajan, Christina Baek, and J Zico Kolter. Assessing generalization of SGD via disagreement. In *International Conference on Learning Representations*, 2022. URL https://openreview.net/forum?id=WvOGCEAQhxl.

Rie Johnson and Tong Zhang. Inconsistency, instability, and generalization gap of deep neural network training. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023.

Keller Jordan. On the variance of neural network training with respect to test sets and distributions. In *The Twelfth International Conference on Learning Representations*, 2024.

Alkis Kalavasis, Amin Karbasi, Kasper Green Larsen, Grigoris Velegkas, and Felix Zhou. Replicable learning of large-margin halfspaces. *arXiv preprint arXiv:2402.13857*, 2024a.

Alkis Kalavasis, Amin Karbasi, Grigoris Velegkas, and Felix Zhou. On the computational landscape of replicable learning. *Advances in Neural Information Processing Systems*, 37:105887–105927, 2024b.

Jared Kaplan, Sam McCandlish, Tom Henighan, Tom B Brown, Benjamin Chess, Rewon Child, Scott Gray, Alec Radford, Jeffrey Wu, and Dario Amodei. Scaling laws for neural language models. *arXiv preprint arXiv:2001.08361*, 2020.

Amin Karbasi, Grigoris Velegkas, Lin Yang, and Felix Zhou. Replicability in reinforcement learning. *Advances in Neural Information Processing Systems*, 36:74702–74735, 2023.

Michael Kearns. Efficient noise-tolerant learning from statistical queries. *Journal of the ACM (JACM)*, 45(6):983–1006, 1998.

Michael Kearns, Aaron Roth, and Emily Ryu. Networked information aggregation via machine learning. In *Proceedings of the ACM-SIAM Symposium on Discrete Algorithms*, 2026.

Gil Kur, Eli Putterman, and Alexander Rakhlin. On the variance, admissibility, and stability of empirical risk minimization. *Advances in Neural Information Processing Systems*, 36:37527–37539, 2023.

Jaehoon Lee, Lechao Xiao, Samuel Schoenholz, Yasaman Bahri, Roman Novak, Jascha Sohl-Dickstein, and Jeffrey Pennington. Wide neural networks of any depth evolve as linear models under gradient descent. In *Advances in Neural Information Processing Systems*, 2019.

Jialin Mao, Itay Griniasty, Han Kheng Teoh, Rahul Ramesh, Rubing Yang, Mark K. Transtrum, James P. Sethna, and Pratik Chaudhari. The training process of many deep networks explores the same low-dimensional manifold. *Proceedings of the National Academy of Sciences*, 2024.

Charles Marx, Flavio Calmon, and Berk Ustun. Predictive multiplicity in classification. In *International conference on machine learning*, pages 6765–6774. PMLR, 2020.

Mahdi Milani Fard, Quentin Cormier, Kevin Canini, and Maya Gupta. Launch and iterate: Reducing prediction churn. *Advances in Neural Information Processing Systems*, 29, 2016.

Preetum Nakkiran and Yamini Bansal. Distributional generalization: A new kind of generalization. *arXiv preprint arXiv:2009.08092*, 2020.

Kenny Peng, Nikhil Garg, and Jon Kleinberg. A no free lunch theorem for human-ai collaboration. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pages 14369–14376, 2025.

Aaron Roth and Alexander Williams Tolbert. Resolving the reference class problem at scale. *Philosophy of Science*, pages 1–15, 2025.

Gowthami Somepalli, Liam Fowl, Arpit Bansal, Ping Yeh-Chiang, Yehuda Dar, Richard Baraniuk, Micah Goldblum, and Tom Goldstein. Can neural nets learn the same model twice? investigating reproducibility and double descent from the decision boundary perspective. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, June 2022.

Lijing Wang, Dipanjan Ghosh, Maria Teresa Gonzalez Diaz, Ahmed Farahat, Mahbubul Alam, Chetan Gupta, Jiangzhuo Chen, and Madhav Marathe. Wisdom of the ensemble: improving consistency of deep learning models. In *Proceedings of the 34th International Conference on Neural Information Processing Systems*, 2020.

Jamelle Watson-Daniels, Flavio du Pin Calmon, Alexander D'Amour, Carol Long, David C Parkes, and Berk Ustun. Predictive churn with the set of good models. *arXiv preprint arXiv:2402.07745*, 2024.

Danny Wood, Tingting Mu, Andrew M Webb, Henry WJ Reeve, Mikel Luján, and Gavin Brown. A unified theory of diversity in ensemble learning. *Journal of machine learning research*, 24(359): 1–49, 2023.

Zhanpeng Zhou, Yongyi Yang, Xiaojiang Yang, Junchi Yan, and Wei Hu. Going beyond linear mode connectivity: The layerwise linear feature connectivity. In *Advances in Neural Information Processing Systems*, 2023.

## Appendix A. Additional Related Work

**Agreement via Interaction** A line of work inspired by Aumann (1976) aims to give interactive test-time protocols through which two models (trained initially on different observations) can arrive at (accuracy improving) agreement. Initial work in economics (Geanakoplos and Polemarchakis, 1982) focused on exact agreement, but more recent work in computer science focused on interactions of bounded length, leading to approximate agreement of the same form that we study here (Aaronson, 2005; Frongillo et al., 2023). This line of work focused on perfect Bayesian learners until Collina et al. (2025, 2026); Kearns et al. (2026) showed that the same kind of accuracy-improving agreement could be obtained via test-time interaction using computationally and data efficient learning algorithms.

**Agreement as Variance** Our disagreement metric is (twice) the variance of the training procedure. Relatedly, Caponnetto and Rakhlin (2006) study stability of empirical risk minimization over Donsker classes, showing that the $L_2(P)$ diameter of the set of near empirical minimizers converges to zero. Kur et al. (2023) show that for realizable learning problems (with mean zero independent noise), empirical risk minimization over a fixed, convex class leads to variance that is upper bounded by the minimax rate. Our results apply to more general settings: our applications to neural networks and regression trees correspond to non-convex learning problems, our application to stacking does not correspond to optimization over a fixed class, and we do not require any realizability assumptions. Note also, the interest of Kur et al. (2023) is to study generalization through the lens of bias/variance tradeoffs, whereas our starting point is to assume small excess risk in distribution.

**Different Notions of Stability** There are many notions of stability in machine learning. Bousquet and Elisseeff (2002) give notions of leave-one-out stability and connect them to out-of-sample generalization. These notions have been influential, and many authors have proven generalization bounds via this link to stability: for example Hardt et al. (2016) show that stochastic gradient descent is stable in this sense if only run for a small number of iterations, and Charles and Papailiopoulos (2018) study the stability of global optimizers in terms of the geometry of the loss-optimal solution.

These notions of stability are different than the disagreement metric we study here. First, stability in the sense of Bousquet and Elisseeff (2002) is stability only of the loss, not the predictions themselves which is our interest. Second, stability in the sense of Bousquet and Elisseeff (2002) is stability with respect to adding or removing a single training example, whereas we want prediction-level stability over fully independent retraining, in which (in general) every single training example is different — just drawn from the same distribution.

Differential privacy (Dwork et al., 2006; Dwork and Roth, 2014) is a strong notion of algorithmic stability that when applied to machine learning requires that when one training sample is changed, the (randomized) training algorithm induces a near-by distribution on output models. Differential privacy is a much stronger stability condition than those of Bousquet and Elisseeff (2002), and similarly implies strong generalization guarantees (Dwork et al., 2015). When the differential privacy stability parameter is taken to be sufficiently small ($\varepsilon \ll 1/\sqrt{n}$), then it implies stability under resampling of the entire training set from the same distribution, as we study in our paper — this is related to what is called *perfect generalization* by Cummings et al. (2016). Via this connection, differential privacy has been shown to be (information theoretically) reducible to replicability (as defined by Impagliazzo et al. (2022)) and vice-versa (Bun et al., 2023). Replicability is a stronger condition than the kind of agreement that we study: in the context of machine learning, it requires that (under coupled random coins across the two training algorithms), the run of two training algorithms over independently sampled training sets output *exactly identical* models with high probability. In contrast we ask that two independently trained models produce *numerically similar* predictions on *most* examples. However, because replicability asks for more, it also comes with severe limitations that we avoid. Via its connection to differential privacy, there are strong separations between problems that are learnable with the constraint of replicability and without (Alon et al., 2019; Bun et al., 2020). Even for those learning problems that are solvable replicably (e.g. learning problems solvable in the statistical query model of Kearns (1998)), *standard* learning algorithms for these problems are not replicable, and the computational and sample complexity of custom-designed replicable algorithms often far exceeds the complexity of non-replicable learning (see e.g. Eaton et al. (2026)). In contrast, our analyses apply to existing, popular, state of the art learning algorithms (gradient boosting and regression tree and neural network training with architecture search). Since *any* model class can be used together with stacking or gradient boosting, there are no barriers to obtaining our kind of model agreement similar to the information theoretic barriers separating replicable from non-replicable learning. The concurrent work of Hopkins et al. (2025) is similarly motivated to ours: their goal is to relax the strict replicability definition of Impagliazzo et al. (2022) to one that requires that two replicably trained models agree on "most inputs", and thereby circumvent the impossibility results separating PAC learning from replicable learning. They give several definitions of approximate replicability and show that approximately replicable PAC learning has similar sample complexity to unconstrained PAC learning. Our approaches, results, and techniques are quite different, however. Hopkins et al. (2025) focuses on binary hypothesis classes and gives custom training procedures relying on shared randomness that satisfy their notion of approximate replicability. We instead focus on (multi-dimensional) regression problems and give analyses of existing, popular learning algorithms. Our training procedures do not use shared randomness.

**Agreement and Ensembling** Wood et al. (2023) studies the error reduction that can be obtained through ensembling methods and relates it to a notion of model disagreement that is related to ours. Their interests are in some sense dual to ours: for them the primary goal is error reduction, and

model disagreement is a means to that end; our primary goal is model agreement, and we show a general recipe for obtaining it — some of our applications (e.g. stacking) are ensembling methods, but others (e.g. neural network training) are not.

**Empirical Phenomena** Empirical work quantifies prediction-level stability across retrainings via churn, per-example consistency, and related notions (Bhojanapalli et al., 2021; Bahri and Jiang, 2021; Johnson and Zhang, 2023). Based on this, various studies show that simple procedures — e.g., ensembling or co-distillation — can increase agreement (Wang et al., 2020; Bhojanapalli et al., 2021). However, recent work showed that fluctuations in run-to-run test accuracy can be largely explained by finite-sample effects even when the underlying predictors are similar (Jordan, 2024). Relatedly, Somepalli et al. (2022) made the observation that across pairs of models, independently trained neural networks often seem to depict similar decision regions despite their complexity which raises the question of when and whether external methods to encourage agreement are even required. On top of this, Mao et al. (2024) provide evidence that training trajectories lie on a shared low-dimensional manifold in prediction space, pointing to a common structure that could underlie agreement. The latter works only characterize the prediction space based on visualizations and do not provide a formal explanation as to why agreement might occur from independent training. Gorecki and Hardt (2025) recently conducted a large empirical study of model disagreement across 50 large language models used for prediction tasks, and find that empirically they have much higher levels of agreement than one would expect if errors were made at random; our work can be viewed as giving foundations to this kind of empirical observation.

Empirical agreement has also been studied through the lens of generalization. In-distribution pairwise disagreement between independently trained copies on unlabeled test data has been observed to provide an accurate estimate of test error (Jiang et al., 2022). Moreover, a single model's pattern of predictions on the training set closely matches its behavior on the test set as distributions, indicating prediction-space stability that is distinct from inter-run agreement (Nakkiran and Bansal, 2020). Beyond in-distribution, there are cases where even out-of-distribution pairwise agreement scales linearly with in-distribution agreement across many shifts (Baek et al., 2022). None of these works provide prediction-space conditions or rates under which independently trained models will immediately agree in the first place.

A complementary line of work focusing on weight-space studies shows that many independently trained solutions can be connected by low-loss paths (Garipov et al., 2018; Draxler et al., 2018). Even when solutions aren't trivially aligned, applying neuron permutations can align them, enabling low-loss interpolation (Entezari et al., 2022; Ainsworth et al., 2023). It can be shown that their layers are stitchable or exhibit layer-wise linear feature connectivity (Bansal et al., 2021; Zhou et al., 2023), which is consistent with a connected region once permutation symmetries are accounted for. These techniques are post-hoc observations about weight or parameter space and do not provide ex ante, prediction-space guarantees or quantitative rates that independent training will agree without alignment.

Closer to prediction space theory, the neural tangent kernel findings characterize how a model's predictive function evolves under gradient descent (Jacot et al., 2018; Lee et al., 2019). However, these analyses focus on a single training trajectory, primarily analyze the infinite-width regime, and do not directly address whether independently trained models will agree. Our work seeks conditions under which standard training *directly* yields approximate agreement "out of the box," bypassing parameter-space alignment and establishing stability in prediction space itself.

## Appendix B. Proofs for Stacking

**Theorem 17 (Agreement for Stacked Aggregation)** *Let* $G = \{g_1, \dots, g_k\} \overset{i.i.d.}{\sim} Q^k$ *and* $G' = \{g'_1, \dots, g'_k\} \overset{i.i.d.}{\sim} Q^k$ *be independent. Define* $f_1, f_2$ *as follows:*

$$f_1 = \arg\min_{f \in V(G)} \mathrm{MSE}(f), \quad f_2 = \arg\min_{f \in V(G')} \mathrm{MSE}(f)$$

*Then we have that*

$$\mathbb{E}_{f_1,f_2}\big[D(f_1, f_2)\big] \;\le\; 4\big(\bar{R}_k - \bar{R}_{2k}\big).$$

**Proof** Fix realizations of $G$ and $G'$, and let $G^* = G \cup G'$ (multiset union). Throughout this section we will think of $G, G' \sim Q^k$, unless explicitly conditioned. Note that $V(G) \subseteq V(G^*)$ and $V(G') \subseteq V(G^*)$. In our proofs, without loss of generality, we will use the notation $h_G$ to denote the least squares minimizer with respect to subspace $G$. In our theorem statements, this corresponds to $f_1$, but we use this notation in our proofs for the sake of clarity. Let $\bar{h} := \frac{1}{2}(h_G + h_{G'})$. Since $h_G \in V(G)$ and $h_{G'} \in V(G')$ and $V(G), V(G') \subseteq V(G^*)$, we have $\bar{h} \in V(G^*)$. Applying Lemma 3 with $f_1 = h_G$, $f_2 = h_{G'}$, and $\mathcal{H} = V(G^*)$, and using $\mathrm{MSE}(h_G) = R(G)$, $\mathrm{MSE}(h_{G'}) = R(G')$, and $R(V(G^*)) = R(G^*)$, we have the pointwise inequality

$$\|h_G - h_{G'}\|^2 \;\le\; 2\big(R(G) - R(G^*)\big) + 2\big(R(G') - R(G^*)\big). \tag{1}$$

We now take expectations over $G, G', G^*$ to relate the two terms on the RHS of Equation 1. Conditional on $G^*$, we can generate the pair $(G, G')$ by drawing a uniformly random permutation $\pi$ of $\{1, \dots, 2k\}$ and letting $G$ be the first $k$ permuted elements of $G^*$ and $G'$ the remaining $k$. This holds because the $2k$ features in $G^*$ arise from $2k$ i.i.d. draws from $Q$ and the joint law of $(G, G')$ is exchangeable under permutations of these $2k$ draws. Conditioning on the unordered multiset $G^*$, $(G, G')$ is a uniformly random partition into two $k$-submultisets. Therefore, taking the conditional expectation of (1) given $G^*$ and using symmetry of $G$ and $G'$,

$$\mathbb{E}_{(G,G')|G^*}\big[\,\|h_G - h_{G'}\|^2 \,\big|\, G^*\big] \;\le\; 4\Big(\mathbb{E}_{(G,G')|G^*}\big[R(G) \,\big|\, G^*\big] - R(G^*)\Big). \tag{2}$$

We now integrate (2) over $G^*$. We claim that

$$\mathbb{E}_{G^*}\Big[\mathbb{E}_{(G,G')|G^*}\big[R(G) \,\big|\, G^*\big]\Big] \;=\; \bar{R}_k \qquad \text{and} \qquad \mathbb{E}_{G^*}\big[R(G^*)\big] \;=\; \bar{R}_{2k}. \tag{3}$$

The second equality is immediate from the definition of $\bar{R}_{2k}$, since $G^*$ is a collection of $2k$ i.i.d. draws from $Q$. For the first equality in (3), let $U$ be a uniformly random $k$-subset of $\{1, \dots, 2k\}$ independent of the draws $\{g_1, \dots, g_{2k}\} \overset{\text{i.i.d.}}{\sim} Q^{2k}$. Define $G_U := \{g_i\}_{i \in U}$. By the conditional description above,

$$\mathbb{E}_{(G,G')|G^*}\big[R(G) \,\big|\, G^*\big] \;=\; \mathbb{E}_U\big[R(G_U) \,\big|\, G^*\big].$$

$$\mathbb{E}_{G^*}\Big[\mathbb{E}_{(G,G')|G^*}\big[R(G) \,\big|\, G^*\big]\Big] = \mathbb{E}_{G^*}\Big[\mathbb{E}_U\big[R(G_U) \,\big|\, G^*\big]\Big] = \mathbb{E}_{G^*,U}\big[R(G_U)\big].$$

For any fixed $U$, the subcollection $\{g_i\}_{i \in U}$ consists of $k$ i.i.d. draws from $Q$ (since the full family is i.i.d. and $U$ is independent of the draws), hence averaging over $U$ yields $\mathbb{E}_{G^*,U}[R(G_U)] = \bar{R}_k$, proving (3).

Finally, taking expectations in (2) and substituting (3) gives

$$\mathbb{E}_{G,G'}\big[\,\|h_G - h_{G'}\|^2\,\big] \;\le\; 4\big(\bar{R}_k - \bar{R}_{2k}\big),$$

which is the desired bound. ■

**Theorem 18 (Near-tightness of the factor 4)** *Fix an integer $k \ge 1$. For every $\varepsilon > 0$, there exists a data distribution $P$ (equivalently, an $L^2(P)$ Hilbert space model) and a distribution $Q$ over base models such that if $G, G' \overset{i.i.d.}{\sim} Q^k$ are independent $k$–tuples and*

$$f_1 = \arg\min_{f \in V(G)} \mathrm{MSE}(f), \qquad f_2 = \arg\min_{f \in V(G')} \mathrm{MSE}(f),$$

*then*

$$\mathbb{E}_{f_1,f_2}\big[D(f_1,f_2)\big] \;\ge\; (4-\varepsilon)\,\big(\bar{R}_k - \bar{R}_{2k}\big).$$

**Proof** Fix $k \ge 1$ and $\varepsilon > 0$. Since the claim is weaker for larger $\varepsilon$, we may assume $\varepsilon \in (0,1]$. We work in a real Hilbert space $\mathcal{H}$ (equivalently $\mathcal{H} = L^2(P)$ for a suitable data distribution $P$[1]) with an orthonormal family $\{e_0, \dots, e_m\}$, where $m \in \mathbb{N}$ will be chosen later, and set the target $y := e_0$. We construct base models that are "noisy versions" of the target. Fix $\sigma > 0$ and define

$$g_i \;:=\; e_0 + \sigma e_i, \qquad i = 1, \dots, m.$$

Let $Q$ be the uniform distribution over $\{g_1, \dots, g_m\}$.

First, we analyze the predictor and risk for a fixed set of distinct base models. Let $H$ be a multiset of draws from $Q$. Let $S(H)$ be the set of distinct indices of base models in $H$, and let $r(H) = |S(H)|$. By symmetry, the least-squares predictor $f_H \in V(H)$ assigns equal weight to each distinct $g_i \in H$. A straightforward calculation shows that the optimal weights are $1/(r(H)+\sigma^2)$, yielding:

$$f_H \;=\; \sum_{i \in S(H)} \frac{1}{r(H)+\sigma^2}\, g_i \;=\; \frac{r(H)}{r(H)+\sigma^2}\, e_0 \;+\; \frac{\sigma}{r(H)+\sigma^2} \sum_{i \in S(H)} e_i. \tag{4}$$

$$R(H) \;=\; \|y - f_H\|^2 \;=\; \frac{\sigma^2}{r(H)+\sigma^2}. \tag{5}$$

In particular, for $G, G' \overset{\text{i.i.d.}}{\sim} Q^k$, we have $f_1 = f_G$, $f_2 = f_{G'}$, and $R(G) = \mathrm{MSE}(f_1)$, $R(G') = \mathrm{MSE}(f_2)$.

Next, we analyze the disagreement and risk drop on the event where all sampled models are distinct. Let $E$ be the event that the $2k$ draws in $G \cup G'$ are all distinct. On this event, $r(G) = k$, $r(G') = k$, and $r(G \cup G') = 2k$. Using (5), the drop in risk on event $E$ is:

$$\Delta_0 \;:=\; R(G) - R(G \cup G') \;=\; \frac{\sigma^2}{k+\sigma^2} - \frac{\sigma^2}{2k+\sigma^2}. \tag{6}$$

1. For example, take $\mathcal{X} = \{0, 1, \dots, m\}$ and let $P$ be uniform on $\mathcal{X}$. Defining $e_j(x) = \sqrt{m+1}\,\mathbb{I}\{x = j\}$ gives an orthonormal family $\{e_0, \dots, e_m\} \subseteq L^2(P)$.

Using (4) and the fact that $G$ and $G'$ share no indices on $E$ (and thus the $e_0$ coefficients are identical and cancel), the disagreement is:

$$D_0 \;:=\; \|f_G - f_{G'}\|^2 \;=\; \left\| \frac{\sigma}{k+\sigma^2} \left( \sum_{i \in S(G)} e_i - \sum_{j \in S(G')} e_j \right) \right\|^2 \;=\; \frac{2k\sigma^2}{(k+\sigma^2)^2}. \tag{7}$$

Comparing these quantities, we see that for small $\sigma$:

$$\frac{D_0}{\Delta_0} \;=\; 4 - \frac{2\sigma^2}{k+\sigma^2} \;\xrightarrow{\sigma \to 0}\; 4. \tag{8}$$

Finally, we handle the expectations by showing that the event $E$ dominates. The probability of a collision among the $2k$ uniform draws from $m$ items is at most $\binom{2k}{2}/m$, and hence

$$\mathbf{Pr}(E) \;\geq\; 1 - \binom{2k}{2} \frac{1}{m}.$$

Since disagreement is always non-negative:

$$\mathbb{E}_{G,G'}\big[D(f_1, f_2)\big] \;\geq\; \mathbf{Pr}(E)\, D_0. \tag{9}$$

For the expected risk drop, we upper bound the risk when collisions occur. The risk $R(H)$ is maximized when $r(H)$ is minimized (i.e., $r(H) = 1$), bounded by $R_{max} = \sigma^2/(1+\sigma^2)$. The expected risk is:

$$\begin{aligned} \bar{R}_k &= \mathbf{Pr}[r(G) = k] \frac{\sigma^2}{k+\sigma^2} + \mathbb{E}\left[R(G) \mathbb{I}(r(G) < k)\right] \\ &\leq \frac{\sigma^2}{k+\sigma^2} + \mathbf{Pr}\big(r(G) < k\big)\, R_{max} \;\leq\; \frac{\sigma^2}{k+\sigma^2} + \binom{k}{2} \frac{1}{m}\, R_{max}. \end{aligned}$$

On the other hand, since $r(G \cup G') \leq 2k$ always, we have the deterministic lower bound $\bar{R}_{2k} \geq \frac{\sigma^2}{2k+\sigma^2}$. Combining these, the expected drop satisfies:

$$\bar{R}_k - \bar{R}_{2k} \leq \Delta_0 + \frac{k^2}{2m} \frac{\sigma^2}{1+\sigma^2}.$$

Now choose $\sigma^2 = (\varepsilon/8)k$ so that (8) gives $D_0 \geq (4 - \varepsilon/4)\Delta_0$. Choosing $m \geq \left\lceil \frac{96\, k^3}{\varepsilon} \right\rceil$ makes $\mathbf{Pr}(E)$ close to 1 and the collision term in the bound on $\bar{R}_k - \bar{R}_{2k}$ negligible compared to $\Delta_0$. Combining (9) with the upper bound on $\bar{R}_k - \bar{R}_{2k}$ then yields $\mathbb{E}_{f_1,f_2}\big[D(f_1, f_2)\big] \geq (4-\varepsilon)(\bar{R}_k - \bar{R}_{2k})$. ■

## Appendix C. Proofs for Gradient Boosting

**Lemma 19 (Correlation Lower Bound w.r.t. Weak Learning Anchor Gap)** *For any $f$, writing $M(f) := \sup_{g \in \mathcal{C}} |\mathbb{E}[\langle y - f, g \rangle]|$, we have*

$$M(f) \;\geq\; \frac{\mathrm{MSE}(f) - R\big(V(\mathcal{C})\big)}{2\,\tau^*}.$$

**Proof** Recall that $\mathcal{K}_{\tau^*} := \tau^* \operatorname{conv}(\mathcal{C})$. Its support function is $\sigma_{\mathcal{K}_{\tau^*}}(u) := \sup_{s \in \mathcal{K}_{\tau^*}} \mathbb{E}[\langle u, s\rangle] = \tau^* \sup_{g \in \pm\mathcal{C}} \mathbb{E}[\langle u, g\rangle]$. We will ultimately relate this quantity to $M(f)$. For any $s \in \mathcal{K}_{\tau^*}$, the squared loss obeys

$$\mathrm{MSE}(f) - \mathrm{MSE}(s) \;=\; \|y-f\|^2 - \|y-s\|^2 \;=\; 2\mathbb{E}[\langle y-f, s-f\rangle] - \|s-f\|^2 \;\le\; 2\mathbb{E}[\big(\langle y-f, s\rangle - \langle y-f, f\rangle\big)].$$

The second equality uses the fact that $||a||^2 - ||b||^2 = 2\langle a, a-b\rangle - ||a-b||^2$. The inequality uses the fact that we can drop the subtracted nonnegative term $||s-f||^2$. Taking the supremum over $s \in \mathcal{K}_{\tau^*}$ yields

$$\mathrm{MSE}(f) - R(\mathcal{K}_{\tau^*}) \;\le\; 2\,\sigma_{\mathcal{K}_{\tau^*}}(y-f) \;-\; 2\mathbb{E}[\langle y - f(x), f(x)\rangle].$$

Applying the same inequality with $f - y$ in place of $y - f$ yields a second upper bound. Since any $X$ with $X \le A$ and $X \le B$ satisfies $X \le (A+B)/2$, averaging the two bounds cancels the unknown linear term $\langle y - f, f\rangle$. Using evenness of the support function for symmetric sets, $\sigma_{\mathcal{K}_{\tau^*}}(u) = \sigma_{\mathcal{K}_{\tau^*}}(-u)$, we get

$$\begin{aligned}
\mathrm{MSE}(f) - R(\mathcal{K}_{\tau^*}) &\le \sigma_{\mathcal{K}_{\tau^*}}(y-f) \;+\; \sigma_{\mathcal{K}_{\tau^*}}(f-y) \\
&= 2\,\sigma_{\mathcal{K}_{\tau^*}}(y-f) \\
&= 2\,\tau^* \sup_{g \in \pm\mathcal{C}} \mathbb{E}[\langle y - f(x), g(x)\rangle] \\
&= 2\,\tau^* \sup_{g \in \mathcal{C}} |\langle \mathbb{E}[y - f(x), g(x)]\rangle| \\
&= 2\,\tau^*\, M(f).
\end{aligned}$$

where we used symmetry of $\mathcal{K}_{\tau^*}$ and of $\mathcal{C}$. For any $u$, the trivial inequality is $\sup_{g\in\mathcal{C}} |\langle u, g\rangle| \ge \sup_{g\in\mathcal{C}} \langle u, g\rangle$. Conversely, because $\mathcal{C}$ is symmetric, for every $g \in \mathcal{C}$ also $-g \in \mathcal{C}$, so $\max\{\langle u, g\rangle, \langle u, -g\rangle\} = |\langle u, g\rangle|$, implying $\sup_{g\in\mathcal{C}} \langle u, g\rangle \ge \sup_{g\in\mathcal{C}} |\langle u, g\rangle|$. Thus $\sup_{g\in\mathcal{C}} |\langle u, g\rangle| = \sup_{g\in\mathcal{C}} \langle u, g\rangle$. Taking $u = y - f$ identifies the last term with $M(f)$. Since $f^* \in \mathcal{K}_{\tau^*} \cap V(\mathcal{C})$ minimizes MSE over $V(\mathcal{C})$, we have $R(\mathcal{K}_{\tau^*}) = R\big(V(\mathcal{C})\big)$. Rearranging yields

$$M(f) \;\ge\; \frac{\mathrm{MSE}(f) - R\big(V(\mathcal{C})\big)}{2\,\tau^*}.$$

**Theorem 20 (Weak Learning Anchor Gap Upper Bound)** *For all $t \ge 1$,*

$$\mathrm{MSE}(f_t) - R\big(V(\mathcal{C})\big) \;\le\; \frac{8\,(\tau^*)^2}{t} \;+\; \sum_{s=1}^{t} \varepsilon_s^2.$$

**Proof** Let $E_t := \mathrm{MSE}(f_t) - R\big(V(\mathcal{C})\big)$. From Proposition 9,

$$E_{t-1} - E_t \;\ge\; \Big(\tfrac{E_{t-1}}{2\tau^*} - \varepsilon_t\Big)_+^2.$$

For any $a \ge 0$ and $b \in \mathbb{R}$, $(a-b)^2 \ge a^2/2 - b^2$. To see this, consider: $a^2 - 2ab + b^2 - a^2/2 + b^2$. We have that this quantity equals $a^2/2 - 2ab + 2b^2$. Since a multiplicative factor of 2 does not affect

the sign, notice that twice this quantity is equal to $(a-2b)^2$ which is non-negative. In this case the inequality also holds for the quantity $((a-b)_+)^2$. Taking $a = E_{t-1}/(2\tau^*)$ and $b = \varepsilon_t$ yields

$$E_{t-1} - E_t \;\geq\; \frac{E_{t-1}^2}{8\,(\tau^*)^2} \;-\; \varepsilon_t^2.$$

Since $E_t \leq E_{t-1}$,

$$\frac{1}{E_t} - \frac{1}{E_{t-1}} \;=\; \frac{E_{t-1}-E_t}{E_t E_{t-1}} \;\geq\; \frac{E_{t-1}-E_t}{E_{t-1}^2} \;\geq\; \frac{1}{8\,(\tau^*)^2} \;-\; \frac{\varepsilon_t^2}{E_{t-1}^2} \;\geq\; \frac{1}{8\,(\tau^*)^2} \;-\; \frac{\varepsilon_t^2}{E_t^2}.$$

Summing from $s=1$ to $t$ gives

$$\frac{1}{E_t} \;\geq\; \frac{1}{E_0} \;+\; \frac{t}{8\,(\tau^*)^2} \;-\; \sum_{s=1}^{t} \frac{\varepsilon_s^2}{E_s^2} \geq \frac{1}{E_0} \;+\; \frac{t}{8\,(\tau^*)^2} \;-\; \frac{1}{E_t^2}\sum_{s=1}^{t} \varepsilon_s^2.$$

Let $A_t := \sum_{s=1}^{t} \varepsilon_s^2$ and $B_t := \frac{1}{E_0} + \frac{t}{8\,(\tau^*)^2}$. Writing $X := 1/E_t$, the inequality becomes $A_t X^2 + X - B_t \geq 0$. If $A_t = 0$ then $X \geq B_t$ and $E_t \leq 1/B_t \leq 8(\tau^*)^2/t$. If $A_t > 0$, define the quantity $Y = 1/X$. Then, the inequality becomes $-B_t Y^2 + Y + A_t \geq 0$. Then the quadratic inequality implies $Y \leq \frac{1+\sqrt{1+4A_tB_t}}{2B_t}$. Using $\sqrt{1+z} \leq 1 + z/2$ for $z \geq 0$ gives

$$\frac{1}{X} \;\leq\; \frac{1}{B_t} \;+\; A_t.$$

Thus $E_t \leq 1/B_t + A_t \leq 8(\tau^*)^2/t + \sum_{s=1}^{t} \varepsilon_s^2$. ■

■

## Appendix D. Proofs for Neural Networks/Regression Trees

**Lemma 21 (Regression-tree midpoint closure)** *For every $d \geq 0$ and every $f_1, f_2 \in \mathsf{Tree}_d$, the midpoint predictor $\bar{f} := \frac{1}{2}(f_1 + f_2)$ lies in $\mathsf{Tree}_{2d}$.*

**Proof** Fix realizations of $f_1, f_2 \in \mathsf{Tree}_d$ as depth-$d$ trees. Consider the partition of $\mathcal{X}$ induced by the leaves of the tree for $f_1$; on each cell of this partition, $f_1$ is constant. Now refine each such cell further using the splits of the tree for $f_2$ restricted to that cell.

Equivalently, we can construct a single tree as follows: take the tree for $f_1$, and at each leaf, graft a copy of the tree for $f_2$. Along any root-to-leaf path, we traverse at most $d$ splits from $f_1$ and then at most $d$ splits from $f_2$, so the resulting tree has depth at most $2d$. Moreover, on each leaf of the resulting tree, both $f_1$ and $f_2$ take constant values, so we can label that leaf with their average $\frac{1}{2}f_1(x) + \frac{1}{2}f_2(x) \in [0,1]$. This yields a depth-$2d$ regression tree computing $\bar{f}$, i.e., $\bar{f} \in \mathsf{Tree}_{2d}$. ■

## Appendix E. Generalization to Multi-Dimensional Strongly Convex Losses

In this section we generalize our setting to study models that output $d$-dimensional distributions as predictions, and optimize arbitrary strongly convex losses. We show that the midpoint anchoring argument extends directly to this more general setting, which lets us model a wide array of practical machine learning problems. First we define general strongly convex loss functions over $d$ dimensional predictions:

**Definition 22 (Strongly convex losses)** *Let $\mathcal{L} : \mathcal{Y} \times \mathbb{R}^d \to \mathbb{R}$ be a continuously differentiable loss function. We say that $\mathcal{L}$ is* $\mu$-strongly convex *if there exists some $\mu > 0$ such that for every $y \in \mathcal{Y}$, $P_1, P_2 \in \mathbb{R}^d$,*

$$\mathcal{L}(y, P_1) \geq \mathcal{L}(y, P_2) + \langle \nabla_p \mathcal{L}(y, P_2), P_1 - P_2 \rangle + \tfrac{\mu}{2}\|P_1 - P_2\|_2^2.$$

For predictors outputting $d$-dimensional predictions, we define disagreement as follows, straightforwardly generalizing our 1-dimensional expected squared disagreement metric:

**Definition 23 (Generalized disagreement)** *Let $P$ be a distribution on $\mathcal{X} \times \mathcal{Y}$ and let $f_1, f_2 : \mathcal{X} \to \mathbb{R}^d$ be functions. The disagreement between $f_1, f_2$ over $P$ is the expected squared Euclidean distance between their predictions:*

$$D(f_1, f_2) = \mathbb{E}[\|f_1(x) - f_2(x)\|_2^2].$$

We will write $R(f) := \mathbb{E}[\mathcal{L}(y, f(x))]$. We can now generalize our disagreement-via-midpoint-anchoring lemma which drives our analyses.

**Lemma 24 (Disagreement via the midpoint anchor)** *Assume $\mathcal{L}$ is $\mu$-strongly convex. For any two functions $f_1, f_2 : \mathcal{X} \to \mathbb{R}^d$, let $\bar{f}(x) := \frac{1}{2}(f_1(x) + f_2(x))$. Then*

$$D(f_1, f_2) \;\leq\; \tfrac{4}{\mu}\Big(R(f_1) + R(f_2) - 2R(\bar{f})\Big).$$

*In particular, if $\bar{f} \in \mathcal{H}$ for some class of predictors $\mathcal{H}$, then*

$$D(f_1, f_2) \;\leq\; \tfrac{4}{\mu}\big(R(f_1) - R(\mathcal{H})\big) + \tfrac{4}{\mu}\big(R(f_2) - R(\mathcal{H})\big).$$

**Proof** Fix any $x \in \mathcal{X}$ and $y \in \mathcal{Y}$ and abbreviate

$$p_1 := f_1(x), \quad p_2 := f_2(x), \quad \bar{p} := \bar{f}(x) = \tfrac{1}{2}(p_1 + p_2).$$

Applying $\mu$-strong convexity (Definition 22) with $P_1 = p_1$ and $P_2 = \bar{p}$ gives

$$\mathcal{L}(y, p_1) \geq \mathcal{L}(y, \bar{p}) + \langle \nabla_p \mathcal{L}(y, \bar{p}),\, p_1 - \bar{p} \rangle + \tfrac{\mu}{2}\|p_1 - \bar{p}\|_2^2.$$

Similarly, with $P_1 = p_2$ and $P_2 = \bar{p}$,

$$\mathcal{L}(y, p_2) \geq \mathcal{L}(y, \bar{p}) + \langle \nabla_p \mathcal{L}(y, \bar{p}),\, p_2 - \bar{p} \rangle + \tfrac{\mu}{2}\|p_2 - \bar{p}\|_2^2.$$

Adding the two inequalities, and using $(p_1 - \bar{p}) + (p_2 - \bar{p}) = p_1 + p_2 - 2\bar{p} = 0$ to cancel the gradient terms, yields

$$\mathcal{L}(y, p_1) + \mathcal{L}(y, p_2) \geq 2\mathcal{L}(y, \bar{p}) + \tfrac{\mu}{2}\Big(\|p_1 - \bar{p}\|_2^2 + \|p_2 - \bar{p}\|_2^2\Big).$$

Since $p_1 - \bar{p} = \frac{1}{2}(p_1 - p_2)$ and $p_2 - \bar{p} = \frac{1}{2}(p_2 - p_1)$, we have

$$\|p_1 - \bar{p}\|_2^2 + \|p_2 - \bar{p}\|_2^2 = 2\left\|\tfrac{1}{2}(p_1 - p_2)\right\|_2^2 = \tfrac{1}{2}\|p_1 - p_2\|_2^2.$$

Substituting this back and rearranging gives the pointwise bound

$$\|f_1(x) - f_2(x)\|_2^2 \le \tfrac{4}{\mu}\Big(\mathcal{L}(y, f_1(x)) + \mathcal{L}(y, f_2(x)) - 2\mathcal{L}\big(y, \bar{f}(x)\big)\Big).$$

Taking expectations over $(x, y) \sim P$ and using the definitions of $D(\cdot,\cdot)$ and $R(\cdot)$ yields

$$D(f_1, f_2) \le \tfrac{4}{\mu}\Big(R(f_1) + R(f_2) - 2R(\bar{f})\Big).$$

For the second inequality, if $\bar{f} \in \mathcal{H}$ then $R(\bar{f}) \ge R(\mathcal{H})$, so substituting $R(\bar{f})$ by $R(\mathcal{H})$ in the right-hand side yields the claim. ■

We now show how to apply the midpoint anchoring lemma to each of our (generalized) applications.

### E.1. Stacking

Here, we will provide a generalization of our stacking results to multi-dimensional strongly-convex losses. We once again model "base models" as being sampled i.i.d. from an arbitrary distribution $Q$, and under two independent training runs write $G, G' \sim Q^k$ to denote the set of $k$ sampled models. We will consider the stacked predictors $f_1 \in V(G)$ and $f_2 \in V(G')$. Define $G^* = G \cup G'$. The key observation is that the midpoint predictor $\frac{1}{2}(f_1 + f_2)$ lies in $V(G^*)$, so we can apply Lemma 24 and then use the same exchangeability argument as in the single-dimensional case.

**Theorem 25** *(Agreement for Stacked Aggregation Generalization) Assume that $\mathcal{L}$ is $\mu$-strongly convex. Let $G = \{g_1, \dots, g_k\} \overset{i.i.d.}{\sim} Q^k$ and $G' = \{g'_1, \dots, g'_k\} \overset{i.i.d.}{\sim} Q^k$ be independent. Define $f_1, f_2$ as follows:*

$$f_1 = \arg\min_{f \in V(G)} \mathbb{E}[\mathcal{L}(y, f(x))], \quad f_2 = \arg\min_{f \in V(G')} \mathbb{E}[\mathcal{L}(y, f(x))]$$

*Then we have that*

$$\mathbb{E}_{f_1, f_2}\big[D(f_1, f_2)\big] \;\le\; \frac{8}{\mu}\big(\bar{R}_k - \bar{R}_{2k}\big).$$

**Proof** Fix realizations of $G$ and $G'$, and let $G^* = G \cup G'$ (multiset union). Throughout this section we will think of $G, G' \sim Q^k$, unless explicitly conditioned. Note that $V(G) \subseteq V(G^*)$ and $V(G') \subseteq V(G^*)$. In our proofs, without loss of generality, we will use the notation $h_G$ to denote the minimizer of $\mathbb{E}[\mathcal{L}(y, \cdot)]$ with respect to subspace $G$. Similarly, we will use the notation $R(G) = R(h_G)$ in this context. In our theorem statements, this corresponds to $f_1$. Let $\bar{h} := \frac{1}{2}(h_G + h_{G'})$. Since $h_G \in V(G)$ and $h_{G'} \in V(G')$ and $V(G), V(G') \subseteq V(G^*)$, we have $\bar{h} \in V(G^*)$. Applying Lemma 24 with $f_1 = h_G$, $f_2 = h_{G'}$, and $\mathcal{H} = V(G^*)$, and using $R(h_G) = R(G)$, $R(h_{G'}) = R(G')$, and $R(V(G^*)) = R(G^*)$, we have the pointwise inequality

$$\|h_G - h_{G'}\|^2 \;\le\; \frac{4}{\mu}\big(R(G) - R(G^*)\big) + \frac{4}{\mu}\big(R(G') - R(G^*)\big). \tag{10}$$

We now take expectations over $G, G', G^*$ to relate the two terms on the RHS of Equation 10. Conditional on $G^*$, we can generate the pair $(G, G')$ by drawing a uniformly random permutation $\pi$ of $\{1, \ldots, 2k\}$ and letting $G$ be the first $k$ permuted elements of $G^*$ and $G'$ the remaining $k$. This holds because the $2k$ features in $G^*$ arise from $2k$ i.i.d. draws from $Q$ and the joint law of $(G, G')$ is exchangeable under permutations of these $2k$ draws. Conditioning on the unordered multiset $G^*$, $(G, G')$ is a uniformly random partition into two $k$-submultisets. Therefore, taking the conditional expectation of (10) given $G^*$ and using symmetry of $G$ and $G'$,

$$\mathbb{E}_{(G,G')|G^*}\left[\, \|h_G - h_{G'}\|^2 \,\middle|\, G^* \right] \;\le\; \frac{8}{\mu}\Big(\mathbb{E}_{(G,G')|G^*}\left[R(G) \,\middle|\, G^*\right] - R(G^*)\Big). \tag{11}$$

We now integrate (11) over $G^*$. We claim that

$$\mathbb{E}_{G^*}\Big[\mathbb{E}_{(G,G')|G^*}\left[R(G) \,\middle|\, G^*\right]\Big] \;=\; \bar{R}_k \qquad \text{and} \qquad \mathbb{E}_{G^*}\left[R(G^*)\right] \;=\; \bar{R}_{2k}. \tag{12}$$

The second equality is immediate from the definition of $\bar{R}_{2k}$, since $G^*$ is a collection of $2k$ i.i.d. draws from $Q$. For the first equality in (12), let $U$ be a uniformly random $k$-subset of $\{1, \ldots, 2k\}$ independent of the draws $\{g_1, \ldots, g_{2k}\} \overset{\text{i.i.d.}}{\sim} Q^{2k}$. Define $G_U := \{g_i\}_{i \in U}$. By the conditional description above,

$$\mathbb{E}_{(G,G')|G^*}\left[R(G) \,\middle|\, G^*\right] \;=\; \mathbb{E}_U\left[R(G_U) \,\middle|\, G^*\right].$$

$$\mathbb{E}_{G^*}\Big[\mathbb{E}_{(G,G')|G^*}\left[R(G) \,\middle|\, G^*\right]\Big] = \mathbb{E}_{G^*}\Big[\mathbb{E}_U\left[R(G_U) \,\middle|\, G^*\right]\Big] = \mathbb{E}_{G^*,U}\left[R(G_U)\right].$$

For any fixed $U$, the subcollection $\{g_i\}_{i \in U}$ consists of $k$ i.i.d. draws from $Q$ (since the full family is i.i.d. and $U$ is independent of the draws), hence averaging over $U$ yields $\mathbb{E}_{G^*,U}[R(G_U)] = \bar{R}_k$, proving (12).

Finally, taking expectations in (11) and substituting (12) gives

$$\mathbb{E}_{G,G'}\left[\, \|h_G - h_{G'}\|^2 \,\right] \;\le\; \frac{8}{\mu}\big(\bar{R}_k - \bar{R}_{2k}\big),$$

which is the desired bound. ■

As before, we have related the stability of (now generalized) stacking to the local learning curve, which is bounded, non-negative, and non-increasing in $k$. As a result for any desired level of stability $\alpha$, there must be a $k \le 2^{O(1/\alpha)}$ that guarantees that level of stability, independently of the complexity of the learning instance — and once again the local learning curve can be empirically investigated on a holdout set to choose such a value of $k$.

### E.2. Gradient Boosting (via Frank Wolfe)

In this section, we generalize our gradient boosting agreement results to the multi-dimensional setting. Along the way we give another generalization as well. Recall that in the final risk bound of Theorem 11, and correspondingly in the final agreement bound, we had a dependence on the instance-dependent constant $\tau^*$, the atomic norm of the best model in the span of the weak learner class. In this section, we instead analyze a Frank-Wolfe variant of gradient boosting. In this variant, the iterates are constrained to lie within a user-specified atomic-norm budget $\tau$. As a result we are able to carry out our anchoring argument with respect to the best norm $\tau$ model in the span of the

**Algorithm 4** Multi-Dimensional Frank–Wolfe

**Input:** SQ-oracle for weak learner class $\mathcal{C}$, budget $\tau > 0$
$f_0 \equiv 0$, $G_0 = \emptyset$ **for** $t \in [k]$ **do**
  Choose $s_t \in \mathcal{C}$ such that $\mathbb{E}[\langle -\nabla_p \mathcal{L}(y, f_{t-1}(x)), s_t(x)\rangle] \geq \max_{s \in \mathcal{C}} \mathbb{E}[\langle -\nabla_p \mathcal{L}(y, f_{t-1}(x)), s(x)\rangle] - \varepsilon_t$
  Choose $g_t \in \mathcal{K}_\tau$ such that $g_t = \tau s_t / \|s_t\|_{\mathcal{A}}$
  $\alpha_t = \frac{2}{t+1}$
  $f_t := f_{t-1} + \alpha_t(g_t - f_{t-1})$; $G_t := G_{t-1} \cup \{g_t\}$
**end**
**return** $f_k$ and $G := G_k$

weak learner class, rather than the best unconstrained model. This lets us replace the dependence on $\tau^*$ with a dependence on $\tau$, which is specified by the user rather than defined by the instance. In this section we will need to work with $L-$smooth losses (in the prediction $p$). In other words we need to assume that for all $y \in \mathcal{Y}$ and all $p_1, p_2 \in \Delta(\mathcal{Y})$ that our loss satisfies:

$$||\nabla_p \mathcal{L}(y, p_1) - \nabla_p \mathcal{L}(y, p_2)||_2 \leq L||p_1 - p_2||_2.$$

Recall the conditions of the weak learner class $\mathcal{C}$ that we had previously, which we continue to assume in this section: symmetry, normalization, and non-degeneracy. Note in this section, for the sake of clarity, we will use the standard inner product $\langle f, g\rangle = f^T g$. When needed, we will explicitly mention the expectations we are computing. When the norms are marked $||f||$, we still take it to mean the same definition as in Preliminaries of $(\mathbb{E}[f(x)^2])^{1/2}$.

We will define the quantity

$$M(f) := \sup_{g \in \mathcal{C}} \left|\mathbb{E}[\langle \nabla_p \mathcal{L}(y, f(x)), g(x)\rangle]\right|.$$

We will also define the closely related quantity

$$G(f) := \sup_{z \in \mathcal{K}_\tau} \mathbb{E}[\langle \nabla_p \mathcal{L}(y, f(x)), f(x) - z(x)\rangle].$$

We can show that for any $f \in \mathcal{K}_\tau$, $G(f) \leq 2\tau M(f)$. Define $\tilde{f}(x), \tilde{g}(x) \in \mathrm{conv}(\mathcal{C})$, where $f(x) = \tau\tilde{f}(x)$ and $g(x) = \tau\tilde{g}(x)$. One can see this (as shown below) because the weak learner class is normalized, functions in $\mathcal{K}_\tau$ can be scaled up to live in $\mathrm{conv}(\mathcal{C})$, the inside inner product term for $M(f)$ is linear in $g$ and therefore the supremum over $\mathrm{conv}(\mathcal{C})$ matches the supremum over $\mathcal{C}$, and triangle inequality.

$$\begin{aligned}
G(f) &= \sup_{z \in \mathcal{K}_\tau} |\mathbb{E}[\langle \nabla \mathcal{L}(y, f(x)), f(x) - z(x) \rangle|] \\
&= \tau \sup_{\tilde{z} \in \text{conv}(\mathcal{C})} |\mathbb{E}[\langle \nabla \mathcal{L}(y, f(x)), \tilde{f}(x) - \tilde{z}(x) \rangle|] \\
&\leq \tau |\mathbb{E}[\langle \nabla \mathcal{L}(y, f(x)), \tilde{f}(x) \rangle]| + \tau \sup_{\tilde{z} \in \text{conv}(\mathcal{C})} |\mathbb{E}[\langle \nabla \mathcal{L}(y, f(x)), \tilde{z}(x) \rangle|] \\
&\leq \tau \sup_{\tilde{h} \in \text{conv}(\mathcal{C})} |\mathbb{E}[\langle \nabla \mathcal{L}(y, f(x)), \tilde{h}(x) \rangle]| + \tau \sup_{\tilde{z} \in \text{conv}(\mathcal{C})} |\mathbb{E}[\langle \nabla \mathcal{L}(y, f(x)), \tilde{z}(x) \rangle|] \\
&= 2\tau \sup_{\tilde{h} \in \mathcal{C}} |\mathbb{E}[\langle \nabla \mathcal{L}(y, f(x)), \tilde{h}(x) \rangle]| \\
&= 2\tau M(f)
\end{aligned}$$

Broadly, our proof will mirror the analysis of our single-dimensional agreement results for gradient boosting. We will once again make use of the conditions on the weak learner class mentioned for gradient boosting of symmetry, normalization, and non-degeneracy. Also note that we can define $G(f)$ with the absolute value due to symmetry of our class, similar to the argument provided in the gradient boosting section. First, we will lower bound the difference of two iterate's losses. This will give us a lower bound on the progress our algorithm's model is making on a per-iterate basis.

**Lemma 26 (FW single-iterate progress)** *Assume $\mathcal{L}$ is $L$-smooth in the second argument. Let $d_t = g_t - f_{t-1}$ with $\|d_t\|_2 \leq 2\tau$. Then with the oracle above we get that,*

$$R(f_{t-1}) - R(f_t) \geq \alpha_t (G(f_{t-1}) - \tau \varepsilon_t) - 2L\tau^2 \alpha_t^2$$

**Proof** By $L-$smoothness we have the following quadratic upper bound (or descent lemma),

$$\mathcal{L}(y, f_{t-1}(x) + \alpha d_t(x)) \leq \mathcal{L}(y, f_{t-1}(x)) + \alpha \langle \nabla_p \mathcal{L}(y, f_{t-1}(x)), d_t(x)) \rangle + \frac{L}{2} \alpha^2 (d_t(x))^2.$$

Taking expectations, we know that

$$R(f_{t-1}) - R(f_t) \geq \alpha \mathbb{E}[\langle -\nabla_p \mathcal{L}(y, f_{t-1}(x)), d_t \rangle] - \frac{L}{2} \alpha^2 ||d_t||_2^2.$$

Consider the quantity $\langle -\nabla_p \mathcal{L}(y, f_{t-1}), d_t \rangle = \langle -\nabla_p \mathcal{L}(y, f_{t-1}), g_t - f_{t-1} \rangle = \langle -\nabla_p \mathcal{L}(y, f_{t-1}), g_t \rangle + \langle \nabla_p \mathcal{L}(y, f_{t-1}), f_{t-1} \rangle$. We know from the oracle that $\langle -\nabla_p \mathcal{L}(y, f_{t-1}), g_t \rangle \geq \tau \sup_{c \in \mathcal{C}} \langle -\nabla_p \mathcal{L}(y, f_{t-1}), c \rangle - \tau \varepsilon_t = \sup_{g \in \mathcal{K}_\tau} \langle -\nabla_p \mathcal{L}(y, f_{t-1}), g \rangle - \tau \varepsilon_t$. We can combine this back with the term $\langle \nabla_p \mathcal{L}(y, f_{t-1}), f_{t-1} \rangle$ and reapply the expectation to lower bound this term by $G(f_{t-1}) - \tau \varepsilon_t$. Therefore, we know that

$$R(f_{t-1}) - R(f_t) \geq \alpha_t (G(f_{t-1}) - \tau \varepsilon_t) - \frac{L}{2} \alpha_t^2 ||d_t||_2^2$$

Using the bound on $||d_t||_2$ (which we get from the normalization condition on the weak learner class and triangle inequality), we get that

$$R(f_{t-1}) - R(f_t) \geq \alpha_t (G(f_{t-1}) - \tau \varepsilon_t) - 2L\tau^2 \alpha_t^2$$

■

Next we lower bound the progress that the *best* model in the weak learner class could make, in terms of the current loss gap with the anchor model and our chosen atomic norm bound $\tau$:

**Lemma 27** *(FW Correlation Lower Bound w.r.t Weak Learning Anchor Gap) For a given $f$ from our algorithm's iterates $(f_t)$, we have that*

$$M(f) \geq \frac{R(f) - R(\mathcal{K}_\tau)}{2\tau}$$

**Proof** Let $f^* = \arg\min_{f \in \mathcal{K}_\tau} R(f)$. We know by convexity that

$$\mathcal{L}(y, f(x)) - \mathcal{L}(y, f^*(x)) \leq \langle \nabla_p \mathcal{L}(y, f(x)), f(x) - f^*(x) \rangle$$

Taking expectations and by an application of Hölder's inequality and triangle inequality, we get that

$$R(f) - R(\mathcal{K}_\tau) \leq ||\nabla R(f)||_{\mathcal{A}^*}(||f||_{\mathcal{A}} + ||f^*||_{\mathcal{A}}).$$

As shown below, by the definition of the dual norm, atomic norm, linearity of the inner product, normalization of the weak learner class, and the budget $\tau$,

$$\begin{aligned} ||\nabla R(f)||_{A^*} &= \sup_{||c||_{\mathcal{A}} \leq 1} |\langle \nabla R(f), c \rangle| \\ &= \sup_{c \in \text{conv}(\mathcal{C})} |\langle \nabla R(f), c \rangle| \\ &= \sup_{c \in \mathcal{C}} |\langle \nabla R(f), c \rangle|. \end{aligned}$$

Therefore, we get that

$$R(f) - R(\mathcal{K}_\tau) \leq 2\tau M(f).$$

Rearranging this expression gives the final bound. ■

Next we derive a recurrence relation between the error gap of the model at iteration $t$ and the best model in the restricted span of the weak learner class.

**Lemma 28 (FW Gap Recurrence Toward $R(\mathcal{K}_\tau)$)** *Assume $\mathcal{L}$ is $L$-smooth in its second argument. Let $E_t := R(f_t) - R(K_\tau)$. Then for all $t \geq 1$,*

$$E_{t-1} - E_t \geq \alpha_t(G(f_{t-1}) - \tau\varepsilon_t) - 2L\tau^2\alpha_t^2 \geq \alpha_t(E_{t-1} - \tau\varepsilon_t) - 2L\tau^2\alpha_t^2$$

**Proof** By $L$-smoothness and the FW update $f_t = f_{t-1} + \alpha_t d_t$ with $d_t = g_t - f_{t-1}$, Lemma 26 gives

$$R(f_{t-1}) - R(f_t) \geq \alpha_t(G(f_{t-1}) - \tau\varepsilon_t) - 2L\tau^2\alpha_t^2$$

Subtract and add $R(K_\tau)$ to obtain the first inequality:

$$E_{t-1} - E_t \geq \alpha_t(G(f_{t-1}) - \tau\varepsilon_t) - 2L\tau^2\alpha_t^2$$

Let $f^* = \arg\min_{f \in \mathcal{K}_\tau} \mathbb{E}[\mathcal{L}(y, f)]$. By convexity, $E_{t-1} = R(f_{t-1}) - R(f^*) \leq \langle \nabla R(f_{t-1}),\ f_{t-1} - f^* \rangle \leq G(f_{t-1})$, so

$$E_{t-1} - E_t \geq \alpha_t(E_{t-1} - \tau\varepsilon_t) - 2L\tau^2\alpha_t^2$$

which is the second inequality. ■

We will use this recurrence relation to bound the error gap for the model at iterate $t$.

**Lemma 29 (FW Anchor Gap Upper Bound)** *For all* $t \geq 1$,

$$R(f_t) - R(K_\tau) \;\leq\; \frac{8L\tau^2}{t+1} + \frac{2\tau}{(t+1)} \sum_{j=1}^{t} \varepsilon_t.$$

**Proof** From Lemma 28 we have the recursion

$$E_{t-1} - E_t \geq \alpha_t(E_{t-1} - \tau\varepsilon_t) - 2L\tau^2\alpha_t^2$$

which is equivalent to

$$\begin{aligned} E_t &\leq E_{t-1} - \alpha_t(E_{t-1} - \tau\varepsilon_t) + 2L\tau^2\alpha_t^2 \\ &= (1-\alpha_t)E_{t-1} + \alpha_t\tau\varepsilon_t + 2L\tau^2\alpha_t^2 \end{aligned}$$

We use the convention that $[k] = \{1, ..., k\}$. Call $C = 4L\tau^2$ and substitute in $\alpha_t$, then we get

$$\begin{aligned} E_t &\leq \frac{t-1}{t+1}E_{t-1} + \frac{2}{t+1}\tau\varepsilon_t + 2L\tau^2\Big(\frac{2}{t+1}\Big)^2 \\ &= \frac{t-1}{t+1}E_{t-1} + \frac{2C}{(t+1)^2} + \frac{2}{t+1}\tau\varepsilon_t \end{aligned}$$

Define $S_t = \tau\sum_{j=1}^{t} j\varepsilon_j$. Then, we will prove via induction that for all $t \geq 1$,

$$E_t \leq \frac{2Ct + 2S_t}{t(t+1)}$$

First, for the base case consider $t = 1$. We have from the recurrence relation that $E_1 \leq 0 + \frac{C}{2} + \tau\varepsilon_t \leq C + \tau\varepsilon_t$. Next, suppose $E_t \leq \frac{2Ct+2S_t}{t(t+1)}$, we will prove the same relationship holds for $E_{t+1}$.

$$\begin{aligned} E_{t+1} &\leq \frac{t}{t+1}E_t + \frac{2C}{(t+2)^2} + \frac{2}{t+2}\tau\epsilon_{t+1} \\ &\leq \frac{t}{t+2}\frac{2Ct+2S_t}{t(t+1)} + \frac{2C}{(t+2)^2} + \frac{2}{t+2}\tau\epsilon_{t+1} \\ &= \frac{2Ct+2S_t}{(t+1)(t+2)} + \frac{2C}{(t+2)^2} + \frac{2}{t+2}\tau\epsilon_{t+1} \\ &= \frac{2C}{t+2}\Big(\frac{t}{t+1} + \frac{1}{t+2}\Big) + \frac{2S_{t+1}}{(t+1)(t+2)} \\ &\leq \frac{2C}{t+2}\Big(\frac{t}{t+1} + \frac{1}{t+1}\Big) + \frac{2S_{t+1}}{(t+1)(t+2)} \\ &\leq \frac{2C}{t+2}\Big(\frac{t+1}{t+1}\Big) + \frac{2S_{t+1}}{(t+1)(t+2)} \\ &= \frac{2C(t+1) + 2S_{t+1}}{(t+1)(t+2)}. \end{aligned}$$

Therefore, we have that

$$E_t \leq \frac{8L\tau^2}{t+1} + \frac{2\tau}{t(t+1)} \sum_{j=1}^{t} j\varepsilon_j$$

Therefore,

$$E_t \leq \frac{8L\tau^2}{t+1} + \frac{2\tau}{(t+1)} \sum_{j=1}^{t} \varepsilon_j$$

■

**Theorem 30** *(FW Gradient Boosting Agreement Bound) Fix any $\mathcal{L}$ that is $L$-smooth and $\mu$-strongly convex. Let $f_1, f_2$ be the output of any two runs of Algorithm 4 parameterized with the same $\tau, k, \mathcal{C}$ such that the sequence of SQ oracle errors are $\{\varepsilon_t, \varepsilon'_t\}_{t\in[k]}$ respectively. Let $f^* = \arg\min_{f\in\mathcal{K}_\tau} R(f)$. Then, we have that*

$$D(f_1, f_2) \leq \frac{64L\tau^2}{\mu(k+1)} + \frac{8\tau}{\mu(k+1)}\Big(\sum_{j=1}^{k} \varepsilon_j + \sum_{j=1}^{k} \varepsilon'_j\Big)$$

**Proof** Since $f^\star$ minimizes $\mathbb{E}[\mathcal{L}(y, f(x))]$ over the convex set $K_\tau$, by first-order optimality we have the inequality

$$\mathbb{E}[\langle \nabla\mathcal{L}(y, f^\star(x)),\, z(x) - f^\star(x)\rangle] \geq 0 \qquad \forall z \in K_\tau.$$

Combining this with $\mu$–strong convexity of $\mathcal{L}$ gives

$$\mathbb{E}[\mathcal{L}(y, g(x))] \geq \mathbb{E}[\mathcal{L}(y, f^\star(x)) + \langle \nabla\mathcal{L}(y, f^\star(x)),\, g(x) - f^\star(x)\rangle + \tfrac{\mu}{2}\|g(x) - f^\star(x)\|_2^2)].$$

Since $\mathcal{K}_\tau$ is convex, the midpoint $\frac{1}{2}(f_1 + f_2)$ lies in $\mathcal{K}_\tau$. Applying Lemma 24 with $\mathcal{H} = \mathcal{K}_\tau$ gives

$$D(f_1, f_2) \leq \tfrac{4}{\mu}\big(R(f_1) - R(f^*)\big) + \tfrac{4}{\mu}\big(R(f_2) - R(f^*)\big).$$

Finally, applying Lemma 29 to both error gap terms gives us the final bound. ■

### E.3. Neural Networks

We next state the midpoint-anchor analogue of our neural-network and regression-tree agreement bounds for multi-dimensional $\mu$-strongly convex losses.

**Theorem 31 (Agreement from midpoint closure)** *Assume $\mathcal{L}$ is $\mu$-strongly convex.*

1. *If $f_1, f_2 \in \mathrm{NN}_n$ satisfy $R(f_i) \leq R(\mathrm{NN}_n) + \varepsilon$ for $i \in \{1, 2\}$, then*

$$D(f_1, f_2) \;\leq\; \tfrac{8}{\mu}\big(R(\mathrm{NN}_n) - R(\mathrm{NN}_{2n}) + \varepsilon\big).$$

2. *If $f_1, f_2 \in \mathsf{Tree}_d$ satisfy $R(f_i) \leq R(\mathsf{Tree}_d) + \varepsilon$ for $i \in \{1, 2\}$, then*

$$D(f_1, f_2) \;\leq\; \tfrac{8}{\mu}\big(R(\mathsf{Tree}_d) - R(\mathsf{Tree}_{2d}) + \varepsilon\big).$$

**Proof** We prove each part by applying Lemma 24 at the appropriate midpoint-closed level.

*Part (1).* Let $f_1, f_2 \in \mathrm{NN}_n$ and define $\bar{f} := \frac{1}{2}(f_1 + f_2)$. By midpoint closure (Lemma 12), we have $\bar{f} \in \mathrm{NN}_{2n}$. Applying Lemma 24 with $\mathcal{H} = \mathrm{NN}_{2n}$ gives

$$D(f_1, f_2) \le \tfrac{4}{\mu}\big(R(f_1) - R(\mathrm{NN}_{2n})\big) + \tfrac{4}{\mu}\big(R(f_2) - R(\mathrm{NN}_{2n})\big).$$

Using the assumptions $R(f_i) \le R(\mathrm{NN}_n) + \varepsilon$ for $i \in \{1, 2\}$, we obtain

$$R(f_i) - R(\mathrm{NN}_{2n}) \le R(\mathrm{NN}_n) - R(\mathrm{NN}_{2n}) + \varepsilon.$$

Substituting this bound for both $i = 1, 2$ yields

$$D(f_1, f_2) \le \tfrac{8}{\mu}\big(R(\mathrm{NN}_n) - R(\mathrm{NN}_{2n}) + \varepsilon\big),$$

as claimed.

*Part (2).* The proof is identical with $\mathsf{Tree}_d$ in place of $\mathrm{NN}_n$. Let $f_1, f_2 \in \mathsf{Tree}_d$ and $\bar{f} := \frac{1}{2}(f_1 + f_2)$. By midpoint closure (Lemma 14), $\bar{f} \in \mathsf{Tree}_{2d}$. Applying Lemma 24 with $\mathcal{H} = \mathsf{Tree}_{2d}$ gives

$$D(f_1, f_2) \le \tfrac{4}{\mu}\big(R(f_1) - R(\mathsf{Tree}_{2d})\big) + \tfrac{4}{\mu}\big(R(f_2) - R(\mathsf{Tree}_{2d})\big).$$

Using $R(f_i) \le R(\mathsf{Tree}_d) + \varepsilon$ for $i \in \{1, 2\}$ and substituting yields

$$D(f_1, f_2) \le \tfrac{8}{\mu}\big(R(\mathsf{Tree}_d) - R(\mathsf{Tree}_{2d}) + \varepsilon\big).$$

■

## Appendix F. Cross-Entropy Loss

Here, we provide an analysis of how to use our anchoring method to bound prediction disagreement for cross-entropy loss, which is convex but not strongly convex. For cross-entropy loss, rather than using the arithmetic midpoint in probability space, we use the KL centroid, equivalently the normalized geometric mean. This is the two-predictor specialization of the KL/Bregman ambiguity decomposition for ensemble predictors discussed by Wood et al. (2023). For completeness, we prove the identity directly below.

**Lemma 32 (KL-centroid identity for cross-entropy loss)** *Let $q_1, q_2 \in \Delta_m$ have strictly positive coordinates, and define their KL centroid $\bar{q} \in \Delta_m$ by*

$$\bar{q}_c := \frac{\sqrt{q_{1,c} q_{2,c}}}{\sum_{j=1}^{m} \sqrt{q_{1,j} q_{2,j}}}.$$

*For a label distribution $y \in \Delta_m$, let*

$$H(y, q) := -\sum_{c=1}^{m} y_c \log q_c$$

*denote cross-entropy loss. Then*

$$\frac{1}{2}\big(H(y, q_1) + H(y, q_2)\big) - H(y, \bar{q}) = \frac{1}{2}\left(\mathrm{KL}(\bar{q} \| q_1) + \mathrm{KL}(\bar{q} \| q_2)\right).$$

**Proof** Let

$$Z := \sum_{j=1}^{m} \sqrt{q_{1,j} q_{2,j}}.$$

By definition of $\bar{q}$,

$$\log \bar{q}_c = \frac{1}{2} \log q_{1,c} + \frac{1}{2} \log q_{2,c} - \log Z.$$

Therefore, for any $y \in \Delta_m$,

$$\begin{aligned} H(y, \bar{q}) &= -\sum_{c=1}^{m} y_c \log \bar{q}_c \\ &= \frac{1}{2} H(y, q_1) + \frac{1}{2} H(y, q_2) + \log Z, \end{aligned}$$

where we used $\sum_c y_c = 1$. Hence

$$\frac{1}{2}\big(H(y, q_1) + H(y, q_2)\big) - H(y, \bar{q}) = -\log Z.$$

On the other hand,

$$\begin{aligned} \frac{1}{2} \left(\mathrm{KL}(\bar{q} \| q_1) + \mathrm{KL}(\bar{q} \| q_2)\right) &= \sum_{c=1}^{m} \bar{q}_c \left( \log \bar{q}_c - \frac{1}{2} \log q_{1,c} - \frac{1}{2} \log q_{2,c} \right) \\ &= \sum_{c=1}^{m} \bar{q}_c (-\log Z) = -\log Z. \end{aligned}$$

Combining the two equalities proves the claim. ■

**Corollary 33 (Disagreement via the KL-centroid anchor)** *Let $q_1, q_2 : X \to \Delta_m$ be predictors with strictly positive coordinates, and define their pointwise KL centroid $\bar{q} : X \to \Delta_m$ by*

$$\bar{q}_c(x) := \frac{\sqrt{q_{1,c}(x) q_{2,c}(x)}}{\sum_{j=1}^{m} \sqrt{q_{1,j}(x) q_{2,j}(x)}}.$$

*Let*

$$\mathrm{CE}(q) := \mathbb{E}_{(x,y)\sim P}[H(y, q(x))]$$

*and let $R(H) := \inf_{q \in H} \mathrm{CE}(q)$. If $\bar{q} \in H$, then*

$$\begin{aligned} \mathbb{E}_x \left[ \|q_1(x) - q_2(x)\|_2^2 \right] &\le 4\, \mathbb{E}_x \left[ \mathrm{KL}(\bar{q}(x) \| q_1(x)) + \mathrm{KL}(\bar{q}(x) \| q_2(x)) \right] \\ &\le 8 \left[ \frac{1}{2}\big(\mathrm{CE}(q_1) + \mathrm{CE}(q_2)\big) - R(H) \right]. \end{aligned}$$

**Proof** Applying Lemma 32 pointwise and taking expectations gives

$$\mathbb{E}_x \left[ \frac{1}{2} \left( \mathrm{KL}(\bar{q}(x) \| q_1(x)) + \mathrm{KL}(\bar{q}(x) \| q_2(x)) \right) \right] = \frac{1}{2}\big(\mathrm{CE}(q_1) + \mathrm{CE}(q_2)\big) - \mathrm{CE}(\bar{q}).$$

Since $\bar{q} \in H$, we have $\mathrm{CE}(\bar{q}) \geq R(H)$, and hence

$$\mathbb{E}_x \left[\mathrm{KL}(\bar{q}(x)\|q_1(x)) + \mathrm{KL}(\bar{q}(x)\|q_2(x))\right] \leq 2\left[\frac{1}{2}\big(\mathrm{CE}(q_1) + \mathrm{CE}(q_2)\big) - R(H)\right].$$

It remains to relate the KL ambiguity term to prediction disagreement. By Pinsker's inequality,

$$\|\bar{q}(x) - q_i(x)\|_1^2 \leq 2\mathrm{KL}(\bar{q}(x)\|q_i(x)) \qquad i \in \{1,2\}.$$

Thus, by the triangle inequality, $(a+b)^2 \leq 2a^2 + 2b^2$, and $\|v\|_2 \leq \|v\|_1$,

$$\begin{aligned}\|q_1(x) - q_2(x)\|_2^2 &\leq \|q_1(x) - q_2(x)\|_1^2 \\ &\leq 2\|\bar{q}(x) - q_1(x)\|_1^2 + 2\|\bar{q}(x) - q_2(x)\|_1^2 \\ &\leq 4\left(\mathrm{KL}(\bar{q}(x)\|q_1(x)) + \mathrm{KL}(\bar{q}(x)\|q_2(x))\right).\end{aligned}$$

Taking expectations and combining the two displayed inequalities proves the claim. ■

**Lemma 34 (KL-centroid closure for softmax logit classes)** *Let $(\mathcal{Z}_n)_{n\geq 1}$ be a nested sequence of logit classes $z : X \to \mathbb{R}^m$ satisfying midpoint closure: for every $z_1, z_2 \in \mathcal{Z}_n$, the pointwise midpoint*

$$\bar{z}(x) := \frac{z_1(x) + z_2(x)}{2}$$

*lies in $\mathcal{Z}_{2n}$. Define the induced probability classes*

$$\mathcal{F}_n := \{\mathrm{softmax}(z) : z \in \mathcal{Z}_n\}.$$

*Then for every $q_1, q_2 \in \mathcal{F}_n$, their pointwise KL centroid lies in $\mathcal{F}_{2n}$.*

**Proof** Since $q_i \in \mathcal{F}_n$, there exist logits $z_i \in \mathcal{Z}_n$ such that $q_i = \mathrm{softmax}(z_i)$ for $i \in \{1,2\}$. By midpoint closure of the logit class,

$$\bar{z} := \frac{z_1 + z_2}{2} \in \mathcal{Z}_{2n}.$$

We claim that $\mathrm{softmax}(\bar{z})$ is the KL centroid of $q_1$ and $q_2$. Indeed, for each coordinate $c$,

$$\sqrt{q_{1,c}(x)q_{2,c}(x)} = \frac{\exp\left(\frac{z_{1,c}(x)+z_{2,c}(x)}{2}\right)}{\sqrt{\left(\sum_j \exp(z_{1,j}(x))\right)\left(\sum_j \exp(z_{2,j}(x))\right)}}.$$

The denominator is independent of $c$, so it cancels when normalizing over coordinates. Therefore,

$$\bar{q}(x) = \mathrm{softmax}\left(\frac{z_1(x) + z_2(x)}{2}\right) = \mathrm{softmax}(\bar{z}(x)).$$

Since $\bar{z} \in \mathcal{Z}_{2n}$, we have $\bar{q} \in \mathcal{F}_{2n}$. ■

**Lemma 35 (Local learning-curve bound for softmax cross-entropy)** *Let $(\mathcal{Z}_n)_{n\geq 1}$ be a nested sequence of logit classes satisfying the midpoint-closure condition of Lemma 34, and let*

$$\mathcal{F}_n := \{\mathrm{softmax}(z) : z \in \mathcal{Z}_n\}.$$

*Fix $n \geq 1$ and suppose $q_1, q_2 \in \mathcal{F}_n$ satisfy*

$$\mathrm{CE}(q_i) \leq R(\mathcal{F}_n) + \varepsilon \qquad \textit{for } i \in \{1, 2\}.$$

*Then*

$$\mathbb{E}_x \left[\|q_1(x) - q_2(x)\|_2^2\right] \leq 8 \left(R(\mathcal{F}_n) - R(\mathcal{F}_{2n}) + \varepsilon\right).$$

**Proof** Let $\bar{q}$ denote the pointwise KL centroid of $q_1$ and $q_2$. By Lemma 34, we have $\bar{q} \in \mathcal{F}_{2n}$. Applying Corollary 33 with $H = \mathcal{F}_{2n}$ gives

$$\mathbb{E}_x \left[\|q_1(x) - q_2(x)\|_2^2\right] \leq 8 \left[\frac{1}{2}\big(\mathrm{CE}(q_1) + \mathrm{CE}(q_2)\big) - R(\mathcal{F}_{2n})\right].$$

Using $\mathrm{CE}(q_i) \leq R(\mathcal{F}_n) + \varepsilon$ for both $i \in \{1, 2\}$ gives

$$\mathbb{E}_x \left[\|q_1(x) - q_2(x)\|_2^2\right] \leq 8 \left(R(\mathcal{F}_n) - R(\mathcal{F}_{2n}) + \varepsilon\right).$$

■